

# DATA MINING

# INTRO LECTURE

---

Introduction

# Instructors

**Aris** (Aris Anagnostopoulos)



**Teaching Assistant (TA):**

**Gianluca** (Gianluca De Carlo)



# Logistics

- Register: **Send email to Aris**
- Web page
- Class hours
- Office hours
- What do you need to know
- Remote attendance (nope)
- Book
- Homeworks, project, participation, exam
- Collaboration policy
- Protected content:
  - Username: **Send email to Aris**
  - Password: **Send email to Aris**

# 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.

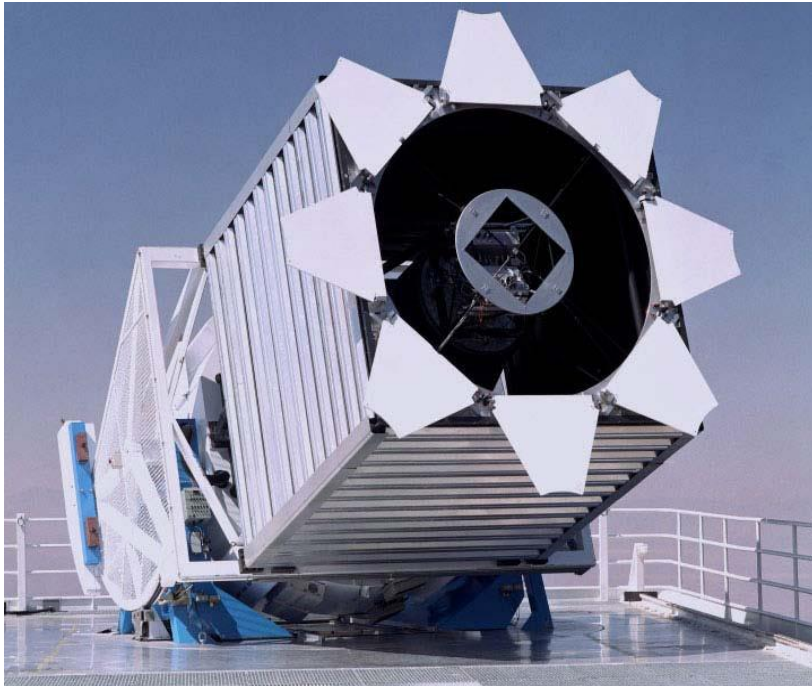


# 1. What is Data Mining?

- Large quantities of data are collected about all aspects of our lives
- This data contains interesting patterns
- Data Mining helps us to
  1. discover these patterns and
  2. use them for decision making across all areas of society, including
    - Business and industry
    - Science and engineering
    - Medicine and biotech
    - Governmen
    - Individuals



# “We are Drowning in Data...”



## Sloan Digital Sky Survey

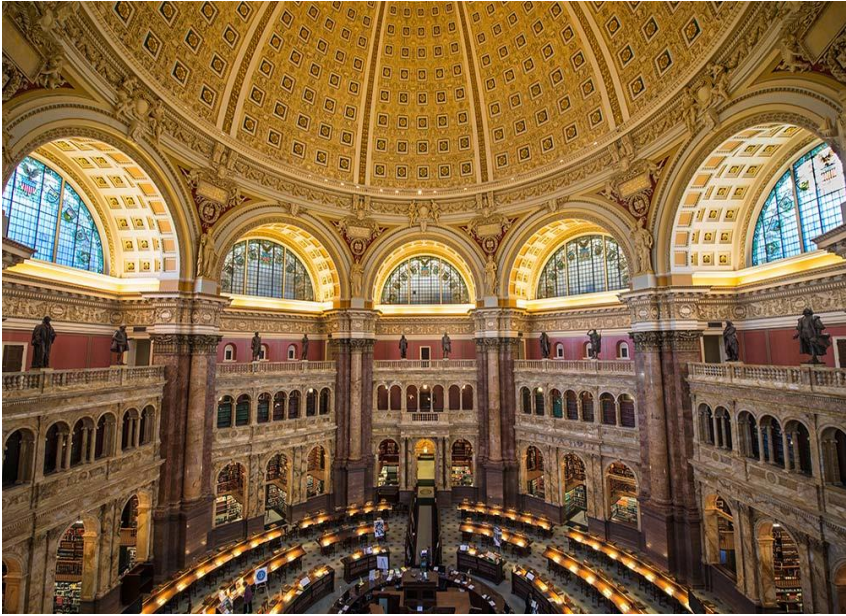
≈ 200 GB/day

≈ 73 TB/year

## Predict

- Type of sky object:  
Star or galaxy?

# “We are Drowning in Data...”



**US Library of Congress**  
≈ 235 TB archived  
≈ 40 Wikipedias

## **Discover**

- Topic distributions
- Historic trends\*
- Citation networks

\* Lansdall-Welfare, et al.: Content analysis of 150 years of British periodicals.PNSA, 2017.



# “We are Drowning in Data...”



## Facebook

- 4 Petabyte of new data generated every day
- over 300 Petabyte in Facebook's data warehouse

## Predict

- Interests and behavior of over one billion people

<https://www.brandwatch.com/blog/facebook-statistics/>

<http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/>



# “We are Drowning in Data...”

## 2019 *This Is What Happens In An Internet Minute*



## Predict

- Interests and behavior of mankind

# “We are Drowning in Data...”

**Law enforcement agencies** collect unknown amounts of data from various sources

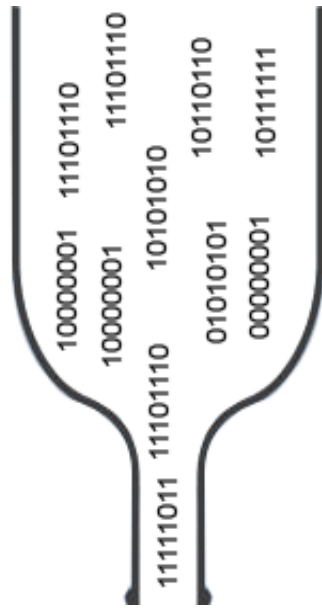
- Cell phone calls
- Location data
- Web browsing behavior
- Credit card transactions
- Online profiles (Facebook)
- ...

**Predict**

- Terrorist or not?
- Trustworthiness



# “...but starving for knowledge!”



← Amount of data that is collected

← Amount of data that can be looked at by humans

We are interested in **the patterns, not the data** itself!

Data Mining methods help us to

- discover interesting patterns in large quantities of data
- take decisions based on the patterns

# Definitions of Data Mining

## – Definitions

**Exploration & analysis, of large quantities of data in order to discover meaningful patterns.**

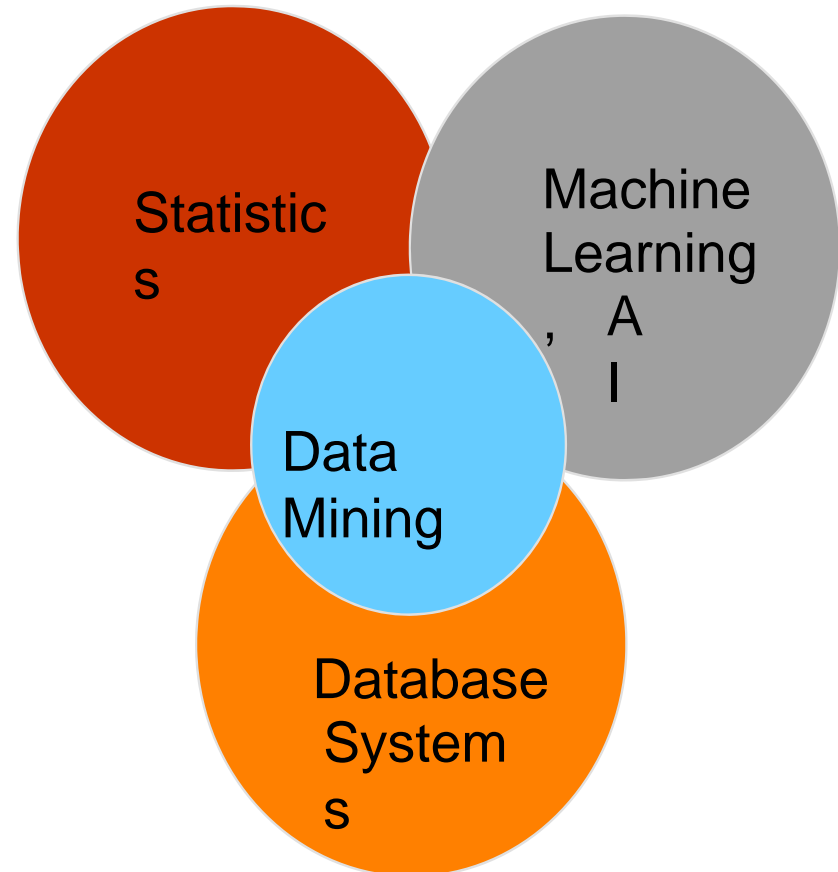
**Non-trivial extraction of**  
– **implicit,**  
– **previously unknown, and**  
– **potentially useful information from data.**

## – Data Mining methods

1. detect interesting patterns in large quantities of data
2. **support** human decision making by providing such patterns
3. **predict** the outcome of a future observation based on the patterns

# Origins of Data Mining

- Data Mining combines ideas from statistics, machine learning, artificial intelligence, and database systems
- Tries to overcome shortcomings of traditional techniques concerning
  - large amount of data
  - high dimensionality of data
  - heterogeneous and complex nature of data
  - explorative analysis beyond hypothesize-and-test paradigm



## 2. Tasks and Applications

### – Descriptive Tasks

- Goal: Find patterns in the data.
- Example: *Which products are often bought together?*

### – Predictive Tasks

- Goal: Predict unknown values of a variable
  - given observations (e.g., from the past)
- Example: *Will a person click a online advertisement?*
  - given her browsing history

### – Machine Learning Terminology

- descriptive = unsupervised
- predictive = supervised



# Politics – Nate Silver (Obama-Romney)





# 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.

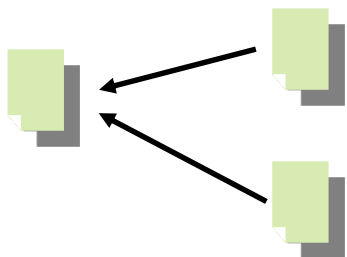
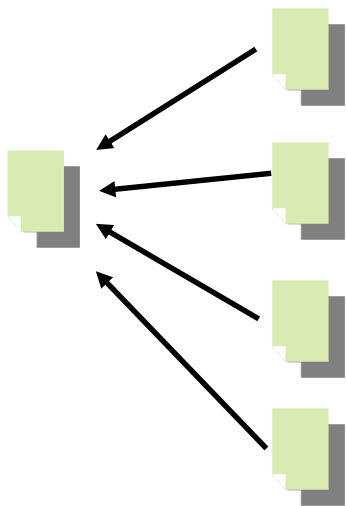
You watch youtube videos, it will recommend others.

You make a google query, it will propose others.

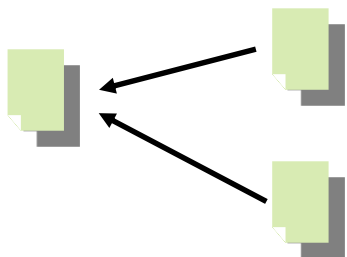
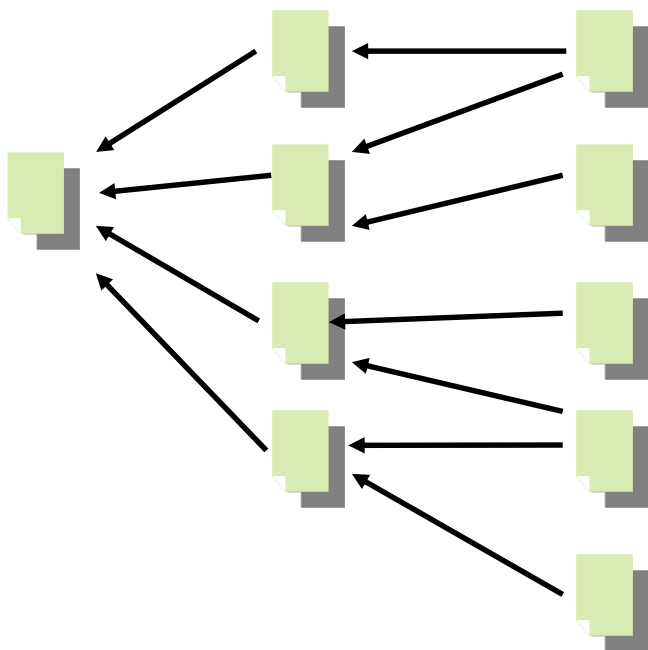
How do they do it?

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

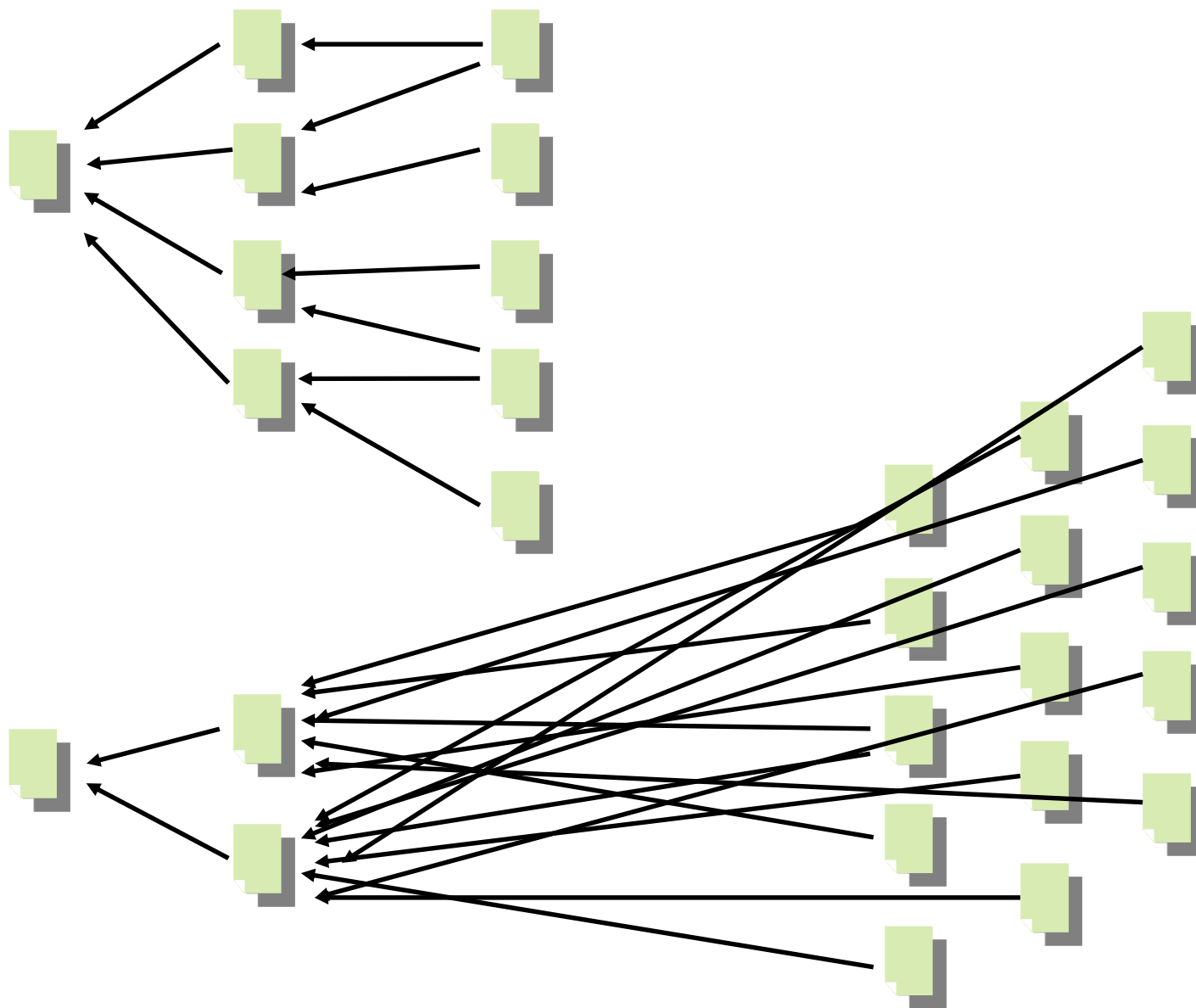
# Google and PageRank



# Google and PageRank



# Google and PageRank

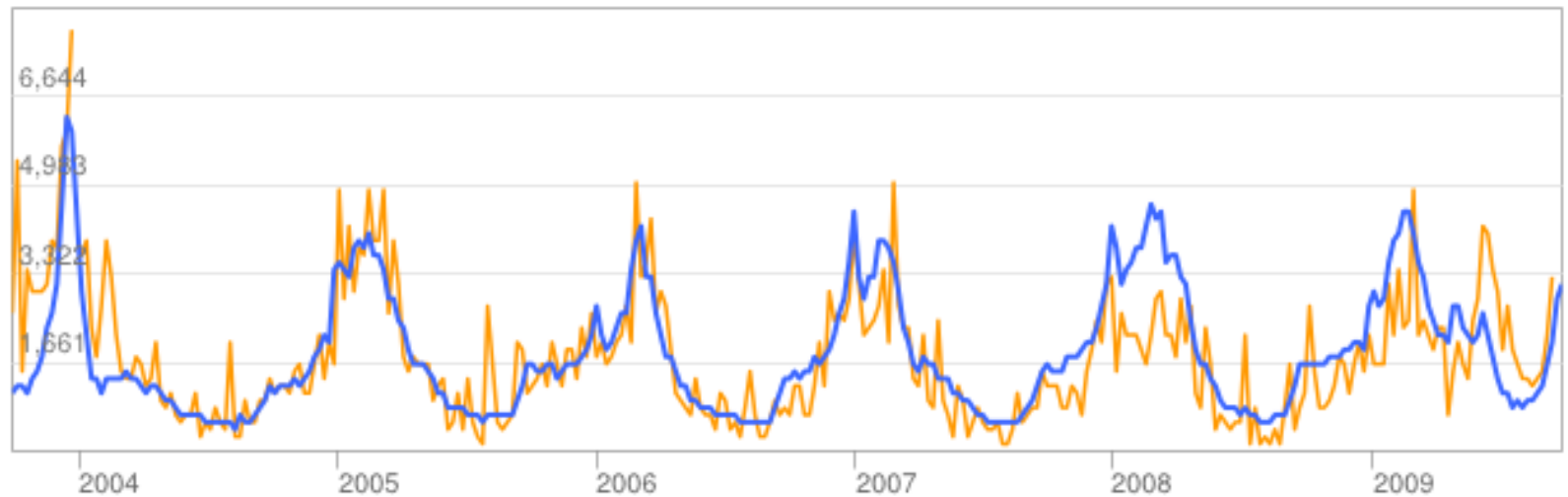


# Google flu

## Canada Flu Activity

Influenza estimate

● Google Flu Trends estimate ● Canada data

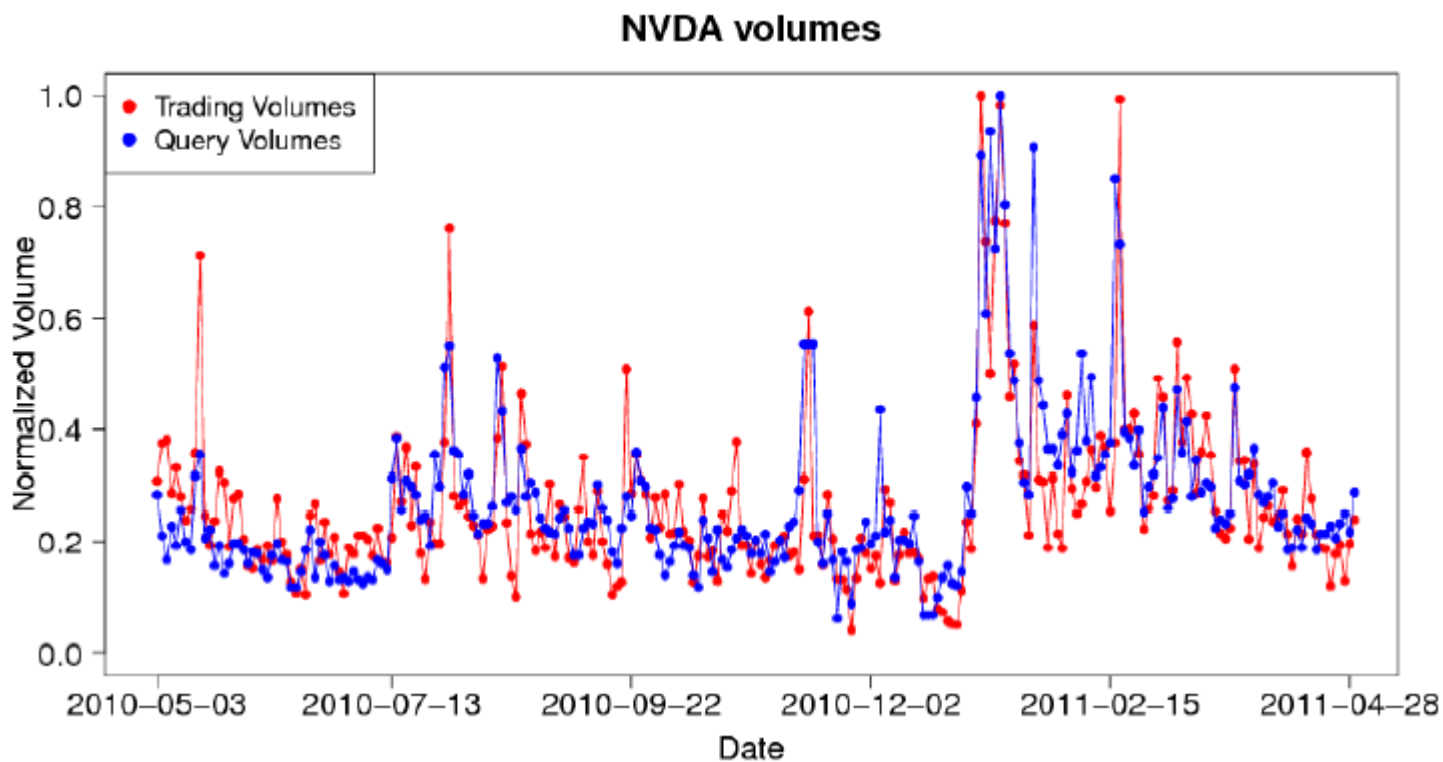


Canada: Influenza-like illness (ILI) data provided publicly by the [Public Health Agency of Canada](#).

# Google and stockmarket

## Web Search Queries Can Predict Stock Market Volumes

Ilaria Bordino<sup>1</sup>, Stefano Battiston<sup>2</sup>, Guido Caldarelli<sup>3,4,5</sup>, Matthieu Cristelli<sup>3\*</sup>, Antti Ukkonen<sup>1</sup>, Ingmar Weber<sup>1</sup>





# Google translate

The image shows a screenshot of the Google Translate website in a Mozilla Firefox browser window. The browser's address bar shows the URL `translate.google.com/#auto/en/I love data mining!%0A%0AI love data science!`. The page title is "Google Translate". The main content area features the Google logo and a "Sign in" button. Below the logo, the word "Translate" is displayed. The interface includes a language selection menu with "German", "English", "Finnish", and "English - detected" (selected). A "Translate" button is visible. The input text box contains "I love data mining!" and "I love data science!". The output text box shows the translated text: "Adoro data mining!" and "Io amo la scienza dati!". At the bottom of the page, there are 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

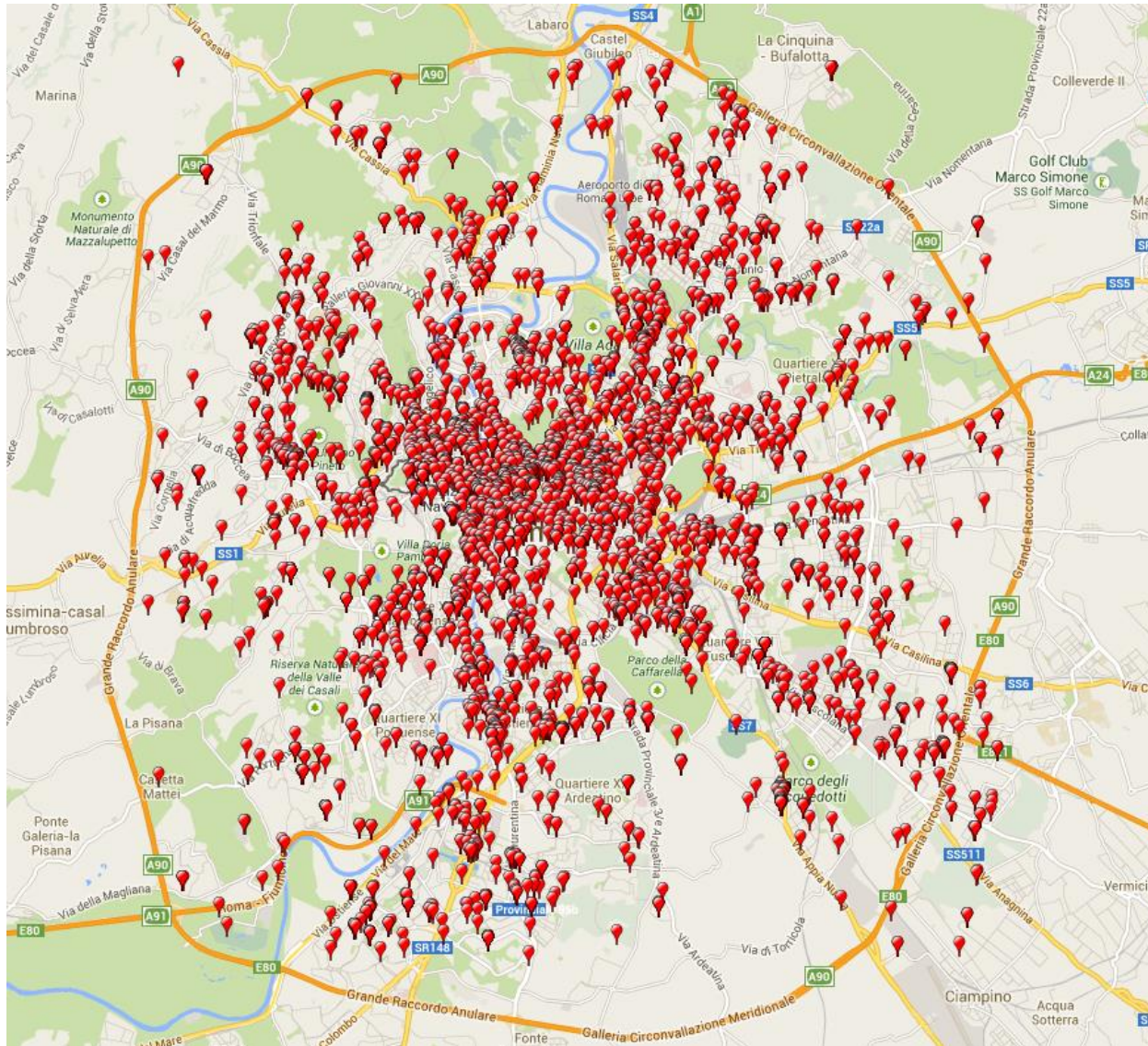
German English Finnish English - detected

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

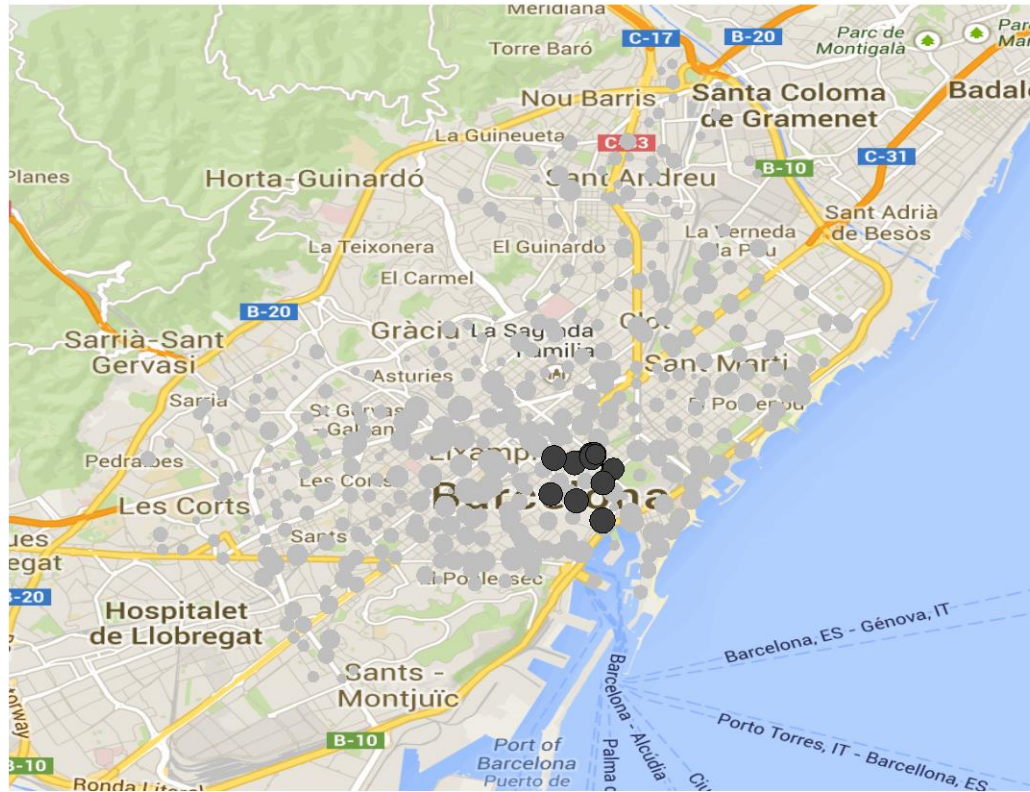


- People tweet about anything...
- Tweets provide a LOT of info
- Can we use it to obtain info about places, events, etc.?





# 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

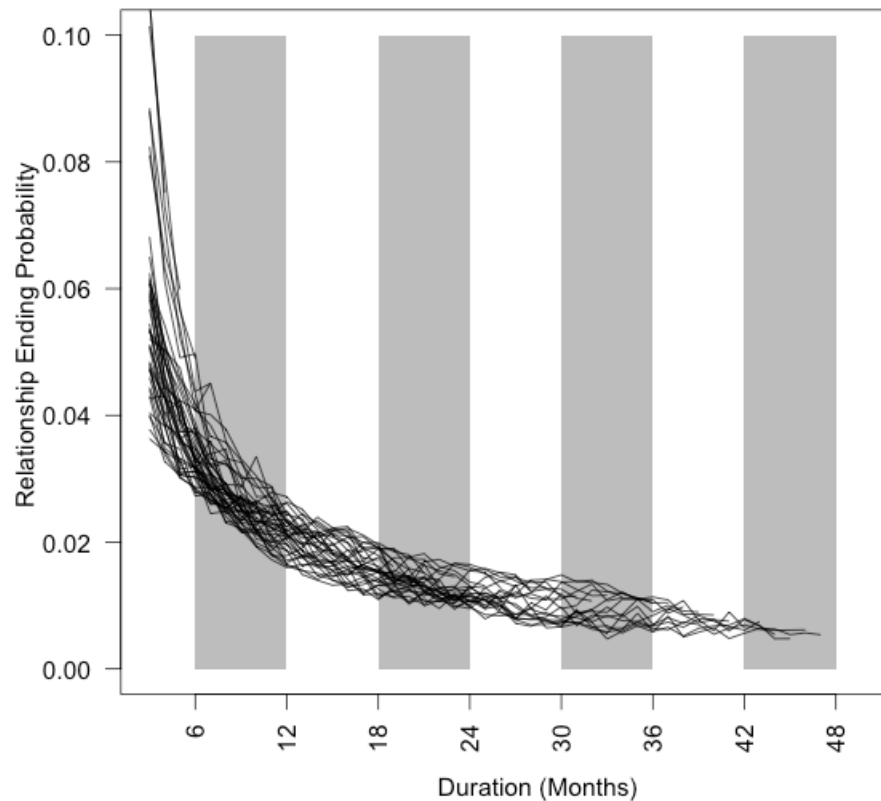
Posted: 02/14/2014 10:37 am EST | Updated: 02/14/2014 4:59 pm EST



Follow

Like

1.7k



# Are emotions contagious?

- In 2014, some FB researchers studied if emotions spread in FB
- They selected 150K users (group P) and they **increased** the number of **positive** posts that they see
- They selected other 150K users (group N) and they **increase** the number of **negative** posts that they see
- They studied what messages do these 300K users post
- Finding: users in **group P**, **increased** the number of **positive** posts and **decreased** the number of **negative**
- The opposite happened to **group N**



# Journalism

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

# Medicine

- Medicine has been completely transformed in the last 20 years

# Types of Data

- Structured
  - 5-10% of the data
  - SQL
- Semi-structured
  - 5-10% of the data
  - XML, CSV, JSON
- Unstructured
  - 80% of the data

# 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: Google indexes over 50 billion pages, linked via hyperlinks
- Facebook: 2.7 billion users
- Twitter: 330 million active users
- Instagram: ~1 billion users
- WhatsApp: 2 billion users
  
- Blogs: 600 million blogs worldwide, presidential candidates run blogs

# Example: genomic sequences

- There exist databases that contain the genome sequence of a lot of people
- Such data can be used to find correlations between diseases and gene mutations
- Example: UKBiobank: Mutations for 500K people



# 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>	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

**Size:** Number of objects

**Dimensionality:** Number of attributes

**Sparsity:** Number of populated object-attribute pairs

# Types of Attributes

There are different types of attributes

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

# 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

	#doors	Horsepower	weight (kg)	price	length (m)	Final speed (km/h)	0-100m (sec)
Car 1	3	120	1520	15,000	3.10	195	13.7
Car 2	4	210	1660	29,000	4.22	248	8.2
Car 3	5	158	2100	32,500	4.92	210	11.4

# 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

	she	loves	pizza	is	delicious	a	good	person	people	are	the	best
She loves pizza, pizza is delicious	1	1	2	1	1	0	0	0	0	0	0	0
She is a good person	1	0	0	1	0	1	1	1	0	0	0	0
good people are the best	0	0	0	0	0	0	1	0	1	1	1	1



# Transaction Data

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

<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

- 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

```
GGTTC CGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

- Data is a long **ordered** string

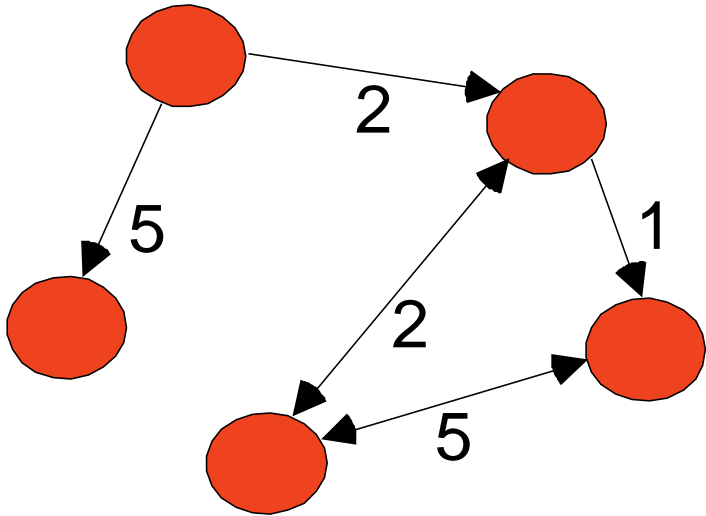
# Ordered Data

- 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?

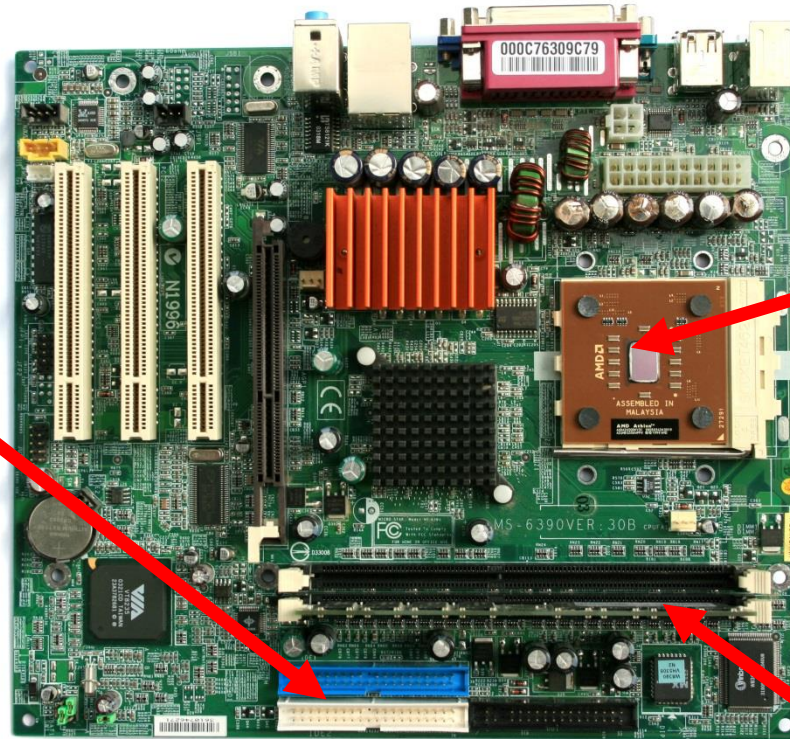
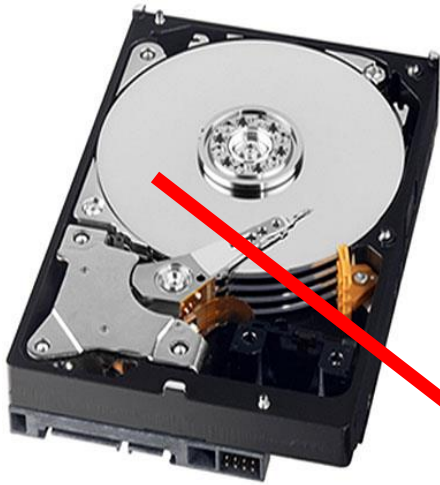
- Suppose you are a stock broker and you observe the fluctuations of multiple stocks over time. What information would you like to get out of your data?





# 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, ...

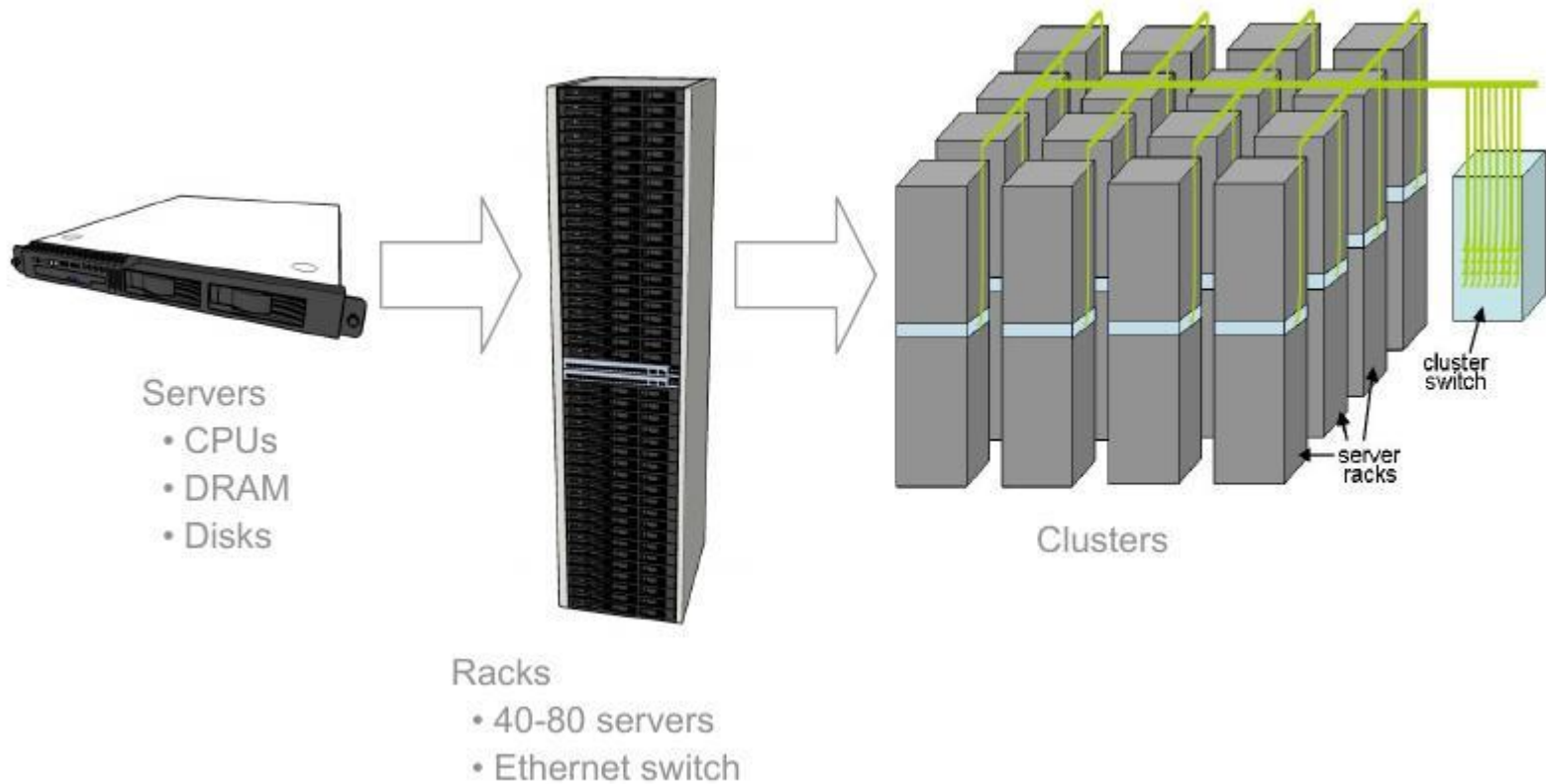




# The Cloud



# The Cloud



# Topics we will cover

- Text mining
- Similarity measures
- Near-neighbor search
- Clustering
- Classification and deep learning
- Feature engineering
- Neural-network embedding
- Graph mining
- Graph neural networks
- Streaming
- Recommender systems
- Social networks
- Models and learning
- Apache Spark
- ...