e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

APPROACHING DATA SCIENCE

Aishwariya Panda*1

*1Department Of IT (Information Technology), S.I.W.S College, Mumbai, India.

ABSTRACT

The amount of data being produced on the internet nowadays is enormous. Due to a data boom, data nature—all data in cyberspace—is currently taking shape. It is important yet challenging to investigate the patterns and laws of data nature. Data Science is a new field that will soon be established. It offers a new research approach (a data-intensive approach) for the scientific and social sciences and extends beyond computer science in terms of data analysis. 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.

I. INTRODUCTION

The big data age has been ushered in by the data explosion, which is the quick expansion of data in cyberspace. Data now have more meaning than before. Data are no longer restricted to measurements' outcomes, the values of qualitative or quantitative variables, or scientific data produced in the setting of experiments and observations. Additionally to all of that, data make up the entirety of cyberspace. According to Zhu, Zhong, and Xiong (2009), the formation and development of Data nature (all the data in cyberspace) occur unintentionally. Data that has no analogs in the natural world is becoming more and more common, including garbage data, online games, and computer viruses, all of which are produced by data nature. The facts found in the natural world have increasingly been exceeded by the information created by data nature, which now displays distinctive patterns.

Since the invention of the computer, we have continuously used and worked with data. In order for us to retrieve the facts of the natural world as needed, computers map and store them as data. But, especially in the realm of research, data has changed from simple data access to huge data analysis (e.g., life science). Big data must confront new needs and challenges as a result, which motivates research on the data itself, such as how to comprehend life through DNA data. The reason for just using data is also shifting. Data analysis comprises investigating the phenomena and patterns included in the data itself and in addition to applying the data to address difficulties in the real world (such as identifying growth trends and estimating the volume of data in cyberspace 10 years from now). By introducing data technologies and methodology to the natural and social sciences as well as by looking into the nature of data, the transition to this new field, known as data science, may and ought to lead. Whether you like it or not, embrace it or not, and are prepared for it or not, data science is here to stay. If you've been doing data science research, you could already be a data scientists could already be a data scientist if you've been doing data science research.

This essay examines why data science is necessary and the problems that data presents. We also discuss how data science is different from established technology and scientific disciplines. We also go through several important problems that data science will confront when it develops into an academic discipline with data as its study subjects, including core theories, fresh approaches, and research areas. We also examine the developments in the field's present research and society and talk about some of the prospects and difficulties it faces. Finally, we provide examples of applying prior knowledge to this new discipline and science.

II. TWO OBSTACLES TO WORKING WITH DATA

The analysis of data confronts increasing difficulties as its importance grows over time. The issues raised below are discussed.

2.1 Data Accuracy:

How can we identify if the information we've been using is accurate or misleading? How do we handle a dataset with erroneous data? How can we gauge a dataset's confidence level if both false and accurate data are present? For instance, certain product reviews may not be legitimate if they are provided by consumers who haven't

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

really utilized the items or even by rivals. As a result, the analytical findings (such as credit ratings) based on a dataset including such data will also be unreliable.

These are significant obstacles in the field of data-related research and will play a significant role in the study of data science. The difficulties are becoming more and more difficult as social networks like Facebook and blogs grow.

2.2 Conflicts with the Online platform Survival:

As the human experience develops, it will eventually embrace cyberspace as well as physical place, thus we will coexist there. How do we live in the era of technology? How to interact online, for instance, is one of the key problems with survival. This might end up being one of the toughest problems in the future of data-related research due to challenges in the communication environment. Actually, there is a fix for this problem available already. One of the computer languages that teenagers use to communicate online is the Martian Language, which serves as an illustration of a cyberspace communication method. Martian Language is exceedingly difficult for most individuals to grasp all at once since it has terminology from many different languages, including English, Chinese, Japanese, and others, which have been adopted and blended.

2.3 Data-Based Scientific Research:

Data are utilized to learn since they represent or map the facts in the natural world. The laws of reality. The study of data-nature phenomena and laws provides evidence for finding patterns and laws in the natural world. Creating techniques in data nature to a promising research area that will be beneficial for scientific study is the exploration of natural laws. Data scientists are aware that, given the tremendous growth in the amount of data in numerous sectors, many traditional approaches to solving issues are ineffective, and they recognize the significance of data in scientific research.

2.4 Data-Based Knowledge Acquisition:

Early computer science historians concentrated on enhancing computing's performance and capabilities. However, a more pressing issue at the moment is how to extract useful information from the growing volume of data being produced, for instance from both the natural and physical sciences. How can we locate relevant info in cyberspace, for instance? How can data be used to get knowledge? These call for a fresh perspective on how we comprehend and analyze data.

We explore below why data science is necessary and where it intersects with other disciplines, keeping in mind the aforementioned problems.

III. WHAT ARE THE DIFFERENCES, THEN?

The process of informationization creates data by storing natural-world items or phenomena in the form of digital data. Data serve as a depiction of nature by documenting human activities, such as labor, means of subsistence, and societal advancement. At the moment, cyberspace is producing a growing amount of data quickly. We refer to this as data explosion. In cyberspace, data explosion creates data nature. Data are a singular thing, thus it is important to examine and understand the data regulations in cyberspace. While investigating the laws of the cosmos, life, human behavior, and social evolution, this research and exploration are significant. We can use data, for instance, to research living things (i.e., bioinformatics) or look into people's behaviors (i.e., behavior informatics; Cao & Yu, 2009).

3.1 Compared to Other Data Technologies, differences:

Since the creation of the computer, approaches to handling data, including data storage, data exchange, and data access, have evolved. Data science challenges have grown and expanded well beyond those in the field of computer science. Data gathering, storage and administration, security, analysis, and visualization are only a few of the tools and methodologies that data science use, although in very different ways from conventional methods. Data mining, information retrieval, data integration, and artificial intelligence are just a few examples of domains where there are overlaps, yet the distinctions are still significant. Fundamental ideas and innovative methods are needed for data science.

In essence, computer science uses computer languages to build models of the actual world, including humans and their activities, so that they may be stored in computer systems. Data in the form of facts are kept in these computer systems. Dealing with data is the work of modeling. As a result, computer science data technologies

were created to be applied to the construction of models for facts and programs as well as the calculation of data using computer systems. The science of data utilization has many more explanations besides this one.

Data processing and data analysis technologies, such as data integration and data mining, are the current emphasis of computer science research. In the area of large-scale data analysis, data mining is a technology that has received a great deal of attention. Researchers there have been creating Explosive transactional and behavioral data explanation and prediction algorithms and tools. Computer science's data mining field focuses on data analysis. However, in the broader subject of data science, "data mining" represents a considerably narrower collection of concepts (Dhar, 2013). Additionally, computer scientists have been at the forefront of data research (such as data mining technologies), therefore papers and conferences connected to this have come from the Association for Computing Machinery or the Institute of Electrical and Electronics Engineers (IEEE) (ACM). In reality, there are more and more academic fields (like bioinformatics) that concentrate on data analysis. The conferences and publications that are connected to these fields are not produced by IEEE or ACM.

One area of data science is the study of how to use data to represent reality, how to organize and use data, and how to create data technology utilizing computers.

3.2 Contrasts with big data:

Data science is driven by big data in the industry. According to Dhar (2013), one of the consequences of data science is how scientists may utilize huge data to their advantage when doing research. Expanding and we have the chance to collect enormous data sets from data nature because of the expanding amounts of data that are kept in cyberspace. We can perform more and better data studies since it is so simple to obtain such large data. Due to their size and complexity, big data cannot be processed using the current data technologies. New data technologies are therefore necessary. Big data technologies have advanced recently. One of the study topics in data science is the development of big data technologies. Data science also includes using big data to address a variety of issues in the social and scientific spheres; big data is one of the top/hottest study topics in data science.

3.3 Differences from Other Sciences:

Differences from Other Sciences: In computer systems, data is the formal representation of the natural world. Although data and knowledge may be regarded as symbols and representations of each other, they are not the same. Research goals, goals, and methods in computer science, information systems, and information science are significantly different from those in data science.

Data science helps social science and natural research, on the one hand. One of the motivating factors behind data science is dealing with data. For natural science and the social sciences, data science offers a new sort of research methodology known as the Scientific Research Method with Data. As a result, data science is often known as a science that uses a lot of data (Hey, Tansley, & Tolle, 2009). For instance, a fundamental experimental course is life science. However, it always takes a long time for scientists to complete an experiment. Today, scientists may do more with their biological data analysis since bioinformatics helps speed up and streamline these laborious procedures. For instance, bioinformatics uses biological data from shotgun sequencing to produce significant discoveries. The field of bioinformatics shows how we may study life using biological data by transforming life science from an experiment-based science to a science integrating computations and experiments. Data-driven biology research also addresses several fresh issues that can't be handled by conventional approaches.

In contrast, more and more scientific research will focus on the data in nature rather than the facts in nature, which will encourage people to identify data and make it easier for them to study both nature and human behavior. Both natural science and social science use elements of the natural world as their research subjects. However, as more and more data emerge without ties to nature or human behavior, they are progressively outpacing and replacing the facts about such things in cyberspace. In contrast to natural science and social science, data researchers typically conduct their study online, using data as their research subjects.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

IV. MAKING THE TRANSFORMATION TO DATA SCIENCE

4.1 Modern technology

Data science has piqued a lot of curiosity. In order to back up his contention that "datalogy" should be used instead of "computer science," Peter Naur (1966) created the word "datalogy" (also known as the "science of data"). The term "data science" approach is capable in the 1990s (Smith, 2006). The term "data science," which was later made a reality by the Data Science Journal in 2002, was first coined by CODATA (the Committee on Data for Science and Technology) (www.codata.org), a representative of the scientific data research sector, to deal with obtained from multiple scientific research domains. Only certain study material, scope, and themes have been mentioned as a formal definition of "data science" (Smith, 2006; Hayashi, 1996; Liu, Zhang, Li, et al., 2009). In 2010, Loukides (2010) discussed what data science is, looked at some aspects of it, such as technologies, businesses that do data science work, and the special skill sets related to it, and made the case that data science should enable the creation of data products rather than just be thought of as an application with data. Data is the research object for data science, according to Zhu et al 2009 (Zhu, Zhong, & Xiong, 2009; Zhu & Xiong, 2009).

Data science is currently moving into a new phase. There are now more organizations dedicated to data science research, including ones in the USA, Canada, Australia, China, the UK, Japan, and Korea. Additionally, journals and proceedings have been released (Zhu & Xiong, 2011). In business, the position of a data scientist is quickly rising in popularity and demand. The EMC Corporation has established a data scientist community and conducted a study of the data science community worldwide (EMC, 2011). The largest professional network in the world, LinkedIn, has established a data science team. Companies like Google, Facebook, IBM, PayPal, and Amazon are among those looking for data scientists to join their data science teams and keep them on the cutting edge of innovation in the big data age.

In order to broaden the subject of statistics in academia, Bell Labs established a data science action plan in 2001. (Cleveland, 2001). The Data Science Journal, the first peer-reviewed publication from CODATA, was released in 2002. The Journal has developed into a gathering place for data scientists and other specialists (Iwata, 2008). The Columbia University-published Journal of Data Science is another work. The first monograph on data science, Dataology and Data Science, was released in 2009. (Zhu & Xiong, 2009). A Springer Open Journal, EPJ Data Science, was released in 2012 by Springer and EPJ.org (www.epjdatascience.com). A growing number of institutions are beginning to establish data science research facilities, such as the Shanghai Key Laboratory of Data Science at Fudan University in China and the Institute for Data Sciences and Engineering at Columbia University. In 2012, UC Berkeley provided a data science course called Introduction to Data Science. 2011 saw the debut of the Introduction to Data Science course at Columbia University. Data science conferences and workshops have also been organized in recent years, including the Data Sciences Summer Institute (DSSI) sponsored by UIUC (the University of Illinois at Urbana-Champaign) in 2011 and 2012 and the yearly workshops on data science held by Fudan University since 2010. The inaugural International Conference on Data Science (ICDS) took place in China in 2014.

4.2 Data Science Research Issues

The cornerstones of scientific inquiry are observation and deductive reasoning. We should concentrate on observational techniques in data nature and data reasoning as well as the underlying ideas and technologies in data science. For example, the presence of data, its measurement, time in cyberspace, data algebra, data similarity and the theory of clusters, data categorization and a data encyclopedia, data camouflage and data perception, data experimentation, data awareness, etc. are all necessary for data science. In order to create new ways and build specialized theories, methods, and technologies in a variety of domains, data science will also enhance the present research methodologies for scientific study. We should place special emphasis on how to recognise data truth, how to encourage related scientific study, and how to get useful knowledge from data.

The fundamental problems in data science are as follows:

1) Data science's fundamental theories:

a) Data similarity is the essential component in determining the links between data in data analysis. The definition of a similarity measure, computing similarities, similarity measure characteristics, evaluating

similarity functions, etc. are all possible research subjects. The development of the similarity theory provides a solution to the fundamental issues with large data analysis and data mining. The advancement of data technology will be impacted by success in this study area.

b) A thorough and accurate theory of data processing is essential to data science, as are data measurements and data algebra. The Relational Model of Data was acknowledged to be flawed from the start, yet the RDBMS (Relational Database Management System) worked great when data fit into tables naturally. The challenges encountered while using the relational database (RDBMS) with certain data structures made the model's flaws very clear. This subject should build an algebraic framework for different kinds of data.

c) Data exploration, data experimentation, and data perception are some of the fundamental research techniques used in data science. Data exploration is investigating the structure and properties of data sets so that we can determine their worth and choose the best methods for evaluating them. Data experiments test and validate hypotheses as well as natural laws or the nature of data. The five senses—vision, hearing, touch, smell, and taste—transmit data in detectable ways.

ii) Examining the nature of data

a) The key tenet of data is that it must be saved in cyberspace in the form of data. As was already said, this is the essence of data. The research of data nature will take place at a deeper level than before, allowing us to investigate if numerous natural laws and principles, such as prime numbers, the Fibonacci sequence, the golden ratio, the Pareto principle, etc., may also be discovered in data nature. Research on data size, growth patterns, veracity, and the effects of data growth on human civilization (such as how data growth influence energy sources?) are all included in this area. The scientific and social sciences do not address these issues.

b) Data categorization and a data encyclopedia - Data classification aids in the comprehension of the data's nature. This topic will investigate data categorization standards, data ontology, the creation of a data encyclopedia, etc.

iii) Information technology and its uses

a) Data-driven scientific research techniques - Almost all scientific research is conducted using computers, and massive volumes of data are saved in computer systems. There is a critical need for research strategy improvement in scientific research nowadays. Data techniques are brand-new approaches to increase the effectiveness and output of scientific research.

b) Domain-driven data approach - Modern scientific research necessitates the fusion of many methodologies; for instance, the fusion of biological investigations with computation results in bioinformatics. How to incorporate data approaches into a particular study topic is one crucial challenge. Instead of being generic technologies, new data technologies will be specialized technologies directed at various sectors and contexts.

c) The needs of diverse applications are explored in this subject, which also abstracts new kinds of data analysis jobs. Topic c: Big data technology and its applications. The main priority is increasing efficiency while handling huge data.

4.3 Future Plans and Directions

More academics are becoming eager participants and enthusiastic advocates of data science. They fervently concur that since there are still numerous issues that need to be resolved and more issues may develop as a result of our efforts, we should all put more time and effort into investigating fundamental theories and cutting-edge technologies of data science. We should also expand our communication and cooperation between various disciplines and backgrounds. This endeavor will take at least fifty years to complete; it is not a short-term strategy. When concentrating on this emerging field of study, scientists should:

- Engage in the development of data science as a new field of study and allow it to demonstrate its potential rather than just creating a few unique or distinct data analysis methods and procedures,
- Make definitions of data science (including context and boundary) clearer and better,
- Build up the ideas of data science,
- Define and explain research topics, themes, directions, and major concerns. Explore the distinctions and connections between data science and other related fields,
- Research data science methodologies,

• Develop data science in conjunction with subject-specific expertise (such as bioinformatics and social networks),

• Build additional data science research institutions and centers;

• Host a workshop on the subject once a year and periodically plan relevant international conferences,

• Include individuals with relevant backgrounds (such as those in mathematics, statistics, physical science, neurology, and systems theory),

• Train graduate students and offer possibilities for student exchanges,

• Establish an open worldwide research platform,

•Seek collaboration between universities and businesses,

• Submit funding applications together; publish workshop proceedings and an international refereed journal on data science.

V. CONCLUSION

Everyone agrees that data science is distinct from current technologies and established sciences and that it will be an important and fruitful study area in the future. Data-related research should guide the development of this emerging field of study, called data science. Instead of independently establishing unique or distinct data analysis methodologies and processes, data researchers should transition to data science. We predict that data science will develop into a distinct branch of knowledge, much like the scientific and social sciences.

ACKNOWLEDGEMENT

Shanghai Science and Technology Development Funds (13dz2260200, 13511504300), as well as NSFC-61170096, have contributed to this effort.

VI. REFERENCES

[1] The authors of Behavior Informatics: An Informatics Perspective for Behavior Studies are Cao, L. B., and Yu, P. S. IEEE Intelligent Informatics Bulletin, volume 10, issue 1, pages 6–11.

[2] In 2001, W. S. Cleveland published Data Science: An Action Plan for Increasing the Technical Aspects of the Statistics Field. 21–26 are included in International Statistical Review, 69(1).

[3] Data science and prediction, Dhar, V. (2013). CACM 56, p 12. EMC (2011). Data Science Revealed: A Data-Driven View into the Emerging New Field. The following was taken from the Internet on November 11, 2014: http://www.emc.com/collateral/about/news/emc-data-sciencestudy-wp.pdf

[4] What Is Data Science, 1996, C. Hayashi Basic Ideas and an Example of a Hayashi, C? The 5th Conference of the International Federation of Classification Societies (IFCS'96) was documented in its proceedings.

[5] The Fourth Paradigm: Data-Intensive Scientific Discovery is a 2009 book by T. Hey, S. Tansley, and K. Tolle. Google Research.

[6] S. C. Iwata (2008) Editor's Note Science's "agenda" for data analysis. Journal of Data Science 7, pages 54–56.

[7] Zhang, H., Li, J. H., and others with Liu, L. (2009) An Exploratory Analysis of the Construction of a Community of Data Scientists. 8, page 24, Data Science Journal.

[8] What is Data Science, according to M. Loukides (2010)? A Radar Report from O'Reilly.

[9] Naur, P. (1966) Datalogy as a Science. Communications of the ACM, volume 9 (7), page 485.

[10] Jack Smith, F. (2006) Academically speaking, data science. Journal of Data Science 5, pages 163–164.

[11] In 2009, Zhu, Y. Y., and Xiong published Dataology and Data Science (in Chinese with English abstract). Press of Fudan University.

[12] Y. Y. Zhu and Y. Xiong (2011) Data science and data ology thus far. On November 16, 2014, this information was obtained from the Internet at:
http://www.paper.edu.cn/en release paper/content/4432156.

[13] Data Explosion, Data Nature, and Dataology. Zhu, Y. Y., N. Zhong, & Y. Xiong. International Conference on Brain Informatics (BI'09) Proceedings.