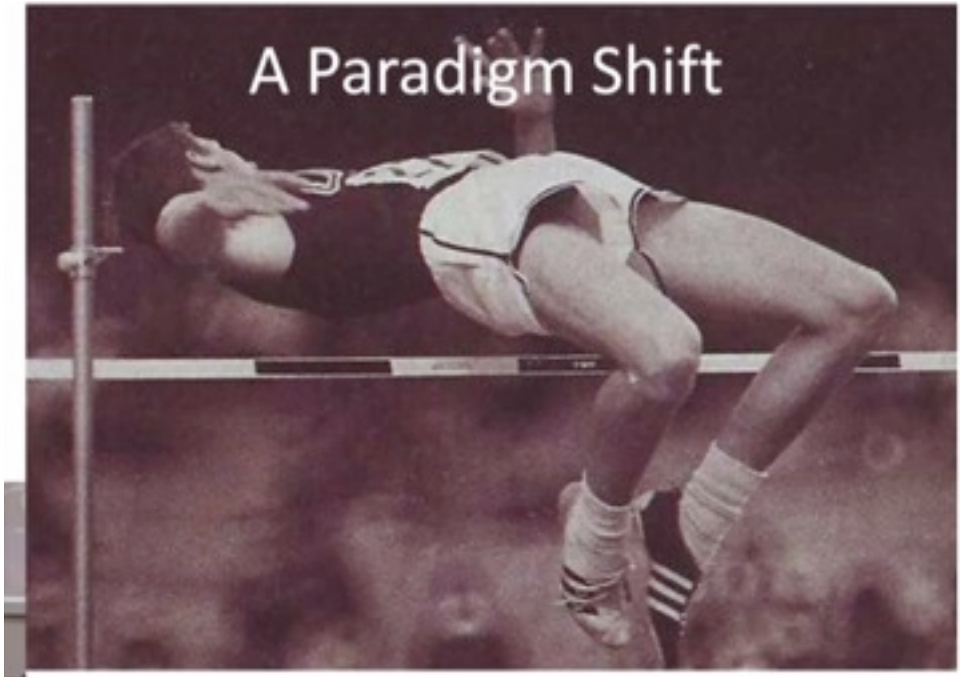
# SD201 Large-Scale Data Mining

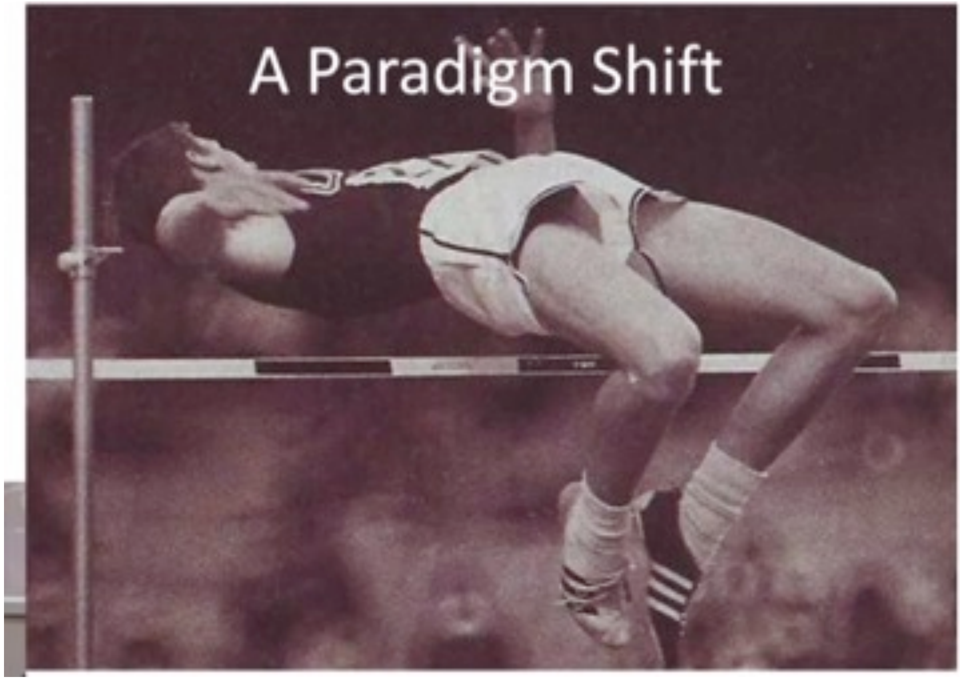
Mauro Sozio

sozio@telecom-paristech.fr

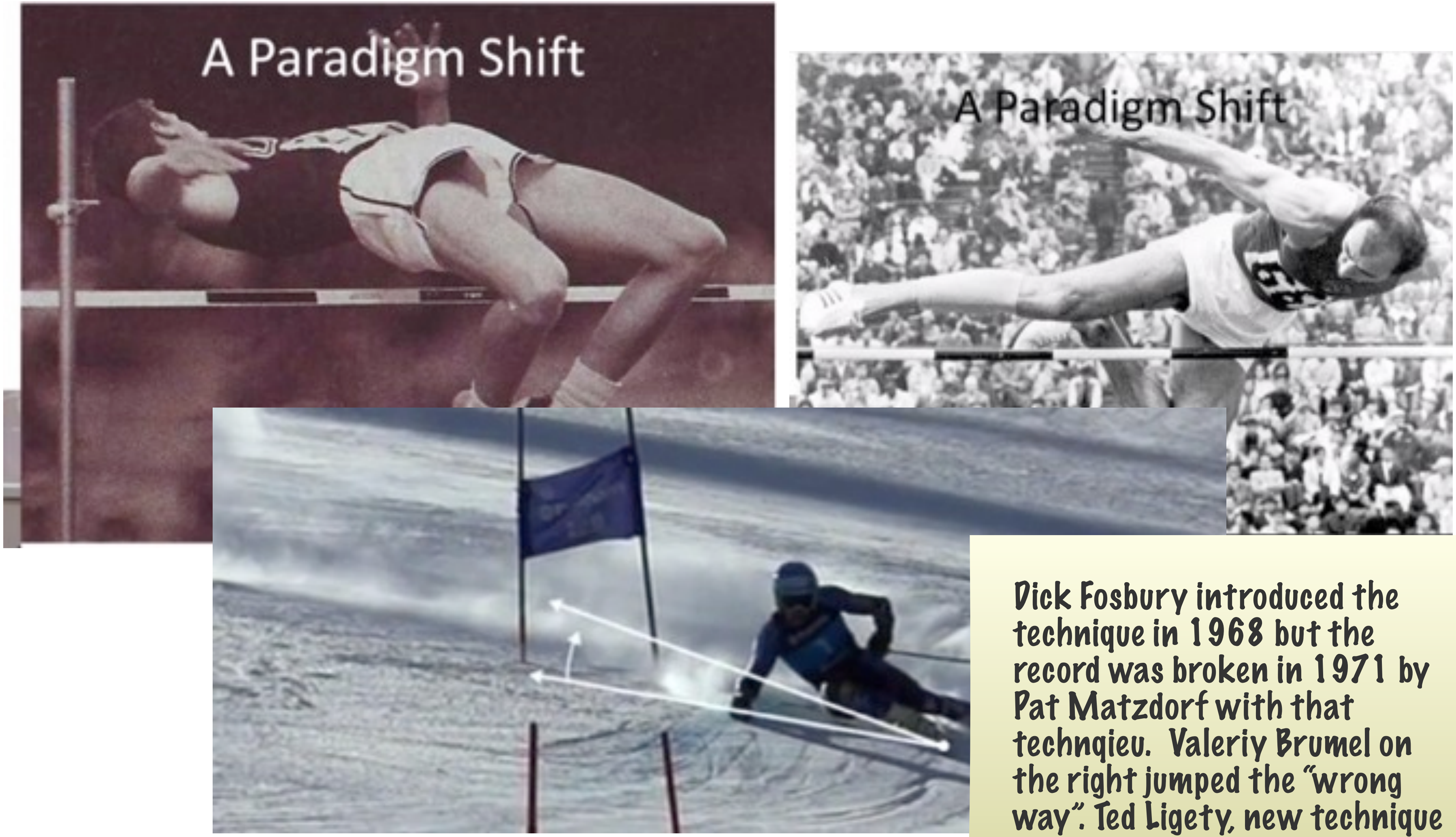
# BigData: A Paradigm Shift



# BigData: A Paradigm Shift



# BigData: A Paradigm Shift



Dick Fosbury introduced the technique in 1968 but the record was broken in 1971 by Pat Matzdorf with that technqieu. Valeriy Brumel on the right jumped the "wrong way". Ted Ligety, new technique with small angle in ski.

# Small-world experiment and six degree of separation

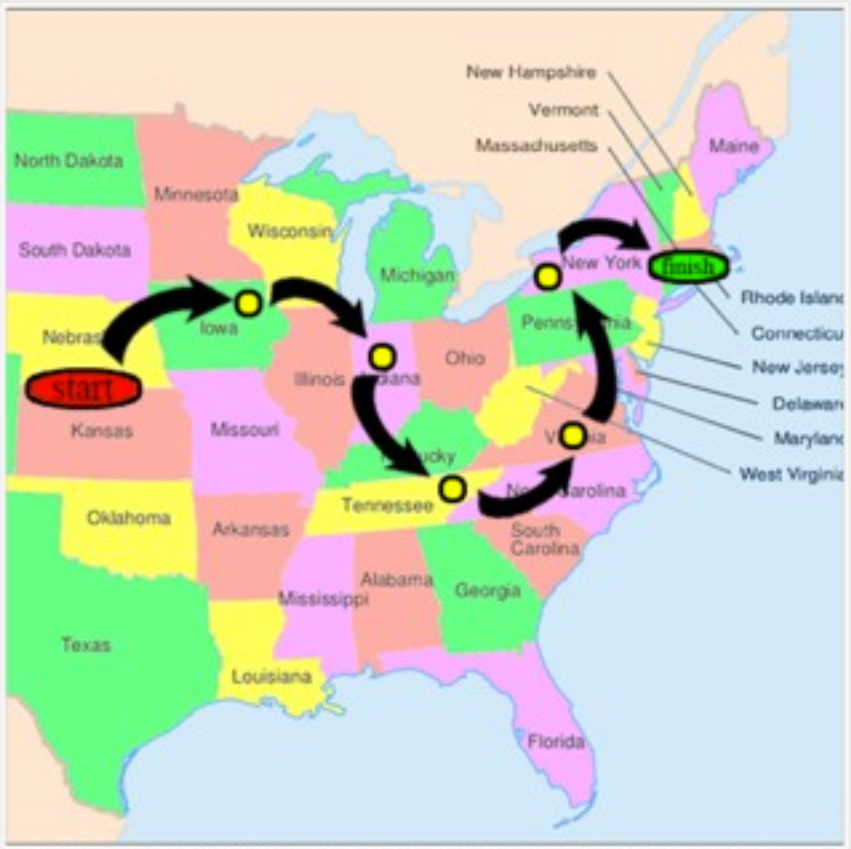
- Study conducted by Stanley Milgram in 1969.

- Questions:
  - What is the probability that two random people in the world know each other?
  - How many hops between them? (e.g. friend of friend of friend = 3 hops.)

# Small-world experiment and six degree of separation

- Experiment:
  - Random people from Nebraska, Kansas,..., were sent a letter with the goal of forwarding it to a random person in Boston.
  - If the person knew that person then he/she could send him/her the letter directly.
  - Otherwise she could forward the letter to a relative or a friend who might know the person.
  - Some basic information about the target person were included.

# Small-world experiment and six degree of separation



One possible path of a message in the "Small World" experiment by Stanley Milgram.

Results:

- only 64 out of 296 letters reach the destination (some people refused to participate)
- among those reaching the destination, the average number of hops was ~5-6.

-> **six degree of separation**

# Six degree of separation in the BigData era

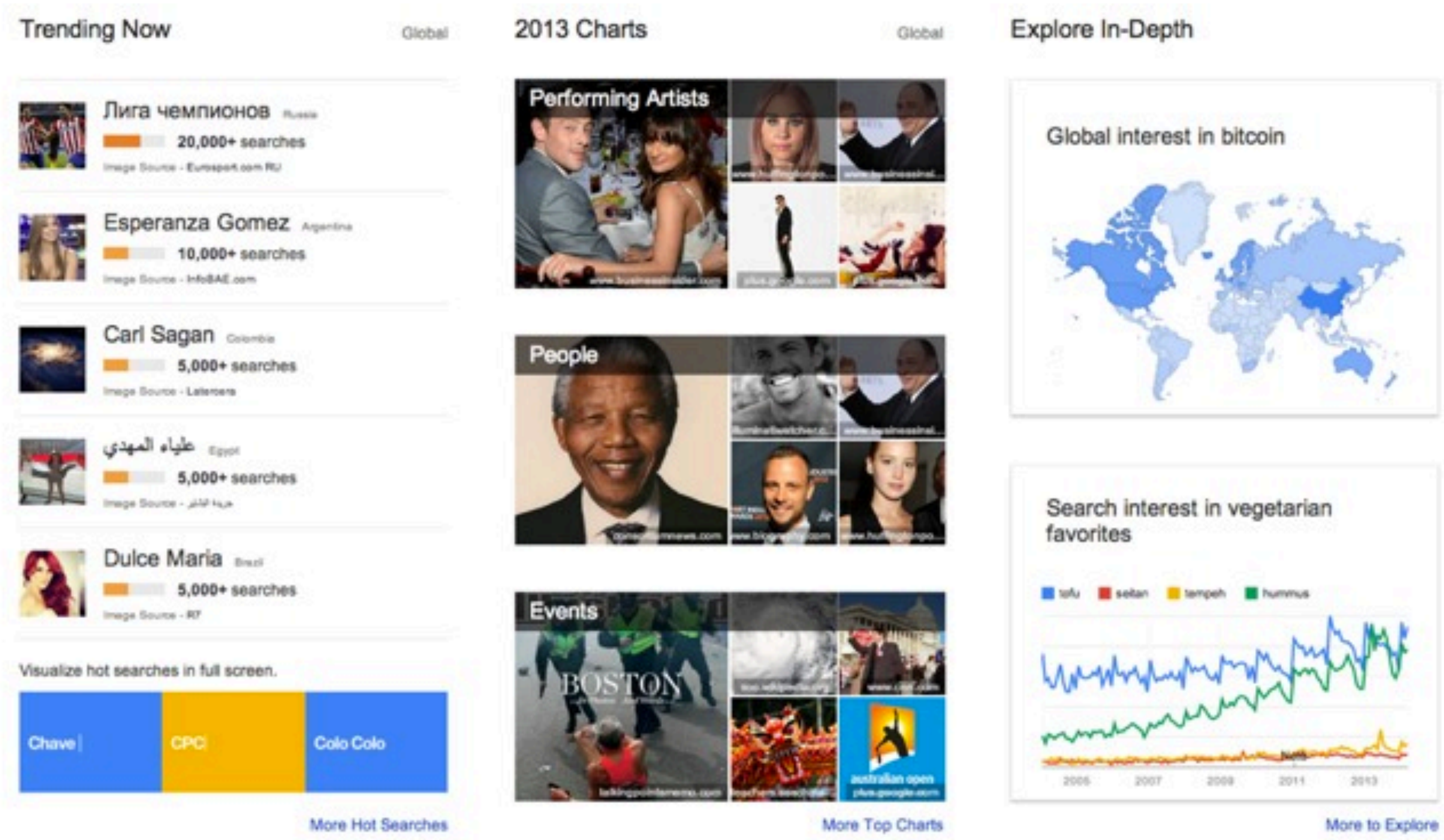
- Similar study on Facebook with more than 1 billion users!
- Sophisticated algorithms estimated the average path length between users: 4!

**References:**

Travers, Jeffrey & Stanley Milgram. 1969. "An Experimental Study of the Small World Problem." *Sociometry*, Vol. 32, No. 4, pp. 425-443.

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, Sebastiano Vigna: Four degrees of separation. WebSci 2012:33-42

# Google Trends



Trending Now — Global

Лига чемпионов — Russia
20,000+ searches
Image Source - Eurosport.com RU

Esperanza Gomez — Argentina
10,000+ searches
Image Source - InfoBAE.com

Carl Sagan — Colombia
5,000+ searches
Image Source - Laterbera

علياء المهدي — Egypt
5,000+ searches
Image Source - [illegible]

Dulce Maria — Brazil
5,000+ searches
Image Source - R7

Visualize hot searches in full screen.

Chave | CPC | Colo Colo

More Hot Searches

2013 Charts — Global

Performing Artists

People

Events

More Top Charts

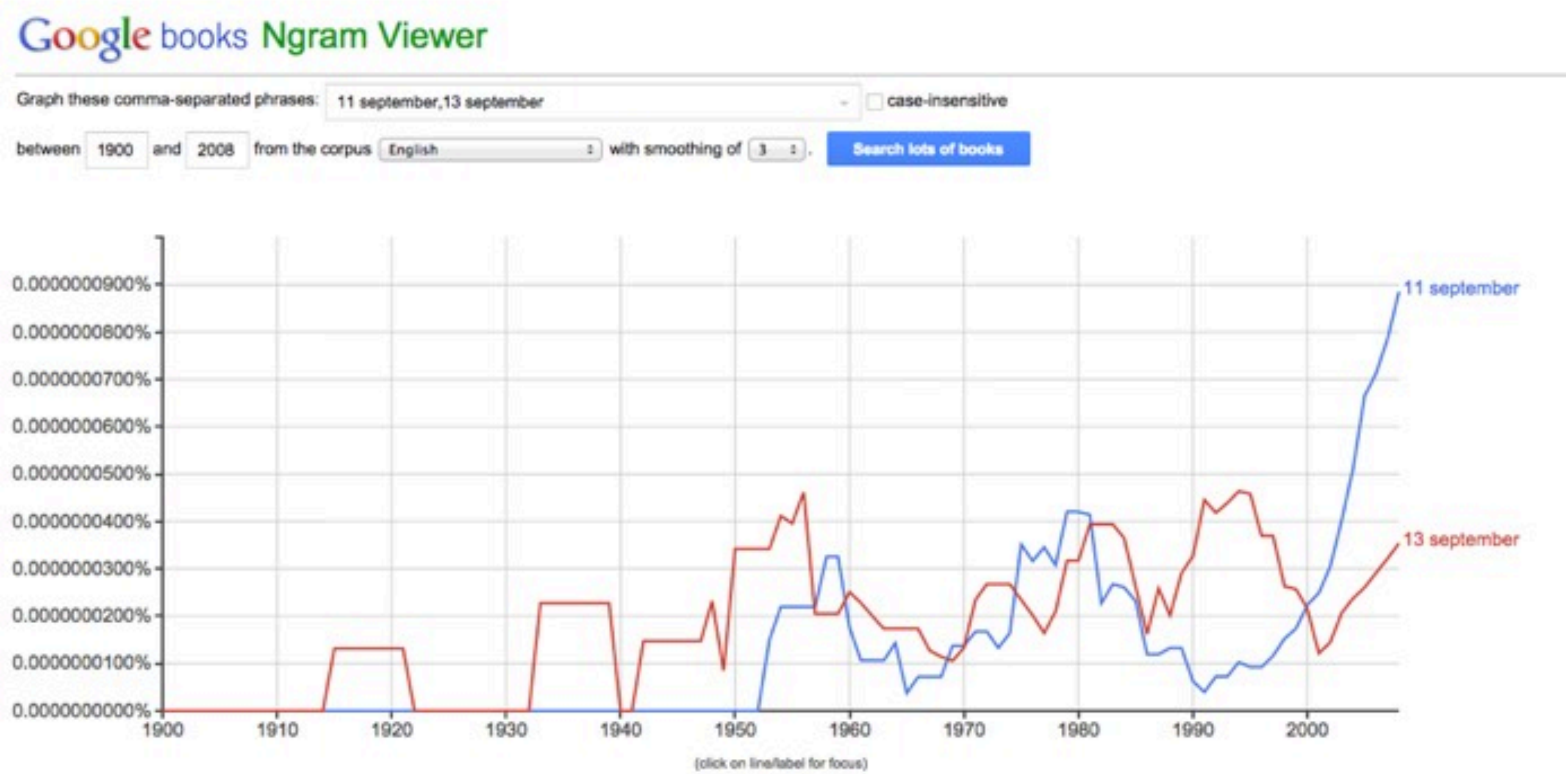
Explore In-Depth

Global interest in bitcoin

Search interest in vegetarian favorites

More to Explore

# Google Books Ngram



Google books Ngram Viewer

Graph these comma-separated phrases: 11 september,13 september □ case-insensitive

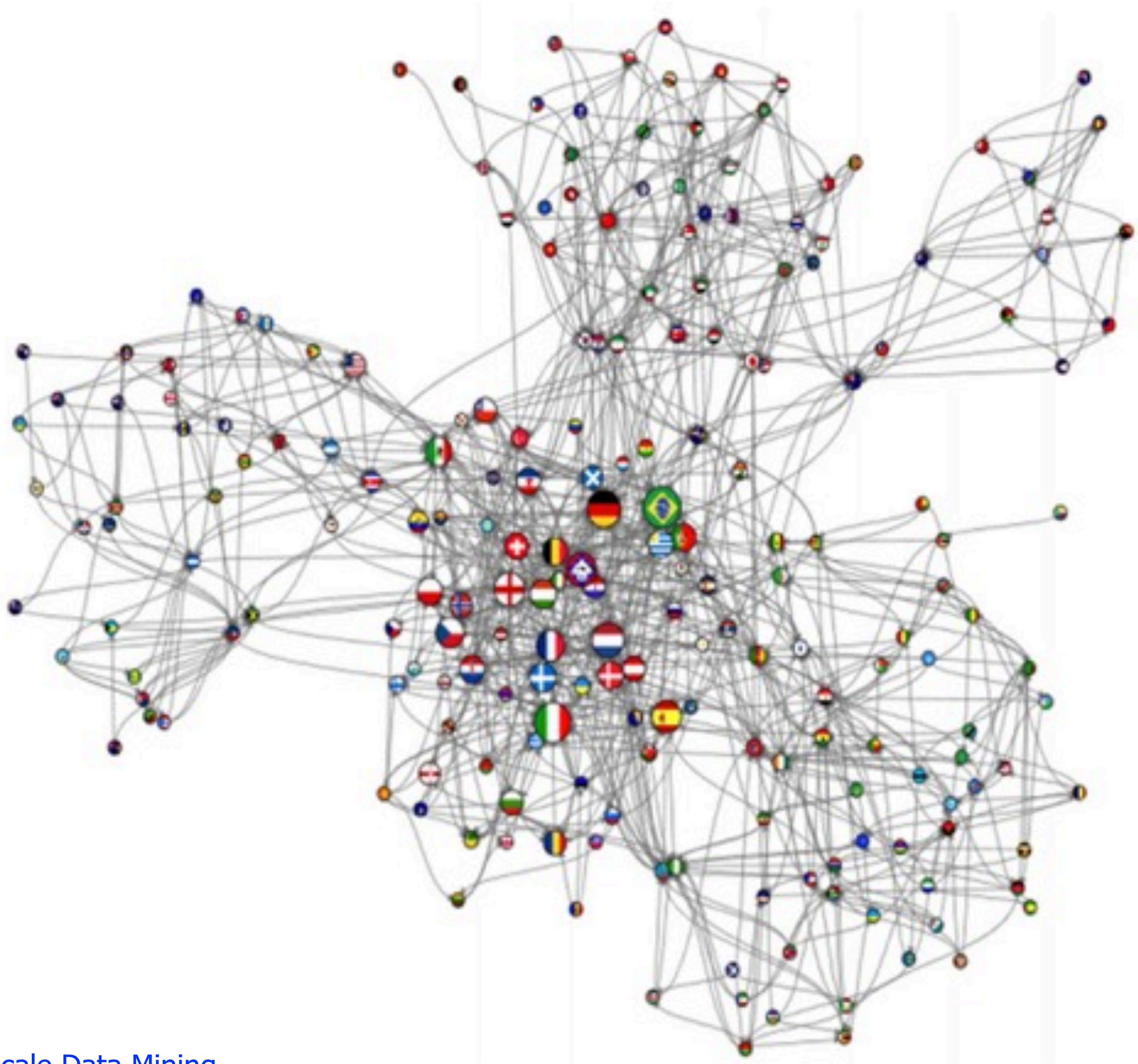
between 1900 and 2008 from the corpus English with smoothing of 3 . Search lots of books

# Ranking Web Pages

- Googling “Stanford University”
  - Stanford University Official web page
  - Stanford University Press...
  - Stanford health care...
  - ...

- Pages are ranked according to their importance... How? PageRank Algorithm...

# Finding best football teams with PageRank



# The BigData Revolution

- BigData is revolutionizing:
  - Crime prevention. We can predict crimes by mining past data.
  - Healthcare. Mining query logs and Twitter for finding flu trends.
  - Detecting earthquakes with Twitter.

google.org Flu Trends around the world

# Why Mine Data? Commercial Viewpoint

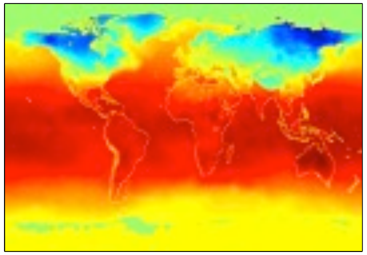
- Lots of data is generated:
  - Web data, e-commerce
  - purchases at department/grocery store
  - Bank/Credit card transactions

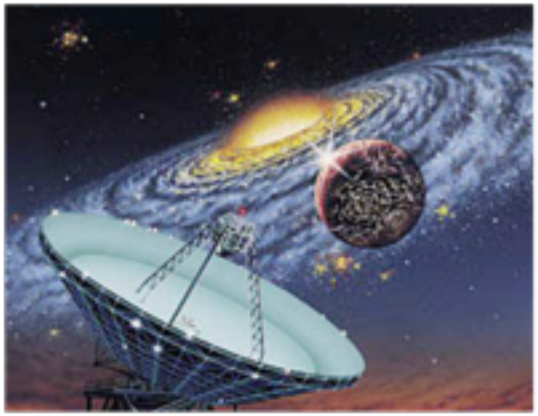
[NEWSFACTOR NETWORK]

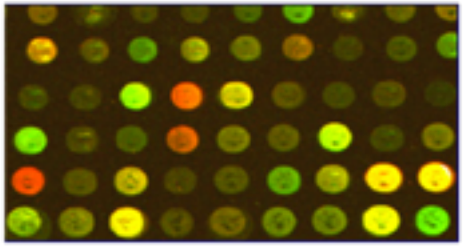
Google

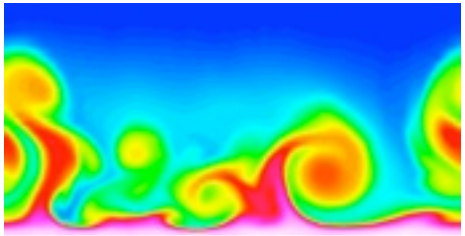
# Why Mine Data? Scientific Viewpoint

- Data collected from:
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating expression data
  - scientific simulations
  - people using Facebook, Twitter, Google
- Data mining may help scientists
  - in classifying and segmenting data
  - in hypothesis formations
  - modeling real-world phenomenon, human behaviour







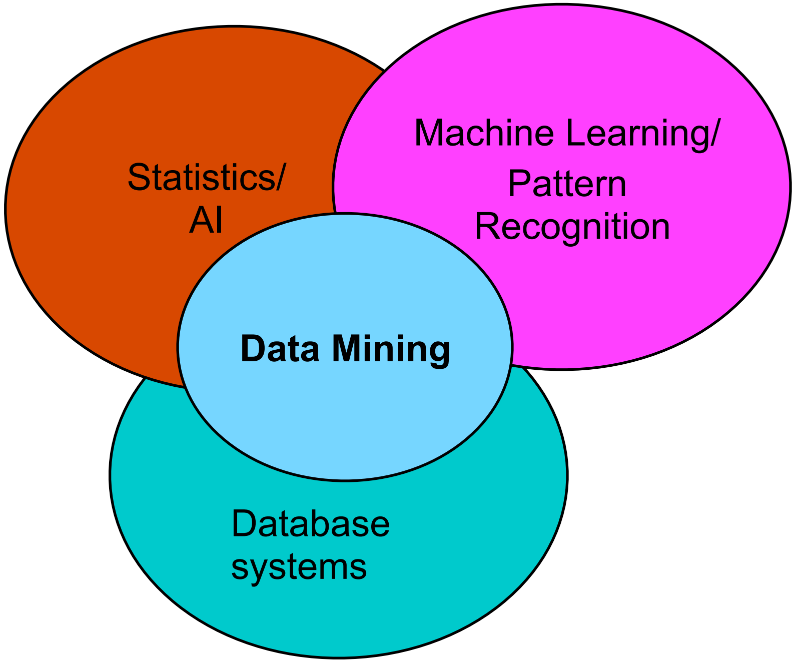


# Data Mining: Definition

- 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.

# Origins of Data Mining

- Uses ideas and techniques from: machine learning/AI, pattern recognition, statistics, theory of algorithms, database systems,...



- Issues:
  - Massive amount of data
  - High dimensionality
  - Heterogenous, distributed nature of data

# Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

# Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Ranking [Descriptive]
- Recommendation Systems [Predictive]

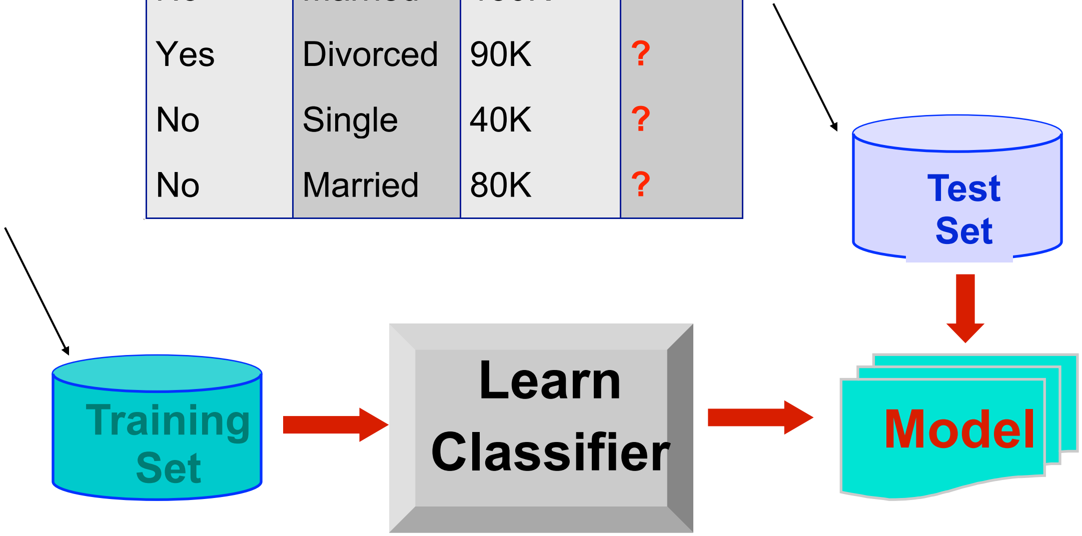
# Classification: Definition

- 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 | **?** |



# Classification: Application 1

- Direct Marketing
  - Goal: *Target* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don’t buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

# Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

# Classification: Application 3

- Customer Attrition/Churn:
  - 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.

From [Berry & Linoff] Data Mining Techniques, 1997

# Classification: Application 4

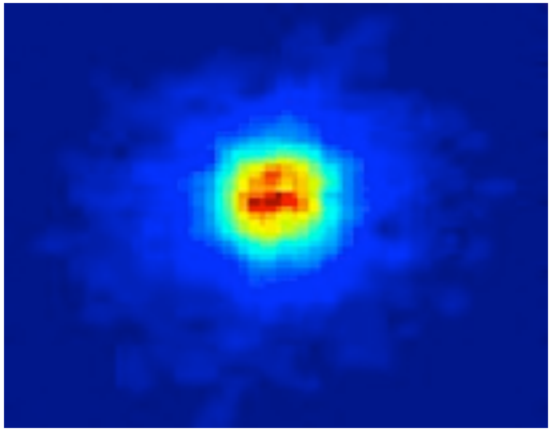
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

# Classifying Galaxies

**Courtesy: http://aps.umn.edu**

***Early***



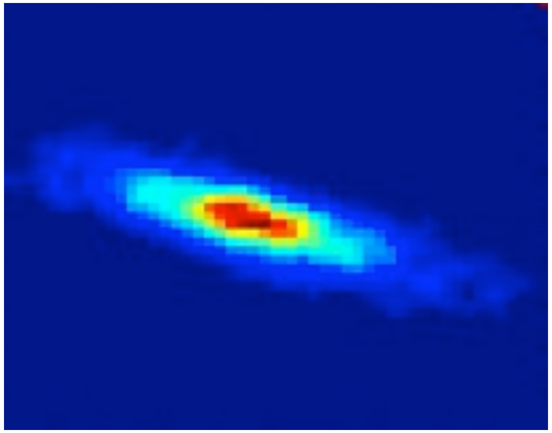
**Class:**

- **Stages of Formation**

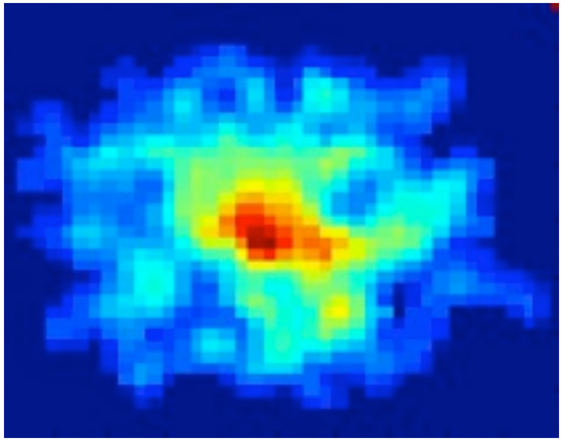
**Attributes:**

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

***Intermediate***



***Late***



**Data Size:**

- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

# Clustering Definition

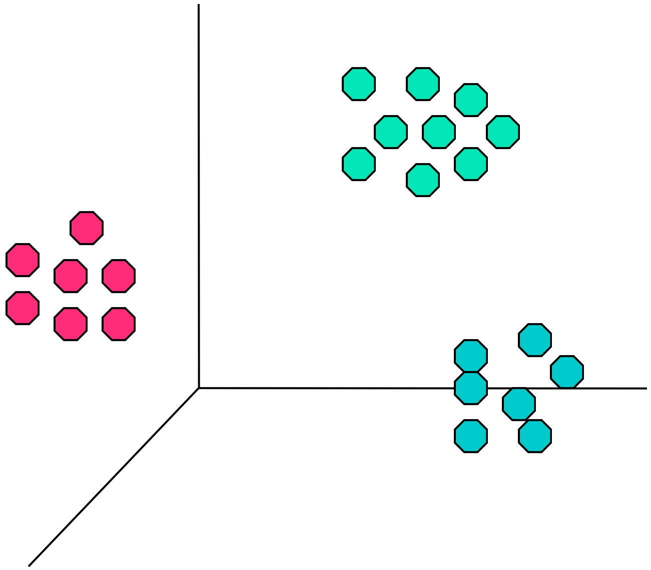
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance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

# Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



# Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

# Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

# Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

| Category | Total Articles | Correctly Placed |
|---|---|---|
| **Financial** | 555 | 364 |
| **Foreign** | 341 | 260 |
| **National** | 273 | 36 |
| **Metro** | 943 | 746 |
| **Sports** | 738 | 573 |
| **Entertainment** | 354 | 278 |

# Clustering of S&P 500 Stock

Observe Stock Movements every day.
Clustering points: Stock-{UP/DOWN}
Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
We used association rules to quantify a similarity measure.

| | *Discovered Clusters* | *Industry Group* |
|---|---|---|
| **1** | Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN | Technology1-DOWN |
| **2** | Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN | Technology2-DOWN |
| **3** | Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN | Financial-DOWN |
| **4** | Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP | Oil-UP |

# Association Rule Discovery:

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

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

Rules Discovered:

**{Milk} --> {Coke}**

**{Diaper, Milk} --> {Beer}**

# Association Rule Discovery: Application 1

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

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

# Association Rule Discovery: Application 2

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

# Association Rule Discovery: Application 3

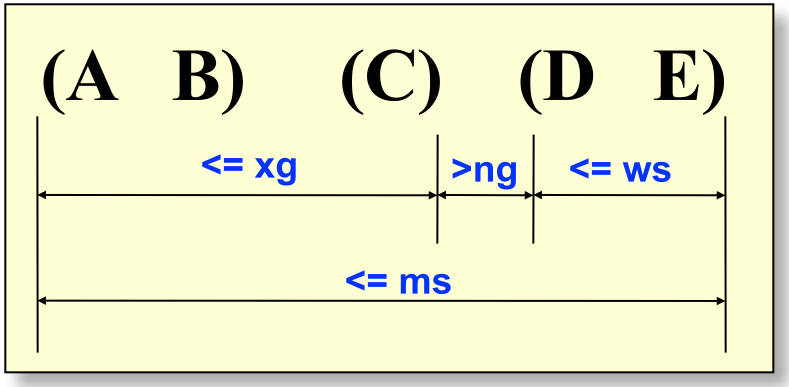
- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

# Sequential Pattern Discovery:

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.

$$(A\ \ B)\quad (C) \longrightarrow (D\ \ E)$$

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



# Sequential Pattern Discovery:

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)

    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:

    (Intro_To_Visual_C) (C++_Primer) -->

    (Python_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:

    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

# Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising costs.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

# Deviation/Anomaly Detection

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

  - Network Intrusion Detection

Internet
DOS
probe
U2R

# Challenges of Data Mining

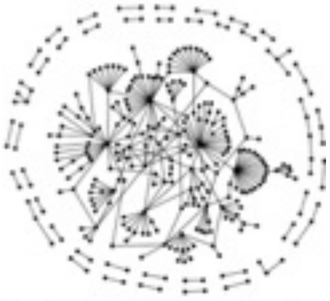
- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

# Real-World Graphs

- Graphs represent pairwise relations between entities in the real world (web pages, proteins, social network users,..)

- Real-world graphs are big! Facebook, 1.2 billion users, Twitter 600M active users, 5k tweets per sec.

Internet network

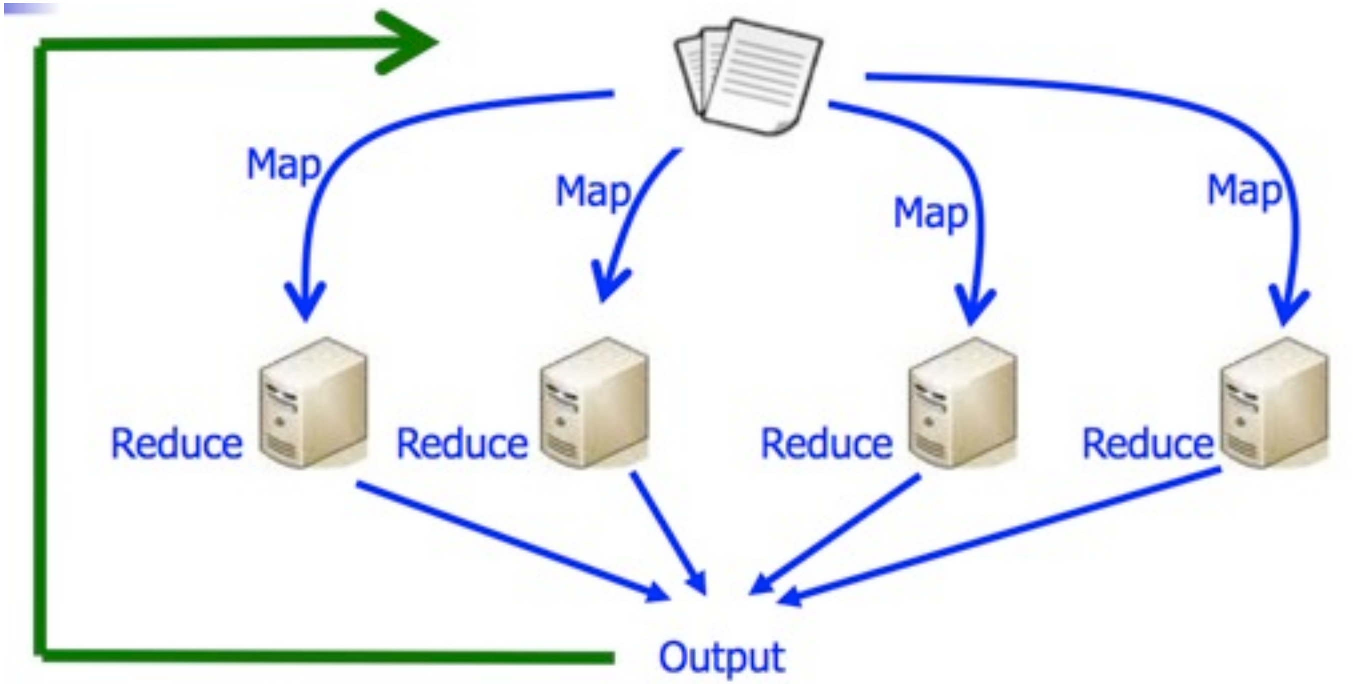


Yeast protein-protein interactions

High school dating network

Twitter follower graph

# MapReduce



Initially developed by Google, it is used by several companies (Yahoo!, IBM) and universities (Cornell, CMU... Telecom ParisTech).

# Our class

- We will cover:
  - PageRank
  - Clustering
  - Decision Trees, Classifiers
  - Spark
  - …

# Books

- Mining of Massive Datasets. J. Leskovec, A. Rajaraman, J. D. Ullman (available online).

- Introduction to Data Mining, Pang-Ning Tan,  Michael Steinbach, Vipin Kumar. Addison-Wesley.

- Data Mining and Analysy. Mohammed J. Zaki, Wagner Meir Jr. Cambridge University Press. (more advanced topics).

# Evaluation

- Max grade: 20.
  - Lab session 1, October 5th (data analysis in Python). Max 5 points
  - Lab session 2, October 19th (data analysis in Python). Max 5 points
  - Two more sessions in Spark (not evaluated).
  - **Final exam** (written): November 9th, max 10 points.