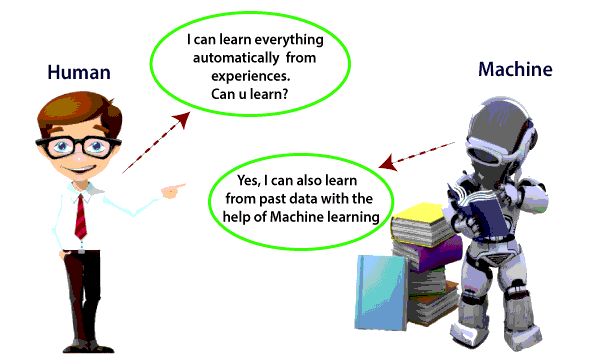
Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**. Currently, it is being used for various tasks such as **image recognition**, **speech recognition**, **email filtering**, **Facebook auto-tagging**, **recommender system**, and many more.

What is Machine Learning

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of **Machine Learning**.



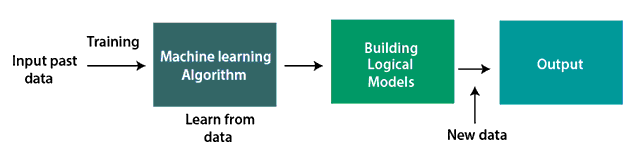
Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by **Arthur Samuel** in **1959**. We can define it in a summarized way as:

Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

How does Machine Learning work

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

The below block diagram explains the working of Machine Learning algorithm:



Features of Machine Learning:

* Machine learning uses data to detect various patterns in a given dataset.
* It can learn from past data and improve automatically.
* It is a data-driven technology.
* Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Need for Machine Learning

It is capable of doing tasks that are too complex for a person to implement directly.

The importance of machine learning can be easily understood by its uses cases, Currently, machine learning is used in **self-driving cars**, **cyber fraud detection**, **face recognition**, and **friend suggestion by Facebook**, etc.

Various top companies such as Netflix and Amazon have build machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

**Following are some key points which show the importance of Machine Learning:**

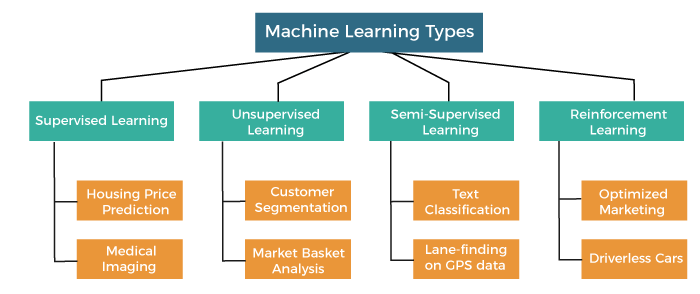
* Rapid increment in the production of data
* Solving complex problems, which are difficult for a human
* Decision making in various sector including finance
* Finding hidden patterns and extracting useful information from data.

Types of Machine Learning

**Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions**. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.

machine learning is divided into mainly four types, which are:

* Supervised Machine Learning
* Unsupervised Machine Learning
* Semi-Supervised Machine Learning
* Reinforcement Learning



In this topic, we will provide a detailed description of the types of Machine Learning along with their respective algorithms:

1. Supervised Machine Learning

As its name suggests, [Supervised machine learning](https://www.javatpoint.com/supervised-machine-learning) is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the **shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc.** After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how the machine identifies the objects in Supervised Learning.

**The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).** Some real-world applications of supervised learning are **Risk Assessment, Fraud Detection, Spam filtering,** etc.

Categories of Supervised Machine Learning

Supervised machine learning can be classified into two types of problems, which are given below:

* **Classification**
* **Regression**

a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "**Yes" or No, Male or Female, Red or Blue, etc**. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

Some popular classification algorithms are given below:

* **Random Forest Algorithm**
* **Decision Tree Algorithm**
* **Logistic Regression Algorithm**
* **Support Vector Machine Algorithm**

b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

* **Simple Linear Regression Algorithm**
* **Multivariate Regression Algorithm**
* **Decision Tree Algorithm**
* **Lasso Regression**

Advantages and Disadvantages of Supervised Learning

**Advantages:**

* Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
* These algorithms are helpful in predicting the output on the basis of prior experience.

**Disadvantages:**

* These algorithms are not able to solve complex tasks.
* It may predict the wrong output if the test data is different from the training data.
* It requires lots of computational time to train the algorithm.

Applications of Supervised Learning

Some common applications of Supervised Learning are given below:

* **Image Segmentation:**  
  Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
* **Medical Diagnosis:**  
  Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.
* **Fraud Detection -** Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
* **Spam detection -** In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
* **Speech Recognition -** Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

2. Unsupervised Machine Learning

[Unsupervised learnin](https://www.javatpoint.com/unsupervised-machine-learning)g is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

**The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.** Machines are instructed to find the hidden patterns from the input dataset.

Let's take an example to understand it more preciously; suppose there is a basket of fruit images, and we input it into the machine learning model. The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects.

So, now the machine will discover its patterns and differences, such as colour difference, shape difference, and predict the output when it is tested with the test dataset.

Categories of Unsupervised Machine Learning

Unsupervised Learning can be further classified into two types, which are given below:

* **Clustering**
* **Association**

1) Clustering

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Some of the popular clustering algorithms are given below:

* **K-Means Clustering algorithm**
* **Mean-shift algorithm**
* **DBSCAN Algorithm**
* **Principal Component Analysis**
* **Independent Component Analysis**

2) Association

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in **Market Basket analysis, Web usage mining, continuous production**, etc.

Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

Advantages and Disadvantages of Unsupervised Learning Algorithm

**Advantages:**

* These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
* Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

**Disadvantages:**

* The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.
* Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

Applications of Unsupervised Learning

* **Network Analysis:** Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.
* **Recommendation Systems:** Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.
* **Anomaly Detection:** Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.
* **Singular Value Decomposition:** Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location.

3. Semi-Supervised Learning

**Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning**. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

**A**lthough Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.

**To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced**. The main aim of [semi-supervised learning](https://www.javatpoint.com/semi-supervised-learning) is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.

We can imagine these algorithms with an example. Supervised learning is where a student is under the supervision of an instructor at home and college. Further, if that student is self-analysing the same concept without any help from the instructor, it comes under unsupervised learning. Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of an instructor at college.

Advantages and disadvantages of Semi-supervised Learning

**Advantages:**

* It is simple and easy to understand the algorithm.
* It is highly efficient.
* It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

**Disadvantages:**

* Iterations results may not be stable.
* We cannot apply these algorithms to network-level data.
* Accuracy is low.

4. Reinforcement Learning

**Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.** Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

The [reinforcement learning](https://www.javatpoint.com/reinforcement-learning) process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.

Due to its way of working, reinforcement learning is employed in different fields such as **Game theory, Operation Research, Information theory, multi-agent systems.**

A reinforcement learning problem can be formalized using **Markov Decision Process(MDP).** In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.

Categories of Reinforcement Learning

Reinforcement learning is categorized mainly into two types of methods/algorithms:

* **Positive Reinforcement Learning:** Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.
* **Negative Reinforcement Learning:** Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

Real-world Use cases of Reinforcement Learning

* **Video Games:** RL algorithms are much popular in gaming applications. It is used to gain super-human performance. Some popular games that use RL algorithms are **AlphaGO** and **AlphaGO Zero**.
* **Resource Management:** The "Resource Management with Deep Reinforcement Learning" paper showed that how to use RL in computer to automatically learn and schedule resources to wait for different jobs in order to minimize average job slowdown.
* **Robotics:**  
  RL is widely being used in Robotics applications. Robots are used in the industrial and manufacturing area, and these robots are made more powerful with reinforcement learning. There are different industries that have their vision of building intelligent robots using AI and Machine learning technology.
* **Text Mining:** Text-mining, one of the great applications of NLP, is now being implemented with the help of Reinforcement Learning by Salesforce company.

**Advantages and Disadvantages of Reinforcement Learning**

**Advantages**

* It helps in solving complex real-world problems which are difficult to be solved by general techniques.
* The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
* Helps in achieving long term results.

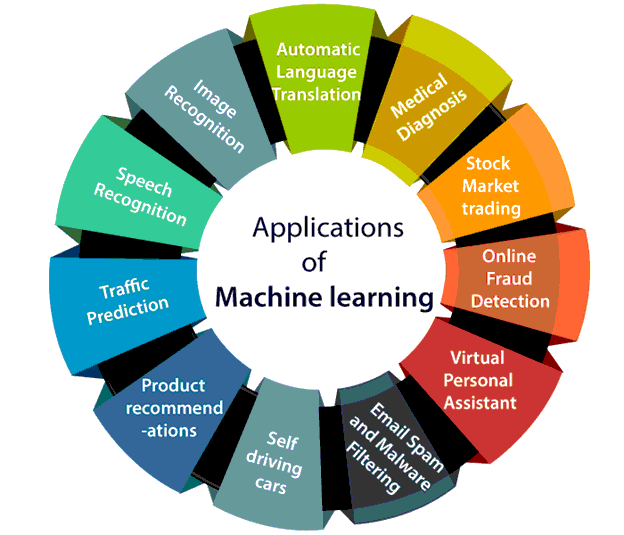
**Disadvantage**

* RL algorithms are not preferred for simple problems.
* RL algorithms require huge data and computations.
* Too much reinforcement learning can lead to an overload of states which can weaken the results.

The curse of dimensionality limits reinforcement learning for real physical systems.

* Applications of Machine learning

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications of Machine Learning:



1. Image Recognition:

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's **face detection** and **recognition algorithm**.

It is based on the Facebook project named "**Deep Face**," which is responsible for face recognition and person identification in the picture.

2. Speech Recognition

While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, machine learning algorithms are widely used by various applications of speech recognition. **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

* **Real Time location** of the vehicle form Google Map app and sensors
* **Average time has taken** on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

4. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. Self-driving cars:

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

6. Email Spam and Malware Filtering:

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:

* Content Filter
* Header filter
* General blacklists filter
* Rules-based filters
* Permission filters

Some machine learning algorithms such as **Multi-Layer Perceptron**, **Decision tree**, and **Naïve Bayes classifier** are used for email spam filtering and malware detection.

7. Virtual Personal Assistant:

We have various virtual personal assistants such as **Google assistant**, **Alexa**, **Cortana**, **Siri**. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.

These virtual assistants use machine learning algorithms as an important part.

These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

8. Online Fraud Detection:

Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as **fake accounts**, **fake ids**, and **steal money** in the middle of a transaction. So to detect this, **Feed Forward Neural network** helps us by checking whether it is a genuine transaction or a fraud transaction.

For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

9. Stock Market trading:

Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's **long short term memory neural network** is used for the prediction of stock market trends.

10. Medical Diagnosis:

In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.

It helps in finding brain tumors and other brain-related diseases easily.

11. Automatic Language Translation:

Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.

* Challenges in Machine Learning

A few of challenges in ML are:

1. Not enough training data :

Let’s say for a child, to make him learn what an apple is, all it takes for you to point to an apple and say apple repeatedly. Now the child can recognize all sorts of apples.

Machine learning is still not up to that level yet; it takes a lot of data for most of the algorithms to function properly. For a simple task, it needs thousands of examples to make something out of it, and for advanced tasks like image or speech recognition, it may need lakhs(millions) of examples.

2. Poor Quality of data:

Obviously, if your training data has lots of errors, outliers, and noise, it will make it impossible for your machine learning model to detect a proper underlying pattern. Hence, it will not perform well.

3. Irrelevant Features:

“Garbage in, garbage out (GIGO).” Training data must always contain more relevant and less to none irrelevant features**.**

4. Nonrepresentative training data:

To make sure that our model generalizes well, we have to make sure that our training data should be representative of the new cases that we want to generalize to.

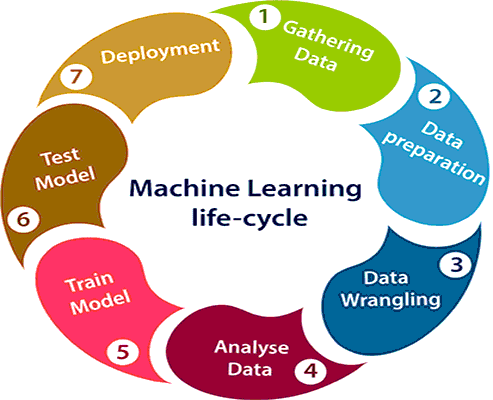
Ex: we assume that youtube’s search engine is providing representative data but in reality, the search will be biased towards popular artists and maybe even the artists that are popular in your location(if you live in India you will be getting the music of Arijit Singh, Sonu Nigam or etc).

5. Overfitting and Underfitting :

Overfitting happens when our model is too complex. Things which we can do to overcome this problem:

* Simplify the model by selecting one with fewer parameters.
* By reducing the number of attributes in training data.
* Constraining the model.
* Gather more training data.
* Reduce the noise.
* Steps involved in Building a ML model:

Machine learning life cycle involves seven major steps, which are given below:



1. Gathering Data:

Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output

2. Data preparation

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

In this step, first, we put all data together, and then randomize the ordering of data.

3. Data Wrangling

Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.

4. Data Analysis

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.

Hence, in this step, we take the data and use machine learning algorithms to build the model.

5. Train Model

Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.

We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

6. Test Model

Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.

Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7. Deployment

The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.

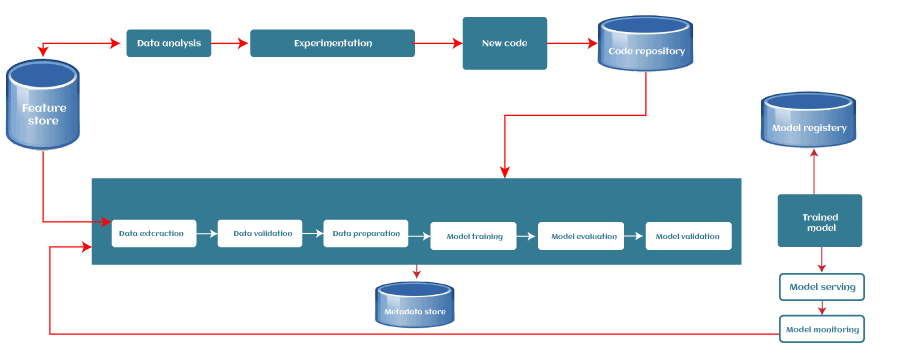
* What is Machine Learning Pipeline?

A Machine Learning pipeline is a process of automating the workflow of a complete machine learning task.

It means that in the pipeline, each step is designed as an independent module, and all these modules are tied together to get the final result.

Machine Learning Pipeline Steps

A typical ML pipeline includes the following stages:



1. Data Ingestion

Each ML pipeline starts with the Data ingestion step. In this step, the data is processed into a well-organized format, which could be suitable to apply for further steps.

2. Data Validation

The next step is data validation, which is required to perform before training a new model. Data validation focuses on statistics of the new data, e.g., range, number of categories, distribution of categories, etc. In this step, data scientists can detect if any anomaly present in the data. There are various data validation tools that enable us to compare different datasets to detect anomalies.

3. Data Pre-processing

The pre-processing step involves preparing the raw data and making it suitable for the ML model. The process includes different sub-steps, such as Data cleaning, feature scaling, etc. The product or output of the data pre-processing step becomes the final dataset that can be used for model training and testing. There are different tools in ML for data pre-processing that can range from simple Python scripts to graph models.

4. Model Training & Tuning

The model training step is the core of each ML pipeline. In this step, the model is trained to take the input (pre-processed dataset) and predicts an output with the highest possible accuracy.

5. Model Analysis

An in-depth analysis of the model's performance is crucial for the final version of the model. The in-depth analysis includes calculating other metrics such as precision, recall etc. This will also help us in determining the dependency of the model on features used in training and explore how the model's predictions would change if we altered the features of a single training example.

6. Model Versioning

The model versioning step keeps track of which model, set of hyperparameters, and datasets have been selected as the next version to be deployed. For various situations, there could occur a significant difference in model performance just by applying more/better training data and without changing any model parameter. Hence, it is important to document all inputs into a new model version and track them.

7. Model Deployment

After training and analyzing the model, it's time to deploy the model. An ML model can be deployed in three ways, which are:

* Using the Model server,
* In a Browser
* On Edge device

The common way to deploy the model is using a model server. Model servers allow to host multiple versions simultaneously, which helps to run tests on models and can provide valuable feedback for model improvement.

8. Feedback Loop

Each pipeline forms a closed-loop to provide feedback. With this close loop, data scientists can determine the effectiveness and performance of the deployed models. This step could be automated or manual depending on the requirement.

* 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.

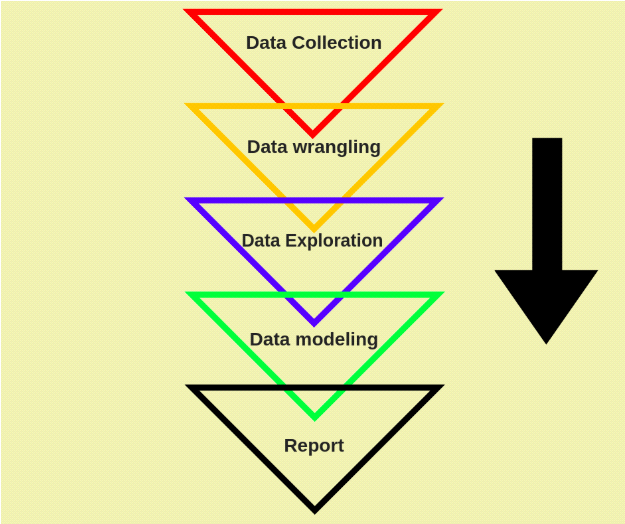
Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

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.

* **How data science works**?

The chart below shows the different steps that are part of a data scientist’s workflow.



1. Data collection:

Data collection is the first step. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output

**Step 2:** **Scrubbing or cleaning the Data**

This is an important step because before you are able to read the data, you must make sure it is in a perfectly readable state, without any mistakes, with no missing values or wrong values.

**Step 3:** **Exploratory Data Analytics**

Now that your Data is clean and readable, it’s time to get to the real work – Analyzing the data. This is done by visualizing the data in various ways and identifying patterns

 In order to be able to analyse the data, you must have high attention to detail to identify if anything is out of place.

**Step 4: Modeling or Machine Learning**

Machine Learning is an application of Artificial Intelligence, in which, a machine can follow commands and rules (algorithms) and come up with predictive solutions without any human supervision.

**Step 5: Interpreting**

This is the final step, in which you uncover your findings and present them to the organization. The most important skill in this would be your ability to explain your results.

* **Major Applications of Data Science**

**1. In Search Engines**

The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.

**2. In Transport**

Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

**For Example,**In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what is the speed limit in Highway, Busy Streets, Narrow Roads, etc. And how to handle different situations while driving etc.

**3. In Finance**

Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses.

**For Example,**In Stock Market, Data Science is the main part. In the Stock Market, Data Science is used to examine past behavior with past data and their goal is to examine the future outcome. Data is analyzed in such a way that it makes it possible to predict future stock prices over a set timetable.

**4. In E-Commerce**

E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.

**For Example,**When we search for something on the E-commerce websites we get suggestions similar to choices according to our past data and also we get recommendations according to most buy the product, most rated, most searched, etc. This is all done with the help of Data Science.

**5. In Health Care**

In the Healthcare Industry data science act as a boon. Data Science is used for:

* Detecting Tumor.
* Drug discoveries.
* Medical Image Analysis.
* Virtual Medical Bots.
* Genetics and Genomics.
* Predictive Modeling for Diagnosis etc.

**6. Image Recognition**

Currently, Data Science is also used in Image Recognition. **For Example,**When we upload our image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science.

**7. Targeting Recommendation**

Targeting Recommendation is the most important application of Data Science. Whatever the user searches on the Internet, he/she will see numerous posts everywhere.

**8. Airline Routing Planning**

With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between .

**9. Medicine and Drug Development**

The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone’s life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors. The algorithms based on data science will forecast how this will react to the human body without lab experiments.

**10. In Delivery Logistics**

Various Logistics companies like DHL, FedEx, etc. make use of Data Science. Data Science helps these companies to find the best route for the Shipment of their Products, the best time suited for delivery, the best mode of transport to reach the destination, etc.

**12. Auto complete**

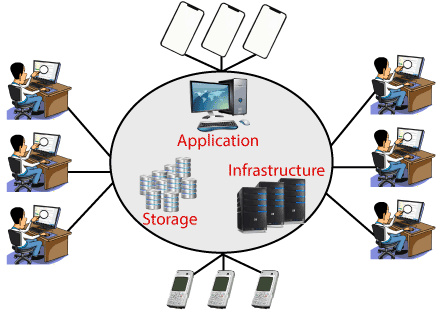
AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line. In Google Mail, when we are writing formal mail to someone so at that time data science concept of Autocomplete feature is used where he/she is an efficient choice to auto-complete the whole line.  Also in Search Engines in social media, in various apps, AutoComplete feature is widely used.

**What’s the difference between data science, artificial intelligence, and machine learning**?

Artificial Intelligence makes a computer act/think like a human. Data science is an AI subset that deals with data methods, scientific analysis, and statistics, all used to gain insight and meaning from data. Machine learning is a subset of AI that teaches computers to learn things from provided data.

* Introduction to Cloud Computing

Cloud Computing is the delivery of computing services such as servers, storage, databases, networking, software, analytics, intelligence, and more, over the Cloud (Internet)



* Cloud Computing is a technology that uses computing resources, including hardware and software, offering services over a network.
* In simple terms, cloud computing stores or accesses data or applications over the internet instead of a hard drive.
* By using the cloud as a technology, services like software, storage, and infrastructure can be provided.
* We can take any required services on rent. The cloud computing services will be charged based on usage.

Benefit of cloud computing:

Screen shot

* Characteristics/Essentials of Cloud Computing

Cloud computing is composed of 5 essential characteristics, viz:

* On-demand Self Service
* Broad Network Access
* Resource Pooling
* Rapid Elasticity
* Measured Service

**On-demand Self Service**

Cloud Computing services are available on-demand and do not require much human interaction.

 For example, for booking a ticket on a travel portal, a passenger gets the flexibility to book his ticket by himself without any human interaction. Right from choosing the flight to preference class, the process is entirely automated and does not require any salesperson in between.

**Broad Network Access**

Cloud computing is accessible from a network, generally over the internet. Similarly, private cloud services can be accessed from anywhere within the enterprise. The services are provided over heterogeneous devices such as mobile phones, laptops, tablets, office computers, etc.

The user can access the existing data on a cloud platform or upload new data on the cloud from anywhere using a device and internet connection.

**Resource Pooling and Multi-tenancy**

Computing resources like networks, servers, storage, applications, and service can be pooled(shared) to serve multiple consumers by securely separating the resources on a logical level.

It is the same as an apartment building where many people share the same building infrastructure but still have the privacy of their apartment.

**Rapid Elasticity and Scalability**

Resource capabilities can be elastically provisioned and released to meet immediate requirements. Similarly, they can be removed or scaled-down when not required. In many cases, this can even happen automatically in response to business demands. This makes sure that the application has the capacity it needs at any point in time.

For example, to meet the demand of the increasing number of passengers, an airline can increase the number of flights for a particular time and stop the flights when the demand goes down.

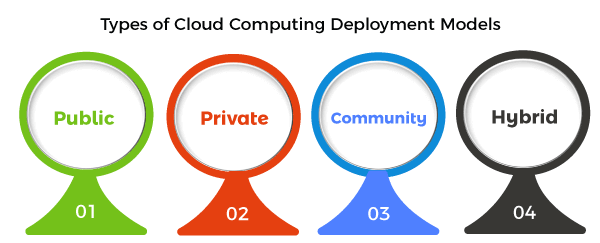
**Measured Service**

The utilization of resources is tracked, monitored, controlled, and reported for each occupant. This gives transparency to both the service provider and the consumer. The cloud system has a metering capability, which is leveraged to monitor billing, use of resources, and pay only for what has been used.

For example: When a passenger is traveling by train, he has to pay only for the distance traveled by him and not for the entire journey that the train take

Most cloud hubs have tens of thousands of servers and storage devices to enable fast loading. It is often possible to choose a geographic area to put the data "closer" to users. Thus, deployment models for cloud computing are categorized based on their location. To know which model would best fit the requirements of your organization, let us first learn about the various types.

* Cloud Deployment Model



* **Public Cloud:**The cloud resources that are owned and operated by a third-party cloud service provider are termed as public clouds. It delivers computing resources such as servers, software, and storage over the internet
* **Private Cloud:**The cloud computing resources that are exclusively used inside a single business or organization are termed as a private cloud. A private cloud may physically be located on the company’s on-site datacentre or hosted by a third-party service provider.
* **Hybrid Cloud:**It is the combination of public and private clouds, which is bounded together by technology that allows data applications to be shared between them. Hybrid cloud provides flexibility and more deployment options to the business.
* Cloud Service Models

There are the following three types of cloud service models -

* [Infrastructure as a Service (IaaS)](https://www.javatpoint.com/cloud-service-models)
* [Platform as a Service (PaaS)](https://www.javatpoint.com/cloud-service-models)
* [Software as a Service (SaaS)](https://www.javatpoint.com/cloud-service-models)
* **Infrastructure as a Service (IaaS):**In IaaS, we can rent IT infrastructures like servers and virtual machines (VMs), storage, networks, operating systems from a cloud service vendor. We can create VM running Windows or Linux and install anything we want on it. Using IaaS, we don’t need to care about the hardware or virtualization software, but other than that, we do have to manage everything else. Using IaaS, we get maximum flexibility, but still, we need to put more effort into maintenance.

**Example:** DigitalOcean, Linode, Amazon Web Services (AWS), Microsoft Azure, Google Compute Engine (GCE), Rackspace, and Cisco Metacloud.

* **Platform as a Service (PaaS):** This service provides an on-demand environment for developing, testing, delivering, and managing software applications. The developer is responsible for the application, and the PaaS vendor provides the ability to deploy and run it. Using PaaS, the flexibility gets reduce, but the management of the environment is taken care of by the cloud vendors.

**Example:** AWS Elastic Beanstalk, Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos, Magento Commerce Cloud, and OpenShift.

* **Software as a Service (SaaS):** It provides a centrally hosted and managed software services to the end-users. It delivers software over the internet, on-demand, and typically on a subscription basis. E.g., Microsoft One Drive, Dropbox, WordPress, Office 365, and Amazon Kindle. SaaS is used to minimize the operational cost to the maximum extent.

Advantages of cloud computing

* **Cost:** It reduces the huge capital costs of buying hardware and software.
* **Speed:** Resources can be accessed in minutes, typically within a few clicks.
* **Scalability:**We can increase or decrease the requirement of resources according to the business requirements.
* **Productivity:**While using cloud computing, we put less operational effort. We do not need to apply patching, as well as no need to maintain hardware and software. So, in this way, the IT team can be more productive and focus on achieving business goals.
* **Reliability:**Backup and recovery of data are less expensive and very fast for business continuity.
* **Security:**Many cloud vendors offer a broad set of policies, technologies, and controls that strengthen our data security.
* Serverless Services

Serverless is a [cloud-native](https://www.redhat.com/en/topics/cloud-native-apps) development model that allows developers to build and run applications without having to manage servers.

There are still servers in serverless, but they are abstracted away from app development. A [cloud provider](https://www.redhat.com/en/topics/cloud-computing/what-are-cloud-providers) handles the routine work of [provisioning](https://www.redhat.com/en/topics/automation/what-is-provisioning), maintaining, and scaling the server [infrastructure](https://www.redhat.com/en/topics/cloud-computing/what-is-it-infrastructure)

The main advantages of the serverless applications are:

* The regular maintenance of any servers is not required. Also, we there is no software or platform to install for it.
* The scaling of the application is very flexible in serverless computing. They can be scaled automatically by adjusting the capacity of the configurations required to run the application. Ex:- Memory, processor requirement etc.
* They have inbuilt availability and fault tolerance.
* You don’t have to pay for idle capacity. For example, there is no charge when your code is not running.
* Serverless computing is more cost-effective than purchasing the servers with required configurations

Some famous platforms for serverless computing:AWS lambda,Microsoft Azur, Google Cloud Platform etc.

* Cloud Service Provider Companies

Cloud Service providers (CSP) offers various services such as **Software as a Service**, **Platform as a service**, **Infrastructure as a service**, **network services**, **business applications**, **mobile applications**, and **infrastructure** in the cloud. The cloud service providers host these services in a data center, and users can access these services through cloud provider companies using an Internet connection.

There are the following Cloud Service Providers Companies -

Amazon Web Services (AWS)

[AWS](https://www.javatpoint.com/aws-tutorial) (Amazon Web Services) is a **secure cloud service platform** provided by **Amazon**. It offers various services such as database storage, computing power, content delivery, Relational Database, Simple Email, Simple Queue, and other functionality to increase the organization's growth.



**Features of AWS**

AWS provides various powerful features for building scalable, cost-effective, enterprise applications. Some important [features of AWS](https://www.javatpoint.com/features-of-aws) is given below-

* AWS is **scalable** because it has an ability to scale the computing resources up or down according to the organization's demand.
* AWS is **cost-effective** as it works on a **pay-as-you-go** pricing model.
* It provides various flexible storage options.
* It offers various **security services** such as infrastructure security, data encryption, monitoring , penetration testing, and DDoS attacks.
* It can efficiently manage and secure Windows workloads.

2. Microsoft Azure

[Microsoft Azure](https://www.javatpoint.com/microsoft-azure) is also known as **Windows Azure**. It supports various operating systems, databases, programming languages, frameworks that allow [IT](https://www.javatpoint.com/it-full-form) professionals to easily build, deploy, and manage applications through a worldwide network. It also allows users to create different groups for related utilities.



Features of Microsoft Azure

* Microsoft Azure provides **scalable**, **flexible**, and **cost-effective**
* It allows developers to quickly manage applications and websites.
* It managed each resource individually.
* It offers a **Content Delivery System (CDS)** for delivering the Images, videos, audios, and applications.

3. Google Cloud Platform (GCP)

Google cloud platform is a product of **Google**. It consists of a set of physical devices, such as computers, hard disk drives, and virtual machines. It also helps organizations to simplify the migration process.



Features of Google Cloud

* It provides various services related to **networking**, including Google Virtual Private It offers various **scalable** and **high-performance**
* GCP provides various **serverless services** such as Messaging, Data Warehouse, Database, Compute, Storage, Data Processing, and Machine learning (ML)

4. IBM Cloud Services

IBM Cloud is an open-source, faster, and more reliable platform. It is built with a suite of advanced data and [AI](https://www.javatpoint.com/artificial-intelligence-tutorial) tools. It offers various services such as [Infrastructure as a service](https://www.javatpoint.com/infrastructure-as-a-service), [Software as a service](https://www.javatpoint.com/software-as-a-service), and [platform as a service](https://www.javatpoint.com/platform-as-a-service).



Feature of IBM Cloud

* IBM cloud improves operational efficiency.
* Its speed and improve the customer's satisfaction.
* It offers Infrastructure as a Service (IaaS), Platform as a Service (PaaS), as well as Software as a Service (SaaS)

5. VMware Cloud

VMware cloud is a Software-Defined Data Center (SSDC) unified platform for the Hybrid Cloud. It allows cloud providers to build flexible, efficient, and robust cloud services.



Features of VMware

* VMware cloud works on the **pay-as-per-use** model and **monthly subscription**
* It provides better customer satisfaction by protecting the user's data.
* It can easily create a new VMware **Software**-**Defined Data Center (SDDC)** cluster on AWS cloud by utilizing a RESTful API.
* It provides flexible storage options.
* It provides a dedicated high-performance network for managing the application traffic and also supports multicast networking.
* It eliminates the time and cost complexity.

6. Oracle cloud

[Oracle](https://www.javatpoint.com/oracle-tutorial) cloud platform is offered by the **Oracle Corporation**. It combines Platform as a Service, Infrastructure as a Service, Software as a Service, and Data as a Service with cloud infrastructure.



Features of Oracle cloud

* Oracle cloud provides various tools for build, integrate, monitor, and secure the applications.
* Its infrastructure uses various languages including, Java, Ruby, PHP, Node.js.
* It offers customizable Virtual Cloud Networks, firewalls, and IP addresses to securely

Virtualization in Cloud Computing

**Virtualization** is the "creation of a virtual (rather than actual) version of something, such as a server, a desktop, a storage device, an operating system or network resources".

In other words, Virtualization is a technique, which allows to share a single physical instance of a resource or an application among multiple customers and organizations.

What is the concept behind the Virtualization?

Creation of a virtual machine over existing operating system and hardware is known as Hardware Virtualization. A Virtual machine provides an environment that is logically separated from the underlying hardware.

The machine on which the virtual machine is going to create is known as **Host Machine** and that virtual machine is referred as a **Guest Machine**

Benefits of Virtualization:

* More flexible and efficient allocation of resources.
* Enhance development productivity.
* It lowers the cost of IT infrastructure.
* Remote access and rapid scalability.
* High availability and disaster recovery.
* Pay peruse of the IT infrastructure on demand.
* Enables running multiple OS

Types of Virtualization:

* Hardware Virtualization.
* Operating system Virtualization.
* Server Virtualization.
* Storage Virtualization.

1) Hardware Virtualization:

When the virtual machine software or virtual machine manager *(VMM) is directly installed on the hardware system* is known as hardware virtualization.

The main job of hypervisor is to control and monitoring the processor, memory and other hardware resources.

After virtualization of hardware system we can install different operating system on it and run different applications on those OS.

**Usage:**

Hardware virtualization is mainly done for the server platforms, because controlling virtual machines is much easier than controlling a physical server.

2) Operating System Virtualization:

When the virtual machine software or virtual machine manager *(VMM) is installed on the Host operating system* instead of directly on the hardware system is known as operating system virtualization.

**Usage:**

Operating System Virtualization is mainly used for testing the applications on different platforms of OS.

3) Server Virtualization:

When the virtual machine software or virtual machine manager *(VMM) is directly installed on the Server system* is known as server virtualization.

**Usage:**

Server virtualization is done because a single physical server can be divided into multiple servers on the demand basis and for balancing the load.

4) Storage Virtualization:

Storage virtualization is the *process of grouping the physical storage from multiple network storage devices so that it looks like a single storage device*.

Storage virtualization is also implemented by using software applications.

**Usage:**

Storage virtualization is mainly done for back-up and recovery purposes.

* Containers:

Containers are lightweight packages of your application code together with dependencies such as specific versions of programming language runtimes and libraries required to run your software services.

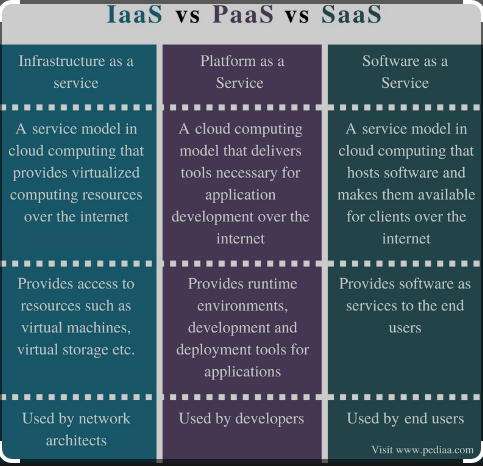
Differences between Containers & VMs

* Containers are much more lightweight than VMs
* Containers virtualize at the OS level while VMs virtualize at the hardware level
* Containers share the OS kernel and use a fraction of the memory VMs require

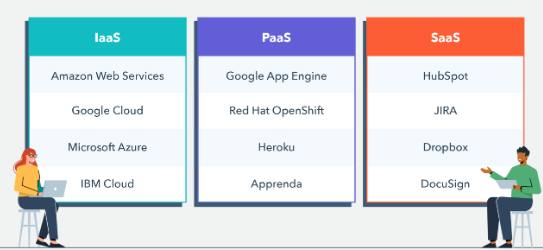
What are the benefits of containers?

* **Separation of responsibility:** Containerization provides a clear separation of responsibility, as developers focus on application logic and dependencie.
* **Workload portability**: Containers can run virtually anywhere, greatly easing development and deployment: on Linux, Windows, and Mac operating systems.
* **Application isolation:** Containers virtualize CPU, memory, storage, and network resources at the operating system level.

Comparision between PaaS , IaaS and SaaS:



Example of IaaS, PaaS and SaaS:



Essentials of Cloud Billing concepts:

## Creating a Free Tier Account on GCP

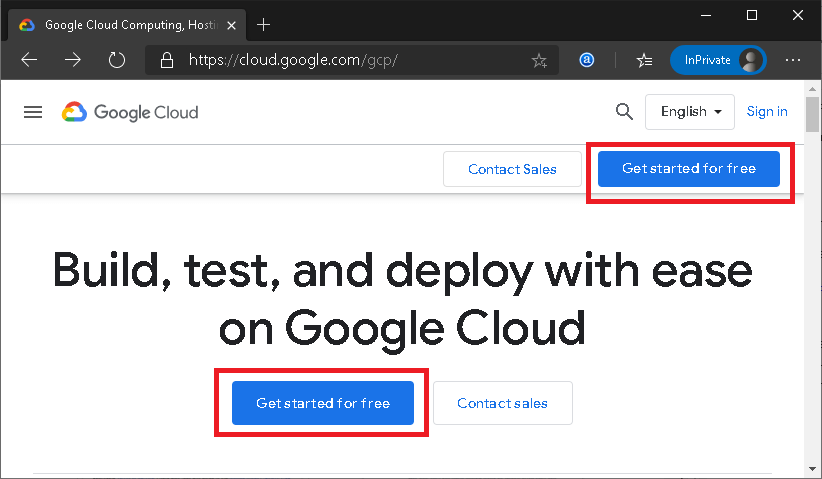
**To start using Google Cloud Platform, we are first required to create an account GCP. Here, we will create a free tier account for explaining the upcoming topic of this tutorial. The best thing about free account is that Google provides $300 worth credit to spend over the next 90 days after the date of account creation. Google offers all the core services of GCP with a free account for the next 90 days.**

**However, users must have a credit card to start a free tier account. Google asks for the credit card details to make sure that it is a genuine human request. Google does not charge automatically even after the 90 days or when we have exhausted the $300 free credit. The amount will only be charged** **when we will be upgrading our free account to a paid account manually.**

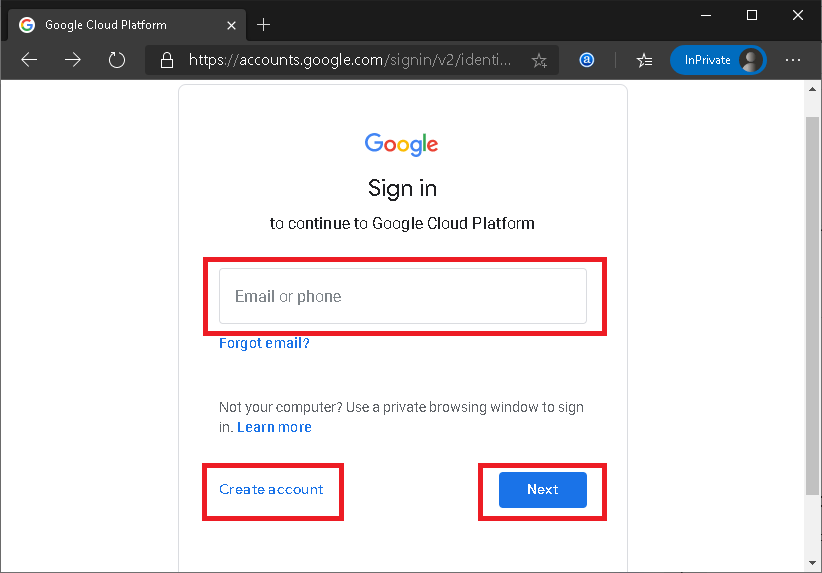
**Let's start with the steps of creating a free tier account on Google Cloud Platform:**

Step 1**: First, we are required to navigate to the following link:**[**https://cloud.google.com/gcp/**](https://cloud.google.com/gcp/)

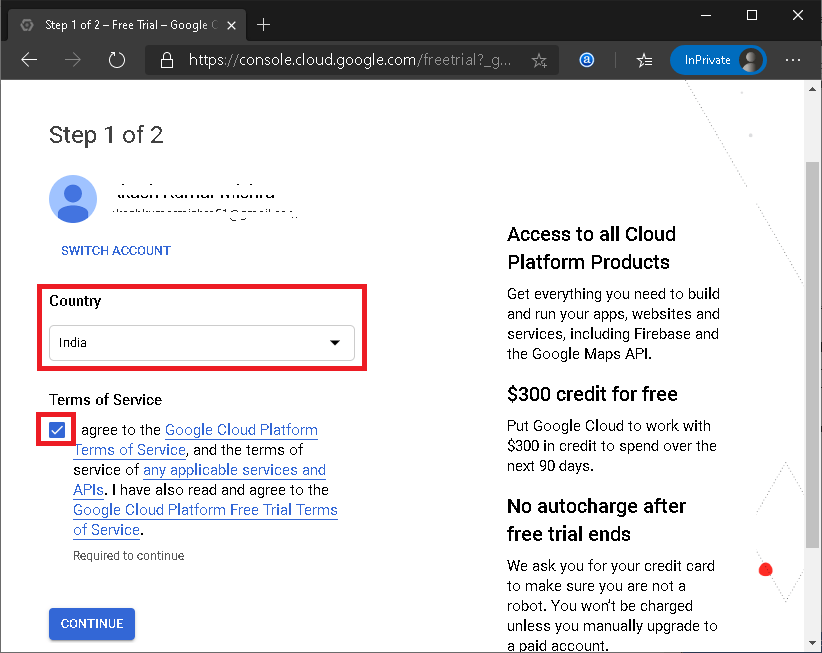
Step 2**: On the next screen, we need to click on 'Get started for free', as shown below:**

****

Step 3**: Next, we are required to login to the Google Account. We can use the 'create an account' button if we don't have an existing Google** account.

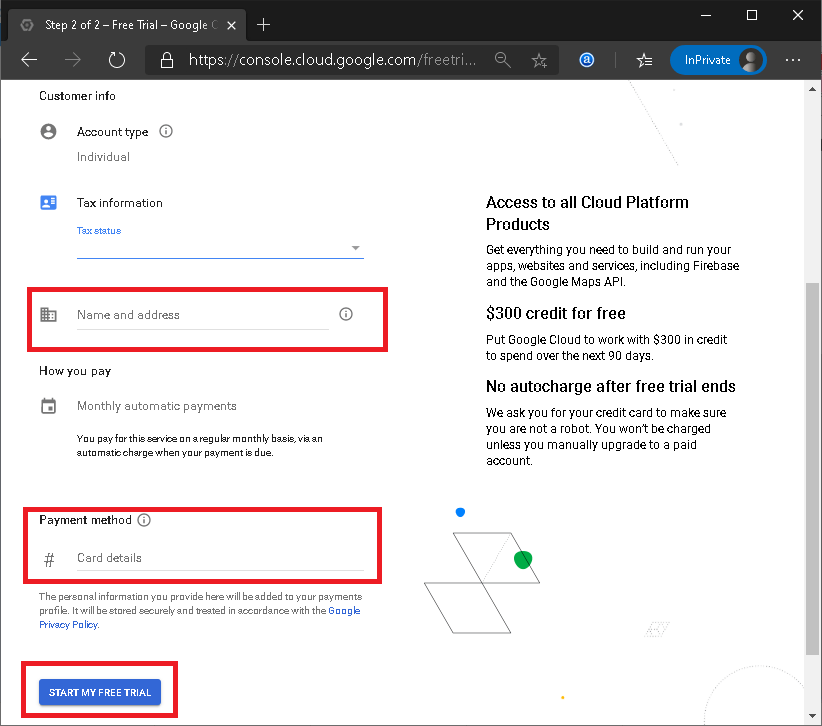


**Step 4**: Once we have logged in, we will get to the following screen:

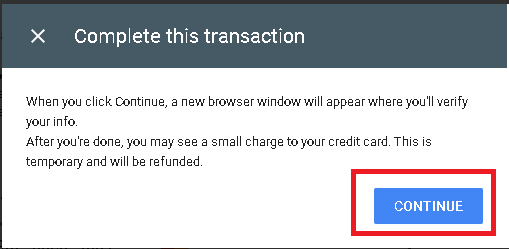


Here, we must select the Country, agree to the Terms of Service, and then click on the 'CONTINUE' button.

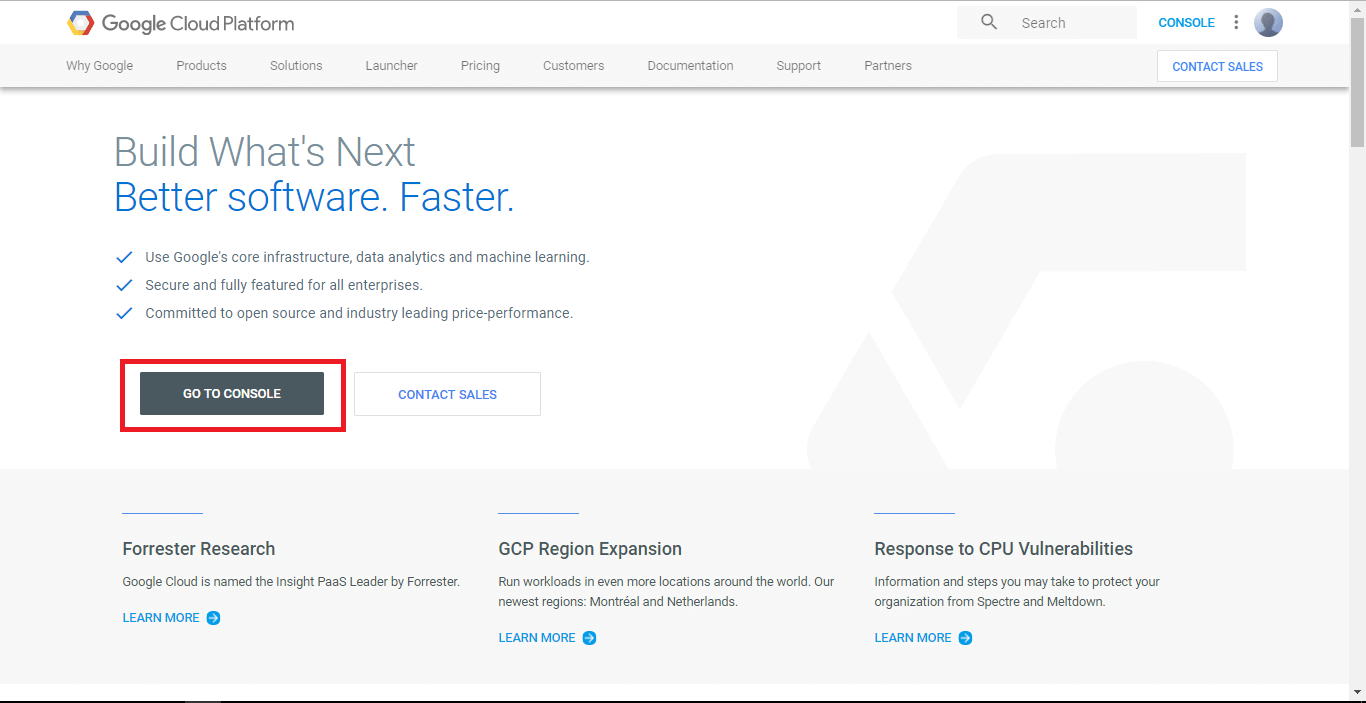
**Step 5**: On the next screen, we have to enter some necessary details such as name and address details. Also, we have to enter payment details like the method of payments and credit card details. After filling all the details, we need to click on the button 'START MY FREE TRIAL' from the bottom of the page:



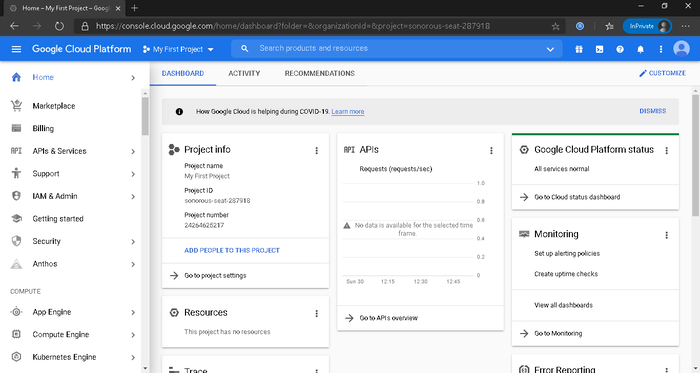
**Step 6: Google asks for the confirmation to use the credit card for the small deduction to ensure that the card information is correct. However, the amount is** refunded back to the same account. Here, we need to click on the 'CONTINUE' button:



**Step 7**: On the next screen, we must click on the 'GO TO CONSOLE' button:



**After clicking on the 'GO TO CONSOLE' button, we will be redirected to the Dashboard that includes a** **summary of GCP services along with projects and other insights. It looks like this:**



**To be specific, a Dashboards of GCP displays the summarized view of the followings:**

Project Info**: contains project details such as project name, ID, and number.**

Resources**: contains a list of resources being used in the related project.**

APIs**: contains various API requests running with the project (in request/sec form).**

**Google Cloud Platform Status**: displays an overall summary of services that are part of GCP.

**Monitoring**: displays alerts, performance stats, Uptime, etc. to ensure that systems are running reliably.

**Error Reporting**: displays errors occurring in the projects, but it needs to be configured first.

**Trace**: displays latency data of existing applications across a distributed tracing system.

**Compute Engine**: displays the insights of CPU usage in percentage (%).

**Tutorials**: contains Getting Started guides (basic guides) to explain how the GCP features work.

**News**: displays news and other important updates regarding Google Cloud Platform.

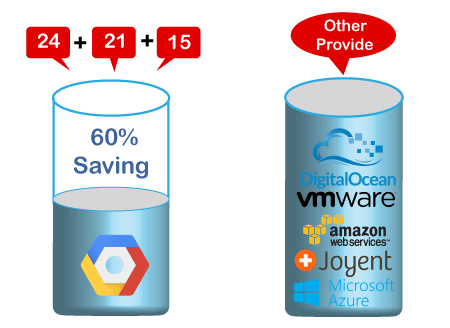
**Documentation**: contains in-depth guides to teach more about Compute Engine, Cloud Storage, and App Engine.

## Google Cloud Platform Pricing

When it comes to pricing, Google Cloud Platform is the cheapest solution in the market. GCP is not only low on price but also offers more features and services than other providers.

When comparing GCP with other leading competitors, it has more benefits over them. Google provides its users a massive 60% savings, including:

* 15% rightsizing recommendation
* 21% list price differences
* 24% of sustained usage discounts



Some of the main benefits of GCP pricing are:

**No Hidden Charges**: There are no hidden charges behind the GCP pricing. Google's pricing structure is straightforward and can be easily understood.

**Pay-as-you-go**: Google offers its customer 'use now, pay later' option. So, users will have to pay only for those services which they want to use or already using.

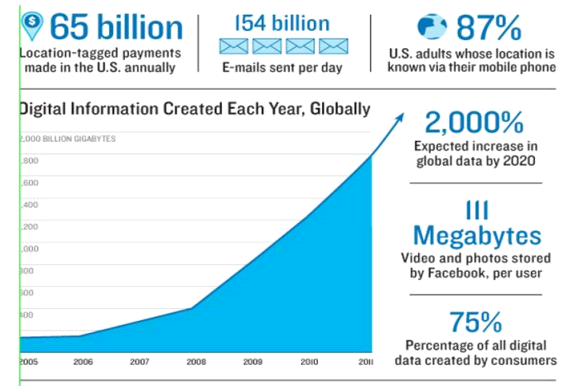
**No Termination Fee**: Users are free to stop using Google services whenever they want, and there will not have to pay any termination fee. That means the moment users stop using Google services; they stop paying for it.

**BIG DATA**

**How data is generated?**

People are generating large amount of data on social networking sites like Facebook, Twitter and Linked in, online photo sharing sites like Instagram, Picasa and video sharing site like YouTube. Large scale data is generated **using blogging sites, email, mobile text messages and personal documents**

Social Media gained popularity from 2008 and gave birth to the term Big Data. Due to popularity & success of different social networks, many more sites are launched and contribute to Big Data.



**Fig. Digital information created each year globally**

Definition of Big data:

***Big data is huge, large, or voluminous data, information, or the relevant statistics acquired by large organizations that are difficult to process by traditional tools***

Sources of Big Data

These data come from many sources like

* **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
* **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
* **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
* **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
* **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

**Types of Big Data**

Following are the types of Big Data:

* **Structured**
* **Unstructured**
* **Semi-structured**

**Structured**

Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data.

**Unstructured**

 A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.

**Semi-structured**

Semi-structured data can contain both the forms of data.

**V’s of Big Data**

Big data can be described by the following characteristics:

* Volume
* Variety
* Velocity

***(i) Volume –*** The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **‘Volume’** is one characteristic which needs to be considered while dealing with Big Data solutions.

***(ii) Variety –*** The next aspect of Big Data is its **variety**.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

***(iii) Velocity –*** The term **‘velocity’** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors,[Mobile](https://www.guru99.com/mobile-testing.html)devices, etc. The flow of data is massive and continuous.

Ex- every minute over 200 million emails are sent, around 20 million photos are viewed and 30.000 uploaded on Flickr, almost 300.000 tweets are sent and almost 2.5 million queries on Google are performed.

***Role of big data in Artificial Intelligence and machine learning:***

[Big data analytics](https://www.qlik.com/us/data-analytics/big-data-analytics) is the use of processes and technologies, including AI and machine learning, to combine and analyze massive datasets with the goal of identifying patterns and developing actionable insights. This helps you make faster, better, data-driven decisions that can increase efficiency, revenue and profits.

Data and AI are merging into a collaborative relationship, where AI is useless without data, and mastering data is impossible without AI.

**How AI is used in big data**

AI makes big data analytics simpler by automating and enhancing data preparation, data visualization, predictive modeling, and other complex analytical tasks that would otherwise be labor-intensive and time-consuming.

**How ML is used in big data**

Big data analytics can make sense of the data by uncovering trends and patterns. Machine learning can accelerate this process with the help of decision-making algorithms. **It can categorize the incoming data, recognize patterns and translate the data into insights helpful for business operations**.

**Python packages/libraries for Machine Learning and deep learning:**

Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reasons is its vast collection of libraries. Python libraries that are used in Machine Learning are: 

* Numpy
* Scipy
* Scikit-learn
* Theano
* TensorFlow
* Keras
* PyTorch
* Pandas
* Matplotlib

Numpy

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

# Python program using NumPy

# for some basic mathematical

# operations

import numpy as np

# Creating two arrays of rank 2

x = np.array([[1, 2], [3, 4]])

y = np.array([[5, 6], [7, 8]])

# Creating two arrays of rank 1

v = np.array([9, 10])

w = np.array([11, 12])

# Inner product of vectors

print(np.dot(v, w), "\n")

# Matrix and Vector product

print(np.dot(x, v), "\n")

# Matrix and matrix product

print(np.dot(x, y))

Output: 

219

[29 67]

[[19 22]

[43 50]]

SciPy

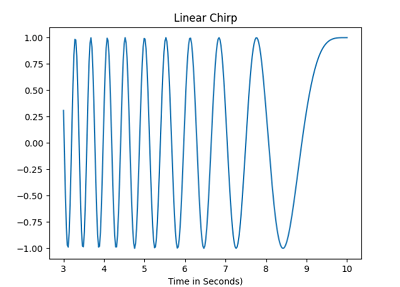


[SciPy](https://www.javatpoint.com/python-scipy) is a popular library among Machine Learning developers as it contains numerous modules for performing optimization, linear algebra, integration, and statistics. SciPy library is different from SciPy stack, as SciPy library is one of the core packages which made up the SciPy stack. SciPy library is used for image manipulation tasks.

**Example 1:**

1. from scipy.signal **import** chirp as cp
2. from scipy.signal **import** spectrogram as sp
3. **import** matplotlib.pyplot as plot
4. **import** numpy as nup
5. t\_T = nup.linspace(3, 10, 300)
6. w\_W = cp(t\_T, f0 = 4, f1 = 2, t1 = 5, method = 'linear')
7. plot.plot(t\_T, w\_W)
8. plot.title ("Linear Chirp")
9. plot.xlabel ('Time in Seconds)')
10. plot.show()

**Output:**



Pandas



[Pandas](https://www.javatpoint.com/python-pandas) is a Python library that is mainly used for data analysis. The users have to prepare the dataset before using it for training the machine learning. Pandas make it easy for the developers as it is developed specifically for data extraction. It has a wide variety of tools for analysing data in detail, providing high-level data structures.

**Example:**

1. **import** pandas as pad
3. data\_1 = {"Countries": ["Bhutan", "Cape Verde", "Chad", "Estonia", "Guinea", "Kenya", "Libya", "Mexico"],
4. "capital": ["Thimphu", "Praia", "N'Djamena", "Tallinn", "Conakry", "Nairobi", "Tripoli", "Mexico City"],
5. "Currency": ["Ngultrum", "Cape Verdean escudo", "CFA Franc", "Estonia Kroon; Euro", "Guinean franc", "Kenya shilling", "Libyan dinar", "Mexican peso"],
6. "population": [20.4, 143.5, 12.52, 135.7, 52.98, 76.21, 34.28, 54.32] }
8. data\_1\_table = pad.DataFrame(data\_1)
9. print(data\_1\_table)

**Output:**

Countries capital Currency population

0 Bhutan Thimphu Ngultrum 20.40

1 Cape Verde Praia Cape Verdean escudo 143.50

2 Chad N'Djamena CFA Franc 12.52

3 Estonia Tallinn Estonia Kroon; Euro 135.70

4 Guinea Conakry Guinean franc 52.98

5 Kenya Nairobi Kenya shilling 76.21

6 Libya Tripoli Libyan dinar 34.28

7 Mexico Mexico City Mexican peso 54.32

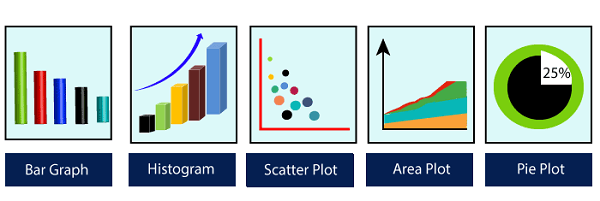
# Python Libraries for Data Visualization

Python programming language has different types of libraries for all kind of projects. Likewise, python has various libraries for visualization of Data, so that user can understand the dataset in very detailed way and analyze it properly.

The Libraries for Data Visualization in Python programming are given below:

* Matplotlib
* Ggplot
* Pygal
* Missingno
* Seaborn
* Plotly
* Gleam
* Leather
* Geoplotlib
* Bokeh
* Folium

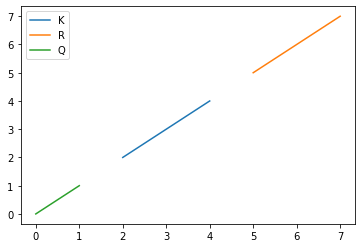
### Matplotlib

* 
* [Matplotlib](https://www.javatpoint.com/matplotlib) is a Python library that is used for data visualization. It is used by developers when they want to visualize the data and its patterns. It is a 2-D plotting library that is used to create 2-D graphs and plots.
* It has a module pyplot which is used for plotting graphs, and it provides different features for control line styles, font properties, formatting axes and many more. Matplotlib provides different types of graphs and plots such as histograms, error charts, bar charts and many more.

**Example 1:**

1. **import** matplotlib.pyplot as plot
2. **import** numpy as nup
4. # Prepare the data
5. K = nup.linspace(2, 4, 8)
6. R = nup.linspace(5, 7, 9)
7. Q = nup.linspace(0, 1, 3)
9. # Plot the data
10. plot.plot(K, K, label = 'K')
11. plot.plot(R, R, label = 'R')
12. plot.plot(Q, Q, label = 'Q')
14. # Add a legend
15. plot.legend()
17. # Show the plot
18. plot.show()

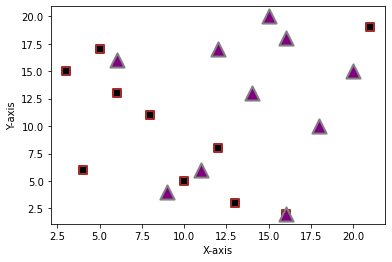
**Output:**



Example 2:

1. **import** matplotlib.pyplot as plot
3. # Creating dataset-1
4. K\_1 = [8, 4, 6, 3, 5, 10,
5. 13, 16, 12, 21]
7. R\_1 = [11, 6, 13, 15, 17, 5,
8. 3, 2, 8, 19]
10. # Creating dataset2
11. K\_2 = [6, 9, 18, 14, 16, 15,
12. 11, 16, 12, 20]
14. R\_2 = [16, 4, 10, 13, 18,
15. 20, 6, 2, 17, 15]
17. plot.scatter(K\_1, R\_1, c = "Black",
18. linewidths = 2,
19. marker = "s",
20. edgecolor = "Brown",
21. s = 50)
23. plot.scatter(K\_2, R\_2, c = "Purple",
24. linewidths = 2,
25. marker = "^",
26. edgecolor = "Grey",
27. s = 200)
29. plt.xlabel ("X-axis")
30. plt.ylabel ("Y-axis")
31. print ("Scatter Plot")
32. plt.show()

**Output:**



## Seaborn

Seaborn is a library of Python programming basically used for making statistical graphics of the dataset. This library is built on top of the Matplotlib library. It is also integrated closely with Pandas, which is used for the data structure of Datasets.

Seaborn is very helpful to explore and understand data in a better way. It provides a high level of a crossing point for sketching attractive and informative algebraic graphics. Let's understand the following example.

**Example -**

1. **import** seaborn as sns
2. sns.set ( color\_codes = True)
3. **import** matplotlib.pyplot as plt
4. **for** i in range(n\_rows):
5. fg,ax = plt.subplots(nrows=1,ncols=n\_cols,sharey=True,figsize=(8, 2.4))
6. **for** j in range(n\_cols):
7. sns.violinplot(x = data.Outcome, y=data[columns[idx]], ax=ax[j])
8. idx += 1
9. **if** idx >= 8:
10. **break**

**Output:**