LINGUISTICA ANTVERPIENSIA, 2021 Issue-2
www.hivt.be
ISSN: 0304-2294

Research Challenges and Opportunities in Data Science - An Intensive Review

[1]Dr M Srinivasa Narayana, [2]Swapna Donepudi, [3]Dr. Deepmala Singh, [4]Dr. B. S. Kiruthika Devi, [5]Dr. Ankireddy.yenireddy, [6]M Geetha

[1]professor , KL business school , KL university Vijayawada

[2]Assistant Professor, Department of CSE, PVPSIT.

[3]Assistant Professor, Management Department, LBEF Campus, Kathmandu (with academic collaboration to Asia Pacific University, Malaysia)

[4]Assistant Professor,Department of Information Technology,Sri Sairam Engineering College, Chennai.

[5]Prof. andHoD Loyola Institute of Technology AND management

[6]Research scholar ,KL business school kl university Vijayawada

Issue Details
Issue Title: Issue 2
Received: 25 March, 2021
Accepted: 27 April, 2021
Published: 15 May, 2021
Pages: 75 - 80

Abstract

Significant amounts of data are obtained and prepared for research, visualisation, and management in the field of data science. Data science is inextricably linked to databases and computer science. Since data science is more complicated than data analysis, it necessitates a wide range of skills. A variety of tasks are required in data science to interpret histograms and averages, which is an essential part of the skill set.Data scientists take an active role in data engineering, data collection, data processing, and data archiving design and implementation. The focus of this paper is on a thorough examination of processes, various problems, open issues, and research perspectives in data science.

Keywords: Data science, Big data, Data analytics, Data scientists, Visualization.

I.INTRODUCTION

Data science is an extension of data mining in that it collects information from vast quantities of unstructured data. Messages, functionality, images, and online networking content are examples of unstructured data. Information science is concerned with the measurement of data and the composition of measurements in order to concentrate the knowledge gained from this data. Data science is therefore a mixture of data and an unquenchable appetite for a response derived from data [1][6].Data science is described as the ability to take data, comprehend it, process it, extract value from it, visualise it, and communicate it. Information planning and machine learning are often used in data science to boost ads, extortion detection, and setting open strategy. It also makes use of the capability to locate information sources, code them, and combine them. Data science also ensures the accuracy of datasets to aid in data comprehension and builds scientific models to make use of the data [4]. The display of knowledge discoveries is depicted in Figure 1 as the basic flow control of a data science operation.

ISSN: 0304-2294

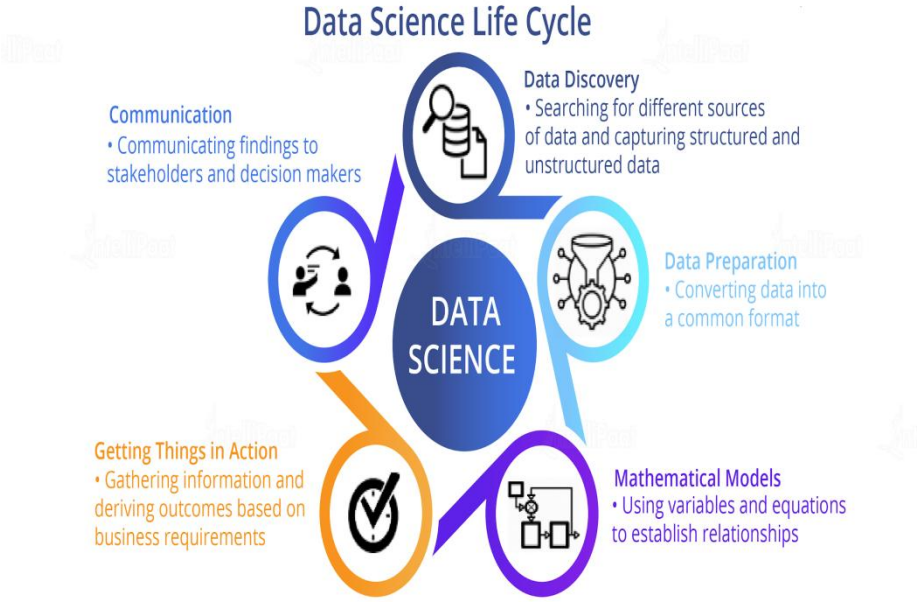

Figure 1. Data science Life cycle

This paper may also benefit from a broad discussion of the problems and future directions. The topic of data science and its various aspects is discussed.

The paper's key contribution is that

1. Discussing the evolution of data science.
2. Evolving application details are thoroughly defined.
3. Data science's most challenging directions are explored in depth.

The above-said points clearly distinguish this survey from other recent surveys. It gives the detail as broad as earlier works. The paper is organized as follows: Section II reviews the process of data science. Section III discusses data science evolution. Section IV discusses the emerging applications and Section V concludes with possible research directions.

II. DATA SCIENCE PROCESSES

Arranging, bundling, and conveying knowledge are three segments of the data science process. Data collection and sorting are part of the bundling process. A data science researcher must understand what the data science transform's yield will be, how it will be achieved, and who will be involved in the yield's development. Collection and preparation of data, analysis to interpret outputs, and distribution of findings in the form of written reports or executable code are all steps in the data science process [3]. Figure 2 depicts the basic steps involved in data science.

ISSN: 0304-2294

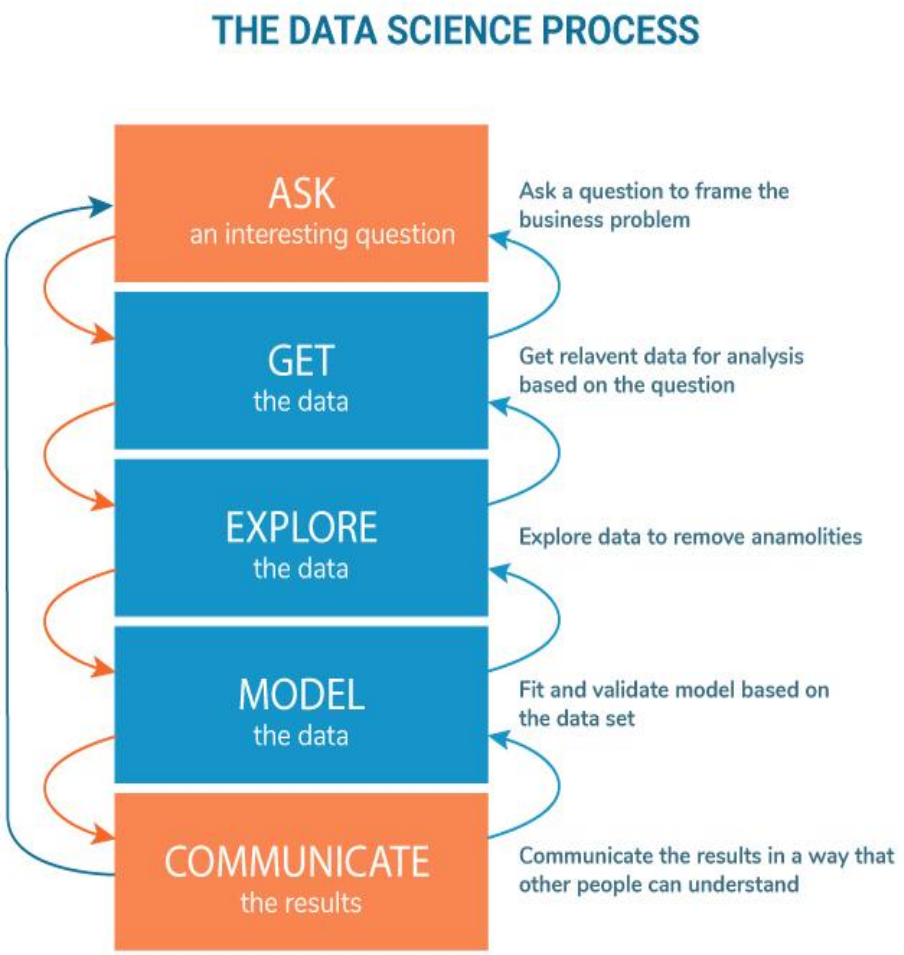

Figure 2. Data science processes

2.1 Data wrangling

It is the method of collecting data from relevant areas and transforming raw data into another format for manipulation using semi-automated tools including data wrangling[4]. Data sorting is the process of arranging data according to best practises in information management. It essentially entails accelerating growth and skills.

2.2 Data bundling

The process of controlling fundamental knowledge into another representation and package is known as data bundling. It is the polar opposite of data sorting. It also entails transitioning people and systems from new to existing. Data analysis is the study of data for the purpose of evaluating, modifying, and explaining knowledge in order to find useful data for decision-making. To extract meaningful knowledge from large quantities of data, various statistics and machine learning algorithms are used [8].

2.3 Data conveying

Conveying data is the method of converting statistical results derived from data into a format that can be easily understood by users. Data transmission is a growth empowerment strategy that begins with one viewpoint on existing technology, making it look as if it is an integral part of the system [2].

III EVOLUTION OF DATA SCIENCE

Tukey coined the word bit in 1962, as seen in [7], and Claude Shannon used it in his 1948 article, A Mathematical Theory of Communications. Exploratory data analysis argued in [9] that more focus should be put on using data to test hypotheses, and that exploratory and confirmatory data analysis should be done concurrently. [5] discusses a study of contemporary data processing models. “The science of dealing with data, once they have been created, whereas the relation of the data to what they represent is assigned to other fields and sciences,” according to another concept of data science.The International Association for Statistical Computing (IASC) was established to bridge the gap between conventional statistical methods and modern computer technology in order to transform data into information. In 1995, the first workshop on information discovery in databases was held, and it has since evolved into the annual conference on knowledge discovery and data mining [10].

In September 1994, database marketing was featured on the cover of Business Week. Companies gathered massive quantities of data in order to determine how likely a consumer is to purchase a product. The data is also used to create a marketing message that will elicit the desired response from customers[11]. Due to being overwhelmed by the volume of data, a previous burst of excitement in the 1980s ended in widespread disappointment. For the first time in 1996, data science was included in the title of a conference, “From Data Mining to Information Discovery in Databases” [12].Several publications in the 1990s illustrated the growing importance of data mining and information discovery in databases [13]. [12] proposes merging the contexts of computer science and data mining in order to broaden the main fields of statistics and to suggest significant improvement in the field of data science. [13] proposes computational data modelling - stochastic data models and algorithmic models. Data science's evolution demonstrates its importance in modern informatics. In the following part, we'll look at some of the applications of data science.

IV EMERGING APPLICATIONS

Data science has spread from statistics and analytics to all fields of science and business in real-world applications. The most popular data science applications are currently in use. Business analytics is a type of data science that is used to collect data about a company's past and current performance in order to improve decision-making processes and forecast future performance [14]. With vast volumes of data obtained and analyzed, data science is used in prediction. It can be used to find patterns and create predictive models using machine learning [15], where information is discovered by induction algorithms. Machine learning is being used to build security predictive models. Data from user logs is obtained using data science to detect fraud [9]. Machine learning algorithms are used to discover patterns that can be used to find malicious insiders and to avoid fraud.

Computer vision is a branch of computer science that uses data from image and video processing to apply computer vision and teach computers to see for the

purpose of collecting and analyzing images. In robots, vehicular ad hoc networks, and human-computer interaction systems, data science implementations of computer vision are used. Natural language processing computer science employs massive quantities of textual data to statistically model linguistic data for tasks such as machine translation, parsing, and sentiment analysis [11][16].In bioinformatics, data science is the use of computers to understand biological data, such as genetics and genomics [14][17]. Many types of data with genomic data in disease research are examined as data science is combined with new informatics approaches. Other applications of data science include astronomical data from modern telescopes and climatic data for climate modelling [12][18].

A lack of data points hampered sales management, and the retail industry uses data science to increase overall revenue by delivering the right product to the right consumer. Data science is used by the government to prevent waste, fraud, and violence, as well as to fight cyberattacks and protect sensitive data. Business intelligence is also used to make better financial choices in order to strengthen defence systems. Most governments also employ data science models for a number of purposes. Data science is a method of transforming data into useful knowledge that has been popular for a long time. The need for statisticians and programmers to efficiently use data according to value and use it for knowledge mining [15] is what drives them.

V. CONCLUSION AND FUTURE DIRECTIONS

People will apply data science to all problems in the near future, and it is possible that some truly amazing applications, apart from online applications (recommendations, ad targeting, and so on), will arise. The skills required for visualization, client interaction, and the development of sellable algorithms can change. Enough businesses will begin to consider task diversification. Companies will form teams with complementary abilities, and data analytics will be used to customise services.Person-level customization can be accomplished using data science in services such as healthcare, insurance, public services, banking, and so on. Information makes future tasks simpler, and the research process becomes more effective as more data is collected. Material sciences, drug discovery, quantum mechanics, neuroscience, nanotechnology, and other fields use data science. People will have more options and computerization, and computers will have more expertise, thanks to the application of data science in large data examinations. These advances can be used to improve the intelligence of both people and objects.

REFERENCES

1. Akshi Kumar and Teeja Mary Sebastian 2012, “Sentiment Analysis on Twitter”, International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July.
2. Brodie, Michael. (2019). What Is Data Science?. 10.1007/978-3-030-11821-1_8.
3. Brumfiel and Geoff 2011,"High-energy physics: Down the petabyte highway", Nature 469, pp. 282–83.
4. Cao, Longbing. (2017). Data Science: A Comprehensive Overview. ACM Computing Surveys. 50. 1-42. 10.1145/3076253.
5. Davenport and Thomas H 2006, "Competing on Analytics", Harvard Business Review.
6. Dhar, V 2013, "Data science and prediction", Communications of the ACM 56.
7. Jeff Leek 2013, "The key word in 'Data Science' is not Data, it is Science", Simply Statistics.

8. Mantri, Sanket. (2016). Data Science: Literature Review & State of Art. 10.13140/RG.2.2.18149.70880.
9. Parsons, MA, MJ Brodzik, and NJ Rutter 2004, "Data management for the cold land processes experiment: improving hydrological science", HYDROL PROCESS. 18:3637-653.
10. Raul Isea 2015, "The Present-Day Meaning of the Word Bioinformatics", Global Journal of Advanced Research, Vol-2, Issue-1 PP. 70-73.
11. ReinhardKlette 2014, " Concise Computer Vision", Springer, ISBN 978-1-4471-6320-6.
12. Salazar-Reyna, Roberto & Gonzalez Aleu, Fernando &Granda-Gutiérrez, Edgar & Díaz-Ramírez, Jenny & Garza-Reyes, Jose Arturo & Kumar, Anil. (2020). Systematic literature review of data science, data analytics, and machine learning applied to healthcare engineering systems. Management Decision. ahead-of-print. 10.1108/MD-01-2020-0035.
13. Song, Il-Yeol& Zhu, Yongjun. (2015). Big data and data science: What should we teach?. Expert Systems. 33. 10.1111/exsy.12130.
14. Statistical Modelling: the two cultures Leo Breiman. Statistical Science. Vol. 16 No.3 199-215, 2001.
15. Weihs, Claus &Ickstadt, Katja. (2018). Data Science: the impact of statistics. International Journal of Data Science and Analytics. 6. 10.1007/s41060-018-0102-5.
16. Murali. K and Siva Perumal, S. "Error Rate Performance Analysis of Power Domain NOMA over AWGN and Fading Channels with Generalized Space Shift Keying in Wireless 5G". Journal of Intelligent & Fuzzy Systems, vol. 40, no. 4, pp. 5771-5776, 2021. DOI: 10.3233/JIFS- 189416
17. Murali. K and Siva Perumal, S. (2020) "A novel rate improvement technique of power domain NOMA in wireless 5G" Journal of Scientific and Industrial Research, 79(10), pp. 902–904
18. Murali, K. and Siva Perumal, S. (2019) "Power signal based multiple access schemes for 5G and beyond- survey and its challenges" Journal ofAdvanced Research in Dynamical and Control Systems, 11(5), pp. 184–191