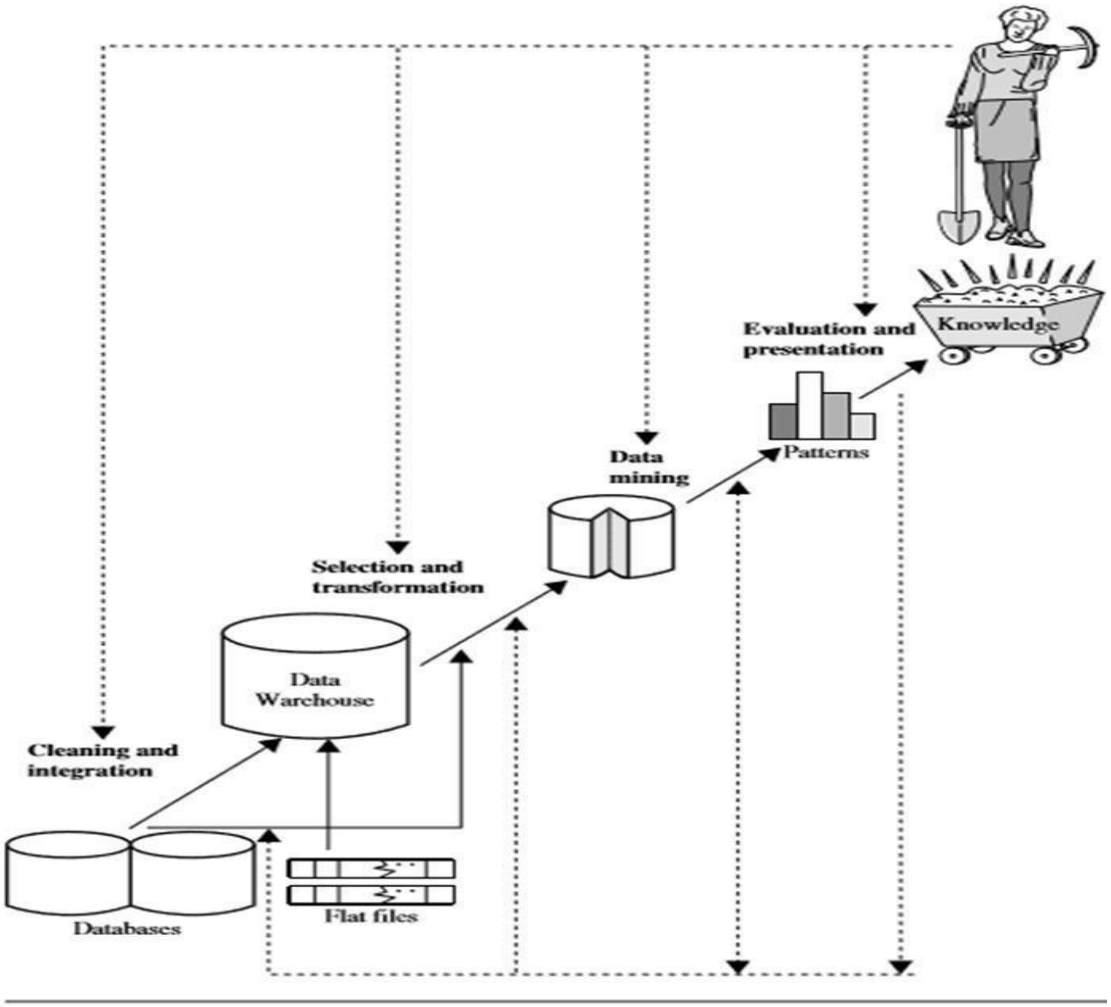
CHAPTER 1: DATA-MINING

What Is Data-Mining?

- **Definition:** The process of discovering patterns, trends and insights from huge volumes of data.
- Data-mining aims to extract valuable knowledge to
  - inform decision-making
  - predict future-outcomes
  - drive innovation
- Through data-mining, companies can uncover hidden-patterns and relationships within the dataset.
- Data-mining finds applications in business, healthcare, finance, marketing, etc.

KDD

- KDD stands for **K**nowledge **D**iscovery in **D**atabases
- **Definition:** A multi-step process of extracting useful knowledge or insights from large datasets.
- Steps in Knowledge Discovery Process includes
  1) **Data-cleaning:** Remove noise and inconsistencies.
  2) **Data-integration:** Combine multiple data-sources.
  3) **Data-selection:** Retrieve relevant-data.
  4) **Data-transformation:** Prepare data for mining.
  5) **Data-mining:** Extract patterns using intelligent methods.
  6) **Pattern evaluation:** Assess patterns' interestingness.
  7) **Knowledge presentation:** Visualize and represent mined knowledge.



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

1) **Data-Cleaning**
   - **Definition:** The process of identifying & correcting errors and inconsistencies in the data.
   - This may include
     - handling missing-values
     - removing duplicate records
     - correcting typographical errors
     - resolving inconsistencies in data-formats
2) **Data-Integration**
   - **Definition:** The process of combining data from multiple sources into a unified format.
   - This may include
     - merging databases
     - integrating data from different departments within an company
     - combining data from external-sources.
3) **Data-Selection**
   - **Definition:** The process of identifying and retrieving relevant-data for analysis.

- This may also involve filtering out irrelevant-data that are not needed for the analysis.

4) **Data-Transformation**
   - **Definition:** The process of converting raw data into a format that is suitable for analysis.
   - This may include
     - Normalization
     - Aggregation
     - Encoding categorical variables
     - Feature engineering.

5) **Data-Mining**
   - **Definition:** The process of discovering patterns, relationships, and insights from large datasets.
   - DM Techniques include
     - Classification
     - Clustering
     - Anomaly-detection

6) **Pattern Evaluation**
   - **Definition:** The process of assessing the quality, significance, and usefulness of discovered patterns.
   - Evaluate accuracy, reliability, and relevance to the problem domain.

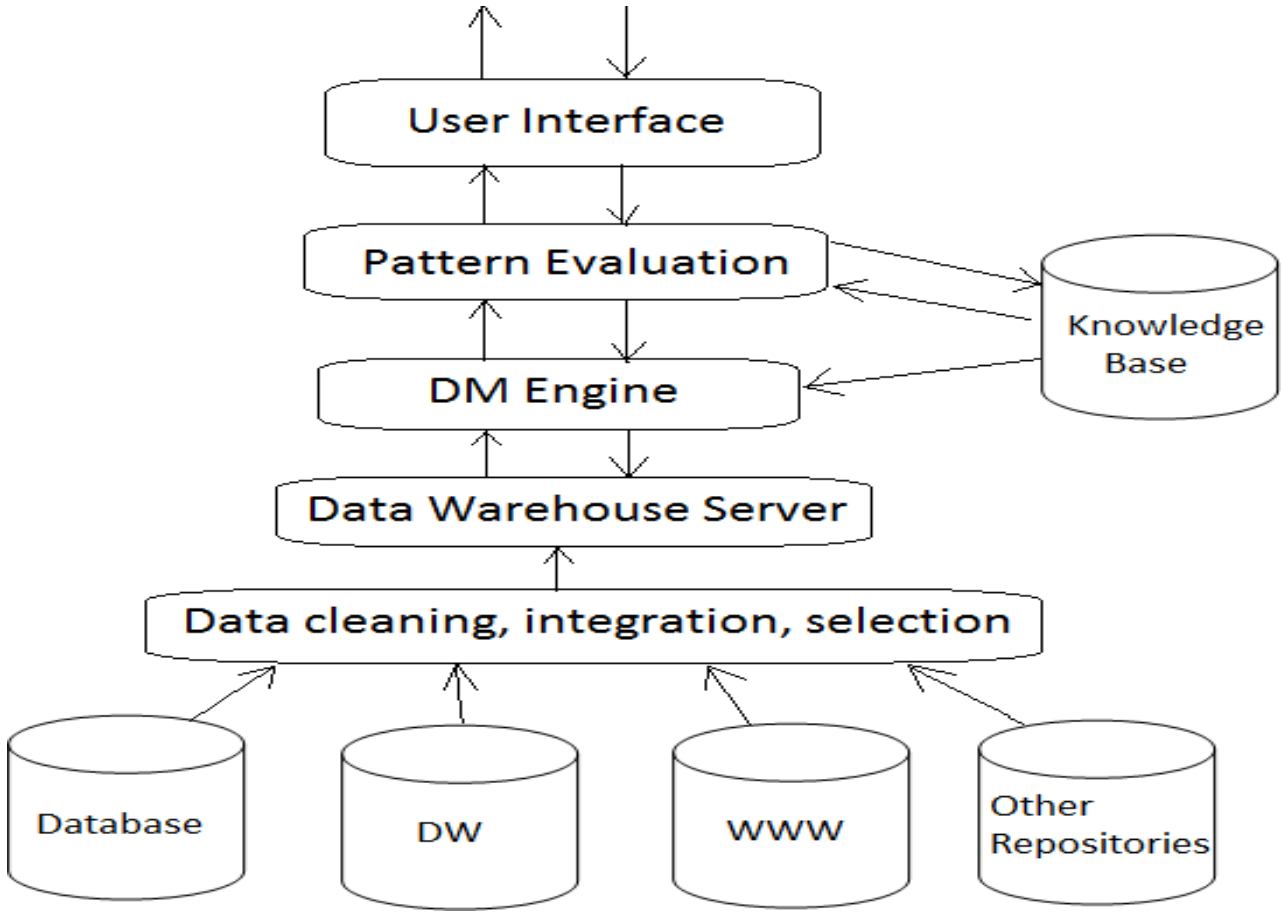
7) **Knowledge Presentation**
   - **Definition:** The process of communicating the insights obtained from the DM process to stakeholders
   - This may include creating visualizations, reports, dashboards etc.

Architecture of DM System

Typically DMS consists of following components:

- Database, data warehouse, WWW, or other information repository(spread sheets, files)
- Data Warehouse Server:
  This server is responsible for fetching the relevant data, based on user's data mining request.
- Knowledge base:
  This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting pattern
- DM Engine:
  Consists of set of methods/functions like characterization, association, correlation analysis, classification, cluster analysis, prediction, outlier analysis, etc.
- Pattern Evaluation:
  Employs interestingness measure and interacts with the data mining modules so as to focus the search toward interesting patterns
- User Interface:

Interface between user & DMS. User specifies query, task, etc. User browse data, visualize output.



Which technologies are used for DM?

Data mining involves an integration of techniques from multiple discipline such as database, data warehouse, statistics, machine learning, pattern recognition, neural networks, data visualization, information retrieval, image/signal processing, spatial & temporal data analysis.

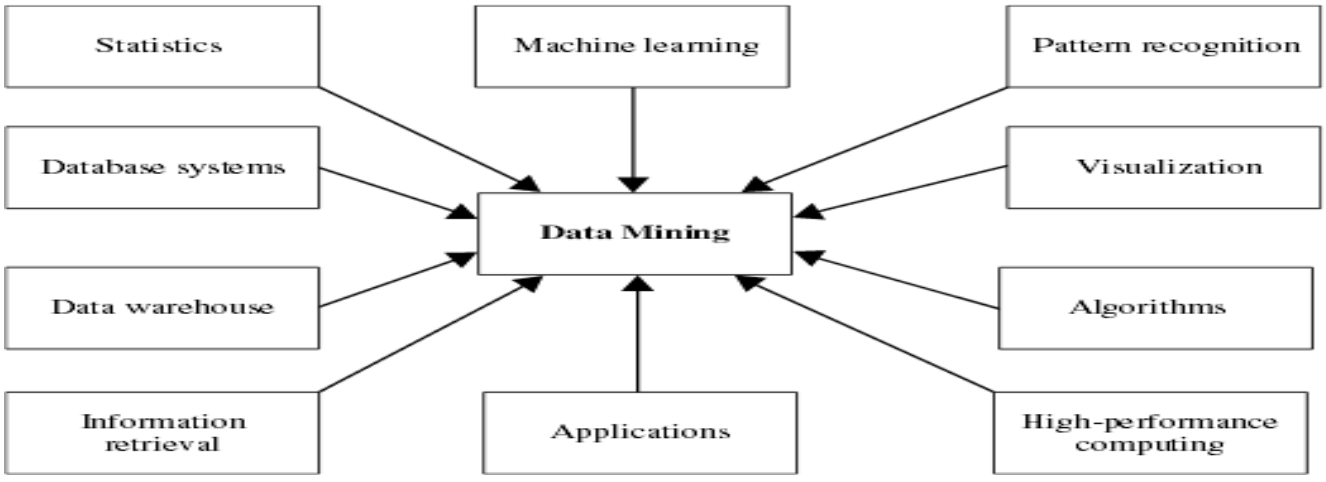


Fig: Data Mining adopts many domains

What kind of data can be mined?

1. Database data:

- A **database system** has a **database** (collection of related data) and a **DBMS** (database management system).
- **DBMS** helps to store, manage, and protect data.
- It allows **multiple users** to access data at the same time.
- It ensures **data security** and prevents **unauthorized access**.
- A **relational database** stores data in **tables**.
- Each table has **rows (tuples)** and **columns (attributes)**.
- A **primary key** uniquely identifies each row.
- The **ER model** represents data using **entities and relationships**.

2. Data Warehouse:

Data Warehouse: A central storage for data from different sources.
Purpose: Helps in decision-making by organizing and summarizing data.
Data: Historical, integrated, and stored in one place.
Structure: Uses **data cubes** with **dimensions** (e.g., time, region) and **cells** storing values.
Example: AIElectronics can analyze sales easily using a data warehouse.

3. Transactional Data:

- T**ransactional Data** - Records transactions like purchases, bookings, or clicks.
- **Transactional Database** - Stores transactional data.
- **Transaction Components**:
  - Unique **transaction ID** (trans_ID).
  - List of **items** in the transaction.
- **Additional Tables** may store:
  - **Item descriptions**.
  - **Salesperson** details.
  - **Branch** information.

Mining Other Kinds of Data

- Mining Spatial Data
  - Spatial frequent/co-located patterns, spatial clustering and classification
- Mining Spatiotemporal and Moving Object Data
  - Spatiotemporal data mining, trajectory mining, swarm, …
- ber-Physical System Data

- Applications: healthcare, air-traffic control, flood simulation

- Mining Multimedia Data
  - Social media data, geo-tagged spatial clustering, periodicity analysis, …
- Mining Text Data
  - Topic modeling, i-topic model, integration with geo- and networked data
- Mining Web Data
  - Web content, web structure, and web usage mining
- Mining Data Streams
  - Dynamics, one-pass, patterns, clustering, classification, outlier detection

Difference Between Data Mining and KDD

Aspect	Data Mining	KDD (Knowledge Discovery in Databases)
Definition	Process of discovering patterns and extracting useful information from large datasets	Overall process of extracting knowledge from data through multiple steps
Objective	To find hidden patterns, relationships, and insights in data	To transform raw data into useful knowledge for decision-making
Scope	Narrower in scope	Broader in scope
Focus	Focuses on algorithms and techniques for data analysis	Includes data preprocessing, data mining, evaluation, and interpretation
Role in Process	A single step within KDD	Complete methodology that includes data mining
Process Stage	Fourth phase of the KDD process	Entire multi-step process
Techniques Used	Classification, clustering, regression, association rules, etc.	Data cleaning, integration, selection, transformation, data mining, pattern evaluation
Application Areas	Marketing, finance, healthcare, fraud detection, etc.	Business intelligence, scientific research, engineering, decision support
Main Emphasis	Pattern extraction	End-to-end knowledge discovery

DBMS vs. Data-Mining

Aspect	DBMS (Database Management System)	Data Mining
Definition	A software system used to store, manage, and access databases	Process of discovering patterns and extracting useful information from large datasets
Main Purpose	Efficient storage, management, and retrieval of data	Extract knowledge, insights, and hidden patterns from data
Type of Data	Mainly structured data stored in tables with predefined schemas	Structured, semi-structured, and unstructured data
Operations	Supports CRUD operations (Create, Read, Update, Delete)	Performs analytical operations like clustering, classification, regression
Focus	Data organization and transaction processing	Pattern discovery and predictive analysis
Techniques Used	Query processing, indexing, transaction management	Machine learning, statistics, and pattern recognition
Languages Used	SQL (Structured Query Language)	SQL, Python, R, and specialized mining tools
Output	Stored and retrieved data	Knowledge, trends, rules, and predictions
Usage	Day-to-day operational data management	Research, business intelligence, and decision-making
Examples	Oracle, MySQL, SQL Server	Weka, RapidMiner, Apache Spark

DM Techniques

- DM operates on various data-repositories. This leads to the extraction of diverse patterns.
- Patterns include
  - Characterization & discrimination
  - Frequent-patterns, associations & correlations
  - Classification & regression
  - Clustering
  - Outliers

Class Description: Characterization and Discrimination

- Data can be associated with classes or concepts.

Characterization

- **Definition:** Describing the general properties of a class or concept by summarizing associated data.
- Output forms include charts, cubes or tables.

Discrimination

- **Definition:** Distinguishing between classes or concepts based on their characteristics.
- Output forms are similar to characterization but include measures for comparison.

Mining Frequent-Patterns, Associations, and Correlations

- **Definition:** Frequent-patterns are those that occur frequently in data.
- Types of Frequent-patterns include

i) Frequent Item-Sets

- Sets of items that frequently co-occur in a dataset.

ii) Frequent Subsequences

- Sequences of items/events that frequently occur while preserving their order.

iii) Frequent Substructures

- Extend frequent Item-sets to more complex structures like trees or graphs.

Association-Rule Mining

- **Definition:** Identifies interesting relationships between variables in a dataset.
- Expresses relationships as association-rules of the form "if X then Y."

Frequent Item-Set Mining

- **Definition:** Process of discovering frequent Item-sets in a dataset.
- Used for various applications like market basket analysis and recommendation-systems.

Classification and Regression for Predictive Analysis

Classification

Definition: Process of finding a model that distinguishes data-classes or concepts.

- Classification is a supervised-learning-task.

- Model derived from analysis of training-data with known class-labels.
- Model used to predict class-labels of objects with unknown labels.
- Examples include spam detection, image classification, etc.

Model Representation

- Different algorithms use different forms of representation.
- Examples:

  Classification rules use
  IF-THEN Decision
  trees use a tree-like
  structure

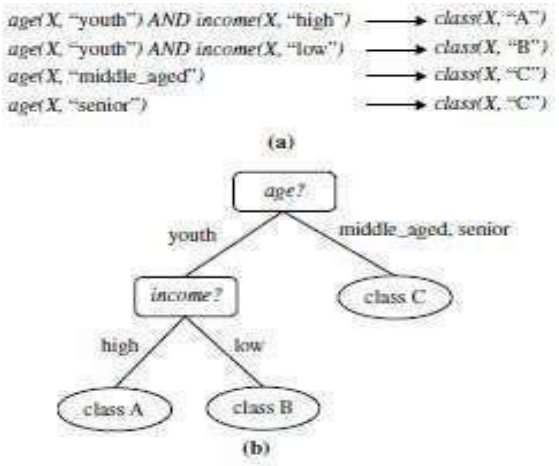


Fig:A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree,

<u>Regression</u>

Definition: Models continuous-valued functions to predict missing or unavailable data-values

- Regression is another supervised-learning-task.
- Used for numeric prediction rather than discrete class-labels.
- Examples include house price prediction, stock price prediction, etc.

Prediction

- Using models derived from classification to make predictions about new or unseen objects.

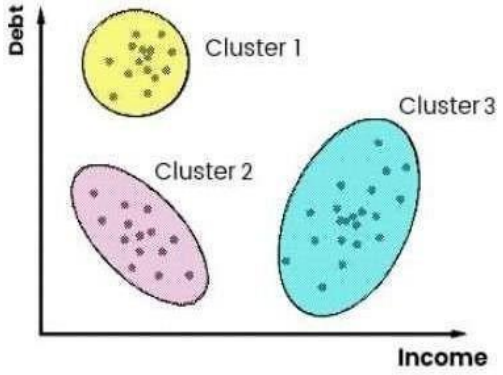
Relevance Analysis

- **Definition:** Assessing the significance and relevance of attributes or features in predictive modeling.
- Selected-attributes used in classification and regression

  Whereas irrelevant-attributes are excluded from consideration.

Cluster Analysis (Clustering)

- **Definition:** Clustering is the process of dividing a set of data-objects into groups.
- Clusters consist of data-objects
  - similar to one another within the same group
  - dissimilar to the objects in other groups
- Clustering is unsupervised-learning because there is no predefined class-labels.
- Classification analyzes class-labeled data sets
  Whereas clustering analyzes data-objects without consulting class-labels.
- Applications include customer segmentation, image segmentation, anomaly-detection, etc.



Outlier Analysis

- **Definition:** Process of identifying data-points that deviate significantly from the rest of the dataset.
- Also known as anomaly-detection.
- Outliers can represent errors, noise, or rare events.

Detection Methods

i) Statistical Tests

- Using statistical-measures to identify objects that are unlikely to occur under normal conditions.

ii) Distance Measures

- Calculating distances between objects to identify those that are farthest from the rest of the data.

iii) Density-Based Methods

- Identifying outliers based on their deviation from the local-density of surrounding data-objects.

Problems and Issues in DM

Data mining systems face a lot of data mining challenges and issues in today's world some of them are:

- Mining methodology and user interaction issues
- Performance issues
- Issues relating to the diversity of database types

1. **Mining methodology and user interaction issues**:

Mining different types of knowledge in the database:
Different users – different knowledge – different ways. This means that different customers need different types of information, so it is difficult to cover the large amount of data that can meet customer needs.

Interactive mining of multi-level abstract knowledge:
Interactive mining allows users to intensively search for patterns from different perspectives. The data mining process should be interactive because it is difficult to know what to find in the database.

The integration of background knowledge:
Background knowledge is used to guide the discovery process and to express discovered patterns.

Query language and special mining:
Relational query languages such as SQL allow users to ask special queries for data retrieval. The data mining query language should exactly match the query language of the data warehouse.

Handling noisy or incomplete data:
In large databases, many attribute values are incorrect. This could be due to human error or any instrument malfunction. Data cleaning methods and data analysis methods are used to deal with noisy data.

2. Performance issues

Efficiency and scalability of data mining algorithms:
In order to effectively extract information from a large amount of data in the database, data mining algorithms must be efficient and scalable.

Parallel, distributed, and incremental mining algorithms:
The enormous size of many databases, the wide distribution of data, and the complexity of some data mining methods are factors driving the development of parallel and distributed data mining algorithms. This algorithm divides data into multiple partitions and processes them in parallel.

3. Issues relating to the diversity of database types:

Handling relational and complex types of data:
Databases and data warehouses store a variety of data. It is impossible for a system to mine all these types of data. Therefore, different data mining systems should be explained for different types of data.

Mining information from heterogeneous databases and global information systems: As data is obtained from different data sources on Local Area Network (LAN) and Wide Area Network (WAN). Discovering knowledge from different structured resources is a big challenge in data mining

Challenges in Data Mining

1. **Data Quality**

    - The quality of data used in data mining is one of the most significant challenges.
    - The accuracy, completeness, and consistency of the data affect the accuracy of the results obtained.
    - The data may contain errors, omissions, duplications, or inconsistencies, which may lead to inaccurate results.
    - Moreover, the data may be incomplete, meaning that some attributes or values are missing, making it challenging to obtain a complete understanding of the data.
    - To address these challenges, data mining practitioners must apply data cleaning and data pre-processing techniques to improve the quality of the data.

2. **Data Complexity**

    - Data complexity refers to the vast amounts of data generated by various sources, such as sensors, social media, and the internet of things (IoT).
    - The complexity of the data may make it challenging to process, analyze, and understand.
    - In addition, the data may be in different formats, making it challenging to integrate into a single dataset.
    - To address this challenge, data mining practitioners use advanced techniques such as clustering, classification, and association rule mining.

3. **Data Privacy and Security**

    - Data privacy and security is another significant challenge in data mining. As more data is collected, stored, and analyzed, the risk of data breaches and cyber-attacks increases.
    - The data may contain personal, sensitive, or confidential information that must be protected.
    - To address this challenge, data mining practitioners must apply data anonymization and data encryption techniques to protect the privacy and security of the data.

4. **Scalability**

    - Data mining algorithms must be scalable to handle large datasets efficiently.
    - As the size of the dataset increases, the time and computational resources required to perform data mining operations also increase.
    - Moreover, the algorithms must be able to handle streaming data, which is generated continuously and must be processed in real-time.
    - To address this challenge, data mining practitioners use distributed computing frameworks such as Hadoop and Spark.

5. **Interpretability**

- Data mining algorithms can produce complex models that are difficult to interpret.
- This is because the algorithms use a combination of statistical and mathematical techniques to identify patterns and relationships in the data.
- Moreover, the models may not be intuitive, making it challenging to understand how the model arrived at a particular conclusion.
- To address this challenge, data mining practitioners use visualization techniques to represent the data and the models visually.

Application of data mining

1. **Scientific Analysis:**

- Scientific simulations are generating bulks of data every day.
- This includes data collected from nuclear laboratories, data about human psychology, etc.
- Data mining techniques are capable of the analysis of these data. Example of scientific analysis:
  - Sequence analysis in bioinformatics
  - Classification of astronomical objects
  - Medical decision support.

2. **Intrusion Detection:**
   - A network intrusion refers to any unauthorized activity on a digital network.
   - Network intrusions often involve stealing valuable network resources.
   - Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies.
   - For example:
     - Detect security violations
     - Misuse Detection
     - Anomaly Detection
3. **Business Transactions**:
   - Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations.
   - The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world.
   - Data mining helps to analyze these business transactions and identify marketing approaches and decision-making.
   - Example :
     - Direct mail targeting
     - Stock trading
     - Customer segmentation
4. **Market Basket Analysis:**
   - Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket.
   - This concept identifies the pattern of frequent purchase items by customers.
   - This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task.

- Example:
  - Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
  - Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

5. **Education:**
   - For analyzing the education sector, data mining uses Educational Data Mining (EDM) method.
   - This method generates patterns that can be used both by learners and educators.
   - By using data mining EDM we can perform some educational task:
     - Predicting students admission in higher education
     - Predicting students profiling
     - Predicting student performance
     - Teachers teaching performance
     - Curriculum development
     - Predicting student placement opportunities
6. **Healthcare and Insurance**:
   - Claims analysis i.e which medical procedures are claimed together.
   - Identify successful medical therapies for different illnesses.
   - Characterizes patient behavior to predict office visits.
7. **Financial/Banking Sector:**
   - A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.
     - Credit card fraud detection.
     - Identify 'Loyal' customers.
     - Extraction of information related to customers.
     - Determine credit card spending by customer groups.
8. **Transportation:**
   - A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services.
   - A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.
     - Determine the distribution schedules among outlets.
     - Analyze loading patterns.