Exploratory Data Analysis

Types of Data

Structured Data

Databases

Datawarehouse

Un-Structured Data

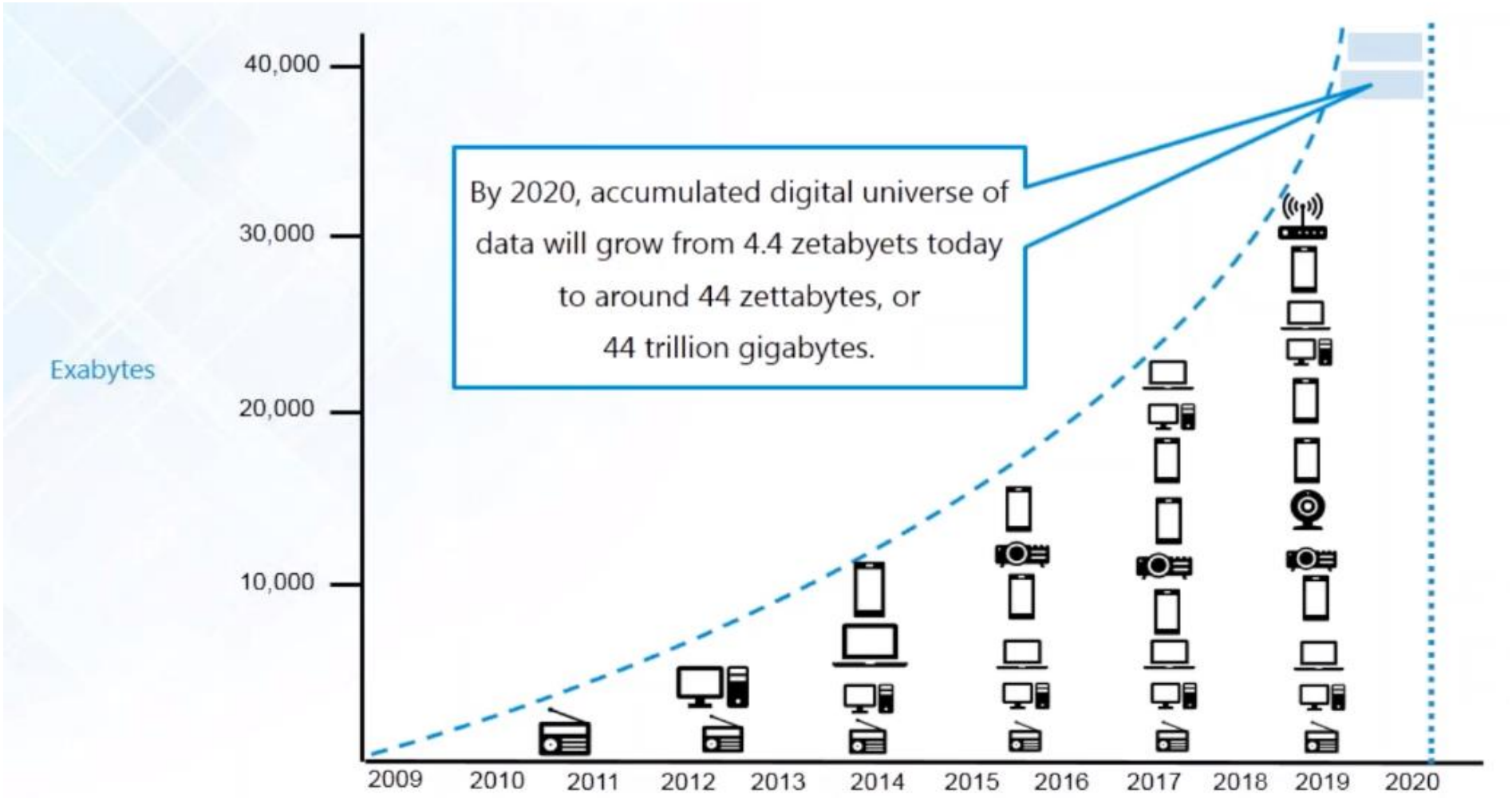
The concept of **labeled data is different** from the types of data.

	Structured Data	Unstructured Data
Labeled	✓	✓

Importance of Data

2.5 Quintillion Bytes is equal to 25,00,000 Tera Bytes

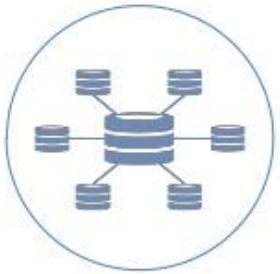
Volume of Unstructured Data



Why Big Data ?

Big Data?

"Information is the oil of the 21st century, and analytics is the combustion engine."
Peter Sondergaard, Senior Vice President, Gartner Research

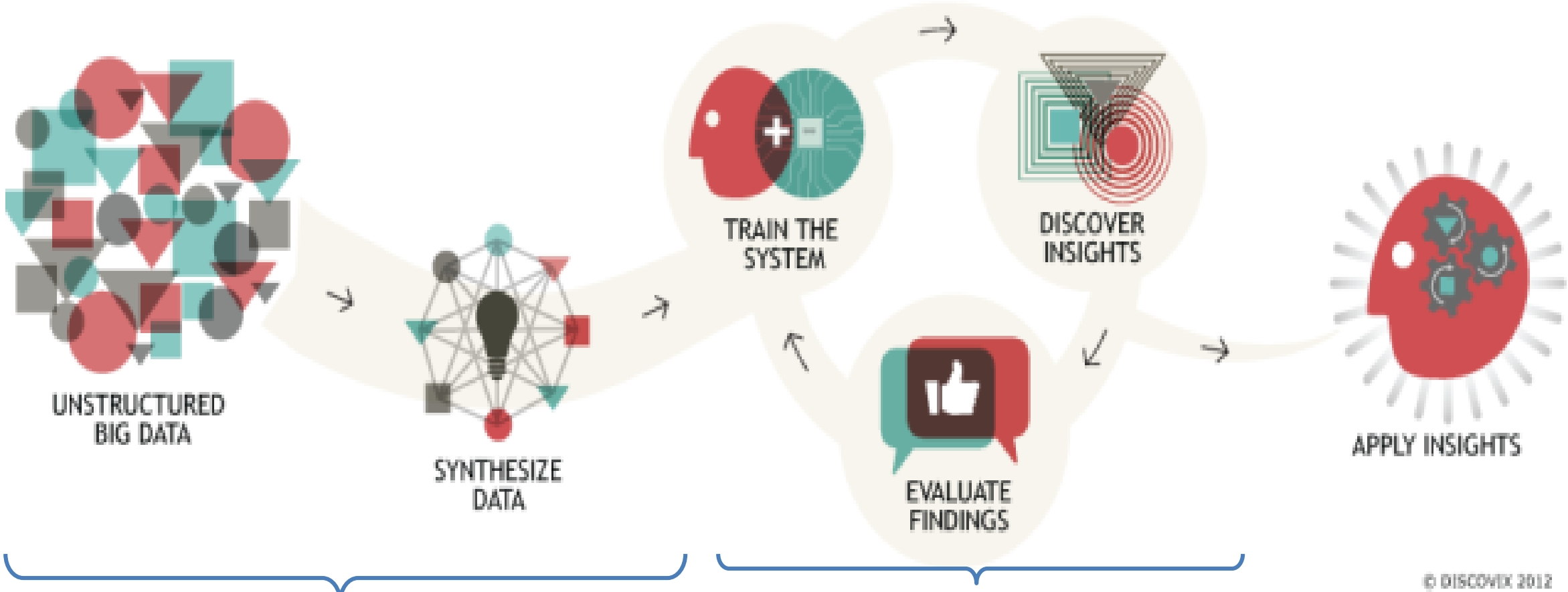


Explosion of data size

Falling cost of data storage

Increase of computing power

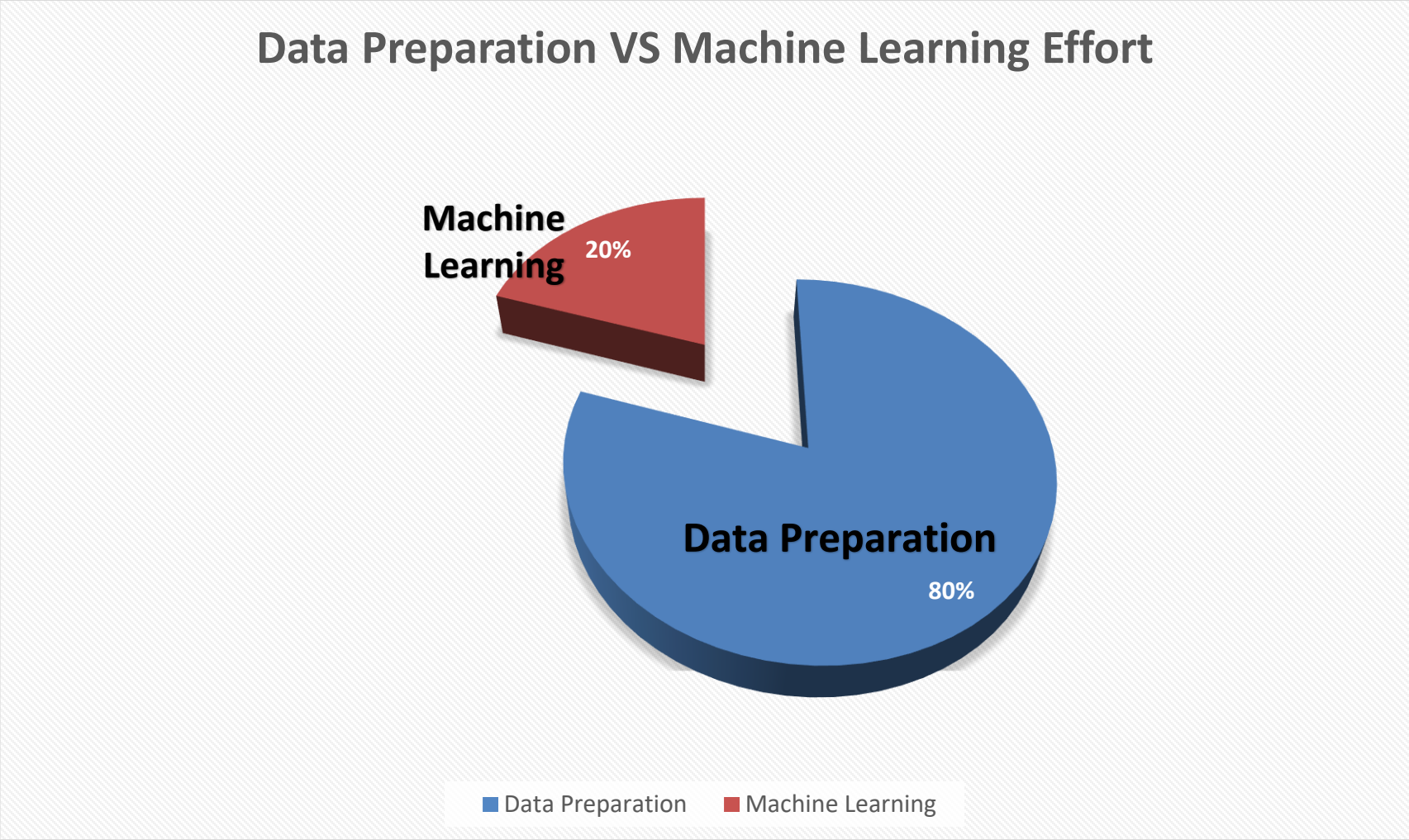
Big Data & Machine Learning



Data Preparation

Machine Learning

Data Preparation Vs ML effort



Data Exploration & Preparation

Steps in Data Exploration & Preparation

1. Variables
2. Missing Values
3. Binning
4. Univariate Analysis
5. Bi-variate Analysis
6. Preparation of Training & Test Data

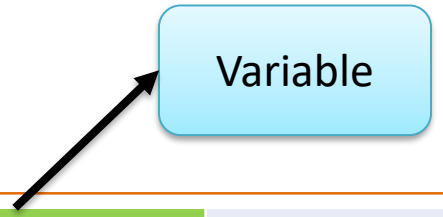
Cases and Variables

We obtain information about ***cases*** or ***units***.

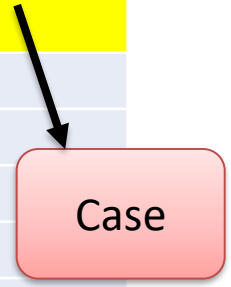
A ***variable*** is any characteristic that is recorded for each case.

Generally each case makes up a row in a dataset, and each variable makes up a column

Example Data Set



ID	Salary	Years	Courses
1	67483	5.8	13
2	77204	7.4	18
3	64972	6.8	23
4	94143	8.3	35
5	78954	8.1	19
6	65154	6.5	12
7	80849	8.2	29
8	83860	9.5	21
9	81909	7	27
10	71335	8.4	19
11	93141	8.6	34
12	79678	6.8	19
13	67545	6	15
14	48424	5.1	4
15	88499	9	28
16	74461	7.1	21
17	74806	6.5	23
18	86326	8	27



Types of Variables

Quantitative (Continuous/Discrete) variables:

- Always numeric
- Can be any number, positive or negative
- Examples: age in years, weight, blood pressure readings, temperature, concentrations of pollutants and other measurements

Qualitative (Categorical) variables:

- Information that can be sorted into categories
- Types of categorical variables – ordinal, nominal

A quick check on knowledge

	A	B	C	D	E	F	G	H	I	J	K	L	
1	Year	Gender	HigherSAT	SAT	GPA	Sibling	Height	Weight	Exercise	TV	Pulse	Award	Pie
2	Senior	M	Math	1210	3.1	4	71	180	10	1	54	Olympic	
3	Sophomore	F	Math	1150	2.5	2	66	120	4	7	66	Academy	
4	FirstYear	M	Math	1110	2.6	2	72	208	14	5	130	Nobel	
5	Junior	M	Math	1120	3.1	1	63	110	3	1	78	Nobel	
6	Sophomore	F	Verbal	1170	2.7	1	65	150	3	3	40	Nobel	
7	Sophomore	F	Verbal	1150	3.2	2	65	114	5	4	80	Nobel	
8	FirstYear	F	Math	1320	2.8	1	66	128	10	10	94	Olympic	
9	Sophomore	M	Math	1370	3.3	1	74	235	13	8	77	Olympic	
10	Junior	F	Verbal	1100	2.8	2	61		3	6	60	Nobel	
11	FirstYear	F	Math	1370	3.7	7	60	115	12	1	94	Nobel	
12	Sophomore	F	Math	1170	2.1	1	65	140	12	6	63	Olympic	
13	First-year	M	Math	1180		2	63	200	10	5	72	Olympic	
14	Sophomore	M	Math	1150	2.9	3	68	162	12	8	54	Olympic	
15	Junior	F	Verbal	1300	3.1	2	68	135	6	1	66	Nobel	
16	First-year	M	Verbal	1350		1	68	193	9	5	72	Nobel	
17	FirstYear	F	Math	1200	3.9	1	63	110	10	2	59	Olympic	
18	FirstYear	F	Verbal	1200	3	2	63	99	3	15	88	Olympic	
19	Sophomore	M	Verbal	1350	3	2	72	165	7	3	59	Nobel	
20	Sophomore	F	Math	1410	3.4	1	62	120	2	1	64	Nobel	

Categorical

Quantitative

Predictor VS Target variables

Predictor – Features

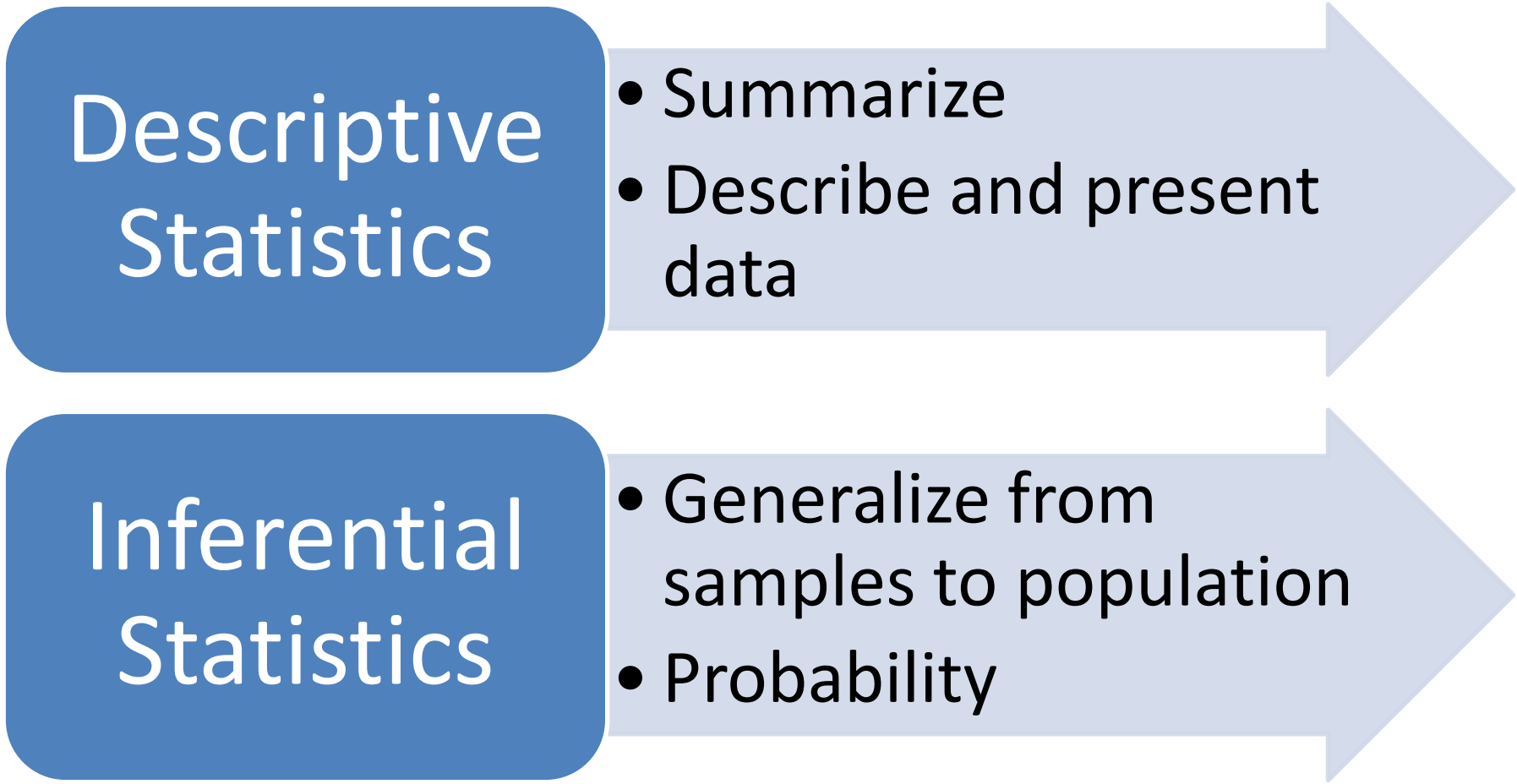
Target – Label

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

Features

Label

Descriptive Vs Inferential



Missing Value Treatment

- Missing data in the training data set can reduce the power / fit of a model.
- Can lead to a biased model because we have not analyzed the behavior .
- Reason for missing values
  - Data Extraction
  - Data Collection

Month	Y	X1	X2	X3	X4
Jan	1766	15	44	114	463
Feb	1634	11	50		400
Mar	1631	20	50	123	466
Apr	1990		45	128	426
May	1791	10	47	133	429
Jun	1491	13	50	132	407
Jul	1713	19		133	408
Aug	1708	18	47	111	489
Sep	1641	19	48	140	402
Oct	1721	13	44	111	
Nov	1436	16	49	149	408
Dec	1151	15	45	133	456

Missing Value Treatment

Methods to treat missing values

1. Deletion

2. Imputation

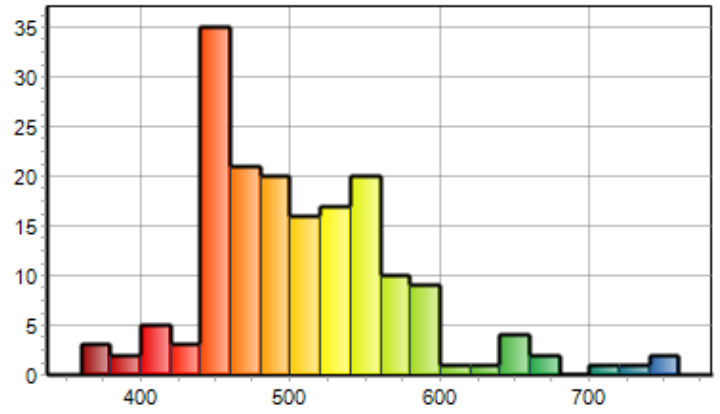
    - Generalized Imputation
    - Similar Case Imputation

Univariate Analysis

- Simplest form of analyzing data.
- It doesn’t deal with causes or relationships, it’s major purpose is to describe about the variable.

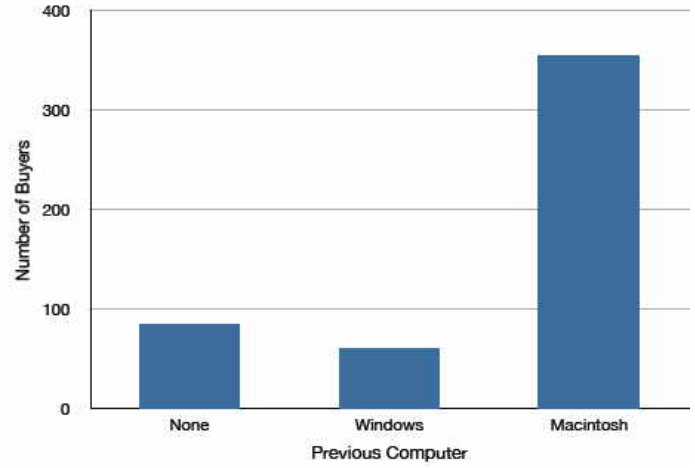
Quantitative

- Histograms



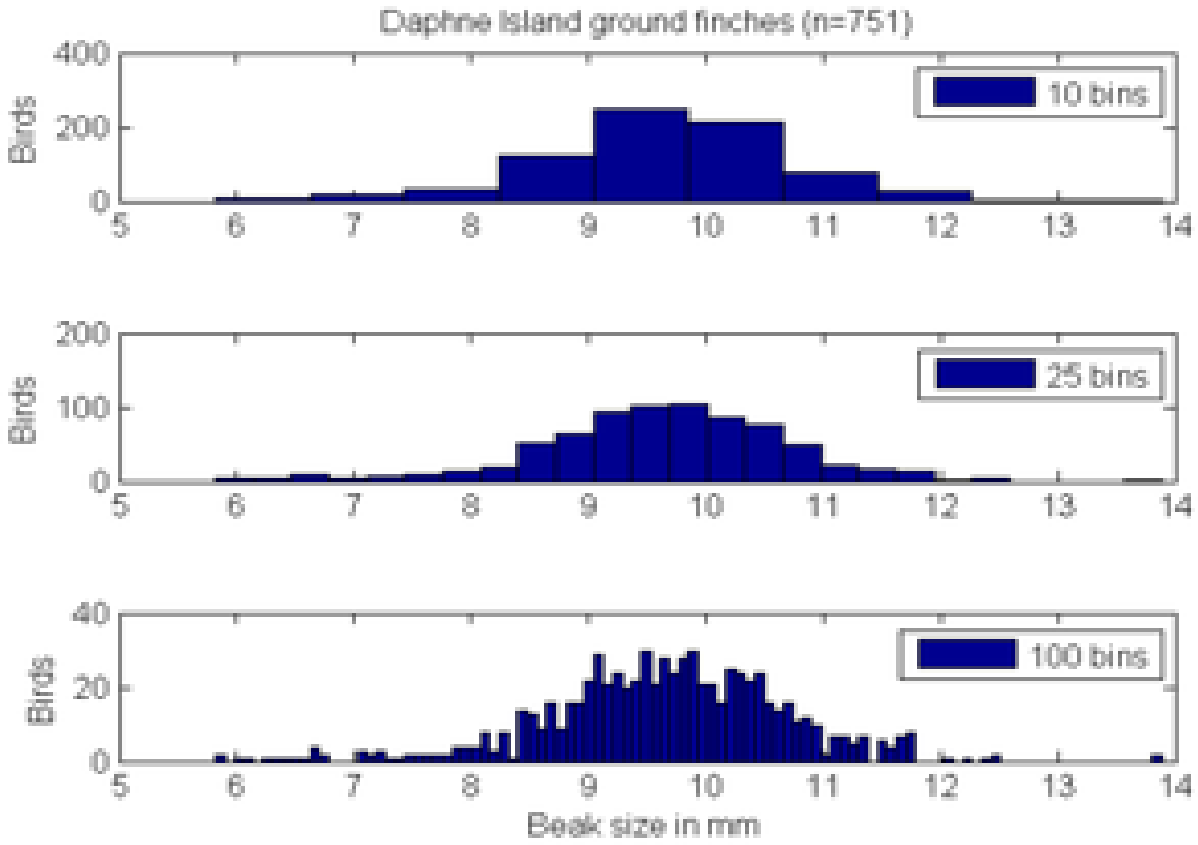
Qualitative

- Bar Chart



Binning

- Divides continuous variables into groups.
- Bins are easy to analyze and interpret.
- Information gets compressed into groups which later affects the final model.
- Hence, it is advisable to create small bins initially.
- Consider distribution of data prior to deciding bin size.

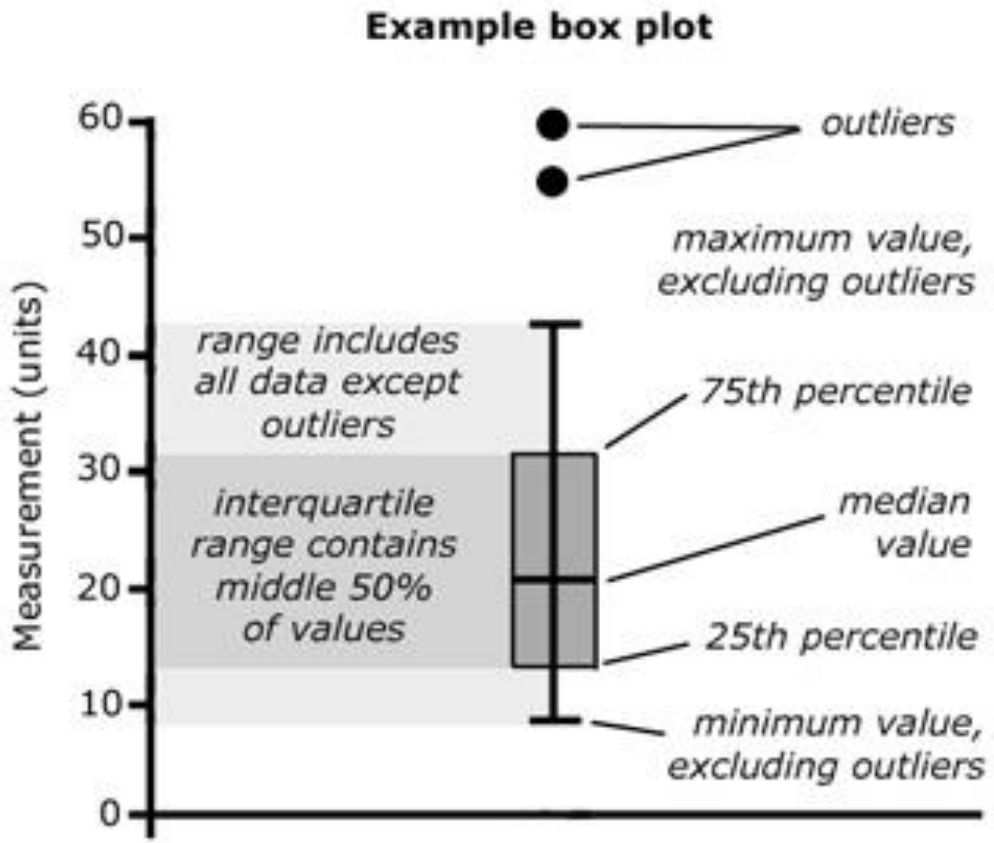


Bi-Variate Analysis

- Finds out the relationship between two variables.
- Bi-variate analysis between two continuous variables, we use scatter plot.

Box Plot

- The box-and-whisker plot is an exploratory graphic, created by John W. Tukey, used to show the distribution of a dataset.



Sampling

Sample is a set of data collected and/or selected from a statistical population by a defined procedure.

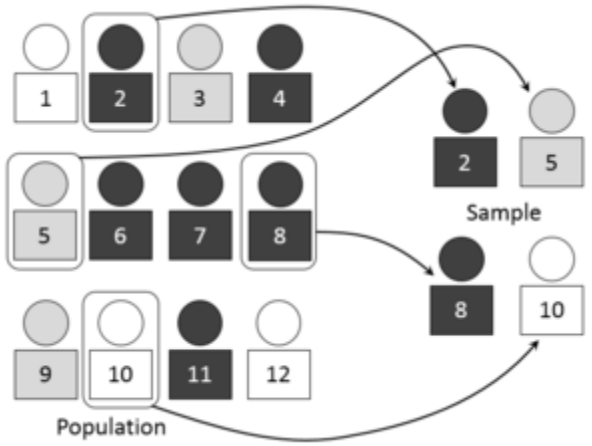
Types of Samples:

- Probability Sampling – (Representative Samples)

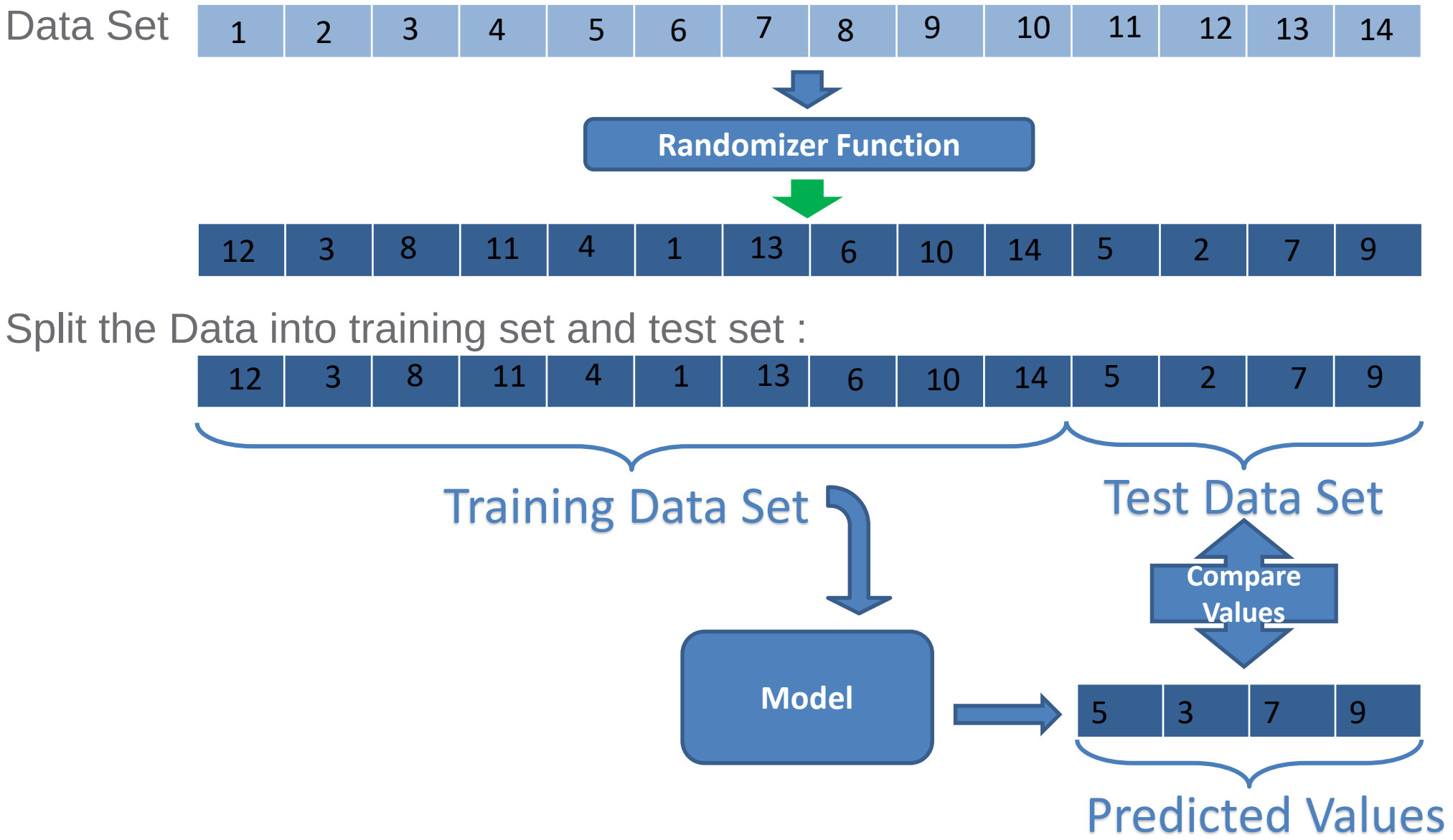
  The subjects of the population get an equal opportunity to be selected as a representative sample.

- Non Probability Sampling – (Non Representative Samples)

  It is not known that which individual from the population will be selected as a sample.

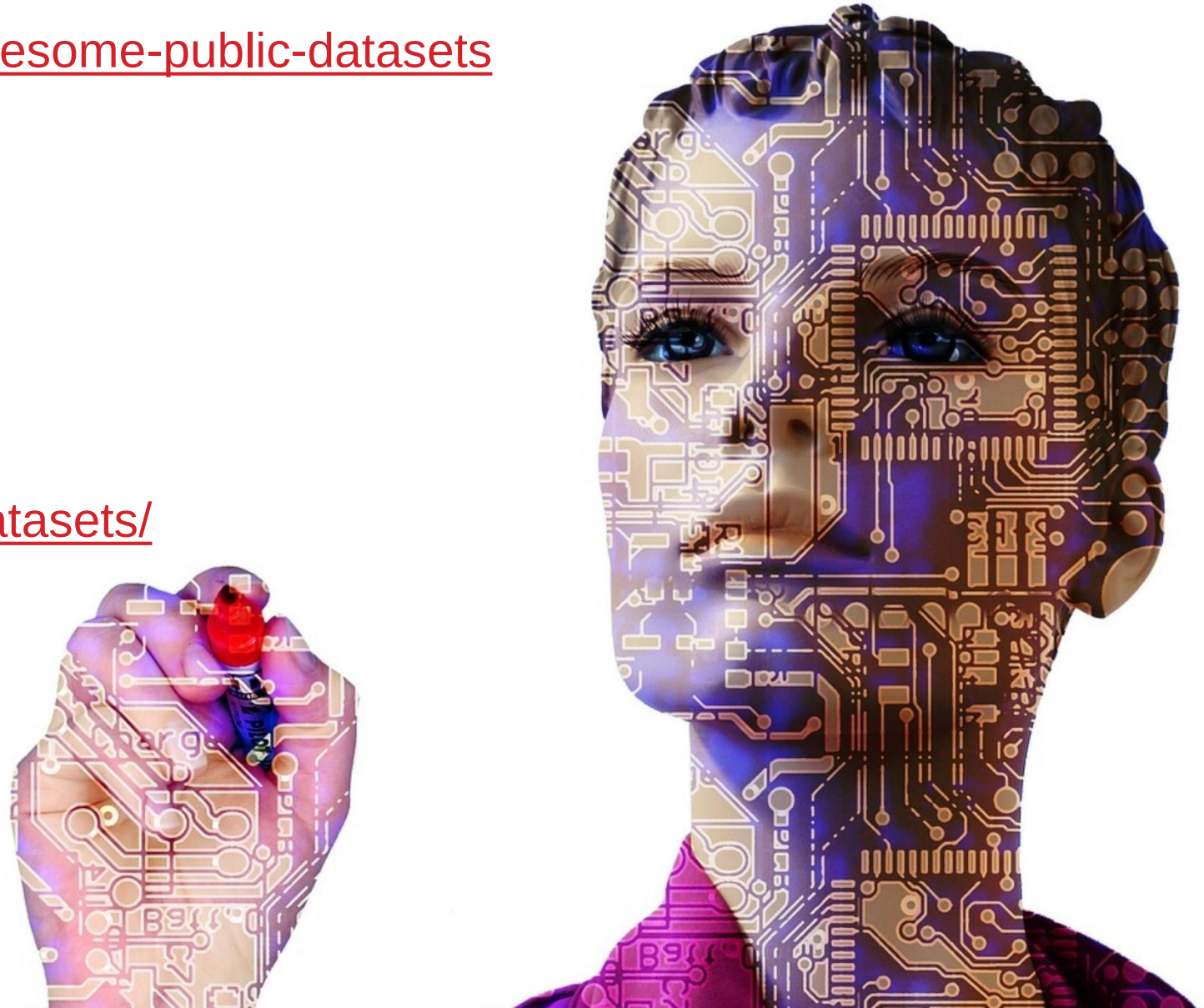


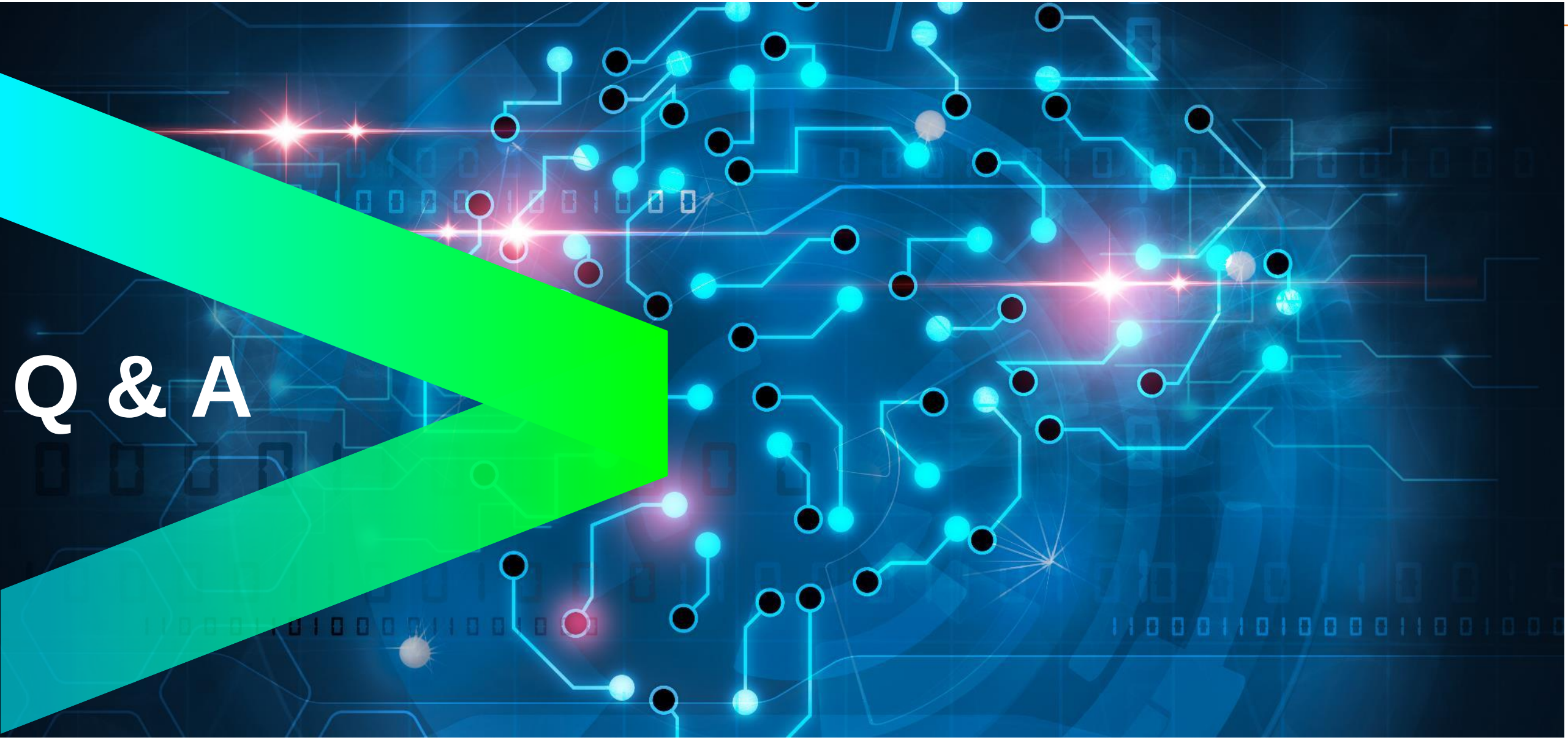
Train & Test Data Set



Public Datasets

1. https://github.com/caesar0301/awesome-public-datasets
2. https://data.gov.in/
3. https://www.data.gov/
4. https://www.kaggle.com/datasets
5. https://aws.amazon.com/public-datasets/




Q & A


Thank You

Standard Deviation Example

Sample Problem: Find the standard deviation for the following set of numbers: 6,2,3,1

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Step 1: Find the mean (6+2+3+1)/4 = 3

Step 2: Find the square of the distance from each data point to the mean

x	$\lvert x - \bar{x}\rvert^2$
6	$\lvert 6 - 3\rvert^2 = 3^2 = 9$
2	$\lvert 2 - 3\rvert^2 = 1^2 = 1$
3	$\lvert 3 - 3\rvert^2 = 0^2 = 0$
1	$\lvert 1 - 3\rvert^2 = 2^2 = 4$

Standard Deviation Example

Step 3: Sum the squares of the distances – (9+1+0+4) = 14

Step 4: Divide the Step 3 by the number of data points - 3.5

Step 5: Take the square root of Step 4 = ~1.84