DATA MINING
INTRO LECTURE

Introduction

Instructors

Aris (Aris Anagnostopoulos)

Teaching Assistant (TA):

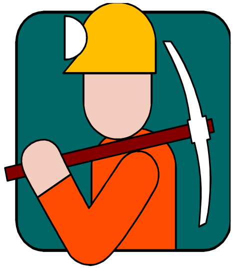
Andrea (Andrea Mastropietro)

Logistics

- Register: **Send email to Aris**
- Web page
- Class hours
- Office hours
- What do you need to know
- Book
- 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.

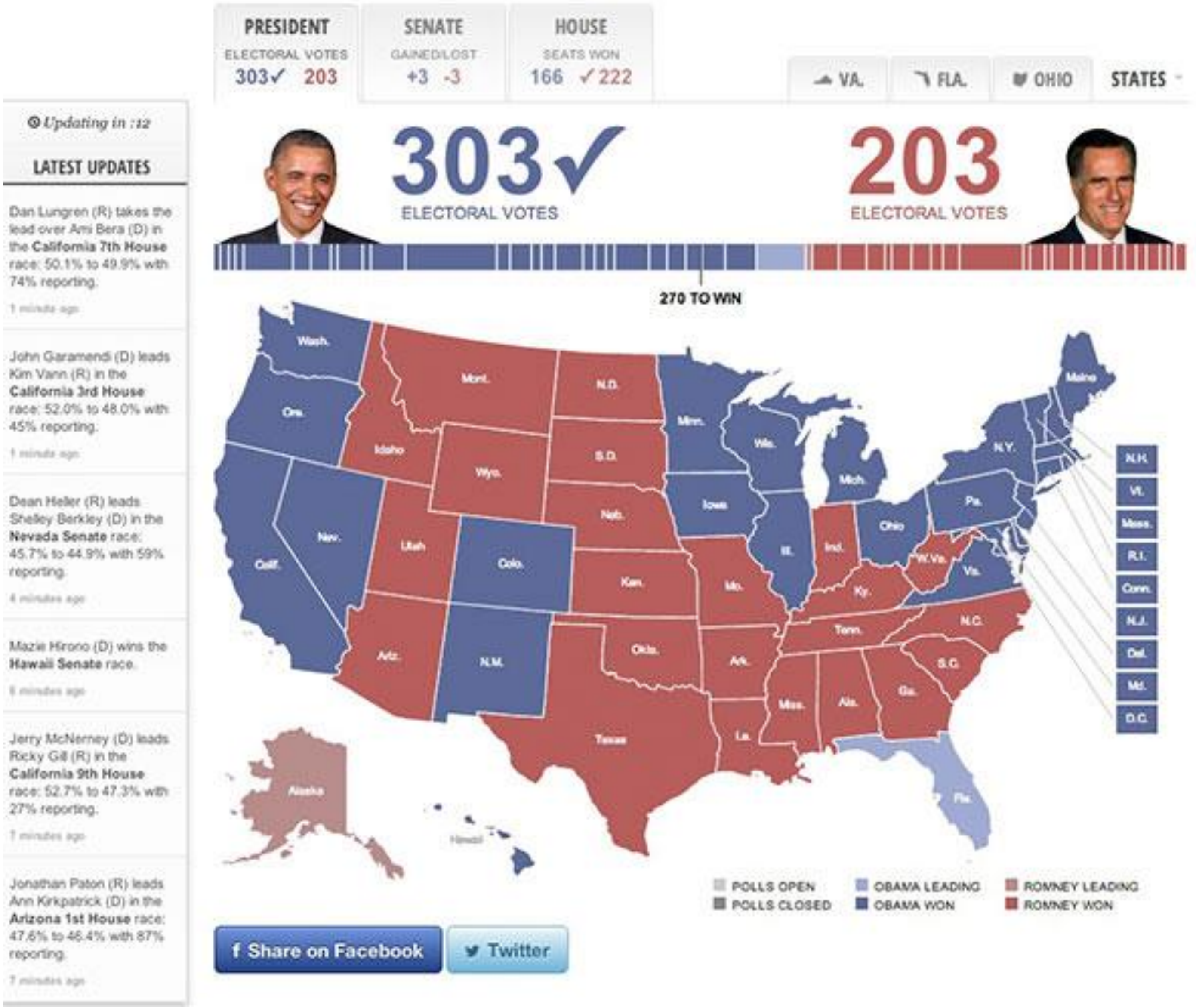
Why do we need data mining?

- 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, marketting

Politics – Nate Silver



Updating in :12

LATEST UPDATES

Dan Lungren (R) takes the lead over Ami Bera (D) in the **California 7th House** race: 50.1% to 49.9% with 74% reporting.

1 minute ago

John Garamendi (D) leads Kim Vann (R) in the **California 3rd House** race: 52.0% to 48.0% with 45% reporting.

1 minute ago

Dean Heller (R) leads Shelley Berkley (D) in the **Nevada Senate** race: 45.7% to 44.9% with 59% reporting.

4 minutes ago

Mazie Hirono (D) wins the **Hawaii Senate** race.

6 minutes ago

Jerry McNerney (D) leads Ricky Gill (R) in the **California 9th House** race: 52.7% to 47.3% with 27% reporting.

7 minutes ago

Jonathan Paton (R) leads Ann Kirkpatrick (D) in the **Arizona 1st House** race: 47.6% to 46.4% with 87% reporting.

7 minutes ago

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.

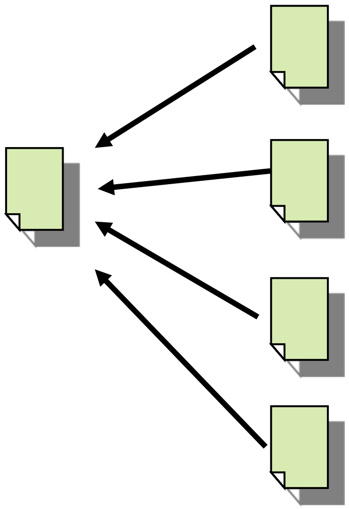
You watch youtube videos, it will recommend others.

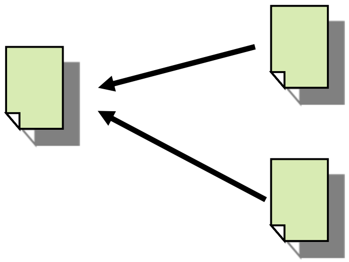
You make a google query, it will propose others.

How do they do it?

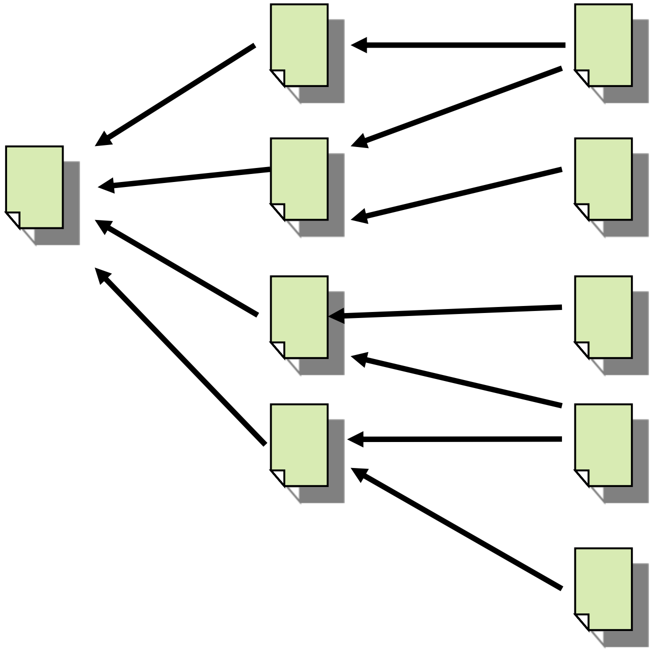
(They analyze what previous **similar** users have done!)

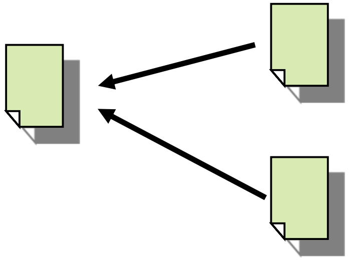
Google and PageRank



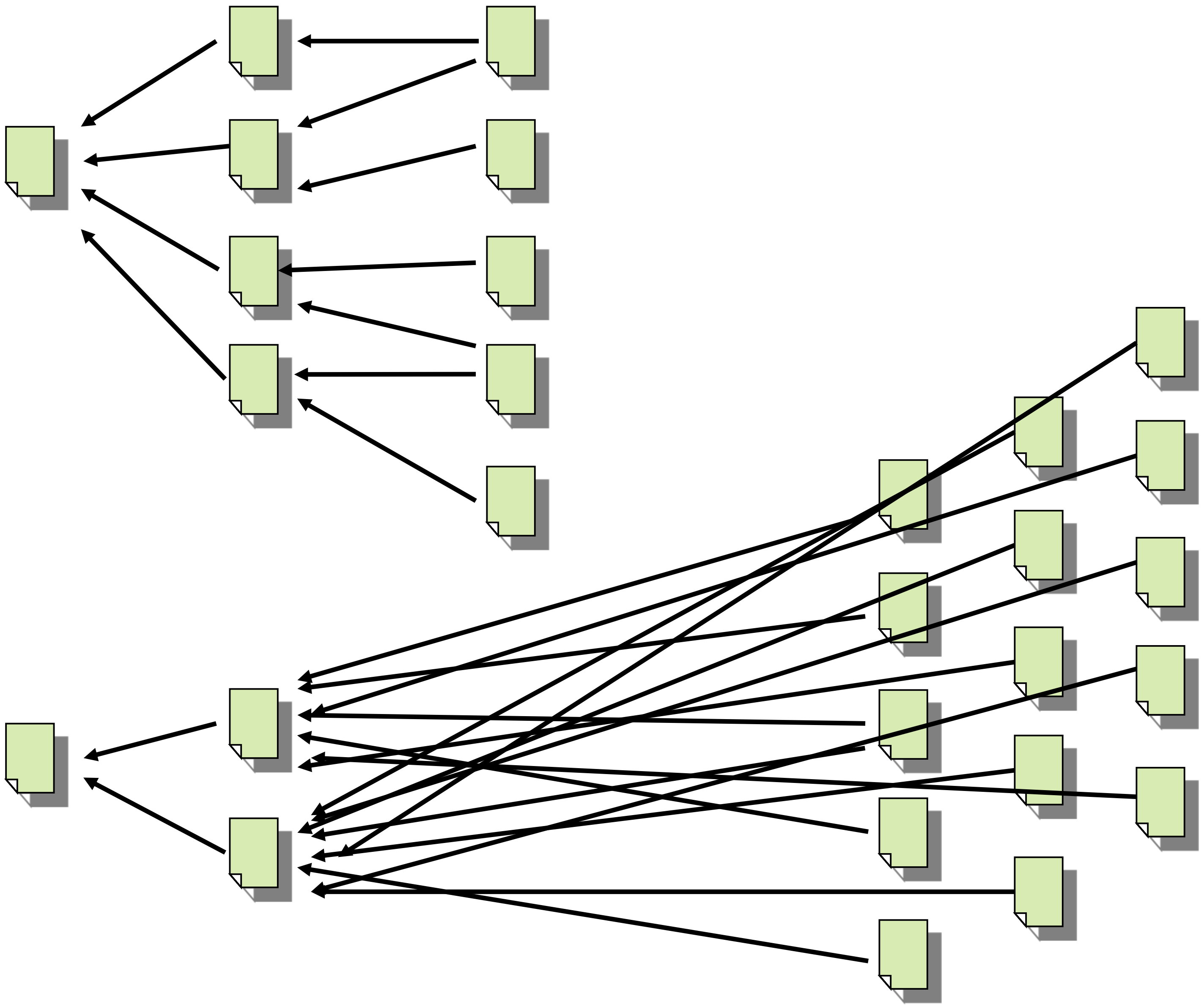


Google and PageRank



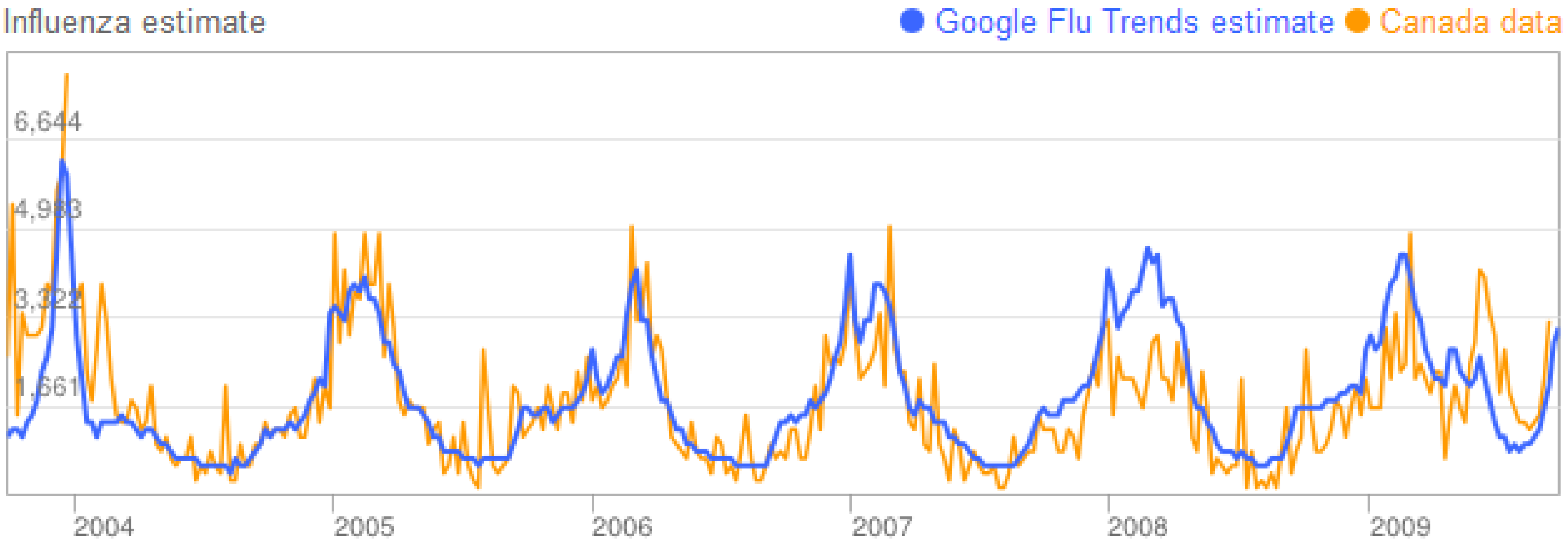


Google and PageRank



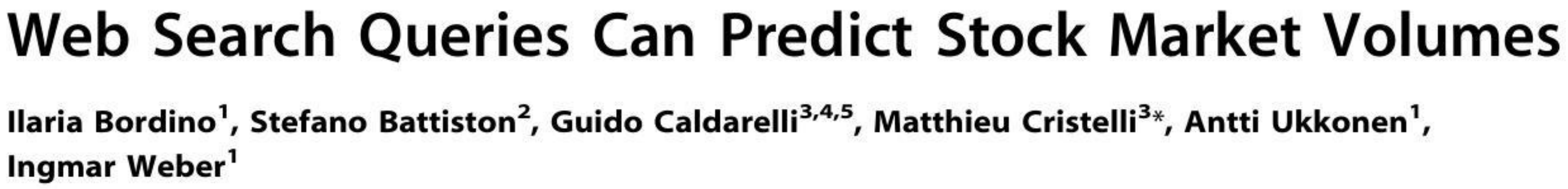
Google flu

Canada Flu Activity



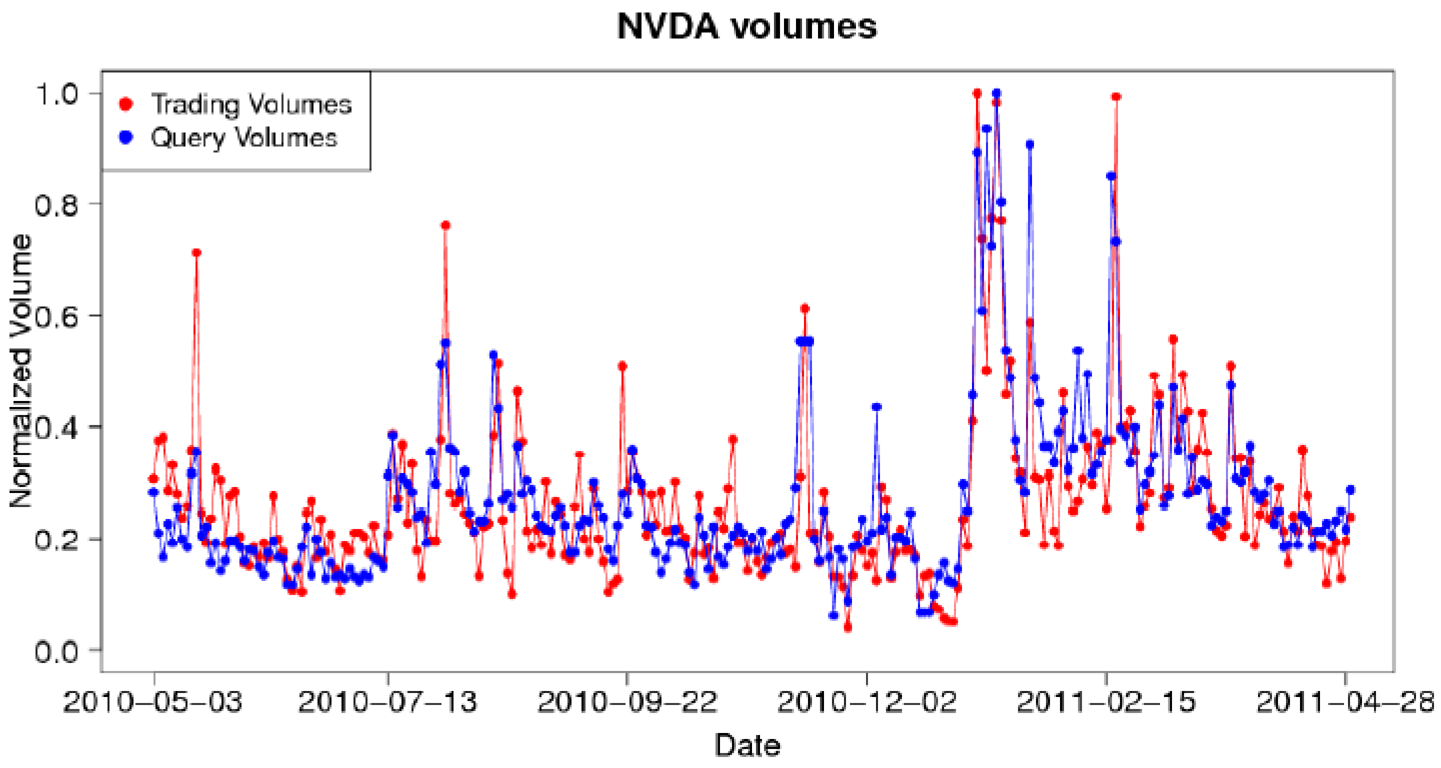
Canada: Influenza-like illness (ILI) data provided publicly by the Public Health Agency of Canada.

Google and stockmarket

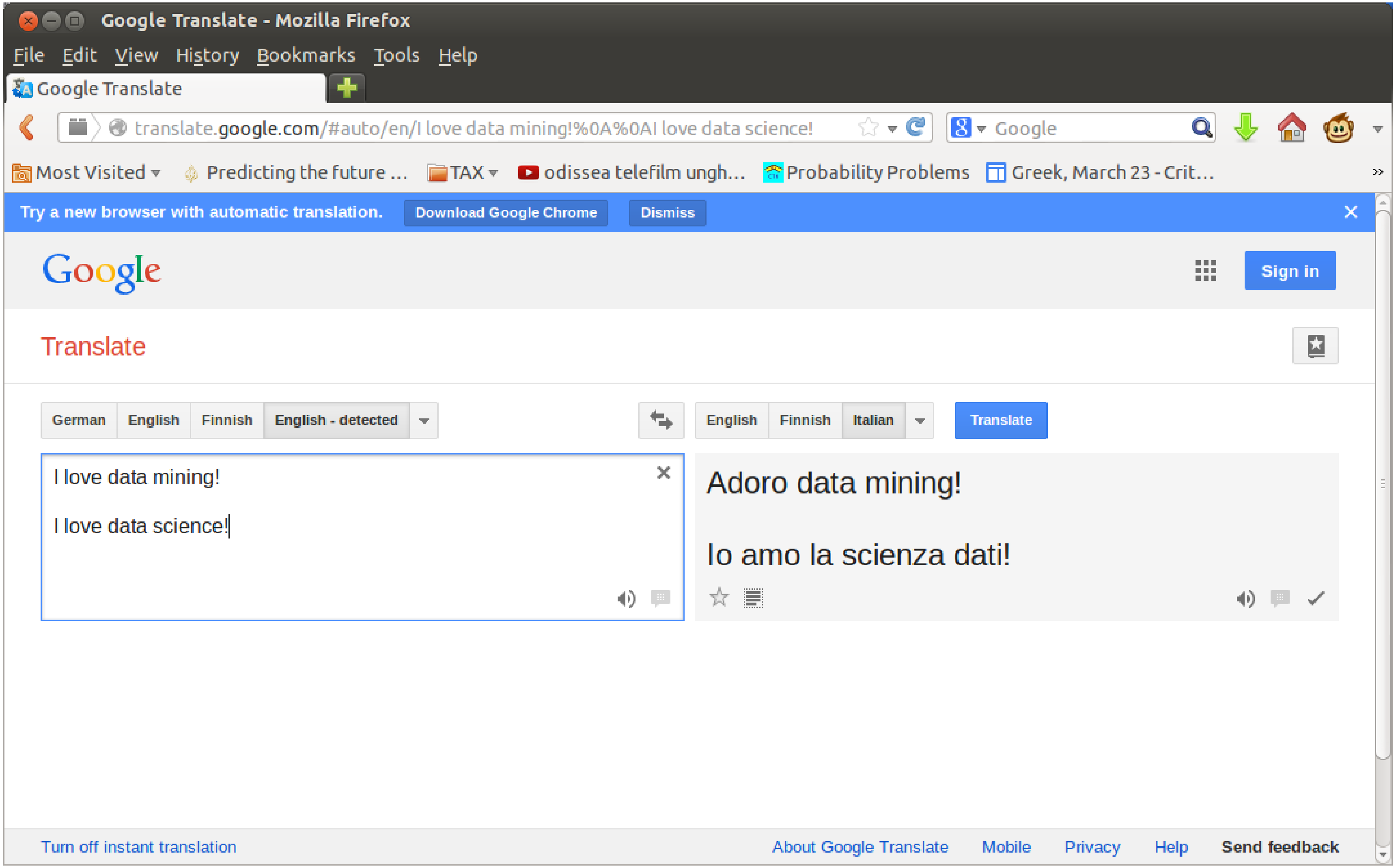


Web Search Queries Can Predict Stock Market Volumes

Ilaria Bordino[1], Stefano Battiston[2], Guido Caldarelli[3,4,5], Matthieu Cristelli[3]*, Antti Ukkonen[1], Ingmar Weber[1]



Google translate



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!

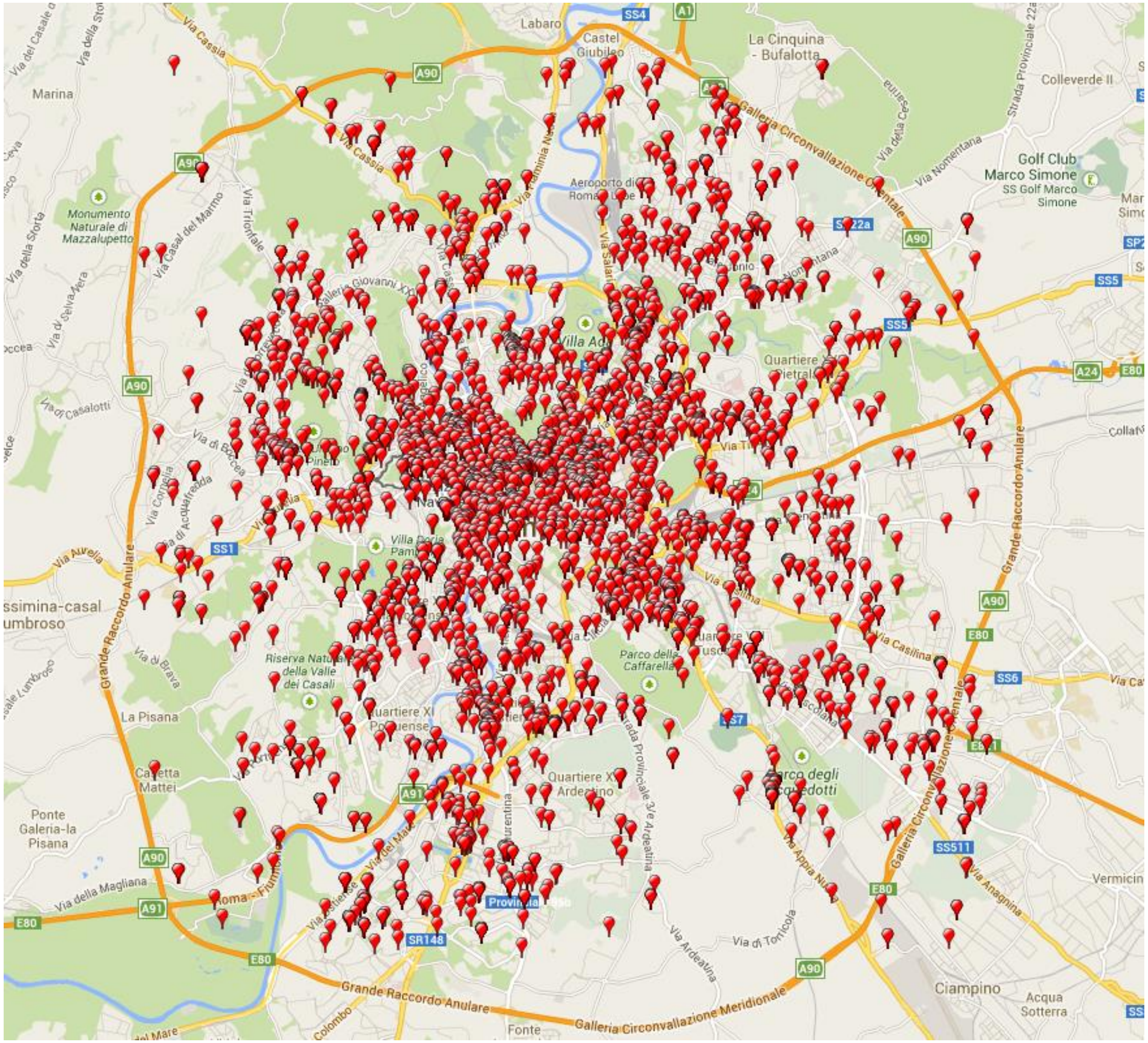
I love data science!

Adoro data mining!

Io amo la scienza dati!

Turn off instant translation

About Google Translate Mobile Privacy Help **Send feedback** twitter

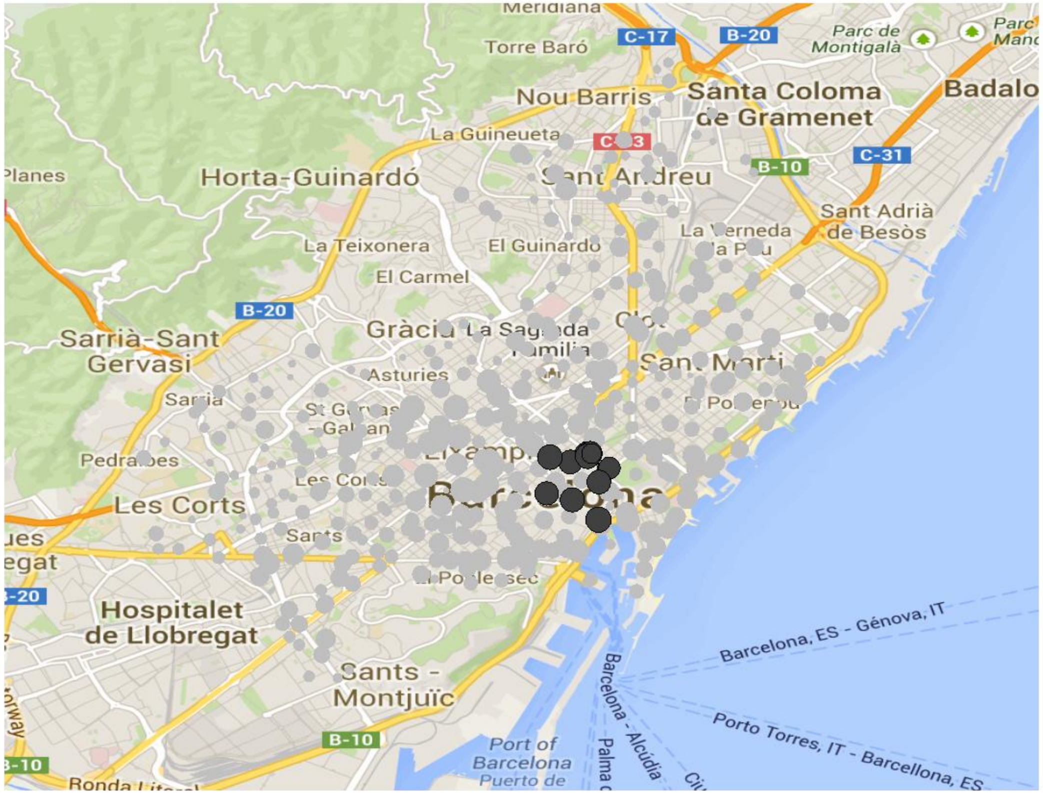

Labaro
Castel Giubileo
La Cinquina - Bufalotta
Colleverde II
Marina
Golf Club Marco Simone
SS Golf Marco Simone
Monumento Naturale di Mazzalupetto
Aeroporto di Roma Urbe
Via Cassia
Via Trionfale
Via Salaria
Via Nomentana
Galleria Circonvallazione Orientale
Villa Ada
Quartiere Pietralata
Via Tiburtina
Via Aurelia
Villa Doria Pamphili
Via Casilina
Parco della Caffarella
Riserva Naturale della Valle dei Casali
Quartiere XI Portuense
Quartiere XX Ardeatino
Parco degli Acquedotti
La Pisana
Casetta Mattei
Ponte Galeria-la Pisana
Via della Magliana
Roma - Fiumicino
Via Appia Nuova
Via Anagnina
Via Ardeatina
Via di Torricola
Vermicino
Ciampino
Acqua Sotterra
Grande Raccordo Anulare
Galleria Circonvallazione Meridionale
Fonte
Collatina
ssimina-casal umbroso
A90
A91
A1
A24
E80
SS1
SS4
SS5
SS6
SS7
SS511
SR148
SS22a
SP2
Provinciale

twitter

- 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

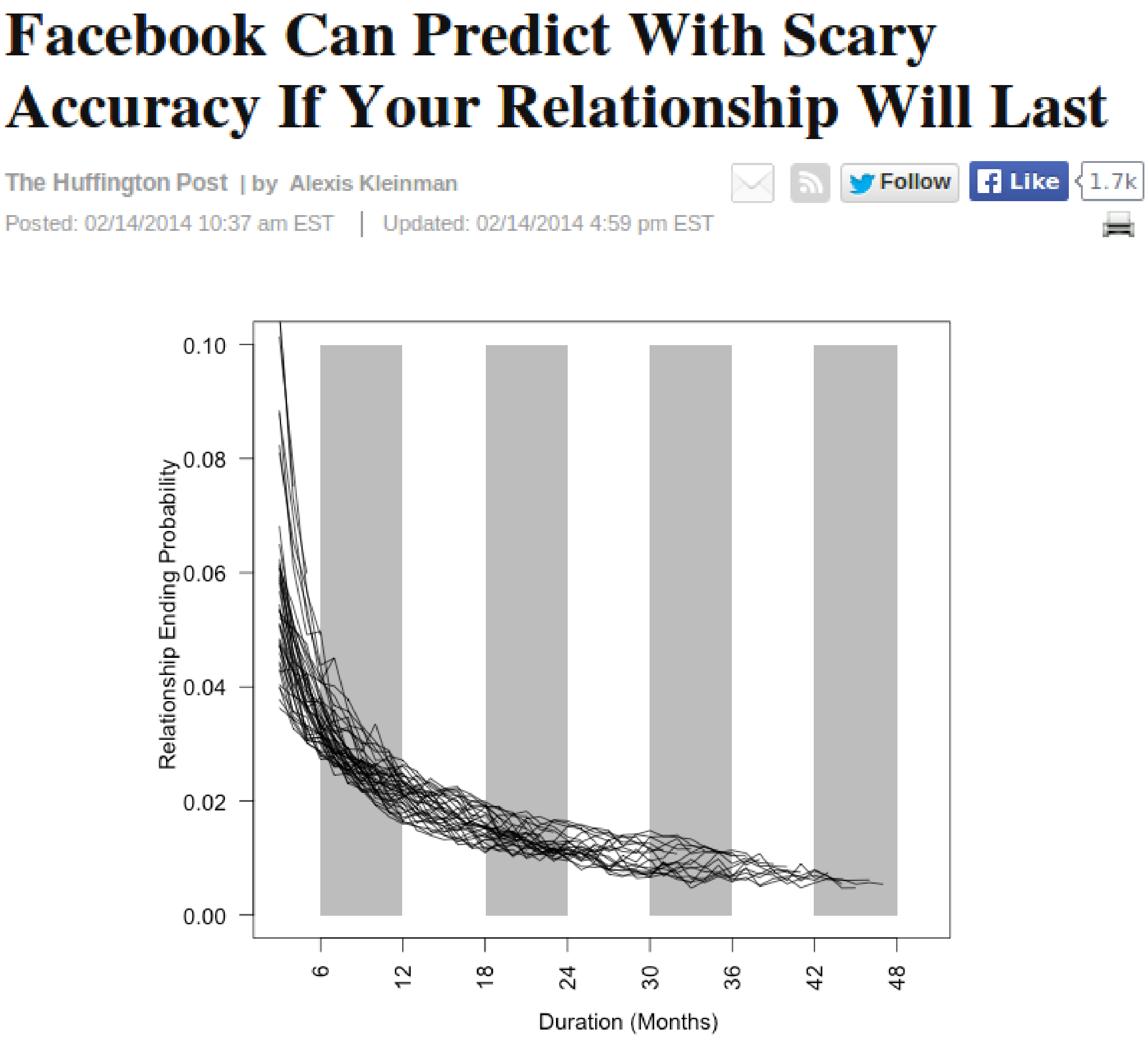


Psychology and Sociology

- 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

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

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 though 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

- http://www.1000genomes.org
  - 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

- 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

Attributes

Objects

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**

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

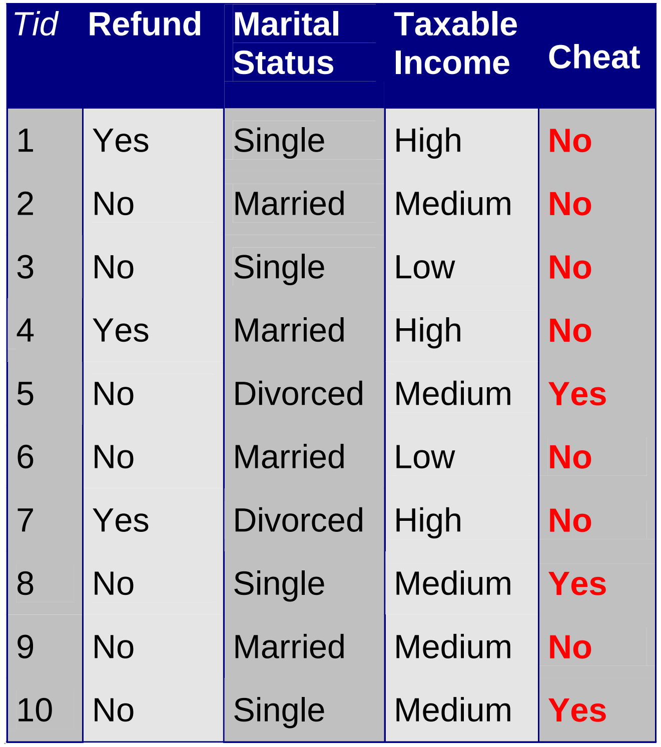
- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- 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

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



Tid	**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**

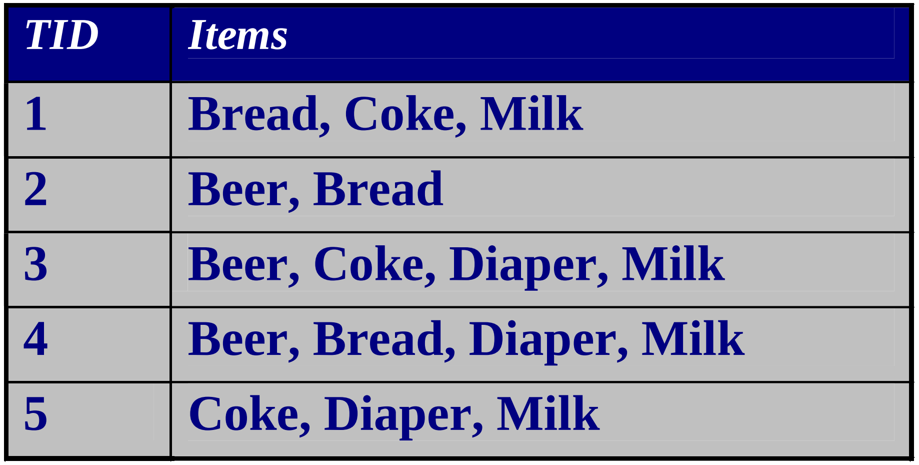
Document Data

- 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.



TID	*Items*
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

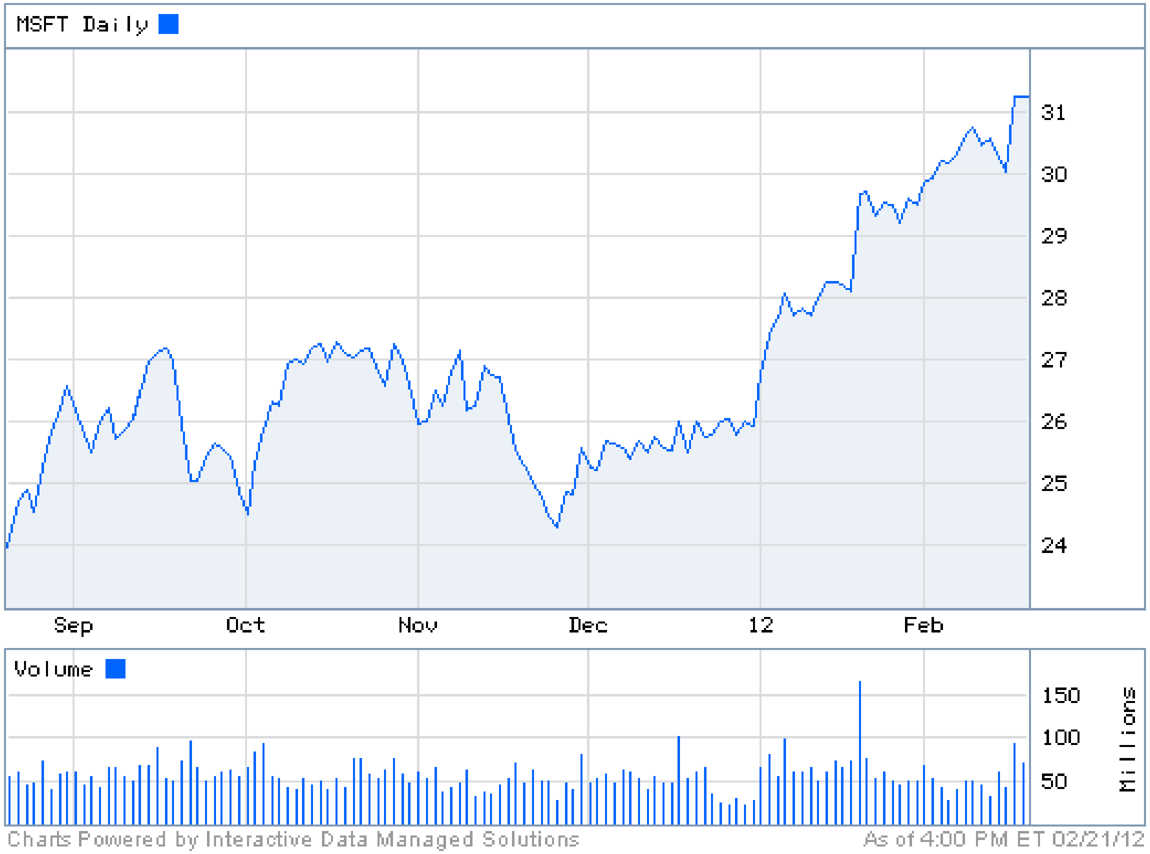
- Genomic sequence data

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

- Data is a long ordered string

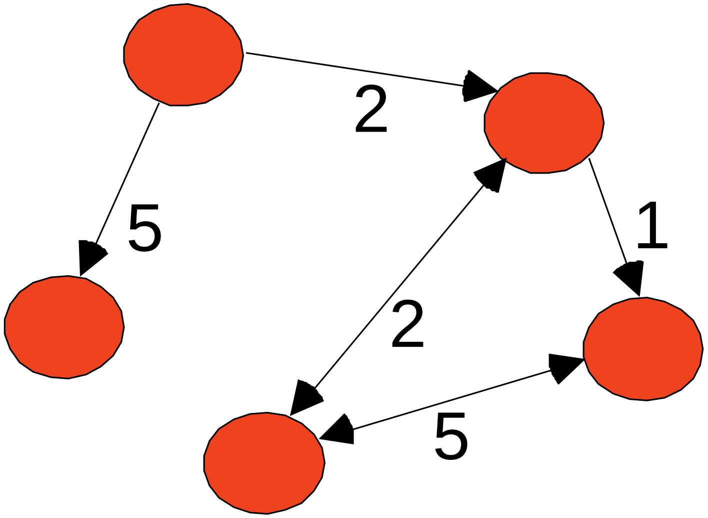
Ordered Data

- 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?

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

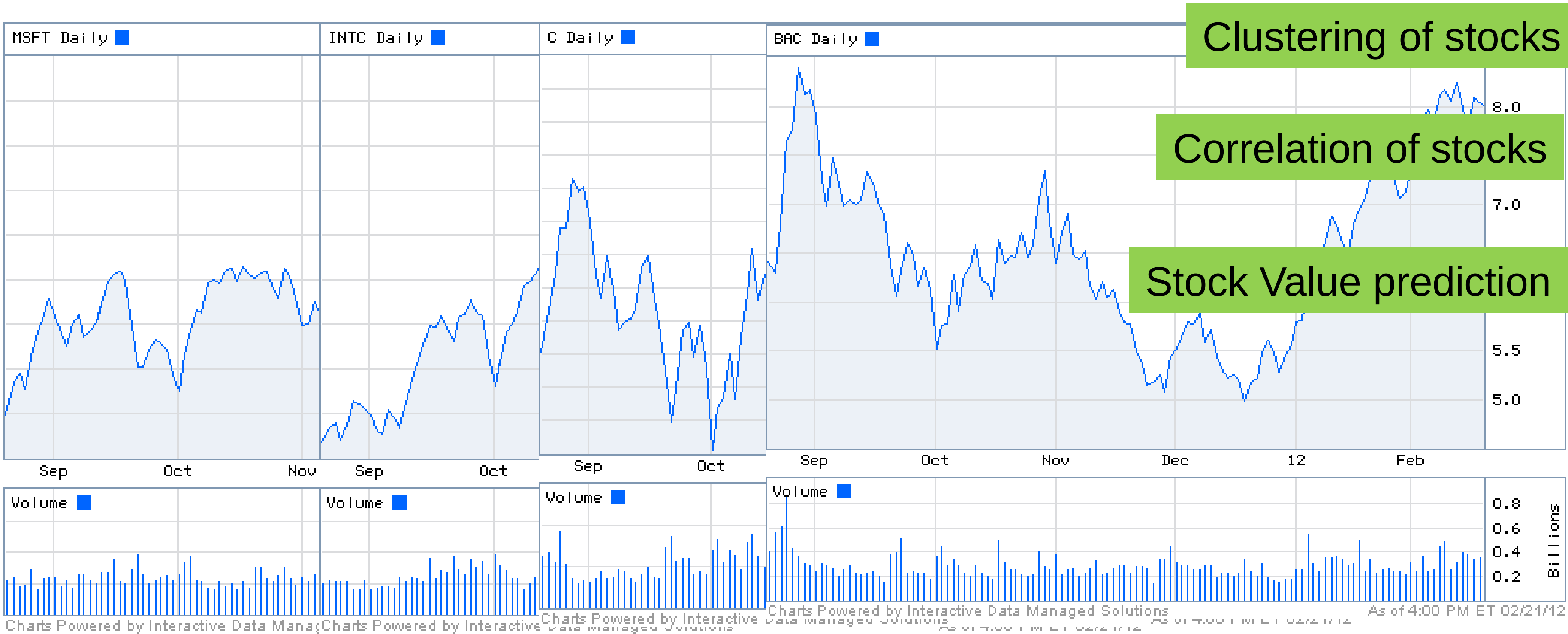
each with a user id and a timestamp. What information would you like to get our of the data?

Ad click prediction

Query reformulations

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

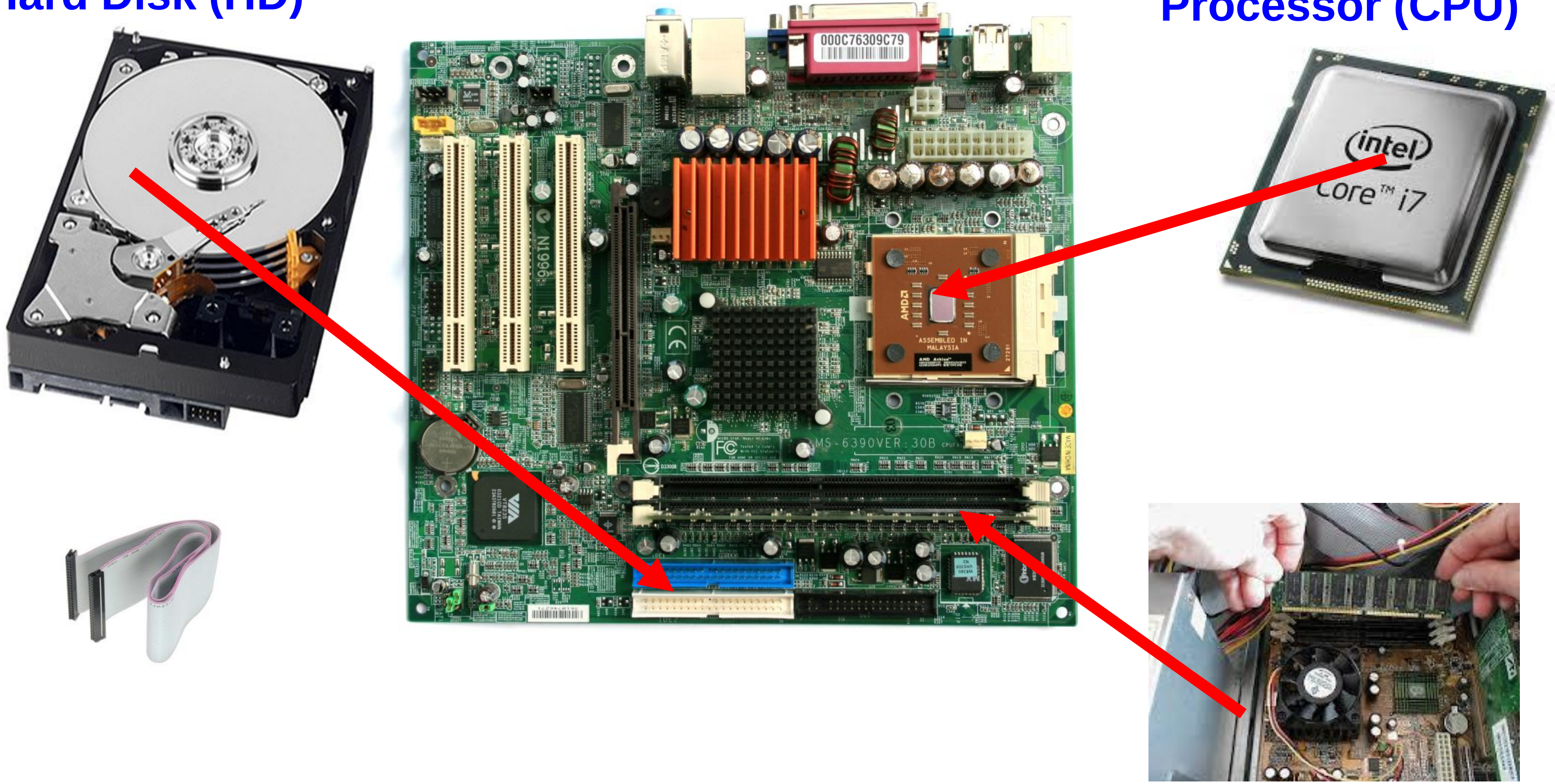
Correlation of stocks

Stock Value prediction

Basics of Computer Architecture

Hard Disk (HD)

Processor (CPU)



Memory (RAM)

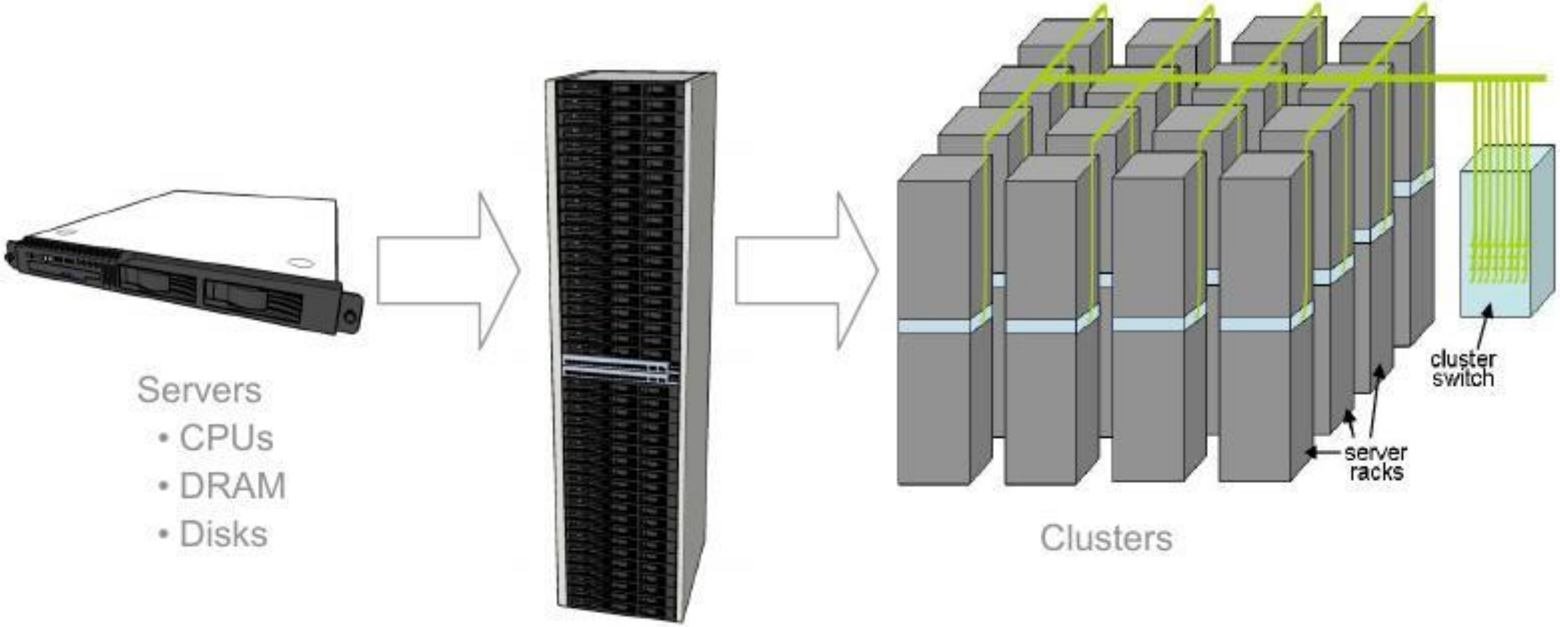
The Cloud

There exist large datacenters for storing data and making computations

- Gmail, dropbox, …

The Cloud

The Cloud



Servers

- CPUs
- DRAM
- Disks

Racks

- 40-80 servers
- Ethernet switch

Clusters

Topics we will cover

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