Google Tools for Data

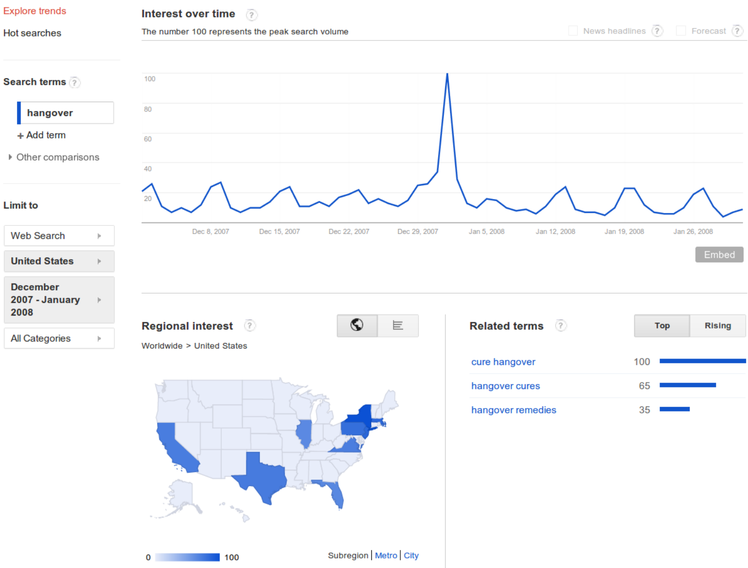
Hal Varian
Google

March 28, 2014

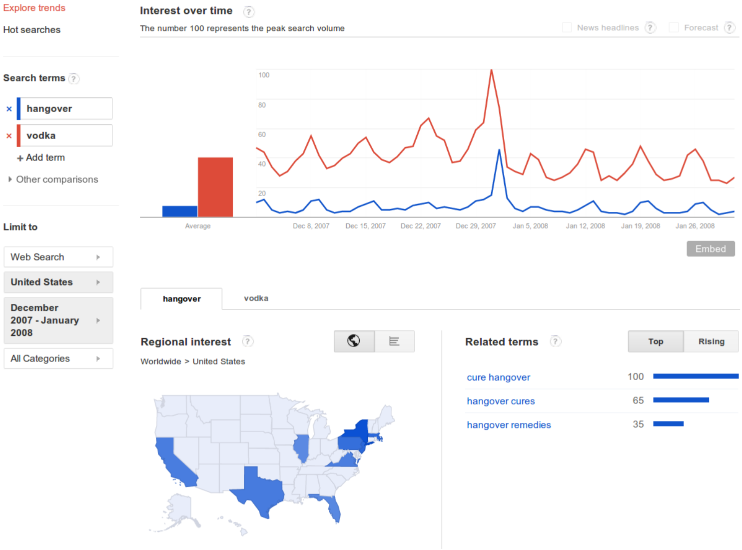
What day of the week are there the most searches for [hangover]?

1. Sunday
2. Monday
3. Tuesday
4. Wednesday
5. Thursday
6. Friday
7. Saturday

Searches for [hangover]

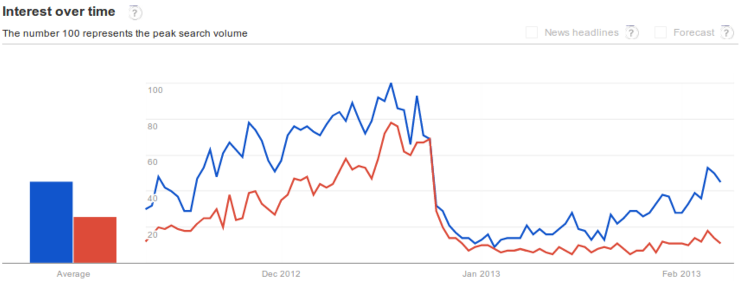


Searches for [hangover] and [vodka]



Looking for gifts when single

1. [gift for boyfriend]
2. [gift for girlfriend]



Looking for gifts when married

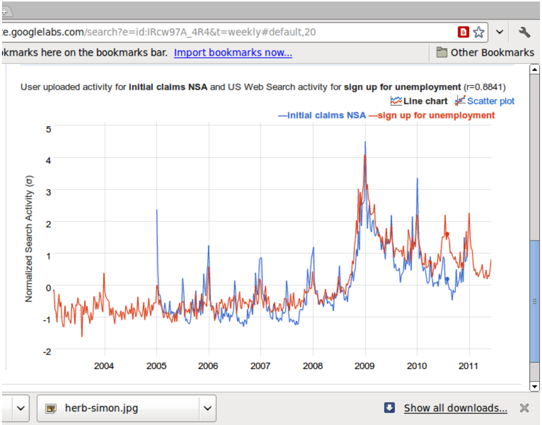
1. [gift for husband]
2. [gift for wife]

Problem motivation

- Want to use Google Trends data to nowcast economic series
  - unemployment may be predicted by "job search" queries
  - auto purchases may be predicted by "vehicle shopping" queries
  - often a contemporaneous relationship, hence "nowcasting"
  - useful due to reporting lags and revisions
- Fat regression problem: there are many more predictors than observations
- Millions of queries, hundreds of categories
  - number of observations ~ 100 for monthly economic data
  - number of predictors ~ 150 for "economic" categories in Trends
- How do we choose which variables to include?

Example: unemployment

- Sometimes Google Correlate works
- Load in: data on initial claims for unemployment benefits
- Returns: 100 queries, including [sign up for unemployment]



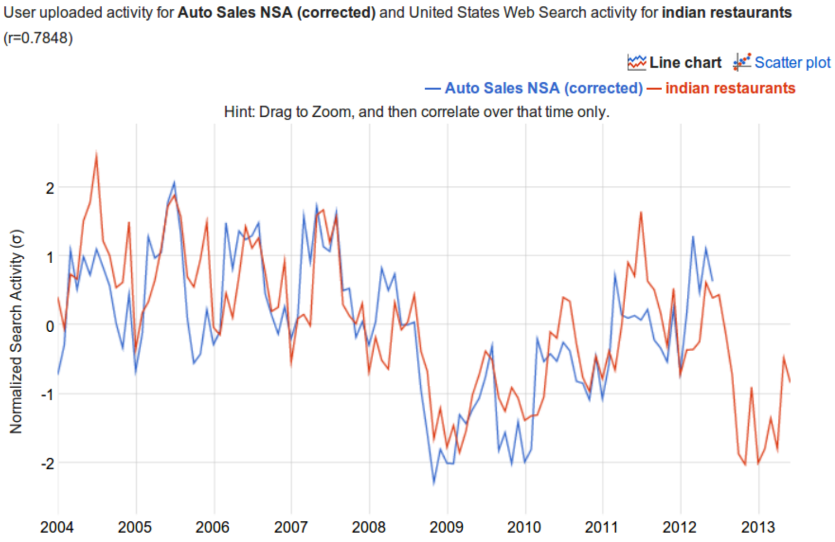
Build a simple AR model

- Use deseasonalized initial claims (y_t)
- Use deseasonalized, detrended searches for [sign up for unemployment] (x_t)

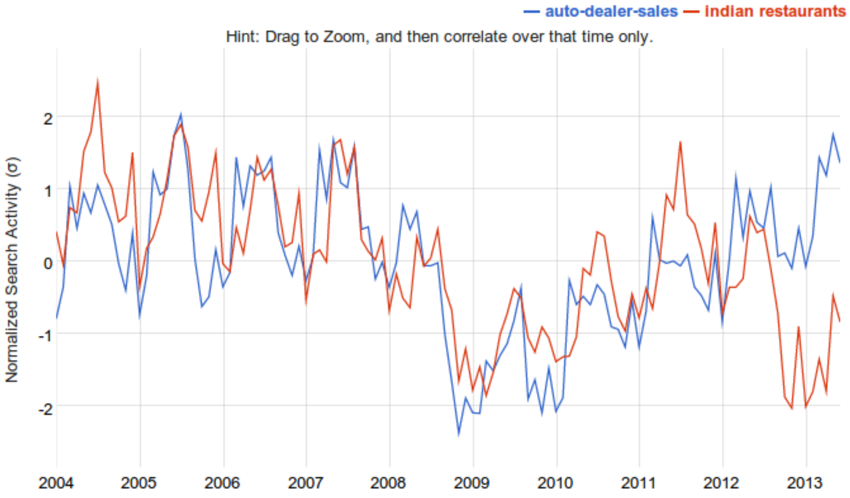
$$\begin{aligned}
\text{base: } y_t &= a_0 + a_1 y_{t-1} + e_t \\
\text{regr: } y_t &= a_0 + a_1 y_{t-1} + b x_t + e_t
\end{aligned}$$

- Estimate regressions using rolling window
- One-step-ahead MAE during recession is about 8.7% lower when [sign up for unemployment] query is included

But simple correlation has limits . . .



Spurious correlation is a danger



How to avoid spurious correlation?

- Control for trend and seasonality
  - Build a model for the *predictable* (trend + seasonality) part of time series
  - In time series this is called *whitening* or *prewhitening*
  - Find regressors that predict the *residuals* after removing trend and seasonality
- How to choose regressors?
  - Simple correlation is too limited
  - Human judgment doesn't scale

Some approaches to variable selection

- Human judgment: what we mostly do
- Significance testing: forward and backward stepwise regression
- Complexity criteria: AIC, BIC, etc
- Dimensionality reduction: principle component, factor models, partial least squares
- Machine learning: Penalized regression, lasso, LARS, ridge regression, elastic net

Our approach

- Bayesian Structural Time Series (BSTS)
  - Decompose time series into trend + seasonality + regression
  - Use Kalman filter for trend + seasonality (whiten time series)
  - Spike and slab regression for variable selection
  - Estimate via Markov Chain Monte Carlo simulation of posterior distribution
  - Bayesian model averaging for final forecast

How BSTS helps reduce overfitting

- Kalman filter used to whiten the series
  - Remove common seasonality and trend, regressors chosen to predict residuals
  - Estimation of (seasonality, trend, regression) is simultaneous
  - Same spirit as Granger causality
- Overfitting due to spurious correlation with regressors
  - Remove "one time" events (can be automated)
  - Apply human judgment
- Overfitting due to many regressors
  - Informative prior to suggest likely number of regressions or regressor categories
  - Bayesian model averaging over many small regressions ("ensemble estimation")

Basic structural model with regression

- Consider classic time series model with *constant* level, linear time trend, and regressors
  - $y_t = \mu + bt + \beta x_t + e_t$
- "Local linear trend" is a stochastic generalization of this
  - Observation: $y_t = \mu_t + z_t + e_{1t}$ = level + regression
  - State 1: $\mu_t = \mu_{t-1} + b_{t-1} + e_{2t}$ = random walk + trend
  - State 2: $b_t = b_{t-1} + e_{3t}$ = random walk for trend
  - State 3: $z_t = \beta x_t$ = regression
- Parameters to estimate: regression coefficients β and variances of (e_{it}) for $i = 1, \ldots, 3$
- Use these variances to construct optimal Kalman forecast: $\hat{y}_t = \hat{y}_{t-1} + k_t \times (y_{t-1} - \hat{y}_{t-1}) + x_t\beta$
- k_t depends on the estimated variances

Intuition for Kalman filter

- Consider simple case without regressors and trend
  - Observation equation: $y_t = \mu_t + e_{1t}$
  - State equation: $\mu_t = \mu_{t-1} + e_{2t}$
- Two extreme cases
  - $e_{2t} = 0$ is constant mean model where best estimate is sample average through $t-1$: $\bar{y}_{t-1} = \sum_{s=1}^{t-1} y_s$
  - $e_{1t} = 0$ is random walk where best estimate is current value y_{t-1}
- For general case take weighted average of current and past observations, where weight depends on estimated variances

Nice features of Kalman approach

- No problem with unit roots or other kinds of nonstationarity
- No problem with missing observations
- No problem with mixed frequency
- No differencing or identification stage (easy to automate)
- Nice Bayesian interpretation
- Easy to compute estimates (particularly in Bayesian case)
- Nice interpretation of structural components
- Easy to add seasonality
- Good forecast performance

Spike and slab regression for variable choice

- Spike
  - Define vector γ that indicates variable inclusion
  - $\gamma_i = 1$ if variable i has non-zero coefficient in regression, 0 otherwise
  - Bernoulli prior distribution, $p(\gamma)$, for γ_i
  - Can use an informative prior; e.g., expected number of predictors
- Slab
  - Conditional on being in regression ($\gamma_i = 1$) put a (weak) prior on β_i, $p(\beta|\gamma)$.
- Estimate posterior distribution of (γ, β) using MCMC

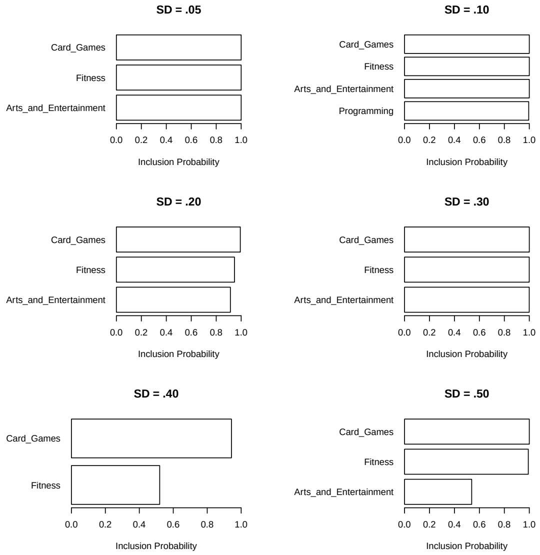
Bayesian model averaging

- We simulate draws from posterior using MCMC
- Each draw has a set of variables in the regression (γ) and a set of regression coefficients (β)
- Make a forecast of y_t using these coefficients
- This gives the posterior forecast distribution for y_t
- Can take average over all the forecasts for final prediction
- Can take average over draws of γ to see which predictors have high probability of being in regression

Torture test simulation for BSTS

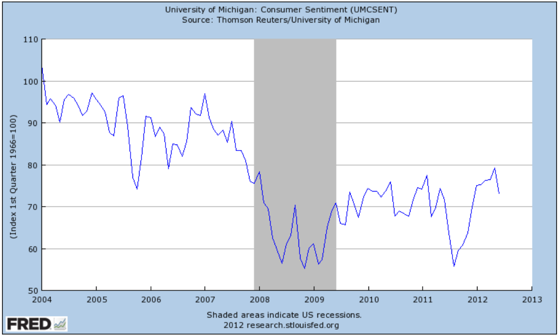
- Pick $k = 3$ categories (out of 150) and their associated time series
- Construct artificial time series = sum of these k + noise
- See if BSTS picks the right categories
  - 0 noise = perfect
  - 5% noise = perfect
  - 10% noise = misses one, but still does good forecast
  - performance deteriorates for higher noise levels
  - ... but it degrades gracefully

Example of torture test



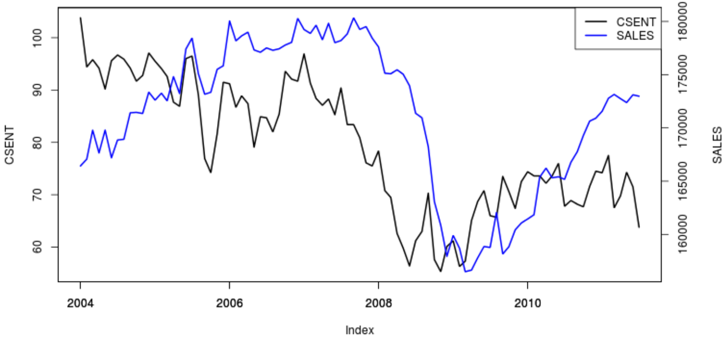
Example 1: Consumer Sentiment

- Monthly UM Consumer sentiment from Jan 2004 to Apr 2012 ($n = 100$)
- Google Insights for Search categories related to economics ($k = 150$)
- No compelling intuition about what predictors should be



Consumer sentiment as leading indicator

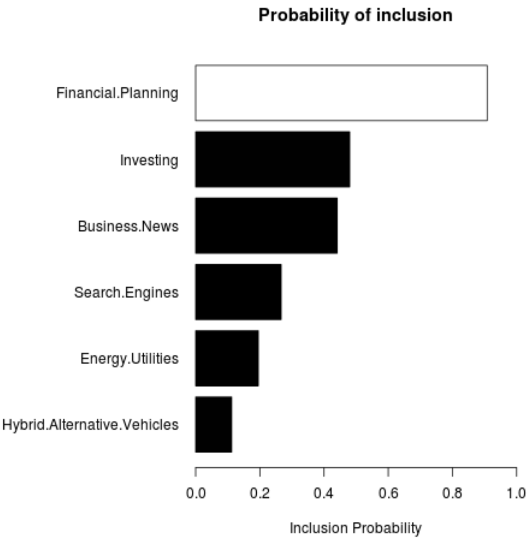
- Leading indicator of retail sales in last recession



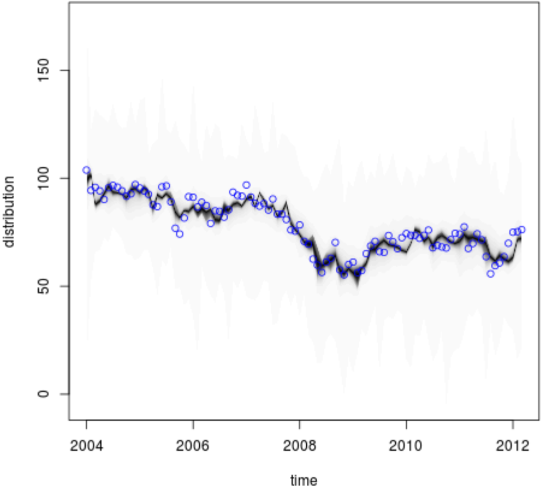
Variable selection

- Google Insights for Search categories related to economics ($k = 150$)
- Deseasonalize predictors using R command `stl`
- Detrend predictors using simple linear regression
- Let `bsts` choose predictors

UM Consumer Sentiment Predictors



Posterior distribution of one-step ahead forecast



State decomposition

Recall observation equation:

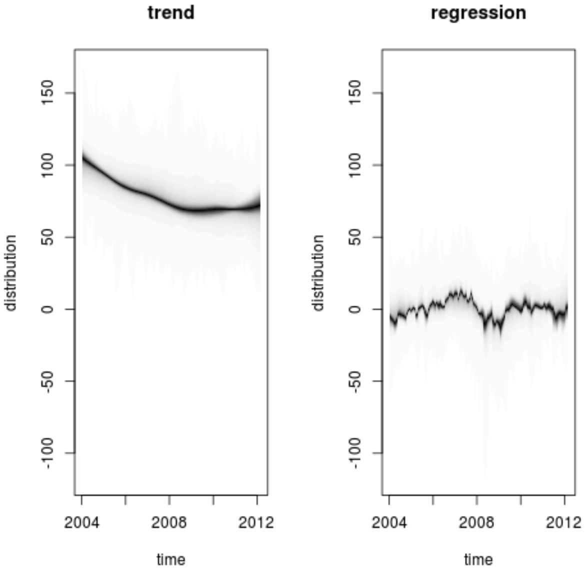
$$y_t = \mu_t + x_t \beta + e_{1t}$$

We can plot the posterior distribution of each of these components. The regression component can be further expanded

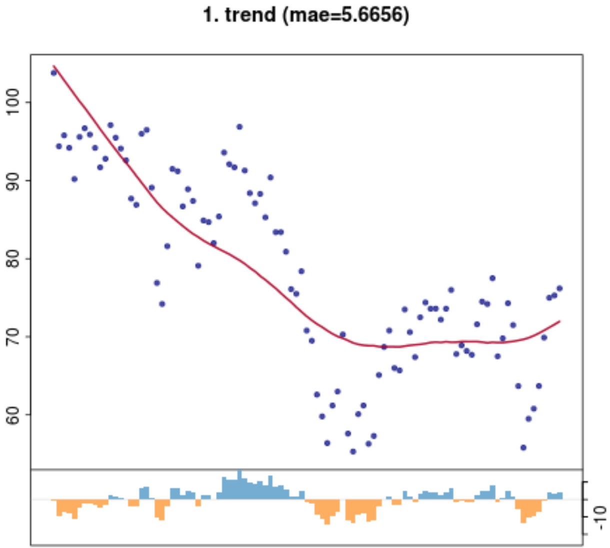
$$y_t = \mu_t + x_{1t}\beta_1 + \cdots + x_{pt}\beta_p + e_{1t}$$

Natural to order predictors by probability of inclusion and look at cumulative plot.

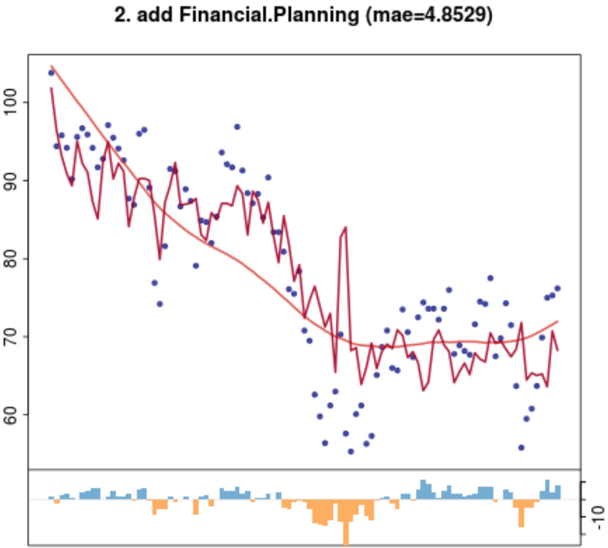
Trend and regression decomposition



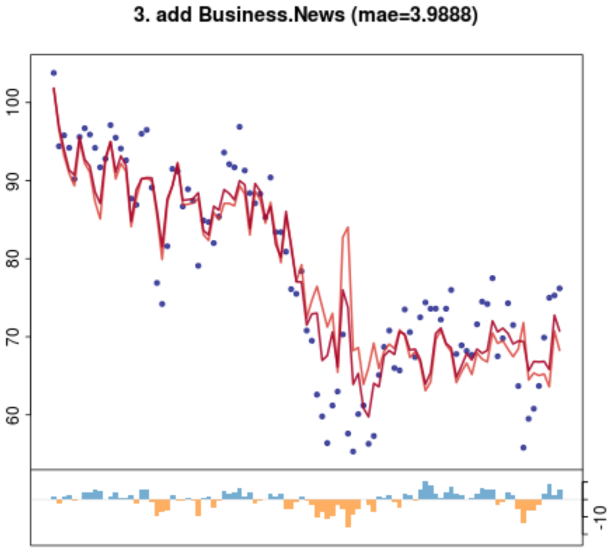
Trend



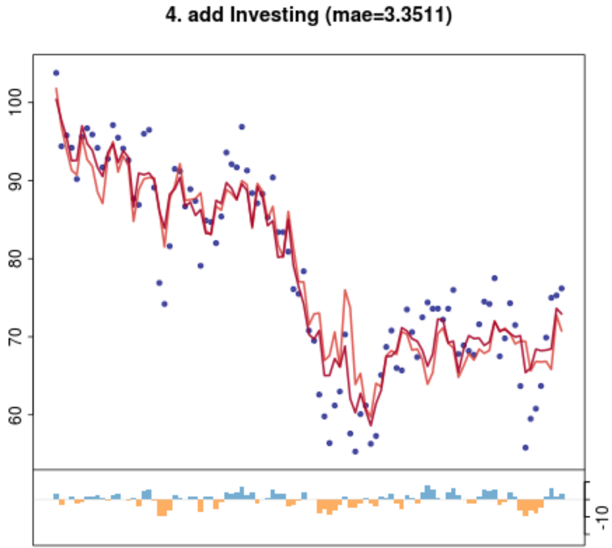
add Financial Planning



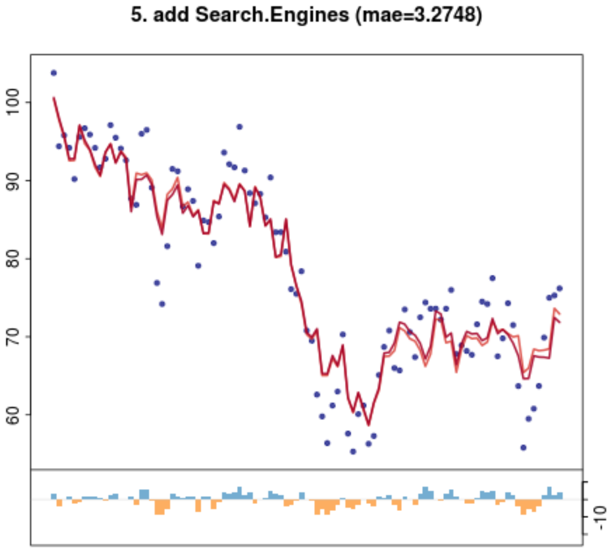
add Business News



add Investing

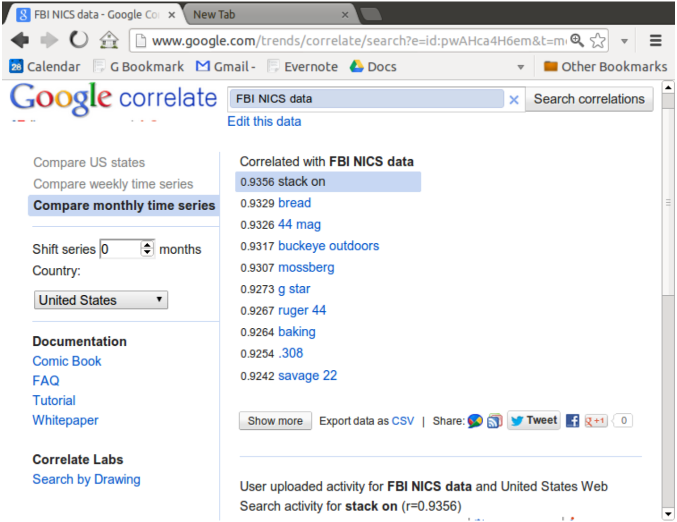


add Search Engines



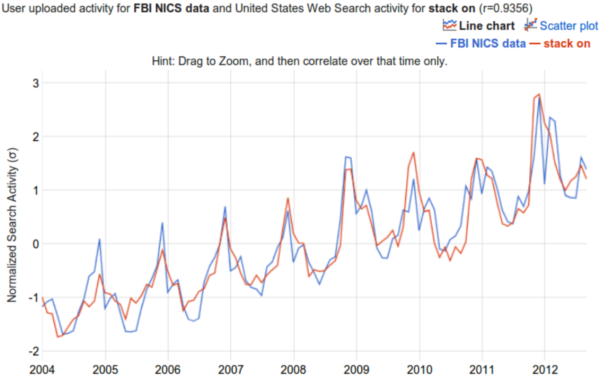
Example 2: gun sales

Use FBI's National Instant Criminal Background Check

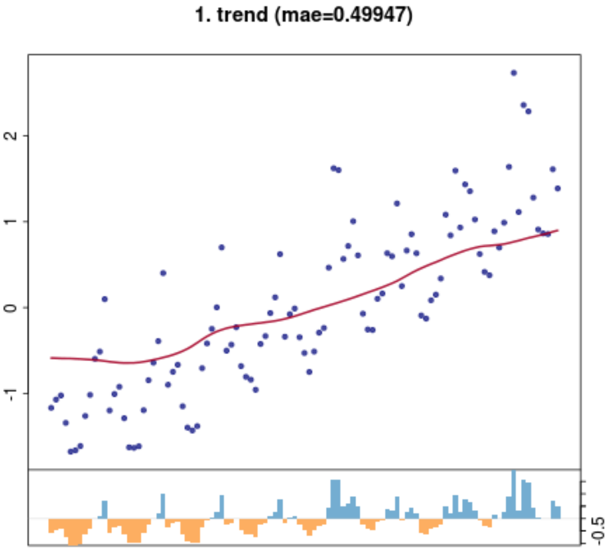


Google Correlate Results

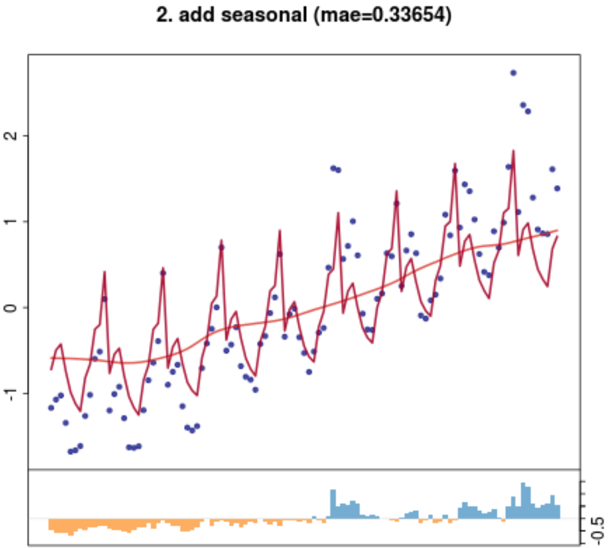
- [stack on] has highest correlation
- [gun shops] is chosen by bsts
- Regression model gives 11% improvement in one-step ahead MAE



Trend



Seasonal



Gun Shops

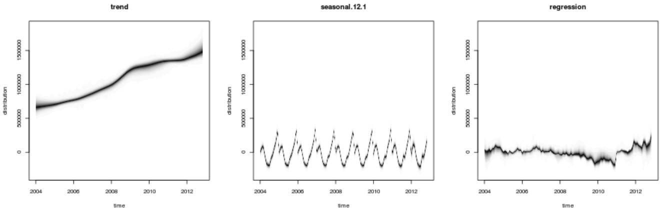
Google Trends predictors

- 586 Google Trends verticals, deseasonalized and detrended
- 107 monthly observations

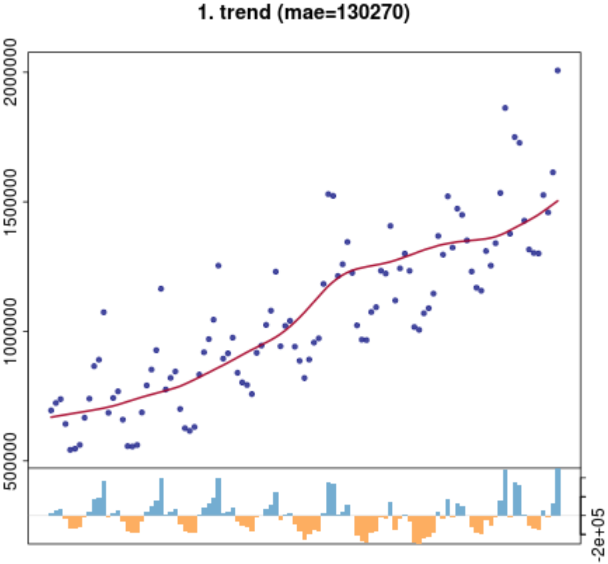
Category	mean	inc.prob
Recreation::Outdoors::Hunting:and:Shooting	1,056,208	0.97
Travel::Adventure:Travel	-84,467	0.09

Table: Google Trends predictors for NICS checks.

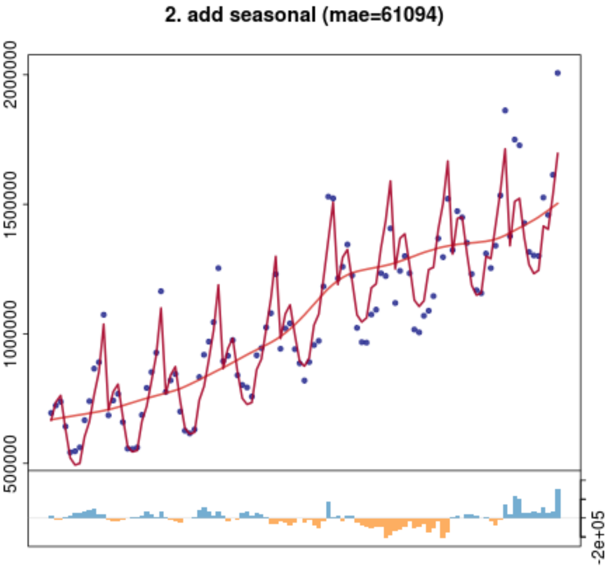
State decomposition



Trend

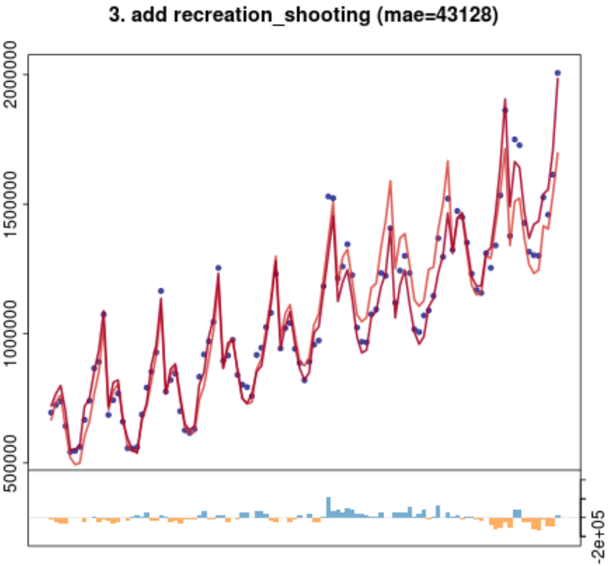


Seasonal

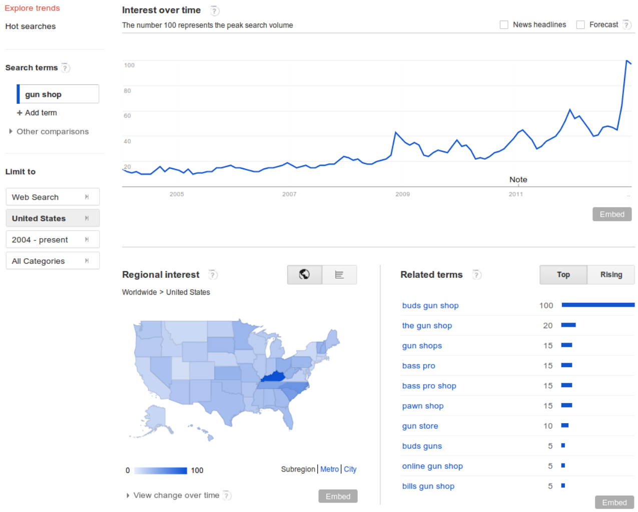


2. add seasonal (mae=61094)
2000000
1500000
1000000
500000
-2e+05

Hunting and Shooting



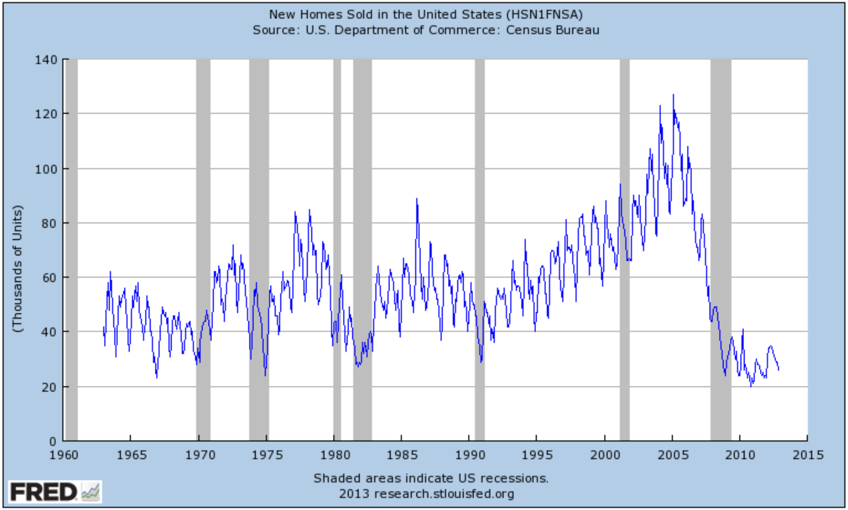
Searches for [gun shop]



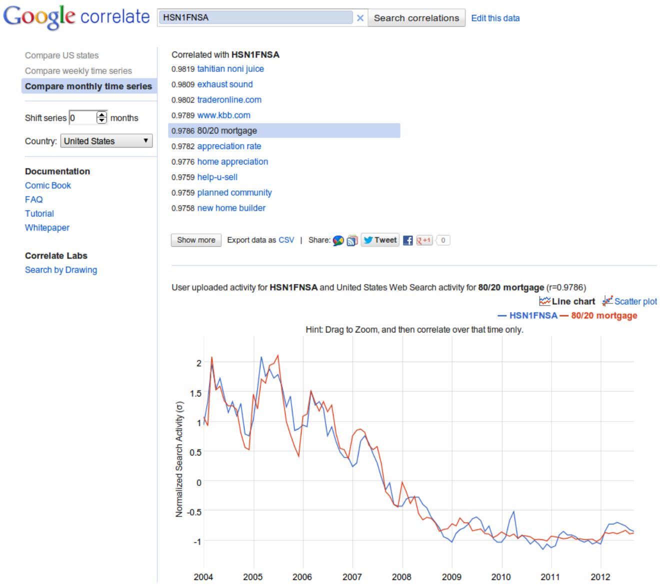
Fun with priors

- Can use prior to improve estimate of trend component
  - Google data starts in 2004, only one recession
  - Can estimate parameters of trend model with no regressors
  - Use this as prior for estimate of trend in estimation period
- Can use prior to influence variable choice in regression
  - Influence the expected number of variables in regression (parsimony)
  - Give higher weight to certain verticals (e.g., economics related)
  - Exclude obvious spurious correlation (e.g., pop song titles)

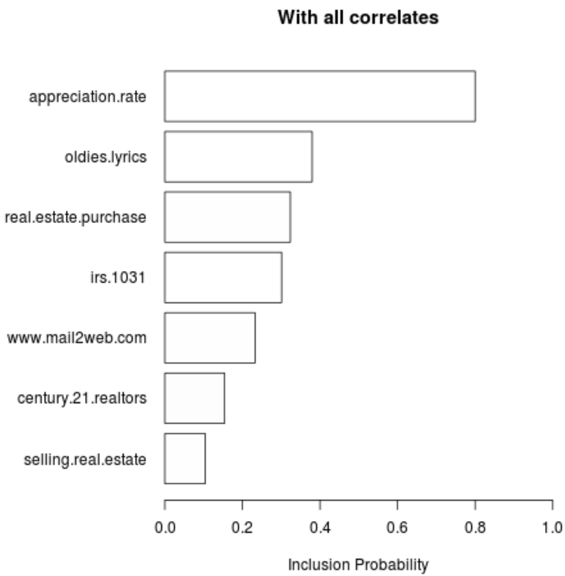
New Homes Sold in the US



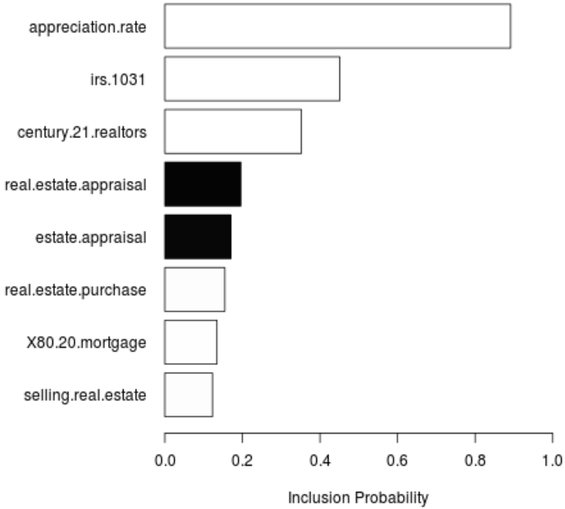
Run correlate



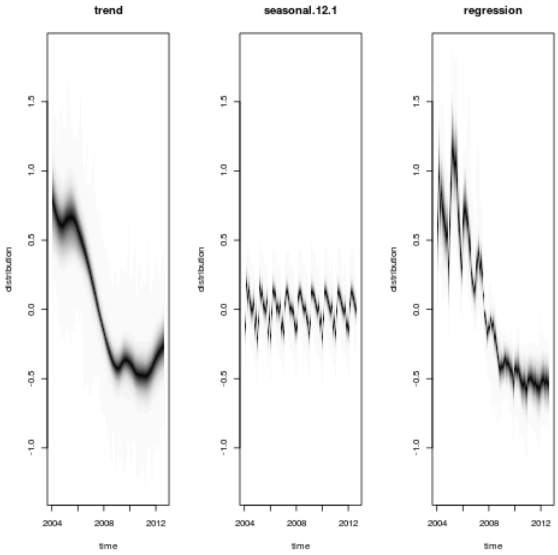
BSTS variable selection



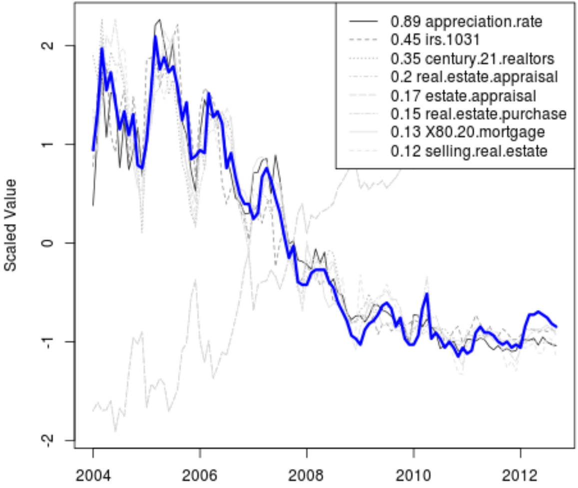
Eliminate spurious correlates



State decomposition

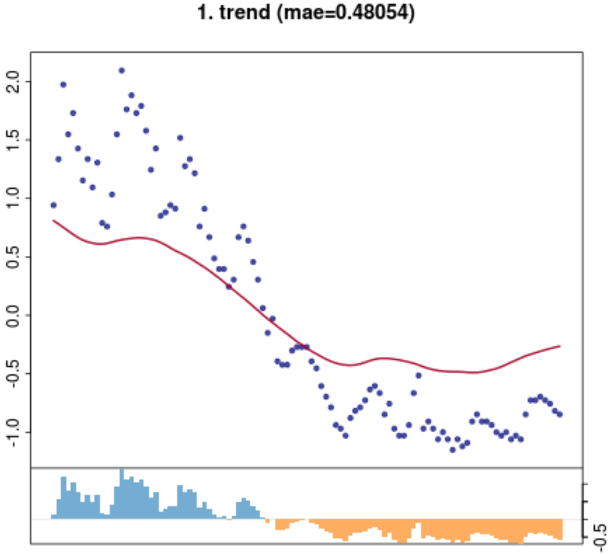


Predictors

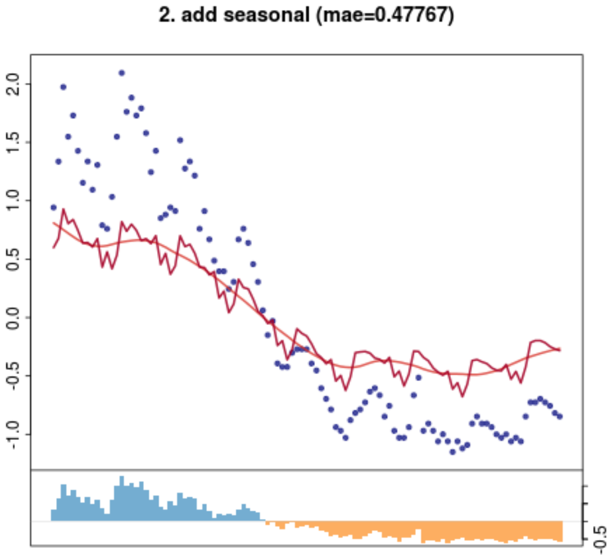


0.89 appreciation.rate
0.45 irs.1031
0.35 century.21.realtors
0.2 real.estate.appraisal
0.17 estate.appraisal
0.15 real.estate.purchase
0.13 X80.20.mortgage
0.12 selling.real.estate
Scaled Value
2
1
0
-1
-2
2004
2006
2008
2010
2012

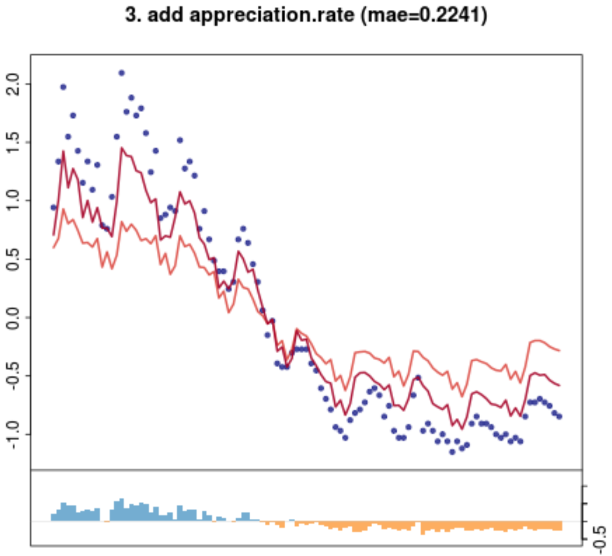
Trend



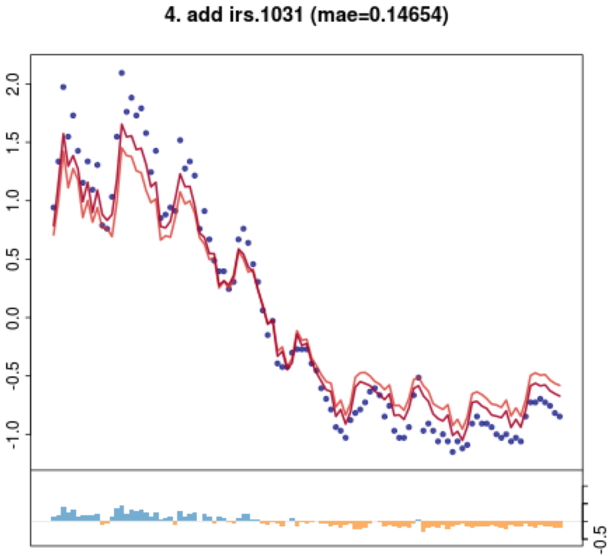
Seasonal



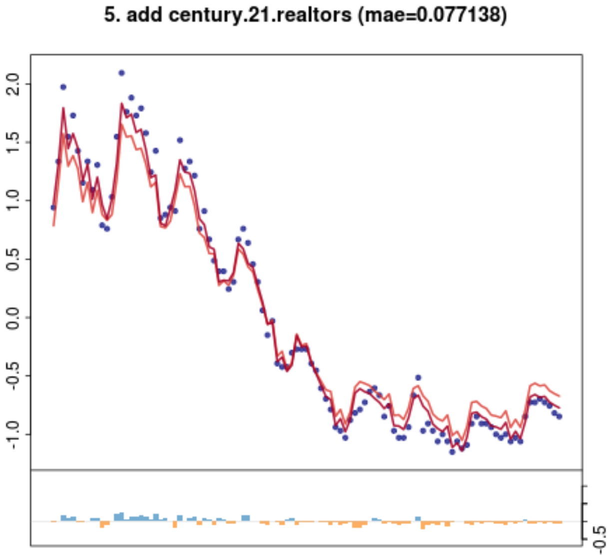
Appreciation rate



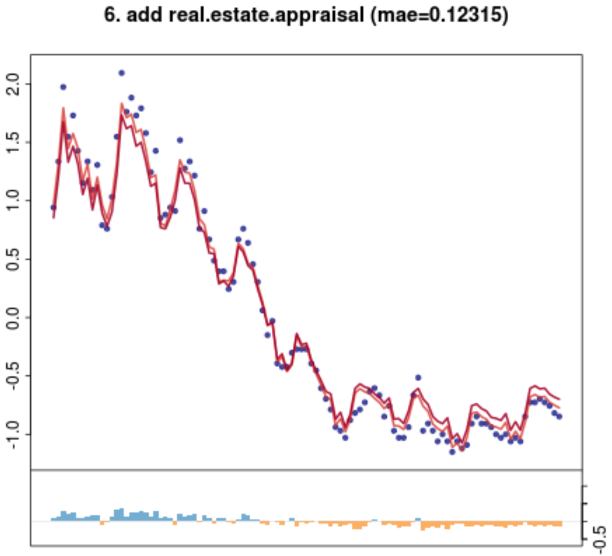
IRS 1031



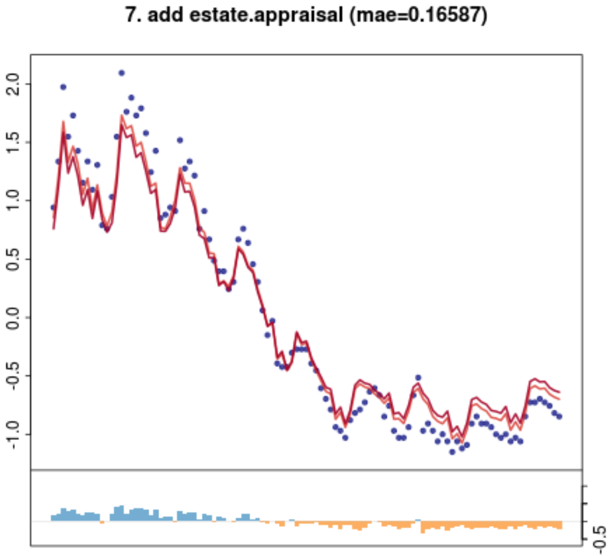
Century 21 realtors



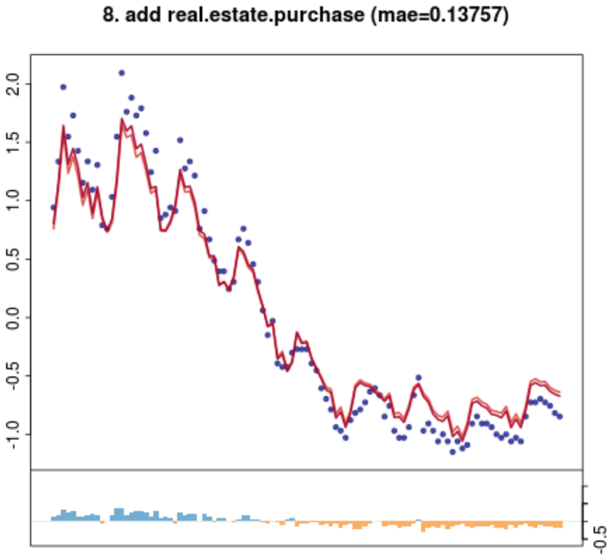
Real estate appraisal



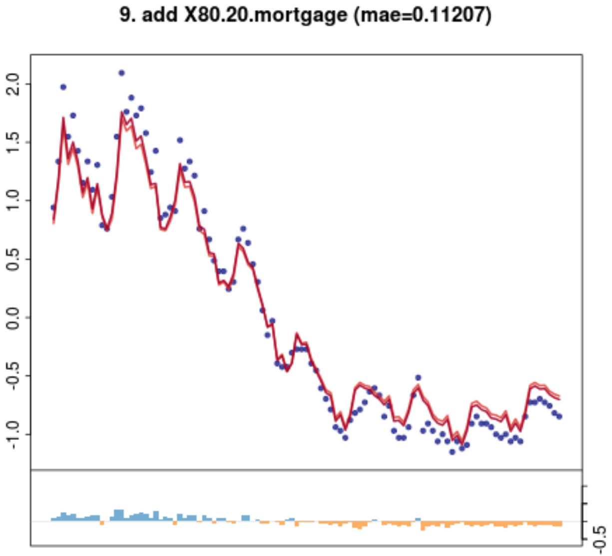
Estate appraisal



Real estate purchase



80-20 mortgage



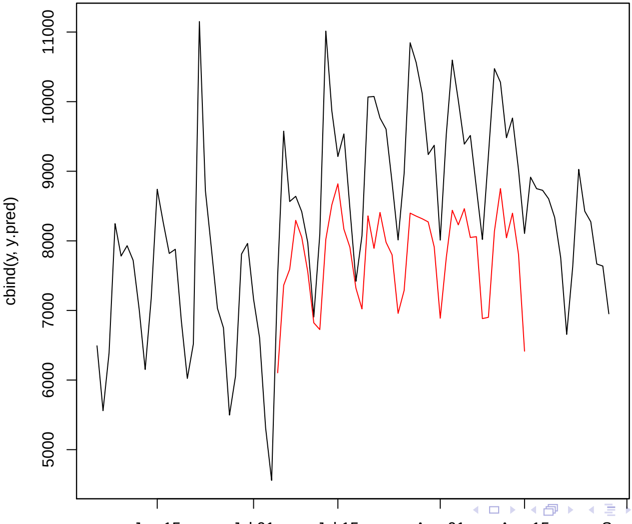
Causal inference

- In order to determine the causal impact of an intervention
  - Estimate what would have happened without the intervention ("the counterfactual")
  - Compare counterfactual to actual outcome
- We can use BSTS (and related tools) to build predictive model for counterfactual

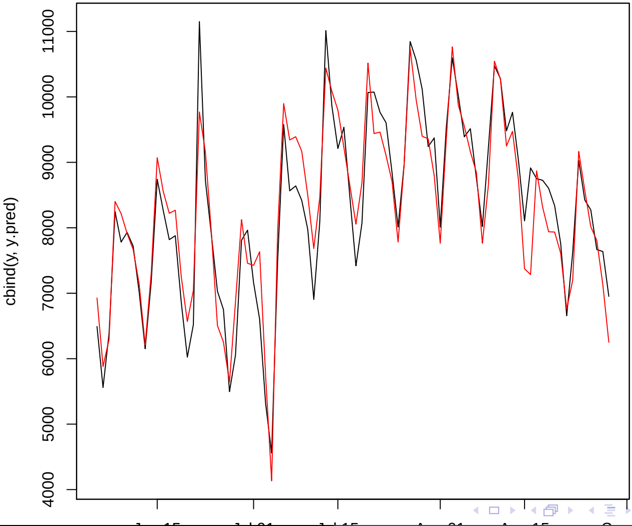
Example: impact of ad spend

- An online advertiser increased ad spending for 6 weeks. What was the impact on visitors to its web site? Two approaches:
  1. Extrapolation
     - Use BSTS to determine best predictors using Trends categories.
     - Turned out to be `Photo and Video Sharing` and `Photo and Video Services`.
     - Build predictive model of counterfactual using these predictors
  2. Dummy variable for campaign period

Extrapolation



Dummy variable



Survey Amplification

- Google Consumer Surveys uses an online survey in place of an ad
- User completes survey to get access to online content
- Win, Win, Win
  - Survey writer: pays about 10 cents a survey
  - Publisher gets 5 cents per response
  - User gets access to premium content

Example of survey

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...

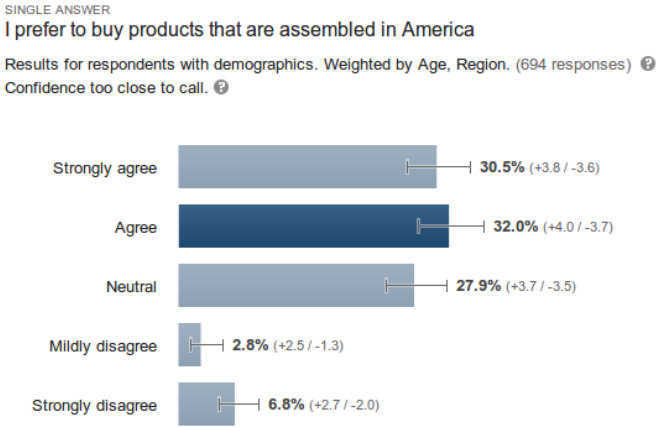
Answer a question to continue reading this page

question 2 of 2:
Have you ever purchased anything from (check all that apply):

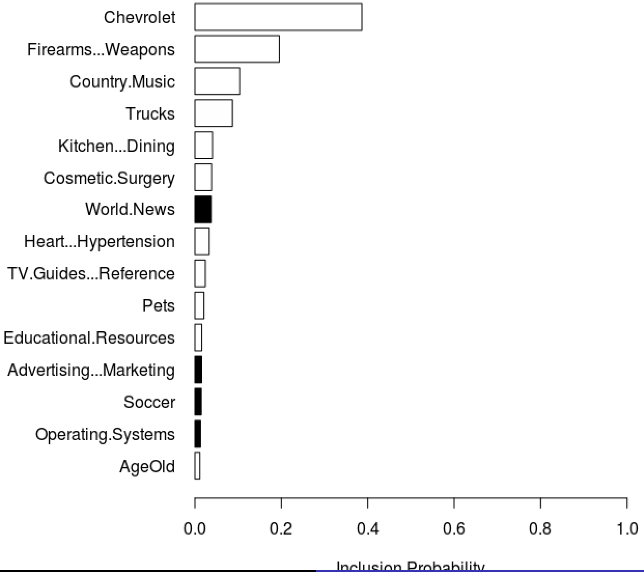
Check all answers that apply

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

Assembled in America?



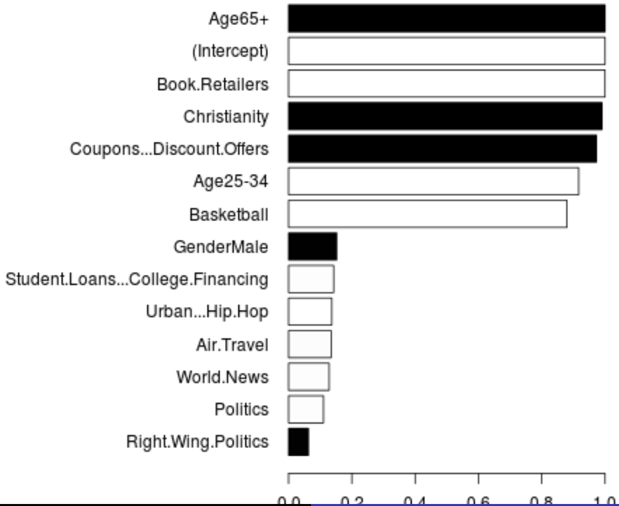
Predictors from Trends



Where to advertise

Top cities		Bottom cities	
1. Kershaw, SC	83.2	Calipatria, CA	40.2
2. Summersville, WV	82.8	Fremont, CA	40.2
3. Grundy, VA	82.8	Mountain View, CA	40.8
4. Chesnee, SC	82.7	San Jose, CA	41.4
5. Duffield, VA	82.5	Berkeley, CA	41.4
6. Norton, VA	82.3	Redmond, WA	41.5
7. Jonesville, VA	82.2	Glendale, CA	41.5
8. Walnut Cove, NC	82.2	Cupertino, CA	41.6

Predictors of Obama Vote



Vote by State