Can Studies on Data Related Issues Be "Data Science"?

Yong Shi
Research on Fictitious Economy & Data Science
Chinese Academy of Sciences, Beijing 100190, China

Our culture has developed the ability to generate masses of data. Computer systems expand much faster than the human ability to absorb. Furthermore, Internet connections make it possible to share data in real time on a global basis. Studies on data related issues from data analysis, data structure and database management rapidly expanded to hot areas, such as data mining, knowledge discovery, intelligent knowledge and knowledge management. A challenging question has been opened to all of scholars and practitioners: can these studies on data related issues form "Data Science"?

This question may be first discussed through data, data format, information, and knowledge in the notions of "science." Data can be defined as a certain form of the representation of facts. Data only presents one side or "angle" of a fact, but not all. Completely relying on the data representation of the fact can be biased. Data formats can be structured and unstructured. If data has a format shown by fixed fields or identified within a data model, it is structured. The examples are relational tables, spreadsheets and XML In contrast with structured data, unstructured data cannot be shown by a pre-defined data model. Books, journals, e-mail message, Web page, documents, audio, video, and image are the examples of unstructured data. Information is any data that has been pre-processed to all aspects of human's interests [1]. Knowledge is human understanding of a subject. Science is viewed as two groups: empirical sciences (including natural and social sciences) and formal science [2, 3]. Data observed or recorded by empirical sciences are used to gain knowledge about nature or human beings. Although a formal science, like mathematics, emphasizes the formation of hypotheses, derivation of theories and discovery of laws, it also uses data to show the natural changes or human behaviors. Therefore, studies on data related issues, including data format, information, and knowledge, inhere in scientific study.

Secondly, studies on data related issues can also interface with the notions of basic science and applied science. While basic science finds knowledge, applied science uses knowledge to solve a real problem. Engineering is a classic example of applied science. It designs and/or builds a system or structure to upgrade the life quality of human beings by using knowledge found from natural and social sciences [4]. The scholars often use a term "Data Engineering" which explains that processes of searching the target knowledge from data, such as data acquisition, data storage, data management, data mining algorithms and knowledge discovery belong clearly to a category of applied science. However, if finding knowledge from data is considered as a whole step, Data Engineering, at least some of its components, can be also viewed as part of basic science.

There could be a common agreement to treat the studies on data related issues broadly as "Data Science", which covers the known notions of data analysis, data structure and database management, data mining, and knowledge discovery and go beyond the terms of intelligent knowledge and knowledge management.

Given much has done by known data mining and knowledge discovery from databases in the language of empirical science and/or applied science, Data Science should pay more attention on how to address its problems of formal science and/or basic science. Instead of validating the theoretical results based on data observed from the real world, formal science focuses on describing the properties of an abstract formation from definitions and/or rules [3, 5]. Similarly, basic science deals with the relationships or logics of the basic objects or forces and discovers the principles or laws among them. Its outcome may not immediately relate to the real-life applications [6]. Data Science can take advantages of approaches used in formal science and basic science in studies on data related issues. For instance, what are the axioms of data acquisition in terms of a given objective? Are there any principles in building data storage? What properties does a model of data management have? Does any globally optimal classifier exist in a data mining algorithm? Can data mining have a law of finding knowledge? These theoretical questions have never been fully addressed in current works of data mining and knowledge discovery from databases, in which many approaches are based on heuristic computing and empirical tests. In fact, a fundamental structure of bridging data mining results and intelligent knowledge with useful properties has been initiated [1]. Given certain conditions, the existence of locally optimal classifier by using an optimization-based method can be proved without any computation or empirical tests [7]. These evidences show that studies on data related issues have a great chance to be updated as a real “Data Science” or science of data.

Finally, can massive data form a “Data Nature”, which differs from the real nature or human society [8]? Massive data is certainly a projection or reflection of a side from many sides of the real nature or human society. However, the picture of “Data Nature” can be understood or interpreted in different ways. Just like the “second life”, “Data Nature”, if exists, is a virtual world. It is a “one-side story” and may not be as much colorful, lively, dynamic, and evolutionary as the real nature or human society. Nevertheless, understanding “Data Science” is as hard as the understanding of the real nature and human society.

References

[1] Zhang, L, J. Li, Y. Shi and X. Liu (2009) “Foundations of Intelligent Knowledge Management”, *The Journal of Human Systems Management*, Vol. 28 (4), 145-161.
[2] Popper, Karl (2002) [1959]. The Logic of Scientific Discovery (2nd English ed.). New York, NY: Routledge Classics.
[3] Lȍw, Benedikt (2002) “The Formal Sciences: Their Scope, Their Foundations, and Their Unity”.
[4] “Engineering” (2011), http://en.wikipedia.org/wiki/Engineering
[5] “Formal Science” (2011), http://en.wikipedia.org/wiki/Formal_sciences
[6] “Basic Science” (2011), http://en.wikipedia.org/wiki/Fundamental_science

[7] Shi, Y, Y. Tian, X. Chen and P. Zhang (2009) "Regularized Multiple Criteria Linear Programs for Classification", *Science in China Series F: Information Sciences*, Vol. 52, 1812-1820.

[8] Zhu, Y. and Y. Xiong (2009), Dataology and Data Science, Fudan University Press (in Chinese).

[9] "Second Life", http://secondlife.com/