An Introduction to Web Science

RuSSIR, Aug 6-10, 2012

Please interrupt at any point!!

Ingmar Weber
ingmar@yahoo-inc.com
Yahoo! Research Barcelona

What is "Web Science"?

"Web Science embraces the study of the Web as a vast information network of people and communities. It also includes the *study of people* and communities using the digital records of user activity mediated by the Web. An understanding of *human behavior* and *social interaction* can contribute to our understanding of the Web, and data obtained from the Web can contribute to our understanding of human behavior and social interaction." [ACM Web Science conference site]

Studying the online world to understand the offline world.

Who the Hell is that Guy?!?

Research Scientist at Yahoo! Research in Barcelona

Working on "Web Mining"
- Query logs: demographic, political, …
- Browsing logs: pre-edit behavior, political, …
- Email: migration patterns, …
- Twitter: political, …

Couchsurfing, endurance sports, candy, …

Course Outline

- **Day 1: Introduction to the Introduction**
  - **Examples, data sets, presentation of the competition**
- Day 2: Web Search and Society
  - Demographics, economy and more
- Day 3: Blogs and Twitter
  - Gender, moods, politics, stock market and more
- Day 4: Social Networks and Online Dating
  - Attractiveness, FB&GPA, FB&Personality and more
- Day 5: E-commerce and Marketing Studies
  - Brand congruence, Groupon Effect, social ads

Examples

Motivating Example: Google Flu Trends

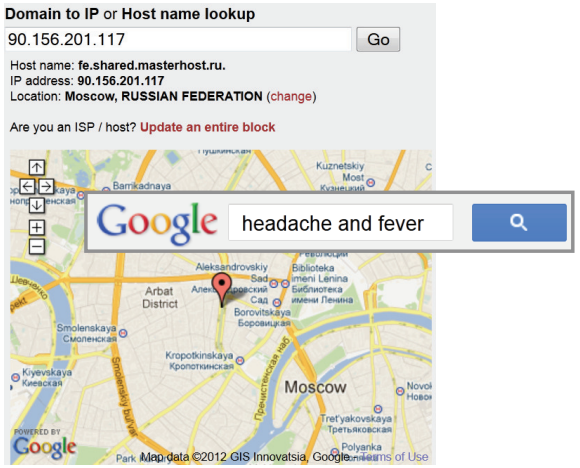
- Show live:
  http://www.google.org/flutrends/

Example 1: Google Flu Trends

- Flu pandemics can potentially kill millions
  - In 1918 >3% of the world's population died
- Health officials are often slow to know:
  - Several days before patients go and see a doctor
  - Several days/weeks until data is aggregated
- People having the flu might search online for:
  - fever, flu, running nose, headache, ...
  - They search before going to the doctor
  - Search engines know immediately

How Does it Work (1/3)

- Queries issued to the search engine are mapped to locations using the IP address



How Does it Work (2/3)

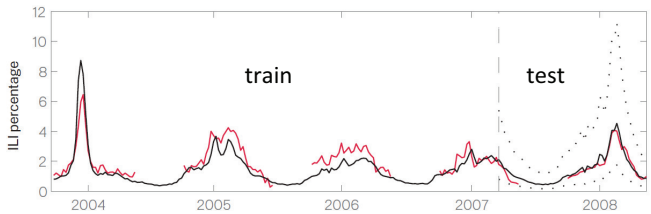
- Aggregate queries for each region, say, U.S. county, and for each week
- Model it as regression problem:
  - Predict probability P of a physician's visit in the region related to an influenza-like illness (ILI)
  - Use search volume as a single explantory variable X_1
- Use logistic regression
  - P = $\frac{e^{(\beta_0+\beta_1 X_1+e)}}{e^{(\beta_0+\beta 1 X_1+e)}+1}$ (always in [0,1])
  - logit(P) = $\beta_0 + \beta_1 X_1 + e$, where logit(P) = P/(1-P)

A Prime Example of Web Science

- Studying the online world ...
  - Use search activity on a large web search engine
- ... to understand the offline world
  - Have real-time "nowcasting" of flu activity
- Good science
  - Validation against historic flu records
  - Using machine learning techiques
- Available for wider user
  - Web demo and also Google Correlate (later)

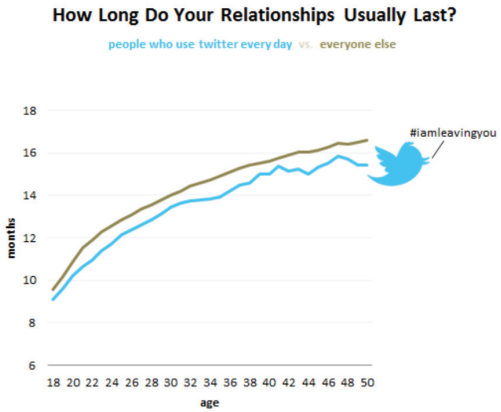
How Does it Work (3/3)

- Use historic data, e.g. two years, to fit the model
  - For each query with sufficient frequency find the best b_0 and b_1, e.g. minimizing the sum of squared errors
  - Select the queries which have the smallest average error across several regions
  - Combine their volume (c.f. boosting) and find a good set of top queries
- Test the model on held-out data



Example 2: Web Science is Attractive

- OkCupid.com: A large online dating site http://blog.okcupid.com/
- Users answer lots of profile questions



Don't forget the selection bias!

Will a First Date Lead to a Relationship?

- Look at users who delete their account
  - Reason given: "I met somebody on OkCupid"
  - Get a list of *couples* this way
- Find set of profile questions with agreement for couples much higher than chance
- "Hidden" machine learning motivation
  - Build classifier to tell couples from non-couples

Is God important in your life?
Is sex the most important part of a relationship?
Does smoking disgust you?

15% couples vs. 8% non-couples

Wouldn't it be fun to chuck it all and go live on a sailboat?
Do you like horror movies?
Have you ever traveled around another country alone?
32% couples vs. 9% non-couples

Data

Example 3: Web Science is Delicious

- What do people cook on Thanksgiving?
- Analysis of search logs of allrecipes.com

http://www.nytimes.com/interactive/2009/11/26/us/20091126-search-graphic.html

- Analyzed up- and down-regulation with respect to average regional volumes
- Temporal: searches for gravy peak later

Web Science Requires Data

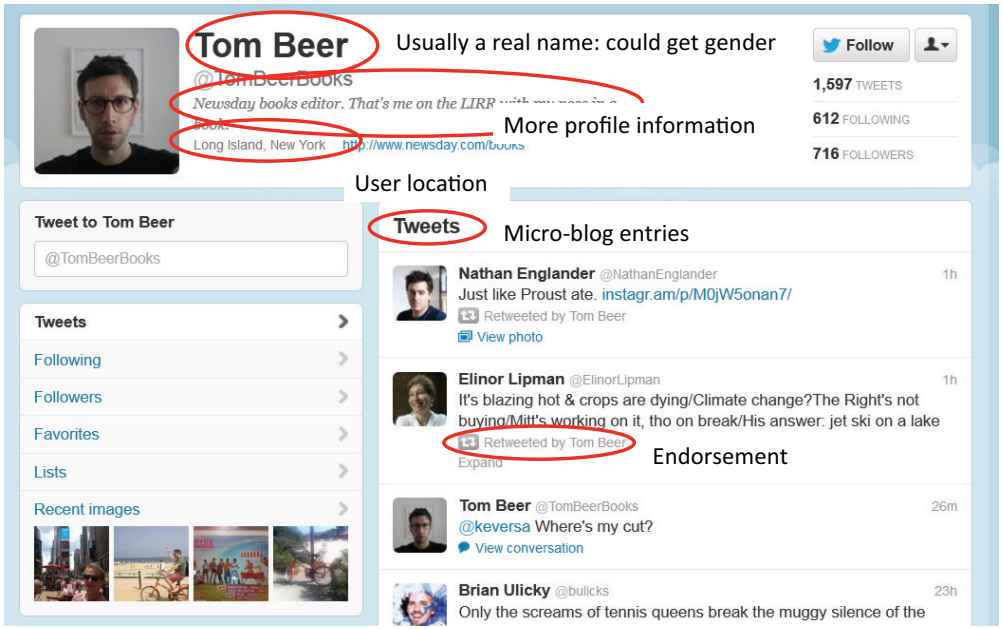
- To study the online world ...
  - ... you need data about the online world
- To understand the offline world ...
  - ... your data needs to be linked to the real world

- Bare web graph (only vertices and edges)
- Measurement of disk throughput under load
+ Blogs with detailed, trustworthy user profiles
+ Wikipedia edits with geolocated IP addresses

Where's the Data?

- A lot of data is locked away inside companies
  - Facebook, Yahoo!, Google, Вконтакте, Яндекс, ...
- Those companies are not that closed ...
  - Many have research-oriented summer internships
- Also lots of data publicly available online
  - Twitter, blogs, Y! Answers, product reviews, ...
- But careful with scraping
  - Read the Terms of Service!
  - Papers get rejected + potential legal trouble

Data: Twitter



Don't Scrape – Even if it's Tempting

"This license does not include any resale or commercial use of this site or its contents; any collection and use of any product listings, descriptions, or prices; any derivative use of this site or its contents; any downloading or copying of account information for the benefit of another merchant; or any use of data mining, robots, or similar data gathering and extraction tools." Amazon's Conditions of Use

Data: Twitter

- Different ways to access the data
- Streaming API, 1% of public tweets
  https://dev.twitter.com/docs/streaming-apis/streams/public
- Search API, 350 requests per hour
  https://dev.twitter.com/docs/api/1/get/search
- Ways to improve rate limit
  http://apigee.com/

Data: Twitter

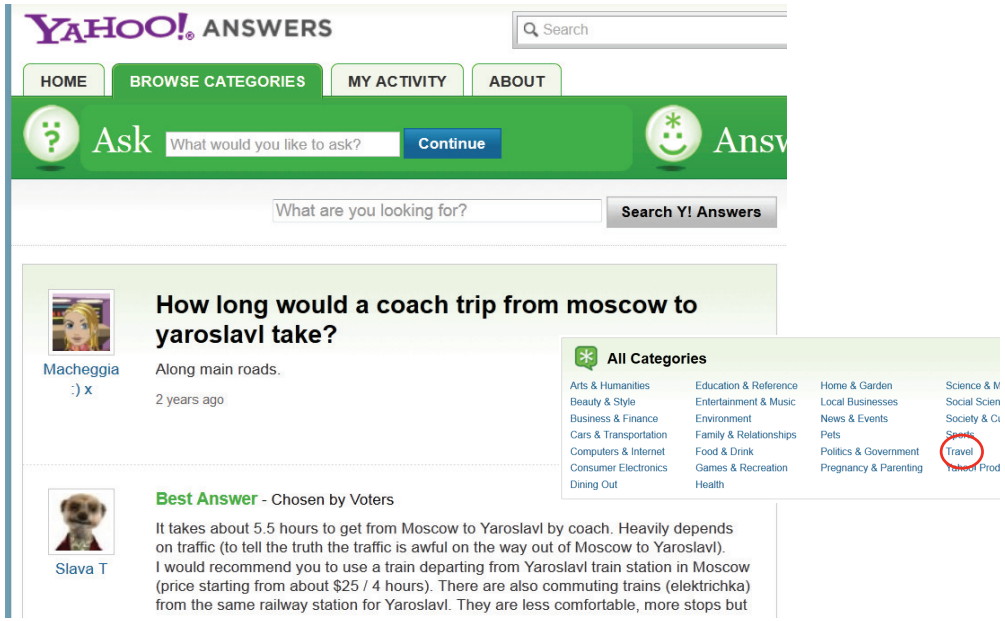
- User directories
  http://www.twellow.com/
  http://wefollow.com/
- Services to get an influence score
  http://developer.klout.com/iodocs
- Services to get sentiments

  https://sites.google.com/site/twittersentimenthelp/api
- Services to get named entities
  http://code.google.com/p/iestwitter/

Data: Yahoo! Answers

- Questions organized by topic
- Users can have profiles
- Lots of user information in questions/answers
- “I‘m Russian.”, “I‘m left-handed.”, …
- Search-based Yahoo! Answers API
  http://developer.yahoo.com/answers/
  10,000 calls per hour

Data: Yahoo! Answers



Data: Wikipedia

Data: Wikipedia

- Complete edit history for all articles
- Some articles are gelocated
- IP addresses for anonymous edits
- IP addresses can be geolocated
- Editors' have profiles

  User:Ellen541167: "I am a democrat and supported Barack Obama for the 2008 general election."

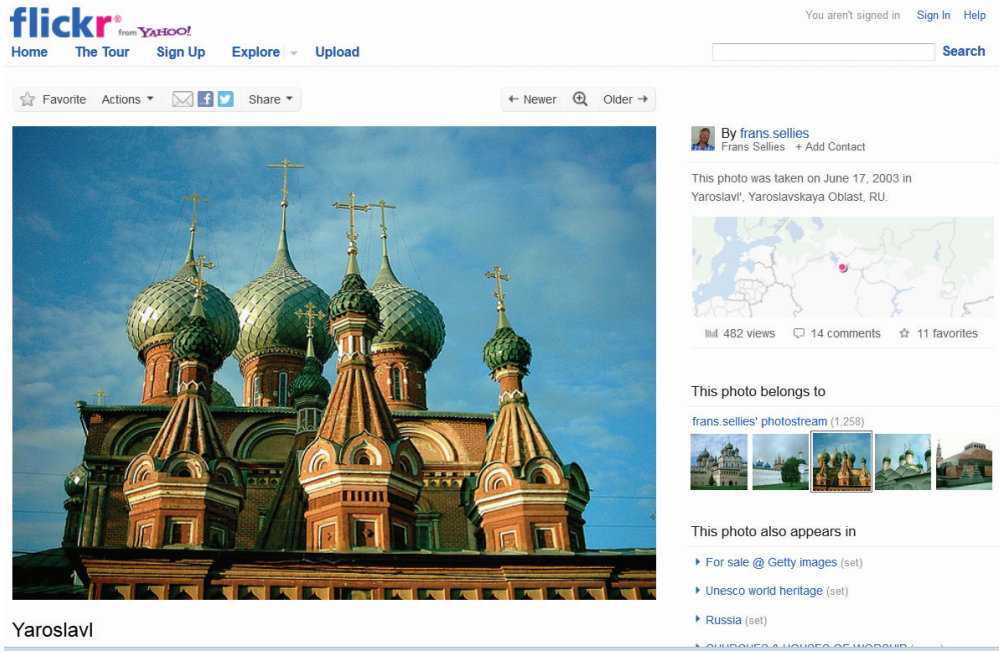
Data: Flickr

- Images come with
  - Tags, location, date, camera information, view count, ...
- Users come with
  - Profile page, list of contacts, group memberships, ...
- Search-based Flickr API

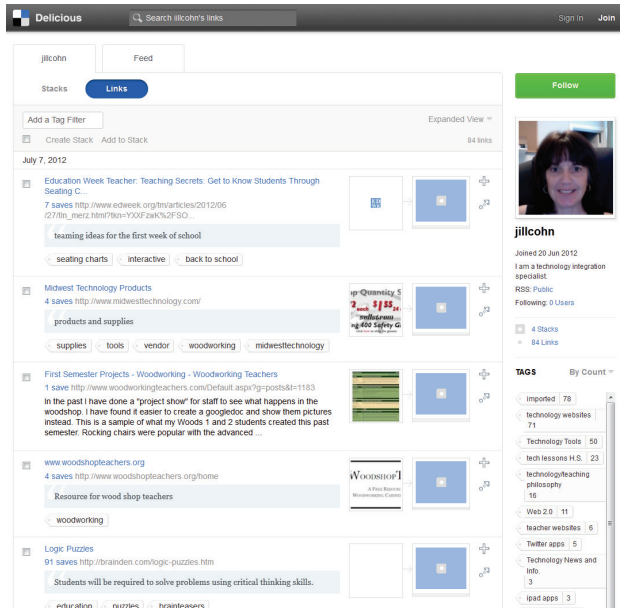
  http://www.flickr.com/services/api/

  10,000 queries per hour

Data: Flickr



Data: Delicious



Data: Delicious

- URLs come with

  Tags, number of “saves”
- Users come with

  URLs with timestamp, URLs with comments, profile information, “stacks”, …
- Search-based Delicious API

  http://delicious.com/developers

  10,000 queries per hour

Data: Youtube

- Videos come with

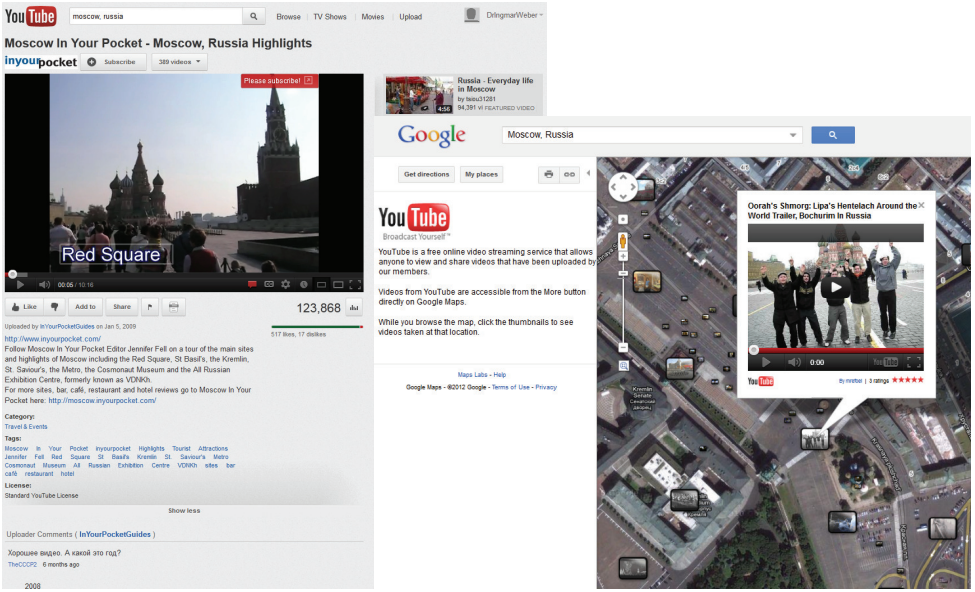
  Tags, description, date, rating, category information, comments, view count, location, …
- User come with

  Profile description, list of videos, …
- Search-based API

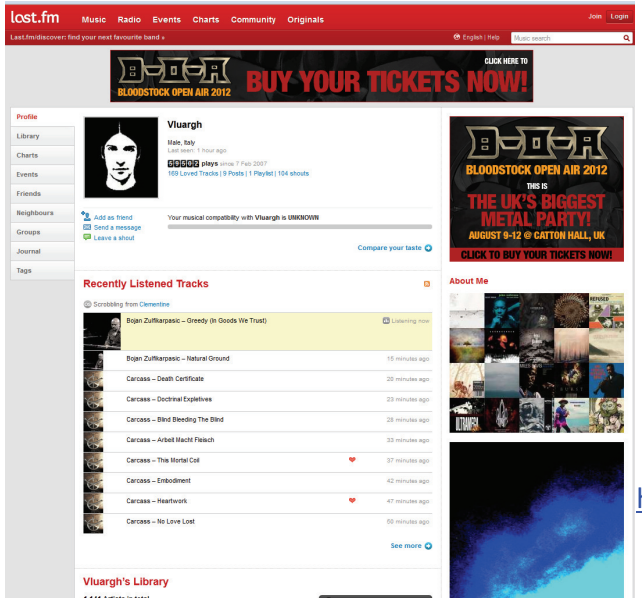
  https://developers.google.com/youtube/

  No formal limit, but they will block you if you misbehave

Data: Youtube



Data: last.fm



http://www.last.fm/api

Data: Google Correlate

- Which web search queries correlate strongest
  - With number of flu cases over time
  - With stock prices over time
  - With unemployment rates across regions
  - With hour of sunshine across regions
- Google Correlate
  http://www.google.com/trends/correlate/
  Upload your own time (or region) series

Data: Tools

- Mapping location names to geo-coordinates
  - http://developer.yahoo.com/geo/placemaker/
- Mapping IP addresses to locations
  - http://www.maxmind.com/app/geolite
- Assigning a sentiment to a short piece of text
  - http://sentistrength.wlv.ac.uk/
- Getting lots of manually labeled data
  - https://www.mturk.com/mturk/welcome

Data: Country Statistics

- CIA World Factbook
  - https://www.cia.gov/library/publications/the-world-factbook/
- U.N. Statistical Database
  - http://unstats.un.org/unsd/databases.htm
- U.S. census data
  - http://www.census.gov/main/www/access.html
- Many country-specific sources

Data: More Links

- Bing web search API
  - http://www.bing.com/developers/s/APIBasics.html
- Get Alexa's site traffic information
  - http://aws.amazon.com/awis/
- A large AOL web search query log (2006)
  - http://www.gregsadetsky.com/aol-data/
- A large set of topically classified URLs
  - http://www.dmoz.org/help/getdata.html

Data: Even More Links

- The Internet "Wayback" Machine
  - http://archive.org/web/web.php
  - Yandex in 1998: 

- Rewiring and combining RSS feeds
  - http://pipes.yahoo.com/pipes/
- A large directory of topic-specific blogs
  - http://technorati.com/blogs/directory/

Competition

Your Web Science Proposal

- A small research-related competition
  - A non-monetary, digestible price
  - Participate as teams or individuals
- Come up with a mini research proposal
  - Any interesting, funny, creative, odd idea
  - Using publicly available data sources

- Are cat or dog owners more intelligent?
- Do left-handed people listen to different music?
- Do color preferences change as people grow older?

Timeline of the Competition

- Today+Tomorrow: Start thinking, discussing, reading, exploring, ...

- Before Wed. 11h00: Submit/edit your proposal (one paragraph only):

http://tinyurl.com/RuSSIR-Research-Proposals

- Before Thu. 11h00 (and after Wed. 14h00): Cast your vote for one submitted proposal:

http://tinyurl.com/RuSSIR-Proposal-Voting

Timeline of Competition (ctd.)

- During Thu. lecture: the top three proposals (according to online votes) are announced
- During Fri. lecture: the top three proposals are presented in person (2 min each, max of 3 slides)

  Winner is determined using an Applause-o-Meter
- Everyone: follow through and submit your research results to "The Journal of Irreproducible Results" http://www.jir.com/ (peer-reviewed)

References for Day 1

- "Detecting influenza epidemics using search engine query data"; J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant; Nature, vol. 457, no. 7232, pp. 1012-1014, 2008.
- "Dating Research from OkCupid"; http://blog.okcupid.com/, 2011.
- "Butterballs or Cheese Balls, an Online Barometer"; Kim Severson; The New York Times, http://www.nytimes.com/2009/11/26/dining/26search.html, 2009.

Questions?

End of Day 1

ingmar@yahoo-inc.com