Course: Data Mining
Lecturer: Hend Muslim Jasim **Date: October ,2025**
Reference: " Data Mining Concepts and Techniques" Third Edition, Jiawei Han, et al.

Lecture 1: Introduction to Data Mining

1. What is Data Mining?

Data mining is the process of discovering interesting, useful, and previously unknown patterns and knowledge from large amounts of data. It is also known as **Knowledge Discovery in Databases (KDD)**. The term *data mining* is often used interchangeably with KDD, though technically, data mining is a core step within the broader KDD process.

The KDD Process:

1. **Data Cleaning** – Remove noise and inconsistencies.
2. **Data Integration** – Combine data from multiple sources.
3. **Data Selection** – Retrieve relevant data for analysis.
4. **Data Transformation** – Convert data into appropriate forms for mining.
5. **Data Mining** – Apply intelligent methods to extract patterns.
6. **Pattern Evaluation** – Identify interesting patterns.
7. **Knowledge Presentation** – Visualize and present the discovered knowledge.

KDD Processes are depicted in Fig. 1.

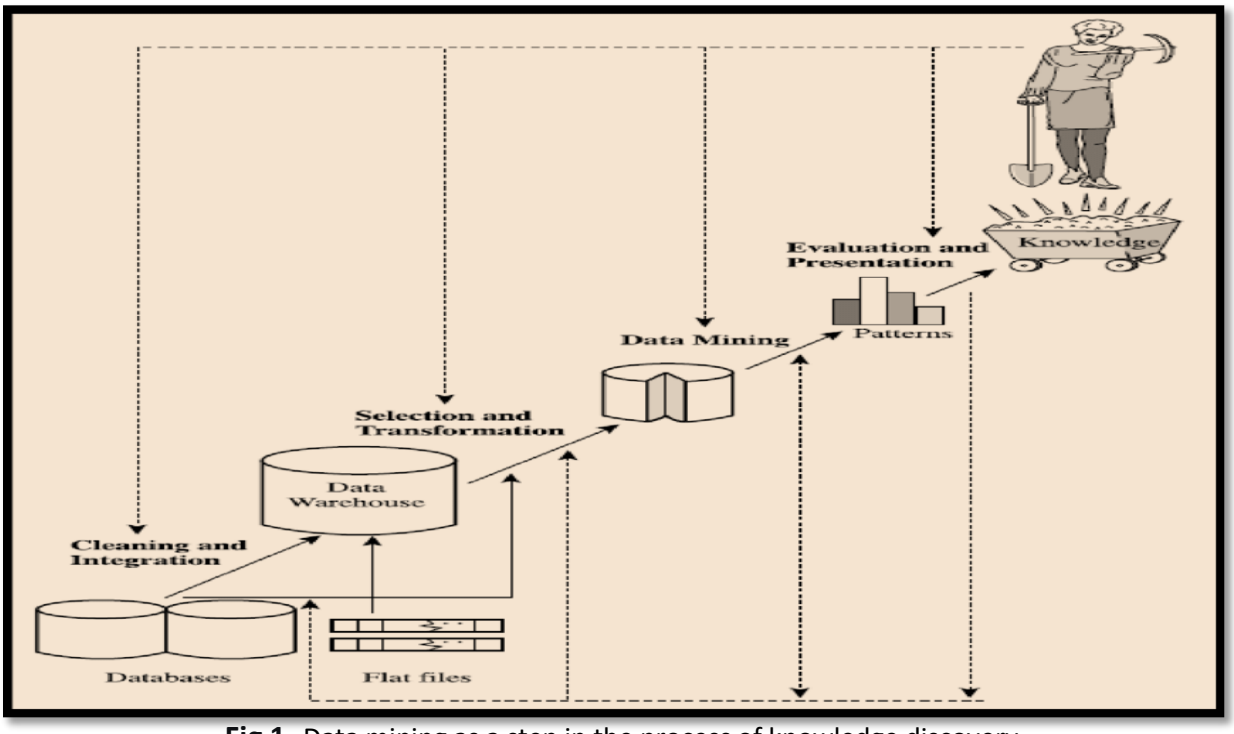

Fig 1- Data mining as a step in the process of knowledge discovery.

2. Why Data Mining?

We live in the **data age**. Massive amounts of data are generated daily from:

- Business transactions
- Social media
- Scientific experiments
- Medical records
- Web searches

This data is often stored in large databases or data warehouses, but without powerful tools, it remains underutilized—a **"data tomb."** Data mining turns these data tombs into **actionable knowledge**.
Example: Google's *Flu Trends* uses aggregated search queries to estimate flu activity faster than traditional systems.

3. What Kinds of Data Can Be Mined?

Data mining can be applied to various types of data:

- **Relational Databases**
- **Data Warehouses**
- **Transactional Data**
- **Advanced Data Types**: time-series, sequences, data streams, spatial, multimedia, text, graphs, and web data.

4. What Kinds of Patterns Can Be Mined?

Data mining functionalities can be categorized into **descriptive** and **predictive** tasks.

Major Data Mining Tasks:

a. Class/Concept Description

- **Characterization**: Summarizing general features of a target class.
- **Discrimination**: Comparing target and contrasting classes.

b. Frequent Pattern Mining

- Frequent Item sets: Items that often occur together (e.g., milk and bread).
- Sequential Patterns: Sequences that occur frequently (e.g., laptop → camera → memory card).
- Association Rules: e.g., buys(X, "computer") → buys(X, "software").

c. Classification and Regression

- Classification: Predicts categorical labels (e.g., "yes" or "no").
- Regression: Predicts continuous numeric values.

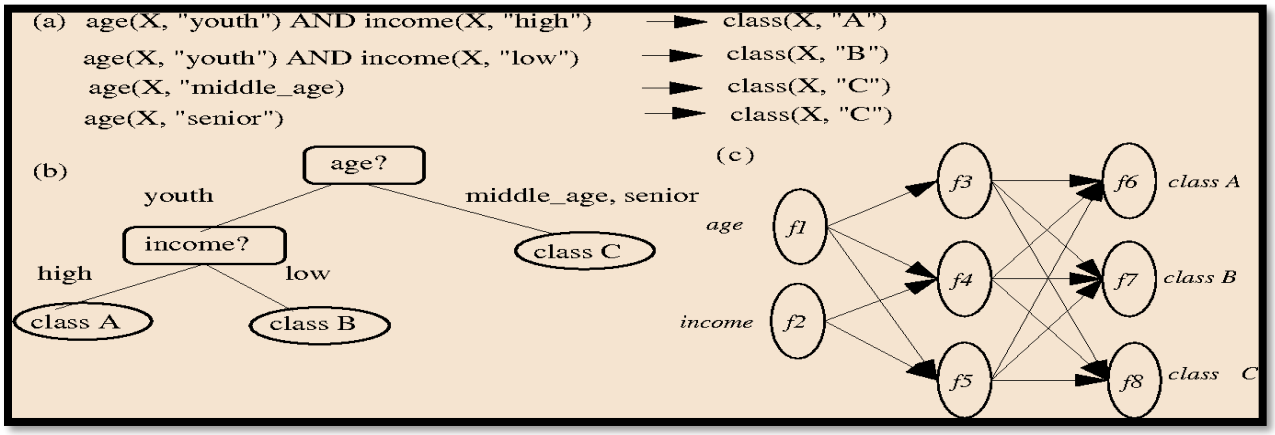

Fig 2- A classification model can be represented in various forms:
(a) IF-THEN rules, (b) a decision tree, or (c) a neural network

d. Cluster Analysis

- Groups data into clusters so that objects in the same cluster are similar and dissimilar to objects in other clusters.

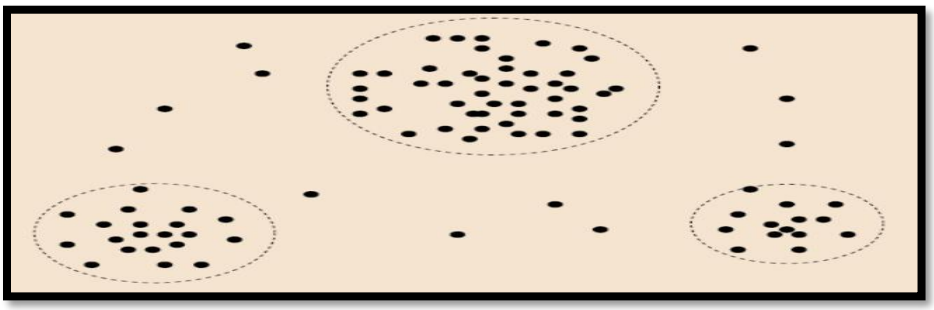

Fig 3- A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

e. Outlier Analysis

Identifies anomalies or exceptions (e.g., fraud detection).

5. Which Technologies Are Used?

Data mining is a **multidisciplinary field** that draws from:

- **Statistics**
- **Machine Learning**
- **Database Systems**
- **Information Retrieval**
- **Visualization**
- **Pattern Recognition**
- **High-Performance Computing**

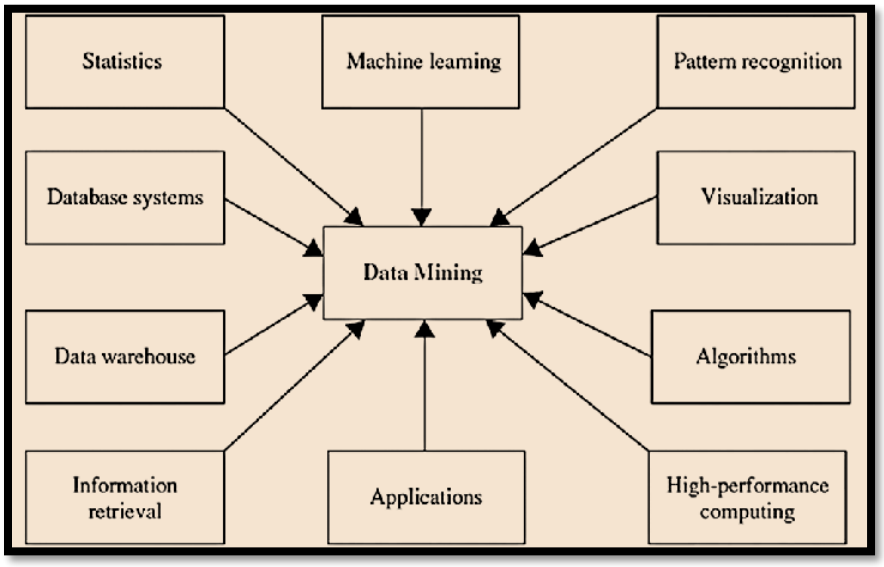

Fig 4 - Data mining adopts techniques from many domains

6. Major Issues in Data Mining

a. Mining Methodology

- Mining diverse and new types of knowledge
- Handling uncertainty, noise, and incomplete data
- Pattern evaluation and interestingness measures

b. User Interaction

- Interactive mining
- Incorporation of background knowledge
- Visualization of results

c. Efficiency and Scalability

- Efficient algorithms for large datasets
- Parallel, distributed, and incremental mining

d. Diversity of Data Types

- Handling complex data (e.g., graphs, multimedia, streams)

e. Data Mining and Society

- Privacy and security
- Social impacts
- Invisible data mining (e.g., recommender systems)

7. Applications of Data Mining

- Business Intelligence
- Web Search Engines
- Bioinformatics
- Healthcare
- Finance
- Retail
- Social Network Analysis

Summary

- Data mining is the **automated extraction of hidden knowledge** from large datasets.
- It is a **natural evolution** of information technology.
- It involves **multiple steps**—from data preparation to pattern evaluation.
- It supports a **wide range of tasks** including classification, clustering, and association.
- It is used in **diverse domains** and relies on **multiple disciplines**.

Data mining turns data into **knowledge**, and knowledge into **action**.