Data Mining:

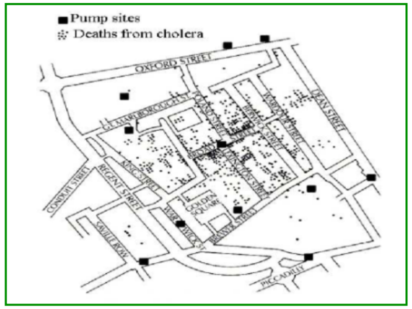
1. Einführung

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
  - Collect 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.

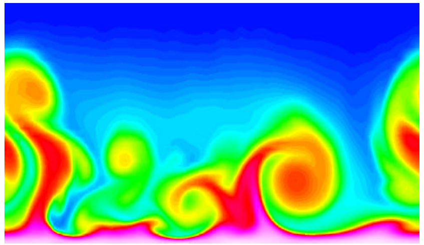
Homeland Security

Business Data



Geo-spatial data

Sensor Networks

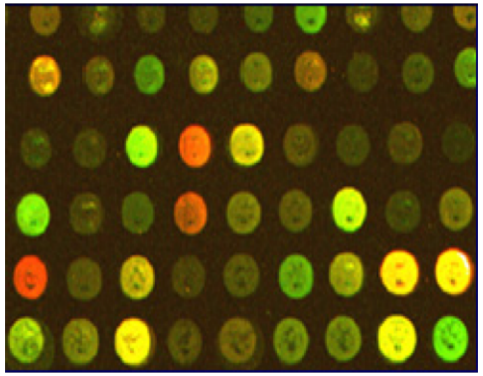


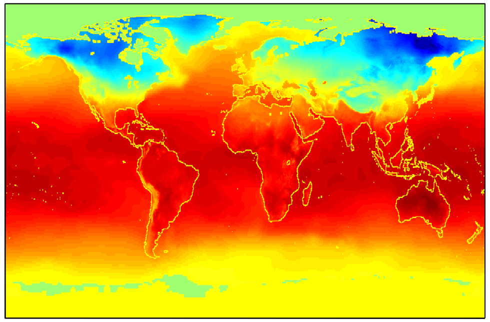
Computational Simulations

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has over a billion active users
  - Purchases at department/ grocery stores, E-commerce
    - Amazon records millions of items/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

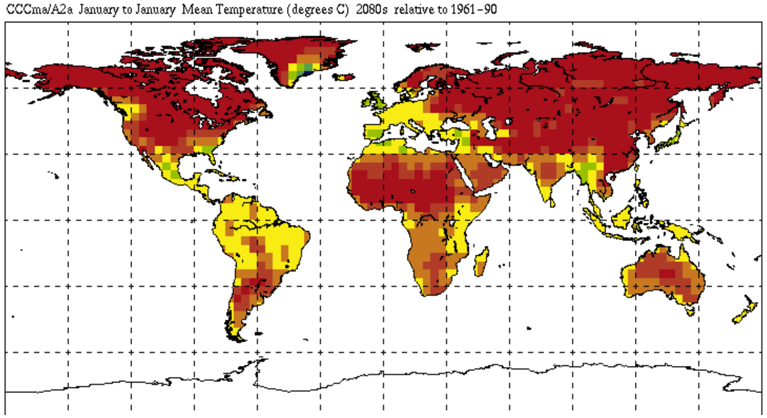




- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - high-throughput biological data
  - scientific simulations
    - terabytes of data generated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

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

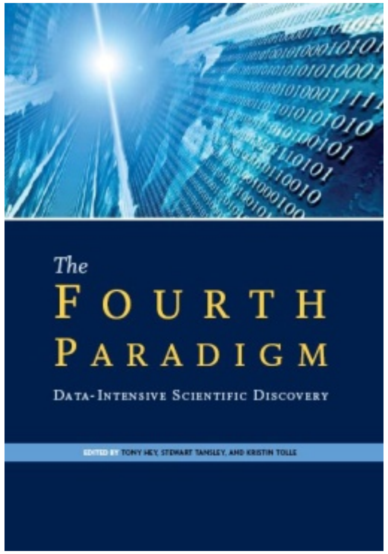
Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

Data Guided Discovery - A new paradigm

"... data-intensive science [is] …a new, fourth paradigm for scientific exploration." - Jim Gray


The FOURTH PARADIGM

Data-Intensive Scientific Discovery

Edited by Tony Hey, Stewart Tansley, and Kristin Tolle

McKinsey Global Institute

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

WIRED MAGAZINE: 16.07

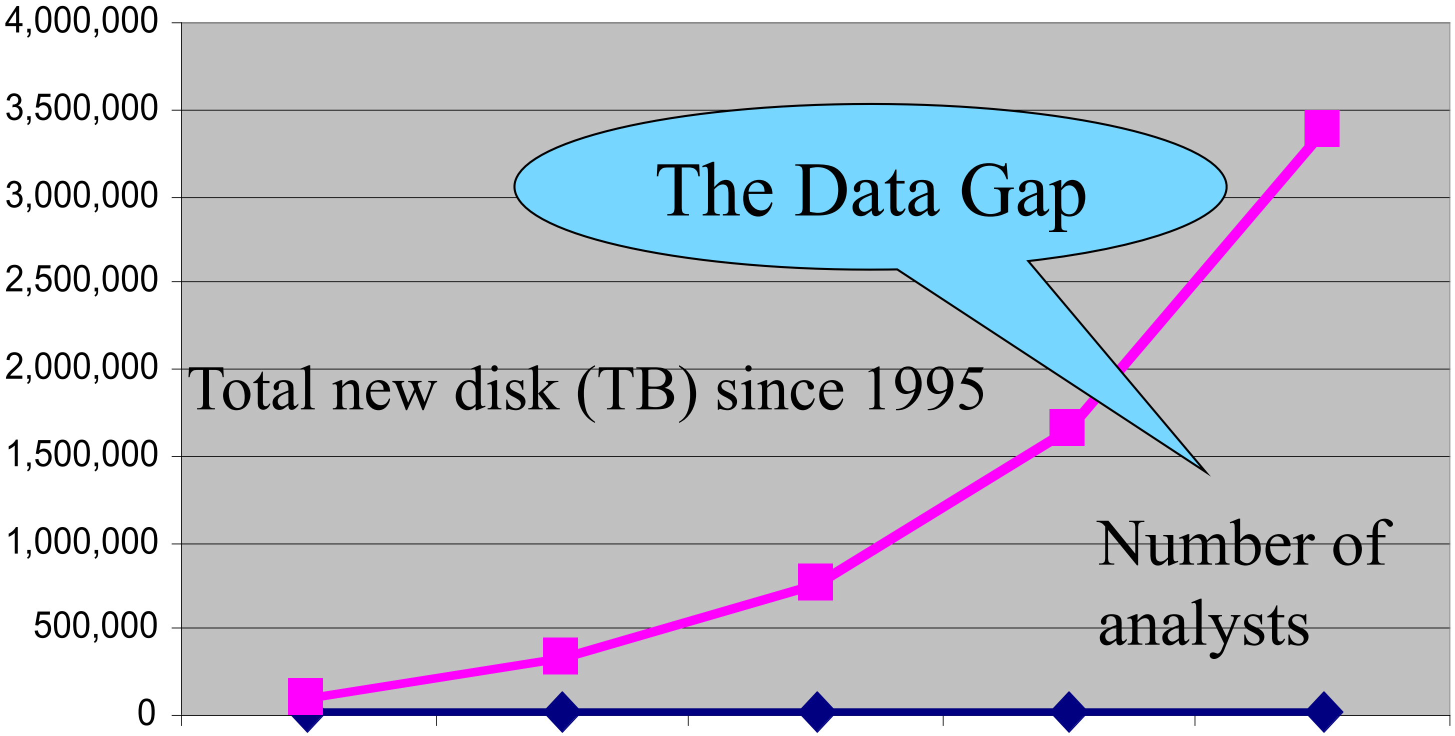
SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Mining Large Data Sets - Motivation

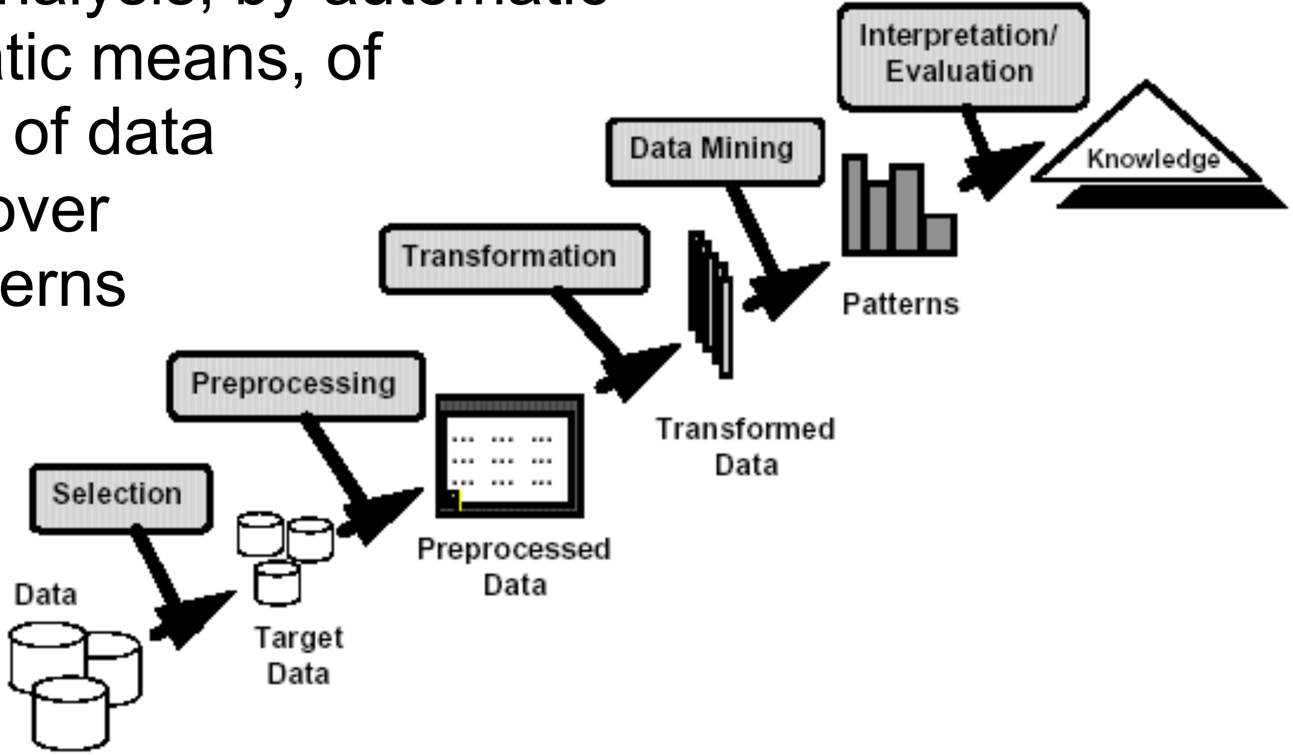
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications” [1995-1999]

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
  - Core of the knowledge discovery process :



What is (not) Data Mining?

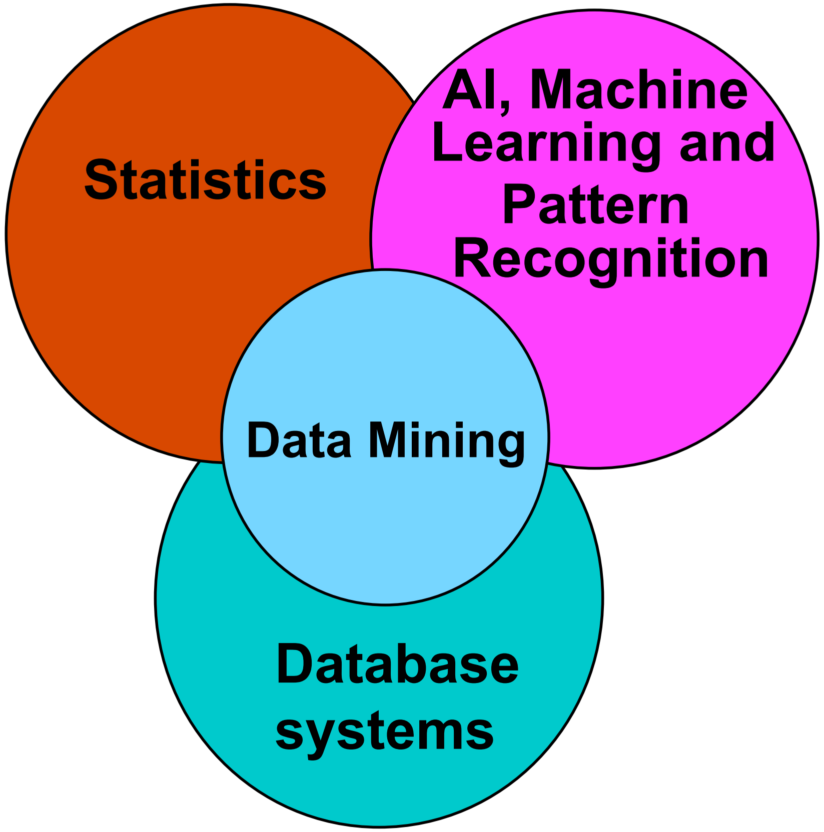
- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **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 to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed



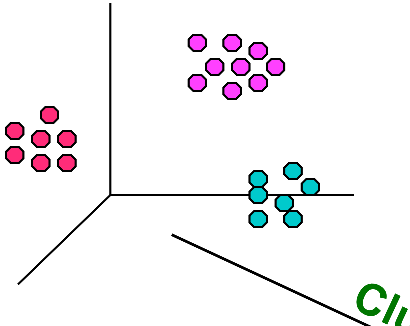
Data Mining Tasks

- 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

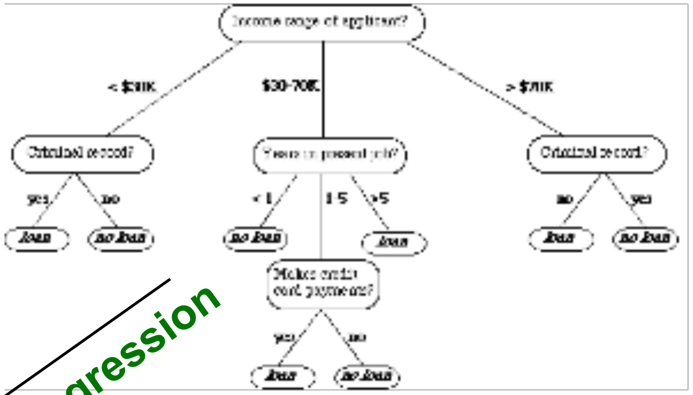
Core Data Mining Tasks …



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

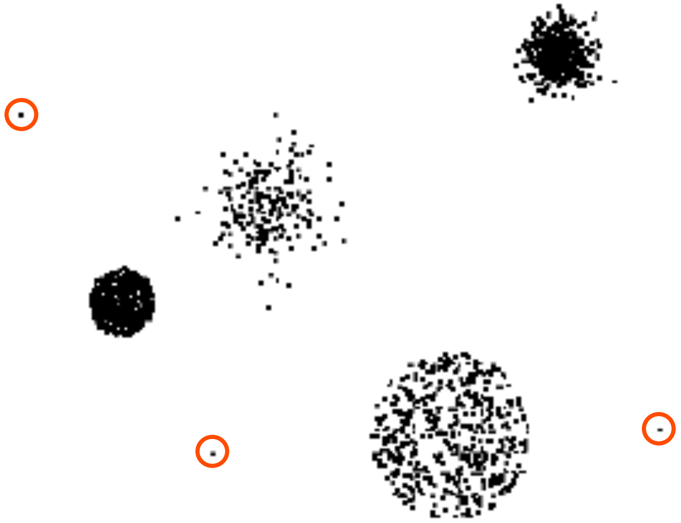


Classification/Regression

Association Rules

Anomaly Detection

Milk
Pampers
Baby-Dry
Mega
6
48



Core Data Mining Tasks... [in dieser Vorlesung]

- Frequent Pattern Discovery [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Classification [Predictive]
- Regression [Predictive]
- Clustering [Descriptive]
- Anomaly/Outlier Detection [Predictive]

Association Rule Discovery: Definition

- Given a set of records ("transactions", e.g. individual purchases) each of which contain some number of items from a given collection.
- Produce *dependency rules* which describe occurrence of an item based on occurrences of other items.
- This requires discovery of *frequent patterns* of itemsets.

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 Rule Discovery: Classic Application

- 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.

    {Diaper, Milk} --> {Beer}

    [So, don't be surprised if you find six-packs stacked next to diapers!]

Association Rule Discovery: Classic Application++

- 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 Analysis: More Applications

- Market-basket analysis
  - Rules are used for shelf management, sales promotion, and also for inventory management, e.g., of repair services

- 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

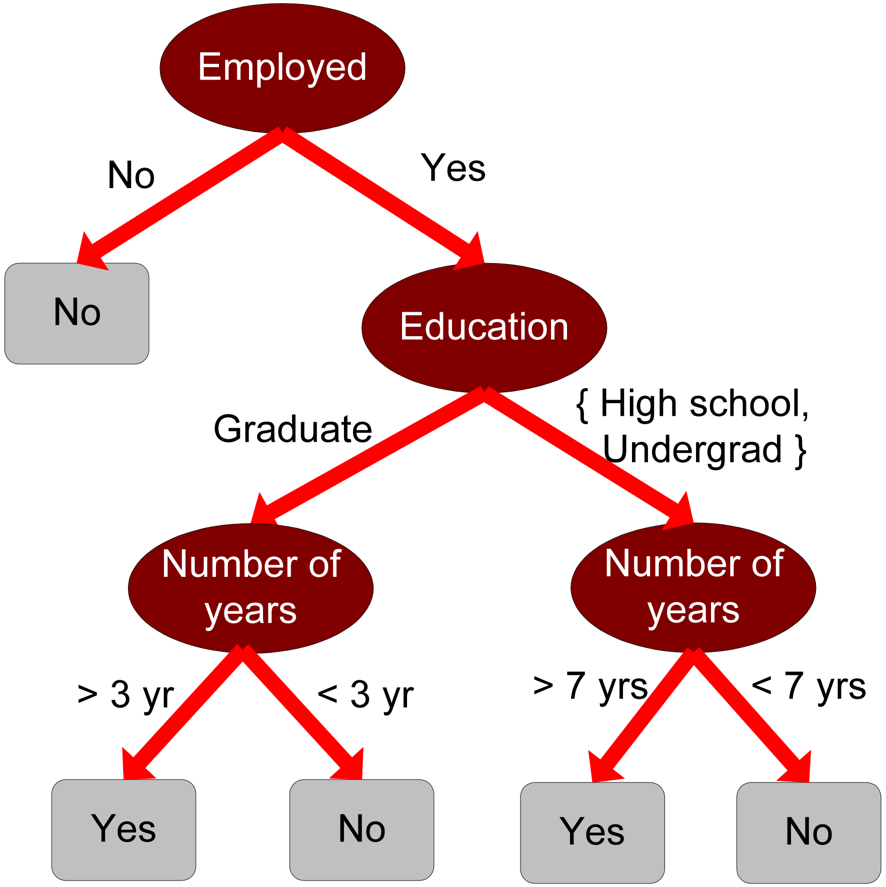
Predictive Modeling: Classification

- Find a model for a 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



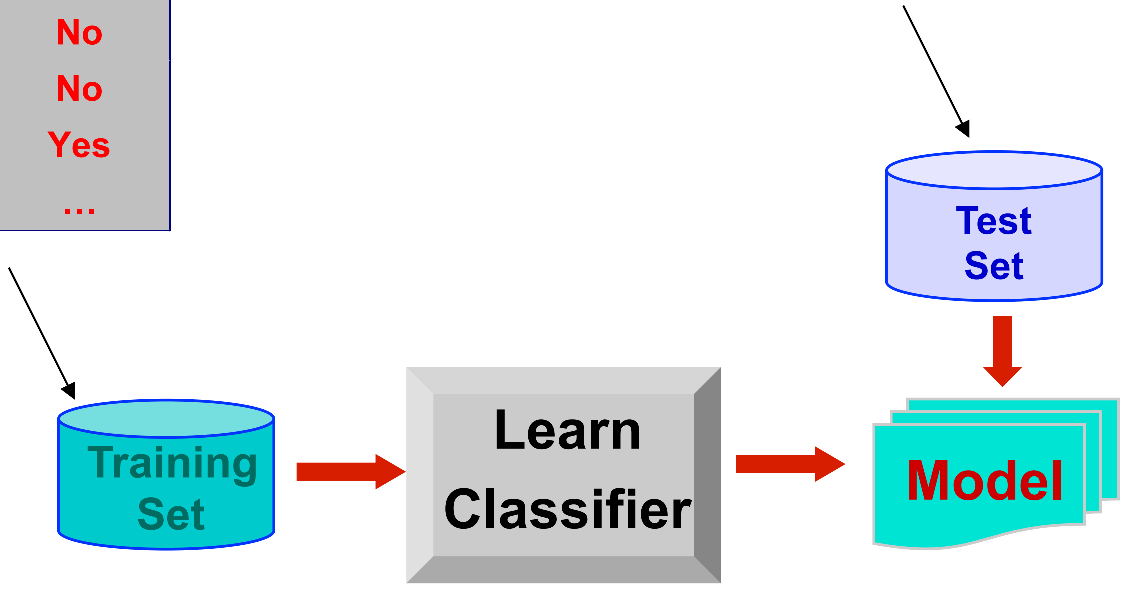
Classification: Definition

- 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.

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	?
…	…	…	…	…



Classification: Application 1

- 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.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

- Customer Attrition/Churn:
  - **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

- 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 4

- 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!

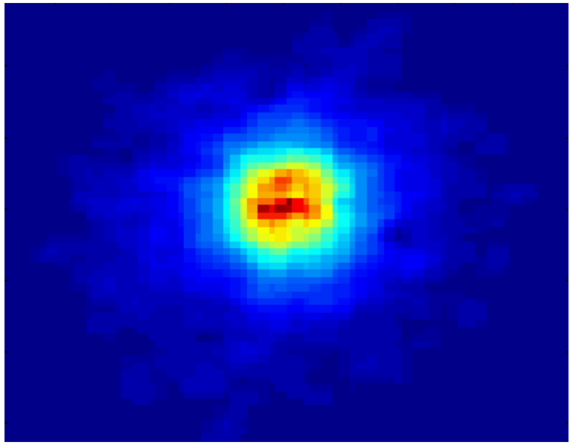
From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classification: Application 5

- Classifying Galaxies

Courtesy: http://aps.umn.edu

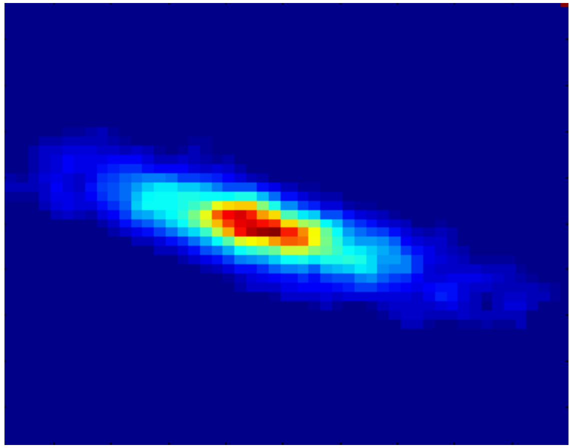
Early



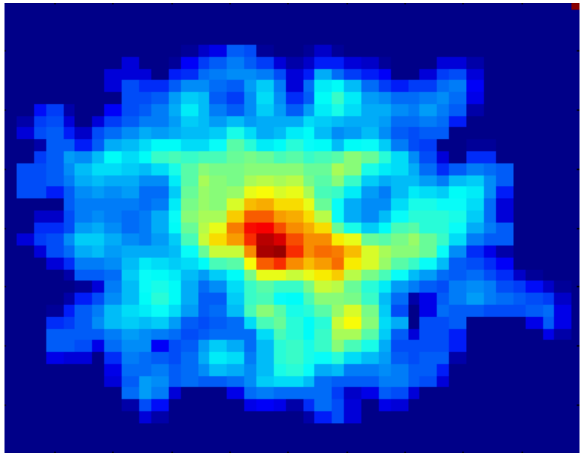
Class:
- **Stages of Formation**

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

Intermediate



Late



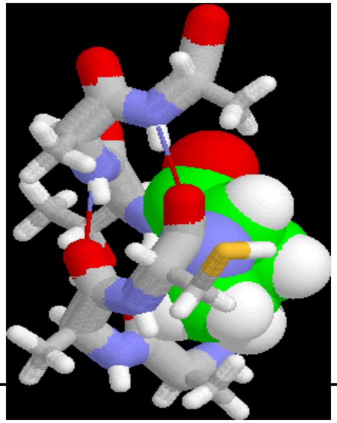
Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Classification: More Applications

- 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



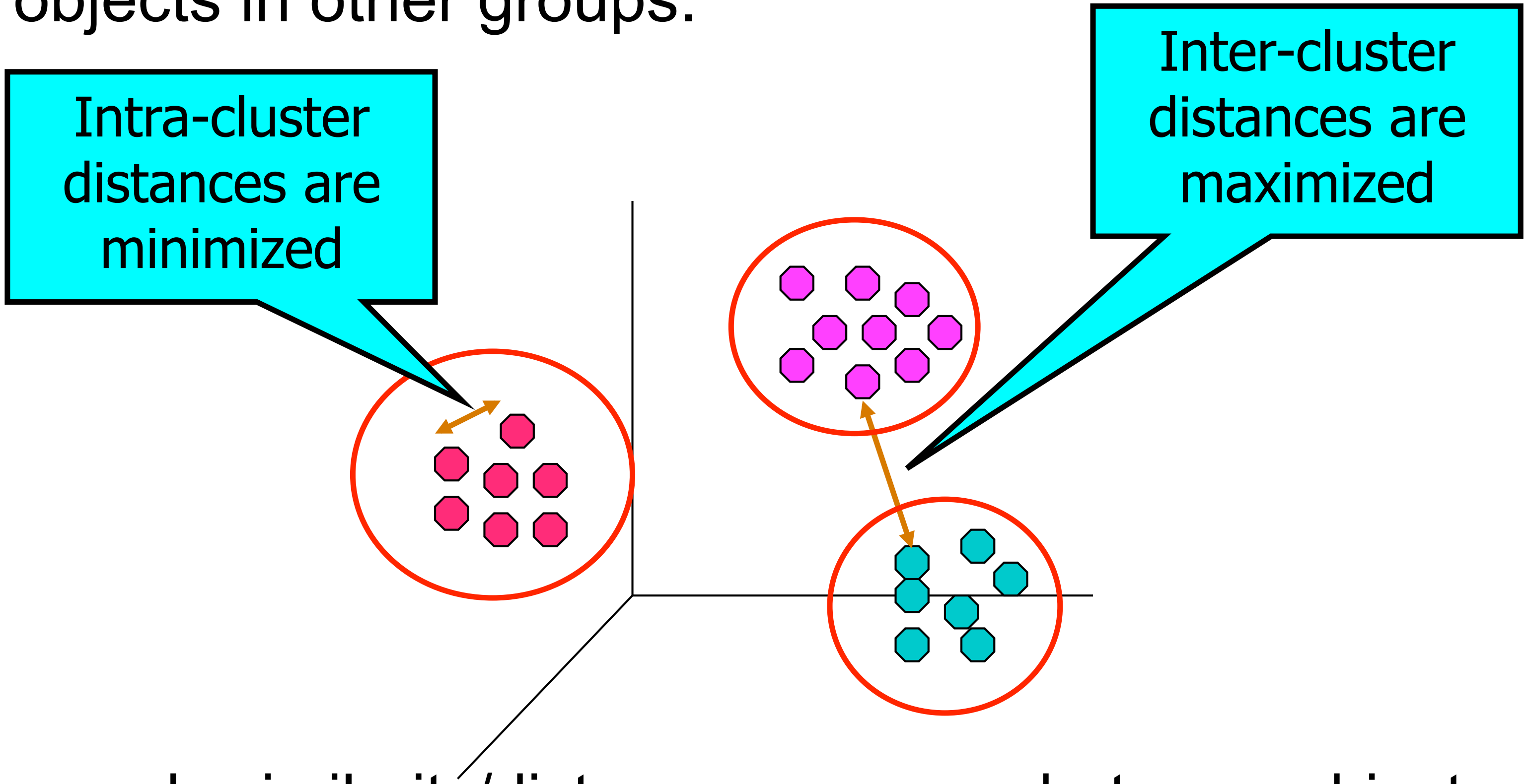
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

Regression

- Classification delivers boolean/categorical class values.
- Here: 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: Definition

- Find groups (*clusters*) of objects such that the objects within a group will be more similar to (or related to) one another and less similar (or unrelated to/different from) the objects in other groups.


Intra-cluster distances are minimized
Inter-cluster distances are maximized

- This needs similarity/distance measures between objects.

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:** 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.

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

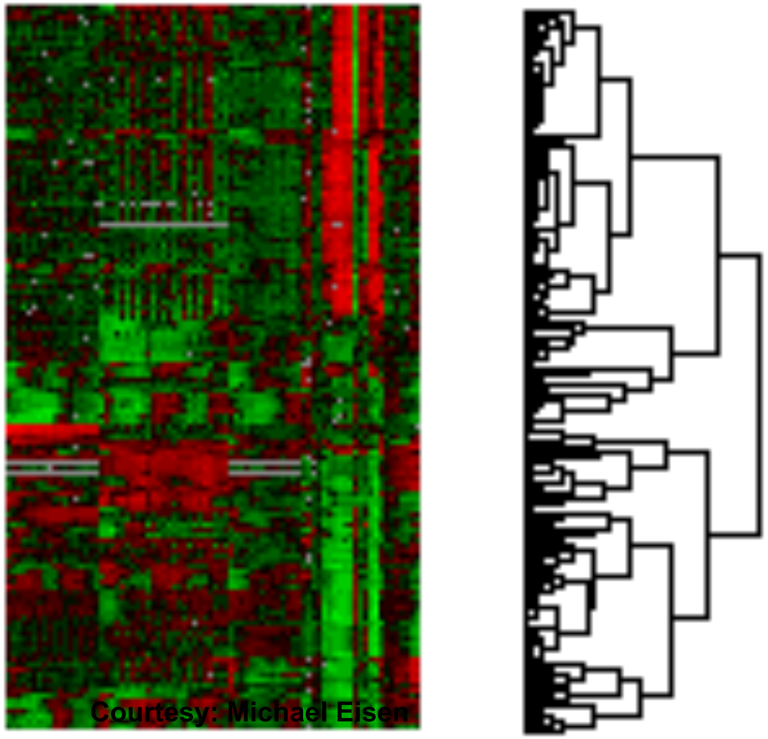
Clustering of S&P 500 Stock Data

- 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.

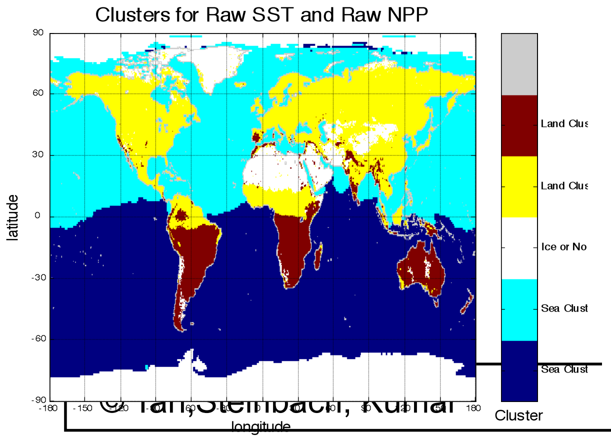
	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, 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

Cluster Analysis: More Applications

- **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

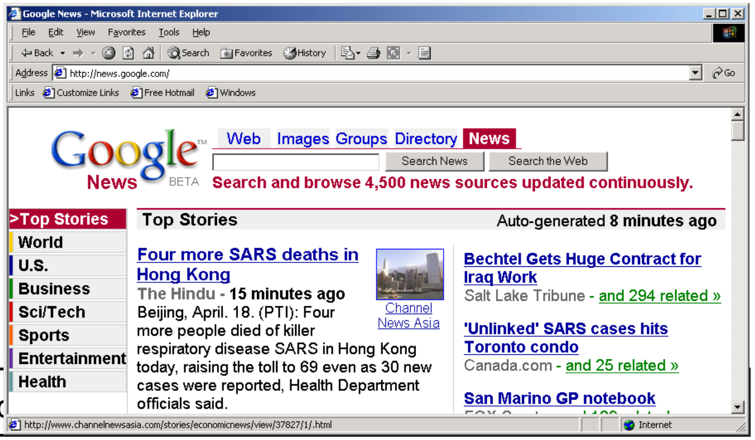


Courtesy: Michael Eisen



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.



http://news.google.com/

Google News BETA

Search and browse 4,500 news sources updated continuously.

>Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health

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 »

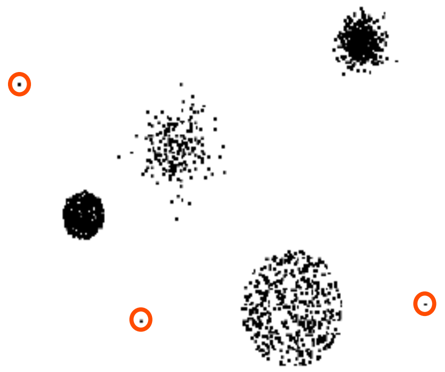
'Unlinked' SARS cases hits Toronto condo
Canada.com - and 25 related »

San Marino GP notebook

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

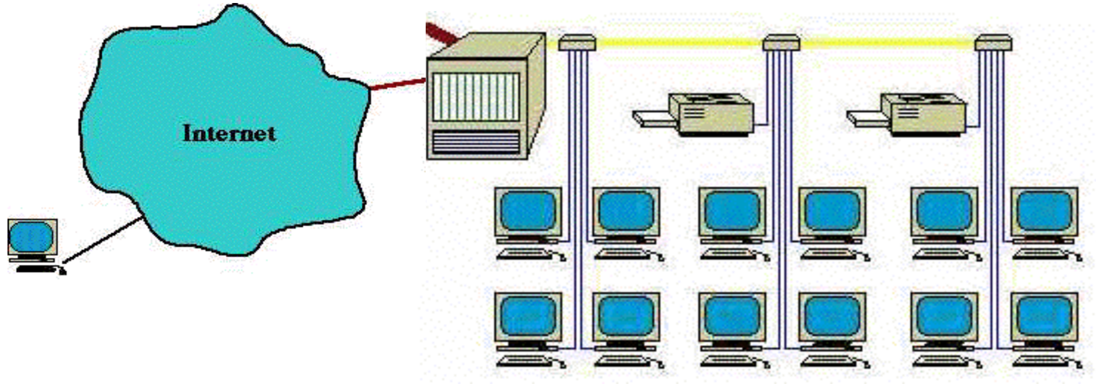
Anomaly/Outlier Detection

- 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



  - Identify anomalous behavior from any kind of sensor network.

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data
- Non-traditional Analysis