

DATA MINING INTRO LECTURE

Introduction

Instructors

Aris (Aris Anagnostopoulos)



Yiannis (Ioannis Chatzigiannakis)



Evimaria (Evimaria Terzi)



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 are generated by the second
 - Mobile devices, digital photographs, web documents.
 - Facebook updates, Tweets, Blogs, User-generated content
 - Transactions, sensor data, surveillance data
 - Queries, clicks, browsing
 - Cheap storage has made possible to maintain this data
- Need to analyze the raw data to extract knowledge

Why do we need data mining?

- 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, 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
 - We need a way to harness the **collective intelligence**
 - **Data are transforming many other fields: politics, biology, sociology, marketing**

Politics – Nate Silver



Politics – Obama campaign

Obama performed a targeted campaign.

They gathered data and demographic info from voters

They controlled tweets

They would send related messages to voters

Recommender systems

You buy something in Amazon and they propose other items you may be interested in.

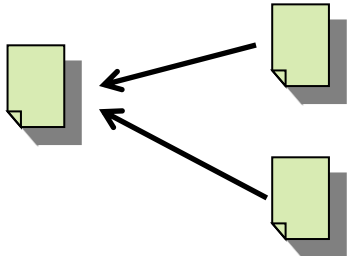
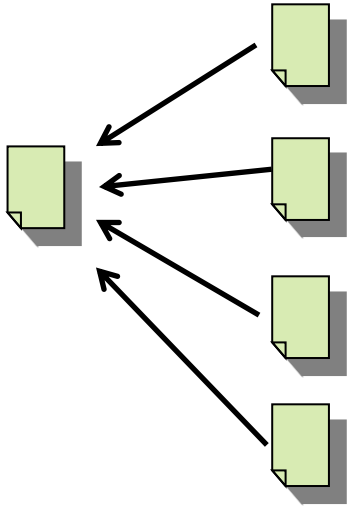
You watch youtube videos, it will recommend others.

You make a google query, it will propose others.

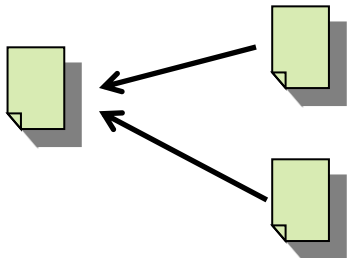
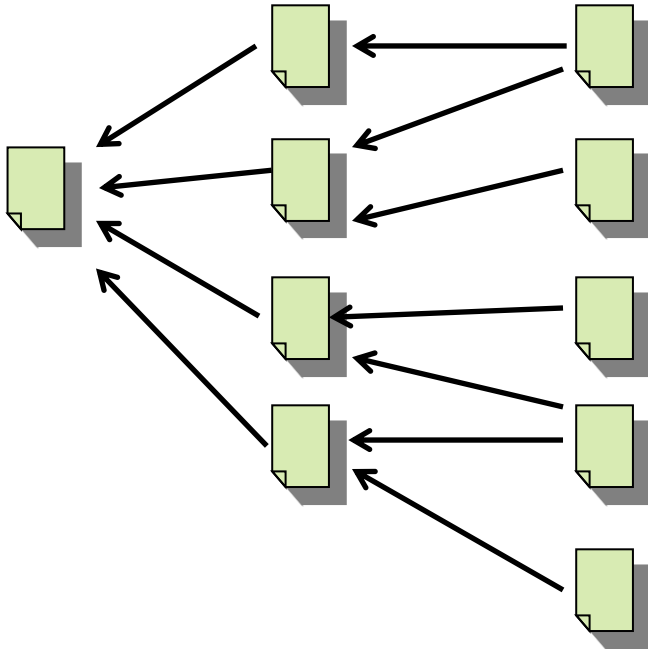
How do they do it?

(They analyze what previous **similar** users have done!)

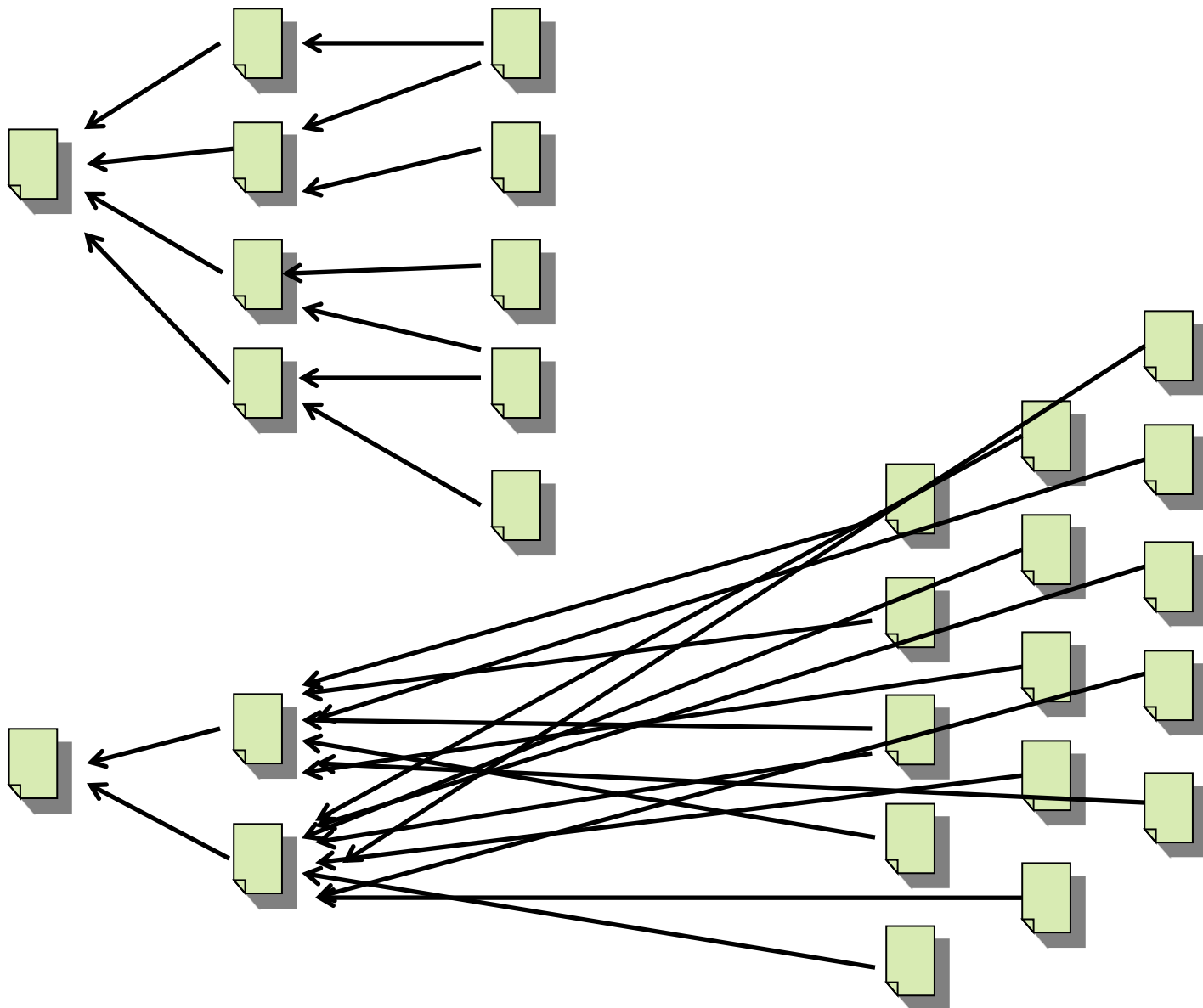
Google and PageRank



Google and PageRank



Google and PageRank

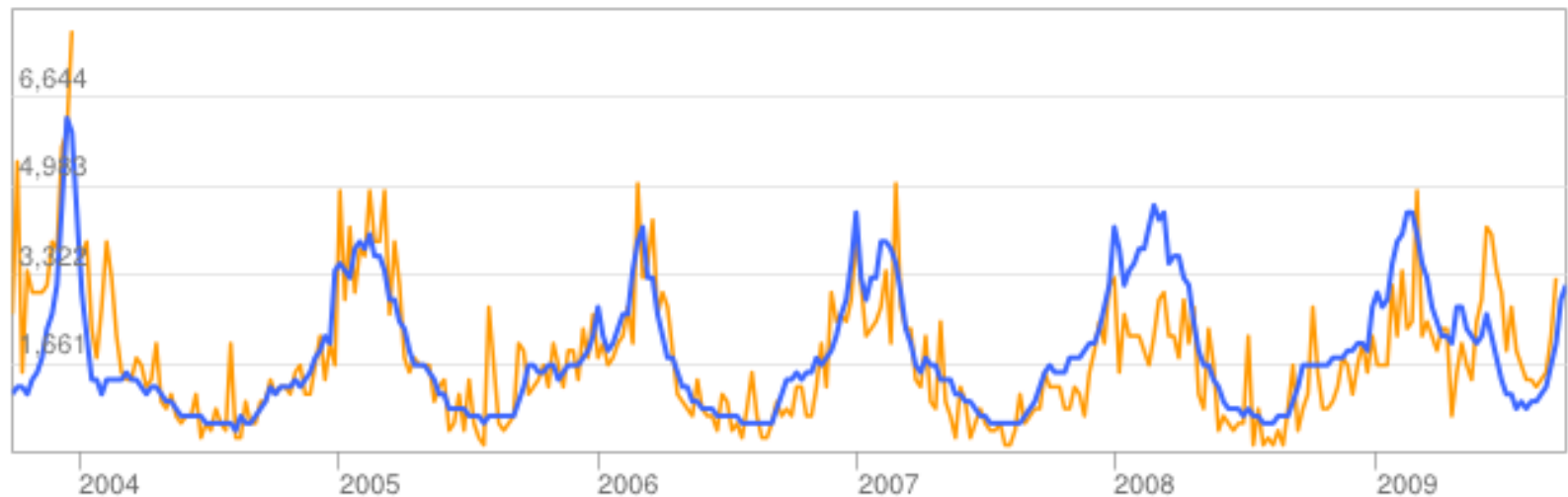


Google flu

Canada Flu Activity

Influenza estimate

● Google Flu Trends estimate ● Canada data

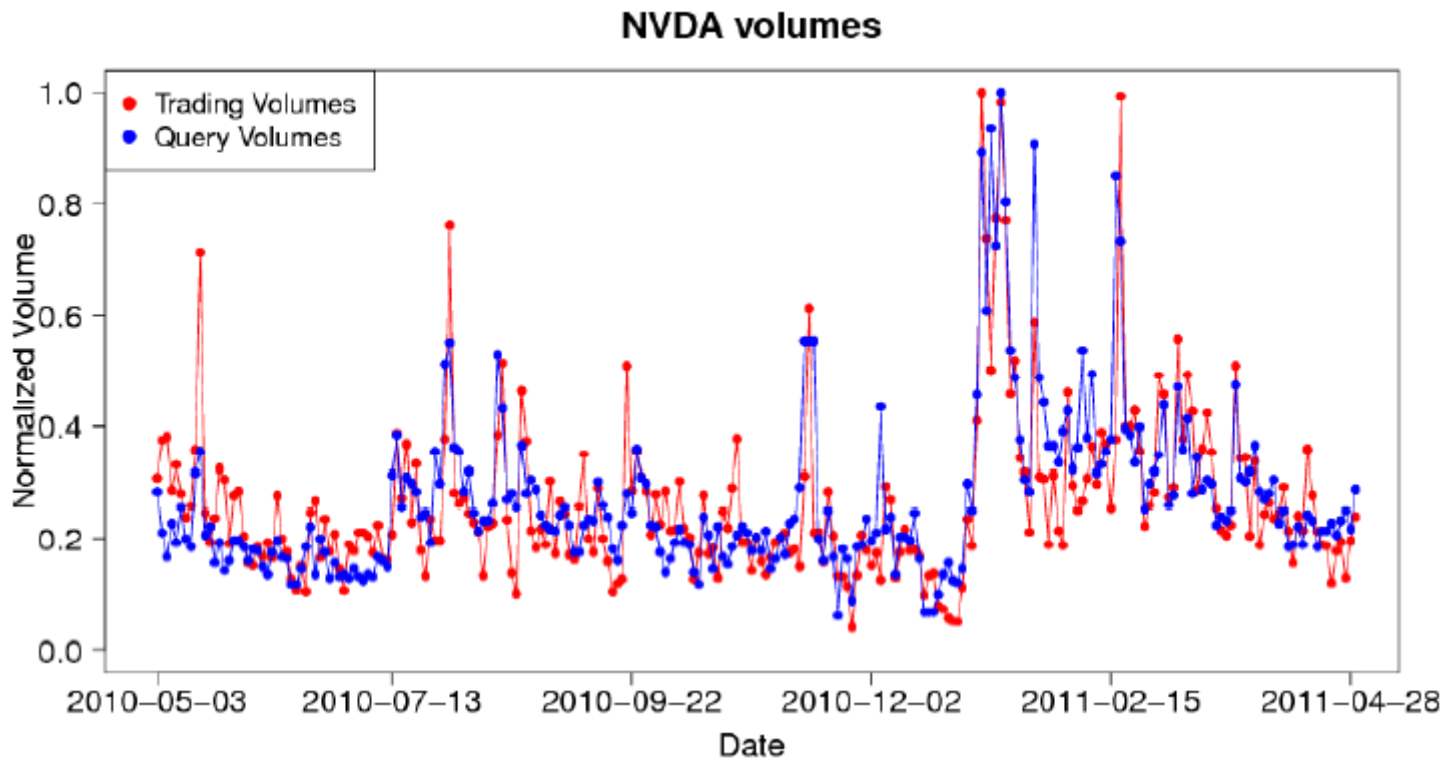


Canada: Influenza-like illness (ILI) data provided publicly by the [Public Health Agency of Canada](#).

Google and stockmarket

Web Search Queries Can Predict Stock Market Volumes

Ilaria Bordino¹, Stefano Battiston², Guido Caldarelli^{3,4,5}, Matthieu Cristelli^{3*}, Antti Ukkonen¹, Ingmar Weber¹



Google translate

The screenshot shows the Google Translate website in a Mozilla Firefox browser. The address bar displays the URL `translate.google.com/#auto/en/I love data mining!%0A%0AI love data science!`. A blue banner at the top of the page reads "Try a new browser with automatic translation." with buttons for "Download Google Chrome" and "Dismiss". The Google logo and a "Sign in" button are in the top right. The "Translate" section features a language selector with "English - detected" selected, and a "Translate" button. The input text box contains "I love data mining!" and "I love data science!". The output box shows the Italian translation: "Adoro data mining!" and "Io amo la scienza dati!". The footer includes links for "Turn off instant translation", "About Google Translate", "Mobile", "Privacy", "Help", and "Send feedback".

Google Translate - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Google Translate

translate.google.com/#auto/en/I love data mining!%0A%0AI love data science!

Most Visited Predicting the future ... TAX odissea telefilm ungh... Probability Problems Greek, March 23 - Crit...

Try a new browser with automatic translation. Download Google Chrome Dismiss

Google Sign in

Translate

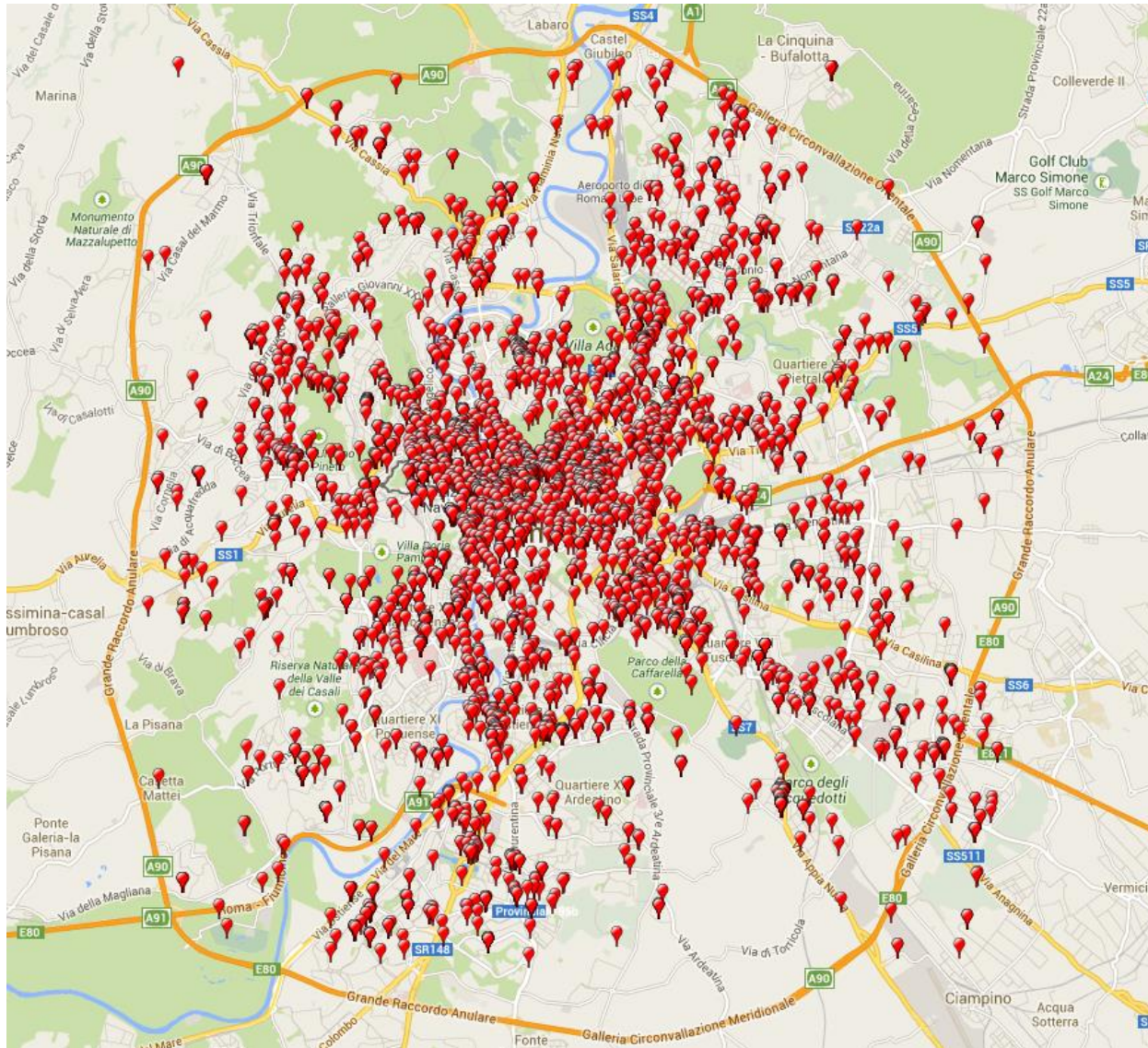
German English Finnish English - detected

English Finnish Italian Translate

I love data mining!
I love data science!

Adoro data mining!
Io amo la scienza dati!

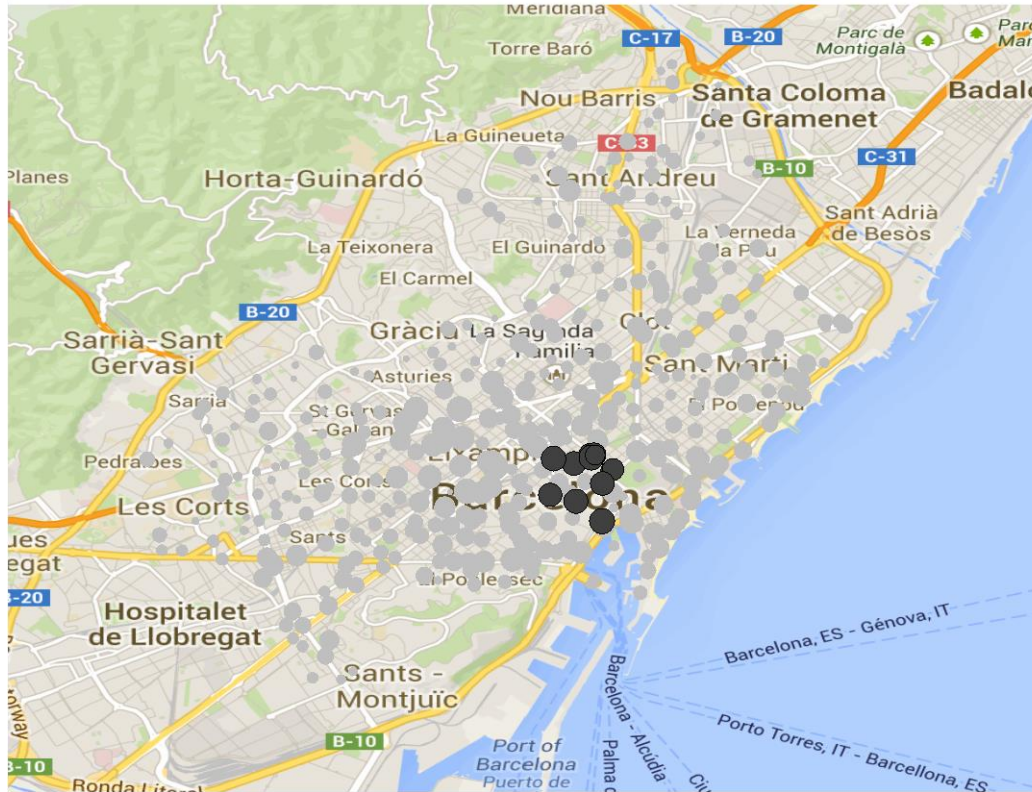
Turn off instant translation About Google Translate Mobile Privacy Help Send feedback



- People tweet about anything...
- Tweets provide a LOT of info
- Can we use it to obtain info about places, events, etc.?



Event detection with twitter



Psychology and Sociology

- Psychological and sociology studies have been revolutionalized with the incorporation of data science techniques
- Before based on surveys
- Now, with systems such as facebook, online games, etc. we can observe the behavior of hundreds of millions of people

What can fb say about relationships?

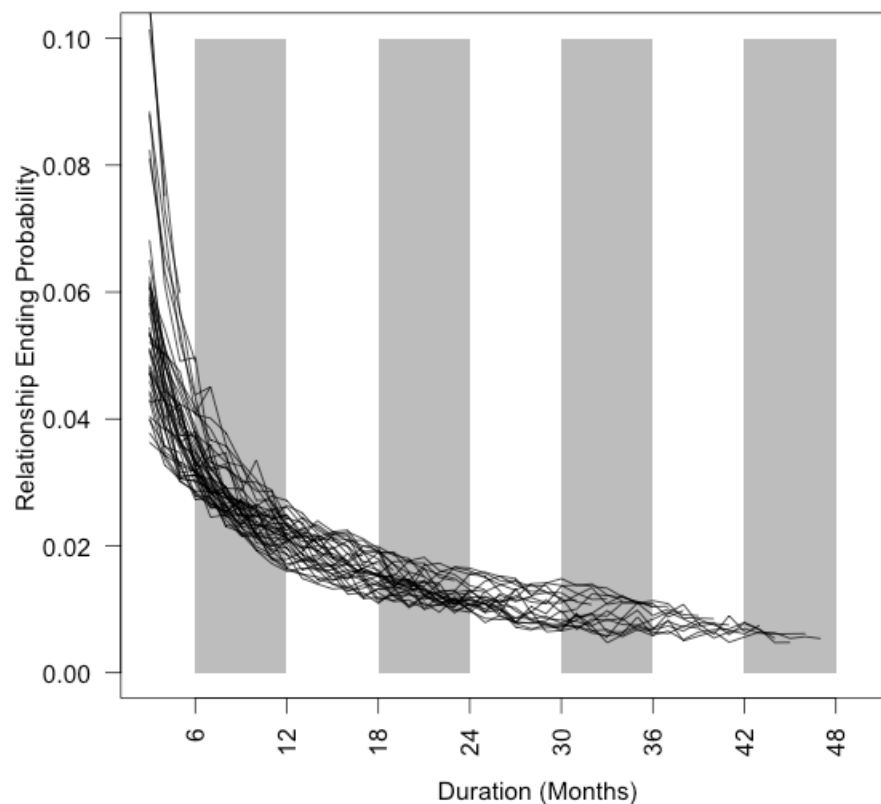
Facebook Can Predict With Scary Accuracy If Your Relationship Will Last

The Huffington Post | by Alexis Kleinman



1.7k

Posted: 02/14/2014 10:37 am EST | Updated: 02/14/2014 4:59 pm EST



Journalism

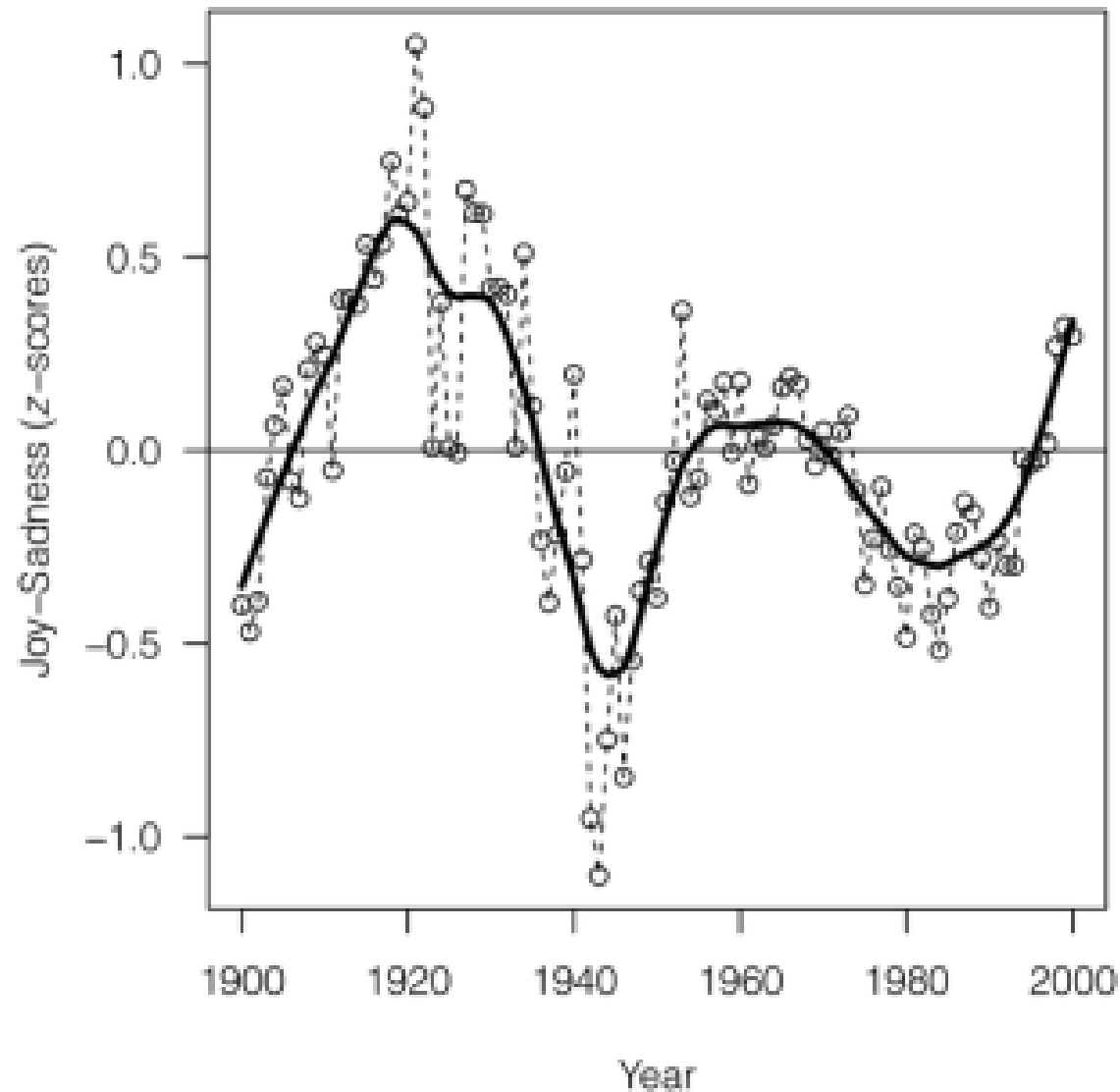
- Journalism is based on more and more data
- Twitter
- Wikileaks

Literature and history

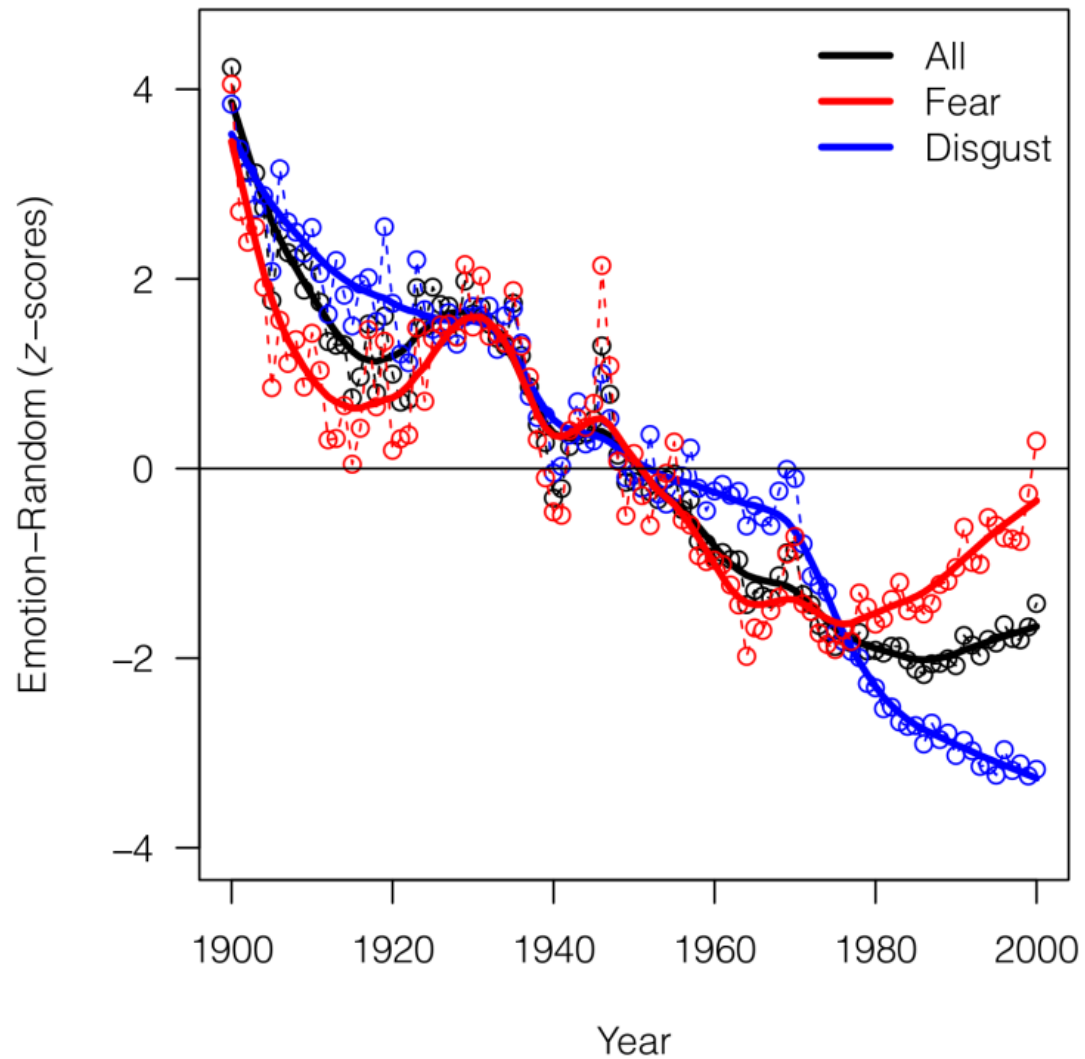
Researchers analyzed the words of thousands of books written in the 20th century.

The studied the words that express emotions over time.

Literature and history



Literature and history



Intro

Web page: <http://aris.me>

Protected:

- User: *****
- Pwd: *****

Register to the mailing list

Lectures

Books

What do you need to know

Office hours

Homeworks

Exams

Collaboration policy

The data are 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 through cameras, queries to search engines

Example: transaction data

- Billions of real-life customers:
 - WALMART: 20 million transactions per day
 - AT&T 300 million 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: 5 million english 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 active users
- Instant messenger: ~1 billion users
- WhatsApp: 900 million 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 $\rightarrow 3 \times 10^{12}$ 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.noaa.gov/ghcnm/>

- “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

Example: 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

So, what is “Data”?

- Collection of data **objects** and their **attributes**
- An attribute is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as **variable**, **field**, **characteristic**, or **feature**
- A collection of attributes describe an object
 - Object is also known as **record**, **point**, **case**, **sample**, **entity**, or **instance**

Objects

Attributes



<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
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

Size: Number of objects

Dimensionality: Number of attributes

Sparsity: Number of populated object-attribute pairs

Types of Attributes

There are different types of attributes

- **Categorical**
 - Examples: eye color, zip codes, words, rankings (e.g, good, fair, bad), height in {tall, medium, short}
 - **Nominal** (no order or comparison) vs **Ordinal** (order but not comparable)
- **Numeric**
 - Examples: dates, temperature, time, length, value, count.
 - **Discrete** (counts) vs **Continuous** (temperature)
 - Special case: **Binary** attributes (yes/no, exists/not exists)

Numeric Record Data

- If data objects have the same **fixed set** of **numeric attributes**, then the data objects can be thought of as **points** in a multi-dimensional space, where each **dimension** represents a distinct attribute
- Such data set can be represented by an **n-by-d data matrix**, where there are **n** rows, one for each object, and **d** columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Categorical Data

- Data that consists of a collection of records, each of which consists of a **fixed set** of **categorical** attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

Document Data

- Each document becomes a 'term' vector,
 - each term is a component (attribute) of the vector,
 - the value of each component is the number of times the corresponding term occurs in the document.
- **Bag-of-words** representation – no ordering

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

- Each record (transaction) is a **set of items**.

<i>TID</i>	<i>Items</i>
1	B read, C oke, M ilk
2	B eer, B read
3	B eer, C oke, D iaper, M ilk
4	B eer, B read, D iaper, M ilk
5	C oke, D iaper, M ilk

- A set of items can also be represented as a **binary vector**, where each attribute is an item.
- A document can also be represented as a **set of words** (no counts)

Sparsity: average number of products bought by a customer

Ordered Data

- Genomic **sequence** data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

- Data is a long **ordered** string

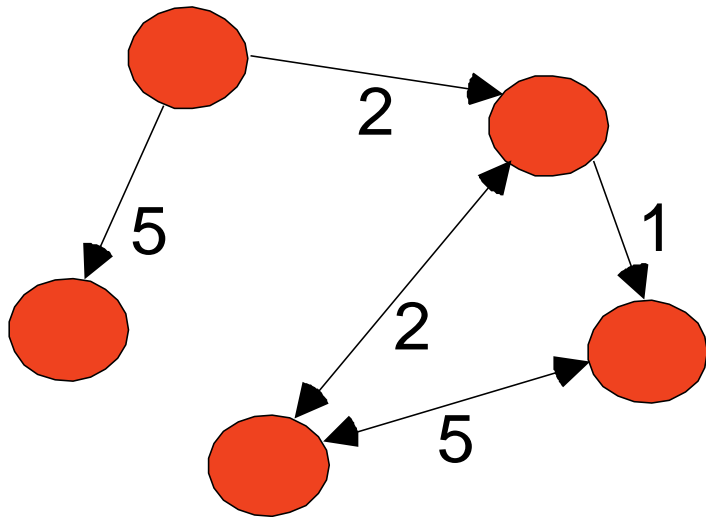
Ordered Data

- Time series
 - Sequence of ordered (over “time”) numeric values.



Graph Data

- Examples: Web graph and HTML Links



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>
```

```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>
```

```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>
```

```
<li>
```

```
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```


Types of data

- **Numeric data:** Each object is a point in a multidimensional space
- **Categorical data:** Each object is a vector of categorical values
- **Set data:** Each object is a set of values (with or without counts)
 - Sets can also be represented as binary vectors, or vectors of counts
- **Ordered sequences:** Each object is an ordered sequence of values.
- **Graph data**

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?

<i>TID</i>	<i>Items</i>
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 out of the data?

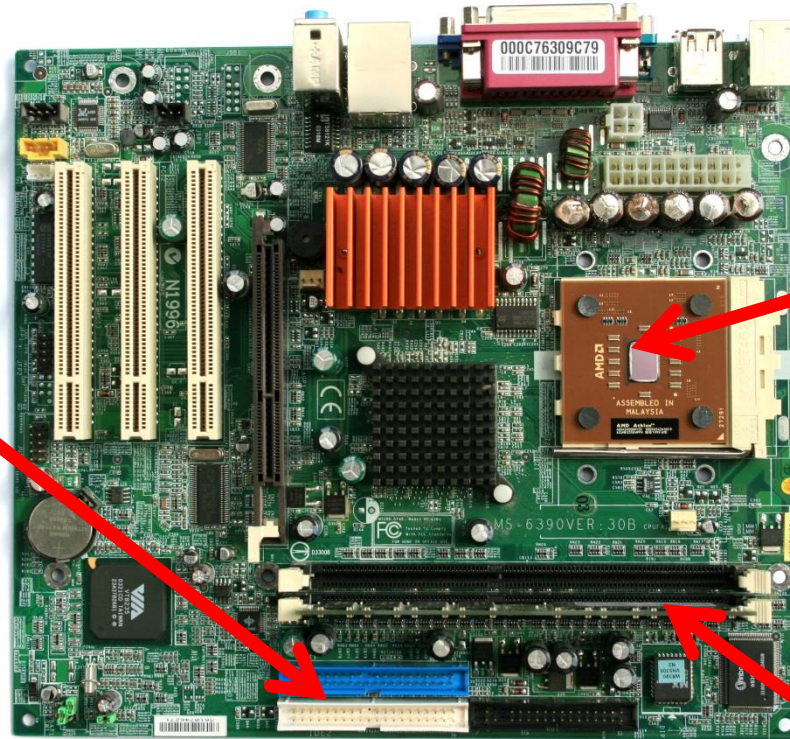
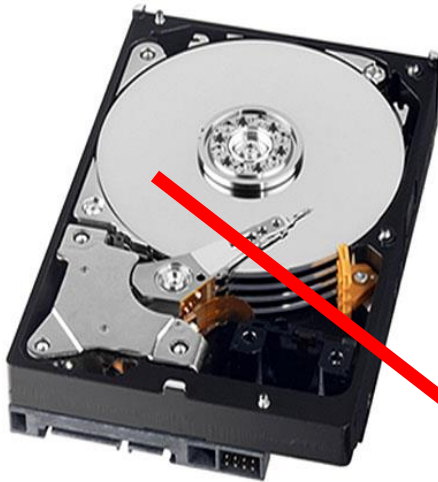
What can you do with the data?

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



Basics of Computer Architecture

Hard Disk (HD)



Processor (CPU)



Memory (RAM)

The Cloud

There exist large datacenters for storing data and making computations

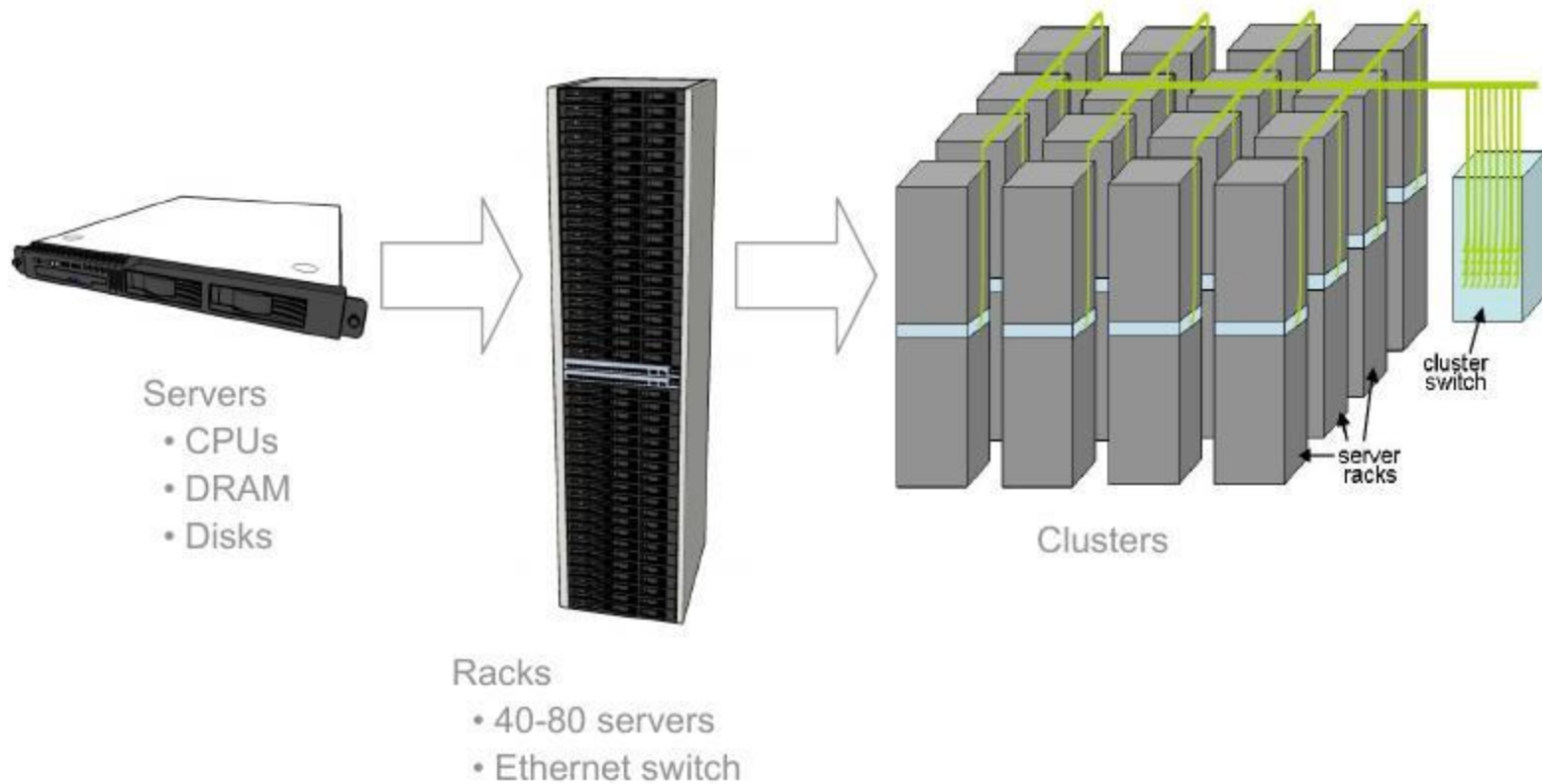
- Gmail, dropbox, ...



The Cloud



The Cloud



Some useful numbers

Operation	Time
Main memory reference	100ns
Send 2K bytes over 1 Gbps network	250ns
Read 1 MB sequentially from memory	150μs
Round trip within same datacenter	500μs
Disk seek	4ms
Read 1 MB sequentially from disk	2ms
Send packet CA->Netherlands->CA	150ms

Topics we will cover

- Text mining
 - Similarity measures
 - Near-neighbor search
 - Clustering
 - Classification
 - Graph mining
 - Frequent itemsets
 - Streaming
 - Recommender systems
 - Social networks
 - Models and learning
 - MapReduce
-
- We will start with probability, python