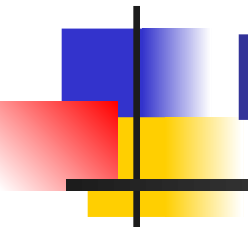


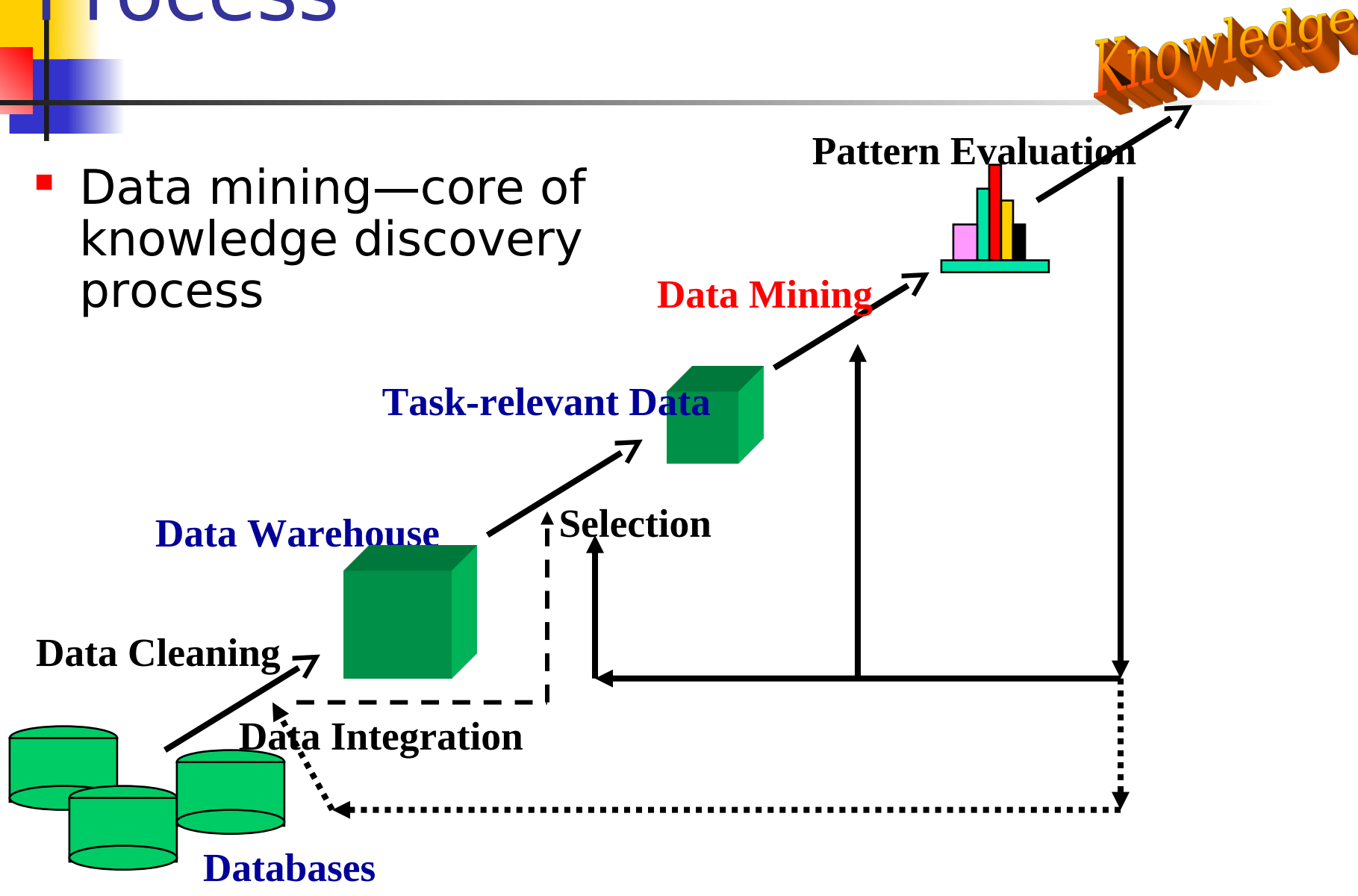
# Introduction to Data Mining

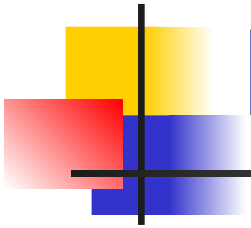
Dr. Bernard Chen

Troy University

# Knowledge Discovery (KDD) Process

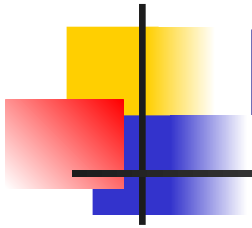
- Data mining—core of knowledge discovery process





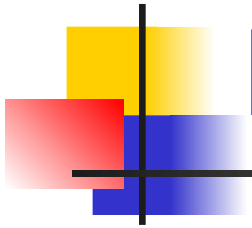
# Knowledge Process

1. **Data cleaning** – to remove noise and inconsistent data
2. **Data integration** – to combine multiple source
3. **Data selection** – to retrieve relevant data for analysis
4. **Data transformation** – to transform data into appropriate form for data mining
5. **Data mining**
6. **Evaluation**
7. **Knowledge presentation**



# Knowledge Process

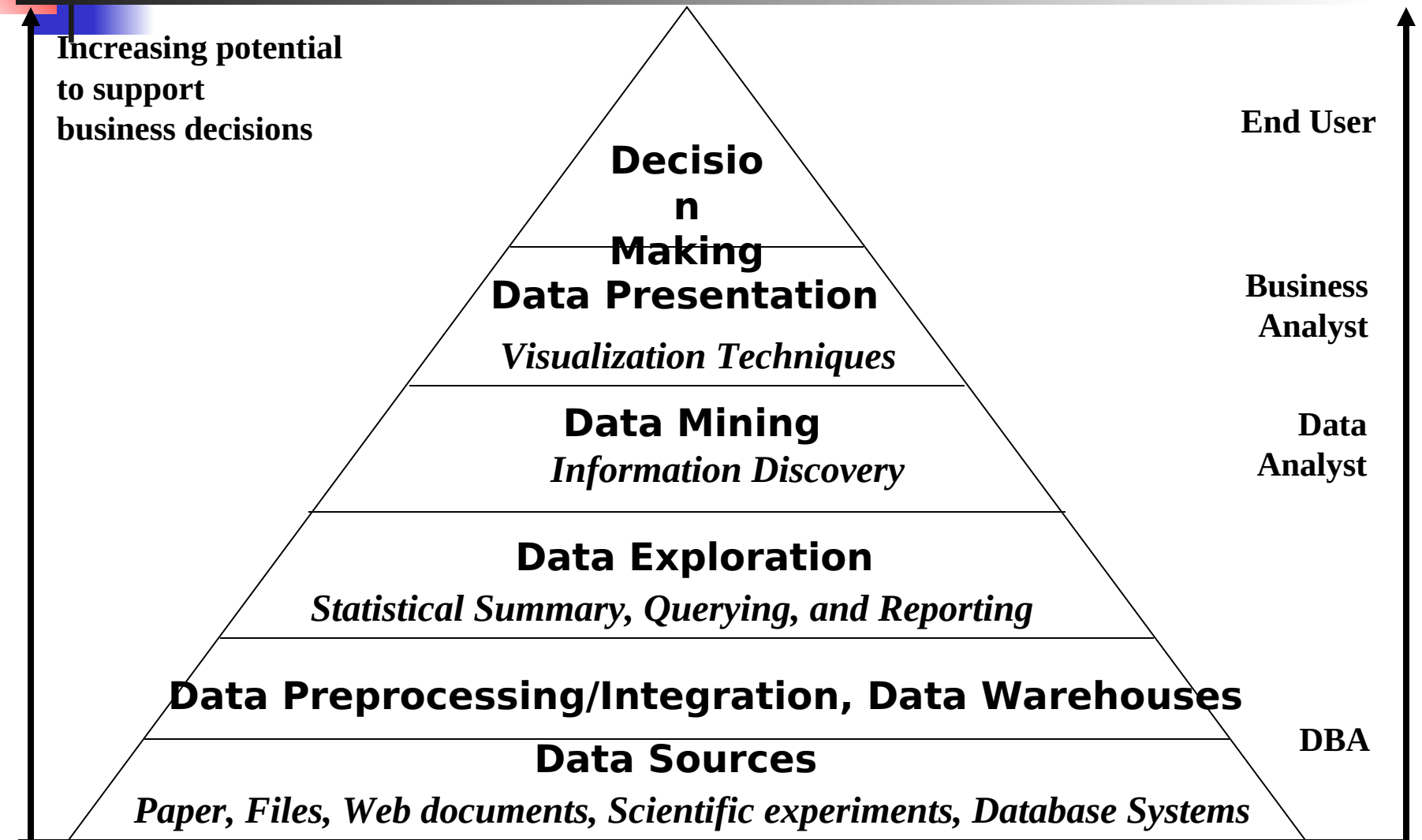
- Step 1 to 4 are different forms of data **preprocessing**
- Although data mining is only one step in the entire process, it is an essential one since it uncovers hidden patterns for evaluation



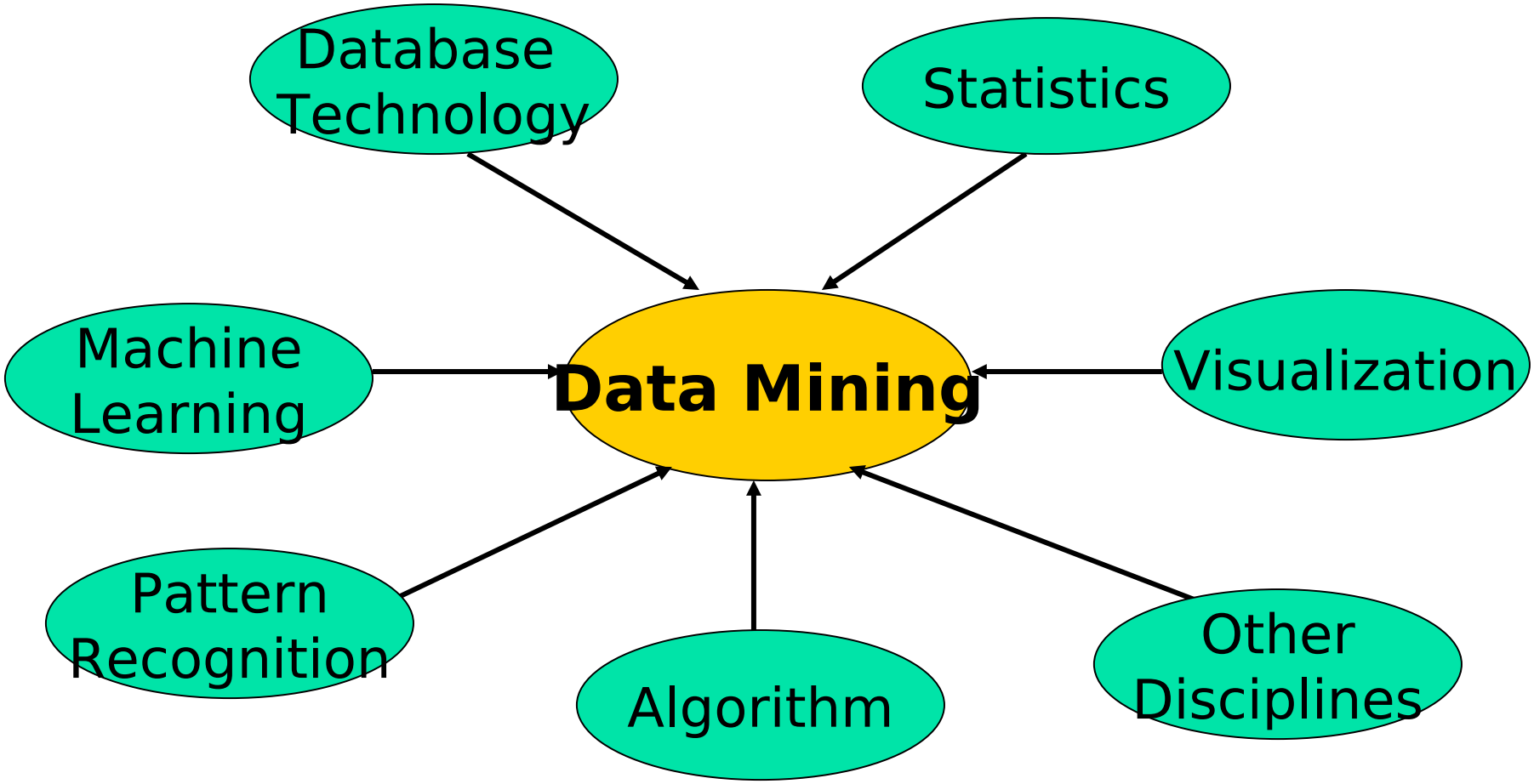
# Knowledge Process

- Based on this view, the architecture of a typical data mining system may have the following major components:
  - Database, data warehouse, world wide web, or other information repository
  - Database or data warehouse server
  - Data mining engine
  - Pattern evaluation model
  - User interface

# Data Mining and Business Intelligence

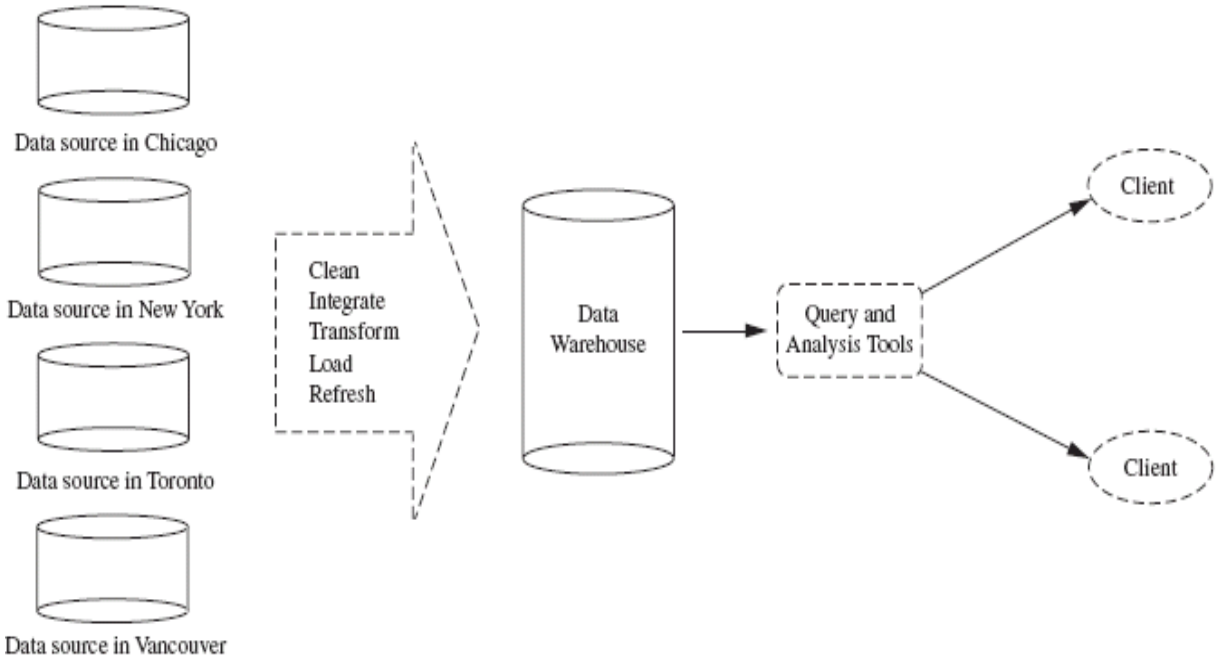


# Data Mining: Confluence of Multipl Disciplines



# Data Mining – on what kind of data?

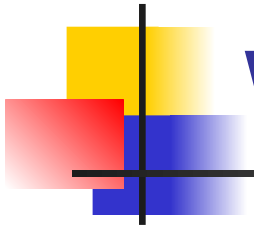
- Data Warehouses



# Data Mining – on what kind of data?

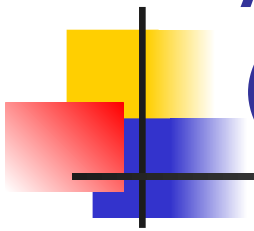
- Transactional Databases
- Advanced data and information systems
  - Object-oriented database
  - Temporal DB, Sequence DB and Time serious DB
  - Spatial DB
  - Text DB and Multimedia DB
  - … and WWW

# What kind of patterns can we mined?



- In general, data mining tasks can be classified into two categories: **descriptive** and **predictive**
  - Descriptive mining tasks characterize the general properties of the data in database
  - Predictive mining tasks performs inference on the current data in order to make predictions

# Mining frequent patterns, Associations, and Correlations (Ch4)

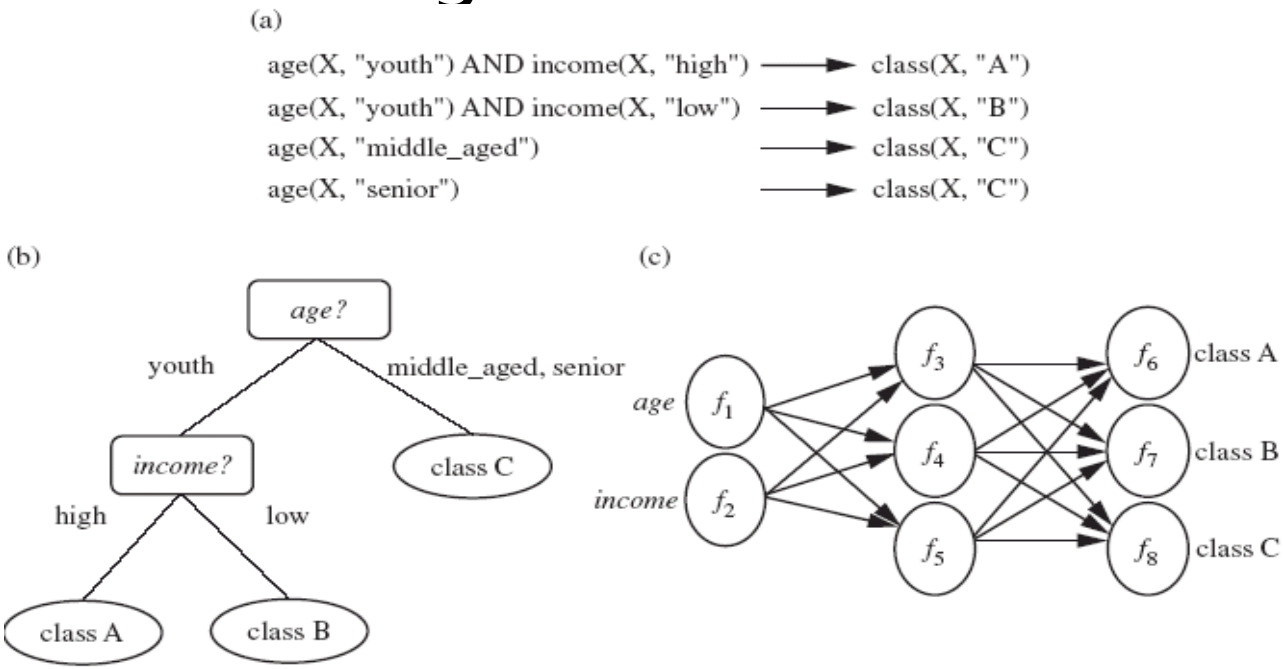


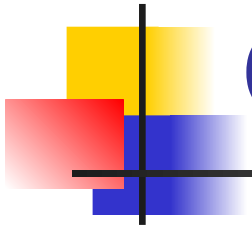
- **Frequent patterns** are patterns that occur frequently in data

- Association analysis:
  - Example: buys(X,”computer”) => buys(X,”software”)  [support = 1%, confidence = 50%]

# Classification and Prediction (Ch 5)

- **Classification** is the process of finding a MODEL that describes and distinguish data classes or



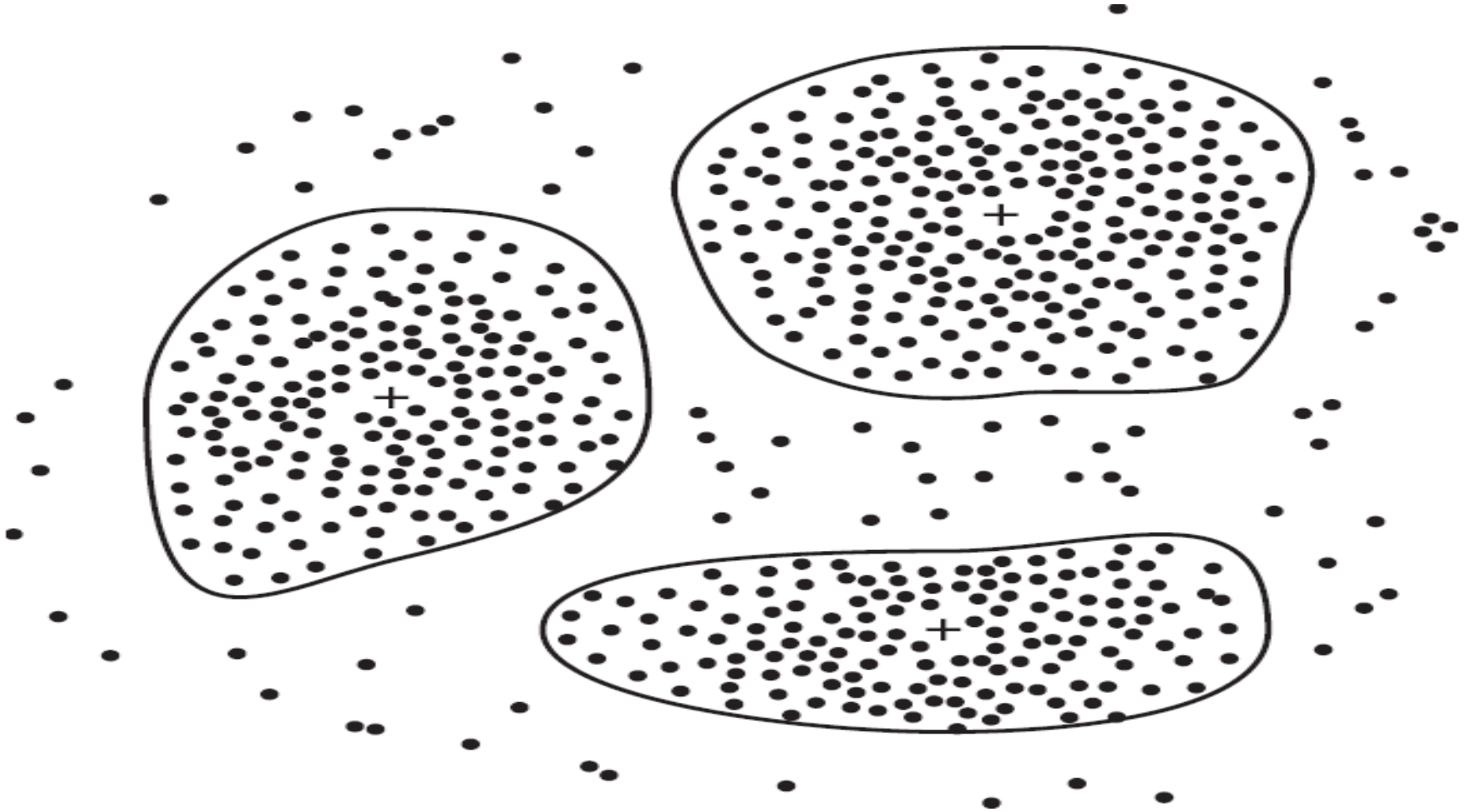


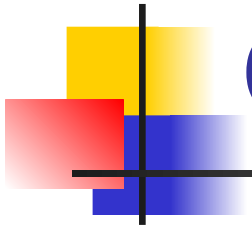
# Cluster analysis (Ch 6)

- In general, the class label are not present in the training data simply they are not known to begin with

- The objects are clustered or grouped based on the principle of *maximizing the intra-cluster similarity* and *minimizing the inter-cluster similarity*

# Cluster analysis





# Outlier Analysis (Ch 7)

- Most data mining methods discard outliers as noise or exceptions.
- However, in some application such as fraud detection, the rare event can be more interesting than regularly occurring ones