# Unit 2
# Introduction to Data Mining

**<u>Prepared By</u>**

**Arjun Singh Saud, Asst. Prof. CDCSIT**

# Data Mining and KDD

- Simply stated, data mining refers to *extracting or mining knowledge from large amounts of data stored in databases, data warehouses, or other information repositories*.
- Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data or KDD.
- Alternatively, others view data mining as simply an essential step in the process of knowledge discovery from Data. KDD consists of an iterative sequence of the following steps:

# Data Mining and KDD

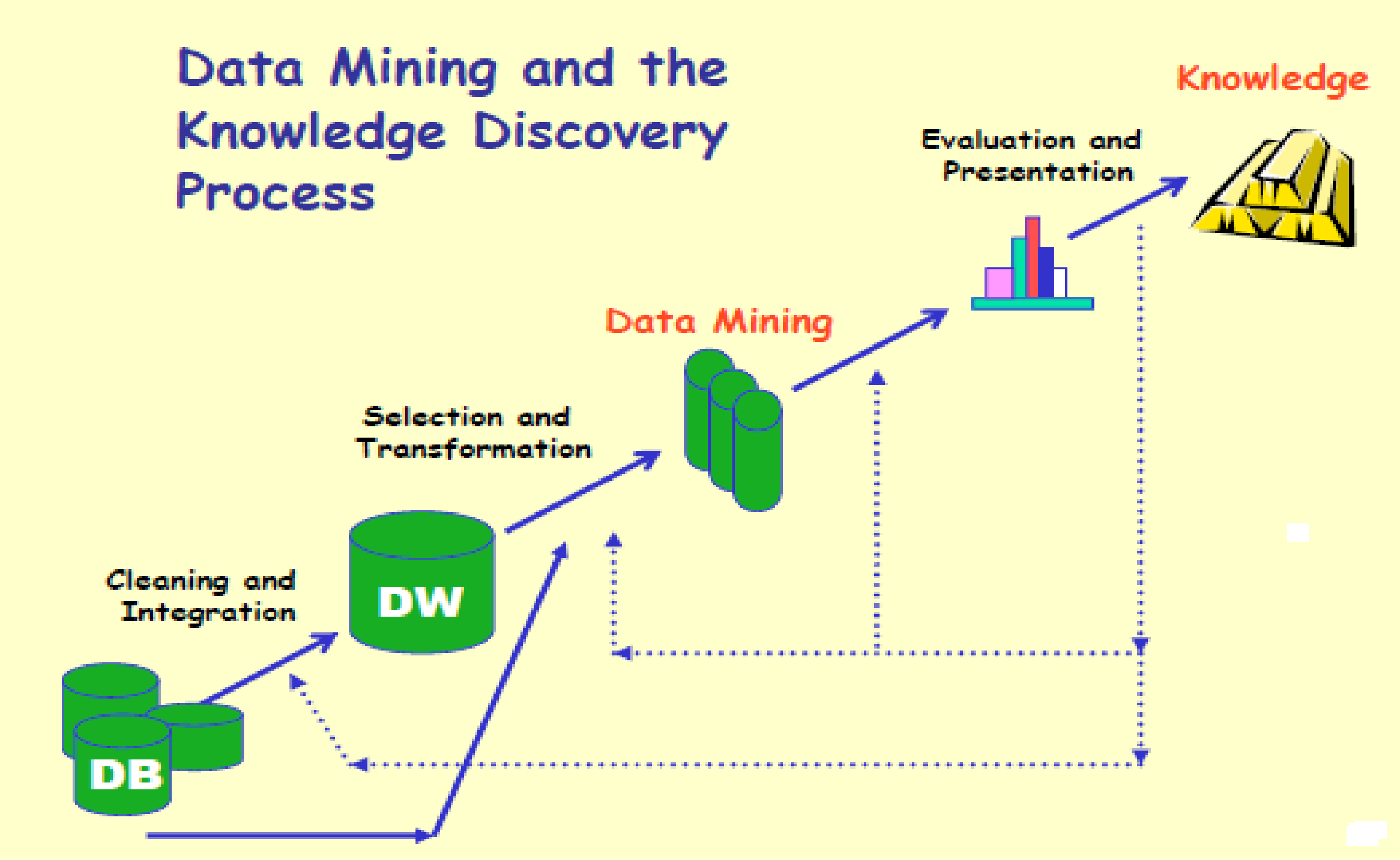

Data Mining and the Knowledge Discovery Process
Knowledge
Evaluation and Presentation
Data Mining
Selection and Transformation
Cleaning and Integration
DW
DB

Figure: Stages of KDD

# Data Mining and KDD

- **Data Cleaning**: Data cleaning is a process of removing unnecessary and inconsistent data from the databases. The main purpose of cleaning is to improve the quality of the data by filling the missing values, configuring the data to make sure that it in consistent format.
- **Data Integration**: In this stage multiple data sources may be combined (i.e. integrated) to form a large database.
- **Data Selection**: Data which is required for data mining process can be extracted from multiple and heterogeneous data sources such as databases, files etc. Data selection is a process where the appropriate data required for analysis is fetched from the databases.

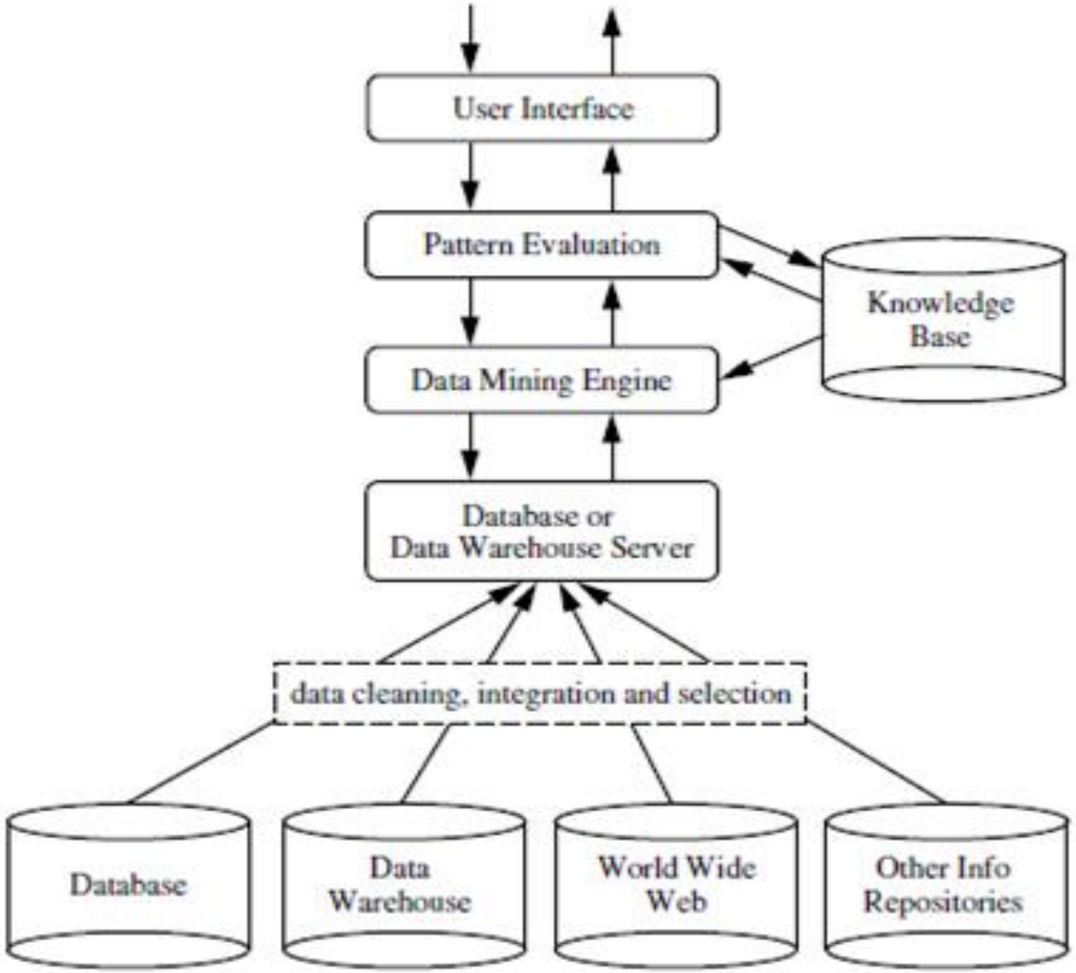
# Data Mining and KDD

- **Data Transformation**: In the transformation stage data extracted from multiple data sources are converted into an appropriate format for data mining process. Data reduction or summarization is used to decrease the number of possible values of data without affecting the integrity of data.
- **Data Mining**: It is the most essential step of KDD process where intelligent methods are applied in order to extract hidden patterns from data stored in databases.

# Data Mining and KDD

- **Pattern Evaluation**: This step Identifies the truly interesting patterns representing knowledge on the basis of some interestingness measures. Support and confidence are two widely used interestingness measures. These patterns are helpful for decision support systems.
- **Knowledge Presentation**: Knowledge representation and visualization techniques are used to present the mined knowledge to the user so that it will be easily understandable to them.

# Architecture of Data Mining System



# Architecture of Data Mining System

- **Data Sources or Repositories**: This component represents multiple data sources such as database, data warehouse, or any other information repository. Data cleaning and data integration techniques may be performed on the data.
- **Database Server or Data Warehouse Server**: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
- **Knowledge Base**: It is the area of knowledge that is used to guide the search, or to perform analysis of the resulting patterns.

# Architecture of Data Mining System

- **Data Mining Engine**: This is core component of the data mining system and consists of a set of functional modules for tasks such as association analysis, classification, Clustering, Evolution analysis, etc.
- **Pattern Evaluation Module:** This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns.

# Architecture of Data Mining System

- **Graphical User Interface**: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task. This component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

# Data Mining Functionalities

- Data mining functionalities or the kinds of patterns that can be discovered are described below.
  - Concept/Class Description
  - Association and Correlation
  - Classification and Regression
  - Clustering Analysis
  - Outlier Analysis
  - Evolution Analysis

# Data Mining Functionalities

**Concept/Class Description:**

- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via data characterization or data discrimination.
- Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

# Data Mining Functionalities

**Concept/Class Description**

- For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing a query on the sales database.
- There are several methods for effective data summarization and characterization. Simple data summaries can be generated based on statistical measures. The data cube-based OLAP roll-up operation can also be used to perform user-controlled data summarization along a specified dimension.

# Data Mining Functionalities

**Concept/Class Description**

- Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.
- For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period. The methods used for data discrimination are similar to those used for data characterization.

# Data Mining Functionalities

## Association and Correlation

- Frequent patterns are patterns that occur frequently in data. Frequent patterns may include frequent itemsets, and subsequences.
- A *frequent itemsets* typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.
- A pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a frequent subsequence. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

# Data Mining Functionalities

**Association and Correlation**

- For example, a marketing manager of an *Electronics store* would like to determine which items are frequently purchased together within the same transactions. For this mining rule can be

  *buys*(*X*; "*computer*"))=>*buys*(*X*; "*software*") [*support* = 20%; *confidence* = 50%]

  *Where X is a variable representing a customer*

- 20% **support** means that 20% of all the transactions under analysis show that computer and software are purchased together.

# Data Mining Functionalities

**Association and Correlation**

- A **confidence** of 50% means that if a customer buys a computer, there is a 50% chance that he/she will buy software as well.
- Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold. Additional analysis can be performed to uncover interesting statistical correlations between associated attribute-value pairs.

# Data Mining Functionalities

## Classification and Regression

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.
- Mainly it is used to predict the class of objects whose class label is unknown.
- The derived model is based on the analysis of a set of training data. Data object whose class label is known is considered as training data.
- The derived model may be represented in various forms, such as *classification rules*, *decision trees*, *mathematical formulae*, *neural networks* etc.

# Data Mining Functionalities

**Classification and Regression**

- Whereas classification predicts categorical labels, regression or prediction models are continuous-valued functions. That is, it is used to predict missing or unavailable *numerical data values* rather than class labels.
- Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.
- Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded.

# Data Mining Functionalities

## Classification and Regression

- If we want to classify a large set of items in the store, based on three kinds of responses to a sales campaign: *good response*, *mild response* and *no response*. We want to derive a model for each of these three classes based on the descriptive features of the items, such as *price, brand, place made, type*, and *category*. This type of problem can be solved using classification.
- Suppose instead, that rather than predicting categorical response labels for each store item, we would like to predict the amount of revenue that each item will generate during an upcoming sale, based on the previous sales data. This is an example of regression analysis.

# Data Mining Functionalities

## Cluster Analysis

- Unlike classification and prediction, which analyze output-labeled data objects, clustering analyzes data objects without consulting a known output-label.
- Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of *maximizing the intra-class similarity and minimizing the interclass similarity*.
- That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

# Data Mining Functionalities

## Cluster Analysis

- For example, cluster analysis can be performed on customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.

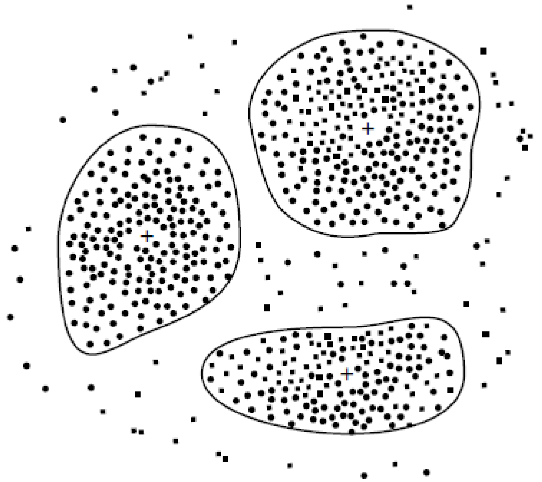


Figure: Three Data Clusters

# Data Mining Functionalities

## Outlier Analysis

- A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers.
- Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones.

# Data Mining Functionalities

## Outlier Analysis

- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers.
- For example, Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account.

# Data Mining Functionalities

**Evolution Analysis**

- Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.
- Distinct features of such an analysis include time-series data analysis, sequence or pattern matching.
- For example, you have the major stock market (time-series) data of the last several years available from the Nepal Stock Exchange. A data mining study of stock exchange may identify stock evolution regularities for overall stocks and for the stocks of particular companies. Such regularities may help predict future trends in stock market prices, contributing to our decision making regarding stock investments.

# Data Objects and Attribute Types

- Data sets are made up of data objects. A data object represents an entity.
- In a sales database, the objects may be customers, store items, and sales; in a university database, the objects may be students, professors, and courses.
- Data objects are typically described by attributes. Data objects can also be referred to as *samples, examples, instances, data points*, or *objects*.
- If the data objects are stored in a database, they are *data tuples*. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.

# Data Objects and Attribute Types

**What is an Attribute?**

- An attribute is a data field, representing a characteristic or feature of a data object. The nouns *attribute*, *dimension*, *feature*, and *variable* are often used interchangeably in the literature.
- The term *dimension* is commonly used in data warehousing. Machine learning literature tends to use the term *feature,* while statisticians prefer the term *variable*. Data mining and database professionals commonly use the term *attribute*.
- Attributes describing a customer object can include, for example*, customer ID*, *name,* and *address*.

# Data Objects and Attribute Types

## Types of Attributes

On the basis of  set of possible values attributes can be divided into following types

- Nominal Attributes
- Ordinal Attributes
- Interval-scaled Attributes
- Ratio-scaled Attributes

# Data Objects and Attribute Types

## Nominal Attributes

- The values of a **nominal attribute** are symbols or *names of things*. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as **categorical**. The values do not have any meaningful order.
- Examples of nominal attributes:
  - ✓Hair_color: possible values are: {black, brown, red, grey, white}
  - ✓Marital_status: possible values are:{Married, Single, Divorced, Widowed}
  - ✓Customer_ID: possible values are: Combination of numbers

# Data Objects and Attribute Types

**Nominal Attributes**

- It is possible to represent such symbols with numbers. With *hair_color*, for instance, we can assign a code of 0 for *black*, 1 for *brown*, and so on.
- However, in such cases, the numbers are not intended to be used quantitatively. That is, mathematical operations on values of nominal attributes are not meaningful. It makes no sense to subtract one customer ID number from another.
- A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present.

# Data Objects and Attribute Types

## Ordinal Attributes

- An **ordinal attribute** is an attribute with possible values that have a meaningful order or *ranking* among them, but the magnitude between successive values is not known.
- Examples of ordinal attributes:
  - ✓Grades: possible values are: {A+, A, A-, B+, B, B- and so on}
  - ✓Height: possible values are:{Tall, Medium, Short}
- The values have a meaningful sequence (which corresponds to increasing height ); however, we cannot tell from the values *how much* bigger, say, a medium is than a short.

# Data Objects and Attribute Types

**Ordinal Attributes**

- Ordinal attributes may also be obtained from the discretization of numeric quantities by splitting the value range into a finite number of ordered categories.
- Note that nominal, and ordinal attributes are *qualitative*. That is, they *describe* a feature of an object without giving an actual size or quantity.
- We can compute median and mode of ordinal attributes. However, we cannot compute mean.
- But, we can only compute mode of nominal attributes.

# Data Objects and Attribute Types

## Interval-Scaled Attributes

- Interval-scaled attributes are numeric attributes. A numeric attribute is *quantitative*; that is, it is a measurable quantity, represented in integer or real values.
- The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the *difference* between values.
- Because interval-scaled attributes are numeric, we can compute their mean value, in addition to the median and mode measures of central tendency.

# Data Objects and Attribute Types

## Interval-Scaled Attributes

- A *temperature* attribute is interval-scaled. Suppose that we have the outdoor *temperature* value for a number of different days, where each day is an object. By ordering the values, we obtain a ranking of the objects with respect to *temperature*.
- In addition, we can quantify the difference between values. For example, a temperature of $20^0$C is five degrees higher than a temperature of $15^0$C.
- Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart.

# Data Objects and Attribute Types

## Interval-Scaled Attributes

- Temperatures in Celsius and Fahrenheit do not have a true zero-point, that is, neither $0^0$C nor $0^0$F indicates “no temperature.”
- Although we can compute the *difference* between temperature values, we cannot talk of one temperature value as being a *multiple* of another.
- Without a true zero, we cannot say, for instance, that $10^0$C is twice as warm as $5^0$C. That is, we cannot speak of the values in terms of ratios. Similarly, there is no true zero-point for calendar dates.

# Data Objects and Attribute Types

## Ratio-Scaled Attributes

- A **ratio-scaled attribute** is a numeric attribute with an inherent zero-point.
- That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.
- In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.
- Temperature in Kelvin, length, counts, elapsed time, etc. are examples of ratio scaled attributes

# Discrete vs. Continuous Attributes

- A **discrete attribute** has a finite or countably infinite set of values, which may or may not be represented as integers.
- The attributes *hair_color, marital_status,* gender, etc. are examples of discrete attributes.
- If the set of possible values for an attribute is infinite it is said to be continuous attribute.
- Attributes *Customer_ID, temperature, etc.* are examples of continuous attributes.

# Statistical Description of Data

- For data preprocessing to be successful, it is essential to have an overall picture of your data.
- Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.
- Basic statistical descriptions include *Measure of Central Tendency and Measure of Dispersion*.

# Statistical Description of Data

## Measure of Central Tendency

- The most common and effective numeric measure of the "center" of a set of data is the *(arithmetic) mean*.
- Let $x_1$,$x_2$, …… ,$x_N$ be a set of *N* values or *observations*, such as for some numeric attribute *X*, like *salary*. The **mean** of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

# Statistical Description of Data

## Measure of Central Tendency

- Sometimes, each value *xi* in a set may be associated with a weight $w_i$ for *i* =1, ….. ,*N*. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

# Statistical Description of Data

## Measure of Central Tendency

- Although the mean is the useful quantity for describing a data set, it is not always the best way of measuring the center of the data. A major problem with the mean is its sensitivity to extreme (e.g., outlier) values. Even a small number of extreme values can corrupt the mean.
- For example, the mean salary at a company may be substantially pushed up by that of a few highly paid managers.

# Statistical Description of Data

**Measure of Central Tendency**

- For skewed (asymmetric) data, a better measure of the center of data is the median, which is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half.
- In probability and statistics, the median generally applies to numeric data; however, we may extend the concept to ordinal data.
- Suppose that a given data set of $N$ values for an attribute $X$ is sorted in increasing order. If $N$ is odd, then the median is the *middle value* of the ordered set.

# Statistical Description of Data

## Measure of Central Tendency

- If *N* is even, then the median is not unique; it is the two middlemost values and any value in between. If *X* is a numeric attribute in this case, by convention, the median is taken as the average of the two middlemost values.
- The *mode* is another measure of central tendency. The **mode** for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes.
- It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.

# Statistical Description of Data

## Measure of Central Tendency

- Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**. In general, a data set with two or more modes is **multimodal**. At the other extreme, if each data value occurs only once, then there is no mode.
- The **midrange** can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set.

# Statistical Description of Data

## Measure of Central Tendency

**Example**

- Suppose we have the following values for salary in thousands: 50, 52, 52, 56, 60, 63, 70, 70, 110, 30, 36, 47. Calculate mean, median, mode, and midrange for the above data.

# Statistical Description of Data

## Measure of Dispersion

- Let $x_1$,$x_2$, ……. ,$x_N$ be a set of observations for some numeric attribute, $X$. The **range** of the set is the difference between the largest and smallest values.
- Suppose that the data for attribute $X$ are sorted in increasing numeric order. **Quartiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.
- The 2-quartile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median.

# Statistical Description of Data

## Measure of Dispersion

- The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**.
- The 100-quantiles are more commonly referred to as **percentiles**; they divide the data distribution into 100 equal-sized consecutive sets.

# Statistical Description of Data

## Measure of Dispersion

- The quartiles give an indication of a distribution's center, spread, and shape. The **first quartile**, denoted by *Q*1, is the 25th percentile. It cuts off the lowest 25% of the data.
- The **third quartile**, denoted by *Q*3, is the 75th percentile—it cuts off the lowest 75% (or highest 25%) of the data.
- The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

# Statistical Description of Data

**Measure of Dispersion**

- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **interquartile range** (**IQR**) and is defined as

$$IQR = Q3 - Q1$$

- Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is.

# Statistical Description of Data

## Measure of Dispersion

- A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.
- The **variance** of *N* observations, $x_1, x_2, \ldots\ldots, x_N$, for a numeric attribute *X* is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- The **standard deviation**, σ , of the observations is the square root of the variance, $\sigma^2$.

# Statistical Description of Data

## Measure of Central Tendency

**Example**

- Suppose we have the following values for salary in thousands: 50, 52, 52, 56, 60, 63, 70, 70, 110, 30, 36, 47. Calculate range, 4-Quantiles, IQR, and variance for the above data.

# Applications of Data Mining

- Data mining can be applied in almost every field. Some of the major applications of data mining are briefly discussed below.
- **Business Intelligence**: Data mining help businesses perform effective market analysis, compare customer feedback on similar products, discover the strengths and weaknesses of their competitors, retain highly valuable customers, and make smart business decisions.

# Applications of Data Mining

- **Market Basket Analysis:** Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly.
- **Fraud Detection:** Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances. For example, detect suspicious money transactions.

# Applications of Data Mining

- **Intrusion Detection:** Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity.
- **Customer Segmentation:** Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. The business could offer them with special offers and enhance satisfaction.

# Applications of Data Mining

- **Bio Informatics:** Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease treatment optimization, etc.
- **Web Search Engines:** Web search engines are essentially very large data mining applications. Various data mining techniques are used in all aspects of search engines, ranging from *crawling*, indexing, and searching.

# Applications of Data Mining

- **Social Web and Networks**: There are a growing number of highly-popular user-centric applications such as blogs, wikis and Web communities that generate a lot of structured and semi-structured information. In these applications data mining can be used to explain and predict the evolution of social networks, personalized search for social interaction, user behavior prediction etc.

# Issue in Data Mining

- The major issues in data mining research, partitioning them into five groups
  - *Mining methodology*
  - *User interaction*
  - *Efficiency and scalability*
  - *Diversity of data types*, and
  - *Data mining and society*

# Issue in Data Mining

## Mining Methodology

- **Mining various and new kinds of knowledge:** As there are diverse applications, new mining tasks continue to emerge. These tasks can use the same database in different ways and require the development of new data mining techniques.
- **Mining knowledge in multidimensional space**: While searching for knowledge in large datasets, we need to explore multidimensional space. To find interesting patterns, various combinations of dimensions need to be applied.

# Issue in Data Mining

## Mining Methodology

- **Data mining—an interdisciplinary effort:** The power of data mining can be substantially enhanced by integrating methods from multiple disciplines. For example, to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and natural language processing.
- **Boosting the power of discovery in a networked environment:** Knowledge derived in one set of objects can be used to boost the discovery of knowledge in a "related" or semantically linked set of objects.

# Issue in Data Mining

## Mining Methodology

- **Handling uncertainty, noise, or incompleteness of data**: Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns. Therefore, techniques like data cleaning, data preprocessing, outlier detection and removal need to be integrated with the data mining process.
- **Pattern evaluation and pattern- or constraint-guided mining**: Not all the patterns generated by data mining processes are interesting. Therefore, techniques are needed to assess the interestingness of discovered patterns.

# Issue in Data Mining

## User Interaction

- **Interactive Mining:** Interactive mining allows users to focus the search for patterns from different angles. The data mining process should be interactive because it is difficult to know what can be discovered within a database.
- **Incorporation of background knowledge:** Background knowledge, constraints, rules, and other information regarding the domain under study should be incorporated into the knowledge discovery process. Such knowledge can be used for pattern evaluation as well as to guide the search toward interesting patterns.

# Issue in Data Mining

**User Interaction**

- **Ad hoc data mining and data mining query languages**: High-level data mining query languages will give users the freedom to define ad hoc data mining tasks. Optimization of the processing of such flexible mining requests is another promising area of study.
- **Presentation and visualization of data mining results**: The knowledge discovered by mining the data should be usable for humans. The system should adopt an expressive representation of knowledge, user-friendly visualization techniques, etc.

# Issue in Data Mining

## Efficiency and Scalability

- **Efficiency and scalability of data mining algorithms:** To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms:** The huge size of many databases, the wide distribution of data, and complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel.

# Issue in Data Mining

## Diversity of Data Types

- **Handling of relational and complex types of data:** There are many kinds of data stored in databases and data warehouses. It is not possible for one system to mine all these kind of data. So different data mining system should be construed for different kinds data.
- **Mining information from heterogeneous databases and global information systems:** Since data is fetched from different data sources on Local Area Network (LAN) and Wide Area Network (WAN).The discovery of knowledge from different sources of structured is a great challenge to data mining.