Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining

by

Tan, Steinbach, Kumar

Lecturers and Tutors:

- Lecturer: Markus Hagenbuchner
  - http://www.hagenbuchner.com/
- Lecturer: Jia Zeng
  - http://www.comp.hkbu.edu.hk/~jiazeng/
- Tutor: Benyun Shi
  - byshi@comp.hkbu.edu.hk

Course Website and Office Hours

- Course Website
  - http://www.comp.hkbu.edu.hk/~markus/teaching/comp7650/
- Office Hours
  - Thursday 9:00AM – 11:00AM

Text Book & Reference Books

- Text book:
  - P. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006.
- Reference books:
  - J. W. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2001.
  - D. Hand, H. Mannila and P. Smyth, Principles of Data Mining, MIT Press, 2001.
  - Alex Berson, Stephen J. Smith, Data Warehousing, Data Mining, & OLAP, McGraw Hill, 2001.
  - Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, San Francisco, CA, 2000.

Time Table

- Sep 3rd – Nov 27th
- 13 weeks of lectures. Some lecture include lab sessions to support your studies.
- The first four lectures (Sep 3rd – Sep 25th) are taught by Dr. Jia Zeng.
- The rest lectures and lab sessions (Sep 30th – Nov 27th) will be delivered by Dr. Markus Hagenbuchner.

Course Outline

Week	Content
1	Introduction
2	Data (types, quality, preprocessing and similarity measures)
3	Classification (Decision Trees and Basic concepts)
4	Classification (Gaussian mixture models and hidden Markov models)
5	Classification (Artificial-Neural Networks)
6	Classification (other methods, and overview)
7	Clustering (introduction, K-means clustering)

Course Outline

Week	Content
8	Clustering (Self-Organizing Maps)
9	Clustering (graph based clustering)
10	Data Exploration (Projection, visualization, statistics, etc.)
11	Anomaly detection
12	Association Rules
13	Association Rules, Summary, Revision,

Course Assessment

- Continuous Assessment (40%)
  - Three assignments (18%)
  - Group project (22%)
- Examination (60%)

Learning Outcomes

- **Knowledge**
  - Identify and distinguish data mining applications from other IT applications.
  - Describe data mining algorithmsDescribe applicability of data mining applications.
- **Professional Skill**
  - Suggest appropriate solutions to data mining problems.
  - Analyze data mining algorithms and techniques.
- **Attitude**
  - Build up team spirit in solving challenging data mining problems.

Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

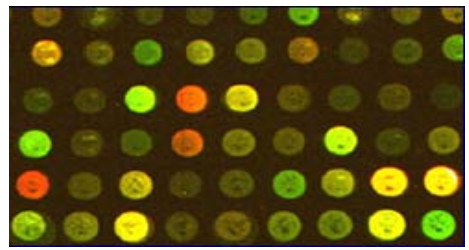
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

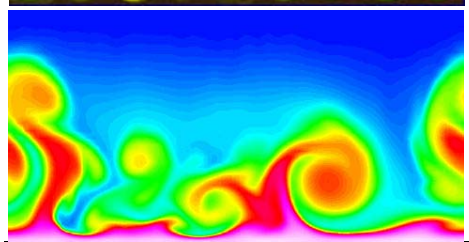
Examples

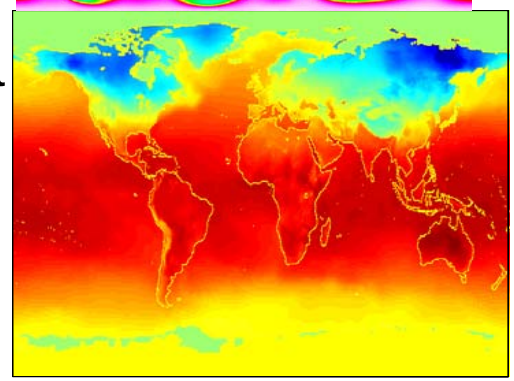
- Who are the best data mining companies?
- Web data and E-commerce:
  - Facebook (http://www.facebook.com/)
  - Twitter (http://twitter.com/)
  - Wikipedia (http://www.wikipedia.org/)
  - Blogs (http://blogsearch.google.com/)
  - Alibaba (http://www.alibaba.com/)
  - Taobao (http://www.taobao.com/)
- Stock market (algorithmic trading: http://en.wikipedia.org/wiki/Algorithmic_trading)
- How cheap are computers? (http://www.hkgolden.com/)

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation







Examples:

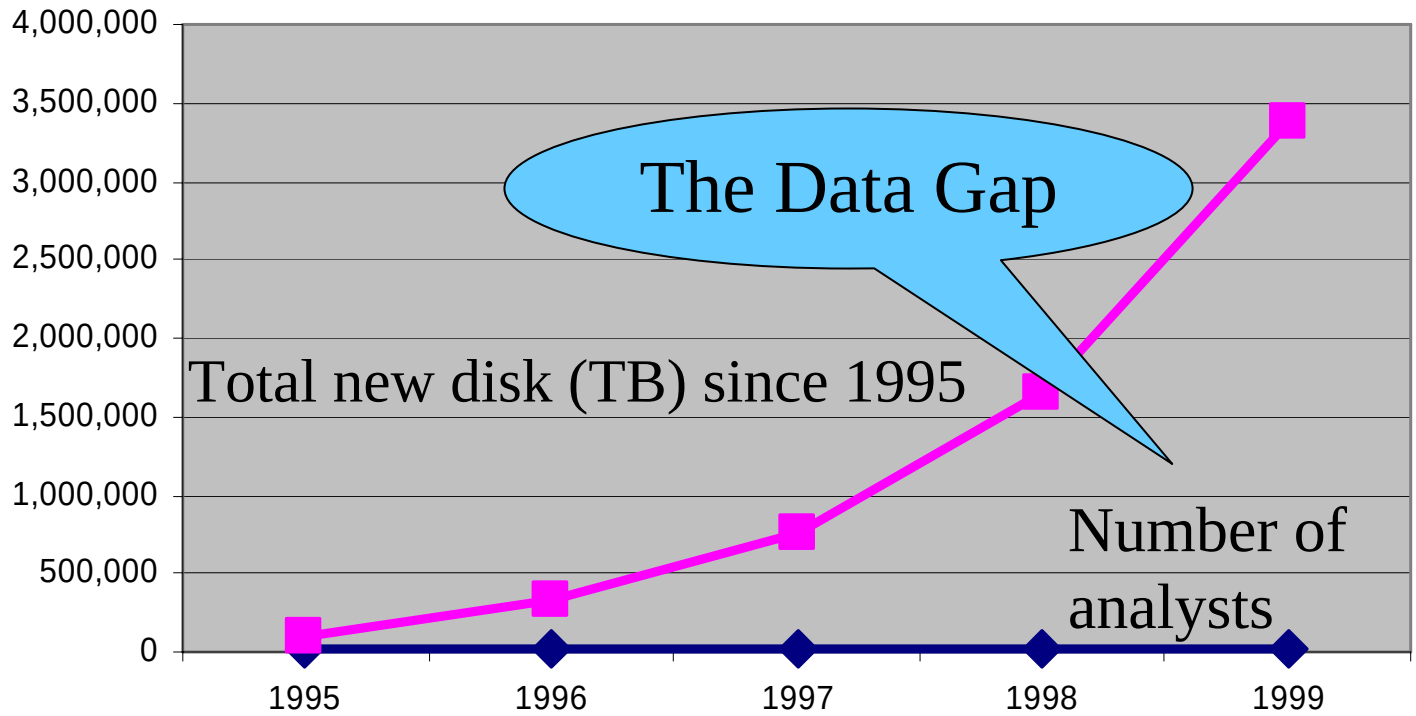
- Data collection (Gigabyte GB)
  - Google map (http://maps.google.com.hk/)
  - Google sky (http://www.google.com/sky/)
  - National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/)
  - National Aeronautics and Space Administration (NASA) (http://www.nasa.gov/)
- 1024byte = 1K; 1024K = 1M; 1024M = 1G; 1024G = 1T

Knowledge Discovery (Case Study)

- Tycho Brahe (第谷): http://www.nada.kth.se/~fred/tycho/index.html
  - He spent around 20 years observing heavens.
  - But how to figure out the astronomical principles from these observed data?
- Johannes Kepler (开普勒): http://en.wikipedia.org/wiki/Johannes_Kepler
  - Data mining and complex inference based on geometric knowledge.
  - Three astronomical principles → Newton's law of universal gravitation (knowledge discovery).

Mining Large Data Sets - Motivation

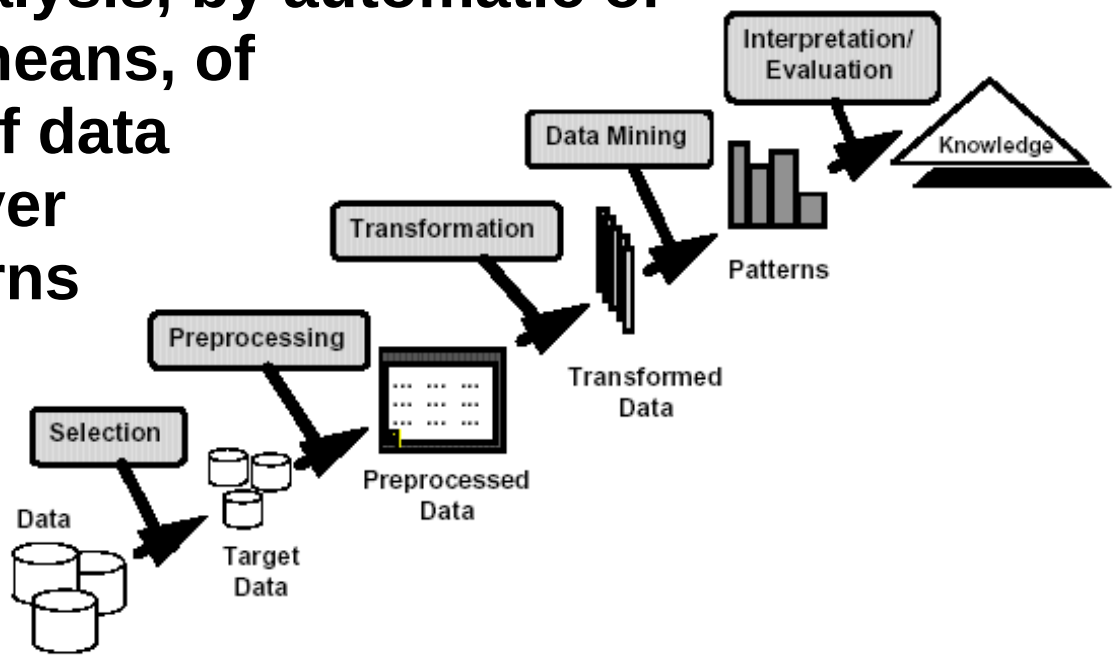
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”

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**



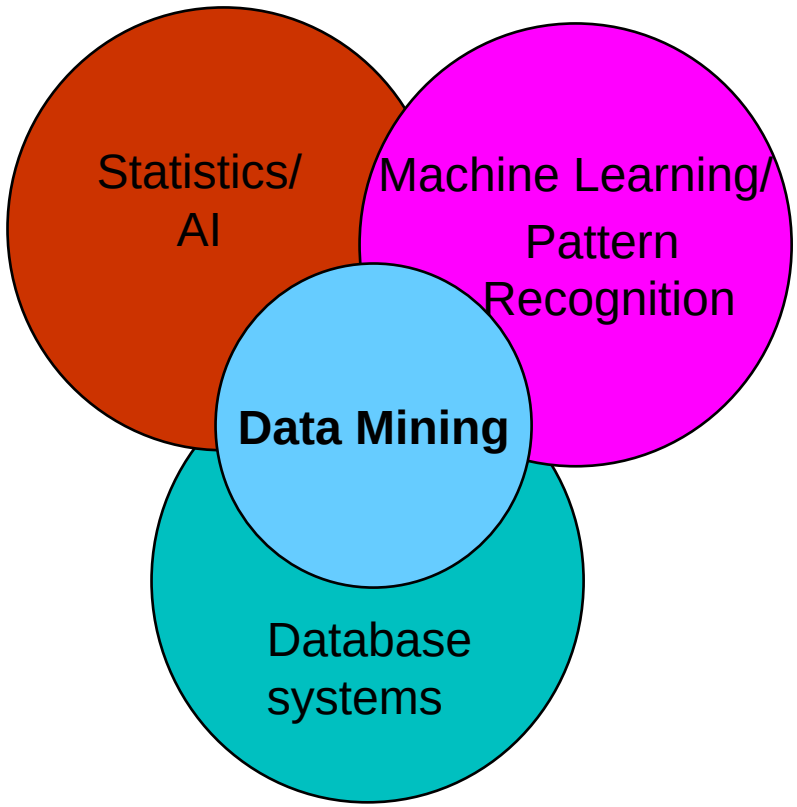
What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon" (Information Retrieval)

- **What is Data Mining?**
  - Recommend a book to the potential users based on their relatives' and friends' preference. (Collaborative filtering)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon.com) (Document clustering)

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



As a summary, what is data mining?

- Algorithms
  - They are soft strategies for processing data.
- Large volume of data
  - Data mining does not focus on small datasets.
- Knowledge discovery
  - Through data mining, we expect to find new knowledge to support or denial our hypothesis.
  - We expect to find “hidden” relations under these data.

Current Data Mining

- Conferences:
  - KDD: http://www.sigkdd.org/kdd2009/
  - ICDM: http://www.cs.umbc.edu/ICDM09/
  - SDM: http://www.siam.org/meetings/sdm10/
- Journals:
  - IEEE Transactions on Knowledge and Data Engineering (TKDE): http://www2.computer.org/portal/web/tkde/about;jsessionid=E19962782EA80EA8661DFFB342EE98E4
  - Data mining and knowledge discovery: http://www.springerlink.com/content/100254/

Data Mining Tasks

- Classification
- Clustering
- Sequential Pattern Discovery
- Regression
- Deviation Detection

Classification: Definition

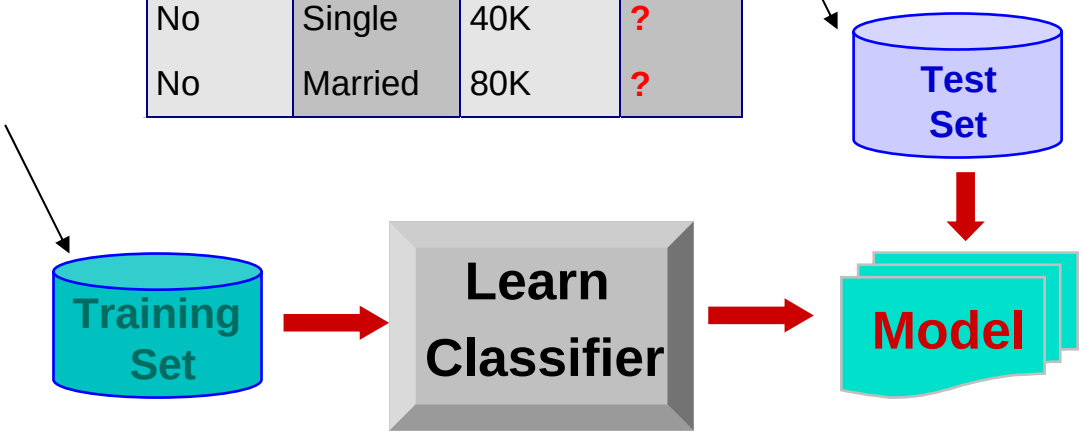
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - 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

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

categorical categorical continuous class

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



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

- 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 3

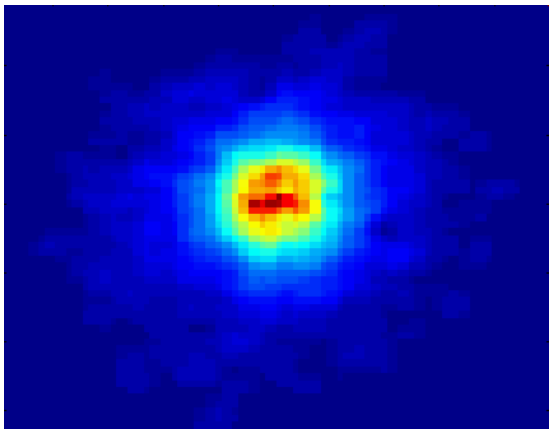
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes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

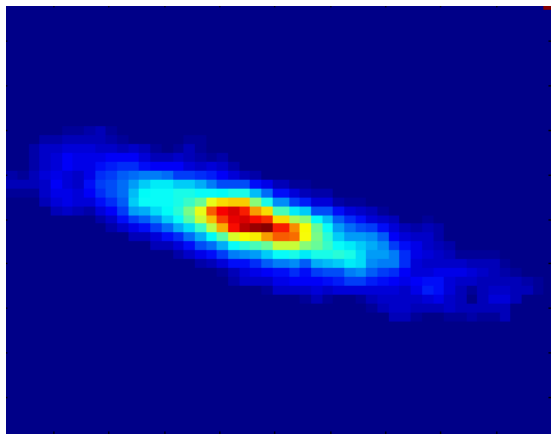
Early



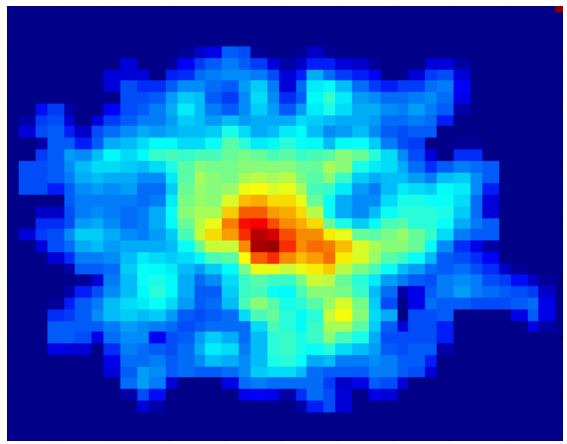
Class:
- **Stages of Formation**

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**

Clustering Definition

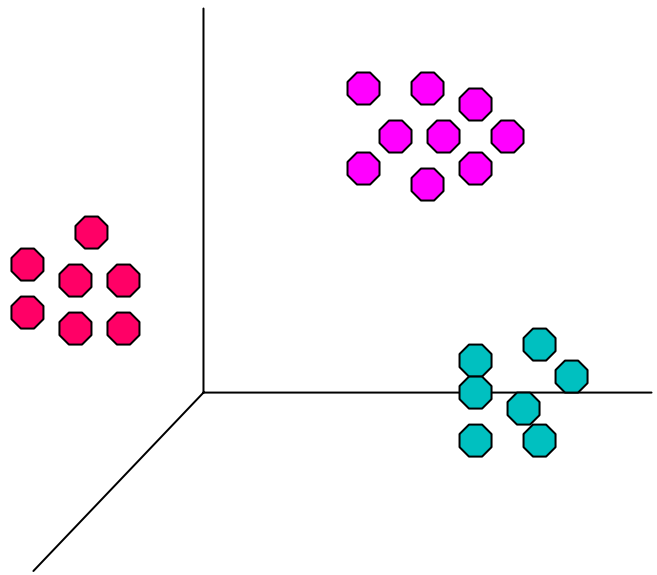
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



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.
  - 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

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 Rule Discovery: Application 1

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

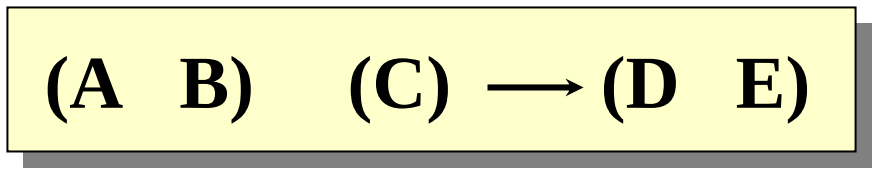
    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips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

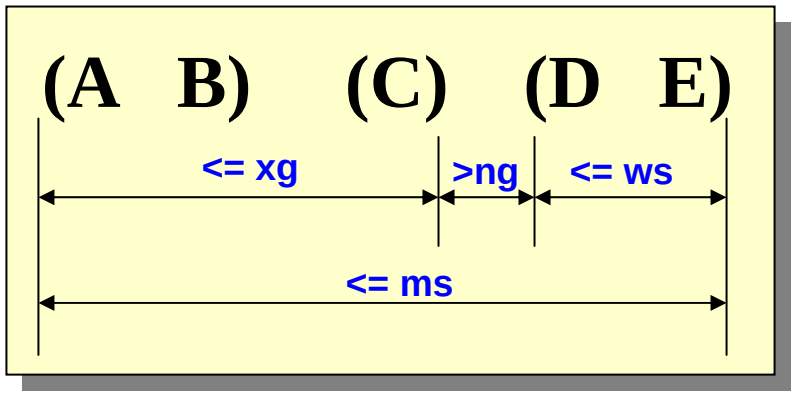
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

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.



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



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.
- Greatly 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 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

Challenges of Data Mining

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