Unit - I

Introduction: - what is Data Science?

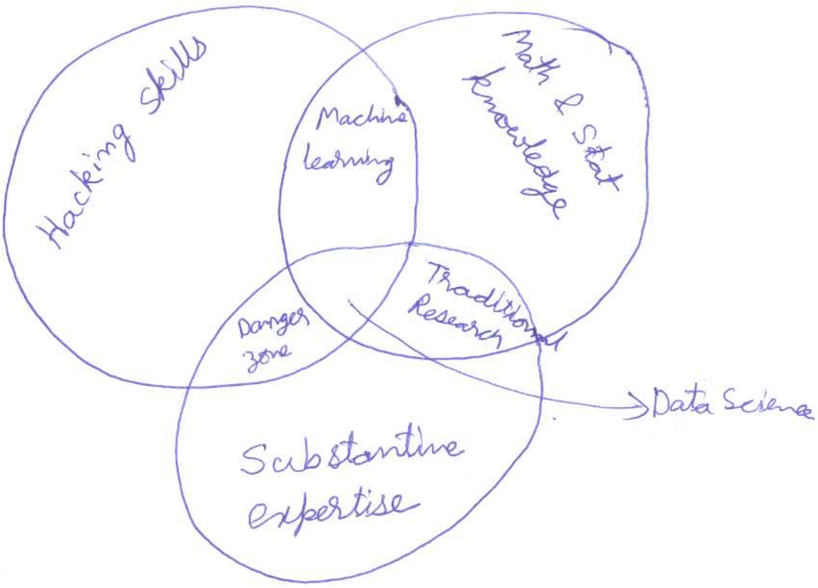
There is lack of clear definitions around the most basic terminology.

Statisticions feel that they are studying and working on the Science of Data.

what data science represents is more of a craft.

Datafication:- is defined as a process of taking all aspects of life and turning them into data.

Drew comway's Venn diagram of data Science



Math & Stat skills:- Significance of Math& stat stems from the fact that it enables you to select the methods for solving issues based on the available facts.

Hacking skills:- coding expertise is required for hacking

A persone with coding ability can apply sophisticated algorithms.

Substantive Expertise:- refers to domain expertise. Knowledge of the topic will facilitate efficient use of data science.

A data Science Profile:- A data scientist should have skills in the following domains.

1) computer Science
2) Math & Statistics
3) Machine Learning
4) Domain expertise
5) communication and presentation skill
6) Data visualization

People working data Science can be divided into four clusters.

Data Business people:- These individuals are focused on the product and profit aspects of data Science. Then Eg :- leaders managers etc

Data creatives:- These people give insights into data.

Data developers:- These data scientists specialize in writing software, managing data infrastructure, scalability.

Data Researchers:- These cluster applies their scientific training and academic tools to organizational data. They generate valuable insights and products.

Unit - I

Chapter 2 - Statistical Inference, Exploratory Data Analysis

Statistical Inference:-

The world we live in is complex, random and uncertain. It is one big data-generating machine.

Statistical inference is the discipline that concerns itself with the development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes.

Population:- A population is the entire group that you want to draw conclusions about. A population refers to the entire set of individuals, objects or data points that you want to study.

Sample:- A sample is a subset of population that is selected for analysis. Sampling allows for inferences about the population using ~~satis~~ statistical techniques A sample is the specific group that you will collect data from.

Bias:- The term bias is used to describe statistics that dont provide an accurate representation of the population. Bias is a statistical term which means a systematic deviation from the actual value.

New kinds of Data:- In olden days we have bunch of numbers and categorical values. Nowadays data includes

1) Traditional data 2) Text 3) Records 4) Geo-based location data

5) Network 6) Sensor data 7) Images etc.

Big Data :- Big data refers to the vast volumes of data generated at high velocity from a variety of sources It is characterized by volume, Variety, Velocity and Value.

Volume - Big data involves large datasets that are too complex for traditional data processing tools to handle.

Velocity :- Big data is generated in real time & near real time requiring fast processing to extract meaningful insights

Variety :- The data comes in multiple forms, including structured data (like database), semi structured data (like XML files) and unstructured data (like text, images and videos).

In the article 'The rise of big data' argue that the Big data revolution consists of three things.

Big data revolution consists of three things.

- → collecting and using a lot of data rather than small samples.
- → Accepting messiness in your data.
- → Giving upon ~~ko~~ knowing the causes.— This is so

Can ~~n~~ N= ALL

An article — election night polls — is in itself a great counter example : even if we poll absolutely everyon who leaves the polling stations, we still dont count people who decided not to vote in the first place. And those might be the very people we'd need to talk to understand our country's voting problems.

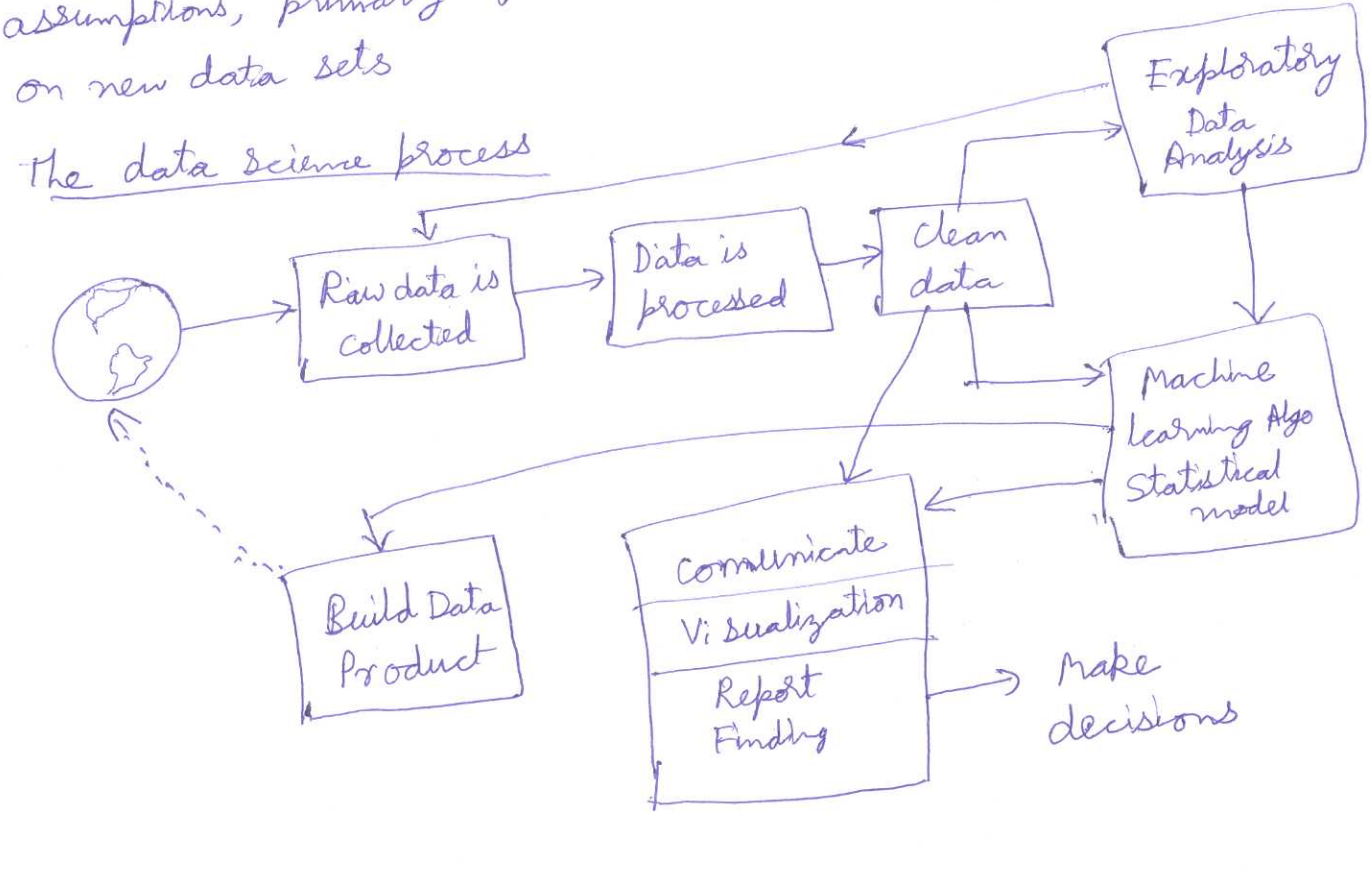
Data is not objective:- In a company women have tended to leave more often, get promoted less often and give more negative feedback on their environments when compared to men. So the model is likely to hire man over the women. Ignoring causation can be a flaw rather than a feature.

What is model? A model is our attempt to understand and represent the nature of reality through a paticular lens. be it architectural, biological or mathematical.

A statistical model is a mathematical representation of data based on explict assumptions about the underlying relationships between variables, often used to test hypotheses and understand the data generating process.

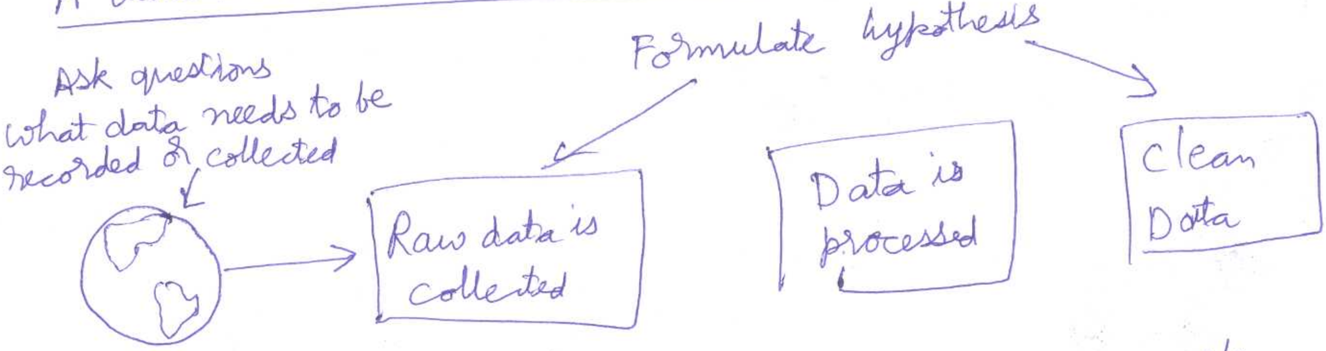
A machine learning algorithm is a computational procedure that learns patters directly from data without requiring explicit assumptions, primarly focused on making accurate predictions on new data sets

The data science process



We got data on real world process. we will start with raw data Eg- logs, employee emails, etc we want to process this to make it clean for analysis.

A data scientist's role in the process



Humans behaviour
Biology
Finance
Internet
Medicine
Sociology
olympics

Email
logs
Medical records
Surveys
Blood drawn
olympic records
- NYT articles

Pipelines
web scraping
cleaning
Munging
Joining
wrangling

clean
outliers
Missing values
Debugging
Table