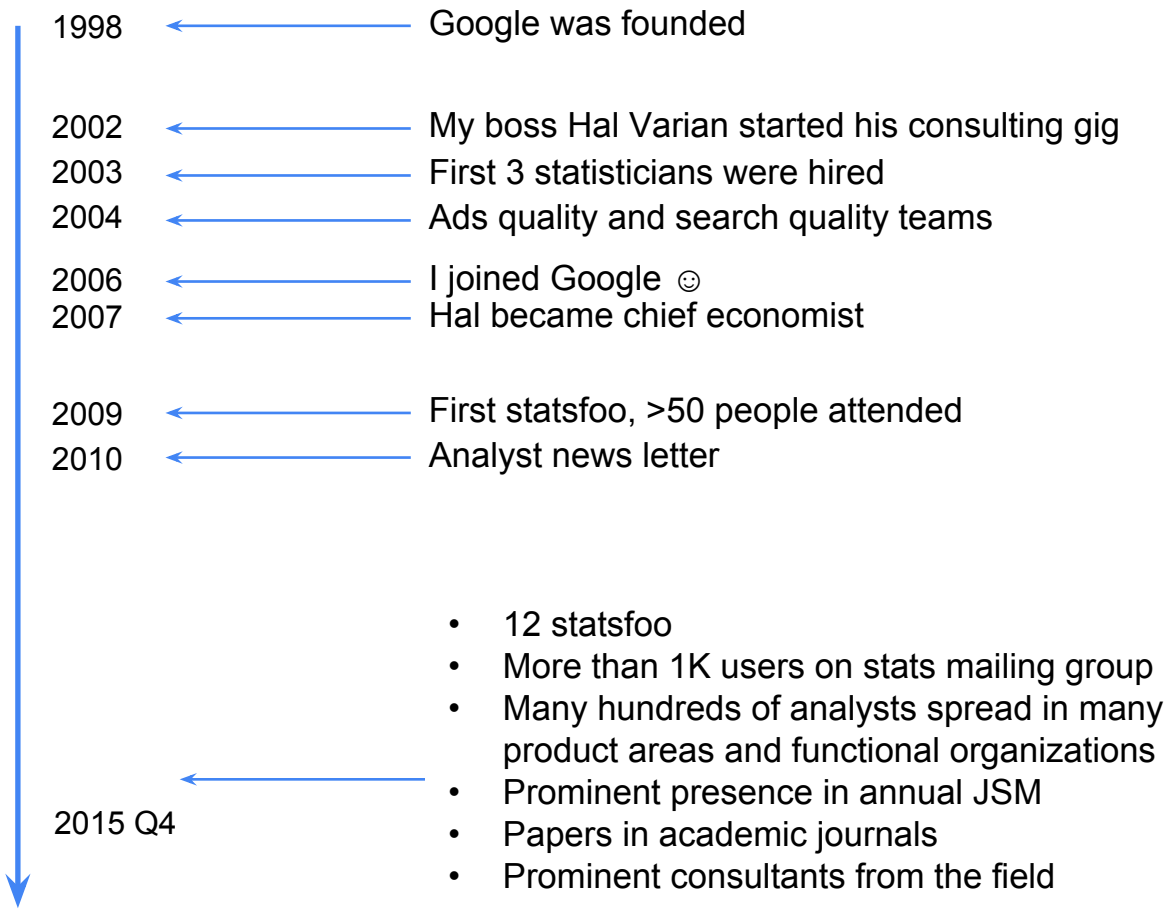
Google

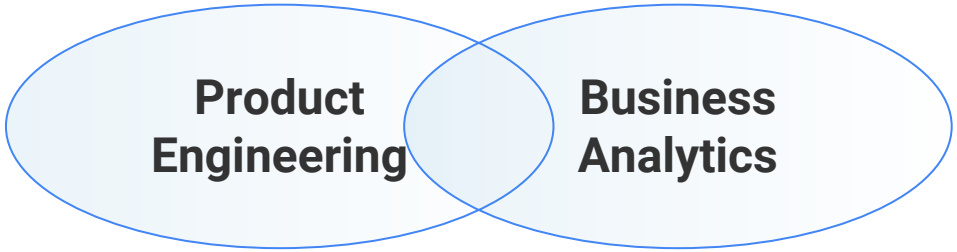
Statistics at Google

Statistics Day In Puerto Rico
9 October 2015
Qing Wu, Senior Economist

Milestones of Statistics in Google



What do Statisticians Do?



- Search
- Ads
- Youtube
- Geo/Map
- Google now
- Google play
- Google consumer survey
- Google analytics
- ...

- Advertising sales analytics
- Quantitative marketing
- Financial analytics
- People analytics
- Quantitative user experiences
- Hardware sales analytics
- Operation analytics (Infrastructure, Shopping Express, Fiber)

Problems We Solve (A few Examples)

Forecast daily/hourly search traffic by country/datacenter

What is impact of weather/recessions/holidays on revenue?

How to show best ads with best formats?

Do users gain ads blindness if we serve them bad ads?

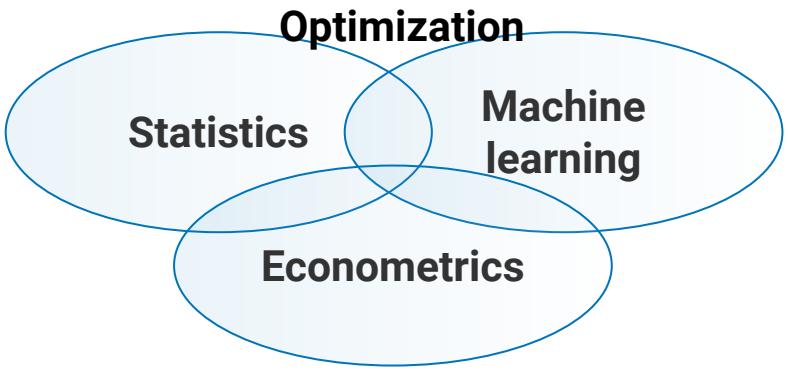
Do advertisers increase spending after adopting a new feature?

Predict hardware sales: life cycle, impact of promotion, etc

Recommendation in play store

Do Youtube ads improve the brand awareness?

Models We Use



Classic Statistics

Statistical Learning

Bayesian Statistics

Forecasting and Time-Series

Spatial Data Analysis

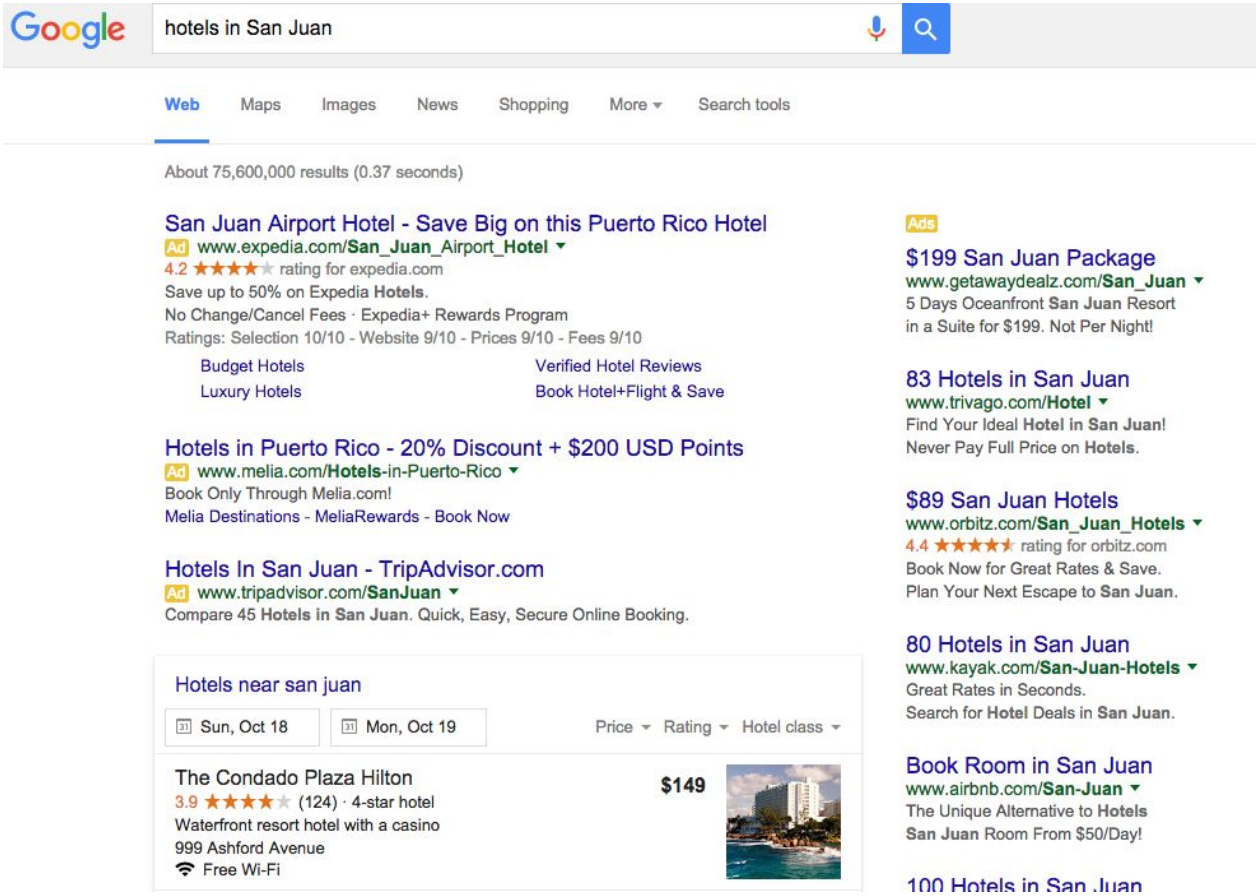
Survey Statistics

Causal Inferences

Experiment Design

Biostatistics

How Does Google Make Money?



Ad Auctions

Generalized Second Price (GSP)

- Order ads by bid per click
- Each ad pays bid of the advertiser below it

"Modified" GSP

Ads ranking depend on

- Bid
- **Prob(clicks | impressions)**
- Quality score
- Formats

Pricing depends on

- Second bid
- Reserved price

Logistic Regression in Parallel

Logistic regression: Prob(click | impr) ~ X b

X: query-ads pairs, ad slots and positions, matching types, etc

Estimate the world's largest logistic regression

- Billions of records
- Update it in real time

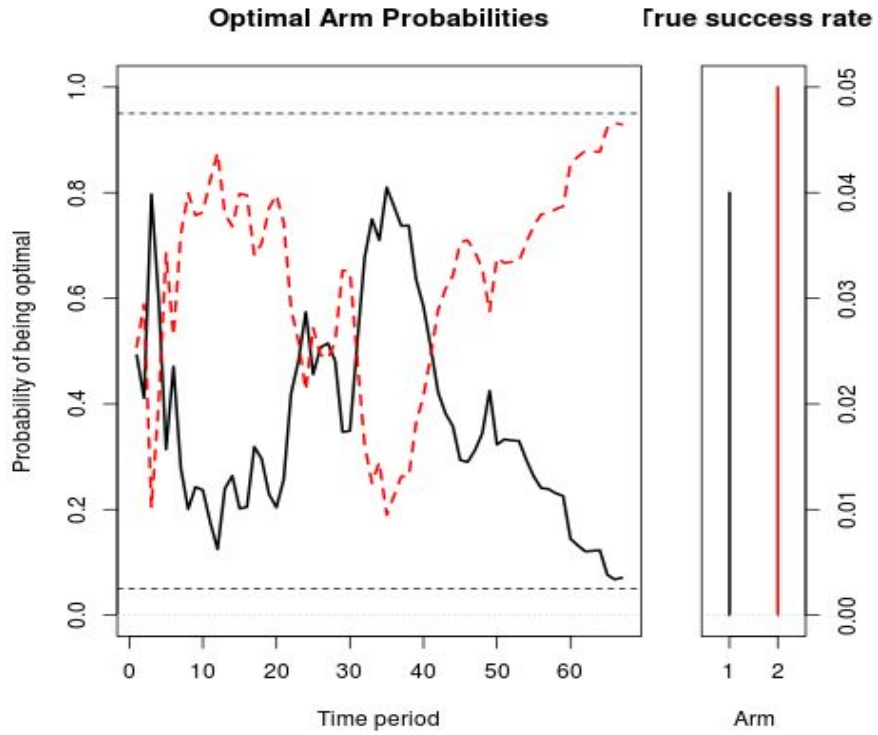
Experimentation

- Experimentation answers the causal questions!
- Experimentation is far easier online.
- Scope of experiments
  - Queries
  - Cookies
  - Geographic
  - Temporal

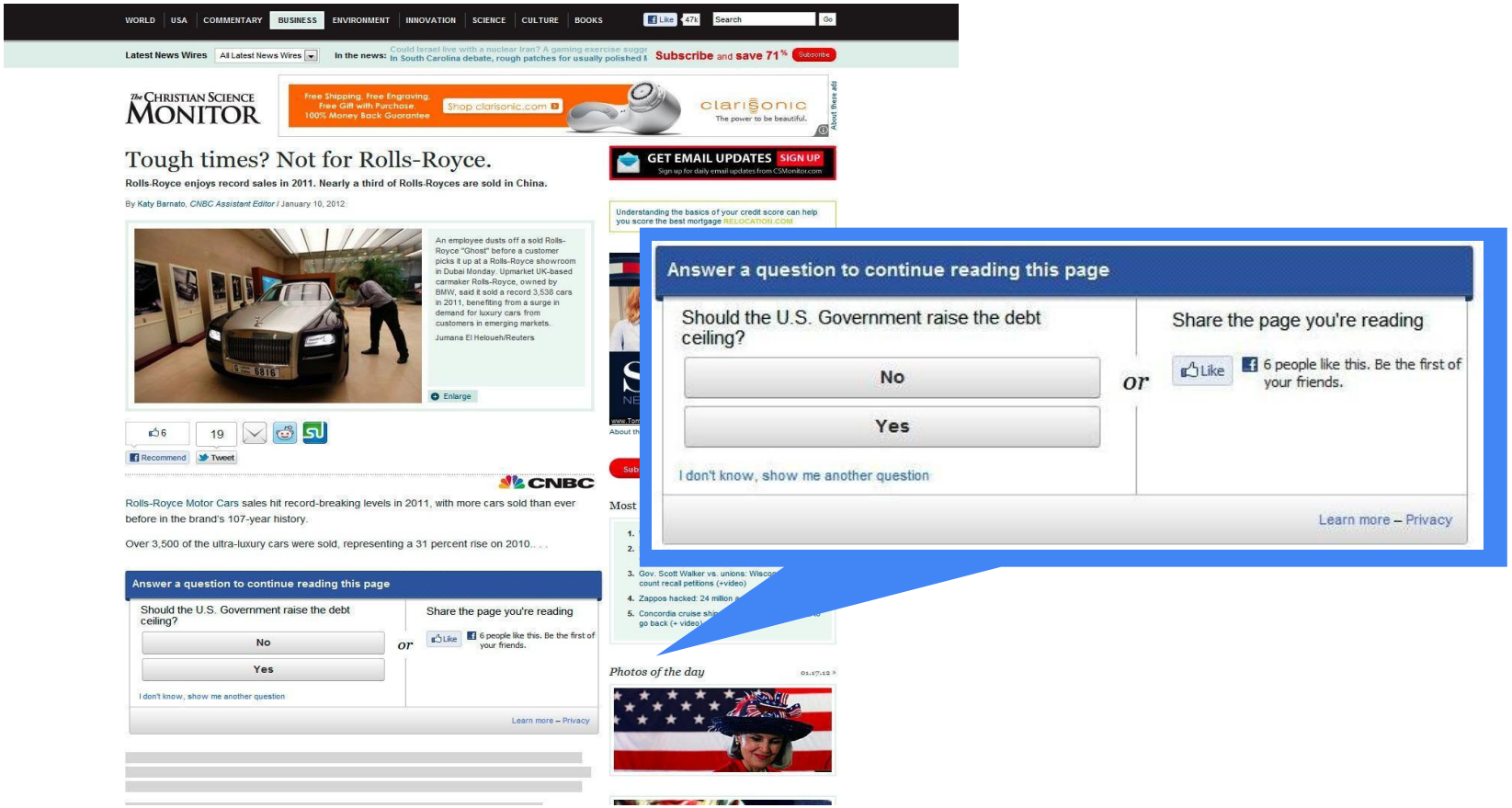
In 2010 Google ran about 10,000 experiments: 5000 in search and 5000 in ads. Implemented 400 improvements in search and a similar number in ads. At any one time on Google you are in a dozen or more experiments.

Efficient Experimentation: Multi-Armed Bandits

Google Analytics Content Experiments: Reassign traffic weights each day based on Bayesian update to accelerate the experiment.



Google Consumer Survey



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

Underlying Statistics

- Bias adjustment
  - Time of day and day of week
  - Non-response table
- Error bars
- Partial ordering
- Hypothesis generation
- Confidence in winner
- And much more!

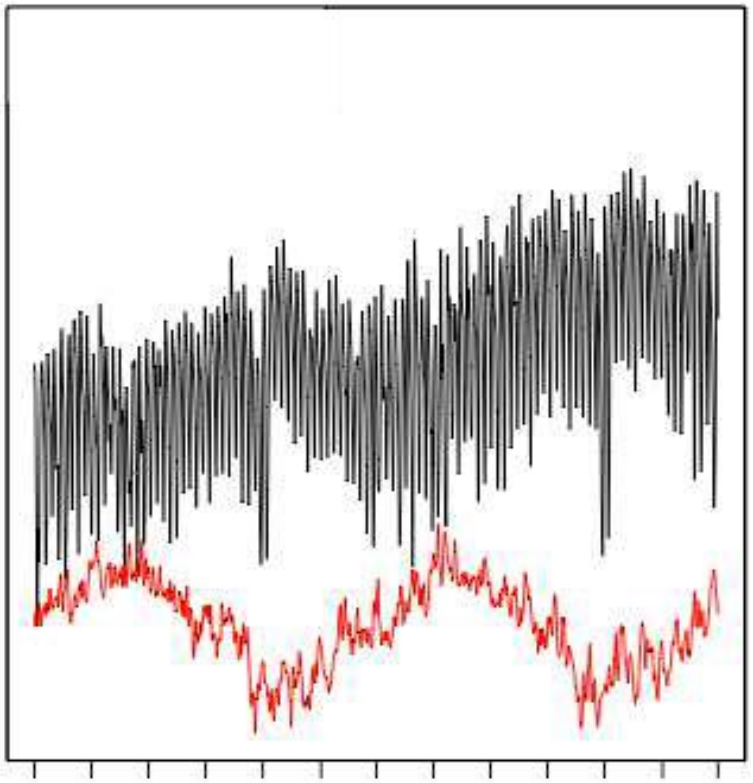
Showcasing Variable Selections for Fat Regressions

In the Google data world, we often have a lot of signals for prediction modeling

- Human Judgment
- Significance Testing
- Dimension Reduction
- Machine Learning
- Bayesian Structural Time Series (BSTS)

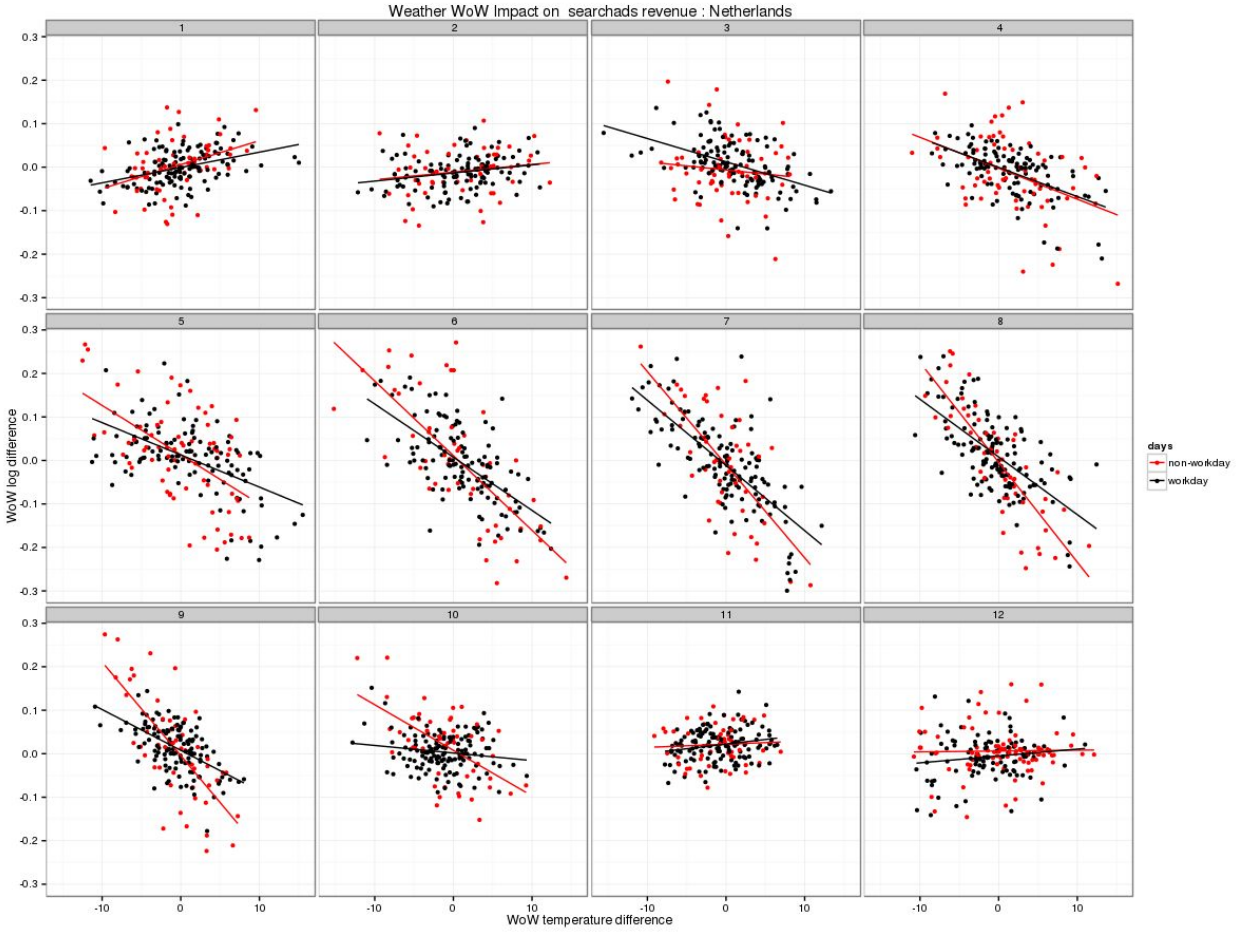
Case A: Weather's Impact on Search Traffic

Search traffic vs daily maximum temperature



At first sight the relationship between two time series does not seem obvious

WoW Difference Scatter Plot by Month

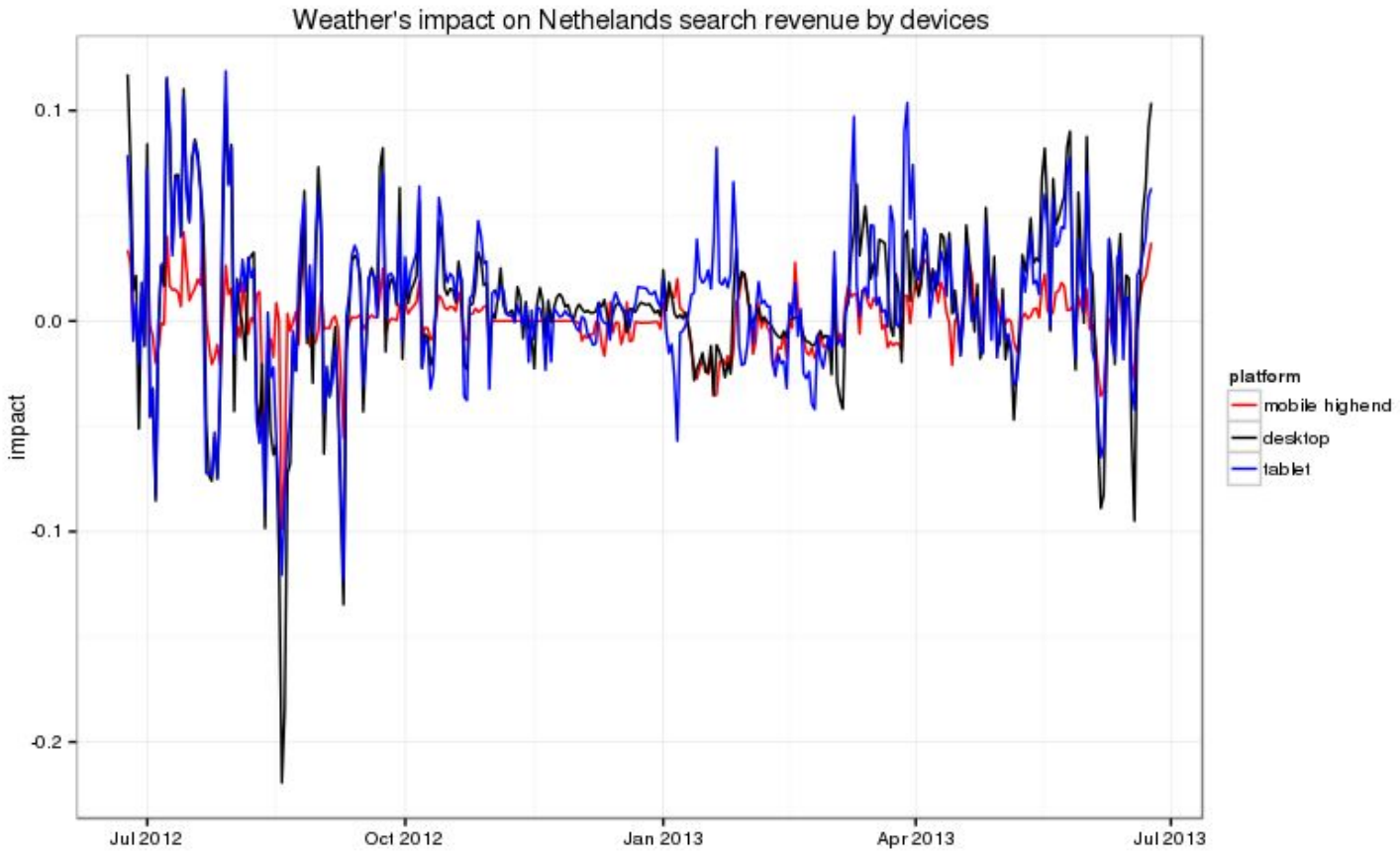


Simple Step-Wise Regression Model

$$
\begin{aligned}
X(n) - X(n-7) \quad = \quad & \alpha_j + (\gamma_{tsj} d_{nsj} + \gamma_{twj} d_{nwj})(T(n) - T(n-7)) \\
& + (\gamma_{tpj} d_{nsj} + \gamma_{tpj} d_{nwj})(P(n) - P(n-7)) \\
& j = 1, \ldots, 12
\end{aligned}
$$

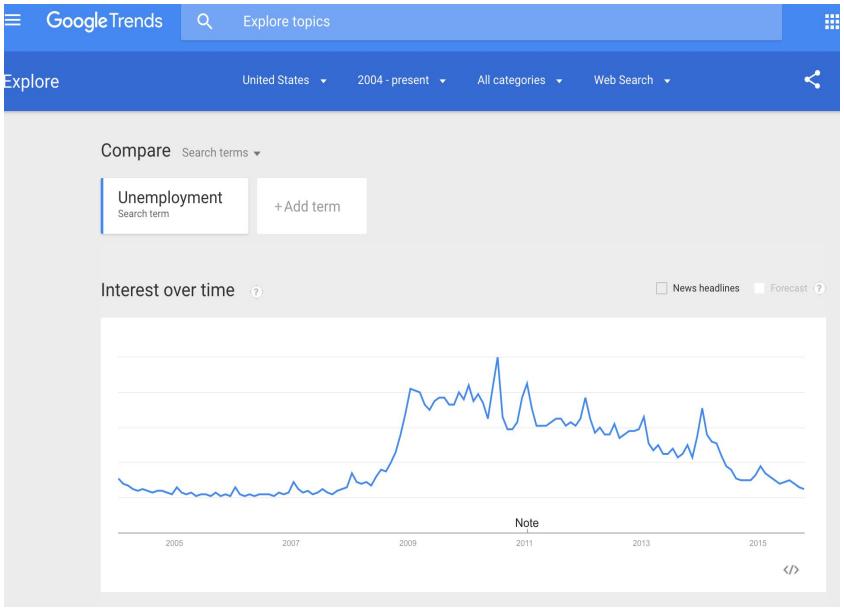
- Model is run for each month
- Workday vs non-workday
- Adding exponents to model non-linear relationship

Weather Impact Dashboard

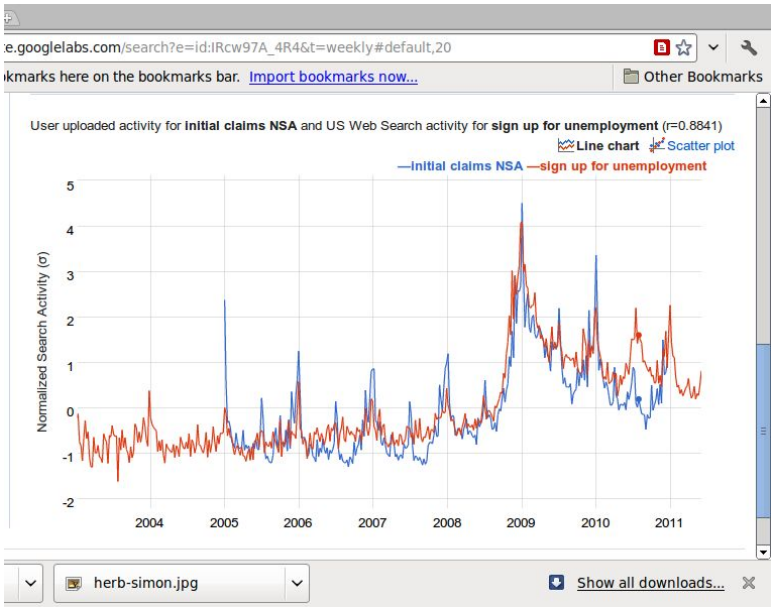


Case B: Predict to Present for Initial Claims

Google Trend

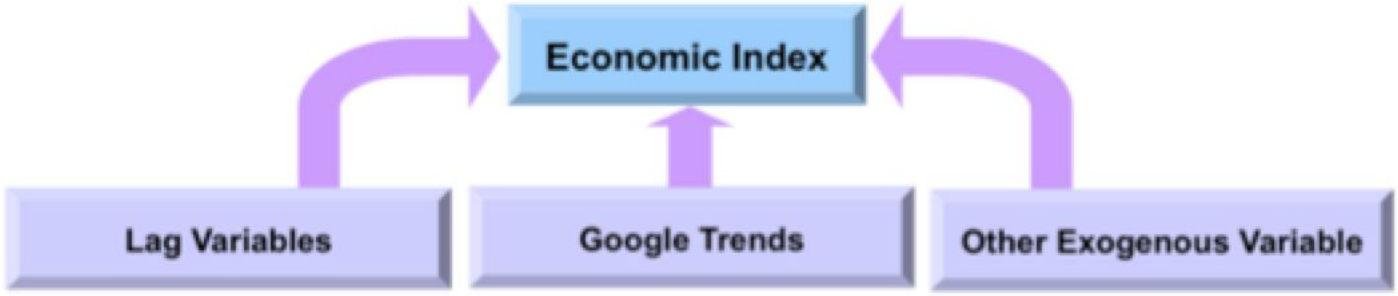


Google Correlate



Can we use related queries to predict/nowcast initial claims?

Predict-to-Present: The Framework and Method



- Use Google correlate to find the query candidates
- Model validation: Use out-of-sample fit to validate that Google trends do better
- Fat regression: Millions of queries, hundreds of categories

Bayesian Structural Time Series (BSTS)

$$y_t = \underbrace{\mu_t}_{\text{trend}} + \underbrace{\gamma_t}_{\text{seasonal}} + \underbrace{\beta^T \mathbf{x}_t}_{\text{regression}} + \epsilon_t$$

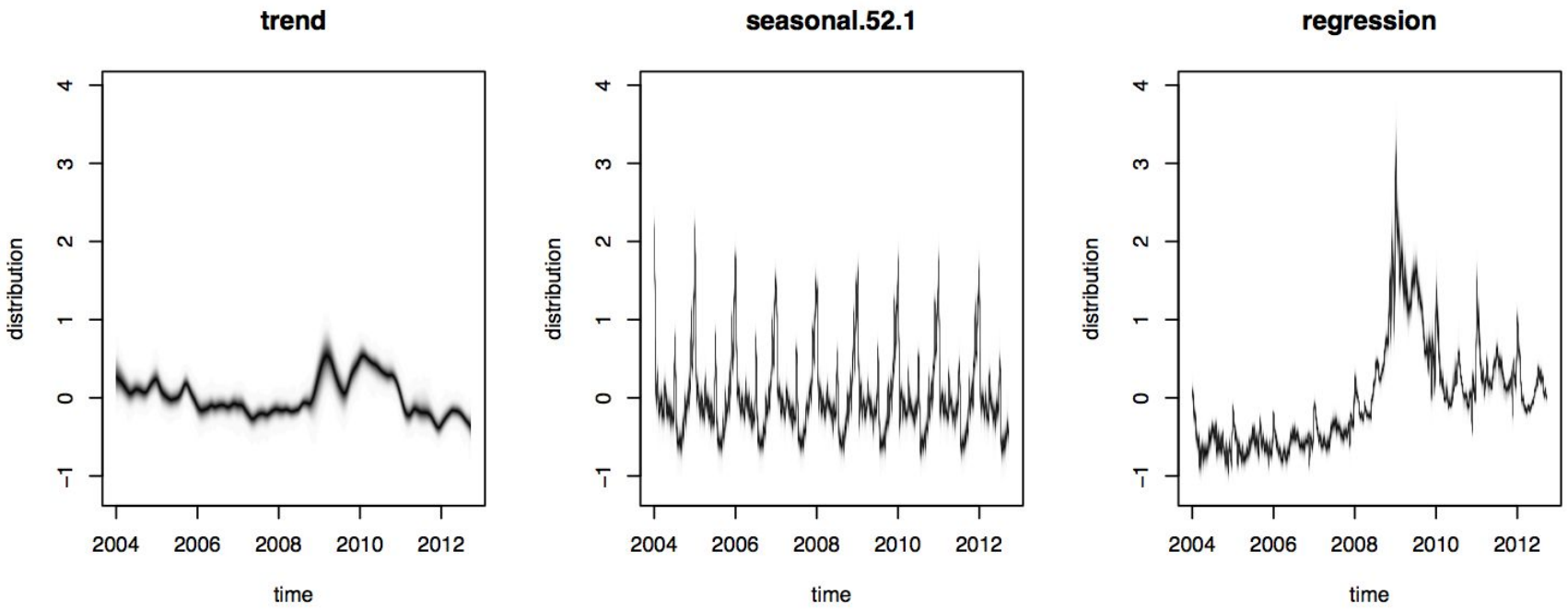
$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t$$

$$\delta_t = \delta_{t-1} + v_t$$

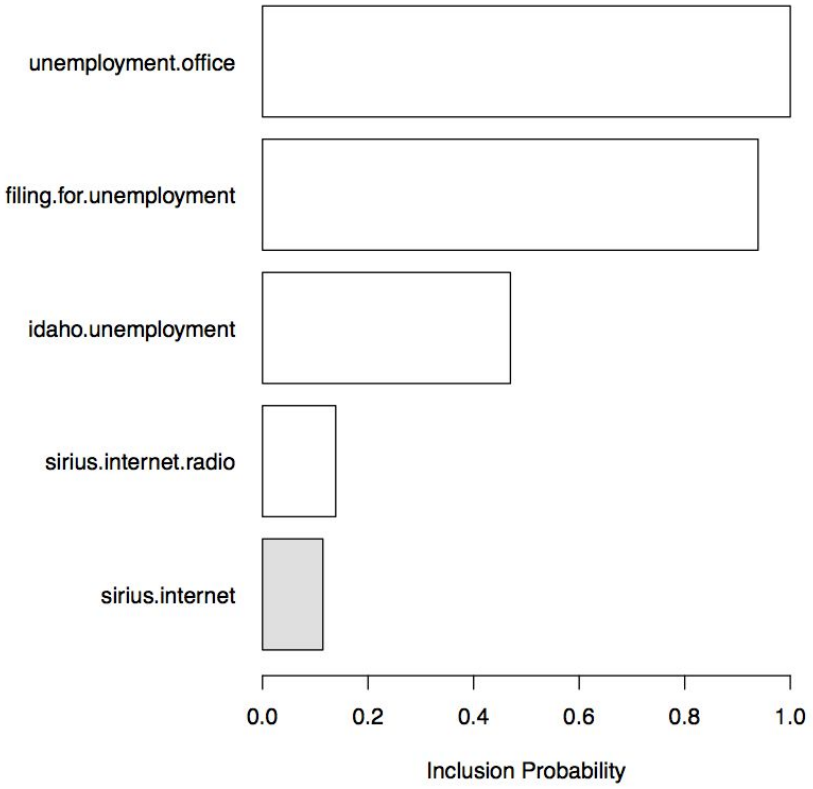
$$\gamma_t = -\sum_{s=1}^{S-1} \gamma_{t-s} + w_t$$

- Decompose time series into trend + seasonality + regression
- Use Kalman filter for trend + seasonality
- Spike and slab regression for variable selection
- Estimate via MCMC for posterior distribution

Decomposition

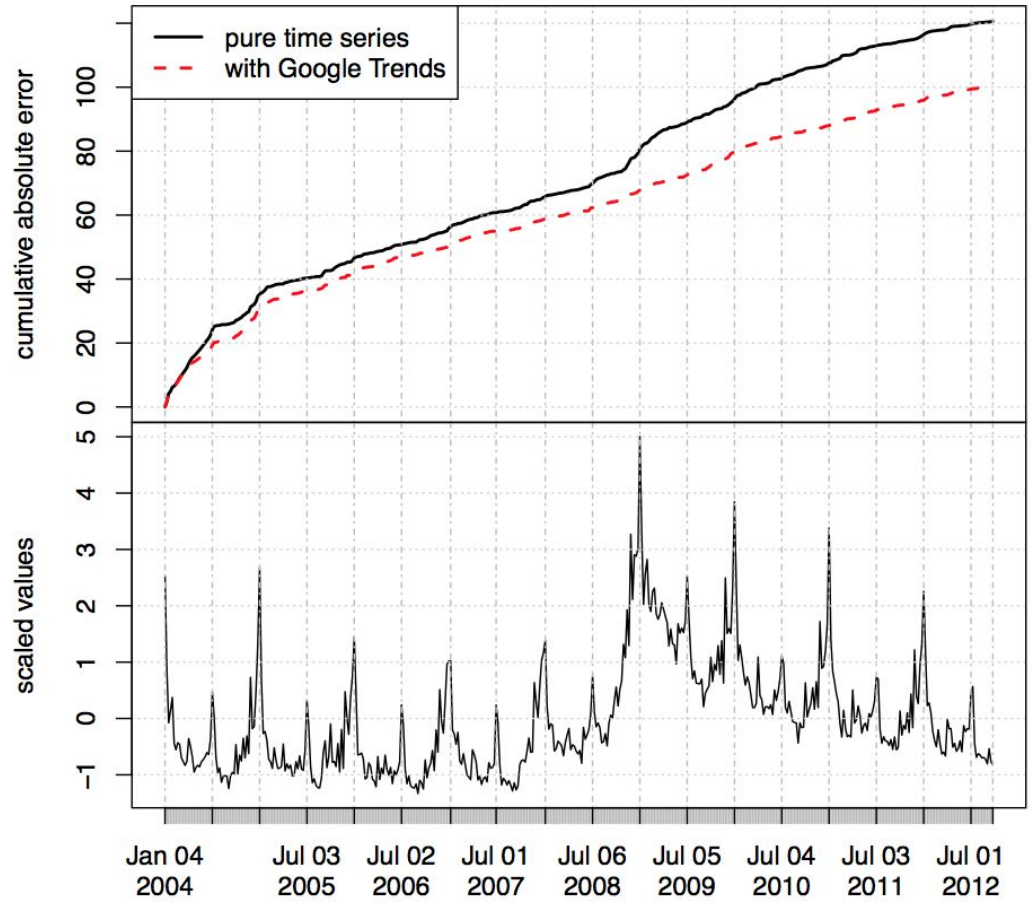


Posterior Inclusion Probabilities



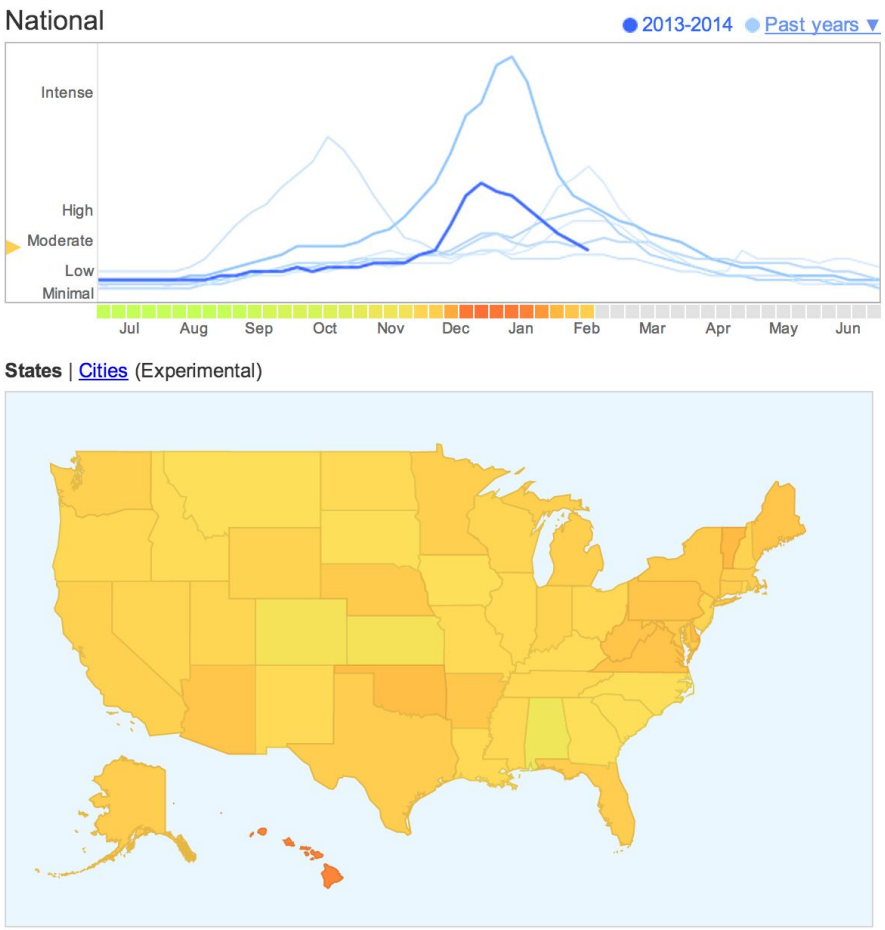
- Showing variables with inclusion probability < 0.1
- White: positive coefficients
- Black: negative coefficients

Performance



- Plot shows one-step ahead prediction error
- The Google query trend did not help much during the normal economic times, but they do after the the big recession started

Case C: Google Flu Trend



Methodology

- Initial query selection
  - Use only health-related queries
  - Use queries that contain a few keywords (flu/fever/etc)
  - Some manual pruning

- Final query selection and combination
  - Lasso (L1 penalty)
  - Elastic Net (L1,L2 combined)
  - BSTS

Penalizing Additional Predictors

$$y = X\beta + \varepsilon$$

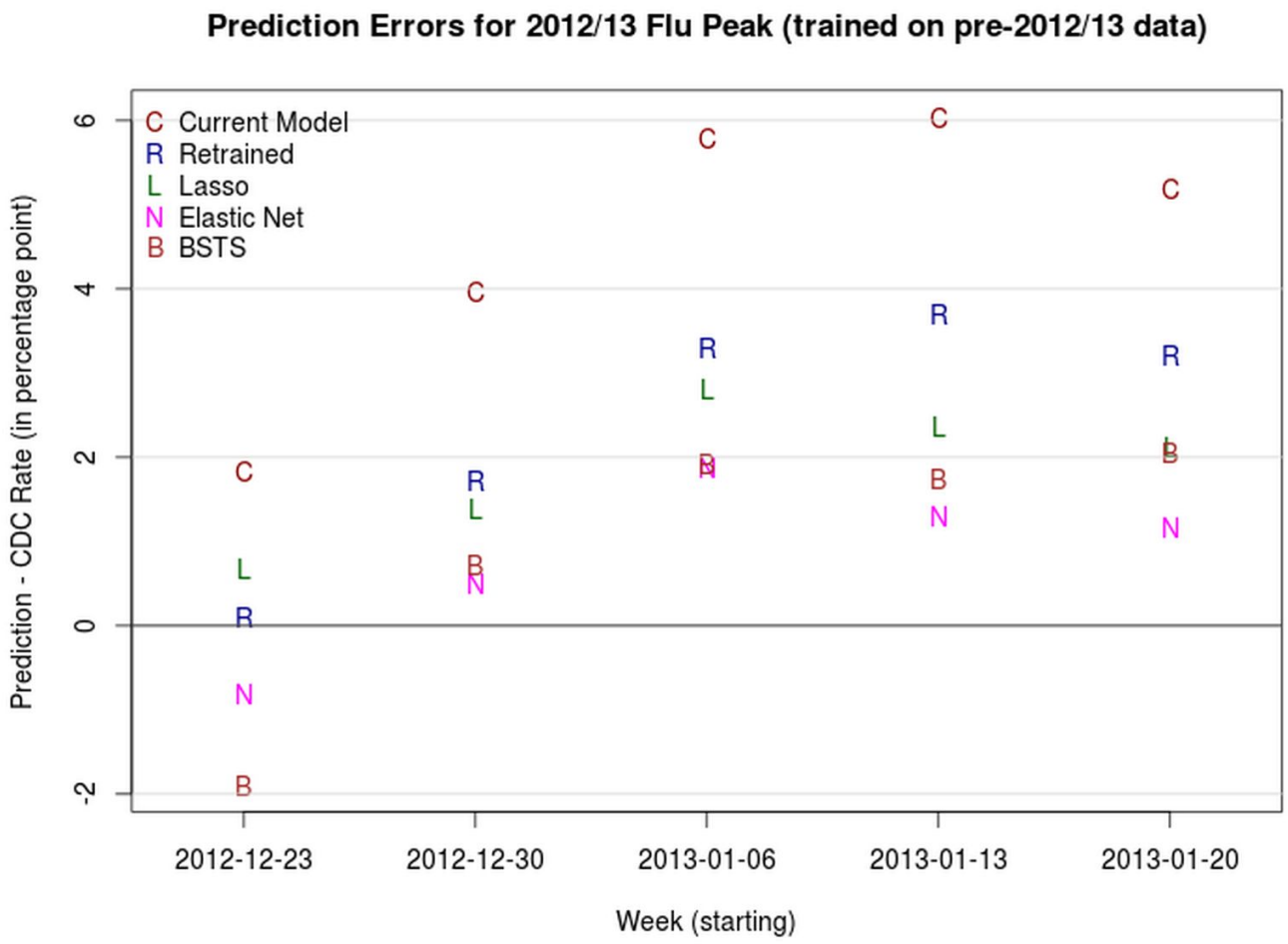
$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1)$$

Techniques to reduce number of predictors and prevent overfitting

- Ridge: $\lambda_1 = 0$
- Lasso: $\lambda_2 = 0$
- Elastic Net

Performance



Big Data, Big Computing Power, Big Opportunities

Parallelization

Bayesianization

Advertising effectiveness and attrition

Causal inference and learning (user/advertiser retention/engagement/acquisition)

Spatial-temporal analysis (map/earth/GPS/auto/loon/fiber)

...