

# Data Mining: Introduction

---

## Lecture Notes for Chapter 1

Introduction to Data Mining, 2<sup>nd</sup> Edition

by

Tan, Steinbach, Karpatne, Kumar

# Large-scale Data is Everywhere!

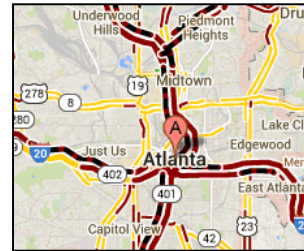
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected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



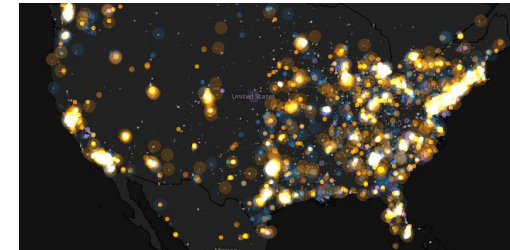
**Cyber Security**



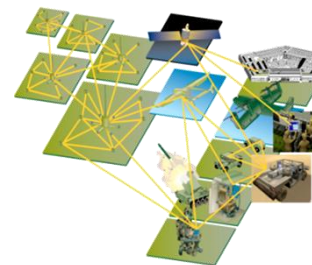
**E-Commerce**



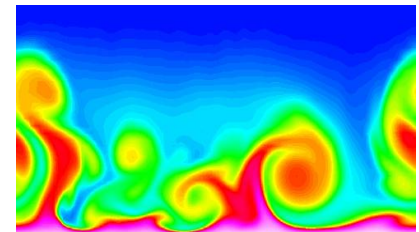
**Traffic Patterns**



**Social Networking: Twitter**



**Sensor Networks**



**Computational Simulations**

# Why Data Mining? Commercial Viewpoint

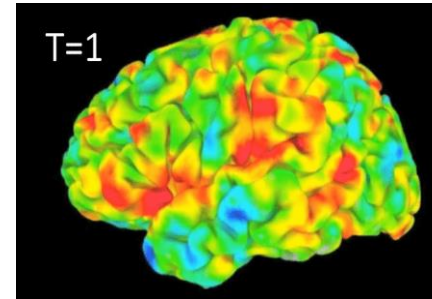
---

- 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

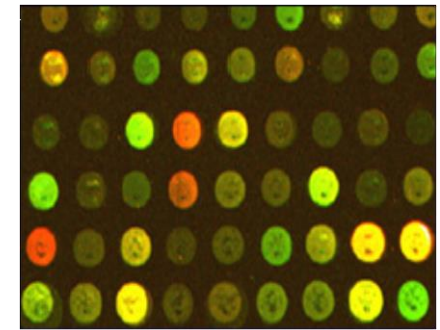
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - ◆ NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - ◆ Sky survey data
  - High-throughput biological data
  - scientific simulations
    - ◆ terabytes of data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



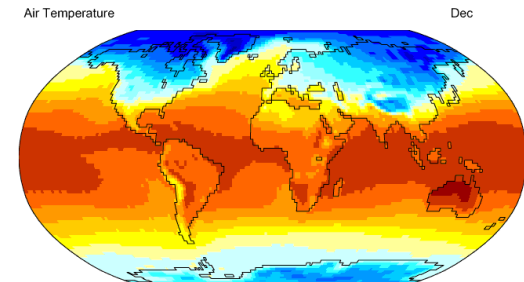
fMRI Data from Brain



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

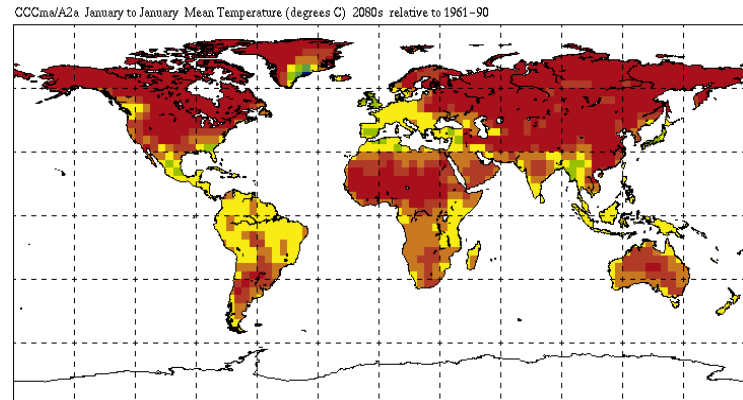
*Big data—capturing its value*

- \$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**



**Predicting the impact of climate change**



**Finding alternative/ green energy sources**

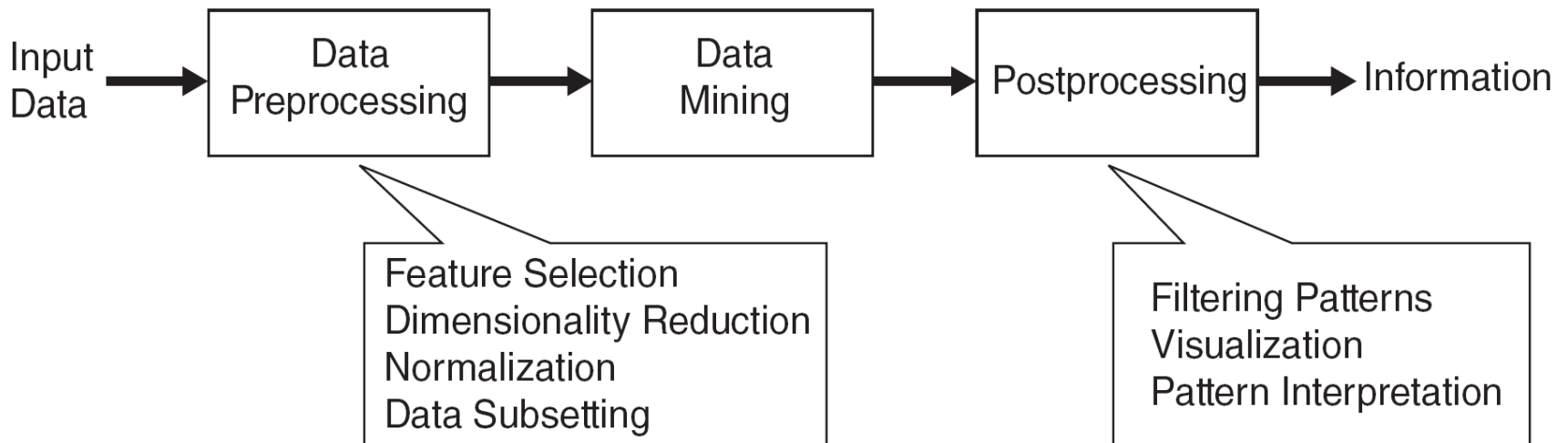


**Reducing hunger and poverty by increasing agriculture production**

# 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



# 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?**

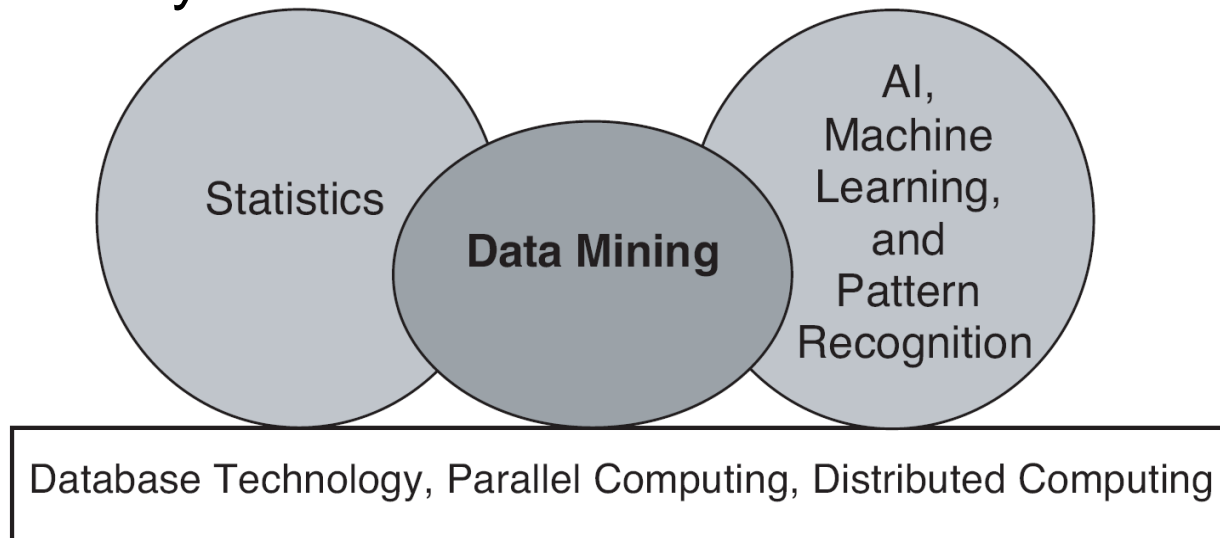
- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)



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



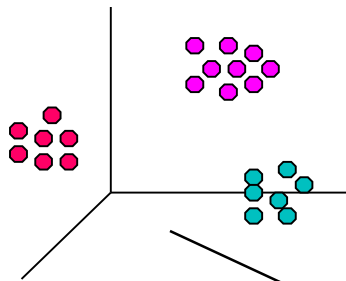
# Data Mining Tasks

---

- 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 ...



Clustering

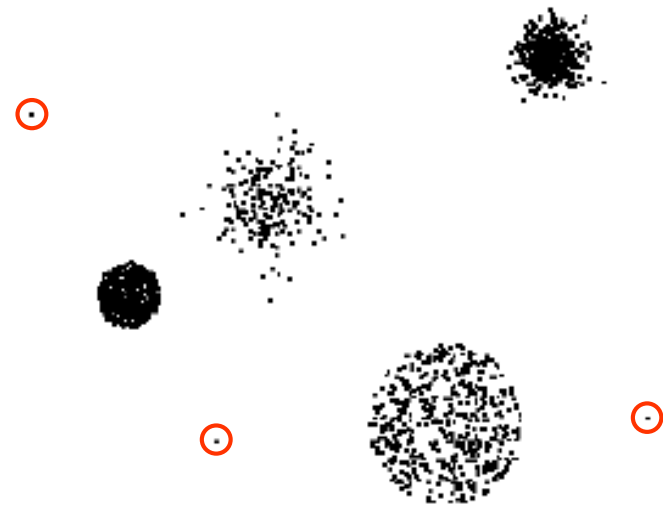
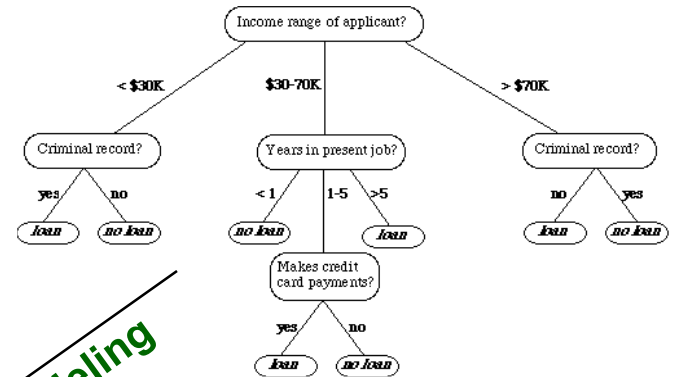
## Data

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

Association Rules

Predictive Modeling

Anomaly Detection



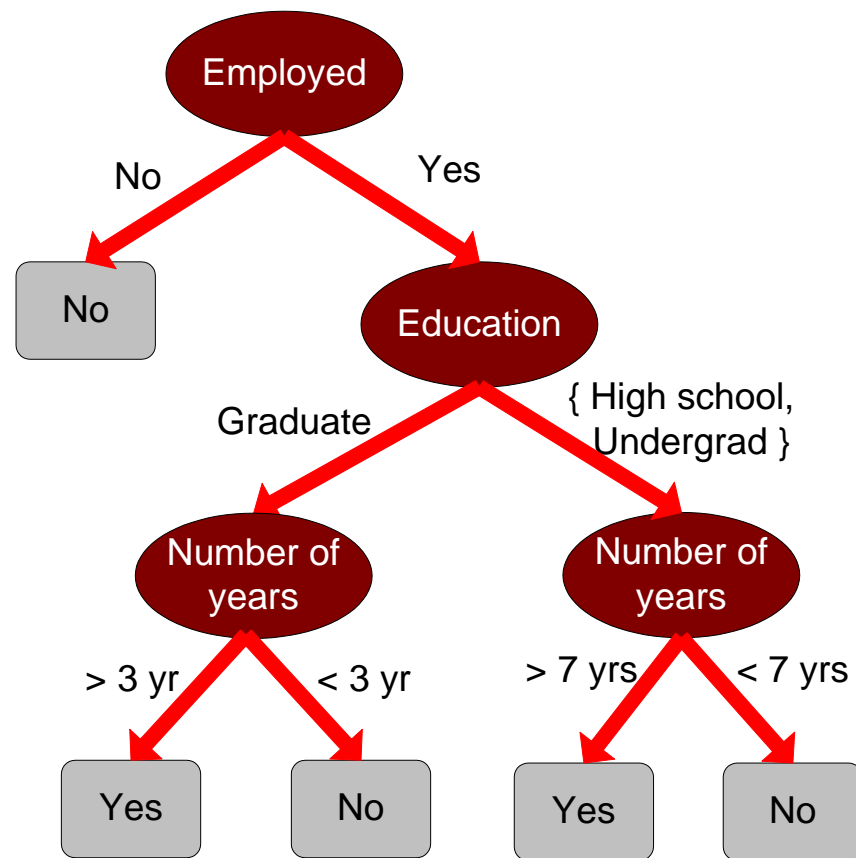
# Predictive Modeling: Classification

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

**Class**

<i>Tid</i>	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**

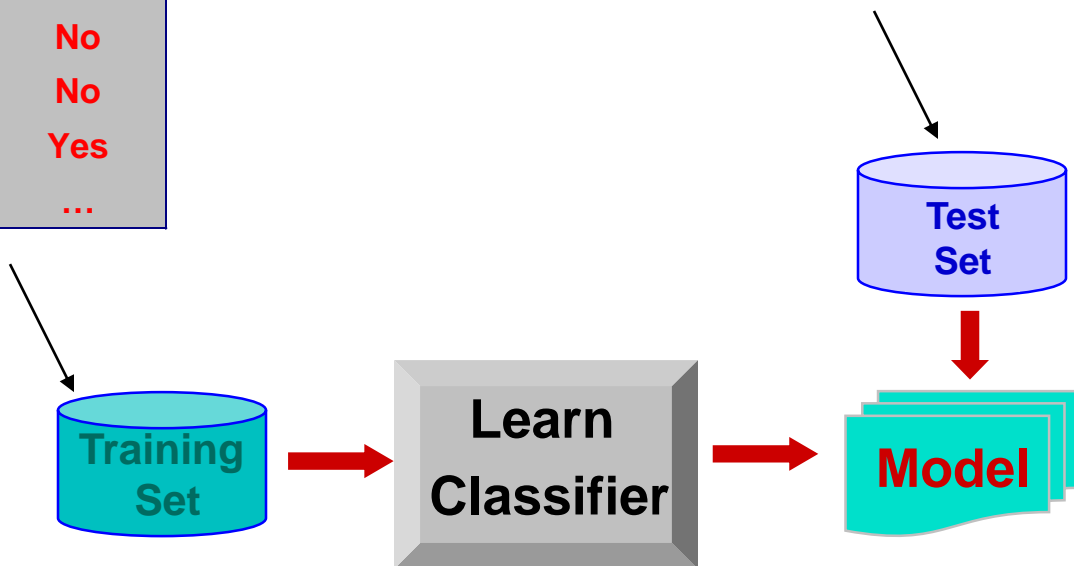


# Classification Example

categorical      categorical      quantitative      class

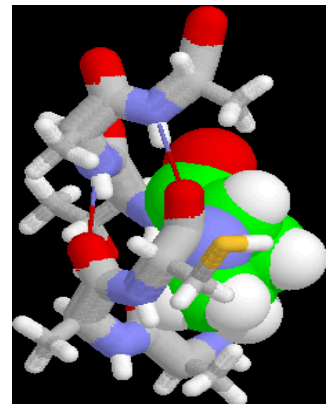
<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...	...	...	...	...

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...	...	...	...	...



# 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 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

---

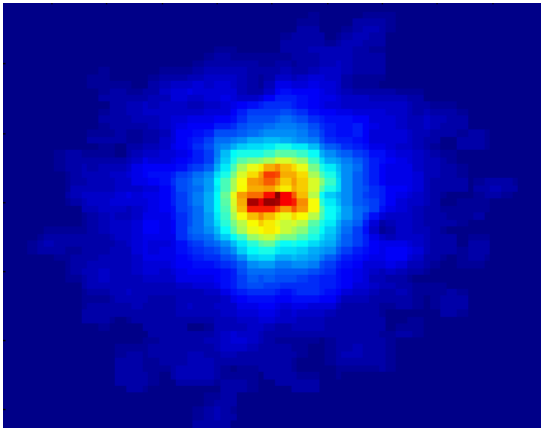
- Sky Survey Cataloging
  - **Goal:** To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Approach:**
    - ◆ Segment the image.
    - ◆ Measure 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>

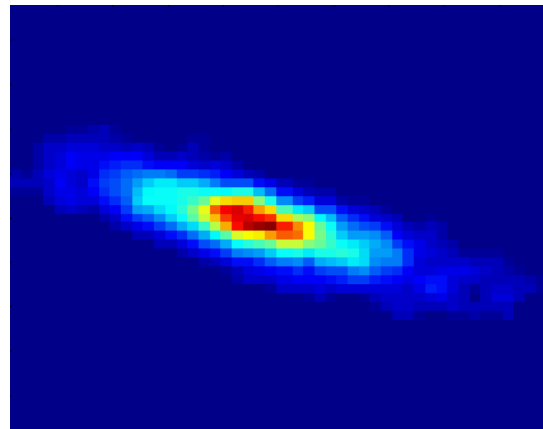
*Early*



**Class:**

- Stages of Formation

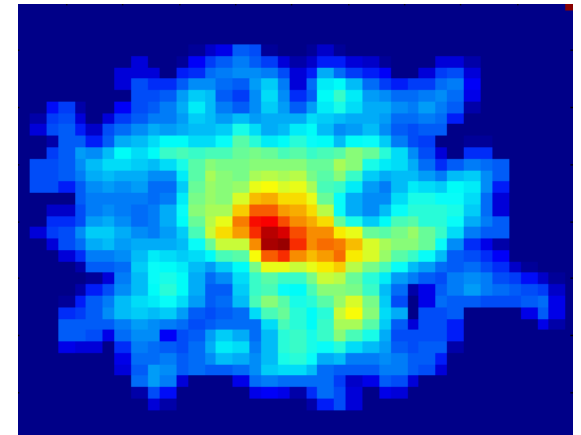
*Intermediate*



**Attributes:**

- Image features,
- Characteristics of light waves received, etc.

*Late*



**Data Size:**

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

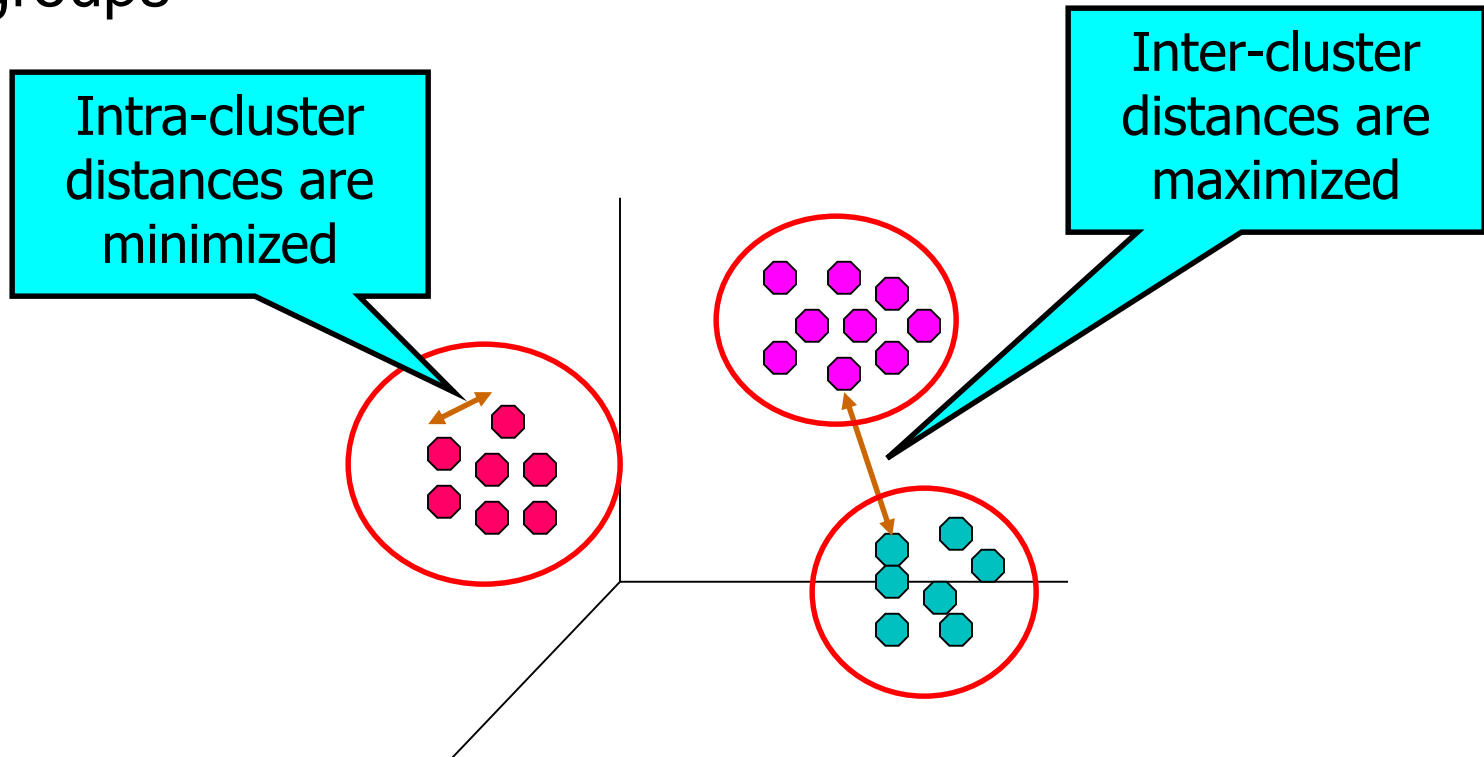
# Regression

---

- 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 advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

# 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



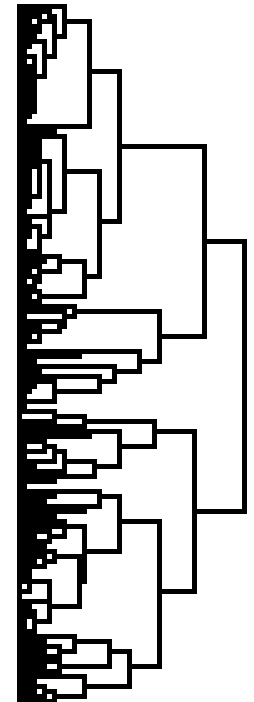
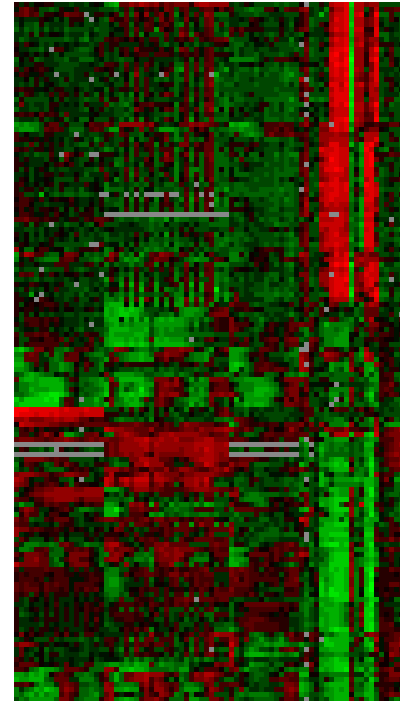
# Applications of Cluster Analysis

- **Understanding**

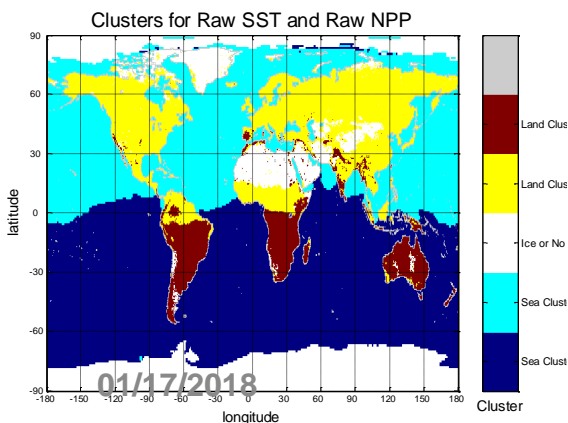
- Custom profiling for targeted marketing
- Group related documents for browsing
- Group genes and proteins that have similar functionality
- Group stocks with similar price fluctuations

- **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.

Introduction to Data Mining, 2nd Edition

A screenshot of a Google News page displayed in a Microsoft Internet Explorer browser window. The page shows the Google News logo, search bars, and a list of top stories. The main story is titled "Four more SARS deaths in Hong Kong" with a sub-headline "The Hindu - 15 minutes ago" and a snippet: "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." Other stories include "Bechtel Gets Huge Contract for Iraq Work" and "'Unlinked' SARS cases hits Toronto condo". The page is dated "Auto-generated 8 minutes ago".

# Clustering: Application 1

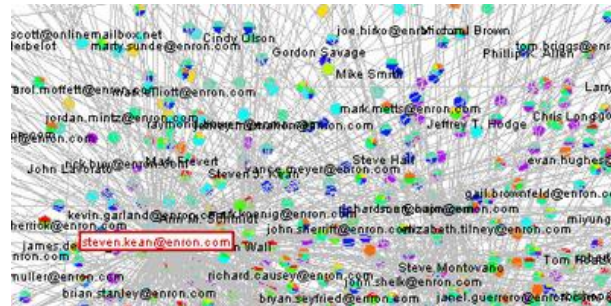
---

- 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



# 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.

<i>TID</i>	<i>Items</i>
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

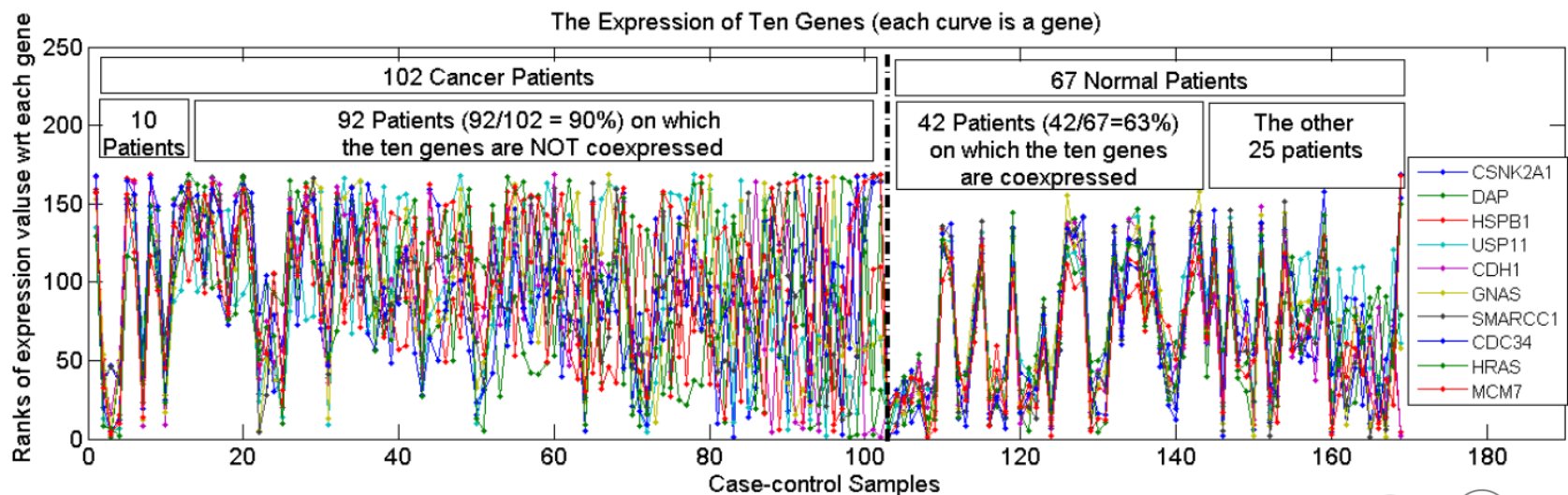
---

- 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

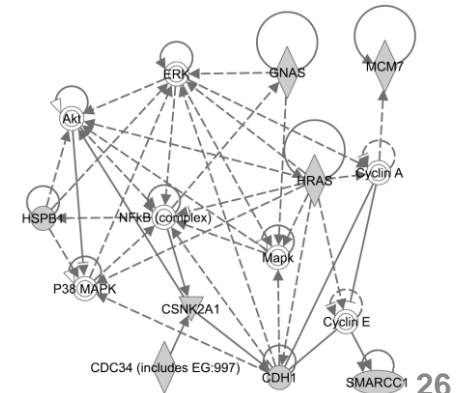
- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

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



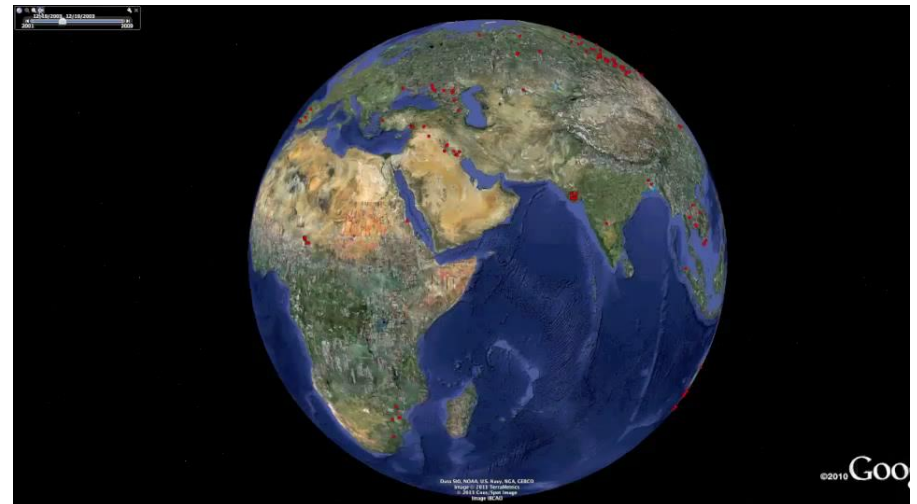
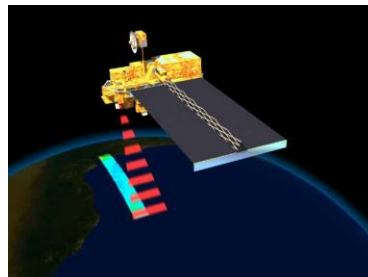
Enriched with the TNF/NFB signaling pathway  
 which is well-known to be related to lung cancer  
 P-value:  $1.4 \cdot 10^{-5}$  (6/10 overlap with the pathway)

[Fang et al PSB 2010]



# 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