Data Mining

(Professional Elective Course – IV)

Course objectives:

1. To introduce students to the basic concepts and techniques of Data Mining.
2. To develop skills of using recent data mining software for solving practical problems.
3. To gain experience of doing independent study and research.
4. To study the methodology of engineering legacy databases for data warehousing and data mining to derive business rules for decision support systems.
5. Develop and apply critical thinking, problem-solving, and decision-making skills.

Course outcomes

1. Understand Data Warehouse fundamentals, Data Mining Principles.
2. Describe different steps in data preprocessing used for data mining.
3. Characterize the kinds of patterns that can be discovered by mining.
4. Apply different data-mining technique for classification of data.
5. Categorize and carefully differentiate between cluster and outlier analysis.

Unit I

- **Introduction:**
- What Is a Data Warehouse?
- Differences between Operational Database Systems and Data Warehouses
- Why Have a Separate Data Warehouse?
- What Is Data Mining?
- What Kinds of Patterns Can Be Mined?: Class/Concept Description: Characterization and Discrimination, Mining Frequent Patterns, Associations, and Correlations, Classification and Regression for Predictive Analysis, Outlier Analysis
- Major Issues in Data Mining: Mining Methodology, User Interaction, Efficiency and Scalability, Diversity of Database Types, Data Mining and Society.

Definition of a Data Warehouse

- "A data warehouse is an enterprise structured repository of subject-oriented, time-variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data." - *Oracle Data Warehouse Method*

- “a subject-oriented, integrated, time variant and non-volatile collection of data used in strategic decision making” [Imnon, 1980]

What is Data Warehouse?

- Defined in many different ways, but not rigorously.
  - 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 warehousing:
  - The process of constructing and using data warehouses

Data Warehouse—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.

Data Warehouse—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.

Data Warehouse—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".

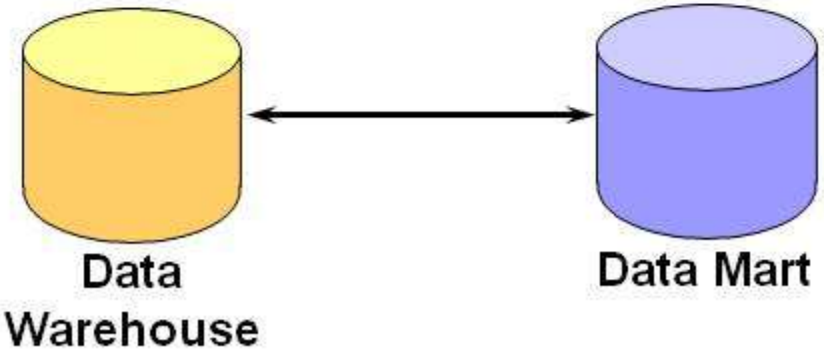
Data Warehouse—Non-Volatile

- 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*.

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Data Warehouse Compared to Data Mart

Property	Data Warehouse	Data Mart
Scope	Enterprise	Department
Subjects	Multiple	Single-subject, line of business (LOB)
Data Source	Many	Few
Size (typical)	100 GB to > 1 TB	< 100 GB
Implementation time	Months to years	Months



Why separate data warehouse

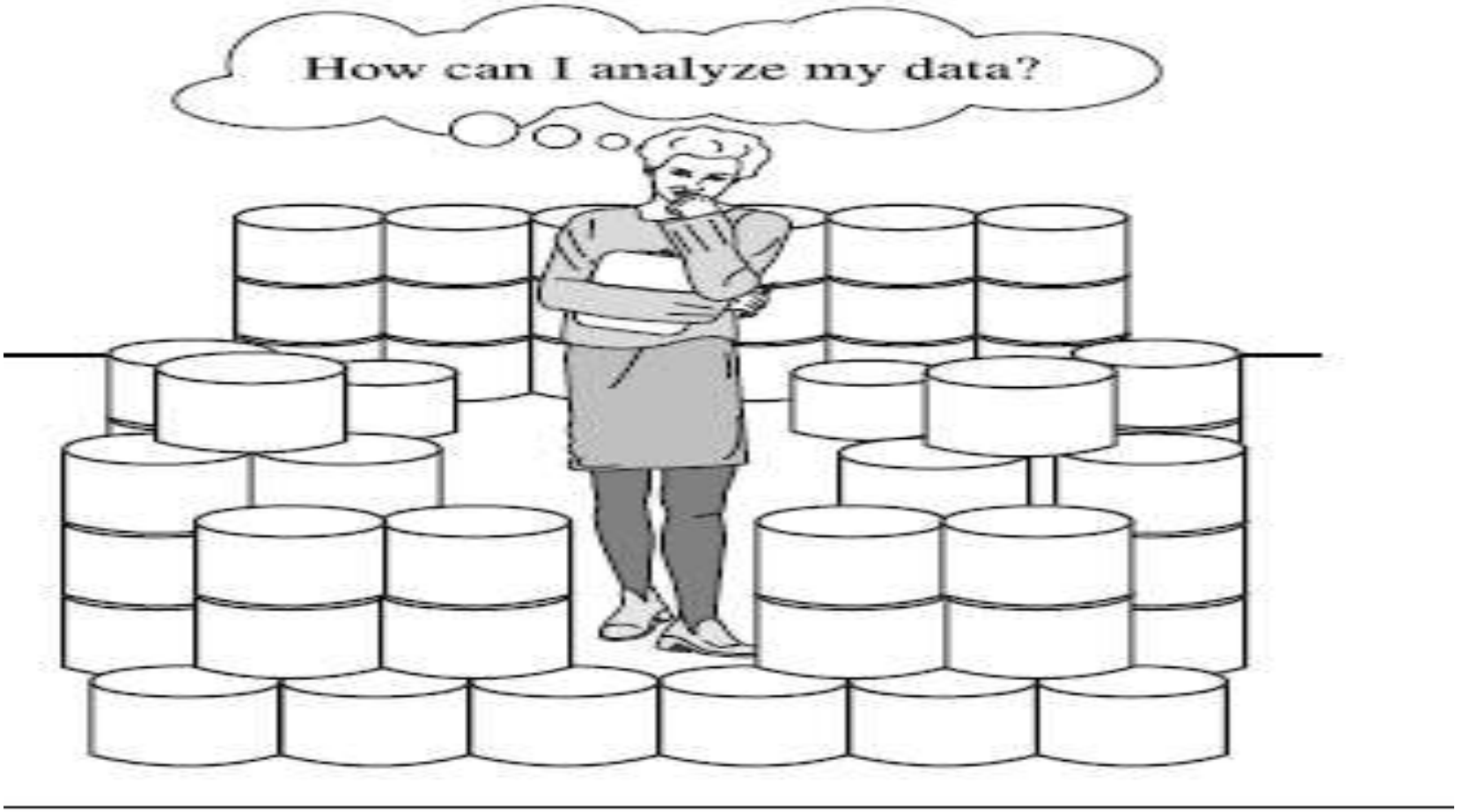
1.High performance

2.OLAP involves large complex queries.

3.Concurrency control and recovery mechanism is not required.

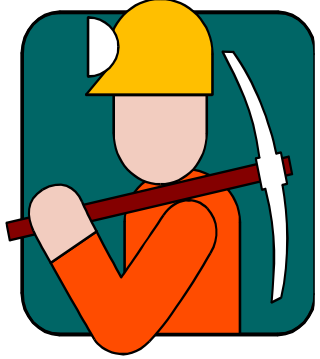
4.Maintain historical data.

Need to find information



We are data rich, but information poor.

What Is Data Mining?



- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or 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.

What is data Mining ?

- Data Mining is the extraction (mining) of knowledge from large amount of data , mining of gold from rocks or sand is called as gold mining not rock mining or sand mining , so the proper name of data mining is —Knowledge Mining‖ from data. But today's business trends calls it data mining.

What is data Mining ?

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.

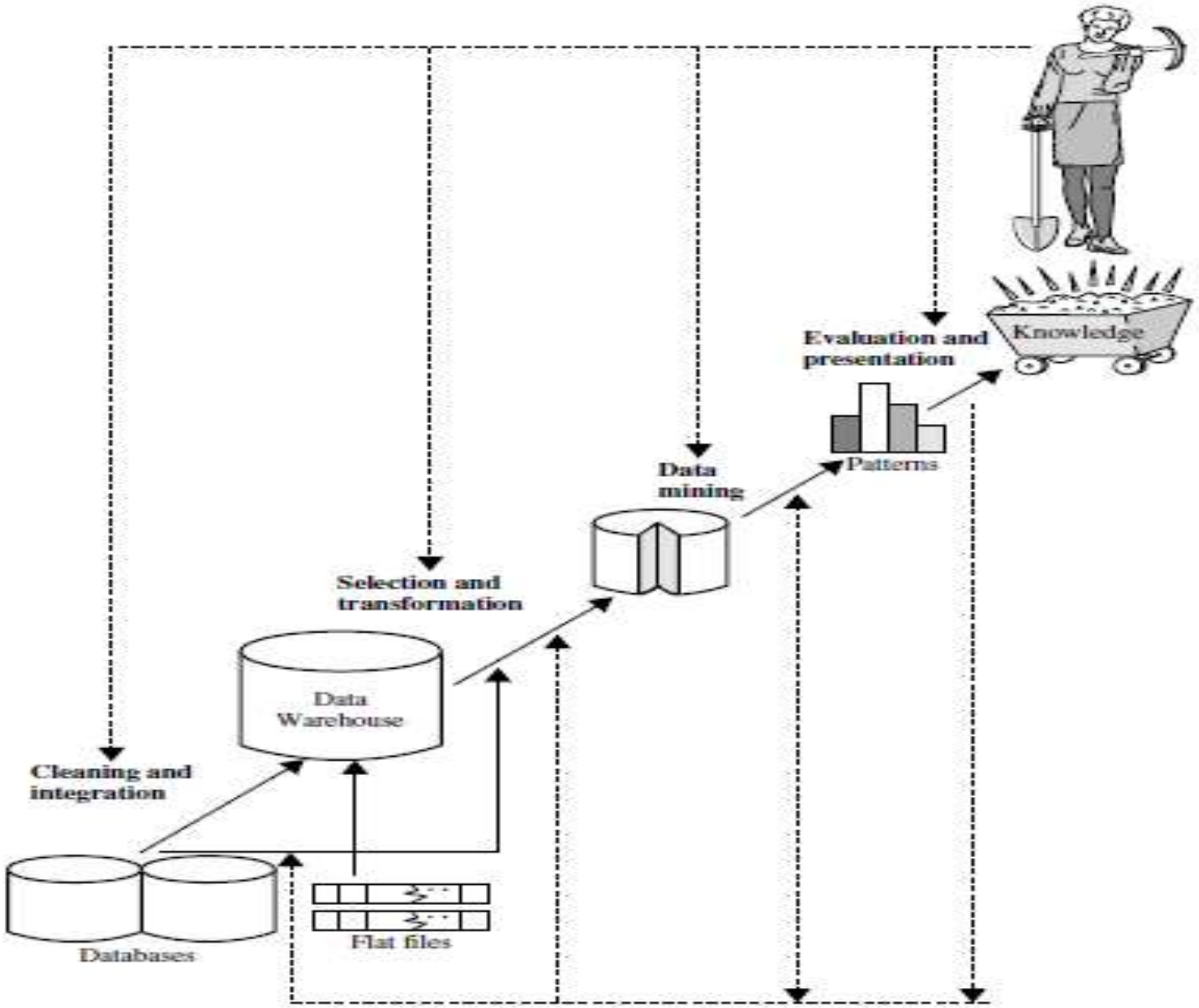
What is Data Mining?

- Efficient automated discovery of previously unknown patterns in large volumes of data.
- Patterns must be valid, novel, useful and understandable.
- Businesses are mostly interested in discovering past patterns to predict future behaviour.

Necessity Is the Mother of Invention

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, and other information repositories
- We are drowning in data, but starving for knowledge!
- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing
  - Miing interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

KDD Process


Evaluation and presentation
Knowledge
Data mining
Patterns
Selection and transformation
Data Warehouse
Cleaning and integration
Databases
Flat files

Steps in KDD

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)

Data Mining Functionalities

Used to specify what kind of patterns to be found in data mining tasks

Data mining task -

Descriptive

Predictive

What Kinds of Patterns Can Be Mined? (Data Mining Functionalities) (1)

- Class/Concept Description: Characterization and Discrimination:

  Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions

- Mining Frequent Patterns, Associations, and Correlations

  buys(X, "computer") => buys(X, "software") [support = 1%, confidence = 50%]

- Classification and Regression for Predictive Analysis

  Construct models (functions) that describe and distinguish classes or concepts for future prediction

  - E.g., classify countries based on (climate), or classify cars based on (gas mileage)

  – Predict some unknown or missing numerical values

What Kinds of Patterns Can Be Mined? (Data Mining Functionalities) (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis

Major Issues in Data Mining

A) Mining methodology :

1) Mining various and new kinds of knowledge.
2) Mining knowledge in multidimensional space
3) Data mining—an interdisciplinary effort
4) Boosting the power of discovery in a networked environment
5) Handling uncertainty, noise, or incompleteness of data
6) Pattern evaluation and pattern- or constraint-guided mining

B) User Interaction

1) Interactive mining
2) Incorporation of background knowledge
3) Ad hoc data mining and data mining query languages
4) Presentation and visualization of data mining results

Major Issues in Data Mining

B) Efficiency and Scalability

1) Efficiency and scalability of data mining algorithms
2) Parallel, distributed, and incremental mining algorithms

C) Diversity of Database Types

1) Handling complex types of data
2) Mining dynamic, networked, and global data repositories

D)Data Mining and Society

1) Social impacts of data mining
2) Privacy-preserving data mining
3) Invisible data mining