SILVER OAK UNIVERSITY
EDUCATION TO INNOVATION

2040233341
Data Mining
Unit 1 : Introduction to Data Mining (DM)

By Kinjal Bhavsar

Topics to be covered

- Why Data Mining?
- What is Data Mining?
- What kind of data is mined?
- KDD Process (**K**nowledge **D**iscovery in **D**atabases)
- What kind of Patterns can be mined?
- Which Technologies are used?
- Major Issues in DM

Why Data Mining?

- Section - 1

Why Data Mining?

- We are living in the data age.
- The Explosive Growth of Data: from terabytes to petabytes
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets
- The abundance of data, coupled with the need for powerful data analysis tools, has been described as a "data rich but information poor" situation.
- As a result, data collected in large data repositories become "data tombs"
- The widening gap between data and information calls for systematic development of "data mining tools" that can turn "data tombs" into "golden nuggets" of knowledge.

Evolution of Database Technology

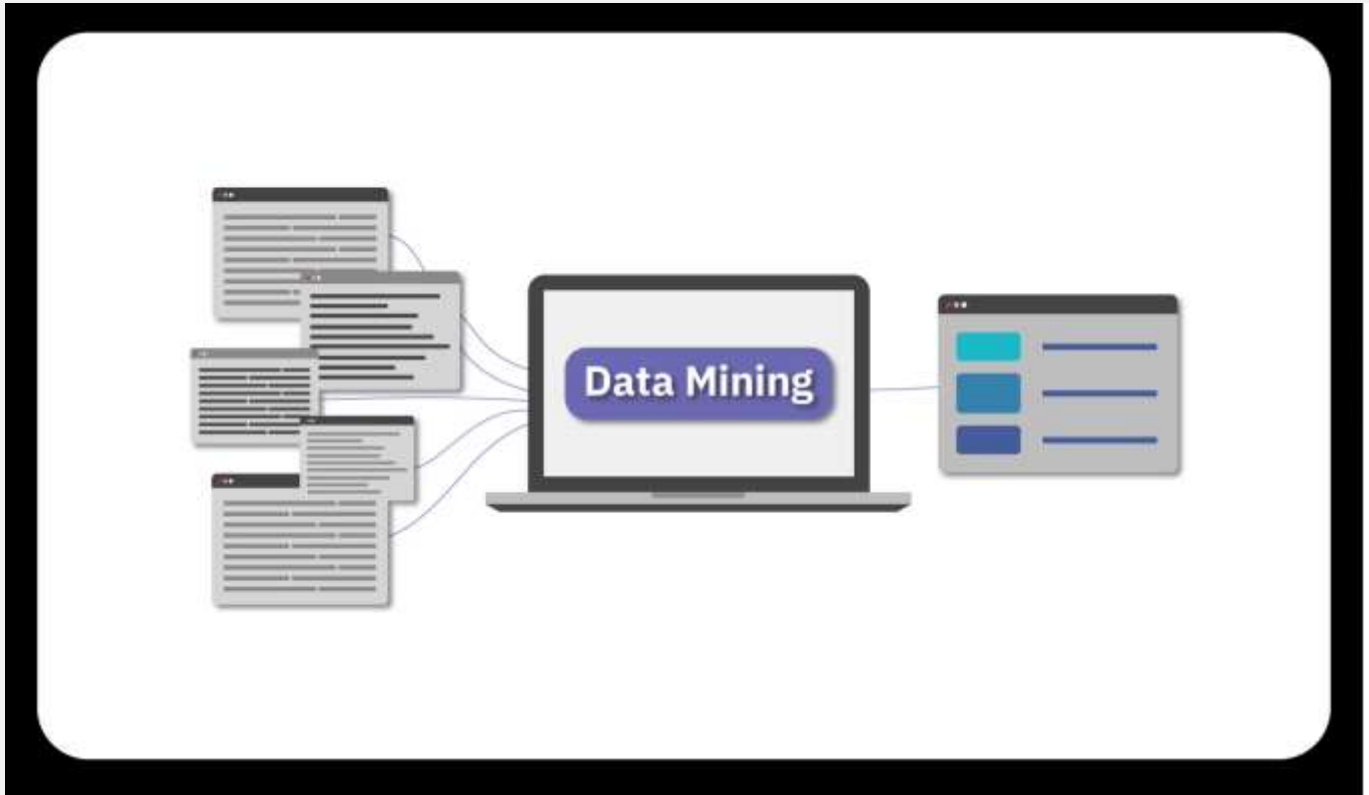
- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - 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?

- Section - 2

What is Data Mining?

- Data Mining refers to “extracting” or “mining” knowledge from large amount of data.
- Definition : “The goal of Data Mining is to extract information from a large dataset and transform it into an understandable structure for further use.”



what kind of data can be mined?

- Section - 3

Data Mining Functionalities

- Data mining functionalities can be classified into two categories:
  1. **Descriptive**
  2. **Predictive**

- **Descriptive**
  - This task presents the general properties of data stored in a database.
  - The descriptive tasks are used to find out patterns in data.
  - **E.g.**: Cluster, Trends, etc.

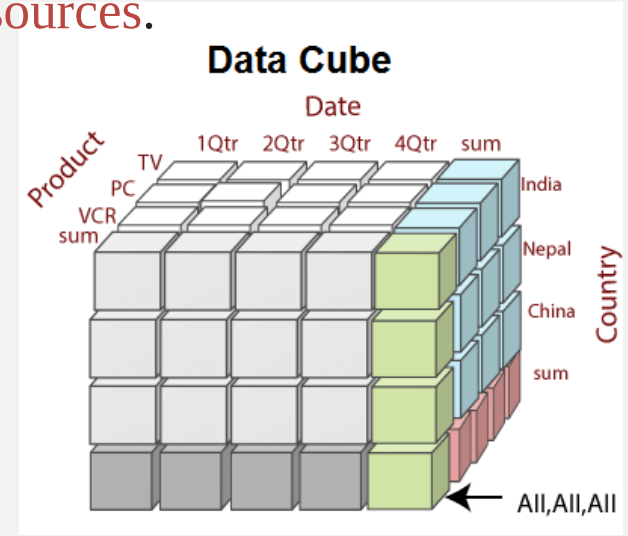
- **Predictive**
  - These tasks predict the value of one attribute on the basis of values of other attributes.
  - **E.g.**: Festival Customer/Product Sell prediction at store

Data Mining - On what kind of data?

- **Relational Databases:**
  - A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database tables, and a set of software programs to manage and access these data.
  - **E.g.** : SQL Server, Oracle etc.
- **Data Warehouses:**
  - A data warehouse is a repository of information collected from multiple sources.
  - It is constructed after pre-processing of data.

    (Data cleaning, Data integration, Data transformation,

    Data loading, and Periodic data refreshing etc.)
  - **E.g.** : Stock Market, D-Mart, Big Bazar etc.



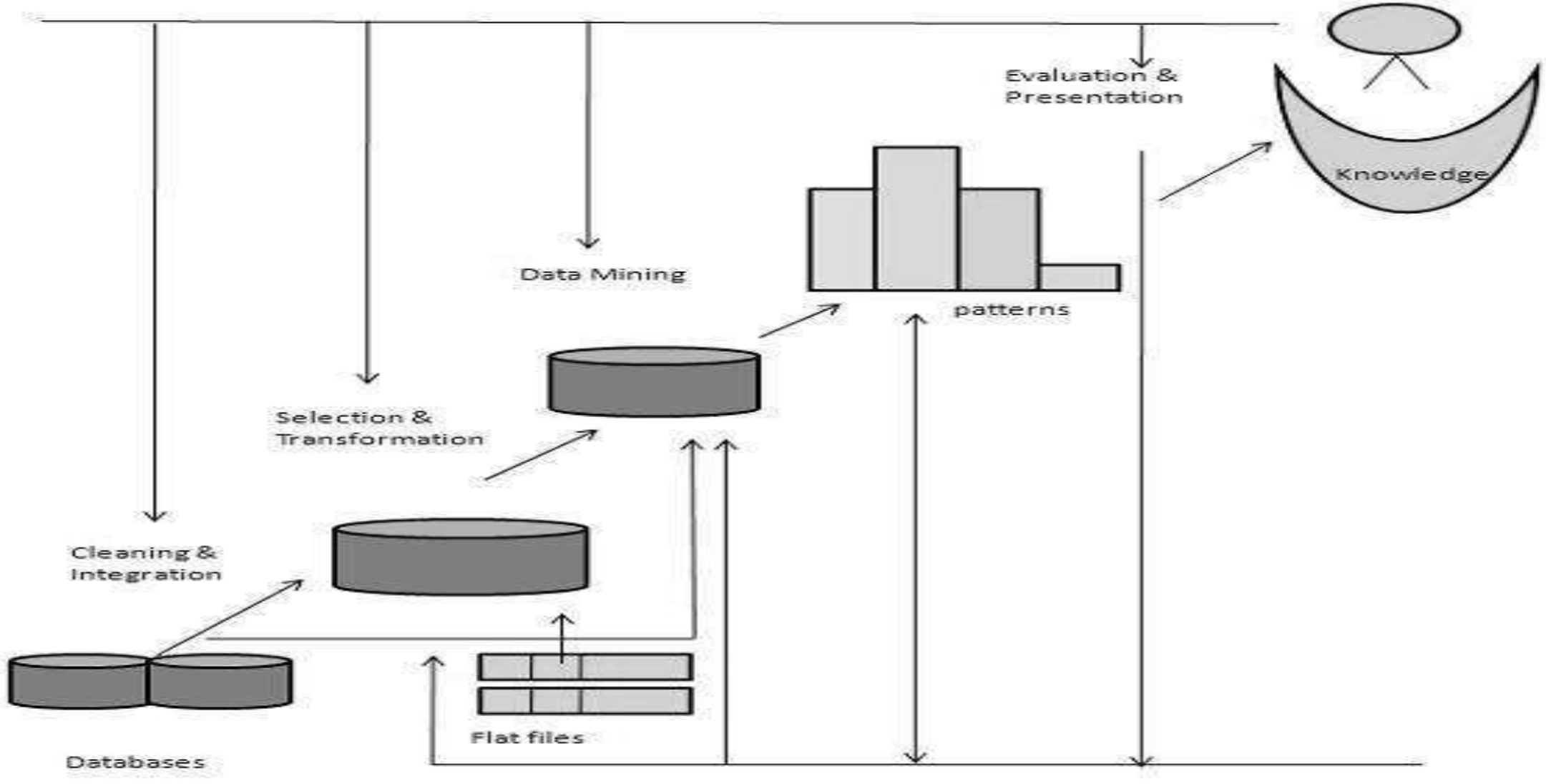
Data Mining—On what kind of data? (Cont..)

- **Transactional Databases:**
  - Transactional database consists of a file where each record represents a transaction.
  - A transaction typically includes a unique transaction identity number (TID) and a list of the items making up the transaction (such as items purchased in a store).
  - **E.g.** : Online shopping on Flipkart, Amazon etc.
- **Other Data/Databases**
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

KDD Process

- Section - 4

The following diagram shows the process of knowledge discovery −



KDD (Knowledge Discovery in Databases) Process

- Knowledge discovery in databases is a process of an iterative sequence of the following steps:
  1. **Data Cleaning**
     - In this step, the noise and inconsistent data is removed.
  2. **Data Integration**
     - In this step, multiple data sources are combined.
  3. **Data Selection**
     - In this step, data relevant to the analysis task are retrieved from the database.
  4. **Data Transformation**
     - In this step, data is transformed or consolidated into forms appropriate for mining.
  5. **Data Mining**
     - In this step, intelligent methods are applied in order to extract data patterns.
  6. **Pattern Evaluation**
     - In this step, the truly interesting patterns representing knowledge based.
  7. **Knowledge Representation**
     - In this step, visualization and knowledge representation techniques are used to present the mined knowledge to the user.

what kind of Pattern can be mined?

- Section - 5

what kind of patterns ?

- **Class / Concept Descriptions:**
  - Data entries are associated with labels or classes.
  - For example, in a company,
    - The classes of items for sales include computer and printer.
    - Concept of customers include big spenders and budget spenders.
  - These descriptions can be derived using Data Characterization and Data Discrimination.
    1. Data Characterization :
       - This refers to summarizing data of class under study.
       - It's also called **Target** class.
       - E.g. Summarize the characteristics of customers who purchase more items on the shopping website.
       - The data related to such items can be collected by executing an SQL query on the purchase database.
       - Output of Data characterization can be presented in various forms. Like, pie chart, bar chart, curves, multidimensional data cubes and multidimensional data tables.
       - The resulting description can also be presented as generalized relations or in rule form is called Characteristic Rules.
    2. Data Discrimination:
       - It refers to the comparison of the target class with one or set of comparative classes.
       - It's also called **Contrasting** classes.

what kind of patterns ?(Cont..)

- E.g. A customer relationship manager want to compare two groups of customers - who shop for computer products regularly and who shop for computer products rarely.
- Output of Data descrimination can be presented in various forms. Like, pie chart, bar chart, curves, multidimensional data cubes and multidimensional data tables.
- The discrimination description expressed in the form of rules are referred as Discrimination Rules.

- **Mining frequent Patterns, Associations and Correlations:**
  - **Frequent patterns**: Frequent Patterns are those patterns that occur frequently in data.
    - There are many kinds of frequent patterns, such as frequent itemsets, frequent subsequence, and frequent sub-structure.
    - Frequent Itemsets: It refers to a set of items that frequently appear together . For example, milk & bread.
    - Frequent Subsequence : A sequence of patterns that occur frequently, such as purchasing a camera is followed by Memory card.
    - Frequent Substructure : It refers to different structural forms, such as graphs,trees or lattices, which may be combined with itemsets or subsequence.

what kind of patterns ?(Cont..)

- **Associations**: It shows the relationships between data and pre-defined association rules.
  - For instance, a shopkeeper makes an association rule that 70% of the time, milk is sold with bread and only 70% of the time, milk is not sold with bread. These two items can be combined together to make an association.
- **Correlations**: This is performed to find the statistical correlations between two data points to find if they have positive, negative, or no effect on each other.

- **Classification and Regression for Predictive Analysis:**
  - **Classification**: Classification is the process of finding a set of models that describes and distinguishes data classes or concepts.
    - The model is derived based on the analysis of a set of training data.
    - Example, Student (Data Set)

  - **Regression :** Unlike classification, regression is used to find the missing numeric values from the dataset. It is also used to predict future numeric values as well. For instance, we can find the behavior of the next year's sales based on the past twenty years' sales by finding the relation between the data.

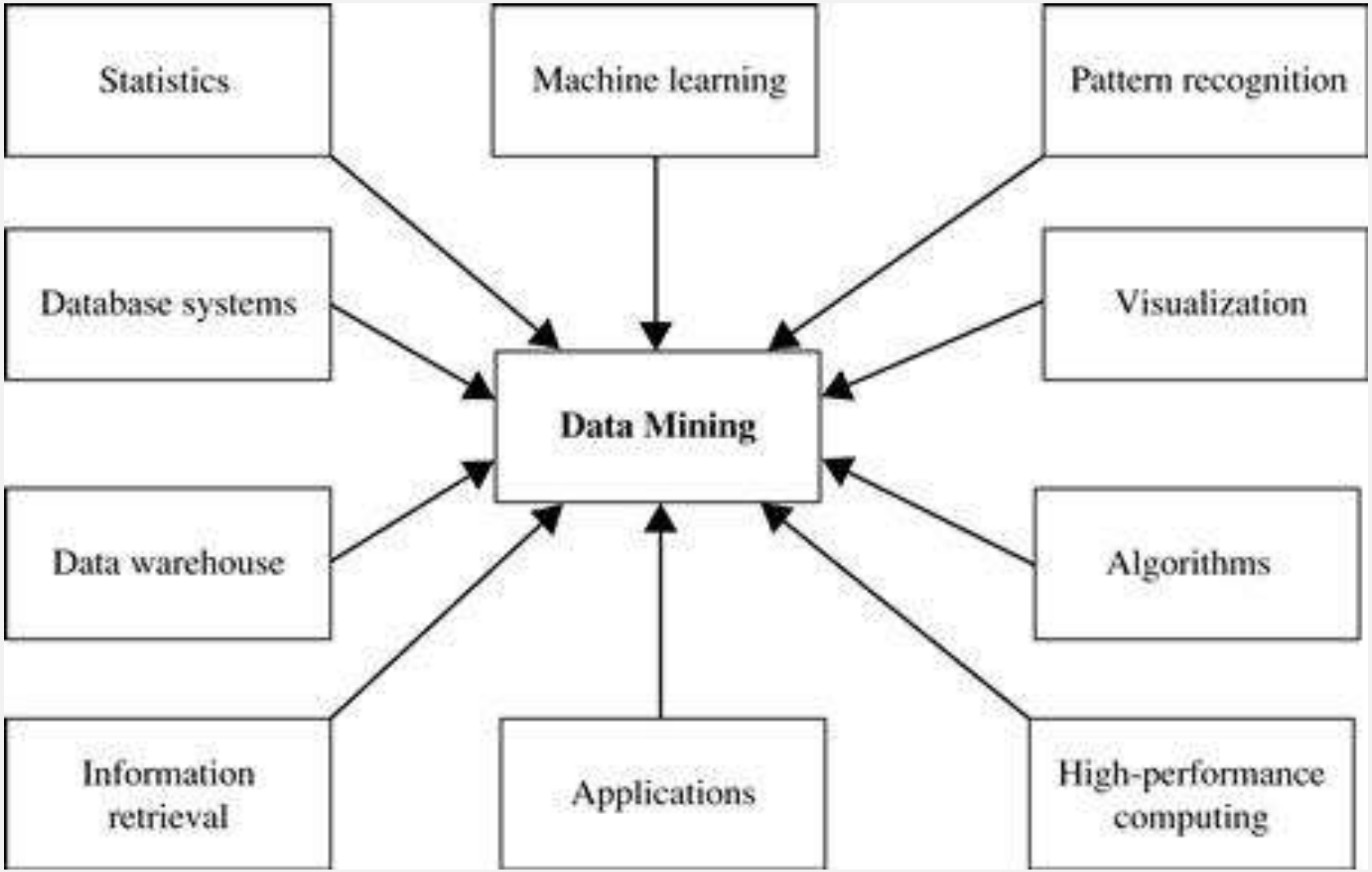
what kind of Patterns? (Cont..)

- **Cluster Analysis:**
  - Clustering is a method of grouping data into different groups ,so that in each group share similar trends & patterns.
  - Clustering can be used to generate class labels for a group a data.
  - The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity.
- **Outlier Analysis:**
  - Not all data points in the dataset need to follow the same behavior. Data points that don't follow the usual behavior are called outliers. Analysis of these outliers is called outlier analysis. These outliers are not considered while working on the data.
  - Most data mining methods discard outlier as noise or exceptions.
  - Finding such type of applications are fraud detection is referred as outlier mining.

Technologies are used

- Section - 6

Technologies are used



Technologies are used (Cont..)

- **Statistics** : Statistics are used to the collection, analysis, interpretation or explanation and representation of data.
  - Data mining has an inherent connection with statistics.
  - A statistical model is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes.
  - For example, in data mining tasks like data characterization and classification, statistical models of target classes can be built.
- **Machine Learning** : Machine learning investigates how computers can learn (or improve their performance) based on data.
  - A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data.
  - Machine learning is a fast-growing discipline. Here, we illustrate classic problems in machine learning that are highly related to data mining.

1. Supervised learning that makes use of class labels to predict information
2. Unsupervised learning doesn't use class labels similar to clustering but it will discover new classes within data.
3. Semi-supervised learning will redefine the boundaries between two classes and makes use of both labeled and unlabeled examples.
4. Active learning will ask the user to label the classes that may be from unlabeled examples. It will optimize

Technologies are used (Cont..)

- **Database systems and Data Warehouse**
  - Database systems research focuses on the creation, maintenance, and use of databases for organizations and end-users. Particularly, database systems researchers have established highly recognized principles in data models, query languages, query processing etc.
  - Many data mining tasks need to handle large data sets or even real-time, fast streaming data.
  - A data warehouse integrates data originating from multiple sources and various timeframes. It consolidates data in multidimensional space to form partially materialized data cubes.
  - The data cube model not only facilitates OLAP in multidimensional databases but also promotes multidimensional data mining.
- **Information Retrieval**
  - Information retrieval (IR) is the science of searching for documents or information in documents. Documents can be text or multimedia, and may reside on the Web.
  - The differences between traditional information retrieval and database systems are twofold: Information retrieval assumes that (1) the data under search are unstructured; and (2) the queries are formed mainly by keywords, which do not have complex structures (unlike SQL queries in database systems).

Technologies are used (Cont..)

- **Pattern Recognition**
  - Pattern recognition is the process of recognizing pattern using machine learning algorithm.
  - Example: speech recognition, multimedia document recognition (MDR), automatic medical diagnosis.
- **Data Visualization**
  - People understand the significance of data by placing it in visual text.
  - Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.
- **Algorithms**
  - An algorithm in datamining is a set of heuristics and calculations that creates a model from data.
- **High Performance Computing**
  - HPC framework which can abstract the increased complexity in current computing systems.
- **Applications**
  - The list of areas where data mining is widely used – Financial Data Analysis, Retail Industry, Biological Data Analysis, Scientific Applications

Major Issues in Data Mining

- Section - 7

Data Mining Issues

- Data mining issues can be classified into five categories:
  1. **Mining Methodology**
  2. **User Interaction**
  3. **Efficiency and Scalability (Algorithms)**
  4. **Diversity of Database Types**
  5. **Data Mining and Society**

Data Mining Issues(Cont..)

- Data mining issues can be classified into five categories:
  - Mining Methodology
    - Mining various and new kinds of knowledge
    - Mining knowledge in multi-dimensional space
    - Data mining: An interdisciplinary effort
    - Boosting the power of discovery in a networked environment
    - Handling noise, uncertainty, and incompleteness of data
    - Pattern evaluation and pattern- or constraint-guided mining
  - User Interaction
    - Interactive mining
    - Incorporation of background knowledge

Data Mining Issues(Cont..)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

Thank You