DMDW

UNIT-1

I. What Motivated Data Mining? Why Is It Important?

- The Explosive Growth of Data: from terabytes to petabytes
    - Data collection and data availability
        - Automated data collection tools, database systems, Web, computerized society
    - Major sources of abundant data
        - Business: Web, e-commerce, transactions, stocks, …
        - Science: Remote sensing, bioinformatics, scientific simulation, …
        - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

- In recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

    - For applications ranging from market analysis, fraud detection, and customer retention, to production control

- Data mining can be viewed as a result of the natural evolution of information technology.
    - The database system industry has witnessed an evolutionary path in the development of the following functionalities: *data collection and database creation, data management* (including data storage and retrieval, and database transaction processing), and *advanced data analysis* (involving data warehousing and data mining).

- The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful and affordable computers, data collection equipment, and storage media.

- Data can now be stored in many different kinds of databases and information repositories. One data repository architecture that has emerged is the data warehouse , a repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to *facilitate management decision making.*
    - On-line analytical processing (OLAP) that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation as well as the ability to view information from different angles.
    - Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data changes over time

- The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation.

Evolution of Database Technology

- 1960s and earlier:
  Data Collection and Database Creation
    - Primitive file processing

- 1970s - early 1980s:
  Data Base Management Systems
  - Hieratical and network database systems
  - Relational database Systems
  - Query languages: SQL
  - Transactions, concurrency control and recovery.
  - On-line transaction processing (OLTP)

- Mid -1980s - present:
  - Advanced data models
    - Extended relational, object-relational
  - Advanced application-oriented DBMS
    - spatial, scientific, engineering, temporal, multimedia, active, stream and sensor, knowledge-based
- Late 1980s-present
  - Advanced Data Analysis
    - Data warehouse and OLAP
    - Data mining and knowledge discovery
    - Advanced data mining appliations
    - Data mining and socity
- 1990s-present:
  - XML-based database systems
  - Integration with information retrieval
  - Data and information integreation
- Present – future:
  - New generation of integrated data and information system.

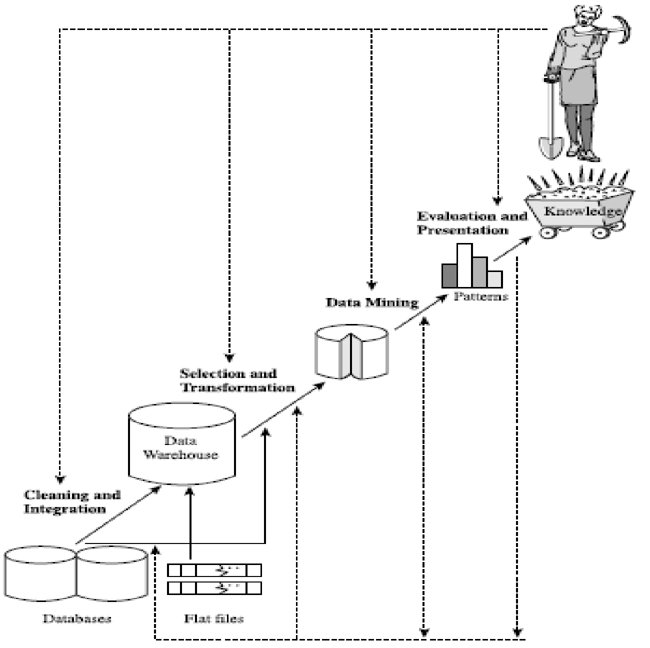
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Data mining refers to extracting or mining knowledge from large amounts of data.
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Knowledge Discovery (KDD) Process

Knowledge discovery as a process is depicted in Figure and consists of 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 or consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)



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

This is a view from typical machine learning and statistics communities

Architecture of a Typical Data Mining System

The architecture of a typical data mining system may have the following major components

- Database, data warehouse, WorldWideWeb, or other information repository: Data cleaning and data integration techniques may be performed on the data.

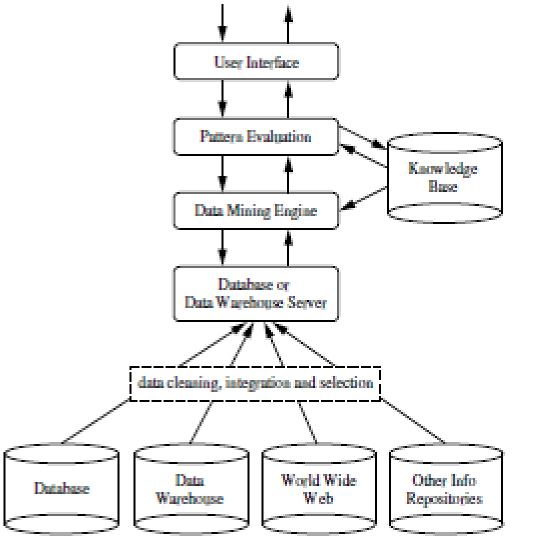
- Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request

- Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.(concept hierarchies, user beliefs, and additional interestingness constraints or thresholds)

- Data mining engine: This is essential to the data mining systemand ideally consists of a set of functional modules for tasks such as characterization, association etc.

- Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns.

- User interface: Thismodule communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query



Architecture of a typical data mining system.

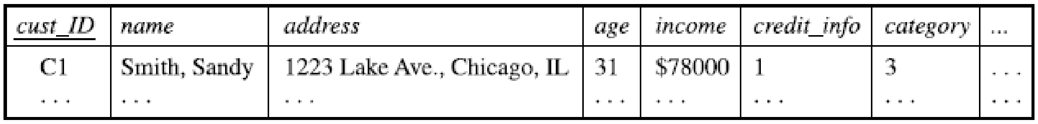
II. Data Mining—On What Kind of Data?

- In principle, data mining should be applicable to any kind of data repository.
- Data repositories include relational databases, data warehouses, transactional databases, advanced database systems, flat files, data streams, and the World Wide Web.
- Advanced database systems include object-relational databases and specific application-oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases. The challenges and techniques of mining may differ for each of the repository systems.

1. **Relational databases** (A relational database is a collection of tables)
   Eg. A relational database for *AllElectronics*. The *AllElectronics* company is described by the following relation tables: *customer, item, employee*, and *branch*.

   - 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.
   - The software programs involve mechanisms for the definition of database structures; for data storage; for concurrent, shared, or distributed data access; and for ensuring the consistency and security of the information stored, despite system crashes or attempts at unauthorized access.

customer



cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1 ...	Smith, Sandy ...	1223 Lake Ave., Chicago, IL ...	31 ...	$78000 ...	1 ...	3 ...	... ...

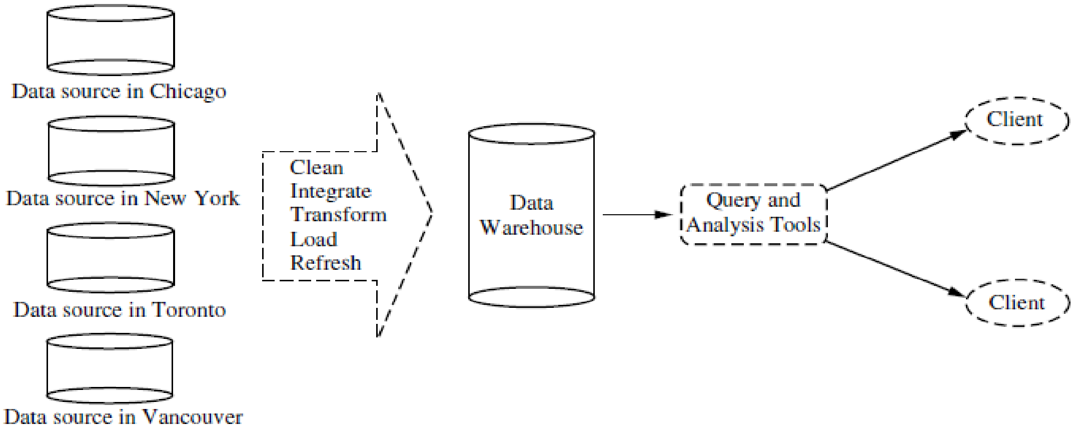
2. **Data warehouses**

   - Suppose that *AllElectronics* is a successful international company, with branches around the world. *Each branch has its own set of databases.*
   - The president of *AllElectronics* 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 *AllElectronics* had a data warehouse, this task would be easy.

Definition: A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.

Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

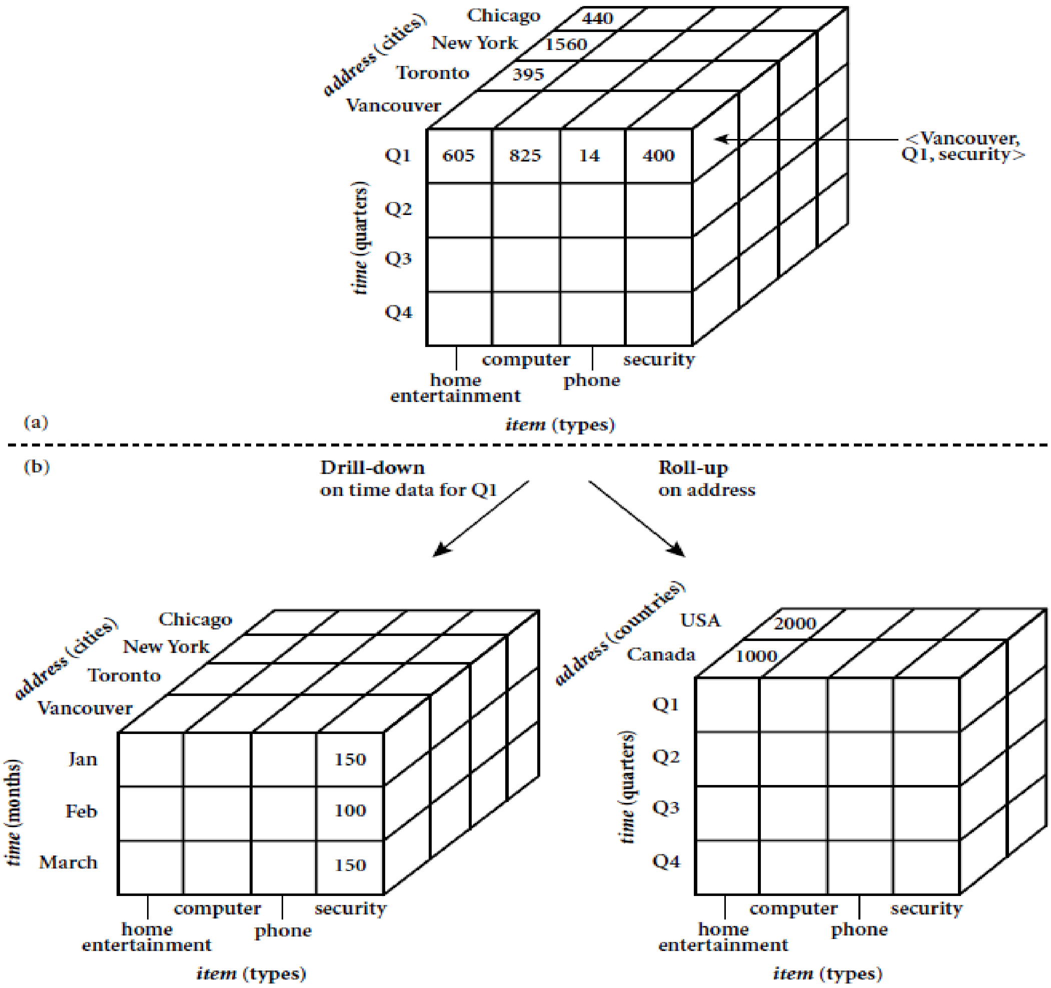


Typical framework of a data warehouse for *AllElectronics*.

- To facilitate decision making, the data in a data warehouse are *organized around major subjects*, such as customer, item, supplier, and activity.
- The data are stored to provide information from a *historical perspective* (such as from the past 5–10 years) 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 database structure, where 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 *sales amount*.
- The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube.
- A data cube provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data.

Eg. A data cube for *AllElectronics*.
A data cube for summarized sales data of *AllElectronics* 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 itemtype 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 items relating to security systems in Vancouver is $400,000, as stored in cell <*Vancouver, Q1, security*>



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.

What is the difference between a data warehouse and a data mart?

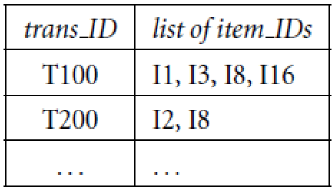
- A data warehouse collects information about subjects that span an *entire organization*, and thus its scope is *enterprise-wide*.
- A data mart, on the other hand, is a department subset of a data warehouse. It focuses on selected subjects, and thus its scope is *department-wide*.

By providing multidimensional data views and the precomputation of summarized data, data warehouse systems are well suited for on-line analytical processing, or OLAP.

Although data warehouse tools help support data analysis, additional tools for data mining are required to allow more in-depth and automated analysis.

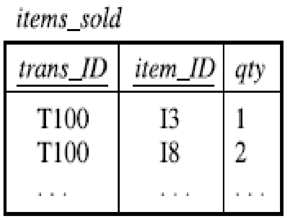
3. Transactional databases

- In general, a transactional database consists of a file where each record represents a transaction.
- A transaction typically includes a unique transaction identity number (*trans ID*) and a list of the items



trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Fragment of a transactional database for sales at *AllElectronics*.



items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

- Suppose you would like to dig deeper into the data by asking, "Which items sold well together?"
- This kind of *market basket data analysis* would enable you to bundle groups of items together as a strategy for maximizing sales.
- For example, given the knowledge that printers are commonly purchased together with computers, you could offer an expensive model of printers at a discount to customers buying selected computers, in the hopes of selling more of the expensive printers.
- *A regular data retrieval system is not able to answer queries like the one above.*
- However, data mining systems for transactional data can do so by identifying *frequent itemsets*, that is, sets of items that are frequently sold together.

4. Advanced Data and Information Systems and Advanced Applications

- The new database applications include handling
- Spatial data (such as maps),
- Engineering design data (such as the design of buildings, system components, or integrated circuits),
- Hypertext and multimedia data (including text, image, video, and audio data),
- Time-related data (such as historical records or stock exchange data),
- Stream data (such as video surveillance and sensor data, where data flow in and out like streams),
- WorldWideWeb (a huge, widely distributed information repository made available by the Internet).

These applications require efficient data structures and scalable methods for handling complex object structures; variable-length records; semistructured or unstructured data; text, spatiotemporal, and multimedia data; and database schemas with complex structures and dynamic changes.

a) Object-Relational Databases

- Object-relational databases are constructed based on an object-relational data model.
- Conceptually, the object-relational data model inherits the essential concepts of object-oriented databases, where, in general terms, each entity is considered as an object.

- A set of variables that describe the objects
- Objects that share a common set of properties can be grouped into an object class.
- Each object is an instance of its class. Object classes can be organized into class/subclass hierarchies so that each class represents properties that are common to objects in that class.
- For data mining in object-relational systems, techniques need to be developed for handling complex object structures, complex data types, class and subclass hierarchies, property inheritance, and methods and procedures.

b) **Temporal Databases, Sequence Databases, and Time-Series Databases**

- A temporal database typically stores relational data that include time-related attributes. These attributes may involve several timestamps, each having different semantics.
- A sequence database stores sequences of ordered events, with or without a concrete notion of time. Examples include customer shopping sequences,
- A time-series database stores sequences of values or events obtained over repeated measurements of time (e.g., hourly, daily, weekly). Examples include data collected from the stock exchange, inventory control, and the observation of natural phenomena (like temperature and wind).
- Data mining techniques can be used to find the characteristics of object evolution, or the trend of changes for objects in the database. Such information can be useful in decision making and strategy planning.

c) **Spatial Databases and Spatiotemporal Databases**

- Spatial databases contain spatial-related information. Examples include geographic (map) databases, very large-scale integration (VLSI) or computed-aided design databases, and medical and satellite image databases.
- Geographic databases have numerous applications, ranging from forestry and ecology planning to providing public service information regarding the location of telephone and electric cables, pipes, and sewage systems. In addition, geographic databases are commonly used in vehicle navigation
- A spatial database that stores spatial objects that change with time is called a spatiotemporal database, from which interesting information can be mined. For example, we may be able to group the trends of moving objects and identify some strangely moving vehicles

d) **Text Databases and Multimedia Databases**

- Text databases are databases that contain word descriptions for objects. such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.
- Multimedia databases store image, audio, and video data. They are used in applications such as picture content-based retrieval, voice-mail systems and recognize spoken commands

e) **Heterogeneous Databases and Legacy Databases**

- A heterogeneous database consists of a set of interconnected, autonomous component databases.
- Many enterprises acquire legacy databases as a result of the long history of information technology development
- A legacy database is a group of *heterogeneous databases* that combines different kinds of data systems, such as relational or object-oriented databases, hierarchical databases, network databases, spreadsheets, multimedia databases, or file systems.

f) **Data Streams**

- Many applications involve the generation and analysis of a newkind of data, called stream data, where data flow in and out of an observation platform (or window) dynamically

- Typical examples of data streams include various kinds of scientific and engineering data, time-series data, and data produced in other dynamic environments, such as power supply, network traffic, stock exchange, telecommunications, Web click streams, video surveillance, and weather or environment monitoring.

g) **The World Wide Web**

- The World Wide Web and its associated distributed information services, such as Yahoo!, Google, America Online, and AltaVista, provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access.
- Users seeking information of interest traverse from one object via links to another.

III. Data Mining Functionalities

There are a number of ***data mining functionalities***. These include

- Characterization and Discrimination
- Classification and Regression
- Associations analysis
- Clustering analysis
- Outlier analysis
- Trend and evolution analysis

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 *bigSpenders* and *budgetSpenders*. 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.

Data characterization is a summarization of the general characteristics or features of a target class of data.

Eg. A data mining system should be able to produce a description summarizing the characteristics of customers who spend more than $1,000 a year at AllElectronics. The result could be a general profile of the customers, such as they are 40–50 years old, employed, and have excellent credit ratings. The system should allow users to drill down on any dimension, such as on occupation in order to view these customers according to their type of employment.

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.

Eg. A data mining system should be able to compare two groups of AllElectronics customers, such as those who shop for computer products regularly (more than two times a month) versus those who rarely shop for such products (i.e., less than three times a year). The resulting description provides a general comparative profile of the customers, such as 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree. Drilling down on a dimension, such as occupation, or adding new dimensions, such as income level, may help in finding even more discriminative features between the two classes.

Classification and Prediction

- Finding models (functions) that describe and distinguish data classes or concepts for predict the class whose label is unknown
- E.g., classify countries based on climate, or classify cars based on gas mileage
- Models: decision-tree, classification rules (if-then), neural network
- Prediction: Predict some unknown or missing numerical values

Association Analysis

- 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.
- Multi-dimensional vs. single-dimensional association
- age(X, “20..29”) ^ income(X, “20..29K”) => buys(X, “PC”) [support = 2%, confidence = 60%]
- contains(T, “computer”) => contains(x, “software”) [support=1%, confidence=75%]

Cluster analysis

- Analyze class-labeled data objects, clustering analyze data objects without consulting a known class label.
- Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

Outlier_analysis

- Outlier: a data object that does not comply with the general behavior of the model of the data
- It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

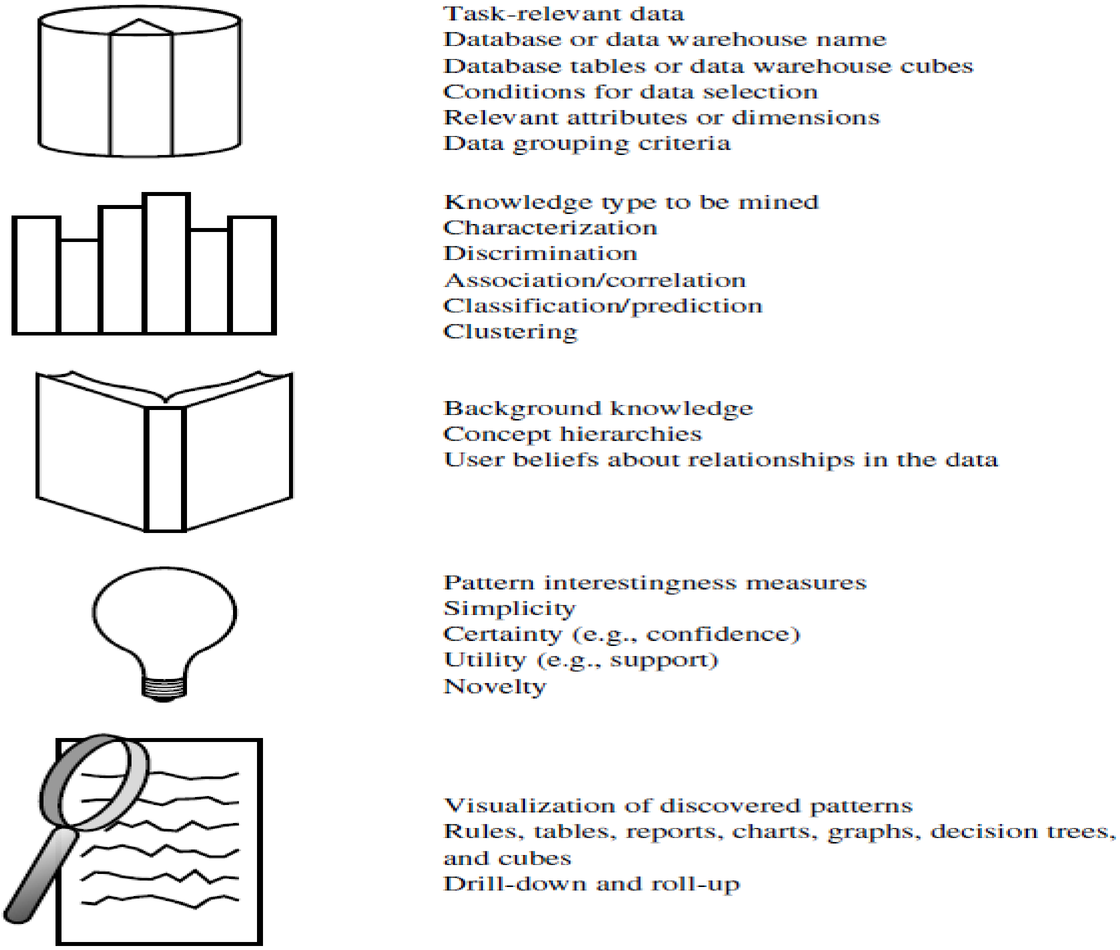
Trend and evolution analysis

- Trend and deviation: regression analysis
- Sequential pattern mining, periodicity analysis
- Similarity-based analysis

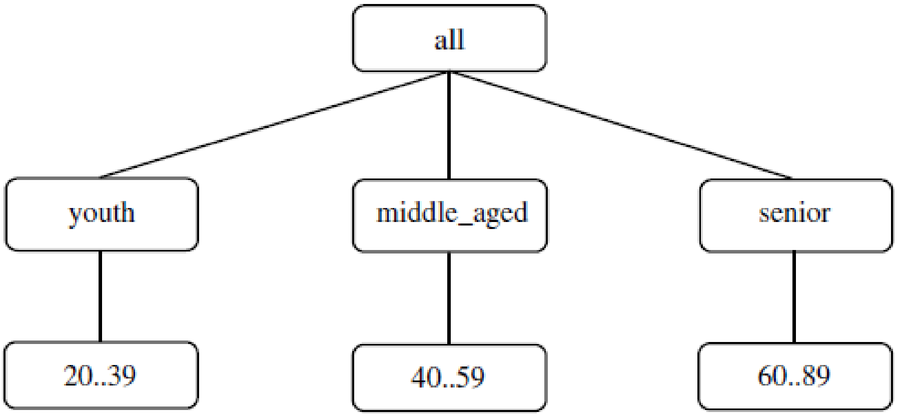
V. Data Mining Task Primitives

Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed. 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. The data mining primitives specify the following, as illustrated in Figure


Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria
Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering
Background knowledge
Concept hierarchies
User beliefs about relationships in the data
Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty
Visualization of discovered patterns
Rules, tables, reports, charts, graphs, decision trees, and cubes
Drill-down and roll-up

Primitives for specifying a data mining task.



A concept hierarchy for the attribute (or dimension) *age*. The root node represents the most general abstraction level, denoted as all.

Example : Mining classification rules. Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than $40,000, and who have bought more than $1,000 worth of items, each of which is priced at no

less than $100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules.
This data mining query is expressed in DMQL3 as follows, where each line of the query has been enumerated to aid in our discussion.

```
(1) use database AllElectronics_db
(2) use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
(3) mine classification as promising_customers
(4) in relevance to C.age, C.income, I.type, I.place_made, T.branch
(5) from customer C, item I, transaction T
(6) where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
        and C.income >=40,000 and I.price >=100
(7) group by T.cust_ID
(8) having sum(I.price) >=1,000
(9) display as rules
```

VI. Are All of the Patterns Interesting?

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

Are all of the patterns interesting? Typically not—only a small fraction of the patterns potentially generated would actually be of interest to any given user.

What makes a pattern interesting?
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 =>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 U Y),where X U 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

support(X =>Y) = P(X U Y):
confidence(X =>Y) = P(Y|X):

In general, each interestingness measure is associated with a threshold, which may be controlled by the user. For example, rules that do not satisfy a confidence threshold of, say, 50% can be considered uninteresting. Rules below the threshold likely reflect noise, exceptions, or minority cases and are probably of less value.

Can a data mining system generate all of the interesting patterns?—refers to the completeness of a data mining algorithm. It is often unrealistic and inefficient for data mining systems to generate all of the possible patterns. Instead, user-provided constraints and interestingness measures should be used to focus the search.

Can a data mining system generate only interesting patterns?— is an optimization problem in data mining. It is highly desirable for data mining systems to generate only interesting patterns. This would be much more efficient for users and data mining systems, because neither would have to search through the patterns generated in order to identify the truly interesting ones. Progress has been made in
this direction; however, such optimization remains a challenging issue in data mining

VII. Classification of Data Mining Systems

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science.

1. ***Classification according to the kinds of databases mined:*** data mining system can be classified according to the kinds of databases mined. For instance, if classifying according to data models, we may have a relational, transactional, object-relational, or data warehouse mining system.

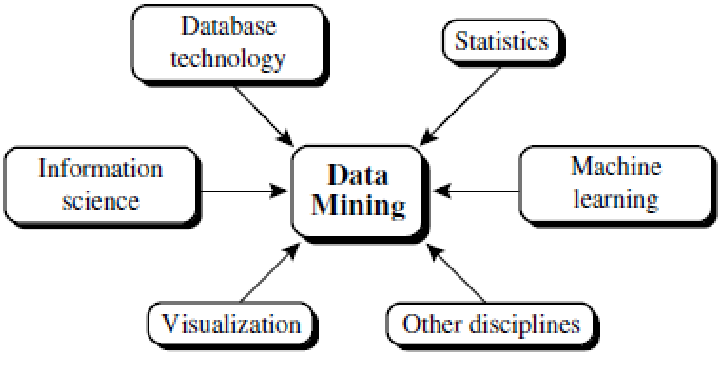
2. ***Classification according to the kinds of knowledge mined:*** Data mining systems can be categorized according to the kinds of knowledge they mine, that is, based on data mining functionalities, such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.

   Moreover, data mining systems can be distinguished based on the granularity or levels of abstraction of the knowledge mined, including generalized knowledge (at a highlevel of abstraction),primitive-level knowledge (at a rawdata level), or knowledge atmultiple levels (considering several levels of abstraction).

3. ***Classification according to the kinds of techniques utilized:*** Data mining systems can be categorized according to the underlying data mining techniques employed.

   These techniques can be described according to the *degree of user interaction involved* (e.g., autonomous systems, interactive exploratory systems, query-driven systems) or the ***methods of data analysis employed*** (e.g., database-oriented or data warehouse oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks, and so on).

4. ***Classification according to the applications adapted:*** Data mining systems can also be categorized according to the applications they adapt. For example, data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and so on.



Data mining as a confluence of multiple disciplines.

VIII. Major Issues in Data Mining

The Major issues in data mining regarding mining methodology, user interaction, performance, and diverse data types.

Mining methodology and user interaction issues:

1. ***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 tasks, including association, classification etc.

2. ***Interactive mining of knowledge at multiple levels of abstraction:*** Because it is difficult to know exactly what can be discovered within a database, the data mining process should be *interactive*. The user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

3. ***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.

4. ***Data mining query languages and ad hoc data mining:*** Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval.

5. ***Presentation and visualization of data mining results:*** Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.

6. ***Handling noisy or incomplete data:*** The data stored in a database may reflect noise, exceptional cases, or incomplete data objects.

7. ***Pattern evaluation—the interestingness problem:*** A data mining systemcan uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty.

Performance issues: These include efficiency, scalability, and parallelization of data mining algorithms.

1. ***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.

2. ***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.

Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "from scratch." Such algorithms perform knowledge modification incrementally to amend and strengthen what was previously discovered.

Issues relating to the diversity of database types:

1. ***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.

   It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data. Therefore, one may expect to have different data mining systems for different kinds of data.

2. ***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.