
Data Science and Machine Learning (ML): An Overview and Scope of Data Science

HIGHLIGHTS

- Introduction
 - What is Data Science?
 - Data Science and Associated Fields
 - Role of Statistics in Data Science
 - Conflicting Definitions of Data Science and its relation to Statistics
 - A brief History and Predictions for Data Science
 - Data Science and Data Analytics
 - Difference between Data Science and Data Analytics
 - Knowledge and Skills for Data Science Professionals
 - Broad View and Data Science Body of Knowledge
 - Data Science and Machine Learning (ML)
 - Application Areas of Machine Learning (ML)
 - Career Path for Data Science Professionals and Data Scientists
 - Outlook
 - Summary
-

Introduction

Data science can be viewed as a multidisciplinary field focused on finding actionable insights from large sets of raw, structured, and unstructured data. The field primarily uses different tools and techniques to unearth answers to the things we don't know. Data science experts use several different areas, including statistical analysis, programming, predictive analytics, data visualization, and machine learning, to parse through massive datasets in an effort to find solutions to problems that haven't been thought of yet.

What Is Data Science?

In this chapter, we explore the field of data science and its relation to machine learning (ML). Data science is about extracting knowledge and insights from data while ML, an integral part of data science, is about analyzing, processing, and creating appropriate models from the data to make decisions. Due to its eclectic definition, data science is a multidisciplinary field that involves the applications of several subjects including, but not limited to, statistics, modelling, mathematics, and computer science to address and solve analytically complex problems using data. In this text, we will explore how to use data science in effective ways to make informed data-driven decisions. In doing so, we will draw insights from the several disciplines that intersect with data science.

Data Science may be thought of as a data driven decision making approach that uses different areas, methods, algorithms, models, and disciplines with the purpose of extracting insights and knowledge from structured and unstructured data. The models in data science are used in predictive analytics and machine learning to predict future outcomes.

Companies now collect massive amounts of data from exabytes to zettabytes which are both structured and unstructured. The advancement in technology and computing capabilities have made it possible to store, process, and analyze this huge data with smarter storage spaces.

Data science is applied to extract information from both structured and [unstructured data](#).^{[1][2]}

Unstructured data is usually not organized in a structured manner and may contain qualitative or categorical elements, such as dates, categories, etc. and are text heavy. They also contain numbers and other forms of measurements. Compared to structured data, the unstructured data contain irregularities. The ambiguities in unstructured data make it difficult to apply traditional tools of statistics and data analysis. Structured data are usually stored in clearly defined fields in databases. In recent years, a number of newly developed tools and software programs have emerged that are capable of analyzing big and unstructured data. One of the earliest applications of unstructured data is in analyzing text data using text mining and other methods.

Data scientists' emphasis is on asking the right questions with a goal of seeking the right or acceptable solutions. The emphasis is on asking the right questions and not seeking specific answers. This is done by predicting potential *trends*, exploring disparate and disconnected data sources, and finding better ways to analyze information. [<https://www.sisense.com/blog/data-science-vs-data-analytics/>].

Data Science and Associated Fields

Data science employs techniques and methods from many other fields, such as mathematics, statistics, computer science, and information science. Besides the methods and theories drawn from several fields, data science also uses data visualization techniques using specially designed software - Tableau and other big data software. The concepts of relational databases (such as SQL), R-statistical software, and programming language – Python are all used in different applications to analyze, extract information, and draw conclusions from data. These are the tools of data science. These tools, techniques, and programming languages provide a unifying approach to explore, analyze, draw conclusions, and make decisions from massive amounts of data companies collect.

Role of Statistics in Data Science

Data scientists should have a strong background in statistics, mathematics, and computer applications. Good analytical and statistical skills are a pre-requisite to successful application and implementation of data science tools. Besides the simple statistical tools, data science also uses statistical modelling including descriptive analytics and predictive and prescriptive analytics tools modelling for decision making and predicting future business outcomes. Thus, a combination of

mathematical methods along with computational algorithms and statistical models is needed for generating successful Data Science solutions. Some key statistical concepts that every data scientist should know include:

- Descriptive Statistics and Data Visualization
- Inferential Statistics Concepts and tools of inferential statistics
- Concepts of probability and probability distributions
- Concepts of Sampling and Sampling Distribution/ Over and Under-Sampling
- Bayesian Statistics
- Hypothesis and other inferential procedures.

Conflicting Definitions of Data Science and its relation to Statistics

Three simplistic and misleading definitions of data science were rejected by Stanford professor David Donoho. These are (1) data science does not equate to big data, in that the size of the data set is not a criterion to distinguish data science and statistics, (2) data science is not defined by the computing skills of sorting big data sets, in that these skills are already generally used for analyses across all disciplines, (3) data science is a heavily applied field where academic programs right now do not sufficiently prepare data scientists for the jobs, in that many graduate programs misleadingly advertise their analytics and statistics training as the data science program.[28][29]. As a statistician, Donoho, following many in his field, champions the broadening of learning scope in the form of data science,[28]. John Chambers urges statisticians to adopt an inclusive concept of learning from data [30]. Together, these statisticians envision an increasingly inclusive applied field that grows out of traditional statistics and beyond.

A Brief History and Predictions for Data Science

The term “Data Science” was created in the early 1960s to describe a new profession that would support the understanding and interpretation of large amounts of data that were being amassed at the time. [<https://www.dataversity.net/brief-history-data-science/>]

The term Data Science is the result of discussions by scientists, statisticians, researchers, computer scientists, and industry pioneers over the years. The following timeline traces the evolution of Data Science from the 1950s. Needless to say, Data Science today categorizes a new profession comprising of the fields of statistics, applied mathematics, computer science, and programming that provides insights based on large amounts of complex data or big data. The following is a discussion on the evolution of Data Science that can be traced back to 1957.

1957	The evolution of Data Science may be traced back to 1957 when Arthur Samuel coined the term “machine learning.”
1962	John Tukey coined the term ‘Data Analysis.’ In 1962, he wrote a paper titled <i>The Future of Data Analysis</i> . He described a shift in the world of statistics and the merging of statistics and computers. John Tukey has several inventions to his name including Box Plot and Exploratory Data Analysis.
1974	In 1974, Peter Naur used the term “Data Science” in his paper <i>Concise Survey of Computer Methods</i> , using the term “Data Science,” repeatedly. Naur presented his own definition of Data Science as: “The usefulness of data and data processes derives from their application in building and handling models of reality.”
1977	The International Association for Statistical Computing (IASC) was formed. The objectives of the Association are to foster world-wide interest in effective statistical computing and to exchange technical knowledge through international contacts and meetings between statisticians, computing professionals, organizations, institutions, governments and the general public. https://iasc-isi.org/about-iasc2/
1986	Carnegie Mellon University professor Hinton co-authors a paper with David E. Rumelhart and Ronald J. Williams on applying the backpropagation algorithm to multi-layer neural networks. This application was a milestone in AI because it allowed the networks to learn internal representations of data.

1989	First Data Science workshop by the Knowledge Discovery in Databases organization.
1997	In November 1997, C.F. Jeff Wu gave the inaugural lecture entitled "Statistics = Data Science?" [10] for his appointment to the H. C. Carver Professorship at the University of Michigan.[11] In this lecture, he characterized statistical work as a trilogy of data collection, data modeling and analysis, and decision-making. In his conclusion, he initiated the modern, non-computer science usage of the term "data science" and advocated that statistics be renamed data science and statisticians data scientists.[10] Later, he presented his lecture entitled "Statistics = Data Science?" as the first of his 1998 P.C. Mahalanobis Memorial Lectures.[12]
2001	William S. Cleveland introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data" in his article "Data Science."
2002	In April 2002, the International Council for Science (ICSU): Committee on Data for Science and Technology (CODATA) [14] started the <i>Data Science Journal</i> , [15] a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues. [16]
2003	in January 2003, Columbia University began publishing The Journal of Data Science, [17] which provided a platform for all data workers to present their views and exchange ideas. The journal was largely devoted to the application of statistical methods and quantitative research
2005	The National Science Board published "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" defining data scientists as "the information and computer scientists, database and software and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a

	digital data collection" whose primary activity is to "conduct creative inquiry and analysis." ^[18]
2006/2007	Around 2007, Turing Award winner Jim Gray envisioned "data-driven science" as a "fourth paradigm" of science that uses the computational analysis of large data as the primary scientific method ^{[4][5]} and "to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other." ^[19]
2012	In the 2012 Harvard Business Review article "Data Scientist: The Sexiest Job of the 21st Century", ^[20] DJ Patil claims to have coined this term in 2008 with Jeff Hammerbacher to define their jobs at LinkedIn and Facebook, respectively. He asserts that a data scientist is "a new breed", and that a "shortage of data scientists is becoming a serious constraint in some sectors" but describes a much more business-oriented role.
2014	The first international conference: IEEE International Conference on Data Science and Advanced Analytics was launched in 2014. ^[22]
	In 2014, the American Statistical Association (ASA) section on Statistical Learning and Data Mining renamed its journal to "Statistical Analysis and Data Mining: The ASA Data Science Journal"
2015	In 2015, the International Journal on Data Science and Analytics ^[25] was launched by Springer to publish original work on data science and big data analytics. Google uses Deep Learning to launch speech recognition, Google Voice, and saw a 49 percent increase in performance. Google launched open-sourced TensorFlow , an artificial intelligence engine to enact Deep Learning using Big Data and Cloud.
2016	In 2016, The ASA changed its section name to "Statistical Learning and Data Science". ^[24]
2017	The DeepMind team released AlphaZero that achieved a superhuman level of play in Chess, Shogi, and Go.

	PricewaterhouseCoopers (PwC) forecasts job listings for data science and analytics will surge to 2.7 million by 2020
<i>2020-present</i>	• Data science jobs will increase by around 28% by 2026. • The global machine learning market was valued at \$8 billion in 2021 and is anticipated to grow at a 39 percent compound annual growth rate (CAGR) by 2027. • The market for Big Data analytics in banking could rise to \$62.10 billion by 2025. • Data creation will grow to more than 180 zettabytes by 2025. • Predictions estimate the world will generate 181 zettabytes of data by 2025. [https://onlinestemprograms.wpi.edu/blog/history-data-science-and-pioneers-you-should-know]

In the past 30 years, Data Science has grown to include businesses and organizations worldwide. It is now being used by governments, geneticists, engineers, and even astronomers. During its evolution, Data Science's use of big data was not simply a "scaling up" of the data, but included shifting to new systems for processing data and the ways data gets studied and analyzed.

Difference between Data Science and Data Analytics?

While the terms data science and data analytics are used interchangeably, data science and big data analytics are unique fields with the major difference being the scope. Data science is an umbrella term for a group of fields that are used to mine large datasets. Data Science has much broader scope compared to data analytics, analytics, and business analytics. Data analytics is a more focused version of data science and focuses more on data analysis and statistics and can even be

considered part of the larger process that uses simple to advanced statistical tools. Analytics is devoted to realizing actionable insights that can be applied immediately based on existing queries.

Another significant difference between the two fields is a question of exploration. Data science isn't concerned with answering specific queries, instead parsing through massive datasets in sometimes unstructured ways to expose insights. Data analysis works better when it is focused, having questions in mind that need answers based on existing data.

Data science produces broader insights that concentrate on which questions should be asked, while big data analytics emphasizes discovering answers to questions being asked.

More importantly, data science is more concerned about asking questions than finding specific answers. The field is focused on establishing potential trends based on existing data, as well as realizing better ways to analyze and model the data. Table 1.2 outlines the difference between Data Science and Data Analytics.

Table 1.2: Difference between Data Science and Data Science

	Data Science	Data Analytics
Scope	Macro	Micro
Goal	Ask the right questions	Find actionable data
Major Fields	Machine learning, AI, search engine engineering, statistics, analytics	Healthcare, gaming, travel, industries with immediate data needs
Analysis of Data and Big Data	Yes	Yes

Some argue that the two fields – Data science and Data analytics-can be considered different sides of the same coin, and their functions are highly interconnected. Data science lays important foundations and parses big datasets to create initial observations, future trends, and potential insights that can be important. This information by itself is useful for some fields, especially

modeling, improving machine learning, and enhancing AI algorithms as it can improve how information is sorted and understood. However, data science asks important questions that we were unaware of before while providing little in the way of answers. By combining data analytics with Data Science, we have additional insights, prediction capabilities, and tools to apply in practical applications.

When thinking of these two disciplines, it's important to forget about viewing them as data science vs data analytics. Instead, we should see them as parts of a whole that are vital to understanding not just the information we have, but how.

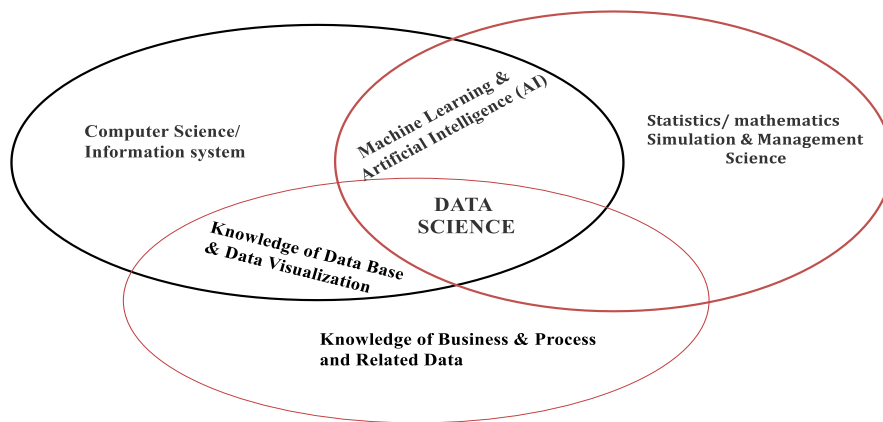
Knowledge and Skills for Data Science Professionals

The key function of the data science professional or a data scientist is to understand the data and identify the correct method or methods that will lead to desired solution. These methods are drawn from different fields including data and big data analysis (visualization techniques) statistics (statistical modeling) and probability, computer science and information systems, programming skills, and an understanding of data bases including querying and data base management.

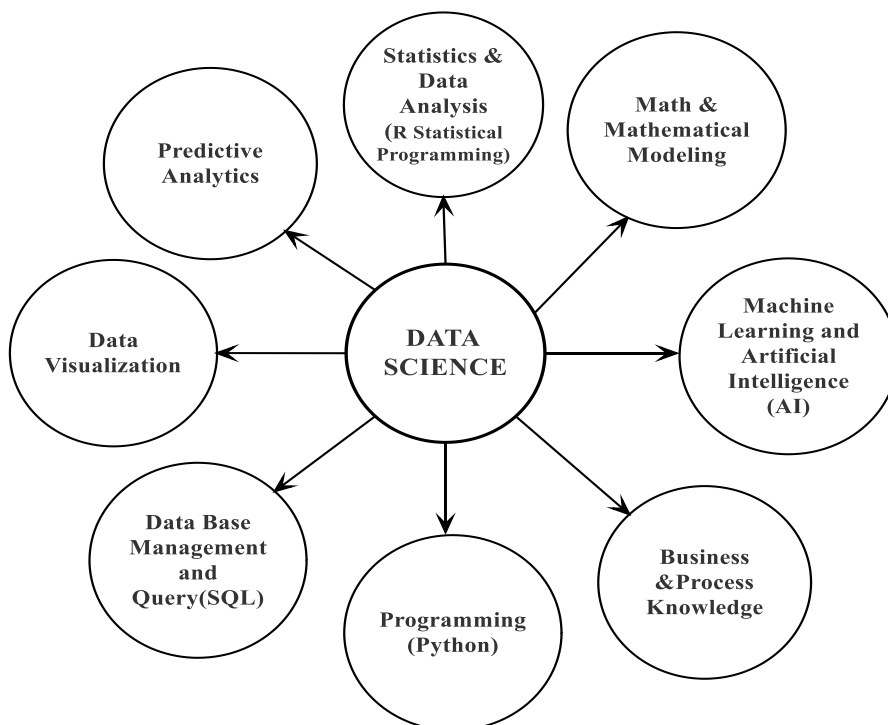
Data science professionals should also have the knowledge of many of the software packages that can be used to solve different types of problems. Some of the commonly used programs are statistical packages (R- statistical computing software), SAS and other statistical packages, relational data base packages – (SQL, MySQL, Oracle and others), Machine learning libraries (recently, many software programs have emerged to automate machine learning tasks are available from software vendors). The two known auto machine learning software are Azur by Microsoft and SAS auto ML.

Broad View of Data Science and Body of Knowledge

Figure 1.1 provides a broader view and key areas of data science. Figure 1.2 outlines the body of knowledge a data science professional is expected to have.



Chapter 1 Figure 1.1: Broad View of Data Science with associated areas



Chapter 1 Figure 1.2: Data Science Body of Knowledge

Data Science and Machine Learning (ML)

Data science can be viewed as an umbrella term that constitutes multidisciplinary fields focused on finding actionable insights from large sets of raw, structured, and unstructured data. The field

primarily uses different tools and techniques in unearthing answers to the things we don't know. (See Figure 1.2). Data Science can be viewed as any process or system used to extract knowledge and insights from data. **Machine learning (ML)** is one of the key elements of Data Science. It may be considered a subset of artificial intelligence (AI). The field of machine learning can be seen as one of the most important developments in data science. Machine learning focuses on building a set of algorithms designed to find patterns in large data sets (big data). These algorithms are applied solve various problems by learning from large quantities of data. Machine learning has become very popular because of recent advancement in computing technologies, improved computing and storage capabilities that makes handling, storing, and using vast amounts of data. The other reasons for the successful applications of machine learning are difficulty and complexity in writing programs for applications like facial recognition that use vast amounts of data. Conventional programming cannot handle huge amounts of data required for facial recognition, stock market predictions, and other applications where real time dynamic data are needed. Abundance of high-quality data can be used efficiently in machine learning models. They also improve performance as new data becomes available.

Application Areas of Machine Learning (ML)

Machine learning algorithms are used in a wide variety of applications, such as in engineering, medicine, email filtering, speech recognition, agriculture, computer vision, facial recognition, signal processing, and cyber security to name a few. In these applications, it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.^{[3][4]} The focus of this book is on machine learning and its applications. We devote Chapter 2 to discussing the details of machine learning.

Career Path for Data Science Professionals and Data Scientists

In order to pursue a career in Data science, a significant amount of education and experience is required. As evident from Figure 1.2 above, a data scientist requires knowledge and expertise from varied fields. The field of data science provides a unifying approach by combining varied areas ranging from statistics, mathematics, analytics, business intelligence, computer science, programming, and information systems. It is rare to find a data science professional with knowledge and background in all these areas. It is often the case that a data scientist has specialization in a sub-field. The minimum education requirement for a data science professional is a bachelor's degree in mathematics, statistics, or computer science. A number of data scientists possess a master's or a PhD degree in data science with adequate experience in the field. The application of data science tools varies depending on the field it is applied to. Note that data science tools and applications when applied to engineering may be different from computer science or business. Therefore, successful application of tools of data science requires expertise and knowledge of the process. Because of the recent boom and interest in data science, machine learning, and artificial intelligence (AI), many universities are now offering degrees and/or specializations in data science and machine learning ranging from undergraduate to graduate degrees.

Outlook

Data science is a growing field. It continues to evolve as one of the most sought-after areas by companies. A career in data science is ranked at the third best job in America for 2020 by Glassdoor, and was ranked the number one best job from 2016-2019.^[29] Data scientists have a median salary of \$118,370 per year or \$56.91 per hour.^[30] These are based on level of education and experience in the field. Job growth in this field is also above average, with a projected increase

of 16% from 2018 to 2028.^[3] The largest employer of data scientists in the US is the federal government, employing 28% of the data science workforce. Other large employers of data scientists are computer system design services, research and development laboratories, big technology companies, and colleges and universities. Typically, data scientists work full time, and some work more than 40 hours a week.^[3]

The outlook for the data science field looks promising. It is estimated that 2.5 to 2.0 million jobs will be created in this area in the next ten years. The data science area is vast and requires knowledge and training from different fields. It is one of the fastest growing areas. Data scientists can have a major positive impact on a business' success,

Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.

Much of the data collected by companies is underutilized. This data, through meaningful information extraction and discovery can be used to make critical business decisions and drive significant business change. It can also be used to optimize customer success and subsequent acquisition, retention, and growth.

Business and research treat their data as an asset. The businesses, processes, and companies are run using their data. The data and variables collected are highly dynamic and continuously change. Data science professionals are needed to process, analyze, and model the data which is usually in

the big data form to be able to visualize, and help companies in making timely data-driven decisions. Data science professionals must be trained to understand, clean, process and analyze the data to extract value from it. It is also important to be able to visualize the data using conventional and big data software in order to communicate data in a meaningful way. This will enable applying proper statistical, modeling, and programming techniques to be able to draw conclusions. All these require knowledge and skills from different areas, and these are hugely important skills in the next decades.” - Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics ³ The increase in demand for data science jobs is expected to grow by 28% by 2020. [<https://datascience.berkeley.edu/about/what-is-data-science/>]. Here are some recent statistics on Data Science and Data Analytics

[<https://techjury.net/blog/big-data-statistics/>]

- Data science jobs will increase by around 31% by 2031(*Source: Towards Data Science*)
- The big data analytics market is set to reach \$103 billion by 2023.
- The demand for composite data analytics professionals will grow by 31% by 2030. (*Source: Forbes*).
- Estimates by the US Bureau of Labor point to a 31% growth by 2030. Composable analytics will be at the forefront of adoption, with 60% of organizations embracing it because of the cost reduction benefit.
- According to Wikibon, the big data analytics market (BDA) is expected to reach \$49 billion with a compounded annual growth rate (CAGR) of 11%. by 2025 So, each year, the market will gain \$7 billion in value. According to this forecast, the BDA market should reach \$103 billion by the end of the prediction period.
- 96% of companies plan to hire job seekers with big data skills. (*Source: The Economist Times*)

Summary:

Data Science is a data-driven decision-making approach that uses several areas, methods, algorithms, models, and disciplines with the purpose of extracting insights and knowledge from structured and unstructured data. These insights are helpful in applying algorithms and models to make decisions.

The chapter explores Data Science and associated areas. Data science as a multidisciplinary field involves the applications of several subjects including, but not limited to, statistics, modelling, mathematics, and computer science to address and solve analytically complex problems using data. The field has a close relation to statistical modeling, and programming.

The chapter provided an overview of Data Science by defining and outlining the tools and techniques and explained the differences and similarities between Data Science and Data Analytics. Data Science continues to evolve as one of the most sought-after areas by companies. The chapter provided a brief history of data science and outlined the career path and job-outlook for this area which continues to be one of the highest of all fields. The field is promising and is showing tremendous job growth. Machine learning (ML) as one of the key elements of Data Science was discussed briefly. The field of machine learning can be seen as one of the most important developments in data science.