What Kinds of Data Can Be Mined?

The most basic forms of data for mining applications are:

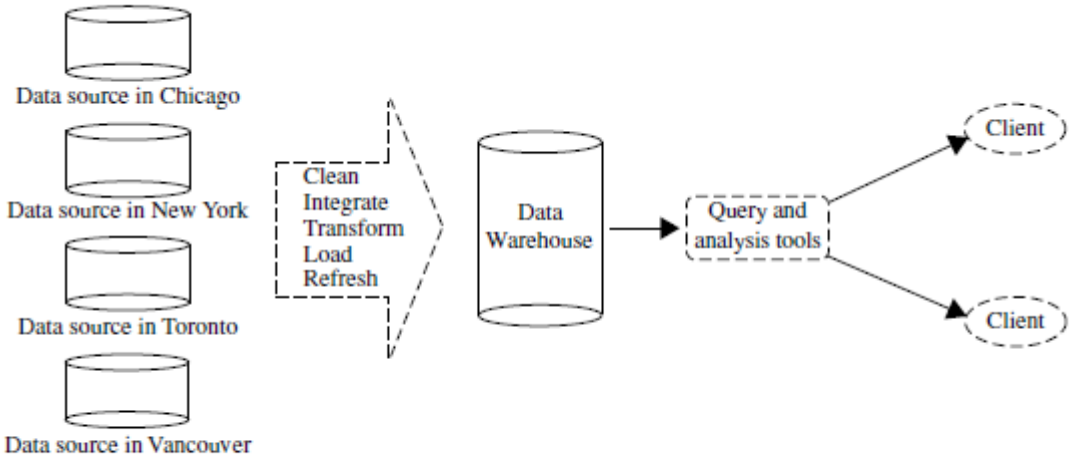
1. Database data
2. Data warehouse data
3. Transactional data
4. The Data mining can also be applied to other forms of data (e.g. data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW).

1. Database Data

- A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (*columns* or *fields*) and usually stores a large set of tuples (*records* or *rows*). Each tuple in a relational table represents an object identified by a unique *key* and described by a set of attribute values. A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and
  their relationships.
- Relational data can be accessed by database queries written in a relational query language (e.g., SQL) or with the assistance of graphical user interfaces. A given query is transformed into a set of relational operations
- When mining relational databases, we can go further by *searching for trends* or *data patterns*. For example: data mining systems can analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information.
- Data mining systems may also detect deviations
- Relational databases are one of the most commonly available and richest information repositories, and thus they are a major data form in the study of data mining.

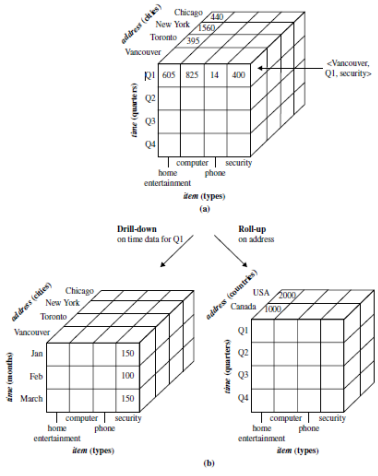
2. Data Warehouses

- 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.
- 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 precomputation and fast access of summarized data.



Typical framework of a data warehouse for *AllElectronics*.

- By providing multidimensional data views and the precomputation 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
- 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



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.

3. Transactional Data

- Each record in a **transactional database** captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page. A transaction typically includes a unique transaction identity number (*trans ID*) and a list of the **items** making up the transaction, such as the items purchased in the transaction.
- A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on

trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Figure 2: Fragment of a transactional database for sales at *AllElectronics*.

4. Other Kinds of Data

- Besides relational database data, data warehouse data, and transaction data, there are many other kinds of data that have versatile forms and structures and rather different semantic meanings

Other kinds of data can be seen in many applications:

- Time-related or sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data)
- Data streams (e.g., video surveillance and sensor data, which are continuously transmitted),
- Spatial data (e.g., maps),
- Engineering design data (e.g., the design of buildings, system components, or integrated circuits),
- Hypertext and multimedia data (including text, image, video, and audio data),
- Graph and networked data (e.g., social and information networks),
- Web (a huge, widely distributed information repository made available by the Internet).

- These applications bring about new challenges, like how to handle data carrying special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity), and how to mine patterns that carry rich structures and semantics.
- Various kinds of knowledge can be mined from these kinds of data

Example: Regarding temporal data, for instance, we can mine banking data for changing trends, which may aid in the *scheduling of bank tellers* according to the volume of customer traffic. Stock exchange data can be mined to uncover trends that could help you plan investment strategies

5. What Kinds of Patterns Can Be Mined?

*There are a number of **data mining functionalities are:***

- Characterization and discrimination (**Class/Concept Description)**
- The mining of frequent patterns, associations, and correlations
- Clustering analysis
- Outlier analysis

- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks
- Such tasks can be classified into two categories: **descriptive** and **predictive**.
- Descriptive mining tasks characterize properties of the data in a target data set.
- Predictive mining tasks perform induction on the current data in order to make predictions.

Class/Concept Description: Characterization and Discrimination

- Data entries can be associated with classes or concepts
  For example: In the *AllElectronics* store, classes of items for sale include *computers* and *printers*, and concepts of customers include *big Spenders* and *budget Spenders*.
- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms.
- Such descriptions of a class or a concept are called **class/concept descriptions**. These descriptions can be derived using :

  (1) *Data characterization*, by summarizing the data of the class under study (often called the **target class**) in general terms, or
  (2) *Data discrimination*, by comparison of the target class with one or a set of comparative classes (often called the **contrasting classes**) or
  (3) Both data characterization and discrimination. **classes**)

- **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
- **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

Mining Frequent Patterns, Associations, and Correlations

- **Frequent patterns** 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
- 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.

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 are derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known).
- The classification model is used to predict the class label of objects for which the the class label is unknown.
- The derived model may be represented in various forms, such as *classification rules* (i.e., *IF-THEN rules*), *decision trees*, *mathematical formulae*, or *neural networks*

- 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. Decision trees can easily be converted to classification rules.
- 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 classificationmodels, such as na¨ıve Bayesian classification, support vector machines, and *k*-nearest-neighbor classification
- Classification predicts categorical (discrete, unordered) labels
- **Regression** models continuous-valued functions.
- 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.
- Classification and regression may need to be preceded by **relevance analysis**, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process.
- Other attributes, which are irrelevant, can then be excluded from consideration

Cluster Analysis

- **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 (i. e.* 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.

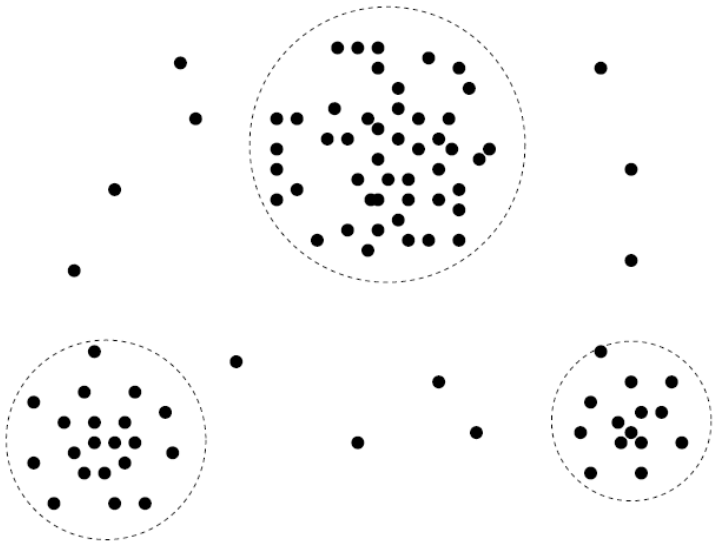


Figure 1 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

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
- The analysis of outlier data is referred to as **outlier analysis** or **anomaly mining**.
- Outliers may be detected using statistical tests that assume a distribution or probability model for the data or using distance measures where objects that are remote from any other cluster are considered outliers