Data Mining

Dr Na Yao

Objectives

- Understand and be able to explain the concept of data mining.
- Understand and be able to explain different applications of data mining.
- Understand some techniques associated with data mining operations, including:
  - predictive modelling,
  - database segmentation,
  - link analysis, and
  - deviation detection.
- Understand and be able to explain the relationship between data mining and data warehousing.

Data Mining, Definition

- The process of extracting **valid**, previously **unknown**, **comprehensible**, and **actionable** information from large databases and using it to make crucial **business decisions**.

- Involves the analysis of data and the use of software techniques for finding **hidden and unexpected** patterns and relationships in sets of data.

Data Mining, Characteristics

- Reveals information that **is hidden and unexpected**.
- Patterns and relationships are identified by examining the **underlying rules and features** in the data.
- Most accurate results normally require **large volumes of data** to deliver reliable conclusions.
- First step: optimal representation of **structure** of sample data, during which time knowledge is acquired, this is then extended to larger sets of data. Underlying assumption: same structure

Data Mining, Use and Advantages

- Data mining can provide huge paybacks for companies who have made a significant investment in data warehousing.

- Relatively new technology, however already used in more and more industries.

Examples of Applications of Data Mining

- Retail / Marketing
  - Identifying buying patterns of customers
  - Finding associations among customer demographic characteristics
  - Predicting response to mailing campaigns
  - Market basket analysis
- Banking
  - Detecting patterns of fraudulent credit card use
  - Identifying loyal customers
  - Predicting customers likely to change their credit card affiliation
  - Determining credit card spending by customer groups

Examples of Applications of Data Mining

- Insurance
  - Claims analysis
  - Predicting which customers will buy new policies

- Medicine
  - Characterizing patient behaviour to predict surgery visits
  - Identifying successful medical therapies for different illnesses

Data Mining Operations

- Four main operations include:
  - Predictive modeling
  - Database segmentation
  - Link analysis
  - Deviation detection

- There are recognized associations between the applications and the corresponding operations.
  - e.g. Direct marketing strategies use database segmentation.

Data Mining Techniques

- Techniques are specific implementations of the data mining operations.

- Each operation has its own strengths and weaknesses.

Data Mining Techniques

- Data mining tools sometimes offer a choice of operations to implement a technique.

- Criteria for selection of tool includes
  - Suitability for certain input data types
  - Transparency of the mining output
  - Tolerance of missing variable values
  - Level of accuracy possible
  - Ability to handle large volumes of data

Data Mining Operations and Associated Techniques

Operations	Data mining techniques
Predictive modeling	Classification
	Value prediction
Database segmentation	Demographic clustering
	Neural clustering
Link analysis	Association discovery
	Sequential pattern discovery
	Similar time sequence discovery
Deviation detection	Statistics
	Visualization

Predictive Modelling

- Similar to the human learning experience
  - uses observations to form a model of the important characteristics of some phenomenon.

- Uses generalizations of ‘real world’ and ability to fit new data into a general framework.

- Can analyze a database to determine essential characteristics (model) about the data set.

Predictive Modelling, an example

- A bank from a developing country wants to give credits to consumers
  - There is very limited data on consumers
  - The bank talk with the local telecom company
  - The paying behaviour for the telecom company is actually a great predictive indicator for the credit behaviour with the bank
  - They bought the data, appended that to the bank data

Predictive Modeling

- The model is developed using a supervised learning approach

- Two phases:
  - **Training**: builds a model using a large sample of historical data called a training set.

  - **Testing**: involves trying out the model on new, previously unseen data to determine its accuracy and physical performance characteristics.

Predictive Modelling

- Applications of predictive modelling include customer retention management, credit approval, cross selling, and direct marketing.

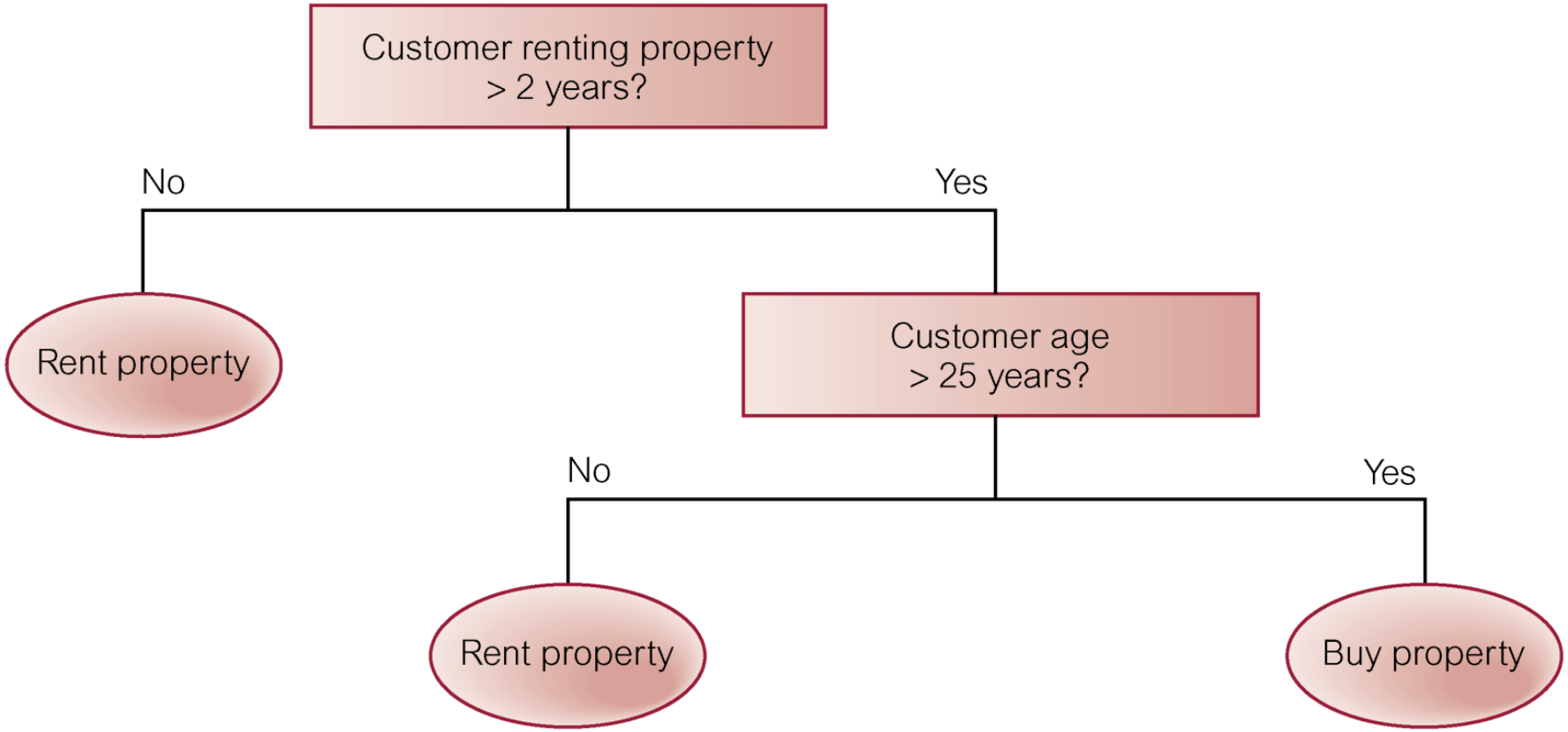
- Two techniques associated with predictive modelling:
  - Classification
  - Value prediction

Predictive Modelling - Classification

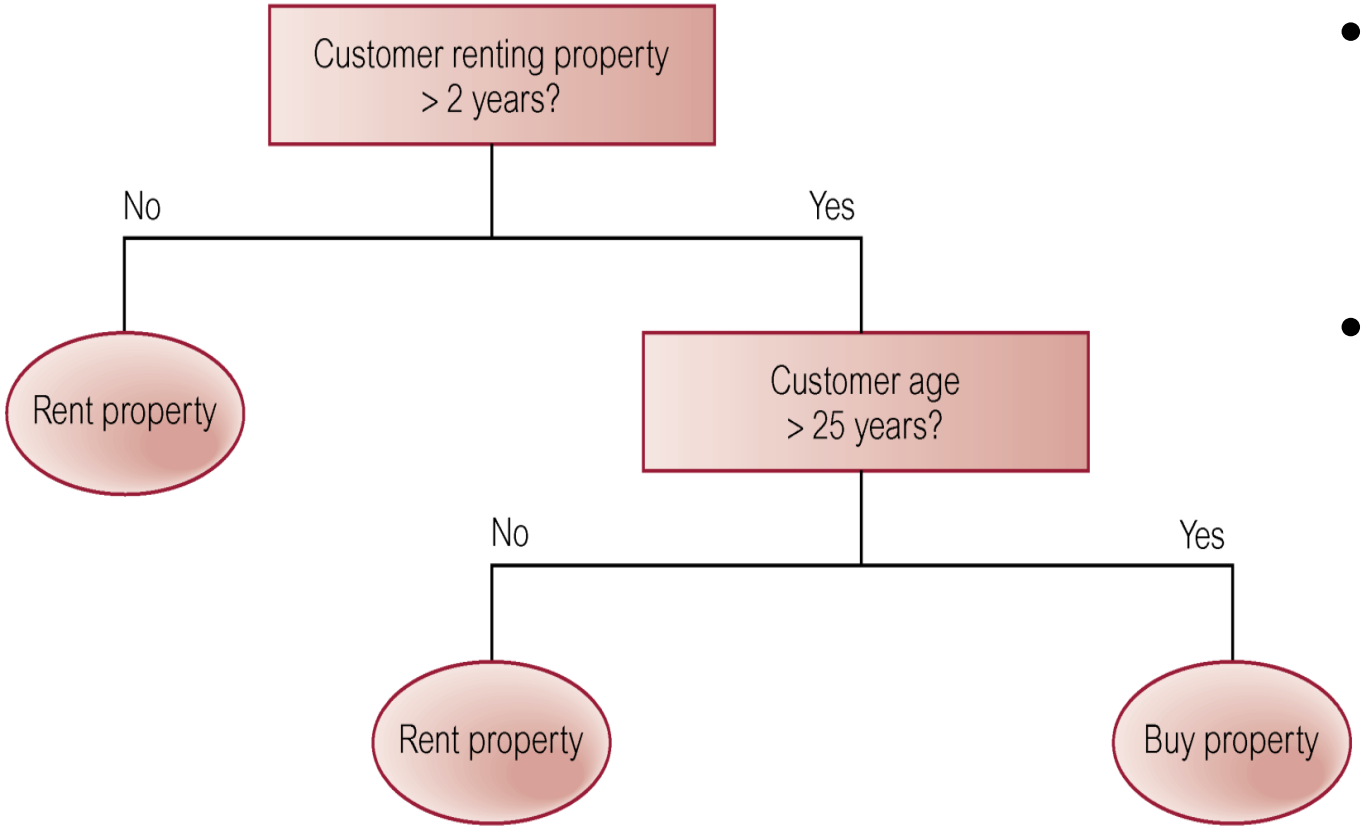
- Used to establish a specific predetermined class for each record in a database from a finite set of possible, class values.

- Two specializations of classification:
  - tree induction
  - neural induction.

Example of Classification using Tree Induction



Example of Classification using Tree Induction



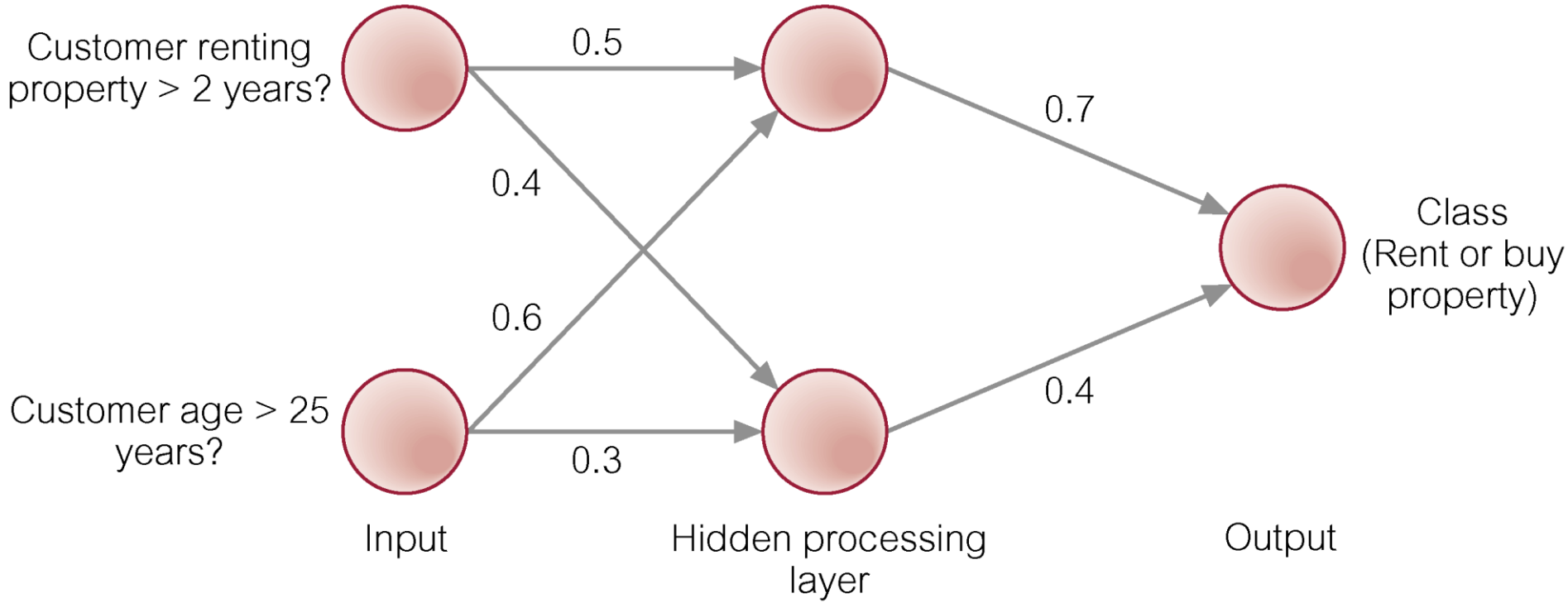
- The decision tree presents the analysis in an intuitive way.
- The model predicts that:
  - customers who have rented for more than two years and
  - are over 25 years old
  - are the most likely to be interested in buying property.

Example of Classification using Neural Induction

- The network attempts to mirror the way the human brain works in recognizing patterns by arithmetically combining all the variables associated with a given data point.

- A neural network contains:
  - A collections of connected nodes
  - There is input, output, and processing at each node.
- Between the visible input and output layers may be a number of hidden processing layers.
- Each processing unit (circle) in one layer is connected to each processing unit in the next layer by a weighted value (= strength of the relationship).

Example of Classification using Neural Induction



- Between the visible input and output layers may be a number of hidden processing layers.
- Each processing unit (circle) in one layer is connected to each processing unit in the next layer by a weighted value (= strength of the relationship).

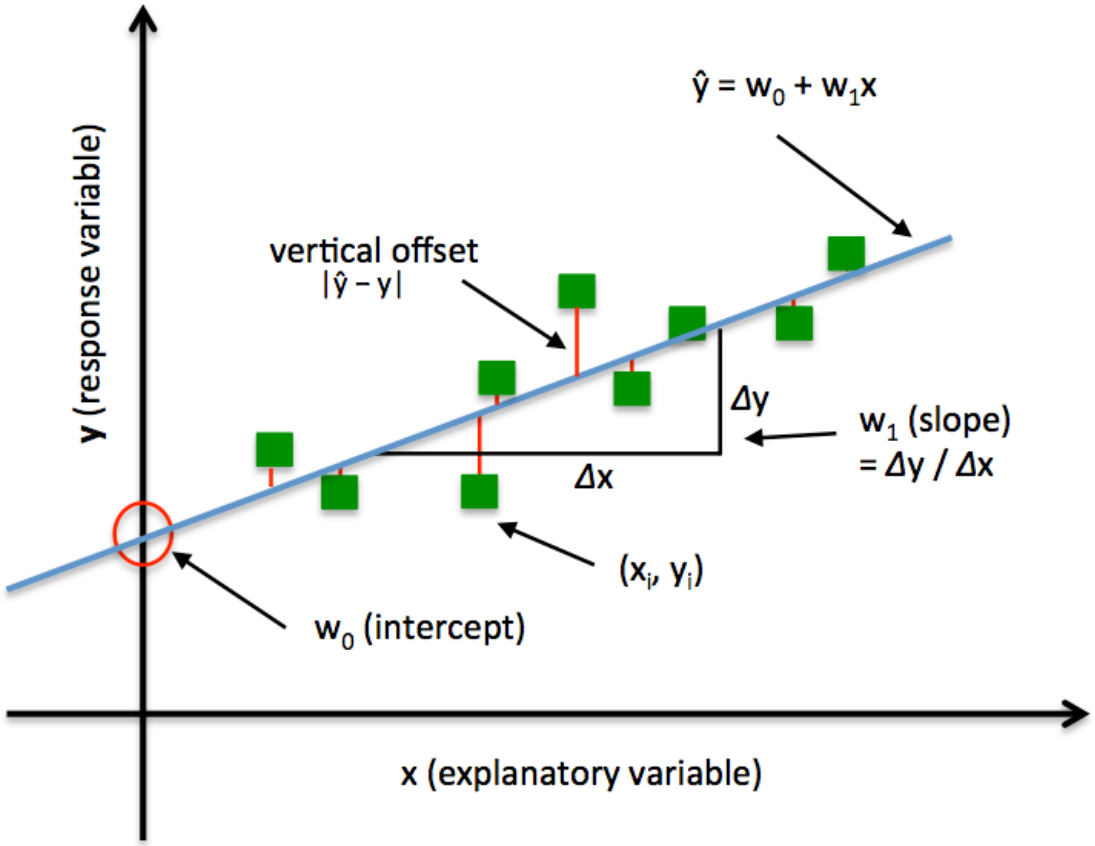
Predictive Modelling - Value Prediction

- Used to estimate a continuous numeric value that is associated with a database record.
- Uses the traditional statistical techniques of linear/non-linear regression.
- Relatively easy-to-use and understand.

Linear regression

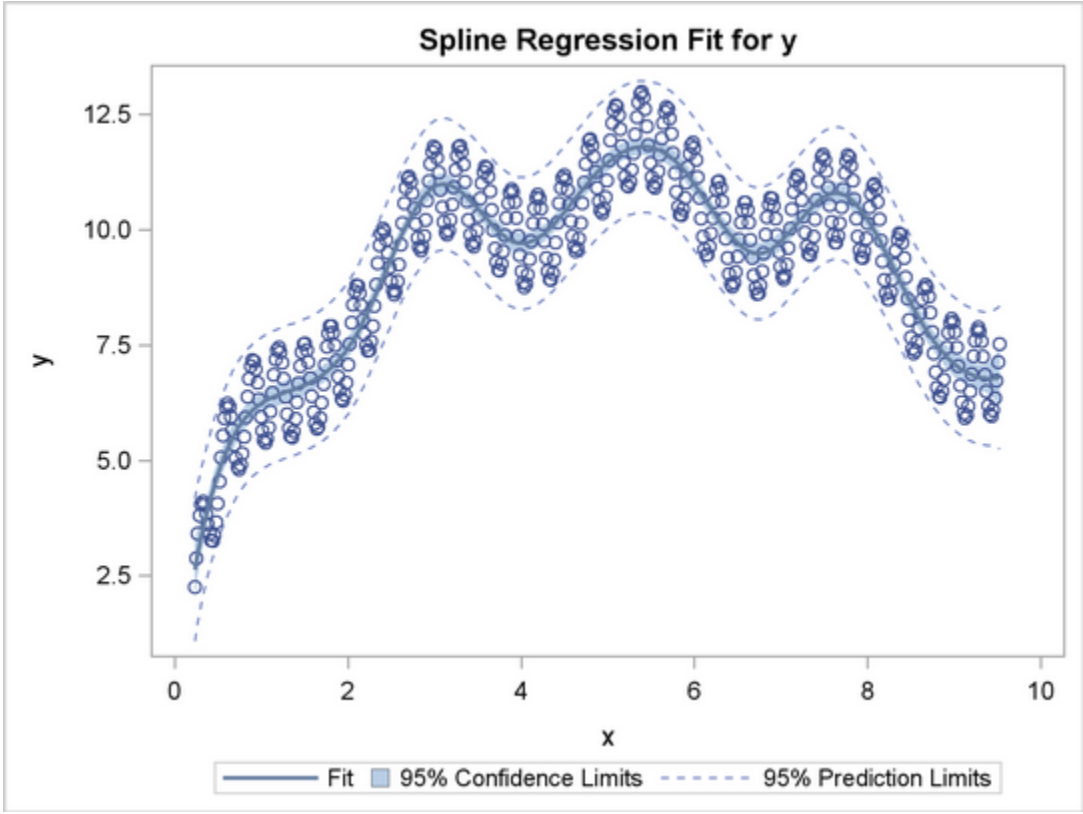
- Attempts to fit a straight line through a plot of the data, such that the line is the best representation of the average of all observations at that point in the plot.

- **Problem:**
  - only works well with linear data
  - sensitive to the presence of outliers.



Non-linear regression

- **Nonlinear regression:** avoids the main problems of linear regression, but not flexible enough to handle all possible shapes of the data plot.
- Statistical measurements are fine for building linear models that describe predictable data points, however, most data is not linear in nature.



Predictive Modelling - Value Prediction

- Data mining requires statistical methods that can accommodate non-linearity, outliers, and non-numeric data.

- Applications of value prediction include credit card fraud detection or target mailing list identification.

Database segmentation

- Uses unsupervised learning to discover homogeneous sub-populations in a database to improve the accuracy of the profiles.

- Less precise than other operations → less sensitive to redundant and irrelevant features.

- Associated with demographic or neural clustering techniques

Link Analysis

- Establishing links, called associations, between the individual records, or sets of records, in a database.

- Types of link analysis:
  - associations discovery ,
  - sequential pattern discovery , and
  - similar time sequence discovery

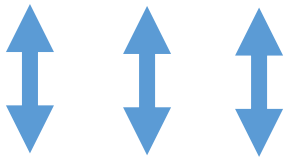
Deviation Detection

- Identifies outliers, which express deviation from some previously known expectation and norm.

- Can be performed using statistics and visualization techniques, e.g., using regression to identify outliers.

Data Mining and Data Warehousing

- A data warehouse is well equipped for providing data for mining:
  - Data quality and consistency is a pre-requisite for mining to ensure the accuracy of the predictive models.



  - Data warehouses are populated with clean, consistent data.

Data Mining and Data Warehousing

- It is advantageous to mine data from multiple sources to discover as many interrelationships as possible. Data warehouses contain data from a number of sources.

- Selecting the relevant subsets of records and fields for data mining requires the query capabilities of the data warehouse.

- The results of a data mining study are useful if there is some way to further investigate the uncovered patterns. Data warehouses provide the capability to go back to the data source.

So is data mining = data warehouse??

Data warehouse relates to the **storage** of the data used for data mining