

Data mining

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. “We are living in the information age” is a popular saying; however, we are actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business society, science and engineering, medicine, and almost every other aspect of daily life

What Is Data Mining?

Data mining refers to extracting or mining knowledge from large amounts of data. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

KDD steps – STEPS IN KNOWLEDGE DISCOVERY FROM DATA

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 terms knowledge discovery in databases (KDD) and data mining are often used interchangeably.

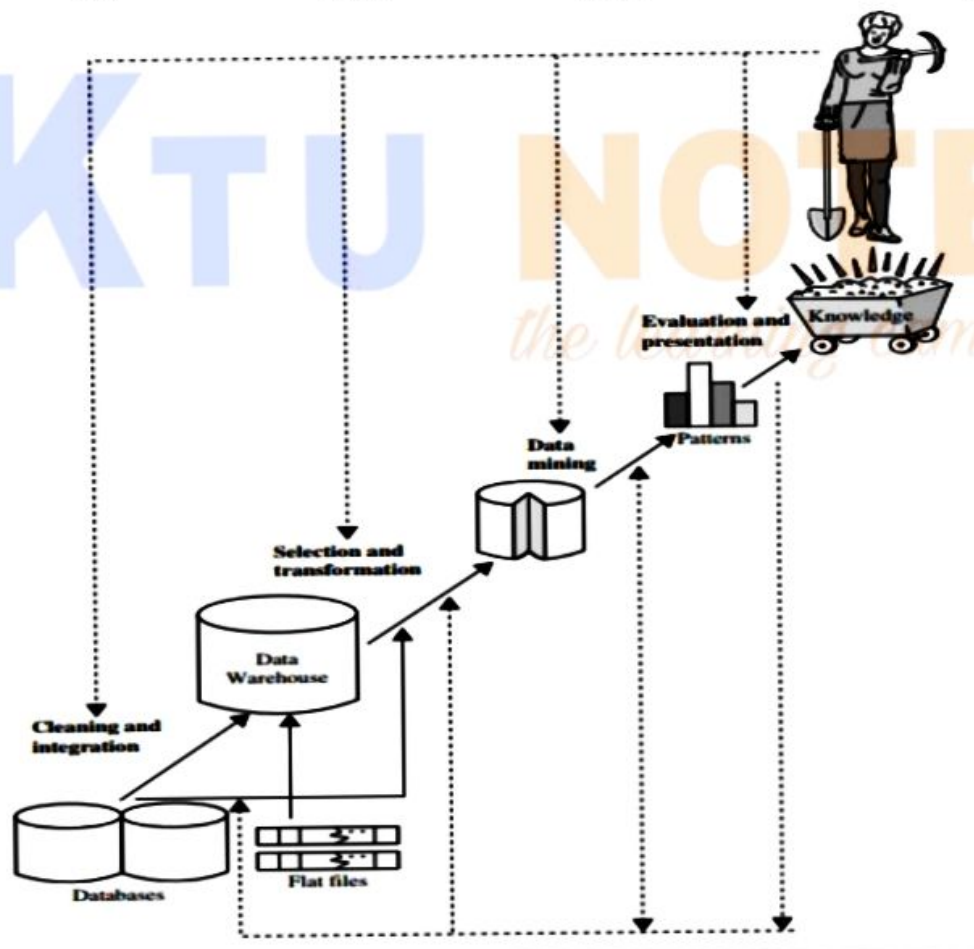


Figure 1.4 Data mining as a step in the process of knowledge discovery.

Over the last few years KDD has been used to refer to a process consisting of many steps, while data mining is only one of these steps.

Knowledge discovery in databases (KDD) is the process of finding useful information and patterns in data. Data mining is the use of algorithms to extract the information and patterns derived by the KDD process

Data mining stages:



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.

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.

Data mining applications:

1. Classification: Eg: In loan database, to classify an applicant as a prospective or defaulter, given his various personal and demographic features along with previous purchase characteristics.
2. Estimation: Predict the attribute of a data instance. Eg: estimate the percentage of marks of a student, whose previous marks are already known.
3. Prediction: Predictive model predicts a future outcome rather than the current behaviour. Eg: Predict next week's closing price for the Google share price per unit.
4. Market basket analysis(association rule mining)

Analyses hidden rules called association rule in a large transactional database.

{pen, pencil-> book} – whenever pen and pencil are purchased together, book is also purchased.

5. Clustering Classification into different classes based on some similarities but the target classes are unknown.
6. Business intelligence
7. Business data analytics
8. Bioinformatics
9. Web mining
10. Text mining
11. Social network data analysis

Data mining models

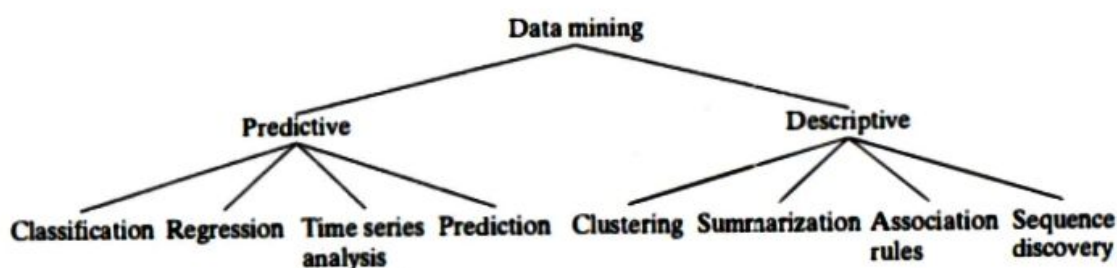


FIGURE 1.5 Data mining models and tasks.

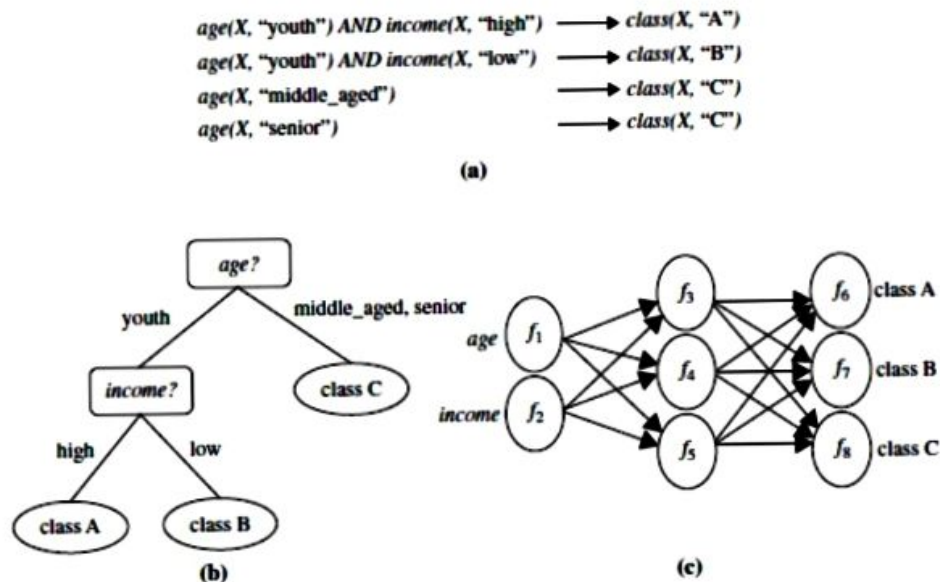


Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Data mining applications:

1. Classification: Eg: In loan database, to classify an applicant as a prospective or defaulter, given his various personal and demographic features along with previous purchase characteristics.
2. Estimation: Predict the attribute of a data instance. Eg: estimate the percentage of marks of a student, whose previous marks are already known.
3. Prediction: Predictive model predicts a future outcome rather than the current behaviour. Eg: Predict next week's closing price for the Google share price per unit.
4. Market basket analysis(association rule mining)

Analyses hidden rules called association rule in a large transactional database.

{pen, pencil-> book} – whenever pen and pencil are purchased together, book is also purchased.

5. Clustering Classification into different classes based on some similarities but the target classes are unknown.
6. Business intelligence
7. Business data analytics
8. Bioinformatics
9. Web mining
10. Text mining
11. Social network data analysis

Data mining Functionalities

1. Class/Concept Description: Characterization and Discrimination: **Data characterization** is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query. For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries. For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.

2. Mining Frequent Patterns, Associations, and Correlations: **Frequent patterns**, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures. A *frequent itemset* typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers. A frequently occurring subsequence, such as the pattern that customers tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a *(frequent) sequential pattern*. A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a *(frequent) structured pattern*. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Association analysis:

$buys(X, \text{"computer"}) \Rightarrow buys(X, \text{"software"})$ [support = 1%, confidence = 50%],

A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% **support** means that 1% of all the transactions under analysis show that computer and software are purchased together. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as "*computer* *software* [1%, 50%]."

Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule**.

3. Classification and Regression for Predictive Analysis:

Classification is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

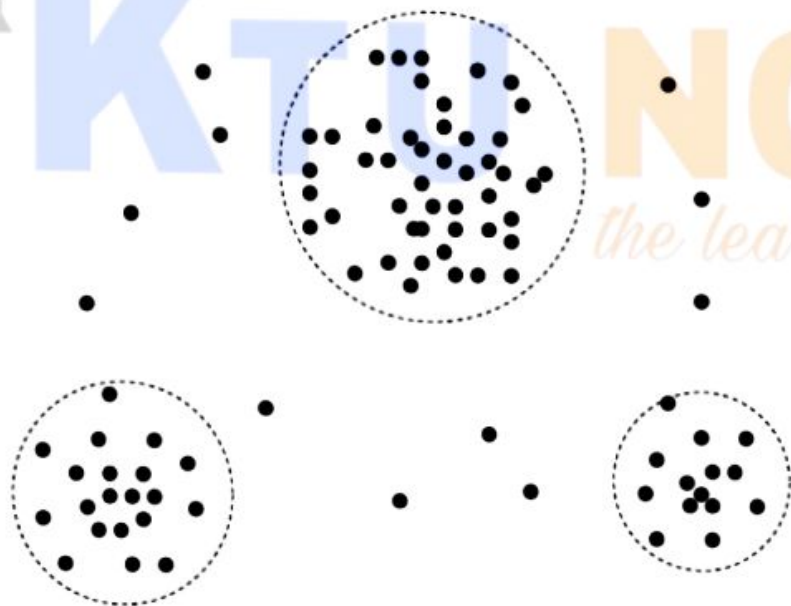
A **decision tree** is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network. A **neural network**, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and k -nearest-neighbor classification. Whereas classification predicts categorical (discrete, unordered) labels, **regression** models continuous-valued functions. That is, regression is used to predict missing or unavailable *numerical data values* rather than (discrete) class labels. The term *prediction* refers to both numeric prediction and class label prediction.

Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution *trends* based on the available data.

Unlike classification and regression, which analyze class-labeled (training) data sets,

4. Clustering analyzes data objects without consulting class labels. In many cases, class-labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity*. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate **taxonomy formation**, that is, the organization of observations into a hierarchy of classes that group similar events together.



6. Outlier Analysis

A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are **outliers**. Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more

interesting than the more regularly occurring ones. The analysis of outlier data is referred to as **outlier analysis** or **anomaly mining**.

Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

Are All Patterns Interesting?

A data mining system has the potential to generate thousands or even millions of patterns, or rules.

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*, and (4) *novel*. A pattern is also interesting if it validates a hypothesis that the user *sought to confirm*. An interesting pattern represents **knowledge**.

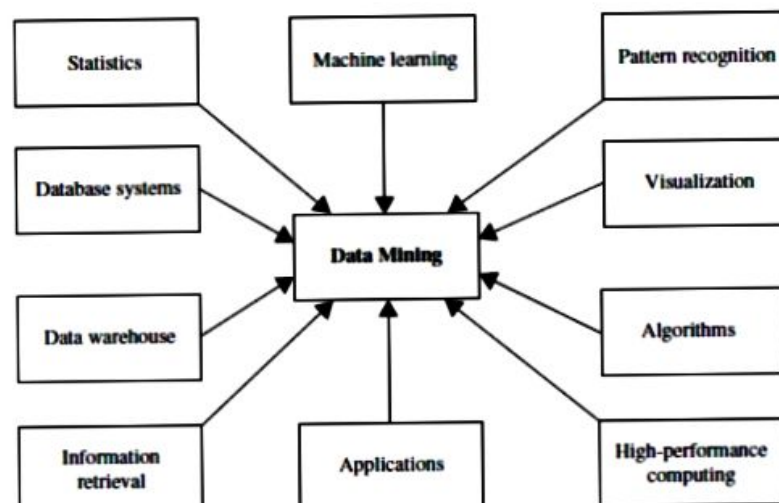
Several **objective measures of pattern interestingness** exist. These are based on the structure of discovered patterns and the statistics underlying them. An objective

measure for association rules of the form $X \Rightarrow Y$ is rule **support**, representing the percentage of transactions from a transaction database that the given rule satisfies. This is taken to be the probability $P(X \cup Y)$, where $X \cup Y$ indicates that a transaction contains both X and Y , that is, the union of itemsets 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 . More formally, support and confidence are defined as

$$\text{support}(X \Rightarrow Y) = P(X \cup Y),$$

$$\text{confidence}(X \Rightarrow Y) = P(Y|X).$$

Technologies for data mining



Statistics:

Statistics studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics. A **statistical model** is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes.

In other words, such statistical models can be the outcome of a data mining task. Alternatively, data mining tasks can be built on top of statistical models. For example, we can use statistics to model noise and missing data values. Then, when mining patterns in a large data set, the data mining process can use the model to help identify and handle noisy or missing values in the data. Statistics research develops tools for prediction and forecasting using data and statistical models. Statistical methods can be used to summarize or describe a collection of data.

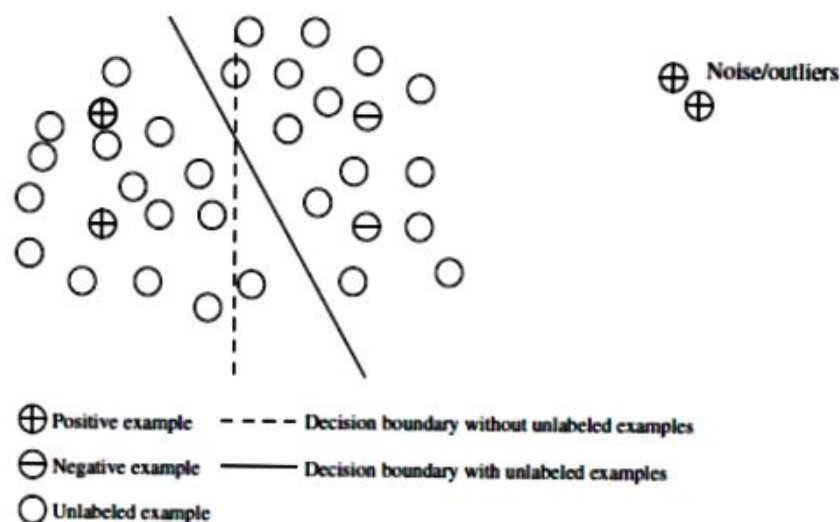
Statistical methods can also be used to verify data mining results. For example, after a classification or prediction model is mined, the model should be verified by statistical hypothesis testing. A **statistical hypothesis test** (sometimes called *confirmatory data analysis*) makes statistical decisions using experimental data. A result is called *statistically significant* if it is unlikely to have occurred by chance. If the classification or prediction model holds true, then the descriptive statistics of the model increases the soundness of the model.

Machine learning: investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to *automatically* learn to recognize complex patterns and make intelligent decisions based on data. For example, a typical machine learning problem is to program a computer so that it can automatically recognize handwritten postal codes on mail after learning from a set of examples. Machine learning is a fast-growing discipline. Here, we illustrate classic problems in machine learning that are highly related to data mining.

Supervised learning is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.

Unsupervised learning is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labeled. Typically, we may use clustering to discover classes within the data. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits. Suppose that it finds 10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to 9, respectively. However, since the training data are not labeled, the learned model cannot tell us the semantic meaning of the clusters found.

Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model. In one approach, labeled examples are used to learn class models and unlabeled examples are used to refine the boundaries between classes. For a two-class problem, we can think of the set of examples belonging to one class as the *positive examples* and those belonging to the other class as the *negative examples*. In Figure if we do not consider the unlabeled examples, the dashed line is the decision boundary that best partitions the positive examples from the negative examples. Using the unlabeled examples, we can refine the decision boundary to the solid line. Moreover, we can detect that the two positive examples at the top right corner, though labeled, are likely noise or outliers.



Active learning is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an example, which may be from a set of unlabeled examples or synthesized by the learning program. The goal is to optimize the model quality by actively acquiring knowledge from human users, given a constraint on how many examples they can be asked to label.

Major issues in data mining

1. Mining methodology and user interaction issues

● *Mining different kinds of knowledge in databases:*

Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery task. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

● *Interactive mining of knowledge at multiple levels of abstraction:*

Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

● *Incorporation of background knowledge:*

Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction

2. Mining methodology and user interaction issues

● *Pattern evaluation—the interestingness problem:*

A data mining system can uncover thousands of patterns. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measures or

user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

3. Performance issues

- *Efficiency and scalability of data mining algorithms:*

To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. The running time of a data mining algorithm must be predictable and acceptable in large databases.

- *Parallel, distributed, and incremental mining algorithms:*

The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

4. Issues relating to the diversity of database types:

- *Handling of relational and complex types of data:*

Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. Specific data mining systems should be constructed for mining specific kinds of data.

- *Mining information from heterogeneous databases and global information systems:*

Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining

