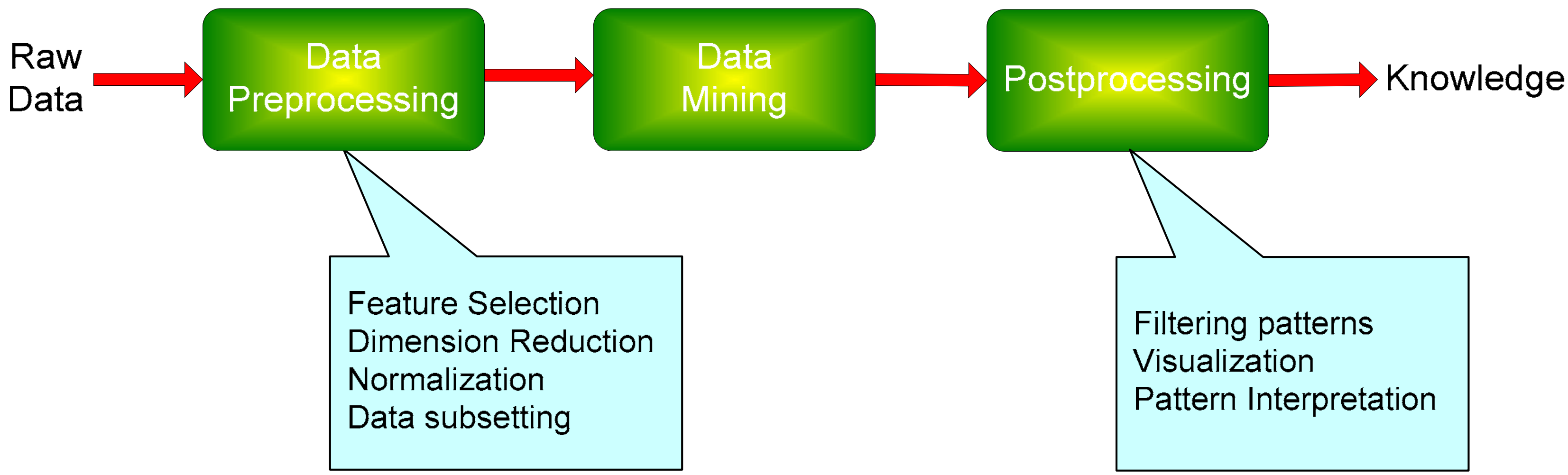
What is Data Mining?

- **Non-trivial extraction of implicit, previously unknown and potentially useful information from data**



Knowledge Discovery in Databases (KDD)

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'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data Explosion

"We are drowning in data, but starving for knowledge"

"The amount of data stored in various media has doubled in three years, from 1999 to 2002. The amount of data put into storage in 2002, five exabytes (one quintillion bytes), was equal to the contents of a half a million new libraries, each containing a digitised version of the print collection of the entire US Library of Congress" (Lyman and Varian, UC Berkeley, 2003)

Scale of Data

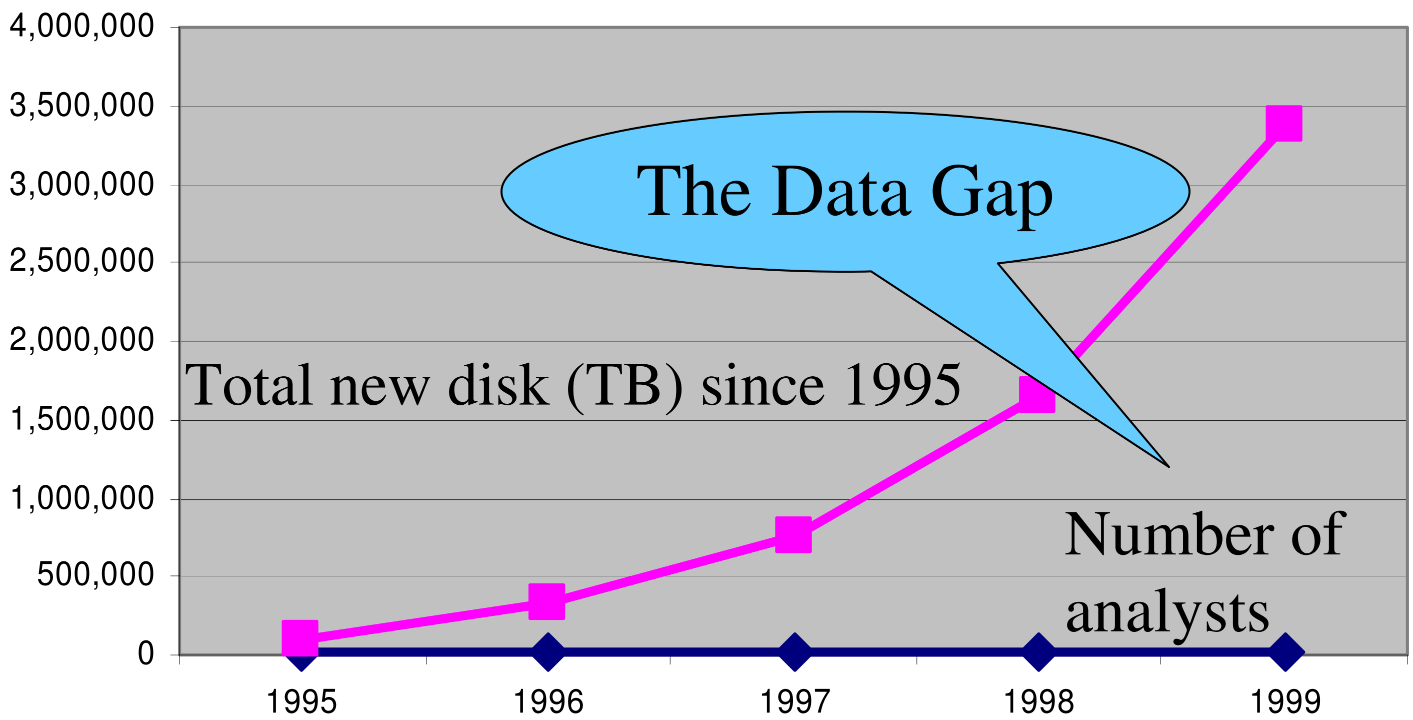
Organization	Scale of Data
Walmart	~ 20 million transactions/day
Google	> 4.2 billion Web pages
Yahoo	~10 GB Web data/hr
NASA satellites	~ 1.2 TB/day
NCBI GenBank	~ 22 million genetic sequences
France Telecom	29.2 TB
UK Land Registry	18.3 TB
AT&T Corp	26.2 TB

"The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don't have a clue as to what any of that data actually means"

(S. Cass, IEEE Spectrum, Jan 2004)

Why Mine Data?

- 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"

Why Mine Data?

"More often, data mining yields unexpected nuggets of information that open the company's eyes to new markets, new ways of reaching customers and new ways of doing business."

[M.Betts, ComputerWorld, April 2003]

"The concept of data mining is one of those things that applies across the spectrum, from business looking at financial data to scientists looking at scientific data…Homeland Security Department will mine data for information from biological sensors, for example… Once we do get a dense enough sensor network out there, we are going to be inundated with data and a lot of the data mining techniques that have been used in industry ... particularly the financial one, will be applied to those data sets." [D.Bolka, Director of HSARPA, 2004]

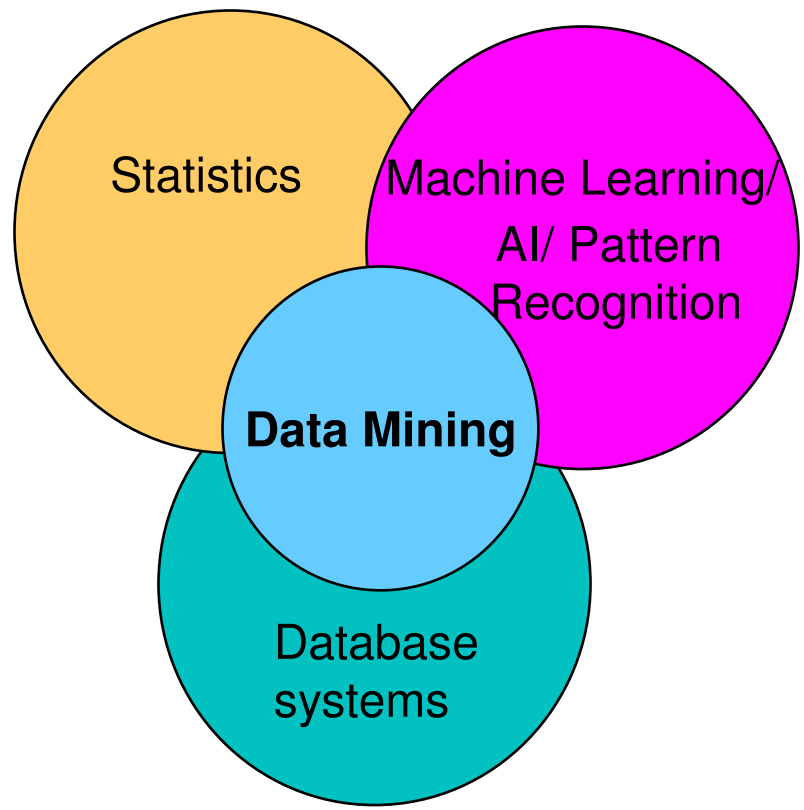
Data Mining Applications

Application	Input	Output
Business Intelligence	Customer purchase history, credit card information	What products are frequently bought together by customers
Collaborative Filtering	User-provided ratings for movies, or other products	Recommended movies or other products
Network Intrusion Detection	TCPdump trace or Cisco NetFlow logs	Anomaly score assigned to each network connection
Web search	Query provided by user	Documents ranked based on their relevance to user input
Medical Diagnosis	Patient history, physiological, and demographic data	Diagnosis of patient as sick or healthy
Climate Research	Measurements from sensors aboard NASA Earth observing satellites	Relationships among Earth Science events, trends in time series, etc
Process Mining	Event-based data from workflow logs	Discrepancies between prescribed models and actual process executions

Origins of Data Mining

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

- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data Mining Tasks

- Predictive Methods
  - Use some variables to predict unknown or values of other variables.

- Descriptive Methods
  - Find human-interpretable patterns that describe the data.

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

Data Mining Tasks...

- Predictive Modeling [Predictive]
  - Classification
  - Regression
- Clustering [Descriptive]
- Pattern Discovery [Descriptive]
  - Association Rule Mining
  - Sequential Pattern Discovery
  - Tree/Subgraph Mining
- Anomaly Detection [Predictive]

Classification: Definition

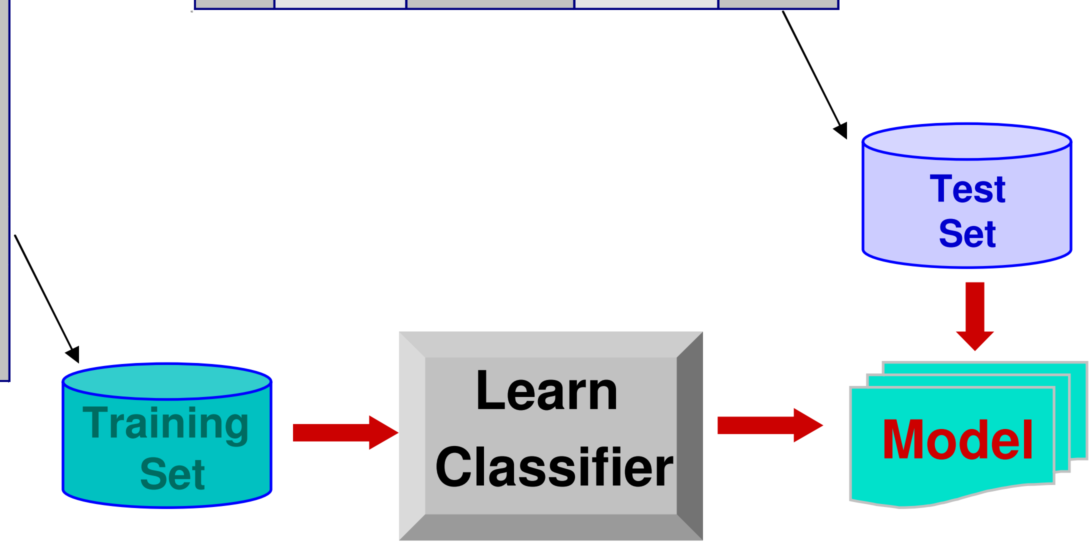
- Given:
  - A collection of records (training set)
    - Each record contains a set of attributes
    - One of the nominal attributes is designated as the class attribute
- Task:
  - Find a model for the class attribute as a function of other attributes
  - Use the model to predict the class for previously unseen records
- Goal:
  - Model should accurately predict the class for previously unseen records
    - 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.

Illustrating Classification

categorical | categorical | continuous | class

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?



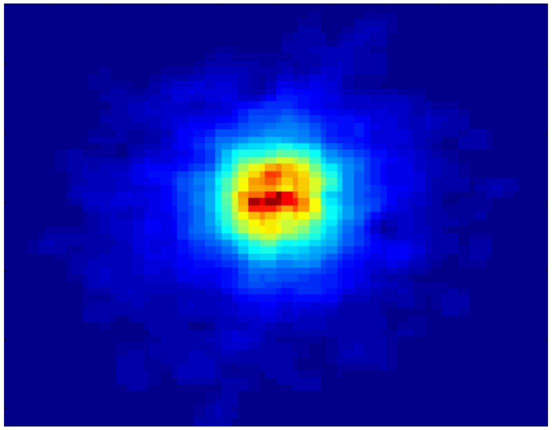
Classification: Applications

- Direct marketing
  - Predict consumers who will most likely buy a new product based on their demographic, lifestyle, and previous buying behavior
- Spam detection
  - Categorize email messages as spam or non-spam based on message header and content
- Functional classification of proteins
  - Assign sequences of unknown proteins to their respective functional classes
- Galaxy classification
  - Classify galaxies based on their image features
- Automated target recognition
  - Identify target objects (enemy tanks, trucks, etc) based on signals gathered from sensor arrays

Classifying Galaxies

Courtesy: http://aps.umn.edu

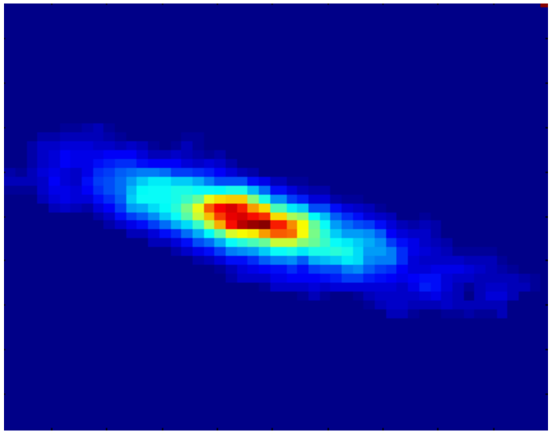
Early



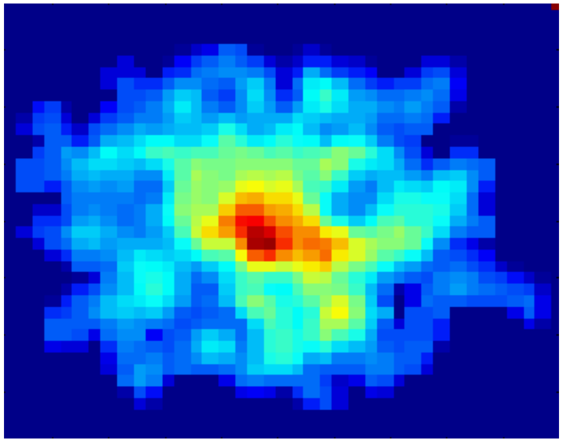
Class:
- Stages of Formation

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

Regression: Definition

- Given:
  - A collection of records (training set)
    - Each record contains a set of attributes
    - One of the continuous-valued attributes is designated as the target variable

- Task:
  - Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency

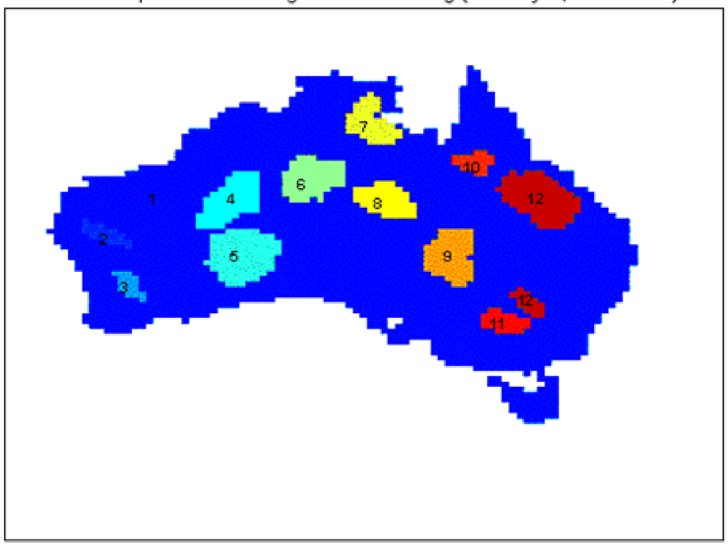
- Greatly studied in statistics, neural network fields

Regression: Applications

- Marketing
  - Predicting sales amounts of new product based on advertising expenditure
- Earth Science
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc
- Finance
  - Time series prediction of stock market indices
- Agriculture
  - Predicting crop yield based on soil fertility and weather information
- Socio-economy
  - Predicting electricity consumption in single family homes based on outdoor temperatures

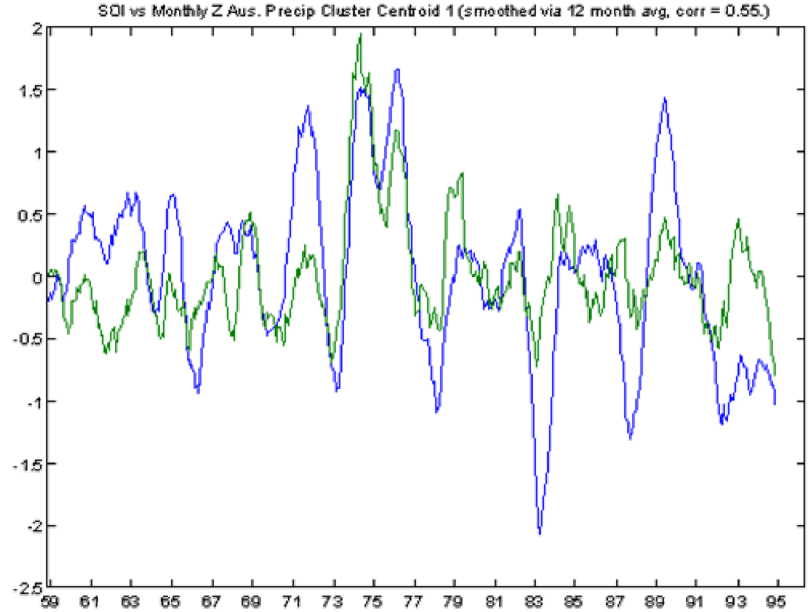
Regression Analysis for Climate Data

13 Precip Clusters Using SNN Clustering (monthly Z, NN = 100)



SOI : a climate index related to the El-Nino phenomenon

SOI vs Monthly Z Aus. Precip Cluster Centroid 1 (smoothed via 12 month avg, corr = 0.55.)



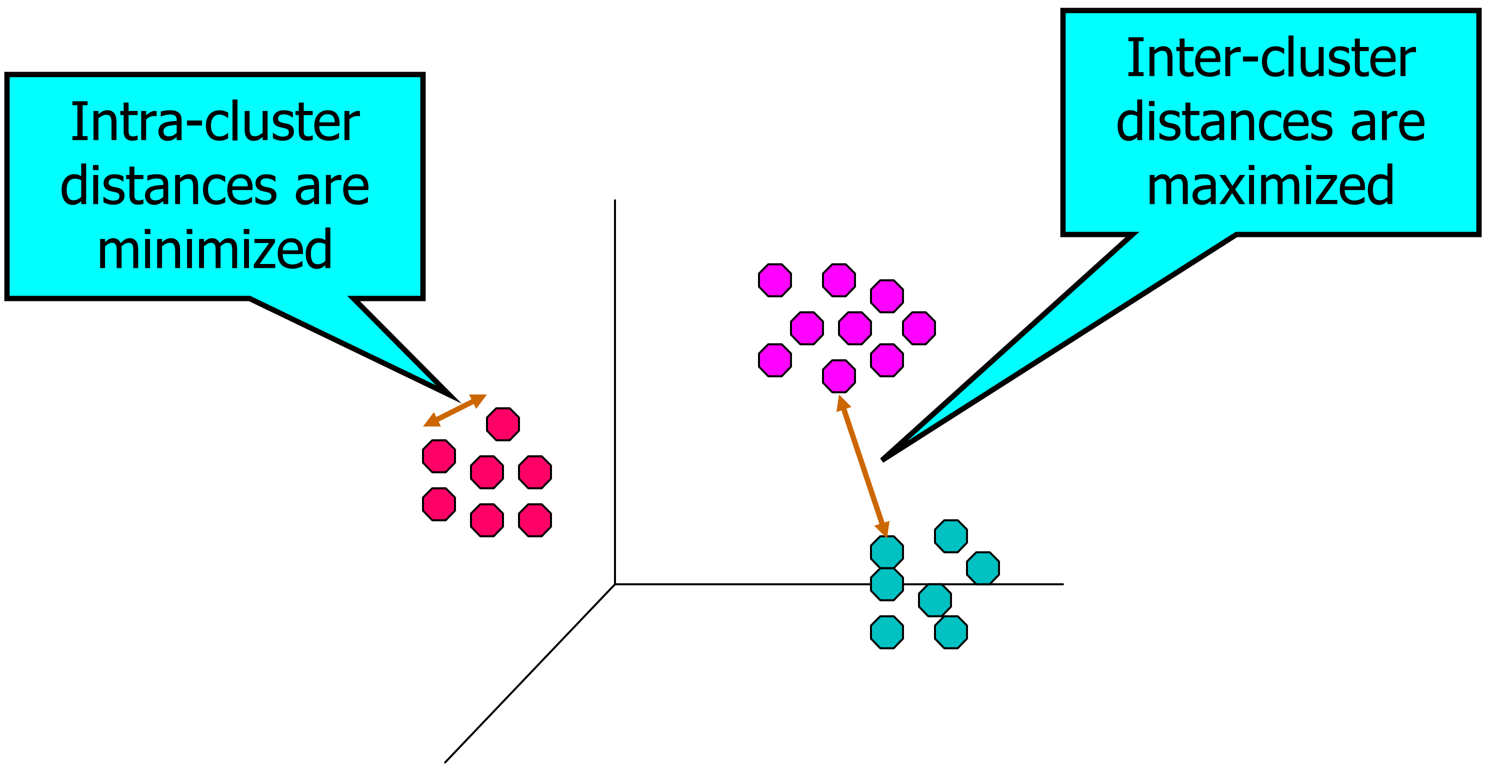
Using SOI to predict precipitation in Australia

Clustering: Definition

- Given:
  - A set of data points
  - Each data point has a set of attributes
  - A distance/similarity measure between data points
    - E.g., Euclidean distance, cosine similarity, and edit distance
- Task
  - Partition the data points into separate groups (clusters)
- Goal:
  - Data points that belong to the same cluster are very similar to one another
  - Data points that belong to different clusters are less similar to one another

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.



Clustering: Applications

- Market Segmentation
  - Subdivide customers based on their geographical and lifestyle related information
- Document clustering
  - Find groups of documents that are similar to each other based on the important terms appearing in them
- Time series clustering
  - Find groups of similar time series (e.g., stock prices, ECG, seismic waves) based on their shapes
- Sequence clustering
  - Find groups of sequences (e.g., Web or protein sequences) with similar features

Association Rule Mining: Definition

- Given:
  - A collection of trainsactions
  - Each transaction contains a set of items
- Task:
  - Discover dependency rules that will predict the presence of an item in a record based on the presence of other items
- Goal:
  - Rules must have high support, i.e., applicable to sufficiently large number of records
  - Rules must have high confidence, i.e., make accurate prediction

Illustrating Association Rule Mining

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 Mining: 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

- World-Wide Web
  - Rules are used to develop Web caching and prefetching techniques

Association Rules in Election Survey Data

Data from 2000 American National Election Studies (NEC) conducted by Center of Political Studies at U of Michigan

Source: M. MacDougall, In Proc of SUGI, 2003

CONF	SUP-PORT	LIFT	_RHAND
40.30	10.42	1.87	WHITE & USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS
32.34	8.36	1.87	USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS & COLLEGE
22.14	5.72	1.83	USE SURPLUS FOR TAX CUTS & MIDDLE CLASS & FOR SCHOOL VOUCHERS
15.92	4.12	1.79	SOUTH & FOR SCHOOL VOUCHERS & COLLEGE
15.67	4.05	1.79	WHITE & UPPER MID CLASS & FOR DEATH PENALTY
16.92	4.37	1.78	WHITE & RURAL & FOR SCHOOL VOUCHERS
22.39	5.79	1.78	USE SURPLUS FOR TAX CUTS & MALE & FOR SCHOOL VOUCHERS
36.32	9.39	1.78	USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS & FOR DEATH PENALTY
16.42	4.24	1.77	USE SURPLUS FOR TAX CUTS & FOR SCHOOL VOUCHERS & AGE 45 - 64
18.91	4.89	1.76	WHITE & SOUTH & FOR SCHOOL VOUCHERS

The highest-lift rules for Republicans tended to repeat the same items: tax cuts, school vouchers, death penalty, college-educated, middle to upper-middle class, as shown in Figure 4.

CONF	SUP-PORT	LIFT	_RHAND
19.96	6.75	2.13	USE SURPLUS FOR SOC. SEC. & BLACK
21.86	7.40	2.10	BLACK
15.40	5.21	2.08	USE SURPLUS FOR TAX CUTS & BLACK
15.02	5.08	2.05	RACE BIASED & BLACK
16.92	5.72	1.74	FOR GUN CONTROL & FOR ABORTION RIGHTS & AGE 45-64
15.59	5.27	1.65	URBAN & FOR GUN CONTROL & FOR ABORTION RIGHTS
17.49	5.92	1.62	URBAN & FOR GUN CONTROL & FOR ENVIR. PROTECTION
19.20	6.50	1.58	URBAN & NO CHILDREN AT HOME & FOR GUN CONTROL
17.87	6.05	1.58	USE SURPLUS FOR SOC. SEC. & FOR ABORTION RIGHTS & AGE 45 - 64
21.10	7.14	1.58	NO CHILDREN AT HOME & HS OR LESS & FOR GUN CONTROL

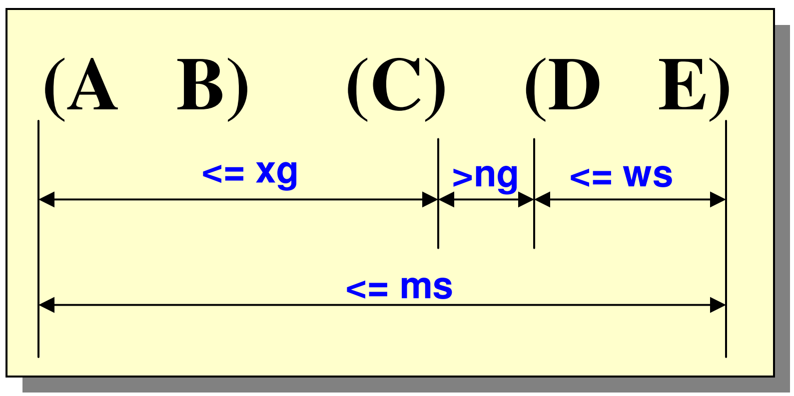
Top 10 rules for democrats, by lift, involved black, age 45-64, and urban, and support of gun control, abortion rights, environmental protection and gays in the military, as listed in Figure 5

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.

$$(A\quad B)\qquad (C) \longrightarrow (D\quad E)$$

- 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)

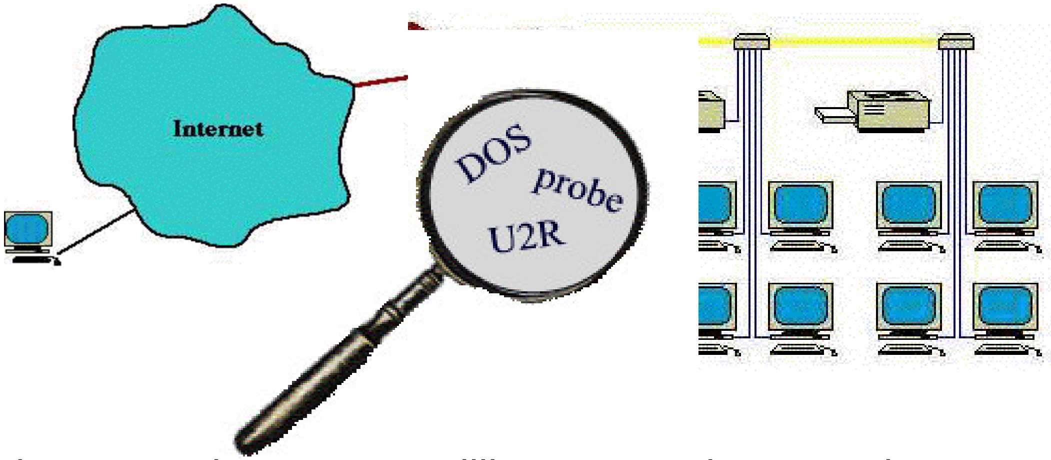
- In point-of-sale transaction sequences,
  - Computer Bookstore:

    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:

    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

Deviation/Anomaly 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