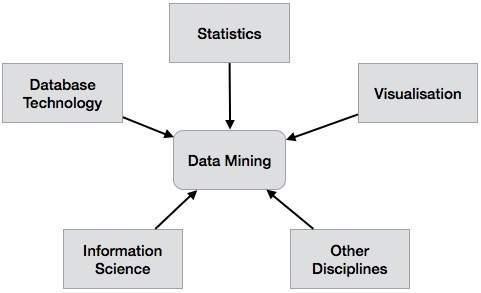
**CLASSIFICATION OF DATA MINING SYSTEMS**

There is a large variety of data mining systems available. Data mining systems may integrate techniques from the following −

* Spatial Data Analysis
* Information Retrieval
* Pattern Recognition
* Image Analysis
* Signal Processing
* Computer Graphics
* Web Technology
* Business
* Bioinformatics

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

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

Apart fraom these, a data mining system can also be classified based on the kind of (a) databases mined, (b) knowledge mined, (c) techniques utilized, and (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 primitives**

A data mining query is defined in terms of the following primitives

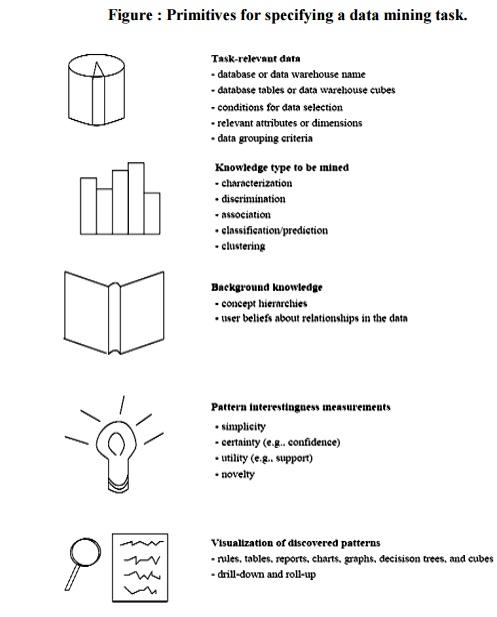
**Task-relevant data:** This is the database portion to be investigated. For example, suppose that you are a manager of All Electronics in charge of sales in the United States and Canada. In particular, you would like to study the buying trends of customers in Canada. Rather than mining on the entire database. These are referred to as relevant attributes

**The kinds of knowledge to be mined:** This specifies the data mining functions to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. For instance, if studying the buying habits of customers in Canada, you may choose to mine associations between customer profiles and the items that these customers like to buy

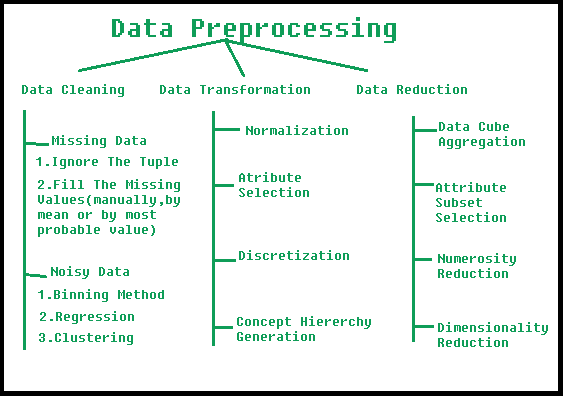
**Background knowledge**: Users can specify background knowledge, or knowledge about the domain to be mined. This knowledge is useful for guiding the knowledge discovery process, and for evaluating the patterns found. There are several kinds of background knowledge.

**Interestingness measures:** These functions are used to separate uninteresting patterns from knowledge. They may be used to guide the mining process, or after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures.

**Presentation and visualization of discovered patterns:** This refers to the form in which discovered patterns are to be displayed. Users can choose from different forms for knowledge presentation, such as rules, tables, charts, graphs, decision trees, and cubes.

****

**PREPROCESSING IN DATA MINING:**  
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



**steps Involved in Data Preprocessing:**

**1. Data Cleaning:**  
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

* **a). Missing Data:**  
  This situation arises when some data is missing in the data. It can be handled in various ways.  
  Some of them are:
  1. **Ignore the tuples:**  
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**  
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**  
  Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**  
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**  
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**  
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**  
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**  
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**  
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**  
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**  
   Here attributes are converted from level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**  
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**  
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**  
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**  
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**  
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Componenet Analysis).

**DISCRETIZATION AND CONCEPT HIERARCHY GENERALIZATION**

A **concept hierarchy** defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Consider a concept hierarchy for the dimension *location*. City values for *location* include Vancouver, Toronto, New York, and Chicago. Each city, however, can be mapped to the province or state to which it belongs. For example, Vancouver can be mapped to British Columbia, and Chicago to Illinois. The provinces and states can in turn be mapped to the country (e.g., Canada or the United States) to which they belong. These mappings form a concept hierarchy for the dimension *location*, mapping a set of low-level concepts (i.e., cities) to higher-level, more general concepts (i.e., countries). This concept hierarchy is illustrated in Figure 4.9.

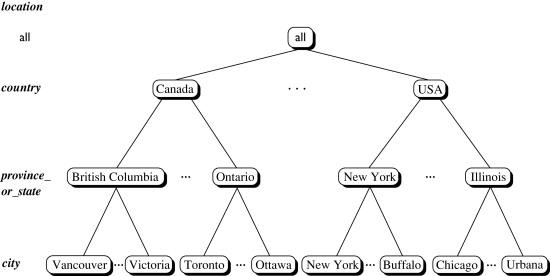


Figure 4.9. A concept hierarchy for *location*. Due to space limitations, not all of the hierarchy nodes are shown, indicated by ellipses between nodes.

Many concept hierarchies are implicit within the database schema. For example, suppose that the dimension *location* is described by the attributes *number, street, city, province\_or\_state, zip\_code*, and *country*. These attributes are related by a total order, forming a concept hierarchy such as *“street < city < province\_or\_state < country.”* This hierarchy is shown in Figure 4.10(a). Alternatively, the attributes of a dimension may be organized in a partial order, forming a lattice. An example of a partial order for the *time* dimension based on the attributes *day, week, month, quarter*, and *year* is *“day <*{*month < quarter; week*} < *year.”*1 This lattice structure is shown in Figure 4.10(b). A concept hierarchy that is a total or partial order among attributes in a database schema is called a **schema hierarchy**. Concept hierarchies that are common to many applications (e.g., *for time*) may be predefined in the [data mining system](https://www.sciencedirect.com/topics/computer-science/data-mining-system). [Data mining systems](https://www.sciencedirect.com/topics/computer-science/data-mining-system) should provide users with the flexibility to tailor predefined hierarchies according to their particular needs. For example, users may want to define a fiscal year starting on April 1 or an academic year starting on September 1.

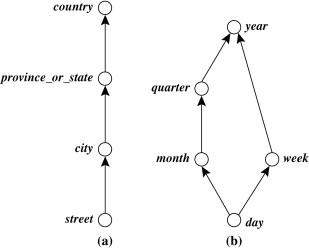


Figure 4.10. Hierarchical and lattice structures of attributes in warehouse dimensions: (a) a hierarchy for *location* and (b) a lattice for *time*.

Concept hierarchies may also be defined by discretizing or grouping values for a given dimension or attribute, resulting in a **set-grouping hierarchy**. A total or partial order can be defined among groups of values. An example of a set-grouping hierarchy is shown in Figure 4.11 for the dimension *price*, where an interval ($*X*…$*Y*] denotes the range from $*X* (exclusive) to $*Y* (inclusive).

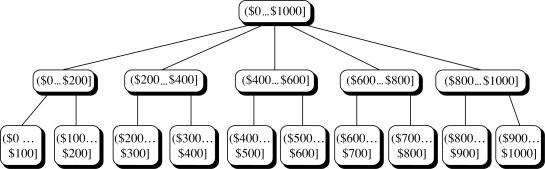


Figure 4.11. A concept hierarchy for *price*.

There may be more than one concept hierarchy for a given attribute or dimension, based on different user viewpoints. For instance, a user may prefer to organize *price* by defining ranges for *inexpensive, moderately\_priced*, and *expensive*.

Concept hierarchies may be provided manually by system users, domain experts, or knowledge engineers, or may be automatically generated based on statistical analysis of the data distribution.

# DATA GENERALIZATION IN DATA MINING - SUMMARIZATION BASED CHARACTERIZATION

From Data Analysis point of view, data mining can be classified into two categories: Descriptive mining and  predictive mining  
**Descriptive mining:** It describes the data set in a concise and summative manner and presents interesting general properties of data.  
**Predictive mining:** It analyzes the data to construct one or a set of models, and attempts to predict the behavior of new data sets.

Databases usually stores a large amount of data in great detail. However, users often like to view sets of summarized data in concise, descriptive terms.

Such data descriptions may provide an overall picture of a class of data or distinguish it from a set of comparative classes.

Such descriptive data mining is called concept descriptions and forms an important component of data mining.

## What Is Concept Description

The simplest kind of descriptive data mining is called concept description. A concept usually refers to a collection of data such as frequent\_buyers, graduate\_students and so on.

As data mining task concept description is not a simple enumeration of the data. Instead, concept description generates descriptions for characterization and comparison of the data.

It is sometimes called class description when the concept to be described refers to a class of objects

* **Characterization**: It provides a concise and succinct summarization of the given collection of data.
* **Comparison**: It provides descriptions comparing two or more collections of data.

## Data Generalization & Summarization

Data and objects in databases contain detailed information at primitive concept level.  
For example, the item relation in a sales database may contain attributes describing low-level item information such as item\_ID, name, brand, category, supplier, place\_made and price.  
It is useful to be able to summarize a large set of data and present it at a high conceptual level.  
For example, summarizing a large set of items relating to Christmas season sales provides a general description of such data, which can be very helpful for sales and marketing managers.  
This requires an important functionality called data generalization.

## Data Generalization

A process that abstracts a large set of task-relevant data in a database from a low conceptual level to higher ones.

Data Generalization is a summarization of general features of objects in a target class and produces what is called characteristic rules.

The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

For example, one may want to characterize the "OurVideoStore" customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the **attribute-oriented induction** method can be used, for example, to carry out data summarization.

Note that with a data cube containing a summarization of data, simple OLAP operations fit the purpose of data characterization.  
**Approaches:**

* Data cube approach(OLAP approach).
* Attribute-oriented induction approach.

## Presentation Of Generalized Results

**Generalized Relation:**

* Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.

**Cross-Tabulation:**

* Mapping results into cross-tabulation form (similar to contingency tables).

**Visualization Techniques:**

* Pie charts, bar charts, curves, cubes, and other visual forms.

**Quantitative characteristic rules:**

* Mapping generalized results in characteristic rules with quantitative information associated with it.

## Data Cube Approach

It is nothing but performing computations and storing results in data cubes.  
**Strength**

* An efficient implementation of data generalization.
* Computation of various kinds of measures, e.g., count( ), sum( ), average( ), max( ).
* Generalization and specialization can be performed on a data cube by roll-up and drill-down.

**Limitations**

* It handles only dimensions of simple non-numeric data and measures of simple aggregated numeric values.
* Lack of intelligent analysis, can’t tell which dimensions should be used and what levels should the generalization reach.