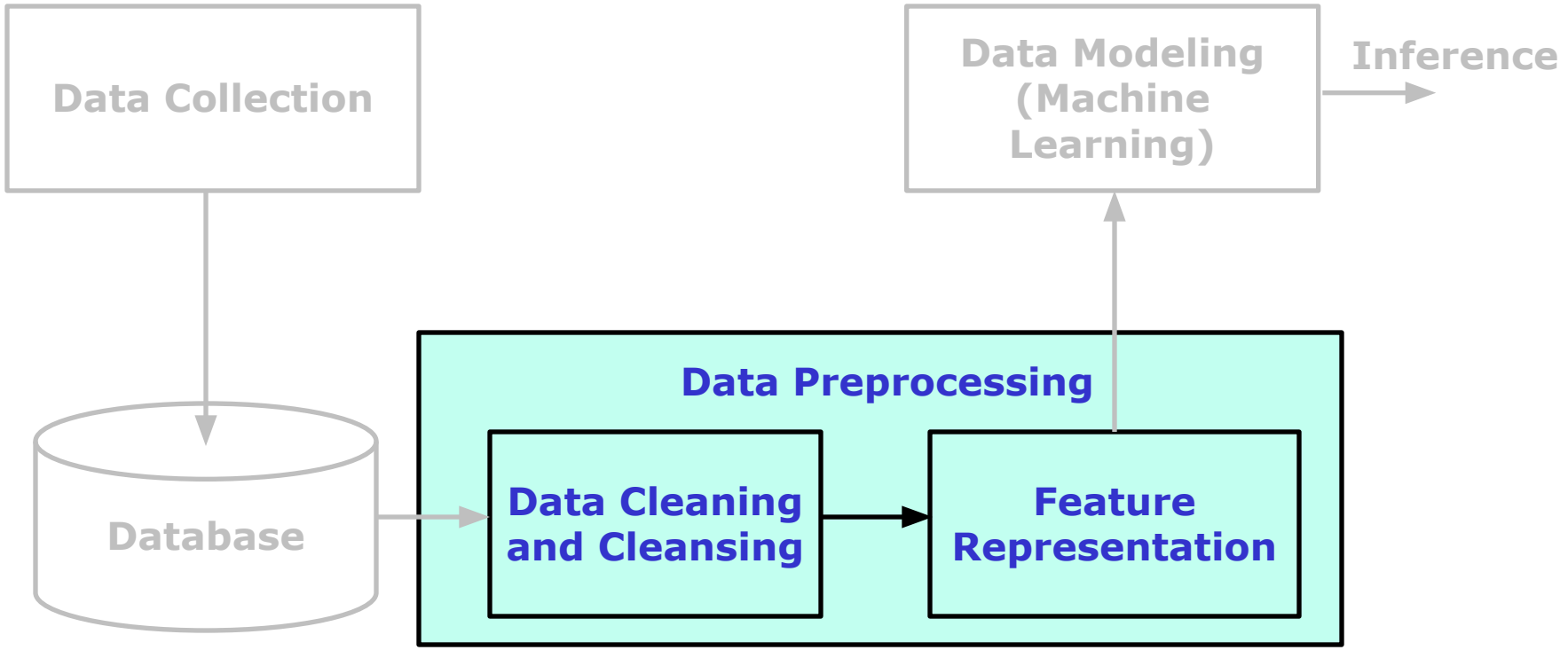
Data Preprocessing

Data Science

- Multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from structured and unstructured data
- Central concept is gaining insight from data
- Machine learning uses data to extract knowledge



Need for Data Preprocessing

- Real world data are tend to be incomplete, noisy and inconsistent due to their huge size and their likely origin from multiple heterogeneous sources
- Preprocessing is important to clean the data
- Low quality data will lead to low quality of analysis results
- If the users believe the data is of low quality (dirty), they are unlikely to trust the results of any data analytics that has been applied to
- Low quality data can cause confusion for analytic procedure using machine learning techniques, resulting in unreliable output
- Data could be
  - Incomplete,
  - noisy and
  - inconsistent
  - These are common properties of large real world databases

Data Preprocessing Techniques

- Data cleaning:
- Data integration:
- Data transformation:
- Data reduction :

Data Preprocessing Techniques

- Data cleaning:
  - Applied to
    - identify the missing values,
    - fill in missing values,
    - remove noise and
    - correct inconsistency in the data
- Data integration:
  - It merges data from multiple sources in to a coherent data source
- Data transformation:
  - Transforming the entries of data to a common format
  - Techniques like normalization and standardization applied to transform the data to another form to improve the accuracy and efficiency of machine learning (ML) algorithms involving distance measures

Data Preprocessing Techniques

- Data reduction:
  - Applied to obtain a reduced representation that is much smaller in volume, yet producing almost same analytical results
  - It can reduce the data size by
    - Aggregation
    - Eliminating irrelevant and redundant features (attributes) through correlation analysis
    - Reducing dimension
- *These techniques are not mutually exclusive; they may work together*

Descriptive Data Summarization (Descriptive Analytics)

- It serves as a foundation for data preprocessing
- It helps us to study the general characteristics of data and identify the presence of noise or outliers
- Data characteristics:
  - Central tendency of data
    - Centre of the data
    - Measuring mean, median and mode
  - Dispersion of data
    - The degree to which numerical data tend to spread
    - Measuring range, quartiles, interquartile range (IQR), the five-number summary and standard deviation

Descriptive Analytics: Measuring Central Tendency

- Mean:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. Mean of this set of values is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Number of records (tuples), N = 10

Years of experience	Salary (in Rs 1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83
Sum:	**91**

Descriptive Analytics: Measuring Central Tendency

- Mean:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. Mean of this set of values is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Number of records (tuples), N = 10

Years of experience	Salary (in Rs 1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Mean Years of experience: Sum/10 **9.1**

Descriptive Analytics: Measuring Central Tendency

- Mean:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. Mean of this set of values is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Number of records (tuples), N = 10

Years of experience	Salary (in Rs 1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Mean Salary: Sum/10

55.4

Descriptive Analytics: Measuring Central Tendency

- Mean:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. Mean of this set of values is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - Mean is a better measure of central tendency for the symmetric data (symmetrically distributed data)

Number of records (tuples), N = 10

Years of experience	Salary (in Rs 1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83
Mean:	**9.1**	**55.4**

Descriptive Analytics: Measuring Central Tendency

- Median:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. The median is the "middle" number (value), when those numbers are listed in order from smallest to greatest.
  - Median is the value separating the higher half from the lower half of a data sample
  - For a given data of N values in sorted order
    - If N is odd, then median is the middle value of the ordered list
    - If N is even, then median is the average of middle two values

Number of records (tuples), N = 10

Years of experience	Salary (in Rs 1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Illustration: Median of attribute "Years of experience"

Descriptive Analytics: Measuring Central Tendency

- Median:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. The median is the "middle" number (value), when those numbers are listed in order from smallest to greatest.
  - Median is the value separating the higher half from the lower half of a data sample
  - For a given data of N values in sorted order
    - If N is odd, then median is the middle value of the ordered list
    - If N is even, then median is the average of middle two values

Sort the values in "Years of experience"

Years of experience
1
3
6
8
9
11
13
16
16
21

Descriptive Analytics: Measuring Central Tendency

- Median:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. The median is the "middle" number (value), when those numbers are listed in order from smallest to greatest.
  - Median is the value separating the higher half from the lower half of a data sample
  - For a given data of N values in sorted order
    - If N is odd, then median is the middle value of the ordered list
    - If N is even, then median is the average of middle two values

Sort the values in "Years of experience"

Years of experience
1
3
6
8
9
11
13
16
16
21

Median: $\frac{9+11}{2}$

Descriptive Analytics: Measuring Central Tendency

- Median:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. The median is the "middle" number (value), when those numbers are listed in order from smallest to greatest.
  - Median is the value separating the higher half from the lower half of a data sample
  - For a given data of N values in sorted order
    - If N is odd, then median is the middle value of the ordered list
    - If N is even, then median is the average of middle two values
  - For asymmetrically distributed (skewed) data, a better measure of centre of data is median

Sort the values in "Years of experience"

Years of experience
1
3
6
8
9
11
13
16
16
21

Median: **10**

Descriptive Analytics: Measuring Central Tendency

- Mode: Most frequent value in an attribute in the data

Illustration: Mode of attribute "Years of experience"

Assume that values are discrete numerical

Number of records (tuples), N = 10

Years of experience	Salary (in Rs 1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Mode: **3**

Descriptive Analytics: Measuring Central Tendency

- Mode: Most frequent value in an attribute in the data

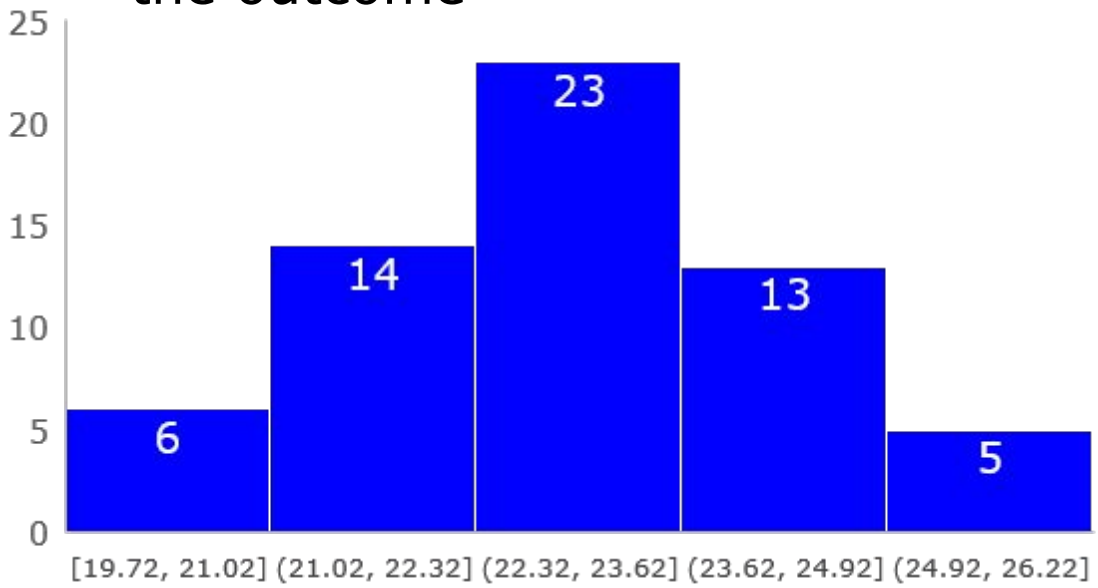
Number of samples, N = 61

Date	Temperature
Sept 1	25.47
Sept 2	26.19
Sept 3	25.17
Sept 4	24.30
Sept 5	24.07
Sept 6	21.21
Sept 7	23.49
Sept 8	21.79
Sept 9	25.09
Sept 10	25.39
---	---
Oct 29	23.06
Oct 30	23.72
Oct 31	23.02

Mean: **22.85**

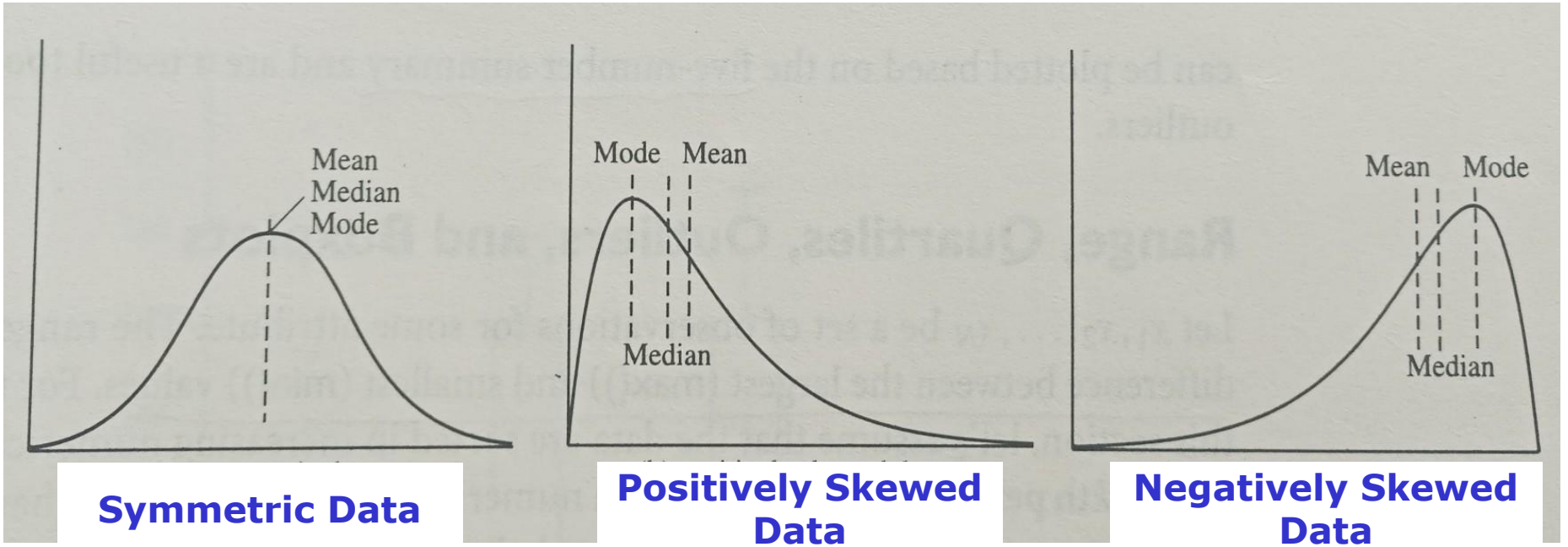
Median: **22.89**

- The mode of a continuous variable is the value at which the probability density function, f(x) , is at a maximum.
- It is a value that is most likely to lie within the same interval as the outcome



Mode: (22.32 – 23.62] ~=22.97

Descriptive Analytics: Measuring Central Tendency



Symmetric Data

Positively Skewed Data

Negatively Skewed Data

Descriptive Analytics: Measuring Dispersion of Data

- The degree to which numerical data tend to spread
- It is also called as variance (in symmetrically distributed data)
- Common measures of data dispersion:
  - Range
  - The five-number summery (based on quartiles)
  - The inter quartile range (IQR)
  - Standard deviation
- Range: The range of a finite set of values is the difference between the maximum and minimum values

Descriptive Analytics: Measuring Dispersion of Data

- Quartiles:
  - The k^{th} percentile:
    - Let x_1, x_2, …, x_N be a set of N values in an attribute
    - The k^{th} percentile of a set of data in numerical order is the value of x_n having the property that k percent of data entries lie at or below x_n
  - Example: 50^{th} percentile
    - The value (number) below which 50% of the data entries (values) lie
      - Those 50% of entries have values equal to or less that 50^{th} percentile

Number of records (tuples), N = 10

Years of experience	Salary (in Rs 1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Illustration: 50th percentile of attribute "Years of experience"

Descriptive Analytics: Measuring Dispersion of Data

- Quartiles:
  - The k^{th} percentile:
    - Let $x_1, x_2, \ldots, x_N$ be a set of N values in an attribute
    - The k^{th} percentile of a set of data in numerical order is the value of x_n having the property that k percent of data entries lie at or below x_n
  - Example: 50^{th} percentile
    - The value (number) below which 50% of the data entries (values) lie
      - Those 50% of entries have values equal to or less that 50^{th} percentile

Sort the values in "Years of experience"

Years of experience
1
3
6
8
9
11
13
16
16
21

50^{th} Percentile: 10

Illustration: 50^{th} percentile of attribute "Years of experience"

Descriptive Analytics: Measuring Dispersion of Data

- Quartiles:
  - The k^{th} percentile:
    - Let x_1, x_2, …, x_N be a set of N values in an attribute
    - The k^{th} percentile of a set of data in numerical order is the value of x_n having the property that k percent of data entries lie at or below x_n
  - Example: 25th percentile
    - The value (number) below which 25% of the data entries (values) lie
      - Those 25% of entries have values equal to or less that 25th percentile
    - Middle element between minimum and 50th percentile

Sort the values in "Years of experience"

Years of experience
1
3
6
8
9
11
13
16
16
21

25th Percentile: **6**

Illustration: 25th percentile of attribute "Years of experience"

Descriptive Analytics: Measuring Dispersion of Data

- Quartiles:
  - The k^{th} percentile:
    - Let x_1, x_2, …, x_N be a set of N values in an attribute
    - The k^{th} percentile of a set of data in numerical order is the value of x_n having the property that k percent of data entries lie at or below x_n
  - Example: 75^{th} percentile
    - The value (number) below which 75% of the data entries (values) lie
      - Those 75% of entries have values equal to or less that 75^{th} percentile
    - Middle element between maximum and 50^{th} percentile

Sort the values in "Years of experience"

Years of experience
1
3
6
8
9
11
13
16
16
21

75^{th} Percentile: 16

Illustration: 75^{th} percentile of attribute "Years of experience"

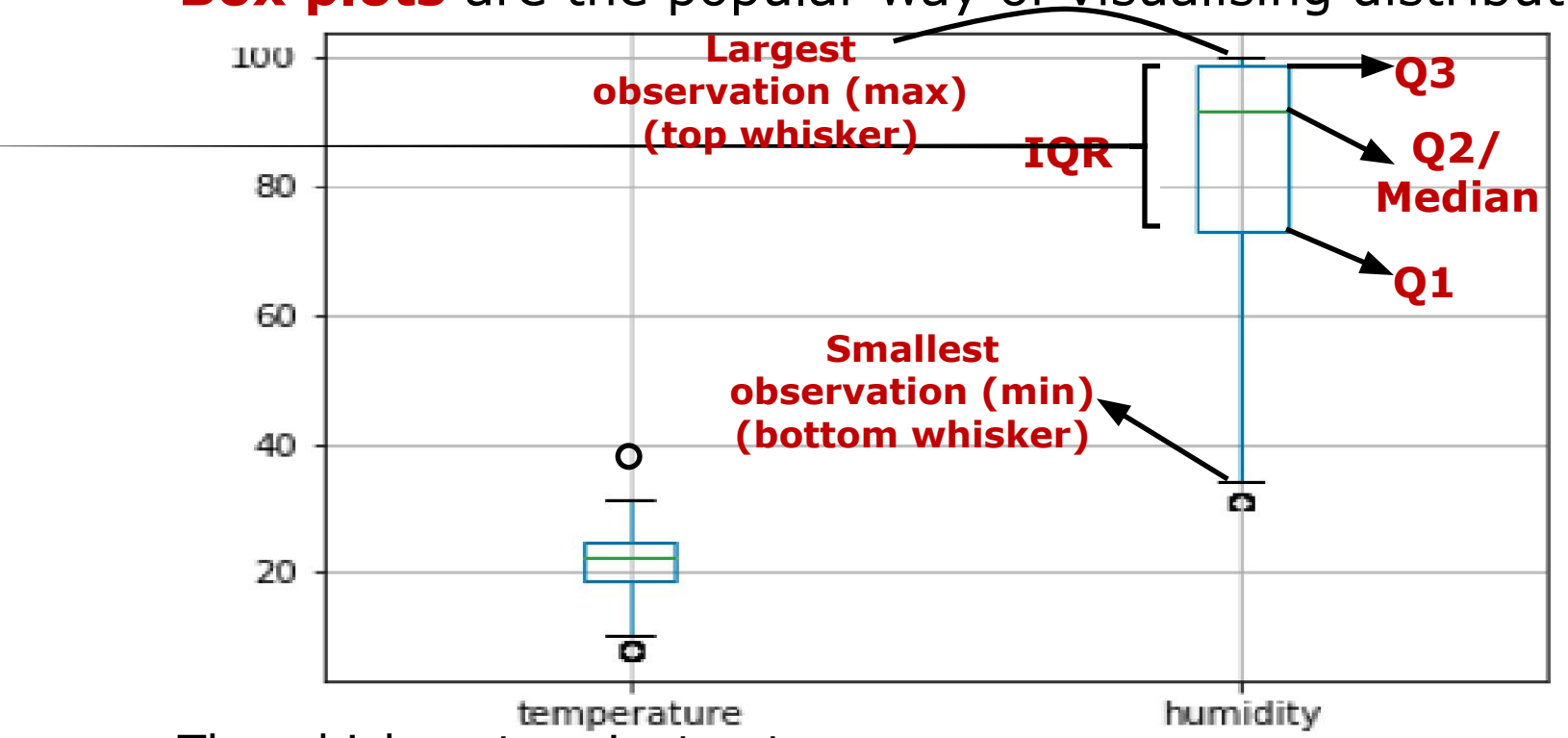
Descriptive Analytics: Measuring Dispersion of Data

- Quartiles:
  - The k^{th} percentile:
    - Let $x_1, x_2, \ldots, x_N$ be a set of N values in an attribute
    - The k^{th} percentile of a set of data in numerical order is the value of x_n having the property that k percent of data entries lie at or below x_n
    - Median is the 50^{th} percentile (the second quartile (Q2))
    - The first quartile (Q1): It is the 25^{th} percentile
    - The third quartile (Q3): It is the 75^{th} percentile
  - The quartiles including median give some indication of centre, spread and shape of distribution
- The distance between the Q1 and Q3 is a simple measure of spread
- Interquartile range (IQR): Distance between the first quartile (Q1) and third quartile (Q2)

$$IQR = Q3 - Q1$$

Descriptive Analytics: Measuring Dispersion of Data

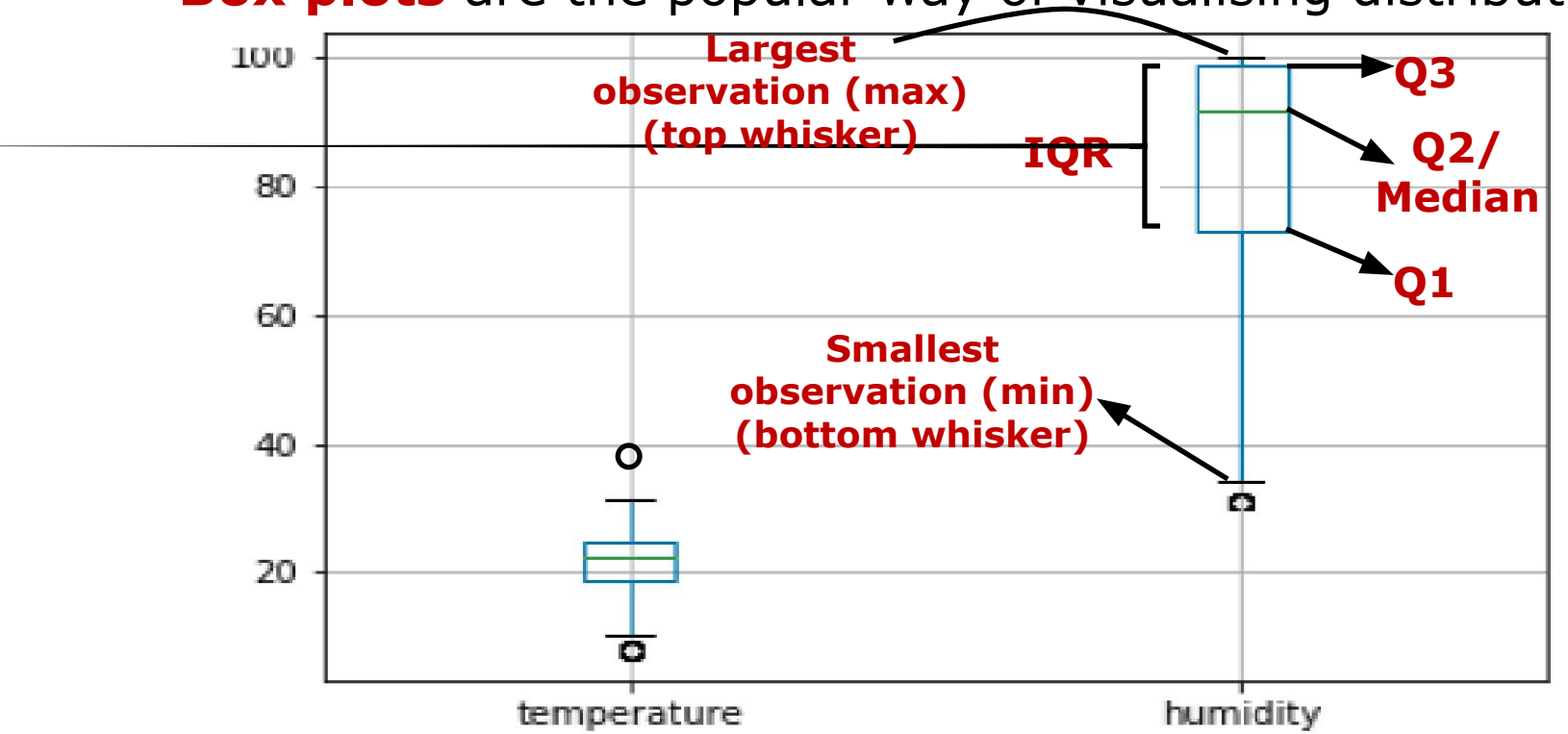
- The five-number summary of distribution:
  - It consists of minimum value, Q1, median, Q3 and maximum value
- **Box plots** are the popular way of visualising distribution



- The whiskers terminate at
  - Smallest (minimum) or largest (maximum) observations **or**
  - the most extreme observations occurring within 1.5 x IQR of respective quartiles (Q1 and Q3)

Descriptive Analytics: Measuring Dispersion of Data

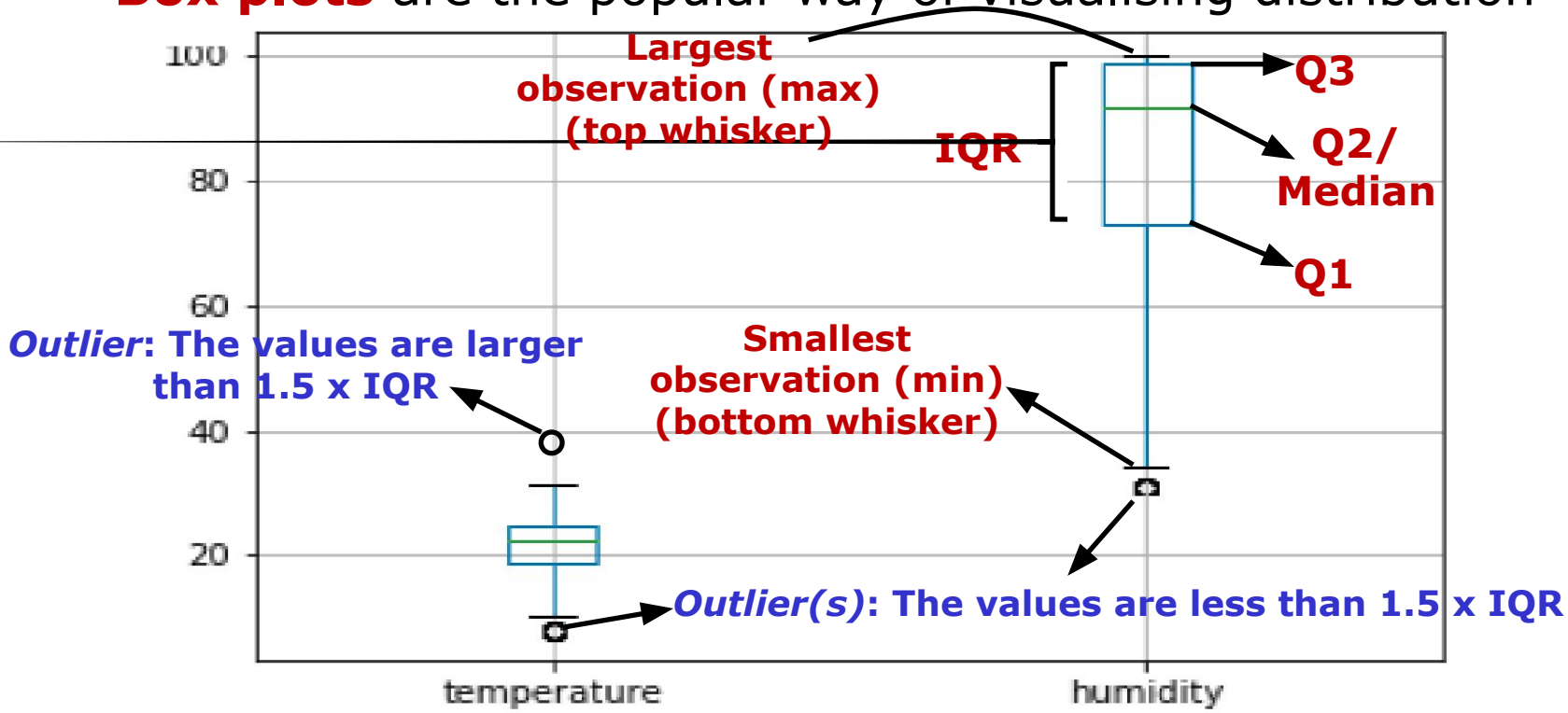
- The five-number summary of distribution:
  - It consists of minimum value, Q1, median, Q3 and maximum value
- **Box plots** are the popular way of visualising distribution



- 1.5 x IQR is equivalent to 2.7σ from mean if the distribution is normal distribution
  - It is close to 3σ from mean which is a standard in normal distribution

Descriptive Analytics: Measuring Dispersion of Data

- The five-number summary of distribution:
  - It consists of minimum value, Q1, median, Q3 and maximum value
- **Box plots** are the popular way of visualising distribution



- **Lower bound: Q1 – (1.5 x IQR)** **Upper bound: Q3 + (1.5 x IQR)**
- Outliers: Any datapoint less than the lower bound and larger than the upper bound

Descriptive Analytics: Measuring Dispersion of Data

- Variance (σ^2):
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. variance (σ^2) of this set of values is given by

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \mu = \text{mean}$$

- Standard deviation (σ):
  - The square root of variance $\sigma = \sqrt{\text{Variance}}$
- Standard deviation measures the spread about the mean
  - It is used when the mean is chosen as the measure of centre, especially in symmetric distribution
- The quartiles Q1 and Q3 measure the spread about median
  - Q1 and Q3 are used when the median is chosen as the measure of centre, especially in skewed distribution