Google™

Google Tools for Data

Hal Varian
Univ of Oregon
Oct 2013

Google™

Google Trends

Google Correlate

Google Consumer Surveys

Searches for [hangover]

Google

Which day of the week are there the most searches for [hangover]?

1: Sunday

2: Monday

3: Tuesday

4: Wednesday

5: Thursday

6: Friday

7: Saturday

Search index for [hangover]

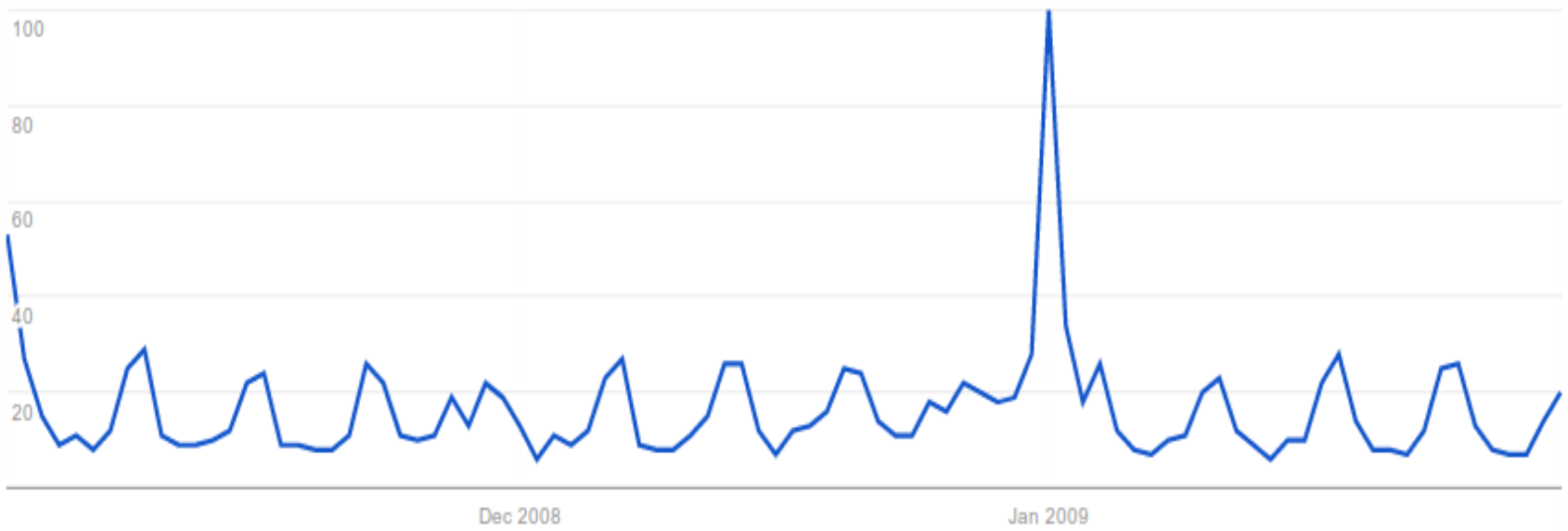
Google

Web Search Interest: **hangover**. United States, Nov 2008 - Jan 2009.

Interest over time

The number 100 represents the peak search volume

News headlines Forecast

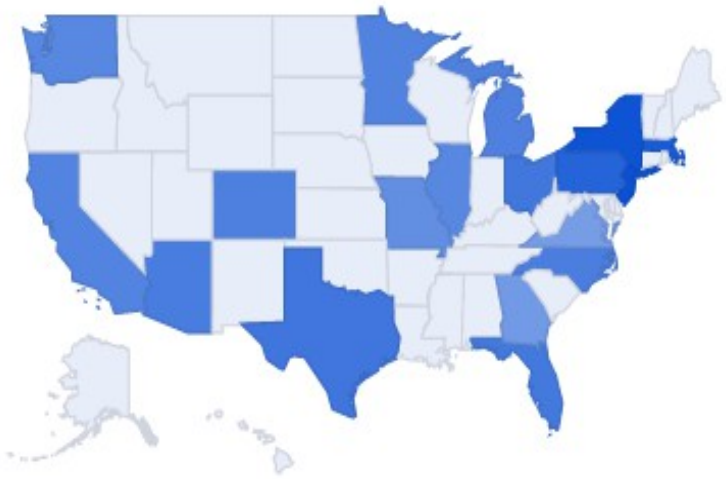


Hangover by geography

Google

Regional interest

Worldwide > United States



0 100

Subregion | Metro | City

View change over time

Embed

Related terms

Top | Rising

cure hangover	100
hangover cures	65
the hangover	50
cure a hangover	40
hangover remedies	35
hangover food	15
hangover symptoms	15
love hangover	15
cure for hangover	15
best hangover cure	15

Embed

Hangover-vodka time series

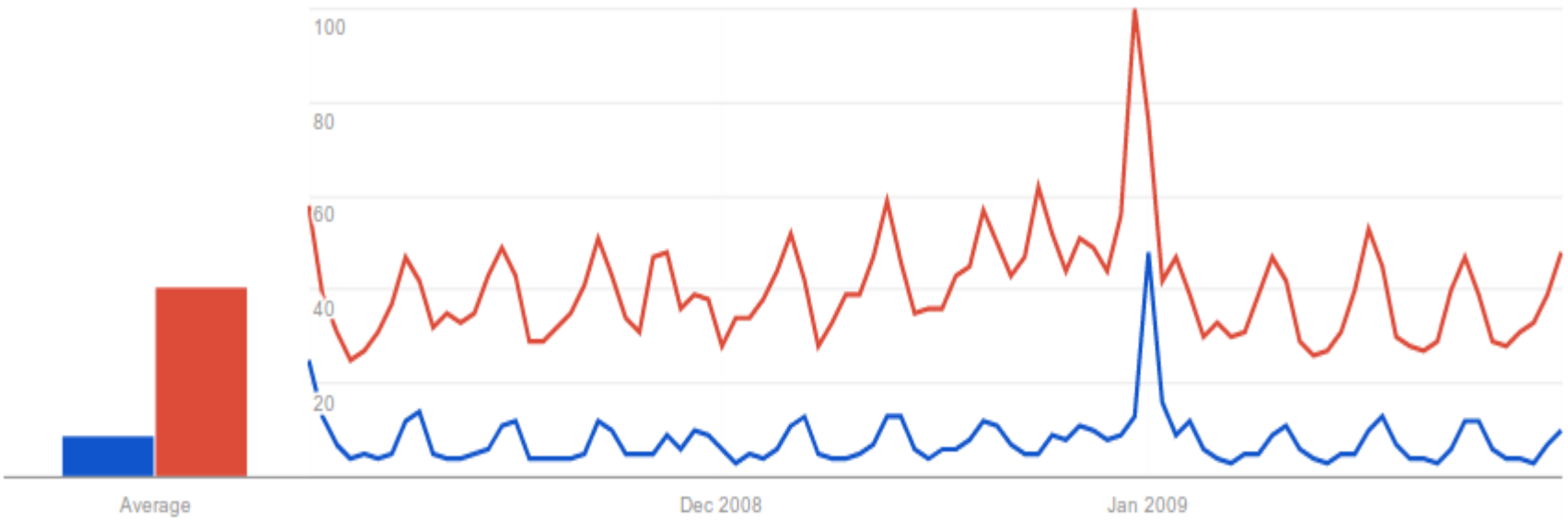
Google

Web Search Interest: **hangover, vodka**. United States, Nov 2008 - Jan 2009.

Interest over time

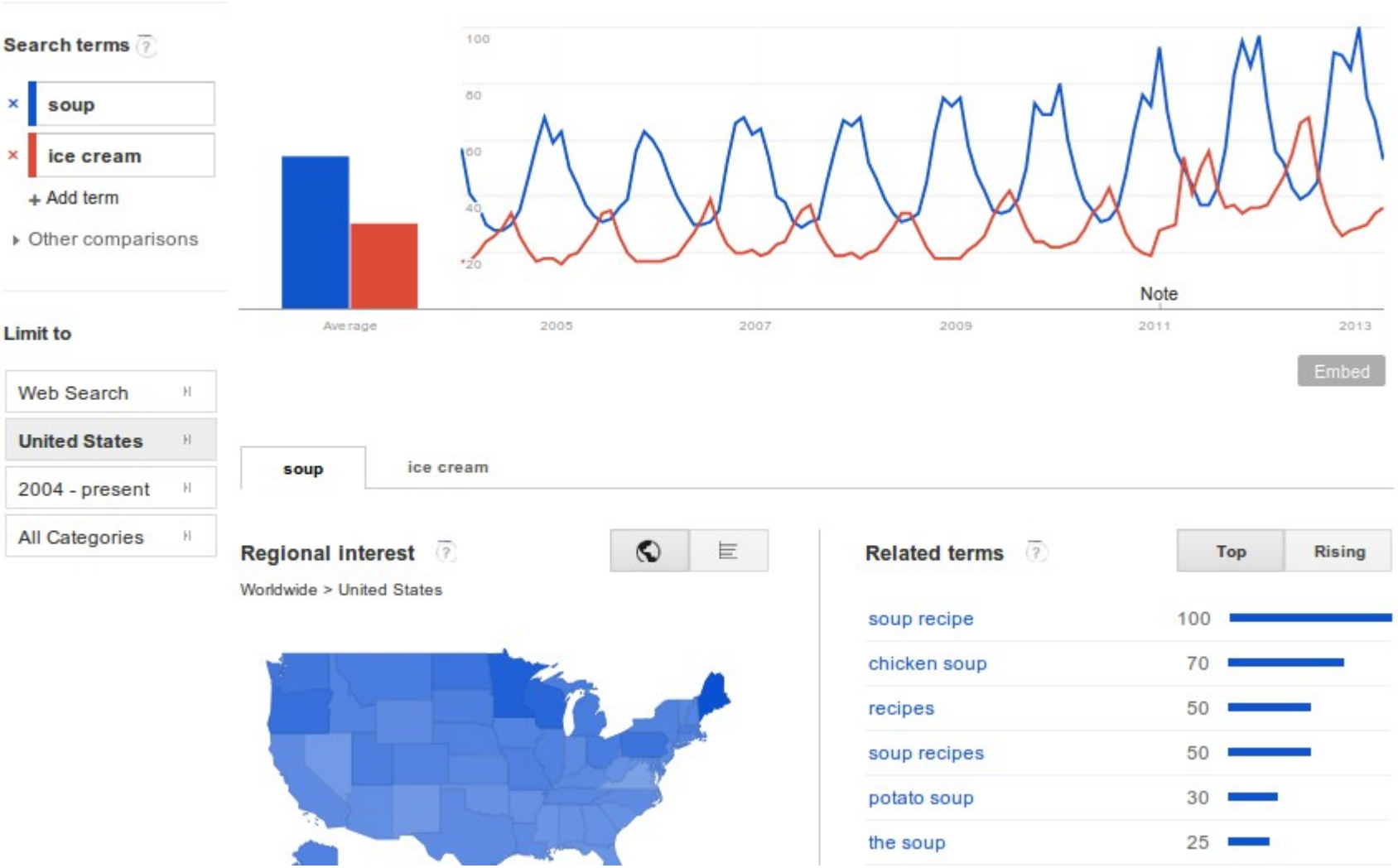
The number 100 represents the peak search volume

News headlines Forecast



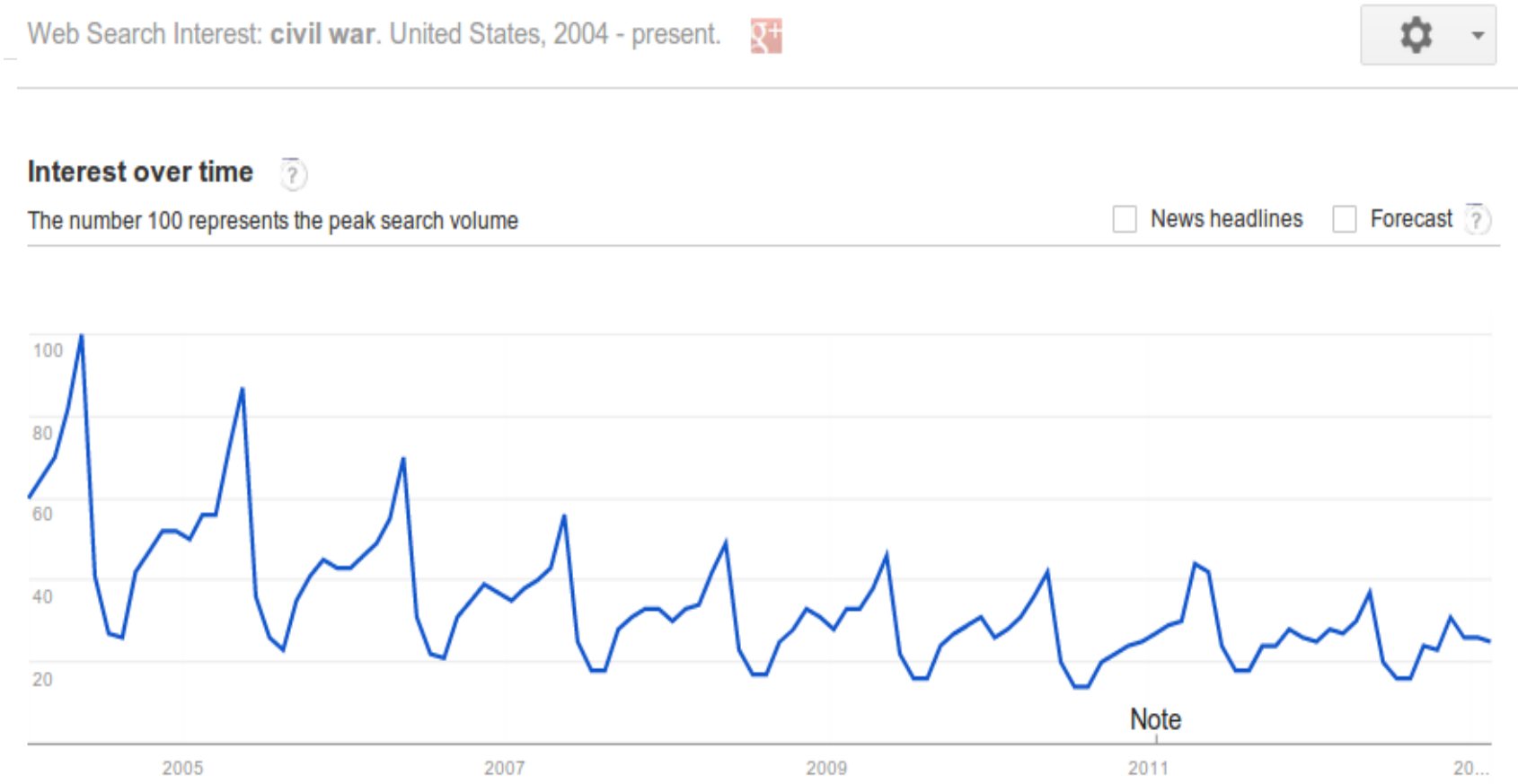
Soup and ice cream

Google



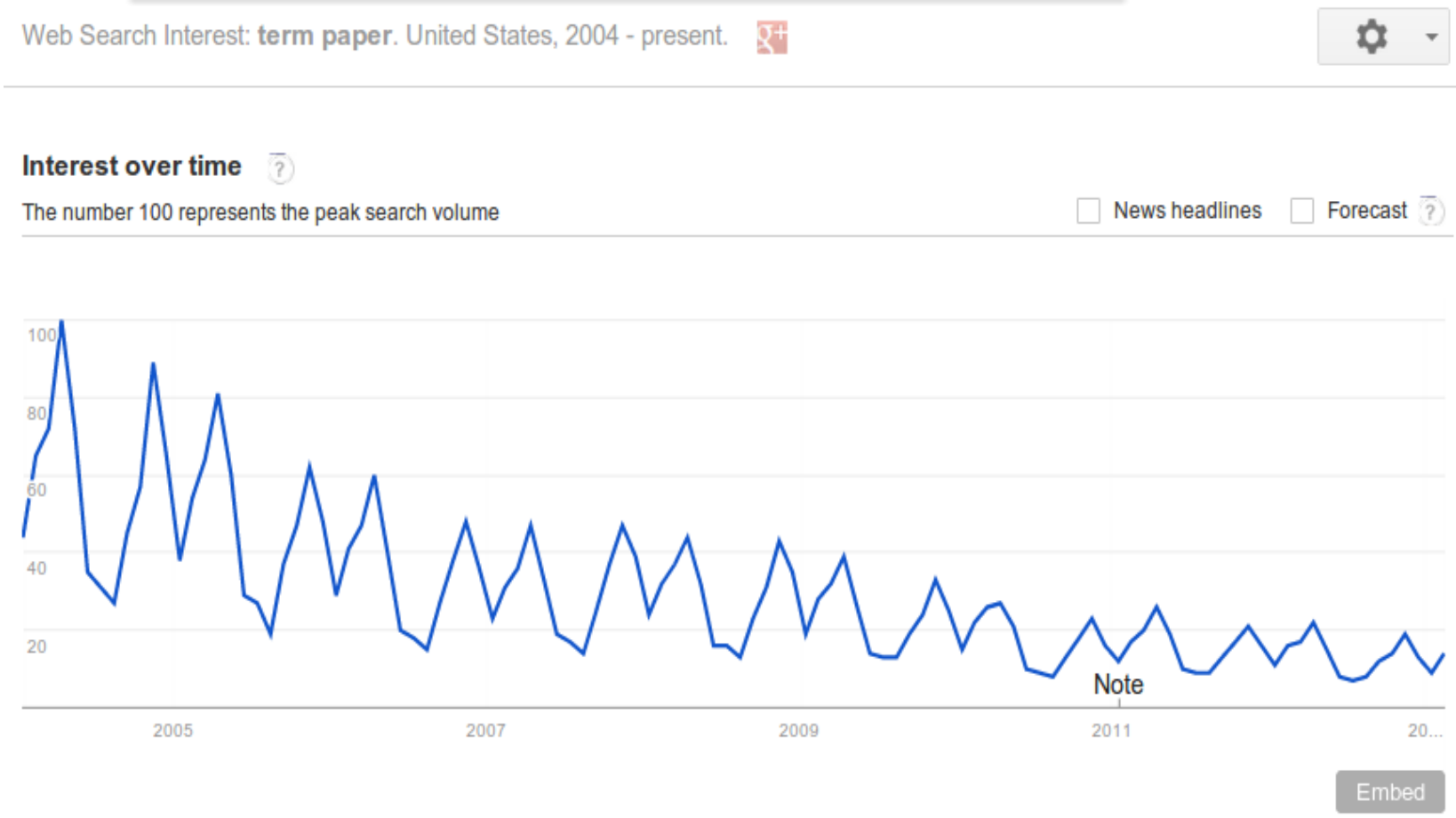
Searches for [civil war]

Google



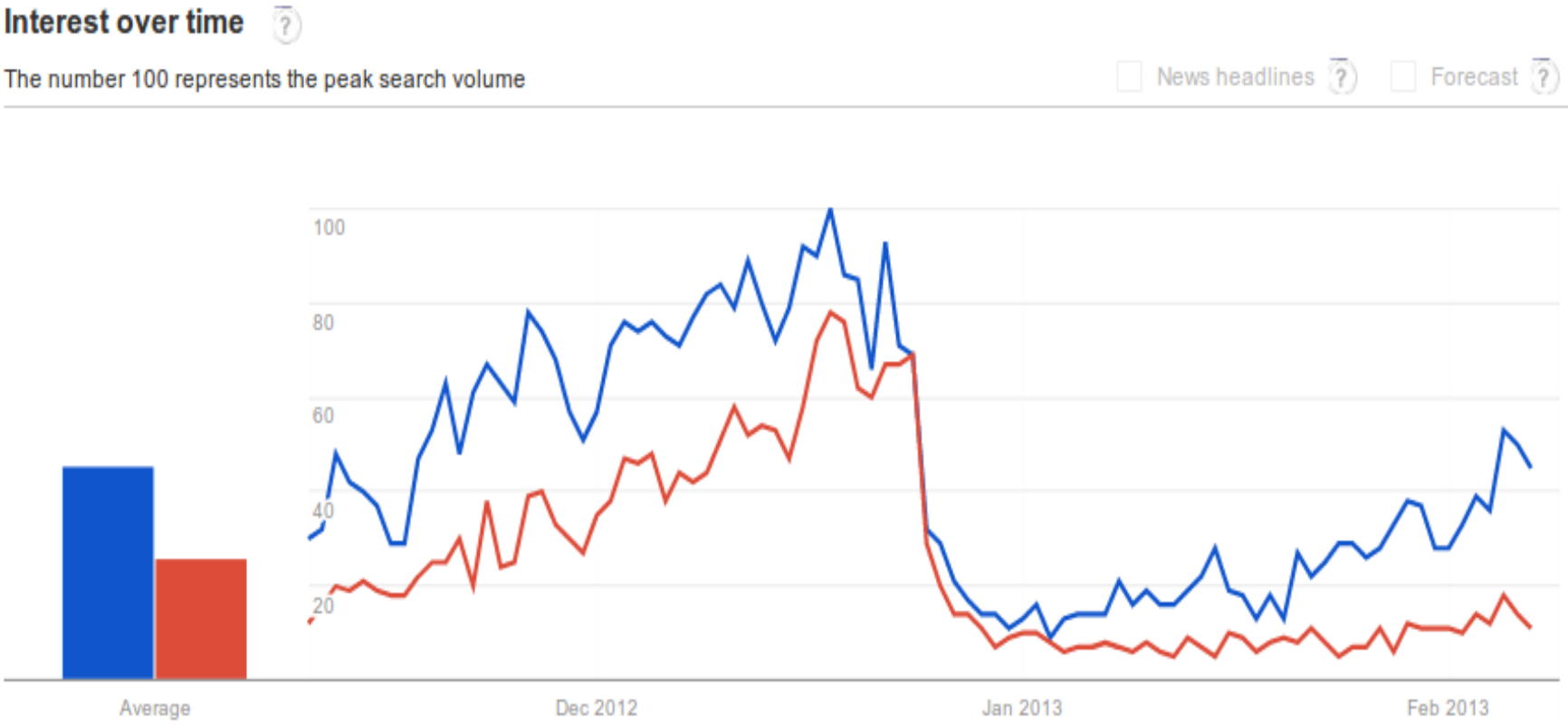
Searches for [term paper]

Google



Gift for boyfriend v Gift for girlfriend

Google



For boyfriend

For girlfriend

Gift for husband v Gift for wife

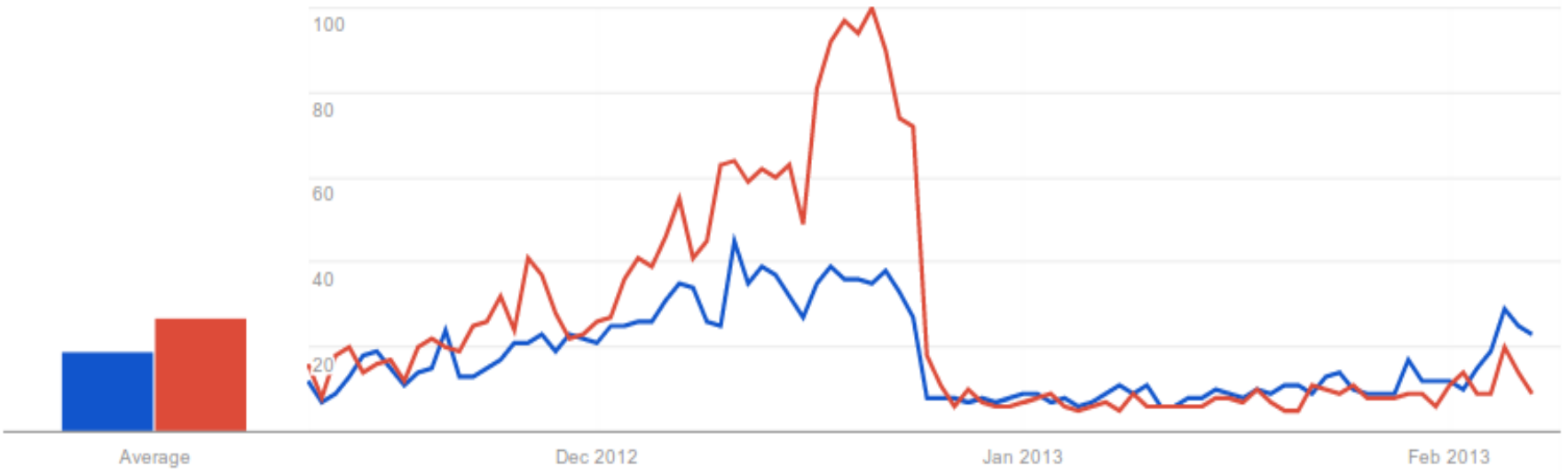
Google

Web Search Interest: **gift for husband, gift for wife**. United States, Past 90 days.

Interest over time

The number 100 represents the peak search volume

News headlines | Forecast



For husband

For wife

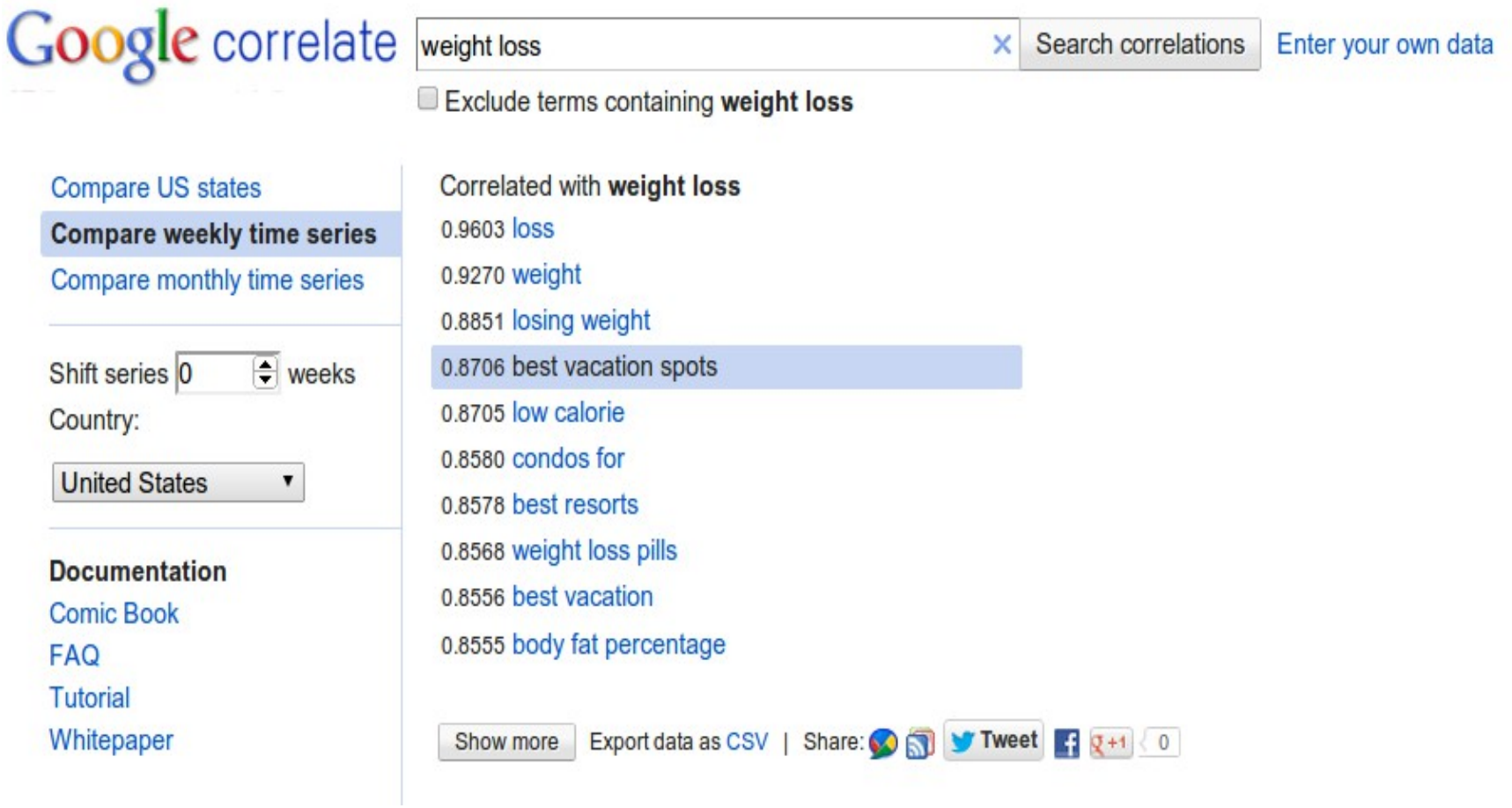
Google Trends

Google Correlate

Google Consumer Surveys

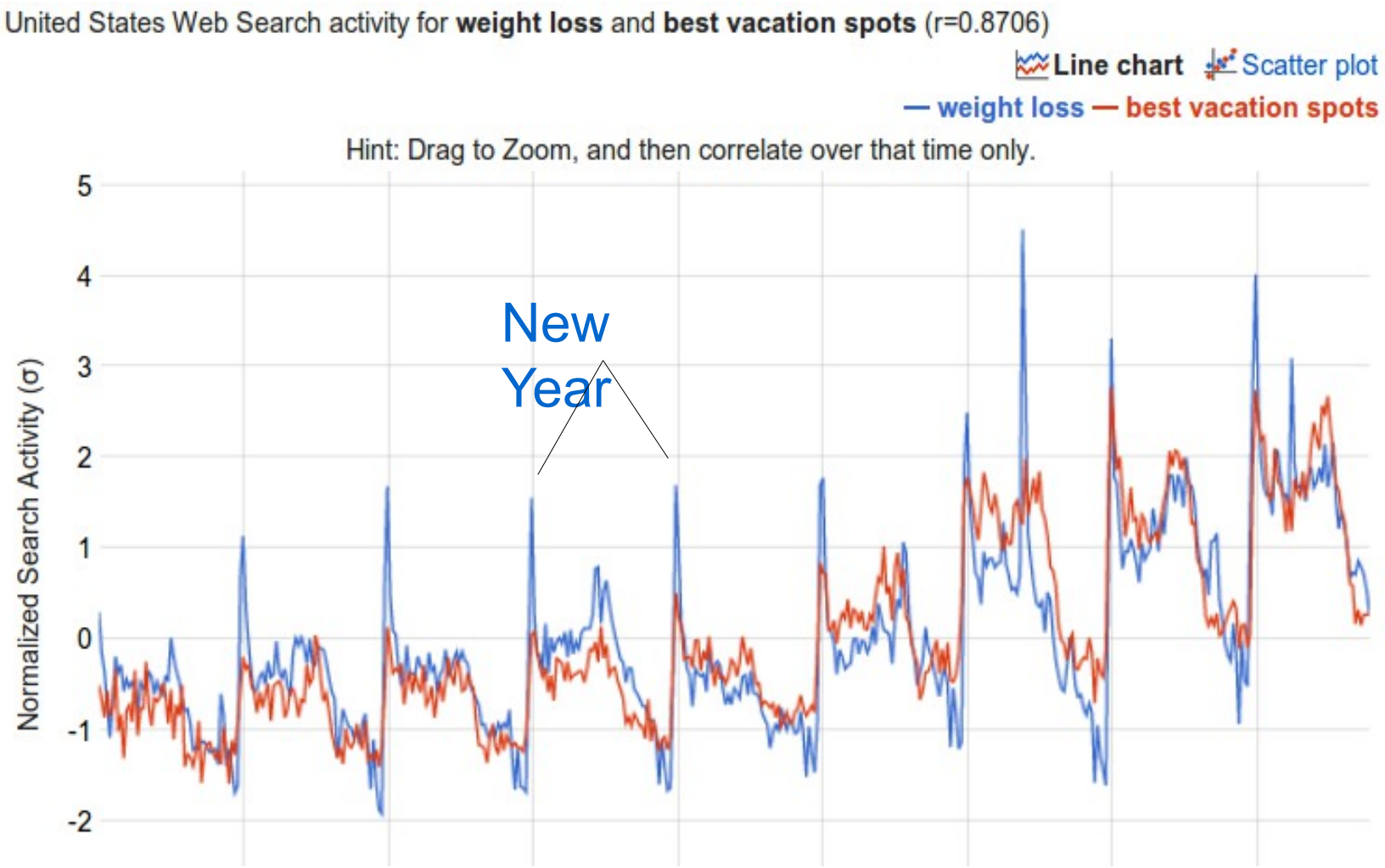
Searches correlated with [weight loss]

Google



Plot of [weight loss] and [best vacation spots]

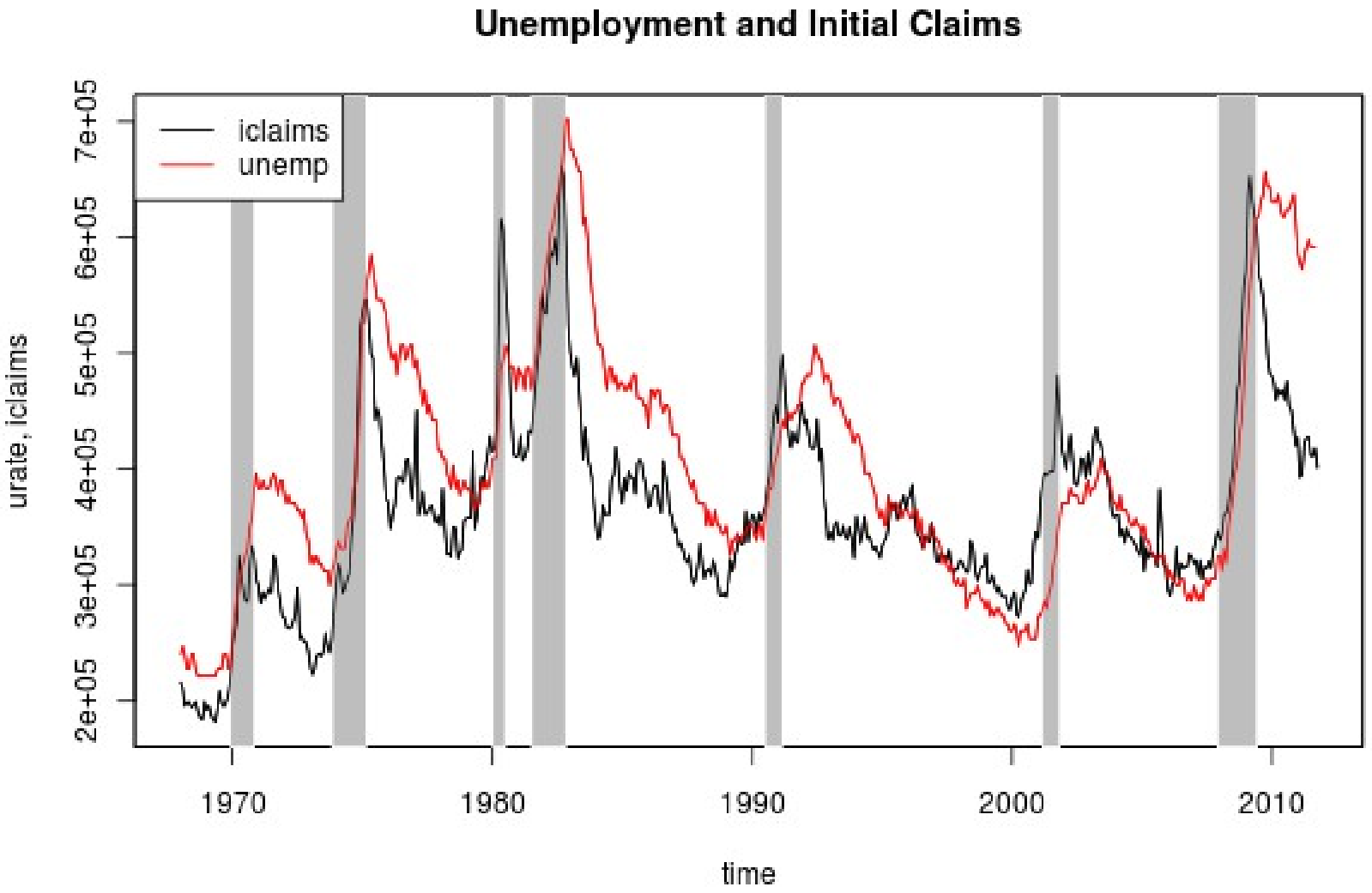
Google



Correlated with [weight loss] 3 weeks later

Google

Initial claims: good leading indicator for recessions



Grey bars indicate recessions

Google Correlate with initial claims data

Google

Compare US states

Compare weekly time series

Compare monthly time series

Shift series 0 weeks

Country:

United States

Documentation

Comic Book

FAQ

Tutorial

Whitepaper

Correlated with **Initial claims NSA**

0.8679 michigan unemployment

0.8273 idaho unemployment

0.8222 pennsylvania unemployment

0.8114 unemployment filing

0.8061 new jersey unemployment

0.8020 illinois unemployment

0.8017 department of unemployment

0.8012 rhode island unemployment

0.7939 unemployment office

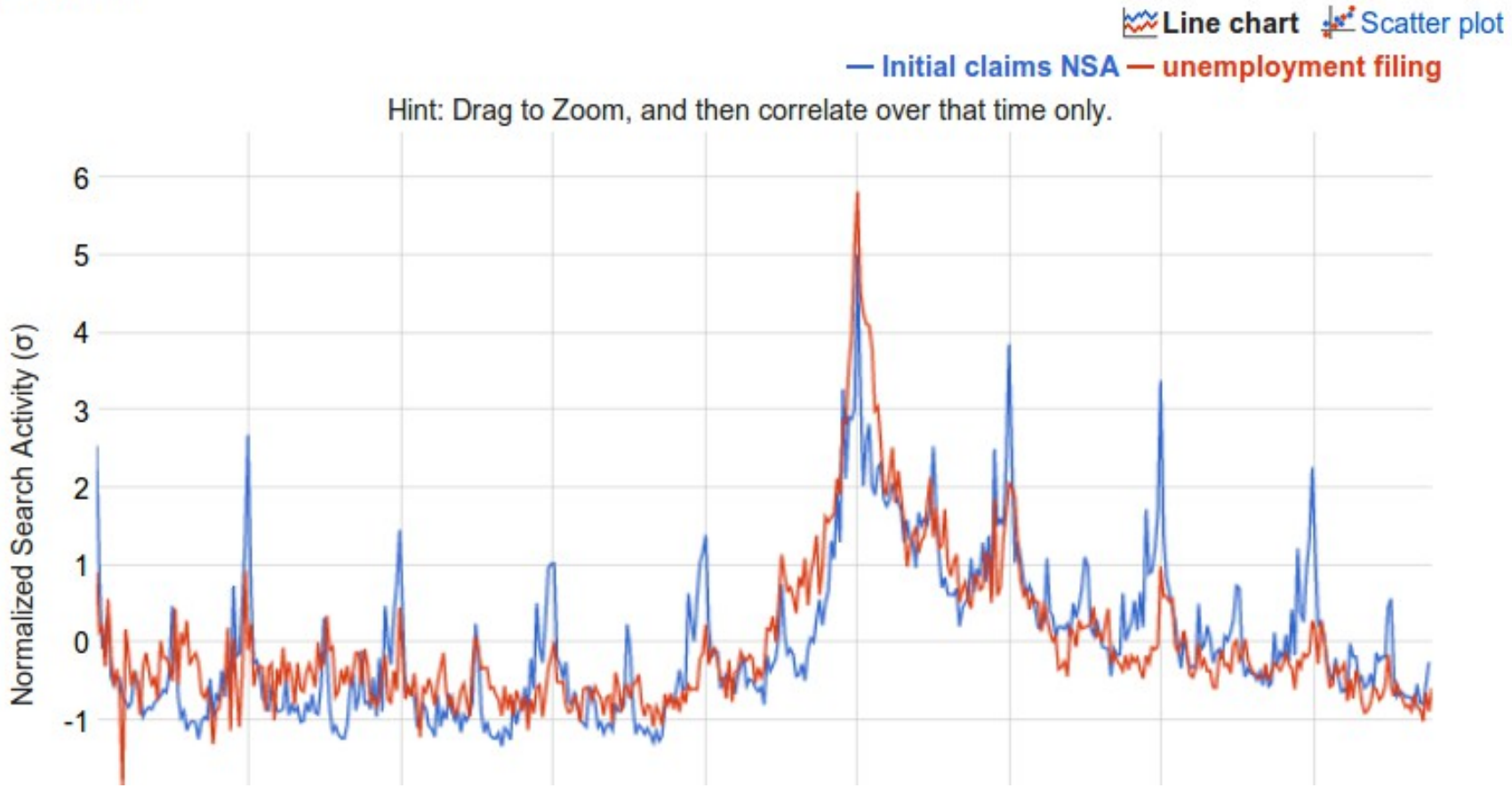
0.7933 filing unemployment

Show more | Export data as CSV | Share: Tweet +1 0

Initial claims and [unemployment filing]

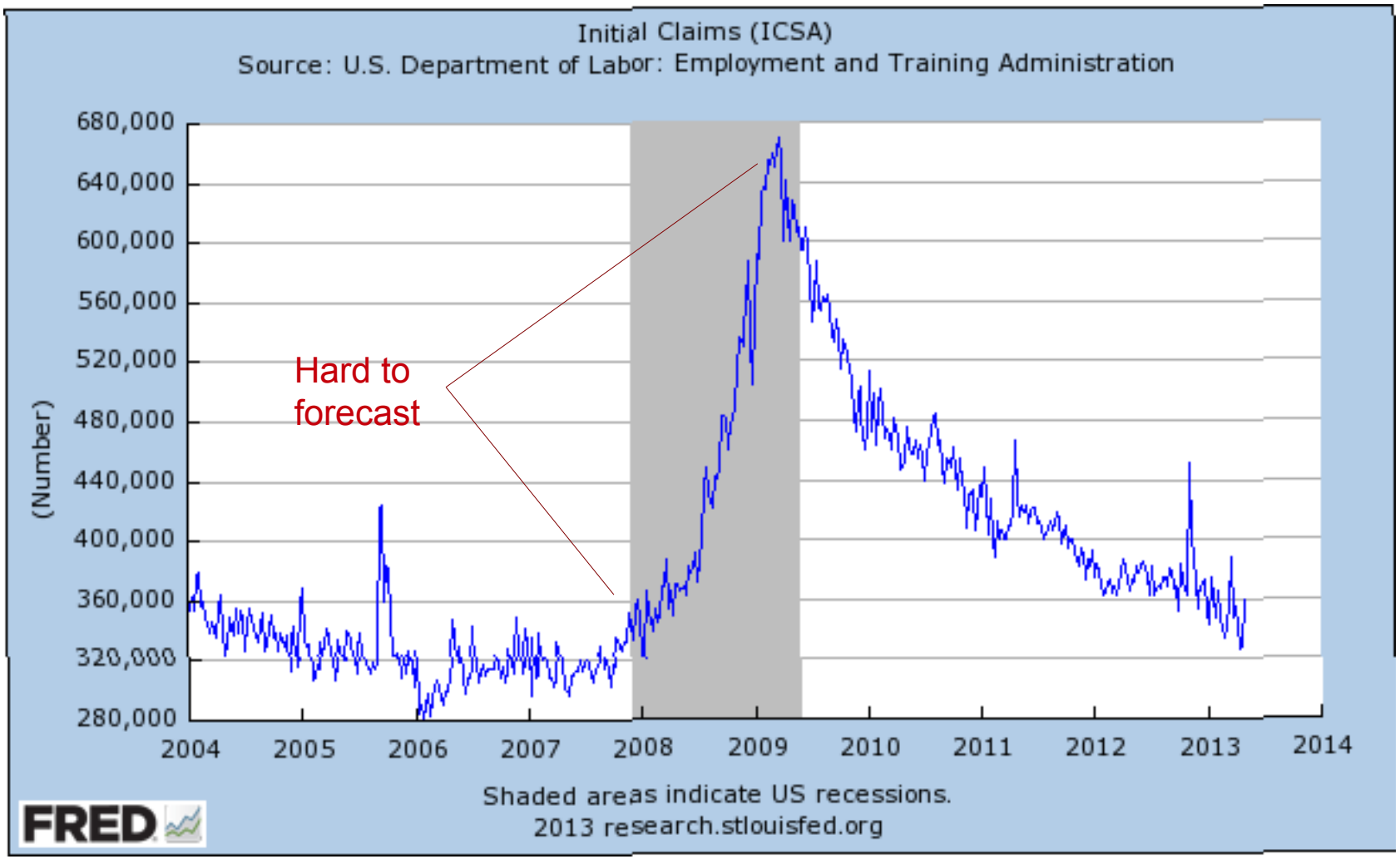
Google

User uploaded activity for **Initial claims NSA** and United States Web Search activity for **unemployment filing** (r=0.8114)



Initial claims, seasonally adjusted

Google



Regression models

Google

Baseline model

$y_t = a\, y_{t-1} + c + e_t$ gives an in-sample MAE of 3.1%

Adding the "unemployment filing" query

$y_t = a\, y_{t-1} + b\, q_t + c + e_t$ gives an in-sample MAE of 3.0%

Train using t weeks, forecast t+1 (rolling window forecast)

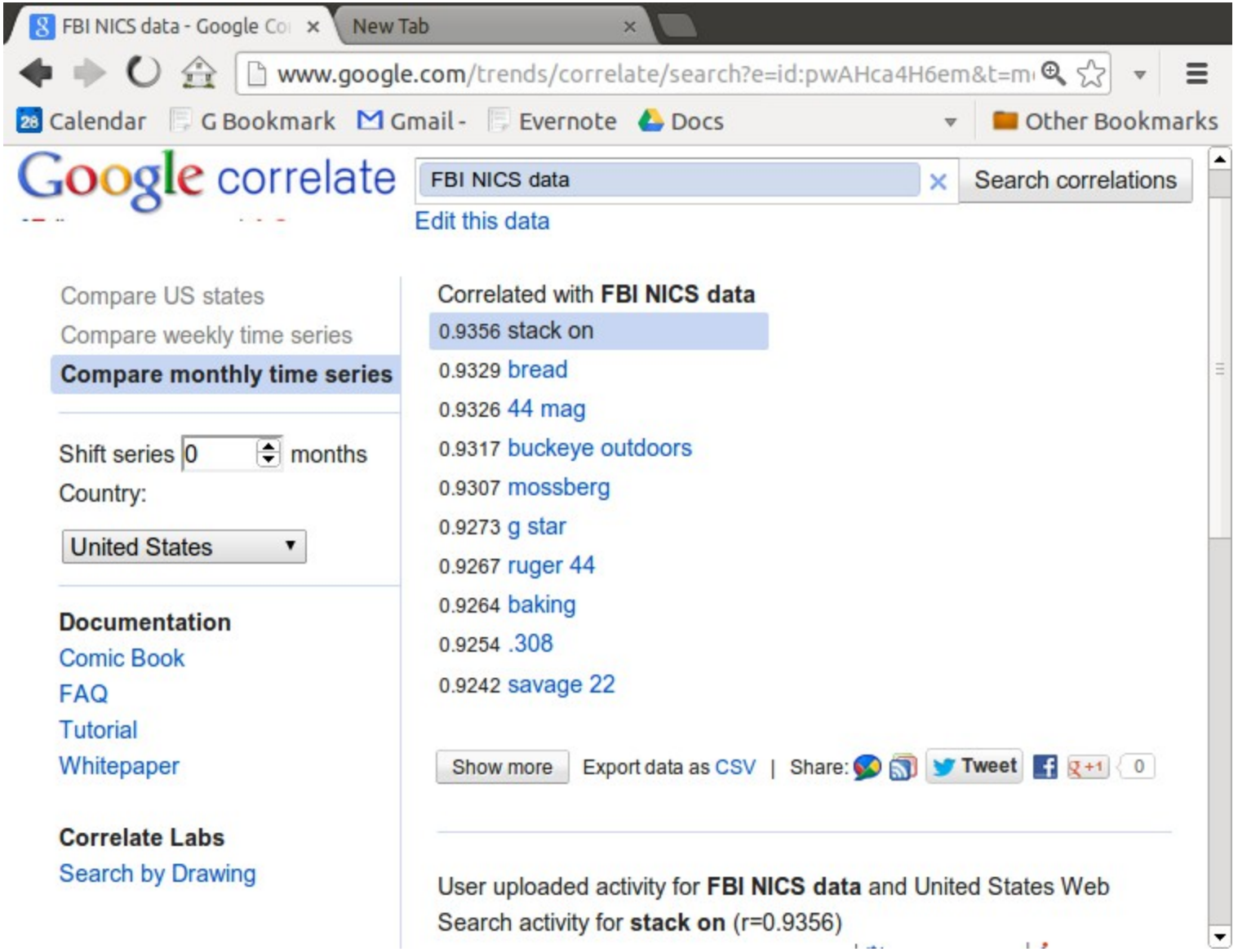
MAE of baseline = 3.2%, MAE with query = 3.2%,
0% improvement

During recession

MAE of baseline = 3.7%, MAE with query = 3.3%,
8.7% improvement

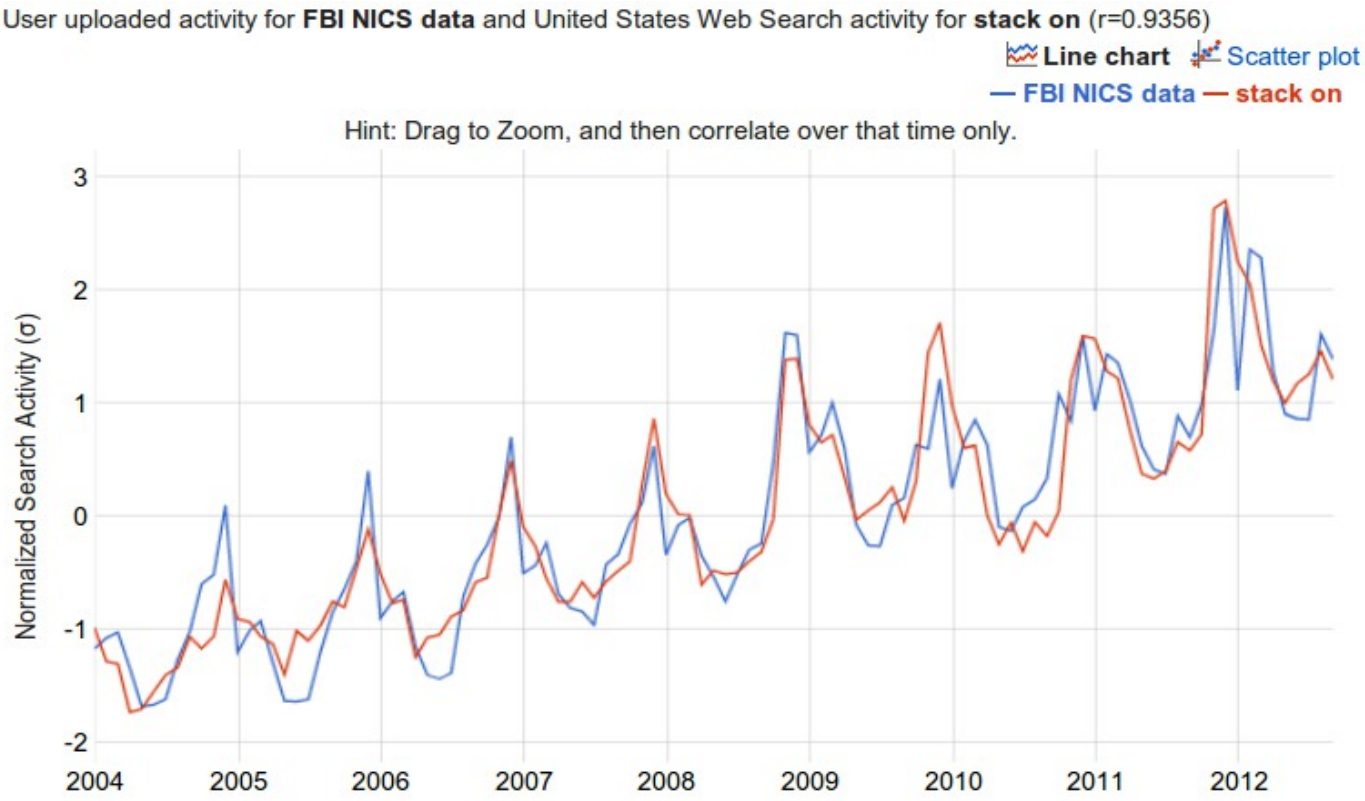
Gun sales background check

Google

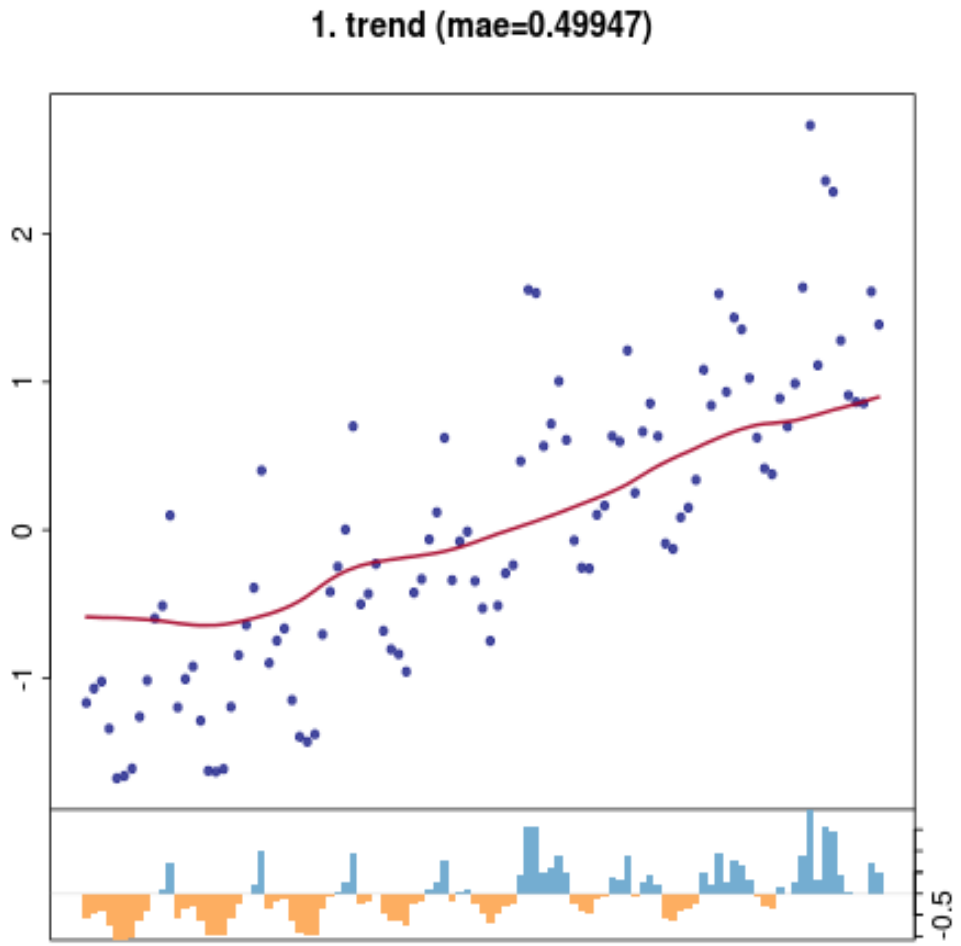


NICS time series

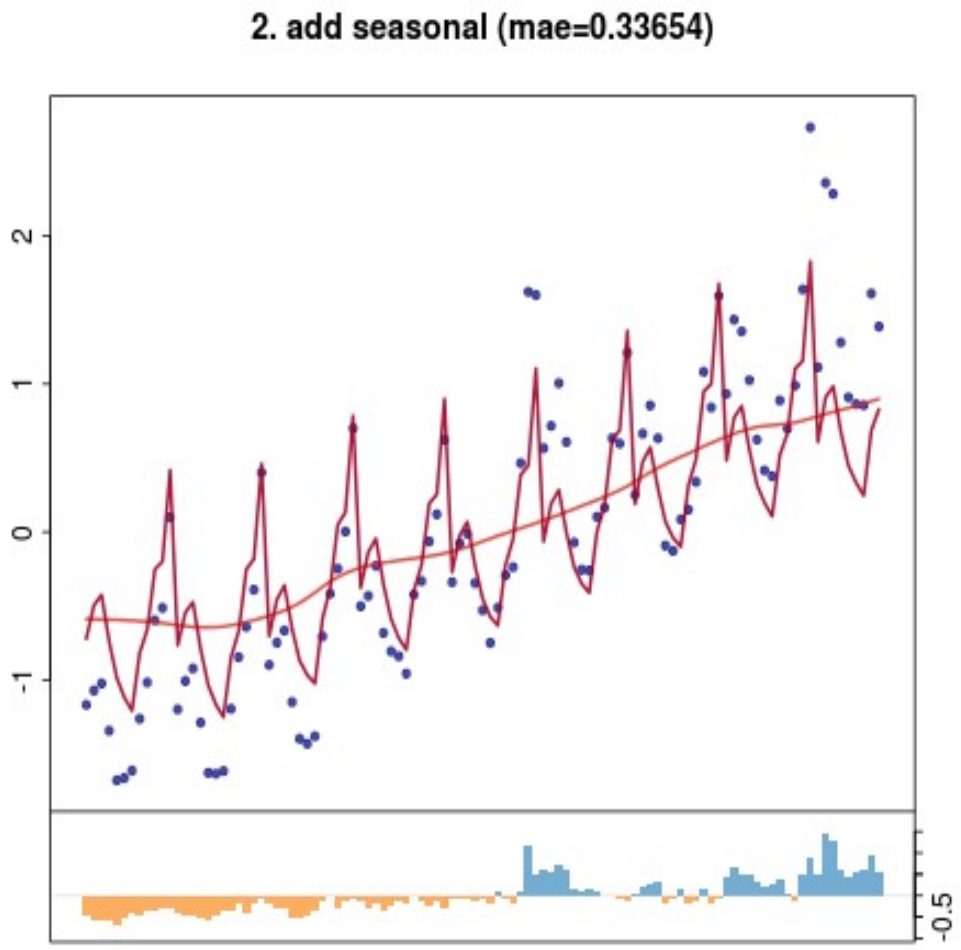
[stack on] has highest correlation
[gun shops] is chosen by BSTS



Trend

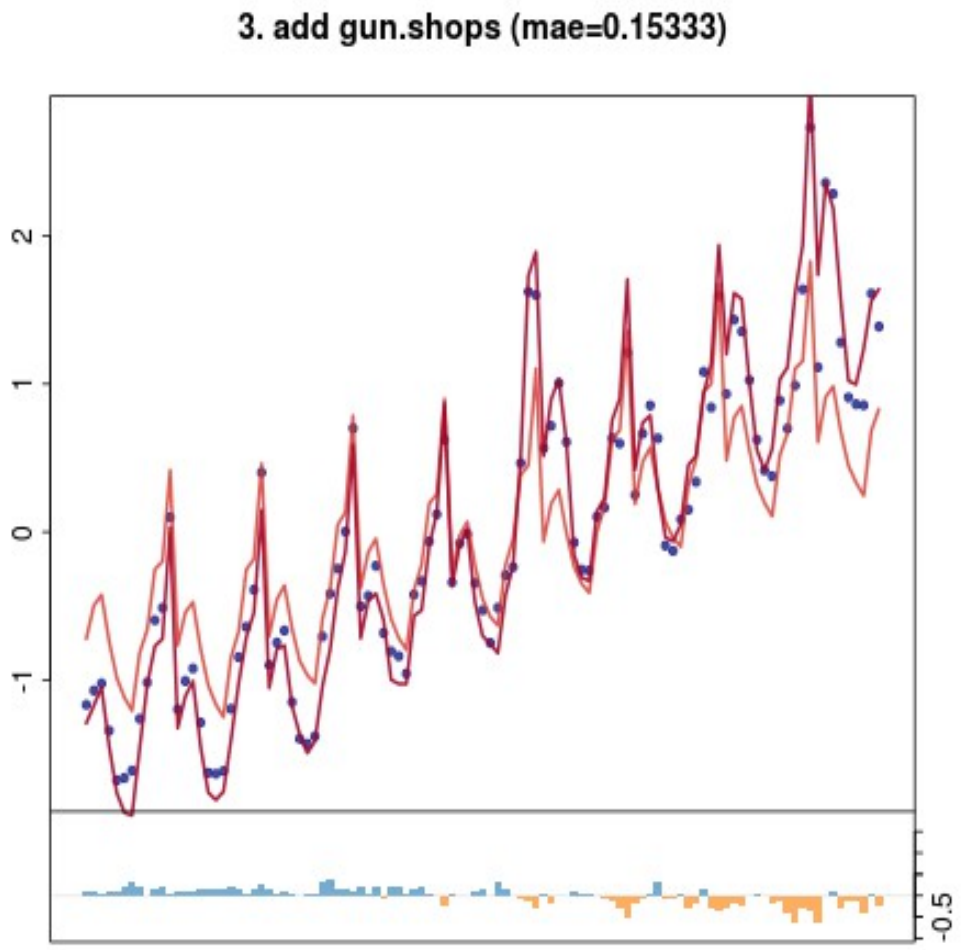


Seasonal



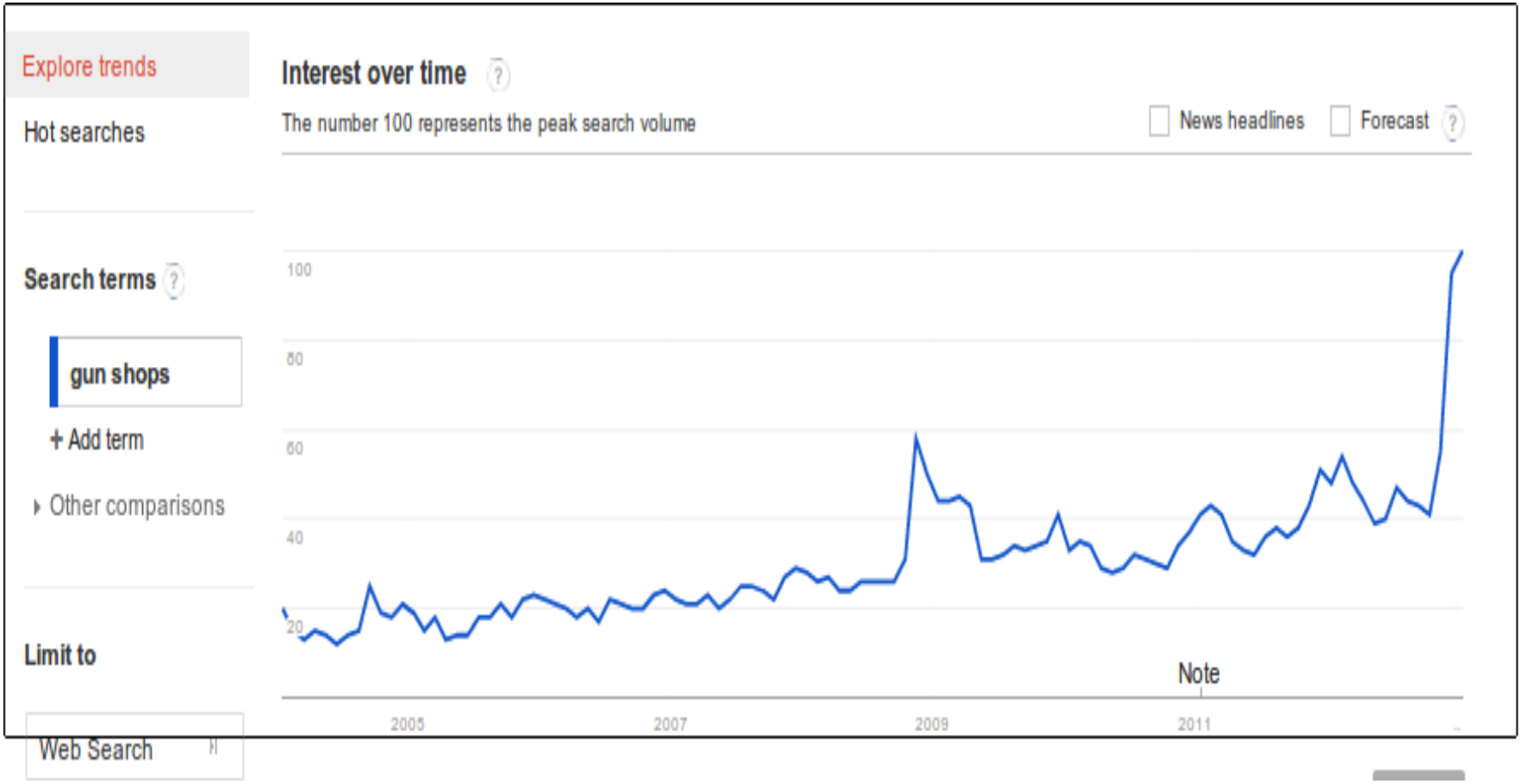
[gun shops]

Google



Searches on [gun shop]

Google



Our goal: automate model discovery

Google

Challenge 1: spurious correlation

Sometimes find correlations due simply to common seasonality or trend

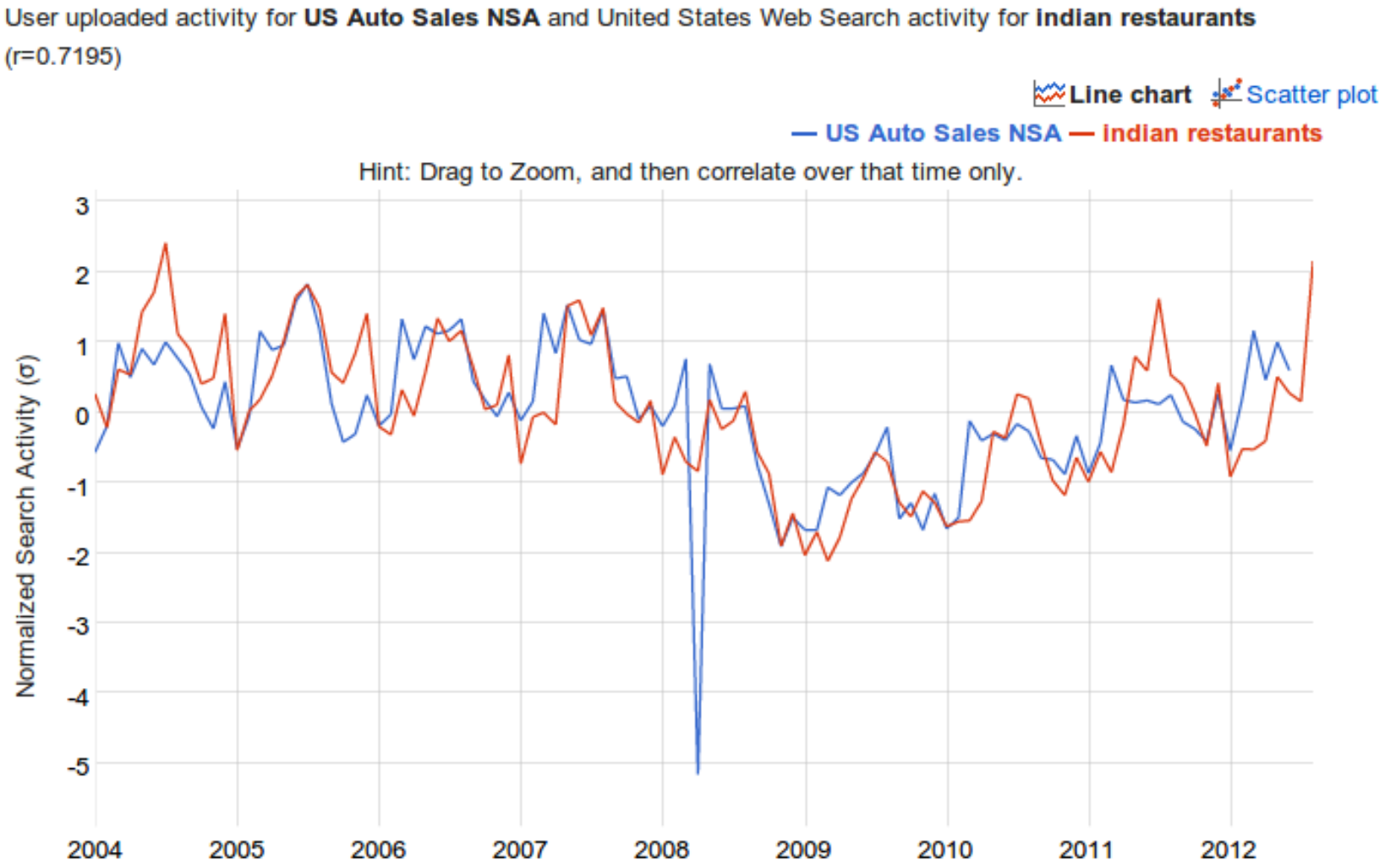
Challenge 2: fat regression

With more predictors than observations can always find a good fit

Challenge 3: overfitting

Within sample fits typically look better than out of sample fits

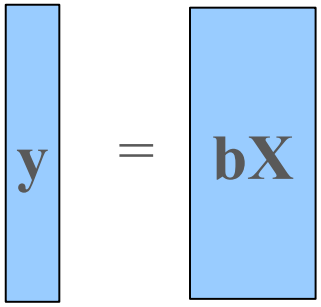
Challenge 1: Spurious Correlation



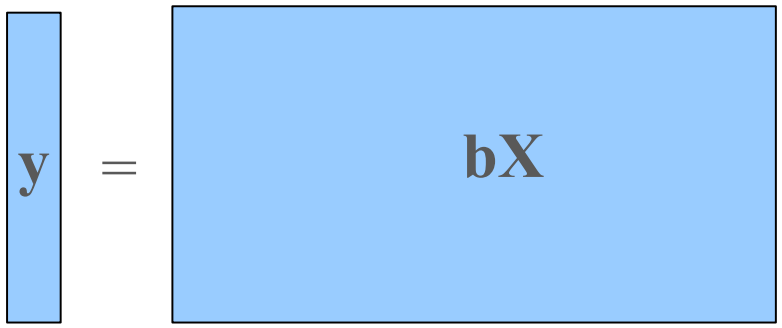
Challenge 2: Fat regression

Google

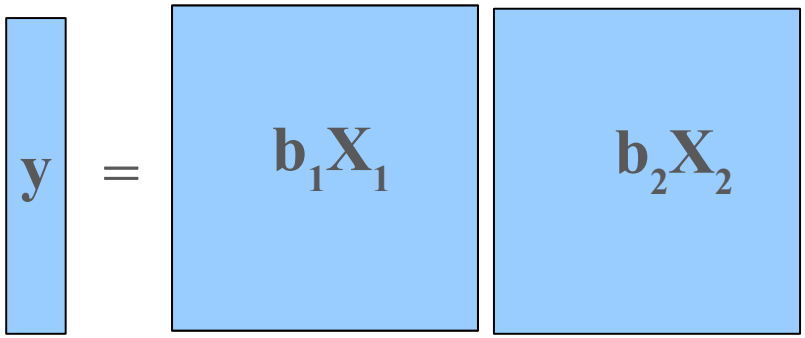
Slim regression



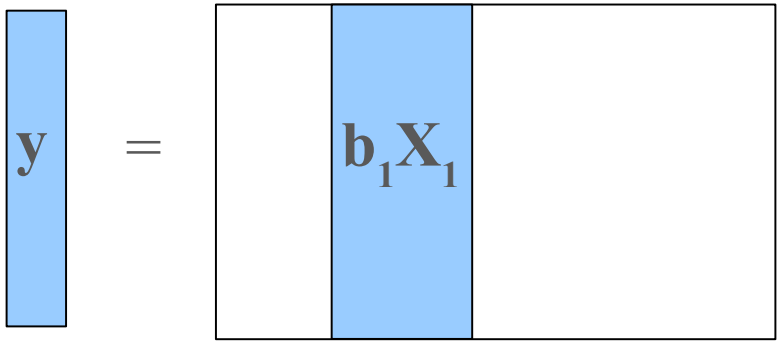
Fat regression



Any square subset of regressors will fit perfectly



A subset of regressors might fit well by chance



Our approach

Google

Estimating time series: use Kalman filter techniques

- Express time series as trend + seasonal + noise (“basic structural model”)
- Forecast univariate model using Kalman filter
- Advantages: flexibility, adaptive, interpretable, handles non-stationarity well

Model selection using “spike and slab” Bayesian regression

- Spike: prior probability that coefficient is included in regression
- Slab: diffuse prior for coefficient, conditional on inclusion
- Estimate a posterior probability that variable is in model
- Combines well with Kalman techniques

Final forecast is weighted average of many models, with weights given by posterior probabilities (Bayesian model averaging)

- Example of “ensemble estimation”
- Agnostic with respect to “true model”
- Tends to avoid overfitting by avoiding choice of “best” single model

Checklist

Google

Kalman filter: handles seasonality and trend

Spike and slab: handles variable selection

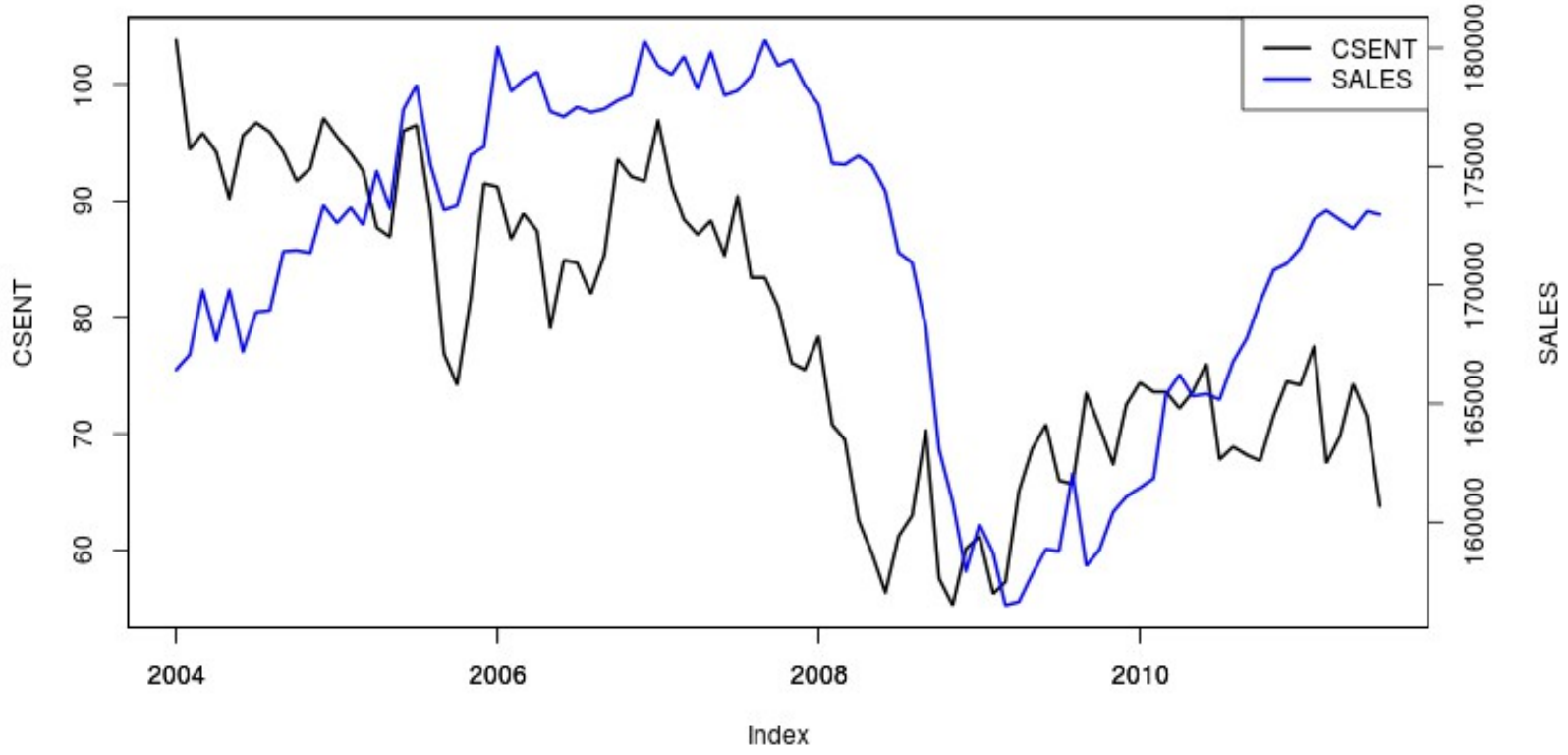
Model averaging: averages over many small models to avoid overfitting

UM consumer sentiment index

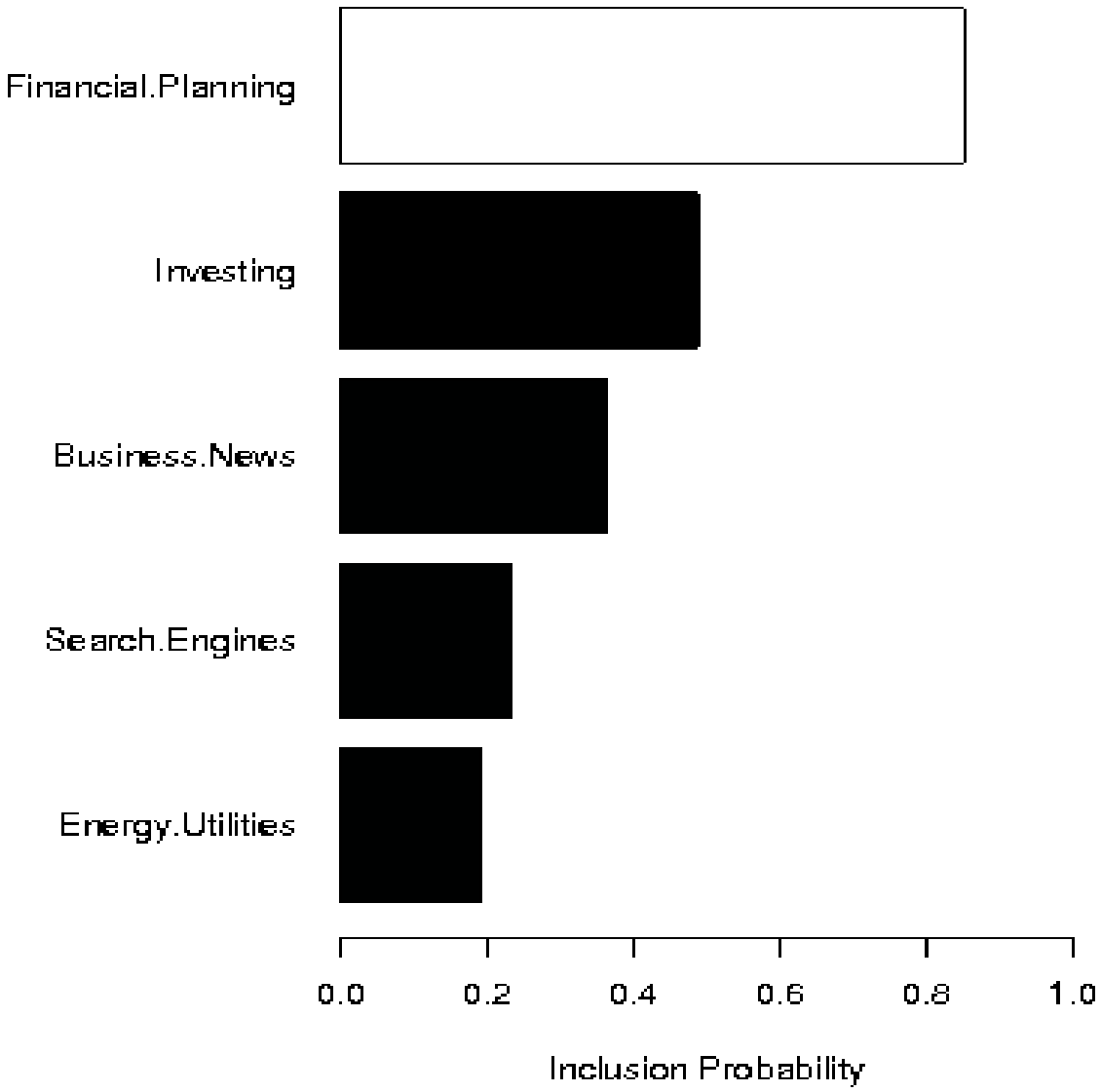
Google

Monthly data from University of Michigan survey

Select predictors using spike and slab from 157 Google economic verticals, using average value for first 2-weeks of month (about 3 weeks before data is released).



Probability of inclusion of predictor (n=98, k= 195)



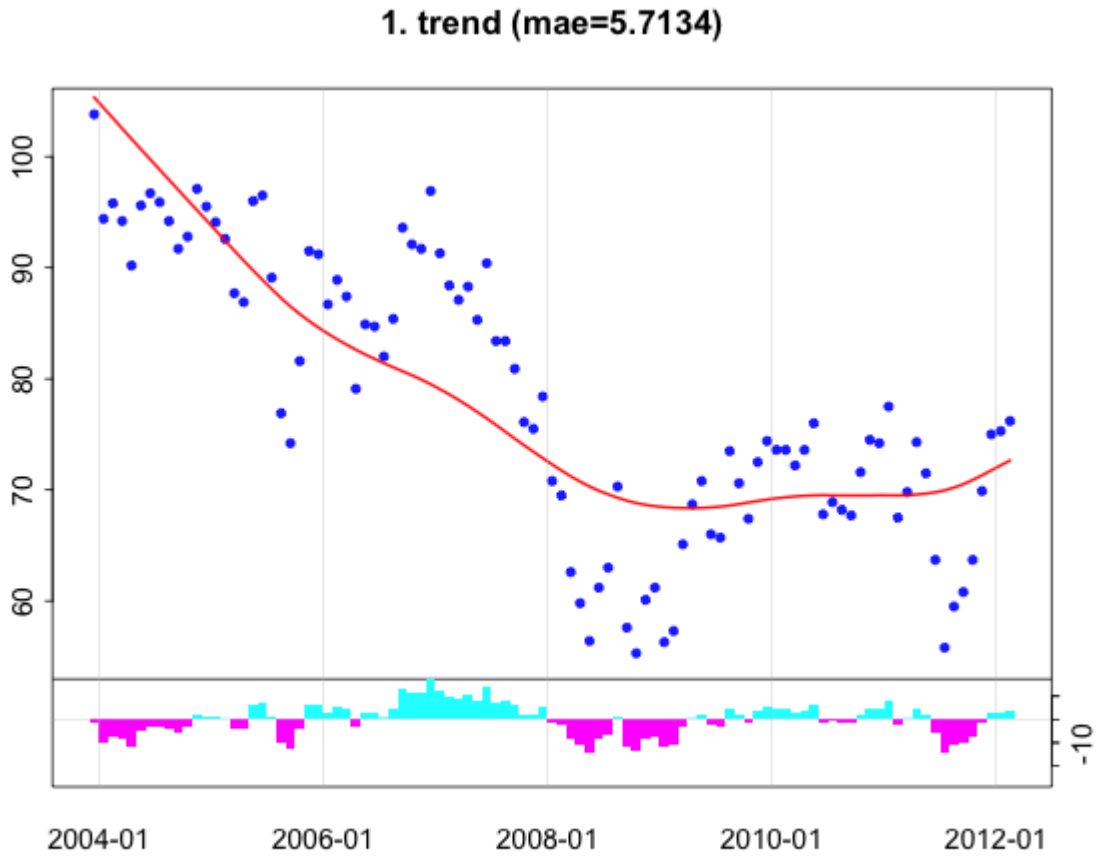
White: positive predictor

Black: negative predictor

Financial planning: personal finance, finance education, finance planning, finance schwab, financial literacy

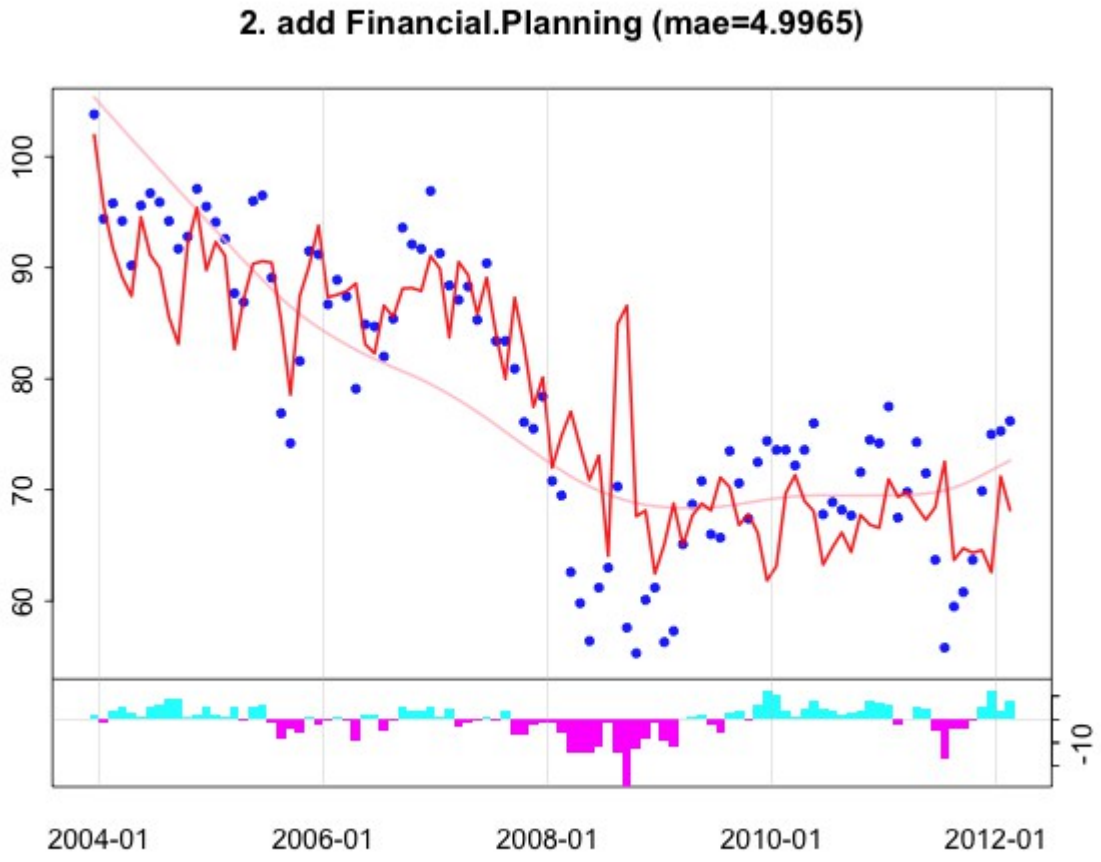
Investing: finance google, stock finance, stocks, etrade, ameritrade, gold

Start with “trend”

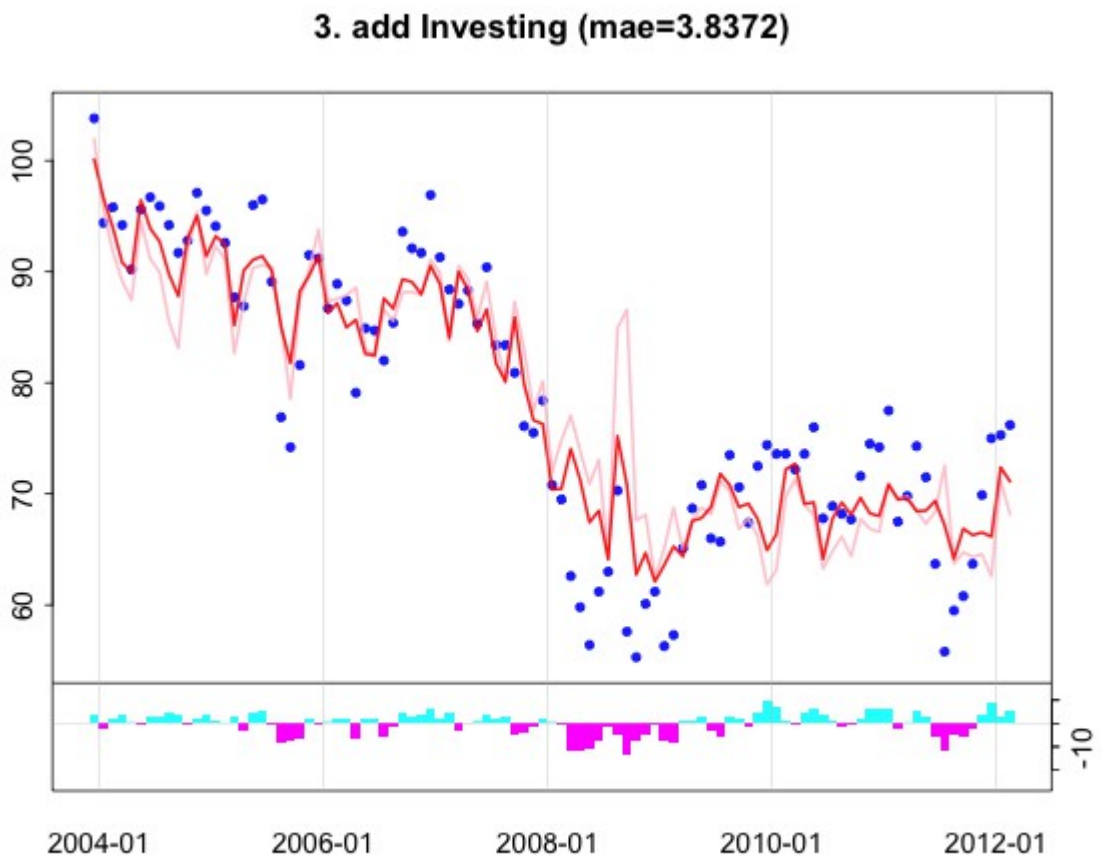


Add “financial planning”

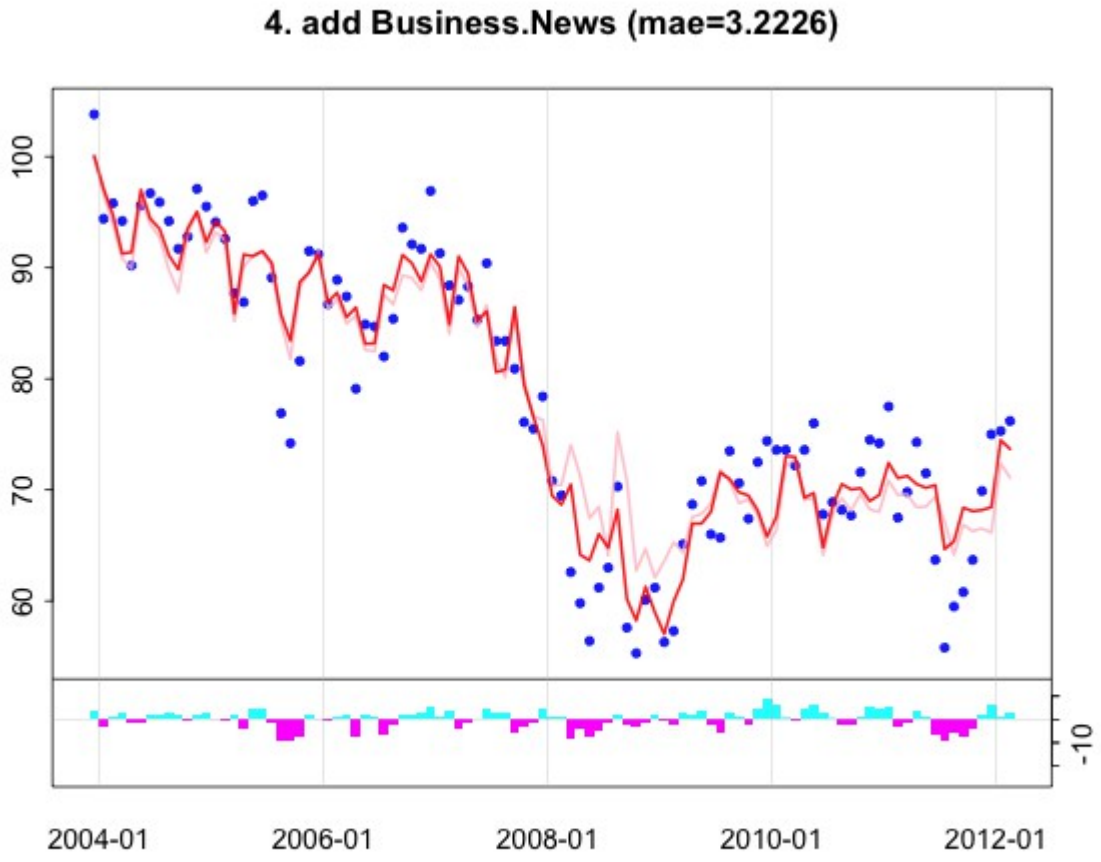
Google



Add “Investing”

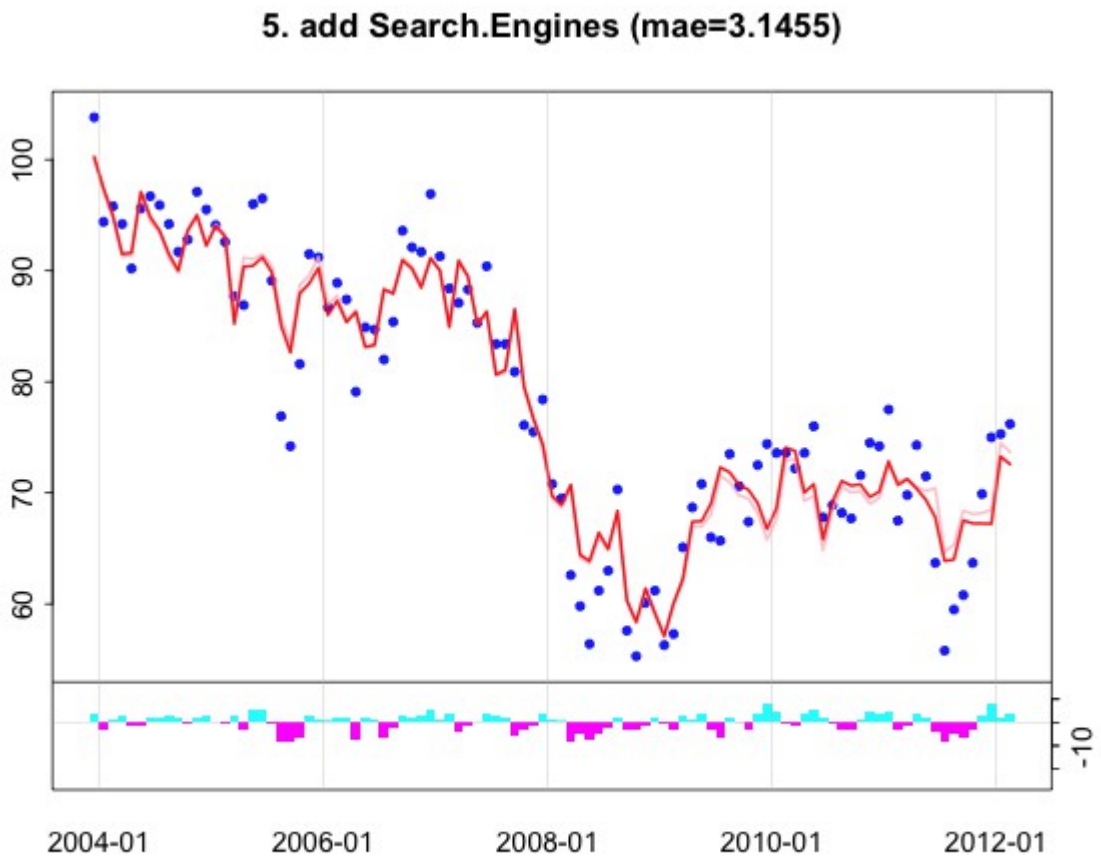


Add “Business News”

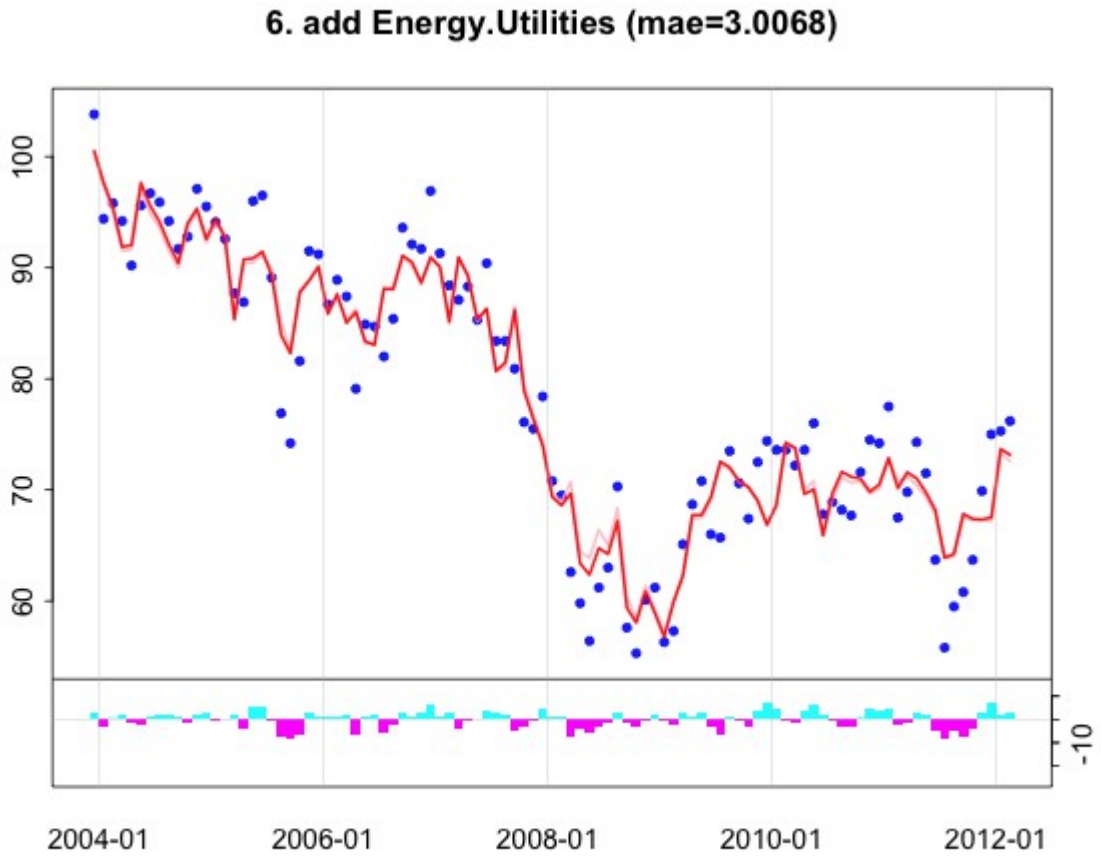


Add “Search Engines”

Google



Add Energy and Utilities



Google

Google Trends

Google Correlate

Google Consumer Surveys

How it works

You create online surveys to gain consumer insight

People complete questions to access premium content

Publishers get paid as their visitors answer

You get nicely aggregated and analyzed data

Bloomberg Businessweek
Businessweek Archives

Global Economics | Companies & Industries | Politics & Policy | Technology | Markets & Finance | Innovation & Design | Lifest

Data Mining: The Big Dig

Posted on June 11, 2000 | 0 Comments

More from Businessweek

Congress on the Couch, Budget Office Stuck Listening

No One Remembers When Bonds Went Truly Bad

To Add Variety and Control Cost, Fast Foods Go Small

The U.S. Economy Probably Grew After All, Thanks to Oil

HP Investors Face a Lonelier Road Ahead

Frontier: Instant Expert

Data Mining: The Big Dig

Your databases and Web sites hold vast stores of information on customer buying habits and market trends--if you know how to analyze the patterns. Some entrepreneurs are intimidated by technical issues or price: Hiring a pro for sophis...

Answer a question to continue reading this page

question 2 of 2:

Have you ever purchased anything from (check all that apply):

Check all answers that apply

- [ ] An email newsletter or ad
- [ ] A YouTube video
- [ ] An ad on your mobile phone
- [ ] An ad on your tablet
- [ ] None of the above

Submit answer(s)

Show me another question

Google

Screening question from Privacy concerns

Export

Report | Custom | Insights

Inferred Gender

Sum	Compare
Male	Female

Inferred Age

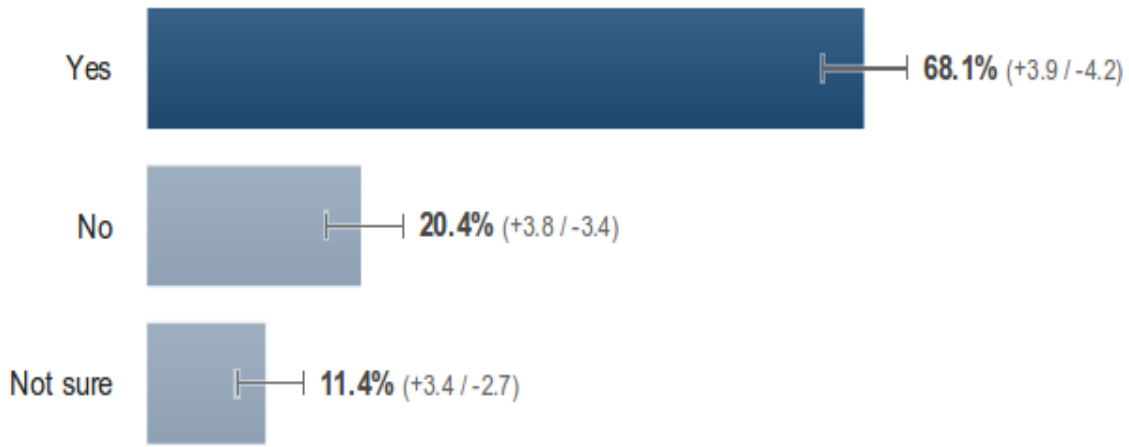
Sum	Compare
18-24	25-34
35-44	45-54
55-64	65+

Geography

SINGLE ANSWER

Are you concerned about your privacy online?

Results for respondents with demographics. Weighted by Age, Gender, Region. (575 responses)
Order statistically significant.



Google

Report | Custom | Insights

Inferred Gender

Sum	Compare
Male	Female

Inferred Age

Sum	Compare
18-24	25-34
35-44	45-54
55-64	65+

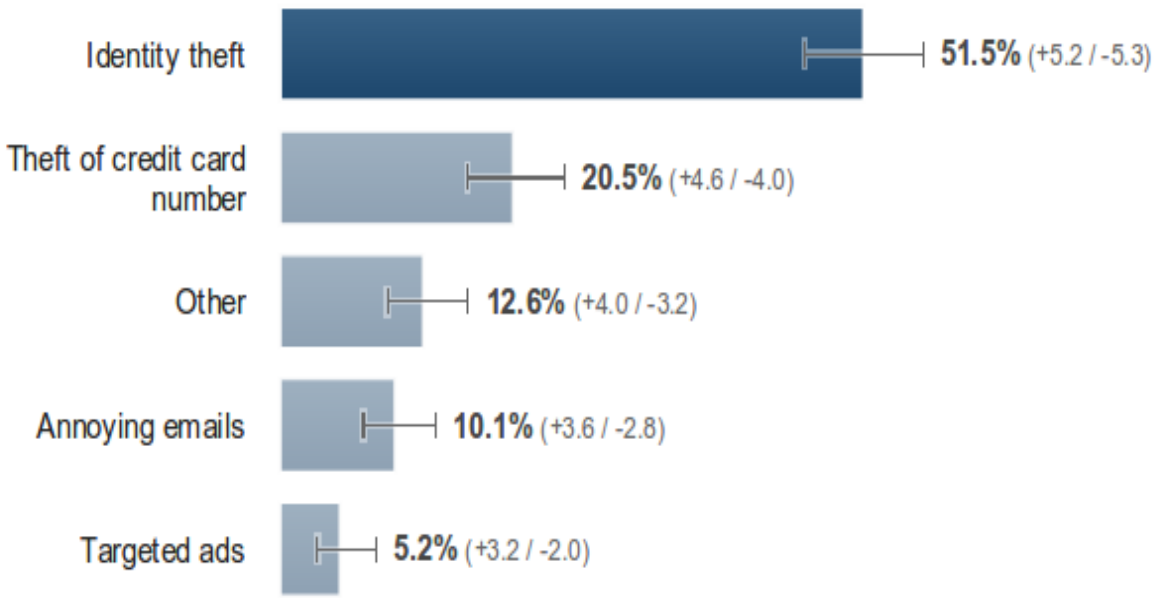
Geography

All of the USA

SINGLE ANSWER

What aspect of online privacy concerns you the most?

Results for respondents with demographics. Weighted by Age, Gender. (383 responses)
Winner statistically significant.



Google

Pollster Accuracy and Bias, 2012 Presidential Election

Likely Voters Polls in Last 21 Days of Campaign

Minimum 5 Polls

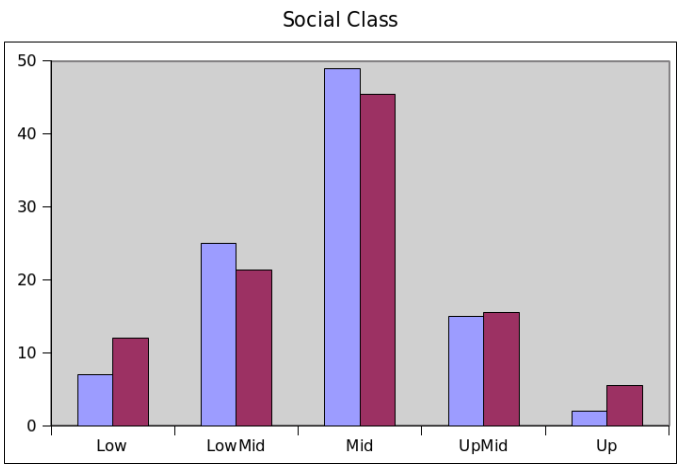
Pollster	# Polls	Avg. Error	Bias	Mode	Cell?
IBD / TIPP	11	0.9	R +0.1	Live Phone	Yes
Google Consumer Surveys	12	1.6	R +1.0	Internet	N/A
Mellman	9	1.6	R +0.0	Live Phone	Yes
RAND Corporation	17	1.8	D +1.5	Internet	N/A
CNN / Opinion Research	10	1.9	R +0.6	Live Phone	Yes
Ipsos / Reuters (online)	42	1.9	R +1.4	Internet	N/A
Angus Reid	11	1.9	R +0.8	Internet	N/A
CVOTER International / UPI	13	2.0	R +2.0	Live Phone	??
Grove Insight	18	2.0	R +0.1	Live Phone	Yes
SurveyUSA	17	2.2	R +0.5	Robodial	Yes
Quinnipiac	5	2.3	D +0.3	Live Phone	Yes
Marist	11	2.5	R +1.0	Live Phone	Yes
YouGov	30	2.6	R +1.1	Internet	N/A
We Ask America	9	2.6	D +0.1	Robodial	No
Public Policy Polling	71	2.7	R +1.6	Robodial	No
Gravis Marketing	16	2.7	R +2.7	Robodial	No
JZ Analytics*	17	2.8	R +0.1	Internet	N/A
Washington Post / ABC News	16	2.8	R +2.7	Live Phone	Yes
Pharos Research Group*	14	4.0	D +2.5	Live Phone	No
Rasmussen Reports	60	4.2	R +3.7	Robo + Internet	No
American Research Group	9	4.5	R +4.5	Live Phone	Yes
Mason-Dixon	8	5.4	R +2.2	Live Phone	Yes
Gallup	11	7.2	R +7.2	Live Phone	Yes

* Not used in FiveThirtyEight forecast.

3rd party analysis

Pew Foundation “A Comparison of Results from Surveys by the Pew Research Center and Google Consumer Surveys”

Nate Silver “Which Polls Fared Best (and Worst) in the 2012 Presidential Race”



Red: Google
Blue: Pew

How this changes surveys

Anyone can do them

The cost is dramatically lower

Results come back in a few hours

Surveys can be replicated … or not

You can detect sensitivity due to wording

Challenges for the future

Private sector has high-frequency, real time data and a lot of it!

- Visa, Mastercard, American Express
- UPS and FedEx
- Wal-Mart, Target, etc
- Supermarket scanner data
- Search engines

Government agencies

- Long historical series, but usually low frequency
- Carefully constructed but labor intensive, with delayed release and periodic revisions

How to combine the public and private data?

- How to integrate massive amounts of private sector real-time information with traditional government statistics