LIVEWIRE
FOR LIVE CAREERS

DATA SCIENCE USING PYTHON

Use of technological devices

=

Generation of such a huge amount of data.

LIVEWIRE
FOR LIVE CAREERS

Session_1

LIVEWIRE
FOR LIVE CAREERS

INTRODUCTION TO DATA SCIENCE

DATA GENERATION

- 2.5 quintillion bytes of data generated per year


Job Trends from Indeed.com
"data scientist"
Percentage of Matching Job Postings
0.02
0.015
0.01
0.005
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11
Jan '12
Jan '13
indeed
one search. all jobs.

No matter how extremely unpleasant your algorithm is, they can often be beaten simply by having more data (and a less sophisticated algorithm).

LIVEWIRE
FOR LIVE CAREERS

How to analyse and collect these Data

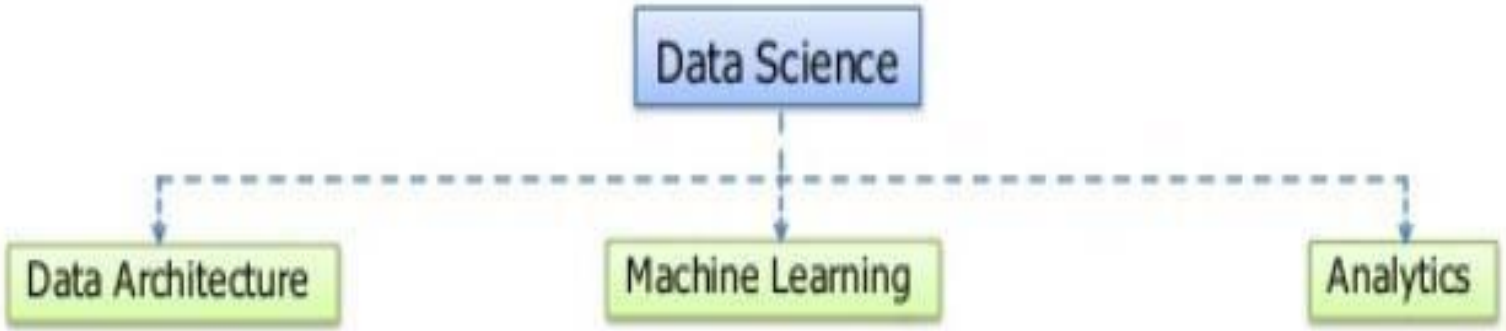
LIVEWIRE
FOR LIVE CAREERS

Data Science

Data science – Discovery of data insight

LIVEWIRE
FOR LIVE CAREERS

Core Components



LIVEWIRE
FOR LIVE CAREERS

Google's search engine is a product of data science

Data Science Job

data science jobs
data science jobs in bangalore
data science jobs in canada
data science jobs in bangalore for freshers
data science jobs in india
data science job description
data science jobs for freshers
data science jobs in dubai
data science jobs in germany
data science jobs in australia

Google Search I'm Feeling Lucky


SALE
10%
BUY
BIG
SALE
BUY
BIG
SALE

https://www.amazon.in/gp/product/B019PIOJYU/ref=s9_acsd_top_hd_bw_b1RCr7P_c_x_w?pf_rd_m=A1VBAL9TL5WCBF&pf_rd_s=merchandised-search-4&pf_rd_r=H76R...

J. K. Rowling + Follow

Get new release updates & improved recommendations

‹ Read more

Length: 345 pages Word Wise: Enabled Enhanced Typesetting: Enabled

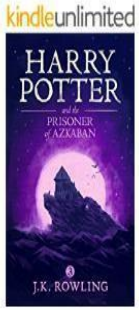
Page Flip: Enabled Language: English

Similar books to Harry Potter and the Philosopher's Stone

Customers who bought this item also bought

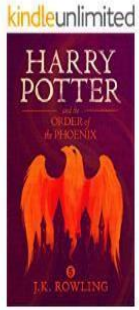
Page 1 of 9

Harry Potter and the Chamber of Secrets
› J.K. Rowling
★★★★½ 536
Kindle Edition
₹ 159.00



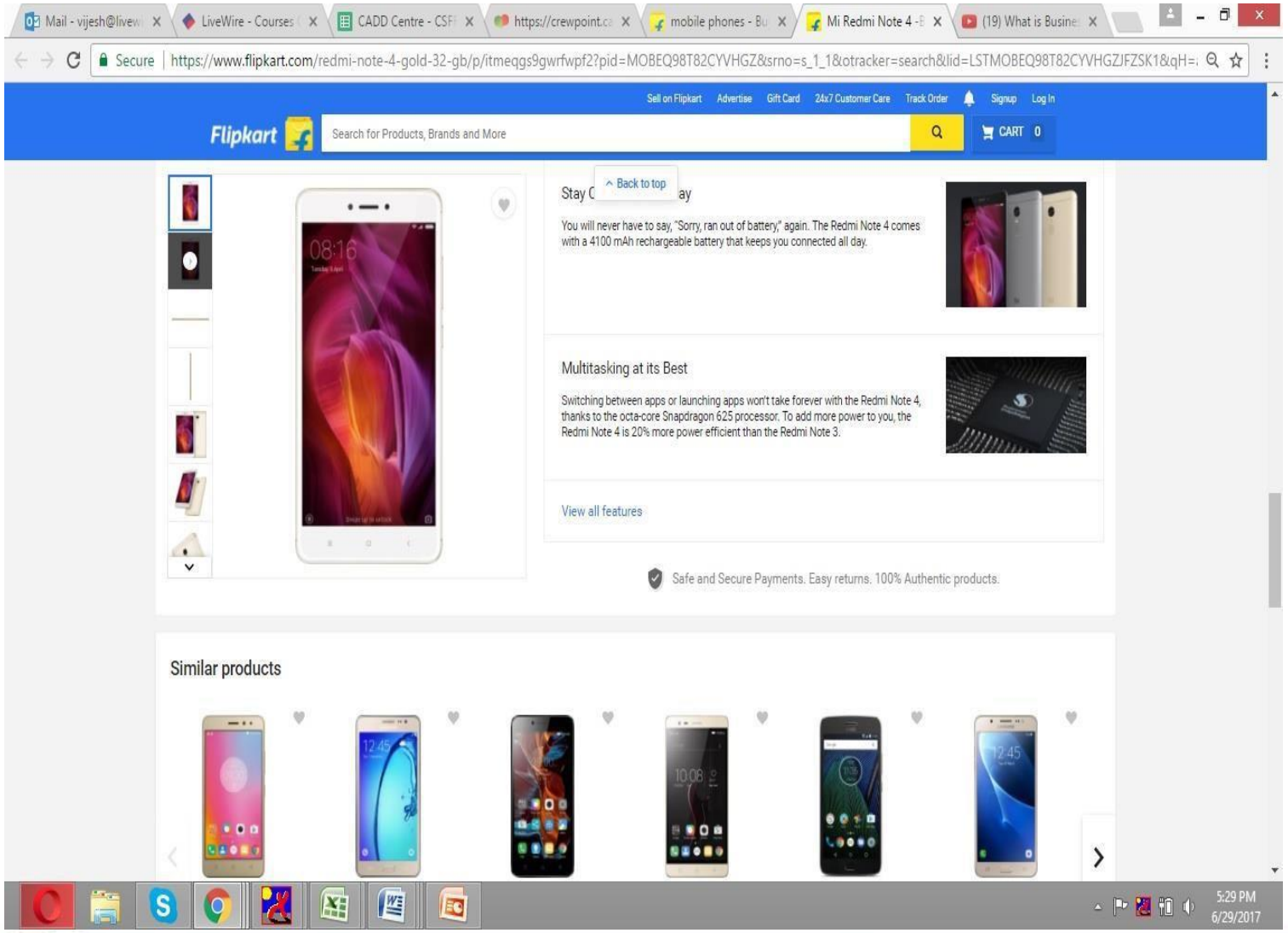
Harry Potter and the Prisoner of Azkaban
› J.K. Rowling
★★★★½ 458
Kindle Edition
₹ 159.00

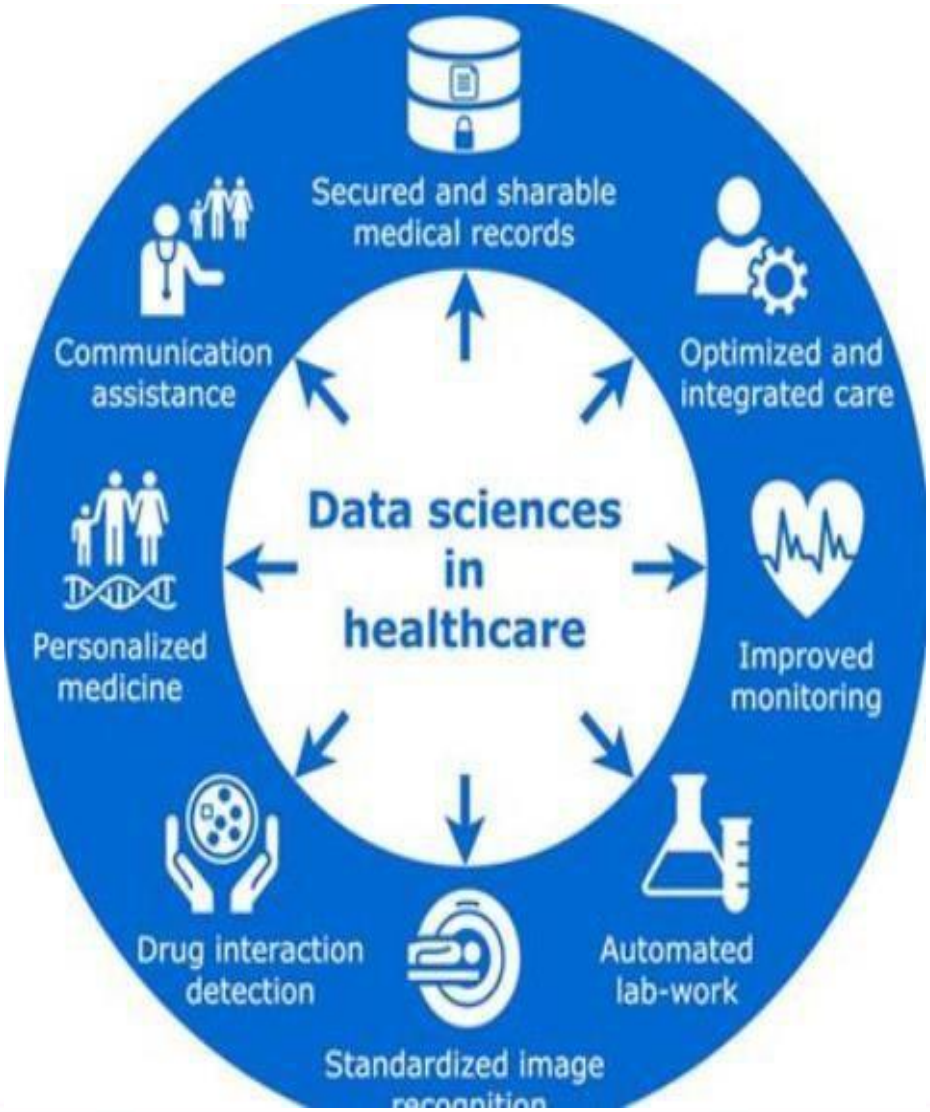
Harry Potter and the Goblet of Fire
› J.K. Rowling
★★★★½ 313
Kindle Edition
₹ 159.00



Harry Potter and the Order of the Phoenix
› J.K. Rowling
★★★★½ 295
Kindle Edition
₹ 159.00

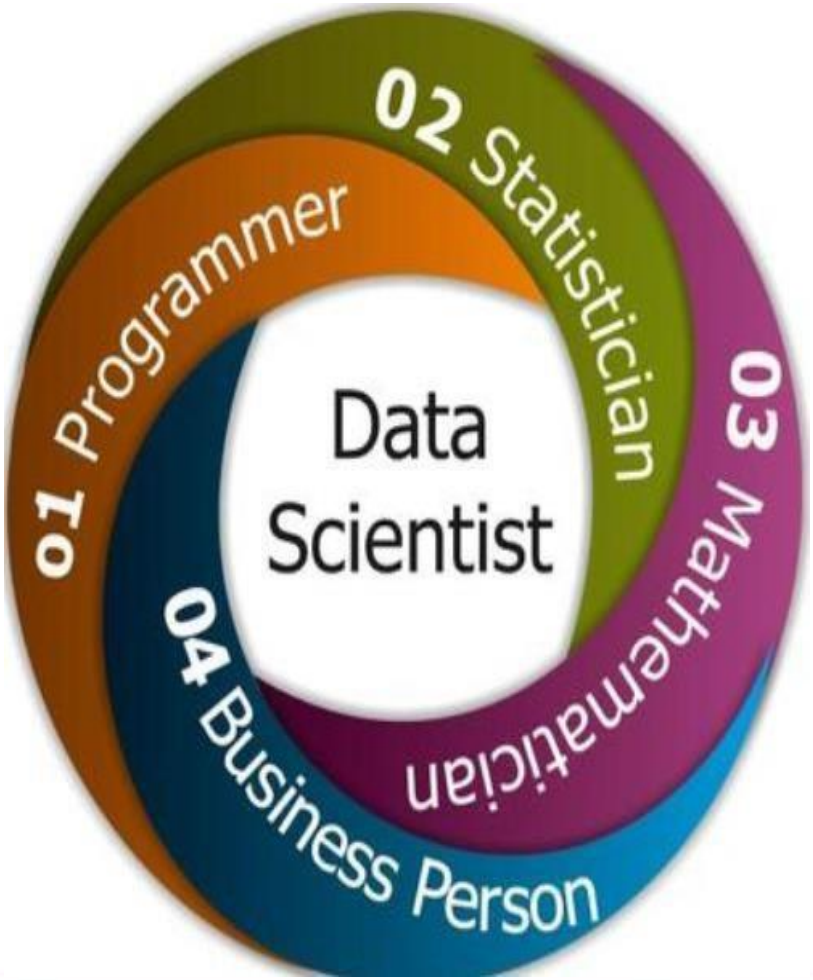
Harry Potter and the Half-Blood Prince
› J.K. Rowling
★★★★½ 235
Kindle Edition
₹ 159.00


Mail - vijesh@livew
LiveWire - Courses
CADD Centre - CSF
https://crewpoint.c
mobile phones - Bu
Mi Redmi Note 4 -
(19) What is Busine
Secure
https://www.flipkart.com/redmi-note-4-gold-32-gb/p/itmeqgs9gwrfwpf2?pid=MOBEQ98T82CYVHGZ&srno=s_1_1&otracker=search&lid=LSTMOBEQ98T82CYVHGZJFZSK1&qH=
Sell on Flipkart
Advertise
Gift Card
24x7 Customer Care
Track Order
Signup
Log In
Flipkart
Search for Products, Brands and More
CART 0
Back to top
Stay C ay
You will never have to say, "Sorry, ran out of battery," again. The Redmi Note 4 comes with a 4100 mAh rechargeable battery that keeps you connected all day.
Multitasking at its Best
Switching between apps or launching apps won't take forever with the Redmi Note 4, thanks to the octa-core Snapdragon 625 processor. To add more power to you, the Redmi Note 4 is 20% more power efficient than the Redmi Note 3.
View all features
Safe and Secure Payments. Easy returns. 100% Authentic products.
Similar products
5:29 PM
6/29/2017

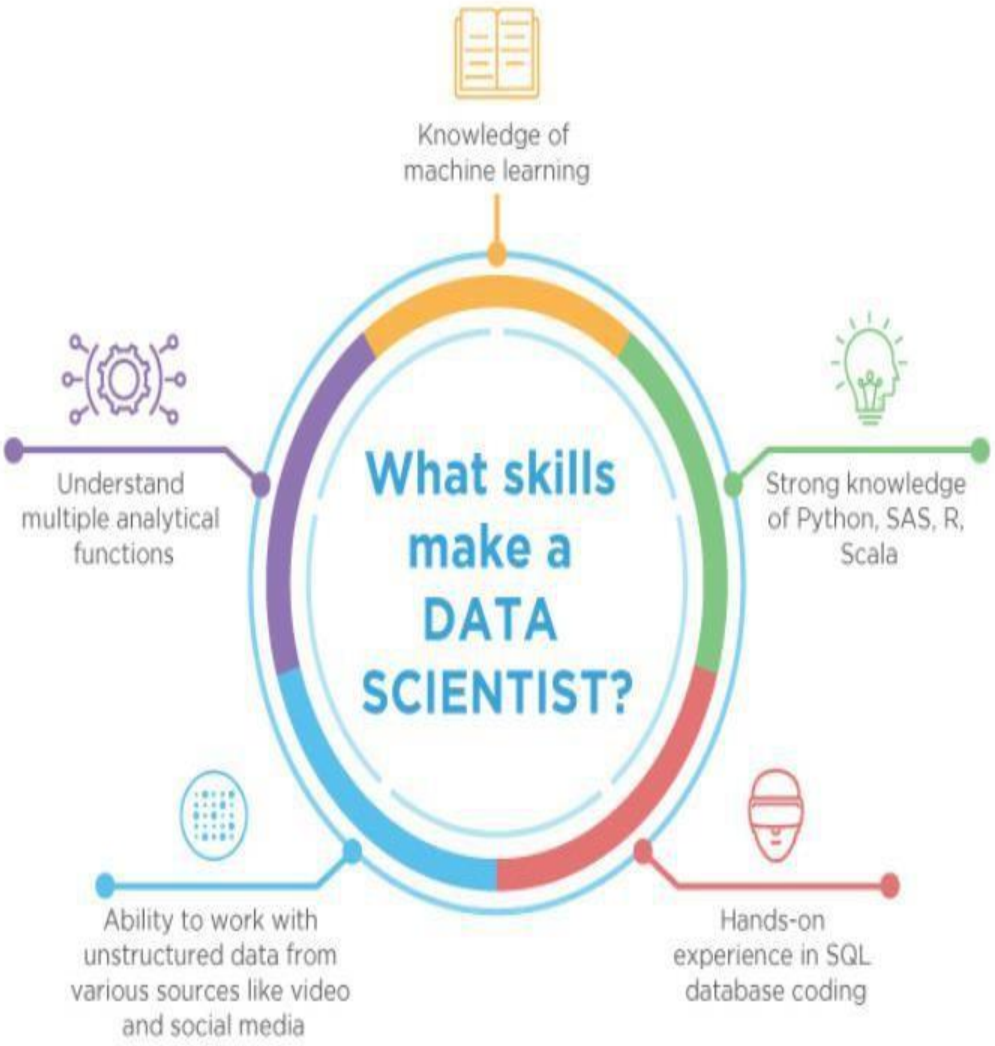

Secured and sharable medical records
Optimized and integrated care
Improved monitoring
Automated lab-work
Standardized image
Drug interaction detection
Personalized medicine
Communication assistance
Data sciences in healthcare

LIVEWIRE
FOR LIVE CAREERS

Who will discover data insights



LIVEWIRE
FOR LIVE CAREERS



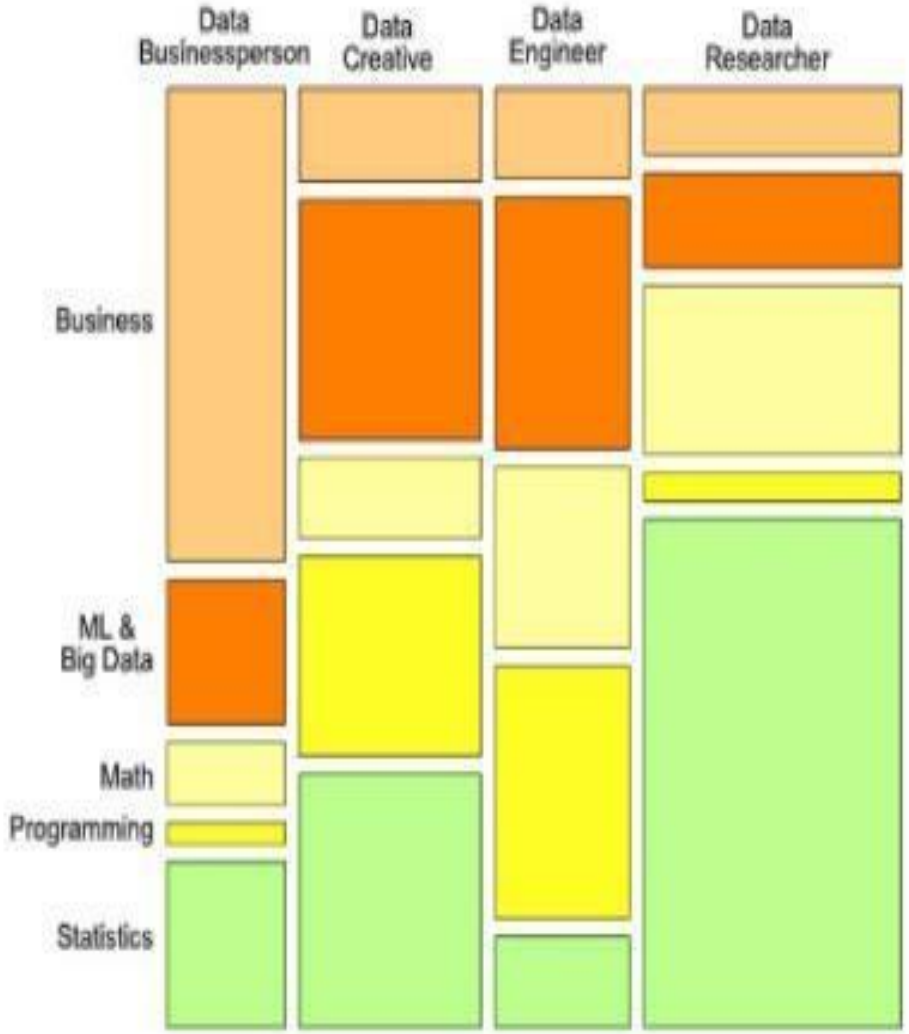
LIVEWIRE
FOR LIVE CAREERS

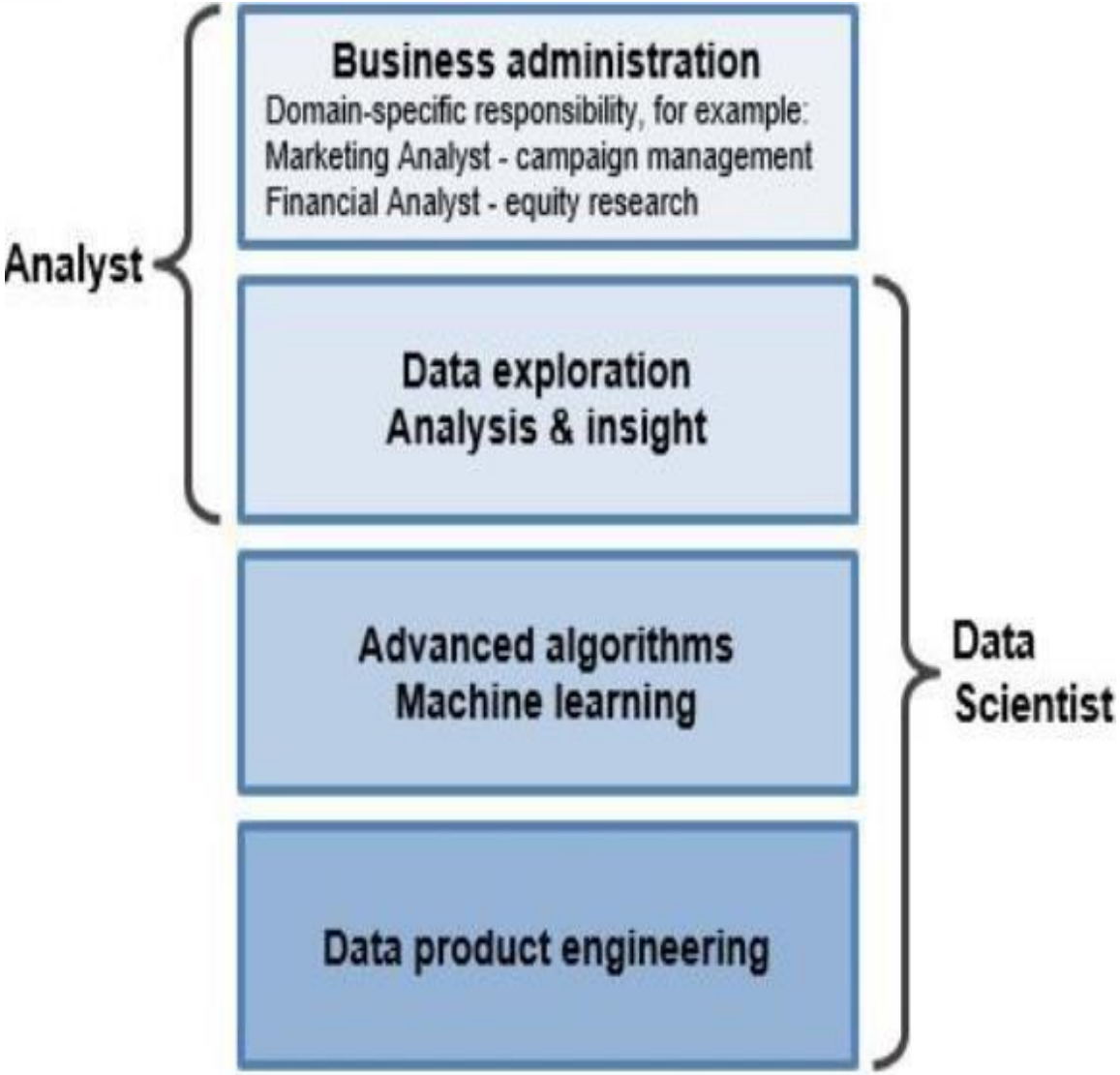
What a Data Scientist do

LIVEWIRE
FOR LIVE CAREERS

Types of Data Scientist

- Data Businesspeople
- Data Creatives
- Data Developers
- Data Researchers


Data Businessperson
Data Creative
Data Engineer
Data Researcher
Business
ML & Big Data
Math
Programming
Statistics


Business administration
Domain-specific responsibility, for example:
Marketing Analyst - campaign management
Financial Analyst - equity research
Analyst
Data exploration
Analysis & insight
Advanced algorithms
Machine learning
Data product engineering
Data
Scientist

Data is collection of facts or information

LIVEWIRE
FOR LIVE CAREERS

Walmart

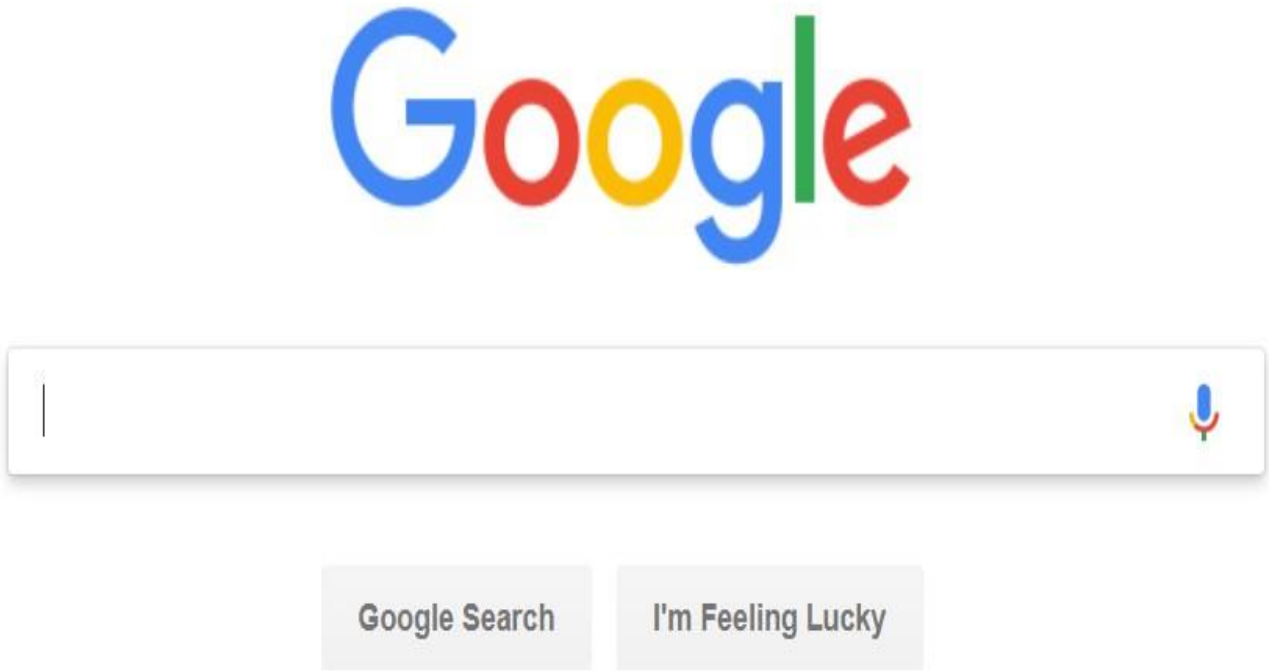
Flipkart

How many Facebook active users

2.27 billion

LIVEWIRE
FOR LIVE CAREERS

How many Google searches per day



Approximately 2 trillion searches

Data collection

- 16 million text messages (sent)
- 2.9 billion email user
- 154200 video calls
- every minute there are 103,447,520 spam emails sent
- 400 million instragrammers (active)
- 95 million photos and videos are shared on instagram
- 4,146,600 user watch utube vidoes
- 300 million videos get uploaded per minute

Why data ?

- data generation
- data-->collection of information

- dataset-->collection of information that related to particular subject

- structured data
- unstructured data

STRUCTURED DATA

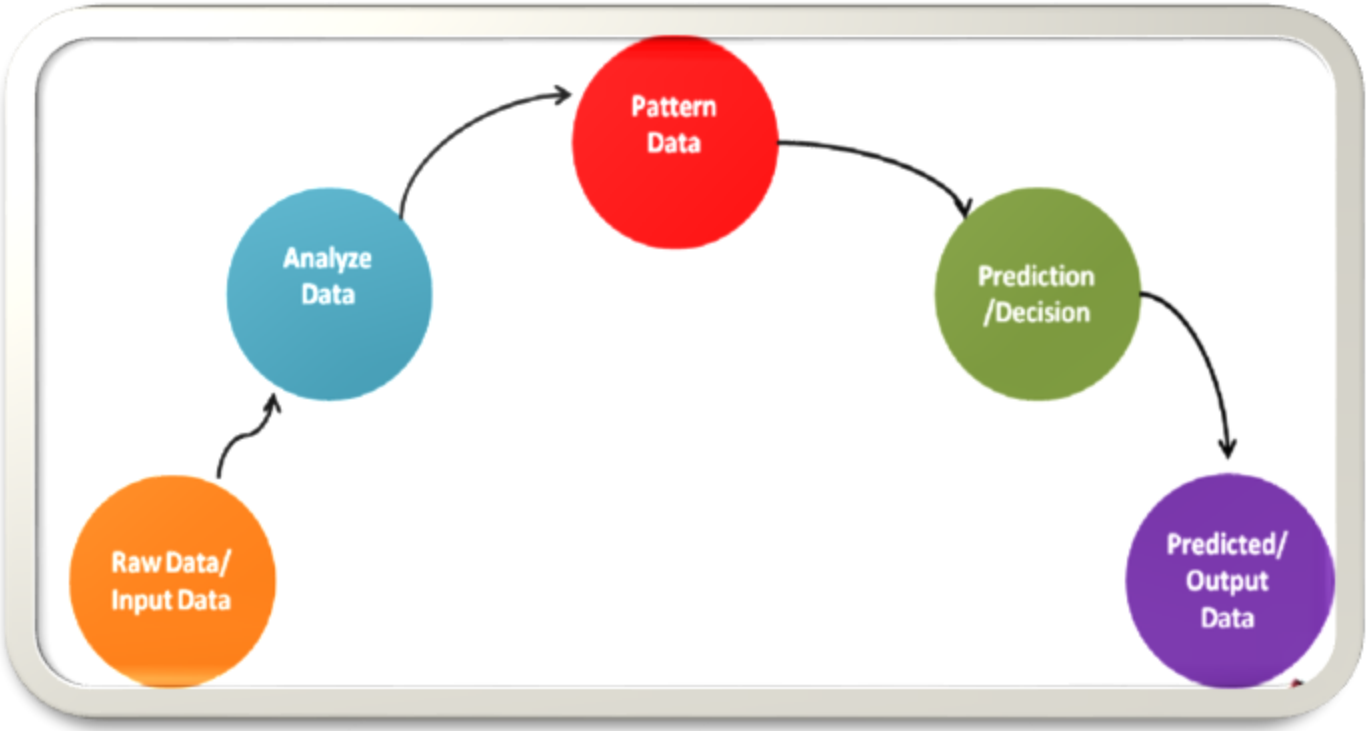
- Fixed in row and columns
- Structured data are those type of data which are stored already in an order
- There are nearly 20 % of total existing data are structured data.
- All the data generated from sensors ,weblogs,these are all machine generated structured data.
- The human generated structured data are those which are taken as information from a human . Like their names,addresses etc..,
- Example :- database

Unstructured data

- The unstructured data have no clear format in storage.
- Atleast 80% of data are unstructured .
- There are various ypes of human-generated unstructured data.
- These are images , videos , social media data etc..,
- Example:- text document, pdfs…,

LIVEWIRE
FOR LIVE CAREERS

DATA SCIENCE



LIVEWIRE
FOR LIVE CAREERS

INTRODUCTION TO PYTHON

Programming languages

- Types:
  - High level language(user friendly)
  - Low level language(machine friendly)
  - Assembly language
- Converters
  - Compiler
  - Interpreter
  - Assembler
- Standard of conversion
  - ASCII

Why Python?

- Syntax free
- Reference based language
- Huge number of library makes every task very easy
- Can integrate it with different technologies and hardware

LIVEWIRE
FOR LIVE CAREERS

Session_2

LIVEWIRE
FOR LIVE CAREERS

BASICS OF PYTHON

Topics

- Variables

```
>>> a=10
>>> type(a)
<class 'int'>
>>> a=20.30
>>> type(a)
<class 'float'>
>>> a="hello"
>>> type(a)
<class 'str'>
```

Topics

- Operators(Symbols use for computation purpose)
- Inputting

```
>>> a=input("enter your name")
enter your nameLiveWire
>>> a
'LiveWire'
```

- Outputting

```
>>> print("hello world")
hello world
```

Topics

- Array
  - List

```
>>> arr=[3,2,4,5,6,7,1,2,4]
>>> arr
[3, 2, 4, 5, 6, 7, 1, 2, 4]
```

  - Tuple

```
>>> tup=(1,4,5,8,9,2)
>>> tup
(1, 4, 5, 8, 9, 2)
```

  - Set

```
>>> se={4,2,5,7,8,9}
>>> se
{2, 4, 5, 7, 8, 9}
```

Topics

- Array
  - Dictionary

```
>>> dic={'Office':'LiveWire','Place':'Jaipur'}
>>> dic
{'Office': 'LiveWire', 'Place': 'Jaipur'}
```

- String

```
>>> st="LiveWire"
>>> type(st)
<class 'str'>
```

Topics

- Loops

```
>>> for var in range(5):
        print("hello")  #print hello for 5 times
>>>a=0
>>>while(a<5):
        print(“hello”)  #print hello 5 times
        a+=1
```

Topics

- Conditional statements

```
>>> a=5
>>> if(a==5):
        print("value is 5")
>>> else:
        print(“value is not 5”)
```

- Functions

```
>>> def funName():
        print("simple function")
```

Topics

- File handling

```
>>>fil=open('file.txt','w')
>>>fil.write("this is for test")
>>>fil.close()
```

- Modules and packages(Predefined and user defined)
- Pip and conda package

```
>>>pip install packageName
>>>conda install packageName
```

Topics

- Numpy

```
>>>import numpy as np
>>>np.sum([1,2,34,5,6,7,8])
```

- Pandas(Library for data manipulation and analysis)
- Matplotlib(Library for making graphical display)
- Regex(Library for matching expression in sentence)

Numpy

- NumPy is a Python library used for working with arrays.
- It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- NumPy stands for Numerical Python.

Advantages of numpy over list:

- Less memory
- Fast
- Convenient

Pandas

- Data scientists make use of Pandas in Python for its following advantages:

1. Easily handles missing data
2. It uses Series for one-dimensional data structure and DataFrame for multi-dimensional data structure
3. It provides an efficient way to slice the dataIt provides a flexible way to merge, concatenate or reshape the data
4. It includes a powerful time series tool to work with

Matplotlib

- Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.
- As such, it offers a viable open source alternative to MATLAB.
- Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications.
- Axes include the X-Axis, Y-Axis, and possibly a Z-Axis, as well.

Regex

- It can detect the presence or absence of a text by matching with a particular pattern, and also can split a pattern into one or more sub-patterns.
- Python provides a re module that supports the use of regex in Python.
- Its primary function is to offer a search, where it takes a regular expression and a string.

LIVEWIRE
FOR LIVE CAREERS

Session_3

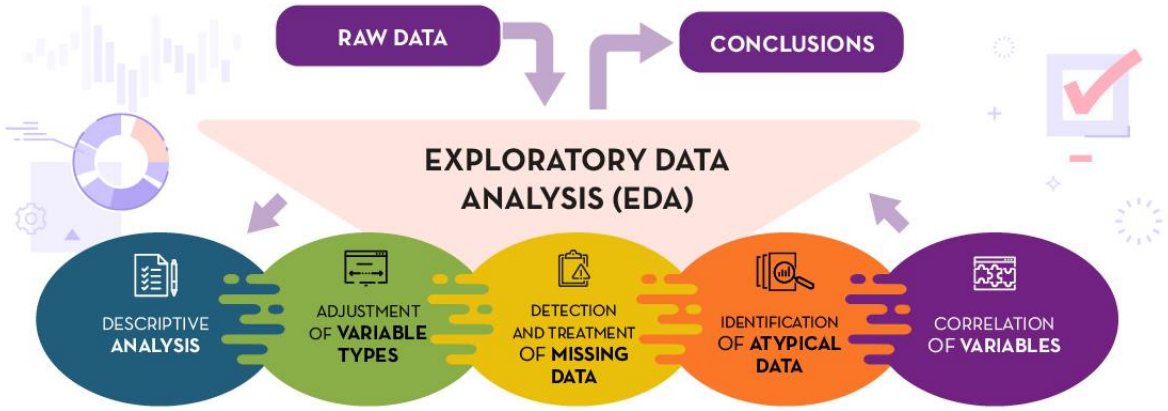
LIVEWIRE
FOR LIVE CAREERS

EDA (Exploratory Data Analysis)

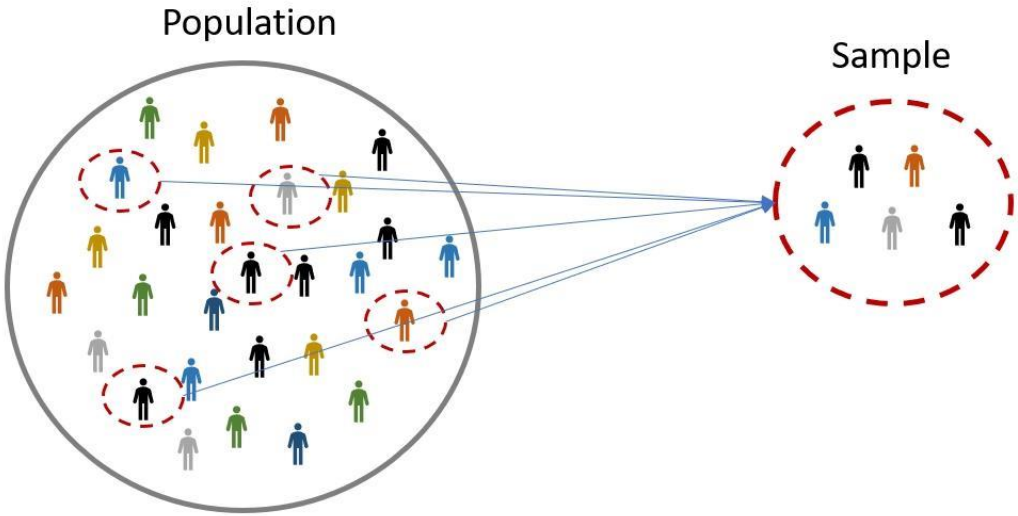
Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is **an approach to analyze the data using visual techniques**.

- It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

Sample vs Population

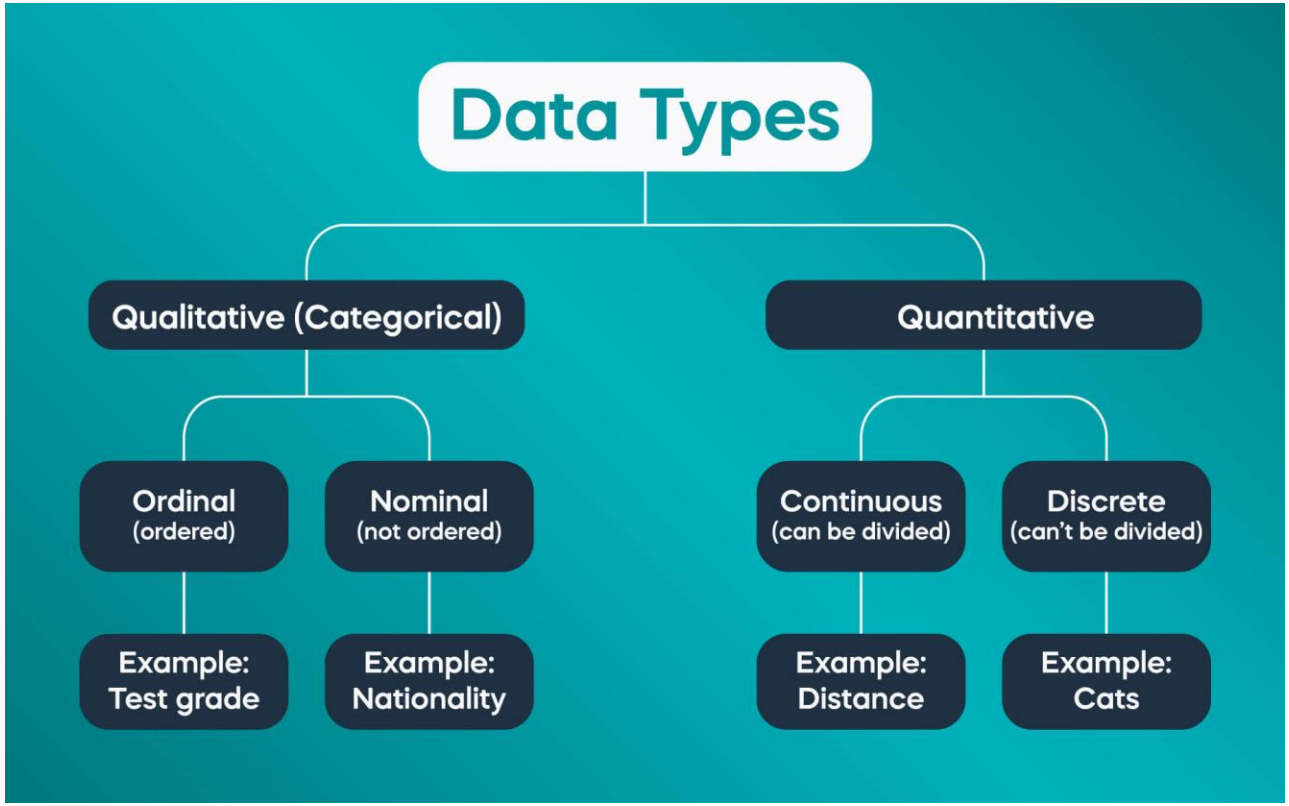


- EDA is statisticians way of *story telling*

LIVEWIRE
FOR LIVE CAREERS

Types of Data

- Qualitative
- Quantitative


Data Types
Qualitative (Categorical)
Quantitative
Ordinal
(ordered)
Nominal
(not ordered)
Continuous
(can be divided)
Discrete
(can't be divided)
Example:
Test grade
Example:
Nationality
Example:
Distance
Example:
Cats

Descriptive Statistics

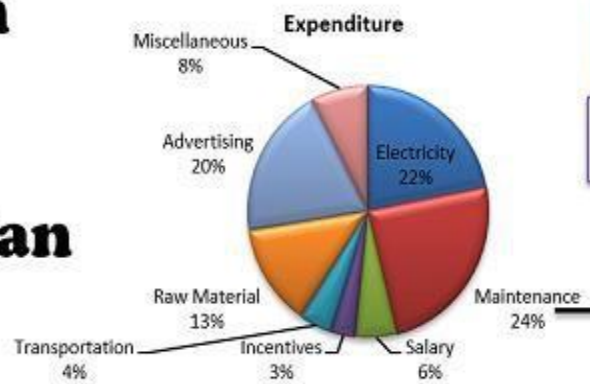
- Summarizing and Organizing data

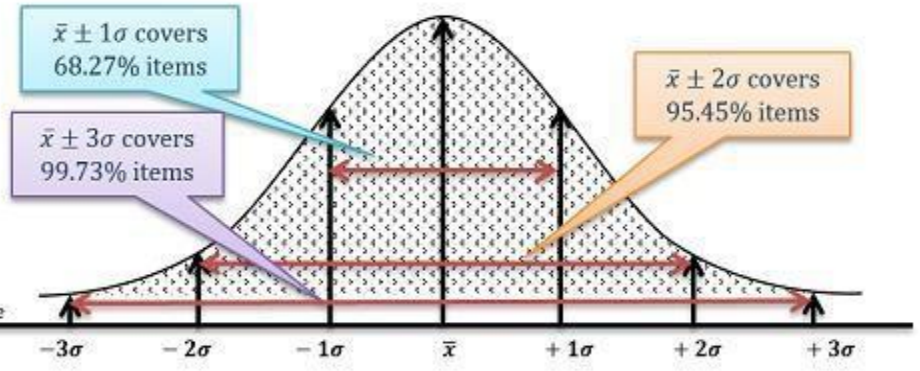
Descriptive Statistics

Mean

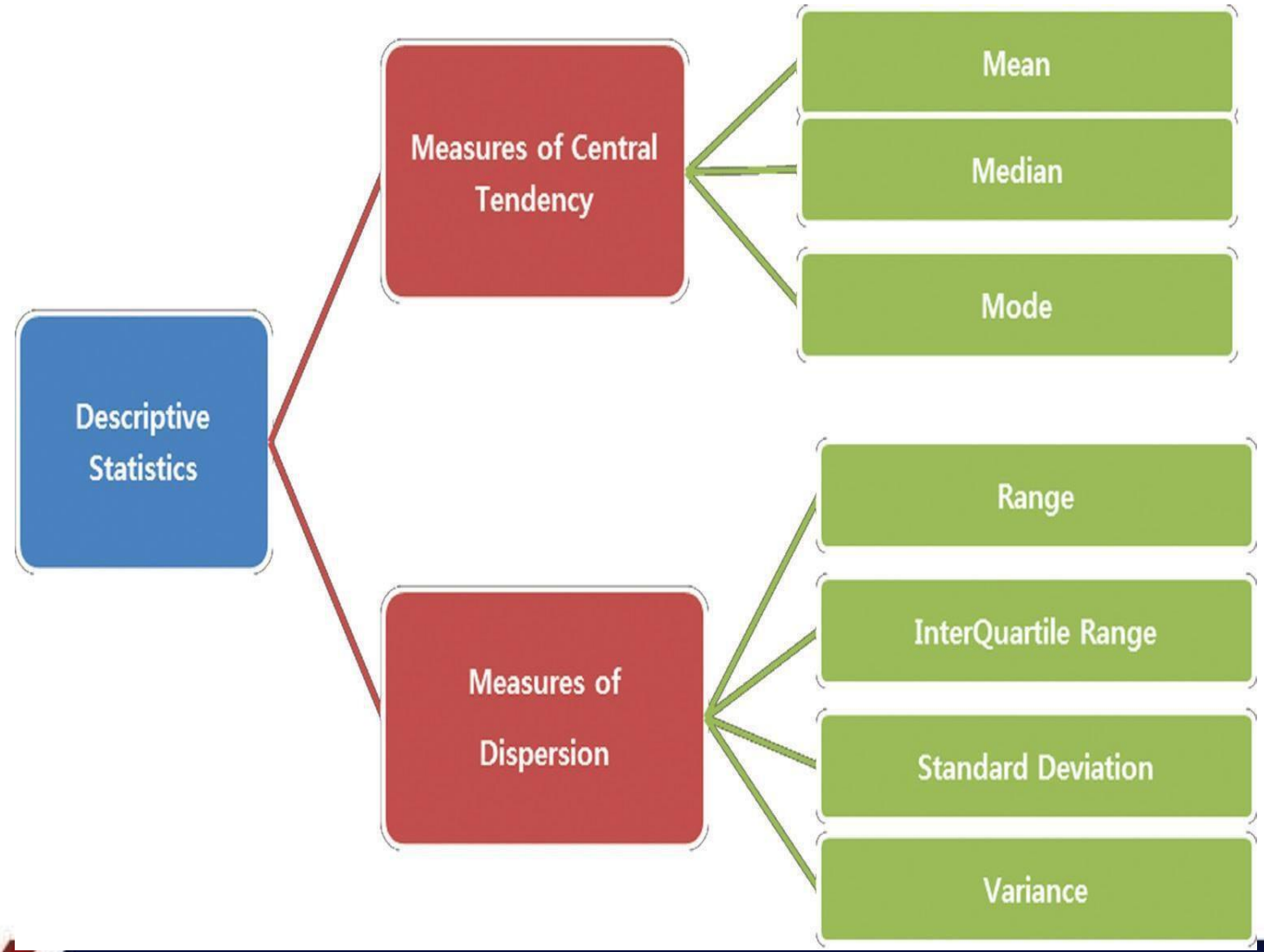
Median

Mode





$$Std.\,Dev.\,\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$


Descriptive Statistics
Measures of Central Tendency
Mean
Median
Mode
Measures of Dispersion
Range
InterQuartile Range
Standard Deviation
Variance

Measures of Central Tendency

- One number that best summarizes the entire set of measurements
- A number that is in some way “central” to the set.

Mean / Average

- Mean or Average is a central tendency of the data

$$x = \frac{12+24+41+51+67+67+85+99}{8} = 55.75$$

Median

- Median is the value which divides the data in 2 equal parts
- Median will be a middle term, if number of terms is odd
- Median will be average of middle 2 terms, if number of terms is even.

12+24+41+51+67+67+85+99 = 59

When Mean = Median

- When the numbers are in arithmetic progression
  - 2 ,4,6,8,10

    Mean = 6

    Median = 6

Mode

- Mode is the term appearing maximum time in data set.
- **Bimodal**
- **Trimodal**
- **Multimodal**

12, 24, 41, 51, 67, 67, 85, 99

Measures of Spread/Deviation

RANGE

Quartiles

- **Inter-quartile Range -** Divided by 4 (ie. 25%, 50%, 75% and 100%)
- **Inter-turtle Range -** Divided by 3
- **Inter-quantile Range -** Divided by 5

Variance

- Variance is an average of squared deviations about the mean

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Note: To avoid getting Zero, the deviation values are squared before they are added up.

Standard deviation

- Standard deviation is the squared root of variance:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n}}$$

Note: If the data points are too far from the mean, there is higher deviation within the data set.

Frequency

- Frequency of an event
- Represented as “n” number of times the event occurred in an experiment
- Frequencies are often graphically represented in histograms

Number of marks	Tally marks	Frequency
1	𝍸 ||	7
2	𝍸	5
3	𝍸 |	6
4	𝍸	5
5	|||	3
Total		**26**

KURTOSIS

KURTOSIS

- Kurtosis refers to **the degree of presence of outliers in the distribution**.
- Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.

- There are three types of kurtosis:
  1. mesokurtic,
  2. leptokurtic,
  3. and platykurtic.

Why do we need kurtosis?

- It is used **to describe the extreme values in one versus the other tail**.
- It is actually the measure of outliers present in the distribution .
- High kurtosis in a data set is an indicator that data has heavy tails or outliers.
- If there is a high kurtosis, then, we need to investigate why do we have so many outliers.

Program:

```
import pandas as pd
returns =[2,4,1,34,22,1,32,55,1,3,8,67,98,10]
series  = pds.Series(returns)
print("Kurtosis:")
print(round(series.kurtosis(), 2))
```

LIVEWIRE
FOR LIVE CAREERS

Session_4

SKEWNESS

SKEWNESS

- The skewness is **a measure of symmetry or asymmetry of data distribution**

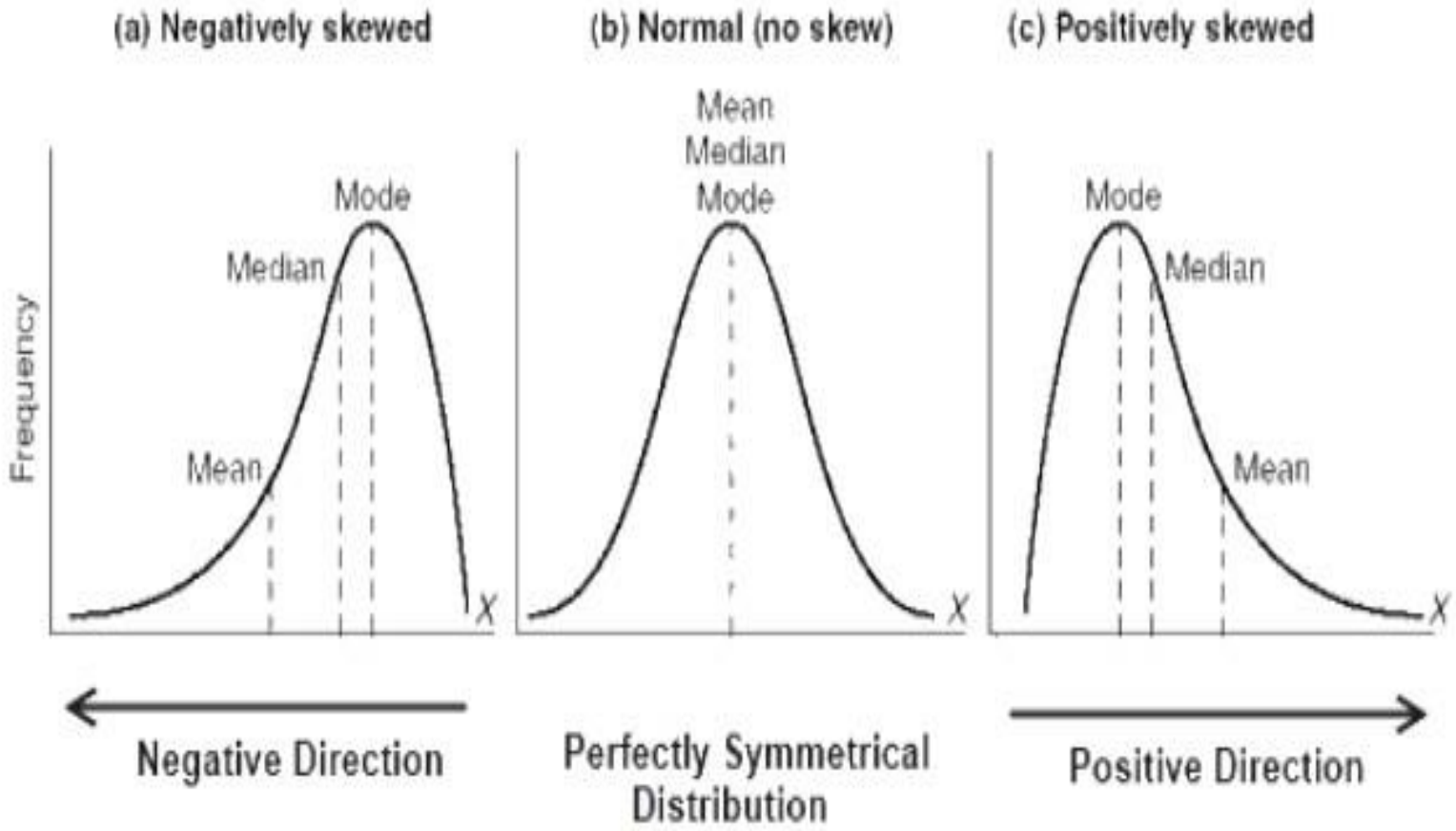
- Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

The three types of skewness are:

•Right skew (also called positive skew). A right-skewed distribution is longer on the right side of its peak than on its left.

•Left skew (also called negative skew). A left-skewed distribution is longer on the left side of its peak than on its right.

•Zero skew.


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
Mean
Median
Mode
Mode
Median
Mean
X
X
X
Negative Direction
Perfectly Symmetrical Distribution
Positive Direction

Program :

```
from scipy.stats import skew
x =[55, 78, 65, 98, 97, 60, 67, 65, 83, 65]
print(skew(x))
```

Manual:

X = [55, 78, 65, 98, 97, 60, 67, 65, 83, 65]

Calculating the mean of X we get: $\bar{x} = 73.3$.

Solving for m_3:

$$m_3 = \frac{1}{10}\sum_{n=1}^{10}(x_n - \bar{x})^3$$

$$m_3 = \frac{(55-73.3)^3 - (78-73.3)^3 - \ldots - (65-73.3)^3}{10}$$

Solving for m_2:

$$m_2 = \frac{1}{10}\sum_{n=1}^{10}(x_n - \bar{x})^2$$

$$m_2 = \frac{(55-73.3)^2 - (78-73.3)^2 - \ldots - (65-73.3)^2}{10}$$

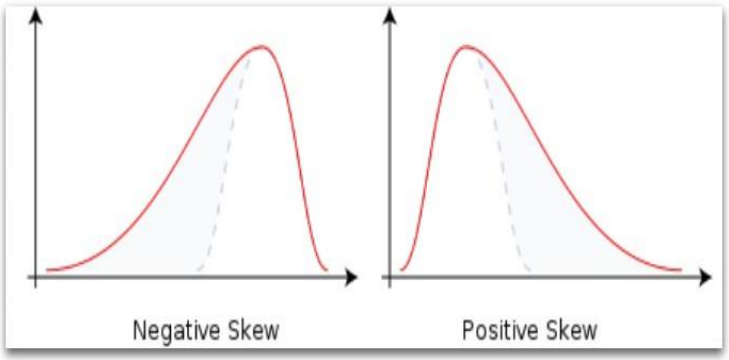
Solving for g_1:

$$g_1 = \frac{m_3}{(m_2)^{\frac{3}{2}}} = \frac{1,895.124}{(204.61)^{\frac{3}{2}}} = 0.647511$$

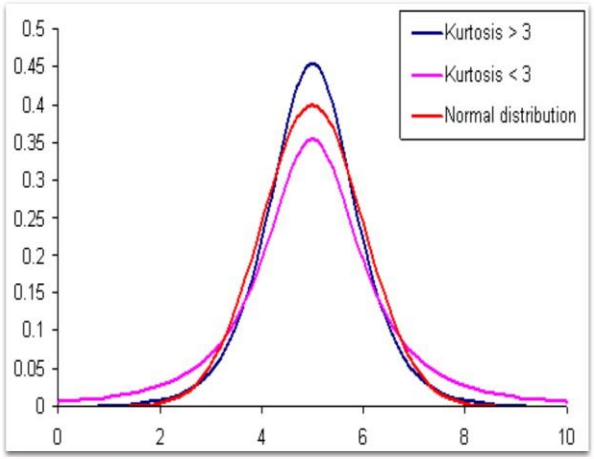
The Fisher–Pearson coefficient of skewness is equal to 0.647511 in this example and show that there is a positive skew in the data. Another way to check it is to look at the mode, median, and mean for these values. Here we

Skewness & Kurtosis

- Measure of symmetry
- Symmetrical distribution has a skewness of zero
- Normal, Right or Left tail



- Measure of flatness
- Gaussian distribution has a kurtosis of 0
- By Prism, Gaussian distribution is expected to have 3 as kurtosis (High, Depth or Flat)



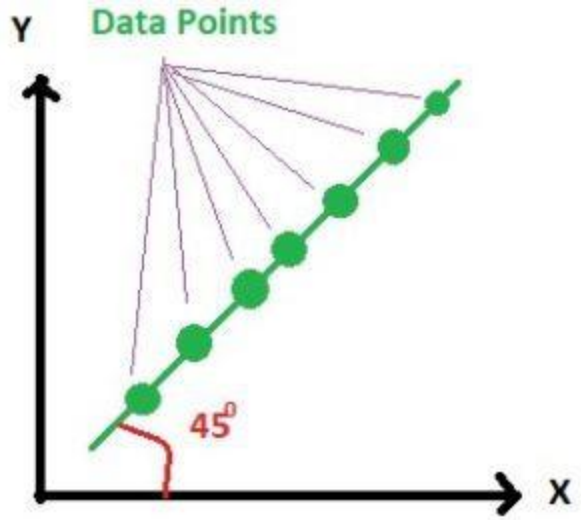
Q-Q PLOT

- qqplot (Quantile-Quantile Plot) in Python

- When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot.
- This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

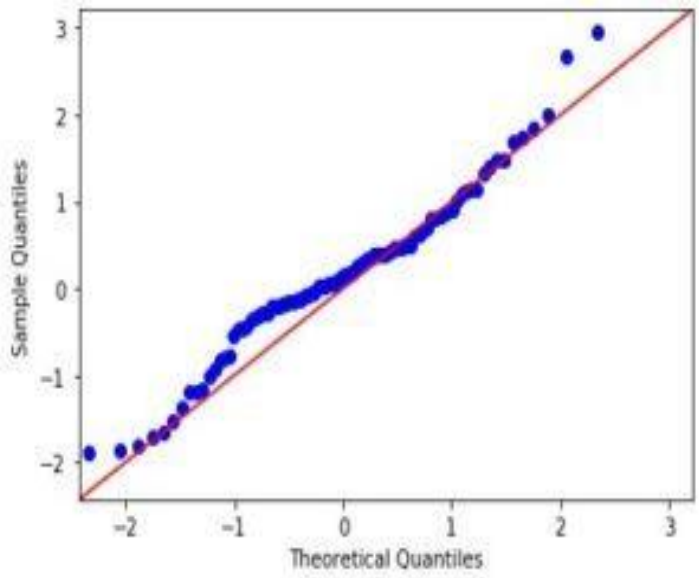
- Interpretations

•All point of quantiles lie on or close to straight line at an angle of 45 degree from x – axis. It indicates that two samples have similar distributions.

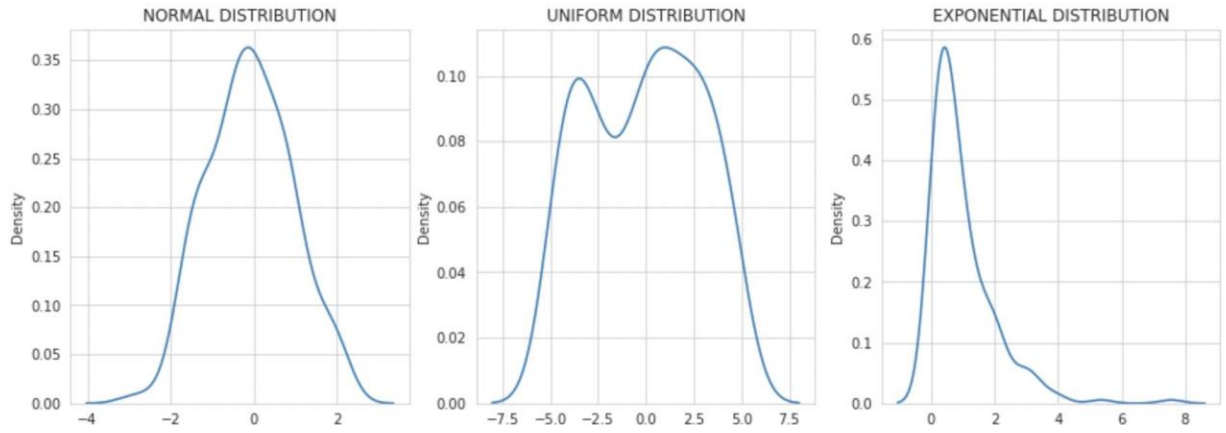


```
import numpy as np
import statsmodels.api as sm
import pylab as py
data_points = np.random.normal(0,1, 100)

sm.qqplot(data_points, line ='45')
py.show()
```



There are various probability distribution types like Gaussian or Normal Distribution, Uniform distribution, Exponential distribution, Binomial distribution, etc



Normal distributions	Uniform distribution	Exponential distributions
• Normal distributions are the most popular ones. • They are a probability distribution that peaks at the middle and decreases at the end of the axis. • It is also known as a bell curve or Gaussian Distribution.	• Uniform distribution is a probability distribution type where the probability of occurrence of x is constant. • For instance, if you throw a dice, the probability of any number is uniform	• Exponential distributions are the ones in which an event occurs continuously and independently at a constant rate. • It is commonly used to measure the expected time for an event to occur.

LIVEWIRE
FOR LIVE CAREERS

Session_5

LIVEWIRE
FOR LIVE CAREERS

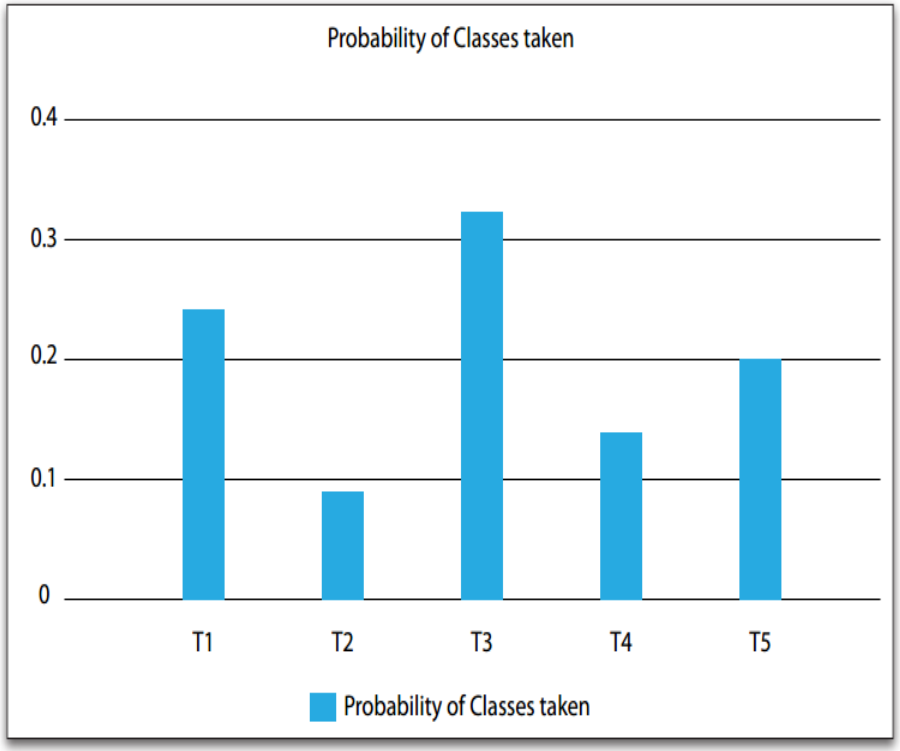
PROBABILITY THEORY

PROBABILITY DISTRIBUTION

How the probabilities are distributed?

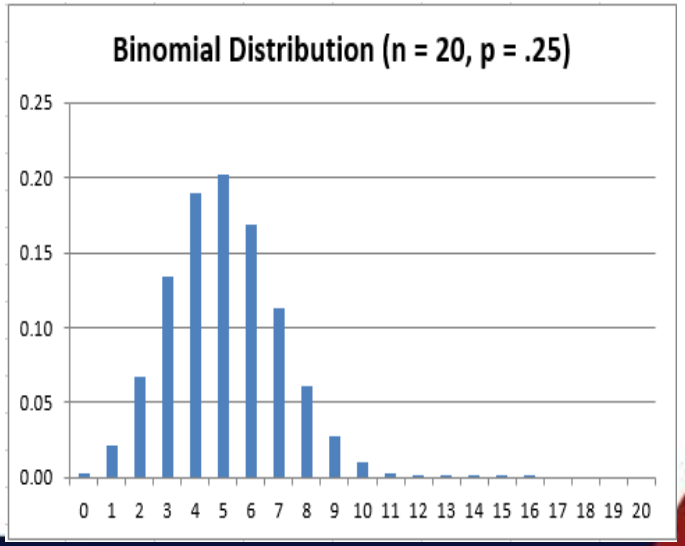
$$\sum_{r} P(r) - 1 \text{ for discrete and } \int_{-\infty}^{\infty} f(x)\,dx - 1 \text{ for continuous}$$

Event	No of classes	Probability
Trainer 1	20	0.25
Trainer 2	7	0.0875
Trainer 3	26	0.325
Trainer 4	11	0.1375
Trainer 5	16	0.2
Total probability		1



Binomial Distribution

$$p(r,N,p)=\binom{N}{r}p^{r}(1-p)^{N-r}$$



LIVEWIRE
FOR LIVE CAREERS

Binomial Distribution

- The binomial distribution is one of the most commonly used distributions in statistics. It describes the probability of obtaining *k* successes in *n* binomial experiments.

- If a random variable *X* follows a binomial distribution, then the probability that $X = k$ successes can be found by the following formula:
- $\mathbf{P(X=k) = {}_nC_k * p^k * (1-p)^{n-k}}$

- where:
- **n:** number of trials
- **k:** number of successes
- **p:** probability of success on a given trial
- $\mathbf{{}_nC_k}$**:** the number of ways to obtain *k* successes in *n* trials

Question 1: Nathan makes 60% of his free-throw attempts. If he shoots 12 free throws, what is the probability that he makes exactly 10?

```
From scipy.stats import binom
#calculate binomial probability
binom.pmf(k=10, n=12, p=0.6)
```

Output:

```
0.0639
```

The probability that Nathan makes exactly 10 free throws is 0.0639.

Question 2: Marty flips a fair coin 5 times. What is the probability that the coin lands on heads 2 times or fewer?

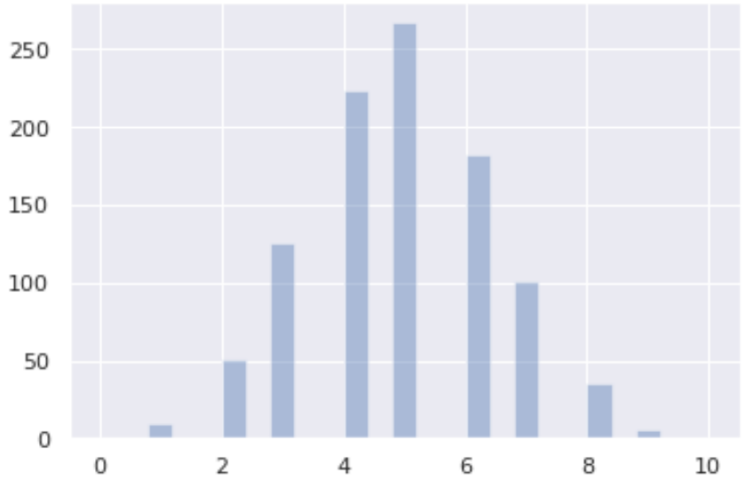
```
from scipy.stats import binom
#calculate binomial probability
binom.cdf(k=2, n=5, p=0.5)
```

Output:
0.5

The probability that the coin lands on heads 2 times or fewer is 0.5.

You can visualize a binomial distribution in Python by using
the **seaborn** and **matplotlib** libraries:

```
from numpy import random import
matplotlib.pyplot as plt
import seaborn as sns
x = random.binomial(n=10, p=0.5, size=1000)
sns.distplot(x, hist=True, kde=False)
plt.show()
```



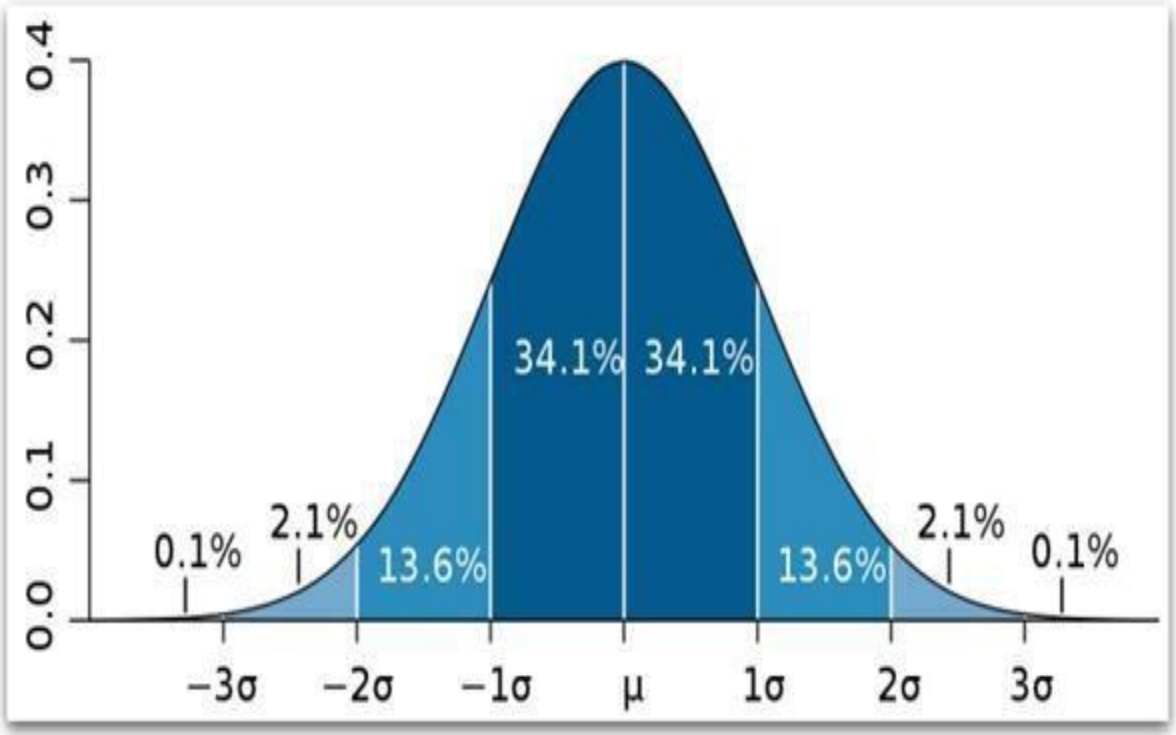
The x-axis describes the number of successes during 10 trials and the y-axis displays the number of times each number of successes occurred during 1,000 experiments.

Example:

- Tossing a coin 10 times for occurrences of head
- Rolling a dice to check for occurrence of a 2
- Surveying a population of 100 people to know if they watch television or not

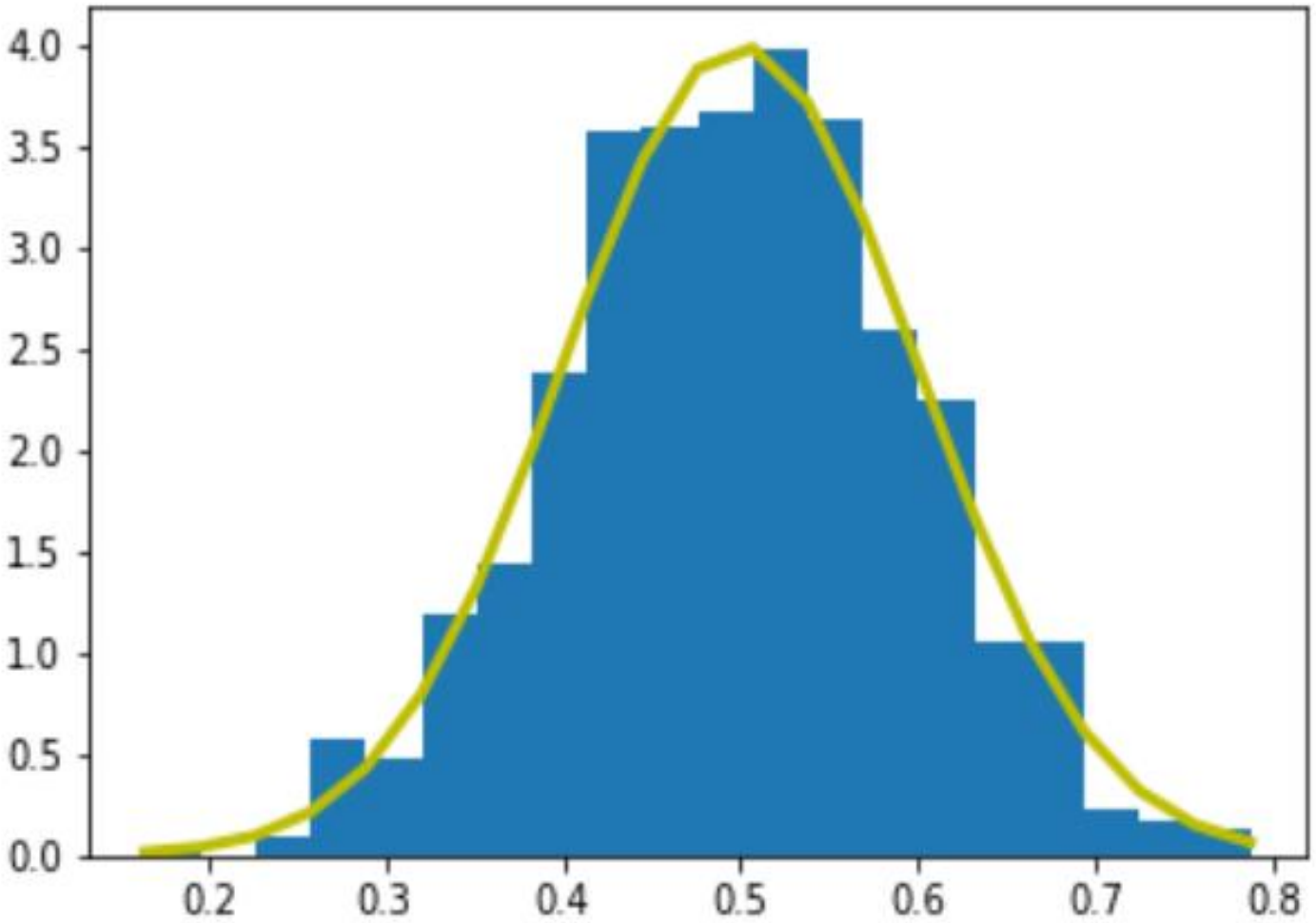
Normal Distribution

$$f\left(x:\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



The normal distribution is a form presenting data by arranging the probability distribution of each value in the data.Most values remain around the mean value making the arrangement symmetric.

```
Import matplotlib.pyplot as plt
import numpy as np
mu, sigma = 0.5, 0.1
s = np.random.normal(mu, sigma, 1000)
# Create the bins and histogram
count, bins, ignored = plt.hist(s, 20, normed=True)
# Plot the distribution curve
plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) * np.exp( - (bins – mu)**2 / (2 *
sigma**2) ), linewidth=3, color='y')
plt.show()
```

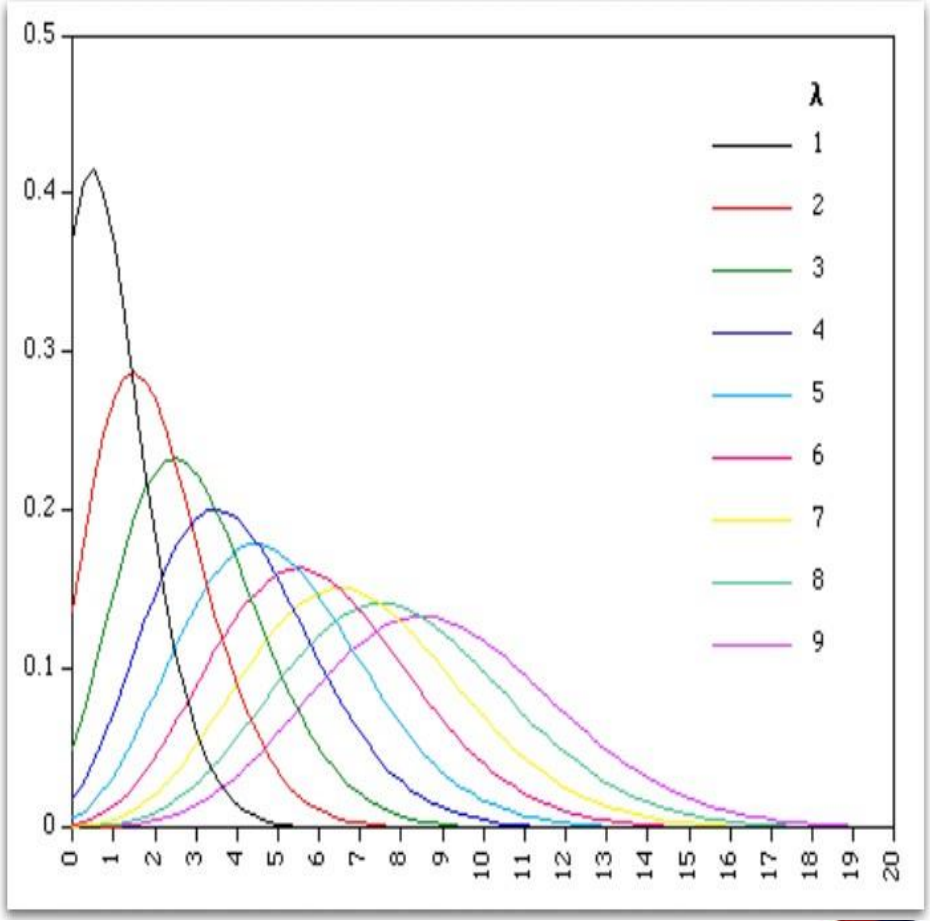

4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Example:

- Heights of people
- Size of things
- Errors in measurements
- Blood pressure
- Marks on a test

Poisson Distribution

$$p(r:\mu) = \frac{\mu^r e^{-\mu}}{r!}$$



- The Poisson distribution describes the probability of obtaining *k* successes during a given time interval.
- If a random variable *X* follows a Poisson distribution, then the probability that *X* = *k* successes can be found by the following formula:
- **$P(X=k) = \lambda^k * e^{-\lambda} / k!$**

where:

•**λ:** mean number of successes that occur during a specific interval

•**k:** number of successes

•**e:** a constant equal to approximately 2.71828

You can use the **poisson.rvs(mu, size)** function to generate random values from a Poisson distribution with a specific mean value and sample size:

```
from scipy.stats import poisson
#generate random values from Poisson distribution with mean=3 and sample size=10
 poisson.rvs(mu=3, size=10)
```

Output:

```
 array([2, 2, 2, 0, 7, 2, 1, 2, 5, 5])
```

You can use the **poisson.pmf(k, mu)** and **poisson.cdf(k, mu)** functions to calculate probabilities related to the Poisson distribution.

Example 1: Probability Equal to Some Value

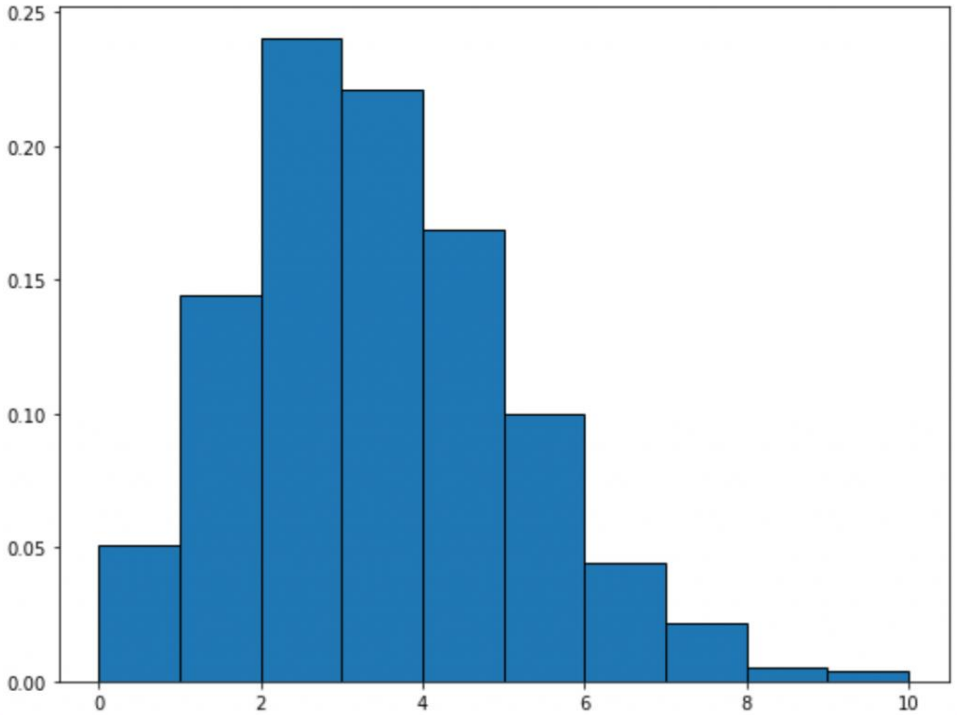
A store sells 3 apples per day on average. What is the probability that they will sell 5 apples on a given day?

```
from scipy.stats import poisson
 #calculate probability
poisson.pmf(k=5, mu=3)
```

Output:

0.100819

```
from scipy.stats import poisson
import matplotlib.pyplot as plt
#generate Poisson distribution with sample size 10000
 x = poisson.rvs(mu=3, size=10000)
#create plot of Poisson distribution
 plt.hist(x, density=True, edgecolor='black')
```



Example:

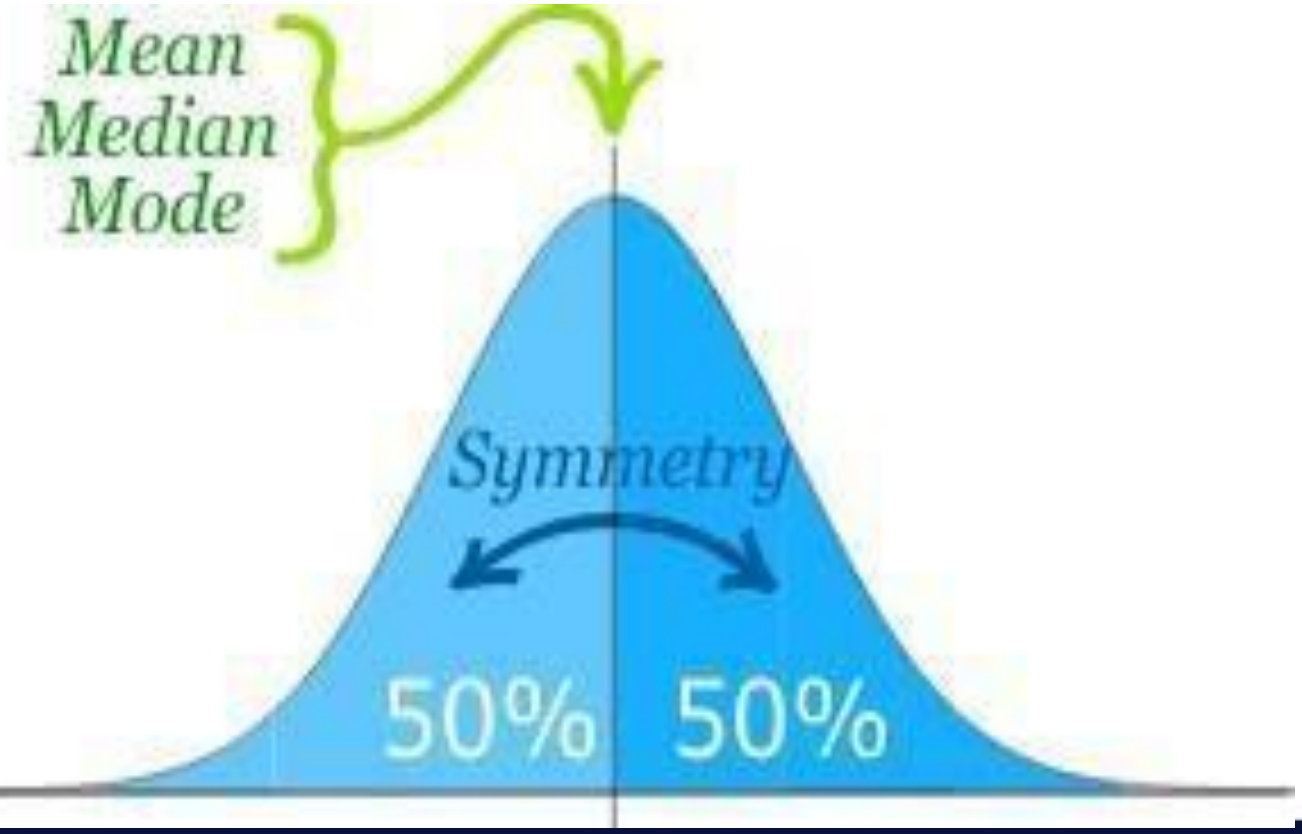
- Car accidents
- Traffic flow and ideal gap distance
- Number of typing errors on a page
- Failure of a machine in one month

LIVEWIRE
FOR LIVE CAREERS

NORMAL DISTRIBUTION

Normal Distribution

- Standard Normal distribution

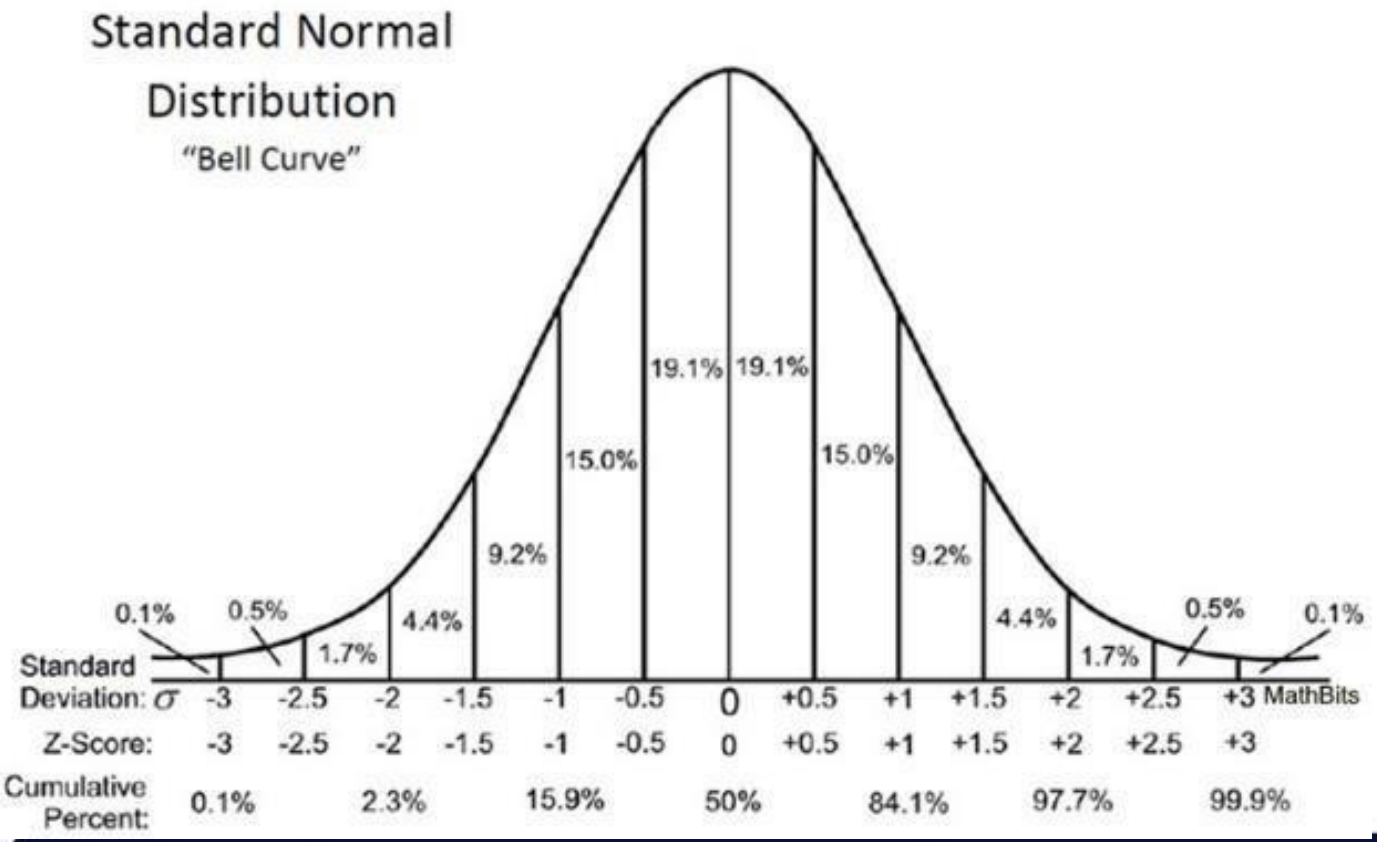


LIVEWIRE
FOR LIVE CAREERS

Z-SCORE

Z-Score

- Converts data into frequency distribution with mean 0 and standard deviation 1.



- In statistics, a z-score tells us how many standard deviations away a value is from the mean. We use the following formula to calculate a z-score:

$z = (X – \mu) / \sigma$

where:

X is a single raw data valueμ is the population meanσ is the population standard deviation

- We can calculate z-scores in Python using scipy.stats.zscore, which uses the following syntax:

- Scipy.stats.zscore(a, axis=0, ddof=0, nan_policy='propagate')
- where:
- a: an array like object containing data

- axis: the axis along which to calculate the z-scores. Default is 0.

- ddof: degrees of freedom correction in the calculation of the standard deviation. Default is 0.

- nan_policy: how to handle when input contains nan.

- Default is propagate, which returns nan. 'raise' throws an error and 'omit' performs calculations ignoring nan values.

```
Import pandas as pd
import numpy as np
import scipy.stats as stats
Data = np.array([6, 7, 7, 12, 13, 13, 15, 16, 19, 22])
Stats.zscore(data)
```

Output:

[-1.394, -1.195, -1.195, -0.199, 0, 0, 0.398, 0.598, 1.195, 1.793]

- Each z-score tells us how many standard deviations away an individual value is from the mean. For example:
- The first value of “6” in the array is **1.394** standard deviations *below* the mean.
- The fifth value of “13” in the array is **0** standard deviations away from the mean, i.e. it is equal to the mean.
- The last value of “22” in the array is **1.793** standard deviations *above* the mean.

LIVEWIRE
FOR LIVE CAREERS

Session_6

CONFIDENCE INTERVAL

CONFIDENCE INTERVAL

- Describe the amount of uncertainty associated with a sample estimate of a population parameter
- Confidence Interval Data Requirements
  - Margin of error

    (Margin of error = Critical value * Standard deviation of statistic)

    (Margin of error = Critical value * Standard error of statistic)

Critical value

Margin of Error

Statistic

- A **confidence interval for a mean** is a range of values that is likely to contain a population mean with a certain level of confidence.
- It is calculated as:
- **Confidence Interval = x +/- t*(s/√n)**
- where:
- **x:** sample mean
- **t:** t-value that corresponds to the confidence level
- **s:** sample standard deviation
- **n:** sample size

The following example shows how to calculate a confidence interval for the true population mean height (in inches) of a certain species of plant, using a sample of 15 plants:

```
import numpy as np
 import scipy.stats as st
#define sample data
data = [12, 12, 13, 13, 15, 16, 17, 22, 23, 25, 26, 27, 28, 28, 29]
#create 95% confidence interval for population mean weight
st.t.interval(alpha=0.95, df=len(data)-1, loc=np.mean(data),
scale=st.sem(data))
```

Output:

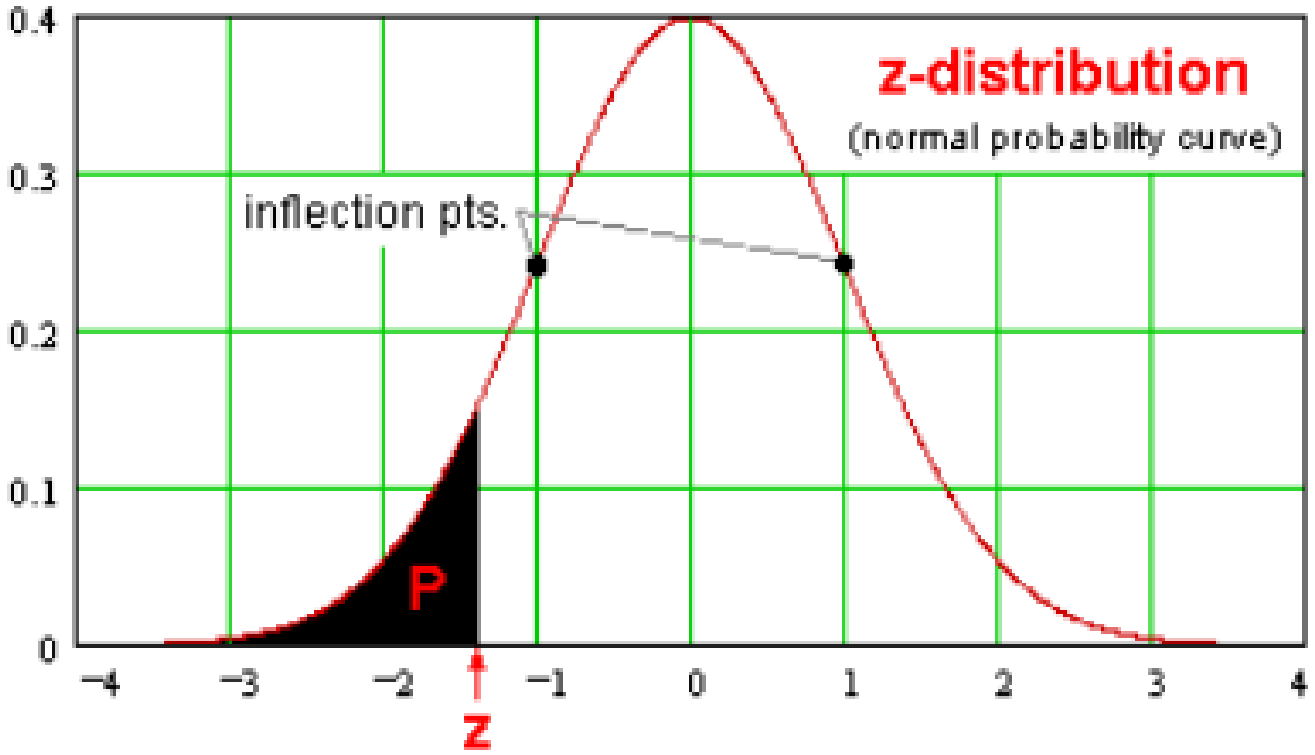
(16.758, 24.042)

LIVEWIRE
FOR LIVE CAREERS

Z-DISTRIBUTION

Z Distribution

- Z-distribution is used to help find probabilities and percentiles for regular normal distributions



- Confidence Interval with Z-Statistics
- **Given:**
- population variance, σ2σ2, is given
- confidence level of the interval: 95% (two-tailed) and 99% (two-tailed)
- **Assumptions:**
- population variance is given
- population is normally distributed,
- even if it is not sample size is large (>30)

- **Formula:**

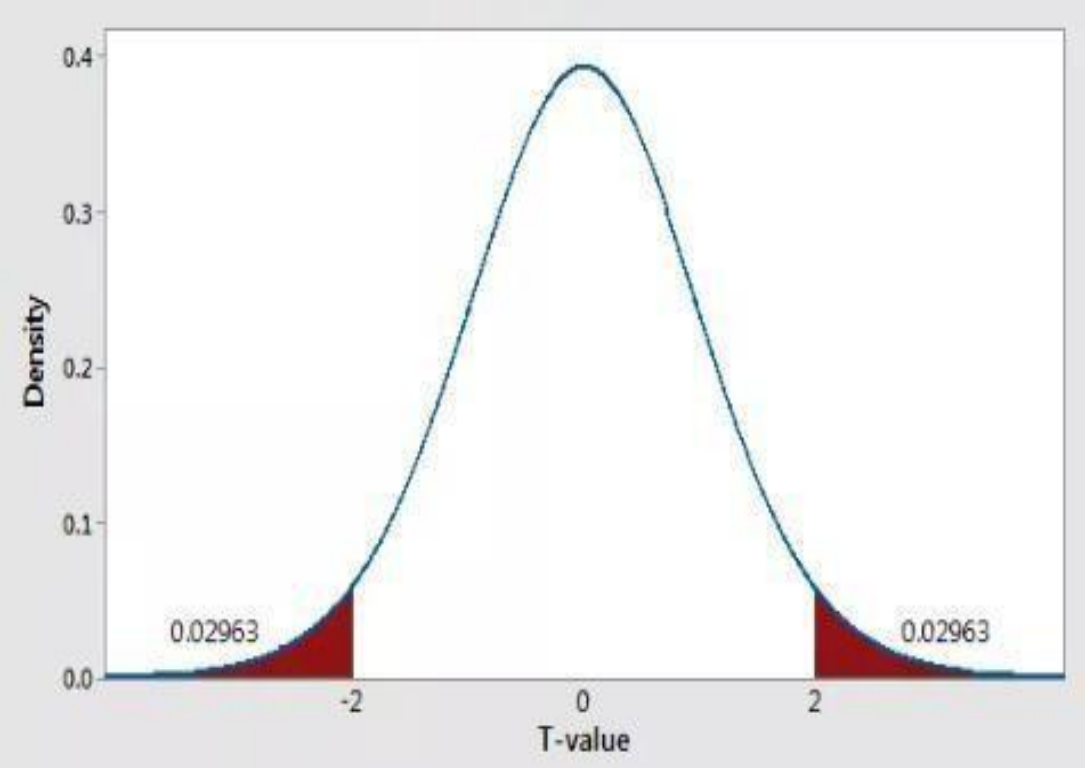
- Confidence Interval (z-score)=[x¯−zα/2×σn−−√,x¯+zα/2×σn−−√]
-

LIVEWIRE
FOR LIVE CAREERS

T-DISTRIBUTION

T Distribution

- Utilized when
  - Unknown standard deviation / Small sample size



- Confidence Interval with T-Statistics
- **Given:**
- confidence level of the interval: 95% (two-tailed) and 99% (two-tailed)
- **Assumptions:**
- population variance is given
- population is normally distributed, even if it is not
- sample size is 20 (<30)
- population variance, σ2σ2, is NOT given

- **Formula:**

- Confidence Interval (t-score)=[x¯−tn−1,α/2×sn−−√,x¯+tn−1,α/2×sn−−√]

LIVEWIRE
FOR LIVE CAREERS

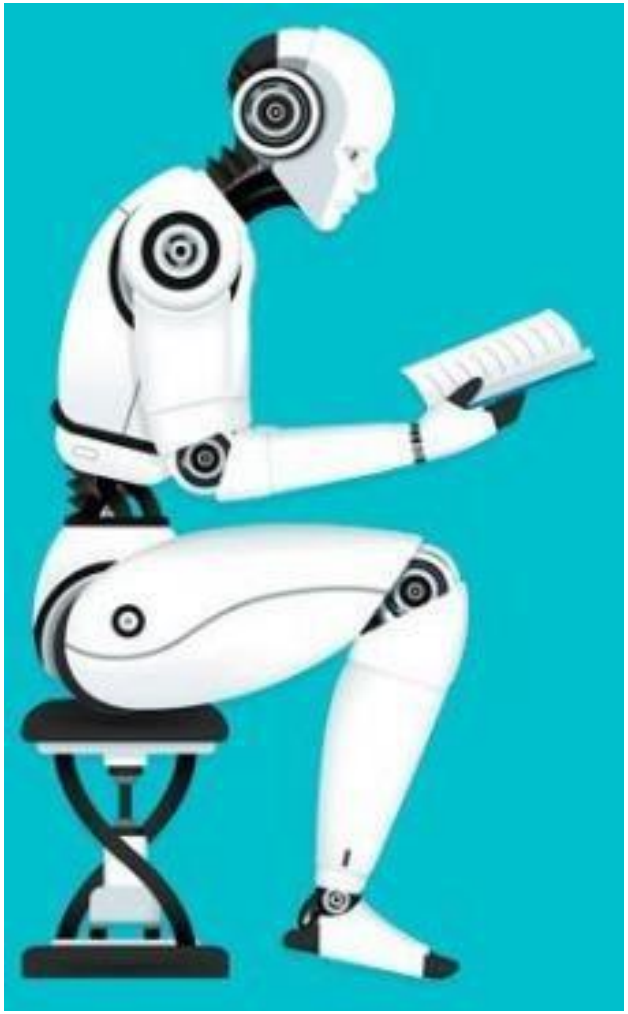
Session_7

LIVEWIRE
FOR LIVE CAREERS

SINGLE and MULTI LINEAR REGRESSION

Introduction to ML

- Learning is the conversion of experience into expertise or knowledge. Its an algorithm based technology to solve any problem.

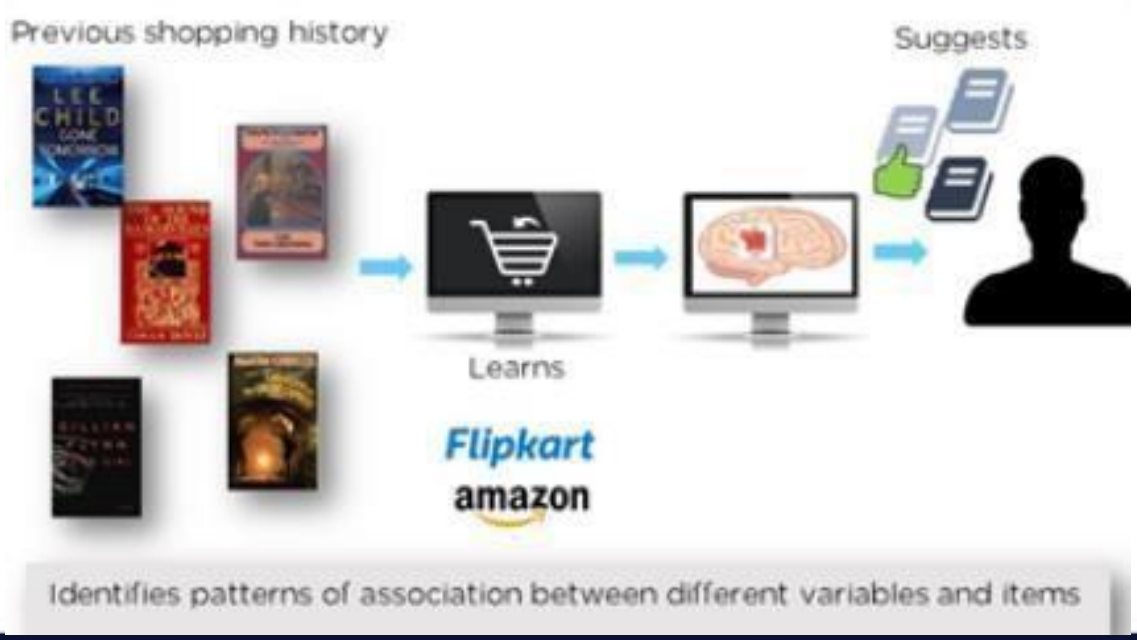


Requirement of ML

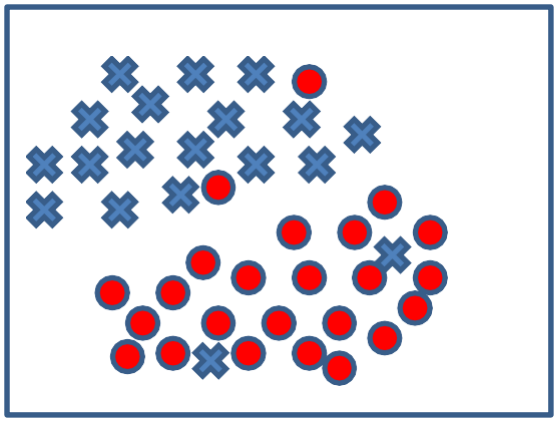
- Programming too complex task
  - Tasks performed by humans
    - whatever task we do in routine are not that easy to elaborate well to transform it into a program.
  - Tasks beyond human capacities
    - Highly complex task
- Adaptivity
  - Changes in the prediction as per the change in data

What is ML

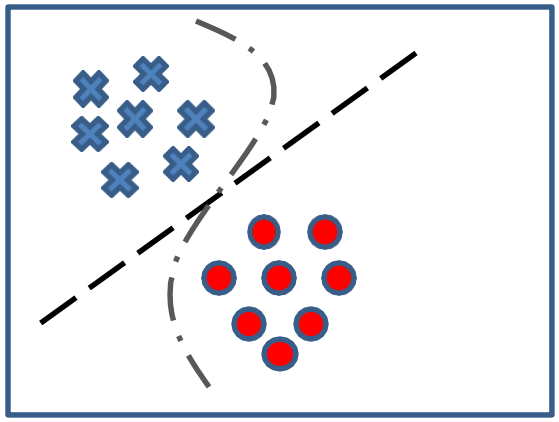
- It is a branch of artificial intelligence, focus on the design and development of algorithm that allow computers to evolve behaviours based on empirical data.
- Example: The person who interested in buying a T-shirt can also buy jeans(**www.amazon.in**)



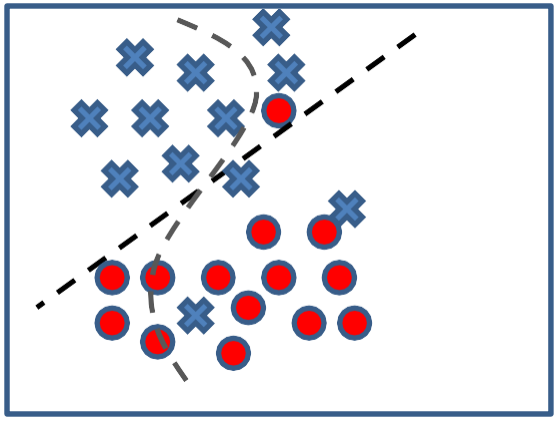
Training and Testing set



Universal set
(unobserved)

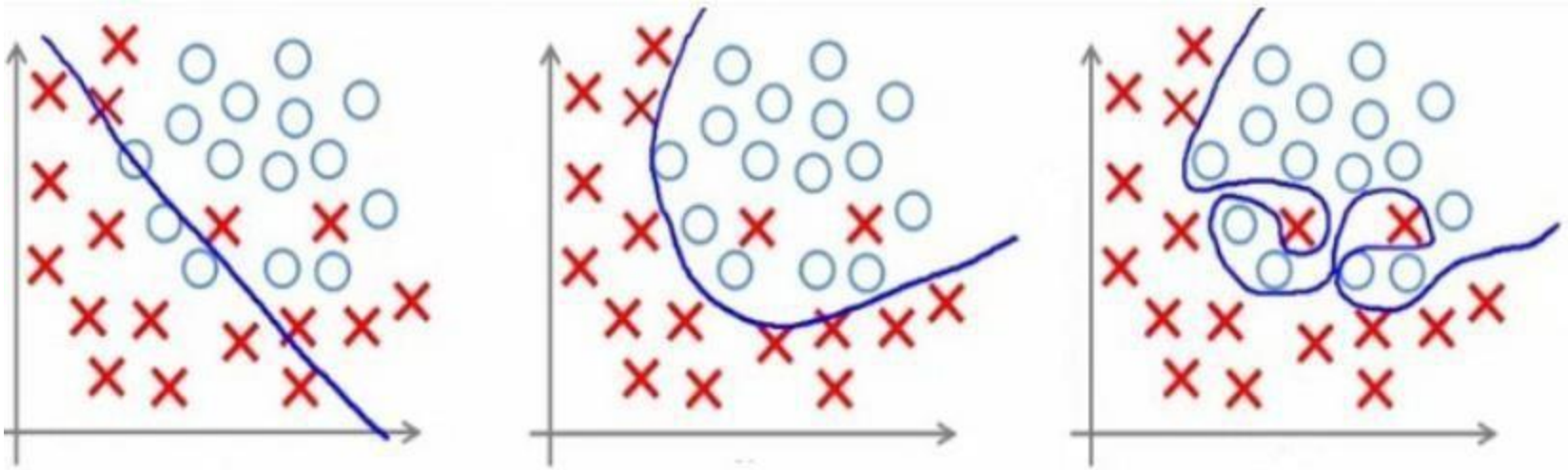


Training set (observed)



Testing set (unobserved)

Over-fitting & Under-fitting



Under-fitting

(too simple to explain the variance)

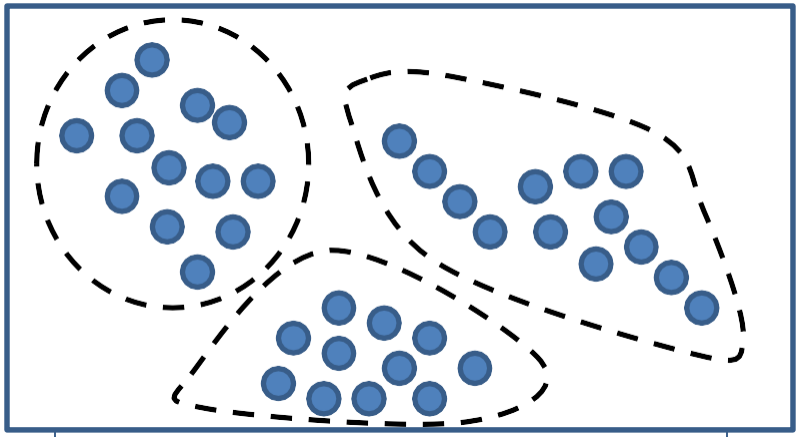
Appropriate-fitting

Over-fitting

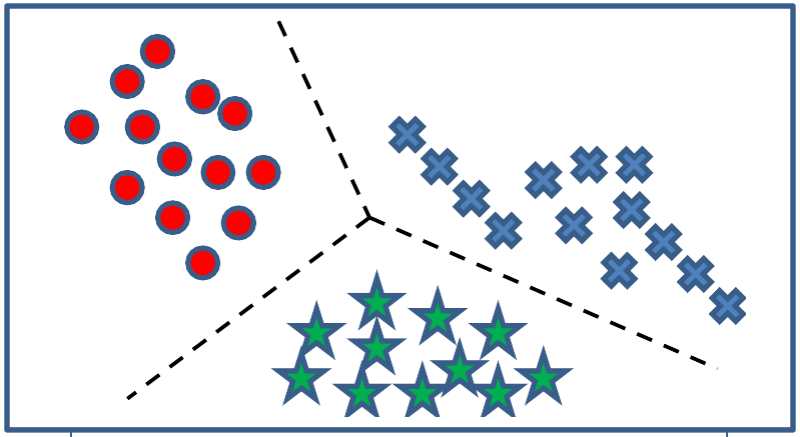
(forcefitting -- too good to be true)

Types of learning

- Supervised learning
  - Classification(For dataset with binary dependent variable)
  - Regression(For dataset with continuous dependent variable)
- Unsupervised learning



Unsupervised learning



Supervised learning

REGRESSION ANALYSIS

- Regression model relates *Y* to a function of X and β:

  Y~f(X, β)

  β=Unknown parameter

  X=Independent Variable

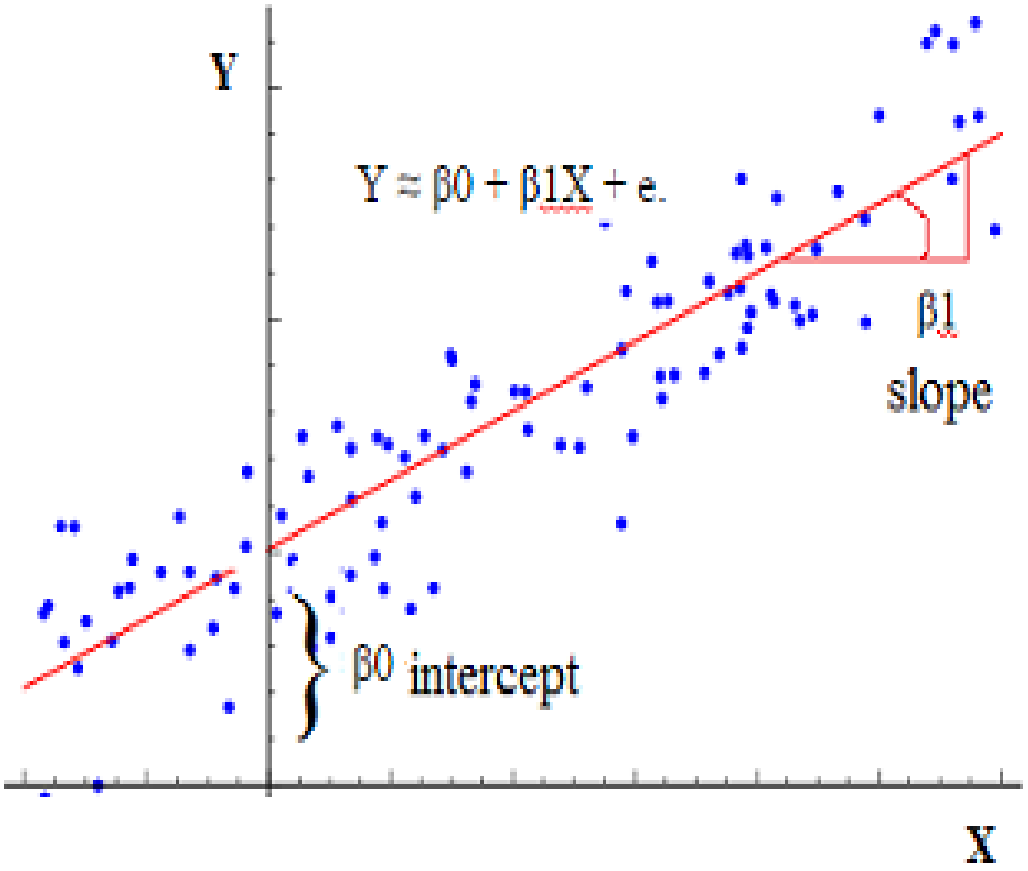
  Y=Dependent Variable

- It will give 2 information:

  β_0 - Intercept

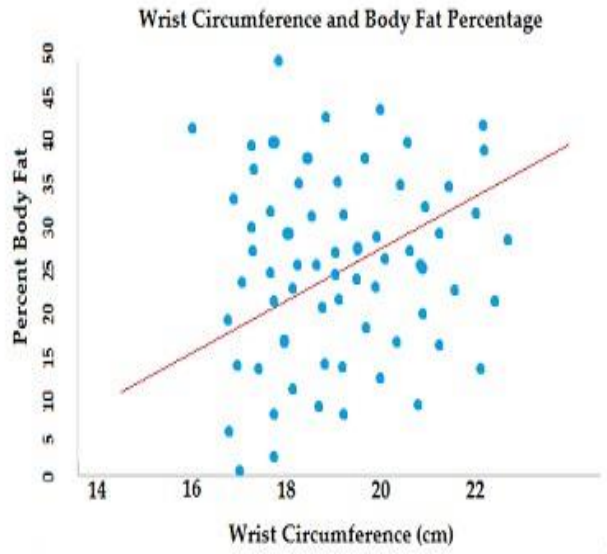
  β_1 _ Slop of the line

Simple Linear Regression - Model



Multi Linear Regression - Model

- Attempts to **model** the relationship between two or more explanatory (independent) variables and a response variable by fitting a **linear** equation to observed data
- Every value of the independent variable x is associated with a value of the dependent variable y



case study

LIVEWIRE
FOR LIVE CAREERS

Session_8

LIVEWIRE
FOR LIVE CAREERS

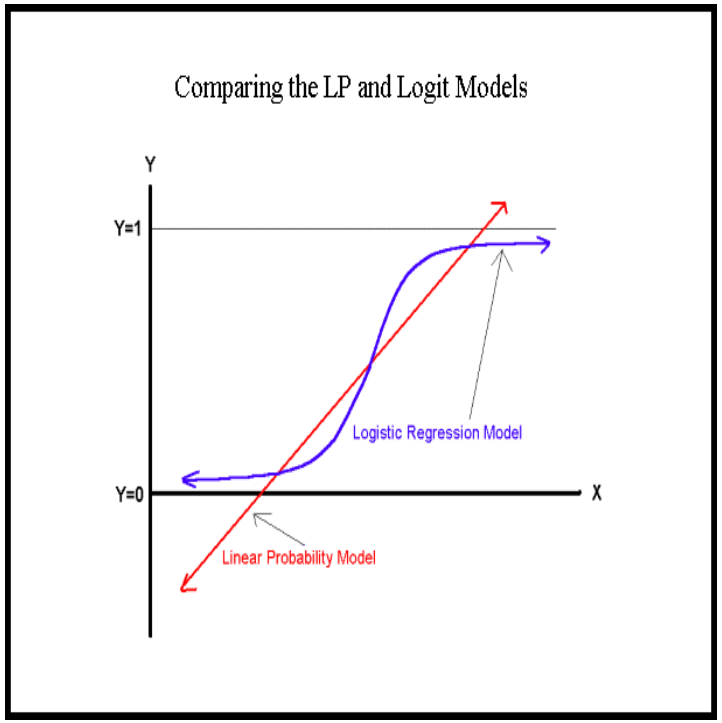
LOGISTIC REGRESSION

Logistic Regression

- Logistic regression is a predictive analysis
- Explain the relationship between one dependent binary variable (Category) and one or more independent variables
- Example of binary variable: pass/fail, win/lose, alive/dead or healthy/sick

Logistic Regression…..

- To draw linear line, convert the data points into linear format then convert into original probability



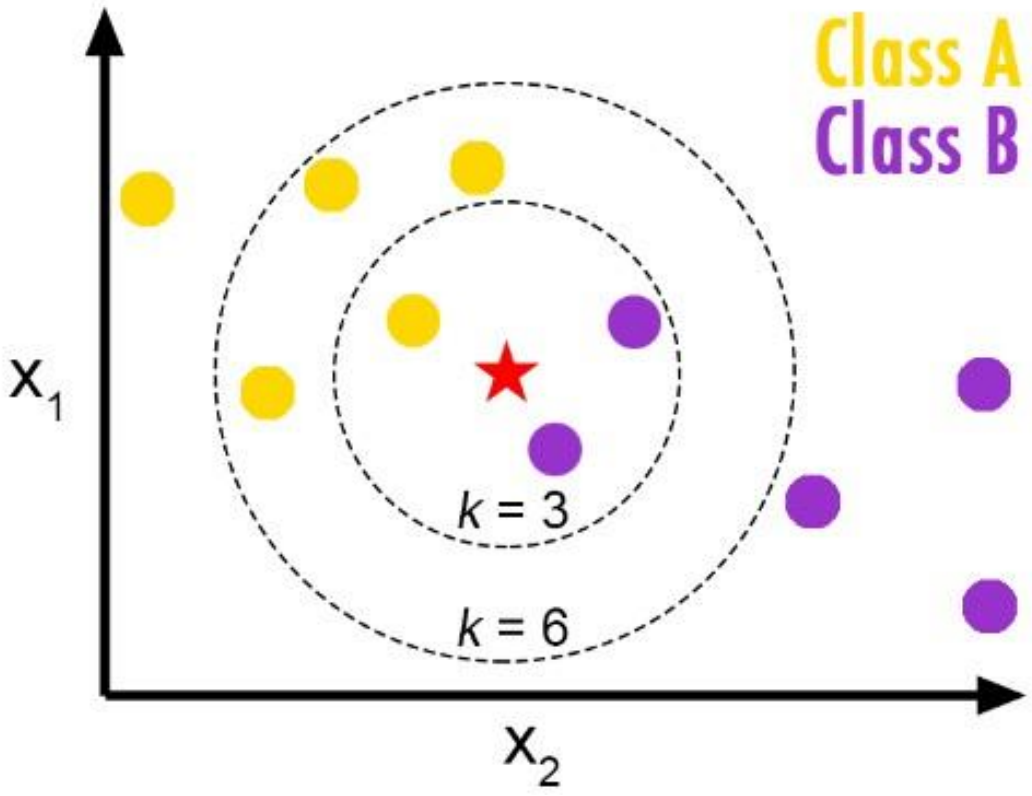
case study

LIVEWIRE
FOR LIVE CAREERS

Session_9

KNN

- Classification algorith
  - KNN

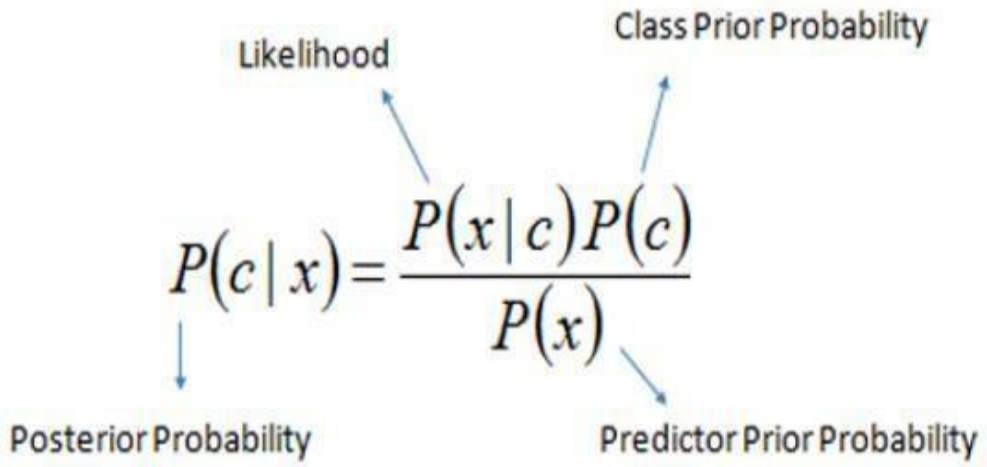


case study

LIVEWIRE
FOR LIVE CAREERS

NAÏVE BAYES

- Linear classifier: Naïve Bayes classifier
- GaussianNB
- MultinomialNB



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

case study

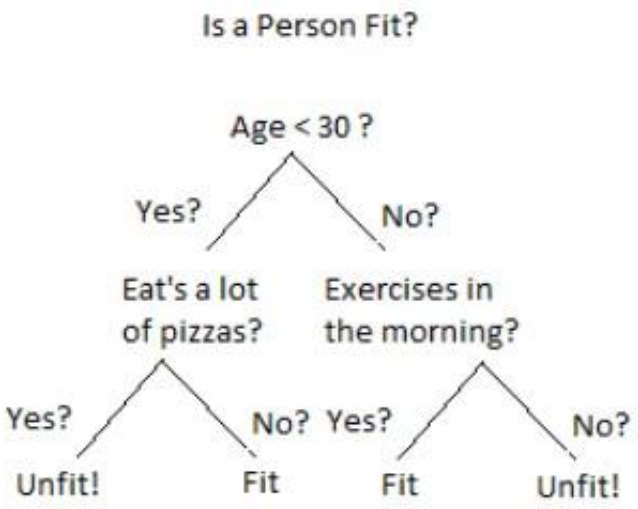
LIVEWIRE
FOR LIVE CAREERS

Session_10

LIVEWIRE
FOR LIVE CAREERS

DECISION TREE

- Decision Tree
  - ENTROPY



$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

- GINI

C1	0
C2	6

$p(C1) = 0/6 = 0 \quad p(C2) = 6/6 = 1$

$Gini = 1 - p(C1)^2 - p(C2)^2 = 1 - 0 - 1 = 0$

C1	1
C2	5

$p(C1) = 1/6 \quad p(C2) = 5/6$

$Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278$

C1	2
C2	4

$p(C1) = 2/6 \quad p(C2) = 4/6$

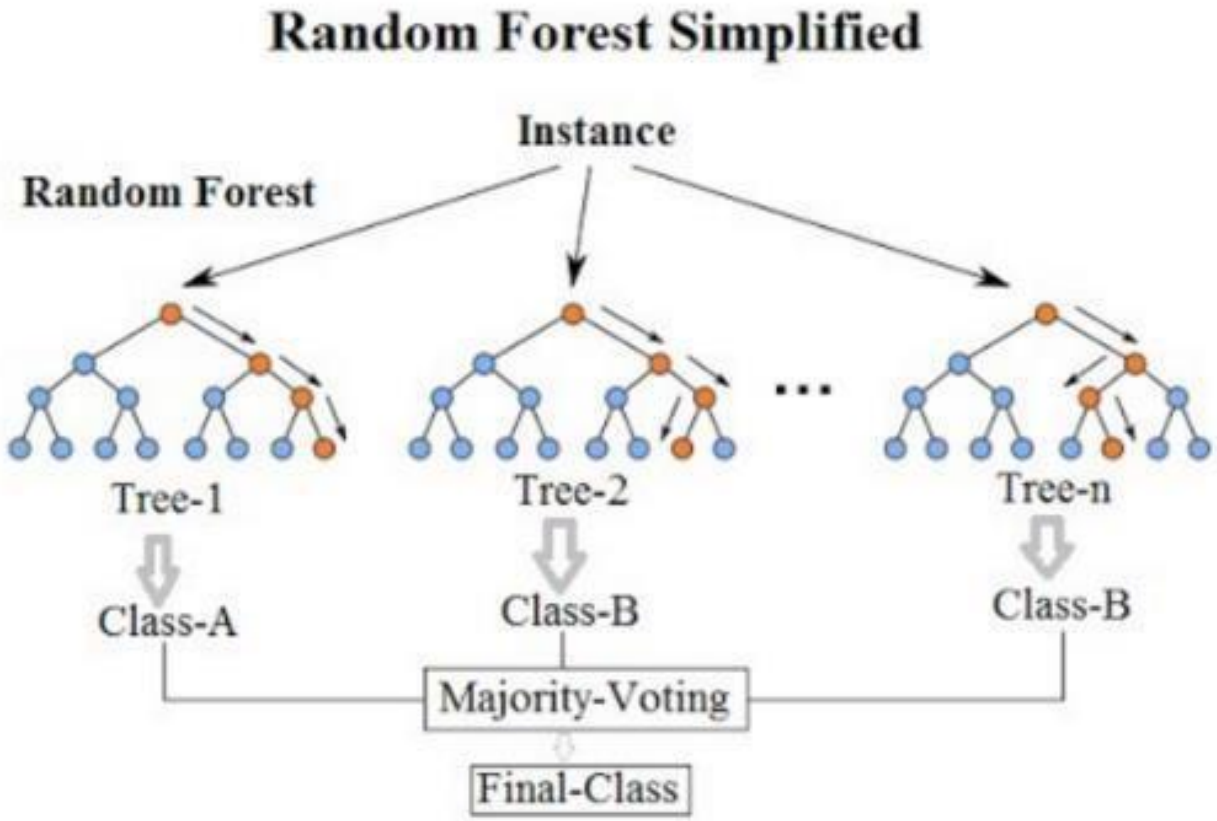
$Gini = 1 - (2/6)^2 - (4/6)^2 = 0.444$

case study

LIVEWIRE
FOR LIVE CAREERS

RANDOM FOREST

- Random Forest



case study

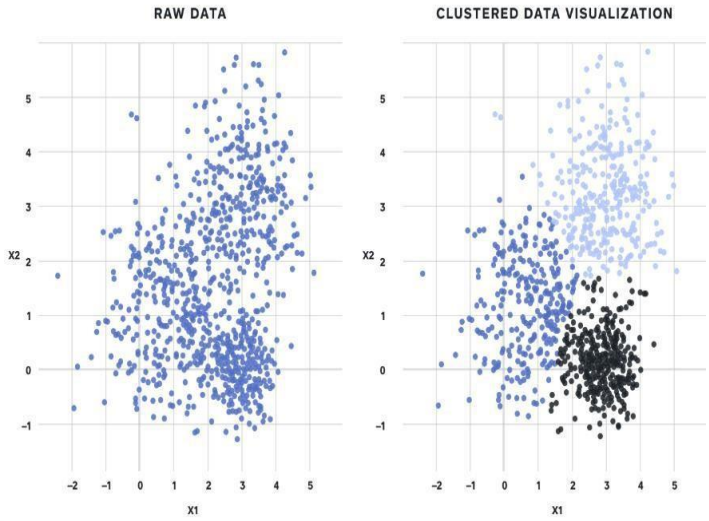
LIVEWIRE
FOR LIVE CAREERS

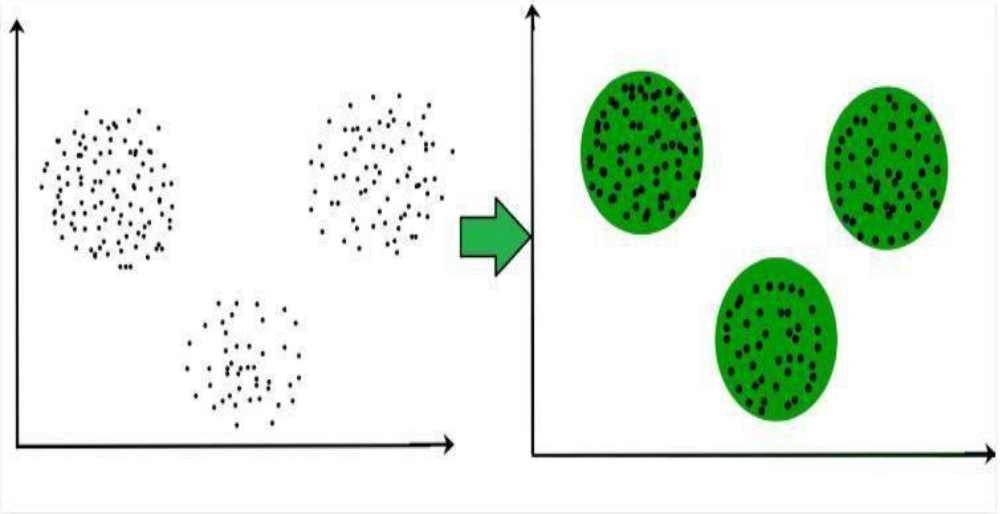
Session_11

LIVEWIRE
FOR LIVE CAREERS

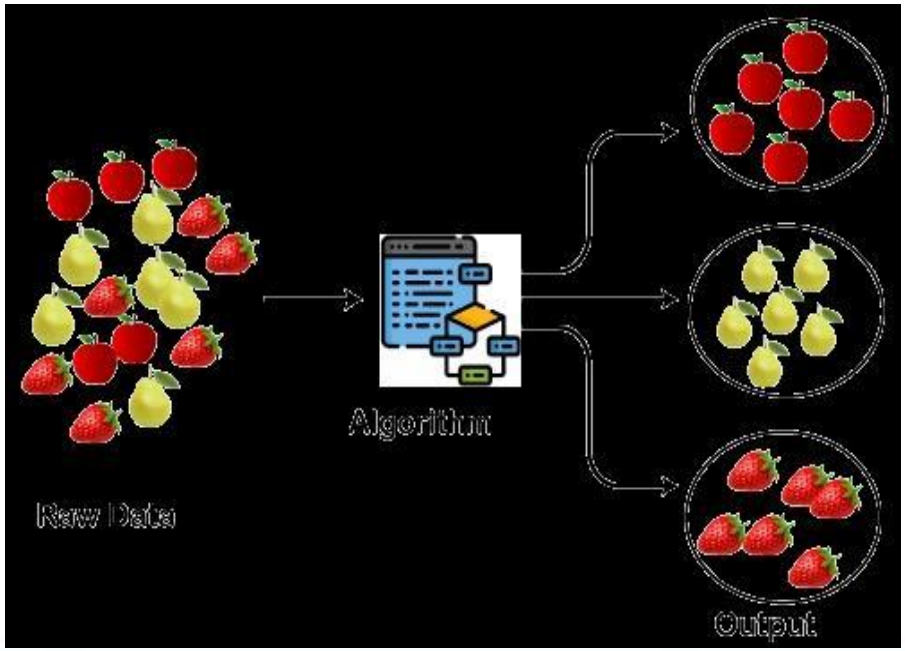
CLUSTERING

Clustering





- Clustering or cluster analysis is a machine learning technique, which **groups the unlabelled dataset**. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points.



Types of Clustering

Roughly clustering can be divided into two subgroups

- Hard Clustering-
  - In hard clustering, each data point either belongs to a cluster completely or not.
  - Strict partitioning clustering: each object belongs to exactly one cluster.
- Soft Clustering-
  - Instead of putting each data point into a separate cluster, a probability of that data point to be in those clusters is assigned.
  - Each object belongs to each cluster to a certain degree.

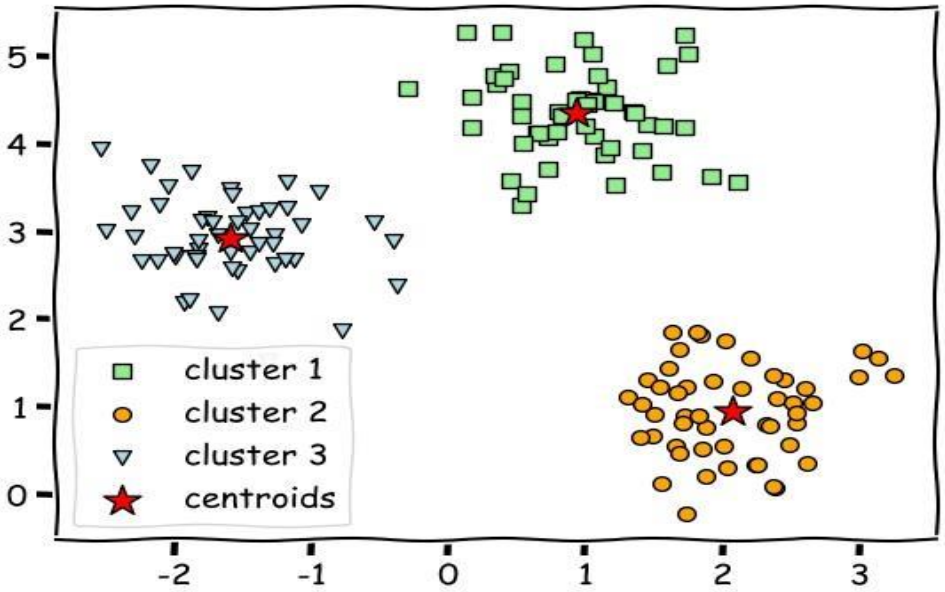
Clustering Algorithms

- Hierarchial Clustering
- K-means Clustering
- DBSCAN
- K-medoids

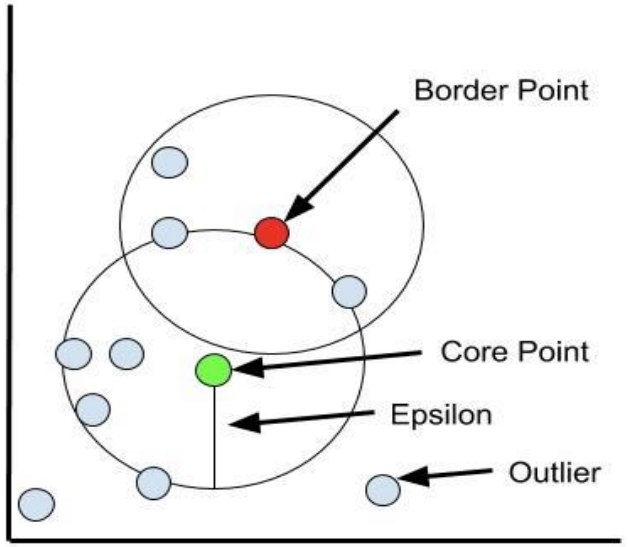
- Hierarchical clustering is **another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics**

- **There are two types of hierarchical clustering: divisive (top-down) and agglomerative (bottom-up).**
- **A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.**

k-means is **a technique for data clustering that may be used for unsupervised machine learning**. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k).



Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is **a base algorithm for density-based clustering**. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.



k -medoids is **a classical partitioning technique of clustering that splits the data set of n objects into k clusters, where the number k of clusters assumed known a priori** (which implies that the programmer must specify k before the execution of a k -medoids algorithm).

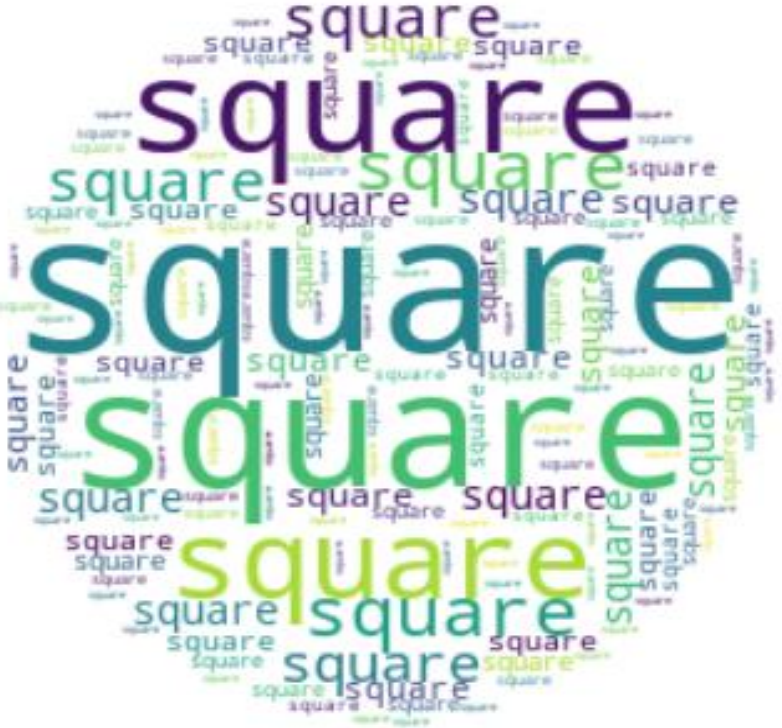
```
from sklearn_extra.cluster import KMedoids
import numpy as np
X = np.asarray([[1, 2], [1, 4], [1, 0], ... [4, 2], [4, 4], [4, 0]])
kmedoids = KMedoids(n_clusters=2, random_state=0).fit(X)
kmedoids.labels_
array([0, 0, 0, 1, 1, 1])
kmedoids.predict([[0,0], [4,4]]) array([0, 1])
kmedoids.cluster_centers_ array([[1, 2], [4, 2]])
kmedoids.inertia_
8.0
```

case study

LIVEWIRE
FOR LIVE CAREERS

WORDCLOUD

```
Import numpy as np
import matplotlib.pyplot as plt
from wordcloud import WordCloud
text = “square”
x, y = np.ogrid[:300, :300]
mask = (x – 150) ** 2 + (y – 150) ** 2 > 130 ** 2
mask = 255 * mask.astype(int)
wc = WordCloud(background_color=“white”, repeat=True, mask=mask)
wc.generate(text)
plt.axis(“off”)
plt.imshow(wc, interpolation=“bilinear”)
plt.show()
```


square

Word Cloud

- Novelty visual representation of text data

LIVEWIRE
FOR LIVE CAREERS

Session_12

LIVEWIRE
FOR LIVE CAREERS

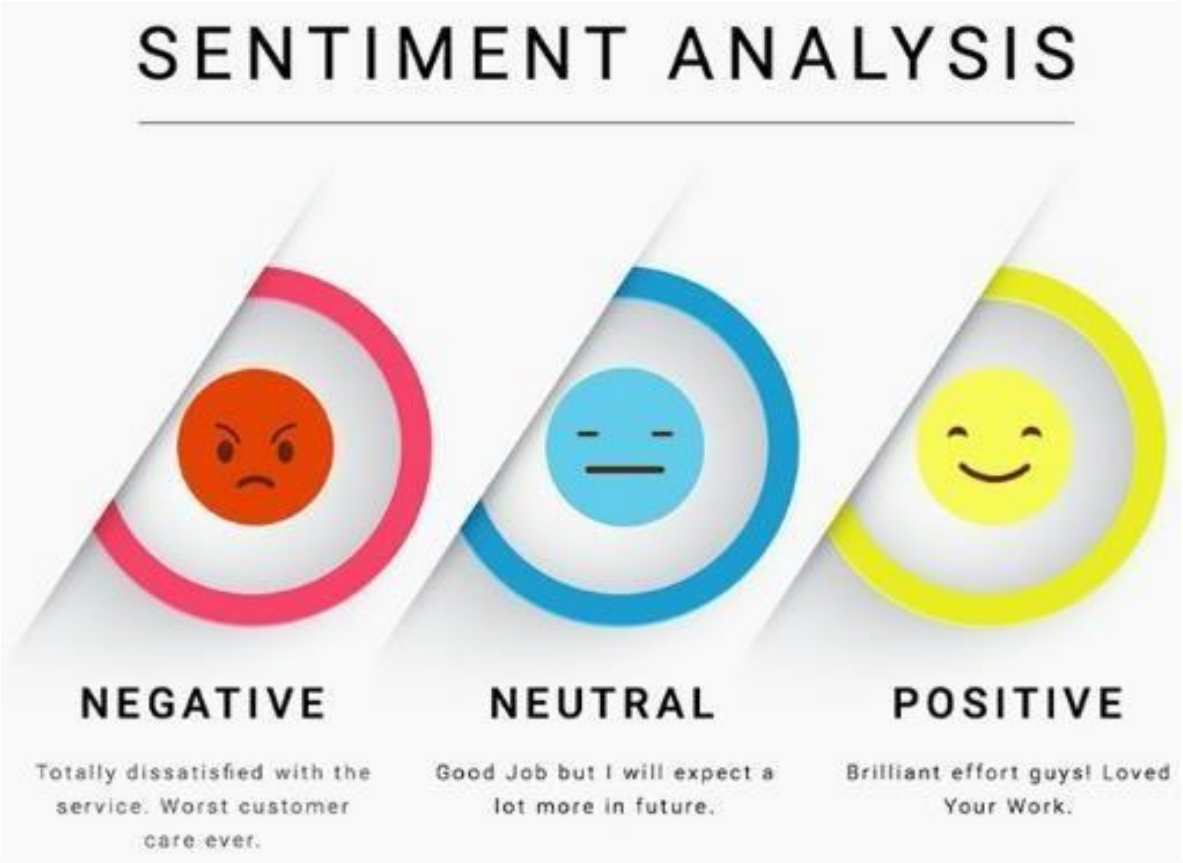
TEXT MINING

Text mining

- It is the process of extracting important information from a natural text. It is unstructured, amorphous and difficult to deal with algorithmically.
- Goal: turn the text into data for analysis using natural language processing.

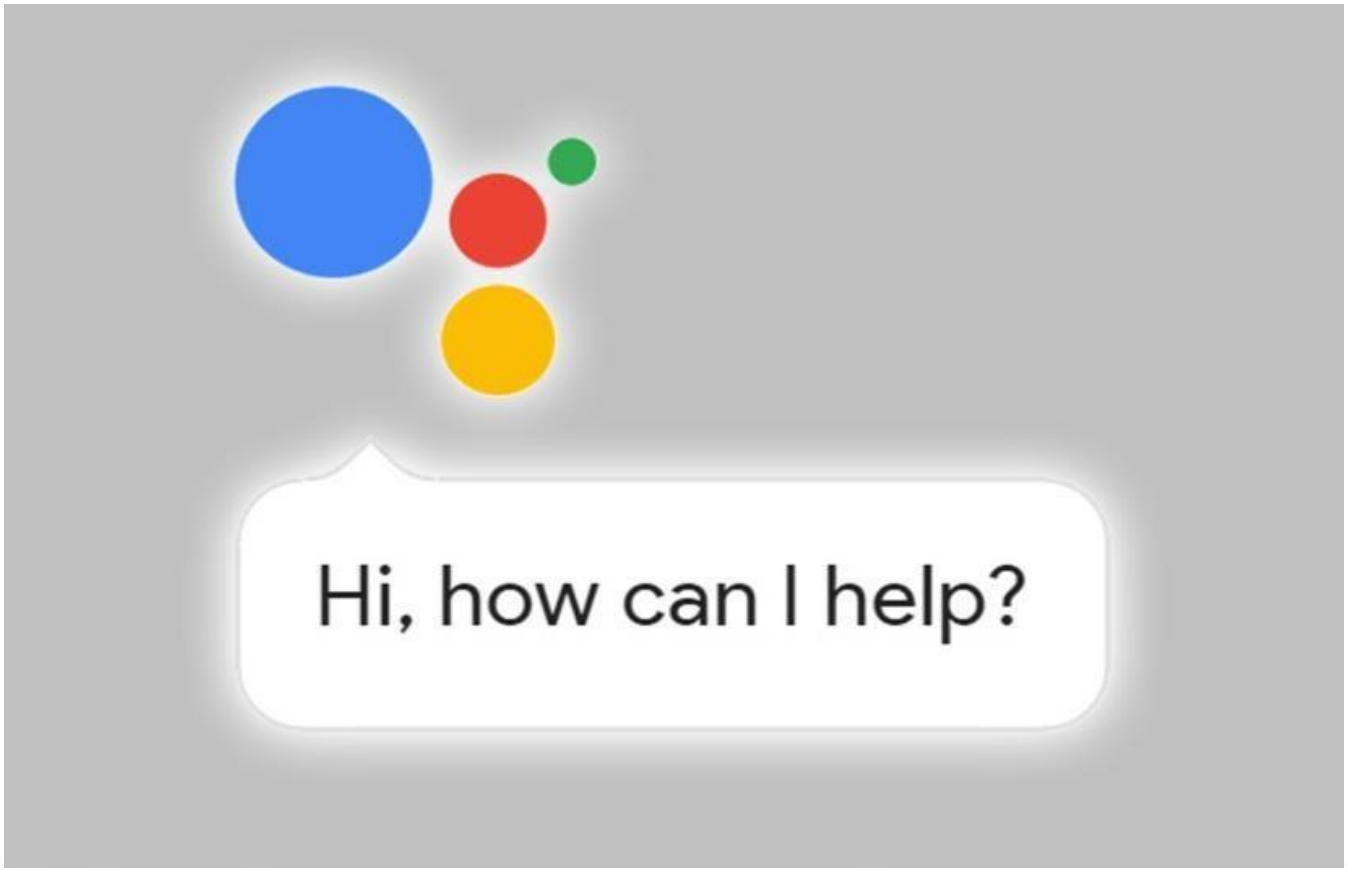
Application of text mining

- Sentimental analysis(Positive, negative, neutral views)



Application of text mining

- Speech recognition(speech to understandable text)



case study

LIVEWIRE
FOR LIVE CAREERS

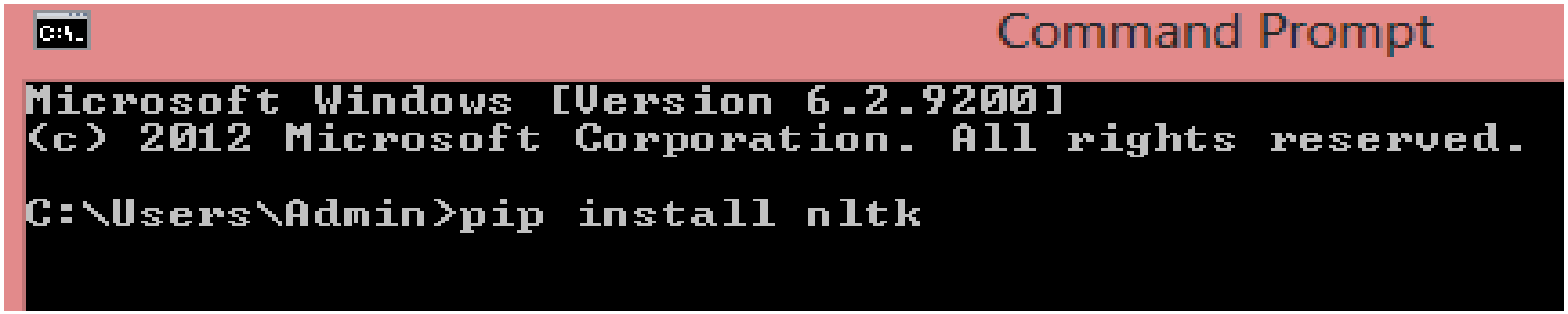
NLP

LIVEWIRE
FOR LIVE CAREERS

Introduction to NLP

Libraries and modules

- Download

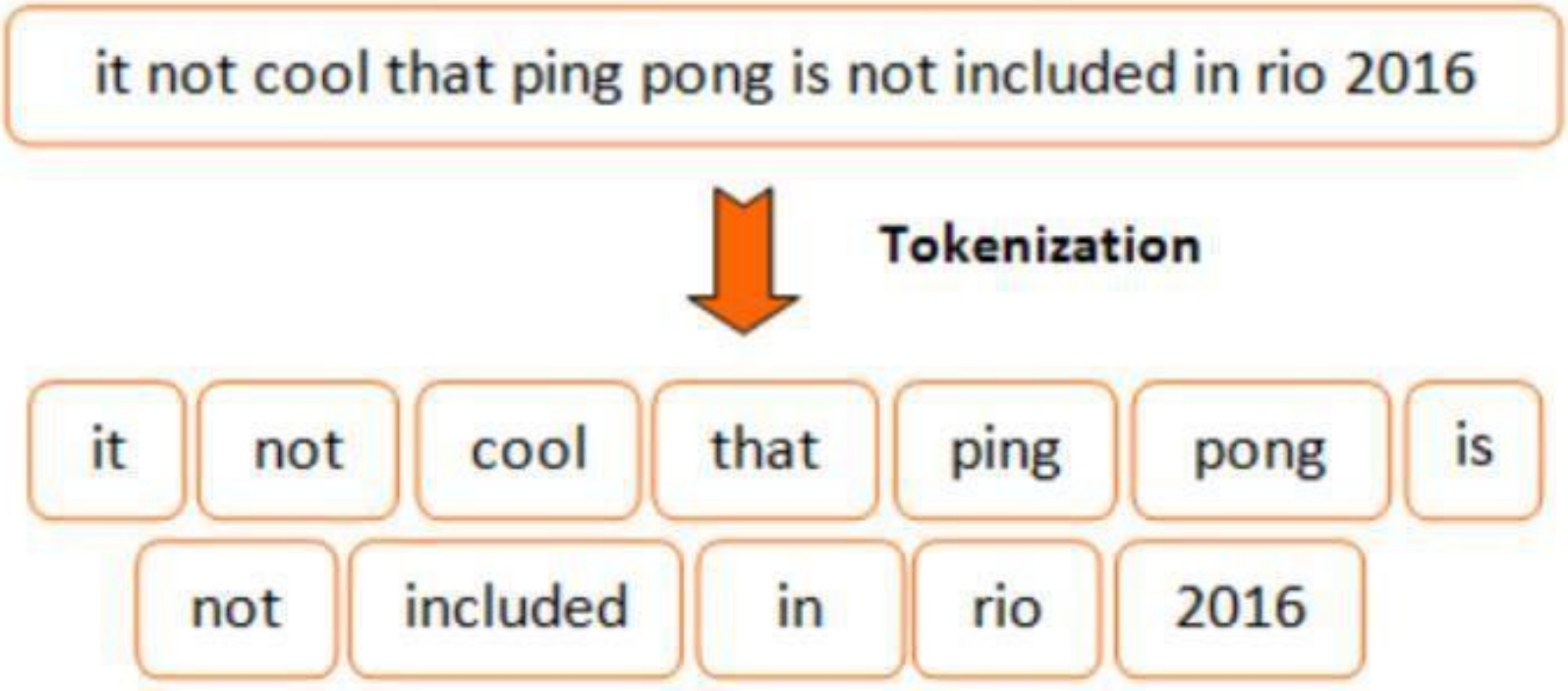


- Import using python tool and download rest packages

```
>>>import nltk
>>>nltk.download()
```

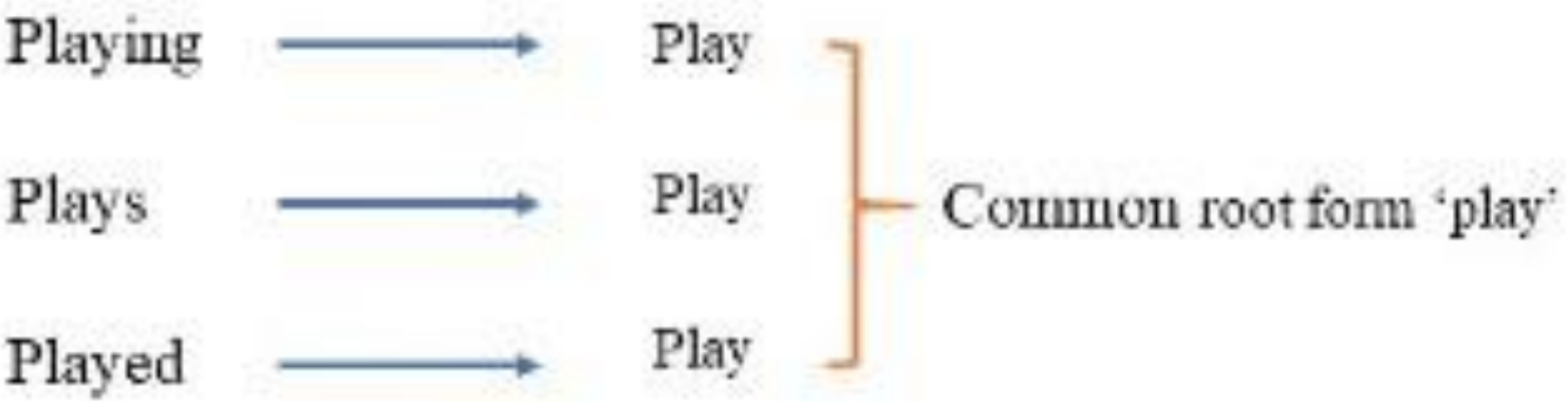
NLP process

- Tokenization
  - Covert a string into small tokens.
  - Library to import nltk.tokenize.word_tokenize



NLP process

- Stemming
  - PORTERSTEMMER
  - LANCASTERSTEMMER



Porter stemmer

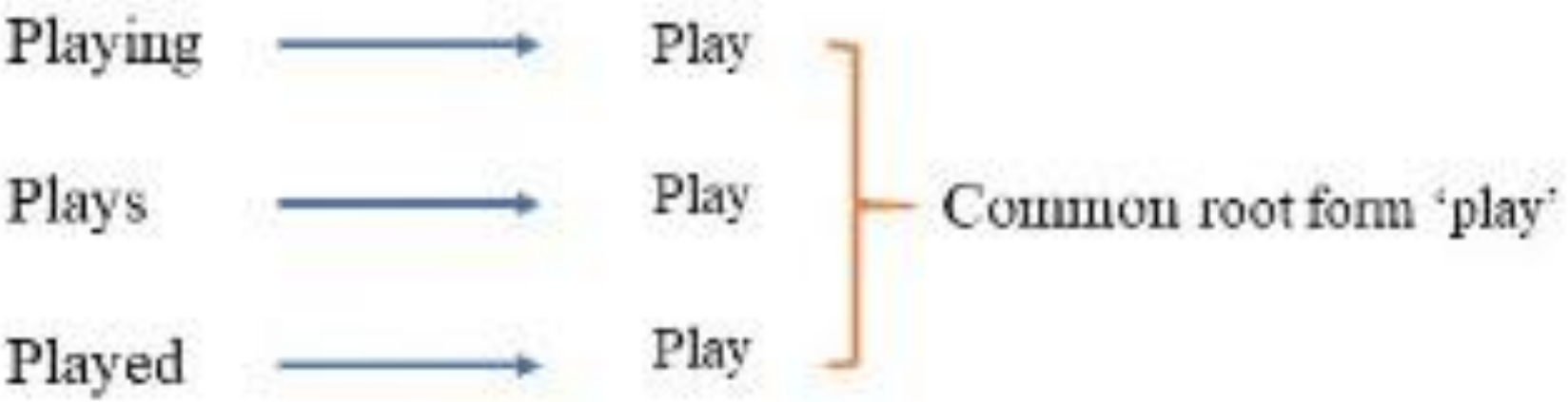
The Porter stemming algorithm (or ‘Porter stemmer’) is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems

LancasterStemmer

LancasterStemmer is a module in NLTK that implements the Lancaster stemming technique. Allow me to illustrate this with an example. In the example below, we construct an instance of LancasterStemmer() and then use the Lancaster algorithm to stem the list of words

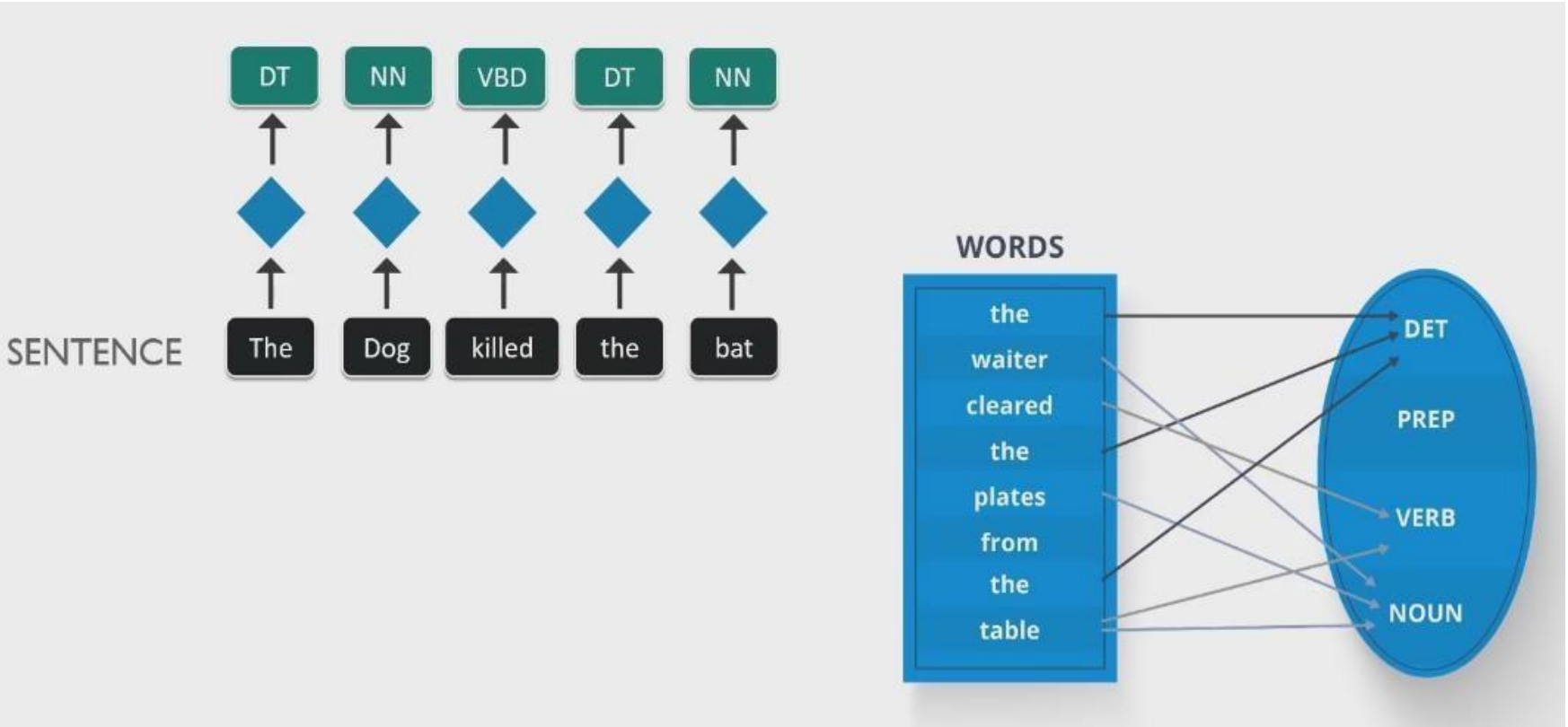
NLP process

- Lemmanization
  - It produces a proper word using POS tags.



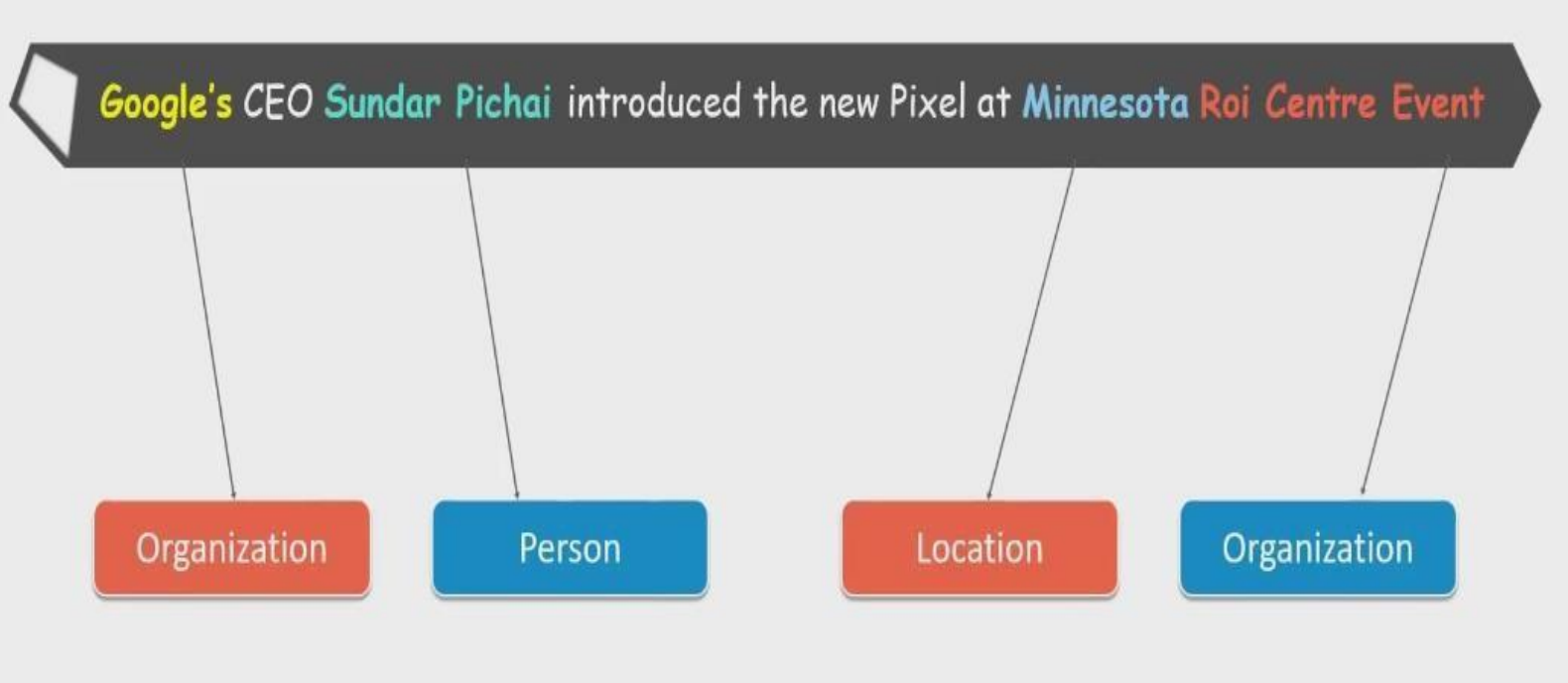
NLP process

- Parts of speech



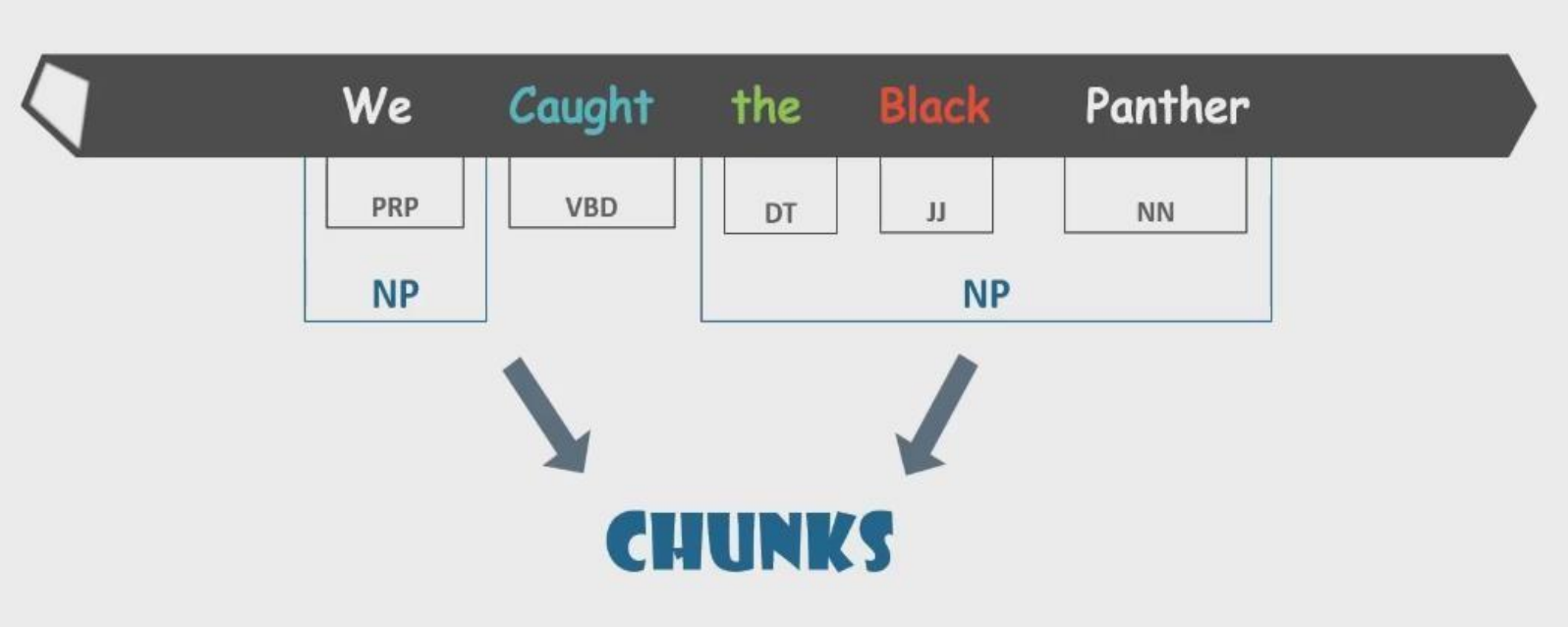
NLP process

- Name entity recognition



NLP process

- Chunking



case study

LIVEWIRE
FOR LIVE CAREERS

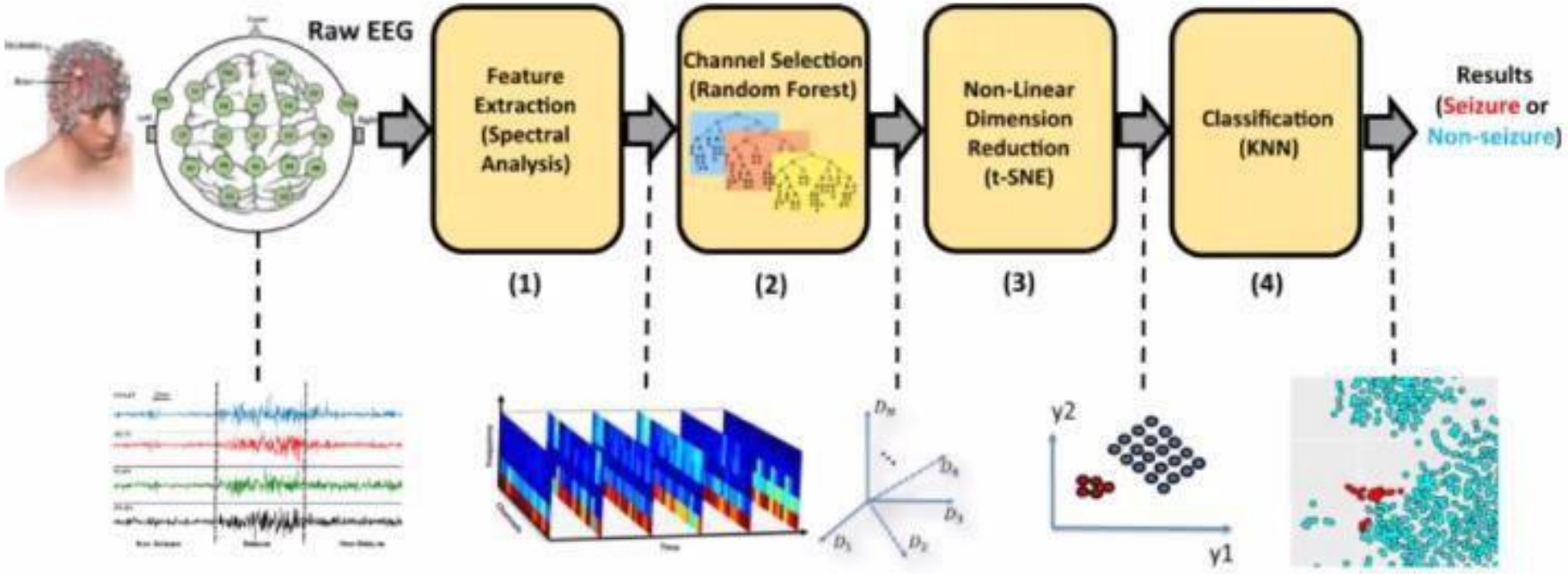
Session_13

LIVEWIRE
FOR LIVE CAREERS

DIMENSIONAL REDUCTION

Dimension Reduction

- Reduction of number of random variables.



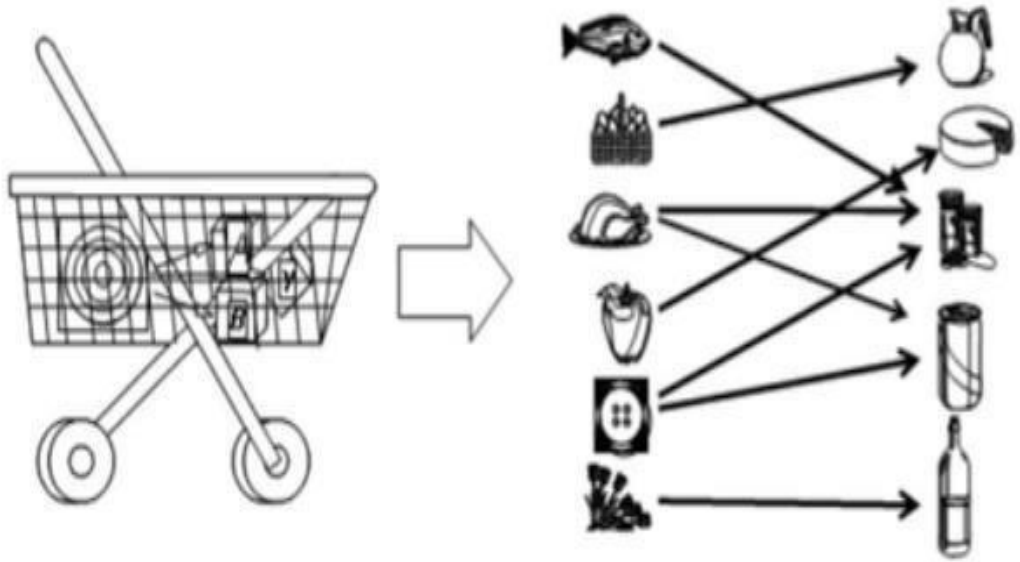
Dimensionality reduction, or **dimension reduction**, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its Intrinsic dimension.

case study

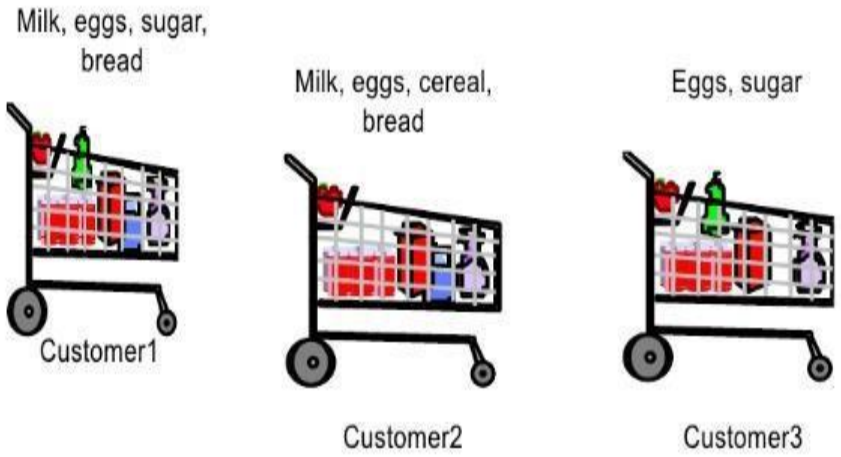
LIVEWIRE
FOR LIVE CAREERS

ASSOCIATION RULE MINING

- Association rules are if-then statements that help to show the probability of relationships
- an antecedent (if)
- a consequent (then)



98% of people who purchased items A and B also purchased item C



- What do my customer buy? Which product are bought together?
- **Aim:** Find associations and correlations between the different items that customers place in their shopping basket

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

frequent item-set = | 3 | 4 | 7 | count = 3

[0] [1] [2]

(min conf = 0.700)

	combination	antecedent		candidate rule	confidence	
k = 1	0	3	count = 6	"if 3 then (4 7)"	conf = 3 / 6 = 0.50	✗
	1	4	count = 5	"if 4 then (3 7)"	conf = 3 / 5 = 0.60	✗
	2	7	count = 4	"if 7 then (3 4)"	conf = 3 / 4 = 0.75	✓
k = 2	0 1	3 4	count = 4	"if (3 4) then 7"	conf = 3 / 4 = 0.75	✓
	0 2	3 7	count = 4	"if (3 7) then 4"	conf = 3 / 4 = 0.75	✓
	1 2	4 7	count = 3	"if (4 7) then 3"	conf = 3 / 3 = 1.00	✓

case study

LIVEWIRE
FOR LIVE CAREERS

Session_14

LIVEWIRE
FOR LIVE CAREERS

FORECASTING

Forecasting

- Time series Analysis

  - Analyse the changes happened in one variables over period of time

  - More suitable for short-term projections (ie. 5 to 6 year time series data)

  - Should have clear trend, pattern and stability

Decomposition of components

- When we decompose the time series data we will get all the time series components
- Two methods of decomposing:

- Time series additive model (Follow the constant pattern)

(ie. Trend + Seasonality + Regular)

- Multiplicative model (Do not follow constant pattern)

(ie. Trend x Seasonality x Irregular)

Techniques/method of forecasting

- Simple Moving Average
- Weighted Moving Average
- Exponent Smoothing
- Holt'z Exponent Smoothing
- Holt'z Winter Smoothing
- Autoregressive Integrated Moving Average (ARIMA)

case study

LIVEWIRE
FOR LIVE CAREERS

THANKS