Welcome
to

Three week

Summer Project Training and Internship Programme

on

"Machine Learning and Data Science using Python"

14 May 2018 – 30 May 2018

Day-1

Module-1

Introduction to Data Science and Machine Learning

What is Data Science?- *The History and Evolution*

- **Data science** is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.
- The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy".
- In 1974, Naur published *Concise Survey of Computer Methods*, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.
- In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference ("Data Science, classification, and related methods"), after the term was introduced in a roundtable discussion by Chikio Hayashi

Is Data Science so Powerful?

- Most information in the world being shared digitally, and more still being stored on cloud, almost every industry is sitting on a treasure trove of data.
- If mined properly, such data can reveal information on customer preferences, tastes, usage in real-time, and much more.
- Most organizations still don't know how to use up to 80% of their data, since most data exists in an unstructured format.
- The key concern is how organisations can use their imagination to pick what data to correlate and analyse.

Use Case 1 : Real Time Analytics by Chicago

- The City of Chicago cuts crime and improves citizen welfare with a real-time geospatial analytics platform called Windy Grid, which pulls together seven million different pieces of data from city departments every day.
- With real-time data becoming more and more available and easily affordable, real time analytics can be what a company may need to push it from just about okay to exceptionally good at understanding themselves, how they're being perceived, their customers' wants and needs and much more.

- Banks are making out of real time analytics to better engage with their customers.

➢ 81% of large banks treat customer centricity as their top priority.

➢ 1 out of every 2 executives believe that they do not have mature capabilities to support their customer strategies.

The solution to the above:

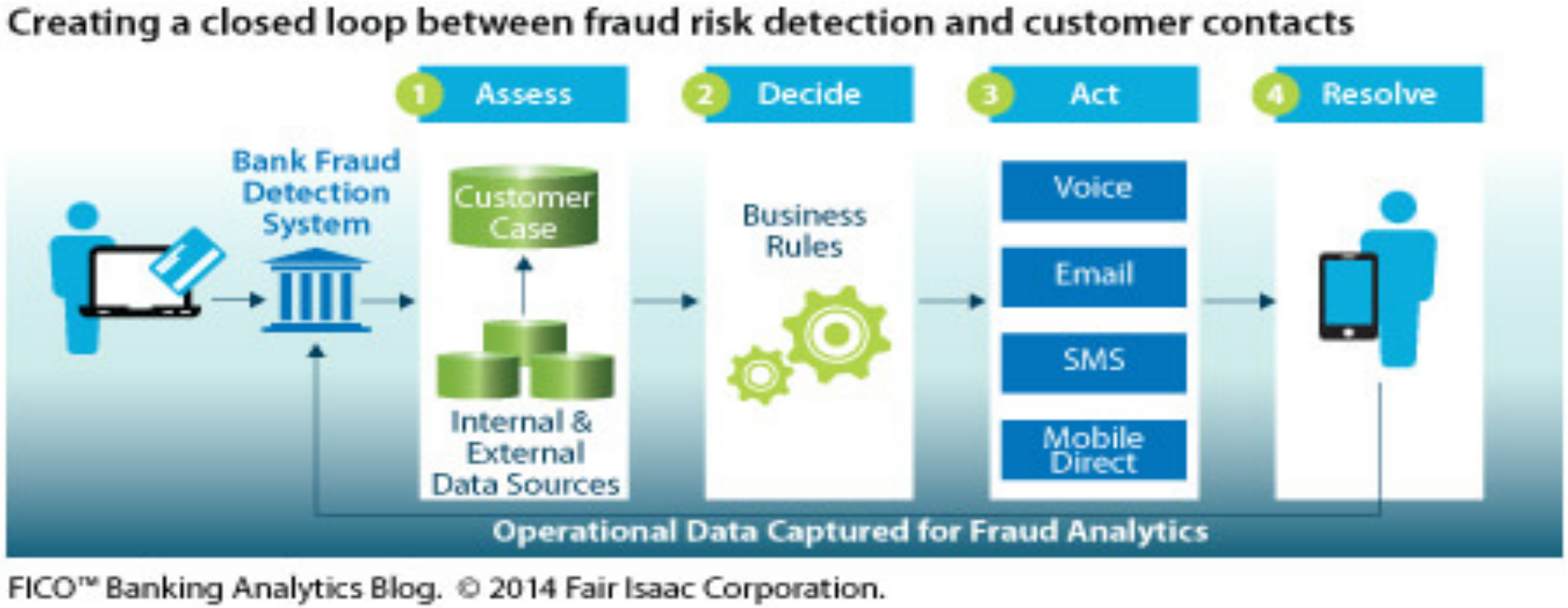
1. Don’t only act on data from months, weeks or even days ago but also respond to changes that occur minute by minute.
2. Real time analytics helps to understand how to better engage with customer.

How does technology help?

1. Real time data processing and analysis
2. Data visualization
3. Predictive Analytics

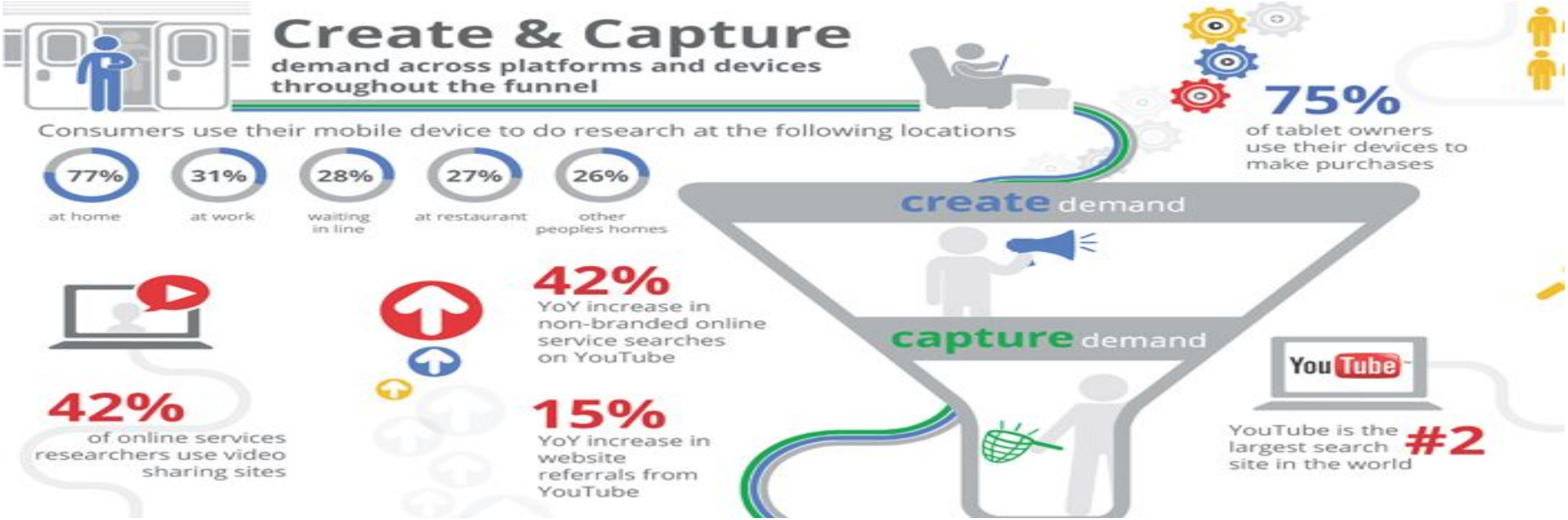
Use Case 2: Fraud Detection

- Financial institutions monitor people spending habits on a real-time basis.
- Banks and credit card companies collect a lot more information from location, life style, people tastes, income, account balances, employment details, credit history and transaction history.
- Fraud instances can be similar in content and appearance but rarely are identical.
- Data companies have to constantly update themselves with the new techniques of fraud detection.



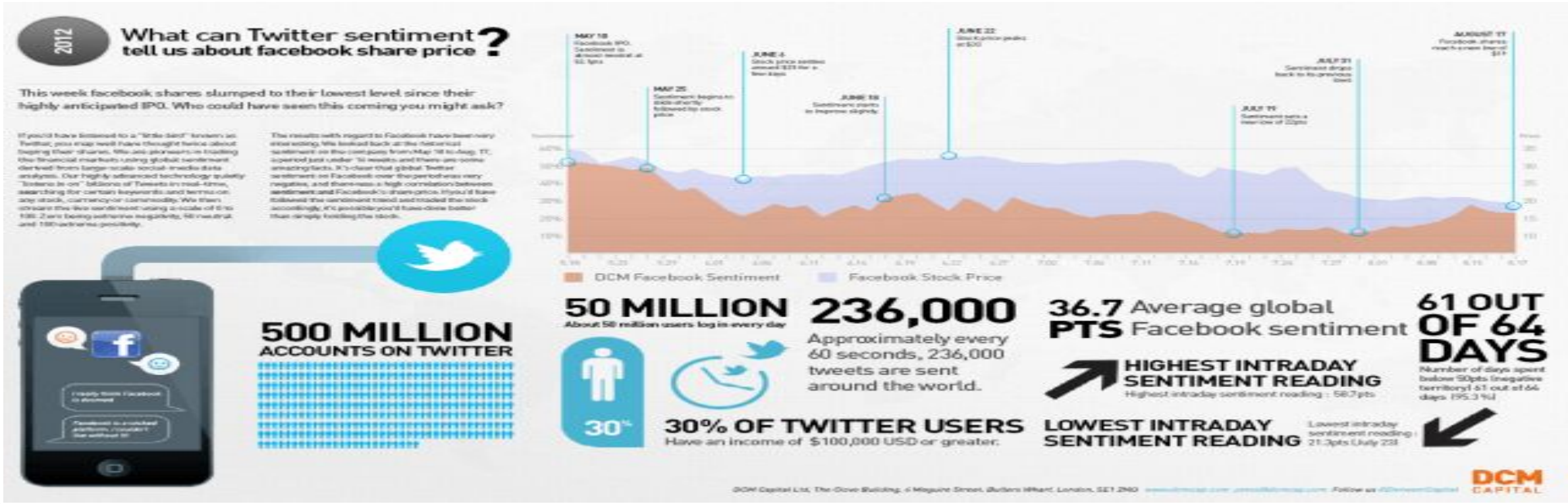
Use Case 3: Click Stream Analysis

- Clickstream data is the trail of digital breadcrumbs left by users as they click their way through a website, and it's loaded with valuable customer information for businesses.
- Clickstream analysis is the process of collecting, analyzing, and reporting aggregate data about which pages visitors visit, and in what order.
- This reveals usage patterns, which in turn gives a heightened understanding of customer behavior. This use of the analysis creates a user profile that aids in understanding the types of people that visit a company's website.



Use Case 4: Sentiment Analysis

- Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.
- Using sentiment analysis techniques, companies can respond to negative (or positive) brand perception.
- When a company releases a new product, monitoring and analyzing social media content can play a large role in quickly remediating bugs and errors.
- PR for political figures and celebrities depends heavily on sentiment analysis and how the person is perceived by people on social media.



Use Case5: Analytics for Customer Loyalty Program

- Retaining current customers has become a high priority for every business.
- Loyalty programs have sprung up in every consumer-related industry, from retail to restaurants, cruise lines to charge cards.
- Simply a loyalty program cannot differentiate a company from its competitors.
- Customers expecting discounts and deals wait for special offer sales and use membership cards and simulates a price-shopping environment for the business rather than increasing their loyalty.
- Companies that wish to have strong loyalty programs can rely on customer analytics to drive their strategies and create measurable business impact.

Business Case

- Analytics for Motor and Pump Monitoring

Click here

Life of a Data Scientist

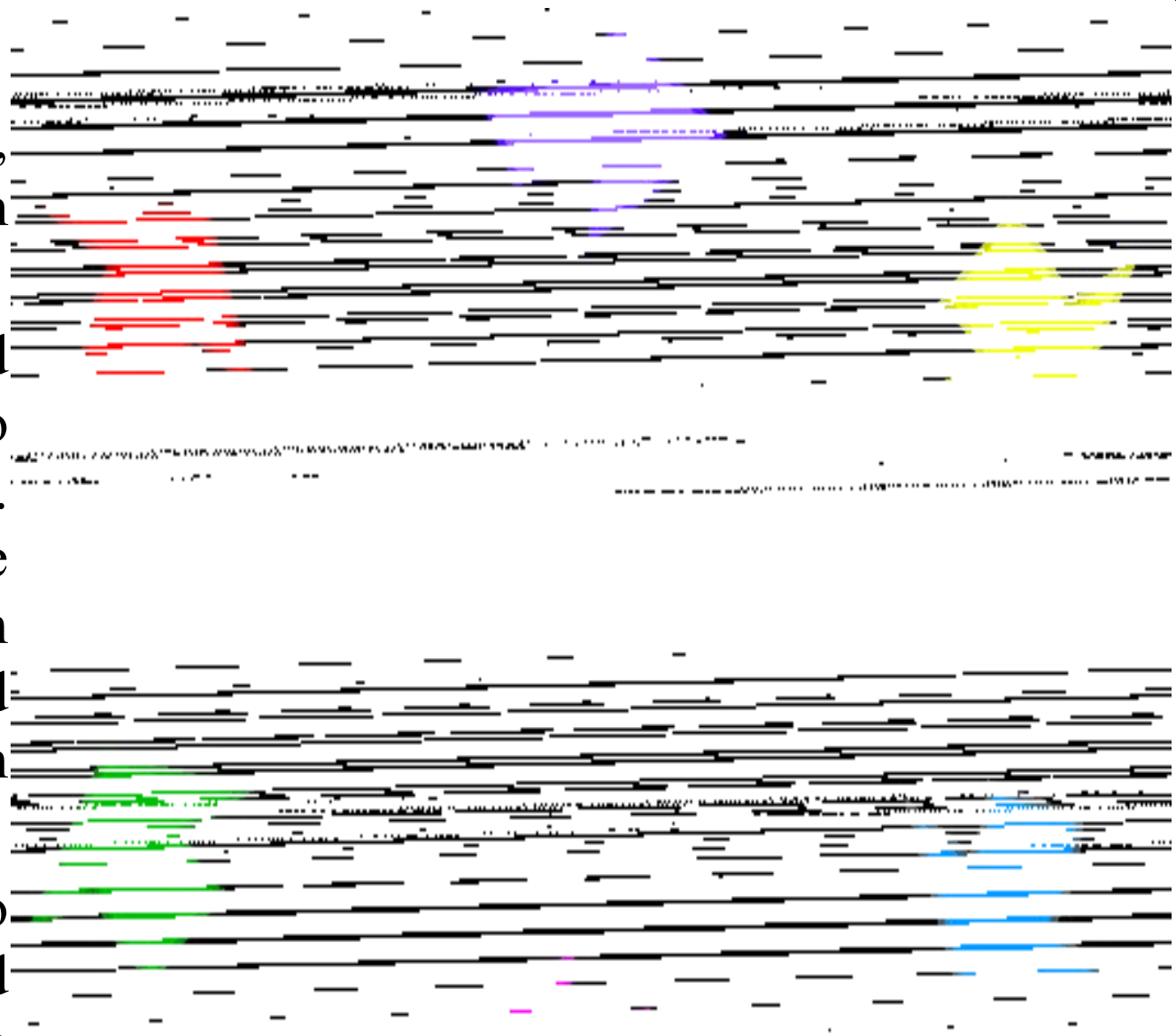
Click here

Module-2

Statistics

Introduction to Statistics

- Statistics is the science of collecting, analyzing and making inference from data.
- A useful branch of mathematics used by researchers in many fields to organize, analyze, and summarize data. Statistical methods and analyses are often used to communicate research findings and to support hypotheses and give credibility to research methodology and conclusions.
- It is important for researchers and also consumers of research to understand statistics so that they can be informed, evaluate the credibility and usefulness of information, and make



About Data

- Data is a vital element that hits any organization(Business/Government/Scientific sector).
- Data is a 1st class citizen.
- Data becomes Unstructured in Enterprise Scale when observed in terms of volume, velocity, variety and veracity.
- Demand for skills in ‘Data Science’ is unprecedented in sectors where value, competitiveness and efficiency are driven by data.

Statistics-Basic Terminology

Term	Meaning
Population	It is the collection of all individuals or items under consideration in a statistical study.
Sample	It is a subgroup from the population that the researcher studies in order to make inferences about the population.
Parameter	A descriptive number about a population. Features of the population can be summarized by *numerical parameters*.
Descriptive Statistics	The branch of statistics devoted to the summarization and description of data is called descriptive statistics. ***Tools used:*** Graphs, Charts and table and descriptive measures such as averages, measures of variation and percentiles.
Inferential Statistics	The branch of statistics concerned with using sample data to make an inference about a population of data is called inferential statistics. ***Tools Used:*** Point estimation, interval estimation and hypothesis testing.
Independent Sampling	Occurs when multiple samples are taken, but each sample has no effect on

Term	Meaning
Random Sampling	Occurs when subjects for the sample are picked totally at random with no other factors influencing their selection.
Stratified Random Sampling	In this process the population is divided into layers based on some criteria and a number of random subjects are taken from each strata

Central Tendency

- Simplest sort of descriptive statistics involves measures of central tendency.
- It is a way of seeing what the aggregate of the data tells us about the data.
- Three most simple measures of central tendency are the mean, median, and mode.

Mean

Mean is simply the arithmetic average.

Mode

Mode is the item in the sample that appears most often

Median

Median is the item that appears at the middle

Ex: Consider a set of test scores	Mean(x)= $\sum x/n$	Mode	Median
65,74,84,84,89,91,93,99,100	Mean = 86.55	84(appears more than once)	89(value at the center)

For even data set median is the mean of two middle most elements

65,74,84,89

Median= (74+84)/2 =79

Spread

- Measures of spread (also called measures of *dispersion*) tells something about how wide the set of data is.
- Common measures of spread are:

1. The range (including the interquartile range and the interdecile range),
2. The standard deviation,
3. The variance,
4. Quartiles.

Range

The range is a basic statistic that tells you the range of values.

Ex:

If your minimum value is $10 and the maximum value is $100 then the range is $90 ($100 – $10).

A similar statistic is the interquartile range, which tells you the range in the middle fifty percent of a set of data; in other words, it's where the bulk of data tends to lie.

Variance

- Variance is the average of the squared differences from the mean.
- Variance is the expectation of the squared deviation of a random variable from its mean.
- It measures how far a set of (random) numbers are spread out from their average value.
- The variance is the square of the standard deviation

Ex:

Consider the following heights of four dogs in mm:

600,470,170,430,300

$$\text{Mean} = \frac{600+470+170+430+300}{5} = 1970/5 = 394$$

$$\text{Variance} = \frac{(600-394)^2+(470-394)^2+(430-394)^2+(300-394)^2}{5} = 108520/5 = 21704$$

Quartiles

- Quartiles divide your data set into quarters according to where those numbers falls on the number line.
- They divide your data into four segments according to where the numbers fall on the number line.
- The four quarters that divide a data set into quartiles are:

1. The lowest 25% of numbers.
2. The next lowest 25% of numbers (up to the median).
3. The second highest 25% of numbers (above the median).
4. The highest 25% of numbers

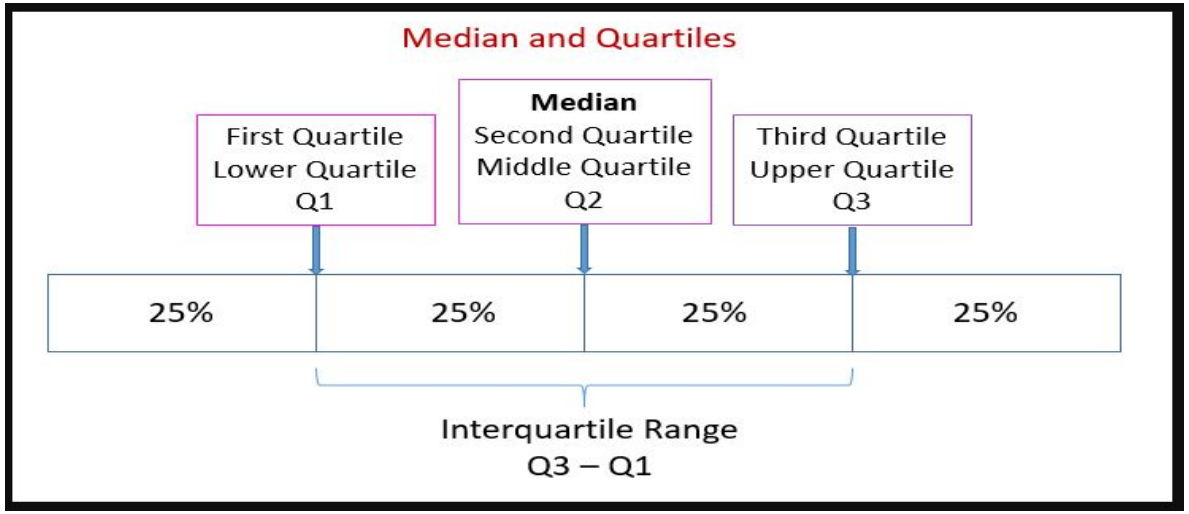
Ex:

Example: Divide the following data set into quartiles: 2, 5, 6, 7, 10, 22, 13, 14, 16, 65, 45, 12.

Step 1: Put the numbers in order: 2, 5, 6, 7, 10, 12 13, 14, 16, 22, 45, 65.

Step 2: Count how many numbers are there in the set and then divide by 4 to cut the list of numbers into quarters. There are 12 numbers in this set, so you would have 3 numbers in each quartile.
2, 5, 6, | 7, 10, 12 | 13, 14, 16, | 22, 45, 65

The **median** divides the data into a lower half and an upper half.
The **lower quartile** is the middle value of the lower half.
The **upper quartile** is the middle value of the upper half.



Ex:

Find the median, lower quartile and upper quartile of the following numbers.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

First, arrange the data in ascending order:

5,7,12,14,15,22,25,30,36,42,53

5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53

lower quartile median upper quartile

If there is an even number of data items, then we need to get the average of the middle numbers.

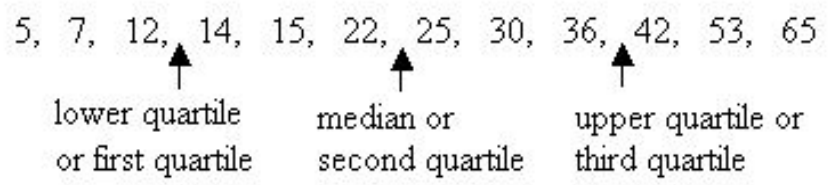
Ex:

Find the median, lower quartile, upper quartile, interquartile range and range of the following numbers.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25, 65

Solution:

Fir



- **Lower quartile or first quartile** = (12+14)/2 = 13
- **Median or second quartile** = (22+25) / 2 =23.5
- **Upper quartile or third quartile** = (36+42) / 2 =39
- **Interquartile range** = Upper quartile – lower quartile = 39 – 13 = 26
- **Range** = largest value – smallest value = 65 – 5 = 60

Standard Deviation

- It tells how tightly all the various data/values in the dataset are clustered around the mean.
- If values in the dataset are pretty tightly bunched together and the bell-shaped curve is steep, then standard deviation is small.
- If values in the dataset are spread apart and the bell curve is relatively flat, then the standard deviation is relatively large.

Ex: Consider the test scores

65,74,84,84,89,91,93,99,100

Mean for above data = 86.5

Standard Deviation = $\sqrt{}$variance/ n - 1

SD= $(65 - 86.5)^2 + (74 - 86.5)^2 + (84 - 86.5)^2 + (84 - 86.5)^2 + (89 - 86.5)^2 + (91 - 86.5)^2 + (93 - 86.5)^2 + (99 - 86.5)^2 + (100 - 86.5)^2 / 9 - 1 = 882/8 = 110.25$

$\sqrt{}110.25 = 10.5$ i.e. various test scores are about 10.5 units from the mean.

End of Day 1