# Chapter 1

# Introduction to Data Mining

**Basic concepts of data mining**

## WHAT IS DATA MINING?

Data mining also called as data archeology, data dredging, data harvesting, is the process of extracting hidden knowledge from large volumes of raw data and using it to make crucial business decisions.

"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, non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large database".

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.

## DATA MINING DEFINITIONS

The term data mining has been stretched beyond its limits to apply to any form of data analysis. Some of the numerous definitions of data mining, or knowledge discovery in databases are:

- "Extraction of interesting information or patterns from data in large databases is known as data mining."
- According to **William J. Frawley, Gregory Piatetsky-Shapiro and Christopher J. Matheus** "*Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies"*
- According to **Marcel Holshemier and Arno Siebes** *"Data mining is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database".*
- Data mining refers to "*using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful".*
- Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.
  - **Valid:** The patterns hold in general. I
  - **Novel:** We did not know the pattern beforehand.
  - **Useful:** We can devise actions from the patterns.
  - **Understandable:** We can interpret and comprehend the patterns.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses. Three steps involved are:

- Exploration
- Pattern identification

- Deployment

***Exploration:*** In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

***Pattern identification:*** Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

***Deployment:*** Patterns are deployed for desired outcome.

Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data."Mining" is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

**Why Use Data Mining Today?**

1. Human analysis skills are inadequate:
   - Volume and dimensionality of the data
   - High data growth rate
2. Availability of:
   - Data
   - Storage
   - Computational Power
   - Off-the-shelf software
   - Expertise
3. An Abundance of data through
   - Supermarket scanners, POS data
   - Preferred customer cards
   - Credit card transactions
   - Direct mail response
   - Call center records.
   - ATM machines
   - Demographic data
   - Sensor networks
   - Cameras
   - Web server logs
   - Customer web site trails
4. Competitive pressure!
   - ***"The secret of success is to know something thal nobody else knows."***
     -Aristotle Onassis
   - Competition on service, not only on price (Banks. phone companies, hotel chains, rental car companies)
5. Personalization, CRM
6. The real-time enterprise
7. "Systemic listening"
8. Security homeland defense

**What is NOT Data Mining?**

- Searching a phone number in a phone book
- Searching a keyword on Google
- Generating histograms of salaries for different age groups
- Issuing SQL query to a database and reading the reply'

Measurable benefits from data mining have been achieved in many different domains:

- ***Fraud management -*** e.g. telecommunications, financial, insurance industries
- ***Market analysis -*** customer, competition, trend analyses
- ***Product development -*** biotechnology, pharmaceutical industry
- ***Entertainment -*** digital convergence, sports
- ***Diagnosis and monitoring -*** medical, aerospace, automotive.

## APPLICATIONS OF DATA MINING

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Data mining has many and varied fields of application some of which are listed below.

1. **Sales/Marketing**
   a) Identify buying patterns from customers
   b) Find associations among customer demographic characteristics
   c) Predict response to mailing campaigns
   d) Market basket analysis.
2. **Banking**
   a) Credit card fraudulent detection
   b) Identify 'loyal' customers
   c) Predict customers likely to change their credit card affiliation
   d) Determine credit card spending by customer groups
   e) Find hidden correlation's between different financial indicators
   f) Identify stock trading rules from historical market data
3. **Insurance and Health Care**
   a) Claims analysis i.e., which medical procedures are claimed together
   b) Predict which customers will buy new policies
   c) Identify behavior patterns of risky customers
   d) Identify fraudulent behavior
4. **Transportation**
   a) Determine the distribution schedules among outlets
   b) Analyze loading patterns
5. **Medicine**
   a) Characterize patient behavior to predict office visits
   b) Identify successful medical therapies for different illnesses

## DISADVANTAGES OF DATA MINING

### Privacy issues

The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs etc. Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of trouble. 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 employee and customers including social security number, birthday, payroll and etc. However how properly this information is taken is still in questions. There have been a lot of cases that hackers were accesses and stole big data of customers from 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 collected through data mining intended for marketing or ethical purposes can be misused. This information is exploited by unethical people or business to take benefit of vulnerable people or business to take benefit 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 will cause serious consequence.

**Functions of Data mining**

Data mining has five main functions:

- ***Classification:*** infers the defining characteristics of a certain group (such as customers who have been lost to competitors).
- ***Clustering:*** identifies groups of items that share a particular characteristic. (Clustering differs from classification in that no predefining characteristic is given in classification.)
- ***Association:*** identifies relationships between events that occur at one time (such as the contents of a shopping basket).
- ***Sequencing:*** similar to association, except that the relationship exists over a-period of time (such as repeat visits to a supermarket or use of a financial planning product).
- ***Forecasting:*** estimates future values based on patterns within large sets of data (such as demand forecasting).

Many people treat data mining as a synonym for another popularly used term, **"Knowledge discovery in databases"**, or **KDD.**

> ***Note:*** *Knowledge discovery in databases is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data. Data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or models in data.*

Data mining, or knowledge discovery in databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies.

THE PROCESS OF KNOWLEDGE DISCOVERY

The main steps of knowledge discovery process are:

1. Identify business Problem
2. Data mining
3. Action
4. Evaluation and measurement
5. Deployment and integration into businesses processes.

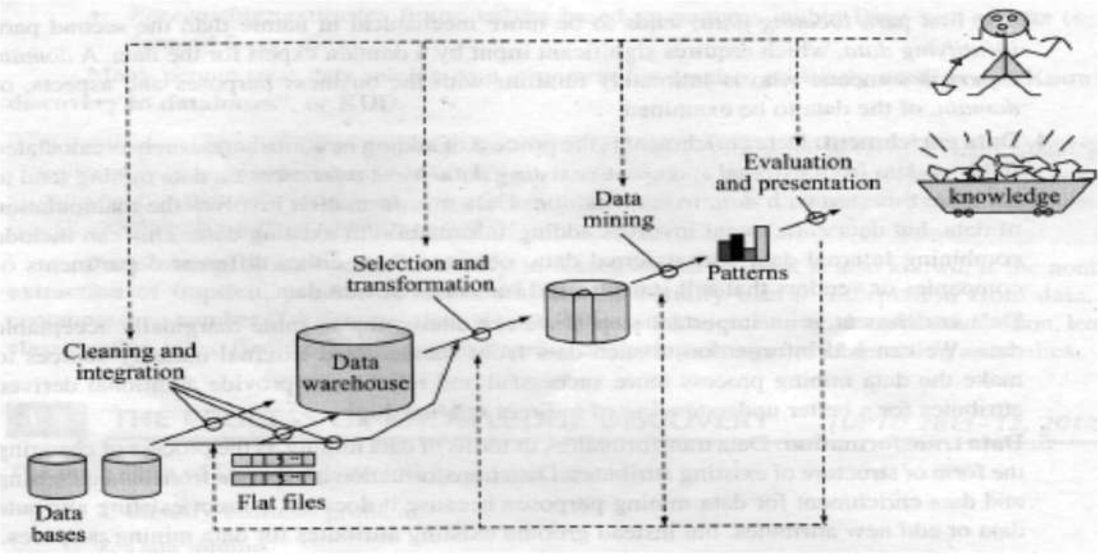

Figure 1.0

Knowledge discovery as a process is depicted in above Figure, and consists of an iterative sequence of the following steps:
Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining.

1. **Data cleaning** (to remove noise or irrelevant data): Data cleaning is the process of ensuring that, for data mining purposes, the data is uniform in terms of key and attributes usage.
   Data cleaning is separate from data enrichment and data transformation because data cleaning attempts to correct misused or incorrect attributes in existing data. Data enrichment, by contrast, adds new attributes to existing data, while data transformation changes the form or structure of attributes in existing data to meet specific data mining requirements.

   There are several types of cleaning process, some of which can be executed in advance while others are invoked only after pollution is detected at the coding or the discovery stage. An important element in a cleaning operation is the de-duplication of records. In a normal database some clients will be represented by several records, although in many cases this will be the result of negligence, such as people making typing errors, or of clients moving from one place to another without notifying change of address. Although data mining and data cleaning are two different disciplines, they have a lot in common and pattern recognition algorithm can be applied in cleaning data.

2. **Data integration** (where multiple data sources may be combined)

3. **Data selection:** There are two parts to selecting data for data mining:
   *i.* *locating data*
   *ii.* *identifying data*

The first part, *locating data*, tends to be more mechanical in nature than the second part, *identifying data*, which requires significant input by a domain expert for the data. A *domain expert* is someone who is intimately familiar with the business purposes and aspects, or *domain*, of the data to be examined.

4. **Data enrichment:** Data enrichment is the process of adding new attributes, such as calculated fields or data from external sources, to existing data. Most references on data mining tend to combine this step with data transformation. Data transformation involves the manipulation of data, but data enrichment involves adding information to existing data. This can include combining internal data

with external data, obtained from either different departments or companies or vendors that sell standardized industry-relevant data.

Data enrichment is an important step if we are attempting to mine marginally acceptable data. We can add information to such data from standardized external industry sources to make the data mining process more successful and reliable, or provide additional derived attributes for a better understanding of indirect relationships.

5. **Data transformation:** Data transformation, in terms of data mining, is the process of changing the form or structure of existing attributes. Data transformation is separate from data cleansing and data enrichment for data mining purposes because it does not correct existing attribute data or add new attributes, but instead grooms existing attributes for data mining purposes.

6. **Data mining:** Here we shall discuss some of the most important machine-learning and pattern recognition algorithms, and in this way get an idea of the opportunities that are available as well as some of the problems that occur during the discovery stage. We shall see that some learning algorithms do well on one part of the set where others fail, and this clearly indicates the need for hybrid learning.

   The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

7. **Pattern evaluation:** Pattern evaluation is used to identify the truly interesting patterns representing knowledge based on some interestingness measures.

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

   ***Visualization techniques*** are a very useful method of discovering patterns in datasets, and may be used at the beginning of a data mining process to get a rough feeling of the quality of the data set and where patterns are to be found. Interesting possibilities are offered by object oriented three dimensional tool kits, such as Inventor, which enable to user to explore three dimensional structures interactively. Advanced graphical techniques in virtual reality enable people to wander through artificial data spaces, while historic development of data sets can be displayed as a kind of animated movie. These simple methods can provide us with a wealth of information. An elementary technique that can be of great value is the so called **scatter diagram**. Scatter diagrams can be used to identify interesting subsets of the data sets so that we can focus on the rest of the data mining process. There is a whole field of research dedicated to the search for interesting projections for data sets that is called *projection pursuit*.

So, data mining steps are:

1. Data preprocessing
   a) Data selection: Identify target datasets and relevant fields
   b) Data cleaning
      - Remove noise and outliers
      - Data transformation
      - Create common units
      - Generate new fields
2. Data mining model construction
3. Model evaluation

***# Describe the steps involved in data mining when viewed as a process of knowledge discovery.***

The steps involved in data mining when viewed as a process of knowledge discovery are as follows:

- **Data cleaning,** a process that removes or transforms noise and inconsistent data.
- **Data integration,** where multiple data sources may be combined.
- **Data selection,** where data relevant to the analysis task are retrieved from the database.
- **Data transformation,** where data are transformed or consolidated into forms appropriate for mining.
- **Data mining,** an essential process where intelligent and efficient methods are applied in order to extract patterns.
- **Pattern evaluation,** a process that identifies the truly interesting patterns representing knowledge, based on some interestingness measures.
- **Knowledge presentation,** where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

**WHAT ARE THE ISSUES IN DATA MINING?**

One of the key issues raised by data mining technology is not a business or a technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and give a significant amount of information about individuals buying habits and preferences.

While data mining is still in its immaturity, it is becoming a trend. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below.

a) **Security and social issues:** Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential distribution of discovered information.

   Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

b) **User interface issues:** The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks' as well as to picture the discovered knowledge from different angles and at different conceptual levels.

c) **Mining methodology issues:** These issues are relevant to the data mining approaches applied and their limitations. Topics such as

- Versatility of the mining approaches
- Diversity of data available
- Dimensionality of the domain
- Broad analysis needs (when known)
- Assessment of the knowledge discovered
- Exploitation of background knowledge and metadata
- Control and handling of noise in data

are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently. Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the accuracy of the results. As a consequence, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information. More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the *curse of dimensionality.* This "curse" affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

d) **Performance issues:** Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets which data mining is dealing with today. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

e) **Data source issues:** We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the explosion of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.