UNIT-3

ISSUES IN DM – KDD PROCESS

Motivation for Data Mining - Data Mining-Definition and Functionalities – Classification of DM Systems - DM task primitives - Integration of a Data Mining system with a Database or a Data Warehouse - Issues in DM – KDD Process

DATA MINING – DEFINITION

Data Mining is the process of discovering interesting patterns (or Knowledge) from large amounts of data.

Data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data Mining refers to extracting or "Mining" Knowledge from large amount of data.

Data Mining Functionalities- What kinds of patterns can be mined?

Data mining functionalities are used to represent the type of patterns that have to be discovered in **data mining tasks**. In general,

Data mining tasks can be classified into **two types** including

1.**Descriptive** - Descriptive task Characterize the general properties of the data in the database.

2. **Predictive** - Predictive mining tasks act inference on the current information to develop predictions.

Data mining functionalities, and the kinds of patterns they can discover,are described below:

1. Class/Concept Descriptions : Characterization and Discrimination

2, Mining Frequent Patterns, Associations, and Correlations

3. Classification and Predictions

4. Cluster analysis

5.Outlier Analysis

6.Evolution Analysis

1. Class/Concept Descriptions : Characterization and Discrimination**:**

Data can be associated with classes or concepts, concept refers to a collection of data items such as Computers,printers etc.

Concepts of Customers-bigSpenders,budgetSpenders,...

How to describe these items or concepts?

Descriptions can be derived as

- **Data Characterization**
- **Data Discrimination**
- **Or both of the Data Characterization & Data Discrimination**

Data Characterization: This refers to the summarizing the general characteristics of a target class of data. The output of the data characterization can be presented in various forms include pie charts, bar charts, curves, multidimensional data cubes.

Example: To study the characteristics of software products with sales increased by 10% in the previous years. To summarize the characteristics of the customer who spend more than $5000 a year at AllElectronics, the result is general profile of those customers such as that they are 40-50 years old, employee and have excellent credit rating.

Data Discrimination: It comparing the target class with one or set of classes. It is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

Example: Compare the general features of software products,whose sales increase by 10% in the last year with those whose sales decrease by 30% during the same period,

2. Mining Frequent Patterns, Associations, and Correlations: Frequent patterns are nothing but things that are found to be most common in the data. There are different kinds of frequencies that can be observed in the dataset.

- **Frequent item set:** This applies to a number of items that can be seen together regularly for eg: milk and sugar.
- **Frequent Subsequence:** This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.
- **Frequent Substructure:** It refers to the different kinds of data structures such as trees and graphs that may be combined with the itemset or subsequence.

Association Analysis: The process involves uncovering the relationship between data and deciding the rules of the association. It is a way of discovering the relationship between various items.

Example: Suppose we want to know which items are frequently purchased together. An example for such a rule mined from a transactional database is,
buys (X, "computer") ⇒ buys (X, "software") [support = 1%, confidence = 50%],
where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all the transactions under analysis show that computer and software are purchased together.

Correlation Analysis: Correlation is a mathematical technique that can show whether and how **strongly the pairs of attributes are related to each other**. For example, Highted people tend to have more weight.

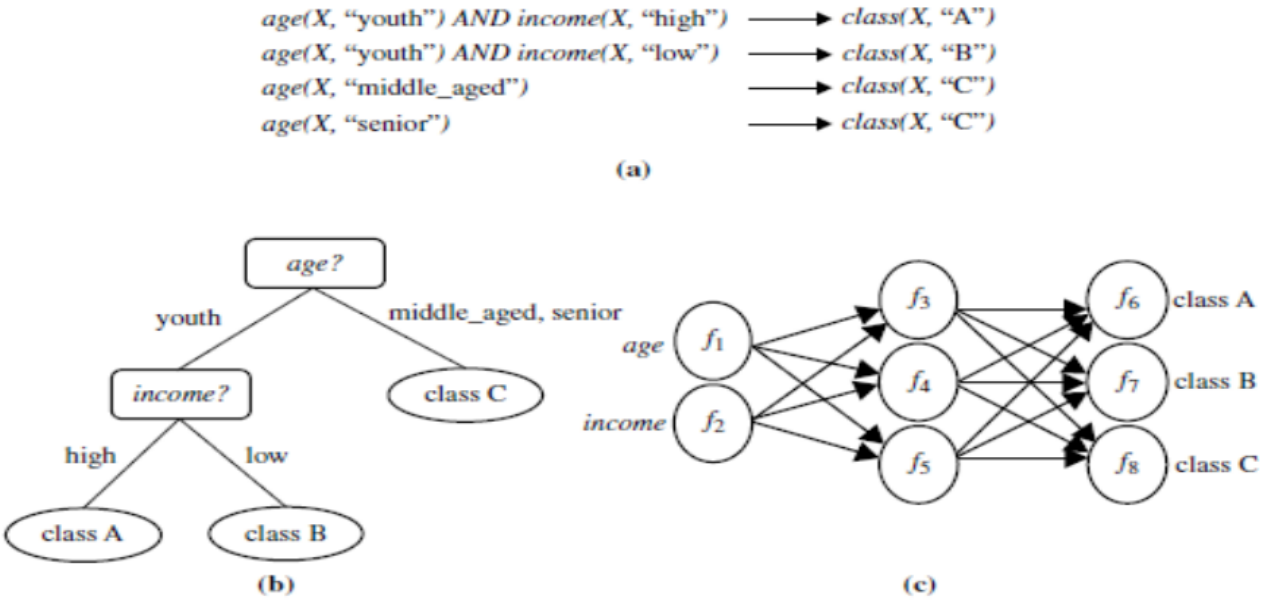
3. Classification and Predictions

Classification

- The process of finding a model that describes and distinguishes the data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- The derived model is based on the analysis of a set of training data(data objects whose class label is known)
- The model can be represented in Classification (IF-THEN) rules, Decision tree, Neural networks,etc.,

A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules.

A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naive Bayesian classification, Support vector machines, and k-nearest-neighbor classification. Whereas classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.



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

- **Prediction** − It defines predict some unavailable data values or pending trends. An object can be anticipated based on the attribute values of the object and attribute values of the classes. It can be a prediction of missing numerical values or increase/decrease trends in time-related information.

4. Cluster analysis

Class label is unknown group data to form new classes. Clusters of objects are formed based on the principle of maximizing intra-class similarity and minimizing interclass similarity Unlike classification and

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

regression, which analyze class-labeled (training) data sets, clustering analyzes data objects without consulting class labels. In many cases, class labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data.

5. Outlier Analysis

A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers. 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. The analysis of outlier data is referred to as outlier analysis or anomaly mining. For example, Outlier analysis. 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 .

6. Evolution analysis

Evolution analysis is the study of data sets that may have undergone a stage of transformation or change.

Evolution analysis describes and models regularities or trends for objects whose behavior changes over time

It provides time-related data clustering and assists in finding trends or changes with features like periodicity, time-series data, and trend similarity.

In addition to aiding in data classification, characterisation, discrimination, and grouping for multivariate time series, the evolution analysis model represents evolving trends in data.

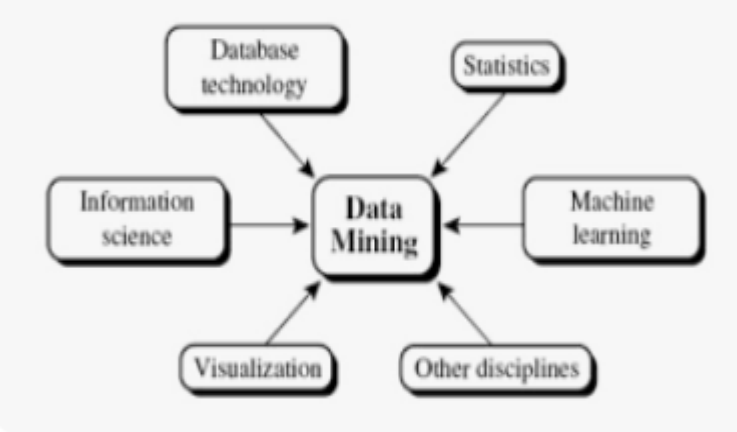
Example: Identify Stock evolution regularities for overall stocks and for the stocks of particular companies.

Classification of Data mining Systems

Data Mining is considered as an interdisciplinary field. It includes a set of various disciplines such as statistics, database systems, machine learning, visualization and information sciences.

Classification of the data mining system helps users to understand the system and match their requirements with such systems.

The data mining system can be classified according to the following criteria:



Some Other Classification Criteria:

- Classification according to kind of databases mined
- Classification according to kind of knowledge mined
- Classification according to kinds of techniques utilized
- Classification according to applications adapted

Classification according to kind of databases mined

We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object- relational, or data warehouse mining system.

- Relational
- Transactional
- Object-relational
- Data warehouse mining system

Classification according to kind of knowledge mined

We can classify the data mining system according to kind of knowledge mined. It is means data mining system are classified on the basis of functionalities such as: Characterization

- Data Characterization & Data Discrimination

- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Classification according to kinds of techniques utilized

We can classify the data mining system according to kind of techniques used. We can describes these techniques according to degree of user interaction involved or the methods of analysis employed.

Classification according to applications adapted

We can classify the data mining system according to application adapted. These applications are as follows:

- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail

Data mining Task primitives

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during the mining process to discover interesting patterns.

Here is the list of Data Mining Task Primitives

- Set of task relevant data to be mined.
- Kind of knowledge to be mined.
- Background knowledge to be used in discovery process.
- Interestingness measures and thresholds for pattern evaluation.
- Representation for visualizing the discovered patterns.

Primitives for specifying a data mining task

• Task-relevant data: This primitive specifies the data upon which mining is to be performed.

• It involves specifying– the database and tables or data warehouse

containing the relevant data, conditions for selecting the relevant data, the relevant attributes or dimensions for exploration, and instructions regarding the ordering or grouping of the data retrieved.

Knowledge type to be mined: This primitive specifies the specific data mining function to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis.

• As well, the user can be more specific and provide pattern templates that all discovered patterns must match. These templates or meta patterns (also called meta rules or meta queries), can be used to guide the discovery.

Background knowledge: This primitive allows users to specify knowledge they have about the domain to be mined.

• Such knowledge can be used to guide the knowledge discovery process and evaluate the patterns that are found.

• The several kinds of background knowledge, this chapter focuses on concept hierarchies.

Pattern interestingness measure: This primitive allows users to specify functions that are used to separate uninteresting patterns from knowledge and may be used to guide the mining process, as well as to evaluate the discovered patterns.

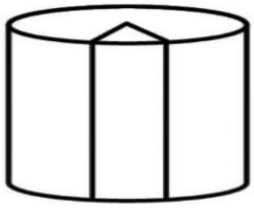
• This allows the user to confine the number of uninteresting patterns returned by the process, as a data mining process may generate a large number of patterns

Interestingness measures can be specified for such pattern characteristics as simplicity, certainty, utility and novelty.

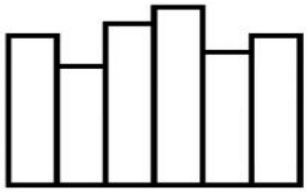
• **Visualization of discovered patterns:** This primitive refers to the form in which discovered patterns are to be displayed.

• In order for data mining to be effective in conveying knowledge to users, data mining systems should be able to display the discovered patterns in multiple forms such as rules, tables, cross tabs (cross-tabulations), pie or bar charts, decision trees, cubes or other visual representations.

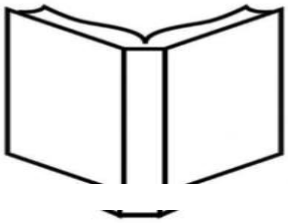
A data mining query language can be designed to incorporate these primitives, allowing users to flexibly interact with data mining systems. Having a data mining query language provides a foundation on which user-friendly graphical interfaces can be built.



Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions



Knowledge type to be mined
Characterization & Discrimination
Association
Classification
prediction
Clustering



Background knowledge
Concept hierarchies
User beliefs about relationships in the data

Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty

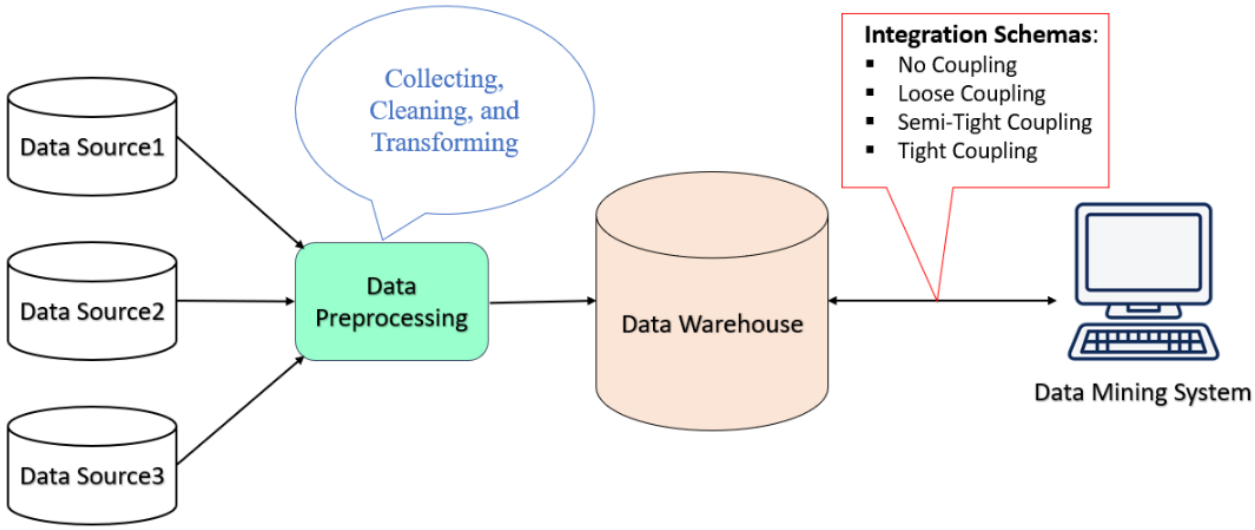
Visualization of discovered patterns
Rules, tables, reports,
charts, graphs,
decision trees, and cubes

Integration of Data mining system with a Data warehouse

The data mining system is **integrated** with a database or data warehouse system so that it can do its **tasks in an effective mode**. A data mining system operates in an environment that needs to **communicate** with other data systems like a Database or Datawarehouse system.

There are different possible integration (coupling) schemes as follows:

- No Coupling
- Loose Coupling
- Semi-Tight Coupling
- Tight Coupling



No Coupling

No coupling means that a Data Mining system will not utilize any function of a Data Base or Data Warehouse system.

It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file.

Drawbacks of No Coupling

First, without using a Database/Data Warehouse system, a Data Mining system may spend a substantial amount of time finding, collecting, cleaning, and transforming data.

Second, there are many tested, Scalable algorithms and data structures implemented in Database and Data Warehouse systems.

Loose Coupling

In this Loose coupling, the data mining system uses some facilities / services of a database or data warehouse system. The data is fetched from a data repository managed by these (DB/DW) systems.

Data mining approaches are used to process the data and then the processed data is saved either in a file or in a designated area in a database or data warehouse.

Loose coupling is better than no coupling because it can fetch any portion of data stored in Databases or Data Warehouses by using query processing, indexing, and other system facilities.

Drawbacks of Loose Coupling

It is difficult for loose coupling to achieve high scalability and good performance with large data sets.

Semi-Tight Coupling

Semi-tight coupling means that besides linking a Data Mining system to a Data Base/Data Warehouse system, efficient implementations of a few essential data mining primitives can be provided in the DB/DW system. These primitives can include (**sorting, indexing, aggregation, histogram analysis, multi way join, and pre computation of some essential statistical measures, such as sum, count, max, min, standard deviation).**

Advantage of Semi-Tight Coupling

This Coupling will enhance the performance of Data Mining systems

Tight Coupling

Tight coupling means that a Data Mining system is smoothly integrated into the Data Base/Data Warehouse system. The data mining subsystem is treated as one functional component of information system. Data mining **queries and functions** are optimized based on **mining query analysis, data structures, indexing schemes, and query processing methods of a DB or DW system.**

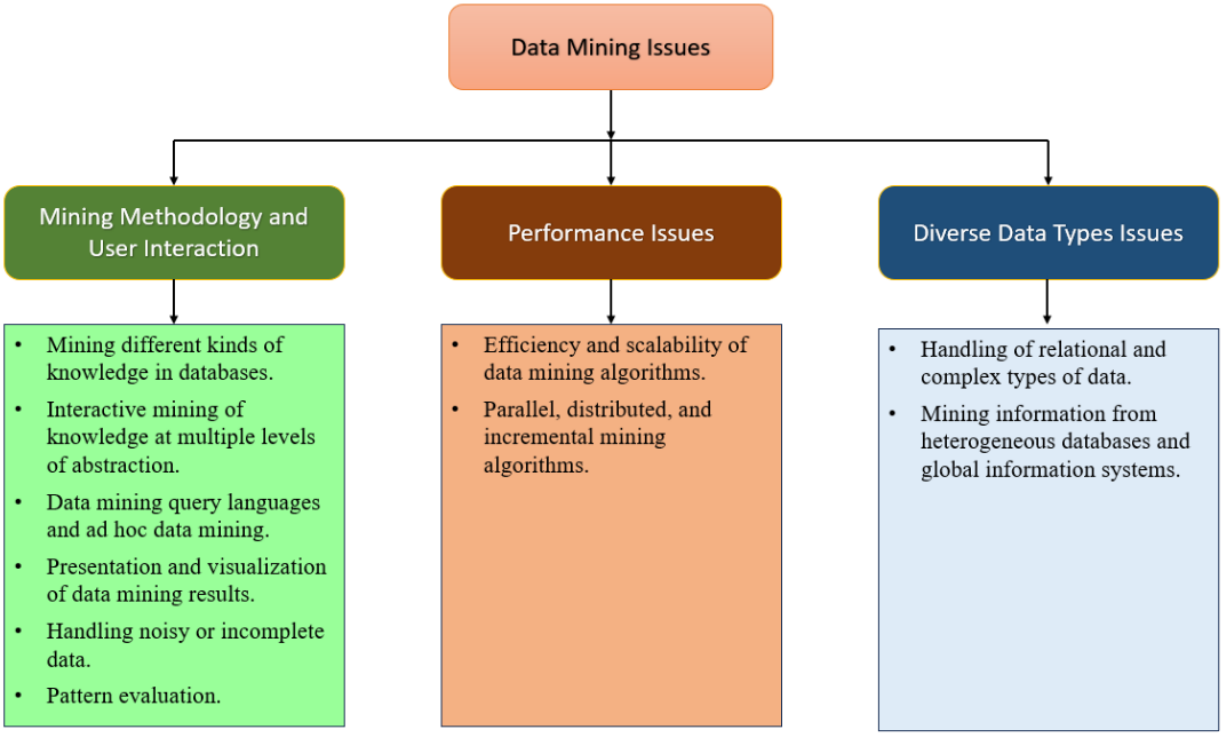
- Issues in DM

Major issues in Data Mining

Data mining, the process of extracting knowledge from data, has become increasingly important as the amount of data generated by individuals, organizations, and machines has grown exponentially. Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources.

The above factors may lead to some issues in data mining. These issues are mainly divided into three categories, which are given below:

1. **Mining Methodology and User Interaction**
2. **Performance Issues**
3. **Diverse Data Types Issues**



Mining Methodology and User Interaction

It refers to the following kinds of issues

- **Mining different kinds of knowledge in databases** – Different users may be interested in different kinds of knowledge. Therefore, it is necessary for data mining to cover a broad range of knowledge discovery task.

- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kinds of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi

structured or unstructured. Therefore, mining the knowledge from them adds challenges to data mining.

Knowledge Discovery from Data (KDD)

The need of data mining is to extract useful information from large datasets and use it to make predictions or better decision-making. Nowadays, data mining is used in almost all places where a large amount of data is stored and processed.

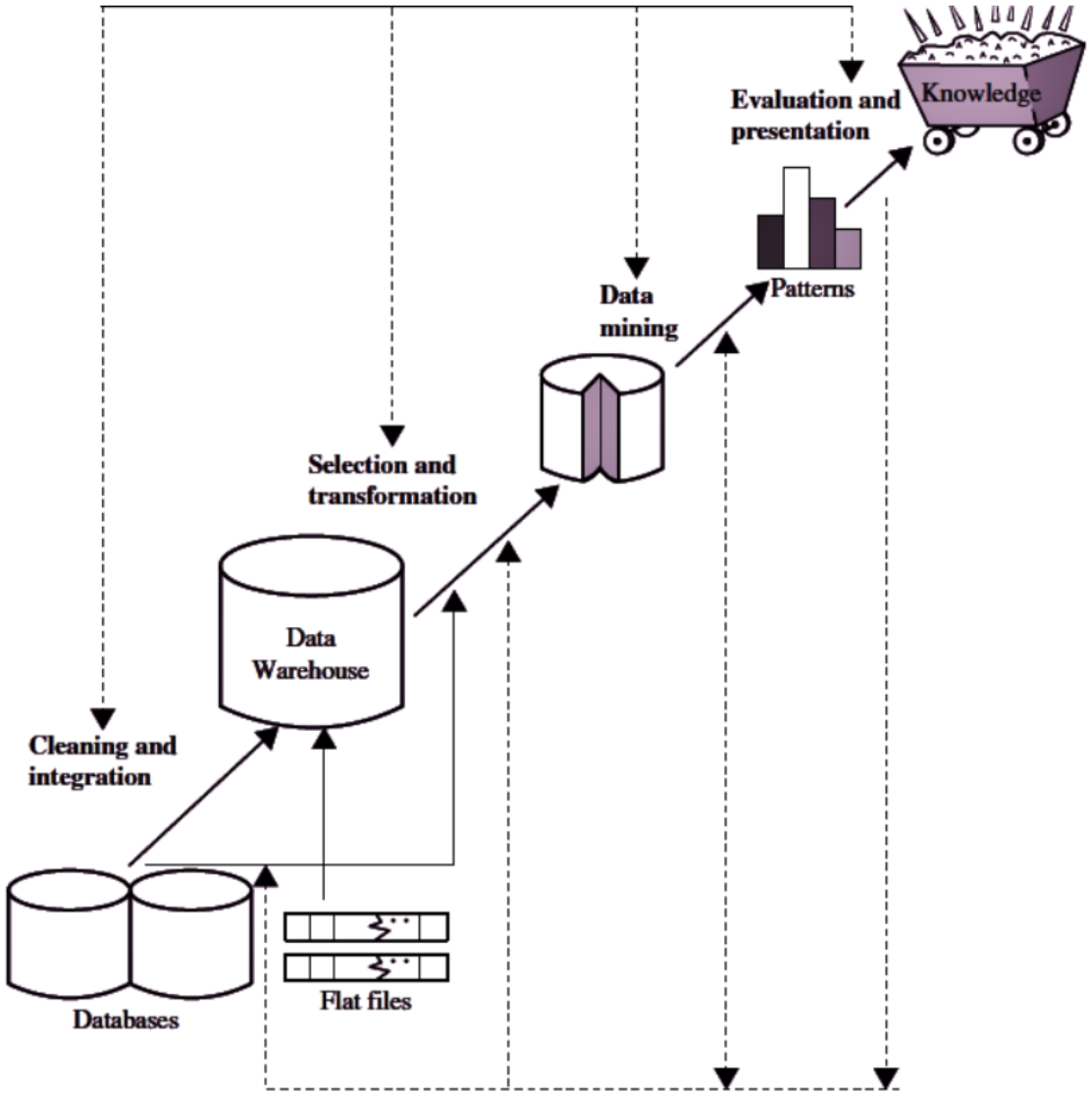
For examples: Banking sector, Market Basket Analysis, Network Intrusion Detection.

Data Mining also known as **Knowledge Discovery from Data or KDD**.

Knowledge Discovery from Data (KDD) Process

KDD is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets.

The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.



The following steps are included in KDD process:

1. **Data Cleaning**
2. **Data Integration**
3. **Data Selection**
4. **Data Transformation**
5. **Data Mining**
6. **Pattern Evaluation**
7. **Knowledge Representation**

1. Data Cleaning

Data cleaning is defined as removal of **noisy** and **irrelevant/ inconsistent** data from data collection.

- Cleaning in case of **Missing values**.
- Cleaning **noisy data**, where noise is a **random** or **variance error**.

In this step, the **noise** and **inconsistent** data is removed.

Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from the database collection.

- ✓ Cleaning in case of **Missing values**.
- ✓ Cleaning **noisy** data, where noise is a random or variance error.
- ✓ Cleaning with **Data Discrepancy Detection** and **Data Transformation Tools**.

2. Data Integration

Data integration is defined as heterogeneous **data from multiple data sources** combined in a common source (Data Warehouse).

i.e., In this step, multiple data sources may be combined as single data source.

A popular trend in the information industry is to perform **data cleaning** and **data integration** as a **data pre processing** step, where the resulting data are stored in a **data warehouse**.

2. Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).

Data integration using

- ✓ **Data Migration tools**
- ✓ **Data Synchronization tools**
- ✓ **ETL** (Extract-Load-Transformation) process.

3. Data Selection

Data selection is defined as the process where **data relevant to the analysis** is decided and retrieved from the data collection. This step in the KDD process is **identifying** and **selecting** the relevant data for analysis.

3. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

Data Selection Methods

- ✓ **Decision Trees**
- ✓ **Naive Bayes**
- ✓ **Regression**
- ✓ **Clustering**
- ✓ **Neural Network**

4. Data Transformation

Data Transformation is defined as the process of **transforming data into appropriate form** required by mining procedure. This step involves reducing the data dimensionality, aggregating the data, normalizing it, and discretizing it to prepare it for further analysis.

4. Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two-step process:

- ✓ **Data Mapping**: Assigning elements from source base to destination to capture transformations.
- ✓ **Code Generation**: Creation of the actual data transformation program.

5. Data Mining

This is the heart of the KDD process and involves applying **various data mining techniques** to the transformed data to discover **hidden patterns, trends, relationships**, and **insights**. A few of the most common data mining techniques include clustering, classification, association rule mining, and anomaly detection.

5. Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

- ✓ Transforms task relevant data into **patterns**.
- ✓ A specific Data mining model to be selected according to the functionality needed. Eg: **Classification, Association or Clustering**.

6. Pattern Evaluation

After the data mining, the next step is to evaluate the **discovered patterns** to determine their usefulness and relevance. This involves assessing the quality of the patterns, evaluating their significance, and selecting the most promising patterns for further analysis.

6. Pattern Evaluation: Pattern Evaluation is defined as identifying patterns representing knowledge based on given measures.

- ✓ Find **interestingness score** of each pattern.
- ✓ Uses **summarization** and **Visualization** to make data understandable by user.

7. Knowledge Representation

This step involves **representing the knowledge** extracted from the data in a way humans can easily understand and use. This can be done through visualizations, reports, or other forms of communication that provide meaningful insights into the data.

7. Knowledge Representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

Generate
- ✓ **Reports**
- ✓ **Tables**
- ✓ **Discriminant rules**
- ✓ **Classification rules**
- ✓ **Characterization rules**, etc.