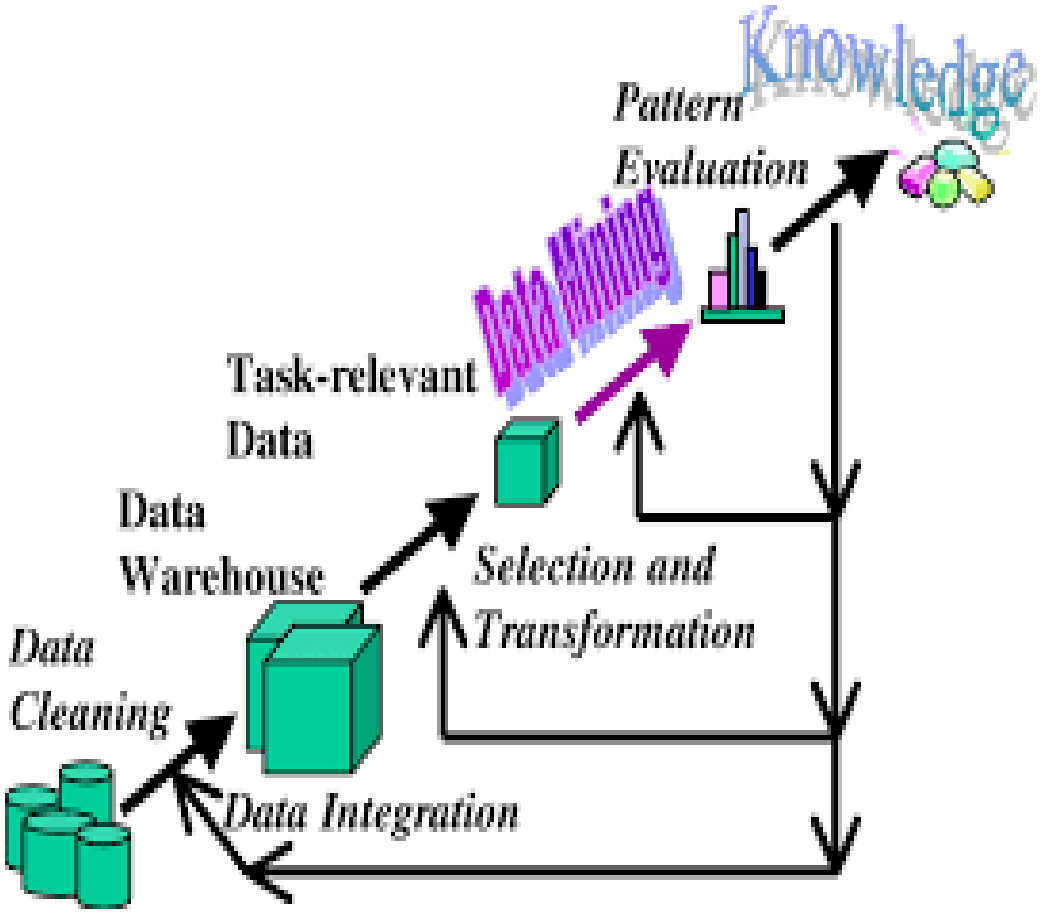
Data mining

- **What is Data mining?**

➢ **Data mining** is defined as a process used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research.

➢ **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.



- The knowledge discovery process is shown in above diagram as an iterative sequence of the following steps:

1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)
3. **Data selection** (where data relevant to the analysis task are retrieved from the database)
4. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

- **Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining.** The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

- Data mining is highly useful in the following domains −

  - Market Analysis and Management
  - Corporate Analysis & Risk Management
  - Fraud Detection

- **What Kinds of Data Can Be Mined?**

- Data mining can be applied to any kind of data as long as the data are meaningful for a target application.

- The most basic forms of data for mining applications are **database data, data warehouse data, and transactional data.** Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW).

1. Database data

- 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. A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.

- Relational data can be accessed by database queries written in a relational query language (e.g., SQL) or with the assistance of graphical user interfaces.

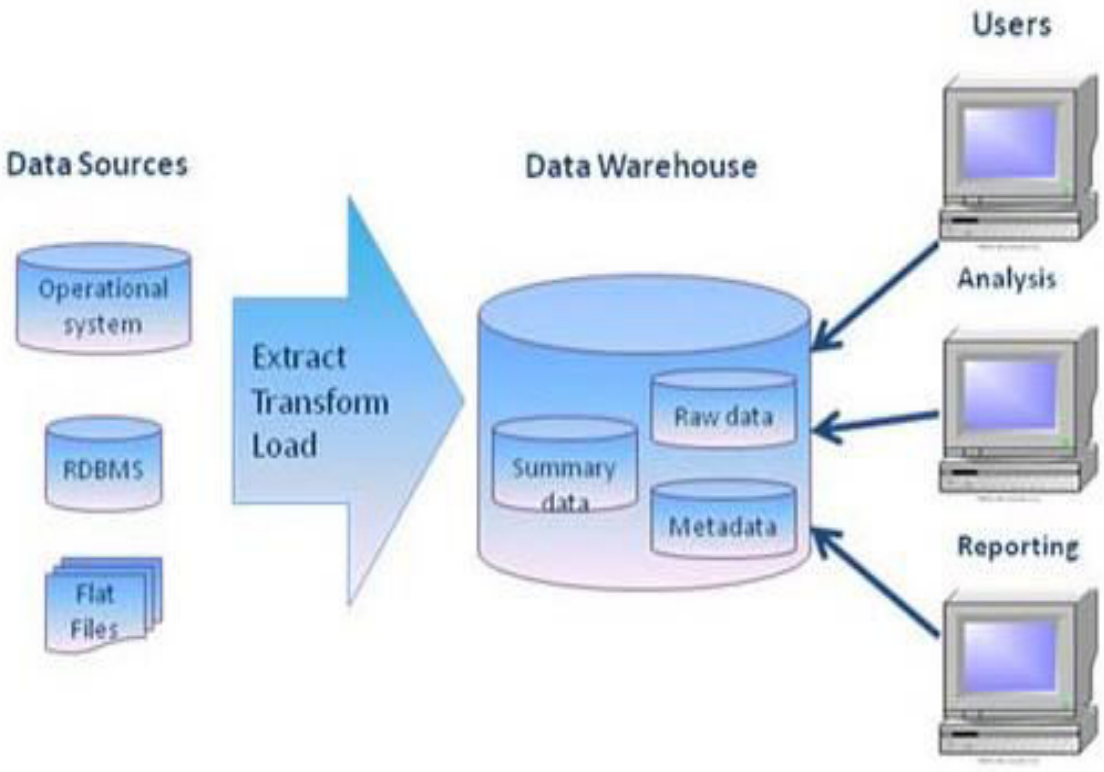
- When mining relational databases, we can go further by searching for trends or data patterns. **For example**, data mining

systems can analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information. Data mining systems may also detect deviations—that is, items with sales that are far from those expected in comparison with the previous year.

2. Data Warehouses

- 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.



Example or Applications of Data Warehousing

- Data Warehousing can be applicable anywhere where we have huge amount of data and we want to see statistical results that help in decision making.

- **Social Media Websites**: The social networking websites like Facebook, Twitter etc. are based on analyzing large data sets. These sites gather data related to members, groups, locations etc. and store it in a single central repository. Being large amount of data, Data Warehouse is needed for implementing the same.

- **Banking**: Most of the banks these days use warehouses to see spending patterns of account/card holders. They use this to provide them special offers, deals, etc.

- **Government**: Government uses data warehouse to store and analyze tax payment which is used to detect tax thefts.

- There can be many more applications in different sectors like E-Commerce, Telecommunication, Transportation Services, Marketing and Distribution, Healthcare and Retail.

DBMS vs Data Warehouse

S.No.	Database	Data Warehouse
1.	A common Database is based on operational or transactional processing. Each operation is an indivisible transaction.	A Data Warehouse is based on analytical processing.
2.	Generally, a Database stores current and up-to-date data which is used for daily operations.	A Data Warehouse maintains historical data over time. Historical data is the data kept over years and can be used for trend analysis, make future predictions and decision support.
3.	A database is generally application specific. Example - A database stores related data, such as the student details in a school.	A Data Warehouse is integrated generally at the organization level, by combining data from different databases. Example - A data warehouse integrates the data from one or more databases, so that analysis can be done to get results, such as the best performing school in a city.
4.	Constructing a Database is not so expensive.	Constructing a Data Warehouse can be expensive.

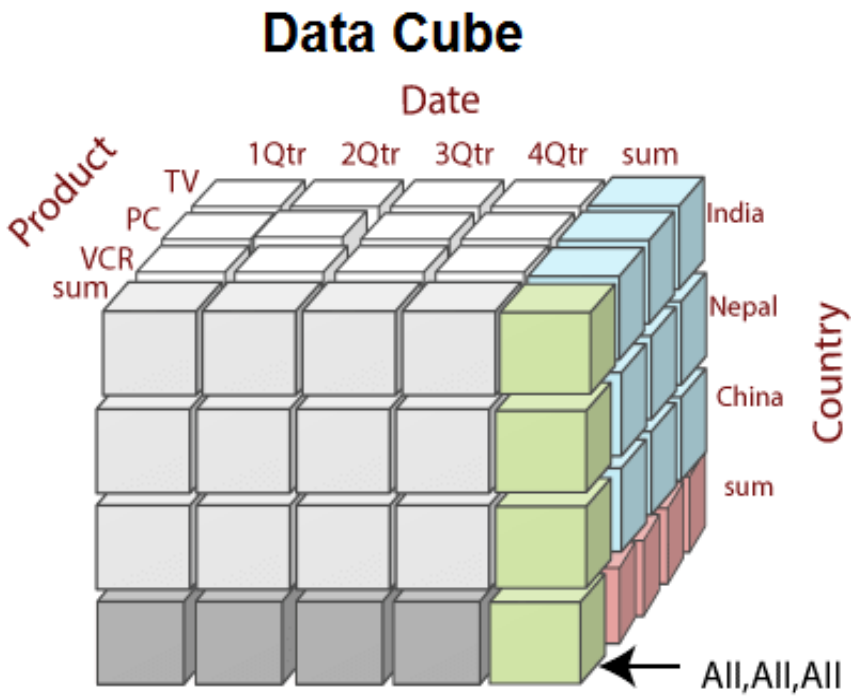
Data Cube

- **A data warehouse is usually modeled by a multidimensional data structure, called a data cube.**

- Data cube is a multi-dimensional structure. Data cube is a data abstraction to view aggregated data from a number of perspectives. The dimensions are aggregated as the 'measure' attribute, as the remaining dimensions are known as the 'feature' attributes.

- A data cube provides a multidimensional view of data and allows the pre computation and fast access of summarized data.

- Data cubes are commonly used for easy interpretation of data. It is used to represent data along with dimensions as some measures of business needs. Each dimension of the cube represents some attribute of the database, e.g Sales per day, month or year.



- By providing multidimensional data views and the precomputation of summarized data, data warehouse systems can provide inherent support for OLAP (Online analytical processing)
- The OLAP cube is a data structure optimized for very quick data analysis.

How does it work?

- A Data warehouse would extract information from multiple data sources and formats like text files, excel sheet, multimedia files, etc.

- The extracted data is cleaned and transformed. Data is loaded into an OLAP server (or OLAP cube) where information is pre-calculated in advance for further analysis.

Basic analytical operations of OLAP

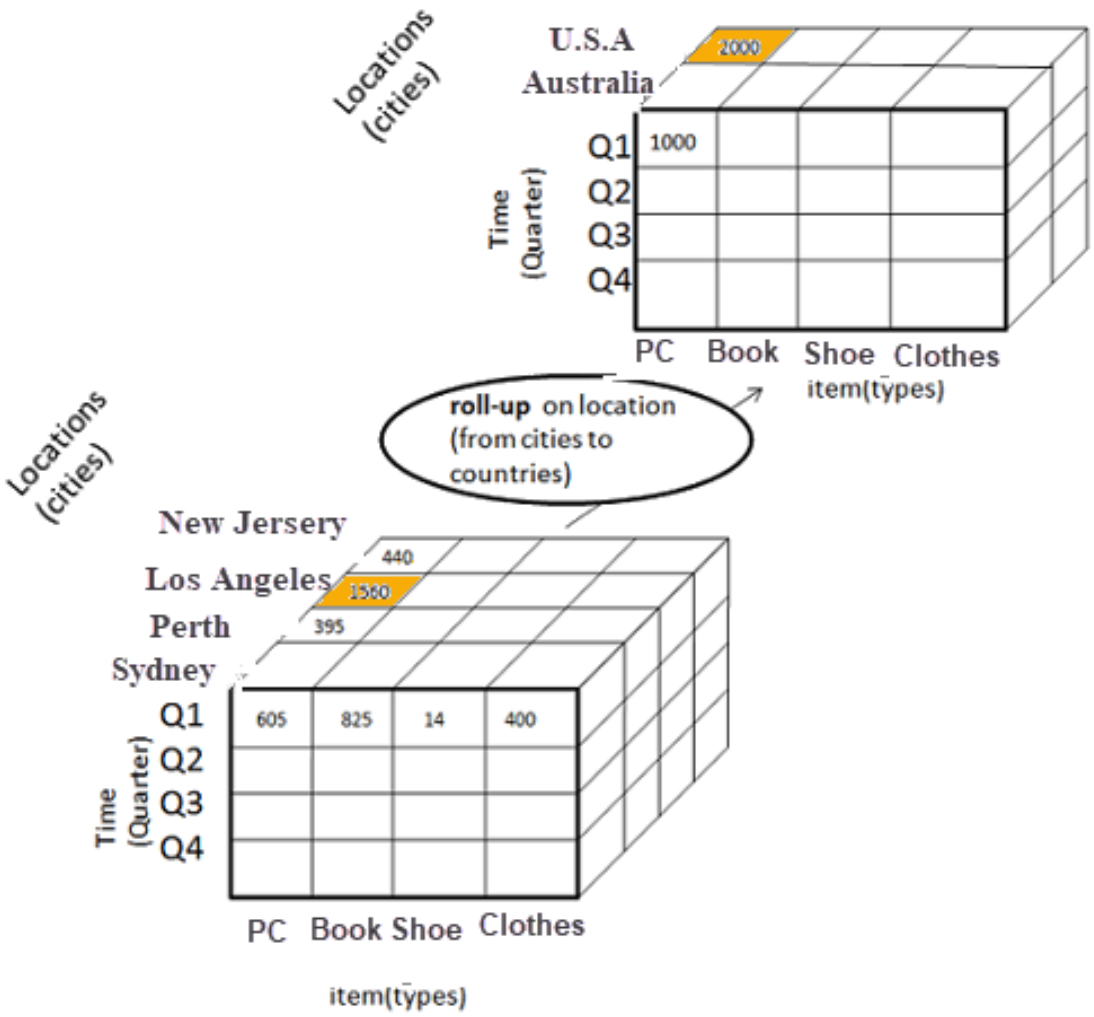
1. Roll-up
2. Drill-down

1) Roll-up:

- Roll-up is also known as "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways

  1. Reducing dimensions
  2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.

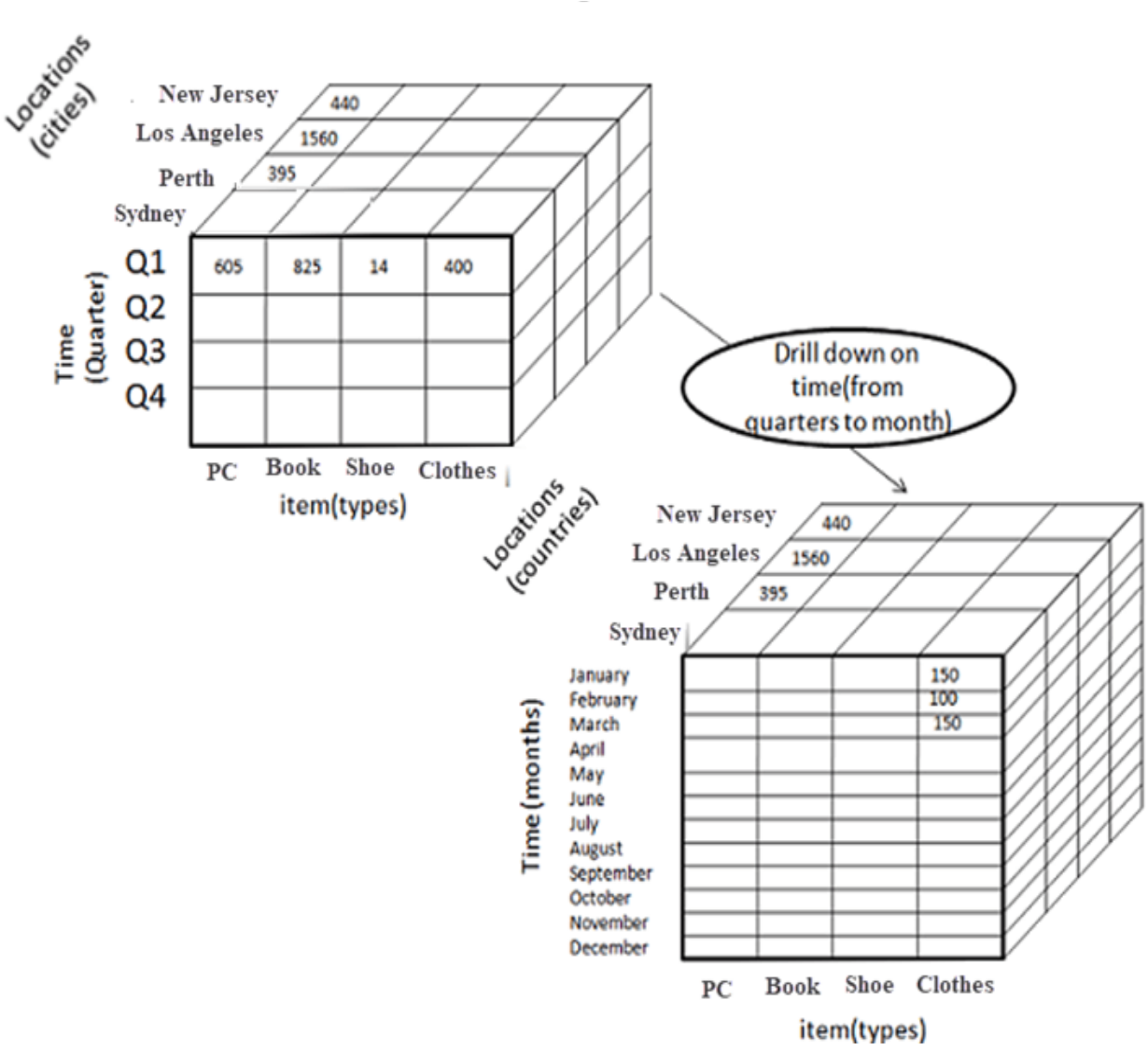
Consider the following diagram



- In this example, cities New jersey and Lost Angles and rolled up into country USA
- The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up
- In this aggregation process, data is location hierarchy moves up from city to the country.
- In the roll-up process at least one or more dimensions need to be removed. In this example, Quater dimension is removed.

2) Drill-down

- In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via
    - Moving down the concept hierarchy
    - Increasing a dimension



- In this example Quarter Q1 is drilled down to months January, February, and March. Corresponding sales are also registers.
- In this example, dimension months are added.

3. Transactional Data

- 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 For example, if one customer purchases multiple products at different times, a transaction record needs to be created for each sale, but the data about the customer stays the same.

- As an analyst of electronic product company, you may ask,"Which items sold well together?" This kind of market basket data analysis would enable you to bundle groups of items together as a strategy for boosting sales.

- For example, given the knowledge that printers are commonly purchased together with computers, you could offer certain printers at a steep discount (or even for free) to customers buying selected computers, in the hopes of selling more computers

4. Other Kinds of Data

- There are many other kinds of data that have versatile forms and structures and rather different semantic meanings.

- Such kinds of data can be seen in many applications: time-related or sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data), data streams (e.g., video surveillance and sensor data, which are continuously transmitted), spatial data (e.g., maps), engineering design data

(e.g., the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), graph and networked data (e.g., social and information networks), and the Web (a huge, widely distributed information repository made available by the Internet).

Examples

1) We can mine banking data for changing trends, which may aid in the scheduling of bank tellers according to the volume of customer traffic.
2) Stock exchange data can be mined to uncover trends that could help you plan investment strategies.
3) By mining text data, such as literature on data mining from the past ten years, we can identify the evolution of hot topics in the field.
4) By mining user comments on products (which are often submitted as short text messages), we can assess customer sentiments and understand how well a product is embraced by a market.
5) From multimedia data, we can mine images to identify objects and classify them by assigning semantic labels or tags. By mining video data of a hockey game, we can detect video sequences corresponding to goals.

- **What Kinds of Patterns Can Be Mined?**

- There are a number of data mining functionalities. These include characterization and discrimination, **the mining of frequent patterns, associations, and correlations, classification and regression, clustering analysis and outlier analysis.**

- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

- In general, **such tasks can be classified into two categories: descriptive and predictive**. **Descriptive mining** tasks characterize properties of the data in a target data set. **Predictive mining** tasks perform induction on the current data in order to make predictions.

1) Class/Concept Description: Characterization and Discrimination

- Class/Concept refers to the data to be associated with the classes or concepts. **For example**, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions.

- **These descriptions can be derived by the following two ways**

  - **Data Characterization**

    - This refers to summarizing data of class under study. This class under study is called as Target Class.

    - **For example**, to study the characteristics of software products with sales that increased by 10% in the previous year

- A customer relationship manager at Electronics product shop may order the following data mining task: **Summarize the characteristics of customers** who spend more than $5000 a year at shop. The data mining system should allow the customer relationship manager to drill down on any dimension, such as on occupation to view these customers according to their type of employment.

- **The output of data characterization can be presented in various forms**. Examples include pie charts, bar charts, curves, multi dimensional data cubes, and multidimensional tables

- **Data Discrimination**

- It refers to the mapping or classification of a class with some predefined group or class.
  - **For example**, A customer relationship manager at any Electronics shop 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).

2) Mining Frequent Patterns, Associations, and Correlations

1) Mining of Frequent Patterns

- Frequent patterns are those patterns that occur frequently in transactional data.
- Here is the list of kind of frequent patterns −

  Frequent Item Set − It refers to a set of items that frequently appear together, for example, milk and bread which are frequently bought together in grocery stores by many customers

- **Frequent Subsequence** − A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- **Frequent Sub Structure** − Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

2) Mining of Association

- Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.
- **For example**, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.
- **For example** Association analysis. Suppose that, as a marketing manager at AllElectronics, you want to know which items are frequently purchased together (i.e., within the same transaction).
- An example of such a rule, mined from the AllElectronics transactional database, is
- buys.X, “computer”/)buys.X, “software”/ [support D 1%, confidence D 50%],where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that he/she will buy software as well. A 1% support means that 1% of all the transactions under analysis show that computer and software are purchased together.

3) Mining of Correlations

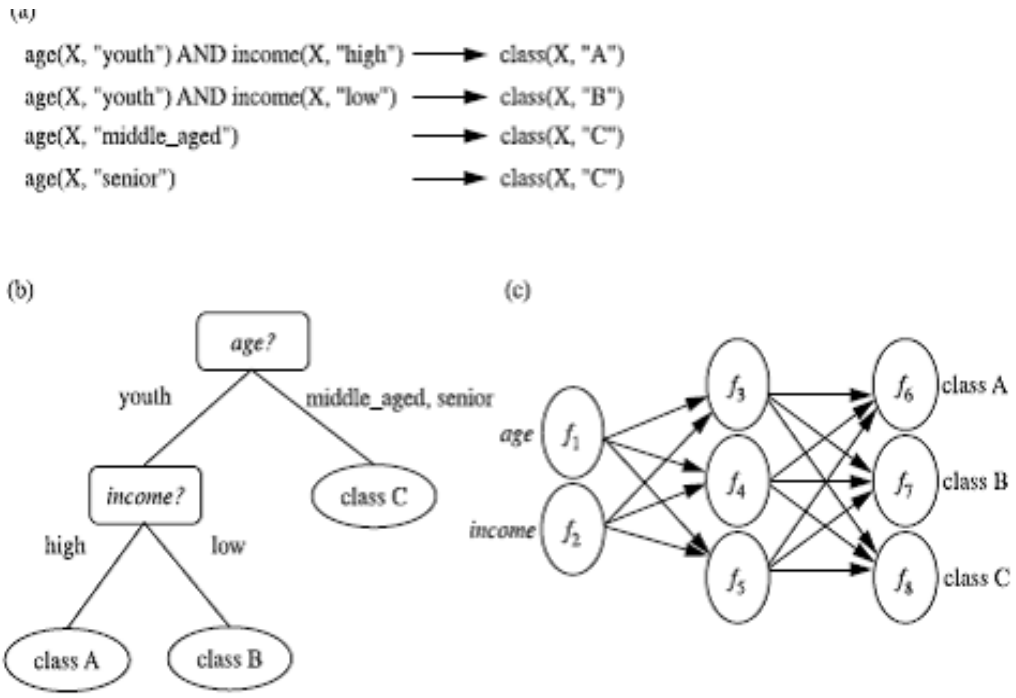
- It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute

value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

4) Classification and Regression for Predictive Analysis

- **Classification is the process of finding a model that describes the data classes or concepts**.
- The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms −

  - Classification (IF-THEN) Rules
  - Decision Trees
  - Mathematical Formulae
  - Neural Networks

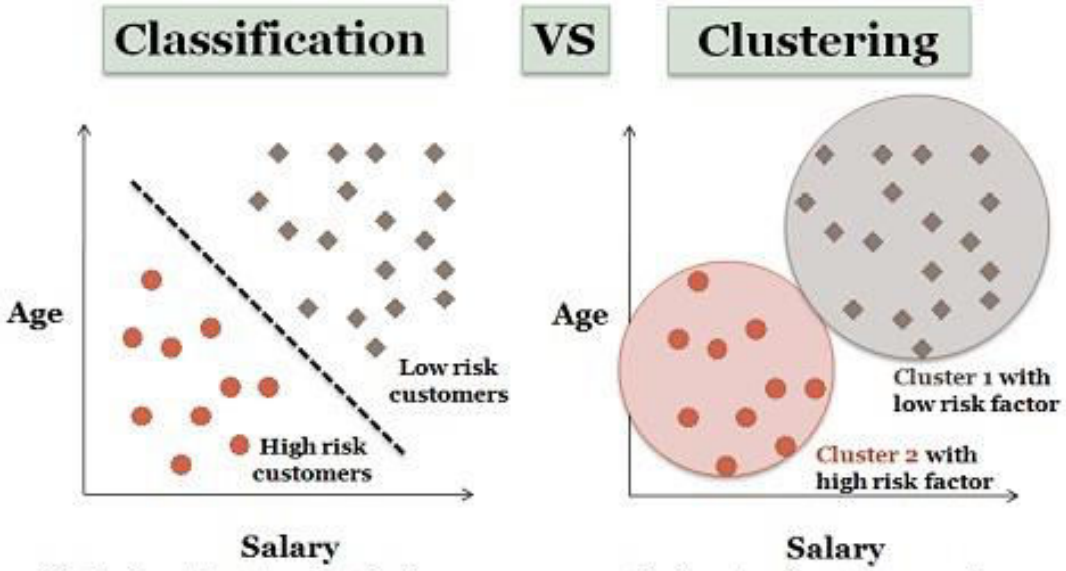


A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

- **The list of functions involved in these processes are as follows**
  - **Classification** − It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
  - **Prediction** − It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
  - **Outlier Analysis** − Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
  - **Evolution Analysis** − Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.
  - **Regression analysis** is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data.

5) Cluster Analysis

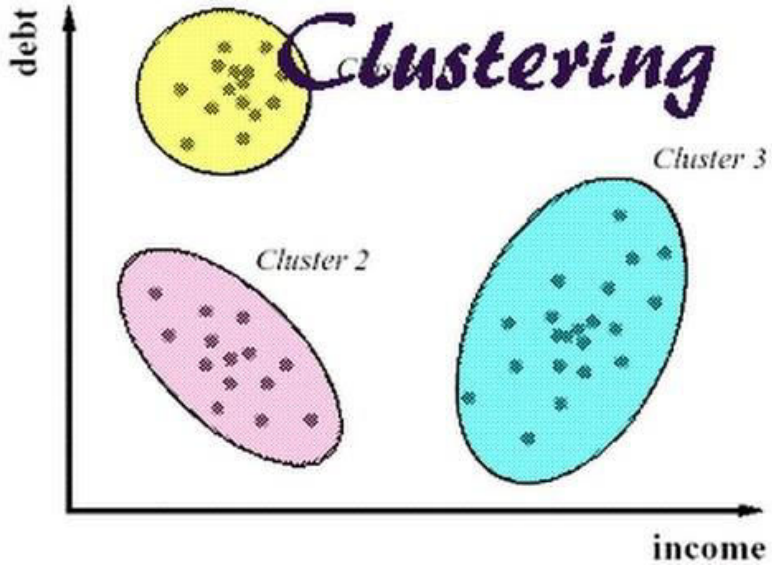
- Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.



Risk classification for the loan payees on the basis of customer salary

- **Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.**

- Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.



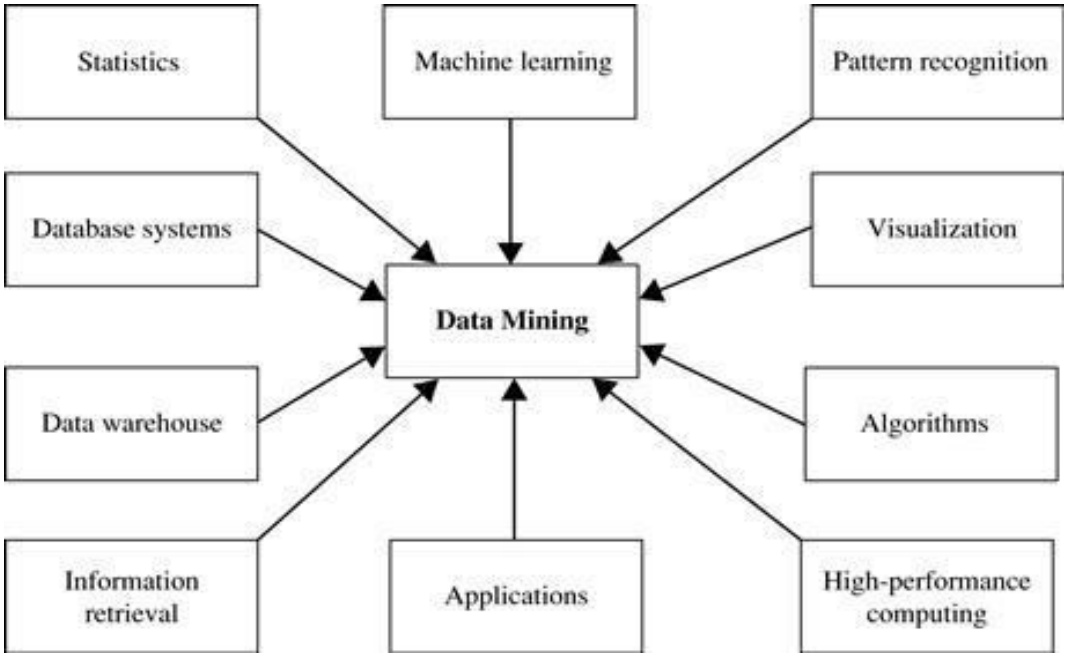
6) Outlier Analysis

- **A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers**.

- Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier analysis or anomaly mining.

▪ Which Technologies Are Used?

- Many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high performance computing, and many application domains as shown in diagram.



1. Statistics

- **Statistics studies the collection, analysis, interpretation or explanation, and presentation of data.** Data mining has an inherent connection with statistics.

- **Statistics is useful for mining various patterns from data as well as for understanding the underlying mechanisms generating and affecting the patterns.**

- A statistical model is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. For example, we can use statistics to model noise and missing data values.

- Statistics research develops tools for prediction and forecasting using data and statistical models. Statistical methods can be used to summarize or describe a collection of data.

- **Predictive statistics models** data in a way that accounts for randomness and uncertainty in the observations. Statistical methods can also be used to verify data mining results.

- **For example**, after a classification or prediction model is mined, the model should be verified by statistical **hypothesis testing**. **A statistical hypothesis test** makes statistical decisions using experimental data.

2. Machine Learning

- **Machine learning** investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data.

- **Examples**

  - **Virtual Personal Assistants** Siri, Alexa, Google Now are some of the popular examples of virtual personal assistants. As the name suggests, they assist in finding information, when asked over voice

  - **Predictions while Commuting**

Traffic Predictions: We all have been using GPS navigation services. While we do that, our current locations and velocities are being saved at a central server for managing traffic.
Online Transportation Networks: When booking a cab, the app estimates the price of the ride.

Videos Surveillance: The video surveillance system nowadays are powered by AI that makes it possible to detect crime before they happen.

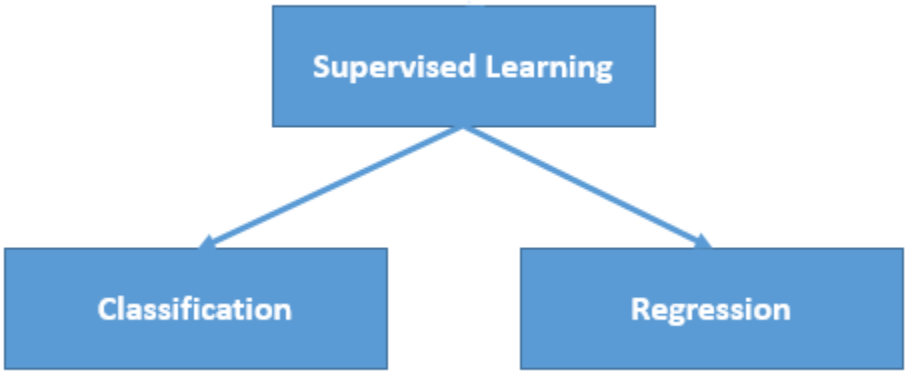
Other examples are **Face Detection**, **Credit Card Fraud Detection, spam Detection etc.**

Here, we illustrate **classic problems in machine learning** that are highly related to data mining.

1. Supervised learning

- Supervised learning is a type of ML where the model is provided with **labeled** training data

- **Types**



Classification

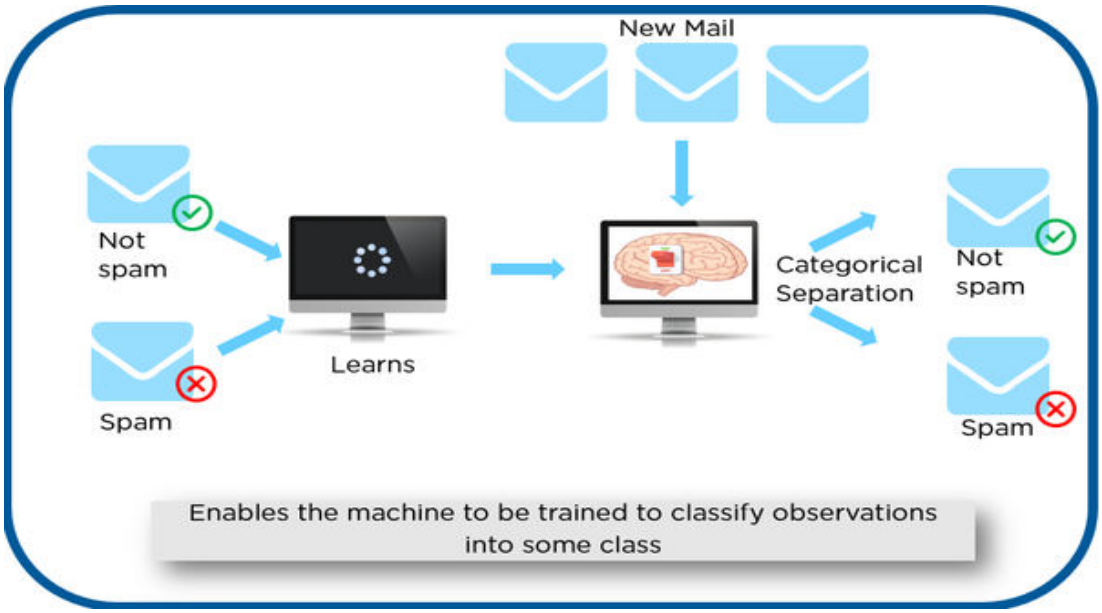
- Data is labeled meaning it is assigned a class, **for example** spam/non-spam or fraud/non-fraud. The decision being modeled is to assign labels to new unlabelled pieces of data. This can be thought of as a discrimination problem, modeling the differences or similarities between groups.

Regression

- Regression problems on the other hand are problems where we try to make a prediction on a continuous scale. Data is labelled with a real value (think floating point) rather then a label.

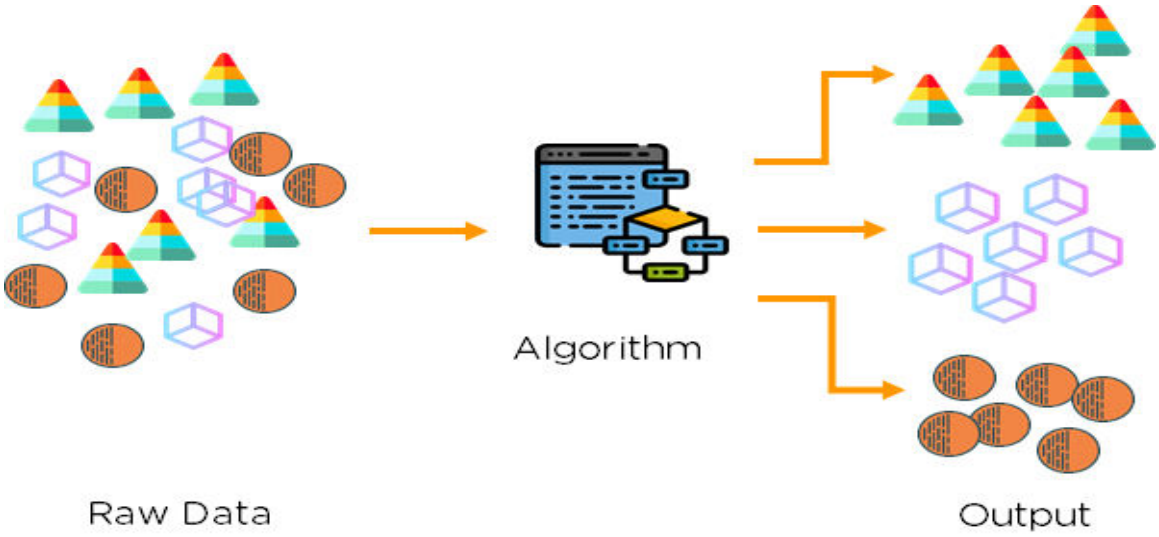
- **Example**

  Time series data like the price of a stock over time, the decision being modeled is what value to predict for new unpredicted data.

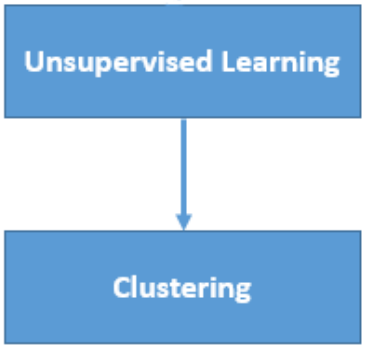


2. Unsupervised learning

- In unsupervised learning, the goal is to identify meaningful patterns in the data.



- Types



Clustering:

- This is a type of problem where we group similar things together.Data is not labelled, but can be divided into groups based on similarity and other measures of natural structure in the data.

- **Examples**

  - organizing pictures by faces without names
  - given news articles, cluster into different types of news
  - given a set of tweets ,cluster based on content of tweet
  - given a set of images, cluster them into different objects

3. Semi-supervised learning

- In this type of learning, the algorithm is trained upon a combination of labeled and unlabeled data.

- Practical applications of Semi-Supervised Learning

- **Speech Analysis**: Since labeling of audio files is a very intensive task, Semi-Supervised learning is a very natural approach to solve this problem.

- **Internet Content Classification**: Labeling each webpage is an impractical and unfeasible process and thus uses Semi-Supervised learning algorithms. Even the Google search algorithm uses a variant of Semi-Supervised learning to rank the relevance of a webpage for a given query.

- **Protein Sequence Classification:** Since DNA strands are typically very large in size, the rise of Semi-Supervised learning has been imminent in this field.

Common Example to all 3 techniques

one may imagine the three types of learning algorithms as Supervised learning where a student is under the supervision of a teacher at both home and school, Unsupervised learning where a student has to figure out a concept himself and Semi-Supervised learning where a teacher teaches a few concepts in class and gives questions as homework which are based on similar concepts.

4. Active learning

- Active Learning is a special case of Supervised Machine Learning. It is a machine learning approach that lets users play an active role in the learning process.

- In active learning, the algorithm gets a lot of data, but not the labels. The algorithm can then explicitly request labels to individual examples. This can be helpful when we have a large amount of unlabeled data

- **The goal of this iterative learning** approach is to speed along the learning process, especially if you don't have a large labeled dataset.

- Example you want to find faces in YouTube videos. In active learning, the algorithm tries to both learn the task and tell us what labels would be most useful at the current state. We can then label just those frames, so that the manual effort are reduced

3. Database Systems and Data Warehouses

4. Information Retrieval

- Information retrieval system is a network of algorithms, which facilitate the search of relevant data / documents as per the user requirement. Documents can be text or multimedia, and may reside on the Web.

- By integrating information retrieval models and data mining techniques, we can find the major topics in a collection of documents and, for each document in the collection, the major topics involved.

Major Issues in Data Mining

Data mining systems face a lot of challenges and issues in today's world some of them are:

1) Mining methodology and user interaction issues
2) Performance issues
3) Issues relating to the diversity of database types

1) Mining methodology and user interaction issues

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.
- **Incorporation of background knowledge** - To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **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.

2) 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. Then the results from the partitions are merged. The incremental algorithms, update databases without mining the data again from scratch.

3) Issues relating to the diversity of database types

- **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 kind 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.