Introduction to Data Mining

Dr. Hui Xiong
Rutgers University

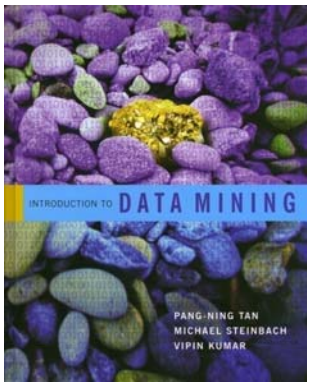
THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Questions ?

- Instructor: Dr. Hui Xiong
- Office Hours: Ackerson 200K
  Wednesday 11:00AM-12:00pm
- Office Phone: 973-353-5261

Email: hxiong@rutgers.edu
WEB: http://datamining.rutgers.edu

Required Textbook



Pang-Ning Tan, Michael Steinbach, Vipin Kumar,
Addison Wesley,
ISBN: 0-321-32136-7, 2005.

Course Objectives

- To teach the fundamental concepts of data mining
- To provide hands-on experience in applying the concepts to real-world applications.

Course Web Site

http://datamining.rutgers.edu/teaching/spring2009/DM/685.html

- This web site is the location for course documents, assignments, announcements and other information. You should check it frequently to remain updated.
- Note that You are responsible for keeping aware of the announcements at the course web site.

Grading Policy

In-class work (including attendance)	10%
Assignments	20%
Projects	20%
Exam I	25%
Exam II	25%

Note that the final letter grade is based on a curve.

Attendance

Regular attendance is compulsory. You are not allowed to check your emails, access Web sites not related to the course or work on something that is beyond the scope of this course during the class time.

Assignments

You may have discussions with your class members, but you have to submit your own work. Please be sure to keep a copy of the assignment by yourself in case that there is any problem with your hand-in or you have to use it later this semester.

Exams

There will be no make-up exams. You are required to present a written proof for situations, such as going to an emergency room due to unexpected and serious illness.

Chatting during the exam is not allowed. No collaboration between class members will be allowed during any exam. There will be no extra-credit project.

Scholastic Dishonesty

The University defines academic dishonesty as cheating, plagiarism, unauthorized collaboration, falsifying academic records, and any act designed to avoid participating honestly in the learning process. Scholastic dishonesty also includes, but not limited to, providing false or misleading information to receive a postponement or an extension on assignments, and submission of essentially the same written assignment for two different courses without the permission of faculty members.

The purpose of assignments is to provide individual feedback as well to get you thinking. Interaction for the purpose of understanding a problem is not considered cheating and will be encouraged. However, the actual solution to problems must be one's own.

Helpful Comments

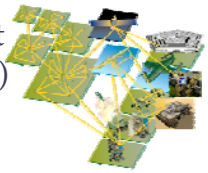
To get full benefit out of the class you have to work regularly. Read the textbook regularly and start working on the assignments soon after they are handed out. Plan to spend at least 15 hours a week on this class doing assignments or reading.

Why Mine Data? Commercial Viewpoint

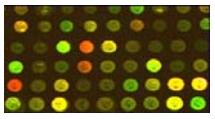
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

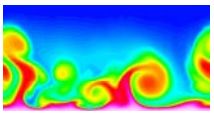
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

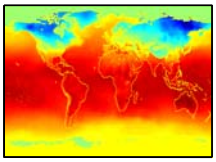


- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data



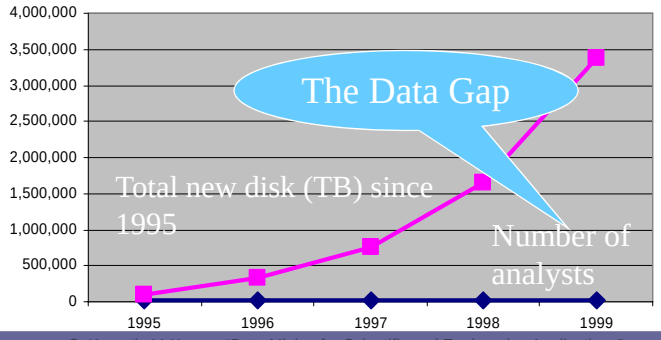


- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



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"

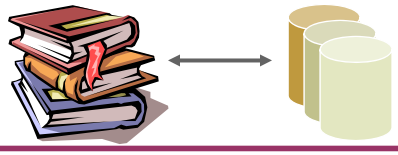
Scale of Data

Organization	Scale of Data
Walmart	~ 20 million transactions/day
Google	~ 8.2 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"

Why Do We Need Data Mining ?

- Leverage organization's data assets
  - Only a small portion (typically - 5%-10%) of the collected data is ever analyzed
  - Data that may never be analyzed continues to be collected, at a great expense, out of fear that something which may prove important in the future is missing.
  - Growth rates of data precludes traditional "manually intensive" approach

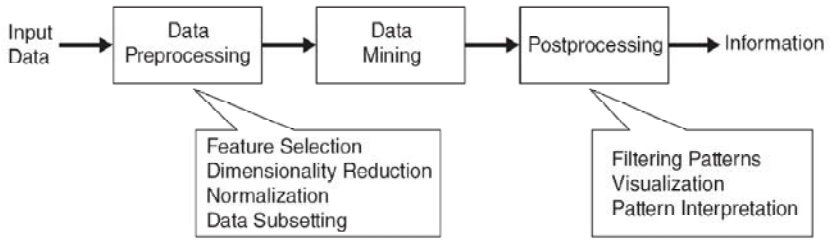


Why Do We Need Data Mining?

- As databases grow, the ability to support the decision support process using traditional query languages becomes infeasible
  - Many queries of interest are difficult to state in a query language (Query formulation problem)
  - "find all cases of fraud"
  - "find all individuals likely to buy a FORD expedition"
  - "find all documents that are similar to this customers problem"

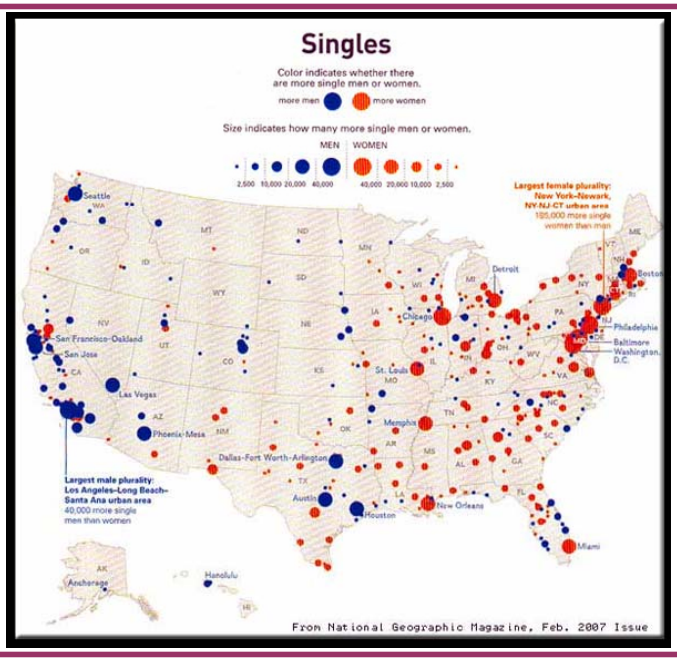
What is Data Mining?

- 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

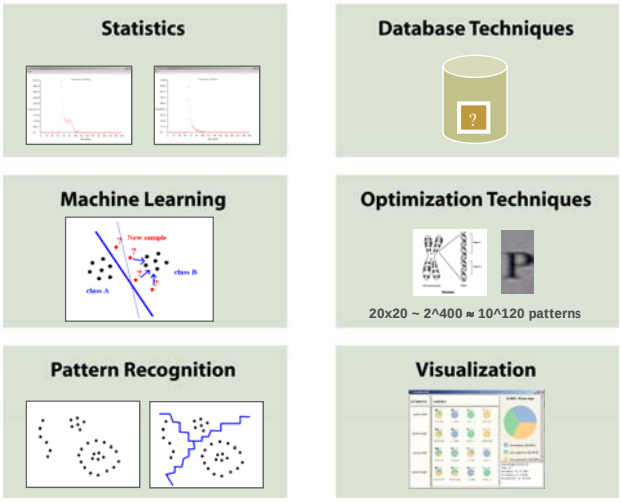


What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Check the dictionary for the meaning of a word
- 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,)



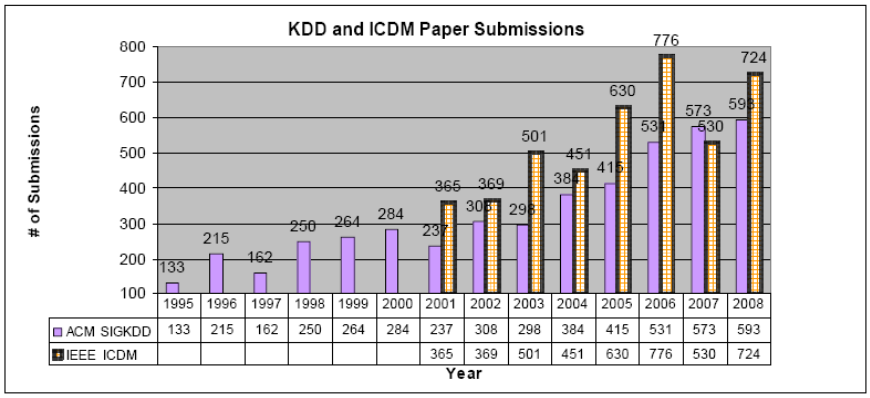
Data Mining: Confluence of Multiple Disciplines



Main Forums in Data Mining

- Conferences:
  - The birth of data mining/KDD: 1989 IJCAI Workshop on Knowledge Discovery in Databases
    - 1991-1994 Workshops on Knowledge Discovery in Databases
  - 1995 – date: International Conferences on Knowledge Discovery and Data Mining (KDD)
  - 2001 – date: IEEE ICDM and SIAM-DM (SDM)
  - Several regional conferences, incl. PAKDD (since 1997) & PKDD (since 1997)
- Journals:
  - Data Mining and Knowledge Discovery (DMKD, since 1997)
  - Knowledge and Information Systems (KAIS, since 1999)
  - IEEE Trans. on Knowledge and Data Engineering (TKDE)
  - Many others, incl. TPAMI, TKDD, ML, MLR, VLDBJ …

ACM KDD vs. IEEE ICDM



TKDE Submission Numbers and Acceptance Rate

Year	New Submission #	(Current) Accpt Rate
2000	195 - Regular	34.40%
2001	294	25.50%
2002	233	24.00%
2003	355	26.40%
2004	347	21.00%
2005	480	30.00%
2006	588	23.00%
2007	625	being accpt'd

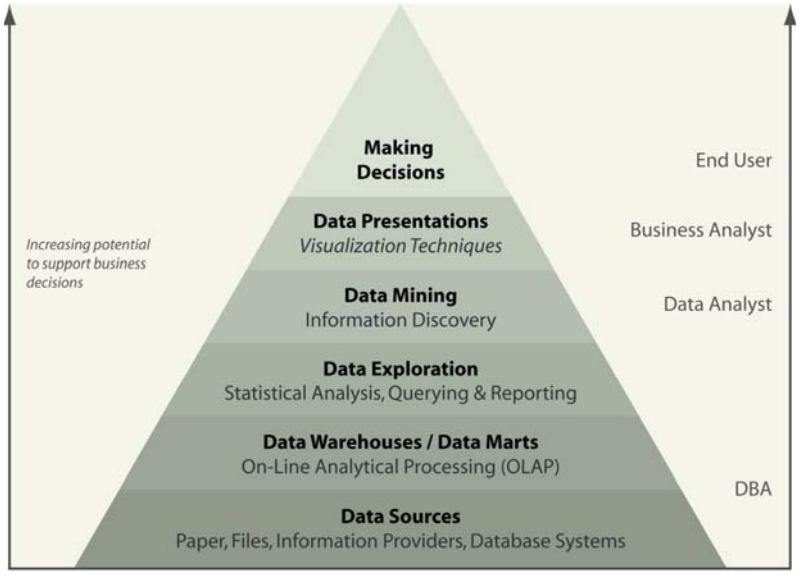
Data Mining Applications

- Market analysis
- Risk analysis and management
- Fraud detection and detection of unusual patterns (outliers)
- Text mining (news group, email, documents) and Web mining
- Stream data mining
- DNA and bio-data analysis

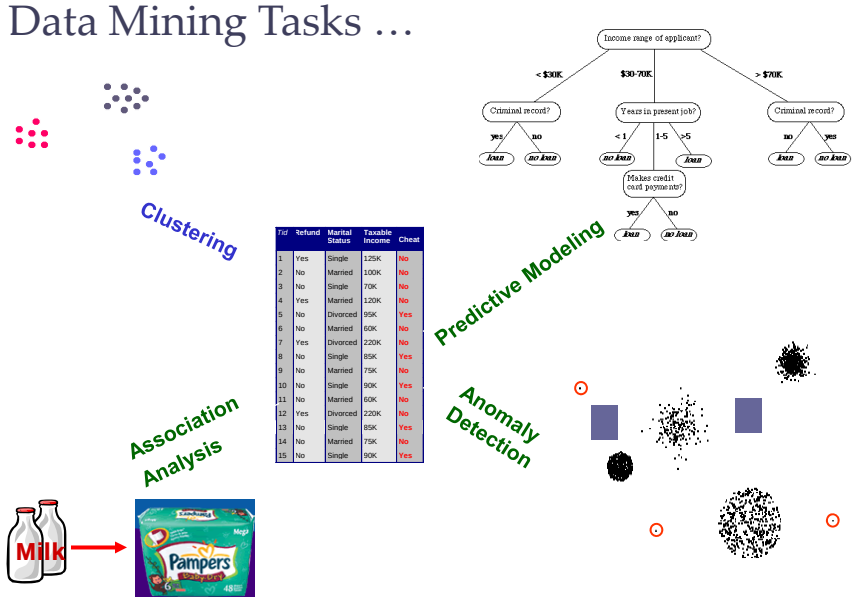
Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, …
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

Data Mining and Business Intelligence



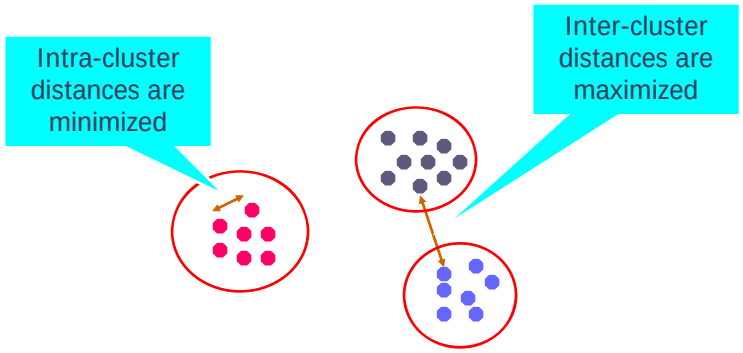
Data Mining Tasks …



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

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

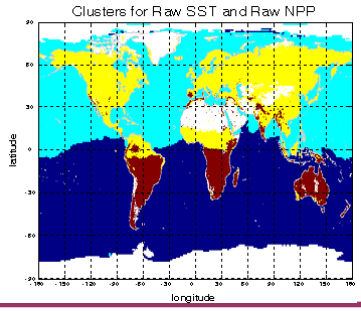


Applications of Cluster Analysis

- Understanding
  - 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

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

	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



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.

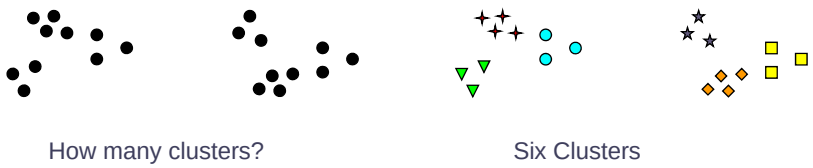
Clustering: Application 2

- 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.

What is not Cluster Analysis?

- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
  - Clustering is a grouping of objects based on the data
- Supervised classification
  - Have class label information
- Association Analysis
  - Local vs. global connections

Notion of a Cluster can be Ambiguous



How many clusters?

Six Clusters

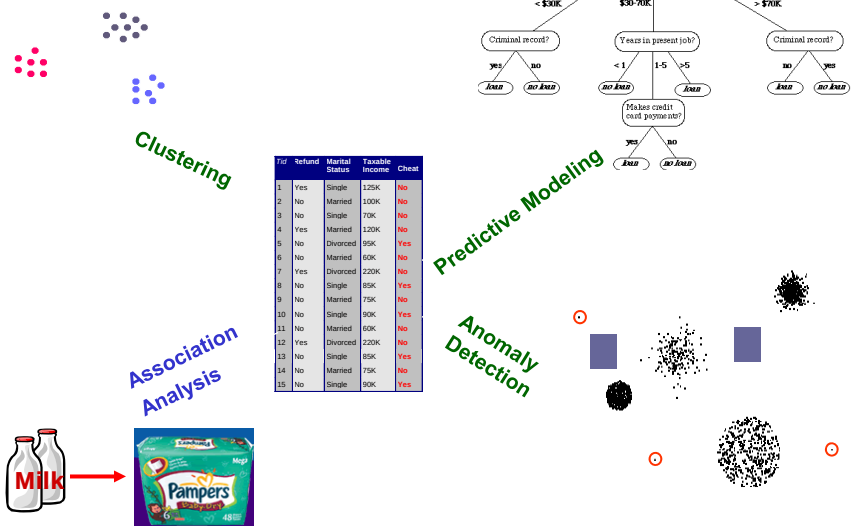
Two Clusters

Four Clusters

Characteristics of the Input Data Are Important

- Type of proximity or density measure
  - This is a derived measure, but central to clustering
- Sparseness
  - Dictates type of similarity
  - Adds to efficiency
- Attribute type
  - Dictates type of similarity
- Type of Data
  - Dictates type of similarity
  - Other characteristics, e.g., autocorrelation
- Dimensionality
- Noise and Outliers
- Type of Distribution

Data Mining Tasks …



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 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 complaints associated with certain diseases

Correlation Computing

- Various Applications of Correlation Analysis
  - i.e. Marketing Data Study, Web Search, Bioinformatics, Public Health
- A Gap between Association Rule Mining and Correlation Computing
  - A lack of precise relationship between support (or confidence) based association measures and correlation measures.
- Statistical Computing
  - Expect to apply statistical techniques more flexibly, efficiently, easily, and with minimal mathematical assumptions.

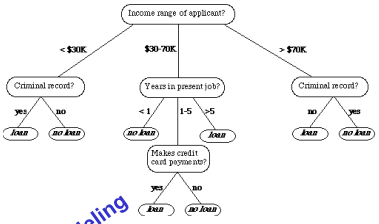
Application Deployment Challenge

- AMAZON.COM: Product Promotion
- Answer the question: Customers who bought this book also bought?

- Computing Challenge!
  - For a database of 10^6 items, 10^{12} possible item pairs
  - Several million transactions will make things worse!

Data Mining Tasks …

Clustering

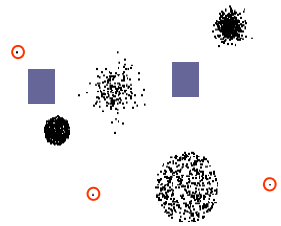


Predictive Modeling

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 Analysis

Anomaly Detection



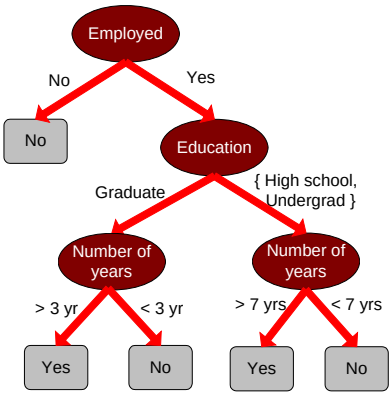
Predictive Modeling: Classification

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

Model for predicting credit worthiness

Class

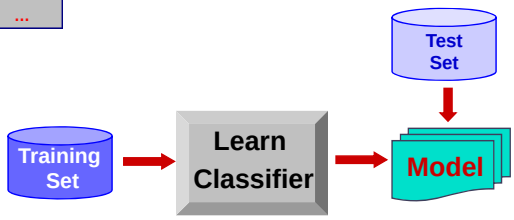
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…



Classification Example

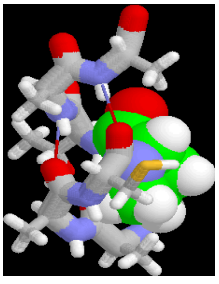
Tid	Employed	Level of Education	# years at present address	Credit Worthy
	categorical	categorical	quantitative	class
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

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



Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace



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.

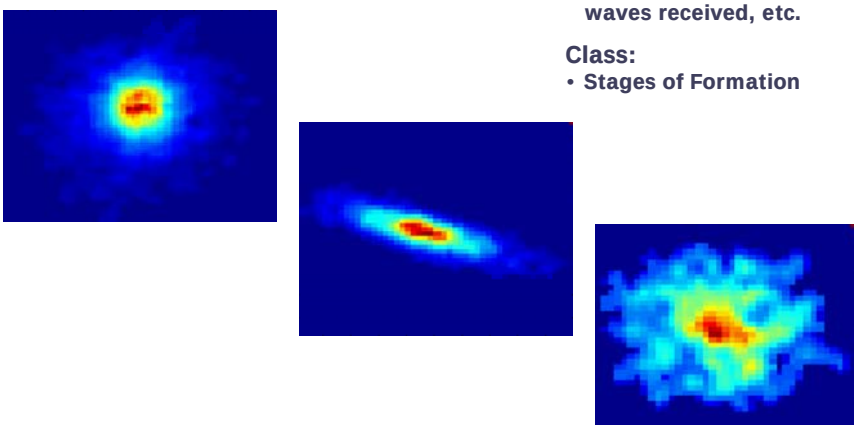
Classification: Application 3

- 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!

Classifying Galaxies

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

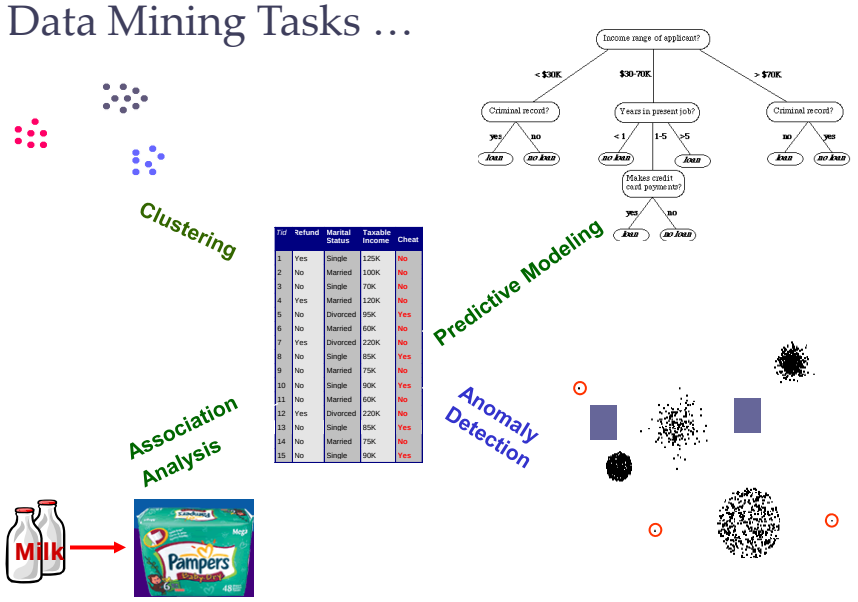
Class:
- **Stages of Formation**



Classification Techniques

- Base Classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines
- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

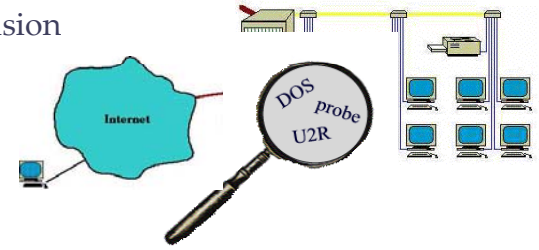
Data Mining Tasks …



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

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection

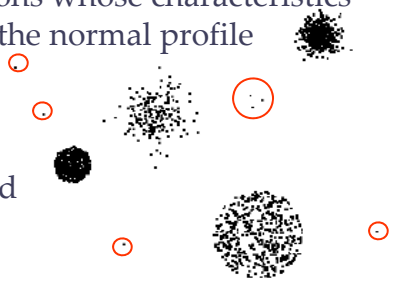


Anomaly Detection

- Challenges
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging (just like for clustering)
  - Finding needle in a haystack
- Working assumption
  - There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

Anomaly Detection Schemes

- General Steps
  - Build a profile of the "normal" behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the "normal" profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile
- Types of anomaly detection schemes
  - Graphical & Statistical-based
  - Distance-based
  - Model-based



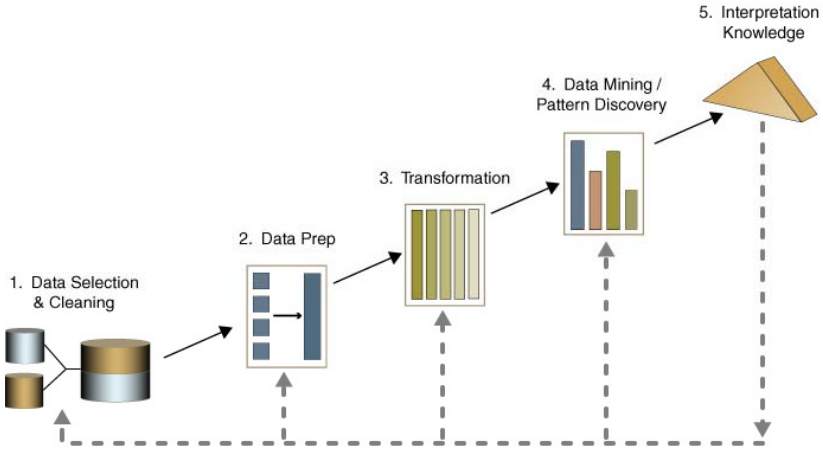
KDD Process

- Develop an understanding of the application domain
  - Relevant prior knowledge, problem objectives, success criteria, current solution, inventory resources, constraints, terminology, cost and benefits
- Create target data set
  - Collect initial data, describe, focus on a subset of variables, verify data quality
- Data cleaning and preprocessing
  - Remove noise, outliers, missing fields, time sequence information, known trends, integrate data
- Data Reduction and projection
  - Feature subset selection, feature construction, discretizations, aggregations

KDD Process

- Selection of data mining task
  - Classification, segmentation, deviation detection, link analysis
- Select data mining approach
- Data mining to extract patterns or models
- Interpretation and evaluation of patterns/models
- Consolidating discovered knowledge

Knowledge Discovery



An Overview of the Steps That Compose the KDD Process

Challenges of Data Mining

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

Commercial and Research Tools

WEKA:
http://www.cs.waikato.ac.nz/ml/weka/

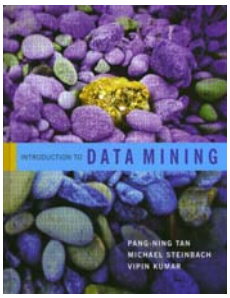
SAS:
http://www.sas.com/

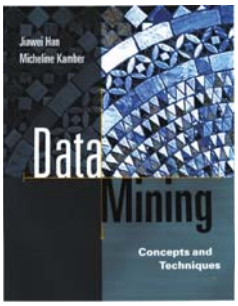
Clementine:
http://www.spss.com/spssbi/clementine/

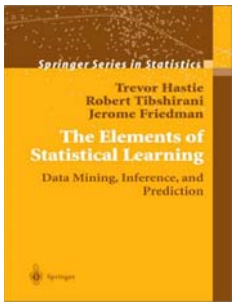
Intelligent Miner
http://www-3.ibm.com/software/data/iminer/

Insightful Miner
http://www.insightful.com/products/product.asp?PID=26

Textbooks



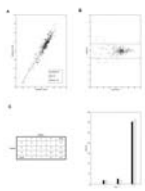




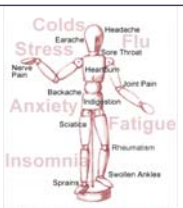
Hans Rosling:
No more boring data:
TEDTalks

http://www.youtube.com/watch?v=hVimVzgtD6w

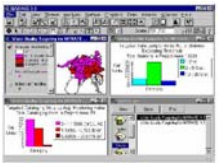
Similarities Between Data Miners and Doctors



Data Characteristics



Your Symptoms?



Data Mining Techniques

Medical Devices

Thank You!

http://datamining.rutgers.edu