An Introduction to Web Science

RuSSIR, Aug 6-10, 2012

Please interrupt at any point!!

Ingmar Weber

ingmar@yahoo-inc.com

Yahoo! Research Barcelona

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

"I know what you did last summer”

Query logs and user privacy

Rosie Jones, Ravi Kumar, Bo Pang and Andrew Tomkins

CIKM '07

The AOL Search Data Leak

- August 2006: AOL Research releases search log for 650,000 users – "anonymized"

```
AnonID  Query   ItemRank        ClickURL
1636218 www.airtime500.com      2       http://www.airtime500.com
2272416 theunorthodoxjew.blogspot.com   1       http://theunorthodoxjew.blogspot.com
172627  www.yahoolagins.com
2569723 www.homesforsale
1196769 zip codes       1       http://www.usps.com
724416  propertytaxsales.com
30011   schwab learning 1       http://www.schwablearning.org
[...]
```

- NY Times reporter identified

  AnonID 4417749 = Thelma Arnold, 62yrs old

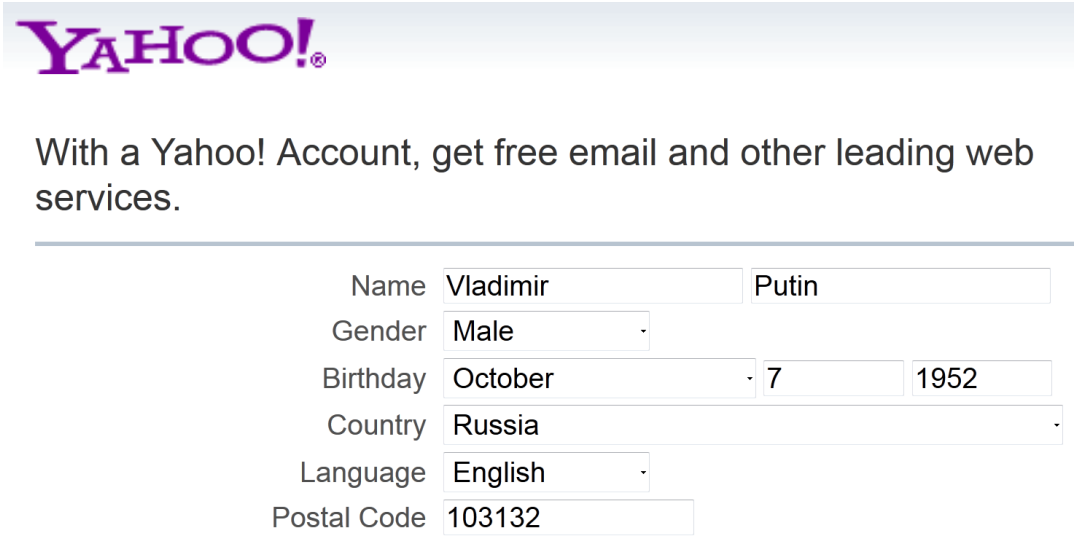
  "landscapers in Lilburn, Ga", "* Arnold", …

- Other example queries

  "dog that urinates on everything", "60 single men", "numb fingers", …

What Can We Infer from Queries?

- Goal
  - Age
  - Gender
  - ZIP code
- Data
  - Yahoo! query logs
  - Self-reported registration information
  - Users with >100 queries
  - Bag-of word representation



Bag of Words

Aug 7, 2012, 12h05: “restaurants in barcelona”

Aug 7, 2012, 12h08: “weather in catalunya”

~~Aug 7, 2012, 12h15: “restaurants in barcelona”~~

Aug 8, 2012, 10h15: “paella food poisoning”

Aug 8, 2012, 11h15: “doctors in barcelona”

Aug 9, 2012, 13h30: “people born in 1978”

Aug 9, 2012, 15h00: “music from the 90s”

…

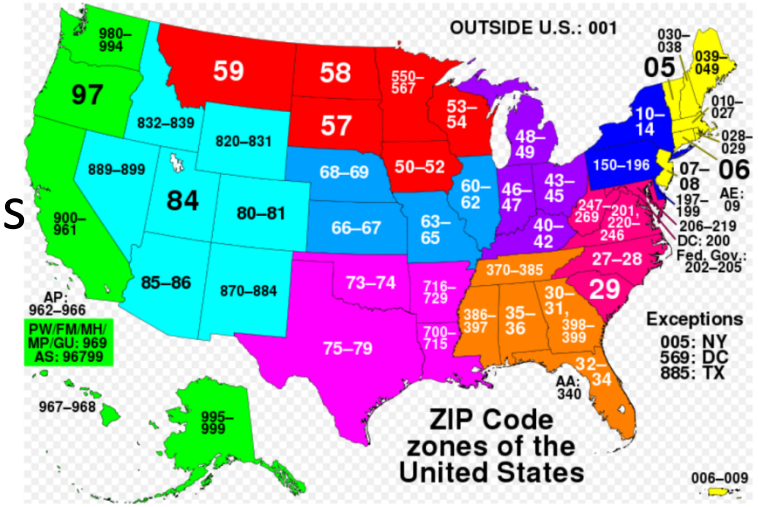
Bag of Words

restaurants: 1, in: 4, barcelona: 2, weather: 1, catalunya: 1, paella: 1, food: 1, poisoning: 1, doctors: 1, people: 1, born: 1, 1978:1, … (actually, used binary presence in paper)

User	restaurants	in	barcelona	..	privacy	90s
Id1	1	4	2	…	0	1
Id2	0	2	0	…	20	8
id3	2	2	0	…	0	0

Gender	birthyear	ZIP
Id1	1978	ES
Id2	1985	34561
id3	1961	90210

ZIP code prefixes correspond to regions



Experimental Results

- Trained an SVM (classification and regression)
  - http://svmlight.joachims.org/
  - 50-50 train-test split
- Gender
  - Baseline: always "male", 57% accuracy
  - SVM: 84% accuracy
  - bridal, makeup, hair vs. poker, football, ---
- Age
  - "… outperforming a baseline of always guessing the middle point"

δ	1	3	7	10
% users with $\epsilon < \delta$	14.7	33.4	63.9	79.0

  - lyrics, pregnancy, mall vs. lottery, retirement, repair

Experimental Results

- Location
  - Used Y! Placemaker to detect location names in queries
  - http://developer.yahoo.com/geo/placemaker/
  - Outputs guess with granularity (ZIP, county, city, …)
  - Aggregate guesses and construct list of locations

Zip	ZIP5	ZIP4	ZIP3
Accuracy top guess (%)	6.27	13.7	34.9
Accuracy top-3 guesses (%)	13.1	25.1	54.1

Related concept:

K-anonymity

Privacy preserving data publishing

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
Alice	22	14000	bronchitis
Andy	24	18000	flu
David	23	25000	gastritis
Gary	41	20000	flu
Helen	36	27000	gastritis
Jane	37	33000	dyspepsia
Ken	40	35000	flu
Linda	43	26000	gastritis
Paul	52	33000	dyspepsia
Steve	56	34000	gastritis

Microdata

Slide by: S. Sundaresan

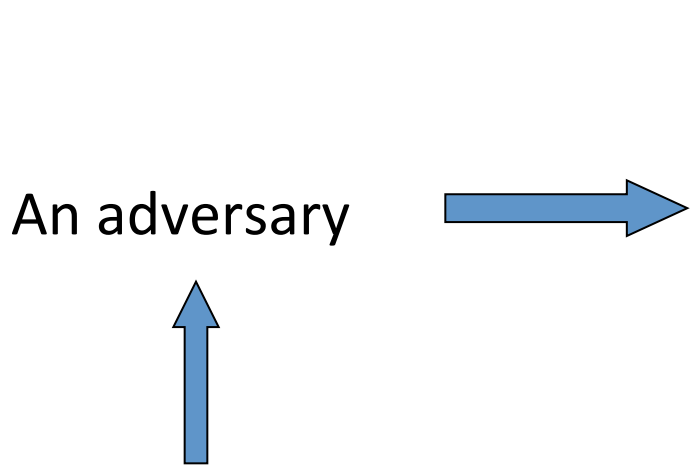
Privacy preserving data publishing

Name	Age	Zipcode	Disease
	21	12000	dyspepsia
	22	14000	bronchitis
	24	18000	flu
	23	25000	gastritis
	41	20000	flu
	36	27000	gastritis
	37	33000	dyspepsia
	40	35000	flu
	43	26000	gastritis
	52	33000	dyspepsia
	56	34000	gastritis

Microdata

Slide by: S. Sundaresan

Inference attack



Name	Age	Zipcode
Bob	21	12000

Published table

Age	Zipcode	Disease
21	12000	dyspepsia
22	14000	bronchitis
24	18000	flu
23	25000	gastritis
41	20000	flu
36	27000	gastritis
37	33000	dyspepsia
40	35000	flu
43	26000	gastritis
52	33000	dyspepsia
56	34000	gastritis

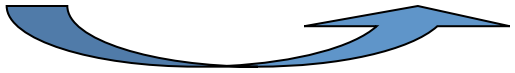
Quasi-identifier (QI) attributes

Slide by: S. Sundaresan

Generalization

- Transform the QI values into less specific forms

Age	Zipcode	Disease
21	12000	dyspepsia
22	14000	bronchitis
24	18000	flu
23	25000	gastritis
41	20000	flu
36	27000	gastritis
37	33000	dyspepsia
40	35000	flu
43	26000	gastritis
52	33000	dyspepsia
56	34000	gastritis



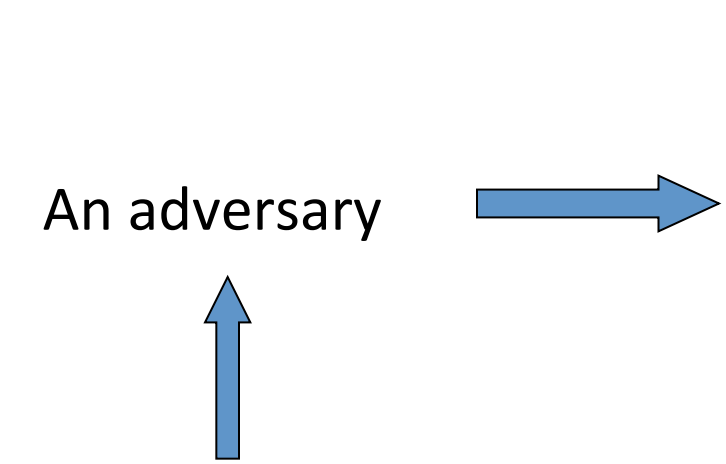
generalize

Age	Zipcode	Disease
[21, 22]	[12k, 14k]	dyspepsia
[21, 22]	[12k, 14k]	bronchitis
[23, 24]	[18k, 25k]	flu
[23, 24]	[18k, 25k]	gastritis
[36, 41]	[20k, 27k]	flu
[36, 41]	[20k, 27k]	gastritis
[37, 43]	[26k, 35k]	dyspepsia
[37, 43]	[26k, 35k]	flu
[37, 43]	[26k, 35k]	gastritis
[52, 56]	[33k, 34k]	dyspepsia
[52, 56]	[33k, 34k]	gastritis

Slide by: S. Sundaresan

Generalization

- Transform each QI value into a less specific form



Name	Age	Zipcode
Bob	21	12000

A generalized table

Age	Zipcode	Disease
[21, 22]	[12k, 14k]	dyspepsia
[21, 22]	[12k, 14k]	bronchitis
[23, 24]	[18k, 25k]	flu
[23, 24]	[18k, 25k]	gastritis
[36, 41]	[20k, 27k]	flu
[36, 41]	[20k, 27k]	gastritis
[37, 43]	[26k, 35k]	dyspepsia
[37, 43]	[26k, 35k]	flu
[37, 43]	[26k, 35k]	gastritis
[52, 56]	[33k, 34k]	dyspepsia
[52, 56]	[33k, 34k]	gastritis

Slide by: S. Sundaresan

K-Anonymity

Latanya Sweeny (1998)

- What is K-Anonymity?
  - If the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release.
  - Example:

    If you try to identify a man from a release, but the only information you have is his birth date and gender. There are k people meet the requirement. This is k-Anonymity.

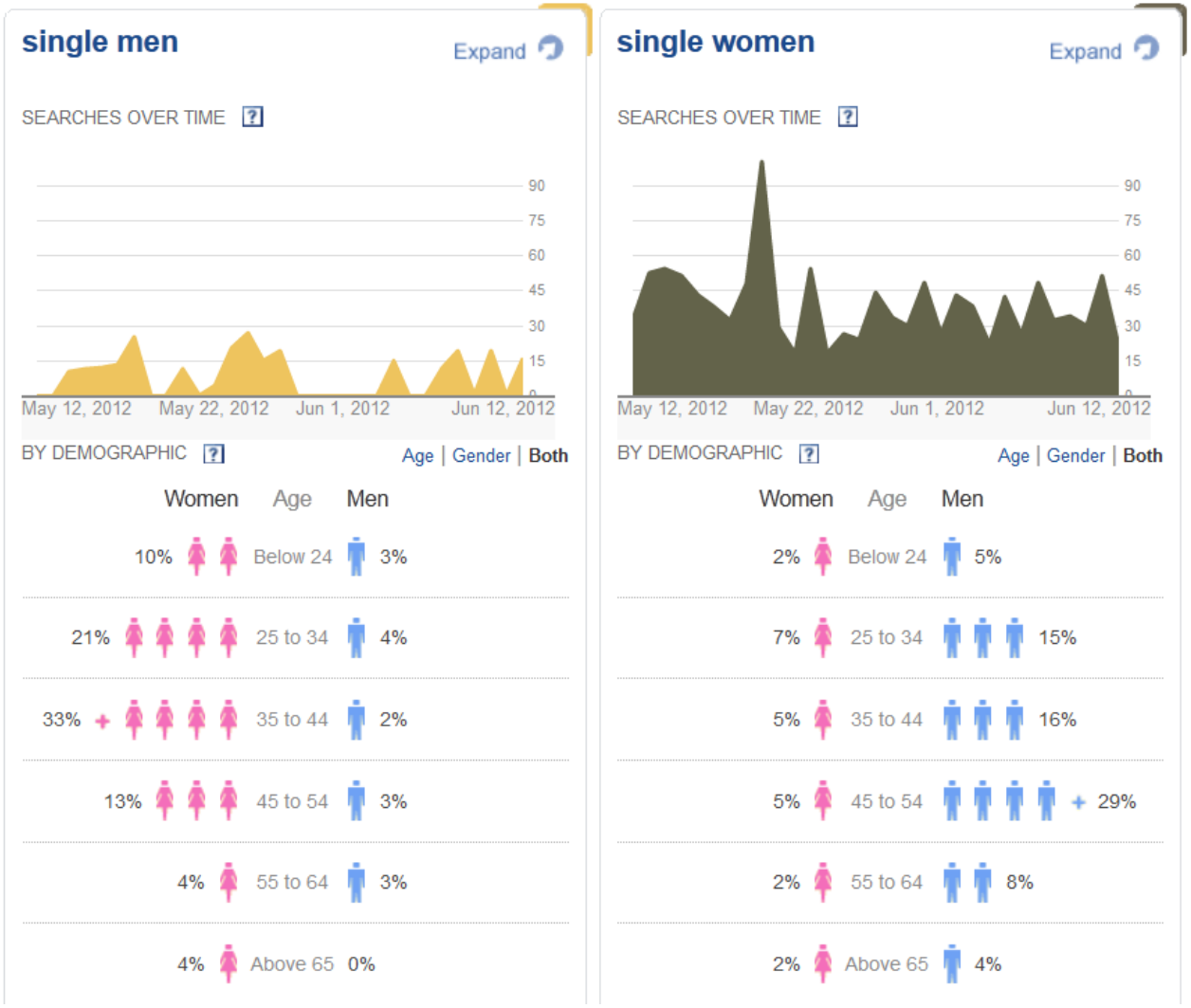
Slide by: S. Sundaresan

Web Search and Demographics

Ingmar Weber, Carlos Castillo, Alejandro Jaimes

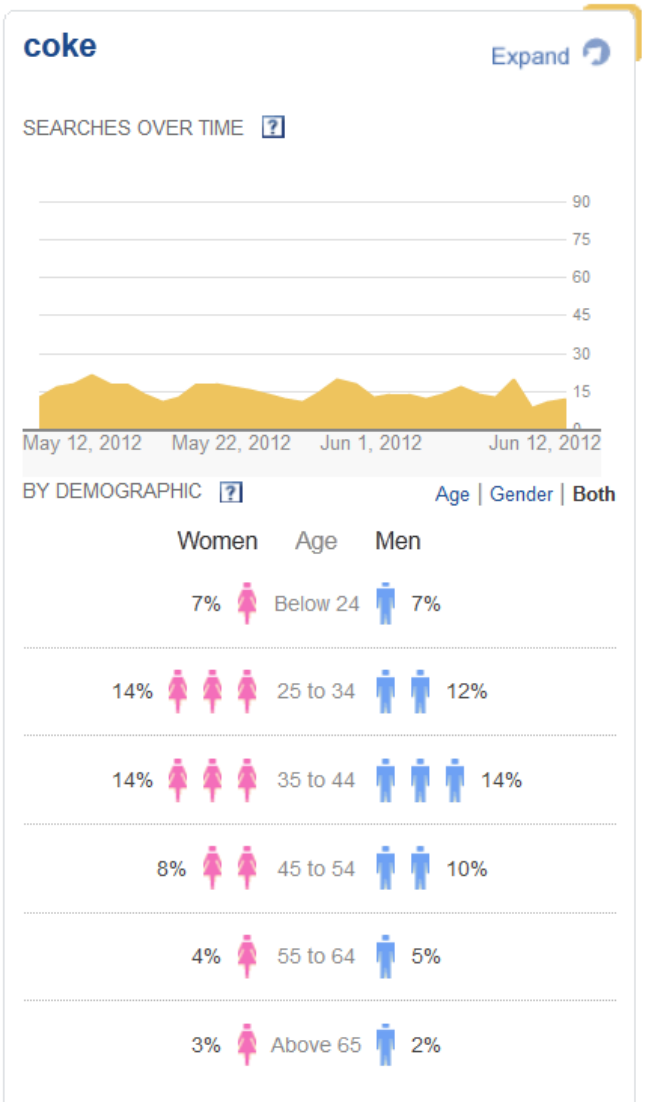
SIGIR'10, CIKM'10, WSDM'11

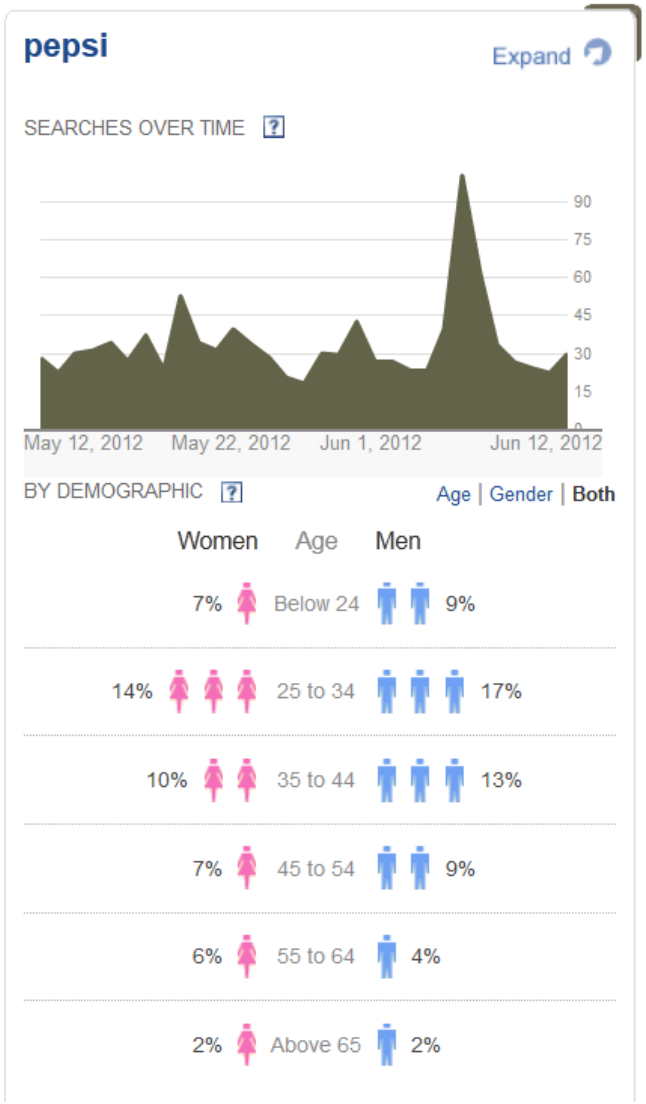
Yahoo! Clues



Women are more popular than men (online)
Single men are sought after by women and vice versa
Women start looking considerably earlier

Yahoo! Clues





Pepsi has slightly younger people searching "The Choice of a New Generation"

How the Data was Obtained

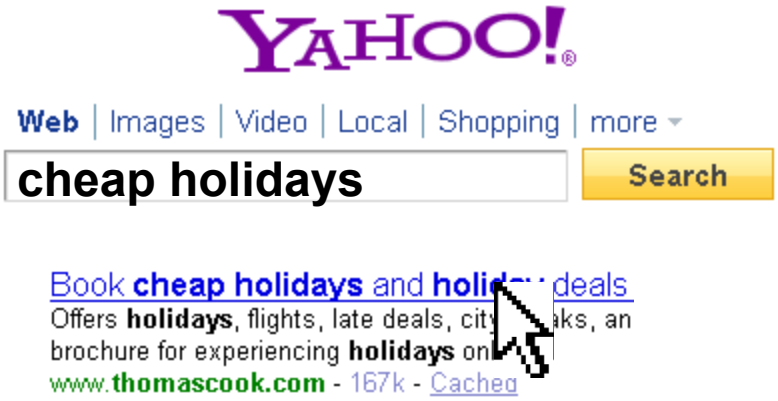
Gender: Male
Birth year: 1978
ZIP code: 95054

US Census Data
factfinder2.census.gov

Expected income: $ 31k

Expected education: 45% BA

Race distribution: 27% w, 57% A



Q

D

Label (Q,D) with $31k, 45%BA, ...
Income_5, education_5, white_1, …

quintiles

Yahoo! Users vs. US population

Feature	Y! aver.	10%	90%	US aver.
P-c income $k	22.9	14.3	33.5	21.6

Pretty good match, but

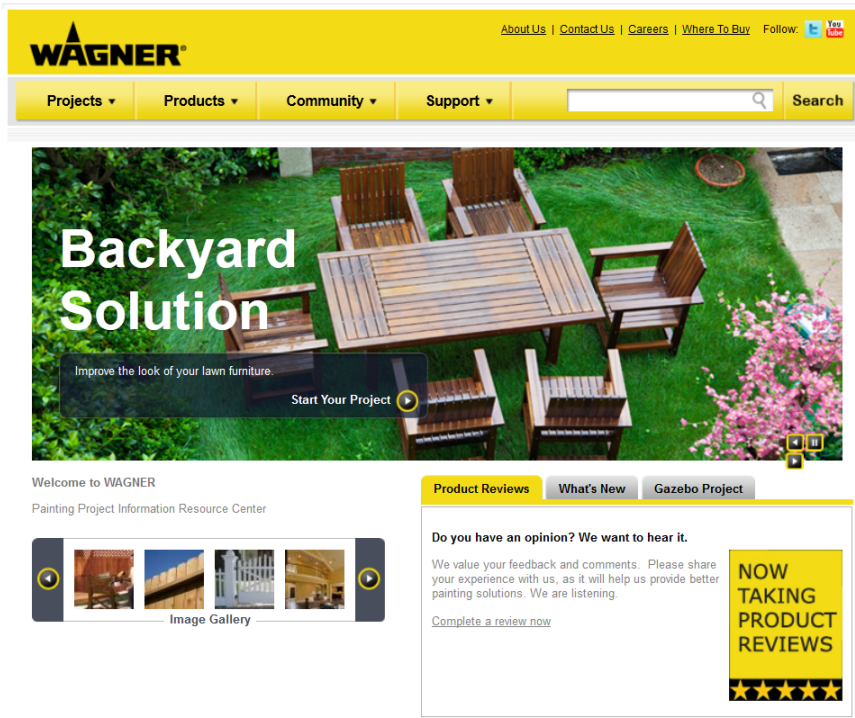
- Slightly higher income
- Slightly more educated
- Slightly more white
- Slightly older

Experiments

- Want to rank a *target* for a certain *input*
  - P("wiki.org/Richard_Wagner" | "wagner")

target = URL U

input = query Q

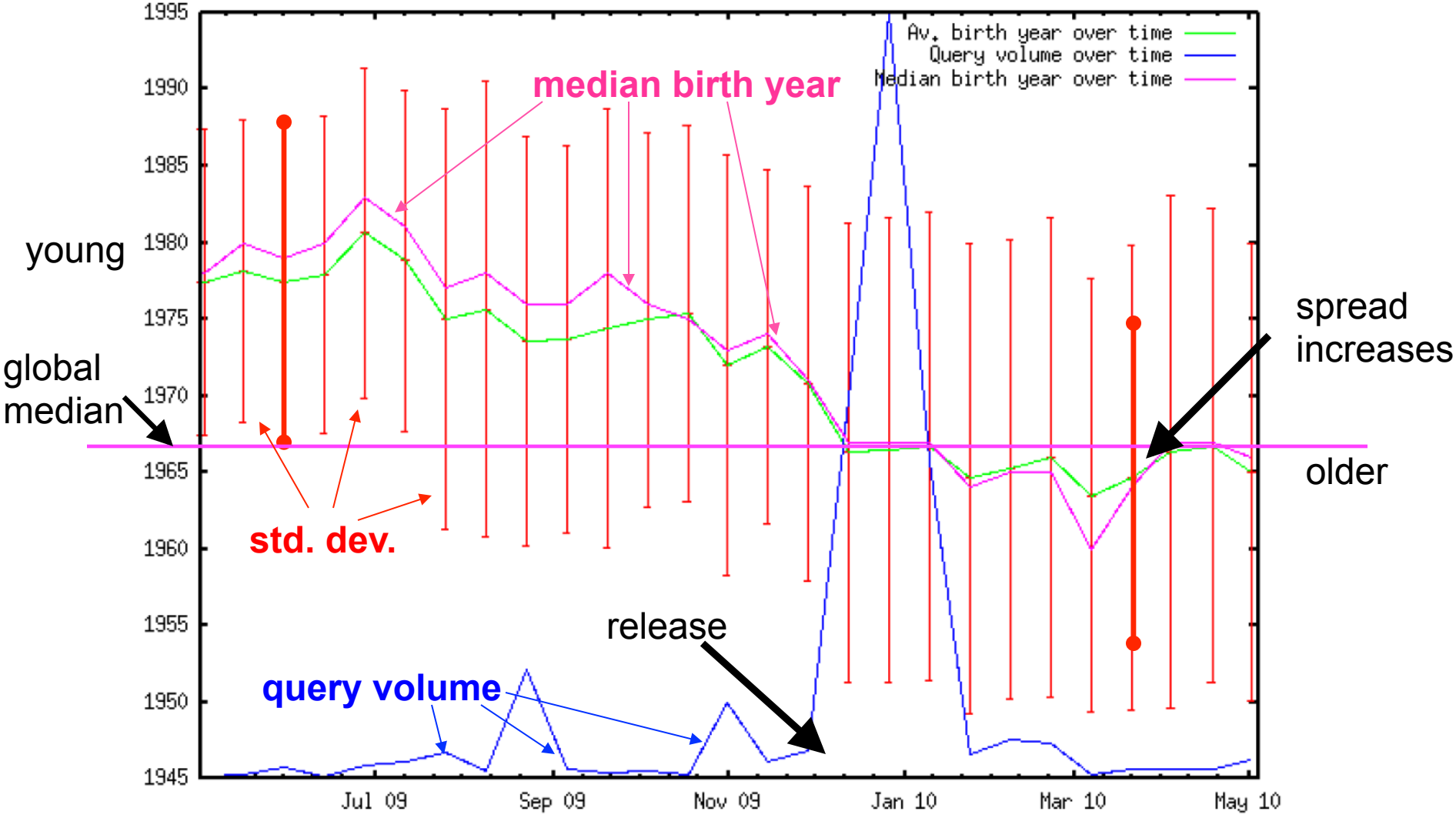


Web Search

- Click behavior can depend on demographics
  - R. Wagner (female) vs. Wagner Spray Tech (male)
  - ESL Federal Credit Union vs. English as a Sec. L.

	# pairs	P@1 w/o F	P@1 with F
all (>500 occs)	207 Mio	.703	.713
H(D|Q) >= 1.0	123 Mio	.557	.574
H(D|Q) >= 2.0	60.6 Mio	.381	.408

Information Flows: “avatar movie”

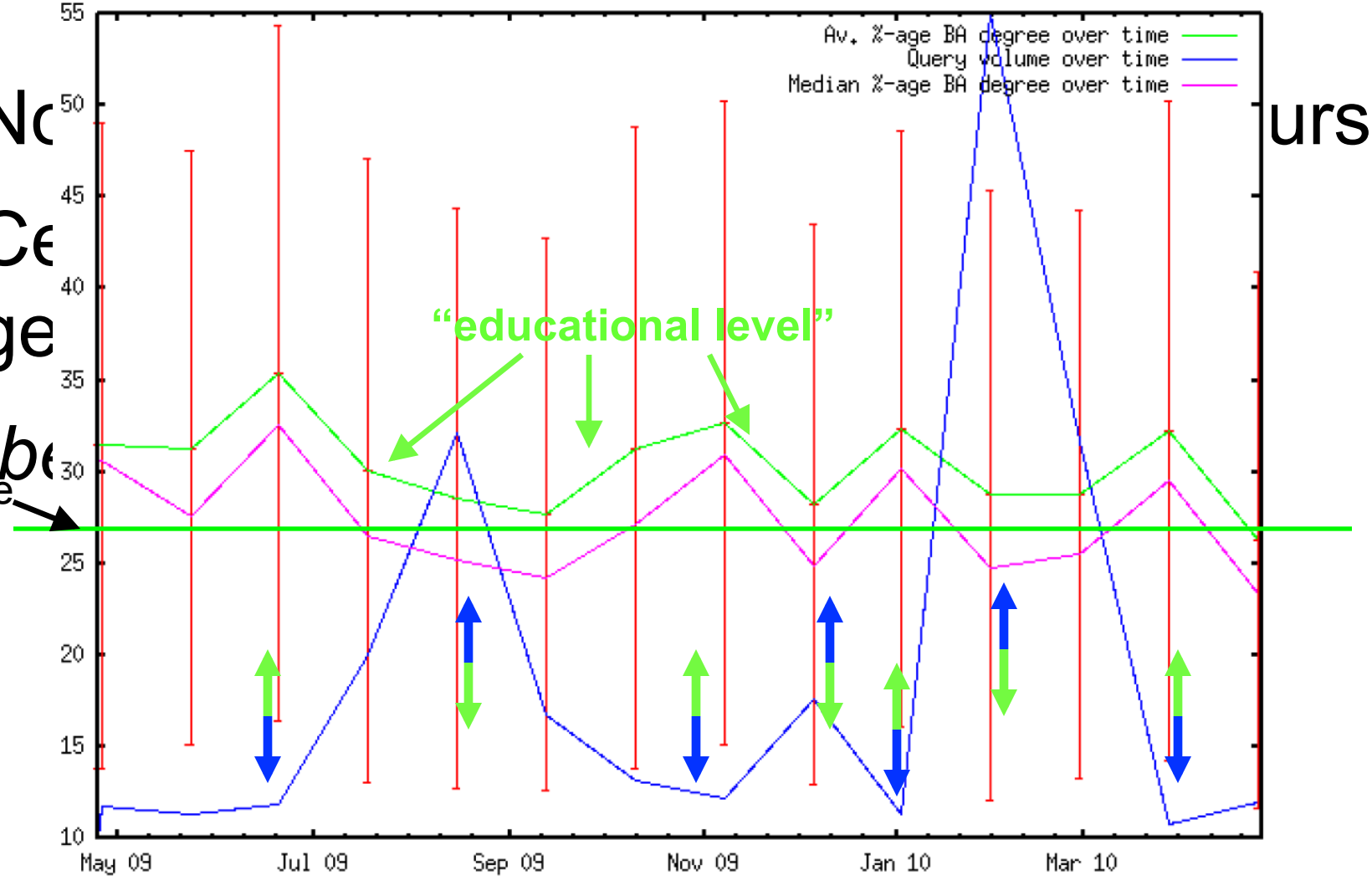


Applications

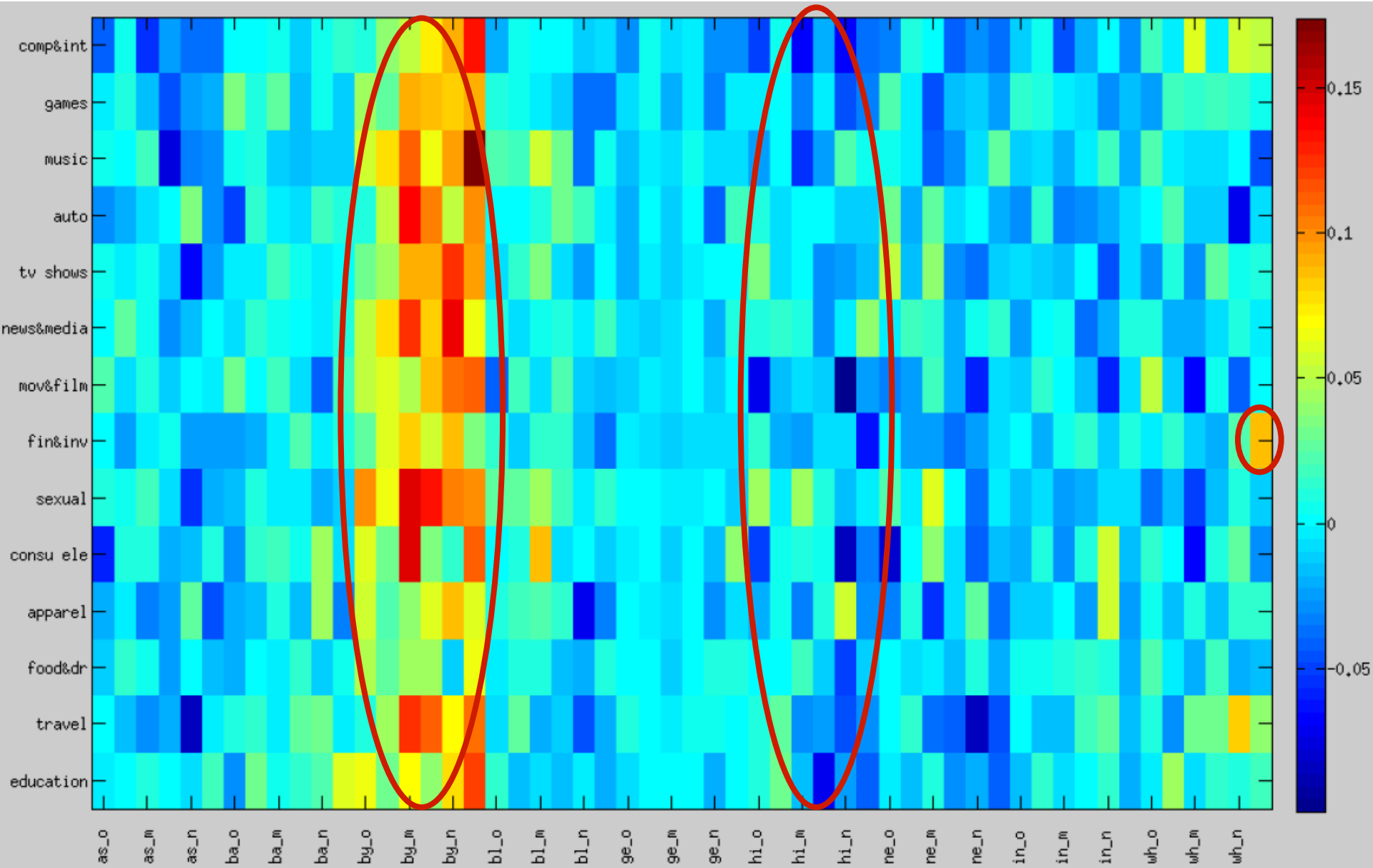
- Targeted advertising
  - Let advertisers plan whole campaign
  - First target X, then Y, then Z
- Information relevancy prediction
  - "You probably know this by now."
- Identify information hermits
  - Who's last to search for vaccination?

From Niche to Masses and Back to Niche

- Nc urst
- Ce
  ge
  be



Flow Visualization (154k queries)

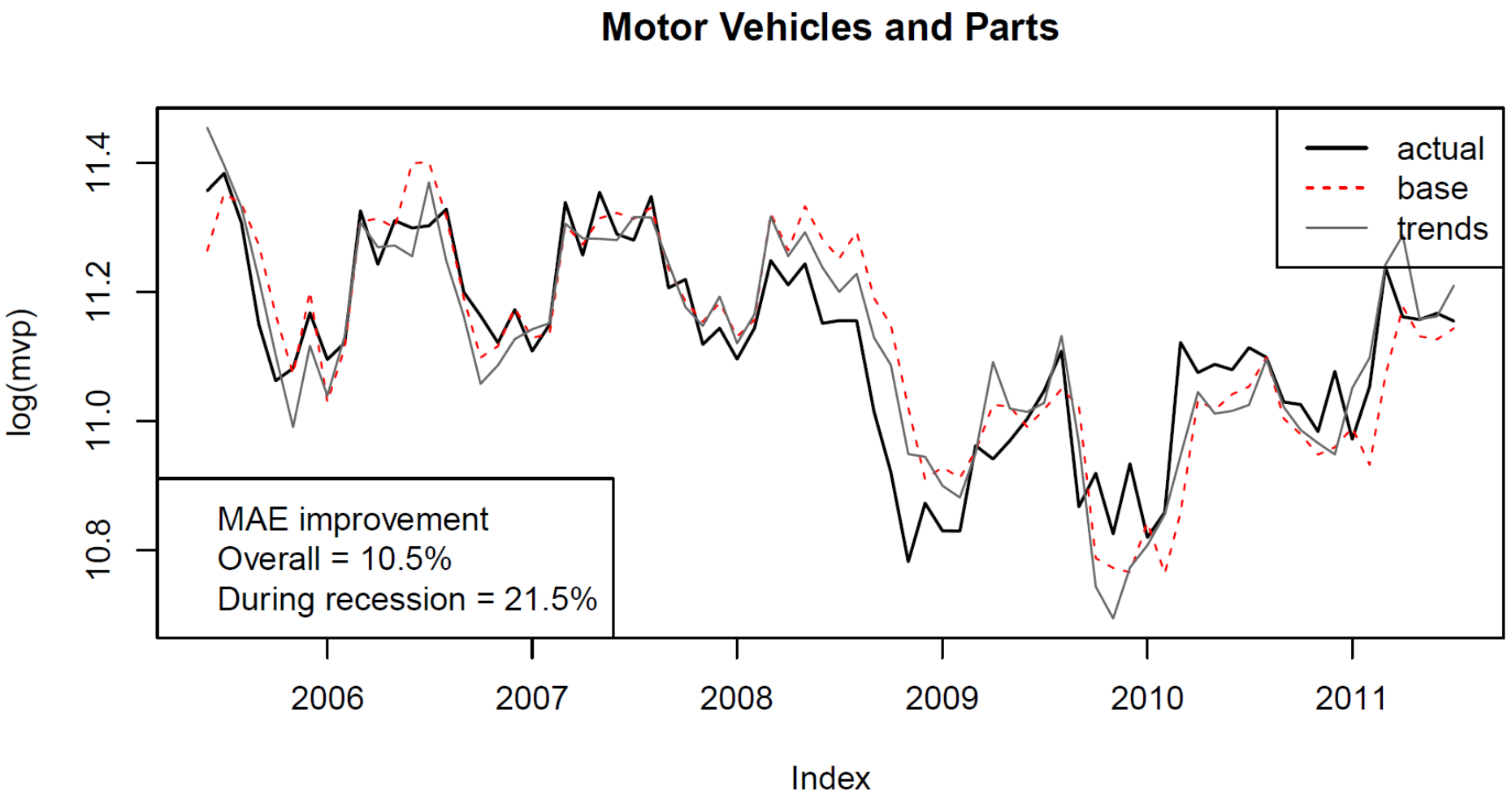


Predicting the Present with Google Trends

Hal Varian and Hyunyoung Choi

Google Research Blog, April 2009

Index of Car Retailers



Google Trends

Google Trends

car sales

Search Trends

Tip: Use commas to compare multiple search terms.

Searches Websites

United States

- Scale is based on the average traffic of **car sales** from United States in all years. Learn more
- An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

car sales 1.00

Google Trends

Google Trends

car sales, barack obama, sex

Search Trends

Tip: Use commas to compare multiple search terms.

Searches Websites

United States

- Scale is based on the average traffic of **car sales** from United States in all years. Learn more
- An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

car sales 1.00 **barack obama** 2.00 **sex** 74.0

Time Series Analysis

- Given a discrete series of values over time
  - Maximum daily temperature for the last week
  - Value of the Dow Jones at closing time
  - Monthly unemployment rates
  - $X_{t-1}, X_{t-2}, X_{t-3}, \ldots$
- We want to predict tomorrow's value
  - Linear regression = autoregression

$$X_t = c + \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t$$

Autoregression

Examples where X_t = tomorrow's temperature

"Tomorrow's temperature is the year's average."

$X_t = c$

"Tomorrow's temperature is the average of today's temperature and the year's average."

$X_t = (c + X_{t-1})/2 = 0.5*c + 0.5 * X_{t-1}$

"The trend since yesterday will continue."

$X_t = X_{t-2} + (X_{t-1} - X_{t-2}) + (X_{t-1} - X_{t-2})$

$= 2*X_{t-1} - X_{t-2}$

Predicting Automobile Sales

- The "Motor Vehicles and Parts Dealers" time series from the census bureau

  http://www.census.gov/retail/marts/www/timeseries.html
- Survey based index, released two weeks after the end of each month
- Raw form used (not "seasonally adjusted")

Applying Autoregression

- Baseline:

  $X_t = Á_1 X_{t-1} + Á_{12} X_{t-12} + ²_t$

- Find best values for Á minimizing the errors
  - Standard linear regression
- Incorporating query volume:

  $X_t = Á_1 X_{t-1} + Á'_1 Q_{t-1} + Á_{12} X_{t-12} + ²_t$

- Trivially a better fit for training data
  - Richer model with more parameters to fit
  - Also for test data?

Applying Autoregression

- Whole test period:
  - AR-1 model: MAE 6.34%
  - AR-1+GT: MAE 5.66%

- Period of recession (Dec. 2007 – Jun. 2009)
  - AR-1 model: MAE 8.86%
  - AR-1+GT: MAE 6.96%

Predicting consumer behavior with Web search

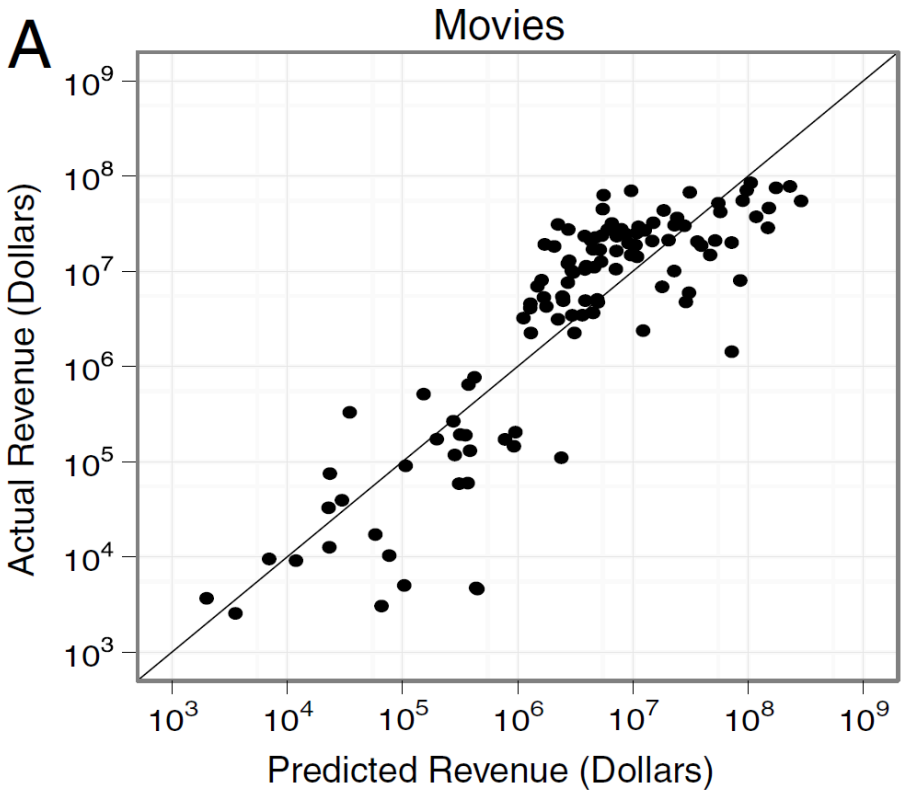
Sharad Goel, Jake Hofman, Sébastien Lahaie, David Pennock and Duncan Watts

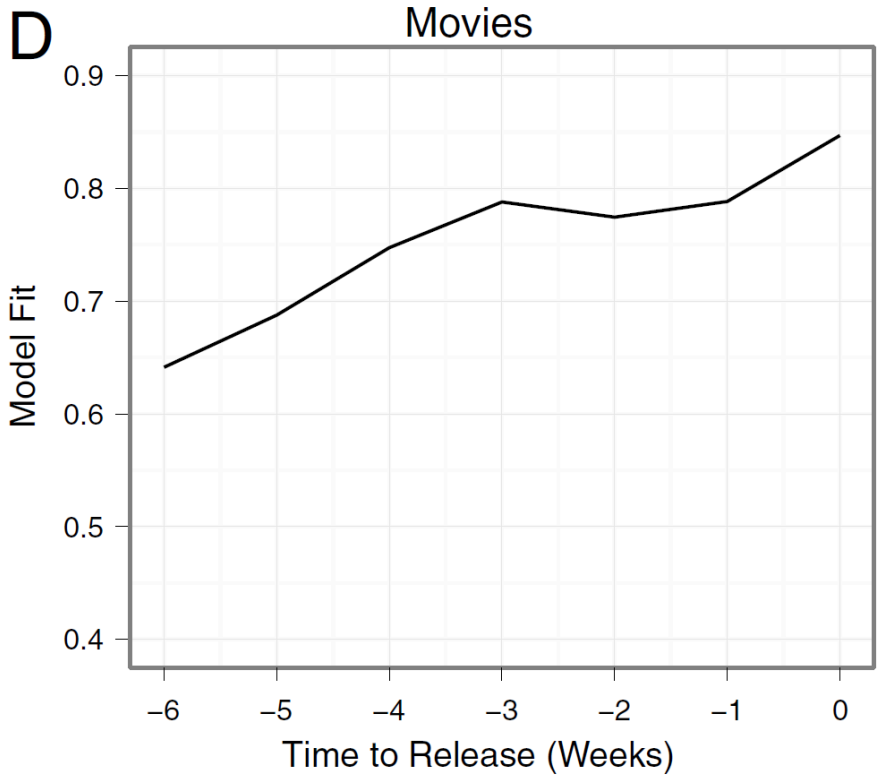
PNAS 2010

Predicting Movie Sales

- Opening week movies box office sales

$$\log(\text{revenue}) = \bar{\ }_0 + \bar{\ }_1 \log(\text{search}) + {}^2$$





Having a Strong Baseline

$$\log(\text{rev.}) = \bar{}_0 + \bar{}_1 \log(\text{budget}) + \bar{}_2 \log(\text{screens}) + \bar{}_3 \log(\text{HSX}) + {}^2$$

Ice Age: Continental Drift

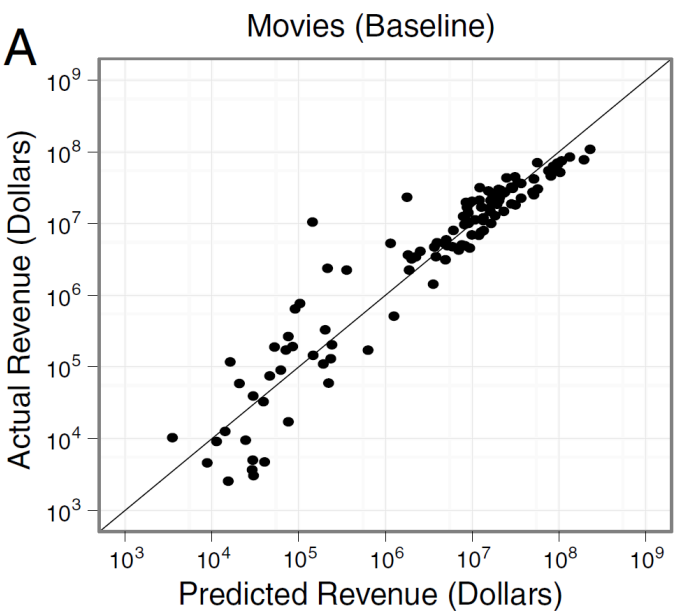
Symbol:	**ICEA4**	Phase:	**Wrap**
Status:	**Active**	Release Date:	**Jul 13, 2012**
IPO Date:	**May 24, 2010**	Release Pattern:	**wide**
Genre:	**Animated**	Gross:	**n/a**
MPAA Rating:	**PG**	Theaters:	**3800**

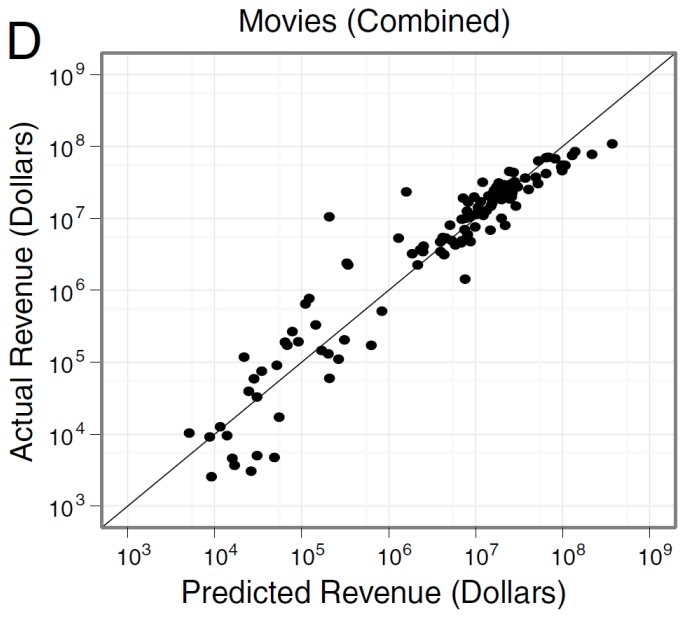
Ice Age: Continental Drift (ICEA4)

H$169.78 CURRENT VALUE

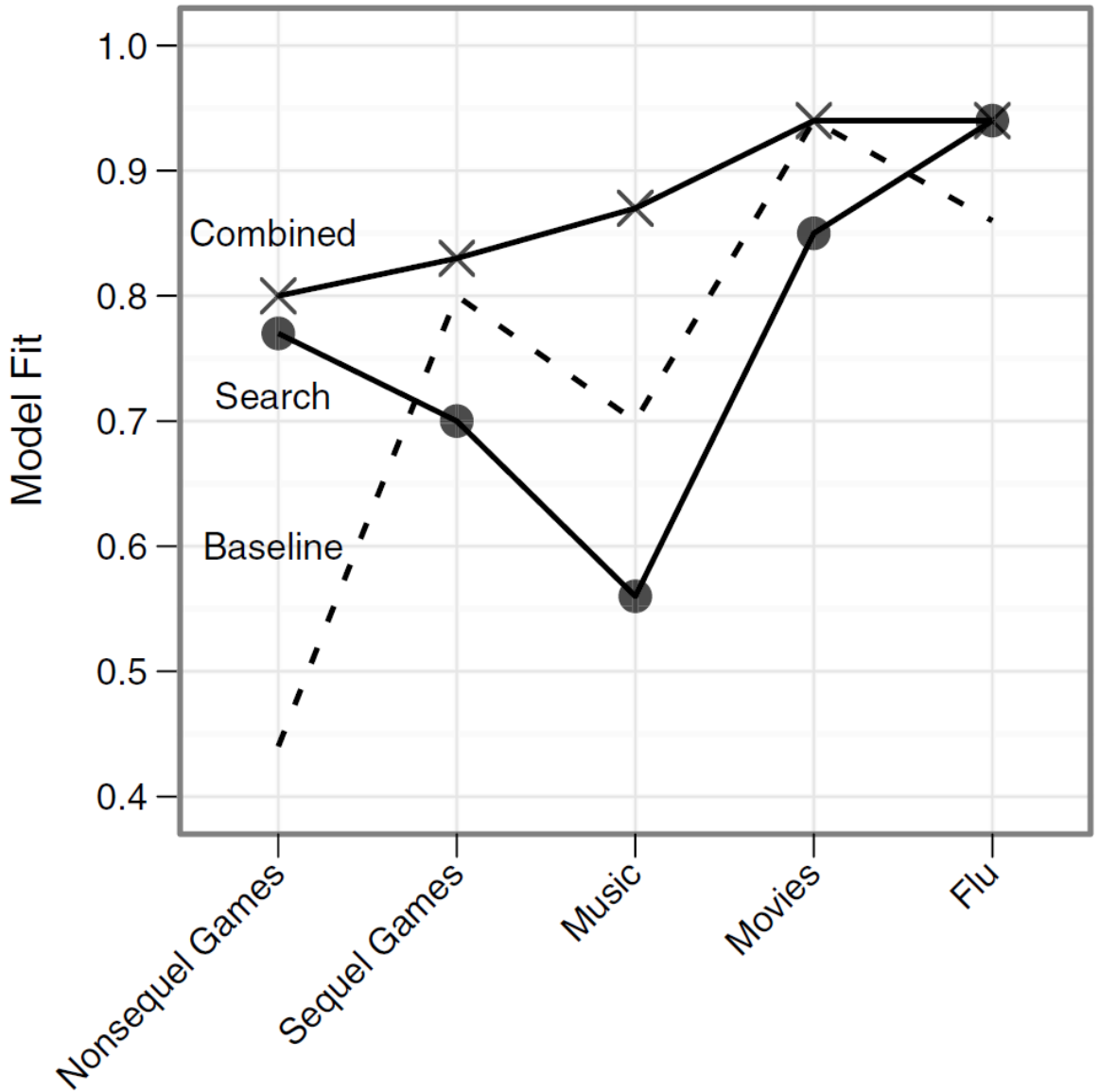
H$1.45 (0.86%) CHANGE TODAY

Shares Held Long on HSX:	163,782,152
Shares Held Short on HSX:	16,350,443
Trading Volume on HSX (Today):	1,849,778





Google Flu Trends - revisited



Internet search behavior as an economic forecasting tool

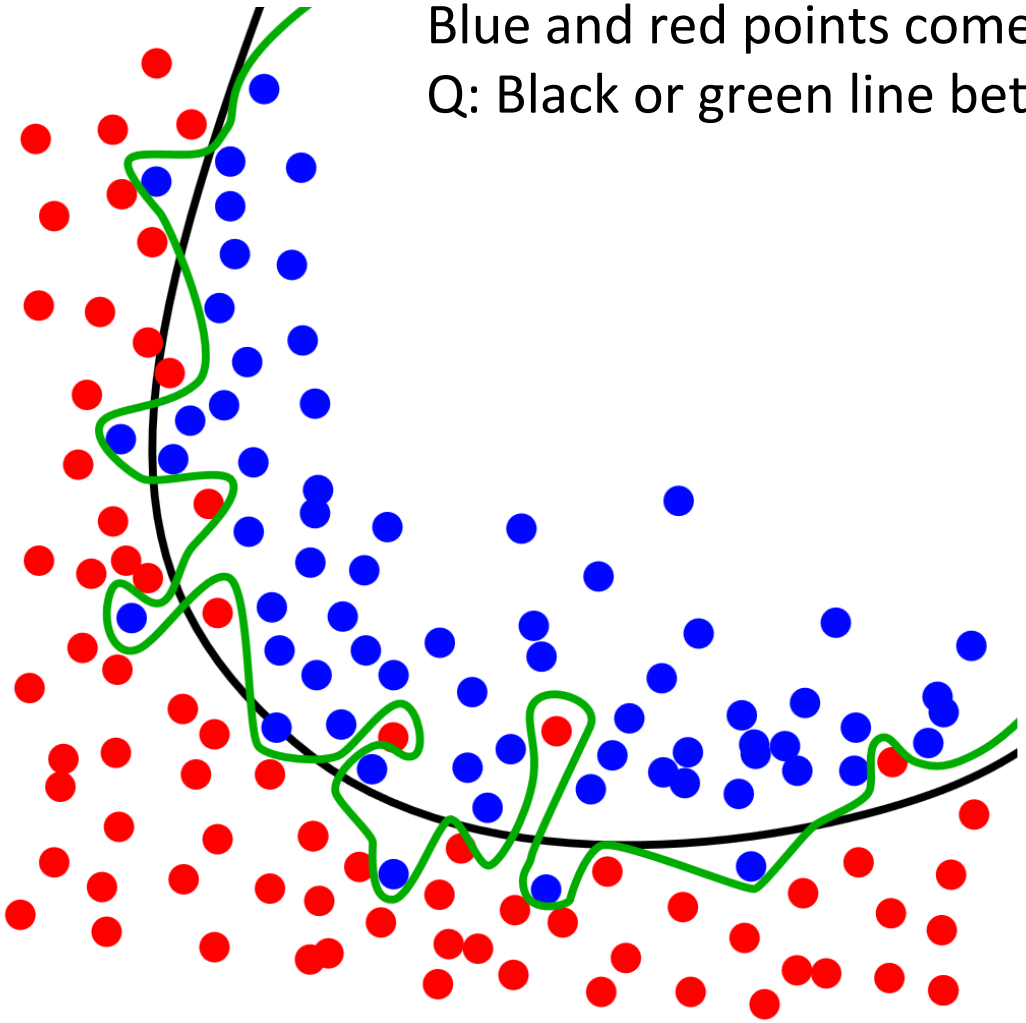
Giselle Guzman

JESM 2011

Predicting Inflation Rates

This paper proposes a measure of real-time inflation expectations based on metadata, i.e., data about data, constructed from internet search queries performed on the search engine Google. The forecasting performance of the Google Inflation Search Index (GISI) is assessed relative to 37 other indicators of inflation expectations – 36 survey measures and the TIPS spread. For decades, the academic literature has focused on three measures of inflation expectations: the Livingston Survey, Survey of Professional Forecasters, and the Michigan Survey. While useful in developing models of forecasting inflation, these low frequency measures appear anachronistic in the modern era of higher frequency and real-time data. I demonstrate that higher frequency measures tend to outperform lower frequency measures in tests of accuracy, predictive power, and rationality. Furthermore, Granger Causality tests indicate that the GISI metadata indicator anticipates the inflation rate by 12 months, and out-of-sample forecasts show that the GISI has the lowest forecast error of all the inflation expectations indicators tested.

What is Meant by Accuracy?



Blue and red points come from a random distribution
Q: Black or green line better at telling them apart?

Green line (zig-zag)
Better “accuracy”
Better “training error”

Black line (smooth)
Better “predictive power”
Better “test error”

Overfitting

What is a "Rational" Model?

- Unbiased: Should be correct "in expectation"

$$\pi_t = \alpha + \beta \pi_t^e + e_t$$

  - Test the joint hypothesis $\alpha = 0$ and $\beta = 1$

- Efficient: not ignoring available information
  - Hypothesis testing:

    Do available "information sets" correlate with error

Granger Causality

What is causality? Can you define it?

"The statement about causality has just two components:

1. The cause occurs before the effect; and
2. The cause contains information about the effect that that is unique, and is in no other variable"

Clive Granger, accepting the 2003 Nobel prize in Economic

A "pragmatic" definition that can be implemented …

Granger Causality

Given a time series, find the longest "lag"

$X_t = c + Á_1 X_{t-1} + ²_t$ Then test $H_0: Á_1 = 0$

$X_t = c + Á_1 X_{t-1} + Á_2 X_{t-2} + ²_t$ Then test $H_0: Á_2 = 0$

…

$X_t = c + Á_1 X_{t-1} + Á_2 X_{t-2} + … + Á_m X_{t-m} + ²_t$

Then add the (supposed) predictive Y_t

$X_t = c + Á_1 X_{t-1} + … + Á_m X_{t-m} + Ã_p Y_{t-p} + … + Ã_q Y_{t-q} + ²_t$

Say "Y Granger-causes X" when the $Ã_p$ are non-zero in a hypothesis test

What to Do With All of This?

- Optional homework for after the course:
- Get data for a time series of interest
  - Stock trading prices
  - Weather reports
  - Movie sales
  - …
- Use Google Correlate and Google Trends to get correlated query volume
- Try out the various concepts introduced
- Write a paper about it ☺
- More time series examples later in course

Reminder:

Competition

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

Questions?

End of Day 2

ingmar@yahoo-inc.com