Unit 1: Introduction to Data Science

Dipesh Koirala

Outline

- Introduction to data science, Applications of data science; Limitations of data science

- Commonly used tools in data science, their strengths and common use-cases: R/RStudio, Python/Pandas/Jupyter Notebooks, Excel/Tableau/PowerBI;

- Data Science life-cycle/Common methodologies for data science: CRISP-DM, OSEMN Framework, TDSP lifecycle;

- Review of statistics and probability: Probability distributions, compound events and independence. Statistics: Centrality measures, variability measures, interpreting variance. Correlation analysis: Correlation coefficients, autocorrelation

History

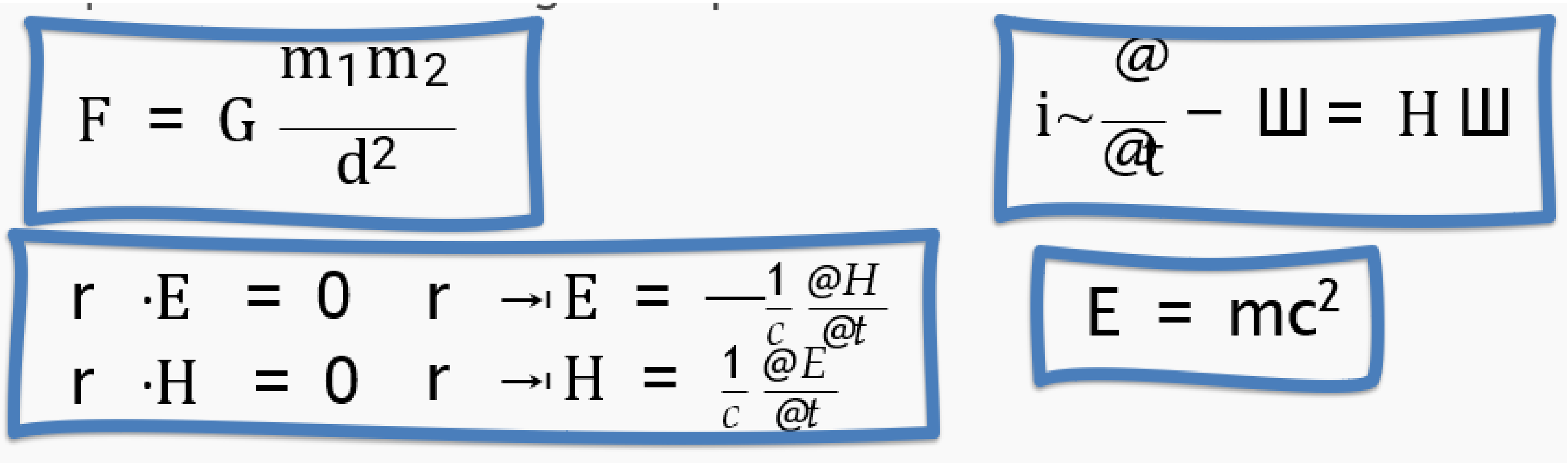
- Long time ago (thousands of years) science was only empirical and people counted stars or crops

History

- Long time ago (thousands of years) science was only empirical and people counted stars or crops and used the data to create machines to describe the phenomena

History

- Few hundred years ago: theoretical approaches, try to derive equations to describe general phenomena.



$$F = G\frac{m_1 m_2}{d^2}$$

$$i\sim\frac{@}{@t} - Ш = H Ш$$

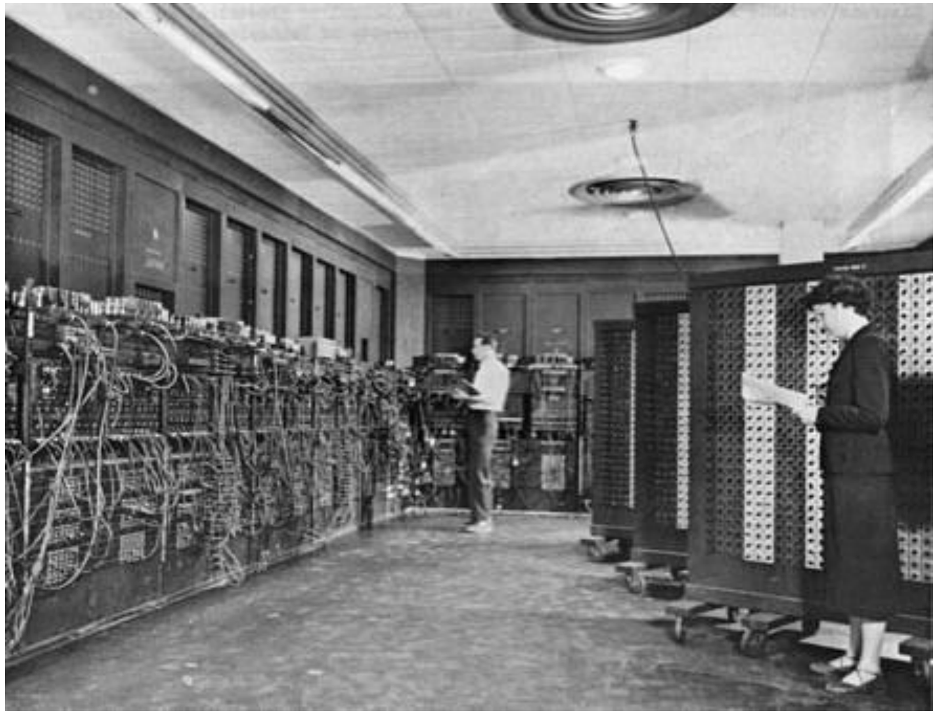
$$r \cdot E = 0 \qquad r \rightarrow E = -\frac{1}{c}\frac{@H}{@t}$$

$$r \cdot H = 0 \qquad r \rightarrow H = \frac{1}{c}\frac{@E}{@t}$$

$$E = mc^2$$

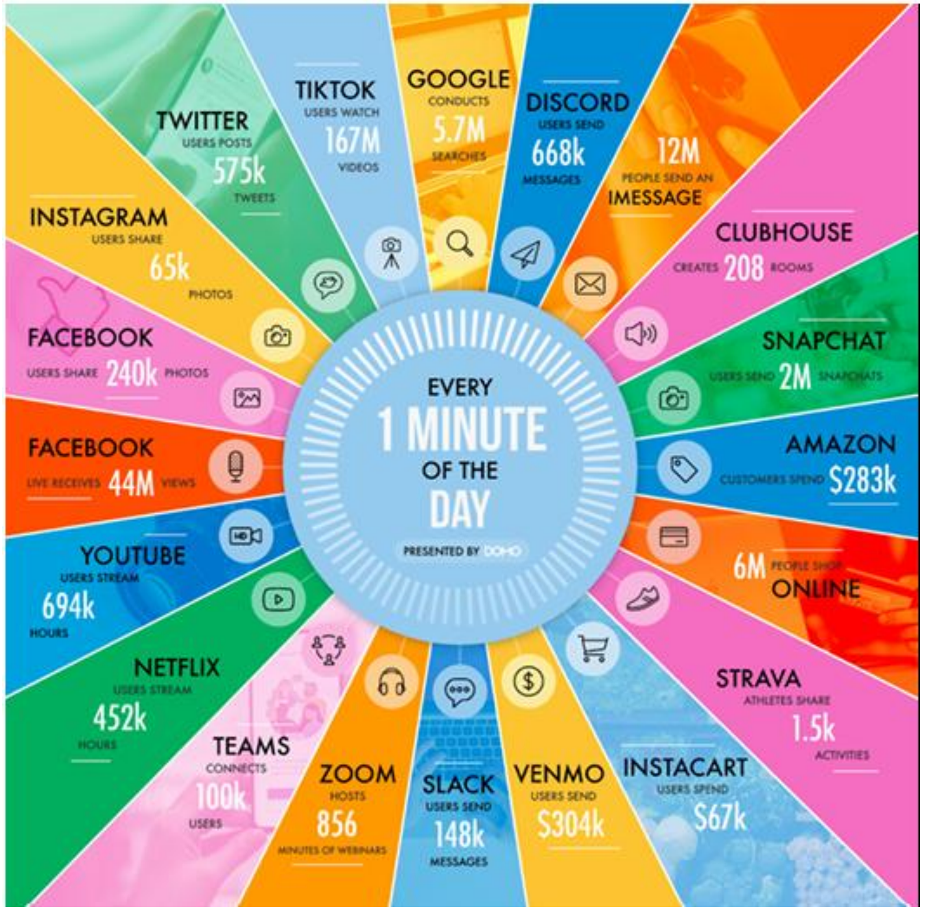
History

- About a hundred years ago: computational approaches appeared



Present

- And the big data began


EVERY
1 MINUTE
OF THE
DAY
PRESENTED BY DOMO
TWITTER USERS POSTS 575k TWEETS
TIKTOK USERS WATCH 167M VIDEOS
GOOGLE CONDUCTS 5.7M SEARCHES
DISCORD USERS SEND 668k MESSAGES
12M PEOPLE SEND AN IMESSAGE
CLUBHOUSE CREATES 208 ROOMS
SNAPCHAT USERS SEND 2M SNAPCHATS
AMAZON CUSTOMERS SPEND $283k
6M PEOPLE SHOP ONLINE
STRAVA ATHLETES SHARE 1.5k ACTIVITIES
INSTACART USERS SPEND $67k
VENMO USERS SEND $304k
SLACK USERS SEND 148k MESSAGES
ZOOM HOSTS 856 MINUTES OF WEBINARS
TEAMS CONNECTS 100k USERS
NETFLIX USERS STREAM 452k HOURS
YOUTUBE USERS STREAM 694k HOURS
FACEBOOK LIVE RECEIVES 44M VIEWS
FACEBOOK USERS SHARE 240k PHOTOS
INSTAGRAM USERS SHARE 65k PHOTOS

Intro

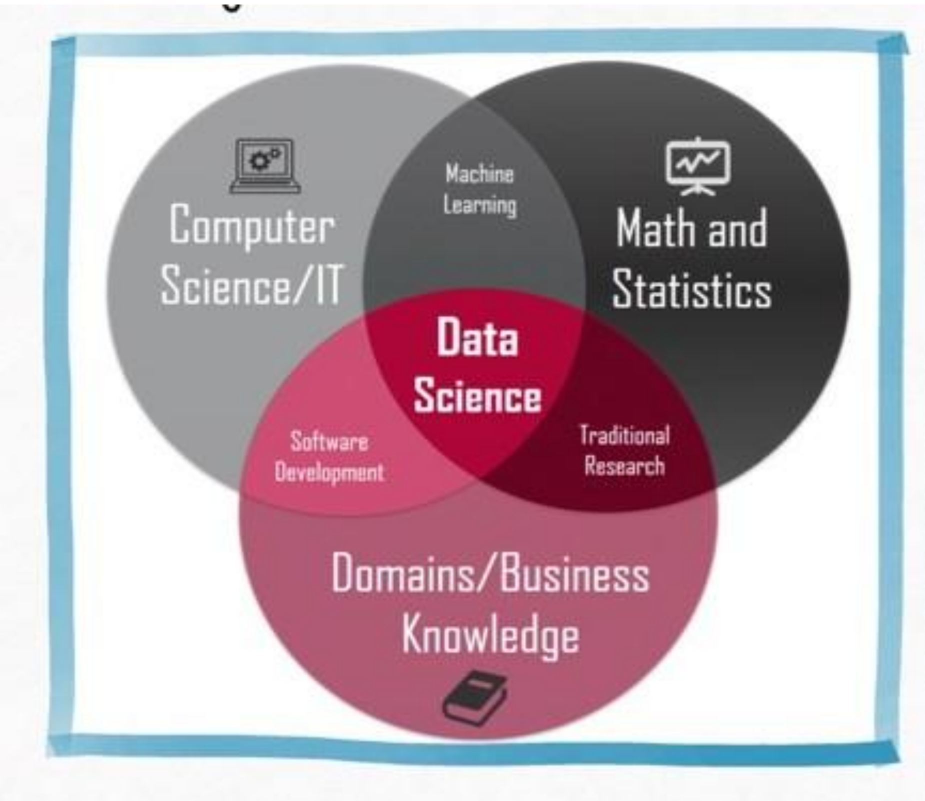
- *Statistics. Math. Computer Science. Physics. Long ago, the four disciplines lived together in harmony. Then, everything changed when the Computer Science attacked(Became dominant).*

- *In this evolving world, a new kind of master emerged: **the data scientist**, skilled in all four disciplines.*

- *With the ability to analyze data, build models, write algorithms, and understand complex systems, the data scientist became essential in solving the world's most pressing problems.*

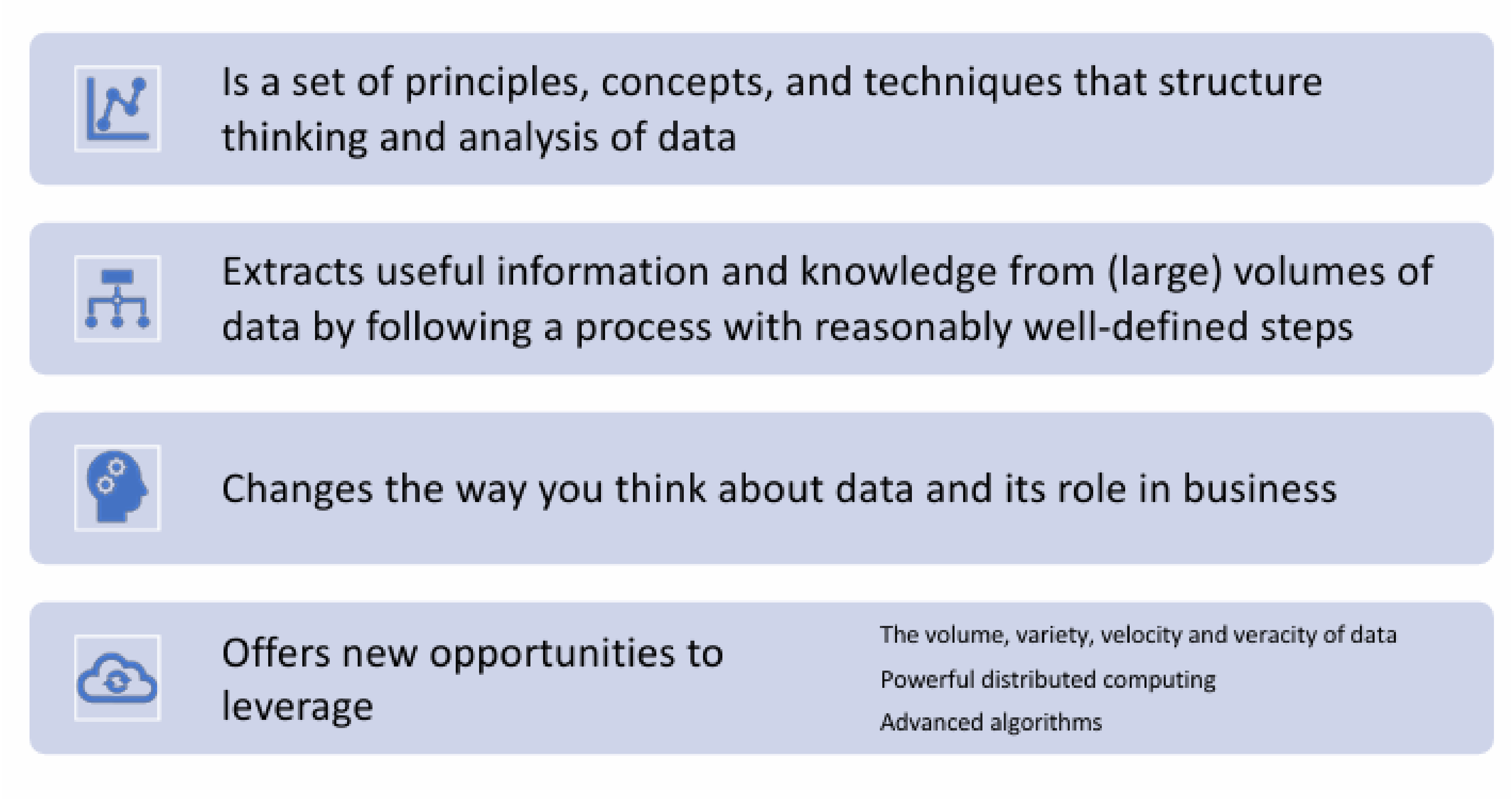
Intro

- Information is what we want, but data are what we've got
- In data science we extract pattern and meaningful insights from data.



- Inter-disciplinary
- Data and task focused
- Resource aware
- Adaptable to changes in the environment and needs

Intro



Is a set of principles, concepts, and techniques that structure thinking and analysis of data

Extracts useful information and knowledge from (large) volumes of data by following a process with reasonably well-defined steps

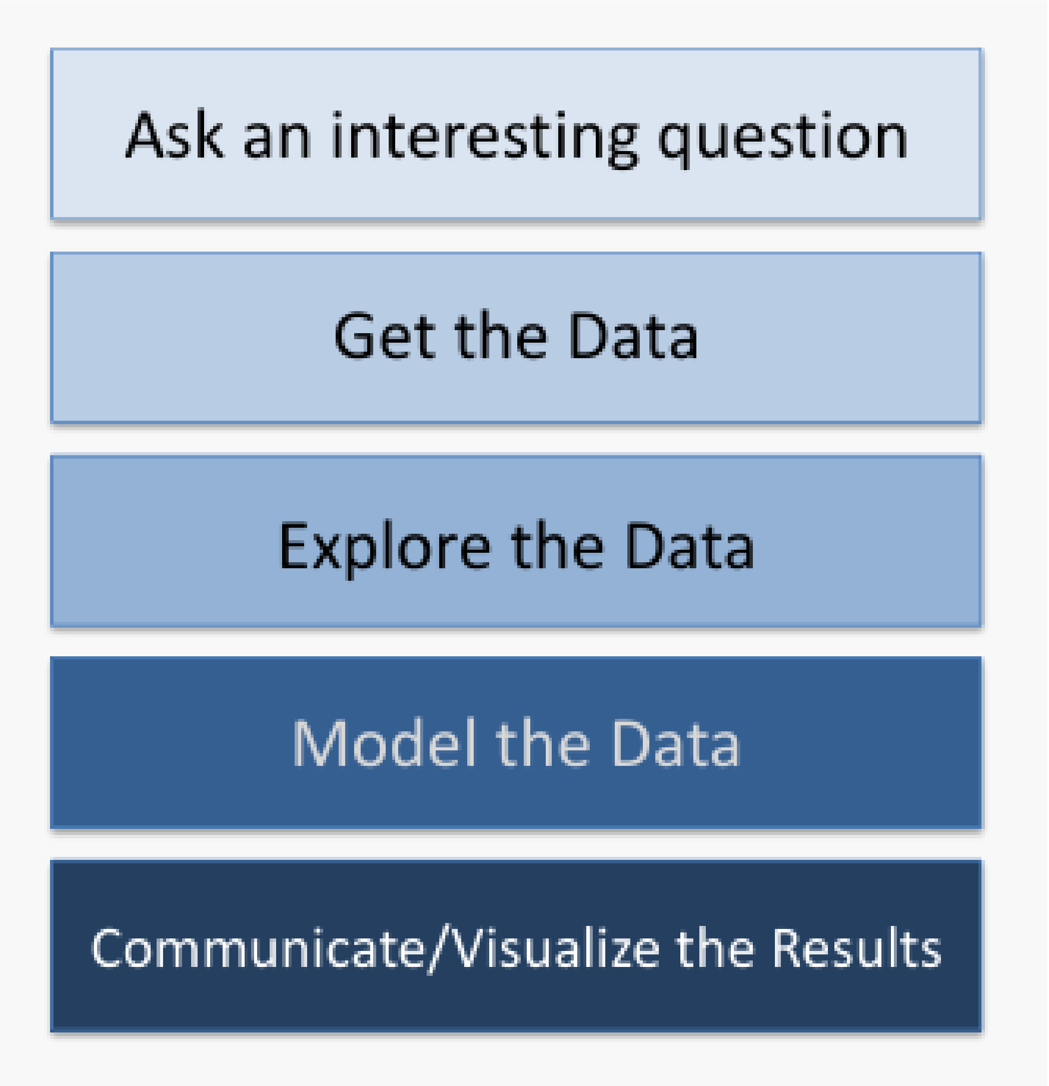
Changes the way you think about data and its role in business

Offers new opportunities to leverage

- The volume, variety, velocity and veracity of data
- Powerful distributed computing
- Advanced algorithms

Intro

The Data Science Process



Ask an interesting question

Get the Data

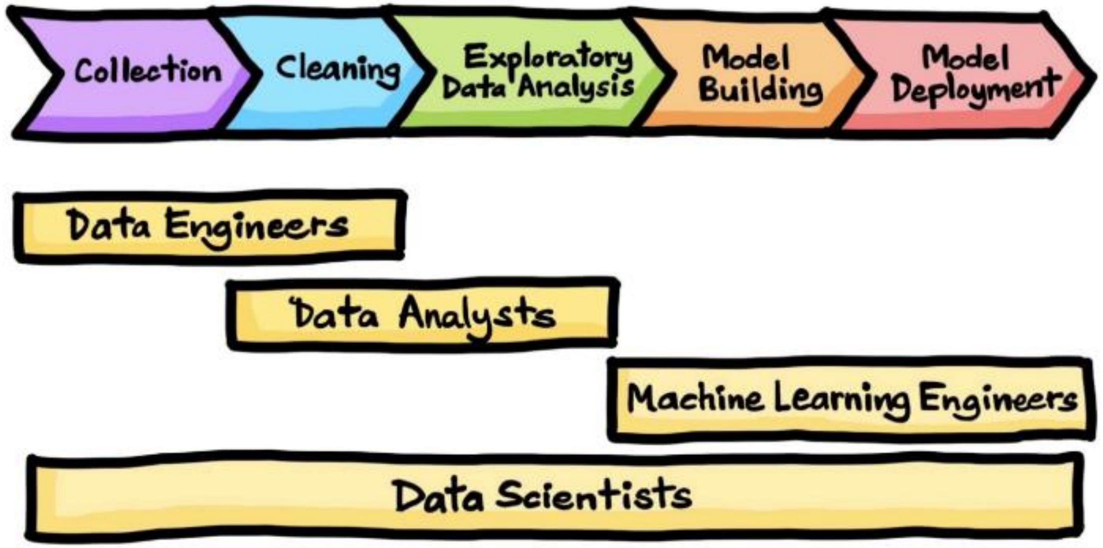
Explore the Data

Model the Data

Communicate/Visualize the Results

Intro

Data Science Process generally follows these steps:



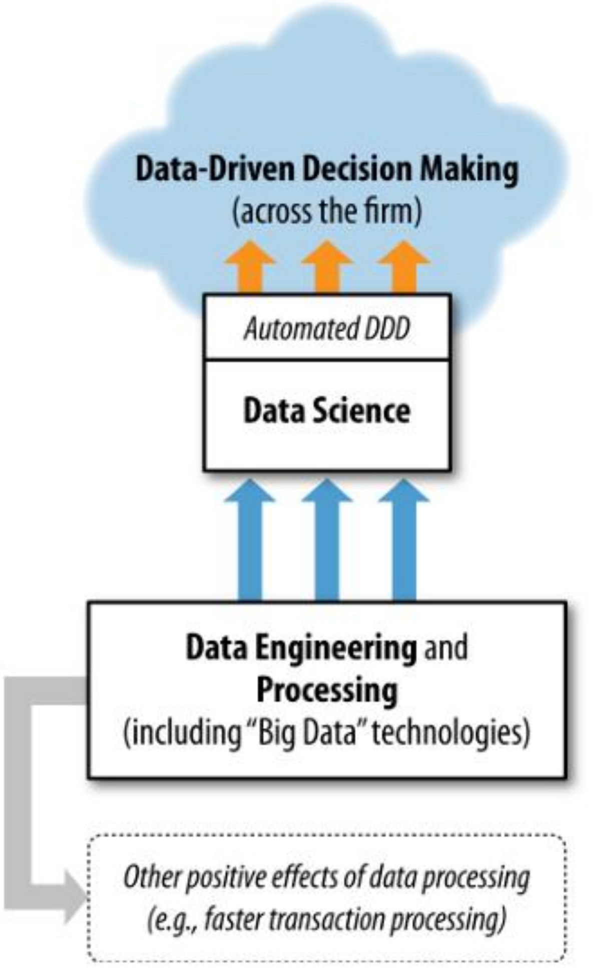
Source: Towards Data Science

Intro

- Data science involves *principles, processes, and techniques for understanding phenomena* via the (automated) analysis of data.

- The ultimate goal of data science as improving decision making, as this generally is of direct interest to business.

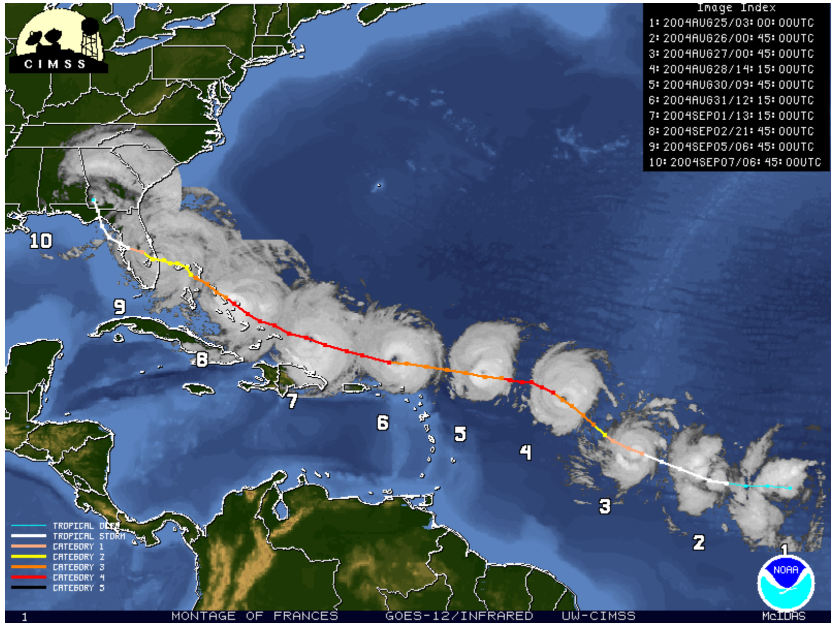
- Data-Driven Decision Making: The practice of making *decisions based on data analysis rather than intuition or observation.* This approach enhances accuracy and effectiveness in business strategies.



Data-Driven Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and Processing
(including "Big Data" technologies)
Other positive effects of data processing
(e.g., faster transaction processing)

Example

- In 2004, as Hurricane Frances approached Florida threatening a direct head
- **Wal-Mart Store decided** that the situation offered a great opportunity for one of their newest data-driven weapons … predictive technology.
- Walmart used predictive analytics to anticipate customer needs and optimize inventory in affected areas.



Example

- It would be more valuable to discover patterns due to the hurricane that were not obvious.
- They didn’t know in the past that strawberry Pop Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane,’

- And the **pre-hurricane top-selling item was beer.**

Example

- In 2012, **Walmart's competitor Target** was in the news for a data-driven decision-making case of its own.
- Like most retailers, Target cares about consumers' shopping habits.
- *Target knew that the arrival of a new baby in a family* is one point where people do change their shopping habits significantly.
- In the Target analyst's words, "*As soon as we get them buying diapers from us, they're going to start buying everything else too*."

Example

- Most retailers know this and *so they compete with each other trying to sell baby-related products to new parents.*

- However, Target wanted to get a jump on their competition.

- They were interested in whether **they could predict that people are expecting a baby.**

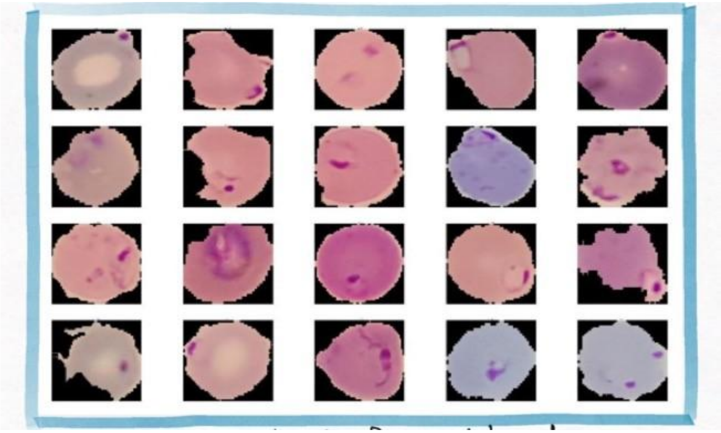
- Target assigned a '**pregnancy prediction score'** to sent customized ads and coupons for baby products

Example

- Using techniques of data science, Target *analyzed historical data on customers who later were revealed to have been pregnant*
- And were able to extract information that could predict which consumers were pregnant.

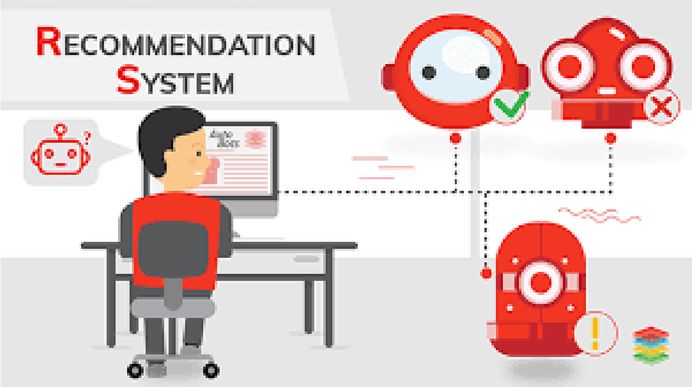
- **E.g.,** pregnant mothers often **change their diets, their wardrobes, their vitamin regimens**, and so on.

Application



Disease Diagnosis

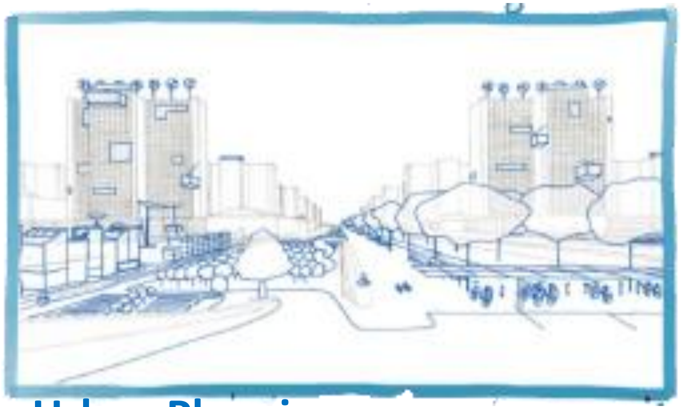
Detecting malaria from blood smears



Recommendation System

Which item next?

Tailor Marketing strategies



Urban Planning

Predicting and planning for resource needs

Agriculture

Precision agriculture

Limitations

- Data Quality and Availability: Incomplete Data, Biased Data
- Cost and Computational Resources: Requires significant computational power
- Ethical and Privacy Concerns: Should follow norms
- Dynamic and Evolving Data: Concept Drift, Real-Time Processing
- Skill Gap: Multidiscipline Knowledge
- Overfitting and Generalization
- Uncertainty and Risk: Model Uncertainty, Risk of Misinterpretation

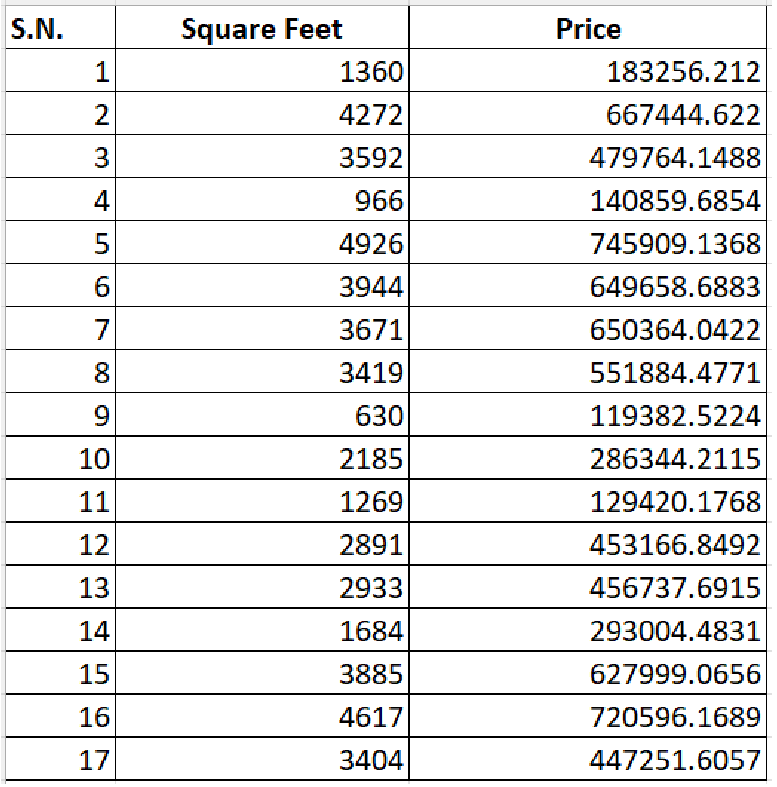
Commonly used tools

- R/Rstudio
- Python/ Pandas/ Jupyter Notebooks
- Excel/ Tableau/ PowerBI

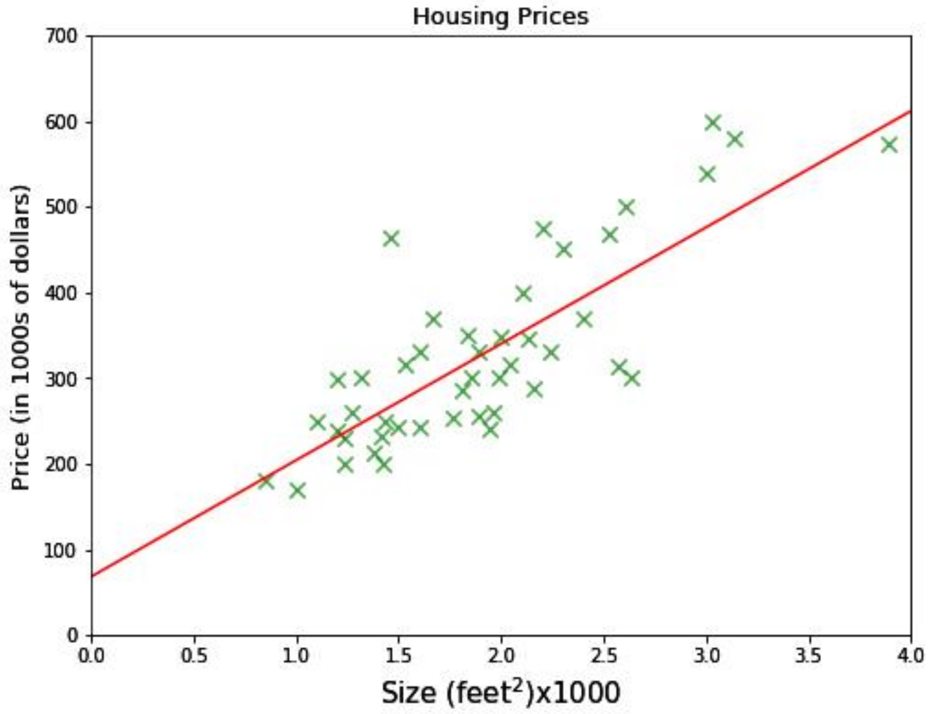
- Programming Language: Python, R
- Data Visualization: Tableau, PowerBi, Python, R
- Machine Learning Libraries: Scikit-learn, Tensorflow, Pytorch
- Data Processing Frameworks: Apache Spark, Hadoop

Example

House Price Prediction

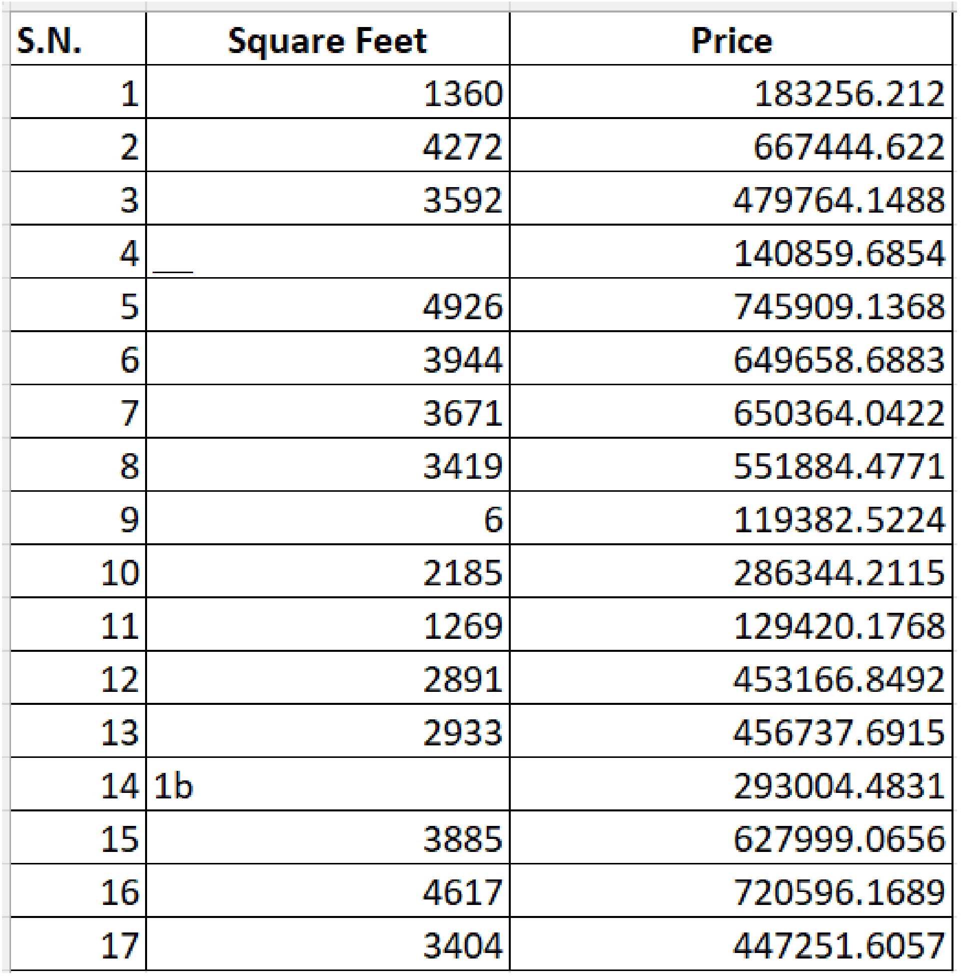


S.N.	Square Feet	Price
1	1360	183256.212
2	4272	667444.622
3	3592	479764.1488
4	966	140859.6854
5	4926	745909.1368
6	3944	649658.6883
7	3671	650364.0422
8	3419	551884.4771
9	630	119382.5224
10	2185	286344.2115
11	1269	129420.1768
12	2891	453166.8492
13	2933	456737.6915
14	1684	293004.4831
15	3885	627999.0656
16	4617	720596.1689
17	3404	447251.6057



Example

House Price Prediction



S.N.	Square Feet	Price
1	1360	183256.212
2	4272	667444.622
3	3592	479764.1488
4	__	140859.6854
5	4926	745909.1368
6	3944	649658.6883
7	3671	650364.0422
8	3419	551884.4771
9	6	119382.5224
10	2185	286344.2115
11	1269	129420.1768
12	2891	453166.8492
13	2933	456737.6915
14	1b	293004.4831
15	3885	627999.0656
16	4617	720596.1689
17	3404	447251.6057

Data Science Life-Cycle

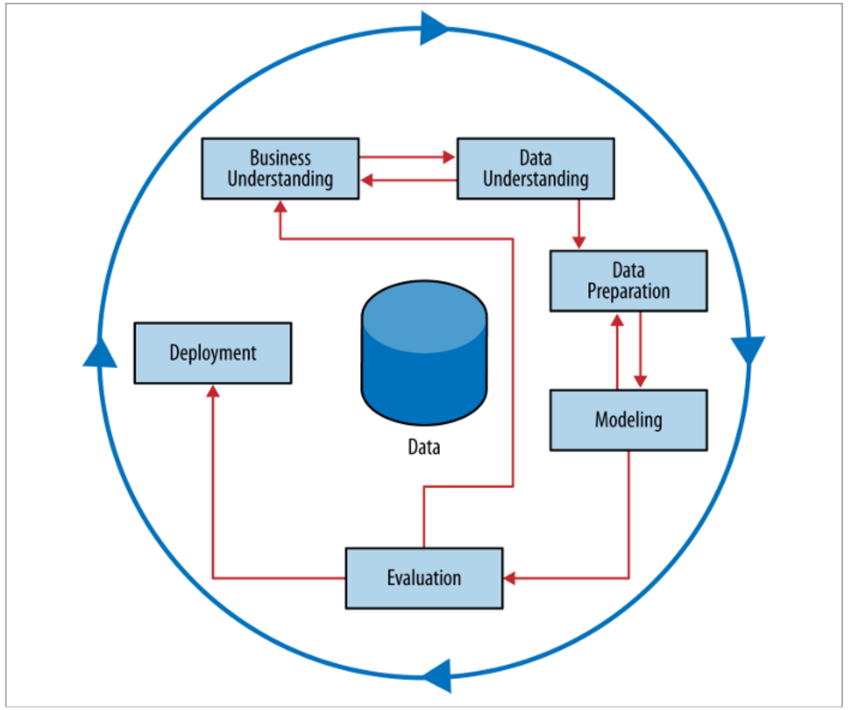
- Different data scientists have different processes for conducting their projects.
- And different types of projects require different steps.

- However, most data science projects flow through a similar workflow.

- *The representation of this workflow is lifecycle.*
  i. CRISP-DM
  ii. OSEMN
  iii. TDSP

Data Science Life-Cycle (CRISP-DM)

- CRISP – DM was published in 1999 to standardize data mining processes across industries, then it has since *become the most common methodology for data mining, analytics, and data science projects.*
- The CRoss Industry Standard Process for Data Mining (**CRISP-DM**) is a process model with six phases that naturally describes the data science life cycle.



1. Business Understanding:

- The *Business Understanding* phase focuses on understanding the objectives and requirements of the project.

Data Science Life-Cycle (CRISP-DM)

2. Data Understanding:

- Drives the focus to *identify, collect, and analyze the data* sets that can help you accomplish the project goals.
- Phases: Collect initial data, Describe data, Explore data, Verify data quality

3. Data Preparation:

- Make sure the fuel (data) is ready to be used in your model.
- Phases: Select data, Clean data, Construct data, Integrate data, Format data

4. Modeling

- Build and evaluate different models based on *several various modelling techniques*.
- Phases: Select modeling techniques, Generate test design, Build model, Assess model

Data Science Life-Cycle (CRISP-DM)

5. Evaluation

- Evaluation phase looks more broadly at which model best meets the business and what to do next.
- Phases: Evaluate results, Review process, Determine next steps

6. Deployment

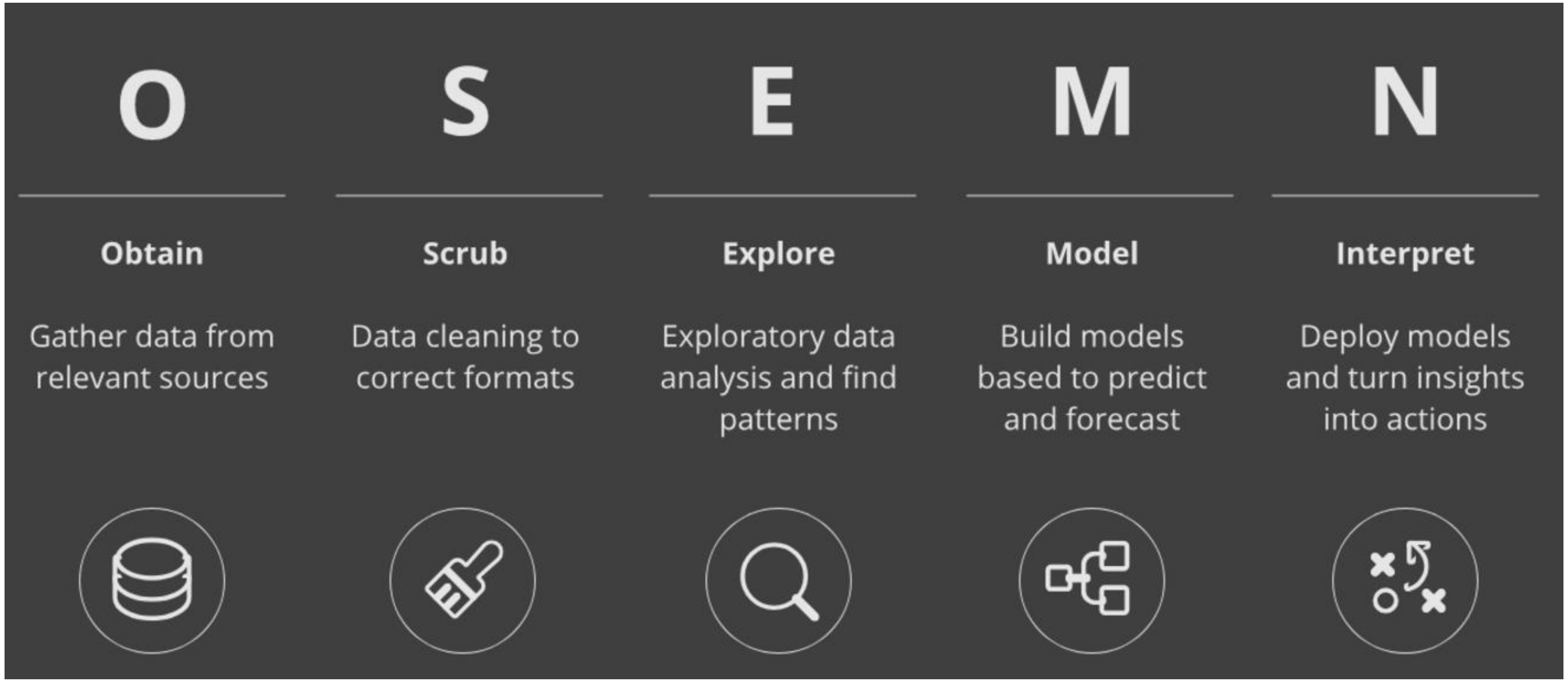
- A *model is not particularly useful unless the customer can access its results.*
- Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Sources:

- ➢ What is CRISP DM? - Data Science Process Alliance
- ➢ Data Science Life Cycle: CRISP-DM and OSEMN frameworks

Data Science Life-Cycle (OSEMN)

- In a 2010 post called "*A Taxonomy of Data Science*" on the dataists blog, Hilary Mason and Chris Wiggins *introduced the OSEMN framework.*
- That basically included categorizing the general workflow that a data scientists typically perform.



O

Obtain

Gather data from relevant sources

S

Scrub

Data cleaning to correct formats

E

Explore

Exploratory data analysis and find patterns

M

Model

Build models based to predict and forecast

N

Interpret

Deploy models and turn insights into actions

Data Science Life-Cycle (OSEMN)

1. Obtain – Data Collection

- First step *is that to obtain the data needed* from available data sources.
- Remember – “Garbage in, Garbage Out”

2. Scrub – Data Cleaning

- Data cleaning is the process of *fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data* within a dataset.
- Although this is often the most time-consuming (and for most the boring) task, it’s a crucial step
-

3. Explore – Data Analysis

- This phase *allows to understand the data* to figure out the course of actions and areas that can be explored in the modeling phase

Data Science Life-Cycle (OSEMN)

- 4. Model – Modelling Data
  - In the fourth step, machine learning techniques is used to help make sense of data and *acquire important insights for data-driven decision-making*.
  - “where the magic happens”.

- 5. Interpret – Model Deployment
  - In the final step, try to make sense of the data by simplifying and summarizing results from *all the models built and by communicating findings.*

Sources:

- OSEMN Data Science Life Cycle - Data Science Process Alliance
- Data Science Life Cycle: CRISP-DM and OSEMN frameworks

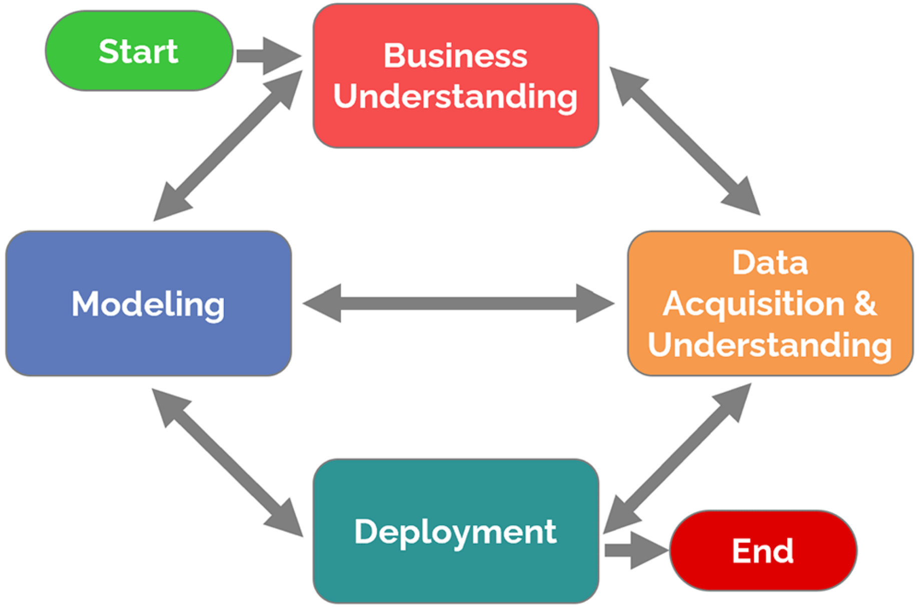
Data Science Life-Cycle – (TDSP)

- Team Data Science Process is launched in 2016 by Microsoft.

- TDSP is "*an agile, iterative data science methodology* to deliver predictive analytics solutions and intelligent applications efficiently." (Microsoft, 2020).

- This is a modern data science process that combines elements of the *core data science life cycle, software engineering, and Agile processes.*

- *Developed to standardize and improve the efficiency of data Science projects*



Data Science Life-Cycle – (TDSP)

- TDSP's project lifecycle includes five iterative stages:

1. **Business Understanding:** define objectives and identify data sources

2. **Data Acquisition and Understanding:** acquire data and determine if it can answer the presenting question

3. **Modeling:** working on data, features and *model training* (combines *Modeling* and *Evaluation*)

4. **Deployment:** deploy into a production environment

5. **Customer Acceptance:** customer validation *if the system meets business needs* (a phase not explicitly covered by CRISP-DM)

Data Science Life-Cycle – (TDSP)

- TDSP addresses the weakness of CRISP-DM’s lack of team definition by defining six roles:
    - Solution architect
    - Project manager
    - Data engineer
    - Data scientist
    - Application developer
    - Project lead

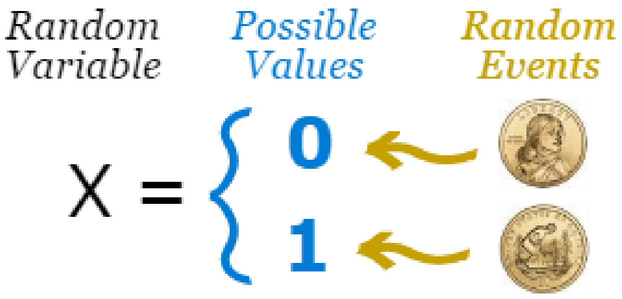
- It provides a structured, iterative approach to building and deploying data science solutions, emphasizing collaboration among team members and stakeholders.

Statistics and Probability

- **Random variables** are functions that maps outcomes of a random experiment to real numbers.
- In probability theory and statistics, a random variable (or stochastic variable) is a way of assigning a value (often a real number) to each possible outcome of a random event
- For example:

There are two possible outcomes for a coin toss: heads, or tails.

The possible outcomes for one fair coin toss can be described using the following random variable:



X: Random Variable

x: a specific value the random variable can take

P(X = x): Probability that X takes the value x

Random Variables

- Tossing 2 coins simultaneously

  Sample Space = {HH, HT, TH, TT}

- Let the random variable be getting number of heads then

  X = {0, 1, 2}

Types

1. Discrete Random Variables: A Random Variable X is said to be discrete if it takes only the values of the set {0, 1, 2, ….. N}

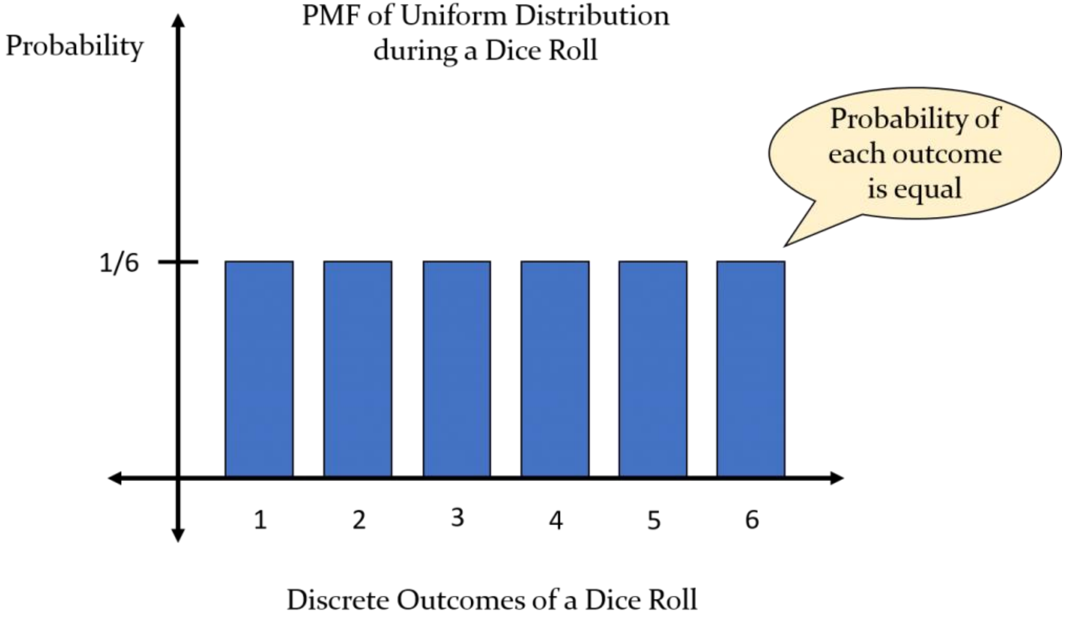
   E.g., Tossing a coin, throwing a dice, number of people born in last year

2. Continuous Random Variables: A Random Variable X which takes all possible values in a given interval of domain

   E.g., Height of an individual, amount of rainfall in rainy seasons

Probability Distribution

- A probability distribution describes *how probabilities are distributed over the values of a random variable.*
- It gives the probability of each outcome of a random experiment or event



Probability Distribution

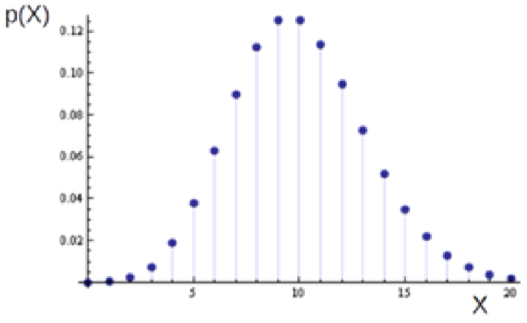
1. Discrete Probability Distribution:

- Let x is a Discrete Random Variable with possible outcomes $x_1, x_2, \ldots x_n$ having probabilities $p(x\text{i})\ for\ \ i = 1{,}2\ \ldots\ n$. Then the function p(xi) is called Probability Mass function of a X.

- Probability distribution is described by Probability Mass Function (PMF)
- {xi, p(xi)} is called Discrete Probability Distribution:

$$\begin{aligned} p(x) &\geq 0 \\ p(x) &\leq 1 \\ \sum_x p(x) &= 1 \end{aligned}$$

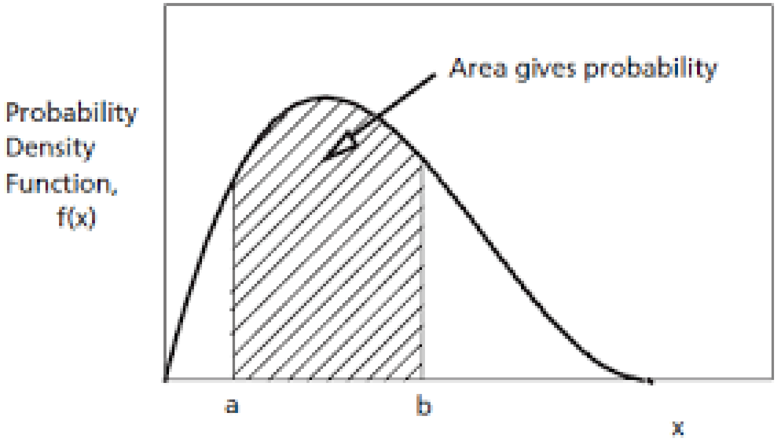
$$\rho_X(x) = \begin{cases} \frac{1}{2}, & \text{if } x = \text{head}, \\ \frac{1}{2}, & \text{if } x = \text{tail}. \end{cases}$$



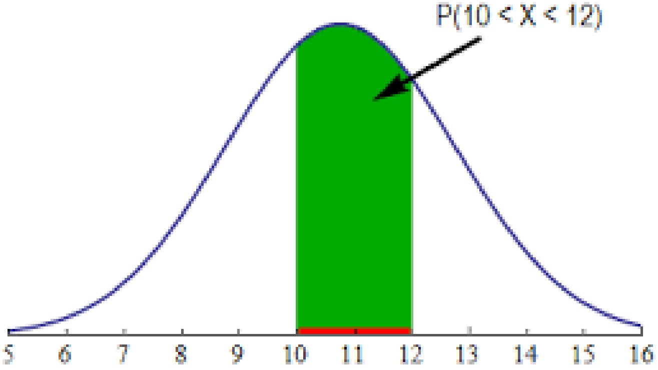
Probability Distribution

2. Continuous Probability Distribution:

- Let X be a continuous random variable taking values on the interval (a, b).
- Then f(x) is the **probability density function (pdf)** which gives the relative likelihood of the variable taking on a specific value.
- The area under the PDF curve over an interval represents the probability of the variable falling within that interval.



$$F(x) = P(a \le x \le b) = \int_a^b f(x)dx \ge 0$$



Probability Distribution

1. Discrete Probability Distribution

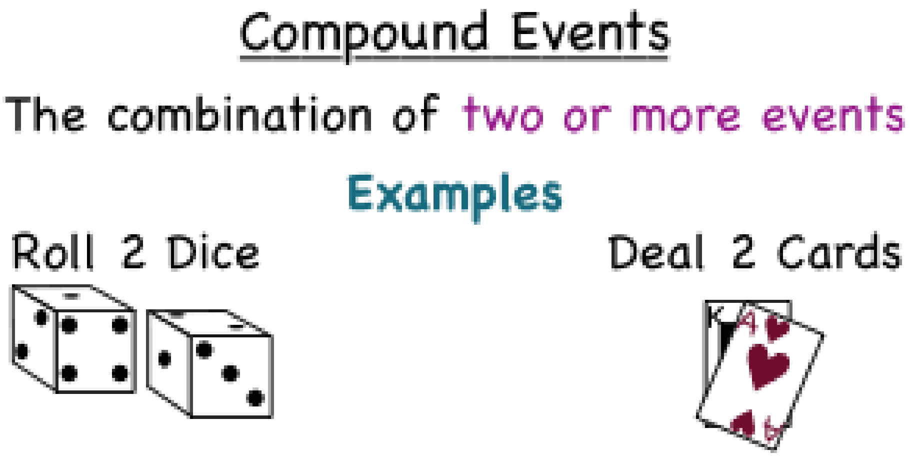
- Defined for discrete random variables
  - Binomial Distribution
  - Poisson Distribution

2. Continuous Probability Distribution

- Defined for continuous random variables
  - Normal Distribution: Symmetric, bell-shaped distribution.
  - Uniform Distribution: All outcomes are equally likely

Compound Events and Independence

- A "compound event" refers to a situation where two or more events occur simultaneously.



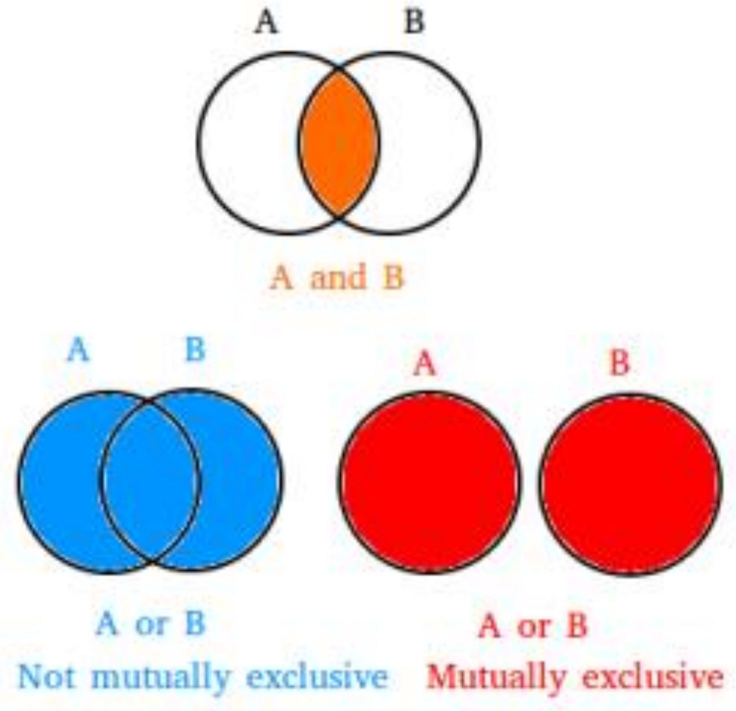
Compound Events

The combination of two or more events

Examples

Roll 2 Dice

Deal 2 Cards



Compound Events

Compound Events and Independence

Compound Events

1. Union of Events (OR Probability)

- Probability that either event A or event B (or both) occurs. (at least one occurs)

Formula For Mutually Exclusive: $P(A \cup B) = P(A) + P(B)$

For non-mutually Exclusive: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2. Intersection of Events (AND Probability)

- Probability that both events A and B occur.

Formula $P(A \cap B) = P(A) \times P(B)$

$P(A \cap B) = P(A) \times P(B|A)$

Compound Events and Independence

- Probability under conditions of Statistical Independence

  1. Marginal: Occurrence of Single Event
  2. Joint: $P(AB) = P(A) \times P(B)$
  3. Conditional Probabilities: $P(A|B) = P(A)$

Statistics: Centrality Measures

Central Tendency:

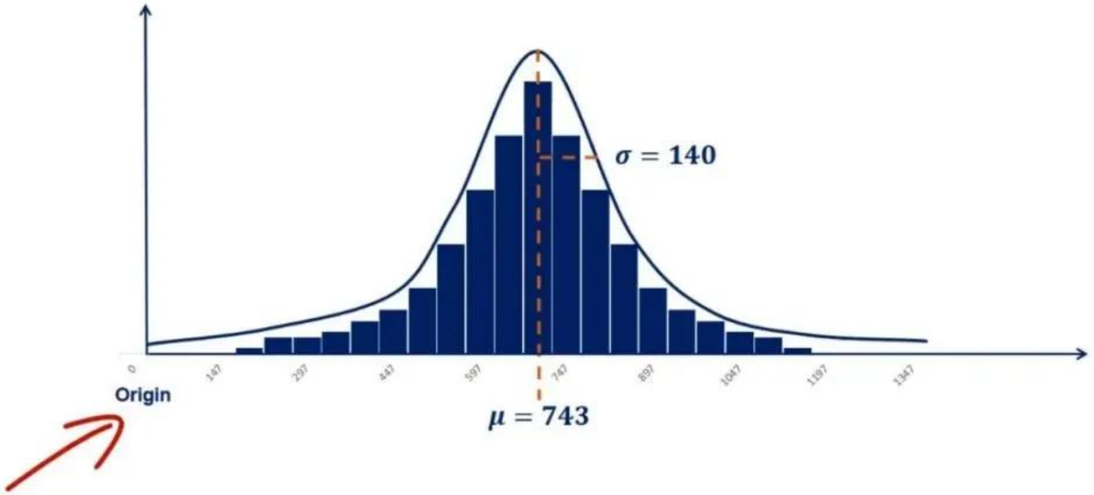
- Identify one value which can be used as being representative of a whole set of data.
- It is the virtue that large amount of quantitative values-move toward center.
- Gives middle point of the distribution
- Allows comparison between two or more groups

- Mean
- Median
- Mode

Statistics: Centrality Measures

Mean

- Average values of the distribution
- Based on all observations
- Affected by extreme values
- Not suggested for qualitative data
- Not suggested for open-ended class



Statistics: Centrality Measures

Median

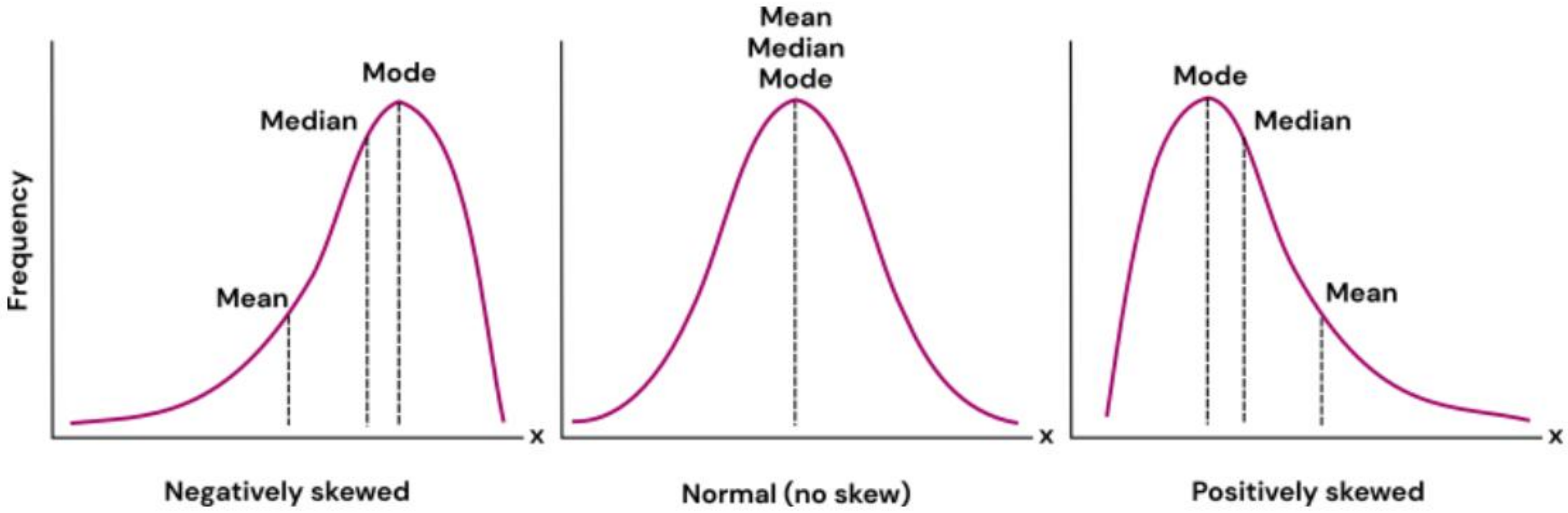
- Middlemost or most central item
- Median is the value which divides the distribution into two equal parts.
- Can be used in open ended class
- Is not affected by extreme values
- Can be used in ordinal qualitative data: extremely sharp, very sharp, sharp, slightly blurred, very blurred

Mode

- Mode is the value, whose repetition is maximum i.e. frequency is maximum.
- It can be used for qualitative as well as quantitative data.

Statistics: Centrality Measures

Mean , Median, Mode



Statistics: Variability Measures

- **Dispersion** is the spread of the data in a distribution, that is, the extent to which the observations are scattered.
- To increase our understanding of the pattern of the data, we must also measure its dispersion
- If the data is widely spread the central location is less representative of the data

1. Range
2. Interquartile Range: $Q_3 - Q_1$
3. Standard deviation
4. Variance

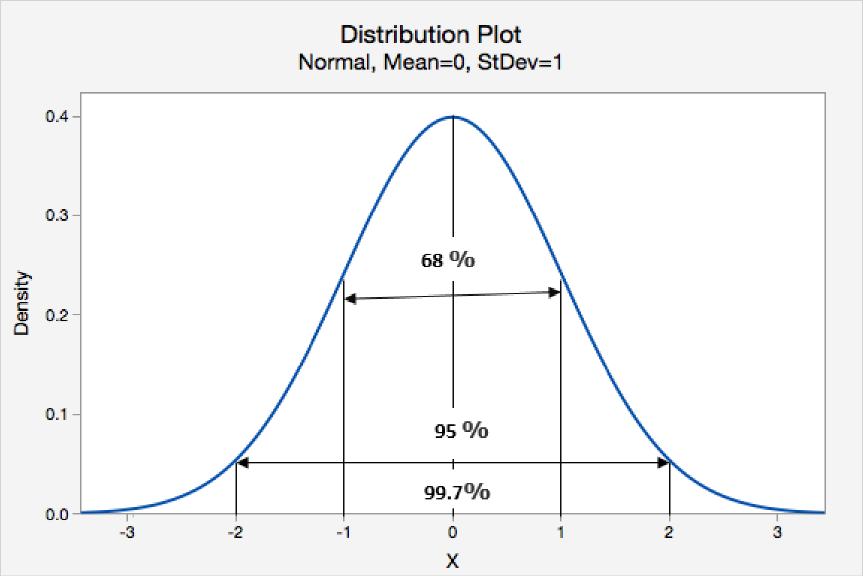
Statistics: Variability Measures

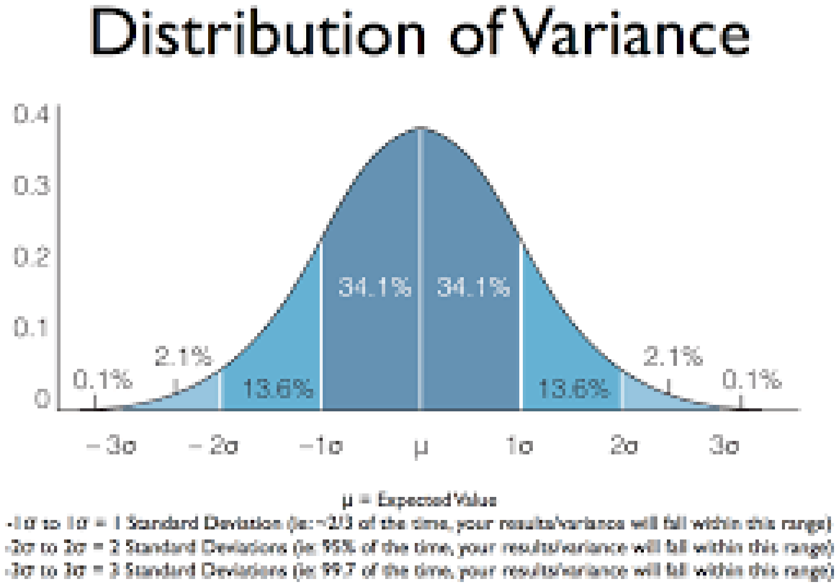
- **Variance is** a measure of how far a set of *data are dispersed out from their mean* or average value.
- It is denoted as ‘σ^2’.
- Variance is the average of squared deviations

- Define **standard deviation** as the “spread of the statistical data from the mean or average position”.
- It is denoted as ‘σ’.
- In test scores, if **mean = 75** and **σ = 10**, most students score between **65 and 85** (1σ range).

Statistics: Variability Measures

Variance and Standard Deviation


Distribution Plot
Normal, Mean=0, StDev=1


Distribution of Variance

μ = Expected Value
-1σ to 1σ = 1 Standard Deviation (ie:~2/3 of the time, your results/variance will fall within this range)
-2σ to 2σ = 2 Standard Deviations (ie: 95% of the time, your results/variance will fall within this range)
-3σ to 3σ = 3 Standard Deviations (ie: 99.7 of the time, your results/variance will fall within this range)

Correlation Coefficient

- Is a tool to determine the degree of associations between variables.
- The measure of relation between two or more variables is correlation

- If one is interested to know the relation between two variables, statistical measures calculated separately is not sufficient to characterize these pairs of data.
- **E.g.,** Height vs weight, income vs expenditure

Methods of Studying Correlation

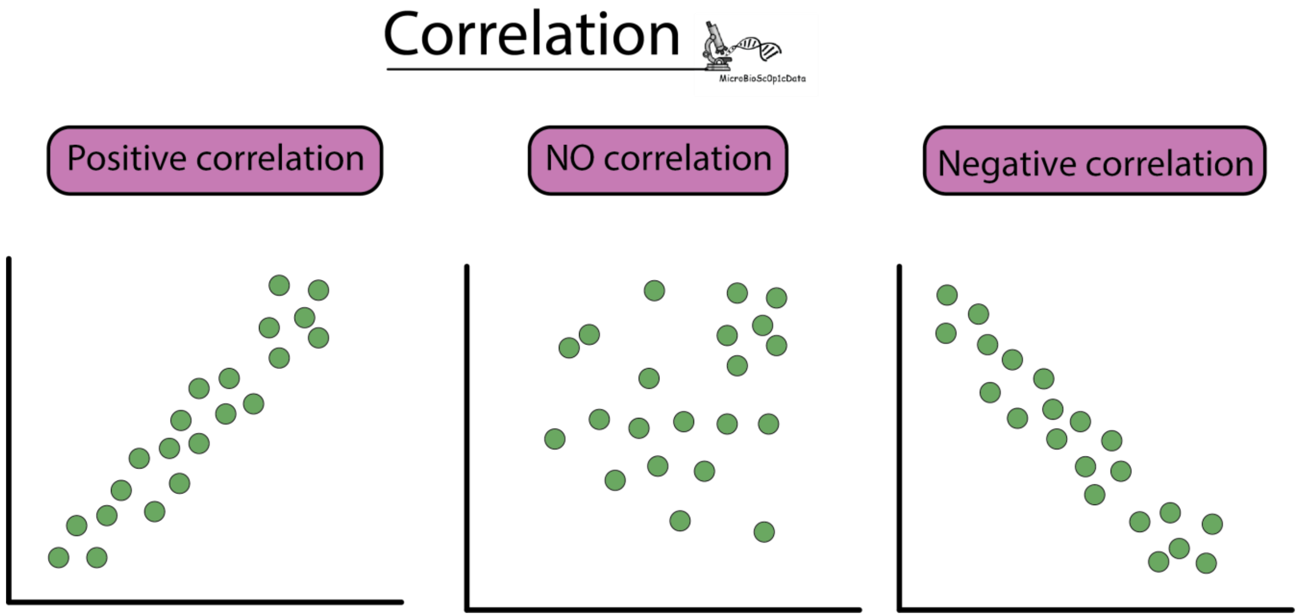
- Scatter Plot method
- Calculation

Correlation Coefficient

- It value ranges from -1 to 1

Correlation can be:

- Positive Correlation
- Negative correlation
- No correlation



Autocorrelation

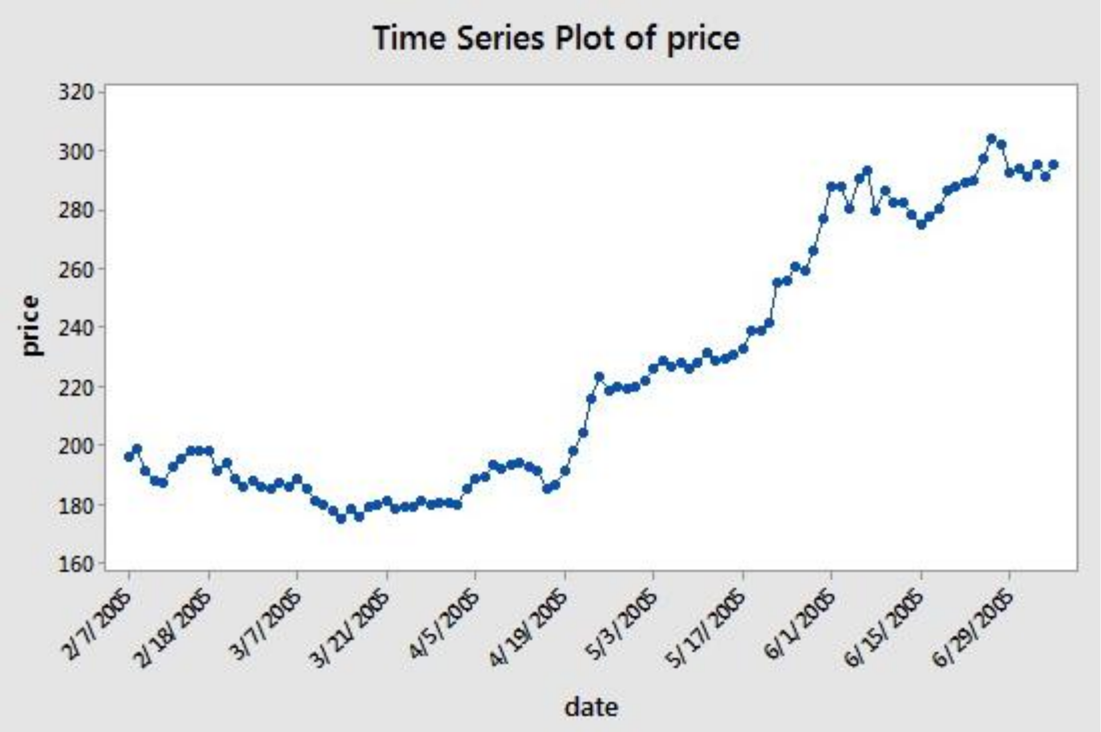
- Autocorrelation/Serial correlation measures the *degree of similarity between a given time series and a lagged version of itself* over successive time intervals.
- It is a tool used to identify repeating **patterns, trends, or dependencies** in data over time.

- A **time series** is a sequence of measurements of the same variable(s) made over time.
- Used for time series forecasting, to analyze economic data, such as GDP, inflation

Autocorrelation

- Autoregressive Model

$$y_t = \beta_0 + \beta_1 \times y_{t-1} + \epsilon_t$$



Types:

- Positive Autocorrelation
- Negative autocorrelation
- No autocorrelation

Note:

- A lag refers to the time difference between the current observation and a past observation. E.g., if today's temperature is compared with yesterday's temperature, the lag is 1.

End of Unit 1

Dipesh Koirala