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!!
  - In the digital age, TB of data is generated by the second.
    - Web, Wikipedia, Mobile devices, Digital photographs and videos, Facebook, Twitter, Instagram, Transactions, sensor data, behavioral data, scientific measurements, wearable computing
  - New ways of generating data are constantly created.
  - Cheap storage has made possible to maintain this data
- Need to analyze the data to extract knowledge

Why do we need data mining?

- "The data is the computer"
  - Large amounts of data can be more powerful than complex algorithms and models
    - Google has solved many Natural Language Processing problems, simply by looking at the data
    - Example: misspellings, synonyms
  - Data is power!
    - Today, the collected data is one of the biggest assets of an online company
      - Query logs of Google, The friendship and updates of Facebook, Tweets and follows of Twitter, Amazon transactions
  - Data for the people:
    - Using data from the people activity we can improve their individual lives but also the overall society life.
  - We need a way to harness the collective intelligence

- From Data mining to Data Science

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.
  - Amazon: millions of purchases 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 in Google index
  - Several trillions overall
- Wikipedia: 4.9 million articles (and counting)
- Online news portals: steady stream of 100’s of new articles every day
- Twitter: ~500 million tweets every day

Example: network data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 1.5 billion users
- Twitter: 300 million users
- LinkedIn: 300 million users

Example: genomic sequences

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

Medical data

- Wearable devices can measure your heart rate, blood sugar, blood pressure, and other signals about your health. Medical records are becoming available to individuals
  - Wearable computing
- Brain imaging
  - Images that monitor the activity in different areas of the brain under different stimuli
    - TB of data that need to be analyzed.
- Gene and Protein interaction networks
  - It is rare that a single gene regulates deterministically the expression of a condition.
  - There are complex networks and probabilistic models that govern the protein expression.

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”
  - Spatiotemporal data

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
- 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.
- Data collected for millions of users on a daily basis

The data is also very complex

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

- 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, status updates in FB, images though cameras, queries to search engines

- Spatial and temporal aspects

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

What can you do with the data?

- In online sites such as Amazon, YouTube, Netflix the data collected can be at the individual level: we know for every user the items they have looked at, reviewed, selected.
- What information can you get from this data?

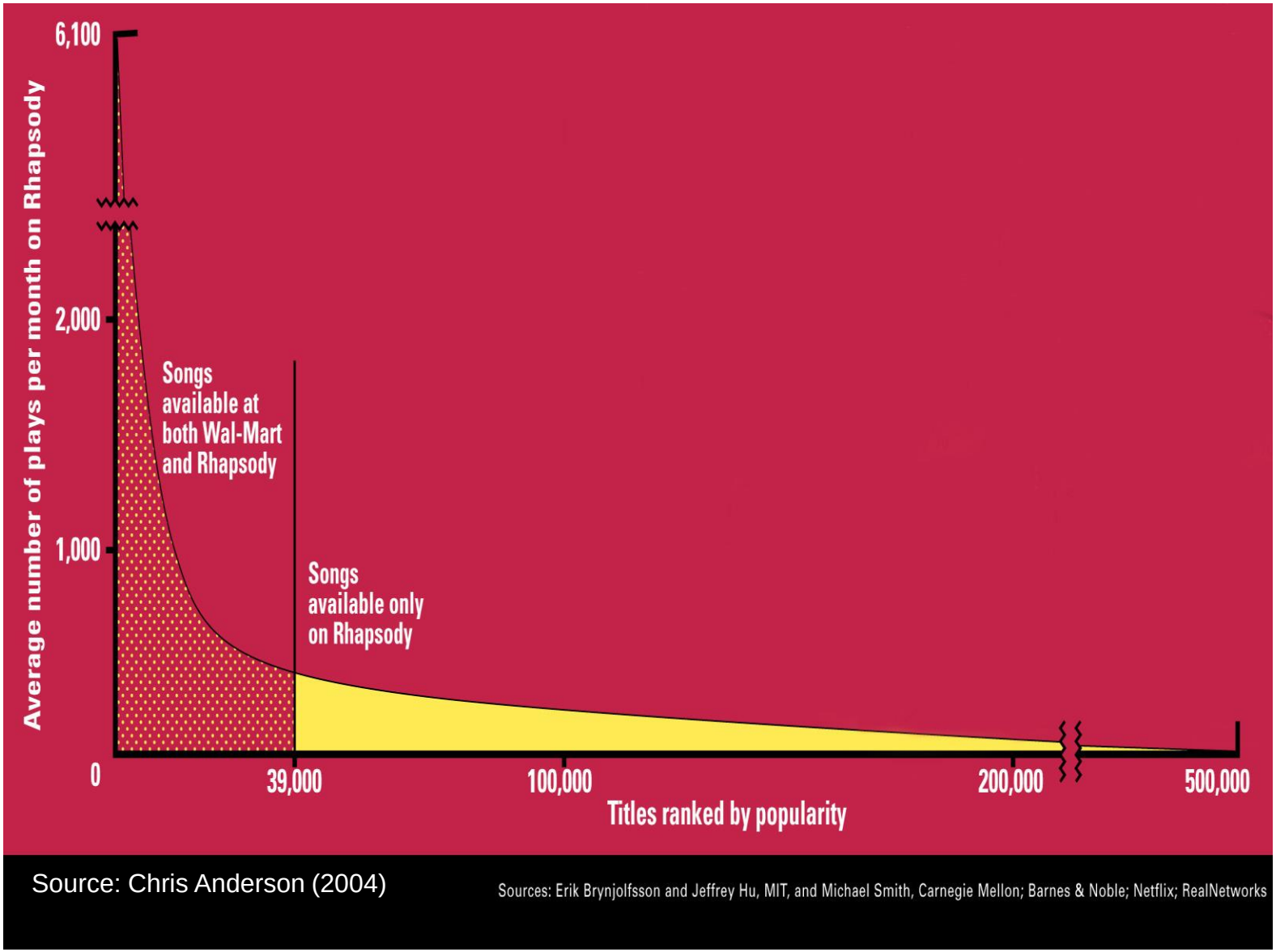
Recommendations

Amazon Recommendations

- “People who have bought this also bought…”

- A huge breakthrough for amazon
  - Took advantage of the big tail
- A big breakthrough for data mining in general

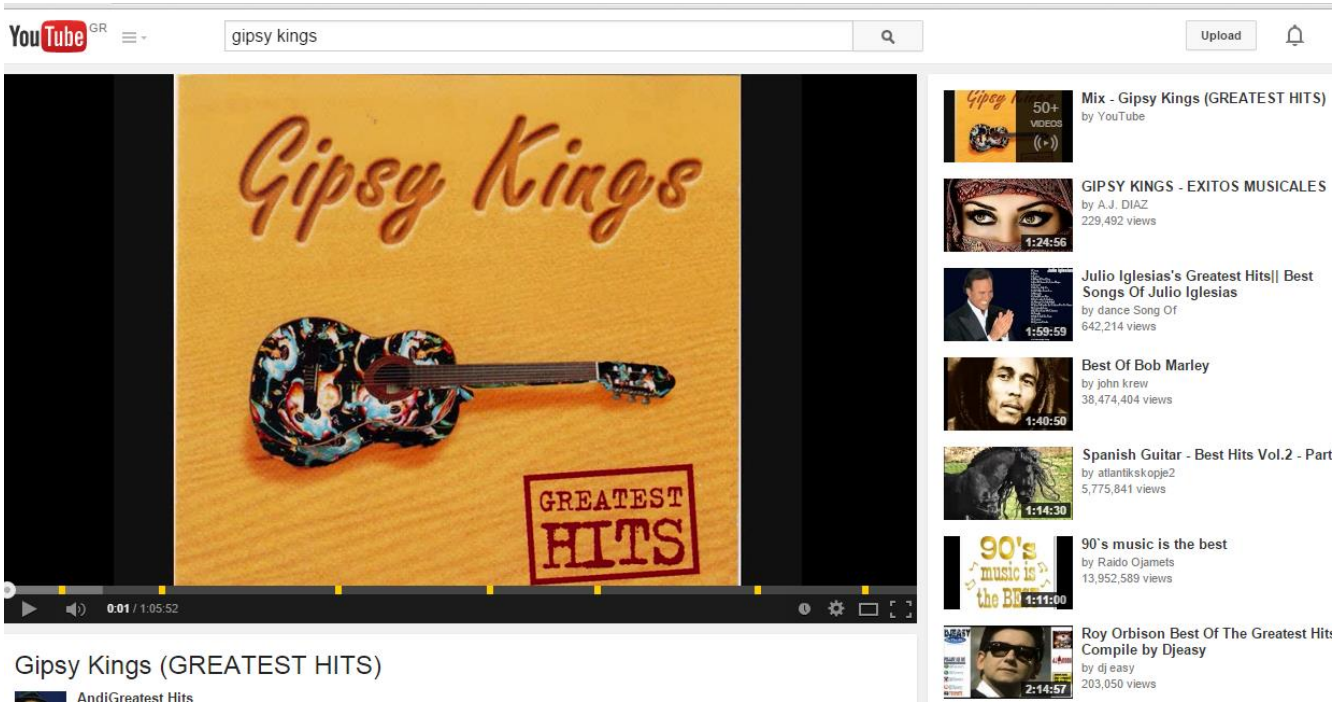
The Long Tail



Source: Chris Anderson (2004)

Other recommendation systems

- Netflix, YouTube, Pandora, etc



- Recommendation systems exploit the collective behavior of users to draw conclusions for an individual
  - Collaborative filtering

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

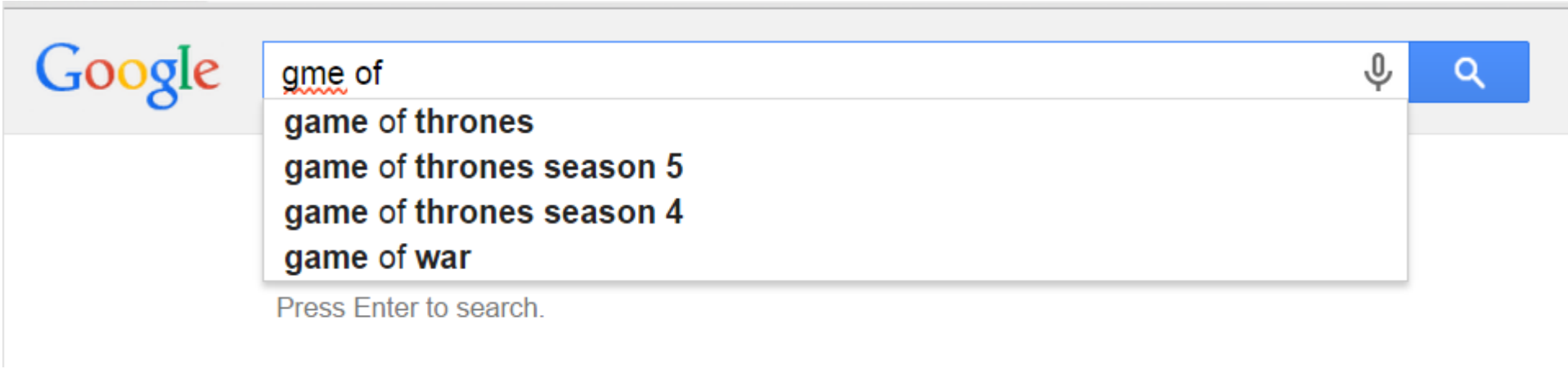
Ad click prediction

Query auto-completion and spelling correction

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

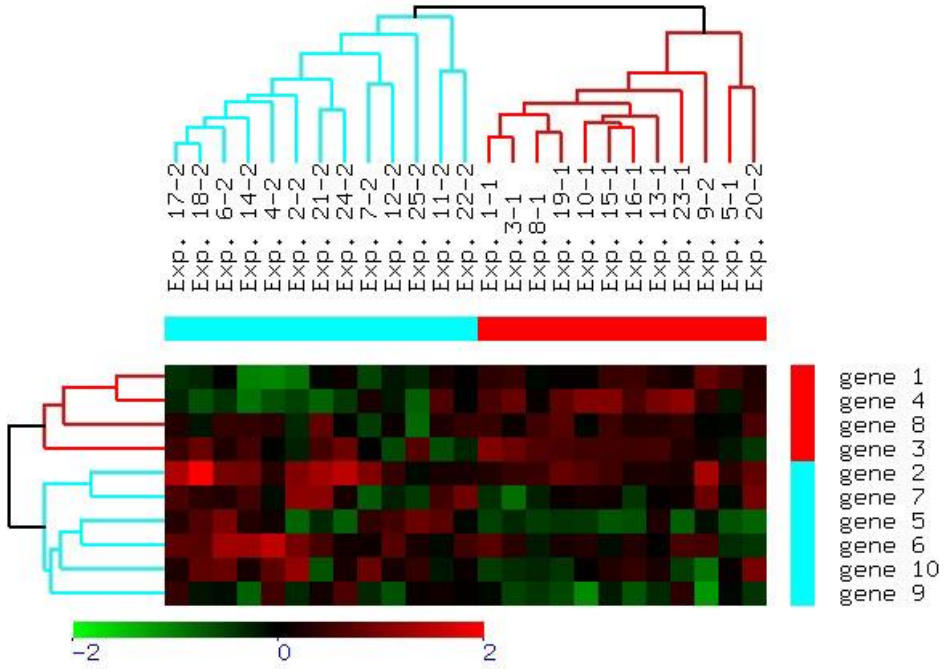
Example Application

- Google auto-complete and spelling correction



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?

Groups of similar stocks

Correlation of stocks

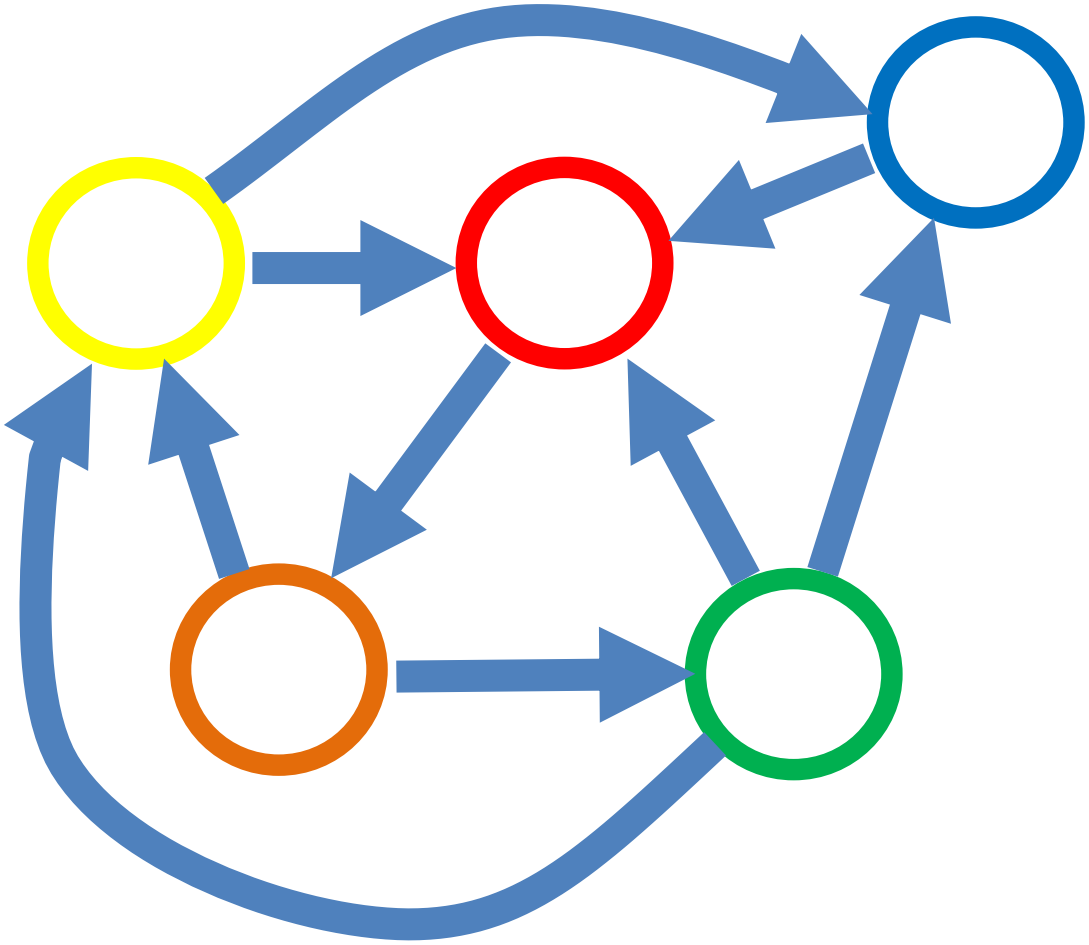
Stock Value prediction

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 important node in the graph?
- What is the shortest path between two nodes?
- How many friends two nodes have in common?
- How does information spread in the network?
- How likely are two nodes to become friends?

What is the most important node in this graph?



The Web as a graph

- Application: When retrieving pages, the authoritativeness is factored in the ranking.
  - This is the idea that made Google a success around 2000
- Today a lot more information is used, like clicks, browsing behavior, etc
  - Ranking of the pages is a very complex task that requires sophisticated techniques

Google game of thrones Search +Panayi

Web Images News Videos Books More Search tools

About 161,000,000 results (0.28 seconds)

Game of Thrones (TV Series 2011–) - IMDb
www.imdb.com/title/tt0944947/
★★★★★ Rating: 9.5/10 - 724,340 votes
The IMDB page for HBO's "**Game of Thrones**" television series, based on A Song of Ice Fire. Contains information on cast and crew.
Full Cast & Crew - Episodes - Season 4 - Emilia Clarke

Game of Thrones - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Game_of_Thrones**
Game of Thrones is an American fantasy drama television series created for HBO by David Benioff and D. B. Weiss as showrunners and main writers. It is an ...
List game thrones episodes - Season 5 - Characters - Season 1

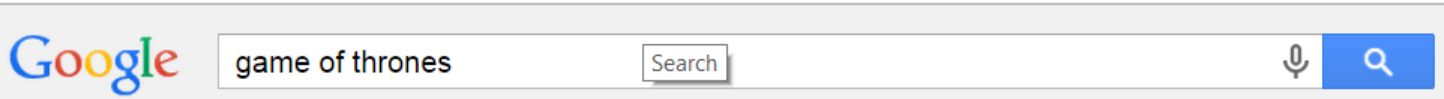
The Official Website for the HBO Series Game of Thrones ...
www.hbo.com/**game-of-thrones**
The official website for **Game of Thrones** on HBO, featuring videos, images, schedule information and episode guides.

In the news
Filming 'Game of Thrones' where winter never comes

Game of Thrones
American Television Series

★★★★★ 9.5/10 · IMDb
★★★★½ 9/10 · TV.com

George R.R. Martin's best-selling book series "A Song
brought to the screen as HBO sinks its considerable st
the medieval fantasy epic. It's the depiction of two pow
kings and queens, knights and renegades, liars and ho
playing a d... More



Web Images News Videos Books More Search tools

Page 10 of about 159,000,000 results (0.45 seconds)

Game of Thrones Show Summary and Episode Schedule ...
www.pogdesign.co.uk/cat/**Game-of-Thrones**-summary
Game of Thrones. Seven noble families fight for control of the mythical land of Westeros. Political and sexual intrigue abound. The primary families are the Stark, ...

Will Bibi's Doomsday Speech Matter? - The Daily Beast
www.thedailybeast.com/.../bibi-israel-in-deadly-**game-of-thrones**-with-ir...
2 days ago - "In this deadly **game of thrones**, there's no place for America or for Israel, no peace for Christians, Jews or Muslims who don't share the Islamist ...

Is 'Winds of Winter' finished? 'Game of Thrones' Nikolaj ...
www.zap2it.com/.../is_winds_of_winter_finished_**game_of_thrones**_nik...
6 hours ago - Nikolaj Coster-Waldau of **Game of Thrones** Is "**Game of Thrones**" fans' impatient wait for George R.R. Martin's next book, "The Winds of Winter," ...

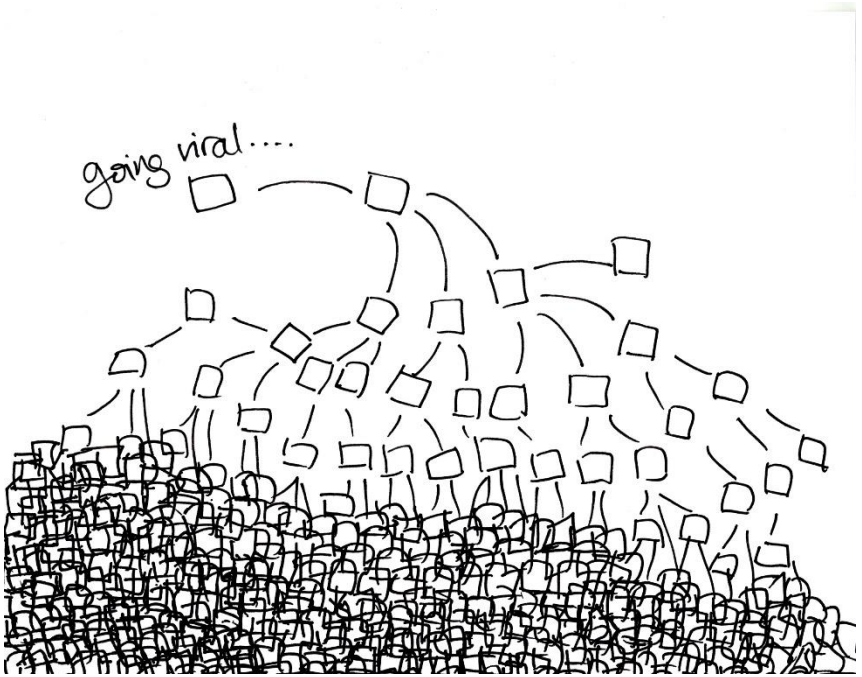
Sand Snakes or Snow Snakes? Not Everyone Is Happy With ...
www.styleite.com/.../sand-snakes-or-snow-snakes-new-**game-of-thrones**-...
2 days ago - **Game of Thrones** is getting a trio of badass new female characters next season. Obara (Keisha Castle-Hughes), Tyene (Rosabell Laurenti ...

OMG, The 'Game Of Thrones' Sand Snakes Look Amazing

Viral Marketing

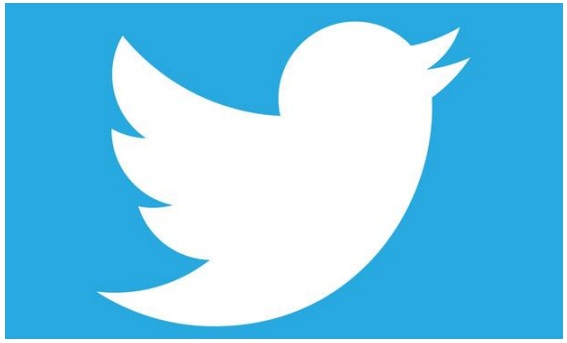
- Word-of-Mouth marketing where the network of users do the advertising themselves.
  - It is considered the most effective, but also the hardest way to advertise

To which users should we advertise in order to maximize the spread of the message?



Friendship suggestions

- LinkedIn, Twitter, Facebook friendship suggestions
  - Useful for the users to discover their friends, but also useful for the network in order to grow, and increase engagement
    - LinkedIn success story



Big data

- The new trend in data mining…
  - An all-encompassing term to describe problems in science, industry, everyday life where there are huge amounts of data that need to be stored, maintained and analyzed to produce value.
- The overall idea:
  - Every activity generates data
    - Wearable computing, Internet of Things, Brain Imaging, Urban behavior
  - If we collect and understand this data we can improve life for the individual and the world
    - E.g., Urban computing, Health informatics.
- Deep Learning:
  - New techniques that can extract useful information (learn) from massive amounts of data.

Example: Big Data and Sports

- All major soccer and basketball teams use data mining to make decisions.

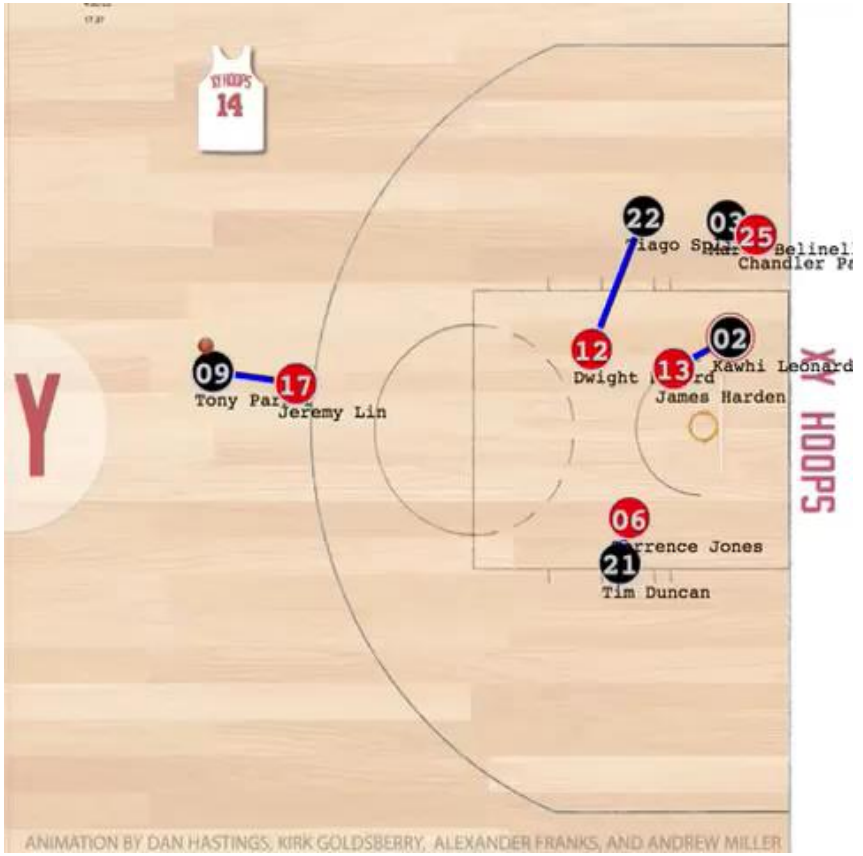
The national team of Germany had a special software for the analysis of video.

They concluded that the possession time per player should be reduced.

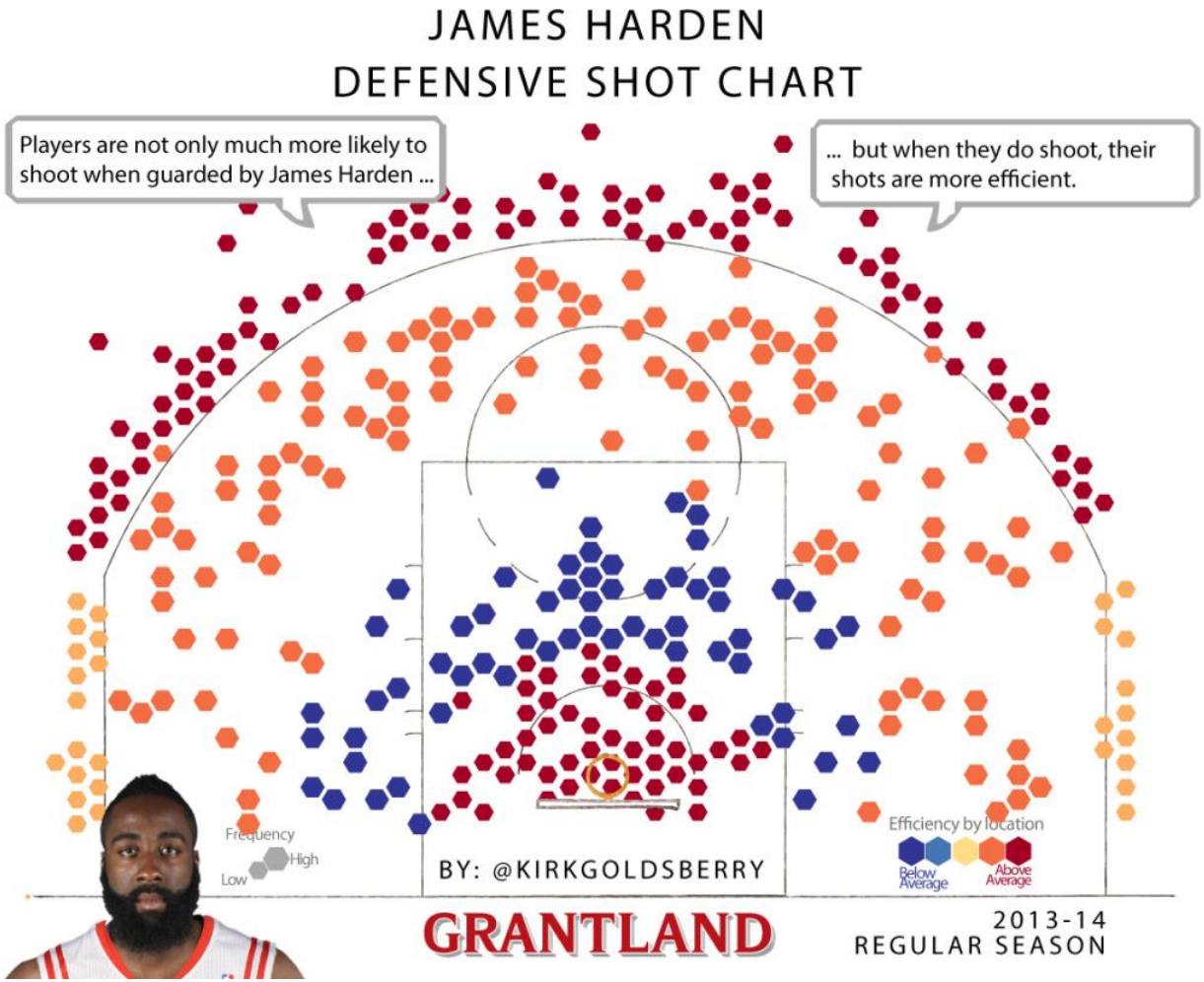
Germany won the 2014 word cup

NBA and Data Mining

In NBA there are special conferences for data science



James Harden defence



From BMI to TMI: The NBA Is Leaning Toward Wearable Tech

NBA

SEPTEMBER 17, 2015

by ZACH LOWE

The NBA is putting its own money into the study of wearable GPS devices, with the likely end goal of outfitting players during games, according to several league sources. The league is funding a study, at the Mayo Clinic in Minnesota, of products from two leading device-makers: Catapult and STATSports.

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 to get a better understanding of the world and advance science and help people
- Scale (in data size and feature dimension)
  - Why not use traditional analytic methods?
  - Enormity of data, curse of dimensionality
  - 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 is data mining again?

- The industry point of view: The analysis of huge amounts of data for extracting useful and actionable information, which is then integrated into production systems in the form of new features of products
  - Data Scientists should be good at data analysis, math, statistics, but also be able to code with huge amounts of data and use the extracted information to build products.

What can we do with data mining?

- Some Data Mining topics:
  - Frequent itemsets and Association Rules extraction
  - Recommendation systems
  - 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.
- Challenge: Do this efficiently for millions of records and 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}

Example 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!

Frequent Itemsets: Applications

- Text mining: finding associated phrases in text
  - There are lots of documents that contain the phrases "association rules", "data mining" and "efficient algorithm"
  - Can be used to define key phrases, correct spelling mistakes, associate different concepts.
- Recommendations:
  - Users who buy this item often buy this item as well
  - Users who watched James Bond movies, also watched Jason Bourne movies.

  - Recommendations make use of item and user similarity

Recommender systems

Collaborative filtering: Use the collective behavior of the users to draw conclusions for an individual

	Harry Potter 1	Harry Potter 2	Harry Potter 3	Twilight	Star Wars 1	Star Wars 2	Star Wars 3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Fill the empty entries of the matrix.
Use the fact that similar users will behave similarly
And similar items will be rated similarly
But how do we define similarity?
How do we make use of the collective behavior?
Big problem – Complicated math using probabilities/linear algerba

Clustering Definition

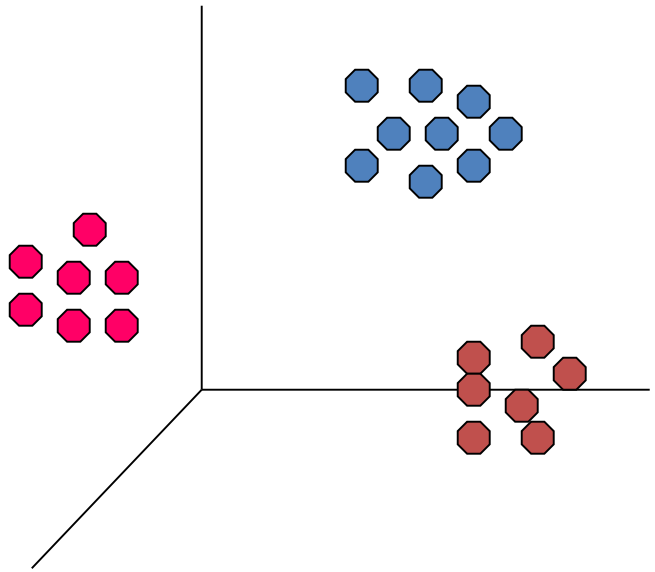
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures?
  - Euclidean Distance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

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

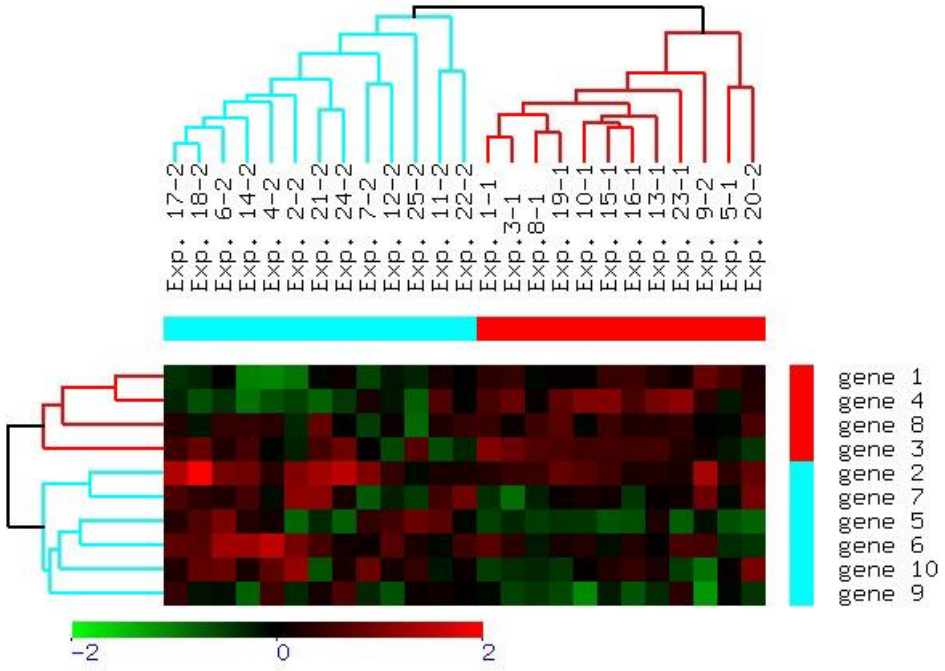
Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- Bioinformatics applications:
  - Goal: Group genes and tissues together such that genes are coexpressed on the same tissues



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.

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.

- 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.

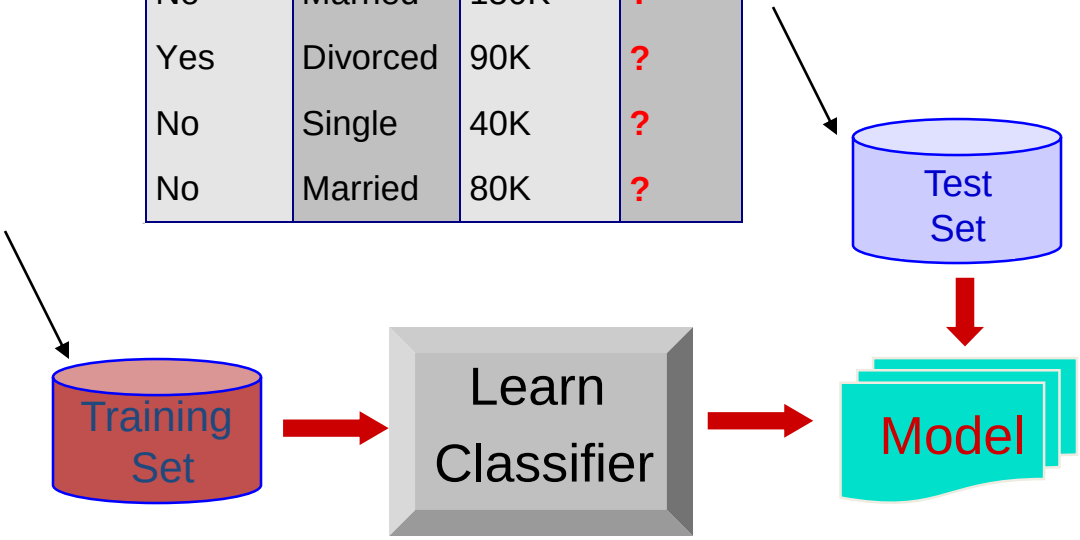
- Goal: previously unseen records should be assigned a class as accurately as possible.

- In simple terms: Create a model that predicts a specific property of the data

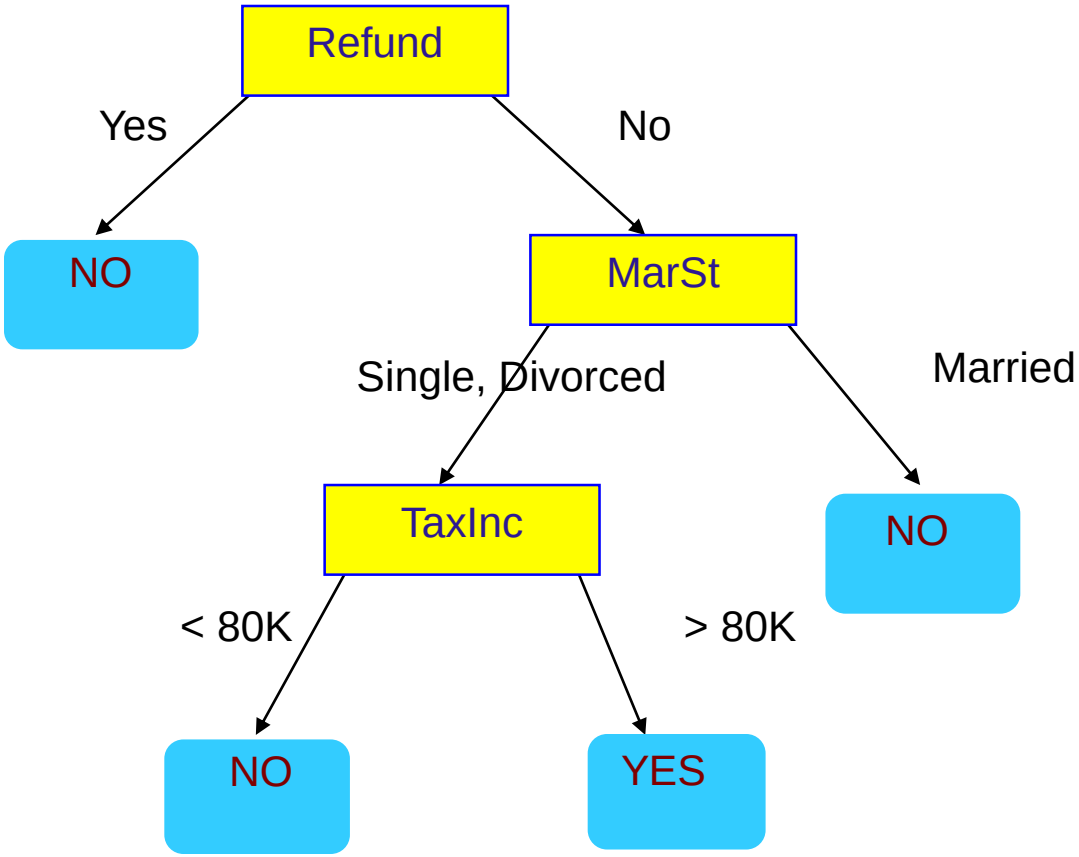
Classification Example: Tax Fraud

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
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	?



Model Example: Decision Trees



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.

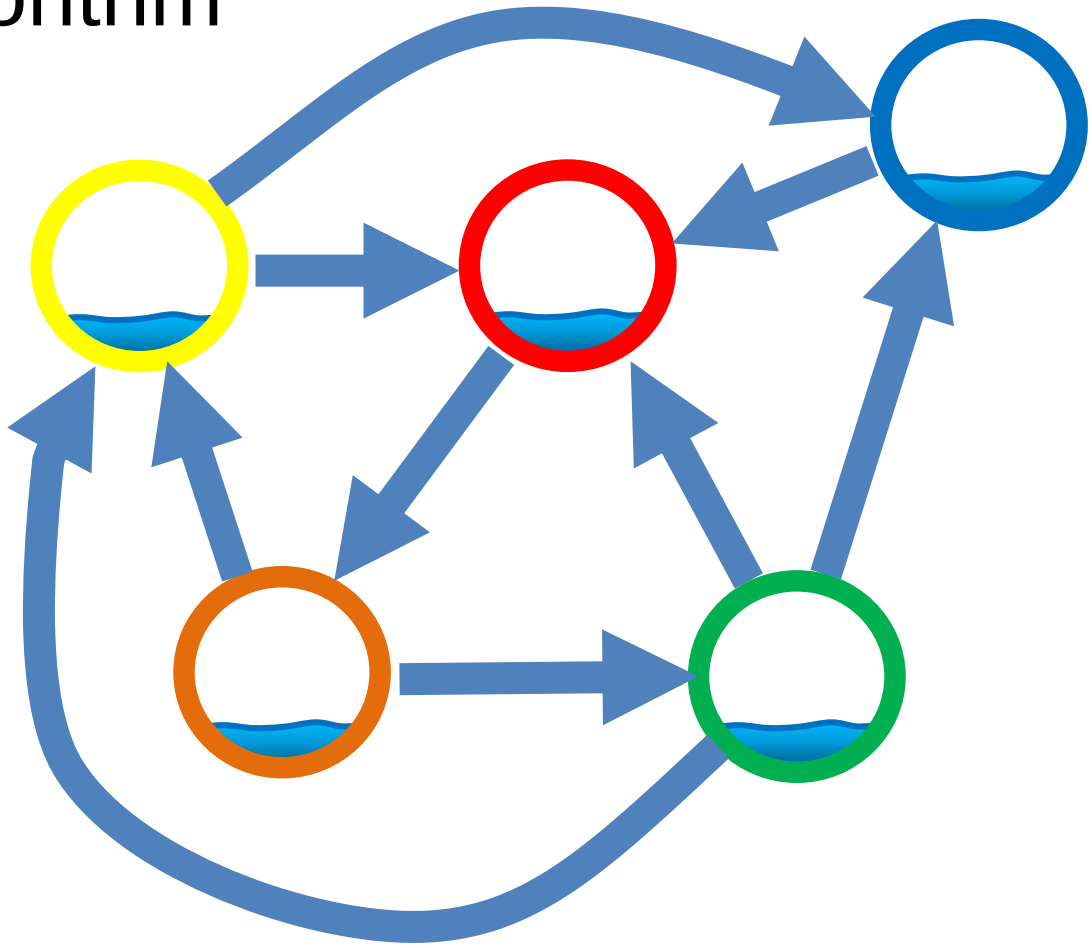
Network data analysis

- 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 authoritative page.

The PageRank algorithm

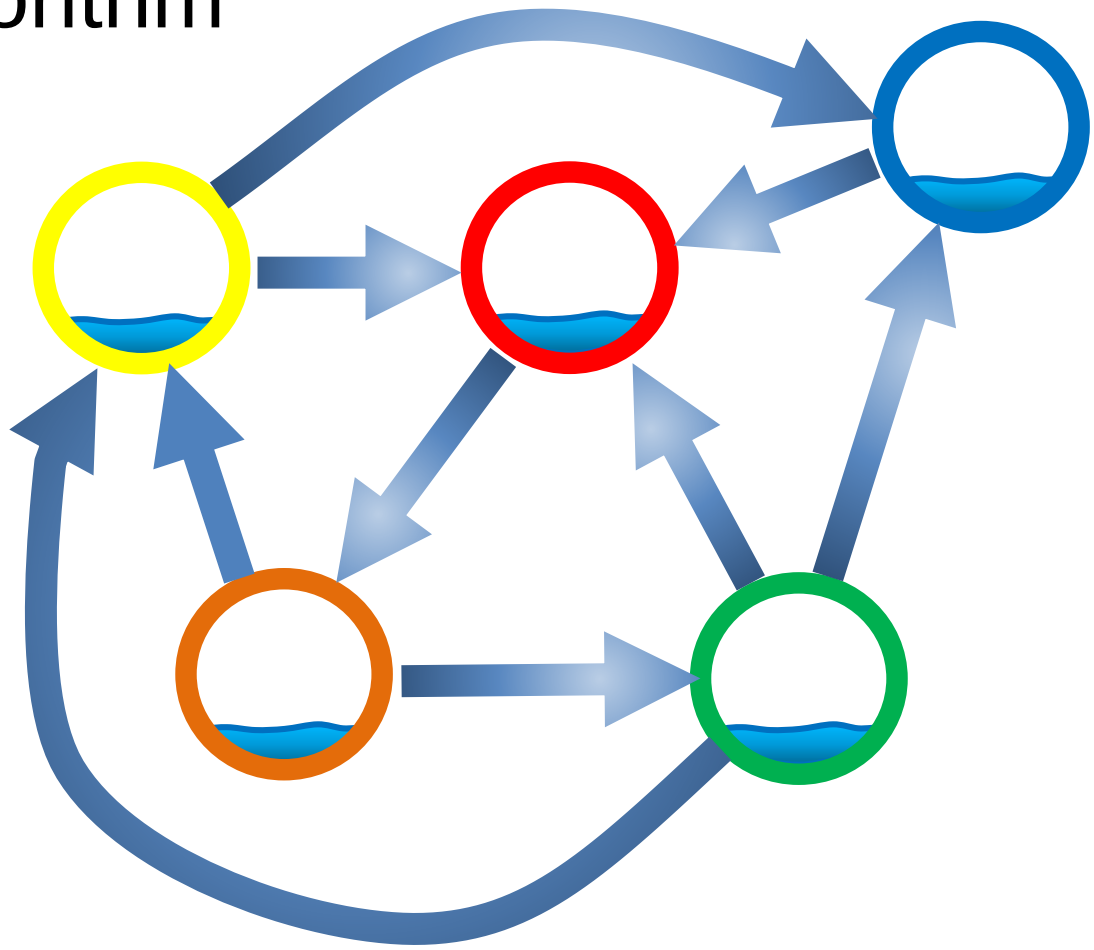
Think of the nodes in the graph as containers of capacity of 1 liter.

We distribute a liter of liquid equally to all containers



The PageRank algorithm

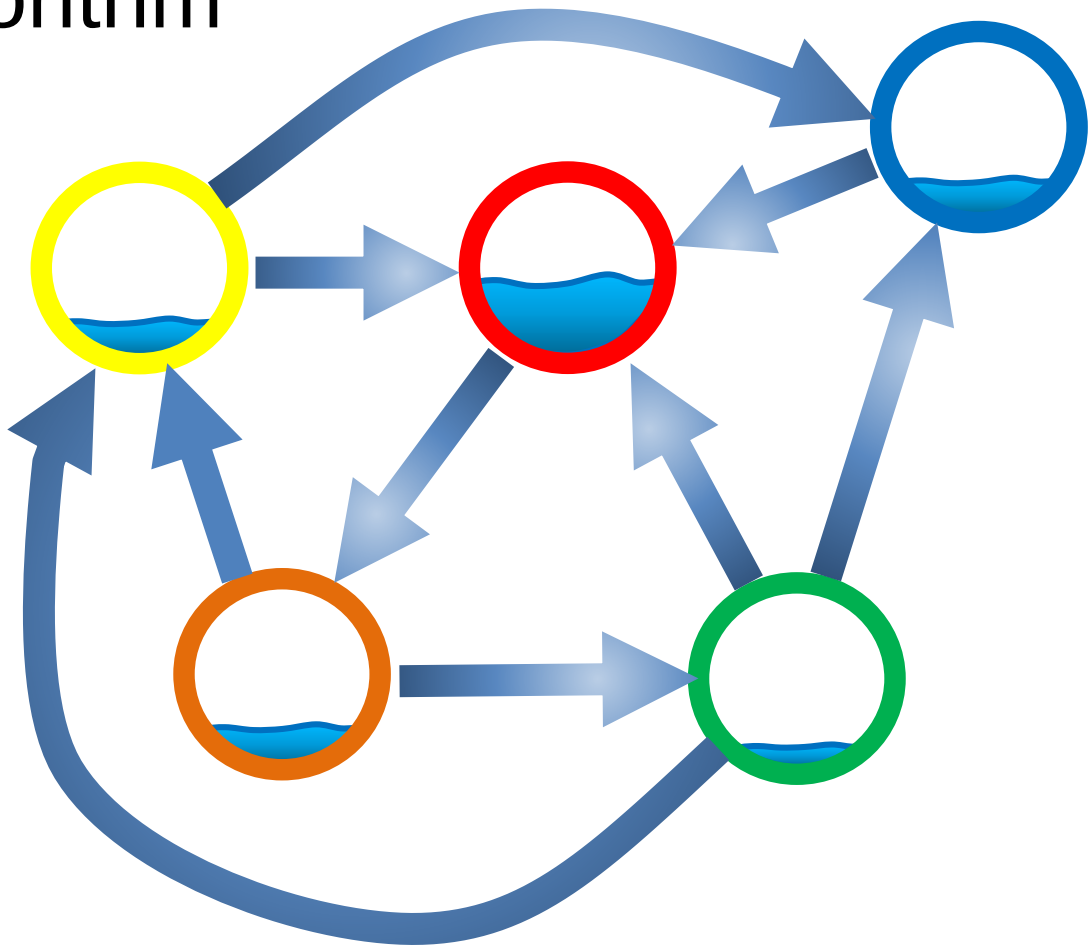
The edges act like pipes that transfer liquid between nodes.



The PageRank algorithm

The edges act like pipes that transfer liquid between nodes.

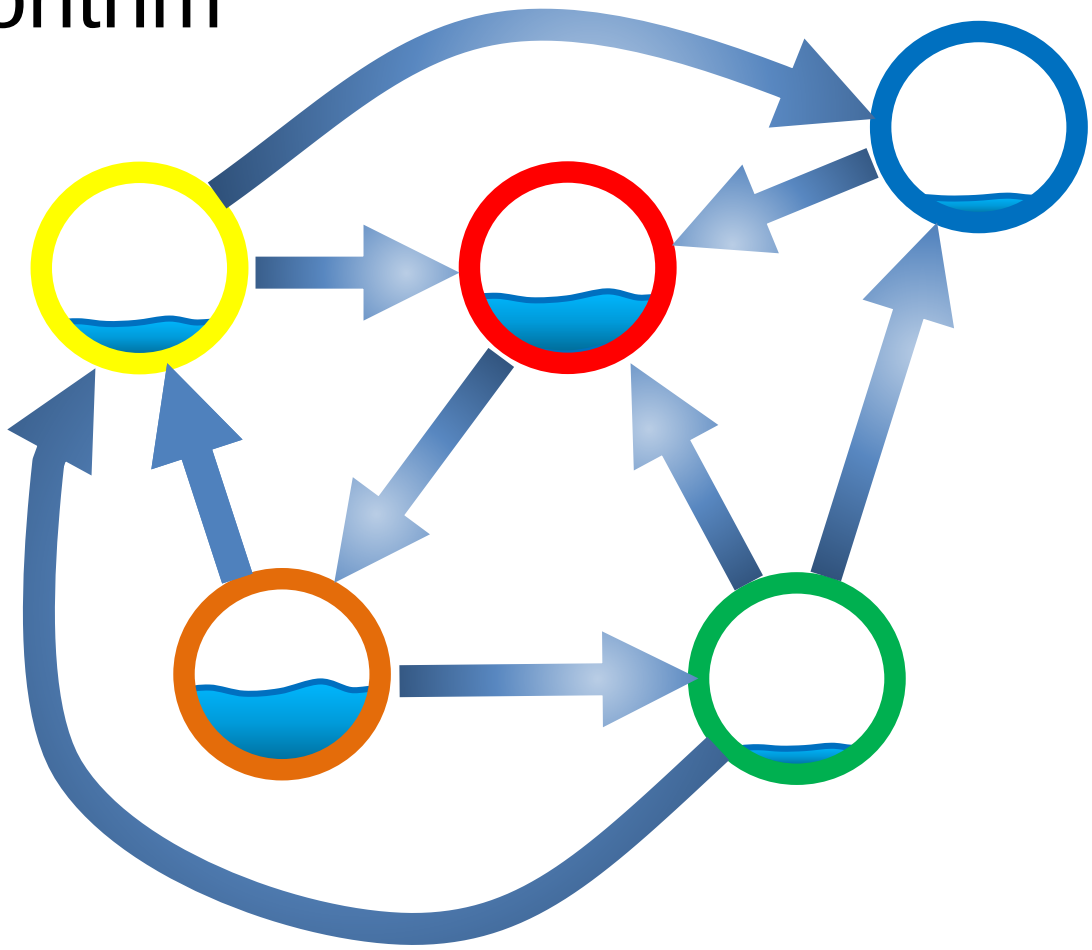
The contents of each node are distributed to its neighbors.



The PageRank algorithm

The edges act like pipes that transfer liquid between nodes.

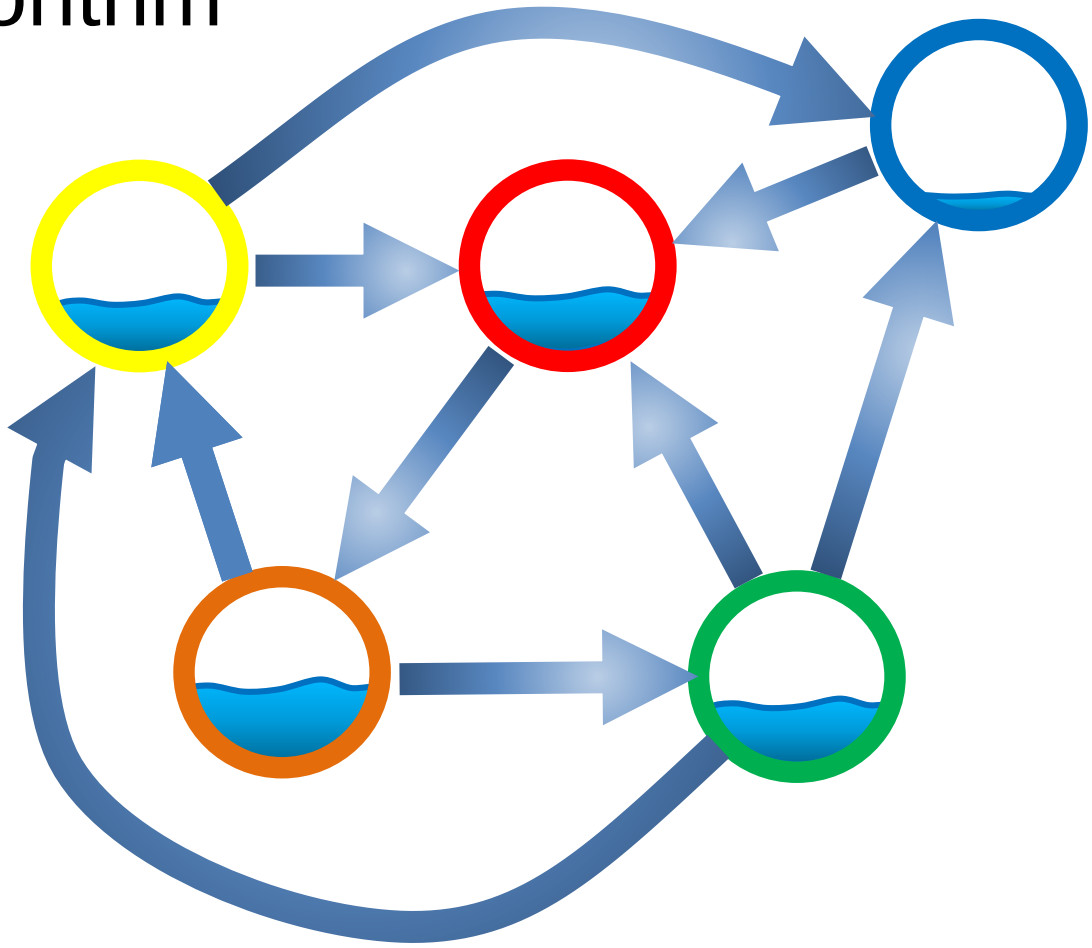
The contents of each node are distributed to its neighbors.



The PageRank algorithm

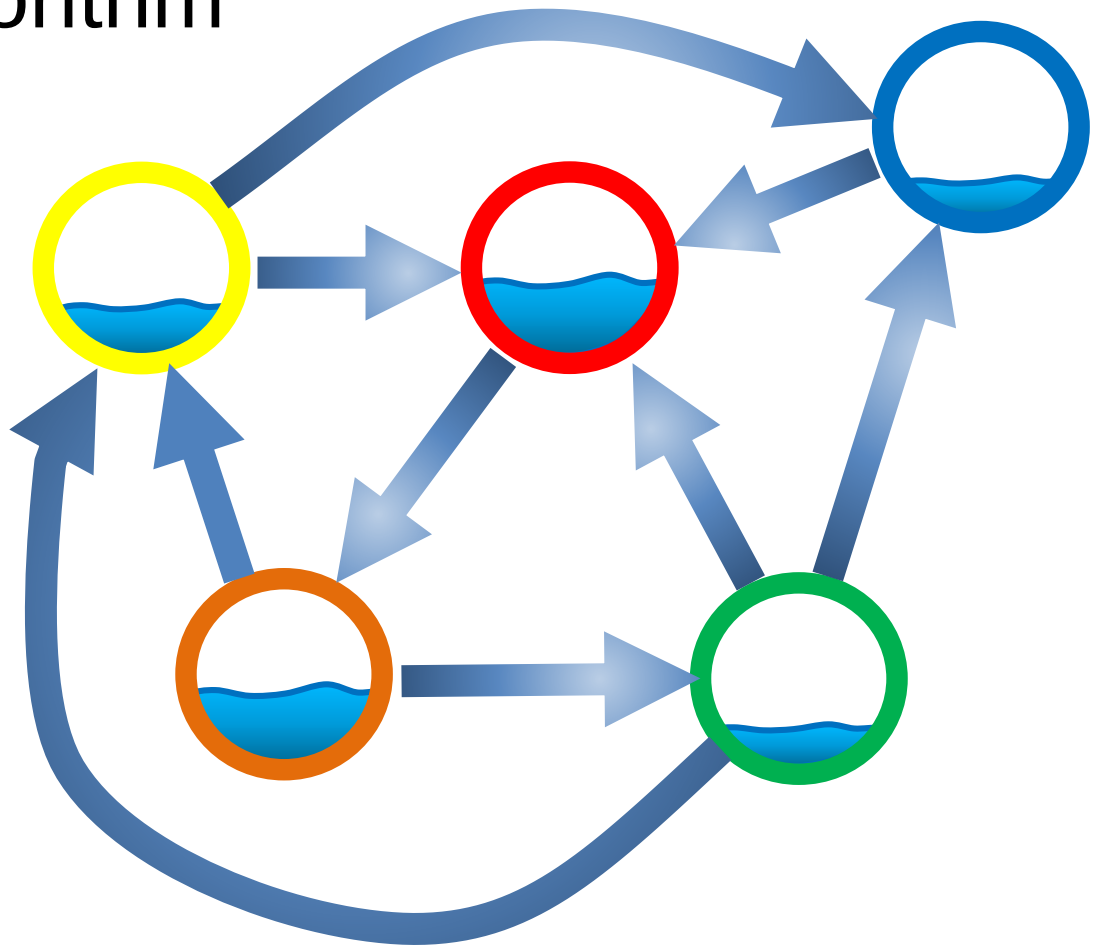
The edges act like pipes that transfer liquid between nodes.

The contents of each node are distributed to its neighbors.



The PageRank algorithm

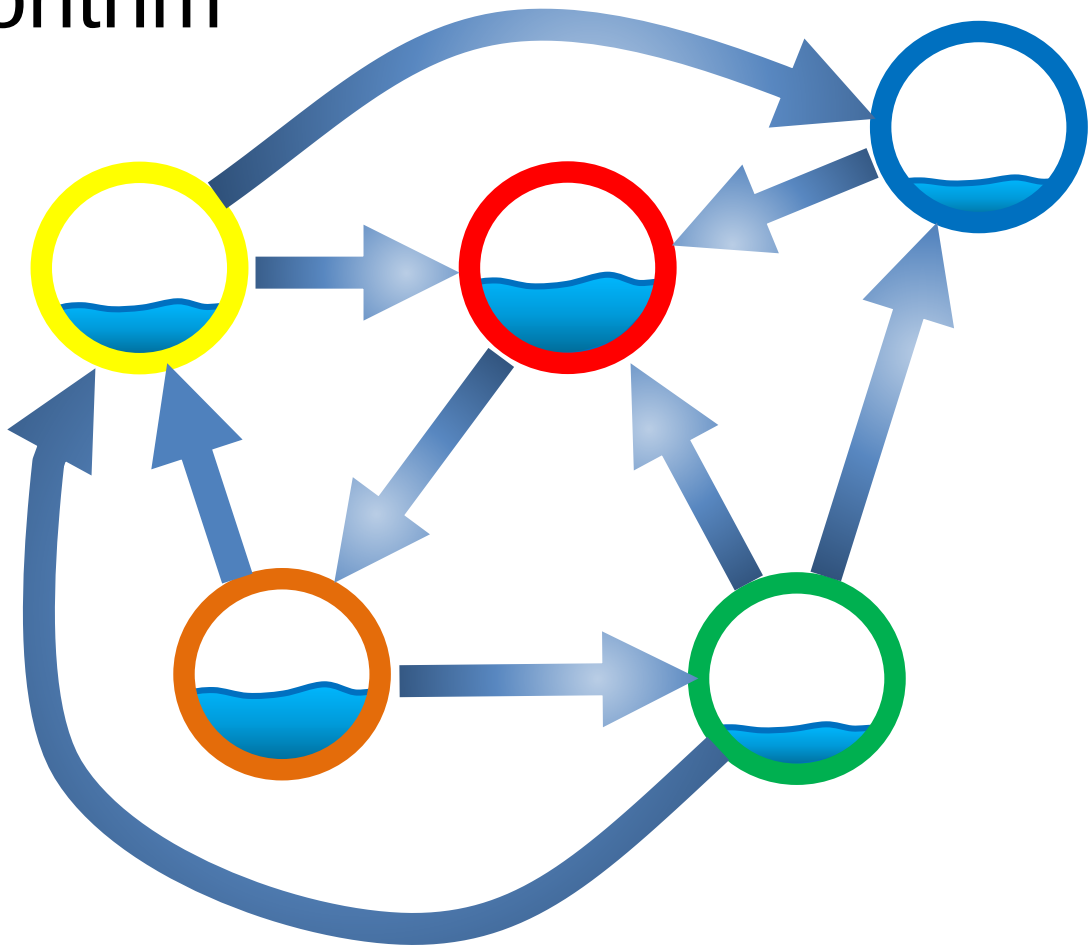
The system will reach an equilibrium state where the amount of liquid in each node remains constant.



The PageRank algorithm

The amount of liquid in each node determines the importance of the node.

Large quantity means large incoming flow from nodes with large quantity of liquid.



Mathematically, we compute an eigenvector of a matrix defined by the adjacency matrix of the graph

Network data analysis

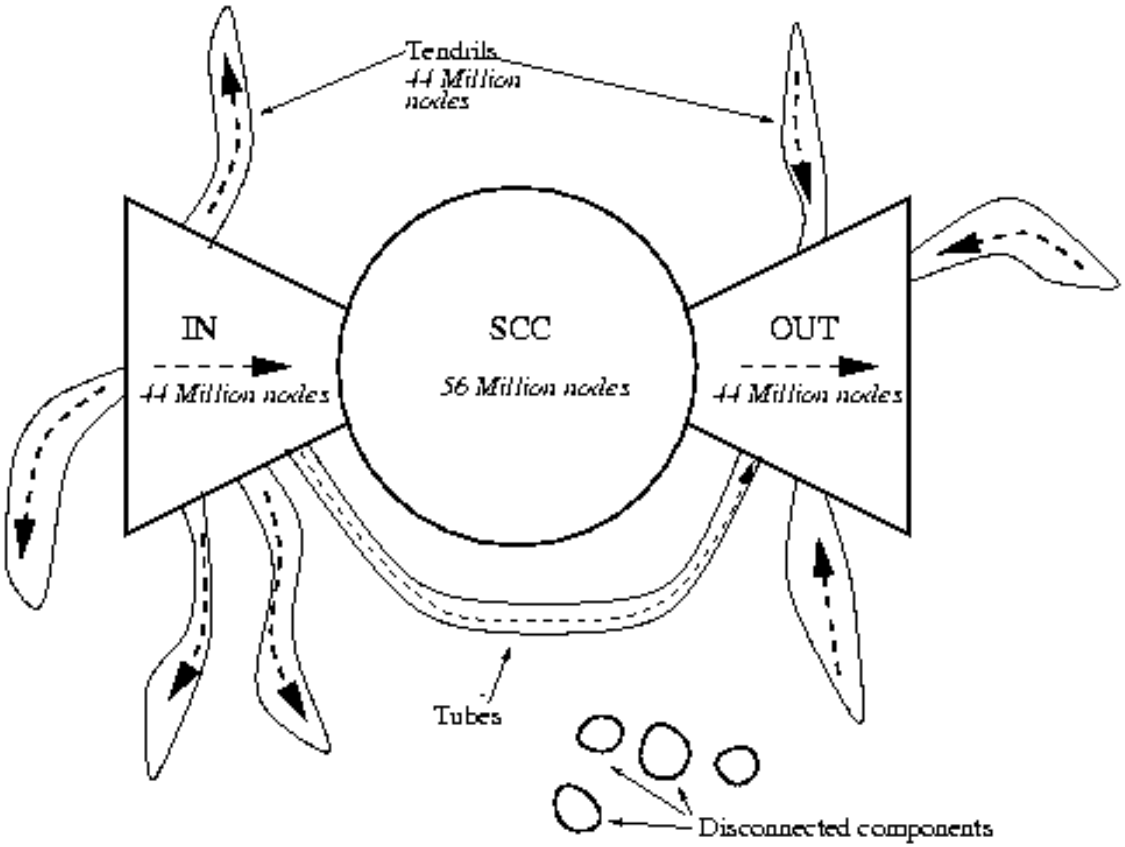
- Given a social network can you predict which individuals will connect in the future?
  - Triadic closure principle: Links are created in a way that usually closes a triangle
    - If both Bob and Charlie know Alice, then they are likely to meet at some point.
- Application: Friend/Connection suggestions in social networks

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.

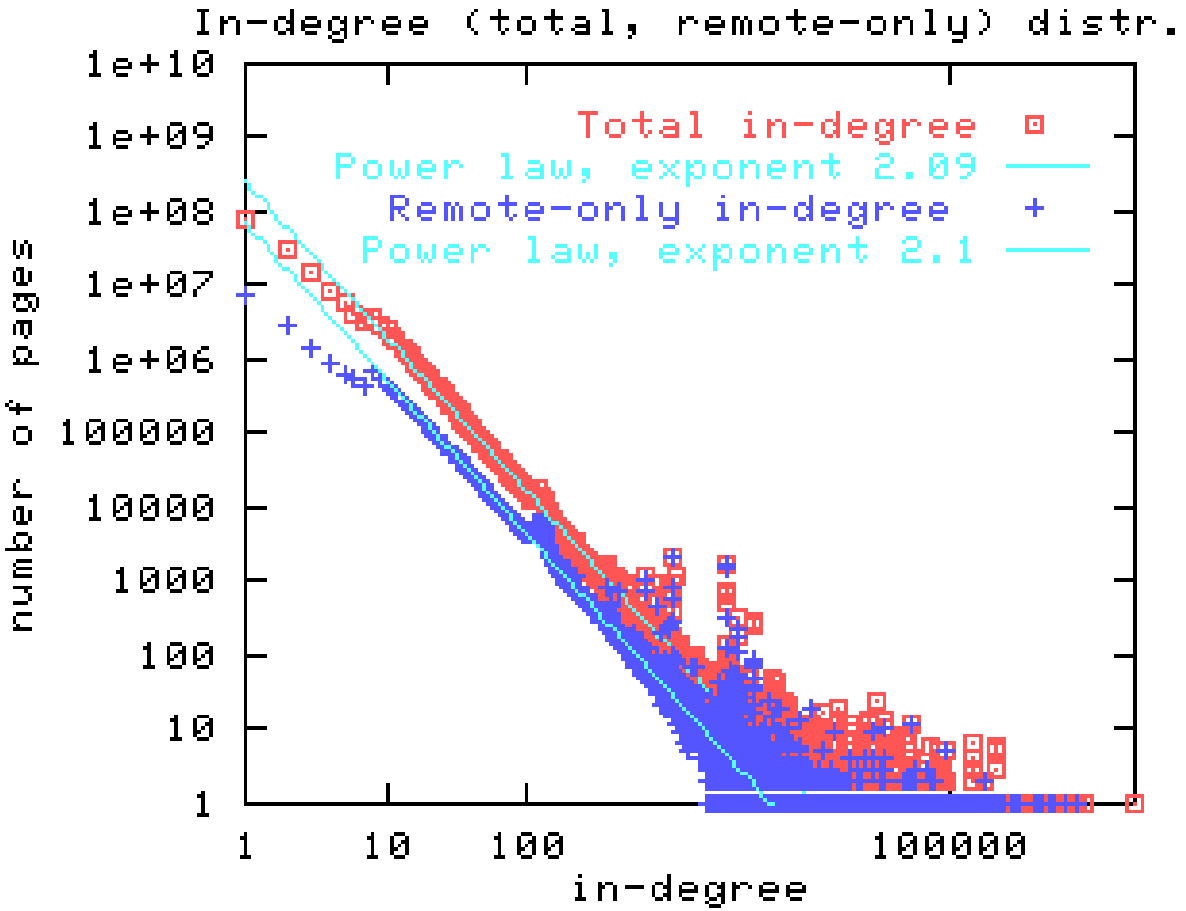
Exploratory Analysis: The Web

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



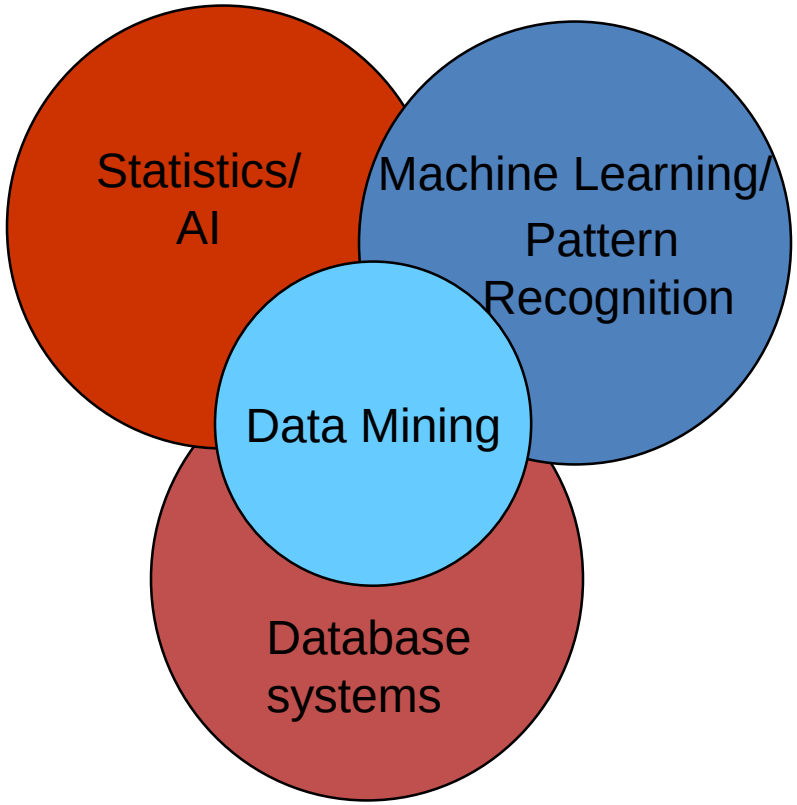
Exploratory Analysis: The Web

- 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



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.

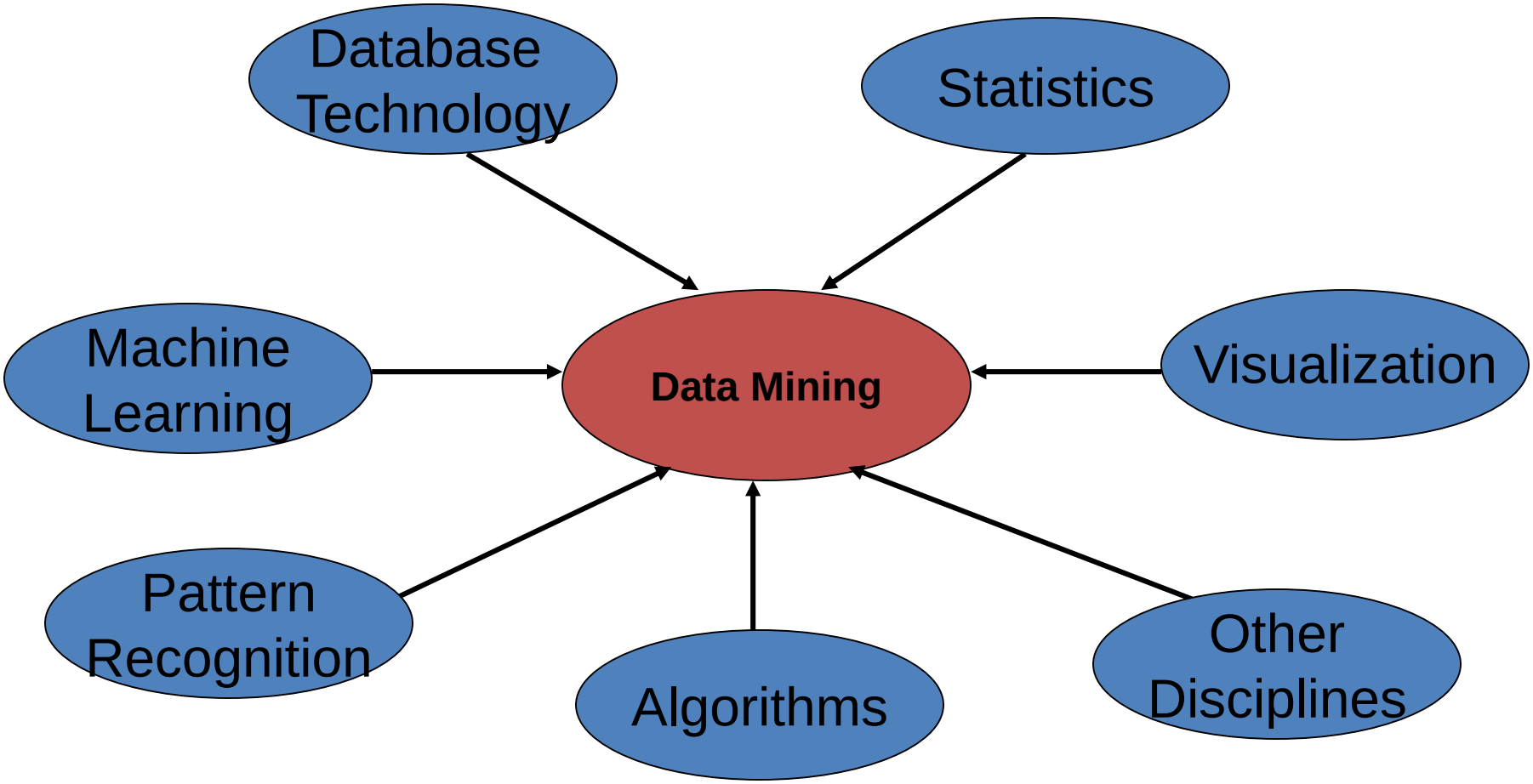
Models vs. Analytic Processing

- 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.

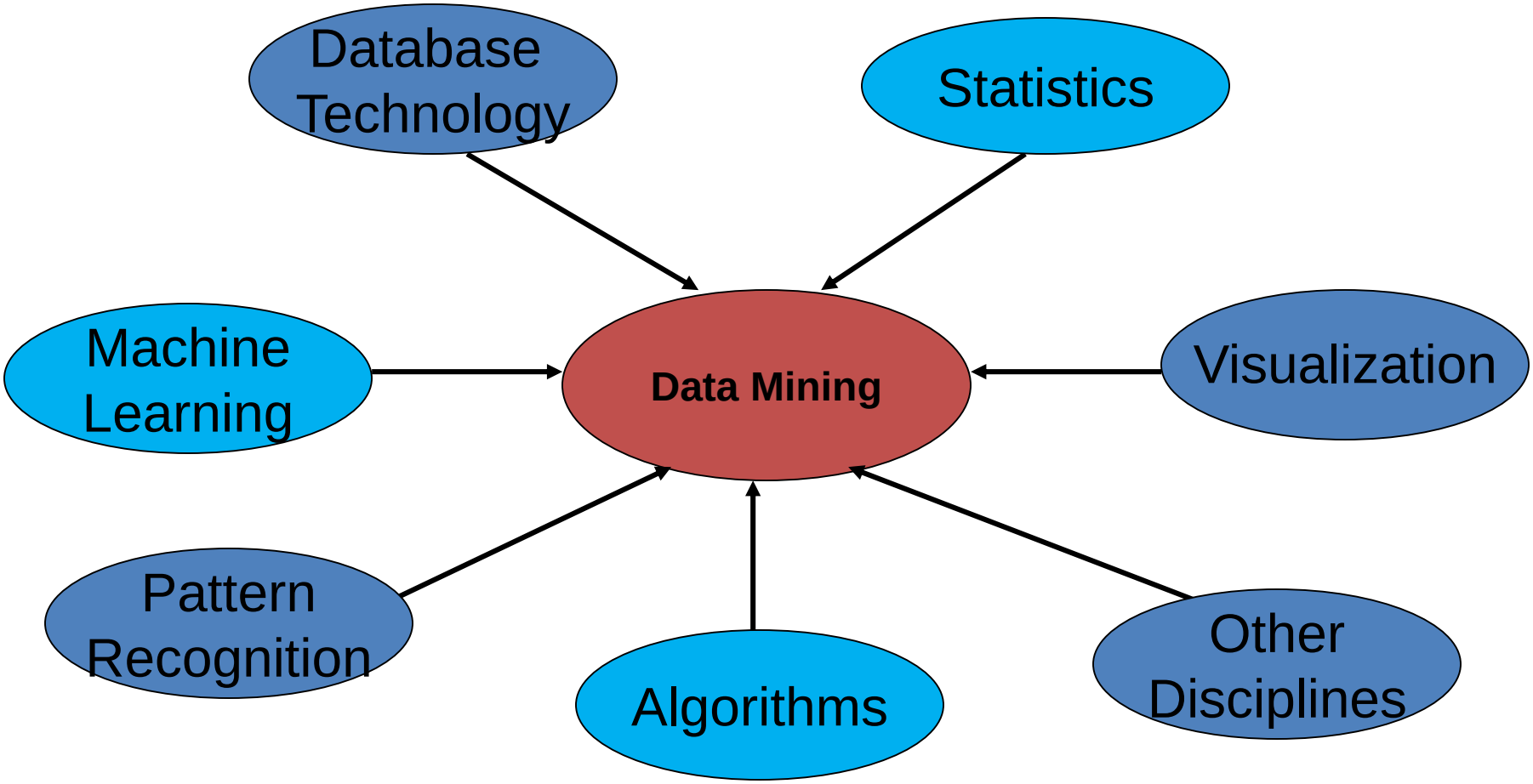
New era of data mining

- Boundaries are becoming less clear
  - Today data mining and machine learning are synonymous. It is assumed that there algorithms should scale. It is clear that statistical inference is used for building the models.

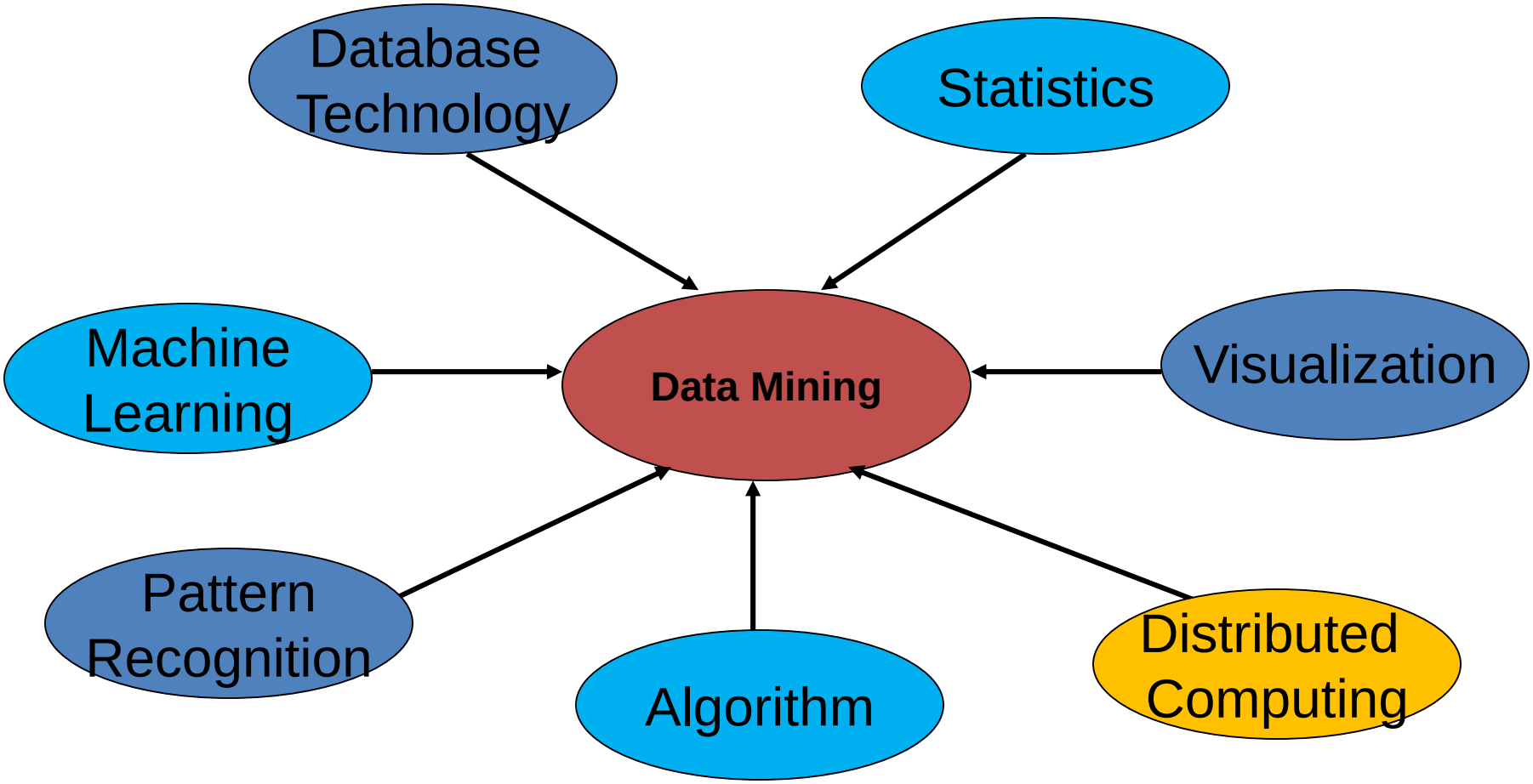
Data Mining: Confluence of Multiple Disciplines



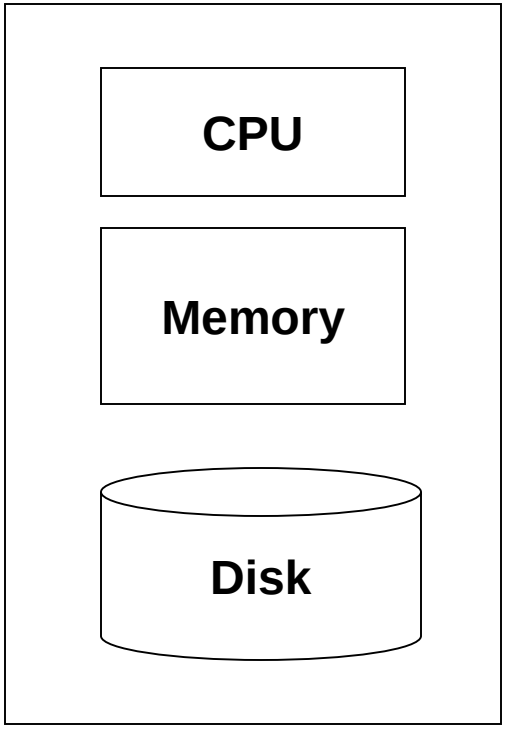
Data Mining: Confluence of Multiple Disciplines



Data Mining: Confluence of Multiple Disciplines



Single-node architecture



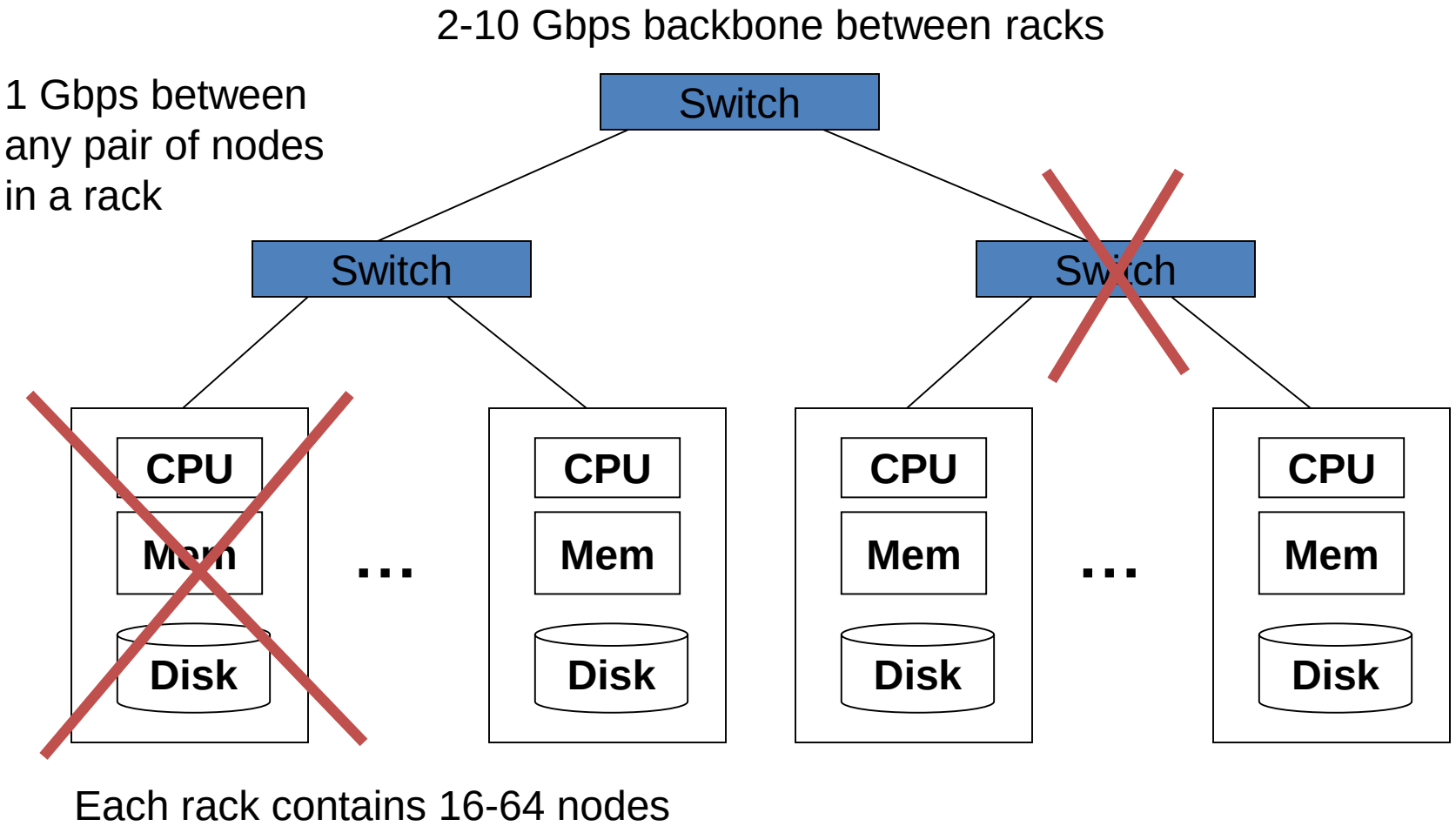
Machine Learning, Statistics

“Classical” Data Mining

Commodity Clusters

- Web data sets can be very large
  - Tens to hundreds of terabytes
  - Cannot mine on a single server
- Standard architecture emerging:
  - Cluster of commodity Linux nodes, Gigabit ethernet interconnect
  - Google GFS; Hadoop HDFS; Kosmix KFS
- Typical usage pattern
  - Huge files (100s of GB to TB)
  - Data is rarely updated in place
  - Reads and appends are common
- How to organize computations on this architecture?
  - Map-Reduce paradigm

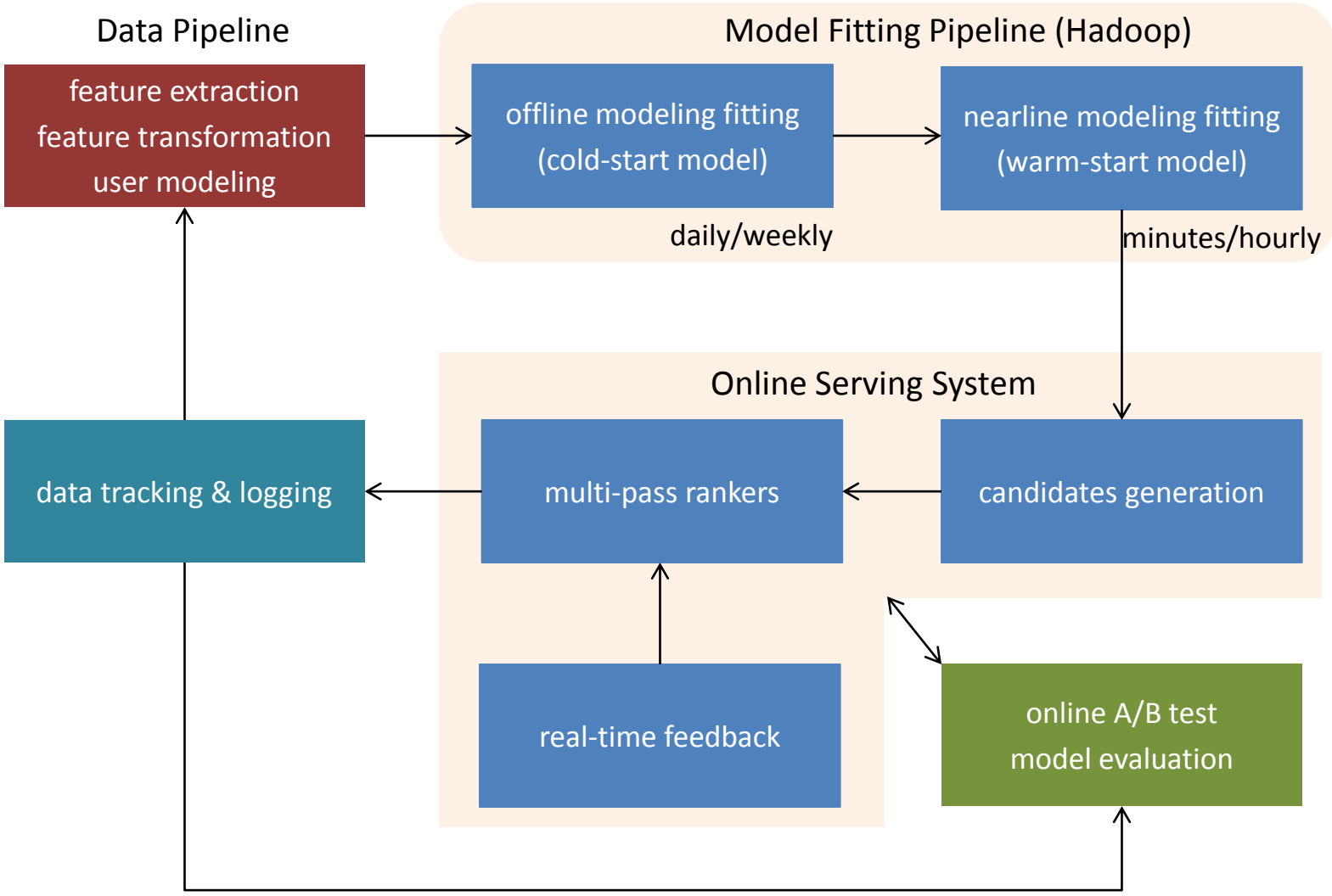
Cluster Architecture



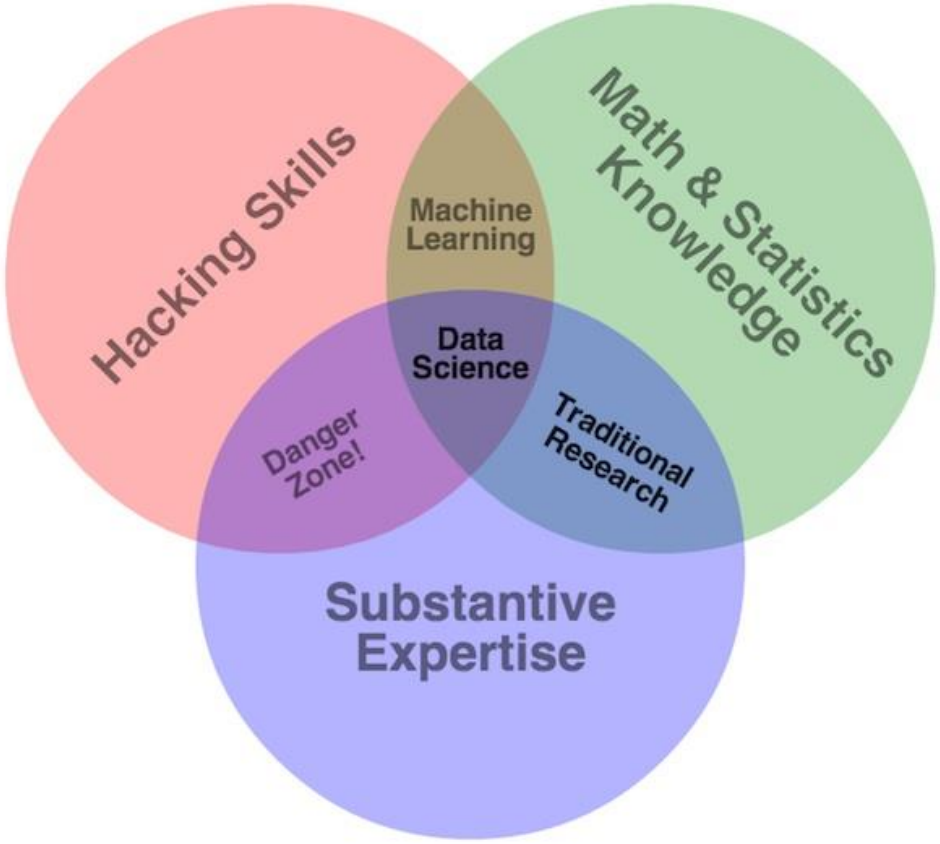
Map-Reduce paradigm

- Map the data into key-value pairs
  - E.g., map a document to word-count pairs
- Group by key
  - Group all pairs of the same word, with lists of counts
- Reduce by aggregating
  - E.g. sum all the counts to produce the total count.

Putting it all together: The LinkedIn Data Mining Pipeline



The Skills of a Data Miner – Data Scientist



It is a hard job

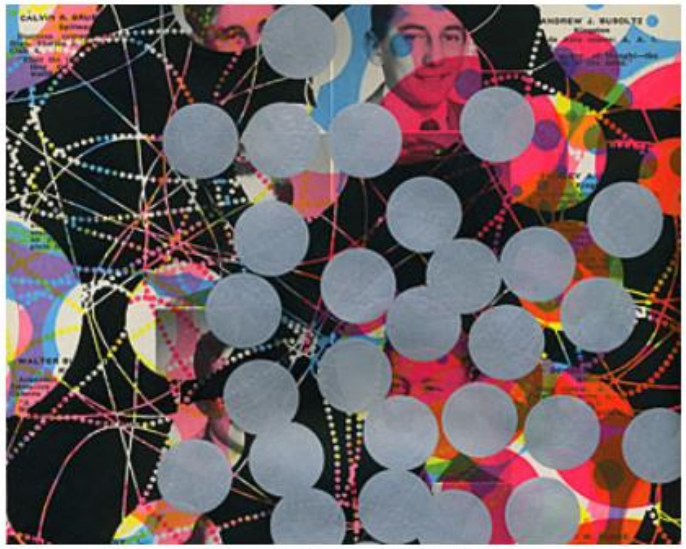
But also a rewarding one

"The success of companies like Google, Facebook, Amazon, and Netflix, not to mention Wall Street firms and industries from manufacturing and retail to healthcare, is increasingly driven by better tools for extracting meaning from very large quantities of data. 'Data Scientist' is now the hottest job title in Silicon Valley." – Tim O'Reilly

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (87)



RELATED

Executive Summary

ALSO AVAILABLE

- Buy PDF

Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high school