Unit 1

Introduction

Data, Information, and Knowledge

- **Data:**

  Data are any facts, numbers, or text that can be processed by a computer.

  Today organizations are accumulating vast and growing amounts of data in different formats and databases.

Data, Information, and Knowledge

Types of data may include

- Operational or Transactional data such as sales, cost, inventory (simply says what is happening)
- Payroll, and accounting.
- Non operational data like industry sales, forecast data, and macroeconomic data.(simply says why things are happening)
- Metadata

  Data about the data itself such as logical database design or data dictionary definitions.

Data, Information, and Knowledge

- **Information:**

  The patterns, associations, or relationships among the data can provide information.

  For example, analysis of retail point-of-sale transaction data can yield information on which products are selling and when.

Data, Information, and Knowledge

- **Knowledge:**

  Information can be converted into knowledge about historical patterns and future trends.

  For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge or consumer buying behavior.

Data mining: What, Who, Why, How, Where?

- What is data mining?
  - Use historical (large-scale) data to uncover regularities and improve future decisions
  - Everybody has some data:
    - Science: physics, chemistry, biology
    - Health care: patients, diseases, images
    - Business: sales, marketing
    - Internet: web

Introduction: What is Data Mining?

- "The process of discovering meaningful patterns and trends often previously unknown by using some mathematical algorithm on huge amount of stored data"
- "Extraction of interesting, significant, implicit, previously unknown and potentially useful information or patterns from data in large database."

Introduction: What is Data Mining?

- Data mining is basically concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.
- Hence, data mining is **Knowledge Discovery from Database.**

Data mining: What, Who, Why, How, Where?

- WHO is doing data mining?

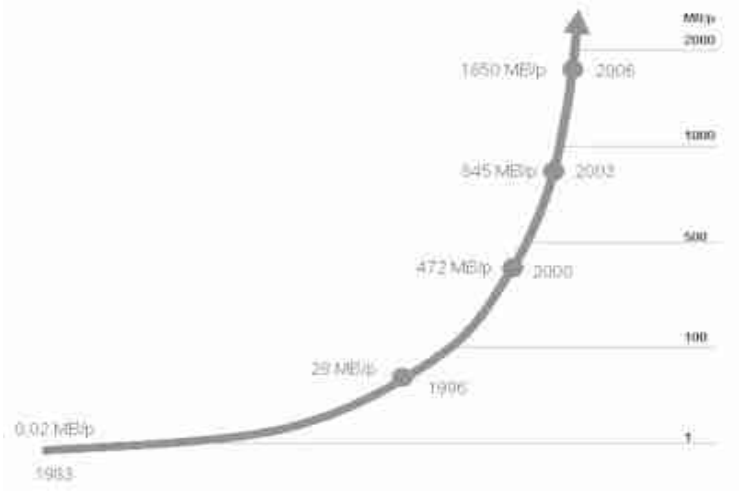
  retailers, financial intuitions, communication, and marketing organizations

Data mining: What, Who, Why, How, Where?

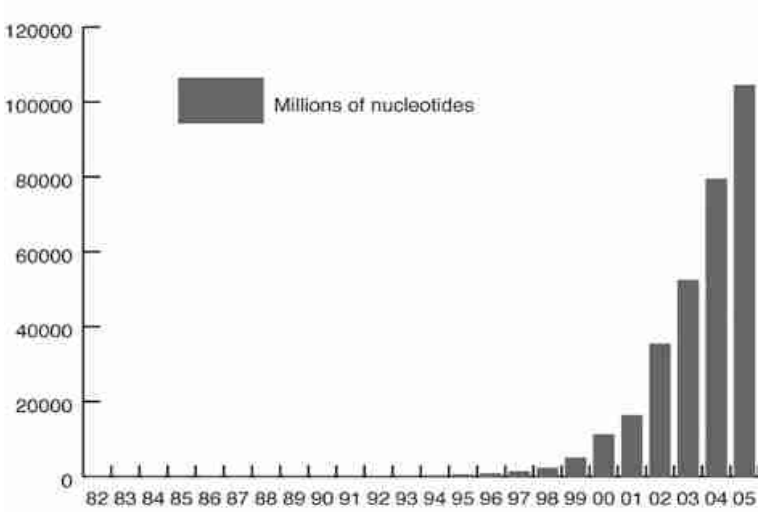
- Why data mining/What Motivated Data Mining?

  A. Information explosion:

  Data → Knowledge/Decision/Understanding/Profit



Personal Information storage



Exponential growth of database

Data mining: What, Who, Why, How, Where?

B. 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 data
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bio-informatics, scientific simulation, …
  - Society and everyone: news, digital cameras, YouTube…

C. Information Crisis

- We are drowning in data, but starving for knowledge.

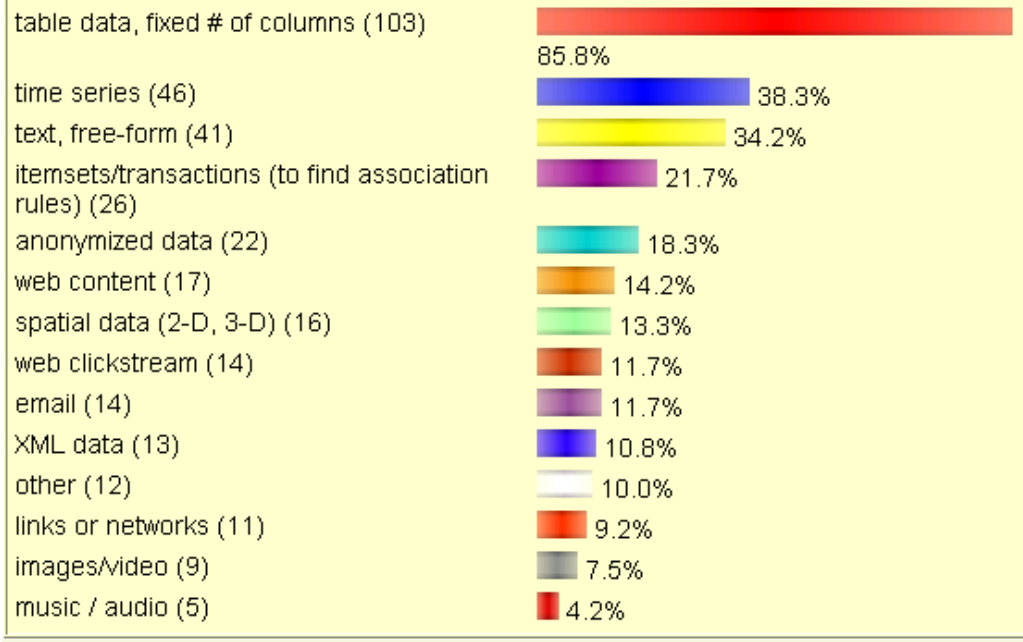
Data mining: What, Who, Why, How, Where?

1. **Collect Your Data**

KDnuggets : **Polls** : Data types analyzed (July 2007)

Poll

Types of Data Analyzed/Mined in the past 12 months: [120 voters total]



Data mining: What, Who, Why, How, Where?

2. Determining the patterns you want to mine: data mining tasks

- Two main types of tasks

  - **Prediction Methods**
    Use some variables to predict unknown or future values of other variables.

  - **Description Methods**
    Find human interpretative patterns/rules that describe the data.

Data mining: What, Who, Why, How, Where?

Basis for Comparison	Descriptive	Predictive
Describes	What happened in the past? By using the stored data.	What might happen in the future? By using the past data and analyzing it.
Process Involved	Involves Data Aggregation.	Involves Statistics and forecast techniques.
Definition	The process of finding useful and important information by analyzing the huge data.	This process involves in forecasting the future of the company, which are very useful.

Data mining: What, Who, Why, How, Where?

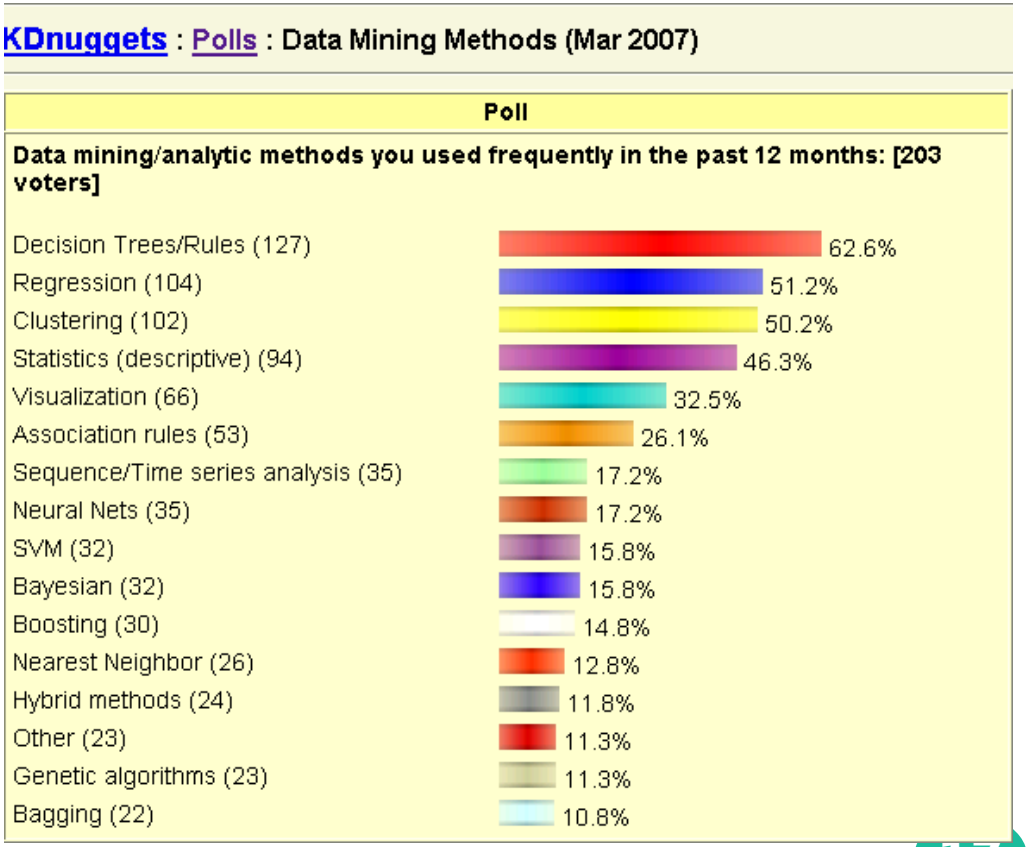
Basis for Comparison	Descriptive	Predictive
Data Volume	It involves in processing huge data that are stored in data warehouses. Limited to past data.	It involves analyzing large past data and then predicts the future using advance techniques.
Examples	Sales report, revenue of a company, performance analysis, etc.	Sentimental analysis, credit score analysis, forecast reports for a company, etc.
Accuracy	It provides accurate data in the reports using past data.	Results are not accurate, it will not tell exactly what will happen
Approach	It allows the reactive approach(responding to events after they have happened)	This a proactive approach (eliminating problems before they have a chance to appear)

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Regression [Predictive]
- Association Rule [Descriptive]
- Sequential Pattern [Descriptive]
- Deviation Detection [Predictive]

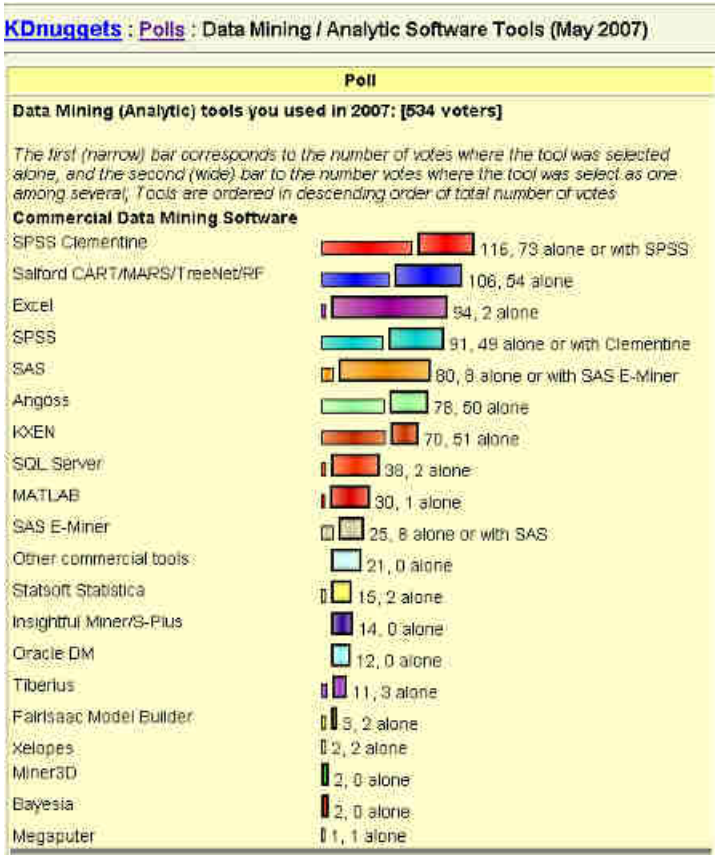
Data mining: What, Who, Why, How, Where?

3. Choose the algorithm(s)



Data mining: What, Who, Why, How, Where?

4. Select existing data mining software/packages



KDnuggets : Polls : Data Mining / Analytic Software Tools (May 2007)

Poll

Data Mining (Analytic) tools you used in 2007: [534 voters]

The first (narrow) bar corresponds to the number of votes where the tool was selected alone, and the second (wide) bar to the number votes where the tool was select as one among several; Tools are ordered in descending order of total number of votes

Commercial Data Mining Software

SPSS Clementine	116, 73 alone or with SPSS
Salford CART/MARS/TreeNet/RF	106, 54 alone
Excel	94, 2 alone
SPSS	91, 49 alone or with Clementine
SAS	80, 8 alone or with SAS E-Miner
Angoss	78, 50 alone
KXEN	70, 51 alone
SQL Server	38, 2 alone
MATLAB	30, 1 alone
SAS E-Miner	25, 8 alone or with SAS
Other commercial tools	21, 0 alone
Statsoft Statistica	15, 2 alone
Insightful Miner/S-Plus	14, 0 alone
Oracle DM	12, 0 alone
Tiberius	11, 3 alone
Fairlsaac Model Builder	3, 2 alone
Xelopes	2, 2 alone
Miner3D	2, 0 alone
Bayesia	2, 0 alone
Megaputer	1, 1 alone

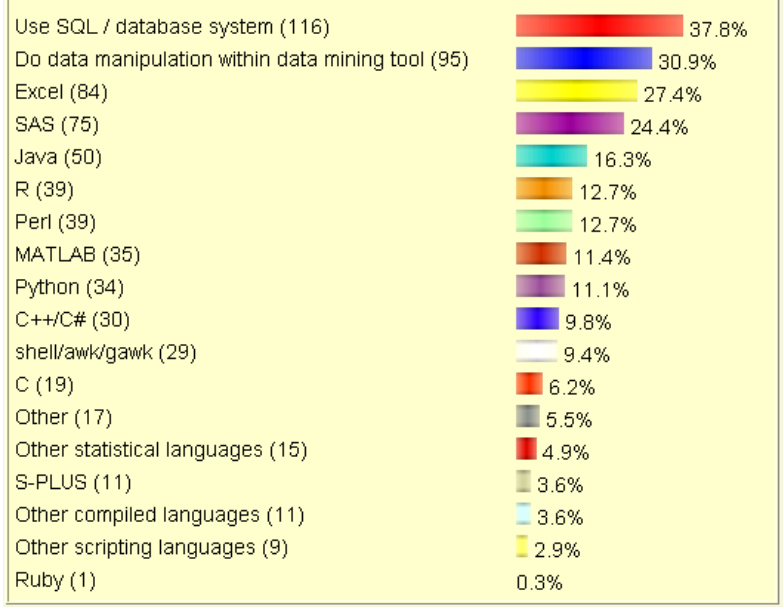
Data mining: What, Who, Why, How, Where?

5. Choose the implementation platform / programming languages

KDnuggets : **Polls** : Data Manipulation Tools/Languages (June 2007)

Poll

What tools/languages you typically use for data manipulation [307 voters]



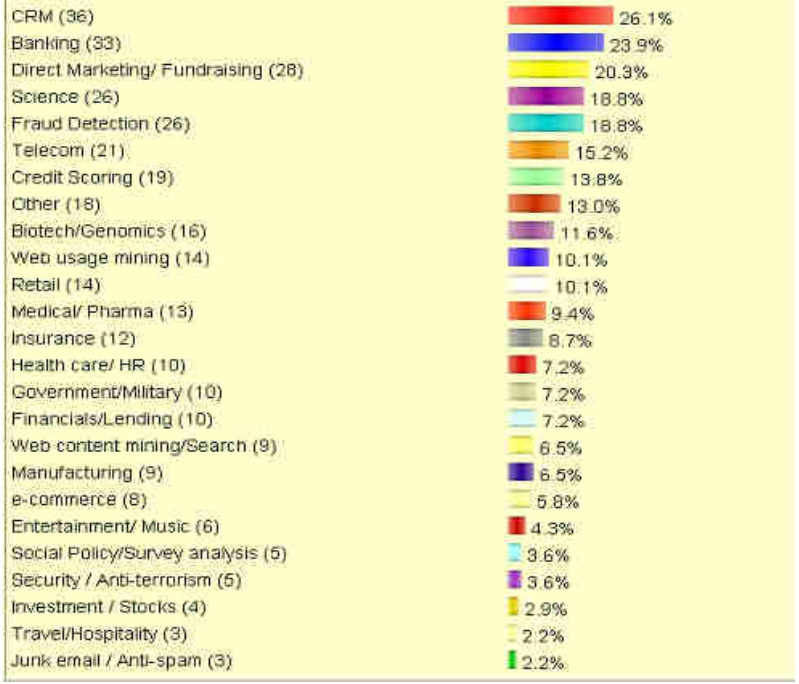
Data mining: What, Who, Why, How, Where?

Where to work?

KDnuggets : Polls : Data Mining Applications by Industry (June 2007)

Poll

Industries/fields where you applied data mining in the past 12 months [138 voters]



Motivation: Evolution of Database Technology

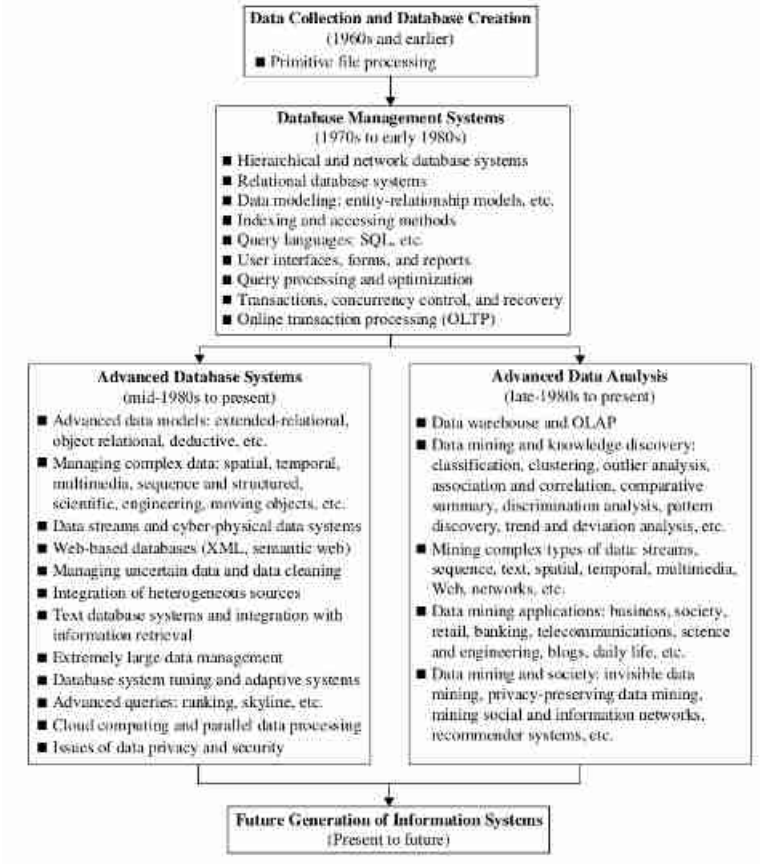
- The major reason for using data mining techniques is requirement of useful information and knowledge from huge amounts of data.
- The information and knowledge gained can be used in many applications such as business management, production control etc.
- Data mining came into existence as a result of the natural evolution of information technology.

Database → Evolutionary Path

Evolutionary path in the database industry has developed the following functionalities:

1. Data collection and database creation.
2. Data management (including data storage and retrieval, and database transaction processing)
3. Data analysis and understanding (involving data warehousing and data mining)

Database → Evolutionary Path



Data Mining as the Evolution of Information Technology

- During 1960's database and information technology has been evolving from primitive file processing systems to powerful database systems.
- During 1970's relational database systems were developed. In addition users access data through query languages. Efficient methods for on-line transaction processing (OLTP) were developed.
- During the mid-1980s many advanced database systems and application-oriented database systems were developed.
- In 1990's Heterogeneous database systems and Internet-based global information systems such as the World-Wide Web (WWW) also emerged and play a vital role in the information industry

Data Mining → On What Kind of Data ?

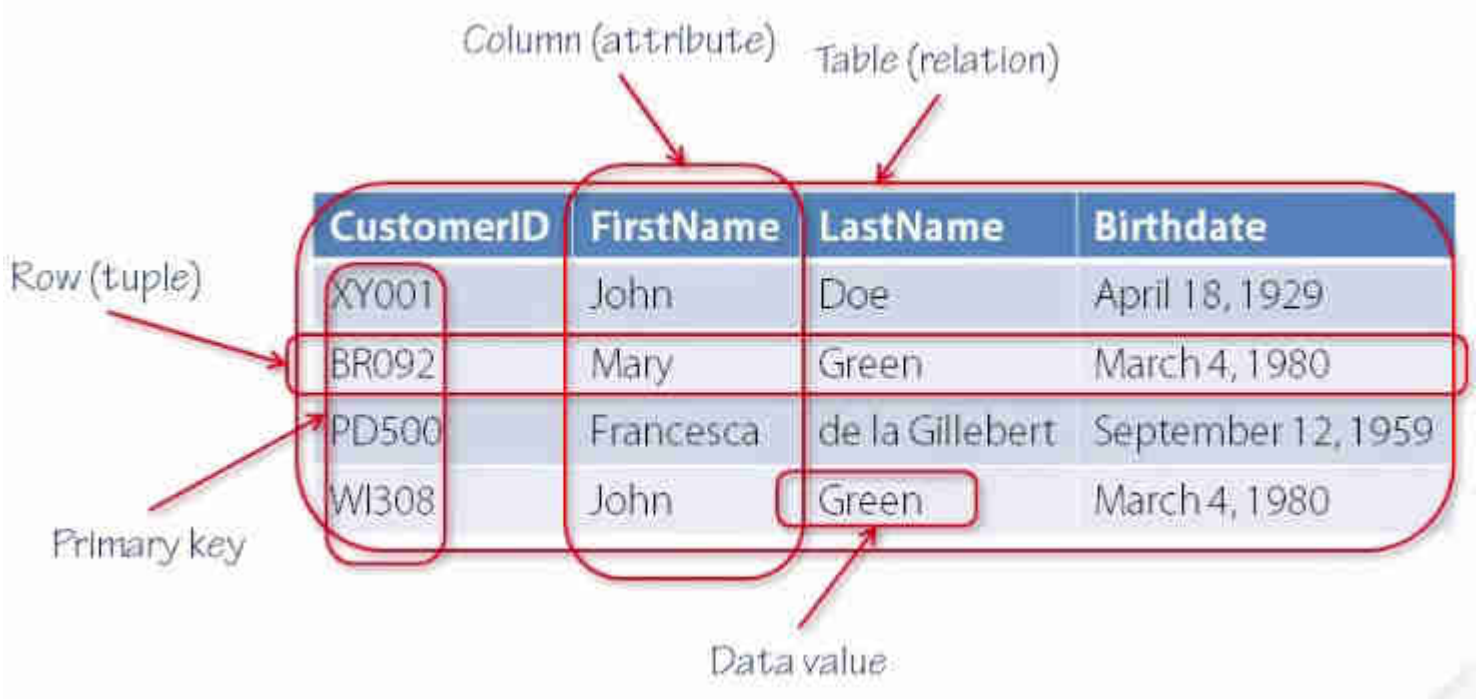
- There are number of different data stores on which mining can be performed. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World-Wide Web.
- Advanced database systems include object-oriented and object relational databases, and specific application-oriented databases such as spatial databases, time-series databases, text databases, and multimedia databases

Relational Database

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 number of tuples (records or rows). Each tuple in a relational table represents an record identified by a unique key and described by a set of attribute values

Relational Database



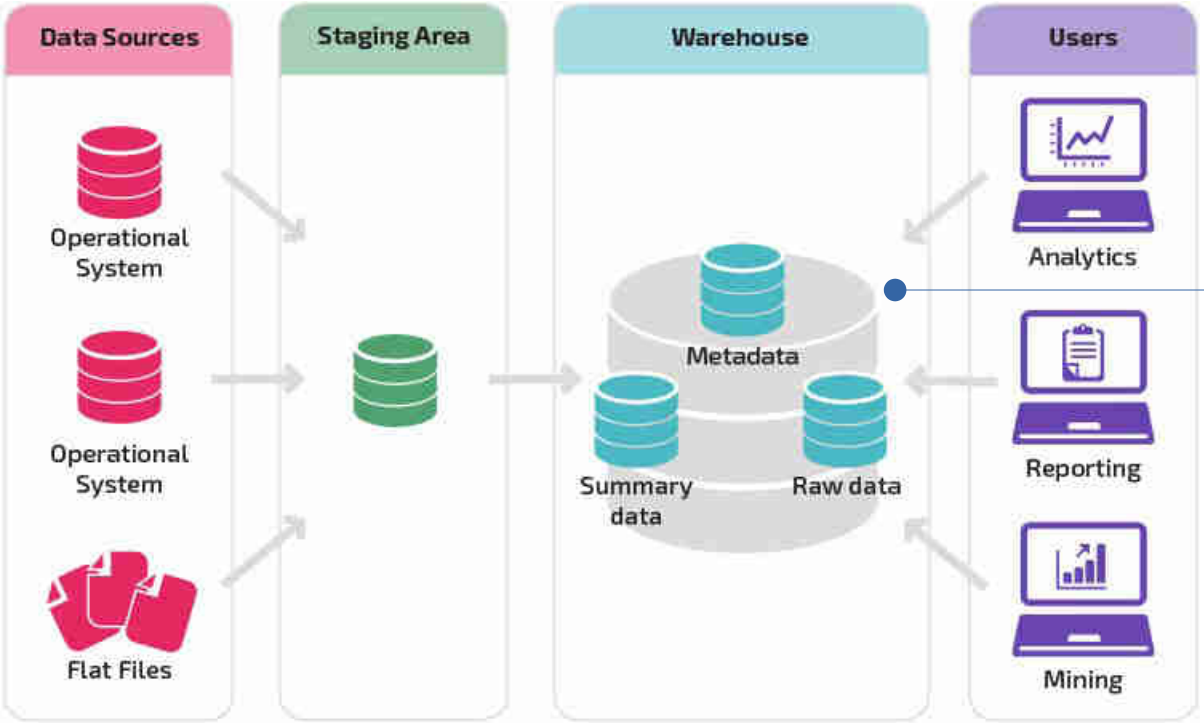
Relational Data Mining

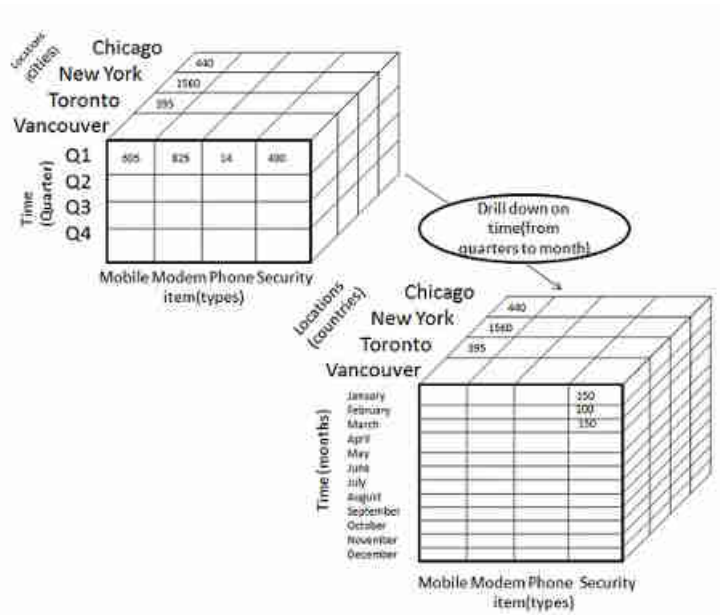
- Multi-disciplinary field dealing with knowledge discovery from relational databases consisting of multiple tables (relations).
- Mining data which consists of complex/structured objects also falls within the scope of this field: the normalized representation of such objects in a relational database requires multiple tables.
- Aims at integrating results from existing fields such as inductive logic programming (ILP), KDD, data mining, machine learning and relational databases; producing new techniques for mining multi-relational data; and practical applications of such techniques.

Data-ware house

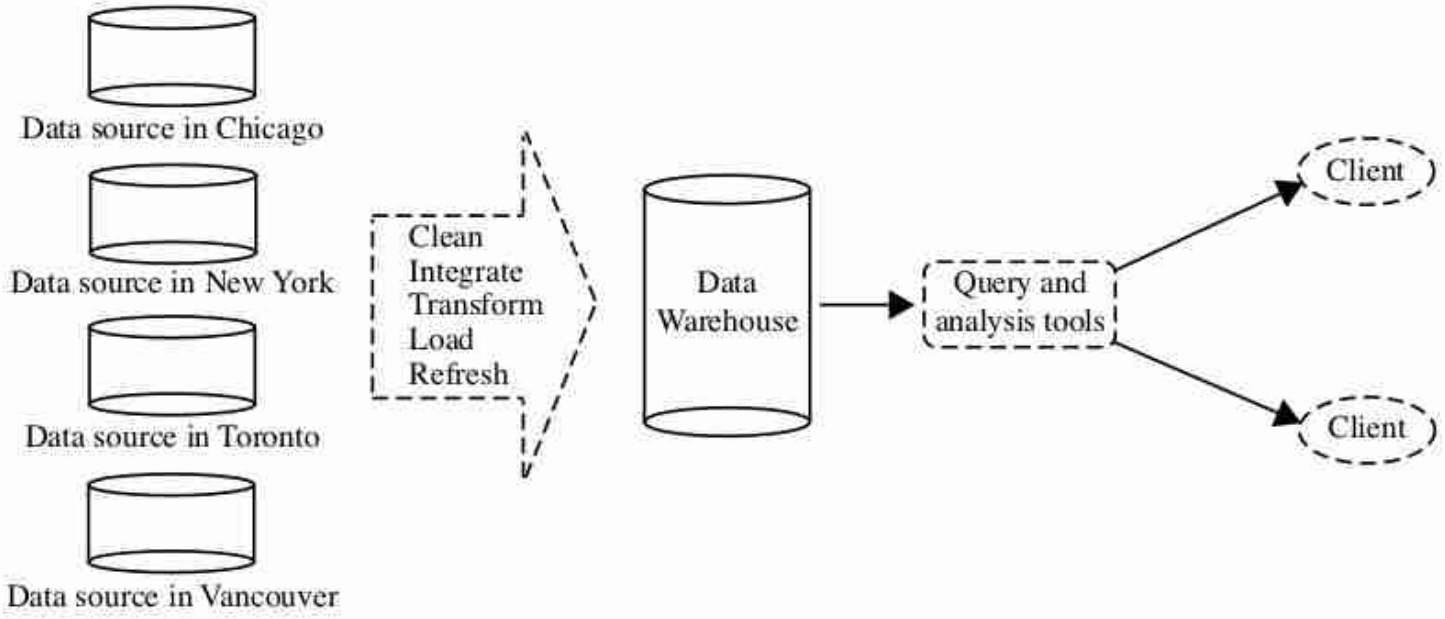
- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing.
- 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.

Data-ware house





Dataware House → Example



Typical framework of a data warehouse for *AllElectronics*.

Transactional databases

- 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 making up the transaction.
- The transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the sales person, and of the branch at which the sale occurred, and so on.

TID	Transaction (item, quantity)
T_1	a:3, c:2, d:3
T_2	a:2, d:1, e:2
T_3	b:3, c:5
T_4	a:1, c:3, e:1, f:2
T_5	b:1, d:3, e:2
T_6	b:2, d:2
T_7	b:3, c:2, e:1, f:1
T_8	a:2, f:2
T_9	c:3, d:2, f:1
T_{10}	a:2, c:2, d:1

Data mining: Uses?

- **Market segmentation** : Identify the common characteristics of customers who buy the same products from your company.
- **Fraud detection** : Identify transactions that are most likely to be fraudulent.
- **Direct marketing** : Identify the prospects who should be included in a mailing list to obtain the highest response rate.
- **Interactive marketing** : Predict what each individual accessing a web site is most likely interested in seeing.
- **Market basket analysis** : Understand what products or services are commonly purchased together, e.g beer and diapers.
- **Trend analysis** : Reveal the difference in a typical customer between the current month and the previous one

Data mining: Generate new business opportunities?

- **Automated prediction of trends and behaviors:**

  Data mining automates the process of finding predictive information in large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is targeted marketing.

- **Automated discovery of previously unknown patterns:**

  Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify unrelated products that are often purchased together.

Data mining: Generate new business opportunities?

- **Market management :**

  Target marketing, customer relationship management,market basket analysis, cross-selling, market segmentation

- **Risk management :**

  Forecasting, customer retention, improved underwriting, quality control, competitive analysis.

- **Fraud management :**

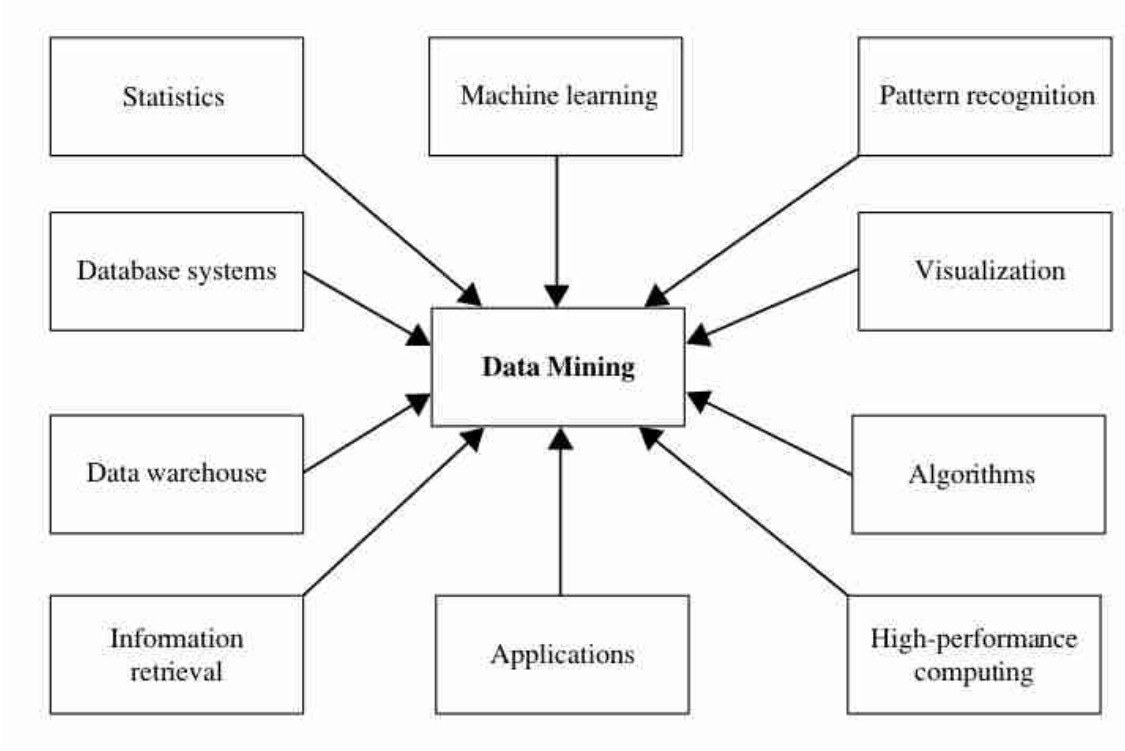
  Fraud Detection

Data mining: Generate new business opportunities?

- **Industrial-specific applications :**

  Banking, finance, and securities: Profitability analysis (for individual officer branch, product, product group, monitoring marketing programs and channels, customer data analysis customer segmentation profiling).

Data mining → Tools Used



Data mining → Tools Used(Machine Learning)

- **Supervised learning**
  The supervision in the learning comes from the labeled examples in the training data set

- **Unsupervised learning**
  The learning process is unsupervised since the input examples are not class labeled

- **Semi-supervised learning**
  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

Data mining → Major Issues

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 tasks

- **Handling noisy or incomplete data**

  The data stored in a database may reflect noise, exceptional cases, or incomplete data objects

- **Pattern evaluation—the interestingness problem**

  A data mining system can 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.

2. Performance issues

- **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

- **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. In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases

3. Diversity of database

- **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

- **Mining information from heterogeneous databases**

  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.

Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

1. Class/Concept Description: Characterization and Discrimination

- Data characterization

  Summarizing the data of the class under study (often called the target class)

- Data discrimination

  Comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

2. Mining Frequent Patterns, Associations, and Correlations

- Frequent patterns

  A frequent item-set 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.

Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

- Associations Analysis

  Once the frequent item-set from transactions in a database have been found, it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence)

  Suppose, as a marketing manager, you would like to determine which items are frequently purchased together within the same transactions.

  buys(X,"computer") → buys(X,"software") [support=1%,confidence=50%]

  where X is a variable representing a customer. Confidence=50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.

  Support=1% means that 1% of all of the transactions under analysis showed that computer and software were purchased together.

Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

- Correlations Analysis

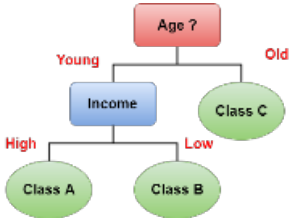
  Correlation analysis is used to find the association between the variables in data mining. Correlation methods are Pearson's product-moment correlation coefficient, Kendall and Spearman rank correlations, etc.

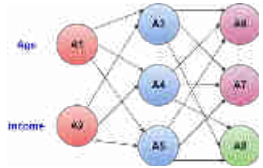
Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

3. Classification and Regression for Predictive Analysis

- Classification

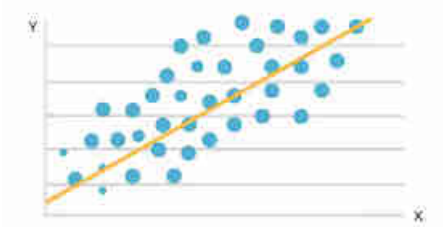
  Process of finding a model (or function) that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of training data. Example, decision tree, Neural Networks





- Regression

  It helps to predict the value of future outcomes by using the past data. This method helps in forecasting the data and Time-series Analysis

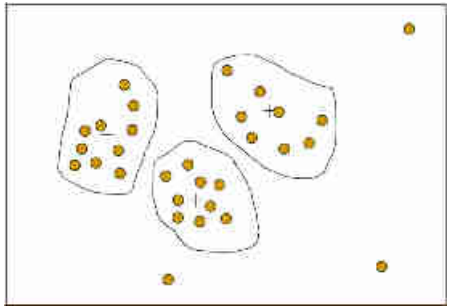


Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

4. Cluster Analysis, Outlier analysis, Evolution analysis

- Cluster Analysis

The objects are 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 very dissimilar to objects in other clusters.

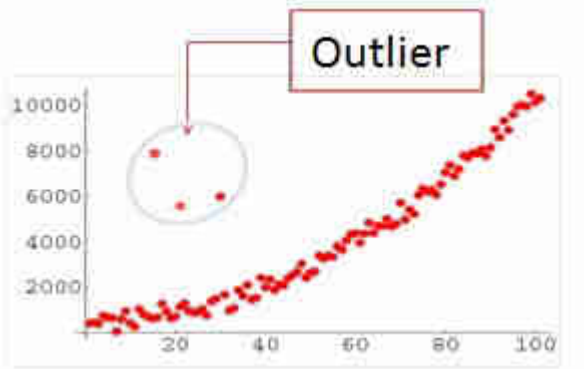


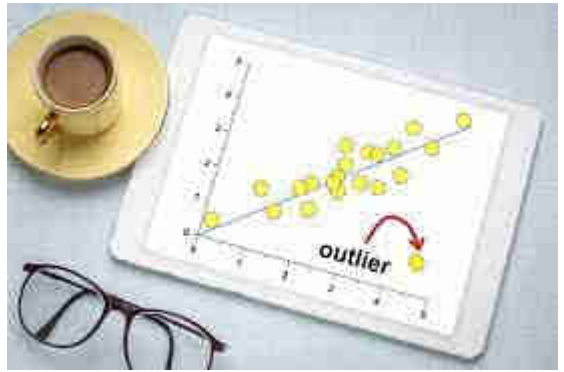
Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

4. Cluster Analysis, Outlier analysis, Evolution analysis

- Outlier Analysis

Outlier analysis is the process of identifying outliers, or abnormal observations, in a dataset.
Also known as outlier detection, it's an important step in data analysis, as it removes erroneous or inaccurate observations which might otherwise skew conclusions.





Data Mining Functionalities → What Kinds of Patterns Can Be Mined?

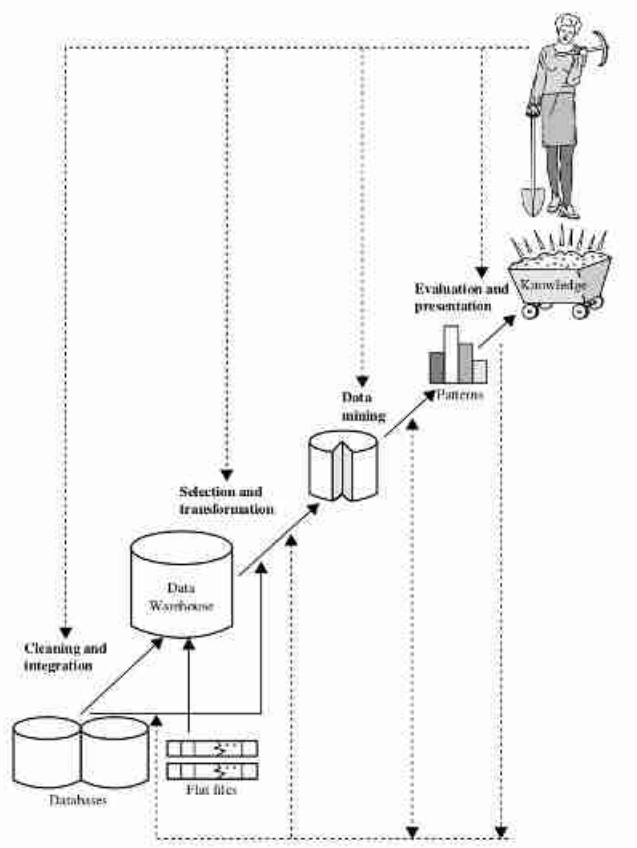
4. Cluster Analysis, Outlier analysis, Evolution analysis

- Evolution and Deviation analysis:

  Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values

Stages of knowledge discovery in database (KDD)

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



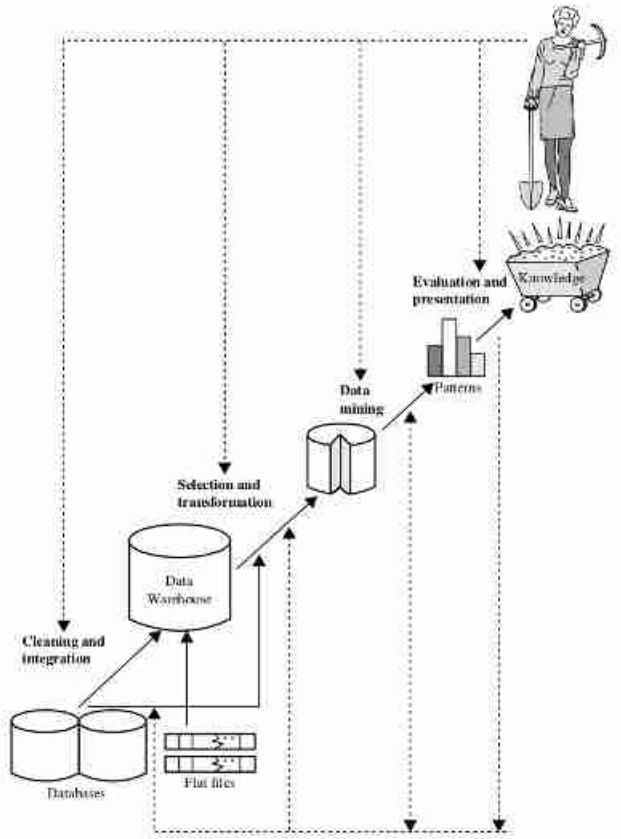
Stages of knowledge discovery in database (KDD)

1. **Data Cleaning and Integration**

   - Data cleaning in which noise and inconsistent data is removed.
   - Data Integration in which multiple data sources are combined.

2. **Data Selection and Transformation**

   - Data Selection in which data relevant to the analysis task are retrieved from the database.
   - In Data Transformation, data are transformed into forms appropriate for mining by performing summary or aggregation operations



Stages of knowledge discovery in database (KDD)

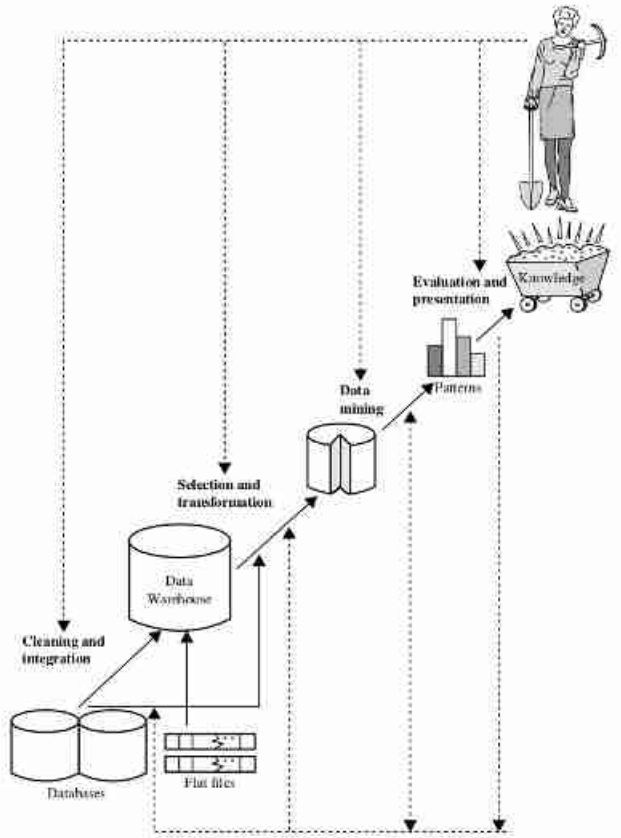
3. **Data Mining**

- data mining methods (algorithms) are applied in order to extract data patterns.

4. **Evaluation and Presentation**

- In Pattern Evaluation, data patterns are identified based on some interesting measures.

- Knowledge is represented to user using many knowledge representation techniques.



Setting up KDD Environment

Main goal of KDD is to obtain better understanding of changing organizational environment. It includes a data mining process but using data mining as single generic tool is neither realistic nor desirable.

KDD environment needs a suit of data mining tools that needs to be selected and tuned carefully for each organization according to their need.

Setting up KDD Environment

1. **Support for Extremely Large Data Set:** Data mining deals with billions of records. Thus fast and flexible way of storing and handling such large volume of data is requires along with capacity of storing intermediate results.

2. **Support Hybrid Learning:** Leaning can be divided into three categories: Classification, Knowledge engineering, & problem solving. Complex data mining projects needs hybrid learning algorithms that possess capabilities of all three categories.

3. **Establish a Data Warehouse:** Data mining mainly depends upon availability and analysis of historic data and hence establishing data warehouse is important for this.

4. **Introduce Data Cleaning Facility:** If data stored in databases contains noise, data processing suffers from pollution. Thus data cleaning facility is necessary to avoid such pollution.

5. **Facilitate Working with Dynamic Coding:** A KDD Environment should enable the user to experiment with different coding schemes. It must keep the track of genealogy of different samples and tables as well as the semantics and transformations of the different attributes are vital.

6. **Beware of False Predictors:** If the results are too good to be true, you probably have found false predictors.

7. **Verify Result:** Examine the results carefully and repeat and refine the knowledge discovery process until you are confident.

Critical Thinking…. !!!!

1. **Are all of the patterns interesting? What makes a pattern interesting? Can a data mining system generate all of the interesting patterns?**
2. **DBMS vs Data-mining**