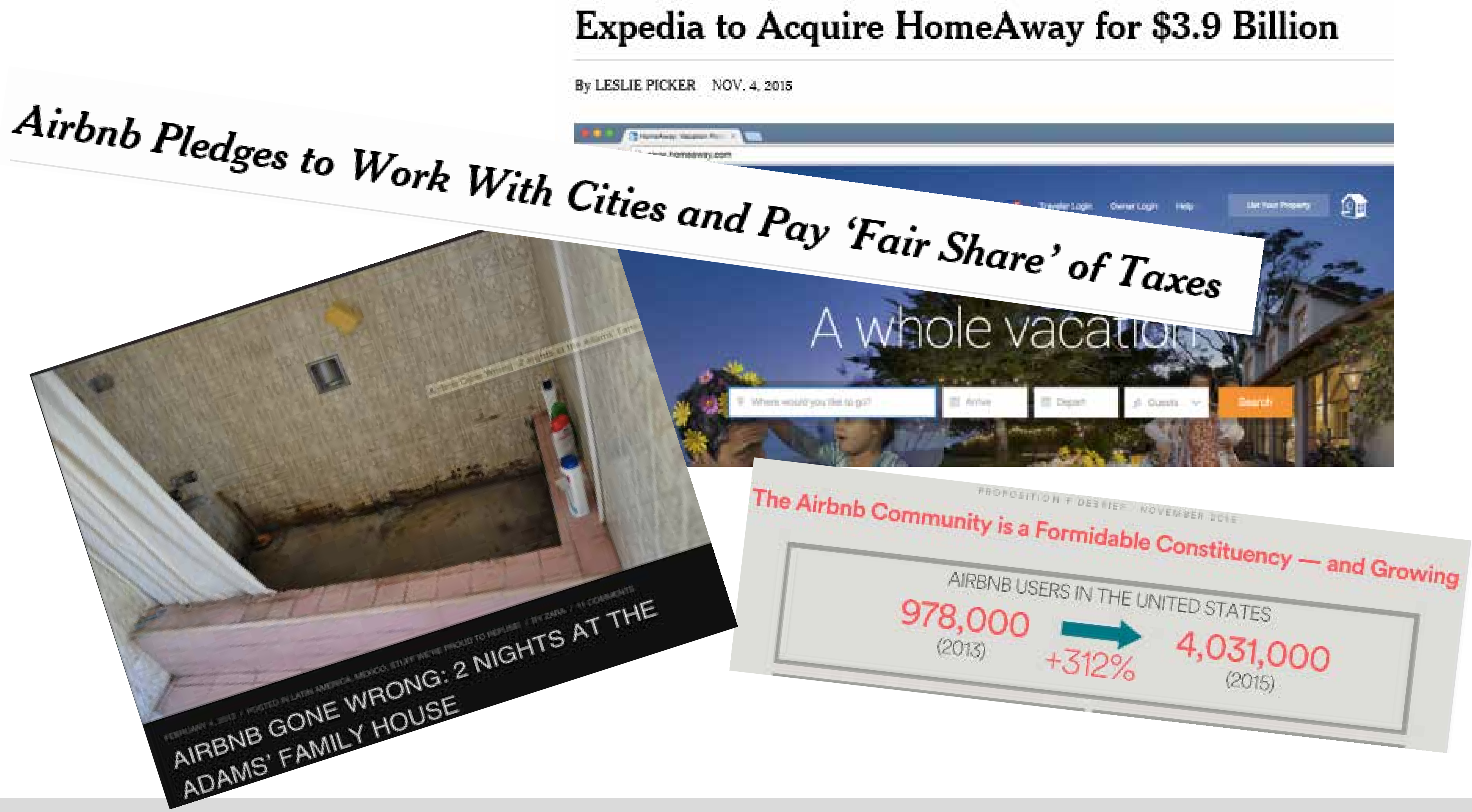
STANFORD GRADUATE SCHOOL OF BUSINESS

The New Economics of Data

Susan Athey
The Economics of Technology Professor

Change lives. Change organizations. Change the world. ge the world.

Marketplaces in the News



On-Demand Services

THE WALL STREET JOURNAL.

http://www.wsj.com/articles/theres-an-uber-for-everything-now-1430845789

TECH | PERSONAL TECH | PERSONAL TECHNOLOGY

There’s an Uber for Everything Now

Apps do your chores: shopping, parking, cooking, cleaning, packing, shipping and more

ILLUSTRATION: ROBERT NEUBECKER

Technology platforms

TECH

Facebook, Amazon and Other Tech Giants Tighten Grip on Internet Economy

Online search, messaging, advertising, applications, computing and storage are delivered on demand

The Cloud Is Raining Cash on Amazon, Google, and Microsoft

Each company's impressive earnings can be attributed to a shift in the industry that's punishing a slew of legacy firms.

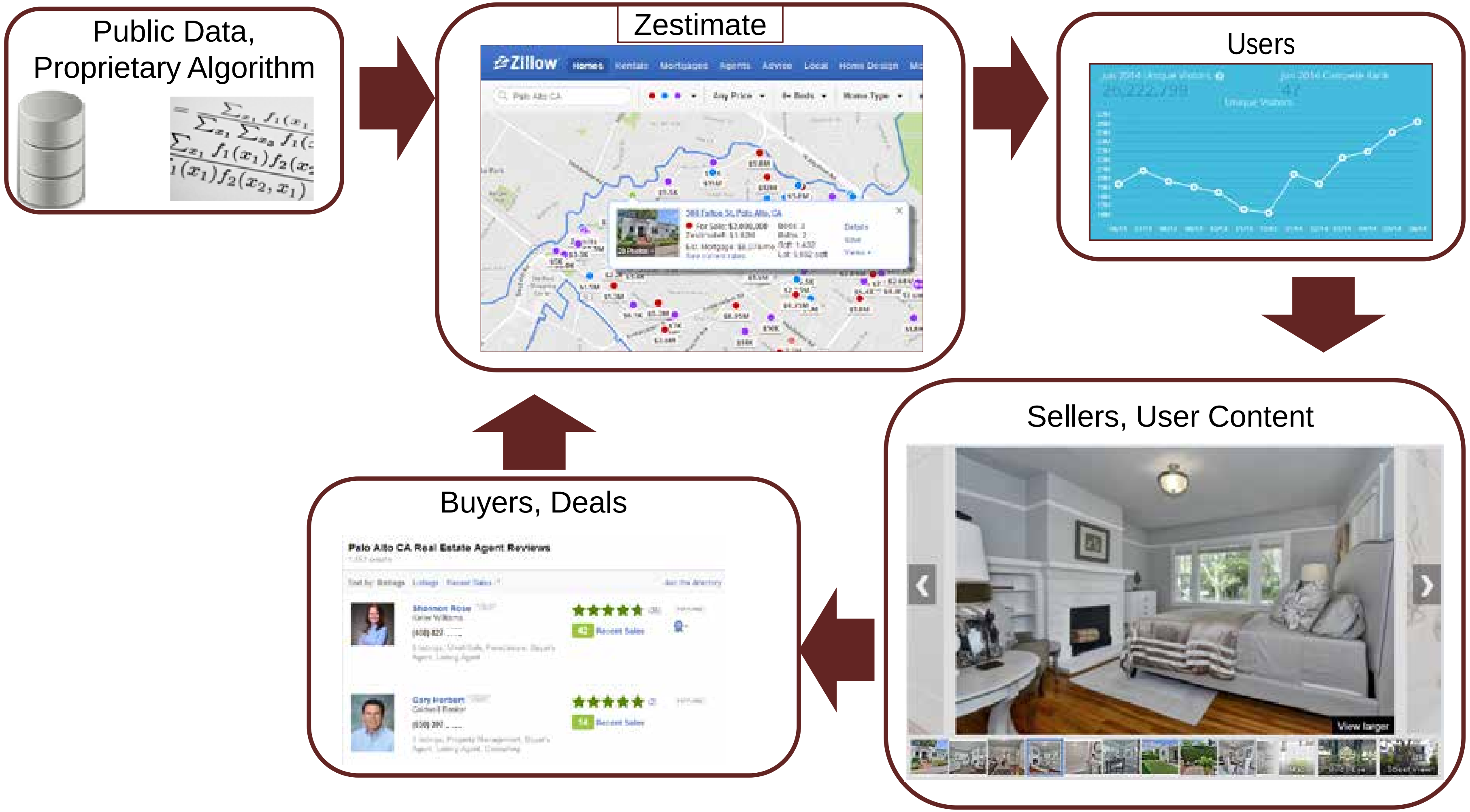
Platforms enjoy new forms of economies of scale

Classic Economies of Scale

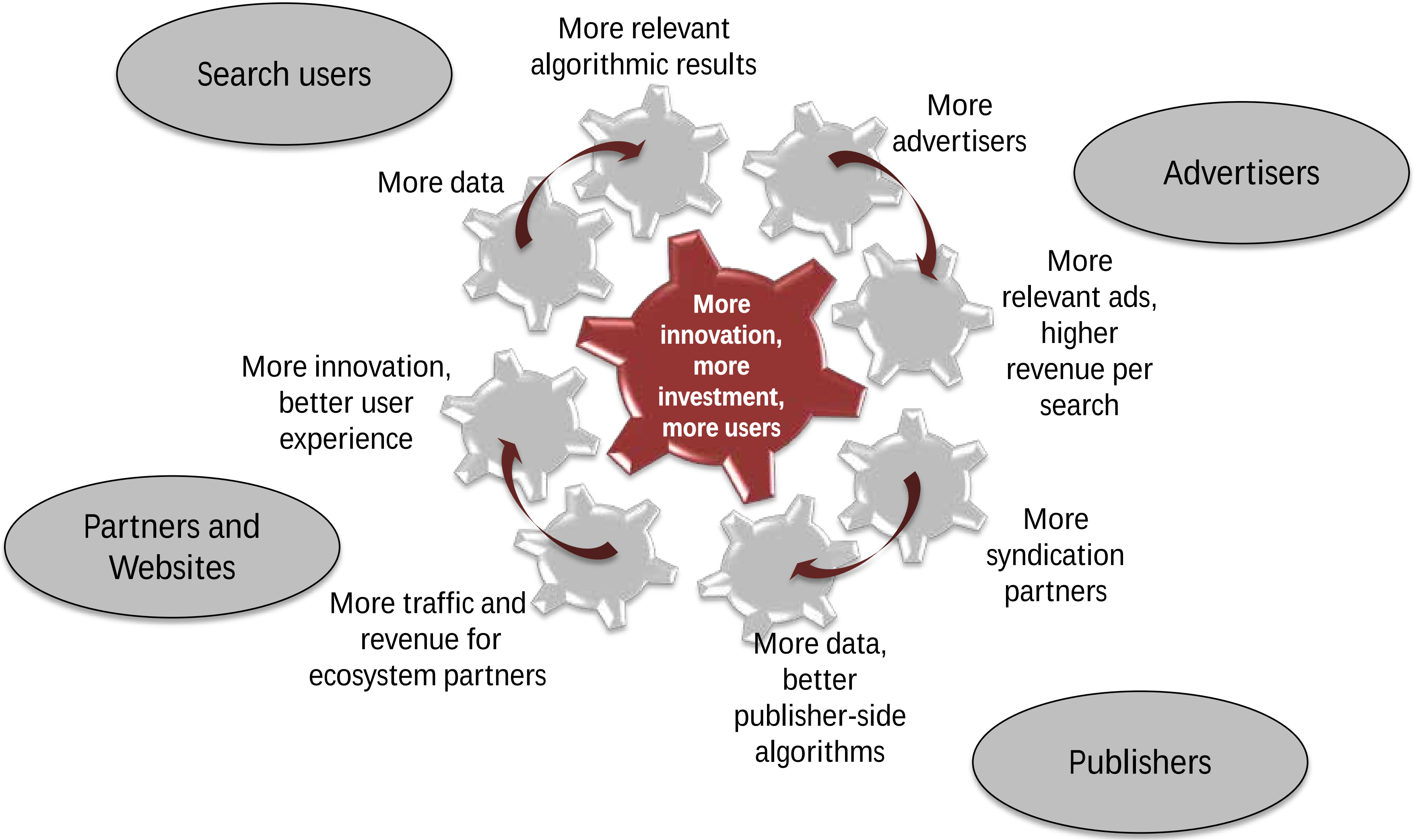
Platform Economies of Scale

- Supply/cost side
  - Fixed costs
    - Setting up & operating the platform
    - Creating content, R&D that attracts users (internet, media, gaming systems)
    - Match-making algorithms (eBay, online advertising, dating)
  - Learning by doing
  - Developing data-driven algorithms
  - “Endogenous sunk costs”
- Demand side
  - Indirect network effects
  - Absolute v. relative scale

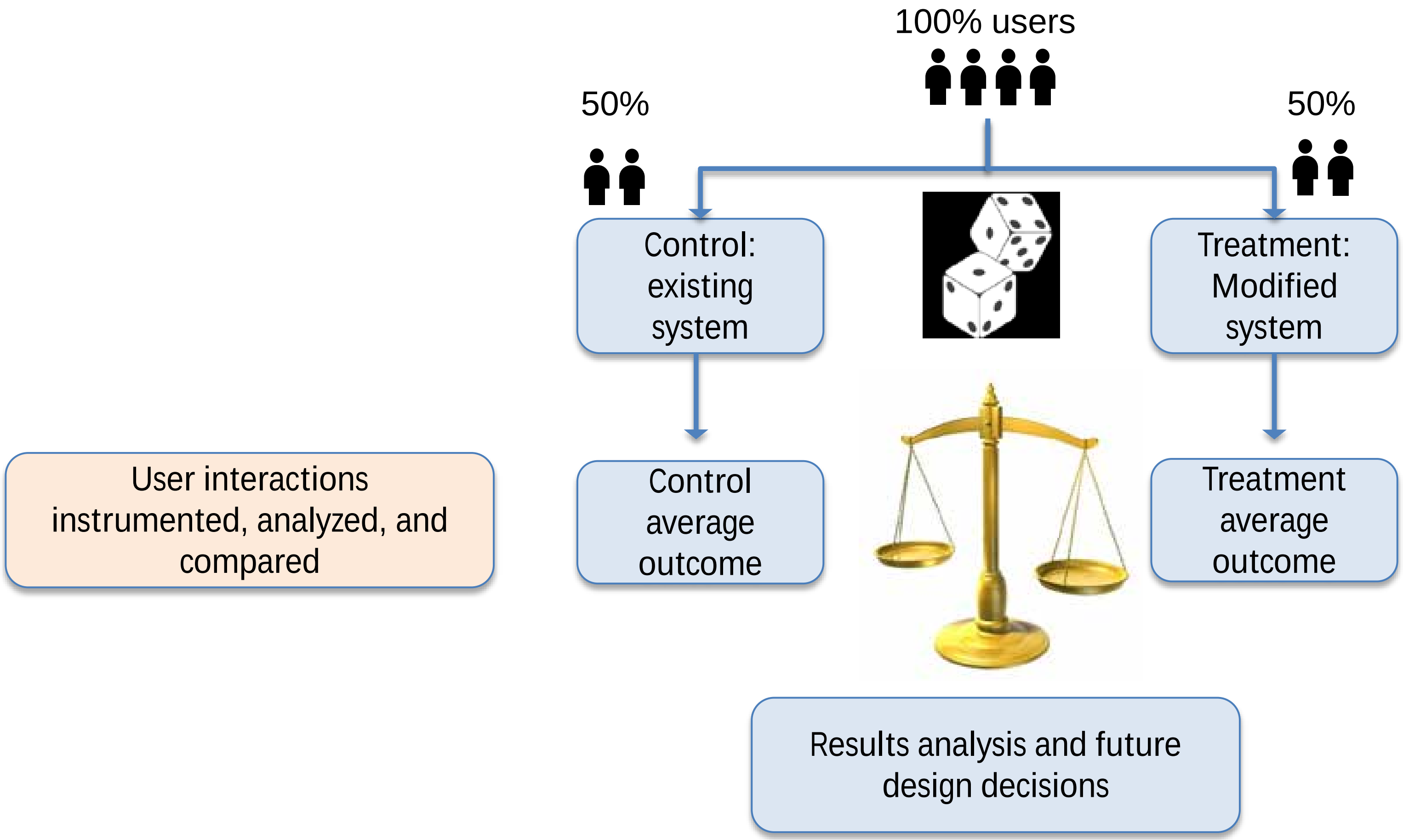
Data as a business model and source of competitive advantage: Zillow



Data as a business model and source of competitive advantage: Search



A/B Testing


100% users
50%
50%
Control: existing system
Treatment: Modified system
Control average outcome
Treatment average outcome
User interactions instrumented, analyzed, and compared
Results analysis and future design decisions

Experimentation increases pace of innovation

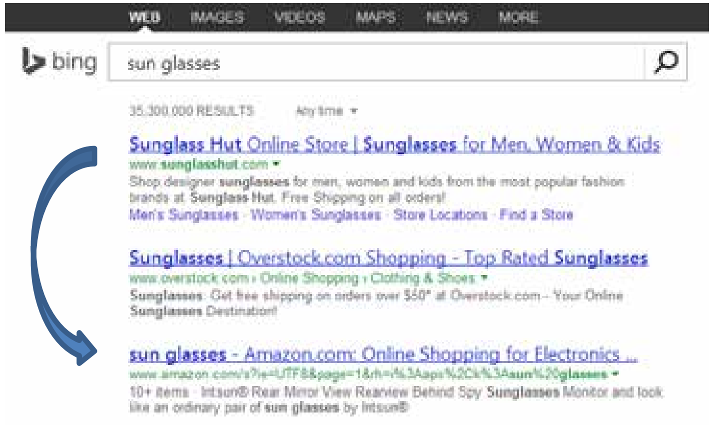
- Prove the value of small changes quickly
- Ship immediately without dedicated meetings and qualitative assessments
- Automate the evaluation process

(Mark Lucovsky, Amazon, 2005): “What is the lag time between the engineer completing the work, and the software reaching its intended customers? A good friend of mine investigated a performance problem one morning, he saw an obvious defect and fixed it. His code was trivial, it was tested during the day, and **rolled out that evening**.”

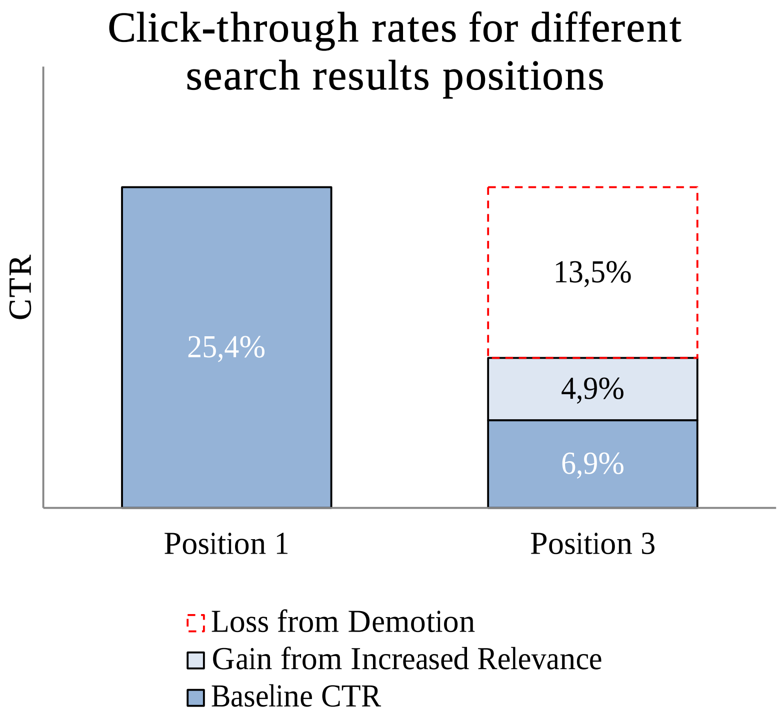
Google Apps’ product manager Rishi Chandra said in an interview (Boulton, 2009): “In terms of the innovation curve that we have, **we release features every two weeks**. That is fundamentally what is going to be Google's differentiation here.

Experimental design

- Take, say, 1% of page views and rerank results or demote vertical links
- Compare metrics for control and treatment, exclude “navigational” queries
- Example metrics: clicks, “good” clicks, click quality



Results



Correlation: Top link gets clicks because search engines put best link in top position. Position is correlated with quality.
Causality: Top link gets clicks because of its position.
Empirical question: How much is position, and how much quality?

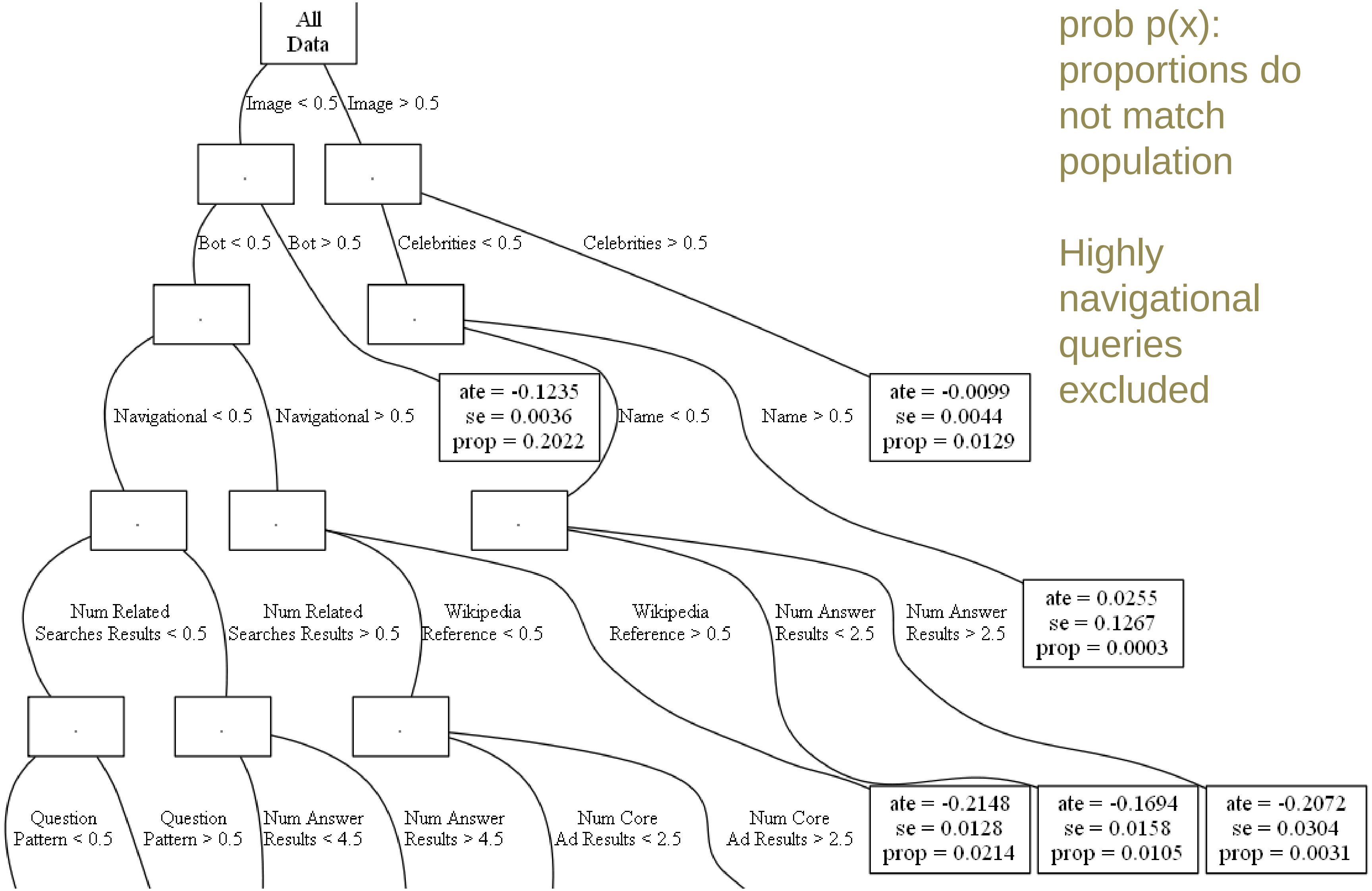
Granular Predictions

Effects of any algorithm are heterogeneous

Customized predictions are more accurate

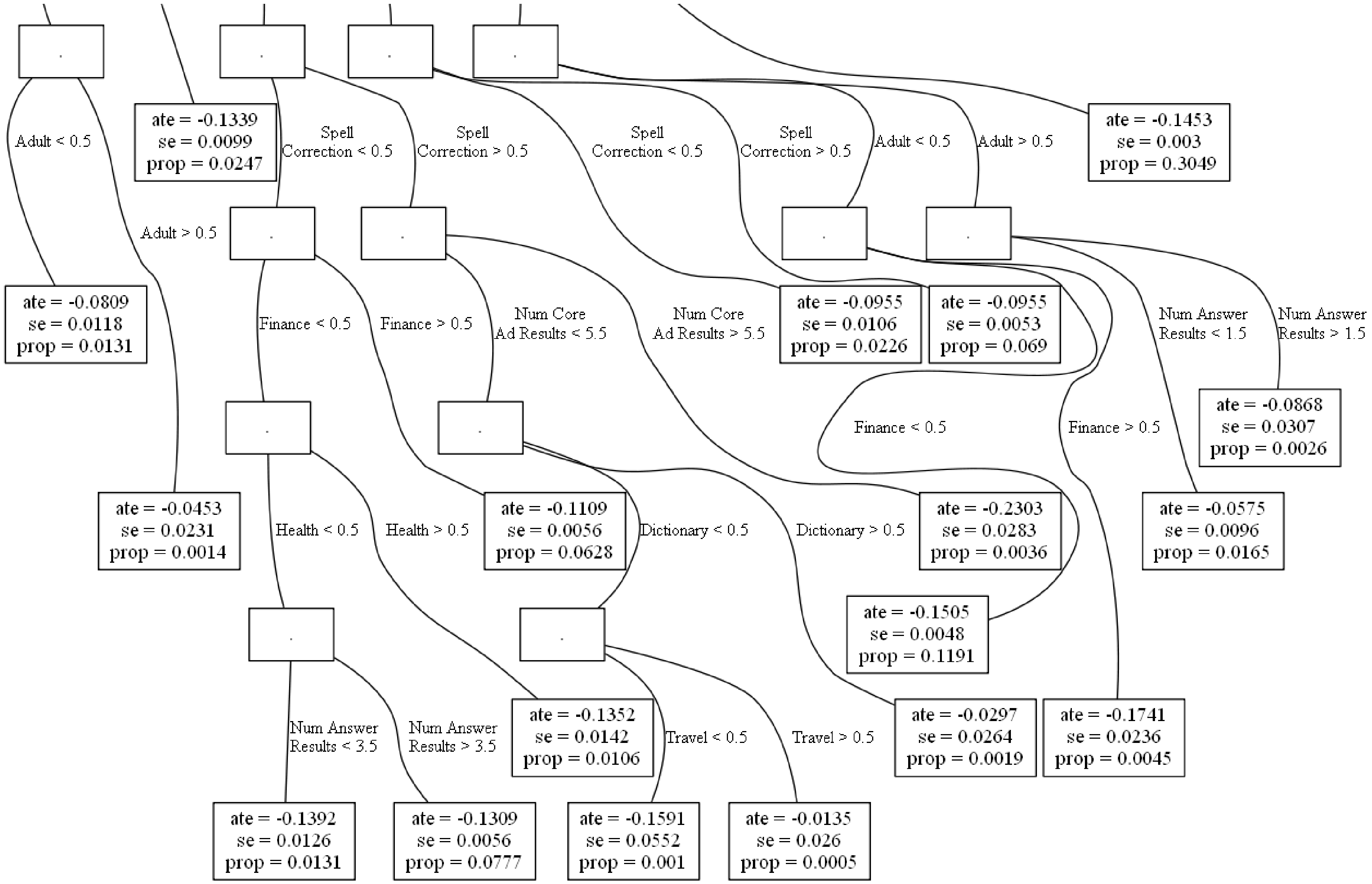
- By user history
- Query characteristics
- Device, OS, browser
- Location
- Etc.

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



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

Highly navigational queries excluded


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

Industrial-Scale Click Prediction Algorithms

Click prediction at Google:

- “A typical industrial model may provide predictions on billions of events per day, using a correspondingly large feature space, and then learn from the resulting mass of data.”
- “It is necessary to make predictions many billions of times per day and to quickly update the model as new clicks and non-clicks are observed. Of course, this data rate means that training data sets are enormous.”
- “The features used in our system are drawn from a variety of sources, including the query, the text of the ad creative, and various ad-related metadata. Data tends to be extremely sparse, with typically only a tiny fraction of non-zero feature values per example.”
- “[We] …handle significantly larger data sets and larger models than have been reported elsewhere to our knowledge, with billions of coefficients.”
- “[We] found that we were unable to project down lower than several billion features without observable loss.”

Source: McMahan et al, 2013, “Ad Click Prediction: A View from the Trenches,” KDD, ACM.

Economics of Data & Scale

- Historical data versus live users for experimentation & optimization
- Importance of learning by doing
- Data that can be bought (generic) versus data in context
- For advertising
  - Advertisers and publishers would like to be able to identify users across sites and when allocating ads to users

- Likely that many data-driven markets will be concentrated
  - How much competition and how effective the competition is depends on fundamentals
  - Where diminishing returns hit depends on fundamentals of markets

STANFORD
GRADUATE SCHOOL OF BUSINESS

Presented by: Susan Athey