DATA MINING
Data & Attributes

Dr. Edi Surya Negara,M.Kom.

Postgraduate Program, Informatics Engineering (S2)

April 23, 2018

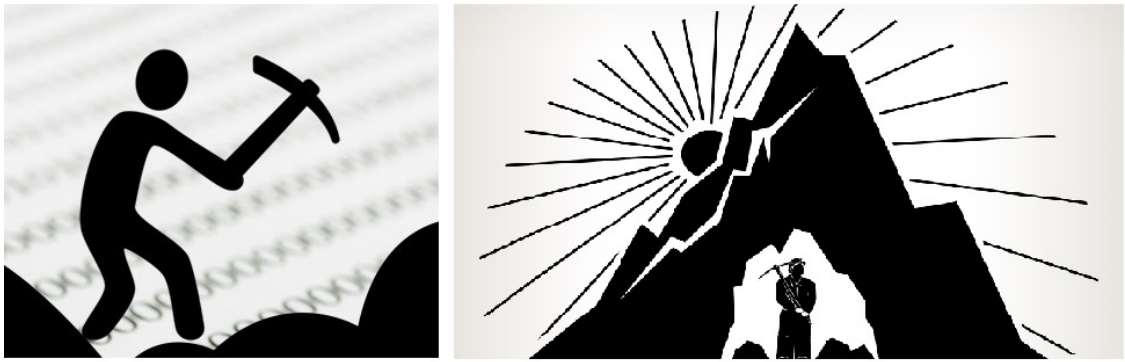
Chapter 1: Objectives

After completing this chapter, students will be able to:

- Explain what is data mining.
- Explain task of data mining.
- Explain benefit of data mining.
- Explain process of data mining.
- Explain about data, attributes, similarity, and dissimilarity data.

Reference: Tan et al: Introduction to data mining. Some slides are adopted from Tan et al.

What is data mining?



- 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.

What is data mining (cont.)?

- Valid: The patterns hold in general.
- Novel: We did not know the pattern beforehand.
- Useful: We can devise actions from the patterns.
- Understandable: We can interpret and comprehend the patterns.

Why Use Data Mining Today?

- Human analysis skills are inadequate:
  - Volume and dimensionality of the data
  - High data growth rate
- Availability of:
  - Data
  - Storage
  - Computational power
  - Off-the-shelf software
  - Expertise

Sources of Data

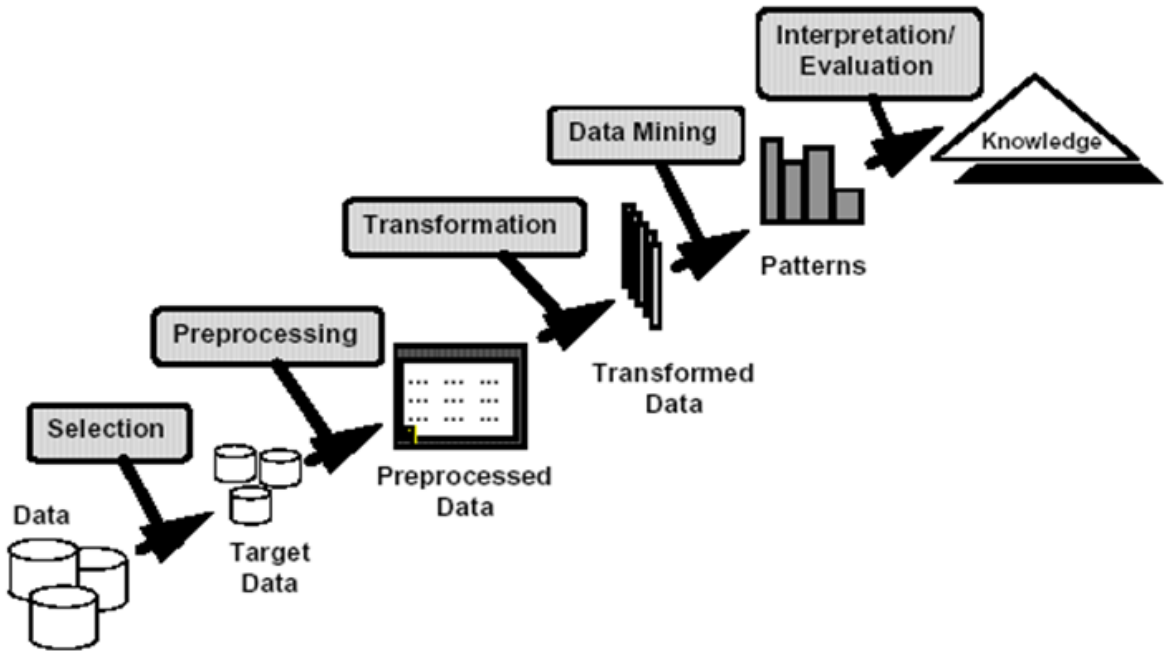
- Supermarket scanners, POS data
- Credit card transactions
- Direct mail response
- Call center records
- ATM machines
- Demographic data
- Sensor networks
- Cameras
- Web server logs
- Customer web site trails

The Knowledge Discovery Process

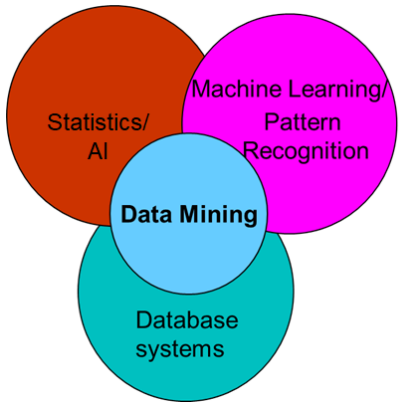
Steps:

- Identify business problem
- Data mining
- Action
- Evaluation and measurement
- Deployment and integration into businesses processes

The data mining process



Origins of Data Mining



- Draws ideas from machine learning/Artificial Intelligence, pattern recognition, statistics, and database systems.
- Human analysis and traditional Techniques may be unsuitable due to: Enormity of data, High dimensionality of data, Heterogeneous data, Distributed nature of data.

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks..

- Association Rule Discovery [Descriptive]
- Clustering [Descriptive]
- Classification [Predictive] (for discrete variables)
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive] (for continuous variable)
- Deviation Detection [Predictive]

Association Rule Discovery: Definition

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

- Given a set of records each of which contain some number of items from a given collection;
- Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

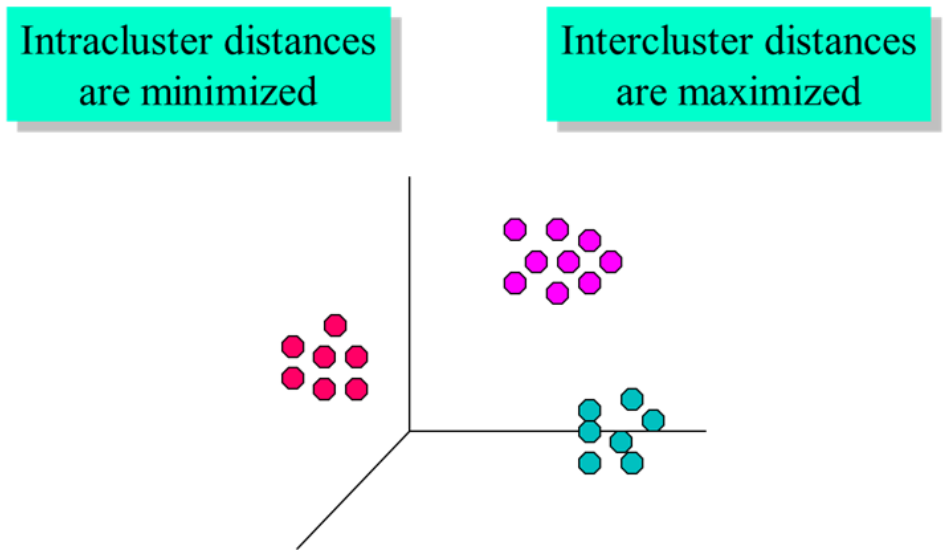
Association Rule Discovery: Sample Application

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.

Clustering: Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering



- Euclidean Distance Based Clustering in 3-D space.

Clustering: Sample Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Sample Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

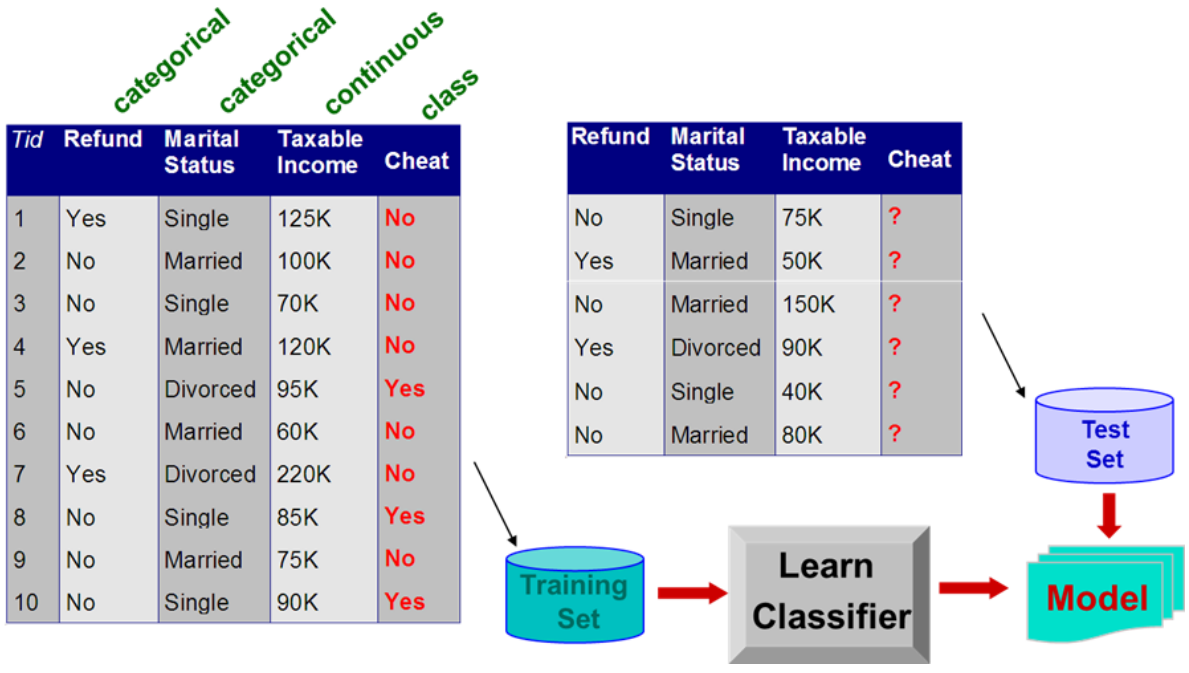
Classification: Definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification and Clustering

- Classification:
  - Classes pre-defined
  - Uses training set (thus also known as supervised learning)
- Clustering:
  - Classes not defined in advance
  - No training set (thus also known as unsupervised learning)

Classification: Example



Classification: sample application 1

- Direct Marketing:
  - Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.
  - Approach:
  - Use the data for a similar product introduced before.
  - We know which customers decided to buy and which decided otherwise. This buy, dont buy decision forms the class attribute.
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers. Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model.

Classification: sample application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
  - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions.
  - Use this model to detect fraud by observing credit card transactions on an account.

Sequential Pattern Discovery: Definition

- Given a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.
- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior.
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection

Universitas Bina Darma

Road to 2025

IT Based International Standard