
Data Science and its Scope

HIGHLIGHTS

What Is Data Science?

Data Science (M)

Data science is about extracting [knowledge](#) and insights from data that can be structured or [unstructured](#).^{[1][2]}

The objective of this book is to provide an introductory overview of data science, understand what data science is, and why data science is such an important field. We will also explore and outline the role of data scientists and what they do.

The initial chapters of the book provide definitions of data, different forms of data and their types followed by some tools and techniques for working with data. One of the major objectives of Data Science is to make sense from the massive amounts of data companies collect. One of the ways of making sense of data is to apply data visualization or graphical techniques used in data analysis. Understanding other tools and techniques for working with data is also important.

Data Science is a vast area. Besides visualization techniques and statistical analysis, it uses statistical programming languages such as R and knowledge of databases (SQL or MySQL) or other database management systems.

One major application of Data Science is in the area of Machine Learning (ML) and Artificial Intelligence. The book provides a detailed overview of Data Science by defining and outlining the techniques of Data Science. It also explains the differences and similarities between Data Science

and Data Analytics. The other concepts related to Data Science, including Analytics, Business Analytics, and Business Intelligence (BI) are discussed in detail. The field of Data Science is also about processing, cleaning, and analyzing data. These concepts and topics important to understanding the field of Data Science are discussed in this book.

Data Science is an emerging field in data analysis and decision-making.

Data Science may be thought of as a data-driven decision-making approach that uses several different 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 models in Data Science are used in predictive analytics to predict future outcomes.

Data science has a much broader scope than analytics, business analytics, and business intelligence. It brings together and uses several disciplines and areas including [statistics](#), [data analysis](#), [data mining](#), [big data](#), [machine learning](#) & artificial intelligence (AI), management science, optimization techniques, and their related methods in order to "understand and analyze actual phenomena" from data.^[3]

Data science employs techniques and methods from 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 the 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.

Data Science employs the tools of information technology, management science (mathematical modeling and simulation), along with data mining and fact-based data to measure past performance

to guide an organization in planning, and predict future outcomes to aid in effective decision-making.

[Turing award](#) winner [Jim Gray](#) viewed data science as a "fourth paradigm" of science ([empirical](#), [theoretical](#), [computational](#), and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the [data deluge](#).^{[4][5]} In 2015, the [American Statistical Association](#) identified [database management](#), statistics and [machine learning](#), and [distributed and parallel systems](#) as the three emerging foundational professional communities.^[6]

Another Look at Data Science

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

Data scientists' main goal is to ask questions and locate potential avenues of study with less concern for specific answers and more emphasis placed on finding the right question to ask. Experts accomplish this 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: Wikipedia.org https://en.wikipedia.org/wiki/Data_science (From Wikipedia, the free encyclopedia)]

Data Science and Statistics

Conflicting Definitions of Data Science and its relation to Statistics

Stanford professor David Donoho, in September 2015, rejected the three simplistic and misleading definitions of data science instead of criticisms.^[28] First, for Donoho, 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.^[28] Second, 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.^[28] Third, 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 who 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.

28. ^ [Jump up to: ^a ^b ^c ^d ^e ^f ^g ^h](#) Donoho, David (September 2015). *"50 Years of Data Science"* (PDF). Based on a Talk at Tukey Centennial Workshop, Princeton NJ Sept 18 2015. [Archived](#) (PDF) from the original on 19 June 2018. Retrieved 11 May 2018.

29. [^] Barlow, Mike (2013). *The Culture of Big Data*. O'Reilly Media, Inc.

30. [^] Chambers, John M. (1 December 1993). "Greater or lesser statistics: a choice for future research". *Statistics and Computing*. **3** (4): 182–184. doi:[10.1007/BF00141776](https://doi.org/10.1007/BF00141776). ISSN [0960-3174](#).

Data Science and Data Analytics [<https://www.sisense.com/blog/data-science-vs-data-analytics/>]

Data analytics focuses on processing and performing statistical analysis on existing datasets. Analysts apply different tools and methods to capture, process, and organize data, and perform data analysis in the databases of companies to uncover actionable insights from data and find ways to present this data. More simply, the field of [data and analytics](#) is directed toward solving problems for questions we know we don't know the answers to. More importantly, it's based on producing results that can lead to immediate improvements.

Data analytics also encompasses a few different branches of broader statistics and analysis which help combine diverse sources of data and locate connections while simplifying the results.

What is the 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 a much broader scope compared 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.

	Data Science	Data Analytics
Scope	Macro	Micro
Goal	Ask the right questions	To find actionable data
Major Fields	Machine learning, AI, search engine engineering, corporate 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 hard answers. *By adding data analytics into the mix, we can turn those things we know we don't know into actionable insights with 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 the 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 databases 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 to automate machine learning asks are available from software vendors. The two known auto machine learning software are Azur by Microsoft and SAS auto ML. Figure 1.1 provides a broader view and the key areas of data science. Figure 5.2 outlines the body of knowledge a data science professional is expected to have.

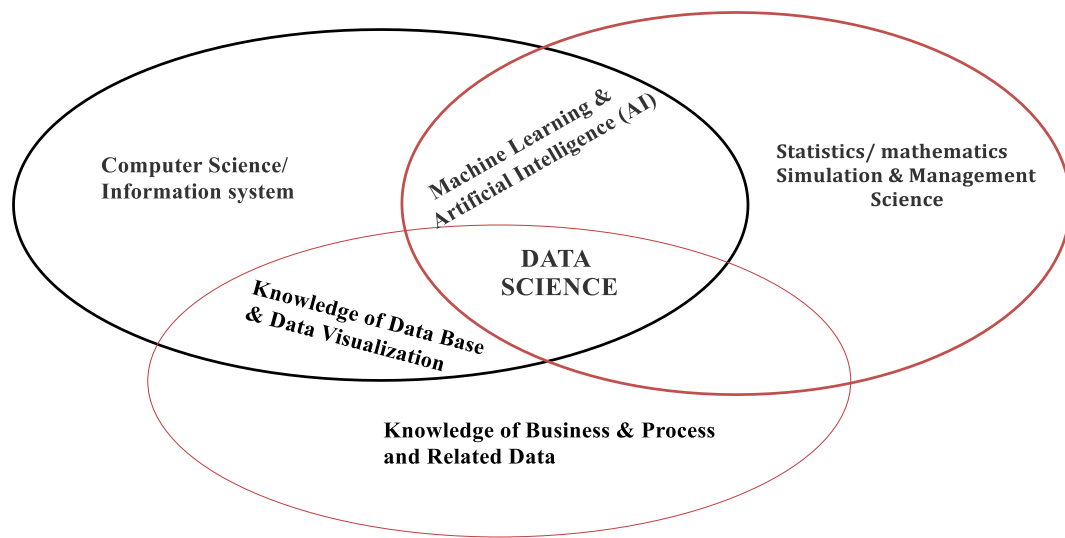


Figure 1.1: Broad View of Data Science with associated areas

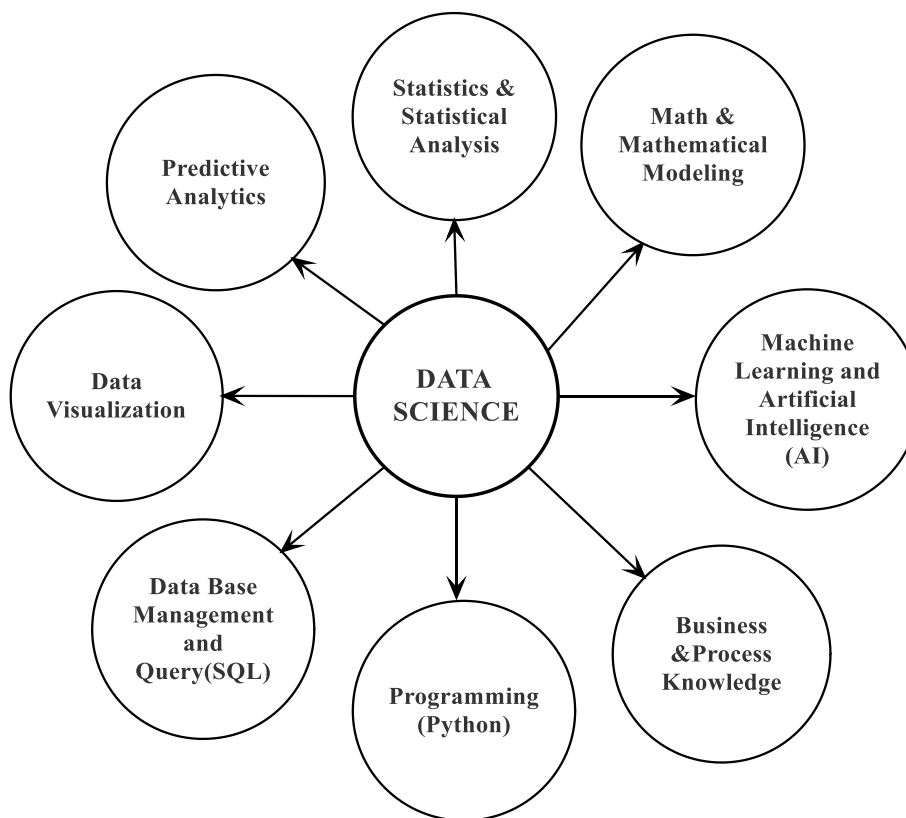


Figure 1.2: Data Science Body of Knowledge

There a number of off-the-shelf data science software and platform in use. The use of these software requires significant knowledge and expertise. Without proper knowledge and background these off-the-shelf software may not be used relatively easily.

Career Path for Data Science Professional and Data Scientist

In order to pursue a career in Data science, a significant amount of education and experience is required. As evident from Figure 1.2, 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 bachelors' degree in mathematics or mathematics. 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 the knowledge of the process.

Some Technologies used in Data Science:

The following is a partial list of technologies used in solving data science problems. Note that the technologies are from different fields including statistics, data visualization, programming, machine learning, and big data.

- [Python](#) is a programming language with simple syntax that is commonly used for data science.^[35] There are a number of python libraries that are used in data science including NumPy, pandas, and scipy.
- [R](#) statistical analysis programming language that was designed for statisticians and data mining^[36] and is optimized for computation.
- [TensorFlow](#) is a framework for creating machine learning models developed by Google.

- [Pytorch](#) is another framework for machine learning developed by Facebook.
- [Jupyter Notebook](#) is an interactive web interface for Python that allows faster experimentation.
- [Tableau](#) makes a variety of software that is used for data visualization^[37].
- [Apache Hadoop](#) is a software framework that is used to process data over large distributed systems.

Data Science Career

Data science is a growing field. A career as a data scientist 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] Job growth in this field is also above average, with a projected increase of 16% from 2018 to 2028.^[30] The largest employer of data scientists in the US is the federal government, employing 28% of the data science workforce.^[30] Other large employers of data scientists are computer system design services, research and development laboratories, and colleges and universities.^[30] Typically, data scientists work full time, and some work more than 40 hours a week.^[30]

Check this site:

<https://intellipaat.com/blog/what-is-data-science/>

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.

“The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”

- Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics ³

Statistics:

Best Job in America 2016, 2017, 2018

<https://datascience.berkeley.edu/about/what-is-data-science/>

[Berkley School of Information]

Companies collect a ton of data, and much of the time it's neglected or underutilized. This data, through meaningful information extraction and discovery of actionable insights, 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.

As mentioned, data scientists can have a major positive impact on a business' success, and sometimes inadvertently cause financial loss, which is one of the many reasons why hiring a top-notch data scientist is critical.

[<https://www.innoarchitech.com/blog/what-is-data-science-does-data-scientist-do>