# Data Science and Its Importance

Chethiya Galkaduwa[1] and Nethmini Ranasinghe[2]

[1]University of Kelaniya
[2]La Trobe University

February 16, 2024

**Abstract**

Data science is an interdisciplinary subject that uses scientific methodologies, data mining techniques, machine-learning algorithms, and large amounts of data to extract information and insights. The article presents an overview of the current state and future possibilities of Data Science in a variety of sectors, explores the benefits, outlines the frameworks and methodologies employed, explains the current obstacles, and offers feasible solutions.

# Data Science and Its Importance

Chethiya Galkaduwa
*University of Kelaniya, Sri Lanka*
*chethiya.galkaduwa@gmail.com*

Nethmini Ranasinghe
*La Trobe University, Australia*
*nethuranasinghe99@gmail.com*

## Abstract

Data science is an interdisciplinary subject that uses scientific methodologies, data mining techniques, machine-learning algorithms, and large amounts of data to extract information and insights. The article presents an overview of the current state and future possibilities of Data Science in a variety of sectors, explores the benefits, outlines the frameworks and methodologies employed, explains the current obstacles, and offers feasible solutions.

**Keywords :** Data Science, Big data, modelling

---

## Introduction

We are in an age of "data science and advanced analytics" in which almost every element of our daily lives is recorded digitally as data [2]. Thus, the present electronic world is a rich mine of many data kinds, including business data, financial data, healthcare data, multimedia data, internet of things (IoT) data, cybersecurity data, and social media data, among others [2]. Every day, the quantity of data that may be structured, semi-structured, or unstructured increases. Data science is often a "concept that unites statistics, data analysis, and their related methods" in order to grasp and assess events via the use of data. [1] Almost every sector or organization is impacted by data; consequently, "Data Science," which includes advanced analytics with machine learning modelling, may be used to business, marketing, finance, IoT systems, information security, urban management, medical services, and policy choices. Data science includes several fields, including statistics, scientific methodologies, and artificial intelligence (AI), and data analysis, to extract value from data. The practitioners of data science are known as data scientists, and they use a range of skills to extract useful information from data collected from the Internet, mobile devices, users, sensors, and other sources. Data science entails the purification, aggregation, and modification of data in order to undertake complex data analysis. The results may then be examined by analytic programs enable corporate leaders to draw informed conclusions using Data science. Data science requires a variety of disciplines and domains of expertise in order to conduct a thorough, extensive, and refined analysis of raw data. Data scientists must be well-versed in data engineering, mathematics, statistics, advanced computing, and data visualization in order to effectively sift through tangled masses of data and communicate just the most critical information in order to boost productivity and creativity. This article gives an overview of the benefits and approaches of data science in several industries.

**Survey**

Researchers in the area of data science use several popular datasets for a number of purposes. These include cybersecurity datasets such as NSL-KDD, UNSW-NB15, Bot-IoT [3]ISCX'12, CIC-DDoS2019, etc., smartphone datasets such as phone call logs, mobile application usages logs, SMS Log, mobile phone notification logs etc., IoT data, health data such as heart disease, diabetes mellitus, COVID-19 etc., agriculture and e-commerce data, and numerous others. In the section under "Different real-world applications." In the real world, the data employed by data-driven applications might be any of the aforementioned types, and they can differ from application to application. In a business setting, data science comprises developing a full understanding of the problem to be addressed, how it impacts the relevant organization or individuals, the ultimate goals for addressing it, and the related project plan. In collaboration with end-users and other stakeholders, data scientists create appropriate queries to identify and grasp business difficulties. This helps acquire a better knowledge of business needs and what information should be extracted from data. Business intelligence [4] refers to the business knowledge that helps companies to make better decisions. Identifying the relevant data sources that may help clarify the formulated questions and choosing what actions should be taken based on the patterns that the data exposes are also essential components of this process. Understanding data is especially important in the business sector, since data science relies largely on data availability [4]. For a data-driven model or system, a comprehensive understanding of the data is necessary. Real-world data sets often include noise, missing figures, conflicts, and other data issues that must be addressed properly. The gathering and purification of relevant data, as well as the data's quality, are crucial to any data science project if useful insights are to be gained. The data science modelling may be used to make modifications and improvements to business practices. Entertaining aspects of the data science process include obtaining a deeper understanding of the business problem to be addressed. Without this, it would be far more challenging to gather accurate data and extract the most useful information for problem-solving decisions. "Data Scientists" often assess and manage data to uncover the answers to crucial questions that aid businesses in making objective decisions and addressing complex problems. In conclusion, a data scientist proactively gathers and analyses data from multiple sources to gain a better understanding of how the business continues to operate, and designs deep learning or data-driven tools/methods, or methodologies, cantered on advanced analytics, which can make the current computing process more smart and intelligent. The term "business data science" is often used to refer to the study of business or e-commerce data in order to get insights into a firm that can then be used to make good decisions and carry them out effectively [5]. In order to predict consumer behaviour, spot patterns and trends in historical company data, and provide recommendations for better decision-making, data scientists may create algorithms or data-driven models. Advanced analytics methods and machine learning models derived from the accumulated data hold the key to company automation, intelligence, and efficiency, which may be attained by following the previously mentioned data science approach. Using predictive modelling with machine learning techniques, online retailers like Amazon may better manage inventories, avoid stock-outs, and streamline logistics. High-stakes business decisions, such as those pertaining to risk management, fraud prevention, credit distribution, customer analytics, and

customized services, electronic trading, and so on rely heavily on historical data, which is collected and used by banking firms. Future applications of data science techniques in business and finance may include automation, cognition, and intelligent decision-making and processes. Several industrial revolutions have occurred in the manufacturing sector in order to keep up with global competition in terms of production capacity, quality, and cost [5]. The term "Industry 4.0," or "the fourth industrial revolution," refers to the current movement toward more production automation and data exchange. This suggests a potential role for industrial data science, defined as "the study of industrial data to acquire insights that typically result in better industrial applications." Modelling from the field of data science may be used to boost efficiency, cut down on expenses, and maximize earnings for manufacturing companies.

Healthcare is one of the most recognized industries in which data science is making substantial progress. Health data science is the extrapolation of practical insights from patient data collected via electronic health records. analysing data from several sources, including as the electronic health record, billing claims, cost predictions, and patient customer surveys, may aid companies in boosting the quality of treatment, decreasing the cost of care, and enhancing the patient experience. In reality, healthcare analytics using machine learning models may save medical costs, predict infectious outbreaks, prevent preventable diseases, and improve life quality [6]. Using machine-learning algorithms, the massive amount of data may be consolidated and properly examined. Analysing the particulars and recognizing the trends in the data may aid in better decision-making, leading to improved patient care. It may assist in comprehending the patterns to improve the result of medical care, life expectancy, early detection, illness diagnosis at an early stage, and inexpensive treatment [10]. In the future decade, data mining tools may facilitate a change from traditional medical databases to a knowledge-rich, evidence-based healthcare environment. With the advent of social media (platforms like Facebook and Twitter) and smartphone applications that can monitor personal health metrics using sensors and analyzers [11], the importance of big data and its usefulness in healthcare and the medical sciences has increased. Data mining's purpose is to enhance the stored user information in order to deliver better treatment and care. This article gives an overview of the benefits and approaches of using big data in health care systems. It emphasizes the vast amounts of data created by these systems, their characteristics, potential security-related issues, data management, and how these analytics facilitate getting considerable insight into big data sets [12]. Some uses of data science in the healthcare industry include disease monitoring, patient data management, image processing of healthcare data from the perspective of big data, data from wearable technology, medical signal analytics, and data storage and cloud computing. Data storage is applied for research, instruction, education, and quality assurance. Using keywords in accordance with the specified patient privacy policy, users may also extract files from a repository holding radiological findings [12]. Various signal-processing methods may be employed to obtain a large number of target features, which are then used by a pre-trained machine-learning model to deliver actionable insight. These remarks may be analytical, prescriptive, or prognostic. These insights may also be used to trigger further strategies, such as alerts and doctor notifications. Disease surveillance includes perception of the illness, comprehension of its state, etiology (the method in which a disease is caused), and prevention. Information gathered through EHRs and the Internet offers enormous potential for disease

analysis. The different techniques of monitoring would help in the planning of services, assessment of treatments, determination of priorities, and development of health policy and practice. Image processing on healthcare data provides vital insights into anatomy and organ function, as well as diagnoses illness and patient health situations. The application of artificial intelligence in image processing will improve areas of health care, such as screening, diagnosis, and prognosis, and the integration of medical pictures with other forms of data and genetic data will boost diagnostic precision and simplify early illness detection [13]. The exponential rise in the number of medical facilities and patients has resulted in improved use of computer-based healthcare diagnostics and decision-making systems in clinical settings. In Patient data management. With the growth of data science and the introduction of several data-driven applications, the health sector continues to be a prominent supplier of data-driven solutions for a better life and customized services for its consumers. Data scientists may acquire valuable insights into enhancing the efficiency of pharmaceutical and medical services by analysing their vast array of data on the healthcare industry, which includes financial, clinical, R&D, administration, and operational information [12].

"The Internet of things" (IoT) is a new technical sector that converts every electronic equipment into a smarter one, and is thus considered as the great frontier that has the potential to enhance almost every aspect of our lives. Machine learning has become an essential technique for IoT applications because it use information to spot patterns and create models that predict future behaviour and occurrences. Smart cities, which use technology to improve municipal services and the quality of life for their citizens, are one of the principal uses of the Internet of Things. Using proper data, for instance, data science methods could be employed to predict traffic in smart cities and estimate the total energy consumption of residents over a certain period of time. On the basis of large-scale IoT datasets, [7] data science models based on deep learning may be constructed. Data science and analytics approaches may enable the modelling of a variety of IoT and smart city services, including smart governance, home automation, learning, communication, transportation, business, farming, health care, and industry, among others.

One of the most crucial areas of study in Industry 4.0 is cybersecurity, or the practice of protecting computer networks, systems, hardware, and data against cyberattacks. Data science techniques, and particularly machine learning, have emerged as an essential cybersecurity tool, with the ability to continually understand to identify trends by analysing data, better detect malware in encrypted traffic, find threats, predict where bad neighbourhoods online are, keep people safe while surfing, or protect data in the cloud by revealing questionable user activity [8].

Information gleaned from a wide range of Internet-connected devices, including personal computers, tablets, and smartphones, is known as behavioural data [2]. Websites, mobile applications, marketing automation systems, contact centres, support desks, billing systems, and so forth are all common sources of data. Behavioural data is not just data since it is not a static set of numbers [9]. In e-commerce and brick-and-mortar retail, foreseeing future market trends and best deals; able to forecast usage pattern, load, and usage patterns in upcoming updates; determining how users utilize an application to accurately predict usage and priorities

in application development; and sectioning consumers into similar organizations to gain a more in-depth understanding of the market are all areas where advanced analytics of these data can be helpful. "next-generation, multifunctional cell phones that facilitate data processing and enhanced wireless connection" [9] is how modern smart mobile phones are described. Users have shown a greater preference for "Mobile Phones" in recent years compared to "Desktop Computer," "Portable Computer," and "Tablet Computer." Email, IM, online shopping, Internet surfing, entertainment, social networking like Facebook, Linkedin, and Twitter, and Internet of Things (IoT) applications including smart cities, health, and transportation services are just a few of the many uses for smartphones [9]. App characteristics including action orientation, adaptability, suggestion and choice orientation, data-driven, context-aware, and cross-platform operation are the foundation on which intelligent apps are constructed [9]. Therefore, mobile data science, the process of gathering a vast amount of mobile data from various sources and evaluating it using machine learning methods to find key ideas or data-driven trends, may play a crucial role in the development of intelligent smartphone applications.

There are various challenges in implementing data science. As discussed briefly in "Understanding data science modelling," the first problem in data science modelling is to grasp real-world business difficulties and their associated data, such as their forms, kind, size, labels, etc. According to the requirements, this is to identify, describe, depict, and evaluate domain-specific business issues and information. A well-defined procedure is necessary for the success of a data-driven business solution prior to commencing actual data analysis. Additionally, it is difficult to acquire business data since data sources may be many and unreliable [1]. Understanding and selecting the right analytical approaches to extract meaningful insights for intelligent decision-making in a given business context is the greatest challenge in the area of data science. Advanced analytics focuses on forecasting the future via the use of data to identify patterns and anticipate what is most likely to occur. Advanced analytics give a better understanding of data and permit thorough data analysis, while basic analytics provide a broad description of data. Understanding complex analytics approaches, especially machine learning and deep learning-based modelling, is crucial [1]. The next step is to extract trustworthy and useful information from the previously gathered data. Data scientists' main purpose is to uncover, explain, represent, and gather data-driven knowledge in order to draw meaningful insights from data. However, real data may include a number of confusing, missing, outlier, and meaningless numbers [101]. The advanced analytics techniques mentioned in "Advanced analytics methods and smart computing," such as machine and deep learning models, have a substantial impact on the information's quality and availability. Consequently, it is necessary to comprehend real-world business scenarios and associated elements in terms of whether, what, and why they are inadequate, missing, or problematic, and then to extend or redevelop conventional systems, such as large-scale hypothesis testing, learning discrepancy, and ambiguity, etc.

## Conclusion

The survey paper shows that data science plays a major role in various sectors to make effective decisions. It is available to almost all industries. Today, there is a vast quantity of data accessible, and brands and organizations' success or failure hinges on their ability to use this data effectively. Utilizing data effectively will be the key to attaining brand objectives in the next years. Consequently, it is playing an increasingly important and central role in the operation and development of brands. Being a data scientist is a prestigious profession, since they are responsible for managing data and giving answers to their challenges, both within and outside of the organization. In terms of experimentation and research, data scientists nowadays are breaking new ground. They are experimenting with intelligence-gathering technology and constructing complex models and algorithms to assist companies in addressing some of their greatest difficulties.

## Acknowledgment

I would like to express my gratitude to all of the participants who contributed to this survey. Their valuable input and insights have greatly enhanced the content of this paper. I would also like to thank Dr. (Ms.) Devindri Perera of University of Kelaniya, Sri Lanka for her support in conducting the survey. Without the assistance, this research would not have been possible.

## References


[1]
I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *SN Computer Science*, vol. 2, no. 5, Jul. 2021, doi: 10.1007/s42979-021-00765-8.

[2]
I. H. Sarker, M. M. Hoque, Md. K. Uddin, and T. Alsanoosy, "Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modeling and Research Directions," *Mobile Networks and Applications*, Sep. 2020, doi: 10.1007/s11036-020-01650-z.

[3]
N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, Nov. 2019, doi: 10.1016/j.future.2019.05.041.

[4]
D. Larson and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science," *International Journal of Information Management*, vol. 36, no. 5, pp. 700–710, Oct. 2016, doi: 10.1016/j.ijinfomgt.2016.04.013.

[5]
J. Li, "Big Research Data and Data Science," *Data Science Journal*, vol. 14, May 2015, doi: 10.5334/dsj-2015-001.

[6]
M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," *Computers & Chemical Engineering*, vol. 106, pp. 212–223, Nov. 2017, doi: 10.1016/j.compchemeng.2017.06.011.

[7]
I. H. Sarker, "Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective," *SN Computer Science*, vol. 2, no. 3, Mar. 2021, doi: 10.1007/s42979-021-00535-6.

[8]
I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, no. 1, Jul. 2020, doi: 10.1186/s40537-020-00318-5.

[9]
I. H. Sarker, A. Colman, and J. Han, "RecencyMiner: mining recency-based personalized behavior from contextual smartphone data," *Journal of Big Data*, vol. 6, no. 1, Jun. 2019, doi: 10.1186/s40537-019-0211-6.

[10]
Y. Ren, R. Werner, N. Pazzi, and A. Boukerche, "Monitoring patients via a secure and mobile healthcare system," *IEEE Wireless Communications*, vol. 17, no. 1, pp. 59–65, Feb. 2010, doi: 10.1109/mwc.2010.5416351.

[11]

W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Health Information Science and Systems*, vol. 2, no. 1, Feb. 2014, doi: 10.1186/2047-2501-2-3.

[12]
S. V. G. Subrahmanya *et al.*, “The role of data science in healthcare advancements: applications, benefits, and future prospects,” *Irish Journal of Medical Science (1971 -)*, vol. 191, no. 4, Aug. 2021, doi: 10.1007/s11845-021-02730-z.

[13]
H. Svanström, T. Callréus, and A. Hviid, “Temporal Data Mining for Adverse Events Following Immunization in Nationwide Danish Healthcare Databases,” *Drug Safety*, vol. 33, no. 11, pp. 1015–1025, Nov. 2010, doi: 10.2165/11537630-000000000-00000.