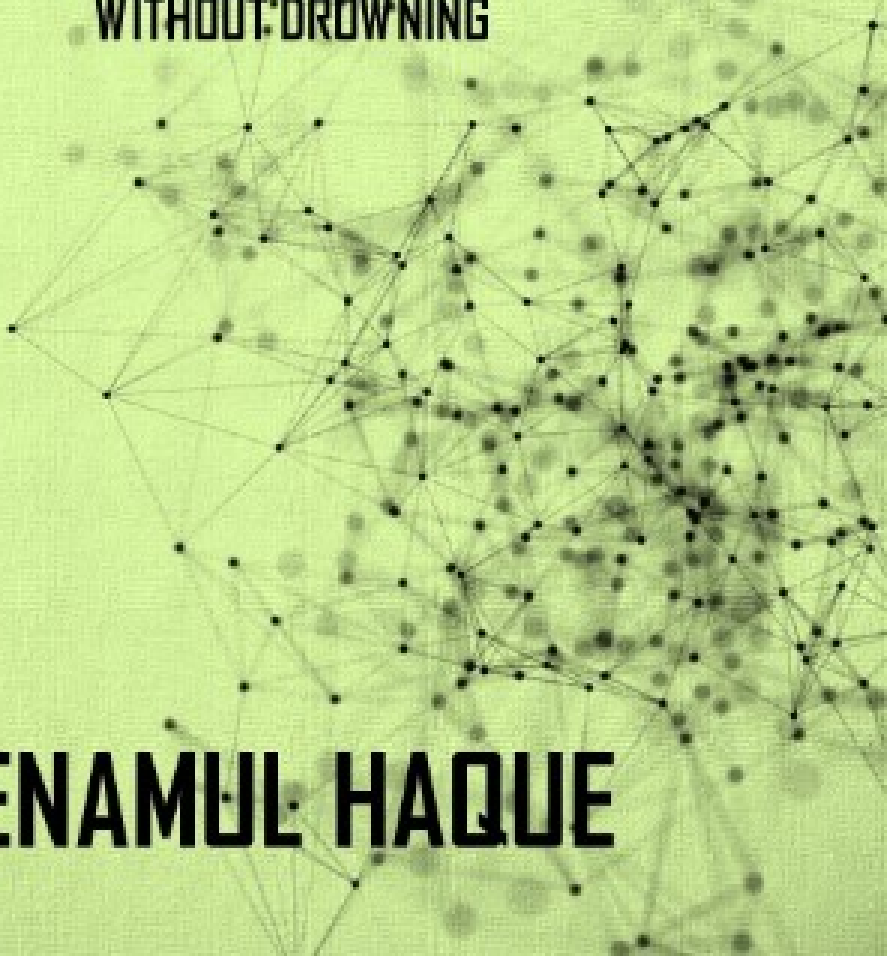
A BEGINNER'S GUIDE TO DATA SCIENCE

HOW TO DIVE INTO THE DATA OCEAN WITHOUT DROWNING



ENAMUL HAQUE

Introduction

We live in a world plagued by an oversupply of data. Websites track any click of any user. Smartphones accumulate information about your location and speed on a daily and every second. "Digitised" selfers wear speedometers on steroids that keep recording their heart rhythms, movement features, eating patterns and sleep patterns. Smart cars collect information about their owners' driving habits, intelligent homes - about their inhabitants' lifestyle, and smart marketers - about our buying habits.

The Internet itself is a massive graph of knowledge, which, among other things, contains an extensive hypertext encyclopedia, specialised databases about films, music, sports results, slot machines, memes and cocktails... and too many statistical reports (some almost true!) from too many government executives, all for you to embrace the immense.

Undoubtfully, the coronavirus pandemic has gripped the world, and people are spending even more time on the Internet, which means more data. In one minute, it turns out that through applications and networks, hundreds of thousands of messages and various online services have spent thousands of dollars. Within a minute, 400,000 hours of movies watched on Netflix, and 500 hours of new videos are uploaded on YouTube. There are 42 million messages sent via WhatsApp, and 347,222 posts are published on Instagram. Facebook users post 147,000 photos in a minute and share 150,000 messages, according to a review by visualcapitalist.com. During the same time, 6,500 parcels ordered through Amazon are delivered, and 208,333 people are conferred in the virtual environment. Mobile apps spend $3,805 per minute, and Doordash orders food 555 times. Also, TikTok has 2,704 people created in a minute, while Twitter has 319 new users. Spotify, meanwhile, is adding 28 new songs to Instagram, with 138,889 people viewing the company's ads. A total of 4.5 billion people use the Internet at the moment. Again, that's a lot of data, Big data in fact.

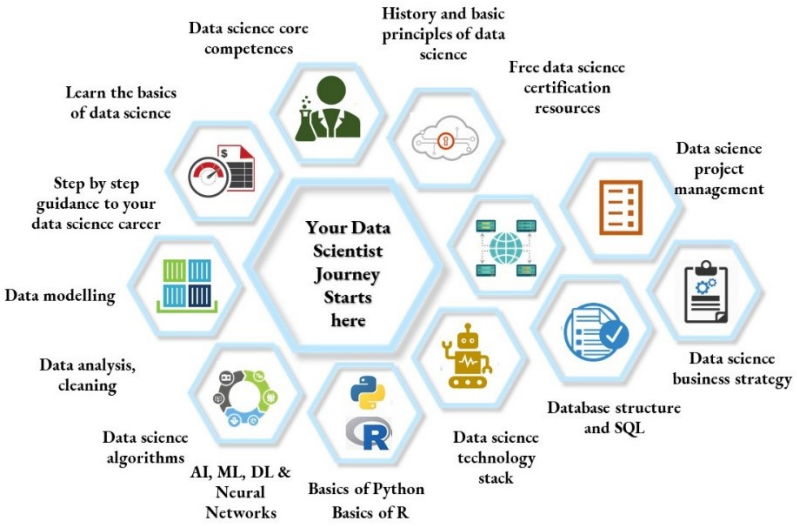


Figure 1 - Highlights of key features of this book

Also, some say data is the new oil, well in many ways, it indeed is. It uses the engines of modern e-commerce, contributes to the development of new products and technologies, is dominated by an extensive network of companies, and is often distributed to natural resources. Overall, it has probably been favourable for human development, as it offers a lot of valuable human insights and allows a wide range of technologies to be distributed for free. On the other hand, science is a systematic enterprise that builds and organises knowledge in the form of testable explanations and predictions about the universe.

Therefore, we can say, data science is a practical discipline that explores methods of generalising the extraction of knowledge from data. Data science consists of various components and is based on methods and theories from many areas of expertise, including signal processing, mathematics, probability models, machine and statistical learning, programming, data technology, image recognition, learning theory, visual analysis, uncertainty modelling, data storage, and high-performance computing to extract meaning from data and create data processing products.

There is a joke that a data analyst is someone who knows statistics better than a computer scientist, and a computer scientist is better than a statistician. Not claiming to be a good joke, but in fact, some data analysts are really experts in mathematical statistics, while others are almost indistinguishable from software engineers. Some are machine learning experts, while others would not be able to learn to find a way out of kindergarten. Some have PhDs with an impressive history of publications, while others have never read academic articles (although they should be ashamed). In short, it does not matter much how to define the concept of data science because you can always find practising data analysts for whom this definition will be completely and utterly incorrect.

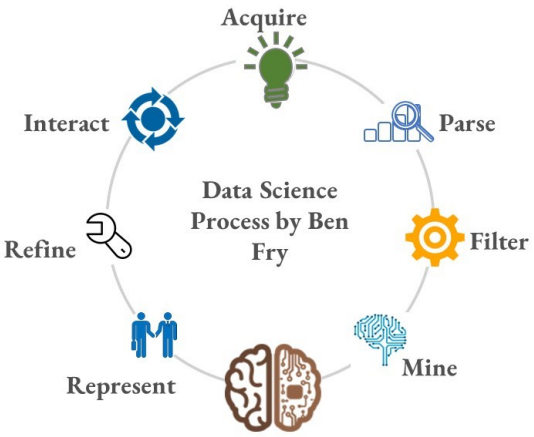


Figure 2 - Ben Fry's data sciecne process model

A data analyst is someone who extracts valuable observations from confusing data. These days, the world is full of people trying to turn data into valuable observations.

For example, the dating site OkCupid asks its members to answer thousands of questions to find the best partner for them. But it also analyses these results to calculate the kinds of innocuous questions you can ask to see how likely intimacy is after the first date.

Facebook asks you to specify your hometown and current location, ostensibly to make it easier for your friends to find you and contact you. But it also analyses these locations to determine global migration patterns and where fans of different football teams live. A major retailer, Target, tracks purchases and interactions online and in-store. It uses data to model predict which customers are pregnant to better sell them baby products.

In 2012, Barack Obama's campaign[i] hired dozens of data analysts who dug and experimented to identify voters who needed extra attention while selecting best appeals and programs to attract financial resources that were directed to specific recipients and focusing efforts to get their opponent out of the race where those efforts might have been most successful. There is a general consensus that these efforts have played an essential role in the President's re-election, making it clear that future political campaigns will be more and more data-driven, leading to a continuous increase in data science and data collection techniques. And before you feel jaded, say a few more words: some data analysts occasionally use their skills for good to make government more efficient, help the homeless, and improve health care. And of course, you won't harm your career if you like to do the best to get people to click on advertising banners.

As the world entered the era of big data, the need for storage increased as well. This was the main task and problem for the enterprises of the industry until 2010. The main focus was on creating storage solutions. Now that Hadoop and other frameworks have successfully solved the storage problem, the focus has shifted to processing this data. Data Science is the secret sauce. All the ideas you see in Hollywood sci-fi movies can really come true thanks to data science. Data science is the future of artificial intelligence. Therefore, it is imperative to understand what data science is and how it can add value to your business.

CHAPTER ONE:
DATA SCIENCE FUNDAMENTALS

"For me, data science was a way to become a detective. For every new case, you have to go into a new field to try to understand how it works, to massage the data until you understand them, to try to acquire all the knowledge of the field without being a specialist. And because I am very curious about very different techniques in science, for me, it is exactly what I wanted." - Godefroy Clair (CTO at Flylab)

Get Started With Data Science

Data science is a discipline that combines statistics, data analysis and related methods to understand and analyse actual events using data. This is a vast area that uses different methods and concepts from other fields, such as mathematics, statistics, and computer science. Data science includes techniques such as machine learning, data engineering, image recognition, visualisation, probability model, signal processing, etc. Over the past few decades, data science has come a long way and has become an essential part of understanding how different industries work.

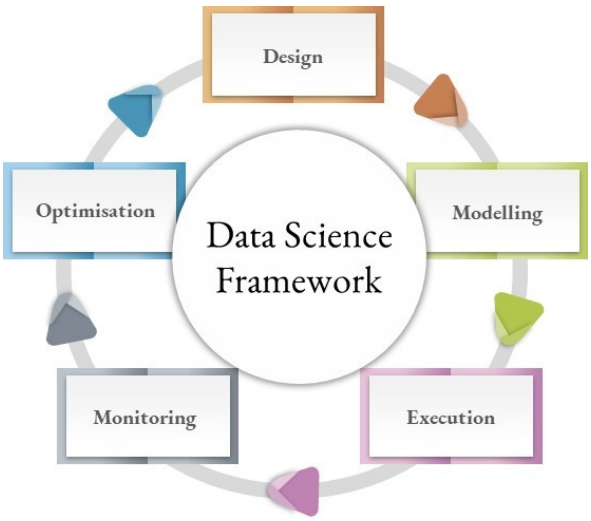


Figure 3 - Data science framework

Data science allows a company to explore and analyse raw data and turn it into valuable information to solve its problems. It enables you to discover insights within the data. By delving into this information at a granular level, the user can find and understand complex trends and behaviours. After all, it's all about bringing information to the surface to help companies make smarter decisions.

For example, Netflix is mining data to discover the viewing patterns of its content to understand what is of interest to users and uses this information to decide which series to produce. Target identifies its key customer segments and purchasing behaviour to be able to address new audiences. Proctor and Gamble rely on data to predict future demand to optimise production.

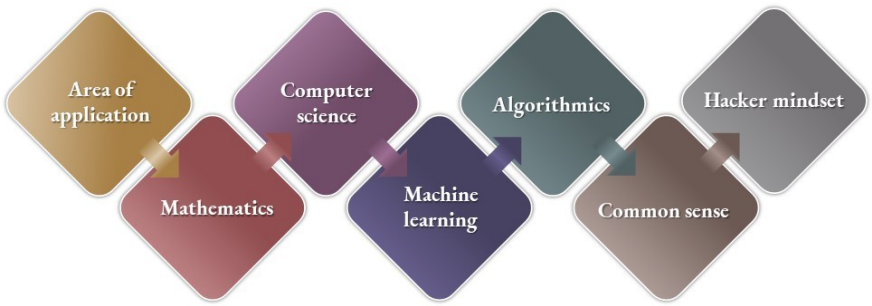


Figure 4 - Various disciplines involved in data science

The range of methods of data science is wide. It includes very non-technical things such as written analyses and simple descriptive statistics, slightly more technical work such as data preparation and visualisation, or mathematically more complex things like predicting time series and automating human activities (such as the pricing of cinema tickets) – and much more.

It is essential to understand that data science's ultimate goal is to solve a problem in a specific area. Having said that, it is necessary to have a very good knowledge of the field of application before embarking on the development of a model. It should also be noted that the areas listed below do not represent an exhaustive list of disciplines involved in data science. In general, data science involves the following fields:

- *The area of application:* The area of application is the sector (The environment) in which you want to make a data product or solve a problem. This could include, for example, the stock market. If we want to establish a predictive model for traders based on past stock prices.
- *Mathematics (Statistics, Probability, Linear Algebra, Analysis, ...):* Mathematics is a significant part of data science. Indeed, problems are very often translated into mathematical models before being solved.
- *Computer science:* Computer science is the basis of data science because models are implemented with code and/or computer tools. Because data is digitally acquired, stored and processed through computing.
- *Machine learning:* Machine learning techniques are increasingly being used in data science.
- *Algorithmics:* Mastery of this science is essential since all models are in the form of algorithms. It is important to understand concepts such as complexity.
- *Common sense:* This is, by far, what is most needed in the face of a complex problem.

Of course, being a data scientist does not mean being an expert in all these areas (even if you have more knowledge in these areas, the better). Indeed, a data science project is very often complex and consists of several steps. So you can find people in a team with different profiles, each in charge of a specific step.

The rise of data science

Traditionally, the data we had was primarily structured and small in size and could be analysed using simple BI tools. Unlike structured data in traditional systems, today, most of the data is unstructured or semi-structured. This data is generated from various sources such as financial journals, text files, media, gauges, and tools. Simple BI tools cannot handle this vast amount and

variety of data that we have today. This is why we need more sophisticated and advanced analytical tools and algorithms to process, analyse and display meaningful ideas. Here are a few reasons that show that data science will always be an essential part of the global economy.

- *Internet search:* Search engines (including Google, Yahoo, Bing and others) use data science algorithms to provide the best possible results for our search queries.
- *Digital advertising:* From banners on websites to digital billboards, almost all of them rely on data provided by scientific algorithms. Online advertising is focused on past user behaviour.
- *Recommendation systems:* Many companies use this system to promote their products and provide suggestions based on the user's interests and relevance.
- *Image recognition:* It is often used to detect certain people, places, or objects inside another larger image.
- *Speech recognition:* This technology is excellent at recognising phonetic sounds and combining them to reproduce spoken words and sentences.
- *Detection of fraud and risks:* Banks and financial institutions have learned to analyse data using customer profiles, past expenditures, and other important variables to predict the likelihood of risk and default.
- *Gaming:* Now, games are created using machine learning algorithms, which rise to a higher level as players advance. In motion games, the computer analyses the players' previous moves and, accordingly, shapes their games.
- *Price comparison:* Algorithms that control price comparisons analyse data and allow you to compare prices for goods sold by different retailers.
- *Planning the airline's itinerary:* Using data science, airlines can predict flight delays, decide whether to land directly at their destination or make intermediate stops, decide which class of planes to buy, and effectively manage customer loyalty programs.
- *Delivery logistics:* Logistics companies use data science to improve their operational efficiency and identify the best delivery routes, the best delivery time, the best mode of transport to choose from, etc.
- *It's different:* Data science is also used in marketing, finance, human resources, health care, government policy, and all possible industries where data is generated.

Data science requires a unique combination of skills and experience. A good data scientist is fluent in programming languages such as R and Python, has knowledge of statistical methods, understands database architecture, and the experience of using those skills to solve real-world problems; we will look into these more details in later chapters of this book.

Progress in data science has been driven by the availability of large data sets and cheap computing power. Without them, data science cannot be effective. A lot of time can be wasted due to small data sets, messy and incorrect data, creating models that give inaccurate or irrelevant results.

The use of data science

Just imagine if you can understand your customers' exact requirements from analysing the existing data such as visitor browsing history, purchase history, age, and income. No doubt you had all this data before, but now with the sheer amount and variety of it, you can train models more

efficiently and recommend the product to your customers with greater accuracy. Isn't this surprising since it will bring more benefits to your organisation?

Let's take another scenario to understand the role of data science in decision making. How about if your car used AI elements to drive you home? The autopilot collects data from sensors, radars, cameras and lasers to create a map of the environment. This data makes decisions, for example, when to accelerate, when to overtake, where to alternate using advanced machine learning algorithms.

Let's also understand how data science can be used in intelligent analytics. Consider an example of weather forecasting. Data from ships, aircraft, radars, satellites can be collected and analysed to create models[ii]. These models not only predict the weather but also help predict the occurrence of any natural disasters. This will help you take the necessary steps ahead of time and save many precious lives.

Data science definitions

Data: The first component of data science, without which the entire further process is impossible, is, in fact, the data itself: how to collect, store and process it, as well as how to extract useful information from the general data array. It is precisely data cleansing and bringing them to the desired form that specialists devote up to 80% of their working time.

Science: We have data; what can we do with it now? Correctly, analyse, extract useful patterns and somehow use them. Disciplines such as statistics, machine learning, optimisation will help us here. They form the next and perhaps most important part of data science - data analysis. Machine learning allows you to find patterns in existing data to then predict the information you need for new objects.

An important part of this point is how to handle data for which standard storage and processing methods are not suitable due to their huge volume and/or variety - the so-called big data. By the way, don't be confused: big data and data science are not synonyms: rather, the first subsection of the second. At the same time, data analysts do not always have to work with big data in practice - small ones can be useful.

Data Science: Data science is a disciplinary mix between data inference, algorithm development and technology, the goal of which is the resolution of complex analytical problems[iii]. At the heart of this great mix is the data, the massive amounts of raw information stored in corporate data warehouses. In practical terms, data science makes it possible to use data creatively to generate value for businesses.

Data science is at the crossroads between technology, so-called pure sciences (mathematics, physics, etc.) and computer science (development). In data science, mathematics and statistics will be used to make probabilistic models or statistical learning. Beyond solid math skills, a layer of computer programming (usually in R or Python) and data engineering is essential. All this knowledge makes it possible to carry out projects around form detection and learning (machine learning), modelling uncertainty, especially in weak signals, data compression, etc.

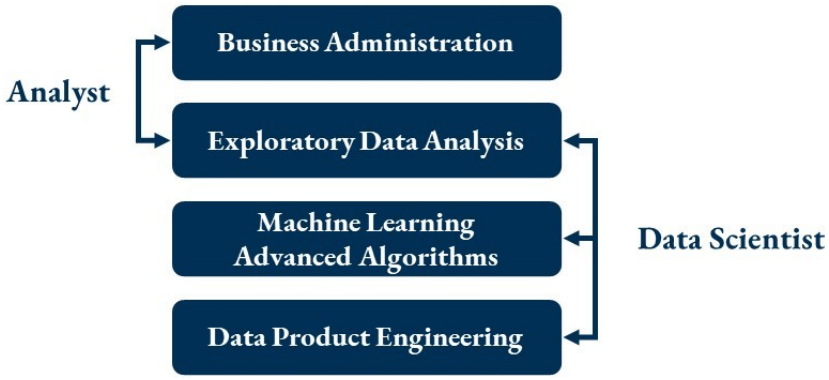


Figure 5 - Differences between Data Scientist and Data Analyst

Data science is a very recent discipline and has been in full development in recent years. The reason? The increase in the volume of data stored by companies, public data and the technical ability to effectively process this data with programming languages allowing extracting value from datasets.

There is no strict definition of data science; the most popular one is: *"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data."*

Data science combines various tools, algorithms, and machine learning principles to discover hidden patterns from raw data. How is this different from what statisticians have been doing for years? The answer lies in the difference between explanation and prediction.

As you can see from the above figure, the data analyst usually explains what is happening while processing the history of the data. On the other hand, data scientist does analysis and uses various advanced machine learning algorithms to identify a particular event's occurrence in the future. The data scientist will view data from many perspectives, sometimes not previously known.

Thus, data science is mainly used for decision making and prediction using predictive causal analytics, prescriptive analytics and machine learning.

Objectives of data science

The purpose of the data scientist is to explore, sort, and analyse big data from various sources to take advantage of it and come to conclusions to optimise business processes or assist in decision-making. For example, we will find machine maintenance or (predictive maintenance) in the fields of marketing and sales with sales prediction according to the weather. The cases of use are almost infinite.

The pillars on which the data scientist relies most often are data mining (data exploration), statistics, machine learning, research algorithms (random forest, decision tree, regression, neural network...), data visualisation with tools such as Matlo. Therefore, data science is revolutionising the processing of corporate data or public data that until then had been difficult to exploit with conventional (so-called structured) technologies. Combining the rapid growth of databases, the emergence of new technologies around machine learning, artificial intelligence and big data now allow semi-structured data analysis to be carried out.

There is a lot of talk about data science regarding big data, but it is not limited to massive data sets.

There will be a strong appetite for data science in areas such as:

- Industry:
    - Predictive maintenance
- Banks and insurance companies with:
    - Process automation
    - Customer knowledge
    - Reducing the attrition rate
- Health:
    - Epidemiology
    - Toxicology
    - Research
- Retail:
    - Sales forecast
    - Customer 360
    - predictive marketing
    - Environments
    - Modelling climate phenomena
    - Impact projection
- Transportation and cities:
    - Smart cities
    - Optimising transport based on passenger flows

There is no shortage of cases of use! Data science's main difficulty is its broadly multidisciplinary aspect at the crossroads between classical sciences, software and programming languages, data security etc.

Benefits of data science

Data science is what makes us human beings as we are today. No, not the computer science of data, but our brain's ability to see connections, draw conclusions from facts and learn from our past experiences. More than any other species on the planet, we depend on our brains for survival. This strategy has already worked for us, and we are unlikely to change it in the near future. But our brains can go that far when it comes to raw computing. Our essence cannot keep up with the amount of data we can collect now and with the level of our curiosity. That's why we turn to machines to do some of the work for us: recognise patterns, create connections, and provide us with answers to our many questions. Finding knowledge in our genes. Relying on computers to do some work for us is a big plus and time savings, especially for big data.

Big data is a generic term for any data set so large or complex that it becomes difficult to process it using traditional data management methods, such as relational database management systems. Widespread databases have long been considered a universal solution, but big data processing requirements have shown the opposite. Data science involves using methods to analyse vast amounts of data and extract the knowledge it contains. You can look at the relationship between big data and data science as the relationship between crude oil and the refinery. Data science and big data came from statistics and traditional data management but are now considered separate disciplines.

Data science and big data are used almost everywhere in both commercial and non-commercial settings. The number of uses is enormous. Businesses in nearly all industries use data science and big data to get an idea of their customers, processes, staff, completions, and products. Many companies use data science to offer customers the best user interface and cross-sales, additional sales and personalise their offerings. A good example of this is Google AdSense, which collects data from Internet users to compare relevant commercial messages to the Internet's user. For example, personalised real-time advertising.

Financial institutions use data science to predict stock markets, determine the risk of lending money, and explore ways to attract new customers for their services. At least half of the world transactions are made automatically on machines based on algorithms, as scientists working on trading algorithms often call using big data and data science techniques. Government organisations are also aware of the value of data. Many government organisations rely on internal data specialists to find valuable information and share their data with the public.

Universities use data science in their research and increase the level of education of their students. Mass open online courses' growth provides a lot of information that allows universities to study how this type of training can complement traditional classes. Massive open online courses are an invaluable asset if you want to become a data scientist and big data specialist, so be sure to look out for some of the best known: Coursera, Udacity and edX; we will also look into more details on this subject later in this book. The situation with big data and data science is changing rapidly, and massive open online courses allow you to study the current disciplines of the best universities. If you are not already familiar with them, take the time to do so now.

Challenges in data science

Data scientists are wrestling with several data-based challenges. The qualitative problems of data are complex and may be related to, for example, data identification or connectivity. In principle, it is worth solving them with business-critical data modelling and standardisation, which are at the core of data management. In addition, data processes as part of business processes need to be fixed. Since data science serves business, it is important to familiarise yourself with what data is used in the business. This is helped by data standards, templates, and dictionary words describing and defining the data your business needs.

It is also possible to study existing data and try to find patterns there. Data mining can be used to identify data that can be useful for business. In this case, you can also specify what data the company lacks. A good tool for mapping existing data is the data catalogue. It describes what data the company uses, where the data is and where it is available. The data catalogue prevents situations so that data scientists do not have visibility and up-to-date information on what data is used in the company.

When you run data science programs in companies, the most significant problems often arise not because of the technology itself but also because of a simple misunderstanding. Misconceptions between departments can lead to severe dissatisfaction between novice data science teams and IT departments.

Data researchers are expected to get wrong, little or no data and turn them into meaningful, actionable predictions is another challenge we may face. Managers may have read articles on the power of machine learning and artificial intelligence and concluded that any data can be fed into an

algorithm and turned into valuable business intelligence. Of course, we know this is not true - your analysis and predictions can be as good as the data you are working with. Of course, statistical techniques can help us fill in our dataset gaps, but there is no magic algorithm that accurately predicts sales in six months when it only learns from a week of data.

Data science specialisation

Data science is a mix of three main fields: mathematical expertise, technology, and business. First, data mining and data product development require an ability to view data through a quantitative prism. The textures, dimensions and correlations between the data can be expressed mathematically. Many of the problems that businesses face can be solved with analytical models based on pure mathematics. Understanding the mechanics of these models is the key to success.

Many people make the mistake of thinking that data science is entirely linked to statistics. Statistics are important but are not the only form of mathematics used. Many machine learning algorithms, for example, rely on linear algebra. In general, a good data scientist must have a solid knowledge of mathematics.

Then, data scientists must be endowed with a form of technological creativity. It uses technology to explore huge datasets and work with complex algorithms to solve complex problems for a good reason. To do this, the data scientist must be able to code, create prototypes of fast solutions, and integrate them into complex data systems. Key languages associated with data science include SQL, Python, R, and SAS. However, knowledge of these languages alone is not enough.

Data science specialists must know how to skillfully navigate between these languages, think algorithmically, and solve complex problems. These faculties are critical because data scientists need to understand the complexity of the data and its flow. A lucidity concerning the connections between these different elements is essential.

A data scientist needs to be a tactical consultant for the company. The data scientist works close to the data to learn more from that data than anyone else. Therefore, it is his responsibility to translate his observations and share his knowledge to help solve the company's problems. He must know how to handle data to tell a coherent story by using insights as a tier.

This relevance to the business is as essential as mastering technology and algorithms. The company's objectives must be aligned with data science projects. In practical terms, the value of a data scientist comes not only from his mastery of mathematics, data and technology but from an association of the three.

For all companies that want to use data to drive business growth, data science is the key. Data science projects can generate significant returns on investment. However, recruiting people with the necessary skills is not an easy task. Once a talented data scientist is hired, it is essential to keep him motivated by offering him the necessary autonomy and offering challenges to match his skills. Learning Data Science requires a reward that is up to the task required.

Understanding data product

A data product is an asset that relies on data and processes it to generate results using an algorithm. The classic example of a data product is a recommendation engine, which ingests user data and generates personalised recommendations based on that data.

Among the most relevant concrete examples are Amazon's recommendation engine or Netflix's. Similarly, Gmail's spam filter is a data product since an algorithm handles incoming emails and determines whether or not they are spam. Computer vision, used by autonomous cars, is also a data product. Learning machine algorithms are capable of recognising traffic lights, detecting other vehicles or pedestrians, etc.

Unlike data insights, data product is not intended to advise a company's executives in their decisions. The accompanying algorithm is designed to be directly integrated into central applications. Examples of data science applications include Amazon's homepage, Gmail's mailbox, or driverless car autopilot software. Data Scientists play a crucial role in the development of data products. They are the ones who develop the algorithms, test them, refine them and deploy them in production systems. This is why data scientists are also technical developers.

Unlocking the true value of data

Today, data is everywhere, and it's no longer just a table of customer and product records. Every time you open a browser, you generate data, every time you click on a website or use an app on your phone, you create data. With the explosion of the digital era comes new challenges around getting data, understanding it, processing it, and just as importantly, when to (and when not to) use it.[iv]

My view is that it is a collection of disciplines used to organise, process, and get the best out of data – whether to meet regulation, gain business understanding, or generate a better customer experience. Whether the action is technical, analytical, or to interpret an output, question it, and understand it. All these things are part of data science.

What Does Data Science Include

Data Science is an interdisciplinary field. It comprises Artificial Intelligence, Machine Learning, and Deep Learning from a bird's eye view.

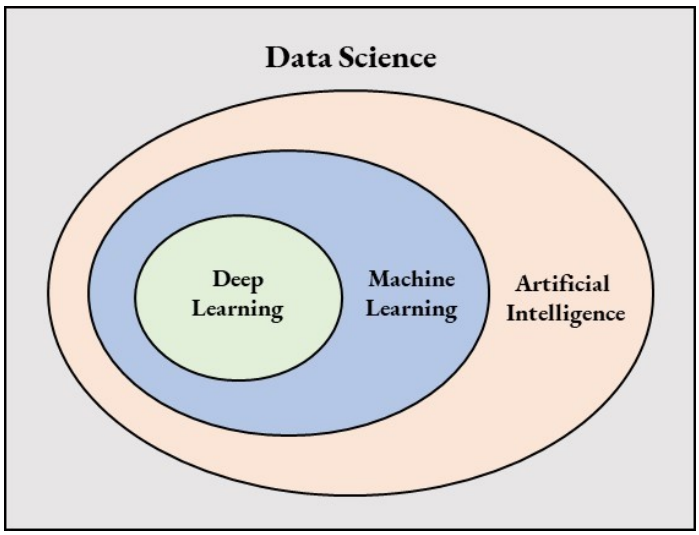


Figure 6 - What does Data Science include?

Artificial intelligence

AI is a field that aims at enabling machines in the replication of human-like intelligence by using Natural Language Processing (NLP), Deep Learning, etc. Chatbots is a recent example of fast gaining wide usage.

Machine learning

Machine Learning provides the machine with the ability to learn without requiring any external programming. The various supervised and unsupervised algorithms are implemented on any classification and regression problem. This makes the field of Data Science even more enjoyable.

Deep learning

Deep Learning is a subset of Machine Learning in the sense that it essentially uses Artificial Neural Networks to mimic how a human brain works. The driverless car is a popular example of Deep Learning.

Mathematics and statistics knowledge

A data scientist must have a strong foundation in Mathematics (Probability) and Statistics concepts. That is the stepping stone into the world of data science.

Descriptive, predictive and prescriptive analytics

All the problems in the realm of Data Science essentially require working on the below areas.

When we try to dig deeper into the past data by applying various methods to understand what has happened, that is called Descriptive Analytics.

When we try to predict the likelihood of any future outcome of a given scenario based on historical data, statistical and Machine learning knowledge is called Predictive Analytics. Understanding about the future in short.

When we try to identify and prescribe the next course of action based on predictive analytics, that is called Prescriptive Analytics.

Technical Knowledge

Implementing the theoretical knowledge with technology is the most exciting part. The journey of a data scientist generally starts with mastering programming languages like R and Python. Additionally, SAS and Excel are also very popular statistical tools that can be learned. Data Visualisation tools like Tableau and Power BI are in high demand being an essential element in the life cycle. Understanding SQL, NoSQL, etc and Big data and Spark, Scala, Hadoop for Data assembly will be an added advantage.

Domain Knowledge

Since every industry demands a data scientist these days, the icing on the cake would be to be an industry-specific data science expert. For example, Healthcare, Finance, HR, Insurance, Energy and Utility, Defense, Education, Media, etc. Nothing like it.

Data Science Foundations

Data Science is a set of specific disciplines from different directions responsible for analysing data and finding optimal solutions. Previously, only mathematical statistics was engaged in this. They began to use machine learning and artificial intelligence, which added optimisation and computer science as data analysis methods to mathematical statistics. Now to get a full picture, you need the understanding of its foundation, which is based on the following pillars:[v]

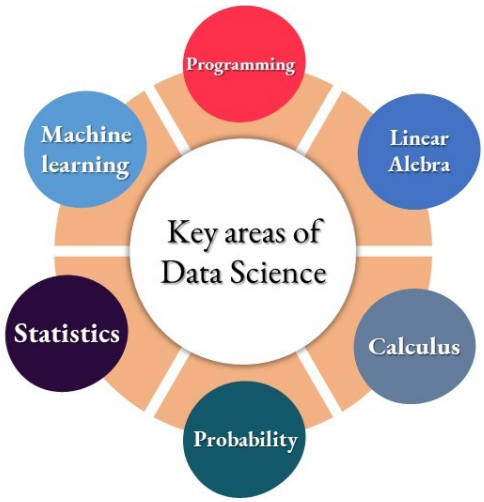


Figure 7 - Data science foundations

Programming

Your first task will be to choose either you'll use Python or R and then immerse yourself into coding.

Linear algebra

As you'll be working with data, you'll want to know how to represent data sets as matrices and understand concepts like vectorisation and orthogonality.

Calculus

Many of the models you'll write and use will use tools like derivatives, integrals and optimisation to compute and find a solution to your problem more rapidly.

Probability

While you use data science, many times, you'll be working to predict something in the future, so you'll want to know how likely something is to happen or why two events are related.

Statistics

To describe the information you'll be analysing, things like the mean or percentiles will come in handy, and tests to check your hypothesis will appear along the way.

Machine learning

Maybe the core of data science is that you'll want to predict something during your project, and that's when machine learning kicks in.

The DIKW Model

Science is examining facts into a theory that has a degree. The basis used is the Data-Information-Knowledge-Wisdom (DIKW) hierarchy. The DIKW pyramid is also known as the DIKW hierarchy, the wisdom hierarchy, the knowledge hierarchy, the information hierarchy, and the data pyramid. Not all versions of the DIKW model reference the four components of science. The DIKW model is often cited and used to define data, information and knowledge in information management, information systems, and literature on knowledge management.

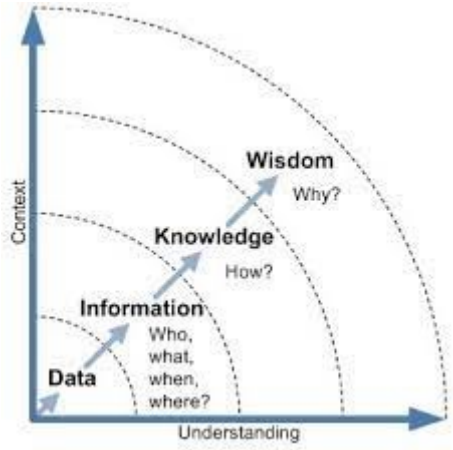


Figure 8 - The DIKW model of knowledge management

Being a data scientists is essentially having a job as a detective, the modern-day Sherlock Holmes. The industry is great for those with curious minds who love solving everyday puzzles. Working with data allows you to play with knowledge[vi]. So let's grow our understanding of this traditional DIKW model.

The relationship of data, information, knowledge and sometimes wisdom in the hierarchy of knowledge has become the language of information science. In 1955 British-American economist and educator Kenneth Boulding presented a hierarchy consisting of signals, messages, information and knowledge using the term knowledge management. In 1987 Milan Zeleny of Czechoslovakia mapped the elements of knowledge into the hierarchy of science formation, namely know-nothing, know-what, know-how and know-why. Zeleny became credited with representing DIKW as a pyramid, although it did not refer to the graphic model.

Data

In the context of DIKW is understood as a symbol or sign representing a stimulus or signal that is useless until it is in use. Zeleny marks as a characteristic of data not used as "know-nothing" Zins refers to it as subjective data where universal data is described as a product of observation, while subjective data is an observation. The definition of data as facts submitted by Rowley data is discrete,

objective fact or observation that is disorganised and has not been processed, so it has no meaning or value because it has not been interpreted in a context. While the data as a signal is understood as sensory stimuli that humans feel through the senses or signal readings, including sensory to the five senses. The opinions Zins expressed were subjective data counts as signals that precede the data in the DIKW.

Information

In the context of DIKW, the information is defined as a description of the knowledge. Distinguished from data helps answer interrogative questions such as 'who', 'what', 'where, how many, 'when'. Information is defined as data that contains meaning and purpose. Rowley describes information as structured data processed so that its information is relevant to a particular context because it is meaningful, valuable, useful, and relevant. In contrast to Ackoff, distinguish between data and structural information is not functional. In the formulation of the proposed hierarchy, Henry defines the information that transforms into functional rather than structural in the difference between data and information.

Knowledge

Knowledge differs from epistemology. The DIKW view in the realm of knowledge refers to information that has been processed, organised or structured in several ways when applied. Zins expresses knowledge of a subjective nature neither universal nor the subject of information science research, although it is often defined in proportionate terms. According to Zeleny, the definition captures knowledge in symbolic form to make information, i.e. all tacit knowledge in human thought. Knowledge as "know-how" and "know-who", and "know-when" is gained through practical experience. Knowledge is an action, not a definition of the action itself. According to Cleveland, Boulding's description of knowledge is a mental structure and definition of knowledge as a result of a person applying informatics to select and organise what is useful to a person.

Wisdom

The wisdom in Zeleny's concept of 'know-why was then refined to distinguish 'why do'. By expanding the definition to include a form of knowledge of what to do. While Ackoff refers to the understanding or appreciation of the 'why'. Wisdom is an evaluated understanding where understanding expresses discrete knowledge and wisdom. Wisdom is the ability to increase the effectiveness of adding value. Cleveland describes wisdom only as composed knowledge information that is made very useful.

History of Data Science

The term data science has appeared relatively recently, but data comprehension has a long history and has been discussed by mathematicians, statisticians and communications professionals for many years. We can also take a good grasp of this from the DIKW model that I described above.

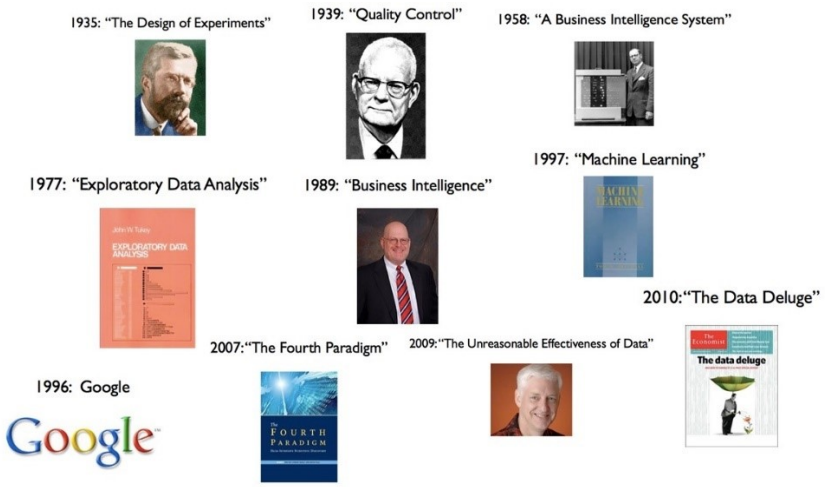


Figure 9 - Data analysis has been around for a while (Abridged Version of Jeff Hammerbacher's timeline for CS 194, 2012)

From statistics to data analysis

The history of data science began even before the advent of computers - in 1948. In "Mathematical Communication Theory," Claude Shannon outlined the basic communication elements through various sources of information. This publication initiated the development of methods for processing, transmitting and storing information. Shannon also developed the concepts of information entropy and redundancy and coined the term bit as a unit of information.

In 1962, mathematician John W. Tuki predicted the impact of modern electronic computing on data analysis as an empirical science. Nevertheless, the modern science of data is far from Tewki's ideas. In 1977, he published a book, Data Research, arguing that more attention should be paid to the use of data to test hypotheses and that research and evidence-based data analysis could and should go side by side. His predictions appeared long before big data and the ability to conduct complex and large-scale computation. The first Programme 101 desktop computer was presented to the public at the World's Fair in New York only in 1964.

By 1981, IBM had released its personal computer, and Apple introduced the first GUI computer two years later. During this decade, computing has evolved rapidly, enabling companies to collect data much more efficiently and efficiently. However, it will be almost two decades before converting this data into information and knowledge.

In 1974, Peter Naur published a book in Sweden and the United States, A Brief Review of Computer Methods. It described the data processing methods of the time that were used in a wide range of applications. They were organised in accordance with the concept defined in the IFIP Guide to Data Processing Concepts and Terms: "Data is a representation of facts or ideas in a "formalised way that can be transmitted or manipulated by a process." The foreword to the book stated that at the IFIP Congress in 1968, a course plan titled "Dataology, Data Science and Data Processing And Its Place in Education" was presented. In the book, the term Data Science was interpreted as a science that deals with data as soon as it is obtained, while the ratio of data to what they represent has been delegated to other areas.

From data analysis to data science

In 1996, Usama Fayyad[vii], Gregory Piatetsky-Shapiro[viii] and Padhraic Smyth[ix] published the book "From Data Mining to Knowledge Discovery in Databases". The authors say that historically, the concept of finding useful patterns in data has received many names, including data mining, knowledge mining, information discovery, data collection, data archaeology, and data patterns. Throughout the 2000s, various scientific journals began to recognise data science as an evolving discipline. In 2005, the National Science Council issued a statement in support of the career development of data science professionals to ensure the availability of experts.

Usama Fayyad

Gregory Piatetsky-Shapiro

Padhraic Smyth

Hal Varian

Drew Conway

Dhanurjay "DJ" Patil

By this time, companies have also begun to view data as a commodity to make money on. Thomas Davenport, Don Cohen and Al Jacobson wrote in a 2005 report by the Babson College Working Knowledge Research Center that instead of competing on traditional factors, companies are beginning to use statistical and quantitative analysis and predictive modelling as key competition elements. In 2009, Google chief economist Hal Varian told McKinsey quarterly that he was concerned about the lack of data to analyse "free and ubiquitous data" of people.

Modern data science

In 2010, Drew Conway[x] published a book in which he wrote, that someone who wants to become a competent data scientist has a lot to learn. Unfortunately, simply listing texts and textbooks does not untangle the knots. As data science evolves and becomes part of the business, so does the need to build strong innovation teams in this area. In 2011, D.J. Patil[xi] published an article entitled "Creating a Data Science Team." He explains what skills, perspectives, tools and processes make such teams successful. In 2021, data science became central to IT amid significant computing advances, as more and more consumers began to master them at lightning speed. With higher processing speeds than ever before, technology has made a giant leap into the new decade. Big data, machine learning and deep learning are central to almost all industries, from business to education and medicine. Today, Data Science specialists are invaluable to any company.

We see that data science's roots lie in statistics and rely on mathematics and computer science. Data science also stems from the practical purpose of using the information to gain knowledge, particularly the idea of using data to solve business problems, as I stated a few times already. Data science will continue to change as human needs change, but one point remains clear, data scientists will be in demand as long as there is data that needs to be analysed. The question is how much data will be available, where it will come from, and what new analysis methods will give them an even deeper understanding.

Data Science Life-Cycle

There are countless interpretations of the life cycle (and of what data science even represents), and I have built this understanding through my research and experience. Data science is a rapidly evolving field, and its terminology is rapidly evolving with it.

Understanding the business

The data scientists in a room are people who keep asking why. These are the people who want to ensure that every decision made in the company is backed by factual data and guarantees (with a high probability) that results will be achieved. Before you can start a data science project, it is imperative to understand the problem you are trying to solve.

According to the Microsoft Azure Blog, we typically use data science to answer five types of questions:

- How much or how much? (Regression)
- What category? (Classification)
- What a group? (Clustering)
- This is strange? (anomaly detection)
- Which option should you choose? (recommendation)

At this stage, you should also define your project's main goals by specifying the variables that need to be predicted. If it's a regression, it might be something like a sales forecast. If it's clustering, it might be a client profile. Understanding the power of data and how you can use it to get results for your business by asking the right questions is more art than science, and it takes a lot of experience to do it well.

Data mining

Now that you've identified your project goals, it's time to start collecting data. Data Mining is the process of collecting data from different sources. Some people tend to group search and data cleansing, but each of these processes is essential in breaking them down into parts. Some questions to consider at this stage are: What data do I need for my project? Where does it live? How can I get this? What's the most efficient way to store and access all of this?

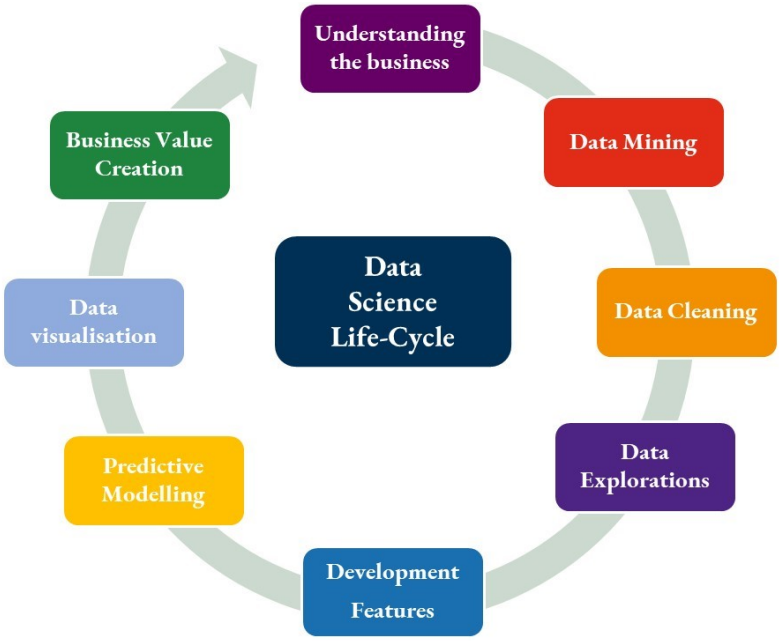


Figure 10 - Data science life-cycle

If all the data required for the project is packaged and transferred to you, you have won the lottery. More often than not, finding the data you want takes time and effort. If the data is stored in databases, your task is relatively simple - you can query the relevant data using SQL queries or manipulate it using a data frame tool like Pandas. However, if your data doesn't actually exist in the dataset, you'll have to clean it up. Beautiful Soup is a popular library used to clean up web pages for data. If you're on a mobile app and want to track user engagement and interactions, there are many tools you can integrate into your app so you can start getting valuable insights from your customers.

Data cleaning

Now that you have received all your data, we move on to the most time-consuming stage - cleaning and preparing the data. This is especially true in big data projects that often use terabytes of data to work with. According to data scientists' interviews, this process (also called "data janitor job") can often take 50 to 80 per cent of their time. So what exactly does this entail, and why does it take so long?

This process takes so long because there are so many possible scenarios that you might need to clean up. For example, data can also have inconsistencies within the same column, which means that some rows can be labelled 0 or 1 and others can be labelled no, or yes, data types can also be inconsistent - some of the 0s can be integers, whereas some of them can be strings. If we are dealing with a categorical data type with several categories, some of the categories may be misspelt or have different cases; for example, they have classes for both male and male. These are just some of the examples in which you can see inconsistencies, and it is important to spot and correct them at this stage.

One of the steps that is often forgotten at this stage and which subsequently cause many problems is the presence of missing data. Lack of data can lead to many errors when building and training a machine learning model. One option is to either ignore instances that have missing values. Depending on your dataset, this may not be realistic if you have a lot of missing data. Another

common approach is to use the so-called mean imputation, which replaces missing values with the mean of all other instances. This is not always recommended because it can reduce your data's volatility, but it makes sense in some cases.

Data exploration

Now that you have a sparkling clean dataset, you're ready to finally start analysing. The data exploration phase is similar to data analysis brainstorming. This is where you understand the patterns and bias in your data. This can include collecting and analysing a random subset of data using Pandas, plotting a histogram or distribution curve to see the overall trend, or even creating an interactive visualisation that lets you dive into each data point and explore the history behind outliers.

With all this information, you begin to form hypotheses about your data and the problem you are solving. For example, if you are predicting student grades, you might try to visualise the relationship between grades and sleep. If you were to predict real estate prices, you could plot prices as a heat map on a spatial graph to see if you can catch any trends.

Development features

In machine learning, a feature is a measurable property or attribute of an observed phenomenon. If we were predicting student outcomes, a possible feature is the amount of sleep they receive. In more complex prediction tasks, such as character recognition, the functions can be histograms that count the number of black pixels.[xii]

According to Andrew Ng, one of the leading experts in machine learning and deep learning: "Working with functions is difficult, time-consuming, and requires expert knowledge. "Applied Machine Learning" is mainly about feature development. "Feature development is the process of using domain knowledge to transform raw data into information features that represent the business problem you are trying to solve. This step will directly affect the accuracy of the predictive model that you create in the next step.

We usually perform two types of tasks in object design - object selection and construction.

Feature selection is the process of reducing features that add more noise than information. This is usually done to avoid the curse of dimensionality, which refers to the increased complexity that occurs in multidimensional spaces (i.e., too many functions). I will not go into details because this topic can be pretty heavy, but we usually use filter methods (we apply a statistical measure to assign a score to each object), wrapper methods (we shape the selection of objects as a search problem and use heuristics to perform the search) or inline methods (use machine learning to figure out which features best affect accuracy).

Building elements involves creating new features from existing ones (and possibly discarding old ones). An example of when you might want to do this is when you have a continuous variable. Still, your domain knowledge informs you that you really only need an indicator variable based on a known threshold. For example, if you have a function for age. Still, your model cares just about whether the person is an adult or a minor; you can set the threshold to 18 and assign different categories for cases above and below this threshold. You can also combine multiple features to make them more informative by taking their amount, difference, or product. For example, if you predicted

student grades and had functions for the number of hours of sleep each night, you might want to create a function.

Predictive modelling

Predictive modelling is where machine learning finally comes into your data science project. I use the term predictive modelling because I believe that a good design is not just a project that trains a model and uses precision but also uses extensive statistical techniques and tests to ensure that the model's results actually make sense and are meaningful. Based on the questions you asked during the business understanding stage, this is where you decide which model to choose for your problem. This is never an easy decision, and there is no single right answer. The model (or models, and you should always test multiple) that you end up training will depend on the size, type and quality of your data, the amount of time and computational resources you are willing to invest. And the kind of output you intend to withdraw. There are several different cheat sheets available online that have a flowchart that helps you choose the correct algorithm based on the type of classification or regression problem you are trying to solve. The two that I really like are Microsoft Azure crib and crib the SAS.

Once you have trained your model, it is very important to measure its success. A process called k-fold cross-validation is commonly used to measure the accuracy of a model. It involves dividing the dataset into k groups of instances of the same size, training in all groups but one, and repeating the process with the various groups omitted. This allows the model to be trained on all data instead of using the typical train test split.

For classification models, we often check accuracy using the PCC (Percentage Correct Classification) together with a confusion matrix that breaks errors into false positives and false negative values. Plots such as ROC curves, which represent actual positive velocity plotted against a false positive velocity background, are also used to evaluate the success of the model. For a regression model, common metrics include the coefficient of determination (which provides information about the confidence of the model), the mean square error (MSE), and the mean absolute error.

Data visualisation

Data visualisation is a tricky area, mainly because it seems simple, but it can perhaps be one of the hardest things to do well. This is because data viz brings together the fields of communication, psychology, statistics and art with the ultimate goal of communicating data in a simple yet effective and visually pleasing way. Once you have gotten the intended ideas from your model, you must present them in a way that the various key stakeholders in the project can understand. Personally, I like working with the parsing and rendering pipeline on an interactive Python notebook like Jupyter, where I can have my code and renderings side by side, allowing fast iteration with libraries like Seaborn and Bokeh. Tools like Tableau and Plotly make it easy to drag and drop your data into visualisation and manipulate it to produce more complex visualisations. If you're creating interactive visualisations for the web, there is no better starting point than D3.js.

Business value creation

Phew. Now that you've gone through your entire life cycle, it's time to get back to the drawing board and extract the business's value creation. Remember, this is a loop, and so it is an iterative process. This is where you measure how your model's success compares to your initial understanding of the business. Does this solve the identified problems? Does the analysis lead to any tangible decisions? Come across any new ideas during the first iteration of the lifecycle (and I assure you that you will). You can now apply that knowledge to the next iteration to generate even more powerful ideas and leverage the power of the data to generate phenomenal results for your business or project.

Case study: diabetes prevention

What if we could foretell the occurrence of diabetes and take appropriate measures beforehand to prevent it?

In this use case, we will predict the occurrence of diabetes by making use of the entire life cycle that we discussed earlier. Let's go through the various steps.

Step 1: First, we will collect the data based on the patient's medical history, as discussed in Phase 1. You can refer to the sample data below.

As you can see, we have the various attributes mentioned below.

Attributes:

- npreg - Number of times you became pregnant
- glucose - plasma glucose concentration
- pb - blood pressure
- skin - Skinfold thickness
- bmi - body mass index
- ped - diabetes function
- age - age
- income - income

Step 2: Now, when we have the data, we need to clean up and prepare the data for data analysis. This data presents many inconsistencies, such as missing values, blank columns, abrupt values, and incorrect data format that need to be cleaned up. Here, we organise the data into a single table under different attributes – making it look more structured. Let's take a look at the sample data below.

This data has many inconsistencies.

Suppose, in the npreg column, "one" is written in words, while it should be in numerical form as 1. Imagine, in the pb column, one of the values is 6600, which is impossible (at least for humans), because bp cannot rise to such a large value.

Also, suppose the "Income" column is blank, and it also makes no sense to predict diabetes. Therefore, it is redundant to have it here and should be removed from the table.

Let's clean up and preprocess this data by removing the discrepant values, filling in the null values, and normalising the data type. If you remember, this is our second phase which is data preprocessing.

Finally, we get the clean data, which can be used for analysis.

Step 3: Now, let's do some analysis. First, let's load the data into the analytical sandbox and apply various statistical functions to it. For example, R has functions like describing, giving us the

number of lost values and unique values. We can also use the summary function to provide us with statistical information such as average, median, range, minimum and maximum values.

Next, we use visualisation techniques such as histograms, line charts, box charts to get a good data distribution idea.

Step 4: Now, based on insights derived from the previous step, the best fit for this type of problem is the decision tree.

We already have the main attributes for analysis like npreg, bmi, etc., so we will use the supervised learning technique to build a model.

In addition, we mainly use the decision tree because it considers all attributes at once, such as those with a linear relationship and those with a nonlinear relationship. In our case, we have a linear relationship between npreg and age, whereas the nonlinear relationship between npreg and ped.

Decision tree models are also very robust because we can use the different combination of attributes to create multiple trees and ultimately implement that with maximum efficiency.

The most critical parameter is the decision tree's glucose level, so it's our root node. Now, the current node and its value determine the next important parameter to take. It continues until we get the result in terms of pos orneg. Pos means that the tendency to have diabetes is positive, and neg means that the tendency to have diabetes is negative.

Step 5: At this stage, we will run a small pilot project to verify that our results are adequate. We'll also look for performance restrictions if any. If the results are not accurate, we need to replan and rebuild the model.

Step 6: After we successfully run the project, we'll share the output for full deployment. Being a data scientist is easier said than done.

[Case study curtesy: Hemant Sharma[xiii]]

Data Science Models

Data modelling is the process of producing a descriptive diagram of relationships between various types of information that are to be stored in a database. One of the data modelling goals is to create the most efficient storage method while still providing for complete access and reporting.

Data modelling is a crucial skill for every data scientist, whether you are researching or architecting a new data store for your company. The ability to think clearly and systematically about the key data points to be stored and retrieved and how they should be grouped and related is what the data modelling component of data science is all about.

Predictive causal analytics

If you want a model that can predict the potential for a particular event in the future, you need to apply predictive analytics. For example, if you lend money, then the likelihood that customers will pay loan payments on time causes you anxiety. Here, you can create a model that can perform analytics on a customer's payment history to predict whether future payments will be timely or not.

Prescriptive analytics

If you want a model with the intelligence to make its own decisions and change it using dynamic parameters, you definitely need analytical forecasting. This is a relatively new field of activity - providing advice. In other words, it not only predicts but also suggests a range of prescribed actions and associated outcomes.

The best example of this is the self-driving car, which I mentioned earlier. The data collected on vehicles can be used to train self-service cars. You can run algorithms on this data to use Artificial Intelligence (AI). This allows your vehicle to decide when to turn, which direction to take when to slow down or accelerate.

Machine learning for making predictions

If you have transactional data from a financial company and need to build a model to determine the future trend, then machine learning algorithms are the best option. This falls under the supervised learning paradigm; we will look into this later. It's called supervised because you already have data from which you can train your machines. For example, a fraud detection model can be trained using a historical record of fraudulent purchases.

Machine learning for pattern discovery

If you do not have parameters from which you can make predictions, you need to figure out the hidden patterns in the dataset in order to be able to make meaningful predictions. This is nothing

more than unsupervised learning since you don't have any predefined categories to the group. The most common algorithm used for pattern detection is clustering.

Let's say you work for a telephone company and you need to create a network by placing towers in a region. You can then use the clustering method to find those towers to ensure that all users receive optimal signal strength.

Business Intelligence (BI) and data science

BI mainly analyses previous data to find an answer in hindsight and uses intuition to describe business trends. BI allows you to take data from external and internal sources, process it, make queries, and create dashboards to answer questions such as quarterly revenue analysis or business problems. BI can assess the impact of certain events in the near future.

It is a more forward-looking approach for data science, an exploratory approach focusing on analysing past or current data and predicting future results to make informed decisions. The data scientist answers open-ended questions about events "what" and "how".

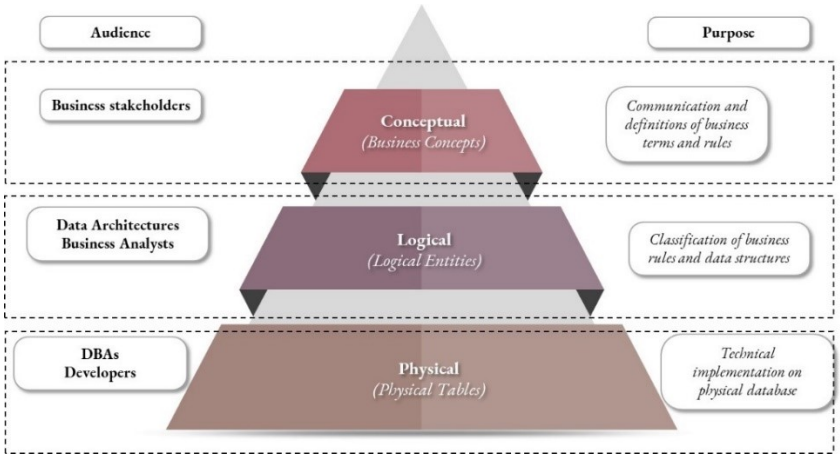


Figure 11 - Levels of data modelling

A common mistake in data science projects is collecting and analysing data, not understanding the requirements without correctly defining the business problem. Therefore, you must follow all stages of the data science lifecycle to ensure the project's smooth operation.

---

[i] **Michael Scherer** - How Obama's data crunchers helped him win - https://edition.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team/index.html

[ii] **Hemant Sharma** - What Is Data Science? A Beginner's Guide To Data Science - https://www.edureka.co/blog/what-is-data-science/

[iii] **DSX HUB** - Data Science: definition, application areas and skills required for data science - https://www.dsxhub.org/data-science-definition-application-areas-and-skills-required-for-data-science/

[iv] **Datascientist.one** - What exactly is data science? - http://datascientist.one/what-exactly-is-data-science/

[v] **Ignacio Montegu**, Mar 7, 2019 - The absolute beginner's guide for data science rookies - https://towardsdatascience.com/the-absolute-beginners-guide-for-data-science-rookies-736e4fcbff0a

[vi] **Orcan Intelligence** - Why Is Data Science So Exciting? - https://medium.com/@Orcanintell/why-is-data-science-so-exciting-de187dcc02c4

[vii] **Usama M. Fayyad** is an American data scientist and co-founder of KDD conferences and ACM SIGKDD association for Knowledge Discovery and Data Mining. He is a speaker on Business Analytics, Data Mining, Data Science, and Big Data. He recently left his role as the Chief Data Officer at Barclays Bank.

[viii] **Gregory I. Piatetsky-Shapiro** is a data scientist and the co-founder of the KDD conferences, and co-founder and past chair of the Association for Computing Machinery SIGKDD group for Knowledge Discovery, Data Mining and Data Science.

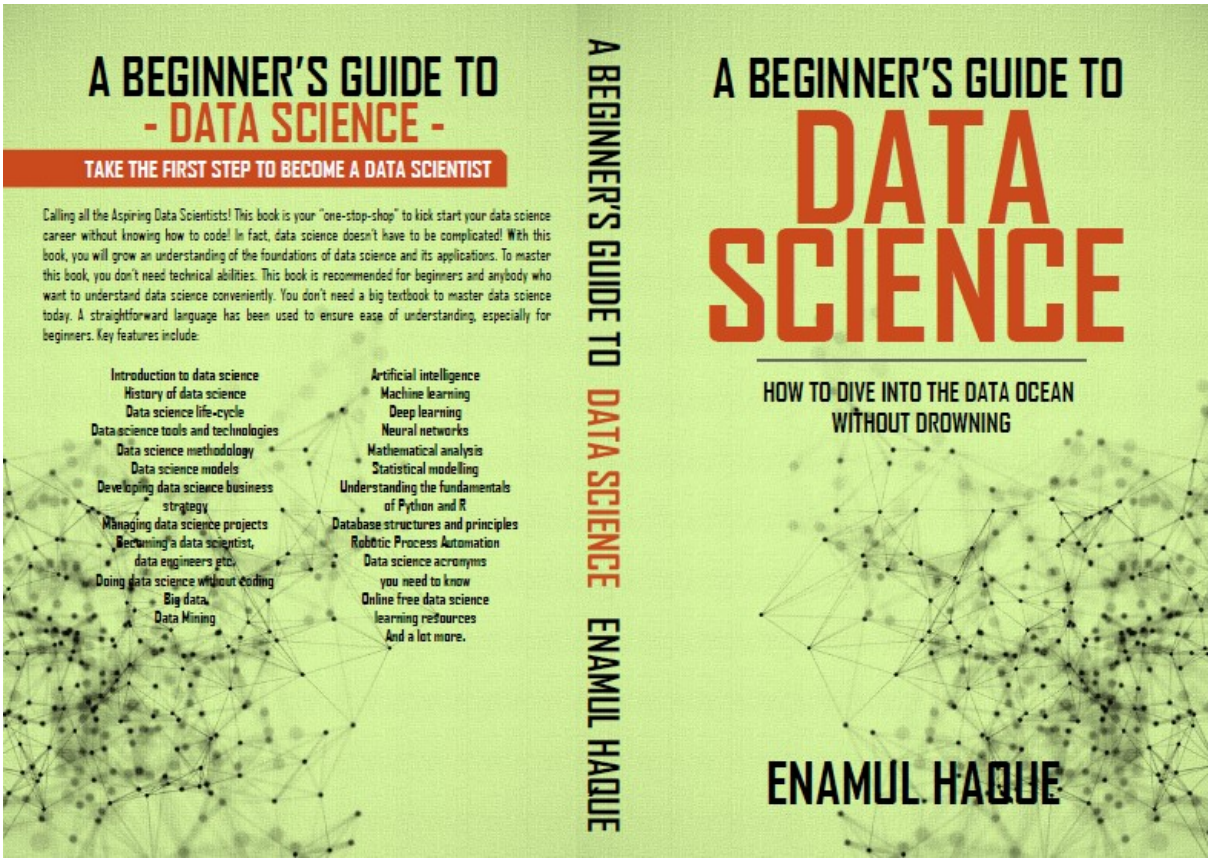
[ix] **Padhraic Smyth** is a Professor of Computer Science in UC Irvine's Donald Bren School of Information and Computer Sciences. He also serves as Director of UC Irvine's Data Science Initiative, and Associate Director for UC Irvine's Center for Machine Learning and Intelligent Systems.

[x] **Drew Conway** is an American data scientist known for his venn diagram definition of data science as well as applying data science to study terrorism. He is currently the founder and CEO at technology startup Alluvium, as well as advisor at multiple technology startups.

[xi] **Dhanurjay "DJ" Patil** is an American mathematician and computer scientist who served as the Chief Data Scientist of the United States Office of Science and Technology Policy. from 2015 to 2017. He is the Head of Technology for Devoted Health.

[xii] **Sudeep Agarwal** - Understanding the Data Science Lifecycle - http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/

[xiii] **Hemant Sharma** - What Is Data Science? A Beginner’s Guide To Data Science - https://www.edureka.co/blog/what-is-data-science/


A BEGINNER'S GUIDE TO
- DATA SCIENCE -
TAKE THE FIRST STEP TO BECOME A DATA SCIENTIST
Calling all the Aspiring Data Scientists! This book is your "one-stop-shop" to kick start your data science career without knowing how to code! In fact, data science doesn't have to be complicated! With this book, you will grow an understanding of the foundations of data science and its applications. To master this book, you don't need technical abilities. This book is recommended for beginners and anybody who want to understand data science conveniently. You don't need a big textbook to master data science today. A straightforward language has been used to ensure ease of understanding, especially for beginners. Key features include:
Introduction to data science
History of data science
Data science life-cycle
Data science tools and technologies
Data science methodology
Data science models
Developing data science business strategy
Managing data science projects
Becoming a data scientist, data engineers etc.
Doing data science without coding
Big data,
Data Mining
Artificial intelligence
Machine learning
Deep learning
Neural networks
Mathematical analysis
Statistical modelling
Understanding the fundamentals of Python and R
Database structures and principles
Robotic Process Automation
Data science acronyms you need to know
Online free data science learning resources
And a lot more.
A BEGINNER'S GUIDE TO DATA SCIENCE ENAMUL HAQUE
A BEGINNER'S GUIDE TO
DATA SCIENCE
HOW TO DIVE INTO THE DATA OCEAN WITHOUT DROWNING
ENAMUL HAQUE