Introduction to Data Mining

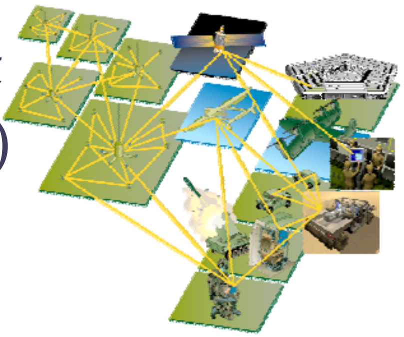
Dr. Hui Xiong
Rutgers University

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

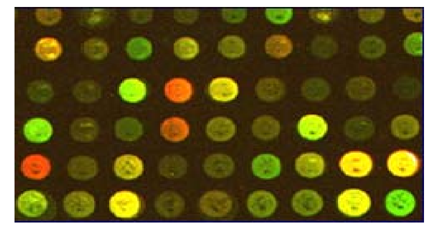
Why Mine Data? Commercial Viewpoint

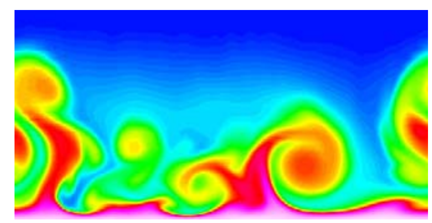
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

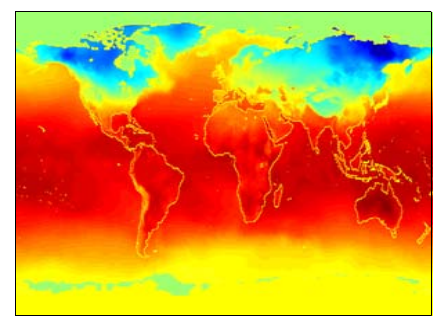


- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data



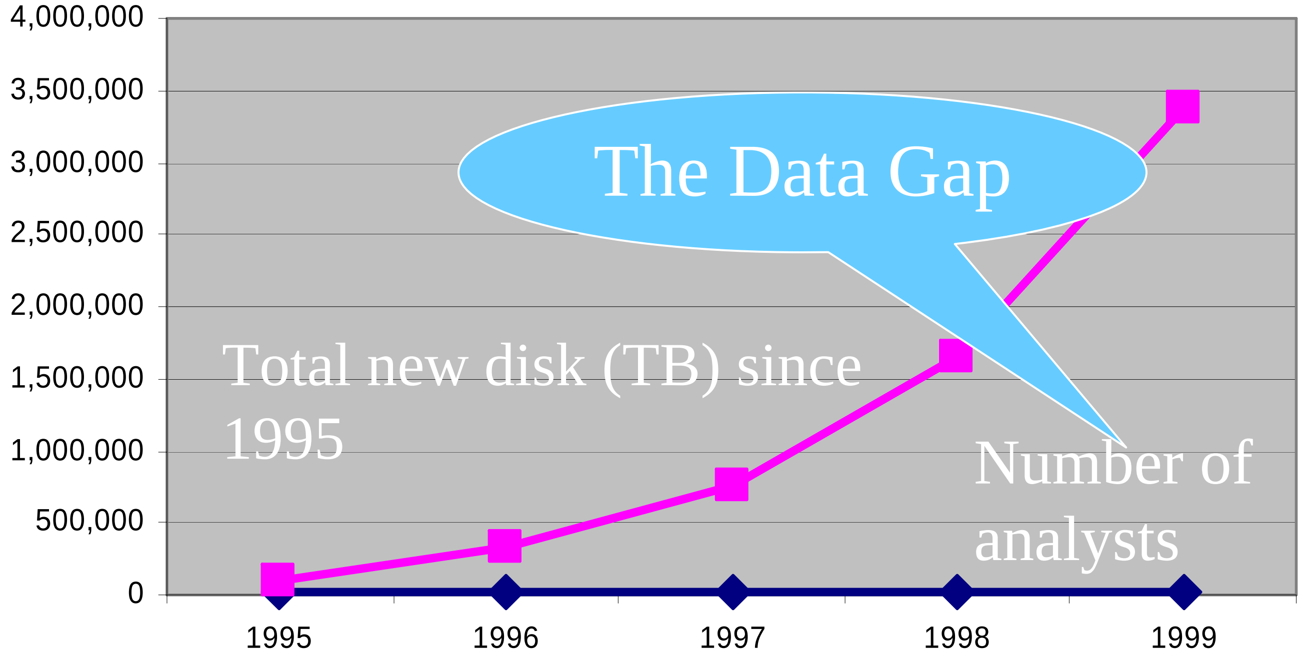


- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



Mining Large Data Sets - Motivation

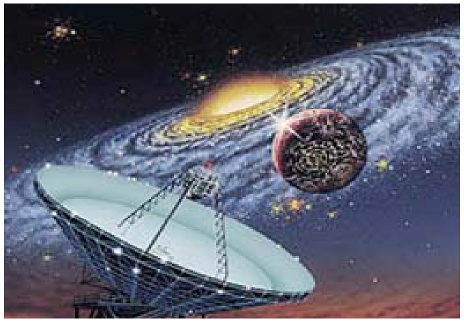
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



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

Scale of Data

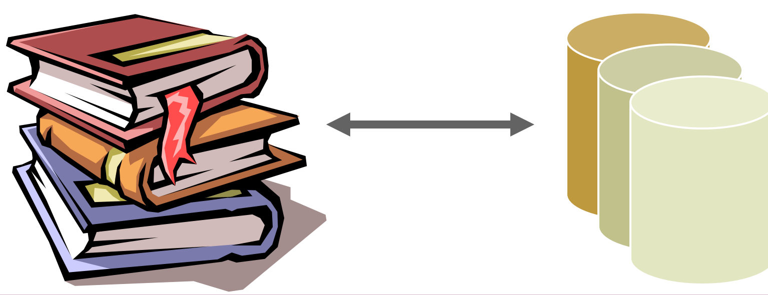
Organization	Scale of Data
Walmart	~ 20 million transactions/day
Google	~ 8.2 billion Web pages
Yahoo	~10 GB Web data/hr
NASA satellites	~ 1.2 TB/day
NCBI GenBank	~ 22 million genetic sequences
France Telecom	29.2 TB
UK Land Registry	18.3 TB
AT&T Corp	26.2 TB



"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

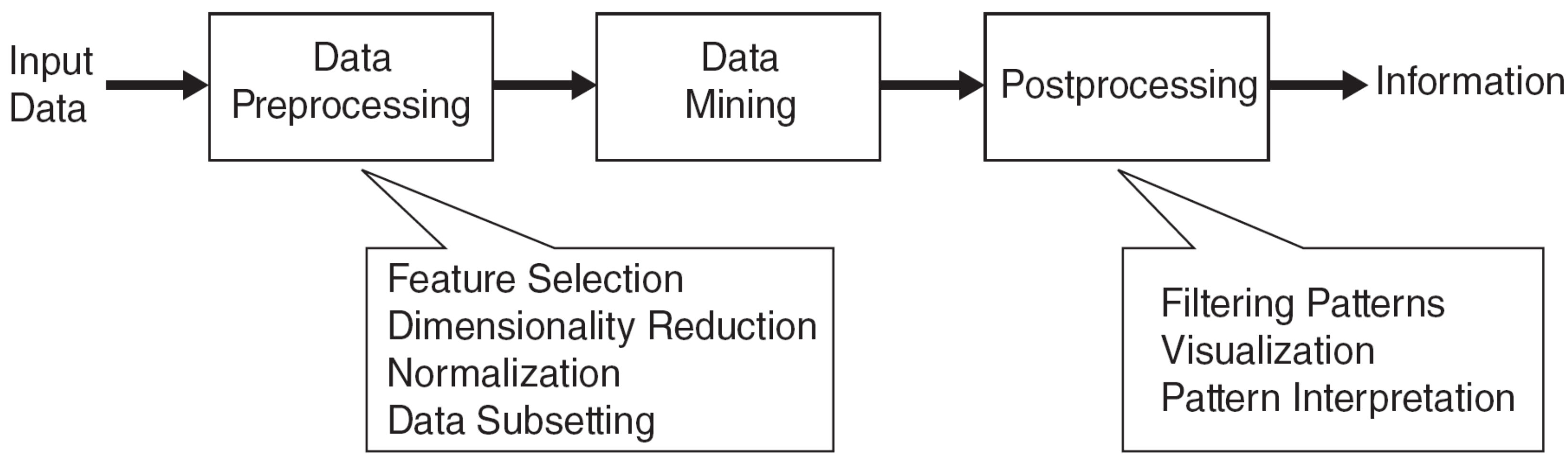


Why Do We Need Data Mining?

- 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?

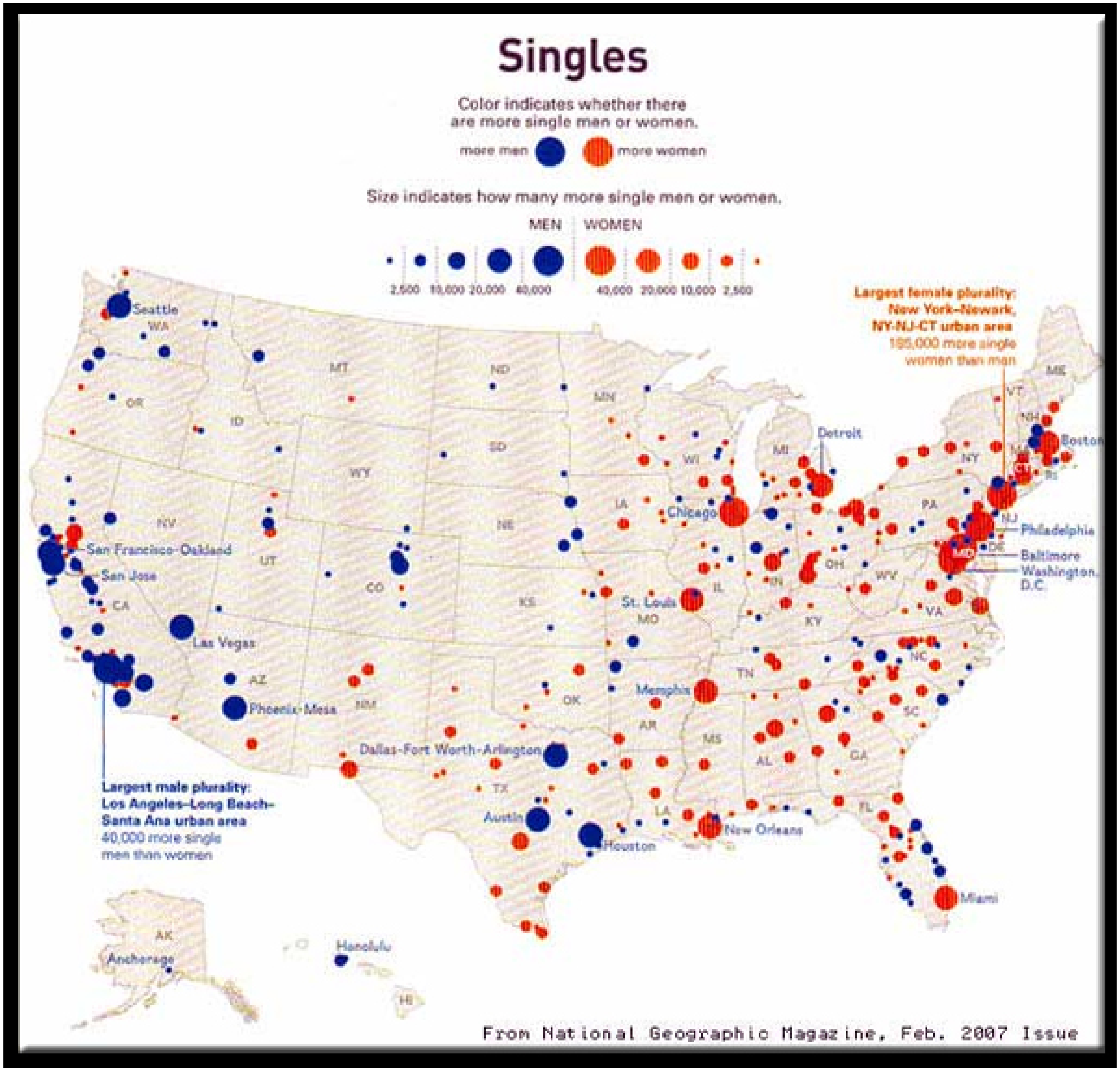
- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



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

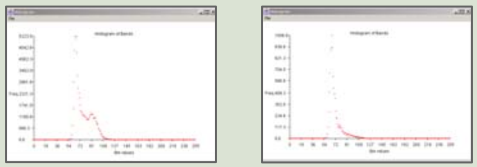
- What is Data Mining?
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)



Singles
Color indicates whether there are more single men or women.
more men
more women
Size indicates how many more single men or women.
MEN
WOMEN
2,500 10,000 20,000 40,000
40,000 20,000 10,000 2,500
Largest female plurality: New York-Newark, NY-NJ-CT urban area
185,000 more single women than men
Largest male plurality: Los Angeles-Long Beach-Santa Ana urban area
40,000 more single men than women
Seattle
San Francisco-Oakland
San Jose
Las Vegas
Phoenix-Mesa
Dallas-Fort Worth-Arlington
Austin
Houston
New Orleans
Memphis
St. Louis
Chicago
Detroit
Boston
Philadelphia
Baltimore
Washington, D.C.
Miami
Anchorage
Honolulu
From National Geographic Magazine, Feb. 2007 Issue

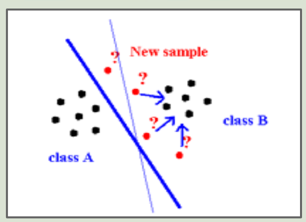
Data Mining: Confluence of Multiple Disciplines

Statistics

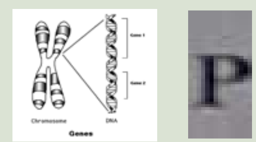


Database Techniques

Machine Learning

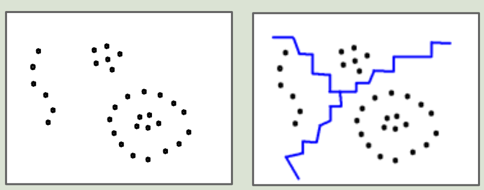


Optimization Techniques



20x20 ~ 2^400 ≈ 10^120 patterns

Pattern Recognition



Visualization

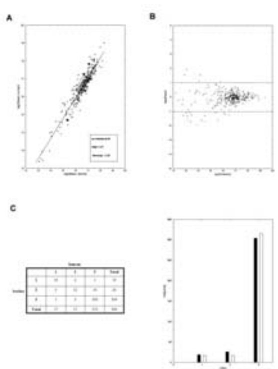
Data Mining Applications

- 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

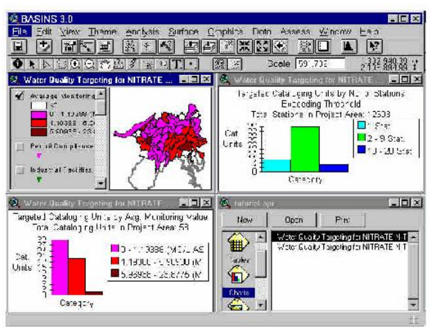
- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, …
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and 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

Similarities Between Data Miners and Doctors



Data Characteristics

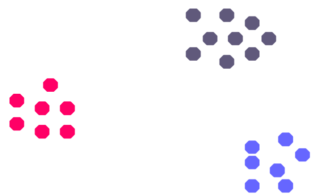
Your Symptoms?



Data Mining Techniques

Medical Devices

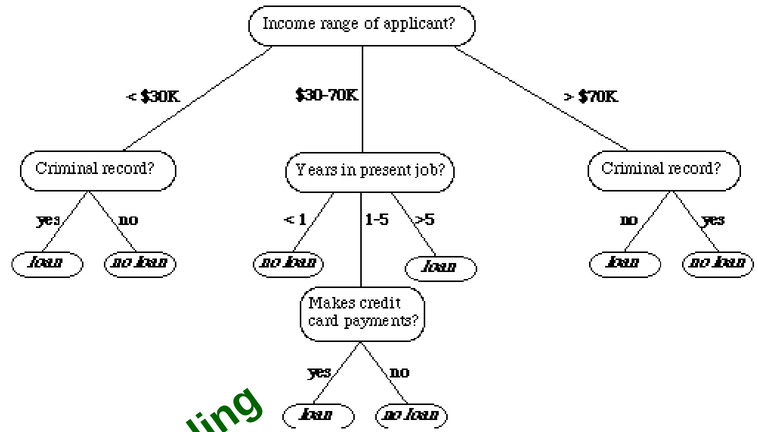
Data Mining Tasks …



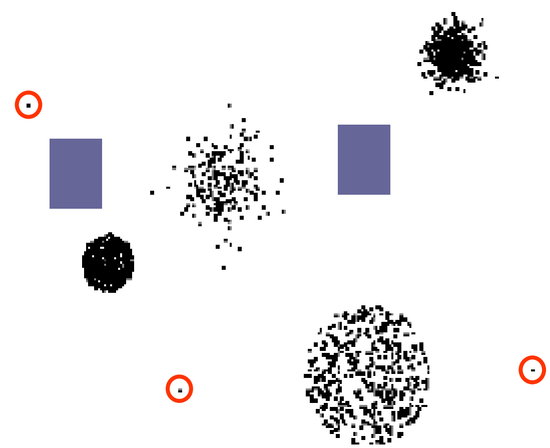
Clustering

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

Predictive Modeling



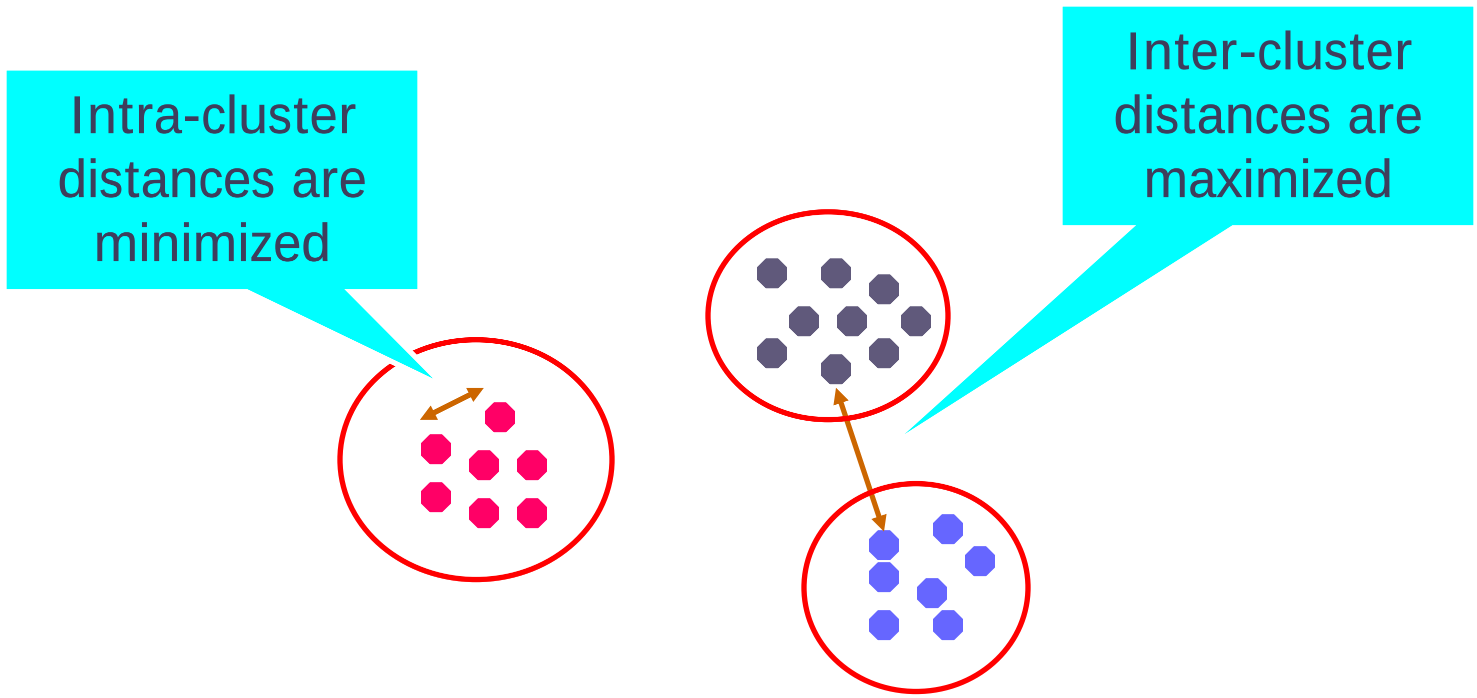
Anomaly Detection



Association Analysis

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



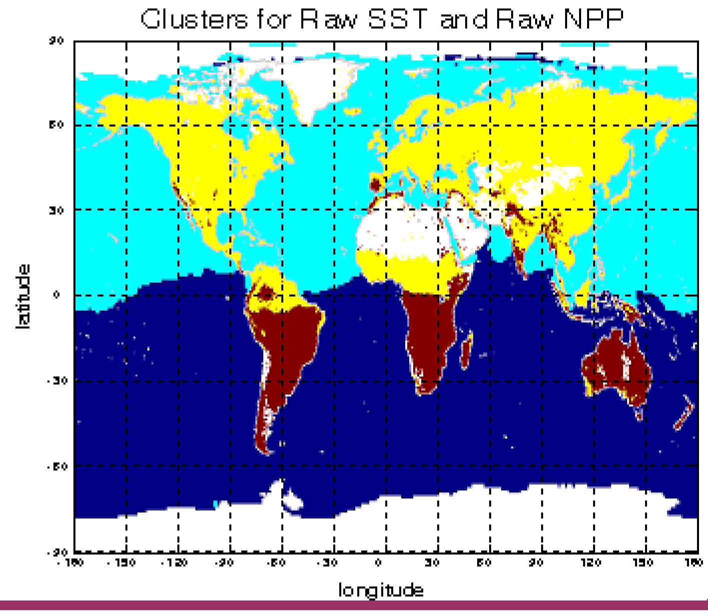
Applications of Cluster Analysis

- Understanding
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations

	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

- 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.



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

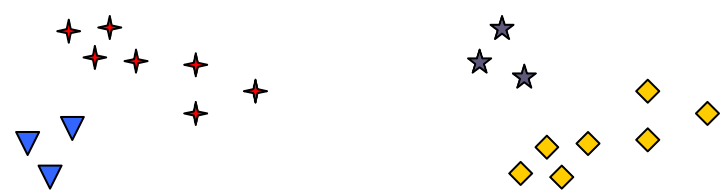
- Document Clustering:
  - **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

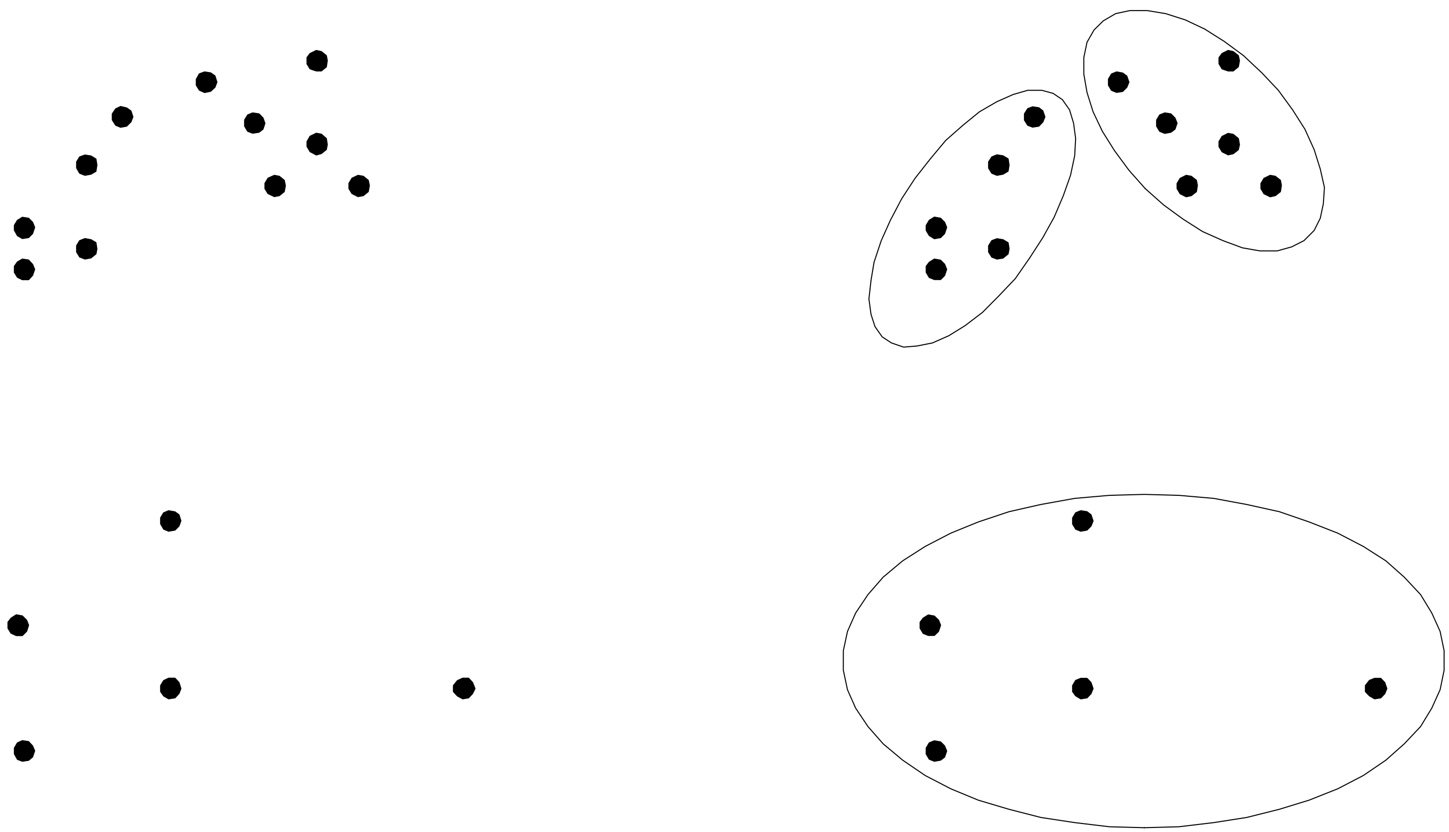


Four Clusters

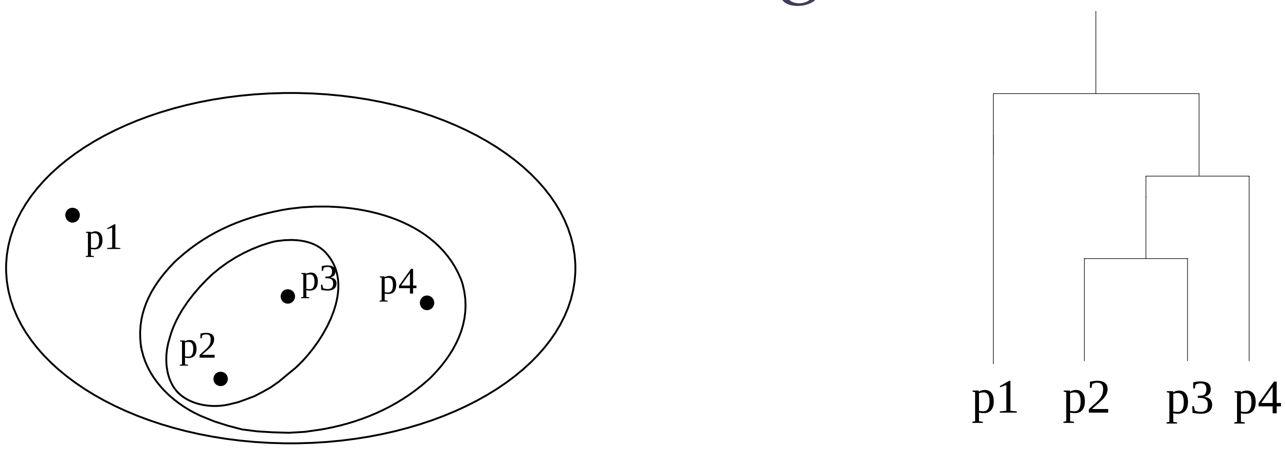
Types of Clusterings

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

Partitional Clustering



Hierarchical Clustering



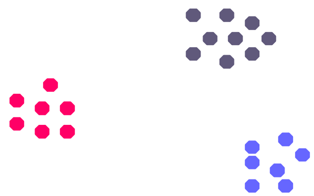
Other Distinctions Between Sets of Clusters

- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or 'border' points
- Fuzzy versus non-fuzzy
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- Partial versus complete
  - In some cases, we only want to cluster some of the data
- Heterogeneous versus homogeneous
  - Clusters of widely different sizes, shapes, and densities

Characteristics of the Input Data Are Important

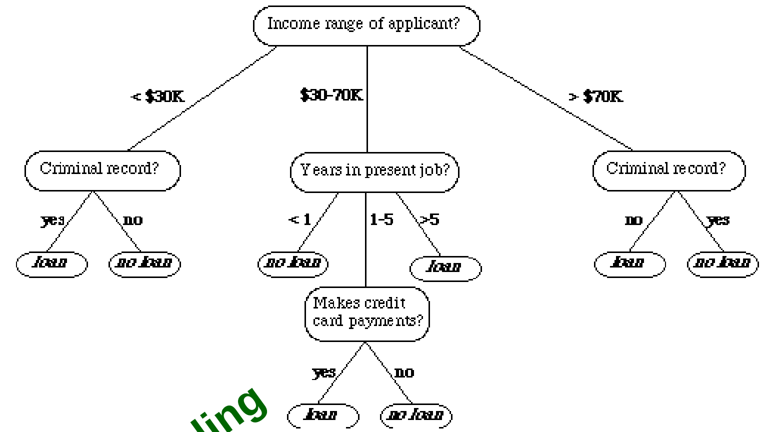
- Type of proximity or density measure
  - This is a derived measure, but central to clustering
- Sparseness
  - Dictates type of similarity
  - Adds to efficiency
- Attribute type
  - Dictates type of similarity
- Type of Data
  - Dictates type of similarity
  - Other characteristics, e.g., autocorrelation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

Data Mining Tasks …



Clustering

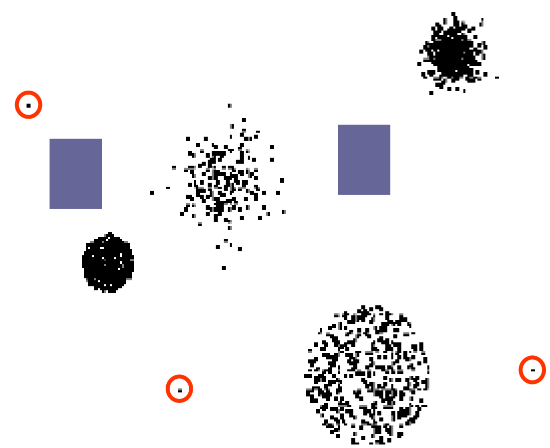
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



Predictive Modeling

Association Analysis

Anomaly Detection



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

Association Rule Mining

-

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Egg
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

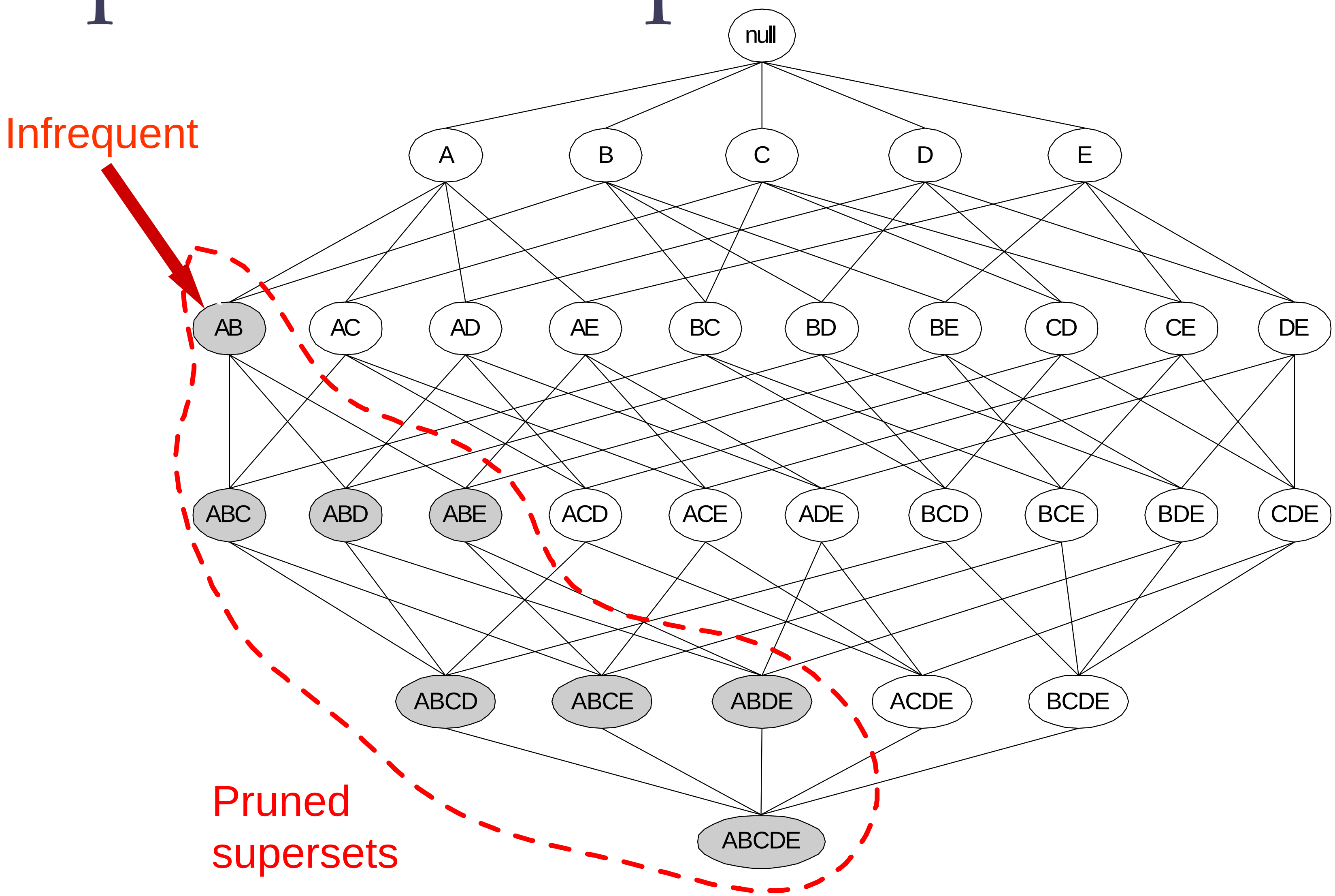
-

Protein Complex	Proteins
c1	p_1, p_2
c2	p_1, p_3, p_4, p_5
c3	p_2, p_3, p_4, p_6

- Pattern
  - A collection of one or more items
    E.g. {Milk}, {Beer, Diaper}
- Support Count (σ)
  - Frequency of occurrence of a pattern.
    E.g. σ({Bread, Milk, Diaper}) = 2
- Support (Agrawal et al. 1993)
  - Fraction of transactions that contain a pattern.
  - E.g. supp({Bread, Milk, Diaper})= 2/5 =40%

- Confidence: its interpretation as conditional probability

Apriori Principle



Correlation Computing

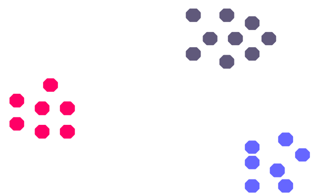
- Various Applications of Correlation Analysis
  - i.e. Marketing Data Study, Web Search, Bioinformatics, Public Health
- A Gap between Association Rule Mining and Correlation Computing
  - A lack of precise relationship between support (or confidence) based association measures and correlation measures.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

Application Deployment Challenge

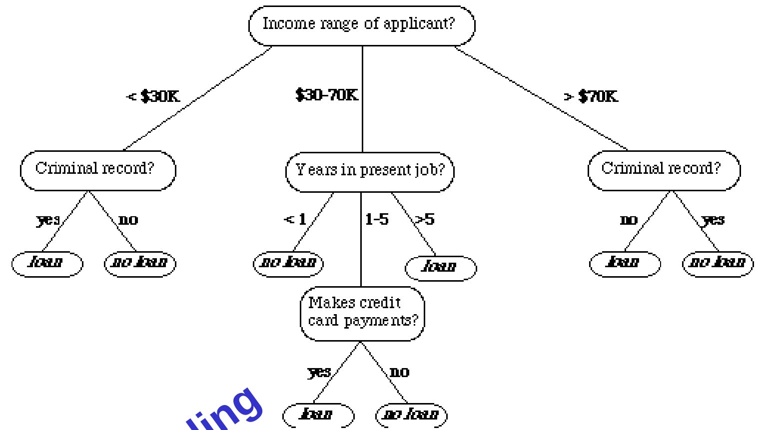
- AMAZON.COM: Product Promotion
- Answer the question: Customers who bought this book also bought?

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

Data Mining Tasks ...



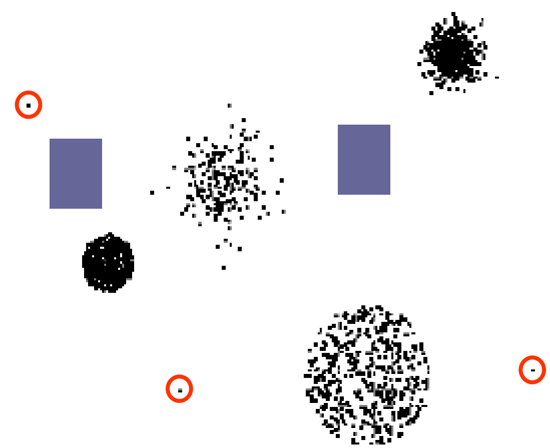
Clustering



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

Anomaly Detection



Association Analysis

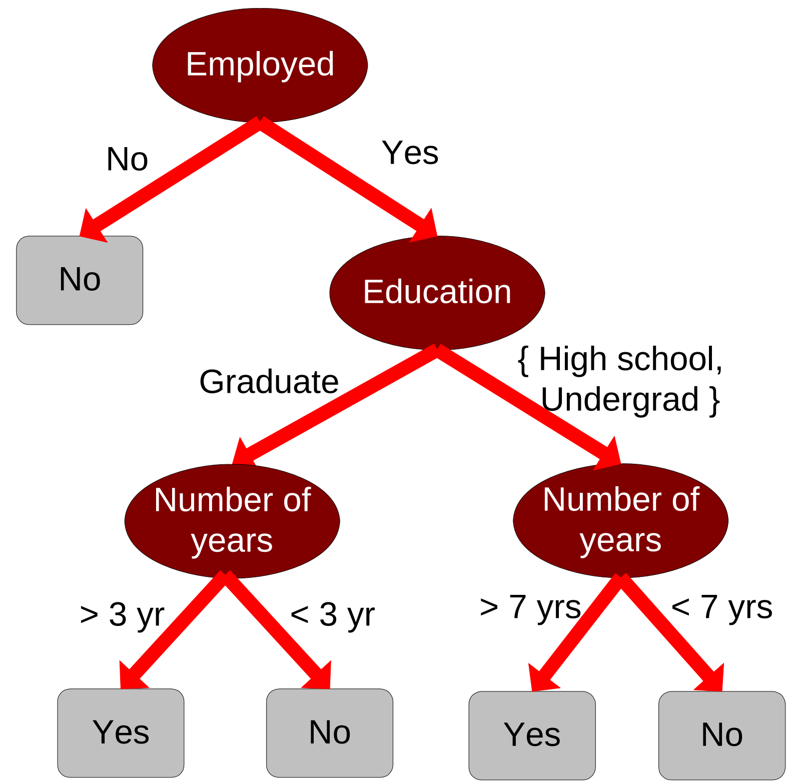
Predictive Modeling: Classification

- 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
…	…	…	…	…

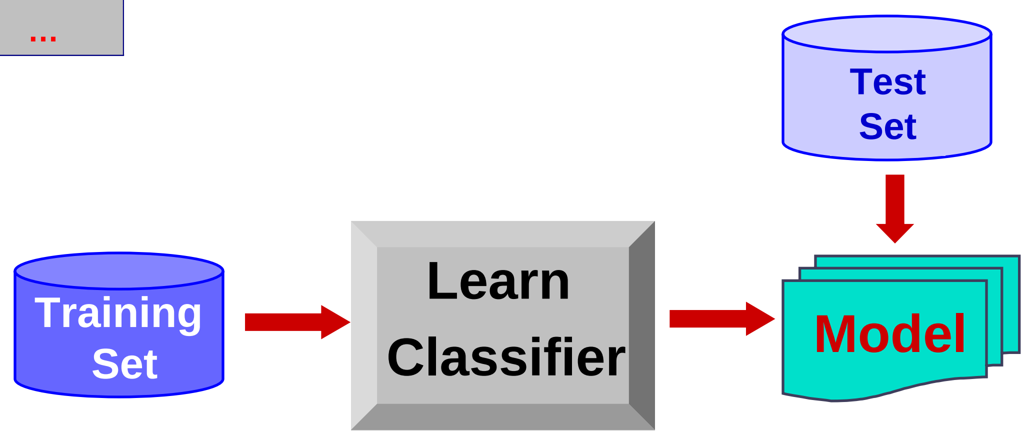


Classification Example

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	?
…	…	…	…	…

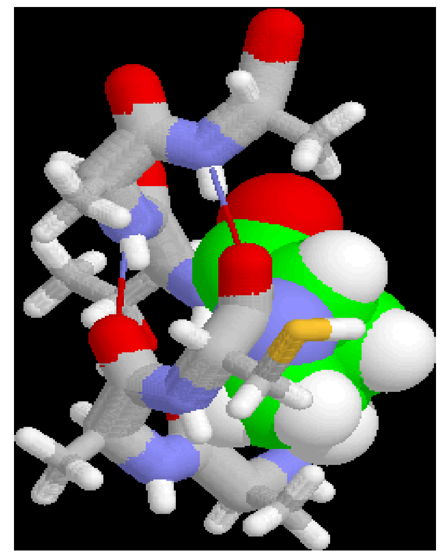


Examples of Classification Task

- 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

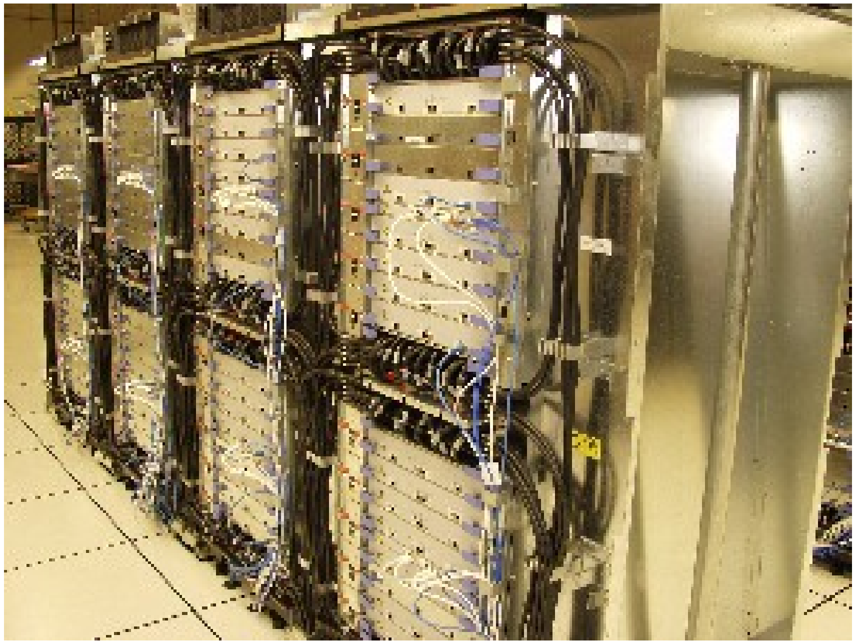
- Fraud Detection
  - **Goal:** Predict fraudulent cases in credit card transactions.
  - **Approach:**
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions 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.

System Event Logs/Job Logs

- Failure Prediction using Event Logs
- Significantly improve Fault Tolerance and Resource Management strategies



Web Usage Mining

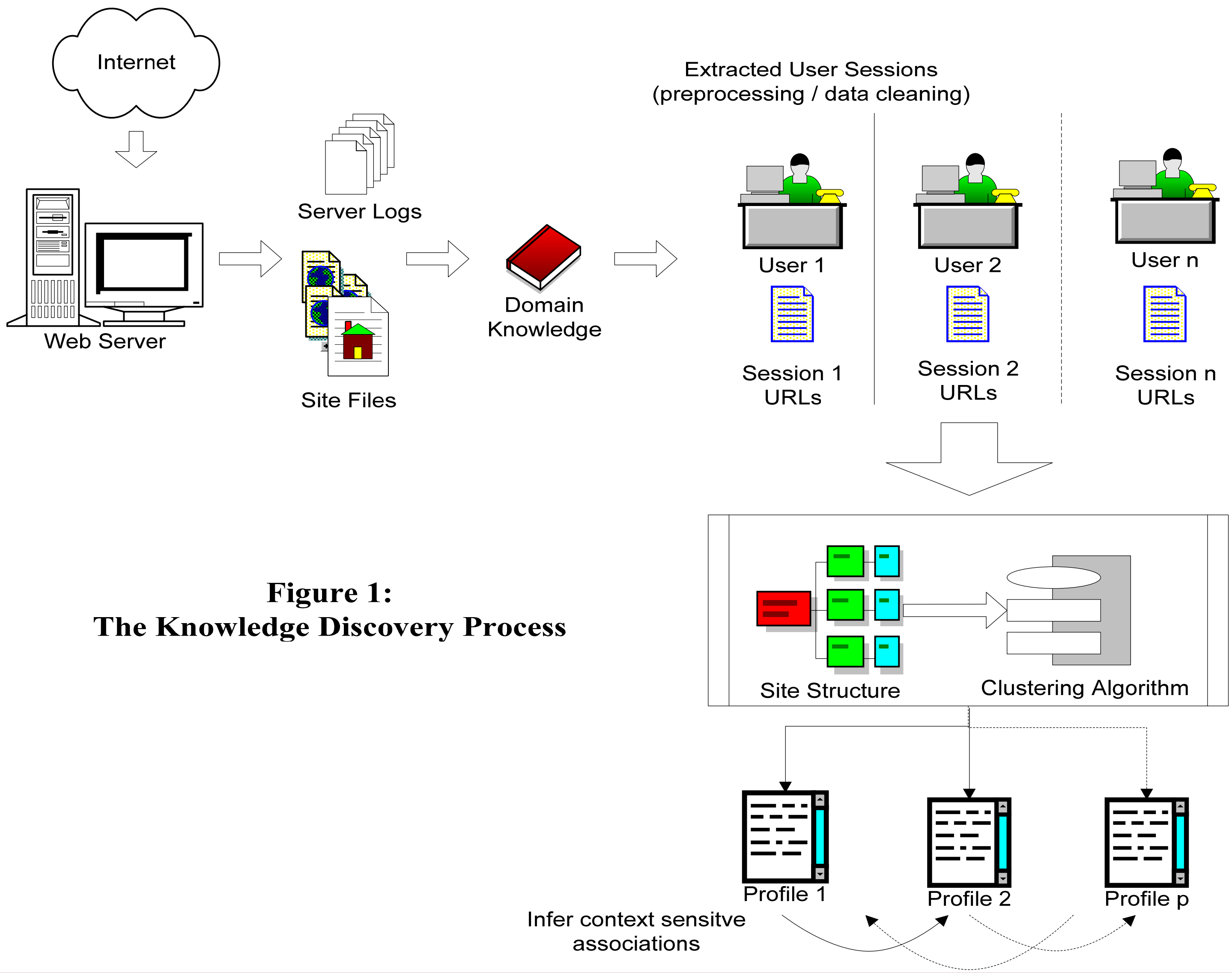


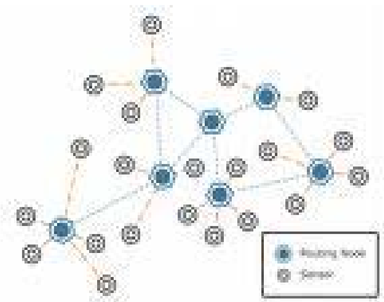
Figure 1:
The Knowledge Discovery Process

User-directed Knowledge Discovery in Wireless Sensor Network

- Learning Active Users Behavior
  - Better Sensor Network Management
  - Identifying Sensor Spoofing

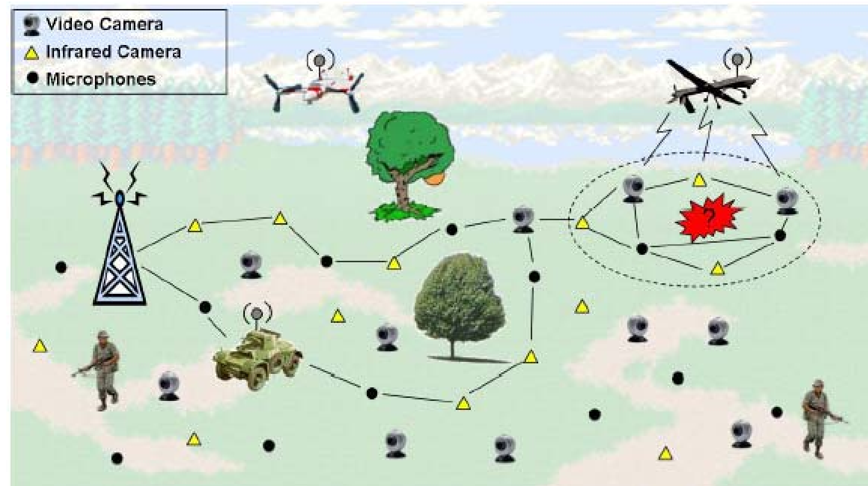
    E.g. Radio-frequency (RF) sensors are vulnerable to spoofing

    the enemy can spoof as friendly forces



Wireless Sensor Networks

- Enemy are the passive users of the system.
- Learn the enemy's usage patterns



- Better Solutions?
  - Enemy Identification ?
  - Where is the enemy?

    Historical Patterns, Joint Learning
  - What are the enemy's goal? (Semantic Constraints)

Classification Techniques

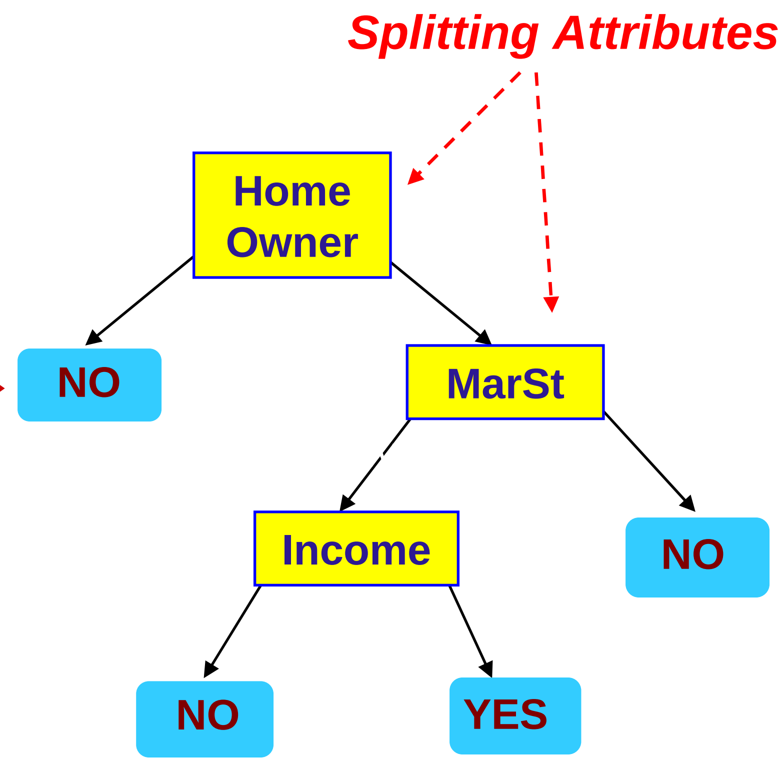
- Base Classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines

- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

Example of a Decision Tree

categorical categorical continuous class

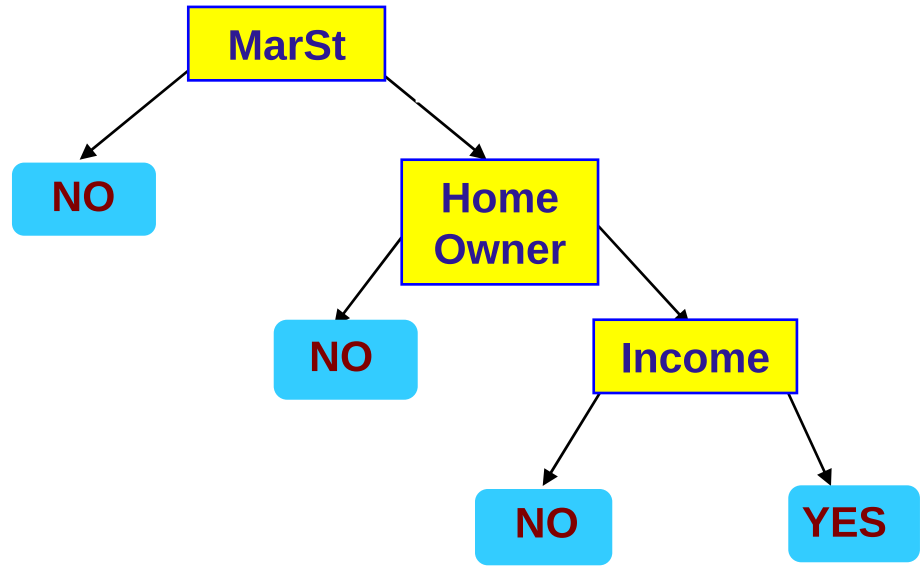
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Another Example of Decision Tree

categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

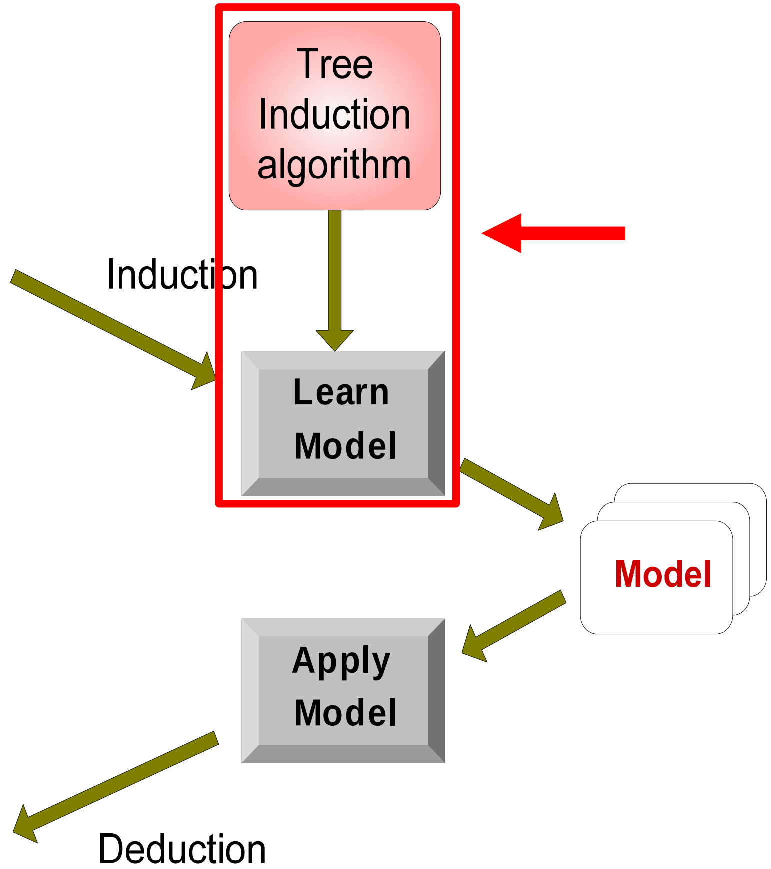


There could be more than one tree that fits the same data!

Decision Tree Classification Task

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

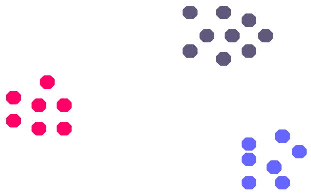
Training Set



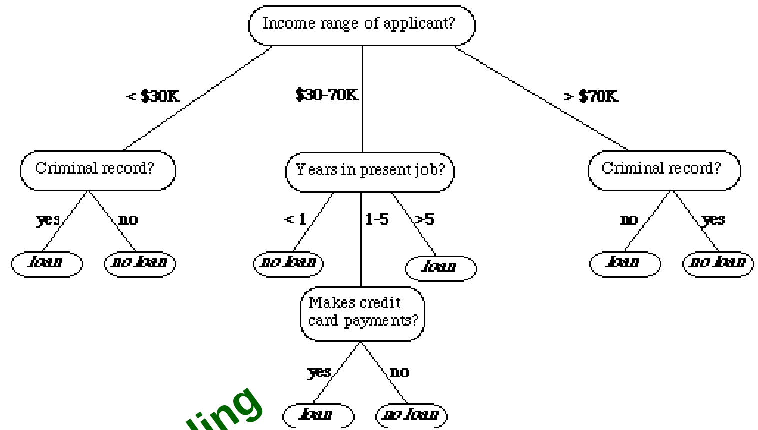
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Data Mining Tasks …



Clustering

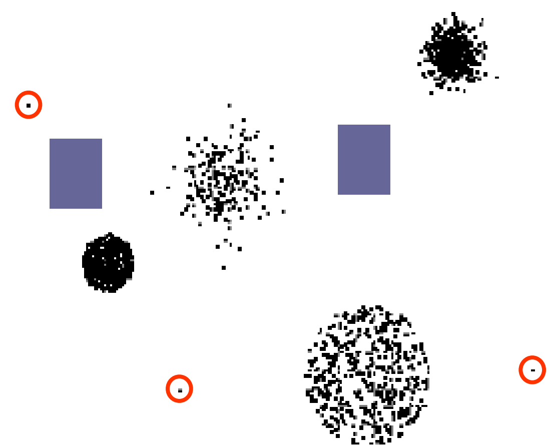


Predictive Modeling

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

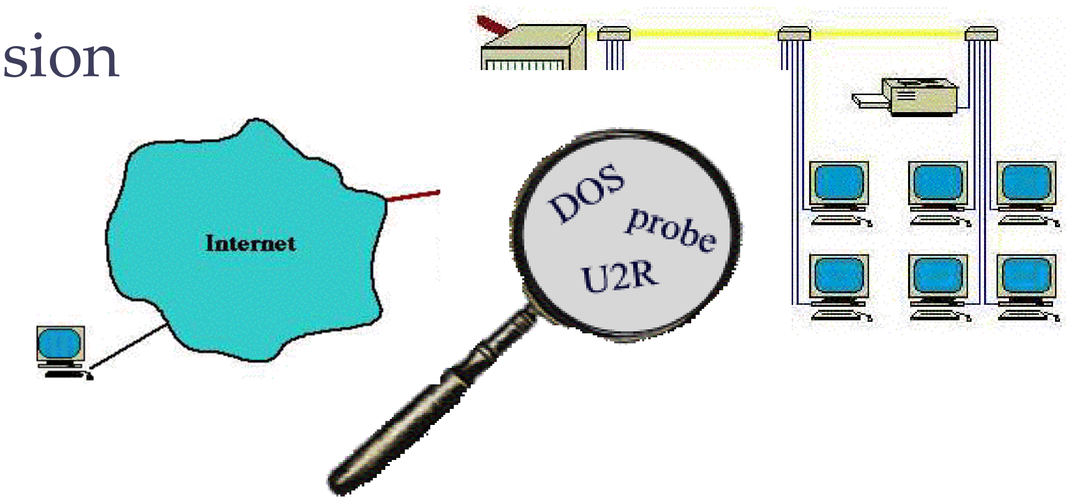
Association Analysis

Anomaly Detection



Deviation/Anomaly Detection

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



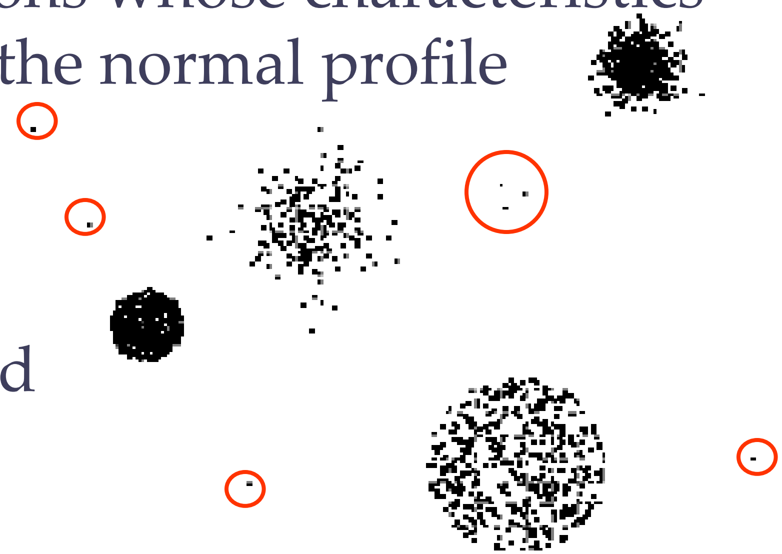
Anomaly Detection

- 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

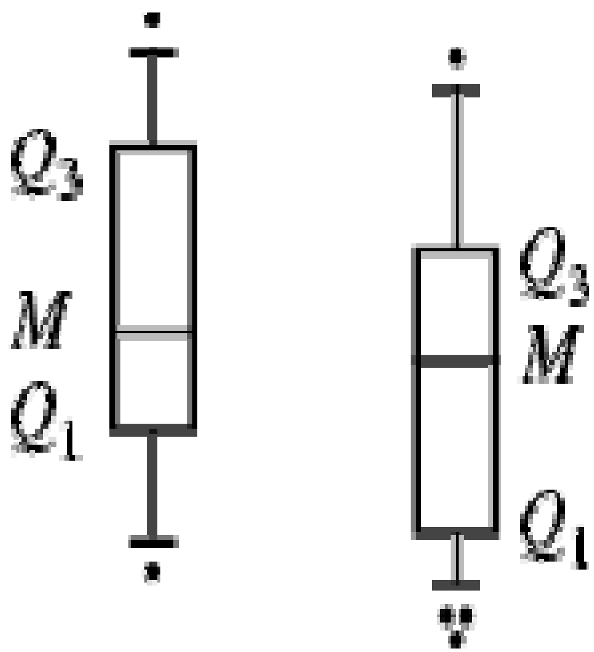
Anomaly Detection Schemes

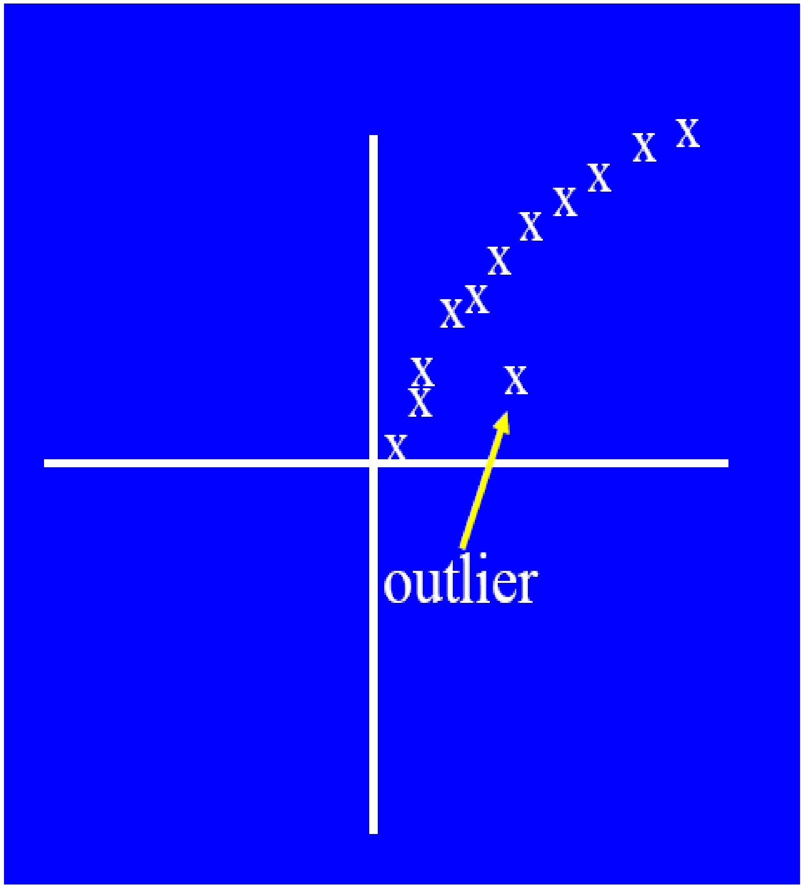
- General Steps
  - Build a profile of the "normal" behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the "normal" profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile
- Types of anomaly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



Graphical Approaches

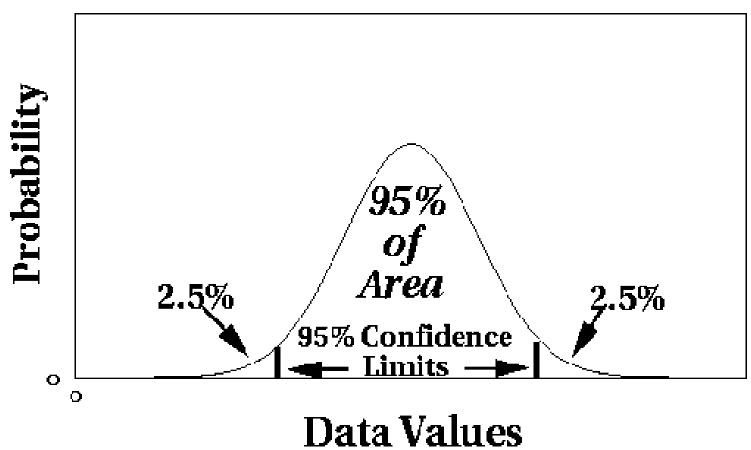
- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)
- Limitations
  - Time consuming
  - Subjective





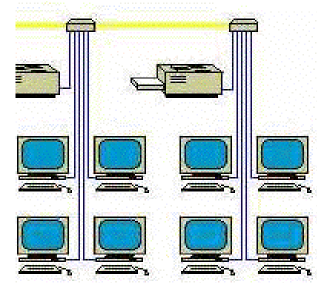
Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Intrusion Detection

- Intrusion Detection System
  - combination of software and hardware that attempts to perform intrusion detection
  - raises the alarm when possible intrusion happens



- Traditional intrusion detection system IDS tools (e.g. SNORT) are based on signatures of known attacks
- Limitations
  - Signature database has to be manually revised for each new type of discovered intrusion
  - They cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system

SNORT

www.snort.org

Data Mining for Network Intrusion Detection

- *Misuse detection*
  - Predictive models are built from labeled labeled data sets (instances are labeled as “normal” or “intrusive”)
  - These models can be more sophisticated and precise than manually created signatures
  - Unable to detect attacks whose instances have not yet been observed
- *Anomaly detection*
  - Identifies anomalies as deviations from “normal” behavior
  - Potential for high false alarm rate - previously unseen (yet legitimate) system behaviors may also be recognized as anomalies

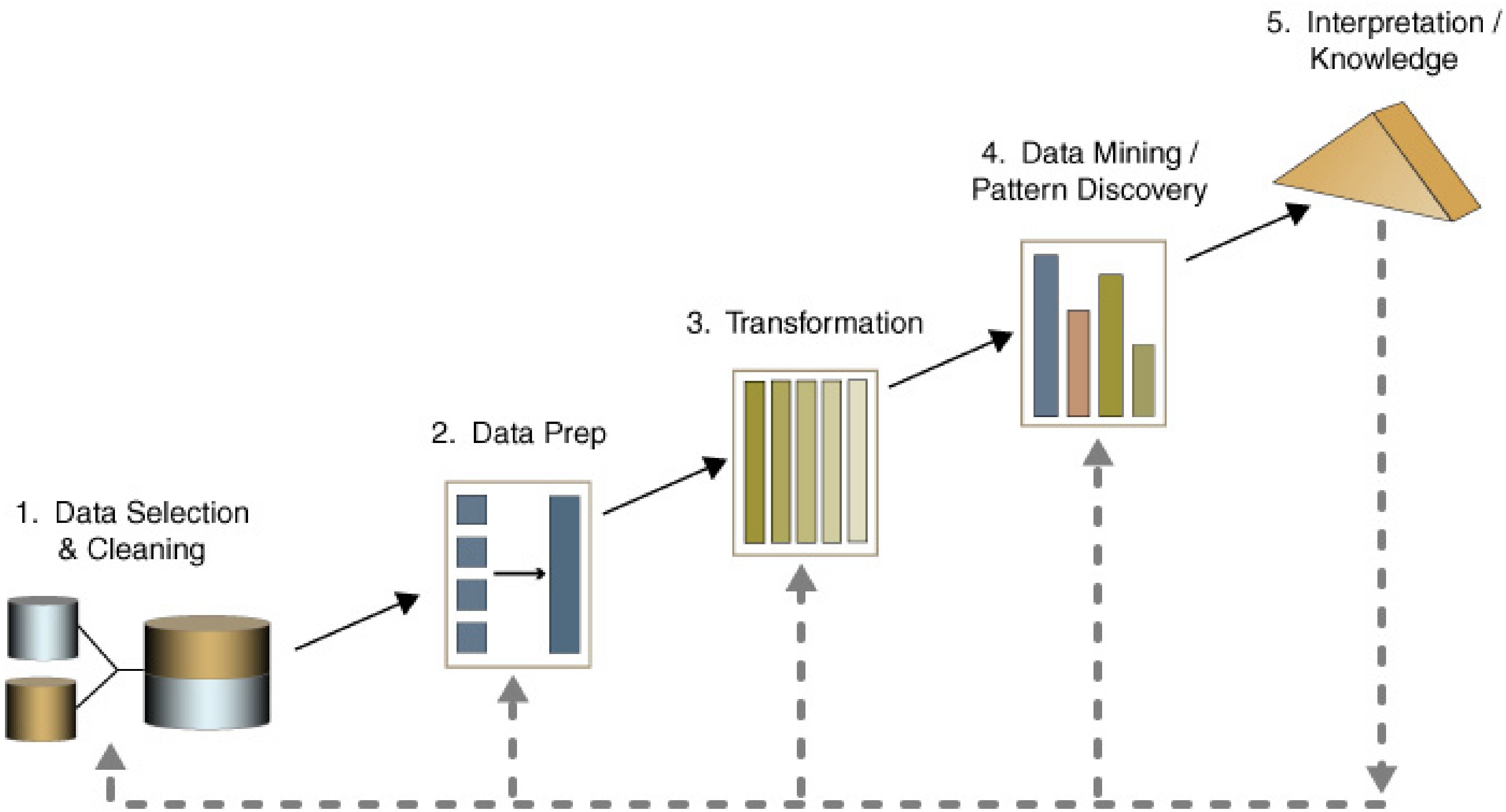
KDD Process

- 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



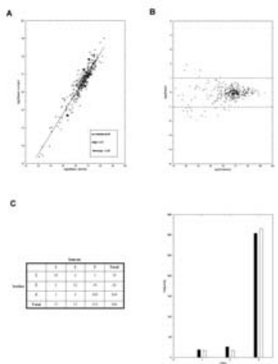
An Overview of the Steps That Compose the KDD Process

Challenges of Data Mining

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

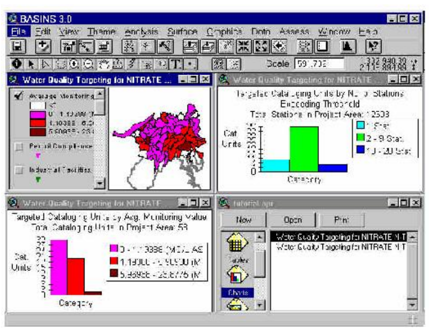
Personal Knowledge Value

Similarities Between Data Miners and Doctors



Data Characteristics

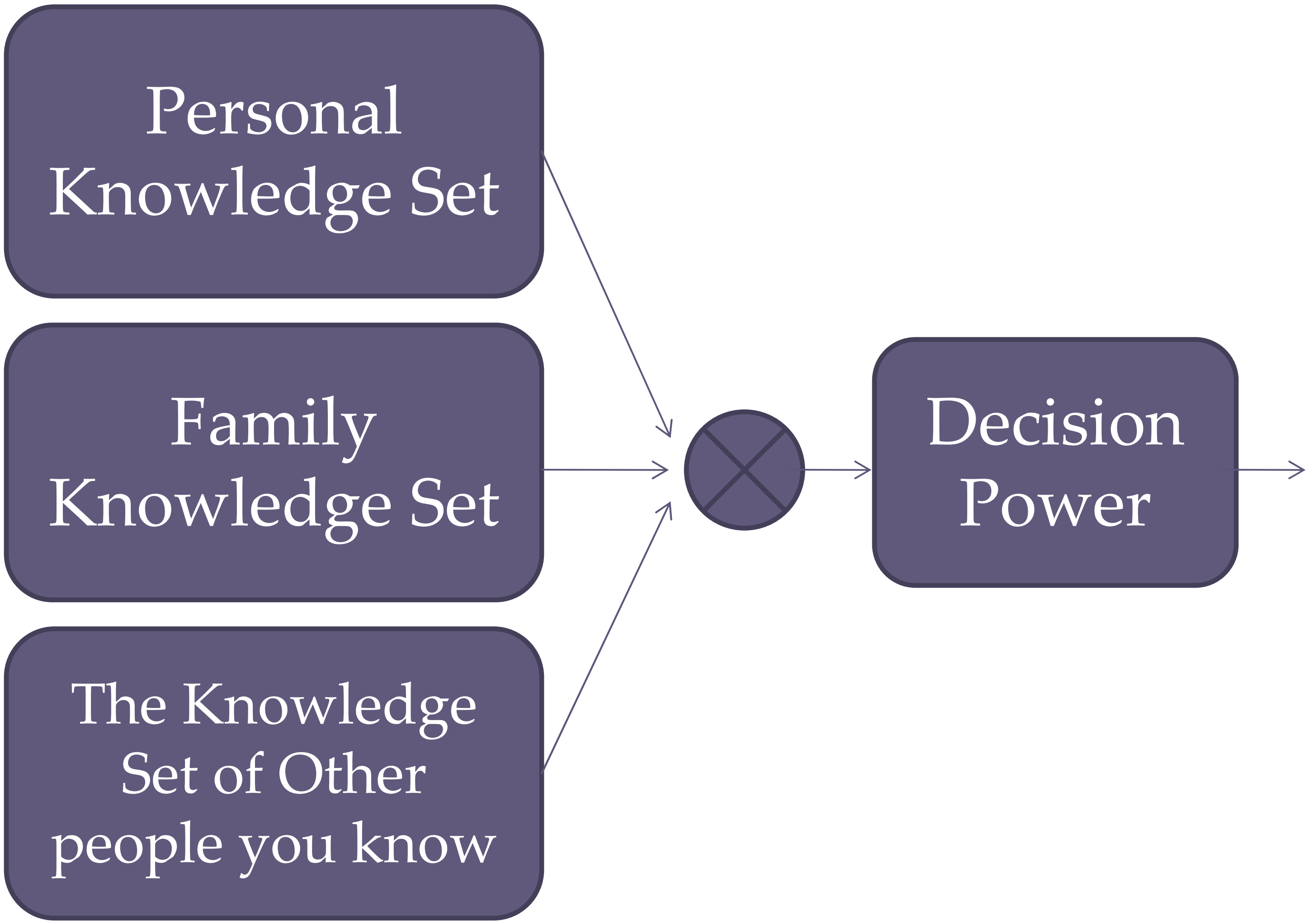
Your Symptoms?



Data Mining Techniques

Medical Devices

Life: A Data Mining Process



Commercial and Research Tools

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

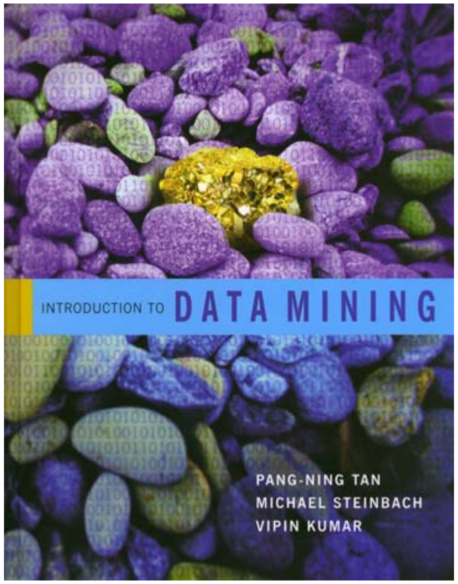
SAS:
http://www.sas.com/

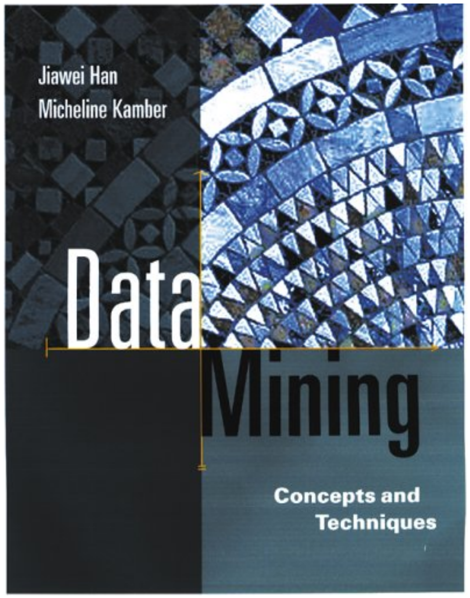
Clementine:
http://www.spss.com/spssbi/clementine/

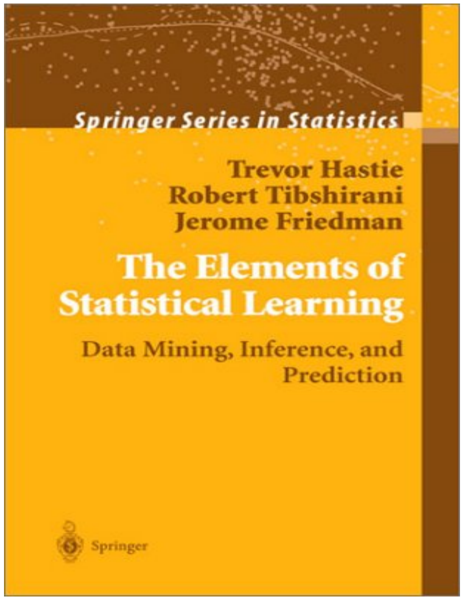
Intelligent Miner
http://www-3.ibm.com/software/data/iminer/

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

Textbooks


INTRODUCTION TO DATA MINING
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR


Jiawei Han
Micheline Kamber
Data
Mining
Concepts and
Techniques


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of
Statistical Learning
Data Mining, Inference, and
Prediction
Springer

Thank You!

http://datamining.rutgers.edu