Identifying the Dimensions of Data Science as a Space

Identifying the Dimensions of Data Science as a Space

Ruxin Huang[1], Xinlin Cai[2], Yuhang Liu[3]*

[1] Parkway Central High School, Missouri, US, huangr0730@gmails.com

[2]University of California, Los Angeles, US, xinlin.cai.melody@gmail.com

[3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, liuyuhang@ict.ac.cn

Abstract: An increasing number of data with great potential value are produced continuously, which exceeds the capacity of the existing computing system and promotes the generation of new basic theories about big data and its processing. Such problems as the relationship between big data and small data and that between big data and the existing scientific system need to be accurately solved by the academic circle through the establishment of data science. However, data science is still in its infancy and the basic dimensions of data science need to be identified. In this paper, five categories demanding for establishment of data science are summarized. A new way to understand big data that gives the most important speculation and research object of data science from a global perspective is developed.

1. Introduction

Big data, a new concept that was recently introduced, has captured a lot of scholars' attention and developed rapidly in these years. But its definition is still largely ambiguous, and there are dozens of different definitions. The characteristic of big data is considered as 3V (large volume, variety, velocity), 4V (added another dimension, veracity), and 5V (added value), respectively. In addition, others added characteristics such as "not pursuing cause and effect" and "a by-product of digital interaction" to previous models. These characteristics are a sectional view of big data and require more research in terms of clarity and universality (e.g. how big the volume makes it big data, whether it is necessary or possible for data to pursue cause and effect, and whether or not all big data by-product).

However, many issues related to big data remain unresolved, including its relationship to existing academic systems. According to the National Science Foundation, data science is still in its infancy. A scientific system of big data has not yet been established globally [4]. As big data seems to hold great potential and may play an important role in future scientific research, it has become one of the most competitive research fields in the world.

The system of data science has not yet been developed. There are a variety of definitions of data science. From an engineering perspective, data science should focus on translating data into action systems. Such understanding would contradict a more common belief in science that "only focusing on process and system is not enough" and a "theorized, formalized, quantified, completed" system must be established. To define each term: "theorized" means that the system must be general and not specific; "formalized" means the system must be rigorous; "quantified" means the system's quantity is measurable; and "completed" means the data science consist of studies with not only big data but also all others. Since big data is a lot more difficult than small data to deal with, in order to empathize with

big data, we call the study with big data as "big data science", and the respective system as a "big data science system". This paper provided the following facets of "big data" from a dialectical perspective:

- We provided multiple sets of the unity of opposites. This is a new approach to understand big data. Instead of focusing on parts, we tried to define big data through a complete "full-view" to find the main object to study within data science;
- We found out the similarity between mineral purification and big data analysis process through using horizontal comparison;
- We used an analogy of "three cosmic velocities" to portray the potential of big data;
- We provided a formal mathematical definition of big data, including three factors that determined the standard of big data;
- We studied multiple different views in both academic and industry fields, such as Schönberger and Pearson's different views on correlation and cause and effect.

In the following paragraphs, we will introduce and analyze multiple contradictory categories of the big data science system, and then discuss the important factors that should be paid attention to in constructing the system.

2. Dimension I: Data vs. Knowledge vs. Anticipation vs. Insight vs. Wisdom

Data is defined as "facts or information usually used to calculate, analyze, or plan something" by Merriam Webster dictionary. As defined, data is objective before it is used to accomplish certain tasks. By processing data, we can effectively turn it into information, knowledge, anticipation, insight, and wisdom, respectively. In the process, the term of insight can be confusing and abstract, but according to Freud, "insight is the process of transforming the unconscious into the conscious." The process of analyzing the data appears to be a similar process to that of purifying the minerals. In both cases, the more steps that it involves, the fewer products can be obtained. Based on the number of products obtained in each step, the whole process from data analysis to mineral purification can be seen as a pyramid structure. Although both processes are in completely different fields, the pyramid pattern must have represented something they have in common.

3. Dimension II: All Data vs. Big Data vs. Small Data

With the help of technological inventions such as telescopes and satellites, human beings are able to perceive the macro world. By using sensors and microscopes, they are able to understand the world at the micro-level. However, the source of big data includes, but does not only include data from these technologies. With the development of big data, people have a deeper understanding of the world, especially the understanding of time and space. To some extent, humans are able to expand their capabilities as they learn more about big data. An analogy of this idea can be found in the physics theory "cosmic velocity". Cosmic velocity refers to the speed at which an object can become a satellite of the Earth (orbiting the Earth) or a satellite of the sun (orbiting the sun) or escape the solar system due to a gravitational field. These three different escape velocities are called the first cosmic velocity, the second cosmic velocity and the third cosmic velocity. Similarly, we can define that as we learn more and more about data, we can acquire our own first, second, and third perceptual abilities within the framework of ethics and rules to achieve transcendence.

All data is the upper limit of big data. In fact, the size of data is often relative. Similar to the concept of relative velocity in physics, the velocity of an object can only be determined when compared with another object. Therefore, when compared with other data sets, the size of the data set can be determined. 1 GB data is not necessarily big data (e.g. 1 GB of genetic sequence data is not considered as big data using computer booster), whereas 1 MB data might be big data (e.g. athlete data). As a result, the standard of determining whether the data is "big" or "small" is fundamental in the problem-solving. Most studies agree that the size of the data can be defined from the following three dimensions:

The first is about the ability of the data processing system. When all data is big data, especially when there are technical advantages in data transmission and analysis, the standard of big data is

higher. To put it simply, a more productive factory (Factory A) produces more products than a less productive factory (Factory B). In this case, the number of rough materials that B Factory B cannot process can be regarded as the "big data" of the factory. However, for Factory A, the same amount of rough materials is not necessarily "big data" because Factory A has a higher capacity in material processing.

The second is about the difficulty of any fundamental steps that exist in the life cycle of the data set. The difficulty of any steps in the data life cycle is also an important determination. As Data Observation Network for Earth stated, “The data life cycle provides an overview of the stages involved in successful management and preservation of data for use and reuse.” For example, when the amount of data collected is close to the size of all data, the data might be considered as big data. In this scenario, the difficulty is not about the insufficient ability of the system but about the difficulty in data collection.

The third is about user demand. Another important dimension of determining the size of a data set is the users’ demand. The higher the user demand, the lower the standard considered big data. For example, when the user is waiting for data processing during his service provided by the system, the shorter the data processing step is, the more satisfying the user experience, which we consider to be a “more demanding choice”. The difficulty of processing data under 1-second demand is greater than that under a 10-second demand. Therefore, in the case of high user demand, data is more easily regarded as big data.

To sum up the above three factors affecting the standard of big data, we could simply understand big data as “difficult data”, that is, “data sets with some difficult steps in the data life cycle under the system”.

In contrast to big data, small data often exists for a purpose. The questionnaire is an example of small data that focuses on a specific part of all the data sets. Big data is often unintentional and automatic, including data collected automatically by sensors and computer programs formed unintentionally. Both examples of big data demonstrate that the formation of big data is most likely spontaneous.

4. Dimension III: Popper’s three worlds

British philosopher Karl Popper introduced a concept that involved three interacting worlds in his lecture in 1978. He described that the world can be split into three categories: the world of physical objects (World I), the world of mental processes (World II), and the world of objective knowledge (World III). Big data generally belongs to world III but can be originated in all these three worlds.

Respective to World I, data is mainly collected by experimental instruments. Correspondingly, “scientific big data” and “IoT big data” are two subjects most studied by scientists. Data collected through traditional scientific experiments are known as "scientific big data", especially in data-rich disciplines such as genomics, proteomics and astrophysics. Scientific instruments are essentially data collectors by purpose. European large Hadron Collider (commonly known as LHC) is the world's largest high-energy particle accelerator. Another example is the five-hundred-meter aperture spherical radio telescope (FAST), which was established by Chinese scientists and can produce a massive amount of data per second.

Respective to World II, human activities, including manufacturing and everyday life, generate a massive amount of data. As enterprise information systems record a great deal of information about commodities, customers, and suppliers, it is difficult to imagine how an enterprise could operate effectively without an entire system. Without such a system, any business activity would be much less efficient and could even result in economic stagnation. Wal-Mart, for example, has thousands of stores around the world and sells billions of products every day. Business data exceeds PB levels and continues to grow. In addition to businesses, human creates tones of data in daily life. One example is “social media data”. The data created by the files, pictures, and texts are vast in variety, massive in amount, and in a chain. Hence human activities play a significant role in data production.

Respective to World III, human creates knowledge. Papers, books, sheet music, and many more are data of human thoughts. All data, from an epistemological point of view, becomes data of the human mind for human decision making.

5. Dimension IV: Correlation vs. Cause and Effect

If a mutual relationship is observed when comparing changes in two variables, the correlation is confirmed. When two variables are correlated, the relationship is not necessarily causal and vice versa. There may be seven different correlations between two events, but only four are causal.

Richard Hamming, a mathematician who received the ACM A.M. Turing Award in 1968, believes that "The purpose of computation is insight, not numbers." The key to "insight" is to discover the relationship between objects. Science listed "How will big pictures emerge from a sea of biological data" as one of the biggest challenges that they are trying to overcome [6]. In the field of big data, great progress has been made in correlation research [1]. However, the study on the causality seems to be less focused. Whether we should pay more attention to causality between data remains a matter of debate. While some people such as Viktor Mayer-Schönberger, the professor of Internet Governance and Regulation at Oxford, believe that "Predictions based on correlations lie at the heart of big data" [3], others believe the study on the causality between data is the key to success. Judea Pearl, known as "the father of Bayesian network", believes the casual relationship is the right path to the real intelligence machines [7].

Connectionism implies correlation while symbolism implies cause and effect. Machine learning has gone through connectionism, symbolism and regression connectionism respectively. After experiencing "the negation of negation", the third attempt finally succeeded. The third attempt at connectionism was more successful than the first, especially with algorithms. However, it is important to note that "the negation of negation" has not yet been completed. In other words, although the third attempt negates the second attempt at symbolism, it has not yet fully absorbed the advantages of symbolism. Connectionism is harder to explain than symbolism. The "negation of negation" will be completed once the interpretability of connectionism is fulfilled.

6. Dimension V: Possibility vs. Necessity

The purpose of data analysis is to maximize the efficiency with which we interpret information in our lives. It is important to note that some data are too large to be accepted without analysis. One example is big data. While the data may contain the information we want, too much information we need and do not need prevents us from getting the information we want in the short term. As Kevin Kelly, the executive editor of Wired Magazine, believes, "the only factor becoming scarce in a world of abundance is human attention." [8] For the process of purifying minerals, the amount of wanted product from the original resources might be limited; for the process of analyzing data, the amount of data that can be analyzed in an ideal time interval is also limited. The final product is always as simple as possible, but contains all the required information. Such products are exactly what we need in our fast-paced life and provide an effective explanation for why we need to study data science, especially the science related to big data.

We also observed that the term with a prefix of "meta-" had important applications in data analysis. For instance, the term "metadata" is defined as "data that provides information about other data"; similarly, "metaphysics" provides reasonable explanations of physics scenarios; "metamathematics" is the field of study of the structure and formal properties of mathematical and similar formal systems. The process of turning data into information, knowledge, anticipation, insight, and wisdom, as mentioned in the previous paragraphs, can be understood as a progressive metadata set. Similar to the second derivative analysis method of the first derivative, each step in the data analysis process can be viewed as the metadata of the previous step.

7. Discussions

We believe that a big data system must have compatible characteristics, not "not" but "not only, but also". The characteristic can be reflected through the following ten aspects: (1) The data is not collected from a single source, but from multiple sources. (2) Focusing not only on structural data but also on non-structural data. (3) Not only focusing on the accuracy of data but also allowing for confounding. (4) That is not to say that causation does not matter. Instead, focus on causation and correlation in general. (5) That is not to say that small volumes are not worth studying. On the contrary, both small data and big data are worth studying (6) The big data is to support the real economy rather than replace it. (7) Not to negate the "Pareto's principle", and we should also focus on secondary majority data rather than primary minority data.

We also found out that "big data" was essentially "difficult data". In addition, we concluded that the methods to overcome the difficulties we listed had several basic characteristics: The first is that the application of modern technologies serves as a symbol of this new era. The second is the basic application of the computer circuit principle. Thirdly, the self-reasoning function is introduced, though it has yet been fully studied. The fourth is the improvement made on chip performance (e.g. increase the number of transistors) and supercomputer property (increase the number of calculation nodes). The fifth is the principle of the memory hierarchy and average memory access time (AMAT), and the sixth is the principle of operating system adjustment and the "tag recommendation system" [9].

8. Related Work

The investment of data science is still in a new stage. Due to the limited adaptability of the traditional data system, the big data science system has not yet been established [4]. Guoliang Chen studied big data computation theory and believed that we "had to completely invest in the scientific problems that brought to us from big data itself such as whether big data was a science and what key aspects we needed to know. Weinan E suggested that "data science mainly includes two aspects: using data methods to study science or using scientific methods to study data". The former aspect includes bioinformatics, astroinformatics, and the Digital Earth, etc. and the latter aspect includes statistics, data mining, and database, etc. All of these disciplines are important components of data science. Once we put all these disciplines together, we will be able to see the full picture of data science. The five dimensions we mentioned in this article focused on the essential characteristics of big data, and highlighted data science as "science of data" instead of "data and science".

9. Conclusions

There has always been a controversy in academia about whether data science is a science or what big data is. Big data system will provide important new theory and valuable practical value. Five dimensions of "to be built" system are proposed. More dimensions can be added upon to the five we proposed. In the end, we hope it can help us take the strategic highlands of data science.

References

[1] Reshef D, Reshef Y, Finucane H K, et al. Detecting Novel Association in Large Data Sets. Science, 2011, 334:1518-1524.
[2] Charis Anderson. The End of Theory: The Data Deluge Makes Scientific Method Obsolete. Wired. 2008,16(7).
[3] Mayer Schnberger V, Cukier K. Big Data: A Revolution that will Transform How We Live, Work, and Think. John Murray, 2013.
[4] Berman F, Stodden V, Szalay A S, et al. Communications of the ACM, Realizing the Potential of Data Science, 2018, 61(4):67~72
[5] Fortnow, Lance. The Golden Ticket: P, NP, and the Search for the Impossible. Princeton University Press, 2013.
[6] Kennedy D, Norman C. What don't we know? Science, 2005, 309(5731):75.
[7] Judea Pearl, Dana Mackenzie. The Book of Why: The New Science of Cause and Effect. Hachette Book Group. 2018.

[8] Kelly K. The inevitable: understanding the 12 technological forces that will shape our future. Journal Technology, Architecture, Design 2016.
[9] Jiuyue Ma, Xiufeng Sui, Ninghui Sun, et al. 2015. Supporting Differentiated Services in Computers via Programmable Architecture for Resourcing-on-Demand (PARD). In Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems. ACM, Istanbul, Turkey, 131–143.