



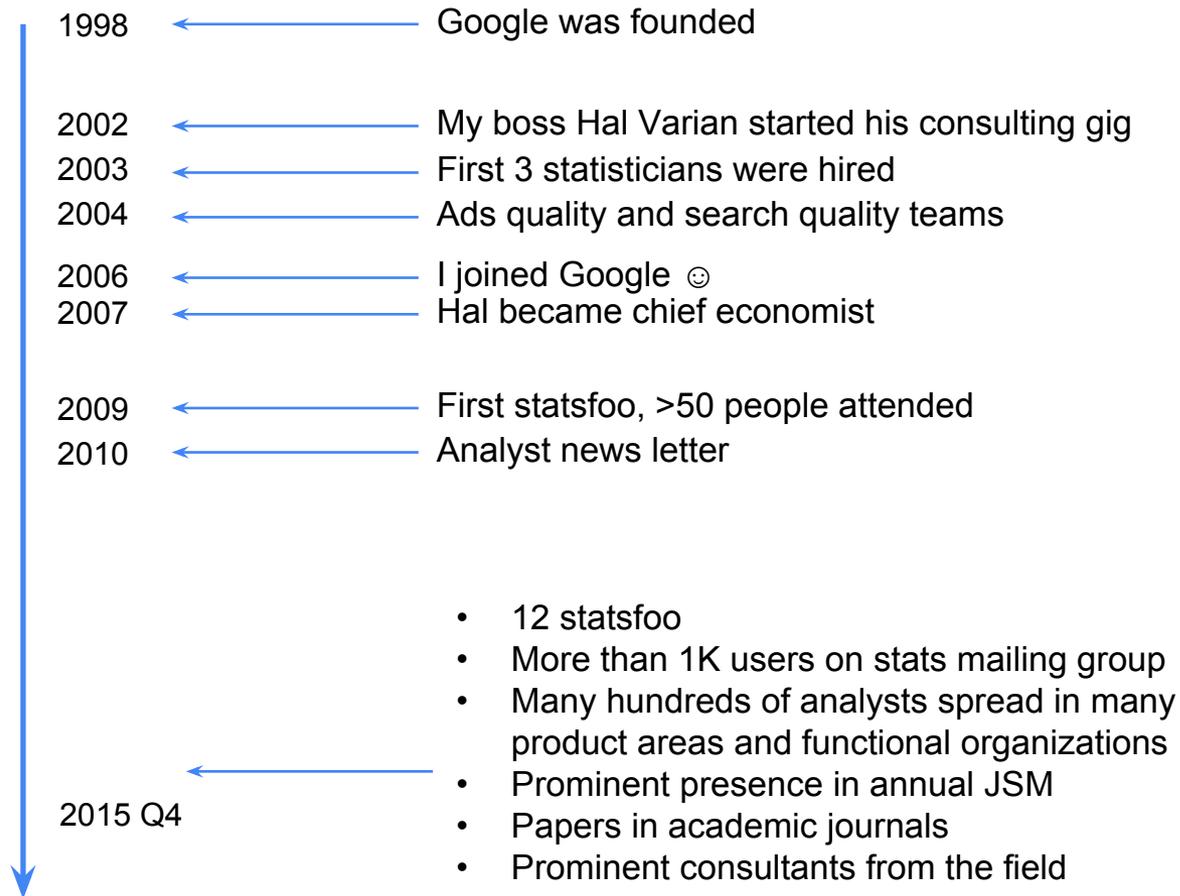
# Statistics at Google

Statistics Day In Puerto Rico

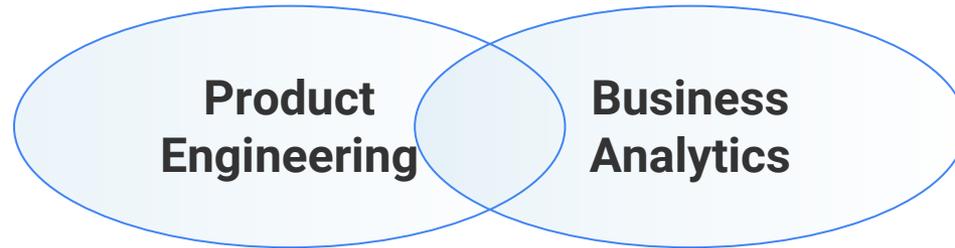
9 October 2015

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?

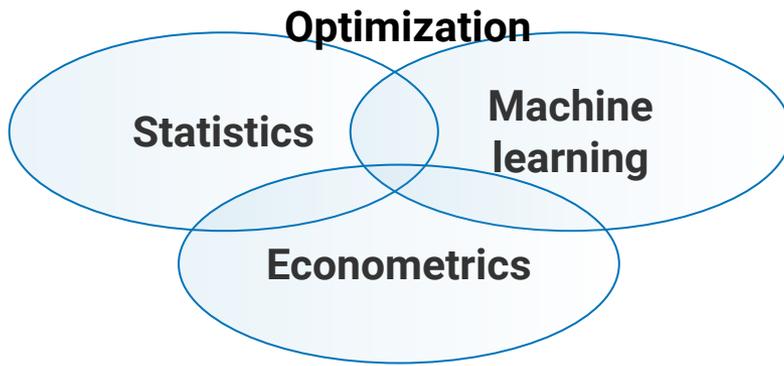
Do advertisers increase spending after adopting a new feature?

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?

Google   

[Web](#) [Maps](#) [Images](#) [News](#) [Shopping](#) [More](#) [Search tools](#)

About 75,600,000 results (0.37 seconds)

**San Juan Airport Hotel - Save Big on this Puerto Rico Hotel**  
**Ad** [www.expedia.com/San\\_Juan\\_Airport\\_Hotel](http://www.expedia.com/San_Juan_Airport_Hotel)   
4.2 ★★★★★ rating for expedia.com  
Save up to 50% on Expedia Hotels.  
No Change/Cancel Fees · Expedia+ Rewards Program  
Ratings: Selection 10/10 - Website 9/10 - Prices 9/10 - Fees 9/10  
Budget Hotels Verified Hotel Reviews  
Luxury Hotels Book Hotel+Flight & Save

**Hotels in Puerto Rico - 20% Discount + \$200 USD Points**  
**Ad** [www.melia.com/Hotels-in-Puerto-Rico](http://www.melia.com/Hotels-in-Puerto-Rico)   
Book Only Through Melia.com!  
Melia Destinations - MeliaRewards - Book Now

**Hotels In San Juan - TripAdvisor.com**  
**Ad** [www.tripadvisor.com/SanJuan](http://www.tripadvisor.com/SanJuan)   
Compare 45 Hotels in San Juan. Quick, Easy, Secure Online Booking.

**Hotels near san juan**  
  Price  Rating  Hotel class 

**The Condado Plaza Hilton**  
3.9 ★★★★★ (124) · 4-star hotel  
Waterfront resort hotel with a casino  
999 Ashford Avenue  
Free Wi-Fi 

**\$149**

**83 Hotels in San Juan**  
[www.trivago.com/Hotel](http://www.trivago.com/Hotel)   
Find Your Ideal Hotel in San Juan!  
Never Pay Full Price on Hotels.

**\$89 San Juan Hotels**  
[www.orbitz.com/San\\_Juan\\_Hotels](http://www.orbitz.com/San_Juan_Hotels)   
4.4 ★★★★★ rating for orbitz.com  
Book Now for Great Rates & Save.  
Plan Your Next Escape to San Juan.

**80 Hotels in San Juan**  
[www.kayak.com/San-Juan-Hotels](http://www.kayak.com/San-Juan-Hotels)   
Great Rates in Seconds.  
Search for Hotel Deals in San Juan.

**Book Room in San Juan**  
[www.airbnb.com/San-Juan](http://www.airbnb.com/San-Juan)   
The Unique Alternative to Hotels  
San Juan Room From \$50/Day!

**100 Hotels in San Juan**

# 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:  $\text{Prob}(\text{click} \mid \text{impr}) \sim X b$

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

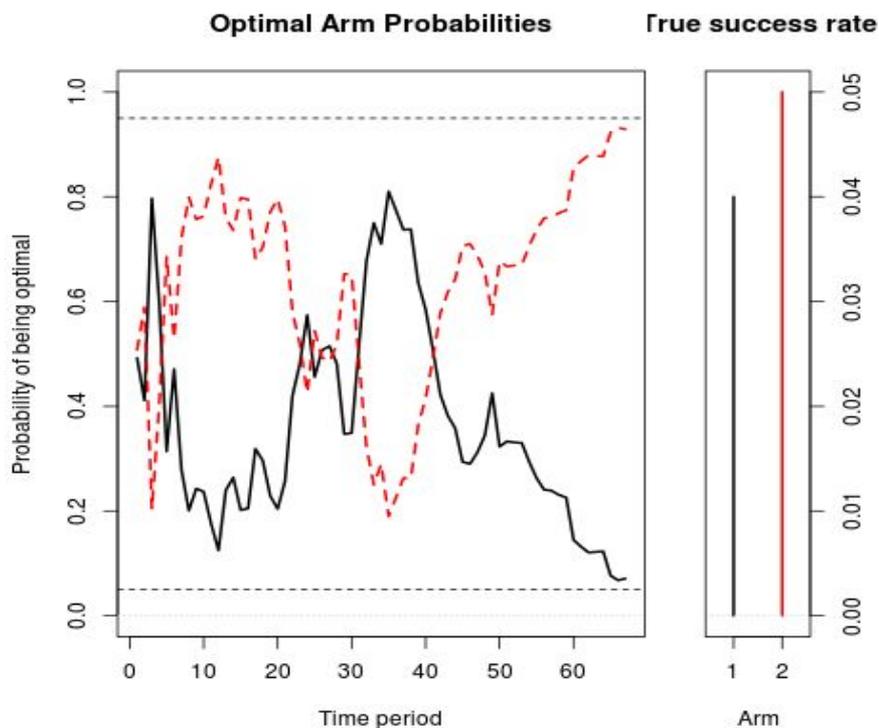
- Experimentation answers the causal questions!
- Experimentation is far easier online.
- Scope of experiments
  - Queries
  - Cookies
  - Geographic
  - Temporal



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

The CHRISTIAN SCIENCE MONITOR

Free Shipping Free Engraving  
Free Gift with Purchase  
100% Money Back Guarantee

Shop clarisonic.com



clarisonic  
The power to be beautiful.

## Tough times? Not for Rolls-Royce.

Rolls-Royce enjoys record sales in 2011. Nearly a third of Rolls-Royces are sold in China.

By Katy Barnato, CNBC Assistant Editor / January 10, 2012



An employee dusts off a sold Rolls-Royce "Onyx" before a customer picks it up at a Rolls-Royce showroom in Dubai Monday. Upmarket UK-based carmaker Rolls-Royce, owned by BMW, said it sold a record 3,530 cars in 2011, benefiting from a surge in demand for luxury cars from customers in emerging markets. Jumana El Hebeueh/Reuters

Enlarge

6 19

Recommend Tweet



Rolls-Royce Motor Cars sales hit record-breaking levels in 2011, with more cars sold than ever before in the brand's 107-year history.

Over 3,500 of the ultra-luxury cars were sold, representing a 31 percent rise on 2010...

Answer a question to continue reading this page

Should the U.S. Government raise the debt ceiling?

I don't know, show me another question

Share the page you're reading

Like **6** people like this. Be the first of your friends.

Learn more - Privacy

GET EMAIL UPDATES SIGN UP  
Sign up for daily email updates from CSMonitor.com

Understanding the basics of your credit score can help you score the best mortgage RELOCATION.COM

Answer a question to continue reading this page

Should the U.S. Government raise the debt ceiling?

I don't know, show me another question

Share the page you're reading

Like **6** people like this. Be the first of your friends.

Learn more - Privacy

- Most
- 1.
  - 2.
  3. Gov. Scott Walker vs. unions, Wisconsin recall petition (video)
  4. Zappos hacked: 24 million accounts exposed
  5. Concordia cruise ship hit by iceberg, goes back (+ video)

Photos of the day



# 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!

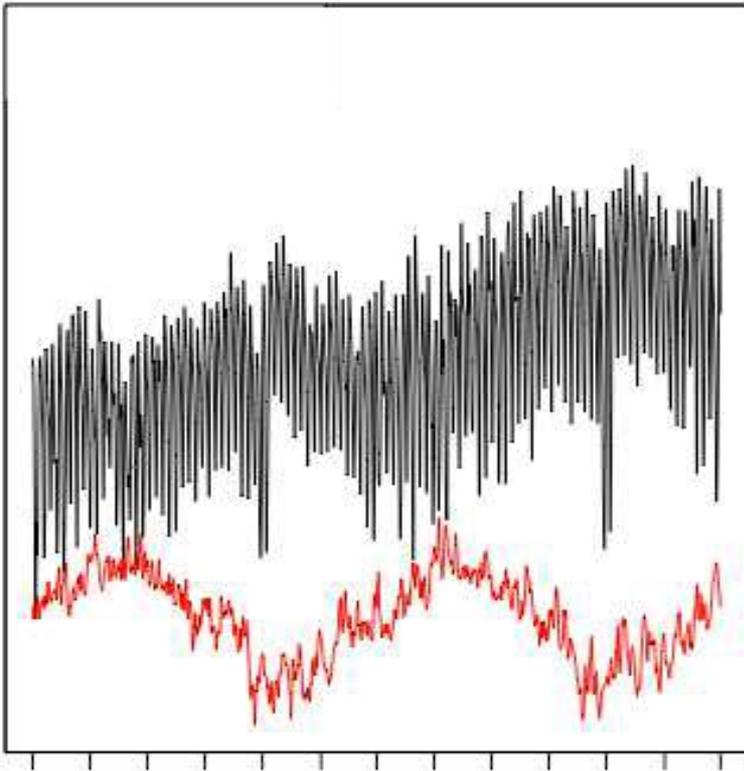
# Showcasing Variable Selections for Fat Regressions

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)

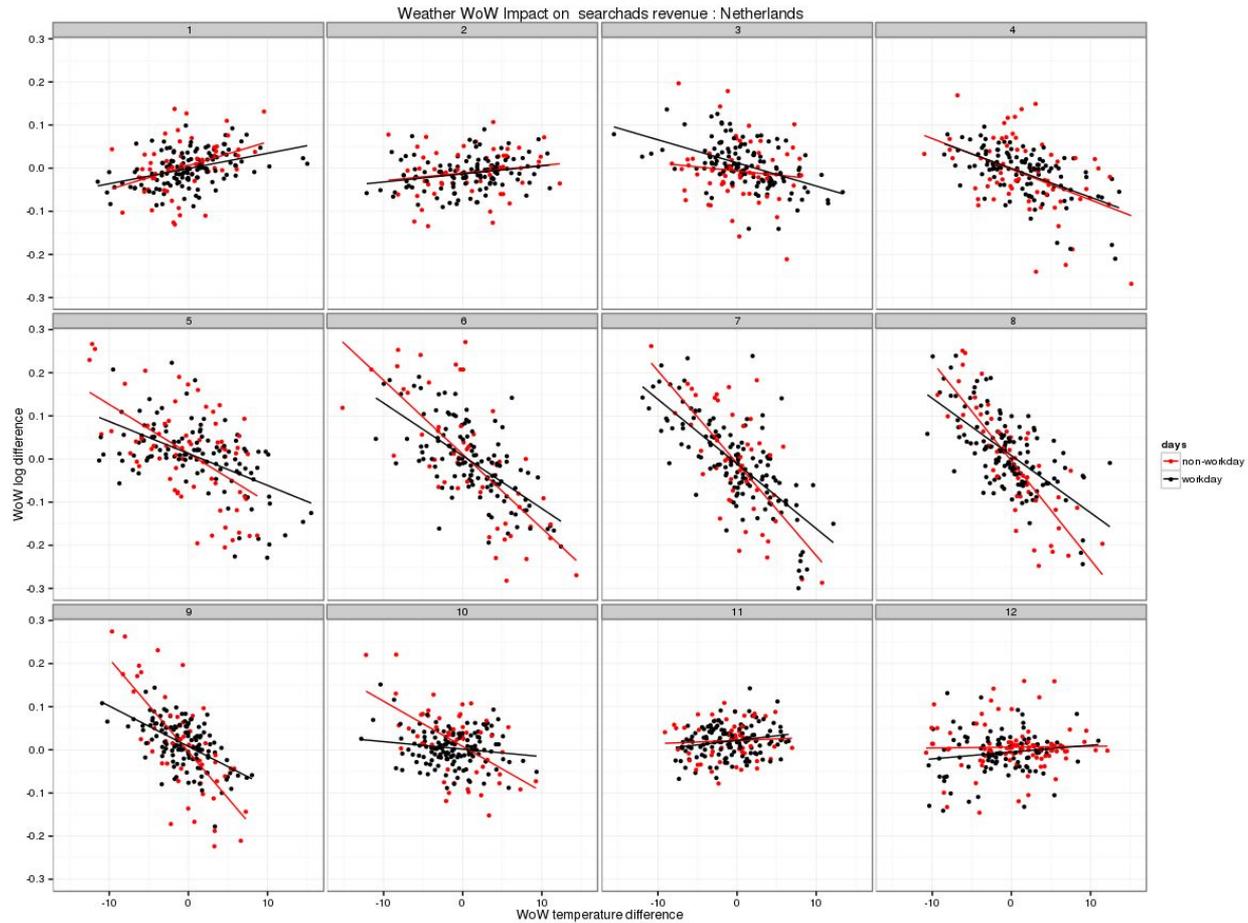
# Case A: Weather's Impact on Search Traffic

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

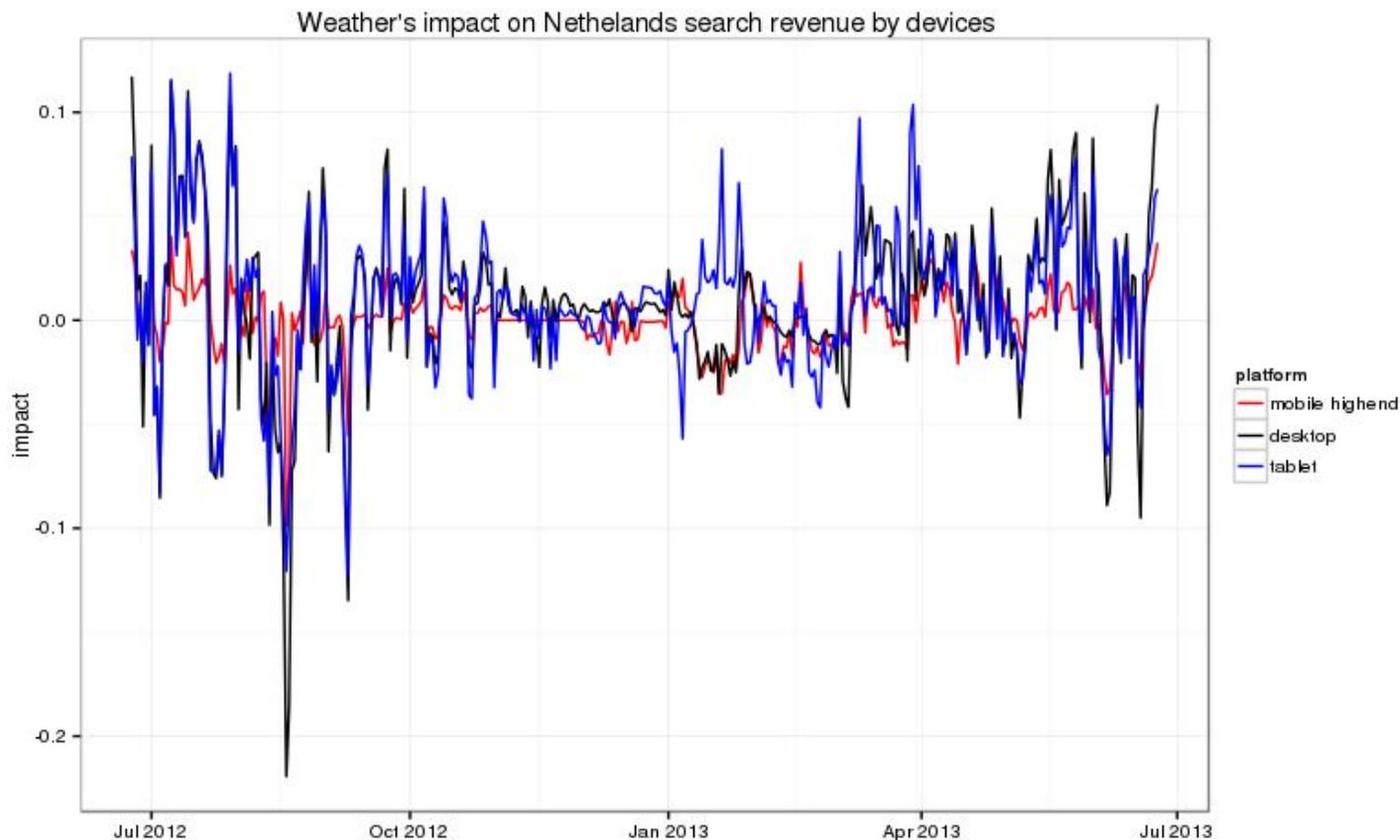


# Simple Step-Wise Regression Model

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

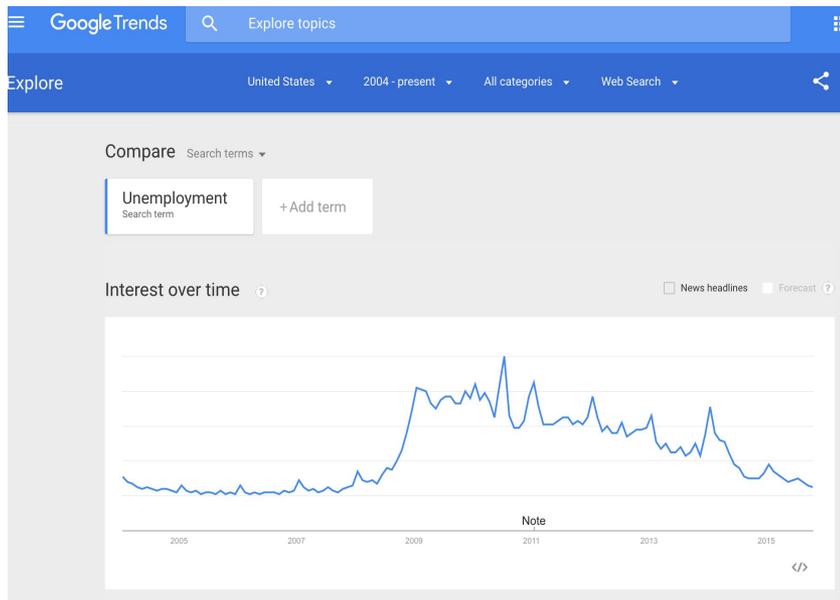
- Model is run for each month
- Workday vs non-workday
- Adding exponents to model non-linear relationship

# Weather Impact Dashboard

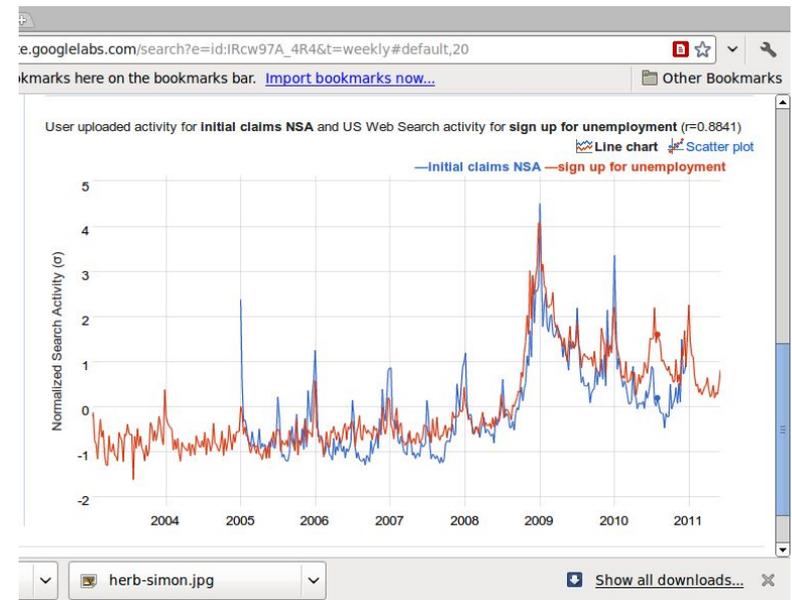


# Case B: Predict to Present for Initial Claims

## 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)

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

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

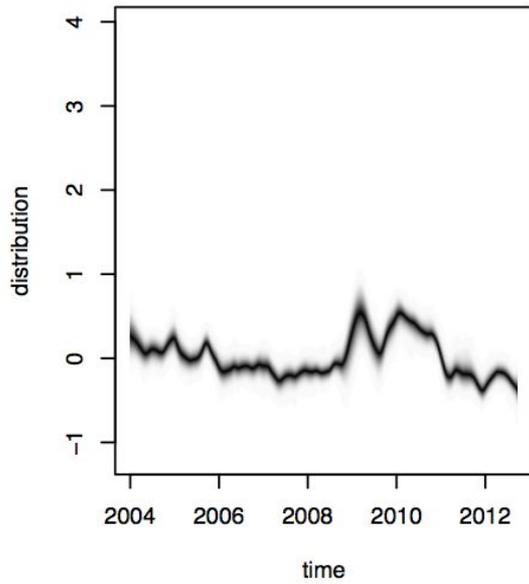
$$\delta_t = \delta_{t-1} + v_t$$

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

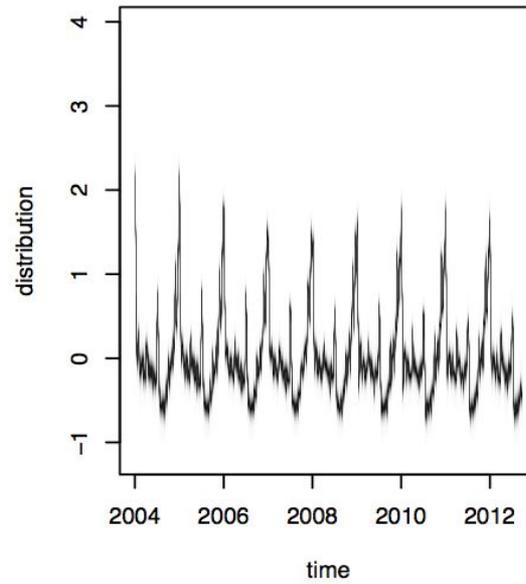
- Decompose time series into trend + seasonality + regression
- Use Kalman filter for trend + seasonality
- Spike and slab regression for variable selection
- Estimate via MCMC for posterior distribution

# Decomposition

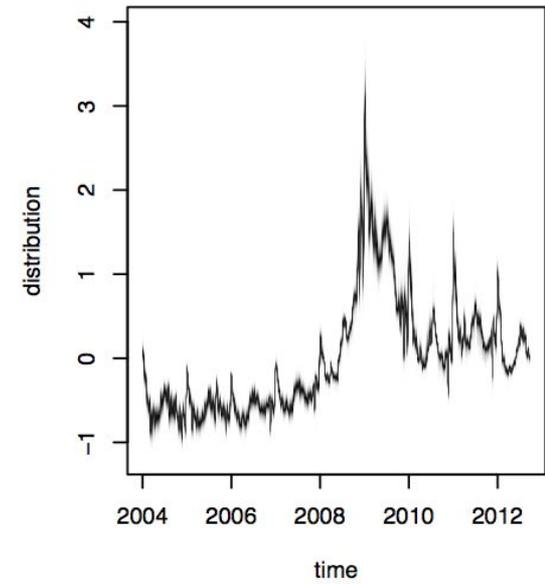
**trend**



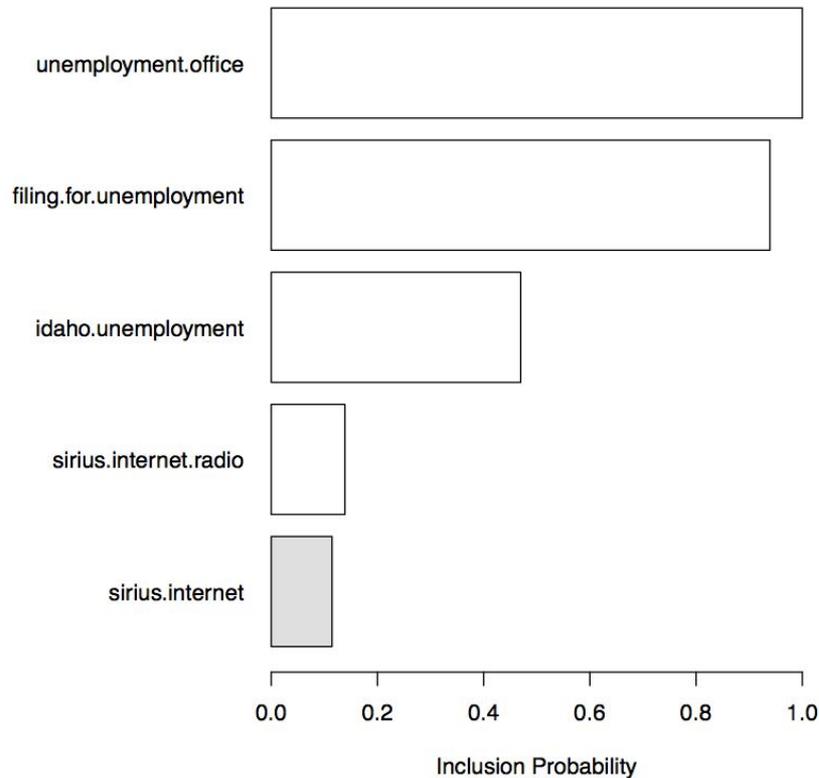
**seasonal.52.1**



**regression**

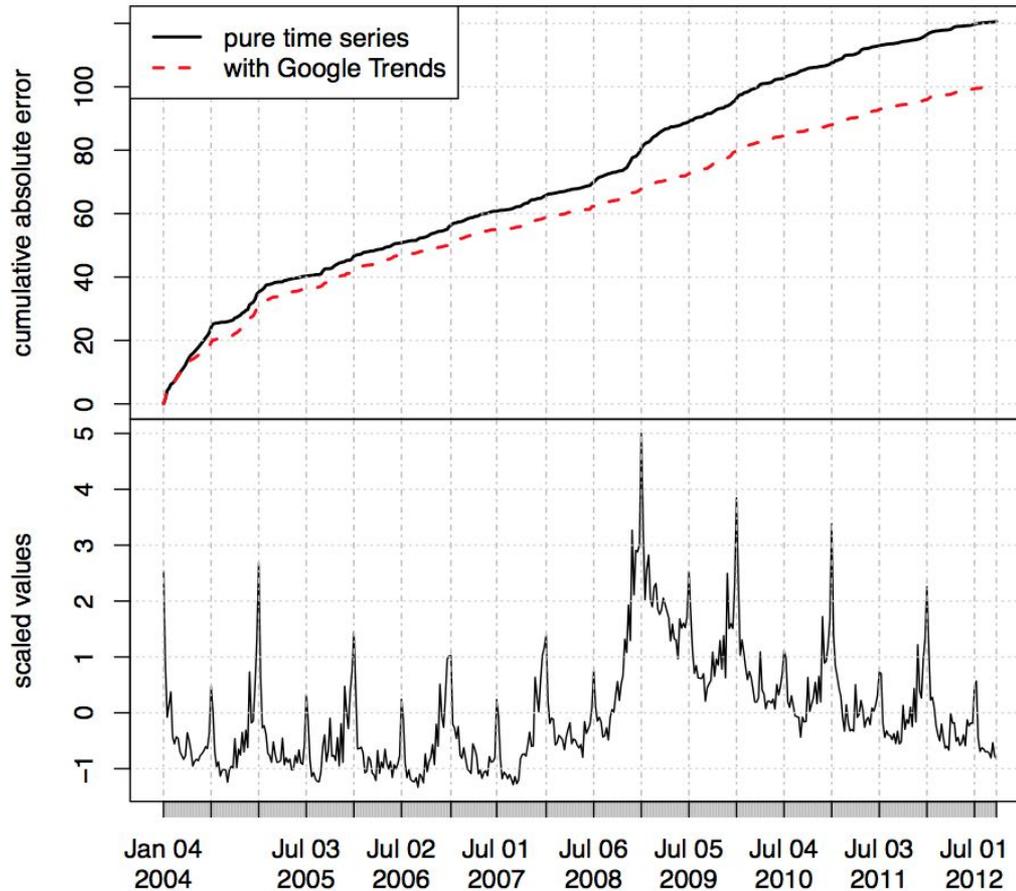


# Posterior Inclusion Probabilities



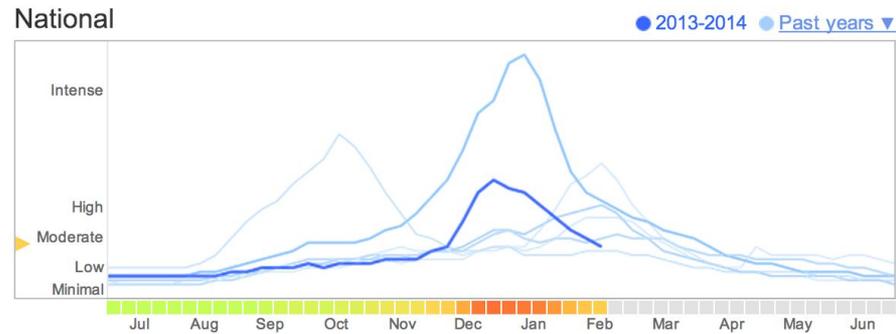
- Showing variables with inclusion probability  $< 0.1$
- White: positive coefficients
- Black: negative coefficients

# Performance

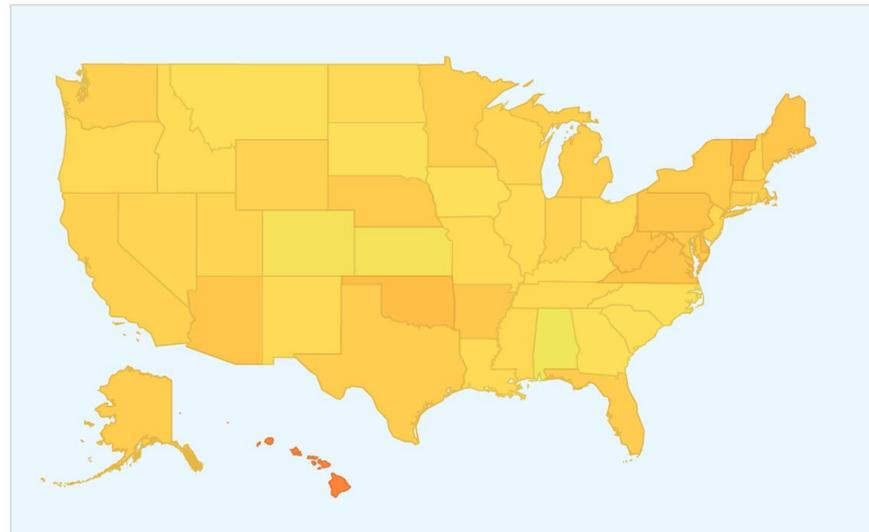


- Plot shows one-step ahead prediction error
- The Google query trend did not help much during the normal economic times, but they do after the the big recession started

# Case C: Google Flu Trend



States | [Cities](#) (Experimental)



# 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

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

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

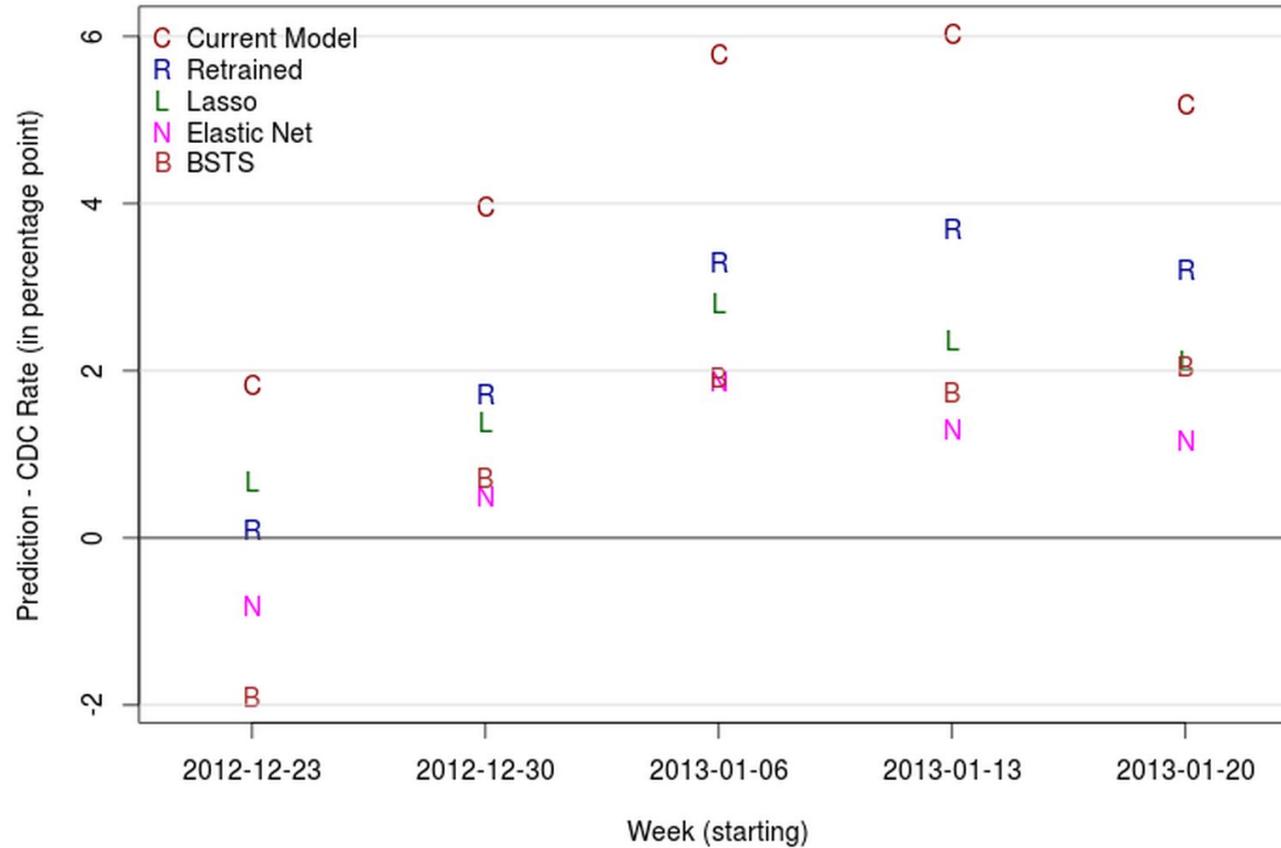
$$\hat{\beta} = \operatorname{argmin}_{\beta} (\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1)$$

Techniques to reduce number of predictors and prevent overfitting

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

# Performance

Prediction Errors for 2012/13 Flu Peak (trained on pre-2012/13 data)



# 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)**

...