DATA WAREHOUSING & MINING

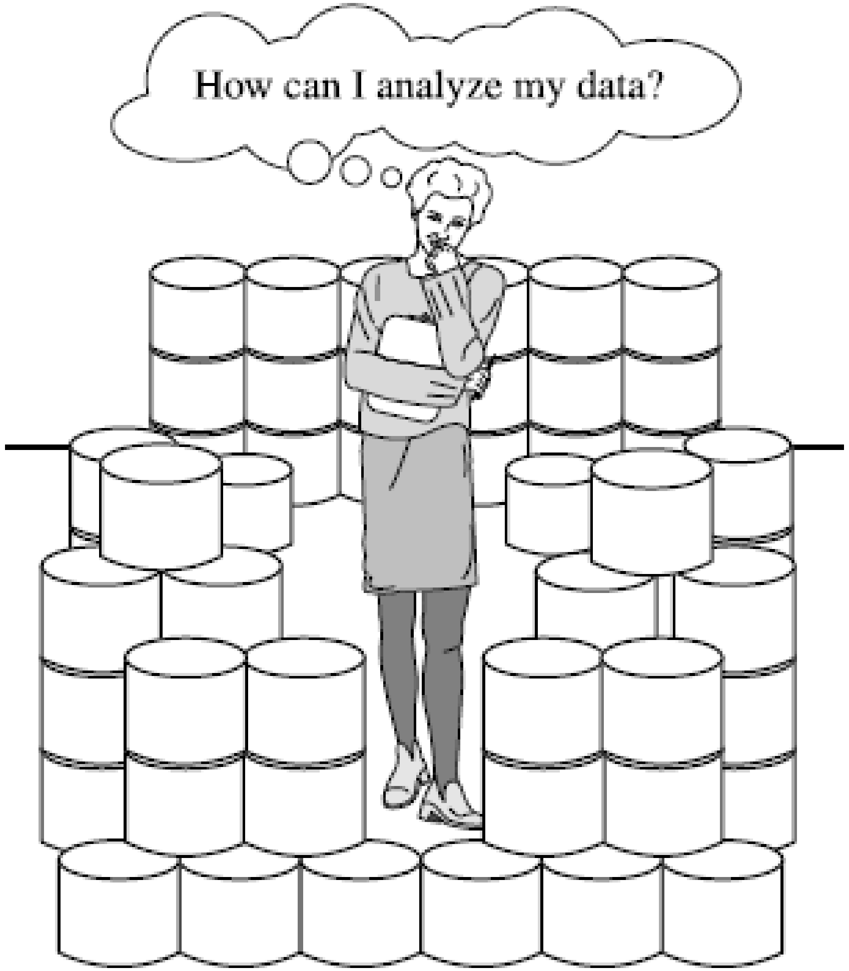
Unit-1: Introduction

Parameshwar R Hegde, PhD

Why Data Mining?

- **Lots of data being collected and stored**
  - Bank transactions
  - E-commerce
  - Hospital
  - Research
  - Social media

- **Strong competitive pressure**
  - Better and customized services (e.g., E-commerce)



Necessity of Data Mining

- **Data explosion**
  - Automated data collection tools and mature database technology led to large amount of data storage

- **Drowning in data, but starving for knowledge**

- **Solution is DATA MINING**
  - Extraction of interesting data/ knowledge from databases and data warehouses

Database Technology

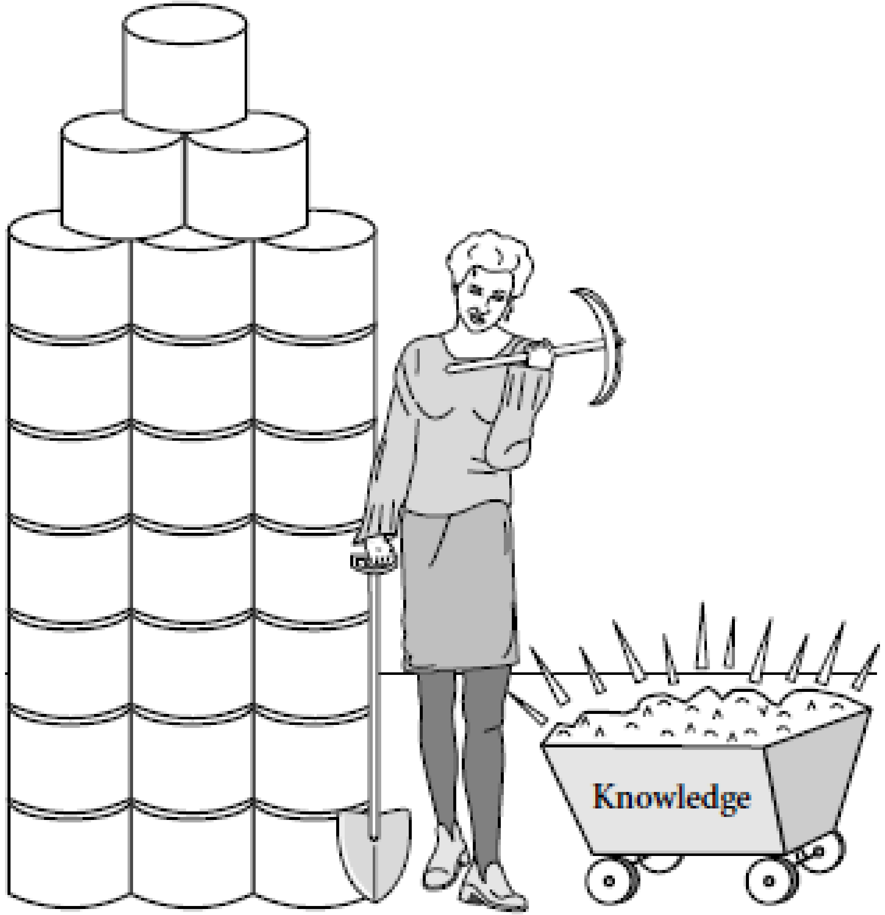
- **1960s**
  - Data collection and database creation
- **1970s**
  - Relational data model and relational DBMS implementation
- **1980s**
  - RDBMS, advanced data models and application oriented DBMS
- **1990s-2000s**
  - Multimedia databases, Web databases, data warehousing and data mining

What is Data Mining?

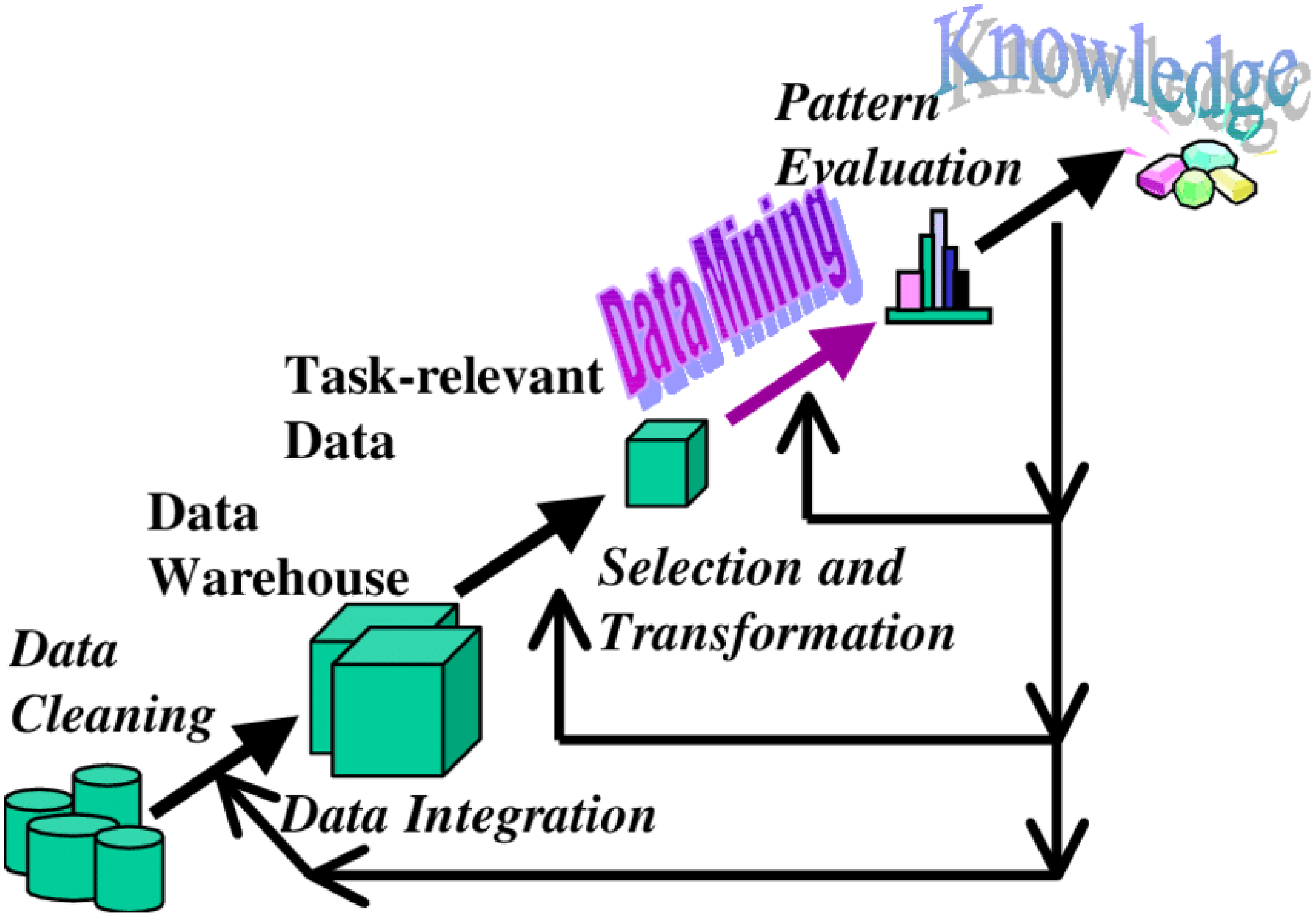
- Extraction of implicit, previously unknown and potentially useful information from data

- Data mining refers to extracting or "mining" knowledge from large amounts of data

- Can also be called as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging



Knowledge Discovery (KDD) Process



https://www.researchgate.net/figure/Data-Mining-is-the-core-of-Knowledge-Discovery-process_fig1_242778793

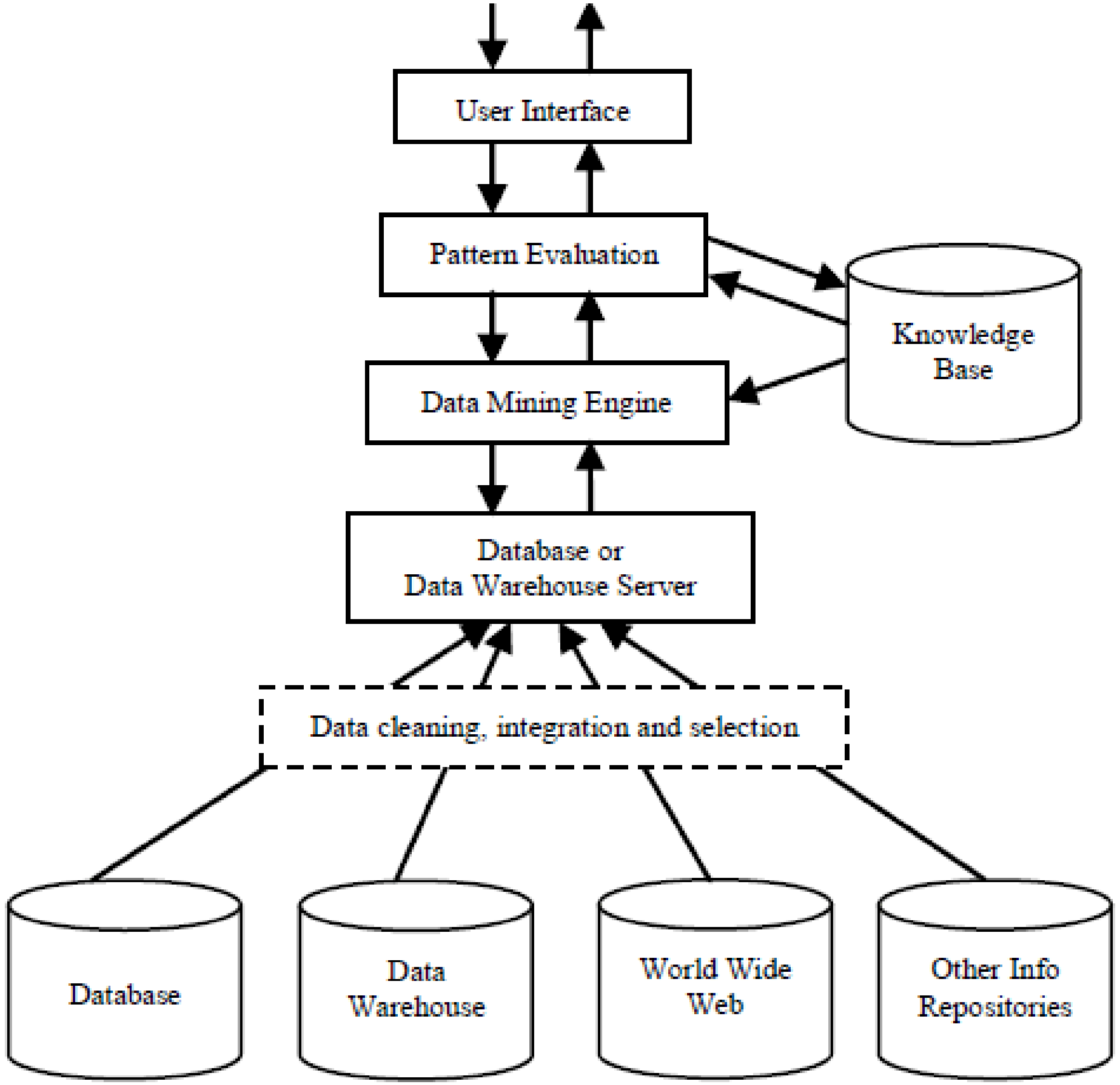
Knowledge Discovery (KDD) Process

- **Learning the application domain**
    - Relevant prior knowledge and goals of application
- **Creating a target dataset:** Data selection
- **Data cleaning:** Remove noise and inconsistent data
- **Data selection and transformation:** Find useful features, dimensionality/ variable reduction
- **Choosing data mining functions:** Summarization, classification, regression, association and clustering

Knowledge Discovery (KDD) Process

- **Data mining:** Search for patterns of interest
- **Pattern evaluation:** Identify the truly interesting patterns representing knowledge based on some interestingness measures
- **Knowledge presentation:** Visualization and knowledge representation techniques used to present the mined knowledge to the user
- **Use of knowledge discovery**

Architecture of Data Mining



Functionalities of Data Mining

- **In general “two” categories**
  1. Descriptive mining
  2. Predictive mining

- **Common data mining tasks**
  - Regression [Predictive]
  - Classification [Predictive]
  - Deviation Detection [Predictive]
  - Clustering [Descriptive]
  - Association Rule Discovery [Descriptive]
  - Sequential Pattern Discovery [Descriptive]

Functionalities of Data Mining

- **Classification and Prediction**
  - ➢ Classification: Finding models that describe and distinguish classes or concepts for future prediction
  - ➢ Prediction: Predict some unknown values
  - ➢ Presentation: Decision tree, classification rule and prediction

- **Cluster analysis**
  - Unknown class label: Group data to form new classes
  - Principle of clustering: maximizing the intra-class similarity, minimizing the inter-class similarity

Functionalities of Data Mining

- **Outlier analysis**
  - Outlier: A data/ object that does not comply with general behavior of the data
  - Quite useful in fraud detection, rare event analysis

- **Trend in evaluation**
  - Trend and deviation: regression analysis
  - Sequential pattern analysis

- **Other pattern-directed or statistical analysis**

Data Objects and Attributes

- **Data objects**
  - Essential part of database that represents the entity
  - Eg., University database: the objects may be students, professors and courses
  - Also referred to as samples, examples, instances and data points
  - Data objects are stored in a database, they are data tuples

- **Data attributes**
  - Data field that represents the characteristics or features of a data object
  - Commonly known as dimension, feature and variable
  - First step of data preprocessing
  - **Basically 2 types**
    1. Qualitative
    2. Quantitative

Qualitative Attributes

- **3 sub-types**
  1. **Nominal (related to names)**
     - ➢ Nominal means “relating to names”
     - ➢ Names of things or some kind of symbols
     - ➢ Categorical attributes and there is no order (rank, position) among values

       Eg., Occupation: with the values *teacher, dentist, programmer, farmer, and so on*
     - ➢ It is possible to represent such symbols or “names” with numbers

       Eg., hair_color, for instance: assign a code of 0 for black, 1 for brown, and so on

Qualitative Attributes

- **3 sub-types**

  2. Binary

  - ➢ Only 2 categories/ states: 0 or 1
  - ➢ Binary attributes are referred to as Boolean (*true and false*)

    Eg., Results: 2 states are pass or fail
  - ➢ 2 types

    1. **Symmetric:** Both values are equally important or carry the same weight;

       Eg., Gender: having the states male and female
    2. **Asymmetric:** Both values are not equally important

       Eg., Results: having the states pass and fail

Qualitative Attributes

- **3 sub-types; cont…**

  3. Ordinal

  - ➢ Values that have a meaningful sequence or ranking(order) between them
  - ➢ Order of values shows what is important but do not indicate how important it is

    Eg., Grade: with values A+, A, A−, B+, and so on
  - ➢ Useful for registering subjective assessments of qualities that cannot be measured objectively; hence, used in surveys
  - ➢ Discretization of numeric quantities by splitting the value range into a finite number of ordered categories gives ordinal attributes
  - ➢ Central tendency of an ordinal attribute can be represented by its mode and its median

Quantitative Attributes

- **Numeric**

  1. **Interval-Scaled**

     - ➢ Measured on a scale of equal-size units
     - ➢ Have order and can be positive, 0, or negative
     - ➢ Thus, such attributes allow us to compare and quantify the difference between values

       Eg., Temperature: with values outdoor temperature value for a number of different days
     - ➢ Can compute their mean value, in addition to the median and mode measures of central tendency

Quantitative Attributes

- **Numeric**
  2. **Ratio-Scaled**
     - ➢ A numeric attribute with an inherent zero-point
     - ➢ Can speak of a value as being a multiple (or ratio) of another value

       Eg., Years of experience: objects are employees

- **Discrete versus Continuous Attributes**
  - ➢ Discrete: A finite or countably infinite set of values, that may or may not be integer values; Eg., Hair color, Medical test
  - ➢ Continuous: A countably infinite values; eg., customer_ID
  - ➢ If an attribute is not discrete, it is continuous

    Terms numeric attribute and continuous attribute are often used interchangeably

Statistical Description of Data

- To identify properties of the data and highlight which data values should be treated as noise or outliers

- **3 Basic statistical descriptions**
  1. Measure of central tendency: Measure the location of the middle or center of a data distribution – mean, median, mode and midrange
  2. Dispersion of the data: how are the data spread out – quartiles, range and interquartile range; the variance and standard deviation (useful to find the outliers)
  3. Graphic displays: To visually inspect our data - bar charts, pie charts, line graphs, quantile plots, histograms, and scatter plots

Statistical Description of Data

- Measure of central tendency: Measure the location of the middle or center of a data distribution – mean, median, mode and midrange
  - ➢ Mean: Most popular measure of central tendency
    - ▪ Can be used for both discrete and continuous data
    - ▪ Equal to the sum of all the values in the data set divided by the number of values in the data set

$$\bar{x} = \frac{\sum x}{n}$$

    - ▪ Problem is – Affected by outliers
  - ➢ Median: Middle score for a set of data that has been arranged in order of magnitude
    - ▪ Less affected by outliers and skewed data

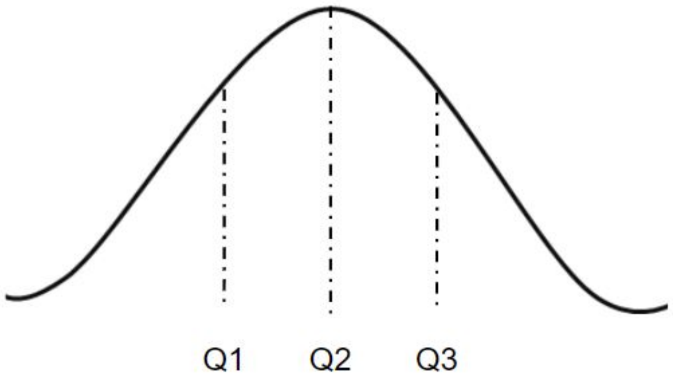
Statistical Description of Data

- Measure of central tendency: Measure the location of the middle or center of a data distribution – mean, median, mode and midrange
  - Mode: Most frequent score in our data set
    - Commonly, mode is used for categorical data
    - Problem is - It is not unique
    - Data set with two or more modes is multimodal
  - Midrange: Average of the largest and smallest values in the set
    - Easy to compute using the SQL aggregate functions, max() and min()

- Unimodal frequency curve with perfect symmetric data distribution, the mean, median, and mode are all at the same center value

Statistical Description of Data

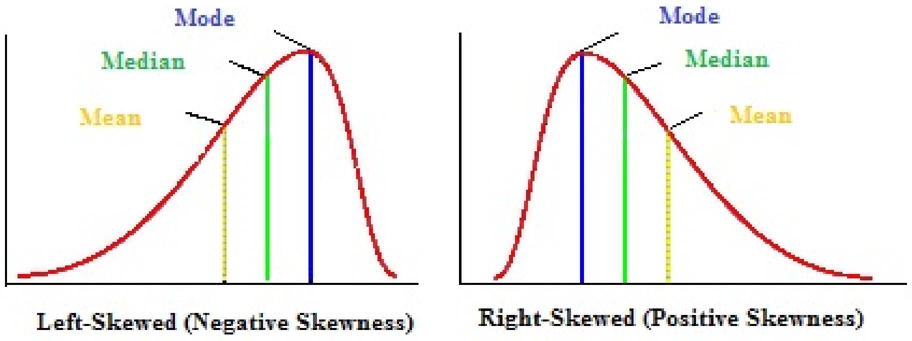
- Dispersion of the data: how are the data spread out – quartiles, range and interquartile range
  - Quartiles: Points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets



    - 2-quantile is the data point dividing the lower and upper halves of the data distribution
    - 4-quantiles are the three data points that split the data distribution into four equal parts
    - Interquartile range (IQR) and is defined as $Q_3 – Q_1$

Statistical Description of Data

- What if data is asymmetric?
  - No single numeric measure of spread (e.g., *IQR*) is very useful for describing skewed distributions
  - Median can not split data into equal-size halves for skewed distributions
  - Hence, it is more informative to provide two quartiles Q1 and Q3, along with the median
  - A fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well

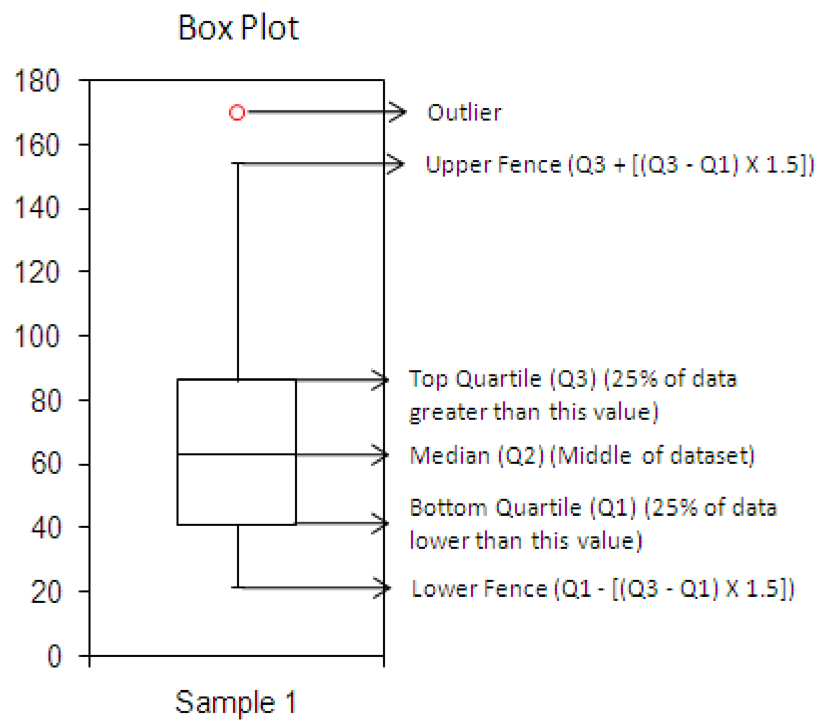

Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Statistical Description of Data

- **Five-number summary** of a distribution consists of the median (Q_2)

- **Boxplots** are a popular way of visualizing a distribution
  - Typically, the ends of the box are at the quartiles so that the box length is the interquartile range
  - The median is marked by a line within the box
  - Two lines (called *whiskers*) outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations

- Boxplots can be used in the comparisons of several sets of compatible data


Box Plot
180
160
140
120
100
80
60
40
20
0
Outlier
Upper Fence (Q3 + [(Q3 - Q1) X 1.5])
Top Quartile (Q3) (25% of data greater than this value)
Median (Q2) (Middle of dataset)
Bottom Quartile (Q1) (25% of data lower than this value)
Lower Fence (Q1 - [(Q3 - Q1) X 1.5])
Sample 1

https://www.listendata.com/2014/08/how-to-read-box-plot.html

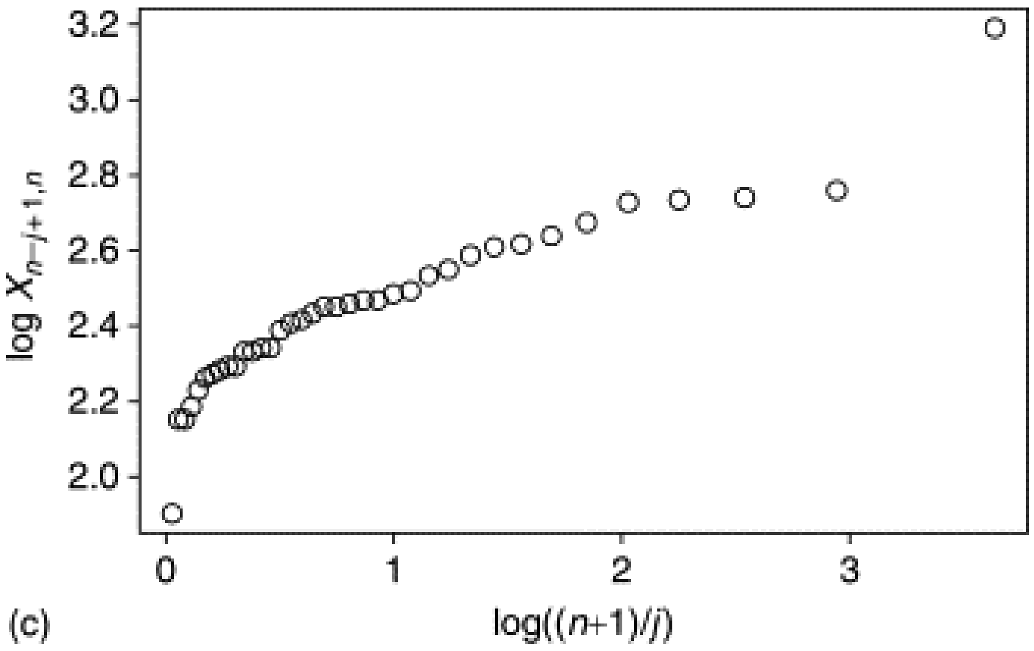
Statistical Description of Data

- Variance and Standard Deviation (SD)
  - Measures of data dispersion
  - σ^2 is variance
  - Indicate how spread out a data distribution is
  - Low SD (σ) means that the data observations tend to be very close to the mean
  - High SD (σ) indicates that the data are spread out over a large range of values
  - SD = 0 only when there is no spread; Otherwise, SD > 0

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

Graphic Displays of Statistical Description of Data

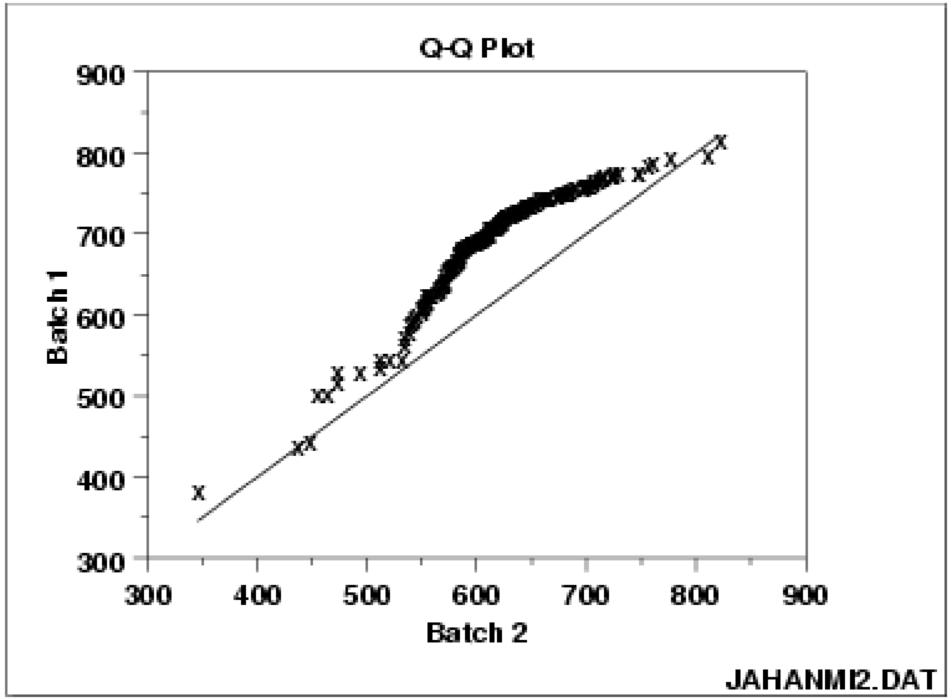
- **Quantile Plot**
  - A simple and effective way to have a first look at a univariate data distribution
  - First, it displays all of the data for the given attribute
  - Second, it plots quantile information



https://www.sciencedirect.com/topics/mathematics/quantile-plot

Graphic Displays of Statistical Description of Data

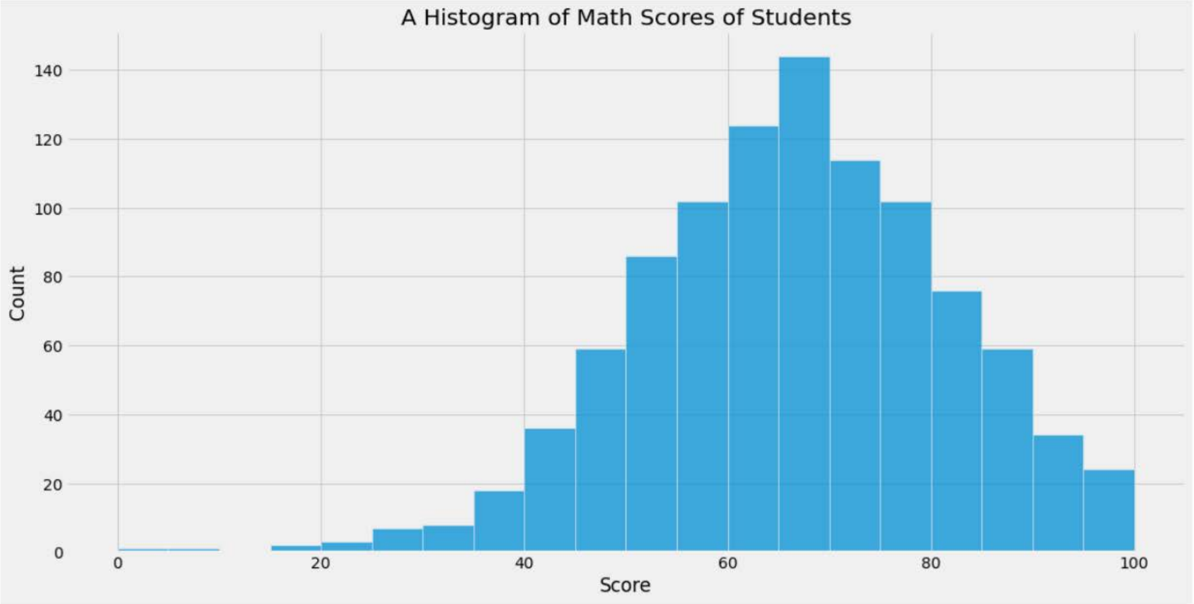
- **Quantile-Quantile Plot**
  - ➢ Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
  - ➢ Allows the user to view whether there is a shift in going from one distribution to another



https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm

Graphic Displays of Statistical Description of Data

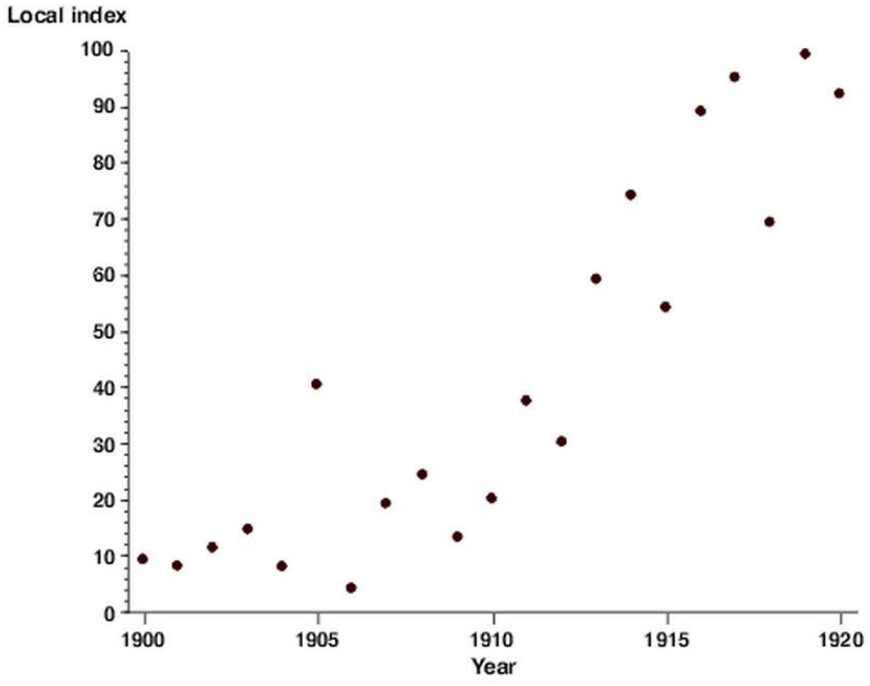
- **Histogram Plot**
  - ➢ “Histos” means pole or mast, and “gram” means chart
  - ➢ A graphical method for summarizing the distribution of a given attribute
  - ➢ May not be as effective as the quantile plot, q-q plot, and boxplot methods in comparing groups of univariate observations



https://towardsdatascience.com/3-best-often-better-alternatives-to-histograms-61ddaec05305

Graphic Displays of Statistical Description of Data

- **Scatter Plot**
  - ➢ Most effective graphical methods for determining relationship, pattern or trend between two numeric attributes
  - ➢ Each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane
  - ➢ Useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships



https://www.betterevaluation.org/en/evaluation-options/scatterplot

Self Study

- **Different data visualization techniques**
- **Measuring data similarity and dissimilarity**