Big Data and Economics, Big Data and Economies

Susan Athey, Stanford University

Disclosure: The author consults for Microsoft.

Lenses on big data

1. The science and practice of using big data
2. Management and organization in the face of big data
3. Implications for industries and economies

Machine Learning Meets Econometrics: New Tools and Methods for Big Data

New Tools and Methods Required

- Getting Data
  - Large & huge scale data extraction & manipulation
  - Scraping, crawling
  - Crowd-sourcing
  - Using APIs & firehoses
- Analyzing Data
  - Parallel computing
  - Cloud computing
  - Text mining
  - Classification
  - Clustering
  - Network analysis
  - Machine learning tools

Supervised Machine Learning v. Econometrics/Statistics Lit. on Causality

Supervised ML

- Well-developed and widely used nonparametric prediction methods that work well with big data
  - Used in technology companies, computer science, statistics, genomics, neuroscience, etc.
  - Rapidly growing in influence
- Cross-validation for model selection
- Focus on prediction and applications of prediction
- Weaknesses
  - Causality (with notable exceptions, including those attending this conference)

Econometrics/Soc Sci/Statistics

- Formal theory of causality
  - See, e.g., Imbens and Rubin book (2015)
- "Structural models" that predict what happens when world changes
  - Used for auctions, anti-trust (e.g. mergers) and business decision-making (e.g. pricing)
- Well-developed and widely used tools for estimation and inference of causal effects in exp. and observational studies
  - Used by social science, policy-makers, development organizations, medicine, business, experimentation
- Weaknesses
  - Non-parametric approaches fail with many covariates
  - Model selection unprincipled
  - Emphasis on standard errors for pre-specified models

A Research Agenda on Causal Inference with Big Data

Problems

- Many problems in social sciences entail a combination of prediction and causal inference
- Existing ML approaches to estimation, model selection and robustness do not directly apply to the problem of estimating causal parameters
- Inference more challenging for some ML methods

Proposals

- Formally model the distinction between causal and predictive parts of the model and treat them differently for both estimation and inference
  - Abadie, Athey, Imbens and Wooldridge (2014, WIP)
- Develop new estimation methods that combine ML approaches for prediction component of models with causal approaches
  - Athey-Imbens (WIP)
- Develop new approaches to cross-validation optimized for causal inference
  - Athey-Imbens (2015, WIP)
- Develop robustness measures for causal parameters inspired by ML
  - Athey-Imbens (*AER* 2015)

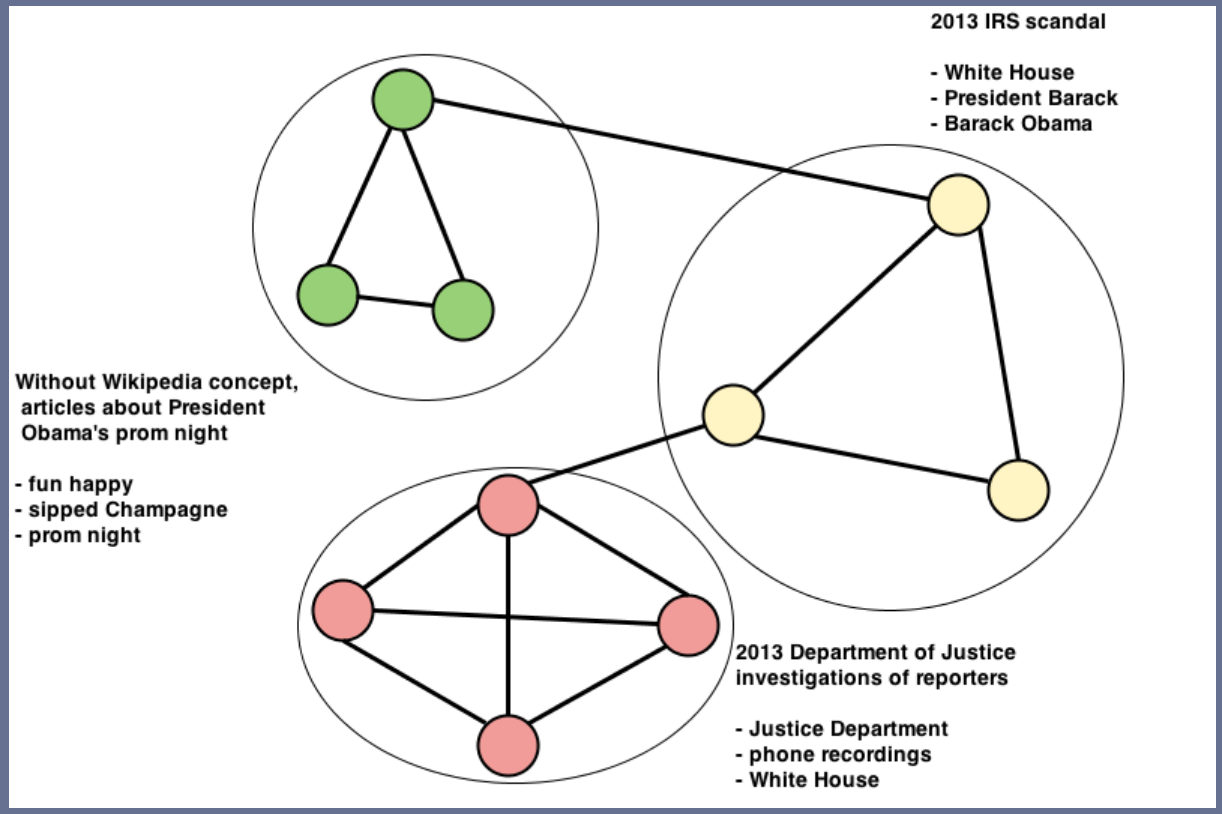
Big Data Methodology: Unsupervised Learning and Network Modeling

- Experimentation on Networks
  - Individuals in a social network are not independent observations
  - Sellers on eBay, advertisers on Google compete
- How do you optimally design experiments and conduct inference?
  - Kleinberg, Ugander KDD
  - Eckles et al
  - Athey, Eckles and Imbens (2015)
- Clustered experimentation
  - Use community detection to find groups with lots of links inside and few outside

Unsupervised learning

- Find related objects, e.g. cats on YouTube
- Find collections of text that are similar

Network Community Detection Example



Topic Examples

Boston Marathon bombings
2013 Korean crisis
2012–13 in English football

Ariel Castro kidnappings

2013 NFL season

2013 in baseball

2013 in American television
2013 NBA Finals

2013 in film

2012–13 NHL season

2013 Masters Tournament

2013 NASCAR Sprint Cup Series

2013 Moore tornado
Death of Lee Rigby
2012 Benghazi attack

Murder of Travis Alexander
List of school shootings in the United States

Death and funeral of Margaret Thatcher

Timeline of the Syrian civil war (May–August 2012)
2013 IRS scandal

NCAA Men's Division I Basketball Championship

Malaysian general election, 2013

Shooting of Trayvon Martin

Phil Mickelson

Using Big Data for Economics Research

Examples

- Public
  - Online News
  - Twitter and Social Media
  - Web Corpus
  - Scraping eBay
  - Wikipedia
  - Reviews
  - Prices and price comparison
  - Google Trends
  - Real estate data
- Restricted Access
  - Scanner data
  - Credit card transactions
  - Facebook behavior
  - Online browsing
  - GPS data
  - Internet search & advertising
  - Health
  - Energy utilization
  - Cities data
    - Transportation
    - Housing
    - Environment
    - Crime
    - Education

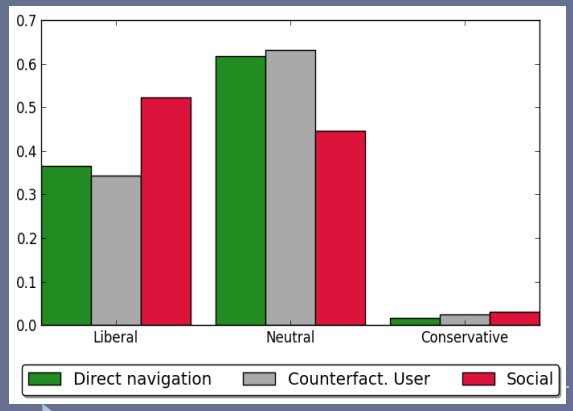
Applied research program in big data

- How does internet and social media impact news? (toolbar data)
  - Aggregators: Athey & Mobius (2012, in progress)
  - Athey, Mobius & Pal (in progress)
  - Athey, Blei, Hoffman, and Mobius (2015): large scale discrete choice
- Do ticker lookups predict stock returns? (toolbar data)
  - What can we learn about individual investors and what drives their behavior?
  - Athey & DellaVigna (in progress)
- Advertiser objectives and behavior in search advertising
  - Structural model to predict behavior (Athey & Nekipelov, 2012)
  - Estimating heterogeneous objectives and response (Athey & Nekipelov, in progress)
- Networks
  - Analytics on the Bitcoin Blockchain as a Network (Athey, Parashkevov, & Xia, 2015)
  - Network effects and adoption (MIT Bitcoin Experiment, Athey, Cattalini, Chandrasekhar, Tucker)

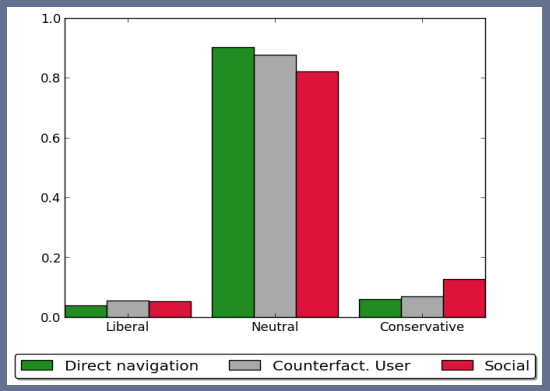
Social Media and Polarization: Evidence from Browsing

- Medium-big data (Gentzkow & Shapiro, 2011): little evidence that internet is associated with polarization
- Micro browsing data (Athey, Mobius & Pal, in progress):
  - Look within internet for social media v. direct navigation
  - Classify articles by topic using text mining, then use Mechanical Turk workers on a sample of articles:
    - Whether topic is *potentially* polarized, and bias of article
  - Adjust for user mix (users equally weighted in both SM & non-SM)

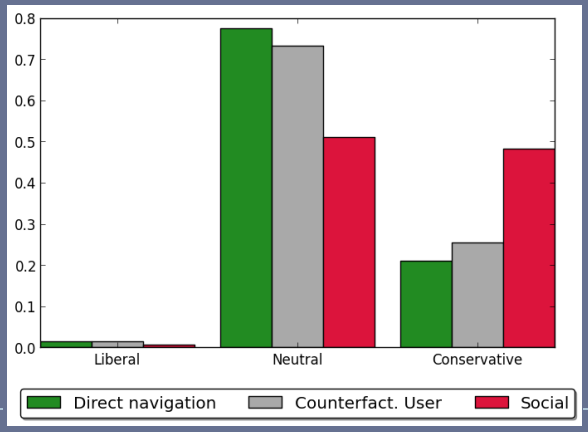
Liberal users



All Users



Conservative users



How much power does a search engine have to manipulate?

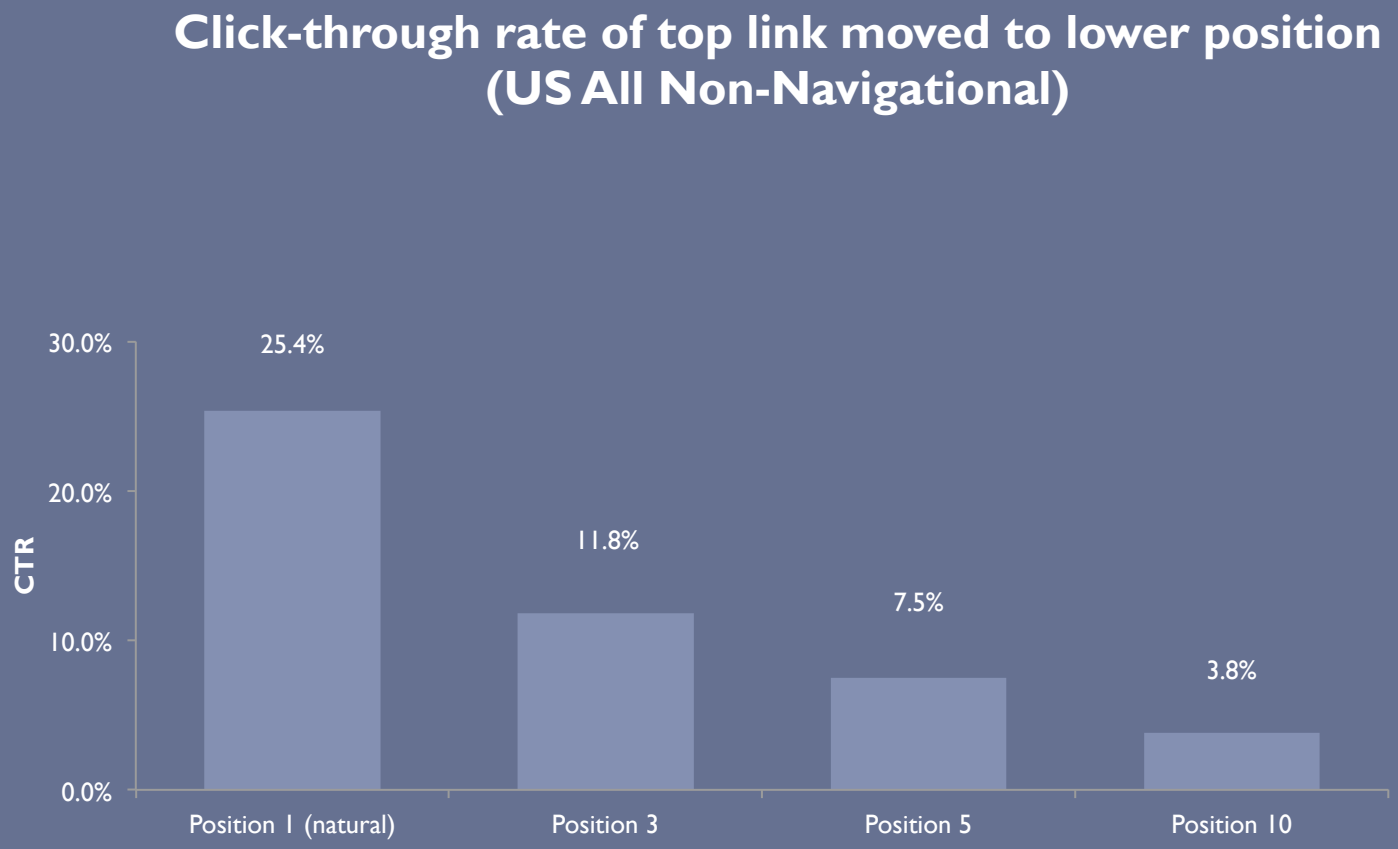
- Big data, but observational: Seabright et al (2012)
  - Look at natural variation in algorithmic results for a few keywords
  - Quantifies impact
- Large-scale experimentation: Athey (2012), Athey and Imbens (2015)
  - Randomize the rankings for a subset of users
  - Measure results
  - Analyze treatment effect heterogeneity
- Repurposing experiments
  - Experiments as instruments: Athey (2011)

Experimental Design for Search Results

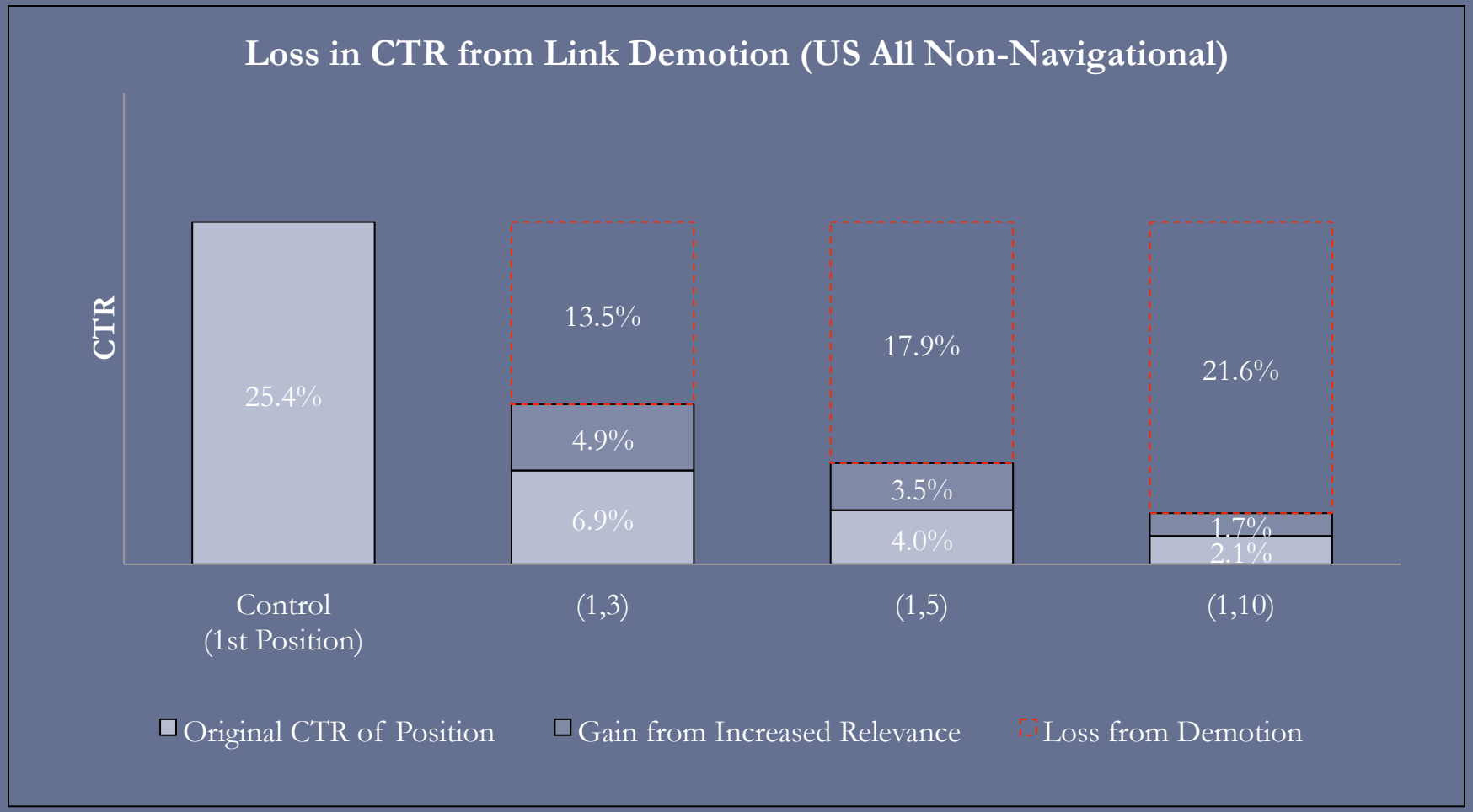
- Can experiment on users or page views
- Take, say, 1% of page views and swap position A with position B
- Compare metrics for control and treatment group
- For algo links, exclude "navigational" queries
- Example metrics:
  - Clicks
  - "Good" clicks
  - Click quality

  - See: http://blogs.technet.com/b/microsoft_on_the_issues/archive/2013/03/25/the-importance-of-search-result-location.aspx

Swapping Positions of Algo Links



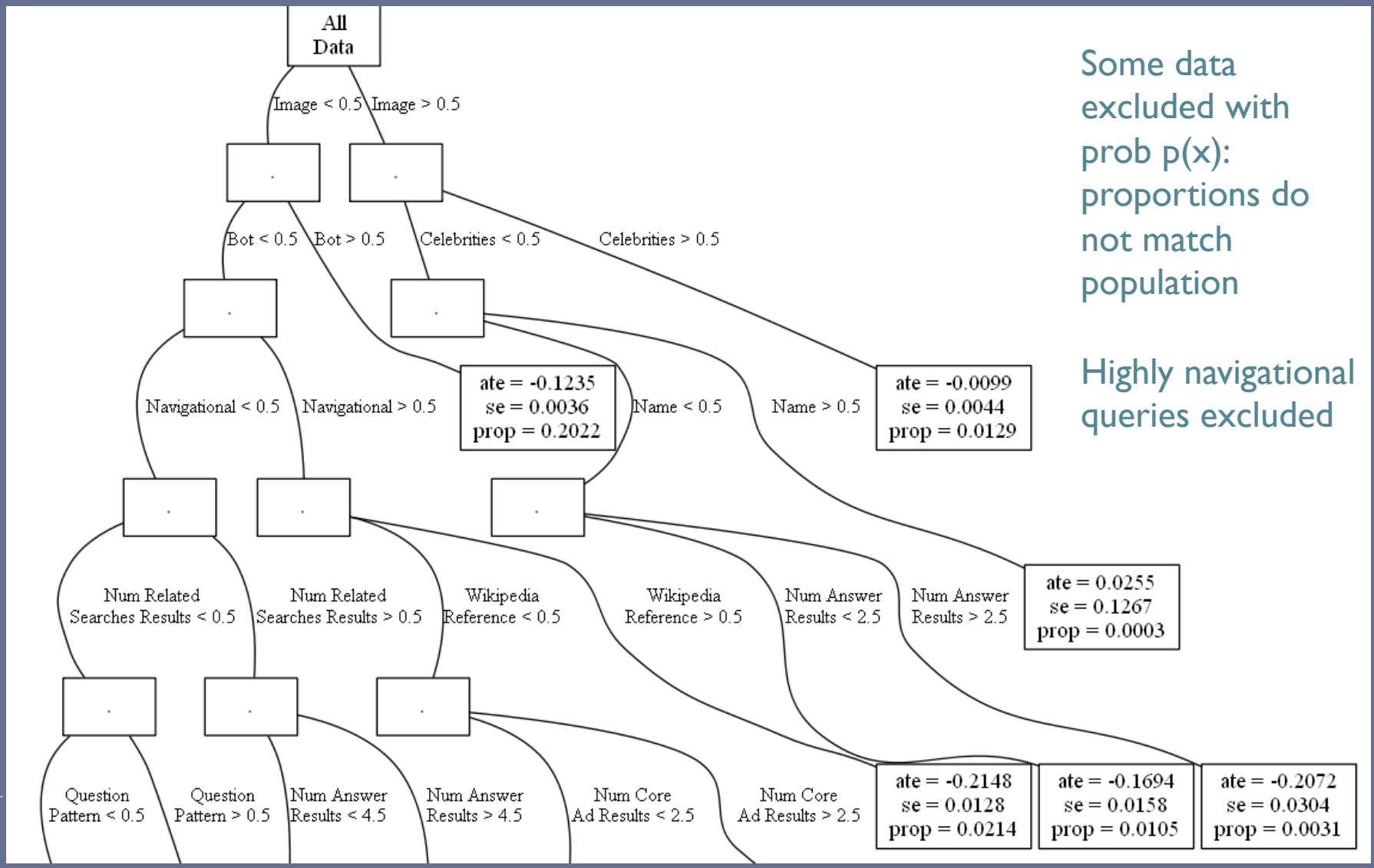
Relevance v. Position



Problem: Treatment Effect Heterogeneity in Estimating Position Effects in Search

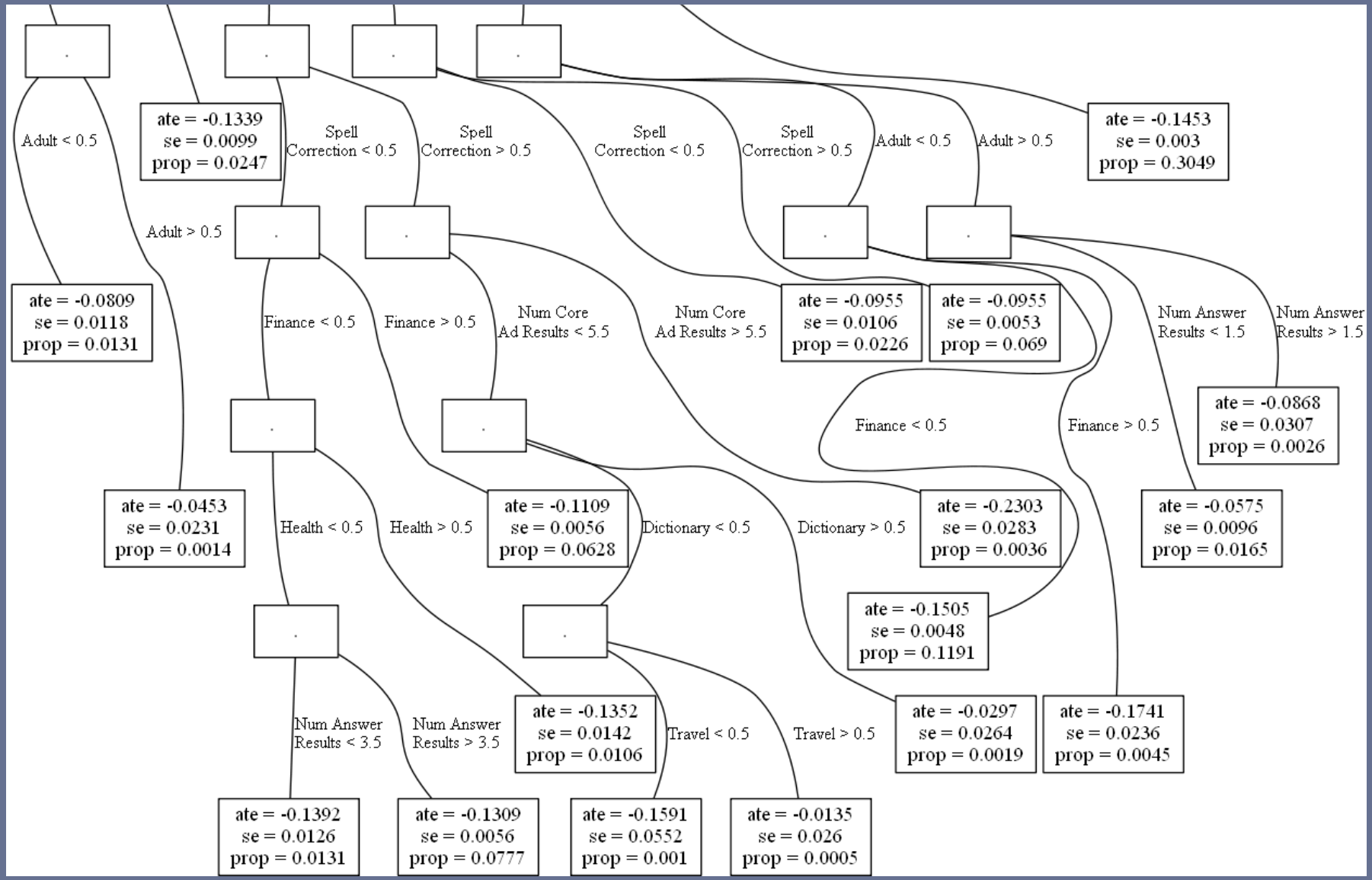
- Queries highly heterogeneous
  - Tens of millions of unique search phrases each month
  - Query mix changes month to month for a variety of reasons
  - Behavior conditional on query is fairly stable
- Desire for segments.
  - Want to understand heterogeneity and make decisions based on it
  - "Tune" algorithms separately by segment
  - Want to predict outcomes if query mix changes
    - For example, bring on new syndication partner with more queries of a certain type

Search Experiment Tree: Effect of Demoting Top Link (Test Sample Effects)



Some data excluded with prob p(x): proportions do not match population

Highly navigational queries excluded


Adult < 0.5
ate = -0.0809
se = 0.0118
prop = 0.0131
ate = -0.1339
se = 0.0099
prop = 0.0247
Adult > 0.5
ate = -0.0453
se = 0.0231
prop = 0.0014
Spell Correction < 0.5
Spell Correction > 0.5
Finance < 0.5
Finance > 0.5
Health < 0.5
Health > 0.5
Num Answer Results < 3.5
Num Answer Results > 3.5
ate = -0.1392
se = 0.0126
prop = 0.0131
ate = -0.1309
se = 0.0056
prop = 0.0777
Num Core Ad Results < 5.5
Num Core Ad Results > 5.5
ate = -0.1109
se = 0.0056
prop = 0.0628
ate = -0.1352
se = 0.0142
prop = 0.0106
Dictionary < 0.5
Dictionary > 0.5
Travel < 0.5
Travel > 0.5
ate = -0.1591
se = 0.0552
prop = 0.001
ate = -0.0135
se = 0.026
prop = 0.0005
Spell Correction < 0.5
Spell Correction > 0.5
ate = -0.0955
se = 0.0106
prop = 0.0226
ate = -0.2303
se = 0.0283
prop = 0.0036
ate = -0.1505
se = 0.0048
prop = 0.1191
ate = -0.0297
se = 0.0264
prop = 0.0019
Adult < 0.5
Adult > 0.5
ate = -0.0955
se = 0.0053
prop = 0.069
Finance < 0.5
Finance > 0.5
ate = -0.1741
se = 0.0236
prop = 0.0045
ate = -0.1453
se = 0.003
prop = 0.3049
Num Answer Results < 1.5
Num Answer Results > 1.5
ate = -0.0575
se = 0.0096
prop = 0.0165
ate = -0.0868
se = 0.0307
prop = 0.0026

Implications for Business and Economies

Business and Management

- Data driven
  - A/B testing
  - Analytics
  - Scarcity of talent that combines conceptual skills (economics, business), statistics, and computing prowess
- Data as a strategic asset
- Economies of scale
  - R&D
  - Machine learning

Economies

- Aggregators & intermediaries
  - Mobile: payments and information
  - News & information
  - Business deals and competition
- Competition/concentration
- Micro labor issues
  - Skills and training
- Macro labor issues
  - Cloud computing, IOT, Automation
  - Economies of scale & capability
  - Concentration of jobs and profits

Appendix

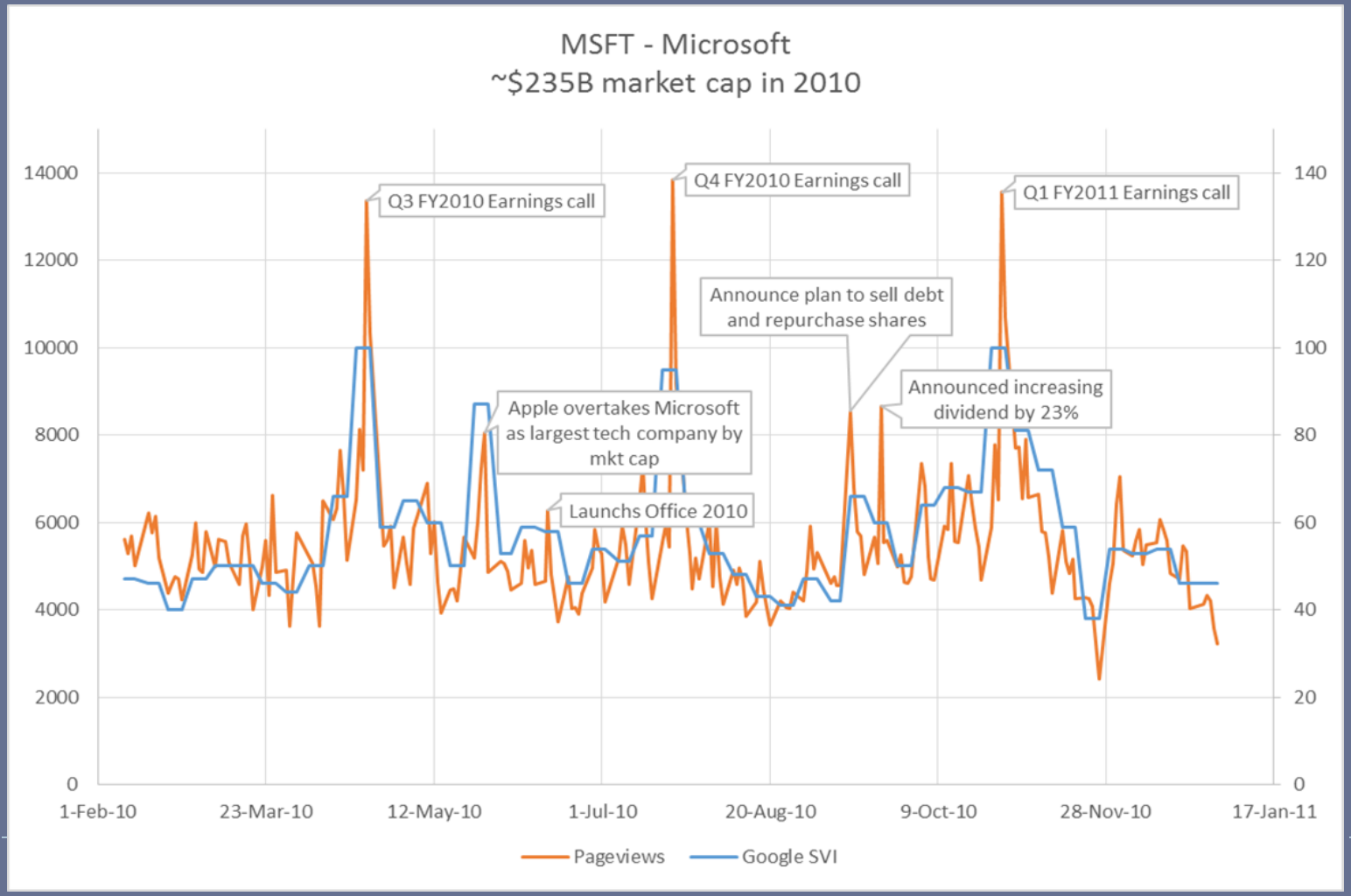
Information and Stock Prices

- Do searches or ticker lookups predict stock returns?
- What kinds of research patterns predict stock returns?
- What kinds of informational events drive ticker lookups and thus stock prices?

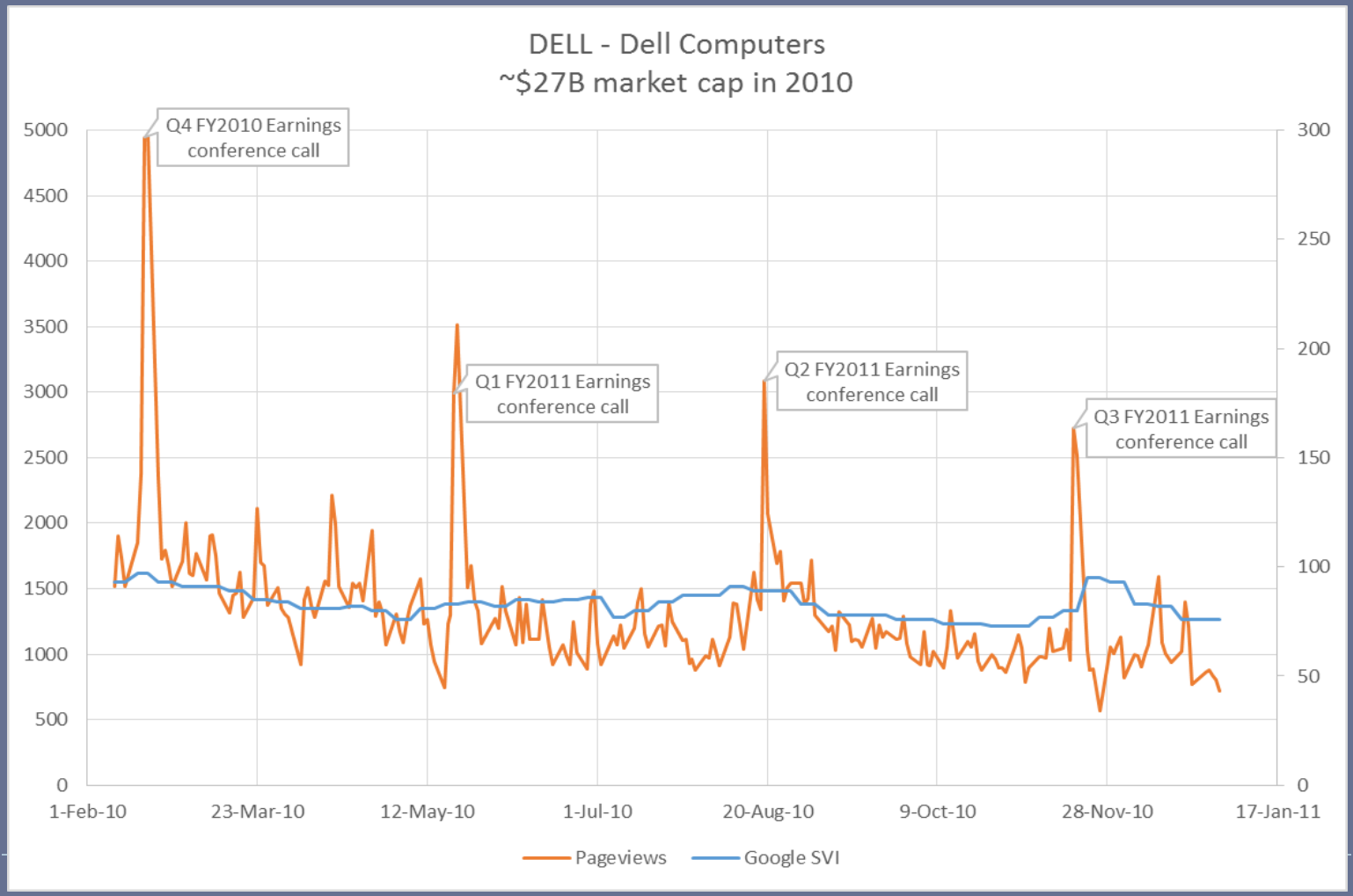
- Medium big data (Google trends):
  - Modest but noisy evidence of relationship between stock prices and searches

- Athey & DellaVigna (in process):
  - Individual investor interest pushes up stock prices temporarily
  - Can construct profitable trading strategies

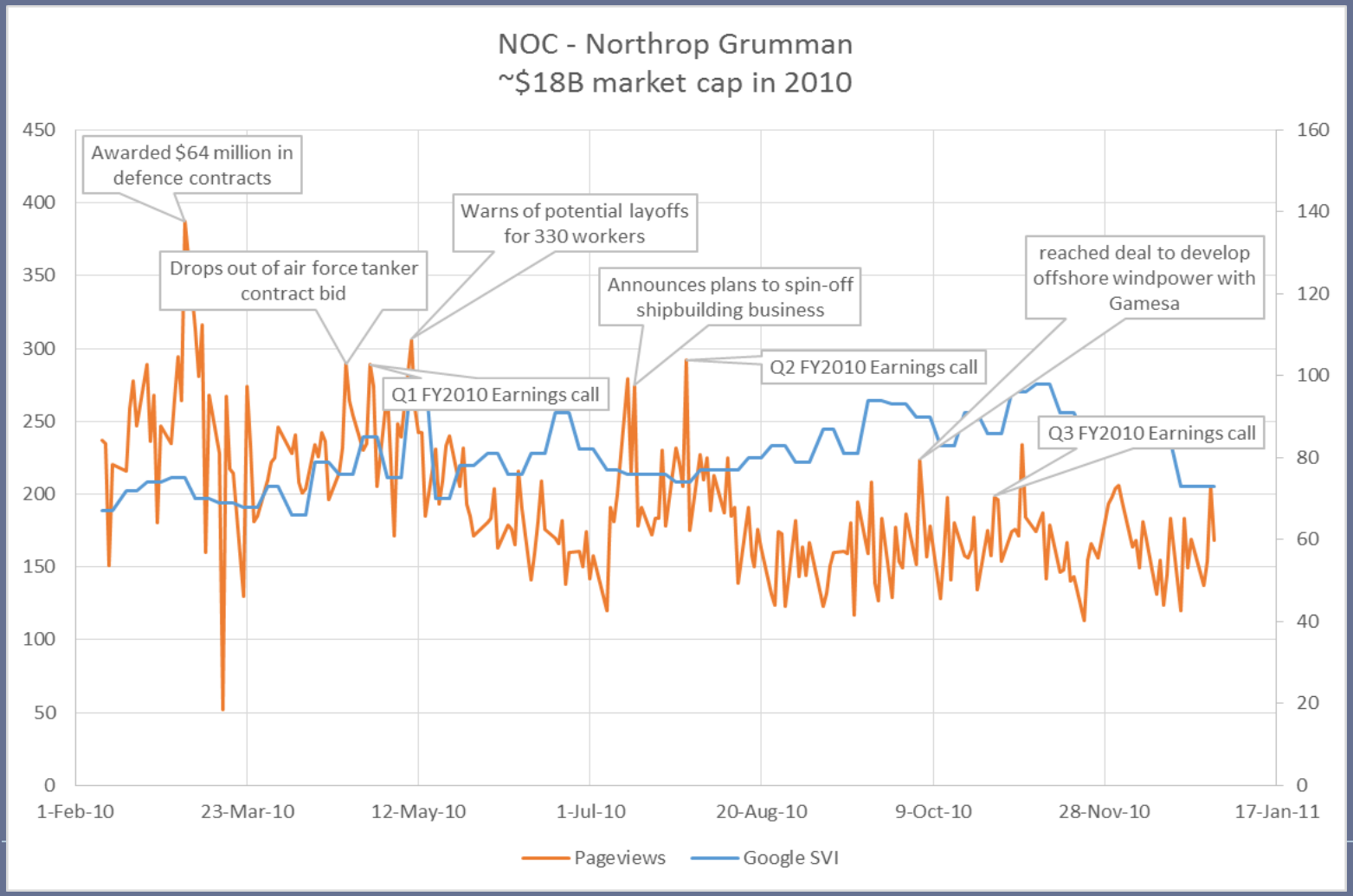
Example of Lookup Series



Example of Lookup Series



Example of Lookup Series



Example of Lookup Series