Introduction to data Science :-

↳ **Data science :-** is a combination of multiple disciplenes like stastics, Data Analytics and Machine learning.
It is basically a data gathering, analyzing the data finding or discovering the pattern in data and predicting the future outcomes by making better decesions.

↳ **Data science need :-**

Data science is used in many industries like - banking
- consultancy.
- healthcare
- Consumer goods.
- E-commerce
- Politics
- logistics Company

In consulting the people with prediction related to the terms.

Analyzing which health exercises and treatment are best

Prediction of future Profits and outcomes of company.

Promotion of Products.

Prediction in election

Analyzing routes / predicting delays in transport.

Basically Datascience is a process of using algorithm, methods and systems to extract knowledge & insight from structured & non-structured data. It uses Analytics, ML to help users make decesion, prediction, enhance optimization and improve opr^n.

Evolution of Data science :-

Stage - I :- In 1962 J.W Turkey published article on "The future of Data Analysis". it was a relationship between Stetistics and Data analysis. At that time bussiness transaction was centralized.

Stage - 2 :- In 1977, J.W turkey published another article related to hypothesis for testing & analyzing the data.

Term Data mining comes into the picture of world.

stage - 3 :- Bussinessmen started using data analysis and using in prediction of their profit and future decesion. They started getting efficiency because of data analysis.

stage - 4 :- Data science started begin recognised by publics. researchers started researshing on D.S. Journal started commenting

on D.S. Big data comes into picture

Stage-5

The term Data science was introduced to World. Acc. to IBM around 90% of total data available to world was produced in last two yrs.

Stage-6

D.S started usig in various society or workig culture all over the world. Everyone's firm started collectig vast data.
- Apple gave credit to Big data and data mining.
- MS, Goog, started usig DS for voice recognitn & speech detectin.

Roles of Data science :-

→ Data Scientist Role :
- collect & Analyze data
- extract insight from patterns using ML model.
- explain & visualize result

→ Data Analyst
- Domain exploration & understandig
- collect & Analyze data
- understand data
- explain & visulize result / optimization
- communicate with stakeholder.

→ Data Engineer
- transform raw data into usable pipeline
- Upgrade the exisisting version of technology with upgraded or newer version.
- Test and Deploy ML models.

→ Database Administrator
- Database backups & recoveries
- responsible for proper functioning of all database.

→ Data Architect
- creates blueprint for data managment
- Also ensures the resources for engineers

→ Data & Analytics Manager.
- oversees the data science opr^n
- Assign duties to their team.

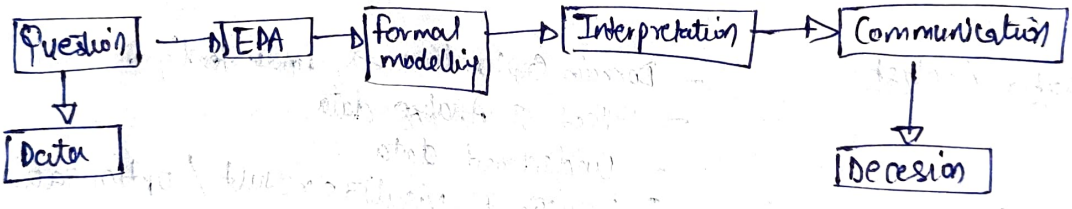
- Bussiness Analyst :
  - Bussiness oriented work
  - identify how Big Data can be linked to actionable bussiness insights for bussiness growth.

- Statician
  - understanding of statical theory
  - data organization
  - create new methodology for engineers

- Machine learning Engineer
  - implement common ML algo.
  - develop pipelines
  - A/B testing.

→ Structure of Data science Project.

QDEFICD



5 phases

1 - Questioning phase :
- most important phase
- helps you understand your data & decide on the type of analysis.
- To extract your data from bigger dataset, one can use distributed storage like Apache Hadoop, Spark etc.

2 - EDA (Exploratory Data Analysis)
- Check if the data you have is suitable to answer your question
- develop the sketch of solution.
- check if your dataset carries all the data that is required.

Formal modelling :
- write down the parameter you are trying to estimate.
- challenge your results through variety of approaches like sensitivity analysis.
- make sure data & algo are reproducible.

Interpretation :
- phase is to assemble all the information you've got after analysis.
- helps to filter out the result you've got.

Communication
- Project communicated to some sort of audience
- Make sure the result of your project are visulized for quick understanding.

Characteristics of data 5 V's

Volume - refers to the vast amount of data generated every second.

variety - refers to the different types of data we can use now.

velocity - refers to speed at which new data is generated and the speed at which data moves around

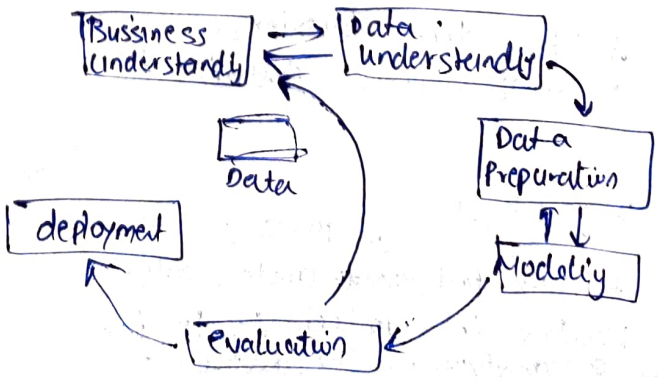
veracity - refers to the messiness or trustworthiness of the data.

Value - refers to having access to big data is no good unless we can turn it into value.

→ Existing Data Analytic Methodologies.

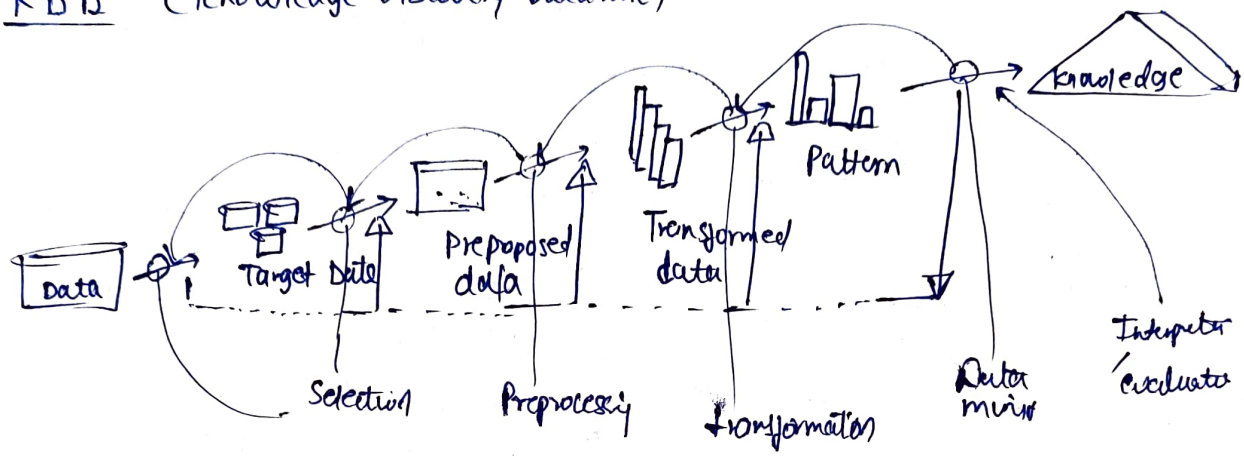
CRISP - DM

Cross-industry standard process for Data mining.



Bussiness understanding - What does bussiness need?
Data understanding - What data do we have / need?
Data preparation - How do we organize data for modeling
Modeling - What modelling techniques should we apply
Evaluation - which model best meet bussiness model
Deployment - How do stakeholders access the result.

→ KDD (Knowledge Discovery Database)



selection - targeted data is determined through database of complied data.

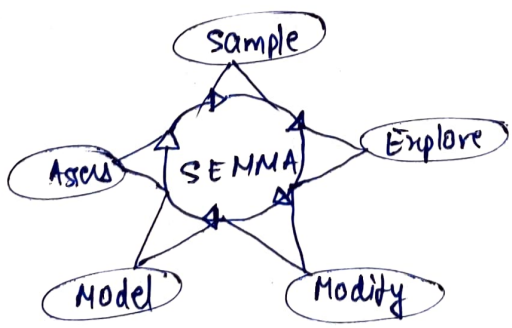
Pre-processing - improving the data

transformation - converting pre-processed data to fully utilized kind.

Data mining - focuses on shifting through transformed data to seek out pattern.

Interpretation / Evaluation :- Cleaned, converted, relevant data framed into visual representation.

→ SEMMA
Sample
Explore
Modify
Model
Assess



Sample :- Generate a representative sample of data

Explore :- visulization & basic description of data.

Modify :- select variables, transform variable representation.

Model :- use variety of statistics and ML model.

Assess :- evaluate accuracy of model.

Appl^ of Data science:

- In search engine :- - Data science is used to get searches faster.
  - for ex. if we search for something on google, google will show the result of that website on the top which is visited most.

- In transport :-
  - Driverless cars
  - training data is fed into the algorithm and with the help of DS techn., the data is analyzed like speed limit in highway n all.

- In finance :-
  - stock market
  - D.S is used to examine past behavior and to examine the future outcome.

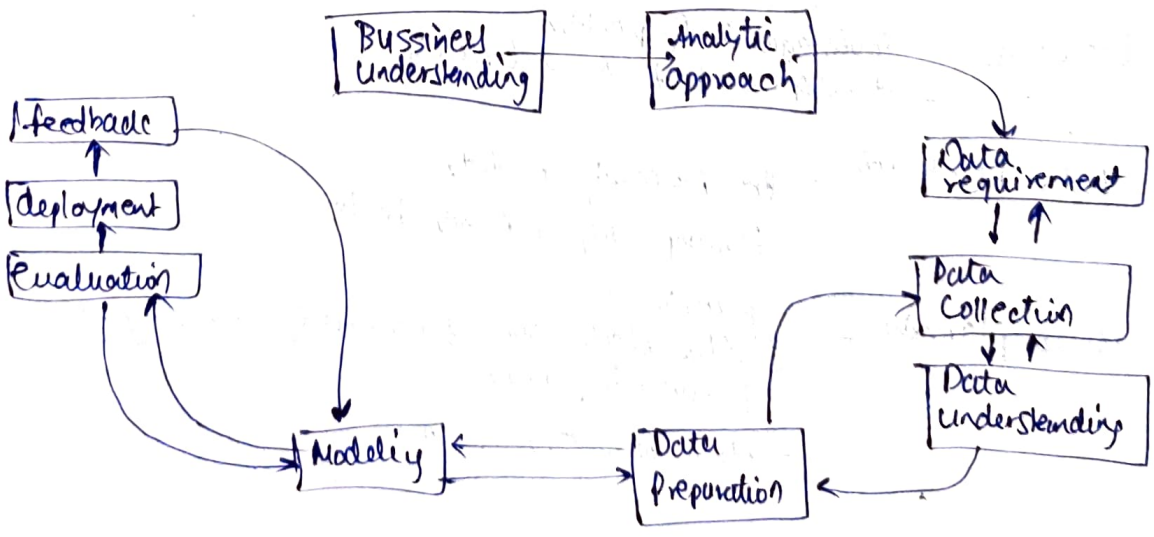
- In E-commerce :-
  - on searching for something on E-commerce website we get suggestion similar to choices acc. to our past data and also we get recommendation

- In healthcare :-
  - detecting tumor
  - Drug discovery
  - Medical bots.

- In Gaming :- used with ML with the help of past data computer player will improve its performance.

- In delivery logistics :- helps to find best route for shipment, best time suited, best mode of transport.

→ Data Science Methodologies.



Bussiness understanding :- problem that you're tryig to solve

Analytic approad :- selecting the right analytical approad

data requirement :- Identifyoy the necessoy data content.

data collection :- requiremants are revised and Stastic and visualization are applied.

data understanding :- activity related to construction of data

data preparation :- process of using domain knowledge of data to make ML model algo work

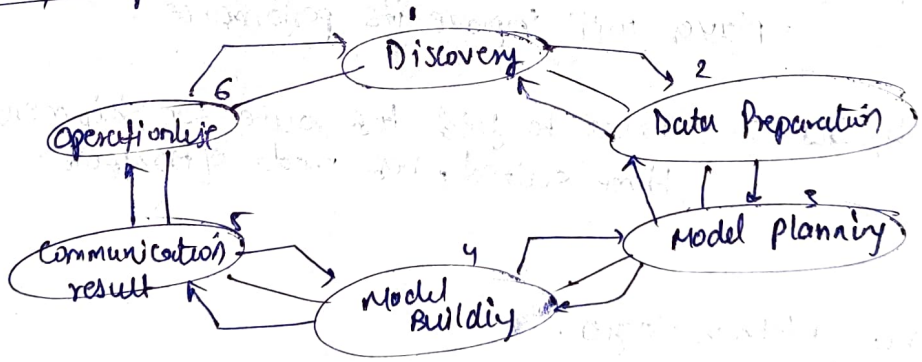
Modeliy :- focusses on developing models that are either descriptive or predictive.

evaluation :- To check wether the model used really answer the question of client

Deployment :- deployed to ultimate test

feedback - users will help refin the model.

→ Life cycle phases of Data Analytics :-



Phase

1 - Discovery - investagation of problem
- develop content & understandy

2 - Data preparation - Pre-processy of data
- Hadoop, Alpine Miner tools.

3 - Model planniy - Planning
develop data set for trainiy, testiy & production purpose
- MATLABS, STASTICA

4. Model Building - development of dataset
   - Matlab, STASTICA, Octave.

5. communication result - identify key finding, summarize and convey to stakeholder

6. operationalize - deployment on small scale
   - MADlib, WSK