The key components of data mining

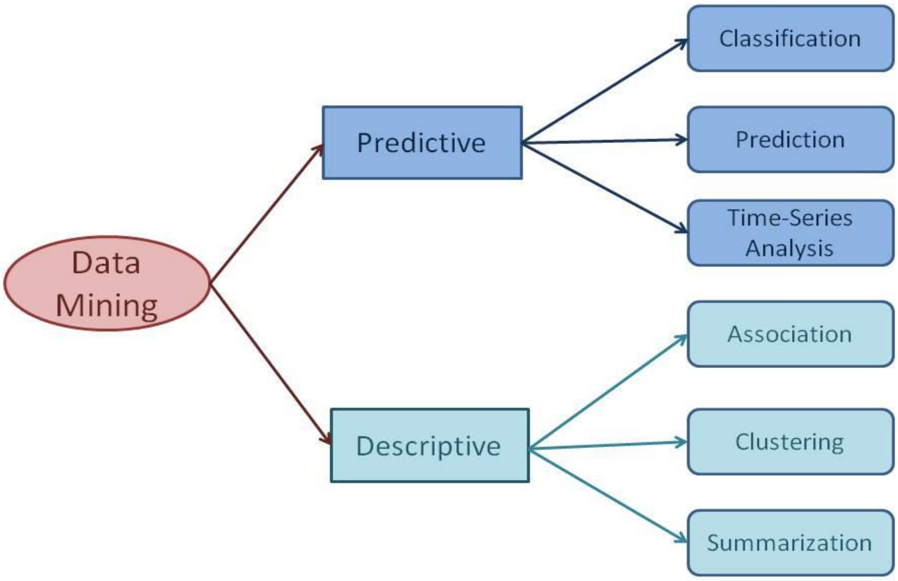
The process of data mining includes several distinct components that address different needs:

- **Preprocessing.** Before you can apply data mining algorithms, you need to build a target data set. One common source for data is a data mart or warehouse. You need to perform preprocessing to be able to analyze the data sets.
- **Data cleansing and preparation.** The target data set must be cleaned and otherwise prepared, to remove “noise,” address missing values, filter outlying data points (for anomaly detection) to remove errors or do further exploration, create segmentation rules, and perform other functions related to data preparation.
- **Association rule learning** (also known as **market basket analysis**). These tools search for relationships among variables in a data set, such as determining which products in a store are often purchased together.
- **Clustering.** This feature of data mining is used to discover groups and structures in data sets that are in some way similar to each other, without using known structures in the data.
- **Classification.** Tools that perform classification generalize known structures to apply to new data points, such as when an email application tries to classify a message as legitimate mail or spam.
- **Regression.** This data mining technique tis used to predict a range of numeric values, such as sales, housing values, temperatures, or prices when given a particular data set.
- **Summarization.** This technique provides a compact representation of a data set, including visualization and report generation.

Data Mining Tasks

The data mining tasks can be classified generally into two types based on what a specific task tries to achieve. Those two categories are **descriptive tasks** and **predictive tasks**. The descriptive data mining tasks characterize the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave.

There are a number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks. A data mining system can execute one or more of the above specified tasks as part of data mining.



a) Classification

Classification derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible.

Classification can be used in direct marketing that is to reduce marketing costs by targeting a set of customers who are likely to buy a new product. Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past. Hence, {purchase, don’t purchase} decision forms the class attribute in this case. Once the class attribute is assigned, demographic and lifestyle information of customers who purchased similar products can be collected and promotion mails can be sent to them directly.

b) Prediction

Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest. For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

c) Time - Series Analysis

Time series is a sequence of events where the next event is determined by one or more of the preceding events. Time series reflects the process being measured and there are certain components that affect the behavior of a process. Time series analysis includes methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time- series analysis.

d) Association

Association discovers the association or connection among a set of items. Association identifies the relationships between objects. Association analysis is used for commodity management, advertising, catalog design, direct marketing etc. A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion of same kind of products. If a retailer finds that beer and nappy are bought together mostly, he can put nappies on sale to promote the sale of beer.

e) Clustering

Clustering is used to identify data objects that are similar to one another. The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on. For example, an insurance company can cluster its customers based on age, residence, income etc. This group information will be helpful to understand the customers better and hence provide better customized services.

f) Summarization

Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data. For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis. Data can be summarized in different abstraction levels and from different angles.

LECTURE 2: Data Mining in Detail

Data Mining Techniques

1. **Classification Analysis**

   This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster. So, in classification analysis you would apply algorithms to decide how new data should be classified. A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.

2. **Association Rule Learning**

It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset. Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout. In IT, programmers use association rules to build programs capable of machine learning.

3. **Anomaly or Outlier Detection**

This refers to the observation for data items in a dataset that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a dataset or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention. This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting eco-system disturbances. Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy.

4. **Clustering Analysis**

The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise. A result of this analysis can be used to create customer profiling.

5. **Regression Analysis**

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting.

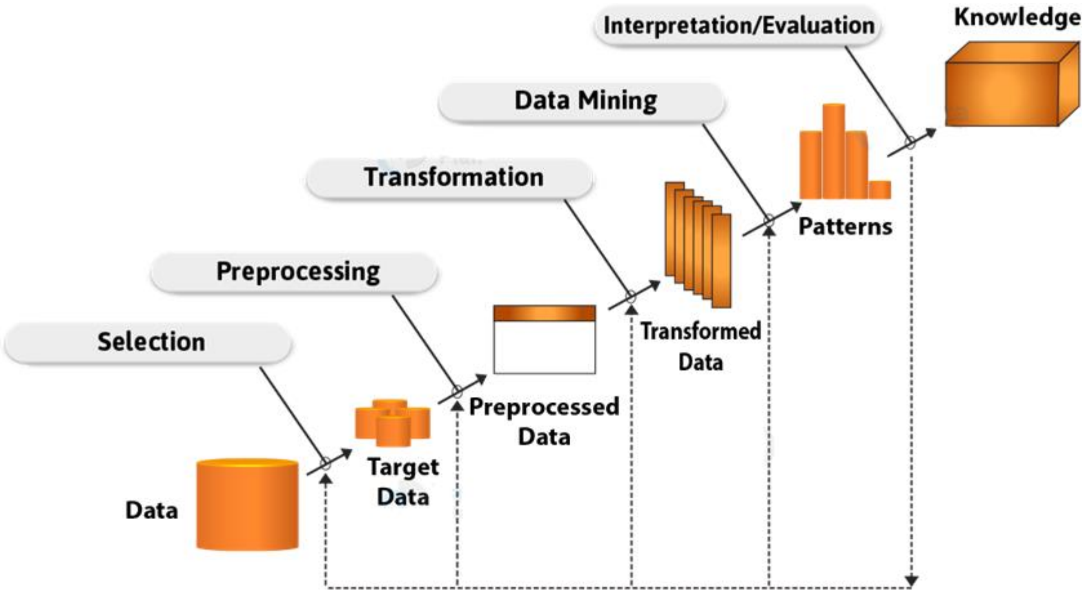
Attributes of Mined Data

1. New (Novel)
2. Correct
3. Potentially useful

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

Data mining is the core part of the knowledge discovery process. In this, process may consist of the following steps Data selection, Data cleaning, Data transformation, pattern searching (data mining), finding presentation, finding interpretation and finding evaluation. The data mining and KDD often used interchangeably because Data mining is the key part of KDD process.



KDD Process Steps

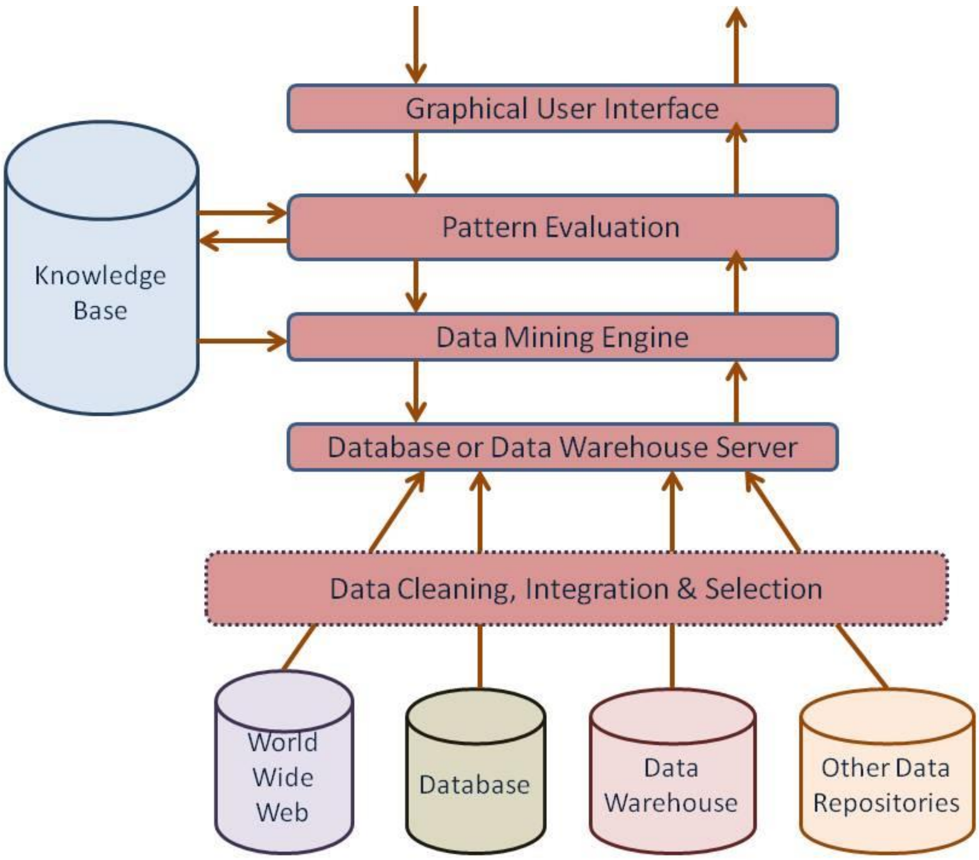
1. **Data Selection:** In this step, data relevant to the analysis task are retrieved from the database. Data can originate from multiple sources or from data warehouse. selected data must be appropriate for mining tasks
2. **Data Pre-processing**: In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
3. **Data Mining**: Intelligent operations such as Clustering, Classification, and regression are applied in order to extract data patterns.
4. **Pattern Evaluation**: Obtained results are evaluated for their accuracy. In this step, data patterns are evaluated.
5. **Knowledge presentation**: Identified patterns are presented in visual forms such as aesthetic graphs

DATA MINING ARCHITECTURE

Data mining is a very important process where potentially useful and previously unknown information is extracted from large volumes of data. There are a number of components involved in the data mining process. These components constitute the architecture of a data mining system.

Data Mining Architecture

The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.



a) Data Sources

Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful. Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

Different Processes

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable. So, first data needs to be cleaned and integrated. Again, more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server. These processes are not as simple as we think. A number of techniques may be performed on the data as part of cleaning, integration and selection.

b) Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

c) Data Mining Engine

The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

d) Pattern Evaluation Modules

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

e) Graphical User Interface

The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

f) Knowledge Base

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

KEY DIFFERENCES BETWEEN DATA WAREHOUSING VS DATA MINING

Some of the major differences between Data Warehousing and Data Mining are mentioned below:

- Data Warehousing is the process of extracting and storing data to allow easier reporting. Whereas Data mining is the use of pattern recognition logic to identify trends within a sample data set, a typical use of data mining is to identify fraud, and to flag unusual patterns in behavior. For Example, Credit Card Company provide you an alert when you are transacting from some other geographical location which you have not used previously. This fraud detection is possible because of data mining.

- The main difference between data warehousing and data mining is that data warehousing is the process of compiling and organizing data into one common database, whereas data mining is the process of extracting meaningful data from that database. Data mining can only be done once data warehousing is complete**.**
- Data warehouse is the repository to store data. On the other hand, data mining is a broad set of activities used to uncover patterns, and give meaning to this data.
- Data warehousing is merely extracting data from different sources, cleaning the data and storing it in the warehouse. Whereas data mining aims to examine or explore the data using queries.

For example A data warehouse of a company store all the relevant information of projects and employees. Using Data mining, one can use this data to generate different reports like profits generated etc.

- Data warehouse is an architecture whereas, data mining is a process that is an outcome of various activities for discovering the new patterns.
- A data warehouse is a technique of organizing data so that there should be corporate credibility and integrity, but, Data mining is helpful in extracting meaningful patterns those are not found, necessarily by only processing data or querying data in the data warehouse.
- Data warehouse contains integrated and processed data to perform data mining at the time of planning and decision making, but data discovered by data mining results in finding patterns that are useful for future predictions.
- Data warehouse supports basic statistical analysis. The information retrieved from data mining is helpful in tasks like Market segmentation, customer profiling, credit risk analysis, fraud detection etc.
- Data warehousing is the process of pooling all relevant data together, whereas Data mining is the process of analyzing unknown patterns of data.
- Data warehouses usually store many months or years of data. This is to support historical analysis. Data mining is the use of pattern recognition logic to identify trend within a sample data set.

Data Warehousing vs. Data Mining Comparison Table

DATA WAREHOUSING	**DATA MINING**
Data warehousing is the process which is used to integrate data from multiple sources and then combine it into a single database	It is the process which is used to extract useful patterns and relationships from a huge amount of data
It provides the organization a mechanism to store huge amount of data	Data mining techniques are applied on data warehouse in order to discover useful patterns

This process must take place before data mining process because it compiles and organizes data into a common database	This process always takes place after data warehousing process because it requires compiled data to extract useful patterns
This process is solely carried out by engineers	This process is carried out by business users with the help of engineers

Why use Data mining?

Some most important reasons for using Data mining are:

- Establish relevance and relationships amongst data. Use this information to generate profitable insights
- Business can make informed decisions quickly
- Helps to find out unusual shopping patterns in grocery stores.
- Optimize website business by providing customize offers to each visitor.
- Helps to measure customer's response rates in business marketing.
- Creating and maintaining new customer groups for marketing purposes.
- Predict customer defections, like which customers are more likely to switch to another supplier in the nearest future.
- Differentiate between profitable and unprofitable customers.
- Identify all kind of suspicious behavior, as part of a fraud detection process.