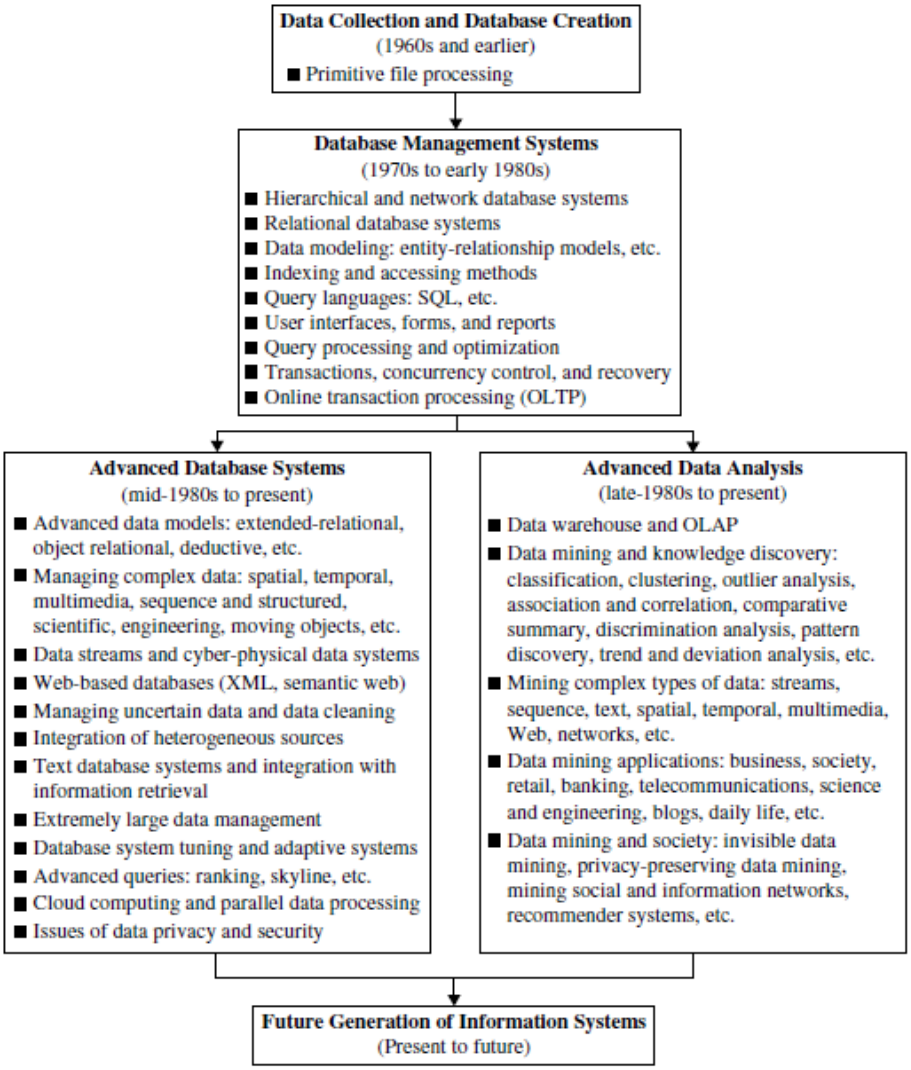
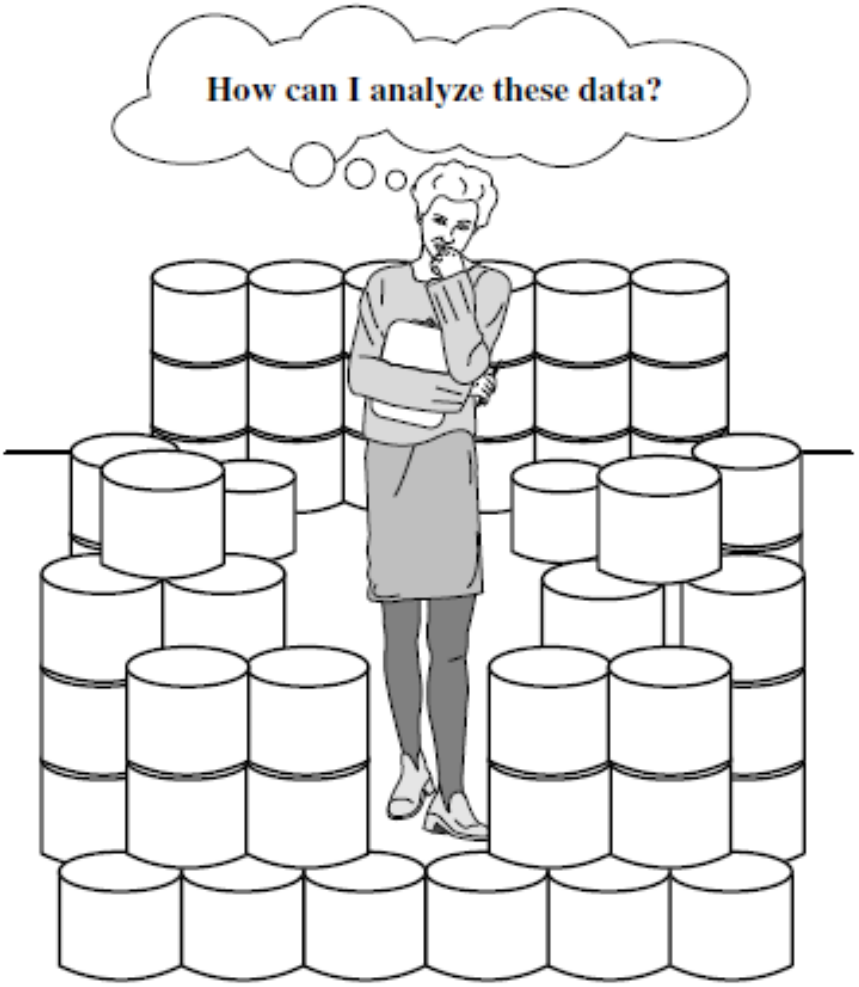
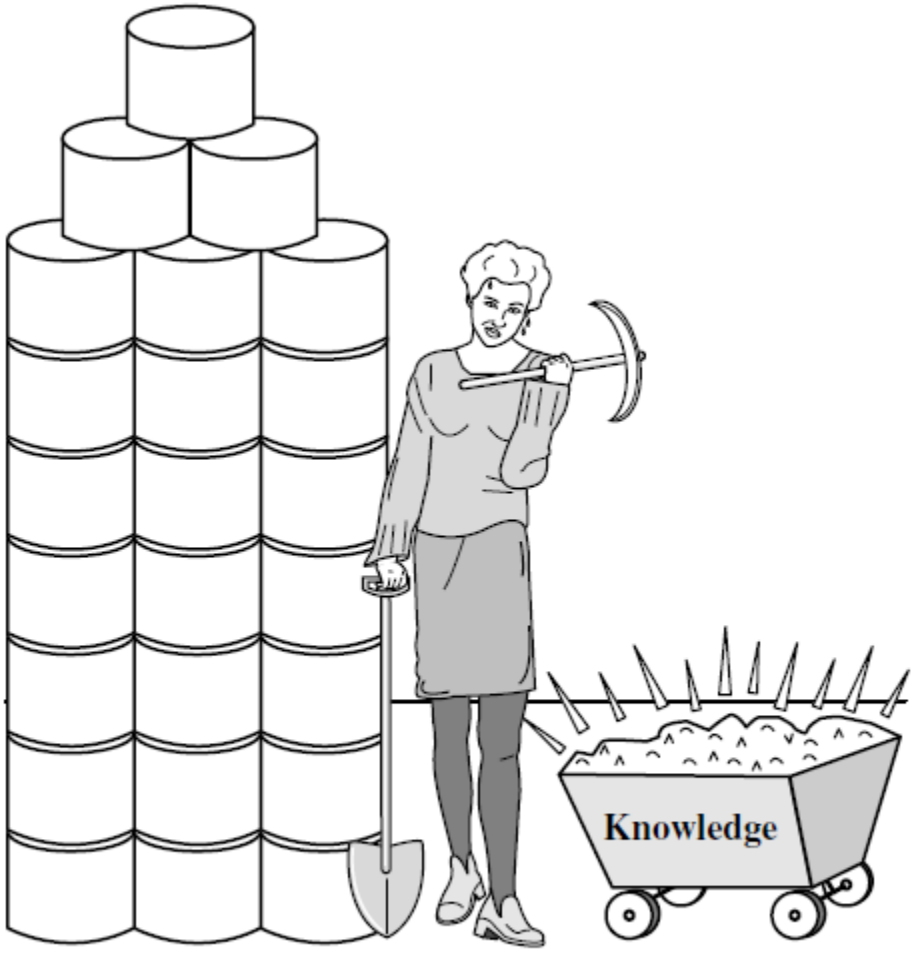
Data Mining as the Evolution of Information Technology





The world is data rich but information poor.

What Is Data Mining?



Data mining—searching for knowledge (interesting patterns) in data.

"**knowledge mining from data**," which is unfortunately somewhat long. However, the shorter term, *knowledge mining* may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer carrying both "data" and "mining" became a popular choice. In addition, many other terms have a similar meaning to data mining—for example, ***knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology,* and *data dredging*.**

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)

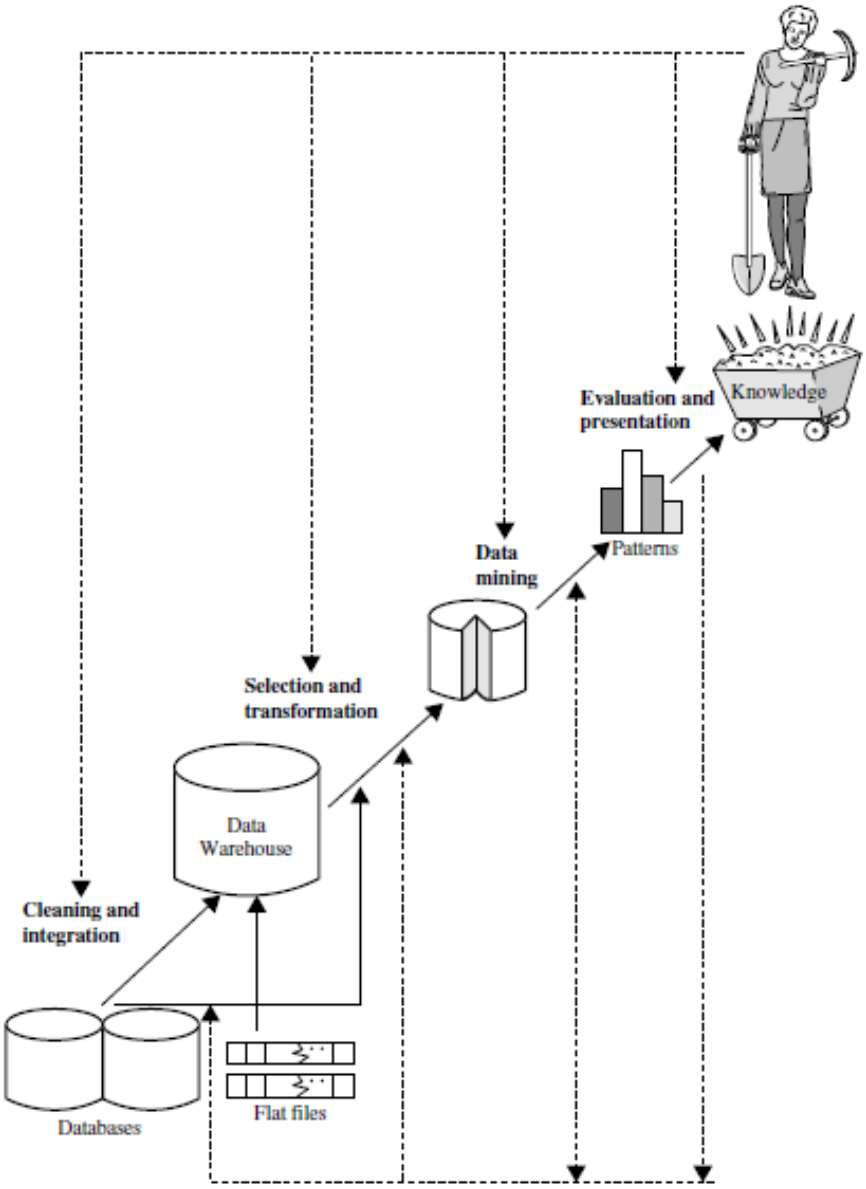
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

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 through 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. The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term *data mining* is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than *knowledge discovery from data*). Therefore, we adopt a broad view of data mining functionality: **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.

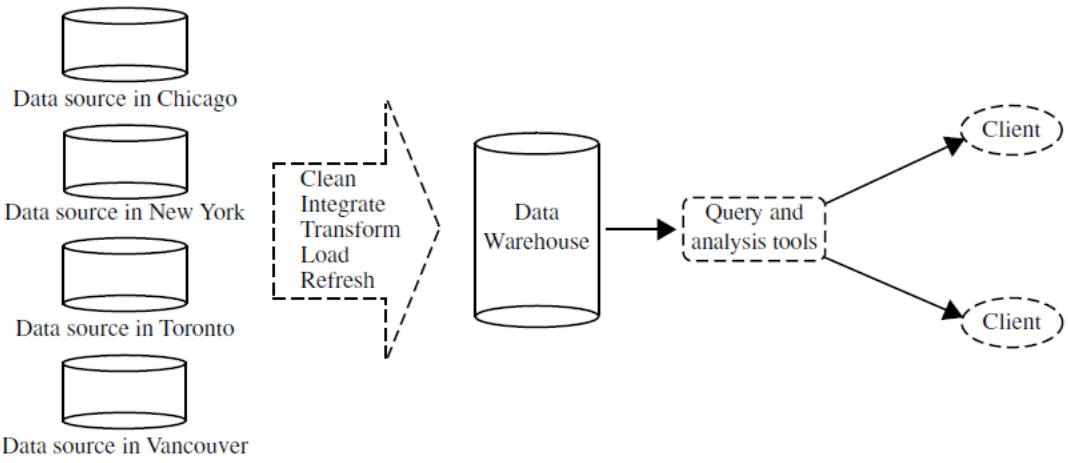


4 Data mining as a step in the process of knowledge discovery.

Data Warehouses

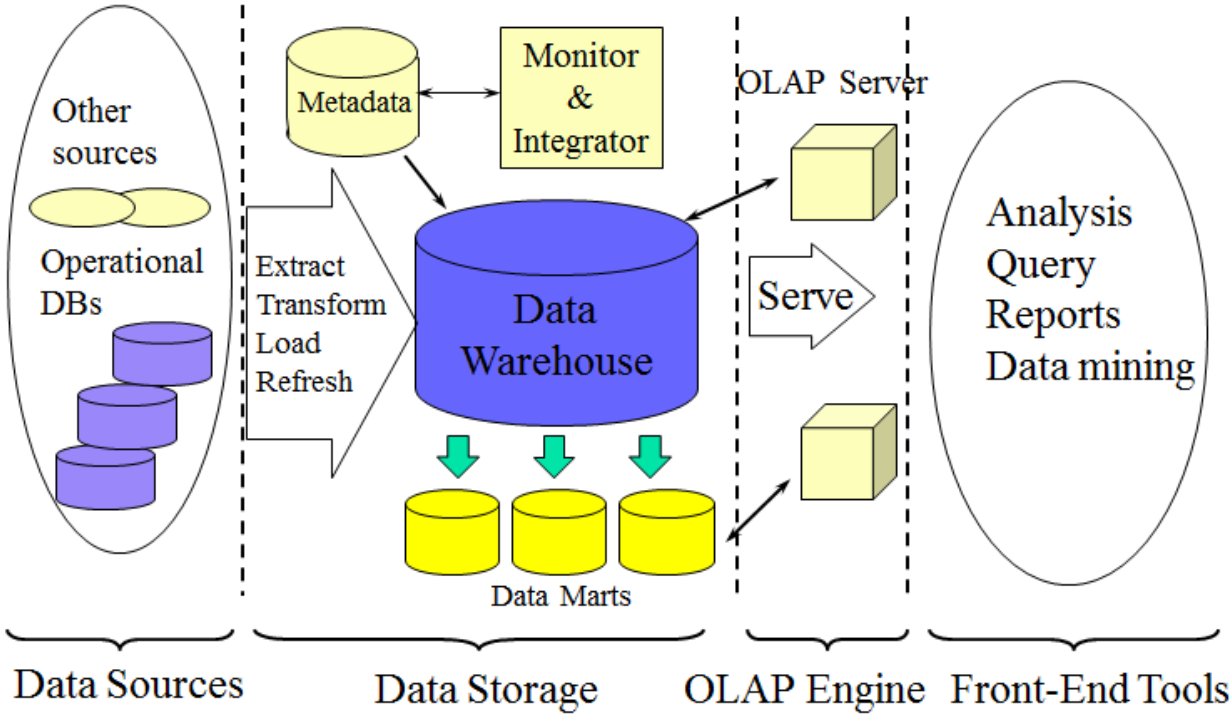
Suppose that *All Electronics* is a successful international company with branches around the world. Each branch has its own set of databases. The president of *All Electronics* has asked you to provide an analysis of the company's sales per item type per branch for the third quarter. This is a difficult task, particularly since the relevant data are spread out over several databases physically located at numerous sites.

If *All Electronics* had a data warehouse, this task would be easy. A **data warehouse** is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. Figure shows the typical framework for construction and use of a data warehouse for *All Electronics*.



Typical framework of a data warehouse for *AllElectronics*.

Data Warehouse: A Multi-Tiered Architecture

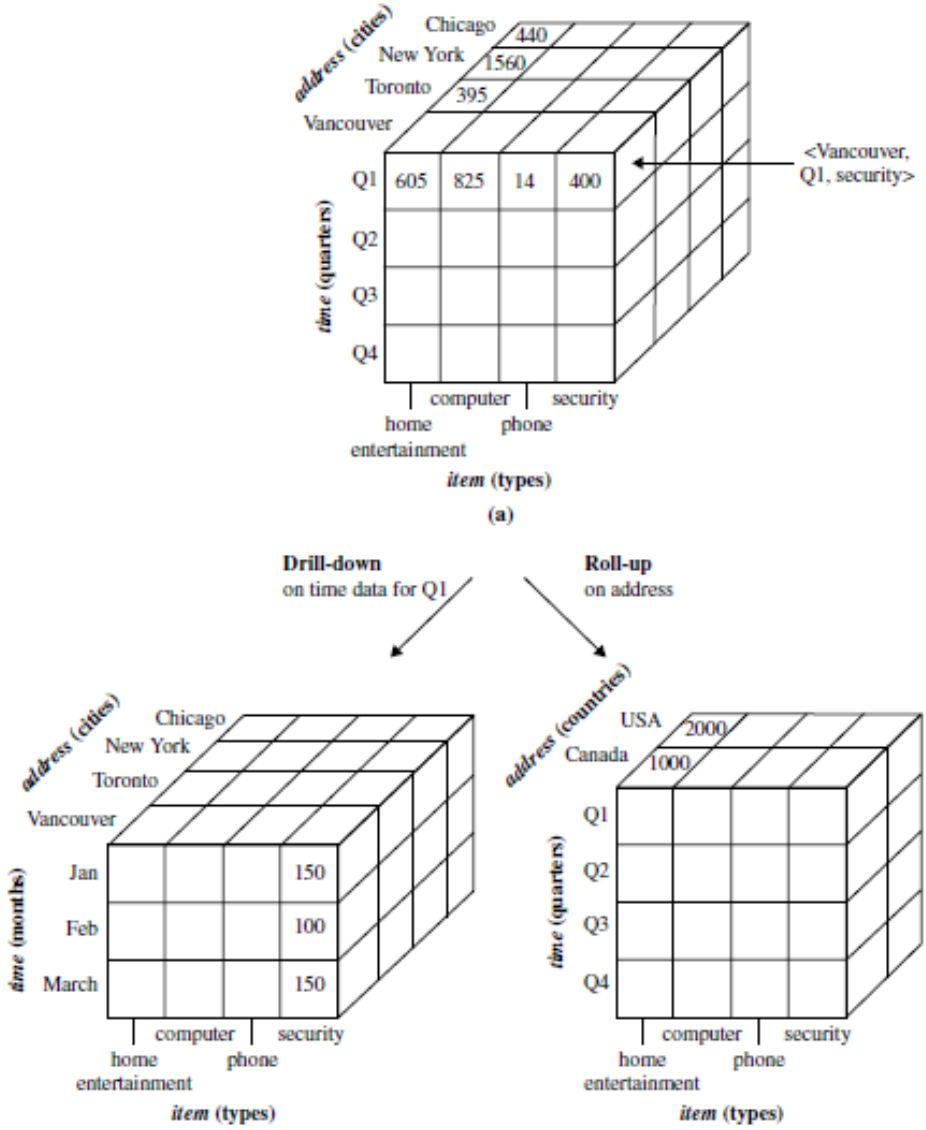


To facilitate decision making, the data in a data warehouse are organized around *major subjects* (e.g., customer, item, supplier, and activity). The data are stored to provide information from a *historical perspective*, such as in the past 6 to 12 months, and are typically *summarized*. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region. A data warehouse is usually modeled by a multidimensional data structure, called a **data cube**, in which 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 *sum*. *sales amount*. A data cube provides a multidimensional view of data and allows the pre-computation and fast access of summarized data.

Example

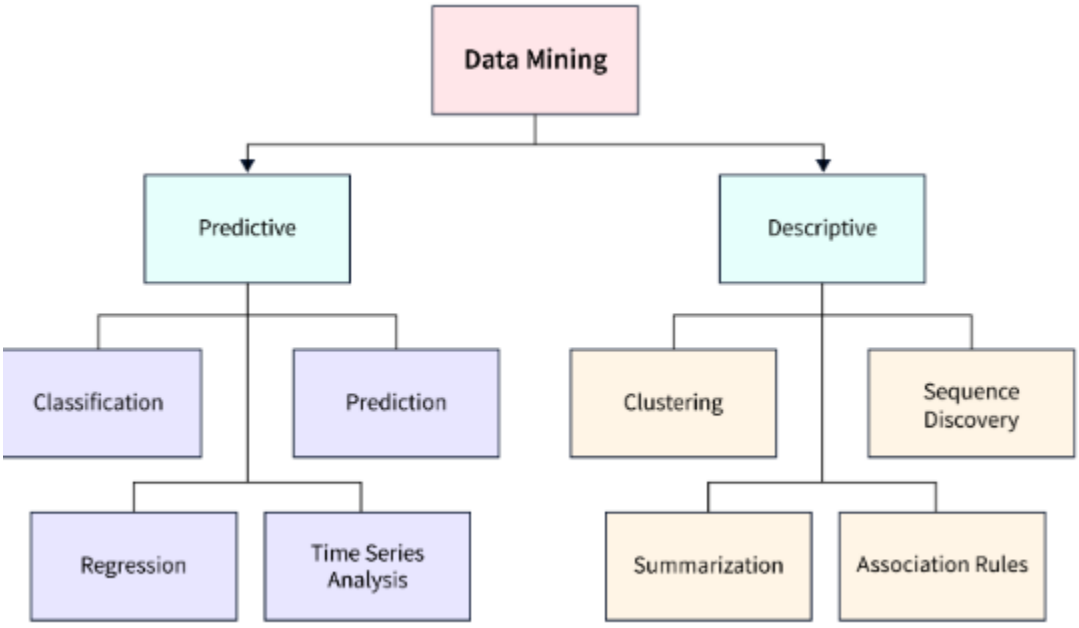
A data cube for *All Electronics*. A data cube for summarized sales data of *All Electronics* is presented in Figure. The cube has three dimensions: *address* (with city values *Chicago, New York, Toronto, Vancouver*), *time* (with quarter values *Q1, Q2, Q3, Q4*), and *item* (with item type values *home entertainment, computer, phone, security*).The aggregate value stored in each cell of the cube is *sales amount* (in thousands). For example, the total sales for the first quarter, *Q1*, for the items related to security systems in Vancouver is $400, 000, as stored in cell (*Vancouver, Q1, security*). Additional cubes may be used to store aggregate sums over each dimension,

corresponding to the aggregate values obtained using different SQL group-bys (e.g., the total sales amount per city and quarter, or per city and item, or per quarter and item, or per each individual dimension). By providing multidimensional data views and the pre-computation of summarized data, data warehouse systems can provide inherent support for OLAP. Online analytical processing operations make use of background knowledge regarding the domain of the data being studied to allow the presentation of data at *different levels of abstraction*. Such operations accommodate different user viewpoints. Examples of OLAP operations include **drill-down** and **roll-up**, which allow the user to view the data at differing degrees of summarization, as illustrated in Figure. For instance, we can drill down on sales data summarized by *quarter* to see data summarized by *month*. Similarly, we can roll up on sales data summarized by *city* to view data summarized by *country*. Although data warehouse tools help support data analysis, additional tools for data mining are often needed for in-depth analysis. **Multidimensional data mining** (also called **exploratory multidimensional data mining**) performs data mining in multidimensional space in an OLAP style. That is, it allows the exploration of multiple combinations of dimensions at varying levels of granularity in data mining, and thus has greater potential for discovering interesting patterns representing knowledge.



7 A multidimensional data cube, commonly used for data warehousing, (a) showing summarized data for *AllElectronics* and (b) showing summarized data resulting from drill-down and roll-up operations on the cube in (a). For improved readability, only some of the cube cell values are shown.

Classification/Types/Activates/functionalities/Task of Data Mining



Data mining encompasses two primary categories of activities - descriptive and predictive data mining.

1. Descriptive Data Mining

Descriptive data mining focuses on uncovering patterns, trends, and relationships within existing data. This category of data mining doesn't aim to make predictions but rather seeks to provide valuable insights into historical or current data. Through techniques like clustering and association rule mining, descriptive data mining can help organizations understand customer behavior, segment their market, or identify anomalies within their datasets. For instance, a retail company might use descriptive data mining to discover customer purchasing patterns, helping it optimize inventory management and marketing strategies.

- ### Clustering

Cluster analysis, or called clustering, is a process of data mining where similar data points are identified and grouped. The idea of clustering is to find homogeneous

groups of data points that shed light on certain group characteristics while minimizing intra-group similarity, i.e., different groups should be distinct.

It is commonly used in customer behavior analysis, fraud detection, etc. Several algorithms allow you to perform clustering, such as K-means clustering, DBSCAN, Gaussian mixture models, mean-shift algorithm, hierarchical clustering using AGNES or DIANA, etc.

- **Visualization/Summarization**

  The primary aim of this type of data mining is to describe data. Therefore, the most common task associated with it is visualization. The idea is to use graphs and charts to represent the data visually. This allows users to summarize the data, identify trends and patterns, and describe the key point in an easy-to-understand medium, which is difficult to do just by looking at the raw data.

  Common visualization schemes used here are histograms, line charts, boxplots, scatterplots for numerical columns, and bar charts and pie charts for categorical and numerical-categorical columns.

- **Sequence discovery and Path Analysis**

  Another task commonly performed in data mining is to find a pattern such that a particular set of events (i.e., values or data points) leads to subsequent events. This identification of the sequence of events or the path that events undertake is called sequence and path analysis. Such analysis is commonly used in e-commerce, online games, etc, to understand how consumers navigate on their platforms.

- **Association Rule Mining**

Association rule mining is used to identify the relationship between two or more variables (attributes/features) in the data. It is also used to identify co-occurring

events. Thus, it discovers relationships between the data points and uncovers the rules that bind them.

One of the most common use cases of association rule mining is in retail, where transaction data is used to find products that are frequently bought together. This task, known as market basket analysis, employs several association rule mining algorithms, such as the Apriori algorithm.

2. Predictive Data Mining

Predictive data mining, on the other hand, goes beyond description to predict future outcomes. It leverages historical data and statistical algorithms to build models that can make predictions or classifications. Common applications include forecasting sales, detecting fraud, or predicting disease outbreaks. For instance, predictive data mining in healthcare can analyze patient records to predict disease risk factors, enabling early intervention and personalized treatment plans. The power of predictive data mining lies in its ability to harness past data to make informed decisions and drive proactive actions, ultimately improving efficiency and competitiveness.

- **Classification**

In classification, well-labeled historical data is used to understand how different data points are associated with different classes. Once a classification model understands this relationship, any new data point can be classified easily.

It is commonly used for churn prediction, loan default risk assessment, item categorization, etc. Common algorithms associated with classification are logistic regression, naïve Bayes, support vector machine classifier, etc.

- **Regression**

Regression is akin to classification, but it differs in that it predicts continuous values instead of classes. Companies often employ this method when predicting

variables like product sales or the success of a marketing campaign. A closely related concept is forecasting, where values are predicted by considering the effects of time.

Common forecasting tasks include weather forecasting, predicting stock prices over the short term, web traffic prediction, etc. Algorithms commonly associated with regression and forecasting are linear regression, ARIMA, Holt-Winters, ARIMAX, GARCH, etc.

- **Prediction**

Prediction is the major functionality of data mining. It encompasses the classification functionality mentioned above but can also be regarded as a stand-alone feature. This functionality analyzes data and comprehends the relationship between data points and their respective classes.

In predictions, however, the core focus is statistical and machine learning techniques to predict any new data point. Here, all types of problems are solved, such as regression, forecasting, and classification.

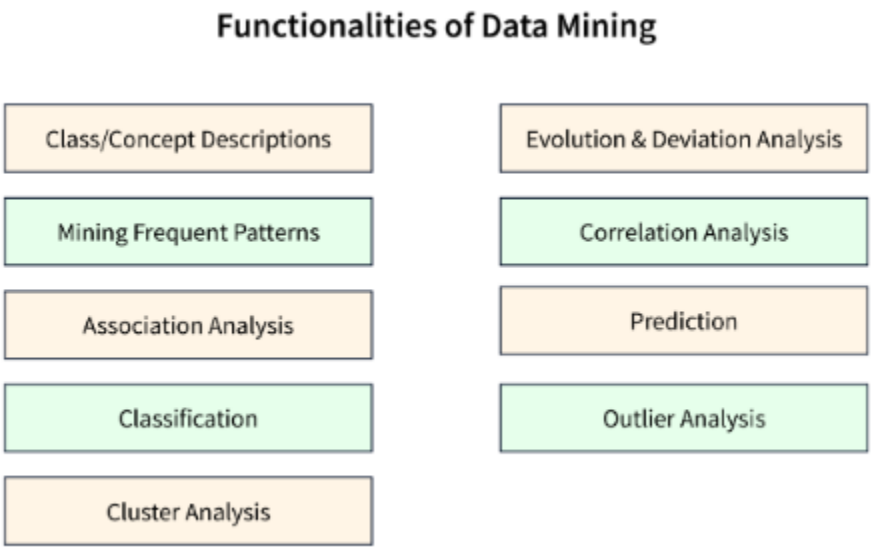
For example, a bank predicts the amount of cash required in an ATM on various days.

- **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.

Functionalities of Data Mining

This section will explore various data mining functionalities as mentioned below.



1. Class/Concept Descriptions

Class or concept descriptions are crucial in understanding and categorizing data in data mining. There are two key categories in this context - data characterization and data discrimination.

- **Data Characterization** - Data characterization summarizes a given dataset's general features and characteristics. It provides a comprehensive view of the data's distribution, central tendencies, and overall structure. This category often employs statistical measures, visualizations, and descriptive techniques to clearly and concisely represent the data's properties. For example, in financial analysis, data characterization might involve generating summary statistics and visualizations to understand the historical performance of a stock portfolio, helping investors assess risk and make informed decisions.
- **Data Discrimination** - Data discrimination, in contrast, is concerned with distinguishing between different classes or categories within a dataset. It

aims to find features or patterns that can effectively separate one class from another. This category is commonly used in classification tasks, such as spam email detection or sentiment analysis in natural language processing. For instance, in email filtering, data discrimination techniques can analyze the content and metadata of emails to determine whether they belong to the "spam" or "not spam" category, helping users manage their inboxes effectively.

2. Mining Frequent Patterns

A key role of data mining involves the identification of data patterns, specifically those that occur frequently within a dataset. These frequent patterns manifest in various forms -

- **Frequent Item Sets** - This term pertains to sets of items that tend to co-occur regularly within the data. For instance, it might reveal that products like milk and sugar are commonly purchased together, providing insights into consumer buying habits.
- **Frequent Substructures** - Frequent substructures refer to diverse data structures that can be associated with item sets or subsequences. Examples include trees and graphs, which often appear in conjunction with certain patterns, unveiling deeper relationships within the data.
- **Frequent Subsequences** - This category involves the identification of recurring sequential patterns. For instance, it may uncover a pattern where customers frequently purchase a phone followed by a phone cover, highlighting sequences of events or actions within the data.

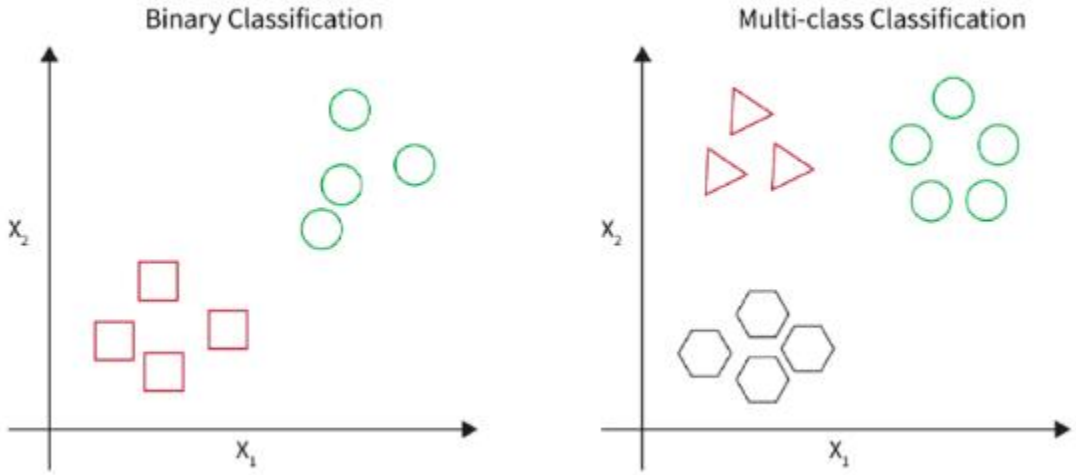
3. Association Analysis

It examines the group of items that commonly appear together within a transactional dataset. This technique is often called Market Basket Analysis due to its prevalent application in the retail industry. To establish association rules, two key parameters are employed:

- **Support** - This parameter identifies the frequency of occurrence of a particular item set within the database.
- **Confidence** - Confidence represents the conditional probability that an item will appear in a transaction, given the occurrence of another item in the same transaction.

4. Classification

Classification in data mining is a technique used to categorize data into predefined classes or categories based on specific attributes or characteristics. It involves the application of algorithms like decision trees, neural networks, or support vector machines to assign objects or records to distinct classes. This process is valuable for tasks such as spam email detection, sentiment analysis, and disease diagnosis, where data needs to be sorted into relevant groups to facilitate decision-making and pattern recognition.



5. Prediction

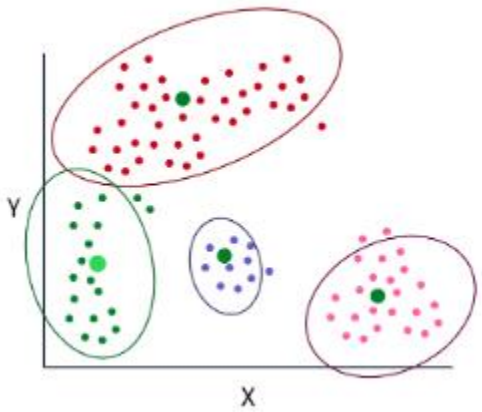
Prediction in data mining is the process of using historical data and patterns to make informed estimates about future or missing data values. It involves the application of various algorithms and techniques to anticipate numerical values, such as sales figures, or to classify items into predefined categories. This predictive capability enables businesses and researchers to make data-driven decisions,

identify trends, and enhance their understanding of complex datasets, ultimately facilitating better planning and strategy development.

- **Numeric Prediction** - Numeric prediction involves forecasting numerical values based on historical data, typically using techniques like linear regression. It helps businesses prepare for future events or trends impacting their operations.
- **Class Prediction** - Class prediction assigns missing class labels to items using a training dataset where class labels are known. It's valuable for categorizing items or objects and improving data completeness and decision-making.

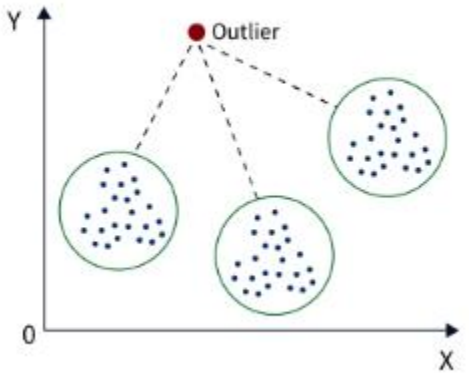
6. Cluster Analysis

Cluster analysis in data mining is a method used to group similar data points together based on their inherent characteristics or attributes. It aims to discover patterns and relationships within data by identifying clusters or clusters of data points that share common features. This technique has various applications, including customer segmentation, anomaly detection, and data compression, and helps uncover hidden structures within datasets, enabling more effective decision-making and data exploration.



7. Outlier Analysis

Outlier analysis in data mining is the process of identifying and examining data points that significantly deviate from the expected or normal patterns within a dataset. These outliers, often anomalies or exceptions, may hold valuable information or indicate errors, fraud, or rare events. Outlier analysis helps detect unusual data instances that can impact data quality, decision-making, and the discovery of novel insights, making it crucial in various domains like fraud detection, quality control, and anomaly detection.



8. Evolution and Deviation Analysis

Evolution analysis in data mining involves tracking changes and patterns over time within a dataset. It aims to identify trends, variations, and evolving relationships in temporal data, enabling businesses to make informed decisions based on historical and current trends. Deviation analysis, on the other hand, focuses on identifying deviations or anomalies in data compared to expected or normative patterns.

9. Correlation Analysis

Correlation analysis in data mining involves examining the statistical relationships between two or more variables within a dataset. It quantifies the degree to which changes in one variable are associated with changes in another, providing insights into their interdependence. This analysis helps identify patterns,

dependencies, and associations between variables, enabling businesses to make data-driven decisions.

Influence from many disciplines

Depending on data mining approach, techniques from other disciplines may be applied such as

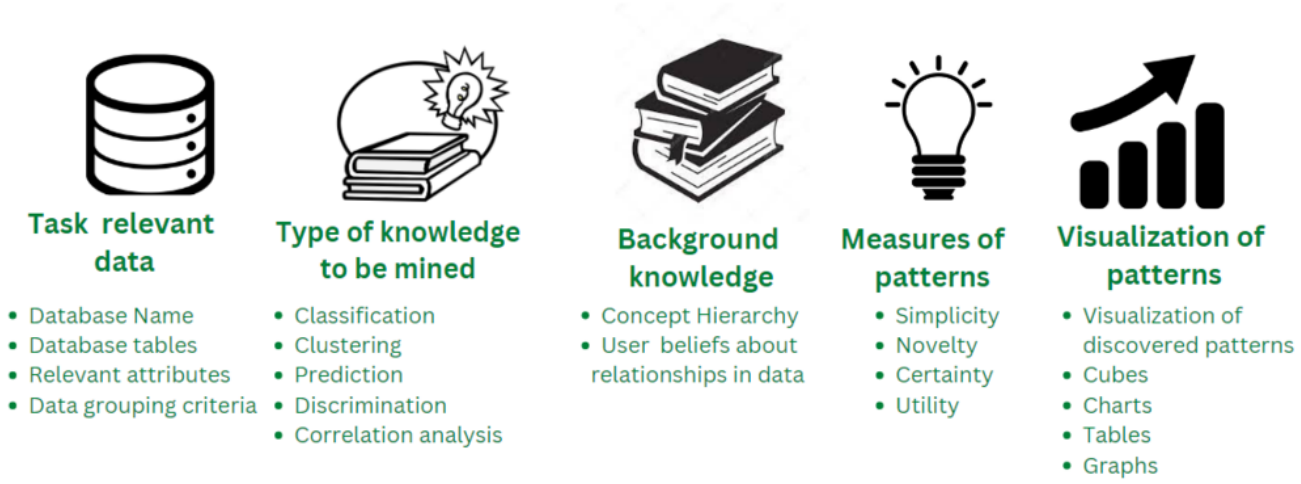
- Information Retrieval
- Artificial Intelligence
- Neural networks
- Fuzzy set theory
- Knowledge representation
- Logic programming
- High performance computing

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.



The Data Mining Task Primitives are as follows:

1. The set of task relevant data to be mined: It refers to the specific data that is relevant and necessary for a particular task or analysis being conducted using data mining techniques. This data may include specific attributes, variables, or characteristics that are relevant to the task at hand, such as customer demographics, sales data, or website usage statistics. The data selected for mining is typically a subset of the overall data available, as not all data may be necessary or relevant for the task. For example: Extracting the database name, database tables, and relevant required attributes from the dataset from the provided input database.

2. Kind of knowledge to be mined: It refers to the type of information or insights that are being sought through the use of data mining techniques. This describes the data mining tasks that must be carried out. It includes various tasks such as classification, clustering, discrimination, characterization, association, and evolution analysis. For example, It determines the task to be performed on the relevant data in order to mine useful information such as classification, clustering, prediction, discrimination, outlier detection, and correlation analysis.

3. Background knowledge to be used in the discovery process: It refers to any prior information or understanding that is used to guide the data mining

process. This can include domain-specific knowledge, such as industry-specific terminology, trends, or best practices, as well as knowledge about the data itself. The use of background knowledge can help to improve the accuracy and relevance of the insights obtained from the data mining process. For example, The use of background knowledge such as concept hierarchies, and user beliefs about relationships in data in order to evaluate and perform more efficiently.

4. Interestingness measures and thresholds for pattern evaluation: It refers to the methods and criteria used to evaluate the quality and relevance of the patterns or insights discovered through data mining. Interestingness measures are used to quantify the degree to which a pattern is considered to be interesting or relevant based on certain criteria, such as its frequency, confidence, or lift. These measures are used to identify patterns that are meaningful or relevant to the task. Thresholds for pattern evaluation, on the other hand, are used to set a minimum level of interestingness that a pattern must meet in order to be considered for further analysis or action. For example: Evaluating the interestingness and interestingness measures such as utility, certainty, and novelty for the data and setting an appropriate threshold value for the pattern evaluation.

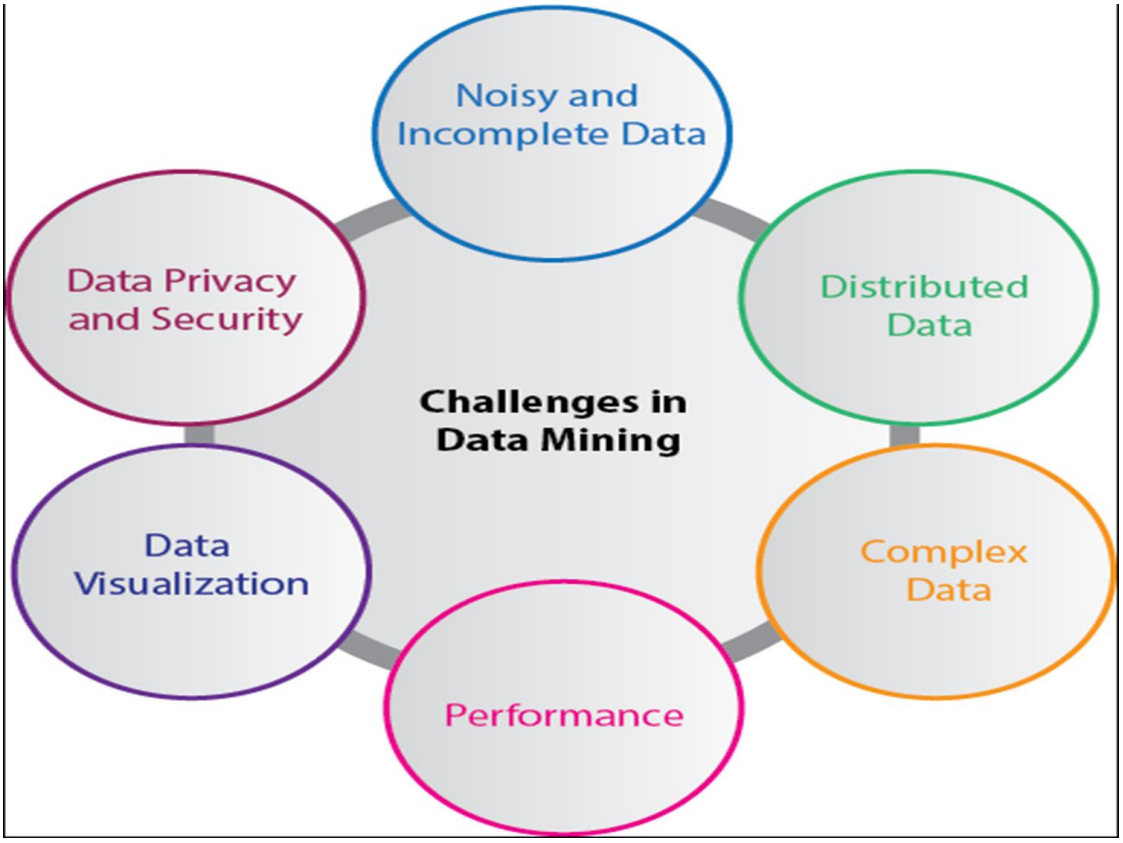
5. Representation for visualizing the discovered pattern: It refers to the methods used to represent the patterns or insights discovered through data mining in a way that is easy to understand and interpret. Visualization techniques such as charts, graphs, and maps are commonly used to represent the data and can help to highlight important trends, patterns, or relationships within the data. Visualizing the discovered pattern helps to make the insights obtained from the data mining process more accessible and understandable to a wider audience, including non-technical stakeholders. For example Presentation and visualization of discovered pattern data using various visualization techniques such as barplot, charts, graphs, tables, etc.

Challenges of Data Mining

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to:

- Performance,
- Data,
- Methods, and techniques, etc.

The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.



- **Incomplete and noisy data**

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than $ 500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data.

Even some customers may not be willing to disclose their phone numbers, which results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data)makes data mining challenging.

- **Data Distribution**

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, It is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

- **Complex Data**

Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

- **Performance**

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

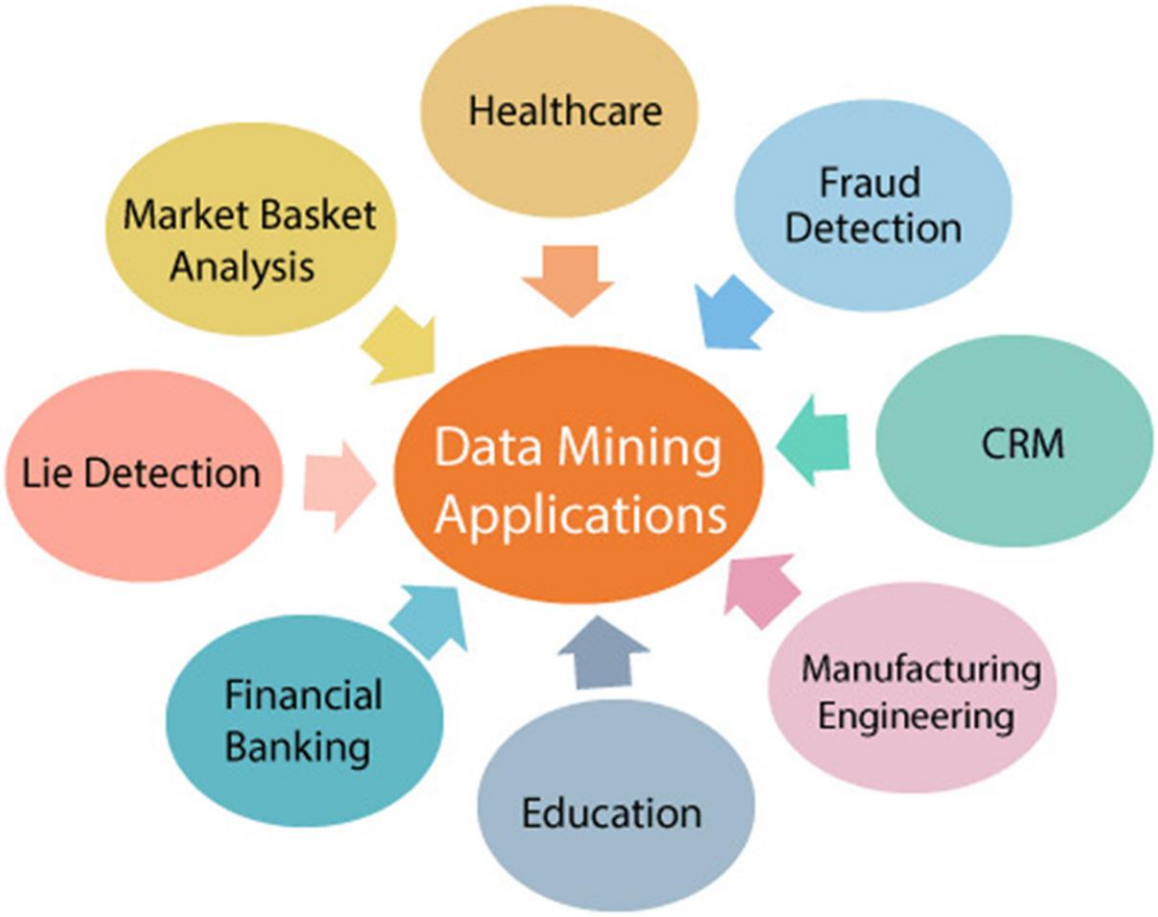
- **Data Privacy and Security**

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

- **Data Visualization**

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

Data Mining Applications



- **Data Mining in Healthcare**

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

- **Data Mining in Market Basket Analysis**

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

- **Data Mining in CRM (Customer Relationship Management)**

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

- **Data Mining in Fraud detection**

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

- **Data Mining in Lie Detection**

Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

- **Data Mining Financial Banking**

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.