102046706 – Data Mining
& Business Intelligence

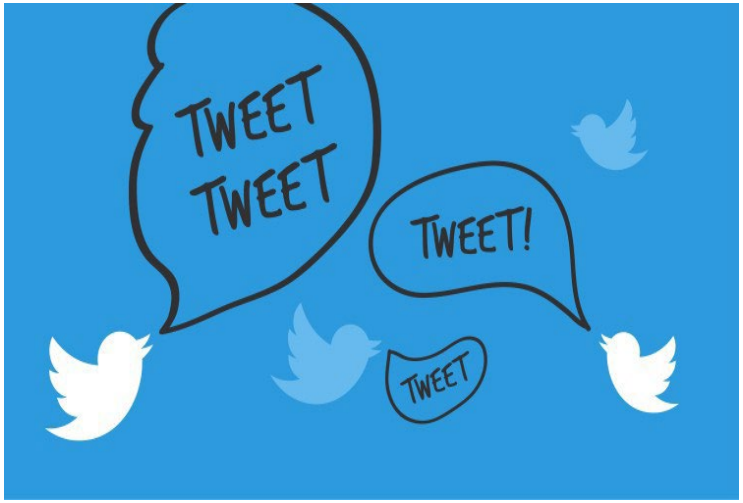
Unit-2

Introduction to Data Mining (DM)

Outline

- Motivation: Why data mining?
- What is data mining?
- Data mining functionalities
- Classification of Data mining systems
- Data Mining: On what kind of data?
- Data Mining Architecture
- KDD Process
- Data mining issues

Motivation : Why Data Mining?



Data Mining

Twitter Trends

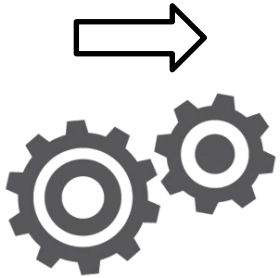
Google Trends

Motivation : Why Data Mining?

"Necessity is the Mother of Invention"

Data Explosion Problem

Solution



"Data Mining"

Extraction of interesting Knowledge from data in large databases

"It has been estimated that the amount of **information** in the world **doubles** every **10** months."

- There is a tremendous increase in the amount of data recorded and stored on digital media as well as individual sources.

Why Data Mining? (Cont..)

"We are drowning in data, but starving for knowledge!"
"Data rich but Information poor"

- Since the 1960's, database and information technology has been changed systematically from primitive file processing systems to powerful database systems.
- The research and development in database systems since the 1970's has led to the development of relational database systems.

Why Data Mining? (Cont..)

Years	Evolution
Since 1960's	Data collection, database creation, IMS (hierarchical database system by IBM) and network DBMS
1970s	Relational data model, relational DBMS implementation
1980s	RDBMS, advanced data models, application-oriented DBMS (spatial, scientific, engineering, etc.)
1990s	Data mining, data warehousing, multimedia databases, and web databases
2000s	Stream data management and mining, Social Networks (Facebook, etc.), web technology (XML) and global information systems
At Present	Heterogeneous database systems, big data

Every day data **grows exponentially**,
but these **all data** are really **important to us**??

What is Data Mining?

1. Data mining refers to extracting or “mining” knowledge from large amounts of data.
2. “Knowledge mining from data” or “Knowledge mining”
3. “Extract knowledge from large data or databases”
4. “Knowledge discovery from database (KDD)”

What is Data Mining? (Cont..)

- It is the **computational process** of **discovering patterns** in **large data sets** involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

The overall goal of the data mining process is to **extract information from a large data sets** and **transform it into an understandable structure** for further use.

What is Data Mining? (Cont..)

Data ☐Knowledge ☐Action ☐Goal

Netflix collects user ratings of movies (**data**) => What types of movies you will like (**knowledge**) => Recommend new movies to you (**action**) => Users stay with Netflix (**goal**)

Gene sequences of cancer patients (**data**) => Which genes lead to cancer? (**knowledge**) => Appropriate treatment (**action**) => Save life (**goal**)

Road traffic (**data**) => Which road is likely to be congested? (**knowledge**) => Suggest better routes to drivers (**action**) => Save time and energy (**goal**)

Data Mining Functionalities

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

- **Descriptive**
  - These tasks present the **general properties** of data stored in database.
  - The descriptive tasks are used to find out patterns in data.
  - **E.g.** : Cluster, correlation, trends etc.

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

Domains of Data Mining Systems

- Data mining is an **interdisciplinary field**, joining of a set of disciplines, including database systems, statistics, machine learning, visualization and information science.

Classification of Data Mining Systems

- **Classification of data mining & Multi-Dimensional View of Data Mining are similar terms.**
- Classification of data mining based on..
  1. **Databases** to be mined
  2. **Knowledge** to be mined
  3. **Techniques/Methods** utilized
  4. **Application** adapted

Classification of Data Mining Systems

1. Classification according to the kinds of **databases** mined:
   - Classified **according to different criteria** (such as data models, or the types of data or applications involved), each of which **may require its own data mining technique**.
   - For instance, if classifying according to data models, we may have a relational, transactional, object-oriented, object-relational, or data warehouse mining system.
   - If classifying according to the **special data types**, we may have a **spatial, time-series, text or multimedia data** mining system or a **world-wide web** mining system.
   - Other system types include **heterogeneous data mining systems** and legacy data mining systems.

Classification of Data Mining Systems (Cont..)

2. Classification according to the kinds of **knowledge** mined:
   - Based on data mining functionalities,
     - Characterization
     - Discrimination
     - Association
     - Correlation analysis
     - Classification & prediction
     - Clustering
     - Outlier analysis

Classification of Data Mining Systems (Cont..)

3. Classification according to the kinds of **techniques** utilized:
   - These techniques can be described according to the **degree of user interaction** involved (e.g., autonomous systems, query-driven systems).
   - The methods of data analysis employed (e.g., database-oriented or data warehouse–oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks etc.)
   - A sophisticated data mining system will often adopt multiple data mining techniques for work out an effective, integrated technique which combines the merits of a few individual approaches.
   - **E.g.** Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

Classification of Data Mining Systems (Cont..)

4. Classification according to the **Applications** adapted:
   - Retail
   - Telecommunication
   - Banking
   - Fraud analysis
   - Stock market analysis
   - Text mining
   - Web mining etc.

Data Mining—On what kind of data?

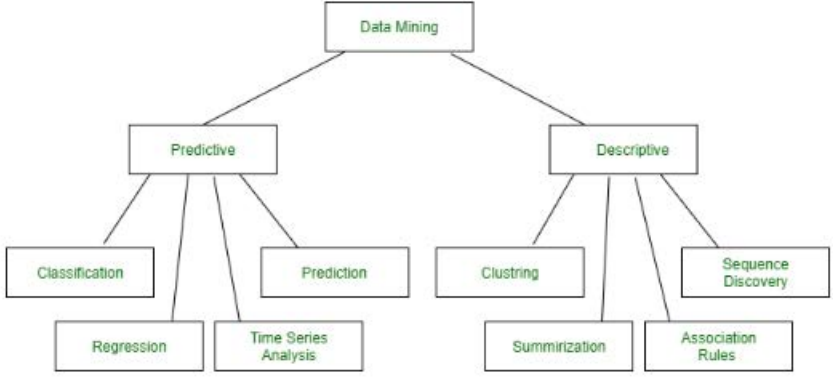
- **Relational Databases:**
  - A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
  - **E.g.** : SQL Server, Oracle etc.
- **Data Warehouses:**
  - A data warehouse is a repository of information collected from multiple sources.
  - Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
  - **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 like Flipkart, Amazon etc.
- **Other Data**
  - Spatial data (Maps or Location)
  - Engineering design data (Design of Buildings, Offices Structures)
  - Hypertext and multimedia data (Including text, image, video, and audio data), the World Wide Web (a huge, widely distributed information repository made available on the Internet).

Types of Data Mining Models –

1. Predictive Models
2. Descriptive Models



Data Mining Models

Predictive Model :

A predictive model constitutes prediction concern values of data using known results found from various data. Predictive modelling may be made based on the use of variant historical data. Predictive model data mining tasks comprise regression, time series analysis, classification, prediction.

The Predictive Model is known as **Statistical Regression**. It is a monitoring learning technique that Incorporates an explication of the dependency of few attribute values upon the values of other attributes In a similar item and the growth of a model that can predict these attribute values for recent cases.

- **Classification** –
  It is the act of assigning objects to one of several predefined categories. Or we can define classification as a learning function of a target function that sets each attribute to a predefined class label.
- **Regression** –
  It is used for appropriate data. It is a technique that verifies data values for a function. There are two types of regression –
  1. **Linear Regression** is associated with the search for the optimal line to fit the two attributes so that one attribute can be applied to predict the other.
  2. **Multi-Linear Regression** involves two or more than two attributes and data are fit to multidimensional space.
- **Time Series Analysis** –
  It is a set of data based on time. Time series analysis serves as an independent variable to estimate the dependent variable in time.
- **Prediction** –
  It predicts some missing or unknown values.

Description Model :

A descriptive model distinguishes relationships or patterns in data. Unlike Predictive Model, a descriptive model serves as a way to explore the properties of data being examined, not to predict new properties, clustering, summarization, associating rules, and sequence discovery are descriptive model data mining tasks.

Descriptive analytics Concentrate on the summarization and conversion of the data into significant information for monitoring and reporting.

- **Clustering** –
  It is the technique of converting a group of abstract objects into classes of identical objects.
- **Summarization** –
  It holds a set of data in a more in-depth, easy-to-understand form.
- **Associative Rules** –
  They find an exciting consistency or causal relationship between a large set of data objects.
- **Sequence** –
  It is the discovery of interesting patterns in the data is in relation to some objective or subjective measurement of how interesting it is.

Data Mining Task Primitives

Data mining task primitives refer to the basic building blocks or components that are used to construct a data mining process. These primitives are used to represent the most common and fundamental tasks that are performed during the data mining process. The use of data mining task primitives can provide a modular and reusable approach, which can improve the performance, efficiency, and understandability of the data mining process.

Task relevant data

- Database Name
- Database tables
- Relevant attributes
- Data grouping criteria

Type of knowledge to be mined

- Classification
- Clustering
- Prediction
- Discrimination
- Correlation analysis

Background knowledge

- Concept Hierarchy
- User beliefs about relationships in data

Measures of patterns

- Simplicity
- Novelty
- Certainty
- Utility

Visualization of patterns

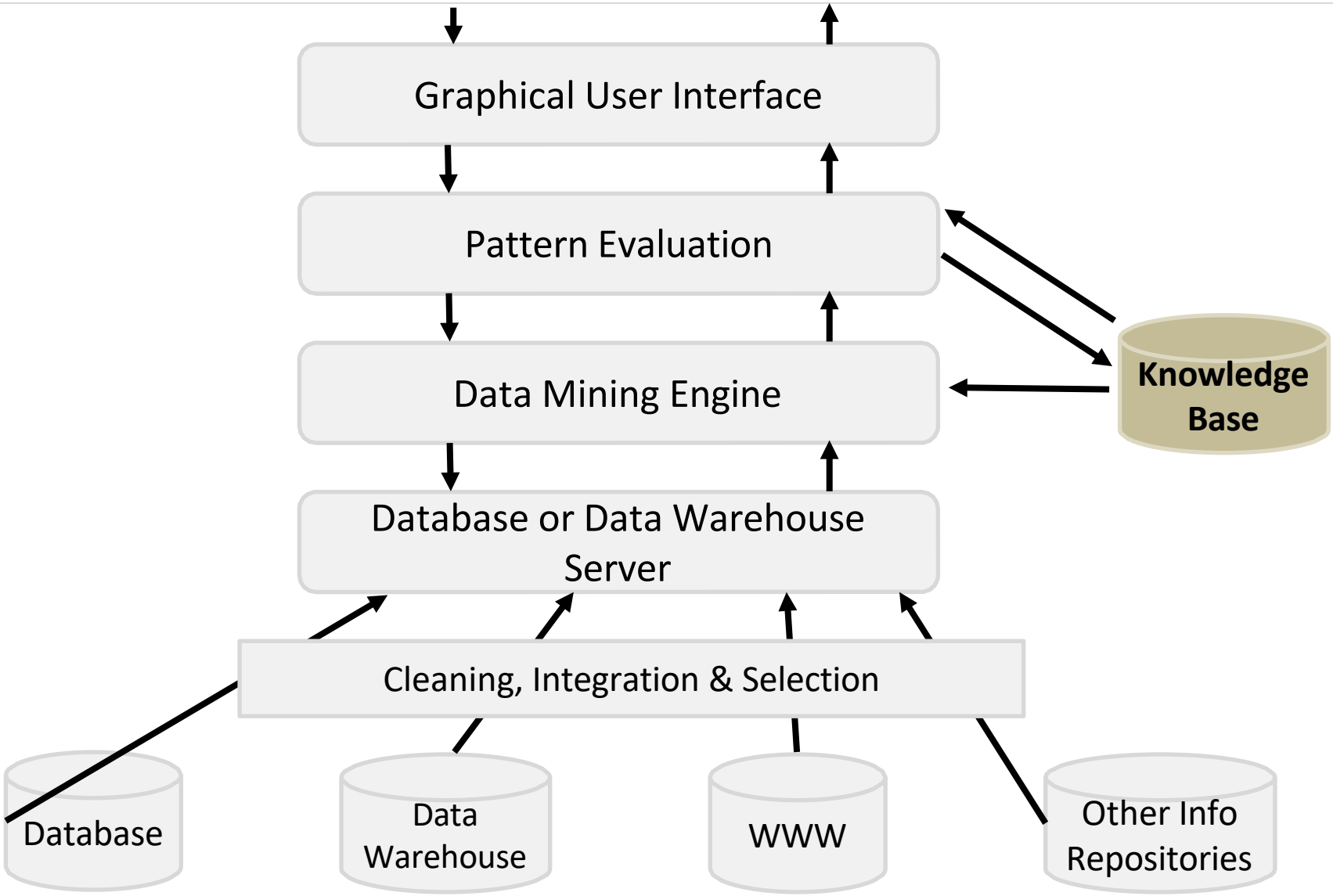
- Visualization of discovered patterns
- Cubes
- Charts
- Tables
- Graphs

The Data Mining Task Primitives are as follows:

1. **The set of task relevant data to be mined:** It refers to the specific data that is relevant and necessary for a particular task or analysis being conducted using data mining techniques. This data may include specific attributes, variables, or characteristics that are relevant to the task at hand, such as customer demographics, sales data, or website usage statistics. The data selected for mining is typically a subset of the overall data available, as not all data may be necessary or relevant for the task. For **example**: Extracting the database name, database tables, and relevant required attributes from the dataset from the provided input database.
2. **Kind of knowledge to be mined**: It refers to the type of information or insights that are being sought through the use of data mining techniques. This describes the data mining tasks that must be carried out. It includes various tasks such as classification, clustering, discrimination, characterization, association, and evolution analysis. For **example**, It determines the task to be performed on the relevant data in order to mine useful information such as classification, clustering, prediction, discrimination, outlier detection, and correlation analysis.
3. **Background knowledge to be used in the discovery process:** It refers to any prior information or understanding that is used to guide the data mining process. This can include domain-specific knowledge, such as industry-specific terminology, trends, or best practices, as well as knowledge about the data itself. The use of background knowledge can help to improve the accuracy and relevance of the insights obtained from the data mining process. For **example**, The use of background knowledge such as concept hierarchies, and user beliefs about relationships in data in order to evaluate and perform more efficiently.

4. **Interestingness measures and thresholds for pattern evaluation:** It refers to the methods and criteria used to evaluate the quality and relevance of the patterns or insights discovered through data mining. Interestingness measures are used to quantify the degree to which a pattern is considered to be interesting or relevant based on certain criteria, such as its frequency, confidence, or lift. These measures are used to identify patterns that are meaningful or relevant to the task. Thresholds for pattern evaluation, on the other hand, are used to set a minimum level of interestingness that a pattern must meet in order to be considered for further analysis or action. For **example**: Evaluating the interestingness and interestingness measures such as utility, certainty, and novelty for the data and setting an appropriate threshold value for the pattern evaluation.
5. **Representation for visualizing the discovered pattern:** It refers to the methods used to represent the patterns or insights discovered through data mining in a way that is easy to understand and interpret. Visualization techniques such as charts, graphs, and maps are commonly used to represent the data and can help to highlight important trends, patterns, or relationships within the data. Visualizing the discovered pattern helps to make the insights obtained from the data mining process more accessible and understandable to a wider audience, including non-technical stakeholders. For **example** Presentation and visualization of discovered pattern data using various visualization techniques such as barplot, charts, graphs, tables, etc.

Data Mining Architecture



Data Mining Architecture (Cont..)

- **Data Mining Engine:**
  - It is essential to the data mining system and ideally consists of a set of **functional modules (knowledge) & methods** for different tasks such as…
    - Characterization
    - Association
    - Correlation analysis
    - Classification & prediction
    - Cluster analysis
    - Outlier analysis

Data Mining Architecture (Cont..)

- **Pattern Evaluation Module:**
  - This component typically employs **interestingness measures** and **interacts with the data mining modules** so it is focus in the search towards **interesting patterns**.
  - The pattern evaluation module integrated with the mining module, depending on the implementation of the data mining method used.

Data Mining Architecture (Cont..)

- **Knowledge base:**
  - Knowledge base is the **domain knowledge** that is used to guide the search or **evaluate the interestingness of resulting patterns**.
  - Such knowledge can include concept hierarchies, used to organize attributes or **attribute values into different levels of abstraction**.
  - Knowledge is such as **user beliefs**, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included.

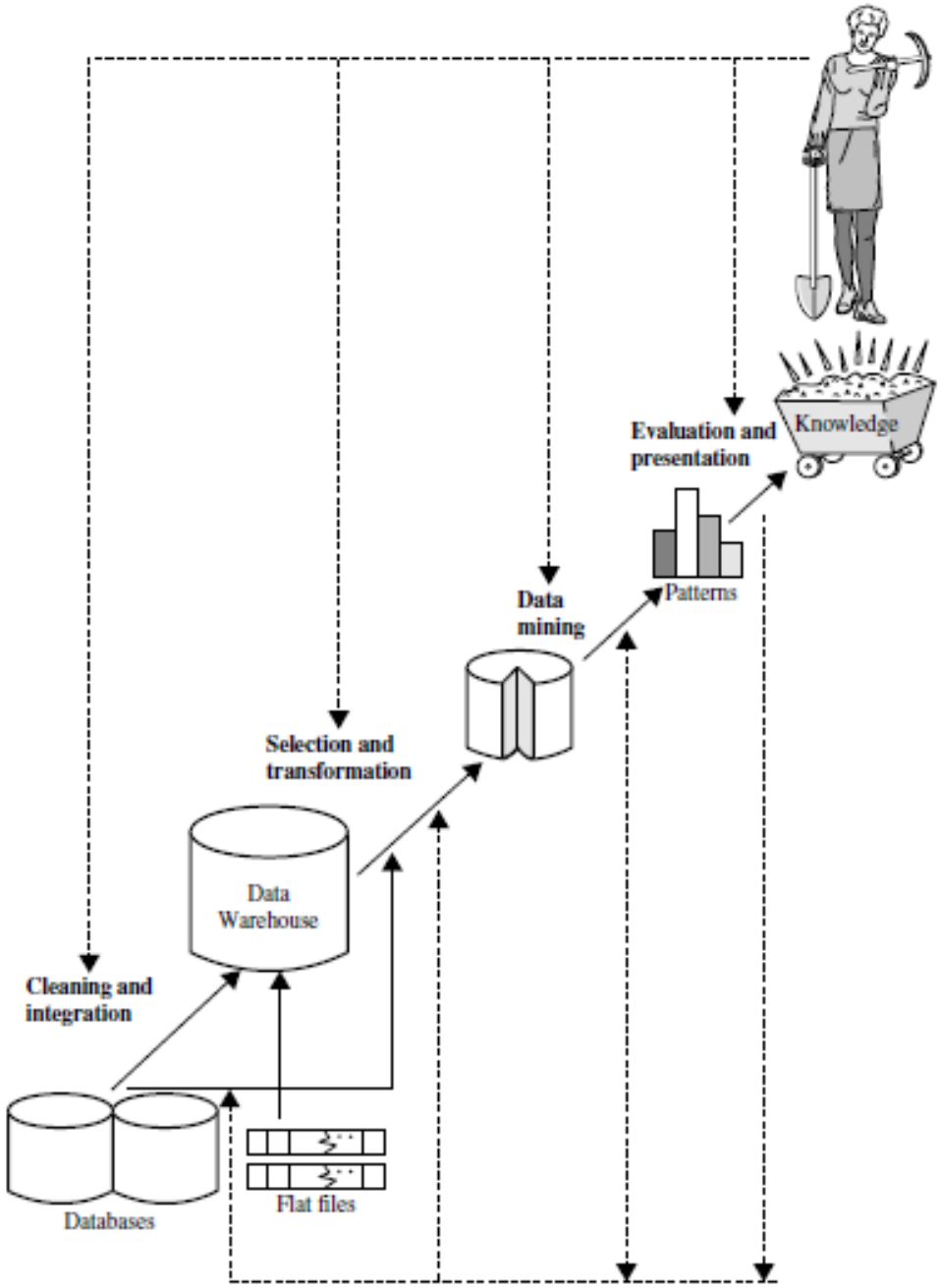
KDD (Knowledge Discovery in Databases) Process

- Knowledge discovery in databases is a process of an iterative sequence of the following steps:
  1. **Selection**
  2. **Preprocessing**
  3. **Transformation**
  4. **Data Mining**
  5. **Pattern Evaluation**
  6. **User Interface (Visualization of Pattern or Knowledge)**

KDD Process

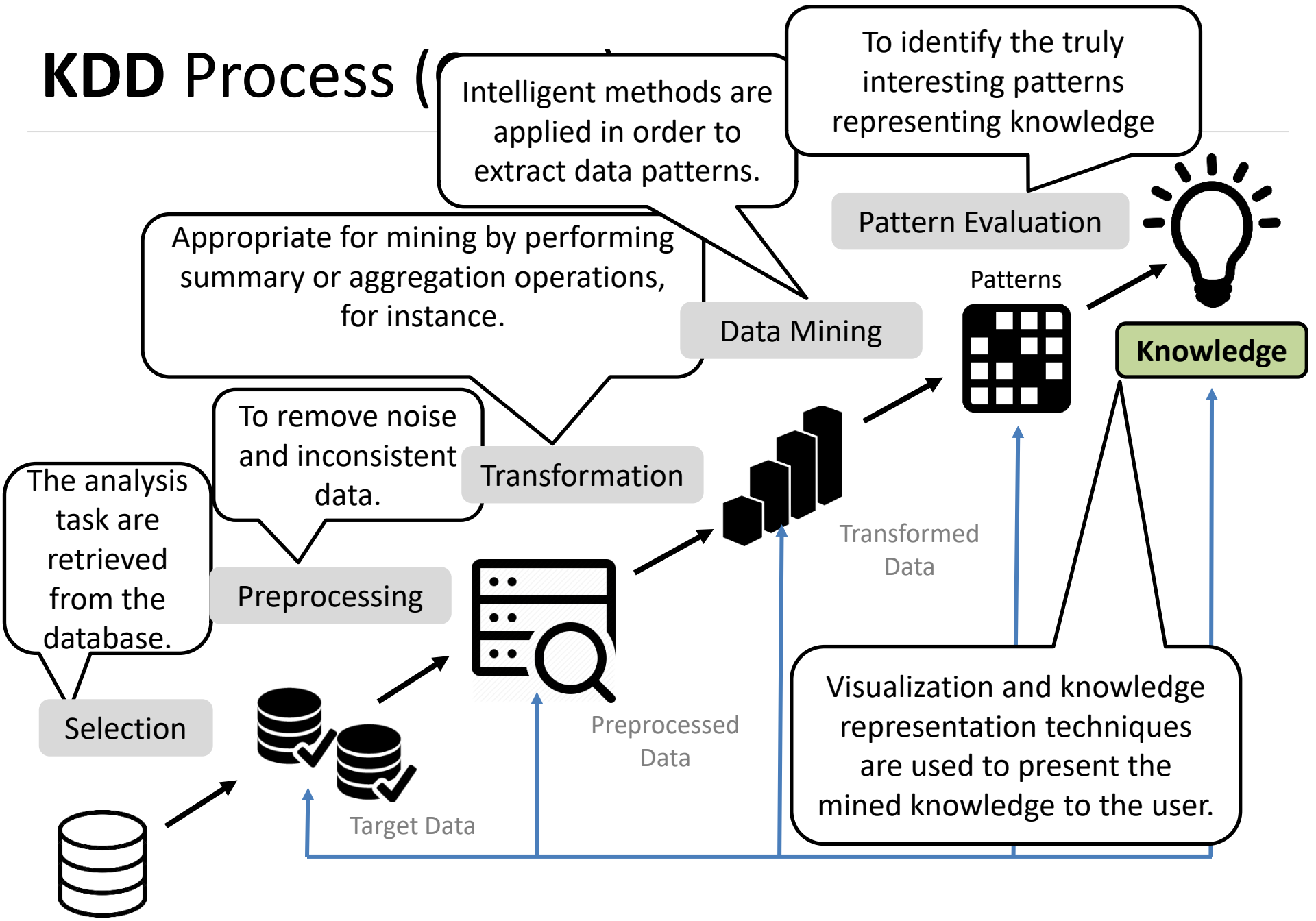
- The knowledge discovery process is an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)



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

KDD Process (


To identify the truly interesting patterns representing knowledge
Intelligent methods are applied in order to extract data patterns.
Pattern Evaluation
Appropriate for mining by performing summary or aggregation operations, for instance.
Patterns
Data Mining
Knowledge
To remove noise and inconsistent data.
Transformation
The analysis task are retrieved from the database.
Transformed Data
Preprocessing
Selection
Preprocessed Data
Visualization and knowledge representation techniques are used to present the mined knowledge to the user.
Target Data

KDD Process (Cont..)

- **Data Selection:** Where data relevant to the analysis task are retrieved from the database.
- **Data Cleaning:** To remove noise and inconsistent data.
- **Data Integration:** Where multiple data sources may be combined.
- **Data Transformation**: Where data are transformed or consolidated into appropriate forms for mining by performing summary or aggregation operations.
- **Data Mining**: An essential process where intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation**: To identify the truly interesting patterns representing knowledge based on some interestingness measures.
- **Knowledge Presentation**: Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data Mining Issues

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

1) Mining Methodology

- **Mining various and new kinds of knowledge**
  - Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, so these tasks may use the same database in **different ways and require the development of numerous data mining techniques**.
- **Mining knowledge in multidimensional space**
  - When searching for knowledge in large data sets, we can explore the data in multidimensional space.
  - That is, we can search for interesting patterns among **combinations of dimensions (attributes)** at varying levels of abstraction. Such mining is known as (exploratory) **multidimensional data mining**.

1) Mining Methodology (Cont..)

- **Data mining—an interdisciplinary effort**
  - The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines.
  - For example, to mine data with natural language **text**, it makes sense to fuse data mining methods of **information retrieval** and **natural language processing**.
- **Handling uncertainty, noise, or incompleteness of data**
  - Data often contain **noise, errors, exceptions, uncertainty** or **incomplete**.
  - Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns.

2) User Interaction

- **Interactive mining**
  - The data mining process should be **highly interactive**. Thus, it is important to build **flexible user interfaces** and an exploratory mining environment, facilitating the user's interaction with the system.
- **Incorporation of background knowledge**
  - **Background knowledge, constraints, rules, and other information** regarding the domain under study should be incorporated into the knowledge discovery process.
- **Presentation and visualization of data mining results**
  - How any system can present data mining results, vividly(clear image in mind) and flexibly ?, so that the **discovered knowledge can be easily understood and directly usable by humans**.

3) Efficiency and Scalability

- **Efficiency and scalability of data mining algorithms**
  - Data mining **algorithms** must be **efficient and scalable** in order to effectively extract information from huge amounts of data lies in many data repositories or in dynamic data streams.
  - In other words, the **running time** of a data mining algorithm must be **predictable, short, and acceptable by applications**.
  - Efficiency, scalability, performance, optimization, and the ability to **execute in real time** are key criteria for **new mining algorithms**.
- **Parallel, distributed, and incremental mining algorithms**
  - The giant size of many data sets, the **wide distribution of data**, and the **computational complexity** of some data mining methods are factors that motivate the development of parallel and distributed data-intensive mining algorithms.

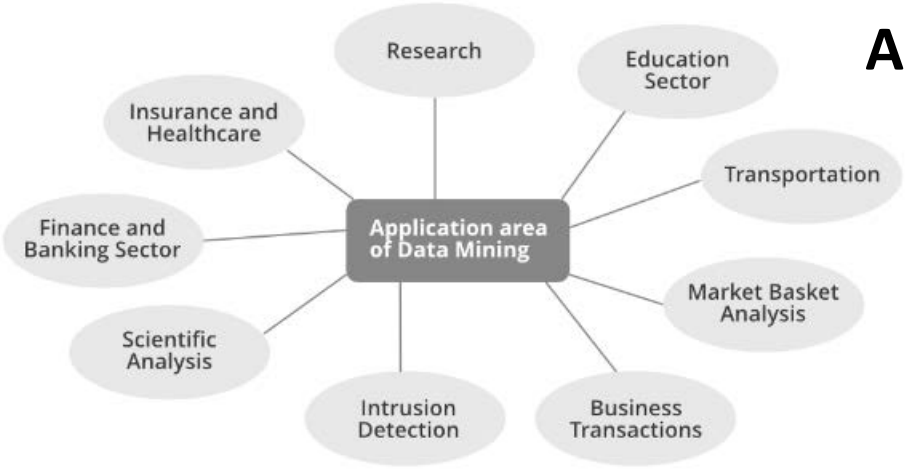
4) Diversity of Database Types

- **Handling complex types of data**
  - Data mining is how to uncover knowledge from **stream**, **time-series**, **sequence**, **graph**, **social network**, and **multirelational data**.
  - In mining various types of attributes are available and also different types of data in database or dataset.
- **Mining dynamic, networked, and global data repositories**
  - Data from multiple sources are connected by the Internet and various **kinds of networks** like **distributed** and **heterogeneous global information systems**.
  - The discovery of knowledge from **different sources** of **structured**, **semi-structured**, or **unstructured** challengeable.
  - **Web Mining**, **multisource data mining** and **information network mining** have become **challenging and fast-evolving data mining fields**.

5) Data Mining and Society

- **Social impacts of data mining**
  - With data mining penetrating our everyday lives, it is important to study the impact of data mining on society, How can we used at a mining technology to **benefit our society**? **How can we guard against its misuse**?
- **Privacy-preserving data mining**
  - Data mining will help in scientific discovery, business management, economy recovery, and security protection (e.g., the real-time discovery of intruders and cyber attacks).
  - However, it poses the risk of **disclosing** an **individual's personal information**.
- **Invisible data mining**
  - We cannot expect everyone in society to learn and master in data mining techniques.
  - For example, when purchasing items online, users may be unaware that the store is likely collecting data on the buying patterns of its customers, which may be **used to recommend other items** for **purchase in the future**.

APPLICATIONS OF DATA MINING

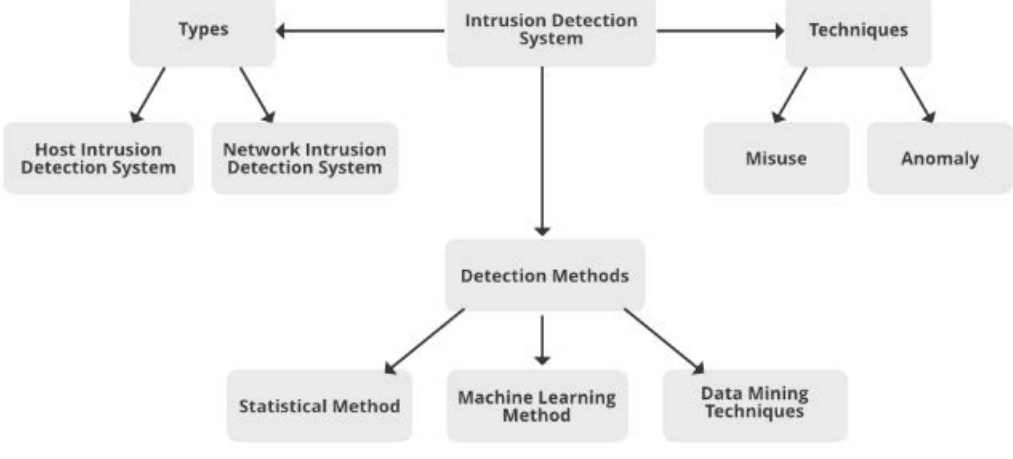


Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

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. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations
- Misuse Detection
- Anomaly Detection



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
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

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.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

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

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things)and Cybersecurity
- Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.
- Characterizes patient behavior to predict office visits.

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.

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.