DATA WARE HOUSE

&

DATA MINING

UNIT 1 PART 1

Yugshakti
Asst. Prof., IINTM

Sources of Data

- Businesses worldwide generate giant data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback.

- Scientific and engineering practices generate high orders of petabytes of data in a continuous manner, from remote sensing, process measuring, scientific experiments, system performance, engineering observations, and environment surveillance.

- Billions of Web searches supported by search engines process tens of petabytes of data daily. Communities and social media have become increasingly important data sources, producing digital pictures and videos, blogs, Web communities, and various kinds of social networks.

- The medical and health industry generates tremendous amounts of data from medical records, patient monitoring, and medical imaging.

- *The list of sources that generate huge amounts of data is endless.*

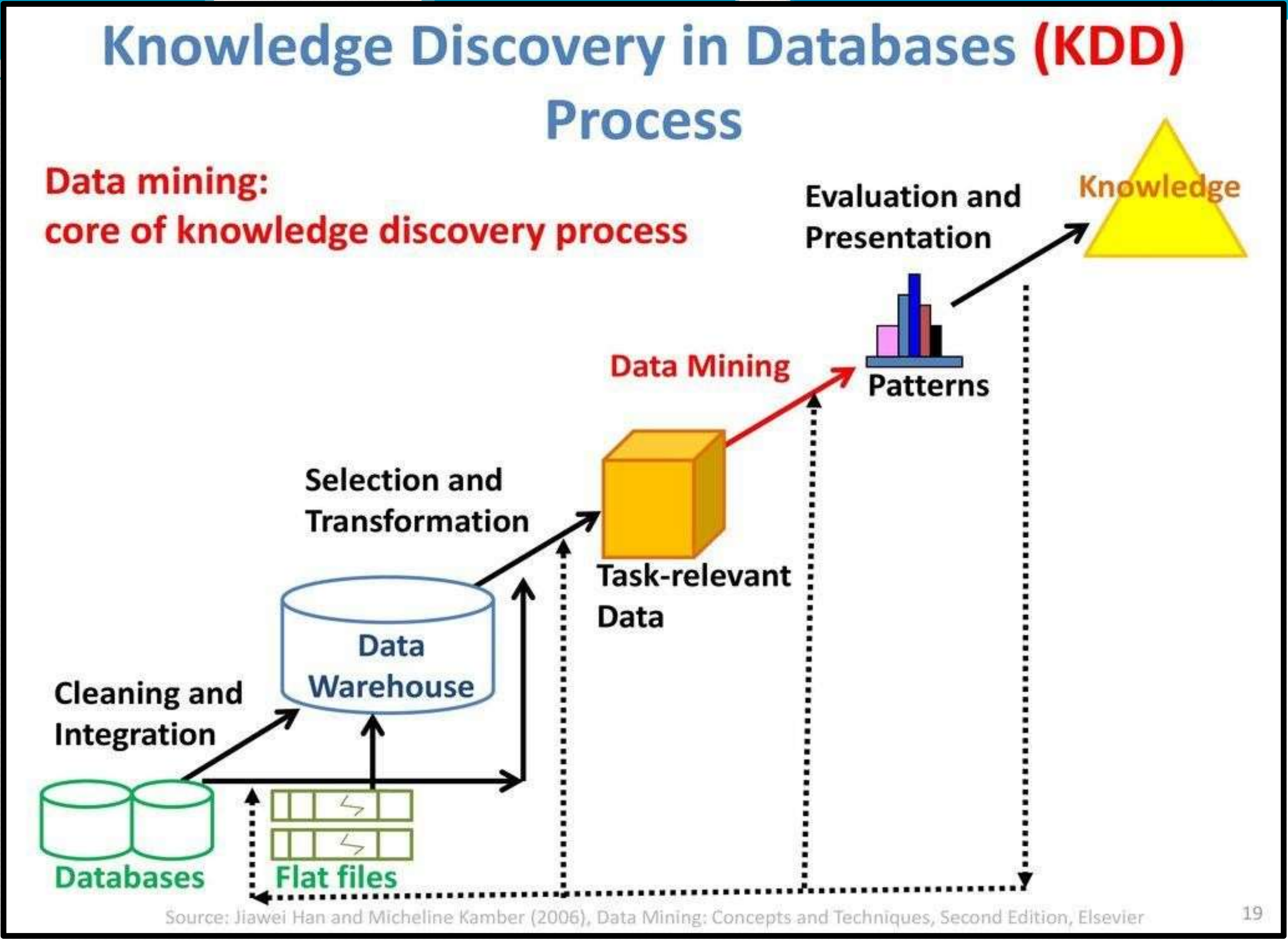
- Powerful and versatile tools are needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge.

- *This necessity has led to the birth of data mining.*

Knowledge Discovery Process(previous year paper)

- **Knowledge discovery concerns the entire knowledge extraction process,**
  - including how data are stored and accessed,
  - how to use efficient and scalable algorithms to analyze massive datasets,
  - how to interpret and visualize the results, and
  - how to model and support the interaction between human and machine.
- **It also concerns support for learning and analyzing the application domain.**

The knowledge discovery process is shown in Figure.



Knowledge Discovery in Databases (KDD) Process

Data mining:
core of knowledge discovery process

Evaluation and Presentation

Knowledge

Data Mining

Patterns

Selection and Transformation

Task-relevant Data

Data Warehouse

Cleaning and Integration

Databases

Flat files

Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

Steps in Knowledge Discovery Process

KDD is 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.

Contents:

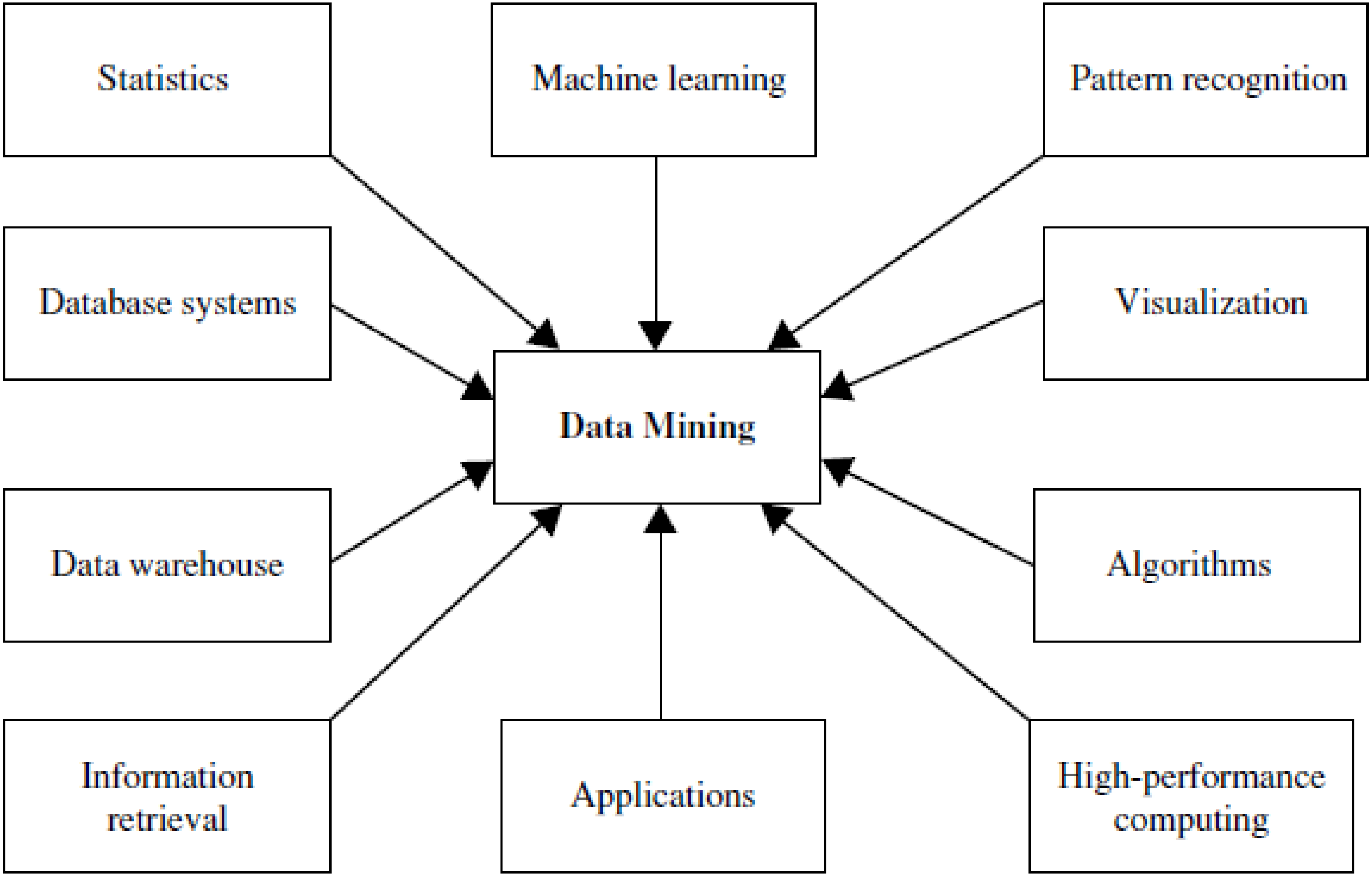
- Introduction to Data Mining
- Data Mining
  - On What Kind of Data
  - What kind of patterns to be mined
- Classification of data mining system
- Data Mining Task primitives
- Major issues in Data Mining

What is Data Mining

- **Data mining is the *process of discovering knowledge in form of interesting patterns and relationships from large amounts of data*.**
- “Data mining,” writes Joseph P. Bigus in his book, *Data Mining with Neural Networks,*
- ***It is the efficient discovery of valuable, non-obvious information from a large collection of data.”***
- Data mining centers around the automated discovery of new facts and relationships in data.

- **It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology**.

- With traditional query tools, you search for known information. Data mining tools enable you to uncover hidden information.



Data mining adopts techniques from many domains.

Difference between KDD and Data Mining

- Although, the two terms KDD and Data Mining are heavily used interchangeably, they refer to two related yet slightly different concepts. **KDD is the overall process of extracting knowledge from data while Data Mining is a step inside the KDD process, which deals with identifying patterns in data**. In other words, Data Mining is only the application of a specific algorithm based on the overall goal of the KDD process.

Classification of Data Mining Systems(previous year paper)

- A data mining system can be classified based on the kind of :

- (a) **Databases mined- what kind of data can be mined?**
  - database data
  - data warehouse data
  - transactional data
- (b)**Knowledge mined- What kind of patterns are extracted?**
  - the mining of frequent patterns, associations, and correlations
  - classification and regression
  - clustering analysis
  - outlier analysis
  - characterization and discrimination

- (c) **Applications adapted**
  - Finance
  - Retail and Telecommunication
  - Science and Engineering
  - Intrusion Detection and Prevention
  - Recommender System

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 most basic forms of data for mining applications are :
- database data
- data warehouse data
- transactional data

Database Data

- 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.

- Suppose that your job is to analyze the *AllElectronics store* data.
- Through the use of relational queries, you can ask things like:

- *"Show me a list of all items that were sold in the last quarter."*
- *"Show me the total sales of the last month, grouped by branch,"*
- *"How many sales transactions occurred in the month of December?"*
- *"Which salesperson had the highest sales?"*

- When **mining relational databases**, we can go further by *searching for trends* or *data patterns*. For example, data mining systems can analyse customer data to predict the credit risk of new customers based on their income, age, and previous credit information.

Data Warehouse

- A Data Warehouse is a repository of information collected from multiple sources, **stored under a unified schema,** and usually residing at a single site.

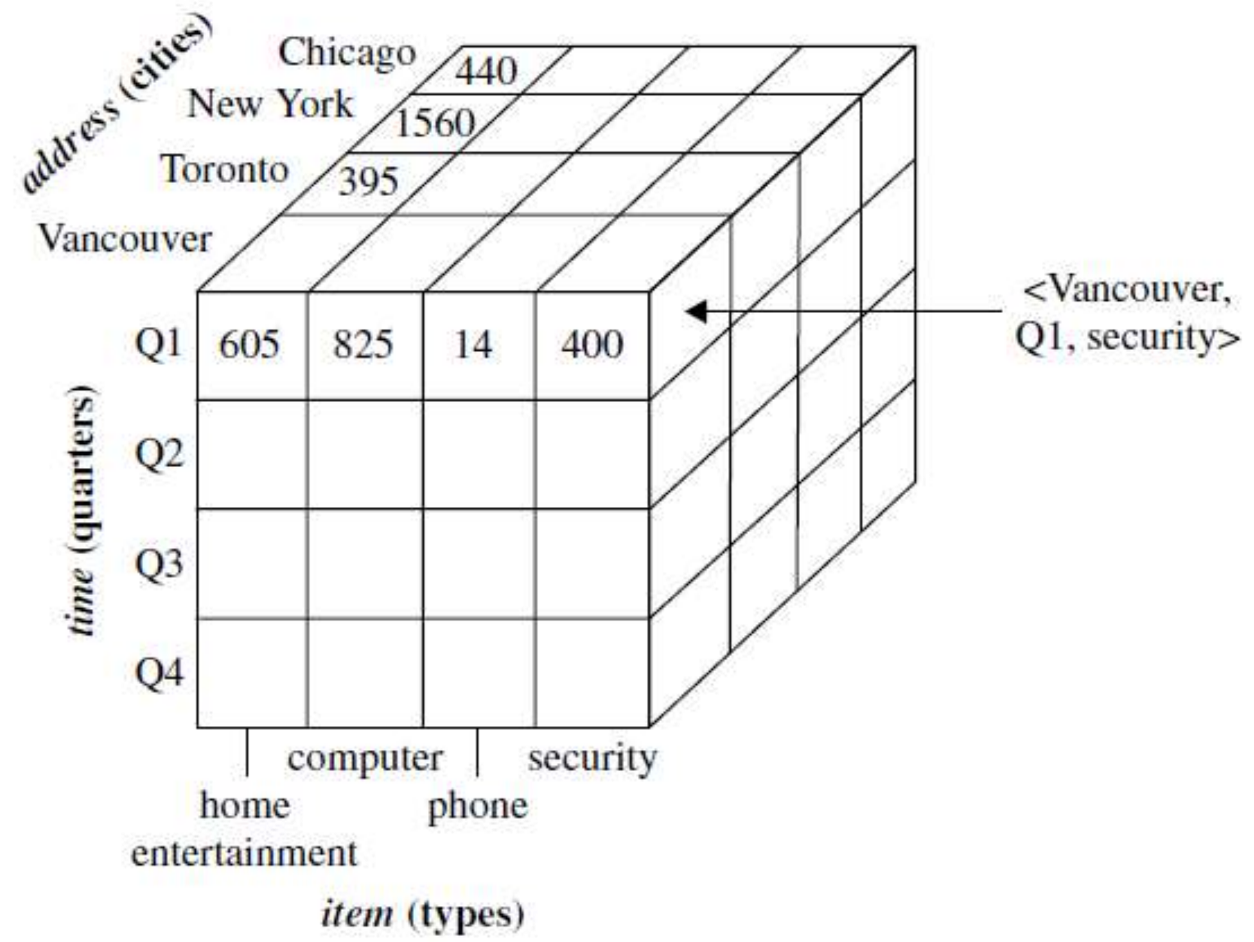
- Data warehouses are **constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.**

- The data are stored to provide information from a ***historical perspective***, such as in the past 6 to 12 months, and are typically ***summarized*.**

- For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item_type for each store or, summarized to a higher level, for each sales region.

- 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*(*sales_amount*).

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



It allows the exploration of multiple combinations of dimensions at **varying levels of granularity in data mining**, and thus has greater potential for discovering interesting patterns representing knowledge.

Transactional Data

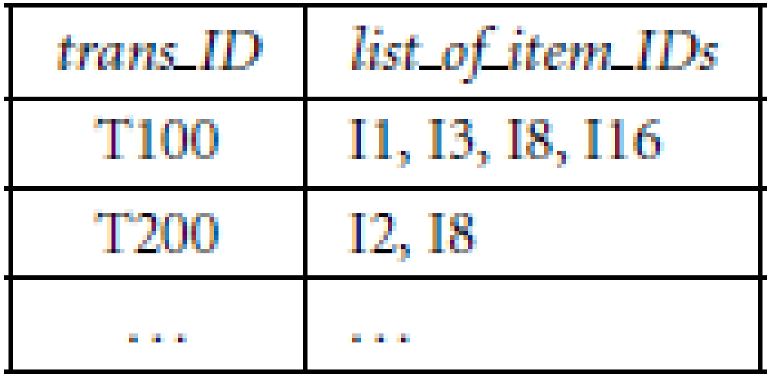
- Each record in a **transactional database** captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.

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

- Transactions can be stored in a table, with one record per transaction.

- Because most relational database systems do not support nested relational structures, the transactional database is usually stored in a flat file.

- Suppose you may want to know, “*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 discounted price.



trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Other Kinds of Data

- 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 accepted by a market.
- By mining **video data** of a hockey game, we can detect video sequences corresponding to goals.
- **Stock exchange data** can be mined to uncover trends that could help you plan investment strategies.
- With **spatial data**, we may look for patterns that describe changes in metropolitan poverty rates based on city distances from major highways.

What kind of patterns can be mined?

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

- 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**.

- **Predictive data mining** tasks come up with a model from the available data set that is helpful in predicting unknown or future values of another data set of interest. A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task.

- **Descriptive data mining** tasks usually finds data describing patterns and comes up with new, significant information from the available data set. A retailer trying to identify products that are purchased together can be considered as a descriptive data mining task

Mining frequent patterns, associations and correlations

- **Frequent patterns, as the name suggests, are patterns that occur frequently in data.**

- There are many kinds of frequent patterns, including frequent itemsets and frequent subsequences.

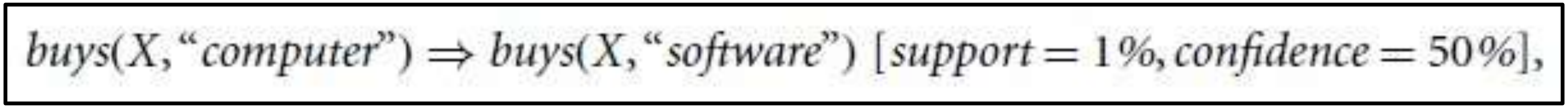
- A **frequent itemset** typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers.

- A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.

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*

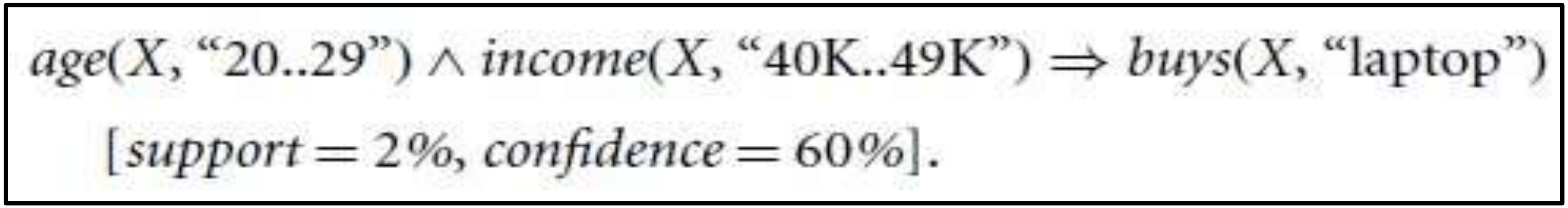


$buys(X, \text{“computer”}) \Rightarrow buys(X, \text{“software”})\ [support = 1\%, confidence = 50\%],$

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 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

The rule indicates that of the *AllElectronics customers under study, 2% are 20 to 29 years* old with an income of $40,000 to $49,000 and have purchased a laptop (computer)at *AllElectronics. There is a 60% probability that a customer in this age and income* group will purchase a laptop.



$$age(X, \text{“20..29”}) \wedge income(X, \text{“40K..49K”}) \Rightarrow buys(X, \text{“laptop”})$$

$$[support = 2\%, confidence = 60\%].$$

As it involves more than one attribute or predicate(age, income, buys), it is known as multidimensional association rule.

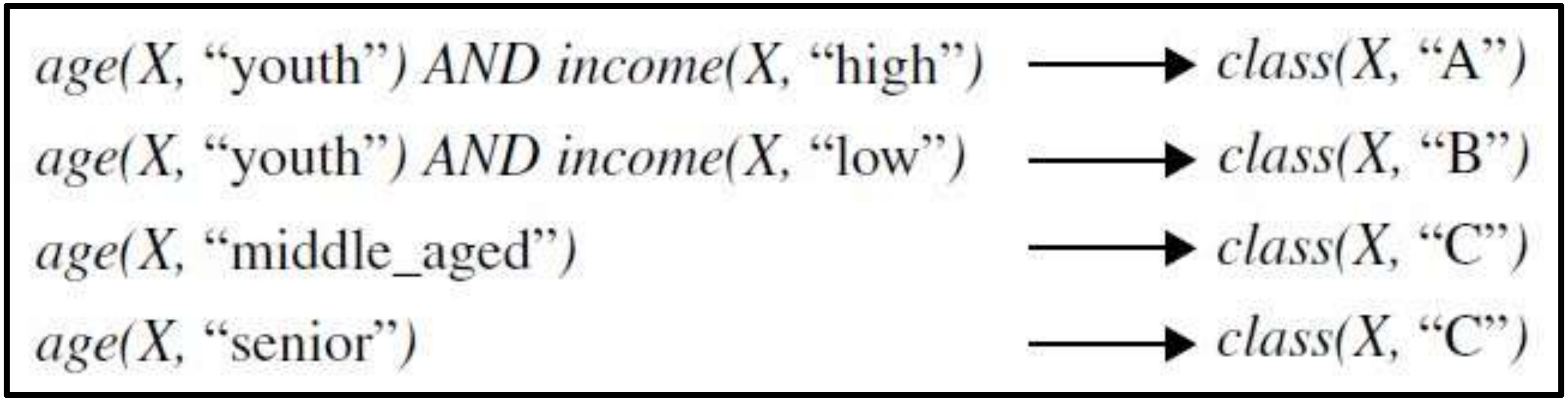
Typically, association rules are discarded as uninteresting if they do not satisfy both a **minimum support threshold and a minimum confidence threshold.**

Classification and Regression for Predictive Analysis

- **Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.**

- The model are derived based on the analysis of a set of **training data (i.e., data objects for which the class labels are known). The model is used** to predict the class label of objects for which the class label is unknown.

- *The derived model may be represented in various* forms, such as *classification rules (i.e., IF-THEN rules), decision trees.*



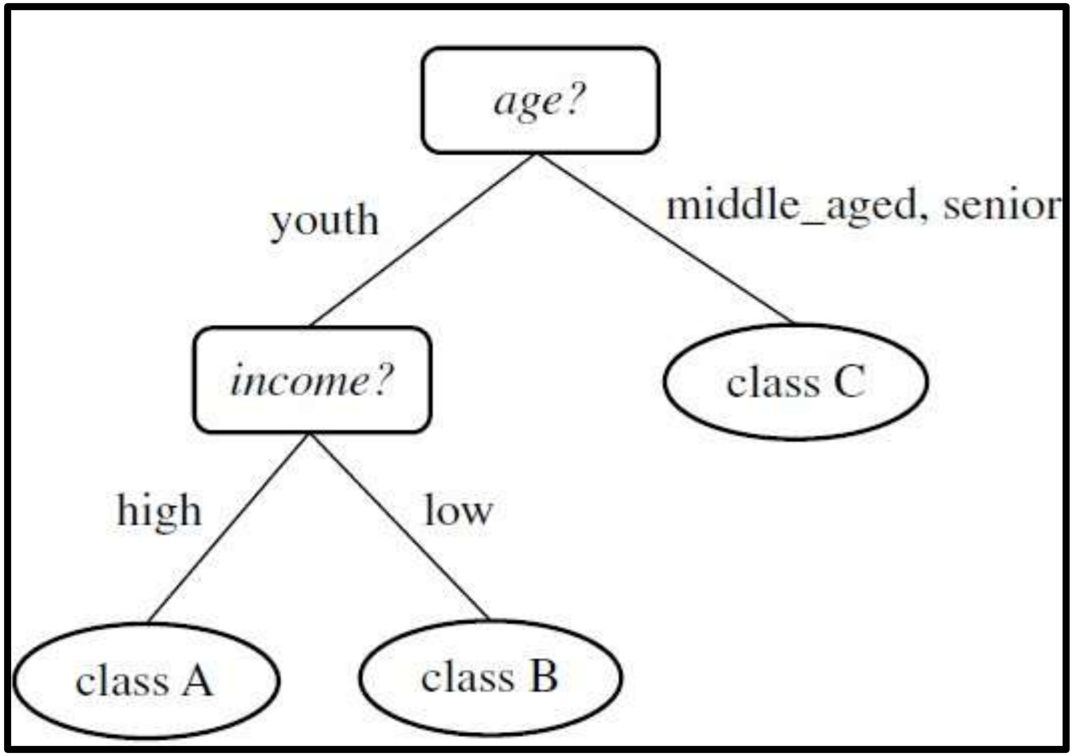
age(*X*, "youth") *AND income*(*X*, "high") → *class*(*X*, "A")

age(*X*, "youth") *AND income*(*X*, "low") → *class*(*X*, "B")

age(*X*, "middle_aged") → *class*(*X*, "C")

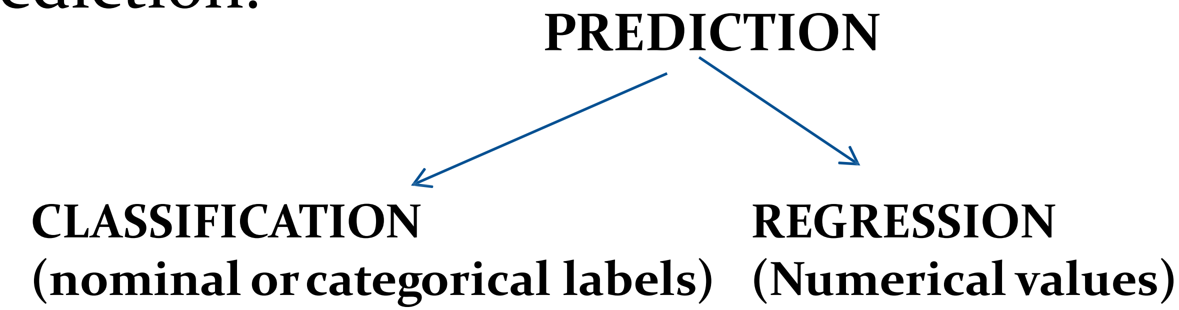
age(*X*, "senior") → *class*(*X*, "C")

- *A* ***decision tree is a flowchart-like tree structure,*** where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions



- Whereas **classification predicts categorical** (discrete, unordered) labels, **regression models continuous-valued functions**. That is, regression is used to predict missing or unavailable *numerical data values rather than (discrete) class labels.*

- *The term prediction* refers to both numeric prediction and class label prediction.



PREDICTION

CLASSIFICATION
(nominal or categorical labels)

REGRESSION
(Numerical values)

Regression Analysis

- **Regression analysis is a** statistical methodology that is most often used for numeric prediction.

- Regression analysis is used in statistics to find trends in data.

- For example, you might guess that there’s a connection between how much you eat and how much you weigh; regression analysis can help you quantify that.

- Regression analysis will provide you with an equation for a graph so that you can make predictions about your data. For example, if you’ve been putting on weight over the last few years, it can predict how much you’ll weigh in ten years time if you continue to put on weight at the same rate.

Clustering

- Unlike classification and regression, which analyze class-labeled (training) data sets, **clustering analyzes data objects without consulting class labels.**

- The objects are clustered or grouped based on the principle of ***maximizing the intraclass similarity and minimizing the interclass similarity***. *That is*, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.

Formation of clusters with similar properties

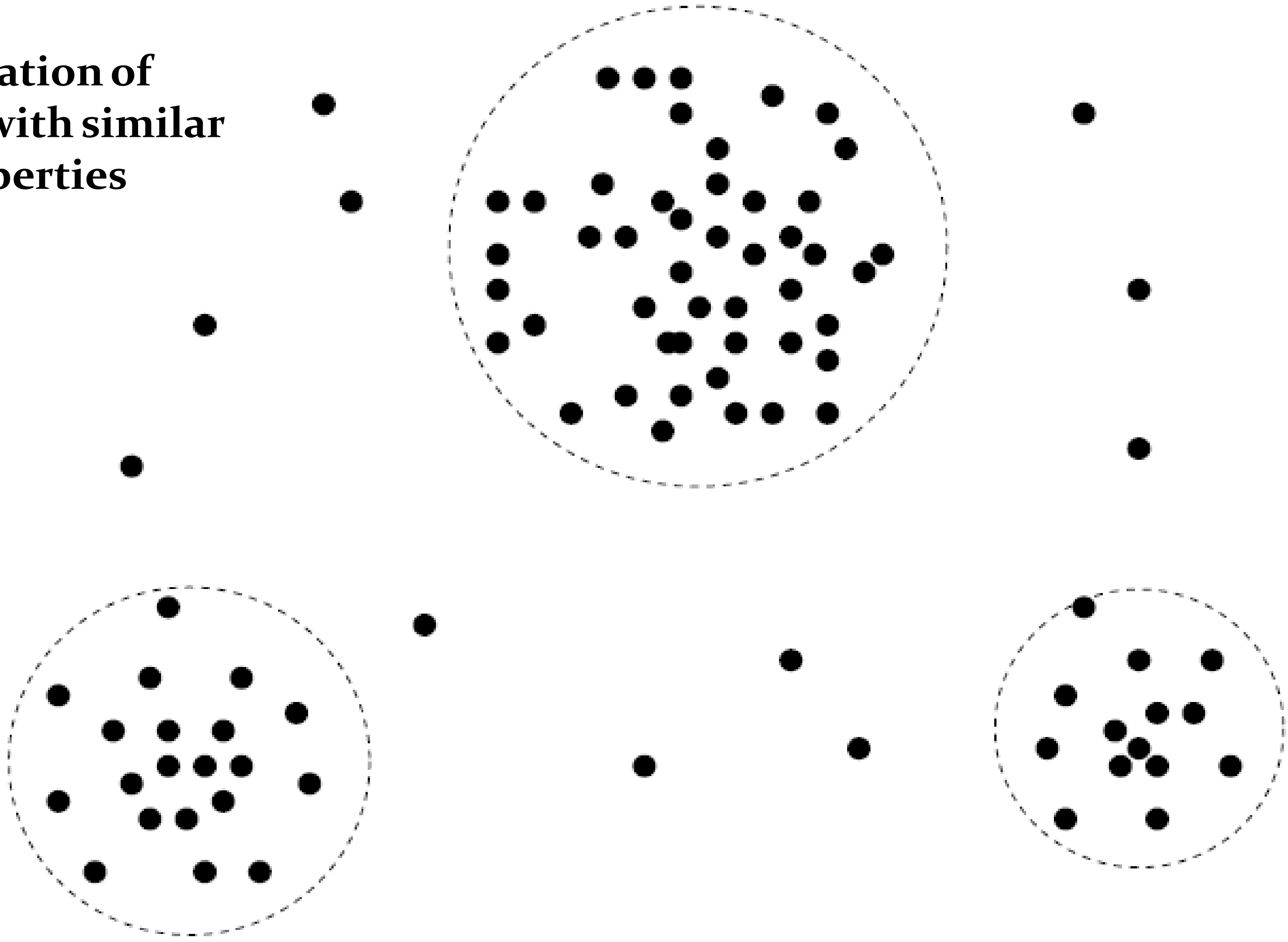


Figure 1.10 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

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.**

- **For example: Outlier analysis may uncover fraudulent usage of credit cards by** detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

Data Characterization

- **Data characterization is a summarization of the general characteristics or features** of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

- For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.

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

Data Discrimination

- **Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.**

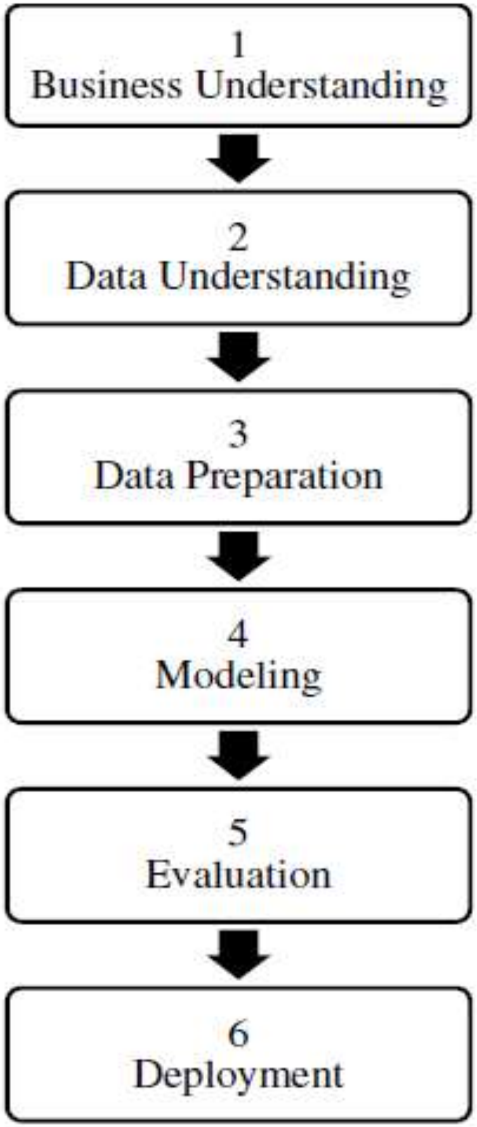
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.

- For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.

- The methods used for data discrimination are similar to those used for data characterization.

Data Mining Life Cycle/Implementation Process

- The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required depending upon the outcome of each phase. The main phases are:



1 Business Understanding

2 Data Understanding

3 Data Preparation

4 Modeling

5 Evaluation

6 Deployment

Figure 3.2: Phases of Data Mining Life Cycle

Business Understanding

In this phase, business and data-mining goals are established.

- First, you need to understand **business and client objectives**. You need to define what your client wants (which many times even they do not know themselves)
- Take stock of the current data mining scenario. Factor in **resources, assumption, constraints**, and other significant factors into your assessment.
- Using business objectives and current scenario, **define your data mining goals.**
- A good **data mining plan** is very detailed and should be developed to accomplish both business and data mining goals.

Data Understanding

- Performed to check whether Data is appropriate for the data mining goals.
- First, **data is collected from multiple data sources** available in the organization.
- These data sources may include multiple databases, flat files or data cubes. There are **issues like object matching and schema integration which can arise** during Data Integration process. It is a quite complex and tricky process as data from various sources unlikely to match easily.
- Next, the step is to search for properties of acquired data. A good way to explore the data is to answer the data mining questions (decided in business phase) using the query, reporting, and visualization tools.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

Data Preparation/Preprocessing

- In this phase, **data is made production ready.**
- The data preparation process consumes about 90% of the time of the project.
- The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required).
- Data cleaning is a process to "clean" the data by smoothing noisy data and filling in missing values.
- For example, for a customer demographics profile, age data is missing. The data is incomplete and should be filled. In some cases, there could be data outliers. For instance, age has a value 100. Data could be inconsistent. For instance, name of the customer is different in different tables.
- Data transformation operations change the data to make it useful in data mining. Following transformation can be applied

Data Transformation

Data transformation operations would contribute toward the success of the mining process.

- **Smoothing:** It helps to remove noise from the data. For eg : using Binning
- **Aggregation:** Summary or aggregation operations are applied to the data. i.e., the weekly sales data is aggregated to calculate the monthly and yearly total.
- **Generalization:** In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.
- **Normalization:** Normalization performed when the attribute data are scaled up or scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.

- **The result of this process is a final data set that can be used in modeling.**

Modelling

- In this phase, mathematical models are used to determine data patterns.
- **Based on the business objectives, suitable modeling techniques should be selected for the prepared dataset.**
- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

Evaluation

- In this phase, patterns identified are evaluated against the business objectives.
- Results generated by the data mining model should be evaluated against the business objectives.
- Gaining business understanding is an iterative process. In fact, while understanding, new business requirements may be raised because of data mining.
- **A go or no-go decision is taken to move the model in the deployment phase.**

Deployment

- In the deployment phase, you ship your data mining discoveries to everyday business operations.

- The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.

- A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created.

- A final project report is created with lessons learned and key experiences during the project. This helps to improve the organization's business policy.

Challenges to Implementation of Data Mining

- **Skilled Experts** are needed to formulate the data mining queries.
- Due to small size training database, a model may not fit future states.
- Data mining needs **large databases** which sometimes are **difficult to manage.**
- **Business practices may need to be modified** to determine to use the information uncovered.
- If the data set is not diverse, data mining results may not be accurate.
- **Integration** information needed from **heterogeneous databases** and global information systems could **be complex.**

Data Mining Tools

Following are 2 popular Data Mining Tools widely used in Industry

- **R-language:**

  R language is an open source tool for statistical computing and graphics. R has a wide variety of statistical, classical statistical tests, time-series analysis, classification and graphical techniques. It offers effective data handing and storage facility.

- **Oracle Data Mining:**

  Oracle Data Mining popularly known as ODM is a module of the Oracle Advanced Analytics Database. This Data mining tool allows data analysts to generate detailed insights and makes predictions. It helps predict customer behavior, develops customer profiles, identifies cross-selling opportunities.

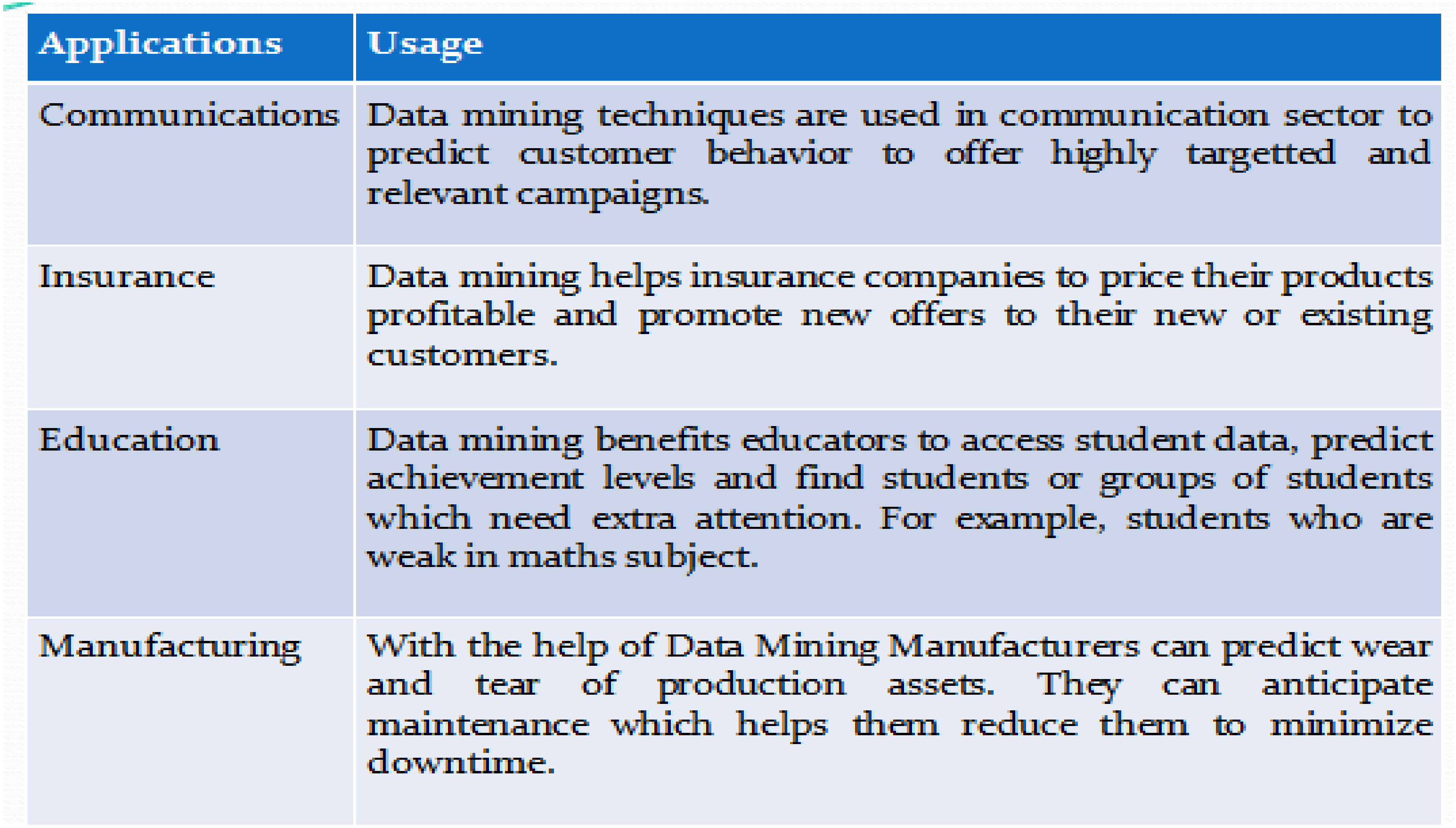
Benefits of Data Mining

- Data mining technique helps companies to get knowledge-based information.
- Data mining **helps organizations to make the profitable adjustments in operation and production**.
- The data mining is a cost-effective and efficient solution compared to other statistical data applications.
- Data mining **helps with the decision-making process**.
- Facilitates **automated prediction of trends and behaviors** as well as automated discovery of hidden patterns.
- It can be implemented in new systems as well as existing platforms
- It is the **speedy process** which makes it easy for the users to analyze huge amount of data in less time.

Limitations of Data Mining

- There are chances of companies may sell useful information of their customers to other companies for money. For example, American Express has sold credit card purchases of their customers to the other companies.
- Many data mining analytics software is difficult to operate and **requires advance training to work on**.
- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, **the selection of correct data mining tool is a very difficult task.**
- If the data mining techniques are not accurate, it can cause serious consequences in certain conditions.

Applications of Data Mining/Where it is used?



Applications	Usage
Communications	Data mining techniques are used in communication sector to predict customer behavior to offer highly targetted and relevant campaigns.
Insurance	Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing 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.
Manufacturing	With the help of Data Mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime.

Applications	Usage
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 comes up with the offer which encourages customers to increase their spending.
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.
Bioinformatics	Data Mining helps to mine biological data from massive datasets gathered in biology and medicine.
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.

Applications	Usage
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 use Data mining techniques to get more customers into their ecommerce store.
Super Markets	Data Mining allows supermarket's 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, diapers and so on.
Crime Investigation	Data Mining helps crime investigation agencies to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border crossing etc.

Data Mining Task Primitives

- Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed.

- **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 inter-actively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

- The five data mining task primitives are:

1) The set of task-relevant data to be mined
2) The kind of knowledge to be mined
3) The background knowledge to be used in the discovery process
4) The interestingness measures and thresholds for pattern evaluation
5) The expected representation for visualizing the discovered patterns

- **The 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 includes the database attributes or data warehouse dimensions of interest.

- **The kind of knowledge to be mined**: This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

- **The background knowledge to be used in the discovery process:** The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction.

- **The interestingness measures and thresholds for pattern evaluation:** 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, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

- **The expected representation for visualizing the discovered patterns:** This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

A data mining query language can be designed to incorporate these primitives, allowing users to flexibly interact with data mining systems. **Having a data mining query language provides a foundation on which user-friendly graphical interfaces can be built.**

Integration of a Data Mining System with a Database or DataWarehouse System

Q) How to integrate or *couple* the DM system with a database (DB) system and/or a data warehouse (DW) system.

- **No coupling**: *No coupling* means that a DM system will not utilize any function of a DB or DW system.
- **Loose coupling**: *Loose coupling* means that a DM system will use some facilities of a DB or DW system, fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place 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,

- **Semi tight coupling**: *Semi tight coupling* means that besides linking a DM system to a DB/DW system, efficient implementations of a few essential data mining primitives.

- **Tight coupling**: *Tight coupling* means that a DM system is smoothly integrated into the DB/DW system. This approach is highly desirable because it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment.

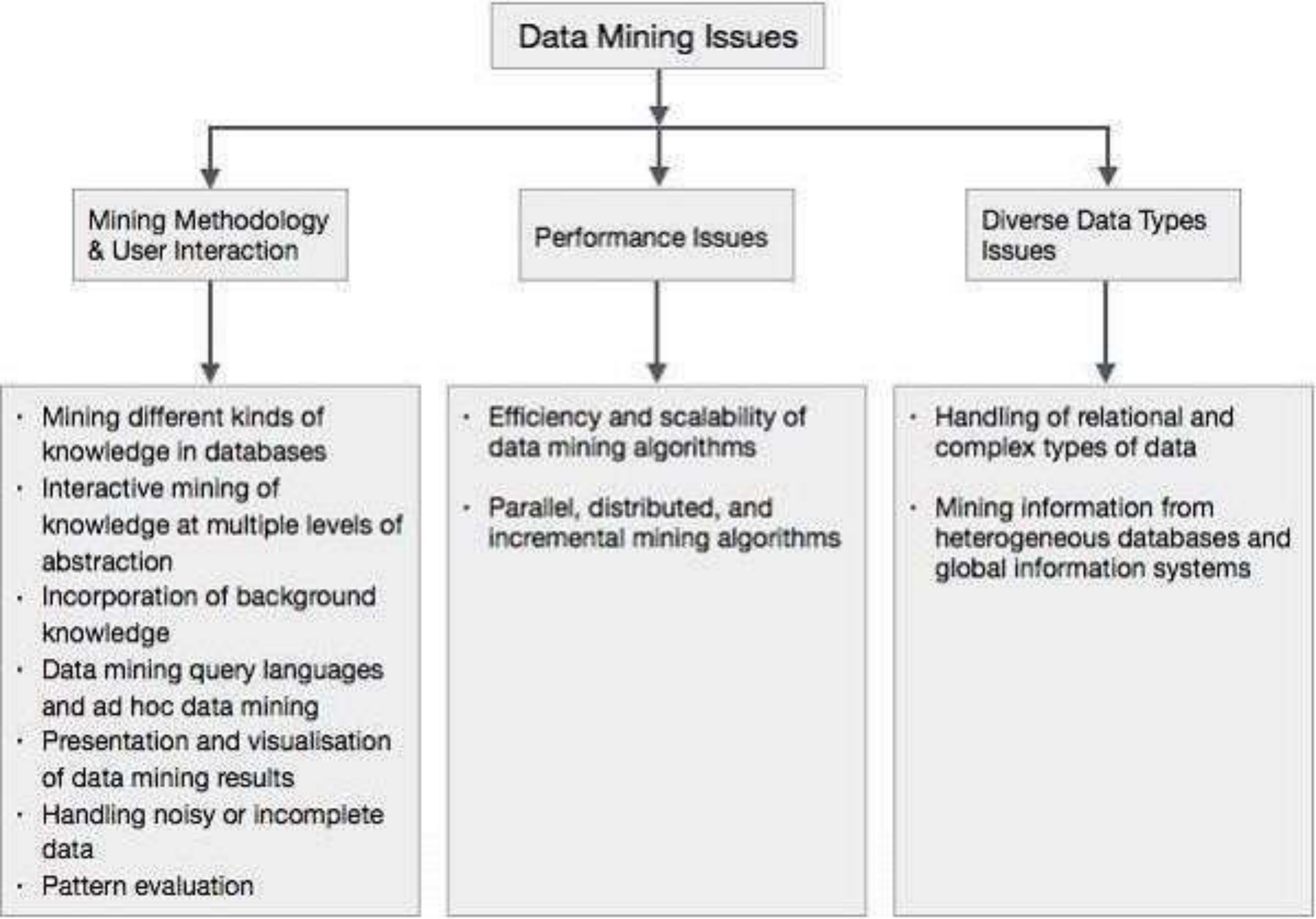
Another Probable Question

Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: *no coupling, loose coupling, semitight coupling,* and *tight coupling.* State which approach you think is the most popular, and why?

Major Issues in Data Mining(previous year paper)

- 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. These factors also create some issues. Here the major issues regarding –

- **Mining Methodology and User Interaction**
- **Performance Issues**
- **Diverse Data Types Issues**



Data Mining Issues

Mining Methodology & User Interaction

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualisation of data mining results
- Handling noisy or incomplete data
- Pattern evaluation

Performance Issues

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining algorithms

Diverse Data Types Issues

- Handling of relational and complex types of data
- Mining information from heterogeneous databases and global information systems

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. So techniques are needed to assess the interestingness of discovered patterns**.**

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.The running time should be predictable,short and acceptable by applications.

- **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 is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Type 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 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.