PRELUDE

- ✓ Data mining is the process of extracting hidden predictive information from a large database. As more data are gathered, with the amount of data doubling every year, data mining is becoming an increasingly important tool to transform this data into information.
- ✓ It is commonly used in a wide range of profiling practices, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size.
- ✓ Data mining tools predict future tends and behaviors, allowing businesses to make proactive, knowledge driven decisions.
- ✓ Data mining sometimes called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information.
- ✓ Data mining software is an analytical tool for analyzing data. It allows users to analyze data from many different dimensions, categorize it, and summarize the relationships identified.
- ✓ Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining consists of five major elements:

- ✓ Extract, transform, and load transaction data onto the data warehouse system.
- ✓ Store and manage the data in a multidimensional database system.
- ✓ Provide data access to business analysts and information technology professionals.
- ✓ Analyze the data by application software.
- ✓ Present the data in a useful format, such as a graph or table.

ADVANTAGES OF DATA MINING

- ✓ Automated prediction of trends and behaviors :
  - o Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data, quickly.
- ✓ Automated Discovery of previously unknown patterns:
  - o Data mining tools sweep through databases and identify previously hidden patterns in one step.
  - o E.g. – analysis of retail sales data to identify apparently unrelated products that are often purchased together.
- ✓ Database can be larger in both depth and breadth:
  - o The databases can have more columns and rows. High performance data mining allows users to explore full depth of a database, without pre-selecting a subset of variables. The data mining database contain larger samples (more rows) as they yield lower estimation errors and variance, and allow users to make conclusion about small but important segments of a population.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms and can be implemented on new systems as existing platforms are upgraded and new products are developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means the users can experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

TECHNOLOGIES USED IN DATA MINING

- ✓ Artificial Neural Networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- ✓ Genetic Algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- ✓ Decision Trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- ✓ Nearest Neighbor Method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is 1). Sometimes called the k-nearest neighbor technique. It is also called case based reasoning.

- ✓ Rule Induction: The extraction of useful if-then rules from data based on statistical significance.

- ✓ Data Visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

- ✓ Evolutionary Programming: This is the most promising branch of data mining at present. The underlying idea of the method is that the system automatically formulates hypothesis about the dependence of the target variable on other variables in the form of programs expressed in an internal programming language.