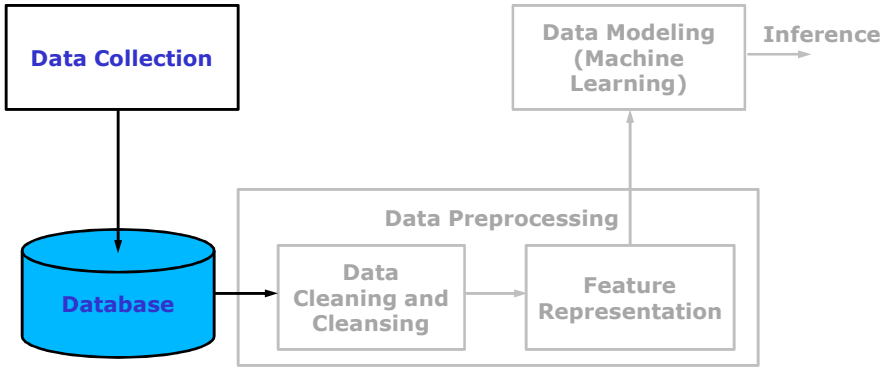
Data Preprocessing

Data Science

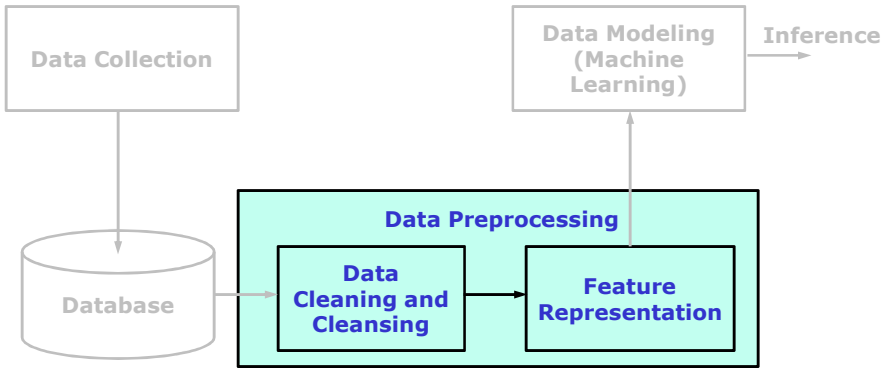
- Multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from structured and unstructured data
- Central concept is gaining insight from data
- Machine learning uses data to extract knowledge



2

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



3

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

Tuple (Record)

- A tuple (record) is finite ordered list (sequence) of elements, where each element is belonging to an attribute

Date/ Time	Temperature (C)/ Humidity (%)	Pressure (Pa)	Rain (inches)	Light Intensity (lux)	Accelerations (g)	Force (N)	Moisture (%)
2017-09-06 18:44:32	23.00,56.00	617.64	0.01	3	0.52,0.31,-0.80,0.00,0.00,0.00,31.36,-159.01	0.02	81.00
2017-09-06 18:33:32	24.00,58.00	619.47	0.01	12	0.52,0.30,-0.79,0.00,0.00,0.00,31.45,-159.12	0.02	82.00
2017-09-06 18:22:39	24.00,58.00	623.37	0.00	71	0.52,0.31,-0.80,0.00,0.00,0.00,31.35,-158.88	0.02	83.00
2017-09-06 18:11:31	25.00,60.00	627.02	0.05	194	0.51,0.31,-0.80,0.00,0.00,0.00,30.80,-159.00	0.02	81.00

Tuple (record)

- Each row is a tuple

Incomplete Data

- Many tuple (records) have no recorded value for several attributes
- Example:

	Dates	Station Id	Temperature	Humidity	Rain
2	08-07-2018	t10	25.46875	82.1875	6.75
3	09-07-2018			83.14912	
4	10-07-2018	t10	25.17021	85.34043	652.5
5	11-07-2018	t10	24.29851	87.68657	963
6	08-07-2018	t11			
7	09-07-2018	t11	26.8494	61.10241	15
8	10-07-2018	t11	27.88806	75.07463	13583.25
9	11-07-2018	t11	27.35915	76.02113	19768.5
10	23-07-2018	t12	24.39024	94.4065	1071
11	24-07-2018	t12	24.16197	97.66901	438.75
12	25-07-2018				
13	26-07-2018	t12	22.19718	99	864

Incomplete Data

- Many tuple (records) have no recorded value for several attributes
- Reasons for incomplete data:
  - User forgot to fill in a field
  - User chose not to fill out the field as it was not considered important at the time of the entry
  - Relevant data may not be recorded due to malfunctioning of equipment
  - Data might have lost while transferring from recorded place
  - Data may not be recorded due to programming error
  - Data might not be recorded due to technology limitations like limited memory

Noisy Data

- Many tuple (records) have incorrect value for several attributes
- Reasons for noisy data:
  - There may be human or computer error occurring in data entry
  - The data collection instruments used may be faulty
  - Error in data transmission
  - There may be technology limitation such as limited buffer size for coordinating synchronised data transfer and consumption

Inconsistent Data

- Data containing discrepancies in stored values for some attributes
- Reasons for inconsistent data:
  - It may result from inconsistencies in name conventions or data codes used or inconsistent formats of input fields such as date
  - Inconsistency in name convention or formats of input fields while integrating
  - Inconsistent data may be due to human or computer error occurring in data entry

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 summery and standard deviation

Descriptive Analytics: Measuring Central Tendency

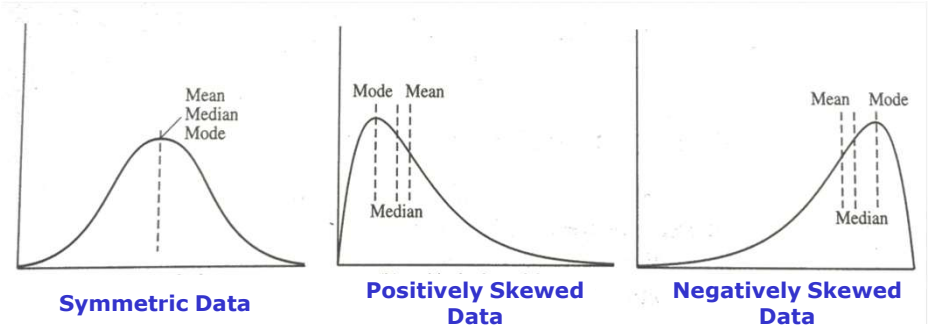
- Mean:
  - Let x_1, x_2, …, x_N be a set of N values in an attribute. Mean of this set of values is given by

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

  - Mean is a better measure of central tendency for the symmetric data (symmetrically distributed data)
- Median:
  - For asymmetrically distributed (skewed) data, a better measure of centre of data is median
  - 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

Descriptive Analytics: Measuring Central Tendency

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



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 set is the difference between the maximum and minimum values

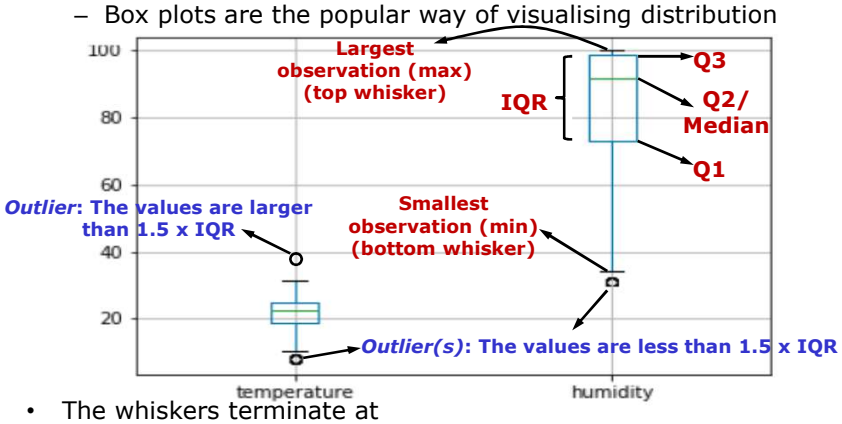
Descriptive Analytics: Measuring Dispersion of Data

- 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_i having the property that k percent of data entries lie at or below x_i
    - Median is the 50th percentile
    - The first quartile (Q1): It is the 25th percentile
    - The third quartile (Q3): It is the 75th 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
- Inter quartile range (IQR): Distance between the first and third quartile

IQR = Q3 – Q1

Descriptive Analytics: Measuring Dispersion of Data

- The five-number summery 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

- 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

19