|  |
| --- |
|  |
| Data Science Course |
|  |



learn.rubiscape.io



Contents

[1. Introduction to Data Science 2](#_Toc48770460)

[1.1 What is Data Science? 2](#_Toc48770461)

[1.1.1 Data is everywhere 3](#_Toc48770462)

[1.1.2 Know your data 3](#_Toc48770463)

[1.1.3 Data Science and Industry 4.0 3](#_Toc48770464)

[1.1.4 Applications of Data Science 4](#_Toc48770465)

[1.2 What do data scientists do? 5](#_Toc48770466)

[1.3 History of Data Science 6](#_Toc48770467)

[1.3.1 Chronology of Events 6](#_Toc48770468)

[1.3.2 Advancement of Big Data 7](#_Toc48770469)

[1.3.3 Evolution of Data Science 7](#_Toc48770470)

[1.4 Careers in Data Science 7](#_Toc48770471)

[1.4.1 Data Science Career Opportunities 7](#_Toc48770472)

[1.4.2 Data Science Job Roles 8](#_Toc48770473)

[1.4.3 Skills required for Data Science 8](#_Toc48770474)

[1.5 Business Application in Data Science 10](#_Toc48770475)

[1.5.1 Importance of Data Science in Business 10](#_Toc48770476)

[1.5.2 Business Applications of Data science 10](#_Toc48770477)

[2. Data Science Methodology 12](#_Toc48770478)

[2.1 Data Science Process 12](#_Toc48770479)

[2.2 Stages of Data Science Process 13](#_Toc48770480)

[2.3 Types of Data 15](#_Toc48770481)

[2.4 How to analyse data 15](#_Toc48770482)

[2.5 Data collection, requirements, understanding, EDA 15](#_Toc48770483)

[2.6 Data Preparation 15](#_Toc48770484)

[2.7 Data Modelling and Introduction to alogorithms 15](#_Toc48770485)

[2.8 Data Visualisation 15](#_Toc48770486)

# Introduction to Data Science

Data science is the latest buzzword in the industry. In this chapter, we will study the concepts related to Data Science.

## What is Data Science?

* Data Science is the study of data.
* It is an interdisciplinary field with a blend of science**,** mathematics, procedures, algorithms, and computer systems.
* It is about exploring data to uncover certain patterns.



Figure 1: What is Data Science?

* It extracts knowledge and gives you insights into raw data.
* It identifies a pattern or a trend in data that is unnoticed by humans.
* It includes methods for collecting, storing, and analyzing data to extract useful information.

### Data is everywhere

Humans contribute about 2.5 exabytes i.e. 1018 bytes of data every day. More than 80% of this data is unstructured, i.e. it does not contain any labels or features. For example, posting a comment on Twitter is a type of unstructured data. By 2025, humans will create 463 exabytes of data per day.

As per the estimation by IBM, roughly 3 million Data Scientist jobs will be created in 2 years.

### Know your data

The key to become a good Data Science professional is to know your data well. To know your data well, you need to get answers to some questions such as:

* What do we do with all this data?
* How do we use this data?
* How do we analyze and process this data with the basic knowledge of algorithms and statistics?
* How a large database can help the business?
* Can a clean data set be applied to any business problem?
* Can we use this data to solve business problems or real-world problems?
* Who is the biggest facilitator of data science and analytics?
* Can business and decision-makers afford to ignore the power of data analytics?
* How can businesses benefit from the analysis of this data?
* Will businesses suffer financially and quality-related losses if they do not employ the power of data analytics? If yes, how and how much?

Data science and data analytics help businesses take data-driven decisions. Data-driven decisions are more helpful than decisions based on mere experience.

An investment in a data science product or a data analysis project will not only be helpful for business but will also give an upper edge over the entire domain or the field of business. If a business does not harness the power of data science, it can incur huge losses and may affect the credibility of the business.

### Data Science and Industry 4.0

Industry 4.0 is the biggest beneficiary of data science. Data science is no longer a passive partner or some simple domain lying at the back end and working for your entire business. It is an active collaborator to boost industrial growth.

**How Industry 4.0 uses this data?**

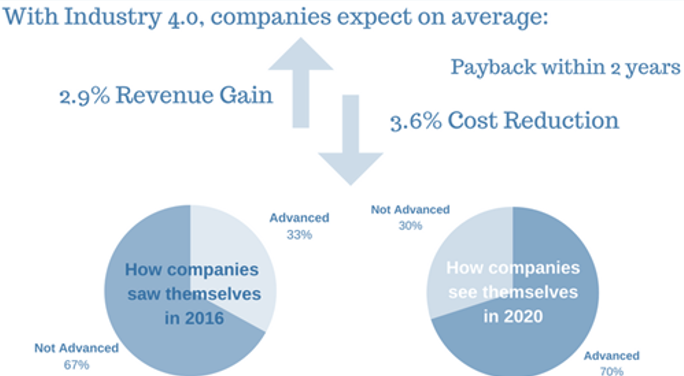


Figure 2: Industry 4.0 Survey conducted by PWC

Industry 4.0 harnesses the power of data science by monitoring, controlling, and optimizing their production. This helps them not only financially but also in the operations. Decision-makers of a company can also benefit from the power of data science by mapping and effectively adapting to the changing needs of the industry.

### Applications of Data Science

* Supply chain and Logistics
* Banking, Financial Services, and Insurance (BFSI) - These domains can identify fraud or a probability of fraud by merely predicting the data patterns.
* Healthcare & Pharmaceuticals
* Bioinformatics
* Development of Smart Cities
* Education
* Energy & Sustainability

**Summary**

Sow your ideas and discover new ways to cultivate them.

Lets a step in the movement towards a knowledge-empowered society.

## What do data scientists do?

In this section, you will read an interview of a data science professional, Mr. Niranjan Dixit. From this interview, you will get an idea about what exactly do data scientists do.

* **Interviewer** – Sir, please give us your brief introduction.
* **Mr Niranjan** – Hello, I am Niranjan Dixit. I am in the field of technology for 20 years. I am interacting with data for a long time now, so I guess you can call me a data scientist.
* **Interviewer** – What drives you as a data science professional?
* **Mr Niranjan** – Interesting question! I turned towards data science quite accidentlly. I was debugging some code in one of my previous projects and we were trying to analyse performance problems. As a regular technologist, instinctively I focussed on a particular piece of code. However, when I actually looked at the data, the problem came from a completely different place. Then, I realised the importance of data and the insights it can give you. That started my data science journey.
* **Interviewer** – What are your daily tasks in a typical workday?
* **Mr Niranjan** – Its nothing different. We are not from Mars; we all belong to Earth. As a data scientist, we typically deal with use cases. Data in itself is just numbers, letters, or characters. What you derive out of it and how you apply it to real-life or business problems makes it more interesting. A typical use case could be end of line testing for manufacturing firms where you need to reduce the amount of testing done or you need to identify problems or to increase productivity. We consider the different problems that businesses face and take the data associated with them and try to derive insights so that we can give some sort of information and help the businesses.
* **Interviewer** – You are a perfect blend of domain expertise as well as technical knowledge. Can you tell us one of your data science success stories?
* **Mr Niranjan** – Sure. A couple of years back, we were working on a telecom project. A network runs on telecom tower sites. These tower sites have typically outgrown metros and have expanded into tier 2, tier 3, and semi-urban, semi-rural areas. Now about 4-6 years back, the power situation was not that good. Power cuts for 8 hours were the norm. This telecom network used to run on diesel because they have a diesel generator there. Consider this is a remote site, where you have to go and fill diesel at each site to ensure it is up and running 24X7. We analysed the usage of diesel across these sites so that we could find a pattern and predict when a site was likely to run out of fuel. In case of a site on hill, where you require 2 hours to reach, we had to inform well in advance so that the diesel can be procured and filled. Therefore, that is one interesting data science story.
* **Interviewer** – To become a successful data science professional, what is your advice to budding data science professionals?
* **Mr Niranjan** – First of all, have a can-do attitude. The sky is the limit. There is no elimination in data science. So, anyone can do data science. If you put your mind to it, study, and focus on the basics and technology behind it, it is not difficult. If I can, then you can.
* **Interviewer** – Thank you for your time. We look forward to interacting with you again.
* **Mr Niranjan** – Thank you so much.

## History of Data Science

In this section, we will see the history and evolution of data science.

### Chronology of Events

### Advancement of Big Data

The dotcom bubble of 1998- 2000 led to the availability of very cheap hard drives. These cheap hard drives were used to store a humongous amount of data. The availability of very cheap hardware in turn led to increased data storage. The accumulation of this data and the availability of cheap hard drives resulted in the rise of Big data. Various calculations and processes were carried out on Big data to get inferences and insights out of them. As the computing speed and processing on Big data increased, it led to the architecture of Big data computing often called Cloud Computing.

### Evolution of Data Science

The advancement of Big data led to the evolution of Big data science. To process Big data on the Cloud, Hadoop software was invented. Though Hadoop was very complicated to work upon, people started developing mass analytical tools. The mass analytical tools were simple interfaces that would run on the Hadoop framework. The major advantage of these mass analytical tools was that they were easy to use and implement. The introduction of mass analytic tools made the analysis of Big data easy. This analysis helped businesses in their faster decision-making and simpler processing of Big data. Thus, evolved the field of data science. It is often defined as the interdisciplinary field of gaining analytical insights out of science, mathematics, scientific algorithms, methods, and processes.

**Summary**

The evolution of data science is a blend of computer statistics and programming languages along with mathematical procedures. It is still an evolving field with continuous innovations that are pushing human frontiers to their extremes.

## Careers in Data Science

Data science is the highest-ranking professions of the 21st century. It is also the most desirable job of the 21st century.

### Data Science Career Opportunities

Given the humongous amount of data generated in the past two years, data science career opportunities are skyrocketing. The shortage of skilled professionals and decision-makers has led to a huge demand for data scientists and data science professionals.

### Data Science Job Roles

Looking at the number of career opportunities available, there are several job roles offered to a data science professional as listed below.

* **Data Scientist**

-A Data Scientist is someone who plays with the science of data. He builds models and derives inferences out of the same.

* **Data Decision Maker**

-A Data Decision Maker is a professional who looks at the data and decides whether to go ahead with a certain decision or a hypothesis or not.

* **Data Engineer**

-A Data Engineer deals with the infrastructure of Big data and data science.

* **Data Architect**

-A Data Architect builds infrastructure as well as security models for data science projects.

* **Data Analyst**

-A Data Analyst is someone who explores data, analyzes data, and derives inferences out of them.

* **Data Analytics Leader**

-A Data Analytics Leader is also a decision-maker who looks at data analysis, its results, and interprets them into data-driven decisions.

* **Data Artist**

-A Data Artist deals with the visualizations of Big data and data science.

* **Statistician**

-A Statistician runs mathematical models and statistical tests on the data.

* **AI or ML Engineer**

-AI is abbreviated as Artificial Intelligence and ML is abbreviated as Machine Learning. AI and machine learning engineers deal with machine learning algorithms and computing.

### Skills required for Data Science

To become a specialized data science professional, one must acquire certain skill sets. These are soft skills and technical skills.

* **Soft skills**
* Creative Thinking
* Problem Solving
* Persuasive Communication

-One should be able to communicate the results or the inferences of the data modeling.

* Critical Thinking

-One must be a critical thinker to solve a certain set of problems.

* **Technical Skills**
* Programming Languages

-One must be well versed in programming languages such as R, Python, and Javascript.

* Statistics and Applied Mathematics

-One must be well versed in linear algebra and calculus to become a great data science professional.

* Knowledge of Big data architecture and infrastructure

-One must have knowledge of Big data infrastructure such as Hadoop.

* Databases

-One should have basic knowledge of databases such as SQL, MySQL, and non-SQL databases such as MongoDB.

* Machine Learning

-One must be well versed in different machine learning algorithms.

* Deep Learning

-Deep learning is an advanced level of machine learning, wherein the machine learns on its own just like a human brain.

* Data Visualization

-Data visualization creates the maximum impact on the inferences drawn out of data science. It is a visual representation of inferences using graphs and charts.

**Summary**

A combination of coding skills and knowledge of statistics, along with the ability of critical and creative thinking, is a full-fledged package of a successful Data Science professional. Data scientists are in constant demand and the need will keep rising with the increasing data.

## Business Application in Data Science

### Importance of Data Science in Business

In this section, we will see the importance of Data Science in Business. Below are some of the challenges in different domains to which Data Science is the solution.

* With the advent of computing technologies, data is growing exponentially in humongous amounts. It is going to be difficult to process such humongous amounts of data.
* The analysis of this Big data and drawing inferences out it is going to be difficult.
* Processing this data will require specialised skill sets and toolsets.
* The decision-making process is going to take a lot of time and become more difficult with the advent of growing data.

### Business Applications of Data science

Let us take a look at the applications of data science.

* **Banking, Finance, and Insurance**
* We use Big data for **Risk modeling**, **Assessment,** and **Compliance**. We try to predict the frauds in credit cards and financial institutions using Big data and patterns generated by data.
* We help in **Revenue Management** in the finance sector by using data science and different tools of data science. We can also help in **Revenue Modelling** using data science. It is an important step for the banking and the finance sector since the entire process is end to end automated.
* **Manufacturing sector**
* Data Science helps in the **Predictive and Conditional Maintenance** of machines in the manufacturing sector. You can anticipate when a machine is going to get out of business through predictive analytics.
* **Demand Forecasting** means managing inventory according to the demand. By harnessing the power of data science, you can forecast the supply and inventory management in the near future. You can predict the future by looking at the older models or older patterns in your inventory.
* **Material Management-** You can manage the entire inventory using Big data and data science tools.
* **Health Care**
* Using Data Science, you can perform **Hypothesis Testing** and **Pattern Identification**. We perform analysis to detect various diseases and their implications using different drugs.
* Healthcare also applies data science to **predict the outcome of a specific disease**. It tries to predict the outcome of disease by predicting the variables of the different patients or the patterns generated by different patients.
* **Retail**
* We can use Data Science for **Price Optimization.** Using algorithms that can learn the patterns from existing data, we can determine the best price of products or services. For example, on amazon.com two different people can see different prices based on their search history.
* **Inventory Management**

You can manage the inventory by forecasting the inventory based on the historical patterns of our behaviour

* **Customer Sentiment Analysis**

Data Science is also helping retailers to perform customer sentiment analysis. For example, when you write a review on an e-commerce platform, the retailer tries to predict whether you've liked the product that you have purchased or you are suffering from post-purchase regret.

Let us take a look at different applications other than the above four domains.

* **Sales and marketing**- You can use data science to predict whether a customer is going to buy a certain product or not and then market it in the same way.
* **Human resource management-** Using data science you can predict whether the employee is going to leave the organization or not.
* **Automation-** Data science can help automate certain processes using previous patterns and history.
* **Consumer Analytics-** You can use consumer analytics to detect the sentiment of the customers and thereby predict whether they are going to be a common customer or not.
* **Risk Compliance-** Using data science you can analyze and mitigate the risk that is going to come in the near future.

**Summary**

It is important to harness the power of data science in different domains. Data Science is a multi-disciplinary field that can help us transform businesses. It is not about transforming a business or digitalizing a business but it is about changing an end-to-end process to harvest profits. Data science can add value to your businesses. It not only changes the older business process but also helps you make data-driven decisions. Data-driven decisions are more helpful since they help in keeping up with the competition and increase profits. The informed decisions save you from the errors made by managerial experiences.

# Data Science Methodology

Let us take a brief look at Data Science process. In this section, we will see the basic steps that constitute the data science process.

## Data Science Process

The prerequisites, before taking a look at this exercise please take a look at data preparation.

The data science process helps you to build a solution for the business or the real world problem. It takes you to a more optimized solution built for your resources and customized to the problem. It also helps in managing the predictive workflow. In other words, it helps you to predict the problems that you might face in the near future.

**Benefits of Data Science process**

* Simplicity

-Data Science process is easy to understand and very simple to implement.

* Speed

-If you follow the data science process, the data science project will speed up to a great limit.

* Repeatability

-You can reuse the same data science process for multiple data science projects.

* Flexibility

- In case you want to alter a certain path or you want to alter certain steps in your process, the process is flexible enough for you to get to a better solution.

* Versioning and tracking

-You can track the entire data science process and save the versions for further reference.

* Quality Assurance

-Data science process ensures that your project is quality based.

* Cost control

-It helps with budgeting the process.

## Stages of Data Science Process

Let us take a look at the stages of Data Science process.



As seen in the figure above, the stages are:

1. **Data Collection** - Data for your business problem can be collected from outside or can be taken from the existing databases in your firm.
2. **Raw Data** – The data collected either from an existing source or from a new source is in its raw format. This data needs to be cleaned.
3. **Data Cleaning** - It is also called as Data Pre-processing. It consists of missing value imputation, outlier detection, and so on.
4. **Data Visualization** – To explore this cleaned data, we need to visualize it. With this, we get a summary of the data. This can be done using different types of charts and plots.
5. **Data Modelling -** After exploring the data, we know which algorithm to use. In Data Modelling, we run an algorithm on the data.
6. **Interpret findings** - Running the algorithm gives some performance metrics. Based on the performance metrics, we evaluate the model and predict the findings. From the insights of the data, we might interpret the findings to solve the business problem.
7. **Revisit the process** - After interpreting the findings, we might want to revisit the model for fresh data since the model has to be generated or run on fresh data.

Let’s look at all the steps in detail.

1. **Data Collection**

The collection of the data is the first stage where the relevant data is collected from different sources. It is important to obtain authentic and reliable data since with very high-quality data, the interpretation and exploration become easier. This data can be collected either from external sources or from an existing database for the business problem. We convert the available dataset into a useable format. We transform the data or raw dataset into an XML or a CSV file so that it can be used for data exploration as well as data cleaning.

1. **Data Preprocessing**

Data pre-processing is the first step in data cleansing. This is the most time-consuming phase. Most data scientists spend 75% of their time on data cleaning and data preprocessing. We do data preprocessing or data cleaning to examine the data and understand its features and the data labels. For example, if you have a column for date of birth you can easily extract the age of the person from that column. To conduct data cleaning or data pre-processing phase, domain- level expertise would be required to understand the impact of these features. For example, to perform data cleaning for the retail sector, one needs to be well versed in jargon such as customer data and customer segmentation or target segmentation.

1. **Data Exploration and Visualisation**

In this step, we explore the cleaned data and visualize it. Data exploration and visualization helps us to uncover certain data patterns or characteristics of data that was uncovered previously. One cannot just look at data and uncover certain patterns. We might have to visualize them to see the descriptive statistics of the data. Data exploration is a blend of manual and automated tools. You can do it either manually on Excel or using tools like Tableau or Rubisite. Data visualization gives a crisp and clear view of your data in just one dashboard. This makes processing and running models much simpler than merely looking at datasheet.

1. **Data Modelling**

In Data Modeling, we conduct an in-depth analysis of the data. Though exploration gives us a clear and crisp view of the data, we have to do some pre-processing as well as some modeling on the data to come up with an inference. Data Modelling involves training the data. For this, we take the data set or segregate the dataset to train the model. Then we implement the machine learning models on the trained data set.

1. **Evaluating the Outcome**

Once we run the model on the trained data, we check the performance metrics and evaluate the model. From the output of the model, we identify the business insight. That helps us to identify the outcome as well. We can correlate the same to the data finding by considering the correlation between features. To interpret the output of the model one requires domain expertise. This helps to identify and interpret the business requirement of the data problem. This also helps to make more data-driven business decisions than the ones based on mere experience.

1. **Revisiting the model**

Depending on the nature of the business, we might want to revisit the module since we might receive new data every six months or within a short period of two days. We would want to run the model again on the new fresh data. The frequency of revisiting the model will be directly proportional to the frequency of the new update that the data has.

**Summary**

1. Start with a reasonable objective. Have a strong foundation for your business problem and start working towards it.
2. Understanding the data intuitively is the key to become a great data science professional. You understand the data to extract some features. Of course, domain knowledge will be required for the same.
3. Run different models on the algorithm to constitute proper performance metrics that will predict outcome properly. Also, revisit your model once you get new data.

## Types of Data

## How to analyse data

## Data collection, requirements, understanding, EDA

## Data Preparation

## Data Modelling and Introduction to alogorithms

## Data Visualisation