facebook

Common Limitations with Big Data in the Field

8/7/2014
Sean J. Taylor
Joint Statistical Meetings

BIG DATA


SAY BIG DATA
ONE MORE TIME
memegenerator.net

Topics

Subscribe

big data problems
Search term

big data solutions
Search term

+Add term

Interest over time

News headlines

Forecast

Source: “big data” from Google

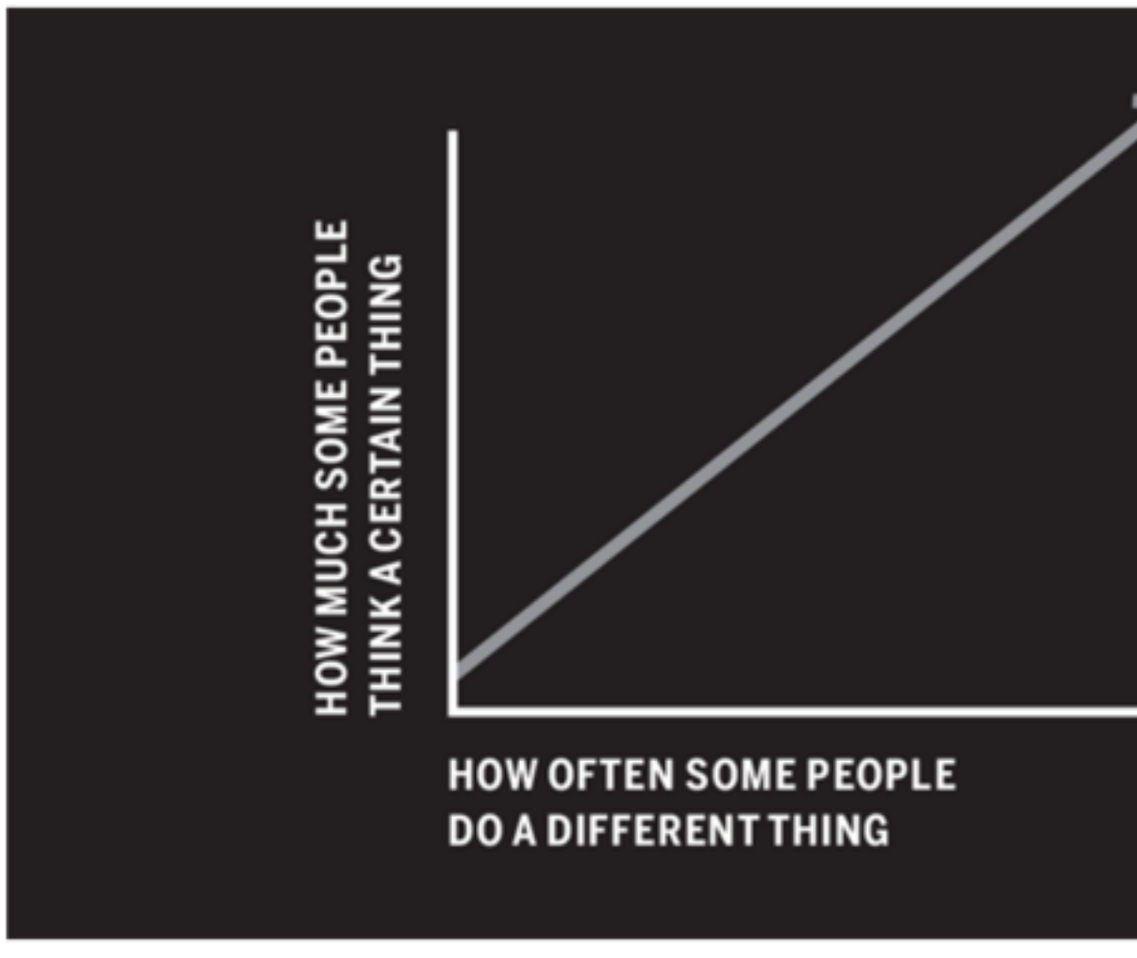
The New York Times

Eight (No, Nine!) Problems With Big Data

By GARY MARCUS and ERNEST DAVIS APRIL 6, 2014

EMAIL
FACEBOOK
TWITTER
SAVE
MORE

BIG data is suddenly everywhere. Everyone seems to be collecting it, analyzing it, making money from it and celebrating (or fearing) its powers. Whether we're talking about analyzing zillions of Google search queries to predict flu outbreaks, or zillions of phone records to detect signs of terrorist activity, or zillions of airline stats to find the best time to buy plane tickets, big data is on the case. By combining the power of modern computing with the plentiful data of the digital era, it promises to solve virtually any problem — crime, public health, the evolution of grammar, the perils of dating — just by crunching the numbers.

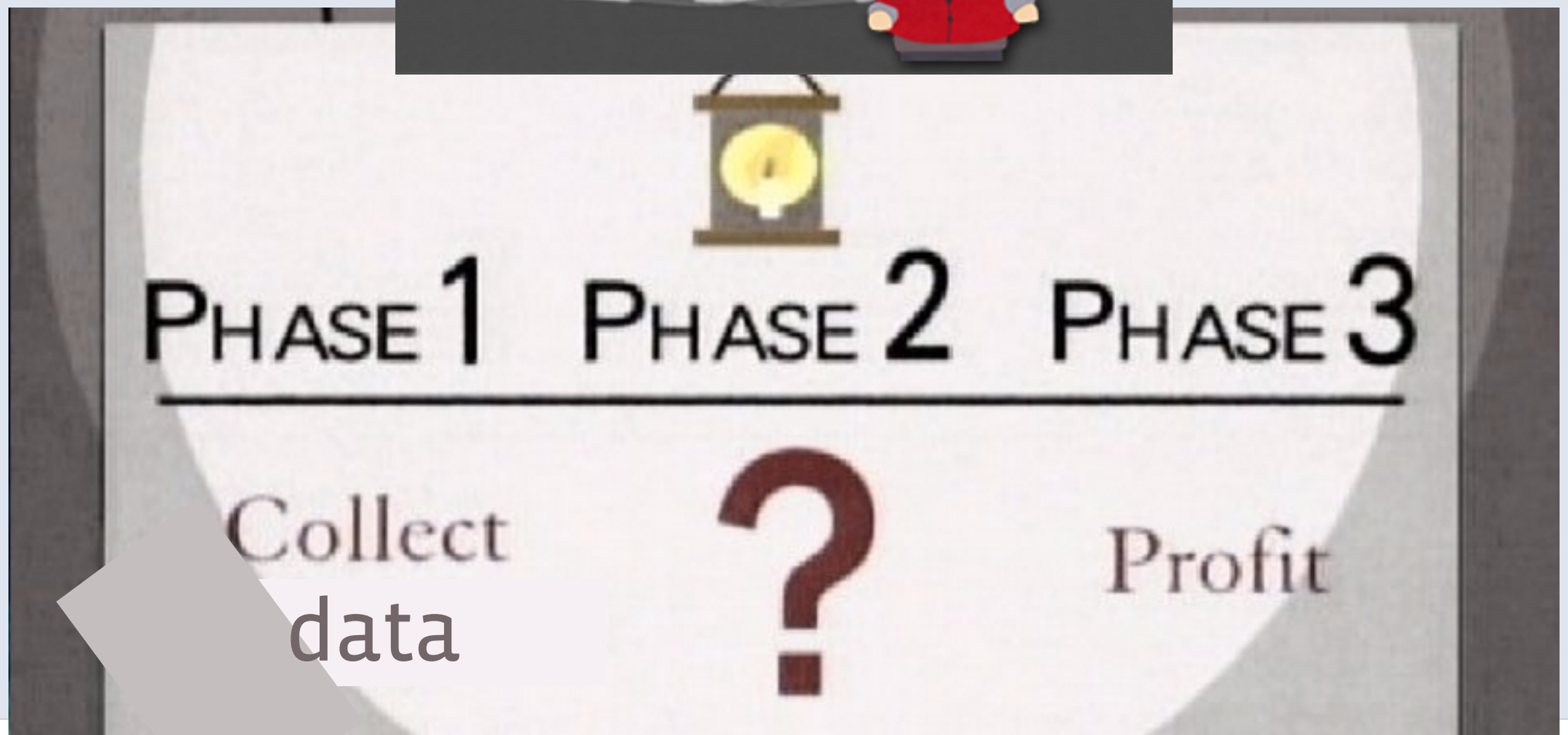


Open, N.Y.

Or so its champions allege. "In the next two

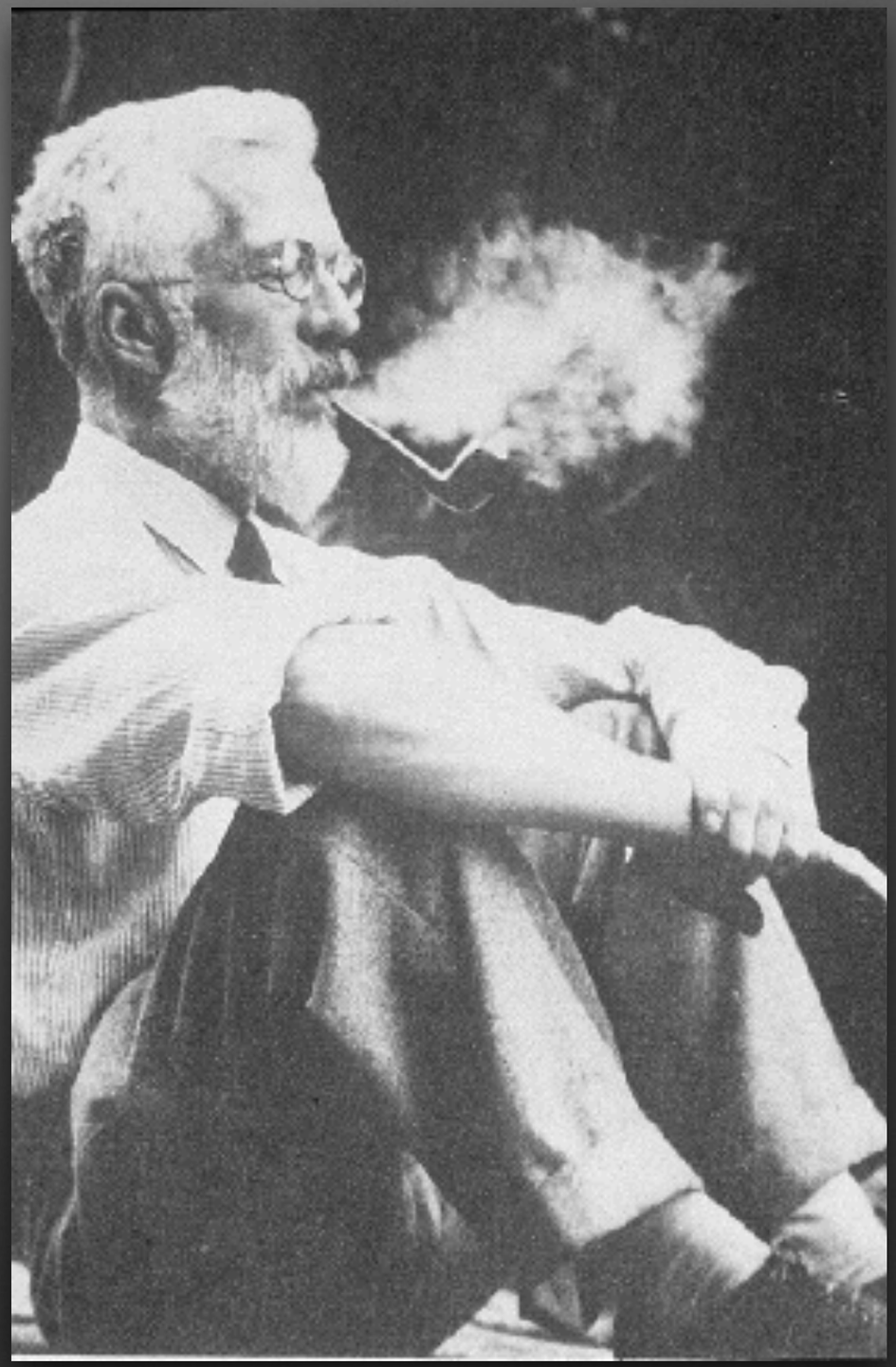

Mo' Data,
Mo' Problems

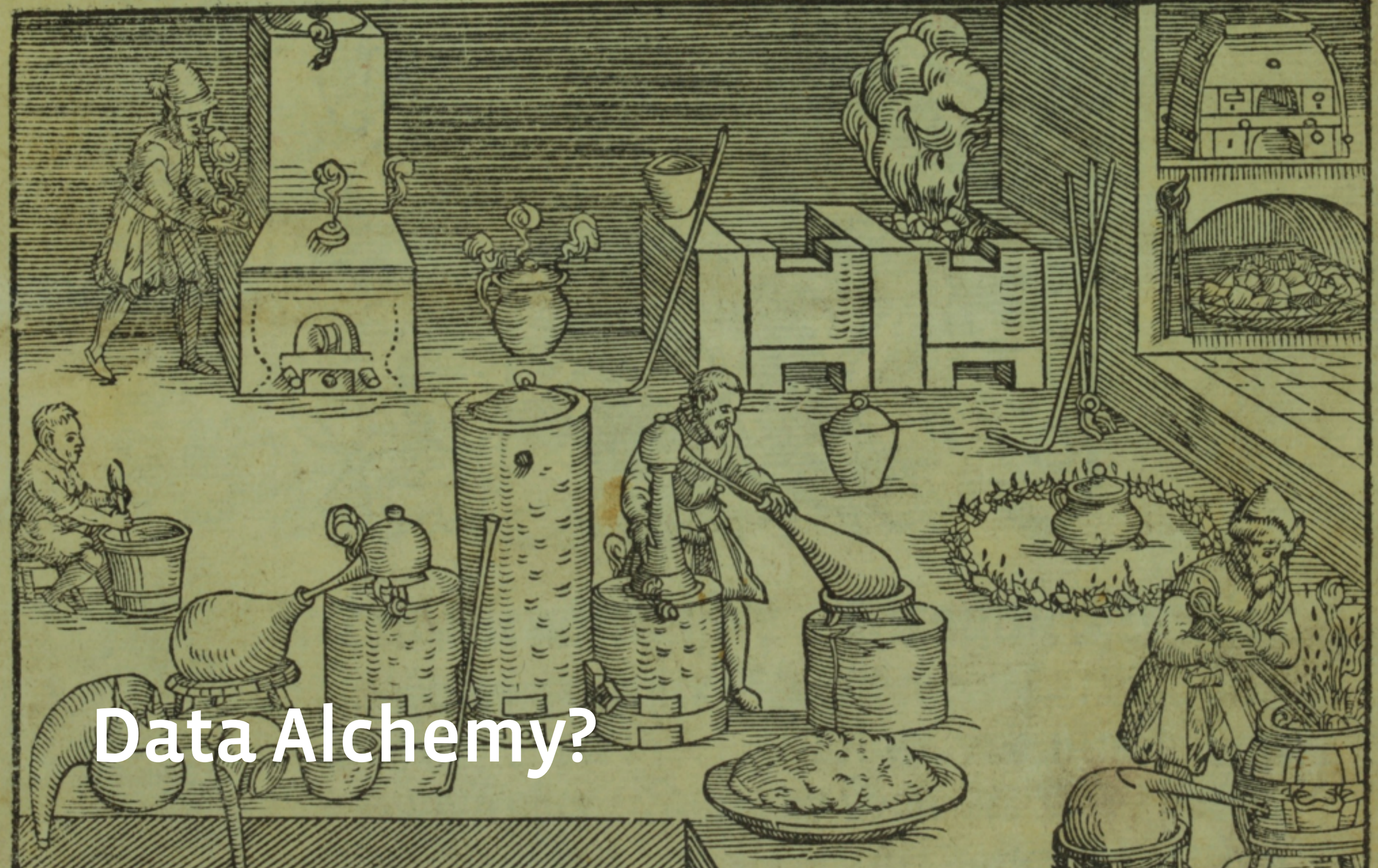
HOARDERS


PHASE 1
PHASE 2
PHASE 3
Collect data
?
Profit

“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.”

— Ronald Fisher





Data Alchemy?

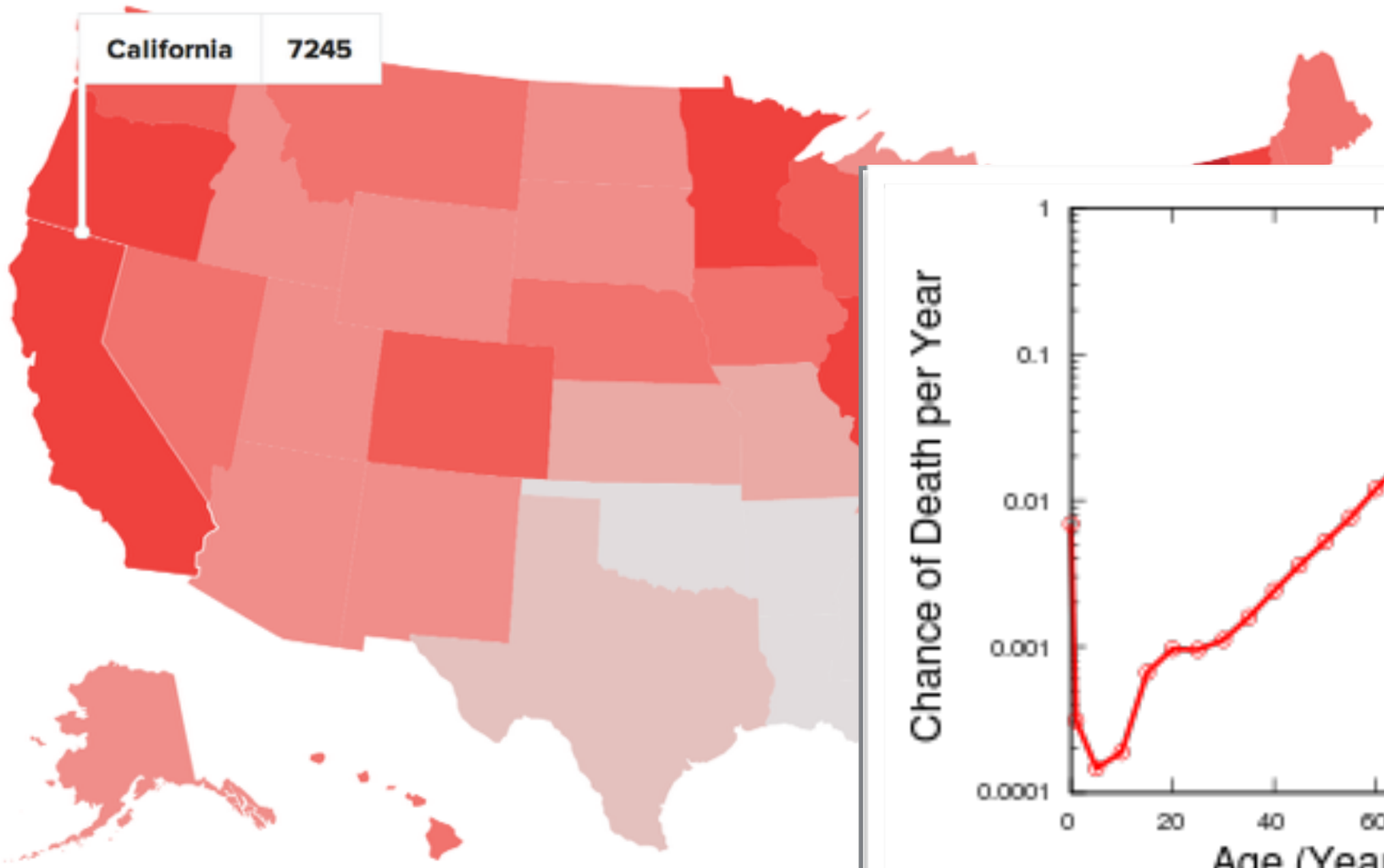
Common Big Data Limitations:

1. Measuring the wrong thing
2. Biased samples
3. Dependence
4. Uncommon support

1. Measuring the wrong thing

The data you have a lot of doesn't measure what you want.

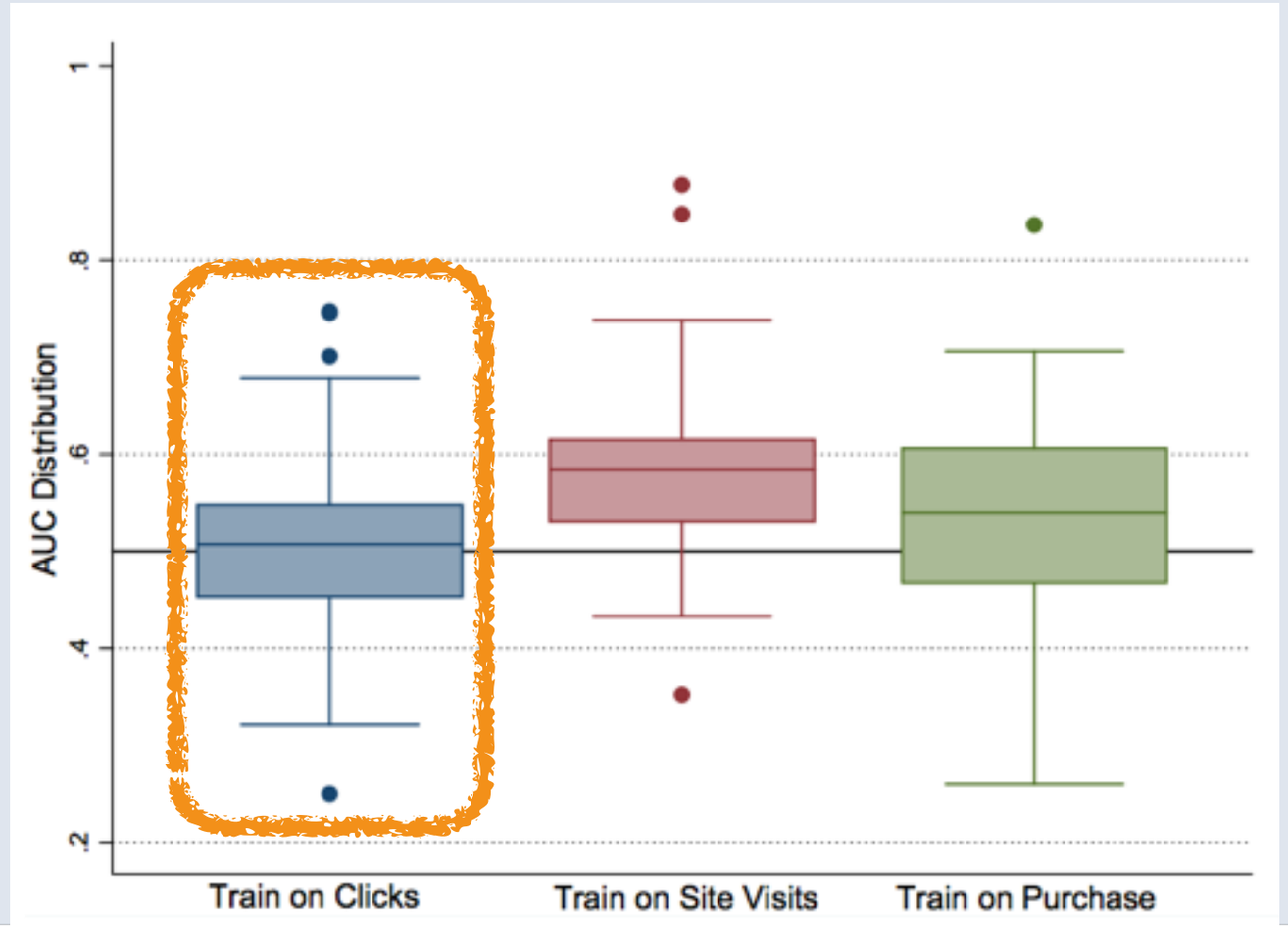
AVERAGE STEPS BY STATE (INTERACTIVE)



Evaluating and Optimizing Online Advertising: Forget the click, but there are good proxies

Brian Dalessandro, Rod Hook, Claudia Perlich
m6d research

Foster Provost
NYU/Stern School of Business and m6d research



Common Pattern

- High volume of of cheap, easy to measure "surrogate" (e.g. steps, clicks)

- Surrogate is correlated with true measurement of interest (e.g. overall health, purchase intention)

- key question: sign and magnitude of "interpretation bias"

- obvious strategy: design better measurement

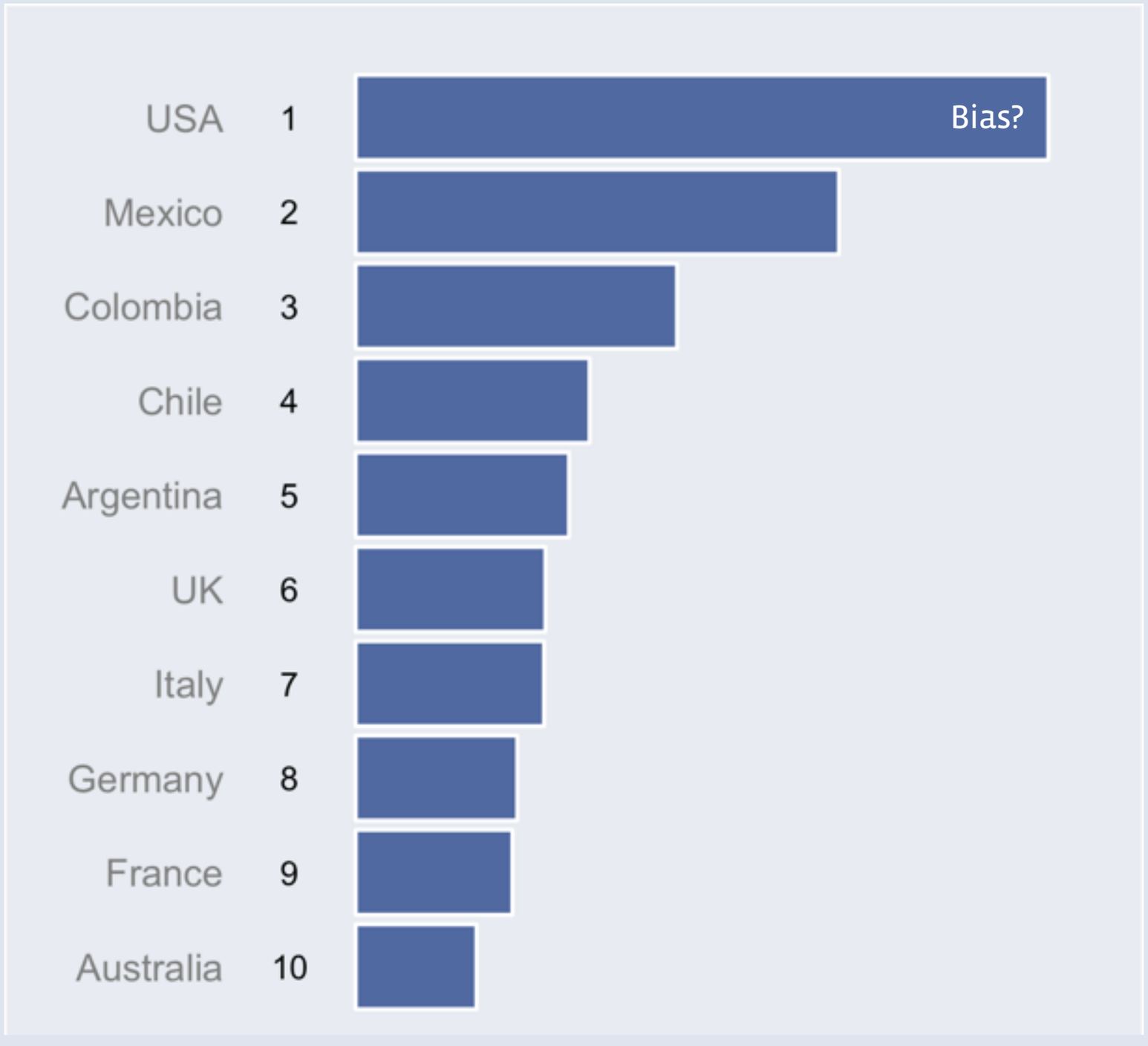
- opportunity: joint, causal modeling of high-resolution surrogate measurement with low-resolution "true" measurement.

2. Biased samples

Size doesn't guarantee your data is representative of the population.

June 5

7:00 UTC



World Cup Arrivals by Country

Having daughters makes you more liberal. No, it makes you more conservative. No, it . . . ??

By Andrew Gelman December 18, 2013

A couple of people pointed me to a recent study by sociologists Dalton Conley and Emily Rauscher that reported that respondents to the 1994 General Social Survey who had daughters were more likely to identify with

Most Read Politics

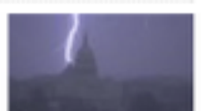
1 Obama signs bill giving Israel $225 million for missile defense system

2 With judge's ruling, shutdown costs could grow

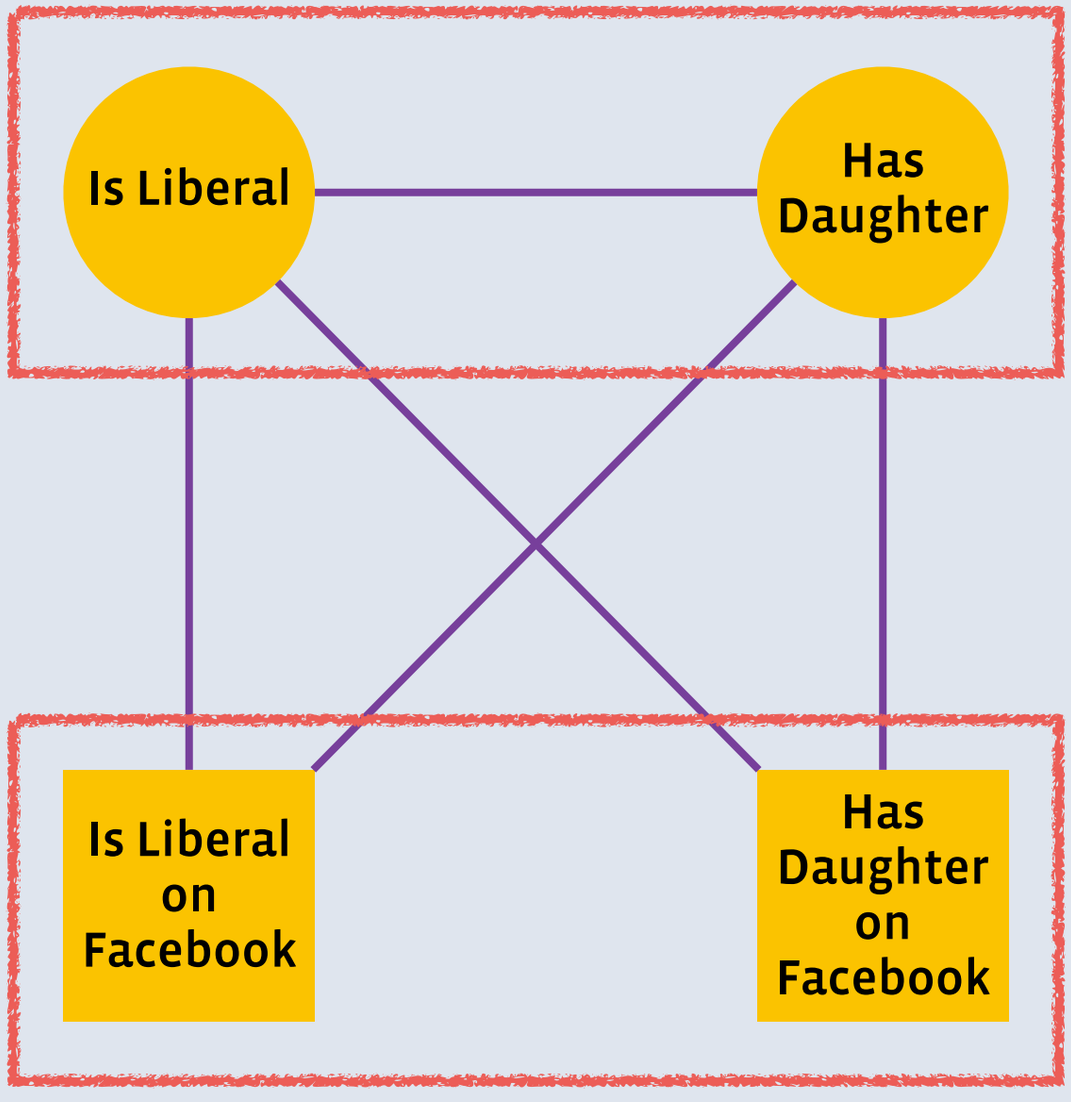
3 The real reason for America's polarization? Look next door.

4 All 274 gifts given to Barack Obama between 2009 and 20...

5 A majority of people don't like their own



What we want



What we get

Common Pattern

- Your sample conditions on your own convenience, user activity level, and/or technological sophistication of participants.

- You'd like to extrapolate your statistic to a more general population of interest.

- key question: (again) sign and magnitude of bias
- obvious strategy: weighting or regression
- opportunity: understand effect of selection on latent variables on bias

3. Dependence

Effective size of data is small due to repeated measurements of units.

(a) (b)

Bakshy et al. (2012)
"Social Influence in Social Advertising"

Person	D	Y
Evan	1	1
Evan	0	0
Ashley	0	1
Ashley	1	1
Ashley	0	1
Greg	1	0
Leena	1	0
Leena	1	1
Ema	0	0
Seamus	1	1

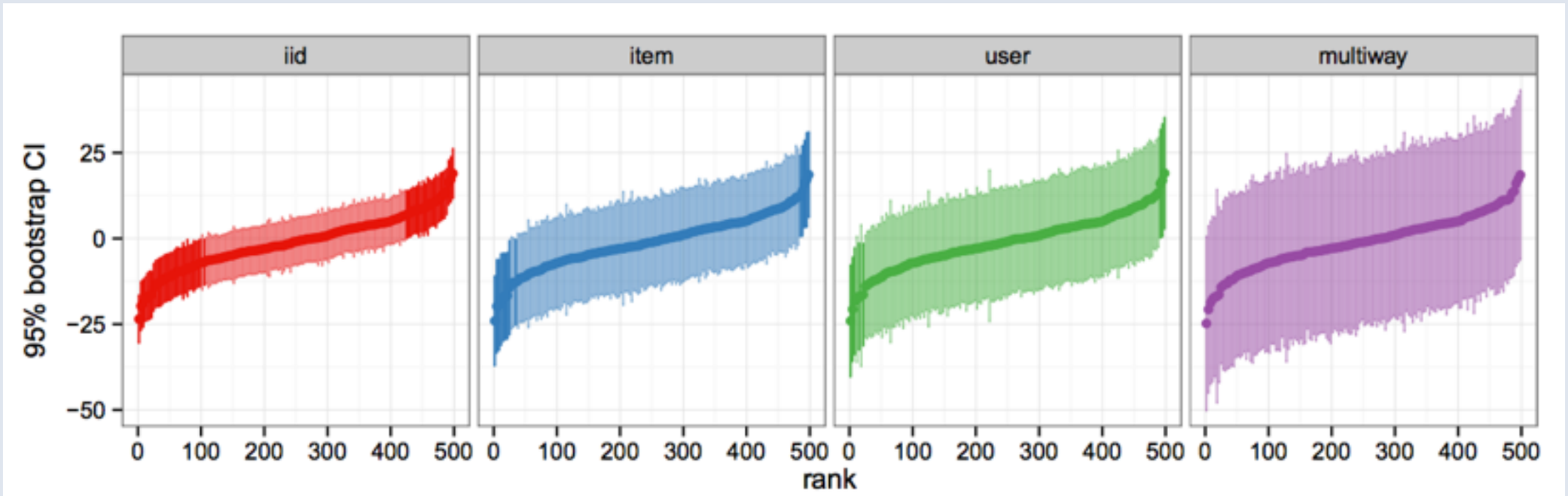
One-way Dependency

$$Y_{ij} = \delta D_{ij} + \alpha_i + \epsilon_{ij}$$

Person	Ad	D	Y
Evan	Sprite	1	1
Evan	Coke	0	0
Ashley	Pepsi	0	1
Ashley	Pepsi	1	1
Ashley	Coke	0	1
Greg	Coke	1	0
Leena	Pepsi	1	0
Leena	Coke	1	1
Ema	Sprite	0	0
Seamus	Sprite	1	1

Two-way Dependency

$$Y_{ij} = \delta D_{ij} + \alpha_i + \beta_j + \epsilon_{ij}$$



Bakshy and Eckles (2013)
“Uncertainty in Online Experiments with Dependent Data”

Common Pattern

- Your sample is large because it includes many repeated measurements of the most active units.

- An inferential procedure that assumes independence will be anti-conservative.

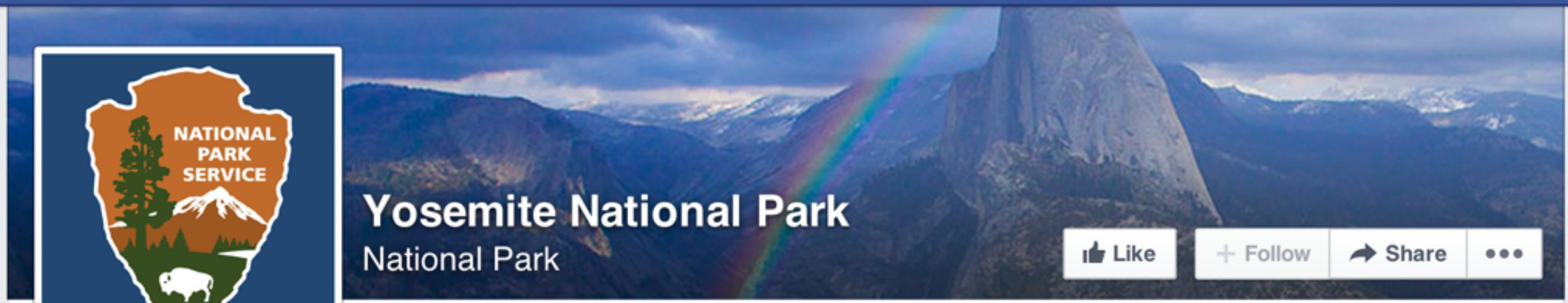
- key question: how to efficiently produce confidence intervals with the right coverage.

- obvious strategy: multi-way bootstrap (Owen and Eckles)

- opportunity: scalable inference for crossed random effects models

4. Uncommon support

Your data are sparse in part of the distribution you care about.



Yosemite National Park

National Park

Like | + Follow | Share | •••

PEOPLE

★★★★★

347,565 likes

1,281,960 visits

Johan Ugander, **Brian Karrer** and **18 other friends** like this or have been here.





+15

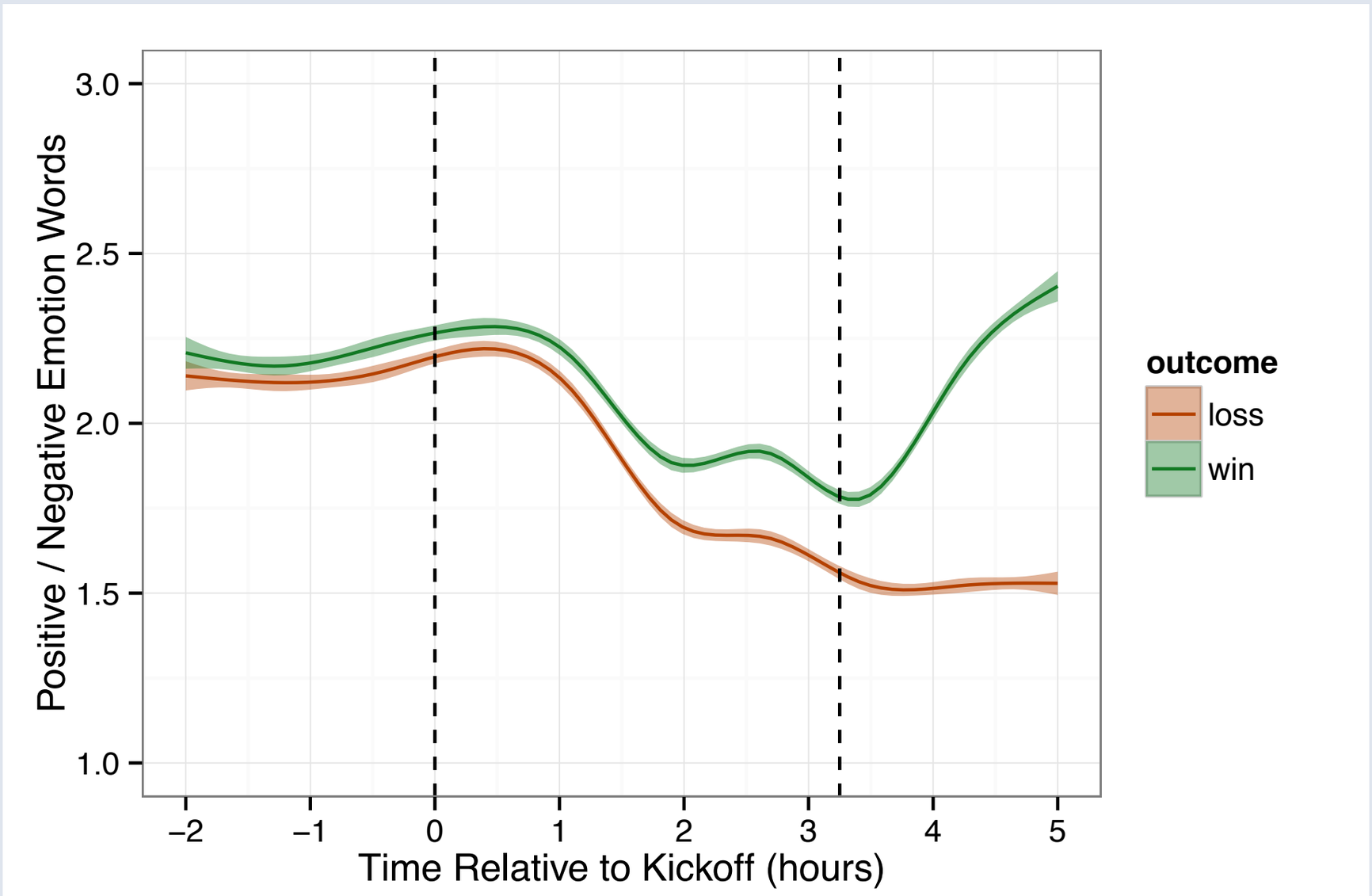
Invite your friends to like this Page

Post | Photo / Video | Review

Write something on this Page...

Yosemite National Park
19 hours ago

This afternoon in Yosemite Valley has been relatively cool and cloudy (thanks to a low pressure system) and smoke-free (thanks to the firefighters, who now have the El Portal Fire 100% contained!)

“The Emotional Highs and Lows of the NFL Season”
Facebook Data Science Blog

Which check-ins are useful?

- We can only study people who post enough text before and after their check-ins to yield reliable measurement.

What is the control check-in?

- Ideally we'd do a within-subjects design. If so we're stuck with only other check-ins (with enough status updates) from the same people.

- National Park visit is likely a vacation, so we need to find other vacation check-ins. Similar day, time, and distance.

  Only a small fraction of the "treatment" cases yield reliable measurement plus control cases.

Common Pattern

- Your sample is large but observations satisfying some set of qualifying criteria (usually matching) are rare.

- Even with strong assumptions, hard to measure precise **differences** due to lack of similar cases.

- key question: how to use as many of the observations as possible without introducing bias.

- obvious strategy: randomized experiments

- opportunity: more efficient causal inference techniques that work at scale (high dimensional regression/matching)

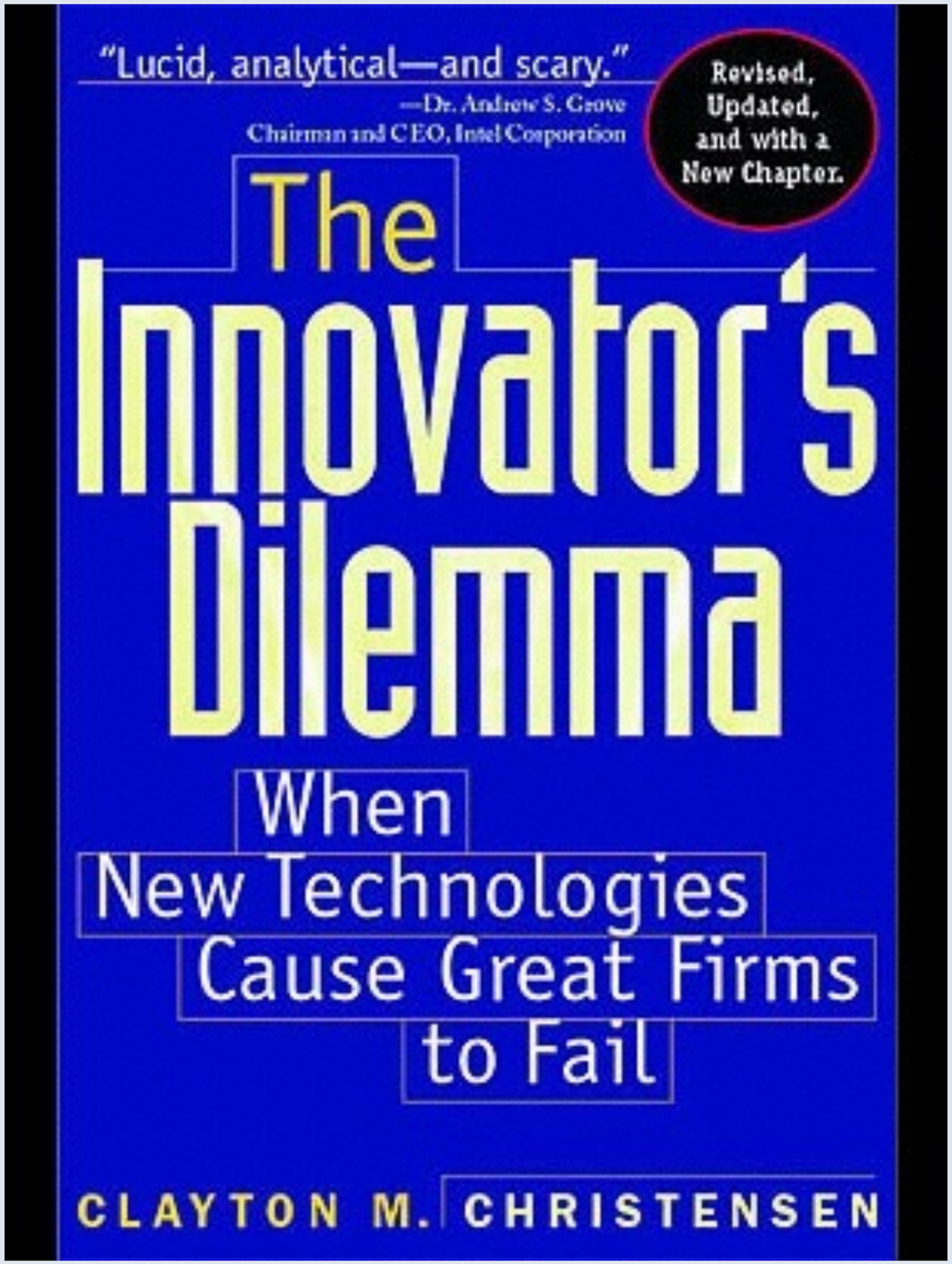
Conclusion 1:
Making your own quality data is better than being a data alchemist.

>

[PDF] **Twitter** mood **predicts the stock market**. - arXiv
arxiv.org/pdf/1010.3003& ▾ arXiv ▾
by J Bollen - 2010 - Cited by 605 - Related articles
Oct 14, 2010 - Index Terms—**stock market** prediction — **twitter** — social media (blogs, **Twitter** feeds, etc) to **predict** changes in vario

Conclusion 2:
Great research opportunities in addressing limitations of "convenience" big data.

(or how I stopped worrying and learned to be a data alchemist)


"Lucid, analytical—and scary."
—Dr. Andrew S. Grove
Chairman and CEO, Intel Corporation
Revised, Updated, and with a New Chapter.
The Innovator's Dilemma
When New Technologies Cause Great Firms to Fail
CLAYTON M. CHRISTENSEN

Sustaining versus Disruptive Innovations

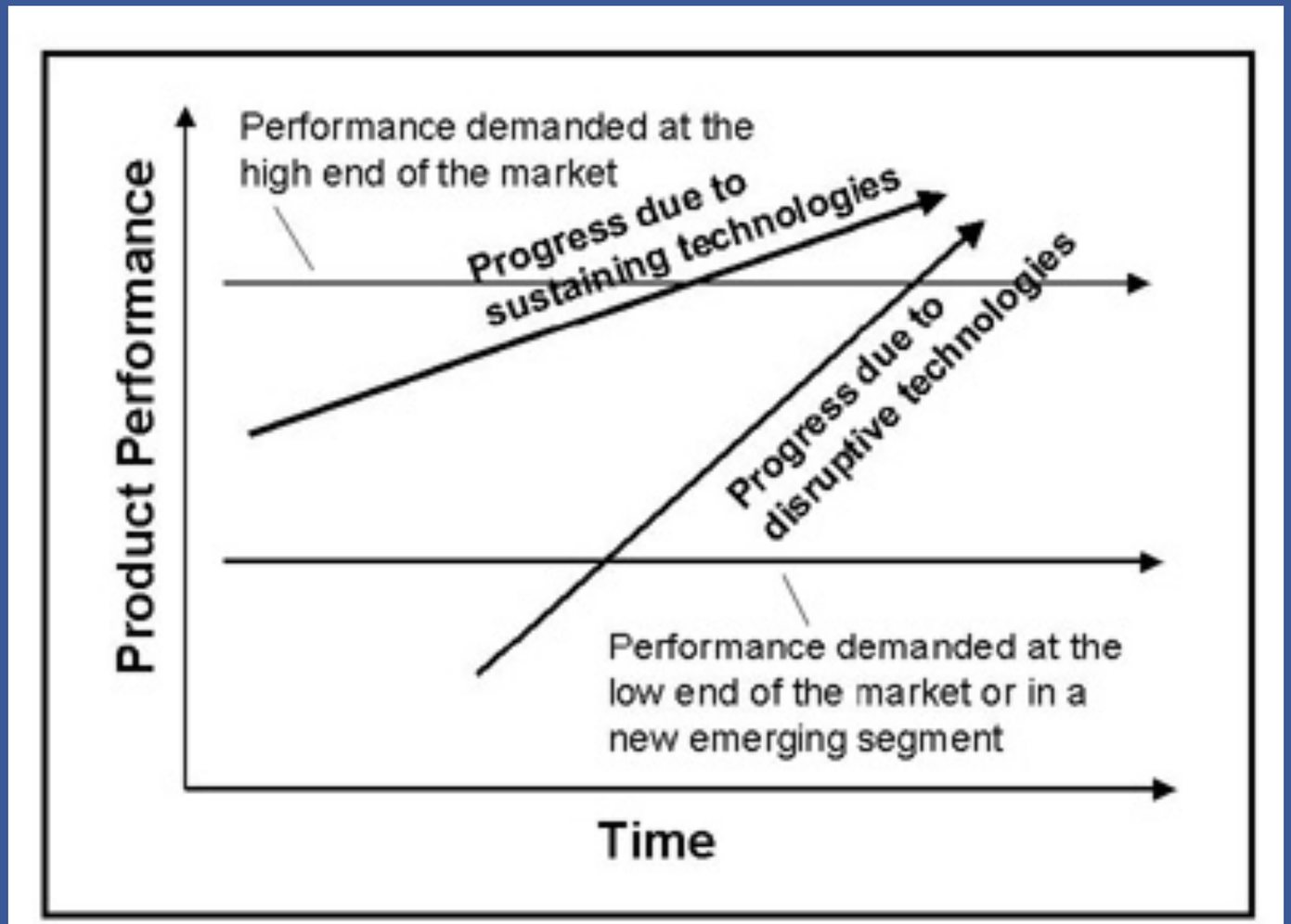
Sustaining innovations

- improve product performance

Disruptive innovations

- Result in worse performance (short-term)
- Eventually surpass sustaining technologies in satisfying market demand with lower costs
- cheaper, simpler, smaller, and frequently more convenient to use

Innovator's Dilemma



sjt@fb.com
http://seanjtaylor.com

facebook