

AN ANALYSIS AND IMPRESSION OF DATA SCIENCE WITH TRENDS

1 Dr. R. Senthilkumar, 2J.Senthil, 3P.M.Manochitra, 4B.BeaulaPinky

1Associate Professor, Computer Science and Engineering,
Shree Venkateshwara Hi-Tech Engineering College, Erode, Tamilnadu, India.
yoursrsk@gmail.com

2Assistant Professor, M.Tech Computer Science and Engineering,
Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu,India.
senthilgobi05@gmail.com

3Assitant Professor, Computer Science and Engineering,
Shree Venkateshwara Hi-Tech Engineering College, Erode, Tamilnadu, India,
manochitra64@gmail.com

4Assistant Professor, Artificial Intelligence and Data Science,
Sri Shanmugha College of Engineering and Technology, Sankari, Tamilnadu, India.
beaulapinky@rocketmail.com

ABSTRACT

The goal of data science, which combines a number of fields, is to build the technology and algorithms necessary to solve complex problems. Today, data science is very important, and scholars are beginning to realize how important it is. Recent years have seen the publication of several research papers on data science, but each one has a unique subject in mind, such as the influence of data science on business, industry, education, and healthcare. An overview of data science will be provided in this article. Additionally, it makes clear how data science operates and what skills a data scientist needs in order to function in this sector. Additionally, it demonstrates current and upcoming data science trends.

Keywords—data science; data science trends; big data.

I. INTRODUCTION

The terms "data science" and "big data" are now among the most often used concepts across a variety of fields, including business, media, economics, education, and social networks [1, 2]. "Data scientist" is one of the most frequently used job titles in this field. The amount of data now in existence is expanding quickly, doubling every two years, and altering the life. According to IBM, 2.5 billion GB of data were produced per day in 2012. According to a Forbes article, data is growing more rapidly than ever before, and by the year 2020, every individual on the planet will produce around 1.7 MB of new information every second. The future is in the area of "data science," thus it is crucial to at least have a basic comprehension of it.

Big data is a crucial component of data science, which applies to both large and small datasets, with the data analysis process covering all outcomes [3]. Data science is based on three

main pillars: data, people, and technology. As a result of rapidly increasing capacity to collect, store, and analyze a constantly expanding range of data created by rising frequency, the discipline of data science is advancing. Given that data science is a relatively young area, it has received a large proportion of research attention, and a large number of studies have been published on it that may result in useful findings [4]. This work suggested a data science road map, including a theoretical foundation, new trends, and the future, to assist researchers in identifying the research gaps currently being investigated.

II. THEORETICAL BACKGROUND

A. Data Science

Data science has gained popularity over the last ten years as a result of the growth of several large internet companies like Google, Yahoo, LinkedIn, Amazon, and Facebook as well as numerous start-ups that have based their businesses on data like Everstring, Climate Corporation, Palantir, and Stitch Fix [2].

The world is getting smarter now as a result of the employment of computational and mathematical methods. Many fields are increasingly concentrating on intelligent interpretation and its analysis in accordance with the need for automation. For that goal, other approaches are in use in addition to "data science". Over the last several years, both the need for data scientists and the importance of their work have increased [5].

[6] note the expansion in the accessibility of enormous data sets as well as advancements in the optimization, machine learning (ML), and artificial intelligence (AI) industries. A new multidisciplinary field called data science is created by statistical advances, applied mathematics, and discrete new algorithms.

In other words, "data science" is a multidisciplinary approach to problem solving that combines data inference, technology, and algorithm creation [5, 7]. The goal of data science is to assist individuals in making wiser choices. For instance, in healthcare settings, the objective is to assist regulators, politicians, and patients in selecting the best course of action among a wide range of alternatives. Predictability is a key factor in data science's ability to enhance decision-making [8].

B. Data Science Workflow

Because big data is exploding, there is an increasing demand for practitioners and a sizable opportunity for academics to recognize the data science process and use new methods to build it [9]. Realizing from the term that "data science" has a broad range. That implies it is difficult to address every area of "data science." Because current "data science" does not have a lengthy history, it is difficult to identify a commonly accepted process. The procedure, however, is shown in Figure 1.

Source data access, data exploration and validation, model creation, output production, data monitoring, exploratory analysis, and deployment are some of the processes that are involved.

A model or visualization might be the outcome. Data scientists must repeat exploratory and experimental analysis until they are pleased with the report or model since the predictive modeling step resembles an inner workflow loop.

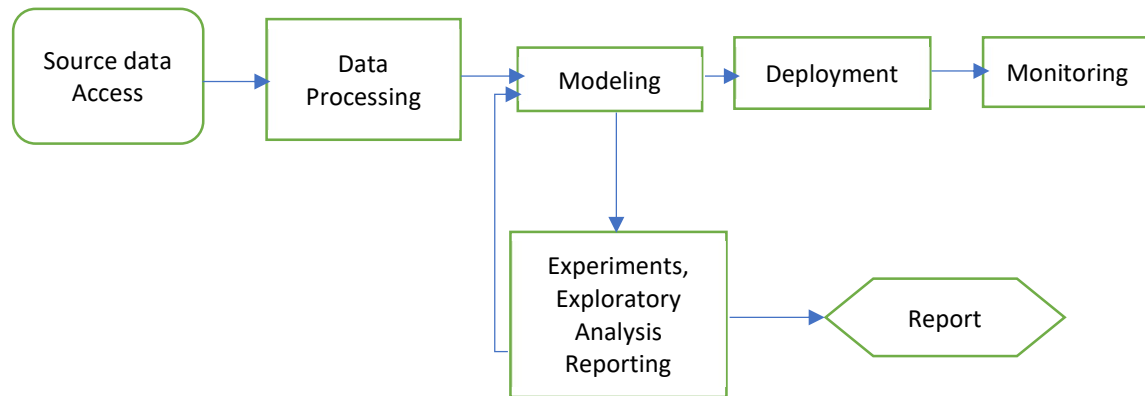


Figure 1: Data Science Workflow

In addition to workflow, data science jobs are a key field of research. Data scientists, data engineers, and data analysts are the three jobs that commonly make up data science. Each position is in charge of managing numerous workflow steps. The phases that "data scientists" and "data analysts" are accountable for, however, overlap. The process steps shown in Figure 1 are all included in the data scientist's work. However, once report satisfaction is attained, the data analysis is finished. Although "data scientists" and "data analysts" refer to two distinct professions, data analysts' work to some degree overlaps with that of data scientists.

C. Team Data Science Process

Utilizing the team data science process (TDSP), which enables the implementation of data science, AI, and ML projects at Microsoft, adopt an agile and iterative approach to your data science projects. By cooperating as a team and using best practices, it demonstrates the data science initiatives agility. Ad hoc and experimental initiatives are eligible for the TDSP. Its five stages are shown in Figure 2 [10].

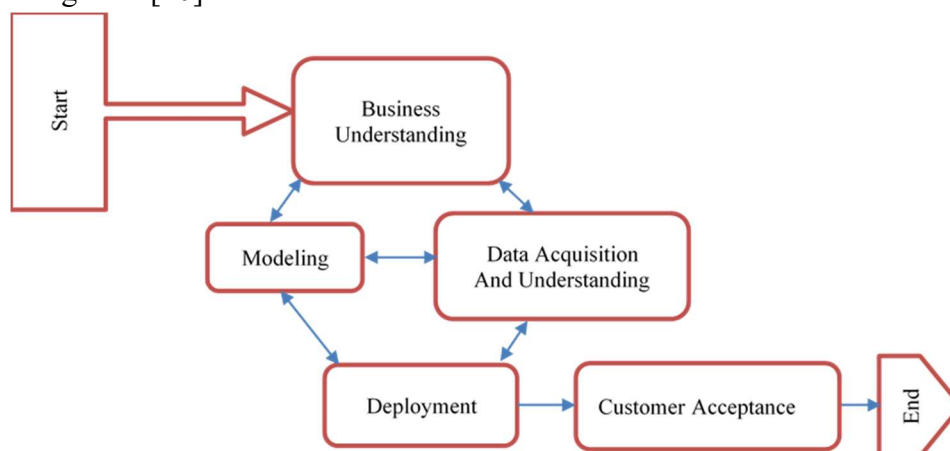


Figure 2: Team Data Science Process

The following are the primary project execution stages as shown in the TDSP [10]:

- Business understanding: Establishing the project objectives, identifying the critical factors that influence the selection of the most appropriate model to fulfill project goals, and pinpointing the related data sources that can be accessed to establish business requirements.
- Data gathering and comprehension: Prioritize creating a clear set of goal variables by defining the relationships between the variables in a clean and crisp dataset. Put the dataset in the appropriate analysis setting.
- Modeling: Choose an ML model that can fit the dataset and anticipate accurate target outcomes while also taking into consideration the model's production performance. To prepare the ML model, identify the best data characteristics.
- Deployment: Turn ML models, together with the data pipeline, into production-level functionality.
- Customer acceptance: To complete the project, satisfy the customer's demands and confirm the model and pipeline they have recommended.

D. Data science vs. big data

Most people mistakenly believe that data science and big data are interchangeable, yet there is a significant distinction between the two. Big data is a large collection of unprocessed digital data that is challenging to manage and analyze using conventional techniques [11]. In other words, it seems that big data may be utilized to analyze insights that can help with improvement and better judgments. While 'data science' is a mix of mathematics, problem-solving, analytics, programming, clever data collection, the capacity to see problems from several angles, data cleansing, planning, and alignment, in a nutshell, it refers to a broad range of techniques for mining big data for knowledge and insight.

III. DATA SCIENCE TRENDS

Many, if not all, areas, including healthcare, manufacturing, business, and academia, have lately been exposed to the trends of data science. In order to offer customers correct traffic data in real-time, Google Maps, as one example of a product made possible by data science, gathers anonymous data from mobile devices [12]. When establishing its recommendation engines, Netflix originally employed data mining methods by asking users to review movies. These ratings were then used to create customized movie suggestions [13]. This demonstrated the effectiveness of leveraging data in recommendation engines. Now that Netflix offers a challenge, the needed project is essentially to create engines that suggest movies by means of past user data, and the award was \$1 million [14]. To succeed in these disciplines, data science, ML, and AI are used. The fields include healthcare, business, education etc.

IV. DATA SCIENCE APPLICATIONS

Data science has considerably helped commercial applications, including ad placement, credit ratings for movies and shopping, recommendations, and stock trading tactics. Other data scientists have applied their knowledge to scientific research via biomedical applications, such as the Facebook algorithm to identify suicidal users, the Microsoft algorithm to identify pancreatic cancer years in advance of its typical diagnosis, and the Google algorithm for recognizing diabetic retinopathy [8]. Following is a list of some of the most significant data science applications like internet search, digital advertisements and recommender systems.

V. DATA SCIENCE CHALLENGES

Big data has recently increased quickly and dramatically due to a variety of factors, including sensors, smart devices, satellite photos, social media, etc., according to [15], who established data storage and analysis as a subfield of data science. Therefore, transmitting and retrieving large amounts of data increases data input and output speeds. Realizing that in certain circumstances, rapid access to data is a top requirement. Scalability and data visualization provide one of the most significant issues since data volume is increasing at a rate that is far faster than CPU processing speed, necessitating the employment of parallel computing approaches to keep up.

Additionally, one of data visualization's main objectives is to plot data more successfully in organizations with millions of customers and reviews, like Amazon, which need adequate skills to show a large and complicated quantity of data rationally [15]. For instance, search engines, social networking platforms, and map apps all need parallel processing power. Data science is a subfield of AI, that includes automated developments that allow computers to do tasks that formerly required human attention in a human-like way. As a result, significant issues may soon arise. However, AI will also greatly raise other challenges, including data privacy, authenticity, and information security, as well as new possibilities for developing, managing, and monitoring systems [15].

VI. CONCLUSION

Data translation into valuable information is the focus of the interdisciplinary topic known as "data science," which is gaining popularity. Basic skills in data mining methods, computational intelligence, predictive analytics, and ML are required in order to get training in this field. Additionally, data science applies approaches from analytics, computer science, applied math, operations research, management science, AI, psychology, and economics. The transdisciplinary character of data science presents difficulties for academic institutions everywhere.

As a result of data science's quick development and significant contributions to several other disciplines, there are countless topics that may be explored and written about. For academics seeking the theoretical basis, trends, and applications of data science and its future, it may be regarded as a very useful reference. What additional qualifications are expected of a data

scientist? As a result, it is crucial now for data scientists to analyze large collections of data, gather useful information, and make the correct choice, which will result in proper accomplishment and success in a variety of disciplines.

REFERENCES

- [1] Khder, M. A., Fujo, S. W., & Sayfi, M. A, “A roadmap to data science: background, future, and trends,” *International Journal of Intelligent Information and Database Systems*, vol. 14, no. 3, pp. 277-293, 2021.
- [2] R. Senthilkumar, B. G. Geetha, "Asymmetric Key Blum-Goldwasser Cryptography for Cloud Services Communication Security," *Journal of Internet Technology*, vol. 21, no. 4 , pp. 929-939, Jul. 2020..
- [3] D. Anusuya, R. Senthilkumar and T. S. Prakash, “Evolutionary feature selection for big data processing using Map reduce and APSO,” *International Journal of Computational Research and Development*, vol. 1, no. 2, pp. 30–35,2017.
- [4] Saltz, J., Shamshurin, I. and Crowston, K, “Comparing data science project management methodologies via a controlled experiment,” *Proceedings of the 50th Hawaii International Conference on System Sciences*, pp.1013–1022, 2017.
- [5] Pedro, F., Márquez, G. and Lev, B, “Introduction to Data Science and Digital Business,” Springer International Publishing, 2019.
- [6] Moscato, P. and Jane, N, “Memetic algorithms for business analytics and data science: a brief survey,” in *Business and Consumer Analytics: New Ideas*, Springer International Publishing, 2019.
- [7] Kenett, R.S., Zonnenshain, A. and Fortuna, G, “A road map for applied data sciences supporting sustainability in advanced manufacturing: the information quality dimensions,” *Procedia Manufacturing*, vol. 21, pp.141–148, 2018.
- [8] Hernán, M.A., Hsu, J. and Healy, B, “A second chance to get causal inference right: a classification of data science tasks,” *Chance*, vol. 32, no. 1, pp.42–49, 2019.
- [9] Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Vera Liao, Q., Dugan, C. and Erickson, T, “How data science workers work with data,” *Conference on Human Factors in Computing Systems – Proceedings*, pp.1–14, 2019.
- [10] Foroughi, F. and Luksch, P, “Data science methodology for cybersecurity projects,” *Computer Science & Information Technology (CS & IT)*, 2018.
- [11] Jan, B., Farman, H., Khan, M., Imran, M., Islam, I.U., Ahmad, A., Ali, S. and Jeon, G, “Deep learning in big data analytics: a comparative study,” *Computers and Electrical Engineering*, September 2018, vol. 75, pp.275–287, 2019.
- [12] Schmidt, E., Rosenberg, J. and Eagle, A, “Google: How Google Works, Grand Central Publishing,” New York, 2014.
- [13] Maddodi, S. and K, K.P, “Netflix bigdata analytics – the emergence of data driven recommendation,” *International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, vol. 3, no. 2, pp.41–51, 2019.

- [14] Donoho, D, “50 years of data science,” Journal of Computational and Graphical Statistics, vol. 26, no. 4, pp.745–766, 2017.
- [15] Acharjya, D.P. and P, K.A, “A survey on big data analytics: challenges, open research issues and tools,” (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 7, no. 2, pp.511–518, 2016.