# Introduction to Data Mining

# 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.
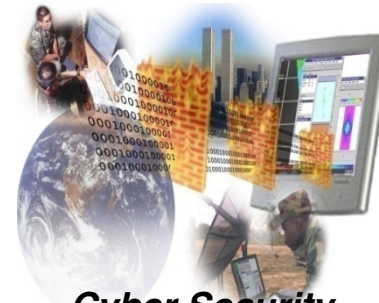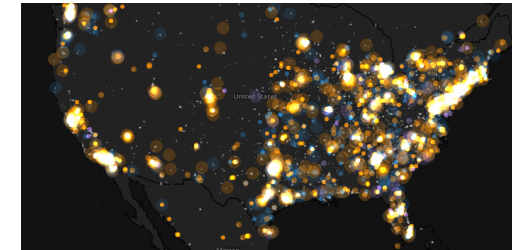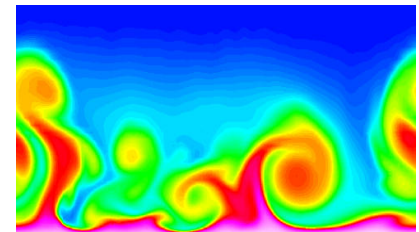

Cyber Security

E-Commerce
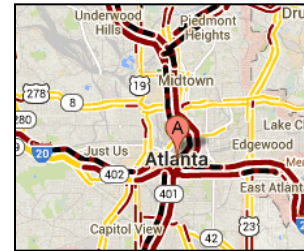
Traffic Patterns

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
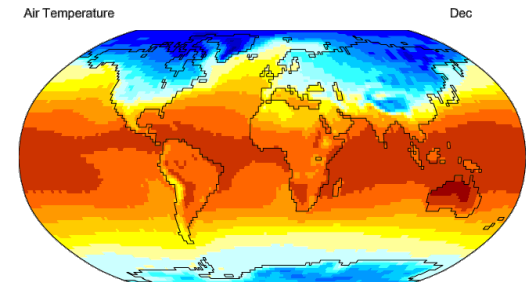 - In hypothesis formation

T=1
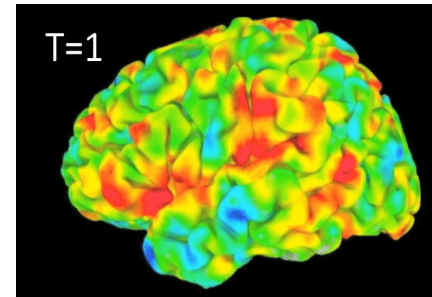
fMRI Data from Brain

Sky Survey Data
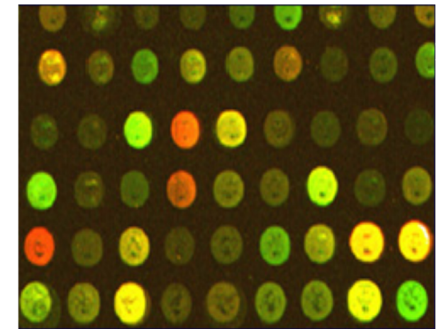
Gene Expression Data

Air Temperature   Dec
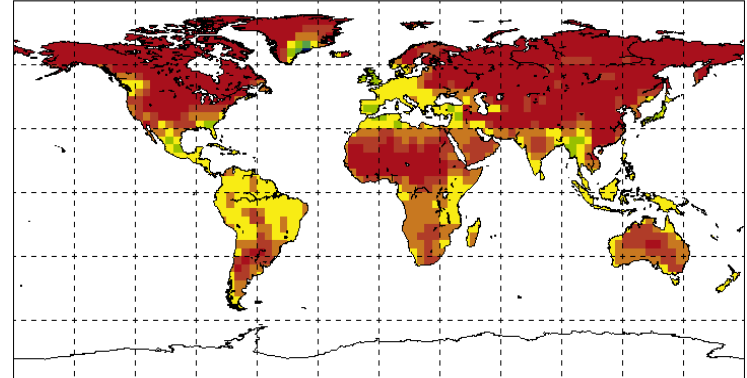
Surface Temperature of Earth

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

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
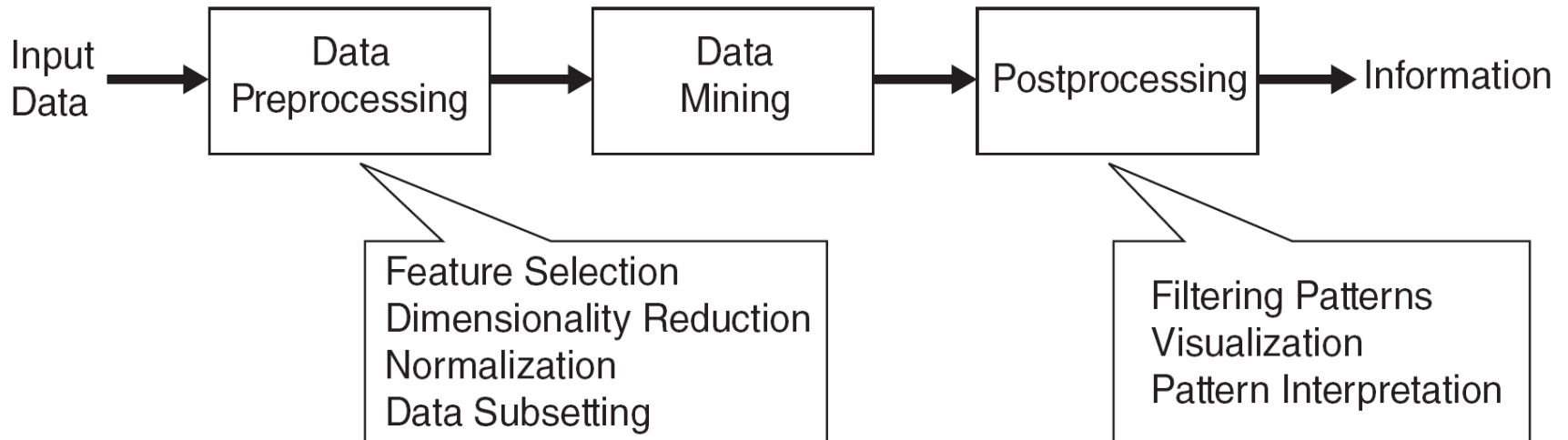
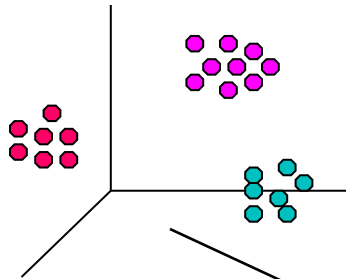# Data Mining Tasks

- 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

Predictive Modeling

Association Rules

Anomaly Detection

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

Milk
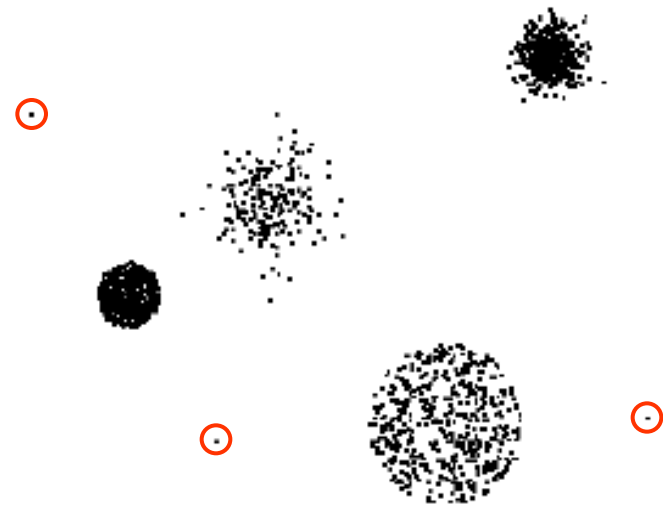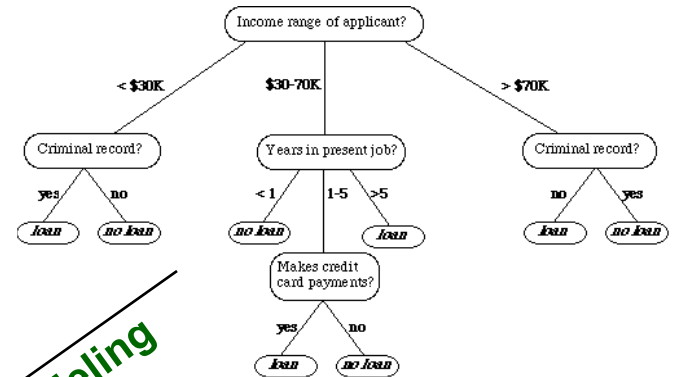
Predictive Modeling: Classification

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

Model for predicting credit worthiness

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
…	…	…	…	…
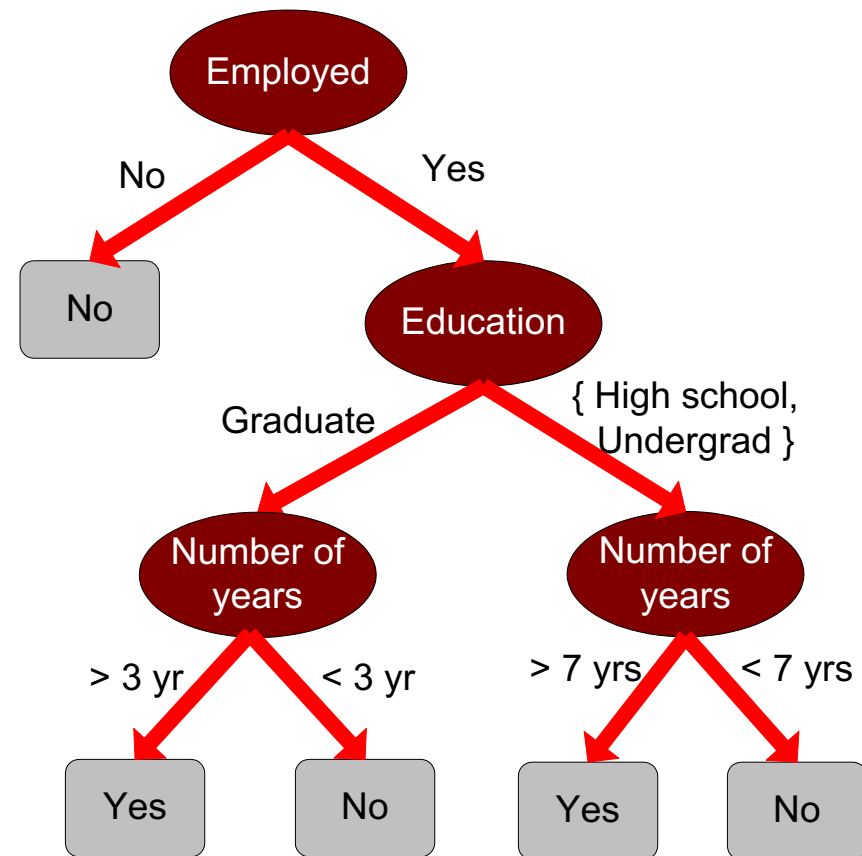
Classification Example

| | 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
…	…	…	…	…

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

Training Set → Learn Classifier → Model

Test Set → Model
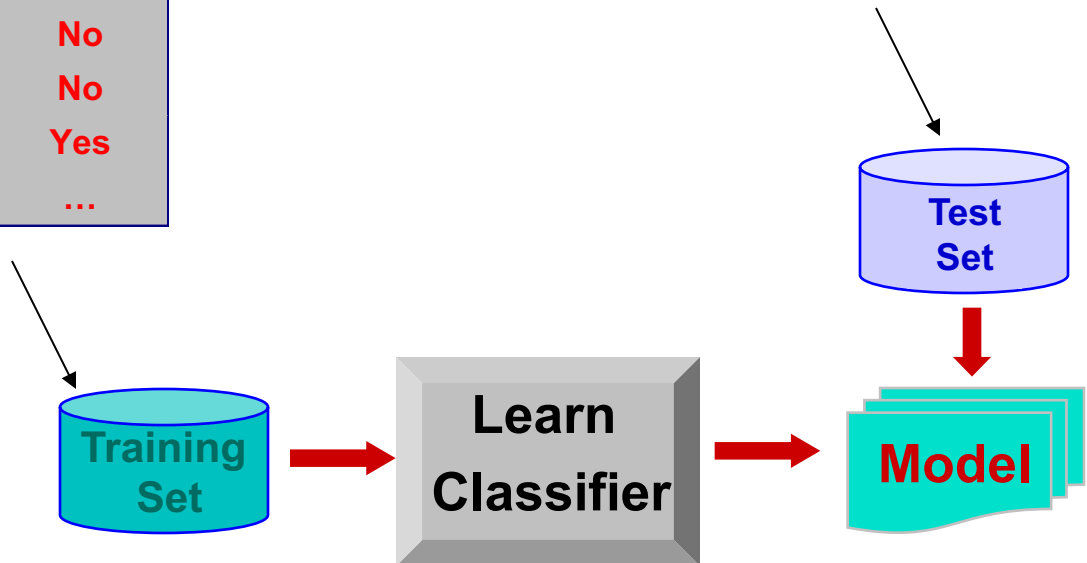
Examples of Classification Task

- 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
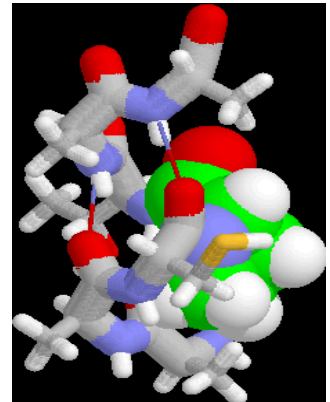
Regression

- 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

Intra-cluster distances are minimized

Inter-cluster distances are maximized
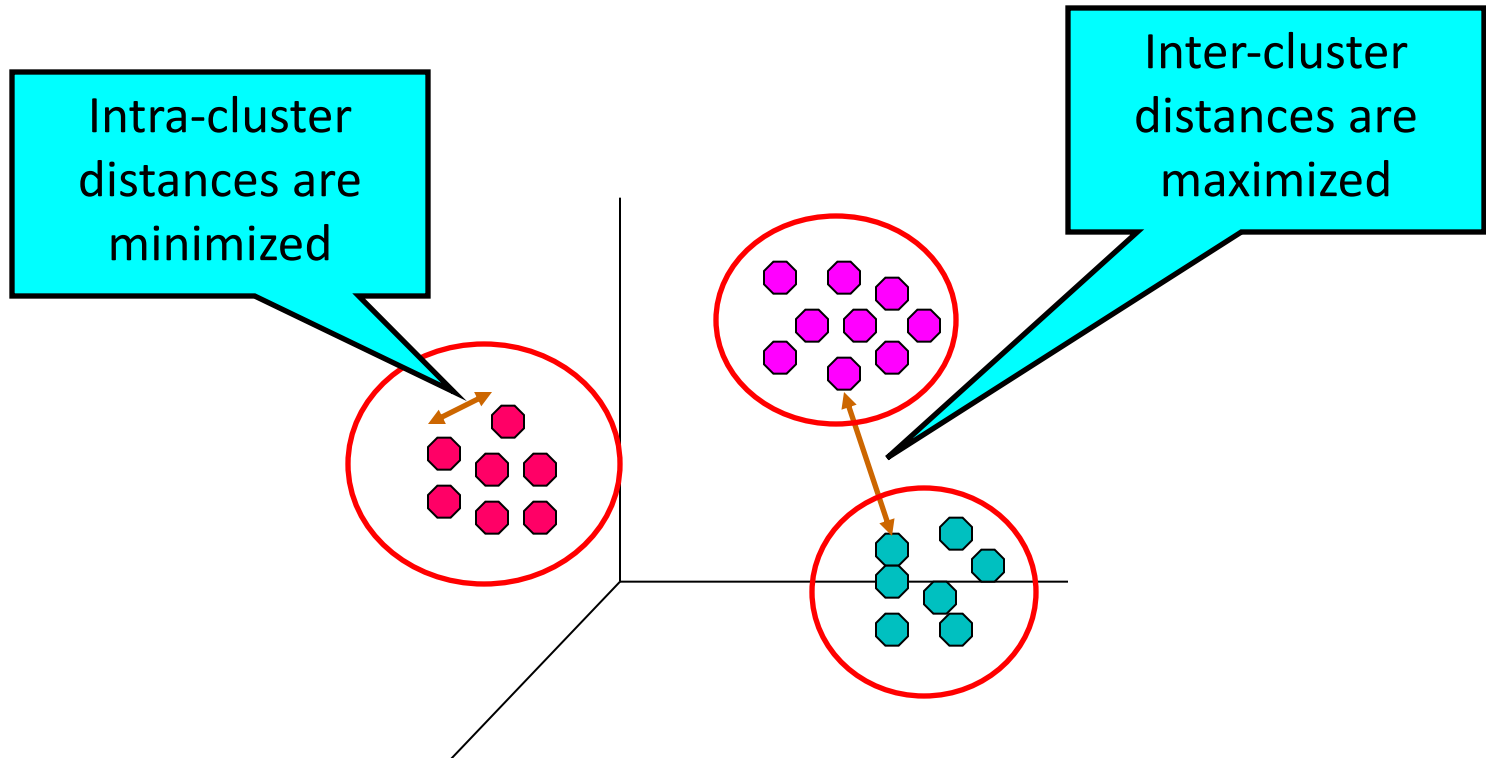
Applications of Cluster Analysis

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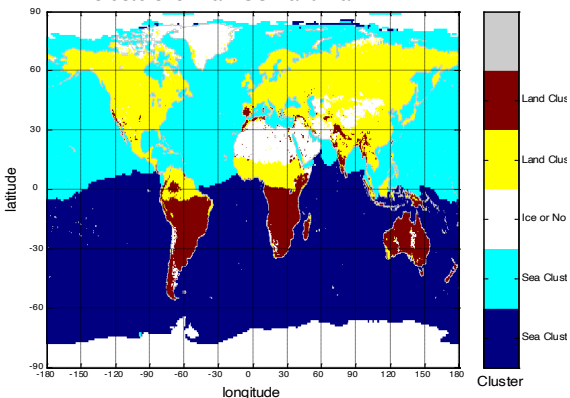 - Group stocks with similar price fluctuations

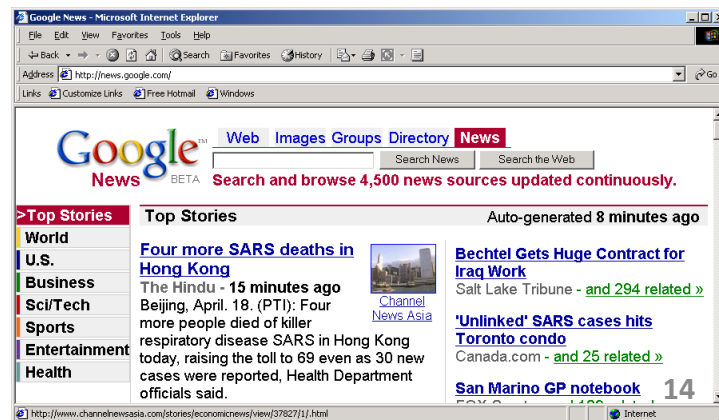- **Summarization**
  - Reduce the size of large data sets
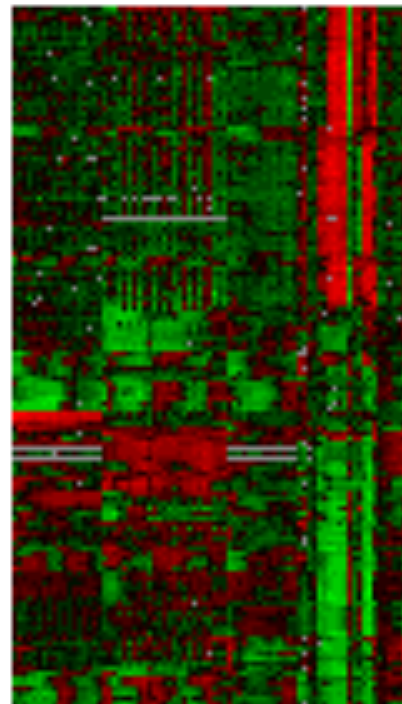
Courtesy: Michael Eisen

Clusters for Raw SST and Raw NPP

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

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.
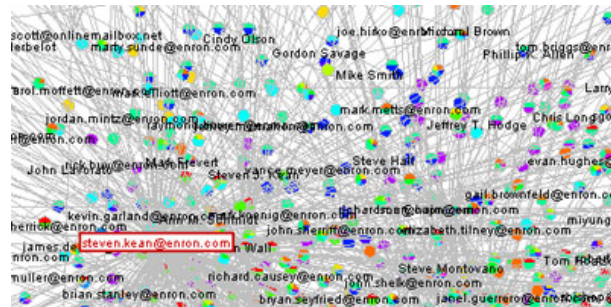
Association Rule Discovery: Definition

- 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

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 fores cover.
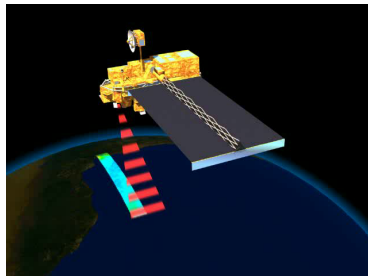
Motivating Challenges

- Scalability

- High Dimensionality

- Heterogeneous and Complex Data

- Data Ownership and Distribution

- Non-traditional Analysis