upGrad

PLACEMENT REFRESHER PROGRAM

Session 7 - ML 1

Basics of ML

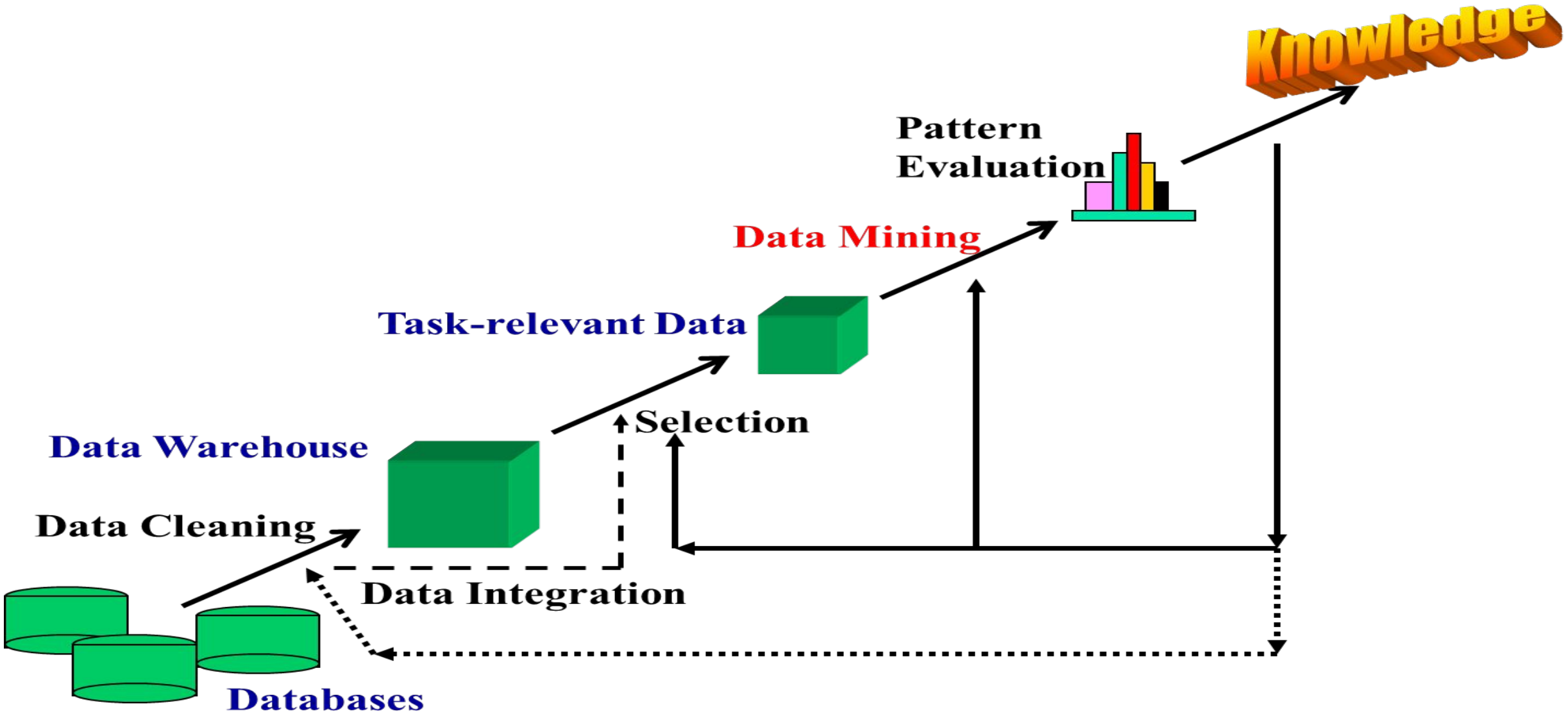
By
Ritesh Kumar Pandey

Agenda

- KDD Process
- Basic Statistical Description of Data
- Normalization of Data
- Hypothesis Testing

Knowledge Discovery from Data (KDD)

upGrad



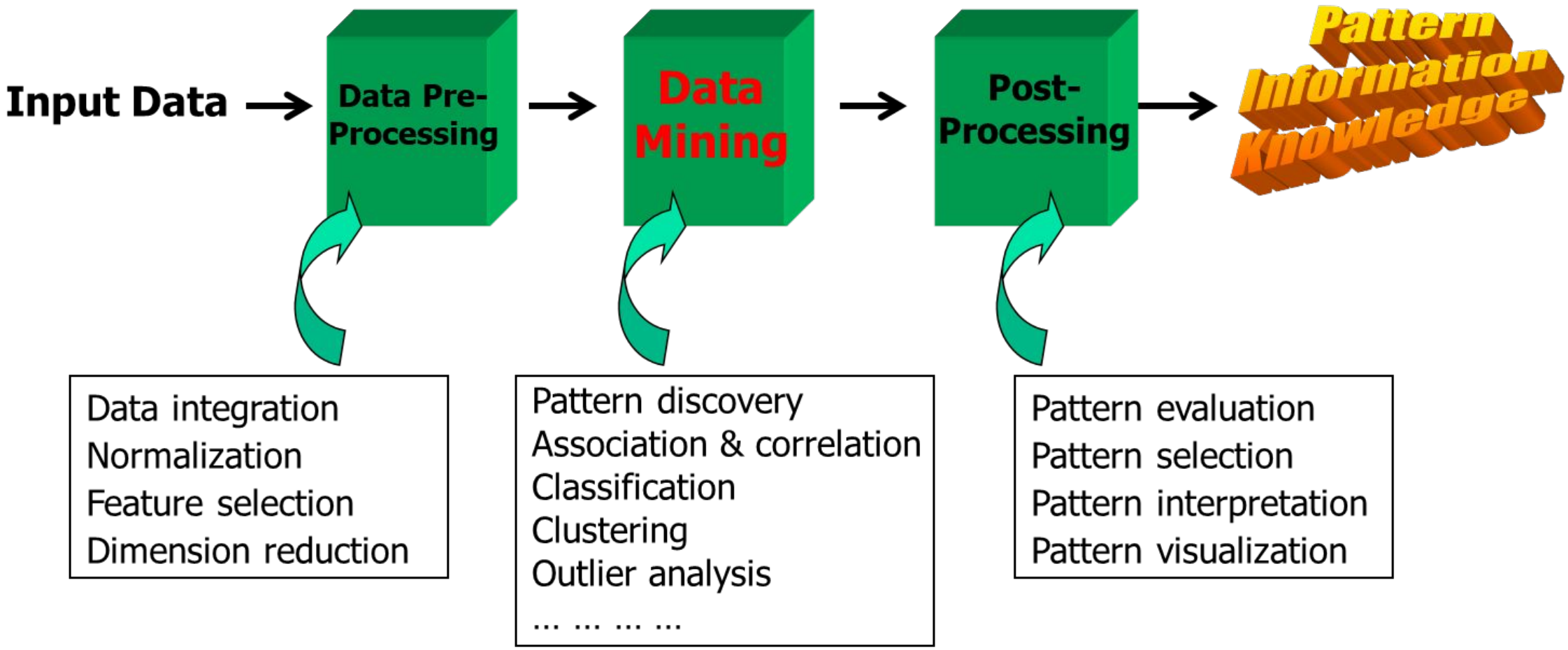
Knowledge Discovery from Data (KDD)

upGrad

- **Data cleaning:** To remove noise and inconsistent data
- **Data integration:** To combine data from multiple data sources
- **Data selection:** Data relevant to the analysis task are retrieved from the database
- **Data transformation:** Data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations
- **Data mining:** An essential process where intelligent methods are applied to extract data patterns
- **Pattern evaluation:** To identify the truly interesting patterns representing knowledge based on interestingness measures
- **Knowledge presentation:** Visualization and knowledge representation techniques are used to present mined knowledge to users

Knowledge Discovery from Data (KDD)

upGrad



- This is a view from typical machine learning and statistics communities

Statistical Description of Data

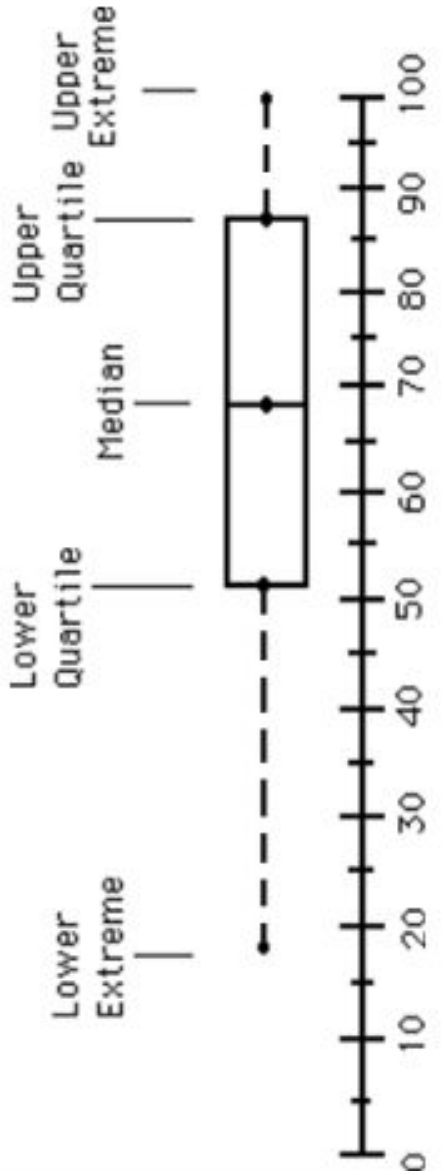
upGrad

- **Quartiles**: The k^{th} percentile of a set of data in numerical order is the value x_i having the property that k percent of the data entries lie at or below x_i
- Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - **Example of age values:** 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
  - Q1 = 20, Q3 = 35
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
  - IQR = 15
- **Outlier**: usually, a value higher/lower than 1.5 x IQR
  - Falling 1.5 x IQR above Q3 = 70
- **Five-number summary** of a distribution: Min, Q1, Median, Q3, Max
  - 13, 20, 25, 35, 70.

Statistical Description of Data

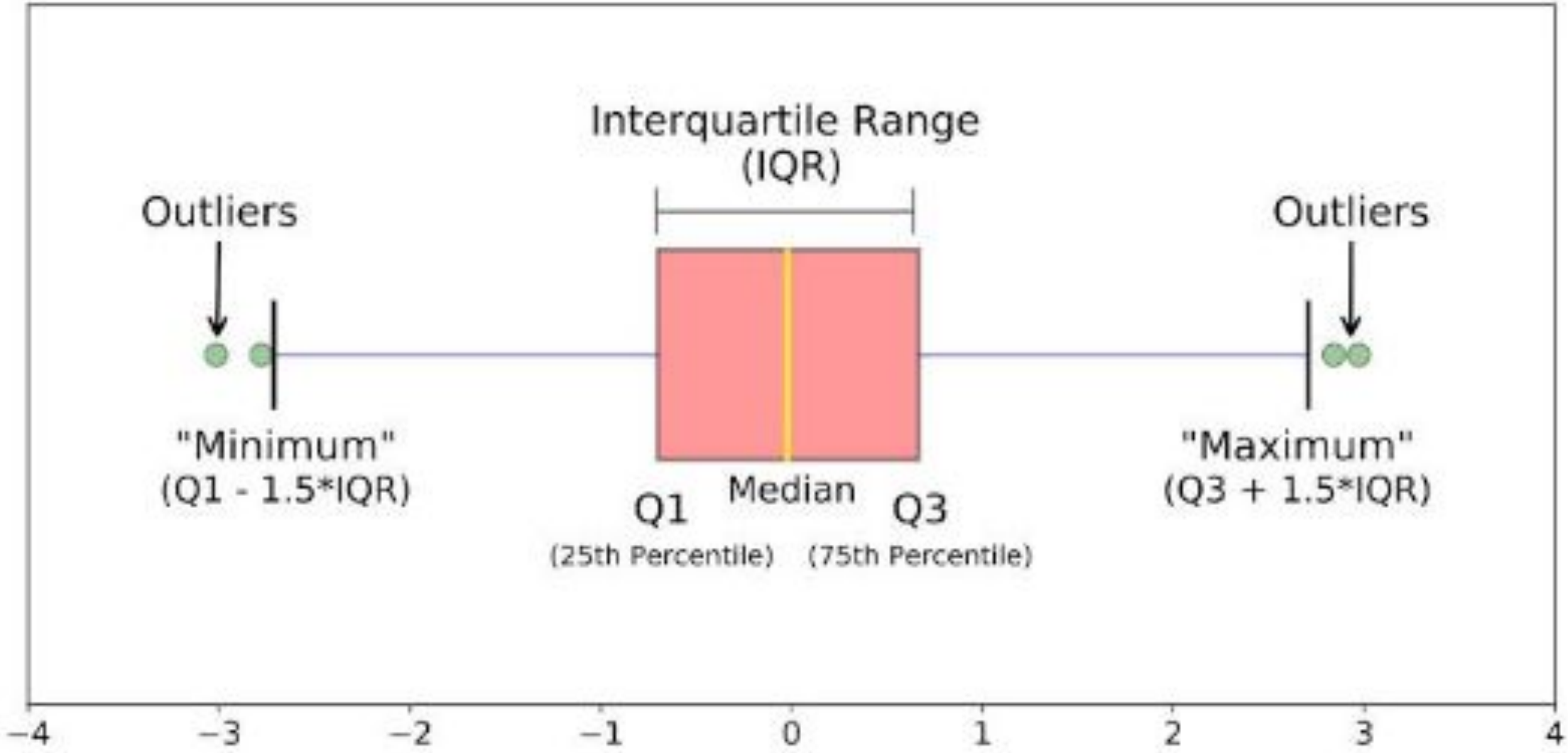
upGrad

- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



Statistical Description of Data

upGrad



Statistical Description of Data - Question

upGrad

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) **5, 19, 25, 29, 30, 30, 31, 32, 32, 36, 36, 36, 36, 43, 43, 45, 45, 45, 45, 46, 50, 57, 85.**

1. What is the mean of the data? What is the median?
2. What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
3. What is the midrange of the data?
4. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
5. Give the five-number summary of the data.
6. Show a boxplot of the data

Statistical Description of Data - Question

upGrad

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) **5, 19, 25, 29, 30, 30, 31, 32, 32, 36, 36, 36, 36, 43, 43, 45, 45, 45, 45, 46, 50, 57, 85.**

1. What is the mean of the data? What is the median?
   Mean = 38, Median = 36
2. What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
   Bimodal (36 and 45 have highest frequency)
3. What is the midrange of the data?
   45
4. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
   Q1 = 30, Q3 = 45, IQR = 15
5. Give the five-number summary of the data.
   5, 30, 36, 45, 80

Normalization of Data

upGrad

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Normalization of Data - Question

upGrad

Use these methods to normalize the following group of data:
200,300,400,600,1000

1. min-max normalization by setting min = 0 and max = 1
2. z-score normalization
3. normalization by decimal scaling

Normalization of Data - Question

upGrad

Use these methods to normalize the following group of data:
200,300,400,600,1000

1. min-max normalization by setting min = 0 and max = 1
2. z-score normalization
3. normalization by decimal scaling

Data	Min-Max Normalized Value	Z-Score Normalized Value	Decimal Scaling Normalized Values
200	0	-1.06	0.2
300	0.125	-0.7	0.3
400	0.25	-0.35	0.4
600	0.5	0.35	0.6
1000	1	1.75	1.0

Hypothesis Testing

upGrad

1. A hypothesis is a starting point that can be confirmed or rejected through evidence

2. The initial belief is called a null hypothesis (H0)
    - Generally the status quo
    - Action taken: Do nothing

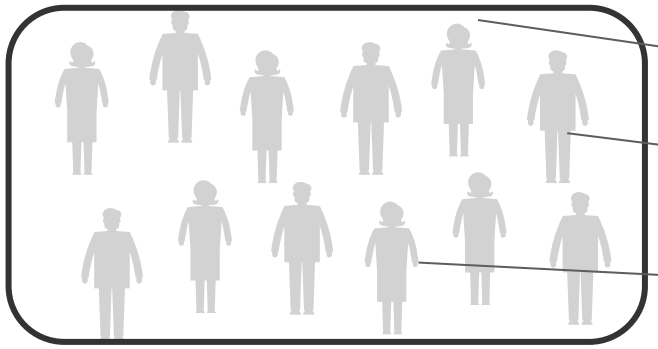
3. Its negation is called the alternate hypothesis (HA, HB, H1)
    - Often a claim to be tested or a change to be detected
    - Action taken: Do something

Hypothesis Testing - Process

upGrad

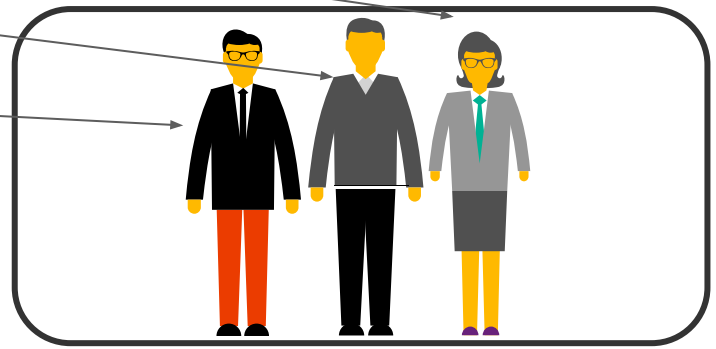
STEP 1: Start with a hypothesis about a population parameter

The parameter could be mean, proportion, median, etc.



STEP 2: Collect sample information

Collect information from a randomly chosen sample and calculate the appropriate sample statistic



STEP 3: Reject/do not reject null hypothesis

Is the sample information strongly consistent with the null hypothesis? If yes, then reject the hypothesis

Hypothesis Testing - Framing a Hypothesis

upGrad

A mobile company claims that their new dash charger will charge the phone to 60% in 30 minutes

If the time taken is less than 30 minutes

Revise the claim to a superior figure

If the time taken is more than 30 minutes

Ask engineers to fix the issue

ACTION: TO BE TAKEN

Hypothesis Testing - Framing a Hypothesis

upGrad

- **Null hypothesis (H_0):** Time taken to charge the phone to 60% = 30 minutes

- **Alternate hypothesis (H_A):** Time taken to charge the phone to 60% ≠ 30 minutes

Hypothesis Testing - Example

upGrad

Status Quo: Person is innocent (Null Hypothesis)
Challenge to the status Quo: Person is not innocent, i.e. guilty (Alternate Hypothesis))

Trial (To test the Hypothesis)

Not enough evidence to prove that the person has committed the crime

Enough evidence to prove that the person has committed the crime

Do nothing (i.e. let the suspect go)

Do something (In this case, give appropriate punishment)

THANK YOU