Thoughts on the Problem of Small Data

Jingtao Cheng*, Wenhui Chen and Yuhong Gong

1Zhejiang Gongshang University,
No. 18, Xuezheng Street, Hangzhou, Zhejiang, China, 310018
*17020040020@pop.zjgsu.edu.cn

Abstract. Marked by Viktor Mayer-Schönberger's "Big Data: A Revolution That Will Transform How We Live, Work, and Think", Big Data swept across the major subject areas, changing human thinking and behavior. When big data is so hot, "small data" has begun to attract the attention of all sectors of society. Due to the high value of density, deep-level mining, single collection and other significant advantages, in many respects, the infinite value of "small data" is not inferior to big data. This paper focuses on the analysis of the connotation, application and development prospects of "small data", and compares it with "big data", creating a certain reference value for the development and research of small data.

1. Introduction
In his book, Schönberger [1] mentioned three major changes in big data: first, analyze all data related to something, rather than rely on analyzing a small volume of sample data; second, accept the complexion of data, rather than pursue accuracy; in the end, focus only on the correlation of things, instead of exploring elusive causality. These three major changes have actually become the three major criteria of big data analysis in the following years, greatly speeding up the development of big data and broadening the concept and scope of data analysis.

At present, big data has become the core of the times. However, due to the "4V" characteristics of big data, namely Volume, Variety, Value, Velocity, it requires a high level data environment, which could not be afforded by small and medium-sized enterprises and small and medium-sized data collection, processing, analysis units and other similar institutions. In other words, only large enterprises with large amounts of social capital resources have the ability to conduct big data research, and their hard requirements for advanced analysis and data mining as well as corresponding infrastructure equipment, talents, and markets also limit the entry of a large number of enterprises. All in all, only the “high-play” companies with strong capital can “play” big data and benefit a lot. Difficult to enter, difficult to develop, and difficult to benefit from, these are the fatal flaws that big data has shown after its development. Compared with big data, small data are relatively simple in these areas and more friendly to all businesses.

The distinction between big data and small data is relative. Undoubtedly, big data is made up of small data[2]. Then, what are the characteristics and attributes of the small data that constitutes big data, and how is it compatible with all enterprises?

2. Definition and Background
So, what exactly is small data? currently, there is no uniform definition of small data at home and abroad. In general, there are two main views on small data[3]: one is that small data is the sample data, which is obtained by sampling in traditional statistical research. These data are the sample data obtained by traditional sampling methods in order to facilitate the operational study of certain problems. The small
data interpreted at this level is obviously the sampling data of our statistics for hundreds of years. The research on this data already has a complete theoretical system, no need to repeat any more.

Another view is that individual data is small data. The individual here can be either a person or a group, as long as the group in question has a common attribute and is being studied. In this view, small data is the all data of a certain aspect of a single individual, that is, individual data sets. By synthesizing small data about all aspects of the individual, an individual data cluster, a small data set, can be formed. Visibly, individual user of small data is concerned[4], is "quantified self"[5], its core is people-oriented. For example, Deborrar Astin, a professor of computer science at Cornell University in the United States, observes and records the daily activity data of his father's death a few months ago - no longer sending e-mails, gradually decreasing number of shopping in supermarket, the shorter and shorter walking distance. These activity data show very clearly his father's physical condition is declining, but these subtle abnormalities could not be reflected in the inspection of the hospital[6]. Thus, Professor Deborrar realized the importance of small data and he was the first to discover the value of small data[7].

Obviously, for the two interpretations of small data, this article prefers the latter. In terms of data, it is different from sampling data such as traditional data, which is all the data of a certain thing; in the scope of objects, it is different from big data, and is comprehensive data for individuals. Small data seems to be a new type of data, which deserves to be treated individually, studied carefully and thought deeply.

3. The Similarities and Differences Between Big Data and Small Data
As we all know, the first change in the era of big data is that we rely on the all data in replace of random sampling. At the same time, from the amount of data, big data is an overall data with "4V" characteristics, and small data is comprehensive data for one aspect of the individual. Then, from the relationship and attribute of the statistical "whole-sampling-sample", both belong to the whole, which is the whole of the research object, different from the sample sample. Thus, at this level, there are several similarities between the two that distinguish the nature of the sample.

The process of traditional statistical analysis can be summarized as “hypothesis-verification”. The specific operation is to estimate the overall situation by taking samples and using the verification results of the samples. After the emergence of big data, the analysis process turned into “discovery-summary”. Through massive and comprehensive data, the internal laws are directly searched for conclusions, and no longer rely on the data analysis of samples.

Sampling causes a big problem – probability. Due to the sample and distribution theory, we must consider the quality of the sample, that is, how to evaluate traditional statistical analysis. The evaluation criteria of statistical analysis mainly lie in two aspects: reliability evaluation and effectiveness evaluation. Reliability is measured in terms of probabilities - sometimes at the level of confidence, sometimes at the level of salience - to infer the degree of mastery of the population from the sample[8]. The significance level of the evaluation indicators vary depending on the sample distribution and statistics. The so-called effectiveness evaluation refers to the authenticity, that is, the size of the error, that is, accuracy and precision. Accuracy refers to the degree consistent with observations of real value, generally can not be measured; precision refers to the degree of dispersion of the sample distribution, sampling distribution of standard deviation to measure.

For traditional statistics, because its data comes from sampling, it means that it has reliability problems and validity problems. For reliability evaluation, how to determine whether it exists? If so, how should the metrics be measured, quantified, and whether there is a scientific basis for reliability evaluation? For effectiveness, it is clear that the sample data of traditional statistics has accuracy and precision problems. Because of the precision, the sampling error and the non-sampling error are also distinguished. The former is affected by sampling and can be controlled by sampling distribution theory.

Our big data and small data are whole data, which is different from traditional data. Obviously there is no sampling problem. Naturally, reliability problems do not exist. One need to pay attention to non-sampling error on the matter of accuracy, which means only accuracy problem need to be considered. However, even if it is the whole data, there is still a difference between the two: because big data is massive data, and the hybridity and diversity of big data, its non-sampling error is difficult to control, and accuracy evaluation is difficult to be perfect. For small data, although it is also the whole
data, since its fundamental purpose is to verify the true state of the individual, its accuracy problem was solved at the beginning, and the true reflection of the individual means accuracy.

Just like the second transformation of big data: not in accuracy, but in hybridity. Small data find another way to trace the high accuracy of the data and truly reflect the connotation and significance of the data. Finally, Schönberger's third shift in the book: no longer keen on finding causal relationships, but instead focusing on the relationships of things. Is it true that the transition from traditional data to big data?

Firstly, for the transformation of data analysis, the traditional statistical analysis process is "qualitative-quantitative-qualitative". The first step is to find the direction of quantitative analysis, and the means are simple, mainly relying on the subjective judgment of the researchers. After quantitative analysis, determine whether the hypothesis of the first step is correct and whether the research is accurate or not. As for the emergence of big data, due to the massive and complete data, we directly omit the first step, and directly from the data analysis can we find the relationship between data characteristics and quantity, which is simplified as "quantitative-qualitative". However, such a change is only brought about by the data structure, in other words, simplification. In my opinion, the so-called third transformation did not happen. Whether it is a traditional data analysis process or a big data analysis process, the final result is a qualitative analysis report, the difference is only in the method and accuracy.

Secondly, let's look at the analysis process of the data. Small data is the whole data, which is similar to big data, the data volume is large, naturally do not need to consider the qualitative of the first step, after quantitative analysis, qualitative research is also needed, but unlike the big data analysis process, this is not the final result. This step is qualitative only to find the relevant, the latter process is the key - feedback. After finding the data relationship, the small data analysis must establish an information feedback mechanism. Through the feedback loop of the information, we can dig deeper from the relevant depth and find the causal relationship of the root cause. This is the core of small data that is different from big data analysis: "Know not only what it is, but also why it is."

Of course, the same point of big data and small data is generated by the whole data. The difference between the two is caused by the different research objects, data types, analytical methods and even data thinking. The above three major changes analyze the similarities and differences between the two from a large point of view, but there are more actual differences, which are reflected in all aspects of statistical analysis.

Then, there is a different understanding of big data and small data for the processing of data noise problems. Noise data refers to errors or abnormal data existing in the data, which cause interference to data analysis. In the eyes of big data, noise is meaningless data. For such data, processing is no different than deleting or replacing. For the understanding of deletion, based on the understanding of the data, it is considered that in the big data environment, the abnormal value such as noise is obviously wrong, and the analysis result will produce an in calculable error. At the same time, the data volume of big data is huge and updated rapidly. Direct deletion of outliers does not have a major impact on the analysis process. Indeed, for occasional outliers, such processing is understandable, but when a large amount of data is abnormal, even if there is noise in the variable, it is still advisable to delete it directly. Compared to deletion, the replacement seems to be more in line with the data preprocessing principle. At the same time, according to different concepts, specific alternative methods emerge in an endless stream. By sub-boxing, data smoothing, integration, data transformation, etc., or using prediction values, smoothing values and even changing data weights to replace the original data can all change the data noise problem. However, the replacement is based on non-anomalous data. Whether the newly generated data column will affect the original data column after replacement, and how much impact it may cause, we are yet to know, after all it is difficult to investigate.

All in all, whether it is deleted or replaced, it is developed under the view that "outliers are wrong." So, does the outlier really mean data errors? Let us review the process of Professor Deborah discovering small data.

In the case of a medical examination without health problems, the social activity information that his father is weakening is opposite to the former. It is completely abnormal, but is the data wrong? Obviously not, it is these abnormal data that truly reflect the physical condition of his father. Such a case,
we have to ponder, the outliers are really wrong, whether these data are also truly reflecting the phenomenon changes. We look at the data noise problem based on the concept of small data. The outliers that bring data noise may be some kind of abnormal fluctuations that occur during the change of things. Such fluctuations represent the changes that are taking place in things. They are still treated as before, which is equivalent to directly ignoring such changes. It runs counter to our original intention of analyzing data. Then, under the premise of small data, it is necessary to take the outliers seriously. Firstly, distinguish between outliers and erroneous data, and filter out the real erroneous data using the big data method. Secondly, the classification of the outliers that affect the change of the phenomenon is classified according to the direction, degree and sequence of the influence on the phenomenon. Thirdly, take appropriate methods for the classification data for analysis and processing, and try to extract the useful value for explaining the analysis phenomenon. Fourthly, synthesize all outliers, quantify the impact of data noise on the problem, and share the final interpretation of the problem.

Finally, for over-fitting problems, the two also have different analytical perspectives. It is true that in the era of big data, the pursuit of precise correlations allows us to discover the truth that has not been discovered before, but whether such reality is in line with reality. We are familiar with such a case. Wal-Mart found that its beer sales were significantly positively correlated with the baby's diaper sales, and explained that the father went to the supermarket to buy baby diapers while buying some beer to take home. The resulting sales strategy has brought considerable benefits. However, if there is a strong correlation between the two because of the large correlation between the two, it is too arbitrary. The value of big data is here to become a roadblock for mining the truth behind the data. Therefore, small data at this time provides the possibility to mine real value.

For specific individuals, special small data research is carried out. It is easy to master more comprehensive data, and it is also easy to master more realistic data. This is extremely important for supplementing relevant analysis and then conducting causal research. It can be seen that the quality of small data is largely decisive for improving the quality of big data. Therefore, before the big data analysis, it is necessary to carry out the data preprocessing process based on the small data concept to evaluate whether the data can be analyzed by big data. Only real data, after analysis, can get real results, in order to sum up the true value from the analysis of the data, and discover the laws, rules and causes of the real things.

In summary, small data has the following advantages. First, it has distinct individual uniqueness. Small data is a collection of data collected around individuals with a very high value density. For different individuals, their behavioral activities are specific, resulting in large differences in collected small data, and individual uniqueness is extremely significant. Second, it has a high degree of privacy and security. Big data mining analysis and value discovery are at the expense of user privacy and security. This is to sacrifice for rapid development, and the cost is too great. Small data, because it is based on the concept of serving the user, attaches great importance to privacy protection and security considerations, on the basis of discovering information and mining value. Third, the biggest advantage of small data is the interpretation of noise and anomalies. Out of the strange point of understanding the abnormal value, small data is very effective in finding the inflection point of the phenomenon change, and can find the "Nino phenomenon" in time [the fishermen in Peru and Ecuador used to call an abnormal climate phenomenon. ]

4. Application of small data
Small data has many significant advantages and ascendancy, and its application has also achieved great results, especially in the field of graphics. In particular, the Chen Chen team's research, the team based on the efficient, real-time, dynamic advantages of small data, developed a user-specific service model. The model consists of three modules: a small data analysis and decision module. The module is responsible for data collection, storage, transmission, and noise filtering and standardization processing, and then small data analysis and decision-making; readers' interest discovery module. The module acquires the knowledge of the reader, and then matches and recommends the interest, and at the same time constructs the readers' interest database; the service content and readers' interest correlation evaluation module. The module is responsible for formulating the recommendation service, through the comparative analysis of the above modules, customizing the personalized service content, making smart
recommendation, and implementing the optimization and update of the service content according to the user feedback.

At the same time, there have been many small data application practices based on different considerations. Z.F.Li[9] designed the library's personalized recommendation service model by establishing an associated database of small data behaviors, and is committed to meeting the common needs and individual needs of readers. L.R.Li[10-11] team analyzed the fit of small data and micro-knowledge services of scientific research users. Based on this, the micro-knowledge service model was proposed. And, a micro-knowledge service system dynamics model guided by the real-time needs of scientific research users was constructed.

5. Conclusion
In summary, we need to explore small data based on traditional statistics and big data thinking. Breaking away from the experience and theory of predecessors to study small data will not last long. Such small data can only be discussed on paper. It is difficult to apply it to reality, and it is gradually drifting away from the concept of "serving the individual". Correspondingly, big data analysis without small data evaluation is only the change and reconstruction of data, and it is only a superficial result, which is not benefit to the research and analysis of the complex relationship of the real world. Research big data and discover small data. Let the data not just the data, let the data speak the truth, present the true relationship, and reveal the real world.

6. References
   Reference to a book:
   [1] Viktor Mayer-Schonberger, Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work and Think. Hodder & Stoughton, 2013.10.
   [2] J.C.Li, Big data analysis and small data research, China Statistics. 2 (2018) 24-26.
   [3] J.C.Li, Some understanding of statistical data, Statistical Research. 11 (2017) 3-14.
   [4] L.F.Chen, Collection, Analysis and Application of Library User Small Data in Big Data Environment, Journal of the National Library of China. 3 (2016) 69-74.
   Reference to a chapter in an edited book:
   [5] Information on https://baike.so.com/doc/25739117-26871562.html.
   [6] D.Estrin, Small data, where n = me, Communications of the ACM. 4 (2014) 32-34.
   [7] D.Estrin, C. W. Thompson, Internet of you: data big and small.IEEE Internet Computing. 6 (2015) 8-10.
   [8] J.C.Li, Big data and statistical new thinking, Statistical Research. 1 (2014) 10-17.
   [9] Z.F.Li, Research on Small Data of Library Users Driven by "Internet +", Library Theory and Practice. 12 (2012) 108-112.
   [10] L.R.Li, Z.H.Deng, Research on micro-knowledge services for small data of scientific research users (I), Library & Information. 3 (2016) 49-54.
   [11] L.R.Li, Z.H.Deng, Research on Micro-Knowledge Service for Small Data of Scientific Research Users (Part II)--Based on System Dynamics, Library & Information. 5 (2016) 110-118.