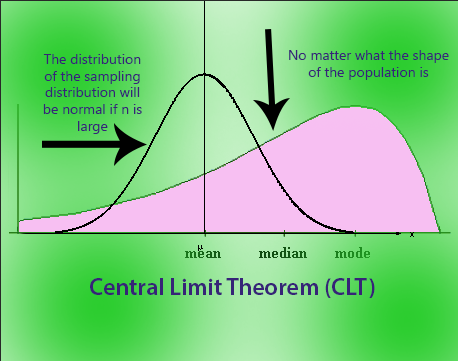
Statistics

# Central Limit Theorem

The normal distribution is achieved when the sample size varies without having an effect on the shape of the population distribution. Widely used in performing hypothesis testing and also to calculate the confidence intervals accurately.

Conditions for central limit theorem

* Data must be sampled randomly
* Sample value must be independent of each other
* Sample size must be sufficiently large, generally it should be greater or equal than 30

Example, we want to calculate the average height of people in the world, and we take some samples from the general population, which serves as the data set. Since it is hard or impossible to obtain data regarding the height of every person in the world, we will simply calculate the mean of our sample. By multiplying it several times, we will obtain the mean and their frequencies which we can plot on the graph and create a normal distribution. It will form a bell-shaped curve that will closely resemble the original data set.

# Hypothesis Testing and Statistical significance of an insight

Hypothesis testing in statistics is used to see if a certain experiment yields meaningful results. It essentially helps to assess the statistical significance of insight by determining the odds of the results occurring by chance.

Hypothesis testing is used to find out the statistical significance of the insight. The null hypothesis and the alternate hypothesis are stated, and the p-value is calculated.

After calculating the p-value, the null hypothesis is assumed true, and the values are determined. To fine-tune the result, the alpha value, which denotes the significance, is tweaked.

If the p-value turns out to be less than the alpha, then the null hypothesis is rejected, but if it is greater than alpha, the null hypothesis is accepted. This ensures that the result obtained is statistically significant.

# Difference between probability and likelihood

Probability attaches to possible results (chances) while likelihood attaches to the hypothesis.

Problem: Captain has to decide to bat first

**Probability**: Only two possibilities

Choose to bat

Doesn’t choose to bat

P (choose to bat) = P (doesn’t choose to bat) = 0.5

**Likelihood**: Choosing to bat first will depend on

Weather Conditions (rainfall, wind speed)

Due pitch

Humidity

# Assumption of Normality

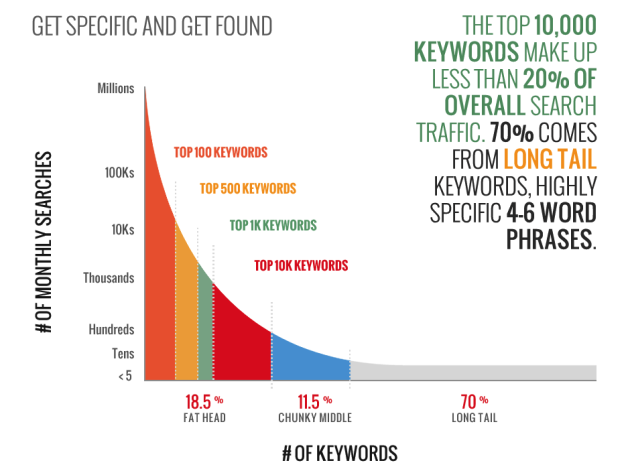
Mean distribution across samples is normal. This is also true across independent samples as well.

# One sample t-test

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population’s mean.

# Long-tailed Distribution

It is a type of distribution where the tail drops off gradually towards the end of the curve. It is widely used in classification and regression. Pareto principle and the product sales distribution are good examples.



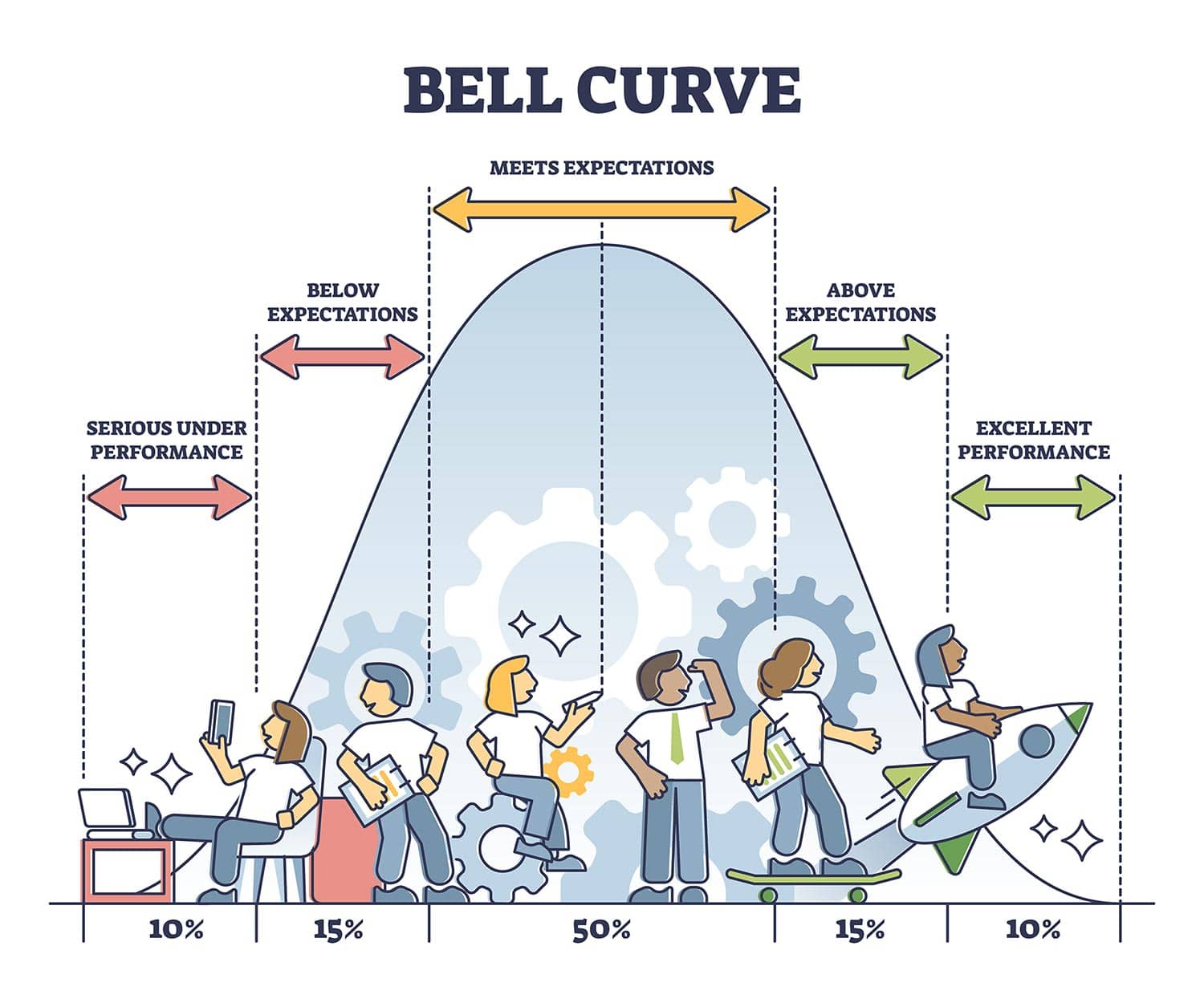
# p-values

It is a measure of the probability of having results equal to or more than the results achieved under a specific hypothesis assuming that the null hypothesis is correct. It represents the probability that the observed difference occurred randomly by chance. It is number that ranges from 0 to 1. It tells us how strong the results are. The claim that is kept for experiment or trial is called Null Hypothesis.

* Low p-value means values ≤ 0.05, null hypothesis can be rejected and the data is unlikely with true null. Results are against the Null Hypothesis
* High p-value means values ≥ 0.05, null hypothesis can be accepted. Results are in favour of the Null Hypothesis.
* P-value = 0.05 means that the hypothesis can go either way.

# Bell-Curve Distribution

Represented by the shape of a bell and indicates normal distribution. The top of the curve shows the mode, mean and median of the data and is perfectly symmetrical. It follows Empirical rule.



# Symmetric Distribution

Data on the left side of the median is same as the one present on the right side of the median.

* Uniform Distribution – When all the outcomes are equally likely
* Binomial Distribution
* Normal Distribution

# Handling Missing Data

* Prediction of the missing values
* Assignment of individual (unique) values
* Deletion of rows, which have the missing data
* Mean imputation or median imputation - Null values in a dataset are replaced directly with the corresponding mean of the data. It is a bad practice as it removes the accountability for feature correlation. The data will have low variance and increased bias, adding to the dip in accuracy of the model.
* Using random forests, it supports the missing values

# Exploratory Data Analysis

The process of performing investigations on data to understand the data better. Initial investigations are done to determine patterns, spot abnormalities, test hypotheses, and also check if the assumptions are right.

# Standard Deviation

Represents the magnitude of how far the data points are from the mean. A high value indicates that the data is spread to extreme ends, far away from the mean.

# Outliers in Stats

Outliers are data points that vary in a large way when compared to other observations in the dataset. An outlier worsens the accuracy of a model and decreases its efficiency.

**In case when there are a lot of outliers that can positively or negatively skew data, the median is preferred as it provides an accurate measure**.

Outlier are determined by using two methods:

* Standard deviation / z-score
* Interquartile range (IQR)

There are some scenarios where outliers are kept in the data, including

* When results are critical
* Outliers add meaning to the data
* Data is highly skewed

# Inlier

A data point that lies at the same level as the rest of the dataset. Finding an inlier in the dataset is difficult when compared to an outlier as it requires external data to do so. Inliers reduce model accuracy; hence they are removed. It is an error.

# Six sigma

Six sigma is a quality assurance methodology used widely in statistics to provide ways to improve processes and functionality when working with data.

A process is considered as six sigma when 99.99966% of the outcomes of the model are considered to be defect-free.

# Quartile

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

* The lower quartile (Q1) is the 25th percentile
* The middle quartile (Q2), also called median, is the 50th percentile
* The upper quartile (Q3) is the 75th percentile

# 5-number summary

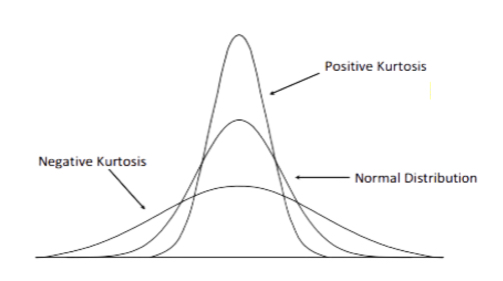
It is a measure of five entities that cover the entire range of data.

* Low extreme (Min)
* First quartile (Q1)
* Median
* Upper quartile (Q3)
* High extreme (Max)

# Quantitative and Qualitative Data

* Quantitative data is known as numeric data.
* Qualitative data is known as categorical data.

# Kurtosis

Kurtosis is used to describe the extreme values present in one tail of distribution or the peaks of frequency distribution versus the other. It is the measure of outliers present in the distribution. A high value of kurtosis represents large amount of outliers in the data. To overcome this, we have to either add more data into the dataset or remove the outliers.

The standard normal distribution has a kurtosis of 3 whereas the values of symmetry and kurtosis between -2 and +2 are considered normal and acceptable.

# Descriptive and Inferential Statistics

* Descriptive statistics is used to summarize a sample set of data like the standard deviation or the mean.
* Inferential statistics is used to draw conclusions from the test data that are subjected to random variations. Inferential statistics is used in research, government operations, quality control and quality assurance teams in MNCs. A population is a large volume of observations. The sample is a small portion of that population.

# Bessel’s Correction

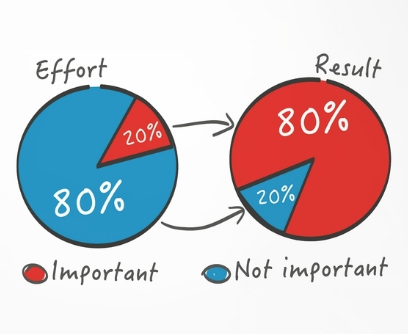
It is a factor that is used to estimate a populations’ standard deviation from its sample. It causes the standard deviation to be less biased, thereby, providing more accurate results. Bessel’s correction advocates the use of n-1 instead of n in the formula of standard deviation.

# DOE

Design of Experiments, is a systematic method that explains the relationship between the factors affecting a process and its output. It is used to infer and predict an outcome by changing the input variables.

# Pareto Principle

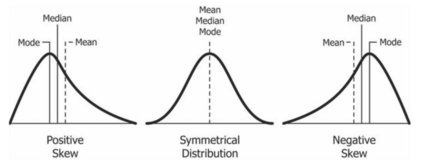
80/20 rule, which means that 80% of the results are obtained from 20% of the causes in an experiment.



# Skewness

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

A distribution can exhibit positive skewness or negative skewness if the tail on the right is longer and the tail on the left side is longer, respectively.



# Alternative Hypothesis

It is a statement used to contradict the null hypothesis. Denoted by H1, is the statement that must be true if the null hypothesis is false. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

# Degree of Freedom

DF is used to define the number of options at hand when performing an analysis. It is mostly used with t-distribution and not with the z-distribution.

If there is an increase in DF, the t-distribution will reach closer to the normal distribution. If DF > 30, this means that the t-distribution at hand is having all of the characteristics of a normal distribution.

Where, Df is degrees of freedom, N is sample size. Degree of freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.

Example, consider a data sample consisting of five positive integers. The values of the five integers must have an average of six. If four of the items within the data set are {3, 8, 5, and 4}, the fifth number must be 10. Because the first four number can be chosen at random, the degree of freedom is four.

Consider a data sample consisting of one integer. That integer must be odd. Because there are constraints on the single item within the data set, the degrees of freedom are zero.

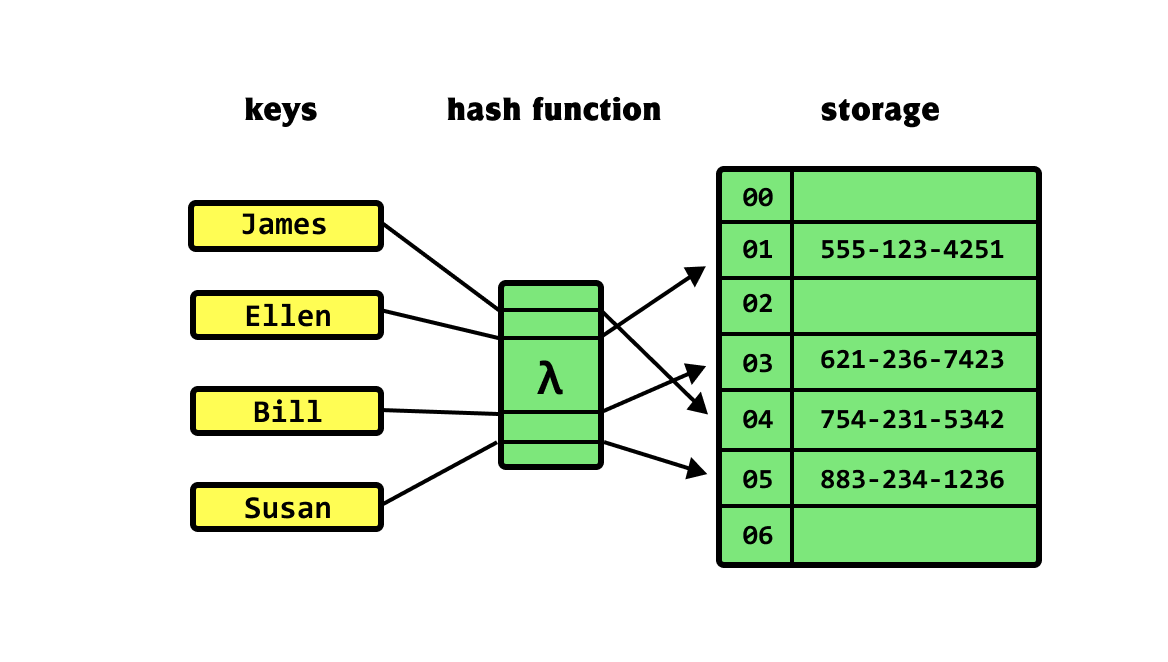
Degree of freedom is used in chi-square test. There are two kinds of chi-square tests: the test of independence, which asks a question of relationship, such as, “Is there a relationship between gender and SAT scores?”; and the goodness-of-fit test, which asks something like “If a coin is tossed 100 times, will it come up heads 50 times and tails 50 times?”

Degrees of freedom tells you how many units within a set can be selected without constraints to still abide by a given rule overseeing the set. For example, consider a set of five items that add to an average value of 20. Degrees of freedom tell you how many of the items (4) can be randomly selected before constraints must be put in place. In this example, once the first four items are picked, you no longer have liberty to randomly select a data point because you must “force balance” to the given average.

# Sensitivity

It is used to determine the accuracy of a classifier (logistic, random forest etc).

# Hash tables

Hash tables are the data structures that are used to store key-value pairs in a structured way. The hashing function is used by a hash table to compute an index that contains all of the details regarding the keys that are mapped to their associated values.

# Empirical rule

Every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68-95-99.7 rule / three-sigma rule. 68% of values will fall within first standard deviation, 95% will fall within two standard deviations, and 99.75% will fall within three standard deviations (µ ± 3σ).

# Statistical Interaction

When two or more variable interact, and this results in a third variable being affected. A real-life example includes the interaction of adding sugar to the stirring of tea. Neither of the two variables has an impact on sweetness, but it is the combination of these two variables that do.

# Heteroscedastic Model

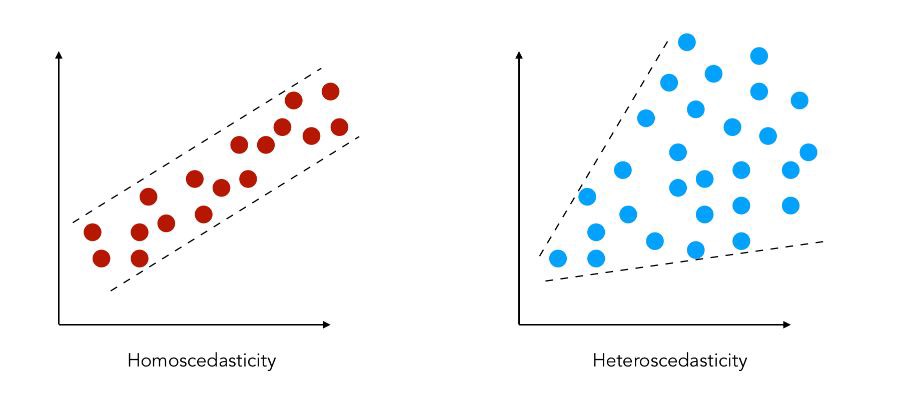
A model is said to be heteroscedastic when the variation in errors comes out to be inconsistent. It often occurs in two forms – conditional and unconditional.

Heteroscedasticity means unequal scattered distribution. Heteroscedasticity is the systematic change in the spread of the residuals or errors over the range of measured values.

It occurs often in datasets, where we have large range between the largest and the smallest observed values. There are two types of heteroscedasticities

Pure heteroscedasticity – It refers to cases where we specify the correct model and let us observe the non-constant variance in residual plots.

Impure heteroscedasticity – It refers to cases where you incorrectly specify the model, and that causes the non-constant variance.



# Autocorrelation

A representation of the degree of correlation between the two variables in a given time series and a lagged version of itself over successive time intervals. It’s conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods. The data is correlated in a way that future outcomes are linked to past outcomes. Autocorrelation makes a model less accurate because even errors follow a sequential pattern.

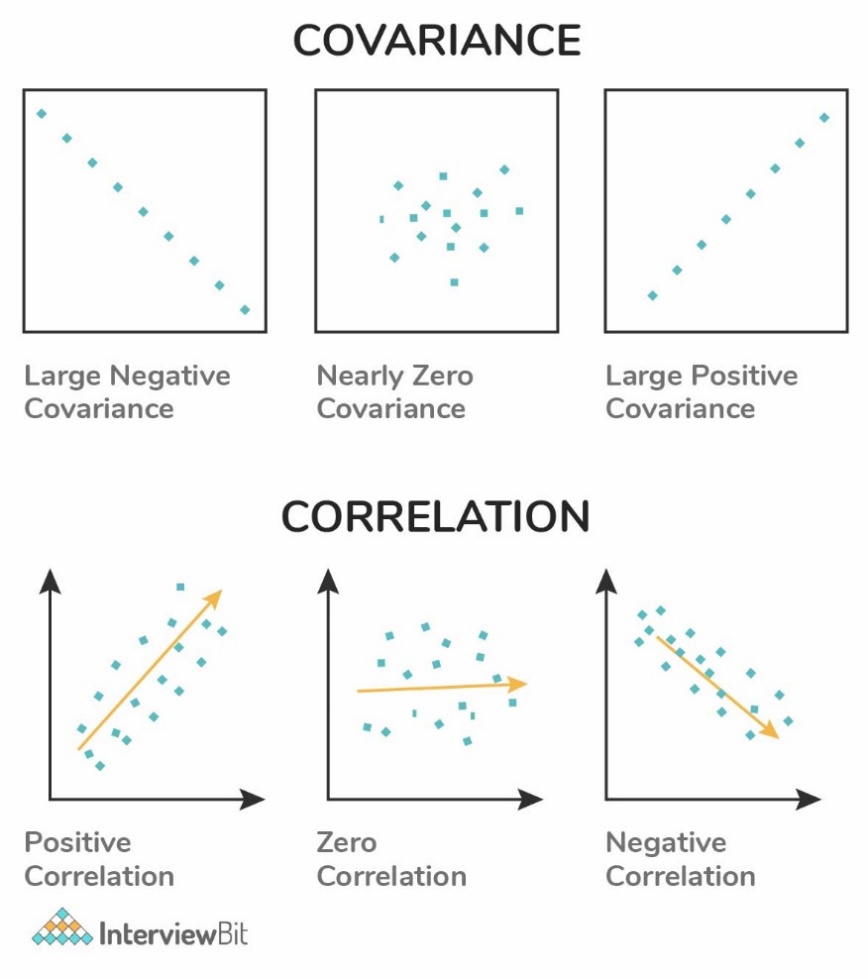
An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of -1 represents a perfect negative correlation. Autocorrelation is also known as lagged correlation/serial correlation.

Autocorrelation can be used to determine if a momentum trading strategy makes sense.

For example, if it's rainy today, the data suggests that it's more likely to rain tomorrow than if it's clear today. When it comes to investing, a stock might have a strong positive autocorrelation of returns, suggesting that if it's "up" today, it's more likely to be up tomorrow, too.

# Correlation and Covariance

Correlation is a technique to measure and estimate the quantitative relationship between two variable and is measured in terms of how strong are the variables related. It is dimensionless. The value of correlation between two variables ranges from -1 to +1. The value -1 represents high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

Covariance is a systematic relationship between pair of variables where changes in one affect changes in another variable. The systematic relation is determined between a pair of random variables to see if the change in one will affect the other variable in the pair or not. Covariance dimension depends on variables.

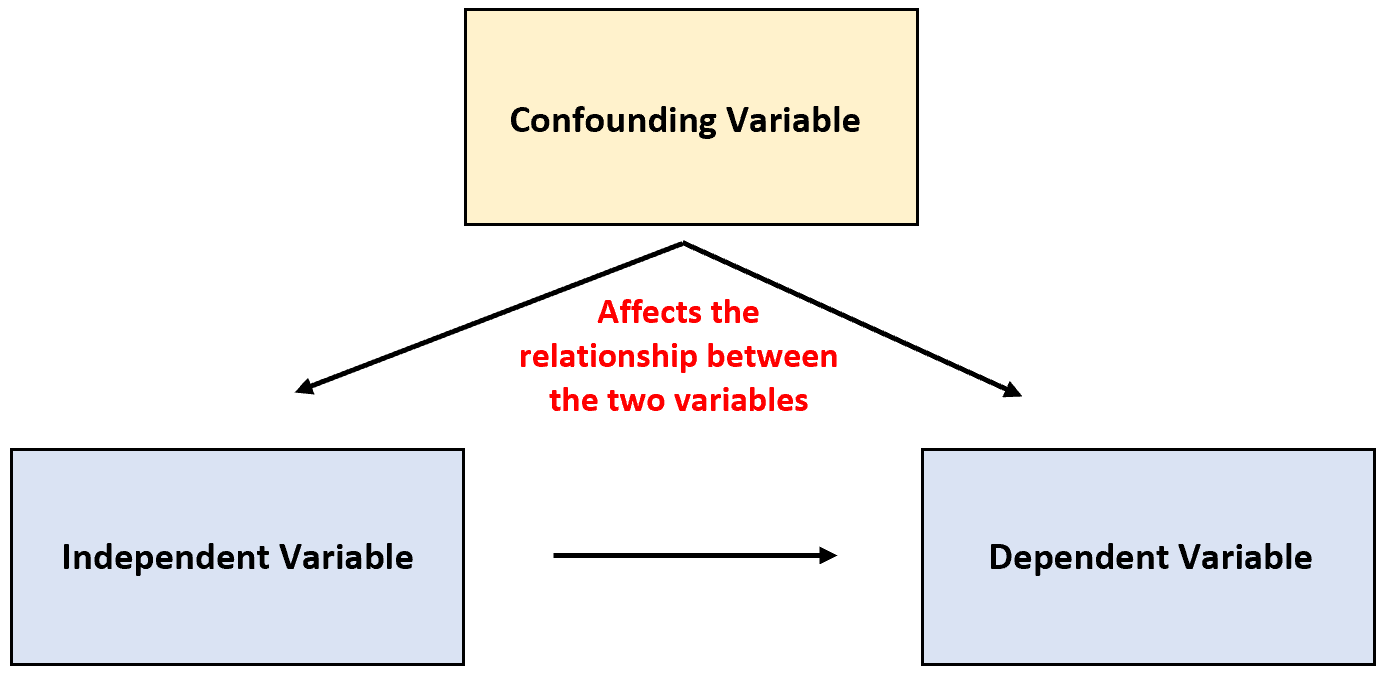
Mathematically, consider 2 random variables, X and Y where the means are represented as and respectively and standard deviations are represented by and respectively and E represents the expected value operator, then:

# Different types of correlation

There are mainly three types of correlation:

* Pearson - Normalized measurement of covariance, assumes both the variables are normally distributed, Measures linear relationship but fail to measure the non-linear relationship between variables.
* Spearman Rank – It is a non-parametric measure, measures both linear and non-linear relationship between two variables.
* Kendall Rank- Non parametric measure for calculating the rank of the correlation coefficient, measures both linear and non-linear relationship between two variables.

# Confounding Variable

A variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provide useless results.

For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather is another confounding variable.

# Cherry-picking

Practise of selecting information which supports a certain claim and ignores any other claim that refutes the desired conclusion. It does not necessarily mean that one side’s information is incorrect, but that a complete picture is not being presented because available evidence to support the other side is not being taken into account.

Cherry picking data happens all the time in politics. For example, in June, 2020, President Trump claimed that the US was doing well in the battle against COVID-19 because the death rate from the disease was declining. This statement totally ignored the information that new records were being set every day for people testing positive for the SARS-CoV-2 virus.

What we have here is a typical case of 'cherry-picking' of data. It isn’t that Trump’s information is incorrect, it just doesn’t present a complete picture of the situation. And politicians are not the only ones guilty of this type of data reporting. Environmentalists, industry representatives, activists of all sorts, and government officials are all in on the cherry-picking harvest!

# P-hacking

Refers to a technique in which data collection or analysis is manipulated until significant patterns can be found who have no underlying effect whatsoever.

# Significance Chasing

Also known as Data Dredging, Data Fishing, or Data Snooping. It refers to the reporting of insignificant results as if they are almost significant.

# Relationship between Confidence level and Significance level

Significance level is the probability of obtaining a result that is extremely different from the condition where the null hypothesis is true. While the confidence level is used as a range of similar values in a population.

Both significance and confidence level are related by the following formula:

# Calculate the score, given that Mean is 160, SD is 15 and Z-score is 1.2

# Difference between Confidence test and Hypothesis test

Confidence interval provides a range of values that helps in capturing the unknown parameter. Confidence interval are important in medical research to provide researchers with a strong bias for their estimations.

Hypothesis testing is used to test an experiment or observation and determine if the results did not occur purely by chance or luck using the below formula where ‘p’ is some parameter.

# Power of Test

The probability of rejecting the null hypothesis when it’s false.

To increase the power of the test

* You can increase alpha, but it also increases the chance of type 1 error.
* Increase the sample size, n. This maintains the type 1 error but reduces type 2.

Type 1 error is a false positive conclusion.

Type 2 error is a false negative conclusion.

# Difference between population parameters and sample statistics

Population parameters are

* Mean =
* Standard deviation =

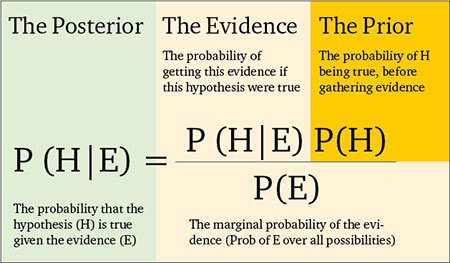
Sample statistics include

* Mean =
* Standard deviation = s

# Bayesian and Frequentist

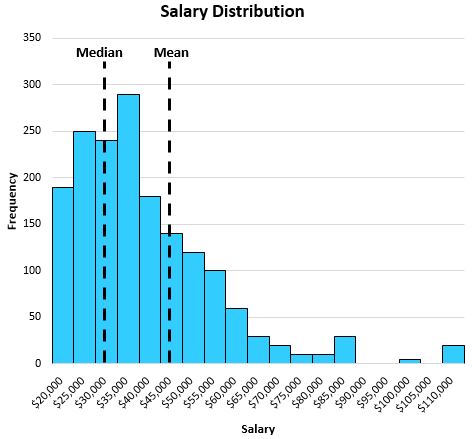
Bayesian rests on the data which is actually observed in reality and further considers the probability distribution on the hypothesis.

Frequentists rest on the hypothesis of choice and further consider the probability distribution on the data, whether it is observed or not.



# Reason of using Median instead of Mean

There are cases in statistical analysis in which the median may provide a better understanding of the data than the mean. I have used median instead of the mean when dealing with datasets that have many outliers. In these cases, the median can provide a better idea of a data trend without the distortion that outliers might cause in the average. It is best to use the mean to describe the centre of a dataset when the distribution is mostly symmetrical and there are no outliers. When a distribution is skewed, the median does a better job of describing the centre of the distribution than the mean.



# When should you use a t-test vs a z-test?

The z-test is used for hypothesis testing in statistics with a normal distribution. It is used to determine population variance in the case where a sample is large.

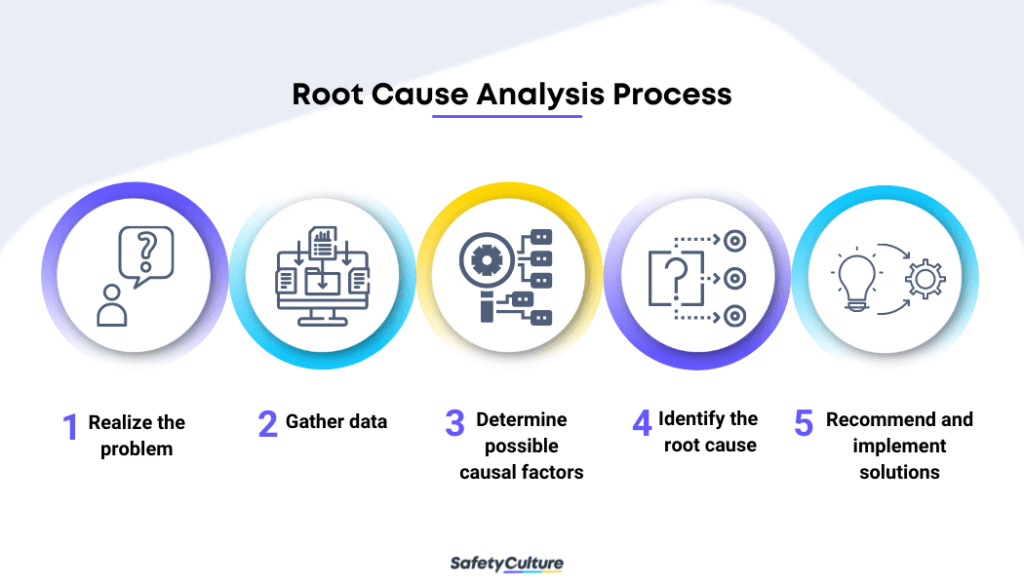
The t-test is used with a t-distribution and used to determine population variance when you have a small sample size.

In case the sample size is large or n>30, a z-test is used. T-tests are helpful when the sample size is small or n<30.

# What is ANOVA test?

Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means or average of different groups. A range of scenarios uses it to determine if there is any difference between the means of different groups.

# Root Cause Analysis

It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring. It was initially developed to analyse industrial accidents.

Example: If the higher crime rate in a city is directly associated with the higher sales in a red-coloured shirt, it means that they are having a positive correlation. However, this does not mean that one causes the other. Causation can be tested using A/B testing or hypothesis testing.

# References

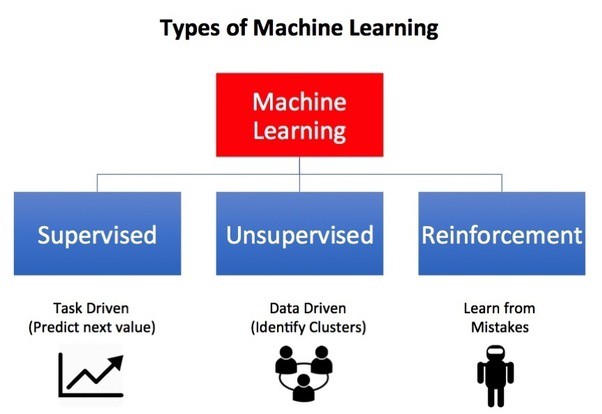
1. Intellipaat
2. Mygreatlearning
3. Analyticsindiamag
4. Indeed.com

Machine Learning and Deep Learning

# Data Science

An interdisciplinary field that includes various scientific processes, algorithms, tools, and machine learning techniques to find common patterns and gather sensible insights from the given raw input data using statistical and mathematical analysis. Statisticians work a posteriori (revised or updated probability of an event occurring after taking into consideration new information), explaining the results and designing a plan, data scientists use historical data to make predictions.

# Machine Learning

ML is a field of computer science that gives computers the ability to learn without being explicitly programmed. ML can be categorized in the following categories

* Supervised ML
* Semi-supervised ML
* Unsupervised ML
* Reinforcement learning (Method based on rewarding desired behaviours and/or punishing undesired ones)

Reason for popularity of DL in recent times

* Increase in the amount of data generated through various sources
* Growth in hardware resources required to run these models.
* GPUs are multiple times faster and they help us build bigger and deeper DL models in comparatively less time than required previously.

# What is Semi-Supervised ML

With more common supervised machine learning methods, you train a machine learning algorithm on a ‘labelled’ dataset in which each record includes the outcome information. This allows the algorithm to deduce patterns and identify relationships between your target variable and the rest of the dataset based on information it already has.

When you don’t have enough labelled data to produce an accurate model and you don’t have the ability or resources to get more data, you can use semi-supervised techniques to increase the size of your training data.

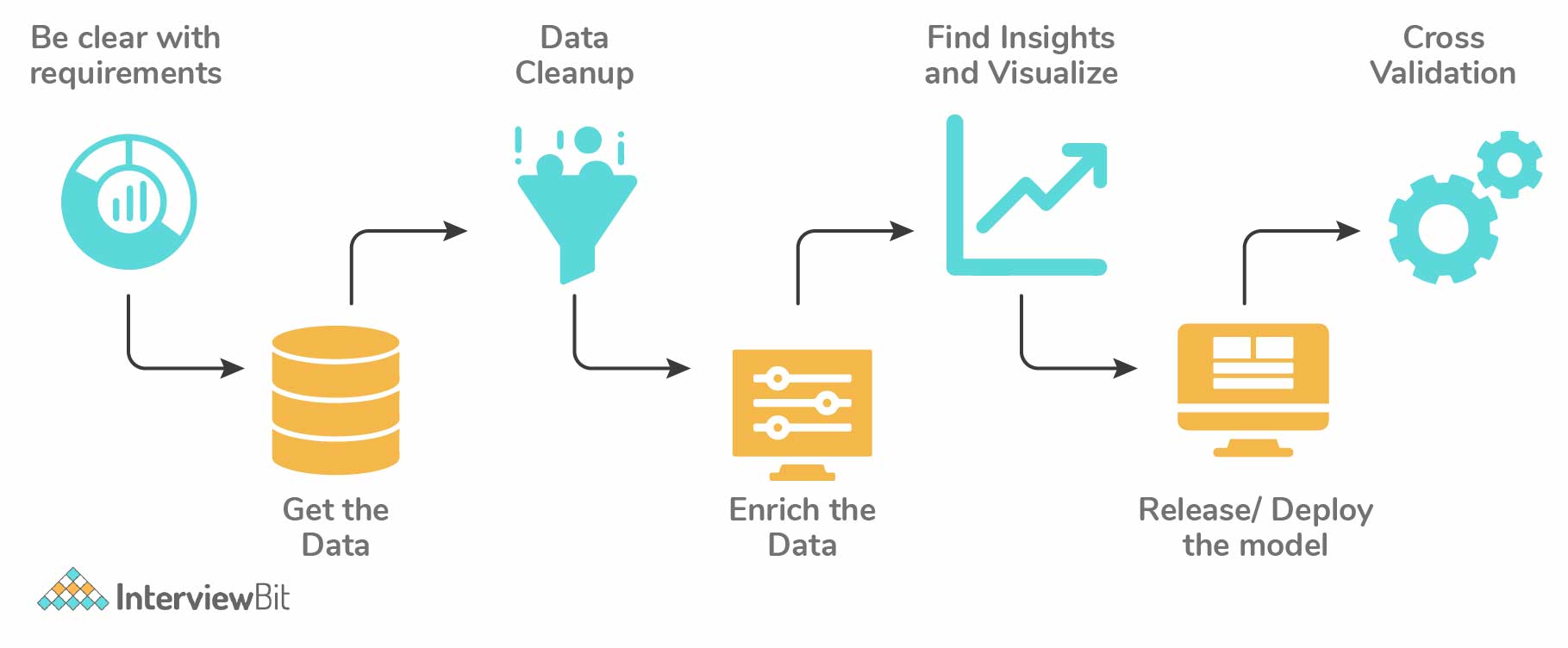
For example, imagine you are developing a model intended to detect fraud for a large bank. Some fraud you know about, but other instances of fraud are slipping by without your knowledge. You can label the dataset with the fraud instances you’re aware of, but the rest of your data will remain unlabelled. You can use a semi-supervised learning algorithm to label the data, and retrain the model with the newly labelled dataset. Then, you apply the retrained model to new data, more accurately identifying fraud using supervised learning techniques. However, there is no way to verify that the algorithm has produced labels that are 100% accurate, resulting in less trustworthy outcomes than traditional supervised techniques.

# Supervised and Unsupervised Learning

|  |  |
| --- | --- |
| **Supervised** | **Unsupervised** |
| Requires labelled data | Works on unlabelled data |
| Used to create models that can be employed to predict or classify things | Used to extract meaningful information out of large volumes of data |
| E.g., Linear Regression, Decision Tree, logistic regression, SVM, Naïve Bayes, K-nearest neighbour, Neural networks | E.g., K-means clustering, Apriori algorithm, hierarchical clustering, Anomaly Detection |
| Has a feedback mechanism | No feedback mechanism |
| Used in Classification and Regression | Dimension reduction |
| E.g., Knowing the height and weight, identifying the gender of the person. Below are the popular supervised learning algorithms. |  |

# Approach for Data Analytics Project

* Thoroughly understand the business requirement/problem
* Explore the given data and analyse it carefully, in case of missing data, get the requirements clarified from the business
* Data clean-up and preparation
* Run the model, build meaningful visualization and analyse the results to get meaningful insights
* Release model implementation, and track the results and performance over a specified period to analyse the usefulness
* Perform cross-validation of the model

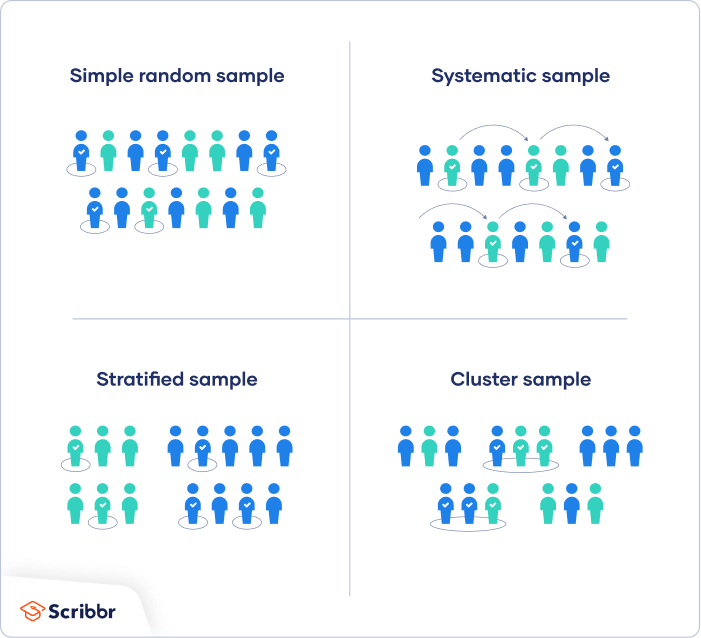


# Data Sampling

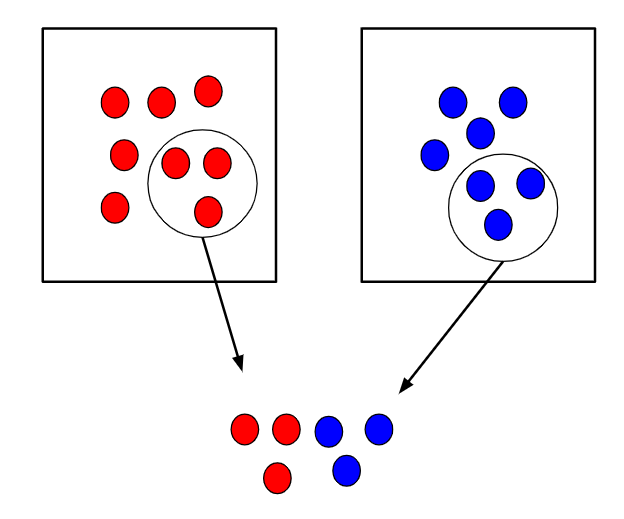
Data analysis cannot be done on the whole volume of data when it involves larger datasets. We take some data samples that can be used for representing the whole population and then perform analysis on it. We should be careful to take sample data out of the huge dataset so that it truly represents the entire dataset. It is one of the most important factors which decides the accuracy of a result.

Mainly there are two types of sampling techniques:

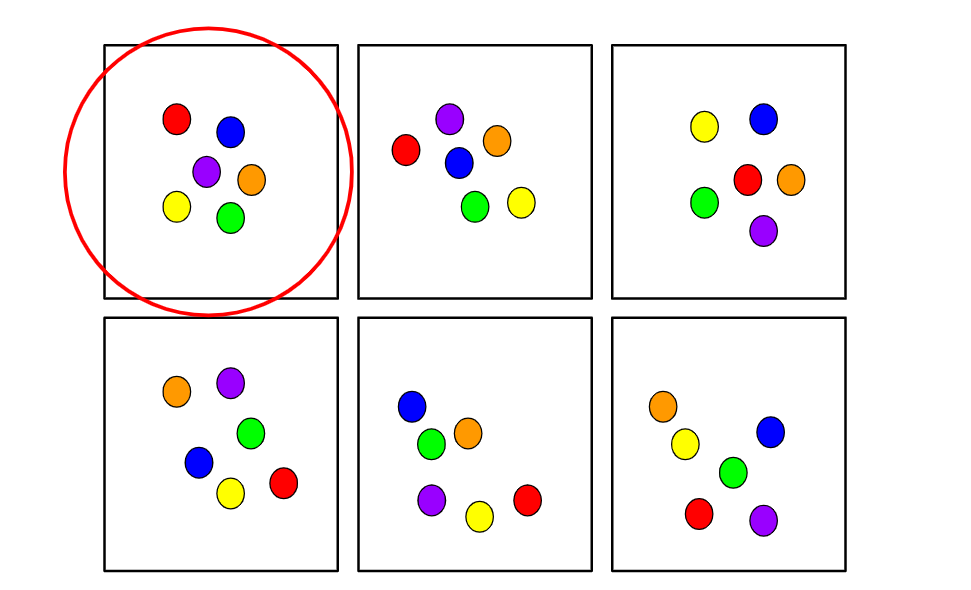
* Probability Sampling: It involves random selection which gives chance to every element to be selected
  + 1. Simple Random Sampling



* + 1. Stratified Sampling: Population is split into sub-populations. It allows you to conclude more precise results by ensuring that every sub-population is represented in the sample.



* + 1. Systematic Sampling: Elements are chosen from an ordered sampling frame. The list is advanced in a circular fashion so once you reach the end of the list, it is progressed from top again.
    2. Cluster Sampling: Each of the sampling units is a collection of clusters of elements. It involves dividing the sample population into separate groups, called clusters. Analysis is conducted on data from the sampled clusters.



* + 1. Multi Stage Sampling
* Non-Probability Sampling: It follows non-random selection which means the selection is done based on your ease or any other required criteria.
  + 1. Convenience Sampling: Method where data is collected from an easily accessible group.
    2. Purposive Sampling: Also known as Judgemental sampling, is where the researchers use their expertise to select a sample that is useful or relevant to the purpose of the research.
    3. Quota Sampling
    4. Snowball Sampling: Used where the population is difficult to access. It can be used to recruit individuals via other individuals.



# Data resampling

Method to sample data for improving accuracy and quantify the uncertainty of population parameters.

* It is done to ensure the model is good enough by training the model on different patterns of a dataset to ensure variations are handled
* It is done in cases where models need to be validated using random subsets or when substituting labels on data points while performing tests

# Parametric & non-parametric models

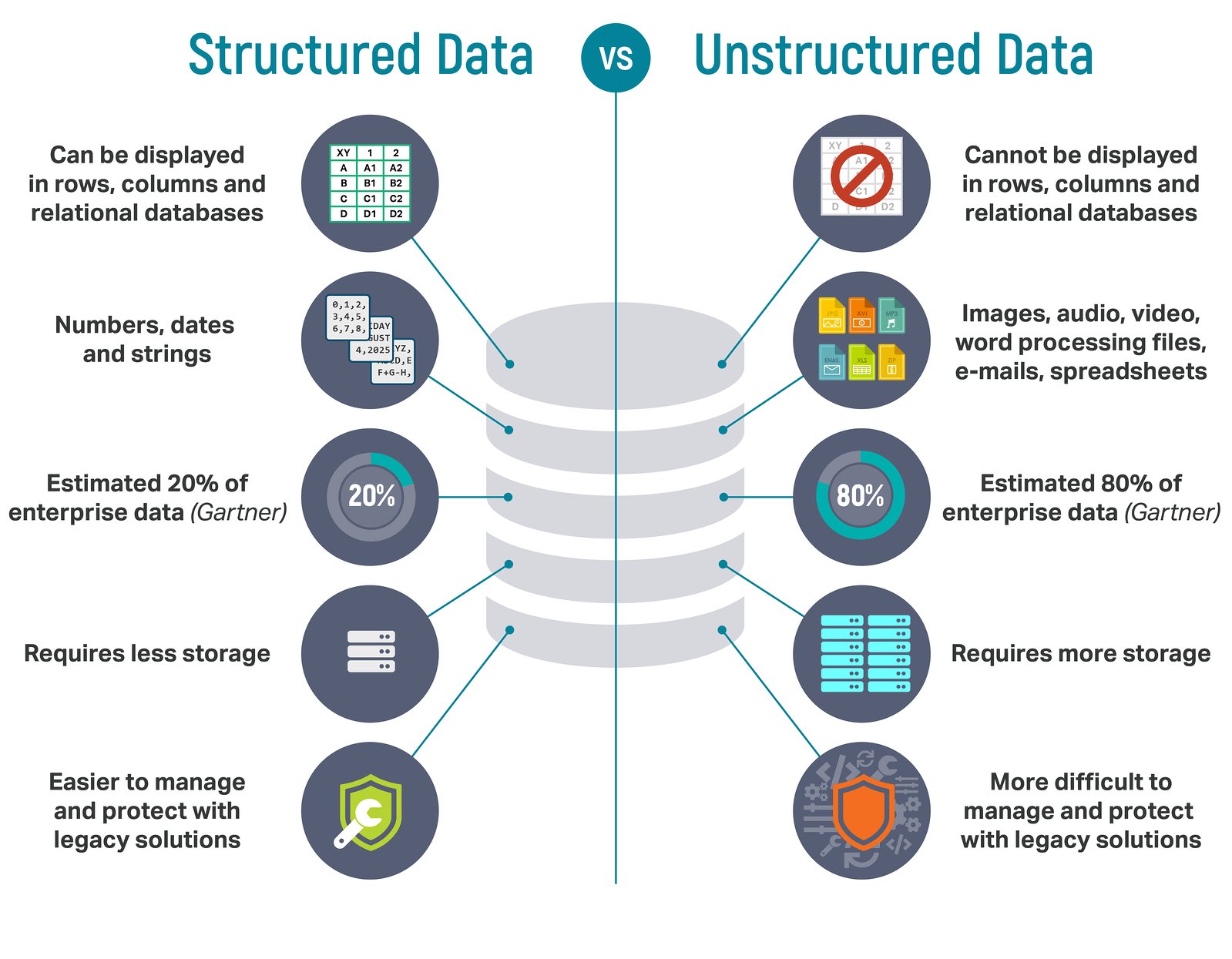
Parametric models are those with a finite number of parameters. To predict new data, you only need to know the parameters of the model. Examples include linear regression, logistic regression, and linear SVMs.

Non parametric models are those with an unbounded number of parameters, allowing for more flexibility. To predict new data, you need to know the parameters of the model and the state of the data that has been observed. Examples include DT, k-nearest neighbours.

# Structured and Unstructured Data

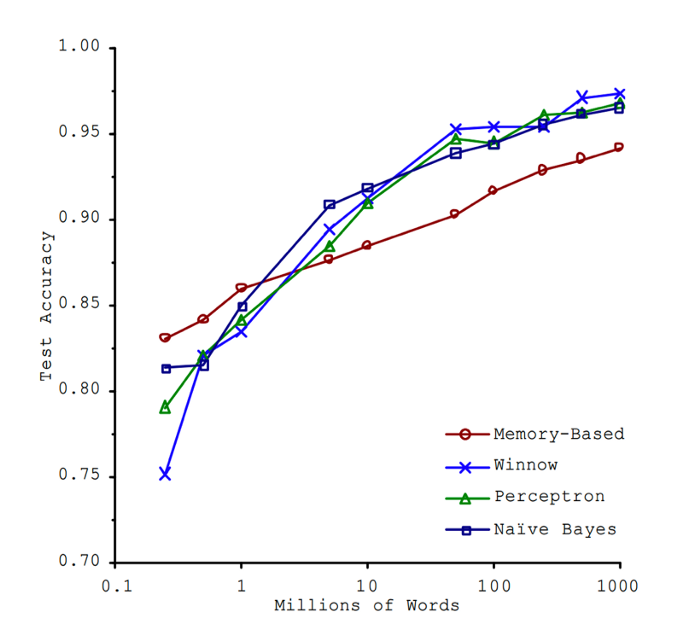
Structured data is data that has been predefined and formatted to a set structure before being placed in data storage, which is often referred to as **schema-on-write**. Example of structured data in the relational database: the data has been formatted into precisely defined fields, such as credit card numbers or address, in order to be easily queried with SQL.

Unstructured data is data stored in its native format and not processed until it is used, which is known as **schema-on-read**. It comes in a myriad of file formats, including email, social media posts, presentations, chats, IoT sensor data, and satellite imagery.



# What is Unreasonable Effectiveness of Data?

The size of the dataset used to train the model mattered far more than the choice of ML approach. And, the performance differences between the models became very small as the dataset grew large. From practical standpoint, the more data is almost always better and ‘more’ gets measured in orders of magnitude.

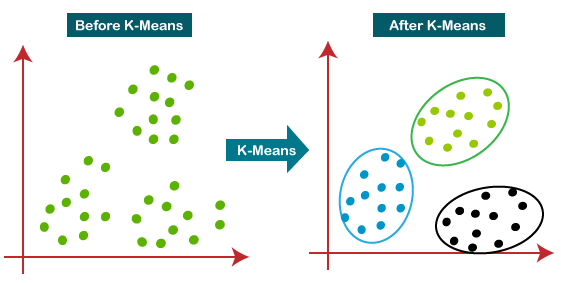


# K-means clustering

Unsupervised learning method. It is the technique of classifying data using a certain set of clusters which is called K clusters.

The two methods to calculate the optimal value of k in k-means are:

* Silhouette score method
* Elbow method: We need to calculate the Within-Cluster-Dum of Squared Errors (WSS) for different k values. The WSS is described as the sum of the squares of the distance between each data value and its centroid. Select the k for which the WSS error starts to become negligible.



# K-Nearest Neighbours

A classification algorithm, which is a subset of supervised learning. K-NN is the number of nearest neighbours used to classify or (predict in case of continuous variable/regression) a test sample.

# Naïve Bayes Classification and Gaussian Naïve Bayes

Bayes theorem finds the probability of an event occurring given the probability of another event has already occurred.



Conditions for Naïve Bayes Classification:

* We assume that no pair of features are dependent. The features are assumed to be independent
* Each feature is given the same weight. None of the attributes is irrelevant and assumed to be contributing equally to the outcome.

Naive Bayes is called naive because it assumes that each input variable is independent. This is a strong assumption and unrealistic for real data; however, the technique is very effective on a large range of complex problems.

Gaussian Naïve Bayes

Continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution.

Standard Naïve Bayes only supports categorical features, while Gaussian Naïve Bayes only supports continuously valued features.

# Bias

Bias is an error introduced in the model because of oversimplification of a machine learning algorithm. It leads to overfitting. Types of biases include

* Selection Bias
* Under coverage Bias - bias that occurs when some members of the population are inadequately represented in the sample
* Survivorship Bias - it is the logical error of focusing on aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.
* Recall Bias
* Exclusion Bias
* Observer Bias

Some of the widely used low and high bias ML algorithms are

Low bias – Decision tress, Support Vector Machines, k-Nearest Neighbours, etc.

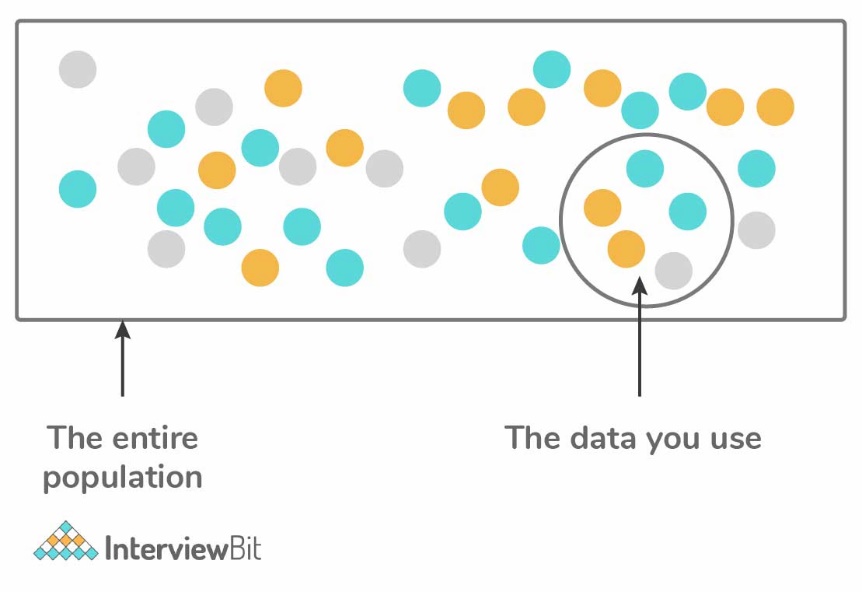
High bias – Linear Regression, Logistic Regression, Linear Discriminant Analysis, etc.

# Selection Bias

Also called as Selection effect, it denotes a situation when selected individuals or a group within a study differ in a manner from the population of interest that they give systematic error in the outcome. Selection Bias leads to skewing results, it can lead to false insights about a particular population group in a study.

* Sampling Bias: In a random population, some members of the population have fewer chances of getting included than others, resulting in a biased sample.
* Time interval bias: Trials may be stopped early if we reach any extreme value but if all variables are of similar invariance, the variables with the highest variance have a higher chance of achieving the extreme value.
* Data Bias: It is when specific data is selected arbitrarily and the generally agreed criteria is not followed.
* Attrition: The loss of the participants which did not complete the trial
* Observer selection: It is a kind of discrepancy or detection bias

When there is no randomization achieved while picking a part of the dataset for analysis. Selection Bias says that the sample analysed does not represent the whole population meant to be analysed.

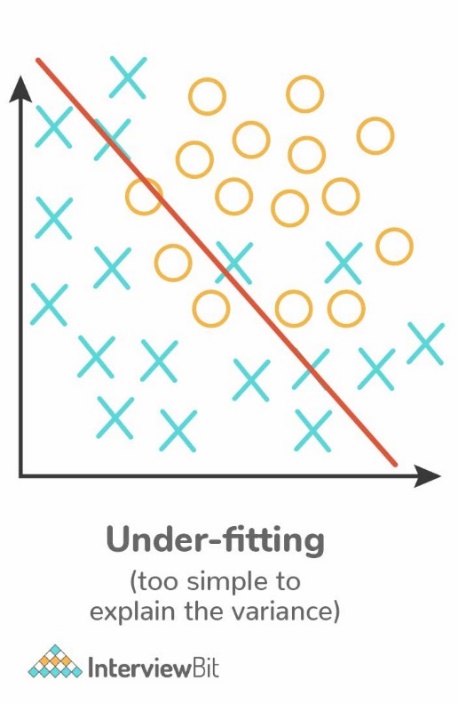
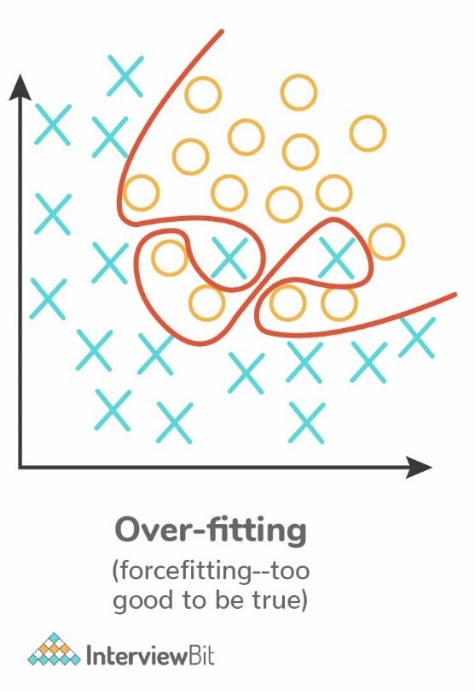


# Variance

Variance is a type of error that occurs when the model ends up being too complex and learns features from the data along with the noise that exists in it. This kind of error can occur if the algorithm used to train the model has high complexity, even though the data and the underlying patterns and trends are quite easy to discover. This makes the model very sensitive, that performs well on the training dataset but poorly on the testing dataset. Variance generally leads to poor accuracy in testing and results in overfitting.

# Overfitting and Underfitting

* Overfitting happens when the model is too robust, low bias and high variance.
* Underfitting happens when the model is too simple, high bias and low variance.



Techniques to reduce Overfitting

* Keep the model simple – take fewer variables into account, thereby removing some of the noise in the training data
* Use cross-validation
* Stop early while training
* Use random dropouts
* Increase training data
* Use regularization such as LASSO, that penalize certain model parameters

Techniques to reduce Underfitting

* Increase model complexity
* Increase the number of features
* Remove noise from the data
* Increase the number of training epochs

# Long and Wide format data

|  |  |
| --- | --- |
| **Long format Data** | **Wide-Format Data** |
| Each row of the data represents the one-time information of a subject. Each subject would have its data in different/ multiple rows. | The repeated responses of a subject are part of separate columns. |
| Data is recognized by considering rows as groups. | Data is recognized by considering columns as groups. |
| Used in R analyses and to write into log files after each trial. | Used in stats packages for repeated measures ANOVAs. |



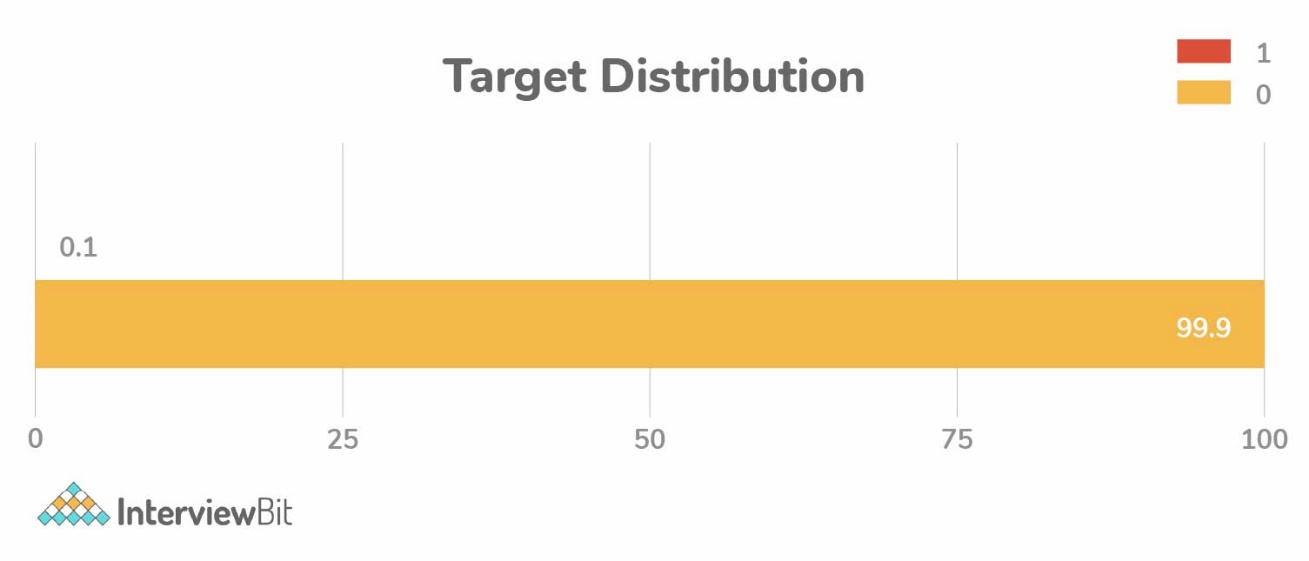
# Imbalanced Data

When data is distributed unequally across different categories leading to error in model performance and inaccuracy in results. One class label has higher observations than the other comparatively.

Techniques to balance imbalanced data include, increasing the sample number of minority classes. The number of samples can be decreased for majority classes. Following are the approaches followed to balance data

* Perform K-fold cross-validation correctly: Cross-Validation needs to be applied properly while using over-sampling. The cross-validation should be done before over-sampling because if it is done later, then it would be like overfitting the model to get a specific result.
* Training set resampling: It is also possible to balance data by working on getting different datasets and this can be achieved by resampling. There are two approaches followed under-sampling that is used based on the use case and the requirements:
  + 1. Under-sampling: This balances the data by reducing the size of the abundant class and is used when the data quantity is sufficient. By performing this, a new dataset that is balanced can be retrieved and this can be used for further modelling.
    2. Over-sampling: This is used when data quantity is not sufficient. This method balances the dataset by trying to increase the samples size. Instead of getting rid of extra samples, new samples are generated and introduced by employing the methods of repetition, bootstrapping, etc.
* Use the right evaluation metrics
  + 1. Specificity / Precision
    2. Sensitivity
    3. F1 score
    4. MCC (Matthews Correlation Coefficient): It represents the correlation coefficient between observed and predicted binary classifications.
    5. AUC (Area Under the Curve)

For example, consider the below graph that illustrates training data: Here, if we measure the accuracy of the model in terms of getting "0"s, then the accuracy of the model would be very high -> 99.9%, but the model does not guarantee any valuable information. In such cases, we can apply different evaluation metrics as stated above.



# KPI

Key Performance Indicator, measures how well the business achieves its objectives. An example of KPI in an organization is the expense ratio.

# Lift

Indicates how good the model is at prediction versus if there was no model.

# Robustness

Represents the system’s capability to handle differences and variances effectively.

# DOE

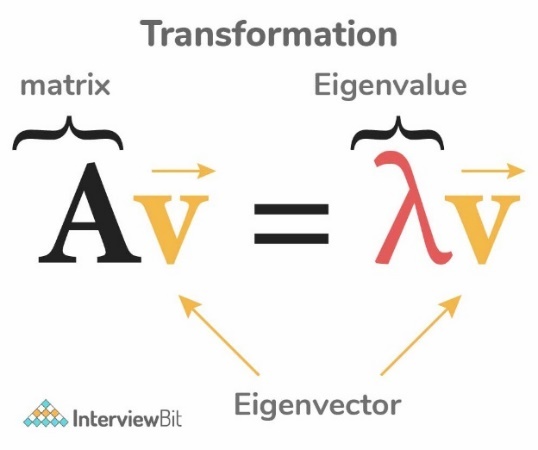
Design of Experiments, represents the task design aiming to describe and explain information variation under hypothesized conditions to reflect variables.

# Eigenvectors and Eigenvalues

Eigenvectors are column vectors or unit vectors whose length/magnitude is equal to 1. These are used in PCA for gathering valuable insights from the given matrix.

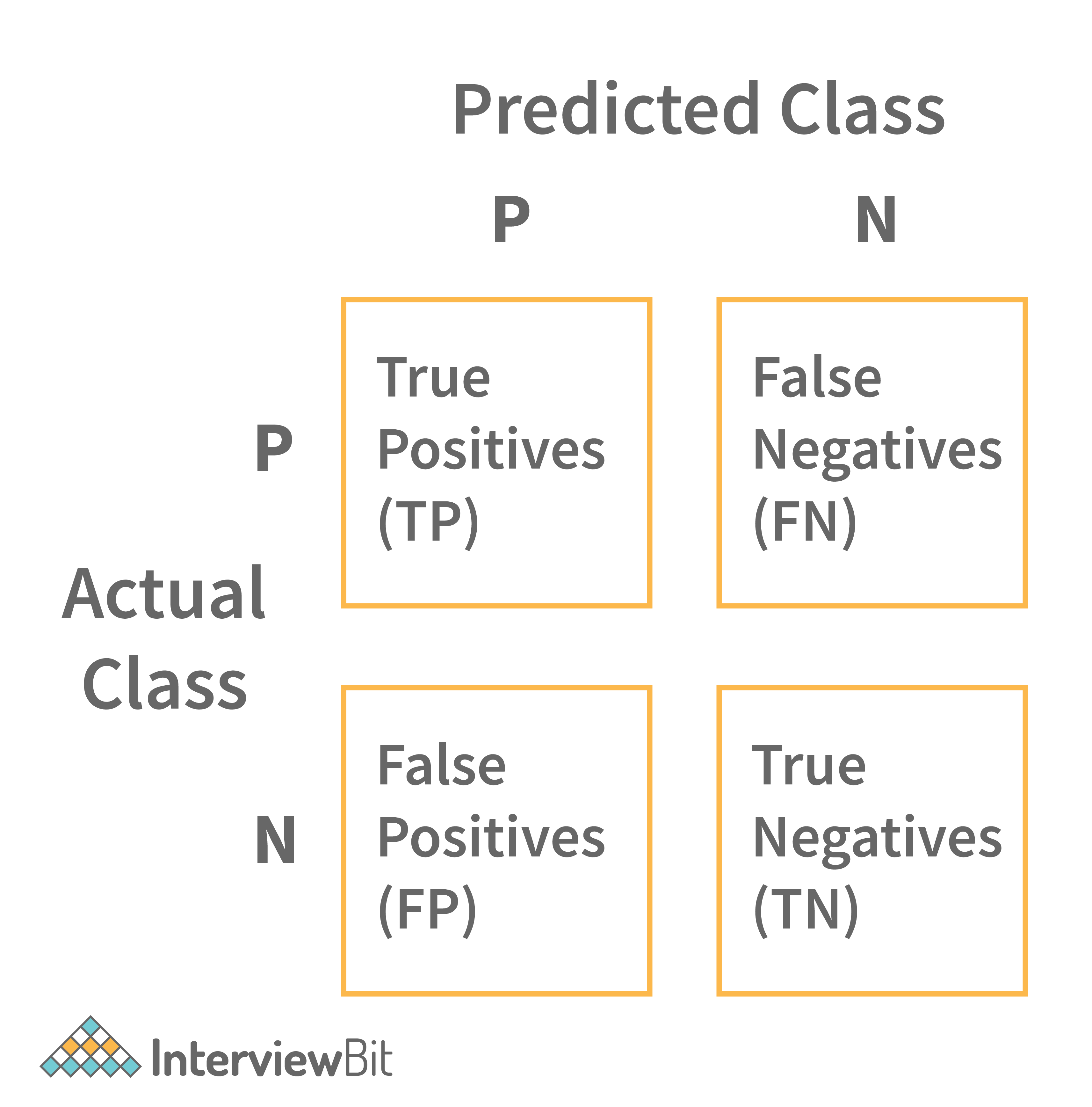
Eigenvectors depict the direction in which a linear transformation moves and acts by compressing, flipping, or stretching. They are used to understand linear transformations and are generally calculated for a correlation or covariance matrix.   
The eigenvalue is the strength of the transformation in the direction of the eigenvector.   
An eigenvector’s direction remains unchanged when a linear transformation is applied to it.

Eigenvalues are coefficients that are applied on eigenvectors which give these vectors different values for length or magnitude. A matrix can be decomposed into Eigenvectors and Eigenvalues and this process is called Eigen decomposition.



# Confusion matrix

Table used to estimate the performance of a model.



* True Positive: Positive Prediction is correct/ Observation is positive, and is predicted to be positive.
* False Positive: Positive Prediction is incorrect/ Observation is negative, but is predicted positive.
* True Negative: Negative Prediction is correct/ Observation is negative, and is predicted to be negative.
* False Negative: Negative Prediction is incorrect/ Observation is positive, but is predicted negative.

Misclassification rate/Error rate

Sensitivity is the measure of True Positive Rate, also called as Recall.

Specificity is the measure of true negative rate.

Recall

Precision is the measure of a positive predicted value.

F-score is the harmonic mean of precision and recall.

High recall indicates the class is correctly recognized (small number of FN).

Low recall indicates the class is incorrectly recognized (large number of FN).

High recall, low precision means that most of the positive examples are correctly recognized (low FN), but there are a lot of false positives.

Low recall, high precision shows that we miss a lot of positive examples (high FN), but those we predict as positive are indeed positive (low FP).

# Type 1 and Type 2 Errors

A Type 1 Error is a False Positive, which means that a positive result was predicted but the actual result is negative.

A Type 2 Error is a False Negative, which means that a negative result was predicted but the actual result is positive.

# False Positive and False Negative Importance Case

* When both FP and FN are equally important

In Banking fields: Lending loans are the main sources of income to the banks. But if the repayment rate isn’t good, then there is a risk of huge losses instead of any profits. So, giving out loans to customers is a gamble as banks can’t risk losing good customers but at the same time, they can’t afford to acquire bad customers. This case is a classic example of equal importance in false positive and false negative scenarios.

* When a FP is important than a FN

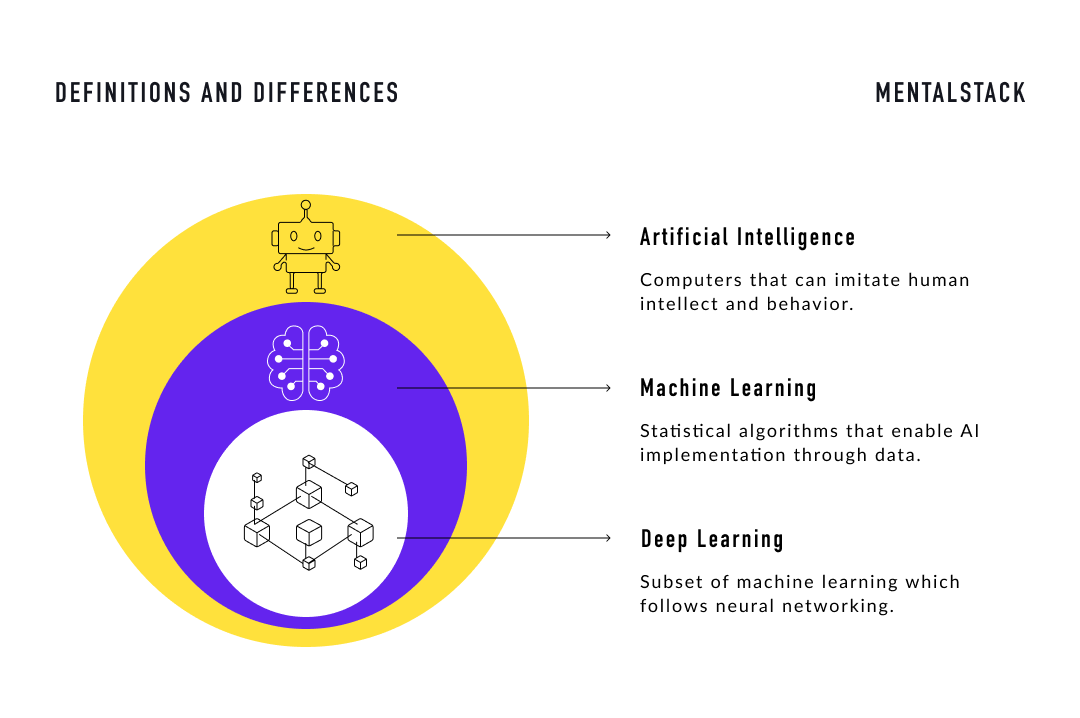
In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn’t have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

# Deep Learning

Multiple layers of processing to extract high level features from the data.

Kind of problems that can be solved by using deep learning:

* Image recognition
* Object detection
* Natural Language Processing



# Neural Networks

Networks that learns the patterns from the data and predicts the output for new data.

# Lazy Learning Algorithms

In machine learning, lazy learning is a learning method in which generalization of the training data is, in theory, delayed until a query is made to the system, as opposed to eager learning, where the system tries to generalize the training data before receiving queries. E.g., KNN.

Training is fast, but prediction is slow. Also known as instance-based learning. Lazy classifiers are very useful when working with large datasets that have a few attributes.

# Weights Initialization in Network

We can either initialize the weights to zero or assign them randomly.

Initializing the weights randomly gives better accuracy to the model since every neuron performs different computation.

# RNN

Recurrent Neural Networks are an artificial neural network that works on a sequence of data, time series and others.

# Gradient

Measure of change in output with little change in input. Measure of change in weights with respect to change in error. Mathematically represented as the slope of a function.

# Exploding Gradients

Exponentially growing error gradients that update the neural network weights to a great extent, causes an overflow, and results in ‘Nan’ values. Exploding Gradients is the problematic scenario where large error gradients accumulate to result in very large updates to the weights of neural network models in the training stage. In an extreme case, the value of weights can overflow and result in ‘Nan’ values. Hence the model becomes unstable and is unable to learn from the training data.

# Vanishing Gradient

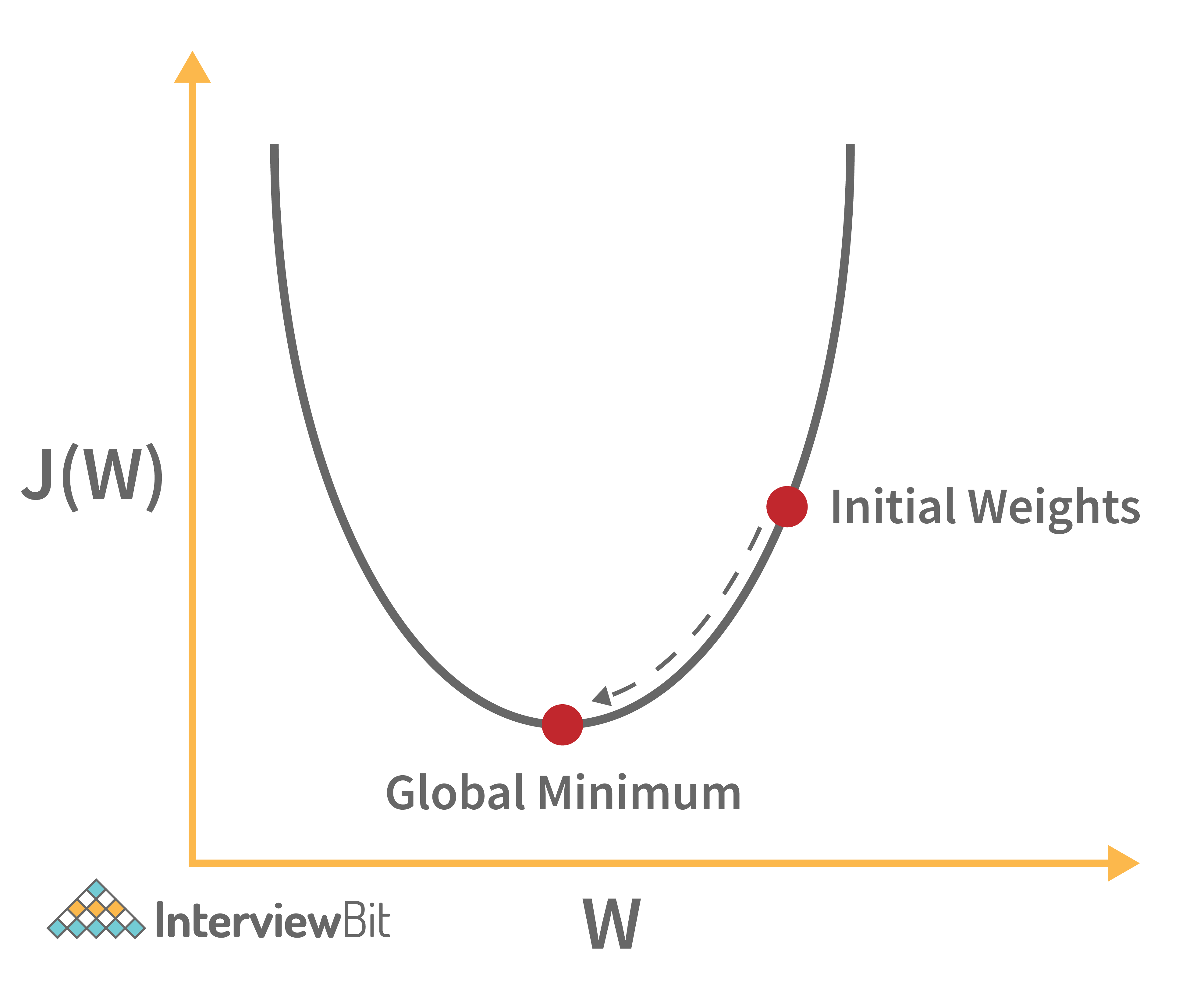
Causes major increase in the training time and poor performance and extremely low accuracy. A condition when the slope is too small during the training of RNN.

# Activation function

It introduces non-linearity in the neural network. This is done to help the leaning process

# Gradient Descent

Minimization function that minimizes the activation function. Gradient descent methods may not converge in some cases, as they reach a local minimum or a local optimum point. This is governed by the data and the starting conditions.



# Batch and Stochastic Gradient Descent

|  |  |
| --- | --- |
| **Batch Gradient Descent** | **Stochastic Gradient Descent** |
| Helps in computing the gradient using the complete data set | Helps in computing the gradient using only a single sample |
| Takes time to converge | Takes less time to converge |
| Volume is large for analysis purposes | Volume is low |
| Update weights infrequently | Update weights frequently |

# SVM Classifier

Support vector machine, used for classification and prediction tasks. SVM consists of a separating plane that discriminates between the two classes of variables. This separating plane is known as hyperplane.

There are two types of SVM classifiers:

* Linear SVM – SVM algorithm predicts a straight hyperplane dividing the two classes. The hyperplane is also called as maximum margin hyperplane.



* Nonlinear SVM



Some of the kernels used in SVM are

* Linear Kernel
* Gaussian Kernel
* Polynomial Kernel
* Laplace RBF Kernel
* Hyperbolic Kernel
* Radial Basis Kernel
* Sigmoid Kernel

Advantages of SVM classifier

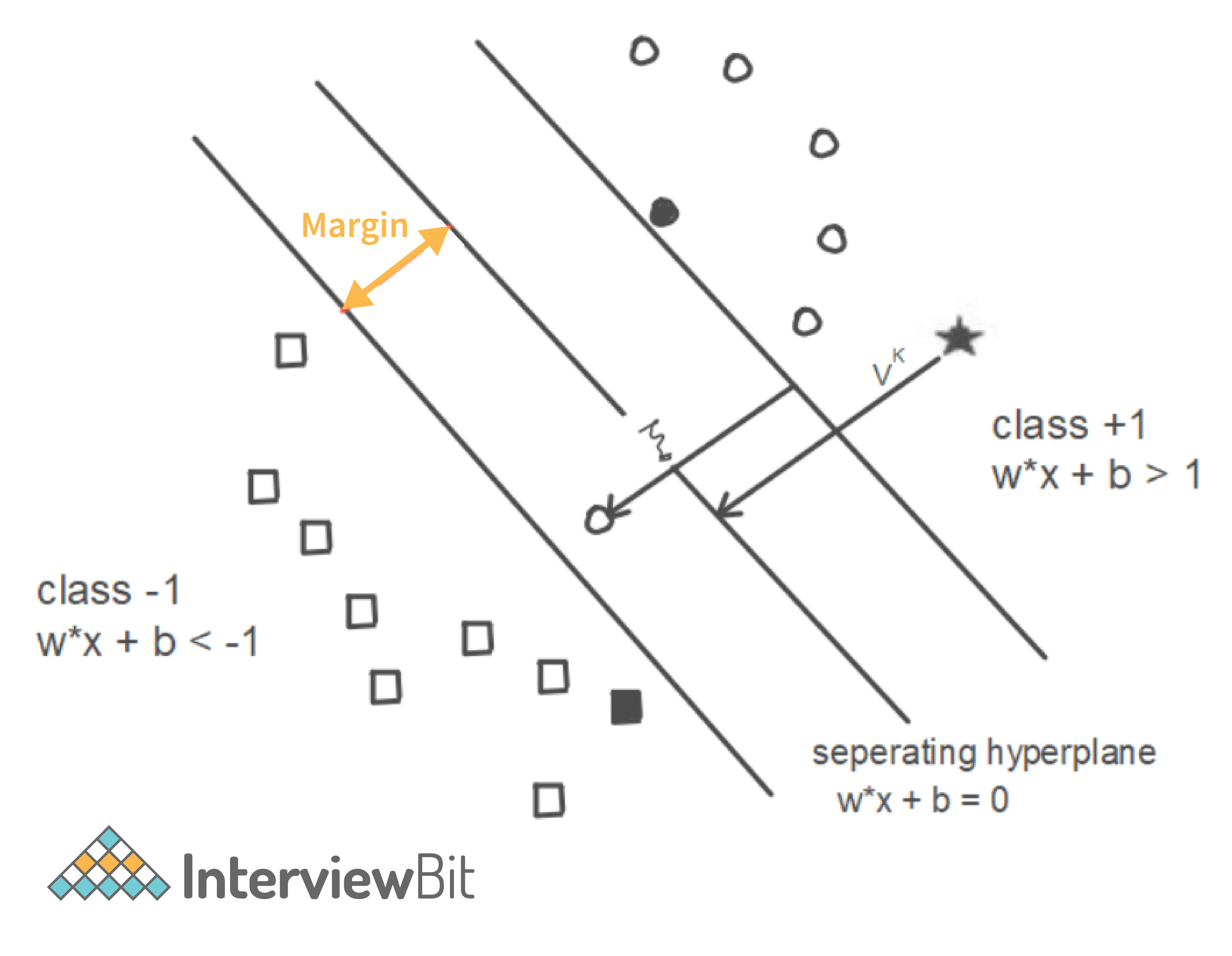
* SVMs are effective when the number of features is quite large.
* Non linear data can also be classified using customized hyperplanes built by using kernel trick

Disadvantages of SVM classifier

* SVM have high algorithmic complexity and extensive memory requirements due to the use of quadratic programming.
* SVMs have good generalization performance, but they can be extremely slow in test phase.

# Support Vectors

Vectors or data points that are nearest to the hyperplane.



# Perceptron

Simplest neural network that contains a single neuron which performs 2 functions. First function is to perform the weighted sum of all the inputs and the second function is an activation function.

# Computational Graph

Dataflow Graph which has a network of nodes where nodes represent operations and the edges represent tensors.

# Cross-Validation

Statistical model validation technique used for improving model’s performance. Model will be trained and tested with rotation using different samples of the training dataset to ensure that the model performs well for unknown data. Most commonly used techniques are:

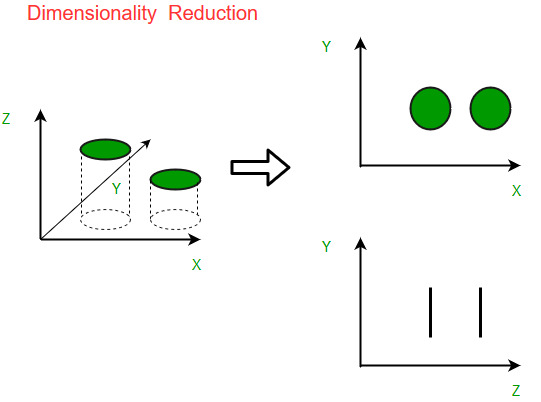
* K-fold method
* Leave p-out method
* Leave-one-out method
* Holdout method



# Star Schema

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

# Dimensionality reduction

Reducing the number of features in a dataset to avoid overfitting and reduce the variance. Dimensionality reduction reduces the dimensions and size of the dataset.

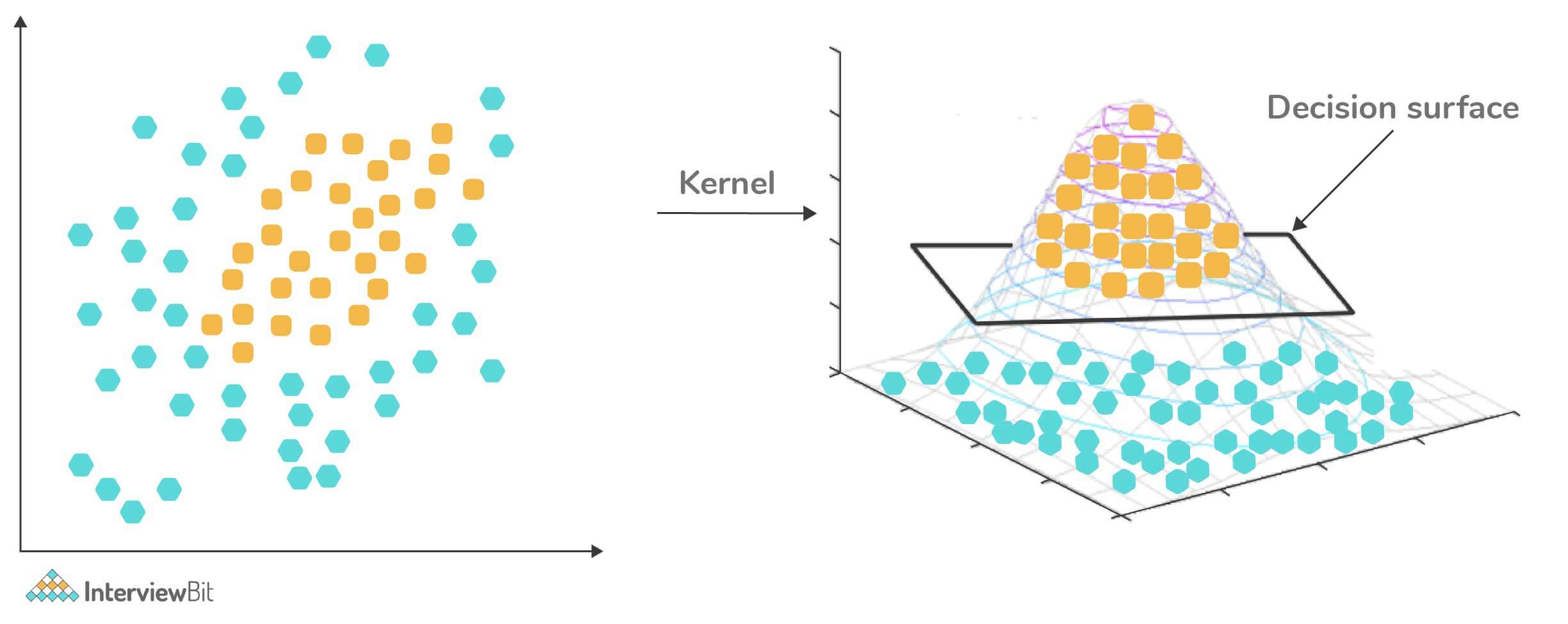
* Reduces the storage space and time for model execution
* Removes the issue of multi-collinearity thereby improving the parameter interpretation of the ML model
* Easier to visualize data when the dimensions are reduced
* Avoids the curse of increased dimensionality

# Handle Missing values

* First, we have to identify the variables with the missing values. If a pattern is identified, it could lead to interesting and meaningful insights. If there are no patterns, substitute the missing values with the median or mean values or we can simply ignore the missing values.
* If the missing value belong to categorical variables, then they are assigned default values such as mean, minimum, and maximum or median. The missing value is assigned the default value. If the data is normally distributed, we give the mean value.
* If 80% of the values are missing for a particular variable, then we would drop the variable instead of treating the missing values.

# Kernel Trick

Kernel functions are generalized dot product functions used for the computing dot product of vectors ‘xx’ and ‘yy’ in high dimensional feature space. Kernel trick is used for solving a non-linear problem by using a linear classifier by transforming linearly inseparable data into separable ones in higher dimensions.



# Identify if a coin is biased

We perform a hypothesis test.

According to the null hypothesis, the coin is unbiased if the probability of head flipping is 50%. Perform the below steps:

* Flip the coin 500 times
* Calculate the p-value
* Compare the p-value against the alpha

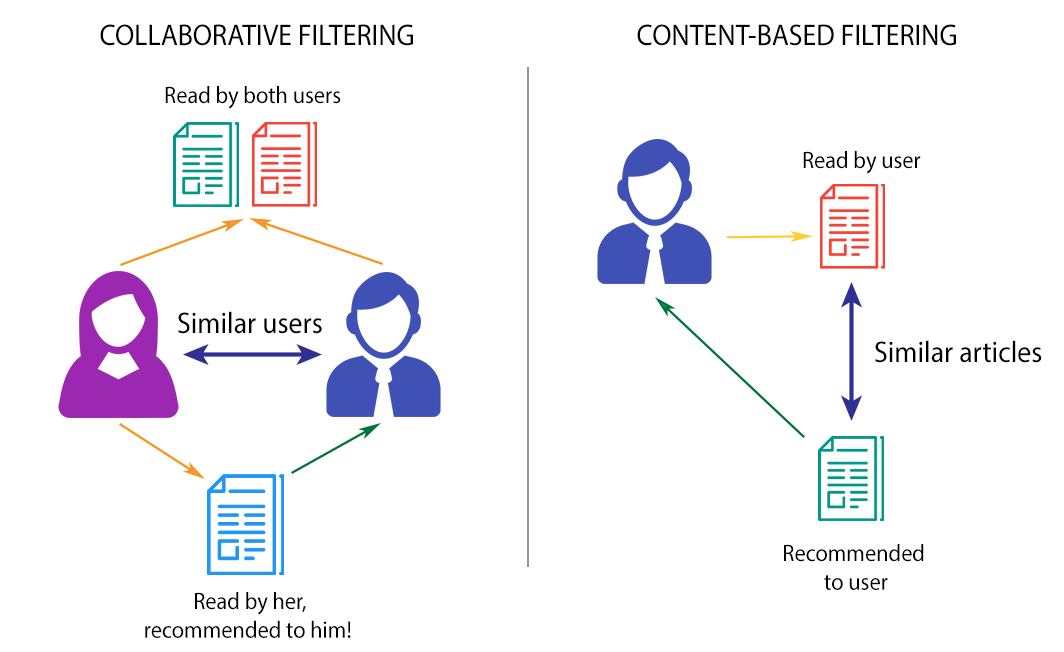
If p-value > alpha: Null hypothesis holds good and the coin is unbiased

If p-value < alpha: Null hypothesis is rejected and the coin is biased

# Treating Outlier in ML

* Drop outlier if it is a garbage value.
* Drop outlier in case of extreme values.
* Try a different model. Data detected as outlier by linear models can be fit by nonlinear models.
* Try normalizing the data. This way, the extreme data points are pulled to a similar range.
* Use algorithms that are less affected by outliers, e.g., random forests

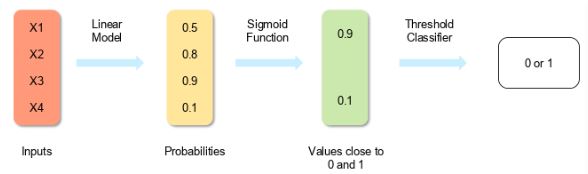
# Recommender system

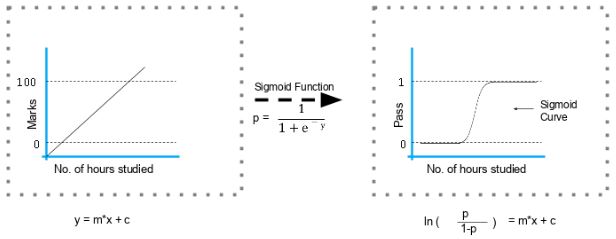
Systems that generate recommendations based on what they know about the user’s tastes from their activities on the platform. These recommendations can also be generated based on what users with a similar taste like watching. Generally, content-based filtering is better than collaborative filtering.

* Collaborative Filtering: Explains the behaviour of other users and their purchase history in terms of ratings, selection etc. The engine makes predictions on what might interest a person based on the preferences of other users. For example, If User A, similar to User B, watched and liked a movie, then that movie will be recommended to User B, and similarly, if User B watched and liked a movie, then that would be recommended to User A.
* Content-based Filtering: Recommendations are generated by making use of the properties of the content that a user is interested in. For example, if a user is watching movies belonging to the action and mystery genre and giving them good ratings, it is a clear indication that the user likes movies of this kind. If shown movies of a similar genre as recommendations, there is a higher probability that the user would like those recommendations as well.

# Logistic Regression

Also known as Logit model. It measures the relationship between the dependent variable and one or more independent variables by estimating probabilities using its underlying logistic function (sigmoid).



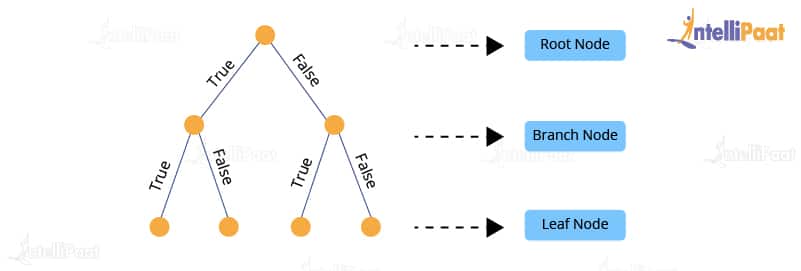


For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e., 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for campaigning of a particular candidate, the amount of time spent in campaigning, etc.

# Decision Tree Algorithm

Supervised learning algorithm that is used for both classification and regression. The dependent variable can be both a numerical value and a categorical value.

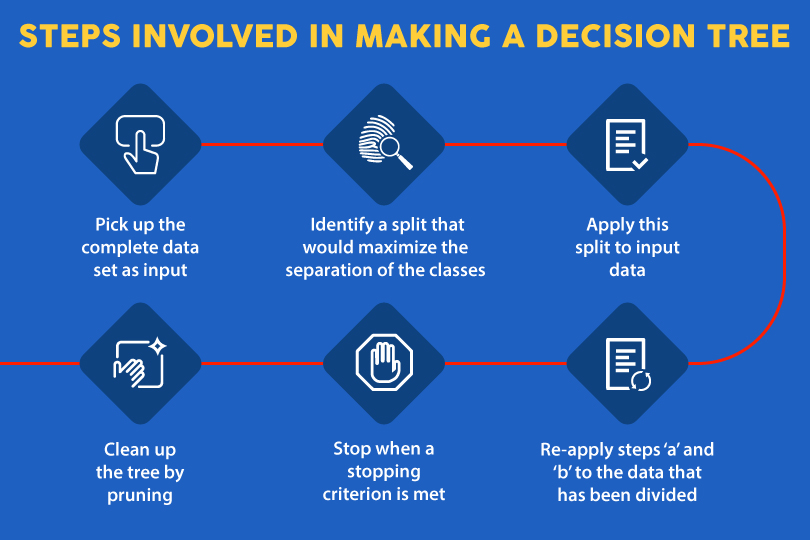
Each node denotes the test on an attribute, each edge denotes the outcome of that attribute, each leaf note holds the class label.



Pruning a DT is the process of removing the sections of the tree that are not necessary or are redundant. It helps to make smaller DT, which performs better, give high accuracy and speed.

Information gain depends on the decrease in entropy after the dataset is split on an attribute. Constructing a DT is always about finding the attributes that return highest information gain.

Entropy is the measure of impurity and randomness in a DT. The entropy of a given dataset tells us how pure or impure the values of the dataset are. For example, suppose we are given a box with 10 blue marbles. Then, the entropy of the box is 0 as it contains marbles of the same colour, i.e., there is no impurity. If we need to draw a marble from the box, the probability of it being blue will be 1.0. However, if we replace 4 of the blue marbles with 4 red marbles in the box, then the entropy increases to 0.4 for drawing blue marbles.

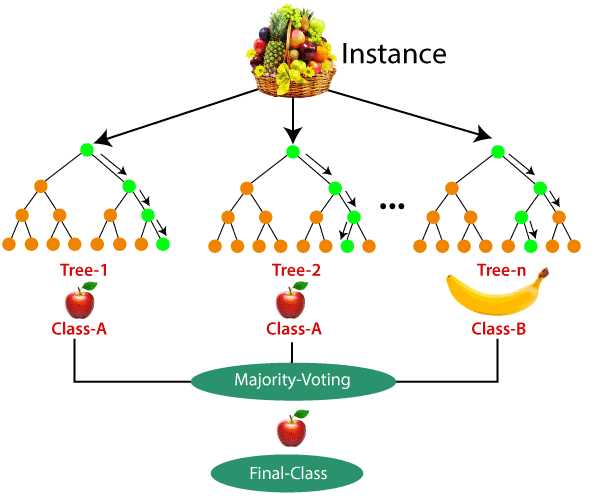


# Steps in making a Decision Tree

* Take the entire dataset as input.
* Calculate entropy of the target variable, as well as the predictor attributes.
* Calculate your information gain of all attributes.
* Choose the attribute with the highest information gain as the root node.
* Repeat the same procedure on every branch until the decision node of each branch is finalized.

# Random Forest

Combination of multiple decision trees. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most votes and in case of regression, it takes the average of outputs by different trees.



Steps in making a Random Forest

Step 1: Randomly select *k* features from a total of *m* features where k << m

Step 2: Among the *k* features, calculate the node D using the best split point.

Step 3: Split the node into daughter nodes using the best split.

Step 4: Repeat steps two and three until leaf nodes are finalized.

Build forest by repeating steps 1-4 for ‘n’ times to create ‘*n’* number of trees.

# Ensemble Learning

Sometimes the datasets are very complex, and it is difficult for one model to be able to grasps the underlying trends in these datasets. In such situations, we combine several individual models together to improve performance. This is called ensemble learning.

# In what scenario would you prefer DT over a RF?

In general, RF is always better than DT. But there are few scenarios

* Explain ability
* Computation
* Features

# Why is it Logistic Regression and not Logistic Classification?

Logistic Regression is closely related to linear regression, it is a generalized linear model. Logistic Regression just uses a sigmoid function and a threshold value to find the probability.

# Bagging and Boosting

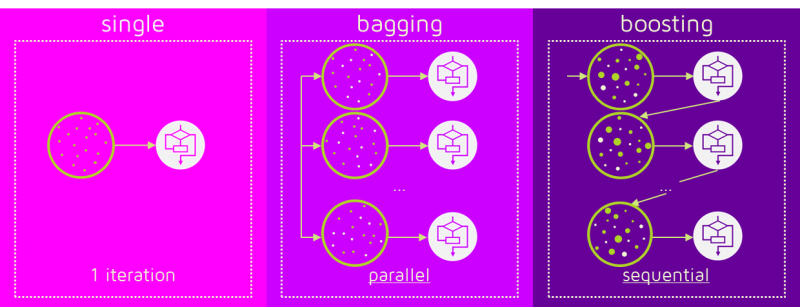
Bagging (bootstrap aggregating) is an ensemble learning method. In this technique, we generate some data using the bootstrap method, in which we use an already existing dataset and generate multiple samples of the N size. This bootstrapped data is then used to train multiple models in parallel, which makes the bagging model more robust than a simple model.

Once all the models are trained, when it’s time to make a prediction, we make predictions using all the trained models and then average the result in the case of regression, and for classification, we choose the result, generated by models, that have the highest frequency.

Boosting is an ensemble learning method. In boosting, we create multiple models and sequentially train them by combining weak models iteratively in a way that training a new model depends on the models trained before it.

In doing so, we take the patterns learned by a previous model and test them on a dataset when training the new model. In each iteration, we give more importance to observations in the dataset that are incorrectly handled or predicted by previous models. Boosting is useful in reducing bias in models as well. Different types of boosting algorithms are

* AdaBoost
* Gradient Boosting
* XGBoost



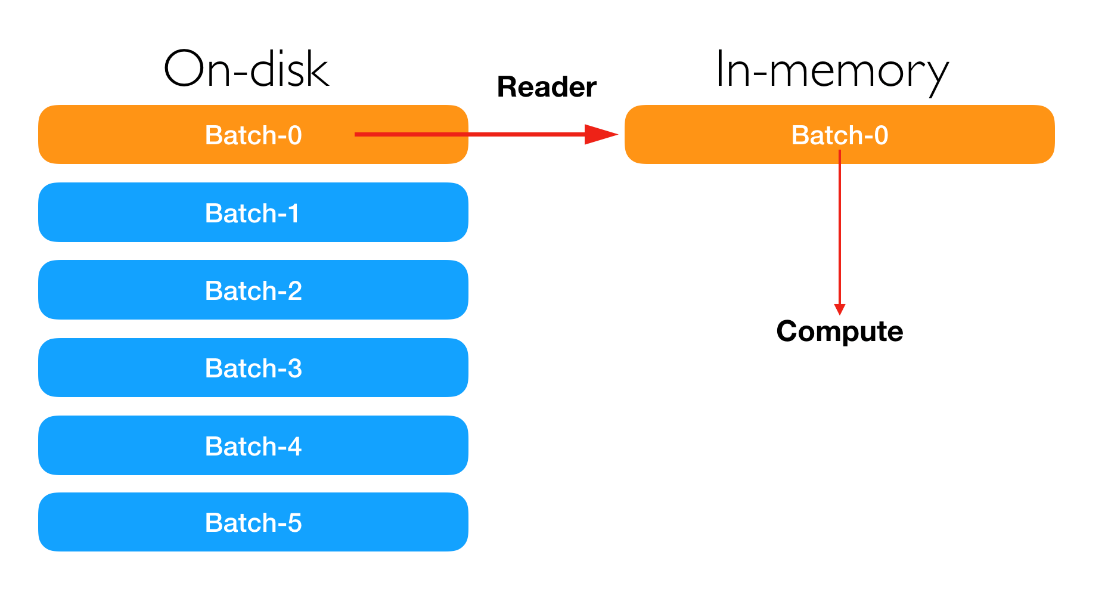
# What is No Free Lunch Theorem?

The No Free Lunch Theorem, abbreviated as NFL, NFLT, is a theoretical finding that suggests all optimisation algorithms perform equally well when their performance is averaged over all possible objective functions.

# Out of Core Learning

Out-of-core learning refers to the machine learning algorithms working with data cannot fit into the memory of a single machine, but that can easily fit into some data storage such as local hard disk or web repository. Your available RAM, the core memory on your single machine, may indeed range from a few gigabytes (sometimes 2 GB, more commonly 4 GB, but we assume that you have 2 GB at maximum) up to 256 GB on large server machines. Large servers are like the ones you can get on cloud computing services such as Amazon Elastic Compute Cloud (EC2), whereas your storage capabilities can easily exceed terabytes of capacity using just an external drive (most likely about 1 TB but it can reach up to 4 TB).

Actually, the ability to learn incrementally from a mini-batch of instances is key to *out-of-core* learning as it guarantees that at any given time there will be only a small amount of data in the main memory. Choose a good size for the mini-batch that balances relevancy and memory footprint could involve some tuning.



# Stacking

Stacking is an ensemble learning method. We can combine weak models that use different learning algorithms. Stacking works by training multiple weak models and then using them together by training another model, called a meta-model, to make predictions based on the multiple outputs of predictions returned by these multiple weak models.

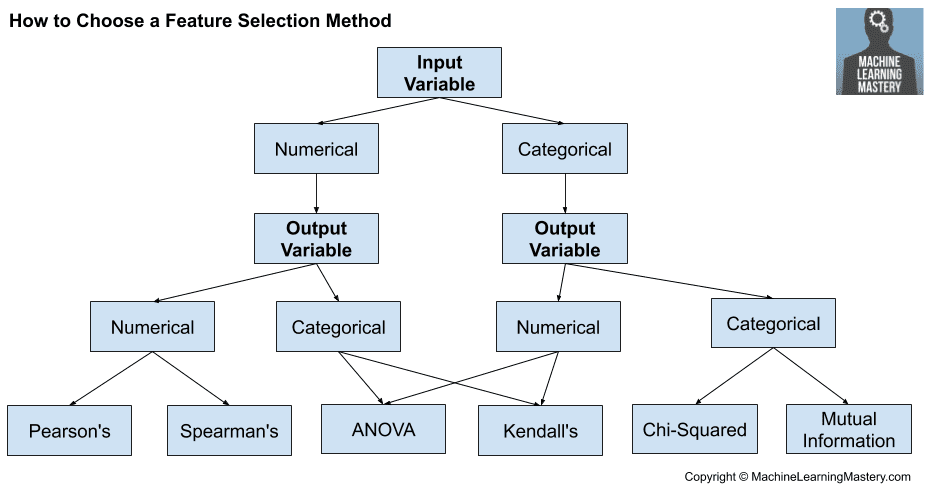
# Feature Vectors

It is an n-dimensional vector of numerical features that represent an object. In ML, feature vectors are used to represent numeric or symbolic characteristics/features of an object in a mathematical way.

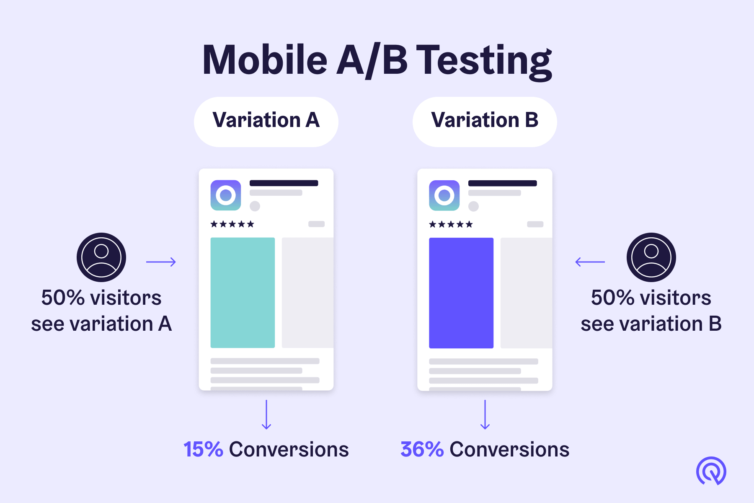
# Feature Selection methods

There are two main methods for feature selection, i.e., filter and wrapper methods

* Filter methods
  + 1. Linear Discrimination Analysis
    2. ANOVA
    3. Chi-Square
    4. Pearson’s Correlation – Variables (both dependent and independent) used for Pearson’s Correlation Coefficient must be quantitative. It will only test for the linear relationship between two variables.
    5. Recursive Feature Elimination
    6. Backward Elimination
    7. Lasso Regression
* Wrapper methods
  + 1. Forward selection: We test one feature at a time and keep adding them until we get a good fit
    2. Backward selection: We test all the features and start removing them to see what works better
    3. Recursive feature selection: Recursively looks through all the different features and how they pair together



# A/B Testing

The goal of A/B testing is to pick the best variant among two hypotheses, the use cases of this kind of testing could be a web page or application responsiveness, landing page redesign, banner testing, marketing campaign performance etc. The first step is to confirm a conversion goal, and then statistical analysis is used to understand which alternative performs better for the given conversion goal.

This is a statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B Testing is to detect any changes to a web page to maximize or increase the outcome of a strategy. A/B testing is used when we wish to test a new feature in a product.

In the A/B test, we give users two variants of the product, and we label these variants as A and B. The A variant can be the product with the new feature added, and the B variant can be the product without the new feature. After users use these two products, we capture their ratings for the product. If the rating of product variant A is statistically and significantly higher, then the new feature is considered an improvement and useful and is accepted. Otherwise, the new feature is removed from the product.

# Chi-Square Test

Chi-squared test, a statistical method, is used by machine learning engineers or statisticians to check the "goodness of fit test". It is used to test if a sample of data came from a population with a specific distribution. Chi-Square test include-

1. Goodness of fit test - Used to determine whether or not a categorical variable follows a hypothesized distribution.
2. Test of independence - Used to determine whether or not there is a significant association between two categorical variables.  
   Let’s suppose you have two random categorical variables X and Y. Now if you want to check if there's any association between X and Y. For example,  
   Researchers want to know if gender(X) is associated with political party(Y) preference in a certain town so they survey 500 voters and record their gender and political party preference.  
   They can perform a Chi-Square Test of Independence to determine if there is a statistically significant association between voting preference and gender.

We can use the Chi-Square test in one of the following situations:  
- When we want to estimate observed distribution matches with the expected distribution, this is also referred to as the goodness of fit test.  
- When we want to estimate whether two random variables are associated or not.

The chi squared test is one type of hypothesis testing, so to solve any problem using the Chi-Square Test we follow the same steps as hypothesis testing.  
  
Hypothesis test steps:  
Define the Null hypothesis (H0)  
Define the alternative hypothesis (H1)  
Design the Test statistic(T)  
Take the T & H0 & H1 find the p-value  
If the p-value is > 5%(5% rule is arbitrary) then we fail to reject h0   
Else reject the H0 and accept the H1

# Normalisation and Standardization

|  |  |
| --- | --- |
| Standardization | Normalization |
| Technique of converting data in such a way that it is normally distributed and has a standard deviation of 1 and a mean of 0. | Technique of converting all data values to lie between 1 and 0. This is also known as min-max scaling |
|  |  |

# Point Estimates

An estimate of the population parameter is given by a particular value called the point estimate. Popular methods used to derive Population Parameter’s Point estimators are – Maximum Likelihood estimator and the Method of Moments.

# Confidence Interval

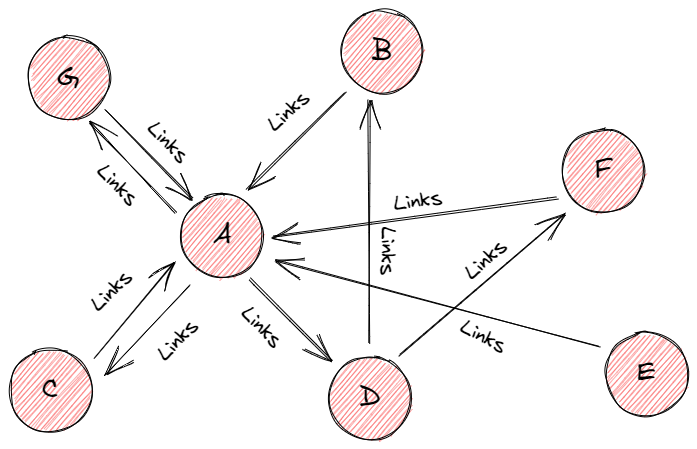
A range of values likely containing the population parameter is given by the confidence interval. The Confidence Coefficient/Confidence level is denoted by , which gives the probability or likeness. The level of significance is given by *α*.

# Markov Chains

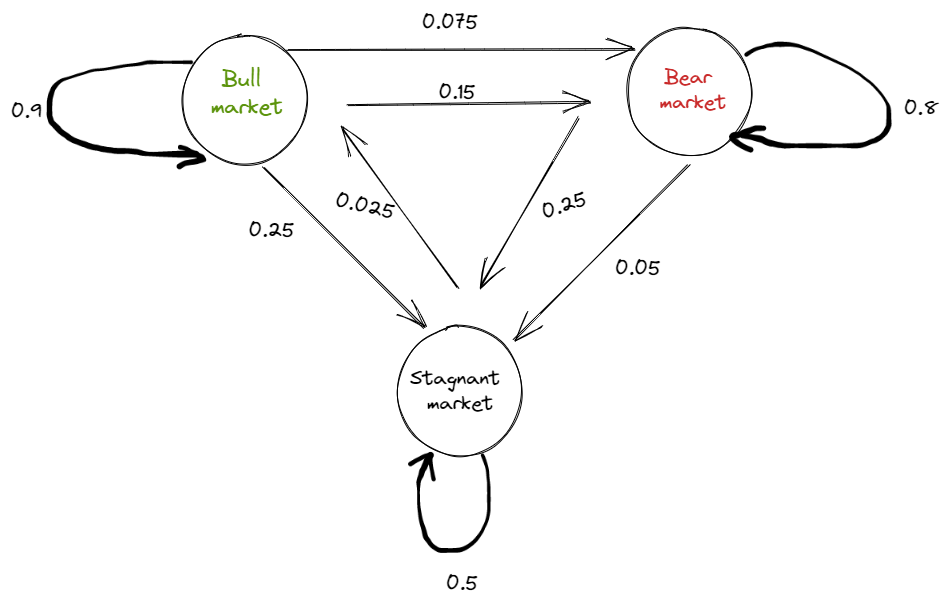
Markov Chains defines that a state’s future probability depends only on its current state. Markov Chain is used in word recommendation. In this system, the model recognizes and recommends the next word based on the immediate previous word.

Markov chains are used to calculate the probability of an event occurring by considering it as a state transitioning to another state or a state transitioning to the same state as before.

* PageRank used by Google – To use the PageRank algorithm, we assume the web to be a directed graph, with web pages acting as nodes and hyperlinks acting as edges. PageRank assigns a value to a page depending on the number of backlinks referring to it. A page that is connected to many other pages earns a higher rank.



* Predicting Market Trends – Markov chains and their associated diagrams may be used to estimate the probability of various financial market climates and so forecast the likelihood of future market circumstances.



# Law of Large Numbers

A theorem that describes the result of performing the same experiment very frequently. It states that the sample mean, sample variance, and standard deviation converge to what we are trying to estimate. If an experiment is repeated independently a large number of times, the average of the individual results is close to the expected value. There are two forms of the law of large numbers, the weak law of large numbers states that as ‘n’ increases, the sample statistic of the sequence converges in probability to the population value.   
The strong law of large numbers describes how a sample statistic converges on the population value as the sample size or the number of trials increases. For example, the sample mean will converge on the population mean as the sample size increases.

An increase in the number of trials in an experiment will result in a positive and proportional increase in the results coming closer to the expected value. As an example, let us check the probability of rolling a six-sided dice three times. The expected value obtained is far from the average value. And if we roll a dice a large number of times, we will obtain the average result closer to the expected value (which is 3.5 in this case).

In a financial context, the law of large numbers indicates that a large entity which is growing rapidly cannot maintain that growth pace forever. In business, it suggests that, as a business expands, the percentage rate of growth becomes increasingly difficult to maintain.

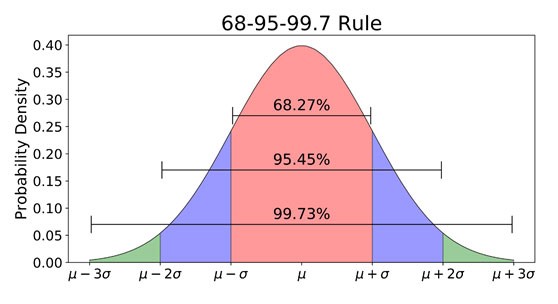
# Normal Distribution

Also known as Gaussian Distribution or Bell-curve distribution, when the data is distributed around a central value, with mean = median, and is in the form of a bell-shaped curve. This kind of distribution has no bias either to the left or to the right.

Exponential distributions do not have a log-normal distribution or a Gaussian Distribution. Any type of data that is categorical will not have this distribution.

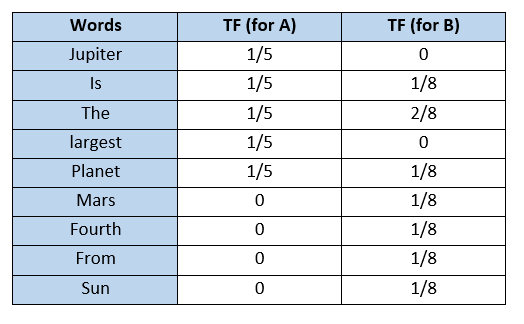
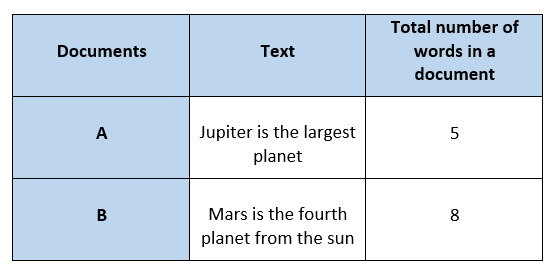
Example: Time until the next earthquake etc.

* Symmetrical (left and right halves are mirror images)
* Unimodal (only one mode)
* Mean, Mode, and Median are all located in the centre
* Central tendency
* Asymptotic

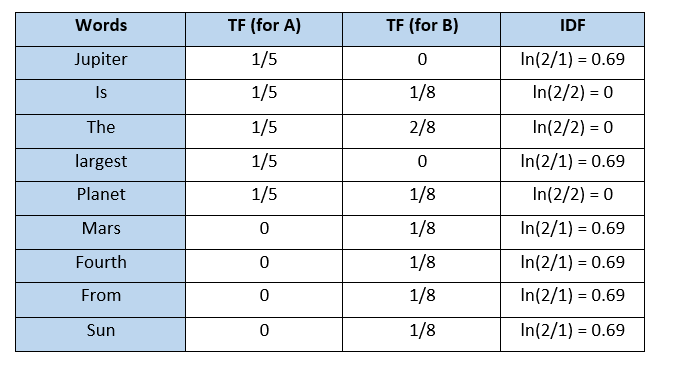


# TF/IDF Vectorization

Term Frequency-Inverse Document Frequency. It is a numerical measure that allows us to determine how important a word is to document in a collection of documents called a corpus. TF/IDF is often used in text mining and information retrieval.

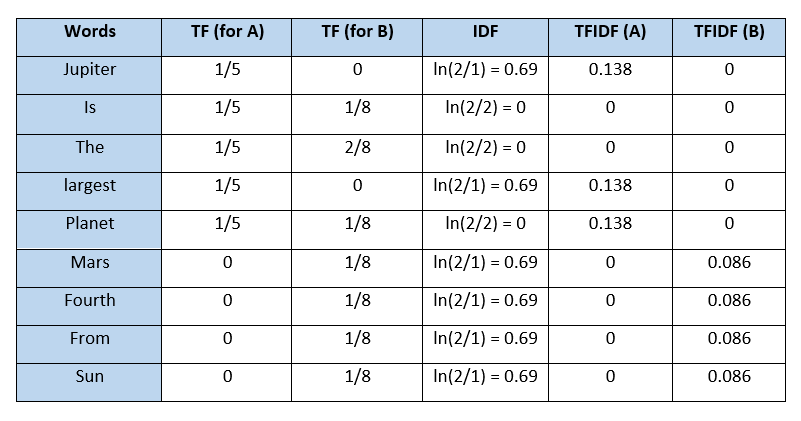


Inverse Document Frequency (IDF) is the measure of the importance of a word. IDF provides weightage to each word based on its frequency in the corpus D. IDF of a word (w) is defined as



The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

TFIDF gives more weightage to the word that is rare in the corpus (all the documents). TFIDF provides more importance to the word that is more frequent in the document.



One problem with TFIDF is that it is unable to capture the semantics. For example, funny and humorous are synonyms, but TFIDF does not capture that. Moreover, TFIDF can be computationally expensive if the vocabulary is vast.

# Dropout

It is used for dropping out hidden and visible units of a network on a random basis. They prevent overfitting.

# Epoch, Batch and Learning Rate

Epoch represents one iteration over the entire dataset.

A batch is a series of broken-down collections of the dataset, which help pass the information into the system. It is used when the developer cannot pass the entire dataset into the neural network at once.

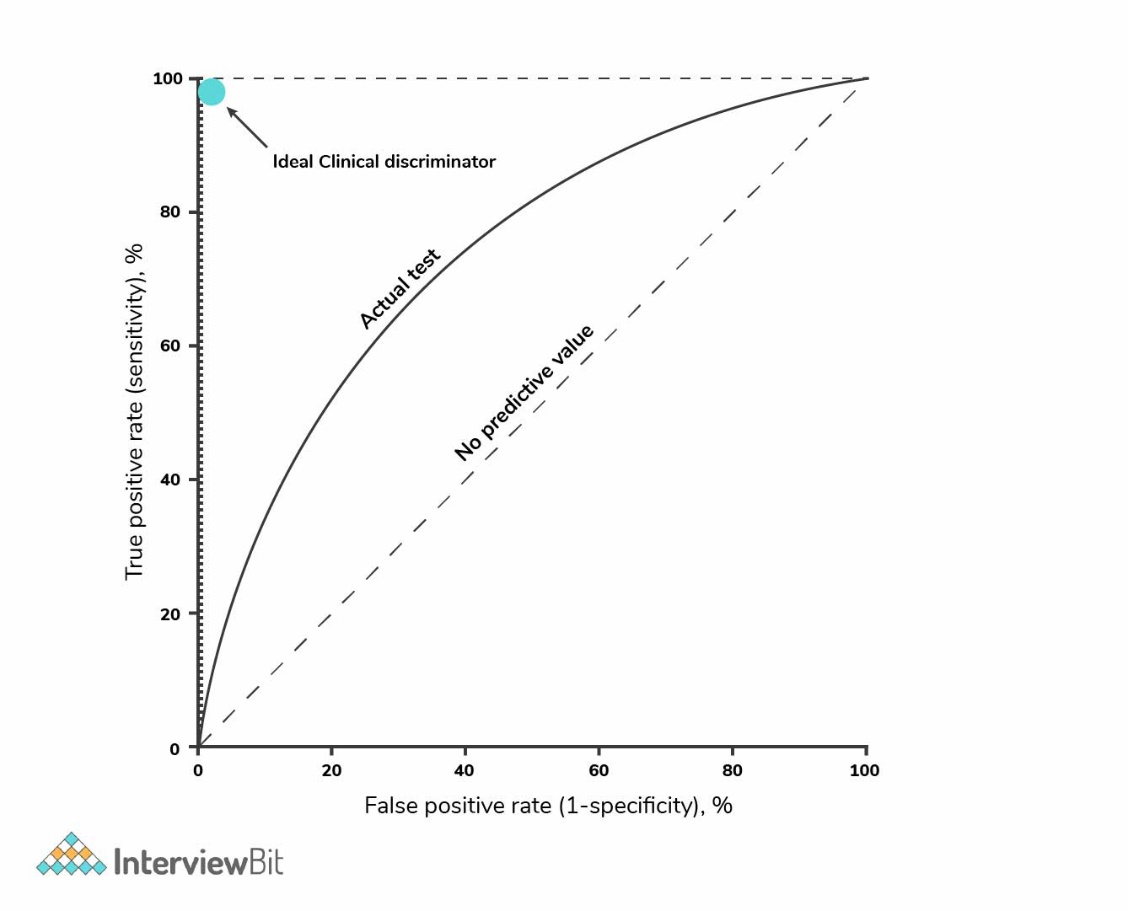
Learning rate controls how much we should adjust the weights concerning the loss gradient. Learning rates are randomly initialised. When the learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point. If the learning rate is too high, this causes undesirable divergent behaviour to the loss function due to drastic updates in weights. It may fail to converge or even diverge.

# Cost function and Loss function

Cost functions are a tool to evaluate how good the model performs. It takes into consideration the errors and losses made in the output layer during the back propagation process.

When calculating loss we consider only a single data point, then we use the term loss function. The loss function is to capture the difference between the actual and predicted values for a single record. Most commonly used loss functions are Mean-squared error and Hinge Loss.

# ROC Curve

Receiver Operating Characteristic curve is a graphical representation between false-positive rates and true positive rates at different thresholds. It is a proxy for trade-off between sensitivity and specificity.

TPR represents the proportion of observations correctly predicted as positive out of overall positive observations.

FPR represents the proportion of observations incorrectly predicted out of overall negative observations.

A completely random model is represented by a straight line, having a 0.5 ROC. The amount of deviation a ROC has from this straight line denotes the efficiency of the model.

# Test set and Validation set

* Validation set is part of training set used to select parameters for avoiding model overfitting.
* Test set is used to evaluate the performance of the trained model.

# Regularization

Regularization entails addition of penalties to different parameters in the ML model for reducing the freedom of the model to avoid the issue of overfitting. E.g., Linear model regularization, Lasso/L1 regularization etc. The model predictions should then minimize the loss function calculated on the regularized training set.

Linear model regularization applies penalty over coefficients that multiplies the predictors.

Lasso/L1 regularization shrinks some coefficients to zero, thereby removing it from the model.

# Hyperparameters

Hyperparameters are parameter whose value is set before the learning process so that the network training requirements can be identified and the structure of the network is improved. It includes learning rate, epochs etc.

# Batch Normalization

A technique which attempts to improve the performance and stability of the neural network. It is done by normalizing the inputs in each layer so that the mean output activation remains 0 with the standard deviation at 1. The model becomes less sensitive to hyperparameter tuning. Weight initialization becomes an easy task. It introduces mild regularisation in the network.

# Layers of CNN

There are four different layers of CNN.

* Convolutional layer
* ReLU layer
* Pooling layer
* Fully connected layer

# Optimiser

Deep learning is an iterative process. With so many hyperparameters to tune or methods to try, it is important to be able to train models fast, to quickly complete the iterative cycle. This is the key to increase the speed and efficiency of a machine learning team. Hence the importance of optimisation algorithms such as stochastic gradient descent, min-batch gradient descent, gradient descent with momentum and the Adam optimiser.

Adam optimiser is the best one.

Given an algorithm *f(x),* it helps in either minimisation or maximisation of the value of *f(x).* We use optimisation algorithms to train the neural network by optimising the cost function.

# Deep Learning Frameworks

* Caffe
* Keras
* TensorFlow
* PyTorch
* Chainer
* Microsoft Cognitive Toolkit

# Online Machine Learning

Online learning is a combination of different techniques of ML where data arrives in sequential order and the learner aims to learn and update the best predictor for future data at every step.

Traditional machine learning techniques run in batch mode. For example, supervised learning tasks where the complete training data is fed in advance to train a model by applying certain algorithms. Such an approach requires entire training data available prior to the learning task and the process is also in offline mode due to expensive training costs. Conventional techniques suffer from some critical drawbacks like low efficiency in both time and space cost; and poor scalability for large-scale applications because the model often has to retrain from scratch for new data.

On the other hand, online learning is a combination of different techniques of ML where data arrives in sequential order and the learner (algorithm/model) aims to learn and update the best predictor for future data at every step. Online learning is able to overcome drawbacks of offline learning like models can be updated instantly for any change in data. Therefore, online learning is far more efficient and scalable for large-scale learning tasks in real-world data, analytics, and various applications where data is not only large in size but also arrives at high velocity.

Training a model can go wrong in lots of different ways: the algorithm itself might not be suitable, the model might fail to generalise well, the learning rate might be wrong, the regularisation might be too low or too high…the list goes on. Why on earth would we even attempt to learn immediately when there are no guarantees on what might happen?

The answer is simple: no matter how good a model is, or how much data you feed it, a model is still an imperfect representation of an environment. To make the best possible decisions right now, we can’t afford to have a model that only knows about things that happened yesterday.

Consider the following example. Let’s say we run a news website. We personalise our news by collecting data on what was clicked or not clicked, and by whom. Based on this information, we predict the types of news different visitors might like, and serve them relevant items.

One day, out of the blue, word gets out that the government is issuing a state of emergency, and will hold a press conference in an hour. Suddenly, everyone is interested in domestic affairs — even those who typically only read about sports or look at the funnies. When presented with a news piece about the conference, a huge percentage of the audience clicks it to learn more.

If you had gone the traditional route and batch trained your recommendation engine once a day, it would still be stuck offering the same type of content, even though the underlying world changed dramatically¹. You should be serving up domestic news right now, but aren’t because your system is too slow.

It gets worse: the following day, after the press conference and following a new training cycle, your engine would start actively recommending domestic news which, after 24 hours, isn’t necessarily interesting any more. It’s made two mistakes, both because it can’t react fast enough.

That’s the power of online learning: done properly, it can react in minutes or even seconds. With it, there is no such thing as “yesterday’s news”.

Online learning also requires an entirely different approach in terms of technical architecture. Since a model can, and will, change from second to second, you can’t just instantiate several instances like you can with traditional techniques. It’s not horizontally scalable. Instead, you are forced to have a single model instance that eats new data as fast as it can, spitting out sets of learned parameters behind an API. And the second that one set in one process gets replaced by a new one, all other processes must follow suit immediately. It’s an engineering challenge, because the most important part (the model) is only vertically scalable. It may not even be feasible to distribute between threads.

Learning immediately also requires fast access to new data. If you’re lucky enough to get all the data you need for a single training example as part of an API call, you’re good to go. But if something is not available client-side, you need to be able to grab that data from somewhere in milliseconds. Typically, that means using an in-memory store like Redis. “Big data” processing frameworks aren’t of much help. If you want to do both batch and online learning, Spark isn’t enough. If you do only online learning, Spark is useless.

# Polling

A method that is used to reduce the spatial dimensions of a CNN. It down samples operations for reducing dimensionality and creating pooled feature maps. Pooling in CNN helps in sliding the filter matrix over the input matrix.

# Boltzmann Machine

It is an algorithm that discover fascinating features representing complex regularities present in the data. It optimizes the quantity and weight.

# LSTM

Long Short-Term Memory. It is a RNN that is capable of learning long term dependencies and recalling information for a longer period as part of its default behaviour.

# Evolution of CNN

It all started with LeNet in 1998 and eventually, after nearly 15 years, lead to ground-breaking models winning the ImageNet Large Scale Visual Recognition Challenge which includes AlexNet in 2012 to Google Net in 2014 to ResNet in 2015 to an ensemble of previous models in 2016. In the last two years, no significant progress has been made, and the new models are an ensemble of previous ground-breaking models.

* LeNet in 1998

LeNet is a 7-level convolutional network by LeCun in 1998 that classifies digits and used by several banks to recognise the hand-written numbers on cheques digitised in 32x32 pixel greyscale input images.

* AlexNet in 2012

AlexNet: It is considered to be the first paper/ model, which rose the interest in CNNs when it won the ImageNet challenge in the year 2012. It is a deep CNN trained on ImageNet and outperformed all the entries that year.

* VGG in 2014

VGG was submitted in the year 2013, and it became a runner up in the ImageNet contest in 2014. It is widely used as a simple architecture compared to AlexNet.

* GoogLeNet in 2014

In 2014, several great models were developed like VGG, but the winner of the ImageNet contest was GoogLeNet. GoogLeNet proposed a module called the inception modules that includes skipping connections in the network, forming a mini-module, and this module is repeated throughout the network.

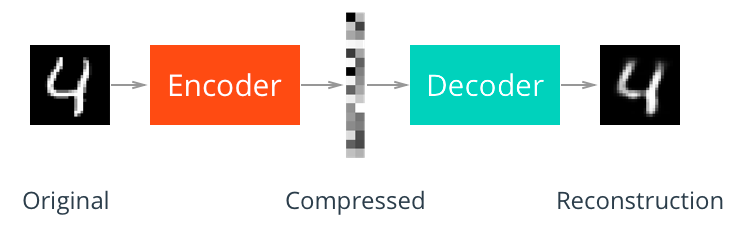
* ResNet in 2015

There are 152 layers in the Microsoft ResNet. The authors showed empirically that if you keep on adding layers, the error rate should keep on decreasing in contrast to “plain nets” we're adding a few layers resulted in higher training and test errors.

# Autoencoders

Autoencoders transform input into output with minimal possible error. An autoencoder receives unlabelled input that is encoded for reconstructing the output. It is a kind of ANN. Autoencoder also tries to generate a representation as close as possible to its original input from the reduced encoding. These models are trained as supervised machine learning models and during inference, they work as unsupervised models that’s why they are called self-supervised models. Application of Autoencoders include:

* File Compression: Autoencoders reduce the dimensionality of input data which is referred to as file compression. Autoencoders work with all kinds of data like images, videos, and audio, this helps in sharing and viewing data faster than we could do with its original file size.
* Image de-noising: Autoencoders does not require any human interaction, once trained on any kind of data it can reproduce that data with less noise than the original image.
* Image Transformation: Autoencoders can transform B/W images to coloured one and vice versa, we can up-sample and down-sample the input data, etc.



# Prior Probability and Likelihood

Prior probability is the proportion of the dependent variable in the dataset.

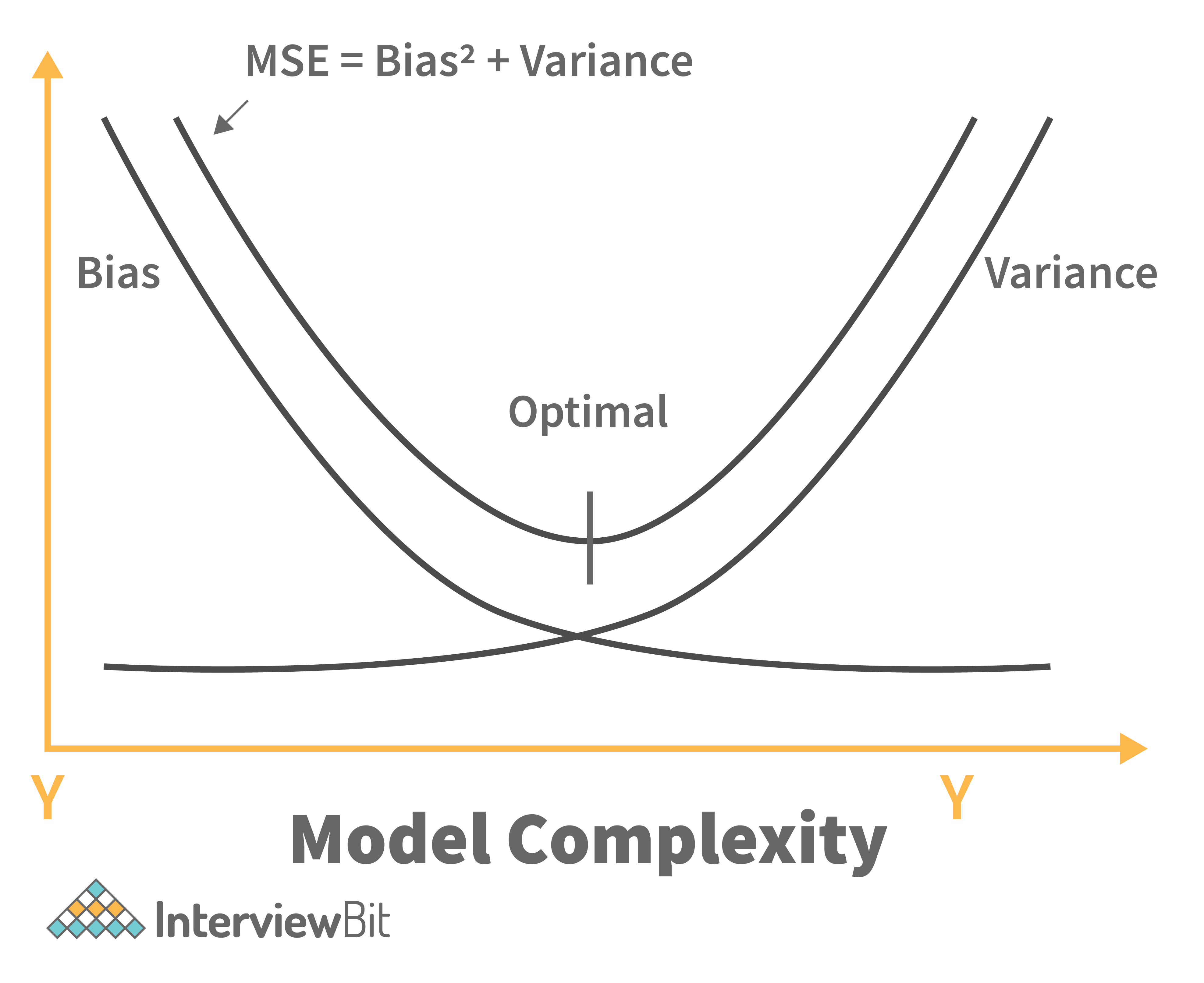
Likelihood is the probability of classifying a given observant in the presence of some other variable.

# Power Analysis

Part of experimental design. It determines the sample size required to find the effect of a given size from a cause with a specific level of assurance.

# Bias-Variance Trade-off

It is the optimum balance between bias and variance in a ML model. If you try to decrease bias, the variance will increase and vice-versa. Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this happens only until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



# Univariate, Bivariate and Multivariate Analysis

Univariate data – Contains one variable.

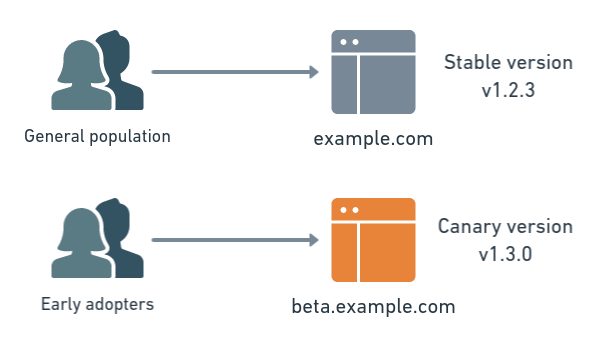
Bivariate data – Contains two different variables.

Multivariate data – Contains three or more variables.

# Maintenance of deployed model

1. Monitor: constant monitoring of all models is needed to determine their performance accuracy.
2. Evaluate: evaluation metrics of the current are calculated to determine if a new algorithm is needed
3. Compare: the new models are compared to each other to determine which model performs the best.
4. Rebuild: the best performing model is rebuilt on the current state of the data.

# Canary Deployment

Canary release allows you to rollout your feature to only a subset of users as an initial test to make sure nothing else in your system broke.

When you deploy the new software version, you shift some percentage – say, 10% - of your user base to the new version while maintaining 90% of users on the old version. If that 10% reports no errors, you can roll it out to gradually more users, until the new version is being used by everyone. If the 10% has problems, though, you can roll it right back, and 90% of your users will have never seen the problem.

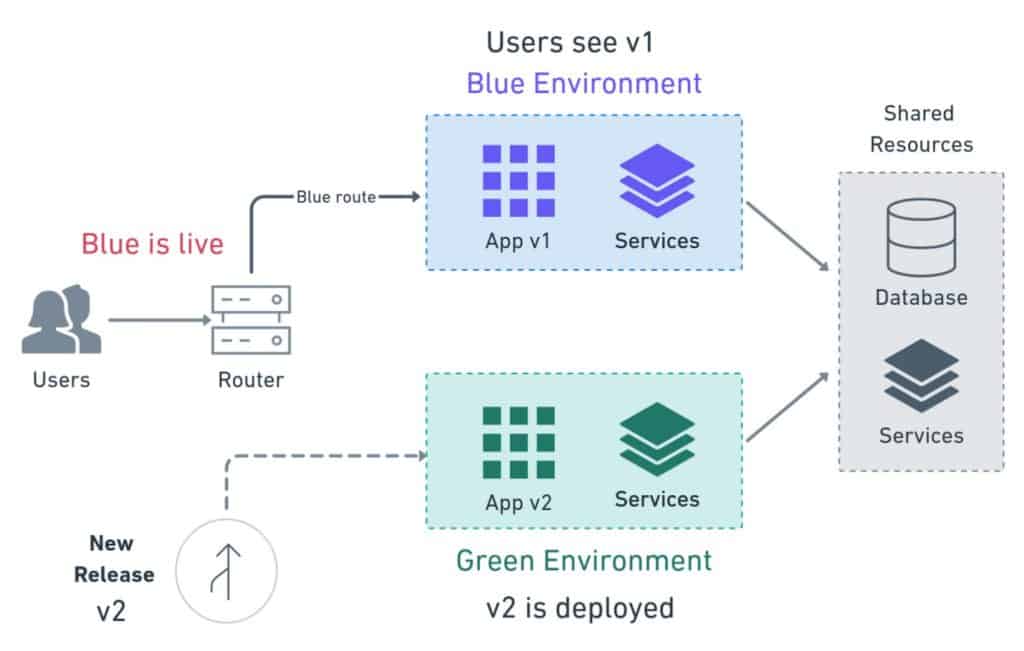
Canary deployment benefits include zero downtime, easy rollout and quick rollback – plus the added safety from gradual rollout process. Drawbacks include – the expense of maintaining multiple server instances, the difficult clone-or-don’t clone database decision.

# Blue Green Deployment

It is a technique that reduces downtime and risk by running two identical production environments called Blue and Green.

At any time, only one of the environments is live, with the live environment serving all production traffic. For this example, Blue is currently live, and Green is idle. As you prepare a new version of your model, deployment and the final stage of testing takes place in the environment that is not live, in this example, Green. Once you have deployed and fully tested the model in Green, you switch the router, so all incoming requests now go to Green instead of Blue. Green is now live, and Blue is idle.

If something unexpected happens with your new version on Green, you can immediately roll back to the last version by switching back to Blue.



# Market Basket Analysis

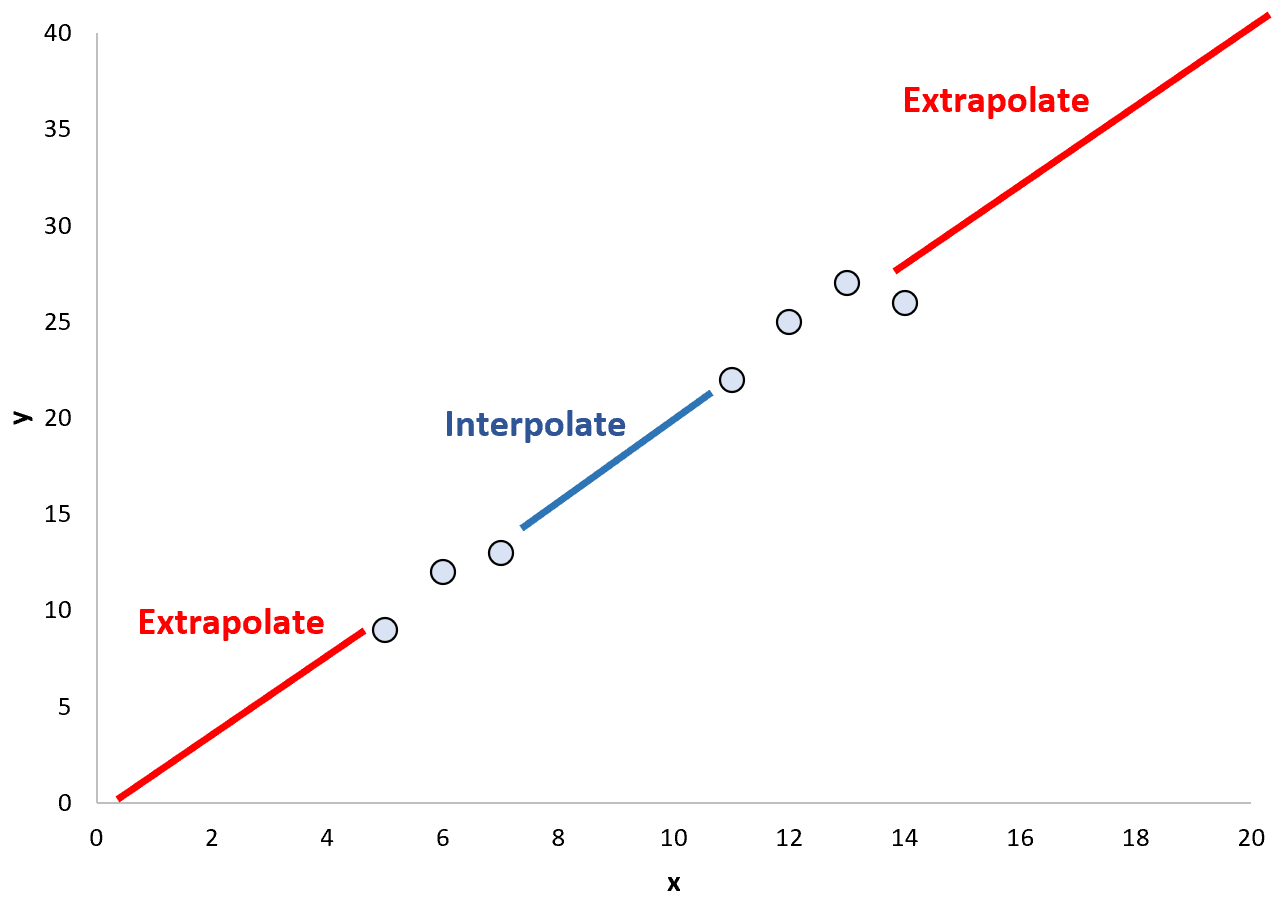
Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. Understanding the relationships and the strength of those relationships is valuable information that can be used to make recommendations, cross-sell, up-sell, offer coupons, etc.

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. It allows retailers to identify relationships between the items that people buy.

# Interpolation and Extrapolation

Interpolation is a prediction made using inputs that lie within the set of observed values. Generally, interpolations are more accurate.

Extrapolation is when a prediction is made using an input that’s outside the set of observed values.



# Time series problems

Time series is extrapolation whereas Regression is interpolation. Time series refers to an organized chain of data. Time-series forecasts what comes next in the sequence.

Regression can be applied to Time-series problems as well as to non-ordered sequences which are termed as features.

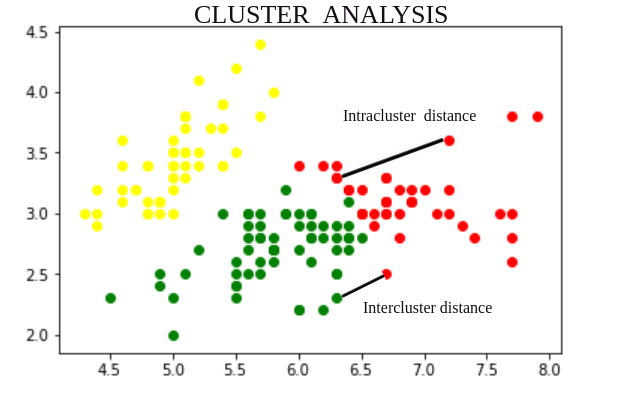
The observations close to one another in time are expected to be similar to the ones far away which provide accountability for seasonality. For instance, today’s weather would be similar to tomorrow’s weather but not similar to weather from 4 months from today. Hence, weather prediction based on past data becomes a time series problem.

# Clustering

Dividing data points into a number of groups. The division is done in a way that all the data points in the same group are more similar to each other than the data points in other groups. Example: Fuzzy clustering, K-means clustering, Density-based clustering etc.

# Intercluster and Intracluster

The aim of the clustering process is to discover overall distribution patterns and interesting correlations among the data attributes. It is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions.

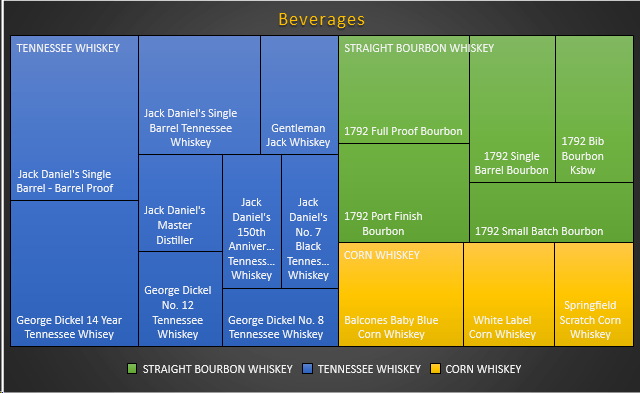


# Heatmap

2-D graphical representation of data containing individual values in a matrix format. The values show the correlation values and are represented by various shades of the same colour.

# Tree Map

It is a chart type that illustrates hierarchical data or part-to-whole relationships.



# Downsides of Visualisation

Few of the downside of visualisation are: It gives estimation not accuracy, a different group of the audience may interpret it differently, improper design can cause confusion.

# In a time-interval of 15-minutes, the probability that you may see a shooting star or a bunch of them is 0.2. What is the percentage chance of you seeing at least one star shooting from the sky if you are under it for about an hour?

Let us say that probability that we may see a minimum of one shooting star in 15 minutes is ‘Prob’. So, Prob = 0.2

Now, the probability that we may not see any shooting star in the time duration of 15 minutes is

The probability that we may not see any shooting star for an hour is

So, the probability that we will see one shooting star in the time interval of an hour is

# You are given two containers - one is 5 and the other one is 7 gallons. How do you use them to measure 4 gallons of water?

Fill the entire 7-gallon container with water. Then use the water in the 7-gallon container in order to fill the entire 5-gallon container. This would leave 2 gal­lons of water in the 7-gallon container. Dump the wa­ter in the 5-gallon container and then pour in it the 2 gallons of water that are in the 7-gallon container. Fill the entire 7-gallon container with water and then start pouring the water in the 5-gallon container. Given that it is already filled with 2 gallons of water, you will be able to pour only 3 gallons, which means that 4 gal­lons would remain within the 7-gallon container. This is how you are able to measure 4 gallons of water.

# How many gallons of white house paint are sold in the US every year?

Find the number of homes in the US: Assum­ing that there are 300 million people in the US and the average household contains 2.5 peo­ple then we can conclude that there are 120 million homes in the US.

* *Number of houses:* Many people live in apart­ments and other types of buildings different than houses. Let’s assume that the percentage of people living in houses is 50%. Hence, there are 60 million houses.
* *Houses that are painted in white:* Although white is the most popular colour, many people choose dif­ferent paint colours for their houses or do not need to paint them (using other types of techniques in order to cover the external surface of the house). Let’s hy­pothesize that 30% of all houses are painted in white, which makes 18 million houses that are painted in white.
* *Repainting:* People need to repaint their houses after a given amount of years. For the purposes of this exercise, let’s hypothesize that people repaint their houses once every 9 years, which means that every year 2 million houses are repainted in white.

I have never painted a house, but let’s assume that in order to repaint a house you need 30 gallons of white paint. This means the total US market for white house paint is 60 million gallons.

# You have eight balls of the same size. Sev­en of them weigh the same, and one of them weighs slightly more. How can you find the ball that is heavier by using a balance and only two attempts at weighing?

You can put six of the balls on the balance. If one of the sides is heavier you will know that the heavier ball is on that side. If not, the heavier ball is among the two that you did not measure and it will be real­ly easy to determine precisely which ball is heavier with your second weighing.

After you determine which side is heavier, you will be left with 3 balls to choose from. You have another attempt at weighing left. You can put two of the balls on the balance and see if one of them is heavier. If it is, then you have found the heavier ball. If it is not, then the third ball is the one that is heavier.

# A windowless room has three light bulbs. You are outside the room with 3 switches, each of them controlling one of the light bulbs. If you were told that you can enter the room only once, how are you going to tell which switch controls which light bulb?

You have to be creative in order to solve this one. You switch on two of the light bulbs and then wait for 30 minutes. Then you switch off one of them and en­ter the room. You will know which switch controls the light bulb that is on. Here is the tough part. How are you going to be able to determine which switch cor­responds to the other two light bulbs? You will have to touch them. Yes. That’s right. Touch them and feel, which one of them is heated. That will be the other bulb that you had turned on for 30 minutes.

In case the bulbs are LED, it is impossible to find the correct answer.

# Data Modelling and Database Design?

Data Modeling can be considered as the first step towards the design of a database. Data modelling creates a conceptual model based on the relationship between various data models. The process involves moving from the conceptual stage to the logical model to the physical schema. It involves the systematic method of applying data modelling techniques.

Database design is the process of designing the database. The database design creates an output which is a detailed data model of the database.

# ZCA Whitening

Zero Component Analysis Whitening. ZCA transforms the data to zero means and makes the features linearly independent of each other. Making the co-variance matrix as the identity matrix is called whitening.

# RandomizedSearchCV, GridSearchCV, BayesianSearchCV

Randomized search CV is used to perform a random search on hyperparameters. Randomized search CV uses a fit and score method, predict probability, transform etc. The parameters of the estimator used to apply these methods are optimized by cross-validated search over parameter settings. A fixed number of parameter settings is sampled from the specified distributions. The number of parameter settings that are tried is given by ‘n\_iter’.

GridSearchCV

Grid search is the process of performing hyperparameter tuning to determine the optimal values for a given model. It runs on all the possible range of hyperparameter values and outputs the best model.

BayesianSearchCV

Bayesian search keeps track of past evaluation results, which they use to form a probabilistic model mapping hyperparameters to a probability of a score on the objective function

# t-SNE

t-Distributed Stochastic Neighbour Embedding is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two or more dimensions suitable for human observation.

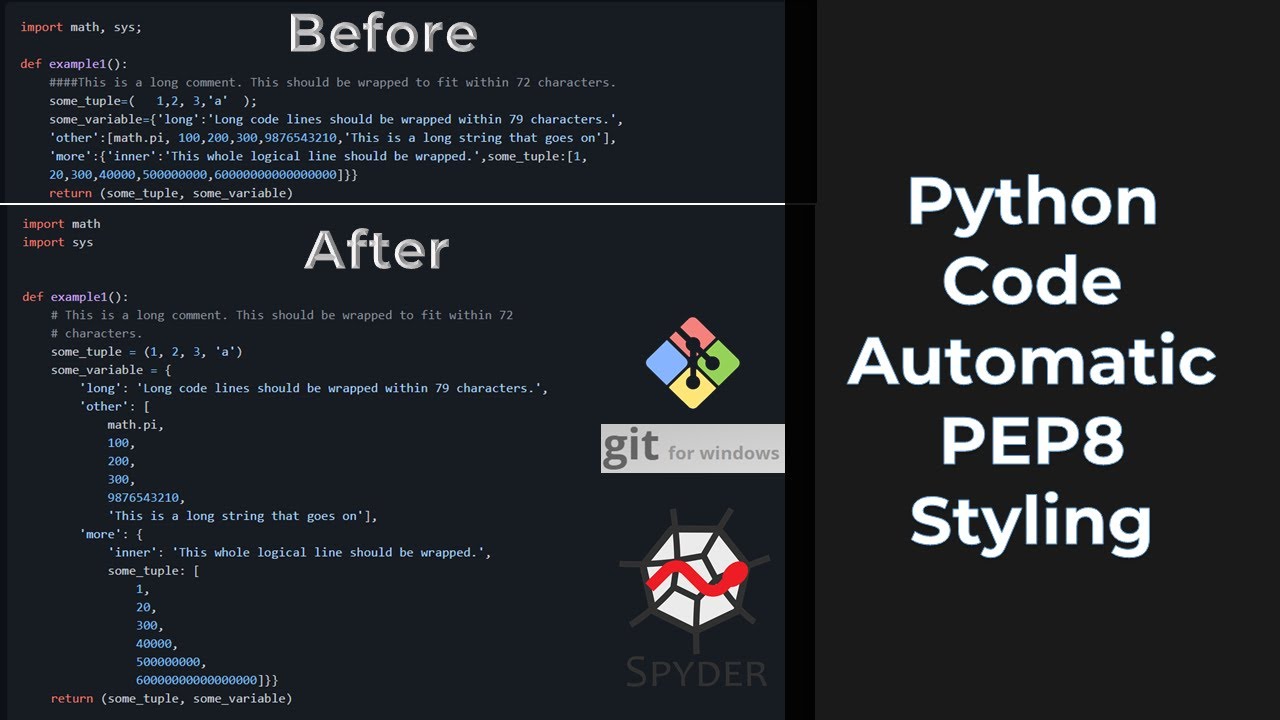
# References

1. Interviewbit
2. Intellipaat
3. Simplilearn
4. Hacker.io
5. Guru99
6. iNeuron
7. Springboard
8. mygreatlearning

Python

# PEP8

PEP 8 is a coding convection. It consists of coding guidelines that are a set of recommendations for Python language about making the Python more readable and usable for another person. PEP stands for Python Enhancement Proposal.



# What type of language is python

Python is capable of scripting, but in general sense, it is considered as a general-purpose programming language.

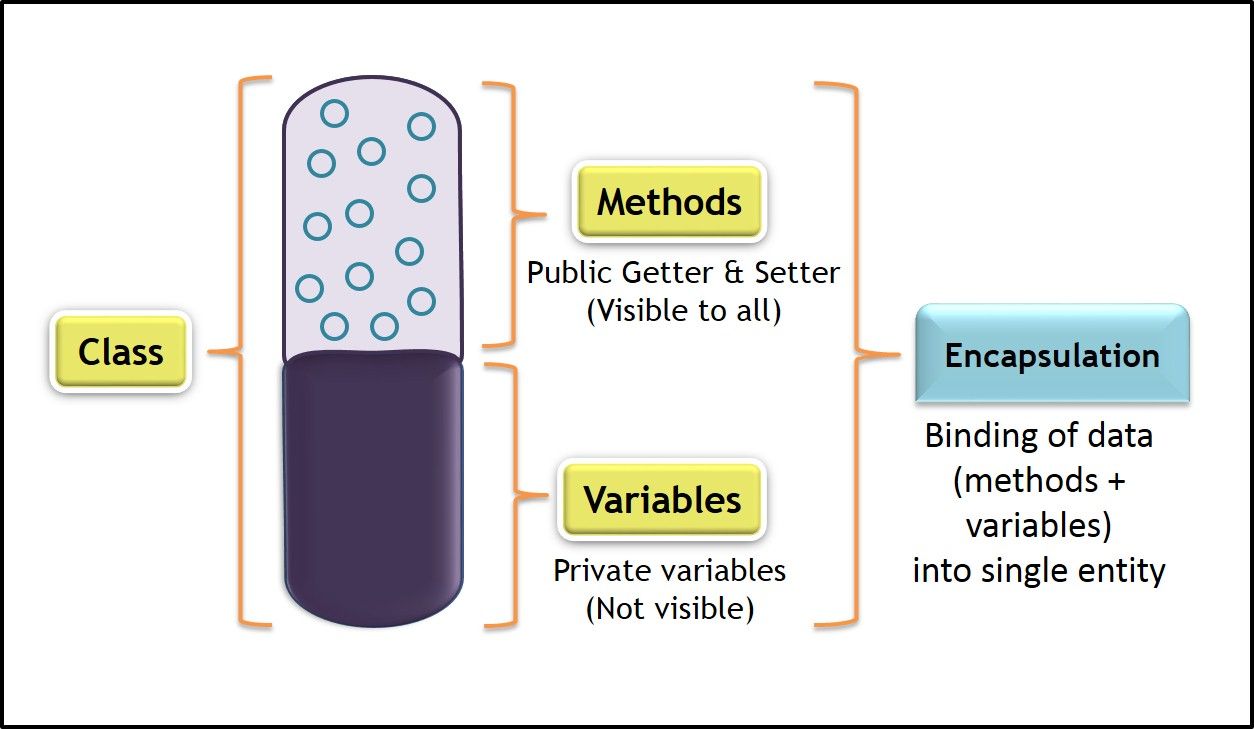
Python is case-sensitive.

Whenever python exits, especially those Python modules which are having circular references to other objects or the objects that are referenced from the global namespaces are not always de-allocated or freed. It is impossible to de-allocate those portions of memory that are reserved by the C library. On exit, because of having its own efficient clean up mechanism, python would try to de-allocate/destroy every other object.

Python lays down the concept of prefixing the name of the variable, function or method with a single or double underscore to imitate the behaviour of protected and private access specifiers.

# Encapsulation

It means binding the code and the data together. Example, a python class.



# What does \*args, \*\*kwargs mean

We use \*args when we aren’t sure how many arguments are going to be passed to a function, or if we want to pass a stored list or tuple of arguments to a function.

\*\*kwargs is used when we don’t know how many keywords argument will be passed to a function, or it can be used to pass the values of a dictionary as keyword arguments.

# Is Python an Object-Oriented language

Yes, Python is an object-oriented programming language meaning it can enclose the codes within the objects. The property allows the storage of the data and the method in a single unit called the object.

# What are docstrings

Docstrings are not actually comments, but they are documentation strings. These docstrings are within triple quotes. They are not assigned to any variable and therefore, at times, serve the purpose of comments as well.

# What is difference between xrange and range

*xrange* and *range* are exact same in terms of functionality. The only difference is that *range* returns a python list object and *xrange* returns an *xrange* object.

# What is a lambda function?

An anonymous function is known as a lambda function. This function can have any number of parameters but can have just one statement.

What are generators

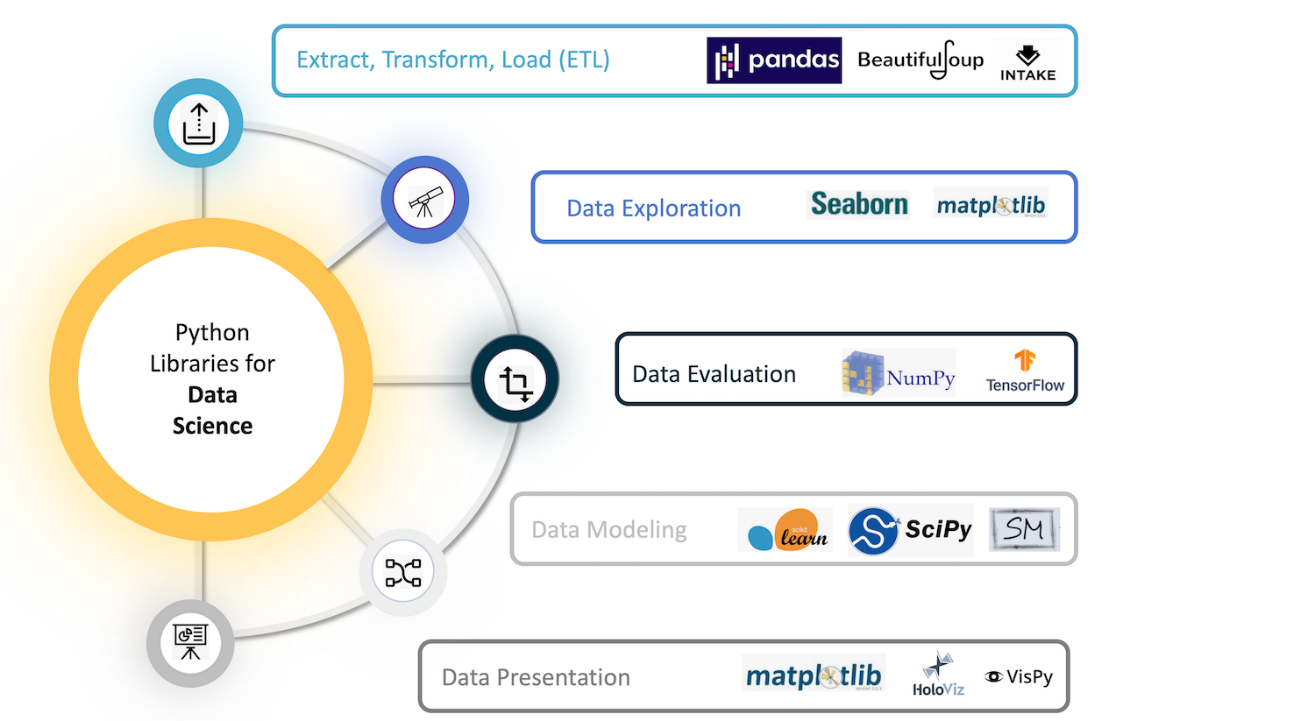
Generators are a way of implementing iterators. A generator function is a normal function except that it contains yield expression in the function definition making it a generator function.

# What is Python module

A module is a single file / files containing functions, definitions, and variables designed to do certain tasks. It is a extension file. It can be imported at any time during a session and needs to be imported only once. To import a python module, there are two ways:

# Library

A library is a collection of reusable functionalities of codes that allows us to perform a variety of tasks without having to write the code. It is a collection of modules. These codes can be used by importing and by calling that library’s method with a period (.).



# Name mutable and immutable objects

The mutability of a data structure is the ability to change the portion of the data structure without having to recreate it. Mutable objects are lists, sets, values in a dictionary.

Immutability is the state of the data structure that cannot be changed after its creation. Immutable objects are integers, strings, float, bool, tuples, keys of a dictionary.

# Compound Data Types

The data type that is constructed using simple, primitive, and basic data types.

# Data Structures

Data structures allow us to store multiple observations. E.g., lists, tuples, sets, and dictionaries.

# Difference between list and tuple

List

* Enclosed within
* Mutable
* Slower than tuples
* E.g. [‘A’, 1, ‘h’]

Tuple

* Enclosed in
* Immutable
* Faster than lists
* Must be used when the order of the elements of a sequence matter.
* E.g. (‘Twenty’, 20, ‘XX’)

# Why is python numpy better than lists

Python numpy arrays should be considered instead of a list because they are fast, consume less memory and convenient with lots of functionality.

# What is the difference between NumPy and SciPy

NumPy would contain nothing but the array data type and the most basic operations: indexing, sorting, reshaping, basic element wise functions, et cetera. All numerical code would reside in SciPy. SciPy contains more fully-featured versions of the linear algebra modules, as well as many other numerical algorithms.

# Features that make pandas a reliable option to store tabular data

* Memory Efficient
* Data Alignment
* Reshaping
* Merge and Join
* Time series

# What is a series in pandas

A series is defined as a 1-D array that is capable of storing various data types. The row labels of the series are called the index. By using a ‘series’ method, we can easily convert the list, tuple, and dictionary into series. A series cannot contain multiple columns.

# What is Time Series in Pandas

A time series is an ordered sequence of data which basically represents how some quantity changes over time. Pandas contain extensive capabilities and features for working with time series data for all domains.

# What is vectorization in Pandas

Vectorization is the process of running operations on the entire array. This is done to reduce the amount of iteration performed by the functions. Pandas have a number of vectorized functions like aggregations, and string functions that are optimized to operate specifically on series and DataFrames. So, it is preferred to use the vectorized panda’s functions to execute the operations quickly.

# What is groupby in Pandas

GroupBy is used to split the data into groups. It groups the data based on some criteria. Grouping also provides a mapping of labels to the group names. It has a lot of variations that can be defined with the parameters and makes the task of splitting the data quick and easy.

# Map function

Map function executes the function given as the first argument on all the elements of the iterable given as the second argument.

# What is the meaning of axis = 0 and axis = 1

Axis = 0 is meant for reading rows, axis = 1 is meant for reading columns.

# String interpolation

* f strings
* % Operator

# Difference between “is” and “==”

‘is’ checks identity and ‘==’ checks equality.

Example: We will create some lists and assign them to names. ‘b’ points to the same object as ‘a’.

Output:

True

True

Output:

True

False

Output:

12345

12345

12456

‘c’ has a different ‘id’ than ‘a’ and ‘b’

# What are Arrays

* Data structure that contains elements of the same data type, e.g., integer, string.
* Elements share the same variable name, but each element has its own unique index number or key.

# List and Dictionary comprehension

Python comprehensions are syntactic constructs providing a way to build a list, dictionary or set based on the existing list, dictionary or set whilst altering or filtering elements.

These are generally more compact and faster than normal functions and loops for creating lists. We must avoid writing very long comprehensions in one line to ensure that code is user-friendly and to maintain the readability.

Example of list comprehension

Output: 10, 20, 30, 40, 50

Example of dictionary comprehension

Output: {‘i’: ‘this’, ‘ii’: ‘month’, ‘iii’: ‘is’, ‘iv’: ‘March’}

# What is tuple unpacking

A tuple can be unpacked in the sense that its elements can be separated in the following manner:

We have a tuple x = (500, 352)

This tuple can be assigned to two new variables in this way: a, b = x

Now, printing a and b will result in: print(a) = 500 and print(b) = 352

Tuple unpacking helps to separate each value one at a time.

# Difference between %, / and //

returns a remainder

returns quotient

// is the floor division that rounds off the quotient to the bottom

Example:

11 % 2 -> 1

11 / 2 -> 5.5

11 // 2 -> 5

# Difference between indexing and slicing

Indexing is extracting or lookup one or particular values in a data structure.

Slicing retrieves a sequence of elements like strings, list, tuples etc.

# Difference between global and local variables

Global variables are the defined and declared outside a function.

Local variables are declared inside the function’s body.

# Why is the “pass” keyword used in python

The “pass” keyword is a no-operation statement in python. It signals that no action is required. It works as a placeholder in compound statements which are intentionally left blank.

Difference between .py and .pyc files

*.py* files contain the source code of a program. Whereas, *.pyc* file contains the bytecode of your program. We get bytecode after compilation of *.py* file (source code).

*.pyc* files are not created for all the files that you run. It is only created for the files that you import.

# Does python allow arguments Pass by Value or Pass by Reference

Neither the arguments are Pass by Value nor does python support Pass by Reference. Instead, they are Pass by Assignment. The parameter which you pass is originally a reference to the object not the reference to a fixed memory location. But the reference is passed by value.

# What is PYTHONPATH in python

PYTHONPATH is an environment variable which you can set to add additional directories where python will look for modules and packages. This is especially useful in maintaining python libraries that you do not wish to install in the global default location.

# Decorator

A decorator allows adding functionality to an existing function by passing that existing function to a decorator, which executes the existing function as well as additional code.

They are represented by the @decorator\_name in python and are called in bottom-up fashion.

Decorator function: This takes , as an argument. It also defines a function, , which calls and executes some code, . Then it returns the function which is defined.

def ():

def logfunction\_called():

print(f’{ } called.’)

return logfunction\_called

We will write other functions that we’ll eventually add with decorator

def my\_name():

print(‘Manish’)

def friends\_name():

print(‘Momo’)

my\_name()

friends\_name()

Output:

Manish

Momo

Now add the decorator to both.

@

def my\_name():

Print(‘Manish’)

@

def driends\_name():

print(‘Momo’)

my\_name()

friends\_name()

Output:

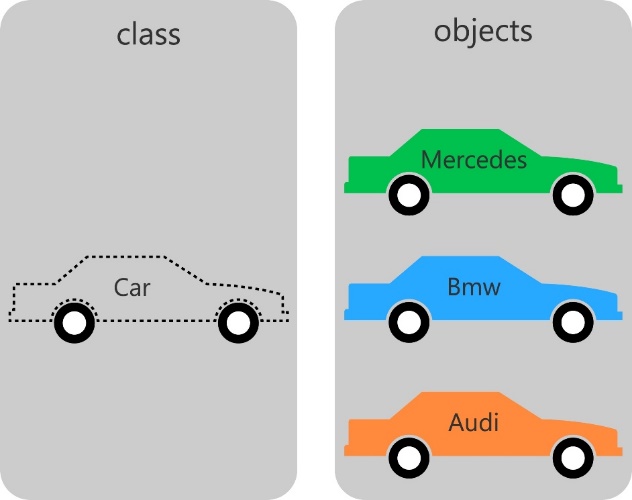
<function my\_name at 0x10f> called

Manish

<function friends\_name at 0x10e> called

Momo

# What is class and object

A class is a user-defined prototype which basically is a blueprint that defines the nature of a future object.

An object is an instance of the class.

Classes can construct instances of objects. This is known as instantiation.

# Difference between data frame and matrices

|  |  |
| --- | --- |
| Data Frame | Matrices |
| A collection of series that share a common index. | A matrix in NumPy is constructed with multiple vectors. |
| It can hold multiple series, which are of different data types. | It can hold only one data type in the entire 2-D structure. |

# Scatter plot

It is 2-D data visualization plot that illustrates relationship between observations of two different variables, plotted against x and y axis respectively.

Explain range function

Range generates a list of integers and there are 3 ways to use it.

The function takes 1 to 3 arguments.

: generate integers from 0 to the “stop” integer.

Output:

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]

generate integers from the “start” to the “stop” integer.

Output:

[2, 3, 4, 5, 6, 7, 8, 9]

generate integers from “start” to “stop” at intervals of “Step”

Output:

[2, 4, 6, 8]

# What does “self” refer to in a class

Self refers to the instance of the class itself. IT’s how we give methods access to and the ability to update the object they belong to.

Below, passing self to \_\_init\_\_() gives us the ability to set the colour of an instance on initialization.

class Shirt:

def \_\_init\_\_(self, colour):

= colour

s = Shirt(‘yellow’)

Output:

Yellow

# Difference between dictionaries and JSON

is python datatype, a collection of indexed but unordered keys and values.

JSON is just a string which follows a specified format and is intended for transferring data.

# Difference between and

is the object representing the function which can be assigned to a variable or passed to another . calls the function and returns it outputs.

def func():

(‘I m a function’)

#=> function \_\_main\_\_.func

#=> I m a function

# Check if a data set or time series is Random

We use the lag plot. If the lag plot for the given dataset does not show any structure, then its random.

# ORM

Object Relational Mapping map data models to database tables and simplifies database transactions.

SQLAlchemy is typically used in the context of Flask, and Django has its own ORM.

# What is namespace in python

In python, every name introduced has a place where it lives and can be hooked for. This is known as namespace. It is like a box where the variable name is mapped to the object placed.

# What is pickling and unpickling

Pickle module accepts any python object and converts it into a string representation and dumps it into a file using the dump function, this process is called pickling.

While the process of retrieving original python objects from the stored string representation is called unpickling.

# Memory status when python exists

No, it is not freed, because the objects that are referenced from global namespaces of Python modules are not always de-allocated when Python exists.

# Difference between append and extend

adds a value to a list while adds values in another list to a list.

#=> [1, 2, 3, 6]

#=> [1, 2, 3, 4, 5]

# inheritance from another in python

In the below example, , inherits from . And with that inheritance comes the instance methods of the parent class.

class Car():

def drive(self):

print(‘vroom’)

class Audi(Car)

pass

audi = Audi()

audi.drive()

# String functions

* To check if a string only contains numbers

–

#=> False

#=> True

* To check if a string contains letters

#=> False

#=> True

* To check if a string contains only numbers and letters

#=> False

#=> True

# Exception handling

Python provides 3 words to handle exceptions, ‘try’, ‘except’ and ‘finally’.

Try:

#Try to do this

Except:

#If try block fails then do this

Finally:

#Always do this

# Difference between the two data series df[‘Name’] and df.loc[:, ‘Name’]

df[‘Name’] is a view of the original data frame and second one is a copy of the original dataframe.

# References

1. Analytixlabs
2. https://towardsdatascience.com/53-python-interview-questions-and-answers-91fa311eec3f

SQL

# Requirements for Data

1. Integrity
2. Availability
3. Security
4. Independent of Application
5. Concurrency

# SQL

Structured Query Language, is the standard language for relational database management systems. It is useful in handling organized data comprised of entities (variables) and relations between different entities of the data.

# SQL Sandbox

SQL Sandbox is a secure environment within SQL Server where untrusted programmes can be run. There are three different types of SQL sandboxes:

Safe Access Sandbox: In this environment, a user may execute SQL activities like as building stored procedures, triggers, and so on, but they can’t access the memory or create files.

Sandbox for External Access: Users can access files without having the ability to alter memory allocation.

Unsafe Access Sandbox: This contains untrustworthy code that allows a user to access memory.

# SQL Skills

SQL skills aid data analysts in the creation, maintenance, and retrieval of data from relational databases, which divide data into columns and rows. Fundamental abilities

* Database Management
* Structuring a Database
* Creating SQL clauses and statements
* SQL system skills like MySQL, PostgreSQL
* PHP expertise is useful
* Analyze SQL data
* OLAP skills

# NoSQL vs SQL

|  |  |
| --- | --- |
| NoSQL | SQL |
| Non-relational database | Relational database |
| Databases use dynamic schemas | Have a specified schema and employ structured query language |
| NoSQL database scale horizontally | SQL databases scale vertically |
| NoSQL databases are document, key-value, graph, or wide-column stores | SQL databases are table-based |
| NoSQL excels at unstructured data such as documents and JSON | SQL databases excel in multi-row transactions. |

# PostgreSQL

PostgreSQL is an enterprise-level, versatile, resilient, open-source, object-relational database management system that supports variable workloads and concurrent users.

# MySQL

It is a relational database management system, like SQL Server, Oracle or IBM DB2, that is used to manage SQL databases. The default port for MySQL server is 3306

# SQL Constraints

Constraints are used to specify the rules concerning data in the table. It can be applied for single or multiple fields in an SQL table during the creation of the table or after creating using the command. The constraints are:

* – Restricts NULL value from being inserted into a column
* – Verifies that all the values in a field satisfy a condition
* – Automatically assigns a default value if no value has been specified for the field
* – Ensures unique values to be inserted into the field. This provides uniqueness for the columns and helps identify each row uniquely. Unlike primary key, there can be multiple unique constraints defined per table.
* – Indexes a field providing faster retrieval of records
* – Uniquely identifies each record in a table. It must contain values and has an implicit constraint. A table in SQL is strictly restricted to have one and only one primary key, which is comprised of single or multiple fields.
* – Ensures referential integrity for a record in another table. A comprises of single or collection of fields in a table that essentially refers to the in another table.

# Limitations of Flat Files

1. Dependency of program on physical structure of data
2. Complex process to retrieve (accessing) data
3. Loss of data on concurrent access
4. Inability to give access based on record (security)
5. Data redundancy and inconsistency

# Use-cases for Files

1. On device access (offline access)
2. Storing frequent incoming data
3. Storing not so important data
4. Low-cost data storage

# Database

A database is a shared collection of logically related data and description of these data, designed to meet the information needs of an organization. A database is an organized collection of data, stored and retrieved digitally from a remote or local computer system.

# Collation

Collation refers to a set of rules that determine how data is sorted and compared. Rules defining the correct character sequence are used to sort the character data. It incorporates options for specifying case sensitivity, accent marks, kana character types, and character width. Below are the different types of collation sensitivity:

* Case sensitivity: A and a are treated differently
* Accent sensitivity: a and á are treated differently
* Kana sensitivity: Japanese kana characters Hiragana and Katakana are treated differently
* Width sensitivity: Same character represented in single-byte (half width) and double-byte (full-width) are treated differently

# Tables and Fields

A table is an organized collection of data stored in the form of rows and columns. Columns can be categorized as vertical and rows as horizontal. The columns in a table are called fields while the rows can be referred to as records.

# ACID

* Atomicity: IT ensures all-or-none rule for database modifications
* Consistency: Data values are consistent across the database
* Isolation: Two transactions are said to be independent of one another
* Durability: Data is not lost even at the time of server failure.

# DBMS

A Database Management System is a software system that enables users to define, create, maintain, and control access to the database. Database systems typically have high cost and they require high end hardware configurations. DBMS is an interface between database application and database.

An Application Program interacts with a database by issuing an appropriate request (typically a SQL statement). It is expensive to use DBMS.

Database Management System, is a software responsible for the creation, retrieval, updating, and management of the database. It ensures that our data is consistent, organized, and is easily accessible by serving as an interface between the database and its end-users or application software. There are two types of DBMS

Relational Database Management System, stores data in the form of a collection of tables, and relations can be defined between the common fields of these tables. Most modern database management systems like MySQL, Microsoft SQL Server, Oracle, IBM DB2, and Amazon Redshift are based on RDBMS.

Non-Relational Database Management System, there is no concept of relations, tuples and attributes. Example – MongoDB

# SQL Comments

SQL Comments are used to clarify portions of SQL statements and to prevent SQL statements from being executed.

Single Line Comments: It starts with two consecutive hyphens (-)

Multi-line Comments: It starts with /\* and ends with \*/

# Cursor

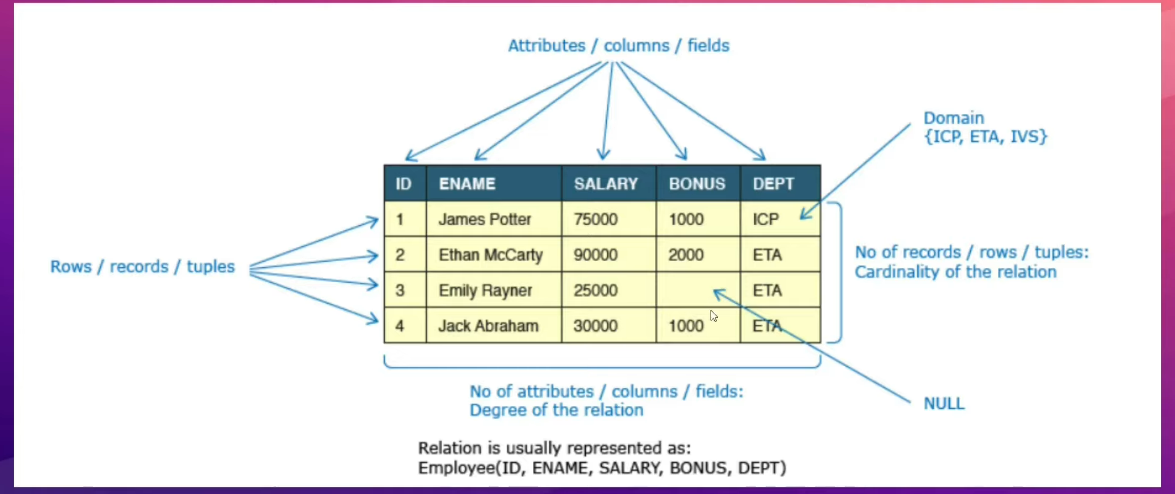
A database cursor is a control structure that allows for the traversal of records in a database. Cursors, facilitates processing after traversal, such as retrieval, addition, and deletion of database records. They can be viewed as a pointer to one row in a set of rows.

# Function of DBMS

1. Data Management – Store, retrieve and modify data
2. Integrity – Maintain accuracy of data
3. Security – Access to authorized users only
4. Concurrency – Simultaneous data access for multiple users
5. Transaction – Modification to database must either be successful or must not happen at all
6. Utilities – Data import/export, user management, backup, logging, performance analysis, recovery mechanism for data

# Types of Databases

1. Relational – Tabular structure (MySQL, Oracle, Microsoft Access, PostgreSQL)
2. NoSQL – Key value pairs (MongoDB)
3. Network – Graph structure
4. Hierarchal – Tree structure



# Data Integrity and Constraints

Data integrity refers to maintaining and assuring the accuracy and consistency of data over its entire life-cycle and is a critical aspect of the design, implementation, and usage of any system which stores, processes, or retrieves data.

Database systems ensure data integrity through constraints which are used to restrict data that can be entered or modified in the database. Database systems offer three types of integrity constraints.

|  |  |  |
| --- | --- | --- |
| Integrity Type | Definition | Enforced Through |
| Entity Integrity | Each table must have a column or a set of columns through which we can uniquely identify a row. These columns cannot have null values. | Primary Key |
| Domain Integrity | All attributes in a table must have a defined domain i.e., a finite set of values which have to be used. When we assign a data type to a column, we limit the values that it can contain. In addition, we can also have value restriction as per business rules e.g., gender must be M or F. | Data types, check constraint |
| Referential Integrity | Every value of a column in a table must exist as a value of another column in a different (or the same) table. | Foreign Key |

# Database Keys

A DBMS key is an attribute or set of attributes which helps us uniquely identify a row in a relation/table.

# Need for keys

1. To uniquely identify a row
2. To enforce data integrity and constraints
3. To establish relationship between tables

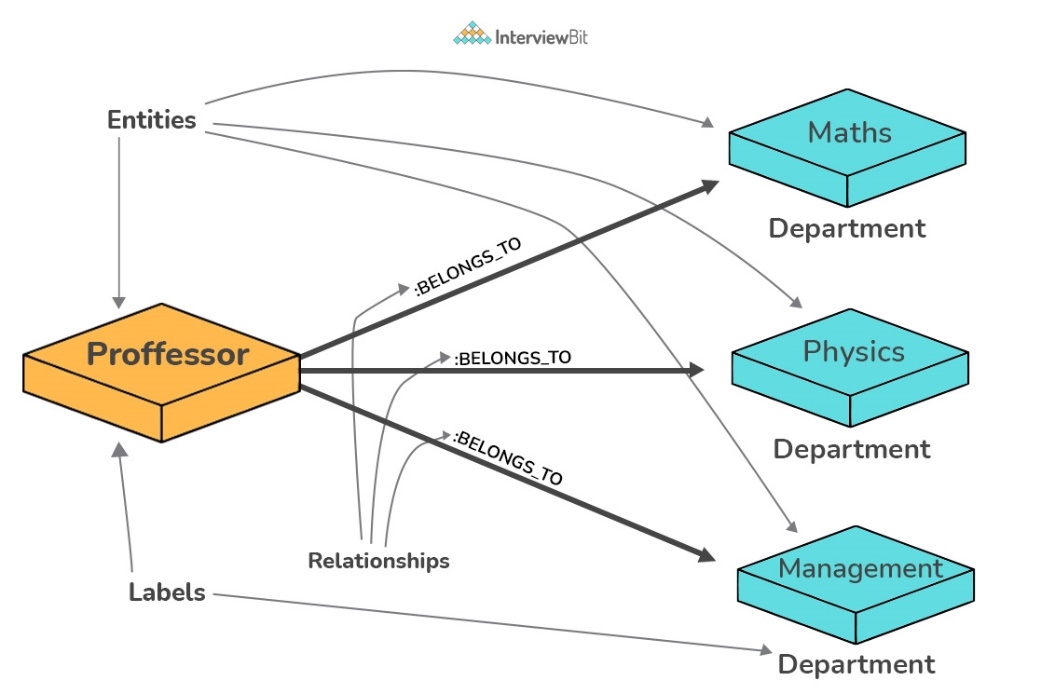
# Types of Keys

|  |  |
| --- | --- |
| Super Key | Foreign Key |
| Candidate Key | Composite Key |
| Primary Key | Compound Key |
| Alternate Key | Surrogate Key |

# Entities and Relationships

An entity can be a real-world object, either tangible or intangible, that can be easily identifiable. For example, in a college database, students, professors, workers, departments, and projects can be referred to as entities. Each entity has some associated properties that provide it an identity.

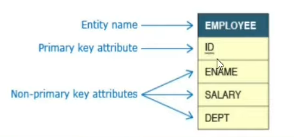
Relations or links between entities that have something to do with each other. For example, the employee’s table in a company’s database can be associated with the salary table in the same database.



# Entity Relationship Model (E-R Model)

ER model is a graphical representation of entities and their relationships which helps in understanding data independent of the actual database implementation.

|  |  |  |
| --- | --- | --- |
| Term | Definition | Examples |
| Entity | Real world objects, either tangible or intangible, which have an independent existence and about which we intend to collect data. | Employee, Computer |
| Attribute | A property that describes an entity. | Name, Salary |

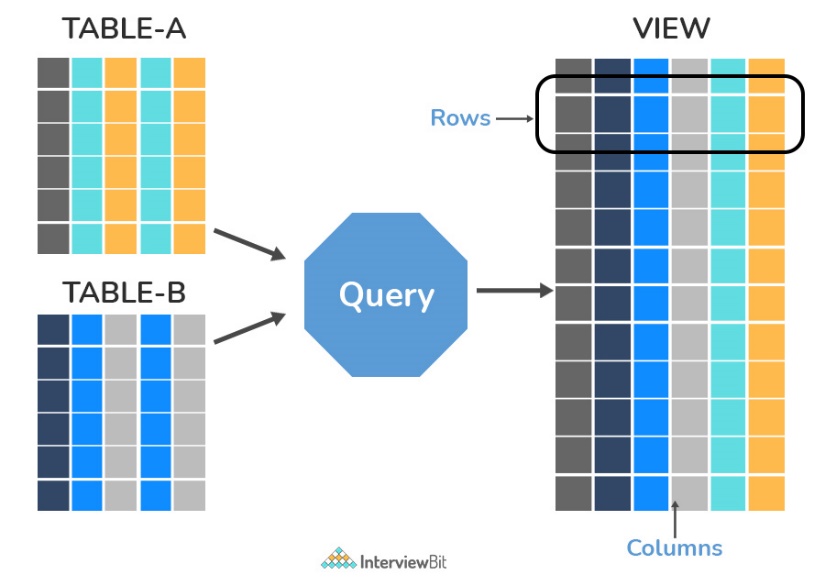


# Cardinality of relationships

Cardinality of relationships is the number of instances in one entity which is associated to the number of instances in another.

# View

A view is SQL is a virtual table based on the result-set of an SQL statement. A view contains rows and columns, just like a real table. The fields in a view are fields from one or more real tables in the database.



# Normalization and Denormalization

Normalization represents the way of organizing structured data in the database efficiently. It includes the creation of tables, establishing relationships between them, and defining rules for those relationships. Inconsistency and redundancy can be kept in check based on these rules, hence, adding flexibility to the database.

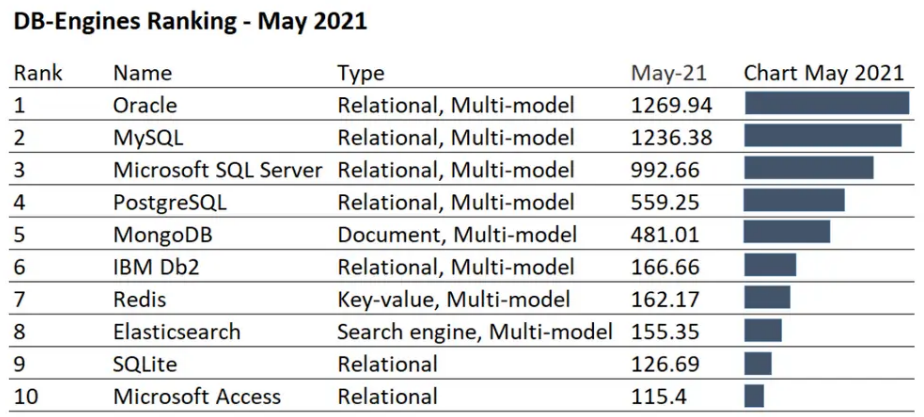
If the same information is repeated in multiple places in the database, there is a risk that it is updated in one place but not the other, leading to data corruption. By having a database with normalization errors, you open the risk of getting invalid or corrupt data into the database.

Denormalization is the inverse process of normalization, where the normalized schema is converted into a schema that has redundant information. The performance is improved by using redundancy and keeping the redundant data consistent. The reason for performing denormalization is the overheads produced in the query processor by an over-normalized structure.

# Prevalence of SQL

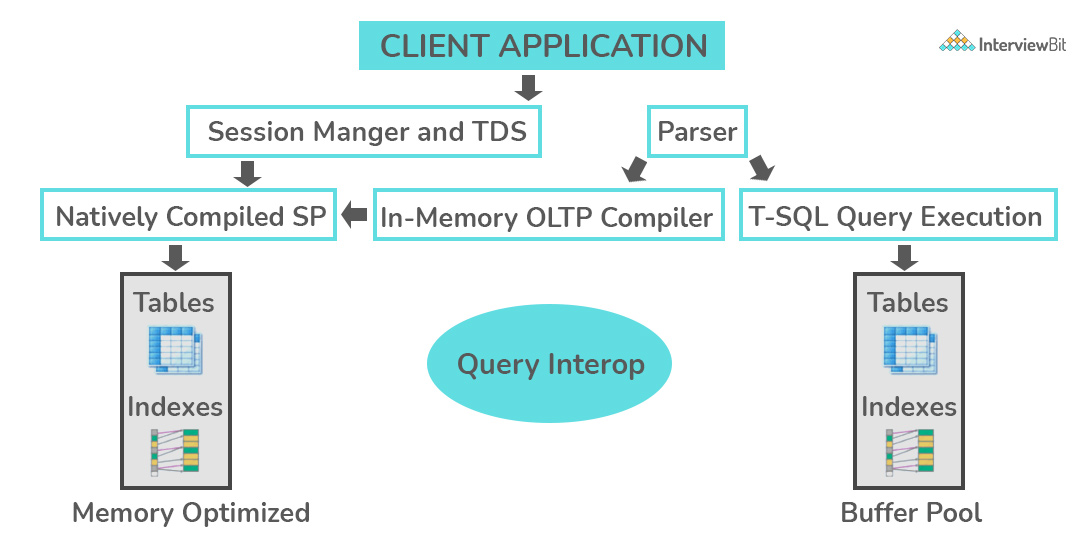
It is easy to visualize databases as large spreadsheets with millions and millions or rows and columns. SQL allows users to quickly manipulate these tables to access information and present the results in the most common formats, tables and associated graphs and visualizations. It is easy to use, syntax is easy to understand.

While No-SQL databases like MongoDB, Cassandra, etc have gained traction with the requirements of Big-Data and real time applications and increasing prevalence of unstructured data, SQL databases still hold the top seven positions of the top ten most popular database engines.



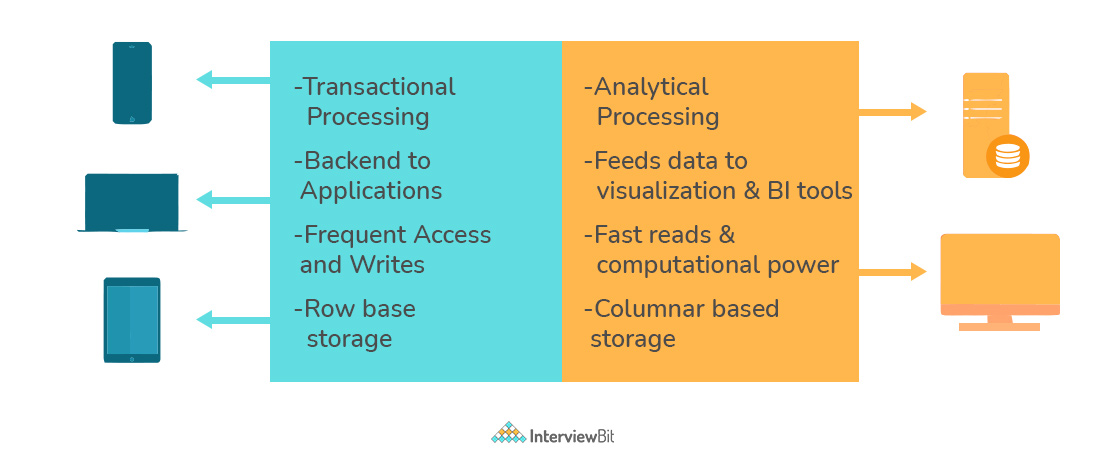
# OLTP

Online Transaction Processing, is a class of software applications capable of supporting transaction-oriented programs. An essential attribute of an OLTP system is its ability to maintain concurrency. To avoid single points of failure, OLTP systems are often decentralized. These systems are usually designed for a large number of users who conduct short transactions. Database queries are usually simple, require sub-second response times, and return relatively few records.



# OLAP

Online Analytical Processing, is a class of software programs that are characterized by the relatively low frequency of online transactions. Queries are often too complex and involve a bunch of aggregations. For OLAP systems, the effectiveness measure relies highly on response time. Such systems are widely used for data mining or maintain aggregated, historical data, usually in multi-dimensional schemas.

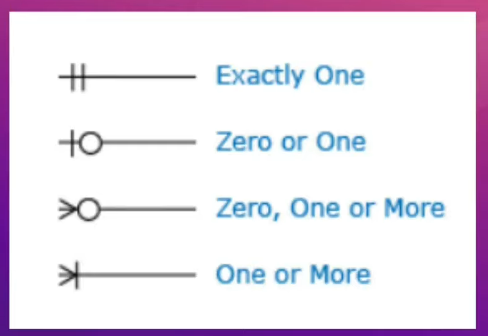


# Types of Relationships

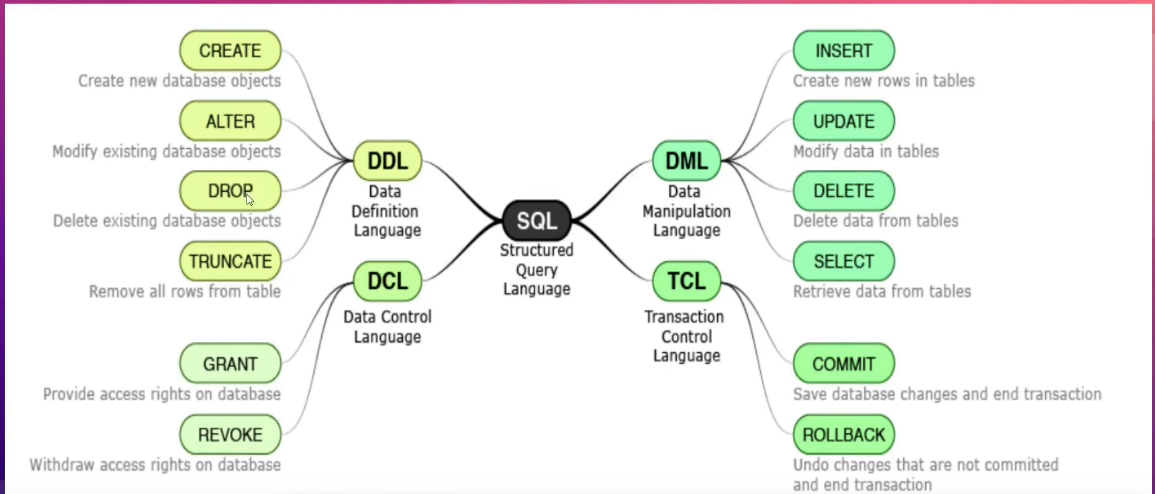
* One-to-One: This can be defined as the relationship between two tables where each record in one table is associated with the maximum of one record in the other table.
* One-to-Many & Many-to-One: Most commonly used relationship where a record in a table is associated with multiple records in the other table.
* Many-to-Many: Used in cases when multiple instances on both sides are needed for defining a relationship.
* Self-Referencing Relationships: Used when a table needs to define a relationship with itself.



# Crow-foot Notation



# Types of SQL Commands



# MySQL Data Types

In MySQL there are three main data types: string, numeric, date and time

String Data Types

|  |  |
| --- | --- |
| **Data type** | **Description** |
| CHAR(size) | A FIXED length string (can contain letters, numbers, and special characters). The size parameter specifies the column length in characters - can be from 0 to 255. Default is 1 |
| VARCHAR(size) | A VARIABLE length string (can contain letters, numbers, and special characters). The size parameter specifies the maximum column length in characters - can be from 0 to 65535 |
| BINARY(size) | Equal to CHAR(), but stores binary byte strings. The size parameter specifies the column length in bytes. Default is 1 |
| VARBINARY(size) | Equal to VARCHAR(), but stores binary byte strings. The size parameter specifies the maximum column length in bytes. |
| TINYBLOB | For BLOBs (Binary Large OBjects). Max length: 255 bytes |
| TINYTEXT | Holds a string with a maximum length of 255 characters |
| TEXT(size) | Holds a string with a maximum length of 65,535 bytes |
| BLOB(size) | For BLOBs (Binary Large OBjects). Holds up to 65,535 bytes of data |
| MEDIUMTEXT | Holds a string with a maximum length of 16,777,215 characters |
| MEDIUMBLOB | For BLOBs (Binary Large OBjects). Holds up to 16,777,215 bytes of data |
| LONGTEXT | Holds a string with a maximum length of 4,294,967,295 characters |
| LONGBLOB | For BLOBs (Binary Large OBjects). Holds up to 4,294,967,295 bytes of data |
| ENUM(val1, val2, val3, ...) | A string object that can have only one value, chosen from a list of possible values. You can list up to 65535 values in an ENUM list. If a value is inserted that is not in the list, a blank value will be inserted. The values are sorted in the order you enter them |
| SET(val1, val2, val3, ...) | A string object that can have 0 or more values, chosen from a list of possible values. You can list up to 64 values in a SET list |

Numeric Data Types

|  |  |
| --- | --- |
| **Data type** | **Description** |
| BIT(size) | A bit-value type. The number of bits per value is specified in size. The size parameter can hold a value from 1 to 64. The default value for size is 1. |
| TINYINT(size) | A very small integer. Signed range is from -128 to 127. Unsigned range is from 0 to 255. The size parameter specifies the maximum display width (which is 255) |
| BOOL | Zero is considered as false, nonzero values are considered as true. |
| BOOLEAN | Equal to BOOL |
| SMALLINT(size) | A small integer. Signed range is from -32768 to 32767. Unsigned range is from 0 to 65535. The size parameter specifies the maximum display width (which is 255) |
| MEDIUMINT(size) | A medium integer. Signed range is from -8388608 to 8388607. Unsigned range is from 0 to 16777215. The size parameter specifies the maximum display width (which is 255) |
| INT(size) | A medium integer. Signed range is from -2147483648 to 2147483647. Unsigned range is from 0 to 4294967295. The size parameter specifies the maximum display width (which is 255) |
| INTEGER(size) | Equal to INT(size) |
| BIGINT(size) | A large integer. Signed range is from -9223372036854775808 to 9223372036854775807. Unsigned range is from 0 to 18446744073709551615. The size parameter specifies the maximum display width (which is 255) |
| FLOAT(size, d) | A floating point number. The total number of digits is specified in size. The number of digits after the decimal point is specified in the d parameter. This syntax is deprecated in MySQL 8.0.17, and it will be removed in future MySQL versions |
| FLOAT(p) | A floating point number. MySQL uses the p value to determine whether to use FLOAT or DOUBLE for the resulting data type. If p is from 0 to 24, the data type becomes FLOAT(). If p is from 25 to 53, the data type becomes DOUBLE() |
| DOUBLE(size, d) | A normal-size floating point number. The total number of digits is specified in size. The number of digits after the decimal point is specified in the d parameter |
| DOUBLE PRECISION(size, d) |  |
| DECIMAL(size, d) | An exact fixed-point number. The total number of digits is specified in size. The number of digits after the decimal point is specified in the d parameter. The maximum number for size is 65. The maximum number for d is 30. The default value for size is 10. The default value for d is 0. |
| DEC(size, d) | Equal to DECIMAL(size,d) |

Date and Time Data Types

|  |  |
| --- | --- |
| **Data type** | **Description** |
| DATE | A date. Format: YYYY-MM-DD. The supported range is from '1000-01-01' to '9999-12-31' |
| DATETIME(fsp) | A date and time combination. Format: YYYY-MM-DD hh:mm:ss. The supported range is from '1000-01-01 00:00:00' to '9999-12-31 23:59:59'. Adding DEFAULT and ON UPDATE in the column definition to get automatic initialization and updating to the current date and time |
| TIMESTAMP(fsp) | A timestamp. TIMESTAMP values are stored as the number of seconds since the Unix epoch ('1970-01-01 00:00:00' UTC). Format: YYYY-MM-DD hh:mm:ss. The supported range is from '1970-01-01 00:00:01' UTC to '2038-01-09 03:14:07' UTC. Automatic initialization and updating to the current date and time can be specified using DEFAULT CURRENT\_TIMESTAMP and ON UPDATE CURRENT\_TIMESTAMP in the column definition |
| TIME(fsp) | A time. Format: hh:mm:ss. The supported range is from '-838:59:59' to '838:59:59' |
| YEAR | A year in four-digit format. Values allowed in four-digit format: 1901 to 2155, and 0000. MySQL 8.0 does not support year in two-digit format. |

# SQL Operators

Arithmetic Operators

|  |  |
| --- | --- |
| Operator | Description |
| + | Add |
| - | Subtract |
| \* | Multiply |
| / | Divide |
| % | Modulo |

Bitwise Operators

|  |  |
| --- | --- |
| Operator | Description |
| & | Bitwise AND |
| | | Bitwise OR |
| ^ | Bitwise exclusive OR |

Comparison Operators

|  |  |
| --- | --- |
| Operator | Description |
| = | Equal to |
| > | Greater than |
| < | Less than |
| >= | Greater than or equal to |
| <= | Less than or equal to |
| <> | Not equal to |

Compound Operators

|  |  |
| --- | --- |
| Operator | Description |
| += | Add equals |
| -= | Subtract equals |
| \*= | Multiply equals |
| /= | Divide equals |
| %= | Modulo equals |
| &= | Bitwise AND equals |
| ^-= | Bitwise exclusive equals |
| |\*= | Bitwise OR equals |

Logical Operators

|  |  |
| --- | --- |
| Operator | Description |
| ALL | TRUE if all of the subquery values meet the condition |
| AND | TRUE if all the conditions separated by AND is TRUE |
| ANY | TRUE if any of the subquery values meet the condition |
| BETWEEN | TRUE if the operand is within the range of comparisons |
| EXISTS | TRUE if the subquery returns one or more records |
| IN | TRUE if the operand is equal to one of a list of expressions |
| LIKE | TRUE if the operand matches a pattern |
| NOT | Displays a record if the condition(s) is NOT TRUE |
| OR | TRUE if any of the conditions separated by OR is TRUE |
| SOME | TRUE if any of the subquery values meet the condition |

# Comments

/\* This is a

Multiline comment

\*/

This is a single line comment

# DDL SQL Queries

1. CREATE DATABASE IF NOT EXISTS dbname
2. DROP DATABASE dbname
3. CREATE TABLE IF NOT EXISTS users(

col1 datatype,

col2 datatype,

email varchar(255) UNIQUE,

name varchar(255),

UNIQUE(name),

CONSTRAINT U\_email UNIQUE (email),

id integer NOT NULL PRIMARY KEY AUTO\_INCREMENT,

PRIMARY KEY (id, email),

age integer CHECK (age > 6 AND age < 25),

gender varchar(255) DEFAULT “Others”

journey\_date datetime DEFAULT CURRENT\_TIMESTAMP

)

1. DROP TABLE tablename
2. TRUNCATE TABLE tablename
3. CREATE TABLE orders(

user\_id integer,

FOREIGN KEY(user\_id) REFERNCES users(id)

)

1. ALTER TABLE students ADD COLUMN college varchar(255) NOT NULL
2. ALTER TABLE students DROP COLUMN age
3. ALTER TABLE students MODIFY COLUMN sname integer
4. ALTER TABLE students DROP CONSTRAINT U\_email
5. ALTER TABLE passenger ADD CONSTRAINT P\_email UNIQUE(email)
6. ALTER TABLE passenger DROP CONSTRAINT P\_email

# DML SQL Queries

1. INSERT INTO students (sid, sname, email, college) VALUES (1, “Nitish”, [abc@gmail.com](mailto:abc@gmail.com), “HIT”)
2. INSERT INTO students VALUES (5, “Rahul”, [abc@gmail.com](mailto:abc@gmail.com), “IEM”)
3. INSERT INTO students VALUES (6, “ewg”, “wrg”, “wrg”), (7, “wrgw”, “wrgw”, “wrgw”)
4. SELECT \* FROM tablename
5. SELECT colname1, colname2 FROM train
6. SELECT Name AS PassengerName, Sex AS Gender, Survived FROM train
7. SELECT Name, SibSp + Parch AS family FROM train
8. SELECT Name, Age + 102 AS CurrentAge FROM train
9. SELECT Name, 100000 AS Compensation FROM train
10. SELECT DISTINCT Sex FROM train
11. SELECT DISTINCT Pclass, Embarked FROM train
12. SELECT \* FROM train WHERE Survived = 0
13. SELECT \* FROM train WHERE Pclass = 3 AND Survived = 0
14. SELECT \* FROM train WHERE Age BETWEEN 10 AND 15
15. SELECT title FROM movies WHERE genre LIKE ‘Comedy’ OR genre LIKE ‘Action’
16. SELECT title, genre FROM movies WHERE genre IN (‘Action’, ‘Horror’, ‘Drama’)
17. SELECT title, genre FROM movies WHERE genre NOT IN (‘Action’, ‘Horror’, ‘Drama’)
18. SELECT title FROM movies WHERE title LIKE ‘A%’
19. SELECT title FROM table\_name WHERE title LIKE ‘%man%’
20. SELECT title FROM movies WHERE actor LIKE ‘%Khan%’ OR actor LIKE ‘%Kapoor%’
21. SELECT title FROM movies WHERE title LIKE ‘\_\_\_\_’
22. SELECT title FROM movies WHERE title LIKE ‘A\_\_\_’
23. UPDATE passenger SET name = ‘Rahul’’
24. UPDATE passenger SET name = ‘Rohit’ WHERE email LIKE ‘%gmail’
25. UPDATE passenger SET name = ‘Ankit’, email =’abc@gmail.com’ WHERE email LIKE ‘%yahoo%’
26. DELETE FROM passenger WHERE id = 1
27. DELETE FROM passenger WHERE id > 2 AND email LIKE ‘%yahoo%’
28. DELETE FROM passenger WHERE 1
29. SELECT title, ABS((india\_gross – budget)) AS profit FROM movies
30. SELECT title, ROUND((runtime/60)) AS runtime\_hrs FROM movies
31. SELECT title, ROUND((runtime/60), 1) AS runtime\_hrs FROM movies
32. SELECT title, CEIL((runtime/60)) AS runtime\_hrs FROM movies
33. SELECT title, FLOOR((runtime/60)) AS runtime\_hrs FROM movies
34. SELECT UPPER(title) FROM movies
35. SELECT title, CONCAT(actor, ‘ ‘, director) AS crew FROM movies
36. SELECT title, LENGTH(title) AS length FROM movies
37. SELECT title, SUBSTR(title, 1, 5) AS short FROM movies
38. SELECT MAX(budget) FROM movies
39. SELECT MIN(india\_gross) FROM movies
40. SELECT SUM(india\_gross) FROM movies
41. SELECT AVG (india\_gross) FROM movies
42. SELECT COUNT (\*) FROM movies
43. SELECT COUNT(DISTINCT(actor)) FROM movies
44. SELECT title, (worldwide\_gross – budget) AS profit FROM movies ORDER BY profit DESC LIMIT 5
45. SELECT title, (worldwide\_gross – budget) AS profit FROM movies ORDER BY profit ASC LIMIT 5
46. SELECT \* FROM movies ORDER BY genre, title DESC
47. SELECT actor, COUNT(\*) AS num\_movies FROM movies GROUP BY actor ORDER BY num\_movies DESC LIMIT 5
48. SELECT genre, SUM (worldwide\_gross – budget) AS total\_profit FROM movies GROUP BY genre ORDER BY total\_profit DESC LIMIT 5
49. SELECT director, AVG (worldwide\_gross – budget) AS avg\_profit FROM movies GROUP BY director ORDER BY avg\_profit DESC LIMIT 5
50. SELECT title, budget FROM movies GROUP BY title ORDER BY budget DESC LIMIT 5
51. SELECT actor, director, SUM (worldwide\_gross – budget) AS profit FROM movies GROUP BY actor, director ORDER BY profit DESC LIMIT 5
52. SELECT actor, SUM (worldwide\_gross – budget) AS total\_profit FROM movies GROUP BY actor ORDER BY total\_profit DESC LIMIT 10
53. SELECT actor, AVG (screens) AS opening

FROM movies

GROUP BY actor

HAVING opening > 1000

ORDER BY opening DESC

1. SELECT title, (worldwide\_gross - budget) profit,

CASE

WHEN (worldwide\_gross – budget) > 100000 THEN “SUPER HIT”

ELSE “FLOP”

END AS verdict

FROM movies

# SQL Joins

1. SELECT \* FROM users CROSS JOIN groups
2. SELECT \* FROM membership m JOIN users u ON m.uid = u.id

# What is

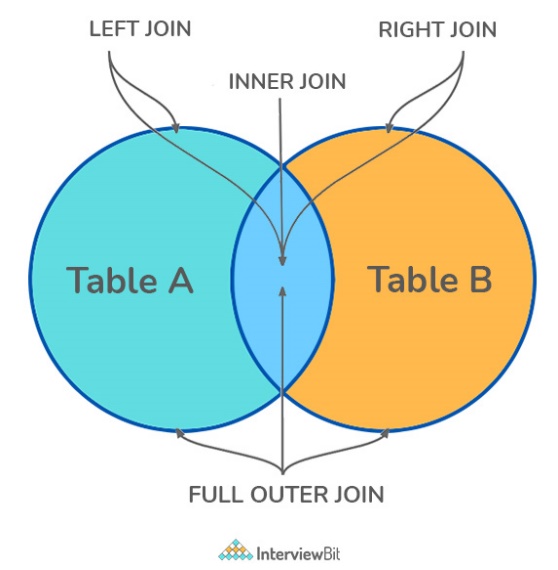
puts lines from queries after each other. removes duplicate records (where all columns in the results are the same). Example

23 bah

45 bah

# What is a

The SQL clause is used to combine records (rows) from two or more tables in a SQL database based on a related column between the two.



There are four different types of JOINs in SQL:

* : Retrieves records that have matching values in both tables involved in the join. This is the widely used join for queries.

Table\_A

Table\_B

Table\_A

Table\_B

* : Retrieves all the records/rows from the left and the matched records/rows from the right table.

Table\_A A

Table\_B B

A.col = B.col

* : Retrieves all the records/rows from the right and the matched records/rows from the left table.

Table\_A A

Table\_B B

A.col = B.col

* : Retrieves all the records where there is a match in either the left or right table.

Table\_A A

Table\_B B

A.col = B.col

# What is Self-Join

A self-JOIN is a case of regular join where a table is joined to itself based on some relation between its own columns. Self-join uses the or clause and a table alias is used to assign different names to the table within the query.

A.emp\_id “Emp\_ID”, A.emp\_name “Employee”,

B.emp\_id “Sup\_ID”, B.emp\_name “Supervisor”

employee A, employee B

A.emp\_sup = B.emp\_id

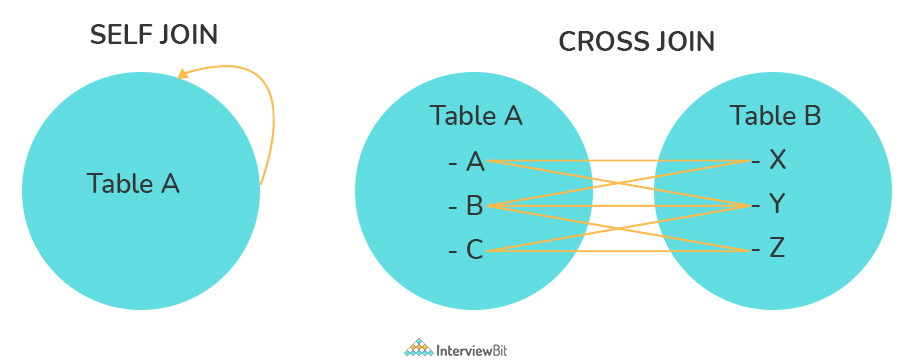
# What is a

Cross join can be defined as a cartesian product of the two tables included in the join. The table after join contains the same number of rows as in the cross-product of the number of rows in the two tables.

stu.name, sub.subject

students stu

subjects sub



# What is an Index

A database index is a data structure that provides a quick lookup of data in a column or columns of a table. It enhances the speed of operations accessing data from a database table at the cost of additional writes and memory to maintain the index data structure.

INDEX index\_name

table\_name (column\_1, column\_2)

INDEX index\_name

There are different types of indexes that can be created for different purposes:

* Unique and Non-unique Index

Unique indexes are indexes that help maintain data integrity by ensuring that no two rows of data in a table have identical key values. Once a unique index has been defined for a table, uniqueness is enforced whenever keys are added or changed within the index.

Non-unique indexes, are not used to enforce constraints on the tables with which they are associated. Instead, non-unique indexes are used solely to improve query performance by maintaining a sorted order of data values that are used frequently.

INDEX myIndex

students (enroll\_no)

* Clustered and Non-Clustered Index

Clustered indexes are those whose order of the rows in the database corresponds to the order of the rows in the index. This is why only one clustered index can exist in a given table, whereas, multiple non-clustered indexes can exist in the table.

The only difference between clustered and non-clustered indexes is that the database manager attempts to keep the data in the database in the same order as the corresponding keys appear in the clustered index.

Clustering indexes can improve the performance of most query operations because they provide a linear-access path to data stored in the database.

# What is a Query

It is a request for data or information from a database table or combination of tables. A database query can be either a select query or an action query.

fname, lname /\* select query \*/

myDb.students

student\_id = 1

myDB.students /\* action query \*/

fname = ‘Captain’, lname = ‘America’

student\_id = 1

# What is a Subquery

A subquery is a query within another query, also known as nested query or inner query. It is used to restrict or enhance the data to be queried by the main query, thus restricting or enhancing the output of the main query respectively. For example, here we fetch the contact information for students who have enrolled for the maths subject:

name, email, mob, address

myDb.contacts

roll\_no ( roll\_no myDb.students subject = ‘Maths’)

There are two types of subqueries – Correlated and Non-Correlated

* A correlated subquery cannot be considered as an independent query, but it can refer to the column in a table listed in the of the main query.
* A non-correlated subquery can be considered as an independent query and the output of the subquery is substituted in the main query.

# What is the statement

operator in SQL is used to select data from a database. The data returned is stored in a result table, called the result-set.

myDB.students

Some common SQL clauses used in conjunction with a query are as follows:

* clause in SQL is used to filter records that are necessary, based on specific conditions
* clause in SQL is used to sort the records based on some fields in ascending () or descending order ()
* clause is used to group records with identical data and can be used in conjunction with some aggregation functions to produce summarized results from the database.
* clause is used to filter records in combination with the . It is different from , since the clause cannot filter aggregated records

myDB.students

graduation\_year = 2019

studentID

COUNT(studentID), country

myDB.students

country != “INDIA”

country

COUNT(studentID) > 5

# What are , and commands

The operator combines and returns the result-set retrieved by two or more statements.

The operator is used to remove duplicates from the result-set.

The clause in SQL combines the result-set fetched by the two statements where records from one match the other and then returns this intersection of result-sets.

name Students

name Contacts

name Students

name Contacts

name Students

name Contacts

name Students

name Contacts

# What is an Alias in SQL

It is a temporary name assigned to the table or table column for the purpose of a particular SQL query. Aliasing can be employed as an obfuscation technique to secure the real names of database fields. A table alias is also called a correlation name.

An alias is represented explicitly by the AS keyword.

A.emp\_name “Employee”

B.emp\_name “Supervisor”

employee A, employee B

A.emp\_sup = B.emp\_id

# What are DELETE, TRUCATE and DROP statements

statement is used to delete rows from a table based on the condition given in the clause or deletes all the rows from the table if no condition is specified. But it does not free the space containing the table.

command is used to delete all the rows from the table and free the space containing the table.

command is used to remove an object from the database. IF you drop a table, all the rows in the table are deleted and the table structure is removed from the database.

Candidates

CandidateID > 1000

Candidates

Candidates

# What is User-defined function

The user-defined functions in SQL are functions that accept parameters, perform complex calculations, and return a value. There are two types of SQL user-defined functions:

* Scalar Function: Return a single scalar value based on the input value.

– Calculates the total length of the given field (column)

– Converts a collection of string values to uppercase characters

– Converts a collection of string values to uppercase characters

– Extracts substrings from a collection of string values in a table

– Concatenates two or more strings

– Generates a random collection of numbers of a given length

– Calculates the round-off integer value for a numeric field (or decimal point value)

– Returns the current date and time

– Sets the format to display a collection of values

* Table-Valued Function: Return a table as output

Inline: returns a table data type based on a single statement.

Multi-statement: returns a tabular result-set but, unlike inline, multiple statements can be used inside the function body.

# What is Aggregate function

An aggregate function performs operations on a collection of values to return a single scalar value. Aggregate funcitons are often used with and clauses of the statement. Following are the widely used SQL aggregate functions

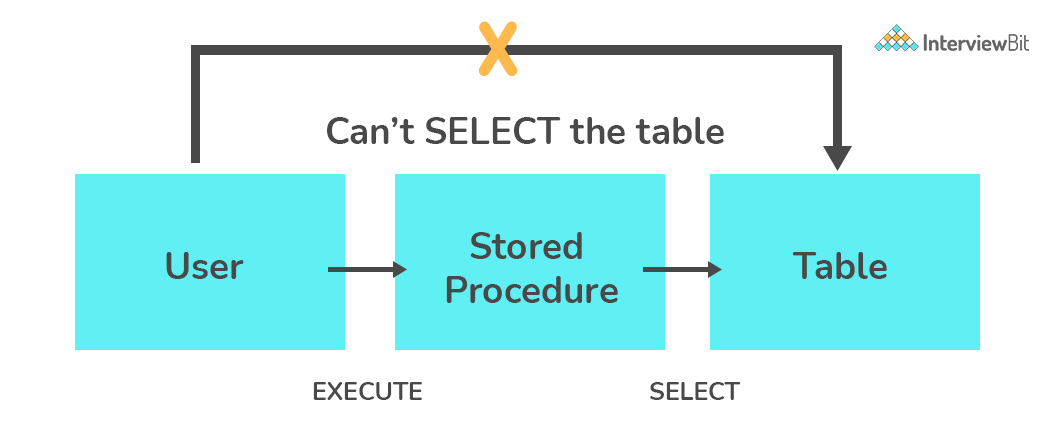
* – Calculates the mean of a collection of values
* – Counts the total number of records in a specific table or view
* – Calculates the minimum of a collection of values
* – Calculates the maximum of a collection of values
* – Fetches the first element in a collection of values
* – Fetches the last element in a collection of values

# What is a Stored Procedure

A stored procedure is a subroutine available to applications that access a RDBMS. Such procedures are stored in the database data dictionary. Disadvantage of stored procedure is that it can be executed nowhere except in the database and occupies who more memory in the database server. It also provides a sense of security and functionality as users who can’t access the data directly can be granted access via stored procedures.

FetchAllStudents()

myDB.students



# What is a Recursive Stored Procedure

A stored procedure that calls itself until a boundary condition is reached, is called a recursive stored procedure. This recursive function helps the programmers to deploy the same set of code several times as and when required. Some SQL programming languages limit the recursion depth to prevent an infinite loop of procedure calls from causing a stack overflow, which slows down the system and may lead to system crashes.

# What is Pattern Matching in SQL

SQL pattern matching provides for pattern search in data. This SQL query uses wildcards to match a string pattern, rather than writing the exact word. The operator is used in conjunction with SQL Wildcards to fetch the required information.

* Using a % wildcard to perform a simple search

The % wildcard matches zero or more characters of any type and can be used to define wildcards both before and after the pattern.

Example: Search a student in database with the first name beginning with the letter K

students

first\_name ‘K%’

* Omitting the patterns using the NOT keyword

Use the keyword to select records that don’t match the pattern.

Example: This query returns all students whose first name does not begin with K

students

first\_name ‘K%’

* Matching a pattern anywhere using the % wildcard twice

Example: Search for a student in database where he/she has a K in his/her first name

students

first\_name ‘%Q%’

* Using the \_ wildcard to match the pattern at a specific position

The \_ wildcard matches exactly one character of any type. It can be used in conjunction with % wildcard. This query fetches all students with letter K at the third position in their first name

students

first\_name ‘\_\_K%’

* Matching patterns for a specific length

The \_ wildcard plays an important role as a limitation when it matches exactly one character. It limits the length and position of the matched results.

/\* Matches first names with three or more letters \*/

students

first\_name ‘\_\_%’

students

first\_name ‘\_\_’

# What is the difference between DDL, DML, DCL, and TCL

These represents four categories into which the SQL commands have been separated. The following are the four significant subsets of the SQL.

* Data Definition Language (DDL) – Involves SQL commands used to define data structures – CREATE, ALTER, TRUNCATE, and DROP
* Data Manipulation Language (DML) – Involves SQL commands used to manipulate data – SELECT, INSERT, UPDATE, DELETE.
* Data Control Language (DCL) – Involves SQL commands used commonly by database administrators (DBAs) to manage permissions – GRANT, REVOKE

GRANT: It enables system administrators to assign privileges and roles to the specific user accounts to perform specific tasks on the database.

REVOKE: It enables system administrators to revoke privileges and roles from the user accounts so that they cannot use the previously assigned permission on the database.

* Transaction Control Language (TCL) – To ensure the transactions that occur in the database to happen in such a way that minimizes the danger of suffering from data loss. The commands in this category are COMMIT, ROLLBACK, SET TRANSACTION, SAVEPOINT, etc.

# References

* Mlstack.cafe