Statistics :

* **Data** — a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process
* Note: Categorical Data can be visualized by Bar Plot, Pie Chart, [Pareto Chart](https://en.wikipedia.org/wiki/Pareto_chart). Numerical Data can be visualized by Histogram, Line Plot, Scatter Plot
* A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features of a collection of information. It helps us in knowing our data better. It is used to describe the characteristics of data.
* **Nominal**: Data at this level is categorized using names, labels or qualities. eg: Brand Name, ZipCode, Gender.
* **Ordinal**: Data at this level can be arranged in order or ranked and can be compared. eg: Grades, Star Reviews, Position in Race, Date
* **Interval**: Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. eg: Temperature in Celsius, Year of Birth.
* **Ratio:** Data at this level is similar to interval level with added property of an inherent zero. Mathematical calculations can be performed on these data points. eg: Height, Age, Weight
* **Population:**Collection of all items (N) and it includes each and every unit of our study. It is hard to define and the measure of characteristic such as mean, mode is called parameter.
* **Sample:** Subset of the population (n) and it includes only a handful units of the population. It is selected at random and the measure of the characteristic is called as statistics.
* The **measure of central tendency** is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.
* **Mean:** The mean is equal to the sum of all the values in the data set divided by the number of values in the data set i.e the calculated average. It susceptible to outliers when unusual values are added it gets skewed i.e deviates from the typical central value.
* **Median:** The median is the middle value for a dataset that has been arranged in order of magnitude. Median is a better alternative to mean as it is less affected by outliers and skewness of the data. The median value is much closer than the typical central value.
* **Mode:** The mode is the most commonly occurring value in the dataset. The mode can, therefore sometimes consider the mode as being the most popular option.
* **Measures of Asymmetry**
* **Skewness:** Skewness is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed towards to the left or to the right. Skewness indicates whether the data is concentrated on one side.
* **Positive Skewness:** Positive Skewness is when the mean>median>mode. The outliers are skewed to the right i.e the tail is skewed to the right.
* **Negative Skewness:** Negative Skewness is when the mean<median<mode. The outliers are skewed to the left i.e the tail is skewed to the left.
* **Skewness** is important as it tells us about where the data is distributed.
* **Measures of Variability(Dispersion)**
* The measure of central tendency gives a single value that represents the whole value; however, the central tendency cannot describe the observation fully. The measure of dispersion helps us to study the variability of the items i.e the spread of data.
* **Remember:** Population Data has N data points and Sample Data has (n-1) data points. (n-1) is called Bessel’s Correction and it is used to reduce bias.
* **Range:** The difference between the largest and the smallest value of a data, is termed as the range of the distribution. Range does not consider all the values of a series, i.e. it takes only the extreme items and middle items are not considered significant. eg: For {13,33,45,67,70} the range is 57 i.e(70–13).
* **Variance:** Variance measures how far is the sum of squared distances from each point to the mean i.e the dispersion around the mean.Variance is the average of all squared deviations.
* **Standard Deviation:**AsVariance suffers from unit difference so standard deviation is used. The square root of the variance is the standard deviation. It tells about the concentration of the data around the mean of the data set.
* **Coefficient of Variation(CV):** It is also called as the relative standard deviation. It is the ratio of standard deviation to the mean of the dataset.
* Standard deviation is the variability of a single dataset. Whereas the coefficient of variance can be used for comparing 2 datasets.
* **Measures of Quartiles**
* Quartiles are better at understanding as every data point considered.
* **Measures of Relationship**
* Measures of relationship are used to find the comparison between 2 variables.
* **Covariance:** Covariance is a measure of the relationship between the variability of 2 variables i.e It measures the degree of change in the variables, when one variable changes, will there be the same/a similar change in the other variable.
* Covariance does not give effective information about the relation between 2 variables as it is not normalized.
* **Correlation:** Correlation gives a better understanding of covariance. It is normalized covariance. Correlation tells us how correlated the variables are to each other. It is also called as Pearson Correlation Coefficient.
* The value of correlation ranges from -1 to 1. -1 indicates negative correlation i.e with an increase in 1 variable independent there is a decrease in the other dependent variable.1 indicates positive correlation i.e with an increase in 1 variable independent there is an increase in the other dependent variable.0 indicates that the variables are independent of each other.
* **Standard deviation** and variance indicate the spread of a data distribution around its mean.
* Standard deviation is a quantity that expresses the value by which the members of a group differ from the mean value for the group.
* Its symbol is σ (the Greek letter sigma). If the data points are further from the mean, there is higher deviation within the data set.
* Standard deviation is used more often than variance because the unit in which it is measured is the same as that of mean, a measure of central tendency.
* **Variance (σ2 )** refers to the spread of the data set, for example, how far the numbers are in relation to the mean. Variance is particularly useful when calculating the probability of future events or performance
* **Covariance** is the measure of how two random variables change together. It is used to calculate the correlation between variables.
* A positive covariance indicates that both variables from the prior line tend to move upward and downward in value at the same time. An inverse or negative covariance means that variables move counter to each other: when one rises, the other falls.
* **The Logistic Sigmoid** is a useful function that follows the S curve. It saturates when input is very large or very small.
* **The distribution** where the data tends to be around a central value with lack of bias or minimal bias toward the left or right is called **Gaussian distribution, also known as normal distribution.**
* A “standard normal distribution” has μ = 0 and σ = 1
* Multivariate normal distribution is the generalization of the univariate normal distribution to multiple variables.
* **The Random sampling** techniques uses for randomization to make sure that every elements of the population gets an equal chance to be part of the selected sample.
* **Outliers** are values that are significantly different from all other observations. any data value that lies more than(1.5 IQR) away from the Q1 and Q3 quartiles is considered an outlier. Visualization methods is boxplot , histogram, scatter plot.
* **Statistics** is a Mathematical **Science** pertaining to **data** collection, analysis, interpretation and presentation. **Statistics** – Math And **Statistics** For **Data Science**. **Statistics** is used to process complex problems in the real world so that **Data Scientists** and Analysts can look for meaningful trends and changes in **Data**.
* **Numerical data** represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.
* **There are two qualitative levels**: nominal and ordinal. The nominal level represents categories that cannot be put in any order, while ordinal represents categories that can be ordered.
* **There are two quantitative levels:** interval and ratio. They both represent “numbers”, however, ratios have a true zero, while intervals don’t.
* **Frequency distribution** tables show the category and its corresponding absolute frequency.
* **Bar charts** are very common. Each bar represents a category. On the y-axis we have the absolute frequency.
* **Pie charts** are used when we want to see the share of an item as a part of the total. Market share is almost always represented with a pie chart.
* **The Pareto diagram** is a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency
* **Histograms** are the one of the most common ways to represent numerical data. Each bar has width equal to the width of the interval. The bars are touching as there is continuation between intervals: where one ends -> the other begins.
* **Cross tables (or contingency tables)** are used to represent categorical variables. One set of categories is labeling the rows and another is labeling the columns. We then fill in the table with the applicable data. It is a good idea to calculate the totals.
* When we want to represent two numerical variables on the same graph, we usually use a **scatter plot**. Scatter plots are useful especially later on, when we talk about regression analysis, as they help us detect patterns (linearity, homoscedasticity).
* The **mean** is the most widely spread measure of central tendency. It is the simple average of the dataset.
* The **median** is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.
* The **mode** is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes.
* **Skewness** is a measure of asymmetry that indicates whether the observations in a dataset are concentrated on one side. Right (positive) skewness looks like the one in the graph. It means that the outliers are to the right (long tail to the right). Left (negative) skewness means that the outliers are to the left.
* **Covariance** is a measure of the joint variability of two variables. A positive covariance means that the two variables move together. A covariance of 0 means that the two variables are independent. A negative covariance means that the two variables move in opposite directions. Covariance can take on values from -∞ to +∞ . This is a problem as it is very hard to put such numbers into perspective.
* **Correlation** is a measure of the joint variability of two variables. Unlike covariance, correlation could be thought of as a standardized measure. It takes on values between -1 and 1, thus it is easy for us to interpret the result. A correlation of 1, known as perfect positive correlation, means that one variable is perfectly explained by the other. A correlation of 0 means that the variables are independent. A correlation of -1, known as perfect negative correlation, means that one variable is explaining the other one perfectly, but they move in opposite directions.
* **To accept the null hypothesis** means that there isn’t enough data to support the change or the innovation brought by the alternative.
* **To reject the null hypothesis** means that there is enough statistical evidence that the status-quo is not representative of the truth.
* **Level of significance (α)**- The probability of rejecting a null hypothesis that is true; the probability of making this error.
* The probability of committing **Type I error (False positive)** (When you reject true null hypothesis) is equal to the significance level (α).
* The probability of committing **Type II error (False negative)**(When you accept false null hypothesis) is equal to the beta (β).
* **The p-value** is the smallest level of significance at which we can still reject the null hypothesis, given the observed sample statistic.
* 0.05 is often the ‘cut-off line’. If our p-value is higher than 0.05 we would normally accept the null hypothesis (equivalent to testing at 5% significance level). If the p-value is lower than 0.05 we would reject the null.
* In probability theory and statistics,a **probability distribution** is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. Examples: Normal distribution, Student’s T distribution, Poisson distribution, Uniform distribution, Binomial distribution.
* **The Normal distribution** is also known as Gaussian distribution or the Bell curve. It is one of the most common distributions due to the following reasons: • It approximates a wide variety of random variables • Distributions of sample means with large enough samples sizes could be approximated to normal • All computable statistics are elegant • Heavily used in regression analysis • Good track record.
* **Controlling for the standard deviation**: Keeping the standard deviation constant, the graph of a normal distribution with: • a smaller mean would look in the same way, but be situated to the left • a larger mean would look in the same way, but be situated to the right.
* **Controlling for the mean:** Keeping the mean constant, a normal distribution with: • a smaller standard deviation would be situated in the same spot, but have a higher peak and thinner tails • a larger standard deviation would be situated in the same spot, but have a lower peak and fatter tails.
* **The Standard Normal distribution** is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1.
* **The Central Limit Theorem (CLT)** is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution. Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples. The CLT applies whenever we have a sum or an average of many variables (e.g. sum of rolled numbers when rolling dice).
* The CLT allows us to assume normality for many different variables. That is very useful for confidence intervals, hypothesis testing, and regression analysis. In fact, the Normal distribution is so predominantly observed around us due to the fact that following the CLT, many variables converge to Normal.
* An **estimator** is a mathematical function that approximates a population parameter depending only on sample information.
* An **estimate** is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.
* A **confidence interval** is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall. We build the confidence interval around the point estimate. (1-α) is the level of confidence. We are (1-α)\*100% confident that the population parameter will fall in the specified interval. Common alphas are: 0.01, 0.05, 0.1.
* **The Student’s T distribution** is used predominantly for creating confidence intervals and testing hypotheses with normally distributed populations when the sample sizes are small. It is particularly useful when we don’t have enough information or it is too costly to obtain it.
* All else equal, the Student’s T distribution has fatter tails than the Normal distribution and a lower peak. This is to reflect the higher level of uncertainty, caused by the small sample size.
* A new discipline that combines the aspects of statistics, mathematics, programming, and visualization to turn data into information. When you combine domain expertise and scientific methods with technology, you get **Data Science**.
* **Probabilit**y gives the information about how likely an event can occur. in technical terms, probability is the measure of how likely an event is when an experiment is conducted.
* **the standard error** is the standard deviation of the sampling distribution. It takes the size of the sample into account.
* **Bias :** An unbiased estimator has an expected value the population parameter. A biased one has an expected value different from the population parameter. The bias is the deviation from the true value.
* **A confidence interval** is the range within which you expect the population parameter to be. You have a certain probability of it being correct, equal to the significance level.
* **Level of confidence:** Shows in what % of cases we expect the population parameter to fall into the confidence interval we obtained. Denoted 1 - α. Example: 95% confidence level means that in 95% of the cases, the population parameter will fall into the specified interval.
* **ANOVA :** Abbreviation of 'analysis of variance'. A statistical framework for analyzing variance of means.
* **Sum of squares total.** SST is the squared differences between the observed dependent variable and its mean.
* **Sum of squares regression**. SSR is the sum of the differences between the predicted value and the mean of the dependent variable. This is the variability explained by our model.
* **Sum of squares error.** SSE is the sum of the differences between the observed value and the predicted value. This is the variability that is NOT explained by our model.
* **R squired** : A measure ranging from 0 to 1 that shows how much of the total variability of the dataset is explained by our regression model.

Data Science With Python :

* **Data Science :** A powerful new approach to make discoveries from data. An automated way to analyze enormous amounts of data and extract information. A new discipline that combines the aspects of statistics, mathematics, programming, and visualization to turn data into information.
* **Data Analytics** is a combination of processes to extract information from datasets.
* **Hypothesis** is used to establish the relationship between dependent and independent variables. Hypothesis building begins in the data exploration stage, but becomes more mature in the conclusion or prediction phase.
* Calculating the difference between the two means is hypothesis testing.
* **Alternative Hypothesis** • Proposed model outcome is accurate and matches the data. • There is a difference between the means of S1 and S2.
* **Null Hypothesis** • Opposite of the alternative hypothesis. • There is no difference between the means of S1 and S2.
* **Statistics** is the study of the collection, analysis, interpretation, presentation, and organization of data.
* There are two major categories of statistics: Descriptive analytics and inferential analytics **Descriptive analytics** organizes the data and focuses on the main characteristics of the data.**Inferential analytics** is valuable when it is not possible to examine each member of the population.
* **Central tendency** indicates data accumulation toward the middle of the distribution or toward the end.
* **The bell curve** is characterized by its bell shape and two parameters, mean and standard deviation. Bell curve is: • Symmetric around the mean • Symmetric on both sides of the center • Having equal mean, median, and mode values • Denser in the center compared to the tails or sides • Defined by mean and standard deviation • Known as the Gaussian curve.
* **Hypothesis testing** is an inferential statistical technique that determines if a certain condition is true for the population.

1. Alternative Hypothesis (H1) Null Hypothesis (H0)
2. A statement that has to be concluded as true. A statement of no effect or no difference.
3. It’s a research hypothesis. It’s the logical opposite of the alternative hypothesis.
4. It needs significant evidence to support the initial hypothesis. It indicates that the alternative hypothesis is incorrect.
5. If the alternative hypothesis garners strong evidence, reject the null hypothesis. Weak evidence of alternative hypothesis indicates that the null hypothesis has to be accepted.

* **Chi – square test-** It is a hypothesis test that compares the observed distribution of your data to an expected distribution of data. Test is usually applied when there are two categorical variables from a single population.
* **Correlation coefficient** measures the extent to which two variables tend to change together. The coefficient describes both the strength and direction of the relationship.
* **Inferential statistics** uses a random sample from the data to make inferences about the population.

Machine Learning :

* **Artificial Intelligence** refers to intelligence displayed by machines that simulates human and animal intelligence.
* **Machine Learning** is an approach or subset of Artificial Intelligence that is based on the idea that machines can be given access to data along with the ability to learn from it. The capability of **Artificial Intelligence** systems to learn by extracting patterns from data is known as Machine Learning
* **Machine Learning** uses a number of theories and techniques from Data Science:
* **classification** is a technique in which the computer program learns from the data input given to it and then uses this learning to classify new observation.
* **Categorization:** A technique of organizing data into categories for its most effective and efficient use.
* **Clustering:** Technique 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.
* **Trend Analysis** is a technique aimed at projecting both current and future movement of events through use of time series data analysis.
* **Anomaly detection** is a technique to identify cases that are unusual within data that is seemingly homogeneous.
* **Visualization:** Technique to present data in a pictorial or graphical format. It enables decision makers to see analytics presented visually.
* **Decision making:** A technique/skill which Machine Learning Techniques provides you with the ability to influence managerial decisions with data as evidence for those possibilities.
* **Seaborn** is a library for making attractive and informative statistical graphics in Python. It is built on top of matplotlib and integrated with the PyData Stack, including support for numpy and pandas data structures, and statistical routines.
* The process of manually converting or mapping data from one raw format into another format is called **data wrangling**. This includes munging and data visualization.

## Supervised Learning

* **Supervised Learning** is a type of machine learning used to train models from labeled training data. It allows you to predict output for future or unseen data.
* If target variable is categorical (classes), then use **classification algorithm**. In other words, classification is applied when the output has finite and discreet values. Example: Predict the class of car given its features like horsepower, mileage, weight, colour, etc. The classifier will build its attributes based on these features. Analysis has three potential outcomes - Sedan, SUV, or Hatchback
* If target variable is a continuous numeric variable (100–2000), then use a **regression algorithm**. Example: Predict the price of a house given its sq. area, location, no of bedrooms, etc. A simple regression algorithm is given below y = w \* x + b This shows relationship between price (y) and sq. area (x) where price is a number from a defined range.
* **Linear Regression** is a statistical model used to predict the relationship between independent and dependent variables denoted by x and y respectively.
* **Multiple linear regression** is a statistical technique used to predict the outcome of a response variable through several explanatory variables and model the relationships between them.
* **Polynomial regression** is applied when data is not formed in a straight line. It is used to fit a linear model to non-linear data by creating new features from powers of non-linear features. Example: Quadratic features
* **R-square** is the most common accuracy metric to judge the performance of regression models.
* The disadvantage with R-squared is that it assumes every independent variable in the model explains variations in the dependent variable. Use **adjusted R-squared** when working on a multiple linear regression problem.
* **Mean-Squared Error (MSE)** is also used to measure the performance of a model. These functions are called the loss function or the cost function, and the value has to be minimized.
* **Gradient descent** is another algorithm used to reduce the loss function. It is an optimization algorithm that tweaks it’s parameters (coefficients) iteratively to minimize a given cost function to its minimum. Model stops learning when the gradient (slope) is zero.
* In regression analysis, **p-values** and coefficients together indicate which relationships in the model are statistically significant and the nature of those relationships. Coefficients describe the mathematical relationship between each independent variable and the dependent variable.
* **p-values** for the coefficients indicate whether these relationships are statistically significant.
* p < 0.05 REJECT the Null hypothesis, meaning variables have some effect and need to be retained p > 0.05 ACCEPT the Null hypothesis, meaning variables have no effect and can be removed.
* If the model learning is poor, you have an **underfitted** situation The algorithm will not work well on test data Retraining may be needed to find a better fit.
* **Overfitting** happens when model accuracy for training data is good, but model does not generalize well to the overall population Algorithm is not able to give good predictions for the new data.
* **Regularization** solves overfitting to the training data
* **Ridge Regression (L2)** is used when there is a problem of multicollinearity. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.
* **Lasso Regression (L1)** is similar to ridge, but it also performs feature selection. It will set the coefficient value for features that do not help in decision making very low, potentially zero. Lasso regression tends to exclude variables that are not required from the equation, whereas ridge tends to do better when all variables are present.
* **ElasticNet** regression combines the strength of lasso and ridge regression. ElasticNet regression combines the strength of lasso and ridge regression.
* **Logistic Regression** is widely used to predict binary outcomes for a given set of independent variables. The dependent variable’s outcome is discrete such as y ϵ {0, 1}. A binary dependent variable can have only two values such as 0 or 1, win or lose, pass or fail, healthy or sick.
* **ROC curve** Compares the model true positive and false positive rates to the ones from a random assignment.
* **AUC (Area under the ROC Curve)** Measures the entire two-dimensional area under the entire ROC curve.

## Feature Engineering

* **Regression** tells the relationship among variables and quantifies the relationship using set of equations.
* **Principal Component Analysis (PCA)** : Extracts hidden factors from the dataset .Defines your data using less number of components, explaining the variance in your data. Reduces the computational complexity. Determines whether a new data point is part of the group of data points from your training set.
* **Linear Discriminant Analysis (LDA**): Reduces Dimensions .Searches for a linear combination of variables that best separates 2 classes. Reduces the degree of overfitting. Determines how to classify a new observation out of a group of classes.

## Supervised Learning–Classification

* **Classification :** A machine learning task that identifies the class to which an instance belongs.
* **Decision Tree Classifier:** A tree-like structure in which the internal node represents the test on an attribute. Each branch represents the outcome of the test, and each leaf node represents the class label. A path from root to leaf represents classification rules.
* **Entropy** measures the impurity of a collection of examples. • It depends on the distribution of the random variable.
* **Information gain** is the expected reduction in entropy caused by partitioning the examples on an attribute. • Higher the information gain, the more effective the attribute in classifying training data.
* **Overfitting** occurs when the learning algorithm continues to develop hypotheses that reduce training set error at the cost of an increased test set error.
* **Bagging:** A technique for reducing the variance of an estimated prediction function.
* **Bootstrapping:** Randomly draws datasets with replacement from the training data.
* **Naive Baye’s classifier** assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

## Unsupervised Learning

* **Clustering:** Grouping objects based on the information found in data that describes the objects or their relationship.
* **Hierarchical Clustering:** Outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering.
* **Dendrogram** ((in Greek, dendro means tree and gramma means drawing) is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical.

## Time Series Modeling

* **Time Series** can be defined as a set of measurements of certain variablemade at regular time intervals.
* **A trend** is a long-term increase or decrease in time series data.
* When factors such as the time of the year or the day of the week affect the dependent variable, repetitive patterns are observed in the time series • **Seasonality** is always of a fixed and known frequency.
* Unlike seasonal patterns, **cyclic patterns** exhibit rise and fall that are not of fixed period • Duration is at least 2 years.
* **Irregular patterns** might occur due to random or unforeseen events • They are often of short duration and non-repeating.
* **A white noise** series is one with a zero mean, a constant variance, and no correlation between its values at different times.
* **Differencing** is performed by subtracting the previous observation from the current observation.
* **Detrending or de-seasonalizing** eliminates the trend and seasonality respectively. **Decomposition** is performed on the original series by regressing the series on time and taking the residuals from the regression.
* In an **AR model**, you predict future values based on a weighted sum of past values.
* **MA model** is used to forecast time series if Yt depends only on the random error terms.
* **ARMA model** is used to forecast time series using both the past values and the error terms.
* **ARIMA model** predicts a value in a response time series as a linear combination of its own past values, past errors, also current and past values of other time series.
* **Autocorrelation** refers to the way the observations in a time series are related to each other.
* **ACF** is the coefficient of correlation between the value of a point at a current time and its value at lag p, that is, correlation between Y(t) and Y(t-p). ACF will identify the order of MA process.
* **PACF** is similar to ACF, but the intermediate lags between t and t-p are removed, that is, correlation between Y(t) and Y(t-p) with p-1 lags excluded. PACF will identify the order of AR process.

## Ensemble Learning

* **Ensemble techniques** combine individual models together to improve the stability and predictive power of the model.
* Certain models do well in modeling one aspect of the data, while others do well in modeling another. Instead of learning a single complex model, learn several simple models and combine their output to produce the final decision. In ensemble learning, other models strength performs offset on individual model variances and biases. Ensemble learning will provide a composite prediction where the final accuracy is better than the accuracy of individual models.
* **Bagging or bootstrap aggregation** reduces variance of an estimate by taking mean of multiple estimates.
* **Boosting** reduces bias by training weak learners sequentially, each trying to correct its predecessor.
* **GBM** minimizes the loss function (MSE) of a model by adding weak learners using a gradient descent procedure.
* **eXtreme Gradient Boosting** is a library for developing fast and high-performance gradient boosting tree models. XGBoost is extensively used in ML competitions as it is almost 10 times faster than other gradient boosting techniques.

## Recommender Systems

* **Recommender system** is an information filtering technique, which provides users with recommendations, which they might be interested in.
* **Personalized Recommendations**: The most relevant item is recommended based on the user profile and contextual parameters.
* **Collaborative:** “Recommends what's popular among your peers".
* **Content-based:** "Displays similar to what you have liked".
* **Knowledge-based:** "Recommends you what fits, based on your needs”
* **Hybrid:** Combinations of various inputs and/or composition of different mechanism.
* **Collaborative Filtering:** It matches people with similar interests as a basis for recommendation.
* **Association Rule Generation:** Apriori Algorithm-Uses frequent itemset to generate association rules. A subset of a frequent itemset must also be a frequent itemset. Frequent itemset is a set of items whose support value > threshold value.

## Text Mining

* **Text mining** utilizes computational techniques to extract and summarize high-quality information from unstructured textual resources.
* **NLTK** is a set of open source Python modules used to work with human language data for applying statistical natural language processing (NLP). Provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet. Provides text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.
* **Tokenization** is a process of breaking running stream of text into words and sentences. It works by separating words using spaces and punctuation.
* **N-grams** are combinations of adjacent words or letters of lengthn in the source text.
* **Stop words** are natural languagewords which have very little meaning, such as “a", “an", "and", “the", and similar words. These are filtered out before processing of natural language data as they don’t reveal much information. Stop words are language dependent.
* **Stemming** involves reducing a word to stem or base (root) form by removing affixes. Various stemming algorithms : Porter stemmer, Lancaster stemmer, Snowball stemmer.
* **Lemmatization** uses vocabulary list and morphological analysis (POS of a word) to get the root word. Lemmatization uses WordNet database which has English words linked together by their semantic relationships.
* **POS tagging** marks words in the corpus to a corresponding part of a speech tag based on its context and definition.
* **NER(Named Entity Recognition)** seeks to extract a real-world entity from the text and sort it into pre-defined categories such as the names of persons, organizations, locations, etc.

Deep Learning with Keras and TensorFlow

* **Artificial intelligence (AI)** is a branch of computer science that attempts to simulate human intelligence in machines.
* The goal of **supervised learning** is to build a mapping function (f) that takes the input data (x) and predicts the output data (Y). The mapping function (f) has to achieve a certain level of performance to be considered fit for production.
* In **unsupervised learning**, only the inputs are known and the model is trained on the inputs to find the underlying patterns in the data.
* **Reinforcement learning** agents are goal-oriented. They learn by trial and error in an environment that provides rewards or penalties in response to the agents’ outputs. The aim is to find the best path that maximizes the likelihood of winning reward.
* **Positive Reinforcement:** Desirable stimulus is added to increase the likelihood of a behavior.
* **Negative Reinforcement:** Undesirable stimulus is removed to increase the likelihood of a behavior.
* **Deep Learning** is a subset of Machine Learning and is used to extract useful patterns from data. Deep Learning models make predictions independent of human intervention.
* The deep in Deep Learning refers to the large number of layers of neurons that help to learn various representations of data.
* Deep Learning models utilize artificial **neural networks (ANN)** that are inspired by the biological neural network of the human brain. ANNs try to imitate neuronal activity in the brain in order to learn to recognize data patterns.
* **The Black Box Problem:** Neural networks do not rely on rules established in advance. They find out patterns and correlations without exposing the reasons leading to them. Human users like to understand how a system made a given decision, as decisions are potential liabilities for domains like finance and medicine.
* **Long short-term memory (LSTM)** is a recurrent neural network (RNN) that selectively remembers patterns for long durations of time. LSTMs are mainly utilized in sequence prediction problems such as speech, images, and video.
* LSTMs have memory and take the order of observations into account. LSTMs need to remember the past to predict the future. Therefore, LSTMs learn any temporal dependence between observations.
* A **LSTM unit** is composed of a cell and three gates which control the flow of information.

## Artificial Neural Network

* An **artificial neuron** is analogous to biological neurons, where each neuron takes inputs, adds weights to them separately, sums them up, and passes this sum through a transfer function to produce a nonlinear output.
* **Perceptron :** Single layer neural network ▪ Consists of weights, the summation processor, and an activation function
* **The XOR Problem:** A perceptron can learn anything that it can represent, i.e., anything separable with a hyperplane. However, it cannot represent Exclusive OR since it is not linearly separable.
* **Learning rate** is used to control the changes made to the weights and biases in order to reduce the error. ▪ It is used to analyze how an error will change when the values of weights and biases are changed by a unit.
* **Epoch :** One epoch consists of one full cycle of training data ▪ Preferred value of number of training epochs is set to 1000.
* **Activation Functions:** Nonlinearities are needed to learn complex (non-linear) representations of data, otherwise the neural networks would be just a linear function.
* Takes a real-valued number and squashes it into a range of 0 to 1 ▪ **Sigmoid** neurons saturate and kill gradients, thus NN will barely learn.
* **Tanh:** Takes a real-valued number and squashes it into a range of -1 to 1 ▪ Like sigmoid, tanh neurons saturate ▪ Unlike sigmoid, output is zero-centered.
* **ReLU:** Takes a real-valued number and thresholds it at zero ▪ Most deep networks use ReLU nowadays ▪ Trains much faster ▪ Prevents the vanishing gradient problem.
* When a neural network contains more than one hidden layer it becomes **a Deep Neural Network.**
* L2 Regularization : Regularization penalizes big weights, in addition to the overall cost function ▪ Weight decay value determines how dominant regularization is during gradient computation ▪ Big weight decay coefficient implies big penalty for big weights.
* **Dropout Regularization:** Randomly drops units (along with their connections) during training ▪ Each unit is retained with fixed probability p, independent of other units ▪ 0 < p < 1 ▪ Hyper-parameter p has to be chosen (tuned) ▪ While testing the entire network gets activated while the weights get scaled by a factor of p.

## Deep Neural Network

* When a neural network contains more than one hidden layer it becomes a **Deep Neural Network**. In deep neural network, each layer recognizes a certain set of features based on the previous layer’s output.
* In a deep learning model, while predicting, the output deviates from the actual value, the quantitative measure of this difference is called **loss**.
* The losses of deep learning models can be evaluated very easily by using Loss Function.
* **MSE** is the average squared difference between actual and predicted value for N number of training data.
* **MAE** is the absolute difference between actual and predicted value for N number of training data.
* In MSE, since each error is squared, it penalizes even small differences in prediction when compared to MAE. In MSE, since each error is squared, it penalizes even small differences in prediction when compared to MAE. Effect of MSE is adverse on outliers. Since each error is squared in MSE, the final MSE also increases. If the data has outliers, MAE will be a better option over MSE. For data without outliers MSE is preferable.
* **Cross entropy** is a way to calculate distance between two probability distributions.
* **Binary cross entropy** assumes a binary value of 0 or 1 to denote negative and positive class respectively, when there is only one output.
* **What Is TensorFlow?** A popular open source library for deep learning and machine learning. Developed by Google Brain Team and released in 2015. Used mainly for classification, perception, understanding, discovering, prediction, and creation.
* **TFlearn** is a modular and transparent deep learning library built on top of Tensorflow. It was designed to provide a higher-level API to TensorFlow in order to facilitate and speed up experimentations, while remaining fully transparent and compatible with it.
* Training functions are another core feature of TFLearn. In Tensorflow, there are no prebuilt API to train a network, so TFLearn integrates a set of functions that can easily handle any neural network training, for any number of inputs, outputs, and optimizers.
* **What Is Keras**? Most powerful and easy to use for developing and evaluating deep learning models .A high-level neural network API, written in Python. Runs seamlessly on CPU and GPU.
* Keras uses TensorFlow, Theano, MxNet, and CNTK (Microsoft) as backends.
* **What Is PyTorch?** A replacement for NumPy to use the power of GPUs .A deep learning research platform that provides maximum flexibility and speed .A product of Facebook's artificial intelligence team.
* **Optimization** is choosing input to obtain the best possible output.
* Algorithms which are used to solve optimization problems are called **optimization algorithm**. In deep learning, optimization algorithms are used to optimize cost function J.
* **Gradient Descent (GD)** GD is used to minimize the cost function J and obtain the optimal weight W and bias b.
* **Stochastic Gradient Descent (SGD)** Single data points are taken to find the optimized weights. The mathematical formulation for the weight evaluation for SGD is same as GD, but the data points are shuffled before using them for optimization. Random data points go into the optimizer and result into random weights, that is the resulted weights are noisy.
* **Stochastic Gradient Descent-Mini Batch (SGD-Mini Batch)** Is combination of vanilla GD and SGD which distributes the whole training data in small mini-batches Divides the training data into small batches, so that the network can easily be trained on the data The mathematical formulation is same as vanilla GD, but the training occurs batch wise For example, training set has 400 training examples which are divided into 10 batches with each batch containing 40 training examples. Thus, the weight evaluation equation will be iterated over 10 times (number of batches).
* **SGD with momentum** or just momentum is an advanced optimization algorithm that uses moving average to update the trainable parameters. SGD with momentum is a very suitable method to overcome the noisy weights of SGD.
* **Nesterov Accelerated Gradient (NAG)** In NAG, interim parameters are observed if the velocity update leads to bad loss. In NAG, an interim velocity weight is calculated which is further used to calculate the weight with the help of a velocity factor. The difference between momentum method and NAG is in the gradient computation phase.
* Adaptive Gradient (AdaGrad) :
* **Root Mean Square Propagation (RMSprop**): RMSprop is developed to take care of the drawback of AdaGrad.
* **Batch Normalization** Weights of the neural network get updated during the training period, in each epoch. Suppose, weight assigned to a neuron suddenly become large, it cascades through all the layers which further causes instability. Normalization of data before feeding into network is not enough, the outputs from the neurons should also be normalized. This is where batch normalization comes into the picture.
* **Gradient** refers to the derivative of loss with respect to weight. It is calculated during the process of back propagation. It is used to update weights of the neural networks.
* When gradient becomes very small, subtracting it from the weight doesn't change the previous weight. Therefore, model stops learning. This problem of neural network is **called vanishing gradient.**
* **Exploding gradients** are a problem, where large error gradients accumulate and result in very large updates to neural network model weights during training.
* **Parameters** are found while training the model. For example, in K-mean clustering, the number of centroids is a model parameter.
* **Hyperparameters** are found before the training. A classic example of hyperparameter is the value of K in K-mean clustering which is decided before creating the model.
* **Hyperparameter tuning** is choosing a set of optimal hyperparameters for a learning algorithm
* Iterating over given hyperparameters using cross validation is called **Grid Search.**
* **Random search** is an optimization method used on functions that are not differentiable or continuous.
* **Gradient-based tuning** is used for algorithms, where it is possible to compute the hyperparameter with respect to the gradient and optimization of the hyperparameter is done by the gradient descent.
* **Evolutionary Optimization:** Uses evolutionary algorithm to find optimal hyperparameters .Used in black box functions with noises for global optimization.
* **Bayesian Optimization** Uses machine learning framework to predict optimal hyperparameters. Finds optimal hyperparameters from the result of previously built models with different hyperparameter configuration through the Gaussian process. Has the inherent property to study the trend in given data set which is not possible for a human.
* **Interpretability** is the degree of human’s ability to predict the model’s result consistently.
* **Intrinsic or Post Hoc:** Achieves interpretability by simplifying the machine learning model and analyzes the method after the training. Refers to models that are considered interpretable due to their simple structure.
* **Model-Specific or Model-Agnostic:** Can be used on any model and are applied after the model has been trained. Works by analyzing feature input and output pairs.

## Convolutional Neural Net (CNN)

* The idea of CNNs was neurobiologically motivated by the findings of locally-sensitive and orientation-selective nerve cells in the visual cortex. ▪ Inventors of CNN designed a network structure that implicitly extracts relevant features. ▪ Convolutional Neural Networks are a special kind of multilayer neural networks.
* **A CNN** is a neural network with convolutional layers (and other layers). A convolutional layer has several filters that perform the convolution operation. The CNN architecture comprises multiple combinations of convolution and pooling layers.
* **The pooling layer** gradually reduces the spatial size of each matrix within the feature map such that the amount of parameters and computation is reduced in the network.

## Recurrent Neural Networks (RNN)

* **What Is Sequential Data?** The dataset is said to be sequential when the data points are dependent on other data points within a dataset.
* **The RNN Model** : The RNN remembers the analysis done upto a given point by maintaining a state.
* **Bidirectional RNNs** are constructed by putting two RNNs (f1 and f2) together. Mathematically, these are defined as y,h=f1(x,h) and z,g = f2(g,x).
* **Deep RNNs** are constructed by adding more layers to simple RNNs. Mathematically, it can be defined as h’,y = f1(h,x), g’,z = f2(g,y)
* **Pyramid RNNs** speed up the training process by reducing the number of timesteps.
* **Deep RNNs** are very hard to train and usually don’t remember data beyond certain timesteps.
* **LSTM:** Tracks long-term dependencies while mitigating the vanishing or exploding gradient problems. It does so via input, forget, and output gates. Controls the exposure of memory content.
* **GRU:** Tracks long-term dependencies using a reset gate and an update gate Exposes the entire cell state to other units in the network.
* **The Attention Model:** Encoder: From word sequence to sentence representation Decoder: From representation to word sequence distribution Universal Representation: .
* representation of meaning.

## Autoencoders

* **Unsupervised learning** is to learn the structure of the data and correlations between features from input data with no labels. For example, clustering, compression, feature and representation learning, dimensionality reduction, generative models, etc.
* **An autoencoder** is a type of artificial neural network used to learn efficient data codings in an unsupervised manner.
* Components of an Autoencoder:
* **1.Encoder:** Reduces the input dimensions and compresses the input data into an encoded representation
* **2.Bottleneck:** Contains the compressed representation of the input data in its lowest possible form.
* **3.Decoder:** Reconstructs the data from the encoded representation to the original as same as possible.
* **4.ReconstructionLoss:** Measures the performance of the decoder and the similarities between the input and the output.
* **Anomaly detection** is the identification of rare items, events, or observations which raise suspicions by differing significantly from the majority of the data.

## An autoencoder, also known as autoassociator or Diabolo networks, is an artificial neural network employed to recreate the given input. It takes a set of unlabeled inputs, encodes them and then tries to extract the most valuable information from them. They are used for feature extraction, learning generative models of data, dimensionality reduction and can be used for compression. Now, autoencoders, based on Restricted Boltzmann Machines, are employed in some of the largest deep learning applications. They are the building blocks of Deep Belief Networks (DBN).

* An autoencoder uses the Loss function to properly train the network. The Loss function will calculate the differences between our output and the expected results. After that, we can minimize this error with gradient descent. There are more than one type of Loss function, it depends on the type of data.

Other: DL

## Deep Belief Network

* One problem with traditional multilayer perceptrons/artificial neural networks is that backpropagation can often lead to “local minima”. This is when your “error surface” contains multiple grooves and you fall into a groove that is not lowest possible groove as you perform gradient descent.
* **Deep belief networks** solve this problem by using an extra step called **pre-training**. Pre-training is done before backpropagation and can lead to an error rate not far from optimal. This puts us in the “neighborhood” of the final solution. Then we use backpropagation to slowly reduce the error rate from there.
* DBNs can be divided in two major parts. The first one are multiple layers of Restricted Boltzmann Machines (RBMs) to pre-train our network. The second one is a feed-forward backpropagation network, that will further refine the results from the RBM stack.

## Constructing the Layers of RBMs

First of all, let's detail Restricted Boltzmann Machines.

#### What are Restricted Boltzmann Machines?

RBMs are shallow neural nets that learn to reconstruct data by themselves in an unsupervised fashion.

#### How it works?

Simply, RBM takes the inputs and translates them to a set of numbers that represents them. Then, these numbers can be translated back to reconstruct the inputs. Through several forward and backward passes, the RBM will be trained, and a trained RBM can reveal which features are the most important ones when detecting patterns.

#### Why are RBMs important?

It can automatically extract **meaningful** features from a given input.

#### What's the RBM's structure?

It only possesses two layers; A visible input layer, and a hidden layer where the features are learned.

## The Restricted Boltzmann Machine model

* The **Restricted Boltzmann Machine** model has two layers of neurons, one of which is what we call a visible input layer and the other is called a hidden layer. The hidden layer is used to learn features from the information fed through the input layer. For our model, the input is going to contain X neurons, where X is the amount of movies in our dataset. Each of these neurons will possess a normalized rating value varying from 0 to 1, where 0 meaning that a user has not watched that movie and the closer the value is to 1, the more the user likes the movie that neuron's representing. These normalized values, of course, will be extracted and normalized from the ratings dataset.
* After passing in the input, we train the RBM on it and have the hidden layer learn its features. These features are what we use to reconstruct the input, which in our case, will predict the ratings for movies that user hasn't watched, which is exactly what we can use to recommend movies!
* We will now begin to format our dataset to follow the model's expected input.
* **Restricted Boltzmann Machine (RBM):**

RBMs are shallow neural nets that learn to reconstruct data by themselves in an unsupervised fashion.

#### Why are RBMs important?

It can automatically extract **meaningful** features from a given input.

#### How does it work?

RBM is a 2 layer neural network. Simply, RBM takes the inputs and translates those into a set of binary values that represents them in the hidden layer. Then, these numbers can be translated back to reconstruct the inputs. Through several forward and backward passes, the RBM will be trained, and a trained RBM can reveal which features are the most important ones when detecting patterns.

#### What are the applications of RBM?

RBM is useful for [Collaborative Filtering](http://www.cs.utoronto.ca/~hinton/absps/netflixICML.pdf), dimensionality reduction, classification, regression, feature learning, topic modeling and even **Deep Belief Networks**.

## Recurrent Neural Networks

* Recurrent Neural Networks are Deep Learning models with simple structures and a feedback mechanism built-in, or in different words, the output of a layer is added to the next input and fed back to the same layer.

The Recurrent Neural Network is a specialized type of Neural Network that solves the issue of **maintaining context for Sequential data** -- such as Weather data, Stocks, Genes, etc. At each iterative step, the processing unit takes in an input and the current state of the network, and produces an output and a new state that is **re-fed into the network**.

## Long Short-Term Memory Model

The Long Short-Term Memory, as it was called, was an abstraction of how computer memory works. It is "bundled" with whatever processing unit is implemented in the Recurrent Network, although outside of its flow, and is responsible for keeping, reading, and outputting information for the model. The way it works is simple: you have a linear unit, which is the information cell itself, surrounded by three logistic gates responsible for maintaining the data. One gate is for inputting data into the information cell, one is for outputting data from the input cell, and the last one is to keep or forget data depending on the needs of the network.

Thanks to that, it not only solves the problem of keeping states, because the network can choose to forget data whenever information is not needed, it also solves the gradient problems, since the Logistic Gates have a very nice derivative.

### Long Short-Term Memory Architecture

The Long Short-Term Memory is composed of a linear unit surrounded by three logistic gates. The name for these gates vary from place to place, but the most usual names for them are:

* the "Input" or "Write" Gate, which handles the writing of data into the information cell
* the "Output" or "Read" Gate, which handles the sending of data back onto the Recurrent Network
* the "Keep" or "Forget" Gate, which handles the maintaining and modification of the data stored in the information cell

## What exactly is Language Modelling?

* Language Modelling, to put it simply, **is the task of assigning probabilities to sequences of words**. This means that, given a context of one or a sequence of words in the language the model was trained on, the model should provide the next most probable words or sequence of words that follows from the given sequence of words the sentence. Language Modelling is one of the most important tasks in Natural Language Processing.

## What is Deep Learning?

**Brief Theory:** Deep learning (also known as deep structured learning, hierarchical learning or deep machine learning) is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers, with complex structures or otherwise, composed of multiple non-linear transformations.

### What is MNIST?

According to LeCun's website, the MNIST is a: "database of handwritten digits that has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image".

### Softmax Regression

Softmax is an activation function that is normally used in classification problems. It generate the probabilities for the output. For example, our model will not be 100% sure that one digit is the number nine, instead, the answer will be a distribution of probabilities where, if the model is right, the nine number will have a larger probability than the other other digits.

### Cost function

It is a function that is used to minimize the difference between the right answers (labels) and estimated outputs by our Network.

### Type of optimization: Gradient Descent

This is the part where you configure the optimizer for your Neural Network. There are several optimizers available, in our case we will use Gradient Descent because it is a well established optimizer.

Batch Gradient Descent is not often used because is too computationally expensive. The good part about this method is that you have the true gradient, but with the expensive computing task of using the whole dataset in one time. Due to this problem, Neural Networks usually use minibatch to train.

## Feature Learning

Feature engineering is the process of extracting useful patterns from input data that will help the prediction model to understand better the real nature of the problem. A good feature learning will present patterns in a way that significantly increase the accuracy and performance of the applied machine learning algorithms in a way that would otherwise be impossible or too expensive by just machine learning itself.

Feature learning algorithms finds the common patterns that are important to distinguish between the wanted classes and extract them automatically. After this process, they are ready to be used in a classification or regression problem.

The great advantage of CNNs is that they are uncommonly good at finding features in images that grow after each level, resulting in high-level features in the end. The final layers (can be one or more) use all these generated features for classification or regression.

Basically, Convolutional Neural Networks is your best friend to **automatically do Feature Engineering** (Feature Learning) without wasting too much time creating your own codes and with no prior need of expertise in the field of Feature Engineering.

* Activation functions are a cornerstone of Machine Learning. In general, Activation Functions define how a processing unit will treat its input -- usually passing this input through it and generating an output through its result. To begin the process of having a more intuitive understanding, let's go through some of the most commonly used functions.

## The Step Functions

The Step function was the first one designed for Machine Learning algorithms. It consists of a simple threshold function that varies the Y value from 0 to 1. This function has been historically utilized for classification problems, like Logistic Regression with two classes. The Step Function simply functions as a limiter. Every input that goes through this function will be applied to gets either assigned a value of 0 or 1. As such, it is easy to see how it can be handy in classification problems.

There are other variations of the Step Function such as the Rectangle Step and others, but those are seldom used.

## The Sigmoid Functions

Sigmoid functions are very useful in the sense that they "squash" their given inputs into a bounded interval. This is exceptionally handy when combining these functions with others such as the Step function.

Most of the Sigmoid functions you should find in applications will be the Logistic, Arctangent, and Hyperbolic Tangent functions.

#### Tanh

The Hyperbolic Tangent, or TanH as it's usually called, is defined as 𝑓(𝑥)=21+𝑒−2𝑥−1f(x)=21+e−2x−1. It produces a sigmoid over the (−1,1)(−1,1) interval. TanH is widely used in a wide range of applications, and is probably the most used function of the Sigmoid family.

## The Linear Unit functions

Linear Units are the next step in activation functions. They take concepts from both Step and Sigmoid functions and behave within the best of the two types of functions. Linear Units in general tend to be variation of what is called the Rectified Linear Unit, or ReLU for short.

The ReLU is a simple function which operates within the [0,∞)[0,∞) interval. For the entirety of the negative value domain, it returns a value of 0, while on the positive value domain, it returns 𝑥x for any 𝑓(𝑥)f(x).s

## What is different between Linear and Logistic Regression?

While Linear Regression is suited for estimating continuous values (e.g. estimating house price), it is n0t the best tool for predicting the class in which an observed data point belongs. In order to provide estimate for classification, we need some sort of guidance on what would be the **most probable class** for that data point. For this, we use **Logistic Regression**.

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, *y*, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

* **Linear Regression**

Defining a linear regression in simple terms, is the approximation of a linear model used to describe the relationship between two or more variables. In a simple linear regression there are two variables, the dependent variable, which can be seen as the "state" or "final goal" that we study and try to predict, and the independent variables, also known as explanatory variables, which can be seen as the "causes" of the "states".

When more than one independent variable is present the process is called multiple linear regression.  
When multiple dependent variables are predicted the process is known as multivariate linear regression.

The equation of a simple linear model is

𝑌=𝑎𝑋+𝑏Y=aX+b

Where Y is the dependent variable and X is the independent variable, and **a** and **b** being the parameters we adjust. **a** is known as "slope" or "gradient" and **b** is the "intercept". You can interpret this equation as Y being a function of X, or Y being dependent on X.

If you plot the model, you will see it is a line, and by adjusting the "slope" parameter you will change the angle between the line and the independent variable axis, and the "intercept parameter" will affect where it crosses the dependent variable's axis.

## How does TensorFlow work?

TensorFlow defines computations as Graphs, and these are made with operations (also know as “ops”). So, when we work with TensorFlow, it is the same as defining a series of operations in a Graph.

To execute these operations as computations, we must launch the Graph into a Session. The session translates and passes the operations represented into the graphs to the device you want to execute them on, be it a GPU or CPU. In fact, TensorFlow's capability to execute the code on different devices such as CPUs and GPUs is a consequence of it's specific structure.

## Why Tensors?

The Tensor structure helps us by giving the freedom to shape the dataset in the way we want.

And it is particularly helpful when dealing with images, due to the nature of how information in images are encoded,