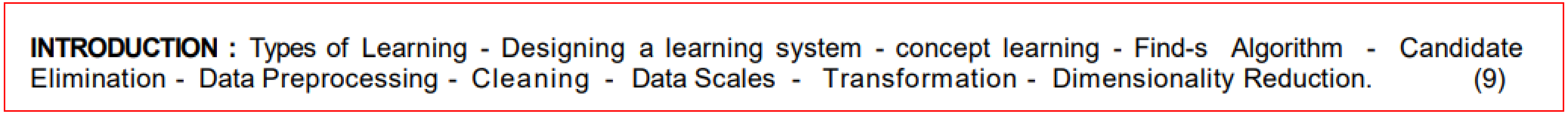
19Z601- MACHINE LEARNING

UNIT- 1 INTRODUCTION



INTRODUCTION : Types of Learning - Designing a learning system - concept learning - Find-s Algorithm - Candidate Elimination - Data Preprocessing - Cleaning - Data Scales - Transformation - Dimensionality Reduction. (9)

Presented by

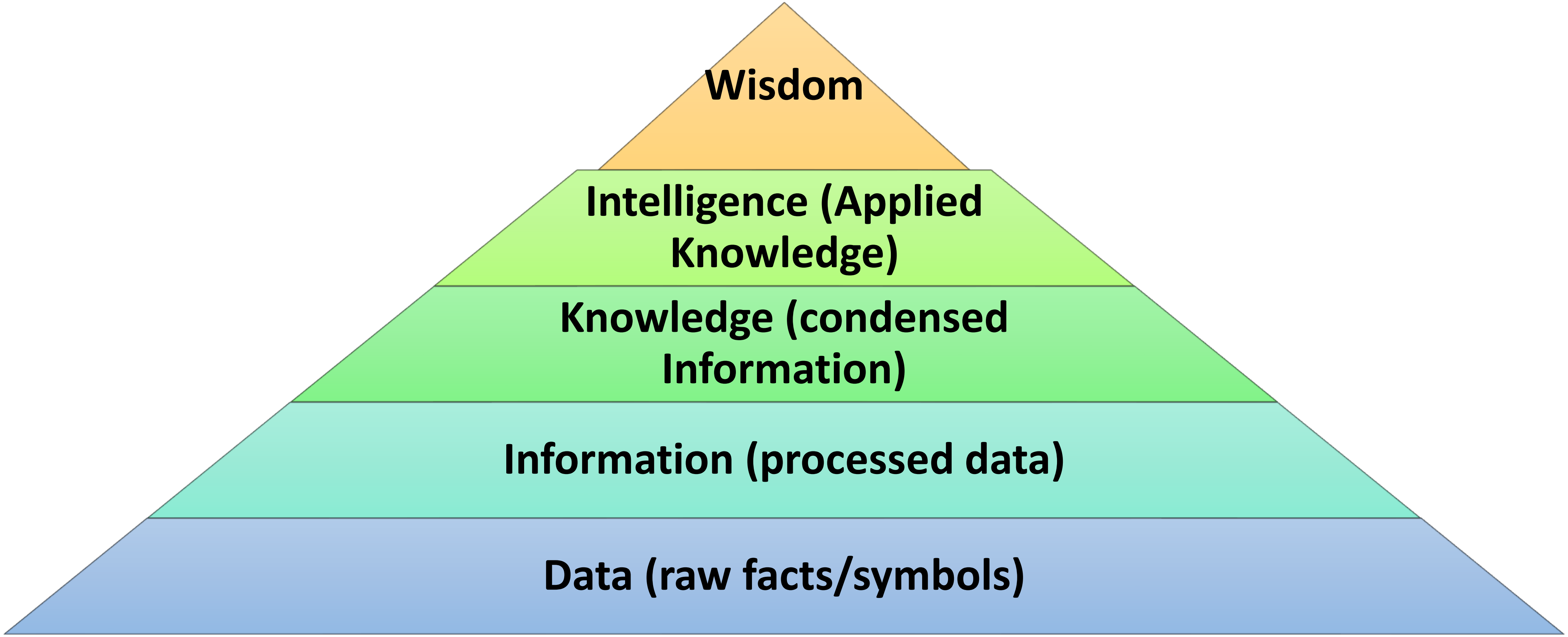
Ms.Anisha.C.D

Assistant Professor

CSE

Data

What is Data? Why it has to be processed?



Wisdom

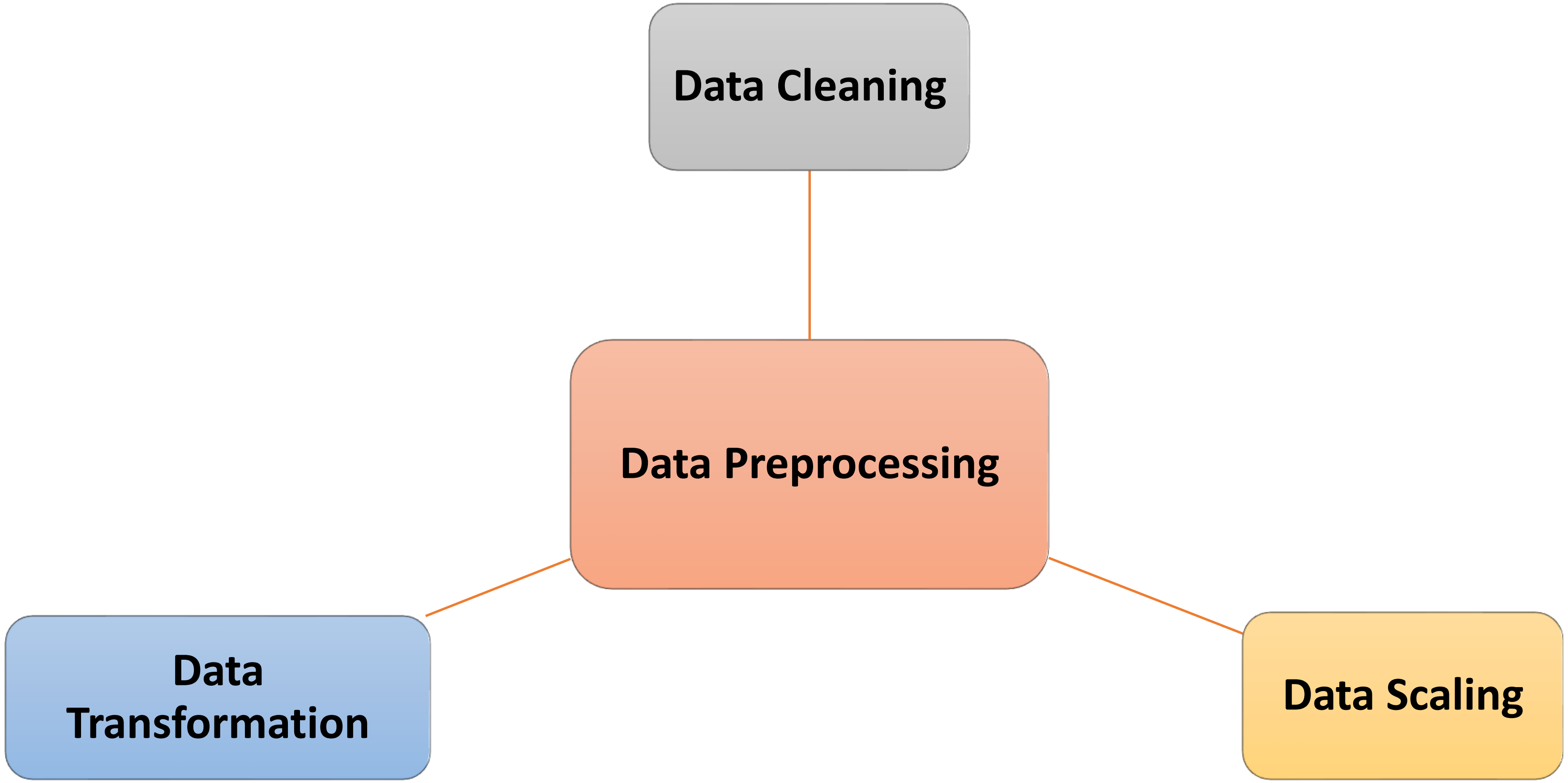
Intelligence (Applied Knowledge)

Knowledge (condensed Information)

Information (processed data)

Data (raw facts/symbols)

Data Preprocessing



Data Cleaning

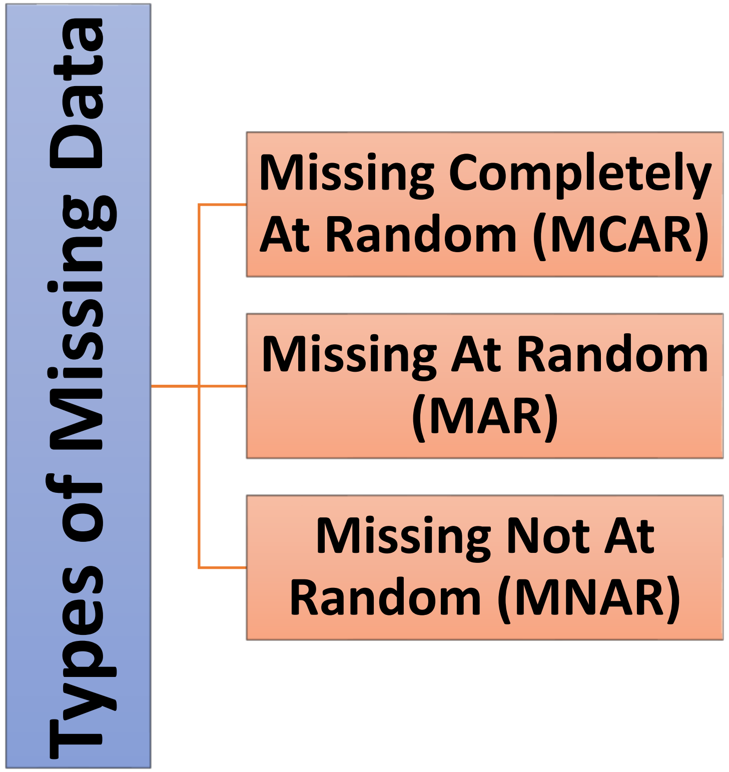
Data Preprocessing

Data Transformation

Data Scaling

Data Cleaning – Handling Missing Data

- **Removing** the **examples with missing features** from the dataset.
- Using a **learning algorithm** that can deal with missing feature values.
- Using a **data imputation** technique.



Types of Missing Data

Missing Completely At Random (MCAR)

Missing At Random (MAR)

Missing Not At Random (MNAR)

Types of Missing Data

➢Missing Completely At Random (MCAR)

- In this type, the **probability of data being missing** is unrelated to both **observed and unobserved data.**
- In other words, **missingness is purely random and occurs by chance.**
- MCAR implies that the **missing data is not systematically related to any variables in the dataset.**
- For example, a sensor failure that results in sporadic missing temperature readings can be considered MCAR.

Types of Missing Data

➢Missing At Random (MAR)

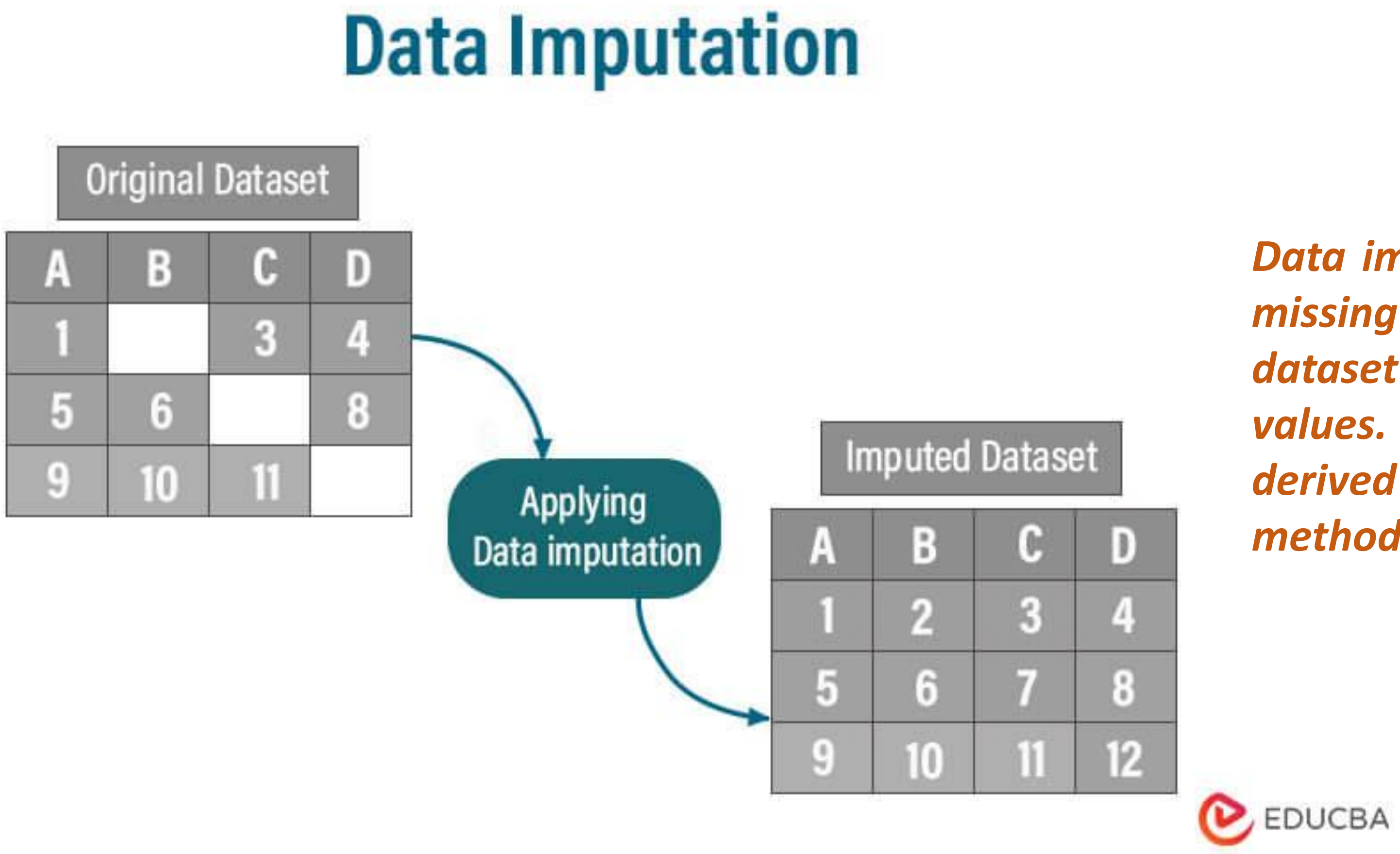
- Missing data is considered MAR **when the probability of data being missing is related to observed data but not directly to unobserved data.**
- In other words, **missingness is dependent on some observed variables.**
- For instance, in a medical study, men might be less likely to report certain health conditions than women, creating missing data related to the gender variable. MAR is a more general and common type of missing data than MCAR.

Types of Missing Data

➢ Missing Not at Random (MNAR)

- MNAR occurs when the **probability of data** being missing is related to **unobserved data or the missing values themselves.**
- This type of missing data can introduce bias into analyses because the missingness is related to the missing values.
- An example of MNAR could be patients with severe symptoms avoiding follow-up appointments, resulting in missing data related to the severity of their condition.

Data Cleaning – Handling Missing Values – Data Imputation Technique



Original Dataset

A	B	C	D
1		3	4
5	6		8
9	10	11	

Imputed Dataset

A	B	C	D
1	2	3	4
5	6	7	8
9	10	11	12

Data imputation is the process of replacing missing or incomplete data points in a dataset with estimated or substituted values. These estimated values are typically derived from the available data, statistical methods, or machine learning algorithms.

Data Cleaning – Handling Missing Values – Data Imputation Technique

- **Mean Imputation:** Replace missing values in numerical variables with the average of the observed values for that variable.
- **Median Imputation:** Replace missing values in numerical variables with the middle value of the observed values for that variable.
- **Mode Imputation:** Replace missing values in categorical variables with the most frequent category among the observed values for that variable.

Example

Bumrah – Cricket Player – Number of Wickets Taken by Last seven games

2	3	1	4	5		2

Mean = sum of all data points / Number of data points
Mean = 17/7 = 2.4 = 2 (Need Discrete value)

2	3	1	4	5	2	2

Median = n is odd , $((n+1)/2)^{th}$ observation

n is even, $$\frac{(n/2)^{th} + ((n/2) +1))^{th}}{2}$$

Median = 4

2	3	1	4	5	4	2

Mode = The data point that appears the most.

2	3	1	4	5	2	2

Applicability of Data Imputation Technique

- Use **mean imputation** for numerical variables when missing data is missing completely at random (MCAR) and the variable has a relatively normal distribution.
- Use **median imputation** when the data is skewed or contains outliers, as it is less sensitive to extreme values.
- Use **mode imputation** for categorical variables when you have missing values that can be reasonably replaced with the most frequent category.

Data Transformation - Binning

- Consider the following set : S = {12,14,19,22,24,26,28,31,32}

- By **equal-frequency bin method**, the data should be distributed across bins. Assume the bins of size 3, then the above data is distributed across the bins as follows:

Bin 1 = 12, 14,19

Bin 2 = 22,24,26

Bin 3 = 28, 31, 32

Data Transformation - Binning

- Consider the following set : S = {12,14,19,22,24,26,28,31,32}

- By **smoothing bins method**, the bins are replaced by the mean of the bin.

Bin 1 = 15,15,15
Bin 2 = 24,24,24
Bin 3 = 30.3,30.3,30.3

Data Transformation - Binning

- Consider the following set : S = {12,14,19,22,24,26,28,31,32}

- By **smoothing by bin boundaries method**, the bins values are replaced by:

Bin 1 = 12,12,19

Bin 2 = 22,22,26

Bin 3 = 28,32,32