Evolving New Trends Towards Data Science

Soumya Ranjan Mishra [1], Tarun Behera [2], Tithirupa Tapaswini [3], Priyanka Shit [4]

SANTOSH KUMAR SHARMA[5]

[1, 2, 3, 4] Gandhi Institute for Education & Technology, Baniatangi, Khordha, Odisha

[5]NM Institute of Engineering & Technology,Bhubaneswar,Odisha

soumyamishra@giet.edu.in, tarunbehera@giet.edu.in, tithirupatapaswini@giet.edu.in, priyankashit@giet.edu.in

The amount of data being produced in internet nowadays is enormous. Due to a data boom, datanature—all data in cyberspace—is currently taking shape. It is important yet challenging to investigate the patterns and laws in data- nature. Data Science is a new field that will soon be established. It offers a new sort of data-intensive research methodology for the natural and social sciences that extends beyond computer science. The issues posed by data are discussed in this essay along with how data science differs from traditional sciences, data technologies, and big data. Our aim is to persuade data-related scholars to shift their attention to this emerging field of study.

Keywords: Data Science; Datanature; Cyberspace; Data; Data scientist

1 Introduction

The data explosion is the rapid increase in the amount of data in cyberspace, which brings humanity into the big data era. The meaning of data has evolved. Data are no longer limited to values of qualitative or quanti- tative variables, or the results of measurements, or scientific data generated within the context of scientific observations and experiments. In addition to all of that, data also are everything found in cyberspace. Data- nature (all the data in cyberspace) forms and develops unconsciously (Zhu, Zhong, & Xiong, 2009: Zhu & Xiong, 2009). There are increasing instances of data that have no references in the natural world, such as computer viruses, online games, and junk data, all of which are generated in datanature. The information generated in datanature has gradually surpassed the facts existing in the natural world and has come to exhibit unique patterns.

Since the computer was invented, we have been constantly utilizing and dealing with data. The facts of the natural world are mapped as data and stored in computers so that we can use them when needed. However, the method of using data has changed from simple data access to big data analysis, especially in the realm of science (e.g., life science). This brings new requirements and challenges for data technologies, which lead to research on the data themselves, such as how to study life through DNA data. The goal of data utilization is also changing. Data analysis not only aims to solve problems based in reality but also extends to analyzing data in order to study the phenomena and rules of the data themselves (e.g., discovering the growth patterns of data and predicting the scale of data in cyberspace ten years into the future). Providing natural and social sciences with data technologies and methods and exploring datanature can and should lead the transition towards this new science, data science. Whether you know it or not; whether you accept it or not; whether you are ready for it or not, data science is coming. If you have been engaging in data science research, you may already have become a data scientist.

In this paper, we present the challenges presented by data and investigate why we need data science. We also include how data science differs from existing technologies and established sciences. Furthermore, we discuss some key issues (e.g., fundamental theories, new methods, and research topics) that will be faced by data science when it becomes an academic discipline having data as its research objects. We also review the progress being made in the current research and society of data science and discuss a few perspectives and challenges found on the agenda of data science. Finally, we illustrate how to transfer existing knowledge to this new science.

2 Challenges When Working with Data

As datanature gradually becomes more important, its study faces more and more challenges. These chal-lenges are discussed below.

2.1 Truth in Data

How do we know whether the data we have are telling the truth or giving false information? How do we deal with a dataset containing false data? If false data are mixed with correct information, how do we measure the confidence

level of a dataset? For example, if some product reviews are given by users not having used the products or even by competitors, those reviews may be not credible. Therefore, the analysis results (e.g., credit ratings) based on a dataset including such data will not be credible either.

These are critical challenges in the data-related research area and will become an important aspect of data science research. With social networks such as Facebook and Blog expanding, the challenges are getting more and more severe.

2.2 Survival Problems in Cyberspace

Cyberspace is becoming a part of humanity's experience, i.e., we will soon live in both physical space and cyberspace. How do we survive in cyberspace? For example, one of the basic survival problems is how to communicate in cyberspace. This may become one of the most difficult problems in the future of data- related research because of problems in communication context. In fact, this problem already exists some- where in cyberspace. For example, the Martian Language, one of the internet languages teenagers use to communicate online, can be regarded as a communication method in cyberspace. For most people, Martian Language is very difficult to understand because it adopts words from various languages (such as English, Chinese, Japanese, etc.) and mixes them all together.

2.3 Scientific Research with Data

Because data are the representation/mapping of the facts in the natural world, they are used to discover rules of the real world. The discovery and exploration of phenomena and rules in datanature support the discovery of phenomena and rules in the natural world. Therefore, developing methods in datanature to explore the rules in the natural world is a potential research field that will be helpful for scientific research. Data researchers recognize that with the explosive increase in the amount of data in various fields, many problems cannot be solved using traditional methods, and they realize the importance of data in scientific research. Consequently, they have been exploring new approaches to deal with data, such as data mining technology and big data analysis.

2.4 Knowledge Acquisition from Data

In the early stages of computer science history, the focus was on how to improve computing's performance and capabilities. Currently, however, a more important problem is how to acquire valuable knowledge from the increasing mass of data being generated, for example from both the natural and physical sciences. Some questions include: How can we find useful data in cyberspace? How can we get knowledge from data? These require us to understand and process data from a new angle.

Focusing on the challenges mentioned above, we discuss below why we need data science and its bounda- ries with other areas.

3 What are the Differences?

Informationization is a data generating process that stores the objects or phenomena of the natural world in the form of data in cyberspace. Data are a representation of nature, recording human behavior, including working, livelihoods, and social development. Presently, the scale of data is increasingly and being rapidly generated in cyberspace. This is called data explosion. Data explosion forms datanature in cyberspace. It is necessary to investigate and explore the data rules in cyberspace as data are a unique entity. Meanwhile, this investigation and exploration are an important way to explore the rules of the universe, life, human behavior, and social development. For example, we can study life (i.e., Bioinformatics) or investigate human behavior (i.e., Behavior Informatics (Cao & Yu, 2009)) using data.

3.1 Differences from Other Data Technologies

The techniques dealing with data, such as data storage, data sharing, and data access, have been develop- ing since the invention of the computer. The formation and development of data science issues extend far beyond those in the area of computer science. Data science uses similar methods and techniques, including data acquisition, data storage and management, data security, data analysis, and data visualization, etc., but in ways very different from traditional methods. There are overlaps in many areas, such as data mining, information retrieval, data integration, and artificial intelligence, but the differences are still profound. Data science requires fundamental theories and new techniques.

In essence, computer science creates models for the real world using computer languages so that real world realities, including humans and their behaviors, can be stored in computer systems. In these com-puter systems, facts are stored in the form of data. The modeling task is a process of dealing with data. Therefore, data

technologies in computer science were intended to be used in building models for facts and programs and for data computation using computer systems. This is only one explanation for the scienceof data usage.

Currently, studies conducted in the field of computer science focus on data processing and data analysis technologies, including data integration and data mining. Data mining is a technique paid an incred- ible amount of attention in the field of large scale data analysis. Its researchers have been developing explosive algorithms and tools for explaining and predicting transactional and behavioral data. Data mining is a branch of computer science that focuses on researching data. However, "data mining" is a much smaller set of concepts in the larger field of data science (Dhar, 2013). Furthermore, computer sci-entists have pioneered research on data (such as data mining technology); therefore the related publica- tions and conferences have come from the Institute of Electrical and Electronics Engineers (IEEE) or theAssociation for Computing Machinery (ACM). In fact, there exist an increasing number of disciplines (suchas Bioinformatics) that focus on data research. In these disciplines, the related publications and confer-ences are not from IEEE or ACM.

Research on how to model realities with data, how to manage and utilize these data, and how to develop data technology using computers belongs to one part of data science.

3.2 Differences from Big Data

Big data in industry impels data science. Dhar (2013) mentions that the implications of data science include the question of how scientists could use big data to their benefit in scientific inquiries. Increasing and explosive quantities of data stored in cyberspace give us the opportunity to acquire big data sets in various areas from datanature. Because it is easy to access such big data, we can conduct more and better researchon data. However, it is difficult to process big data using existing data technologies due to their large scaleand complexity. Therefore, new data technologies are demanded. Nowadays, big data technology has been improving. Developing big data technology itself is one of the research issues in data science. Utilizing big data to solve various problems in scientific and social areas is also one part of data science; big data is one of the top/hot research topics in data science.

3.3 Differences from Other Sciences

Data is the formal representation of nature in computer systems; information is the phenomena of nature, society, and thinking activities; and knowledge is experience gained through practice. Data can be regardedas symbols and representations of information and knowledge; however, they should not be equivalent to information and knowledge. Data science research objects, goals, and methods are essentially different from those of computer science, information science, and knowledge science.

On the one hand, data science supports natural science and social science. Dealing with data is one of the driving forces behind data science. Data science provides a type of novel research method, called theScientific Research Method with Data, for natural science and the social sciences. Hence, data science is also referred to as a data-intensive science (Hey, Tansley, & Tolle, 2009). For example, life science is a basic experi- mental course. However, scientists always take a long time to carry out an experiment. Nowadays, scientists can earn more achievements from their biological data analysis because bioinformatics can decrease these time-consuming experiments and enhance their efficiency. In particular, bioinformatics makes important discoveries from biological data, such as shotgun sequencing. Bioinformatics is a discipline that transferslife science from an experiment-based science to a science combining computations with experiments, demonstrating that we can research life through biological data. Biology research with data also solves some new problems that traditional methods cannot handle.

On the other hand, more and more scientific research will be directly targeted at data in datanature, instead of the facts in nature, which will then promote man to recognize data and facilitate them to explore nature and human behavior. Natural science takes substances in nature as research objects, and social sci- ence takes human behaviors as research objects. However, data in cyberspace are gradually covering and exceeding the facts in nature and human behavior because more and more data exist without references in nature and human behavior. Consequently, data researchers tend to research data in cyberspace, i.e., take data as research objects, which is different from natural science and social science.

4 How to Transfer to Data Science
4.1 The State of the Art

Data Science has been attracting a great deal of attention. The term "datalogy" (also called "science of data") was firstly used by Peter Naur (1966) to suggest that "computer science" should be called "datalogy". Theterm "data science" began to be used in the 1990's (Smith, 2006). CODATA (the Committee on Data for Science and Technology) (www.codata.org), a representative of the scientific data research area, uses theterm data science to

deal with data from various scientific research fields, which was then put into practice through the *Data Science Journal* in 2002. No definition of "data science" has been formally written, only some research content, scope, and topics have been pointed out (Smith, 2006; Hayashi, 1996; Liu, Zhang, Li, et al., 2009). In 2010, Loukides (2010) discussed what data science is, examined some aspects of data science including technologies, companies doing data science work, and the unique skill sets associated with it, and argued that data science should enable the creation of data products not just be considered as an applica-tion with data. In 2009, Zhu et al. defined data science as a new science whose research object is data (Zhu, Zhong, & Xiong, 2009: Zhu & Xiong, 2009).

Currently, data science is entering a brand-new stage. More data science research organizations have been established, including organizations in the USA, Canada, Australia, China, UK, Japan, and Korea; journals and proceedings have also been published (Zhu & Xiong, 2011).

In industry, the data scientist's role is fast becoming an in-demand and sought-after career. The EMC Corporation has built a community of data scientists and issued a survey of the global data science community (EMC, 2011). The LinkedIn data science team has been built by the world's largest professional network, LinkedIn. An increas- ing number of companies, such as Google, Facebook, IBM, PayPal, and Amazon, are also seeking data scientists to join their data science teams and help them maintain an innovative edge in the big data era.

In academia, Bell Labs published a data science action plan to enlarge the field of statistics in 2001 (Cleveland, 2001). In 2002, CODATA published its first refereed journal, the *Data Science Journal*. The Journal has become a salon for data scientists and experts in other fields (Iwata, 2008). Another publication is the *Journal of Data Science*, published by Columbia University. In 2009, the first data science monograph *Dataology and Data Science* was published (Zhu & Xiong, 2009). In 2012, Springer and EPJ.org published a Springer Open Journal, EPJ Data Science (www.epjdatascience.com/). More universities are starting to build data science research centers, such as the Institute for Data Sciences and Engineering at Columbia University and the Shanghai Key Laboratory of Data Science, Fudan University, China. UC Berkeley offered a data sci- ence course, Introduction to Data Science, in 2012. Columbia University began a course named Introduction to Data Science in 2011. Meanwhile, conferences and workshops on data science have been held in recent years, such as the Data Sciences Summer Institute (DSSI) hosted by UIUC (University of Illinois at Urbana- Champaign) in 2011 and 2012; the annual workshops on data science held by Fudan University since 2010. In 2014, the First International Conference on Data Science (ICDS) was held in China.

4.2 Research Issues in Data Science

Observation and logical reasoning are the basis of scientific research. In data science, we should focus on observation methods in data nature and data reasoning as well as the fundamental theories and technol- ogies. Data science requires more fundamental theories and new methods and techniques; for instance, the existence of data, the measurement of data, time in cyberspace, data algebra, data similarities and the theory of clusters, data classification and a data encyclopedia, data camouflage and data perception, data experiments, data awareness, etc. Data science will also improve the current research methods for scientific research in order to form new methods and develop specific theories, methods, and technologies in various fields. We should emphasize how to identify truth in data, how to support other scientific research, and how to acquire valuable knowledge from data.

The main issues in data science include:

i) **Foundational theories of data science**
   a) The theory of data similarity – data similarity is the key element in measuring the relationships among data for data analysis. Research topics include the definition of a similarity measure, computation of similarities, similarity measure properties, evaluation criteria of similarity func- tions, etc. Construction of the similarity theory is a solution to the core problems of data mining and big data analysis. Achievement in this research direction will impact the development of data technology.
   b) Data measurements and data algebra – a complete and correct theory of data computing is vital to data science. The RDBMS (Relational Database Management System) was fine when data natu- rally fit into tables, but it was known from the beginning that the Relational Model of Data was incomplete. The imperfection of the model became obvious primarily because of the difficulties experienced when using the relational database (RDBMS) with particular data structures. This topic should construct an algebraic system for various types of data.
   c) Data science research methods – the basic research methods for data science include data explo- ration, data experiments, and data perception. Data exploration explores the characteristics and structure of data

sets so that we can assess the value of data sets and select methods for analyzing the data sets. Data experiments check and verify hypotheses and the laws of nature or datanature. Data perception transfers data in perceptible ways through the five senses: vision, hearing, touch, smell, and taste.

ii) **Exploration of datanature**
    a) The fundamental rule of data – as mentioned above, research achievements from nature or human society are stored in cyberspace in the form of data, which forms datanature. The exploration of datanature will be on a higher level than before, allowing us to examine whether many princi- ples and laws in nature also can be found in datanature, such as prime numbers, the Fibonacci sequence, the golden ratio, the Pareto principle, etc. This topic includes research on datanature size, data growth patterns, data truth, data growth's impact on human society (e.g., how does data growth affect energy sources?), etc. These problems are not discussed in the natural and social sciences.
    b) The classification of data and a data encyclopedia – classifying data is helpful in understanding datanature. This topic will research standards for data classification, the ontology of data, the con- struction of a data encyclopedia, etc.

iii) **Data technology and its applications**
    a) Scientific research methods with data – computers are used in almost all scientific research and huge amounts of data are stored in computer systems. Scientific research is confronted with a pro- found need for reform in terms of research approaches. Data methods are new ways for scientific research to improve efficiency and results.
    b) Domain-driven data technique – present-day scientific research requires the integration of mul- tiple methods; for example, the combination of biological experiments and computation yields bioinformatics. One important issue is how it is possible to integrate data methods into a specific research area. New data technologies will be the special technologies aimed at different fields and circumstances instead of general technologies.
    c) Big data technique and its applications – this topic explores the requirements from various appli- cations and abstracts new types of data analysis tasks. Improving efficiency in dealing with big data is of primary importance.

4.3 Future Directions and Plans

More and more scholars are willing to participant and actively promote data science. They strongly agree that we all should spend more time and effort to explore fundamental theories and innovative technologies of data science and build up more and wider communication and cooperation among various disciplines and different backgrounds because there are still many problems to be solved and more problems might arise because of our endeavors. This is not a short-term plan but will be a task lasting half a century or even more.

In focusing on this new science, scientists should:

- Engage in developing data science as a new science and let it show its potential rather than only developing some individual or separate data analysis methods and techniques;
- Clarify and improve the definitions (including context and boundary) on data science;
- Explore the differences and relationships between data science and other related areas;
- Build up the theories of data science;
- Define and illustrate research topics, themes, directions, and key issues;
- Explore the methodology of data science;
- Develop data science combined with domain knowledge (e.g., bioinformatics, social networks);
- Construct more research institutes and centers for data science;
- Hold a workshop once a year and organize related international conferences on data science regularly;
- Incorporate people from related backgrounds (e.g., mathematics, statistics, physical science, neuroscience, systems theory);
- Train graduate students and provide student exchange opportunities;
- Seek cooperation between colleges and enterprises and apply for funding jointly;
- Establish an open international research platform;
- Publish workshop proceedings and an international refereed journal on data science.

5 Conclusions

There is unanimous agreement that data science is different from existing technologies and established sciences and will be a meaningful and promising research direction in the future. Data related research can and should lead the transition towards this new science – data science. Meanwhile, data researchersshould transfer into data science rather than developing individual or separate data analysis methods and techniques on their own. We believe that data science will become a new kind of science, which is exactlythe same as the natural sciences and social sciences.

6 Acknowledgment

6 References

Cao, L. B. & Yu, P. S. (2009) Behavior Informatics: An Informatics Perspective for Behavior Studies. *IEEE Intel-ligent Informatics Bulletin 10*(1), pp 6–11.

Cleveland, W. S. (2001) Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statis-tics. *International Statistical Review 69*(1), pp 21–26.

Dhar, V. (2013) Data Science and Prediction. *CACM 56*, p 12.

EMC (2011) Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field. Retrieved from theWorld Wide Web November 11, 2014: http://www.emc.com/collateral/about/news/emc-data-science-study-wp.pdf

Hayashi, C. (1996) What is Data Science? Fundamental Concepts and a Heuristic Example. In *Proceedings of the 5th Conference of the International Federation of Classification Societies (IFCS'96)*.

Hey, T., Tansley, S., & Tolle, K. (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Iwata, S. C. (2008) Editor's Note: Scientific 'Agenda' of Data Science. *Data Science Journal 7*, pp 54–56.

Liu, L., Zhang, H., Li, J. H., et al. (2009) Building a Community of Data Scientists: an Explorative Analysis. *Data Science Journal 8,* p 24.

Loukides, M. (2010) What is Data Science? An O'Reilly Radar Report.

Naur, P. (1966) The Science of Datalogy. *Communications of the ACM 9*(7), p 485.

Smith, F. Jack (2006) Data Science as an academic discipline. *Data Science Journal 5*, pp 163–164.

Zhu, Y. Y. & Xiong, Y. (2009) Datalogy and Data Science (in Chinese with English abstract). Fudan UniversityPress.

Zhu, Y. Y. & Xiong, Y. (2011) Datalogy and Data Science: Up to Now. Retrieved from the World Wide WebNoember 16, 2014: http://www.paper.edu.cn/en_releasepaper/content/4432156

Zhu, Y. Y., Zhong, N., & Xiong, Y. (2009) Data Explosion, Data Nature and Datalogy. In *Proceedings of Inter-national Conference on Brain Informatics (BI'09)*.