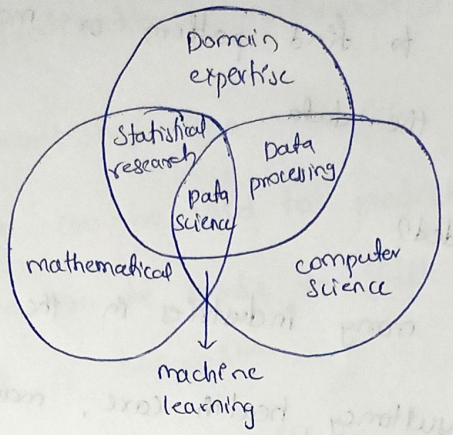
What is Data Science ?



→ Data science combines maths & statistics specialized programming, advanced analytics AI & ML with specific subject matter expertise to uncover actionable insights hidden in an organisation data. These insights can be to guide decision making & planning.

→ Data science is a combination of multiple disciplines that uses statistics, data analysis & machine learning to analyse data to what knowledge & insights from it.

→ Data science is about data gathering data analysis & decision making

→ Data science is about finding pattern in data through analysis & make future predictions by using DS companies are able to make:

1) Best decision

2) Predictive analysis

3) Pattern discoveries to find pattern or may be hidden information in the data.

Where is DS needed?

→ DS is used in many industries in the world today.

→ Eg:- Banking, consultancy, health care, manufacturing etc.

→ Data is a collection of raw facts.

→ Raw facts:- raw facts may be text / images.

→ Information:- which has a meaning for the data or which collectively carries logical data. Eg:- Hindi songs.

→ Data:- which does not have meaning.

eg:- song, text.

Data Science Television:-

The Quant shop:-

→ The DS television show we live in a data driven world, where massive computer analysis the trace of all human activity in our quest to predict the future

→ Data Science is a rapidly emerging discipline at the intersection of statistics, machine learning, data visualization & mathematical modeling.

→ The Quant shop is a television show about data & how data can be used to predict the future. We show how data can be made to talk & explore the limits of what is knowable.

→ This is a program for anyone who has ever made a bet or thought about tomorrow for anyone interested in how the mathematical or computational models that rule the world are conceived built & tested and how well they perform.

Structure :-

→ Each of the eight programs in all initial season are built around a particular real world prediction challenge we watch as a student team comes to grips with the problem learning along with them as they build a forecasting model.

→ They make their predictions & watch along with them to see if they are right or wrong. This first season features episodes on :-

1) Finding Miss Universe:-

→ we can computational models predict who will win a beauty contet? Is beauty just subjective or can algorithm tell who is the fairest one of all?

2) Modelling the movies:-

→ which movie will grass the most on christmas day? which actors deserve can data reveal the next Hollywood star?

3) Winning the body pool:-

→ How accurately can we predict juniors weight before they are born? How can data clarify environmental risks to developing pregnancies?

4) The art of the Action:-

→ How many millions will a particular J.W Turav painting fetch at auction? Can computers have an artistic some of what's worth buying?

5) White Christmas:

→ What places will ~~crack~~ wake up to a snowy Xmas this year? And how can you tell one month in advance?

6) Predicting the playoffs:-

→ Which football team will the super bowl? can Google's page Rank algorithm pick the winner on the field as accurately on the web?

7) The Ghoul pool:-

→ Death comes to all men but when can we apply actuarial models to celebrations decide who shall live & who shall die?

8) Playing the market:-

→ Speculators get rich when right and tomorrows price & poor when wrong how to equants predict?

Computer Science, Data Science, Real Science.

Data & Method Centrism:-

→ Scientist are data driven (Analysis of data)

→ Computer scientists use algorithm driven.

→ Real scientists spend anauromous (very large in size, quality or extent) amounts of efforts collecting data to answer their questions of interest.

Concern about result:-

→ Real scientists care about answers they analyse data to discover something about how the world works.

→ Good science care about whether the result make sense because they care about what the answer mean.

→ Computer scientists worry about producing a valuable looking number.

Robustness:-

→ Robustness is nothing but quality or condition of being strong.

→ Real scientists are comfortable with the idea that data has error.

→ Computer scientists are not comfortable.

→ scientists think abot of possible sources of bias or (error in their data).

→ Good programmers use strong data type and passing methodologies to guard against formating errors.

Percision:-

→ Nothing is completely true or false in science while everything is either true/false in computer science.

Asking Interesting Questions from Data :-?

→ What things might you be able to learn from a given data set.

→ What do you or your people really want to know about the world.

→ What will it mean to you once you find out?

Google N-gram :-

→ Google uses data to improve search result and provide fresh access to out of print book.

→ N-gram is a set of sequencial of N-words

→ The google N-gram displays user to selected word or passes (N-gram) in a graph that shows how those passes have occured in a corps.

→ Google N-gram corps are made up of scanned books available in google books.

Properties of Data :-

→ One purpose of DS is to structure data, may be it interpretable and easy to work with.

→ Data can be characterized into 3 groups.

1) Structured & unstructured
2) Quantitative & categorical
3) Big Data & little data.

1). Structured & Unstructured data.

Unstructured	Structured.
→ Unstructured data is unorganised but we must organise the data for analysis process.	→ Structured data is organised and easier to work with the data.
→ Eg:- collection of tweets from twitter.	→ Data is often represented by a matrix; where the rows of a matrix represented by distinct item or records and the columns represents distinct properties of item or records.
→ Our first step is generally is to build a matrix structure	
→ A bag of words model will construct a matrix with a rows of each tweet and a column for each frequently used vocabulary words.	
→ The no. of times tweet (i) words (j)	

2) Quantitative & categorical data:-

Quantitative	Categorical.
→ Quantitative Data consists of numerical data (or) values like height and weight.	→ categorical data consists of cables describing the properties of the objects under investigation like gender, hair, colour and occupation.
→ The data can be incoporated directly into algebraic formulas and mathematical models (or) displayed in conventional graphs and charts.	→ In other hand the categorical data can represent with a specific data or Ranking.
→ Quantitative data can be analysed using various statistical techniques such as mean, median, mode, variance & standard deviation etc.	→ The difference between them may not consistant or measurable.

3) Big data & Small data:-

Big data	Small data.
→ Big data refers to extremely large and complex data sets that are beyond the capacity of the traditional data processing and analysis tools. to handle effectively.	→ Little data involves relatively small amounts of information typically ranging from a few kilobytes to giga bytes. it can be stored and processed on a single computer or small server.
→ Big data can exhibit inconsistencies in data quality format or structure. This variability can make it challenging to work with and analyze the data effectively.	→ Analysis of little data can be performed using common data analysis and visualization techniques.

Developing scoring system:-

→ Scores are functions that map the features of each entity to a numerical values of merit the basic approaches for building effective scoring system and evaluating them.

Gold standards & proxies :-

→ In data science, a gold standard is a set of labels or answers that we trust to be correct.

→ In the original formulation of BMI, the gold standard was, the body that percentages carefully measured on a small number of subjects of course, such measurements are subject to some error, but by defining these values to be the gold standard for fitness.

→ The presence of a gold standard provides a rigorous or best way to develop a good scoring system.

→ We can use curve fitting technique like linear regression to weight the input features so as to best approximate the "right answers" on the gold standard instances.

→ But it can be hard to find real gold standards proxies are easier to find data that should correlate well with the desired but unobtained ground truth.

→ BMI was designed to be a proxy for body fat percentage. It is easily computer from just height & weight and does a pretty good job corelating with body fat.

→ Proxies are particularly good when evaluating scoring / ranking systems.

Scores Vs Ranking:-

→ Rankings are permutative ordering of entries by merit, generally constructed by sorting the output for some scoring systems. popular examples of ranking / Rating system includes:

(1) Football / Basketball top twenty:-

→ Press agencies, generally rank the top college sports teams by aggregating the votes of coaches or sports writers.

→ Typically each voter provides their own personal ranking of the top 20 teams, and each team get awarded more points the higher they appear on the voters lists.

→ Summing up the points from each voter gives a total score for each team and sorting these score defines the ranking.

(2) University Academic Rankings:-

→ The magazine called Us News and world report publishes annual rankings of the top American colleges and universities.

→ Their methodology is proprietary and changes each year, Presumably to motivate people to buy the news ranking.

→ But it is generally a score produced from statistics like faculty/student-ratio, acceptance ratio, the standardized test score of its students and applications

→ Since scores and rankings are duals of each other, which produces or provides or more meaning full. Representation of the data? As in any comparison the best answer is that it depends on factors like:-

i) will the numbers be presented in isolation?

ii) what is underlying distribution of score?

iii) Do you care about the extremes of middle?

Recognizing Good Scoring Functions:-

→ Good scoring function are good because they are easily interpretable and generally believable. Here we review, the properties of statistics which point in these direction.

(1) Easily Computable:-

→ Good statistics can be easily described and presented. BMI is a excellent example, it contains only two parameter and is ~~ex~~calculated evaluated using only simple algebra.

Some of the other examples are as follows:-

1) easily understandable.
2) Monotonic Interpretation of variables.
3) Produces generally satisfying results on outliers.
4) User systematically normalized variables.

BMI [Body Mass Index]:-

→ BMI is a measure used to access an individuals body weight in relation to their height. It is calculated by dividing a persons weight in kilograms by the square of their height in meters.

→ Formula to calculate BMI:- $\frac{\text{weight (in kg)}}{(\text{height})^2 \text{ (in meter)}}$

→ BMI of an individuals is calculated by the use of a mathematical formula it can also be estimated using weight in pounds to estimate BMI.

BMI	weight status
Below 18.5	under weight
18.5	Normal
25.0 - 29.9	over weight
30.0 above	obese.

agram:

→ BMI is commonly used to categorize individuals into different weight classes such as underweight and obese.

→ BMI is used to provide a good measure of obesity. But BMI fails to provide actual information on body composition like amount of muscle bone fact etc.

→ As single measure BMI is clearly not a perfect measure of health. But its still a useful starting point for important condition that becomes more likely when a person is obese.

Asking Interesting Questions from Data:—

1) What things might you be able to learn from a given data set?

2) What do you/your people really want to know about the world?

3) What will it mean to you once you find out?

The baseball encyclopedia :-

→ A large collection of information about one or many subjects often arranged alphabetically in articles in a book or set of books.

→ Baseball has long and an outsize importance in the world of data science.

→ What makes baseball important to data science is its extensive statistical record of play, dating back for well over a hundred years.

→ Baseball is a sport of descrete events; pitchers throw balls & batters try to hit them, that naturally lends itself to informative statistics.

The Internet movie Data base (ImDb) :-

→ The Internet movie db is an online db containing information and statistics about movies, TV shows and video games as well as actors, directors & other film industry professionals.

→ This information includes list of cast, to crew members, movie release dates & box office information.

→ ImDb was first published in 1990 by Weedham a computer programmer as a group of scripts which allowed users to search atleast of film credits.

Advanced Ranking Techniques:-

→ In the absence of any gold standard, these method produce statistics which are often revealing and informative.

→ When several powerful techniques have been developed to compute ranking from specific type of inputs.

① Elo ranking:-

→ Ranking are often formed by analyzing sequence of binary comparisons which arise naturally in competitive between entities.

* Sports Contest Results:-

→ Typical sorting event

→ Assume that in a football (or) chess matche. pit teams 'A' & 'B' against each other.

→ only one team will win.

→ Thus each match is essentially a binary comparison of merit.

* Implicit Comparisons:-

→ Suppose a student has been accepted by both universities 'A' and 'B' but opts for 'A'. This can be taken as an implicit vote that 'A' is better than B.

→ The Elo system starts by rating all players equally and then incrementally adjusted to each players score in response to the of each match.

$$r'(A) = r(A) + k(S_A - \mu_A)$$

→ where → $r'(A)$ & $r(A)$ represent the previous and updated score for player 'A'.

→ k is a fixed parameter reflecting the maximum possible score adjustment in response to a single match.

→ S_A is the scoring results achieved by player A in the match under consideration, Typically $S_A = 1$ if A won & $S_A = -1$ if A lost.

→ μ_A was the expected once competing against 'B'. if 'A' has exactly the same skill level as B then presumably $\mu_A = 0$.

⇒ What is expected Match score?

if $P(A>B)$ estimates the probability that A beats B,

$$\mu_A = 1 \cdot P_{A>B} + (-1)(1 - P_{A>B})$$

→ If the ranking system is meaningful, this probability should be a function of the difference the score $r(A)$ & $r(B)$.

The target function!-

→ We need a function $f(x)$ that takes 'X' and 'Y' fields a probability

$f(0) = 1/2$

$f(\infty) = 1$

$f(-\infty) = 0$