DATA WAREHOUSE & DATA MINING

BY

DR. PREETHAM KUMAR

HOD

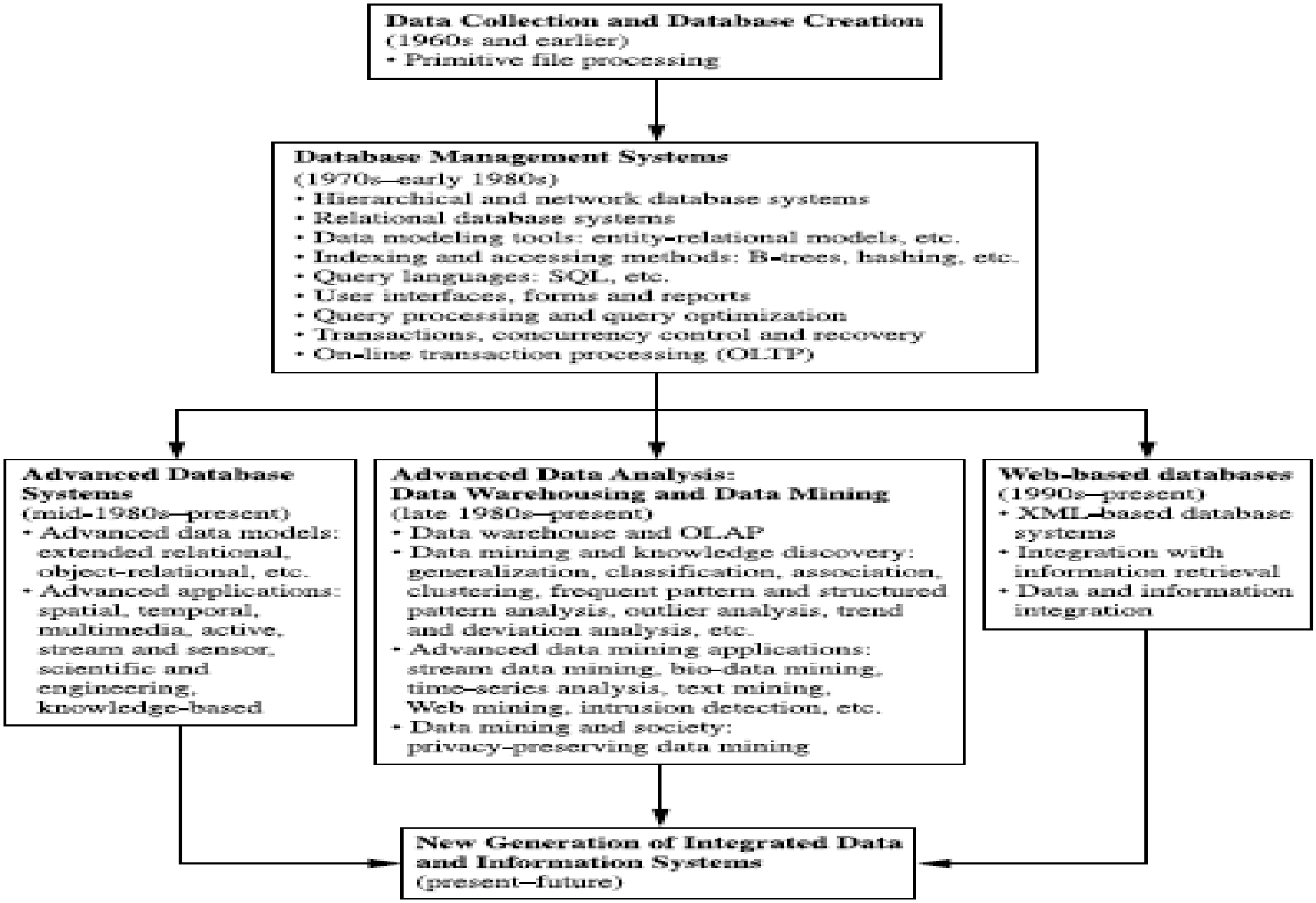
DEPT. OF INFORMATION & COMMUNICATION TECHNOLOGY

Ref : Jiawei Han & Micheline & Kamber

What Motivated Data Mining? Why Is It Important?

- Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.
- The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.
- Data mining can be viewed as a result of the natural evolution of information technology.

- Database system industry has witnessed an evolutionary path in the development of the following functionalities (Figure):
- *Data collection and database creation, data management* (including data storage and retrieval, and database transaction processing), and *advanced data analysis* (involving data warehousing and data mining).



Dr. Preetham Kumar, Dept. of I & CT

The evolution of database system technology.

- For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing.
- With numerous database systems offering query and transaction processing as common practice, advanced data analysis has naturally become the next target.

- Since the 1960s, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems.
- The research and development in database systems since the 1970s has progressed from early hierarchical and network database systems to the development of relational database systems…….

- Data modeling tools, and indexing and accessing methods.
- In addition, users gained convenient and flexible data access through query languages, user interfaces, optimized query processing, and transaction management.
- Efficient methods for on-line transaction processing (OLTP), where a query is viewed as a read-only transaction, have contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data.

- Database technology since the mid-1980s has been characterized by the popular adoption of relational technology and an upsurge of research and development activities on new and powerful database systems.
- These promote the development of advanced data models such as extended-relational, object-oriented, object-relational, and deductive models.
- Application-oriented database systems, including spatial, temporal, multimedia, active, stream, and sensor, and scientific and engineering databases, knowledge bases, and office information bases, have flourished

- Issues related to the distribution, diversification, and sharing of data have been studied extensively.
- Heterogeneous database systems and Internet-based global information systems such as the World Wide Web (WWW) have also emerged and play a vital role in the information industry.

- The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful and affordable computers, data collection equipment, and storage media.
- This technology provides a great boost to the database and information industry, and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis.

- Data can now be stored in many different kinds of databases and information repositories.
- One data repository architecture that has emerged is the data warehouse a repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making.
- Data warehouse technology includes data cleaning, data integration, and on-line analytical processing (OLAP), that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation as well as the ability to view information from different angles data classification, clustering, and the characterization of data changes over time

- Huge volumes of data can be accumulated beyond databases and data warehouses.
- Typical examples include the World Wide Web and *data streams*, where data flow in and out like streams, as in applications like video surveillance, telecommunication, and sensor networks. The effective and efficient analysis of data in such different forms becomes a challenging task.

- The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation.
- The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools (Figure).
- As a result, data collected in large data repositories become "data tombs"—data archives that are seldom visited

- Consequently, important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data.

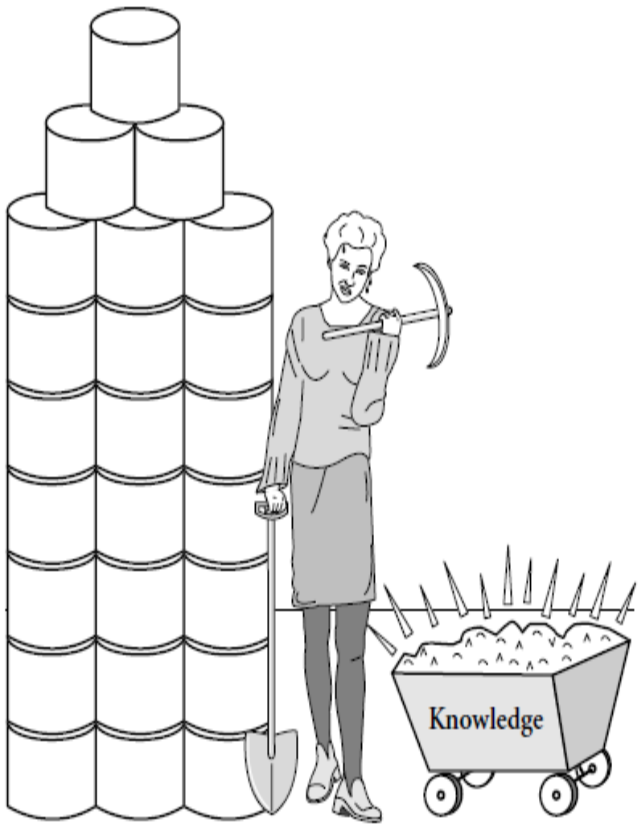
- In addition, consider expert system technologies, which typically rely on users or domain experts to *manually* input knowledge into knowledge bases.
- Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly.
- Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research.
- The widening gap between data and information calls for a systematic development of *data mining tools* that will turn data tombs into "golden nuggets" of knowledge.

What Is Data Mining?

- Simply stated, data mining refers to *extracting* or *"mining" knowledge from large amounts of data*. The term is actually a misnomer.
- Remember that the mining of gold from rocks or sand is referred to as *gold* mining rather than rock or sand mining.
- Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long.
- "Knowledge mining," a shorter term, may not reflect the emphasis on mining from large amounts of data.

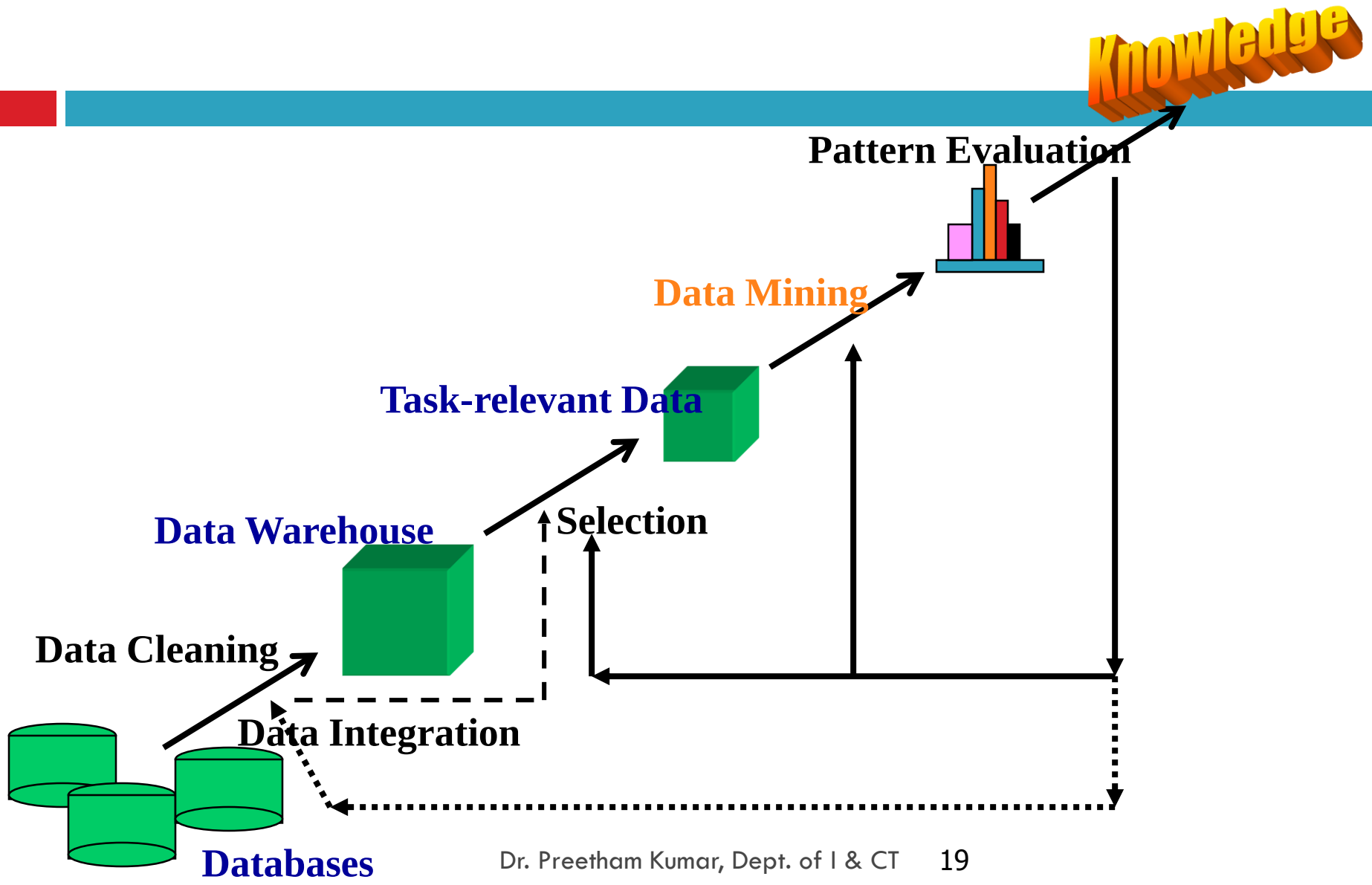
- Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material (Figure).
- Thus, such a misnomer that carries both “data” and “mining” became a popular choice.
- Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

- Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.
- Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.



Data mining—searching for knowledge (interesting patterns) in your data.

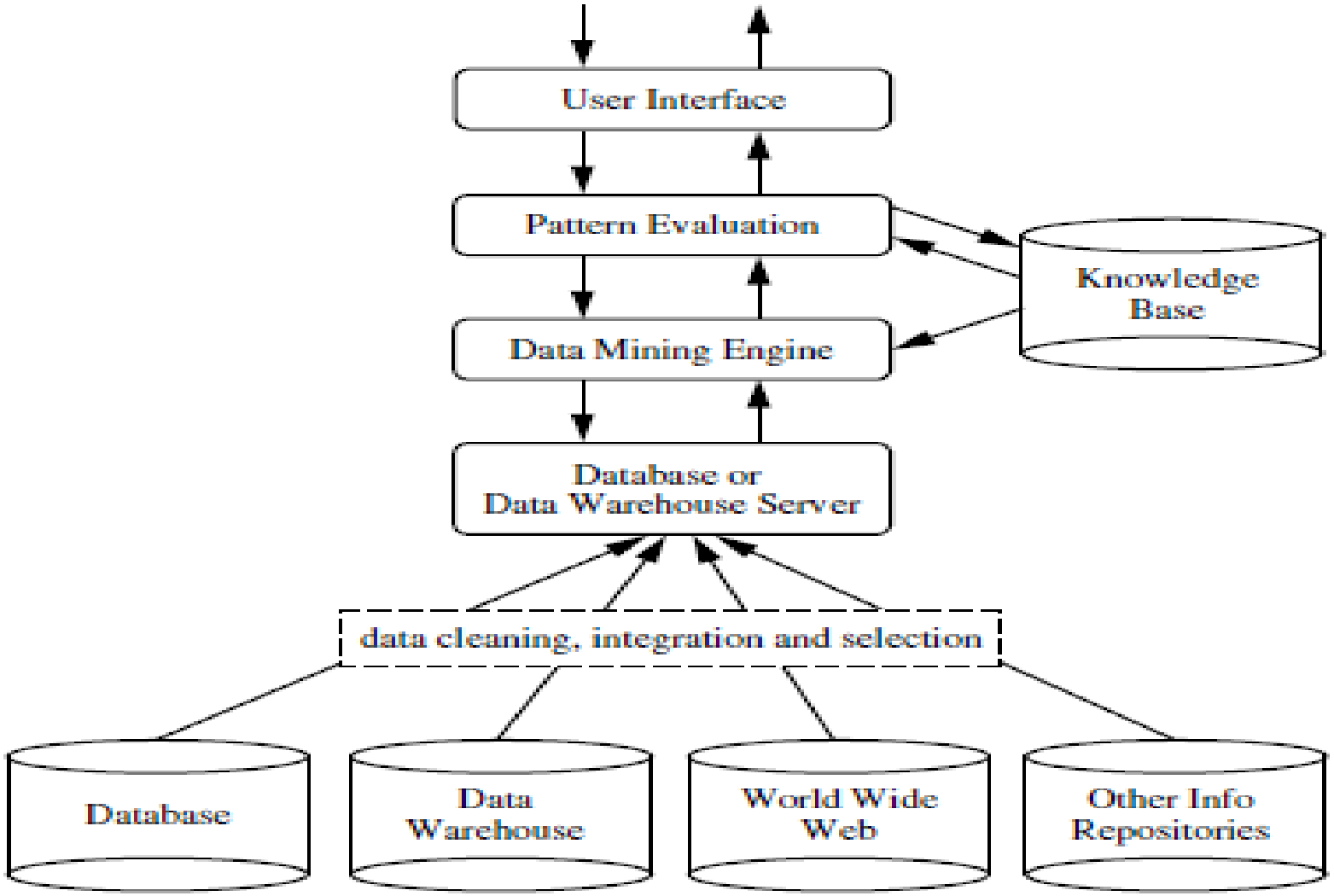
Knowledge Discovery (KDD) Process



- Knowledge discovery as a process is depicted in Figure and consists of 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)1
- **3.** Data selection (where data relevant to the analysis task are retrieved fromthe database)
- **4.** Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

- 5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
- **6.** Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures;)
- **7.** Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

DATA MINING SYSTEM ARCHITECTURE



Architecture of a typical data mining system.

- Based on this view, the architecture of a typical data mining system may have the following major components (Figure):

Database, data warehouse,WorldWideWeb, or other information repository:

- This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories.
- Data cleaning and data integration techniques may be performed on the data.

Database or data warehouse server:

- The database or data warehouse server is responsible

for fetching the relevant data, based on the user's data mining request.

- **Knowledge base:** This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
- Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.
- Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included.
- Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data mining engine:

- This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module:

- This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns.
- It may use interestingness thresholds to filter out discovered patterns.
- Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.
- For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns

- **User interface:** This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.
- In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms

Data Mining—On What Kind of Data?

- Relational Databases, Data Warehouses, Transactional Databases
- Advanced Data and Information Systems and Advanced Applications, Object-Relational Databases
- Temporal Databases, Sequence Databases, and Time-Series Databases, Spatial Databases and Spatiotemporal Databases
- Text Databases and Multimedia Databases,
- Heterogeneous Databases and Legacy Databases
- Data Streams, TheWorld WideWeb

(FOR DETAILED DISCUSSION REFER HAN & KAMBER TEXT)

What Kinds of Patterns Can Be Mined?

- Various types of databases and information repositories on which data mining can be performed.
- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks
- data mining tasks can be classified into two categories: **descriptive and predictive.**

- **Descriptive** mining tasks characterize the general properties of the data in the database.
- **Predictive mining** tasks perform inference on the current data in order to make predictions.

- Users may have no idea regarding what kinds of patterns in their data may be interesting.
- Hence they may like to search for several different kinds of patterns in parallel.
- Thus it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations or applications

- Data mining systems should be able to discover patterns at various granularity (i.e., different levels of abstraction).
- Data mining systems should also allow users to specify hints to guide or focus the search for interesting patterns.
- Because some patterns may not hold for all of the data in the database, a measure of certainty or "trustworthiness" is usually associated with each discovered pattern.

Data mining functionalities, and the kinds of patterns

Description: Characterization and Discrimination:

- Data can be associated with classes or concepts.
- For example, in the *AllElectronics* store, classes of items for sale include ***computers*** and ***printers***, and concepts of customers include ***bigSpenders*** and b***udgetSpenders***.

- The individual classes and concepts may be described in summarized, concise, and precise terms. Such descriptions of a class or a concept are called **class/concept descriptions.**
- These descriptions can be derived via

(1) *data characterization*, by summarizing the data of the class under study (often called the target class) in general terms. Or

(2) *data discrimination*: by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or

(3) both data characterization and discrimination.

Data characterization

- Is a summarization of the general characteristics or features of a target class of data.
- The data corresponding to the user-specified class are typically collected by a database query.
- For example, to study the characteristics of software products whose sales increased by 10% in the last year, the data related to such products can be collected by executing an SQL query.

Other methods for effective data summarization and characterization

- Simple data summaries based on statistical measures and plots.
- The data cube–based OLAP roll-up operation can be used to perform user-controlled data summarization along a specified dimension.

Output of data characterization

- The output of can be presented in various forms.
- Examples include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs.
- The resulting descriptions can also be presented as generalized relations or in rule form(called characteristic rules).

Example for characterization

- A data mining system should be able to produce a description summarizing the characteristics of customers who spend more than $1,000 a year at *AllElectronics*.
- The result could be a general profile of the customers, such as they are **40–50 years old, employed, and have excellent credit ratings.**
- The system should allow users to drill down on any dimension, such as on *occupation* in order to view these customers according to their type of employment.


Google autocomplete results:
"Why is [state] so..."
liberal
good
republican
cold
cold
cold
cold
liberal
cold
windy
boring
democratic
corrupt
republ-
lican
important
expensive
liberal
white
expensive
smart
small
haunted
haunted
bad
boring
rich
expensive
empty
mormon
fit
flat
conservative
poor
poor
strict
racist
cheap
poor
hot
poor
republican
poor
poor
backwards
big
good
racist
hot
expensive
expensive
as of January 2014

Data discrimination

- Is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.
- The target and contrasting classes can be specified by the user.
- The corresponding data objects are retrieved through database queries.

Example

- The user may like to compare the **general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period.**
- The methods used for data discrimination are similar to those used for data characterization.

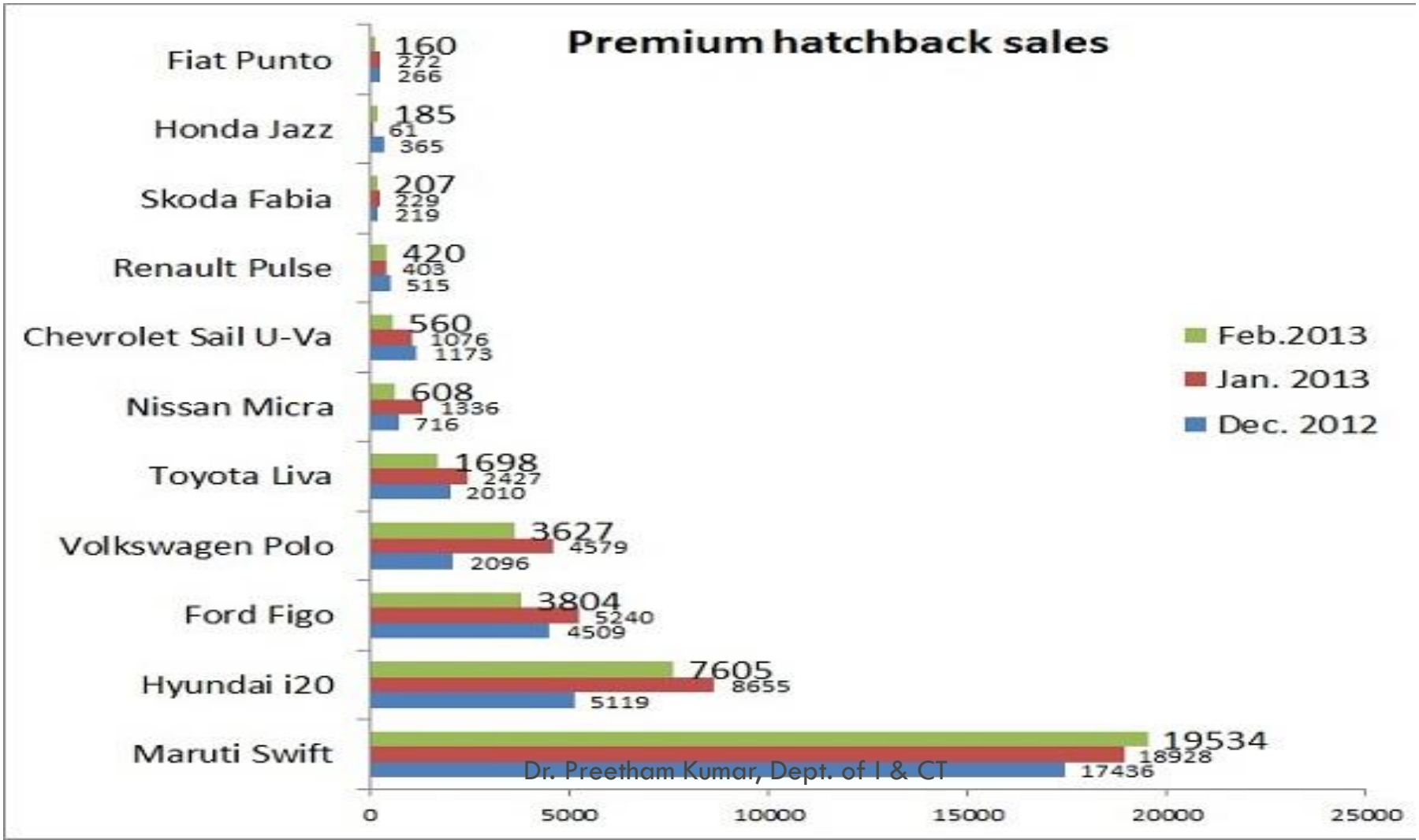
How are discrimination descriptions output

- The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help distinguish between the target and contrasting classes.
- Discrimination descriptions expressed in rule form are referred to as discriminant rules.

Example

- Data mining system should be able to compare two groups of *AllElectronics* customers, such as those who shop for computer products regularly (more than two times a month) versus those who rarely shop for such products (i.e., less than three times a year).
- The resulting description provides a general comparative profile of the customers as follows

- 80% of the customers who frequently purchase computer products are **between 20 and 40 years old and have a university education.**
- Drilling down on a dimension, such as *occupation*, or adding new dimensions, such as *income level*, may help in finding even more discriminative features between the two classes.


Premium hatchback sales
Fiat Punto
160
272
266
Honda Jazz
185
61
365
Skoda Fabia
207
229
219
Renault Pulse
420
403
515
Chevrolet Sail U-Va
560
1076
1173
Nissan Micra
608
1336
716
Toyota Liva
1698
2427
2010
Volkswagen Polo
3627
4579
2096
Ford Figo
3804
5240
4509
Hyundai i20
7605
8655
5119
Maruti Swift
19534
18928
17436
Feb.2013
Jan. 2013
Dec. 2012
0
5000
10000
15000
20000
25000

Mining Frequent Patterns, Associations, and Correlations

- **Frequent patterns:** patterns that occur frequently in data.
- There are many kinds of frequent patterns including
  - Itemsets
  - subsequences and
  - substructures

A *frequent itemset* : typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.

A frequently occurring subsequence: are the patterns that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (*frequent*) *sequential pattern*

- **A substructure :** refer to different structural forms, such as graphs, trees, or lattices, which may be combined with itemsets or subsequences.
- If a substructure occurs frequently, it is called a (*frequent*) *structured pattern*.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Association analysis

- Suppose, a marketing manager of *AllElectronics*, you would like to determine which items are frequently purchased together within the same transactions.
- An example of such a rule, mined from the *AllElectronics* transactional database, is

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

where X is a variable representing a customer

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

- 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 of the transactions under analysis showed that computer and software were purchased together.
- This association rule involves a single attribute or predicate (i.e., *buys*) that repeats.
- Association rules that contain a single predicate are referred to as **single-dimensional association rules.**

- Dropping the predicate notation, the above rule can be written simply as "$computer \Rightarrow software\ [1\%, 50\%]$".

$$age(X, \text{“20...29”}) \wedge income(X, \text{“20K...29K”}) \Rightarrow buys(X, \text{“CD player”})$$
$$[support = 2\%, confidence = 60\%]$$

- The above rule indicates that customers under study, 2% are 20 to 29 years of age with an income of 20,000 to 29,000 and have purchased a CD player.
- There is a 60% probability that a customer in this age and income group will purchase a CD player.

$$age(X, \text{“20...29”}) \wedge income(X, \text{“20K...29K”}) \Rightarrow buys(X, \text{“CD player”})$$
$$[support = 2\%, confidence = 60\%]$$

- The above rule is an association between more than one attribute, or predicate (i.e., *age*, *income*, and *buys*).
- The above rule can be referred to as a **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.**
- Additional analysis can be performed to uncover interesting statistical correlations between associated attribute-value pairs.

Classification and Prediction

- Classification is the process of finding a **model** (or function) that **describes** and **distinguishes** data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).
- Classification predicts categorical (discrete, unordered) labels

How is the derived model presented

- *classification (IF-THEN) rules*
- *decision trees*
- *mathematical formulae*
- *or neural networks*

- **A decision tree** is a flow-chart-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.
- **A decision tree** **can easily** be converted to classification rules.

- **A neural network,** when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.
- There are many other methods for constructing classification models, such as
  - Naïve Bayesian classification
  - support vector machines and
  - *k*-nearest neighbor classification.

(a)

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")

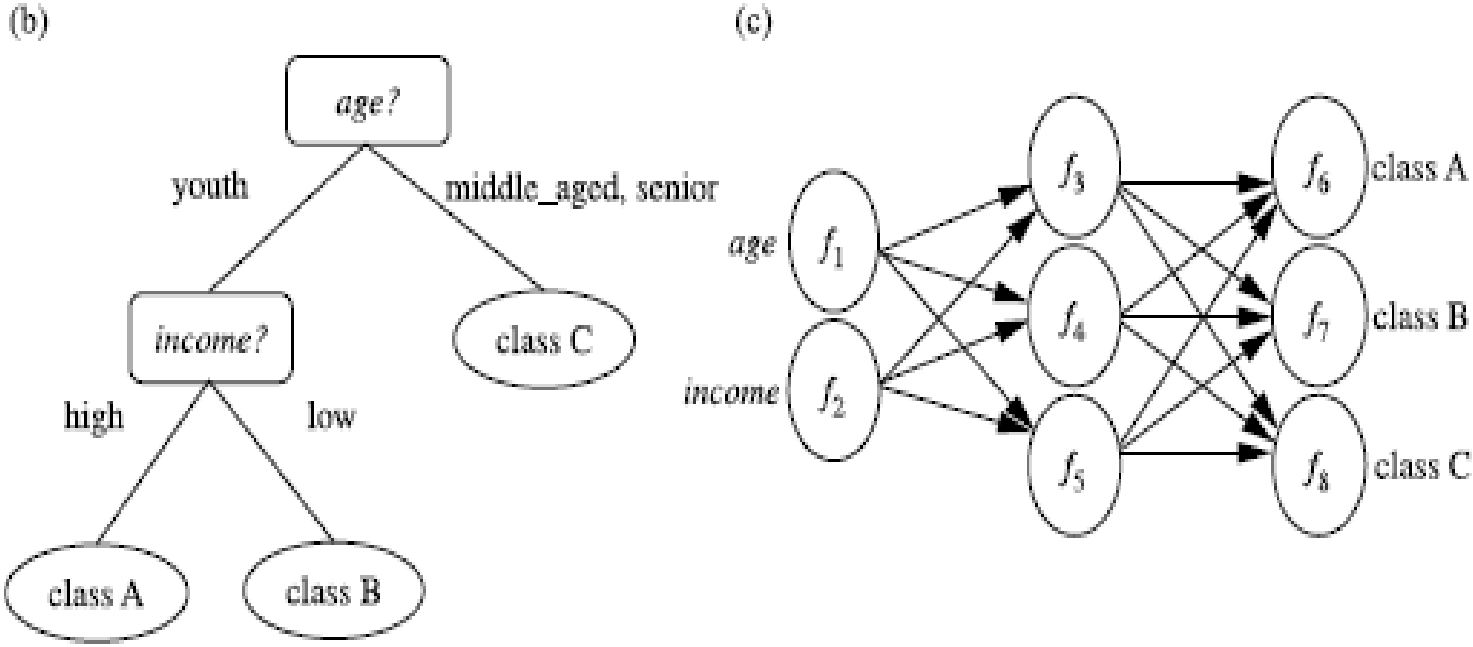


Figure 1.10 A classification model can be represented in various forms, such as (a) IF-THEN rules, (b) a decision tree, or a (c) neural network.

Prediction

- Prediction models continuous-valued functions.
- That is, it is used to predict missing or unavailable *numerical data values* rather than class labels.
- Although the term *prediction* may refer to both numeric prediction and class label prediction, in this we refer primarily to numeric prediction.

- Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.
- Prediction also encompasses the identification of distribution *trends* based on the available data.
- Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded.

Example for Classification and prediction

- Suppose, sales manager of *All Electronics* would like to classify a large set of items in the store, based on three kinds of responses to a sales campaign :
  - *good response*, *mild response*, and *no response*.
- *You would like to derive a model for each of these three classes based on the descriptive features of the items such as*
  - *price, brand, place made, type, and category.*

- The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.
- Suppose that the resulting classification is expressed in the form of a decision tree and
- Tree may identify *price* as being the single factor that best distinguishes the three classes
- The tree may reveal that, after *price*, other features that help further distinguish objects of each class from another include *brand* and *place made*.
- Such a decision tree may help to understand the impact of the given sales campaign and design a more effective campaign for the future.

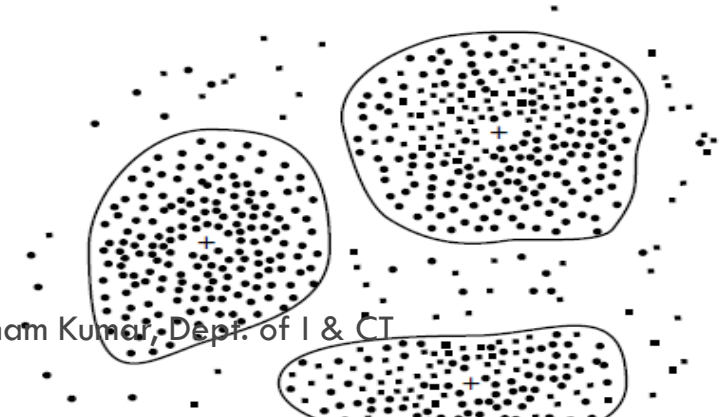
- Suppose instead, that rather than predicting categorical response labels for each store item, you would like to predict the amount of revenue that each item will generate during an upcoming sale at *AllElectronics*, based on previous sales data.
- This is an example of (numeric) prediction because the model constructed will predict a continuous-valued function, or ordered value.

Cluster Analysis

- Classification and prediction analyze class-labeled data objects.
- Clustering analyzes data objects without consulting a known class label.
- The class labels are not present in the training data simply because they are not known to begin with.
- Clustering can be used to generate such labels.

- The objects are clustered or grouped based on the principle of *maximizing the intra class similarity and minimizing the interclass similarity*.
- That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.
- Each cluster that is formed can be viewed as a class of objects from which rules can be derived.

- Cluster analysis : Cluster analysis can be performed on *All Electronics* customer data in order to identify homogeneous subpopulations of customers.
- These clusters may represent individual target groups for marketing.
- Figure shows a 2-D plot of customers with respect to customer locations in a city. Three clusters of data points are evident.



Outlier Analysis

- Database may contain data objects that do not comply with the general behavior or model of the data called as **outliers.**
- Most data mining methods discard outliers as noise or exceptions
- However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones.
- The analysis of outlier data is referred to as outlier mining.

- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or
- using distance measures where objects that are a substantial distance from any other cluster are considered outliers.
- Rather than using statistical or distance measures, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group.

Example

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

Evolution Analysis

- Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.
- Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of *time related* data,
- Distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Example

- Suppose that you have the major stock market (time-series) data of the last several years available from the New York Stock Exchange and you would like to invest in shares of high-tech industrial companies.
- A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies.
- Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

Are All of the Patterns Interesting?

- A data mining system has the potential to generate thousands or even millions of patterns, or rules.
- *Only* a small fraction of the patterns potentially generated would actually be of interest to any given user.
- *"What makes a pattern interesting? Can a data mining system generate all of the interesting patterns? 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*.
- A pattern is also interesting if it validates a hypothesis that the user *sought to confirm*.
- An interesting pattern represents knowledge

Several objective measures of pattern interestingness

- **Support:** An objective measure for association rules of the form $X \rightarrow Y$, **representing the percentage of transactions from a transaction database that the given rule satisfies.**
- This is taken to be the probability $P(X \cup Y)$, where $X \cup Y$ indicates that a transaction contains both X and Y, that is, the union of itemsets X and Y.

- **Confidence:** which assesses the degree of certainty of the detected association.
- This is taken to be the conditional probability $P(Y \mid X)$, that is, the probability that a transaction containing X also contains Y.
- More formally, support and confidence are defined as

$$support(X \Rightarrow Y) = P(X \cup Y).$$

$$confidence(X \Rightarrow Y) = P(Y \mid X).$$

- In general, each interestingness measure is associated with a threshold which may be controlled by the user.
- For example, rules that do not satisfy a confidence threshold of, say, 50% can be considered uninteresting.
- Rules below the threshold likely reflect noise, exceptions, or minority cases and are probably of less value.

Subjective measure

- Although objective measures help identify interesting patterns, they are insufficient unless combined with subjective measures that reflect the needs and interests of a particular user
- For example, patterns describing the characteristics of customers who shop frequently at *AllElectronics* should interest the marketing manager, but may be of little interest to analysts studying the same database for patterns on employee performance.
- Furthermore, many patterns that are interesting by objective standards may represent common knowledge and, therefore, are actually uninteresting.

- **Subjective interestingness:** measures are based on user beliefs in the data.
- These measures find patterns interesting if they are unexpected (contradicting a user's belief) or offer strategic information on which the user can act. In this case, patterns are referred to as actionable.
- Patterns that are expected can be interesting if they confirm a hypothesis that the user wished to validate, or resemble a user's hunch.

Can a data mining system generate all of the interesting patterns

- **Refers to completeness** of a data mining algorithm.
- It is often unrealistic and inefficient for data mining systems to generate all of the possible patterns.
- Instead, user-provided constraints and interestingness measures should be used to focus the search.
- For some mining tasks, such as association, this is often sufficient to ensure the completeness of the algorithm.
- Association rule mining is an example where the use

of constraints and interestingness measures can ensure the completeness of mining.

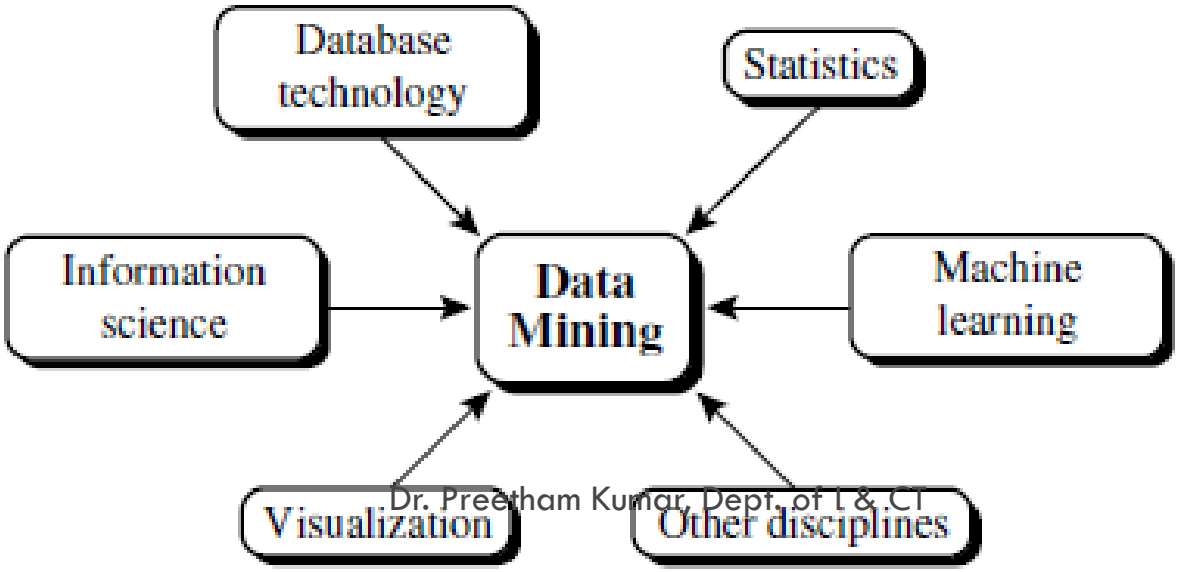
*Can a data mining system generate **only** interesting patterns*

- It is an optimization problem in data mining.
- It is highly desirable for data mining systems to generate only interesting patterns.
- This would be much more efficient for users and data mining systems, because neither would have to search through the patterns generated in order to identify the truly interesting ones.
- Progress has been made in this direction; however, such optimization remains a challenging issue in data mining.

- Measures of pattern interestingness are essential for the efficient discovery of patterns of value to the given user.
- Measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones.
- Measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre-specified interestingness constraints.

Classification of Data Mining Systems

- Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science



- Depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing.
- Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology

- Because of the diversity of disciplines contributing to data mining, data mining research is expected to generate a large variety of data mining systems.
- Therefore, it is necessary to provide a clear classification of data mining systems, which may help potential users distinguish between such systems and identify those that best match their needs.
- Data mining systems can be categorized according to various criteria, as follows:

Classification according to the kinds of databases mined:

- A data mining system can be classified according to the kinds of databases mined.
- Database systems can be classified according to different criteria (such as data models, or the types of data or applications involved), each of which may require its own data mining technique.

For instance,

- If classifying according to data models, we may have a relational, transactional, object-relational, or data warehouse mining system.
- If classifying according to the special types of data handled, we may have a spatial, time-series, text, stream data, multimedia data mining system, or a World Wide Web mining system.

Classification according to the kinds of knowledge mined

- Data mining systems can be categorized according to the kinds of knowledge they mine, that is, based on data mining functionalities, such as characterization, discrimination, association, classification, prediction, clustering, outlier analysis, and evolution analysis.
- A comprehensive data mining system usually provides multiple and/or integrated data mining functionalities.

- Moreover, data mining systems can be distinguished based on the granularity or levels of abstraction of the knowledge mined, including generalized knowledge (at a High level of abstraction),primitive-level knowledge (at a raw data level), or knowledge at multiple levels (considering several levels of abstraction).
- An advanced data mining system should facilitate the discovery of knowledge at multiple levels of abstraction.

- Data mining systems can also be categorized as those that mine data regularities (commonly occurring patterns) versus those that mine data irregularities (such as exceptions, or outliers).
- In general, concept description, association and correlation analysis, classification, prediction, and clustering mine data regularities, rejecting outliers as noise. These methods may also help detect outliers.

Classification according to the kinds of techniques utilized:

- Data mining systems can be categorized according to the underlying data mining techniques employed.
- These techniques can be described according to the
  - **degree of user interaction involved** (e.g., autonomous systems, interactive exploratory systems, query-driven systems) or
  - **The methods of data analysis employed** (e.g., database-oriented or data warehouse– oriented techniques, statistics, visualization, pattern recognition, and so on).

- A sophisticated data mining system will often adopt multiple data mining techniques or work out an effective, integrated technique that combines the merits of a few individual approaches.

Classification according to the applications adapted:

- Data mining systems can also be categorized according to the applications they adapt.
- For example, data mining systems may be tailored specifically for finance, DNA, stock telecommunications, markets, e-mail, and so on.
- Different applications often require the integration of application-specific methods.
- Therefore, a generic, all-purpose data mining system may not fit domain-specific mining tasks.

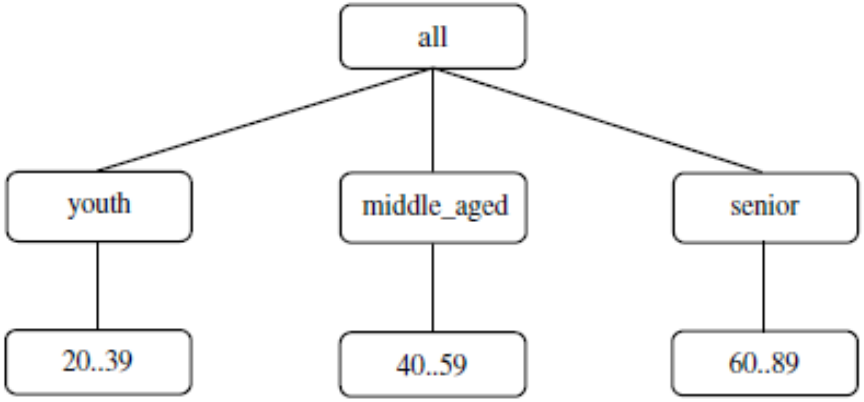
Data Mining Task Primitives

- 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 *interactively* 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 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 (referred to as the *relevant attributes or dimensions*).
- ***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

- This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found.
- *Concept hierarchies* are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.



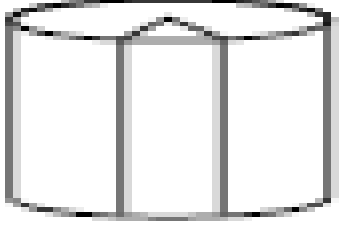
- User beliefs regarding relationships in the data are another form of background knowledge.

The interestingness measures and thresholds for pattern evaluation:

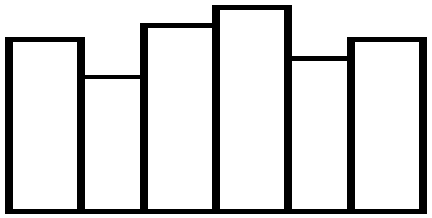
- They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns.
- Different kinds of knowledge may have different interestingness measures.
- For example, interestingness measures for association rules include *support* and *confidence*.
- Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

The expected representation for visualizing the discovered patterns:

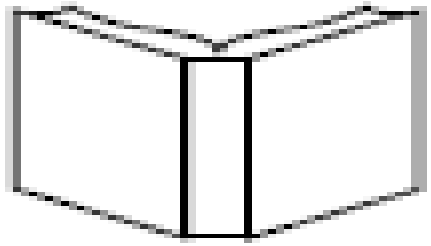
- This refers to the form in which discovered patterns are to be displayed , which may include rules, tables, charts, graphs, decision trees, and cubes.



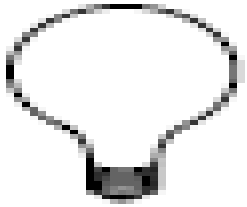
Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria



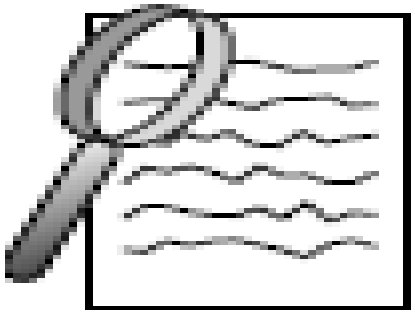
Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering



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
Drill-down and roll-up

- 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.
- This facilitates a data mining system's communication with other information systems and its integration with the overall information processing environment.

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

- **No coupling:** *No coupling* means that a DM system will not utilize any function of a DB or DW 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

- First, a DB system provides a great deal of flexibility and efficiency at storing, organizing, accessing, and processing data. Without using a DB/DWsystem, a DM system may spend a substantial amount of time finding, collecting, cleaning, and transforming data.
- In DB and/or DWsystems, data tend to be well organized, indexed, cleaned, integrated, or consolidated, so that finding the task-relevant, high-quality data becomes an easy task.

- Second , there are many tested, scalable algorithms and data structures implemented in DB and DW systems.
- It is feasible to realize efficient, scalable implementations using such systems. Moreover, most data have been or will be stored in DB/DW systems.
- Without any coupling of such systems, a DM system will need to use other tools to extract data, making it difficult to integrate such a system into an information processing environment.
- Thus, no coupling represents a poor design.

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, indexing, and other system facilities.

- It incurs some advantages of the flexibility, efficiency, and other features provided by such systems.
- However, many loosely coupled mining systems are main memory-based.
- Because mining does not explore data structures and query optimization methods provided by DB or DW systems, it is difficult for loose coupling to achieve high scalability and good performance with large data sets.

Semitight coupling

- *Semitight coupling* means that besides linking a DM system to a DB/DW 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, multiway join, and pre computation of some essential statistical measures, such as sum, count, max, min, standard deviation, and so on.

- Moreover, some frequently used intermediate mining results can be pre-computed and stored in the DB/DW system.
- Because these intermediate mining results are either pre-computed or can be computed efficiently, this design will enhance the performance of a DM system.

Tight coupling

- *Tight coupling* means that a DM system is smoothly integrated into the DB/DW system.
- The data mining subsystem is treated as one functional component of an 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.

- This approach is highly desirable because it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment.

Major Issues in Data Mining

Mining methodology and user interaction issues: These reflect

- the kinds of knowledge mined,
- the ability to mine knowledge at multiple granularities,
- the use of domain knowledge,
- ad hoc mining, and
- knowledge visualization.

Mining different kinds of knowledge in databases

- Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis (which includes trend and similarity analysis).
- These tasks may use the same database in different ways and require the development of numerous data mining techniques.

Interactive mining of knowledge at multiple levels of abstraction

- Difficult to know exactly what can be discovered within a database, the data mining process should be *interactive*.
- For databases containing a huge amount of data, appropriate sampling techniques can first be applied to facilitate interactive data exploration.
- Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

- Specifically, knowledge should be mined by drilling down, rolling up, and pivoting through the data space and knowledge space interactively, similar to what OLAP can do on data cubes.
- In this way, the user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

Incorporation of background knowledge

- Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.
- Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

Data mining query languages and ad hoc data mining

- Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval.
- In a similar vein, high-level data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns.
- Such a language should be integrated with a database or data warehouse query language and optimized for efficient and flexible data mining.

Presentation and visualization of data mining results:

- Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.
- This is especially crucial if the data mining system is to be interactive.
- This requires the system to adopt expressive knowledge representation techniques, such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.

Handling noisy or incomplete data

- The data stored in a database may reflect noise, exceptional cases, or incomplete data objects.
- As a result, the accuracy of the discovered patterns can be poor.
- Data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases.

Pattern evaluation - the interestingness problem

- A data mining system can uncover thousands of patterns.
- Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack of novelty.
- Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations.
- The use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

Performance issues

Efficiency and scalability of data mining algorithms:

- To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.
- The running time of a data mining algorithm must be predictable and acceptable in large databases.

- From a database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems.

Parallel, distributed, and incremental mining algorithms:

- The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel.
- The results from the partitions are then merged.
- Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "from scratch." Such algorithms perform knowledge modification incrementally to amend and strengthen what was previously discovered.

Issues relating to the diversity of database types

Handling of relational and complex types of data:

- Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.
- However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining.
- Specific data mining systems should be constructed for mining specific kinds of data.
- Therefore, one may expect to have different data mining systems for different kinds of data.

Mining information from heterogeneous databases and global information systems

- Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases.
- The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining.
- Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases.
- Web mining, which uncovers interesting knowledge about Web contents, Web structures, Web usage, and Web dynamics, becomes a very challenging and fast-evolving field in data mining


Best
Wishes.
The MAHE
University Building
manipal.edu
search
The Healthsciences
Library