A roadmap to data science: Background, future, and trends

Article *in* International Journal of Intelligent Information and Database Systems · January 2021

DOI: 10.1504/IJIIDS.2021.116459

3 authors, including:

Moaiad Khder
Applied Science University
63 PUBLICATIONS **440** CITATIONS
SEE PROFILE

Samah Fujo
Ahlia University
9 PUBLICATIONS **51** CITATIONS
SEE PROFILE

A roadmap to data science: background, future, and trends

Moaiad Ahmad Khder* and Samah Wael Fujo

Department of Computer Science,
Applied Science University, Bahrain
Email: moaiad.khder@asu.edu.bh
Email: samahwael-1993@hotmail.com
*Corresponding author

Mohammad Adnan Sayfi

Department of Computing, Engineering and Technology,
Faculty of Applied Sciences,
University of Sunderland, UK
Email: msayfe9@gmail.com

Abstract: Data science is a combination of several disciplines that aims to get accurate insights from a bunch of data, develop the technology, and algorithm to solve the complicated problems analytically. Today, data science plays a massive role in our life, and researchers realise how it is essential. Numerous research studies on data science have been published in recent years, but each focus on specific issues, such as data science and its impact on business, manufacturing, academia, and healthcare. This paper will present a roadmap of data science to benefit the readers to know about this field and realise how it is essential in several areas. In addition, it clarifies the mechanism of how data science work and what are the capabilities of the data scientist to be able to work in this field. It also shows the trends and future work of data science.

Keywords: data science; big data; data science trends.

Reference to this paper should be made as follows: Khder, M.A., Fujo, S.W. and Sayfi, M.A. (2021) 'A roadmap to data science: background, future, and trends', *Int. J. Intelligent Information and Database Systems*, Vol. 14, No. 3, pp.277–293.

Biographical notes: Moaiad Ahmad Khder is a senior member of IEEE. He received his PhD degree from the Faculty of Information Science and Technology, The National University of Malaysia, in 2015. He is currently an Assistant Professor with the Computer Science Department, Applied Science University, Bahrain. He has been working on the area of mobile environment, mobile database, data science and cloud computing.

Samah Wael Fujo received her Bachelor's degree in Computer Science on 2019 from Applied Science University, Bahrain. She is currently a Master student in IT and Computer Science and also working as a Research Assistant.

Mohammad Adnan Sayfi received his Master's degree in Data Science on 2019 from University of Sunderland, UK. He is currently working as data science and machine learning analyst.

1 Introduction

Nowadays, 'data science', along with 'big data', became one of the most commonly used phrases in places like industry, economics, media, social networks, and education, with 'data scientist' is among the most common job titles (Yan et al., 2019). The number of data that exists is currently growing at a high rate, tripling every two years and changing the way of our life. In 2012, 2.5 billion gigabytes (GB) of data were generated daily, according to IBM. Forbes's article notes that data is rising exponentially more than ever before, and about 1.7 MB of new knowledge will be generated every second for every human being in the world by the year 2020. Therefore, understanding at least the basics of the 'data science' field extremely important, and that is where our future lies (Arora, 2019). Big data is an integral part of data science where data science applies to all small and large datasets, and then all results are covered by the data analysis process (Chen et al., 2016), Data science built on three primary pillars data, human, and technologies, with the expanding the capacity of gather, store, and examination a consistently developing variety of data that produced by growing the frequency, the field of data science is evolving quickly. Taking into consideration that data science is a new field, which makes a significant share of the attention of the researches, and so many researches have been written about data science that can produce valuable outcomes (Saltz et al., 2017). This paper proposed a roadmap for data science including theoretical background, new trends and future to help researchers to figure out the existing researches gaps and….

2 Theoretical background

2.1 Data science

Data for Technology and Science Committee started publishing the *Data Science Journal* in 2002, and *The Journal of Data Science* was published in 2003 by Columbia University. In the last decade, data science has become popular with the boom of many master online corporations, such as Google, LinkedIn, Yahoo, Facebook and Amazon, and many data-built start-ups, such as Everstring, Climate Corporation, Palantir, and Stitch Fix (Yan et al., 2019).

Today, the world is becoming smarter because of using computational and mathematical techniques. Many of the disciplines are now focused on smart interpretation and their analysis as per the automation requirement. There are many methodologies in practice for that purpose plus the field of 'data science'. The demand for data scientists and its significance has proliferated during the past few years (Pedro et al., 2019).

Moscato and Jane (2019) witnessing the growing availability of vast quantities of data and developments in the fields of artificial intelligence (AI), machine learning (ML), and optimisation. Statistical breakthroughs, discrete new algorithms, and applied mathematics create a new interdisciplinary domain called data science.

In other words, 'data science' is a multidisciplinary combination of data inference, technology, and development of algorithms to solving analytically complicated problems (Kenett et al., 2018; Pedro et al., 2019). Data science is aimed at helping people to make better decisions. In health settings, for example, the goal is to help policymakers, patients,

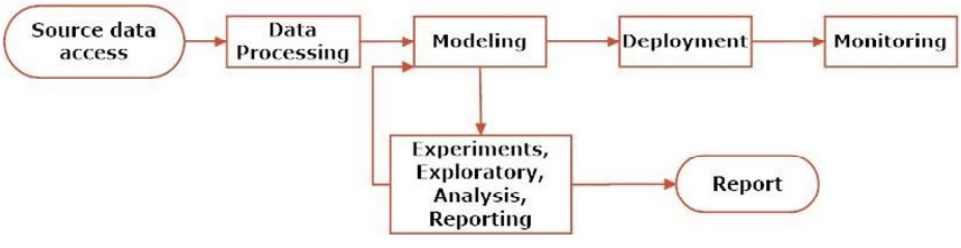
public health officers, clinicians, regulators decide among many possible strategies. Often, the ability of data science to improve decision making depends on its predictability (Hernán et al., 2019).

2.2 Workflow of data science (phases)

There was a growing need for practitioners and a considerable possibility for researchers to know and understand the data science workflow and implement new tools to develop it because big data is overgrowing (Muller et al., 2019).

Knowing from the definition that 'data science' covers a vast spectrum. That means tackling all aspects of 'data science' is an impossible mission. It is hard to find a widely agreed workflow because modern 'data science' does not have a long history. However, the workflow that appears Figure 1. It involves several phases: source data access, data exploration and validation, model development, output generation, exploratory analysis, data monitoring, and deployment. The output may be a model or visualisation. The predictive modelling phase looks like an inner workflow loop, so data scientists have to repeat exploratory and experiment analysis until they are satisfied with the report or model.

Figure 1 Data science workflow (see online version for colours)

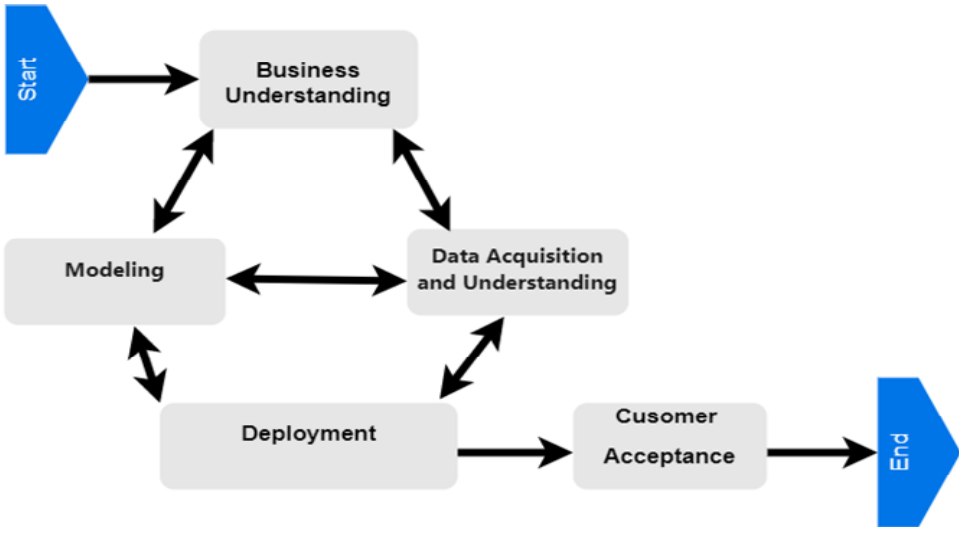


In addition to the workflow, data science roles are another significant area of study. The data science typically has three roles: data scientists, data engineers, and data analysts. Each role is in charge of the various stages of the workflow. However, the stages to which 'data scientists' and 'data analysts' are responsible are overlapping. The work of the data scientist encompasses all the workflow phases shown in Figure 1. Nonetheless, the data analysis work is done when report satisfaction is achieved. Although the 'data scientists and data analysts' are two different job titles, the work of data analysts is part of the work of data scientists to some extent (Zhang, 2020).

2.3 Team data science process

Deploy agile and iterative methodology into data science projects using team data science process (TDSP) that facilitate to implement data science, ML, and AI projects in Microsoft.it reveals the agility of the data science projects, by working as a team and use best practices. TDSP applies on exploratory and ad hoc projects. Figure 2 shows its five phases (Foroughi and Luksch, 2018).

Figure 2 Team data science process (see online version for colours)



The TDSP shows the main phases for projects execution (Foroughi and Luksch, 2018), as follows:

- *Business understanding*: Starting to define the project goals, determine the key variables that help to choose the most relevant model to achieve project goals, identify the relevant source data access that defines the business needs.
- *Data acquisition and understanding*: Make a clean and fine dataset as a priority, define the relationship between variables to give clear target variables. Ingest the dataset into the right analytical environment.
- *Modelling*: Prepare the ML model by finding the best data features, use a ML model that can fit the dataset, and predicts accurate target results, taking into account the production performance for the chosen model.
- *Deployment*: Put ML models into production level, including the data pipeline.
- *Customer acceptance*: Meet the customer's needs, confirm the suggested model and pipeline by the customer to finalise the project.

2.4 Data science vs. big data

Most people think big data and data science are the same, but there is a clear difference between these two terms. Big data is a large set of digital raw data that is difficult to handle and evaluate using traditional methods (Jan et al., 2019). In other words, big data seems to be something that can be used to analyse insights that can lead to better decisions and improvement. While 'data science' is the combination of mathematics, analytics, problem-solving, programming, and ingeniously gathering data, the ability to

look at issues differently, data cleaning, planning, and alignment. Simply put, it is the umbrella of mechanisms used to extract information and insight from big data (Arora, 2019).

2.5 Data science vs. data analytics

Although the terms are used interchangeably by many people, data science and data analytics are distinct fields, with the ranges substantially varying. Data science is an umbrella term for a collection of areas used to mine massive data. While 'data analytics' is a more focused version of this, which can be seen as part of the larger cycle. Analytics is committed to the discovery of actionable insights that can be directly implemented based on the existing queries.

Another essential distinction between the two areas is a matter of exploration. Data science is not concerned with addressing specific questions, but parsing is often unstructured ways of revealing insights across large data sets. Analysis of data works best when it is oriented, keeping in mind questions that need responses based on existing data. Data science creates more in-depth perspectives based on which questions to be posed, while 'data analytics' emphasises the exploration of answers to the questions being posed.

The two fields can be viewed as opposite sides of a similar coin and are strongly intertwined with their functions. Data science lays critical foundations and compiles massive datasets to establish possible significant initial findings, future trends, and potential insights. This knowledge is useful for specific fields on its own, in particular modelling, developing ML, and enhancing AI algorithms as it can improve the sorting and comprehension of the knowledge. Nonetheless, data science raises crucial questions that were previously unaware of while providing little in the way of hard responses. Incorporating data analytics into the mix can transform certain items we know or do not know into practical insights (King, 2019).

2.6 The connection between data science, ML, and AI

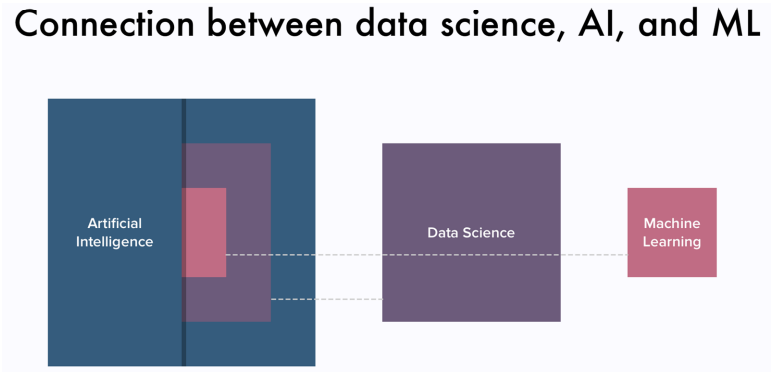
Shabbir and Anwer (2015) define AI today is capable to think like humans and imitates their actions, performing a variety of tasks that need analysing and learning, solve problems, and take decisions.

AI leans towards smart devices such as autonomous vehicles are highly succeeding with high-efficiency features such as analysing traffic, reducing speed; from self-driving cars to personal assistance powered by voice recognition. AI is fast progressing in developing human-like, cognitive characteristics.

ML is used to make machines ready to learn and can handle a big amount of data efficiently and to reveal patterns from a sea of data that was hidden without the use of ML (Dey, 2016).

Rich data sources available increased the use of ML; from search engines to information filtering on social platforms to recommender engines like Netflix also used in smart devices production like cameras and smartphones. Figure 3 shows the connection between data science, AI, and ML.

Figure 3 Connection between data science, AI, and ML (see online version for colours)



2.6.1 Connection between data science and ML

The main connection that is ML is part of the data science. The ML models trained on data given by the data science pipeline, more data fed the model gives more accurate predictions. The difference between them is that data science uses much more than ML technology, so that ML is just a tool in the hand of the data scientist.

The following shows other fields used in data science:

- data engineering
- data visualisation
- model deployment.

2.6.2 Connection between data science and AI

Data science is a more technical part of data management, it uses AI to analyse datasets, reveal hidden patterns towards the predictions part. Eventually, AI and ML powers data scientists to collect more information that leads to having accurate insights.

Data science consists of interpreting datasets, visualising data, and making predictions using various techniques. That helps to build statistical models, AI helps with these models to act like humans.

3 Trends of data science

The trends of data science have been expanded recently to many fields if it is not all, such as healthcare, manufacturing, business, and academia.

As an example of granted data science products – Google Maps, Google collects anonymous data from mobile devices that help real-time algorithms to provide users with accurate traffic data (Schmidt et al., 2014).

Netflix proved the efficiency of using data in its recommendation engines, at first Netflix used data mining techniques by asking users to provide movies rates, to use these ratings in building the recommendation engines that lead to personalised movie recommendations (Maddodi and K, 2019). Now Netflix has its challenge where the

required project is basically to build engines that recommend movies by using user's historical data and the prize was $1 M (Donoho, 2017). Those fields use data science, ML, and AI to reach absolute success.

3.1 Healthcare industry

Thalhah et al. (2019) throughout the world, data scientists are progressively revolutionising the healthcare sector through machine-learning and advanced analytics. They work to improve all facets of healthcare service by leveraging data-gaining organisational expertise and enhancing care delivery; some of these facets are listed below.

3.1.1 COVID-19

In March 2020, the 'World Health Organization' (WHO) announced COVID-19, an illness caused by the 'SARS-CoV-2' virus, a pandemic. Infections rose rapidly, and enormous attempts are being made to contain the illness. Data scientists are already instrumental in addressing emerging COVID-19 challenges (Latif et al., 2020). Some of the recent studies about data science and COVID-19 show in the following:

Scientists at the US Government estimate that COVID-19 could kill thousands of Americans. For those with COVID-19, many preexisting conditions that raise the risk of mortality are the same illnesses that are caused by long-term exposition to air pollution. Wu et al. (2020) researched whether long-term, average exposition to fine particulate matter PM2.5 in the USA is associated with increased risk of COVID-19 death. After contributing to data science and evaluating the data collected, Wu et al. (2020) found that only 1 $\mu g/m^3$ increase in PM2.5 was correlated with an 8% rise in COVID-19 death rates '95% confidence interval {CI}: 2%, 15%'. The tests for secondary and sensibility analyses were statistically relevant and robust. A slight rise in long-term exposition to PM2.5 results in a substantial increase in the risk of death for COVID-19. Although the inherent weaknesses of the nature of a biological sample, the findings highlight the importance of continuing to implement existing regulations on air pollution to save human health during and after that COVID-19 crisis.

The data and code are open to the public for regular updating of the authors' analyses.

Ray et al. (2020) the data science contribution includes an interactive and engaging app (covind19.org) with regular updated short- and long-term forecasts that can help inform COVID-19 policies and practices in India. Anyone can imagine the observational values for India and establish predictions with quantification of uncertainties under hypothetical scenarios. Authors openly made the prediction codes available for reproducible research (https:/github.com/umich-cphds/cov-ind-19) and for other countries affected by COVID-19 to be used for their prediction and data analysis work.

Latif et al. (2020) attempted to systematise the current researches of data science in the area of COVID-19. In addition to reviewing the increasingly growing body of recent study, authors survey public databases and repositories, which can be used to monitor (COVID-19) spread and mitigation strategies for further analysis. The authors present a bibliometric review of the papers published during this short period as part of this. Eventually, depending on these observations, they illustrate common obstacles and pitfalls found in the works being surveyed. Besides, the study created a live content repository1 to keep up to date with the latest tools, including new researches and datasets.

If the researchers keep using data science to analyse the data that comes from current injures to get an accurate prediction, they may achieve the right insights, which can treat this disease 'COVID-19'.

3.1.2 Surgical intensive care unit (SICU)

Tsai et al. (2019) proposed a 'data science framework' for extubation prediction and quantified the value of the data. The results of the study support the consistency of the decision and the valuable knowledge given by the expected model. The implemented framework of data science is general and can be expanded by gathering specified related factors to many other cohorts of ICU patients (e.g., cardiac ICU, medical ICU, neuro ICU, etc.) The methodologies of variable prediction and selection models are more applicable to other industries (e.g., services and manufacturing). In contrast, the decision analysis needs to define constraint and practical knowledge based on the domain applied.

3.1.3 Diagnostics

Data science help in diagnosis decides which form of treatment to be given to the patient and it is seen as a vital part of the clinical care process.

3.1.4 Optimal staffing

For patient care to have enough medical services, it raises the need for healthcare each time, and it finds a challenge many times. In an inflexible patient flow plan, any change often affects the operating units, such as urgent care units and ICUs. It can increase labour costs if there are more than enough personnel available.

3.1.5 Health of public

Numerous healthcare groups have now started using extensive knowledge to improve overall public health. There is a vast number of scattered healthcare data from various sources, such as Google maps, social media, websites, and wearables. These data hold the key to specific geography for understanding general public health. It can be analysed by data scientists to prepare heat maps related to variables such as restorative after-effects of individual citizens in the health, population, geology, conditions, etc.

3.1.6 Discovery of drug

Before a drug is introduced to the market, it takes a lot of money, work, testing, and time-related business. The cost of supplying another drug for general sale is assessed. 'Data science' can use different processes of unstructured and structured biomedical data obtained from various contextual investigations, tests, results of treatment, online networking, and so on. It would then be able to use advanced computational simulations to predict how the drug would interact with body proteins and foresee the accomplishment speed.

3.1.7 Healthcare costs reduction

Healthcare costs already have the earmarks of ascending over time, which proves to be an impacting element in conveying the predominant encounter with patients. Data scientists may examine the charging of knowledge and data from clinical systems concerning variables and charges levels. This helps them access the necessary infrastructure and room use required to take patient needs into account; along these lines, they identify possible regions of revenue loss and organisational gaps.

3.1.8 Wearables device

Wearable devices are quickly becoming omnibus. Besides making frosty ornamentation, they empower the managers in individuals to be self-wellbeing. They record essential readings of well-being, such as rest design, pulse, beat, circulatory strain, and so on. A 'data scientist' can use all those enormous data to analyse it (Latif et al., 2020).

3.2 Advanced manufacturing and business

In Khakifirooz et al. (2019), advances in telecommunications, information technology, and data-enabled decision-making will make advanced manufacturing an integral component of sustainability. Kenett et al. (2018) explored the general role of 'data science', and in particular, the role of analytics in developed manufacturing and sustainability. The authors listed examples of analytical methods and challenges which are available. The information quality framework (InfoQ) is introduced as an infrastructure for the assessment of analytical methods and instruments. The elements of an applied data science roadmap include linking academia and industry and providing testbed environments where advanced analytics technologies can be tested and knowledge hubs where expertise can be shared. In the following, some of the most advanced manufacturing and how data science plays a role to develop them:

3.2.1 Semiconductor industry

In Khakifirooz et al. (2019), due to worldwide demand, the semiconductor industry is one of the few global sectors in a smart growth mode. The significant opportunities that can raise productivity cost reduction and enhance quality in wafer manufacturing are based on real environment simulations in cyber-physical space and sensors. They combine them with deconcentrate decision-making systems. However, the industry faced this convergence with new, unique challenges. Robots, cyber-physical space data stream, can help make the manufacturing intelligent. Therefore, for the delivery of value from manufacturing data, there would be a greater need for modelling, optimisation, and simulation. Khakifirooz et al. (2019) reviewed the great achievement of smart manufacturing in the semiconductor industry, focusing on data-enabled decision-making and data-science-based applications for optimisation. Additionally, this industry discussed future directions for research and new challenges.

3.2.2 TFT-LCD manufacturing

The manufacturers of TFT-LCD panels rely on advanced design and engineering expertise for process control and quality control across the entire production line. To shorten development and minimise labour costs, Lee and Tsai (2019) proposed a three-phase 'data science' system embedded with several 'data mining' and 'ML' techniques. That can classify the variables impacting yield, forecast the metro-logy result of the photo spacer cycle, and suggest cycle control in the manufacturing process of colour filters. In order to validate the proposed framework, an empirical study is being carried out by Taiwan's leading manufacturer of TFT-LCD. The results demonstrate that the proposed framework selects the important variables effectively and quickly, predicts the metrology outcome with better productivity, and highlighted the critical impact and interaction effect of the chosen variables to improve the productivities.

3.2.3 IoT manufacturing

IoT is one of the most vital fields of next-generation technologies that are receiving widespread industry attention. IoT technologies provide enhanced data collection, allowing for real-time responses, improving device access and control, connecting technologies, and growing efficiency and productivity. IoT can be called a smart device deployment that uses connectivity and data. The devices are linked and interacted with one another, and the IoT technologies incorporate the devices' collected data with customer service systems, business analytics instruments, vendor managed inventory systems, and business intelligence apps. The built-in IoT devices efficiently produce vast amounts of data. In IoT, data science may play a significant role in extracting valuable knowledge to predict and analyse this knowledge in order to get better decision-making. Al-badi et al. (2018) presented some of the opportunities IoT and data science require to produce more benefits for academia and industry.

3.2.4 Chemical manufacturing

Piccione (2019) 'data science', smart manufacturing, digitalisation, Industry 4.0: all of these concepts attract significant attention from industry and funding institutions. While some concepts remain hazy, all too many corporations are happy to jump on to a bandwagon because of the economic success of the digital giants. Achievements in the engineering and chemical sciences are also documented, guided by common enablers and the technical specifications of the different applications. High expectations for 'data science' applications in 'chemical engineering' have tends to results, along with a loss of visibility of a purely data-centric approach's limits. At the same time, 'chemical engineers' may not be completely prepared to accept the digital revolution, particularly data science. This short communication, directed at all stakeholders in the chemical industry's digital transformation, points out an inspiring vision for the interplay between 'data science' and chemical engineering, along with challenges, opportunities, and suggested solutions for addressing them. There are several mechanisms for guiding functional strategies: task classes, workflows, and a decision tree for actively deciding what privilege approaches. Piccione (2019) outlined the fundamental problems, and possible solutions, to fruitfully bring chemical engineering and 'data science' together.

3.2.5 Digital business

'Data science' is the hidden ingredient for organisations which have policies or plan to expand and improve their business through data-driven decision making. Information science-based ventures will earn more returns and profits from data-based product creation as well as data-based advice (Pedro et al., 2019). 'Data science' for business and decision-taking brings together the main issues needed to understand and apply decision-making or analytics (Fávero and Belfiore, 2019).

3.3 Education

'Data science' is now a topic of high demand, particularly in academia. Numerous universities are producing new 'data science' majors. Research laboratories are organising head-on workshops besides some of the most common offers like MOOCs and data science-focused coding boot-camps. Despite this increasing attention in the past few years, there are still a few agreements about what a curriculum in data science will include (Kross and Guo, 2019).

In April 2014, the 'Johns Hopkins Data Science' Specialty introduced a nine-course program which has now seen more than 4.2 million student enrolment over the past five years. Kross et al. (2020) have defined and compared the program with standard 'data science' curricula which organised in (2014 and 2015). Kross et al. (2020) showed that in online data science programs, administrative decisions, and novel pedagogical introduced in their program have become standard. It also addresses the effect of the data science discipline on US data science education. Eventually, researchers concluded with some thoughts on the future of 'data science' education in a democratised world of technology.

In Yan et al. (2019), the 'University of Massachusetts Dartmouth' started offering graduate and research programs in data science from 2015. A few articles dealing with graduate data science courses have been published, much less dealing with undergraduate classes. Discussion by the authors focused on the structure and function of the undergraduate curriculum and, precisely, the first data science course. The study analysed the performance of a first-year data science undergraduate course as part of a four-year university level BS in data science. It also elaborated on what core elements are for any beginning data science undergraduate course. The authors' primary purpose is to encourage debate about the related concepts and standards for a constructive introduction to data science.

4 Application of data science

Data science has been outstanding in commercial applications such as shopping and film credit rating, recommendations, stock trading techniques, and ad placement. Some other data scientists have relocated their abilities to scientific research using biomedical applications like the Google algorithm for diagnosing diabetic retinopathy, the Microsoft algorithm for predicting pancreatic cancer years before its regular diagnosis, and the Facebook algorithm for detecting users suicidal (Hernán et al., 2019). Some of the most important applications of data science are listed in the following:

4.1 Internet search

Search engines use 'data science' algorithms to produce the best results in a fraction of seconds for search queries.

4.2 Digital advertisements

The entire digital marketing range uses the algorithms for data science from display banners to digital billboards. That is the main reason why digital ads are getting higher CTRs than traditional ads.

4.3 Recommender systems

The recommendation system helps users find appropriate items from billions of available goods and adds a great deal to user experience. Most companies use this program to advertise their products and suggestions according to the user's demands and knowledge relevance. The recommendations system are based on preceding search results of the consumer (Arora, 2019).

5 Data science open challenges

Acharjya and P (2016) defined that data storage and analysis is part of the data science field, in recent times the big data has grown rapidly and exponentially by many causes such as smart devices, sensors, satellite images, social media, etc. Hence, a massive amount of data leads to the data input/output speed, while sending and retrieving data. Knowing that in some cases the data accessibility is a high priority to be accessed instantly.

One of the most crucial challenges is scalability and data visualisation, as the size of data is fast growing much faster than CPU capabilities, which lead to the use of parallel computing techniques to keep pace with the huge inflation in data volume. For instance, maps applications, social media platforms, search engines need parallel computing powers, moreover, the main objective of the data visualisation is to plot data more effectively, in such companies that have millions of users and reviews such as Amazon that requires decent capabilities to present big and complex amount of data conceptually (Acharjya and P, 2016).

As data science is under the AI umbrella, which AI include automation progress that allows the machines to take human positions in doing such tasks that required human attention, here critical issues arise in the near time, However, AI will create more new opportunities in building, maintaining, and monitoring systems, furthermore, it will significantly increase other issues like data privacy, authenticity, and information security (Shabbir and Anwer, 2015).

6 Data scientist's capabilities

After introducing the data science definition, the question arises: who is the data scientist, or what is supposed to be? Vicario and Coleman (2019) attempted to provide a profile of

data scientists. The 'data scientist' is the person that knows how specific problems relate to the available data, and can thus make the best use of the data to generate added value. The 'data scientist' manages data, ensures its integrity and availability, and mines useful data to provide knowledge and forecasting and support decision making. What are the competencies a data scientist is supposed to have, also? First of all, the 'data scientist' must be able to objectively evaluate problems and consider the actual underlying market or practical issues. They need to be able to manage a great set of techniques. Vicario and Coleman (2019) mentioned the general agreement on statistical approaches that should be available to anyone considered a Data Scientist. Furthermore, expertise in ML is valuable and beneficial for the Data Scientist.

Kross and Guo (2019) presented an interview study of 20 data scientists who teach through industry and academia in different settings. Given the fact that none of them come from formal backgrounds in computer science, they teach a collection of advanced technical skills that make a coherent stack of technologies to allow for transparent and reproducible research.

Data scientist is therefore not expected to be a computer scientist, but must have the proper level of similarity with information technology. In summary, in statistics, mathematics, and computer science, the data scientist has cross-competences. Because we live in a digital era where anything can be a source of digital data, the professional capable of managing data and extracting value from it will be the most appreciated work environment (Vicario and Coleman, 2019).

7 Discussion and future view

The healthcare industry generates a copious amount of data every day. Mathematics is the foundation stone of every present-day scientific order. Practically any modern information technology program, like AI, has strong scientific support, and math plays a vital role in healthcare. Health practitioners must obtain accurate measurements and data for diagnosis, treatment of medical conditions, and presentation. Through ML and advanced analytics, data scientists around the world are rapidly revolutionising the health sector (Thalhah et al., 2019).

For further research, in addition to factors relating to arterial blood gas (ABG), biochemistry, Glasgow coma scale, etc., there are many environmental factors (e.g., family care, doctor review time interval, etc.) and neurological dysfunctions (e.g., dysphagia) that can be considered to improve the decision on extubation and the status of the patient (Tsai et al., 2019).

Data science education programs are required at various levels (universities, schools, colleges, staff, managers, and scientists) to develop human-resource infrastructure and services for data analytics growth. When more companies rely on analytics for their sustainable growth and competitive edge, there are many organisational design patterns emerging around data focus. Businesses which take data thoughtfully are arranged as an asset around the data. Such organisations are democratising data access and giving the right data at the right time to the right user'. Those organisations are encouraging and endorsing data sharing which requires several parts of the company collaborate (Kenett et al., 2018).

Another point, the fast-growing manufacturing of semiconductors requires a 'knowledge management system (KMS)' to support the 'decision support system' (DSS)

management. KMS will identify and analyse trend gaps in research and help organise future research directions for new product innovation (Khakifirooz et al., 2019).

Likewise, (IoT) has a bright and dark side to any technology such as big data, fog computing, and cloud, etc. The research world, however, is presently intended to reduce the IoT-related concerns to make it a reliable, secure, and trusted platform to seek fascinating insight. Research in this field is growing rapidly. It could predict that it will keep going because data is of high value to the organisations and IoT is the primary source for the collection and generation of volumes and data variety. The relationship between IoT and 'data science' is eternal because analytical approaches are needed to convert data into diamonds. There are also many opportunities to apply to the IoT and data science sectors (Al-badi et al., 2018).

Manufacturing is also undergoing a transformation driven by advances in process technology, IT, and data science. A potential manufacturing company is going to be highly interactive. It will open up opportunities for 'ML algorithms' in the spirit of the digital twin idea, to produce predictive models throughout the business. 'Generative adversarial neural networks' gained some attention from the research community in the manufacturing sector. Kusiak (2020) presented descriptive research and implementations of the two principles of ML in manufacturing. The paper addresses the advantages and drawbacks of each neural network. Kusiak (2020) paper may be useful in identifying research gaps, encouraging research on ML in new manufacturing areas, contributing to the creation of effective neural network architectures, and gaining deeper insights into manufacturing data.

Most important of all, Piccione (2019) mentioned some bright potential in the chemical process industries and said that to help achieve this brilliant potential, 'chemical engineers' must collaborate with data scientists, for example, through training, mutual experiences, and joint projects. Data science can also be a great aid to ensure that data is collected and used accurately and efficiently, regardless of the models and methods utilised for their study.

In the coming years, work will be undertaken to provide the same degree of support to data science practitioners and learners in terms of both resources and culture that have been built to date for more conventional programming fields. Besides new technological frameworks for teaching data science, the creation of ways of helping novices understand the social structures that underlie these tools is also essential. Discussions regarding fairness, ethics, and algorithmic bias, for example, are vital to how data science is taught, and who also receives such education. In summary, education of data science is now a rapidly growing type of end-user programming education and computing, distinct from other similar genres traditionally studied in HCI (e.g., end-user programming, learning programming for interaction designers, and conversational programming), with its particular challenges which demand researchers to design new tools and supporting workflows. (Kross and Guo, 2019) the paper was regarded as an invitation to the HCI community – which has already provided several research insights into end-user programming and computer education – to explore further the new frontier of 'data science' learning.

Most significantly, the vast majority of mainstream data science programs target a broad group of early graduate or undergraduate students who are technologically interested. However, there is a significant time to enhance data literacy beyond these groups, extending into the curriculum earlier and reaching across various disciplines. Kross et al. (2020) expect that much of the development in data science education will be

centred on those populations in the coming years. Regardless of the platforms, results, or content data democratisation demands democratisation of 'data science' education that began with the 'Johns Hopkins' Data Science Specialisation.

Lastly, in some ways, the world is in a golden era with fantastic opportunities offered by Data Science expansion before data analytics has become so entrenched that there are rare creative opportunities to develop tailor-made solutions. The need to ensure that the purity and quality of statistics are maintained not only because of its professional sensitivity but also when statistics are applied correctly and sensibly then the optimal benefits of 'data science' are achieved (Vicario and Coleman, 2019).

8 Conclusions

'Data science' is increasingly known as a centre, a multidisciplinary field dedicated to the transformation of data into useful information. To be trained in this area, basic knowledge in data mining algorithms, predictive analytics, computational intelligence, and ML will need to be mastered. Besides, data science incorporates computer science, analytics, applied math, operations analysis, management science, psychology, AI, and economics techniques. Data science's interdisciplinary nature challenges academic institutions all over the world (Moscato and Jane, 2019).

Writing about data science is endless due to the rapid evolution of this field and its vast contribution to multidisciplinary; many data science research directions can still figure out and write about them. In this paper, we tried to cover most of these directions related to the data science concept to help the reader comprehend and realise the importance of data science. It can be considered an excellent reference to the researchers' who is looking for theoretical background, trends, and applications of data science and its future. Moreover, what are the capabilities that supposed to be in any data scientist? In conclusion, as data is the fuel of our futures, it is essential today's need for data scientists to analyse these bundles of data, gain valuable knowledge, and make the right decision, leading to the right achievement and success in many fields.

References

Acharjya, D.P. and P, K.A. (2016) '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.

Al-badi, A., Tarhini, A. and Al-qirim, N. (2018) 'Emerging technologies in computing', in *Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Vol. 200, Issue Illsley 2014, https://doi.org/10.1007/978-3-319-95450-9.

Arora, S. (2019) *Data Science vs. Big Data vs. Data Analytics*, Elérhető [online] http://www.Simplilearn.Com/Data-Science-vs-Big-Data-vs-Data-Analytics-Article (accessed 3 September 2020).

Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y. and Li, L. (2016) 'Big data analytics and big data science : a survey', *Journal of Management Analytics*, No. 1, pp.1–42, https://doi.org/10.1080/23270012.2016.1141332.

Dey, A. (2016) 'Machine learning algorithms: a review', *(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 7, No. 3, pp.1174–1179.

Donoho, D. (2017) '50 years of data science', *Journal of Computational and Graphical Statistics*, Vol. 26, No. 4, pp.745–766.

Fávero, L.P. and Belfiore, P. (2019) *Data Science for Business and Decision Making*, Academic Press, USA.

Foroughi, F. and Luksch, P. (2018) 'Data science methodology for cybersecurity projects', *Computer Science & Information Technology (CS & IT)*.

Hernán, M.A., Hsu, J. and Healy, B. (2019) 'A second chance to get causal inference right: a classification of data science tasks', *Chance*, Vol. 32, No. 1, pp.42–49, https://doi.org/10.1080/09332480.2019.1579578.

Jan, B., Farman, H., Khan, M., Imran, M., Islam, I.U., Ahmad, A., Ali, S. and Jeon, G. (2019) 'Deep learning in big data analytics: a comparative study', *Computers and Electrical Engineering*, September 2018, Vol. 75, pp.275–287, https://doi.org/10.1016/j.compeleceng.2017.12.009.

Kenett, R.S., Zonnenshain, A. and Fortuna, G. (2018) 'A road map for applied data sciences supporting sustainability in advanced manufacturing: the information quality dimensions', *Procedia Manufacturing*, Vol. 21, pp.141–148, https://doi.org/10.1016/j.promfg.2018.02.104.

Khakifirooz, M., Fathi, M. and Wu, K. (2019) 'Development of smart semiconductor manufacturing: operations research and data science perspectives', *IEEE Access*, Vol. 7, pp.108419–108430, https://doi.org/10.1109/ACCESS.2019.2933167.

King, T. (2019) *Data Science vs. Data Analytics – What's the Difference* [online] Https://Solutionsreview.Com/Business-Intelligence/DataScience-vs-Data-Analytics-Whats-the-Difference/ (accessed 14 May 2020).

Kross, S. and Guo, P.J. (2019) 'Practitioners teaching data science in industry and academia: expectations, workflows, and challenges', *Conference on Human Factors in Computing Systems – Proceedings*, pp.1–14, https://doi.org/10.1145/3290605.3300493.

Kross, S., Peng, R.D., Caffo, B.S., Gooding, I., Leek, J.T., Kross, S., Peng, R.D., Caffo, B.S., Gooding, I., Leek, J.T., Kross, S., Peng, R.D., Caffo, B.S., Gooding, I. and Leek, J.T. (2020) 'The democratization of data science education the democratization of data science education abstract', *The American Statistician*, Vol. 74, No. 1, pp.1–7, https://doi.org/10.1080/00031305.2019.1668849.

Kusiak, A. (2020) 'Convolutional and generative adversarial neural networks in manufacturing', *International Journal of Production Research*, Vol. 58, No. 5, pp.1594–1604, https://doi.org/10.1080/00207543.2019.1662133.

Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J. and Tyson, G. (2020) 'Leveraging data science to combat COVID-19: a comprehensive review', *EEE Transactions on Artificial Intelligence*, USA.

Lee, C.Y. and Tsai, T.L. (2019) 'Data science framework for variable selection, metrology prediction, and process control in TFT-LCD manufacturing', *Robotics and Computer-Integrated Manufacturing*, February 2018, Vol. 55, pp.76–87, https://doi.org/10.1016/j.rcim.2018.07.013.

Maddodi, S. and K, K.P. (2019) '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.

Moscato, P. and Jane, N. (2019) 'Memetic algorithms for business analytics and data science: a brief survey', in *Business and Consumer Analytics: New Ideas*, Springer International Publishing.

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Vera Liao, Q., Dugan, C. and Erickson, T. (2019) 'How data science workers work with data', *Conference on Human Factors in Computing Systems – Proceedings*, pp.1–14, https://doi.org/10.1145/3290605.3300356.

Pedro, F., Márquez, G. and Lev, B. (2019) *Introduction to Data Science and Digital Business*, Springer International Publishing, https://doi.org/10.1007/978-3-31995651-0.

Piccione, P.M. (2019) 'Realistic interplays between data science and chemical engineering in the first quarter of the 21st century: facts and a vision', *Chemical Engineering Research and Design*, Vol. 147, pp.668–675, https://doi.org/10.1016/j.cherd.2019.05.046.

Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Mohammed, S., Purkayastha, S., Halder, A., Rix, A., Barker, D., Kleinsasser, M., Zhou, Y., Song, P., Bose, D., Banerjee, M., Baladandayuthapani, V., Ghosh, P. and Mukherjee, B. (2020) *Predictions, Role of Interventions and Effects of a Historic National Lockdown in India's Response to the COVID-19 Pandemic: Data Science Call to Arms*, MedRxiv, 2020.04.15.20067256, https://doi.org/10.1101/2020.04.15.20067256.

Saltz, J., Shamshurin, I. and Crowston, K. (2017) 'Comparing data science project management methodologies via a controlled experiment', *Proceedings of the 50th Hawaii International Conference on System Sciences*, pp.1013–1022, https://doi.org/10.24251/hicss.2017.120.

Schmidt, E., Rosenberg, J. and Eagle, A. (2014) *Google: How Google Works*, Grand Central Publishing, New York.

Shabbir, J. and Anwer, T. (2015) ', "Artificial intelligence and its role in near future', *CoRR*, Vol. abs/1804.01396, 2018 [online] http://arxiv.org/abs/1804.01396.

Thalhah, S.Z., Tohir, M., Nguyen, P.T., Shankar, K. and Rahim, R. (2019) 'Mathematical issues in data science and applications for health care', *International Journal of Recent Technology and Engineering,*, Vol. 8, No. 2, Special Issue 11, pp.4153–4156, https://doi.org/10.35940/ijrte.B1599.0982S1119.

Tsai, T-L., Huang, M-H., Lee, C-Y. and Lai, W-W. (2019) 'Data science for extubation prediction and value of information in surgical intensive care unit', *Journal of Clinical Medicine*, Vol. 8, No. 10, p.1709, https://doi.org/10.3390/jcm8101709.

Vicario, G. and Coleman, S. (2019) 'A review of data science in business and industry and a future view', *Applied Stochastic Models in Business and Industry*, November 2018, pp.1–13, https://doi.org/10.1002/asmb.2488.

Wu, X., Nethery, R.C., Sabath, M.B., Braun, D. and Dominici, F. (2020) *Exposure to Air Pollution and COVID-19 Mortality in the United States*, medRxiv, UK.

Yan, D., Davis, G.E., Yan, D. and Davis, G.E. (2019) 'A first course in data science', *Journal of Statistics Education*, Vol. 27, No. 2, pp.99–109, https://doi.org/10.1080/10691898.2019.1623136.

Zhang, Z. (2020) *DevOps for Data Science System*, School of Electrical Engineering and Computer Science (EECS), p.66 [online] https://www.divaportal.org/smash/record.jsf?pid=diva2%3A1424394&dswid=mainwindow (accessed 3 September 2020).