Analysis of Massive Data Sets

http://www.fer.hr/predmet/avsp

Prof. dr. sc. Siniša Srbljić

Doc. dr. sc. Dejan Škvorc

Doc. dr. sc. Ante Đerek

Faculty of Electrical Engineering and Computing
Consumer Computing Laboratory

Analysis of Massive Data Sets: Introduction Lecture

Marin Šilić, PhD

Overview

- Motivation
- Big Data
- Course Content
- Course Organization
- Mozgalo Contest

Motivation

- **Politics – managing campaigns**
  - Why do campaigns need data?
    - The available resources (time, money, volunteers) are limited
    - The most valuable information – acquire the list of citizens to contact
      - Potential voters, volunteers and donors
      - Which citizens engage in specific campaign supporting actions
    - Campaigns use data to build predictive models
      - Behavior scores
      - Support scores
      - Responsiveness scores

Nickerson, David W., and Todd Rogers. "Political Campaigns and Big Data."
HKS Faculty Research Working Paper Series RWP13-045, Revised February 2014.

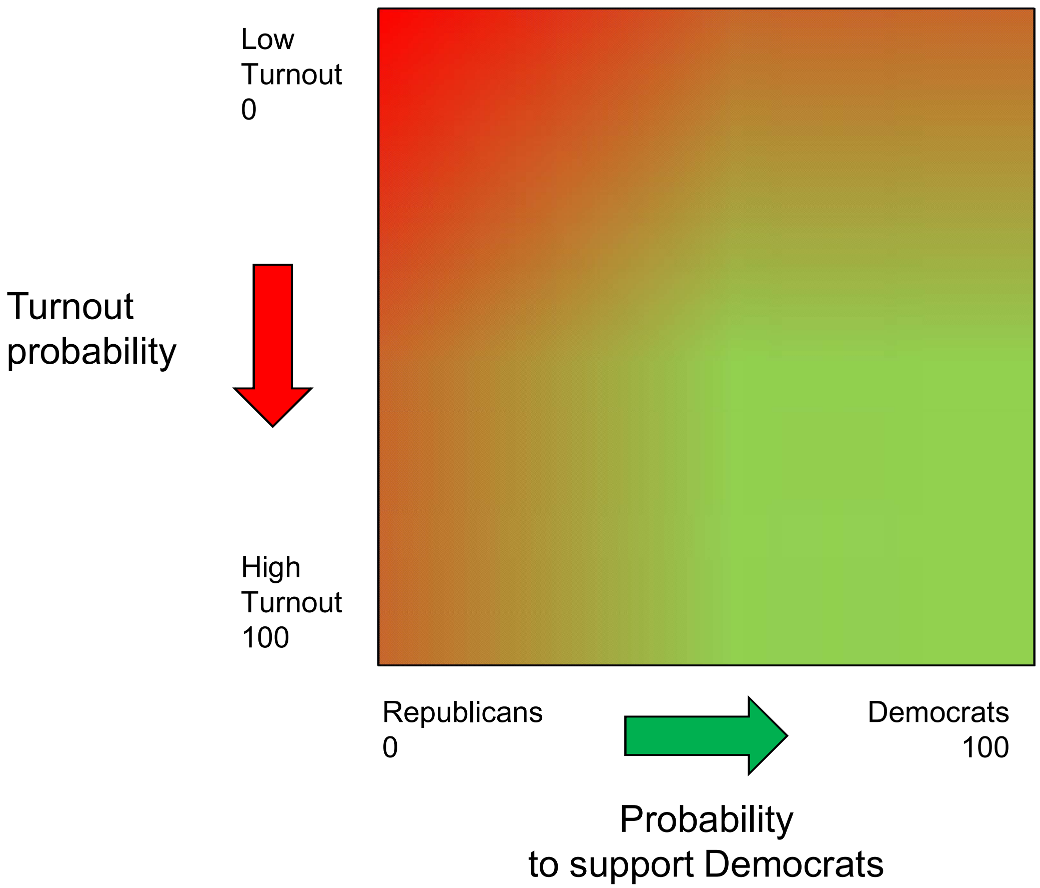
Motivation

- **Politics – managing campaigns**
  - Where does data come from?
    - The goal is to integrate various data sources
      - Digital communications, field operations, canvassing, phone calls, volunteer recruitment, fundraising, etc.
    - The most heralded success of Obama 2012 campaign was the creation of *Narwhal*
      - Program that merged data collected from various sources
      - The re-election campaign began with a 10TB database and ended up with 50TB of data by the end of the election.
    - Voter database
      - Directly: age, gender, address, phone number, voting history
      - Indirectly: census, education, ethic distribution, household details

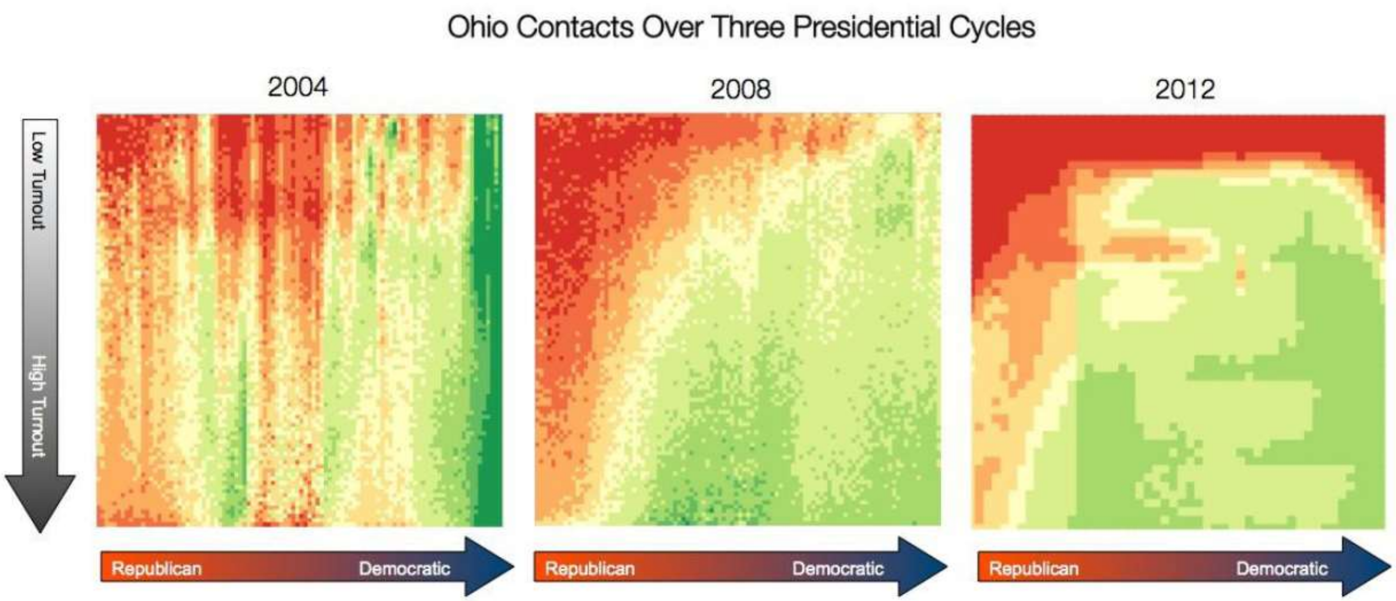
Nickerson, David W., and Todd Rogers. "Political Campaigns and Big Data."
HKS Faculty Research Working Paper Series RWP13-045, Revised February 2014.

Motivation

- Politics – managing campaigns



- **Politics – managing campaigns**

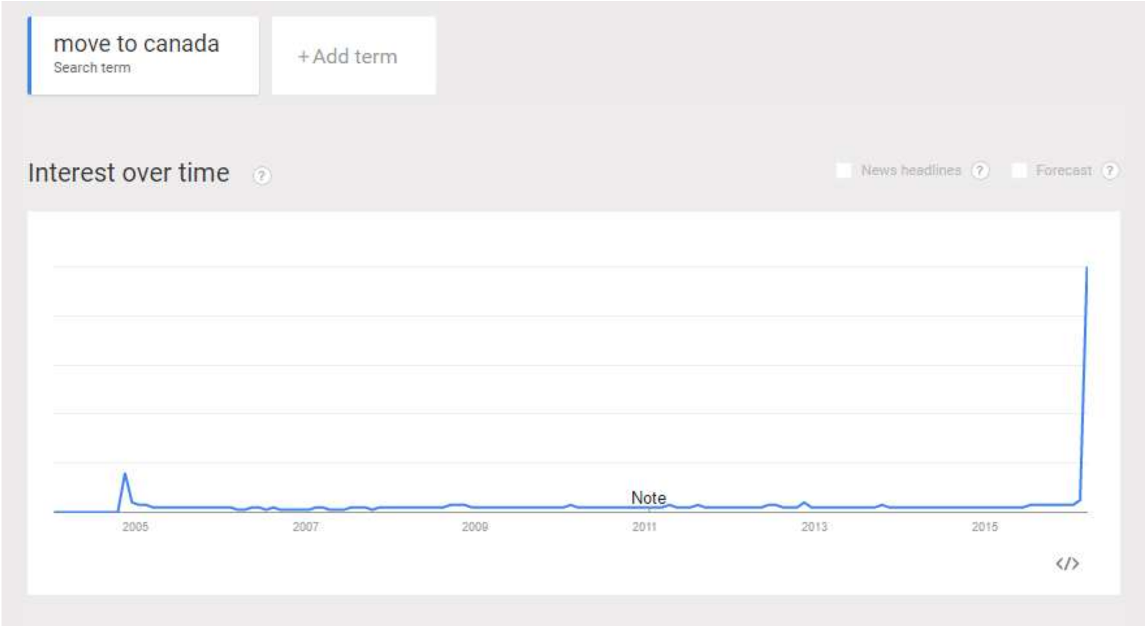


Source: Catalist, LLC
X-axis is likelihood of supporting a Democratic candidate over a Republican candidate, ranging from 0 (left) to 100 (right).
Y-axis is likelihood of voting ranging, ranging from 100 (low) to 0 (high).
Colors represent density/frequency of direct contacts from all Catalist clients over the course of the entire election cycle. Dark red means these citizens received the fewest direct contacts over the election cycle, and dark green means these citizens received the most direct contacts over the election cycle.

Nickerson, David W., and Todd Rogers. "Political Campaigns and Big Data." HKS Faculty Research Working Paper Series RWP13-045, Revised February 2014.

Motivation

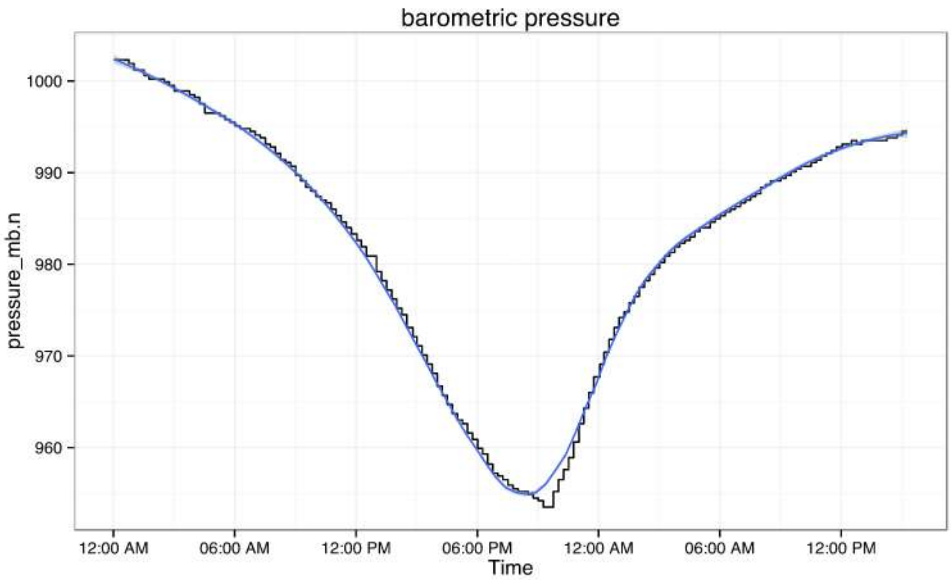
- **Politics – Google Trends**
  - Presidential elections primaries 2016 in US
    - Super Tuesday, March 5 2016



    - Donald Trump won 7/11 states and 256/600 delegates for Republican primaries

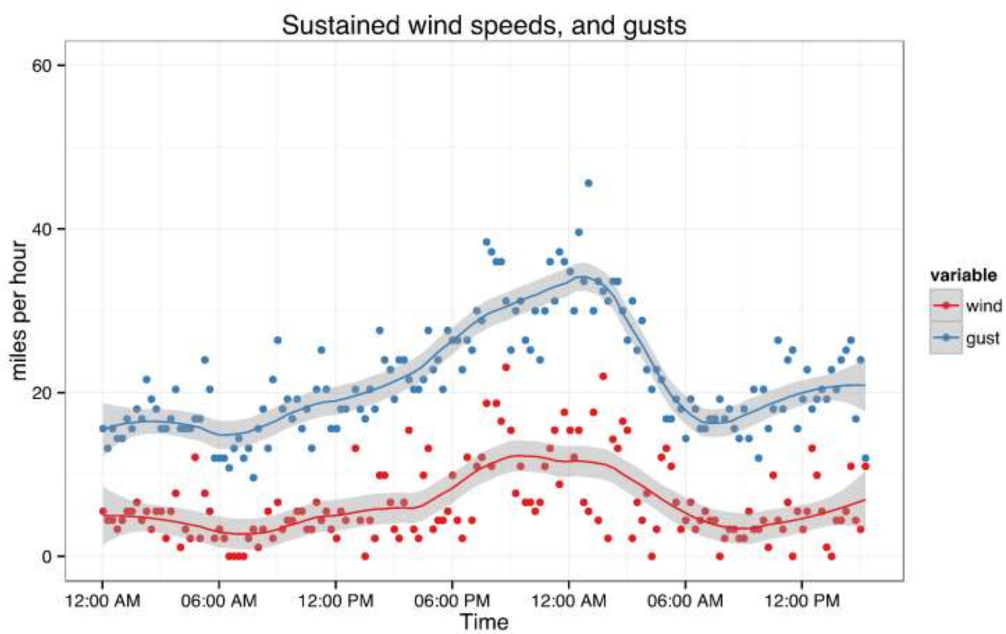
Motivation

- **Weather forecast (natural catastrophes)**
  - Hurricane Sandy east coast of the US
    - Visualization of atmospheric data in real time predicted the dynamics of hurricane movement
    - https://rpubs.com/JoFrhwld/sandy

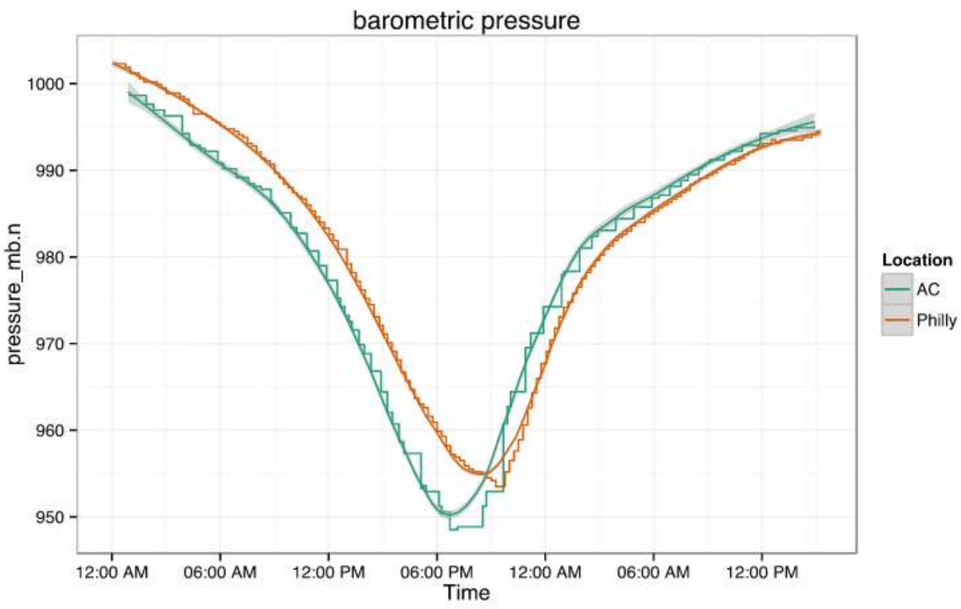


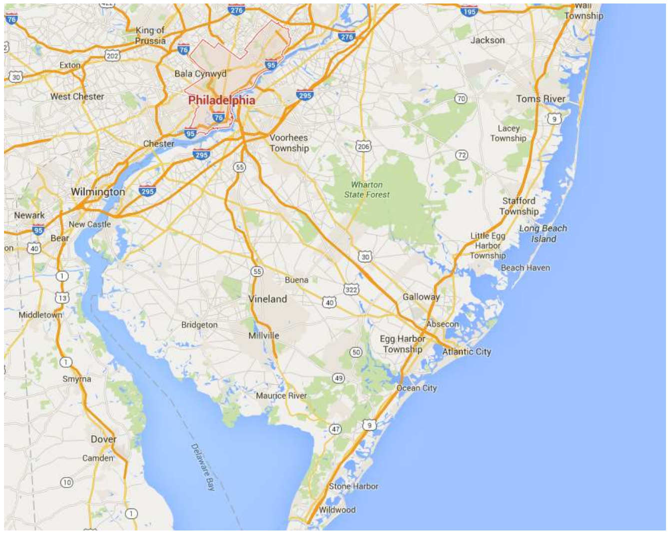
Motivation

- **Weather forecast (natural catastrophes)**
  - Hurricane Sandy east coast of the US
    - Visualization of atmospheric data in real time predicted the dynamics of hurricane movement
    - https://rpubs.com/JoFrhwld/sandy



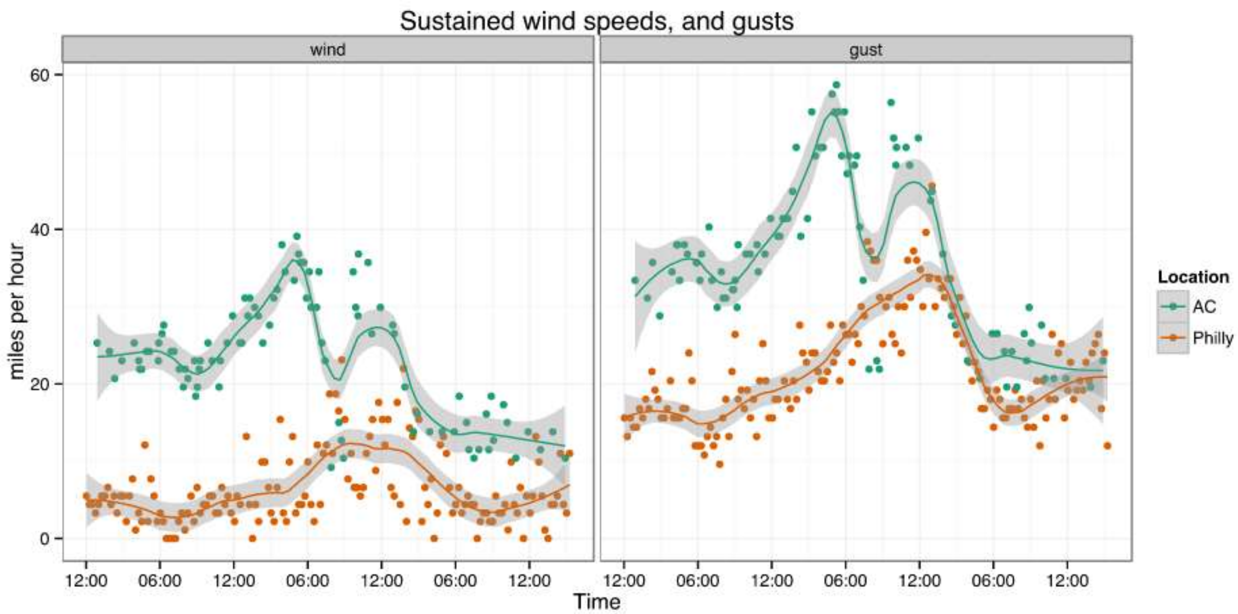
- **Weather forecast (natural catastrophes)**
  - Hurricane Sandy east coast of the US
    - Visualization of atmospheric data in real time predicted the dynamics of hurricane movement
    - https://rpubs.com/JoFrhwld/sandy

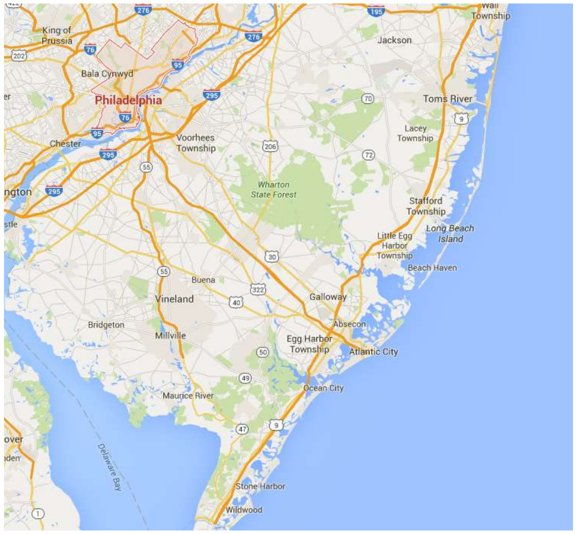




Motivation

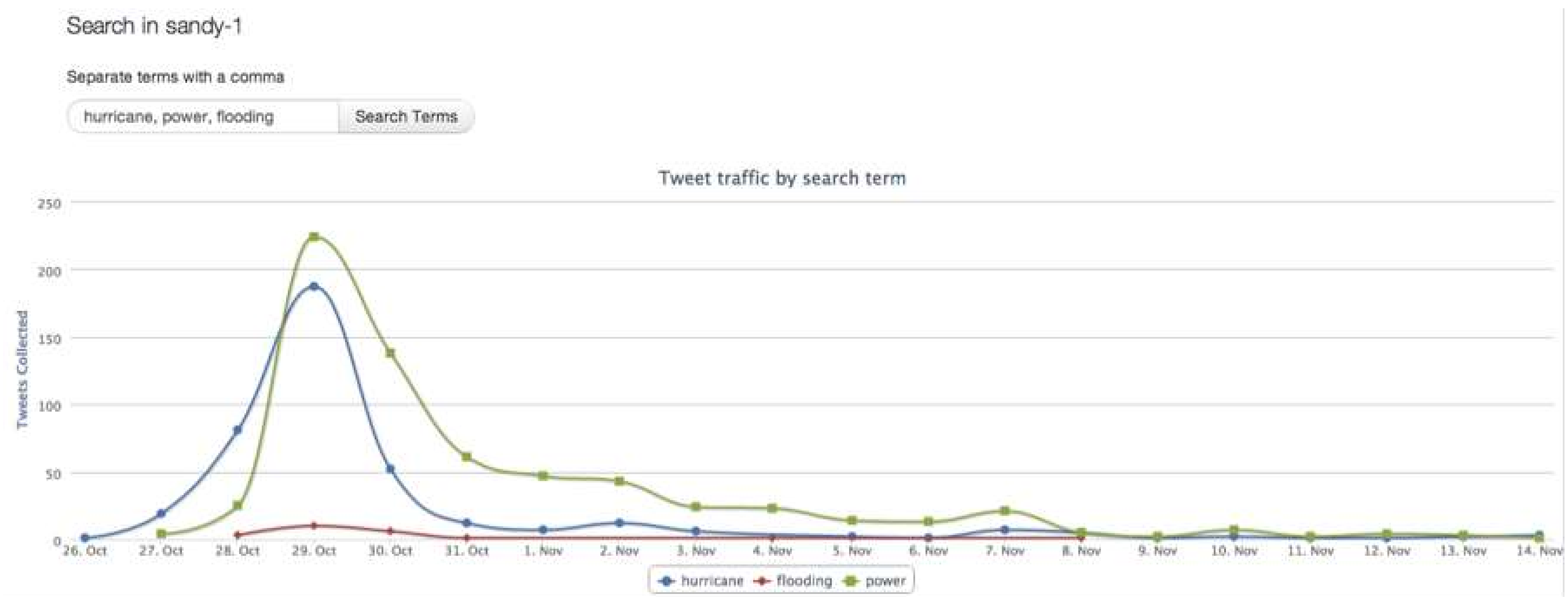
- **Weather forecast (natural catastrophes)**
  - Hurricane Sandy east coast of the US
    - Visualization of atmospheric data in real time predicted the dynamics of hurricane movement
    - https://rpubs.com/JoFrhwld/sandy





Motivation

- **Weather forecast (natural catastrophes)**
  - Hurricane Sandy east coast of the US
    - The analysis of power outage reports on *twitter*
    - The correlation with *tweets* about hurricane and floods



Motivation

- **Weather forecast (natural catastrophes)**
  - Hurricane Sandy east coast of the US
    - The analysis of power outage reports on *twitter*
    - The correlation with *tweets* about hurricane and floods
    - Visualizations of tweets on a map in timeline
      - http://blog.echen.me/hurricane-sandy-outages/
      - https://www.youtube.com/watch?v=iVIvDzC5Wwc

Motivation

- **Digital Humanities**
  - The analysis of emotions in books written in English language through 20th century
    - How did the emotions in books change through 20th century?
      1. The authors used Google's n-gram database
         - The database presents a sample of 4% digitally scanned books through several centuries ~5M books
         - It contains the information how many times each n-gram is used in a given year
      - 1-gram is a string of characters uninterrupted by space (word, number etc.)
      - n-gram is a set of *n* 1-grams

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604170/
https://wordnet.princeton.edu/wordnet/

Motivation

- **Digital Humanities**
  - The analysis of emotions in books written in English language through 20th century
    - How did the emotions in books change through 20th century?
      2. Apparently, there are so called, “mood” words that contain intrinsic emotional character
         - » The authors identified 6 distinct collections (WordNet!) of words that are associated to emotions *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise* respectively

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604170/
https://wordnet.princeton.edu/wordnet/

Motivation

- **Digital Humanities**
  - The analysis of emotions in books written in English language through 20th century
    - How did the emotions in books change through 20th century?
      3. For each emotion and for each given year, the authors computed the mood score
         - » It is an average normalized frequency of occurrence of 1-grams that belong to a given emotion for a given year

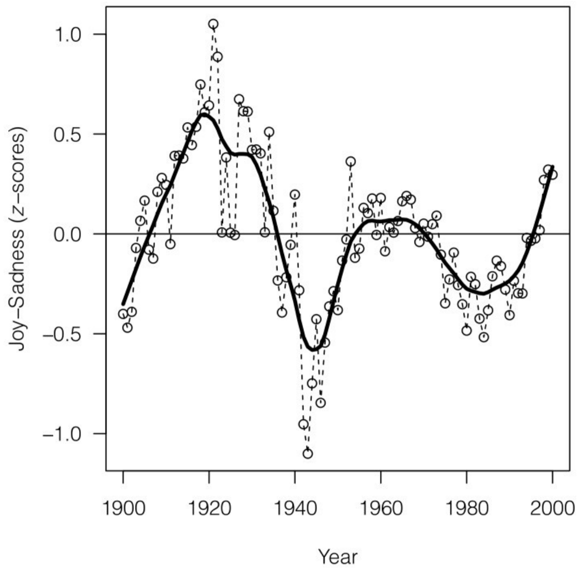
$$M_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{the}}$$

         - » To compare different moods, after computing mood scores for the entire set of years, they converted them to z-scores

$$M_{ZY} = \frac{M_Y - \mu_M}{\sigma_M},$$ where μ_M and σ_M denote mean and standard deviation

Motivation

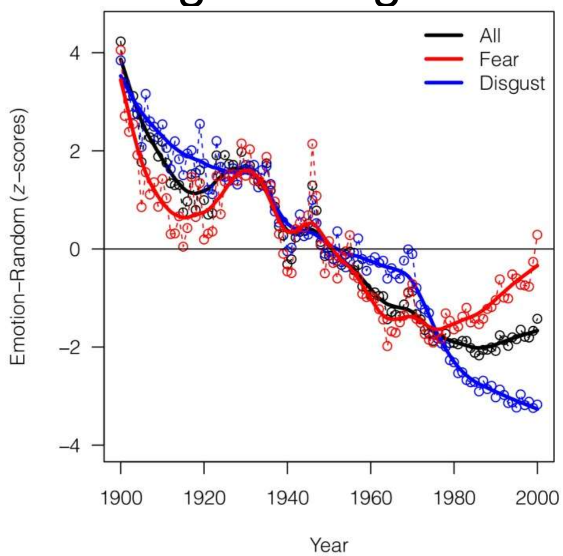
- **Digital Humanities**
  - The analysis of emotions in books written in English language through 20th century
    - How did the emotions in books change through 20th century?
      - Results
        - » Sad peaks in 40s and 70s
        - » Happy peaks in 20s and 60s and also in recent years



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604170/
https://wordnet.princeton.edu/wordnet/

Motivation

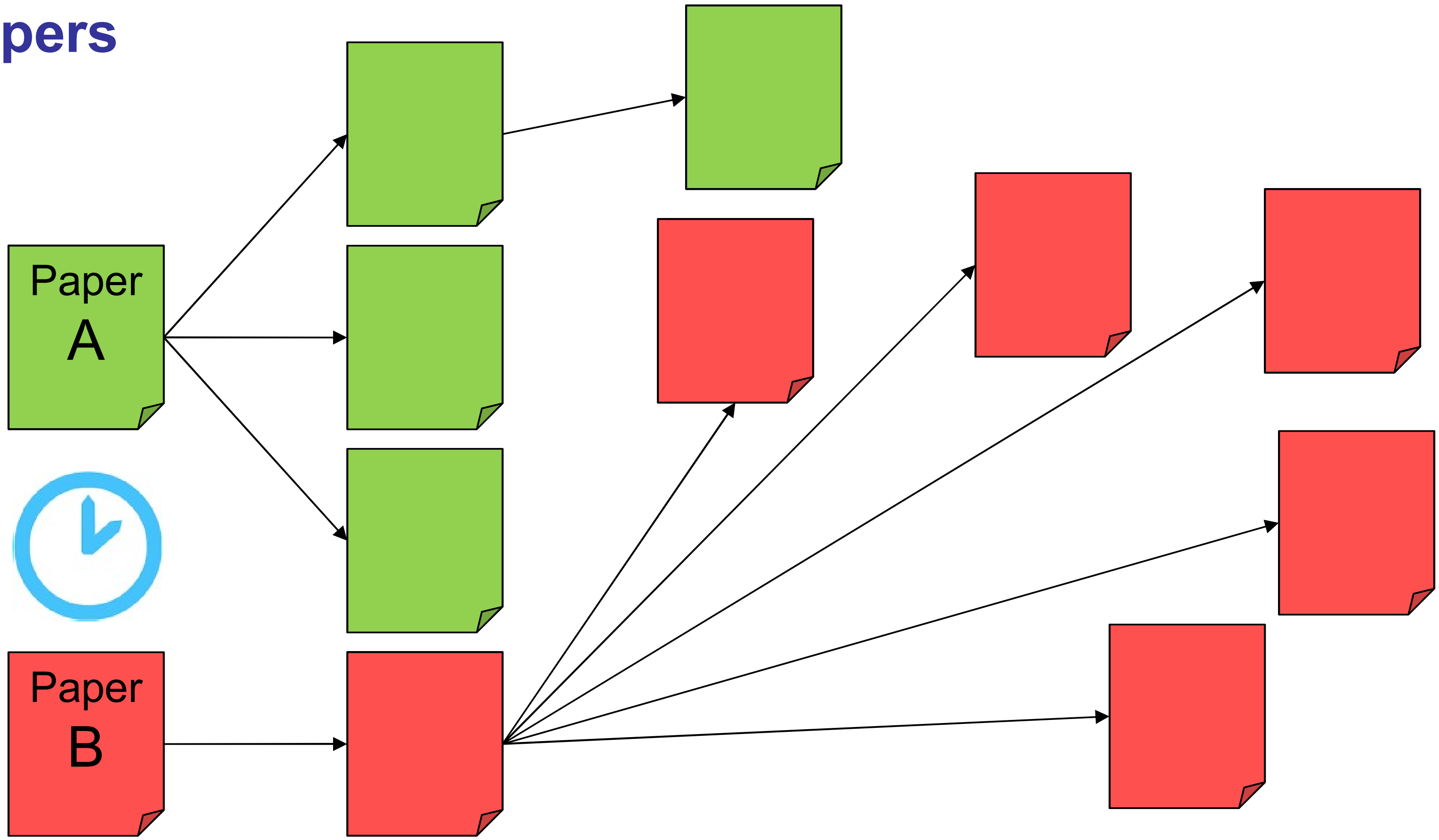
- **Digital Humanities**
  - The analysis of emotions in books written in English language through 20th century
    - How did the emotions in books change through 20th century?
      - Results
        - » Decrease in the use of emotions-related words through time



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604170/
https://wordnet.princeton.edu/wordnet/

Motivation

- Bibliometric – measuring relevance of scientific papers



Motivation

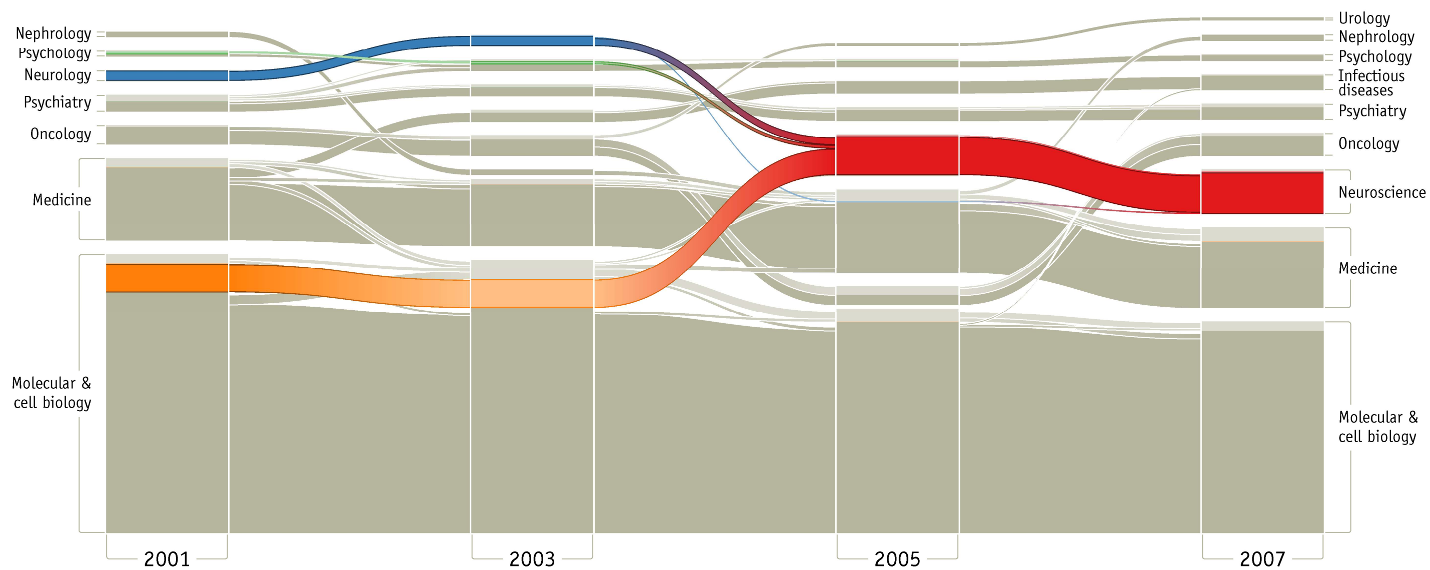
- **Bibliometric – measuring relevance of scientific papers**
  - How to decide which paper is more relevant?
  - This problem is similar to the problem of important pages on the web
    - We can say a webpage is important if a lot of important pages point to it
    - Similarly, we can say a particular paper is relevant if a lot of relevant papers cite it
    - In Google they proposed a method to evaluate the importance of pages on the web

Motivation

- **Bibliometric – measuring relevance of scientific papers**
  - How to decide which paper is more relevant?
  - This problem is similar to the problem of important pages on the web
    - It is called Page rank algorithm, it is a significant part of Google's success
      1. For each node you add up weights of its neighbors (nodes it links to) and assigns them to the node
      2. Then you pass on that weight to every node that links to it
      3. You keeps doing this until some convergence condition is reached
      4. Finally, you end up with the relative importance of these nodes (pages, papers, it can be apply to any graph)

Motivation

- **Clustering scientific papers**
  - Clustering scientific papers in clusters according to their field in timeline
    - Visualization of clusters can show new scientific disciplines emerging



[Mapping Change in Large Networks](#)
Rosvall M, Bergstrom CT (2010) Mapping Change in Large Networks. PLoS ONE 5(1): e8694. doi: 10.1371/journal.pone.0008694

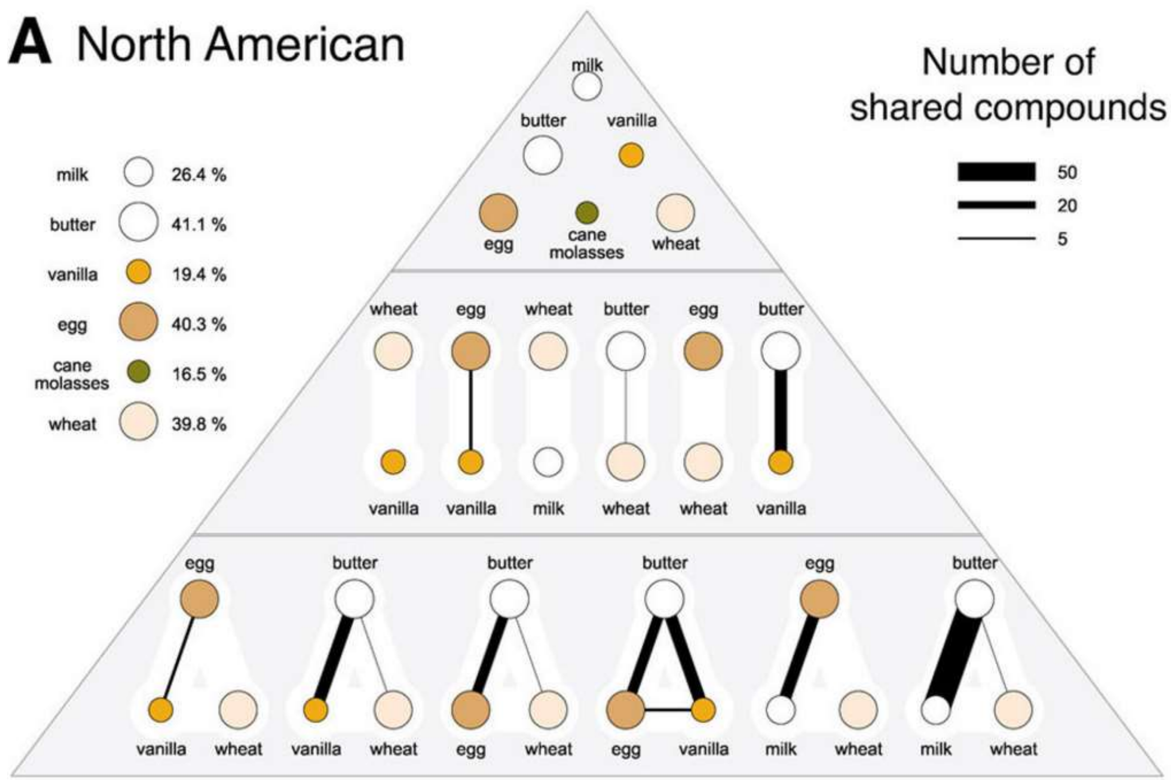
Motivation

- **Food**
  - Flavor network and the principles of food pairing
  - Idea:
    - Analyze the co-occurrence of graph of ingredients in recipes database worldwide
      - North America, Western Europe, Latin America, East Asia, Southern Europe
    - Derive the underlying principles of food pairing in different regional cuisines

Ahn, Yong-Yeol, et al. "Flavor network and the principles of food pairing."*Scientific reports* 1 (2011).

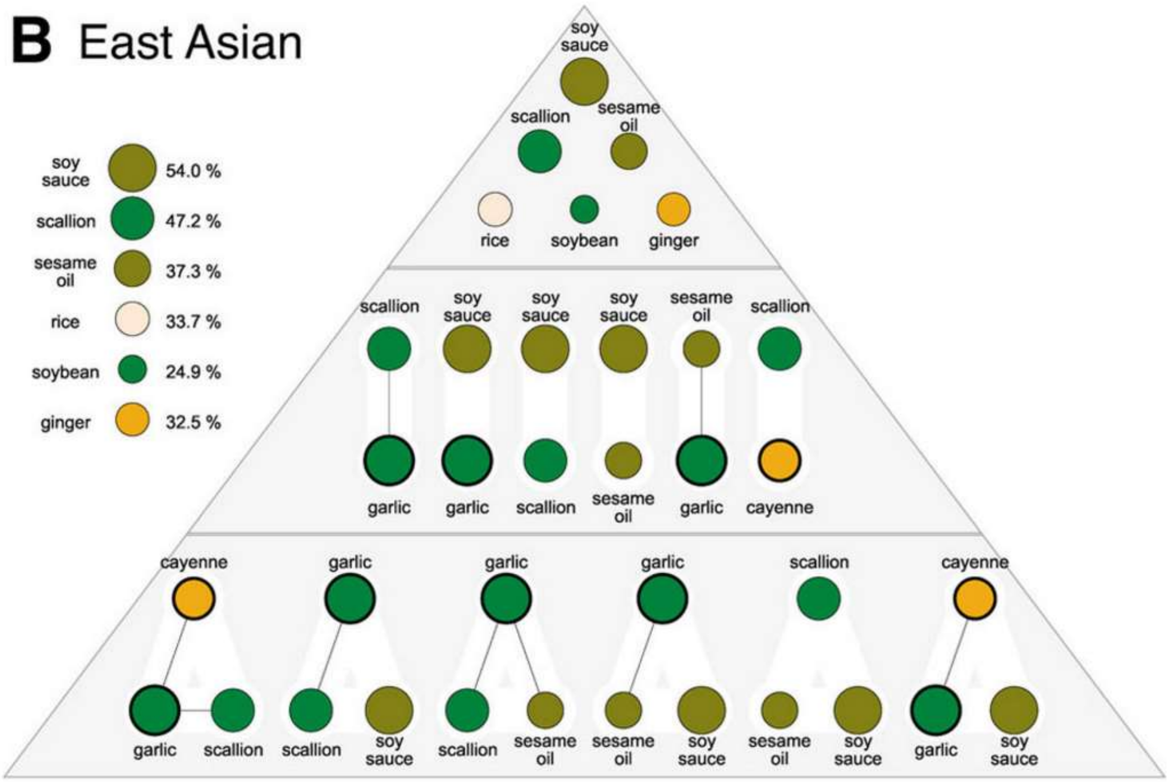
Motivation

- **Food**
  - Flavor network and the principles of food pairing
    - Results by looking into 6 most authentic ingredients



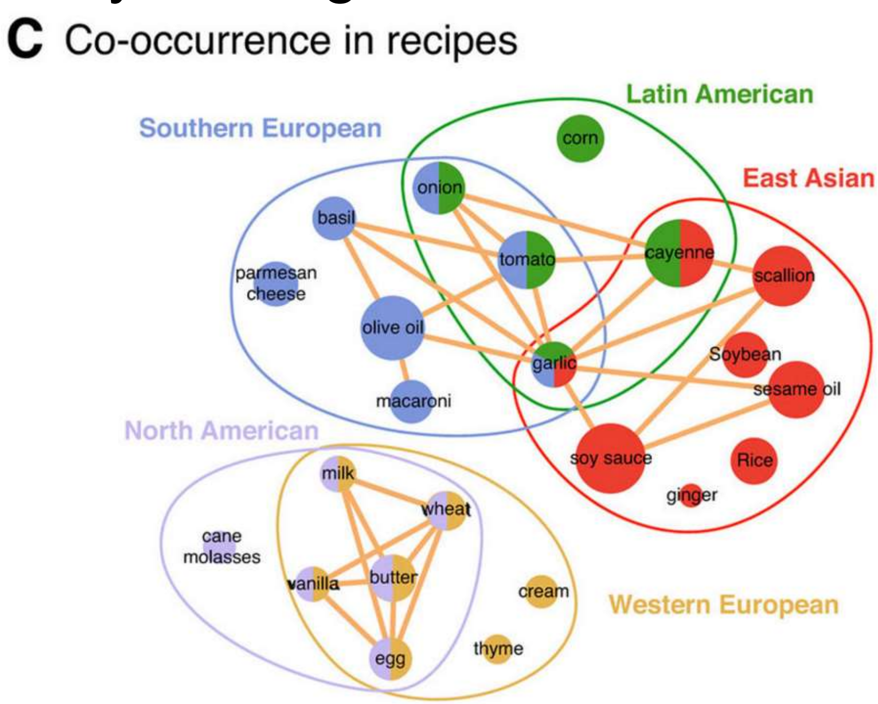
Motivation

- **Food**
  - Flavor network and the principles of food pairing
    - Results by looking into 6 most authentic ingredients



Motivation

- **Food**
  - Flavor network and the principles of food pairing
  - Results by looking into 6 most authentic ingredients



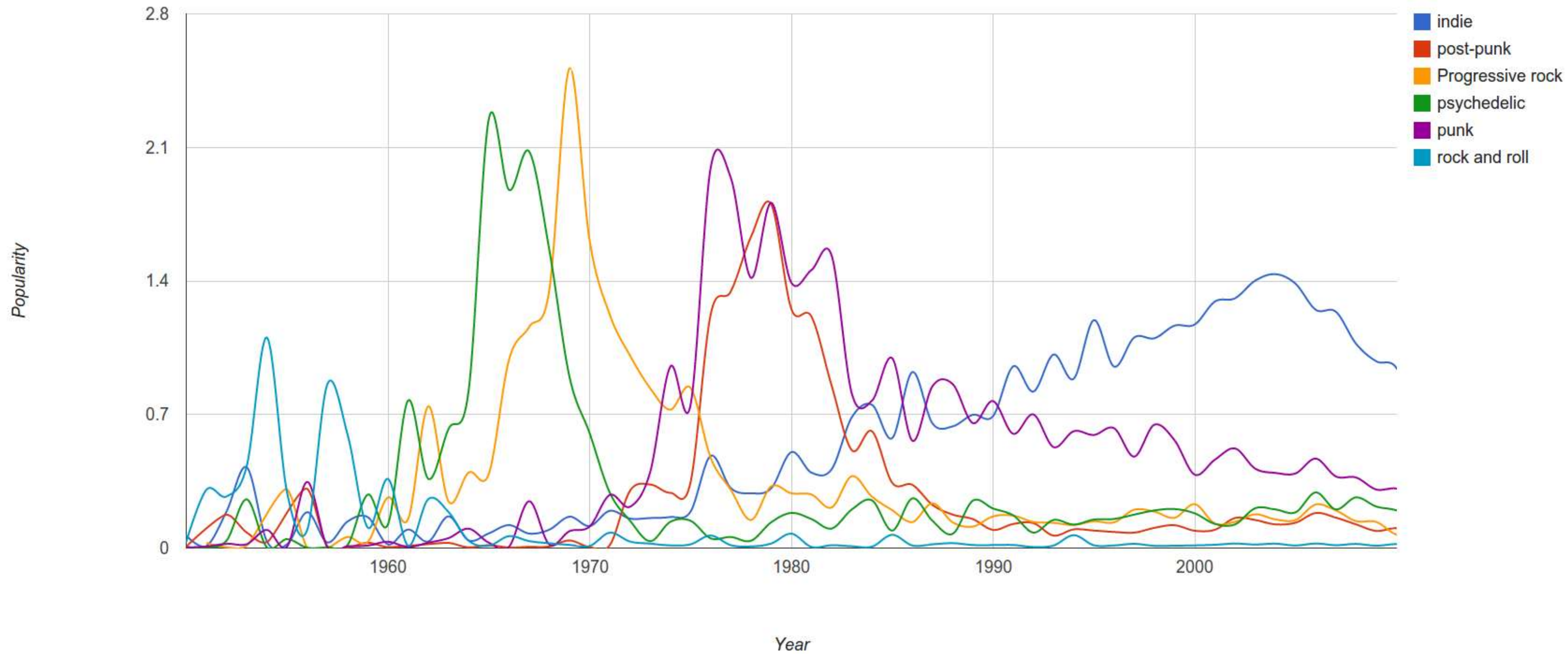
Ahn, Yong-Yeol, et al. "Flavor network and the principles of food pairing."*Scientific reports* 1 (2011).

Motivation

- **Music**
  - The analysis of genre popularities at Last.fm
  - Idea:
    - There is data about artists and bands forming *year and place* available (this data is generated Last.fm's users and attached to artists' wiki pages)
    - At the Last.fm they have massive amount of user tags that correlate artists to genres
    - Derive genres popularities based on the number of artists formed at that time

Motivation

- **Music**
  - The analysis of genre popularities at Last.fm
  - Results



http://blog.last.fm/2012/09/06/genre-timelines-and-more-distinctive-lyrics

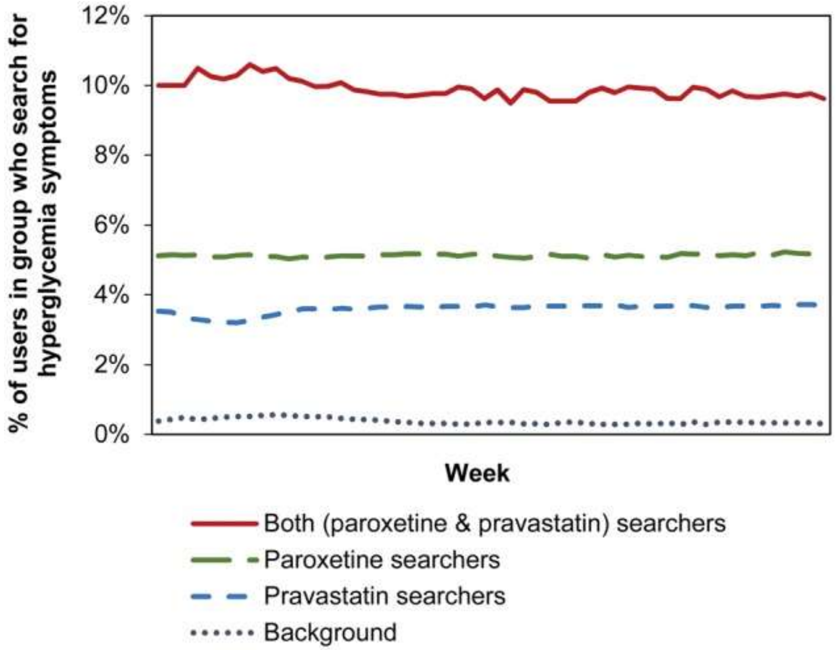
Motivation

- **Health and Pharmacy**
  - Finding hidden side effects associated with particular drugs
  - Idea:
    - Analyze web search traffic for a set of ~1M users over a period of time (about a week)
    - Associate other symptoms (such as hyperglycemia) as a side effect of a particular drug

http://www.ncbi.nlm.nih.gov/pubmed/23467469

Motivation

- **Health and Pharmacy**
  - Finding side effects associated with particular drugs
  - Results



However, hyperglycemia was not associated with these drugs!!!

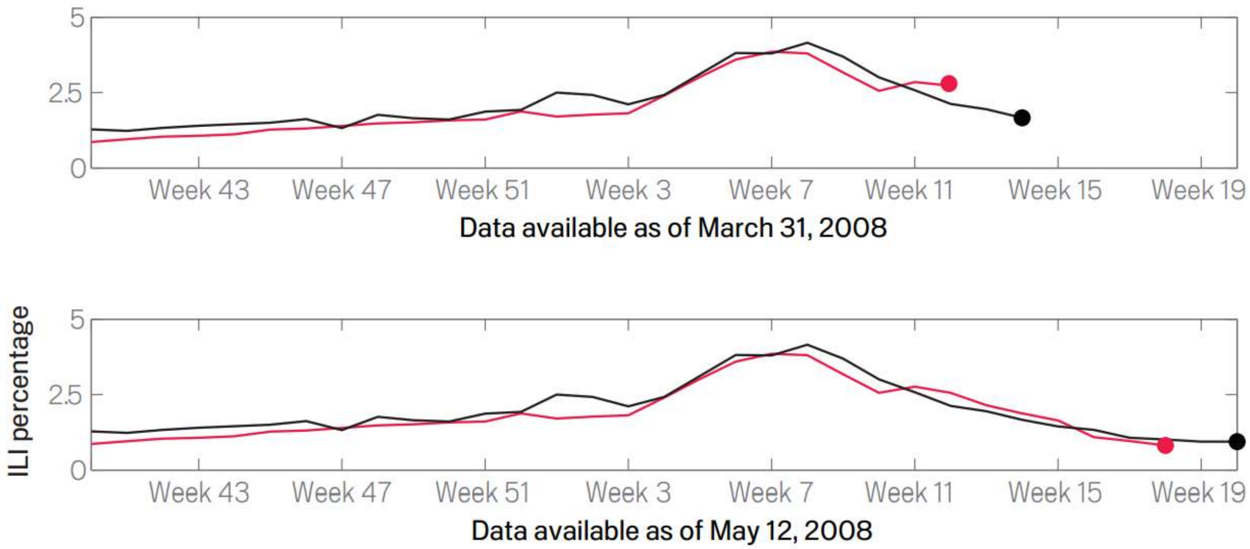
http://www.ncbi.nlm.nih.gov/pubmed/23467469

Motivation

- **Predicting flu season (Google flu trend)**
  - Predict the flu season by monitoring and mining web search traffic
  - Idea:
    - When people notice flu symptoms or catch a flu, they tend to search terms related to flu or flu symptoms
    - In order to predict the flu season one could monitor web search traffic and measure the occurrence of flu related terms
    - By monitoring flu related terms one could predict that there is a flu outbreak coming in case the flu related terms occur more frequently than usual

Motivation

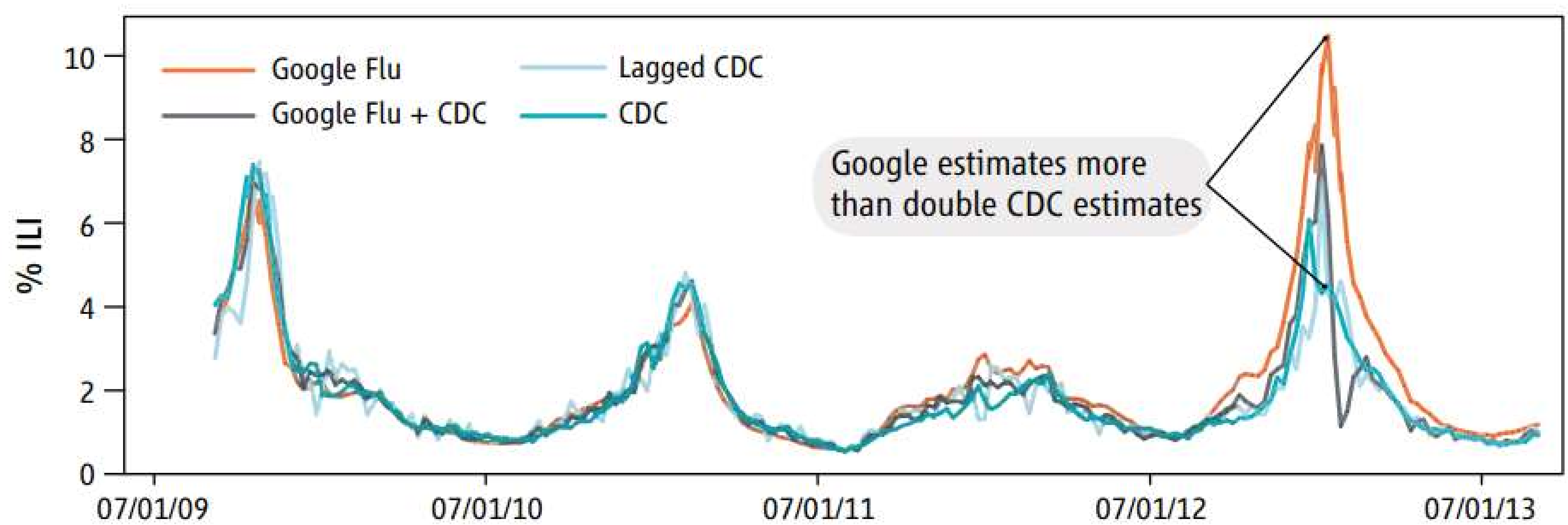
- **Predicting flu season (Google flu trend)**
  - Predict the flu season by monitoring and mining web search logs
    - Results – Google flu was very accurate in flu outbreak prediction, it was able to predict flu earlier than CDC



Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* 457.7232 (2009): 1012-1014.

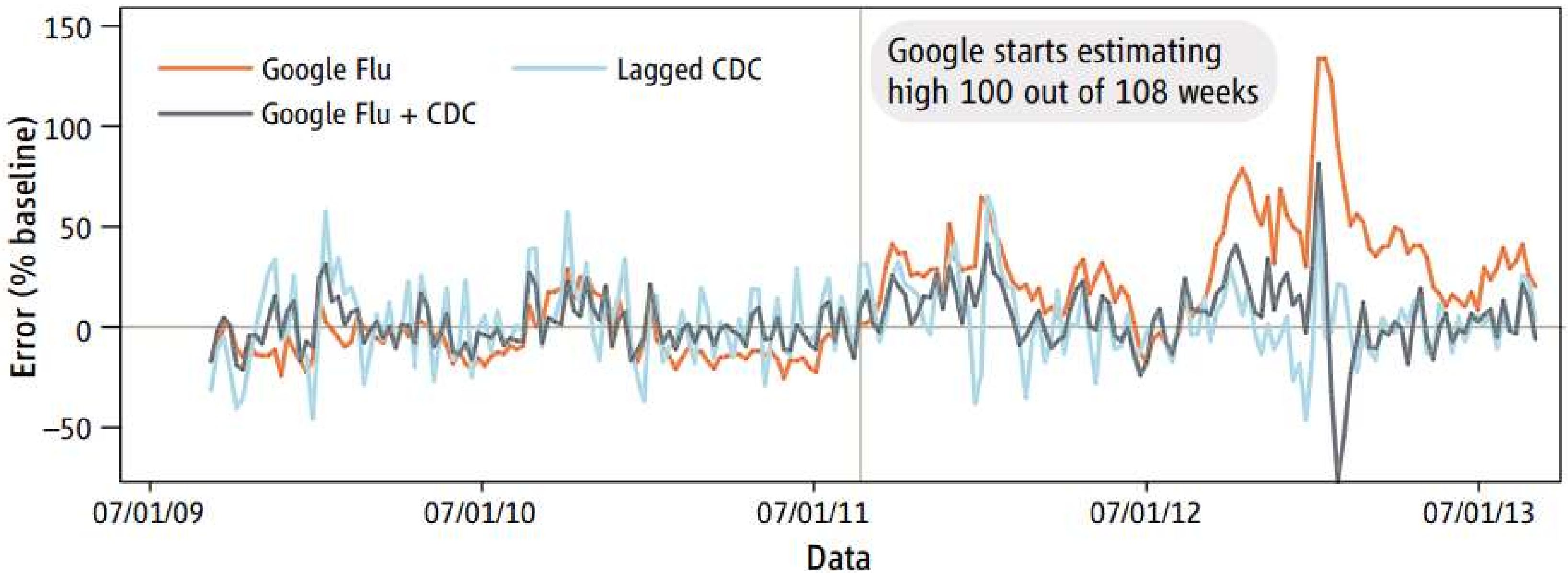
Motivation

- **Predicting flu season (Google flu trend)**
  - Predict the flu season by monitoring web search logs
    - However, as reported in 2013…
      - GFT has been persistently overestimating flu prevalence in 2011 – 2012 and has missed high for 100 out of 108 weeks



Motivation

- **Predicting flu season (Google flu trend)**
  - Predict the flu season by monitoring web search logs
    - However, as reported in 2013…
      - Even traditional models based on only 3-week old data (lagged CDC) were outperforming GFT



Lazer, David, et al. "The parable of Google Flu: traps in big data analysis."*Science* 343.14 March (2014).

Motivation

- **Predicting flu season (Google flu trend)**
  - Predict the flu season by monitoring web search logs
    - However, as reported in 2013…
      - So, what was the issue? Why did “BigData” fail?
        - » “Big Data hubris” is the implicit assumption that big data are a substitute for traditional data collection and analysis
        - » Essentially, the methodology was to find best matches among 50M search terms to fit 1152 data points
          - » The odds of finding search terms that match flu symptoms but are structurally unrelated were quite high
        - » The most common explanation: Media-stoked panic in 2013 flu season – flood of searches containing flu terms

Lazer, David, et al. "The parable of Google Flu: traps in big data analysis."*Science* 343.14 March (2014).

Motivation

- **Predicting flu season (Google flu trend)**
  - Predict the flu season by monitoring web search logs
    - However, as reported in 2013…
      - So, what was the issue? Why did “BigData” fail?
        - » Another explanation: Google modified their search algorithm to better fit their business model
          - » For instance, for searches including terms “fever” and ”cough” it returned suggestions with potential diagnoses
          - » This recommendation business model can increase the occurrence of particular search terms
    - More general
      - All empirical research stands on a foundation of measurement
      - Is the instrumentation actually capturing the theoretical construct of interest?

Lazer, David, et al. "The parable of Google Flu: traps in big data analysis."*Science* 343.14 March (2014).

Motivation

- **We have seen different examples**
    - Graph analytics
    - Databases
    - Visualizations
    - Large datasets
    - Repurposing data

Big Data

- What makes it a Big Data?

> *"Big Data is any data that is expensive to manage and hard to extract value from."*
>
> **Michael Franklin**
>
> **Director of the Algorithms, Machines and People Lab**
>
> **University of Berkeley**

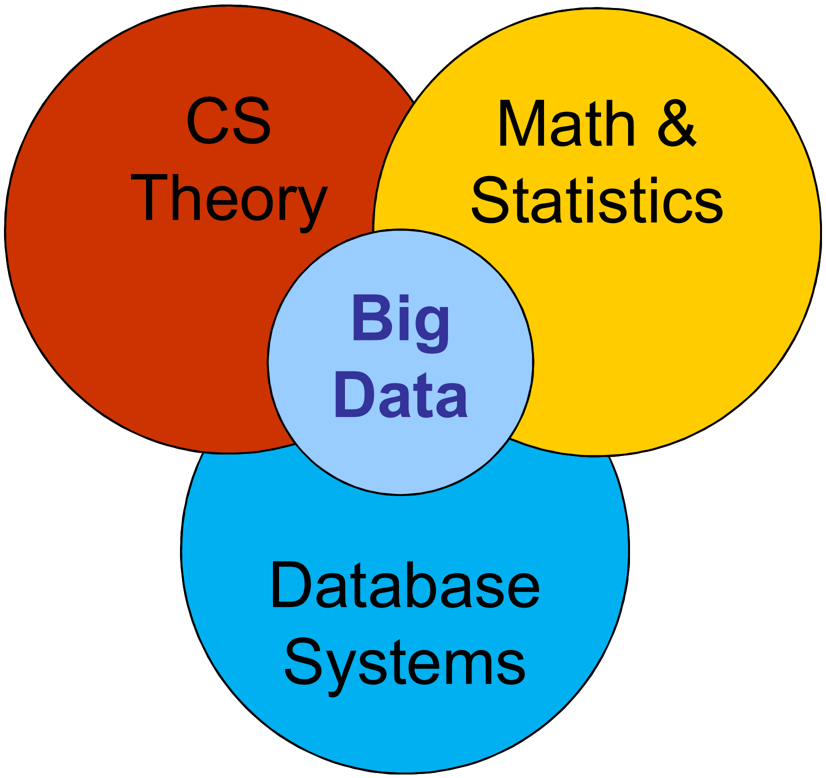
- Idea: "Big" is relative!

Big Data

- **Big Data contains value and knowledge**
- **To extract the knowledge data needs to be:**
  - Stored
  - Managed
  - ANALYZED * this course *

Big Data

- **Where does Big Data stands**
  - It overlaps with
    - Business Intelligence
    - Statistics
    - Database Management
    - Visualization
    - Machine Learning



Big Data

- **Big Data and its relation to…**
  - Business Intelligence
    - Software stack designed for BI is very specific and not very adaptable when requirements change
      - Data warehouse and specific dashboards and reports that consume data from the data warehouse in order to answer specific questions.
    - Software stack designed for BI is not applicable to Big Data problems where changing requirements is a norm
    - BI is a specific tool designed for a specific set of problems, Big Data is much broader term
    - BI engineers do not consumer their own products and make the decisions themselves, while Big Data analysts do

Big Data

- **Big Data and its relation to…**
  - Statistics
    - Statistical methods are the core of what Big Data is today
    - A statistician will typically assume that datasets he deals with will fit into main memory on a single machine
    - Statistics – extract most information from a very sparse and expensive to acquire typically small dataset
    - However, now we move from a data poor regime to a data rich regime
    - The goal is not anymore about to design new fancy mathematical method to squeeze more information from a small dataset
    - The goal is now about to build new engineering tools to process very large datasets

Big Data

- **Big Data and its relation to…**
  - Database Management
    - Database engineers and administrators posses a lot of skills to make them appropriate for Big Data tasks
    - However, they are focused on a particular data model which is usually the relational data model (rows and columns)
    - Big data analysts deal with heterogeneous data sources that may include video, audio, text, graphs (nodes and edges), images, structures and unstructured data, etc.
    - For some of those data sources the relational data model may not be appropriate data model

Big Data

- **Big Data and its relation to…**
  - Visualization
    - Similarly like statisticians, visualization specialist are less concerned with massive datasets that span across hundreds/thousands of machines on the Internet
  - Machine Learning
    - It is perhaps the closest to Big Data
    - However, the proportion of time spent on choosing the appropriate machine learning technique and running it is a very small fraction
    - Big Data analysts spend much more time on the preparation of the data
      - Manipulation, cleaning of the data (Data jujitsu).

Big Data

- **3V in Big Data**
  - Volume
    - The size of the data
      * Terabytes, Records, Transactions, Tables, Files, etc. *

  - Velocity
    - The latency of data processing with respect to the growing demand for data usage
      * Batch, real-time, streams, near-time, etc. *

  - Variety
    - The diversity of sources, formats, quality, structures
      * structured, unstructured, semi structured, logs, text, sensors, images, other media, etc. *

Big Data

- **Big Data sources**
  - The ability to collect a enormous amount of data from customers
    - Everything people click, buy, show particular interest, search along with the actual time is logged and placed somewhere
  - New and pervasive sensors
    - In science, for instance underwater sensors, astronomy telescopes
    - But also in business, devices like car black boxes
    - HydroSense and ElectriSense devices
  - The ability to “keep everything”
    - The disk capacity is growing extremely fast while the price of disk storage is significantly reduced

Big Data

- **Three important skills of Big Data analyst**
  - Statistics
    - Traditional analysis and interpretation
  - Data Munging (Data Ju-jitsu)
    - Parsing, scraping and formatting large and heterogeneous data sets
    - Big Data specialist spend 90% of their time on this task!!!
  - Visualization
    - Communicate the results using graphs, tools, etc.

Course Content

- **What are the expected outcomes of the course?**
  - Deal with different types of data
    - High dimensional data
    - Graphs
    - Infinite/never-ending data
    - Labeled data
  - Use different computation models
    - MapReduce
    - Streams and online algorithms
    - In-memory single machine

Course Content

- **What are the expected outcomes of the course?**
  - Solve real-world problems
    - Recommender Systems
    - Market Basket Analysis
    - Spam filters
    - Duplicate document detection
  - Gain various skills spanning through different areas
    - Linear Algebra (SVD, Rec. Sys., Detecting Communities)
    - Optimization (stochastic gradient descent)
    - Dynamic Programming (Frequent itemsets)
    - Hashing (LSH, Bloom filters)

Course Content

- **Prerequisites**
  - Algorithms
    - Dynamic programming, basic data structures
  - Basic probability
    - Moments, typical distributions, MLE, …
  - Programming
    - You can do programming assignments in any language we support (Python, Java, C, C++, C# it is your choice)
    - We recommend Python, Java and C#

Course Content

- **Lectures schedule (1. cycle)**

1.	Introduction Lecture	10.03.2016.
2.	MapReduce	17.03.2016.
3.	Detecting Near Duplicates	24.03.2016.
4.	Finding Frequent Itemsets	31.03.2016.
5.	Mining Data Streams	07.04.2016.
6.	Link Analysis	14.04.2016.
7.	Clustering	21.04.2016.

  - Midterm – two weeks long break

Course Content

- **Lectures schedule (2. cycle)**

8.	Detecting Communities in SNGs	12.05.2016.
9.	Recommendation Systems	19.05.2016.
○	Corpus – Holiday in Croatia	
10.	Advertising on the Web	02.06.2016.
11.	Finding Similar Items	09.06.2016.
12.	Large-Scale Machine Learning	16.06.2016.
○	Final exam	

Course Organization

- **Course stuff**
  - Lecturers
    - Full professor Siniša Srbljić, PhD
    - Assistant professor Dejan Škvorc, PhD
    - Assistant professor Ante Đerek, PhD
  - Teaching Assistants
    - Goran Delač, PhD
    - Marin Šilić, PhD
    - Klemo Vladimir, PhD
    - Zvonimir Pavlić, MSc
  - Questions
    - avsp@zemris.fer.hr

Course Organization

◻ ECTS points	**4**
◻ Lectures hours	**30**
◻ Programming assignments	**15**
◻ Course duration in weeks	**13**

Course Organization

- **Calendar**
- **Lectures are scheduled**

  On Thursday

  11:00 – 13:00

  B5

- **Questions**
  - avsp@zemris.fer.hr

Academic Calendar 2015. / 2016.
Bachelor and Master Study

	October					November						December				
Mon		5	12	19	26		2	9	16	23	30		7	14	21	28
Tue		6	13	20	27		3	10	17	24		1	8	15	22	29
Wed		7	14	21	28		4	11	18	25		2	9	16	23	30
Thu	1	8	15	22	29		5	12	19	26		3	10	17	24	31
Fri	2	9	16	23	30		6	13	20	27		4	11	18	25	
Sat	3	10	17	24	31		7	14	21	28		5	12	19	26	
Sun	4	11	18	25		1	8	15	22	29		6	13	20	27	

	January					February					March				
Mon		4	11	18	25	1	8	15	22	29		7	14	21	28
Tue		5	12	19	26	2	9	16	23		1	8	15	22	29
Wed		6	13	20	27	3	10	17	24		2	9	16	23	30
Thu		7	14	21	28	4	11	18	25		3	10	17	24	31
Fri	1	8	15	22	29	5	12	19	26		4	11	18	25	
Sat	2	9	16	23	30	6	13	20	27		5	12	19	26	
Sun	3	10	17	24	31	7	14	21	28		6	13	20	27	

	April					May						June				
Mon		4	11	18	25		2	9	16	23	30		6	13	20	27
Tue		5	12	19	26		3	10	17	24	31		7	14	21	28
Wed		6	13	20	27		4	11	18	25		1	8	15	22	29
Thu		7	14	21	28		5	12	19	26		2	9	16	23	30
Fri	1	8	15	22	29		6	13	20	27		3	10	17	24	
Sat	2	9	16	23	30		7	14	21	28		4	11	18	25	
Sun	3	10	17	24		1	8	15	22	29		5	12	19	26	

	July					August					September				
Mon		4	11	18	25	1	8	15	22	29		5	12	19	26
Tue		5	12	19	26	2	9	16	23	30		6	13	20	27
Wed		6	13	20	27	3	10	17	24	31		7	14	21	28
Thu		7	14	21	28	4	11	18	25		1	8	15	22	29
Fri	1	8	15	22	29	5	12	19	26		2	9	16	23	30
Sat	2	9	16	23	30	6	13	20	27		3	10	17	24	
Sun	3	10	17	24	31	7	14	21	28		4	11	18	25	

6	Weeks reserved for teaching	21	Sessions of the Faculty Council
1	Weeks reserved for final exams	5	Beginning of the Academic Year
23	Weeks reserved for mid-term exams	23	Holidays for students
15	Regular exams	20	The Faculty Day
29	Enrolment and Skills	9	Non-working days
8	National holidays	17	Working Saturdays

Course Organization

- **Programming assignments**
  - Several assignments (probably 4 or 5)
  - Students submit their solutions independently
  - The solutions will be evaluated using online judge system
  - Follow the news on the course website, instructions for the first assignment will be posted soon!

Course Organization

- **Grades**

Theory	**65 points**
○ Midterm	30 points
○ Final exam	30 points
○ Lectures participation	5 points
Additional	
○ Student's activity during lectures	-15 do + $\propto$ points
Programming assignments	**35 points**
○ Each assignment will be graded uniformly	

Course Organization

- **The right to take an exam**
  - Programming assignments threshold 17.5 points (50%)
- **To get a positive grade**
  - Theory threshed 32.5 points (50%)
  - Programming assignments threshold 17.5 points (50%)

- **Grades** (*)
  - excellent (5) ≥ 88 points
  - very good (4) ≥ 75 points
  - good (3) ≥ 63 points
  - sufficient (2) ≥ 50 points

Mozgalo Contest

- **Mozgalo Big Data Contest**
  - Important dates
    - Apply by March 14, 2016 (hurry up, 4 more days left!!!)
    - Submission by May 14, 2016
    - Final decisions May 23, 2016
  - Work in teams 2 – 4 students
    - Solve real-world problems
    - Past years
      - What is the best beer in Croatia (2014, 106 contestants in 33 teams)
      - Marketing campaigns in banks (2015, 155 contestants in 51 team)

Mozgalo Contest

- **Mozgalo Big Data contest**
  - Prizes
    - 1st prize 12k HRK
    - 2nd prize 8k HRK
    - 3rd prize 4k HRK
  - Apply
    - https://www.estudent.hr/category/natjecanja/mozgalo
    - https://www.youtube.com/watch?v=U7imrpbfFrA
  - Contest & Course
    - It is tightly related to the course, it covers the same topics