ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΔΙΑΛΕΞΗ 1

Εισαγωγή

Συστάσεις Ι

- Ποιός είμαι εγώ:
  - Email: tsap@cs.uoi.gr
  - Γραφείο: Β.3
  - Προτιμώμενες ώρες γραφείου: απογευματινές/βραδινές.

  - Πράγματα με τα οποία έχω ασχοληθεί στο παρελθόν
    - Σχεδιασμός και ανάλυση αλγορίθμων για ranking χρησιμοποιώντας τους συνδέσμους του παγκόσμιου ιστού (PageRank-like).
    - Αλγορίθμους για clustering, ανάλυση βιολογικών δεδομένων, σημασία των αποτελεσμάτων αλγορίθμων εξόρυξης δεδομένων.
    - Web Information Retrieval, κοινωνικά δίκτυα, User Generated Content.
  - Πράγματα που με ενδιαφέρουν τώρα
    - Web mining, Social networks, User Generated Content
    - Mobile applications, Mining of mobile data.

Συστάσεις ΙΙ

- Ποιοί είσαστε εσείς:
  - Συμπληρώστε τη φόρμα με τα στοιχεία σας για την email λίστα του μαθήματος.
  - Μετά την εισαγωγή θα κάνουμε ένα εισαγωγικό quiz γνώσεων.

Γενικές πληροφορίες για το μάθημα

- Διαλεξεις: Τέταρτη 1-4 μ.μ.
  - Οι διαφάνειες θα είναι στα αγγλικά, αλλά θα προσπαθήσω να βγάζω και μετάφραση.

- Web: http://www.cs.uoi.gr/~tsap/teaching/cs-072/
  - Ανακοινώσεις, ασκήσεις, υλικό για διάβασμα διαφάνειες από τις διαλέξεις

- Βαθμολογία: TBD (To Be Defined)
  - Θα έχει τουλάχιστον 3 ασκήσεις, ίσως να έχει ένα project, ίσως να έχει τελική εξέταση.
  - Πολιτική για καθυστερημένες εργασίες:
    - Μία μέρα καθυστέρηση -10%, δύο μέρες -20%, τρεις μέρες -40%, τέσσερεις μέρες -80%, πέντε μέρες -100%.

«Προαπαιτούμενα»

- Δεν υπάρχουν προαπαιτούμενα αλλά καλό θα είναι να έχετε κάποια άνεση με:
  - Αλγορίθμους: γνώση βασικών αλγορίθμων (π.χ., sorting), και σχεδίασης αλγορίθμων (greedy algorithms, dynamic programming).
  - Πολυπλοκότητα: NP-hardness, ασυμπτωτική ανάλυση πολυπλοκότητας.
  - Δομές δεδομένων: χρήση βασικών δομών δεδομένων.
  - Προγραμματισμός: γρήγορο prototyping για τρέχετε πειράματα (οποιαδήποτε γλώσσα); matlab
  - Πιθανότητες: Γνώσεις πιθανοτήτων.
  - Γραφήματα: βασικές έννοιες γραφημάτων
  - Γραμμική άλγεβρα: πίνακες, διανύσματα, ιδιοδιανύσματα,

Στόχοι του μαθήματος

- Να καταλάβετε το είδος των προβλημάτων που μπορείτε να λύσετε χρησιμοποιώντας τεχνικές data mining.
- Να μάθετε βασικές έννοιες του data mining, που καλύπτουν και το θεωρητικό υπόβαθρο, και την εφαρμογή στην πράξη.
- Να καταλάβουμε τη θεωρία πίσω από τους αλγόριθμους και τις τεχνικές
- Να αποκτήσετε ένα σύνολο από εργαλεία (toolbox) για εξόρυξη δεδομένων.
- Να παίξετε με πραγματικά δεδομένα και να δείτε κάποια ενδιαφέροντα πραγματικά προβλήματα (ελπίζω).
- Να μάθετε διασκεδάζοντας.

Μάθημα

- Η παρακολούθηση και συμμετοχή είναι απαραίτητες
    - Κάνετε ερωτήσεις. Κάποια πράγματα δεν θα είναι ξεκάθαρα και θα πρέπει να τα επαναλάβω.
    - Αν κάτι στηρίζεται σε παλαιότερη γνώση που δεν θυμάστε ζητήστε να κάνουμε μια (σύντομη) επισκόπηση.
    - Αν υπάρχει πρόβλημα με αγγλική ορολογία και τις διαφάνειες μπορούμε να κάνουμε κάποιες ρυθμίσεις.

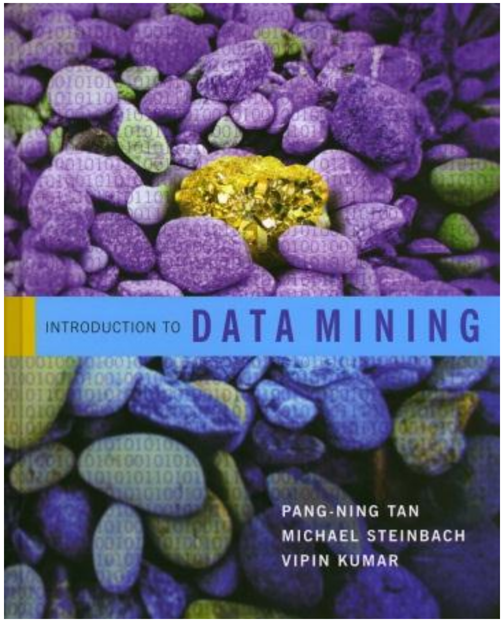
Θέματα που θα καλύψουμε

- Κάποιο υποσύνολο από τα παρακάτω
  - Frequent itemsets and association rules (συσχετισμοί)
  - Covering problems
  - Definitions and Computation of Similarity
  - Clustering (συσταδιοποίηση), co-clustering, compression
  - Classification (κατηγοριοποίηση)
  - Dimensionality Reduction
  - Ranking (ιεραρχηση/ταξινόμηση)
  - Recommendation stystems
  - Graph Analysis
  - Map-Reduce tools
  - Time-series analysis
  - Aggregation
  - Privacy preserving data mining

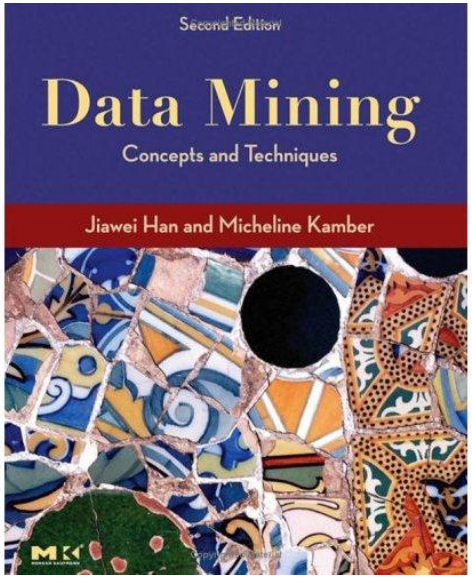
Βιβλιογραφία (ελληνικά)

- *Μ. Βαζιργιάννης και Μ. Χαλκίδη, Εξόρυξη Γνώσης από Βάσεις Δεδομένων. Τυποθήτω, Νοέμβριος 2003*
- *P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining Addison Wesley, 2006, Β. Βερύκιος και Σ. Σουραβλάς, Εκδόσεις Τζιόλα (2010).*
- *M. H. Dunham, Data Mining, Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα. Επιμέλεια Ελληνικής Έκδοσης: Β. Βερύκιος και Γ. Θεοδωρίδης. Εκδόσεις Νέων Τεχνολογιών, 2004.*

Βιβλιογραφία (αγγλικά)



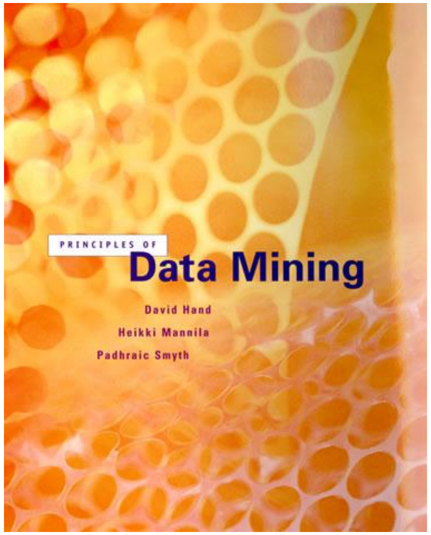
P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006

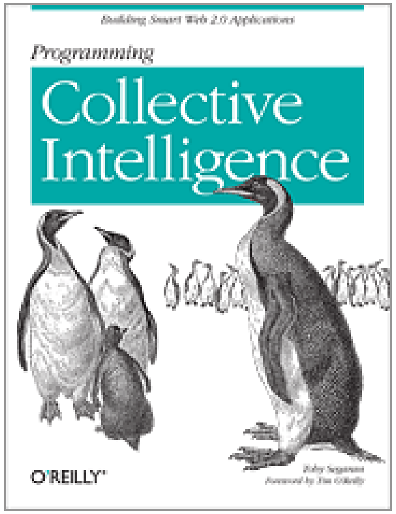


J. Han and M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006

Βιβλιογραφία (αγγλικά)

Hand, Mannila, Smyth. Principles of Data Mining





Toby Segaran, Programming Collective Intelligence. Building Smart Web 2.0 Applications

Anand Rajaraman and Jeff Ullman [Mining Massive Datasets](). Διατίθεται δωρεάν online.

Υλικό

- Εκτός από βιβλία θα χρησιμοποιήσουμε υλικό και από δημοσιευμένα άρθρα

- Για τις διαφάνειες θα δανειστούμε από πολλές πηγές
  - Εξόρυξη δεδομένων, Ε. Πιτρουρά
  - Data Mining, E. Terzi
  - P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006
  - J. Han and M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006
  - Anand Rajaraman and Jeff Ullman Mining Massive Datasets.

Quiz

- Σύντομο quiz με κάποιες βασικές ερωτήσεις γνώσεων
  - Δεν βαθμολογείστε, ο στόχος είναι να ρυθμίσω το επίπεδο του μαθήματος
  - Μπορείτε να το δώσετε ανώνυμο, αν και θα προτιμούσα να ξέρω τις αδυναμίες του καθενός
  - Έχετε όσο χρόνο χρειάζεστε.

DATA MINING
LECTURE 1

Introduction

What is data mining?

- After years of data mining there is still no unique answer to this question.

- A tentative definition:

Data mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data.

Why do we need data mining?

- Really, really huge amounts of raw data!!
  - Moore’s law: more efficient processors, larger memories
  - Communications have improved too
  - Measurement technologies have improved dramatically
  - The web, and mobile devices generate TB of data every minute
  - It possible to store and collect lots of raw data
  - The data-analysis methods are lagging behind
- Need to analyze the raw data to extract knowledge

The data is also very complex

- Multiple types of data: tables, time series, images, graphs, etc

- Spatial and temporal aspects

- Interconnected data of different types:
  - From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, images though cameras, queries to search engines

Example: transaction data

- Billions of real-life customers:
  - WALMART: 20M transactions per day
  - AT&T 300 M calls per day
  - Credit card companies: billions of transactions per day.

- The point cards allow companies to collect information about specific users

Example: document data

- Web as a document repository: estimated 50 billions of web pages
- Wikipedia: 4 million articles (and counting)
- Online collections of scientific articles
- Online news portals: steady stream of new articles every day
- Twitter: ~300 million tweets every day

Example: network data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 500 million users
- Twitter: 300 million users
- Instant messenger: ~1billion users
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

Example: genomic sequences

- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- 3*10^9 nucleotides per person → 3*10^12 nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

Example: environmental data

- Climate data (just an example)

http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php

- “a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center”

- “6000 temperature stations, 7500 precipitation stations, 2000 pressure stations”

Behavioral data

- Mobile phones today record a large amount of information about the user behavior
  - GPS records position
  - Camera produces images
  - Communication via phone and SMS
  - Text via facebook updates
  - Association with entities via check-ins

Online behavioral data

- Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.

- Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.

What can you do with the data?

- Suppose that you are the owner of a supermarket and you have collected billions of market basket data. What information would you extract from it and how would you use it?

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

Product placement

Catalog creation

Recommendations

- What if this was an online store?

What can you do with the data?

- Suppose you are a search engine and you have a toolbar log consisting of
  - pages browsed,
  - queries,
  - pages clicked,
  - ads clicked

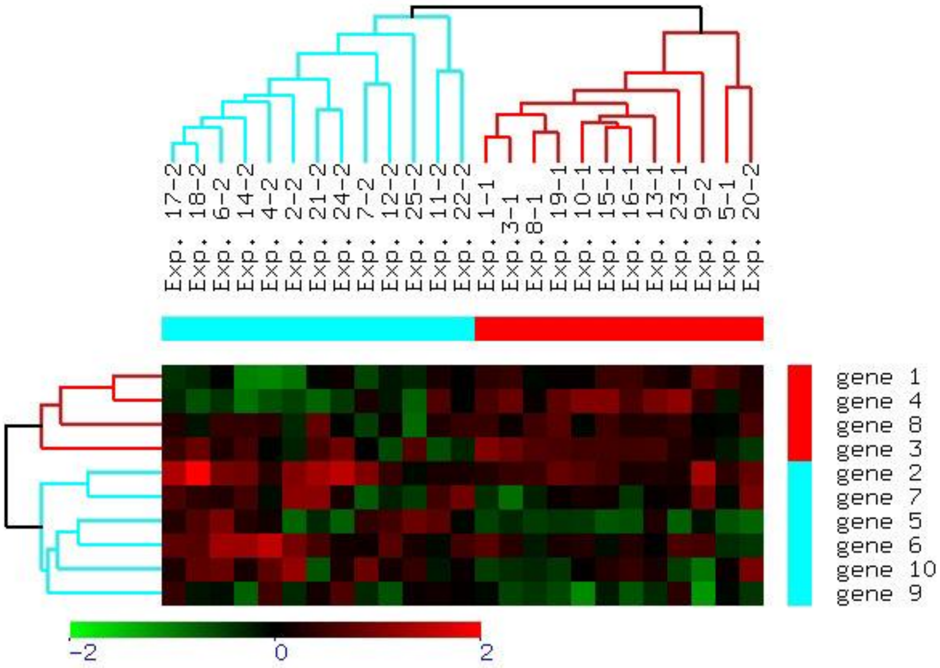
Ad click prediction

Query reformulations

each with a user id and a timestamp. What information would you like to get our of the data?

What can you do with the data?

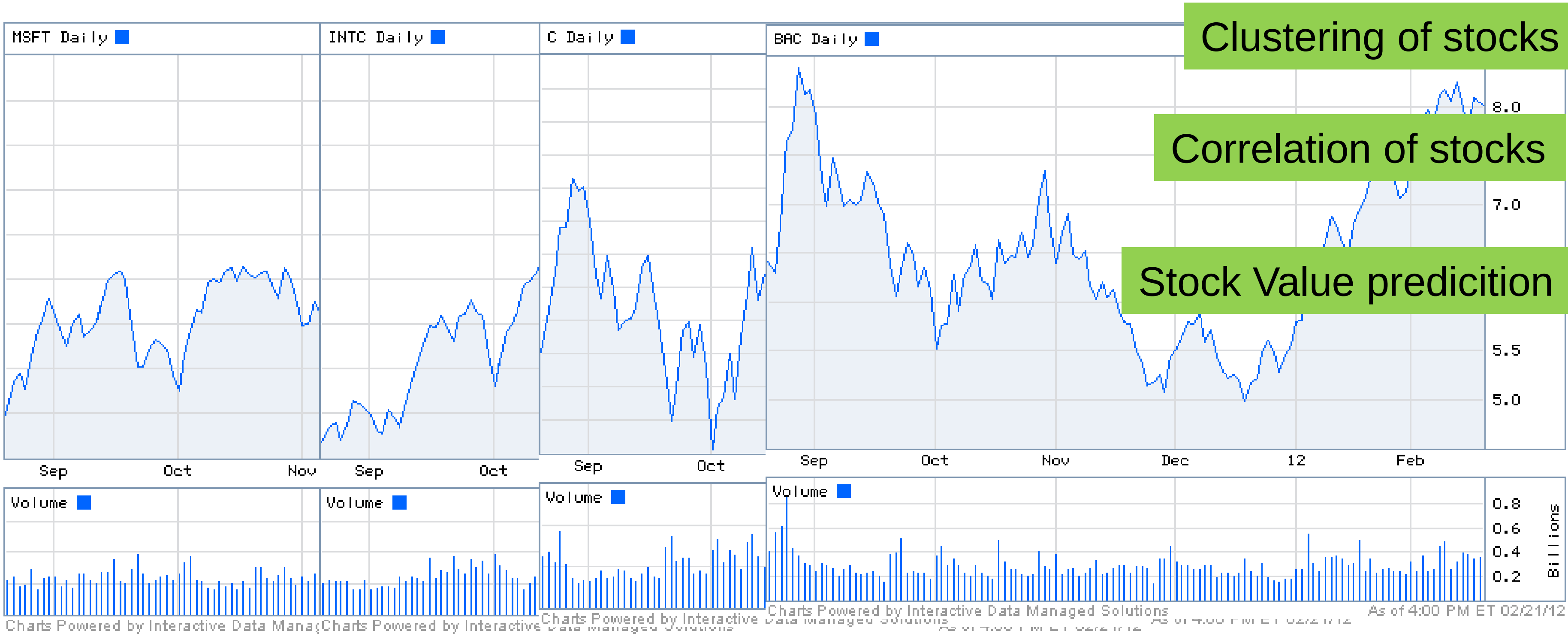
- Suppose you are biologist who has microarray expression data: thousands of genes, and their expression values over thousands of different settings (e.g. tissues). What information would you like to get out of your data?



Groups of genes and tissues

What can you do with the data?

- Suppose you are a stock broker and you observe the fluctuations of multiple stocks over time. What information would you like to get our of your data?



Clustering of stocks

Correlation of stocks

Stock Value predicition

What can you do with the data?

- You are the owner of a social network, and you have full access to the social graph, what kind of information do you want to get out of your graph?

Who is the most central node in the graph?
What is the shortest path between two nodes?
How many paths there are between two nodes?
How does information spread on the network?

Why data mining?

- Commercial point of view
  - Data has become the key competitive advantage of companies
    - Examples: Facebook, Google, Amazon
  - Being able to extract useful information out of the data is key for exploiting them commercially.
- Scientific point of view
  - Scientists are at an unprecedented position where they can collect TB of information
    - Examples: Sensor data, astronomy data, social network data, gene data
  - We need the tools to analyze such data and get a better understanding of the world
- Scale (in data size and feature dimension)
  - Why not use traditional analytic methods?
  - The amount and the complexity of data does not allow for manual processing of the data. We need automated techniques.

What is Data Mining again?

- “Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data analyst” (Hand, Mannila, Smyth)

- “Data mining is the discovery of models for data” (Rajaraman, Ullman)
  - We can have the following types of models
    - Models that explain the data (e.g., a single function)
    - Models that predict the future data instances.
    - Models that summarize the data
    - Models the extract the most prominent features of the data.

What can we do with data mining?

- Some examples:
  - Frequent itemsets and Association Rules extraction
  - Coverage
  - Clustering
  - Classification
  - Ranking
  - Exploratory analysis

Frequent Itemsets and Association Rules

- Given a set of records each of which contain some number of items from a given collection;
  - Identify sets of items (itemsets) occurring frequently together
  - 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**

Itemsets Discovered:

{Milk,Coke}

{Diaper, Milk}

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

Frequent Itemsets: Application

- Text mining: finding associated phrases in text
  - There are lots of documents that contain the phrases “association rules”, “data mining” and “efficient algorithm”

Association Rule Discovery: Application

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

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

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.

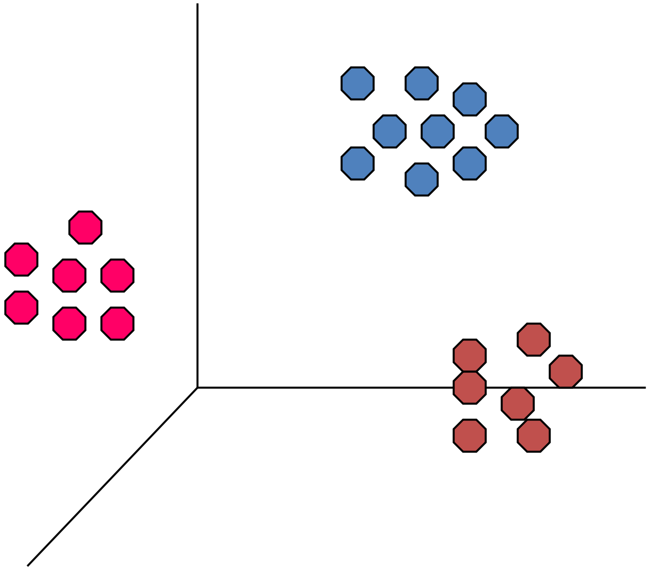
Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

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.

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

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.

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

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

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

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

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

Coverage

- Given a set of customers and items and the transaction relationship between the two, select a small set of items that “covers” all users.
  - For each user there is at least one item in the set that the user has bought.
- This formulation can be generalized for any two types of entities, and it is very useful in practice.
- Application:
  - Create a catalog to send out that has at least one item of interest for every customer.

Classification: Definition

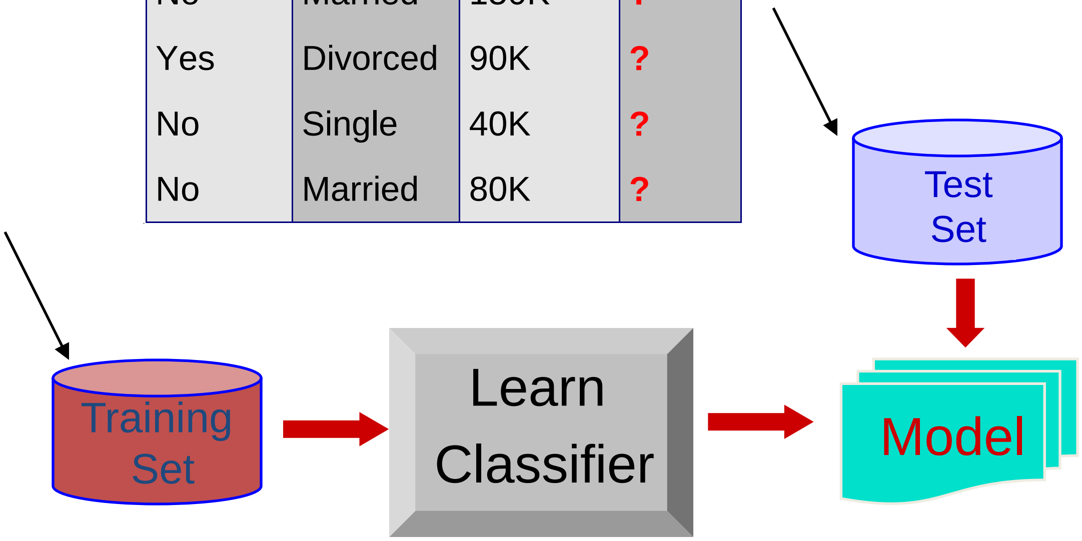
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical categorical continuous class

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

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



Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

Classification: Application 1

- Ad Click Prediction
  - Goal: Predict if a user that visits a web page will click on a displayed ad. Use it to target users with high click probability.
  - Approach:
    - Collect data for users over a period of time and record who clicks and who does not. The {click, no click} information forms the class attribute.
    - Use the history of the user (web pages browsed, queries issued) as the features.
    - Learn a classifier model and test on new users.

Classification: Application 2

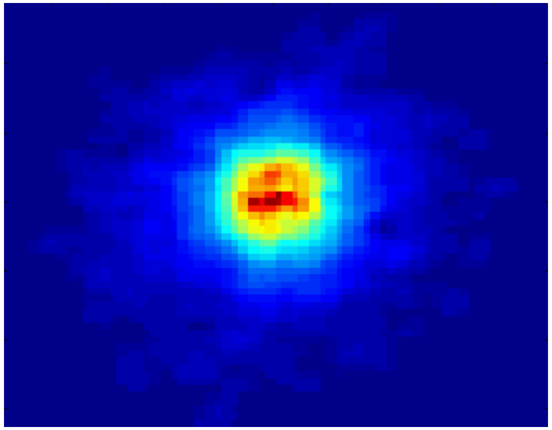
- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as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

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

Classifying Galaxies

Courtesy: http://aps.umn.edu

Early



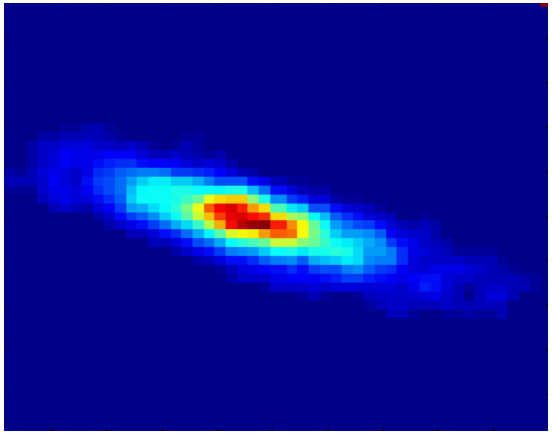
Class:

- Stages of Formation

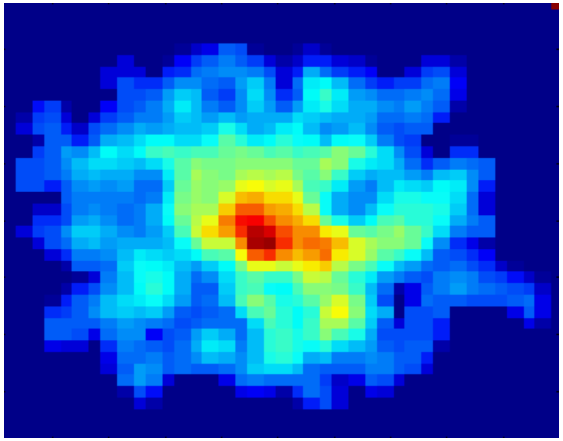
Attributes:

- 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

Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

Link Analysis Ranking

- Given a collection of web pages that are linked to each other, rank the pages according to importance (authoritativeness) in the graph
  - Intuition: A page gains authority if it is linked to by another page.

- Application: When retrieving pages, the authoritativeness is factored in the ranking.

Exploratory Analysis

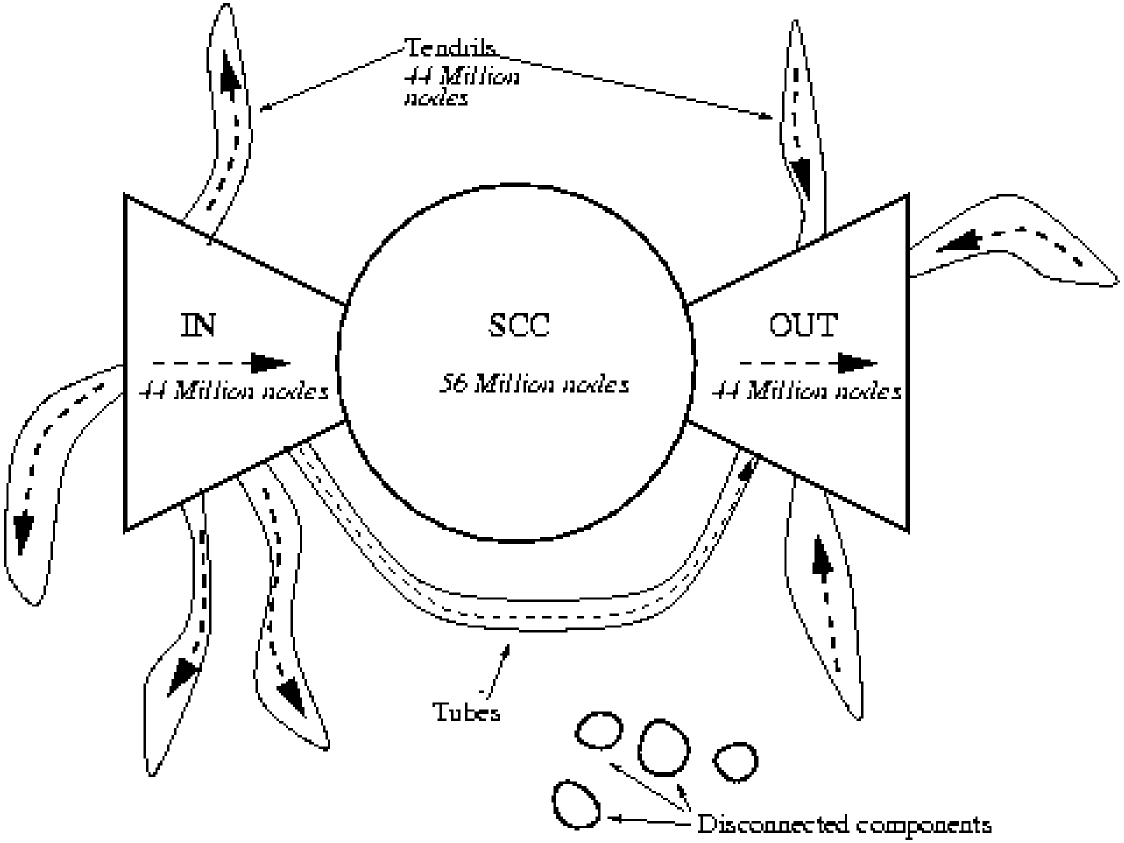
- Trying to understand the data as a physical phenomenon, and describe them with simple metrics
  - What does the web graph look like?
  - How often do people repeat the same query?
  - Are friends in facebook also friends in twitter?

- The important thing is to find the right metrics and ask the right questions

- It helps our understanding of the world, and can lead to models of the phenomena we observe.

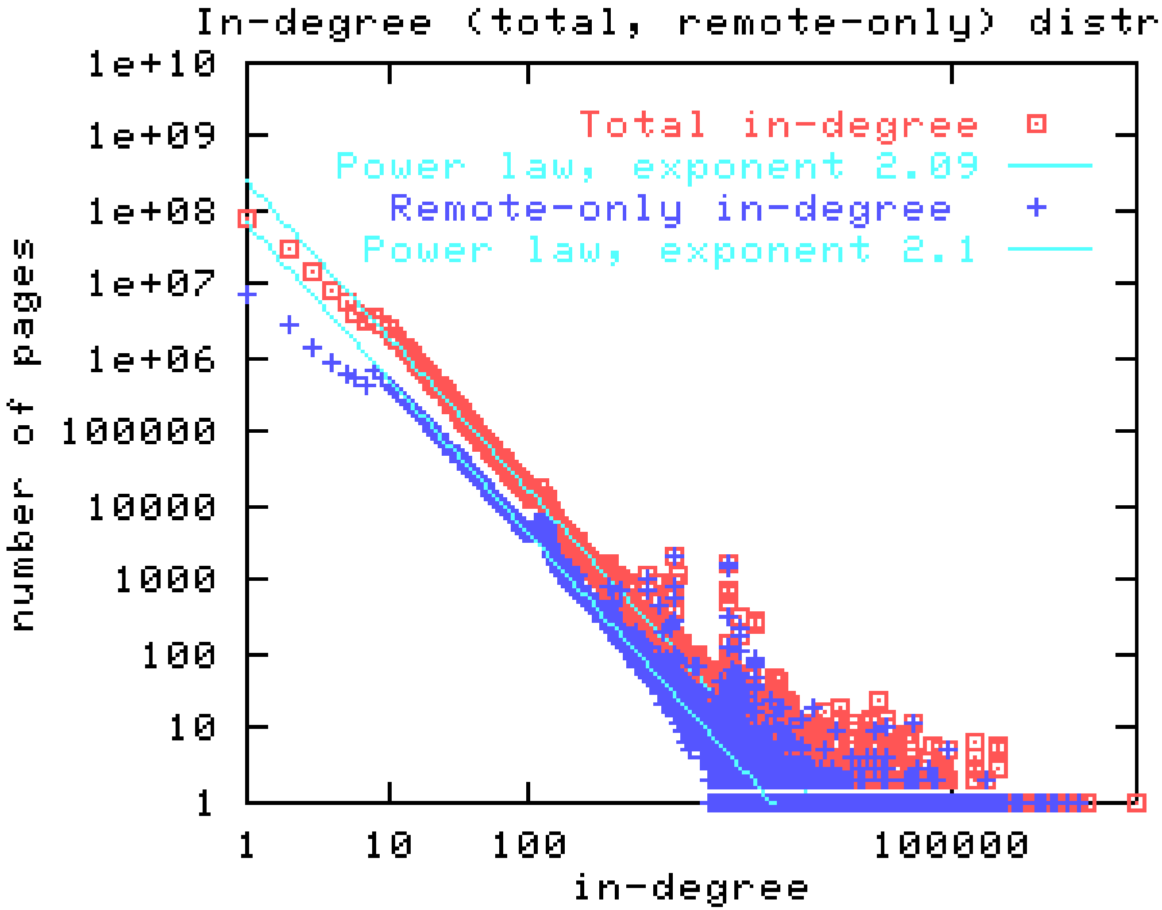
Exploratory Analysis: The Web

- What is the structure and the properties of the web?



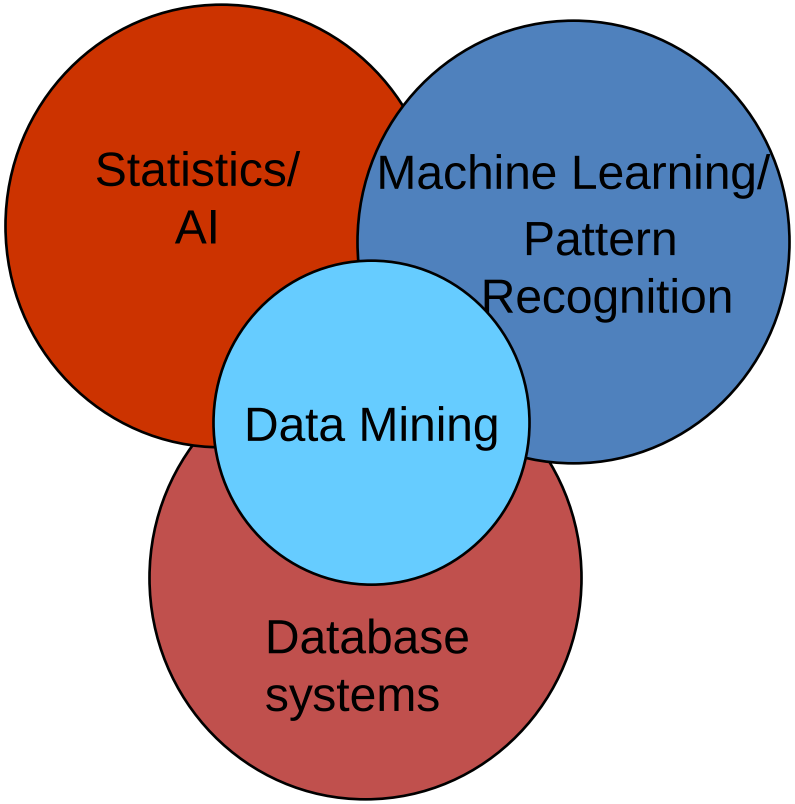
Exploratory Analysis: The Web

- What is the distribution of the incoming links?



Connections of Data Mining with other areas

- 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
  - Emphasis on the use of data



Tan, M. Steinbach and V. Kumar, Introduction to Data Mining

Cultures

- Databases: concentrate on large-scale (non-main-memory) data.
- AI (machine-learning): concentrate on complex methods, small data.
  - In today's world data is more important than algorithms
- Statistics: concentrate on models.

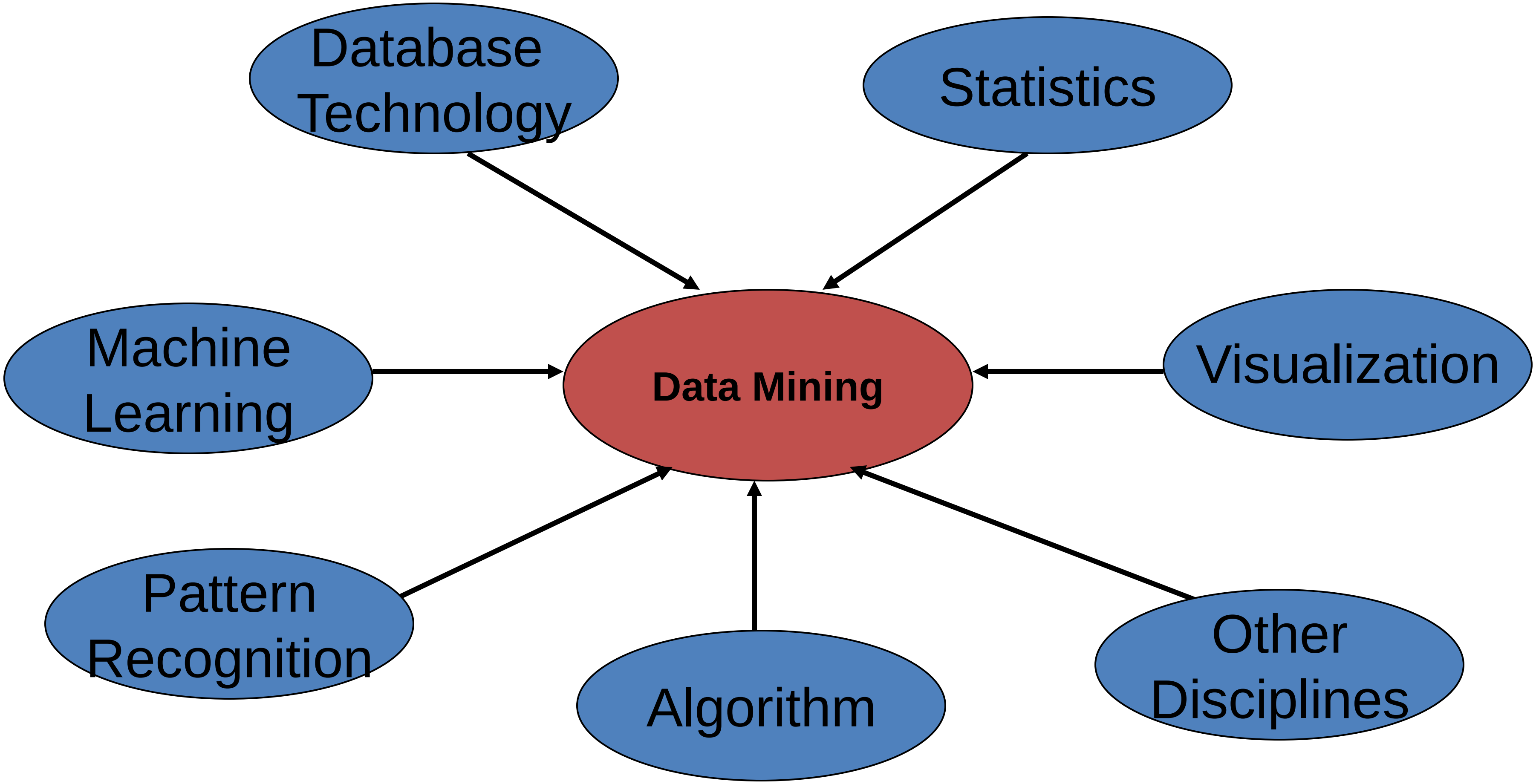
Models vs. Analytic Processing

- To a database person, data-mining is an extreme form of analytic processing – queries that examine large amounts of data.
  - Result is the query answer.
- To a statistician, data-mining is the inference of models.
  - Result is the parameters of the model.

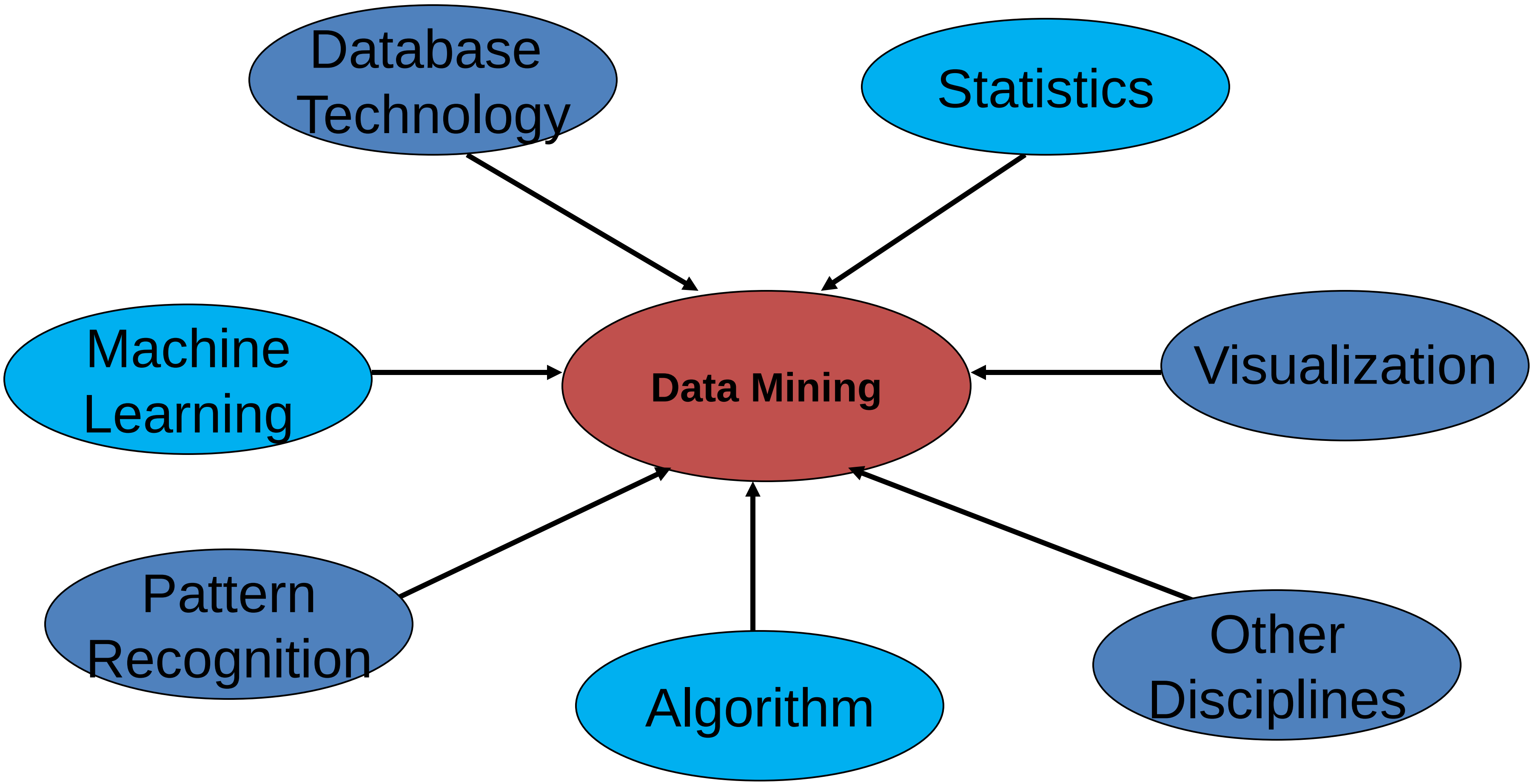
(Way too Simple) Example

- Given a billion numbers, a DB person would compute their average and standard deviation.
- A statistician might fit the billion points to the best Gaussian distribution and report the mean and standard deviation *of that distribution*.

Data Mining: Confluence of Multiple Disciplines



Data Mining: Confluence of Multiple Disciplines



Database Technology

Statistics

Machine Learning

Data Mining

Visualization

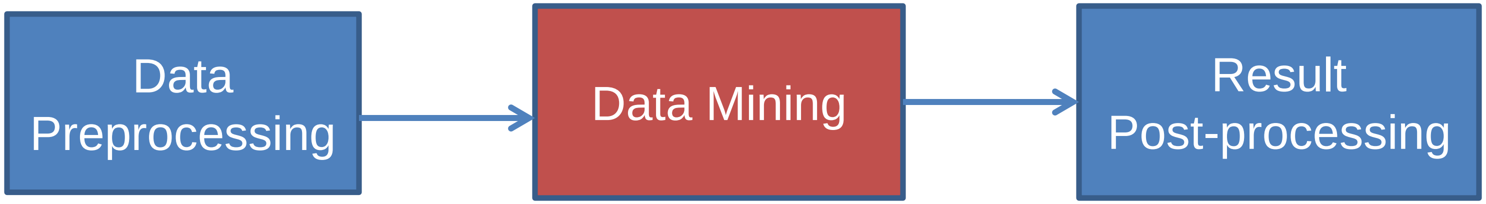
Pattern Recognition

Algorithm

Other Disciplines

The data analysis pipeline

- Mining is not the only step in the analysis process



- Preprocessing: real data is noisy, incomplete and inconsistent. Data cleaning is required to make sense of the data
  - Techniques: Sampling, Dimensionality Reduction, Feature selection.
  - A dirty work, but it is often the most important step for the analysis.
- Post-Processing: Make the data actionable and useful to the user
  - Statistical analysis of importance
  - Visualization.
- Pre- and Post-processing are often data mining tasks as well

Meaningfulness of Answers

- A big data-mining risk is that you will “discover” patterns that are meaningless.
- Statisticians call it Bonferroni’s principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.
- The Rhine Paradox: a great example of how not to conduct scientific research.

Rhine Paradox – (1)

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception.
- He devised (something like) an experiment where subjects were asked to guess 10 hidden cards – red or blue.
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right!

Rhine Paradox – (2)

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he conclude?
  - Answer on next slide.

Rhine Paradox – (3)

- He concluded that you shouldn’t tell people they have ESP; it causes them to lose it.