## Unit 4: New database applications and environments

### What Is Data Mining?

- Extraction of interesting knowledge or patterns of knowledge from huge amount of data
- Alternative names: Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

It is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. Even the term data mining does not really present all the major components in the picture. To refer to the mining of gold from rocks or sand, we say gold mining instead of rock or sand mining. Analogously, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. However, the shorter term, knowledge mining 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. Thus, such a misnomer carrying both "data" and "mining" became a popular choice.

In addition, many other terms have a similar meaning to data mining—for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

The knowledge discovery process:

- **Data cleaning** (to remove noise and inconsistent data)
- **Data integration** (where multiple data sources may be combined) A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step, where the resulting data are stored in a data warehouse.
- **Data selection** (where data relevant to the analysis task are retrieved from the database)
- **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations) Sometimes data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing. Data reduction may also be performed to obtain a smaller representation of the original data without sacrificing its integrity.
- **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures)
- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

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

The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in

media, and in the research milieu, the term data mining is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data). Therefore, we adopt a broad view of data mining functionality: Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

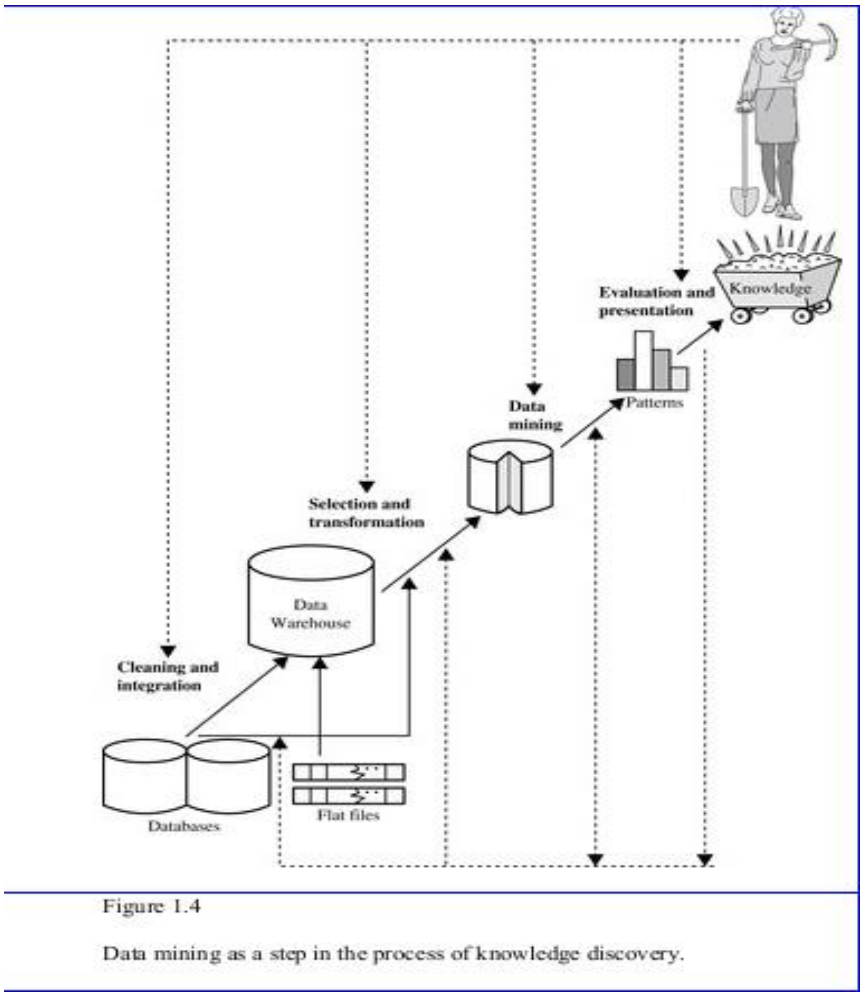



Figure 1.4

Data mining as a step in the process of knowledge discovery.

**Data Mining Functionalities**

- Multidimensional concept description: Characterization and discrimination
- Frequent patterns, association
- Classification and prediction
- Cluster analysis
- Outlier analysis
- Trend and evolution analysis

**Data Association:** combination/coordination between two

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

Single Dimension Association Rule
where X is a variable representing a customer.

-> if a customer buys a computer, there is a 50% chance that he will buy software as well

-> 1% of all the transactions under analysis show that computer and software are purchased together

age (X, “20..29”)^income(X, “40K..49K”)=>buys(X, “laptop”)

Multi-Dimension Association Rule [support = 2%, confidence = 60%]

Association rules are discarded as uninteresting if they do not satisfy minimum support threshold and minimum confidence threshold

- 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 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.

## Classification and prediction

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

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

〉 Predict some unknown or missing numerical values

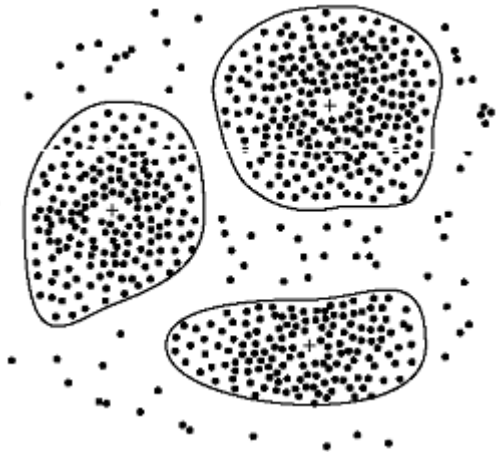
〉 E.g., classify countries based on (climate)

**Cluster Analysis**

Grouping similar instances into clusters

Clustering can be used to generate class labels for a group of data which did not exist at the beginning.

A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster "center" is marked with a "+".



**Applications of data mining**

Data Mining is primarily used today by companies with a strong consumer focus — retail, financial, communication, and marketing organizations, to "drill down" into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments.

Other important areas where data mining is widely used:

**Future Healthcare**

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

**Market Basket Analysis**

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

**Education**
There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

**Manufacturing Engineering**
Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

**CRM**
Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

**Fraud Detection**
Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

**Lie Detection**
Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

**Financial Banking**
With computerized banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

**Corporate Surveillance**

Corporate surveillance is the monitoring of a person or group's behaviour by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

**Research Analysis**

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualization and visual data mining provide us with a clear view of the data.

**Criminal Investigation**

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

**Bio Informatics**

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

**Data Warehousing: Overview of data warehousing**

**What is Data Warehouse?**

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.
- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.

- A decision support database that is maintained separately from the organization's operational database
- Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a **subject-oriented, integrated, time-variant, and nonvolatile** collection of data in support of management's decision-making process."—W. H. Inmon

Data Warehouse Features

**Subject-Oriented:**

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

**Integrated:**

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

**Time Variant**

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

**Nonvolatile:**

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*

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

process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity). 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 modeled 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 pre-computation and fast access of summarized data.

**Typical functionality of a data warehouse**

A data warehouse helps business executives to organize, analyze, and use their data for decision making. A data warehouse serves as a sole part of a plan-execute-assess "closed-loop" feedback system for the enterprise management. Data warehouses are widely used in the following fields:

- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing

The following are the functions of data warehouse tools and utilities:

- **Data Extraction** - Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** - Involves finding and correcting the errors in data.
- **Data Transformation** - Involves converting the data from legacy format to warehouse format.
- **Data Loading** - Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** - Involves updating from data sources to warehouse.

**Note**: Data cleaning and data transformation are important steps in improving the quality of data and data mining results.

| S.N | Data Warehouse (OLAP) | Operational Database(OLTP) |
|---|---|---|
| | **OLAP (on-line analytical processing)** | **OLTP (on-line transaction processing)** |
| 1 | It involves historical processing of information. | It involves day-to-day processing. |

| | | |
|---|---|---|
| 2 | OLAP systems are used by knowledge workers such as executives, managers, and analysts. | OLTP systems are used by clerks, DBAs, or database professionals. |
| 3 | It is used to analyze the business. | It is used to run the business. |
| 4 | It focuses on Information out. | It focuses on Data in. |
| 5 | It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema. | It is based on Entity Relationship Model. |
| 6 | It focuses on Information out. | It is application oriented. |
| 7 | It contains historical data. | It contains current data. |
| 8 | It provides summarized and consolidated data. | It provides primitive and highly detailed data. |
| 9 | It provides summarized and multidimensional view of data. | It provides detailed and flat relational view of data. |
| 10 | The number of users is in hundreds. | The number of users is in thousands. |
| 11 | The number of records accessed is in millions. | The number of records accessed is in tens. |
| 12 | The database size is from 100GB to 100 TB. | The database size is from 100 MB to 100 GB. |
| 13 | These are highly flexible. | It provides high performance. |