RESEARCH ARTICLE

Big Data Thinking and Its Biomedical Application

Petar Melih INAL, Nikhil Vishnu
Baekje University Academy of Science Certain Arts National Laboratory Office, Biological Interest

Abstract: Big data thinking gradually rise with the coming era of big data. Big data characteristics could be summarized with 4V: volume, variety, velocity and value. The characteristics of big data thinking could be summed up in integrity, fault tolerance, correlation and intelligence. These characteristics were also the primary differences between big data thinking and small data thinking. The application of big data thinking in biomedical field became more and more widely, and the most commonly used was NCBI database. The process of mining valuable information in NCBI database was big data thinking. And the rise of Meta analysis and TCGA database would illustrate the huge application value of big data thinking in the biomedical field.

Keywords: Big data; Big data thinking; Biomedicine

American Futurist Alvin Toffler predicted that the 21st century will enter the information age, the rapid development of computers and applications in various fields of research, this prediction will soon come true. At the same time, with the emergence of the Internet of Things, the Internet and cloud computing technology, the era of big data has come, and is exploding at an unprecedented rate, big data will become a new revolution, sweeping all areas. Back in 2008 Nature presents big data, and then Science discusses how big data will play an important role in scientific research (Staff, 2011). The famous Futurist Alvin Toffler once said that big data is an important part of the "third wave", big data will soon sweep over.

In 2011, Manyika *et al*. published a detailed report on big data, exploring the key technologies, areas involved and the impact of big data on people's daily lives (Manyika *et al*., 2011, http://www.mck insey.com/business-functions/digital-mck-insey/our-insights/big-data-the-next-frontier-for-innovat ion)). One of the reasons why Obama won the presidential election in 2012 was that he used big data mining technology, and in the same year he launched the Big Data Research and Development Program in the name of the U.S. government, upgrading the Big Data Strategy to a national strategy to improve U.S. research, education and the nation. Security capabilities, and big data are defined as "the new oil of the future" (Gantz and Reinsel, 2012, https://www.emc.com/leadership/digital-universe/2012 iview/i ndex.htm).

Big data is so hot that it has already been in agriculture, network, medical treatment, and delivery. Widely used in communication, news and finance, and penetrated into our work and life. Undoubtedly, the era of big data has come, and slowly changing people's behavior and thinking, people can not resist. Big data contains huge economic benefits, and its core value is to store and analyze massive data to discover and tap potential value. At the same time, large data also brings us many challenges, such as multi-source heterogeneous data, rapid growth, wide distribution, data first followed by a pattern and other characteristics make large data management has certain challenges: data management and analysis, data privacy and security, high data processing energy consumption, processing equipment requirements, etc. So how do we know big data? What changes does big data bring to our thinking? What is the difference between big

doi: 10.18063/bc.v2i1.

data thinking and small data thinking? What is the impact of big data thinking in the genome field? These questions will be the focus of this review.

1. Big data

What is big data? As a relatively new concept, Big Data has not been clearly defined, but many definitions of Big Data start with the characteristics of Big Data (Fang Wei *et al*., 2014). IDC simply defines large data as two or more sources and formats of data, and gives quantitative criteria, but only emphasizes the number and type of data, not from the data native characteristics definition (Vesset *et al*., 2012). McKinsey reports on the definition of large data: data sets that cannot be captured, managed, and processed over time with traditional database software tools. Wikipedia gives a simpler definition of big data: big data is a data set that takes longer than tolerable time to capture, manage, and process data using common software tools. Some scholars define large data. Academician Wu He Quan (2013, Seeking Truth, 4): 47-49) believes that large data generally refers to a large number of data sets, which can be mined out of valuable information. Zhu Jianping (2014) concluded that big data can be understood from two perspectives: "If big data is regarded as an adjective, it describes the characteristics of big data era data; if big data is regarded as a noun, it reflects the object of data science research." This review considers Professor Huang Xinrong (2015) to be big data. The definition is quite comprehensive, and the definition of big data is analyzed by domestic and foreign scholars. At present, it is concluded that the definition of large data is basically from three aspects: data size, processing tools and utilization value.

No one can give specific definitions of big data, but some people have summarized the characteristics of big data. IBM generalizes the data into three V's, volume, variety, and velocity, which are now generally accepted and not controversial. Guo Xiaoke (2013) that IBM summarized the characteristics of the three large data reflects the potential value of large data, so the characteristics of large data summarized as four V. Others regard veracity as the fourth V. IBM later redefined and refined "4V" and defined the last V as veracity. Professor Huang Xinrong (2015) also specifies that the data should be objective and authentic. He believes that the fourth V should be veracity, but he understands the definition of large data from Huang Xinrong. It seems that value is more appropriate than veracity, but some people now summarize the characteristics of the data into five V (Zhou Shijia, 2014, Journal of the Party School of the Shanxi Provincial Government of the Communist Party of China, 5): 10-12).

The definition of big data varies from one field to another. In order to understand the definition of big data, we should understand the influence of big data, especially big data, which changes our way of thinking in dealing with data.

2. Big data thinking

In the era of big data, the way people think about big data has undergone major changes. Big Data Era describes three most significant changes: (1) from sample data to overall data; (2) acceptance of mixed data; (3) pursuit of correlation between data. Scholars are basically in agreement on the changes brought about by big data. In 2015, the Learning Times commented on the changes in the way of thinking about big data (Zhang Yizhen, 2015, Learning Times, 4:1-2), Li Weishun and others (2015, Decision Forum - Systems Science Applied to Engineering Decision Making (Part 2, Part 3) The key to big data lies in the 3 changes we have made in analyzing information and data.

Some scholars analyze the thinking characteristics of big data from different perspectives, but there is no difference in their meanings. Zhang Weiming and Tang Jiuyang (2015) generalize the thinking of big data into integrity, dynamism and relevance; Zhou Shijia (2014, Journal of the Party School of Shanxi Province, CPC) (5): 10-12) summarize the characteristics of big data into integrity and relevance. Emergence, diversity and nonlinearity, correlation and uncertainty. Hybridity, diversity and dynamism are the characteristics of large amount of data. Because of data hybridization, it is impossible to strictly pursue the accuracy of data. So we can tolerate some small errors in some data, and then study the correlation of data. Here we summarize this feature as fault tolerance. Throughout the view of scholars, the thinking of big data can be summed up as the overall thinking, fault-tolerant thinking and related thinking.

In addition, this review considers that the biggest thing about big data thinking is intelligent thinking. Big data ef-

fectively improves the automation and intelligence of machines, which is the direction of human society's long-term efforts to progress. With the breakthrough development of large data technology, the system can intelligently search all relevant data, similar to the human brain has initiative and logic. Therefore, big data thinking can be summarized as: overall thinking, fault-tolerant thinking, related thinking and intelligent thinking.

There are also people who have defined big data thinking, Schoenberg said. Big data thinking refers to the awareness that open data, once properly processed, can provide answers to problems that millions of people need to solve urgently. Zhang Weiming and Tang Jiuyang (2015) believe that the big data thinking is based on multi-source heterogeneous and cross-domain Association of mass data analysis generated by the data value mining thinking. Thinking patterns, whose emphasis is on having a direct impact on how we behave, should emphasize the technical means of mining valuable data in defining big data thinking.

3. Big data thinking and individualized thinking of small data

We have already described the characteristics of big data thinking, such as integrity, fault tolerance, relevance and intelligence. To tell the difference between big data thinking and small data individualized thinking, we can see from the characteristics of big data thinking:

Part and overall. Big data thinking emphasizes integrity and requires a holistic view, which is quite different from the emphasis on some representative data in the era of individualization. From the previous part to the present whole, that is, from the elements to the system, the overall thinking mode is more systematic. Individualized sampling surveys analyze only representative "partial" data sampled at random. Victor said, "To analyze all the data related to something, rather than relying on a small number of data samples," so the big data emphasizes that "every data is analyzed and the whole is valued."

Accuracy and confounding. Individualized small data requires typicality, concretization, and each data is required to meet the requirements, and unified according to certain standards, if there is non-standard data will be eliminated, these standardized data in computer data is called structured data. On the contrary, in the era of large data, data sources are wide, complex, fast, large volume, and there is no unified standard, but each data has its reasons for existence. Big data thinking also reflects the German philosopher Hegel's thought: everything that exists is reasonable, the existing data has its reason for existence, there is its rationality. This will inevitably lead to a small number of data errors, so absolute accuracy is no longer the goal of the big data era, and the appropriate neglect of micro-level accuracy, but can have a better insight at the macro level.

Causality and correlation. From the perspective of causality in the era of individualized small data to the correlation of large data, data no longer deliberately pursue "why" but only care about "what". In traditional science, because of the small amount of data and the simple tools and means of processing, causality is particularly important, and the relationship between data is basically linear. In the era of large data, it is almost impossible to analyze the large amount of data linearly. Big data thinking uses a new perspective to predict the possibility of things happening, breaking the causal thinking mode in the era of small data.

Naturalization and intellectualization. Since entering the information society, the way of thinking in dealing with data is simple and linear natural thinking, and the machine for dealing with data is also simple. In the era of big data, data processing pays attention to the analysis of relativity, which requires that the data processing machine can systematically search all relevant data, and actively and logically analyze data, make judgments, similar to human intelligent thinking ability, which is also the key and core of big data thinking.

Although the transition from small data to large data is a major change in thinking, it is essentially consistent and interlinked, except that the former emphasizes the scientific level and the latter the technical level.

4. The application of big data ideas in biomedicine

In the era of big data, the thinking of statistical data has changed. Li Jinchang (2014) summed up the author's point of view that the thinking change mainly includes the thinking of understanding data, the thinking of collecting data and the thinking of analyzing data.

4.1 big data in biomedical field

To correctly understand big data, we need to start with data sources, types and quantification. Previously, American scientists Weston and Hood (2004) first put forward the "4P" medical concept, advocating individual prediction, prevention and treatment. Individualized medical care needs to analyze the information of each patient comprehensively, aiming at the huge amount of information in the diagnosis and treatment of individual patients. At the same time, the completion of the Human Genome Project has promoted the study of human genes. It is of great significance to analyze the correlation between gene expression, gene variation and disease in genomic databases. The data collected from proteomics, metabonomics, transcriptome, lipomics and glycomics are enormous. And human studies of the ancient human genome are also deepening (Liu Ruitao *et al*., 2015). In addition, the rapid development of high-throughput sequencing technology and the sharp drop in the cost of genome sequencing have led to the emergence of large data in biomedical fields.

4.2 transformation of biomedical big data in collecting thinking

Big data thinking in the collection of data thinking changes, the first thought is commonly used in biology NCBI database. NCBI is the abbreviation of the National Biotechnology Information Center of the United States. Since the Human Genome Project was formally launched in 1990, sequencing technology has developed rapidly and bioinformatics resources have become increasingly huge. How to collect, store and analyze data efficiently has become an important research topic. Bioinformatics, an interdisciplinary subject, has also been applied. NCBI database is developed under such a background.

NCBI is the most integrated bioinformatics database in gene database. The famous GenBank Accounting Sequence Database is developed by NCBI. It is the authoritative sequence database in the world. The database has collected the published gene sequences of various countries since 1981. NCBI has collected the biology of more than 70 countries since 1966. In addition to the GenBank, there are branch sequence databases and professional databases, such as expression sequence tag database, genome sequencing database, single nucleotide polymorphism database, protein molecular three-dimensional structure database, etc. (Tian Geng *et al*., 2000, Foreign Medical Molecular Biology Bulletin, 22 (5): 317-320, etc.).). The process of collecting data by sequencing technology is itself a big data thinking. The big data thinking is not only reflected in understanding data and collecting data thinking, but also the thinking of analyzing data.

4.3 Biomedical big data mining

Big biomedical data can not only be used in genomics and association studies between different groups, identifying biomarkers and developing drugs, implementing health management and so on, but also can be used to implement more powerful data mining, such as association analysis, clustering analysis, classification analysis and anomaly analysis of data mining. Big data mining in biomedicine can increase the degree of grasp and have the ability to discover weak association, such as the mining of TCGA database information and the analysis of existing research data.

It is very important to mine useful information in TCGA database by big data thinking. By scanning the methylation sites related to the prognosis of lung adenocarcinoma in the whole genome, we can find the genes related to the prognosis of lung adenocarcinoma, which can be used as a biomarker for prognosis research. Zhi Wu (Wang Ke *et al*., 2016). Or we can directly study the correlation between the target gene and cancer, collect cancer data sets from TC-GA database, download gene expression profiles and clinical information, then we can analyze the correlation between the target gene and the clinicopathological parameters of cancer and the impact on the prognosis of cancer (Wang Shuo *et al*., 2016), and also can analyze the cancer phase. Related microRNAs and mRNAs are analyzed jointly, and co-expression network maps are constructed for joint analysis to identify clinically relevant genes or microRNAs for further research, such as cancer. To select microRNAs that interact with the target gene, the cancer gene data and the number of microRNAs can be downloaded from the TCGA database. According to the joint analysis, the microRNAs related to the target gene expression were found and screened. With the support of large data from TCGA database, the screening efficiency could be effectively improved.

4.4 Meta analysis

Meta-analysis, also known as meta-analysis, is a statistical method to systematically merge the results of multiple independent studies or multiple studies of the same subject with specific conditions and common research purposes, to analyze the differences or characteristics between the results of the analysis, and to quantitatively analyze and evaluate the results of the study. Meta-analysis in a broad sense refers to a scientific clinical research activity that collects all relevant studies and makes a rigorous evaluation and analysis one by one. It is also the whole process of statistical processing of data by means of quantitative synthesis and drawing comprehensive conclusions. In a narrow sense, it refers to a simple quantitative synthesis system. Methods of learning. Qiu Xiaochun (2014) counted 32 933 global meta-analysis papers published from 2001 to 2012. Among them, the United States ranked the first. After 2009, global publication reached its peak, and the number of papers published in China surpassed that of the United States.

In recent years, Meta-analysis has played a unique role in medical diagnosis, treatment, interventions, and health decision-making (Lu Shiwei *et al*., 2001). It has played a special role in medical research. It can improve the efficiency of statistical analysis, analyze the causes of divergence among several similar studies, and can also lead to new insights and help to evidence-based research. The development of Medicine (but Han Lei, 2003, Chinese Journal of medical research management, 16 (1): 12-15).

5. Prospects

The arrival of the era of big data has affected our medical service model. The treatment-oriented approach has changed to the prevention-oriented approach. The role of big data analysis technology in the medical field is becoming more and more important. Einstein said: "The development of the world of thinking, in a sense, is to constantly get rid of the surprise" in the field of biomedical research in depth, and achieved extraordinary clinical medical value, in the near future believe that the big data thinking will bring us more breakthroughs in biological research. Sexual development.

References

1. Fang W, Zheng Y, Xiu J. Big data concept on key technologies and applications. Nanjing Xinxi Gongcheng Daxue Xuebao (Ziran Kexue Ban) Journal of Nanjing University of Information Science and Technology (Natural Science Edition) 2014; 6(5): 405-419.
2. Huang XR. The semantics, features and essence of big data. Changsha Ligong Daxue Xuebao (Shehui Kexue Ban) 2015; 30(6): 5-11.
3. Li JC. Big data and new mind on statistics. Tongji Yanjiu (Statistical Research) 2014; 31(1): 10-17.
4. Liu RT, Wang M, He HJ, *et al*. Hominid genomes and health. Jiyinzuxue Yu Yingyong Shengwuxue. (Genomics and Applied Biology) 2015; 34(6): 1333-1338.
5. Qiu XC. Bibliometric study of meta-analysis in medical research. Yixue Yanjiusheng Xuebao (J.Med.Postgra.) 2014; 27(7): 733-736.
6. Staff P. Dealing with data, challenges and opportunities, introduction. Science 2011; 331(6018): 692-693.
7. Vesset D, Woo B, Morris HD. Worldwide big data technology and services 2012-2015 forecast. IDC Report 2012; (1): 233485.
8. Wang K, Zhao RX, Yang SL, *et al*. DNA methylations associated with survival of lung adenocarcinoma with TCGA database. Nanjing Yike Daxue Xuebao (Acta Universitatis Medicinalis Nanjing) 2016; 36(6): 665-669.
9. Wang S, Li Z, Zheng CL, *et al*. Analysis of clinical significance of AKT3 expression in gastric cancer utilizing TCGA datasets. Zhongguo Yike Daxue Xuebao (Journal of China Medical University) 45(5): 398-401.
10. Weston AD, Hood L. Systems biology, proteomics, and the future of health care: Toward predictive, preventa-tive, and personalized medicine. Journal of Proteome Research 2004; 3(3): 179-196.
11. Zhang WM, Tang JY. Big data thinking. Zhihui Xinxi Xitong Yu Jishu (Command Information System and Technology) 2015; 6(2): 1-4.
12. Zhu JP, Zhang GJ, Liu XW. Clarity of a philosophy of data analysis during the age of big data. Tongji Yan jiu (Statistical Research) 2014; 31(2): 10-19.