Introduction to Data Mining

Instructor: Dr.Eng. Rahmat Widyanto

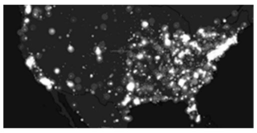
Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

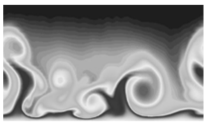
Cyber Security

E-Commerce

Traffic Patterns


Social Networking: Twitter

Sensor Networks

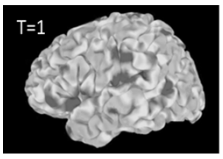

Computational Simulations

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - ◆Yahoo has Peta Bytes of web data
    - ◆Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
    - ◆ Amazon handles millions of visits/day
  - 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)

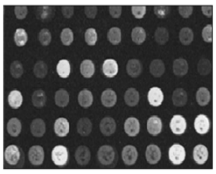
Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - ◆ NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - ◆ Sky survey data
  - High-throughput biological data
  - scientific simulations
    - ◆ terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

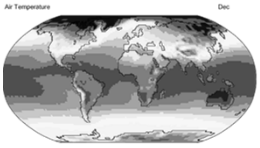


fMRI Data from Brain

Sky Survey Data

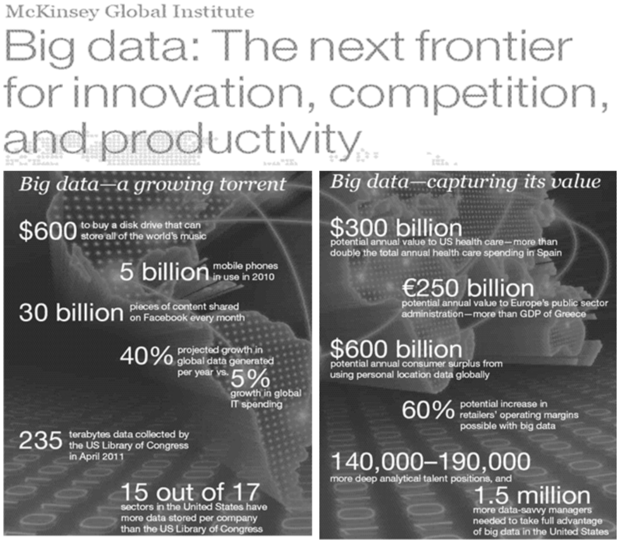


Gene Expression Data



Surface Temperature of Earth

Great opportunities to improve productivity in all walks of life



McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

Big data—a growing torrent

$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

235 terabytes data collected by the US Library of Congress in April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Big data—capturing its value

$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion potential annual value to Europe's public sector administration—more than GDP of Greece

$600 billion potential annual consumer surplus from using personal location data globally

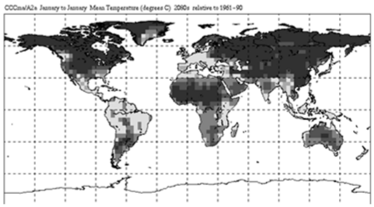
60% potential increase in retailers' operating margins possible with big data

140,000–190,000 more deep analytical talent positions, and

1.5 million more data-savvy managers needed to take full advantage of big data in the United States

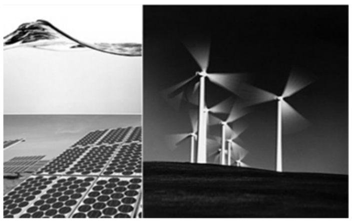
Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



CCCma/A2a January to January Mean Temperature (degrees C) 2080s relative to 1961-90

Predicting the impact of climate change

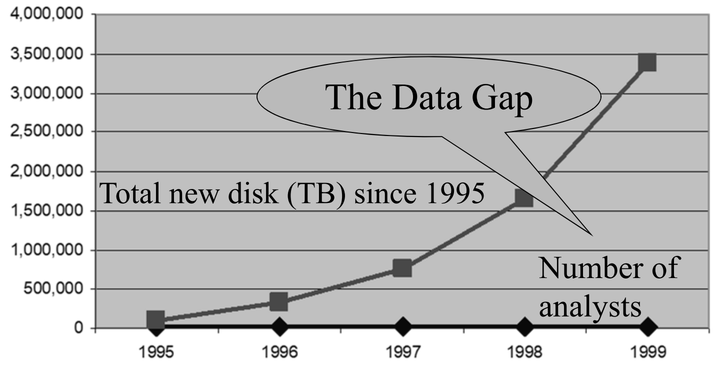


Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

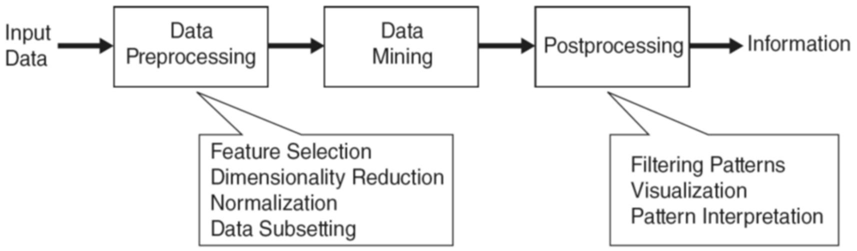
Mining Large Data Sets - Motivation

- **There is often information "hidden" in the data that is not readily evident**
- **Human analysts may take weeks to discover useful information**
- **Much of the data is never analyzed at all**



What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns


Input Data → Data Preprocessing → Data Mining → Postprocessing → Information

Feature Selection
Dimensionality Reduction
Normalization
Data Subsetting

Filtering Patterns
Visualization
Pattern Interpretation

What is (not) Data Mining?

- **What is not Data Mining?**
  - – Look up phone number in phone directory
  - – Query a Web search engine for information about “Amazon”

- **What is Data Mining?**
  - – Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly… in Boston area)
  - – Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)

Some (not so useful) patterns...

- “rules” for American presidents (before 2004 elections)

Some (not so useful) patterns...

- “rules” for American presidents (before 2004 elections)
  - if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected

Some (not so useful) patterns...

- “rules” for American presidents (before 2004 elections)
  - if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
  - no Republican has ever won a presidential election without carrying Ohio

Some (not so useful) patterns...

- “rules” for American presidents (before 2004 elections)
  - if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
  - no Republican has ever won a presidential election without carrying Ohio
  - no incumbent with a four-letter last name has ever been re-elected (Polk, Taft, Ford, Bush Sr.)

Some (not so useful) patterns...

- “rules” for American presidents (before 2004 elections)
  - if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
  - no Republican has ever won a presidential election without carrying Ohio
  - no incumbent with a four-letter last name has ever been re-elected (Polk, Taft, Ford, Bush Sr.)
  - Americans won’t unseat a wartime President

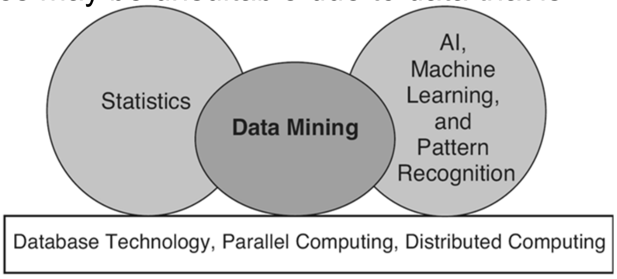
Some (not so useful) patterns...

- “rules” for American presidents (before 2004 elections)
  - if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected (Redskins vs. Panthers: 13-21)
  - no Republican has ever won a presidential election without carrying Ohio
  - no incumbent with a four-letter last name has ever been re-elected (Polk, Taft, Ford, Bush Sr.) (GWB)
  - Americans won’t unseat a wartime President

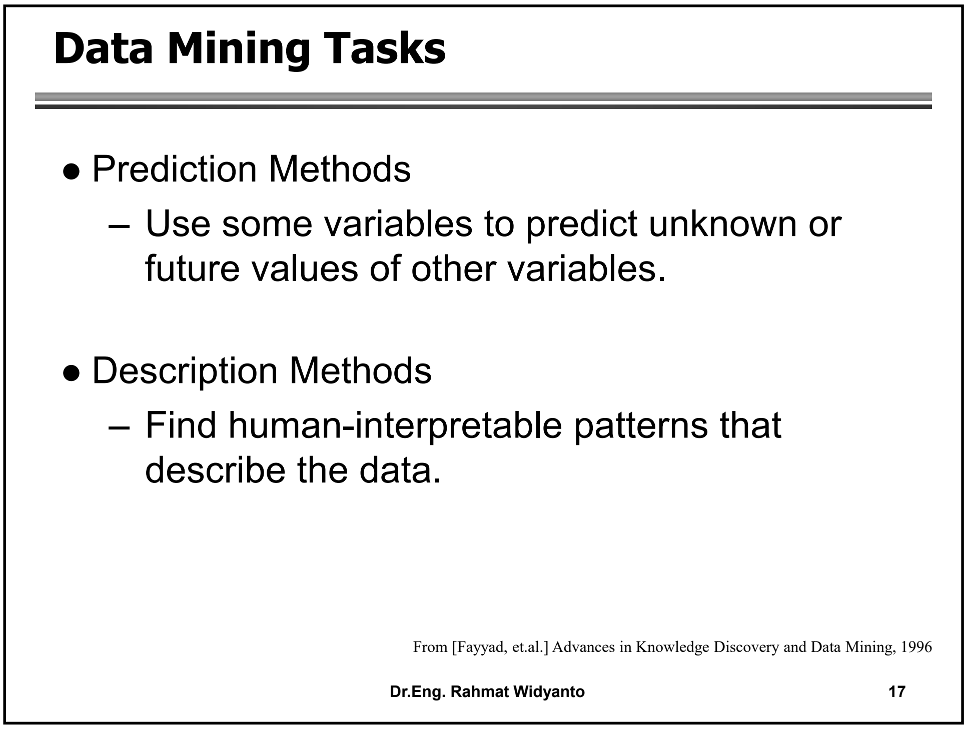
Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed


Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

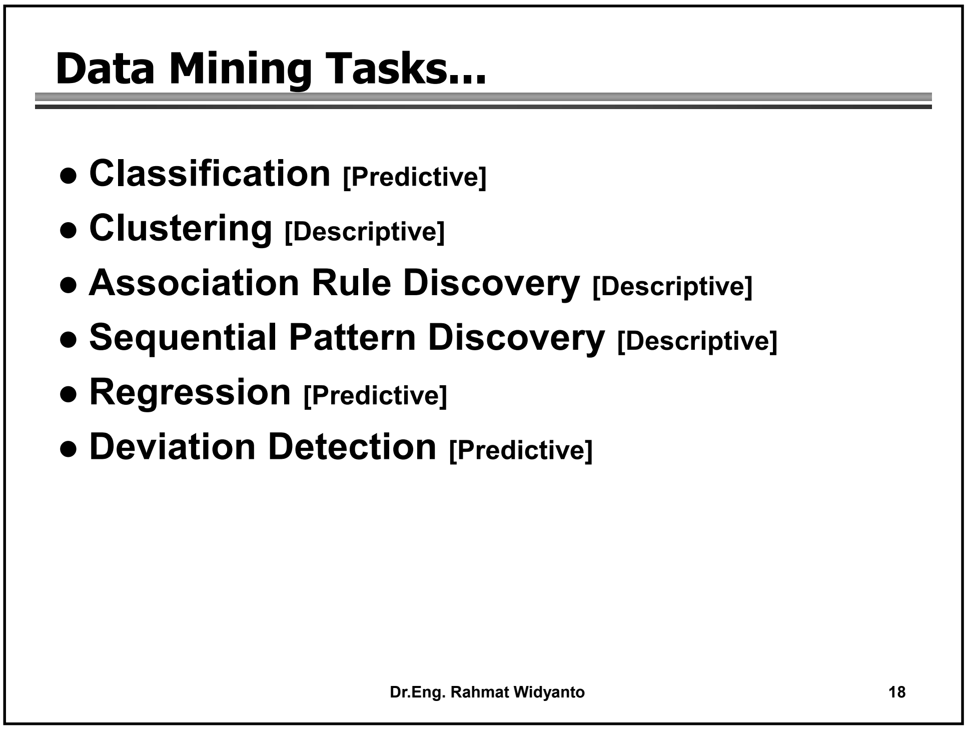
- A key component of the emerging field of data science and data-driven discovery



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.

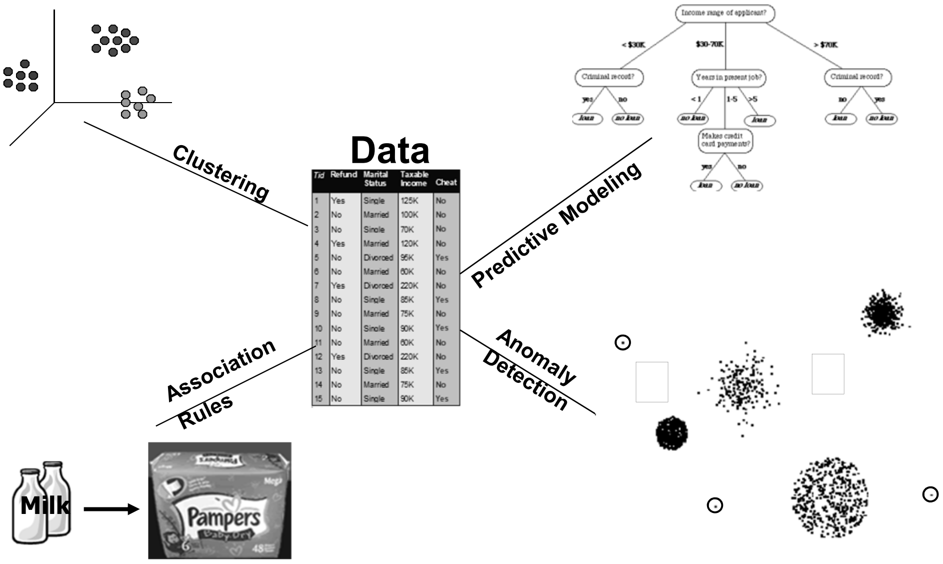
From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996



Data Mining Tasks...

- **Classification** **[Predictive]**
- **Clustering** **[Descriptive]**
- **Association Rule Discovery** **[Descriptive]**
- **Sequential Pattern Discovery** **[Descriptive]**
- **Regression** **[Predictive]**
- **Deviation Detection** **[Predictive]**

Data Mining Tasks ...



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

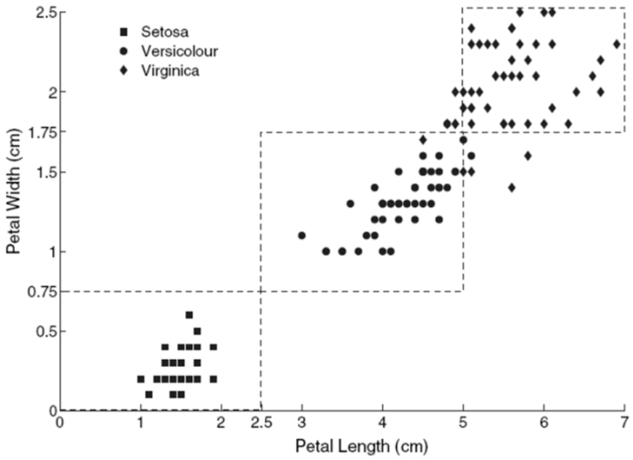


Figure 1.4. Petal width versus petal length for 150 Iris flowers.

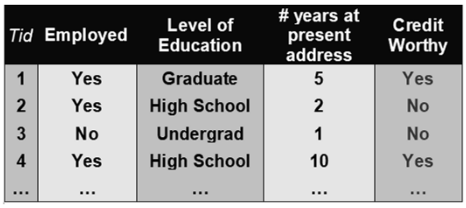
Classification: Definition

- 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.

Predictive Modeling: Classification

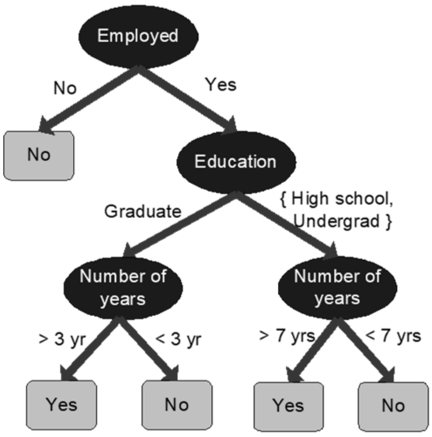
- Find a model for class attribute as a function of the values of other attributes

Class



Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness

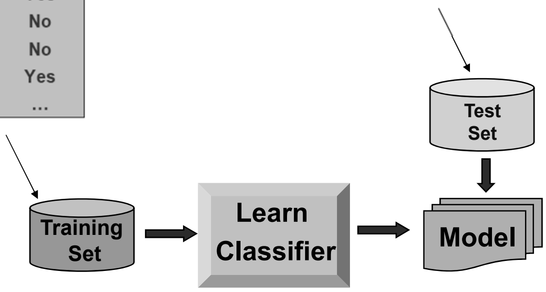


Classification Example

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

categorical categorical quantitative class

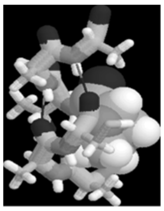
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…



Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



Classification: Application 0

- 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, don't 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: Application 1

- 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.

Classification: Application 2

- Churn prediction for telephone customers
  - **Goal:** To predict whether a customer is likely to be lost to a competitor.
  - **Approach:**
    - ◆ Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - ◆ Label the customers as loyal or disloyal.
    - ◆ Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 3

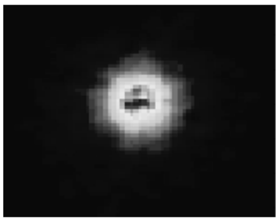
- Sky Survey Cataloging
  - **Goal:** To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Approach:**
    - ◆ Segment the image.
    - ◆ Measure image attributes (features) - 40 of them per object.
    - ◆ Model the class based on these features.
    - ◆ Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

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

Classifying Galaxies

Courtesy: http://aps.umn.edu

Early

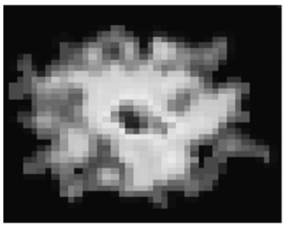


Class:
- **Stages of Formation**

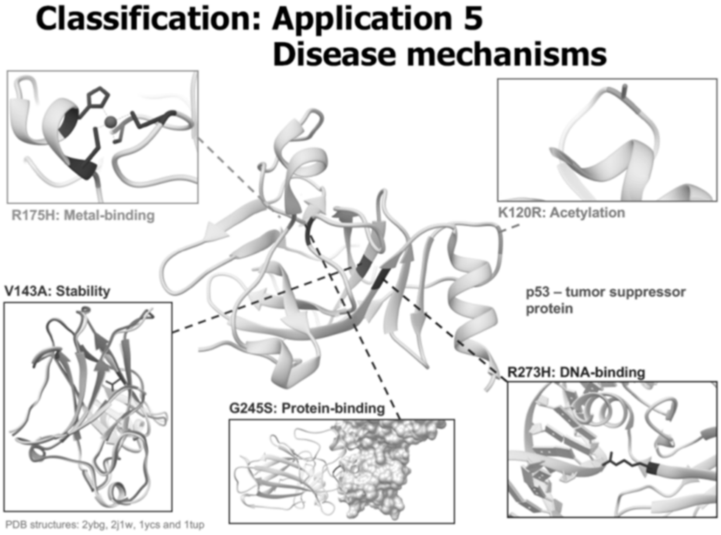
Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate

Late



Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

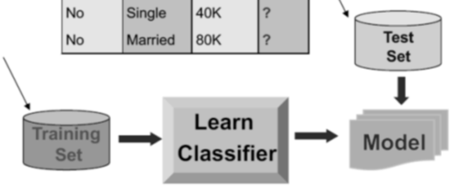


Classification: Application 5
Disease mechanisms

Let's look at classification again!

	categorical	categorical	continuous	class
Tid	**Home Owner**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Home Owner	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



A small digression: why MATLAB?

Home Owner	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

→

No:	**→ 0**
Yes:	**→ 1**
Single:	**→ 0**
Married:	**→ 1**
Divorced:	**→ 2**

↙

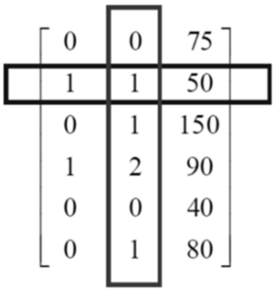
$$\begin{array}{cc} & \text{3 columns} \\ \text{6 rows} & \begin{bmatrix} 0 & 0 & 75 \\ 1 & 1 & 50 \\ 0 & 1 & 150 \\ 1 & 2 & 90 \\ 0 & 0 & 40 \\ 0 & 1 & 80 \end{bmatrix} \end{array}$$

← this is a 6-by-3 MATRIX!!!

6 rows

3 columns

A small digression: why MATLAB?



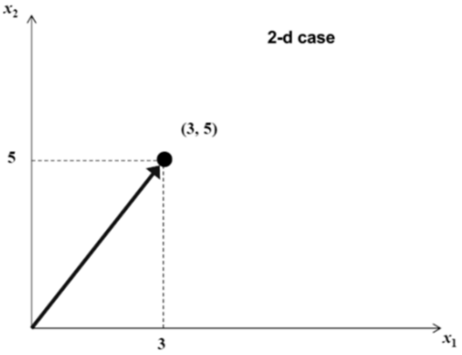
row vector (3-dimensional vector; 1-by-3)

column vector

(6-dimensional vector)

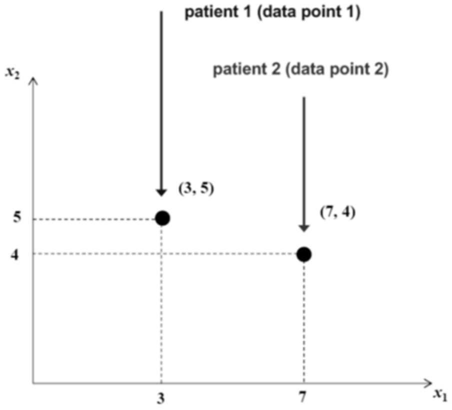
(6-by-1)

2-d case



Let's introduce some notation…

2-dimensional case



Assume medical situation:

x_1 is blood pressure

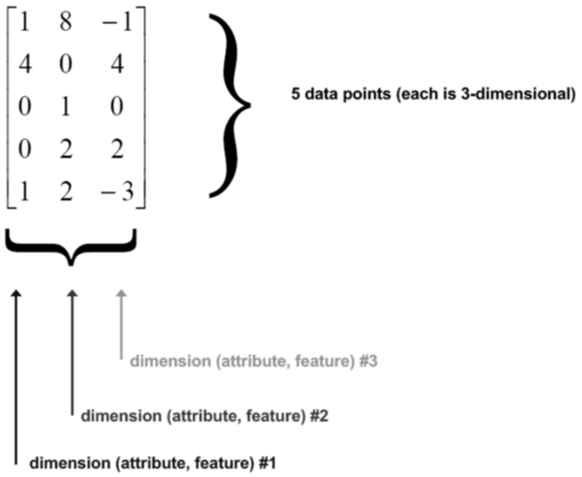
x_2 is temperature

Then, each vector (x_1, x_2) corresponds to <u>one patient</u>

Dataset:

$$\begin{bmatrix} 3 & 5 \\ 7 & 4 \end{bmatrix}$$

Let's digress even more...



$$\begin{bmatrix} 1 & 8 & -1 \\ 4 & 0 & 4 \\ 0 & 1 & 0 \\ 0 & 2 & 2 \\ 1 & 2 & -3 \end{bmatrix}$$

5 data points (each is 3-dimensional)

dimension (attribute, feature) #3

dimension (attribute, feature) #2

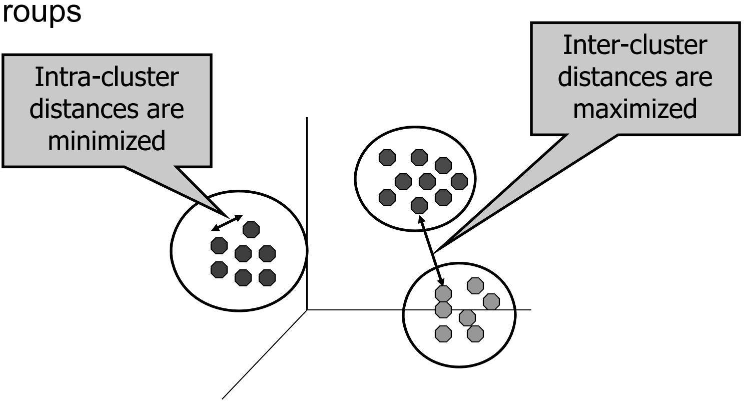
dimension (attribute, feature) #1

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.

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

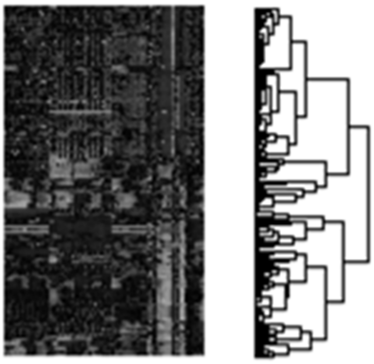


Intra-cluster distances are minimized

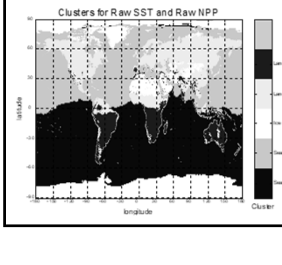
Inter-cluster distances are maximized

Applications of Cluster Analysis

- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.

Google News - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://news.google.com/
Go
Links Customize Links Free Hotmail Windows
Google News BETA
Web Images Groups Directory News
Search News Search the Web
>Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health
Internet

Search and browse 4,500 news sources updated continuously.

Top Stories Auto-generated 8 minutes ago

Four more SARS deaths in Hong Kong
The Hindu - 15 minutes ago
Beijing, April. 18. (PTI): Four more people died of killer respiratory disease SARS in Hong Kong today, raising the toll to 69 even as 30 new cases were reported, Health Department officials said.

Channel News Asia

Bechtel Gets Huge Contract for Iraq Work
Salt Lake Tribune - and 294 related »

'Unlinked' SARS cases hits Toronto condo
Canada.com - and 25 related »

San Marino GP notebook

http://www.channelnewsasia.com/stories/economicnews/view/37827/1/.html

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.

Enron email dataset

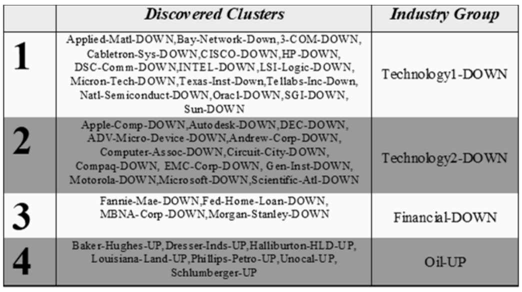
Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering of S&P 500 Stock Data

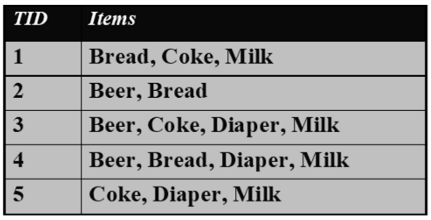
⌘Observe Stock Movements every day.
⌘Clustering points: Stock-{UP/DOWN}
⌘Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
⌘We used association rules to quantify a similarity measure.



	Discovered Clusters	Industry Group
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Association Rule Discovery: Definition

- 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.



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}

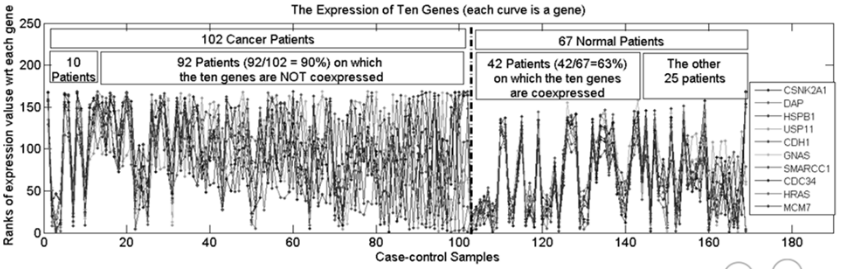
Association Analysis: Applications

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

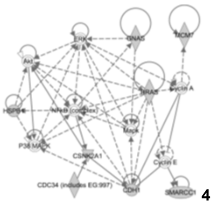
Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et a 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)



[Fang et al PSB 2010]

Association Rule Discovery: Application 1

- **Marketing and Sales Promotion:**
  - **Let the rule discovered be**

    {Bagels, … } --> {Potato Chips}
  - **Potato Chips as consequent => Can be used to determine what should be done to boost its sales.**
  - **Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.**
  - **Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!**

Association Rule Discovery: Application 2

- 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.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

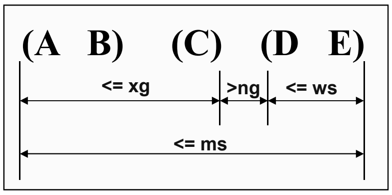
- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery: Definition

- 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)

- Rules are formed by first disovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential Pattern Discovery: Examples

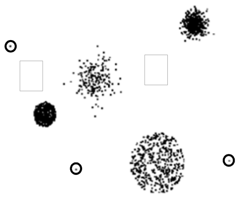
- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)
    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

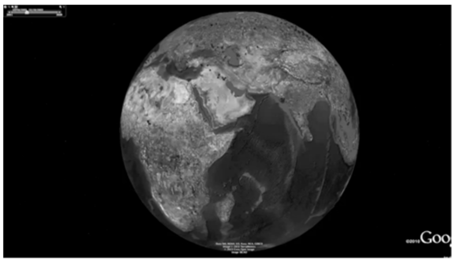
Regression

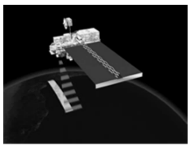
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  -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

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.







Motivating Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis