Data Warehousing and Mining

— Module 1.2—
Introduction to Data Mining

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, sale transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- <u>We are drowning in data, but starving for knowledge!</u>
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Evolution of Sciences

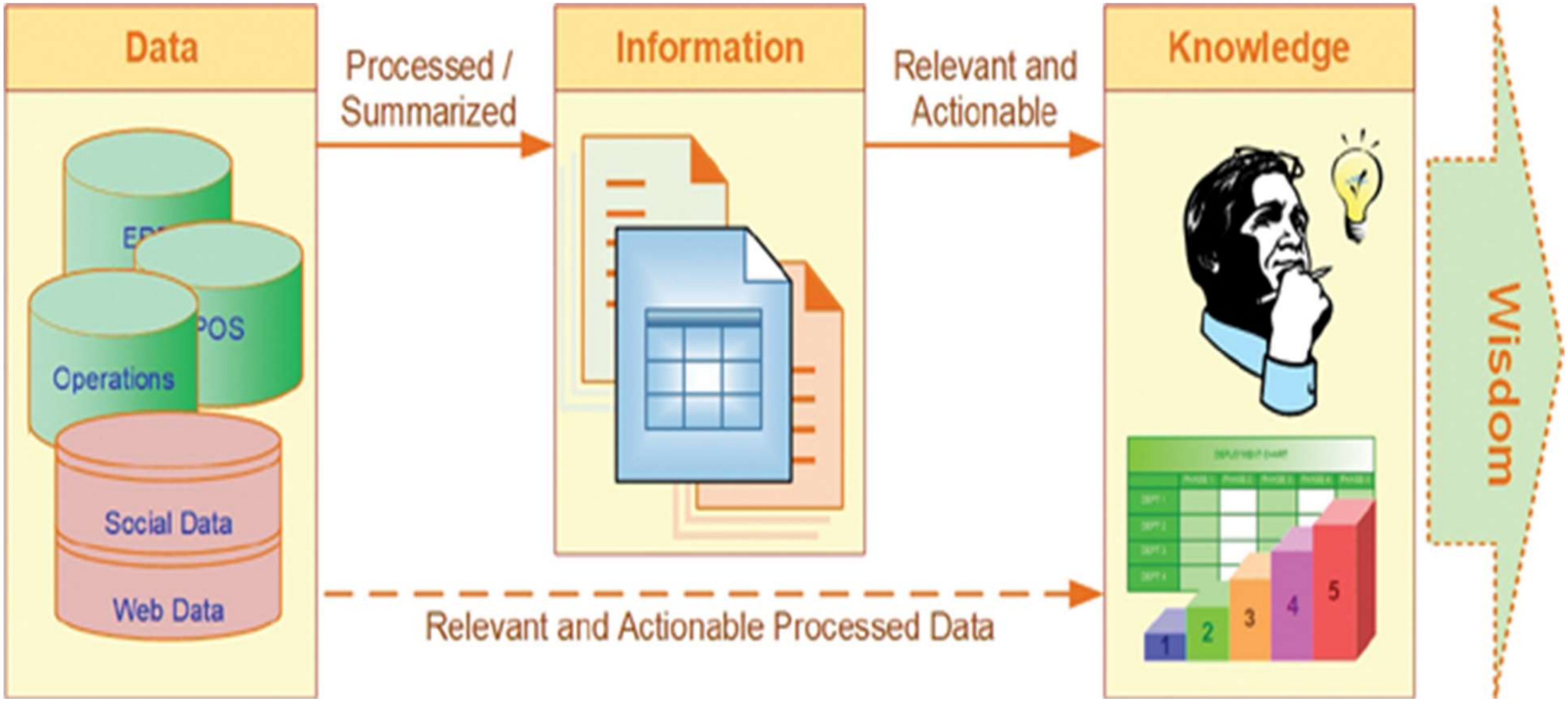
- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

What Is Data Mining?

- **Data mining is the process of converting data into information and then into knowledge.**
- Knowledge is very distinct from data and information
- Knowledge is information that is contextual, relevant, and actionable.
- knowledge has strong experiential and reflective elements that distinguish it from information in a given context.



What Is Data Mining?

- Data mining is a process that involves **using statistical, mathematical, and artificial intelligence techniques and algorithms to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data.**

Fayyad et al. (1996) defined data mining as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases,"

Ø **Process** implies that data mining comprises many iterative steps.

Ø **Nontrivial** means that some experimentation-type search or inference is involved;

Ø **Valid** means that the discovered patterns should hold true on new data with a sufficient degree of certainty.

Ø **Novel** means that the patterns were not previously known to the user in the context of the system being analyzed.

Ø **Potentially useful** means that the discovered patterns should lead to some benefit to the user or task.

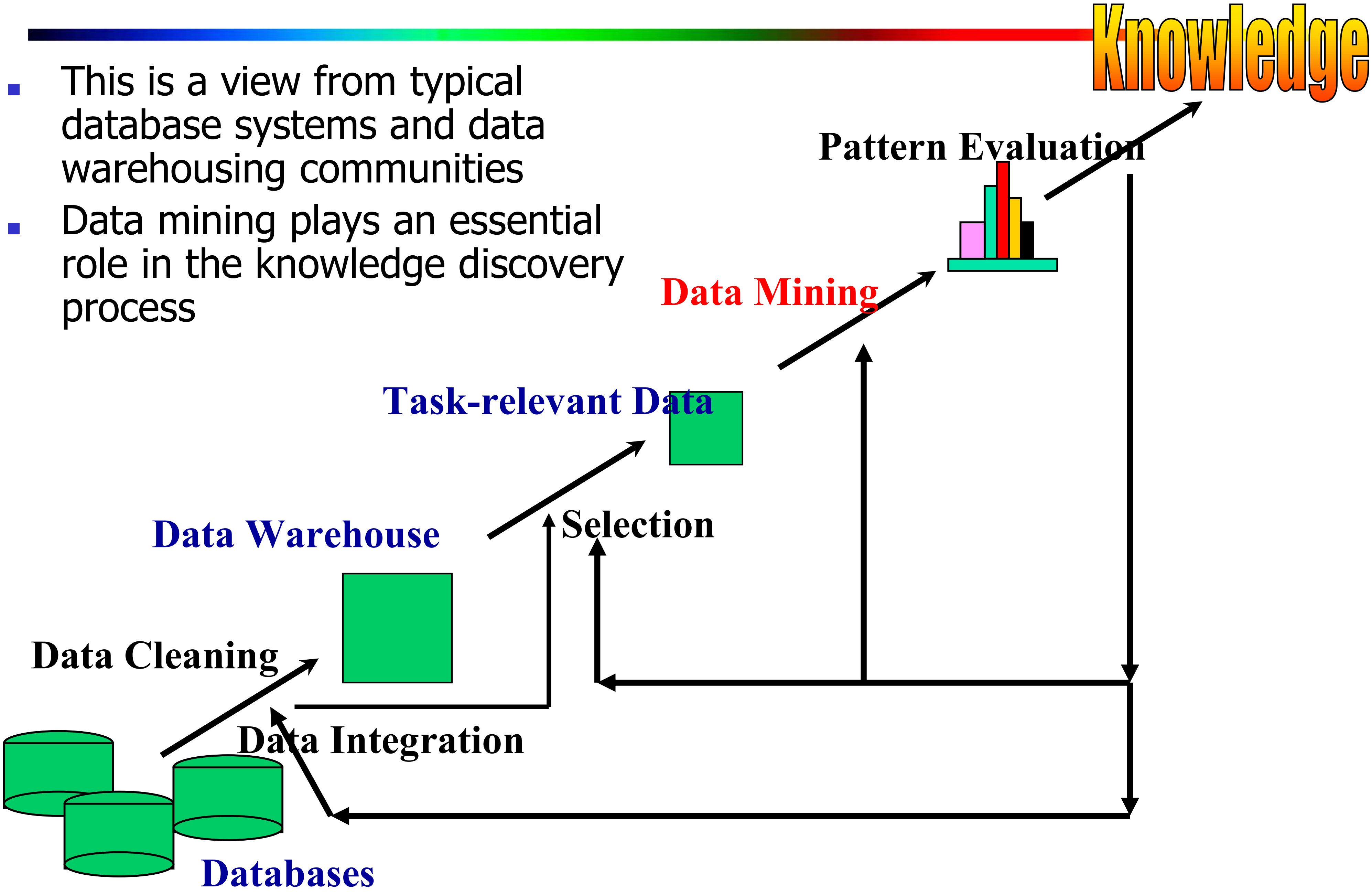
Ø **Ultimately understandable** means that the pattern should make business sense that leads to users saying "This makes sense. Why didn't I think of that?"—if not immediately at least after some processing.

Data Mining: Alternative names

- Alternative names
  - Knowledge discovery (mining) in databases (KDD),
  - knowledge extraction,
  - data/pattern analysis,
  - data archeology, data dredging,
  - information harvesting,
  - business intelligence, etc.

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Spatial data
  - Multimedia database
  - Text databases
  - The World-Wide Web

What Kinds of Patterns Can Be Mined?

- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

  - Descriptive mining tasks : Deals with the General characteristics and converts them into relevant and useful information

  - Predictive mining tasks: Predicts future values by analyzing data patterns and their outcomes based on past data.

Descriptive DM Functionalities

1. **Class/Concept Description:**

❑ Data entries can be associated with the classes or concepts.

❑ These descriptions can be derived using

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

- *Example: At electronic store a Customer relationship manager asks to Summarize the characteristics of customers who spend more than Rs.10000 a year at the store.*

or

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

- *Example: A customer relationship manager at Electronics store may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g. less than three times a year).*

or

(3) both data characterization and discrimination.

Descriptive DM Functionalities

2. Mining of frequent patterns:

- Patterns that occur frequently in data.
- It includes--
  - **Frequent itemset** : refers to a set of items that often appear together in a transactional data set
  - **Frequent subsequences (also known as sequential patterns):** A frequently occurring subsequence like laptop→ digital camera→ memory card
  - **Frequent substructures:**
  - A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that 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.

Descriptive DM Functionalities

3. Association Analysis

- Defines relationships between the data and predefined association rules.
- Suppose that, as a marketing manager at *Electronics store*, you want to know which items are frequently purchased together
- A rule, $buys(X, \text{“computer”}) \Rightarrow buys(X, \text{“software”})\ [support = 1\%, confidence = 50\%]$
- Association rules that contain a single predicate are referred to as **single-dimensional association rules**.
- Suppose, instead, that we are given the *Electronics* relational database related to purchases. A data mining system may find association rules like

$$age(X, \text{“20..29”}) \wedge income(X, \text{“40K..49K”}) \Rightarrow buys(X, \text{“laptop”})$$
$$[support = 2\%, confidence = 60\%].$$

- Association rules that contain a more than one predicate/attributes are referred to as **multi-dimensional association rules**.

Descriptive DM Functionalities

4. Clustering : can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity*.

i.e. clusters of objects are formed so that objects within a cluster have high similarity , but are rather dissimilar to objects in other clusters.

- ❑ Clustering can also facilitate taxonomy formation → Organization of observations into a hierarchy of classes that group similar events together.
- ❑ Example: Cluster analysis can be performed on *Electronics store* customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing

Predictive Data mining functionalities

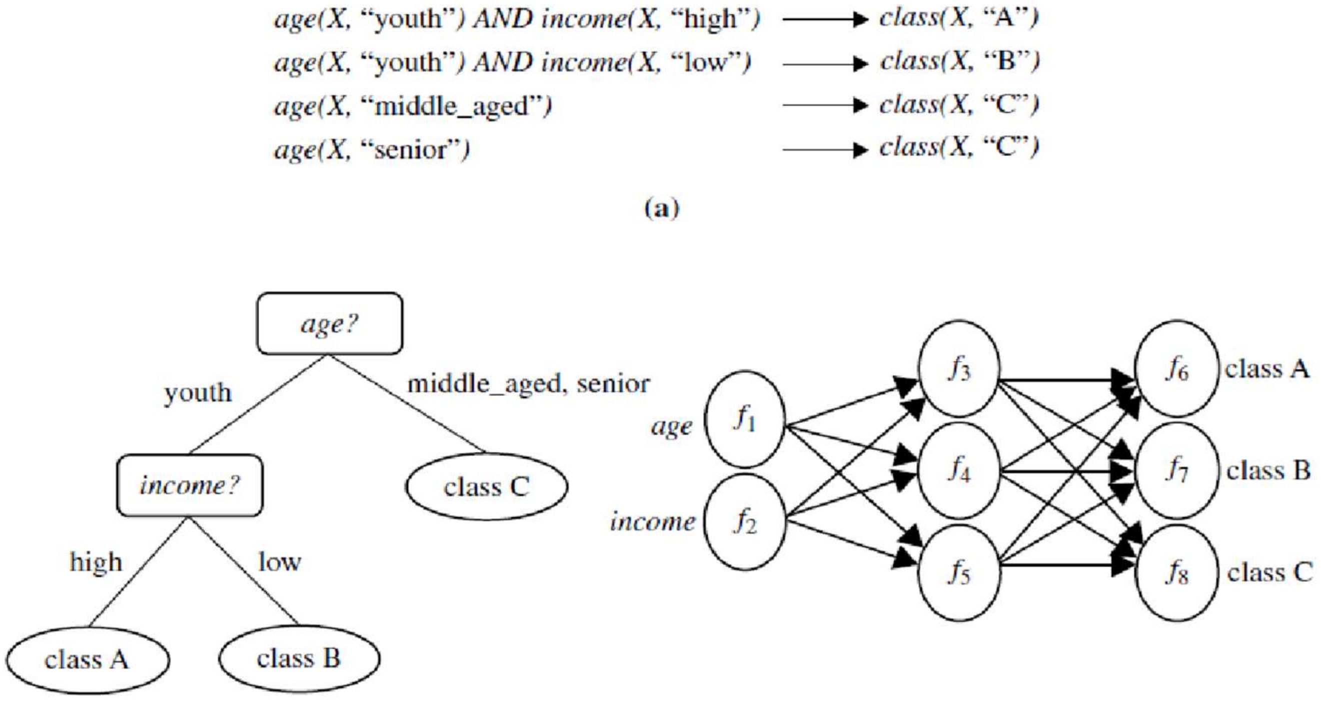
➢ Predicts future values by analyzing data patterns and their outcomes based on past data.

•Classification
•Regression
•Outlier analysis

Predictive Data mining functionalities

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

- The models are derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known).
- The model is used to predict the class label of objects for which the class label is unknown.
- The derived model may be represented in various forms, such as *classification rules* (i.e., *IF-THEN rules*), *decision trees*, *mathematical formulae, or neural networks*



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

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

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

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

(a)

(b) **(c)**

Predictive Data mining functionalities

2. **Regression:**

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

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

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

- Regression also encompasses the identification of distribution *trends* based on the available data.

Predictive Data mining functionalities

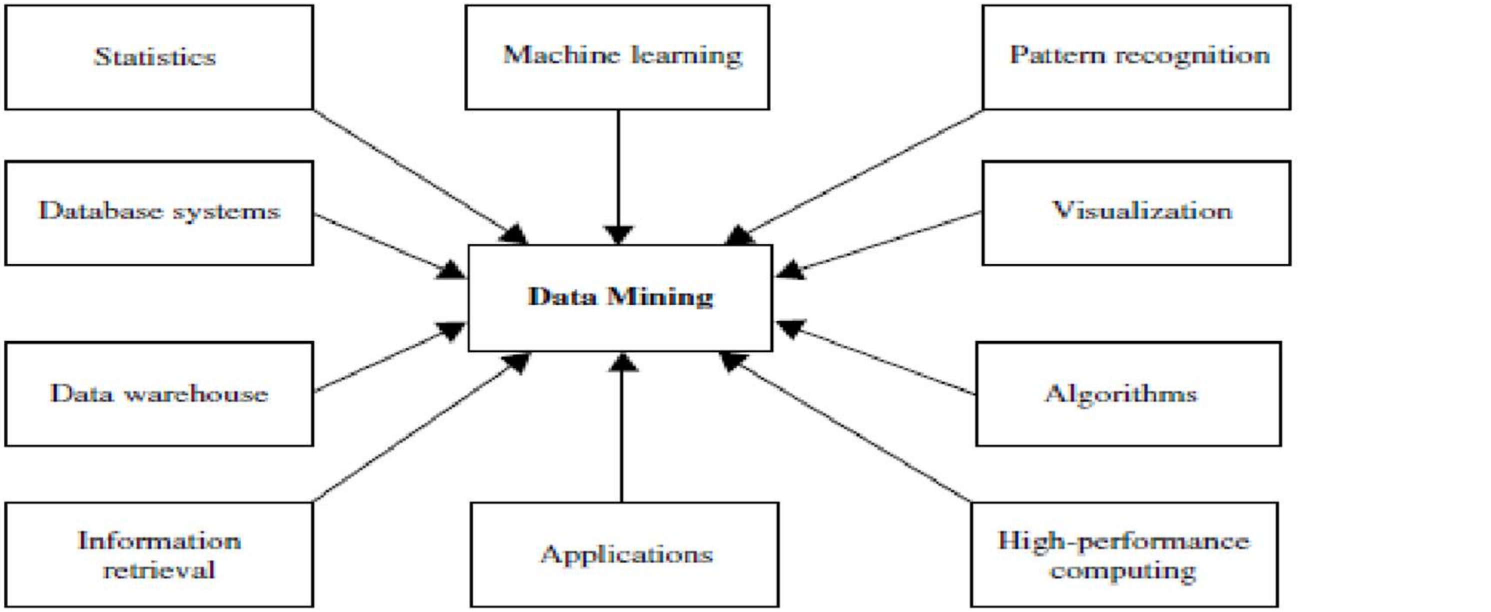
3. Outlier Analysis :

- Outlier: A data object that does not comply with the general behavior of the data
- Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection), the rare events can be more interesting than the more regularly occurring ones.
- 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 remote from any other cluster are considered outliers.
- Example: Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account.

  Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

  - It is used in observing the change in trends of buying patterns of a customer.

Technologies used in Data Mining



Data mining adopts techniques from many domains.

1. Statistics:

- It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.
- Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

Technologies used in Data Mining

2. Machine learning

- **Arthur Samuel** defined machine learning as a field of study that gives computers the ability to learn without being programmed.

- When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.

- In machine learning, an algorithm is constructed to predict the data from the available database **(Predictive analysis).**

- It is related to computational statistics.

Technologies used in Data Mining

The four types of machine learning are:

1. Supervised learning

It is based on the classification.

It is also called as **inductive learning**. In this method, the desired outputs are included in the training dataset.

2. Unsupervised learning

Unsupervised learning is based on clustering. Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.

3. Semi-supervised learning

Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions. This method generally avoids the large number of labeled examples (i.e. desired outputs) .

4. Active learning

Active learning is a powerful approach in analyzing the data efficiently.

The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

Technologies used in Data Mining

3. Database systems and data warehouses

- Databases are used for the purpose of recording the data as well as data warehousing.
- Online Transactional Processing (OLTP) uses databases for day to day transaction
  purpose.
- To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.
- **E**ntity-**R**elational modeling techniques are used for relational database management system design.
- Data warehouses are used to store historical data which helps to take strategical decision for business.
- It is used for online analytical processing (OALP), which helps to analyze the data.

Technologies used in Data Mining

4. Information retrieval

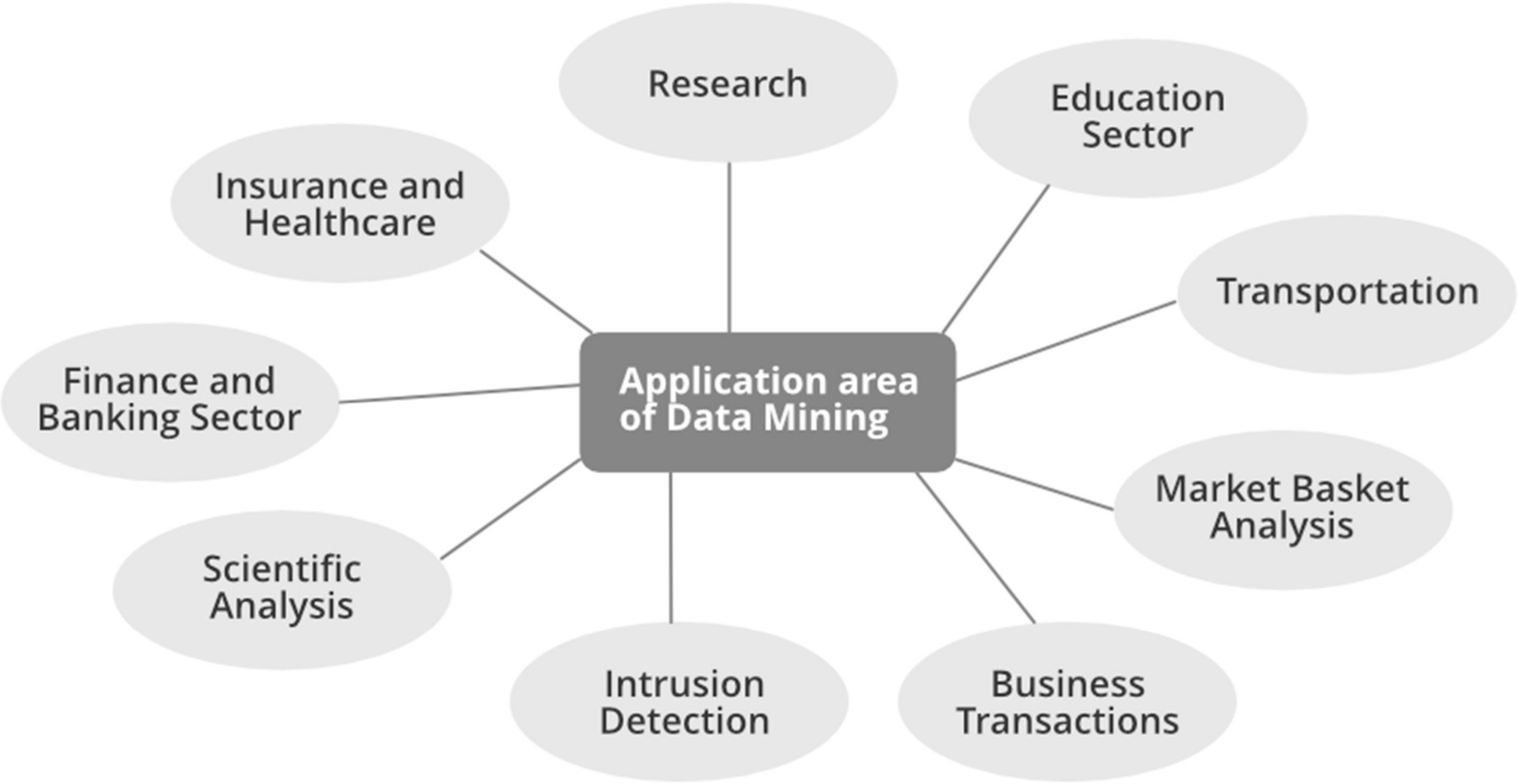
Information deals with uncertain representations of the semantics of objects (text, images).
For example: Finding relevant information from a large document.

Technologies used in Data Mining

5. Decision support system

- Decision support system is a category of information system. It is very useful in decision making for organizations.
- It is an interactive software based system which helps decision makers to extract useful information from the data, documents to make the decision.

Data Mining: Applications



Data Mining: Applications

1. **Marketing and CRM:**
   -To identify most **likely buyers of new products**
   -To identify root causes of customer attrition so as to improve customer retention
   -To discover time variant associations between products and services to maximize sale and find most profitable customers.
1. **Banking and Finance:**
   -To **detect fraudulent credit card** and online banking transactions
   -To optimize the cash return by forecasting cash flow on banking entities
   -To streamline and automate the processing of loan applications by accurately predicting most probable defaulters.
   -To maximize the customer value by identifying and selling the products and services that customers are most likely to buy.
1. **Retailing and Logistics:**
   -To identify accurate sales volume at specific retail locations in order to determine correct inventory levels.
   -To do MBA to improve store layout and optimize sales promotions
   -To forecast consumption levels for different product types.
   -To discover interesting patterns in the movement of products in a supply chain by analyzing sensory and RFID data.

Data Mining: Applications

4. **Manufacturing:**
   - To **predict machine failures** using sensory data
   - To discover novel patterns to identify and improve product quality.
5. **Brokerages and Security Tradings:**
   - To predict when and how much certain stock / bond prices will change.
   - To forecast range of market fluctuations,direction of fluctuations
   - To assess effect of particular issues/events on market movements.
   - To identify and prevent fraudulent activities in security trading.
6. **Insurance:**
   - To predict which **customers will buy new policies**
   - Identify fraudulent behavior of customers
   - Prevent incorrect claim payments
7. **Computer Hardware and Software:**
   - To **predict disk failure**
   - To identify and filter unwanted web contents and email messages
   - To identify potentially unsecured software products
8. **Government and Defense:**
   - To forecast the cost of moving military personnel and equipments.
   - To predict resource consumption for better planning and budgeting

Data Mining: Applications

9. Travel and Lodging:

- To predict sales of different services to optimally price these services.

- To forecast demand at different locations to better allocate limited organizational resources. .

- To identify most profitable customers and provide them with personalised services.

- To retain valuable employees by identifying and acting on the root causes for attrition

9. Health and Healthcare:

- To identify successful **medical therapies for different illnesses.**

- To identify people without health insurance and reasons behind it.

- To forecast the time of demand at different service locations to optimally allocate organizational resources.

- To retain valuable employees by identifying root causes for attrition

9. Entertainment:

-To analyze viewer data to determine which programs to show during prime time.

-To decide where to insert advertisements so as to maximize the returns.

-To predict financial success of the movies before they are produced.

9. Sports:

-To improve performance of NBA teams in US

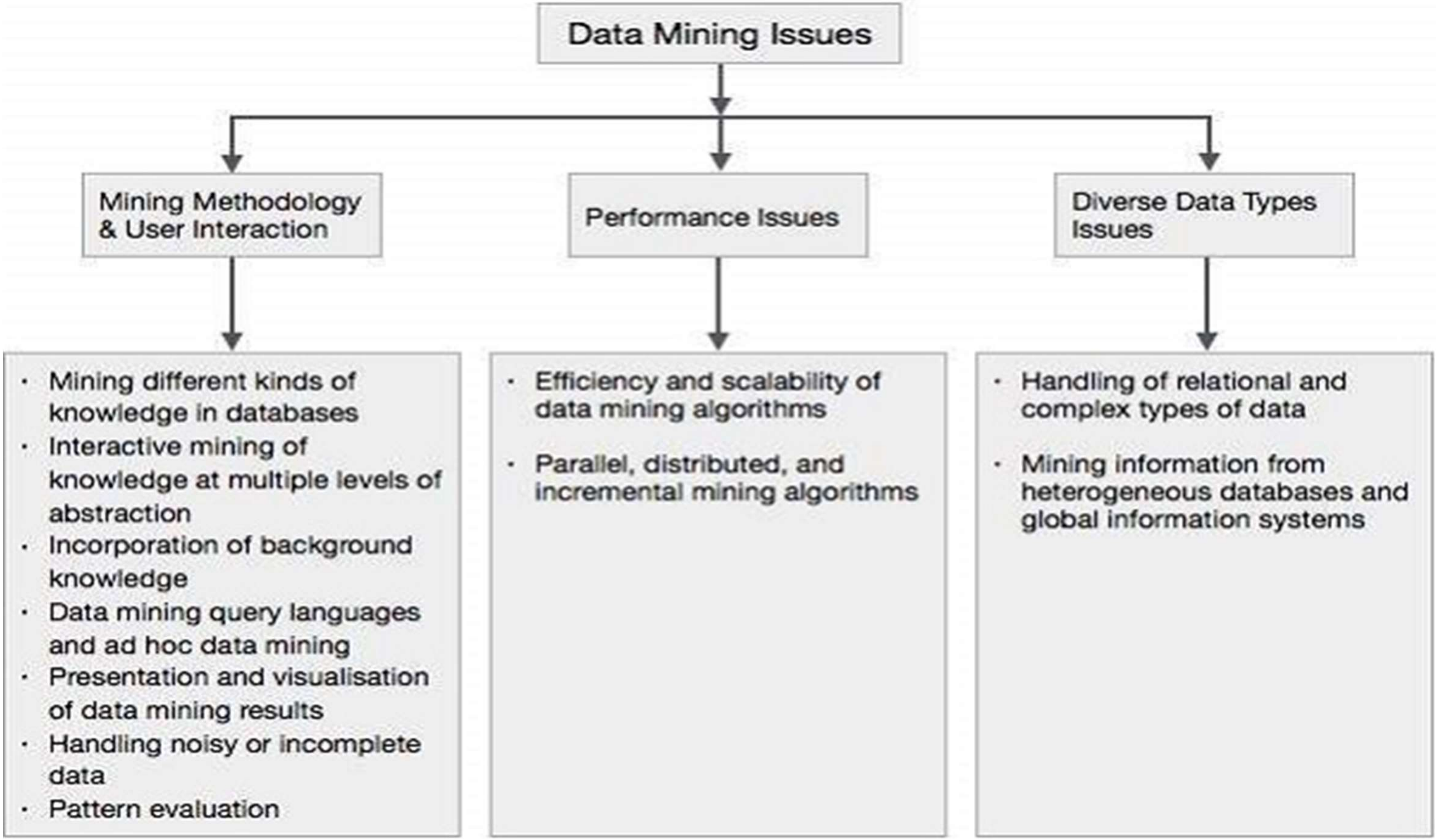
-To increase the chances of winning

Self Learning Topics

1. Data Marts

2. Major issues in Data Mining

Data Mining Issues



Data Mining Issues

Mining Methodology & User Interaction

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

Performance Issues

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

Diverse Data Types Issues

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