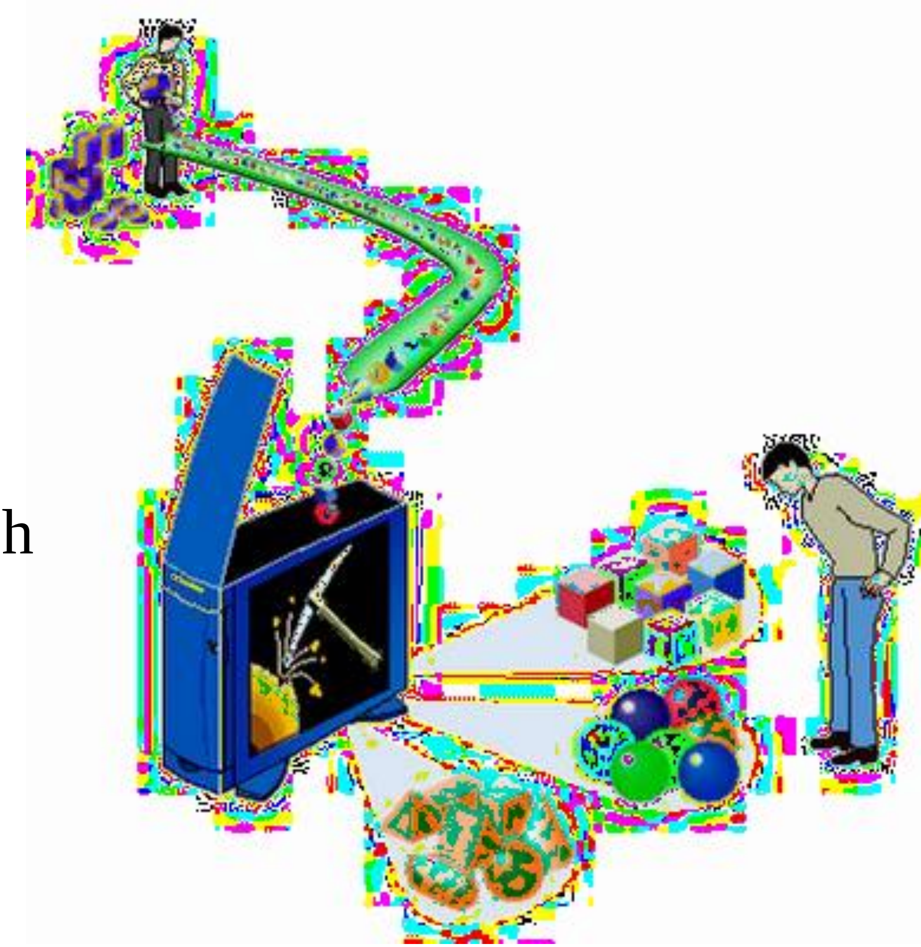
Data Mining

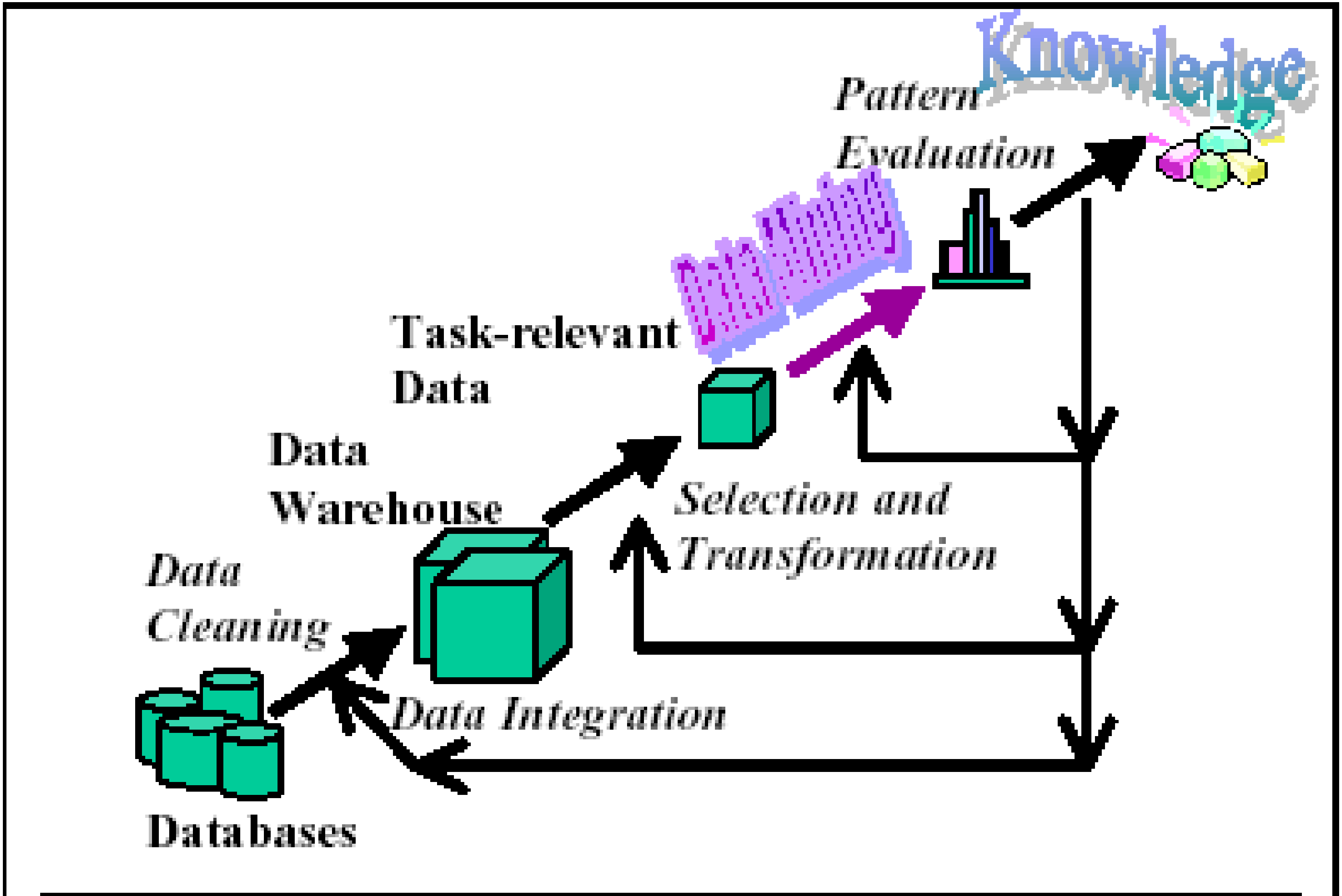
The word "**Mining"** refers to the extraction of valuable things like minerals from the earth.

Data mining is the process by which we can extract interesting patterns and knowledge from huge amounts of data



Data mining is also known as Knowledge Discovery in Databases -KDD

Data Mining


Knowledge
Pattern
Evaluation
Data Mining
Task-relevant
Data
Data
Warehouse
Selection and
Transformation
Data
Cleaning
Data Integration
Databases

Difference between Data Mining and Data Warehousing

Data Mining provides the Enterprise with **intelligence** and

Data Warehousing provides the Enterprise with a **memory.**

Data warehousing is the process that is used to integrate and combine data from multiple sources and format into a single unified schema. So it provides the enterprise with a storage mechanism for its huge amount of data.

On the other hand, Data mining is the process of extracting interesting patterns and knowledge from huge amount of data. So we can apply data mining techniques on the data warehouse of an enterprise to discover useful patterns.

Data Mining

Data mining consists of five major elements:

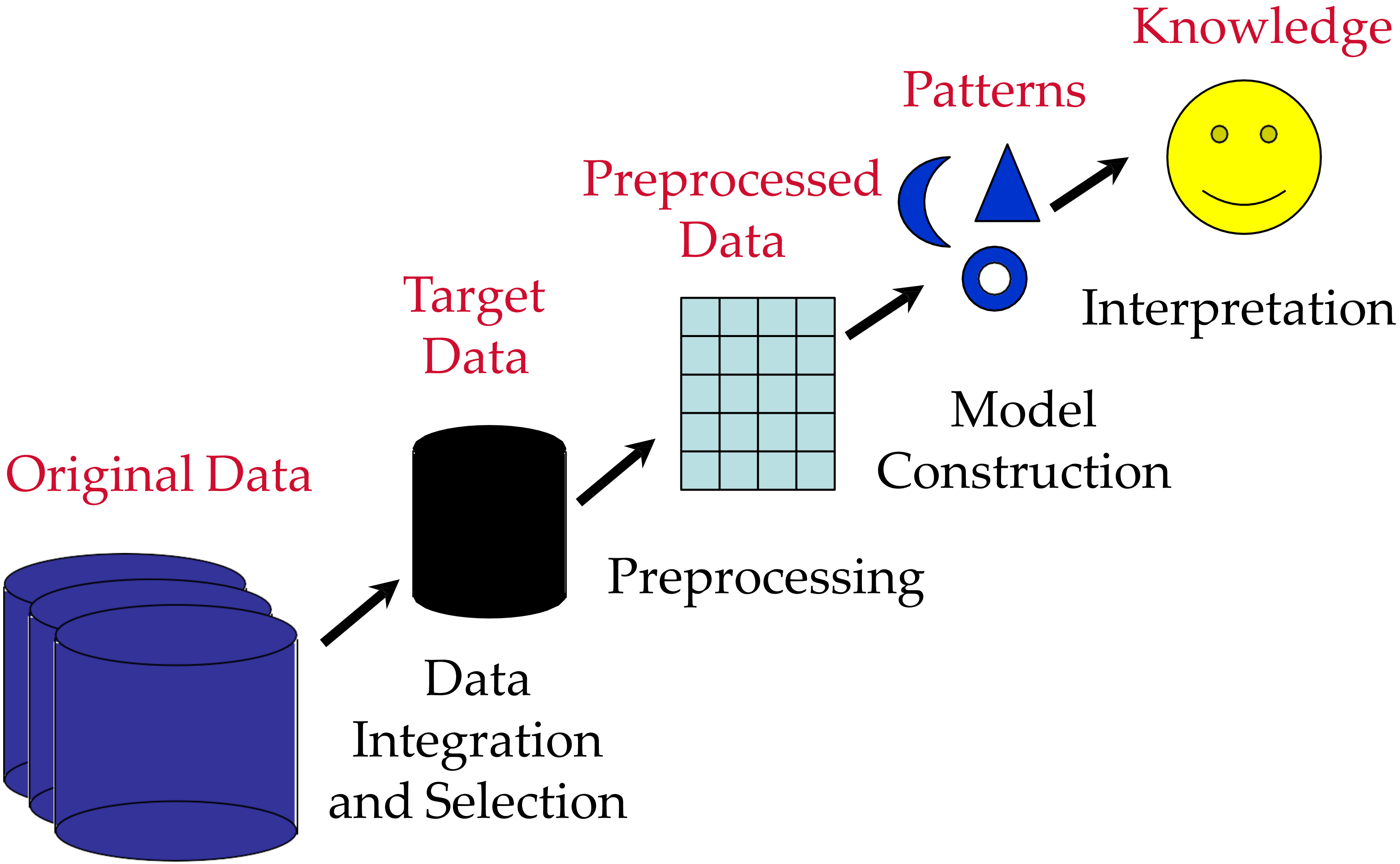
-Extract, transform, and load transaction data onto the data warehouse system.

- Store and manage the data in a multidimensional database system.

-Provide data access to business analysts and information technology professionals.

- Analyze the data by application software.

Preprocessing and Mining



Original Data

Data Integration and Selection

Target Data

Preprocessing

Preprocessed Data

Model Construction

Patterns

Interpretation

Knowledge

Steps of Data Mining

There are various steps that are involved in mining data as shown in the picture.

Data Integration: First of all the data are collected and integrated from all the different sources.

Data Selection: In this step we select only those data which we think useful for data mining.

Data Cleaning: The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get remove such anomalies.

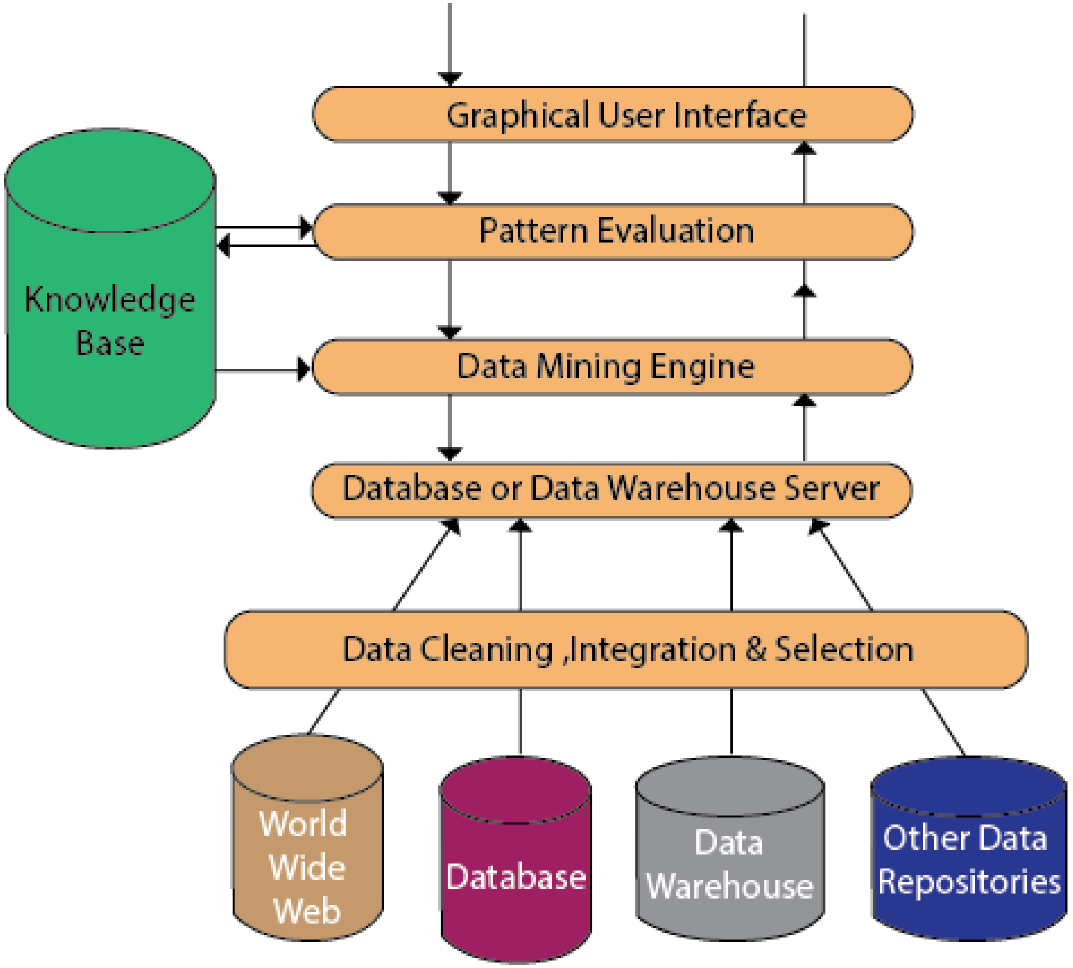
Data Transformation: The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.

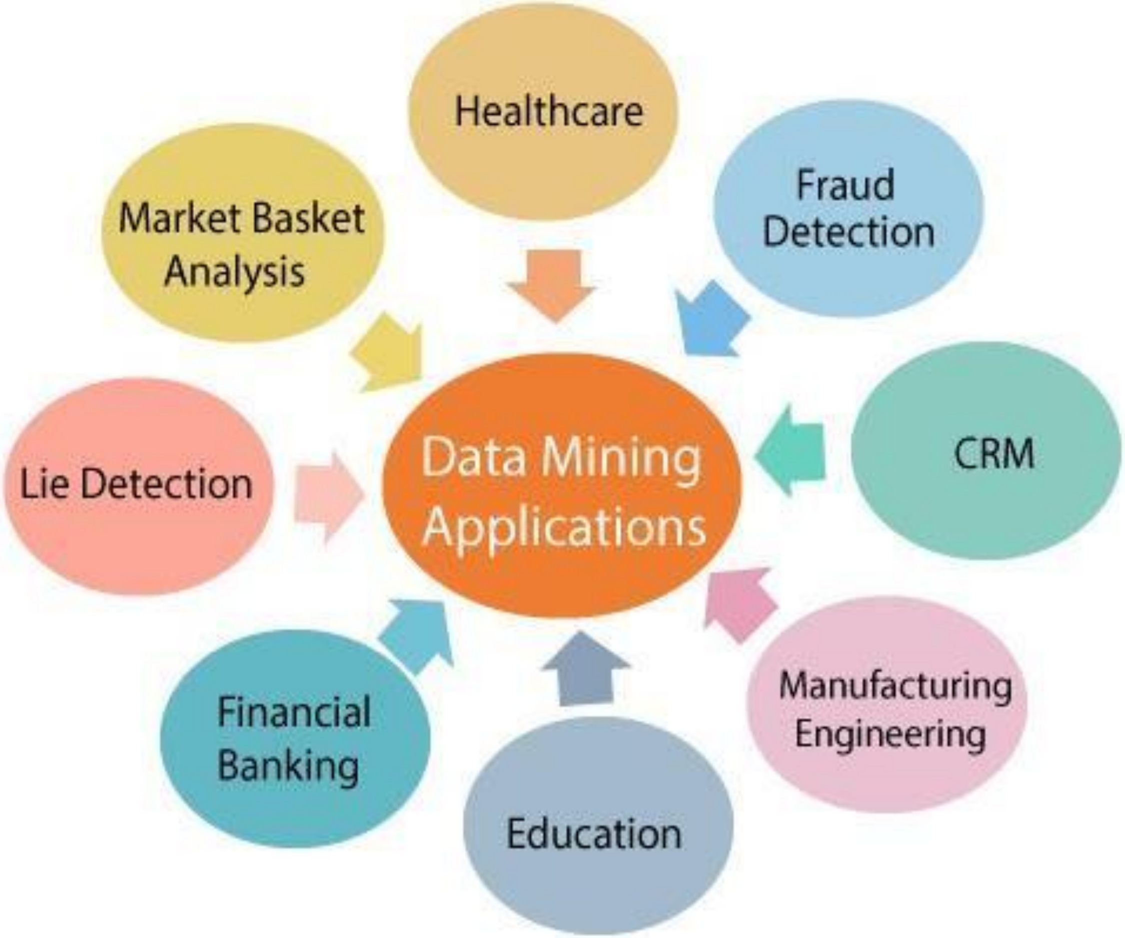
Data Mining: Now we are ready to apply data mining techniques on the data to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.

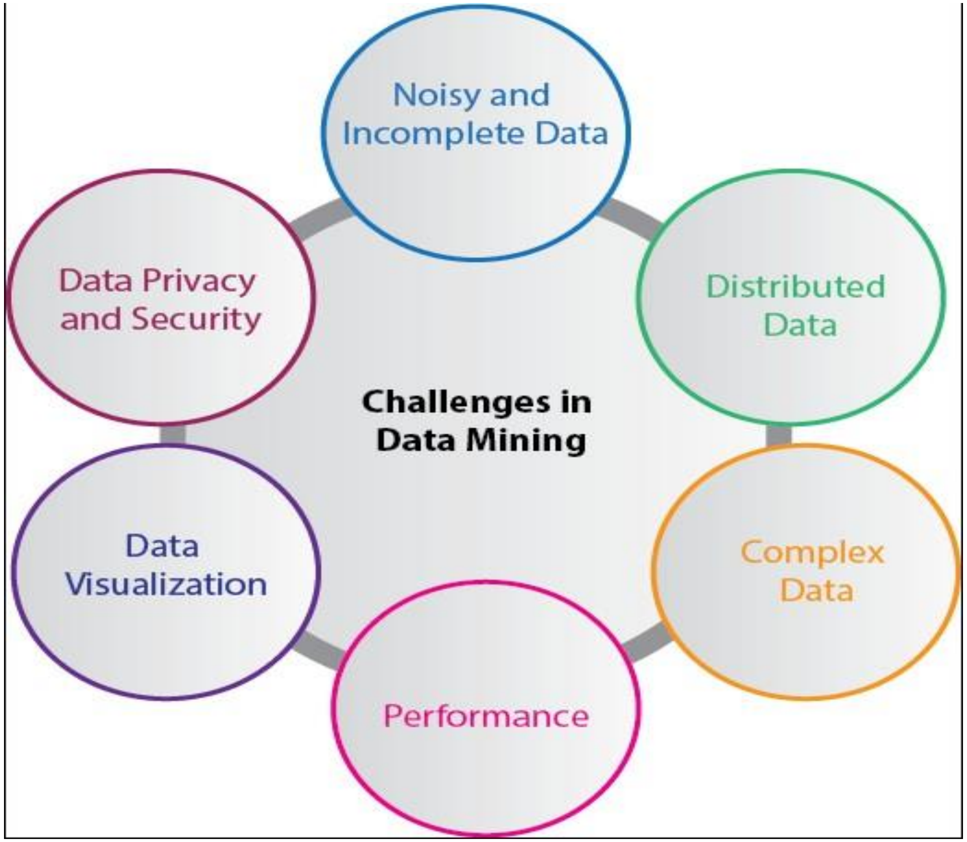
Pattern Evaluation and Knowledge Presentation: This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.

Decisions / Use of Discovered Knowledge: This step helps user to make use of the knowledge acquired to take better decisions.

Data Mining Architecture




Healthcare
Market Basket Analysis
Fraud Detection
Data Mining Applications
Lie Detection
CRM
Financial Banking
Education
Manufacturing Engineering


Noisy and Incomplete Data
Distributed Data
Complex Data
Performance
Data Visualization
Data Privacy and Security
Challenges in Data Mining

The most commonly used techniques in data mining are:

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on

Machine learning

Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to learn based on data, such as from sensor data or databases.

A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data.

Hence, machine learning is closely related to fields such as statistics, probability theory, **data mining**, pattern recognition, artificial intelligence, adaptive control, and theoretical computer science.

Machine learning Application

- Natural Language processing
- Syntactic pattern recognition
- Search engines
- Medical Diagnosis
- Bioinformatics
- Detecting Credit card fraud
- Stock market analysis
- Speech & Handwriting recognition
- Game playing
- Classifying DNA Sequence

Data Mining Techniques

Data mining techniques usually fall into two categories,

Descriptive

Predictive

Descriptive data mining aims to find patterns in the data that provide some information about what the data contains.

Predictive data mining uses historical data to infer something about future events.

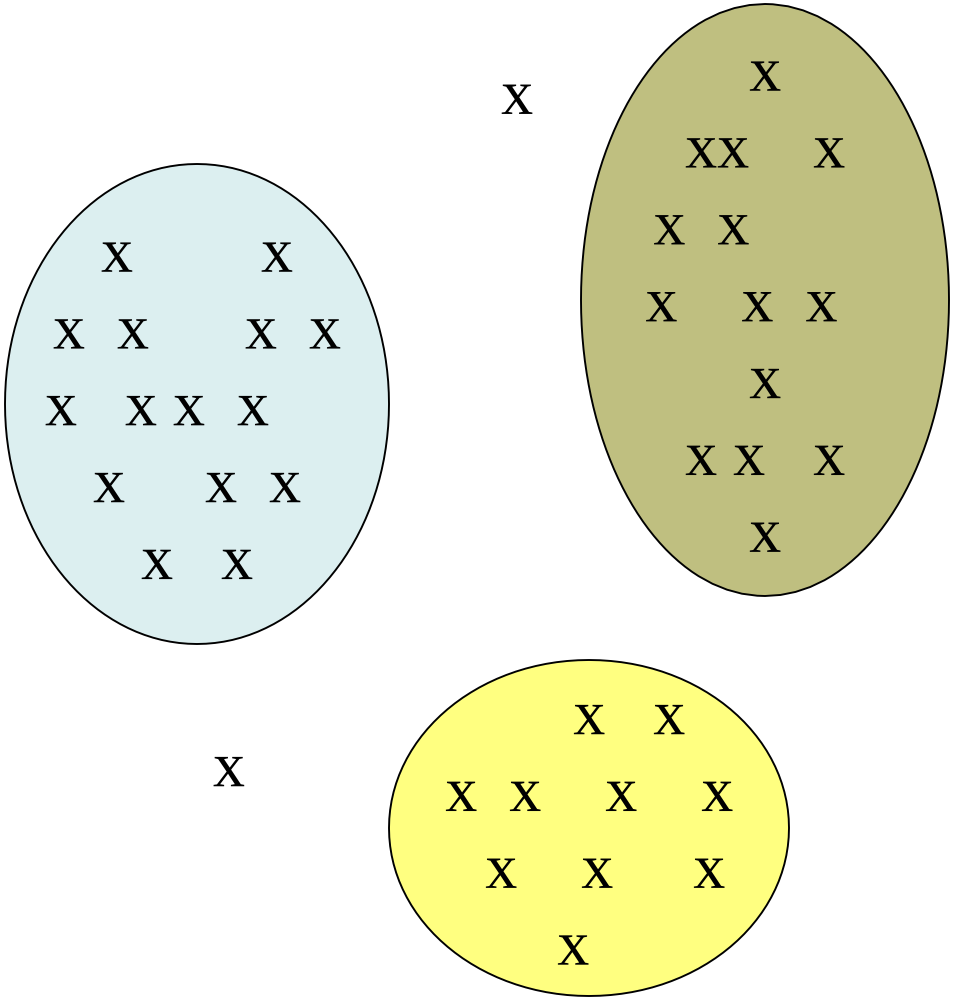
Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
- Most data mining methods discards outliers as noise or exception.
- The analysis of outlier data is referred to as outlier mining
- Outliers may be detected using statistical test that assume a distribution or probability model for data.

Cluster

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters

Example: Clusters

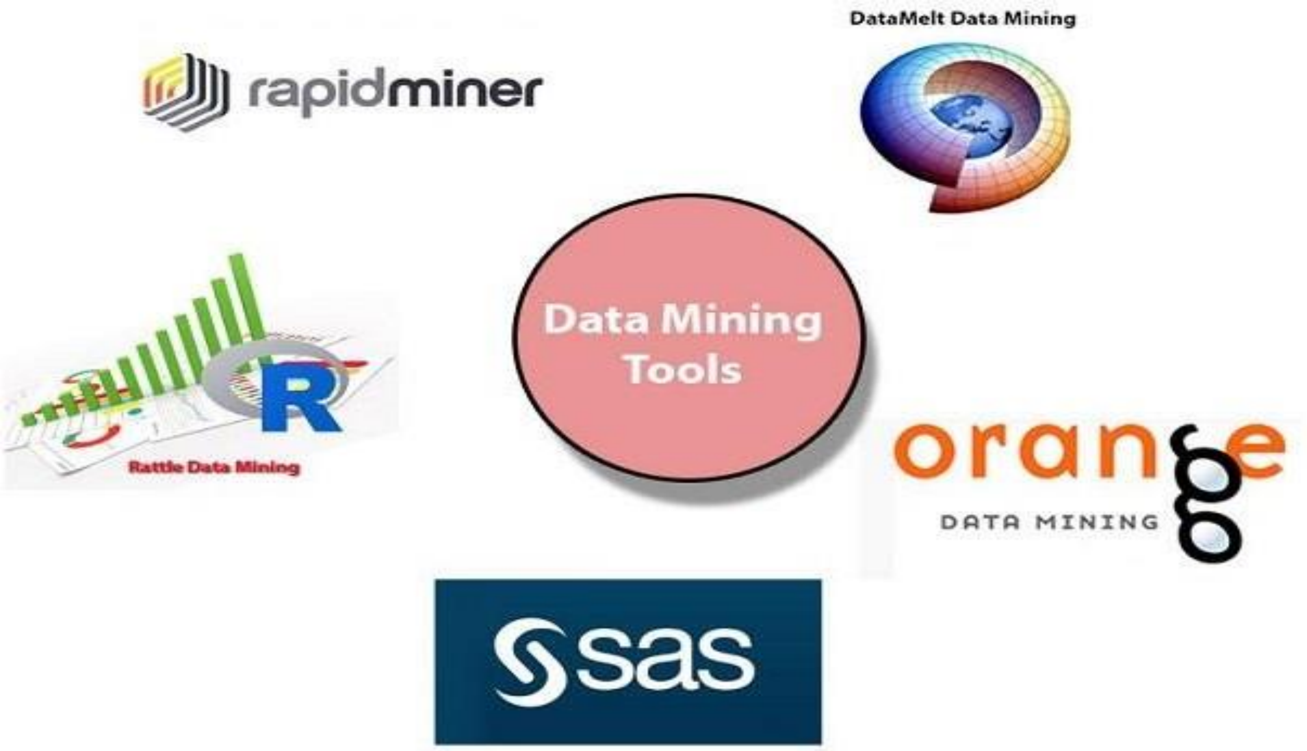


Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Profitable Applications: Data Mining Tools:

- most popular data mining tools


rapidminer
DataMelt Data Mining
Data Mining Tools
R
Rattle Data Mining
orange
DATA MINING
sas