STUDY OF DIFFERENT DATA SCIENCE APPROACHES

Abstract

For the purpose of developing new algorithms, improving data models, and producing sophisticated analytics, data science has made large research investments. However, authors have not frequently addressed the organisational and socio-technical difficulties that arise when carrying out a data science project. These difficulties include the absence of a defined vision and goals, the overemphasis on technical issues, the inadequacy of ad hoc projects, and the uncertainty of responsibilities in data science. There haven't been many methods proposed in the literature to deal with this kind of issue; some of them go as far back as the middle of the 1990s, so they aren't up to speed with the most recent developments in big data and machine learning technology. However, fewer approaches offer a complete framework. We'll discuss the necessity to develop a more thorough technique for working on data science projects in this piece. We first research ways that have been written about in the literature and group them into four categories based on their focuses: project, team, data, and information management. Last but not least, we offer a conceptual framework that describes the essential characteristics that a methodology for managing data science activities from a broad viewpoint should have. This framework could serve as a guide for other academics as they develop new data science methods or update existing ones.

Keywords: Big data, data science methodology, organisational implications, knowledge management.

Authors

Dr. Ajay Lala
Professor,
Department of CSE,
Gyan Ganga College of technology,
Jabalpur.
ajaylala@ggits.org

Pinkal Jain
Assistant Professor,
Department of CSE,
Gyan Ganga College of technology,
Jabalpur.
pinkaljain@ggct.co.in

I. INTRODUCTION

Recent years have seen an increase in interest in the study of data science, leading to considerable research investments in the creation of novel algorithms, improved data models, and advanced analytics. Recent strides in the field are a reliable sign of this endeavour [1]. However, practical data science initiatives may not always make use of these most recent technological advancements. A New Vantage study and VentureBeat [2] report that 87% of data science projects never reach production. Adopting big data and artificial intelligence (AI) efforts continues to be a major obstacle for 77% of businesses. Additionally, [4] stated that by 2022, 80% of analytics insights will not result in profitable outcomes. 80% of projects involving data science will "remain alchemy, run by wizards" through 2020. Given the competitive advantage that such cutting-edge methodologies give academics and practitioners, it is quite astonishing to find such low success rates in data science programmes.

According to Leo Breiman [5], there are two cultures in statistical modelling: one that is more concerned with creating efficient algorithms to produce accurate predictive models to predict the future, and the other that is more interested in comprehending the physical world and its underlying mechanisms. According to the latter perspective, the scientific method is extremely important, as are theory, experience, subject-matter expertise, and causation. Science philosophers from the 20th century engaged in debates over how knowledge and science advance [6] such as Karl Popper, Thomas Kuhn, Imre Lakatos, and Paul Feyerabend. Some authors (references 7 and 8) have discussed the challenges related to organizational and socio-technical aspects encountered during the execution of data science projects. These challenges encompass various issues such as a dearth of vision, strategic direction, and well-defined objectives; an undue emphasis on technical aspects; a deficiency in ensuring reproducibility; and the ambiguity surrounding roles and responsibilities. These difficulties often lead to data science projects being managed in a haphazard manner and remaining in a nascent stage of development.

While the community hasn't shown an overwhelming level of attention to these issues and there hasn't been a sufficient amount of literature dedicated to tackling them, it's essential to acknowledge that these challenges do indeed manifest in real-world data science projects. As Section 4 will show, some writers have suggested project management approaches for data science projects and have developed new tools and procedures to address the aforementioned problems.

Although the suggested solutions are paving the way to solve these issues, it is the case that data science projects do not use these methodologies. A survey [9] of professionals from both the private sector and not-for-profit organisations in 2018 found that while 85% of the respondents believed that using a better and more consistent process would result in more successful data science projects, 82% of the respondents did not follow an explicit process methodology.

According to a poll conducted by Nuggets in 2014 [10], CRISP-DM was the primary methodology employed by 43% of respondents. Since the first Nuggets poll in 2002 to the most recent one in 2014, this methodology has continuously been the one most frequently

employed for analytics, data mining, and data science projects [11]. Despite being widely used, CRISP-DM was developed in the middle of the 1990s and hasn't been updated since.

Consequently, the purpose of this study is to critically review the approaches used to manage data science projects, classify them based on their focus, and assess how well they are able to deal with the issues that are now present. As a result of our research, we offer a conceptual framework that includes potential characteristics of a methodology for overseeing data science initiatives from a comprehensive standpoint. Other researchers can use this plan as a guide to develop already popular approaches or create brand-new ones.

The remainder of the paper is organised as follows: In order to avoid any semantic misunderstandings and to explain what a data science project is about, definitions of the terms "data science" and "big data" are offered in Section 2, which also contextualises the issue. The organisational and socio-technical challenges that arise when carrying out a data science project are also discussed in Section 2. Section 3 introduces the research topics and provides a description of the research methods employed in the work. A critical analysis of data science project approaches is presented in Section 4. Finally, Section 6 offers the guidelines for furthering this research in the future and the key conclusions of the work. Section 5 analyses the data acquired.

II. THEORETICAL FRAMEWORK

2.1 Background

To avoid any conceptual misconceptions, we start by giving the term "data science" a correct definition. Its description will help to clarify the distinctive qualities of data science projects and serve as the foundation for the suggested management method in this study.

1. Datascience

The fields that feed and develop the data science tree are generally acknowledged by authors who construct data science from well-established branches of research. In opposition to the perspective presented in reference [13], which asserts that data science involves a blend of mathematical proficiency, business insight, and hacking abilities, reference [12] defines data science as an amalgamation of computer science, business engineering, statistics, data mining, machine learning, operations research, six sigma, automation, and domain-specific knowledge. Data science requires a mix of skills, including traditional computer science, arithmetic, and art, according to [14]. [15] shows a Venn diagram in which the combination of a) hacking abilities, b) math and statistics know-how, and c) substantive expertise is represented as data science. Contrarily, according to the authors of [16], many problems in data science are statistical engineering problems with larger, more complex data that may require distributed computing and machine learning approaches in addition to statistical modelling.

Although understanding this aim is crucial to comprehending the role of data science in business and industry as well as its possible domain applications, the basic goal of data science is rarely highlighted by writers. In contrast to [17], which asserts that data science uses "statistical and machine learning techniques on big multi-structured data in a distributed computing environment to identify correlations and causal relation- ships, classify and predict

events, identify patterns and anomaly lies, and infer probabilities, interest, and sentiment," data science is defined by the authors of [9] as the analysis of data to find solutions and gain new insights." According to them, data science combines expertise in statistics, data management, and software development. Data science, according to [18], is the study of computing principles, methodologies, and structures.

To better comprehend its place among traditional employment rolls, there has also been an increase in interest in describing the work done by data scientists and outlining the abilities required to become one. While reference [20] portrays a "unicorn" view of data scientists, suggesting that they are responsible for all aspects of the data process, from data discovery to large-scale processing, visualization, and storytelling, reference [19] characterizes a data scientist as an individual who excels in statistics to a greater extent than most software engineers and possesses superior software engineering skills compared to most statisticians.

We use the comments mentioned above as a point of reference and offer our own definition of data science among the thousands of various interpretations that may be found in order to frame the remainder of the article: Data science is an interdisciplinary topic that straddles the lines of computer science, mathematics, and statistics. It entails the application of scientific procedures and methods to derive knowledge and value from vast amounts of structured and/or unstructured data.

So, we infer from this description that the goal of data science projects is to use data-driven methodologies to solve complicated real-world problems. This means that practically every existing industry area and domain can benefit from data science: e-commerce (targeted advertising [33], product recommendation [34], sentiment analysis [35]), manufacturing optimization (failure prediction [30], maintenance scheduling [31], anomaly detection [32], health-care (medical image analysis [27], drug discovery [28], bio-informatics [29],), banking (fraud detection [21], credit risk modelling [22], customer lifetime value [23]), finance (customer segmentation [24], risk analysis [25], algorithmic trading [26]), Despite the fact that data science is an area that may be used to any field, We think it is essential to have knowledge of the application area in order to properly extract value from data.

2. Big Data Technologies

Looking back at the development of data science over the past ten years, it is clear that this field's explosive growth is directly related to our ability to gather, store, and analyse data that is produced more often [39]. The paradigm of data science and big data actually changed in the middle of the 2000s as a result of several fundamental changes that occurred in each of these stages (collection, storage, and analysis).

In terms of collection, the development of reliable and cost-effective interconnected sensors, as well as the inclusion of industrial gear and built-in sensors in smart phones, has fundamentally altered how statistical analysis is conducted. In fact, historically the cost of acquisition was so expensive that statisticians meticulously gathered data to be necessary and adequate to answer a certain issue. The volume of machine-generated data has increased dramatically as a result of this significant shift in data collection. In terms of storage, we must emphasise the advancement of fresh methods for distributing data among cluster nodes as

well as the advancement of distributed compute to process data on those cluster nodes concurrently. New technologies that helped with advancements include the Hadoop and Apache Spark ecosystems. In methods of collecting and storage. In addition, the development of new algorithms and methods for data analysis has greatly benefited from the rise in computing capacity, both in CPUs and GPUs. Deep learning approaches have advanced due in part to the most recent advancements in GPU technology, which are particularly eager for quick matrix operations [40].

Alongside the progress made in data collection, storage, and analysis, a substantial community of developers and researchers from prominent businesses and academic institutions has played a pivotal role in the advancement of data science. Big data, a subset of data science, concentrates on the distribution and parallel processing of data, typically characterized by the "5 V's" (volume, velocity, variety, variability, and value). Over recent years, the data science community has predominantly emphasized achieving excellence and dedicating substantial research efforts to the development of advanced analytics, primarily aimed at addressing technological issues while often overlooking the organizational and socio-technical challenges. The primary issues experienced by data science experts throughout actual business and industry projects are outlined in the following section.

2.2 Current Challenges

Beyond the analytical ones, there are additional problems associated with leveraging data science within a commercial organisational context. The research stated at the beginning of this paper only serve to highlight the current challenges in carrying out data science and big data projects. Below, we've compiled some of the major difficulties and problems that arise during a data science project, both technically and organizationally.

Coordination, collaboration and communication

Data scientists used to operate alone as "lone wolves," but that is changing as the industry develops to include teams with specific skills. Coordination, which is described as "the management of dependencies across task activities," is cited as the main issue for data science projects by [41, 9, 42], who view them as complicated team initiatives. Processes that are poorly coordinated lead to misunderstandings, inefficiencies, and mistakes. Additionally, this lack of effective coordination affects both data analytics teams and the entire business [43].

In addition to a lack of coordination, [44, 45, 46] point out clear problems with collaboration, and [47, 48] emphasise a lack of open communication between the three primary stakeholders: the company (client), the analytics team, and the IT department. For instance, [44] talks about how challenging it is for analytics teams to deploy to production, cooperate with the IT department, and communicate with business partners about data science. [44] also highlights the absence of business assistance in the sense that there is insufficient business input or domain expertise knowledge to provide positive outcomes. Overall, it appears that the data analytics team and data scientists are having difficulty collaborating effectively with the IT division and the business agents.

Moreover, [48] identifies inefficient governance frameworks for data analytics, and [43] emphasises insufficient management and a lack of senior management endorsement. Working in a disorganised, unclear environment can be frustrating and can make it harder for team members to stay motivated and concentrate on the project's goals, according to [49] who state this in this context.

Building data analytic steams

In other words, [50] highlights issues in assembling the best team for the project, while [45, 46, 43, 48, 51] emphasises the dearth of individuals with analytic expertise. Every major institution has introduced new big data, analytics, or data science degrees as a result of these workforce shortages in the specialised analytical field [42]. In this regard, [46] promotes the necessity for a multidisciplinary team because data science projects require management, technological, business, and data science expertise. For instance, [9] claims that because of process immaturity and the lack of a solid team-based methodology, data science teams have a heavy dependence on the senior data scientist.

Defining the data science project

Data science initiatives frequently include substantial back-and-forth between team members and trial-and-error to find the best analysis tools, programmes, and parameters. They also frequently involve highly uncertain inputs and outcomes [52].

Setting appropriate expectations [17], creating realistic project timeframes, and determining how long projects would take to complete [8] are all made difficult by the experimental character of these projects. In this regard, [50, 53] point out that it might be challenging to grasp the project's scope ex ante and that it can be challenging to comprehend the business objectives.

More specifically, the writers in [47, 43, 48] draw attention to the incorrect project scope, lack of clear business objectives, and insufficient ROI or business cases. For [54], there is an unfair focus on technological issues, which has prevented firms from maximising the promise of data analytics. Instead of concentrating on the business issue, data scientists have frequently been fixated on attaining cutting-edge outcomes on benchmarking activities. Yet, this obsession with achieving the best results can actually build models too complex to be of any value. This way of thinking is practical for data science competitions like Kaggle [55], but not for the business world. Although Kaggle contests are excellent for teaching machine learning, Yet, they may incorrectly anticipate what should be demanded in actual commercial situations [56].

Stake holdersvs Analytics

In addition, the project proposal is frequently poorly specified [44] and the business side is not sufficiently involved. The business side may only offer the data and a minimal amount of domain knowledge, thinking that the data analytics team will complete the remaining "magic" on its own. The false idea that these new technologies can accomplish everything the business advises at a very low cost has been caused by the high expectations that machine learning and deep learning techniques have raised [57]. The lack of participation

from the business side may also be a result of a lack of communication between the two sides: data scientists may not be familiar with the subject matter of the data, and the company is typically not knowledgeable about data analysis methods. In order to bridge the communication gap between these two parties and close the data science gap, it may be essential to have an intermediate who is conversant in both the language of data analytics and the application area [58].

A lack of uniform methodologies and processes to tackle the subject of data science [45] and a low level of process maturity may be the causes of the highlighted project management challenges [52]. Moreover, the repercussions of such poor adoption of procedures and techniques may result in "scope creep" [41, 9] and delivering the "wrong thing" [41, 9]. In fact, teams are more likely to deliver something that does not meet stakeholder needs when there are ineffective systems in place for engaging with stakeholders. The lack of influence and customer or company utilisation of the project outputs is the most blatant illustration of this issue [44].

Driving with Data: The primary uniqueness of a data science project is defined by the utilisation of a data-driven strategy. The core of the entire project is data. However, it also results in a few specific problems that are covered below. The primary issues that come up when dealing with data, whether they are caused by the tools, the technology itself, or information management, are compiled below.

The quality of the data is a common complaint made by data scientists in actual data science projects. Whether the data is hard to acquire [46] or "dirty" and has problems, data scientists typically come to the conclusion that it lacks the potential to be suitable for machine learning algorithms. For the project to be successful, it is essential to comprehend the types of data that might be available [50], their representativeness for the issue at hand [46], and their constraints [53]. In reality, [59] asserts that results may be inaccurate if coordinated data cleansing or quality assurance procedures are not performed. Data scientists frequently overlook the validation stage in this regard. In order to ensure a strong validation of the suggested fix under actual industrial and commercial circumstances, It is necessary to collect data and/or domain knowledge with enough anticipation.

It's also crucial to take large data into account [60]. The computing demands are increased by a growth in data volume and velocity, which makes the project more dependent on IT resources [8]. Also, the size of the data amplifies the complexity of the technology, as well as the required architecture and infrastructure [48] and, consequently, the associated expenses [43].

Considering huge data is also essential [60]. A rise in data volume and velocity increases the computing requirements, which increases the project's reliance on IT resources [8]. The amount of the data further increases the complexity of the technology, as well as the costs [43] associated with the necessary architecture and infrastructure [48].

One of the most common problems in regard to the constraints of machine learning algorithms is that popular deep learning techniques demand a lot of useful training data, and their reliability is occasionally questioned. The expensive costs of model training and retraining are raised by [51]. In fact, data scientists frequently utilise four times as much data to train machine learning models, which is expensive and resource-intensive. In addition, [51]

notes that data scientists frequently ignore larger business objectives or trade-offs among several objectives in favour of focusing on the incorrect model performance indicators.

Deliverinsights: [41, 9, 60] draw attention to the problem of sluggish data and information sharing among team members. They contend that the ineffective methods for archiving, retrieving, and sharing data and documents waste time since users must search for information and raise the possibility of utilising an incorrect version. Regarding this, [41, 9, 59] show that data science projects lack reproducibility. Since it may be "impossible to further expand on past initiatives given the inconsistent preservation of relevant artefacts" such data, packages, documentation, and interim findings, they really urge action and the creation of new technologies to address the lack of reproducibility.

Table1: Data science projects main challenges

Team Leadership	Project Administration	Management of Data and Information
a lack of cooperation	Process maturity is low.	inability to reproduce
problems in collaboration across teams	ambiguous business goals	Knowledge retention and accumulation Poor data quality for ML
Lack of openness in the exchange of ideas	Having realistic expectations	absence of quality control inspections
ineffective forms of government	Setting realistic project deadlines is challenging	No validation information
lack of analytically inclined individuals	Biased focus on technical concerns	Data privacy and security
Don't rely just on top data scientists	delivering the incorrect item	funding for IT infrastructure
Put together multidisciplinary teams	Project not Utilized by company	inability to reproduce

This could pose a significant challenge to the long-term viability of data science ventures. In a number of applied data science initiatives, the primary output may not be the machine learning model or the projected quantity of interest, but rather an intangible like the project process itself or the knowledge acquired along its development. While achieving the project's goals is critical, there are times when understanding how the project did so, the route it travelled, and the reasoning behind why it took those steps rather than others is more crucial.For comprehending the results and paving the way for future projects, this created knowledge about the path that a data science project takes is essential. Since this knowledge must be handled and maintained in excellent shape, the capacity to replicate data science jobs and experiments is essential.Retaining institutional expertise is difficult, according to [51], because data scientists and developers are in short supply and might take on other positions.

In order to address this issue, [51] suggests that everything be documented and that a thorough record be created for all new machine learning models, allowing for the easy replication of previous employees' work by new workers. In this context, [53] identified

knowledge-sharing within data science teams and throughout the organisation as a crucial element of project success, while [43, 48] included data & information management.

[51] also discussed the problem of various similar but inconsistent data sets in relation to data management, where numerous variations of the same data sets may be in use inside the organisation without a mechanism to distinguish which is the proper one.

Summary

Based on whether they pertain to the a) team or organisation, b) project management, or c) data & information management, the major concerns that have been presented have been divided into three primary types. This taxonomy's objective is to make it easier to understand the numerous issues that could come up when working on a data science project. This classification will also support the assessment of data science approaches later in the text. On [60], the challenges related to big data are categorised into three groups: problems with data, processes, and management. For them, processing issues occur during data processing, whereas management issues relate to issues like knowledge gaps, privacy concerns, and security gaps. We contend that due to its broader breadth, this classification belongs in the proposed taxonomy of difficulties.

Some author's claims that a number of the issues raised are considered as signs or reflections of a bigger problem. Regarding this matter, the author in reference [9] proposed that improving data science methodologies could enhance the success rate of data science projects. In the same article, the author referred to a 2018 survey involving professionals from both industry and non-profit organizations. The survey revealed that 82% of the respondents admitted to not employing a well-defined process methodology for the development of data science projects. However, interestingly, 85% of the respondents believed that adopting a more robust and dependable process would result in more successful data science projects. Therefore, this article aims to explore the following research questions:

- **RQ1:** What project management techniques are available in the literature for data science projects?
- **RQ2:** Are these approaches now in use meeting the requirements of present challenges?

III. RESEARCH METHODOLOGY

In this post, we have chosen to conduct a critical analysis of the literature in order to examine the most recent data science methodologies. A critical literature review is described by Saunders and Rojon as "a mix of our knowledge and understanding of what has been written, our evaluation and judgement skills, and our ability to structure these clearly and rationally in writing" (p. 61). They also emphasise some essential characteristics of a critical literature review: In addition to discussing and evaluating the most pertinent research on the subject, this chapter also acknowledges the most significant and pertinent theories, frameworks, and experts in the field, contextualises and justifies goals, and identifies knowledge gaps that have not yet been investigated**.** Based on the ideas stated earlier and using a comparison of literature on data science project management, the critical literature review that is being presented was completed. The knowledge surrounding the usage of data

science approaches is dispersed across several sources, including scientific journals, books, blogs, white papers, and open internet publishing platforms. This is the main reason why a critical review was chosen instead of a systematic study.

It's true that the knowledge found in unofficial sources was crucial for comprehending the viewpoint of actual data science projects. A set of selection criteria was established in order to choose articles for inclusion:

1. Google Scholar, Web of Science, Mandalay, and other search engines were used as sources.
2. **Timeframe:** 2010 until 2020 (exception on CRISP- DM, from 1995)
3. Journal papers, conference papers, and white papers are among the document types that are available. Papers from practitioners and academics were also used. A practitioner's viewpoint on the data science approaches might be provided via the practitioner articles, which were also included.
4. Content: either explicitly (in the title or text) or indirectly (in the project procedures (inferred by the content). Each of the papers included in this review presented findings pertaining to at least one of the three categories of techniques (i.e. team, project and data & information dimensions of data science project management).

Finally, 19 papers (spanning the years 1996 to 2019) were chosen for analysis. Because there was a lot of information in each study, it was determined that making a comparative table was the most effective way to compare studies. The following four points were made an effort to differentiate the main study components in the first table:

1. Paper information (Author, Journal, and Year)
2. Key concepts and words
3. The viewpoint (team, project, data & information)
4. Findings

REFERENCES

[1] Vol. 2, Issue 5 pp.43-48 (2020)ISSN: 2668-778X www. techniumscience.com
[2] Ida asadisomeh European Conference on Information Systems (ECIS) At: Istanbul, Turkey
[3] www.mbaknol.com/information-systems-management/ethical-security-legal-and-privacy-concerns-of-data mining
[4] Published by Doctor Erick at March 4, 2017
[5] ivypanda.com/essays/ethical-implications-of-data-mining-by-government-institutions
[6] Social,Ethical and Legal Issues of data Mining by john Wang idea group publishing 2003
[7] Marina Da Bormida978-1-80262-414-4, eISBN: 978-1-80262-411-3ISSN: 2398-6018 Publication date: 9
[8] December 2021
[9] Amin choudhary digital development advisor at USAID published oct 10 ,2015
[10] Innov clin neurosci2020 Oct-Dec; 17(10-12): 24–30. Published online 2020 Oct 1.
[11] Volume 8 Issue 2 february 2022 e08981Deborah Wiltshire

