

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

Motivating Example: Google Flu Trends

Examples

- Show live:

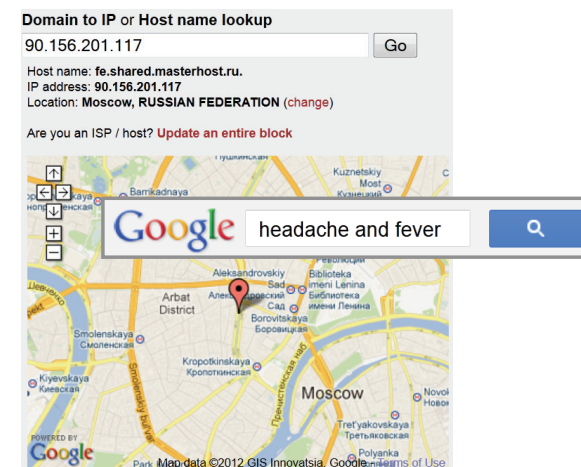
<http://www.google.org/flutrends/>

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

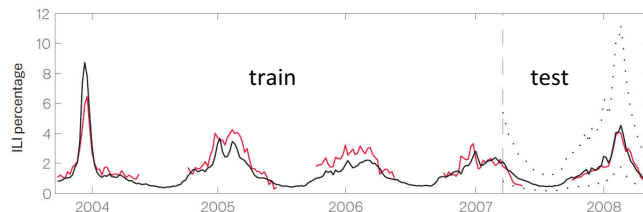


How Does it Work (2/3)

- 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 explanatory 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])
 - $\text{logit}(P) = \beta_0 + \beta_1 X_1 + e$, where $\text{logit}(P) = P/(1-P)$

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

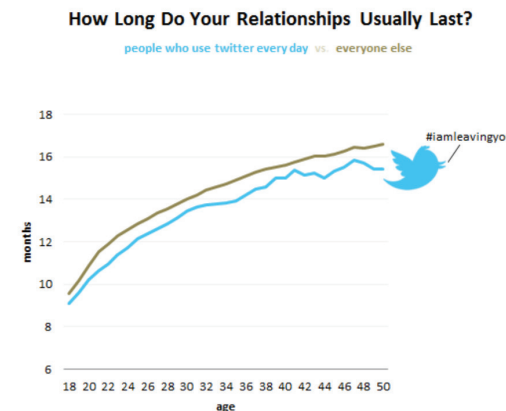


A Prime Example of Web Science

- 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 techniques
- Available for wider user
 - Web demo and also Google Correlate (later)

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?

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?

15% couples vs. 8% non-couples

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

Don't Scrape – Even if it's Tempting

Reviews Written by PghYinzer "more of myself after winter" (Pittsburgh, PA United States)

Customer Reviews: 231
Top Reviewer Ranking: 608
Helpful Votes: 3772

Page: 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11-20

Topeak BabySeat with Rack - TCS2101
Offered by Brands Cycle and Fitness
Price: \$141.95
Availability: In Stock

What do parents buy? Both before and after birth.

★★★★☆ **Seat is great but rack will not work with many bikes**, July 1, 2012

This review is from: **Topeak BabySeat with Rack - TCS2101 (Sports)**

I have the older version of this seat and was looking to get the newer version because it is rated to hold 8 pounds more of kid. One of my girls is right at the 40 pound mark. I had no trouble mounting the old style (Baby Seat I) on my bike and theoretically, if you have the seat stay mounts, this should be just as easy. But my bike doesn't have them! So I am stuck with the old version. The seat itself as far as I can see is pretty similar and we're perfectly happy with it from that end. Even with my 40 pounder, my bike handles just fine and my daughter says she's comfortable.

If you don't have the mounts, try to find an earlier version.

"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

Tom Beer
@TomBeerBooks
Newsday books editor. That's me on the LIRR with my book.
Long Island, New York <http://www.newsday.com/books>

Usually a real name: could get gender

1,597 TWEETS
612 FOLLOWING
716 FOLLOWERS

More profile information

User location

Tweets Micro-blog entries

Nathan Englander @NathanEnglander
Just like Proust ate. [instagr.am/p/MQjW5onan7/](http://instagram.com/p/MQjW5onan7/)
Retweeted by Tom Beer
View photo

Elinor Lipman @ElinorLipman
It's blazing hot & crops are dying/Climate change?The Right's not buying/Mitt's working on it, tho on break/His answer: jet ski on a lake
Retweeted by Tom Beer
Expand

Tom Beer @TomBeerBooks
@keversa Where's my cut?
View conversation

Brian Ulicky @bulicks
Only the screams of tennis queens break the muggy silence of the

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

The screenshot shows the Yahoo! Answers homepage. At the top, there's a search bar and navigation links: HOME, BROWSE CATEGORIES, MY ACTIVITY, and ABOUT. Below this is a green banner with 'Ask' and 'Answer' buttons. A search bar with the text 'What are you looking for?' is also present. The main content area features a question: 'How long would a coach trip from moscow to yaroslavl take?' by user 'Macheggia'. The question has 2 answers. Below the question, there's a 'Best Answer' section chosen by voters, which states: 'It takes about 5.5 hours to get from Moscow to Yaroslavl by coach. Heavily depends on traffic (to tell the truth the traffic is awful on the way out of Moscow to Yaroslavl). I would recommend you to use a train departing from Yaroslavl train station in Moscow (price starting from about \$25 / 4 hours). There are also commuting trains (elektrichka) from the same railway station for Yaroslavl. They are less comfortable, more stops but...'. On the right side, there's a sidebar with 'All Categories' listed, including Arts & Humanities, Education & Reference, Home & Garden, Science & Nature, Beauty & Style, Entertainment & Music, Local Businesses, Social Sciences, Business & Finance, Environment, News & Events, Society & Culture, Cars & Transportation, Family & Relationships, Pets, Sports, Computers & Internet, Food & Drink, Politics & Government, Travel, Consumer Electronics, Games & Recreation, Pregnancy & Parenting, and Yahoo! Products. The 'Travel' category is circled in red.

Data: Wikipedia

The screenshot shows the Wikipedia article for 'Yaroslavl'. The article text states: 'Yaroslavl (Russian: Ярославль, Yaroslavl'; IPA: [jɐrɐˈslavlʲ]) is a city and the administrative center of Yaroslavl Oblast, Russia, located 250 kilometers (160 mi) northeast of Moscow. The'. Below the article text, there's a section for 'Yaroslavl: Revision history'. This section shows a table of revisions with columns for date, time, user, and size. The first revision is by 'Voldemort175' on 01:07, 8 July 2012, with a size of (82,753 bytes) (+18). The second revision is by 'Chris the speller' on 21:35, 7 July 2012, with a size of (82,735 bytes) (-8). The table also includes links for 'cur', 'prev', 'newer', 'older', '50', '100', '250', and '500'.

Data: Wikipedia

- Complete edit history for all articles
- Some articles are geolocated
- 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

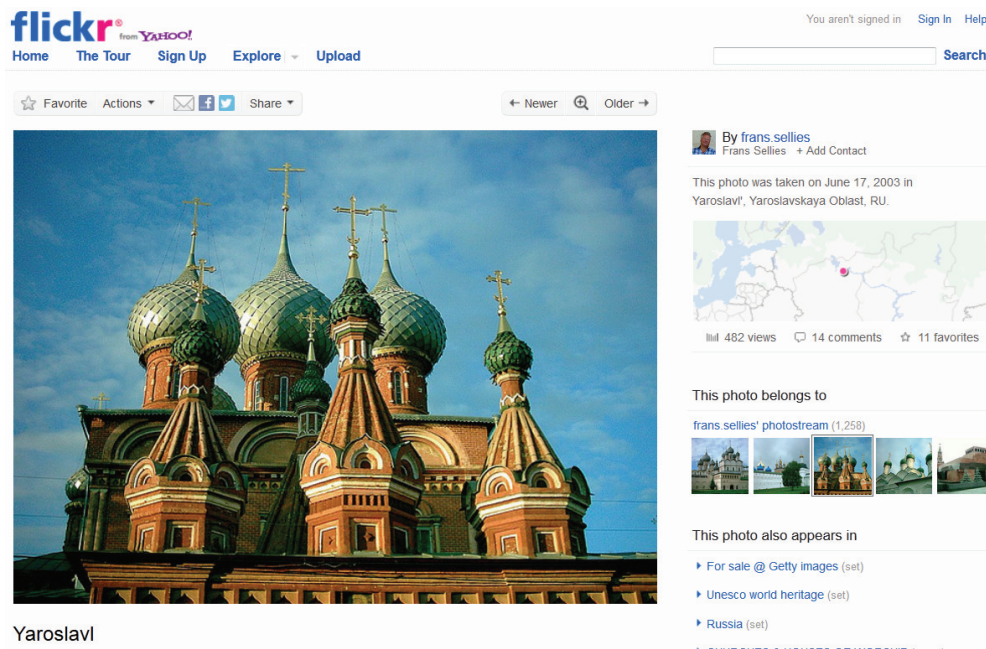
- Images come with
 - Tags, location, date, camera information, view count, ...
- Users come with
 - Profile page, list of contacts, group memberships, ...

- Search-based Flickr API

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

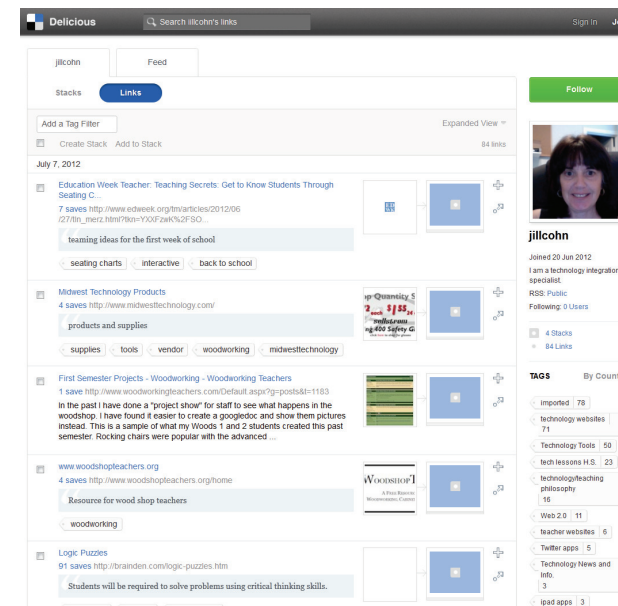
10,000 queries per hour

Data: Flickr



The screenshot shows a Flickr page for a photo of the Yaroslavl Cathedral. The photo is a large, ornate church with multiple green onion-shaped domes and golden crosses. The page includes the Flickr logo, navigation links (Home, The Tour, Sign Up, Explore, Upload), and a search bar. Below the photo, there is a caption: "This photo was taken on June 17, 2003 in Yaroslavl', Yaroslavl'skaya Oblast, RU." and a map showing the location. The photo has 482 views, 14 comments, and 11 favorites. It is attributed to "frans sellies" and is part of a "photostream (1,258)". The page also shows that the photo appears in "For sale @ Getty images (set)", "Unesco world heritage (set)", and "Russia (set)".

Data: Delicious



The screenshot shows a Delicious profile page for a user named "jilcohn". The page includes the Delicious logo, a search bar, and navigation links (Feed, Links, Stacks). The profile shows a photo of the user, a bio, and a list of links. The links are categorized by tags such as "education", "teaching", "technology", "products", "supplies", "tools", "vendor", "woodworking", "midwesttechnology", "first semester projects", "woodworking teachers", "logic puzzles", and "students". The page also shows a "Follow" button and a "Tags" section with a list of tags and their counts.

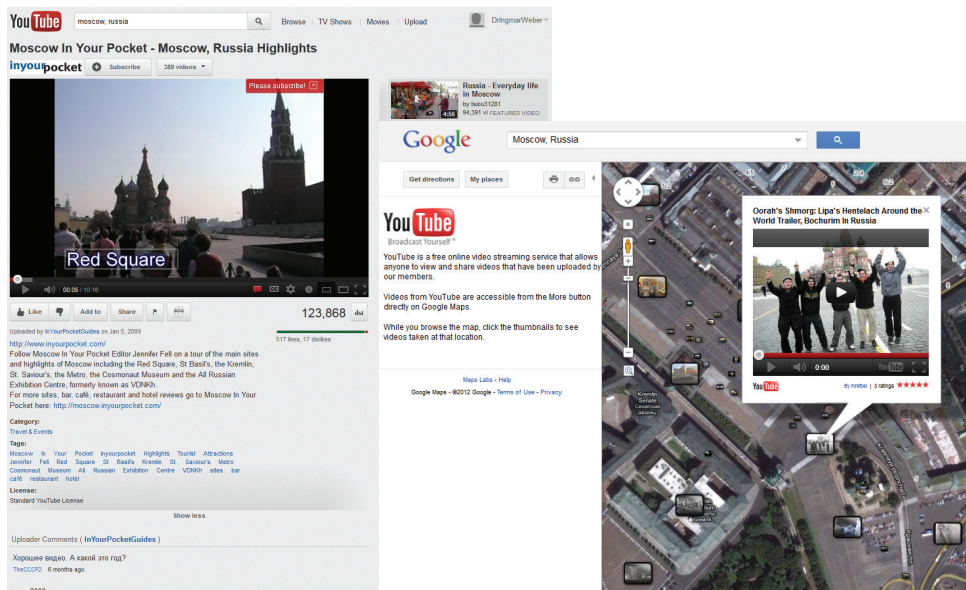
Data: Delicious

- 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

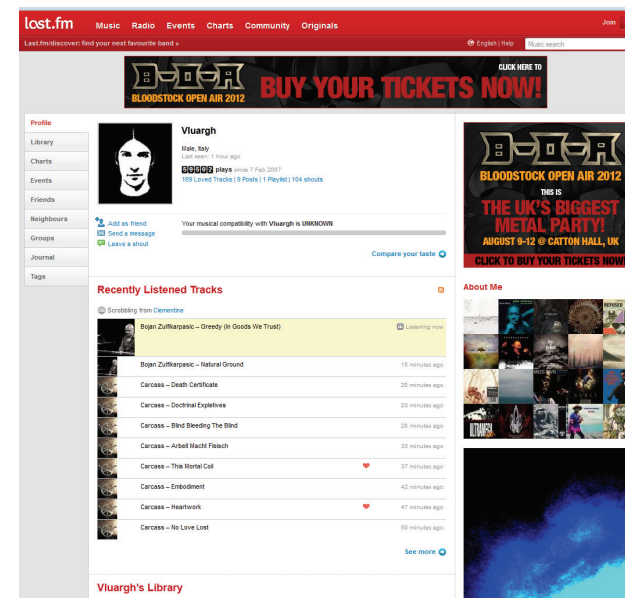
Data: Youtube

- 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: 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: 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: 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)

Questions?

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.

End of Day 1

ingmar@yahoo-inc.com