What is Data Science?

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.

Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

In short, we can say that data science is all about:

* Asking the correct questions and analyzing the raw data.
* Modeling the data using various complex and efficient algorithms.
* Visualizing the data to get a better perspective.
* Understanding the data to make better decisions and finding the final result.

Example:

Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Need for Data Science:

Some years ago, data was less and mostly available in a structured form, which could be easily stored in excel sheets, and processed using BI tools.

But in today's world, data is becoming so vast, i.e., approximately **2.5 quintals bytes** of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single person on earth. Every Company requires data to work, grow, and improve their businesses.

Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science. Following are some main reasons for using data science technology:

* With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
* Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.
* Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
* Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

Data science Jobs:

As per various surveys, data scientist job is becoming the most demanding Job of the 21st century due to increasing demands for data science. Some people also called it "the **hottest job title of the 21st century**". Data scientists are the experts who can use various statistical tools and machine learning algorithms to understand and analyze the data.

The average salary range for data scientist will be approximately **$95,000 to $ 165,000 per annum**, and as per different researches, about **11.5 millions** of job will be created by the year **2026**.

Types of Data Science Job

If you learn data science, then you get the opportunity to find the various exciting job roles in this domain. The main job roles are given below:

1. Data Scientist
2. Data Analyst
3. Machine learning expert
4. Data engineer
5. Data Architect
6. Data Administrator
7. Business Analyst
8. Business Intelligence Manager

Below is the explanation of some critical job titles of data science.

**1. Data Analyst:**

Data analyst is an individual, who performs mining of huge amount of data, models the data, looks for patterns, relationship, trends, and so on. At the end of the day, he comes up with visualization and reporting for analyzing the data for decision making and problem-solving process.

**Skill required:** For becoming a data analyst, you must get a good background in **mathematics, business intelligence, data mining**, and basic knowledge of **statistics**. You should also be familiar with some computer languages and tools such as **MATLAB, Python, SQL, Hive, Pig, Excel, SAS, R, JS, Spark**, etc.

**2. Machine Learning Expert:**

The machine learning expert is the one who works with various machine learning algorithms used in data science such as **regression, clustering, classification, decision tree, random forest**, etc.

**Skill Required:** Computer programming languages such as Python, C++, R, Java, and Hadoop. You should also have an understanding of various algorithms, problem-solving analytical skill, probability, and statistics.

**3. Data Engineer:**

A data engineer works with massive amount of data and responsible for building and maintaining the data architecture of a data science project. Data engineer also works for the creation of data set processes used in modeling, mining, acquisition, and verification.

**Skill required:** Data engineer must have depth knowledge of **SQL, MongoDB, Cassandra, HBase, Apache Spark, Hive, MapReduce**, with language knowledge of **Python, C/C++, Java, Perl**, etc.

**4. Data Scientist:**

A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc.

**Skill required:** To become a data scientist, one should have technical language skills such as **R, SAS, SQL, Python, Hive, Pig, Apache spark, MATLAB**. Data scientists must have an understanding of Statistics, Mathematics, visualization, and communication skills.

Prerequisite for Data Science

Non-Technical Prerequisite:

* **Curiosity:** To learn data science, one must have curiosities. When you have curiosity and ask various questions, then you can understand the business problem easily.
* **Critical Thinking:** It is also required for a data scientist so that you can find multiple new ways to solve the problem with efficiency.
* **Communication skills:** Communication skills are most important for a data scientist because after solving a business problem, you need to communicate it with the team.

Technical Prerequisite:

* **Machine learning:** To understand data science, one needs to understand the concept of machine learning. Data science uses machine learning algorithms to solve various problems.
* **Mathematical modeling:** Mathematical modeling is required to make fast mathematical calculations and predictions from the available data.
* **Statistics:** Basic understanding of statistics is required, such as mean, median, or standard deviation. It is needed to extract knowledge and obtain better results from the data.
* **Computer programming:** For data science, knowledge of at least one programming language is required. R, Python, Spark are some required computer programming languages for data science.
* **Databases:** The depth understanding of Databases such as SQL, is essential for data science to get the data and to work with data.

Difference between BI and Data Science

BI stands for business intelligence, which is also used for data analysis of business information: Below are some differences between BI and Data sciences:

|  |  |  |
| --- | --- | --- |
| **Criterion** | **Business intelligence** | **Data science** |
| **Data Source** | Business intelligence deals with structured data, e.g., data warehouse. | Data science deals with structured and unstructured data, e.g., weblogs, feedback, etc. |
| **Method** | Analytical(historical data) | Scientific(goes deeper to know the reason for the data report) |
| **Skills** | Statistics and Visualization are the two skills required for business intelligence. | Statistics, Visualization, and Machine learning are the required skills for data science. |
| **Focus** | Business intelligence focuses on both Past and present data | Data science focuses on past data, present data, and also future predictions. |

Data Science Components:

The main components of Data Science are given below:

**1. Statistics:** Statistics is one of the most important components of data science. Statistics is a way to collect and analyze the numerical data in a large amount and finding meaningful insights from it.

**2. Domain Expertise:** In data science, domain expertise binds data science together. Domain expertise means specialized knowledge or skills of a particular area. In data science, there are various areas for which we need domain experts.

**3. Data engineering:** Data engineering is a part of data science, which involves acquiring, storing, retrieving, and transforming the data. Data engineering also includes metadata (data about data) to the data.

**4. Visualization:** Data visualization is meant by representing data in a visual context so that people can easily understand the significance of data. Data visualization makes it easy to access the huge amount of data in visuals.

**5. Advanced computing:** Heavy lifting of data science is advanced computing. Advanced computing involves designing, writing, debugging, and maintaining the source code of computer programs.

**6. Mathematics:** Mathematics is the critical part of data science. Mathematics involves the study of quantity, structure, space, and changes. For a data scientist, knowledge of good mathematics is essential.

**7. Machine learning:** Machine learning is backbone of data science. Machine learning is all about to provide training to a machine so that it can act as a human brain. In data science, we use various machine learning algorithms to solve the problems.

Tools for Data Science

Following are some tools required for data science:

* **Data Analysis tools:** R, Python, Statistics, SAS, Jupyter, R Studio, MATLAB, Excel, RapidMiner.
* **Data Warehousing:** ETL, SQL, Hadoop, Informatica/Talend, AWS Redshift
* **Data Visualization tools:** R, Jupyter, Tableau, Cognos.
* **Machine learning tools:** Spark, Mahout, Azure ML studio.

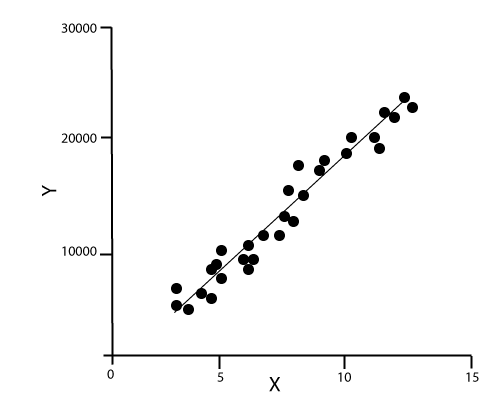
Machine learning in Data Science

To become a data scientist, one should also be aware of machine learning and its algorithms, as in data science, there are various machine learning algorithms which are broadly being used. Following are the name of some machine learning algorithms used in data science:

* Regression
* Decision tree
* Clustering
* Principal component analysis
* Support vector machines
* Naive Bayes
* Artificial neural network
* Apriori

We will provide you some brief introduction for few of the important algorithms here,

**1. Linear Regression Algorithm:** Linear regression is the most popular machine learning algorithm based on supervised learning. This algorithm work on regression, which is a method of modeling target values based on independent variables. It represents the form of the linear equation, which has a relationship between the set of inputs and predictive output. This algorithm is mostly used in forecasting and predictions. Since it shows the linear relationship between input and output variable, hence it is called linear regression.



The below equation can describe the relationship between x and y variables:

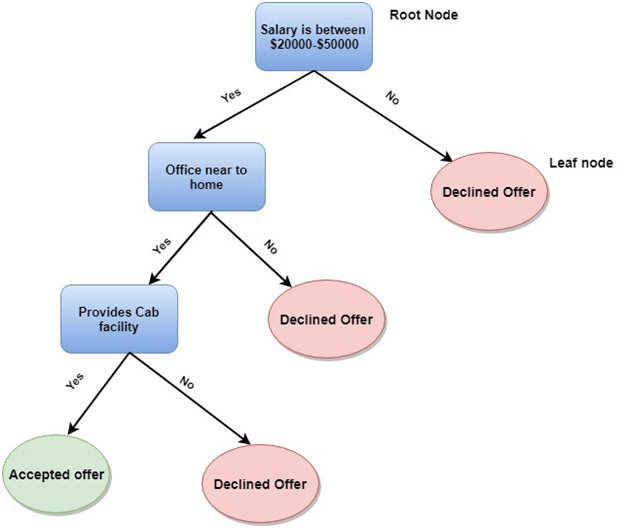
1. Y= mx+c

Where, **y**= Dependent variable  
**X**= independent variable  
**M**= slope  
**C**= intercept.

**2. Decision Tree:** Decision Tree algorithm is another machine learning algorithm, which belongs to the supervised learning algorithm. This is one of the most popular machine learning algorithms. It can be used for both classification and regression problems.

In the decision tree algorithm, we can solve the problem, by using tree representation in which, each node represents a feature, each branch represents a decision, and each leaf represents the outcome.

Following is the example for a Job offer problem:

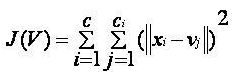


In the decision tree, we start from the root of the tree and compare the values of the root attribute with record attribute. On the basis of this comparison, we follow the branch as per the value and then move to the next node. We continue comparing these values until we reach the leaf node with predicated class value.

**3. K-Means Clustering:** K-means clustering is one of the most popular algorithms of machine learning, which belongs to the unsupervised learning algorithm. It solves the clustering problem.

If we are given a data set of items, with certain features and values, and we need to categorize those set of items into groups, so such type of problems can be solved using k-means clustering algorithm.

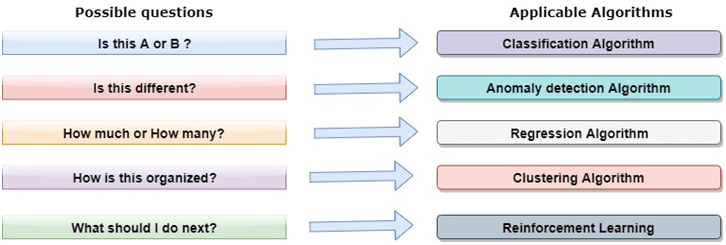
K-means clustering algorithm aims at minimizing an objective function, which known as squared error function, and it is given as:



**Where**, J(V) => Objective function  
'||xi - vj||' => Euclidean distance between xi and vj.  
ci' => Number of data points in ith cluster.  
C => Number of clusters.

How to solve a problem in Data Science using Machine learning algorithms?

Now, let's understand what are the most common types of problems occurred in data science and what is the approach to solving the problems. So in data science, problems are solved using algorithms, and below is the diagram representation for applicable algorithms for possible questions:



**Is this A or B? :**

We can refer to this type of problem which has only two fixed solutions such as Yes or No, 1 or 0, may or may not. And this type of problems can be solved using classification algorithms.

**Is this different? :**

We can refer to this type of question which belongs to various patterns, and we need to find odd from them. Such type of problems can be solved using Anomaly Detection Algorithms.

**How much or how many?**

The other type of problem occurs which ask for numerical values or figures such as what is the time today, what will be the temperature today, can be solved using regression algorithms.

**How is this organized?**

Now if you have a problem which needs to deal with the organization of data, then it can be solved using clustering algorithms.

Clustering algorithm organizes and groups the data based on features, colors, or other common characteristics.

Data Science Lifecycle

**1. Discovery:** The first phase is discovery, which involves asking the right questions. When you start any data science project, you need to determine what are the basic requirements, priorities, and project budget. In this phase, we need to determine all the requirements of the project such as the number of people, technology, time, data, an end goal, and then we can frame the business problem on first hypothesis level.

**2. Data preparation:** Data preparation is also known as Data Munging. In this phase, we need to perform the following tasks:

* Data cleaning
* Data Reduction
* Data integration
* Data transformation,

After performing all the above tasks, we can easily use this data for our further processes.

**3. Model Planning:** In this phase, we need to determine the various methods and techniques to establish the relation between input variables. We will apply Exploratory data analytics(EDA) by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:

* SQL Analysis Services
* R
* SAS
* Python

**4. Model-building:** In this phase, the process of model building starts. We will create datasets for training and testing purpose. We will apply different techniques such as association, classification, and clustering, to build the model.

Following are some common Model building tools:

* SAS Enterprise Miner
* WEKA
* SPCS Modeler
* MATLAB

**5. Operationalize:** In this phase, we will deliver the final reports of the project, along with briefings, code, and technical documents. This phase provides you a clear overview of complete project performance and other components on a small scale before the full deployment.

**6. Communicate results:** In this phase, we will check if we reach the goal, which we have set on the initial phase. We will communicate the findings and final result with the business team.

Applications of Data Science:

* **Image recognition and speech recognition:**  
  Data science is currently using for Image and speech recognition. When you upload an image on Facebook and start getting the suggestion to tag to your friends. This automatic tagging suggestion uses image recognition algorithm, which is part of data science.  
  When you say something using, "Ok Google, Siri, Cortana", etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.
* **Gaming world:**  
  In the gaming world, the use of Machine learning algorithms is increasing day by day. EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.
* **Internet search:**  
  When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc. All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.
* **Transport:**  
  Transport industries also using data science technology to create self-driving cars. With self-driving cars, it will be easy to reduce the number of road accidents.
* **Healthcare:**  
  In the healthcare sector, data science is providing lots of benefits. Data science is being used for tumor detection, drug discovery, medical image analysis, virtual medical bots, etc.
* **Recommendation systems:**  
  Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.
* **Risk detection:**  
  Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this can be rescued.  
  Most of the finance companies are looking for the data scientist to avoid risk and any type of losses with an increase in customer satisfaction.

**Data Analysis**

Raw data aggregated is data that is not oriented. It requires a thoughtful understanding as well as the appropriate questions in order to create sense out of it. Many insights fail to analyse data completely and become difficult for the stakeholders' comprehension. Therefore, it becomes necessary for a data analyst to define and understand data with the right set of initial questions and a standardized workflow for the different types of analysis he needs to perform.

The following words are from Jeff Leek's fascinating book "The Elements of Data Analytic Style," which broadly categorizes various analysis phases based on the type of question and the outcome expected to be achieved for the particular business need.

## Descriptive Data Analysis

The name suggests that this kind of analysis offers basic "descriptions" or summaries about the raw data set accumulated and the observations added to the same.

They can be both visual and quantitative, and the data can be depicted using statistics and simple graphs. This summary does not require any further analysis and is utilized as a summary to make sense of the information.

**Example:** Data on segregation of students enrolled in the same course at college:

The data could be split into various categories such as numbers, gender, residency age, race, and so on. The information summarizes or groups the data into a fixed set that describes all students and the specific information. It doesn't suggest anything and only provides specifics. Thus, it is a type of descriptive analytics.

## Exploratory Data Analysis

Analysis of descriptive data output that is further studied for discoveries patterns, trends, correlations, or inter-relations among different areas of the data in order to develop an interpretation, an idea, or hypotheses. This is the foundation of Exploratory Data Analysis (EDA).

In essence, it's expanding over the description data sets and trying to provide a comprehensive overview of the data. According to Dianne Cook, as well as Deborah F. Swayne rightly refer to in their book, "(EDA is) a 'play-in-the-sand' to allow us to find the unexpected and come to some understanding of our data."

The main focus isn't always the result of the problem statement; rather, to look at the various elements of data in the first place in order to more intimately.

**Example:** A typical EDA application studies the behaviour of traffic patterns in cities around the world. Although the data gathered may vary in terms of its nature, various surprising discoveries may be discovered like the frequency of accidents that occur at traffic signals, the amount of pollution that is produced on a daily basis because of exhaust emissions from vehicles, and even the rates of traffic congestion in a week. The outcome of the real issue isn't always determined by these findings. The information gathered alongside other data may be helpful to determine the result.

## Inferential/Quantified Data Analysis

The distinction between inferential and exploratory analysis could be identified by determining if the analysis offers consistent information across various samples and the ones in the present.

**Example:** Calculating the mean of marks earned by students taking an exam against the difficulty index for 100 students can give valuable information on the students of 100.

This data can assist in understanding the quality of the connection between these two dimensions when studying student performance on exams. Although it's impossible to know the reasons for these relationships, there is a way to determine the significance of a certain connection in determining inferential results.

## Predictive Data Analysis

The predictive analysis predicts the outcomes that could be expected from a small subset of data from the initial population set. This method of predicting new information is mostly built on quantifiable metrics from the existing data set.

Predictive analysis is not able to quantify the relationship between two dimensions as the inferential statistical method. Rather it uses probabilities that they share to predict possible outcomes in the future.

**Example:** Examining the influence and popularity of the nominees running for election to determine the outcome of that election.

In this case, we can determine the likelihood of the success of the candidate based on data about issues he discusses as well as his conservative and liberal views, information on his popularity in the state of his residence and so on. While we can estimate a potential outcome based on these data, however, we can't predict the outcome accurately.

## Causal Data Analysis

Making modifications to one dimension or measurement to create a conclusive version of a different dimension is the foundation of causal analysis. It is designed to determine the extent and direction the measurement takes in contrast to the previous two. It is a predictive analysis and an inferential one.

**Example:** A randomized clinical trial to determine whether faecal transfer decreases the incidence of infections caused by Clostridium di-facile.

Patients in this research were randomly assigned to receive a faecal transfer along with standard care or regular treatment. Based on the results, the researchers found an unambiguous relationship between the outcomes of infections and transplants. Therefore, the study of the causality of patients produced an exact average outcome from raw data.

## Mechanistic Data Analysis

Although causal data provides an accurate average result, the aim isn't just to comprehend that there's an impact of the inferences derived from data but also to understand how the effect is affecting the outcome.

**An example:** Mechanistic analysis that examines the way in which wing design influences the flow of air around a wing, which results in less drag. In the absence of any engineering expertise, mechanical analysis of data is extremely difficult and is rarely done.

## Conclusion

As we can see, harnessing big-data analytics can bring huge benefits to companies, providing the context of data to tell an even more comprehensive story. By converting complex data sets into actionable intelligence, stakeholders can make better business decisions. If we know how to make big data accessible to our clients, the value of our service is now ten times greater.

# **Life Cycle Phases of Data Analytics**

In this tutorial, we're going to talk about the different phases of the life cycle of data analytics, in which we will go over different life cycle phases and then go over them in detail.

## Life Cycle of Data Analytics

The Data analytics lifecycle was designed to address Big Data problems and data science projects. The process is repeated to show the real projects. To address the specific demands for conducting analysis on Big Data, the step-by-step methodology is required to plan the various tasks associated with the acquisition, processing, analysis, and recycling of data.

### Phase 1: Discovery -

* The data science team is trained and researches the issue.
* Create context and gain understanding.
* Learn about the data sources that are needed and accessible to the project.
* The team comes up with an initial hypothesis, which can be later confirmed with evidence.

### Phase 2: Data Preparation -

* Methods to investigate the possibilities of pre-processing, analysing, and preparing data before analysis and modelling.
* It is required to have an analytic sandbox. The team performs, loads, and transforms to bring information to the data sandbox.
* Data preparation tasks can be repeated and not in a predetermined sequence.
* Some of the tools used commonly for this process include - Hadoop, Alpine Miner, Open Refine, etc.

### Phase 3: Model Planning -

* The team studies data to discover the connections between variables. Later, it selects the most significant variables as well as the most effective models.
* In this phase, the data science teams create data sets that can be used for training for testing, production, and training goals.
* The team builds and implements models based on the work completed in the modelling planning phase.
* Some of the tools used commonly for this stage are MATLAB and STASTICA.

### Phase 4: Model Building -

* The team creates datasets for training, testing as well as production use.
* The team is also evaluating whether its current tools are sufficient to run the models or if they require an even more robust environment to run models.
* Tools that are free or open-source or free tools Rand PL/R, Octave, WEKA.
* Commercial tools - MATLAB, STASTICA.

### Phase 5: Communication Results -

* Following the execution of the model, team members will need to evaluate the outcomes of the model to establish criteria for the success or failure of the model.
* The team is considering how best to present findings and outcomes to the various members of the team and other stakeholders while taking into consideration cautionary tales and assumptions.
* The team should determine the most important findings, quantify their value to the business and create a narrative to present findings and summarize them to all stakeholders.

### Phase 6: Operationalize -

* The team distributes the benefits of the project to a wider audience. It sets up a pilot project that will deploy the work in a controlled manner prior to expanding the project to the entire enterprise of users.
* This technique allows the team to gain insight into the performance and constraints related to the model within a production setting at a small scale and then make necessary adjustments before full deployment.
* The team produces the last reports, presentations, and codes.