

# Google Tools for Data

Hal Varian  
Univ of Oregon  
Oct 2013



Google Trends

Google Correlate

Google Consumer Surveys

## **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]



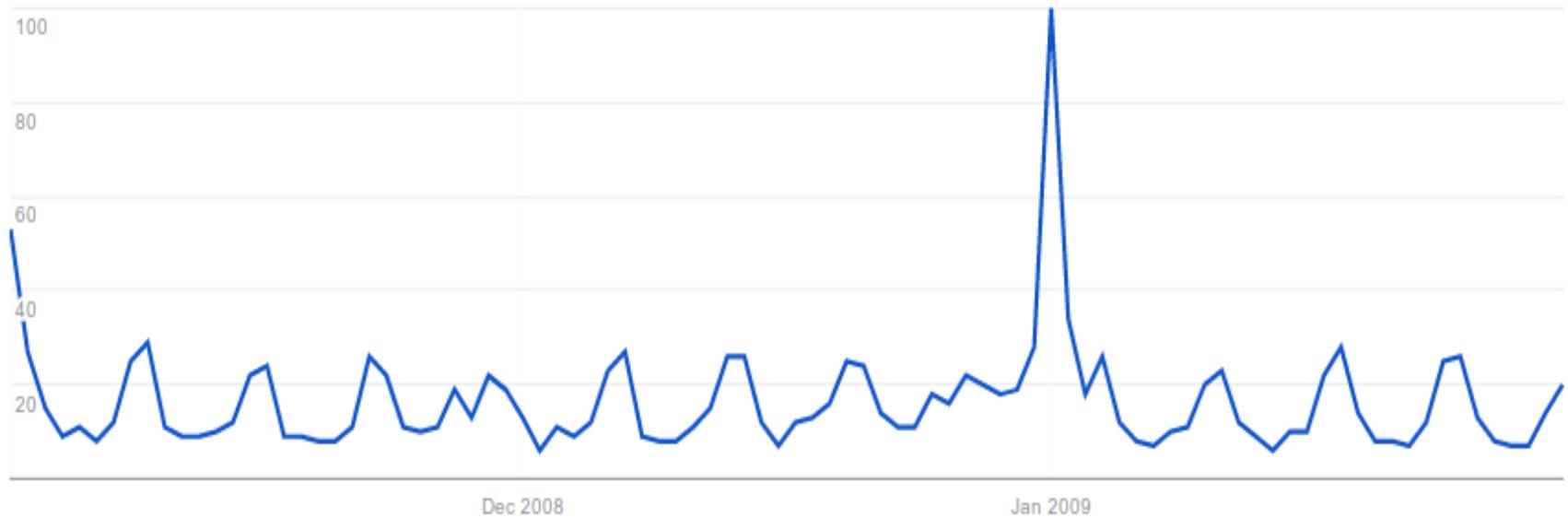
Web Search Interest: **hangover**. United States, Nov 2008 - Jan 2009. 



## Interest over time

The number 100 represents the peak search volume

News headlines  Forecast 

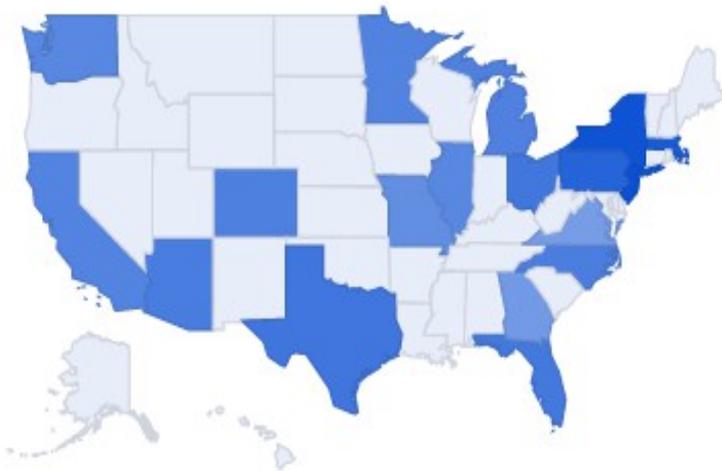


# Hangover by geography

## Regional interest ?



Worldwide > United States



0 100

Subregion | Metro | City

▶ View change over time ?

Embed

## Related terms ?



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



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

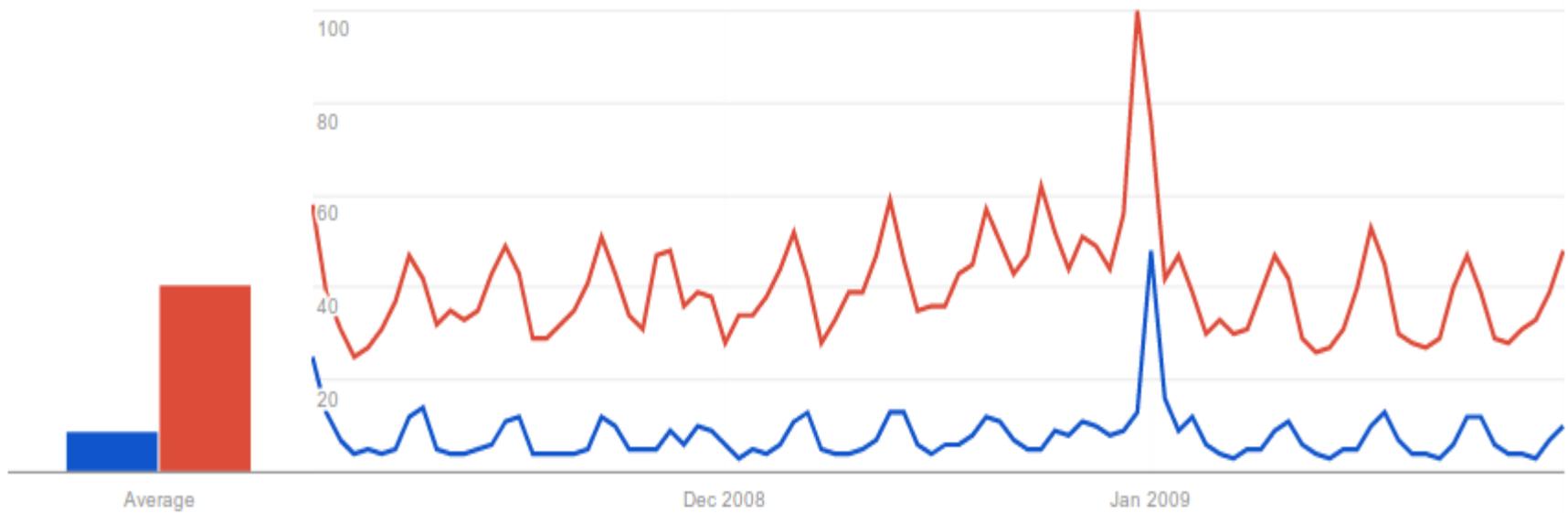


## Interest over time ?

The number 100 represents the peak search volume

News headlines

Forecast ?



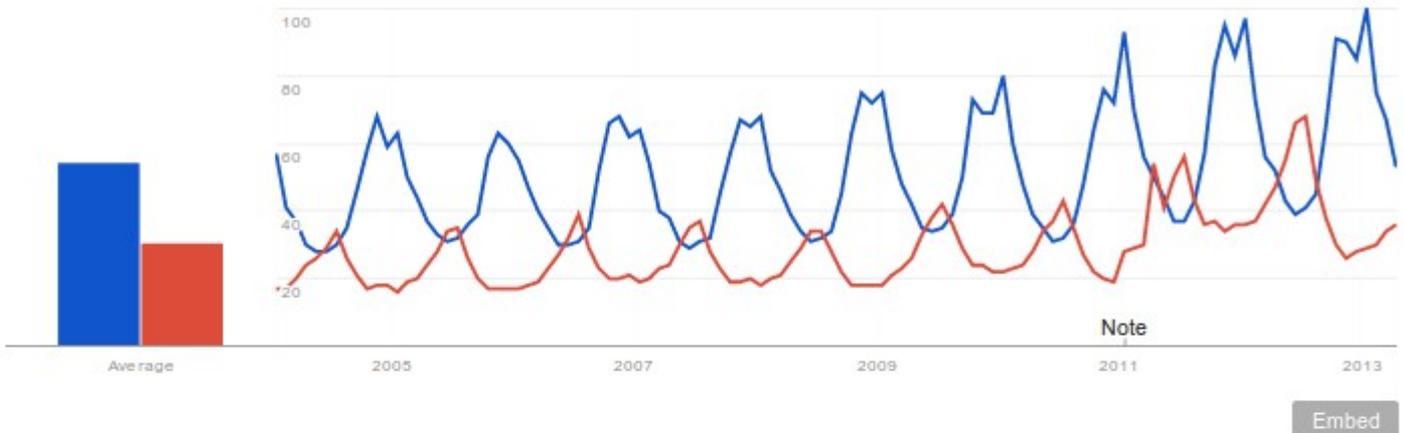
# Soup and ice cream

## Search terms ?

- soup
- ice cream
- + Add term
- Other comparisons

## Limit to

- Web Search H
- United States** H
- 2004 - present H
- All Categories H



## Regional interest ?

Worldwide > United States



## Related terms ?

	Top	Rising
soup recipe	100	<div style="width: 100%;"></div>
chicken soup	70	<div style="width: 70%;"></div>
recipes	50	<div style="width: 50%;"></div>
soup recipes	50	<div style="width: 50%;"></div>
potato soup	30	<div style="width: 30%;"></div>
the soup	25	<div style="width: 25%;"></div>

# Searches for [civil war]



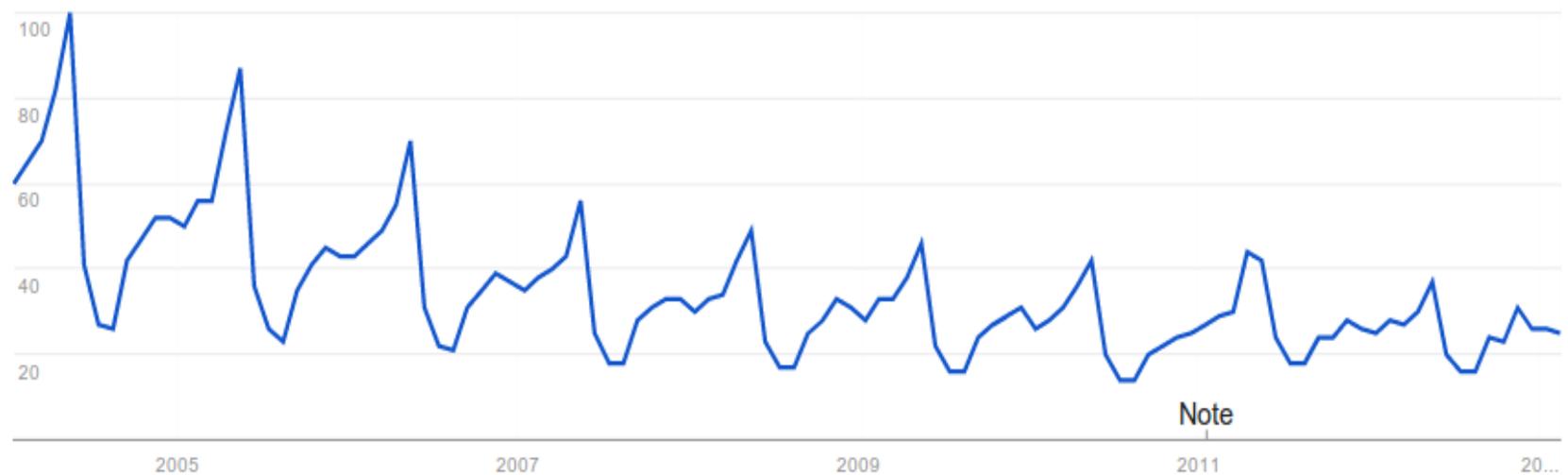
Web Search Interest: **civil war**. United States, 2004 - present. 



## Interest over time

The number 100 represents the peak search volume

News headlines  Forecast 





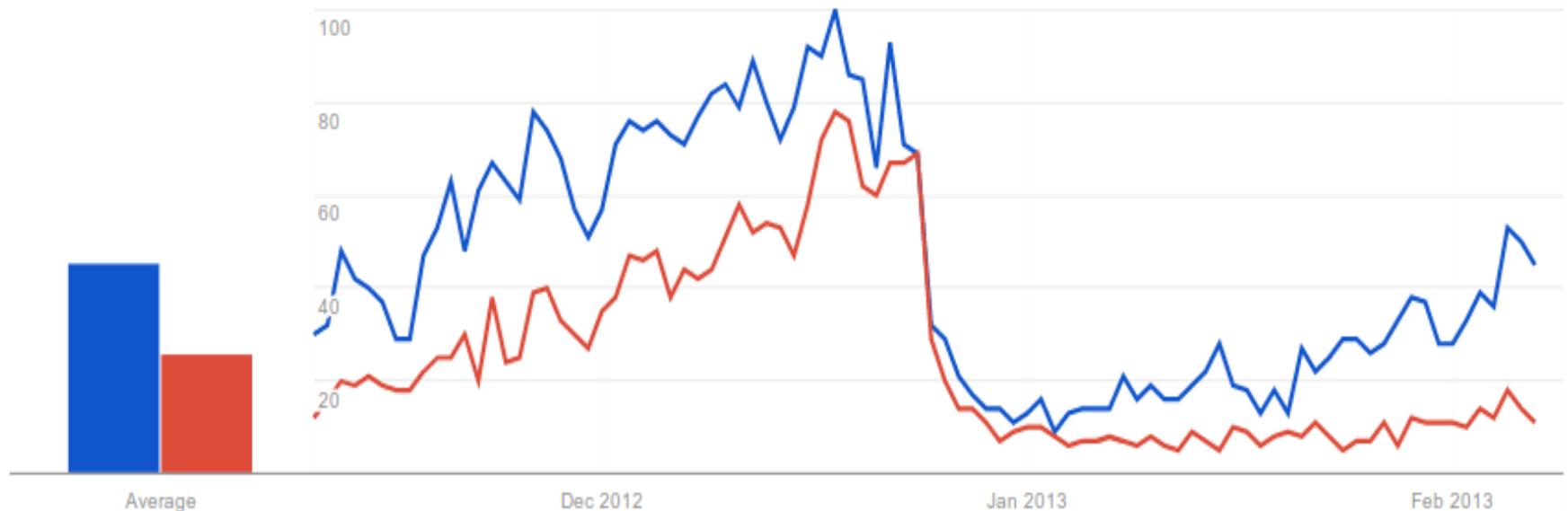
# Gift for boyfriend v Gift for girlfriend



## Interest over time ?

The number 100 represents the peak search volume

News headlines ?  Forecast ?



For boyfriend      For girlfriend

# Gift for husband v Gift for wife



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

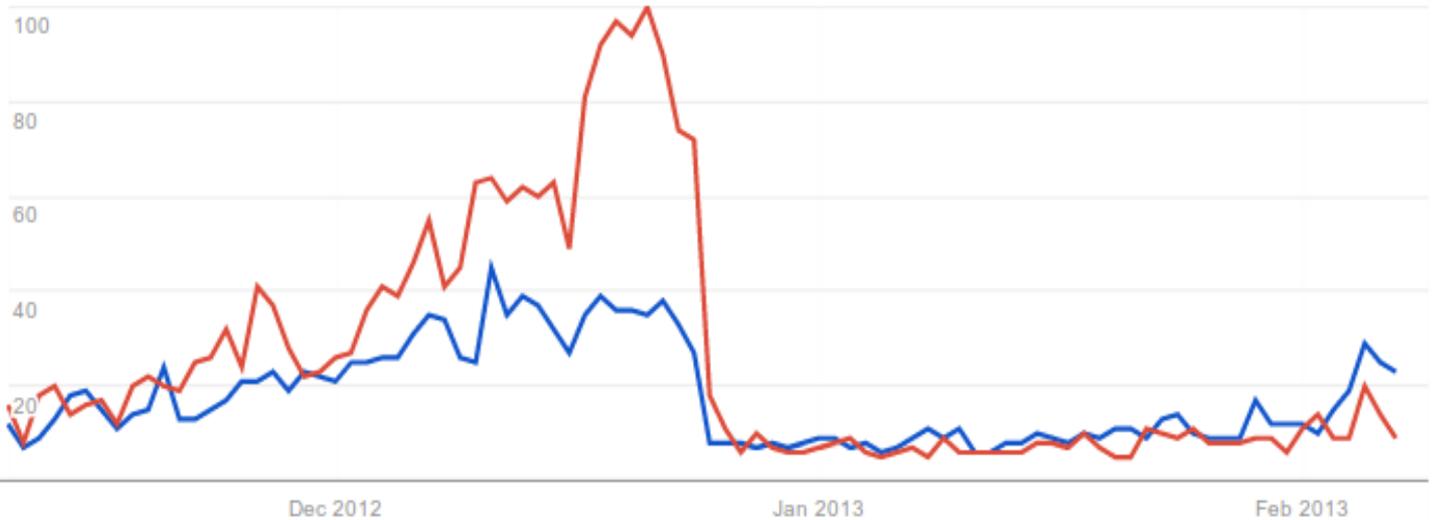


## Interest over time

The number 100 represents the peak search volume

News headlines 

Forecast 



Average   
For husband      For wife



Google Trends

Google Correlate

Google Consumer Surveys

# Searches correlated with [weight loss]



weight loss



Search correlations

Enter your own data

Exclude terms containing **weight loss**

[Compare US states](#)

**Compare weekly time series**

[Compare monthly time series](#)

Shift series  weeks

Country:

[United States](#)

## Documentation

[Comic Book](#)

[FAQ](#)

[Tutorial](#)

[Whitepaper](#)

## Correlated with **weight loss**

0.9603 [loss](#)

0.9270 [weight](#)

0.8851 [losing weight](#)

0.8706 [best vacation spots](#)

0.8705 [low calorie](#)

0.8580 [condos for](#)

0.8578 [best resorts](#)

0.8568 [weight loss pills](#)

0.8556 [best vacation](#)

0.8555 [body fat percentage](#)

Show more

Export data as [CSV](#)

Share:



Tweet



0

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

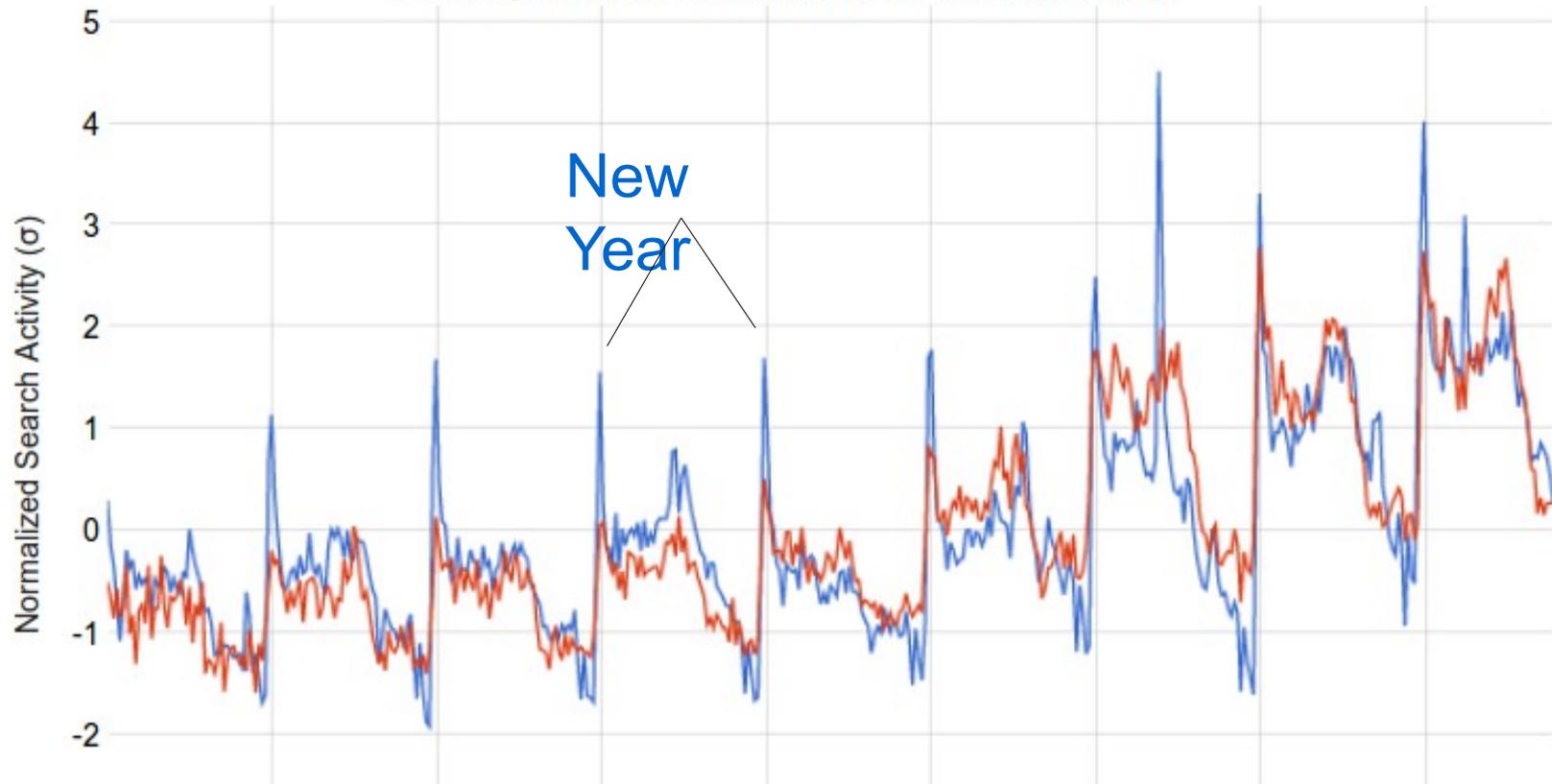


United States Web Search activity for **weight loss** and **best vacation spots** ( $r=0.8706$ )

Line chart Scatter plot

— weight loss — best vacation spots

Hint: Drag to Zoom, and then correlate over that time only.



# Correlated with [weight loss] 3 weeks later



weight loss



Search correlations

Enter your own data

Exclude terms containing **weight loss**

[Compare US states](#)

**[Compare weekly time series](#)**

[Compare monthly time series](#)

Shift series  weeks

Country:

## Documentation

[Comic Book](#)

[FAQ](#)

[Tutorial](#)

[Whitepaper](#)

## Correlated with **weight loss**

0.8124 [not losing weight](#)

0.8074 [protein bars](#)

0.7985 [weight loss plateau](#)

0.7979 [used car dealerships](#)

0.7958 [whey](#)

0.7933 [car dealerships](#)

0.7894 [themed bridal shower](#)

0.7893 [pure protein](#)

0.7893 [school district map](#)

0.7889 [salvage title](#)

Show more

Export data as [CSV](#)

Share:



Tweet



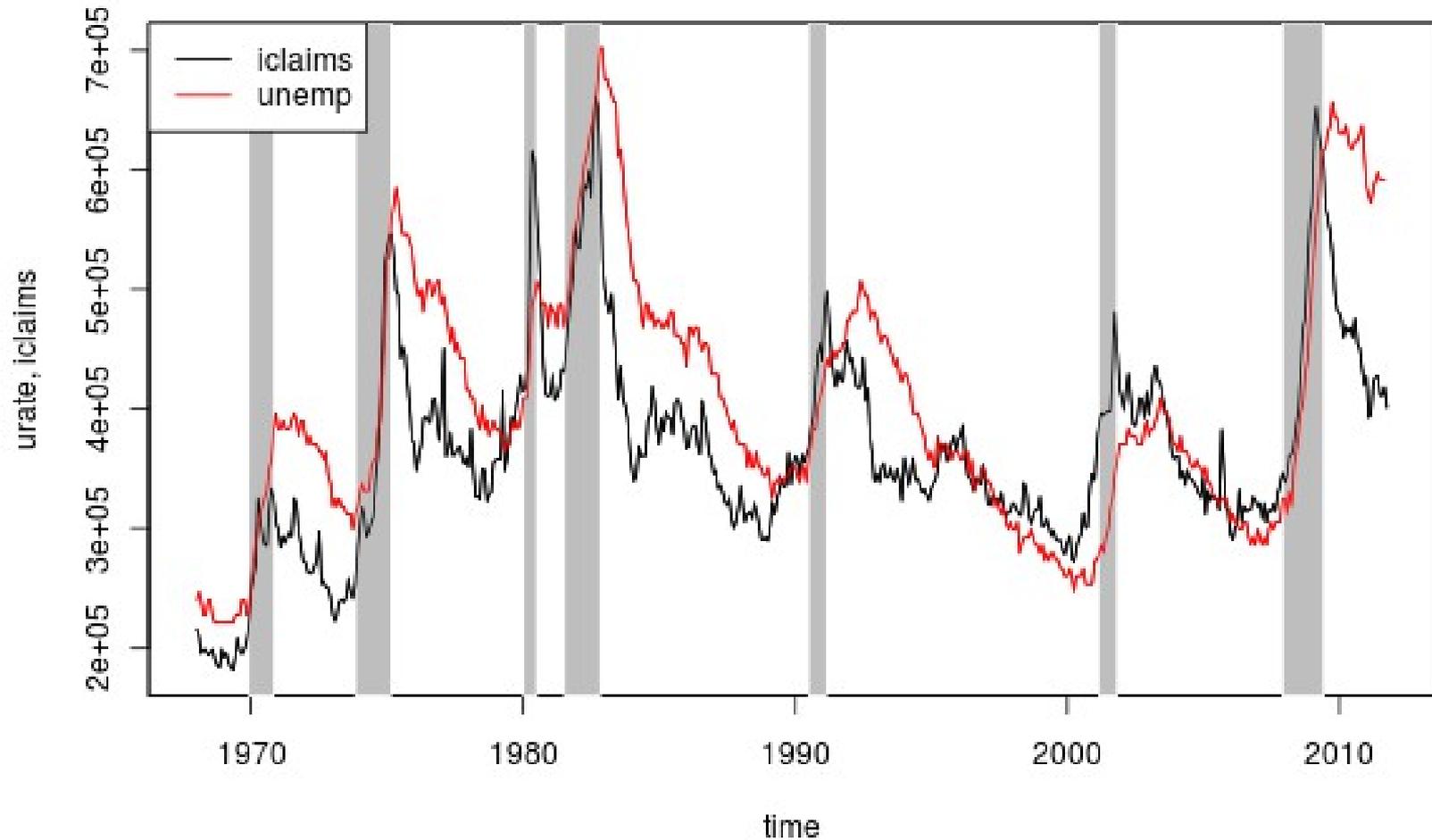
+1

0

# Initial claims: good leading indicator for recessions



## Unemployment and Initial Claims



Grey bars indicate recessions

# Google Correlate with initial claims data



Compare US states

**Compare weekly time series**

Compare monthly time series

Shift series  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



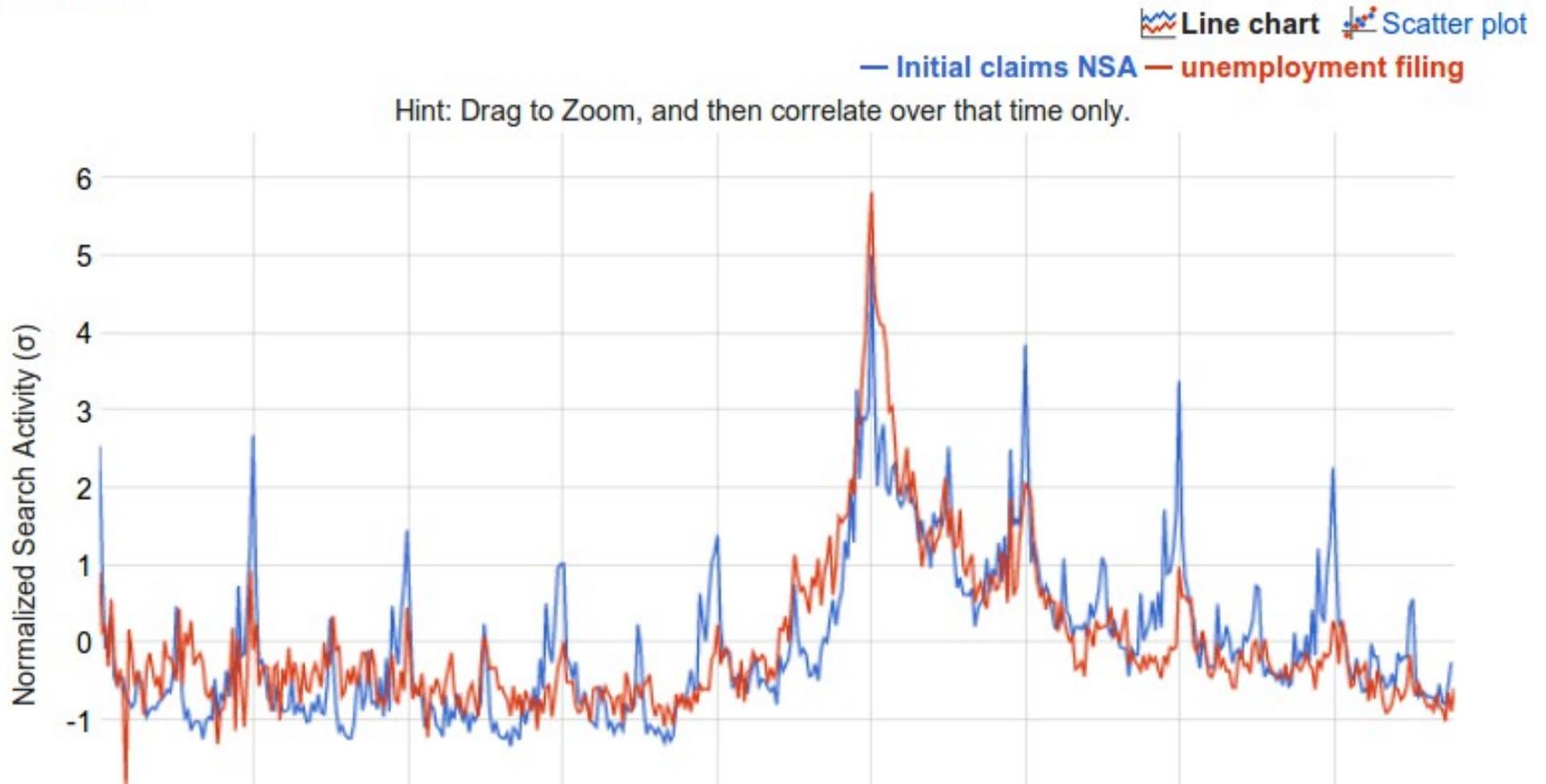
 +1

 0

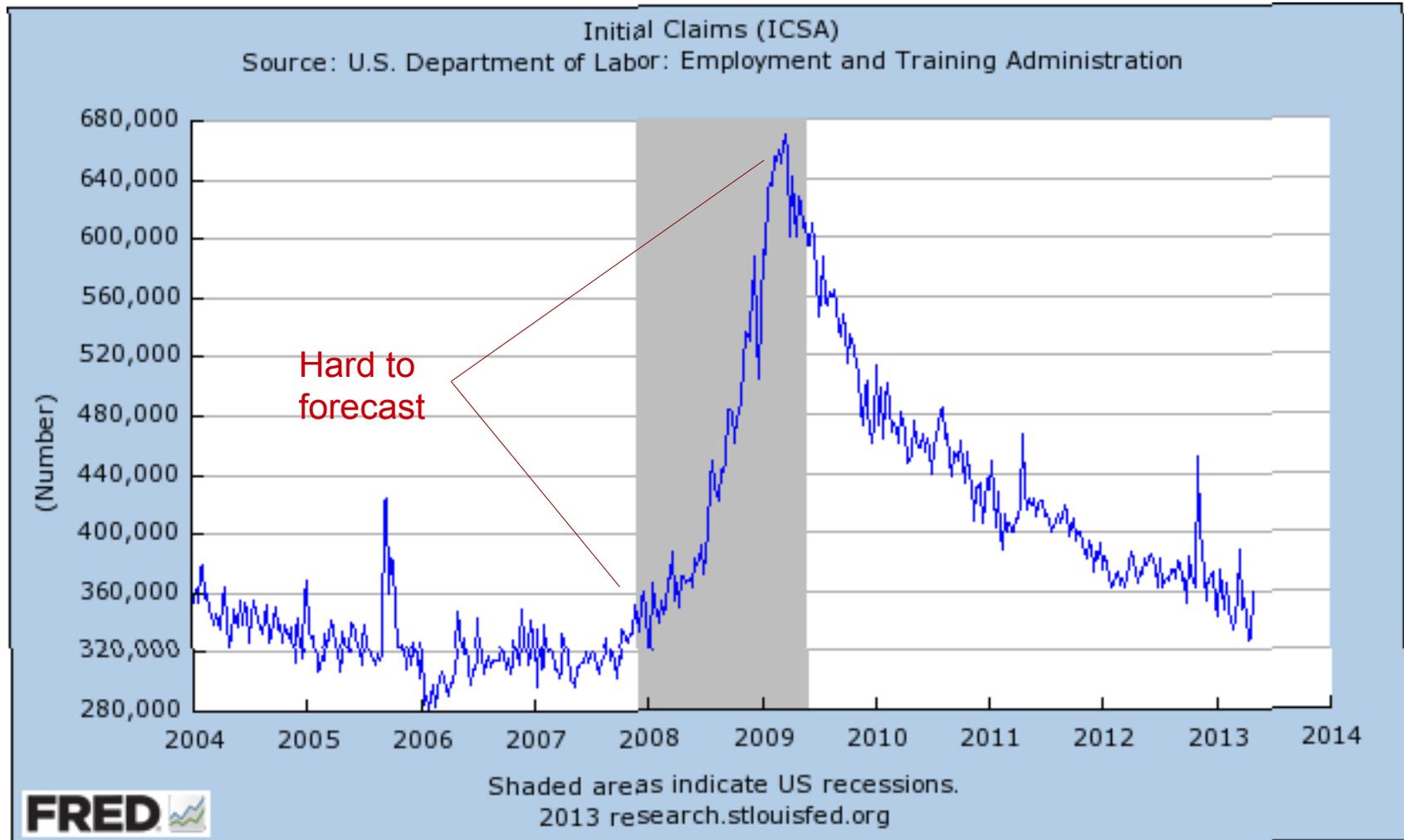
# Initial claims and [unemployment filing]



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



# Initial claims, seasonally adjusted



## 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



The screenshot shows a Google Correlate search for "FBI NICS data". The search results are displayed in a list format, with the top result being "stack on" with a correlation coefficient of 0.9356. Other results include "bread" (0.9329), "44 mag" (0.9326), "buckeye outdoors" (0.9317), "mossberg" (0.9307), "g star" (0.9273), "ruger 44" (0.9267), "baking" (0.9264), ".308" (0.9254), and "savage 22" (0.9242). The interface includes a search bar, a "Search correlations" button, and a list of related terms. The left sidebar contains navigation options like "Compare US states", "Compare weekly time series", and "Compare monthly time series". The bottom of the page shows a "Show more" button, an "Export data as CSV" option, and social media sharing icons for Google+, RSS, Twitter, Facebook, and +1.

FBI NICS data - Google Cor x New Tab x

www.google.com/trends/correlate/search?e=id:pwAHca4H6em&t=m

Calendar G Bookmark Gmail - Evernote Docs Other Bookmarks

Google correlate FBI NICS data Search correlations

Edit this data

Compare US states  
Compare weekly time series  
**Compare monthly time series**

Shift series 0 months  
Country: United States

**Documentation**  
Comic Book  
FAQ  
Tutorial  
Whitepaper

**Correlate Labs**  
Search by Drawing

**Correlated with FBI NICS data**

- 0.9356 stack on
- 0.9329 bread
- 0.9326 44 mag
- 0.9317 buckeye outdoors
- 0.9307 mossberg
- 0.9273 g star
- 0.9267 ruger 44
- 0.9264 baking
- 0.9254 .308
- 0.9242 savage 22

Show more Export data as CSV | Share: Tweet 0

User uploaded activity for **FBI NICS data** and United States Web Search activity for **stack on** (r=0.9356)

# NICS time series

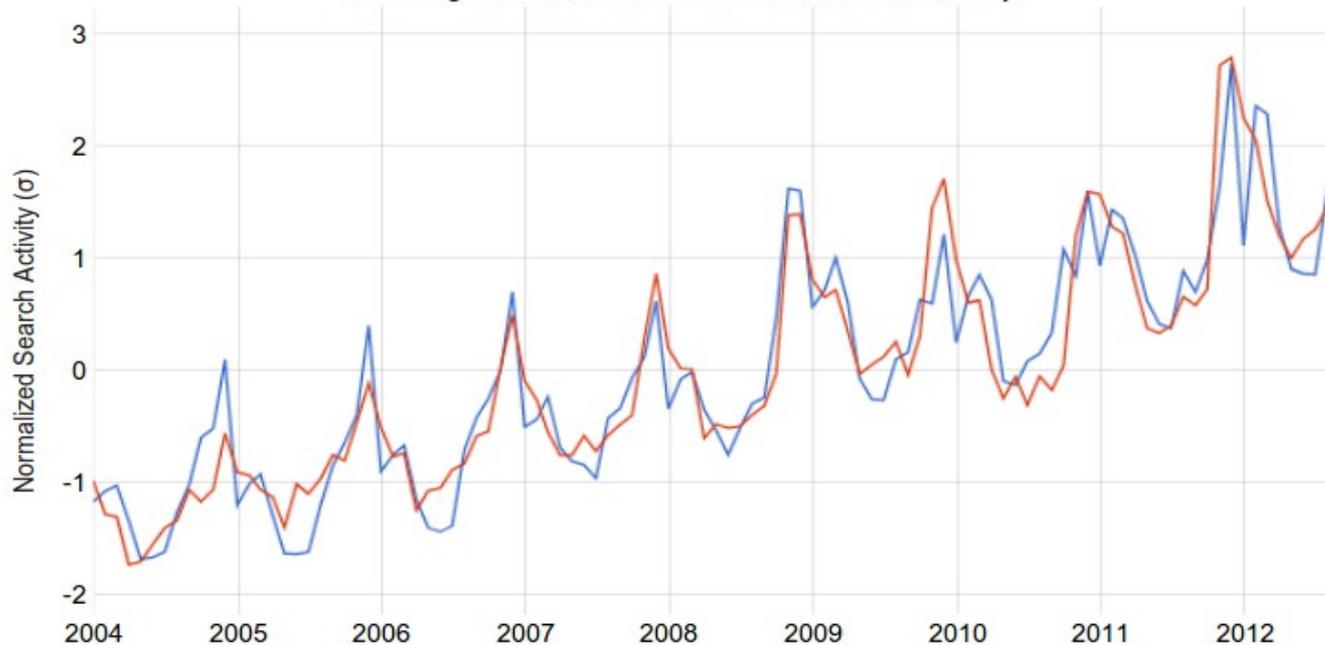
[stack on] has highest correlation  
[gun shops] is chosen by BSTS

User uploaded activity for **FBI NICS data** and United States Web Search activity for **stack on** ( $r=0.9356$ )

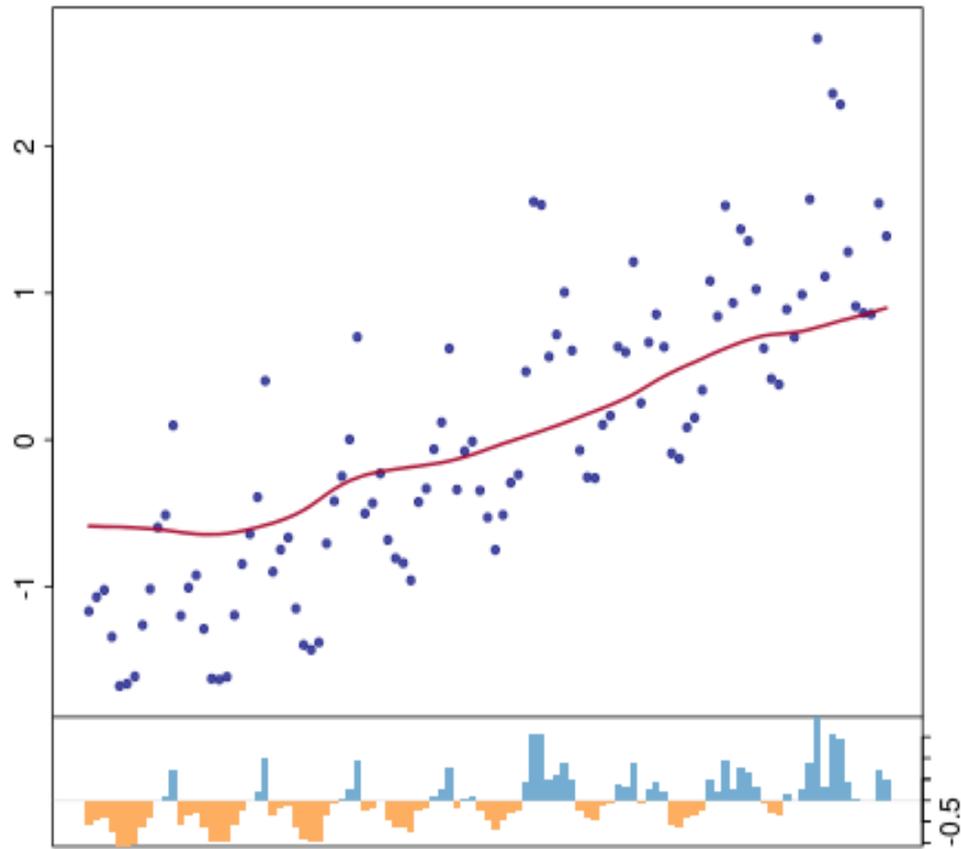
Line chart Scatter plot

— FBI NICS data — stack on

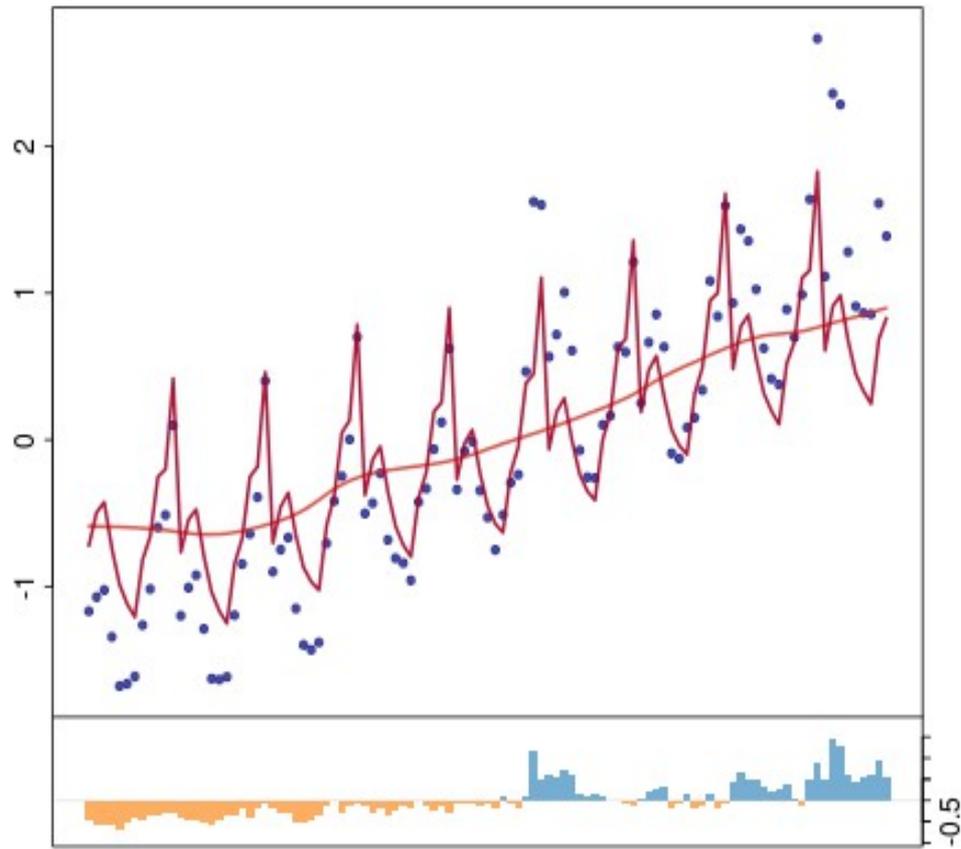
Hint: Drag to Zoom, and then correlate over that time only.



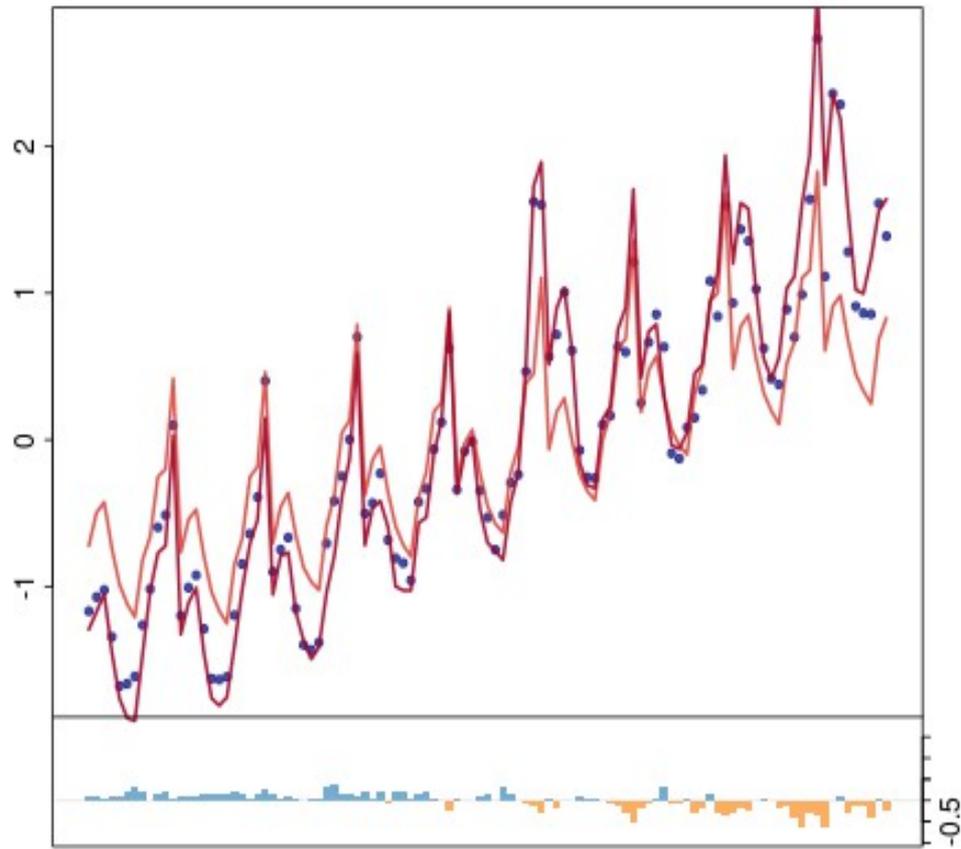
1. trend (mae=0.49947)



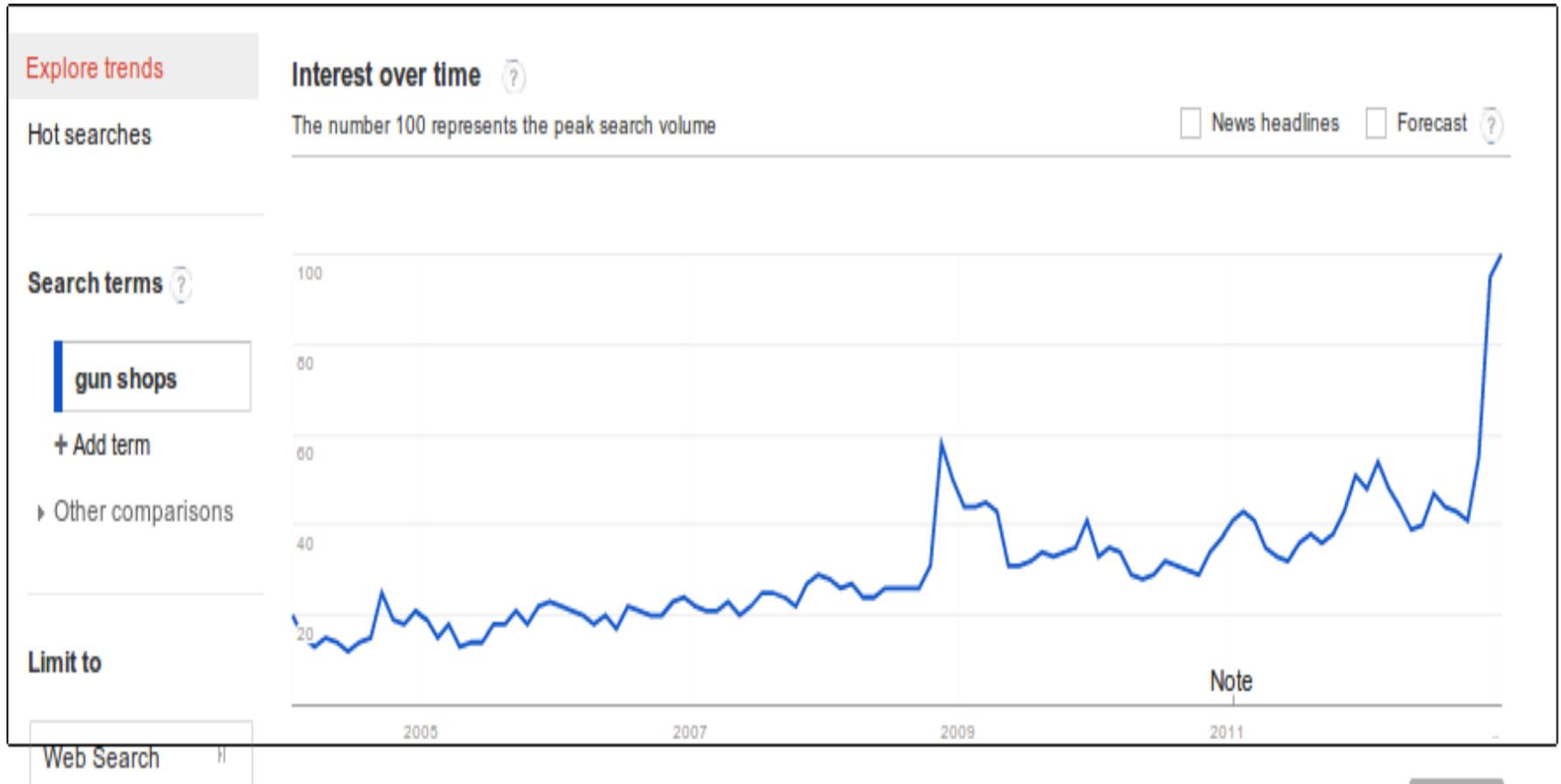
2. add seasonal (mae=0.33654)



3. add gun.shops (mae=0.15333)



# Searches on [gun shop]



## **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

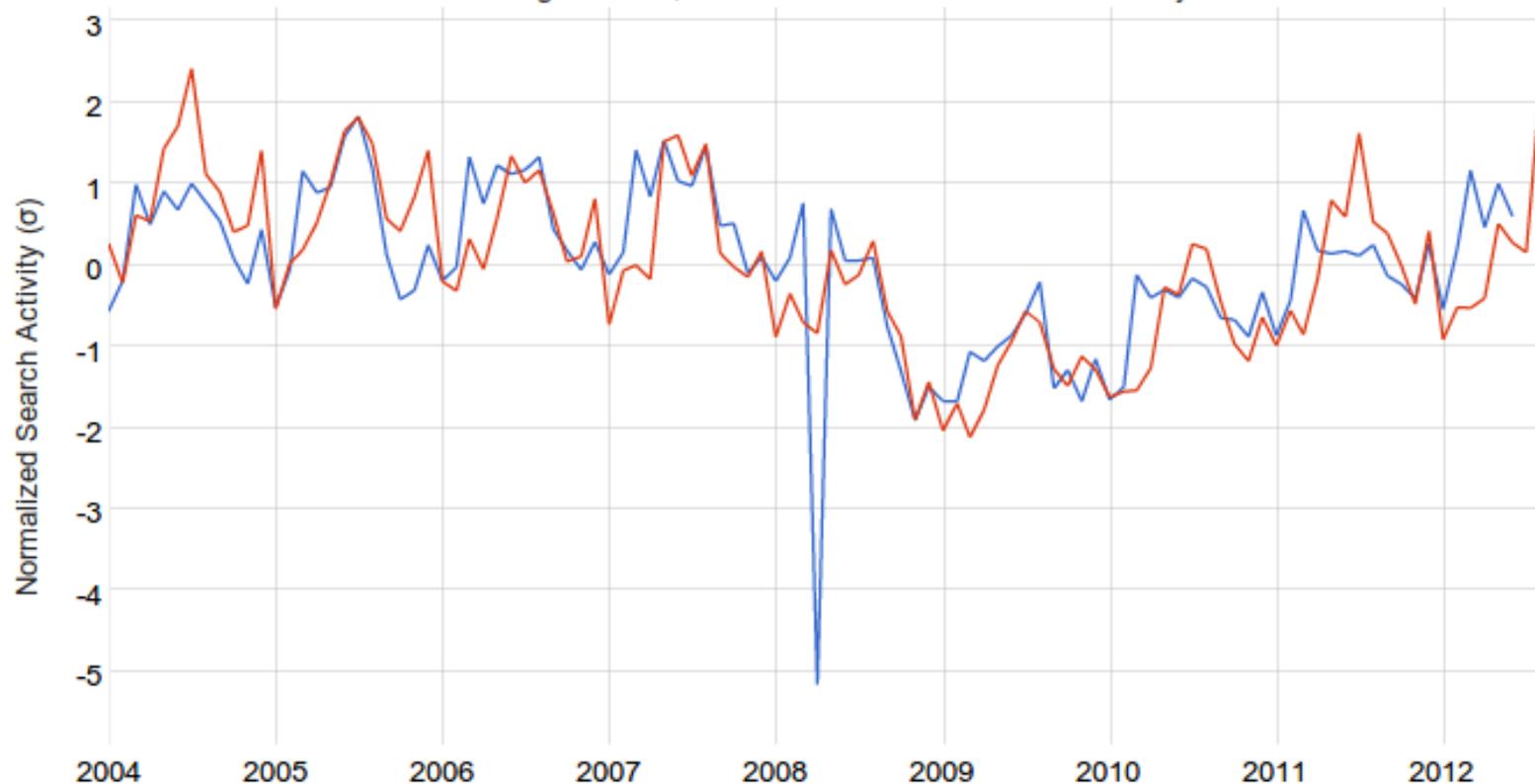


User uploaded activity for **US Auto Sales NSA** and United States Web Search activity for **Indian restaurants**  
( $r=0.7195$ )

 Line chart  Scatter plot

— US Auto Sales NSA — Indian restaurants

Hint: Drag to Zoom, and then correlate over that time only.



# Challenge 2: Fat regression

Slim regression

$$\mathbf{y} = \mathbf{bX}$$

Fat regression

$$\mathbf{y} = \mathbf{bX}$$

Any square subset of regressors will fit perfectly

$$\mathbf{y} = \mathbf{b}_1\mathbf{X}_1 \quad \mathbf{b}_2\mathbf{X}_2$$

A subset of regressors might fit well by chance

$$\mathbf{y} = \mathbf{b}_1\mathbf{X}_1$$

## 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

**Kalman filter:** handles  
seasonality and trend

**Spike and slab:** handles  
variable selection

**Model averaging:**  
averages over many  
small models to avoid  
overfitting

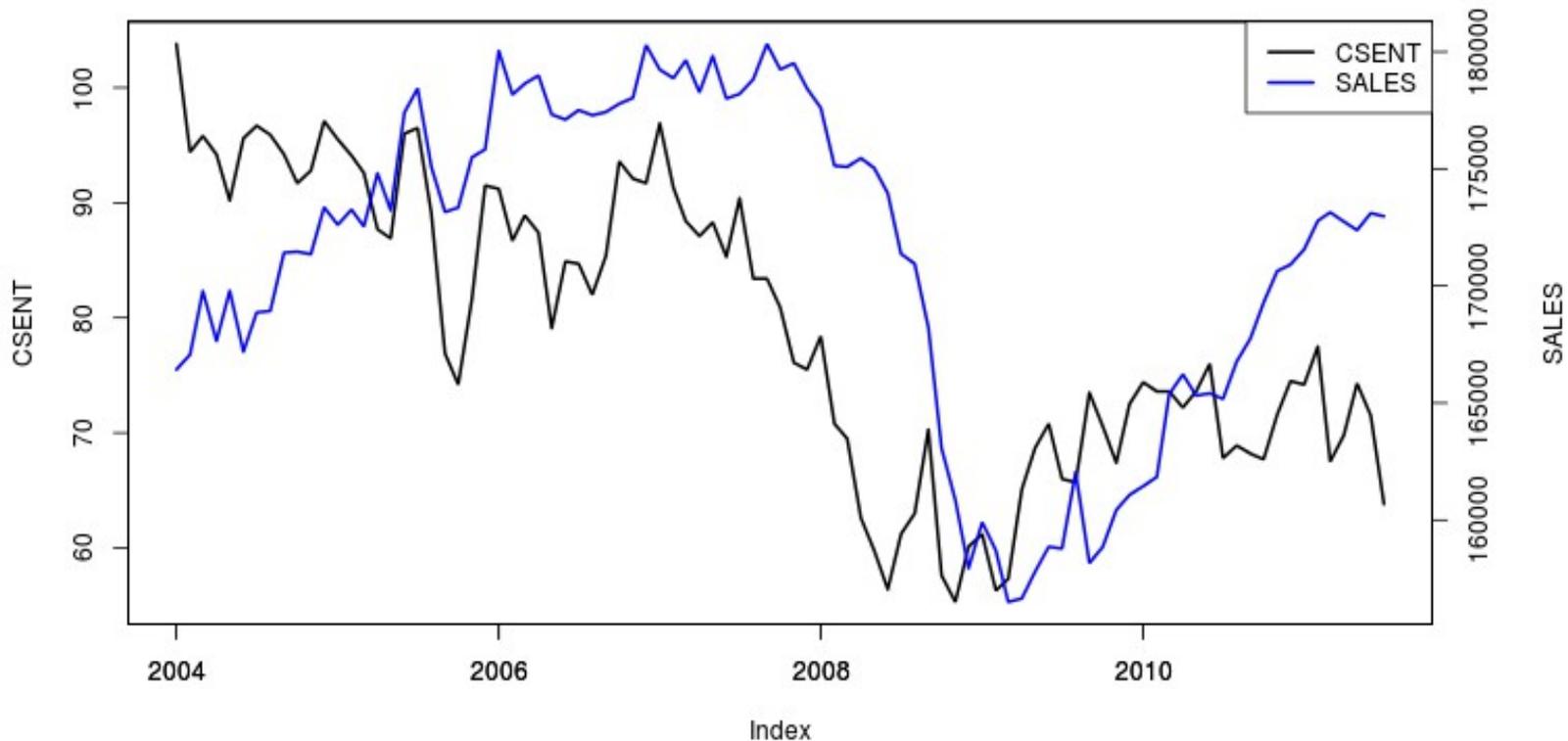


# UM consumer sentiment index

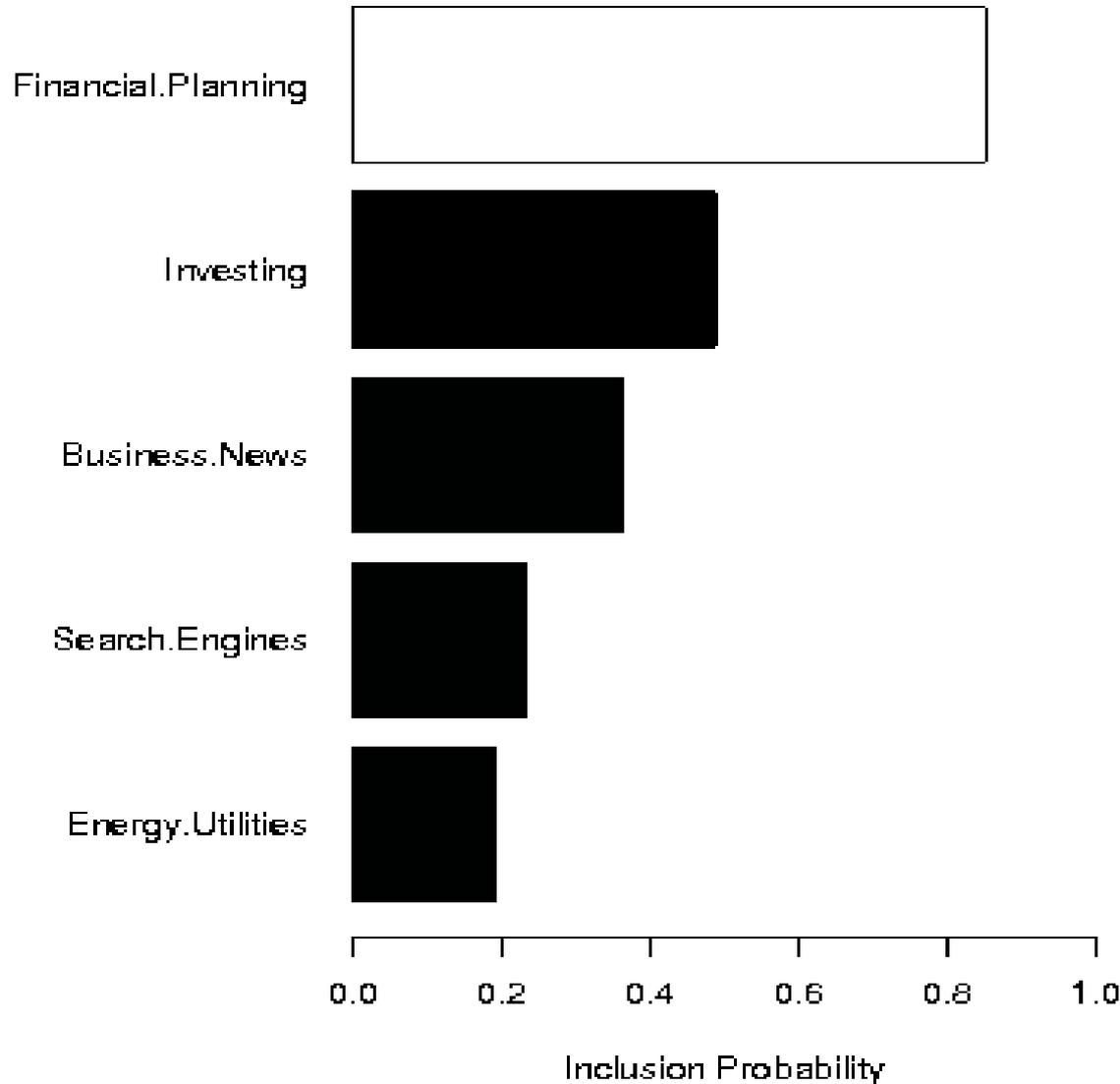


Monthly data from University of Michigan survey

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”

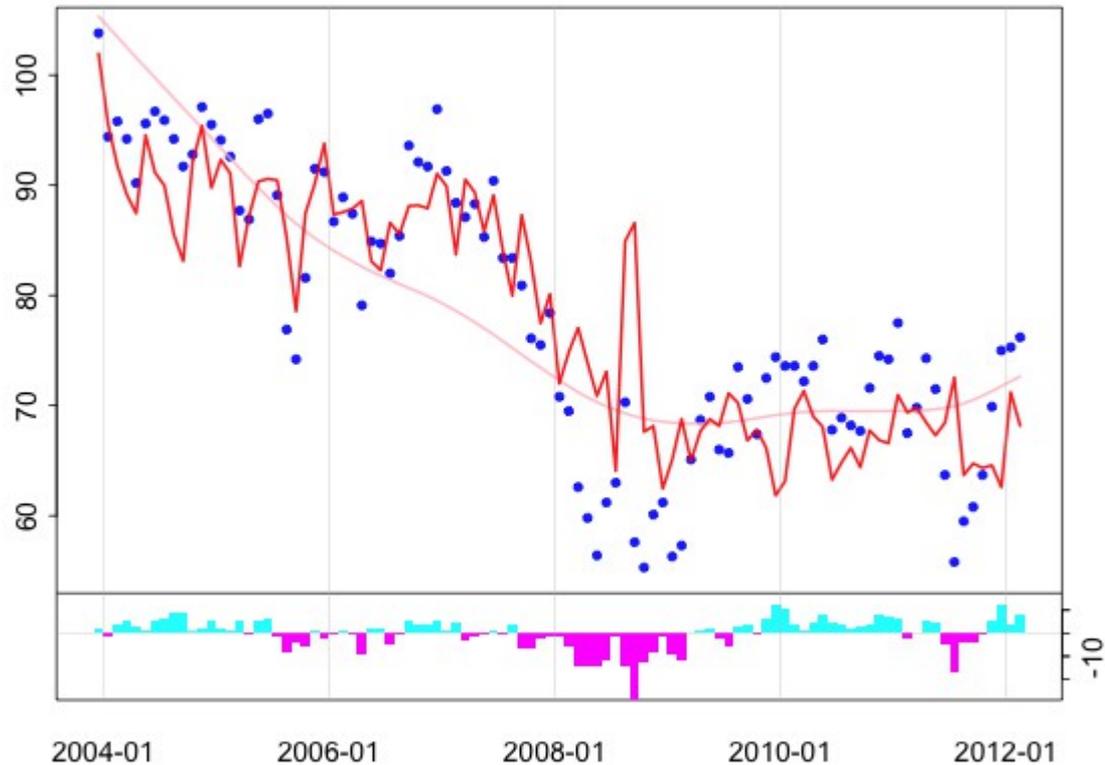
1. trend (mae=5.7134)



# Add “financial planning”

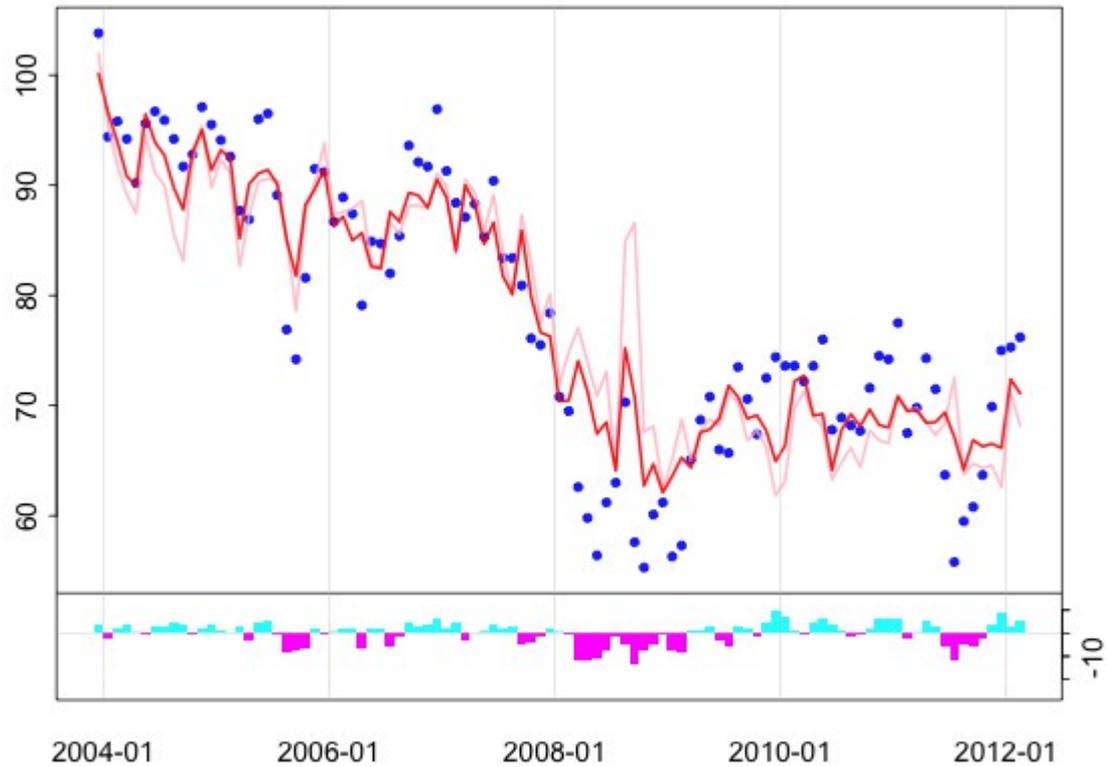


## 2. add Financial.Planning (mae=4.9965)



# Add "Investing"

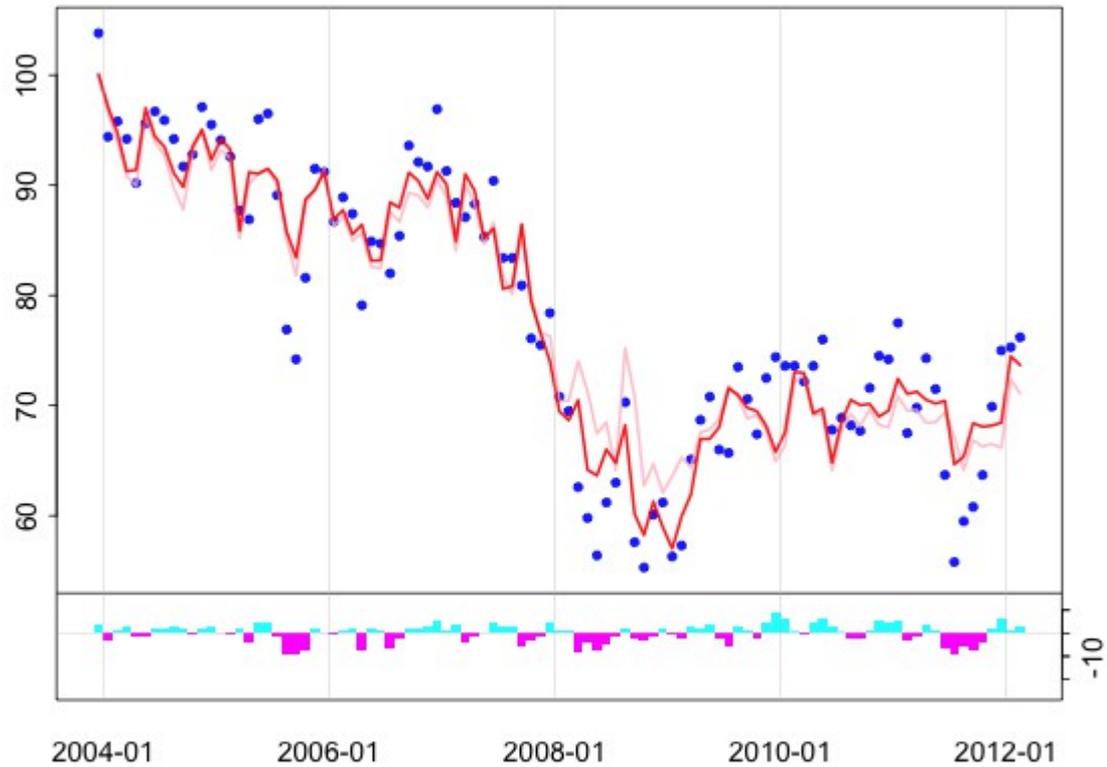
3. add Investing (mae=3.8372)



# Add "Business News"



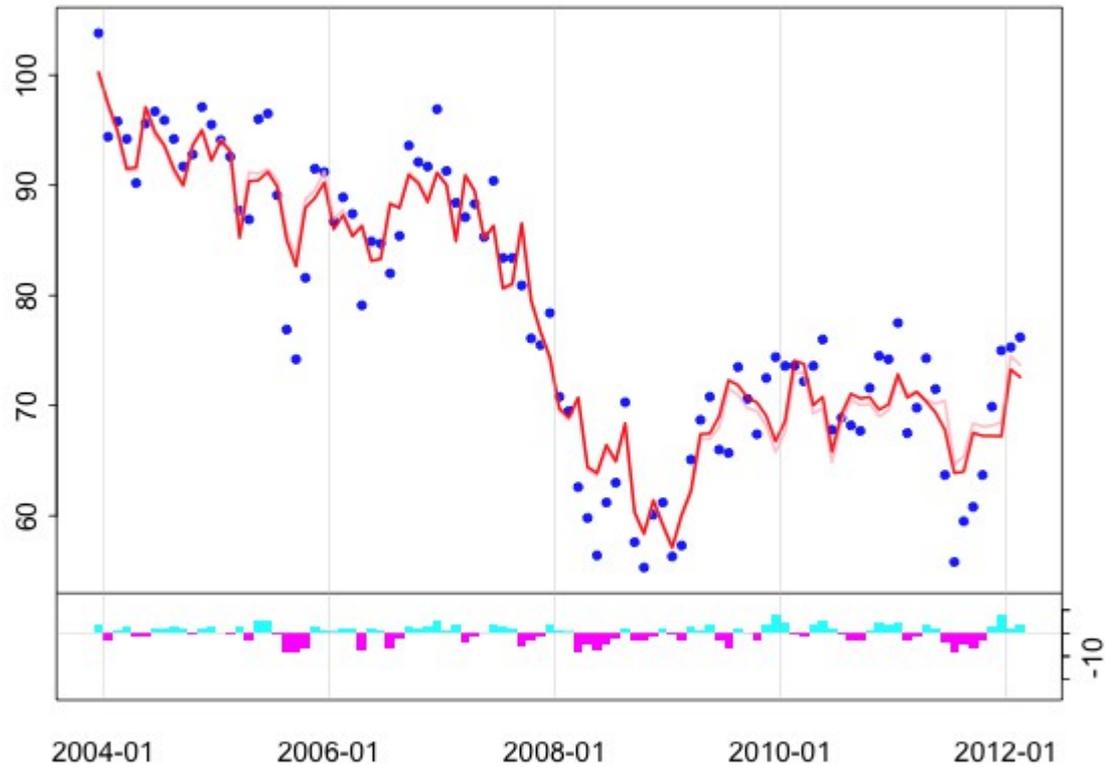
## 4. add Business.News (mae=3.2226)



# Add "Search Engines"

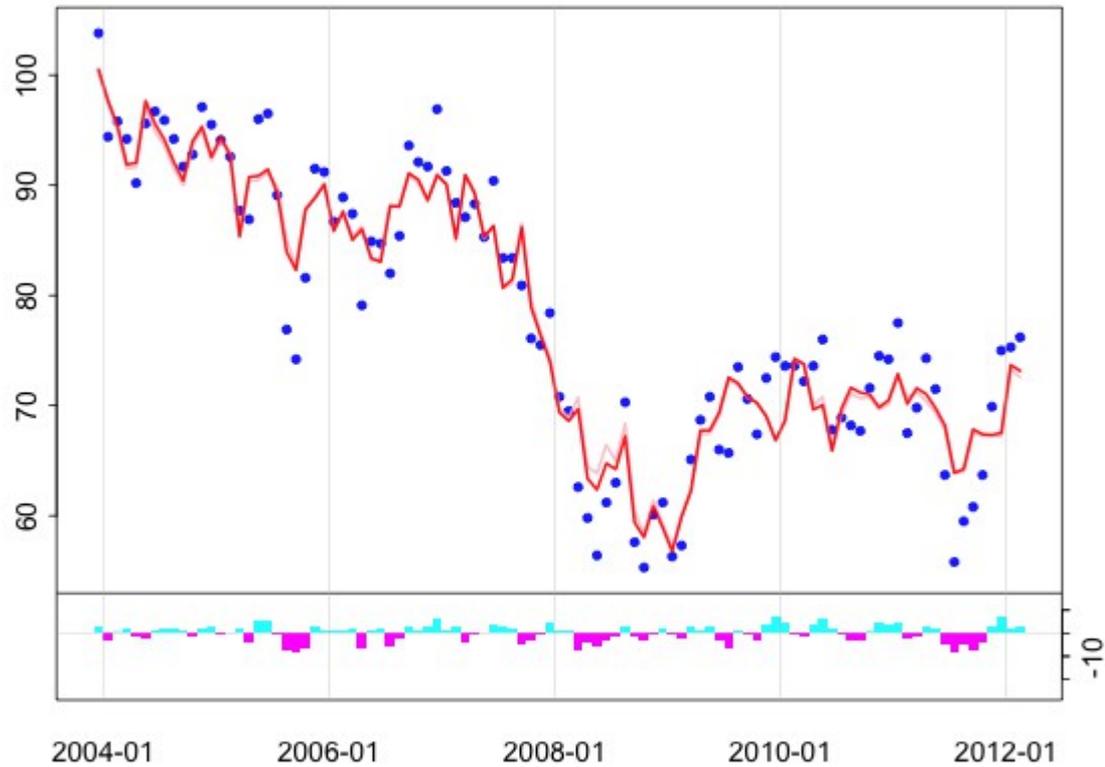


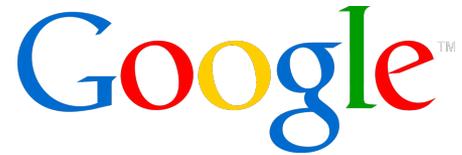
5. add Search.Engines (mae=3.1455)



# Add Energy and Utilities

6. add Energy.Utilities (mae=3.0068)





Google Trends

Google Correlate

Google Consumer  
Surveys



## How it works



1  
You create online surveys to gain consumer insight



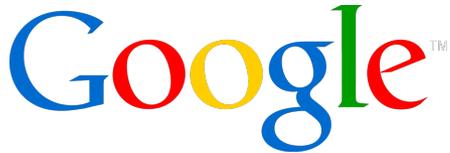
2  
People complete questions to access premium content



3  
Publishers get paid as their visitors answer



4  
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 | [Twitter](#) [Facebook](#) [LinkedIn](#) [Google+](#) [Comments](#) 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:*

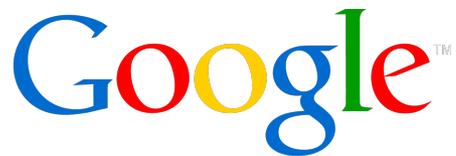
Have you ever purchased anything from (check all that apply):

*Check all answers that apply*

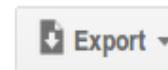
- An email newsletter or ad
- A YouTube video
- An ad on your mobile phone
- An ad on your tablet
- None of the above

Submit answer(s)

[Show me another question](#)



Screening question from Privacy concerns



**Report** Custom Insights <sup>2</sup>

Inferred Gender

Sum	Compare
Male	Female

Inferred Age

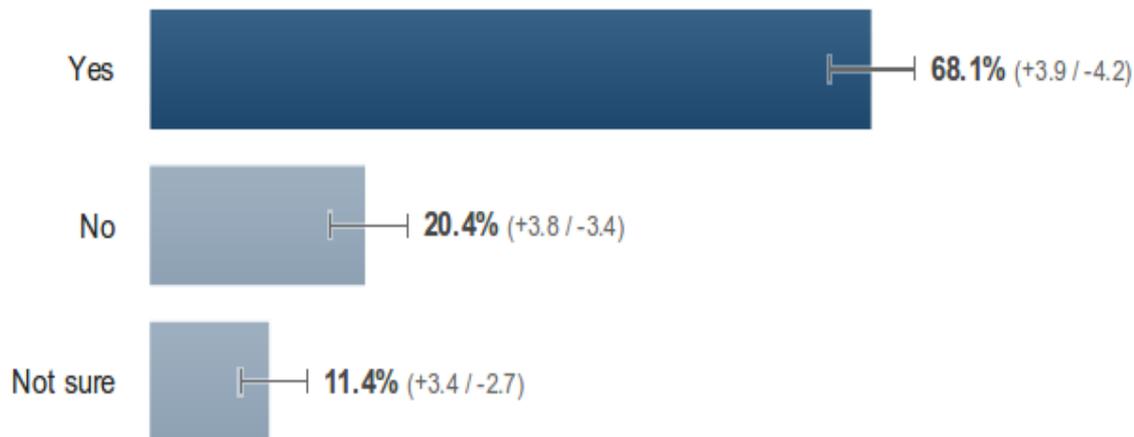
Sum	Compare
18-24	25-34
35-44	45-54
55-64	65+

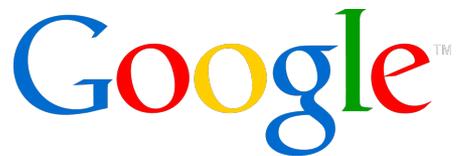
Geography

SINGLE ANSWER

## Are you concerned about your privacy online?

Results for respondents with demographics. Weighted by Age, Gender, Region. (575 responses) <sup>?</sup>  
Order statistically significant. <sup>?</sup>





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

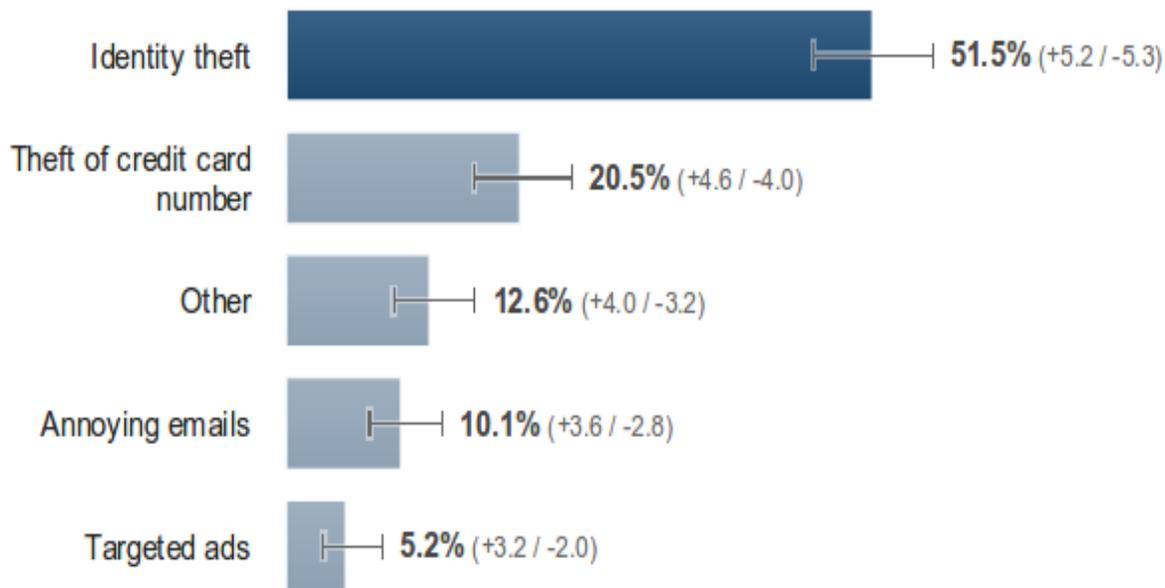
All of the USA

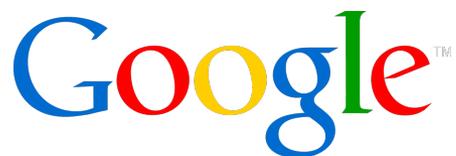
SINGLE ANSWER

## What aspect of online privacy concerns you the most?

Results for respondents with demographics. Weighted by Age, Gender. (383 responses) ?

Winner statistically significant. ?





### Pollster Accuracy and Bias, 2012 Presidential Election

Likely Voters Polls in Last 21 Days of Campaign

Minimum 5 Polls

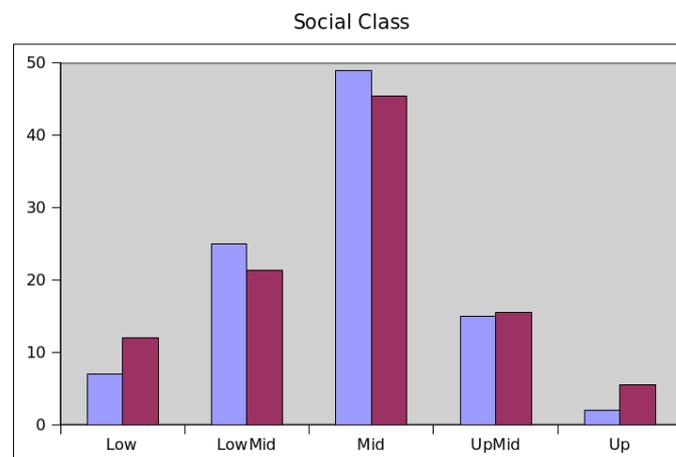
Pollster	# Polls	Avg. Error	Bias	Mode	Cell?
IBD / TIPP	11	0.9	R +0.1	Live Phone	Yes
Google Consumer Surveys	12	1.6	R +1.0	Internet	N/A
Mellman	9	1.6	R +0.0	Live Phone	Yes
RAND Corporation	17	1.8	D +1.5	Internet	N/A
CNN / Opinion Research	10	1.9	R +0.6	Live Phone	Yes
Ipsos / Reuters (online)	42	1.9	R +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.

### 3<sup>rd</sup> 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

# Google™ 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



WAL\*MART



FedEx



TARGET

## 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