Data Mining
(EECS 6412)

Introduction (Cont’d)

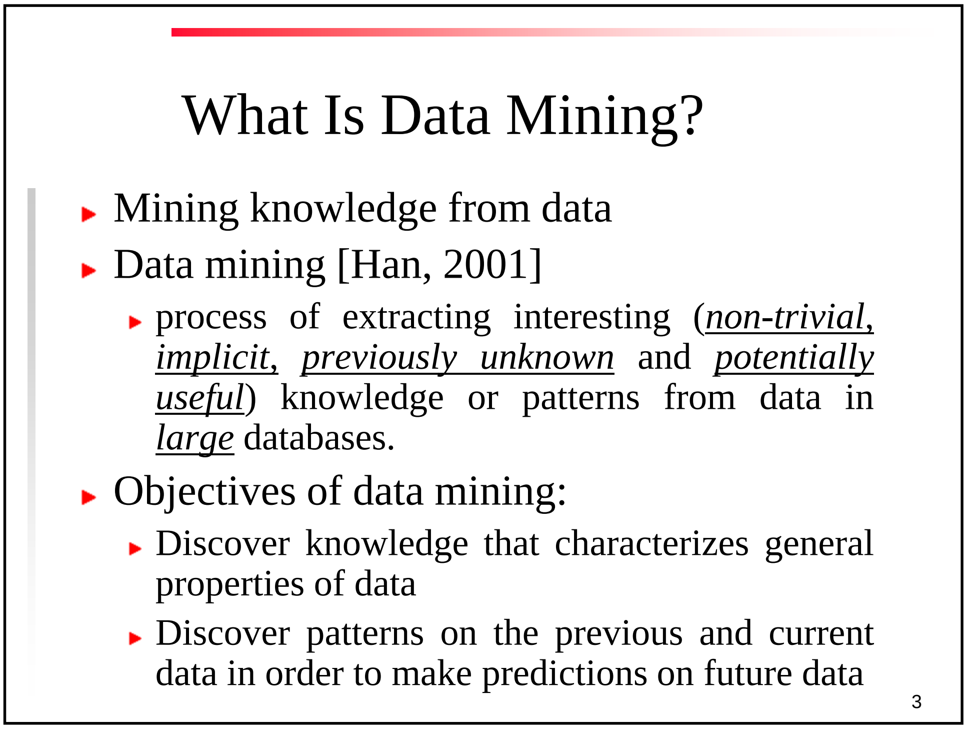
(http://www.eecs.yorku.ca/course/6412/)

Outline of Introduction

- *Why data mining?*
- *What is data mining?*

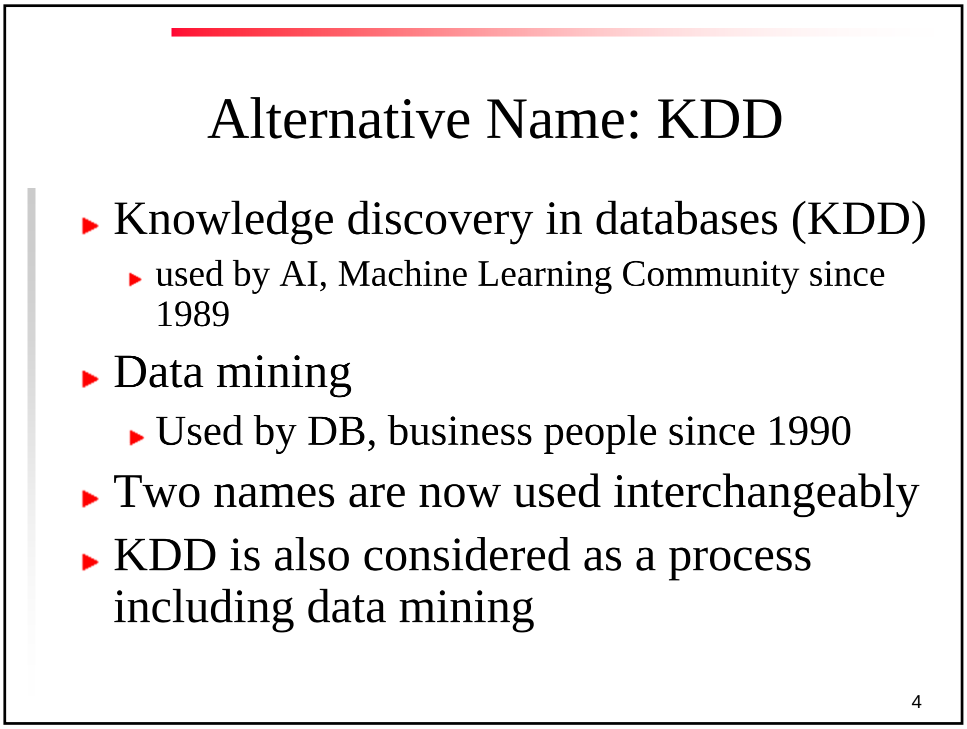
} Covered last time

- Process of KDD or data mining
- What kind of data to mine from?
- What kind of patterns to mine? (Data Mining Tasks)
- Data mining R&D issues
- Data Mining applications



What Is Data Mining?

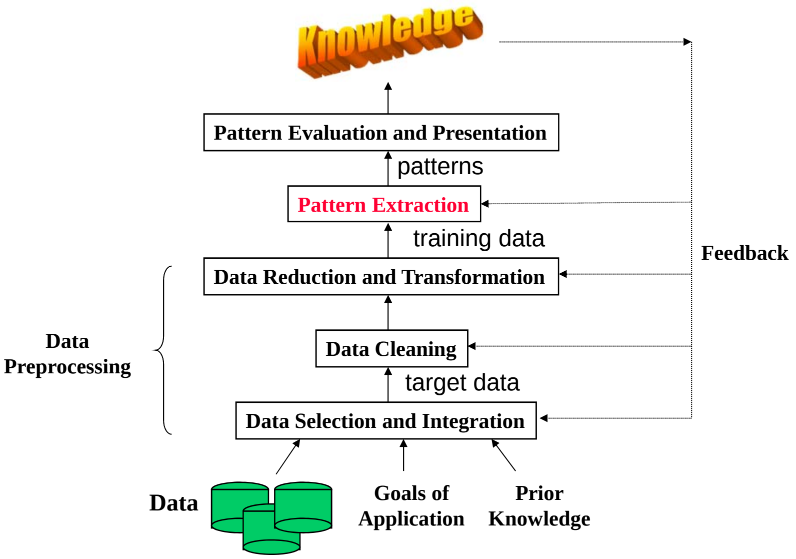
- Mining knowledge from data
- Data mining [Han, 2001]
  - process of extracting interesting (*non-trivial*, *implicit*, *previously unknown* and *potentially useful*) knowledge or patterns from data in *large* databases.
- Objectives of data mining:
  - Discover knowledge that characterizes general properties of data
  - Discover patterns on the previous and current data in order to make predictions on future data



Alternative Name: KDD

- Knowledge discovery in databases (KDD)
  - used by AI, Machine Learning Community since 1989
- Data mining
  - Used by DB, business people since 1990
- Two names are now used interchangeably
- KDD is also considered as a process including data mining

Process of Data Mining and KDD



Knowledge

Pattern Evaluation and Presentation

patterns

Pattern Extraction

training data

Data Reduction and Transformation

Data Cleaning

target data

Data Selection and Integration

Data Preprocessing

Feedback

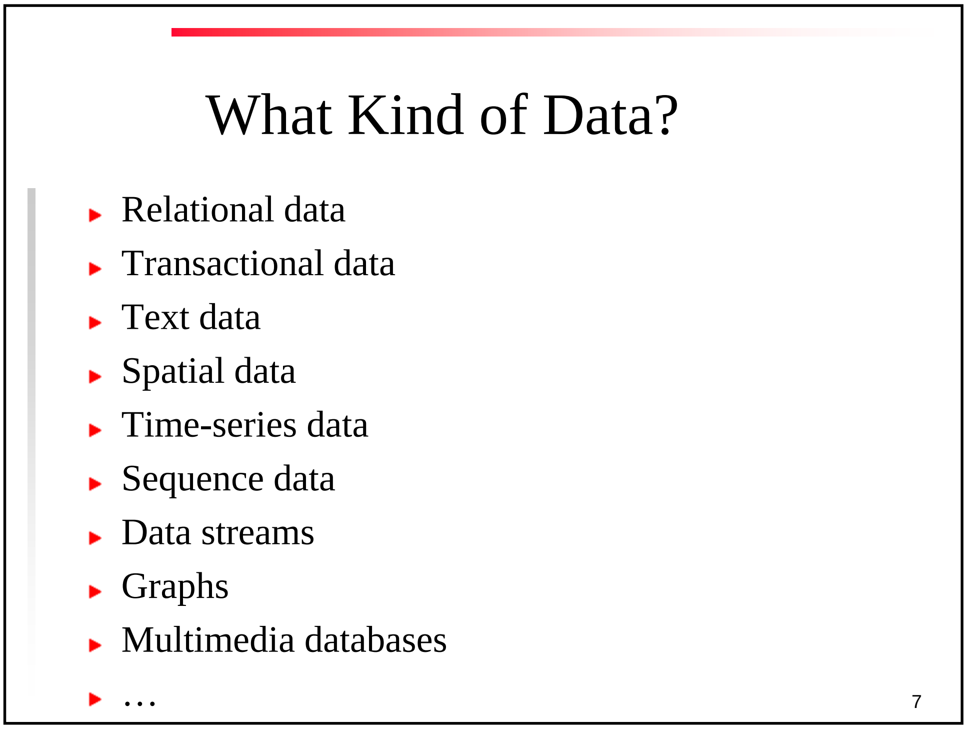
Data

Goals of Application

Prior Knowledge

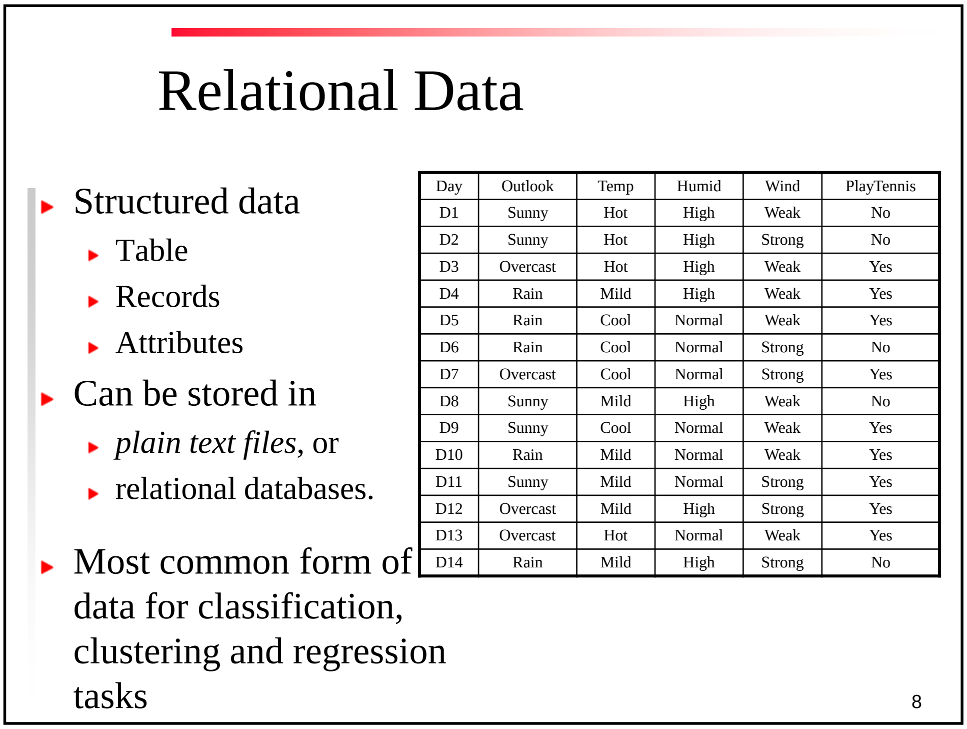
Outline

- *Why data mining? (Done)*
- *What is data mining? (Done)*
- *Process of KDD or data mining (Done)*
- What kind of data to mine from?
- What kind of patterns to mine? (Data Mining Tasks)
- Data mining R&D issues
- Data mining applications



What Kind of Data?

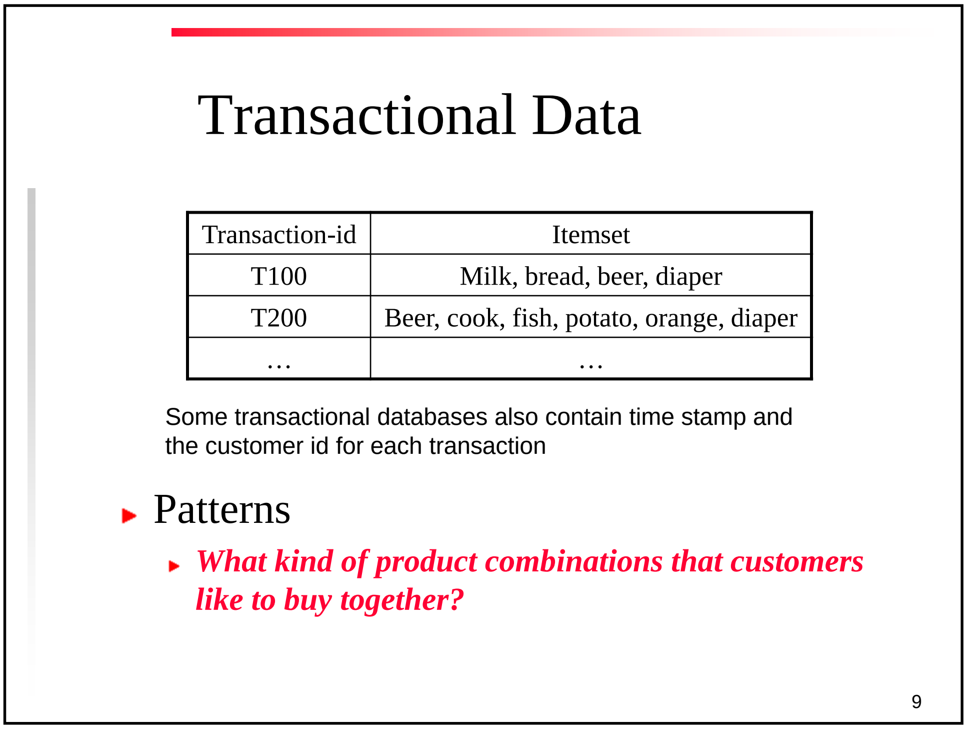
- Relational data
- Transactional data
- Text data
- Spatial data
- Time-series data
- Sequence data
- Data streams
- Graphs
- Multimedia databases
- …



Relational Data

- Structured data
  - Table
  - Records
  - Attributes
- Can be stored in
  - *plain text files*, or
  - relational databases.
- Most common form of data for classification, clustering and regression tasks

Day	Outlook	Temp	Humid	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

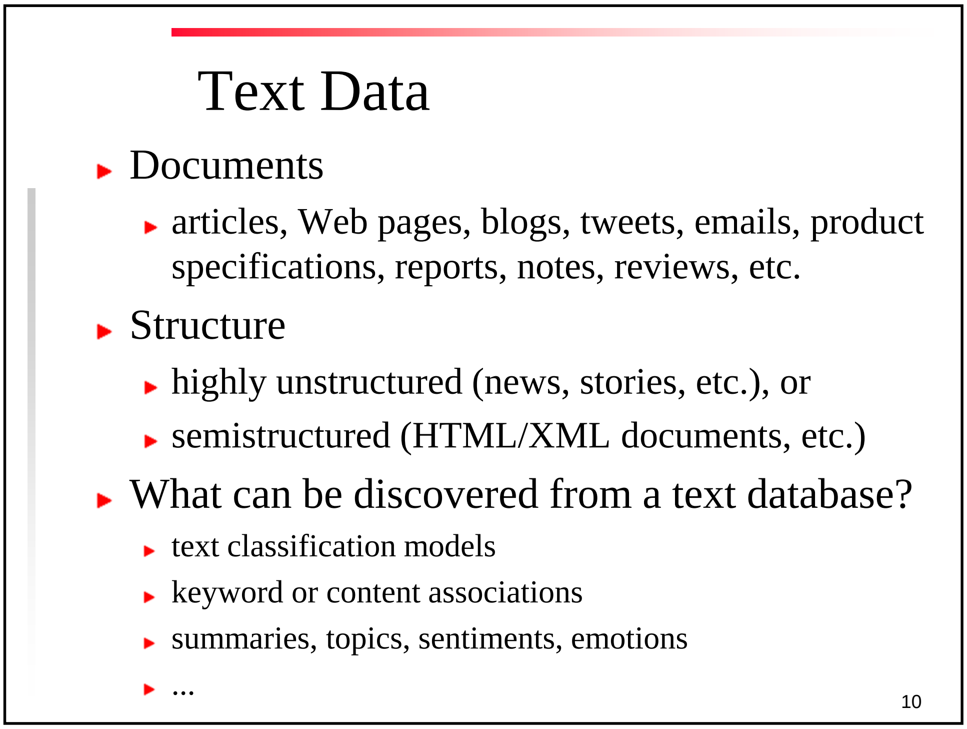


Transactional Data

Transaction-id	Itemset
T100	Milk, bread, beer, diaper
T200	Beer, cook, fish, potato, orange, diaper
…	…

Some transactional databases also contain time stamp and the customer id for each transaction

- Patterns
  - ***What kind of product combinations that customers like to buy together?***



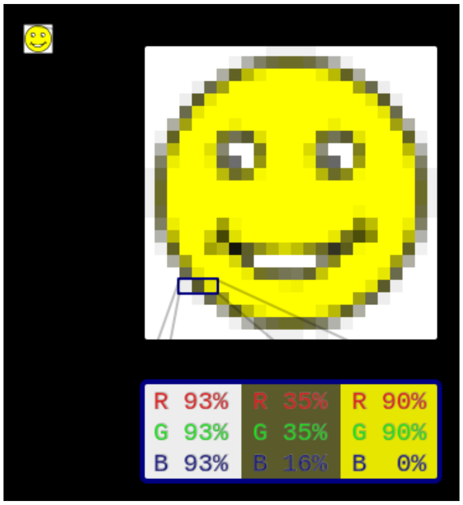
Text Data

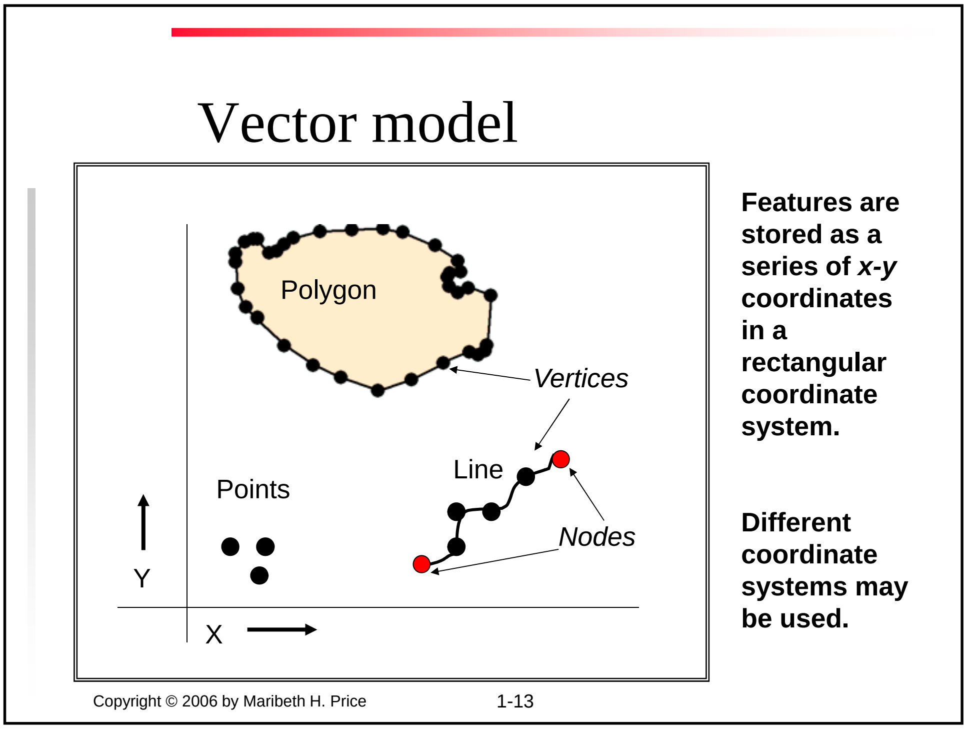
- Documents
  - articles, Web pages, blogs, tweets, emails, product specifications, reports, notes, reviews, etc.
- Structure
  - highly unstructured (news, stories, etc.), or
  - semistructured (HTML/XML documents, etc.)
- What can be discovered from a text database?
  - text classification models
  - keyword or content associations
  - summaries, topics, sentiments, emotions
  - …

Spatial Data

- Spatial related information
  - Information about locations, shapes, characteristics and relationships among geographic objects.
- Examples
  - Maps
  - Geographic databases:
    - Characteristics of attributes/features in a geographic space
    - Linking attributes and locations.
  - Medical or satellite image databases
  - VLSI chip design databases
- Format
  - Raster format (composed of pixels)
  - Vector format (composed of paths)

Raster Graphics



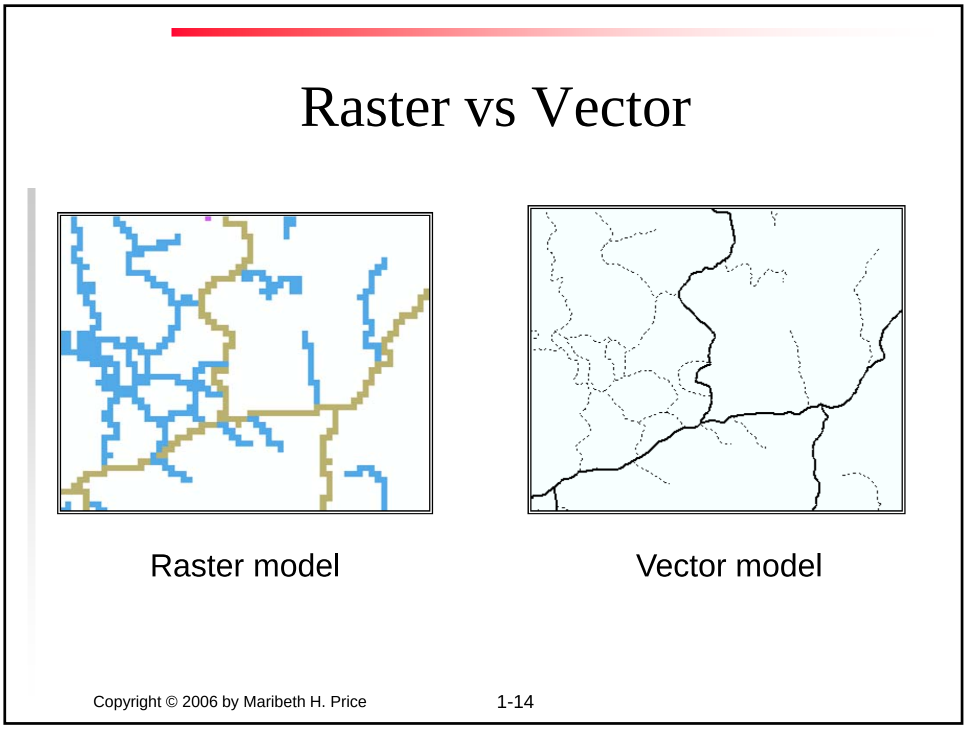


Vector model

Features are stored as a series of *x-y* coordinates in a rectangular coordinate system.

Different coordinate systems may be used.

1-13



Raster vs Vector

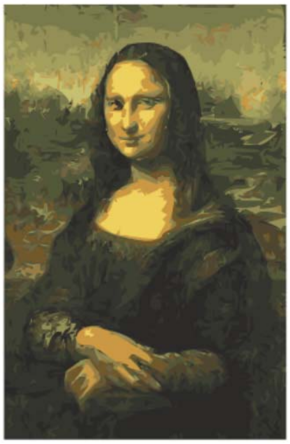
Raster model

Vector model

1-14

Raster vs Vector

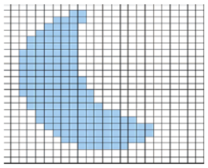
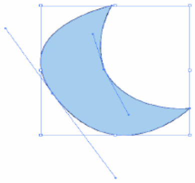
Raster Image

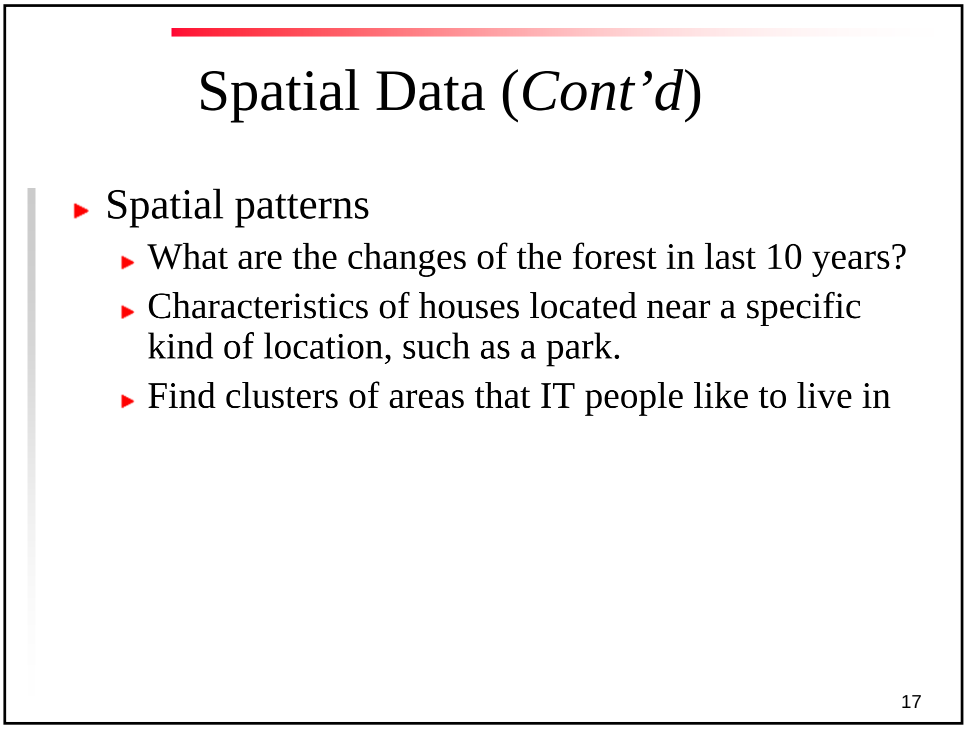


Vector Image

Raster vs Vector

Raster and Vector Graphics

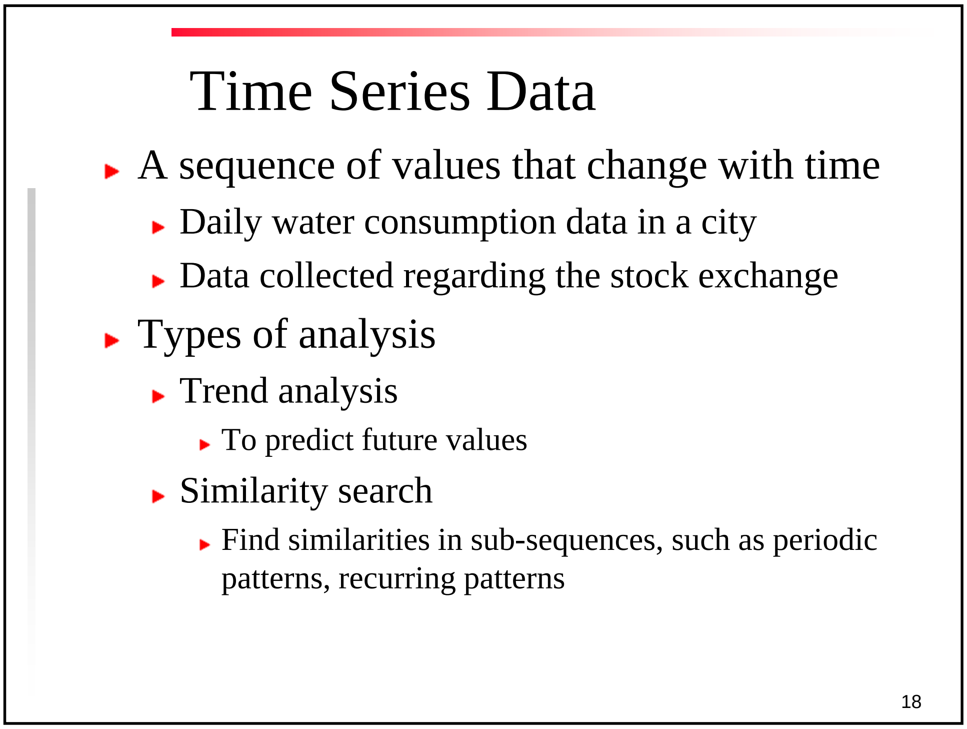
Raster	Vector
	
Made up of a ***grid*** of pixels	Geometric shapes and lines that are defined ***mathematically***
Resolution dependent	Resolution ***independent***
When scaled, visual quality and sharpness is degraded	When scaled, visual quality and sharpness is ***unaffected***
File size is relatively ***large***	File size is relatively ***small***
File Formats: ***GIF, TIF, BMP, PSD***	File Formats: ***EPS, WMF, AI***
Pixel-oriented	***Object*** -oriented



Spatial Data (*Cont’d*)

- Spatial patterns
  - What are the changes of the forest in last 10 years?
  - Characteristics of houses located near a specific kind of location, such as a park.
  - Find clusters of areas that IT people like to live in

17



Time Series Data

- A sequence of values that change with time
  - Daily water consumption data in a city
  - Data collected regarding the stock exchange
- Types of analysis
  - Trend analysis
    - To predict future values
  - Similarity search
    - Find similarities in sub-sequences, such as periodic patterns, recurring patterns

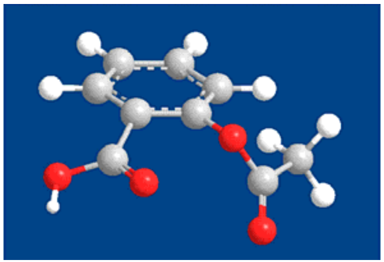
18

Sequence Data

- Sequences of ordered objects or events (with or without concrete notation of time)
  - Bio-sequences
    - DNA, protein
  - Web log data
    - Click stream (web page traversal sequences)
  - Sequences of items bought by a customer
- Patterns
  - Frequent sequences, alignments of two bio-sequences, etc.

Graph Data

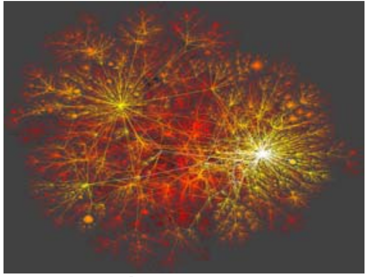
Graphs are everywhere



Aspirin

from H. Jeong et al Nature 411, 41 (2001)

Yeast protein interaction network

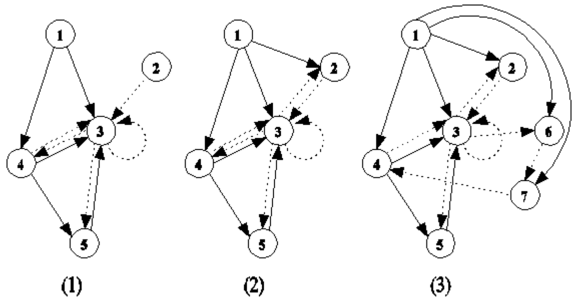


Internet

Co-author network

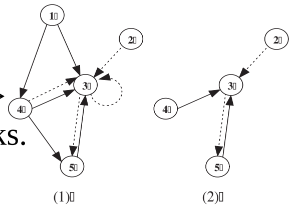
Example of Graph Dataset

Graph data set: containing one or more graphs (3 in this example)



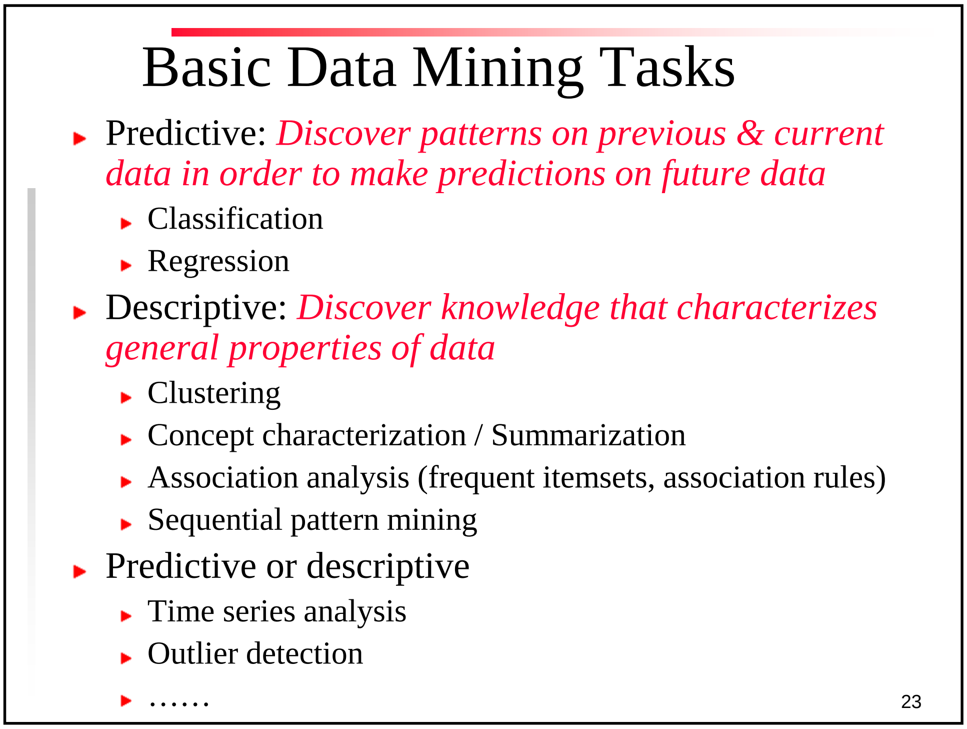
Patterns can be found:

- Frequent subgraphs →
- Find communities in social networks.
- Most influential people in social networks



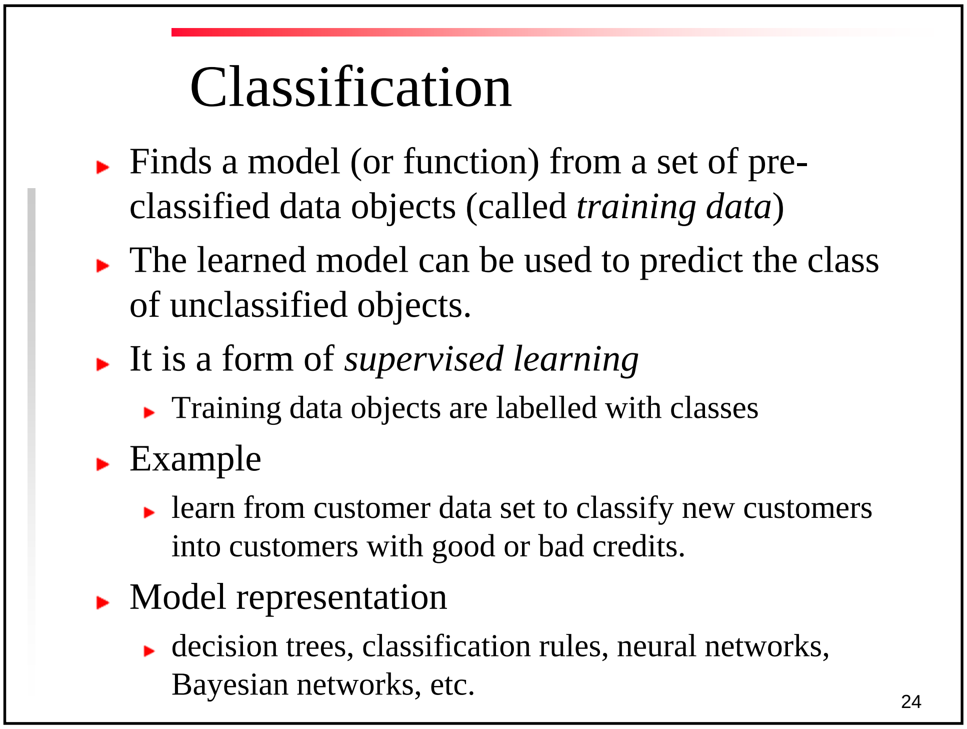
Outline

- *Why data mining? (Done)*
- *What is data mining? (Done)*
- *Process of KDD or data mining (Done)*
- *What kind of data to mine from? (Done)*
- What kind of patterns to mine? (Data Mining Tasks)
- Data mining R&D issues
- Data mining applications



Basic Data Mining Tasks

- Predictive: *Discover patterns on previous & current data in order to make predictions on future data*
  - Classification
  - Regression
- Descriptive: *Discover knowledge that characterizes general properties of data*
  - Clustering
  - Concept characterization / Summarization
  - Association analysis (frequent itemsets, association rules)
  - Sequential pattern mining
- Predictive or descriptive
  - Time series analysis
  - Outlier detection
  - ……



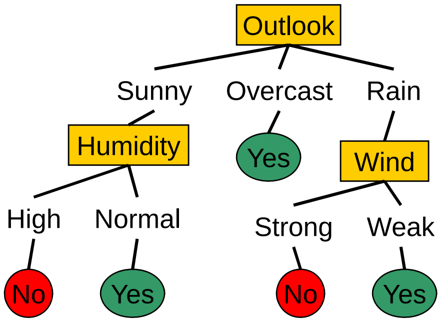
Classification

- Finds a model (or function) from a set of pre-classified data objects (called *training data*)
- The learned model can be used to predict the class of unclassified objects.
- It is a form of *supervised learning*
  - Training data objects are labelled with classes
- Example
  - learn from customer data set to classify new customers into customers with good or bad credits.
- Model representation
  - decision trees, classification rules, neural networks, Bayesian networks, etc.

Classification Models

- A decision tree for *PlayTennis*

Day	Outlook	Temp	Humid	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Regression

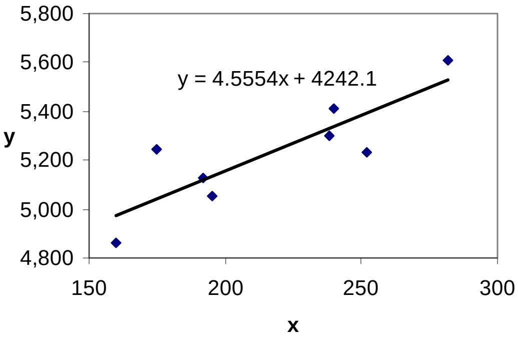
- Finds a function from data which relates a real-valued variable with one or more other variables:

$$y = f(x_1, x_2, \ldots, x_k)$$

- The learned model can be used to predict some unknown or missing numerical values
- It is also supervised learning
- Examples:
  - Predict daily water demand, age of abalone, etc.
- Types of models:
  - Linear regression model
  - Nonlinear regression model

Example of Simple Linear Regression

x	y
160	4,862
175	5244
192	5128
195	5052
238	5298
240	5410
252	5234
282	5608



Example of Multiple Linear Regression

Wire Bond Data

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

Fitted linear regression model:

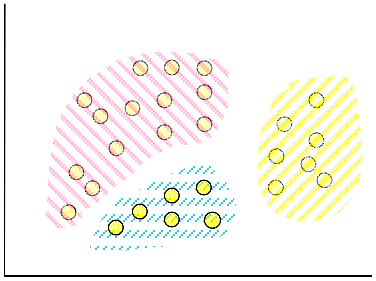
$$y = 2.26379 + 2.74427x_1 + 0.01253x_2$$

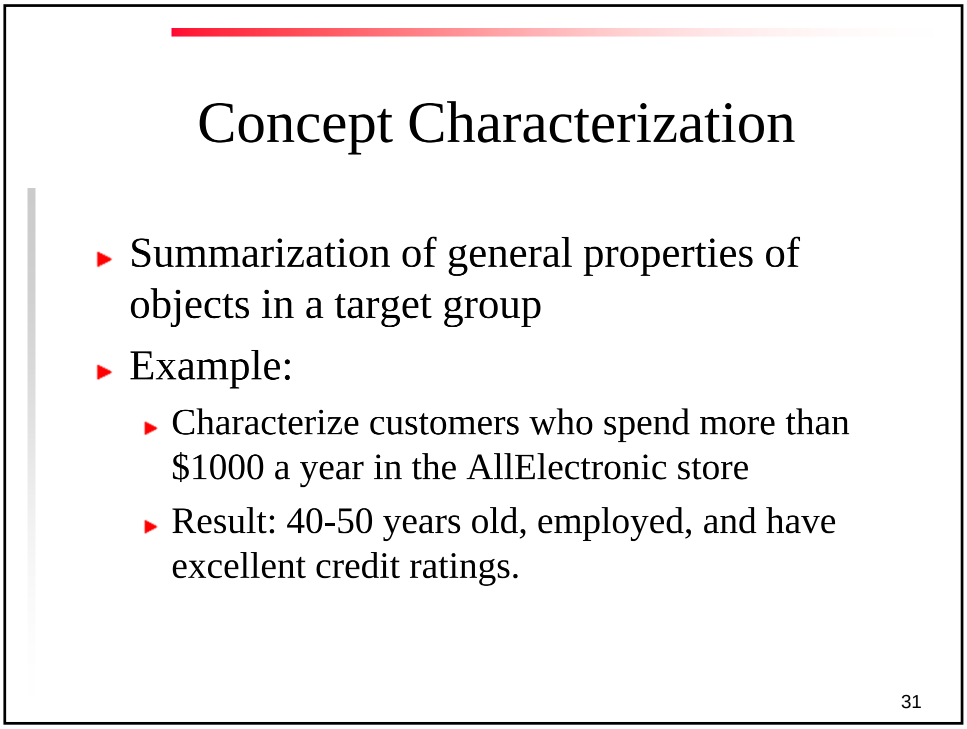
Basic Data Mining Tasks

- Predictive:
  - Classification
  - Regression
- **Descriptive:**
  - Clustering
  - Concept characterization / Summarization
  - Association analysis (frequent itemsets, association rules)
  - Sequential pattern mining
- Predictive or descriptive
  - Time series analysis
  - Outlier detection
  - ……

Clustering

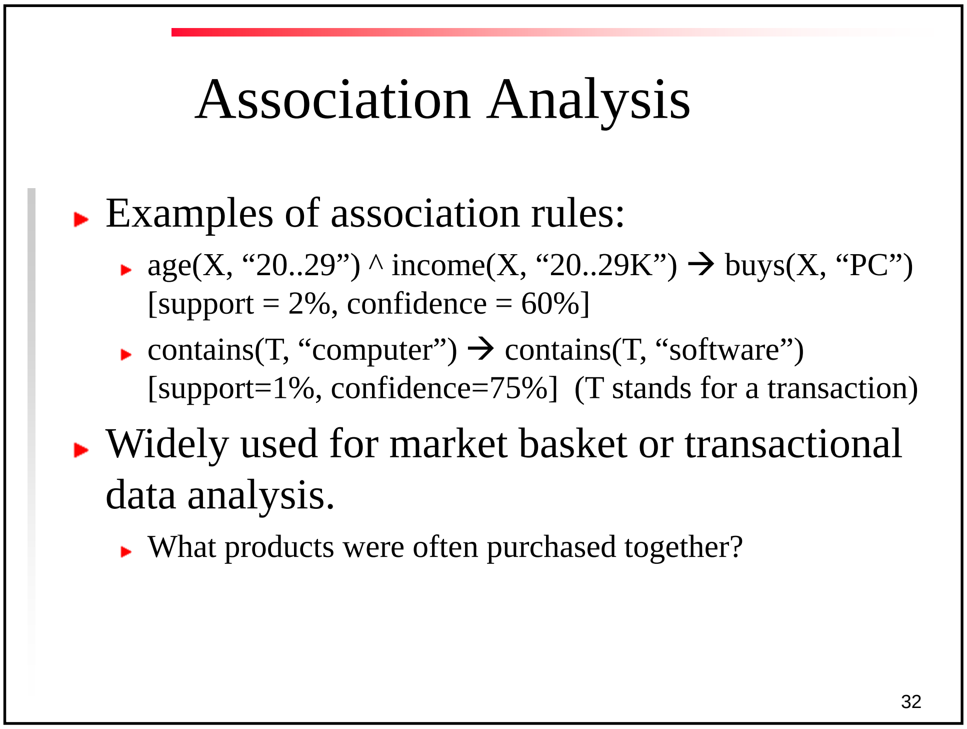
- Class label is unknown in the training data.
- Group data to form classes (clusters)
  - Unsupervised learning.
- Principle: maximizing the intra-class similarity and minimizing the inter-class similarity
- Applications
  - Market/customer segmentation





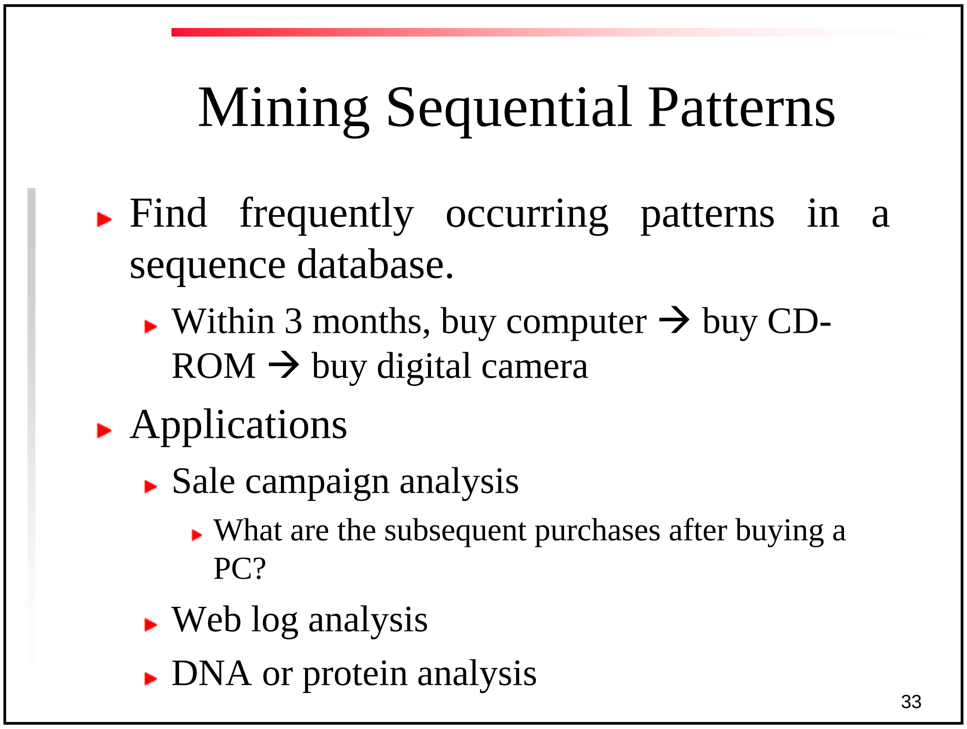
Concept Characterization

- Summarization of general properties of objects in a target group
- Example:
  - Characterize customers who spend more than $1000 a year in the AllElectronic store
  - Result: 40-50 years old, employed, and have excellent credit ratings.



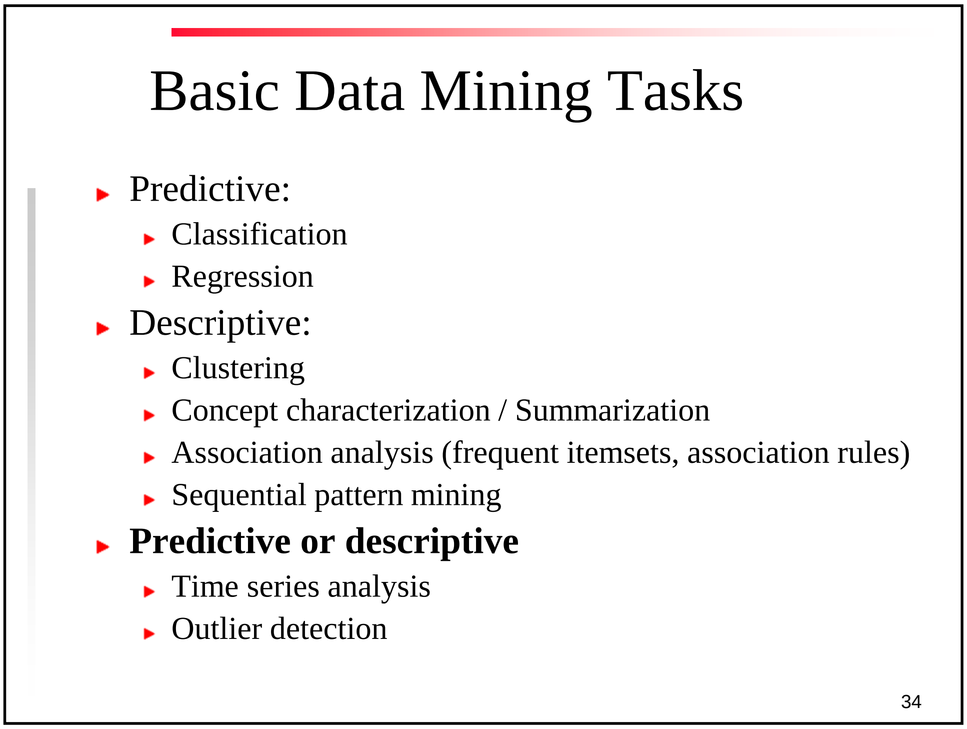
Association Analysis

- Examples of association rules:
  - age(X, “20..29”) ^ income(X, “20..29K”) → buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(T, “computer”) → contains(T, “software”) [support=1%, confidence=75%] (T stands for a transaction)
- Widely used for market basket or transactional data analysis.
  - What products were often purchased together?



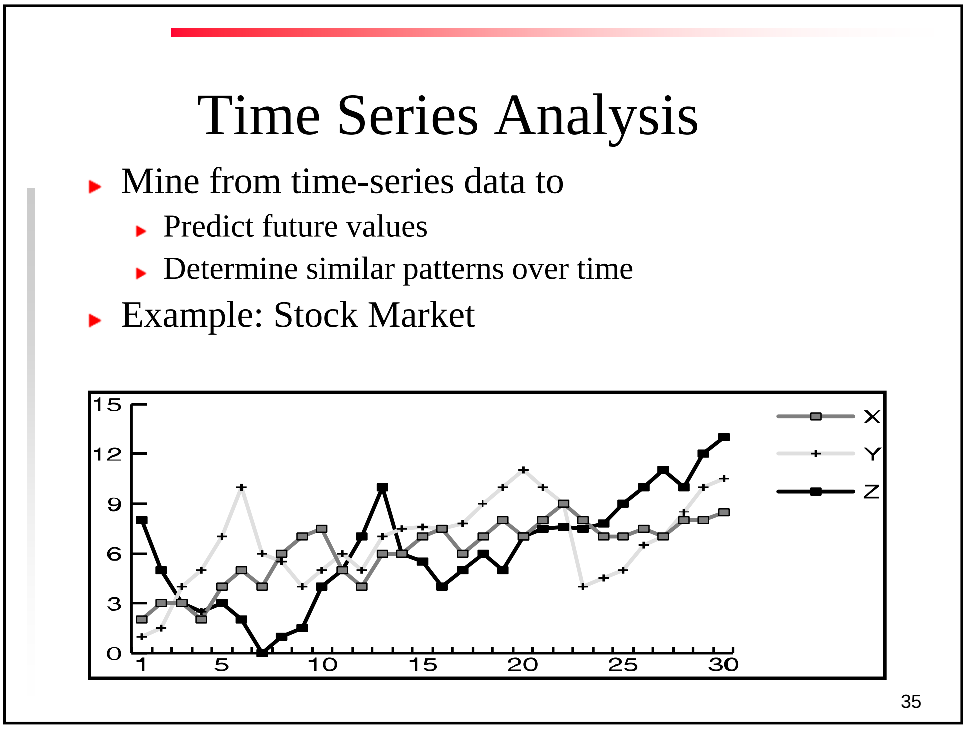
Mining Sequential Patterns

- Find frequently occurring patterns in a sequence database.
  - Within 3 months, buy computer → buy CD-ROM → buy digital camera
- Applications
  - Sale campaign analysis
    - What are the subsequent purchases after buying a PC?
  - Web log analysis
  - DNA or protein analysis



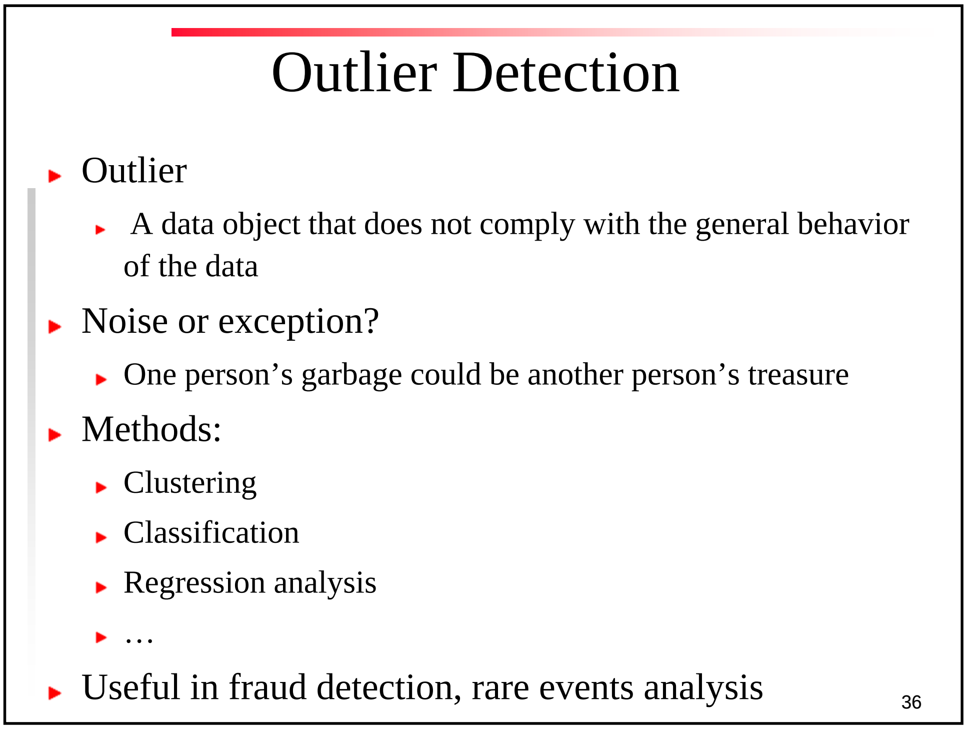
Basic Data Mining Tasks

- Predictive:
  - Classification
  - Regression
- Descriptive:
  - Clustering
  - Concept characterization / Summarization
  - Association analysis (frequent itemsets, association rules)
  - Sequential pattern mining
- **Predictive or descriptive**
  - Time series analysis
  - Outlier detection



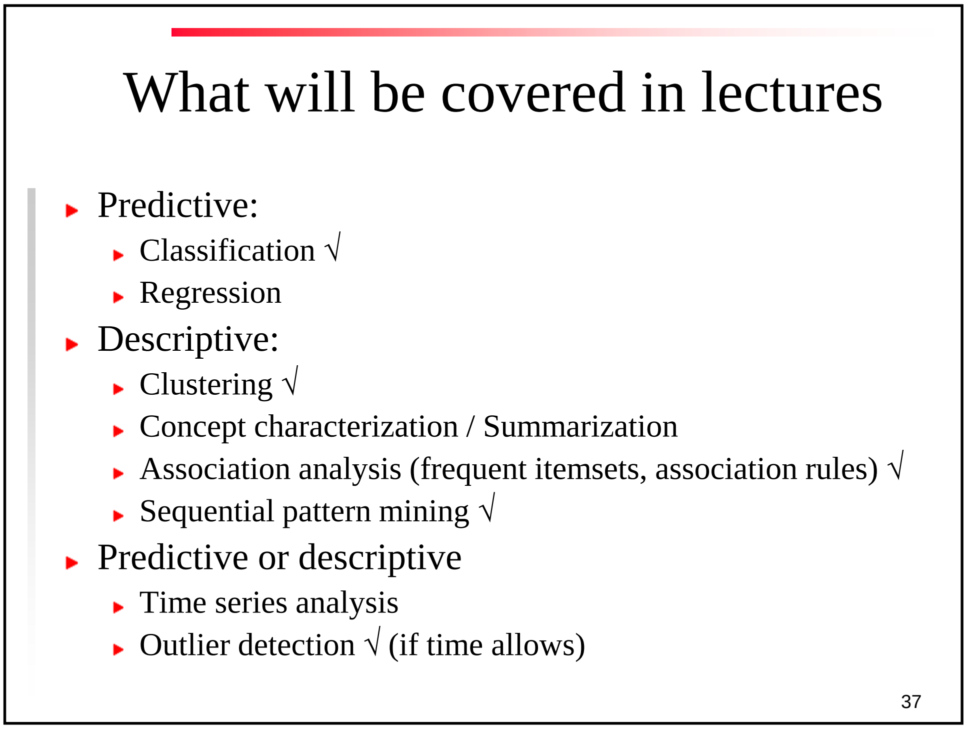
Time Series Analysis

- Mine from time-series data to
  - Predict future values
  - Determine similar patterns over time
- Example: Stock Market



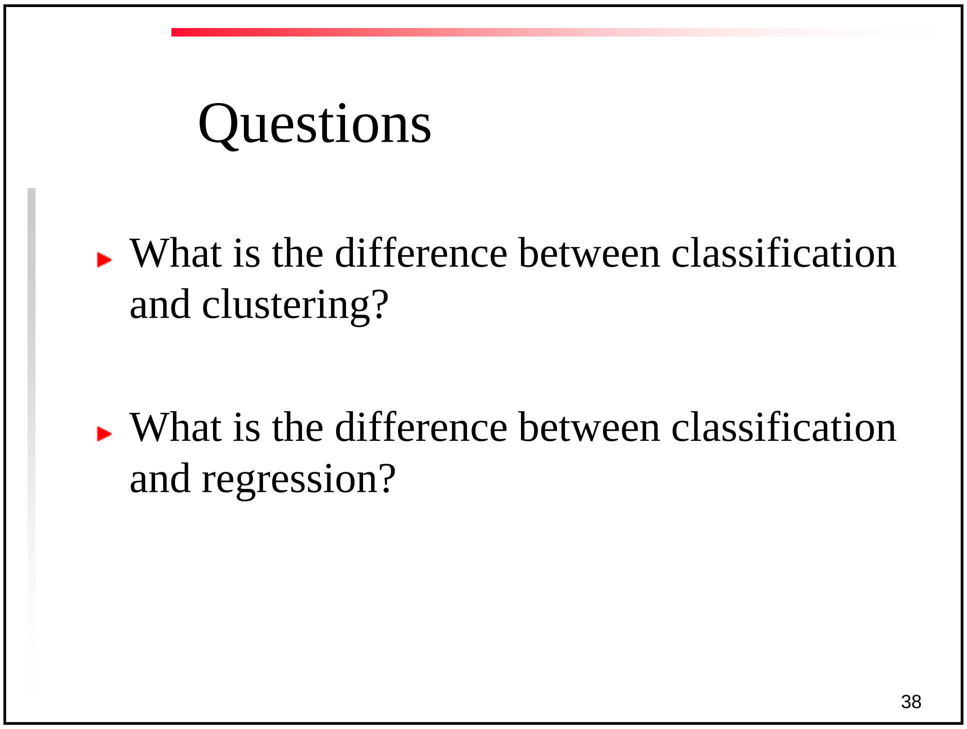
Outlier Detection

- Outlier
  - A data object that does not comply with the general behavior of the data
- Noise or exception?
  - One person’s garbage could be another person’s treasure
- Methods:
  - Clustering
  - Classification
  - Regression analysis
  - …
- Useful in fraud detection, rare events analysis



What will be covered in lectures

- Predictive:
  - Classification √
  - Regression
- Descriptive:
  - Clustering √
  - Concept characterization / Summarization
  - Association analysis (frequent itemsets, association rules) √
  - Sequential pattern mining √
- Predictive or descriptive
  - Time series analysis
  - Outlier detection √ (if time allows)



Questions

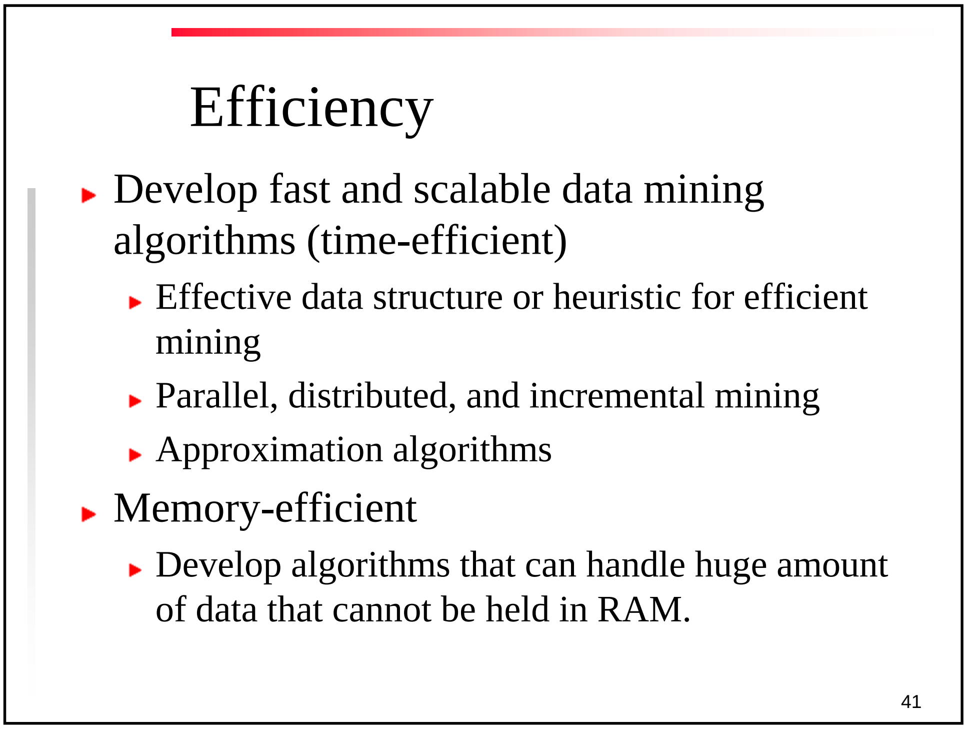
- What is the difference between classification and clustering?
- What is the difference between classification and regression?

Outline

- *Why data mining? (Done)*
- *What is data mining? (Done)*
- *Course information (Done)*
- *Process of KDD or data mining (Done)*
- *What kind of data to mine from? (Done)*
- *What kind of patterns to mine? (Data Mining Tasks) (Done)*
- Data mining R&D issues
- Data mining applications

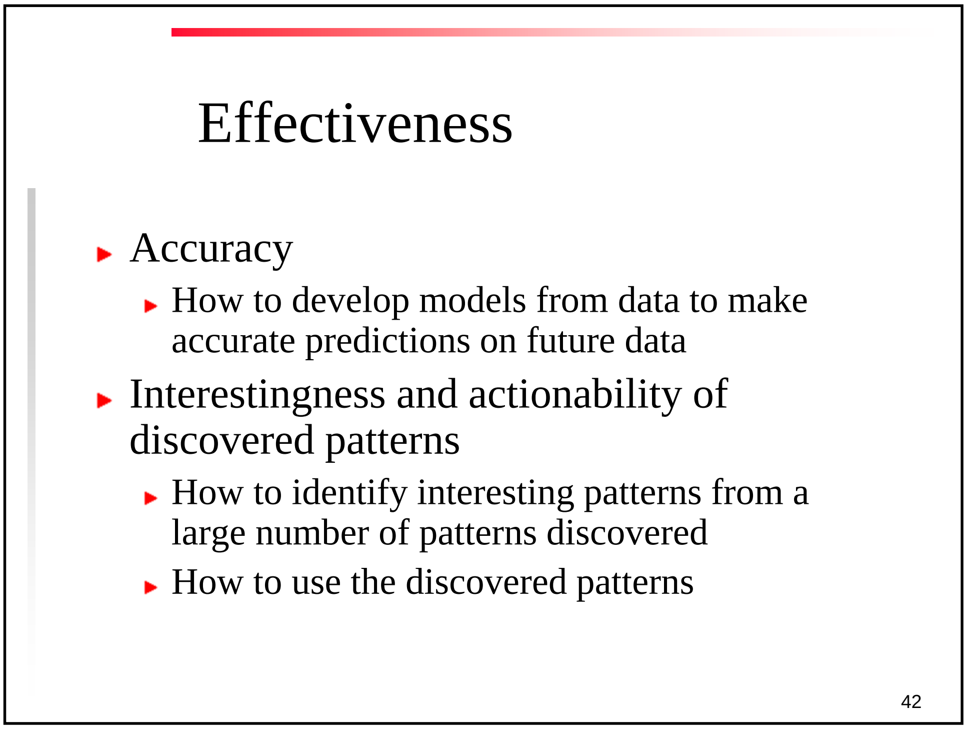
Major Issues in Data Mining

- Efficiency
- Effectiveness
- User interaction
- Privacy preserving



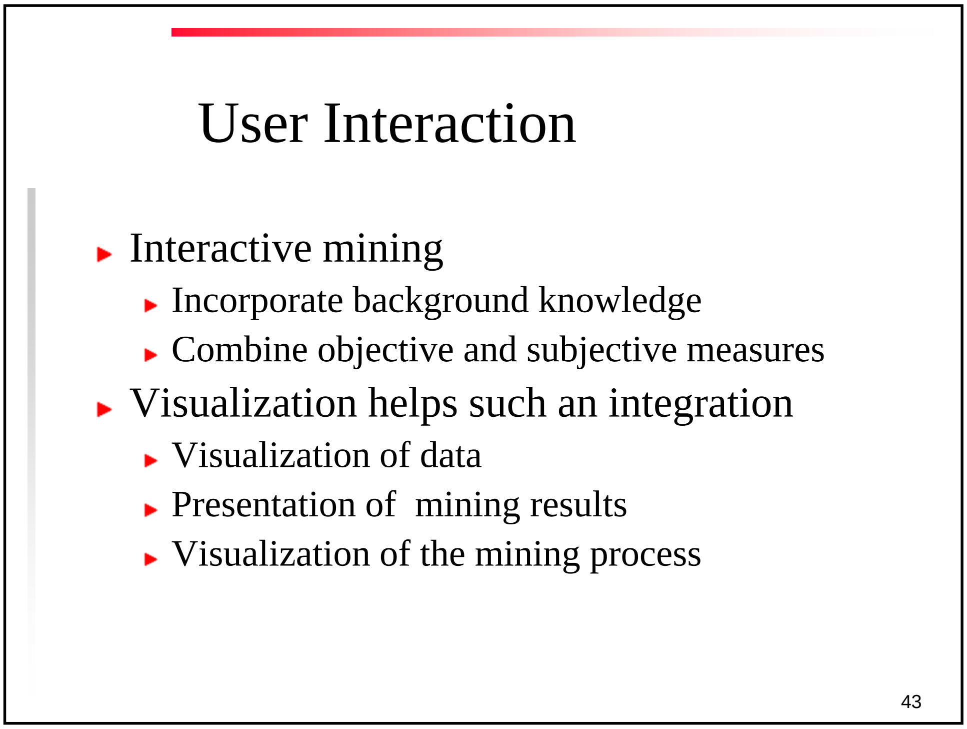
Efficiency

- Develop fast and scalable data mining algorithms (time-efficient)
  - Effective data structure or heuristic for efficient mining
  - Parallel, distributed, and incremental mining
  - Approximation algorithms
- Memory-efficient
  - Develop algorithms that can handle huge amount of data that cannot be held in RAM.



Effectiveness

- Accuracy
  - How to develop models from data to make accurate predictions on future data
- Interestingness and actionability of discovered patterns
  - How to identify interesting patterns from a large number of patterns discovered
  - How to use the discovered patterns

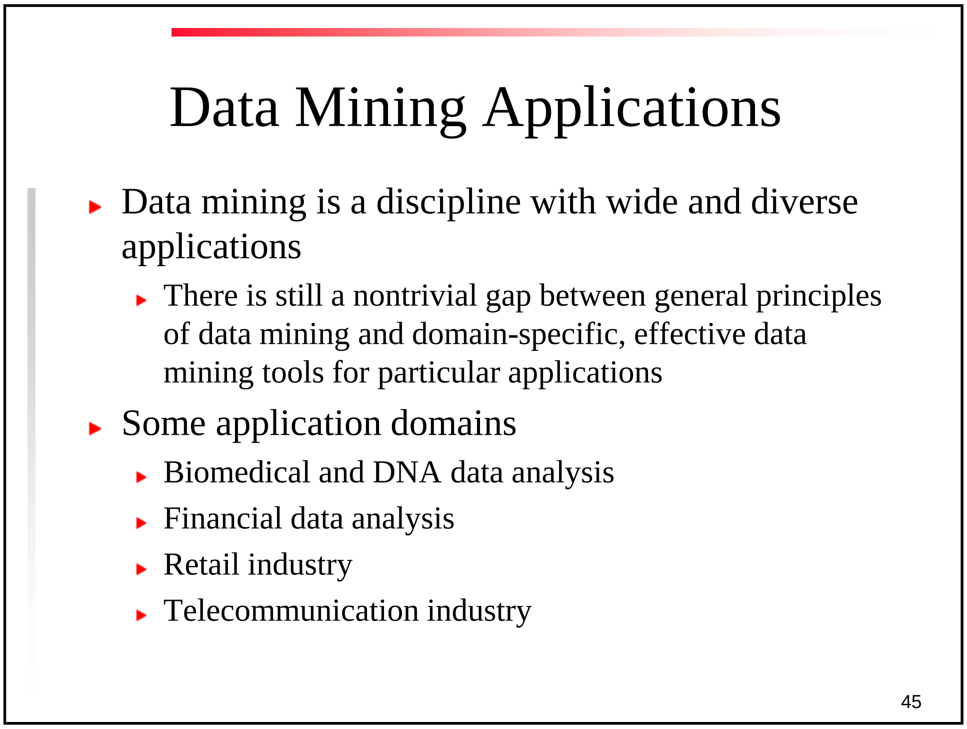


User Interaction

- Interactive mining
  - Incorporate background knowledge
  - Combine objective and subjective measures
- Visualization helps such an integration
  - Visualization of data
  - Presentation of mining results
  - Visualization of the mining process

Data Mining with Privacy

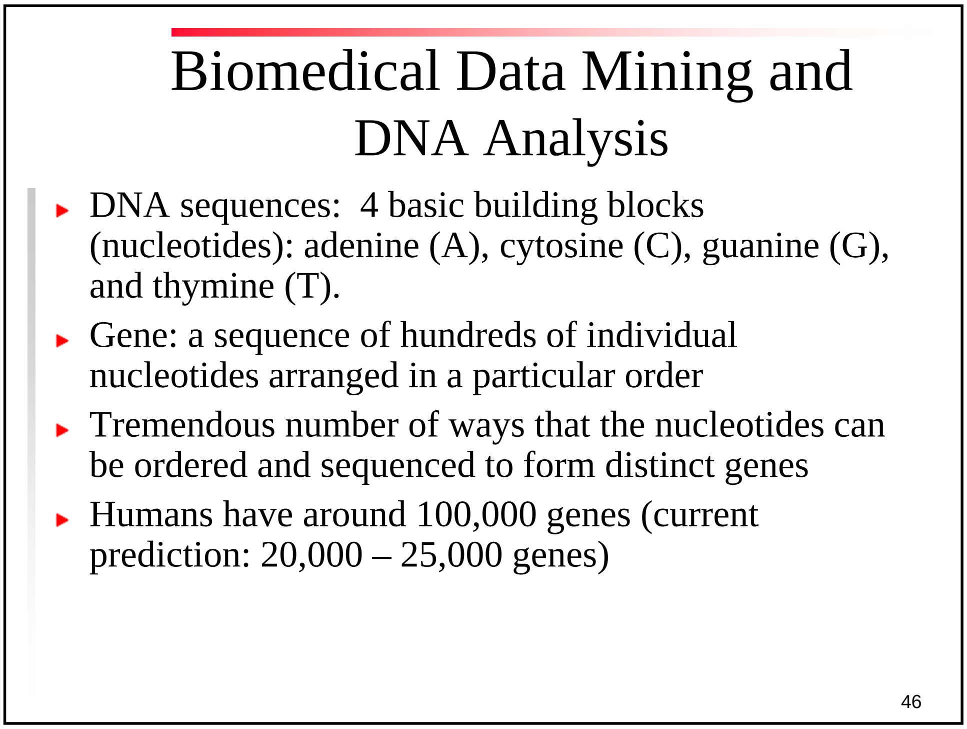
- Data may contain private information
  - Data mining can invade privacy
- Technical solutions can limit privacy invasion
  - Replacing sensitive personal data with anon. ID
  - Alter the data so that real values are obscured
  - Multi-party computation – distributed data
  - …
- Bayardo & Srikant, Technological Solutions for Protecting Privacy, IEEE Computer, Sep 2003



Data Mining Applications

- Data mining is a discipline with wide and diverse applications
  - There is still a nontrivial gap between general principles of data mining and domain-specific, effective data mining tools for particular applications
- Some application domains
  - Biomedical and DNA data analysis
  - Financial data analysis
  - Retail industry
  - Telecommunication industry

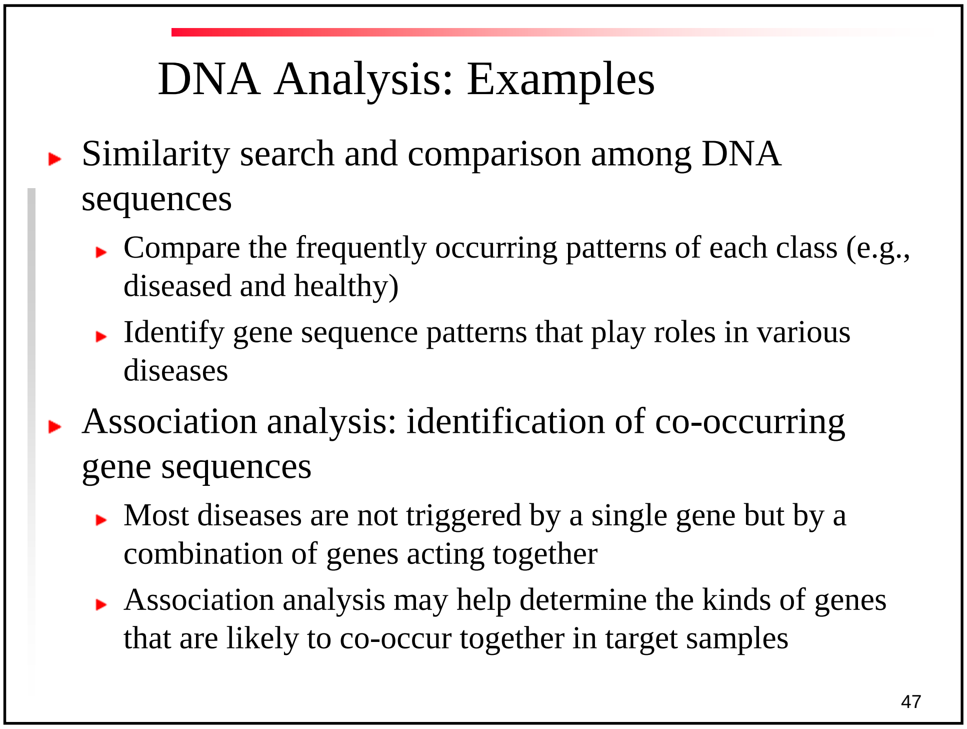
45



Biomedical Data Mining and DNA Analysis

- DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T).
- Gene: a sequence of hundreds of individual nucleotides arranged in a particular order
- Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
- Humans have around 100,000 genes (current prediction: 20,000 – 25,000 genes)

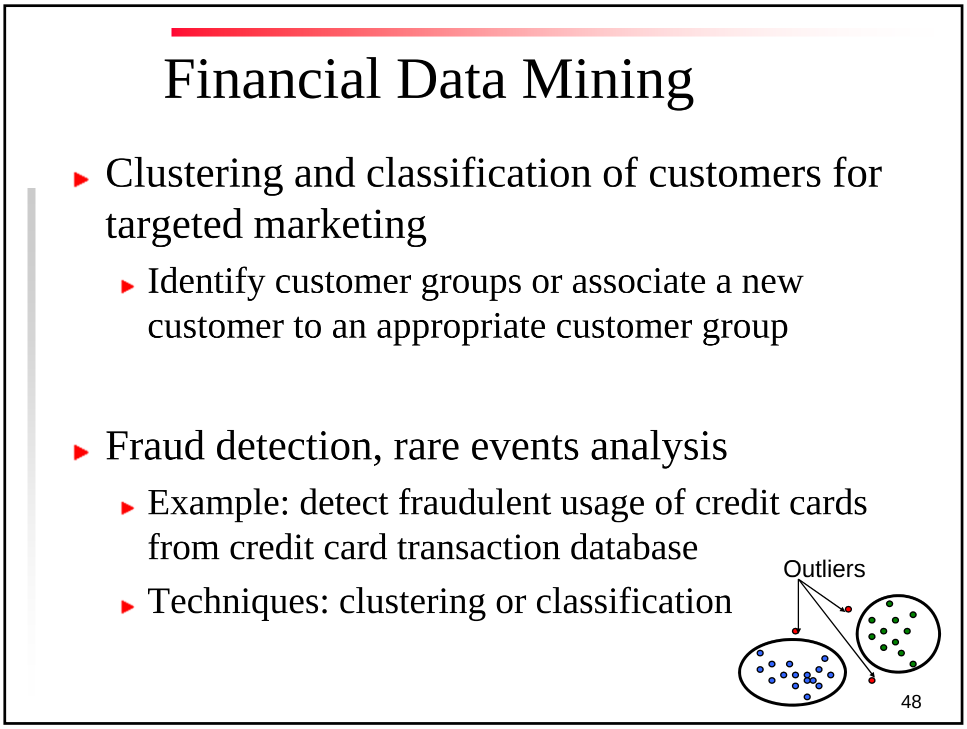
46



DNA Analysis: Examples

- Similarity search and comparison among DNA sequences
  - Compare the frequently occurring patterns of each class (e.g., diseased and healthy)
  - Identify gene sequence patterns that play roles in various diseases
- Association analysis: identification of co-occurring gene sequences
  - Most diseases are not triggered by a single gene but by a combination of genes acting together
  - Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples

47



Financial Data Mining

- Clustering and classification of customers for targeted marketing
  - Identify customer groups or associate a new customer to an appropriate customer group
- Fraud detection, rare events analysis
  - Example: detect fraudulent usage of credit cards from credit card transaction database
  - Techniques: clustering or classification

48

Data Mining in Retail Industry: Examples

- Discover customer shopping patterns and trends
  - Re-arrange store layout
  - Purchase recommendation and cross-reference of items
- Customer retention: Analysis of customer loyalty
  - Use customer loyalty card information to register sequences of purchases of particular customers
  - Use sequential pattern mining to investigate changes in customer consumption or loyalty
  - Suggest adjustments on the pricing and variety of goods

Data Mining for Telecomm. Industry

- Fraudulent pattern analysis and the identification of unusual patterns
  - Identify potentially fraudulent users and their atypical usage patterns
  - Detect attempts to gain fraudulent entry to customer accounts
  - Discover unusual patterns which may need special attention
- Multidimensional association and sequential pattern analysis
  - Find usage patterns for a set of communication services by customer group, by month, etc.
  - Promote the sales of specific services
  - Improve the availability of particular services in a region

Advanced Topics in Data Mining

- Text Mining
  - Text classification and clustering
  - Sentiment analysis
  - Emotion detection
  - Topic detection
- Web Mining
  - Web usage mining
  - Web content mining
- Graph mining
  - Information network analysis
  - Social network analysis
- Spatial data mining

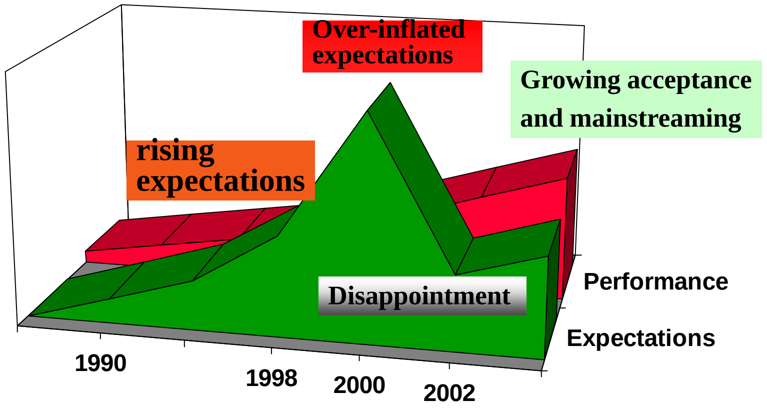
Advanced Topics in Data Mining (Cont'd)

- Data stream mining
- Data mining with high performance computing
  - Parallel, distributed, and cloud-based high performance data mining
  - Data mining with GPU
- Visual data mining
- ……

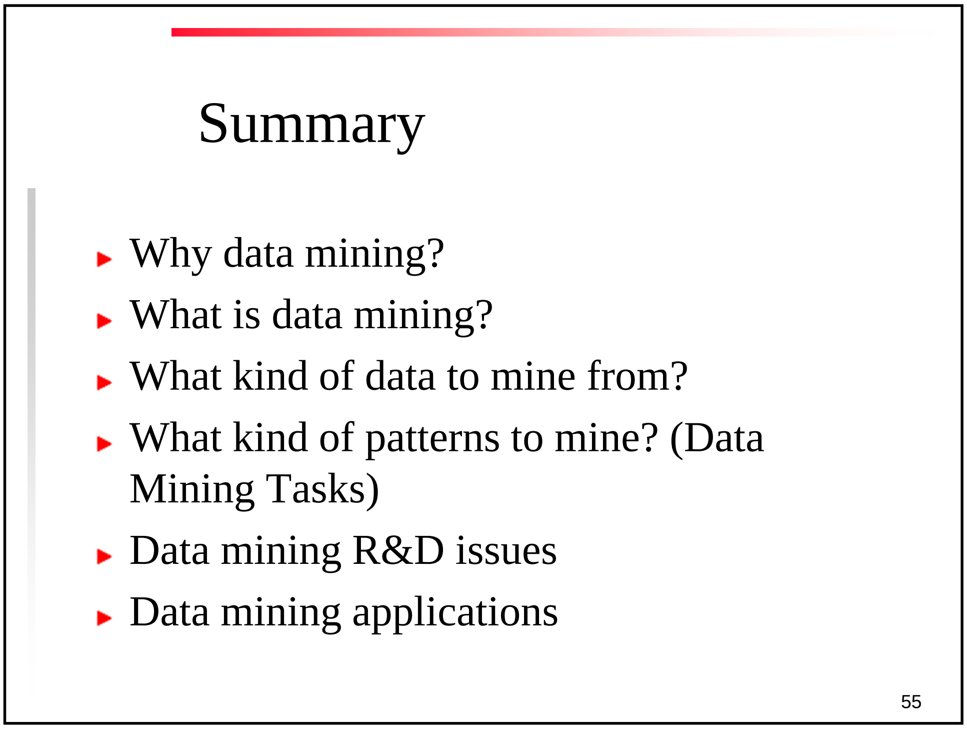
Unsuccessful e-commerce case study (KDD-Cup 2000)

- Data: clickstream and purchase data from Gazelle.com, legwear and legcare e-tailer
- Task: Characterize visitors who spend more than $12 on an average order at the site
- Dataset of 3,465 purchases, 1,831 customers
- Very interesting analysis by Cup participants
  - thousands of hours - $X,000,000 (Millions) of consulting
- Total sales -- $Y,000
- Obituary: Gazelle.com out of business, Aug 2000

The Hype Curve for Data Mining

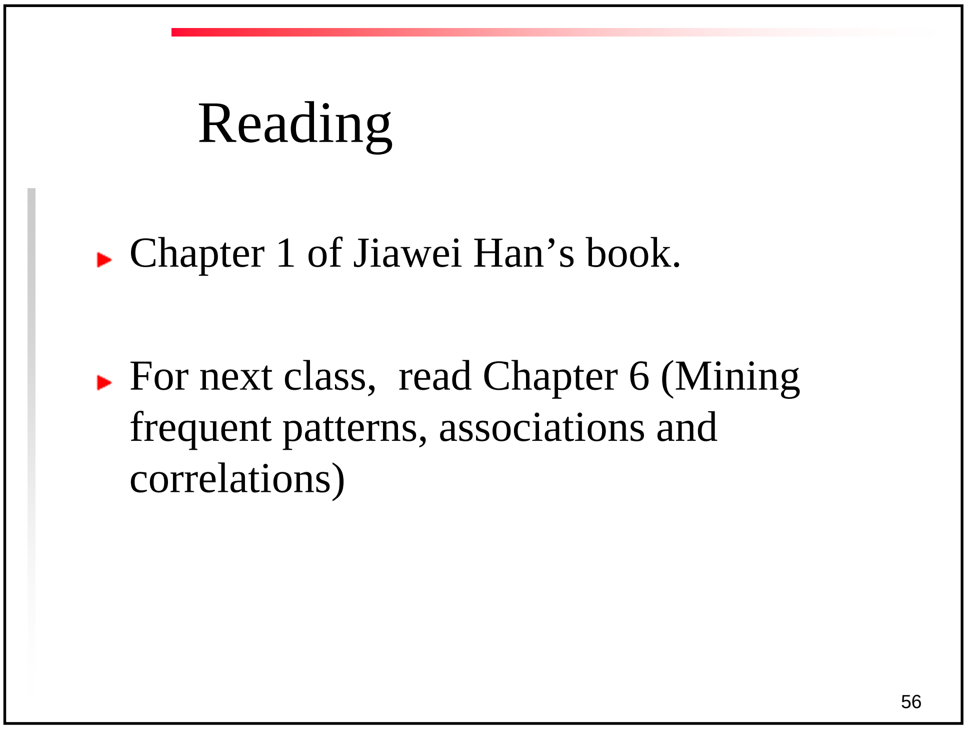


(By Gregory Piatetsky-Shapiro)



Summary

- Why data mining?
- What is data mining?
- What kind of data to mine from?
- What kind of patterns to mine? (Data Mining Tasks)
- Data mining R&D issues
- Data mining applications



Reading

- Chapter 1 of Jiawei Han’s book.
- For next class, read Chapter 6 (Mining frequent patterns, associations and correlations)