**What is Data Science?**

[ IT extract the datasets from past datasets ]

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.

**Data Science** is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data. The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and [big data](https://www.guru99.com/what-is-big-data.html).

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.

Extract meaning full data from past data/present data by using maths, statical concepts, programming skills and algorithms.

## Why Data Science?

Here, are significant advantages of using Data Analytics Technology:

* Data is the oil for today's world. With the right tools, technologies, algorithms, we can use data and convert it into a distinctive business advantage
* Data Science can help you to detect fraud using advanced machine learning algorithms
* It helps you to prevent any significant monetary losses
* Allows to build intelligence ability in machines
* You can perform sentiment analysis to gauge customer brand loyalty
* It enables you to take better and faster decisions
* Helps you to recommend the right product to the right customer to enhance your business

## Data Science Components

### Statistics:

### Statistics is the most critical unit of Data Science basics. It is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.

### Visualization:

### Visualization technique helps you to access huge amounts of data in easy to understand and digestible visuals.

### Machine Learning:

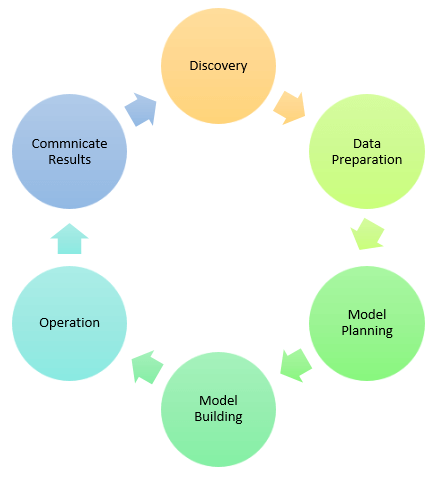
### Machine Learning explores the building and study of algorithms which learn to make predictions about unforeseen/future data.

### Deep Learning:

### Deep Learning method is new machine learning research where the algorithm selects the analysis model to follow.

## Data Science Process

Now in this Data Science Tutorial, we will learn the Data Science Process:



### 1. Discovery:

Discovery step involves acquiring data from all the identified internal & external sources which helps you to answer the business question.

The data can be:

* Logs from webservers
* Data gathered from social media
* Census datasets
* Data streamed from online sources using APIs

### 2. Preparation:

Data can have lots of inconsistencies like missing value, blank columns, incorrect data format which needs to be cleaned. You need to process, explore, and condition data before modeling. The cleaner your data, the better are your predictions.

### 3. Model Planning:

In this stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and [visualization tools](https://www.guru99.com/best-data-visualization-tools.html). SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

### 4. Model Building:

In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the "testing" dataset.

### 5. Operationalize:

In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.

### 6. Communicate Results

In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

## # Applications of Data Science

Now in this Data Science Tutorial, we will learn about Applications of Data Science:

### Internet Search:

Google search use Data science technology to search a specific result within a fraction of a second

### Recommendation Systems:

To create a recommendation system. Example, "suggested friends" on Facebook or suggested videos" on YouTube, everything is done with the help of Data Science.

### Image & Speech Recognition:

Speech recognizes system like Siri, Google assistant, Alexa runs on the technique of Data science. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

### Gaming world:

EA Sports, Sony, Nintendo, are using Data science technology. This enhances your gaming experience. Games are now developed using Machine Learning technique. It can update itself when you move to higher levels.

### Online Price Comparison:

PriceRunner, Junglee, Shopzilla work on the Data science mechanism. Here, data is fetched from the relevant websites using APIs.

**# Challenges of Data science Technology**

* High variety of information & data is required for accurate analysis
* Not adequate data science talent pool available
* Management does not provide financial support for a data science team
* Unavailability of/difficult access to data
* Data Science results not effectively used by business decision makers
* Explaining data science to others is difficult
* Privacy issues
* Lack of significant domain expert
* If an organization is very small, they can't have a Data Science team

Hypothesis

A statistical **hypothesis** is an explanation about the relationship between data populations

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.

**. 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

### Q1. What are the differences between supervised and unsupervised learning?

In a supervised learning model, the algorithm learns on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data.

An unsupervised model, in contrast, provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own.

**Q2. How is logistic regression done?**

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a [logistic function](https://en.wikipedia.org/wiki/Logistic_function), which is the cumulative distribution function of [logistic distribution](https://en.wikipedia.org/wiki/Logistic_distribution).

### ****Q3. What is Selection Bias?****

**Selection bias** occurs when the study population does not represent the target population. It may occur during identification of the study population. ... When a study population is identified, **selection bias** occurs when the criteria used to recruit and enroll patients into study are inherently different.

### ****Q4. What is a confusion matrix?****

\*\*In the field of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and specifically the problem of [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification), a **confusion matrix**, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) one (in [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) it is usually called a matching matrix).

\*\*A **confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The **confusion matrix** itself is relatively simple to understand, but the related terminology can be **confusing**.

### ****#Q. How to combat Overfitting and Underfitting?****

### ****Q5. What is Cluster Sampling?****

**Cluster sampling** is a [sampling](https://en.wikipedia.org/wiki/Sampling_(statistics)) plan used when mutually homogeneous yet internally heterogeneous groupings are evident in a [statistical population](https://en.wikipedia.org/wiki/Statistical_population).

In this sampling plan, the total population is divided into these groups (known as clusters) and a [simple random sample](https://en.wikipedia.org/wiki/Simple_random_sample) of the groups is selected. The elements in each cluster are then sampled.