1
DATA MINING

1.1 Definition of 'Data Mining:

Definition: In simple words, data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).

Description:
Key features of data mining:

• Automatic pattern predictions based on trend and behavior analysis.

• Prediction based on likely outcomes.

• Creation of decision-oriented information.

• Focus on large data sets and databases for analysis.

• Clustering based on finding and visually documented groups of facts not previously known.

1.2 Specific use of data mining

- Market segmentation
    - Data mining helps to identify the common characteristics of customers who buy the same products from your company
- Customer anticipation(expectation)
    - It helps to predict which customers may leave your company and go to a competitor
- Fraud detection- it identifies which transaction are most likely to be fraudulent.
- Direct marketing
    - Direct marketing identifies which prospects should be included to obtain the highest response rate.

- Interactive marketing
    - It is useful for predicting what each user on web site is most likely interested in seeing.
- Market basket analysis
    - It helps to understand what product or services are commonly purchased together.
- Trend analysis
    - Trend analysis identifies the difference between typical customers this month and last.

1.3 Challenges of Data Mining:

- Scalability: Scalable techniques are needed for handling massive size of datasets that are now created.
- Poor efficiency: Such large datasets require the use of efficient method for storing, indexing and retrieving data from secondary or even tertiary storage system.
- Complexity: Such techniques can dramatically increase the size of the datasets which can be handled and for that it may requires new design and algorithm.
- Poor quality: poor quality such as noisy data dirty data missing value, in exact or incorrect data.

1.4 Knowledge discovery in Database

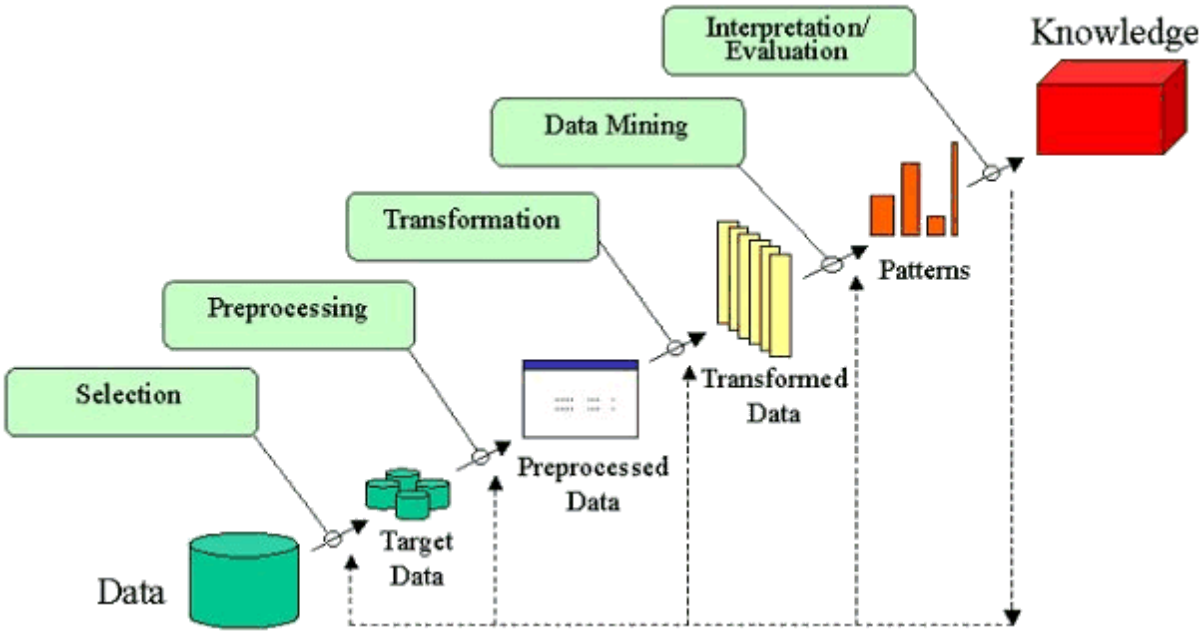


Fig:An Outline of the Steps of the KDD Process

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

1. Developing an understanding of
    - the relevant prior knowledge
    - the goals of the end-user
2. Creating a target data set: (SELECTION)
    - selecting a data set, or
    - Focusing on a subset of variables, or data samples (*a **data sample** is a set of **data** collected and/or selected from a statistical population by a defined procedure. The elements of a **sample** are known as **sample** points, **sampling** units or observations. ... The **sample** usually represents a subset of manageable size*.), on which discovery is to be performed.
3. Data cleaning and preprocessing.(PREPRCESSING)
    - ***Data preprocessing** is a **data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors .**Data preprocessing** is a proven method of resolving such issues.*
    - Removal of noise or outliers.
        1. Fill missing values
        2. Data not entered due to misunderstanding

        How to handle missing data:

    - Strategies for handling missing data fields.
        1. Filling missing value manually
        2. Use of global constant
        3. Imputations(use of attribute mean to fill the missing value
    - Accounting for time sequence information and known changes.
4. Data reduction and projection.(TRANSFORMATION)
    - Finding useful features to represent the data depending on the goal of the task.
    - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
5. Choosing the data mining algorithm(s).
    - Selecting method(s) to be used for searching for patterns in the data.
    - Deciding which models and parameters may be appropriate.

- Matching a particular data mining method with the overall criteria of the KDD process.

6. Data mining.
   - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
7. Interpreting mined patterns.
8. Consolidating discovered knowledge.

The terms *knowledge discovery* and *data mining* are distinct.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.
Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

1.5 Data pre-processing

Why pre-processing the data ?

We pre-process the data because real world data are generally

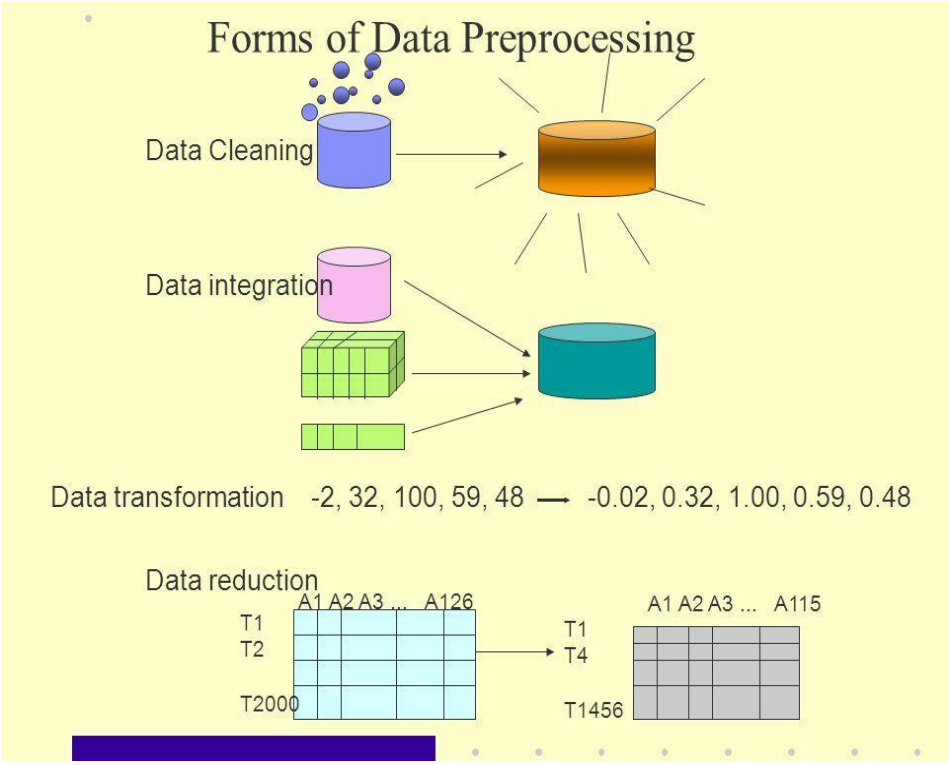
1.1. Incomplete:
Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

1.2. Noisy:
Containing errors or outliers

1.3. Inconsistent:
Containing discrepancies in codes or names

Because of this reason we need to pre-process the data.

Froms of data pre-processing



1. Data cleaning

Real world data tend to be incomplete, noisy and inconsistent; Data cleaning routines attempt to fill the missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. Basic methods of data cleaning are:

1. Fill in missing values

If the values are missed then following steps can be taken.

1.1. Ignore the tuple:

Usually done when class label is missing. This method is not very effective unless the tuples contain several attributes with missing values.

1.2. Use the attribute mean (or majority nominal value) to fill in the missing value.

Suppose that the average income of a company of customer is 2000. Use this value to replace the missing value of income.

1.3. Use the attribute mean (or majority nominal value) for all samples belonging to the same class.

If classifying customer according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

1.4. *Predict the missing value by using a learning algorithm:*

Consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bays or decision tree) to predict the missing value.

2. Identify outliers and smooth out noisy data:

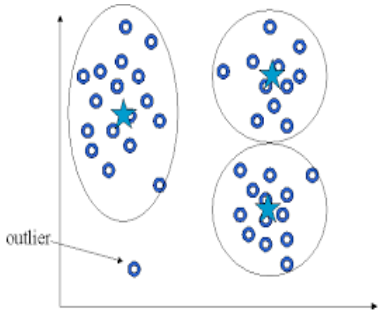
2.1. Binning

Sort the attribute values and partition them into bins (see "Unsupervised discretization" below);

Then smooth by bin means, bin median, or bin boundaries.

2.2. Clustering:

Outliers may be detected by clustering where similar values are organised into groups or clusters. Intuitively, values that fall outside of the site of clusters may be considered outliers.



2.3. Regression:

Data can be smoothed by fitting the data to a function, such as with regression. Linear regression involves finding the “best” line to fit two variables, so that on variable can be used to predict the other.

2.4. Correct inconsistent data:

There may be inconsistencies in data recorded for some transactions. Some data inconsistencies may be corrected manually using external references, for ex, errors made at data entry may be corrected by performing a paper trace. Knowledge engineering tools may also be used to detect the violation of known data constraints.

There may be inconsistencies due to data integration, where a given attribute can have different name in different database.

2. Data transformation

a) Normalization:

a. Scaling attribute values to fall within a specified range.

i. Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)

b. Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev

b) Aggregation: moving up in the concept hierarchy on numeric attributes.
c) Generalization: moving up in the concept hierarchy on nominal attributes.
d) Attribute construction: replacing or adding new attributes inferred by existing attributes.

3. Data reduction

1. Reducing the number of attributes
    - Data cube aggregation: applying roll-up, slice or dice operations.
    - Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space .
    - Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data..
2. Reducing the number of attribute values
    - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
    - Clustering: grouping values in clusters.
    - Aggregation or generalization
3. Reducing the number of tuples
    - Sampling

1.6 Applications of Data Mining

1. E-Commerce:

- For Business Intellegence
    - Offers upsell.example Amazon.com
- Fraud detection:
    - *A problem faced by all e-commerce companies is misuse of our systems and, in some cases, fraud. For example, sellers may deliberately list a product in the wrong category to attract user attention, or the item sold is not as the seller described it. On the buy side, all retailers face problems with users using stolen credit cards to make purchases or register new user accounts.*
    - *Fraud detection involves constant monitoring of online activities, and automatic triggering of internal alarms. Data mining uses statistical analysis and machine learning for the technique of "anomaly detection", that is, detecting abnormal patterns in a data sequence.*

- *Detecting seller fraud requires mining data on seller profile, item category, listing price and auction activities. By combining all of this data, we can have a complete picture and fast detection in real time.*

- Product Search:
  - *When the user searches for a product, how do we find the best results for the user? Typically, a user query of a few keywords can match many products. For example, "Verizon Cell phones" is a popular query at eBay, and it matches more than 34,000 listed items.*
  - *One factor we can use in product ranking is user click-through rates or product sell-through rate. Both indicate a facet of the popularity of a product page. In addition, user behavioral data gives us the link from a query, to a product page view, and all the way to the purchase event. Through large-scale data analysis of query logs, we can create graphs between queries and products, and between different products. For example, the user who searches for "Verizon cell phones" might click on the Samsung SCH U940 Glyde product, and the LG VX10000 Voyager. We now know the query is related to those two products, and the two products have a relationship to each other since a user viewed (and perhaps considered buying) both.*

- **Product recommendation**
  - *Recommending similar products is an important part of eBay. A good product recommendation can save hours of search time and delight our users.*
  - *Typical recommendation systems are built upon the principle of "collaborative filtering", where the aggregated choices of similar, past users can be used to provide insights for the current user. We do this in our new product based experience. Try viewing our Apple iPod touch 2nd generation page and scroll down — you'll see that users who viewed this product also viewed other generations of the iPod touch and the iPod classic.*
  - *Discovering item similarity requires understanding product attributes, price ranges, user purchase patterns, and product categories. Given the hundreds of millions of items sold on eBay, and the diversity of merchandise on our website, this is a challenging computational task. Data mining provides possible tools to tackle this problem, and we are always actively improving our approach to the problem.*

2. Crime Agencies:

- Use to spot trends across the data helping with everything from where to deploy police manpower. (where the crime is mostly likely to happen).
- To search at a border crossing (based on age/type of the vehicle, age of occupation.

- **Data mining and criminal intelligence techniques**
    - ***Entity extraction****: Commonly used to automatically identify people, organizations, vehicles and personal details in unstructured data such as police reports. Even if entity extraction provides only basic information, it can accelerate the investigation by rapidly providing precise details from large amounts of unstructured data.*
    - ***Clustering techniques:*** *Clustering techniques are used to group similar characteristics together in classes in order to gain intelligence by maximizing or minimizing similarities; for example, to identify suspects or criminal groups conducting crimes in similar ways. Clustering techniques could be effectively applied through conceptual space algorithms to discover criminal relations by cross referencing entities in criminal records.*
    - ***Association rules:*** *This data mining technique has been used to discover recurring items in databases in order to create pattern rules and detect potential future events. This technique has been effective in preventing network intrusions and attacks, such as* ***denial of service attacks****(DDoS).*
    - ***Sequential pattern mining:*** *as association rule it is useful to identify sequences or recurring item in order to define patterns and prevent attacks, in network security.*
    - ***Classification:*** *This technique is useful for analyzing unstructured data to discover common properties among criminal entities. Classification has been used together with inferential statistics techniques to predict crime trends. This technique can dramatically narrow down different criminal entities and organize them into predefined classes.*
    - ***String comparison:*** *This technique is used to reveal deceptive information in criminal records by comparing structured text fields. This requires highly intensive computational capabilities.*

3. Telecommunication

- *Telecommunication companies maintain data about the phone calls that traverse their networks in the form of call detail records, which contain descriptive information for each phone call. In 2001, AT&T long distance customers generated over 300 million call detail records per day (Cortes & Pregibon, 2001) and, because call detail records are kept online for several months, this meant that billions of call detail records were readily available for data mining.* ***Call detail data is useful for marketing and fraud detection applications.***

- *Telecommunication companies also maintain extensive customer information, such as billing information, as well as information obtained from outside parties, such as credit score information. This information can be quite useful and often is combined with telecommunication-specific data to improve the results of data mining. For example,* ***while call detail data can be used to identify suspicious calling patterns, a customer's credit score is often incorporated into the analysis before determining the likelihood that fraud is actually taking place.***

- *Telecommunications companies also generate and store an extensive amount of data related to the operation of their networks. This is because the network elements in these large telecommunication networks have some self-diagnostic capabilities that permit them to generate both status and alarm messages.* ***These streams of messages can be mined in order to support network management functions, namely fault isolation and prediction.***

4. Biological Data Analysis:

- *Thousands of genes (~25K in human DNA) function in a complicated and orchestrated way that creates the mystery of life. !*
- *Genomic studies the functionality of specific genes, their relations to diseases, their associated proteins and their participation in biological processes !*
- *Proteins (~1M in human organism) are responsible for many regulatory functions in cells, tissues and organism !*
- *Proteome, the collection of proteins produced, evolves dynamically during time depending on environmental signals. !*
- *Proteomic studies the sequences of proteins and their functionalities*
- *(case of stem cell, haruko obokata)*

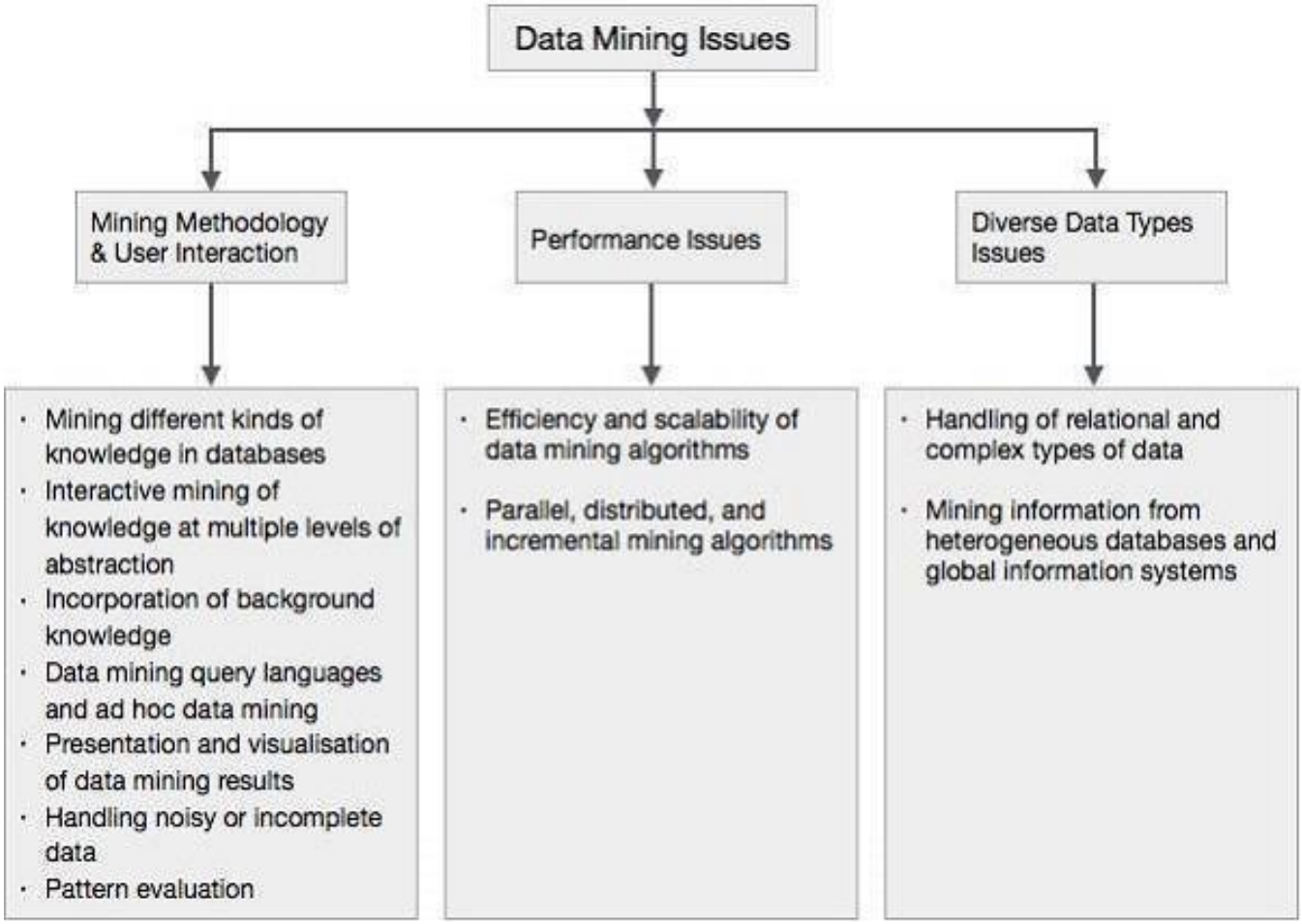
1.7 Data mining issues:

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

- Mining Methodology and User Interaction

- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



1. Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − the data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

2. Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms**− In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions are merged. The incremental algorithms, update databases without mining the data again from scratch.

3. Diverse Data Types Issues

- **Handling of relational and complex types of data** − the database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − the data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

1.8 Disadvantages of data mining

- **Privacy Issues**
    - The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time, the personal information they own probably is sold to other or leak.

- **Security issues**
    - Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from the big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem.
- **Misuse of information/inaccurate information**
    - Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.

In addition,

Data mining technique is not perfectly accurate. Therefore, if inaccurate information is used for decision-making, it will cause serious consequence.

1.9 Problems of Data Mining

The amount of data being generated and stored every day is exponential. A recent study estimated that every minute, Google receives over 2 million queries, e-mail users send over 200 million messages, YouTube users upload 48 hours of video, Face book users share over 680,000 pieces of content, and Twitter users generate 100,000 tweets. Besides, media sharing sites, stock trading sites and news sources continually pile up more new data throughout the day.

The common problems in Data Mining.

1. Poor data quality such as noisy data, dirty data, missing values, inexact or incorrect values, inadequate data size and poor representation in data sampling.

2. Integrating conflicting or redundant data from different sources and forms: multimedia files (audio, video and images), geo data, text, social, numeric, etc…
3. Proliferation of security and privacy concerns by individuals, organizations and governments.
4. Unavailability of data or difficult access to data.
5. Efficiency and scalability of data mining algorithms to effectively extract the information from huge amount of data in databases.
6. Dealing with huge datasets that require distributed approaches.
7. Dealing with non-static, unbalanced and cost-sensitive data.
8. Mining information from heterogeneous databases and global information systems.
9. Constant updation of models to handle data velocity or new incoming data.
10. High cost of buying and maintaining powerful softwares, servers and storage hardwares that handle large amounts of data.
11. Processing of large, complex and unstructured data into a structured format.
12. Sheer quantity of output from many data mining methods.