DA

MCQ's

- https://letsfindcourse.com/data-science/data-analysis-mcq-questions-and-answers

Unit 1

What is Big data?

- Big data is nothign but huge amount of data or collection of large datasets which cannot be processed by using tarditional computing techniques.
- It is very complex to store, manage and access these data using regular tools.

Some sources of big data:

- Banking
- Transport
- Search Engine
- Social Media
- Stock

What is data science?

- Data science is a field of big data
- Data science involves extracting the meaningful information (in structured or unstructured form) from the huge complex info.

Types of data:

- Structured: Data can be stored into relational DBMS in the form of tables
- Semi structured: This data is neither structured nor unstructured. This type of data can be represented using XML and JSON.
- ustructured: This type data do not have any fixed data model. Ex images, videos, text, system logs, etc.

Issues regarding data in traditional file:

- Volume: Huge amount of data is generated
- Variety: Data is generated from various sources
- Velocity: High speed data generation
- Variablility: We cannot predict the flow of data such as size of data because of emerging trends.
- Complexity: It is very difficult to link such a huge data.

Applications of Big data:

- Fraud detection
- Log analysis
- Social media analysis

Data explosion:

- Data expsion is nothign but repid growth of data.
- One reason can be innovation

Data volume and Data velocity

Data volume: Data volume in Big data is nothing but the amount of data being generated.
Data velocity: Flow of data

Big data infrastructure challenges:

- Storage
- Transporatation
- Processing
- Speed and throuhput

Business Intelligence vs Data Science

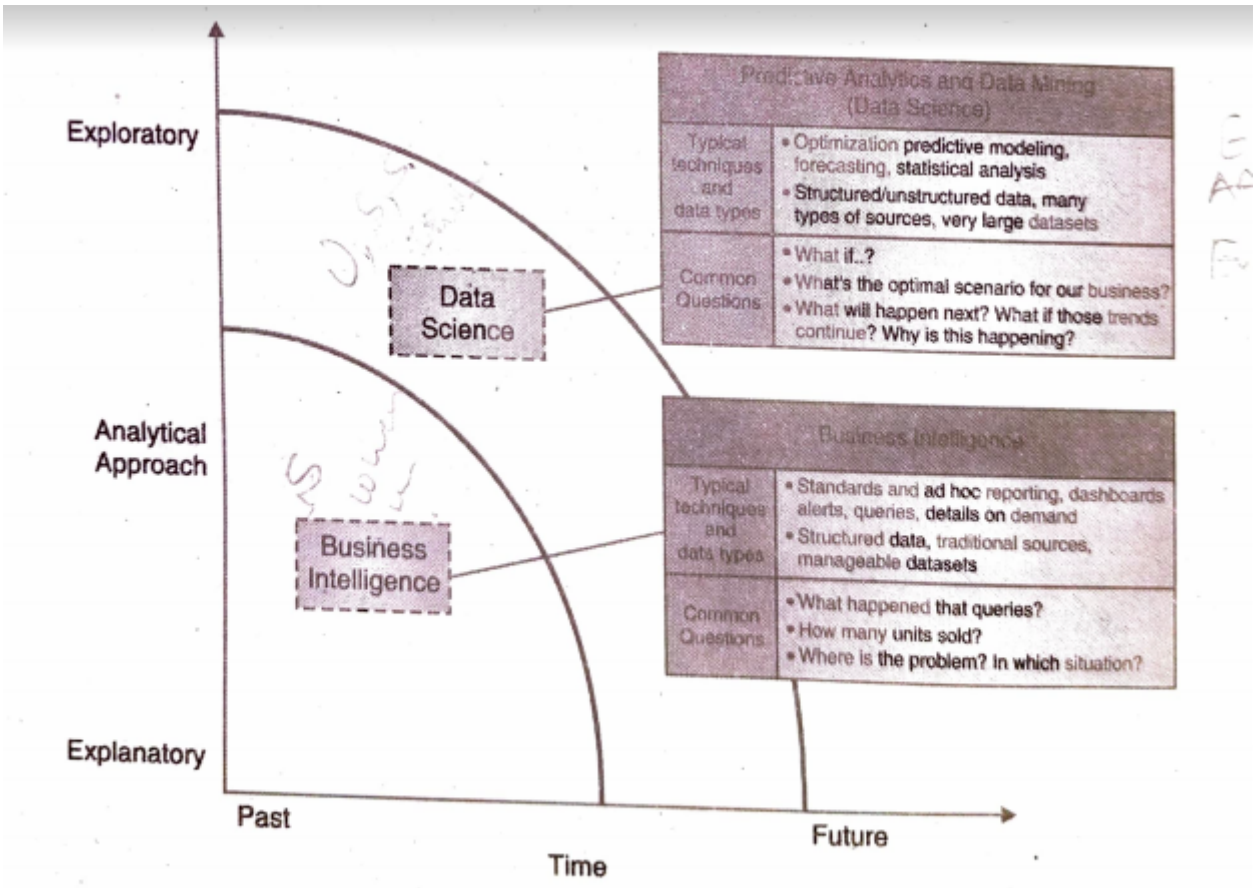


Fig. 1.2.1 : Comparing BI with Data Science

- Business intelligence is basically analyzing the past data and generating records, grpahs, etc. Like how many units of the products are sold, etc.
- BI usually provides information on business questions for the current period or in the past.

- Explanatory
- When and where events occured
- Data science analzes the past data for future use like future prediction of the product.
- This is exploratory
- how and why events occur

Current Analytical Architecture Problems:

- Solution is to ue analytic sanboxes to enable data scientists to carry out advanced analytics in a controlled way.

Four main elements in the interconnected web:

- Data devices
- Data collectors
- Data aggregators
- Data users/buyers

Key Roles for a Successful Analytics Project

- Business User
- Project Sponser
- Project Manager
- Business intelligence Analyst
- Database Adminstrator
- Data Engineer
- Data Scientist

Data Analytic Lifecycle:

- Data Discovery
- Data prep
- Model Planning
- Model Building
- Communicate results
- Operationlize

Important Points

- Data Analysis is defined by the statistician John Tukey in 1961 as "Procedures for analyzing data.
- The word 'Big data' was coined by Roger Mougalas
- big data word coined in year 1998

Unit 2

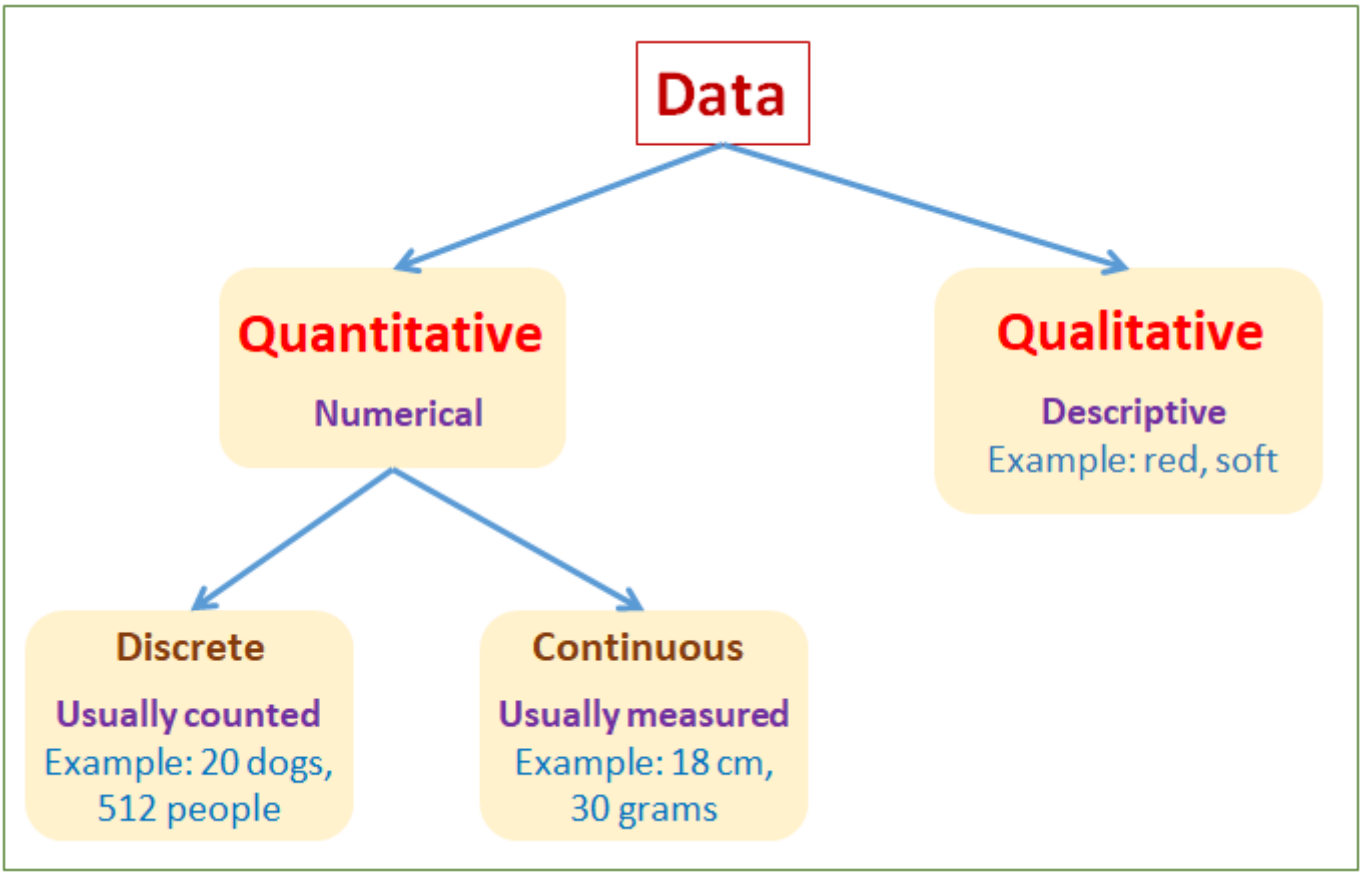
Population

- population is nothing but the entire group that you want to draw conclusins about.

Sample

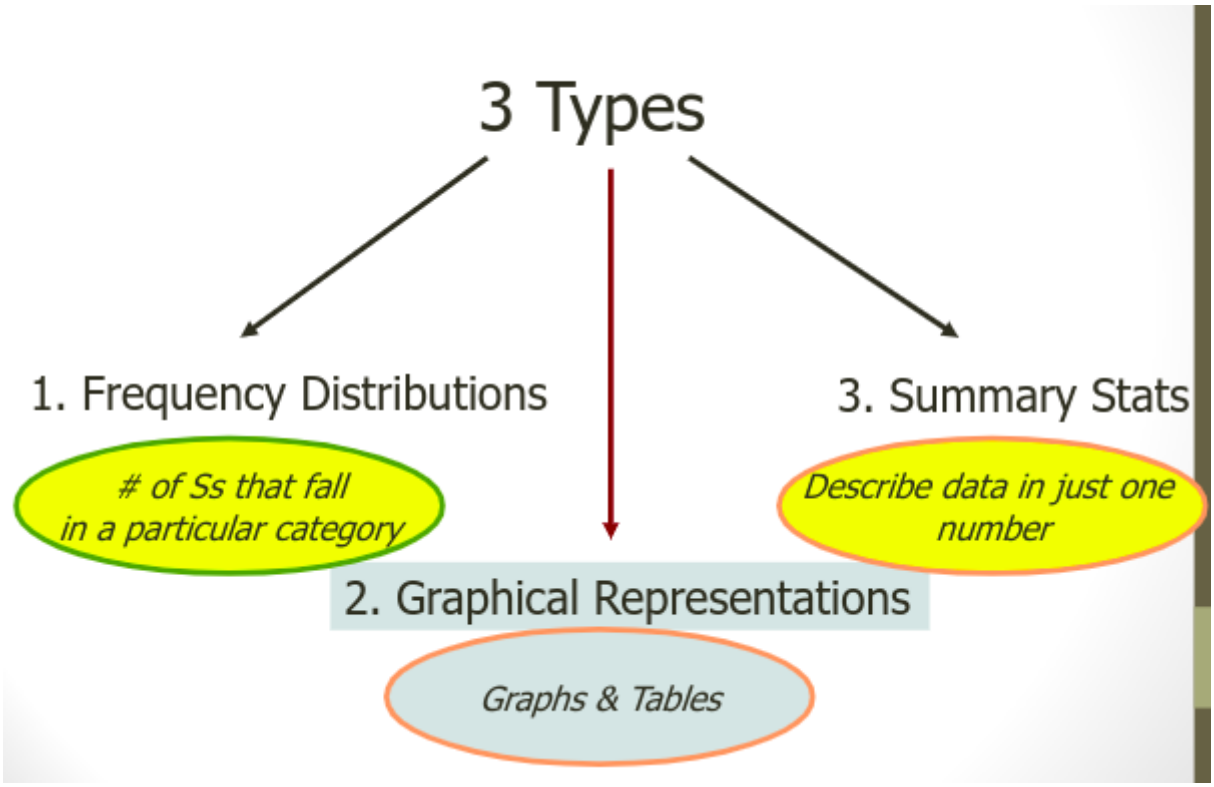
- sample is nothing but the subset of population i.e. the specific group that we will collect data from.
- it is a subset of population
- sameple size is always less then total populations ize
- process of selecting samples is called as sampling
- and the members which are selected in the sample group are sampling units
- the size of sample is nothing but the no of sampling units

Types of data



Types of statistics:

- Descriptive:
  - It describes the populationm, collect and analyze the data. Generate charts, graphs, repots from the data.
  - includes calculation of various mesaures such as average, percentile, center tendency measures, variability measrures, etc.
  - measures of central tendency: mean mode and median
  - measrues of dispersion: range, sd, quatile deviatiom, variance, etc.
  - Threee types of descriptive statistics:

    - frequency distribution
    - graphcal representation
    - summary stats
    - 

- Inferential
    - Inferential statistics analze the sample data to draw conclusions about the populations.
    - Inferential stattistics is used to draw the conclusion from tthe data generated form discriptive statictics.
    - includes methods like point estimation, interval estimation, hypothesis testing.

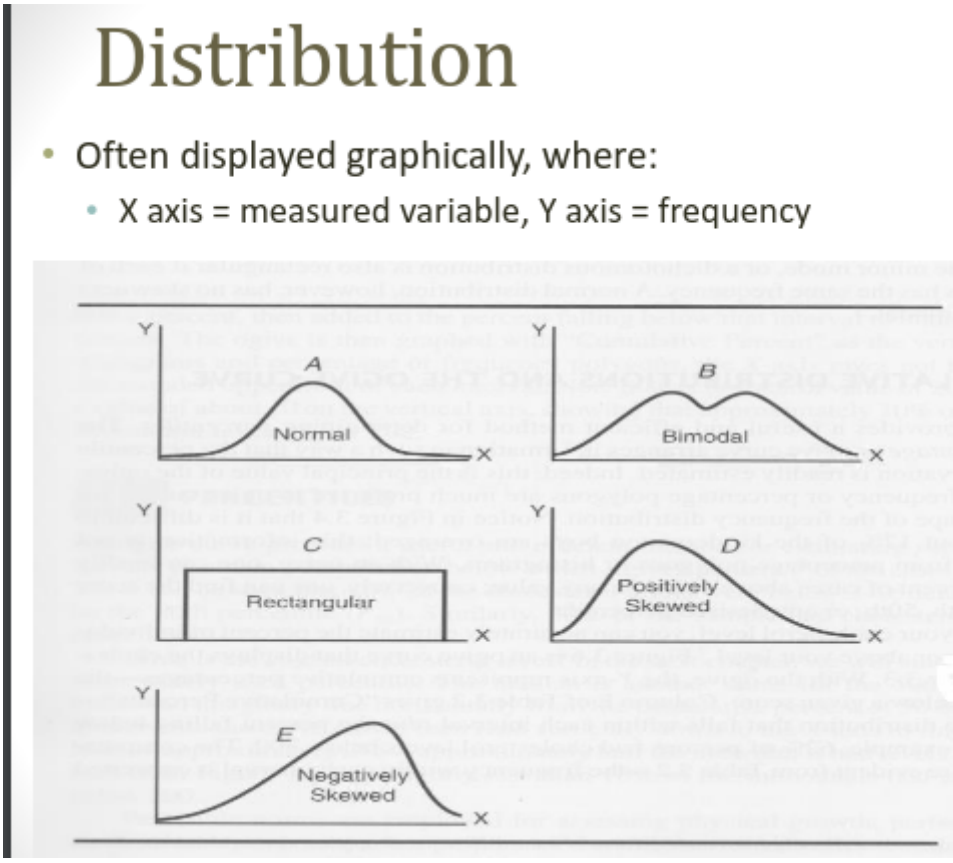
Comparison Chart

Basis for Comparison	Descriptive Statistics	Inferential Statistics
Meaning	Descriptive Statistics is that branch of statistics which is concerned with describing the population under study.	Inferential Statistics is a type of statistics, that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
What it does?	Organize, analyze and present data in a meaningful way.	Compares, test and predicts data.
Form of final Result	Charts, Graphs and Tables	Probability
Usage	To describe a situation.	To explain the chances of occurrence of an event.
Function	It explains the data, which is already known, to summarize sample.	It attempts to reach the conclusion to learn about the population, that extends beyond the data available.

Measures of central tendency:

- Mean: average
- Mode: represents score with the highestfrquency of occurances.
- Median: Median is the middle score for a set of data that has been arranged in order of magnitude. In case of even elements take the avg of two values. The median is less affected by outliers and skewed data.

Distributions



When to use which measure of central tendency:

- If data is categorial use mode because mean and median is not possible.
- If data is quantitative then use mean or median.
- If the data is highly skewed then use median otherwise go for mean.

Measures of dispersion or variability:

- This tells us how the ovservations are sperad out on each side of the center.
- If there is no dispersion then all the obersvations are identical
- If the there is small dispersion means there is high uniformity in the observations.
- Absolute measures of dispersion

- **Absolute measures of dispersion**
  - Range [Max-min]
  - Percentile(Interquartile) range [Q3-Q1]: measures variability based on percentiles
  - Quartile deviation: estimate the spread of a distribution about mean[Q3-Q1]/2
  - Mean deviation: indicates how far, on average, all values are from the middle(mean)
  - Variance and standard deviation

- Relative measures of dispersion

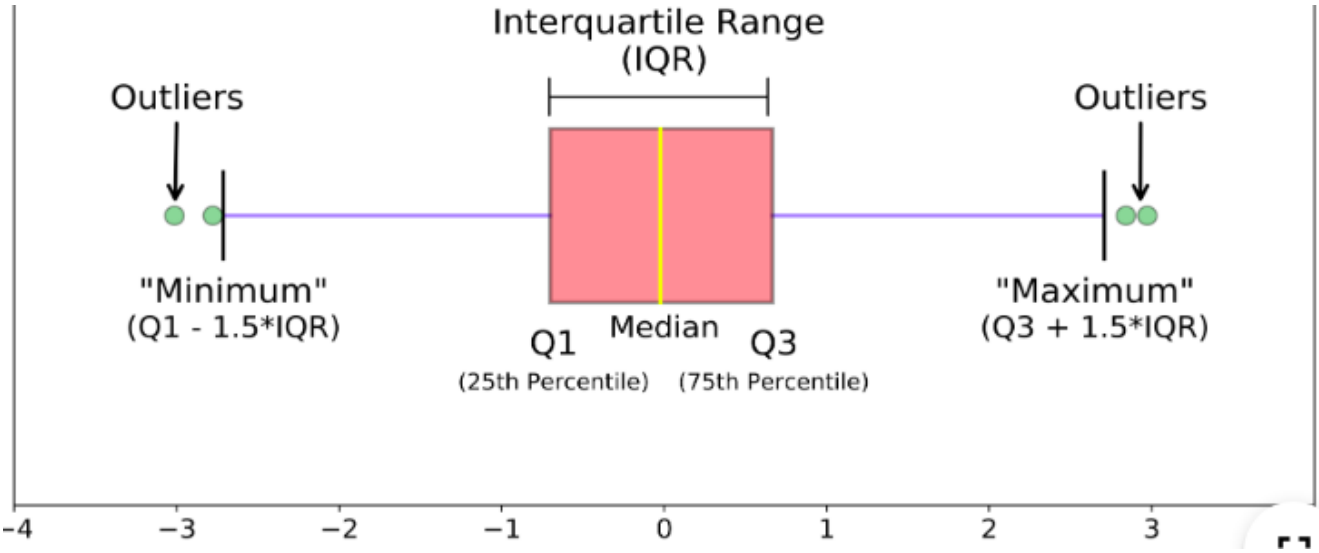
  - **Relative measure of dispersion(Ratio/percentage):**
  - Used to compare two or more dataset and for unit free comparison
    - The coefficients of dispersion (C.D.) based on different measures of dispersion Based on Range = (X max – X min) / (X max + X min).
    - C.D. based on quartile deviation = $(Q_3 - Q1) / (Q_3 + Q1)$.
    - Based on mean deviation = Mean deviation/average
    - For Standard deviation = S.D. / Mean

- Standard Deviation:
  - SD measures the spread of data around mean.
  - High SD indicates widely spread from the mean
  - Low SD indicates very close to mean

Box Plot:

- Box plot gives the five number summary: min, first quatile(Q1), median, third quatile(Q3), max
- Box plot tells about outliers and their values

- **Median (Q2/50th Percentile): the middle value of the dataset.**
- **First quartile (Q1/25th Percentile): the middle number between the smallest number (not the "minimum") and the median of the dataset.**
- **Third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the maximum) of the dataset.**
- **Interquartile range (IQR): 25th to the 75th percentile.**
- **Whiskers (shown in blue)**
- **outliers (shown as green circles)**
- **Maximum: Q3 + 1.5*IQR**
- **Minimum: Q1 -1.5*IQR**



Inferential Statistics:

- point estimation
  - The difference between the point estimate and the true value of the population parameter is called estimation error or sampling error.

Hypothesis Testing

- Also called significance testing

- **Hypothesis originates from the Greek word *hupo* (under) and *thesis* (placing).**

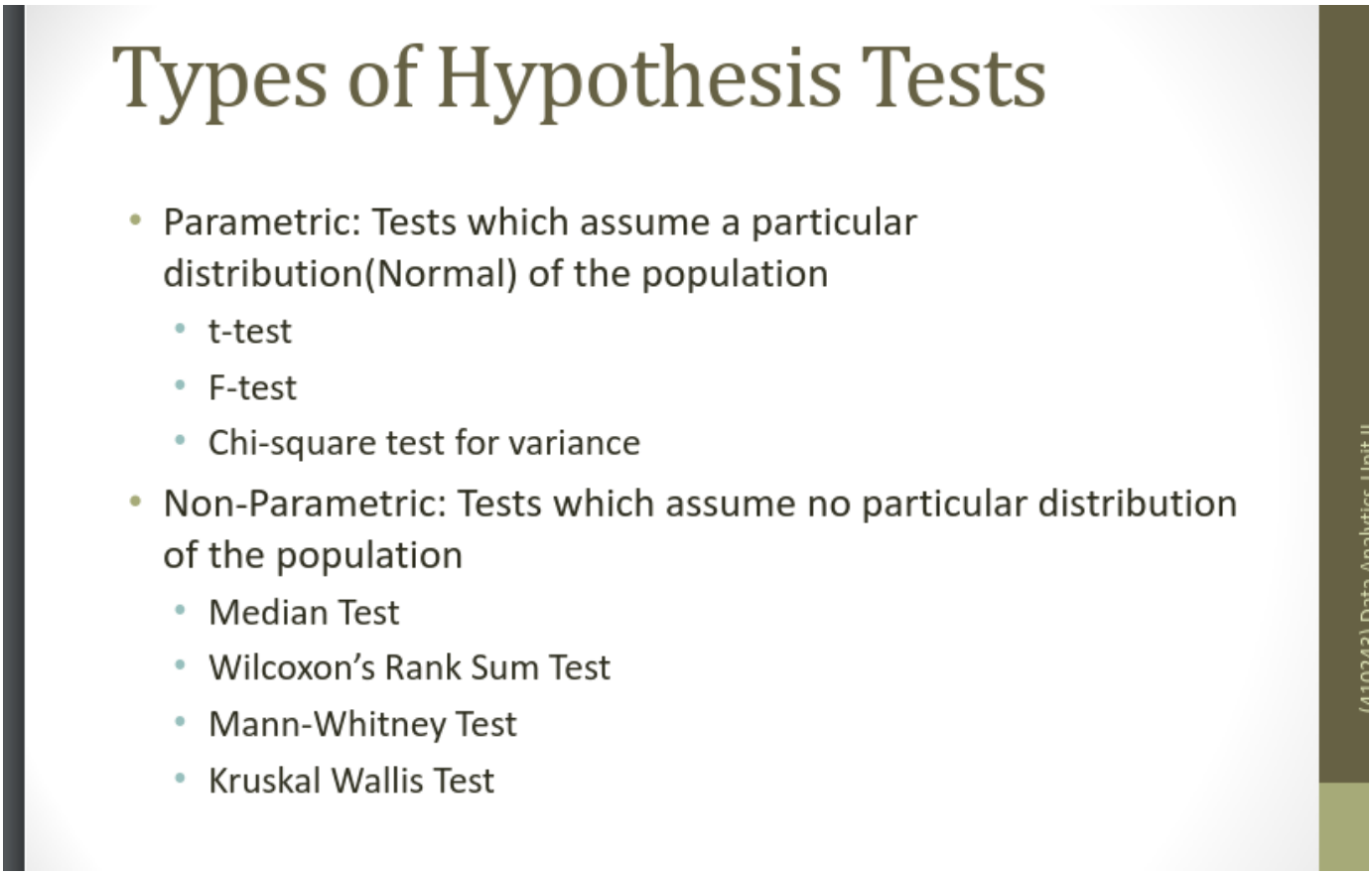
- It is a method for testing an assumption about a paramters in a population with the help of data measured in sample
- It is nothing but forming the claim and then check it with the data.
- Null hypothesis:

    - null hypothesis when there is no differnece between two samples.
    - well accepted / by deafault
    - Ho representatin symbol
- Alternative Hypothesis
    - also called as reshearch hypothesis
    - alternatice when there is a difference
    - oppostite of Ho
    - HA representatin symbol
    - Consider an example of recognizing the teaching effect of Teacher T1 and Teacher T2 on students.

      The corresponding null hypothesis and alternative hypothesis will be as follow:

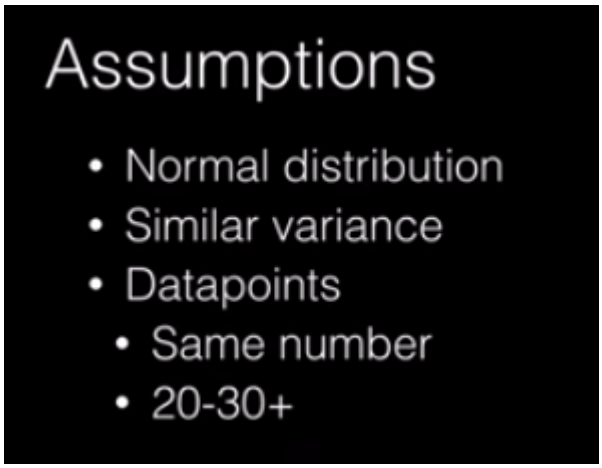
      - **H0 :** Teaching of Teacher T1 and Teacher T2 have the same effect on the students.
      - **HA :** Teaching of Teacher T1 has a better effect than Teacher T2 on the students.
- for performing hypothesis test it is required to define the null and alternative hypothesis because the hypothesis test nothing but accpeting or rejecting the null or alternative hypothesis.

Types of Hypothesis tests



Student's t-test

- Differnce of means is calculated by using student't t-test or welch t-test.

- If scaling terms of both population are know nad when they are following normal distrucbution then the t-test can be performed.
- We cannot performed t-test if the scaling terms is not known.
- degree of freedom = no of samples - 1
- In this test we calculate the value of t by using the given two populations and then in the degree of freedom chart we check the acceptance value if our value < that value then we simply accept otherwise we reject.
- $$T = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ $$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$
- The shape of t-distribution is same as normal distribution
- Infact when the degree of freedom rachees the value 30 or more then the t-distribution is identical to normal distribution.
- In student t'test main supposition is variance of both the population is same if that is not same then we have to use welch's t-test.
- 

welch's t-test

$$T_{welch} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Where, X_i = sample mean

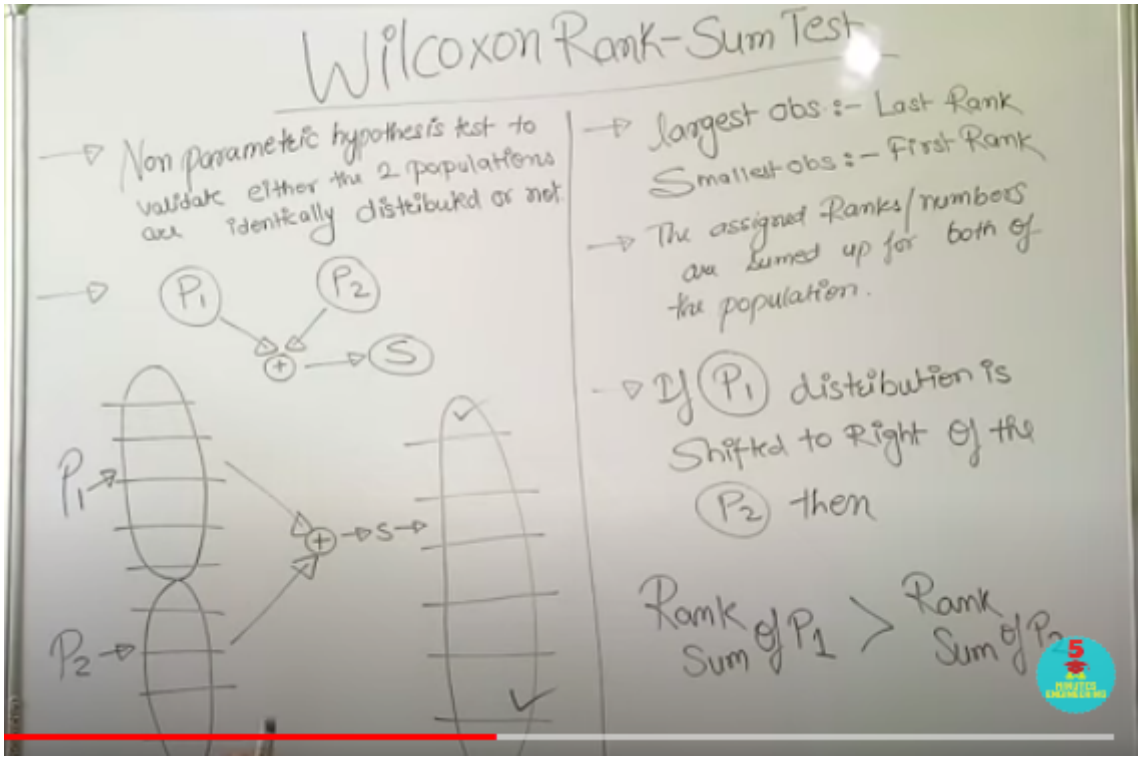
S_i^2 = sample variance

And n_i = sample size.

Wilcoxon Rank-Sum Test

- This test is non-parametric test i.e. it does not requre any type of distribution in the populations.
- This test is used to validate whether the two populations are identical or not.

- This test first group the both the populations observations into one cluster then we assigned rank to the obersevations.
- Larger observations get last rank and smaller observation get first rank and so on.
- In case if two observations are similar then in this case the average of the rank is assigned to observations.
- After that we separet the observations ranks into their respective populations and calculate the sum of ranks.
- If pop1 sum of ranks is > sum of pop2 then the pop1 is shifted to right.



Type I and Type II errors

- Hypothesis test may result in two type of errors depending on whether the test accepts or rejects the null hypothesis.
- If null hypothesis is true but it is rejected then Type I error. Denoted by alpha.
- If null hypothesis is false but it is accepted then Type II error. Denoted by beta.

TABLE 3-6 *Type I and Type II Error*

	H_0 is true	H_0 is false
H_0 is accepted	Correct outcome	*Type II Error*
H_0 is rejected	*Type I error*	Correct outcome

Power and Sample Size

- Power of test is nothing but the possibility of rejecting the null hypothesis.
- Denoted by 1-B where B (beta) is type II error

ANNOVA

- ANNOVA can be used to validate the disimilarity between multiple populations.
- In t-test we can only use two population but here we can use multiple populations for the test.
- ANNOVA stands for Analysis of variance
- This technique is simply extension of hypothesis testing.
- Here null hypothesis is defined as all the population mean are equal
- Here alternative hypothesis is considered as at least one pair of the population is not equal
- The between groups mean sum of squares (Sn^2) is an approximation of the between groups variance.
- The withing groups mean sum of squares (Sw^2) is an approximation of the withing groups variance.
- If the value of Sn^2 > Sw^2 then it means that some of the population means are different from each other.
- Here f-statistic value is calulcated by Sn^2 / Sw^2

Clustering

- Clustering is nothing but grouping of things together
- unsupervised technique
- unsupervised technique means problem of discovering unknown structure inside the unlabelled data
- clsuering cannot be used for prediction or forecasting.
- common clustring technique is KMeans
- clustering is frequently used as an operned to classification.

KMeans algorithm:

- KMeans is an analytical algorithm
- Unsupervised
- A cluster refers to a collection of data points aggregated together because of certain similarities.
- The K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
- The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.
- It halts creating and optimizing clusters when either:
    - The centroids have stabilized — there is no change in their values because the clustering has been successful.
    - The defined number of iterations has been achieved.

How to determine the best value for K?

- For this we can use WSS (Withing sum of squares) metric

- $$WSS = \sum_{i=1}^{M} d(p_I, q^{(i)})^2 = \sum_{i=1}^{M} \sum_{j=1}^{n} (p_{ij} - q_j^{(i)})^2$$
- In simple words WSS is nothing but sum of squared of distance between all the points and their nearest clsuter.

T-test, CHI Square Test, ANNOVA Test

When to use what?

- If we have only one categorial attribute use hypothesis test
- If we have two categorial attributes together then use CHI square test
- If we have categorial and one continuous numerical value then use T test
- If we have two continous numerical value then use ANNOVA Test.

Clustering Techniques:

- Flat
  - KMeans
- Hierarchical: Dendogram graph is used
  - Agglomerative (Bottom Up): Matthamaticaaly easy to solve (Denodogram is used)
  - disisive (Top Down)

Unit 3

Association Rules

- Association rules in unsupervised technique
- It is not predective, it is descriptive method
- Association rules is used to find the relationshipe between items in the large dataset
- Relationship is in the form of rules or frequent itemset
- Association rule mining is a technique to identify underlying relations between different items.

Apirori Algorithm

- This is the most fundamental association rule mining algorithm
- Iterative algorithm and level wise
- Apriori Property
  - All subsets of a frequent itemset must be frequent(Apriori propertry).
  - If an itemset is infrequent, all its supersets will be infrequent.

- Support
  - $$support(I) = \frac{Number\ of\ transactions\ containing\ I}{Total\ number\ of\ transactions}$$
- Confidence
  - Confidence measures the trustiness of a discoverd rule.
  - $$confidence(X \rightarrow Y) = \frac{Number\ of\ transactions\ containing\ X\ and\ Y}{Number\ of\ transactions\ containing\ X}$$
- Lift
  - Lift measures how much more often x and y occur together than expected
  - $$Lift = \frac{Support}{Support(X) * Support(Y)}$$
- Leverage
  - measures the diff in the prob of x and y appearing together
  - $$Leverage(X \rightarrow Y) = Support(X \wedge Y) - Support(X) * Support(Y)$$
- Frequent Item Set
  - An itemset whose support is greater than or equal to a minSup threshold

Association rule mining is a two step process:

- Finding the frequent itemset
  - Appriory
  - FP-growth
- Generating the association Rules

Apriori Algorithm:

Apriori employs an iterative approach known as a *level-wise* search, where *k*-itemsets are used to explore (*k*+1)-itemsets.

Apriori property: All nonempty subsets of a frequent itemset must also be frequent.

CSE GURUS @

A two-step process is followed:

1. The join step: To find L_k, a set of candidate *k*-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k.
2. The prune step: C_k is a superset of L_{k-1}, that is, its members may or may not be frequent, but all of the frequent *k*- itemsets are included in C_k. To reduce the size of C_k the Apriori property is used.

Apriori Algorithm: Example

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Min.support count=2

CSE GURUS @ M

C_2 L_2 C_3 L_3 ...

2-freq

I1, I2
I1, I3
I1, I4
I1, I5
I2

Scan *D* for count of each candidate →

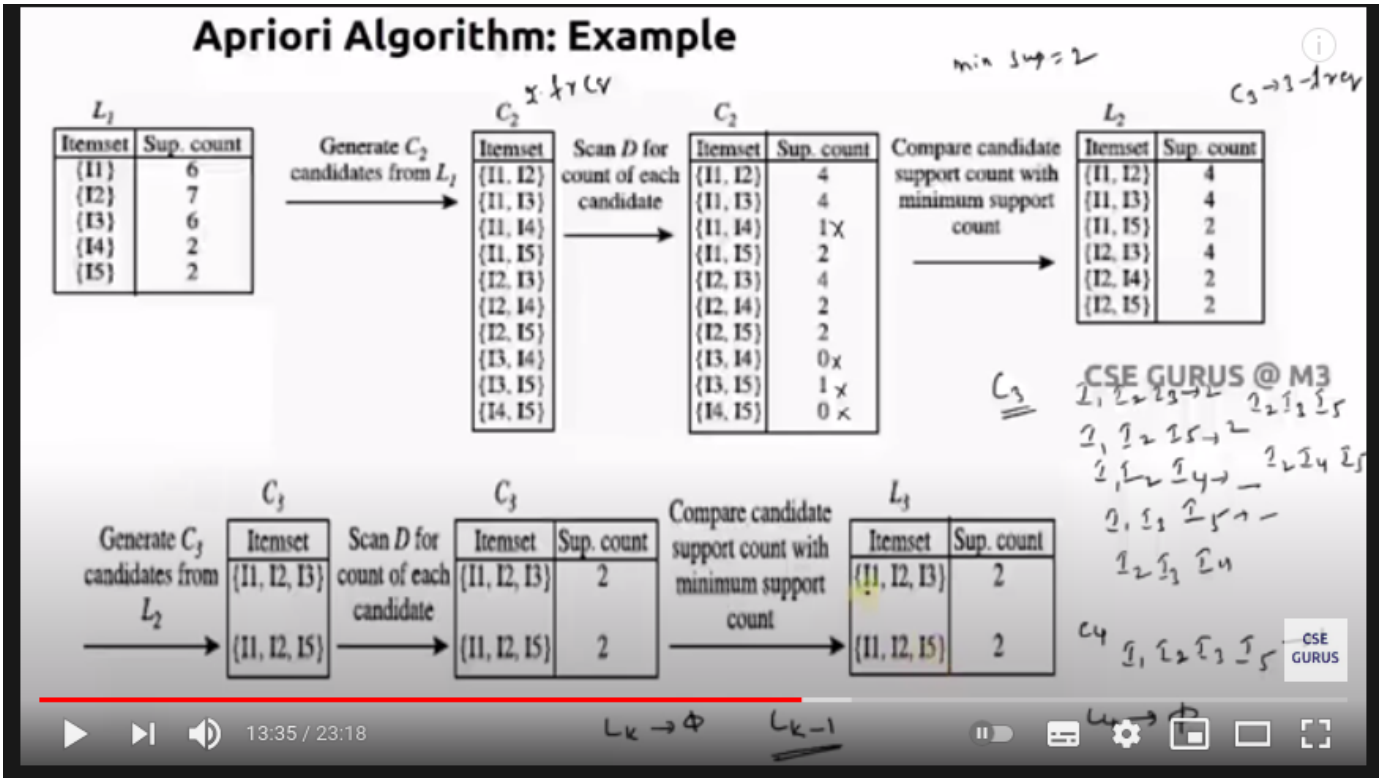
C_1 1-freq

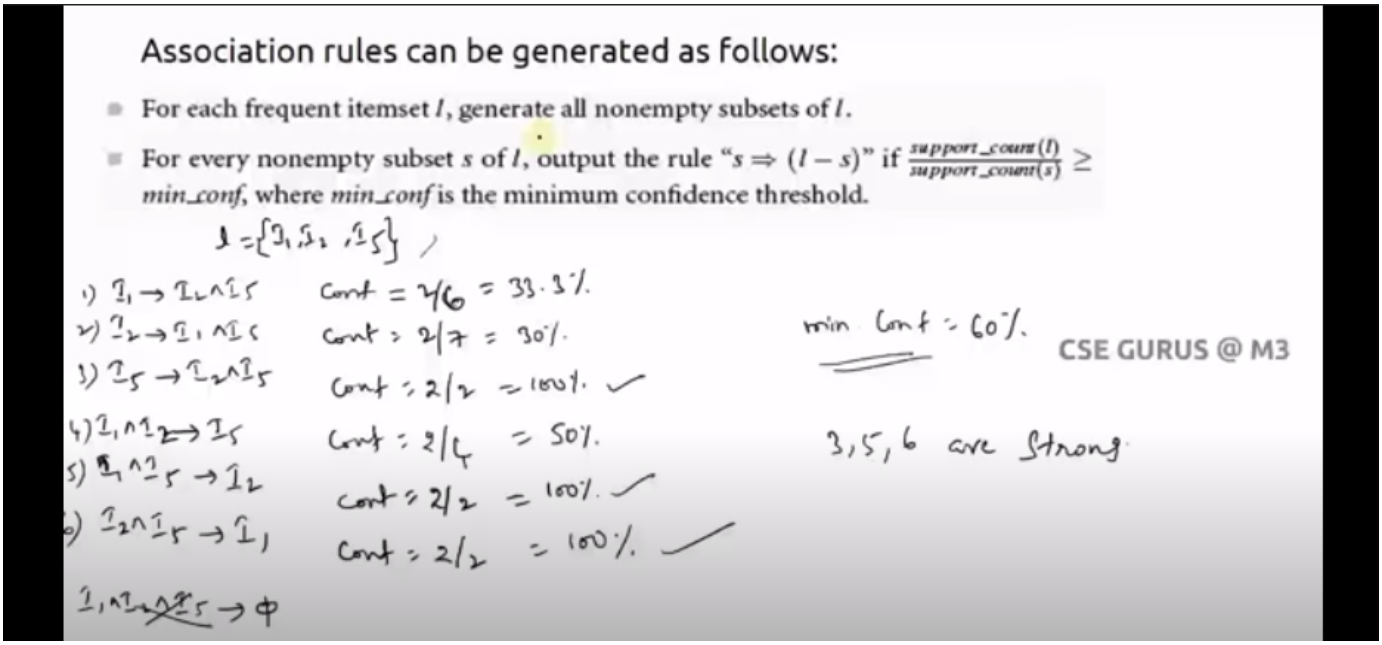
Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Compare candidate support count with minimum support count →

L_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2





For Detailed Apriori algorithm
https://www.youtube.com/watch?v=UP4ezNZfcH0&ab_channel=CSEGURUS

Problem with apriori:

- For large datasets the apriori algorihm can ve omputationally expensive

Techniques to improves the efficiency of apriori:

- Sampling
- Hashbased item set counting
- dynamic item set counting
- Transaction reduction
- Paritioning

Regression Analysis

- Let's say we have to predict the annual sales then the sales is the dependent variable and the attributes on which the sales prices is calculated are independent variables.
- Regression analysis is used to model the relationship between a dependent variable and one or more indepdenent variable.

Outliers: outliers are nothing but the very extream values in the observarions.

Multicolinerity:

- If the indepedent variables are highly correlated to each other then the variables are said to be multicolinear.
- Many types of regression techniques assume that multicolinearity should not be present in the dataset.

Overfitting:

- When we add uncessary explanatory variables it might lead to overfitting
- In overfitting the model works well on our training set but does not works well for testing data.
- This is called as high variance problem.

Underfitting:

- When our model gives poor results even for training dataset then it is called as underfitting.
- This is also called as high bias problem.

Types of Regression

- Linear Regression
- Logistic Regression

Linear Regression

- Linear regression is used to model relationship between independent variabels and continuous depedent variable.
- If we have only 1 depdent and 1 indepdent varible then it is simple linear regression.
- If we have more then 1 indepdent varibles then it is called as multiple linear regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \varepsilon$$

where:

y is the outcome variable

x_j are the input variables, for $j = 1, 2, \ldots, p-1$

β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j, for $j = 1, 2, \ldots, p-1$

ε is a random error term that represents the difference in the linear model and a particular observed value for y

- Here in the above image Bi are the regeression coefficients and the E is the random error term.
- regression coefficinets estimation can be done by OLS (Ordinary Least Squares)

Model with normally distributed errors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \varepsilon$$

where:

y is the outcome variable

x_j are the input variables, for $j = 1, 2, \ldots, p-1$

β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j, for $j = 1, 2, \ldots, p-1$

$\varepsilon \sim N(0, \sigma^2)$ and the εs are independent of each other

Assumption of linear regression:

- Absence of multicolinearity
- No heteroscedasticity
- There must be a linear relation between dependent and independent variables.
- No outliers
- Error term shoud be normally distributed with mean 0 and constant variance

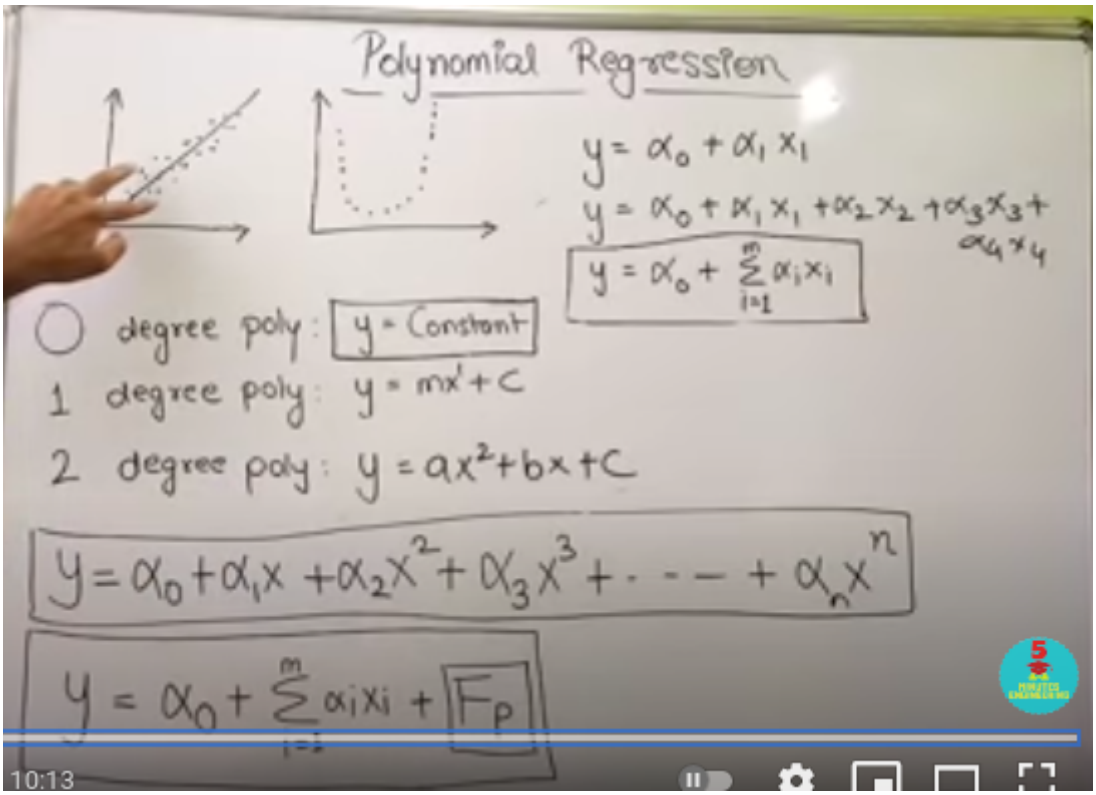
Logistic Regression:

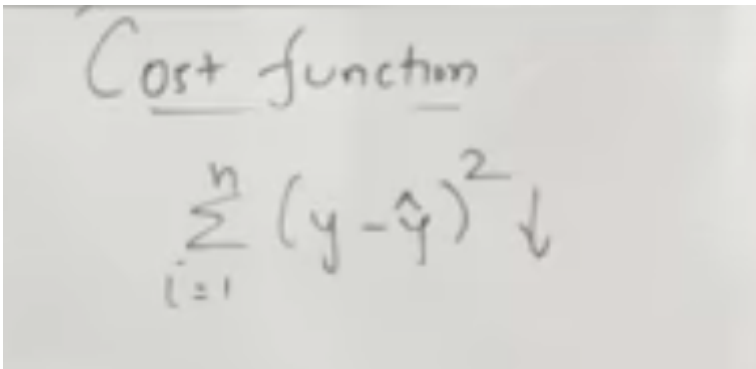
- In logistic regression dependent variable is binary (2 categories)
- Indepdent variabels can be binary or continous
- In multinomial logistic regression there are more than two categories in dependent variable.

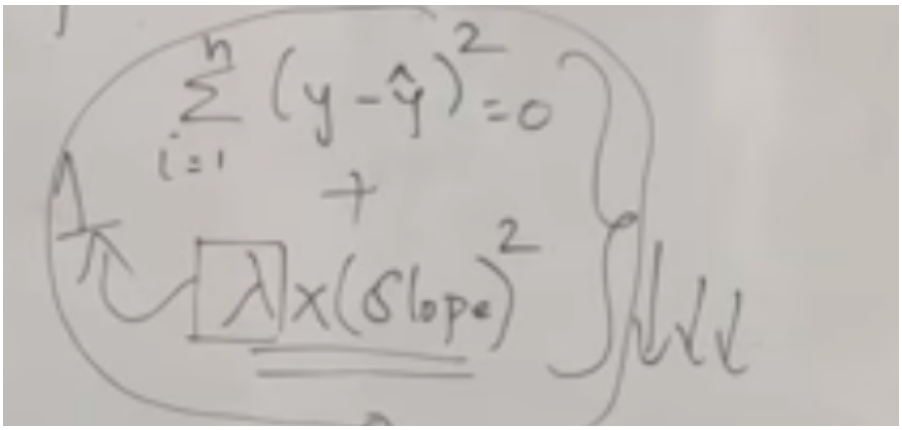
Why we don't use linear regression in some case?

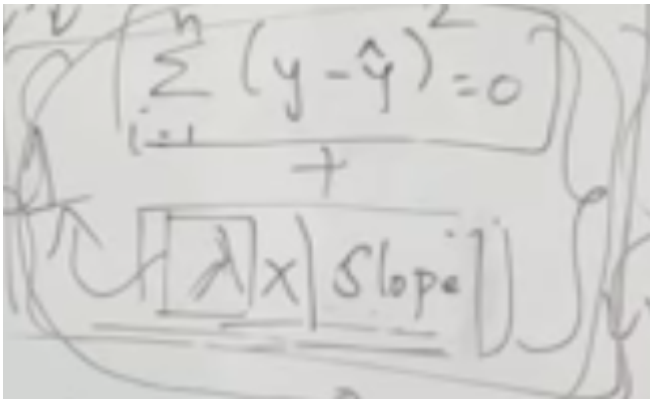
- The homoscedascity assumption is violated
- The errors are not normally distributed
- y follows binomial distribution and hence is not normal.

Other Regression Models:

- Polynomial Regression:
  - can deal with non-linear separable data
  - curve line instead of straight line
  - 

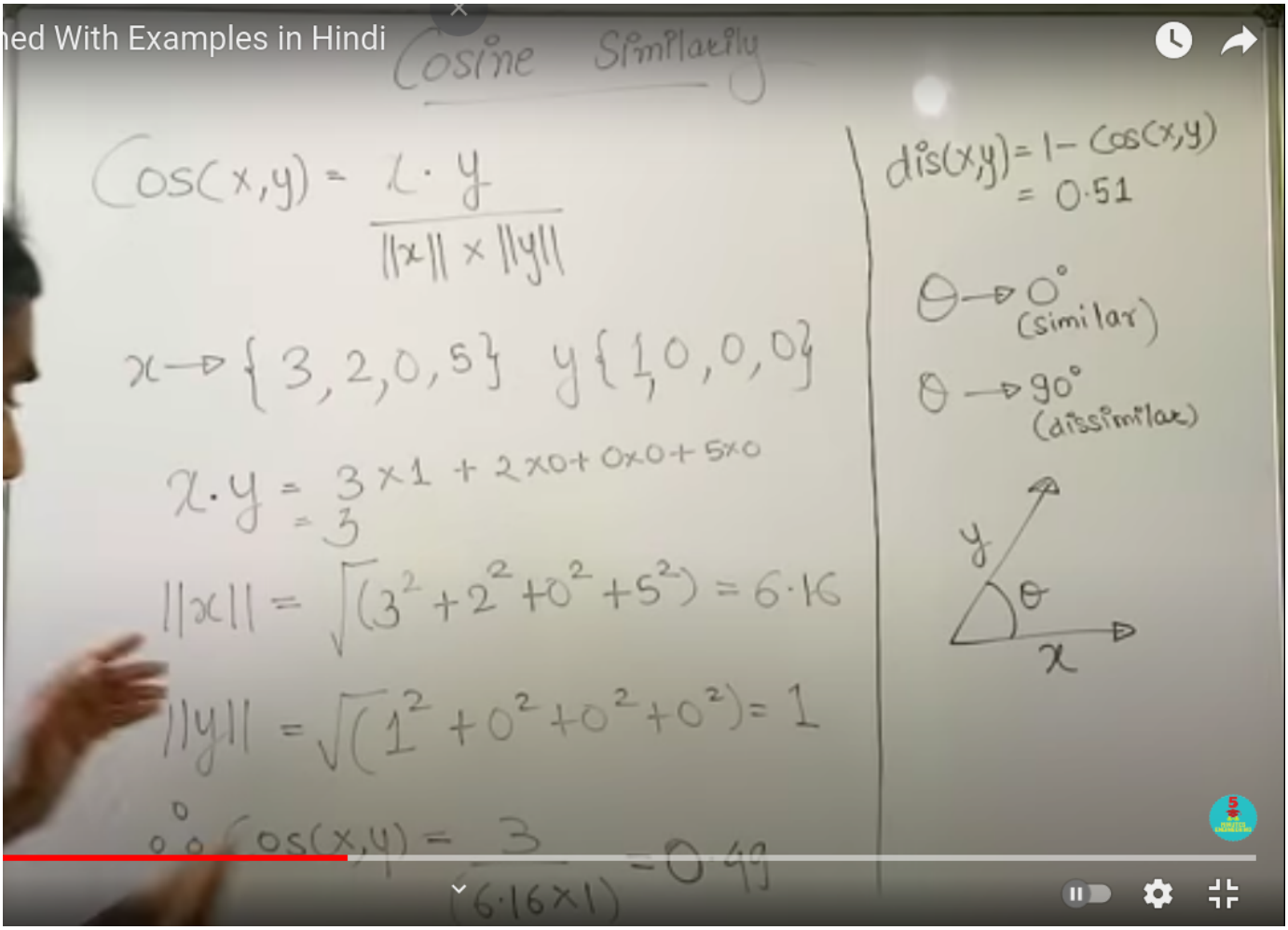
- Ridge Regression
  - when the input variables are highly corerelated this will fail linear or polynomial regression and causes overfitting problem.
  - In this case we have to use ridge or lasso regression.
  - Ridge regression adds extra term in the original cost function and this helps to reduce the overfitting of the model.
  - Linear regression cost function
  - 

- Ridge regression cost function
  - 

  - Value of lambda is from 0 to any positive value
  - Cross validation is used to decide the value of lambda
- Lasso Regression
  - lasso regression is same as ridge regression but instead of square slope we use absolute value of slope.
  - 

  - Lasso regression also helps in feature selection. https://www.youtube.com/watch?v=9lRv01HDU0s&ab_channel=KrishNaik
- ElasticNet Regression
  - Hybrid version of ridge and lasso.

Cosine Similarity Measure between two sentences



https://medium.com/@sumn2u/cosine-similarity-between-two-sentences-8f6630b0ebb7#:~:text=Cosine similarity between two sentences can be found as a,represent sentences%2Fparagraphs as vectors.&text=Here are two very short,more than Linda loves me

Unit 4

Decision Trees

- Also known as prediction trees
- In decision trees we have tree nodes and the branches
- X (x1,x2,...xn) are the input variables to the decision trees and it returns the output variable Y based on the input
- input variable can be categrial or continuous
- leaf nodes in the decision trees do not have any further branches and the leaf nodes basically gives the prediction output
- Two types of decision trees:
  - Classification
  - Regression
- classification trees are used when the output variable is categorial, mostly binary like yes or no, sale or not, etc

- Regression trees are used when the output variable is numeric or continuous
- decision rules are fairly straightforward like if-then rules
- The branching of the node is called as split

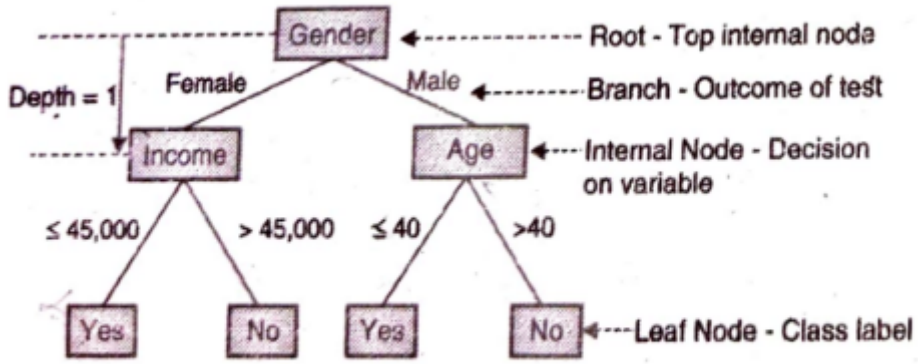


Fig. 4.1.1 : Example of Decision Tree

- At each split, out of all attributes the decision tree select the most informative attribute. The extent to informativeness of an attribute is decided by measures like entropy and information gain.
- The main idea of a decision tree is to identify the features which contain the most information regarding the target feature and then split the dataset along the values of these features such that the target feature values at the resulting nodes are as pure as possible.

Decision trees impurity measures for classification

- Gini Index
- entropy

Decision trees impurity measures for Regression

- varience

Decision trees purity measures

- Information gain

Impurity Criterion

Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

p_j: proportion of the samples that belongs to class c for a particular node

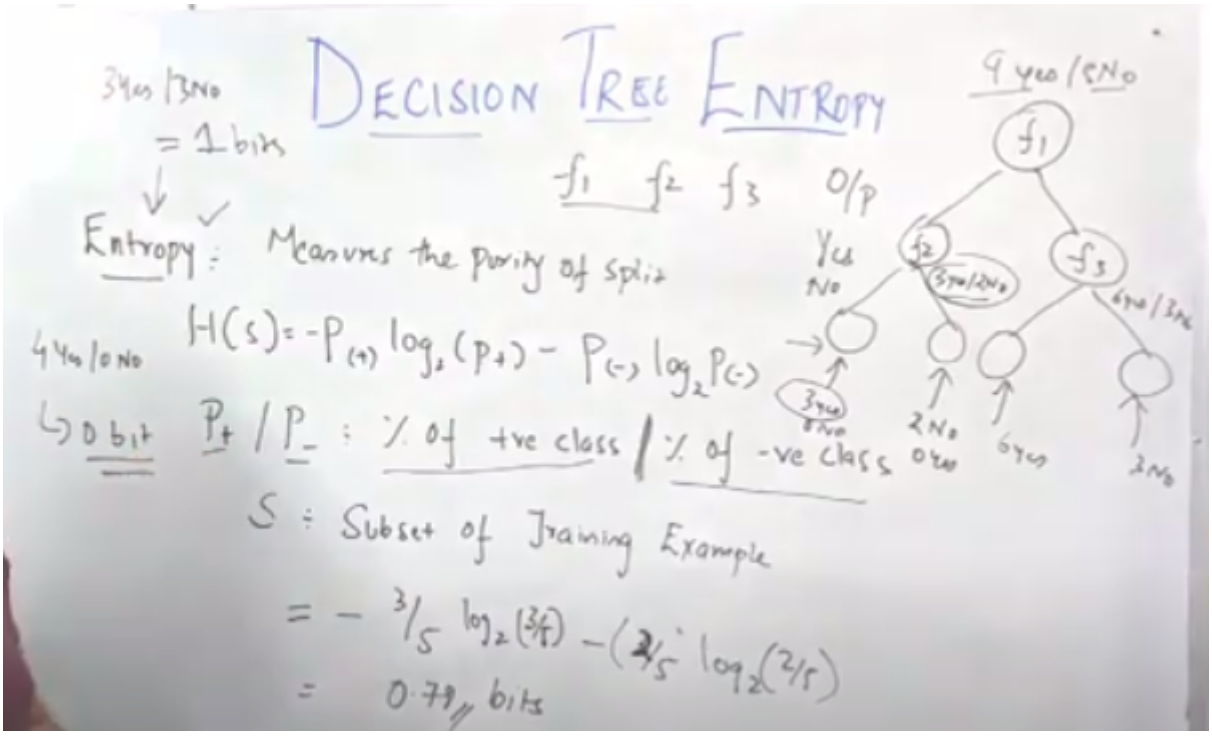
Entropy

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

p_j: proportion of the samples that belongs to class c for a particular node.

*This is the the definition of entropy for all non-empty classes ($p \neq 0$). The entropy is 0 if all samples at a node belong to the same class.

Entropy and Information Gain



$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

InformationGain(feature)= Entropy(Dataset)−Entropy(feature)

Example:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

- The feature with the largest information gain should be used as the root node to start building the decision tree.
- ID3 algorithm uses information gain for constructing the decision tree.
- For 2 categories classes the max value of entropy is 1. So refer to the first image let's we have 3 yes and 3 no then the entropy value is 1. So 1 means totally impure. We have to select the entopy with min value.
- Max entropy values is depends on the no of classes
-
  - 2 classes: Max entropy is 1
  - 4 Classes: Max entropy is 2
  - 8 Classes: Max entropy is 3
  - 16 classes: Max entropy is 4

Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

Example:

Past Trend	Open Interest	Trading Volume	Return
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up
Negative	Low	High	Down
Positive	Low	Low	Down
Negative	High	High	Down
Negative	Low	High	Down
Positive	Low	Low	Down
Positive	High	High	Up

Let's start by calculating the Gini Index for 'Past Trend'.
P(Past Trend=Positive): 6/10
P(Past Trend=Negative): 4/10
If (Past Trend = Positive & Return = Up), probability = 4/6
If (Past Trend = Positive & Return = Down), probability = 2/6

```
Gini index = 1 - ((4/6)^2 + (2/6)^2) = 0.45
```

- Attribute with the minimum gini index will be selected.

Decision Tree Algorithms:

- ID3
  - Iterative Dichotomiser 3
  - invented by Ross Quinlan
  - precursor to the c4.5 algorithm
  - uses the entropy and infromation gain to create the decision tree from the dataset.
- C4.5
  - It is extension to the ID3 algorithm
  - ID3 algorithm only works for categorial and C4.5 works for both categorial and continuous data.
  - Uses gain ration and split info for the decision tree building
- CART
  - CART (Classification and regression trees)
  - Uses Gini index

-
- ASSISTANT

Evaluating decision trees

- Greedy algorithms are used in decision trees in which they all time select the best option.
- Sometimes if bad split is taken then it affects the overall tree performance so that's why we use random forest concept to randomize the data or randomize the splitting to get the diff structure of the trees.
- After then we can evalute all the trees and then we select the most effective tree structure.
- Also we have to observe the depth as well as nodes of the tree because if the depth of the tree or the nodes in the tree are more with the less no of memebers then it is considerd as overfitting.
- 2 approches to avoid overfitting in decision trees:
  - prevent growth of the tree prior to its reaching the point
  - complete the full growth of the tree and then post-prune the tree with approaches like reduced-error pruning and rule-based post pruning.

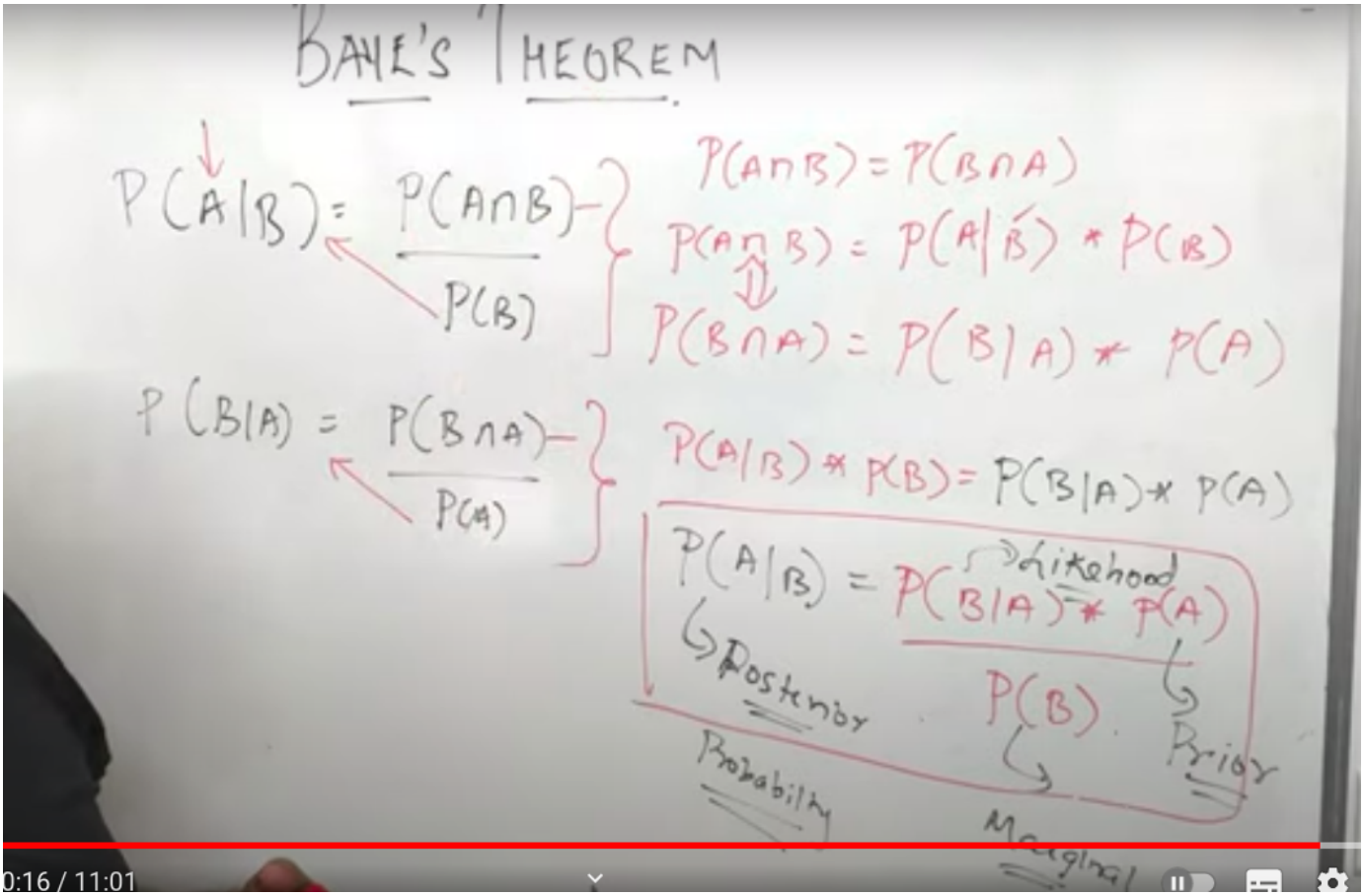
Application of Decision Trees

- decision tree can be used for classification as well as regression problems
- also let's we are usign logistic regression for classification and our dataset contains lots of feature variables so we can use the decision trees to select the most improtant attributes based on the values of information gain.
- prune redundant variables.

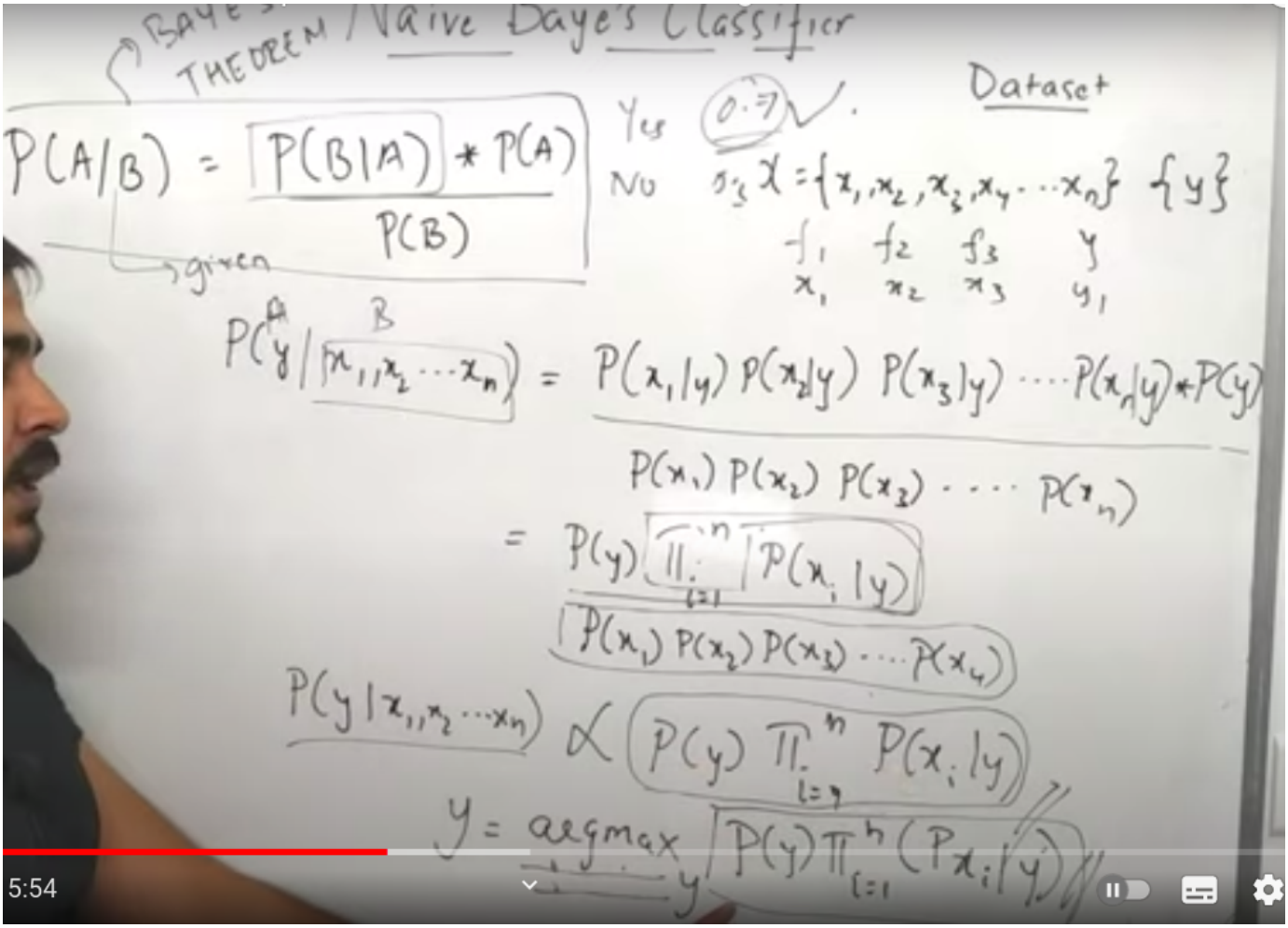
Naive Bayes

- Naive bayes is a probablistic classification method based on bayes theorem or law.
- A naive bayes classifier considers that the existance or nonexistance of a specific feature of a class is not related to the existance or nonexistance of other features.
- Niave bayes supports both categorial as well as continuous input variabls
- Also we can convert the continuous variables into the categorial ones by using the method called discretization of continuous variables.
- Applications:
  - Most popular use case of naive bayes is spam filtering (Email)
  - Fraud detection

Bayes Theorem



Naive Bayes Classifier



- Because of multiplying several smaller probabilityva values will give even more smaller result, close to zero. This problem is called as numerical underflow.
- It can be solved by using log function
- Instead of directly multiplying the prob we have to calculate the log of probablity and then we have mutliply it.
- $$log(a \cdot b) = log(a) + log(b)$$
- As we know multiplication of log ab is equal to loga + log b
- so instead of product we will add all the log of probabilities.
- $$P(c_1|A) \propto \log P(c_1) + \sum_{i=1}^{m} \log P(a_j|c_1) \quad i = 1, 2, \ldots n \qquad \ldots(4.2.8)$$

Smoothing

- If one of the attribute values is not present in one of the class labels in the training set, the seubsequent p(aj|ci) will equal to zero.
- And because of this zero all other values of multiplication result will become zero.
- In order to avoid this smoothing techniques are used.
- Laplace smoothing
- $$P^*(x) = \frac{\text{count}(x) + 1}{\Sigma x\,[\text{count}(x) + 1]}$$

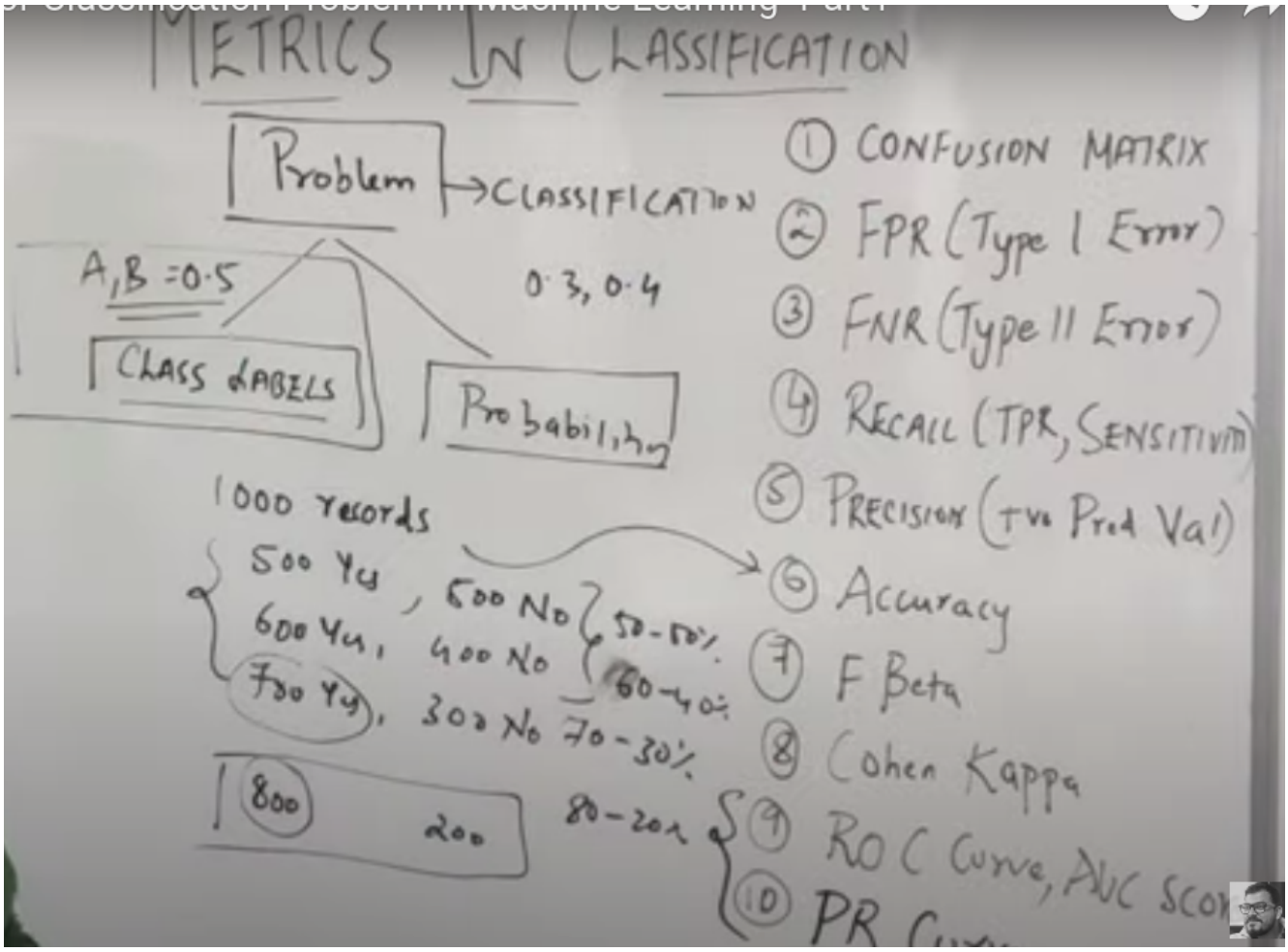
Naive Bayes Diagnostics

- Naive bays classifiers are able to handle missing values which is not possible to logistic regression.
- Naive bayes is very easy to understand and impleent. Also computationally also it is very efficient.
- Drawback:
  - The variables in the data are considered by the Naïve Bayes as conditionally independent. Hence, correlating variables becomes sensitive since the algorithm may double count the effects.

    E.g., consider that individuals having low income and low credit tend to default. If there is requirement of scoring "default" based on both income as well as credit as two different attributes, there may be double-counting effect in naïve Bayes on the default outcome, which will affect the accuracy of the prediction.
- In its simple form we can use the naive bayes only with categorial variables. If there are any continuous variables then they must be first converted into catgorial variables using the dicretization of variables process.

Diagnostics of classifiers

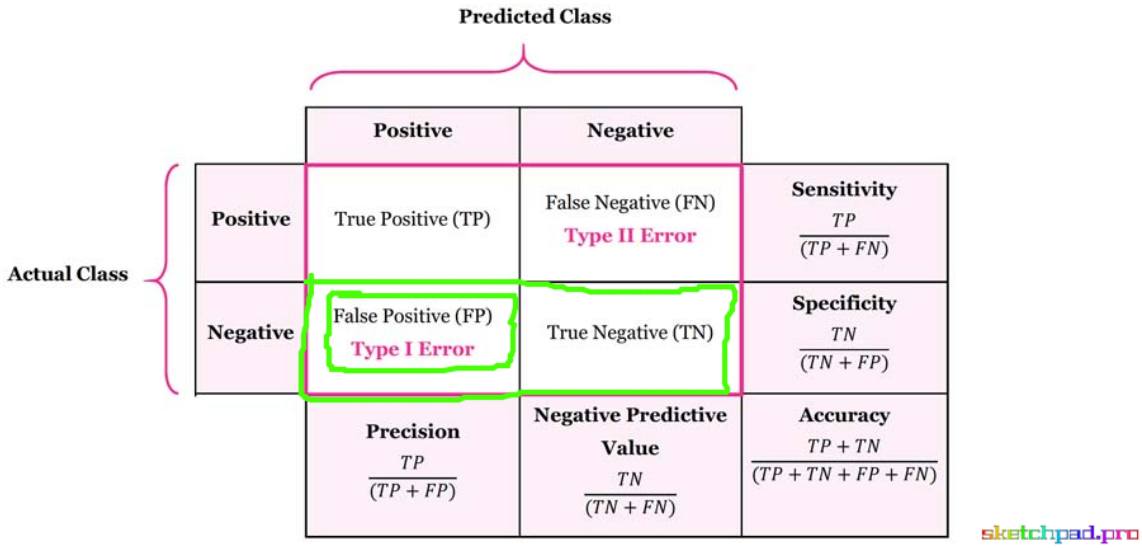
- For balance dataset we can use auuracy measure.
- If we have imbalance dataset then we should not have to use Accuracy measure.
- In case of imbalance dataset we can use recall, precision, FPR, or FNR.
- What is balance dataset? 50-50, 60-40, 70-30 are balanced datasets.



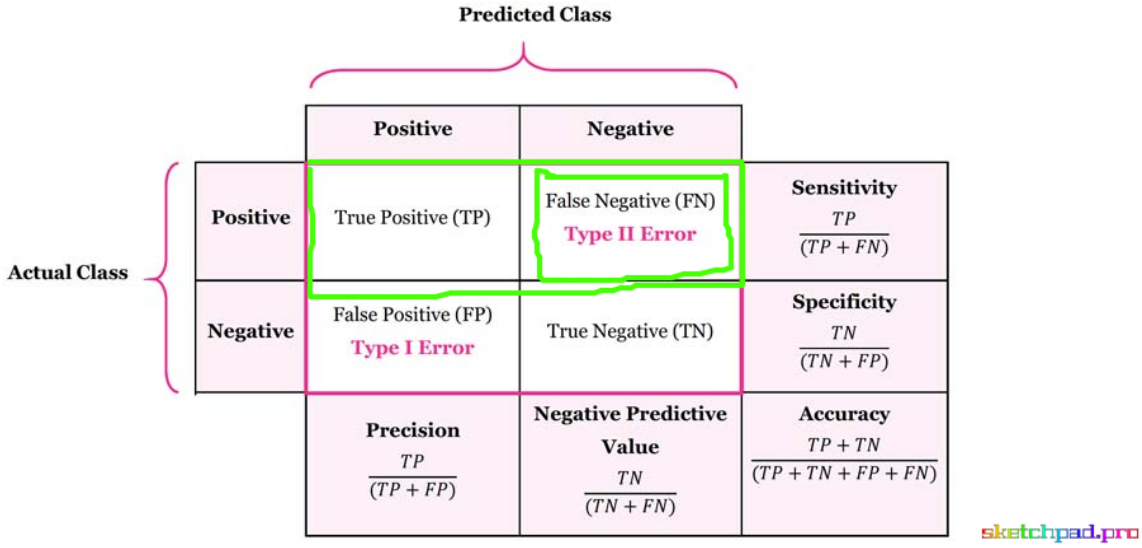
Predicted Class

		Positive	Negative	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

- Accuracy: TP+TN/(TP+TN+FP+FN)
- FPR (False Postive Rate) or TypeI error: FP/(FP+TN)

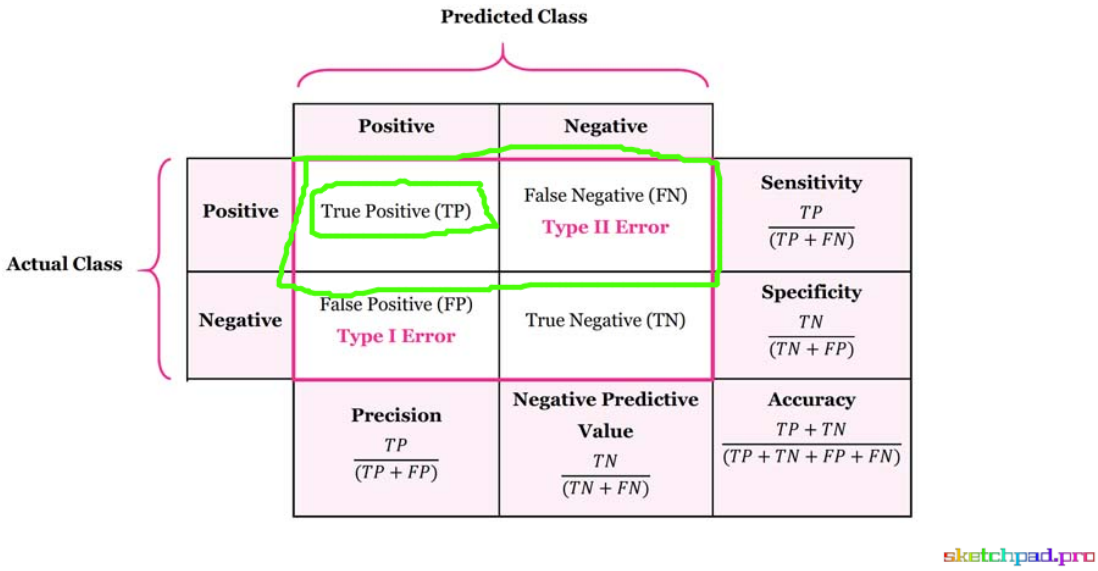


- FNR (False Negetive Rate) or TypeII error: FN/(FN+TP)

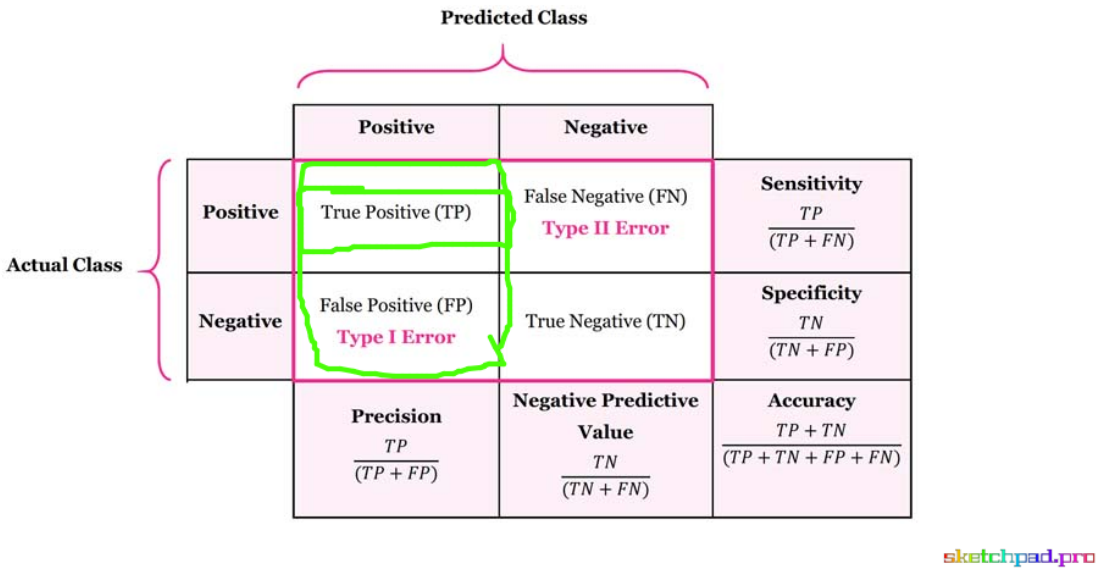


- Error Rate:
  - A percentage of errors made over the whole set of instances used for testing.
  - error rate = 1 - accuracy
  - OR error rate = (FP + FN) / (TP + TN + FP + FN)
- Recall:
  - Out of total positive values how many positives we have predicted correctly
  - It is the measure of completeness
  - TPR
  - Sensitivity

- TP/TP+FN



- Precision:
  - Out of totally actual predicted positive result how many are actually positives
  - It is the measure of exactness
  - Positive prediction value
  - TP/TP+FP

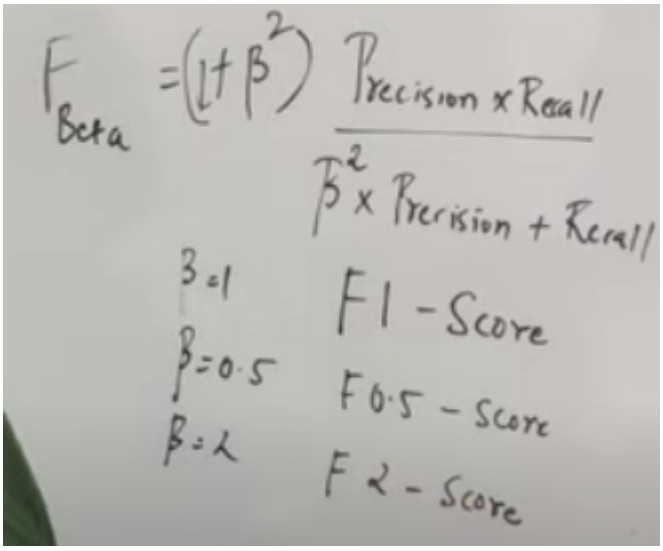


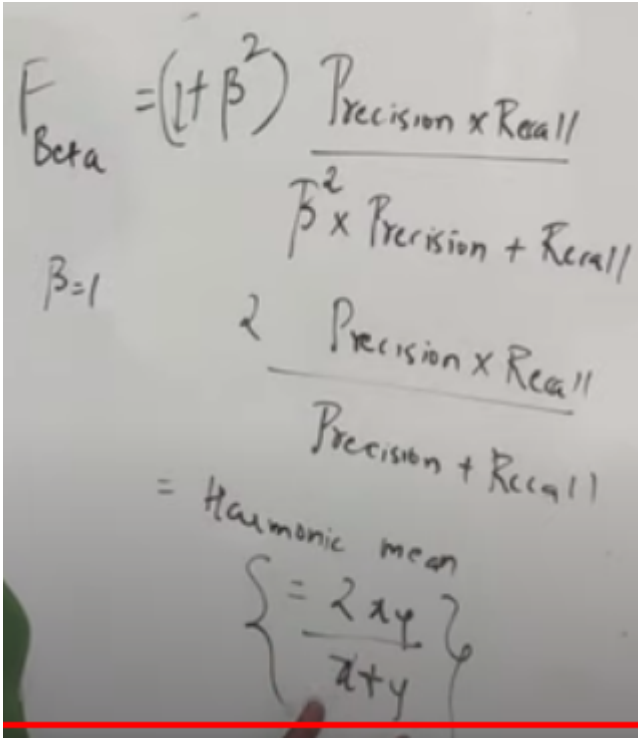
- Specificity:
  - out of total negaive values how many negatives we have predicted correctly.

- TN/TN+FP

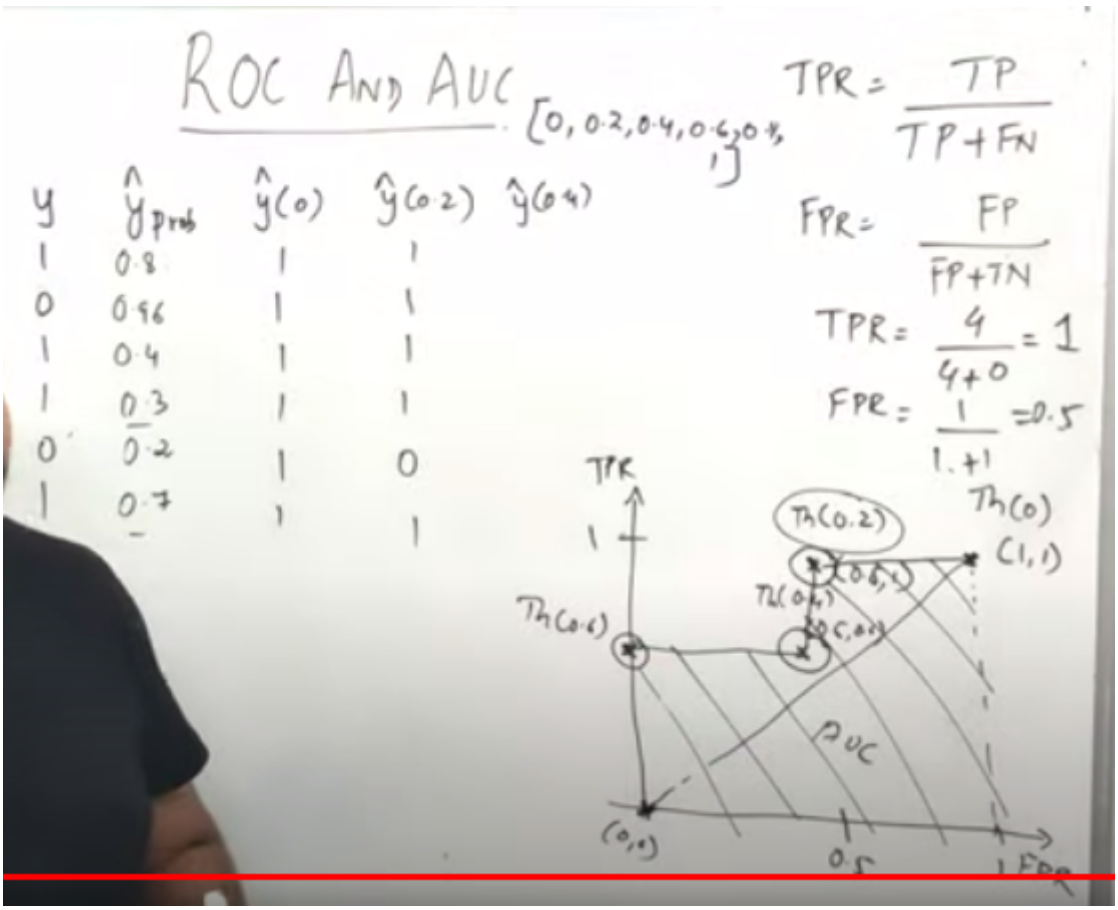
		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

sketchpad.pro

- F Beta
  - when both precision and recall are important then we use F beta
  - 

  - 

    - Select B=1 when FP and FN both are equally important
    - Select B=0.5 when FP has more importance then FN
    - Select B=2 when FN has more importance then FP

- ROC and AUC
    - 

    - ROC stands for (Receiver Operating Chracteristic)
    - AUC (Area under the curve)
    - for creating ROC curve we have to use TPR and FPR values for diff thresholds.

When to use precision and when to use recall?

- So consider an example of email spam detection so if email is not spam but is is detected as spam so FP i.e. email is not spam actually but it is dected as positive so here we try to reuce the FP.
- So whenever FP is more important then we have to use precision
- and when FN is more important then we have to use recall

Detailed Video: https://www.youtube.com/watch?v=aWAnNHXIKww

Other classification Methods

- There are some other classification techniques such as bagging, boosting, support vector meachines (SVM) and random forest.
- These are some ensemble machine learning techniques which use multiple models to get better predictive performance.

Unit 5

Data visualization

- Graphical or pictorial representation of the information is called as data visualization
- Main goal and objective of data visualization is to commnicate information clearly and effectively through graphical means.
- Data visualization is one of the steps in data analysis or data science.
- In the data visualization if the data is in the numerical form, then it is represented using bars, lines or dots.
- metadata i.e. data about the data can be more informative.
- visualization can be static as well as interactive.
- Data visulization is also considered as a branch of desciptive statistics.

Challenges to Big Data Visualization

5V's of Big data:

- Volume: amount of data being generated
- Velocity: data generated very fast
- Variety: diff infomraiton from diff sources
- Varacity: refers to complexity of the data due wo which quality and accurcay of data is reduced.
- Value: use the information for business by analyzing the big data.

Challenges:

- Scalability
- Dynamics
- Diversity and heterogenecity
- Perspective and Interactive Scalability

Some other problems:

- Visual Noise: A dataset contains the object which are very realtive to each other so users unable to separet them on the screen at the time of visulaization.
- Information loss: While performing data reduction there may be chances to information loss
- Large Image perception: The aspect ratio, device resolution and physical perception limits the data visulaization methods.
- High rate of image change
- High performance requriements

Conventional Data Visualization Tools

Some points based on which selectionof most effecive data visulization method is done:

- Size and volume of data
- Cardinality

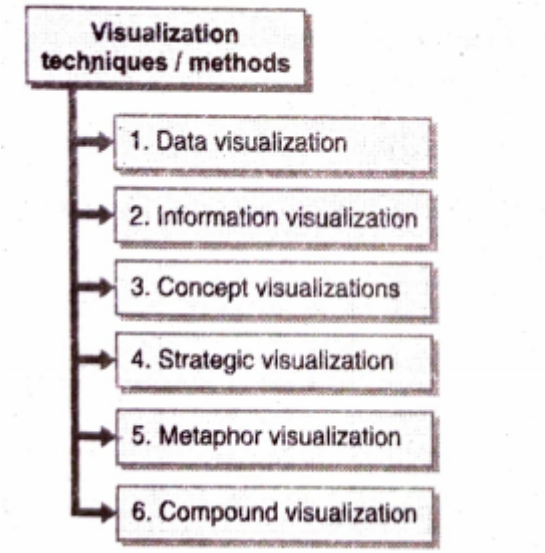
- Portion of dat ato be convey
- Audience
- Types of visual
- Data visulization can be static or interative

Interactive visualizations approaches:

- Zoom in and Zoom out or Zooming (ZUI): Zoomable or Zooming User Interface
- Zoom + PAN
- Overview + Detail
- Focus + Context or Fish Eye

Visualization Techniques or Methods

- Six main categories



[DICSMC]

Data visulization: Helps to visulize the quantitative data in schematic or diagramtic forms. Ex table, line chart, pie chart, hostogram, scatter plot, etc.

Information Visualization:

- Information visualization facilitates interactivity in the data to increase cognition or perception ability.
- Data is transformed into a changeable image (data map, tree map, clustering, semantic network, timeline and venn diagram, etc)

Concept Visulization:

- Used to explain the ideas, plans, concepts in detail and analyze them easily with the help of mindmap, layer chart, decision tree, pert chart, etc

Strategic Visualization:

- Used to represent organization stretegies of development, formulatiom, communication, implementation and some time its analysis with the hep of organizational chart, stretegy map, failure tree and portfolio diagram.

Metaphor Visulization:

- This structures and orgaizaes information graphically.
- Helps to express insight of information using metaphor charactreistics lie metro map, story template, funnel, tree, etc.

Compound Visualization:

- Allows to merge diff graphical formats into single schema or frame.
- Cartoon, Rich picture, knowledge map, learning map, etc

In this chapter we are only focusing data and information visulization.

Data visualization techniques

There are some **data visualization techniques** like :

1. Table
2. Pie Chart
3. Bar Chart
4. Histogram
5. Line Chart
6. Area Chart
7. Scatter Plot
8. Bubble Chart
9. Multiple Data Series

Information visualization techniques

There are some **Information visualization techniques** like : .

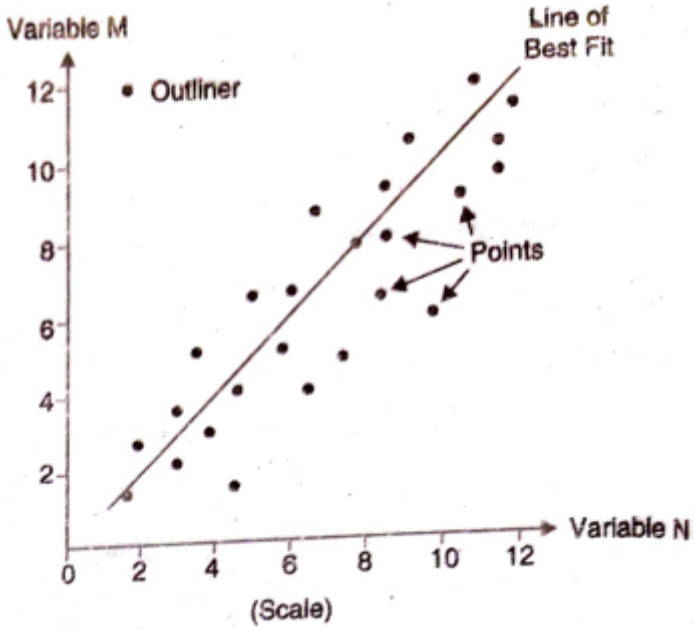
1. Parallel Coordinates
2. Tree Map
3. Entity Relationship Diagram
4. Cone Tree
5. Time Line
6. Flow Chart
7. Data Flow Diagram
8. Venn Diagram
9. Semantic Network

Types of Data visulization

- Table
  - Smallesh unit in the table is cell which is made up of row and column no
  - Rows are called as tuple, record, vector and columns are called as attirbutes, fields, properties, paramters, etc
  - P.E. Hoffman brings a new term for table as "table visulization"
  - Here row is reffred as Dimension that signifies indepedenet variable of a record adn the column is refferd as variates and signifies the dependent variable.
- Histogram
  - most commonly used grpahical representation technique

- A vertical bar chart is used to draw the histogram which represets the distribution of a set of data over a interval or certain time period and relationships of a single variable over a set of classes.

- Scatter Plot
  - scatter plots are also known as XY plots, scatter charts, scatter graphcs, point graphs and scattergrams.
  - Primary use of scatter plot is to represent corelation and determine the outliers in the dataset.
  - Scatter plots helps to detech the existence or absense of a relationship between two variables (M and N)



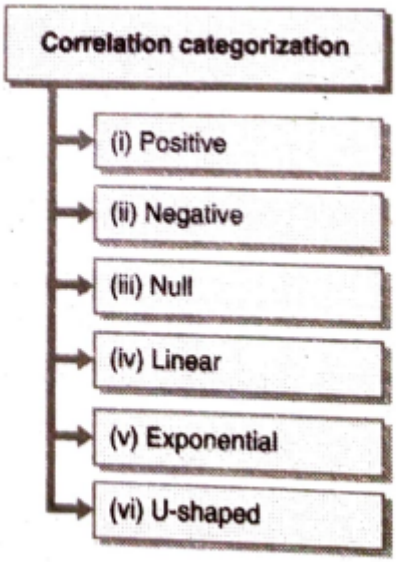
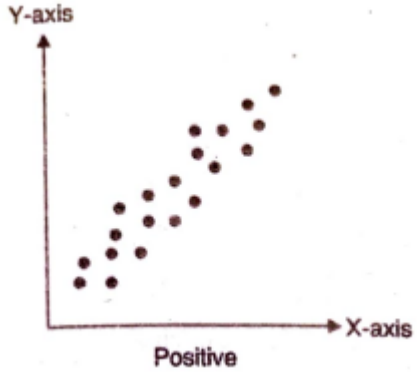


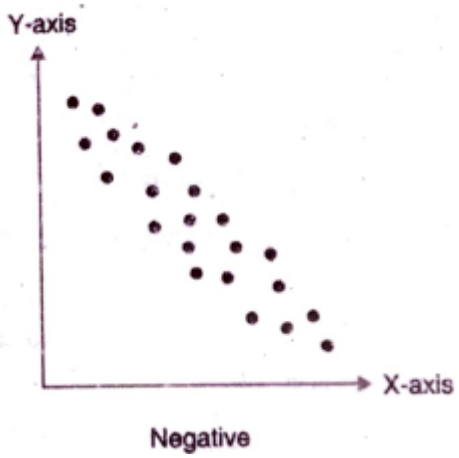
Fig. 5.5.4 : Correlation categorization

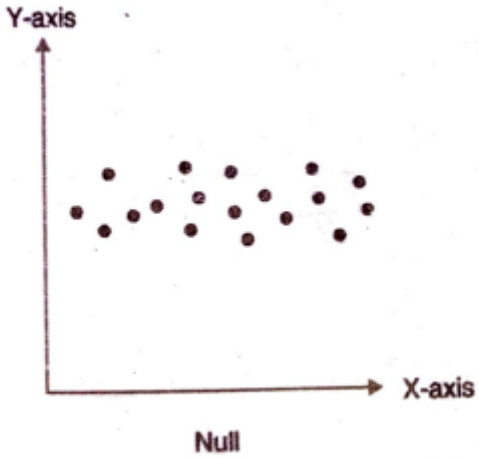
  - Types of correlation

- Positive: X-axis values increases the value at Y-axis also increases.



- Negative: X-axis value incrases and Y-axis value decreases.





- Null: No correlation

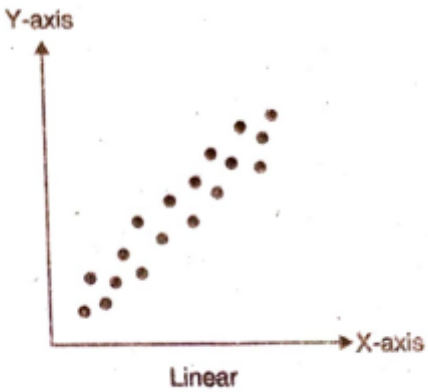
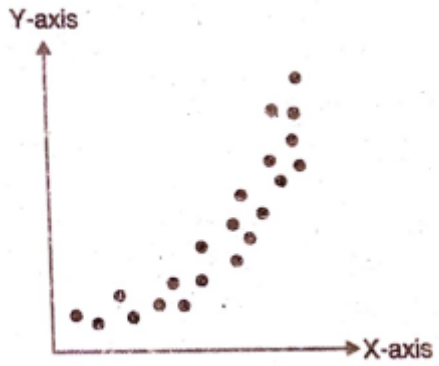


Fig. 5.5.8

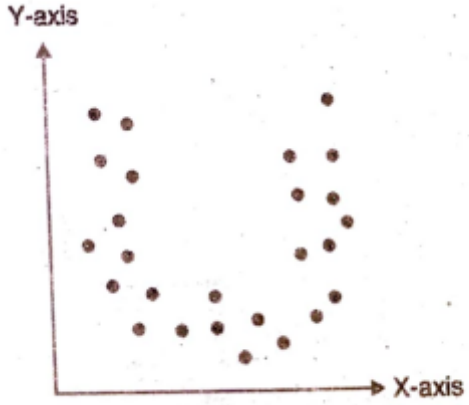
- Linear: linear line

- Exponential:



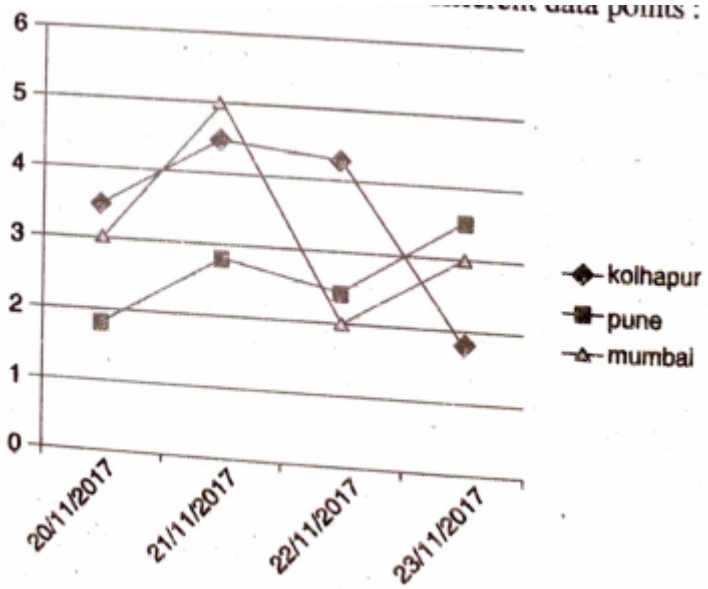
Exponential

- U-Shaped:



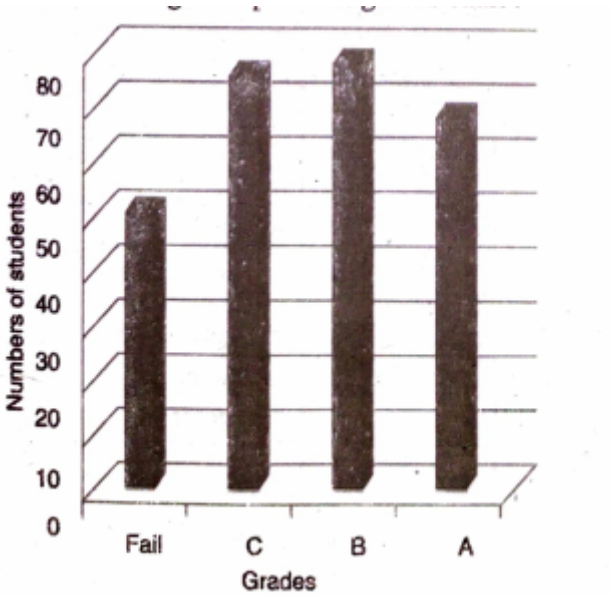
U-Shaped

- Various Charts
  - Line chart:
    - The extension of scatter plot is line graph.
    - Most basic chart type which is formde by connecting a series of data points together with a line
    - Line charts are used when we have a continuous data set.



    - trend based visualizations
  - Bar chart:
    - used to represent discrete data

- generally bar charts look similar as histograms but bar charts cant't be used to represent continuous data over time intervals.
- Example



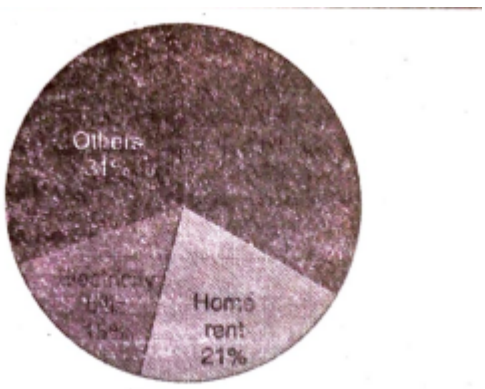
- Pie chart
  - a circle is divided to represent diff catrgoris of a dat set.
  - circle is divided into proportional segments called as sectors.
  - Example
  - 

    Fig. 5.5.15

- Area Chart
  - Area chart is used to graphically represent the development of quantitative values / data over the intervals.
  - both line chart and area chart are used to represent time-series relationip or continuity over a dataset or trends.
  - diff between line chart and area charts is that the area chart fills the area below the line with a specific color or texture.

- Example

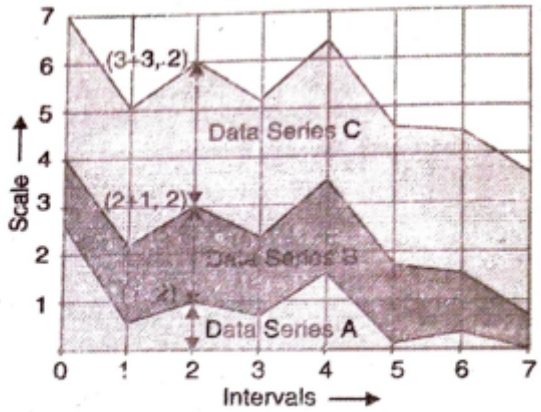
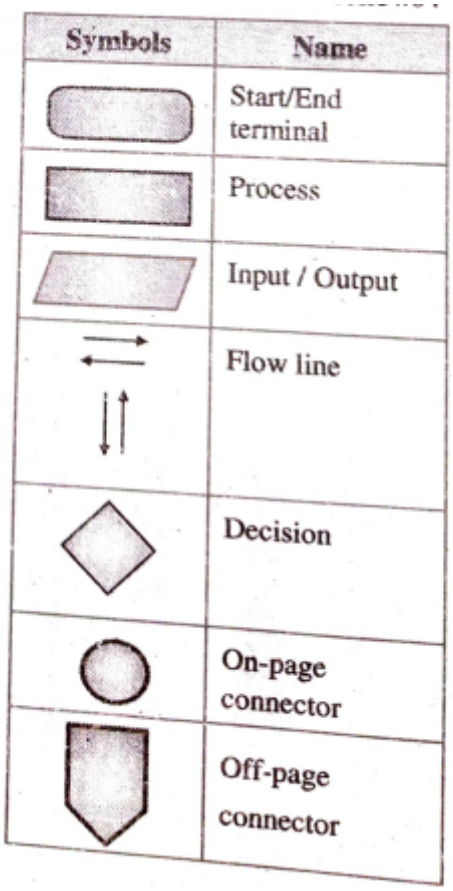


Fig. 5.5.17

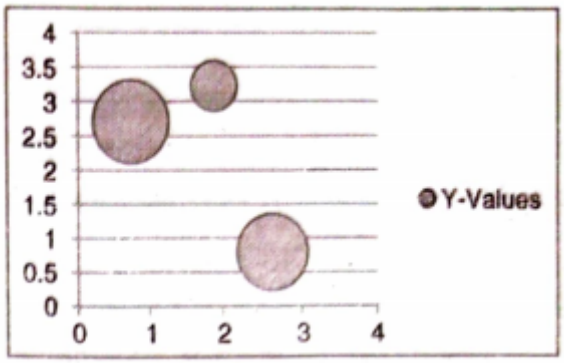
- Flow Chart

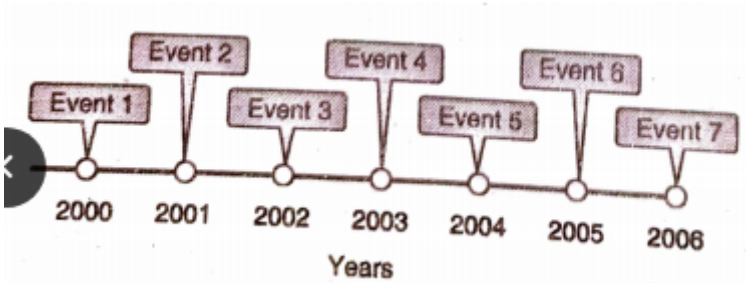


Symbols	Name
	Start/End terminal
	Process
	Input / Output
	Flow line
	Decision
	On-page connector
	Off-page connector

- Flow chart symbols

- Bubble chart
  - bubble chart is variation of scatter plot
  - scatter plot compare two values whereas bubble chart adds one more dimenision size as a the third variable.
  - So one value represents X-axis value, second values represents the Y-axis value and the third value representes the size of the bubble.

- Example 

- Timeline
  - timeline is a pictorial representation of a series of events in chronological sequence along with drawling straight line.
  - Sometime also called as chronology
  - most of the time the timeline is a graphical representaton of past events which are the part of history.
  - categories of timeline:
    - linear timeline
    - comparative timeline
  - Example 

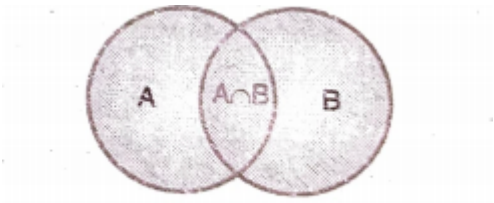
- Other Diagrams: venn, data flow and entitiy relationship diagrams
- Venn Diagram
  - in 1880 john venn introducess venn diagrams whcih are used to represent the relationship among two or more sets so that are also known as set diagrams.
  - Example 

Fig. 5.5.25
- Data Flow Diagram (DFD)
  - Data flow diagram is used to graphically represent the data or information transformation; means it is used to reprsent the data processing, data storing, and flow of data throught the processes.
  - DFD are used to visualize the data processing.

- DFD's can be of diff levels like DFD level0, DFD level1, etc.
- level 0 gives the overview of entire system whereas level1 represents detail view of each proess.
- Example

– Fig. 5.5.26 shows the simple DFD Diagram.

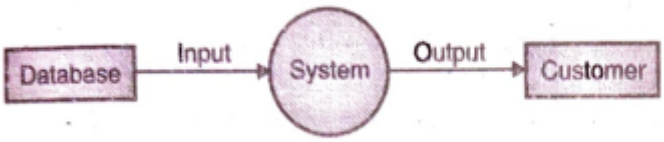


Fig. 5.5.26

- ER Diagram
  - Entity Relationship diagram
  - Mainly used for database modeling.
- Other Data visulization Methods: Parallel Coordinates, treemap, cone tree and semantic networks.
- Parallel Coorindates: used to plot individual data elements over several dimensions. useful for displaying multidimensional data
- Treemap and cone tree: used for visualizing the hierarchical structured data.
- Semantic Networks: To represent logic relationship among several concepts semantic network is used. It generated directed graph.

Tools used in Data Visualization

Mutlidimensional Visulization:

- Two types:
  - First type exmines the ctegory proportions or category counts For ex histogram, pie chart, bar chartm, etc
  - Second type exmines the relationships among the variables. Ex: line chart, area chart, scatter plot, etc.

Mutlidimensional Visulization Tools

- Google Charts
- Many Eyes: introduced by IBM research and devloped by using JAVA and Flash.
- Tableau Public
- Weave: weave stands for web based analysis analysis and visulization environment.
- Wordle: takes the text as input from the user and generated "word clouds"

Hierarchical Visulization tools

- Network Workbench
- Protovis

Open Source Data visulization tools

- Datawrapper
- Chart JS
- Raw
- Charted
- Timeline
- Leaflet

Data visualization with Tableau

- In data analytics and visulization the most popular tool is tableau.
- It is a business intelligence (BI) tool.
- It supports interative visulization of data.
- It provides faster visualization because it has an in-memory data engine. For cacheing and in memory analytics it used Hive.
- The hadoop infrastructure can be embed by tableau.
- Tableau public is a free desktop version software.

Analytical techniques used in Big Data Visulization

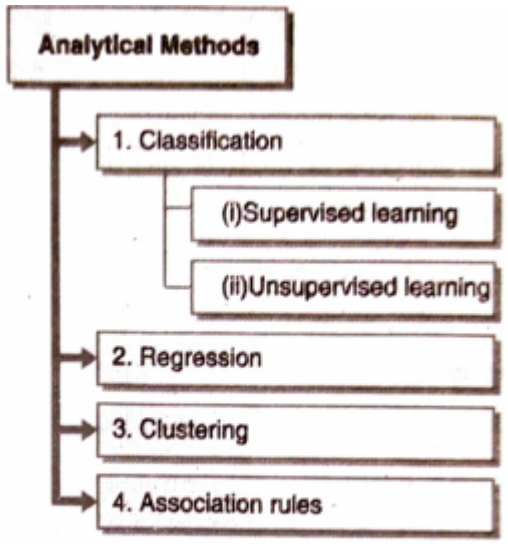


Fig. 5.11.1 : Analytical methods

Classification: The classifier assign the class labels to the new observations. Ex: Decision tree classifier

Regression: The main diff between classification and regression is that, classification predicts something will happen, whereas the regression predicts hwo much of it will happen. Ex linear and logistic regression.

Clustering: Unsupervised technique Ex KMeans

Association Rules: Unsupervised and discriptive technique not predictive. Ex: Apriority and FP Growth.

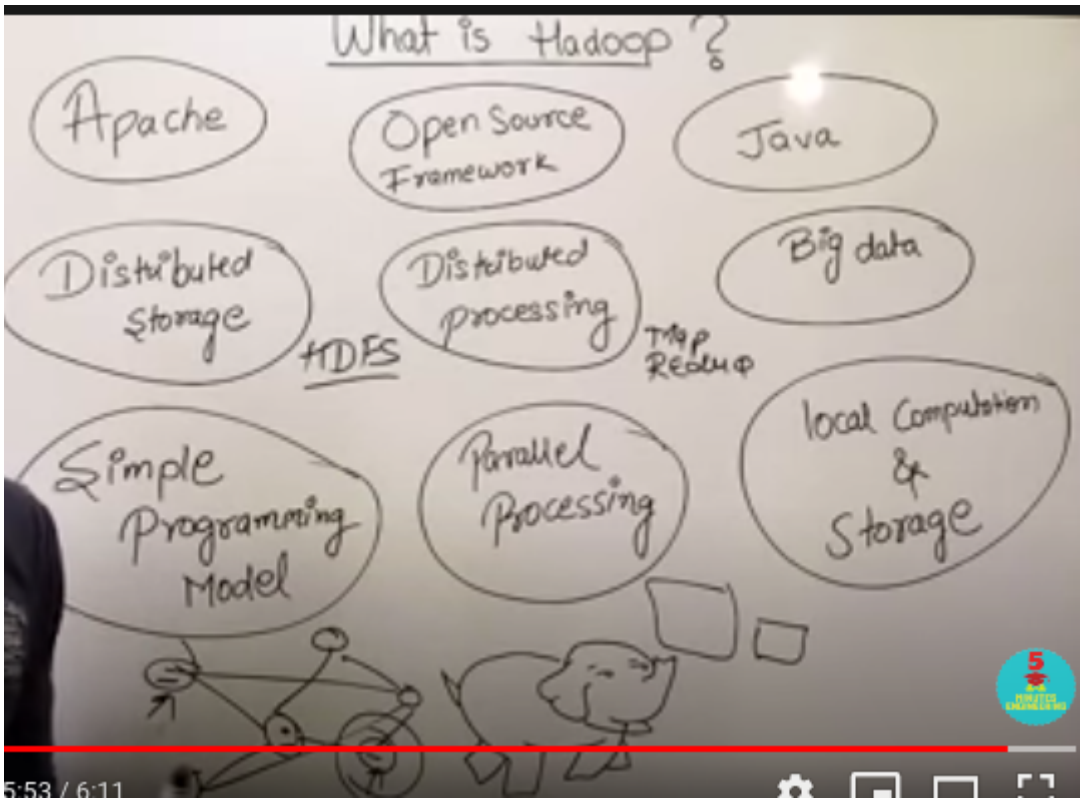
Unit 6

Types of Data:

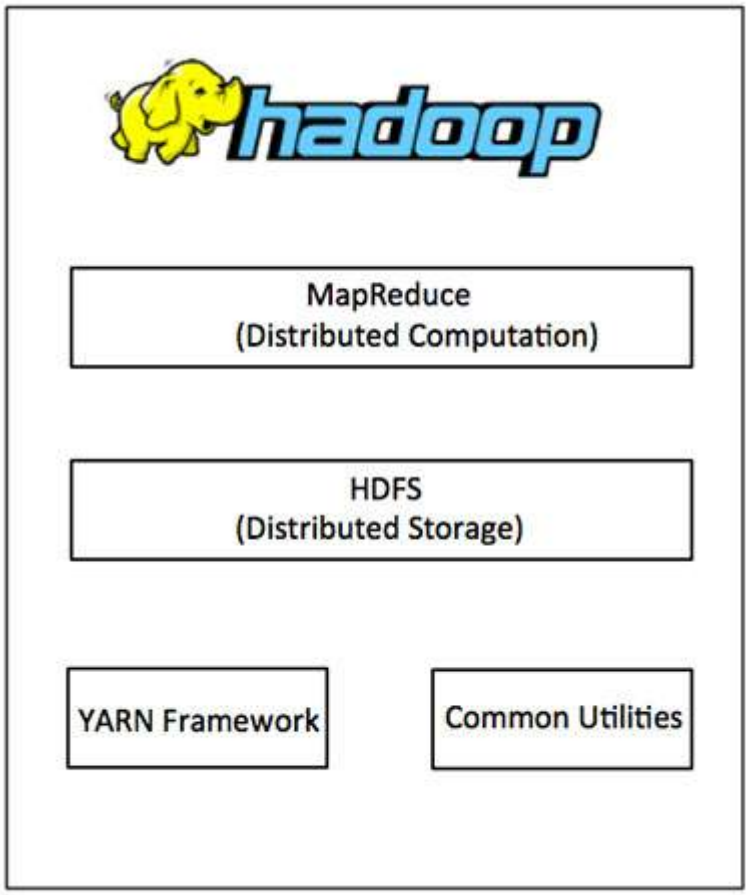
- Structured Data: A precise and steady format (data table)
- Semi-Structured: format which describes itself (XML)
- Quasi-structured: little bit inconsistent format (hyperlink)
- Unstructured: incosistent format (text or video)
- The Map reduce paradigm provides a way to divide the huge task into smaller subtasks and then run those subtasks simultaneously and merge the results of the individual tasks into single final output.
- Apache Hadoop provides a sofwware implementation of Map Reduce.

Apache Hadoop

- Hadoop is an open source java based programmign framwork, which is used for data storage and data processing of large datasets in a distributed computing environment.
- Hadoop is created in 2006 by Doug Cutting and Mike Cafarella to support distribution for the Nutch (Search Engine).
- In 2008 Yahoo launches hadoop as an open source project.
- Now Apache Software Foundation (ASF) manages the hadoop framework.



Modules (Components) of Hadoop

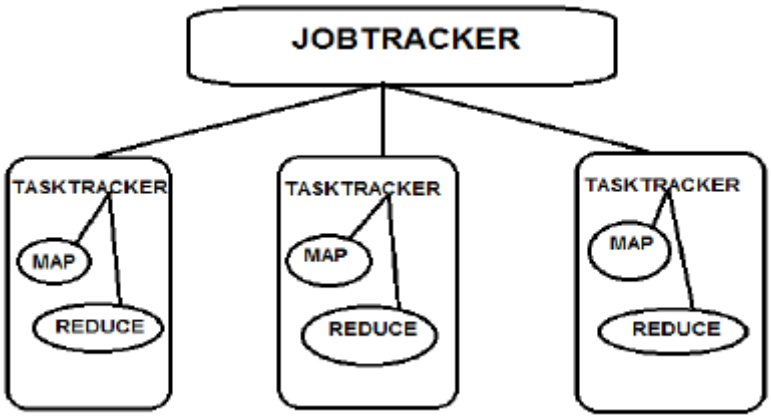


- HDFS:
  - HDFS stands for Hadoop Distributed File System
  - It divides the file into multile blocks and store onto diff nodes in the distributed environment.
- YARN:
  - YARN stands for Yet Another Resource Negotiator.
  - It is used for job scheduling and manging the muliple nodes (clusters)
- MapReduce:
  - MapReduce is YARN based system for parallel processing of large data set using key value pair.
- Hadoop Common:
  - contains java libraries and utilities which are used to start hadoop.

MapReduce

- MapReduce is a software framework which is used to write application easily to process huge amount of data simultaneously on large clusters in reliable and fault tolerant manner.
- In mapreduce therer are two tasks: map task and reduce task
- Map task: This is the first task which gets the input data and convert it into another format. (key value format)

- Reduce task: It takes the output of the map task as input and combine those data tuples into a smaller set of tuples.
- The map task is performed by mappers and the reduce task is performed by reducers.
- Hadoop contains single master JobTracker and one slave TaskTracker per cluster node.



- Master JobTracker is reponsibile for managing resources, asssigning tasks to the slave tasktrackers, monitoring the tasks and re-executing the failed tasks, etc
- TaskTracker executes the task asssigned by master and provide the task status information back to master.
- Here master is very important. If JobTracker goes down, all running tasks get halted.

HDFS

- HDFS stands for Hadoop Distributed File System.
- This is the primary file system that is used in hadoop applications.
- The HDFS is based on Google File System (GFS).

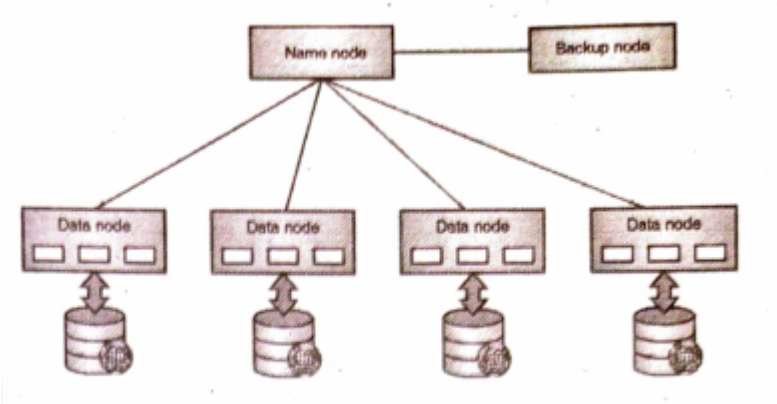


Fig. 6.1.4 : HDFS

- The architecture used by HDFS is master slave architecture.
- Name node stored the metadata information.
- Actual file data is stored on data nodes. So in the HDFS the file is divided into blocks and that blocks are stored on the data node and also replicated to mulitple data nodes.
- The namenode contains the mapping of file block to particular data node.

The Hadoop Ecosystem

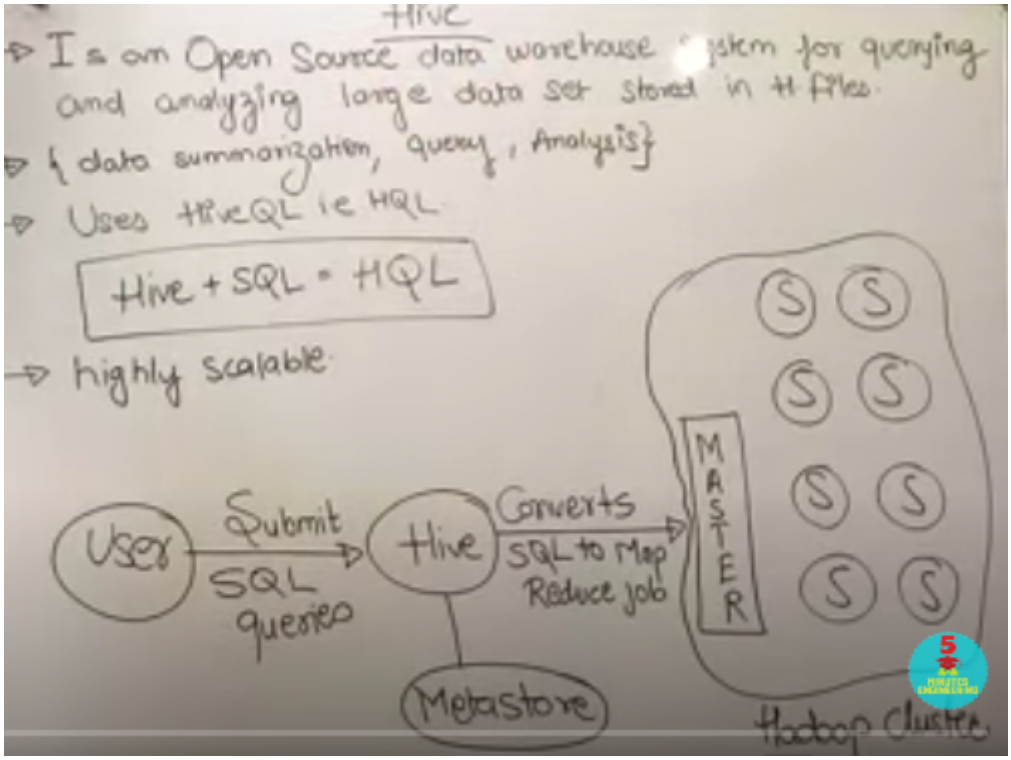
- Hadoop is a platform or framework for Big data applications.
- It contains open source as well as commerical applications or tools.
- Some popular hadoop ecosystem open source tools are pig, HBase, Mahout, Hive, etc.

Pig:

- Apache pig is a high level platform to create program that runs on the Apache Hadoop.
- Pig uses Pig latin programming language to create the programs.
- Pig was originally developed at Yahoo for researcher to have an adhoc way of creating and executing MapRedice jobs on very large data sets.
- In 2007 it was moved into the Apache Software Foundation (ASF)

Hive:

- Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.



Major components of the Hive architecture are :

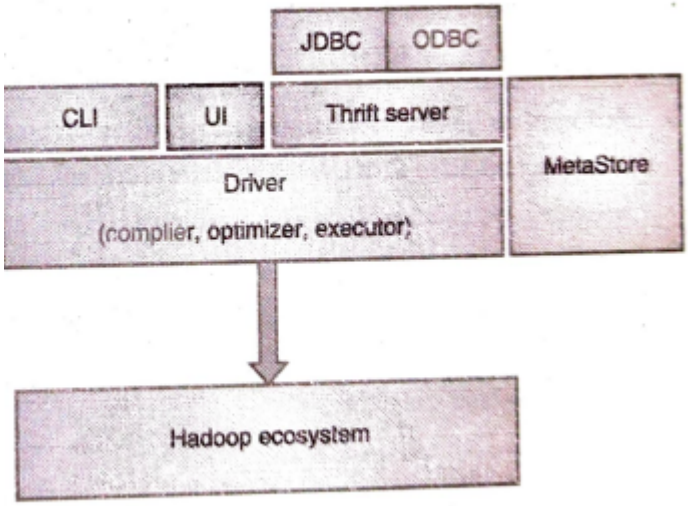


Fig. 6.2.1 : Architecture of HIVE

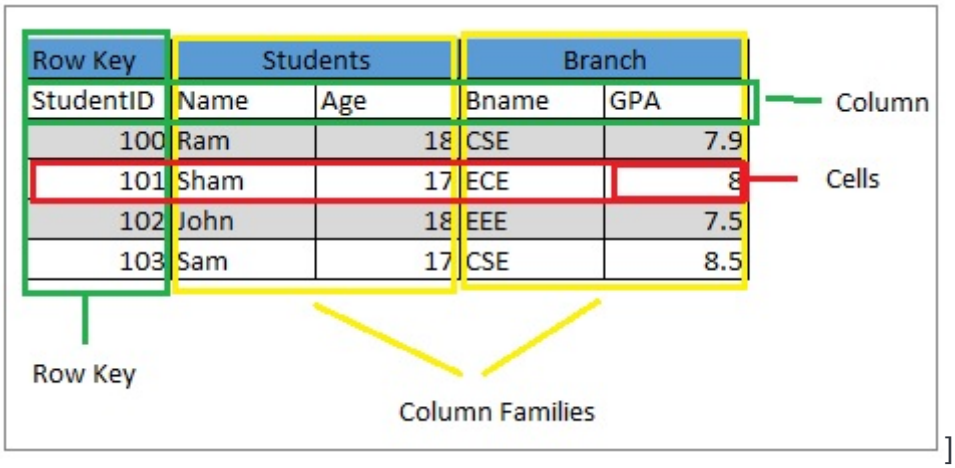
HBase:

- In hadoop big data is stored using HDFS using distributed storage and for data processing mapreduce is used.
- But in hadoop the access to the data is sequential and no random access allowed to the data. That's why HBASE comes into the play.
- HBase built on the top of the HDFS.
- HBase is open source and horizontally scalable.
- It is a distributed column oriented database.

HBase Data Model Terminology

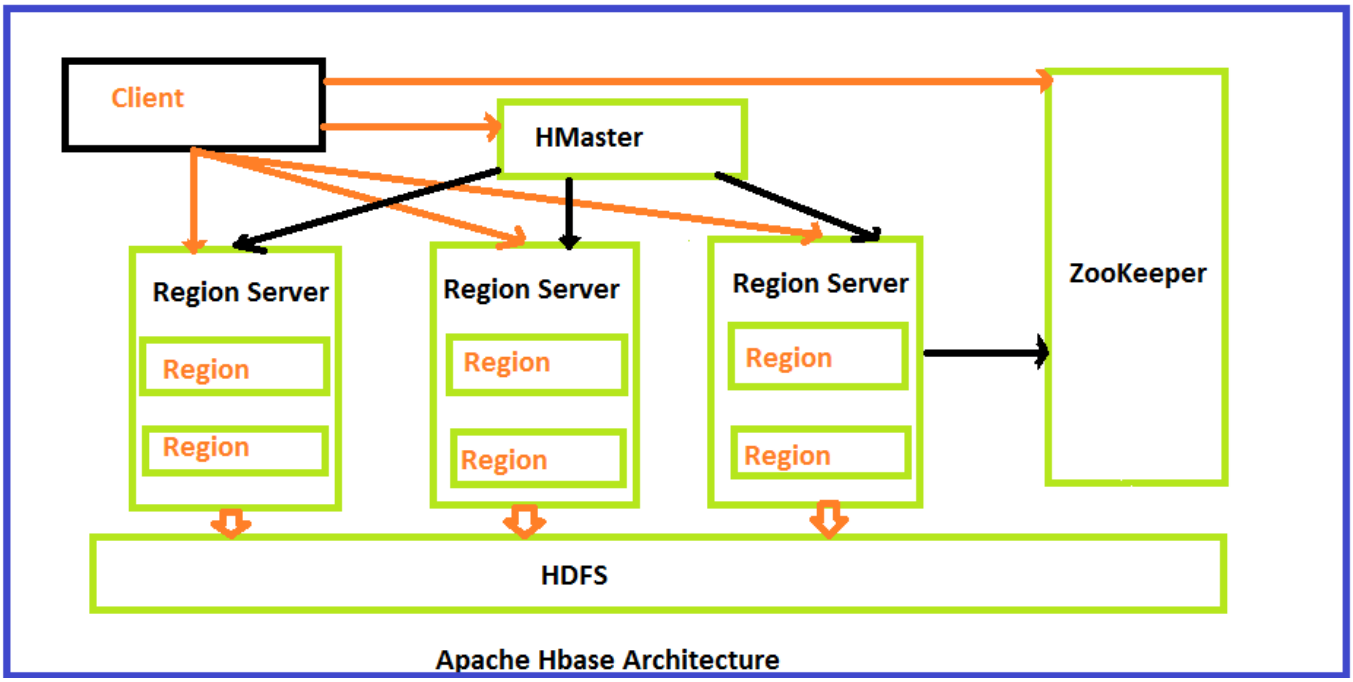
- Table: collection of rows
- Row: rowid and collection of column families

- Column:
- Column family: collection of columns
- Column Quantifier: to provide the index to the data, a column quantifier is added to column family.
- cell: cell contains a value and a timestamp which represents the value's version.
- Timestamp



]

HBase Architecture



HMaster: HMaster is the master server in the HBase. It is resonsbile to manage the region servers and to monitor the health of the region server. Aprat fromt this HMaster is also reponsible for maintaining the meteadata. So for DDL (create table, remove table, modify column, add column, etc) commands HMaster is used. HMaster works on the name node in the Hadoop cluster.

Client: It is the actual client which wants to perform operations. Client communicates with the Zookeeper and HMaster. For read and write operations client directly communicates with the Region severs. For DDL related commands client communicates with HMaster.

Region Server: Region server handles the read and write request from the client and perform that operations on the region. Region is the actual area where actuall column of the tabels are present with

the data.

ZooKeeper: ZooKeeper is the centralized monitoring server which manitains configuration and provides distributed synchronization. For read and write operatiosn client first sends the reqyest to Zookeeper and then to the region servers.

Mahout

- R language is suitable to perform analysis on the smaller data sets. But for large dataets it is not very effective and that's why we use mahout.
- Mahout provides java executable libararies to perform analysis in the hadoop environment.
- Apache Mahout is considered as a toolset which directs Hadoop (the elephant in this case), to get meaningful analytic results.
- Java code is provided by Mahout which implements the algoithms for various techqniues in the following categories:
    - Classification
        - Logistic regression
        - Naive bayes
        - random forest
        - hidden markov models
    - Clsutering
        - canopy clustering
        - KMeans
        - Fuzzy KMeans
        - Expectation Maximization (EM)
    - Recommendation/Collaborative Filtering
        - Nondistributed recommenders
        - distributed item based collaborative filtering

NoSQL

- NoSQL means non sql, non relational or not only SQL.
- NoSQL provides various advantages over realtional systems like here there is not any fixed schema definition which we have to follow while inserting the data.
- Relational SQL databases are not suitabel in terms of scaling and efficient use of commodity hardware and processing.
- NoSQL databases are more concentrated on availability, paritition scheme, speed for which they may compromiae the consistency.
- In 1998 Carlo Strozzi first introduced open source lightweight relational database sytsem named as "NoSQL".

- NoSQL is prefered over SQL for Big data applications.
- SQL support complex quiers but NoSQL cannot support complex queries like SQL.

Types of NoSQL Databases

- Key value store
- Document store
- Column store
- Graph store

Key Value Store:

- key value store databases uses the associative array concept as a fundamental data model.
- We can say like it is table with 2 column where one column is key and the other column is value.
- keu should be unique because key is used to identify the value.
- Examples: Redis, Memcached, Dynamo, Riak, etc
- These all key value store databases are diff in terms of some properties like persistence, in memory, etc.
- Memcache is not persistent while Riak is persistent.

Document Store

- In Document store record are stored as documents.
- Here document contains collections of key value pairs.
- Each document is identified by unique key
- Document store is the extension of key value store where we can perfomr the querying on the documents, indexing, etc.
- Examples: MongoDB, Couchbase, etc.

Column Store

- Column store are column oriented NoSQL databases.
- Here data is stired into cells grouped in the columns
- columns are group together into column family.
- Read and write operations are performed on columns rather then on rows.
- Examples: HBase, Big table, etc.

Graph Store

- Examples: Neo4j, Polyglot

NoSQL Aggregate Model

- The key-value and document store databases are strongly aggregate oriented.
- The Column family has two level aggregate structure. The first key is the row identifier. The second level value are defined to as columns.

An Analytics Project

- A pilot project indicates a project before a full-scale rollout of the newly added algorithms or functionality.