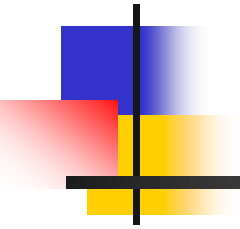


Data Mining

Yi-Cheng Chen (陳以錚)

Dept. of Computer Science & Information Engineering,

Tamkang University

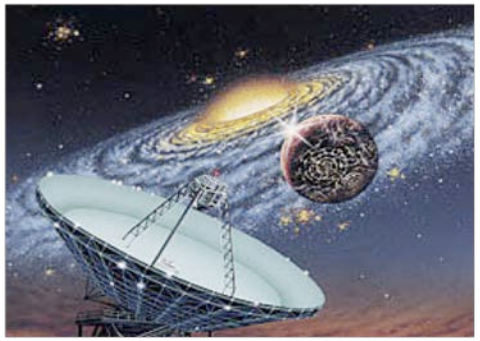
Why Mine Data? Commercial Viewpoint

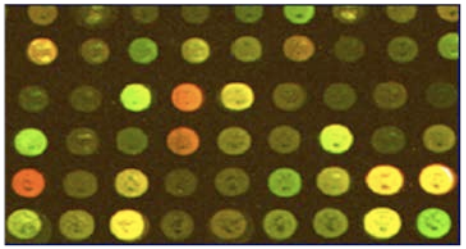
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/ grocery stores
  - Bank/Credit Card transactions

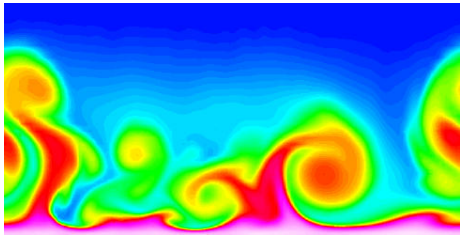
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

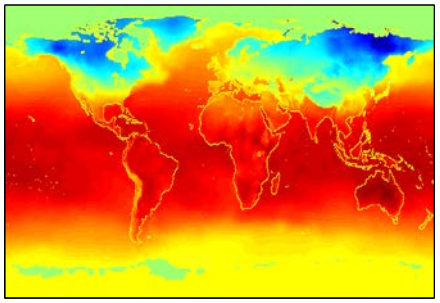
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



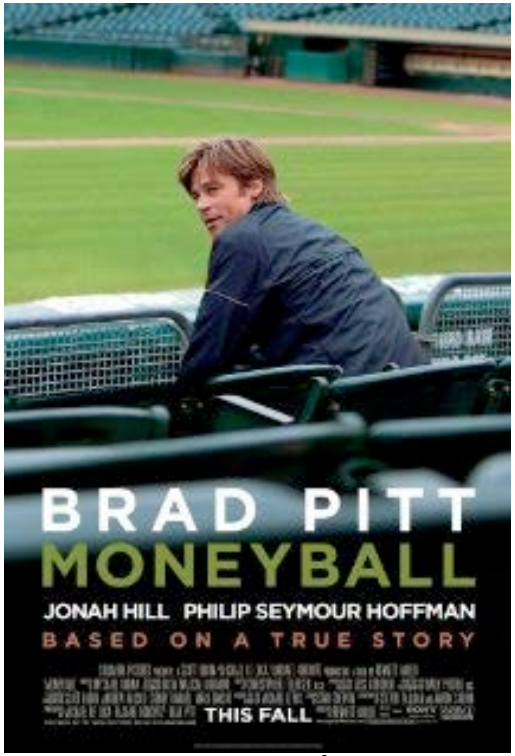


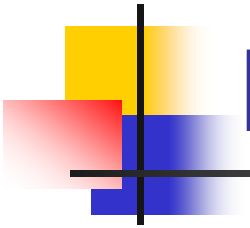




Motivation

- **We are data rich but information poor**

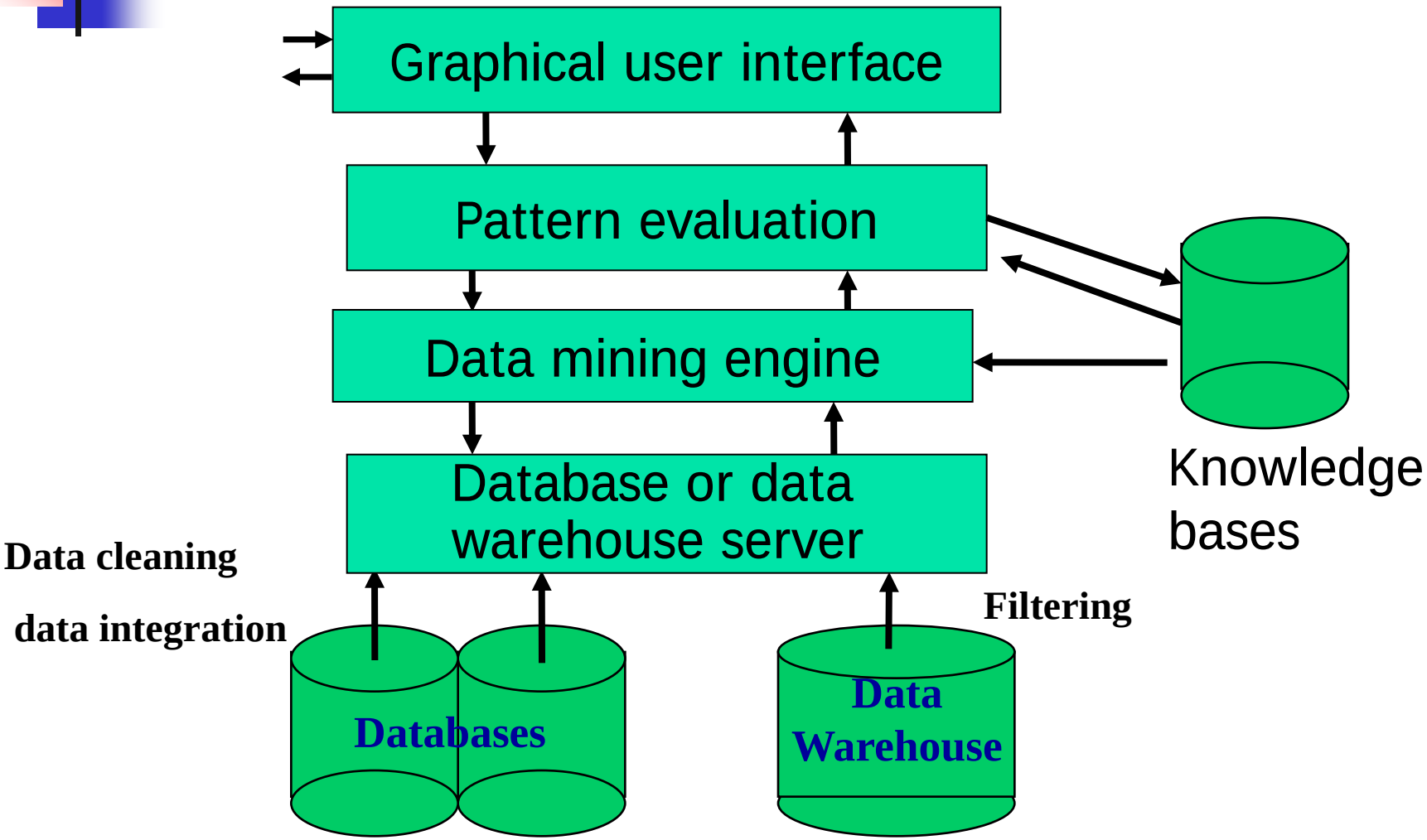




Data Mining

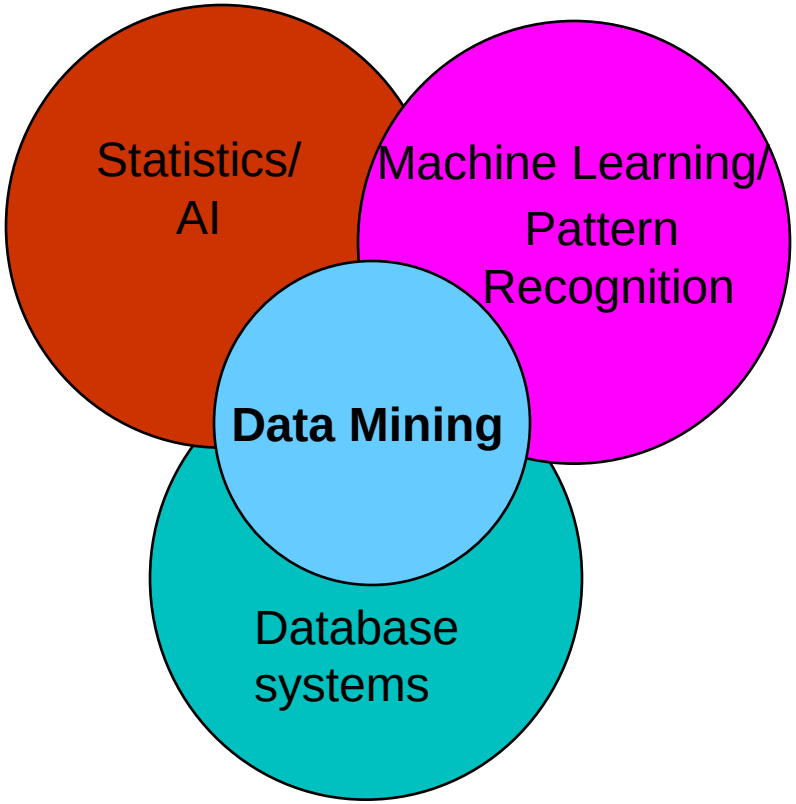
- We are buried in data, but looking for knowledge
- Data mining:Knowledge discovery in databases
  - Extraction of interesting knowledge (rules, regularities, patterns) from data in large databases

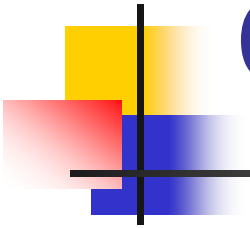
Architecture: Typical Data Mining System



Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



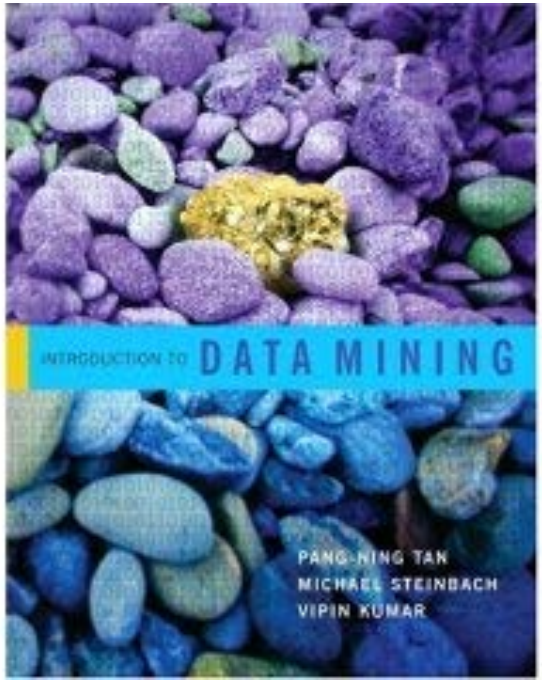


Course Materials

- Introduction to data mining
- Mining association rules
- Mining sequential patterns
- Data classification
- Data clustering
- Web mining
- Stream data mining
- Mining in social network
- Big data/ cloud data mining

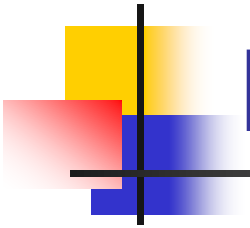
Textbook

- Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Addison-Wesley
- (Reference) Data Mining: Concepts and Techniques, J. Han and M. Kamber, Morgan Kaufmann
- Paper Collection



Evaluation of Database Technology

- 1960s: data collection, database creation
- 1970s: relational model
- 1980s: advanced data model
- 1990s: data mining & data warehousing, digital library, Web databases
- 2000s: Stream data management and mining, data mining with a variety of applications, Web technology and global information systems
- 2010s: Big data and cloud data mining



Data Mining

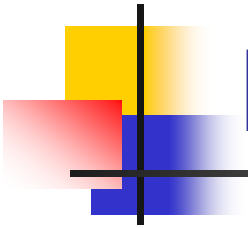
- We are buried in data, but looking for knowledge
- Data mining:Knowledge discovery in databases
  - Extraction of interesting knowledge (rules, regularities, patterns) from data in large databases

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)

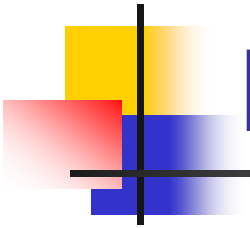
Why Data Mining?—Potential Applications (cont'd)

- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - DNA and bio-data analysis
  - Mining in social network
  - Multimedia or sensor network



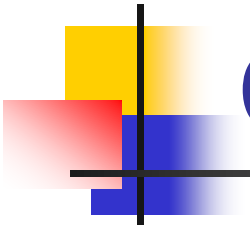
Notes

- Data mining is very application dependent
  - Small team with good skills and domain knowledge
- Emerging issues:
  - Journals: IEEE TKDE, ACM TKDD, KAIS
  - Conferences: ACM SIGKDD, SIGMOD, CIKM
    IEEE ICDM, ICDE
    SDM, VLDB, PAKDD etc.



Data Warehousing

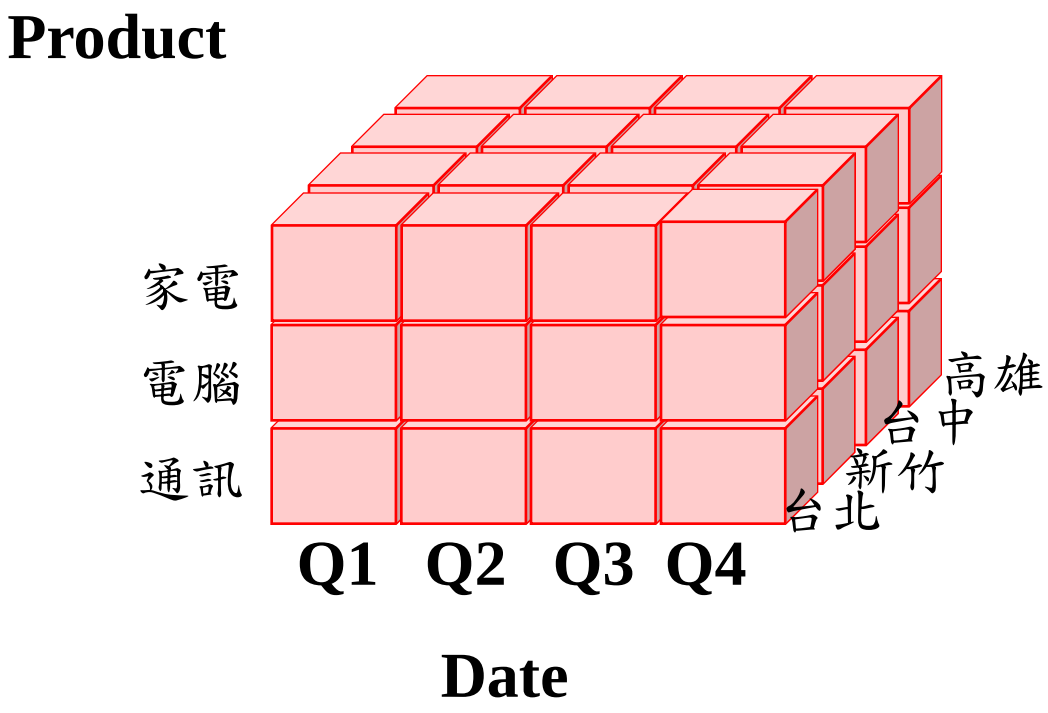
- An architectural foundation for decision support system, consisting of
  - Integrated data, detailed and summarized data, historical data, and metadata
- Set up stages for effective data mining

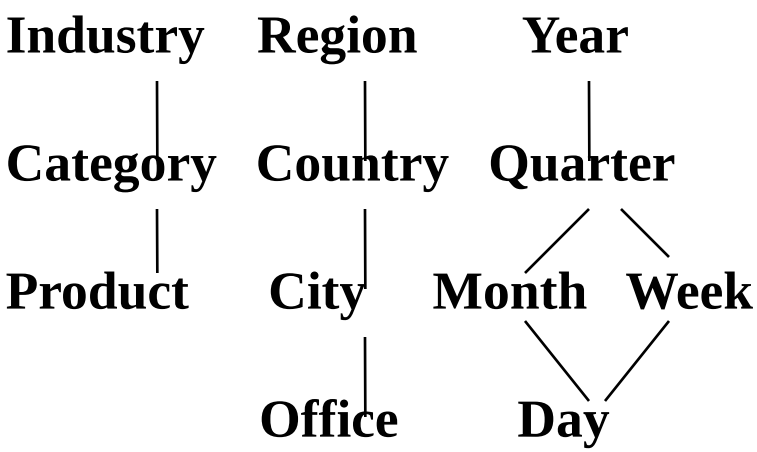


OLAP

- On-Line Analytical Processing: simple data mining facility
  - Responds to queries quickly
- A multidimensional, logical view of the data
- Interactive analysis of the data: drill down and roll-up, etc.

Data Cube


Product
家電
電腦
通訊
Q1
Q2
Q3
Q4
Date
台北
新竹
台中
高雄
City


Industry
Region
Year
Category
Country
Quarter
Product
City
Month
Week
Office
Day

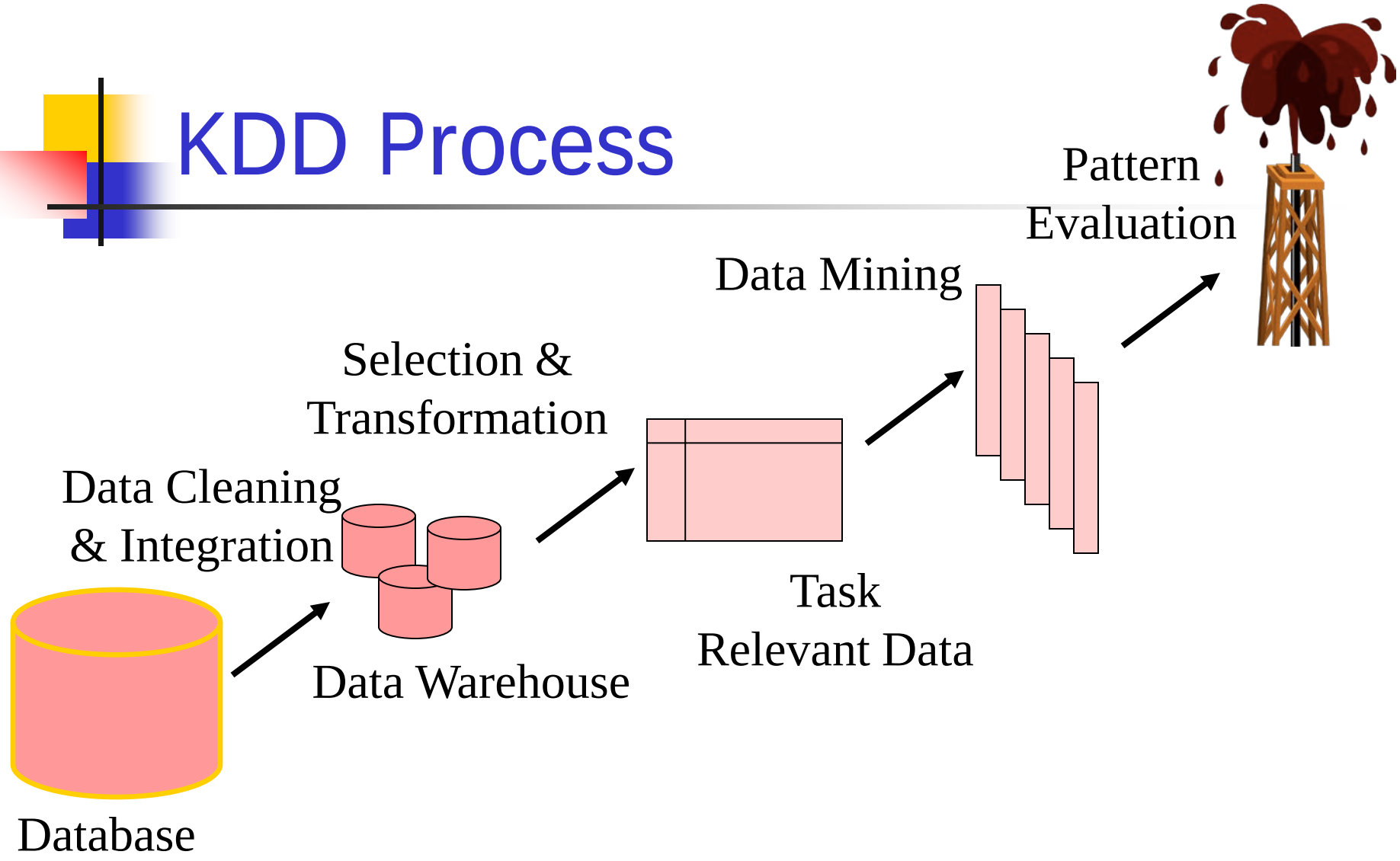
Knowledge Discovery from Databases

- Nontrivial process of extraction of
  - valid (with some degree of certainty)
  - novel (implicit, previously unknown)
  - potential useful
  - ultimately understandable
  - patterns from large collection of data
- Pattern
  - expression in languages describing subset of data
  - model (structure) applicable to subset of data

Similar Terms of KDD

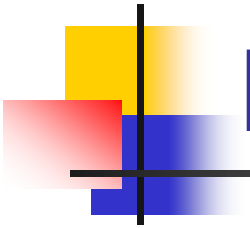
- Knowledge Discovery in Databases (KDD)
- Knowledge mining from databases
- Knowledge extraction
- Regularities
- Data analysis

KDD Process


Pattern
Evaluation
Data Mining
Selection &
Transformation
Data Cleaning
& Integration
Task
Relevant Data
Data Warehouse
Database

Classification of DM Techniques

- What kinds of databases to work on
- What kind of knowledge to be mined
- What kind of techniques to be utilized



Databases to Work on

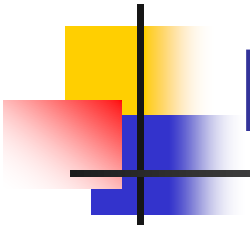
- Relational
- Transactional
- Object-oriented
- Spatial
- Temporal
- Multimedia

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.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996



Knowledge to Be Mined

- Association rules
- Classification
- Clustering
- Trend and deviation analysis
- Outlier

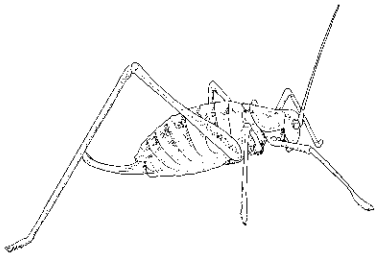
Association Rules

- Buy(bread) ^ Buy(milk) => Buy(butter)
- Age(20~29) ^ Income(20~30k) => Buy(CD player)

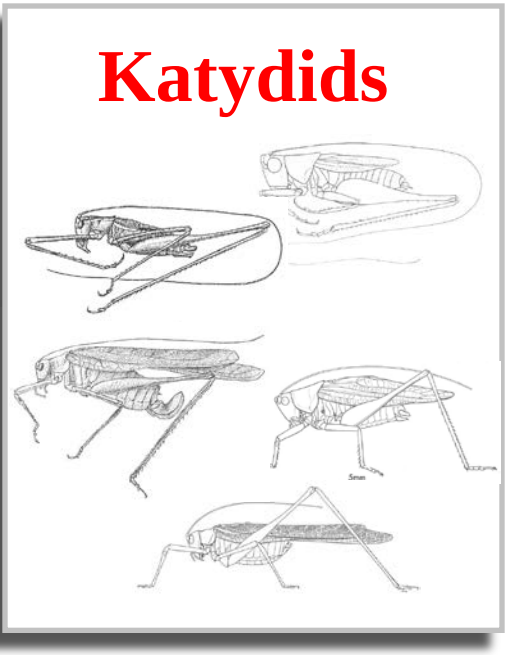
The Classification Problem

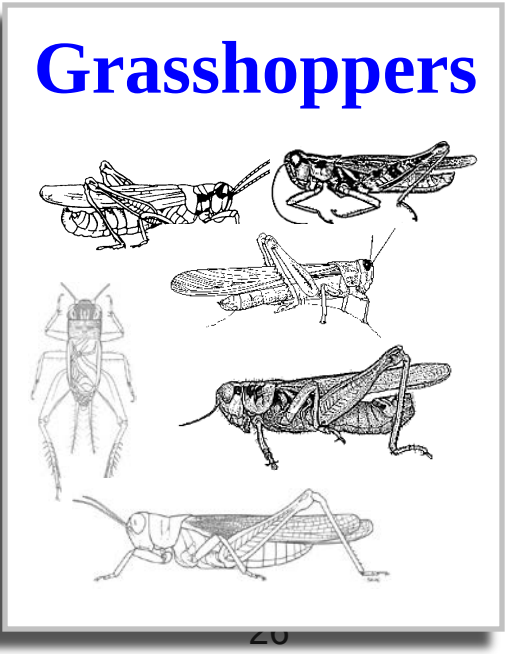
(informal definition)

Given a collection of annotated data. In this case 5 instances **Katydids** of and five of **Grasshoppers**, decide what type of insect the unlabeled example is.

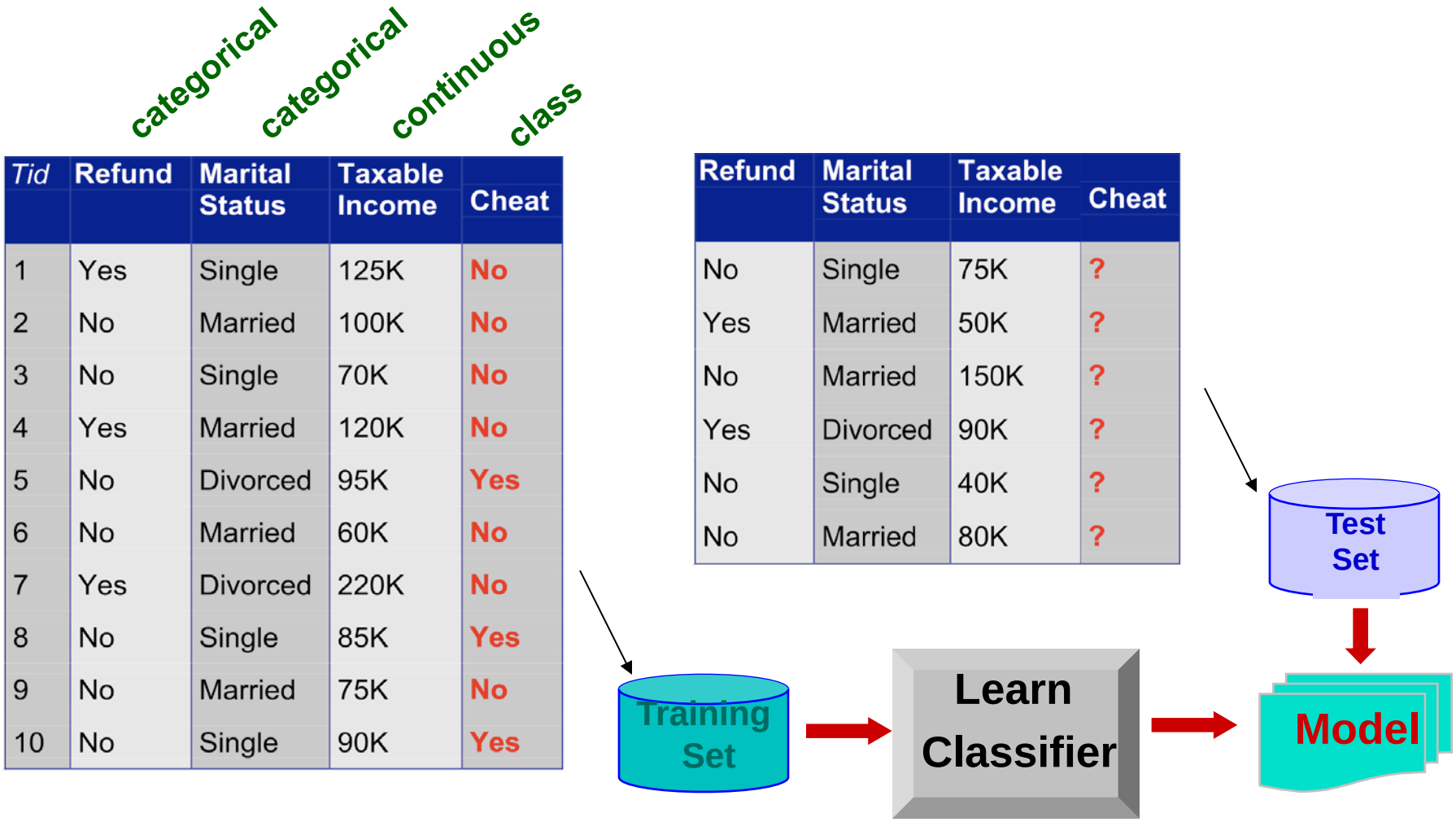


Katydid or **Grasshopper**?



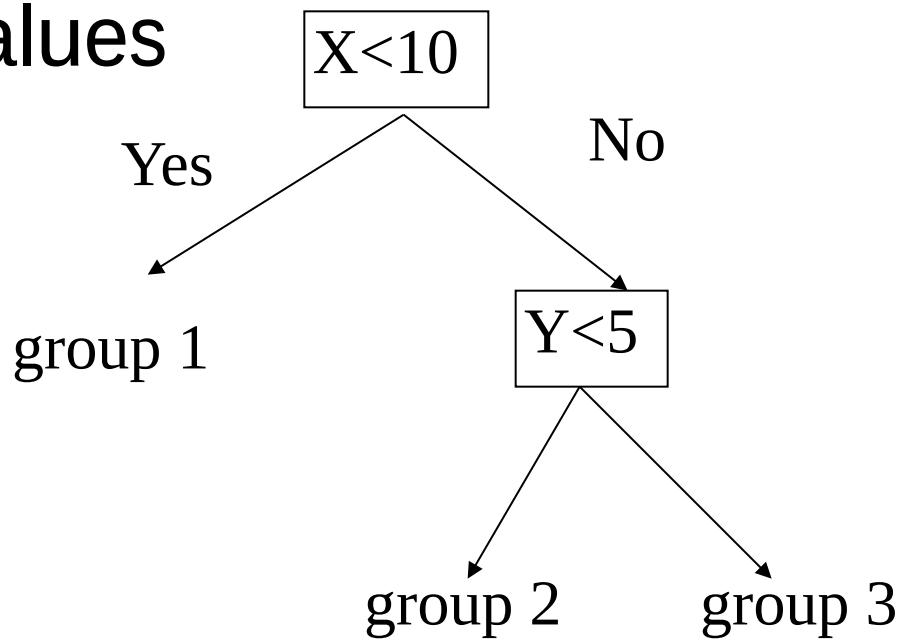


Classification Example



Classification

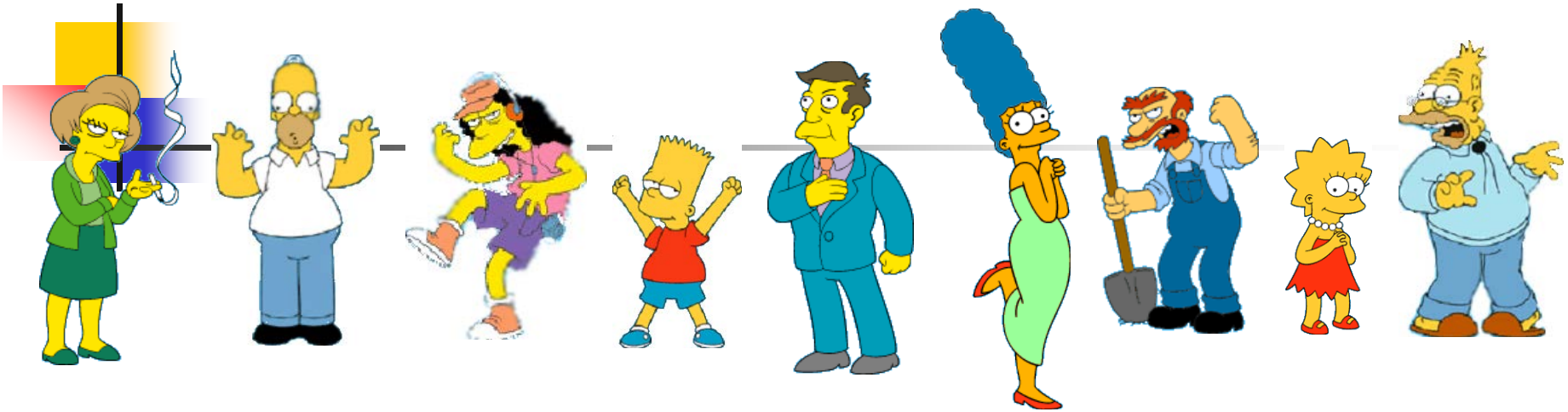
- Supervised classification
- Organizes data into given classes based on attribute values



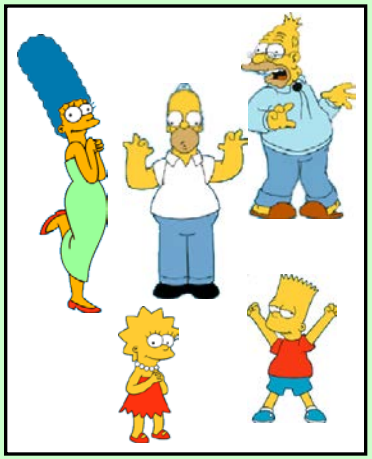
Classification: Application

- 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.

What is a natural grouping among these objects?

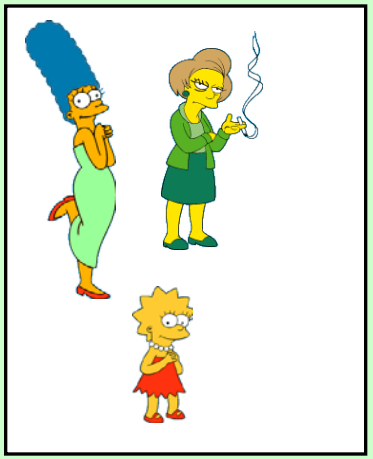


Clustering is subjective



Simpson's Family

School Employees

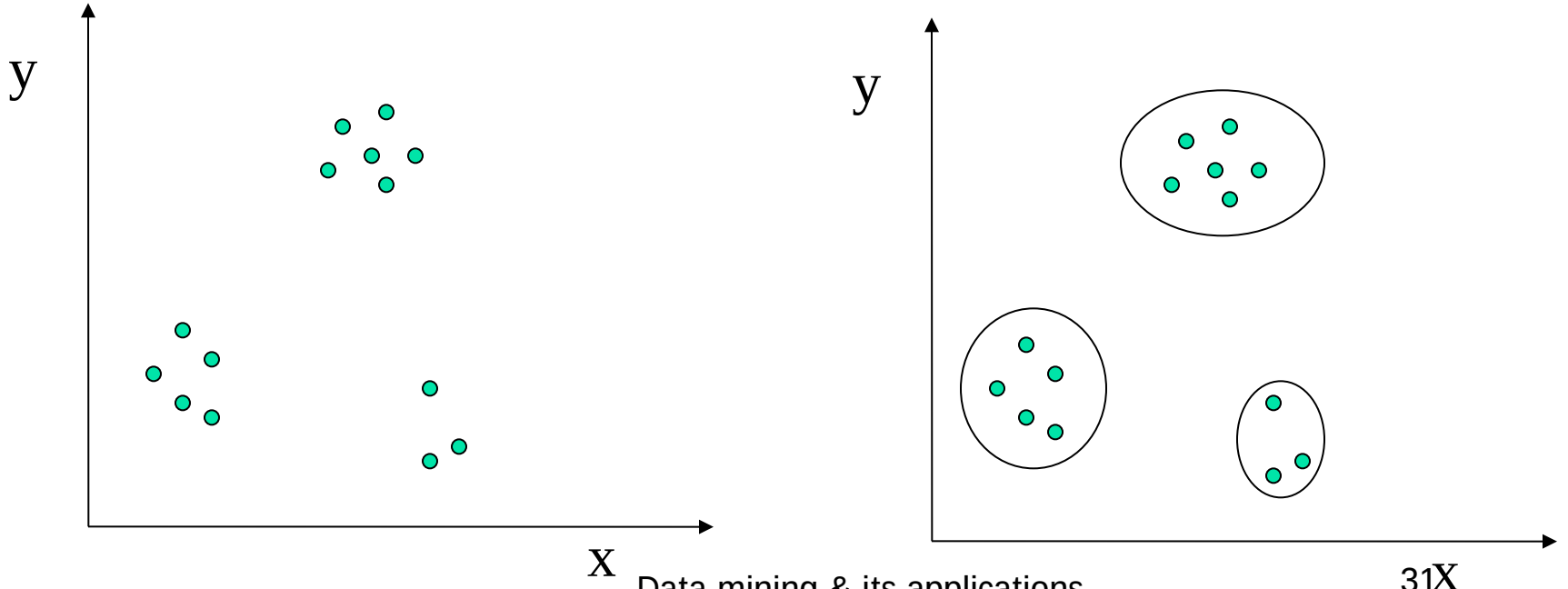


Females

Males

Clustering

- Unsupervised classification
- Organizes data into classes based on attribute values



Clustering: 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: 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.

Time Series Analysis

- Trends analysis
- Regression
- Sequential patterns
- Similar sequences

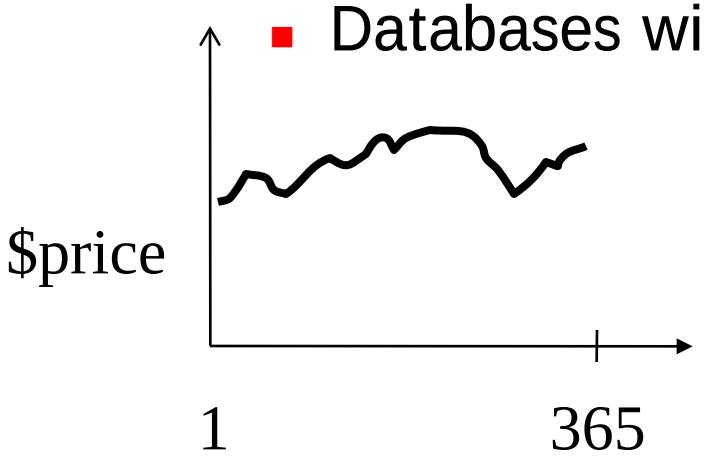
Sequential Patterns

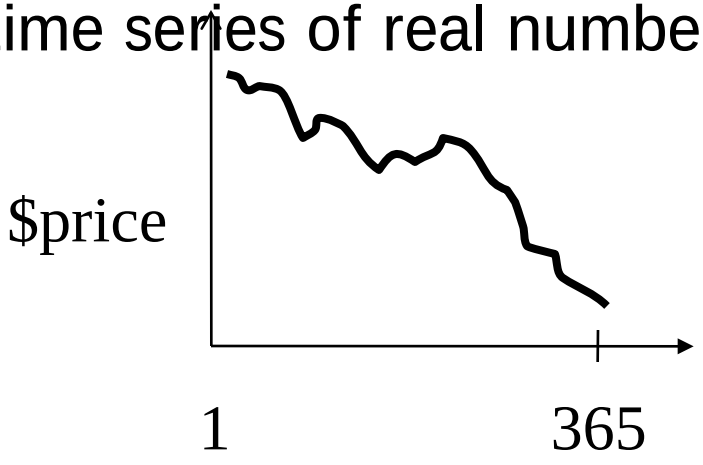
- Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.

(A B) (C) (D E)

Time Series Database

- Time series
  - Financial, marketing & production: stock price, sales number
  - Scientific: weather data, geological, astrophysics
- Time series DB
  - Databases with many time series of real numbers



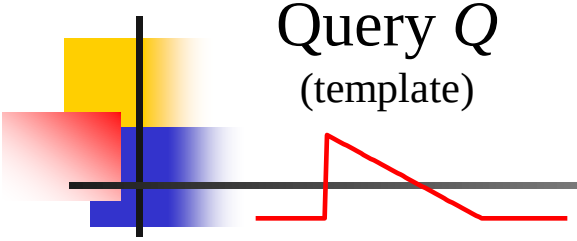


Time Series Database (cont'd)

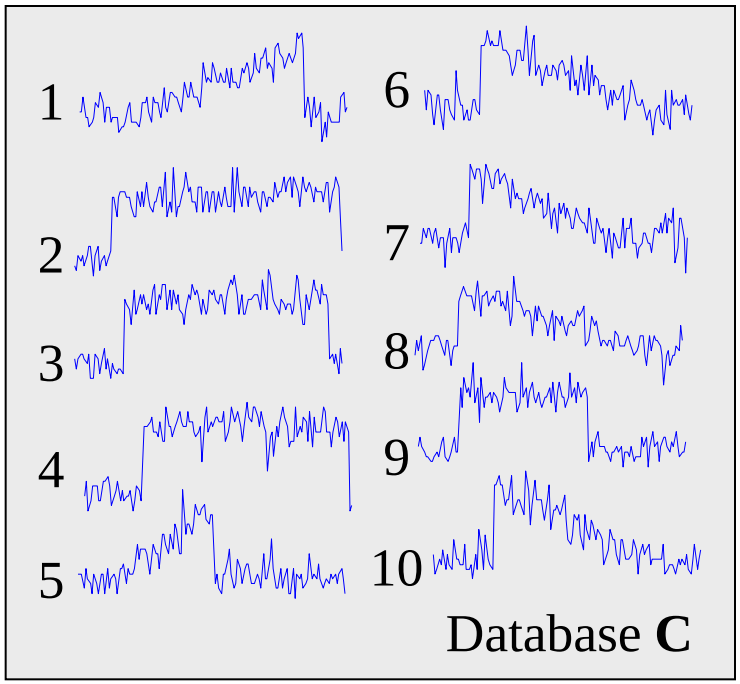
- Query in time series DB
  - Searching for similar patterns
  - Whole matching
  - Subsequence matching
  - Examples
    - Identify companies with similar pattern of growth
    - Determine products with similar selling patterns
    - Discover stocks with similar movement in stock prices
    - Find if a musical score is similar to one of the copyrighted scores

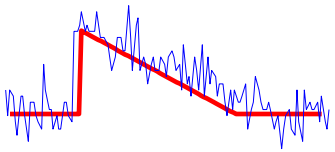
The similarity matching problem can come in two flavors I

Query *Q*
(template)



1: Whole Matching

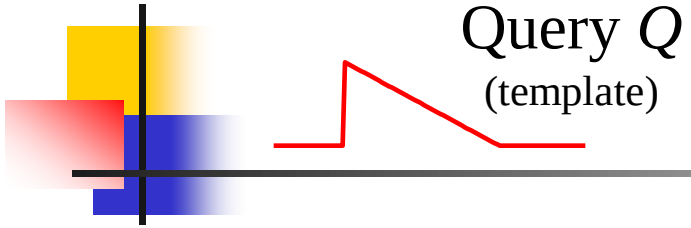




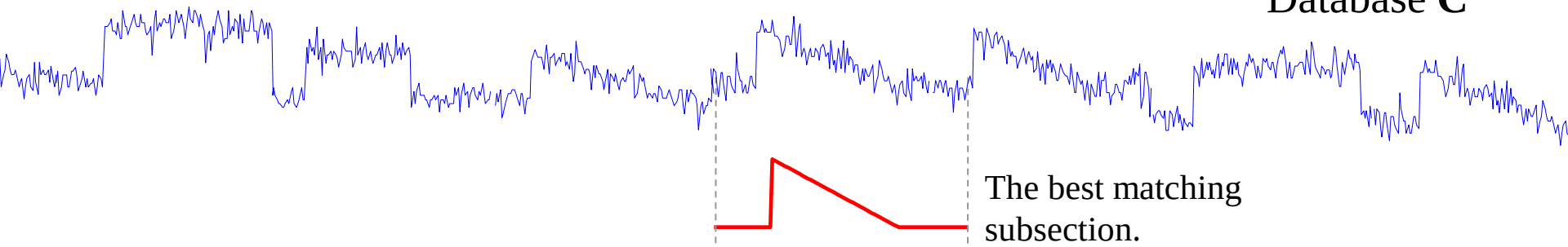
C_6 is the best match.

Given a Query *Q*, a reference database **C** and a distance measure, find the C_i that best matches *Q*.

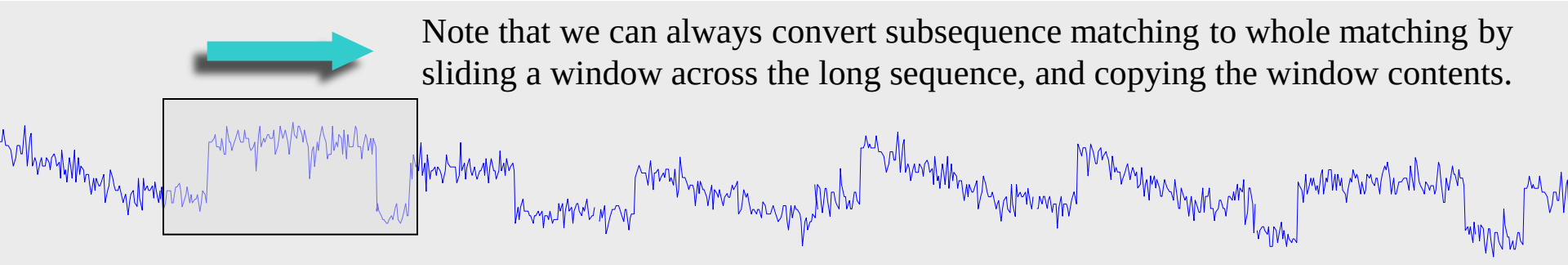
The similarity matching problem can come in two flavors II

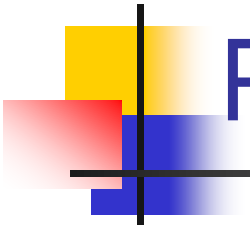


2: Subsequence Matching



Given a Query *Q*, a reference database **C** and a distance measure, find the location that best matches *Q*.





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.
- Time series prediction of stock market indices.

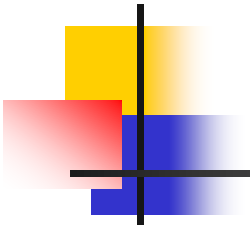
Deviation / Anomaly Detection

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

Typical network traffic at University level may reach over 100 million connections per day

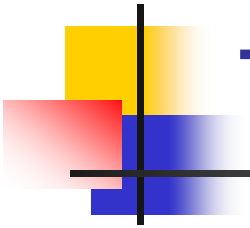
Performance Measurement

- Efficiency
- Effectiveness (interestingness)
  - Objective measures; based on statistics & structures of patterns
  - e.g. support, confidence
  - Subjective: based on user’ s beliefs in data
  - e.g. unexpectedness, novelty



Interestingness

- A pattern is interesting if it is
  - Easily understood by humans
  - Valid on new or test data with some degree of certainty
  - Potentially useful
  - Validates some hypothesis that a user seeks to confirm



Techniques to Be Utilized

- Database-oriented
- Machine learning
- Neural network
- Fuzzy set
- Statistics
- Visualization

Features & Challenges of KDD

- Handling of different types of data
- Efficiency & scalability of data mining algorithm
- Usefulness, certainly & expressiveness of results
- Interactive mining at multiple abstraction levels
- Parallel & distributed data mining
- Protection of privacy & data security