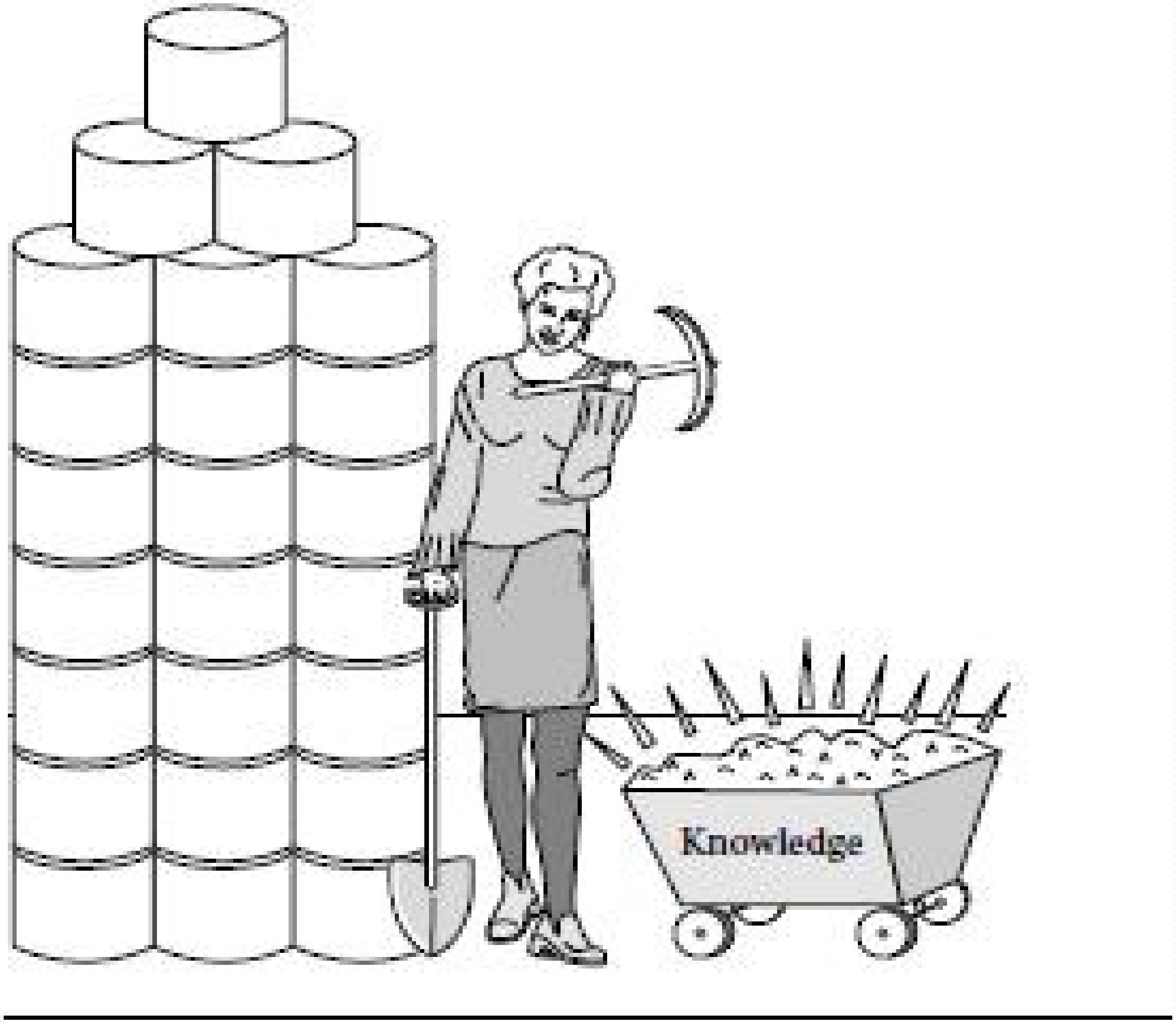
Data Mining and Data Warehousing

By :
Dr. Rinkle Rani
Associate Professor, CSED
TIET, Patiala

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- Data mining—Automated analysis of massive data sets

How can I analyze my data?



Knowledge

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
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Data Mining is:

(1) The efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets

(2) The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner

Examples of Data mining Applications

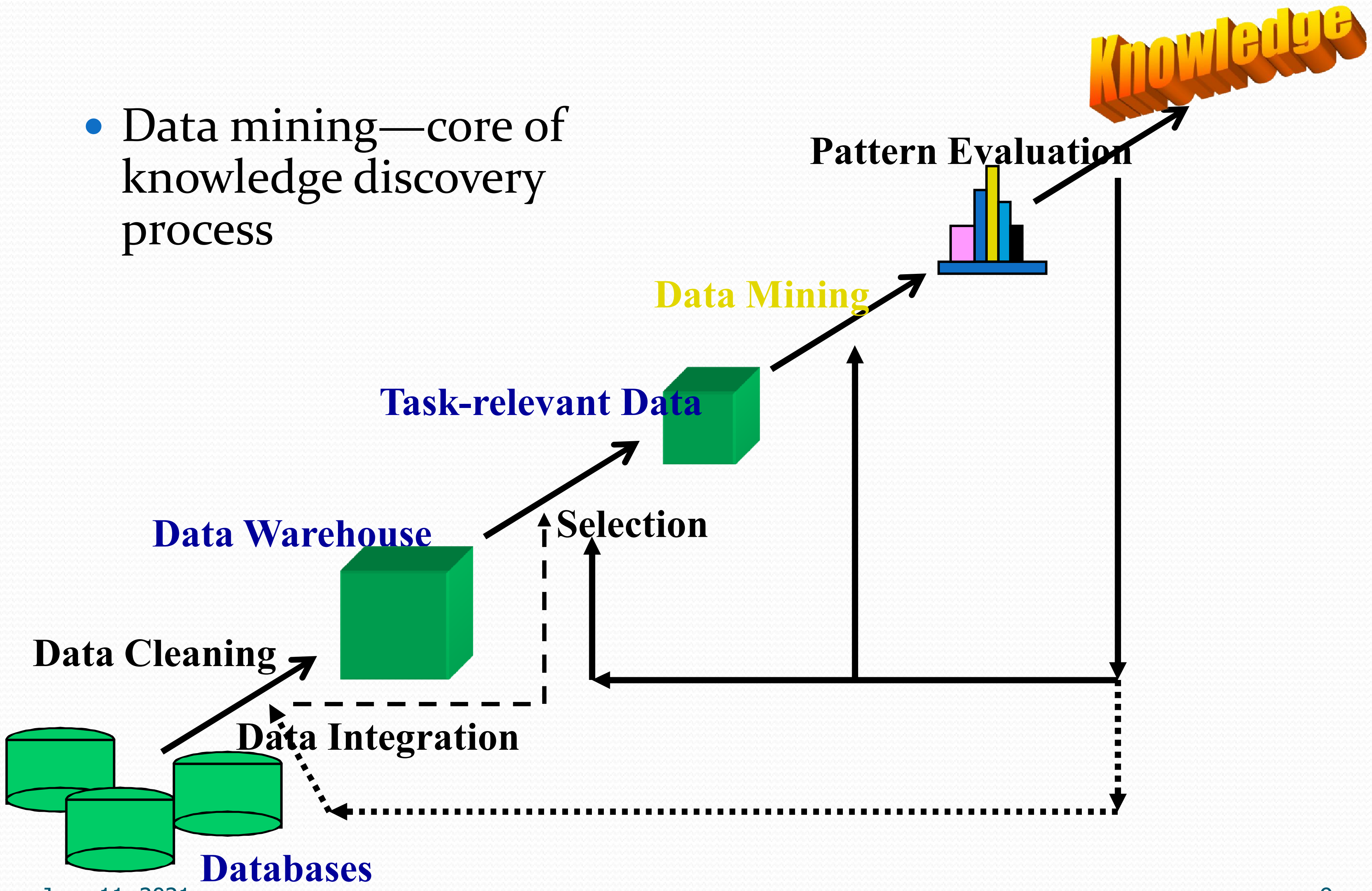
1. Fraud detection: credit cards, phone cards
2. Marketing: customer targeting
3. Data Warehousing: Walmart
4. Astronomy
5. Molecular biology

How Data Mining is used

1. Identify the problem
2. Use data mining techniques to transform the data into information
3. Act on the information
4. Measure the results

Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



What is KDD?

KDD is referred to as Knowledge Discovery in Database and is defined as a method of finding, transforming, and refining meaningful data and patterns from a raw database in order to be utilized in different domains or applications.

KDD Process may consist of the following steps :-

1 Data cleaning -

First step in the Knowledge Discovery Process is Data cleaning in which noise and inconsistent data is removed.

2 Data Integration -

Second step is Data Integration in which multiple data sources are combined.

3 Data Selection -

Next step is Data Selection in which data relevant to the analysis task are retrieved from the database.

4 Data Transformation -

In Data Transformation, data are transformed into forms appropriate for mining by performing summary or aggregation operations.

5 Data Mining -

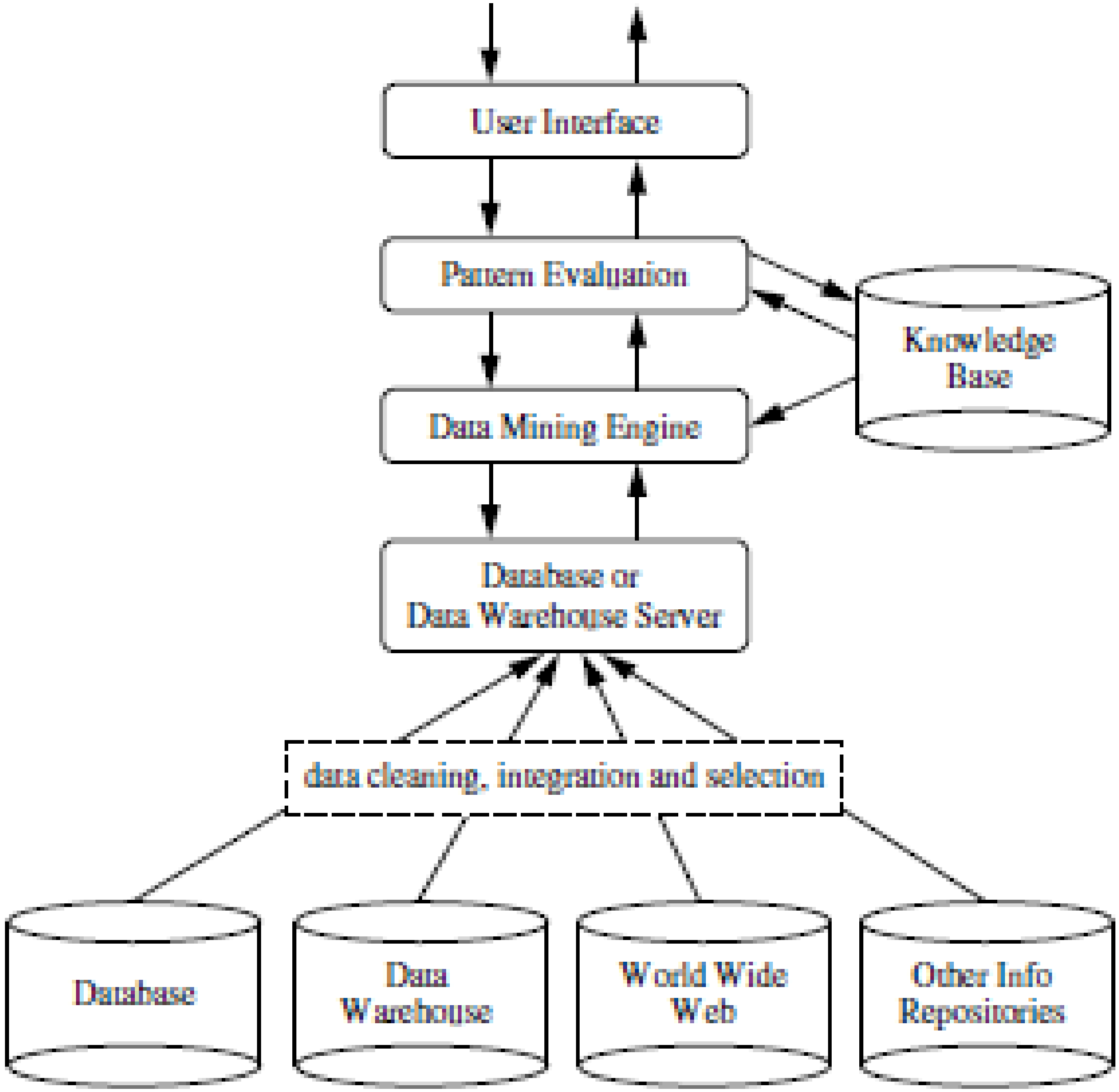
In Data Mining, data mining methods (algorithms) are applied in order to extract data patterns.

6 Pattern Evaluation -

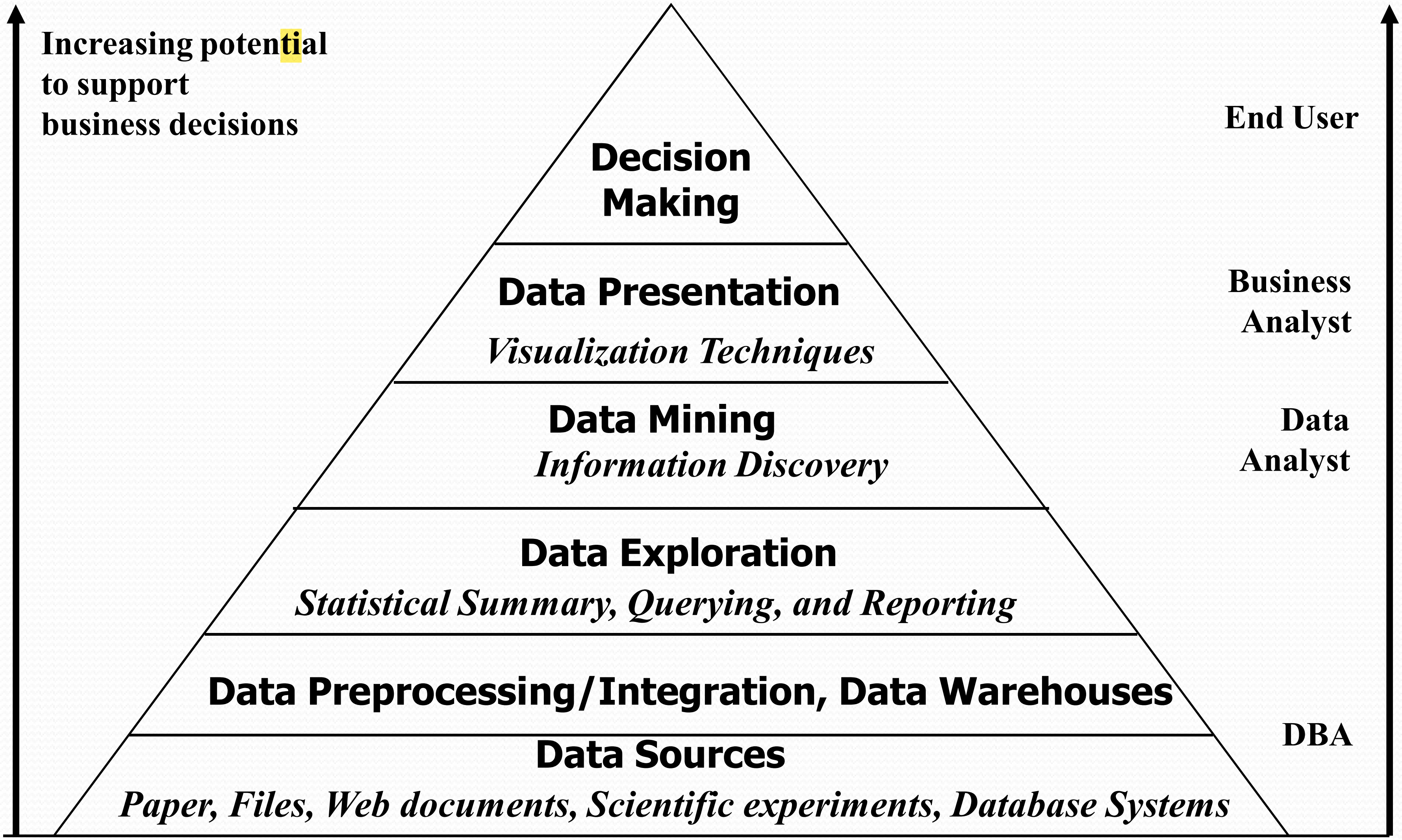
In Pattern Evaluation, data patterns are identified based on some interesting measures.

7 Knowledge Presentation -

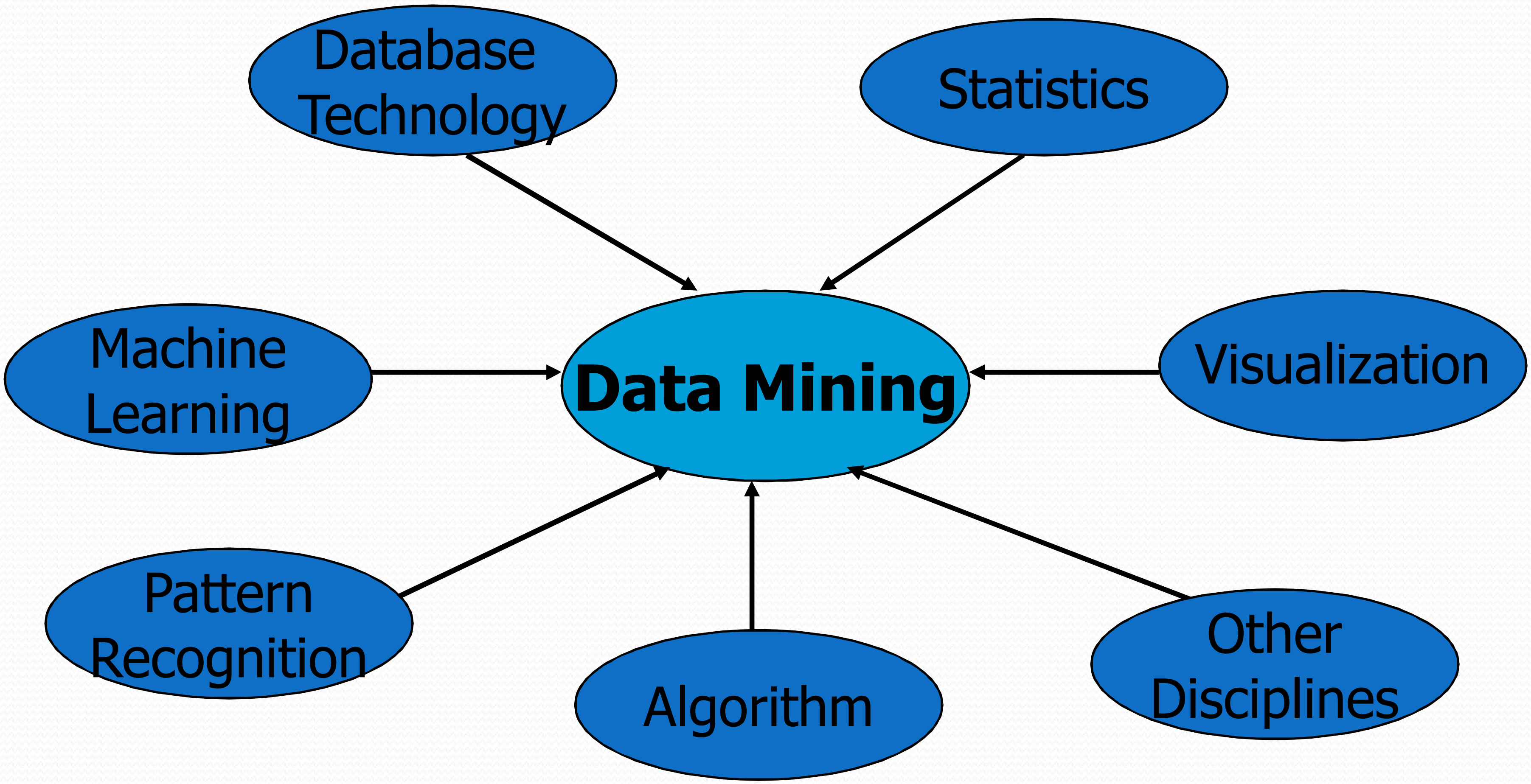
In Knowledge Presentation, knowledge is represented to user using many knowledge representation techniques.


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

Data Mining and Business Intelligence


Increasing potential
to support
business decisions
Decision
Making
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems
End User
Business
Analyst
Data
Analyst
DBA

Data Mining: Confluence of Multiple Disciplines



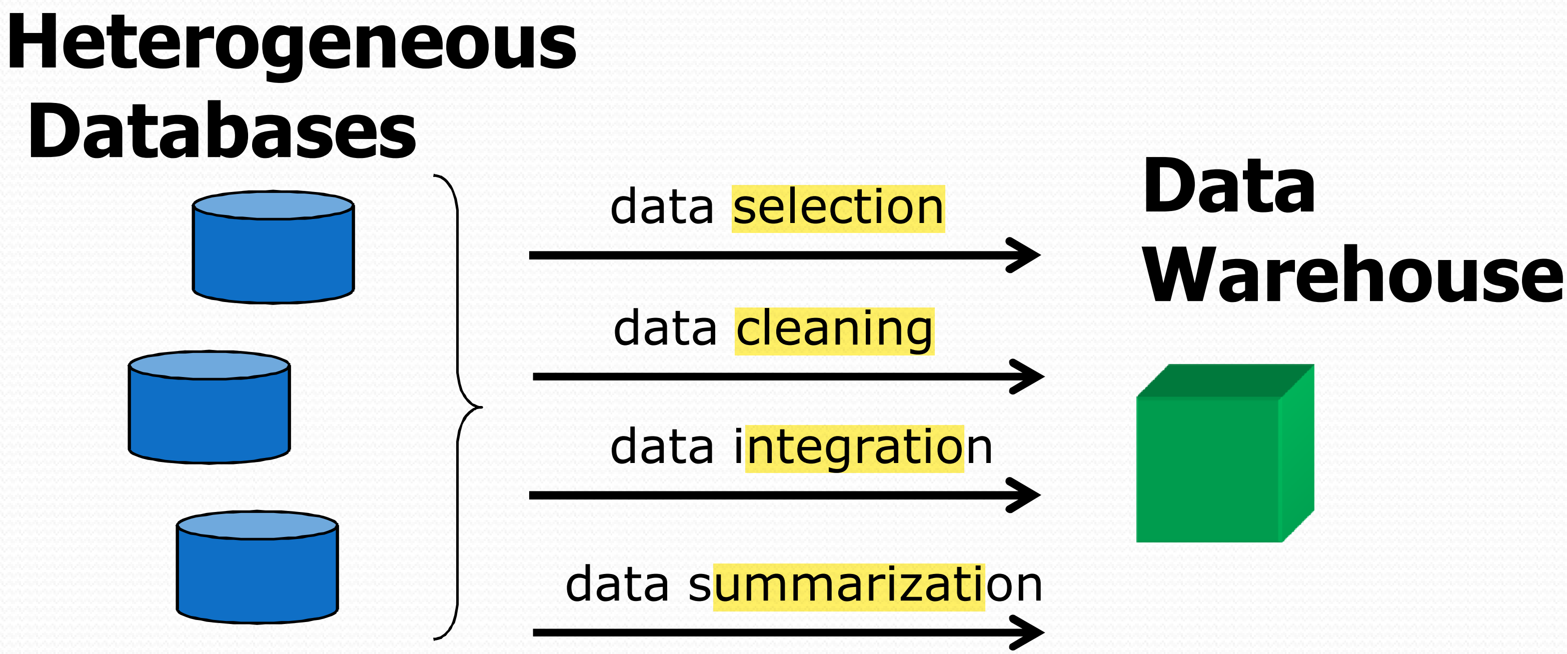
Data Mining Tasks

1. Classification: learning a function that maps an item into one of a set of predefined classes
2. Regression: learning a function that maps an item to a real value
3. Clustering: identify a set of groups of similar items
4. Dependencies and associations:

   identify significant dependencies between data attributes
5. Summarization: find a compact description of the dataset or a subset of the dataset

Why Data Warehousing?

- Data warehousing can be considered as an important preprocessing step for data mining



Heterogeneous Databases

data selection

data cleaning

data integration

data summarization

Data Warehouse

- A data warehouse also provides on-line analytical processing (OLAP) tools for interactive multidimensional data analysis.

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 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" (the time elements could be extracted from log files of transactions)

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*.

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making

OLTP vs. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	manager
function	day to day operations	Decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB (even PB)
metric	transaction throughput	query throughput, response