Profession of the Future

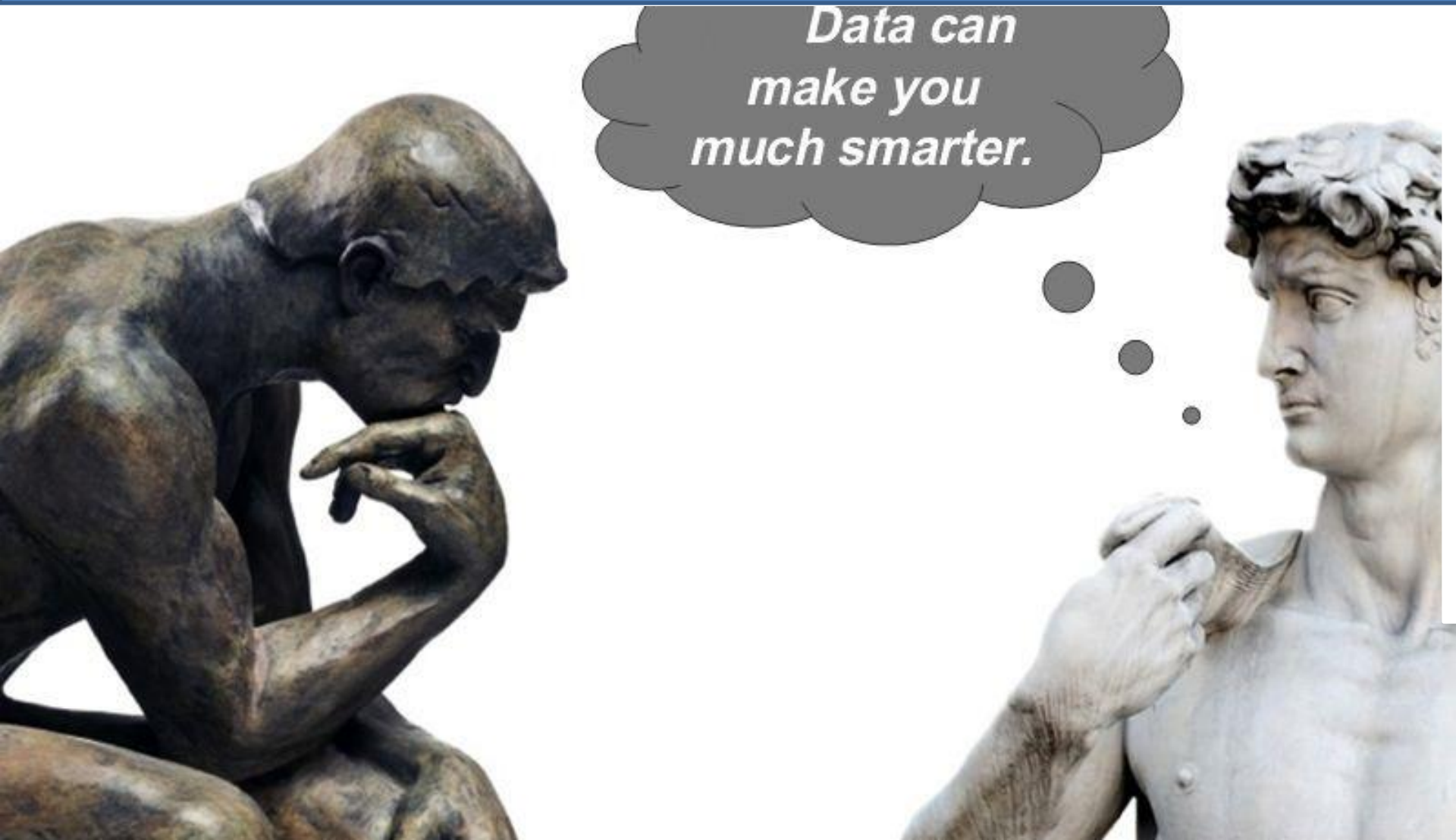
Artificial Intelligence

Data Science & Big Data Analytics


Stay home,
Stay safe
Keep Learning
Become Smarter

An Investment in Knowledge is the Best Investment

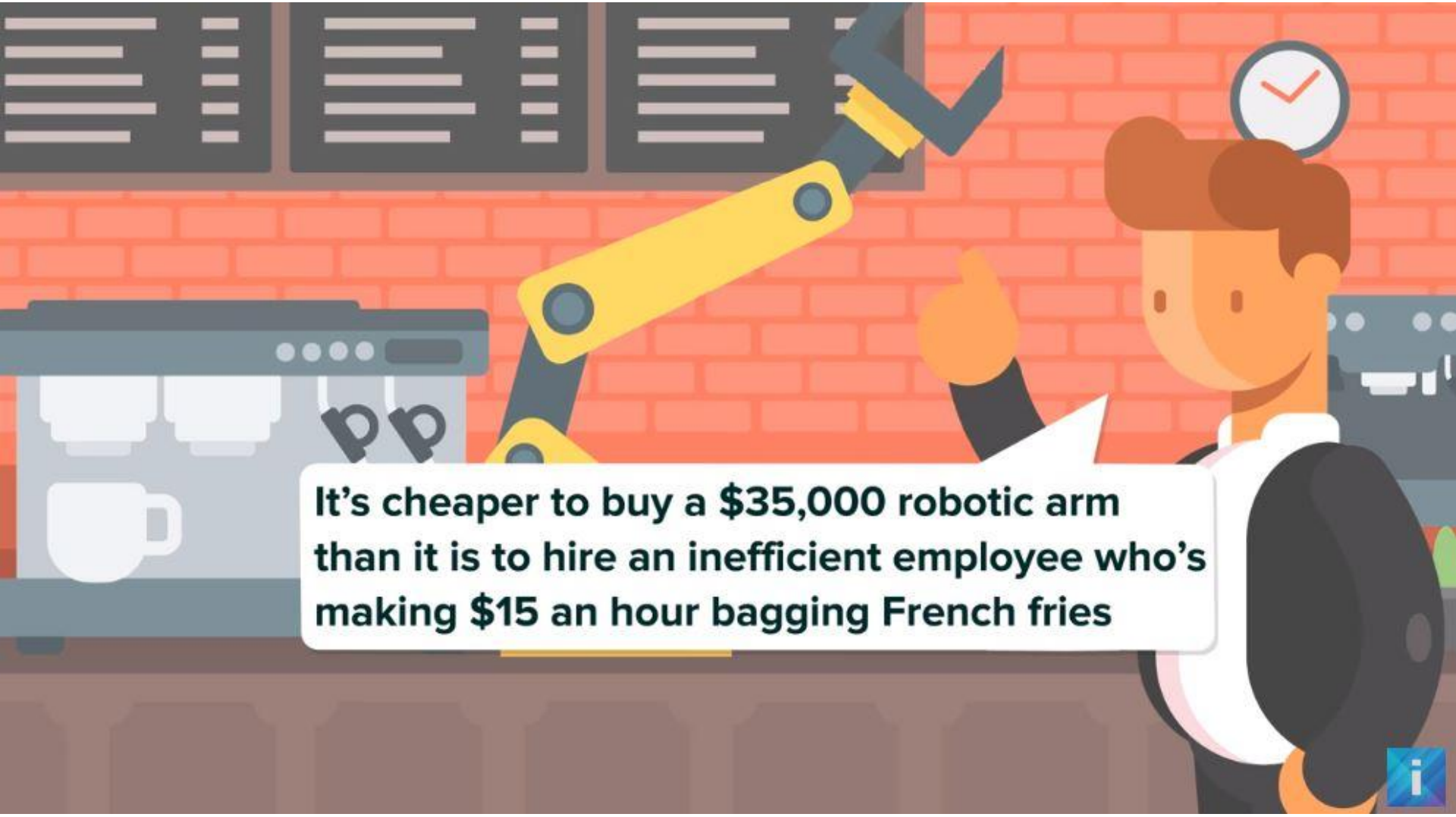
Benjamin Franklin …

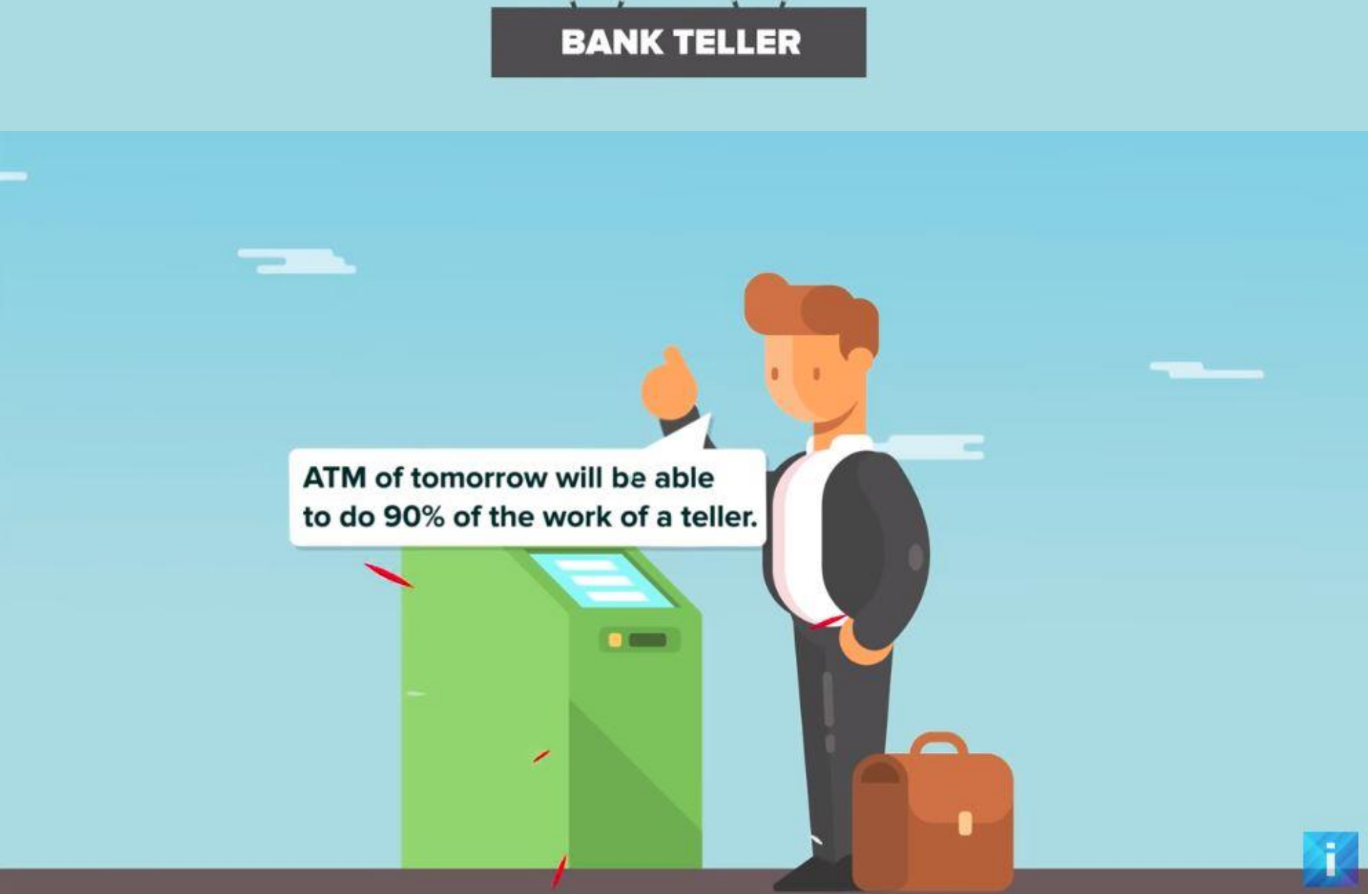


Why ?

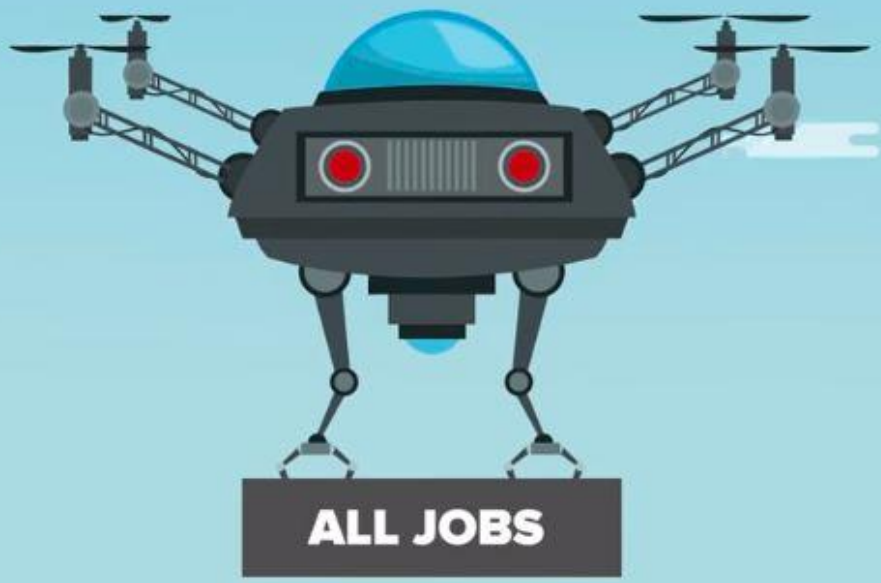
Profession of the Future

Problem?


It's cheaper to buy a $35,000 robotic arm than it is to hire an inefficient employee who's making $15 an hour bagging French fries


BANK TELLER
ATM of tomorrow will be able to do 90% of the work of a teller.


JOURNALISTS


ALL JOBS

What we want?

Better Career

Better Salary

Big Companies Hiring

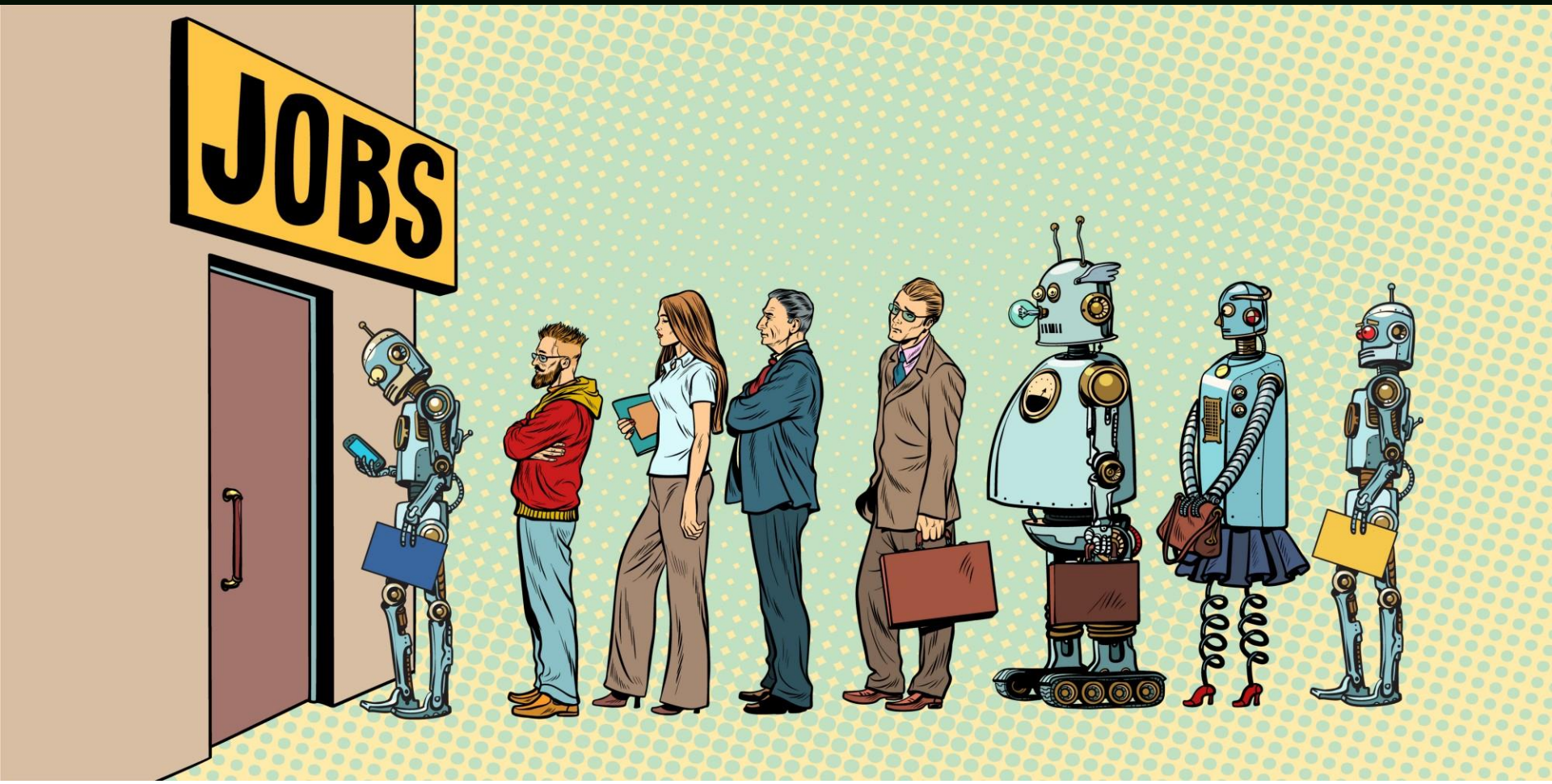
Better Job Opportunities

Big Data everywhere

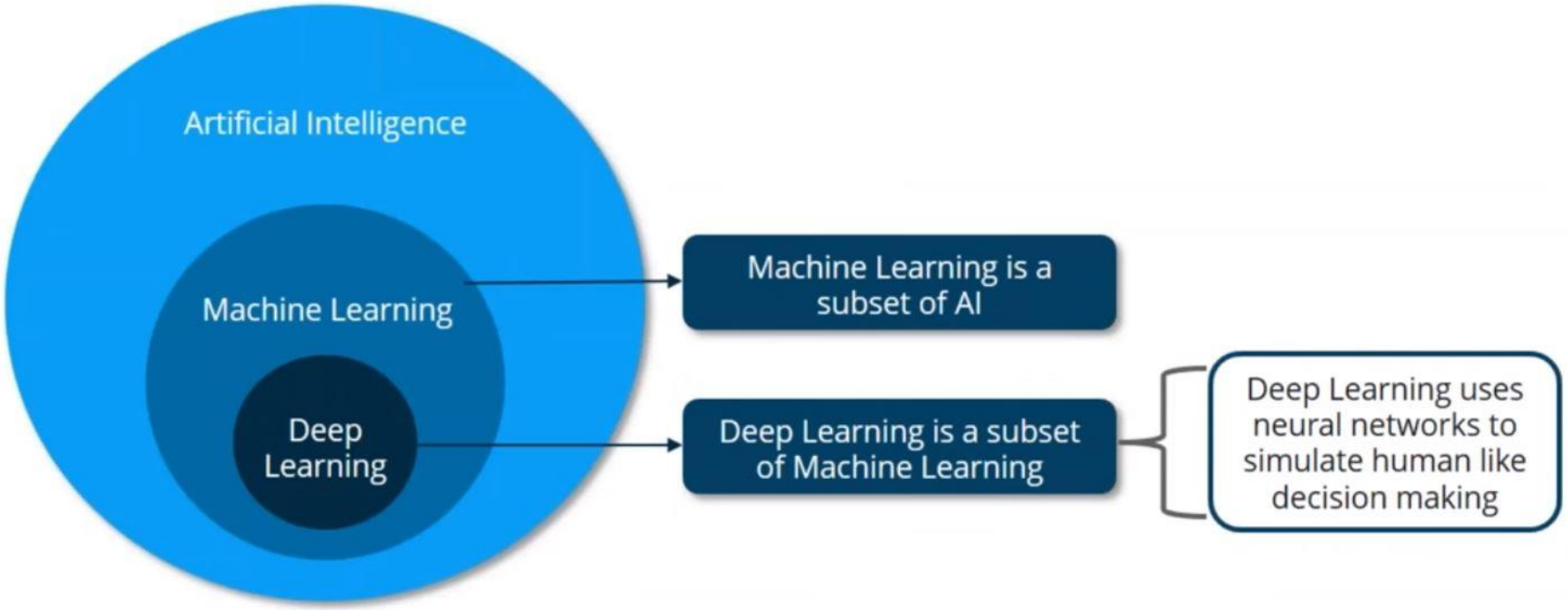
NLP & Machine Learning & Deep Learning & Computer Vision

Present = Gift

Limited Talent



NLP & Machine Learning & Deep Learning & Computer Vision



What is Data Science?

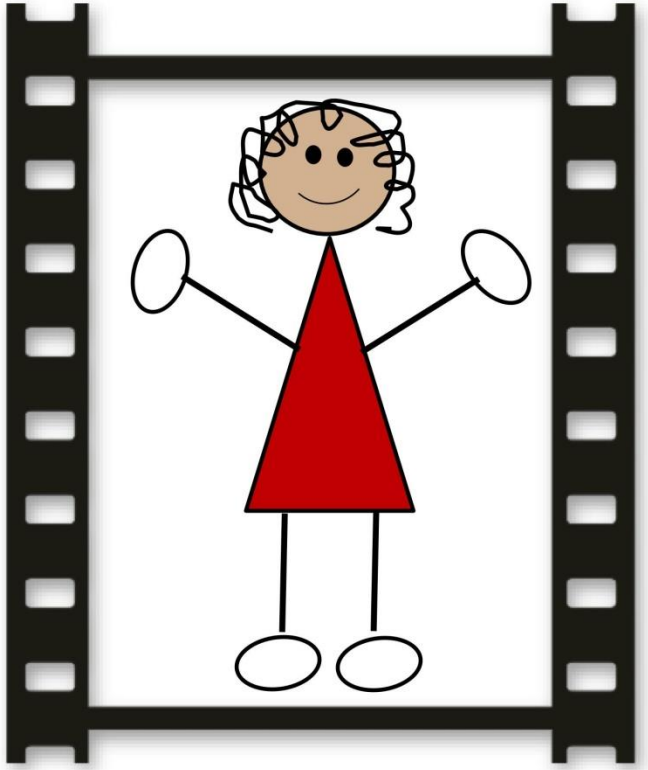
Diploma in Data Science & Big Data Analytics

The Protagonist

Meet Ms. Julie Johnson, an Electrical Engineer who graduated with top grades.

She got a campus offer from Xtelligence Consulting, a leading provider of Analytics consulting and services.

Julie is excited about joining Xtelligence and starting her career in the exciting field of Analytics.



End of Day: What Happened?

End of Day: Buried under “The Jargon Cloud”

Scenario: Busting the Jargon

Vanity Cosmetics is a Beauty Care and Personal Care product company. Currently their operations and market is primarily in Canada, but they would like to expand to other markets.

The board of directors meet with the leadership team of Vanity Cosmetics to discuss the expansion strategy. One of the key areas that emerge as focus area is Analytics. The board asks the leadership to put together an Analytics and Data strategy for the organization to successfully transition to the next level.

What jargon does the Vanity team needs to understand and familiarize? What is the Analytics roadmap for Vanity?

Data Science

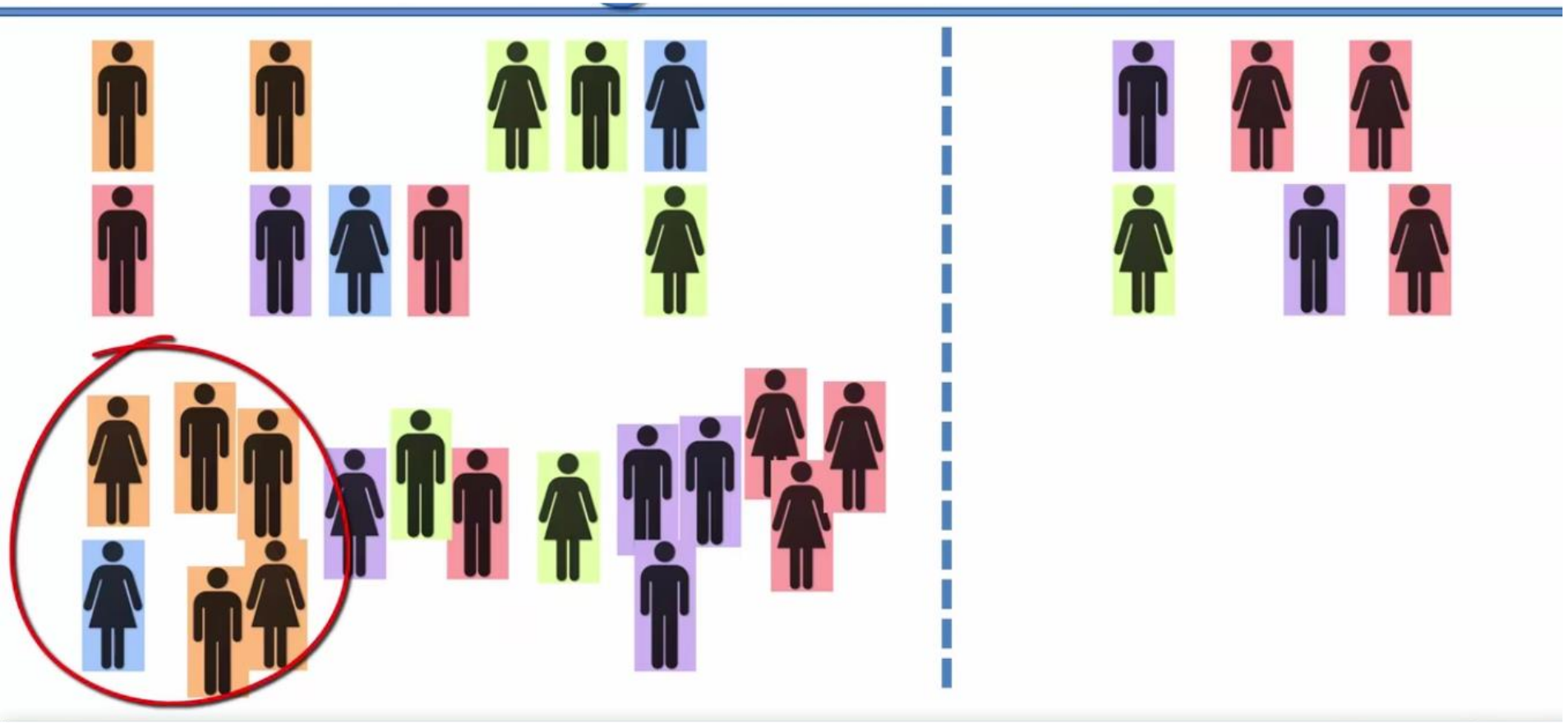
- Skill of extracting of knowledge from data

Insights

Data Science

Diploma in Data Science & Big Data Analytics

Others



Stay Home

Identifying Hidden patterns & hidden relationships

Diploma in Data Science & Big Data Analytics

By Navin Pathuru at Naresh IT Opposite Satyam Theatre, Ameerpet, Hyderabad 040-23746666, 23734842

That's Insight

What Companies want?

Diploma in Data Science & Big Data Analytics

Data Science

- Skill of extracting of knowledge from data
- Using knowledge to predict the unknown
- Improve business outcomes with the power of data

Knowledge

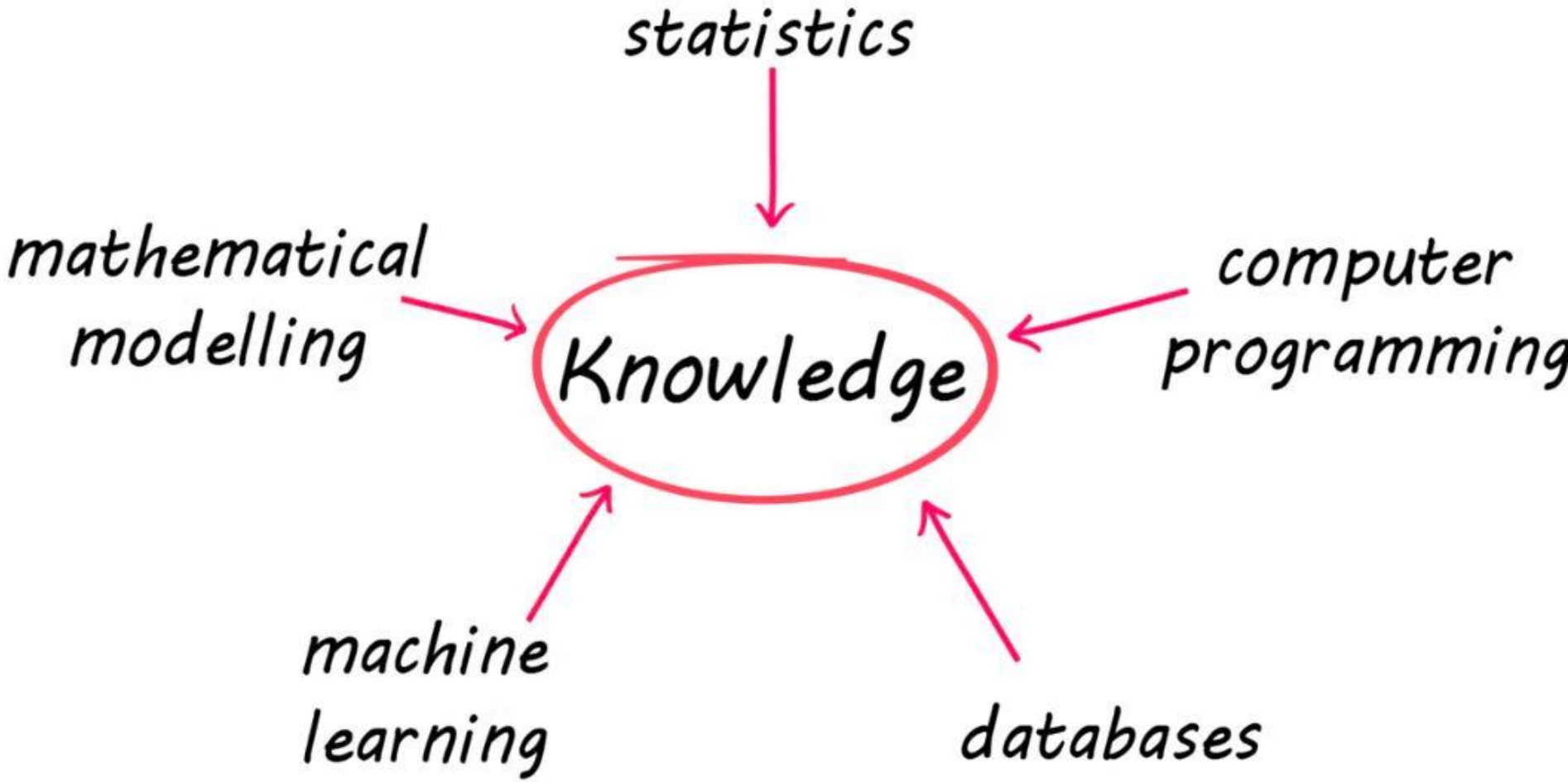
Diploma in Data Science & Big Data Analytics

By Navin Pathuru at Naresh IT Opposite Satyam Theatre, Ameerpet, Hyderabad 040-23746666, 23734842

Data Science

- Skill of extracting of knowledge from data

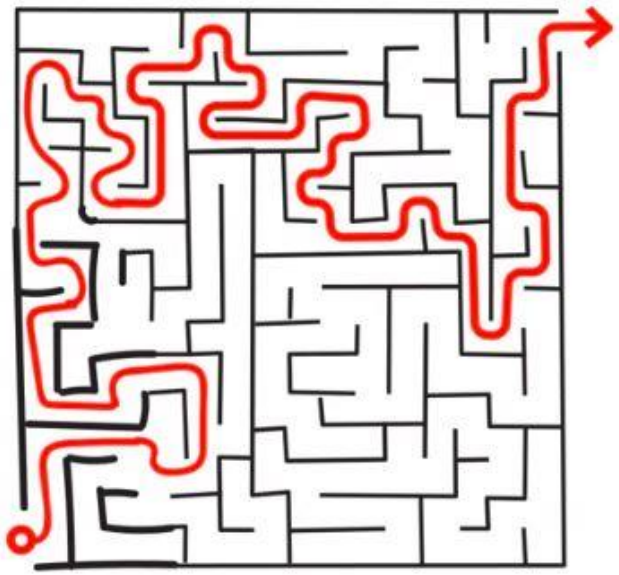
Use Data Analytics to help fight Corona



Diploma in Data Science & Big Data Analytics

Hidden Patterns in Data

Data Scientist



Everyday Recommendations

Shopping

Restaurant

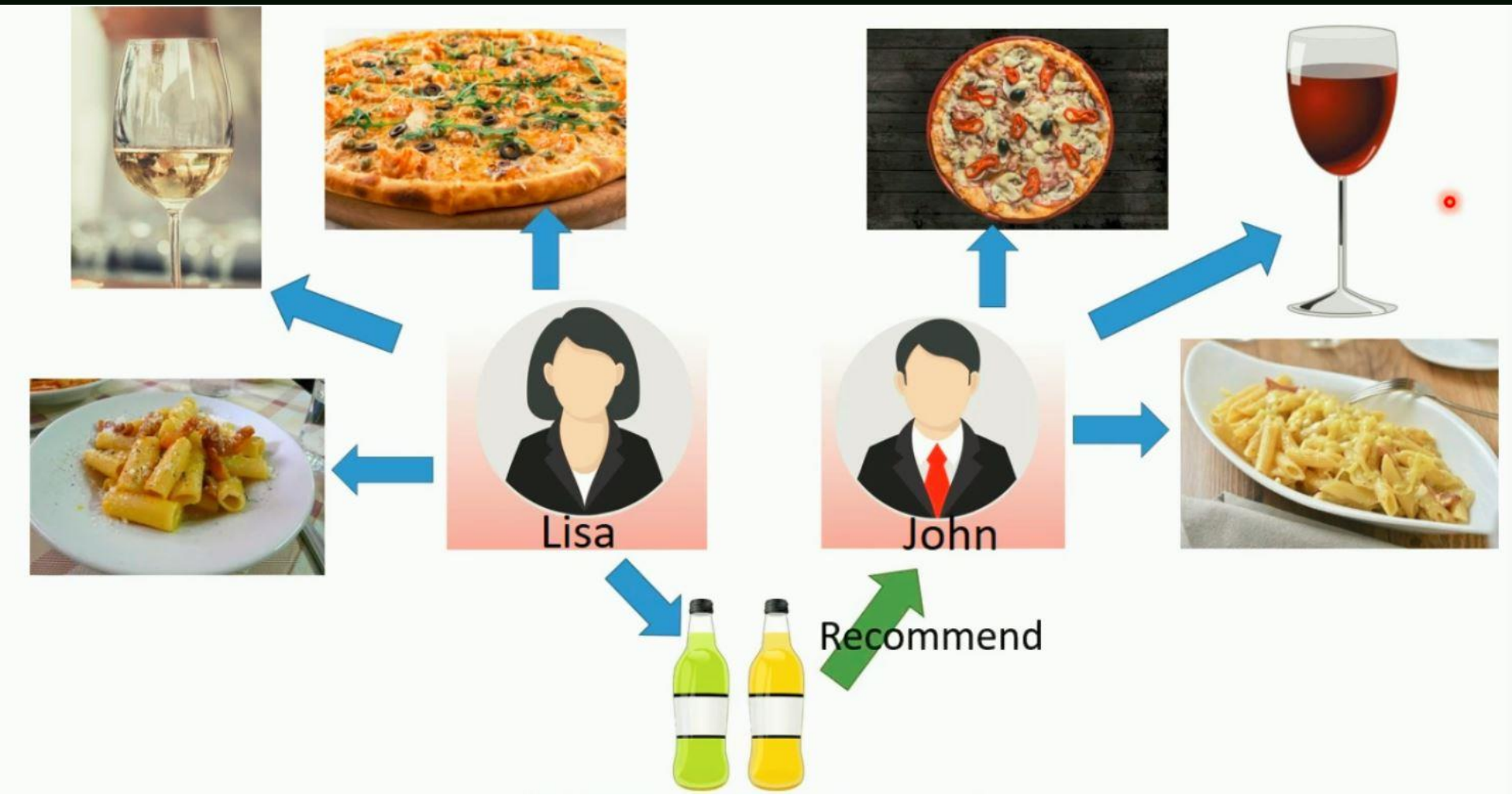
Movie

Books

Travel

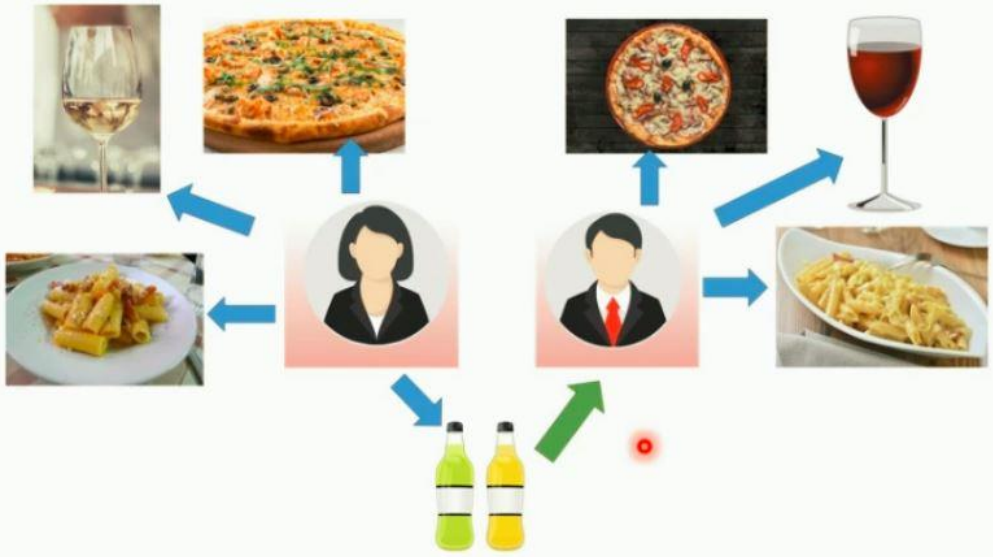
Friends, Family, Colleagues, Professors

Recommendation System



Collaborative Filtering

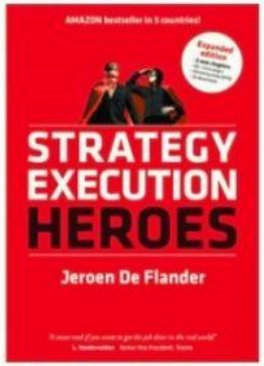
- Analyse User behaviour, Activities and preferences
- Recommend based on similarity to other user
- People who agreed in the past will agree in the future, and that they will like similar kinds of items as they liked in the past.

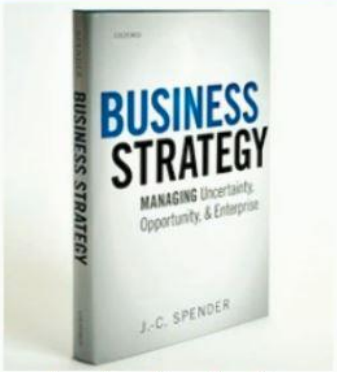


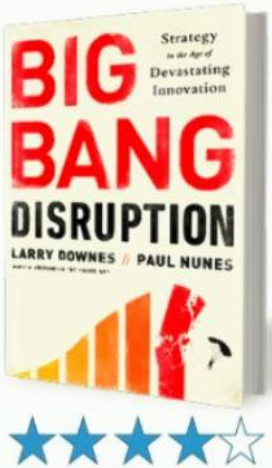
People who buy “x” also buy “y”.

Recommendation Systems

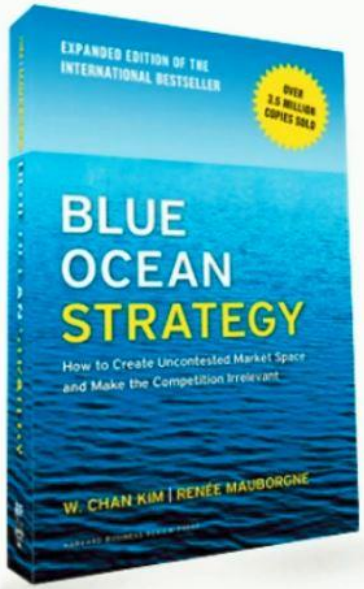
Content Based Filtering







Recommend



Heart of the Businesses

Invest in Knowledge

"Sachin Tendulkar is the Roger Federer of Cricket"

Roger Federer – tennis + cricket = Sachin Tendulkar

Language of the Machines…

Natural Language Processing…

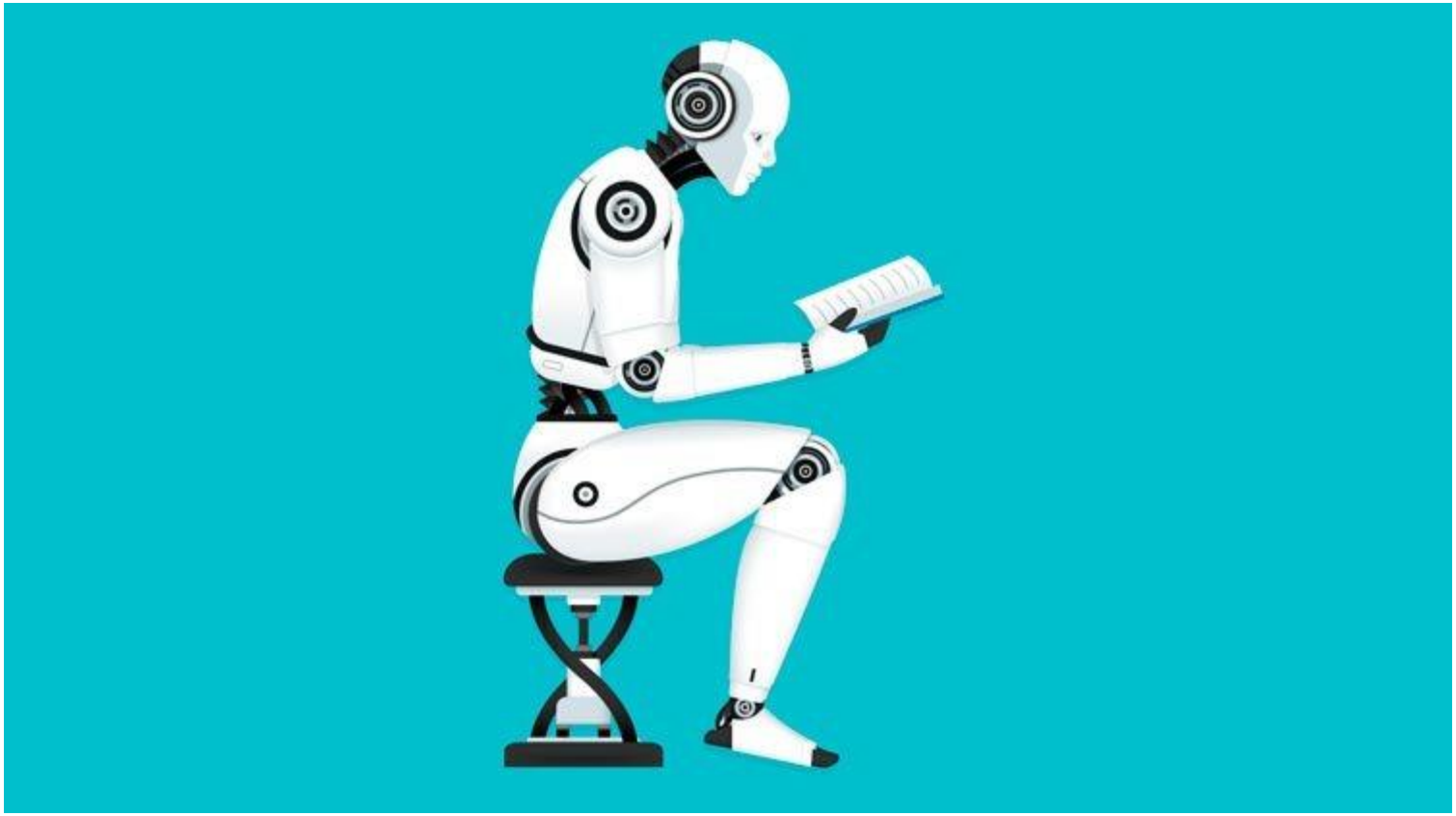
"I fear not the man who has practiced 10,000 kicks once, but I fear the man who has practiced one kick 10,000 times." - **Bruce Lee**

Stay Home + Stay Safe + Learning =
Best Investment =
Best Life

Learning is Life

Learning = Life

Machine + Data = Machine Learning



Humans + Experience = Learning = Life

Diploma in Data Science & Big Data Analytics

By Navin Pathuru at Naresh IT Opposite Satyam Theatre, Ameerpet, Hyderabad 040-23746666, 23734842

Machine Learning

shahrukh khan - YouTube

Secure | https://www.youtube.com/results?search_query=shahrukh+khan

YouTube IN

shahrukh khan

Home
Trending
Subscriptions

LIBRARY
History
Watch later
Lessons in Data ...
Bhakthi
Show more

SUBSCRIPTIONS
SVBC TTD
University of Hel... 1
IMA UMN 1
UpX Academy 5
Yury Kashnitsky 2
ExcelR Solutions 3
Coding Tech 15
Show 994 more

About 6,680,000 results

FILTER

Shahrukh Khan in Kuwait
Kuwait Uptodate • 278K views • 1 week ago
Shahrukh Khan in Kuwait Video by Usman Choudhry.

An Insight, An Idea with Shah Rukh Khan
World Economic Forum ✓ 95K views • Streamed 3 months ago
A conversation with actor and Crystal Award honouree **Shah Rukh Khan** on creating change in India through women's ...

Dr Shah Rukh Khan - Life Lessons
The University of Edinburgh • 1.2M views • 2 years ago
Dr **Shah Rukh Khan**, Bollywood actor, delivers his public lecture entitled Life Lessons. Dr Khan is one of the most influential actors ...

Non stop Hindi songs BEST OF SHAHRUKH KHAN | BEST SONGS OF SHARUKH KHAN
BOLLYWOOD HITS • 26M views • 3 years ago
Best songs of sharukh **khan** This is one of the best songs of sharukh **khan** non stop songs.

KBC 6 NOV 4 Full Episode Shahrukh and Katrina || Amitabh Bachchan
Rabius Sunny • 32K views • 1 month ago

Shah Rukh Khan
Music

Dilwale Dulhania Le Jayenge (Dialogues)
82K views • 1 year ago

World Dance Medley
1.7M views • 3 years ago

Jab Tak Hai Jaan [The Poem] With lyrics & English Transla
526K views • 5 years ago

Korbo Lorbo Jeetbo
20K views • 1 year ago

Chalak Chalak
10K views • 6 months ago

Apun Bola
1.6K views • 2 years ago

Aa Raha Hoon Palat Ke
40K views • 6 years ago

Suraj Hua Madham
1.2K views • 2 months ago

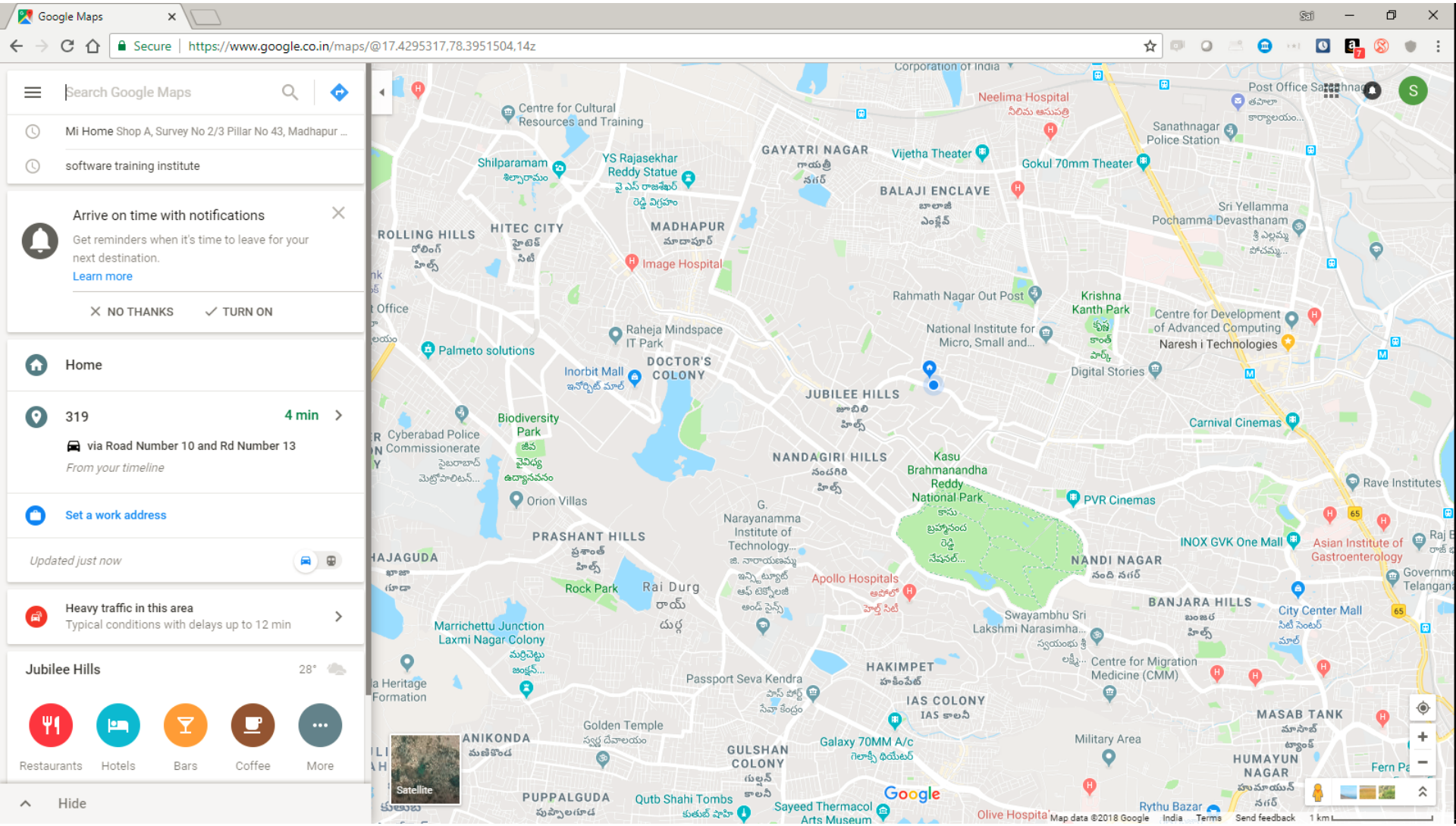
4:19 PM
5/7/2018

Machine Learning to the rescue

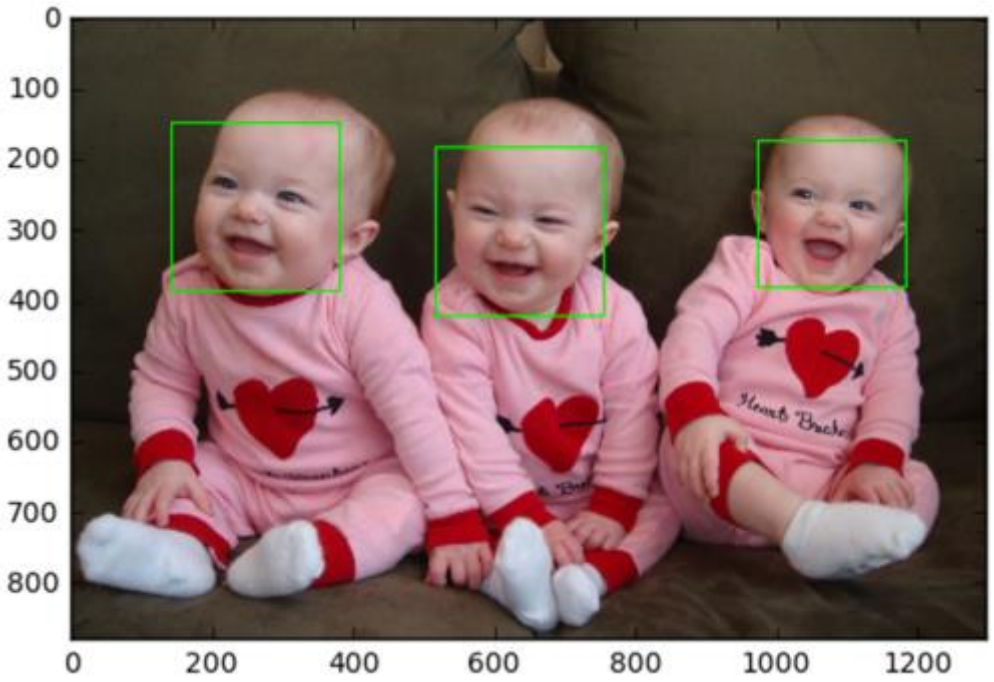
Diploma in Data Science & Big Data Analytics

By Navin Pathuru at Naresh IT Opposite Satyam Theatre, Ameerpet, Hyderabad 040-23746666, 23734842

Machine Learning



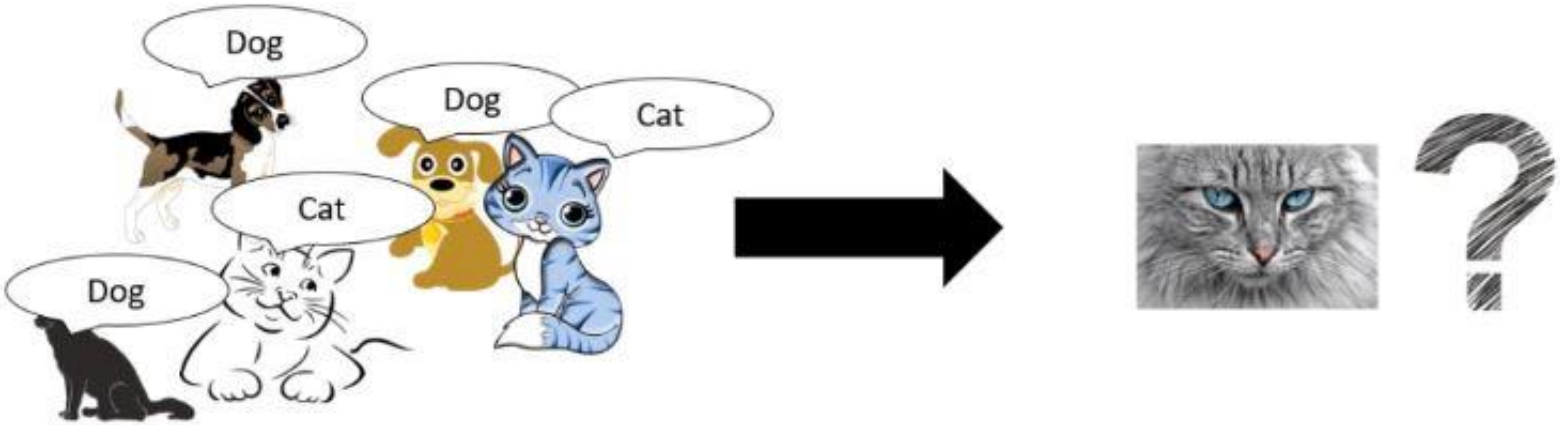
Machine Learning



What is Machine Learning

Machine learning is at the heart
of **global mega-trends & innovations...**

Machine Learning Problem



Training

Testing

Data

- (Test Data/ Live Data)

→ **Model** → **Predicted Output**

Simple Mapping Problem Ctd…

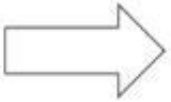
x1	x2	y
1	2	3
2	3	5
3	4	7
4	5	9
6	7	13

Can you think of mathematical relation exist between x1, x2 & y.

$$y = x1 + x2$$

Machine Learning

x1	x2	x3	…	x1000	y
..	..	..	..	..	..
..	..	..	..	..	..
..	..	..	..	..	..
..	..	..	..	..	..
..	..	..	..	..	..



$$y = f(x1, x2, \ldots, x1000)$$

Machine Learning

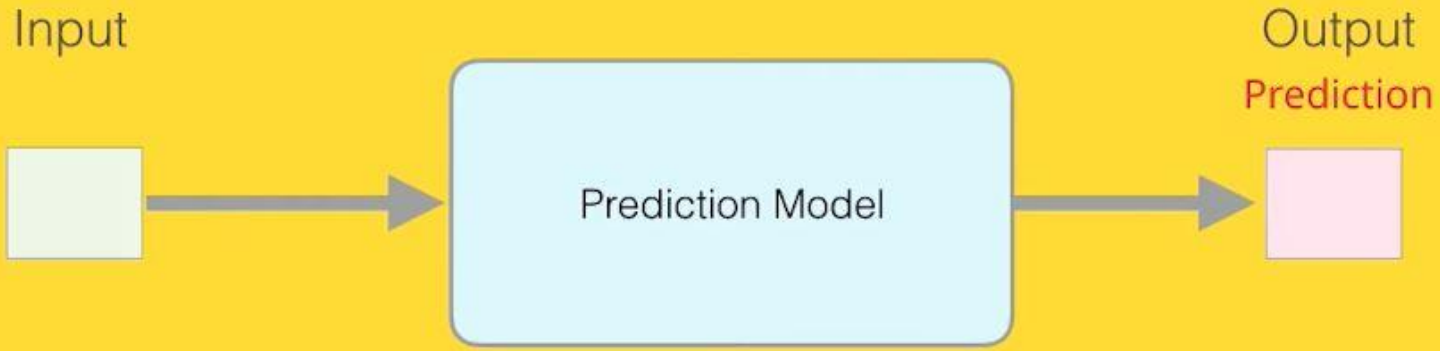
$$x1, x2, x3, \ldots, x1000$$

$$\begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

$$y = Cat$$

$$y = f(x1, x2, \ldots, x1000)$$

Prediction



CONCEPT

Predicting (estimating, calculating) values based on patterns in other, existing values is the most commonly used application of machine learning in practice

Prediction

Features | Label

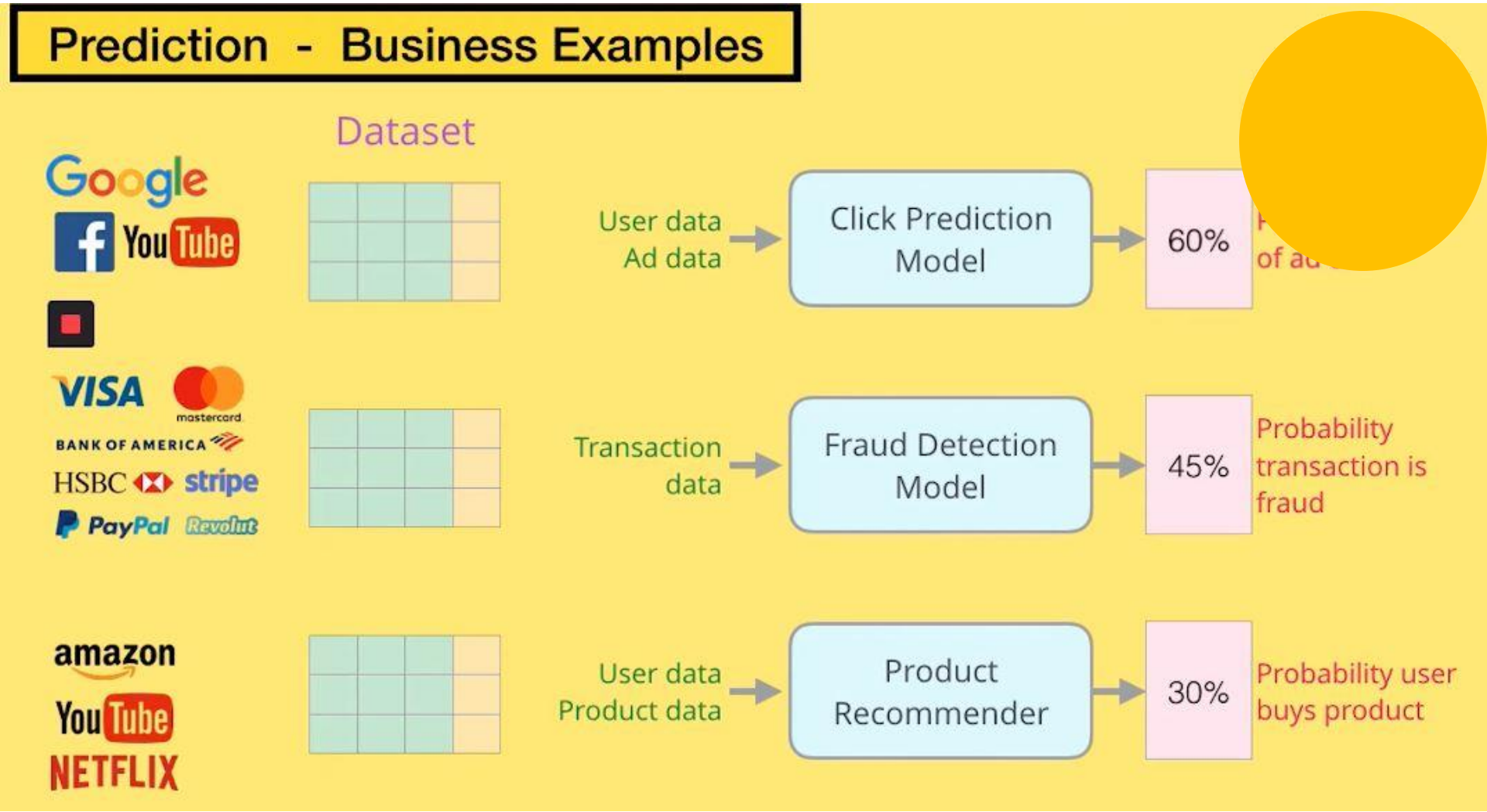
# of people in group	$ paid for ice cream
1	10
2	20
4	40

Labeled Dataset

3	?

10 Weight

Prediction - Business Examples



The Prediction Recipe

1. What to Predict?

Prediction:

$ dollars sale

?

2. Prepare Data

#	Feature #1: # of people in group	Label: $ paid for ice cream
1	1	10
2	2	20
3	4	40

10

3. Train & Evaluate Models

Model

Parameters:

10

4. Predict

Feature

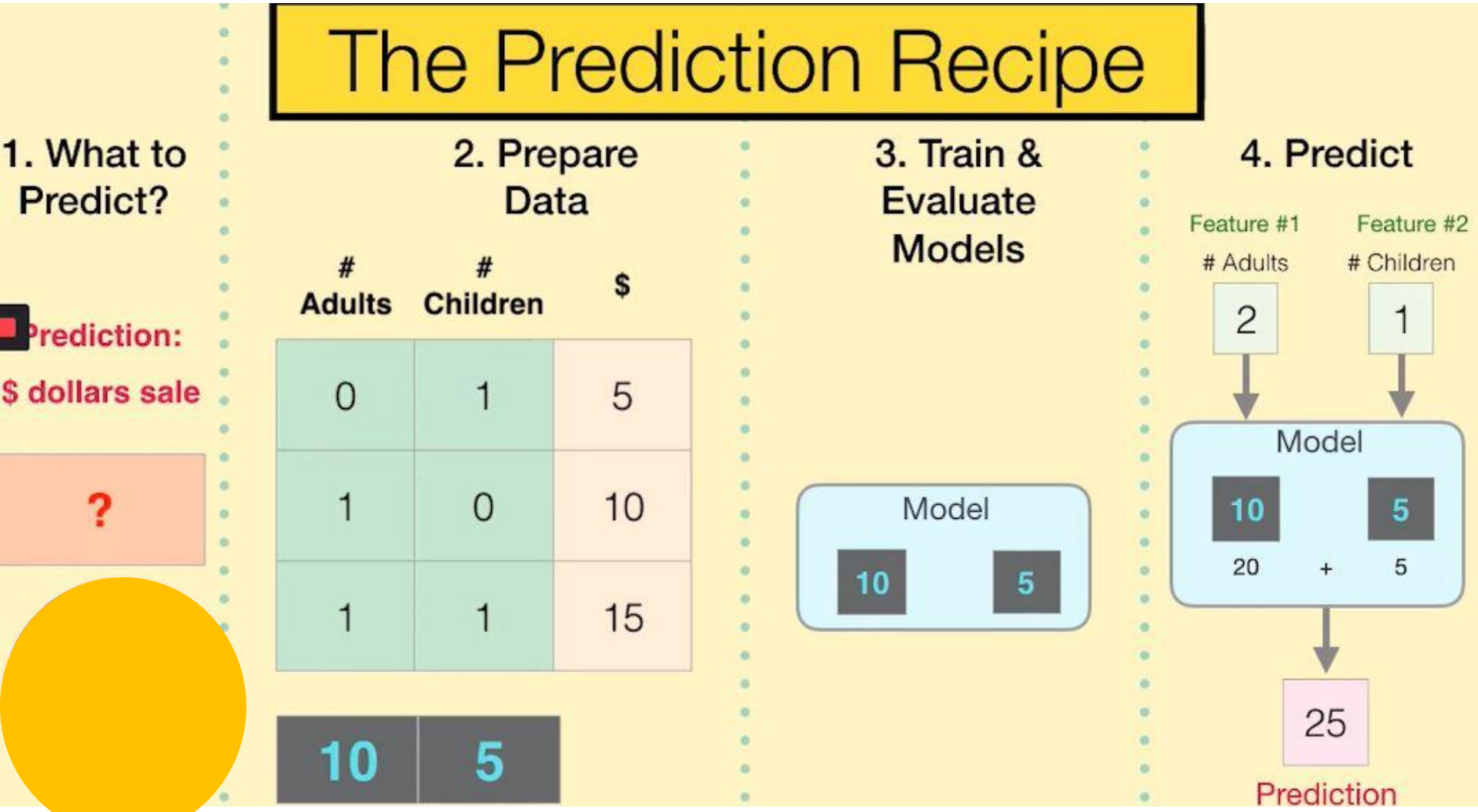
3

Model

10

Prediction

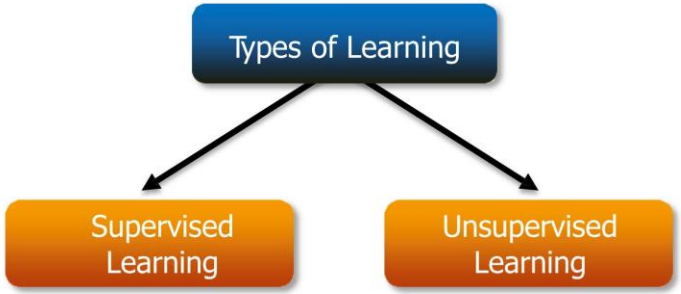
30



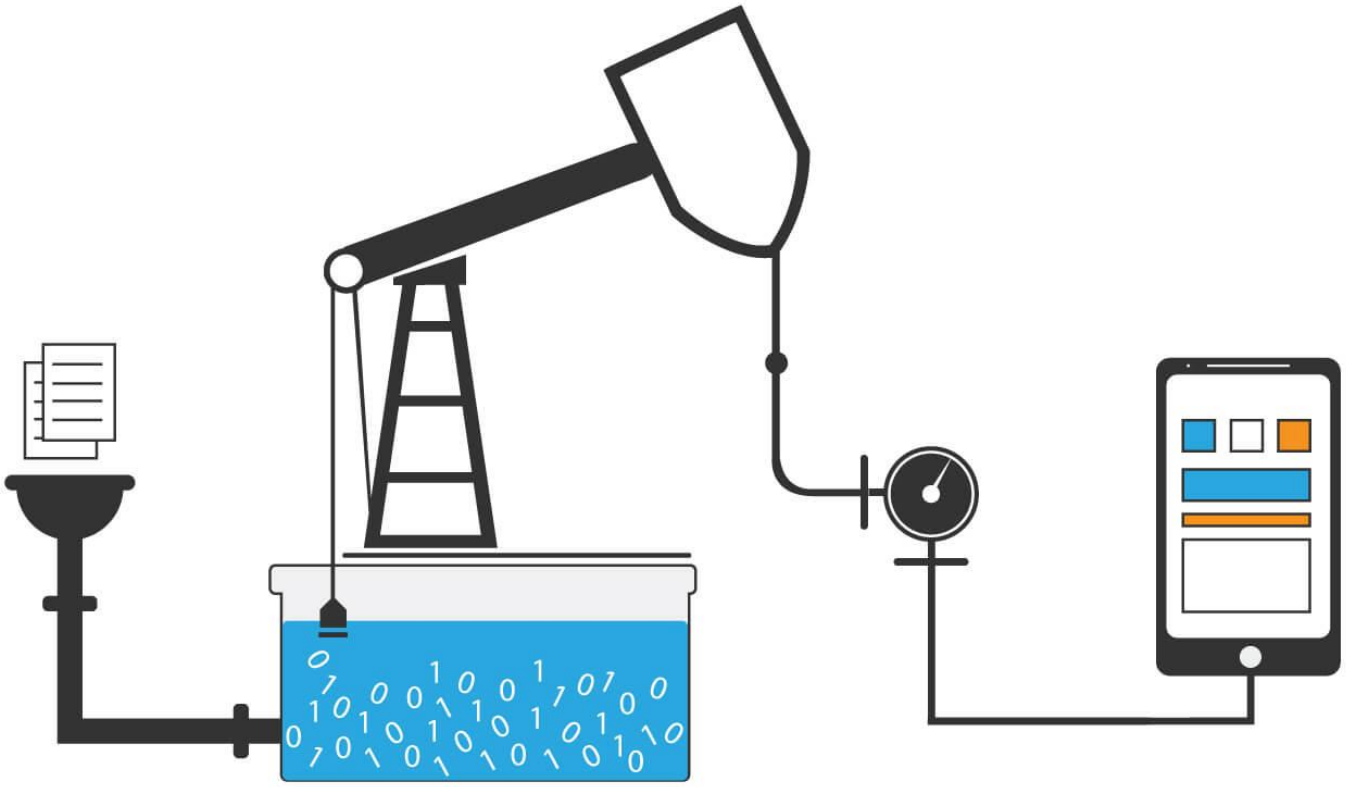
Diploma in Data Science & Big Data Analytics

Learning Techniques

Attain knowledge by study, experience, or by being taught.



What is Data?



Data is the new Oil

Business Problem

A travel booking website books thousands of hotel rooms every day.

With such a huge growth, the task at hand is to improve the booking per customer by showing the best hotels as per the taste and preferences of the user.

What is Data?

Sample Set

- Hotel 1 – Good Gym, Small Pool
- Hotel 2 – Amazing Gym, No Pool
- Hotel 3 – Great Gym, Small Pool
- Hotel 4 – Basic Gym, Large Pool
- Hotel 5 – No Gym, Large and Beautifully designed pool

- John – Prefers Gym over swimming pool
- Kavin – Likes Gym more than pool
- Bill – Needs a pool with basic Gym
- Frans – A pool is a must compared to Gym

Item – Feature Matrix

Hotel	Gym	Pool
Hotel 1	0.8	0.2
Hotel 2	1	0
Hotel 3	0.9	0.1
Hotel 4	0.1	0.9
Hotel 5	0	1

User – Feature Matrix

User	Gym	Pool
John	0.9	0.1
Kavin	0.8	0.2
Bill	0.3	0.7
Frans	0	1

Recommendation System

Collaborative Filtering

User	Hotel 1	Hotel 2	Hotel 3	Hotel 4	Hotel 5
John	4	5	?	?	1
Kavin	?	5	4.5	?	1
Bill	2	3	?	5	4
Frans	1	?	?	?	5

Which hotel to recommend to Frans?

What is Data?

Data is the foundation of Analytics. Before starting any analysis, you need to understand the characteristics of data, its source of origination, and the transformation it has gone through.

What is Data?

Data is a **set of values** of **qualitative** or **quantitative** variables.

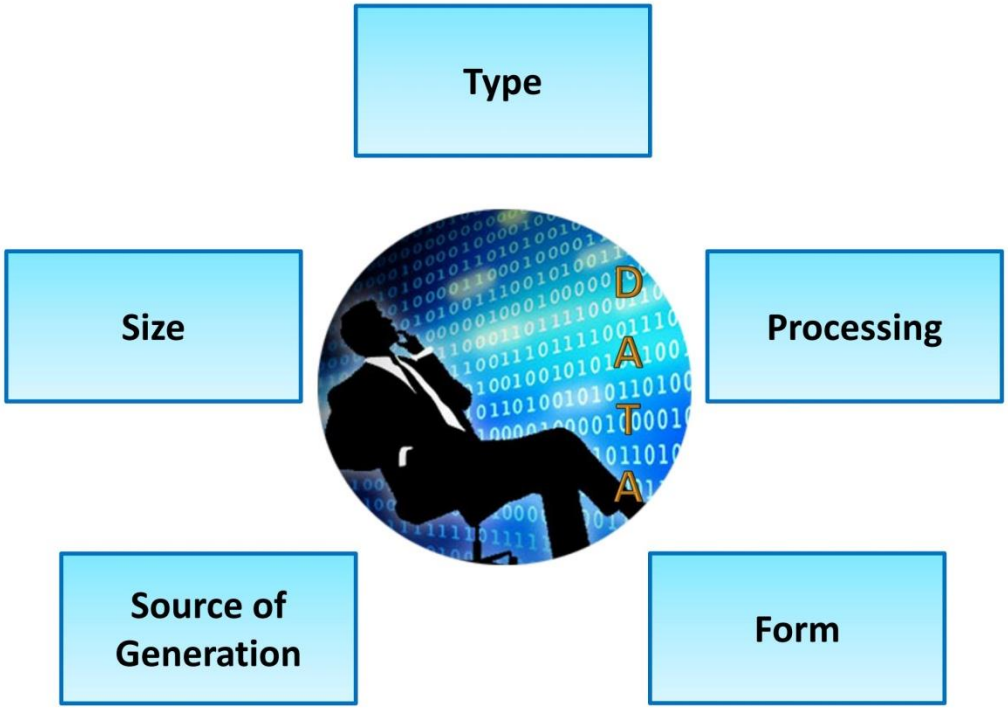
- Data is descriptive in nature; it describes an attribute that can be observed, but not measured
- Examples:
  - Flavors of ice cream = {“Vanilla”, ”Butterscotch”, “Chocolate” }
  - Hair color = { “Blonde”, “Brunette”, “Black” }
  - Profession type = { “Engineer”, “Tailor”, “Consultant” }

Qualitative

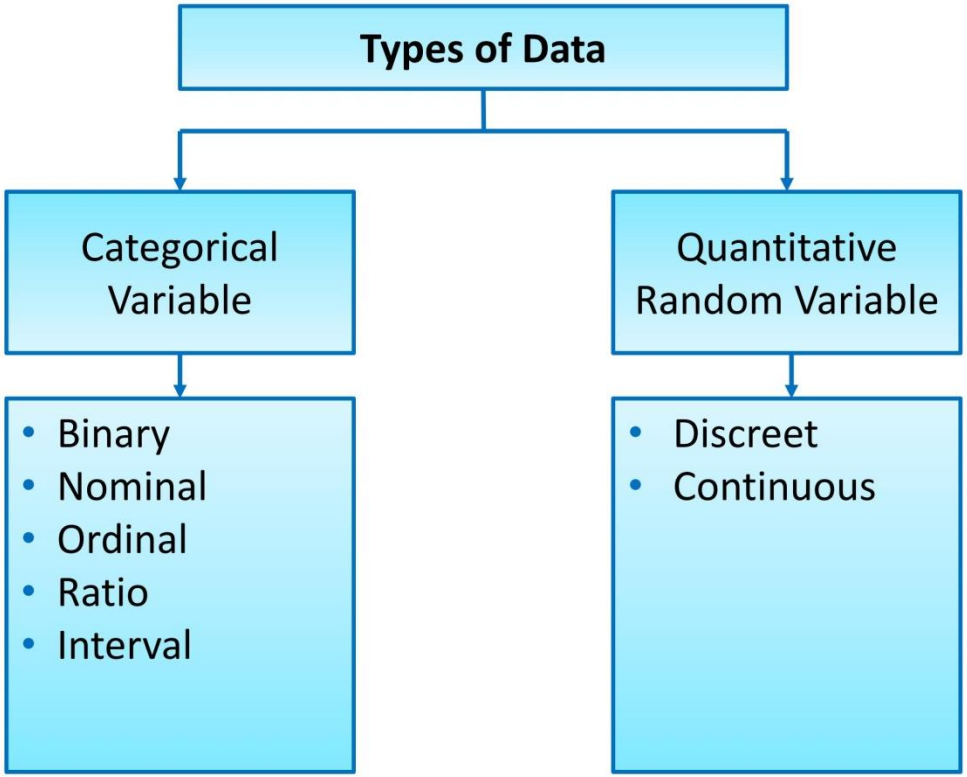
- Data is a numeric measure; it captures the measure of an attribute
- Examples:
  - Heights of students = {5’6”, 5’9”, 5’3”, 5’5” }
  - Cost = {120.5, 130.2, 111.6, 90.8}
  - Age = {34,26,67, 53}

Quantitative

Basis of Data Categorization



Types of Data



Binary Data

Binary data has only two possible states:

- 0 or 1
- Toss of a coin
- Switch On or Off
- Dot and dash of telegraph

Nominal Data

- Categorical data where the data is coded in a manner that it represents a label
- You can only count but cannot order or measure nominal data

Examples: Names of cars, book titles in a library, and marital status

Ordinal Data

- Data is ordered
- It has a natural hierarchy
- The intervals between the ranks may not be necessarily equal (distance between groups can be different)

Examples: Customer satisfaction score and medal tally

Discreet and Continuous Data

- Numerical data
- Finite number of possible values
- Examples:
  - Number of people in a room
  - Number of items in a basket
  - Numbers of hours in a day

Discrete Data

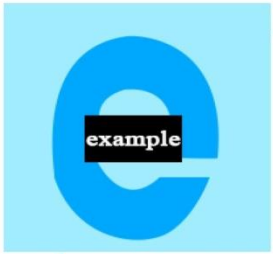
- Numerical data
- Infinite number of possible values
- Usually is in decimals
- Examples:
  - Height
  - Weight
  - Sales
  - Account balance

Continuous Data

Raw Data

RAW Data Definition:

- Data from the source
- Input to the data processing process
- Raw data may:
  - Have errors
  - Not validated
  - Multiple forms
  - Unformatted
  - Dubious, requiring confirmation or citation



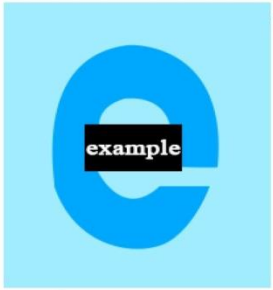
RAW Data Example:

- If the correct format is not specified in an application form, the date of birth data can take many forms, such as "31st January 1990", "31/01/1990", "31/1/90", and "31 Jan 90". This raw data needs to be processed to a common format for further use by systems/humans

Processed Data

Processed Data Definition:

- Data after processing for issues in the raw data
- Analysis ready data
- Processing includes scrubbing, cleansing, merging, formatting, transforming, and so on
- All data processing steps documented



Processed Data Example:

- Recoding: “Number of children” field in a survey form may be left blank by people who don’t have children. This has to be coded as “0”, which is a valid value for this variable
- Deriving: End of day sale amount for a store can be calculated by summing up all the transactions in a day

Data Collection Types

Census

- Systematic collection of data about all members of population

Observational study

- Collection of data to draw inference of outcome of a treatment on subjects. It is not in control of the investigator to assign the subjects either to the test or the control groups

Convenience sample

- Collection of data from a sample where the subjects are selected because of their convenient accessibility and proximity to the researcher

Randomized trial

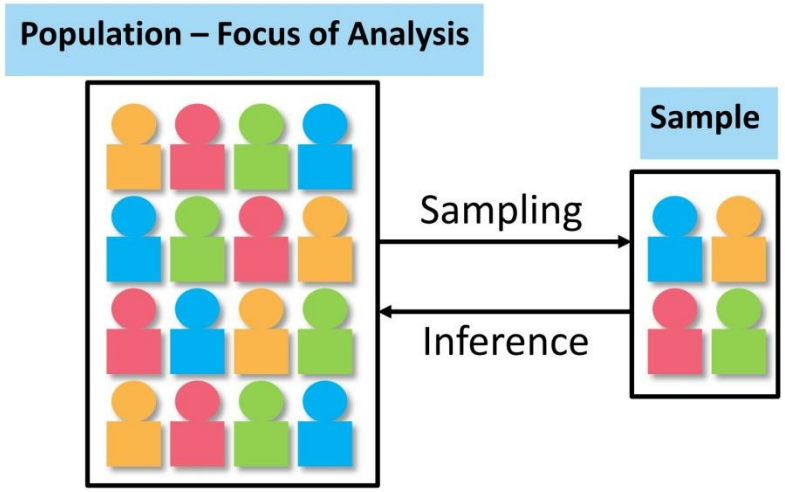
- Collection of data to draw inference of outcome of a treatment on subjects. The investigator randomly allocates the subjects to either the test or the control group

Agenda

- Sampling
  - Methods
  - Estimation of Sample Size

Population and Sample

- Population:
  A complete set of items that have common properties which, are the subject of statistical analysis
- Sample:
  A subset of the population of a manageable size selected through a defined procedure



Why Sampling?

Saves cost	Less expensive to study the sample than the population
Saves time	Less time needed to study the sample than the population
Accuracy	Sampling process is designed and conducted in a systematic manner by skilled personnel so that the expected results are accurate
Complete	No missing units and no duplication

Sampling: Example

- Analysis Requirement: Study the salaries offered to Engineering graduates in 2014 in India
- Population: All Engineering graduates across colleges in the country
- Sampling Frame: Engineering graduates with the listed characteristics:
  - 2014 pass out
  - Placed in campus recruitment
  - All branches
  - All tiers of colleges
  - Company profile
  - Industry and sector

Sampling Methods

Sampling methods are broadly classified as:

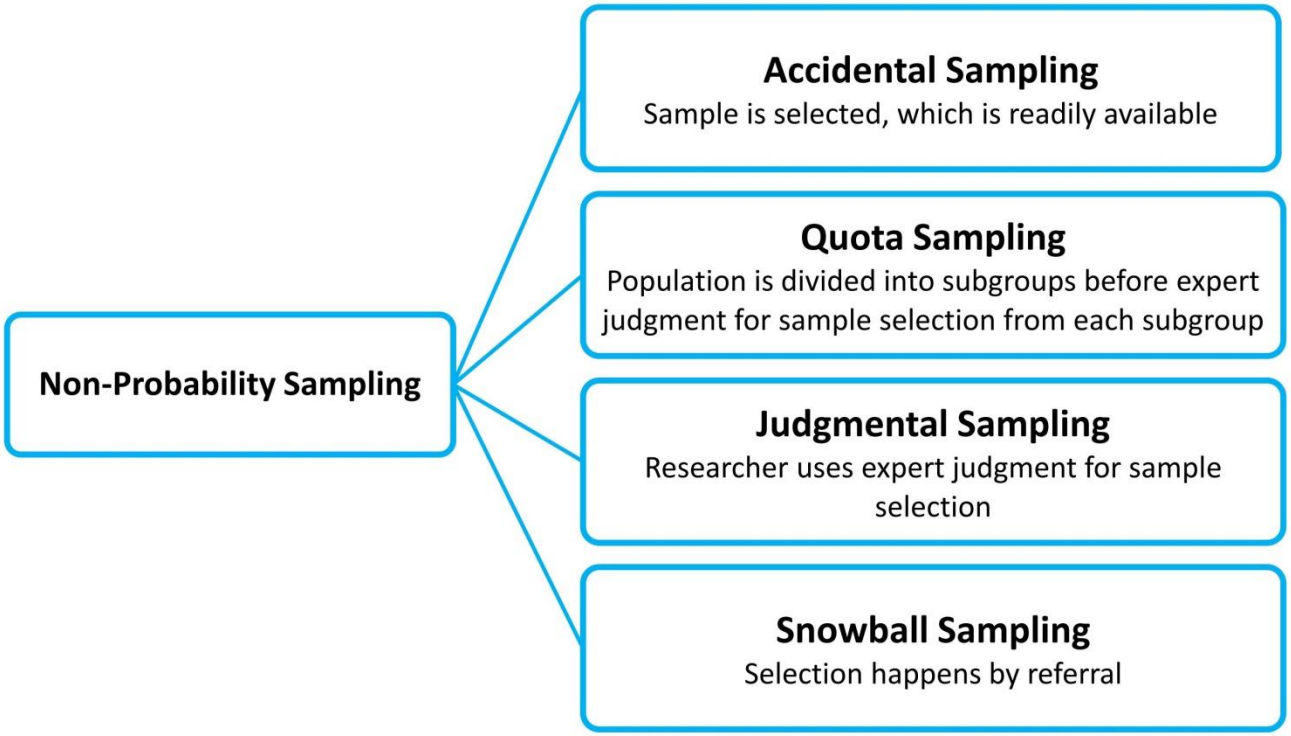
Probability Sampling

- Is the sampling method where each member of the population has a known probability (non-zero) of being selected as a part of the sample
- Sample is not biased

Non-Probability Sampling

- Is a sampling method where some members of the population have no chance of being selected
- Exclusion of samples can lead to sampling bias

Non-Probability Sampling

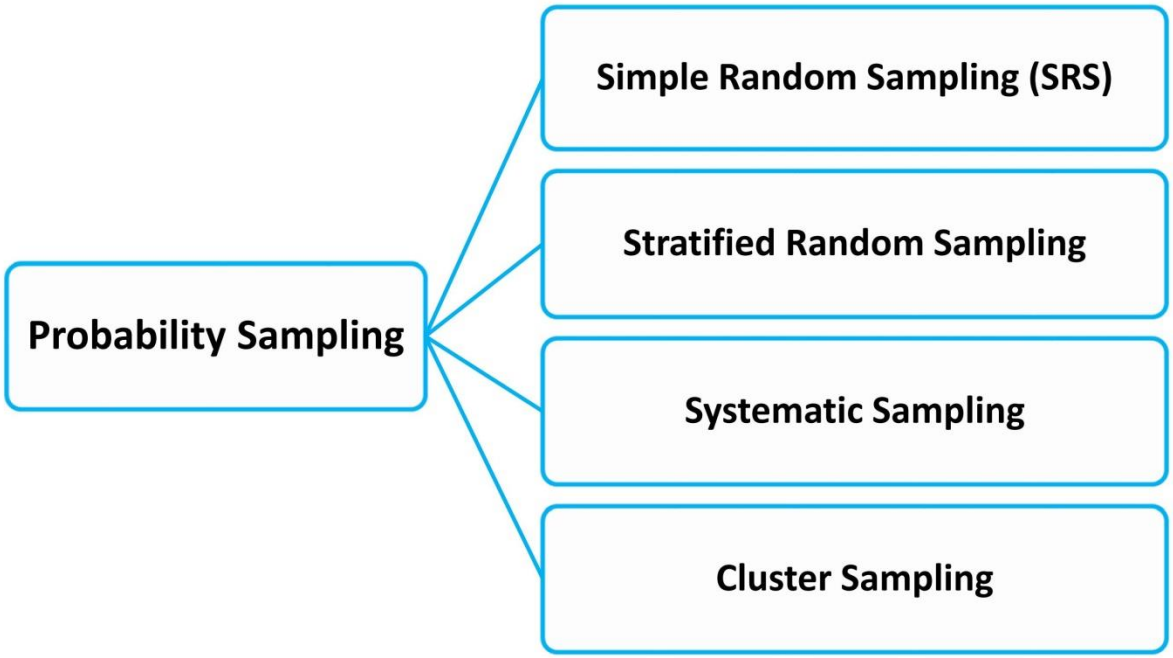


Non-Probability Sampling Examples

Jon, an Android App developer takes help of his friends Ted, Burt, and William to study the market before releasing an App developed by him. They identify the local shopping mall as the best place to do the survey. The table below shows examples of sampling methods.

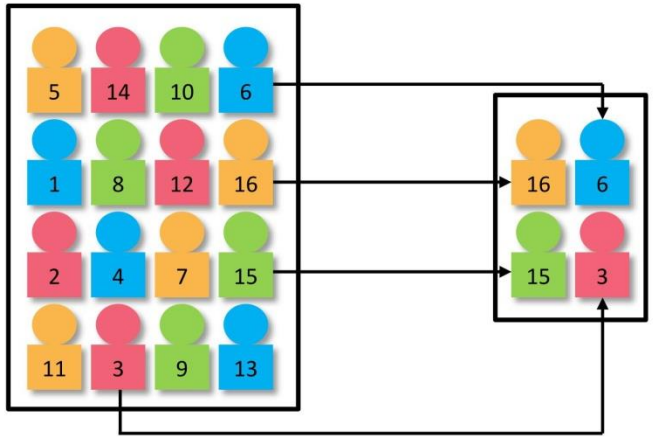
Sampling Method	Example
Accidental (Convenience)	Ted targets all people entering the mall to fill the survey.
Judgmental (Purposive)	Burt targets only college students for the survey.
Quota Sampling	William wants to ensure that there is adequate representation of men and women, so he decides on a quota of 3 men:1 woman.
Snowball Sampling	Jon focusses on getting references from people who fill the form. He also gives them a trial version of the App.

Probability Sampling



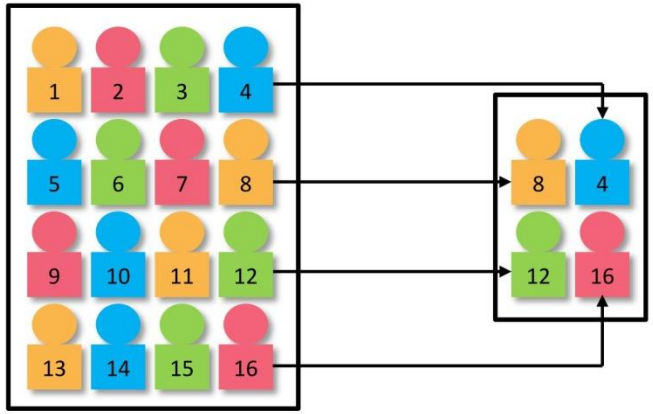
SRS

- All samples have an equal chance of being selected from the population
- SRS is cumbersome and tedious when sampling from a large population
- SRS is vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't represent the population



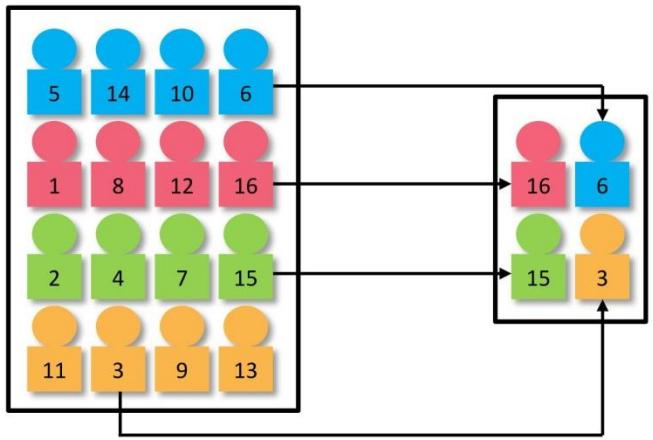
Systematic Sampling

- The population is arranged according to some ordering scheme and then samples are selected at regular intervals through that ordered list
  - Systematic sampling involves a random start and then proceeds with the selection of every k^{th} element from then onwards
  - As long as the starting point is randomized, systematic sampling is a type of probability sampling
  - Systematic sampling is vulnerable to periodicities in the list



Stratified Random Sampling

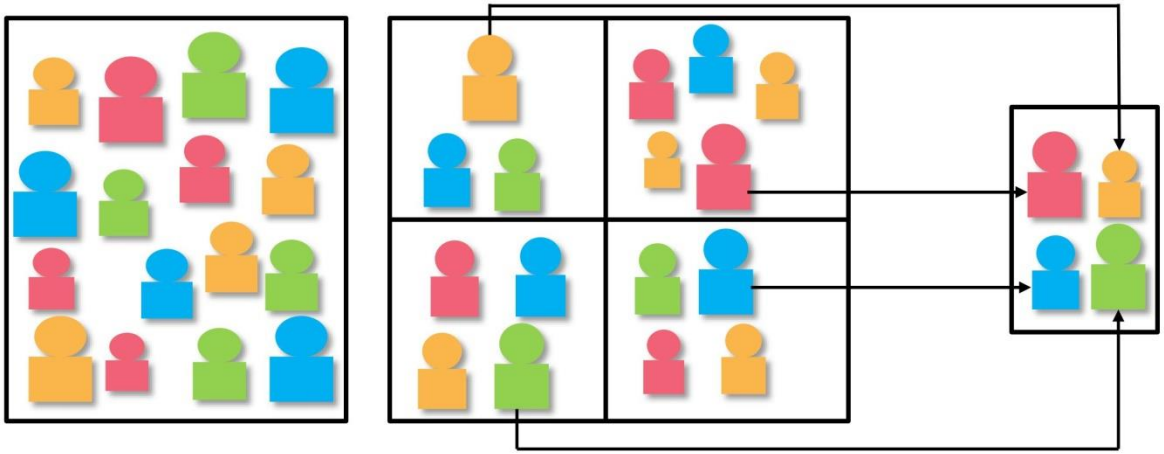
Stratified Random Sampling involves selecting independent samples from a number of subpopulations, groups, or strata within the population.



Cluster Sampling

Cluster sampling involves:

- Step 1: Divide the defined population into number of mutually exclusive and collectively exhaustive subpopulation groups or clusters
- Step 2: Select an independent simple random sample of clusters



Probability Sampling Examples

A nutrition drink manufacturer conducts an analysis on height of children. He selects a few schools and creates a database of all students.

Sampling Method	Data collection
SRS	A random number is assigned to each student and numbers 1-1000 are selected in the sample.
Systematic Sampling	Every 10^{th} student in the database is selected for the sample.
Stratified Random Sampling	Students are separated by schools and random samples are chosen from each school.
Cluster Sampling	Students are divided by age groups and samples are selected from each cluster.

Agenda

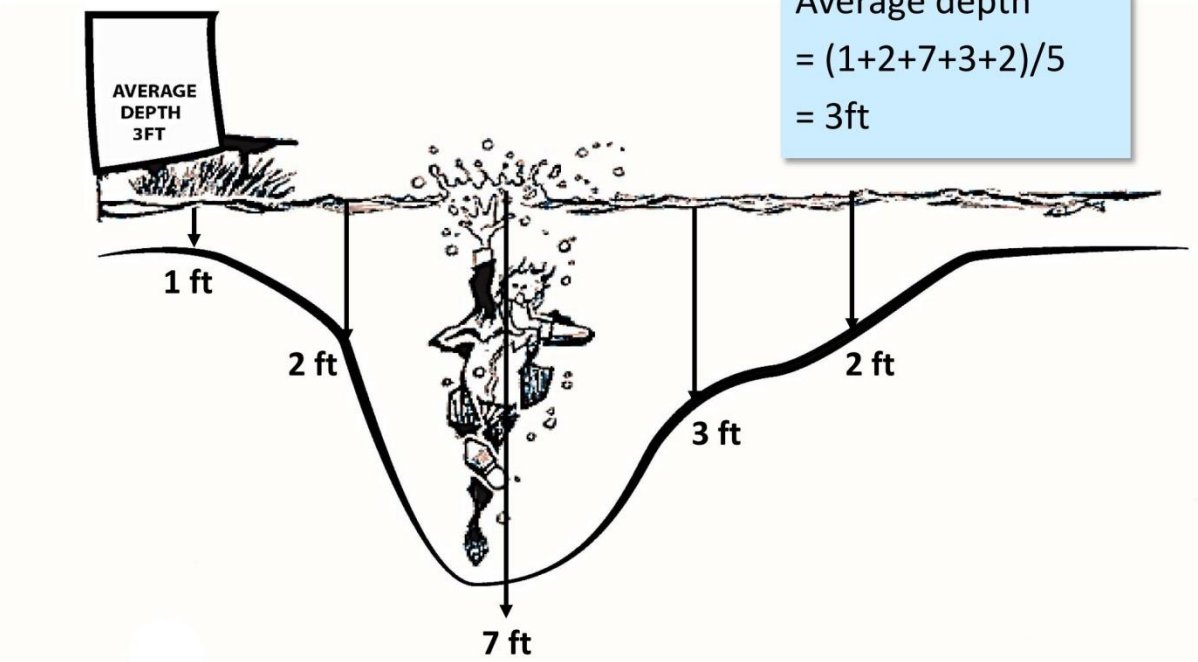
- What is Statistics?
- Central Tendency Measures
- Dispersion Measures
- Data Distributions

The “Average” Story

Alan went for a trek. On the way, he had to cross a stream. As Alan did not know swimming, he started exploring alternate routes to cross over.

Suddenly he saw a sign-post, which said “Average depth 3 feet”. Alan was 5’7” tall and thought he could safely cross the stream.

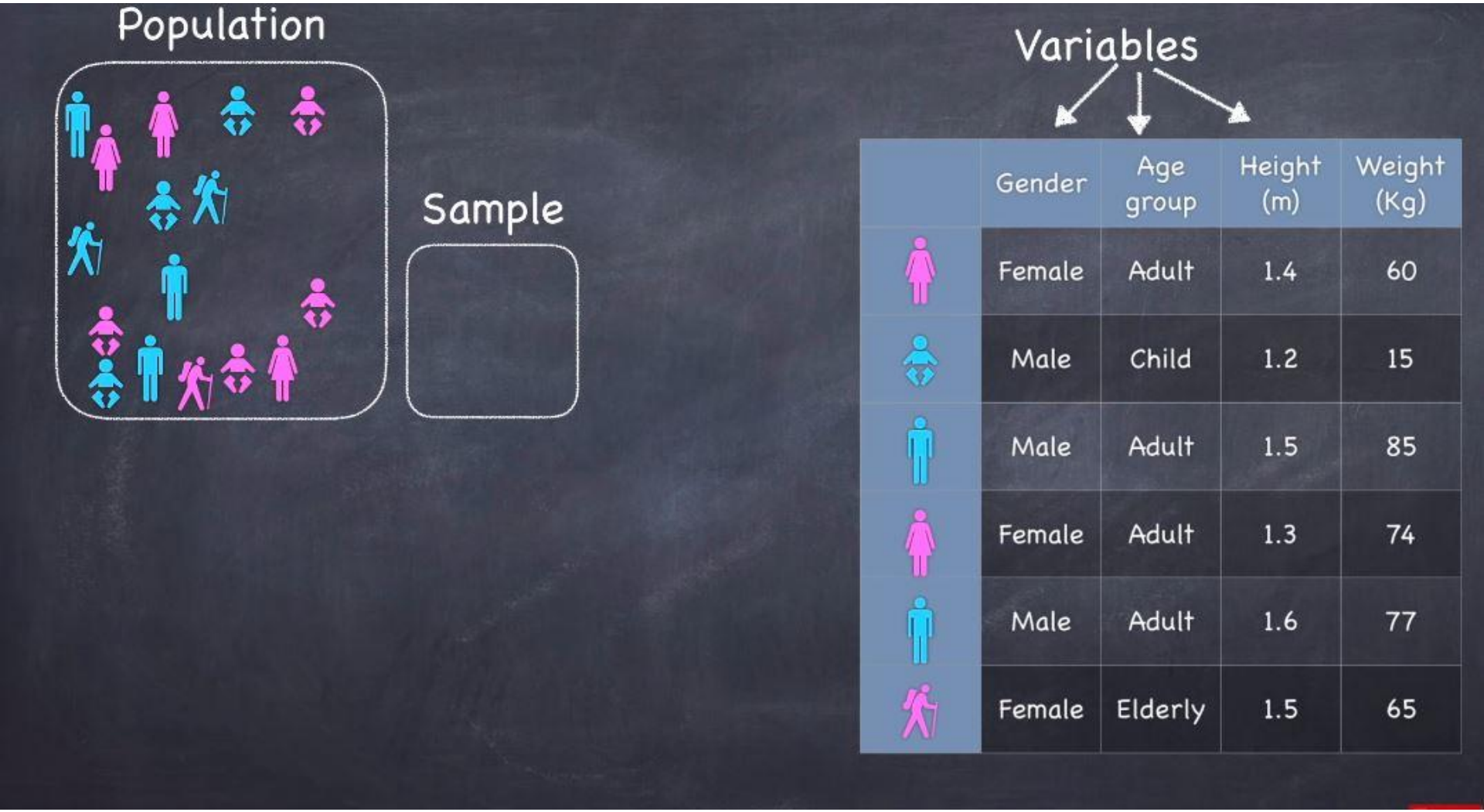
did Alan Drown?



Average depth
= (1+2+7+3+2)/5
= 3ft

Beware of Averages!!!


Population
Sample


Population
Sample
Variables
Gender
Age group
Height (m)
Weight (Kg)
Female
Adult
1.4
60
Male
Child
1.2
15
Male
Adult
1.5
85
Female
Adult
1.3
74
Male
Adult
1.6
77
Female
Elderly
1.5
65

Variables

	Gender	Age group	Height (m)	Weight (Kg)
	Female	Adult	1.4	60
	Male	Child	1.2	15
	Male	Adult	1.5	85
	Female	Adult	1.3	74
	Male	Adult	1.6	77
	Female	Elderly	1.5	65

Categorical

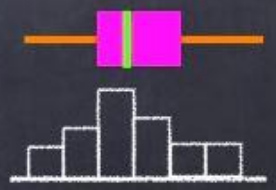
Summarise

		%
males	3	

Visualise

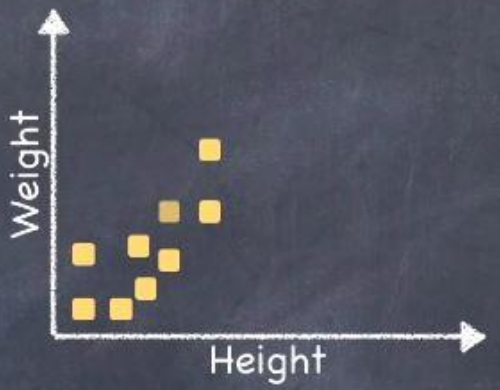
Numeric

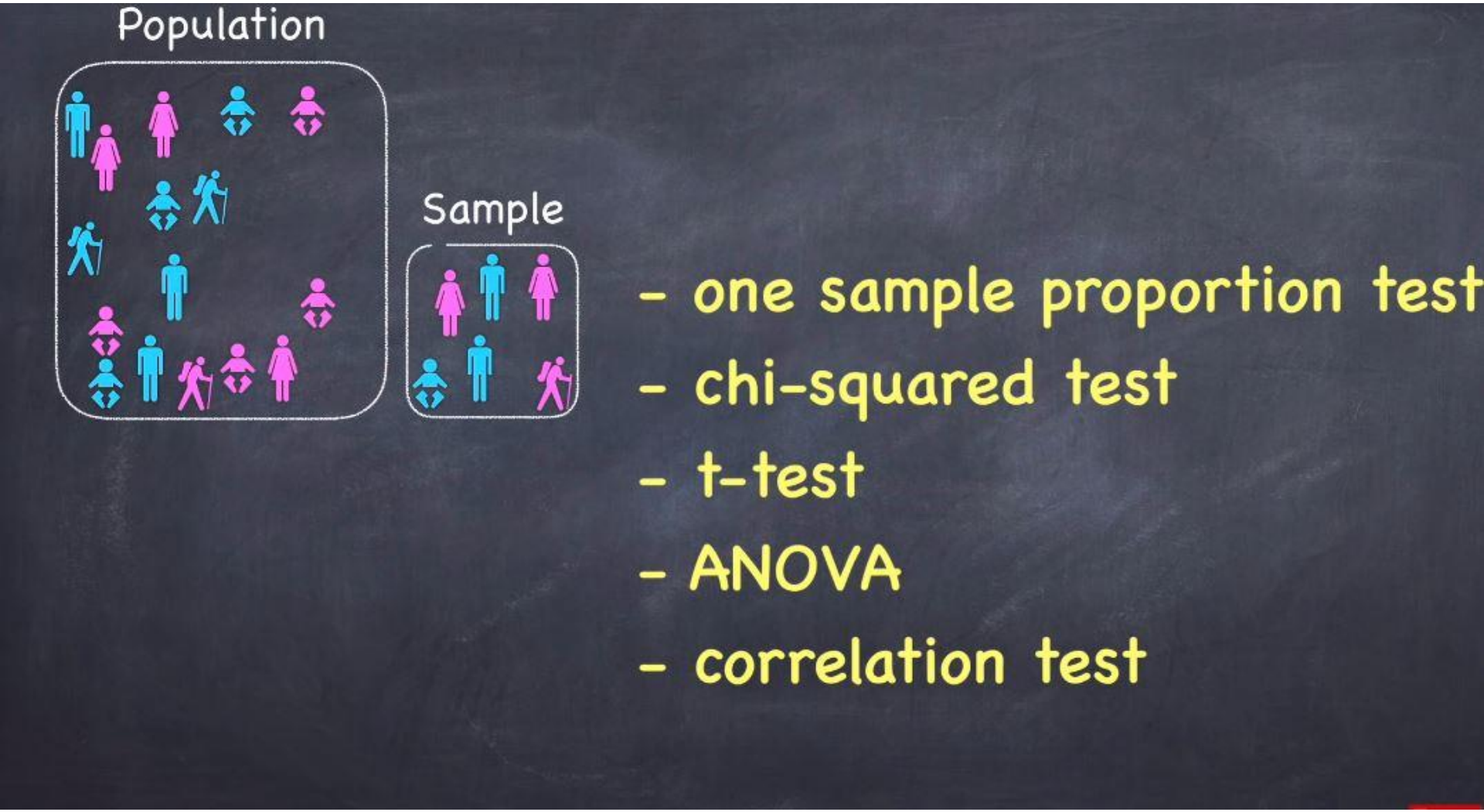
Range	1	6
IQR	2	4
Median	3	
Mean	3.5	



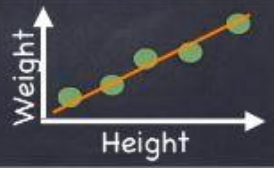
	Gender	Age group	Height (m)	Weight (Kg)
	Female	Adult	1.6	77
	Male	Child	1.2	15
	Male	Adult	1.5	85
	Female	Adult	1.3	74
	Male	Adult	1.4	60
	Female	Elderly	1.5	65


Height
M
F


Weight
Height

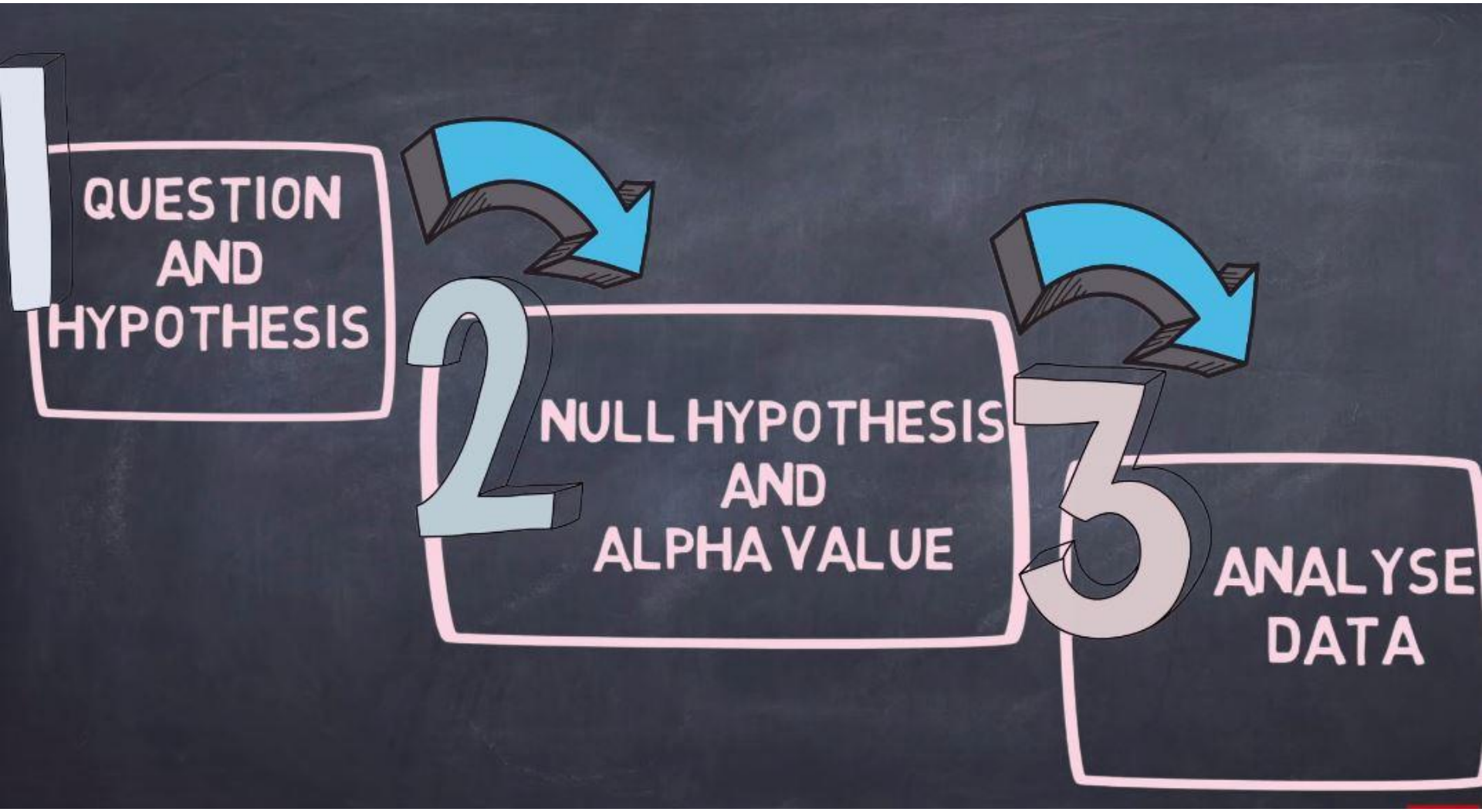

Population
Sample
- one sample proportion test
- chi-squared test
- t-test
- ANOVA
- correlation test

	Gender	Age group	Height (m)	Weight (Kg)
	Female	Adult	1.4	60
	Male	Child	1.2	15
	Male	Adult	1.5	85
	Female	Adult	1.3	74
	Male	Adult	1.6	77
	Female	Elderly	1.5	65

	What we observe in our sample data	Is it real?
One categorical		1 sample proportion test
Two categorical		Chi squared
One numeric		t-test
One numeric and one categorical		t-test or ANOVA
Two numeric		correlation test

Diploma in Data Science & Big Data Analytics

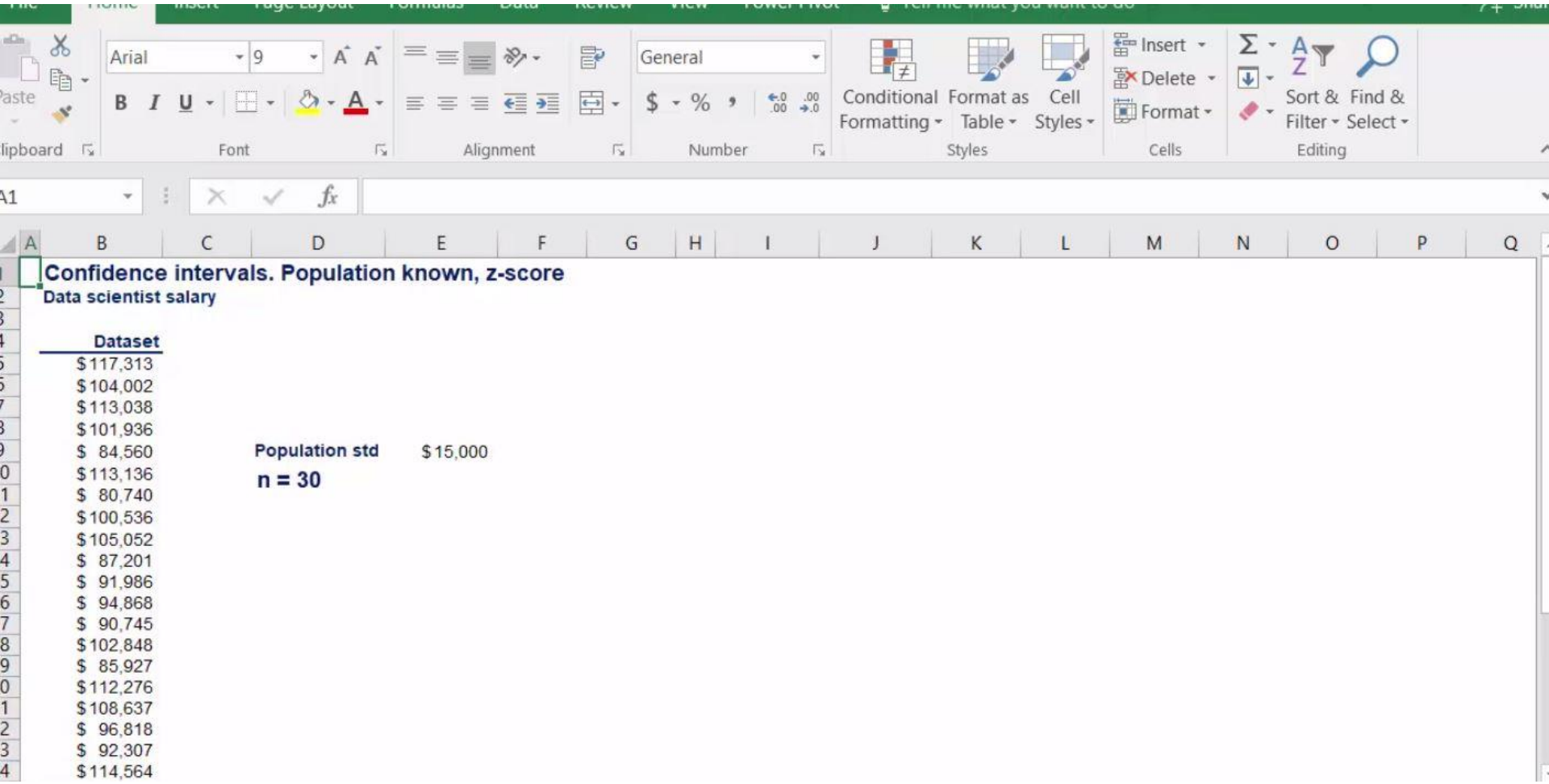
By Navin Pathuru at Naresh IT Opposite Satyam Theatre, Ameerpet, Hyderabad 040-23746666, 23734842

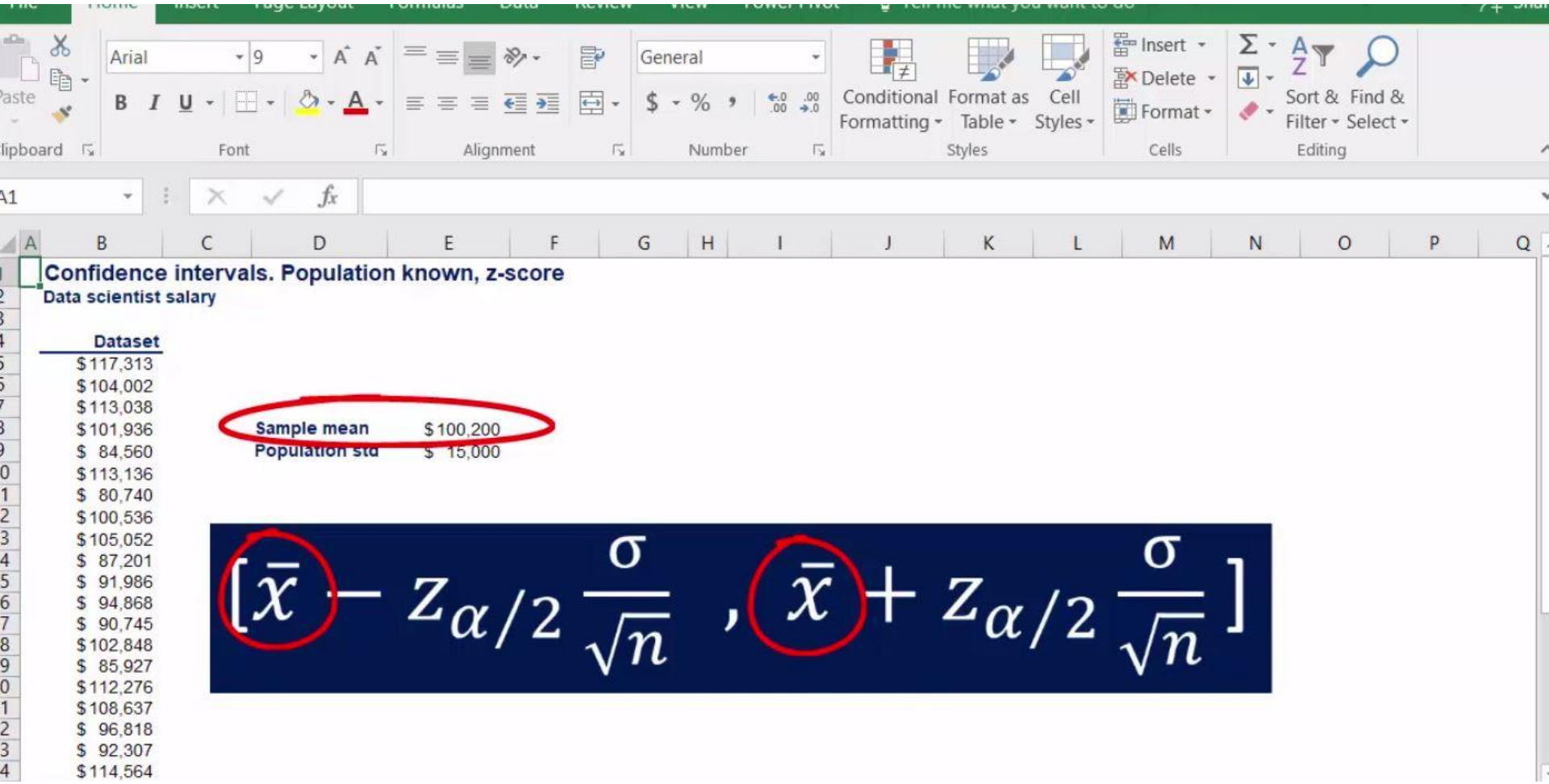

1
QUESTION AND HYPOTHESIS
2
NULL HYPOTHESIS AND ALPHA VALUE
3
ANALYSE DATA

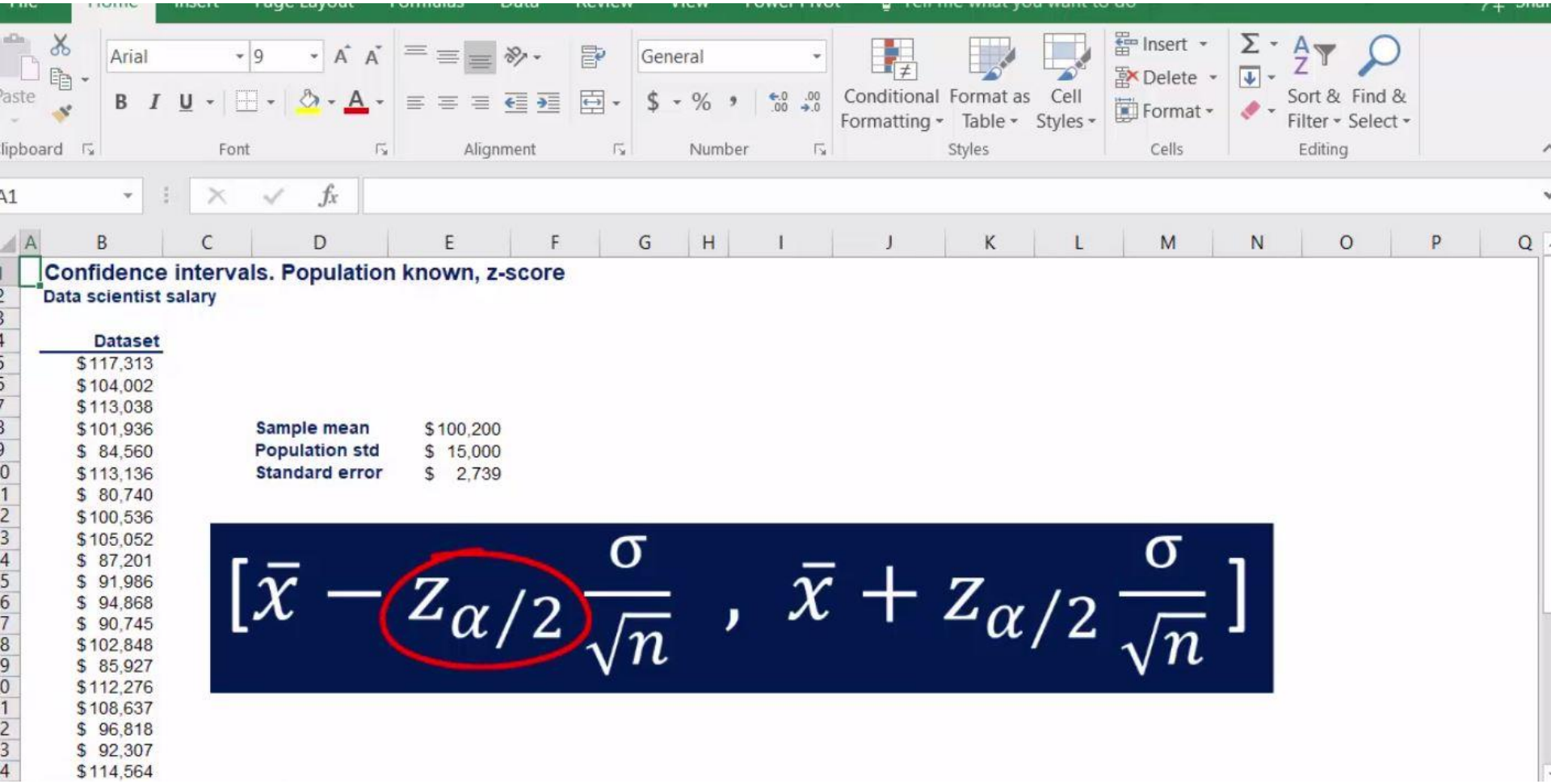
Agenda

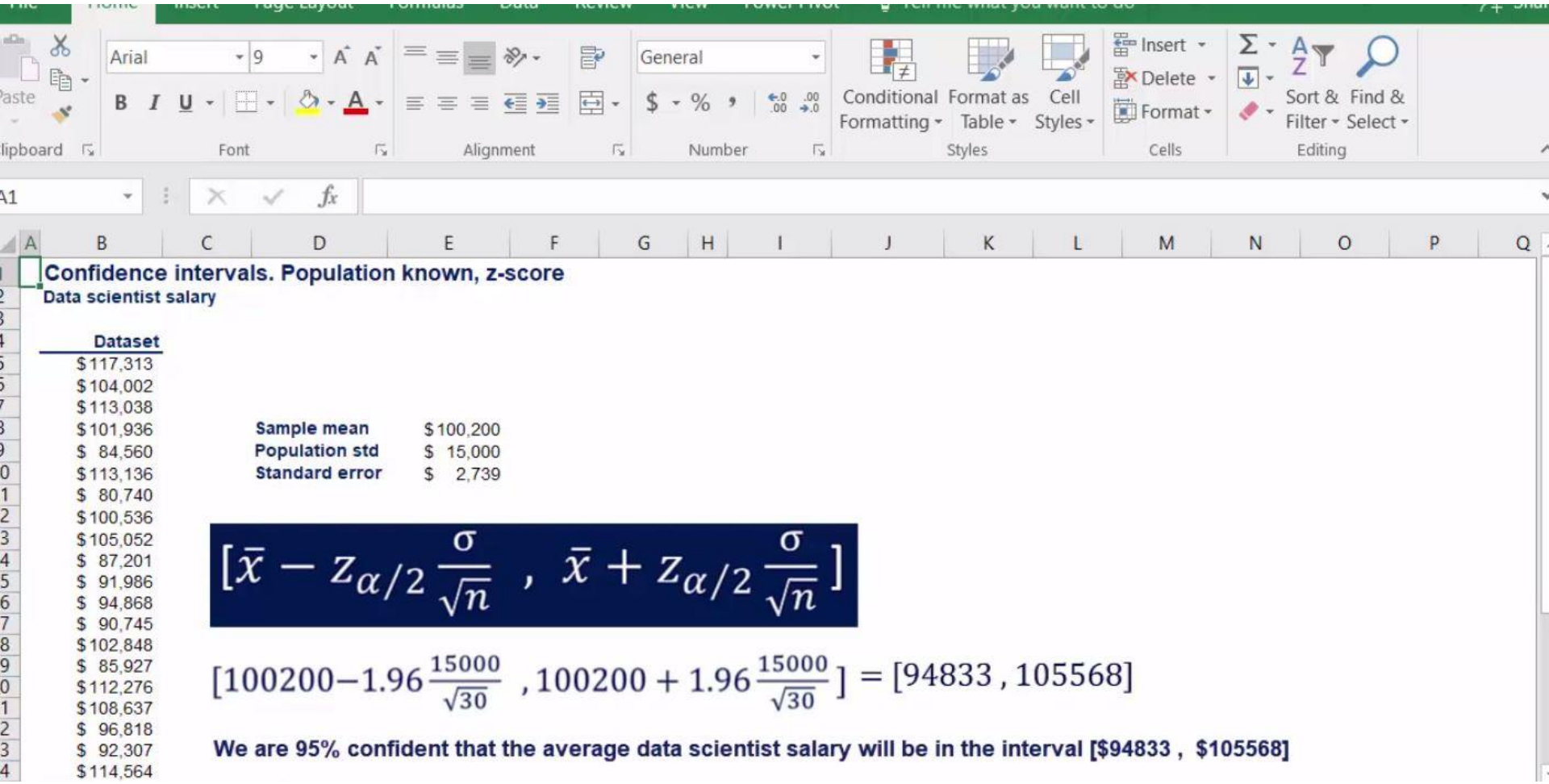
- Overview to Hypothesis Testing
  - Chi-Square Test
  - Test for Continuous Data
  - Non-Normal Data
  - Correlation and Regression

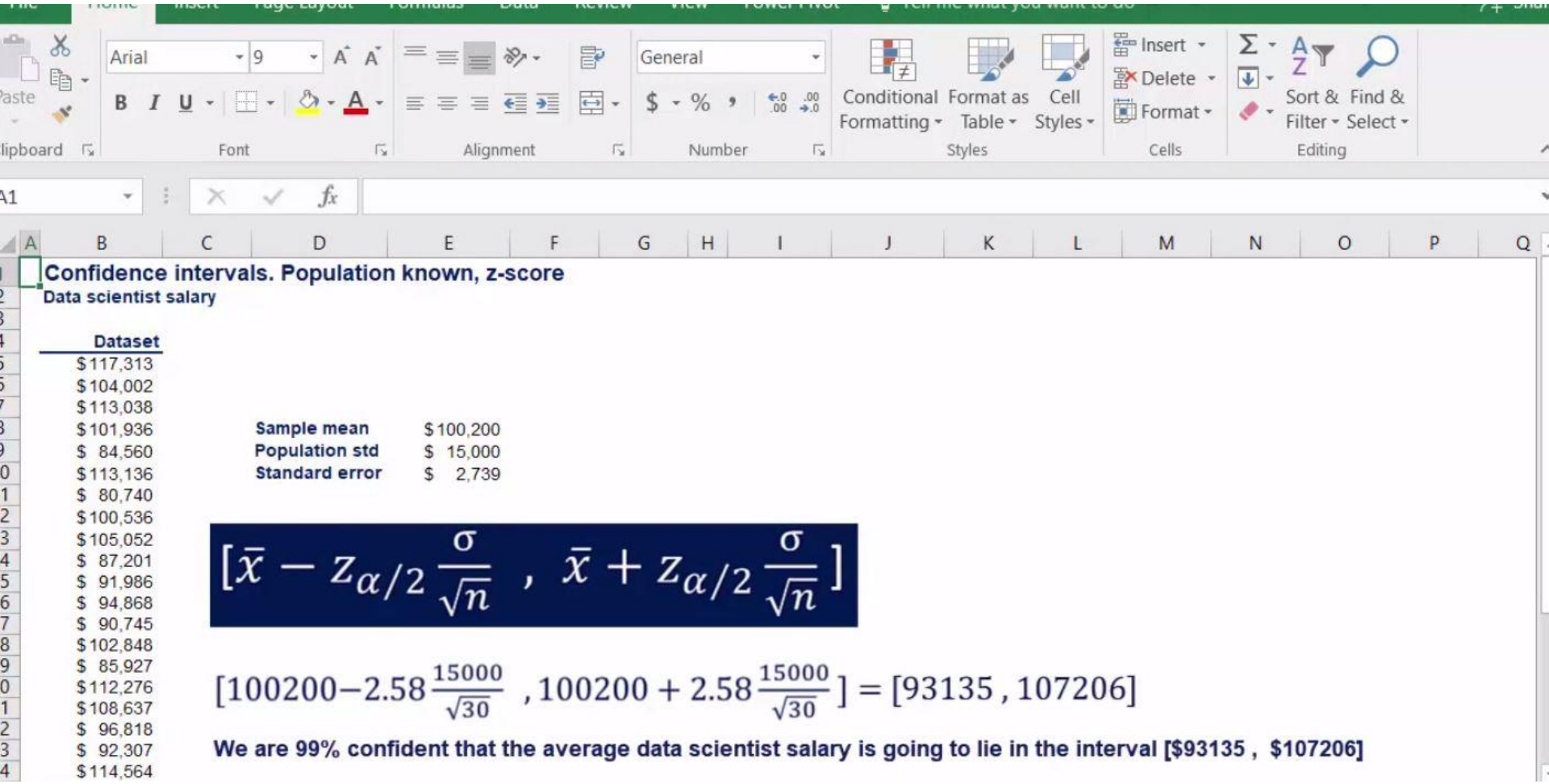
Example


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Population std
$15,000
n = 30


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$
$[100200-1.96\frac{15000}{\sqrt{30}} , 100200 + 1.96\frac{15000}{\sqrt{30}}] = [94833 , 105568]$
We are 95% confident that the average data scientist salary will be in the interval [$94833 , $105568]



Confidence intervals. Population known, z-score

Data scientist salary

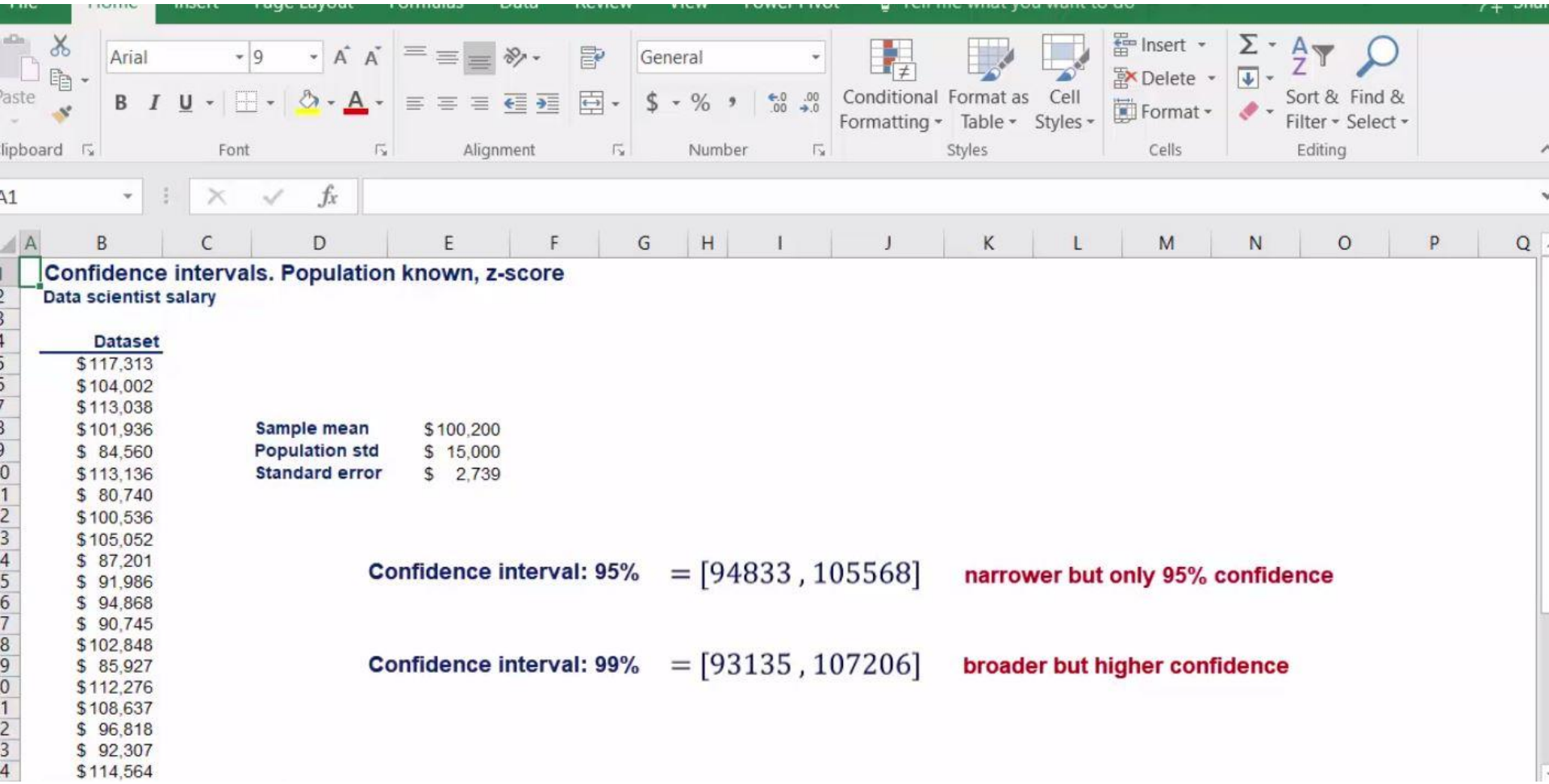
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564

Sample mean	$100,200
Population std	$ 15,000
Standard error	$ 2,739

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$\left[100200 - 2.58\frac{15000}{\sqrt{30}} \;, 100200 + 2.58\frac{15000}{\sqrt{30}}\right] = [93135\,,\,107206]$$

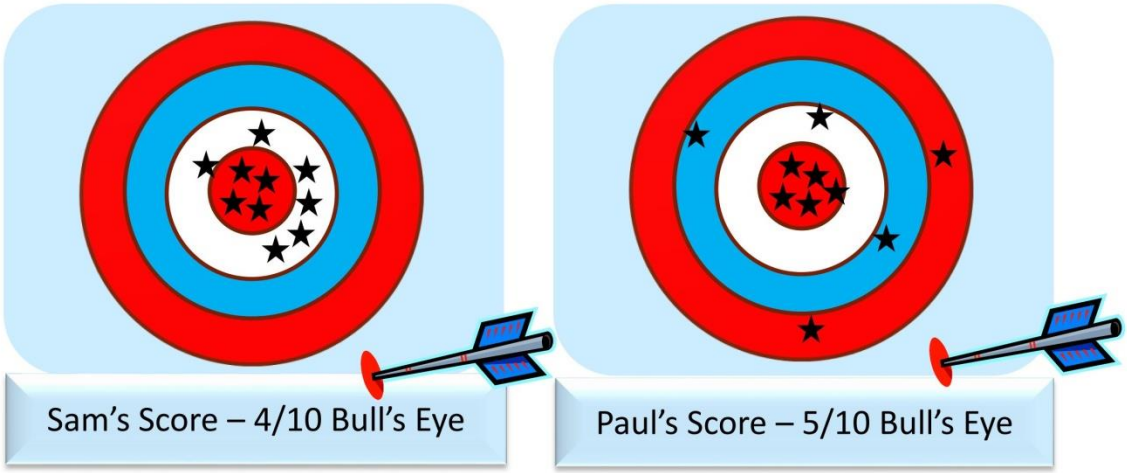
We are 99% confident that the average data scientist salary is going to lie in the interval [$93135 , $107206]



Diploma in Data Science & Big Data Analytics

Bull's Eye

Sam and Paul are throwing darts at the local sports bar. A few of their friends start a betting pool. Both Sam and Paul shoot 10 practice shots each so that their friends can decide their bets.



Discussion

Who is a better bet: Sam or Paul?

Discussion

- Earth is the center of universe
- Earth is flat
- Continents do not move
- Stress causes Ulcers
- 10,000 hours of appropriately guided practice is “the magic number of greatness” (Malcolm Gladwell)
- Men are better drivers than women

What are these statements? How were they proved or disproved?

Hypothesis: Business Examples

- No difference in performance of the sales team across geographies and product lines
- Change in gas price will have no impact on losses in automotive finance
- Change in CEO will have no impact on the stock price
- Real estate yields are the same in all metros
- Compensation changes will not impact attrition

Hypothesis Testing

- Is a method of making an inference about a population parameter based on sample data
- Is statistical analysis used to determine if the difference observed in samples is not a random occurrence but a true difference

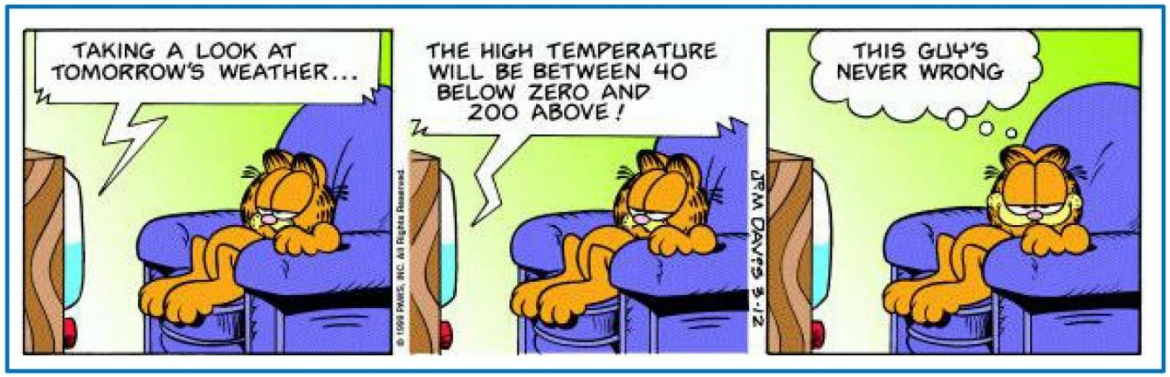
Key Terms

Three key terms that you need to understand in Hypothesis Testing are:

Confidence Interval	Measure for reliability of an estimate; sample is used for estimating a population parameter so we need to know the reliability of that estimate
Degrees of Freedom	Number of values that are free to vary in a study
P-value	Probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the Null Hypothesis is true

Confidence Interval

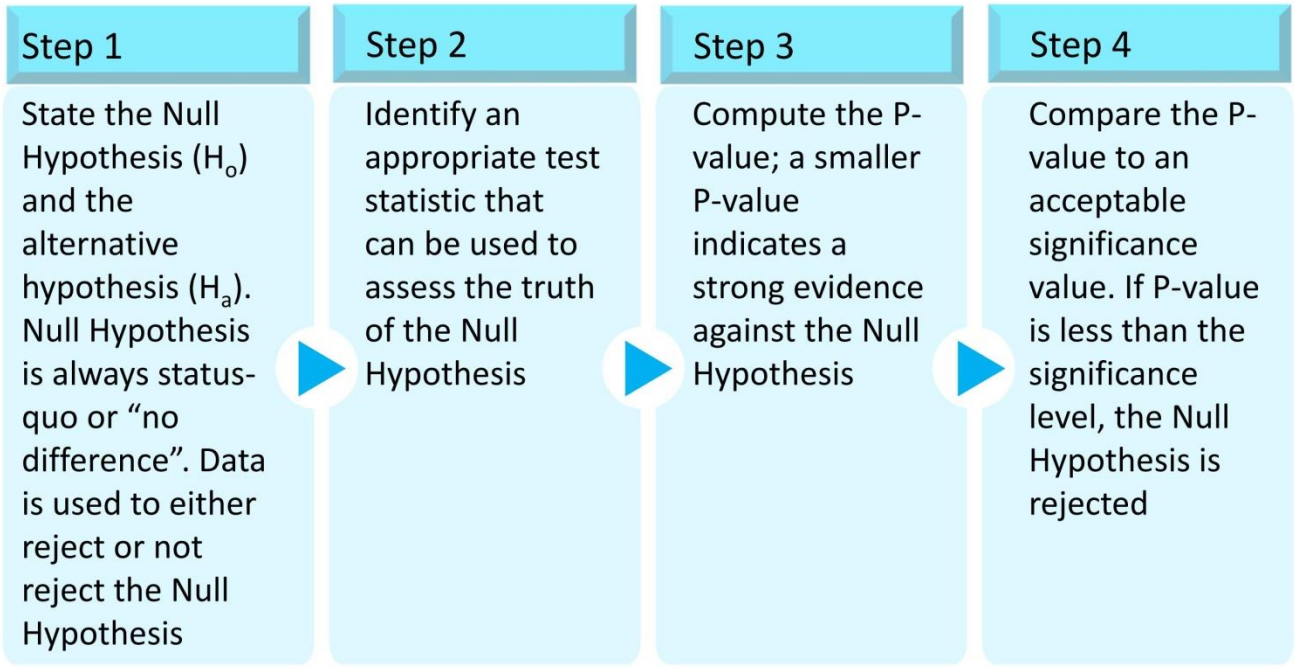
- Confidence interval
  - Describe the reliability of an estimate
  - Range of values (lower and upper boundary) within which the population parameter is included
  - Width of the interval indicates the uncertainty associated with the estimate
- Confidence level
  - Probability associated with the confidence interval



100% Confidence Interval

Process of Hypothesis Testing

The process of Hypothesis Testing consists of four steps:



Justice

Null Hypothesis = “Person is innocent”

		Decision	
		Prison	Set free
True State	Innocent	Type I error	Correct decision
	Guilty	Correct decision	Type II error

Points to Remember

Important points to note regarding Hypothesis Testing:

1. It is always “Reject” or “Do Not Reject” the Null Hypothesis
2. Rejecting the Null Hypothesis means there is evidence that there is a difference (based on the sample data)
3. Failure to reject the Null Hypothesis means data is insufficient to conclude that there is a difference

What is a Hypothesis?

- A hypothesis is an assumption which we make about a population parameter.
- The hypothesis which we wish to test is called the null hypothesis because it implies that there is no difference between the true value and the hypothesized value.

Cont..

- A thesis is something that has been proven to be true. However, a hypothesis is something that has not yet been proven to be true.
- Hypothesis testing is the process of determining whether or not a given hypothesis is true.
- Hypothesis testing along with estimation forms the foundation of inferential statistics.

Measures of Central Tendency: Review Example

House Prices:

$2,000,000
$500,000
$300,000
$100,000
$100,000
Sum $3,000,000

- **Mean:** $3,000,000/5
  = **$600,000**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Median:** middle value of ranked data

$$\frac{n+1}{2}$$

  $(5+1)/2 = 3^{rd}$ position
  = **$300,000**

- **Mode:** most frequent value
  = **$100,000**

Agenda

- What is Statistics?
- **Central Tendency Measures**
- Dispersion Measures
- Data Distributions

Central Tendency

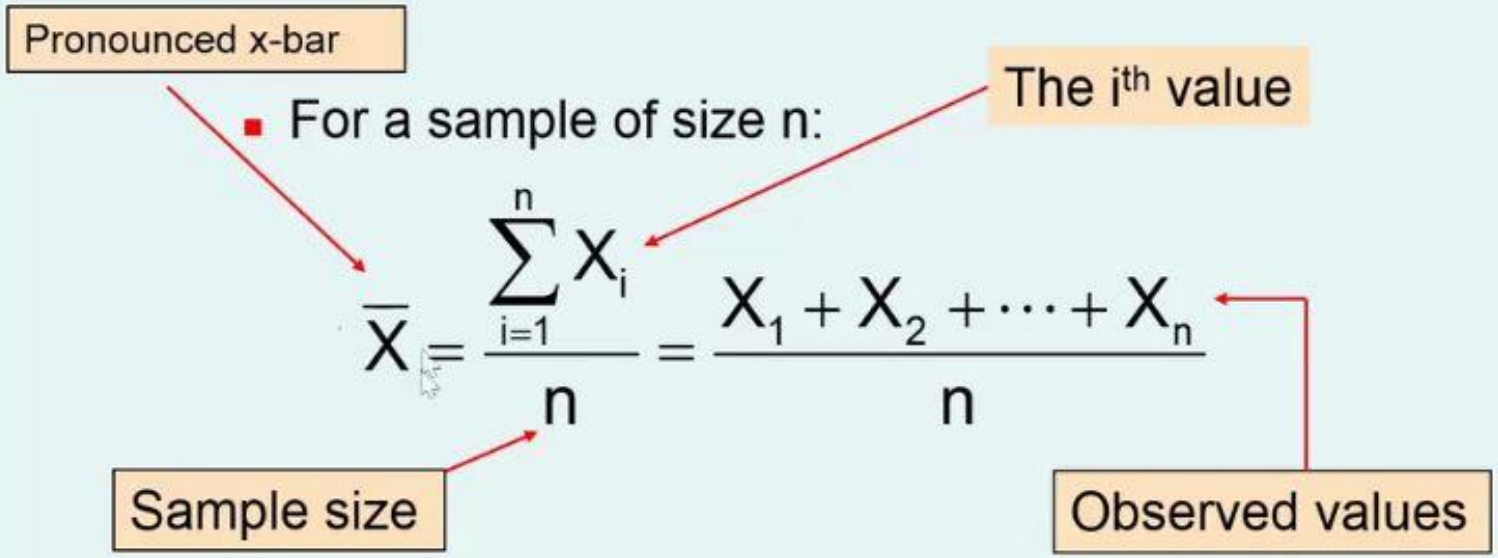
A measure of **Central Tendency** is a single value that attempts to describe a set of data **by identifying the central position** within that set of data. In other words, the Central Tendency computes the "center" around which the data is distributed.

The three measures of Central Tendency are:

- Mean
- Median
- Mode

Measures of Central Tendency: The Mean

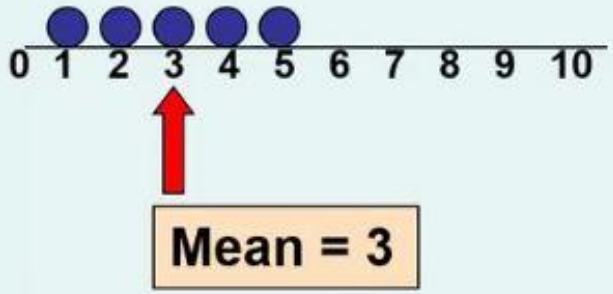
- The arithmetic mean (often just called "**mean**") is the most common ***measure of central tendency***



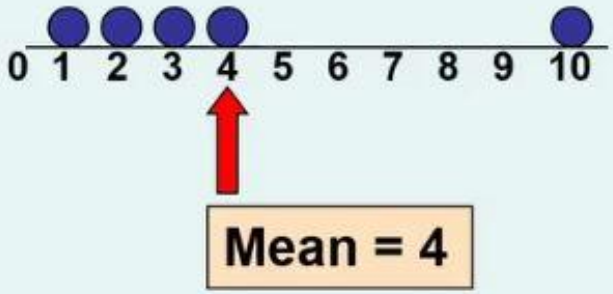
Measures of Central Tendency: The Mean

(continued)

- Mean = sum of values **divided** by the number of values
- Affected by extreme values (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

The "Hotshot" Sales Executive

Kurt works as a sales manager at vsellhomes.com. In the monthly sales review, Kurt reports that he will achieve his quarterly target of $1M.

Kurt claims his average deal size is $100,000 and he has 10 deals in his pipeline. Kurt's boss Ross is very delighted with his numbers.

At the end of quarter, even after closing 8 deals Kurt fails to meet his target number and falls short by more than $500,000.

Discussion

Why did Kurt fail to achieve his quarterly target?

With 10 deals in pipeline and with average deal size of $100,000 and converting 7 of those deals, how did he fail?

The Reality of the "Hotshot" Salesman

- Average deal size in pipeline

  = $100,000

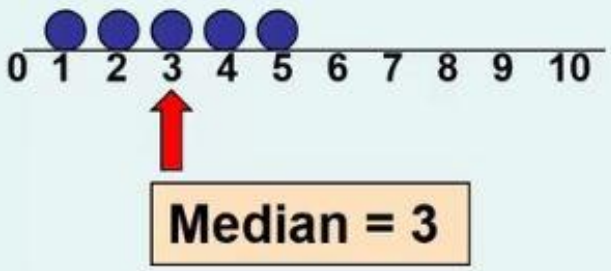
- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation
- Median = $55,000 more realistic measure

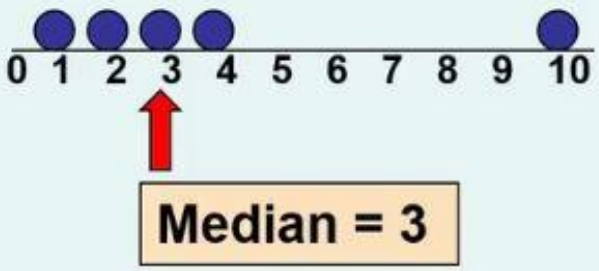
Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

Median is less susceptible to the influence of outliers.

Measures of Central Tendency: The Median

- In an ordered array, the **median** is the "middle" number (50% above, 50% below)





- **Not** affected by extreme values (outliers)

Measures of Central Tendency: Locating the Median

- First arrange the values in *numerical order* (smallest to largest) to find the **median**:

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- If the number of values is odd, the median is the middle number

- If the number of values is even, the median is the average of the two middle numbers

*Note that $\frac{n+1}{2}$ is not the *value* of the median, only the ***position*** of the median in the ranked data.

Median

The middle value

Example: 1, 2, 3, 4, 5

Median = 3

Median

Example: 1, 2, 3, 4, 5, 6

Two middle scores: 3, 4

To find the median, take the average of the two middle scores: (3+4)/2 = 3.5

Median = 3.5

Median

Odd N: When there are an odd number of values, the median is the middle score

(1, 2, 3, 4, 5; N=5) median = 3

Even N: When there are an even number of values, the median is equal to the average of the two middle scores

(1, 2, 3, 4, 5, 6; N=6) median = 3.5

Median

Prior to calculating the median, be sure that the numbers are ordered from smallest to largest (don't pick the middle number of a set of numbers if they are not first ordered)

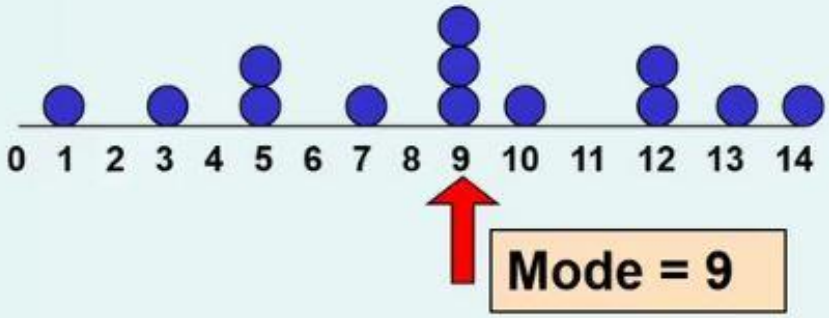
Example: 2, 3, 1, 3, 1, 6

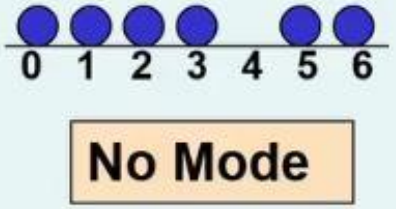
Reordered: 1, 1, 2, 3, 3, 6

Median = 2.5

Measures of Central Tendency: The Mode

- Value that occurs **most often**
- There may be no mode
- There may be several modes





Mode

The most frequently occurring score

Example: 1, 2, 2, 2, 3, 3, 4

Mode = 2

Mode

Example: 1, 2, 2, 3, 3, 4

Bimodal (two modes) = 2, 3

Mode

Example: 1, 2, 2, 3, 3, 4, 4

Multimodal (three or more modes) = 2, 3, 4

Measures of Central Tendency: Review Example

House Prices:

$2,000,000
$500,000
$300,000
$150,000
$130,000
$120,000
Sum $3,200,000

- **Mean:** $3,200,000/6
  = **$533,333.33**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Median:** middle value of ranked data

$$\frac{n+1}{2}$$

  (6+1)/2 = 3.5 position
  = **($300,000 + $150,000)/2**
  = **$225,000**

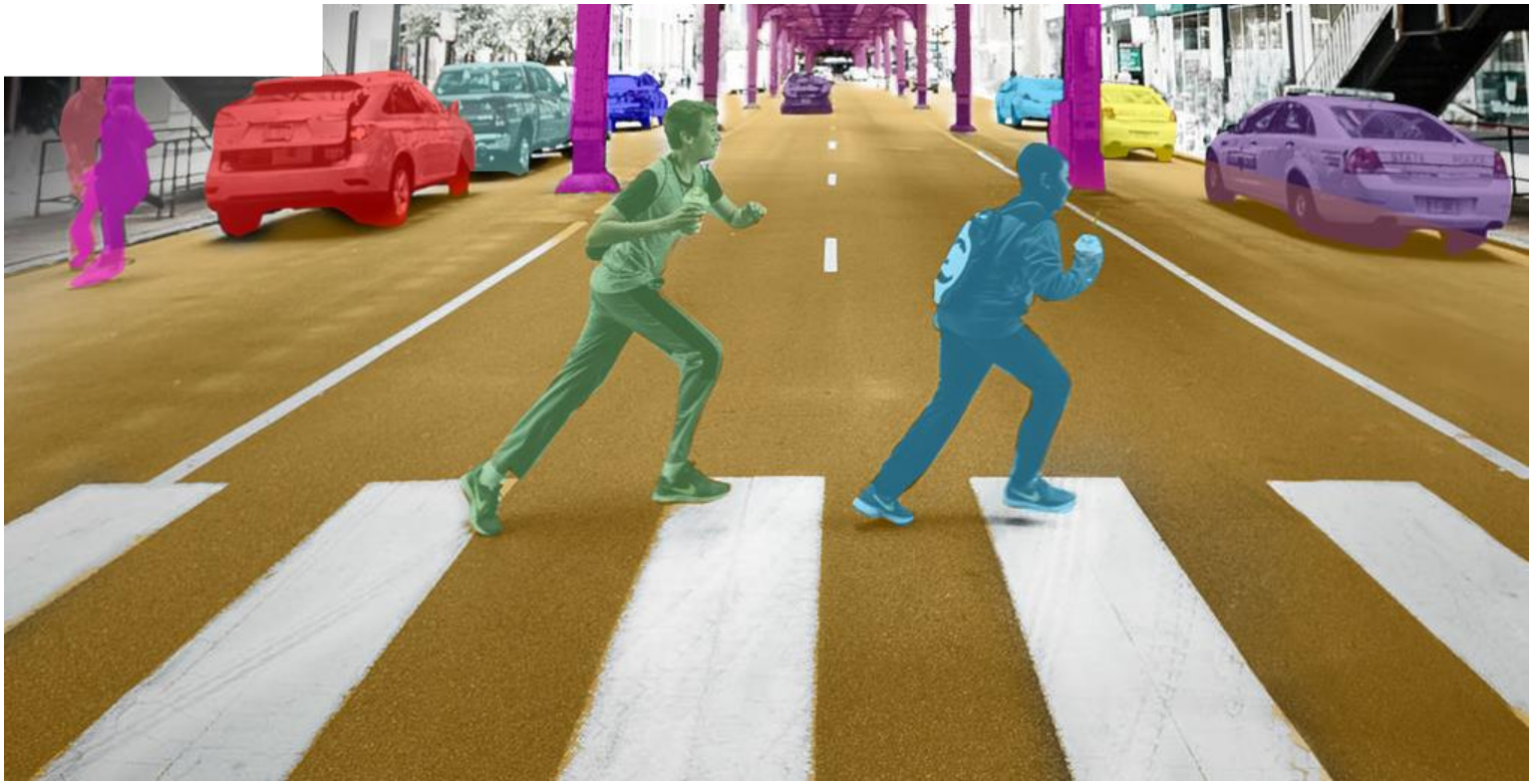
- **Mode:** most frequent value
  = **N/A**

What Data Scientists do?

Word2Vec – Extracting sentence meaning

"Sachin Tendulkar is the Roger Federer of Cricket"

Roger Federer – tennis + cricket = Sachin Tendulkar



What's the first thing you do when you're attempting to cross the road?

Look left and right, take stock of the vehicles on the road, and make a decision.

Our brain is able to analyze, in a matter of milliseconds, what kind of vehicle (car, bus, truck, auto, etc.) is coming towards us.

Can machines do that?

Computer Vision

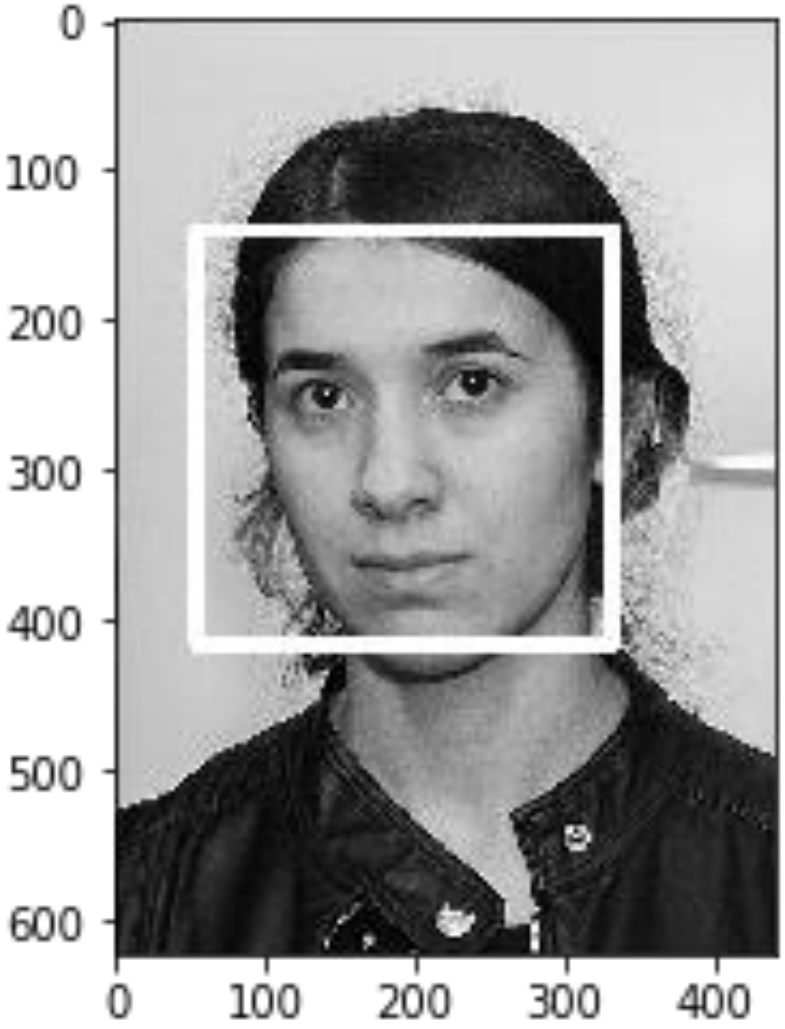
‘no’ till a few years back.

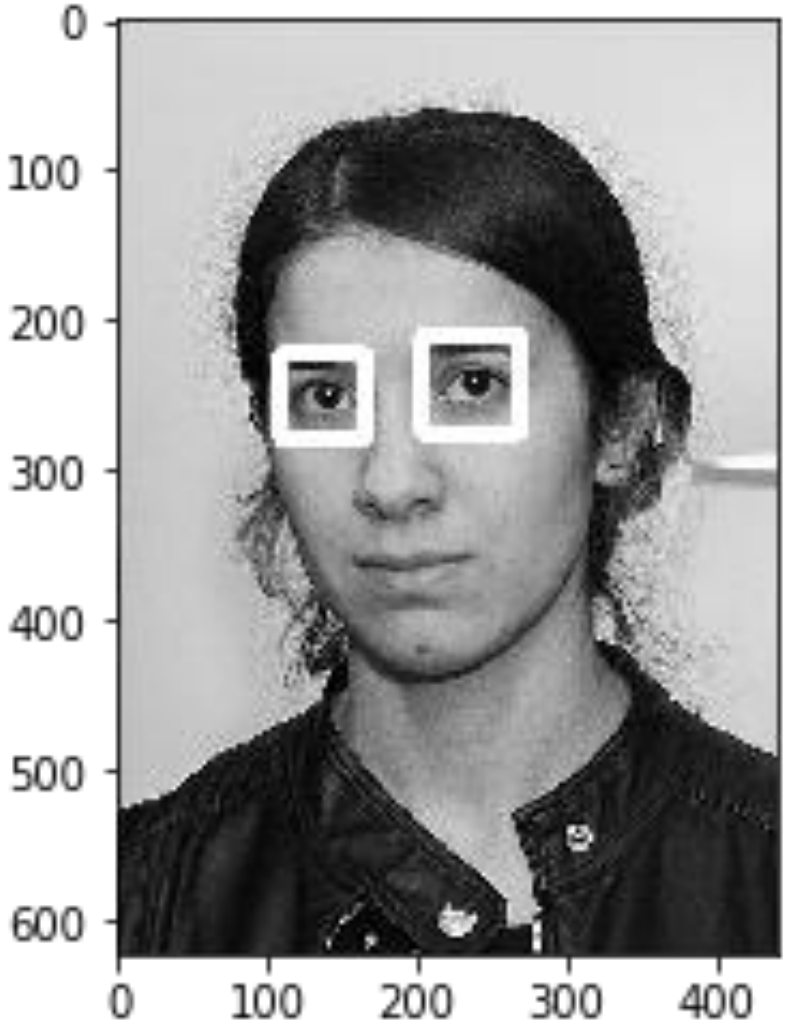
But computer vision has changed the game.

We can build computer vision models that can detect objects, determine their shape, predict the direction the objects will go in, and many other things.

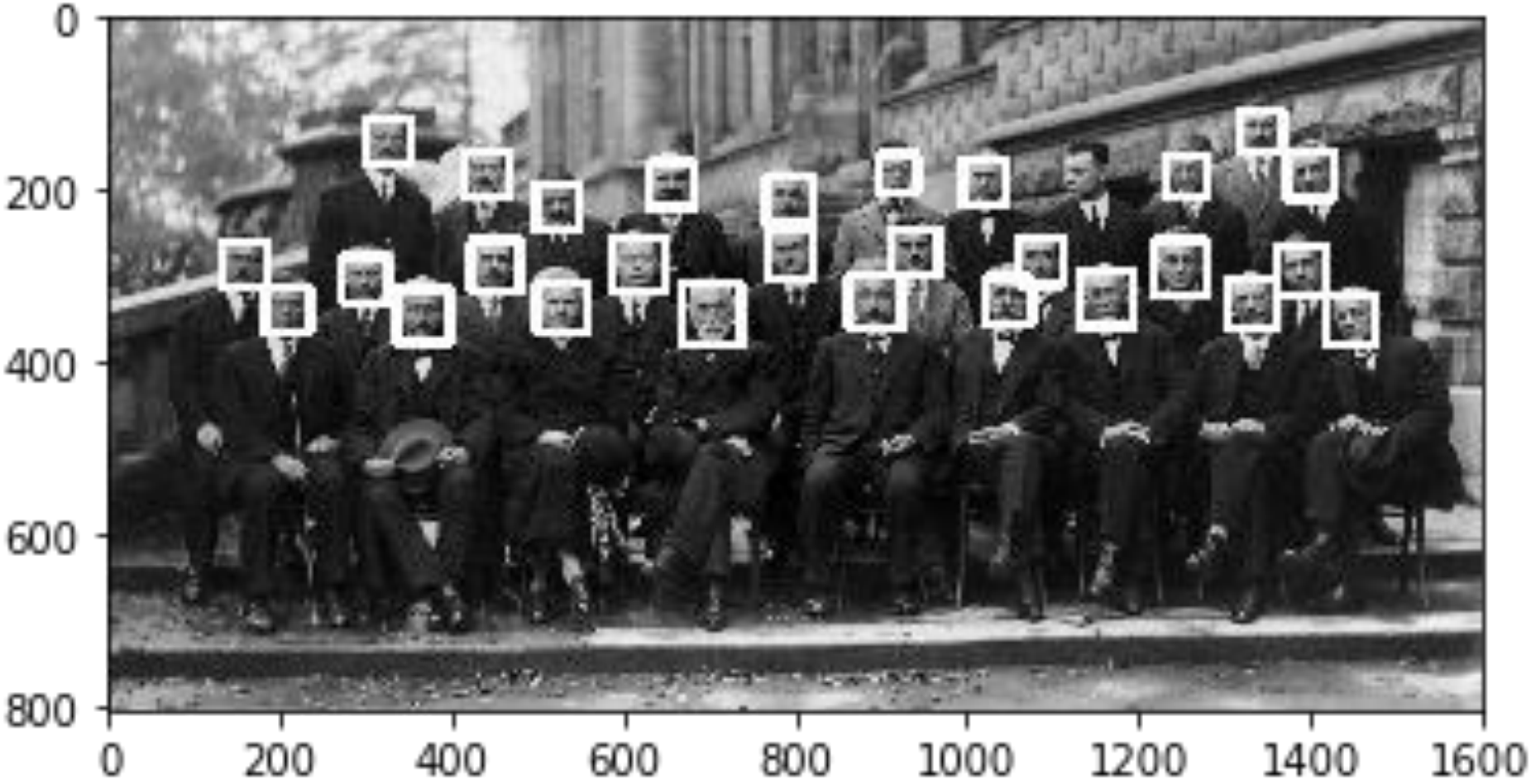
That’s the powerful technology behind self-driving cars!

What Data Scientists do?




0
100
200
300
400
500
600
0
100
200
300
400

What Data Scientists do?



Diploma in Data Science & Big Data Analytics

What Data Scientists do?

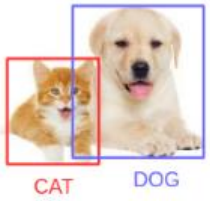
Image localization comes into the picture

It helps us to identify the location of a single object in the given image.

In case we have multiple objects present, we then rely on the concept of object detection (OD).

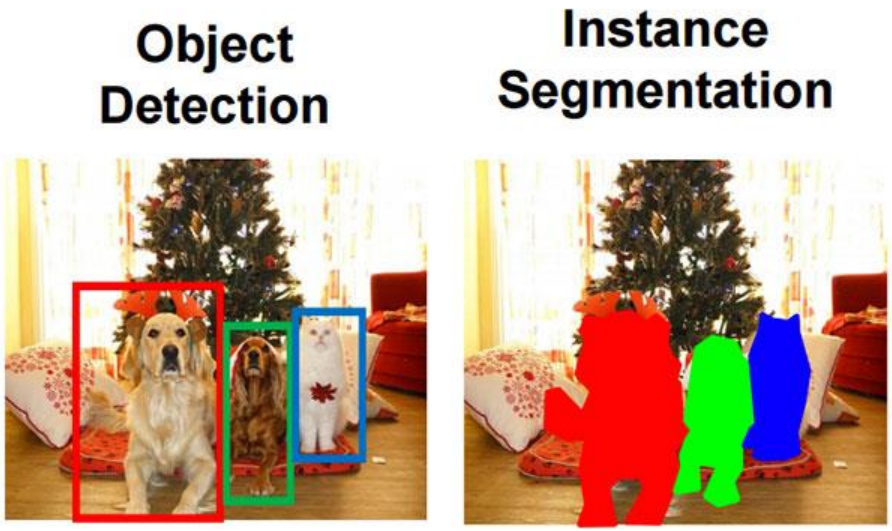
We can predict the location along with the class for each object using OD.

Image Localization



Object Detection

What Data Scientists do?

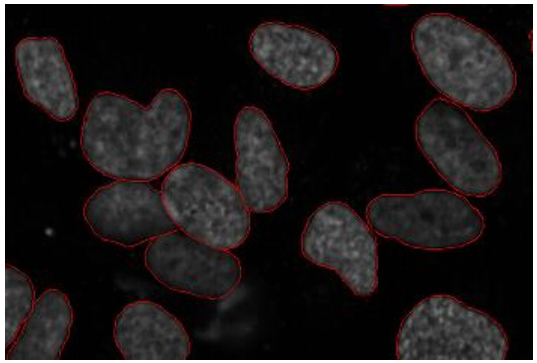


Object detection builds a bounding box corresponding to each class in the image.

But it tells us nothing about the shape of the object.

We only get the set of bounding box coordinates.

Image segmentation creates a pixel-wise mask for each object in the image. This technique gives us a far more granular understanding of the object(s) in the image.



Cancer has long been a deadly illness.

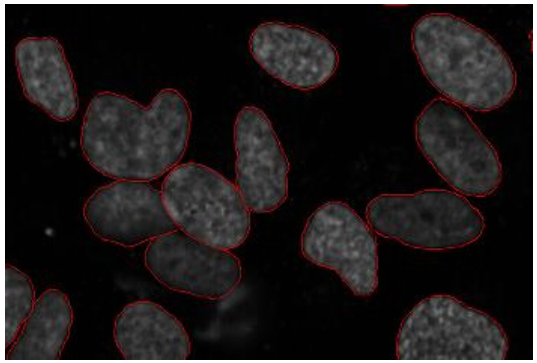
Even in today's age of technological advancements, cancer can be fatal if we don't identify it at an early stage.

Detecting cancerous cell(s) as quickly as possible can potentially save millions of lives.

The shape of the cancerous cells plays a vital role in determining the severity of the cancer.

Object detection will not be very useful here.

What Data Scientists do?



Computer Vision techniques make a MASSIVE impact here.

There are many other applications where vision is transforming industries:

•Traffic Control Systems

•Self Driving Cars

•Locating objects in satellite images

In image 1, every pixel belongs to a particular class (either background or person).

All the pixels belonging to a particular class are represented by the same color (background as black and person as pink). This is semantic segmentation

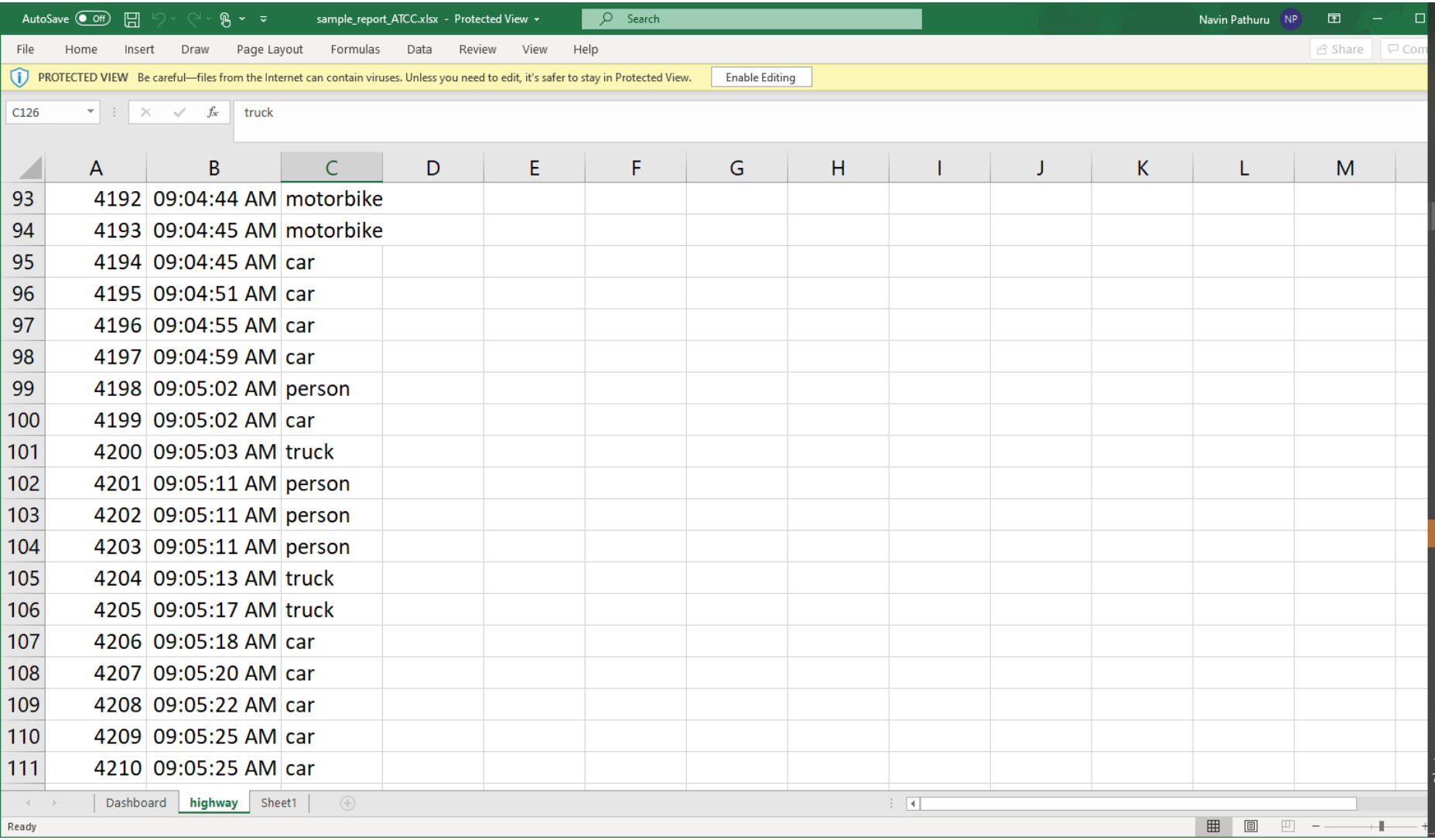
Image 1

Image 2

Mask R-CNN

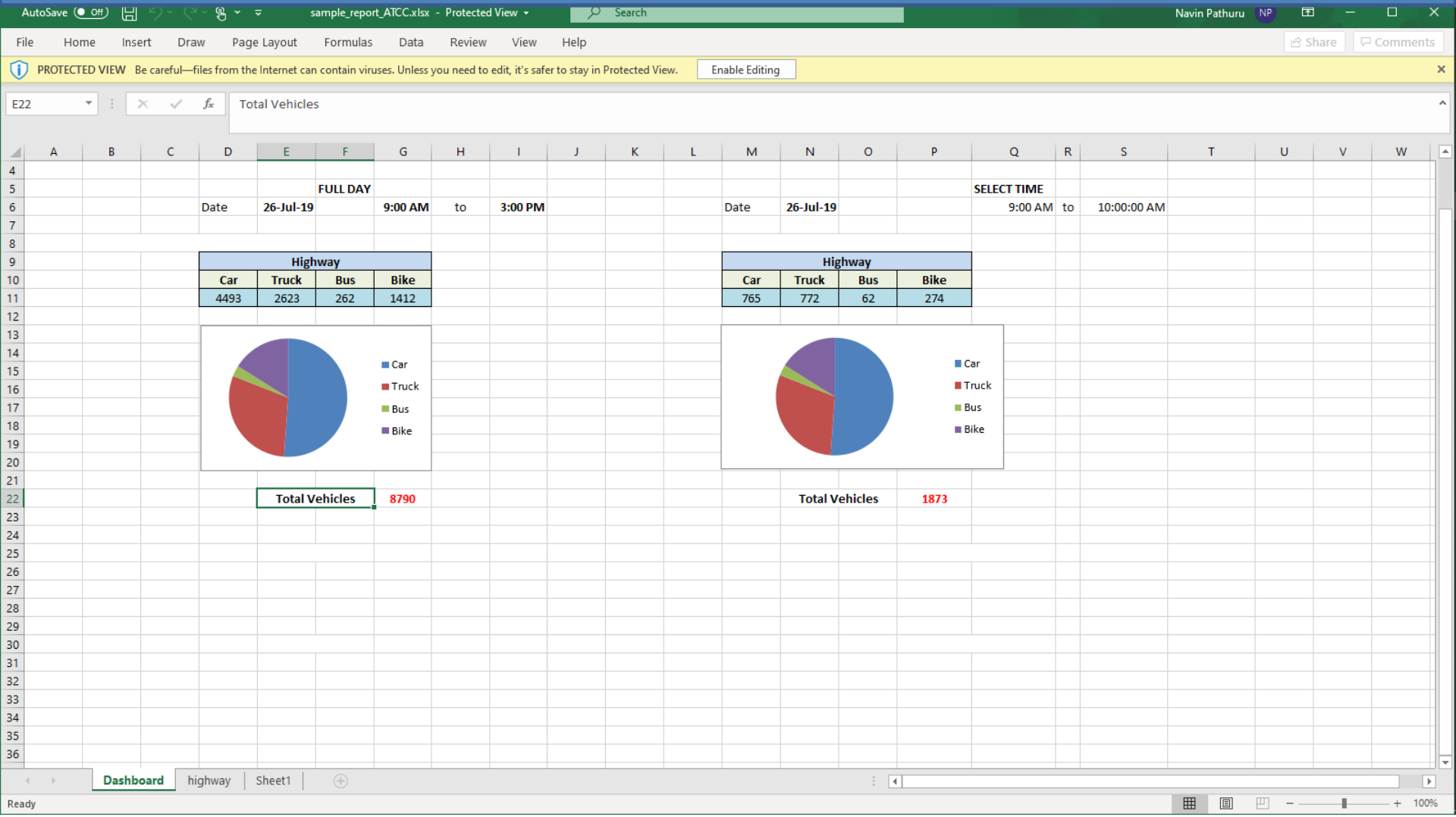

car 0.904
car 0.911
car 0.908
car 0.934
car 0.986
car 0.996
car 0.995
car 0.981
car 0.996
car 0.994
car 0.993
car 0.995
car 0.995

What Data Scientists do?



	A	B	C	D	E	F	G	H	I	J	K	L	M
93	4192	09:04:44 AM	motorbike										
94	4193	09:04:45 AM	motorbike										
95	4194	09:04:45 AM	car										
96	4195	09:04:51 AM	car										
97	4196	09:04:55 AM	car										
98	4197	09:04:59 AM	car										
99	4198	09:05:02 AM	person										
100	4199	09:05:02 AM	car										
101	4200	09:05:03 AM	truck										
102	4201	09:05:11 AM	person										
103	4202	09:05:11 AM	person										
104	4203	09:05:11 AM	person										
105	4204	09:05:13 AM	truck										
106	4205	09:05:17 AM	truck										
107	4206	09:05:18 AM	car										
108	4207	09:05:20 AM	car										
109	4208	09:05:22 AM	car										
110	4209	09:05:25 AM	car										
111	4210	09:05:25 AM	car										

What Data Scientists do?



Diploma in Data Science & Big Data Analytics

By Navin Pathuru at Naresh IT Opposite Satyam Theatre, Ameerpet, Hyderabad 040-23746666, 23734842