A Hands on Guide to Google Data

Hal Varian
Seth Stephens-Davidowitz
Google,Inc.

March 7, 2015

Three Google tools for social science research

Google Correlate Shows the queries that are most correlated with cross-section (state) or time series (weekly/monthly) data.

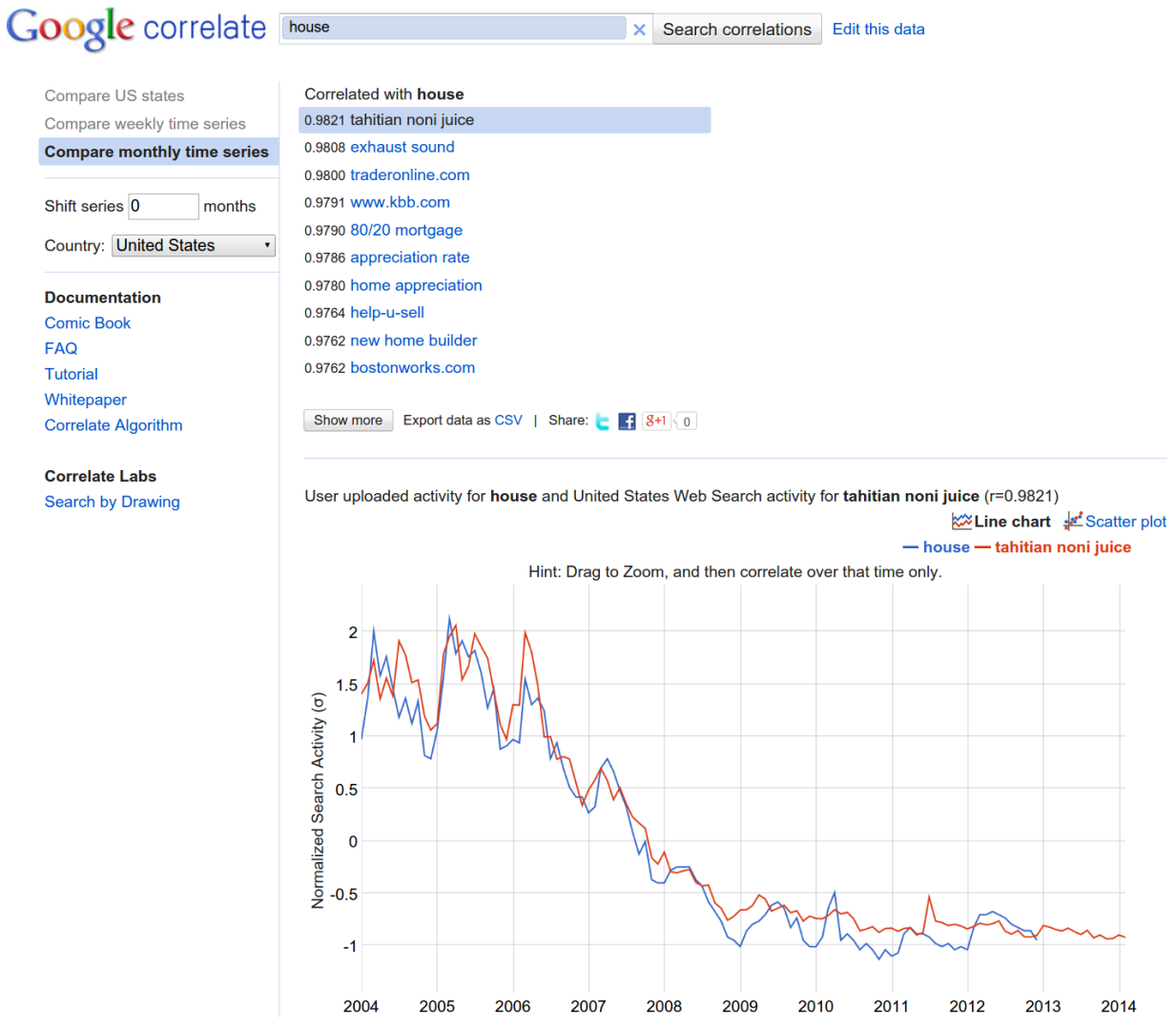
Google Trends Shows an index of activity for specific queries or categories of queries.

Google Consumer Survey Lightweight, quick and inexpensive surveys of internet users.

Correlate: predict house sales over time

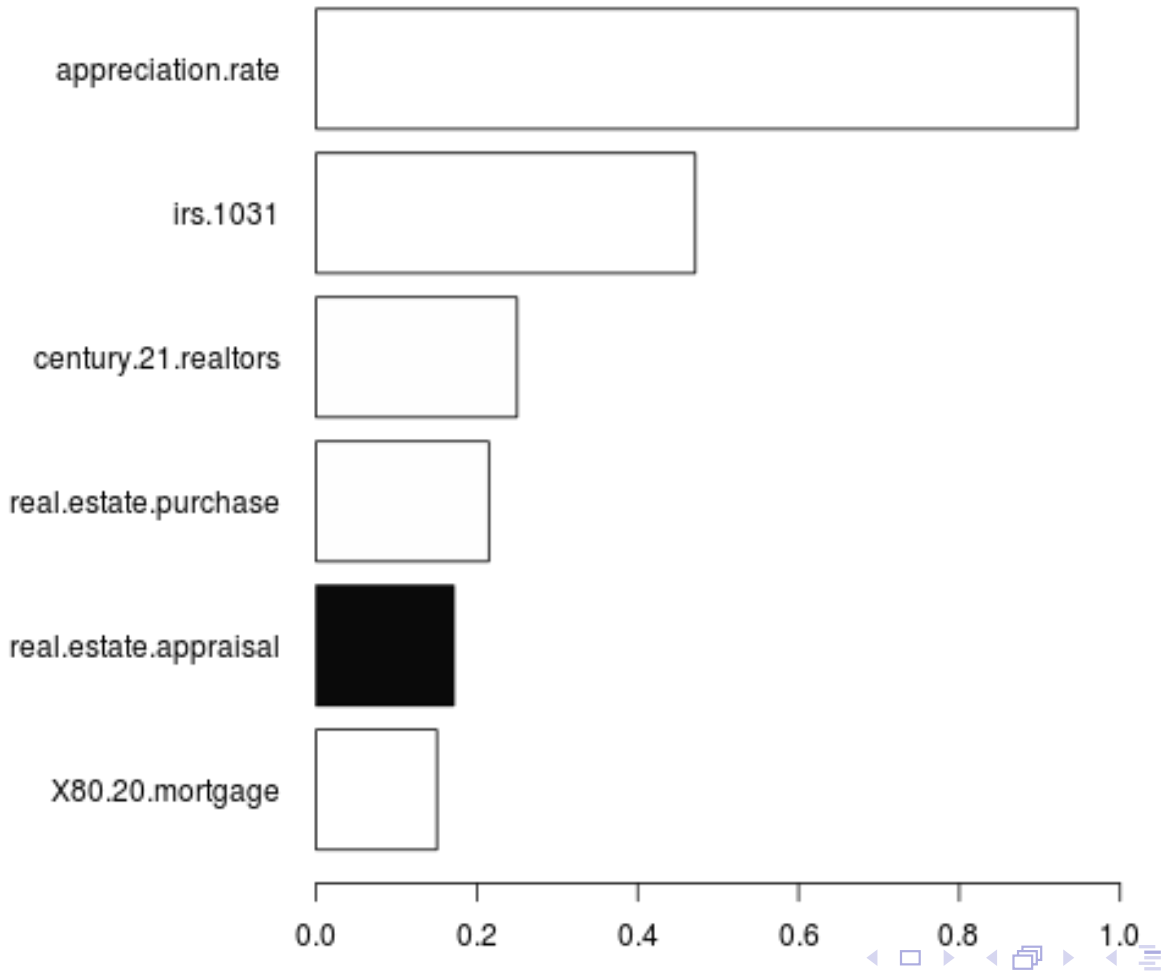
- Go to FRED
- Download "New One Family Houses Sold" from 2004 to present
- Clean up CSV file, upload to Google Correlate
- Examine plots
- Export data from Correlate as CSV file

Google Correlate screenshot



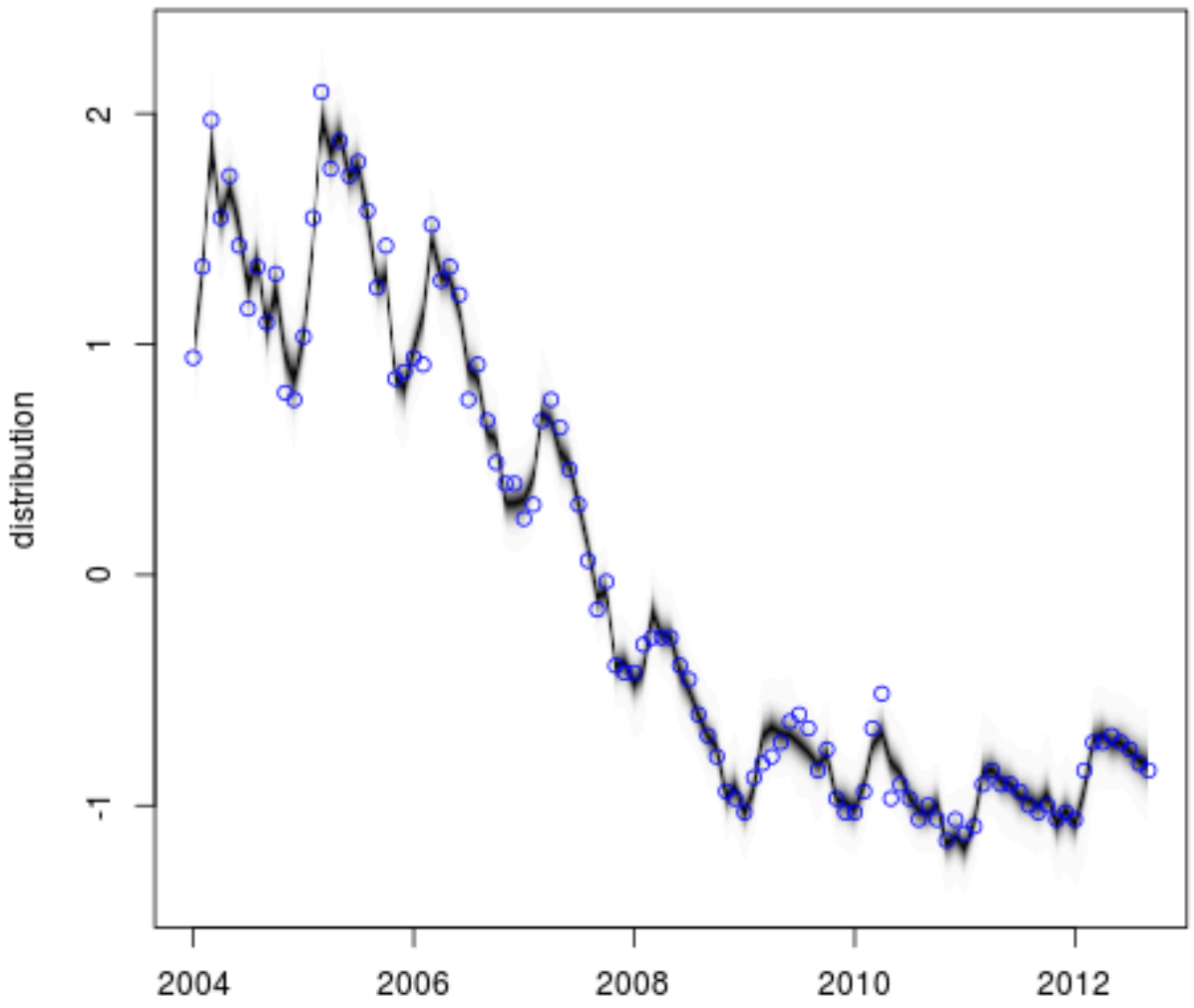
Find best predictors using BSTS

White: positive Black: negative



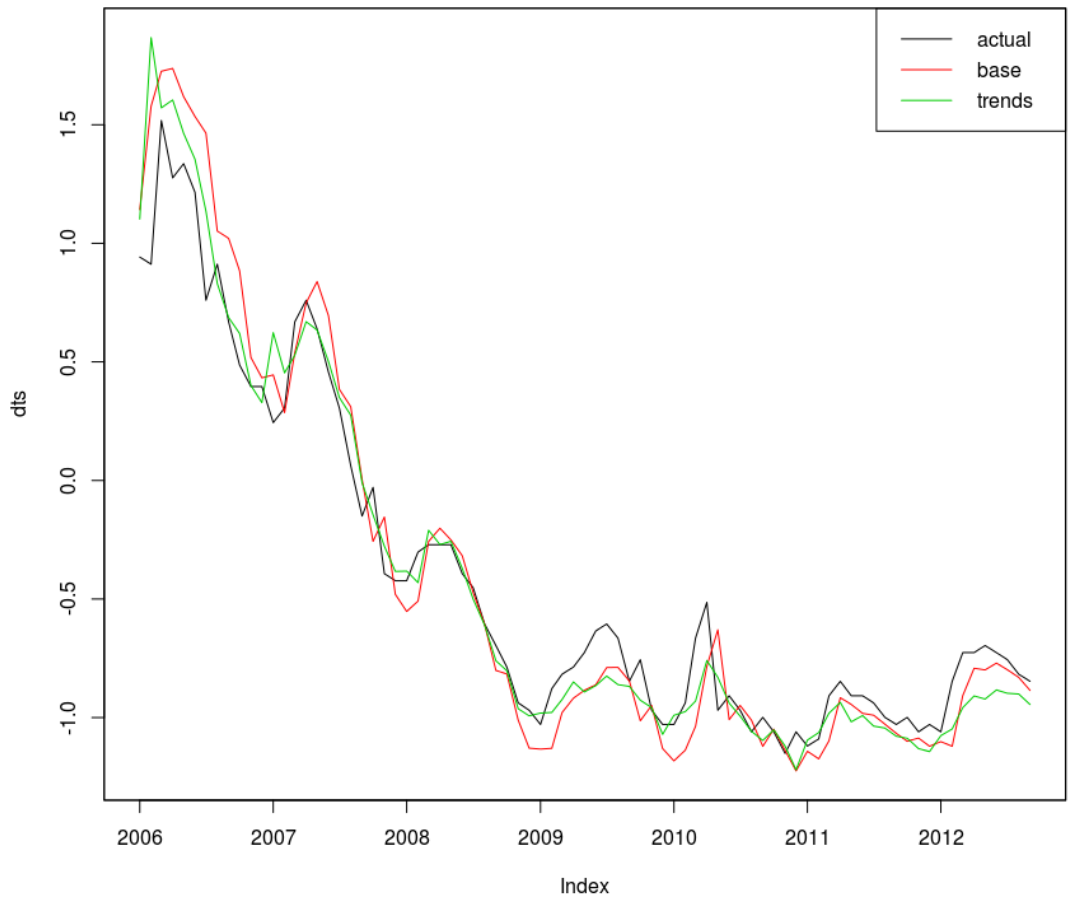
In-sample fit with BSTS

Circles: actual Black: predicted



Out-of-sample prediction using regression

- AR1 regression with contemporaneous query data
  $y_t = y_{t-1} + x_t + e_t$ one-step ahead prediction
- Get 23% improvement in MAPE compared to pure AR1 model

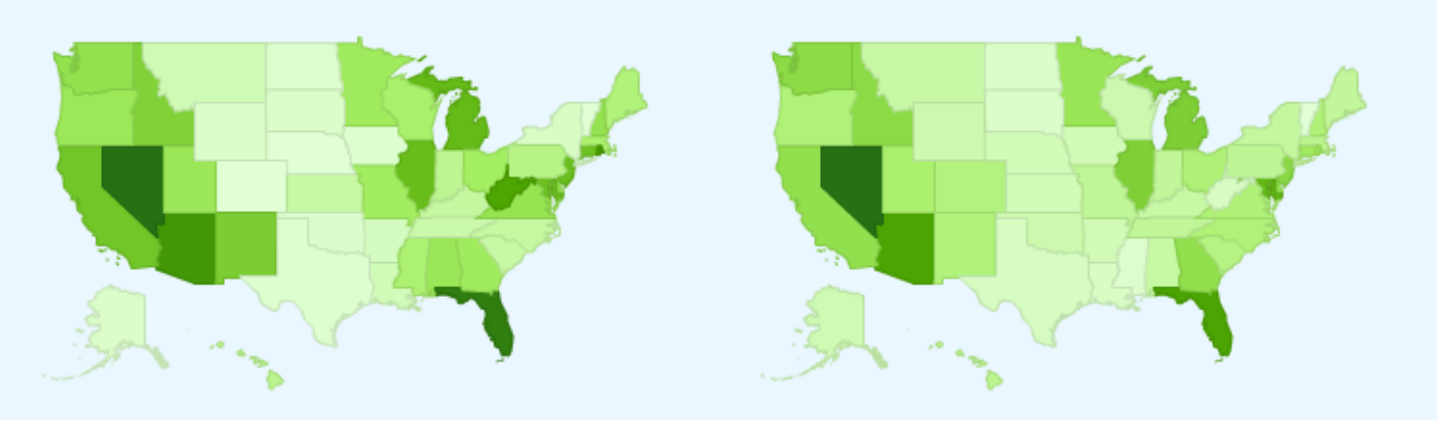


House price declines by state (cross section)

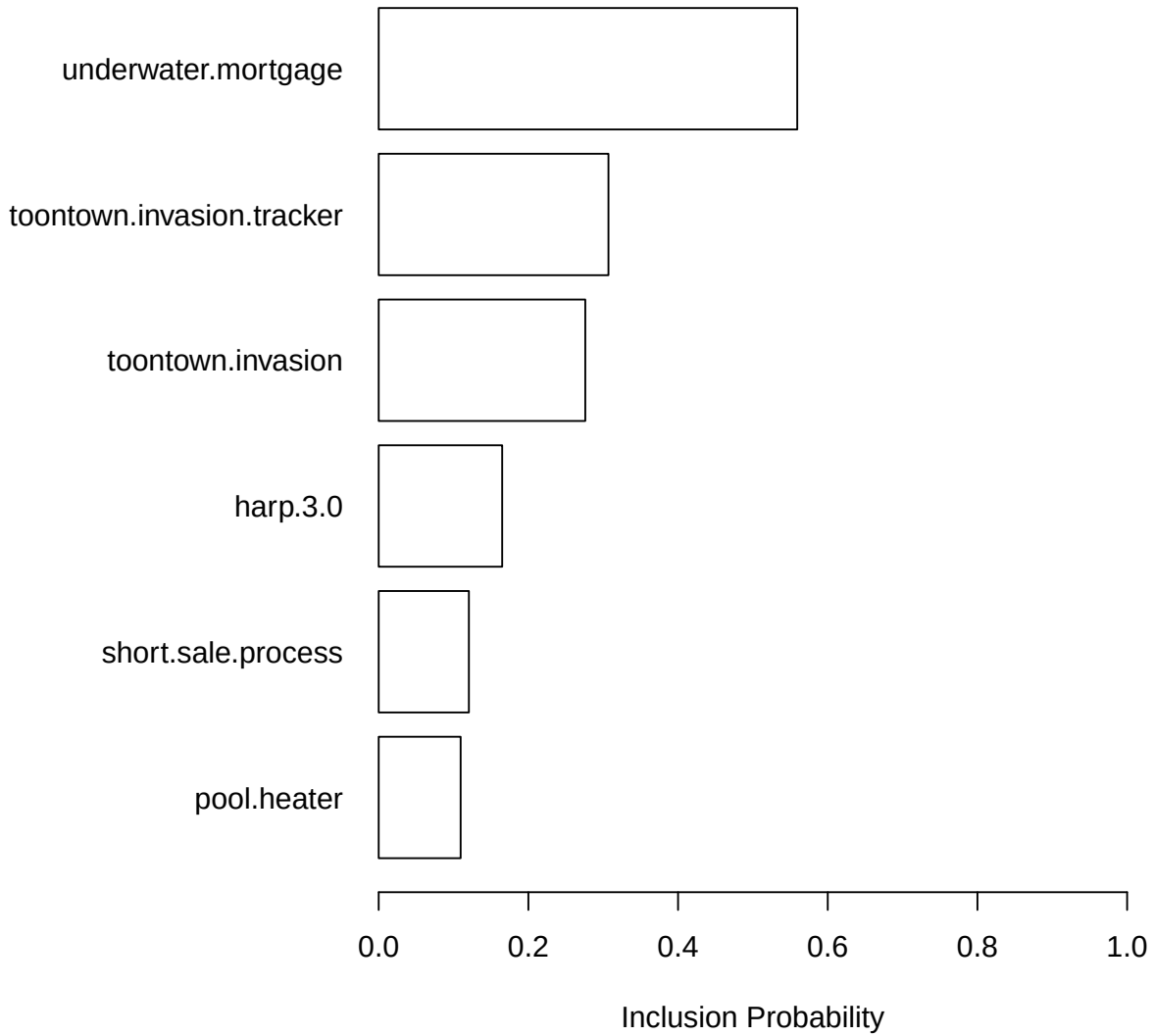
- Use Google Correlate as before, but now by state
- Here is comparison of price declines and query [short sale process]

User uploaded activity for **housing-decline** and United States Web Search activity for **short sale process** (r=0.7888)

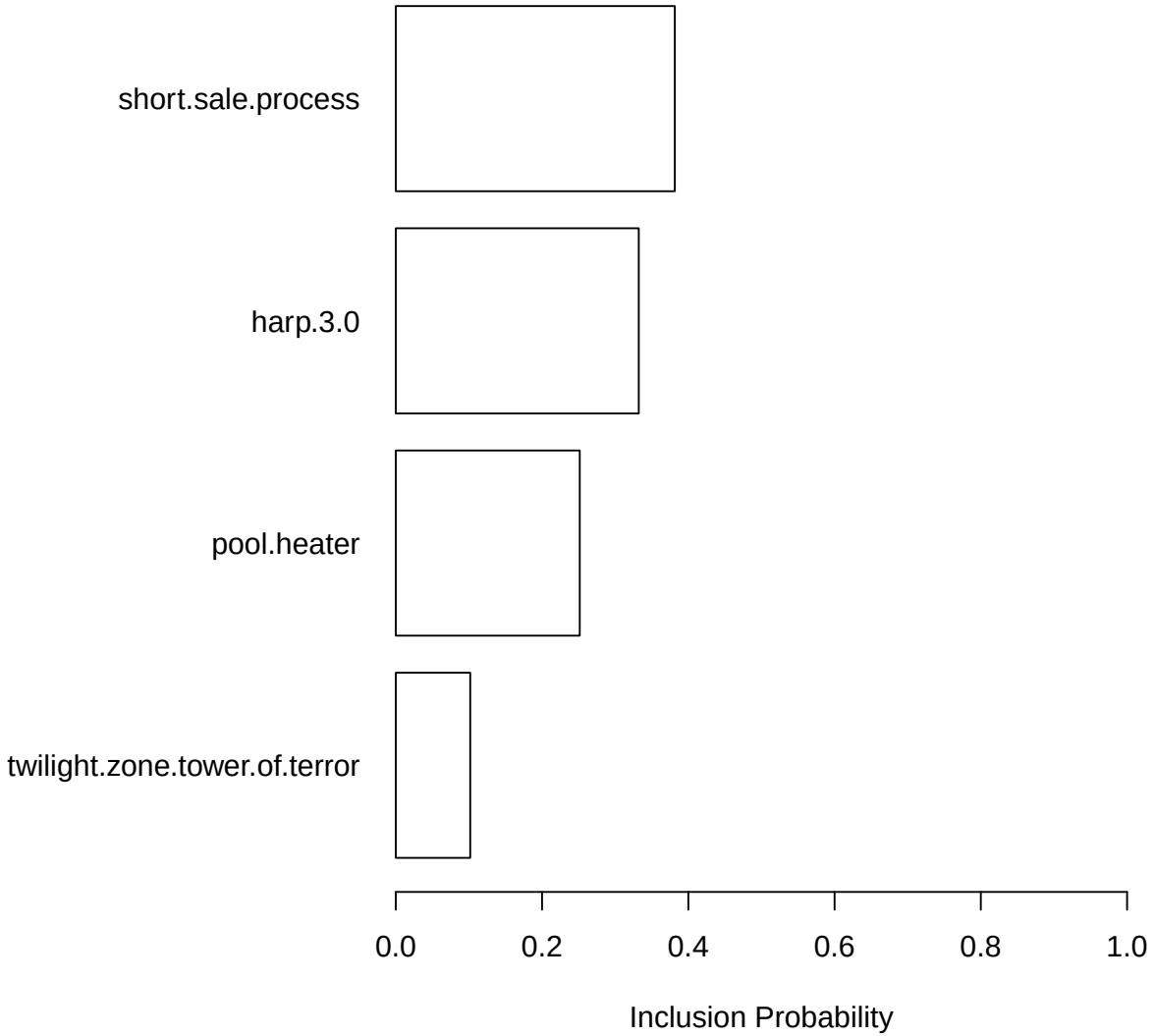
State maps Scatter plot



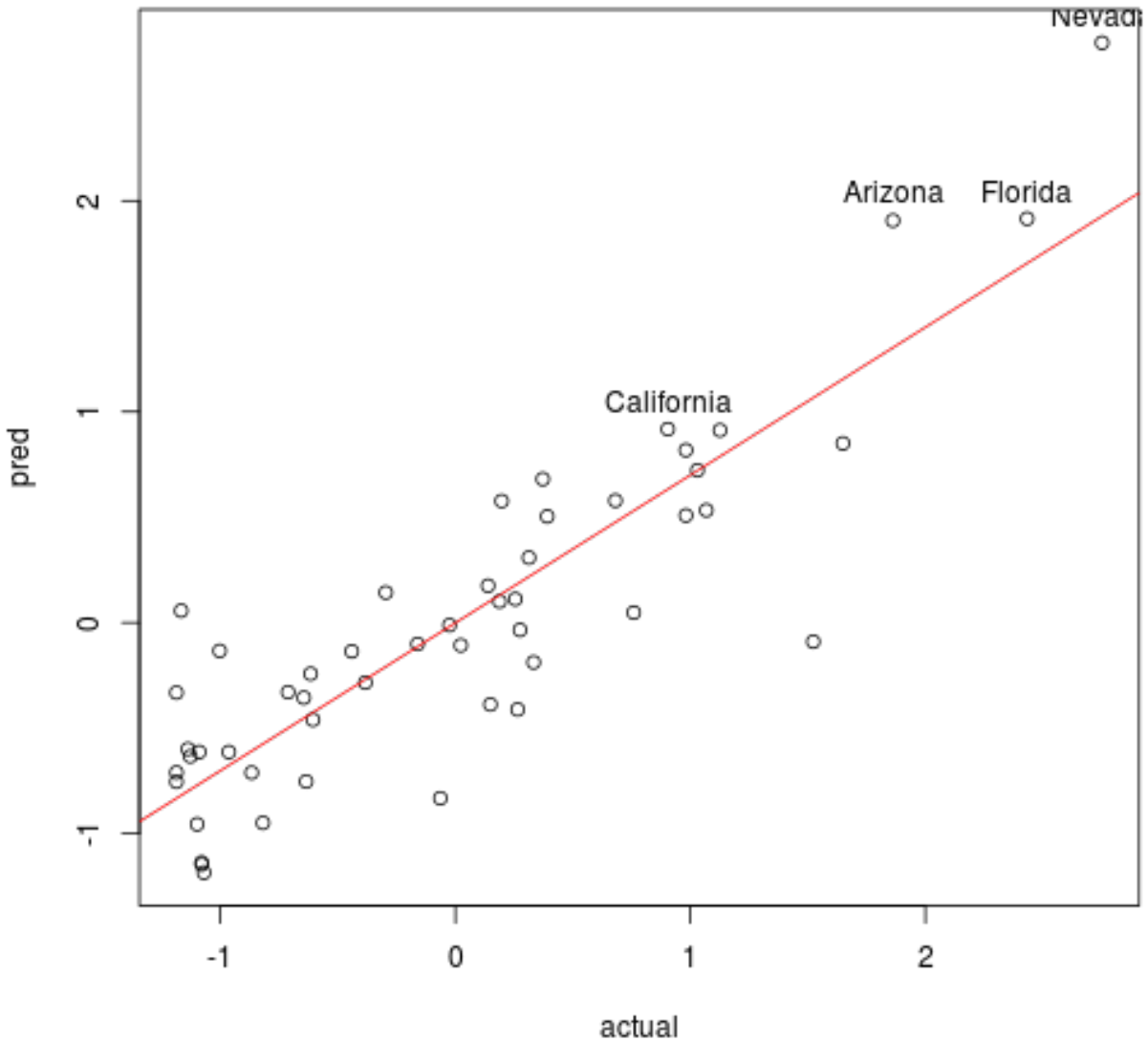
Find predictors using spike-slab



Apply some judgment



Compare actual to fitted across states



What queries are predictive of short lifespan?

negative life expectancy | Search correlations | Edit this data

Compare US states
Compare weekly time series
Compare monthly time series

Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

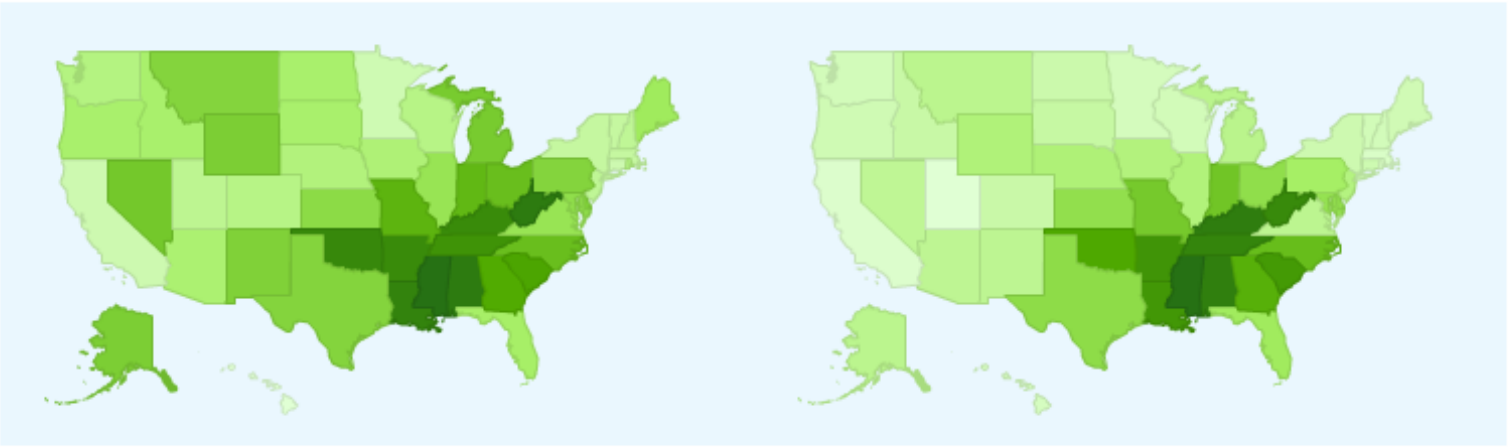
Correlated with **negative life expectancy**

0.9092 blood pressure medicine
0.8985 obama a
0.8978 major payne
0.8975 against obama
0.8936 king james bible online
0.8935 about obama
0.8928 prescription medicine
0.8920 40 caliber
0.8919 .38 revolver
0.8916 reprobate
0.8911 performance track
0.8910 lost books of the bible
0.8905 glock 40 cal
0.8898 lost books
0.8896 the mark of the beast
0.8892 obama says
0.8891 obama said
0.8882 sodom and
0.8882 the antichrist
0.8865 globe life
0.8858 the judge
0.8834 hair pics
0.8833 medicine side effects
0.8829 momma
0.8828 james david
0.8823 flexeril

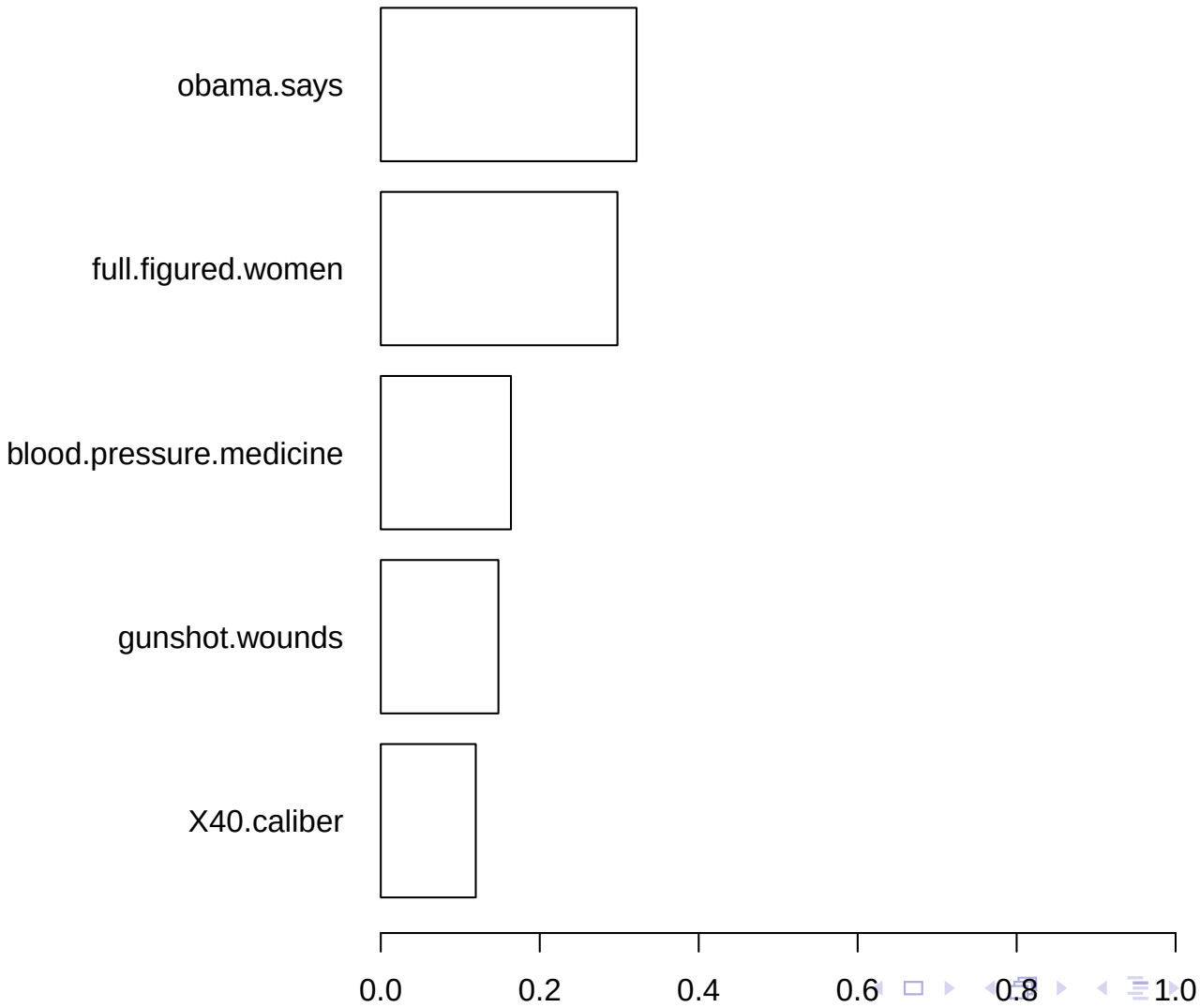
life expectancy vs [blood pressure medication]

User uploaded activity for **negative life expectancy** and United States Web Search activity for **blood pressure medicine** (r=0.9092)

State maps Scatter plot



Predictors of short lifespan?

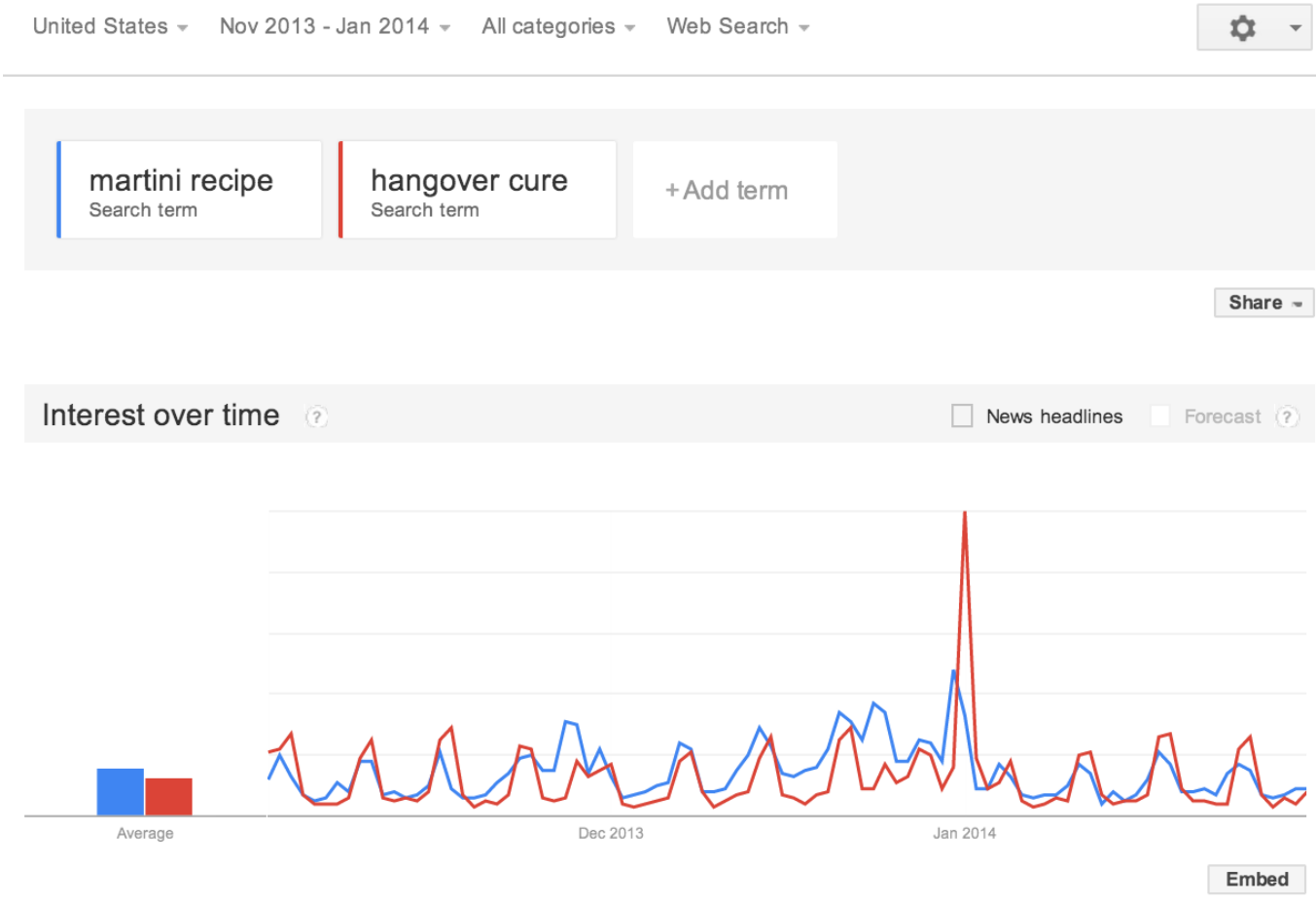


Question about hangovers

What day of the week has the most queries for [hangover]?

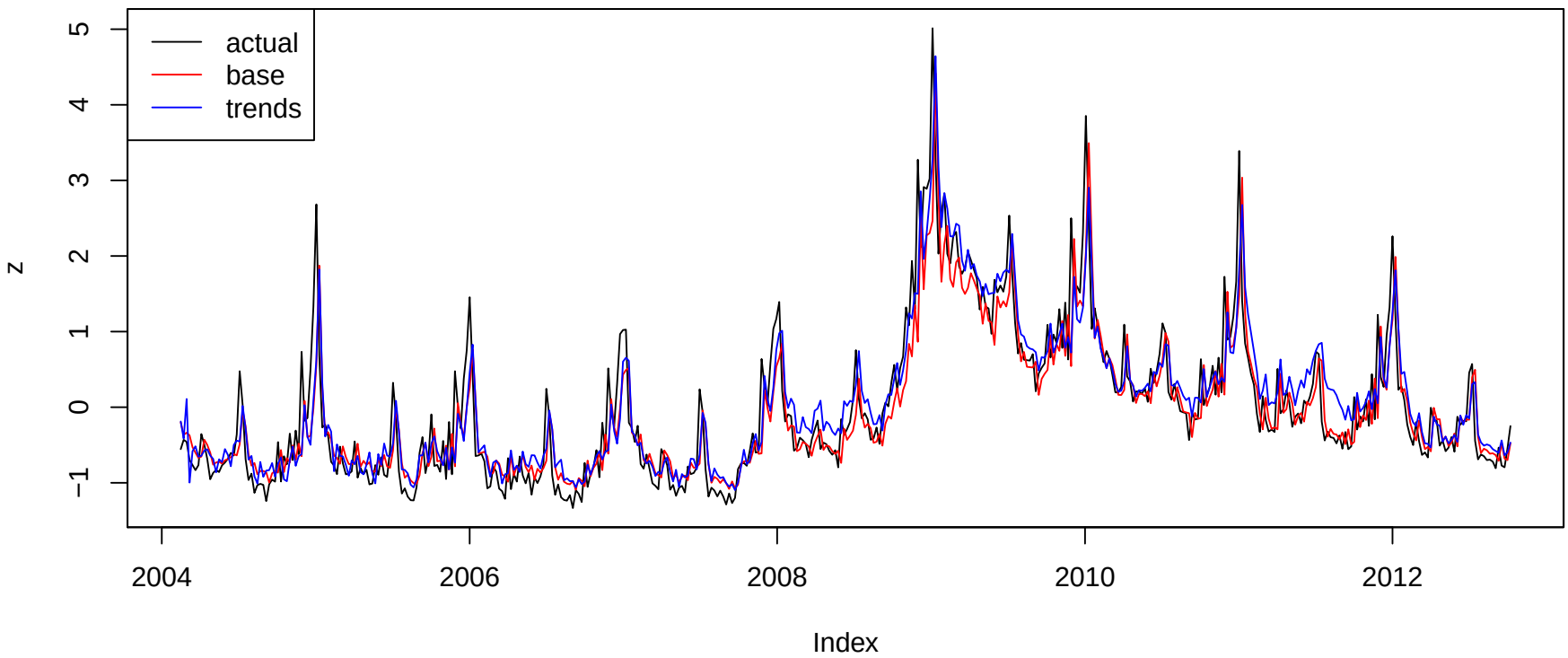
- Monday
- Tuesday
- Wednesday
- Thursday
- Friday
- Saturday
- Sunday

Google Trends: queries for [hangover]



Initial claims for Unemployment Benefits

- One-week ahead forecast using query [filing for unemployment]
- Baseline is AR(1) model



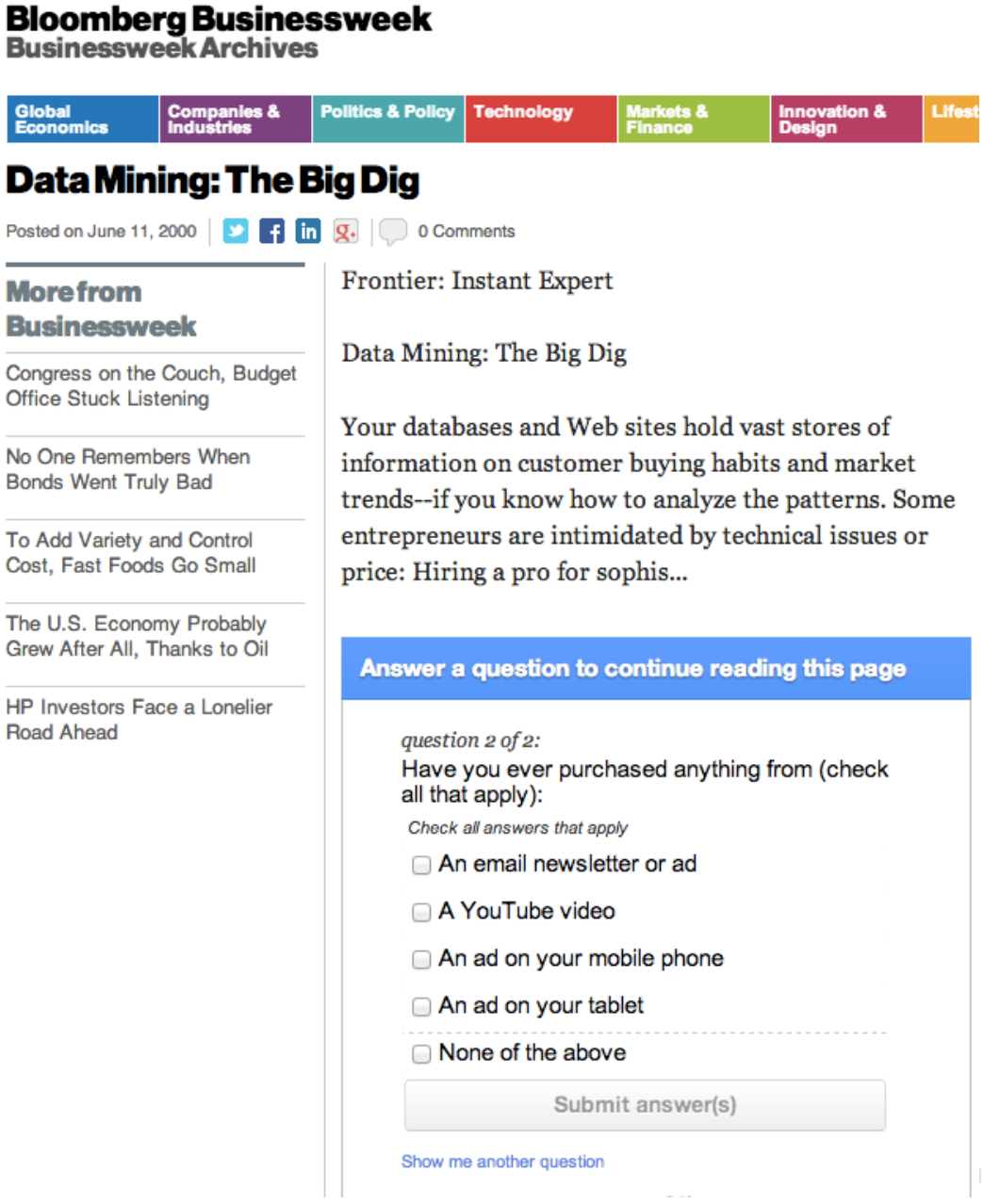
How does unemployment insurance affect job search?

- Scott Baker and Andrey Fradkin, SSRN 2251548
- Measure job search intensity using Google searches for [jobs -steve]
- Individuals on UI search 30% less than unemployed not on UI
- Searches close to UI exhaustion search twice as much as those with 30 weeks remaining
- Decrease in job search due to UI extensions had were small

How many votes did Obama lose due to racism?

- Looks at Democratic votes in 2004 compared to 2008
- Looks at racist queries by location using Google Trends
- "The higher the racially charged search rate in an area, the worse Mr. Obama did."
- Racism cost Obama about 4 percentage points of popular vote

Google Consumer Surveys



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

Example surveys

- If you were asked to use one of these commonly used names for social classes, which would you say you belong in?
- Do you support Obama or Romney in the upcoming election?
- I prefer to buy products that are assembled in America. [Agree or disagree]

Social class Pew Foundation and GCS

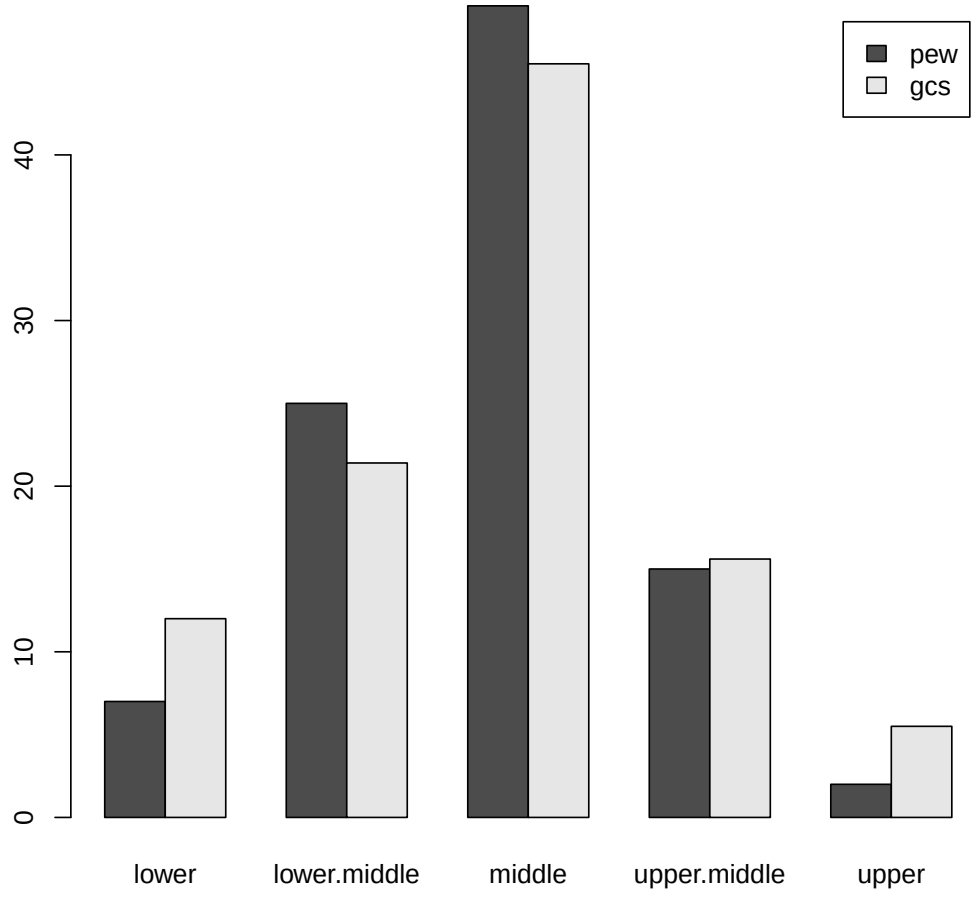
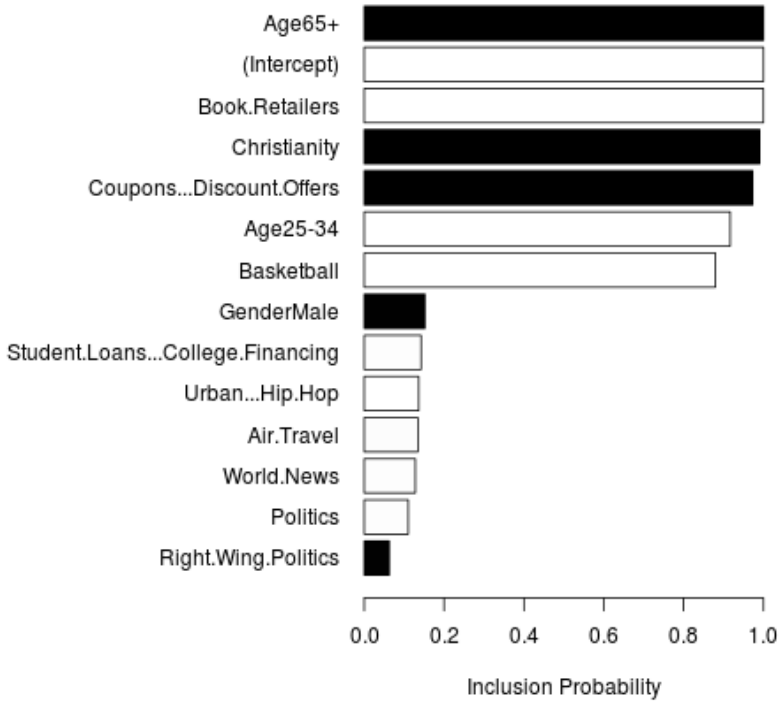


Figure : Comparing Pew and GCS answers to social class question.

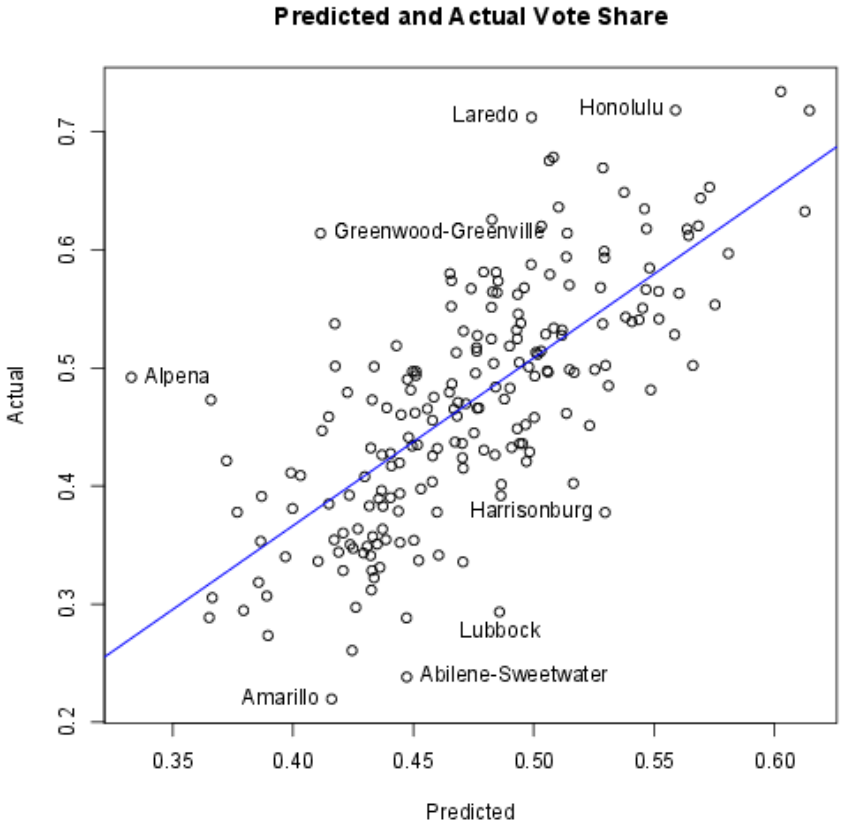
Survey amplification

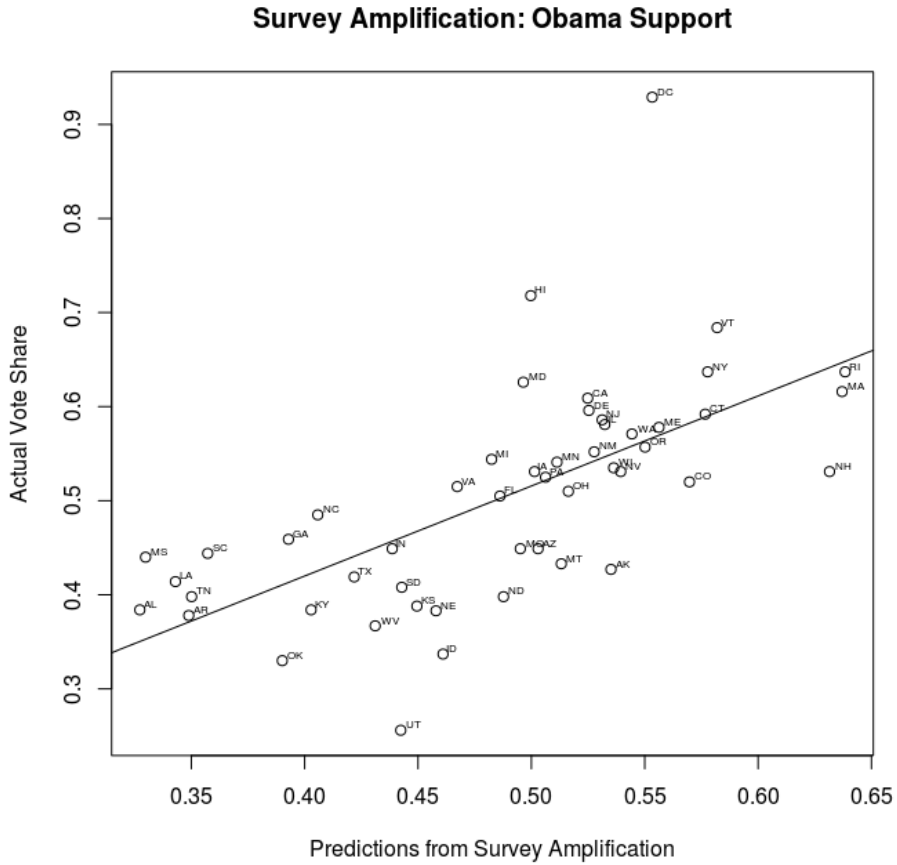
1. Run a GCS asking "Do you support Obama in the upcoming election?"
2. Associate each (yes,no) response in the survey data to the city associated with the respondent.
3. Build a predictive model for the responses using the Trends category data described above.
4. The resulting regression can be used to extrapolate survey responses to any other geographic region using the Google Trends categories associated with that city.

Predictors for Obama support

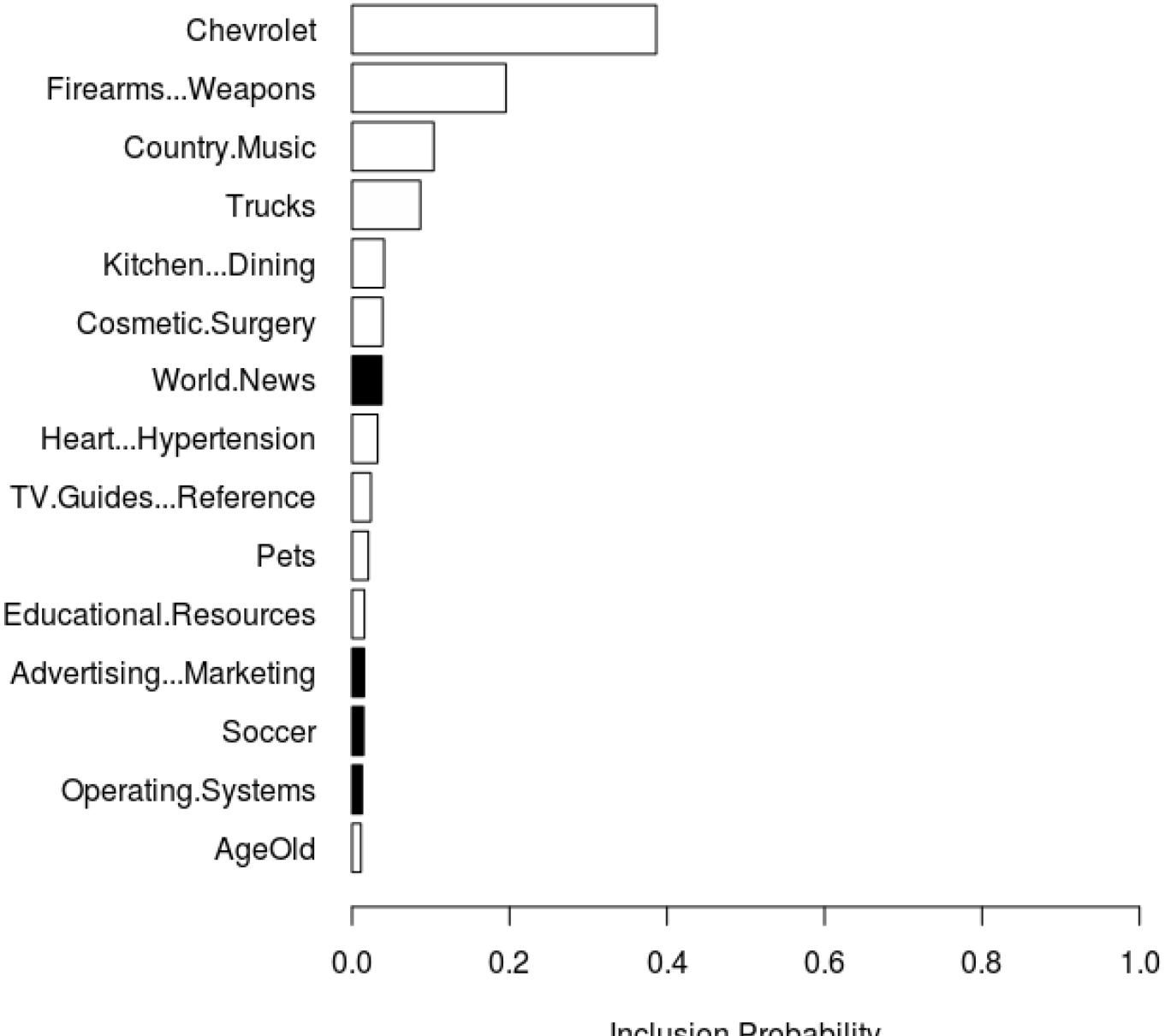


Election outcome by state and DMA





“I prefer to buy products that are assembled in America”



Where is this message predicted to resonate?

Most in favor Kernshaw, SC; Summersville, WV; Grundy, VA; Chesnee, SC; Duffield, VA; Norton, VA; Jonesville, VA; Walnut Cove, NC; Weston, WV; Ennice, NC ...

Least in favor Calipatria, CA; Fremont, CA; Mountain View, CA; San Jose, CA; Berkeley, CA; Redmond, WA; Glendale, CA; Cupertino, CA; Palo Alto, CA; Daggett, CA ...