# Introduction to Data Mining

In today's world, the amount of data is increasing exponentially whether it is biomedical data, security data or online shopping data, many industries preserve the data in order to analyse it, so that they can serve their customers more effectively through the information which they take out from large preserve data. This taking out or digging out information from huge data sets obtained from different sources and industries is known as Data Mining.

## Advantages of Data Mining

The following are the advantages of data mining:

- Easy to analysis huge data at one go
- Profitable decision-making process
- Prediction of trends
- Knowledge-based information
- Profitable production
- Discovery of hidden patterns
- Cost effective, time efficient, and effective prediction

## Applications of Data Mining

### 1. Communications

Data mining techniques are used in the communication sector to predict customer behavior to offer highly targeted and relevant campaigns.

### 2. Insurance

Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing customers.

### 3. Education

Data mining benefits educators to access student data, predict achievement levels and find students or groups of students who need extra attention. For example, students who are weak in a science subject.

### 4. Manufacturing

By using the help of Data Mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime.

**5. Banking**

Data mining helps the finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc.

**6. Retail**

Data Mining techniques help retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to come up with the offer which encourages customers to increase their spending.

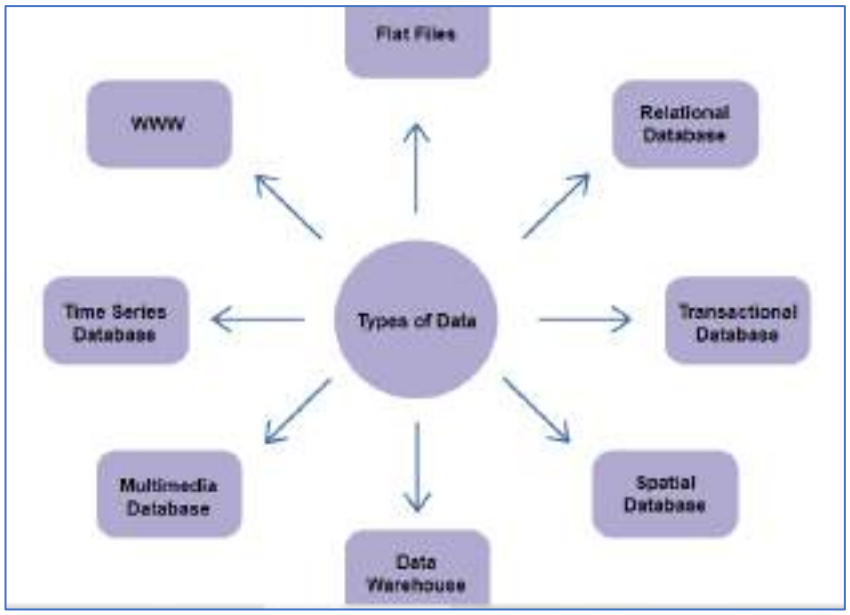
**7. Service Providers**

Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze billing details, customer service interactions, complaints made to the company to assign each customer a probability score and offer incentives.

**8. E-Commerce**

E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, which uses Data mining techniques to get more customers into their eCommerce store.

# Types of Data




**Flat Files:** Flat files are the most common data source for data mining algorithms, especially at the research level.

Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied.

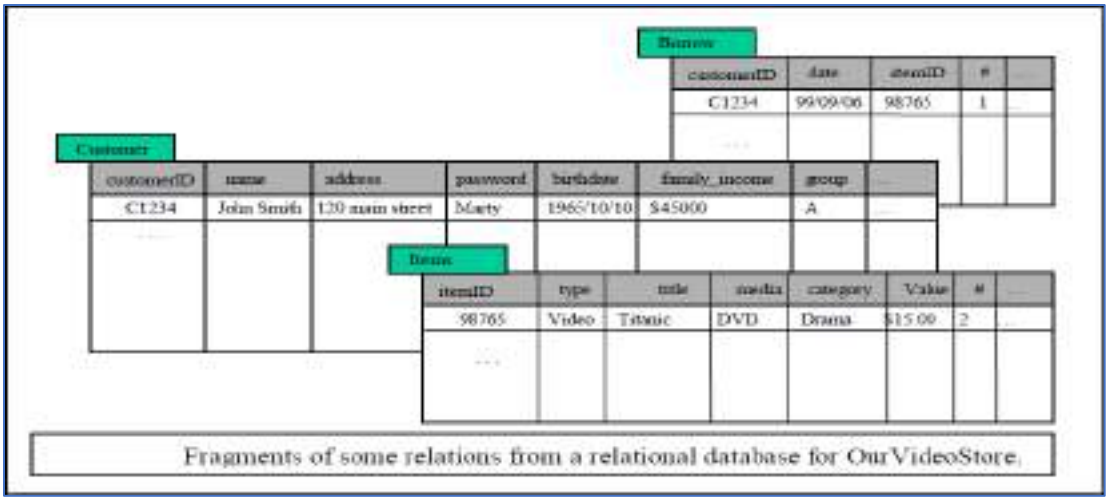
The data in these files can be transactions, time-series data, scientific measurements, etc.

**Relational Databases:** A relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships.

Tables have columns and rows, where columns represent attributes and rows represent tuples.

A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

In following figure, it presents some relations Customer, Items, and Borrow representing business activity in a video store. These relations are just a subset of what could be a database for the video store and is given as an example.




The most used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count. For instance, an SQL query to select the videos grouped by category would be:

**SELECT count (*) FROM Items WHERE type=video GROUP BY category.**

Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files, since they can take advantage of the structure inherent to relational databases.

**Transactional databases:** In general, a transactional database consists of a flat file where each record represents a transaction.

A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction (such as items

purchased in a store) as shown below:

SALES

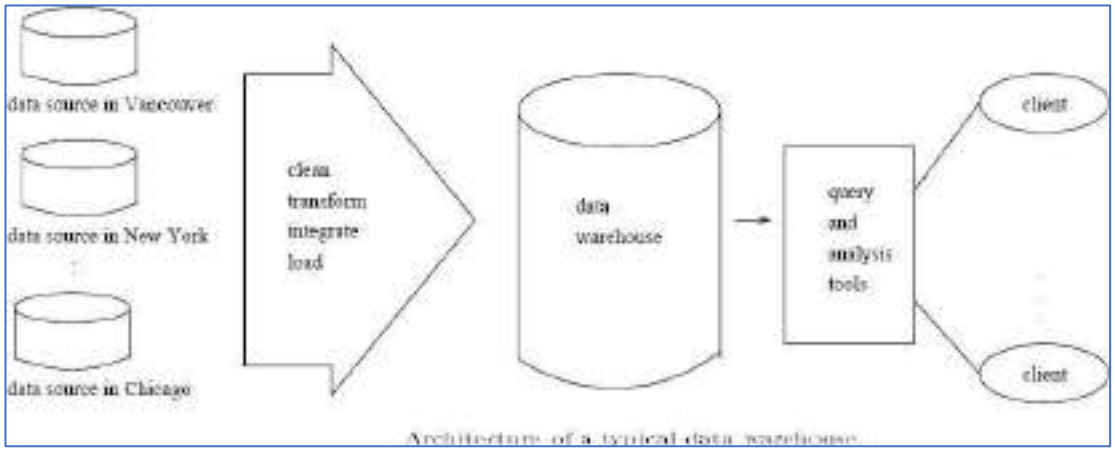
| Trans-ID | List of item ID's |
|---|---|
| T100<br>…….. | I1, I3, I8<br>…… |

**Spatial Database:** It contains spatial-related data, which may be represented in the form of raster or vector data.

Raster data consists of n-dimensional bit maps or pixel maps, and vector data are represented by lines, points, polygons or other kinds of processed primitives.

Examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.

**Data Warehouse:** A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing.



Architecture of a typical data warehouse

To facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity. The data are stored to provide information from a historical perspective and are typically summarized.

A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data.

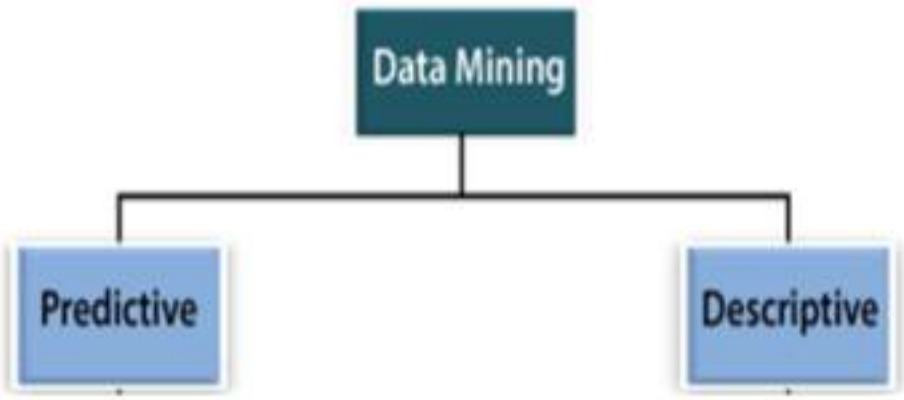
**Multimedia Database:** It stores images, audio, and video data, and is used in applications such as picture content-based retrieval, voice-mail systems, video-on- demand systems, the World Wide Web, and speech-based user interfaces.

**Time-Series Databases:** Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

**World-Wide Web:** WWW provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access. Some examples of distributed information services associated with the World-Wide Web include America Online Yahoo!, AltaVista, and Prodigy.

## Data Mining Functionalities

Data mining functions are used to define the trends or correlations contained in data mining activities. In comparison, data mining activities can be divided into two categories:




**Predictive Data Mining:** It helps developers to provide unlabelled definitions of attributes. With previously available or historical data, data mining can be used to make predictions about critical business metrics based on data's linearity.

For example, predicting the volume of business next quarter based on performance in the previous quarters over several years or judging from the findings of a patient's medical examinations that is he suffering from any disease.

**Descriptive Data Mining:** It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set.

For example, count, average etc.

Data mining functionalities are used to represent the type of patterns that must be discovered in data mining tasks.

Data mining is extensively used in many areas or sectors. It is used to predict and characterize data.

There are several data mining functionalities that the organized and scientific methods offer, such as:


**1. Class/Concept Descriptions**

A class or concept implies there is a data set or set of features that define the class or a concept.

A class can be a category of items on a shop floor, and a concept could be the abstract idea on which data may be categorized like products to be put on clearance sale and non-sale products.

There are two concepts here, one that helps with grouping and the other that helps in differentiating.

**Data Characterization:** This refers to the summary of general characteristics or features of the class, resulting in specific rules that define a target class.

- A data analysis technique called Attribute-oriented Induction is employed on the data set for achieving characterization.

**Data Discrimination:** Discrimination is used to separate distinct data sets based on the disparity in attribute values.

- It compares features of a class with features of one or more contrasting classes.

Eg: bar charts, curves and pie charts.

**2. Mining Frequent Patterns**

One of the functions of data mining is finding data patterns.

Frequent patterns are discovered to be most common in data. Various types of frequency can be found in the dataset.

- **Frequent item set:** This term refers to a group of items that are commonly found together, such as milk and sugar.
- **Frequent substructure:** It refers to the various types of data structures that can be combined with an item set or subsequence's, such as trees and graphs.
- **Frequent Subsequence:** A regular pattern series, such as buying a phone followed by a cover.

**3. Association Analysis**

It analyses the set of items that generally occur together in a transactional dataset. It is also known as Market Basket Analysis for its wide use in retail sales.

Two parameters are used for determining the association rules:

Support: It provides which identifies the common item set in the database.

Confidence: is the conditional probability that an item occurs when another item occurs in a transaction.

## 4. Classification

Classification is a data mining technique that categorizes items in a collection based on some predefined properties.

It uses methods like if-then, decision trees or neural networks to predict a class or essentially classify a collection of items.

A training set containing items whose properties are known is used to train the system to predict the category of items from an unknown collection of items.

## 5. Prediction

It defines predict some unavailable data values or spending 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 or decrease trends in time-related information.

There are primarily two types of predictions in data mining: numeric and class predictions.

- ***Numeric predictions*** are made by creating a linear regression model that is based on historical data. Prediction of numeric values helps businesses ramp up for a future event that might impact the business positively or negatively.
- ***Class predictions*** are used to fill in missing class information for products using a training data set where the class for products is known.

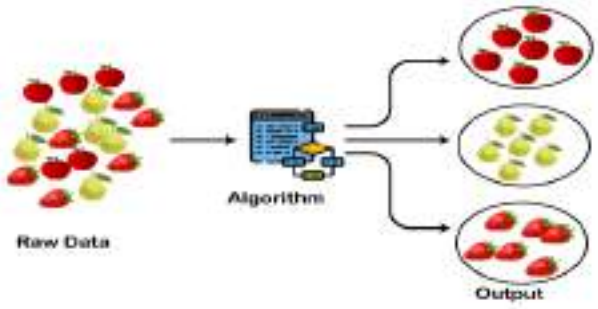
## 6. Cluster Analysis

In image processing, pattern recognition and bioinformatics, clustering is a popular data mining functionality.

It is like classification, but the classes are not predefined. Data attributes represent the classes.

Similar data are grouped together, with the difference being that a class label is not known.

Clustering algorithms group data based on similar features and dissimilarities.

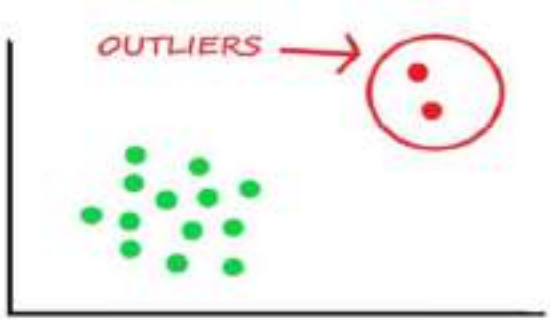


## 7. Outlier Analysis

Outlier analysis is important to understand the quality of data. If there are too many outliers, you cannot trust the data or draw patterns.

An outlier analysis determines if there is something out of turn in the data and whether it indicates a situation that a business needs to consider and take measures to mitigate.

An outlier analysis of the data that cannot be grouped into any classes by the algorithms is pulled up.




## 8. Evolution and Deviation Analysis

Evolution Analysis pertains to the study of data sets that change over time.

Evolution analysis models are designed to capture evolutionary trends in data helping to characterize, classify, cluster or discriminate time-related data.

## 9. Correlation Analysis

Correlation is a mathematical technique for determining whether and how strongly two attributes is related to one another.

It refers to the various types of data structures, such as trees and graphs, that can be combined with an item set or subsequence.

It determines how well two numerically measured continuous variables are linked. Researchers can use this type of analysis to see if there are any possible correlations between variables in their study.

# Interestingness Patterns

A pattern is interesting if it is

(1) Easily understood by humans

(2) Valid on new or test data with some degree of certainty

(3) Potentially useful

(4) Novel

A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge.

**Objective measures of pattern interestingness** are based on the structure of discovered patterns and the statistics underlying them. An objective measure for association rules of the form X Y is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies. This is taken to be the probability P.(X U Y) where X UY indicates that a transaction contains both X and Y, that is, the union of item sets X and Y.

Another objective measure for association rules is confidence, which assesses the degree of certainty of the detected association. This is taken to be the conditional probability P.(Y | X), that is, the probability

that a transaction containing X also contains Y.

$$support(X \Rightarrow Y) = P(X \cup Y),$$
$$confidence(X \Rightarrow Y) = P(Y|X).$$

In general, each interestingness measure is associated with a threshold, which may be controlled by the user. For example, rules that do not satisfy a confidence threshold of, say, 50% can be considered uninteresting. Rules below the threshold likely reflect noise, exceptions, or minority cases and are probably of less value.

Other objective interestingness measures include accuracy and coverage for classification (IF-THEN) rules. In general terms, accuracy tells us the percentage of data that are correctly classified by a rule. Coverage is similar to support, in that it tells us the percentage of data to which a rule applies.

.

**Subjective interestingness measures** are based on user beliefs in the data. These measures find patterns interesting if the patterns are unexpected (contradicting a user's belief) or offer strategic information on which the user can act. In the latter case, such

patterns are referred to as actionable. For example, patterns like “a large earthquake often follows a cluster of small quakes” may be highly actionable if users can act on the information to save lives. Patterns that are expected can be interesting if they confirm a hypothesis that the user wishes to validate or they resemble a user’s hunch.

It is often unrealistic and inefficient for data mining systems to generate all possible patterns. Instead, user provided constraints and interestingness measures should be used to focus the search. For some mining tasks, such as association, this is often sufficient to ensure the completeness of the algorithm. Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining.

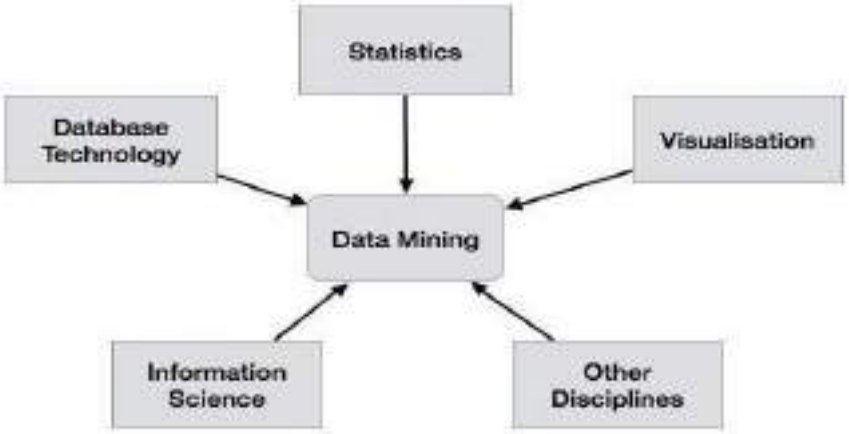
It is highly desirable for data mining systems to generate only interesting patterns. This would be efficient for users and data mining systems because neither would have to search through the patterns generated to identify the truly interesting ones. Progress has been made in this direction; however, such optimization remains a challenging issue in data mining. Measures of pattern interestingness are essential for the efficient discovery of patterns by target users. Such measures can be used after the data mining step to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones.

More important, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy pre specified interestingness constraints.

# Classification of Data Mining Systems

A data mining system can be classified according to the following criteria −

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines



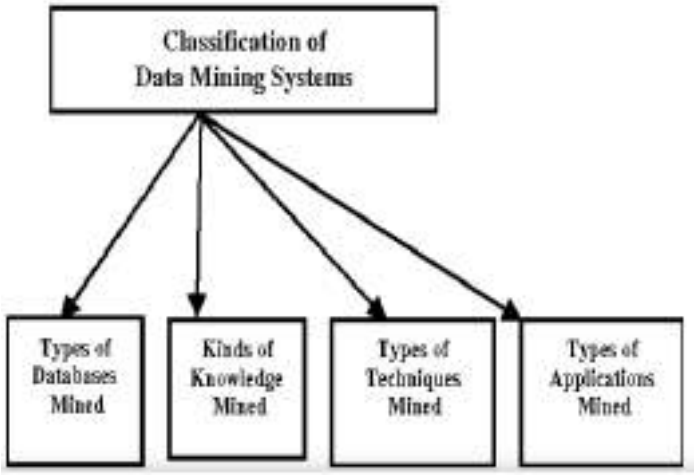

Apart from these, a data mining system can also be classified based on the kind of

(a) Databases mined

(b) Knowledge mined

(c) Techniques utilized

(d) Applications adapted



**Classification Based on the Databases Mined**

We can classify a data mining system according to the 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 a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.

**Classification Based on the kind of Knowledge Mined**

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

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Outlier Analysis
- Evolution Analysis

**Classification Based on the Techniques Utilized**

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

**Classification Based on the Applications Adapted**

We can classify a data mining system according to the applications 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 discovery to direct the mining process or examine the findings from different angles or depths.

The data mining primitives specify the following,

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

**1. The set of task-relevant data to be mined**

This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (the relevant attributes or dimensions).

In a relational database, the set of task-relevant data can be collected via a relational query involving operations like selection, projection, join, and aggregation.

The data collection process results in a new data relational called the ***initial data relation***. The initial data relation can be ordered or grouped according to the conditions specified in the query.

This data retrieval can be thought of as a subtask of the data mining task. This initial relation may or may not correspond to physical relation in the database. Since virtual relations are called Views in the field of databases, the set of task-relevant data for data mining is called a minable view.

**2. The kind of knowledge to be mined**

This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

**3. The background knowledge to be used in the discovery process**

This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns found.

Concept hierarchies are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction. Concept hierarchy defines a sequence of mappings from low-level concepts to higher-level, more general concepts.

**Rolling Up - Generalization of data:** Allow to view data at more meaningful and explicit abstractions and makes it easier to understand. It compresses the data, and it would require fewer input/output operations.

**Drilling Down - Specialization of data:** Concept values replaced by lower-level concepts. Based on different user viewpoints, there may be more than one concept hierarchy for a given attribute or dimension.

An example of a concept hierarchy for the attribute (or dimension) age is shown below. User beliefs regarding relationships in the data are another form of background knowledge.

**4. The interestingness measures and thresholds for pattern evaluation**

Different kinds of knowledge may have different interesting measures. They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. For example, interesting measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

**Simplicity:** A factor contributing to the interestingness of a pattern is the pattern's overall simplicity for human comprehension. For example, the more complex the structure of a

rule is, the more difficult it is to interpret, and hence, the less interesting it is likely to be. Objective measures of pattern simplicity can be viewed as functions of the pattern structure, defined in terms of the pattern size in bits or the number of attributes or operators appearing in the pattern.

**Certainty (Confidence):** Each discovered pattern should have a measure of certainty associated with it that assesses the validity or "trustworthiness" of the pattern. A certainty measure for association rules of the form "A =>B" where A and B are sets of items is confidence. Confidence is a certainty measure. Given a set of task-relevant data tuples, the confidence of "A => B" is defined as
Confidence (A=>B) = # tuples containing both A and B /# tuples containing A

**Utility (Support):** The potential usefulness of a pattern is a factor defining its interestingness. It can be estimated by a utility function, such as support. The support of an association pattern refers to the percentage of task-relevant data tuples (or transactions) for which the pattern is true.

Utility (support) usefulness of pattern
Support (A=>B) = # tuples containing both A and B / total #of tuples

**Novelty:** Novel patterns are those that contribute new information or increased performance to the given pattern set.

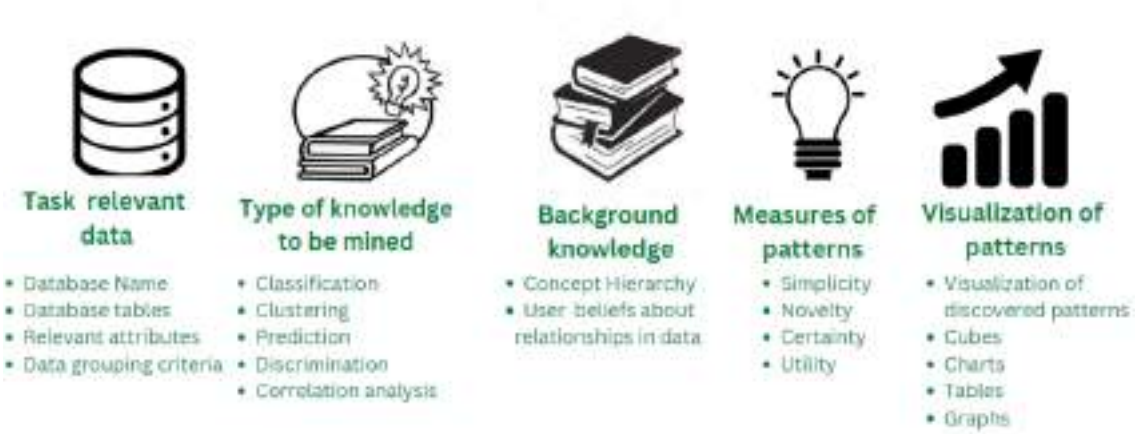
For example: Data exception, another strategy for detecting novelty is to remove redundant patterns.

**5. The expected representation for visualizing the discovered patterns**

This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, cross tabs, charts, graphs, decision trees, cubes, or other visual representations.

Users must be able to specify the forms of presentation to be used for displaying the discovered patterns. Some representation forms may be better suited than others for particular kinds of knowledge.

For example, generalized relations and their corresponding cross tabs or pie/bar charts are good for presenting characteristic descriptions, whereas decision trees are common for classification.



## Integration of Data Mining System with a Data warehouse

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme. In this scheme, the main focus is on data mining design and on developing efficient and effective algorithms for mining the available data sets.

The list of Integration Schemes is as follows −

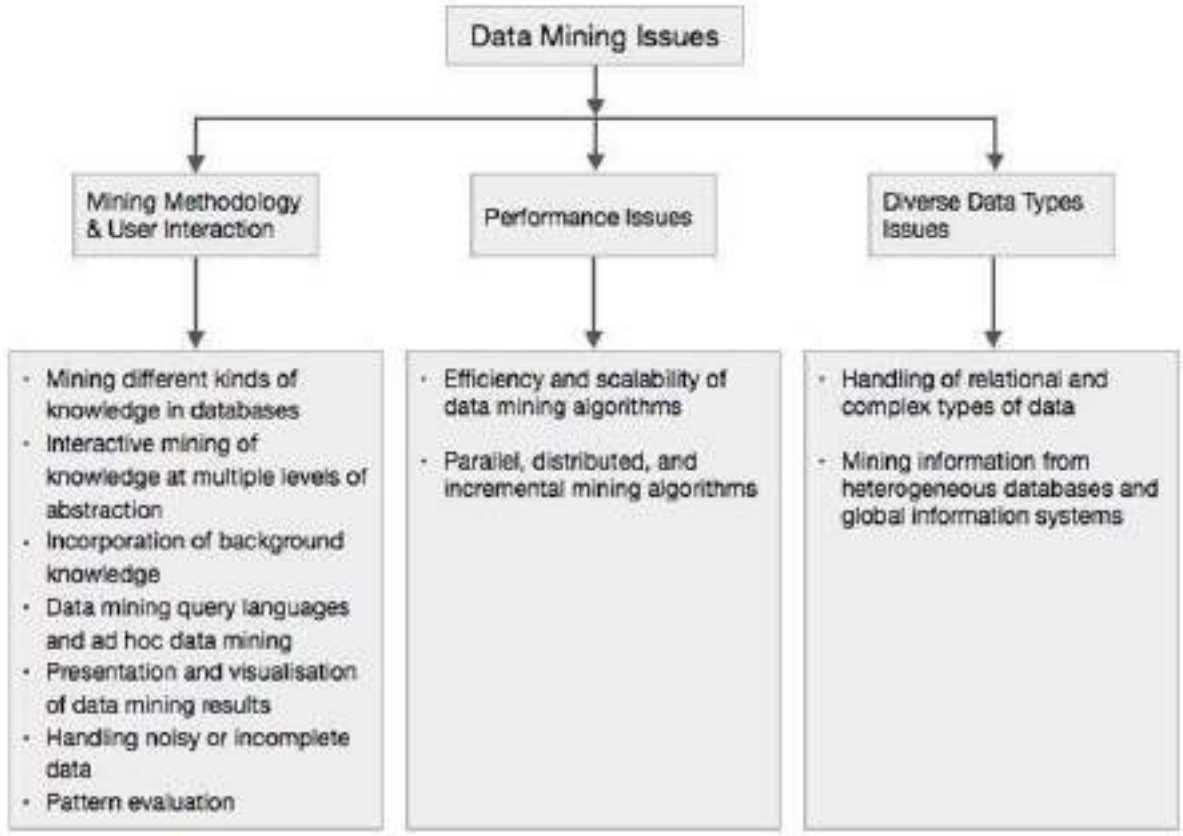
- **No Coupling** − in this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in another file.
- **Loose Coupling** − in this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory managed by these systems and performs data mining on that data. It then stores the mining result either in a file or in a designated place in a database or in a data warehouse.
- **Semi−tight Coupling** − In this scheme, the data mining system is linked with a database or a data warehouse system and in addition to that, efficient implementations of a few data mining primitives can be provided in the database.
- **Tight coupling** − in this coupling scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

# Major Issues in Data Mining

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. These factors also create some issues.

Major issues are:
Mining Methodology and User Interaction
Performance Issues
Diverse Data Types Issues

The following diagram describes the major issues




**Mining Methodology and User Interaction Issues 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.

**Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.

**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.

P**arallel, 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. Then the results from the partitions is merged. 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 kind 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

# Data Preprocessing

Data pre-processing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms. Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

- **Accuracy**: To check whether the data entered is correct or not.
- **Completeness**: To check whether the data is available or not recorded.
- **Consistency:** To check whether the same data is kept in all the places that do or do not match.
- **Timeliness**: The data should be updated correctly.
- **Believability**: The data should be trustable.
- **Interpretability**: The understandability of the data.

There are 4 major tasks in data preprocessing – Data cleaning, Data integration, Data reduction, and Data transformation.

**Data Cleaning**

Data cleaning is the process of removing incorrect data, incomplete data, and inaccurate data from the datasets, and it also replaces the missing values. Here are some techniques for data cleaning:

**Handling Missing Values**

- Standard values like “Not Available” or “NA” can be used to replace the missing values.
- Missing values can also be filled manually, but it is not recommended when that dataset is big.

- The attribute’s mean value can be used to replace the missing value when the data is normally distributed
  wherein in the case of non-normal distribution median value of the attribute can be used.
- While using regression or decision tree algorithms, the missing value can be replaced by the most probable value.

**Handling Noisy Data**

Noisy generally means random error or containing unnecessary data points. Handling noisy data is one of the most important steps as it leads to the optimization of the model we are using Here are some of the methods to handle noisy data.

- **Binning:** This method is to smooth or handle noisy data. First, the data is sorted then, and then the sorted values are separated and stored in the form of bins. There are three methods for smoothing data in the bin.

  **Smoothing by bin mean method**: In this method, the values in the bin are replaced by the mean value of the bin;

  **Smoothing by bin median**: In this method, the values in the bin are replaced by the median value;

  **Smoothing by bin boundary**: In this method, the using minimum and maximum values of the bin values are taken, and the closest boundary value replaces the values.

- **Regression:** This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.
- **Clustering:** This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

## Data Integration

The process of combining multiple sources into a single dataset. The Data integration process is one of the main components of data management. There are some problems to be considered during data integration.

This can be done in two types:

a. Tight coupling: Data is combined together into a physical location. Once the data is combined we cannot access it separately.
b. Loose coupling: The data is not integrated directly, it will be integrated using an interface. The interface is created and then only it will be combined using it. Here the data can be accessed at any point of time individually. Data remains in actual database only.

**Schema integration**: Integrates metadata(a set of data that describes other data) from different sources.

**Entity identification problem*:*** Identifying entities from multiple databases. For example, the system or the user should know the student *id* of one database and student name of another database belonging to the same entity.

**Detecting and resolving data value concepts***:* The data taken from different databases while merging may differ. The attribute values from one database may differ from another database. For example, the date format may differ, like “MM/DD/YYYY” or “DD/MM/YYYY”.

## Data Reduction

This process helps in the reduction of the volume of the data, which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space. Some of the data reduction techniques are dimensionality reduction, numerosity reduction, and data compression.

- **Dimensionality reduction:** This process is necessary for real-world applications as the data size is big. In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced. Combining and merging the attributes of the data without losing its original characteristics. This also helps in the reduction of storage space, and computation time is reduced. When the data is highly dimensional, a problem called the “Curse of Dimensionality” occurs.
- **Numerosity Reduction**: In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.
- **Data compression**: The compressed form of data is called data compression. This compression can be lossless or lossy. When there is no loss of information during compression, it is called lossless compression. Whereas lossy compression reduces information, but it removes only the unnecessary information.

## Data Transformation

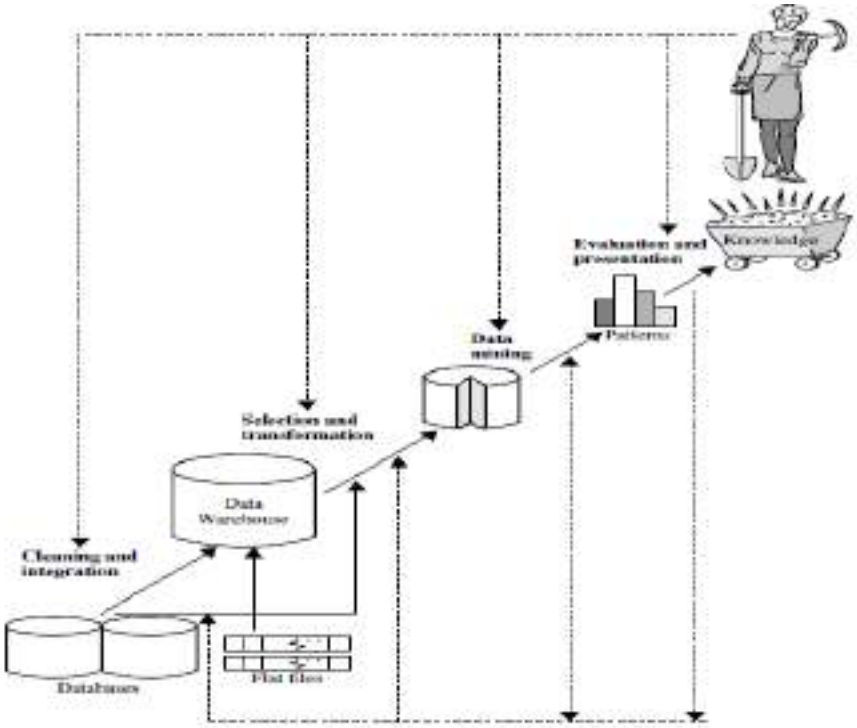
The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods for data transformation.

- **Smoothing:** With the help of algorithms, we can remove noise from the dataset, which helps in knowing the important features of the dataset. By smoothing, we can find even a simple change that helps in prediction.
- **Aggregation:** In this method, the data is stored and presented in the form of a summary. The data set, which is from multiple sources, is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data. When the quality and the quantity of the data are good, the results are more relevant.
- **Discretization:** The continuous data here is split into intervals. Discretization reduces the data size. For example, rather than specifying the class time, we can set an interval like (3 pm-5 pm, or 6 pm-8 pm).
- **Normalization:** It is the method of scaling the data so that it can be represented in a smaller range. Example ranging from -1.0 to 1.0.

**Knowledge Discovery in Database (KDD)**

Data mining is also called **Knowledge Discovery in Database (KDD)**. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Many people treat data mining as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is shown in figure as 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)

Steps 1 to 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically