Grenze International Journal of Engineering and Technology, Jan Issue

A Review Paper on Exploring the Concept of Data Science: A Comprehensive Analysis

Samiksha Gaikwad[1], Parimal Chaudhari[2], Dipali Jadhav[3], Punam Bodade[4] and Dhiraj Shirbhate[5]
[1-4]Student, Department of Computer Engineering, Government College of Engineering, Yavatmal, Maharashtra, India
Email: samikshagaikwad29@gmail.com, parimal9822@gmail.com, dipali.jadhav.184c@gmail.com, punambodade@gmail.com
[5]Assistant Professor, Department of Computer Engineering, Government College of Engineering, Yavatmal, Maharashtra, India
Email: shirbhate.dhiraj@gmail.com

***Abstract*— Data science is a rapidly growing technology in the technical world that fulfills the requirements for data and various data aspects. The core of all emerging technologies is data science, which includes machine learning, artificial intelligence, robotics, edge computing, and blockchain technology. In this review paper, we consider the detailed concept on data science, such as where the data is generated, the skills to handle data, its growth, how it works, and the impact of data science on other technologies. The basic aim of this review paper is to provide a basic summary of data science that everyone easily understands.**

***Index Terms*— Data Science, machine learning, big data, artificial intelligence**

I. INTRODUCTION

In simple words, the term “Data Science” means the science of dealing with various types of data. It is an emerging technology that combines the study of domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data. In the world nowadays, we need a large amount of space to store petabytes and hexabytes of data because data is growing vastly. The number of people use mobile, laptops, etc. It also works in social media, transactions, enterprises, stock marketing, etc.

In 1962, John Tukey described a field he called "data analysis", which resembles modern data science.[9] In the year of 1996, first time the term ‘Data Science’ was introduced in the title of a statistical conference “The International Federation of Classification Societies” (IFCS) known for “Data Science, classification and related methods” [2]. In July 1966, Turing award winner Peter Naur proposed the term datalogy in a letter to the editor of the Communications of the ACM[5]. He defined datalogy as the “science of the nature and use of data” and suggested using this term rather than “computer science”. [5] Since 2012, data has been discussed as a “critical new form of economic currency.

Over the last few years data science has continued evolving and permeated nearly every industry that produces or realizes data.

Data science involves statistical and computational techniques to retrieve specific knowledge from data. It includes a wide range of tasks, including data visualization, statistical modeling, machine learning, and data cleaning and preparation. Data scientists have used these similar techniques to find trends and patterns in data, make predictions,

Grenze ID: 01.GIJET.10.1.150

and support decision-making. These also work with different types of data, such as structured (data stored in relational format) and unstructured (text, images, audio or video). Data science is a multidisciplinary field that involves a combination of skills and knowledge from different sectors such as statistics, computer science, and mathematics. One of the definitions for Data Science was provided by Cao in the form of a formula [6]:
Data Science = (Informatics + Statistics + Communication + Computing + Management + Sociology) | (Environment + Data + Thinking). But now considering today's factors the formula can be modified accordingly as follows:
Data Science = (Statistics and mathematical operations Artificial Intelligence + Machine Learning or Deep Learning) | (Big Data + information + surrounding).
Only, data science obtains meaningful information from structured or unstructured data, followed by extracting the data using various techniques, analysis, programming, and business skills.
Due to the extreme growth of technologies such as Artificial intelligence, Cybersecurity, robotics, Cloud Computing, the Internet of Things, Machine Learning, and Automation in various sectors, the creation of data increases exponentially, which can cause data complexity. Nowadays, data is becoming big data, which affects the enlargement of 9 Vs (Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization, and Value). Handling data is becoming a major challenge. Currently, we are facing challenges such as data privacy and ethics, data quality, in-disciplinary collaboration, and Scalability and Security.
To overcome these challenges, we generally focus on terms such as defining a clear objective, Investing in Scalable Infrastructure, Data Quality and Data Governance, Data Security and Privacy, Data Integration, Data Analysis and Machine Learning, Data Visualization, and Data Visualization. This review paper aims to provide a valuable resource for both beginners and experts in the field that offers insights into the existent state of data science and its impact on various technologies. In addition, we consider the importance of the data science course, which is essential for advancing technology.
This review paper is divided into five sections. Section I gives the introduction of the review paper, and Section II describes the related work regarding the title of the review paper. In Section III we compare Data Science with other emerging technologies and in Section IV we discuss some applications of data Science in various sectors lastly Section V gives the overall conclusion on Data Science.

II. Related Work

A. Key Sources of Data Science

The fundamental requirement of data science is the data itself. In simple words, data is nothing but distinct pieces of information that are essential for various sectors in the world, such as business, industrial, social, and public sectors. the data is the core component of data science that can cause inadequate information.
However, the growth of data is exponentially increasing, and the root cause of this growth is nothing but sources where the data is generated. Here we list some key data sources such as sensor data, customer data, real-time streaming data, public data, scientific data, transnational data, APIs (Application Programming Interfaces), market research data, and historical data. Stakeholders are also responsible for generating data. To understand the concept of data science, it is essential to discuss the life cycle. As shown in Fig. 1, the data includes six stages: identifying and understanding the problem statement, collecting data and defining domain expertise data processing, data analysis and visualization, data modeling, and model development.

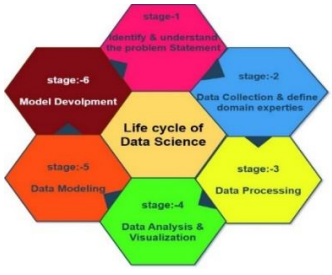

Fig. 1. Life Cycle of Data Science

In the first stage of the life cycle of data science, the problem statement is identified and understood fully, and various solutions are found during this stage.

After the first stage, we collect various data related to the problem and choose the appropriate technology to solve the problem in the second stage. Now, we have a huge amount of data collected at the third stage, and we process the data. Then, we analyze and visualize the data to understand it easily. Visualization is essential in data science because it has the perception capabilities of the human cognitive system that can be exploited by using the right visualizations [1]. After these stages, data modeling is performed, and finally, at six stages, we deploy our model as a solution to the problem statement.

B. Storage: Various Techniques of Storing Data

Data science involves storing and analyzing data by performing various statistical and mathematical operations. Techniques such as relational databases, No-SQL databases, cloud storage, data warehouses, data lakes, file systems, tape storage, and blockchains are commonly used by developers. However, technologies such as relational databases, No-SQL databases, and cloud storage have been mostly used recently. Relational databases store data in structured tables containing rows and columns. It is commonly known as a relational database management system (RDBMS). The most common examples are MySQL, PostgreSQL, and Oracle. No-SQL databases store and handle unstructured and semi-structured data to quickly access data. The most common examples are MongoDB, Redis, and Cassandra. Cloud storage is becoming a major data storage technique. Cloud storage is often used for backup and serving static content for web and Android development. Cloud services such as Google Cloud Storage and Microsoft Azure provide effective storage and easy access to various types of data.

C. Importance

The main use of data science is to extract important data (insight) from large or massive amounts of data using various analysis processes and algorithms.

Data Science acts as a link between the IT and business sectors. Data science opens the door for data scientists to earn more money and provide a huge amount of employment. Data science includes various tools such as Tableau, Rapid Minier, Quik, KNIME, Power BI, and Looker so that by using this data science combines knowledge from various domains. So that it provides the flexibility to the large amount of data. Data science is useful in various fields, but now, with the help of this technology, machines can predict the situation using historical data. If the technology is implemented properly, it can acquire the power to predict the future more accurately and efficiently.

D. Skills to Handle Data

The following diagram shows the skills required to handle various data formats. These skills are essential for dealing with various aspects of data science. Although the diagram shows these skills, such as mathematical and statistics, data analysis, programming, and automation skills, we should consider other skills such as hacking, critical thinking, problem-solving, visualization of data, and communication. These skills are essential to becoming a better data scientist in the technical industries. These skills are essential for the data scientist.

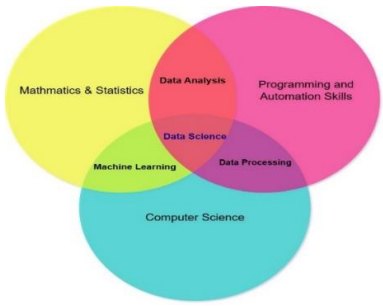

Fig. 2. Relationship Between Data Science and the Skills Required to Handle Data Science

E. Challenges

One of the important challenges in data science is finding a large amount of qualitative data and removing the data's low quality. Challenges for data scientists include communicating effectively with business leaders and managers and identifying business issues. The data scientist and machine learning team spend more time analyzing and preparing a huge amount of real-time data to ensure consistency. Selecting an accurate model is also a challenge in data science because selecting the wrong statistical and ML models can generate poor results. Security is one of the biggest challenges in data science. Security is important for data privacy, data storage, data

transmission, third-party services, and security development. Data science is a rapidly increasing field; therefore, keeping up with new technology and tools is a current challenge.
Some other challenges are mentioned below:

a) Data Volume

Handling and processing large volumes of data, often referred to as "big data," can be challenging as data sources continue to grow continuously.

b) Data Variety

Data can come in various forms, including structured data (e.g., databases), unstructured data (e.g., text and images), and semi-structured data (e.g., JSON).

c) Data Integration

It is a process of combining data from various sources and providing a piece of abstract information to the users.

d) Data Velocity Real-Time

Data processing, required in some applications, adds an extra layer of complexity. Managing real-time data streaming and analysis can be challenging.

e) Ethical Concerns

The ethical use of data is a growing concern. Ensuring that data science projects adhere to ethical standards and principles is an ongoing challenge.

f) Model Selection and Tuning

Choosing the right model for a given problem and optimizing its parameters can be challenging.

g) Feature Engineering

Extracting meaningful features from data can be time-consuming and requires domain knowledge. Effective feature engineering is crucial for model performance.

h) Interpretability

Understanding and explaining model decisions can be challenging, particularly in fields with regulatory requirements.

i) Bias and Fairness

Data and models can perpetuate biases present in the data, leading to unfair outcomes. Ensuring fairness and mitigating bias in data and models is a complex ethical challenge.

j) Keeping Up with Technology

Staying up-to-date with the latest tools, techniques, and algorithms in the data science field can be a continuous challenge.

III. DISCUSSION

A. Data Science and Big Data

Big data is specified to datasets that are too vast or complex to deal with using traditional data processing software. It consists of greater variety, increased volume, and greater velocity. Big data is a related marketing term[10]. Collecting big data is not always useful for achieving better efficiency results; however, it is becoming important to deal with big data. Big data technology is useful for analyzing data more clearly and also it helps to improve the effectiveness of various predicting and analyzing models of data
In this technological world, data is becoming big data because it is generated too fast and it is necessary to handle such data. These rapidly growing types of big data are not properly managed by traditional data software. Additionally, data science is a data-driven technology that plays an important role in finding the hidden patterns in big data, which helps to deal with complex data and organizations to make better decisions. Recently, this technology ecosystem has undergone some radical change due to the advent of Big data techniques [4]. Consideration of big data is becoming an essential part of data science. Now, we can say that big data is the core element of prediction or recommendation systems.

We can say that data science provides direction to big data, which is essential for creating better prediction systems and increases the effectiveness of the program. On the other hand, increasing the growth of big data is helpful for better predictions, which can improve decision power. The overall volume of data recorded by mankind has been growing for decades [3], hence it is necessary to handle big data properly

B. Relationship Between Data Science and Other Technologies

As shown in Fig. 3. Data science is associated with various skills such as statistics, data modeling, data visualization, coding, Math/Probability, and data wrangling. It encompasses a data range in huge amount, including all data-related tasks such as data collection, data visualization, and analysis. Machine learning is associated with various skills such as statistics, data visualization, coding, mathematics, and probability. ML uses techniques such as statistical modeling algorithms and training data to make good predictions. Android is associated with coding skills. Android is an open-source operating system. Android is a mobile operating system that works as a platform for developing mobile applications. Artificial Intelligence (AI) is the combination of science and engineering for making intelligent machines [7]. AI requires a large amount of data to generate solutions, and the skills required to handle data science help uncover the hidden patterns in the data.

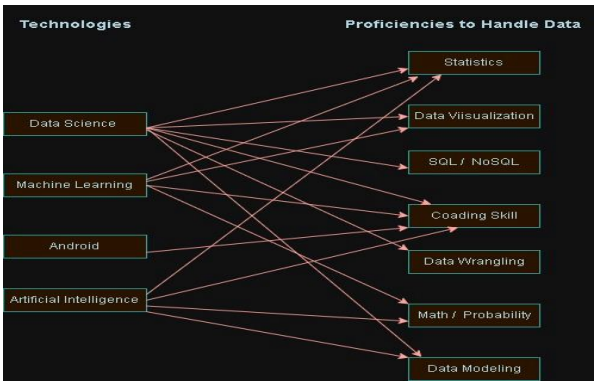

Fig.3 Relationship between Machine Learning, Artificial Intelligence, and Android with Various Skills.

Following table 1. Gives a detailed comparison between Data Science, Machine Learning, Artificial Intelligence, and Android.

TABLE I. COMPARISON OF VARIOUS TECHNOLOGIES WITH DATA SCIENCE

Key Features	Data Science	Machine Learning	Artificial Intelligence	Android
Interpretation	Data Science refers to the science of dealing with various types of data.	ML is a subgroup of AI that is used to predict based on experience data by using various algorithms.	AI is a branch of science that builds computers and various machines that act like human intelligence.	Android is an open-source operating system—designed for various mobile applications.
Data Range	Data Science encompasses a large range of data, including all data-related tasks such as data visualization, collection, and analysis.	ML deals with large amounts of data to develop an algorithm for making predictions or decisions.	AI includes ML but expands further to cover different tasks such as natural language processing, and computer vision.	Android is a mobile operating system that works as a platform for developing mobile applications.
Techniques	Data mining, Data analysis Visualization, Mathematical statistics Techniques.	ML uses techniques such as statistical modeling algorithms and training data to make good predictions.	Intelligent activities such as visual perception, speech recognition, and natural language understanding	It involves Java, Coco, testing, Jetpack, xml layout, and Android API.
Outcome Consists of	Visualization, reports, dashboards, and data-driven insights	Predictive models, Classification, Recommendation.	Natural language understanding, reasoning, and problem-solving.	User interfaces and applications for Android devices
Popular Tools Used for	Software as a service (SAS), Tableau, Apache Spark, MATLAB, Python, R, SQL, and statistical software	Microsoft Azure ML Studio, Google TensorFlow, IBM Watson	PyTorch, Python, and TensorFlow Tableau, GPT3	Java, Kotlin, Android Studio and Android development libraries.

IV. APPLICATIONS

Data science has proven its power and effectiveness through successful applications across various sectors. In this regard, we present a selection of real-life examples and academics that demonstrate the immense impact of data science in different areas such as healthcare, space research programs, finance, marketing, and many more.

A. Healthcare

Data science plays a key role in improving the accuracy of clinical diagnosis. Machine learning models analyze patient data, including medical records and diagnostic images, to predict disease and recommend treatment plans. For example, IBM Watson for Oncology helps oncologists make personalized treatment decisions for cancer patients.

In epidemiology, data science helps identify disease outbreaks and predict their spread. Global Public Health Information Network (GPHIN) technical art data and global monitor detection of news sources for maturity potential health threats. We can consider the recent example of a pandemic, i.e., COVID-19, which was first detected in China at the end of December 2019 and later spread over all the continents of the earth except Antarctica. Data science, along with statistical analysis, computational science, and biology computational, applications from epidemiology drive drug discovery and molecular design for therapeutic purposes. Various data-driven models, such as numerical mathematical models and correlation and predictive models, have been developed to study COVID-19 cases.

B. Space Research Program

In the space research program, data science plays a major role. The success rate of any space research is directly proportional to the proper implementation of data science. In space research programs, a huge amount of data is generated by various sources, including inbuilt sensors, rovers, satellites, and space telescopes. These generated data are properly analyzed by the data scientist and used for making a predictive model that helps to predict the situation according to space. Predictive modeling also helps reduce the risk of failure in space research missions.

C. Marketing

E-commerce businesses and marketers are leveraging data knowledge to segment their customer base and provide personalized recommendations. For example, Amazon's recommendation system, Zomato recommendation. It uses collaborative filtering and machine learning to suggest items in users' browsing and purchase history.

Marketers use A/B testing to optimize website design, email campaigns, and advertising strategies. Data-driven research from A/B testing to support refined marketing efforts, resulting in higher conversion rates and return on investment (ROI).

For example, a data analyst might analyze sales data to identify trends in customer behavior and make recommendations for marketing strategies.[8]

Sectors such as finance, business, research programs, and government organizations use the concept of data science to increase its efficiency.

V. CONCLUSION

Data science is a dynamic technology that generates new concepts and helps in researching new technologies. With the intense growth of data, learning about data science is essential. As we discussed in this review paper, this growth of data itself brings major challenges such as securing data, data loss, and unnecessary data. The management of data is important to handle such challenges, and we must be aware of the data science concept.

In this review paper, we discussed the essential aspects and trends in the field of data science and highlighted key challenges and relevant concerns in the context of its future. First and foremost, there has been rapid growth in the field of data science, where the power of data analytics, discovery, and problem-solving spans across various domains. The main focus of our review is the critical importance of data science, its quality, and the extraction of past data in any successful data science project.

REFERENCES

[1] J. Wijk, "The value of visualization," in Proc. IEEE Vis, 2005, pp. 79–86.
[2] Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery.Microsoft Research

[3] M. Hilbert and P. Lopez, “The world’s technological capacity to store, communicate, and compute information,” Science, vol. 332, no: 6025, pp. 60–65, 2011.
[4] O. Terzo, P. Ruiu, E. Bucci, and F. Xhafa, “Data as a service (DAAS) for sharing and processing of large data collections in the cloud,” in Proc. 7th Int. Conf. Complex, Intell., Softw. Intensive Syst., Jul. 2013, pp. 475–480.
[5] van der Aalst, W., & Damiani, E. (2015). Processes Meet Big Data: Connecting Data Science with Process Science.
[6] Cao, L. (2017): Data science: a comprehensive overview, ACM Computing Survey.
[7] Cao, L. (2017): Data science: a comprehensive overview, ACM Computing Survey.
[8] Gareth, Hastie; Witten, Tibshira (29 September 2017). "An Introduction to Statistical Learning: with Applications in R." Springer.
[9] Donoho, David (18 September 2015). "50 Years of Data Science" (PDF). Retrieved 2 April 2020.
[10] Pham, Peter. "The Impacts of Big Data That You May Not Have Heard Of". Forbes. Retrieved 3 April 2020