DATA MINING AND DATA ANALYTICS

DATA MINING

UNIT-I: Introduction to Data Mining 8

Introduction to Data Mining: Kinds of Data, Data Mining Functionalities - Interesting Patterns, Task Primitives, Issues in Data Mining, Data Preprocessing.

Introduction

What is Data

Data can be defined as a representation of facts, concepts, or instructions in a formalized manner, which should be suitable for communication, interpretation, or processing by human or electronic machine.

In other words, The Data is collection of **objects** defined by **attributes**.

A **data object** represents an entity.

- Also called as record, sample, example, instance, data point, object, tuple.

Examples:

- In a sales database, the objects may be customers, store items, and sales;
- In a medical database, the objects may be patients;
- In a university database, the objects may be students, professors, and courses.

Data objects are described by attributes, in other words, A collection of attributes describes an object.

Attributes

Objects

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Low	Yes
D2	Sunny	Hot	High	High	No
D3	Overcast	Hot	High	Low	Yes
D4	Rainy	Cold	Normal	High	No
D5	Rainy	Cold	Normal	Low	Yes
D6	Sunny	Hot	Normal	Low	Yes
D7	Overcast	Cold	Normal	High	No
D8	Rainy	Normal	Normal	Normal	Yes
D9	Overcast	Normal	Low	High	No
D10	Sunny	Normal	Normal	Medium	Yes

Database rows→dataobjects

database columns → attributes

Attributes

An attribute is a data field, representing property or feature of a data object.

- Also known as **dimension**, **feature**, and **variable**.

Examples:
Weight of a person, height, temperature, customer _ID, name, address etc.

What is Data Mining

Definition

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

(or)
Definition

Data Mining is a process of finding potentially useful patterns and valuable information from huge amount of data.

(or)
Definition

Data Mining is all about discovering hidden, unsuspected, and previously unknown yet valid relationships amongst the data.

(or)
Definition

Transforming tremendous amounts of data into organized knowledge.

(or)
Definition

Data Mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, Information extraction, data dredging, etc.

Data Mining Applications

- **Insurance:** Data mining helps insurance companies to price their products profitable and deciding whether to approve policy applications, including risk modelling and management for prospective customers.
- **Education:** Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.
- **Banking:** Data mining helps finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc. Bank and credit card companies use data mining tools to build financial risk models, detect fraudulent transactions and examine loan and credit applications.
- **Retail:** Data Mining techniques help retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to come up with the offer which encourages customers to increase their spending. Online retailers mine customer data and internet clickstream records to help them target marketing campaigns, ads and promotional offers to individual shoppers.
- **Service Providers:** Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze

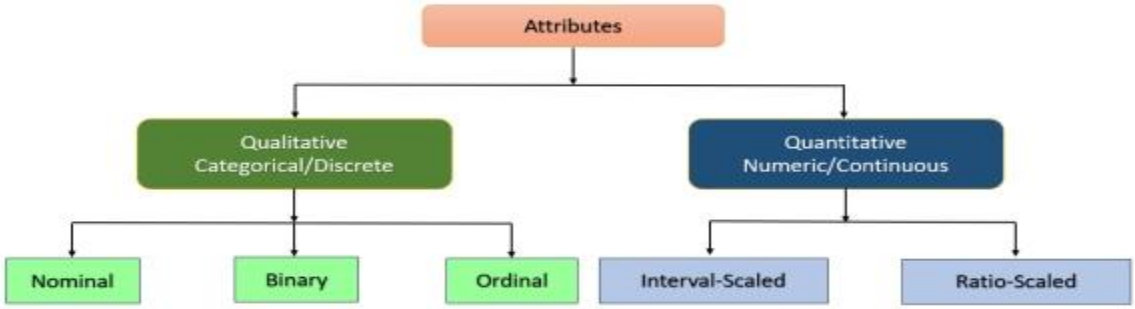
billing details, customer service interactions, complaints made to the company to assign each customer a probability score and offers incentives.

- **E-Commerce:** E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, who uses Data mining techniques to get more customers into their eCommerce store.
- **Super Markets:** Data Mining allows supermarkets develop rules to predict if their shoppers were likely to be expecting. By evaluating their buying pattern, they could find woman customers who are most likely pregnant. They can start targeting products like baby powder, baby shop, and diapers and so on.
- **Entertainment:** Streaming services do data mining to analyze what users are watching or listening to and to make personalized recommendations based on people's viewing and listening habits.
- **Healthcare:** Data mining helps doctors diagnose medical conditions, treat patients and analyze X-rays and other medical imaging results. Medical research also depends heavily on data mining, machine learning and other forms of analytics.

Attribute Types

Attribute values are numbers or symbols assigned to an attribute. The type of the attribute can be determined based on the assigned value.

The set of possible values - nominal, binary, ordinal, or numeric - the attribute can have.



Nominal Attributes

- The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state.
- Nominal attributes are also referred to as Qualitative and Categorical attributes.
- The values of nominal attributes do not have any meaningful order.

Example

Attributes	**Possible Values**
hair_color	black, brown, red, green, and so on.
marital_status	single, married, divorced, and widowed.
occupation	teacher, doctor, farmer, student and so on.

The nominal attribute values do not have any meaningful order about them and they are not quantitative. So

- It makes no sense to find the mean (average) value or median (middle) value for such an attribute.
- However, we can find the attribute’s most commonly occurring value (mode)

Binary Attributes

A binary attribute is a special nominal attribute with only two states: 0 or 1. Where 0 typically means that the attribute is absent, and 1 means that it is present.

Symmetric Binary Attribute

A binary attribute is symmetric if both of its states are equally valuable and carry the same weight.

Example: the attribute gender having the states male and female.

Asymmetric Binary Attribute

A binary attribute is asymmetric if the outcomes of the states are not equally important.

Example: Test results for COVID patient: Positive (1) and Negative (0).

By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., COVID positive) and the other by 0 (e.g., COVID negative).

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.

- Ordinal attributes are also referred to as Qualitative and Categorical attributes.

Example: An ordinal attribute drink size corresponds to the size of drinks available at a fast-food restaurant.

- This attribute has three possible values: small, medium, and large.
- The values have a meaningful sequence (which corresponds to increasing drink size);
- However, we cannot tell from the values how much bigger, say, a medium is than a large.

Ordinal attributes are useful in surveys, In one survey, participants were asked to rate how satisfied they were as customers.

Customer satisfaction had the following ordinal categories:

0: very dissatisfied
1: somewhat dissatisfied
2: neutral
3: satisfied
4: very satisfied.

The central tendency of an ordinal attribute can be represented by its mode and its median (middle value in an ordered sequence), but the mean cannot be defined.

Interval-Scaled Attributes

Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative.We can compare and quantify the difference between values of interval attributes.

Examples:

A temperature attribute is an interval attribute.

We can quantify the difference between values. For example, a temperature of 20^{o}C is five degrees higher than a temperature of 15^{o}C.

Calendar dates is another example for an interval attribute.

Temperatures in Celsius do not have a true zero point, that is, 0^{o}C does not indicate “no temperature.”

Calendar dates do not have a true zero point, that is, the year 0 not the beginning of the time.

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 10oC is twice as warm as 5oC. That is, we cannot speak of the values in terms of ratios.
The central tendency of an interval attribute can be represented by its mode, its median (middle value in an ordered sequence), and its mean Data.

Ratio Attribute

A ratio attribute is a numeric attributes with an inherent zero point.
Examples:

- **number_of_words** in a documents object.
- **count** attribute such as years of experience for employee object.
- Attributes to measure **weight, height, latitude,** and **longitude** coordinates.
- With an **amount** attribute we can say "you are 100 times richer with $100 than with $1".

- If a measurement is ratio scaled, we can speak of a value as being a multiple (or ratio) of another value.
The central tendency of an ratio attribute can be represented by its mode, its median (middle value in an ordered sequence), and its mean
Properties of Attribute Values
The type of an attribute depends on which of the following properties it possesses:

- Distinctness: =, !=
- Order: < >
- Addition: + -
- Multiplication: * /

Nominal attribute: distinctness
Ordinal attribute: distinctness & order
Interval attribute: distinctness, order & addition
Ratio attribute: all 4 properties

Basic Statistical Descriptions of 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.
For data preprocessing tasks, we want to learn about data characteristics regarding **central tendency** of the data.

- Measures of central tendency include **Mean**, **Median**, and **Mode**.

Mean

The most common and effective numeric measure of the **"center"** of a set of data is the (arithmetic) mean. Let x1, x2, ……, xN 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
$\bar{x}=\sum i=1N xiN=x1+x2+ \ldots +xNN$
Example: Mean. Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using above Eq., we have
$\bar{x}=30+36+47+50+52+52+56+60+63+70+70+11012=69612=58$

Thus, the mean salary is $58,000.
Sometimes, each value xi in a set may be associated with a weight wi for i = 1,, ,N. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute

$$\bar{x}=\sum_{i=1}^{N} w_i x_i \Big/ \sum_{i=1}^{N} w_i = w_1 x_1 + w_2 x_2 + \ldots + w_N x_N \Big/ w_1 + w_2 + \ldots + w_N$$

This is called the **weighted arithmetic mean** or the **weighted average**.

Median

Another measure of the center of data is the **median**. Suppose that a given data set of N distinct values is **sorted** in numerical order.

- If **N is odd**, the median is the **middle value** of the ordered set;
- If **N is even**, the median is the **average of the middle two values**.

Example: Median. Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
There is an **even number** of observations (i.e., 12); therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list). By convention, we assign the **average of the two middlemost values** as the median;
that is

$$52+562=1082=54$$

Thus, the median is $54,000.
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.
- If **N is even**, then the **median may not be not unique**.

In this case, the median is the two middlemost values and any value in between.

Mode

Another measure of central tendency is the **mode**. The mode for a set of data is the value that occurs **most frequently** in the set.
It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.

- Data sets with one, two, or three modes: called **unimodal**, **bimodal**, and **trimodal**.
- At the other extreme, if each data value occurs **only once**, then there is **no mode**.

Example: Mode. Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
The above data has **bimodal mode**.i. e The two modes are 52 and 70.

Midrange

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.
Example: Midrange. Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
The midrange of the data is

$$30,000+110,0002=\$70,000$$

Thus, the median is $70,000

Central Tendency Measures for different attributes:

Central Tendency Measures for Numerical Attributes: **Mean, Median, Mode**

Central Tendency Measures for Categorical Attributes:

- Central Tendency Measures for Nominal Attributes: **Mode**
- Central Tendency Measures for Ordinal Attributes: **Mode, Median**

Example:

What are central tendency measures (mean, median, mode) for the following attributes?

Solution:

attr1={2,4,4,6,8,24}

mean=(2+4+4+6+8+24)/6=8averageofallvalues

median = (4+6)/2 = 5 avg. of two middle values

mode = 4 most frequent item

attr2 = {2,4,7,10,12}

mean = (2+4+7+10+12)/5 = 7 average of all values

median = 7 middle value

mode = any of them (no mode) all of them has same freq.

attr3 = {xs, s, s, s, m, m, l}

mean is meaningless for categorical attributes.

median = s middle value

mode = s most frequent item

Knowledge Discovery from Data (KDD)

The need of data mining is to extract useful information from large datasets and use it to make predictions or better decision-making. Nowadays, data mining is used in almost all places where a large amount of data is stored and processed.

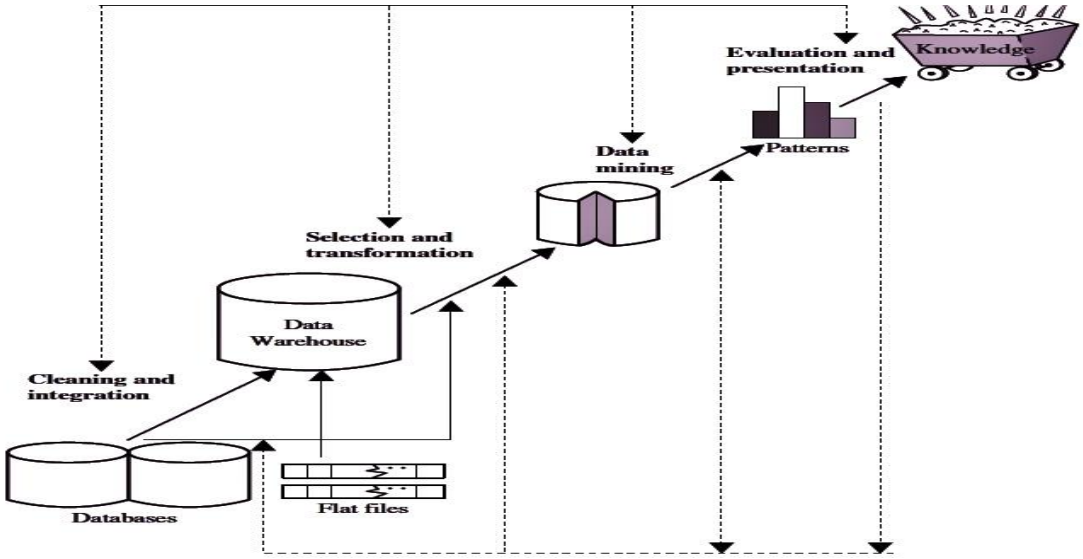
For examples: Banking sector, Market Basket Analysis, Network Intrusion Detection.

Data Mining also known as **K**nowledge **D**iscovery from **D**ata or **KDD**.

Knowledge Discovery from Data (KDD) Process

KDD is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets.

The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.



The following steps are included in KDD process:

1. **Data Cleaning**
2. **Data Integration**
3. **Data Selection**
4. **Data Transformation**
5. **Data Mining**
6. **Pattern Evaluation**
7. **Knowledge Representation**

1. Data Cleaning

Data cleaning is defined as removal of **noisy** and **irrelevant/ inconsistent** data from data collection.

- Cleaning in case of **Missing values**.
- Cleaning **noisy data**, where noise is a **random** or **variance error**.

In this step, the **noise** and **inconsistent** data is removed.

2. Data Integration

Data integration is defined as heterogeneous **data from multiple data sources** combined in a common source (Data Warehouse).
i.e., In this step, multiple data sources may be combined as single data source.
A popular trend in the information industry is to perform **data cleaning** and **data integration** as a **data preprocessing** step, where the resulting data are stored in a **data warehouse**.

3. Data Selection

Data selection is defined as the process where **data relevant to the analysis** is decided and retrieved from the data collection. This step in the KDD process is **identifying** and **selecting** the relevant data for analysis.

4. Data Transformation

Data Transformation is defined as the process of **transforming data into appropriate form** required by mining procedure. This step involves reducing the data dimensionality, aggregating the data, normalizing it, and discretizing it to prepare it for further analysis.

Data Mining

This is the heart of the KDD process and involves applying **various data mining techniques** to the transformed data to discover **hidden patterns, trends, relationships**, and **insights**. A few of the most common data mining techniques include clustering, classification, association rule mining, and anomaly detection.

5. Pattern Evaluation

After the data mining, the next step is to evaluate the **discovered patterns** to determine their usefulness and relevance. This involves assessing the quality of the patterns, evaluating their significance, and selecting the most promising patterns for further analysis.

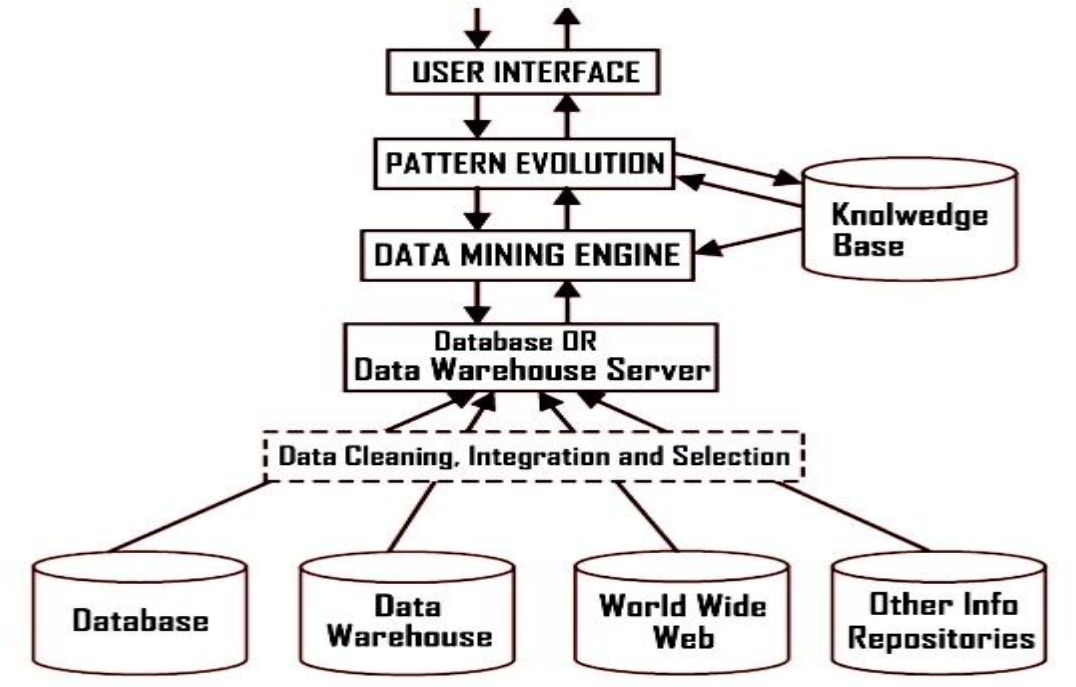
6. Knowledge Representation

This step involves **representing the knowledge** extracted from the data in a way humans can easily understand and use. This can be done through visualizations, reports, or other forms of communication that provide meaningful insights into the data.

Data Mining architecture

Data mining is a very important process where potentially useful and previously unknown information is extracted from large volumes of data. There are several components involved in the data mining process.

The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.



Data Sources

Database, data warehouse, World Wide Web (WWW), text files and **other documents** are the actual sources of data. You need large volumes of historical data for data mining to be successful.

Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets, or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

The data needs to be **cleaned, integrated,** and **selected** before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable. So, first data needs to be cleaned and integrated.

Database or Data Warehouse Server

The database or data warehouse server contains the **actual data** that is **ready to be processed**. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

Data Mining Engine

The data mining engine is the **core component** of any data mining system. It consists of **several modules** for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

Pattern Evaluation Modules

The pattern evaluation module is mainly responsible for the **measure of interesting of the pattern** by using a **threshold value**. It interacts with the data mining engine to focus the search towards interesting patterns.

Graphical User Interface

The graphical user interface module provides the **communication** between the **user** and the **data mining system**. This module helps the user use the system easily
and efficiently without knowing the real complexity behind the process.

When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

Knowledge Base

The knowledge base is **helpful** in the **whole data mining process**. It might be useful for guiding the search or evaluating the interesting of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining.

The data mining engine might get inputs from the knowledge base to make the **result more accurate** and **reliable**. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

Types of Data

What Kinds of Data Can Be Mined

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application.

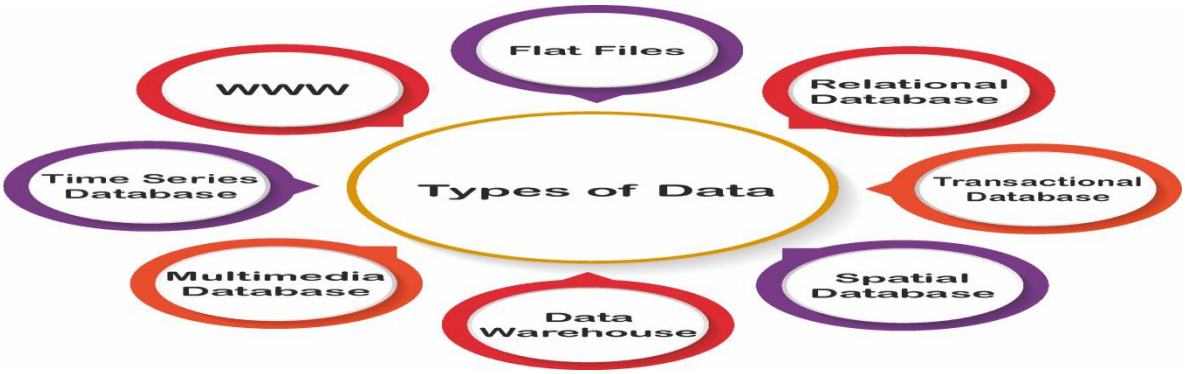
The following are the most basic forms of data for mining.

Basic forms of data for mining

- Database Data (or) Relational database
- Data warehouse data
- Transactional data

Other forms of data for mining

- Multimedia Database
- Spatial Database
- World Wide Web
- Text data (Flat File)
- Time series database

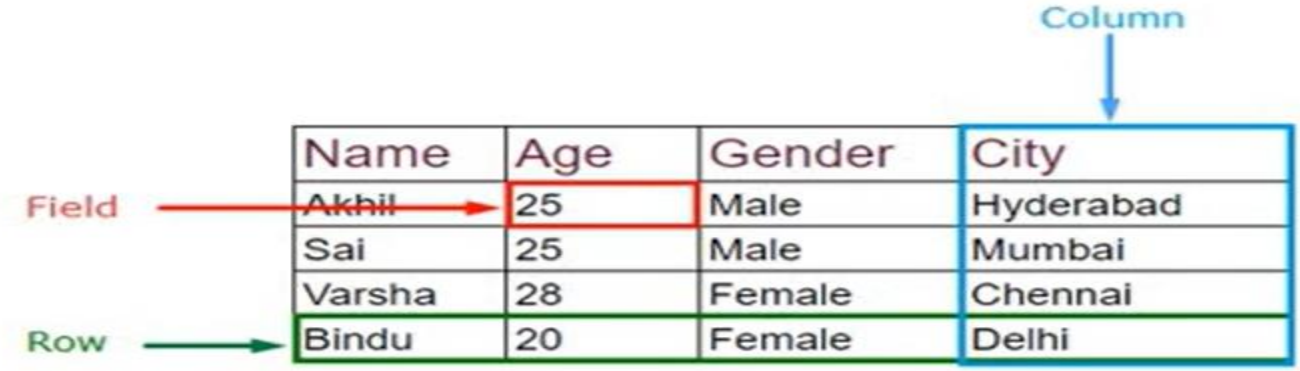


Database Data (or) Relational database

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.

A relational database: is a collection of tables, each of which is assigned a unique name, each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.

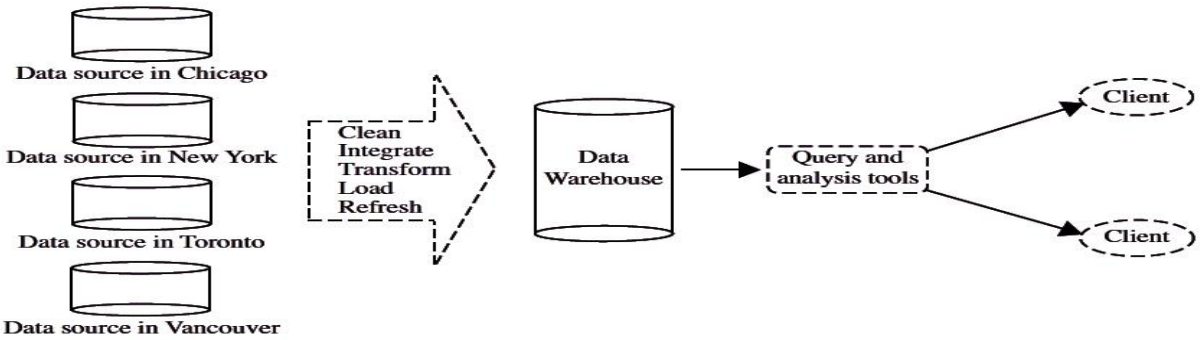
Example:



Name	Age	Gender	City
Akhil	25	Male	Hyderabad
Sai	25	Male	Mumbai
Varsha	28	Female	Chennai
Bindu	20	Female	Delhi

Data warehouse data

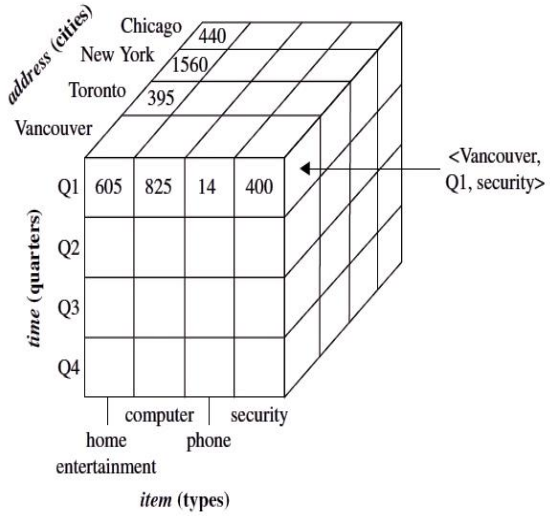
A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.



A data warehouse is defined as the collection of data integrated from multiple sources. Later this data can be mined for decision making.

A data warehouse is usually modelled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum. A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

Example:



Transactional data

Transactional database is a collection of data organized by time stamps, date etc to represent transaction in databases. In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.

A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction.

This type of database has the capability to roll back or undo operation when a transaction is not completed or committed. And it follows ACID property of DBMS.
Example:

TID	Items
T1	Bread, Coke, Milk
T2	Popcorn, Bread
T3	Popcorn, Coke, Egg, Milk
T4	Popcorn, Bread, Egg, Milk
T5	Coke, Egg, Milk

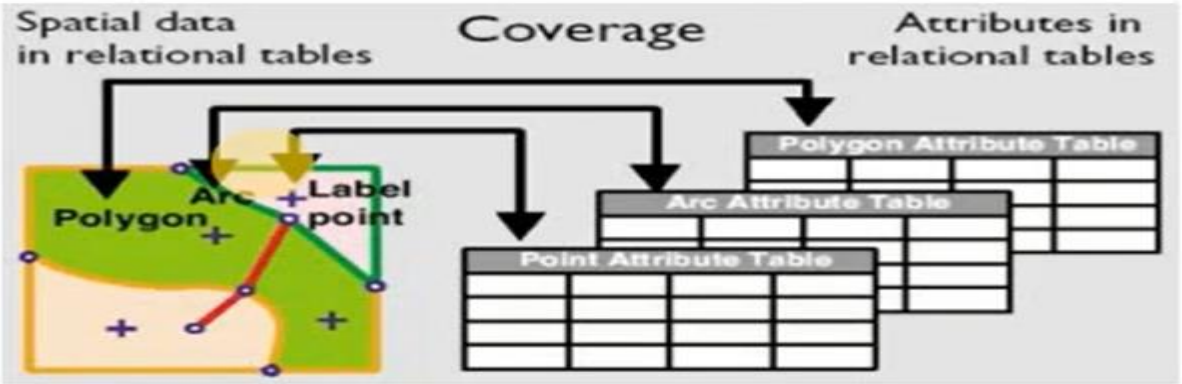
Fig: Transactional data

Multimedia database

The multimedia databases are used to store multimedia data such as images, animation, audio, video along with text. This data is stored in the form of multiple file types like **.txt**(text), **.jpg**(images), **.swf**(videos), **.mp3**(audio) etc.

Spatial database

A spatial database is a database that is enhanced to store and access spatial data or data that defines a geometric space. These data are often associated with geographic locations and features, or constructed features like cities. Data on spatial databases are stored as coordinates, points, lines, polygons and topology.



World Wide Web

The World Wide Web is a collection of documents and resources such as audio, video, and text. It identifies all this by URLs of the web browsers which are linked through HTML pages. Online shopping, job hunting, and research are some uses.

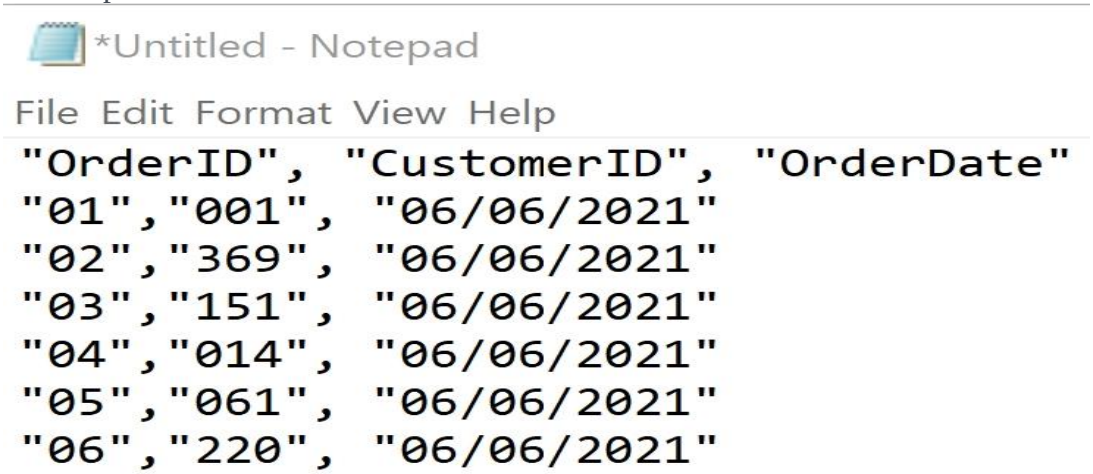
It is the most heterogeneous repository as it collects data from multiple resources. And it is dynamic in nature as Volume of data is continuously increasing and changing.

Text data (Flat File)

Flat files are a type of structured data that are stored in a plain text format. They are called "flat" because they have no hierarchical structure, unlike a relational database table. Flat files typically consist of rows and columns of data, with each row representing a single record and each column representing a field or attribute within that record. They can be stored in various formats such as CSV, tab-separated values (TSV) and fixed-width format.

- Flat files are defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.
- Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables.

Example:



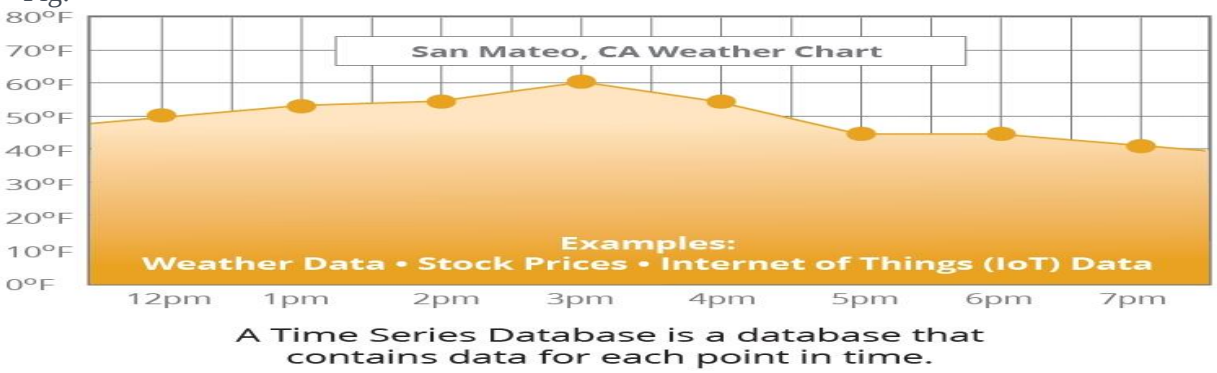
Time series database

Time-series data is a sequence of data points collected over time intervals, allowing us to track changes over time. Time-series data can track changes over milliseconds, days, or even years.

A time series database (TSDB) is a database optimized for time-stamped or time series data. Time series data are simply measurements or events that are tracked, monitored, downsampled, and aggregated over time. This could be server metrics, application performance monitoring, network data, sensor data, events, clicks, trades in a market, and many other types of analytics data.

Example:

Fig:



A Time Series Database is a database that contains data for each point in time.

What is Data Mart?

A Data Mart is focused on a single functional area of an organization and contains a subset of data stored in a Data Warehouse. A Data Mart is a abbreviated version of Data Warehouse and is designed for use by a specific department, unit or set of users in an organization. E.g., Marketing, Sales, HR or finance. It is often controlled by a single department in an organization.

Data Mining Functionalities

Data mining is important because there is so much data out there, and it's impossible for people to look through it all by themselves.

Data mining uses various functionalities to analyze the data and find patterns, trends, and other information that would be hard for people to find on their own.

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such data mining tasks can be classified into two categories: **descriptive** and **predictive**.

Descriptive data mining

Similarities and patterns in data may be discovered using descriptive data mining.

This kind of mining focuses on transforming raw data into information that can be used in reports and analyses. It provides certain knowledge about the data,

for instance, count, average.

It gives information about what is happening inside the data without any previous idea. It exhibits the common features in the data. In simple words, you get to know the general properties of the data present in the database.

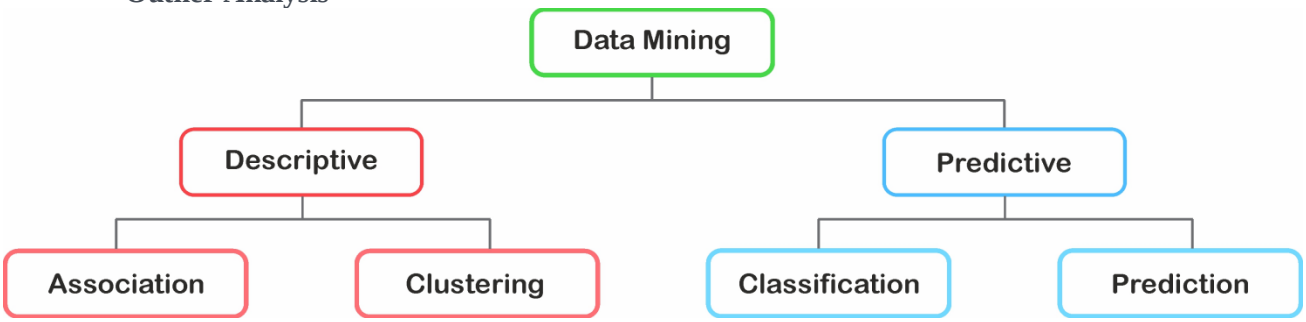
Predictive data mining

These kind of mining tasks perform inference on the current data in order to make predictions.

This helps the developers in understanding the characteristics that are not explicitly available. For instance, the prediction of business analysis in the next quarter with the performance of the previous quarters. In general, the predictive analysis predicts or infers the characteristics with the previously available data.

The following are data mining functionalities:

- **Class/Concept Description (Characterization and Discrimination)**
- **Classification**
- **Prediction**
- **Association Analysis**
- **Cluster Analysis**
- **Outlier Analysis**



Class/Concept Description: Characterization and Discrimination

Data is associated with classes or concepts.

Class: A collection of things sharing a common attribute

Example: Classes of items – computers and printers

Concept: An abstract or general idea derived from specific instances.

Example: Concepts of customers – big Spenders and budget Spenders.

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 using **data characterization** and **data discrimination**, or both.

Data characterization

Data characterization is a summarization of the general characteristics or features of a target class of data.

Data summarization can be done based on statistical measures and plots.

The output of data characterization can be presented in various forms it includes pie charts, bar charts, curves, and multidimensional data cubes.

Example: A customer relationship manager at All Electronics may order the following data mining task: Summarize the characteristics of customers who spend more than **$5000** a year at All Electronics.

The result is a general profile of these customers, such as that they are **40 to 50 years old**, employed, and have **excellent credit ratings**.

Data discrimination

Data discrimination is one of the functionalities of data mining. It compares the data between the two classes. Generally, it maps the target class with a predefined group or class. It compares and contrasts the characteristics of the class with the predefined class using a set of rules called discriminate rules.

Example: A customer relationship manager at All Electronics may want to compare two groups of customers those who shop for computer products regularly(e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year).

The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.

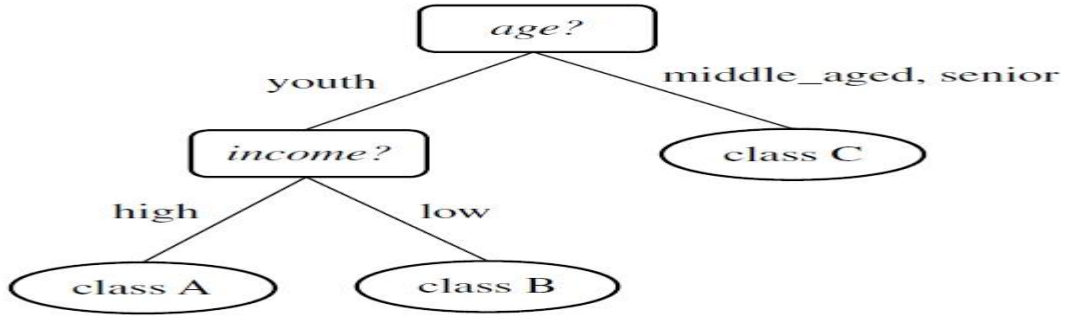
Classification:

Classification is a data mining technique that categorizes items in a collection based on some predefined properties. It uses methods like IF-THEN, Decision trees or Neural networks to predict a class or essentially classify a collection of items.

Classification is a supervised learning technique used to categorize data into predefined classes or labels.

Example:

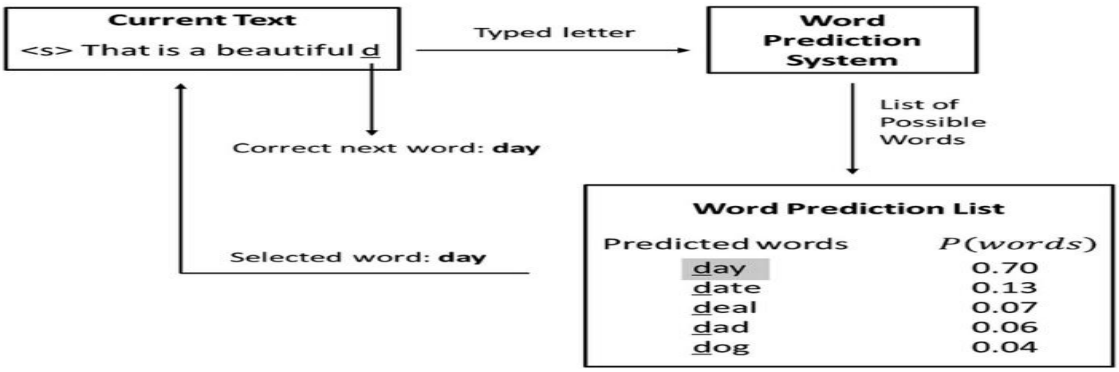
$$age(X, \text{“youth”}) \; AND \; income(X, \text{“high”}) \longrightarrow class(X, \text{“A”})$$
$$age(X, \text{“youth”}) \; AND \; income(X, \text{“low”}) \longrightarrow class(X, \text{“B”})$$
$$age(X, \text{“middle\_aged”}) \longrightarrow class(X, \text{“C”})$$
$$age(X, \text{“senior”}) \longrightarrow class(X, \text{“C”})$$



Prediction

Finding missing data in a database is very important for the accuracy of the analysis. Prediction is one of the data mining functionalities that help the analyst find the missing numeric values. If there is a missing class label, then this function is done using classification. It is very important in business intelligence and is very popular. One of the methods is to predict the missing or unavailable data using prediction analysis.

Example:



Association Analysis

Association Analysis is a functionality of data mining. It relates two or more attributes of the data. It discovers the relationship between the data and the rules that are binding them. It is also known as Market Basket Analysis for its wide use in retail sales.

The suggestion that Amazon shows on the bottom, "Customers who bought this also bought." is a real-time example of association analysis.

It relates two transactions of similar items and finds out the probability of the same happening again. This helps the companies improve their sales of various items.

Transaction Data:

TID	Items
T1	Mobile, HeadPhones, Router
T2	Mobile, Screenguard, Cable
T3	Mobile, Screenguard, Backcover, HDD
T4	Mobile, Powerbank, Mouse
T5	Mobile, Screenguard, Pendrive

Frequent Itemset:
{ Mobile, Screenguard }

Association Rule:
Mobile ⟶ Screenguard

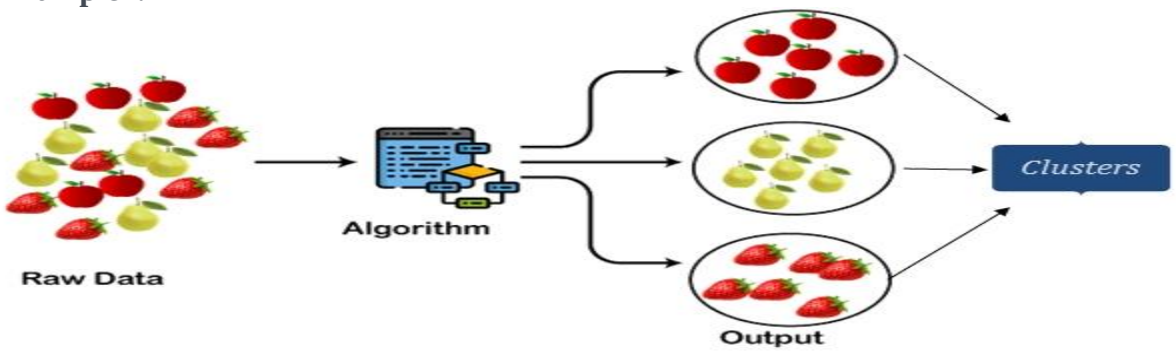
Cluster Analysis

Clustering is an unsupervised learning technique that group's similar data points together based on their features. The goal is to identify underlying structures or patterns in the data. Some common clustering algorithms include K-means, hierarchical clustering, and DBSCAN.

This data mining functionality is similar to classification. But in this case, the class label is unknown.Similar objects are grouped in a cluster. There are vast differences between one cluster and another.

Example1:

Example2:

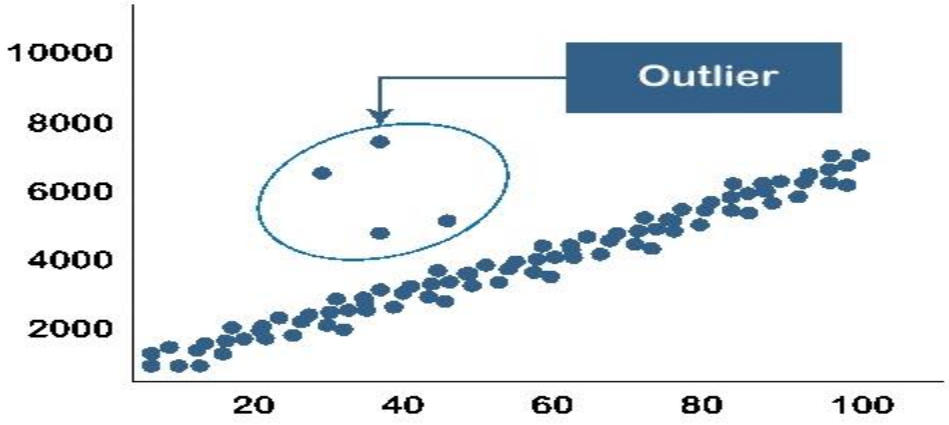


Outlier Analysis

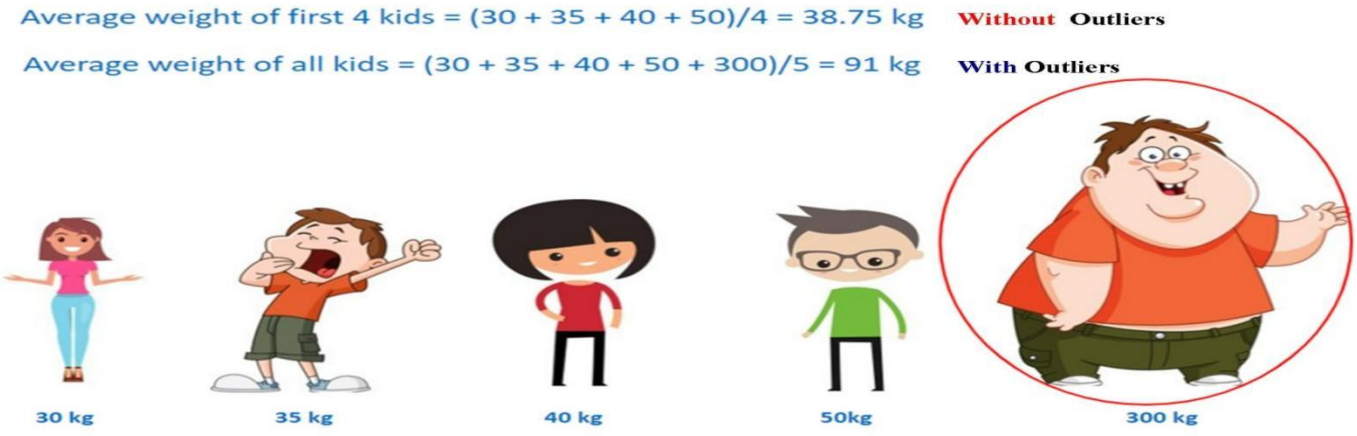
When data that cannot be grouped in any of the class appears, we use outlier analysis. There will be occurrences of data that will have different attributes/features to any of the other classes or clusters. These outstanding data are called outliers. They are usually considered noise or exceptions, and the analysis of these outliers is called outlier mining.

Outlier analysis is important to understand the quality of data. If there are too many outliers, you cannot trust the data or draw patterns out of it.

Example1:



Example2:



Interestingness Patterns

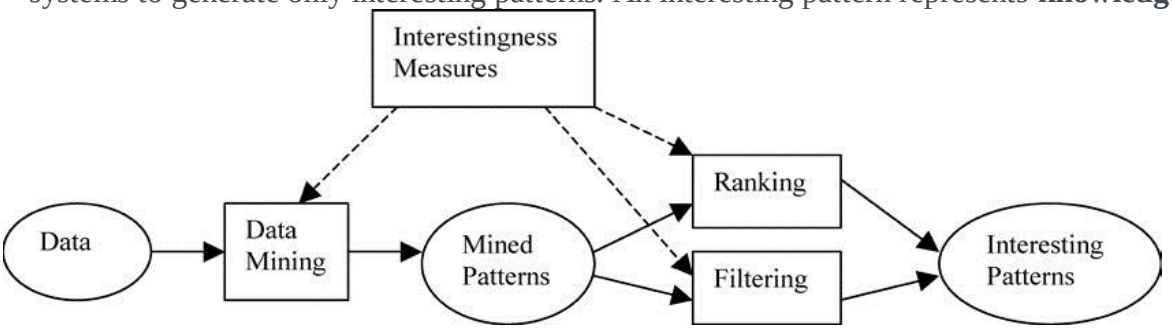
A data mining system has the potential to generate thousands or even millions of patterns, or rules. then **"are all of the patterns interesting?"** Typically, **not**—only a small fraction of the patterns potentially generated would be of interest to any given user.
This raises some serious questions for data mining. You may wonder,

1. What makes a pattern interesting?
2. Can a data mining system generate all the interesting patterns?
3. Can a data mining system generate only interesting patterns?

To answer the **first question**, a pattern is **interesting** if it is

1. easily understood by humans,
2. valid on new or test data with some degree of certainty,
3. potentially useful, and
4. Novel.

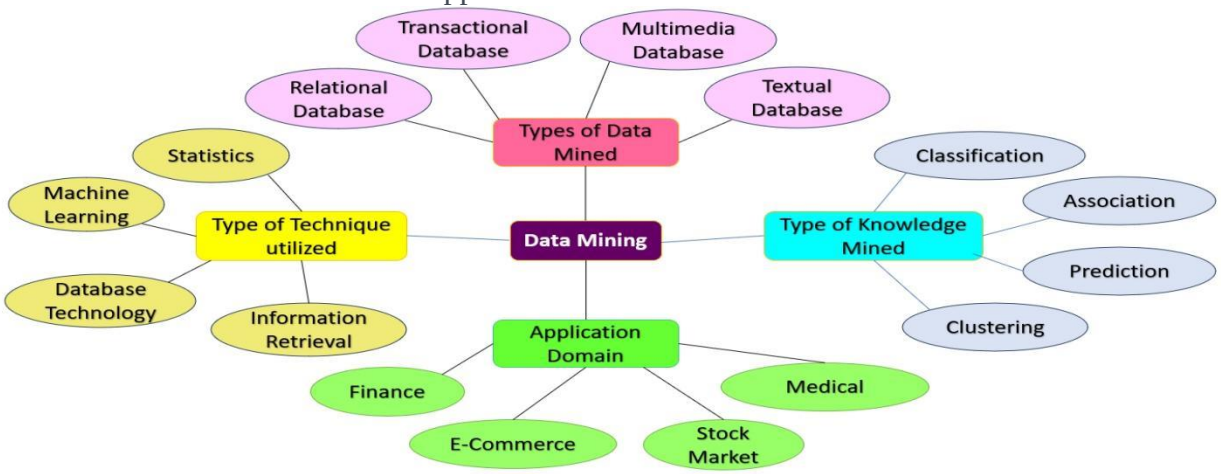
The second question**—Can a data mining system generate all the interesting patterns?**--refers to the completeness of a data mining algorithm. It is often unrealistic and inefficient for data mining systems to generate all the possible patterns. Instead, user-provided constraints and interesting measures should be used to focus the search. A data mining algorithm is complete if it mines all interesting patterns.
Finally, the third question -- **"Can a data mining system generate only interesting patterns?"**— is an optimization problem in data mining. It is highly desirable for data mining systems to generate only interesting patterns. An interesting pattern represents **knowledge**.



Classification of Data Mining systems

Data Mining is considered as an interdisciplinary field. It includes a set of various disciplines such as statistics, database systems, machine learning, visualization, and information sciences. Classification of the data mining system helps users to understand the system and match their requirements with such systems.
Data mining discovers patterns and extracts useful information from large datasets. Organizations need to analyze and interpret data using data mining systems as data grows rapidly. With an exponential increase in data, active data analysis is necessary to make sense of it all.
Data mining (DM) systems can be classified based on various factors.

- Classification based on Types of Data Mined
- Classification based on Type of knowledge Mined
- Classification based on Type of Technique Utilized
- Classification based on Application Domain



Classification based on Types of Data Mined

A database mining system can be classified based on 'type of data' or 'use of data' model or 'application of data.'

For Example: Relational Database, Transactional Database, Multimedia Database, Textual Data, World Wide Web (WWW) and etc,

Classification based on Type of knowledge Mined:

We can classify a data mining system according to the kind of knowledge mined. It means the data mining system is classified based on functionalities such as

- Association Analysis
- Classification
- Prediction
- Cluster Analysis
- Characterization
- Discrimination

Classification based on Type of Technique Utilized

We can classify a data mining system according to the kind of techniques used. We can describe these techniques according to the degree of user interaction involved or the methods of analysis employed.

Data Mining systems use various techniques, including Statistics, Machine Learning, Database Systems, Information retrieval, Visualization, and pattern recognition.

Classification based on Application Domain

We can classify a data mining system according to the applications adapted. These applications are as follows

- Finance
- Telecommunications

- E-Commerce
- Medial Sector
- Stock Markets

Data mining Task primitives

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during the mining process to discover interesting patterns.

Here is the list of Data Mining Task Primitives

- Set of task relevant data to be mined.
- Kind of knowledge to be mined.
- Background knowledge to be used in discovery process.
- Interestingness measures and thresholds for pattern evaluation.
- Representation for visualizing the discovered patterns.

Set of task relevant data to be mined
This specifies the portions of the database or the set of data in which the user is interested. This portion includes the following

- Database Attributes
- Data Warehouse dimensions of interest

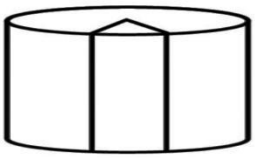
For example, suppose that you are a manager of All Electronics in charge of sales in the United States and Canada. You would like to study the buying trends of customers in Canada. Rather than mining on the entire database. These are referred to as relevant attributes.

Kind of knowledge to be mined

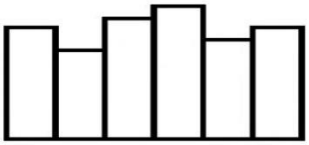
This specifies the data mining functions to be performed, such as

- Characterization& Discrimination
- Association
- Classification
- Clustering
- Prediction
- Outlier analysis

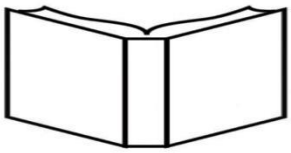
For instance, if studying the buying habits of customers in Canada, you may choose to mine associations between customer profiles and the items that these customers like to buy.



Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions



Knowledge type to be mined
Characterization & Discrimination
Association
Classification
prediction
Clustering



Background knowledge
Concept hierarchies
User beliefs about relationships in the data

Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty

Visualization of discovered patterns
Rules, tables, reports,
charts, graphs,
decision trees, and cubes

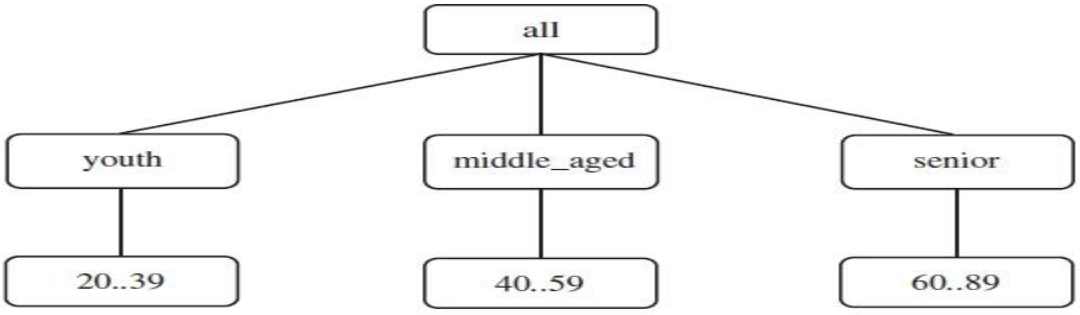
Background knowledge to be used in discovery process

Users can specify background knowledge, or knowledge about the domain to be mined. This knowledge is usefulfor guiding the knowledge discovery process, and for evaluating the patterns found. User beliefs about relationship in the data.

There are several kinds of background knowledge. Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.

Example:

An example of a concept hierarchy for the attribute (or dimension) age is shown in the following Figure.



In the above, the root node represents the mostgeneral abstraction level, denoted as all.

Interestingness measures and thresholds for pattern evaluation

The Interestingness measures are used to separateinteresting and uninteresting patterns from the knowledge.They may be used to guide the mining process, or after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures.

For **example**, interesting measures for association rules include support and confidence.

Representation for visualizing the discovered patterns

This refers to the formin which discovered patterns are to be displayed. Users can choose from different forms for knowledge presentation, such as

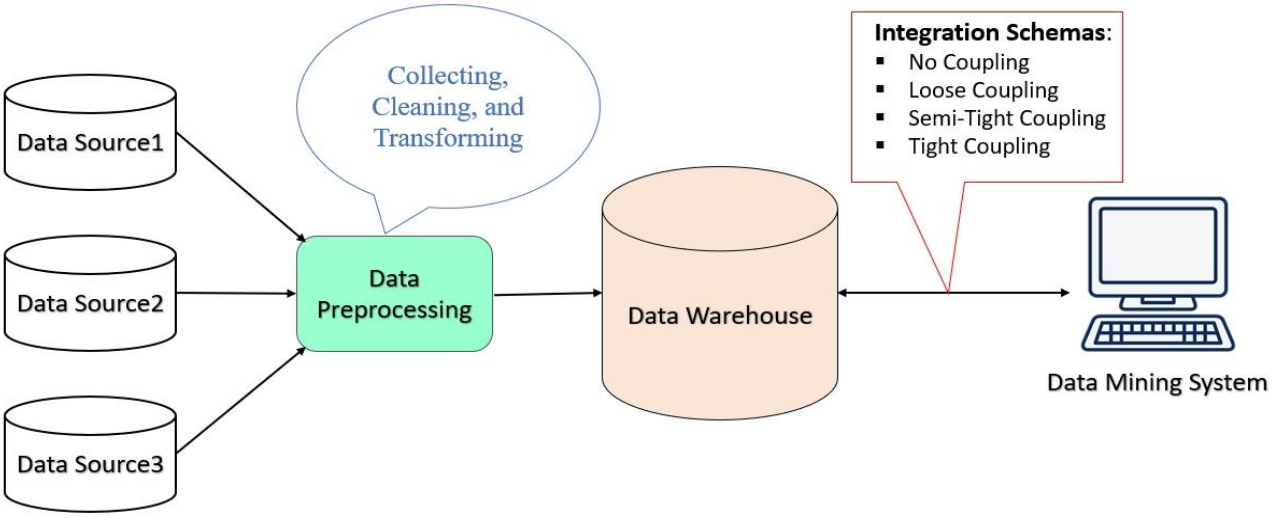
- Rules, tables, reports, charts, graphs, decision trees, and cubes.

Integration of Data mining system with a Data warehouse

The data mining system is **integrated** with a database or data warehouse system so that it can do its **tasks in an effective mode**. A data mining system operates in an environment that needs to **communicate** with other data systems like a Database or Datawarehouse system.

There are differentpossible integration (coupling) schemes as follows:

- No Coupling
- Loose Coupling
- Semi-Tight Coupling
- Tight Coupling



No Coupling

No coupling means that a Data Mining system will **not utilize any function** of a Data Base or Data Warehouse system.

It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file.

Drawbacks of No Coupling

- First, without using a Database/Data Warehouse system, a Data Mining system may spend a substantial amount of time finding, collecting, cleaning, and transforming data.
- Second, there are many tested, scalable algorithms and data structures implemented in Database and Data Warehouse systems.

Loose Coupling

In this loose coupling, the data mining system uses **some facilities / services** of a database or data warehouse system. The data is fetched from a data repository managed by these (DB/DW) systems.

Data mining approaches are used to process the data and then the processed data is saved either in a file or in a designated area in a database or data warehouse.

Loose coupling is better than no coupling because it can fetch any portion of data stored in Databases or Data Warehouses by using query processing, indexing, and other system facilities.

Drawbacks of Loose Coupling

- It is difficult for loose coupling to achieve high scalability and good performance with large data sets.

Semi-Tight Coupling

Semi tight coupling means that besides linking a Data Mining system to a Data Base/Data Warehouse system, efficient implementations of a few essential data mining primitives can be provided in the DB/DW system. These primitives can include sorting, indexing, aggregation, histogram analysis, multi way join, and precomputation of some essential statistical measures, such as sum, count, max, min, and standard deviation.

Advantage of Semi-Tight Coupling

- This Coupling will enhance the performance of Data Mining systems

Tight Coupling

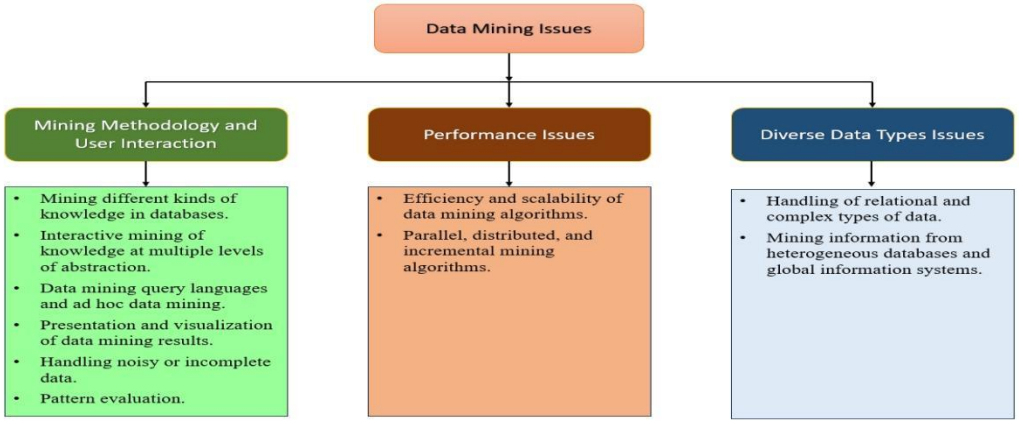
Tight coupling means that a Data Mining system is **smoothly integrated** into the Data Base/Data Warehouse system. The data mining subsystem is treated as **one functional component** of information system. Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of a DB or DW system.

Major issues in Data Mining

Data mining, the process of extracting knowledge from data, has become increasingly important as the amount of data generated by individuals, organizations, and machines has grown exponentially. Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources.

The above factors may lead to some issues in data mining. These issues are mainly divided into three categories, which are given below:

1. **Mining Methodology and User Interaction**
2. **Performance Issues**
3. **Diverse Data Types Issues**



Mining Methodology and User Interaction

It refers to the following kinds of issues

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore, it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion.

The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kinds of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore, mining the knowledge from them adds challenges to data mining.

Data Preprocessing

What is Data Preprocessing?

Data preprocessing is a crucial step in data mining. It involves transforming raw data into a clean, structured, and suitable format for mining. Proper data preprocessing helps improve the quality of the data, enhances the performance of algorithms, and ensures more accurate and reliable results.

Why Preprocess the Data?

In the real world, many databases and data warehouses have **noisy**, **missing**, and **inconsistent** data due to their huge size. Low quality data leads to low quality data mining.

Noisy: Containing errors or outliers. **E.g., Salary = "-10"**

Noisy data may come from

- Human or computer error at data entry.
- Errors in data transmission.

Missing:lacking certain attribute values or containing only aggregate data.**E.g.,Occupation = ""**

Missing (Incomplete) may data come from

- "Not applicable" data value when collected.
- Human/hardware/software problems.

Inconsistent: Data inconsistency meaning is that different versions of the same data appear in different places. For **example**, the ZIP code is saved in one table as **1234-567** numeric data format; while in another table it may be represented in **1234567**.

Inconsistent data may come from

- Errors in data entry.
- Merging data from different sources with varying formats.
- Differences in the data collection process.

Data preprocessing is used to improve the quality of data and mining results. And The goal of data preprocessing is to enhance the accuracy, efficiency, and reliability of data mining algorithms.

Major Tasks in Data Preprocessing

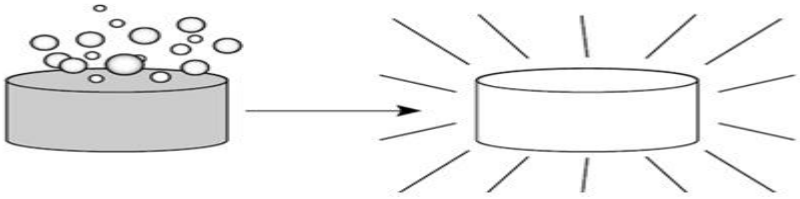
Data preprocessing is an essential step in the knowledge discovery process, because quality decisions must be based on quality data. And Data Preprocessing involves Data Cleaning, Data Integration, Data Reduction and Data Transformation.

Steps in Data Preprocessing

1. Data Cleaning

Data cleaning is a process that "cleans" the data by filling in the missing values, smoothing noisy data, analyzing, and removing outliers, and removing inconsistencies in the data.
If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied.

Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or datacleansing) routines attempt to fill in missing values, smooth out noise while identifyingoutliers, and correct inconsistencies in the data.



Missing Values

Imagine that you need to analyze All Electronics sales and customer data. You note that many tuples have no recorded value for several attributes such as customer income. How can you go about filling in the missing values for this attribute? There are several methods to fill the missing values.
Those are,

1. **Ignore the tuple:** This is usually done when the class label is missing (classification). This method is not very effective, unless the tuple contains several attributes with missing values.
2. **Fill in the missing value manually:** In general, this approach is time consuming and may not be feasible given a large data set with many missing values.
3. **Use a global constant to fill in the missing value:** Replace all missing attribute values by the same constant such as a label like “Unknown” or “- ∞ “.
4. **Use the attribute mean or median to fill in the missing value:** Replace all missing values in the attribute by the mean or median of that attribute values.

Noisy Data:

Noise is a random error or variance in a measured variable. Data smoothing techniques are used to eliminate noise and extract the useful patterns. The different techniques used for data smoothing are:

1.

Binning: Binning methods smooth a sorted data value by consulting its “neighborhood,” that is, the values around it. The sorted values are distributed into several “buckets,” or bins. Because binning methods consult the neighborhood of values, they perform local

smoothing.
There are three kinds of binning. They are:

2.
   - Smoothing by Bin Means: In this method, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9.
   - Smoothing by Bin Medians: In this method, each value in a bin is replaced by the median value of the bin. For example, the median of the values 4, 8, and 15 in Bin 1 is 8. Therefore, each original value in this bin is replaced by the value 8.
   - Smoothing by Bin Boundaries: In this method, the minimum and maximum values in each bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. For example, the middle value of the values 4, 8, and 15 in Bin 1is replaced with nearest boundary i.e., 4.

Example:

Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34
Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34
Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29
Smoothing by bin medians:
Bin 1: 8, 8, 8
Bin 2: 21, 21, 21
Bin 3: 28, 28, 28
Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

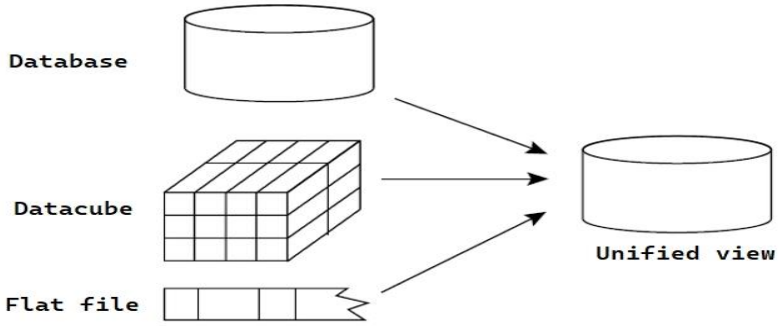
3. Regression: Data smoothing can also be done by regression, a technique that used to predict the numeric values in a given data set. It analyses the relationship between a target variable (dependent) and its predictor variable (independent).
   - Regression is a form of a supervised machine learning technique that tries to predict any continuous valued attribute.
   - Regression done in two ways; Linear regression involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.
4. Clustering: It supports in identifying the outliers. The similar values are organized into clusters and those values which fall outside the cluster are known as outliers.

2. Data Integration

Data integration is the process of combining data from multiple sources into a single, unified view. This process involves identifying and accessing the different data sources, mapping the data to a common format. Different data sources may include multiple data cubes, databases, or flat files.

The goal of data integration is to make it easier to access and analyze data that is spread across multiple systems or platforms, in order to gain a more complete and accurate understanding of the data.

Data integration strategy is typically described using a triple (G, S, M) approach, where G denotes the global schema, S denotes the schema of the heterogeneous data sources, and M represents the mapping between the queries of the source and global schema.



Example: To understand the (G, S, M) approach, let us consider a data integration scenario that aims to combine employee data from two different HR databases, database A and database B. The global schema (G) would define the unified view of employee data, including attributes like EmployeeID, Name, Department, and Salary.

In the schema of heterogeneous sources, database A (S1) might have attributes like EmpID, FullName, Dept, and Pay, while database B's schema (S2) might have attributes like ID, EmployeeName, DepartmentName, and Wage. The mappings (M) would then define how the attributes in S1 and S2 map to the attributes in G, allowing for the integration of employee data from both systems into the global schema.

Issues in Data Integration

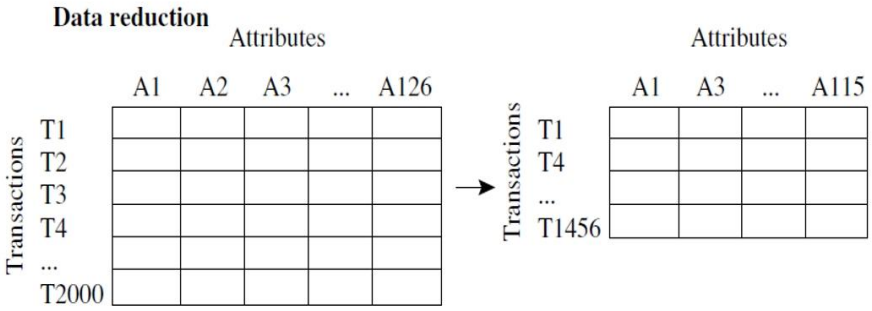
There are several issues that can arise when integrating data from multiple sources, including:

1. Data Quality: Data from different sources may have varying levels of accuracy, completeness, and consistency, which can lead to data quality issues in the integrated data.
2. Data Semantics: Integrating data from different sources can be challenging because the same data element may have different meanings across sources.
3. Data Heterogeneity: Different sources may use different data formats, structures, or schemas, making it difficult to combine and analyze the data.

3. Data Reduction

Imagine that you have selected data from the All Electronics data warehouse for analysis. The data set will likely be huge! Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.
Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

In simple words, Data reduction is a technique used in data mining to reduce the size of a dateset while still preserving the most important information. This can be beneficial in situations where the date set is too large to be processed efficiently, or where the dateset contains a large amount of irrelevant or redundant information.

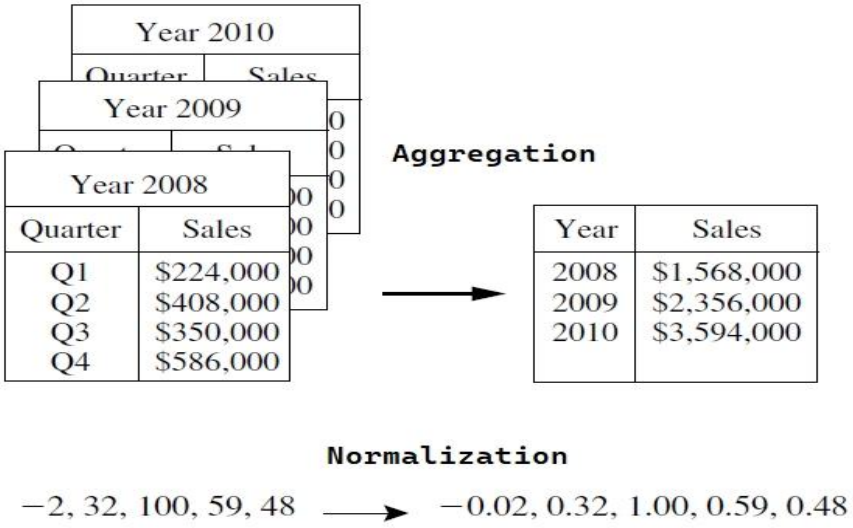


There are several different data reduction techniques that can be used in data mining, including:

1. Data Sampling: This technique involves selecting a subset of the data to work with, rather than using the entire dataset. This can be useful for reducing the size of a dataset while still preserving the overall trends and patterns in the data.
2. Dimensionality Reduction: This technique involves reducing the number of features in the dataset, either by removing features that are not relevant or by combining multiple features into a single feature.
3. Data compression: This is the process of altering, encoding, or transforming the structure of data in order to save space. By reducing duplication and encoding data in binary form, data compression creates a compact representation of information. And it involves the techniques such as loss or lossless compression to reduce the size of a dataset.

4. Data Transformation

Data transformation in data mining refers to the process of converting raw data into a format that is suitable for analysis and modeling. The goal of data transformation is to prepare the data for data mining so that it can be used to extract useful insights and knowledge.



Data transformation typically involves several steps, including:

1. **Smoothing:** It is a process that is used to remove noise from the dataset using techniques includes binning, regression, and clustering.
2. **Attribute construction (or feature construction):** In this, new attributes are constructed and added from the given set of attributes to help the mining process.
3. **Aggregation:** In this, summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated to compute monthly and annual total amounts.
4. **Data normalization:** This process involves converting all data variables into a small range. Such as -1.0 to 1.0, or 0.0 to 1.0.
5. **Generalization:** It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example, Age initially in Numerical form (22,) is converted into categorical value (young, old).

Method Name	Irregularity	Output
Data Cleaning	Missing, Noise, and Inconsistent data	Quality Data before Integration
Data Integration	Different data sources (data cubes, databases, or flat files)	Unified view
Data Reduction	Huge amounts of data can take a long time, making such analysis impractical or infeasible.	Reduce the size of a dataset and maintains the integrity.
Data Transformation	Raw data	Prepare the data for data mining