CS 277, Data Mining

Web Data Analysis: Part 3, Search Queries

Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Web Mining

Web = a potentially enormous "data set" for data mining

Multiple aspects of "Web mining"

1. Web Content
   - e.g., categorizing Web pages based on their text content
2. Web Connectivity/Link Analysis
   - e.g., characterizing distributions on path lengths between pages
   - e.g., determining importance of pages from graph structure
3. Web Usage
   - e.g., understanding user behavior from Web and search logs
4. Web Advertising
   - e.g., algorithms for optimizing which ads to show which users

All are interconnected/interdependent

- E.g., Google (and most search engines) use both content and connectivity

Learning from Search Query Data

Learning to Rank Retrieved Documents given a Query Q

- Early search engines (pre 2005) ranked docs based on heuristics, e.g.,
  - N documents contain the query strong Q
  - Rank the N documents for the user using features such as PageRank of the doc, number of times Q appears in the doc, does Q appear in the title of the doc, etc
  - Features are combined using human-generated weights

- This approach has some significant disadvantages
  - Is very difficult to maintain and tune the manual weights, expensive and labor-intensive
  - Does not take advantage of the billions of human clicks (feedback)

- Alternative: use machine learning to learn which features are relevant
  - This is the currently most common approach
  - Logistic regression and SVMs are widely used
  - Features are based on
    - How well the query matches a document (information-retrieval features)
    - Properties of the document itself (PageRank, length, time-stamp, etc)

Which Results do Users See? Eye-Tracking Data

From Joachims and Radlinksi, 2007

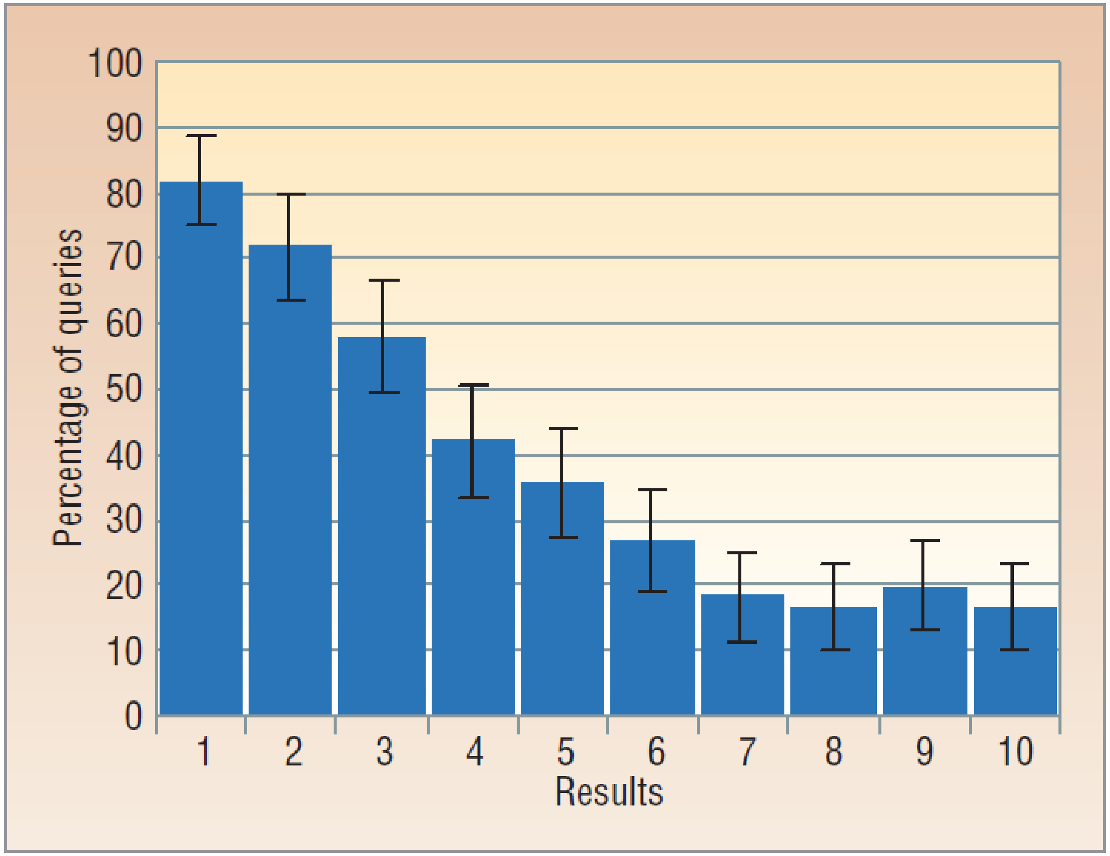


Figure 2. Rank and viewership. Percentage of queries where a user viewed the search result presented at a particular rank.

Presentation Bias: Order of Presentation Affects the Results

From Joachims and Radlinksi, 2007

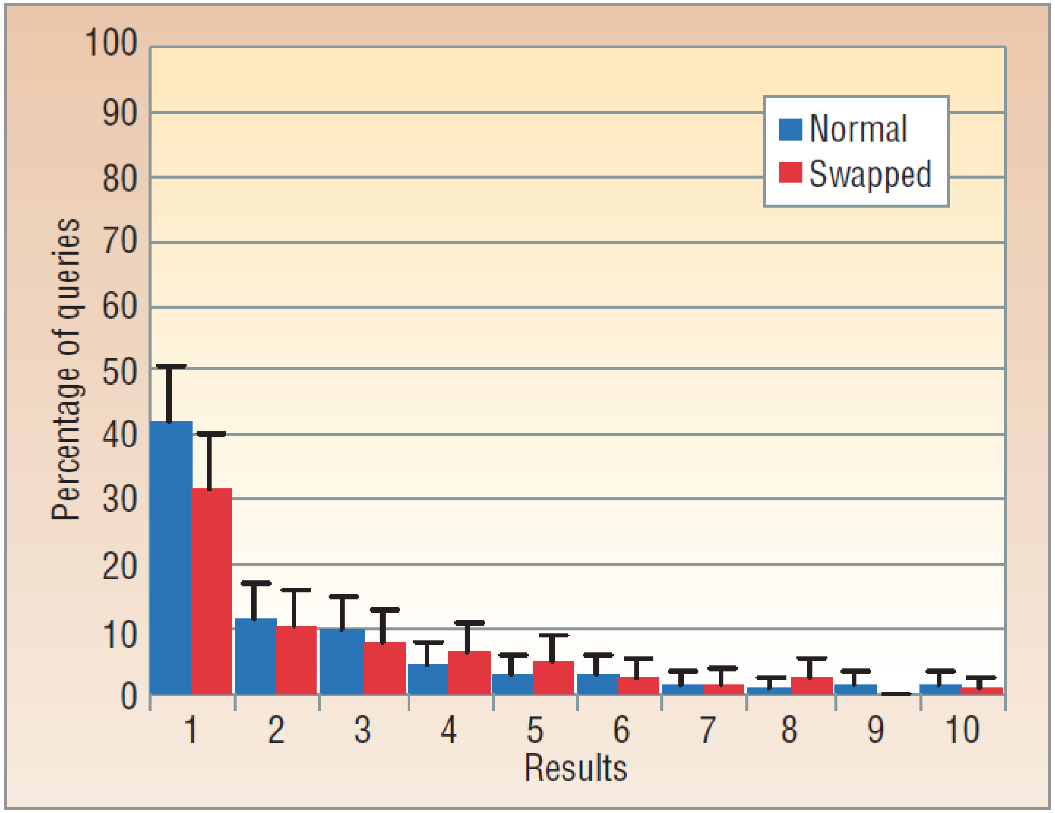


Figure 3. Swapped results. Percentage of queries where a user clicked the result presented at a given rank, both in the normal and swapped conditions.

Percentage of Results that users clicked where:

Blue = original order of results

Red = First and second results swapped

Note that position matters! The 2nd document in the first set of results gets 3 times more clicks when it is promoted to 1st position.

Learning to Rank

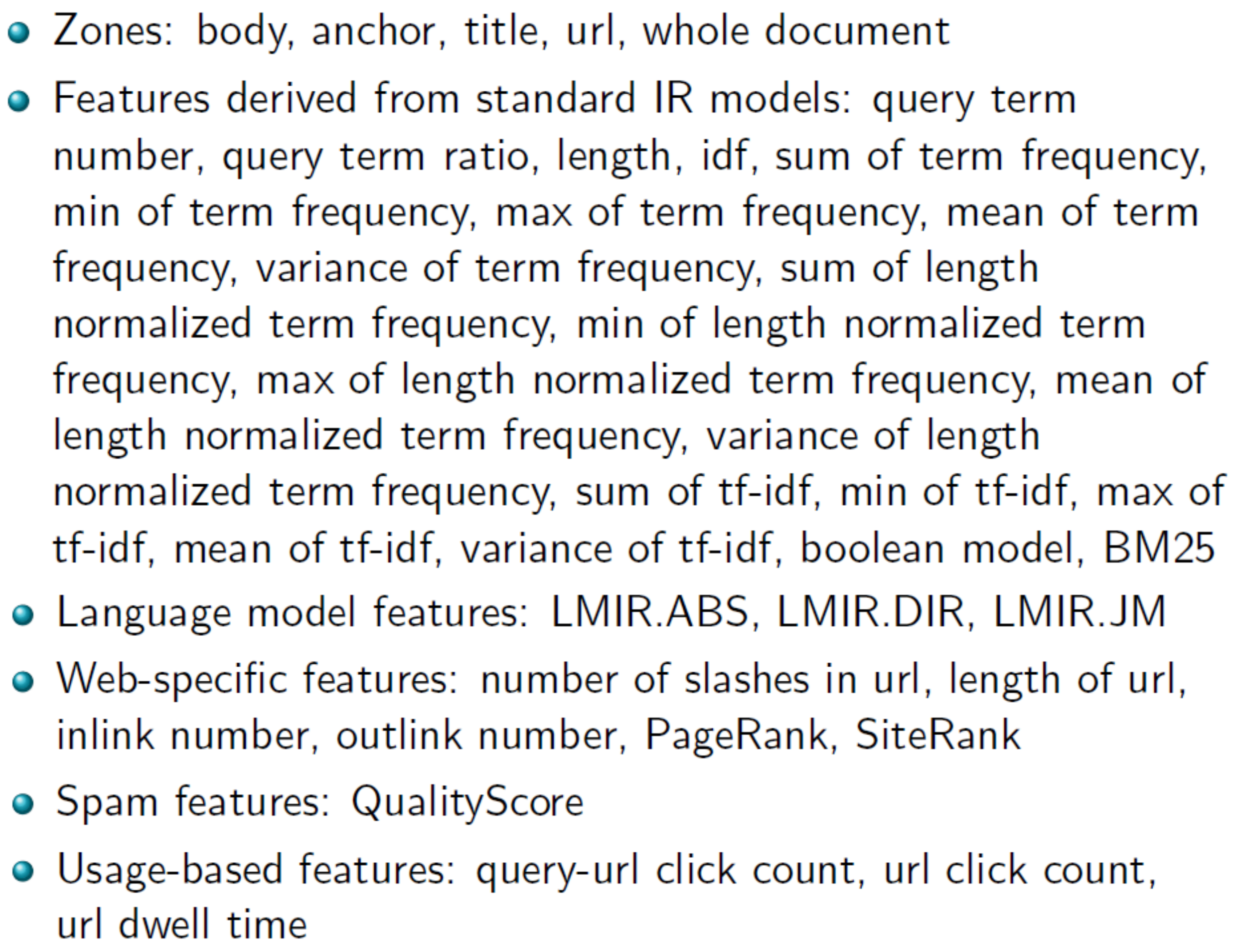
- Key lesson from presentation bias:
  - when evaluating whether a result is clicked on or not, it is important use context
  - context = rank position and other ads

- Example:
  - Say A is position 1, B in position 2, C in position 3
  - User 1 clicks on A in position 1: so we don't learn much about B or C
  - User 2 clicks on B in position 2: we can be much more certain that 2 prefers B over A
  - Thus, better (for user 2) to learn a ranking B > A, rather than "absolute" approach of learning that B is relevant, A not relevant, C not relevant

- Learning to Rank Frameworks
  - Incorporate rankings as constraints into learning algorithm
  - A variety of algorithms have been proposed this, e.g.,
    - Learn to discriminate between pairs of preferences
    - Learn on a full ranked list

Document-Ranking Features used by Microsoft Research

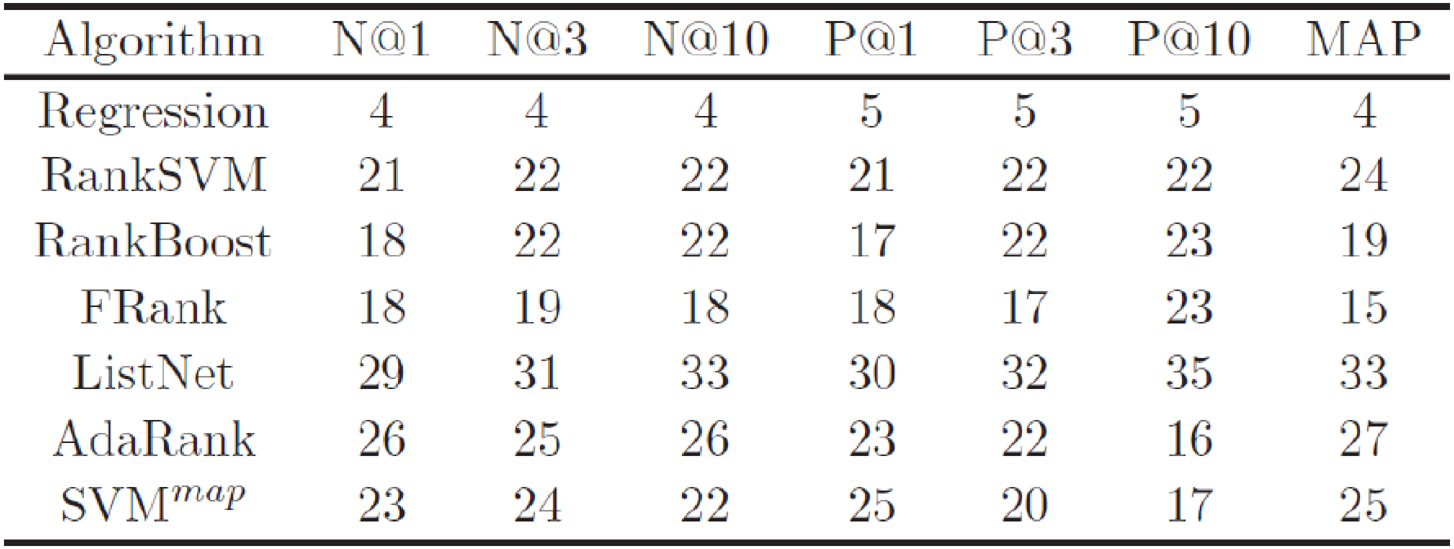
From H. Schutze, Information Retrieval Class, 2013
See also http://research.microsoft.com/en-us/projects/mslr/



- Zones: body, anchor, title, url, whole document
- Features derived from standard IR models: query term number, query term ratio, length, idf, sum of term frequency, min of term frequency, max of term frequency, mean of term frequency, variance of term frequency, sum of length normalized term frequency, min of length normalized term frequency, max of length normalized term frequency, mean of length normalized term frequency, variance of length normalized term frequency, sum of tf-idf, min of tf-idf, max of tf-idf, mean of tf-idf, variance of tf-idf, boolean model, BM25
- Language model features: LMIR.ABS, LMIR.DIR, LMIR.JM
- Web-specific features: number of slashes in url, length of url, inlink number, outlink number, PageRank, SiteRank
- Spam features: QualityScore
- Usage-based features: query-url click count, url click count, url dwell time

Experimental Results: Multiple Algorithms, Multiple Data Sets

Results from Tie-Yan Liu, WWW 2009 Tutorial on Learning to Rank



Algorithm	N@1	N@3	N@10	P@1	P@3	P@10	MAP
Regression	4	4	4	5	5	5	4
RankSVM	21	22	22	21	22	22	24
RankBoost	18	22	22	17	22	23	19
FRank	18	19	18	18	17	23	15
ListNet	29	31	33	30	32	35	33
AdaRank	26	25	26	23	22	16	27
SVM^{map}	23	24	22	25	20	17	25

Rows = different algorithms
Columns = different performance metrics (P = precision, MAP = Mean Average Precision)

Numbers indicate how often each algorithm performed best across multiple data sets

Standard regression algorithm is clearly inferior to ranking-based algorithms

How can Search Query Data be Used?

- Create behavioral profile for individual users
  - Classify search queries into predefined categories/taxonomy
  - Represent user as a “temporally-evolving” category vector
  - Widely used in advertising

- Infer user attributes or intents
  - “Intent classifier”: track search queries over time to try to infer if a user is trying to complete a particular action, e.g.,
    - Purchase a {car, house, camera, ….}
    - Organize a trip to ….
    - Make a will, get a divorce, …
  - Demographic classifier
    - Try to infer gender/age/income for a user from their search queries
  - Predictive modeling used..but these are difficult problems

- Use aggregate search data for forecasting, trend detection

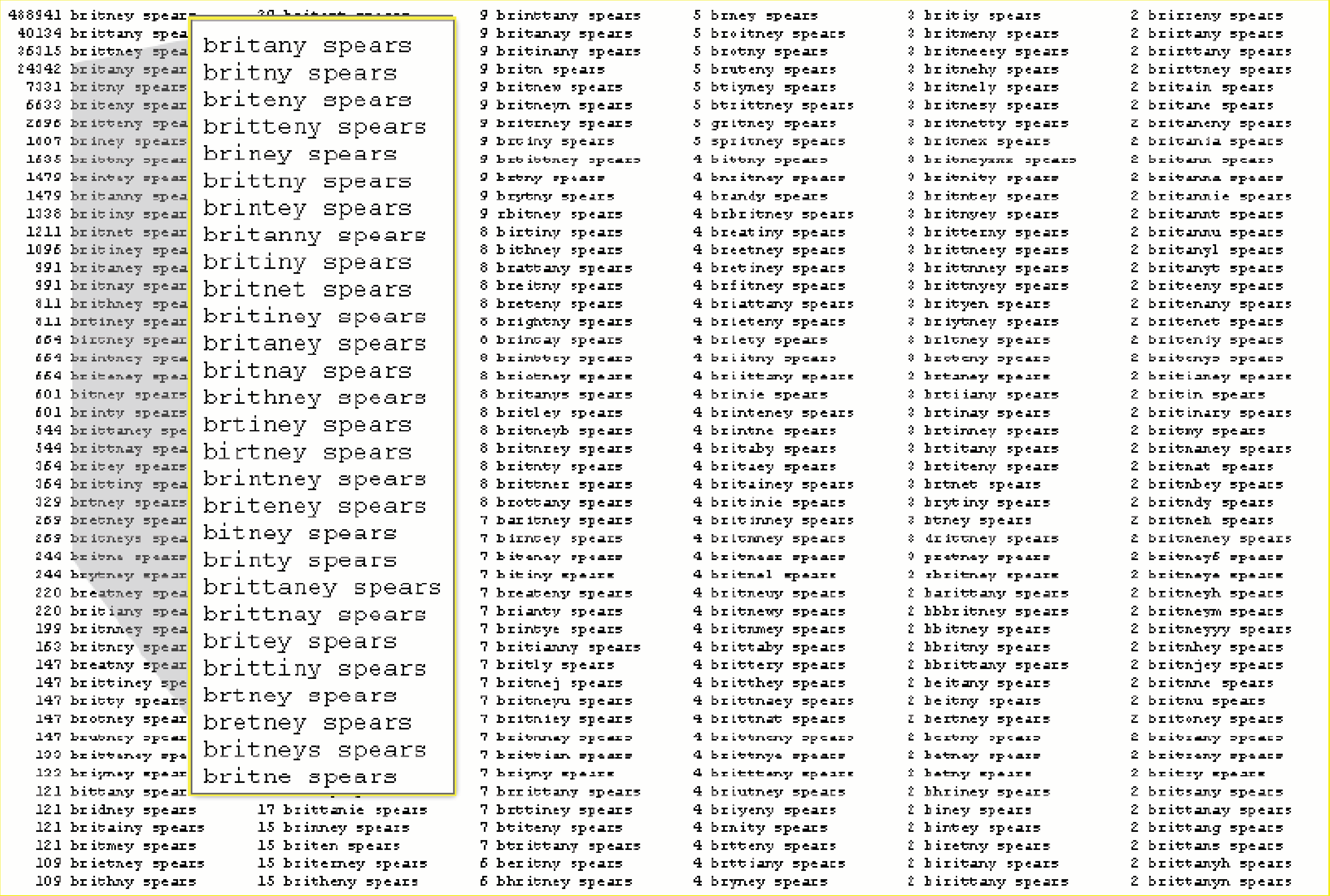
Types of Search Queries

(from Broder, 2002)

- Informational
  - User wants to learn about a specific topic
  - e.g., Query = [Data Privacy]
- Navigational
  - User wants to go to that page
  - Query = [Bank of America]
- Transactional
  - User wants to do something on a Web page, e.g.,
    - Buy a book
    - Plan a trip
- Other
  - E.g., user is “exploring” a topic

Logs of Search Query Data

source: Dan Russell, Google



```
488941 britney spears
 40134 brittany spea
 36315 brittney spea
 24342 britany spear
  7331 britny spears
  6633 briteny spear
  2696 britteny spea
  1807 briney spears
  1635 brittny spear
  1479 brintey spear
  1479 britanny spea
  1338 britiny spear
  1211 britnet spear
  1096 britiney spea
   991 britaney spea
   991 britnay spear
   811 brithney spea
   811 brtiney spear
   664 birtney spear
   664 brintney spea
   664 briteney spea
   601 bitney spears
   601 brinty spears
   544 brittaney spe
   544 brittnay spea
   364 britey spears
   364 brittiny spea
   329 brtney spears
   269 bretney spear
   269 britneys spea
   244 britne spears
   244 brytney spear
   220 breatney spea
   220 britiany spea
   199 britnney spea
   163 britnry spear
   147 breatny spear
   147 brittiney spe
   147 britty spears
   147 brotney spear
   147 brutney spear
   133 britteney spe
   133 briyney spear
   121 bittany spear
   121 bridney spears
   121 britainy spears
   121 britmey spears
   109 brietney spears
   109 brithny spears
```

```
britany spears
britny spears
briteny spears
britteny spears
briney spears
brittny spears
brintey spears
britanny spears
britiny spears
britnet spears
britiney spears
britaney spears
britnay spears
brithney spears
brtiney spears
birtney spears
brintney spears
briteney spears
bitney spears
brinty spears
brittaney spears
brittnay spears
britey spears
brittiny spears
brtney spears
bretney spears
britneys spears
britne spears
```

```
30 britnt spears
17 brittanie spears
15 brinney spears
15 briten spears
15 briterney spears
15 britheny spears
```

```
9 brinttany spears
9 britanay spears
9 britinany spears
9 britn spears
9 britnew spears
9 britneyn spears
9 britrney spears
9 brtiny spears
9 brtittney spears
9 brtny spears
9 brytny spears
9 rbitney spears
8 birtiny spears
8 bithney spears
8 brattany spears
8 breitny spears
8 breteny spears
8 brightny spears
8 brintay spears
8 brinttey spears
8 briotney spears
8 britanys spears
8 britley spears
8 britneyb spears
8 britnrey spears
8 britnty spears
8 brittner spears
8 brottany spears
7 baritney spears
7 birntey spears
7 biteney spears
7 bitiny spears
7 breateny spears
7 brianty spears
7 brintye spears
7 britianny spears
7 britly spears
7 britnej spears
7 britneyu spears
7 britniey spears
7 britnnay spears
7 brittian spears
7 briymy spears
7 brrittany spears
7 brttiney spears
7 btiteny spears
7 btrittany spears
6 beritny spears
6 bhritney spears
```

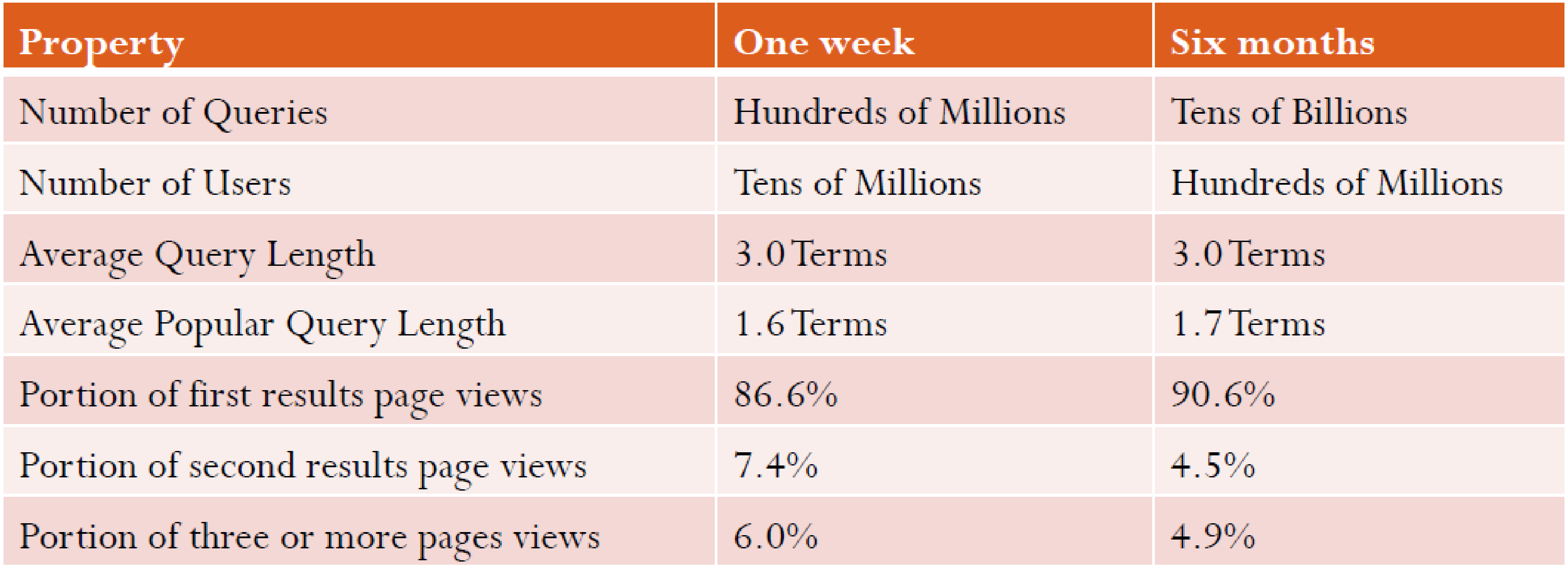
```
5 brney spears
5 broitney spears
5 brotny spears
5 bruteny spears
5 btiyney spears
5 btrittney spears
5 gritney spears
5 spritney spears
4 bittny spears
4 bnritney spears
4 brandy spears
4 brbritney spears
4 breatiny spears
4 breetney spears
4 bretiney spears
4 brfitney spears
4 briattany spears
4 brieteny spears
4 briety spears
4 briitny spears
4 briittany spears
4 brinie spears
4 brinteney spears
4 brintne spears
4 britaby spears
4 britaey spears
4 britainey spears
4 britinie spears
4 britinney spears
4 britmney spears
4 britnear spears
4 britnel spears
4 britneuy spears
4 britnewy spears
4 britnmey spears
4 brittaby spears
4 brittery spears
4 britthey spears
4 brittnaey spears
4 brittnat spears
4 brittneny spears
4 brittnye spears
4 britttany spears
4 briutney spears
4 briyeny spears
4 brnity spears
4 brtteny spears
4 brttiany spears
4 bryney spears
```

```
3 britiy spears
3 britmeny spears
3 britneeey spears
3 britnehy spears
3 britnely spears
3 britnesy spears
3 britnetty spears
3 britnex spears
3 britneyxxx spears
3 britnity spears
3 britntey spears
3 britnyey spears
3 britterny spears
3 brittneey spears
3 brittnney spears
3 brittnyey spears
3 brityen spears
3 briytney spears
3 briltney spears
3 brotony spears
2 brtaney spears
3 brtiiany spears
3 brtinay spears
3 brtinney spears
3 brtitany spears
3 brtiteny spears
3 brtnet spears
3 brytiny spears
3 btney spears
3 drittney spears
3 pretney spears
2 rbritney spears
2 barittany spears
2 bbbritney spears
2 bbitney spears
2 bbritny spears
2 bbrittany spears
2 beitany spears
2 beitny spears
2 bertney spears
2 bertny spears
2 betney spears
2 betny spears
2 bhriney spears
2 biney spears
2 bintey spears
2 biretny spears
2 biritany spears
2 birittany spears
```

```
2 brirreny spears
2 brirtany spears
2 brirttany spears
2 brirttney spears
2 britain spears
2 britane spears
2 britaneny spears
2 britania spears
2 britann spears
2 britanna spears
2 britannie spears
2 britannt spears
2 britannu spears
2 britanyl spears
2 britanyt spears
2 briteeny spears
2 britenany spears
2 britenet spears
2 briteniy spears
2 britenys spears
2 britianey spears
2 britin spears
2 britinary spears
2 britmy spears
2 britnaney spears
2 britnat spears
2 britnbey spears
2 britndy spears
2 britneh spears
2 britneney spears
2 britney6 spears
2 britneye spears
2 britneyh spears
2 britneym spears
2 britneyyy spears
2 britnhey spears
2 britnjey spears
2 britnne spears
2 britnu spears
2 britoney spears
2 britrany spears
2 britreny spears
2 britry spears
2 britsany spears
2 brittanay spears
2 brittang spears
2 brittans spears
2 brittanyh spears
2 brittanym spears
```

Yahoo Search Query Statistics

(from Broder and Josifovski, Computational Advertising, Stanford, 2011)

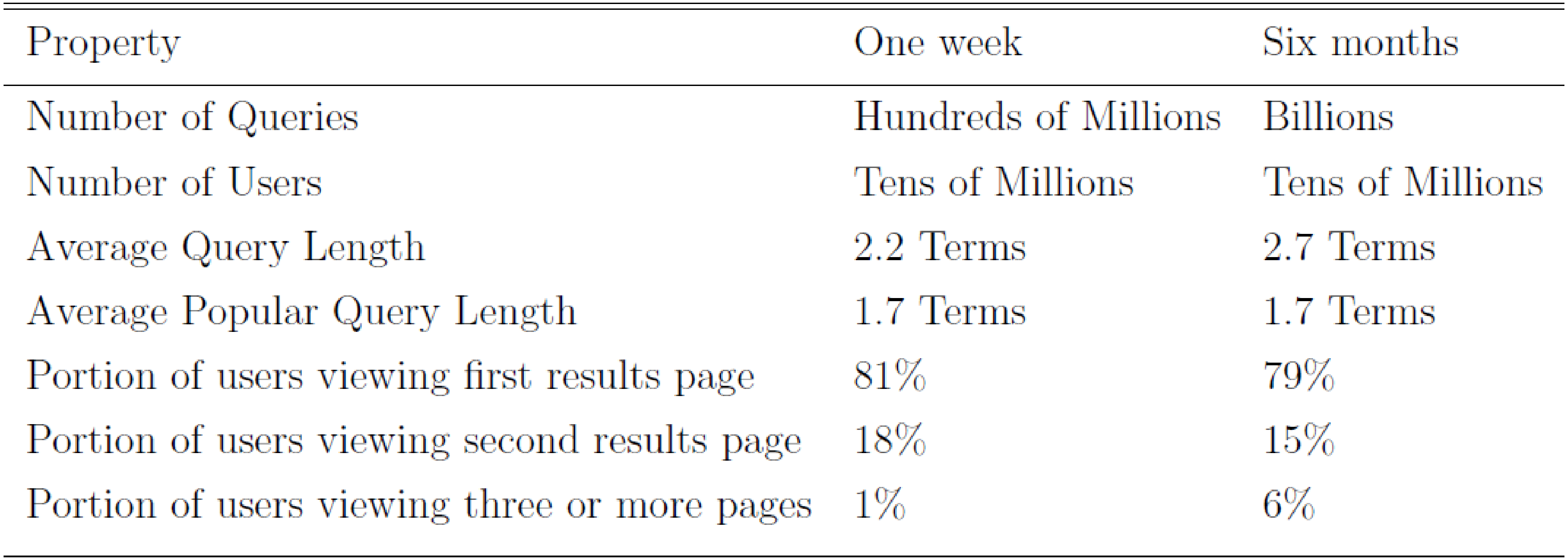


Property	One week	Six months
Number of Queries	Hundreds of Millions	Tens of Billions
Number of Users	Tens of Millions	Hundreds of Millions
Average Query Length	3.0 Terms	3.0 Terms
Average Popular Query Length	1.6 Terms	1.7 Terms
Portion of first results page views	86.6%	90.6%
Portion of second results page views	7.4%	4.5%
Portion of three or more pages views	6.0%	4.9%

AOL Search Query Statistics

(from Steven Beitzel, PhD Thesis, 2006)

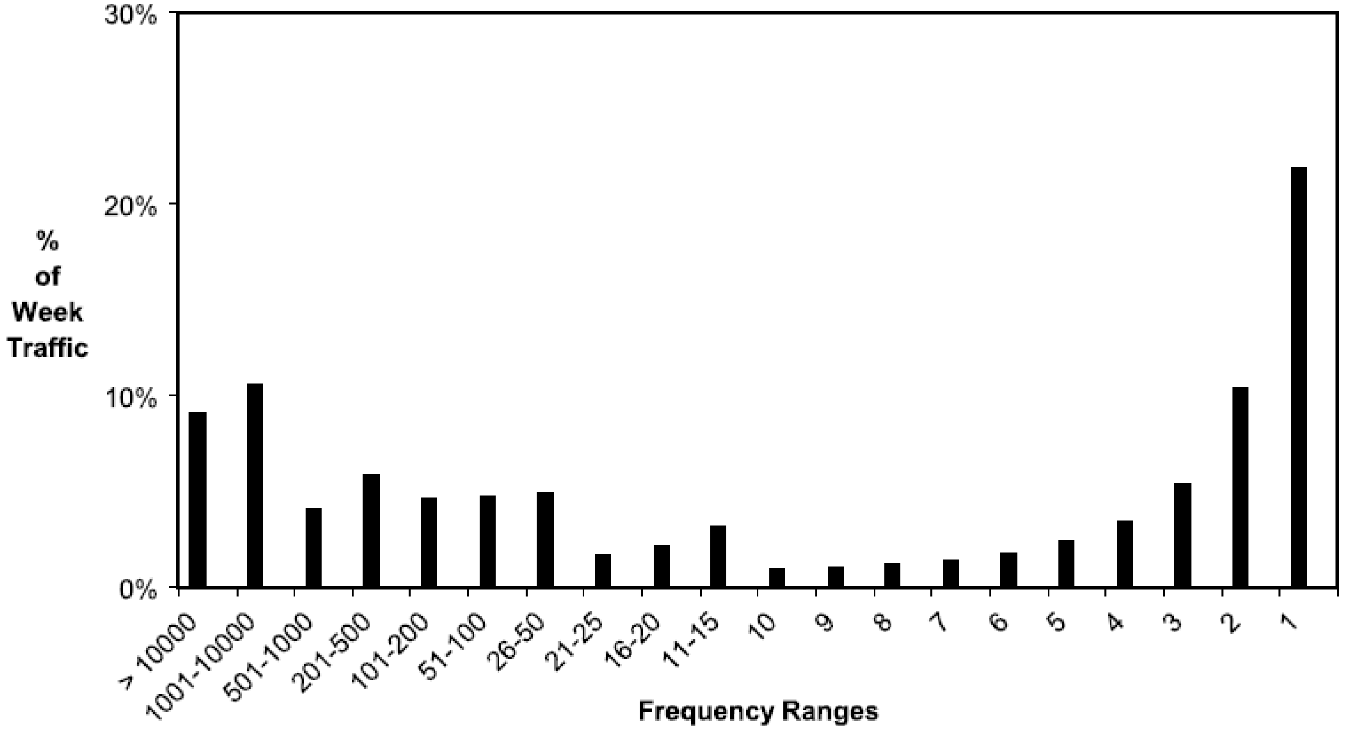
Table 2.1. Aggregate Query Log Statistics



Property	One week	Six months
Number of Queries	Hundreds of Millions	Billions
Number of Users	Tens of Millions	Tens of Millions
Average Query Length	2.2 Terms	2.7 Terms
Average Popular Query Length	1.7 Terms	1.7 Terms
Portion of users viewing first results page	81%	79%
Portion of users viewing second results page	18%	15%
Portion of users viewing three or more pages	1%	6%

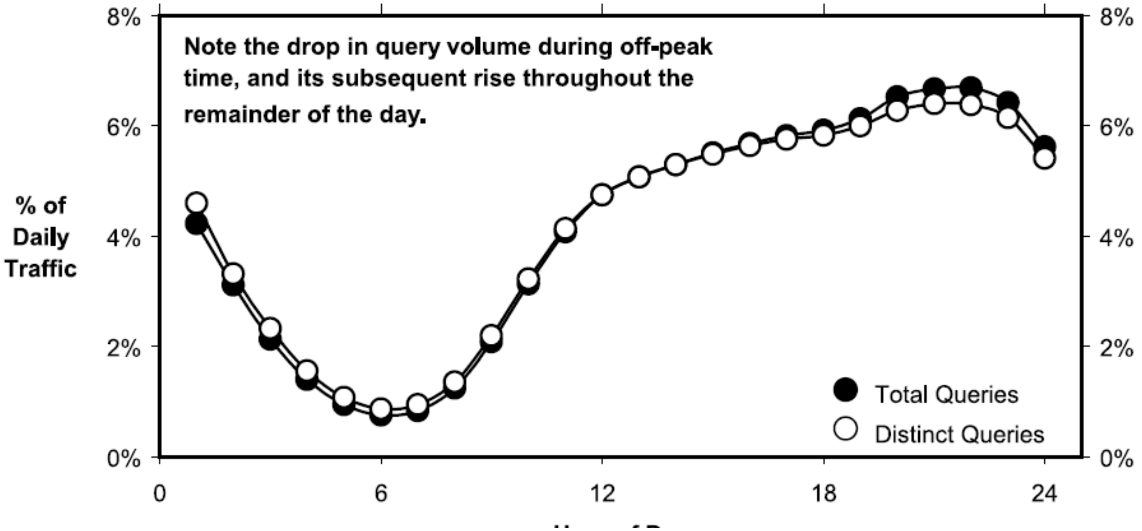
AOL Search Query Statistics: Frequency Distribution

(from Steven Beitzel, PhD Thesis, 2006)

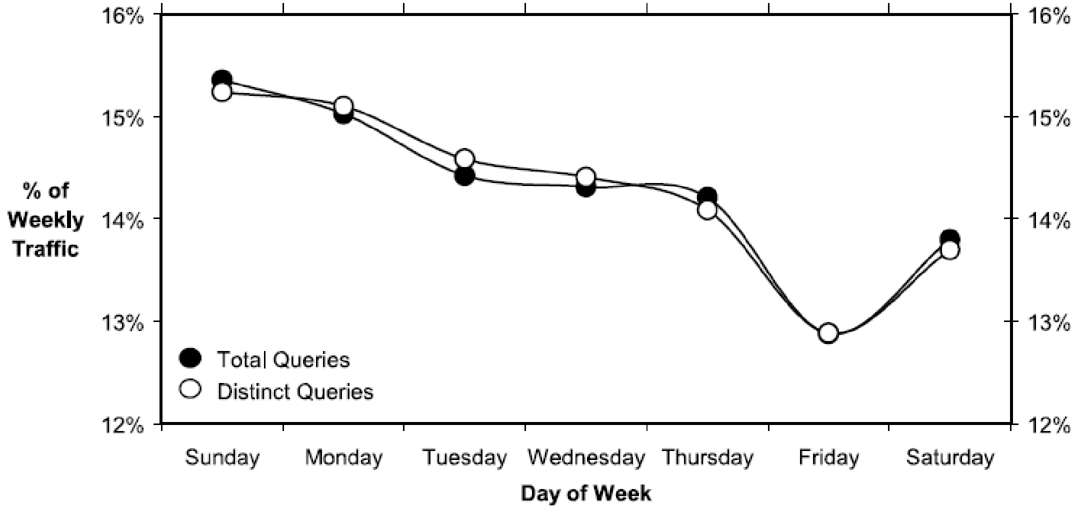


AOL Search Query Statistics: Temporal Patterns

(from Steven Beitzel, PhD Thesis, 2006)

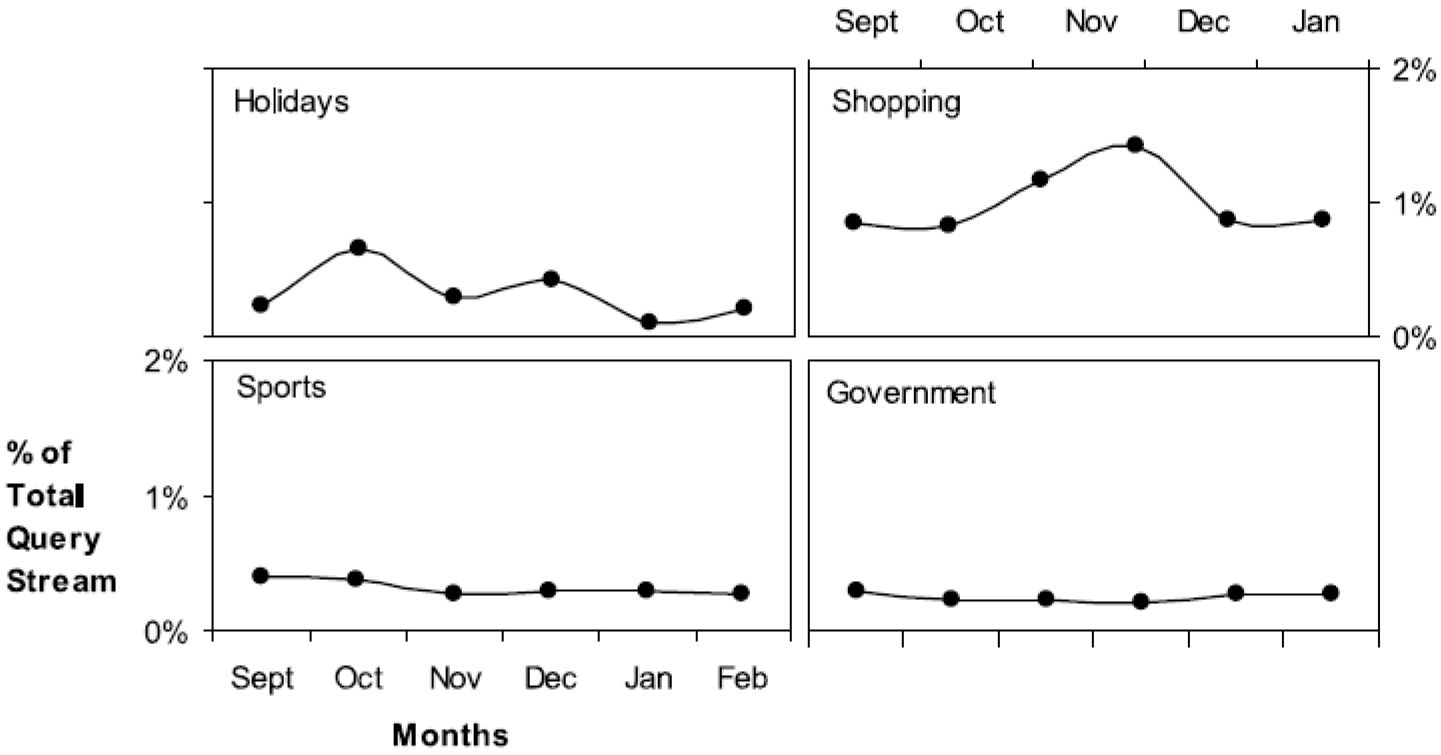


Note the drop in query volume during off-peak time, and its subsequent rise throughout the remainder of the day.



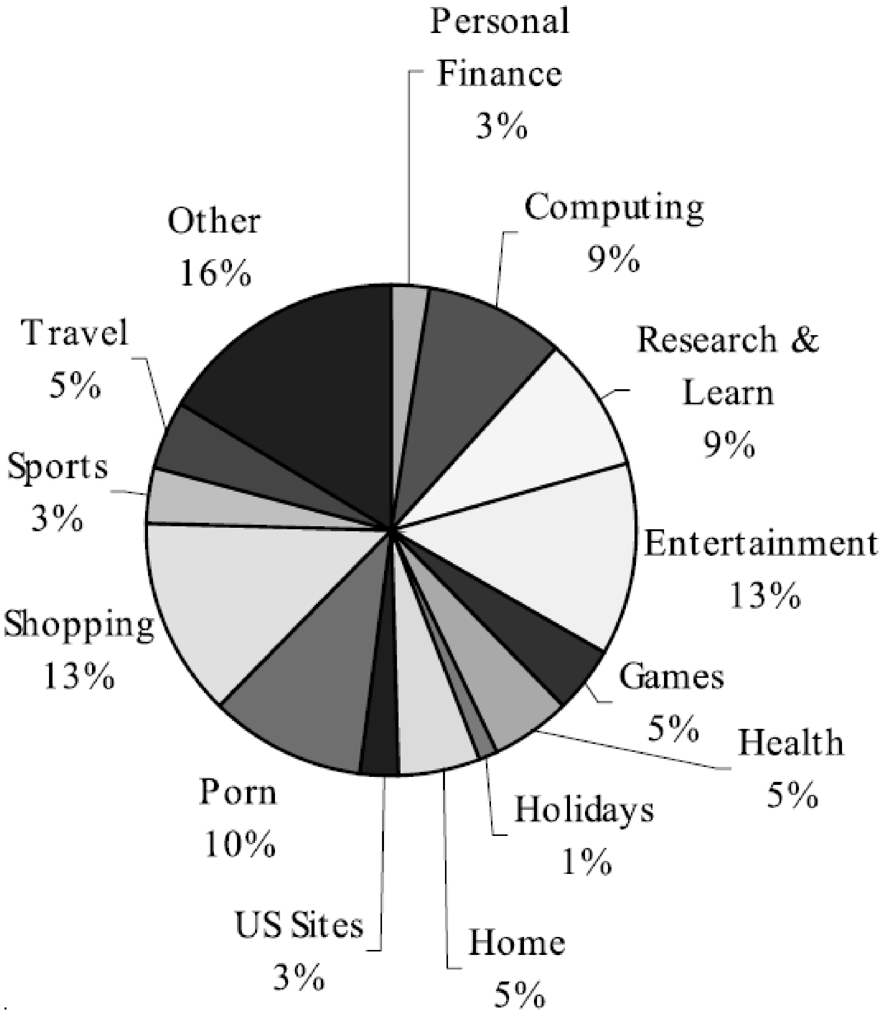
AOL Search Query Statistics: Temporal Patterns

(from Steven Beitzel, PhD Thesis, 2006)



AOL Search Query Statistics: Query Categories

(from Steven Beitzel, PhD Thesis, 2006)



Query categorization performed by matching query strings to lists generated by human editors

Life Cycles from Query Logs

(from Matt Richardson, 2008,
Learning about the World through Long-Term Query Logs)

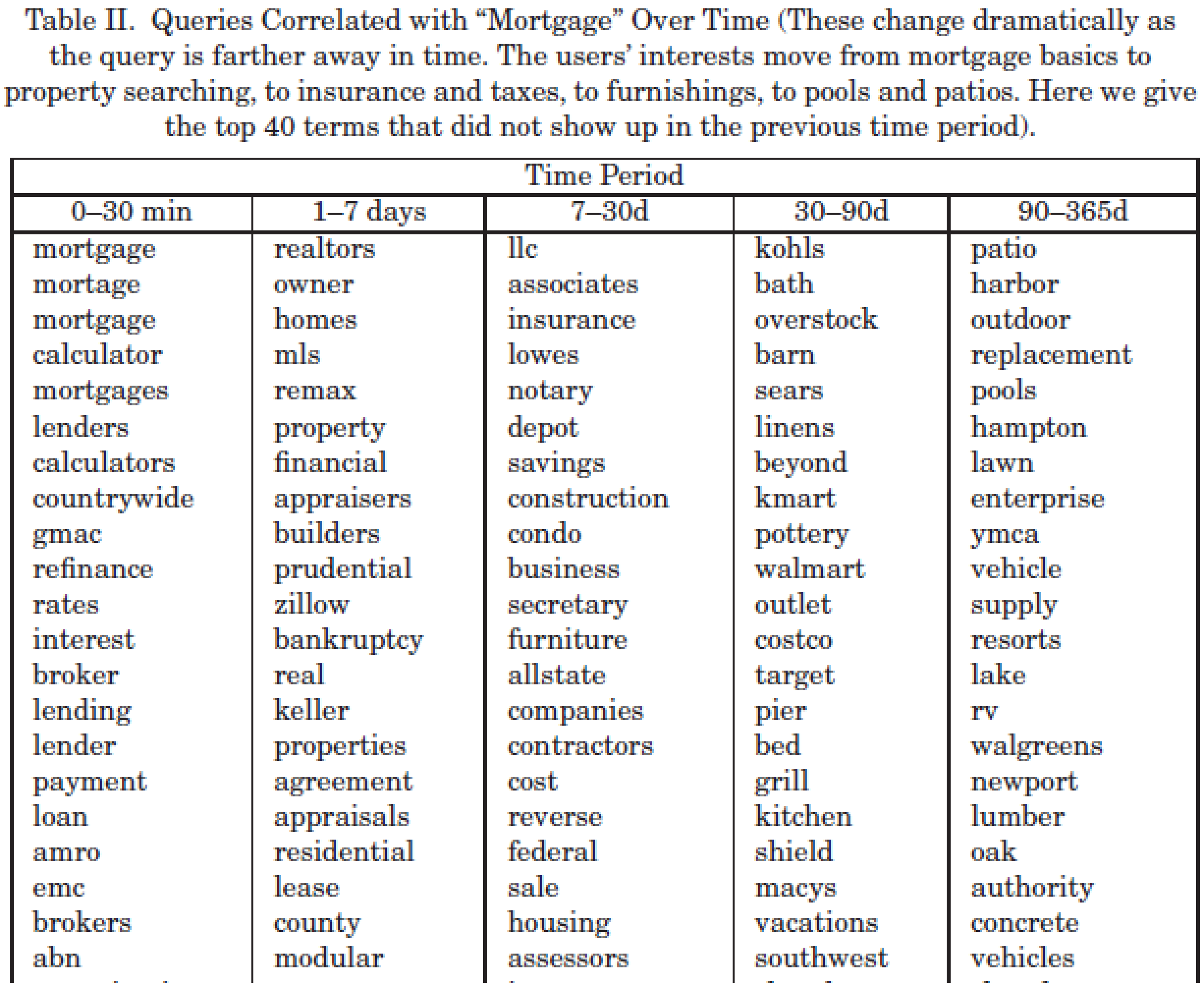
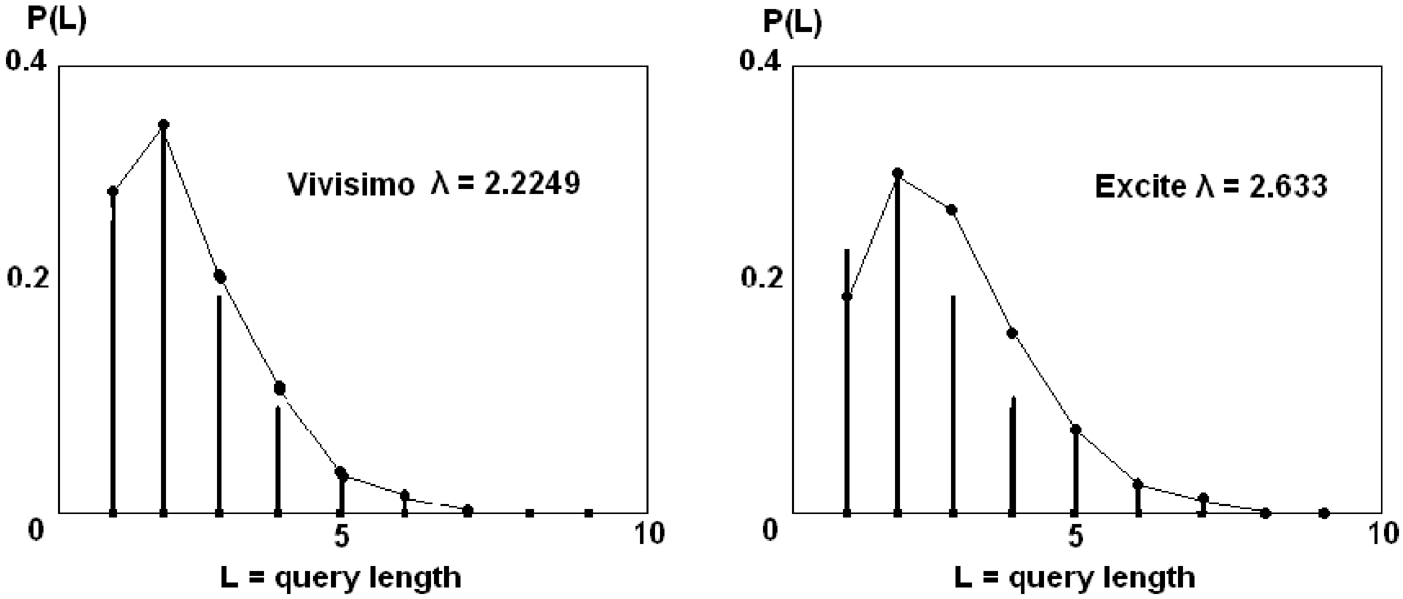


Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles

Typical Distributions of Number of Terms per Query

(from Xie and O Halloran, 2002)



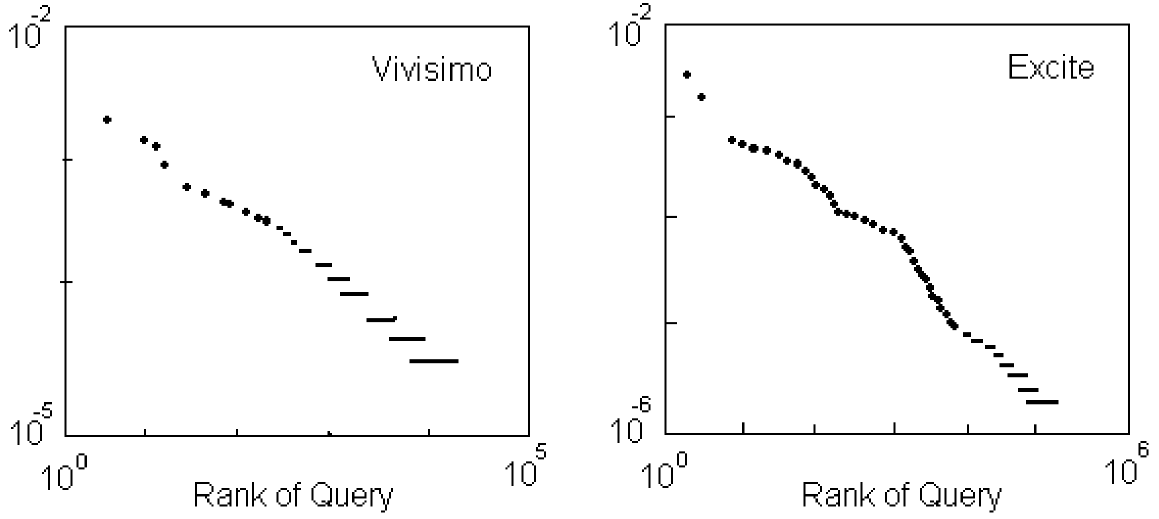
For two search engines, plots show

(a) Empirical query length distributions (bars), and

(b) Fitted Poisson model with parameter λ (dots with lines)

Power-law Characteristics of Distributions of Query Strings

(from Xie and O Halloran, 2002)

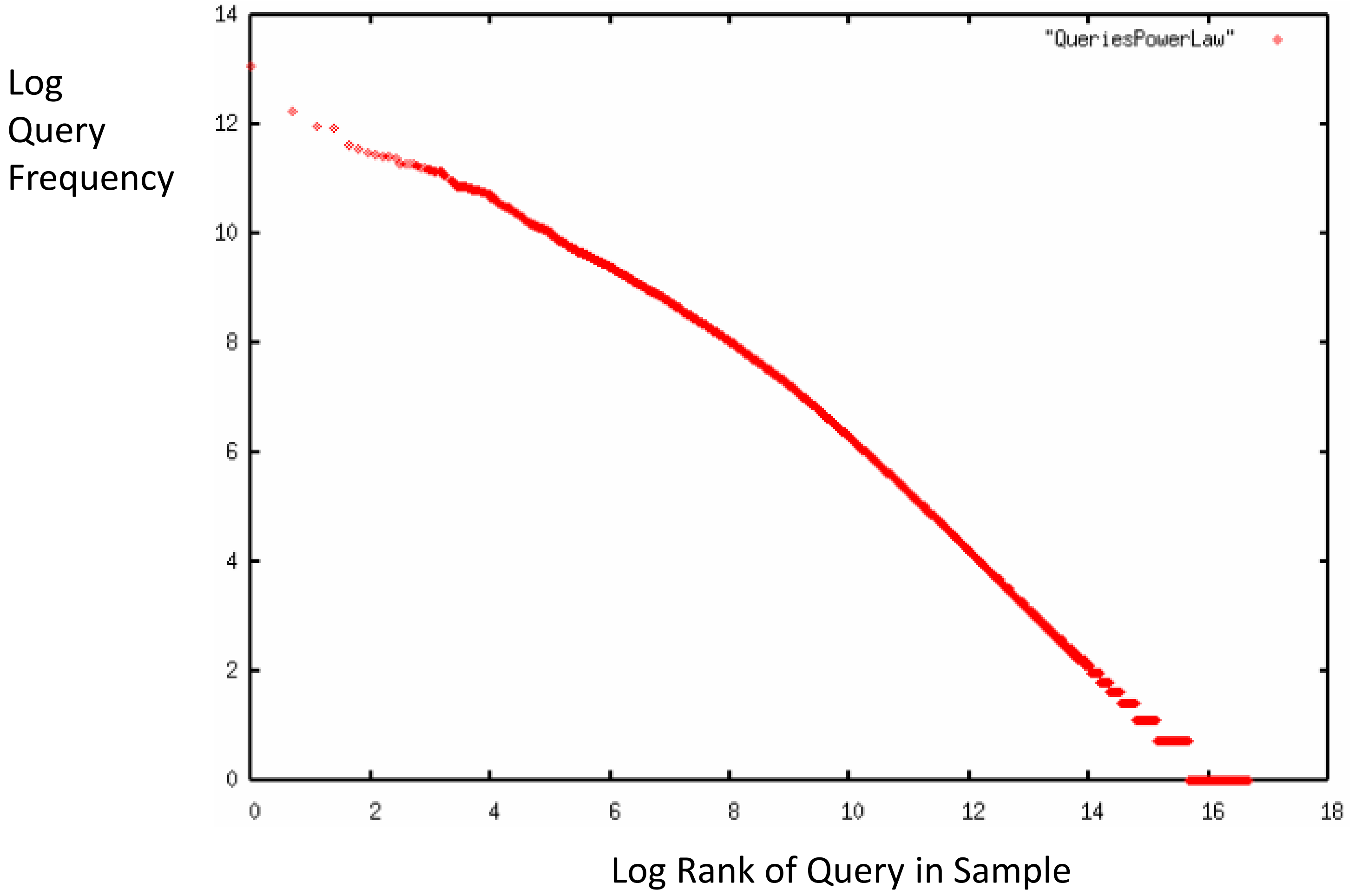


Approximately power-law behavior
(linear on log-log scale)

- Frequency f(r) of queries with rank r from 2 search engines
  - 110k queries from Vivisimo
  - 1.9 Million queries from Excite
- Note “long-tail” of rare queries (log-scale on both axes)
  - Presence of many rare queries makes prediction/ad-matching difficult

Sample of 17 Million Unique Queries from eBay

From N. Parikh and N. Sundaresan
Inferring Semantic Query Relations from Collective User Behavior
Proceedings of CIKM, 2008



From: M. Richardson, Microsoft Research
Learning about the World through Long-Term Query Logs
ACM Transactions on the Web, October 2008

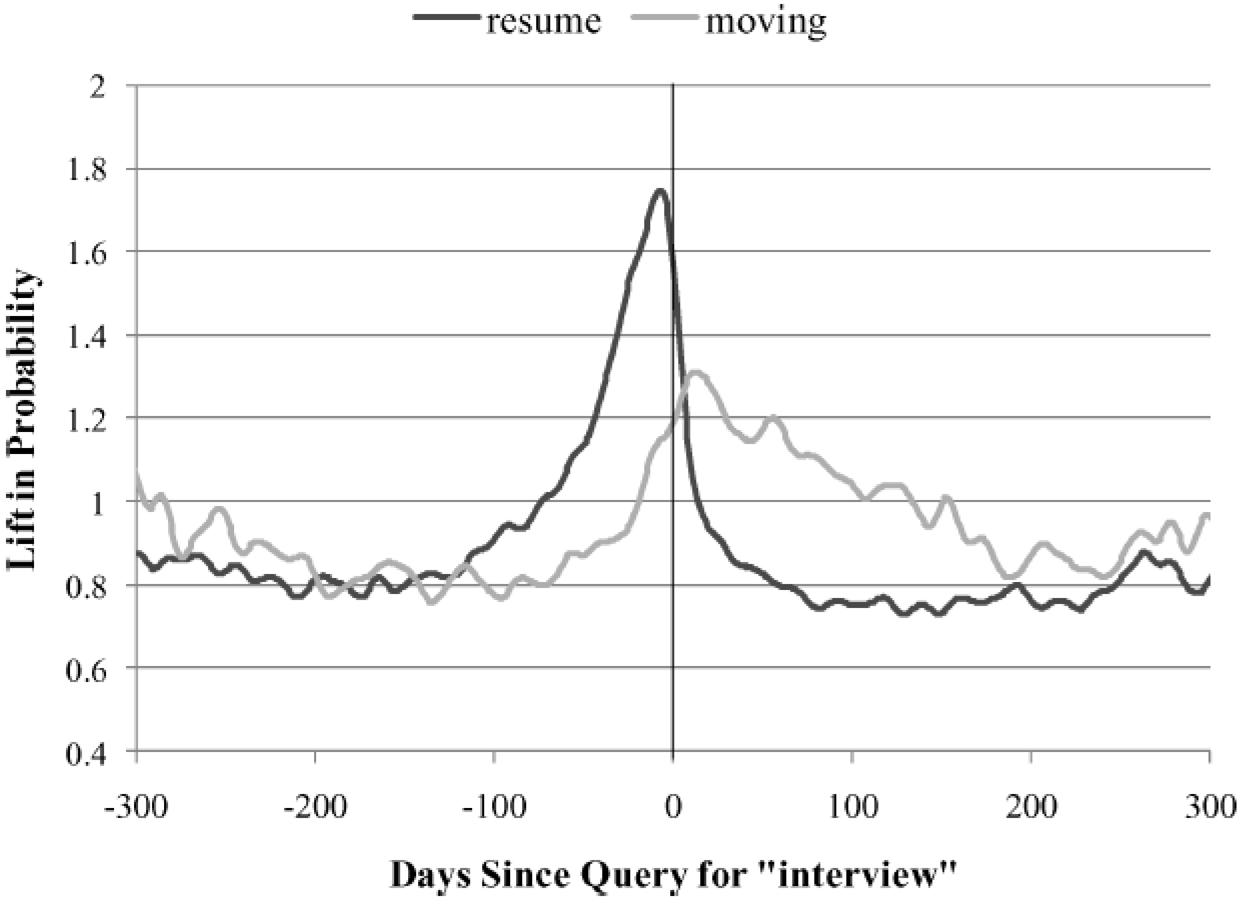
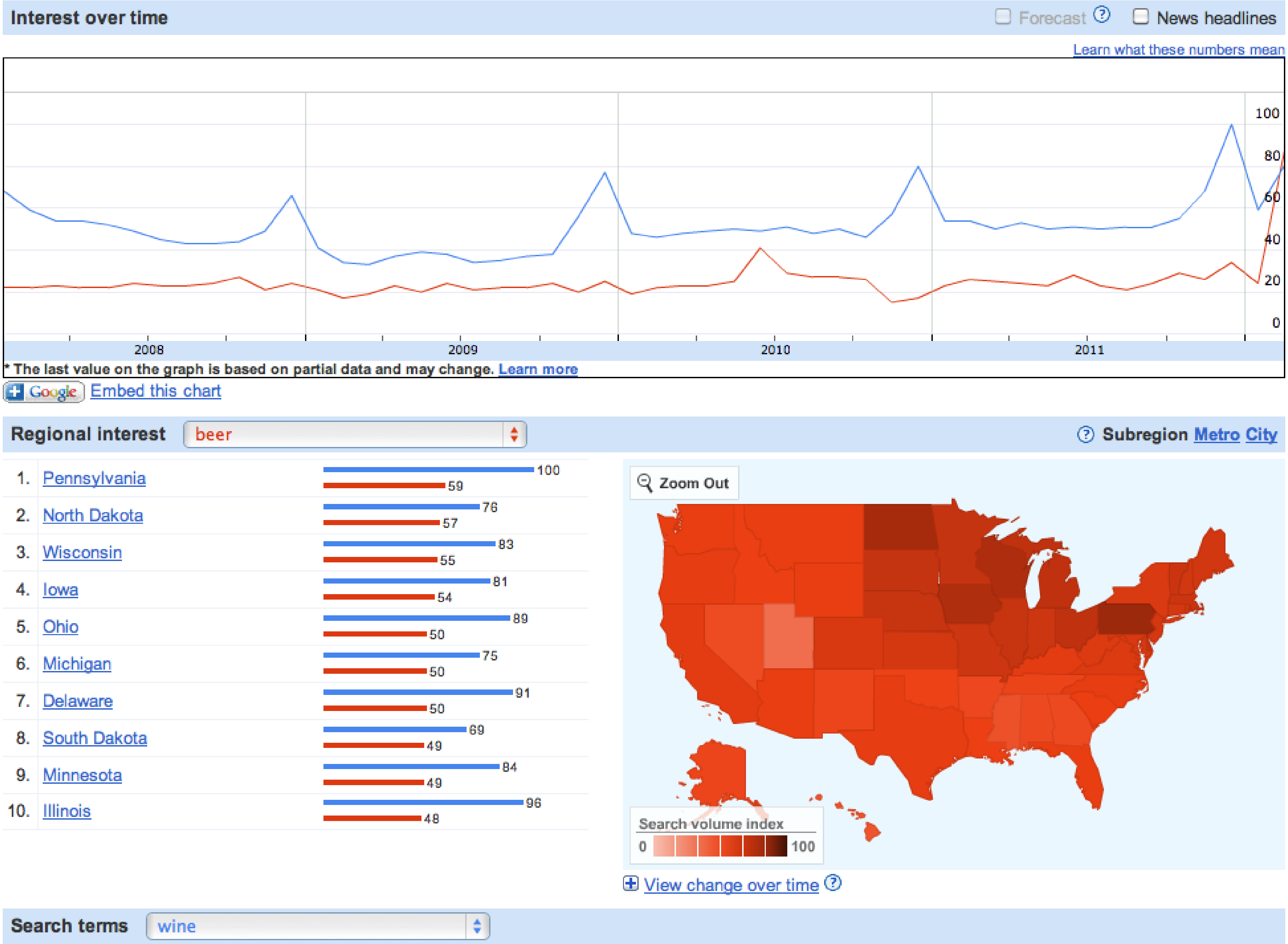


Fig. 4. $S(\delta)$ (lift in probability) for the queries “resume” and “moving” given the reference query “interview”. People begin looking for information on resumes up to 100 days before the interview query; most look immediately before. Users become significantly more interested in moving information after the interview query.

Publicly-Available Search Trend Data

- Google Insights Data
  - Weekly summaries of aggregate query data since 2004
  - Both general query categories and query specific
  - Can be localized to geographic areas
  - Numbers are proportions, not absolute counts



Interest over time

□ Forecast ⓘ □ News headlines

Learn what these numbers mean

* The last value on the graph is based on partial data and may change. Learn more

Google Embed this chart

Regional interest

beer

ⓘ Subregion Metro City

1.	Pennsylvania	100	59
2.	North Dakota	76	57
3.	Wisconsin	83	55
4.	Iowa	81	54
5.	Ohio	89	50
6.	Michigan	75	50
7.	Delaware	91	50
8.	South Dakota	69	49
9.	Minnesota	84	49
10.	Illinois	96	48

View change over time ⓘ

Search terms

wine

Predicting with Aggregated Search Queries

- *Predicting the Present with Google Trends*
  - Choi and Varian, 2009 (Varian is Google's chief economist)
  - Broad range of predictions (e.g., auto, homes, etc)
  - Systematic increases in predictive accuracy using Google queries

- *Forecasting Existing Home Sales using Google Search Engine Queries*
  - Duke economics thesis, B. D. Humphrey, 2010
  - Google queries improve accuracy of monthly forecasts of home sales
  - Up to 30% reduction in error
  - e.g., increase in queries related to unemployment and rentals are predictive of future decreases in home sales

EXAMPLE OF USING SEARCH ENGINE LOGS: DETECTING FLU OUTBREAKS

J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, L. Brilliant
Detecting influenza epidemics using search engine query data
Nature, Februrary 2009

Detecting Flu Outbreaks

- Problem:
  - Influenza epidemics cause 250k to 500k deaths per year
  - Motivates quick detection of an outbreak
- How are flu outbreaks currently detected?
  - CDC gathers counts of “influenza-like illness” (ILI) physician visits
  - Surveillance data published nationally and regionally, weekly basis
  - Problem: 1 to 2 week reporting lag
- Additional data?
  - Monitor over-the-counter flu medication sales
  - Monitor calls to health advice lines

Using Search Queries

- Idea:
  - Use influenza related search queries to predict ILI counts (historically)
  - Should be much faster than 1-2 week lag of CDC data

- Data:
  - Look at all search queries in Google from 2003 to 2008
    - Several hundred billion individual searches in the United States
    - Keep track of only the 50 million most common queries
    - Keep a weekly count for each query
    - Also keep counts of each query by geographic region

      (requires use of geo-location from IP addresses: >95% accurate)

  So counts for 50 million queries x 170 weeks x 9 regions

Building a Predictive Model

Target variable to be predicted:

For each week, for each region

$I(t)$ = percentage of physician visits that are ILI (as compiled by CDC)

Input variables:

$Q(t)$ = sum of top n highest correlated queries / total number of queries that week

Logistic Model:

$$\log(I(t) / [1 - I(t)]) = \alpha \log (Q(t)/ [1 - Q(t)]) + \text{noise}$$

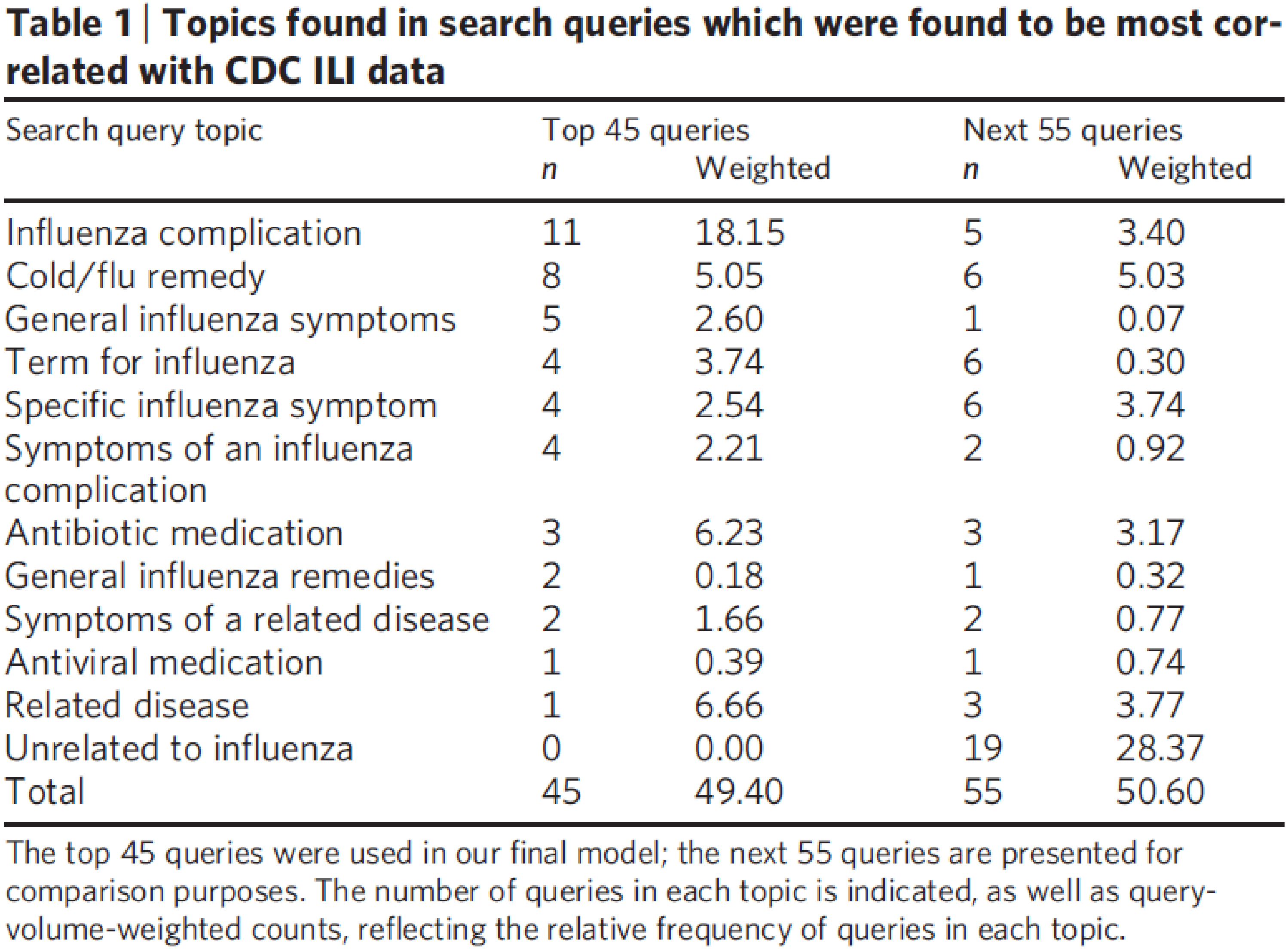


Table 1 | Topics found in search queries which were found to be most correlated with CDC ILI data

Search query topic	Top 45 queries		Next 55 queries	
	n	Weighted	*n*	Weighted
Influenza complication	11	18.15	5	3.40
Cold/flu remedy	8	5.05	6	5.03
General influenza symptoms	5	2.60	1	0.07
Term for influenza	4	3.74	6	0.30
Specific influenza symptom	4	2.54	6	3.74
Symptoms of an influenza complication	4	2.21	2	0.92
Antibiotic medication	3	6.23	3	3.17
General influenza remedies	2	0.18	1	0.32
Symptoms of a related disease	2	1.66	2	0.77
Antiviral medication	1	0.39	1	0.74
Related disease	1	6.66	3	3.77
Unrelated to influenza	0	0.00	19	28.37
Total	45	49.40	55	50.60

The top 45 queries were used in our final model; the next 55 queries are presented for comparison purposes. The number of queries in each topic is indicated, as well as query-volume-weighted counts, reflecting the relative frequency of queries in each topic.

From Ginsberg et al, Nature, 2009

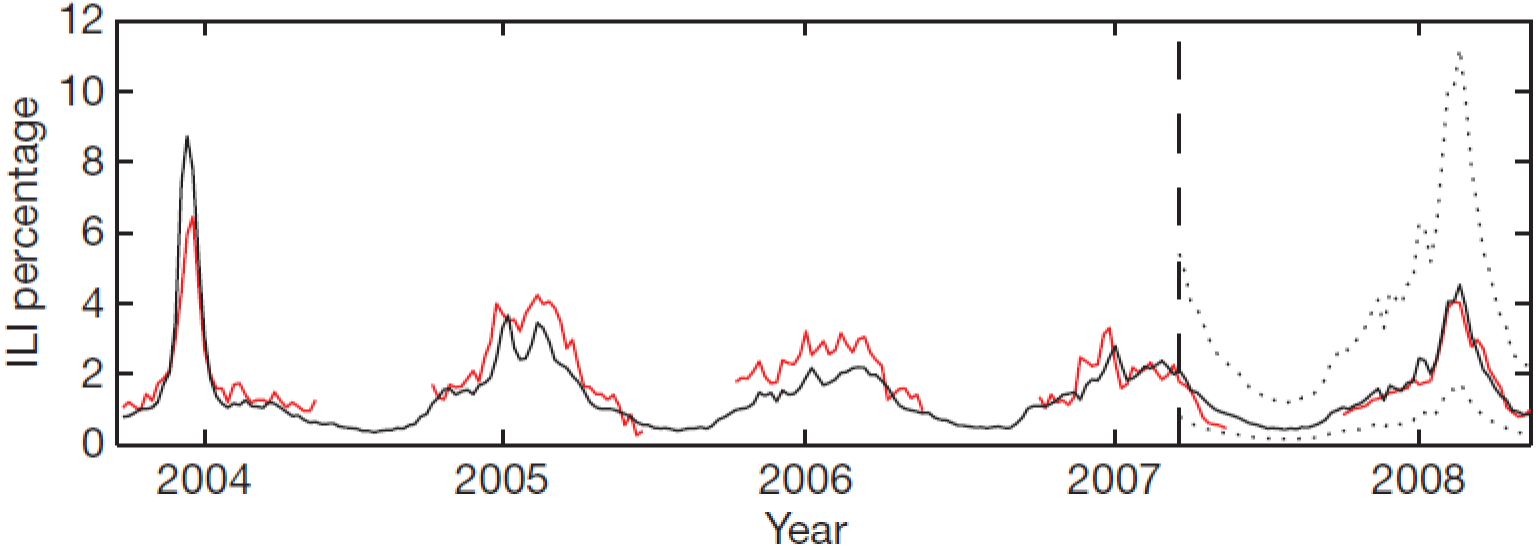


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

From Ginsberg et al, Nature, 2009

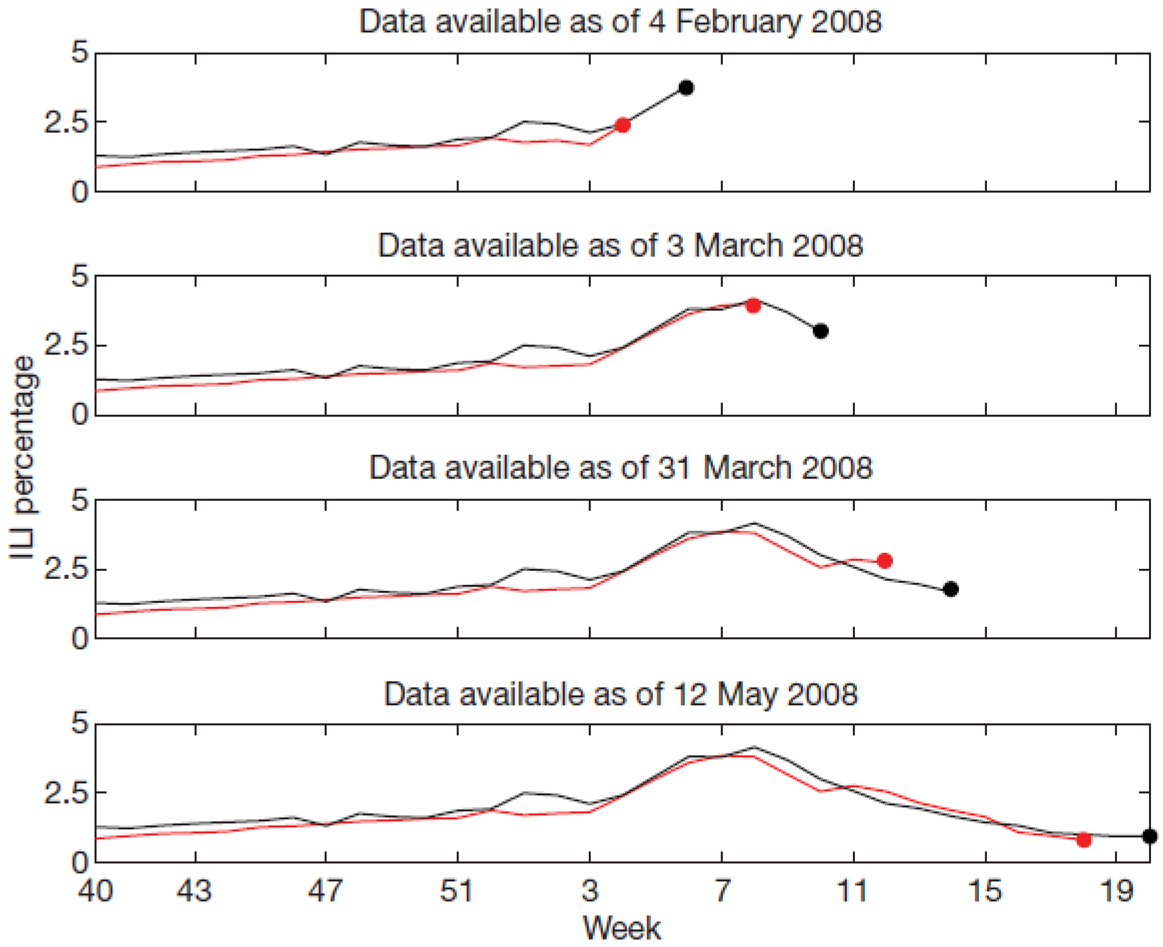


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3 March our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.

Key point in these graphs is that the CDC data was lagging Google predictions by 1 to 2 weeks

From Ginsberg et al, Nature, 2009

Class Presentations in Wednesday's Class

- Brief (5 minute) presentations for 8 selected projects
  - Start of class on Wednesday
  - Participants: please email me your slides by Wednesday afternoon (earlier is better)

- Selected projects chosen for a variety of factors
  - Good reports, interesting projects, diversity of projects
  - If your project was not selected for presentation it does not imply that your project is not highly ranked – the set of projects selected is not the top 8 projects in the class