CSC2552

Topics in Computational Social Science: AI, Data, and Society

Spring 2021

Lecture 2: Introduction to Computational Social Science cont'd

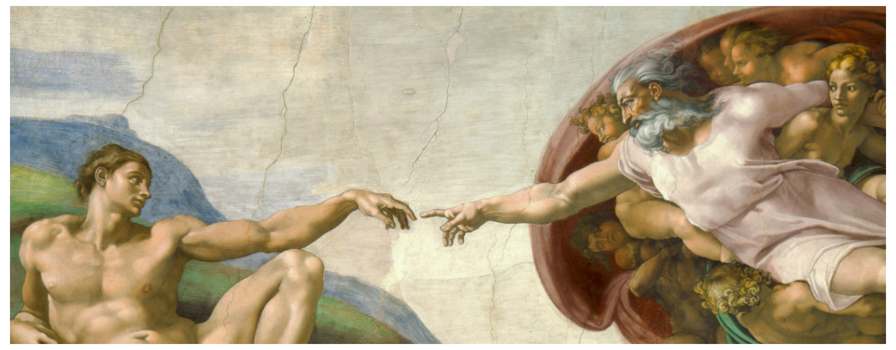
Ashton Anderson
University of Toronto

Computational social science in 7 easy pieces

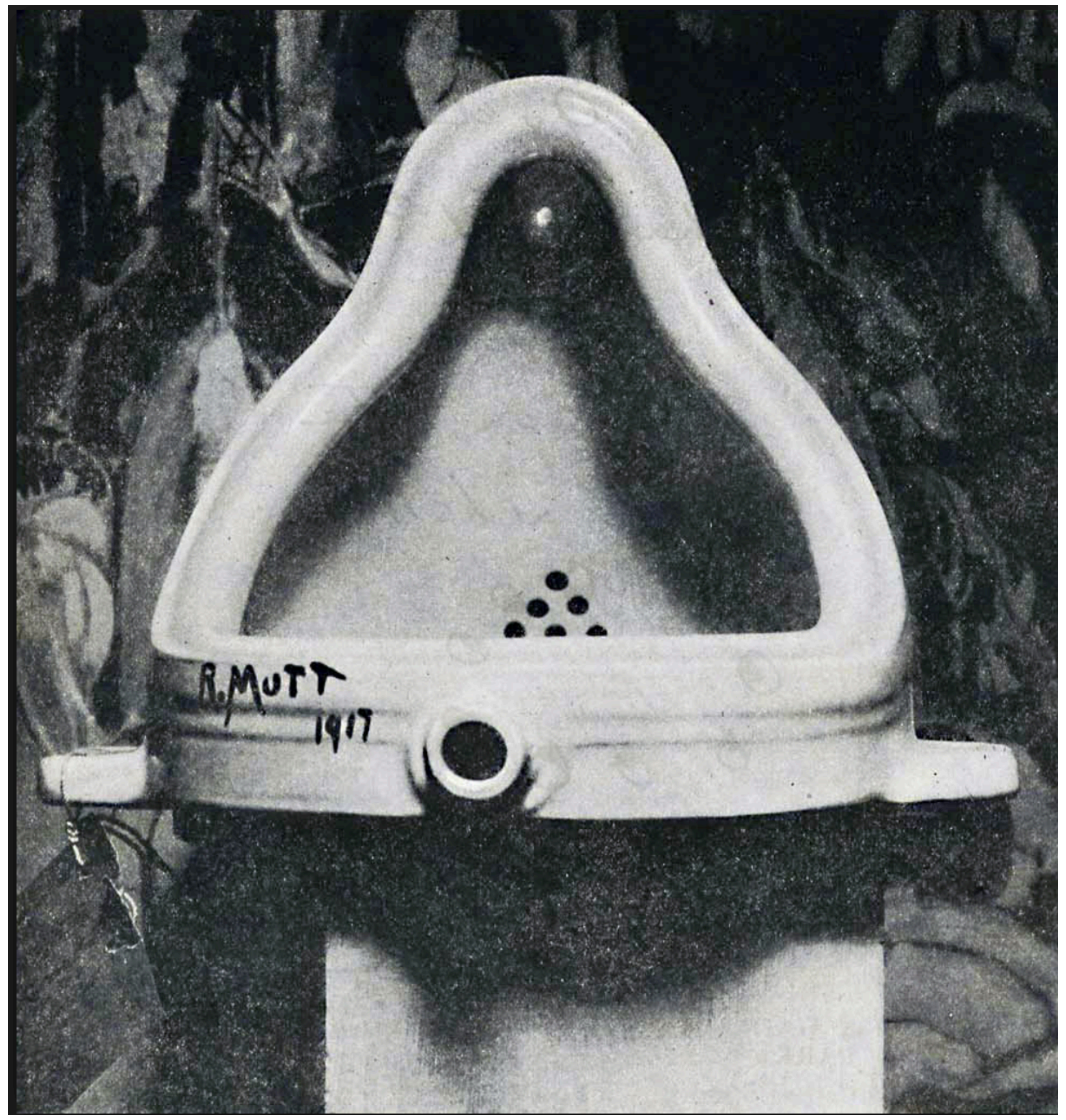
	Week	Date	Topic	Reviews Due	Textbook Readings
	1	1/14	Introduction to computational social science		Ch. 1
	2	1/21	Introduction to computational social science cont'd		Ch. 1
★	3	1/28	Observational studies 1	1/27 9:00pm	Ch. 2
★	4	2/4	Observational studies 2	2/3 9:00pm	Ch. 2
★	5	2/11	Experiments 1	2/10 9:00pm	Ch. 4
★	6	2/25	Experiments 2	2/24 9:00pm	Ch. 4
	7	3/4	Project proposals		
★	8	3/11	Asking questions	3/10 9:00pm	Ch. 3
★	9	3/18	Mass collaboration	3/17 9:00pm	Ch. 5
★	10	3/25	Ethics in computational social science	3/24 9:00pm	Ch. 6
	11	4/1	Project presentations (Part 1)		
	12	4/8	Project presentations (Part 2)		

Readymades

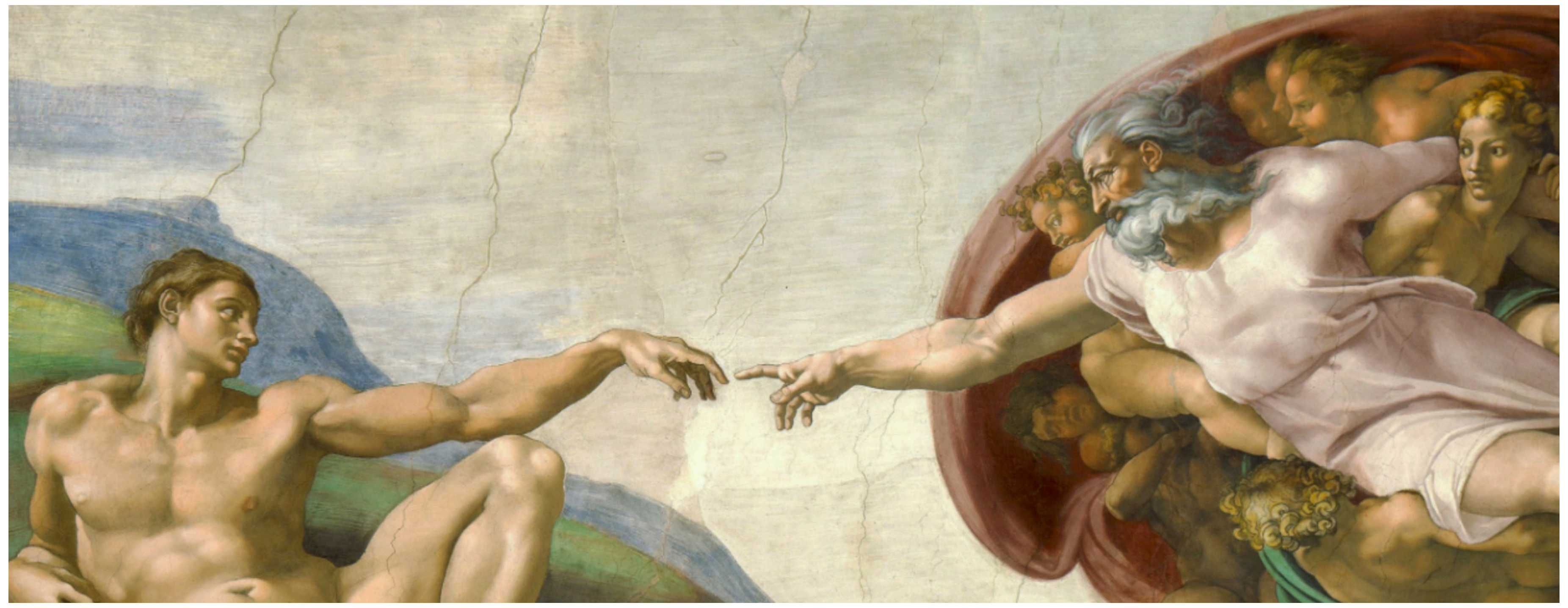
Custommades



Ways of doing computational social science

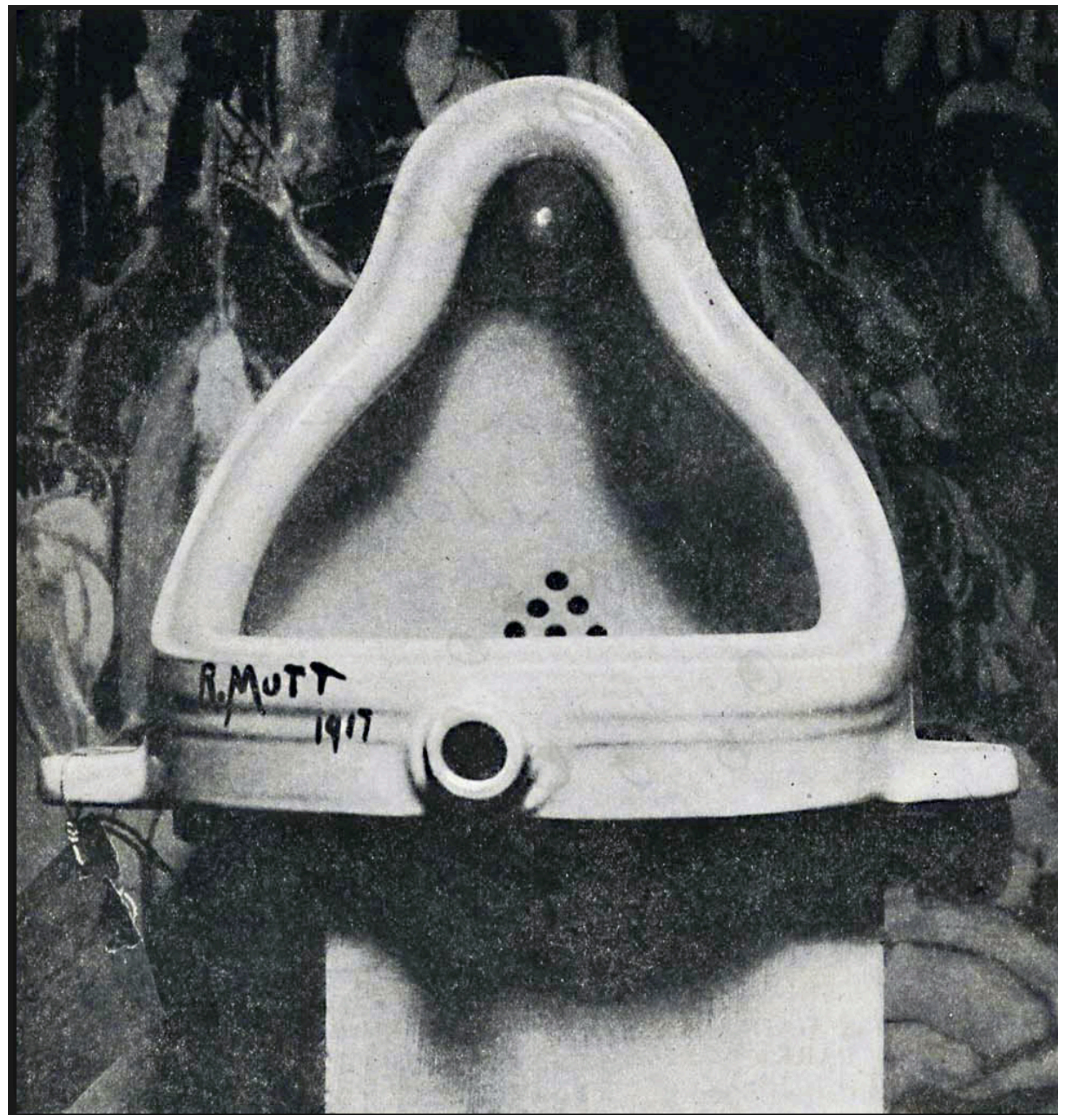


Readymades

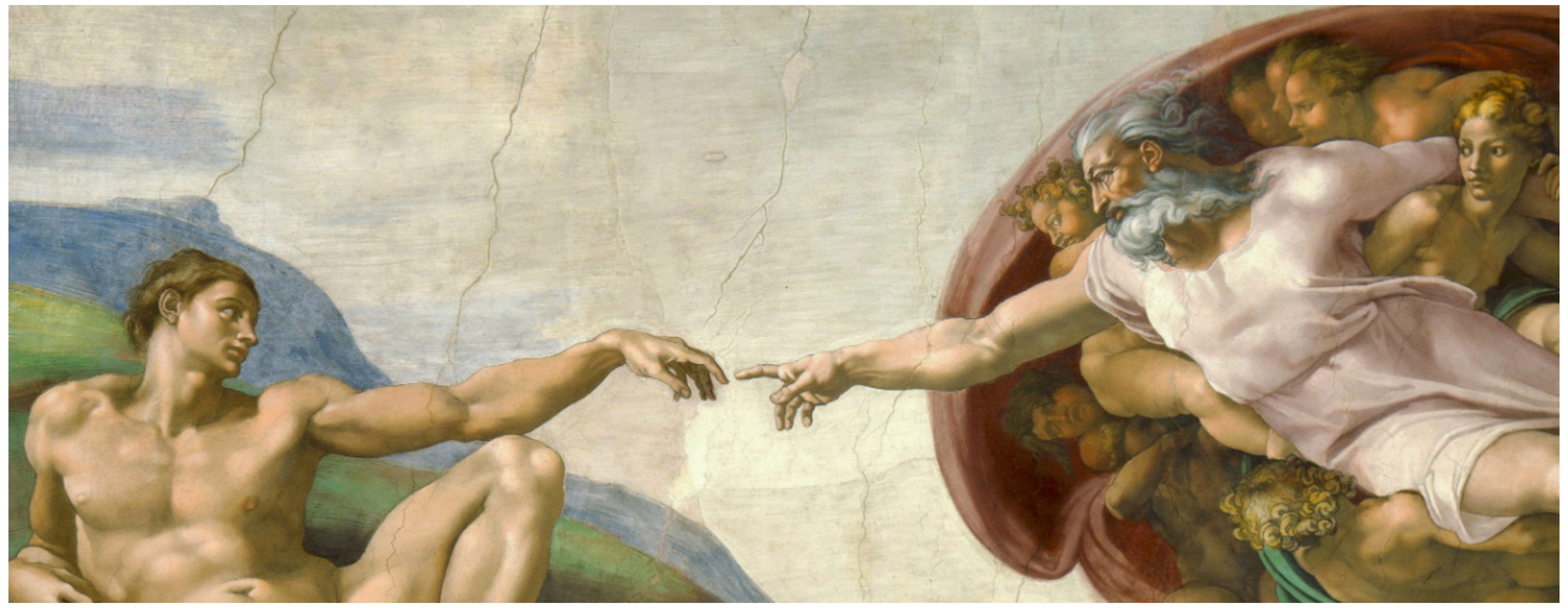


Custommades

Ways of doing computational social science



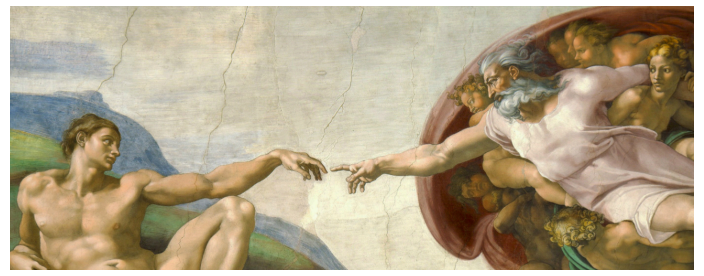
"Found" data



Experiments

A spectrum between the two

Ways of doing computational social science



Observational analyses

Human computation

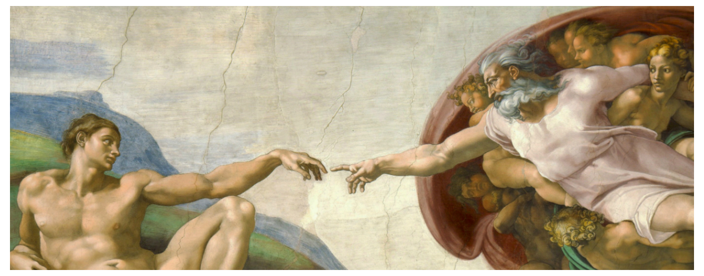
Natural experiments

Surveys

Field experiments

Lab studies

Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

Surveys

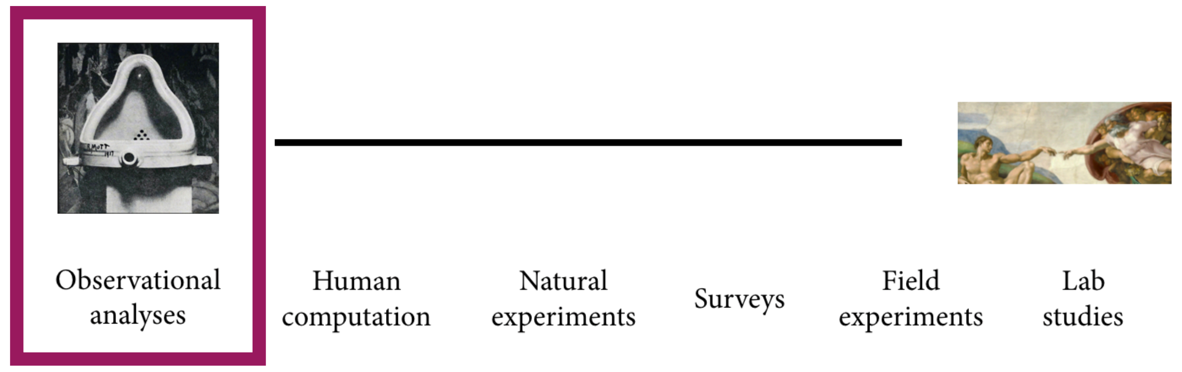
Field experiments

Lab studies

Observational analyses of existing data

- Massive datasets of all kinds of human behaviour are now available for study
  - Wikipedia, GPS traces, health databases, Facebook, Twitter, Reddit, reviews, purchases, dating, invitations, exercise apps, etc., etc…
- Key part of the "socioscope": huge traces of things that we couldn't see before
- Lack of detail/fidelity in individual records is hopefully made up for by large numbers of records (small noisy errors cancel out, big patterns are signal)

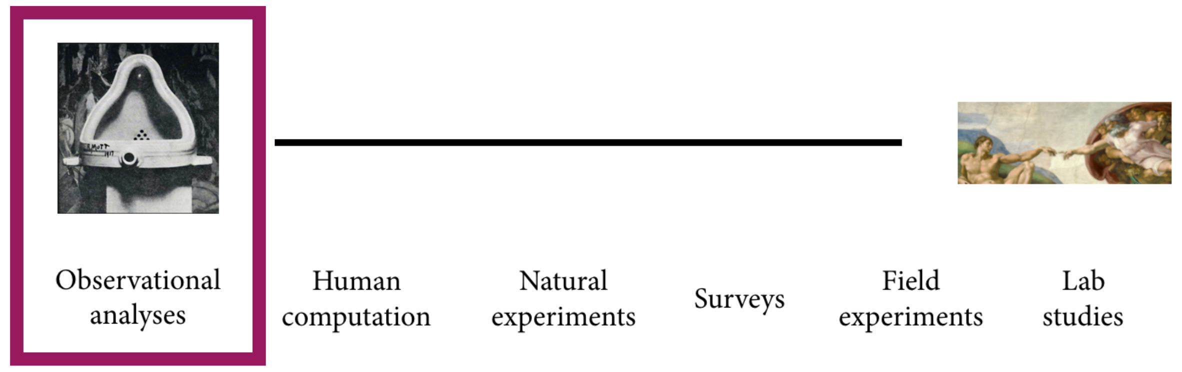
"Big data" / "Found data"



Ten common characteristics of big data

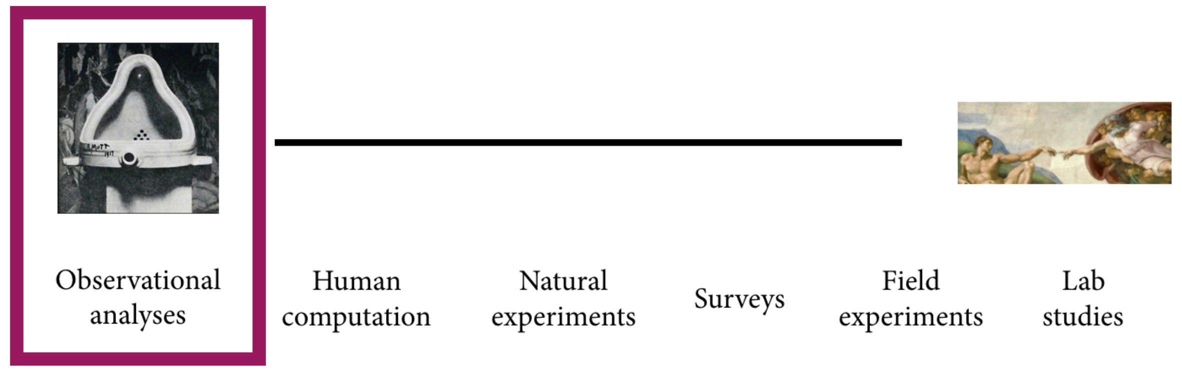
- Big: statistical power, rare events, fine resolution
- Always-on: unexpected events, real-time measurement
- Nonreactive: measurement probably won't change behaviour

- Incomplete: probably won't have the ideal information you want
- Inaccessible: difficult to access (gov't, companies)
- Nonrepresentative: bad out-of-sample generalization (good in-sample)
- Drifting: Population drift, usage drift, system drift
- Algorithmically confounded: want to study behaviour, not an algorithm
- Dirty: Junk, spam
- Sensitive: Private, hard to tell what's sensitive



Observing Behaviour: Three research strategies

1. Counting things
2. Forecasting/nowcasting
3. Approximating experiments



Biases in social data

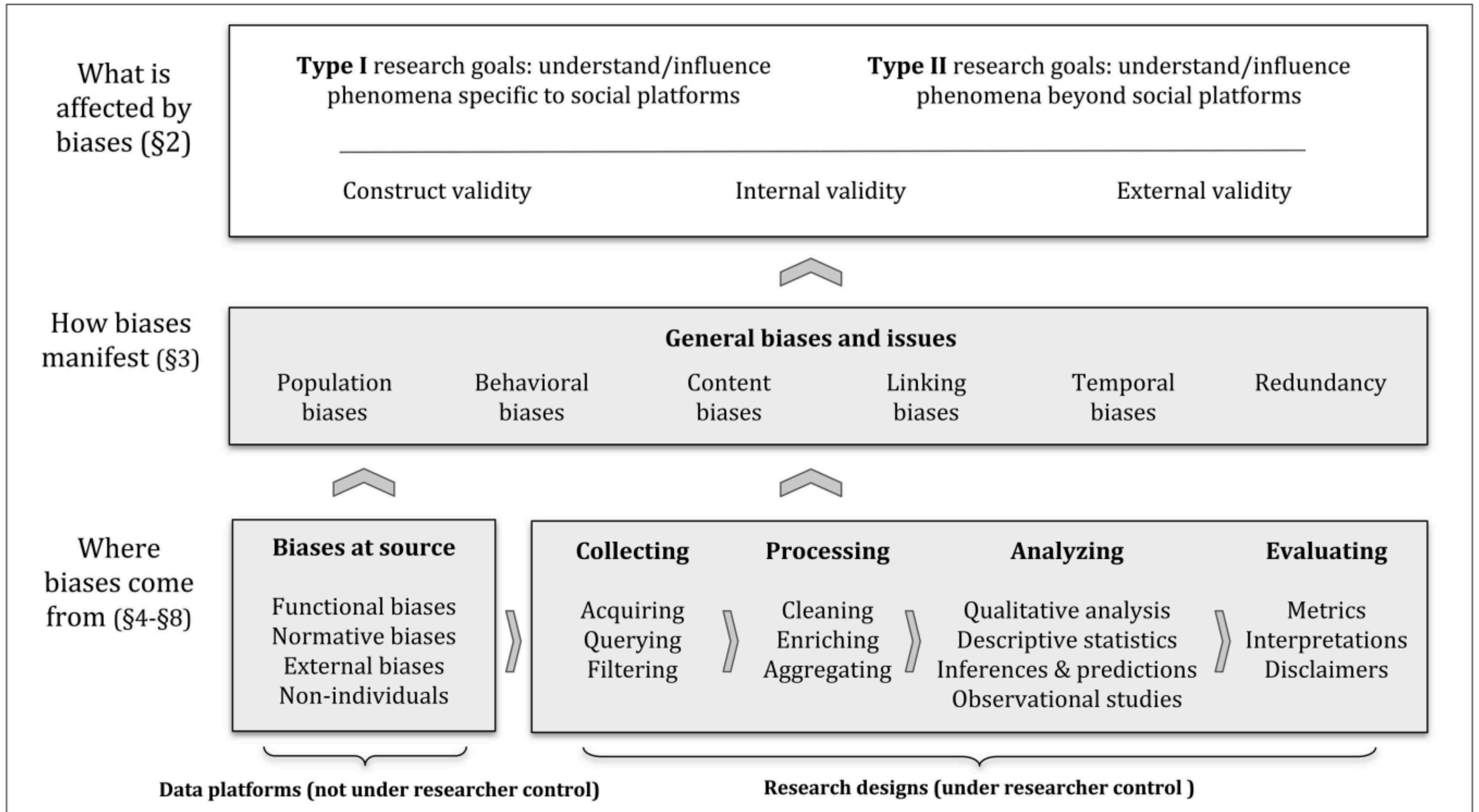
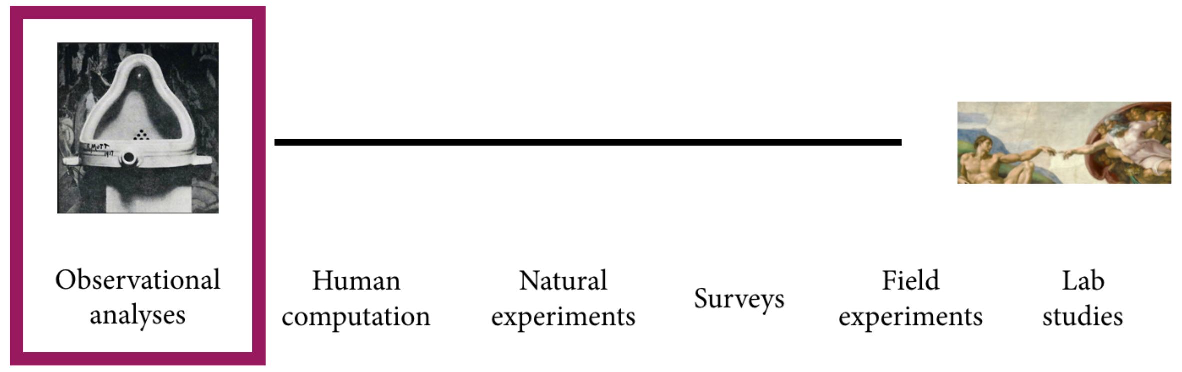


FIGURE 1 | Depiction of the framework we use to describe biases and pitfalls when working with social data. The arrows indicate how components in our framework directly tend to affect others, indicating that reaching certain social data analysis goals (section 2.1) requires research to satisfy certain validity criteria (section 2.2), which can be compromised by biases and other issues with social data (section 3). These biases and issues may occur at the source of the data (section 4), or they may be introduced along the data analysis pipeline (sections 5–8). See section 2.3 for a more detailed description.

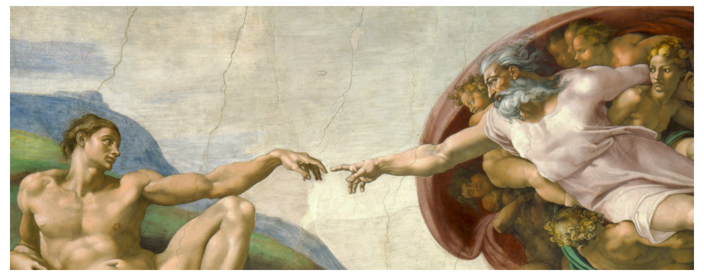
Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]*

[1] Microsoft Research, New York, NY, United States, [2] Microsoft Research, Montreal, QC, Canada, [3] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, [4] Microsoft Research, Redmond, WA, United States



Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

Surveys

Field experiments

Experiments

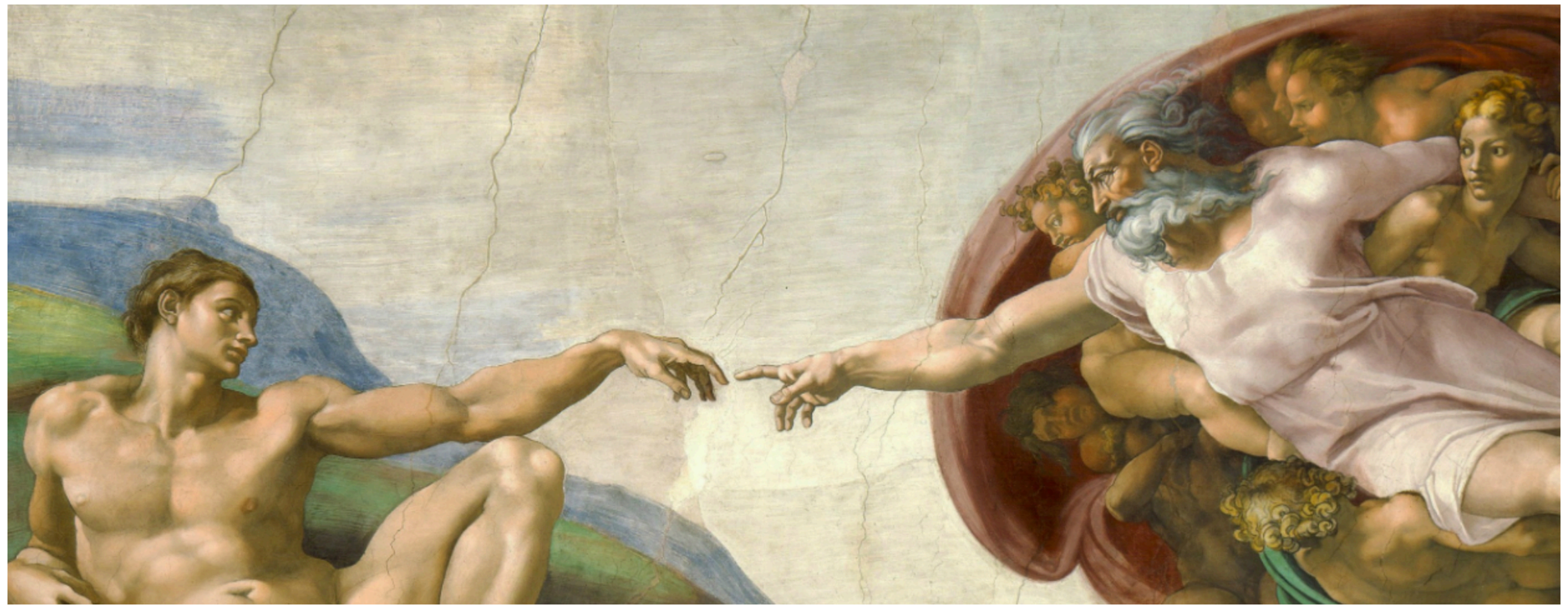
Experiments

On the other end of the spectrum is experimentation

The goal is to learn about causal relationships (cause-and-effect questions)

The strategy is to directly manipulate the environment and observe the consequences

Design the ideal scenario that will create just the data you need to answer your question



Experiments

Here, researchers intervene in the world to isolate and study a specific question

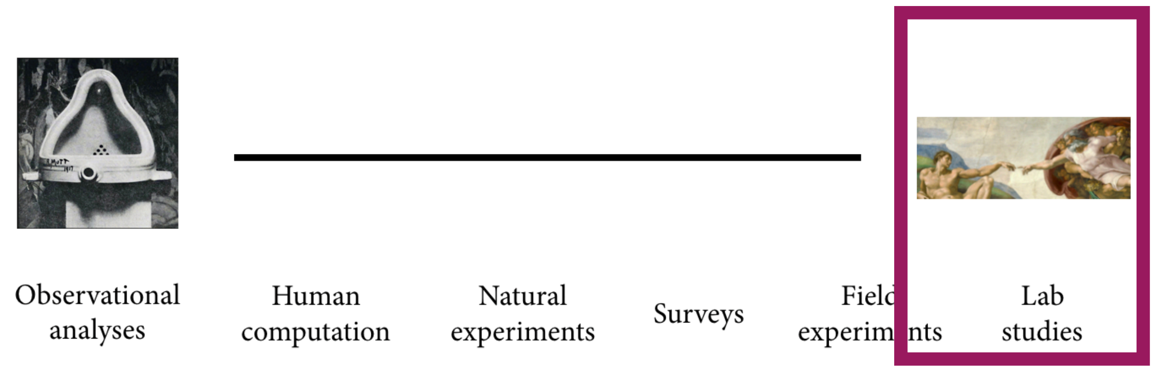
<u>Nomenclature:</u>

“Experiment”: perturb and observe

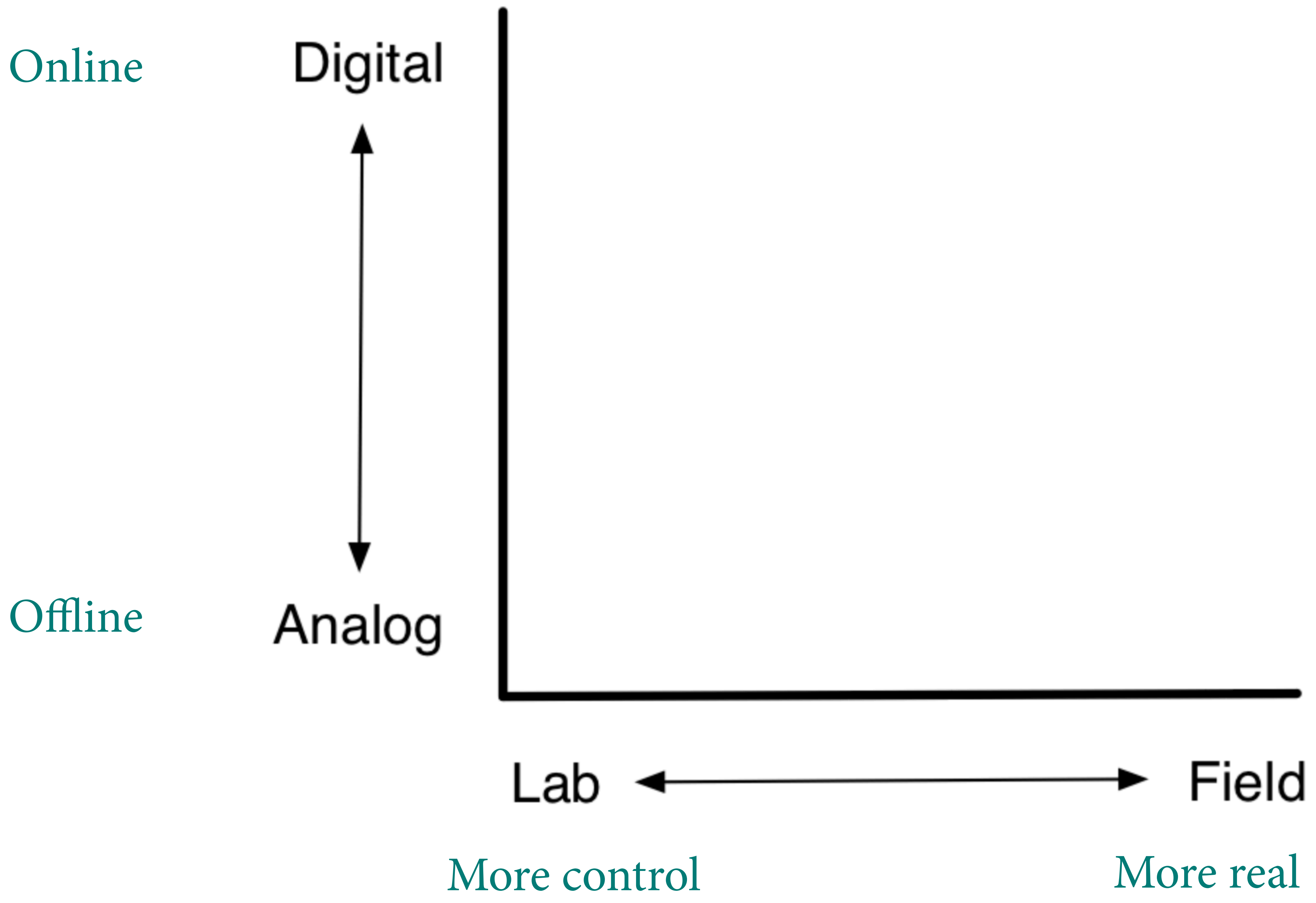
“Randomized controlled experiment”: Intervene for one group, don’t for another (randomly)

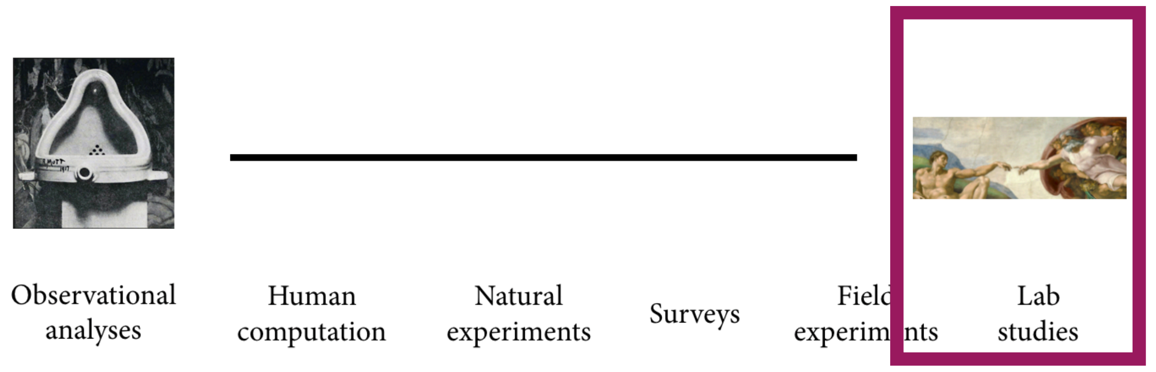
Correlation is not causation

Observational data often plagued by unknown or hard-to-control confounding variables

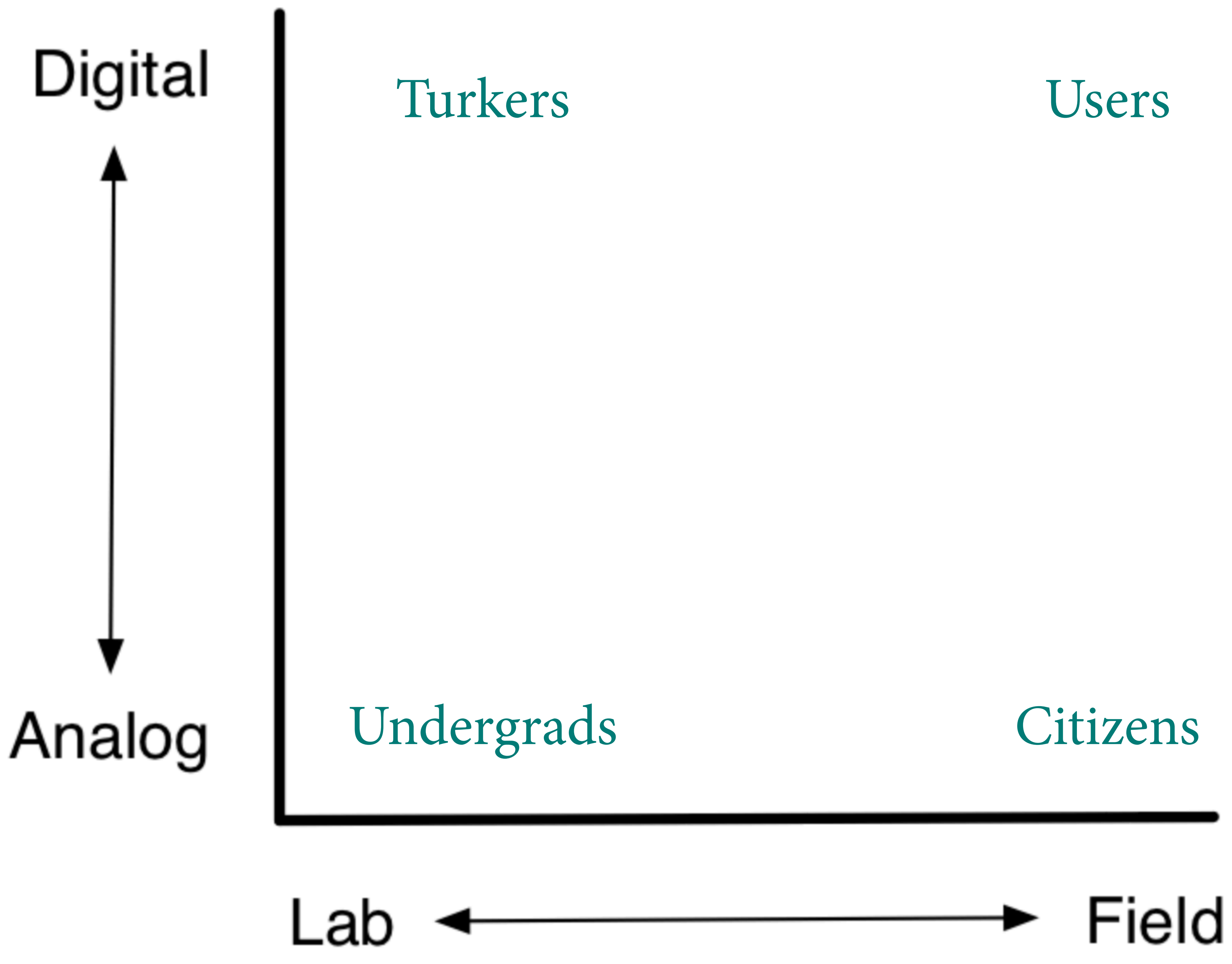


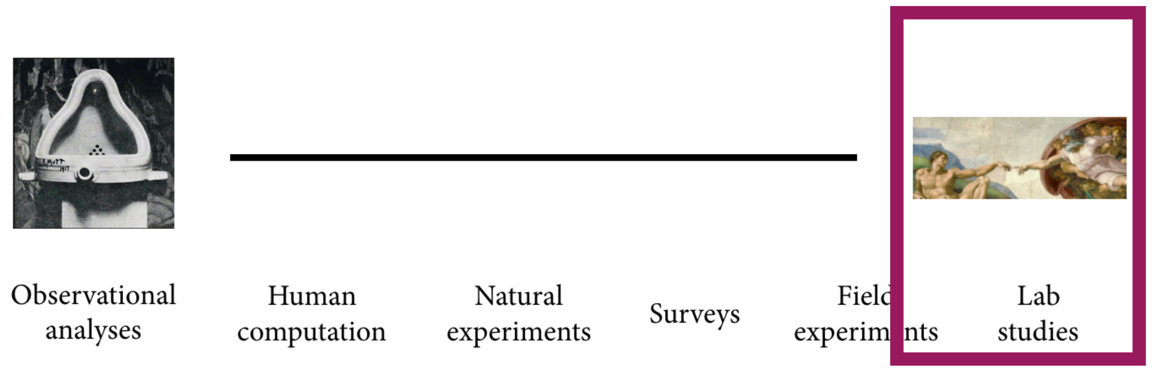
Experiments





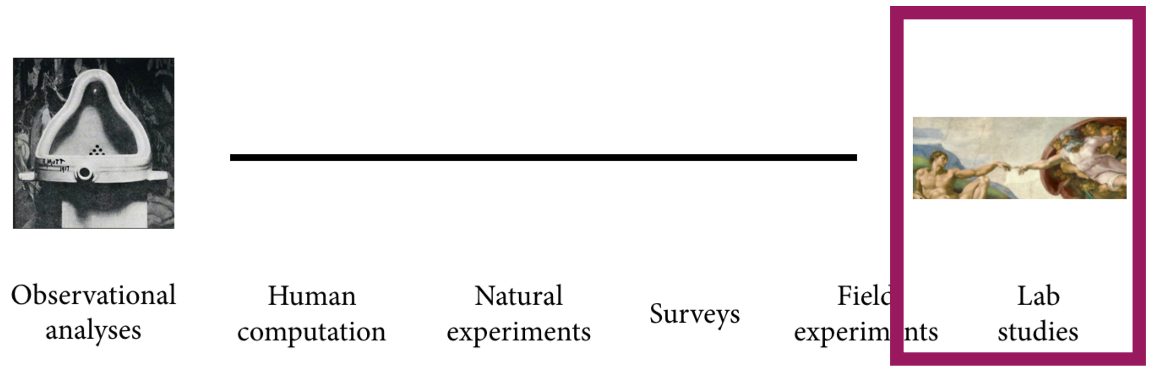
Experiments





Three major components of rich experiments

1. Validity
2. Heterogeneity
3. Mechanisms

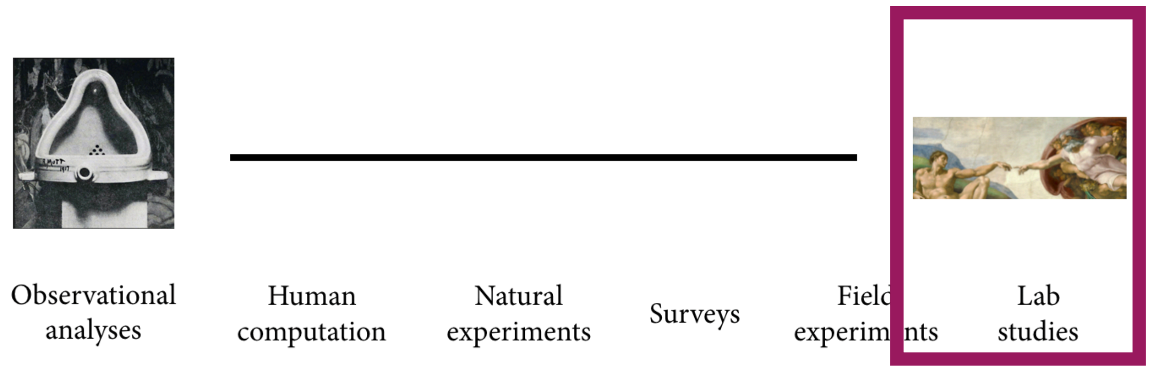


Three major components of rich experiments: 1. Validity

Validity: How general are the results?

Types of validity:

1. Statistical conclusion validity: were the stats done right?
2. Internal validity: was the experiment done right?
3. Construct validity: are we measuring the right thing?
4. External validity: is this applicable in other settings?

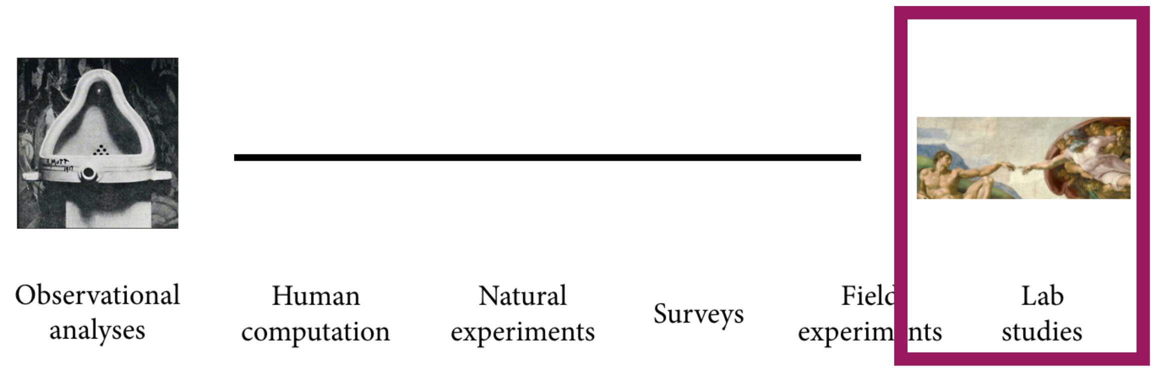


Three major components of rich experiments: 2. Heterogeneity

Barebones experiment: measure the average treatment effect (ATE)

But in social research, people almost always vary.

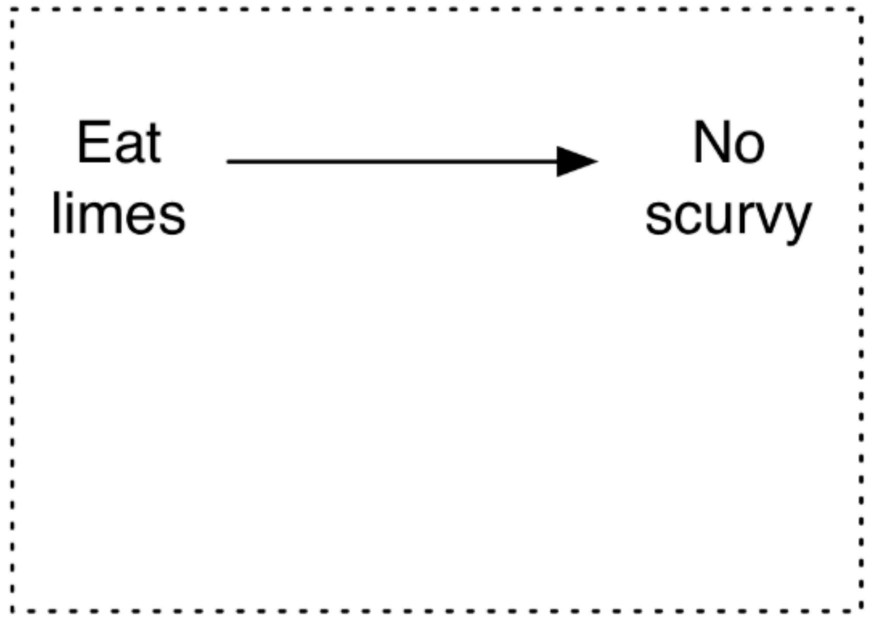
Digital research presents many more opportunities to measure how causes affect people differently



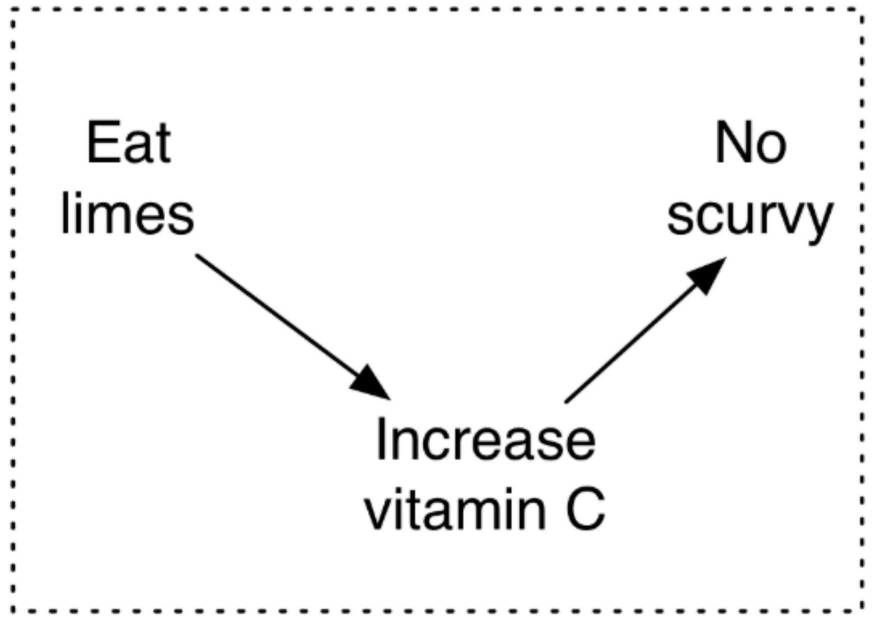
Three major components of rich experiments: 3. Mechanisms

Barebones experiment: measure what happened.

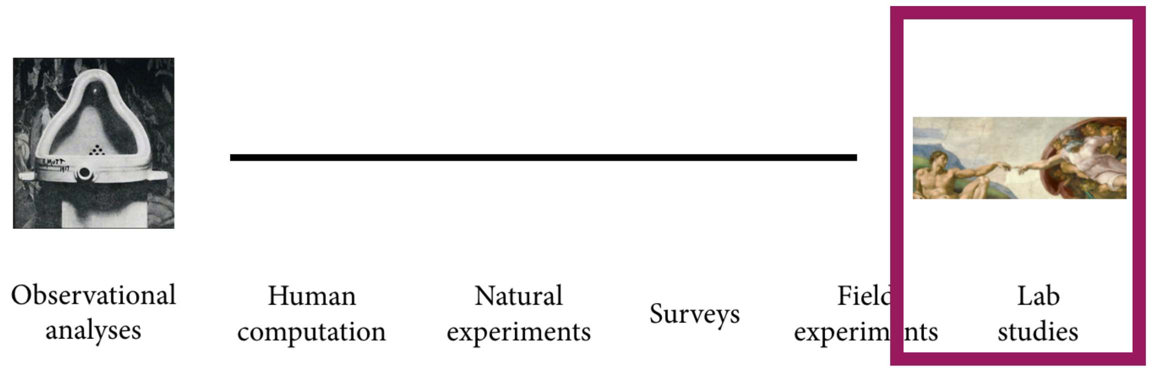
Mechanisms: why and how did it happen?



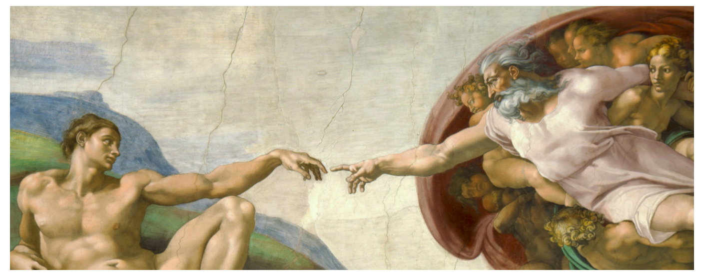
Causal effect without mechanism



Causal effect with mechanism



Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

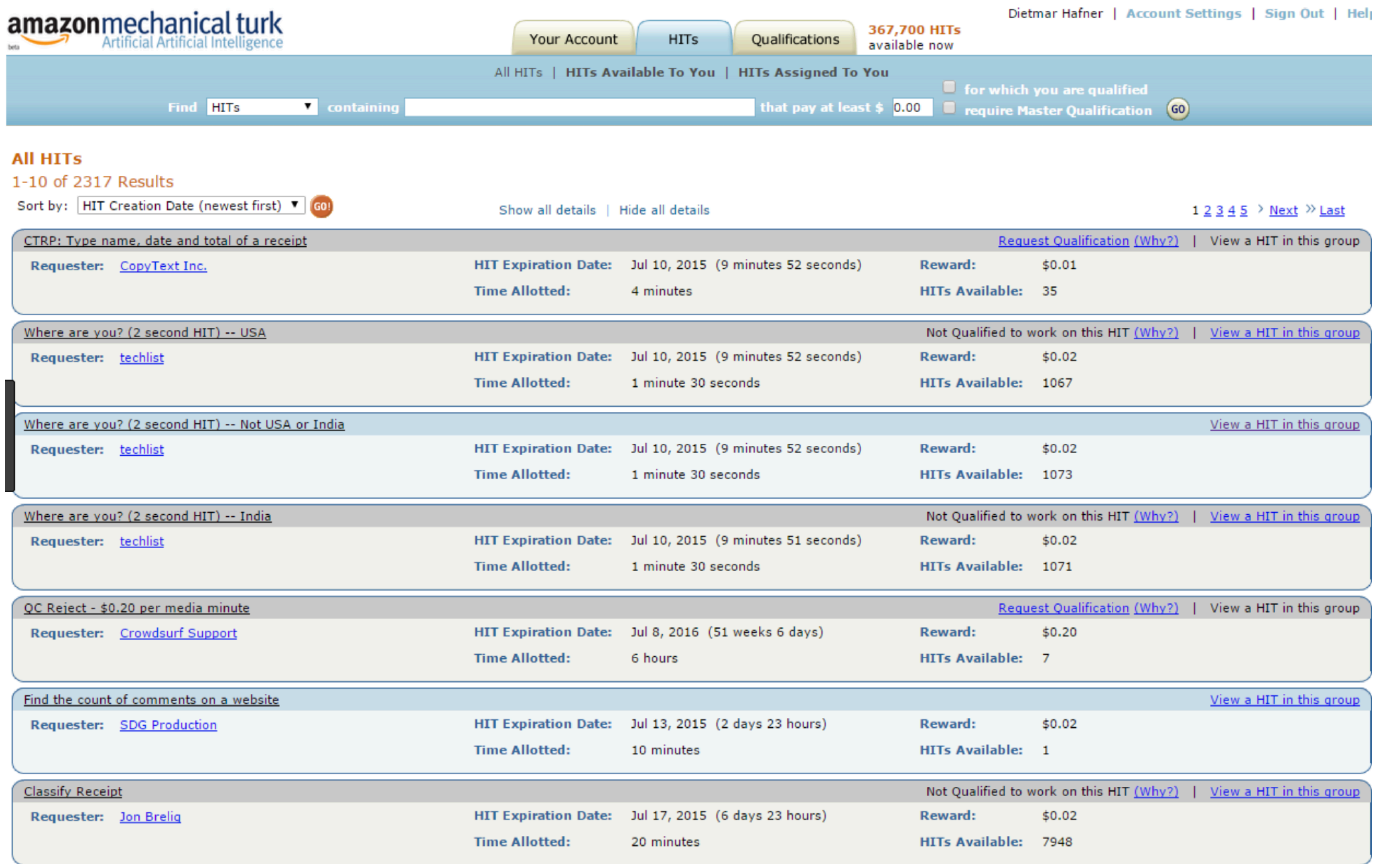
Surveys

Field experiments

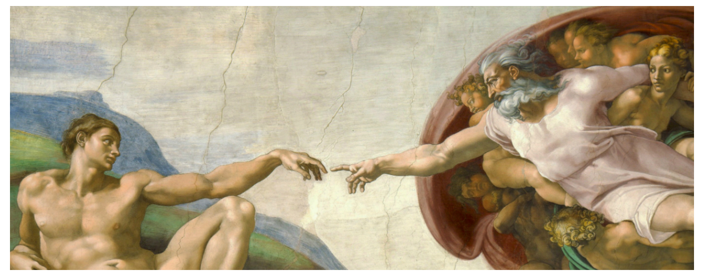
Experiments

Human computation

- Online crowdsourcing platforms allow dividing work into microtasks
- Human-in-the-loop computing, modern-day lab studies, mass collaboration to build big resources (Wikipedia etc.)



Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

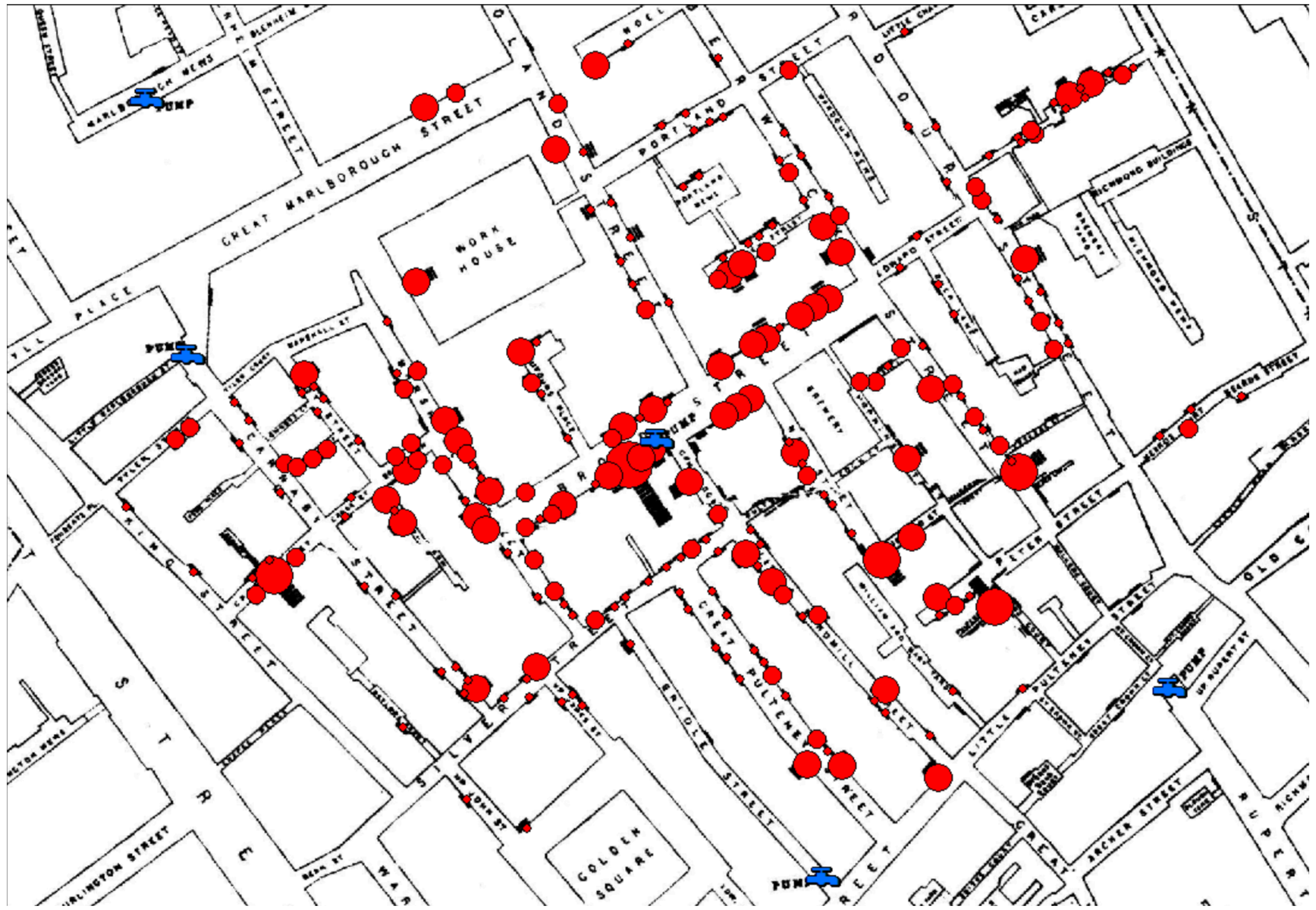
Surveys

Field experiments

Experiments

Natural experiments

Sometimes observational data has some random component you can exploit, and analyze as a “natural” experiment

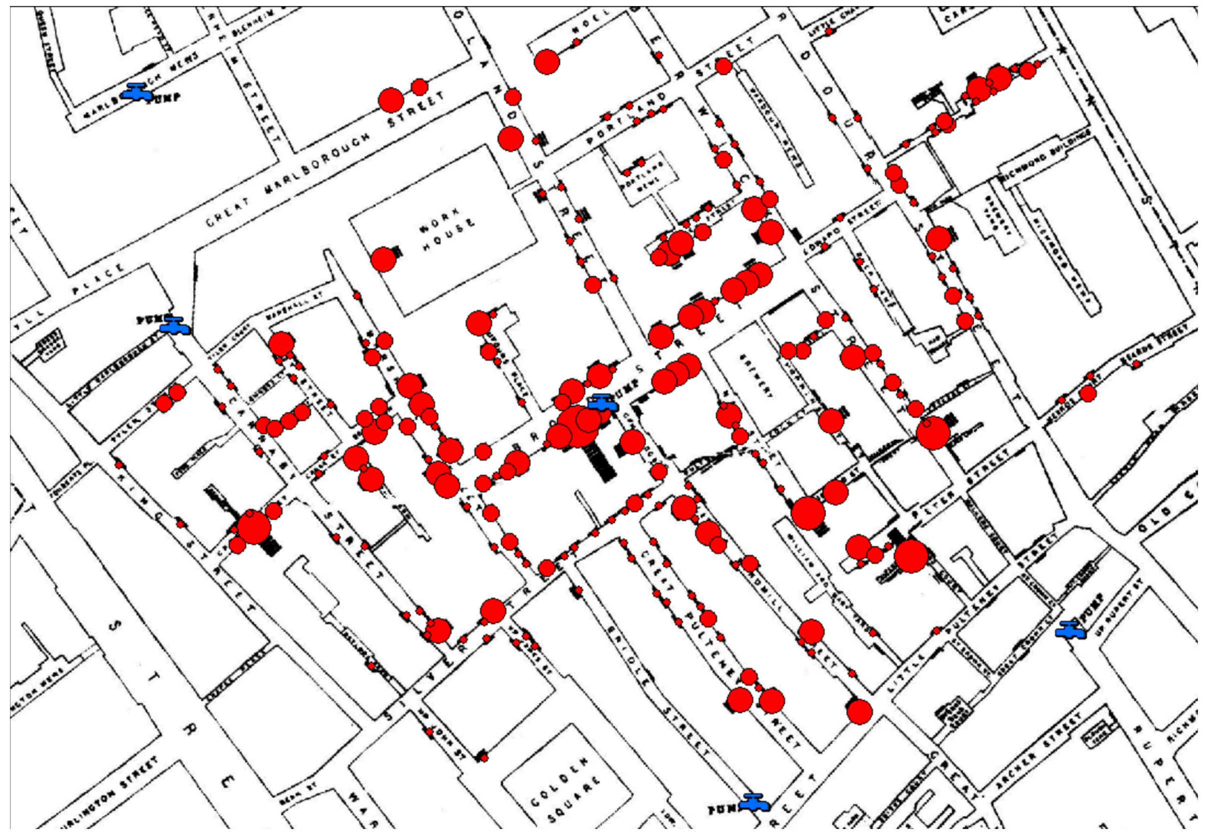


Cholera outbreak in London in 1850s

Natural experiments

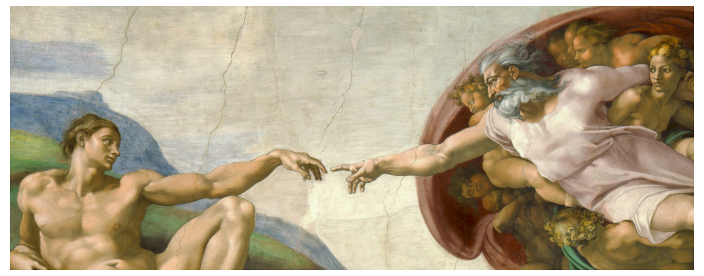
- Physician John Snow produced a map suggesting particular water was the culprit
- Two main water suppliers: one from downstream Thames where raw sewage was dumped in the water (high attack rates), and one from upstream (low attack rates)
- Which supplier you had was pretty arbitrary (varied even within same house, same neighbourhood, etc.)
- Exposure to polluted water was as-if random

Now: in large datasets, more opportunities to identify and argue for as-if random assignment



Cholera outbreak in London in 1850s

Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

Surveys

Field experiments

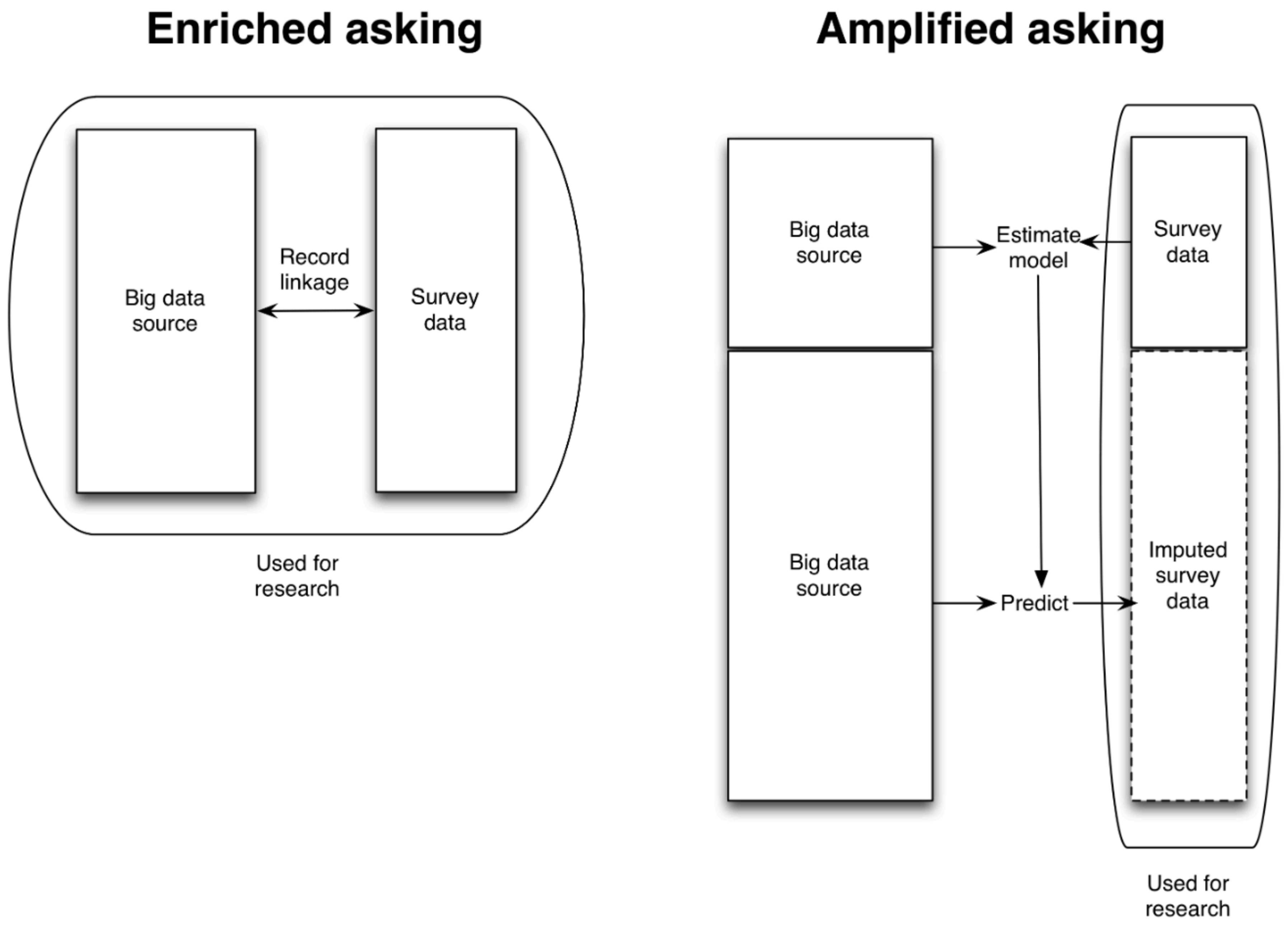
Experiments

Surveys: asking questions

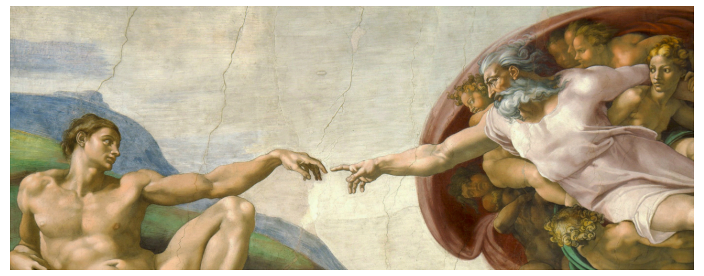
Social research has a unique advantage: we can ask our subjects what they're thinking!

Still the best way to learn the answer to many questions

In the digital era, there are new ways of asking questions



Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

Surveys

Field experiments

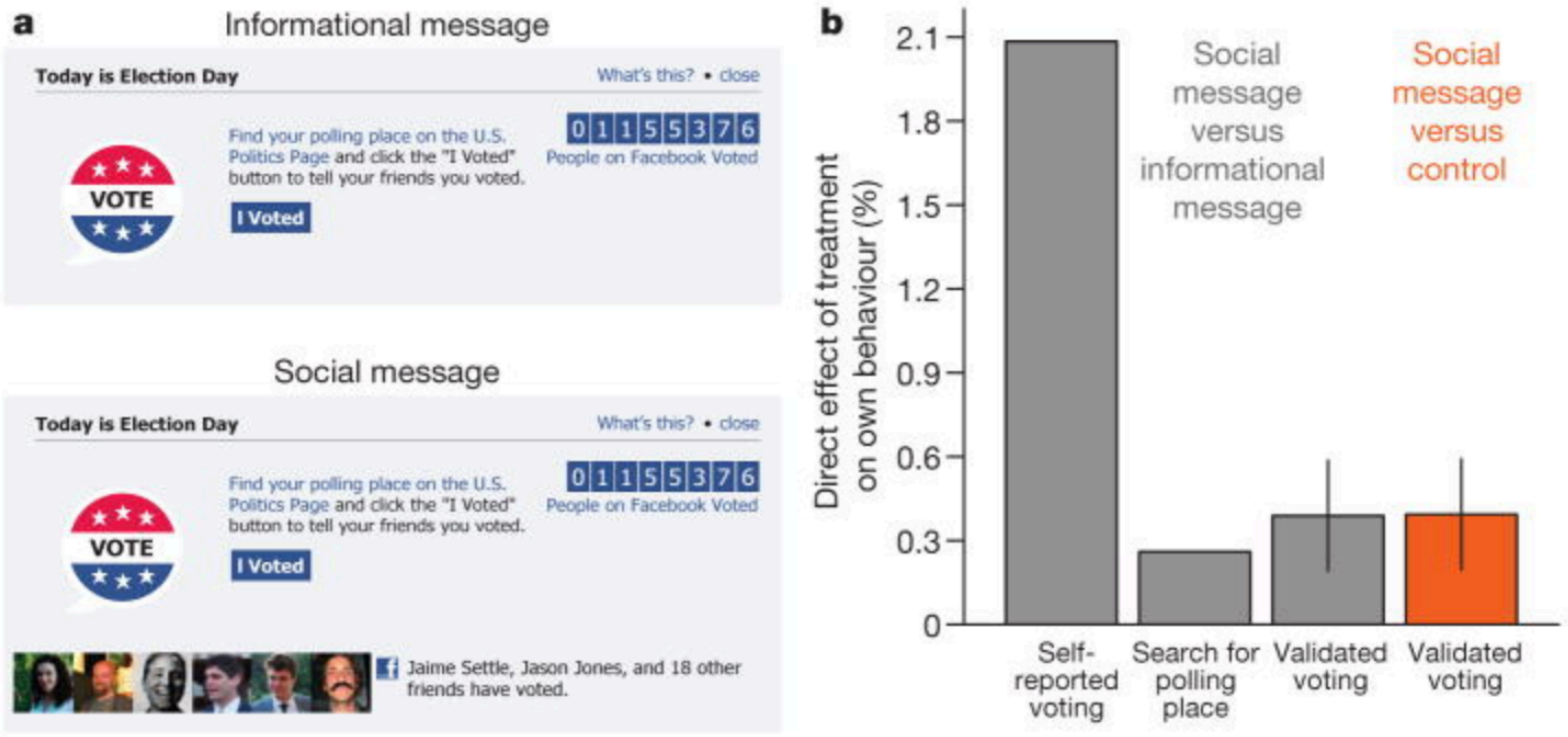
Experiments

Field experiments

- Introducing a treatment into a real system
- Much more possible now with algorithmic systems

Voting experiment on Facebook

Figure 1



The experiment and direct effects

a, **b**, Examples of the informational message and social message Facebook treatments (**a**) and their direct effect on voting behaviour (**b**). Vertical lines indicate s.e.m. (they are too small to be seen for the first two bars).

~300,000 more validated votes

AI, Data, and Society: Algorithmic decision-making

Example: St. George's Hospital in the UK developed an algorithm to sort medical school applicants. Algorithm trained to mimic past admissions decisions made by humans.

But past decisions were biased against women and minorities. It codified discrimination.

Web search ads for “Kristen Haring”

Ads by Google

[We Found:**Kristen Haring**]

1) Contact **Kristen Haring** - Free Info! 2) Current Phone, Address & More.

www.peoplesmart.com/**Kristen**

Search by Phone	Search by Email
Background Checks	Search by Address
Public Records	Criminal Records

Kristen Haring

Public Records Found For: **Kristen Haring**. Search Now.

www.publicrecords.com/

Web search ads for "Latanya Farrell"

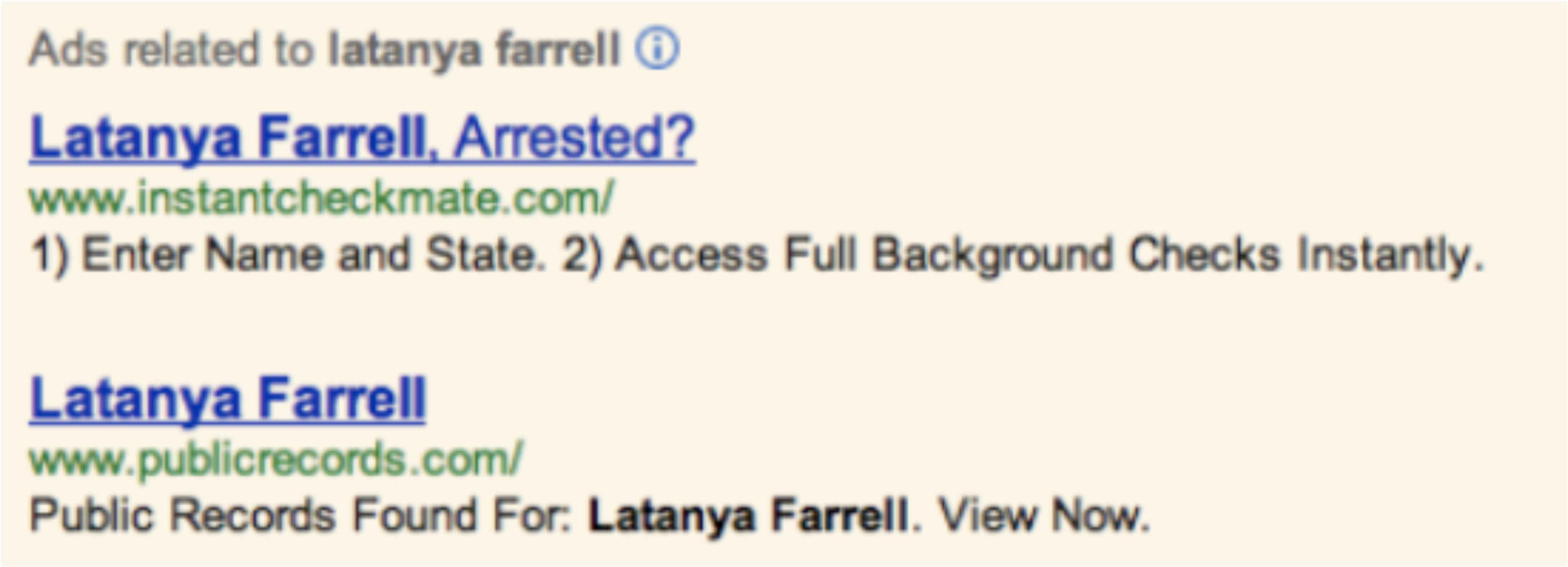


Image labeling gone wrong

Image searching for “CEO”

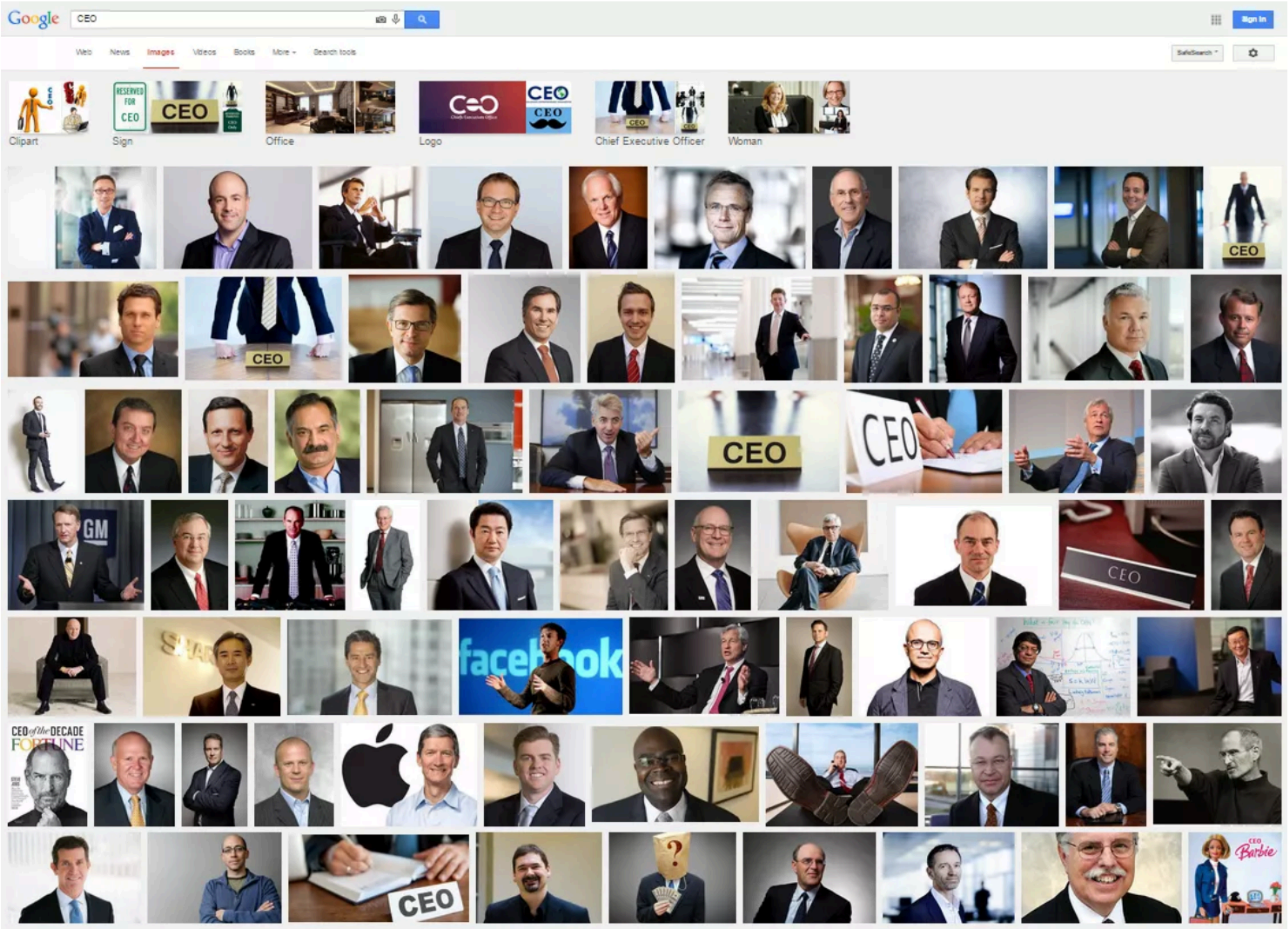


Image searching for “CEO”

By the way: this picture is from an Onion article.

Ethics and privacy

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

Facebook's Users Outraged Over Emotion Experiment

Facebook reveals news feed experiment to control emotions

Facebook emotion experiment sparks criticism

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

Everything We Know About Facebook's Secret Mood Manipulation Experiment

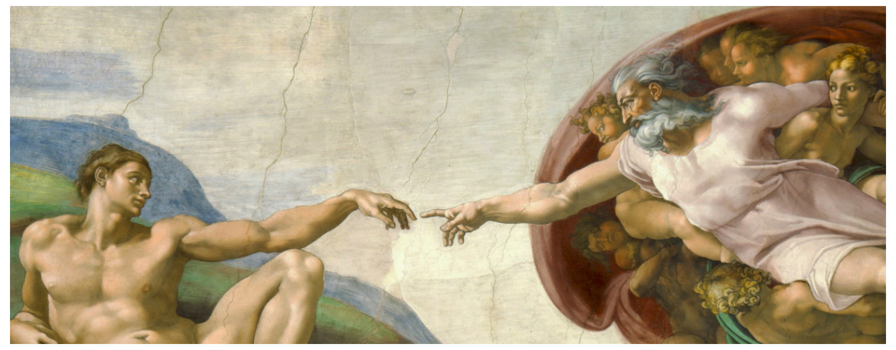
Facebook conducted secret psychology experiment on users' emotions

Computational social science in 7 easy pieces

	Week	Date	Topic	Reviews Due	Textbook Readings
	1	1/14	Introduction to computational social science		Ch. 1
	2	1/21	Introduction to computational social science cont'd		Ch. 1
★	3	1/28	Observational studies 1	1/27 9:00pm	Ch. 2
★	4	2/4	Observational studies 2	2/3 9:00pm	Ch. 2
★	5	2/11	Experiments 1	2/10 9:00pm	Ch. 4
★	6	2/25	Experiments 2	2/24 9:00pm	Ch. 4
	7	3/4	Project proposals		
★	8	3/11	Asking questions	3/10 9:00pm	Ch. 3
★	9	3/18	Mass collaboration	3/17 9:00pm	Ch. 5
★	10	3/25	Ethics in computational social science	3/24 9:00pm	Ch. 6
	11	4/1	Project presentations (Part 1)		
	12	4/8	Project presentations (Part 2)		

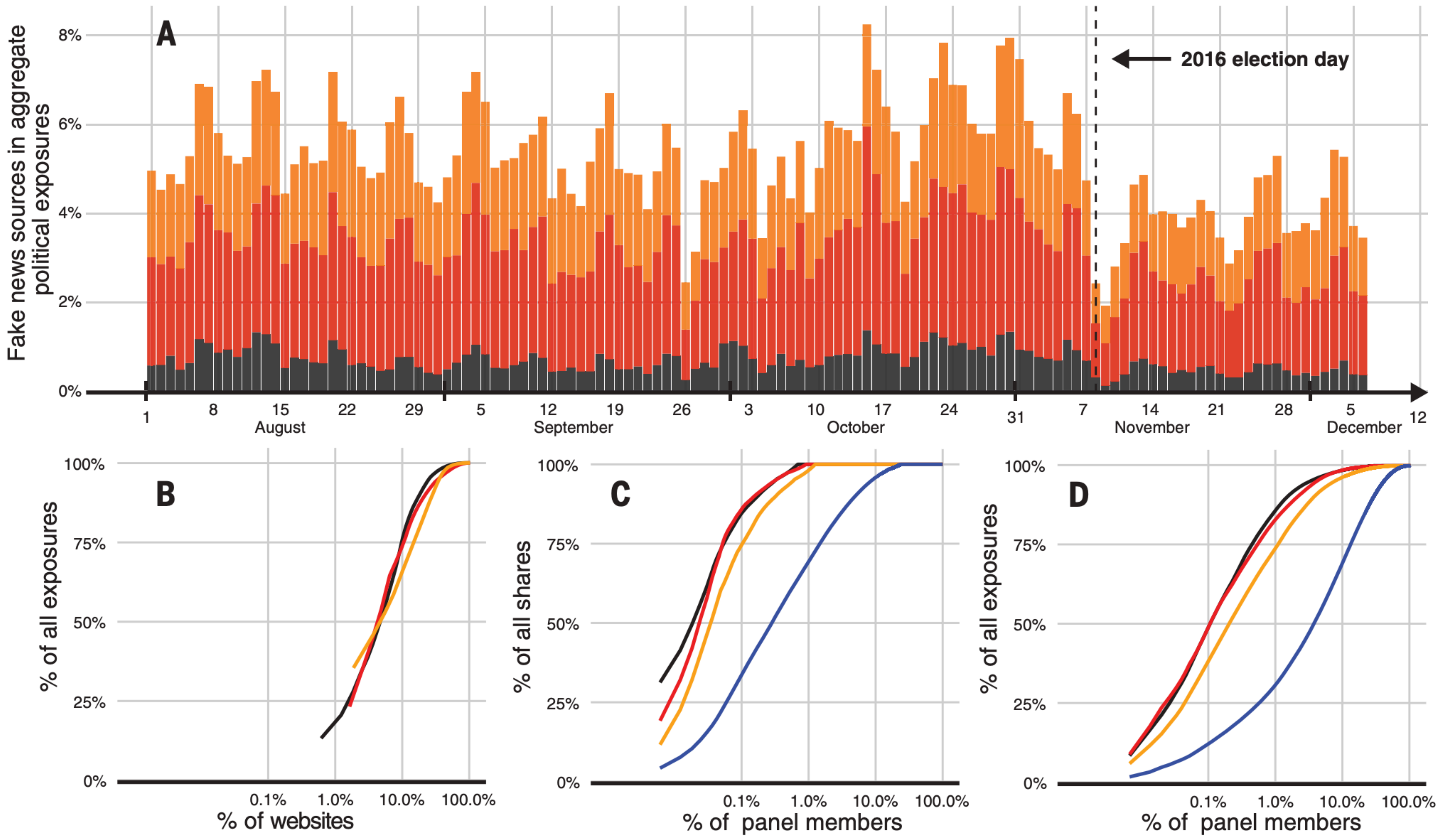
Readymades

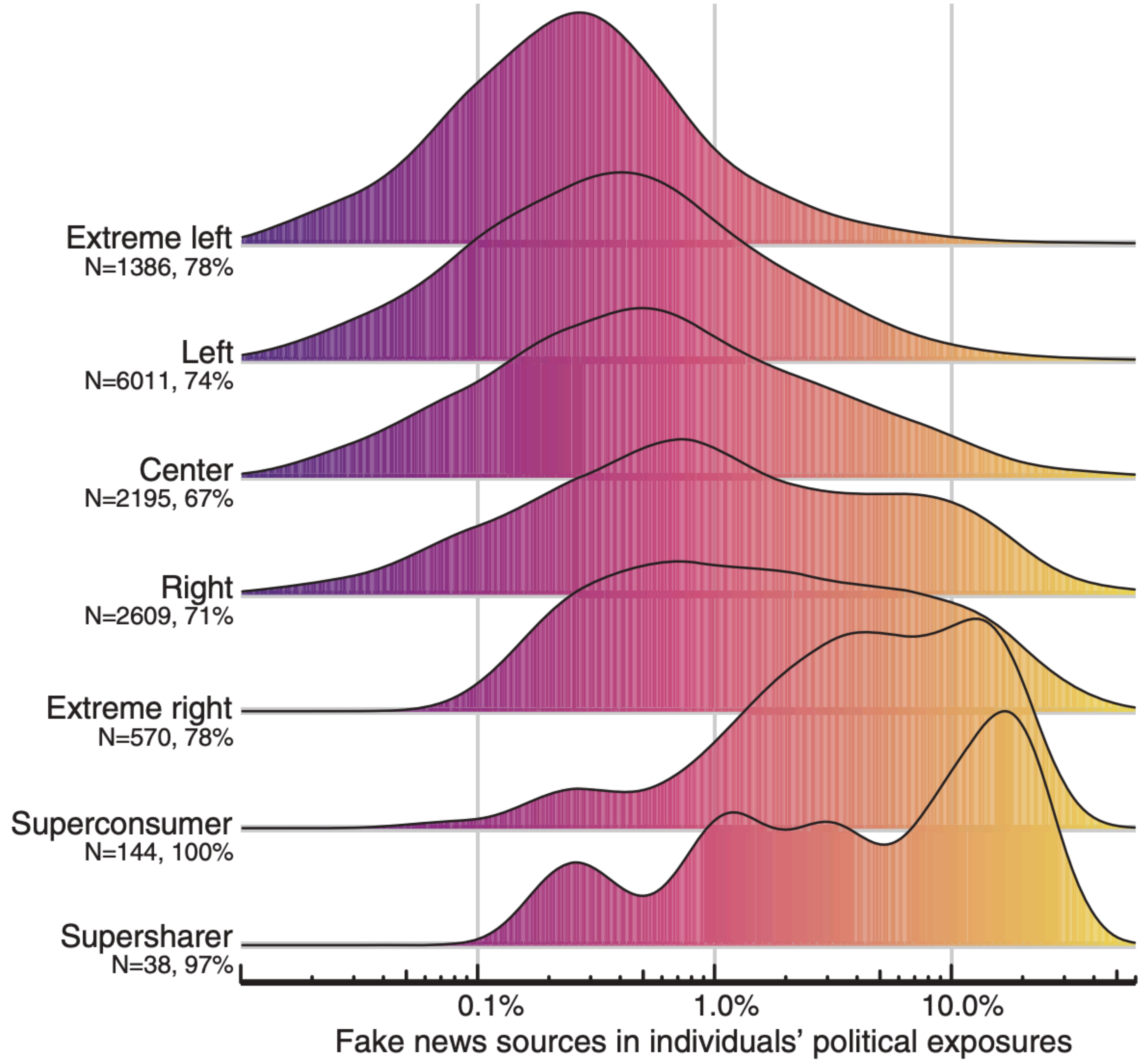
Custommades



Observational studies 1

Analysis of exposure/sharing of fake news by registered voters on Twitter

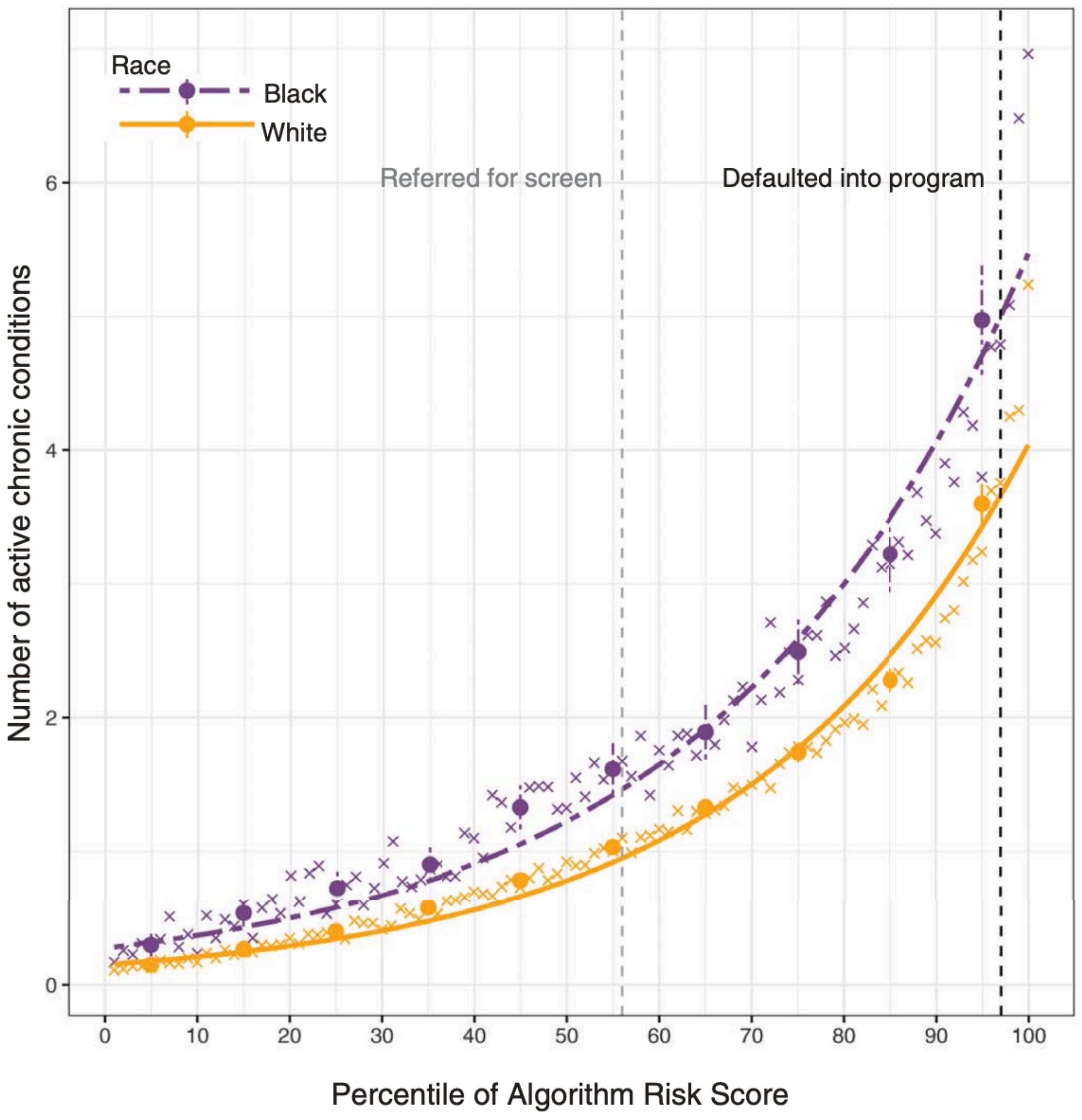




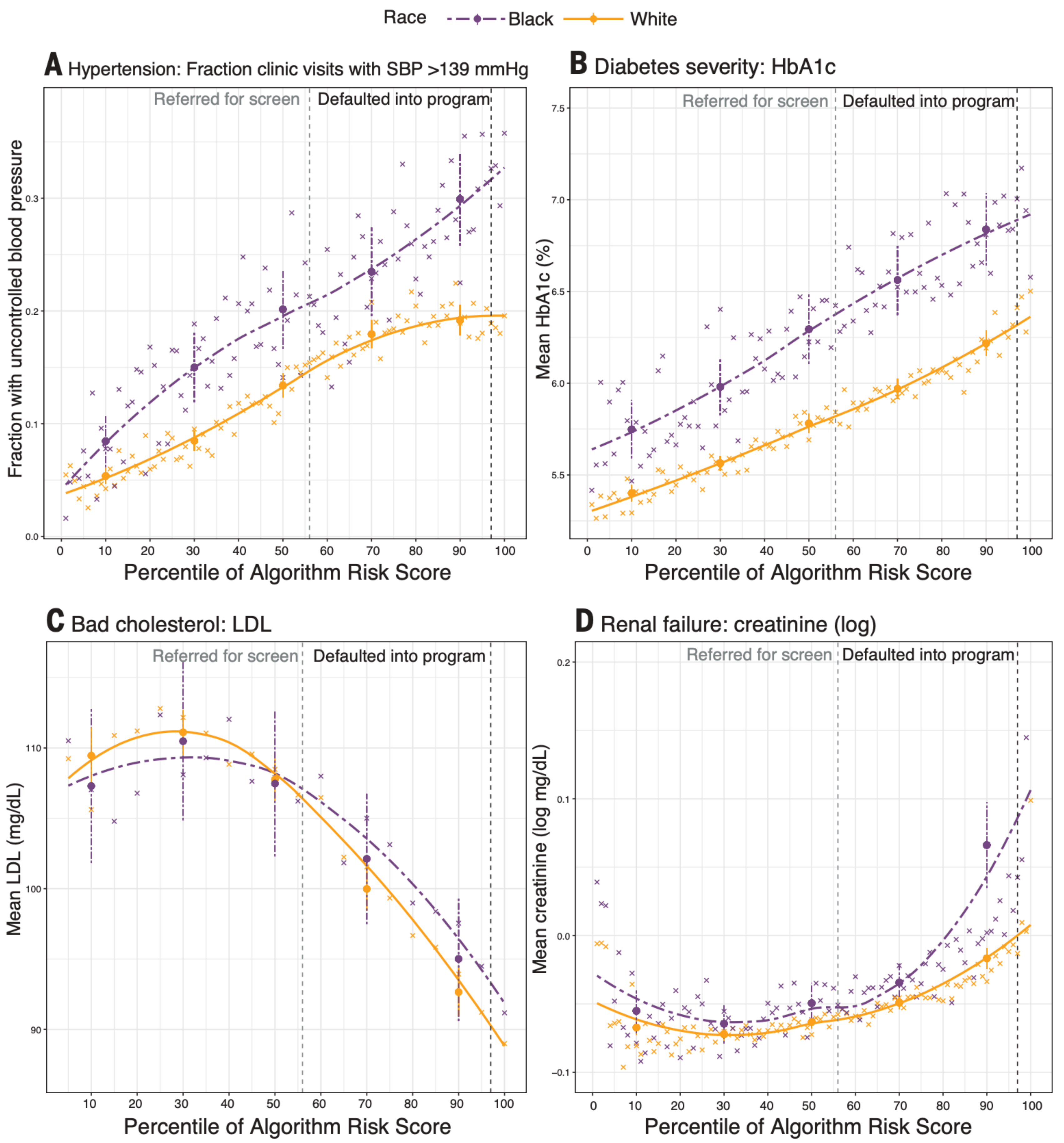
Observational studies 1

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†



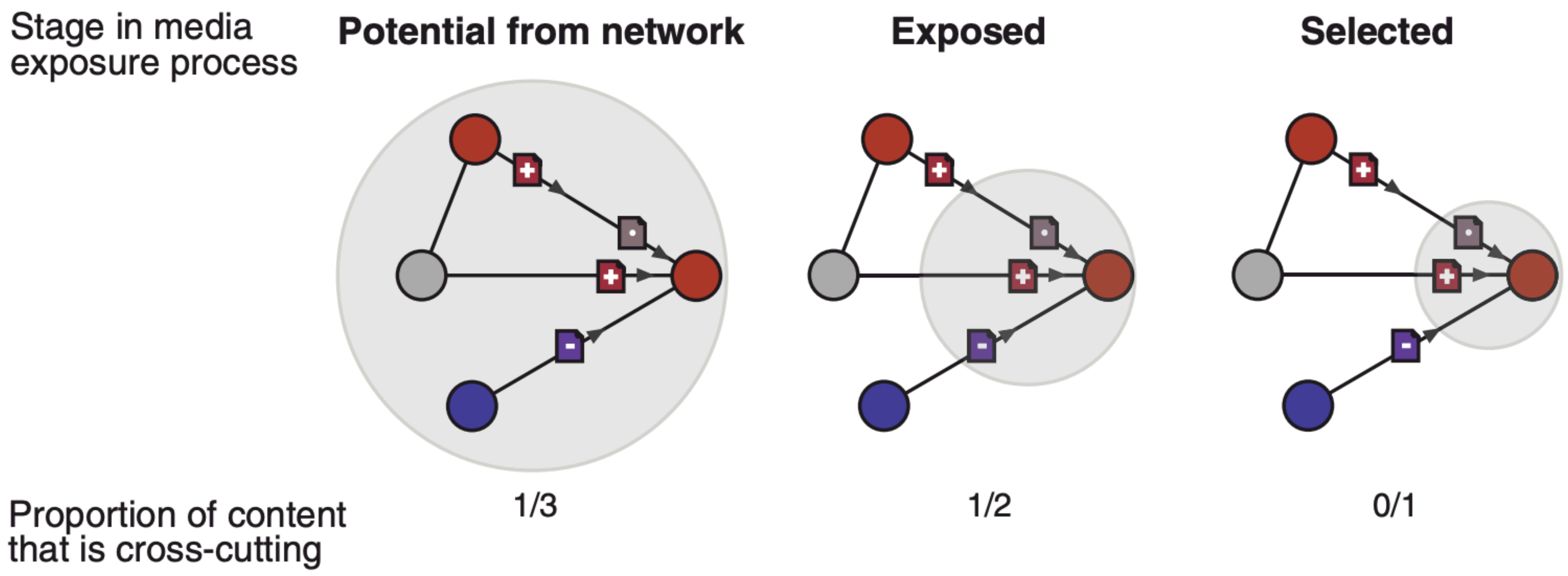
Measuring algorithmic bias in a high-stakes health setting



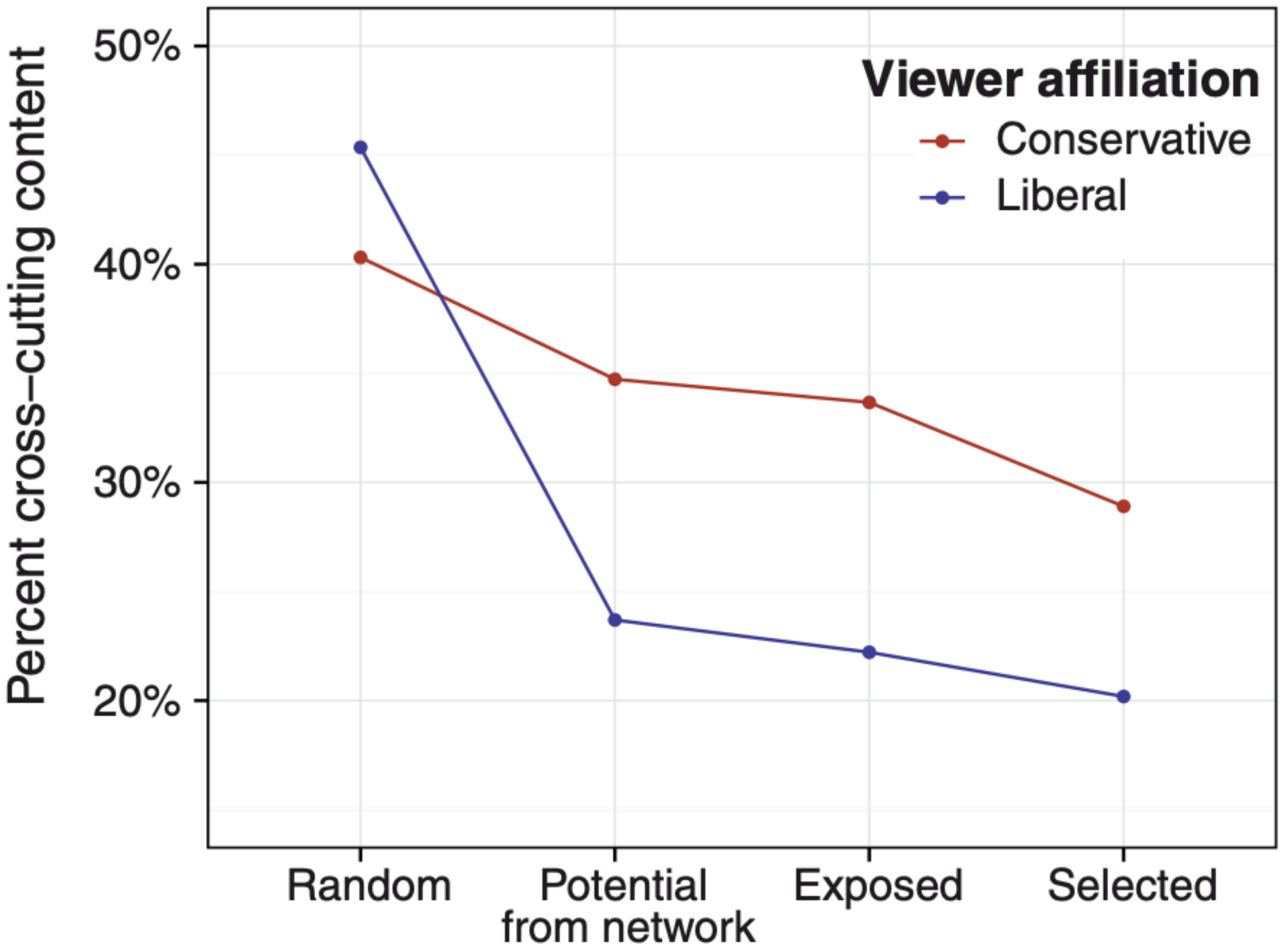
Observational studies 2

Exposure to ideologically diverse news and opinion on Facebook

Eytan Bakshy,[1*†] Solomon Messing,[1†] Lada A. Adamic[1,2]



Measuring algorithmic "filter bubble" effects on Facebook

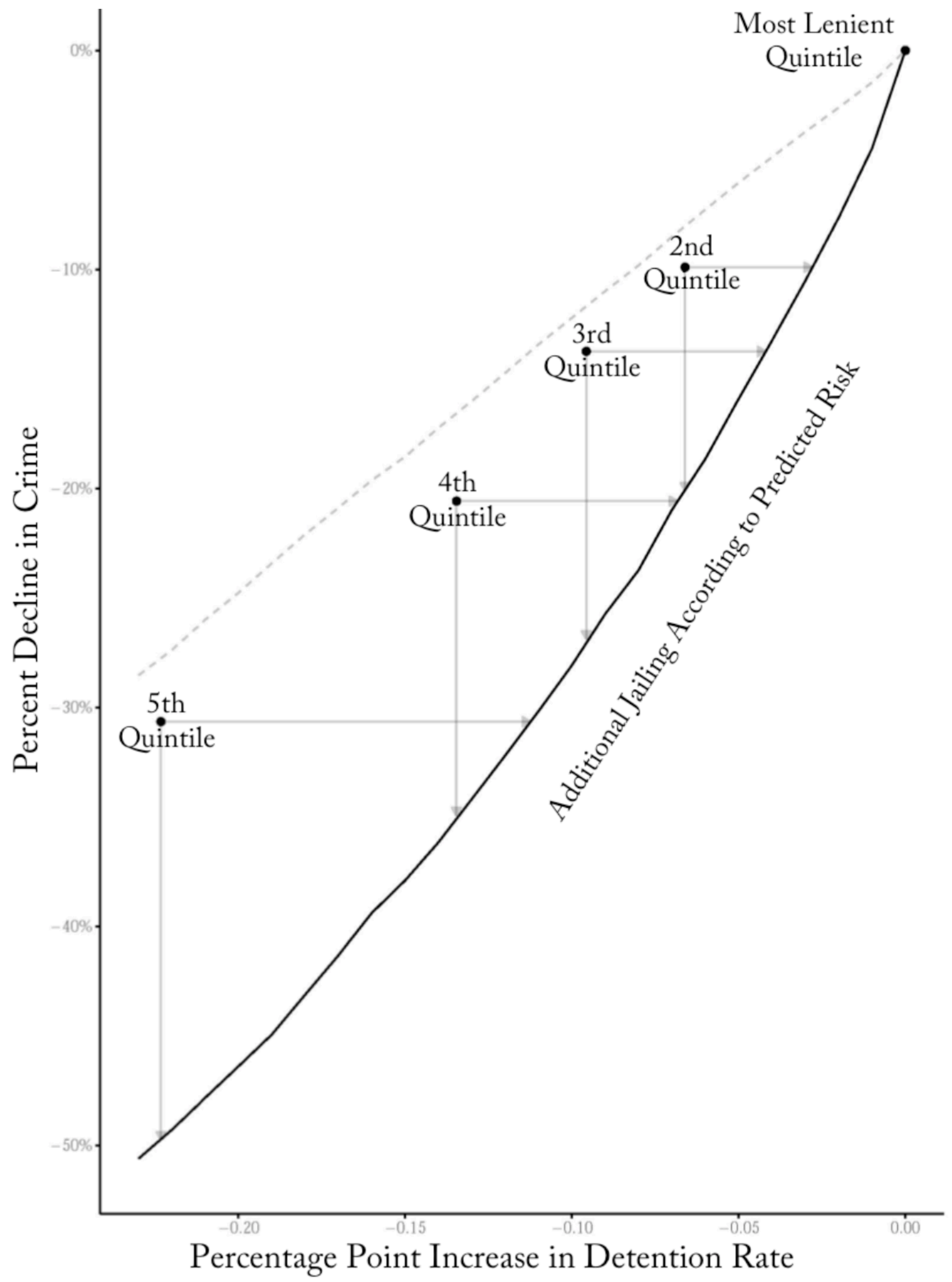


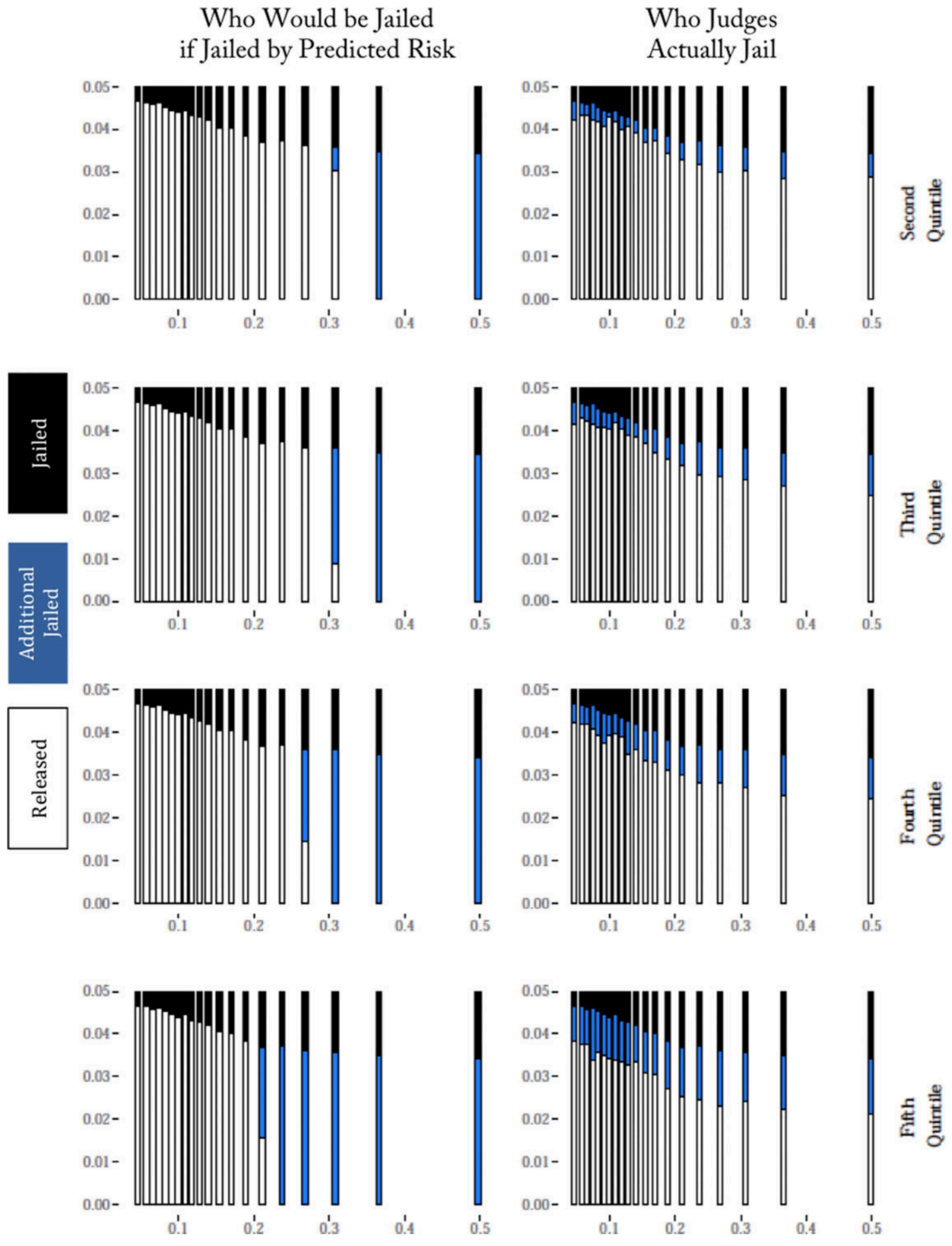
Observational studies 2

HUMAN DECISIONS AND MACHINE PREDICTIONS*

JON KLEINBERG
HIMABINDU LAKKARAJU
JURE LESKOVEC
JENS LUDWIG
SENDHIL MULLAINATHAN

758K pretrial bail decisions after arrests in NYC 2008–2013





Experiments 1

The Role of Social Networks in Information Diffusion

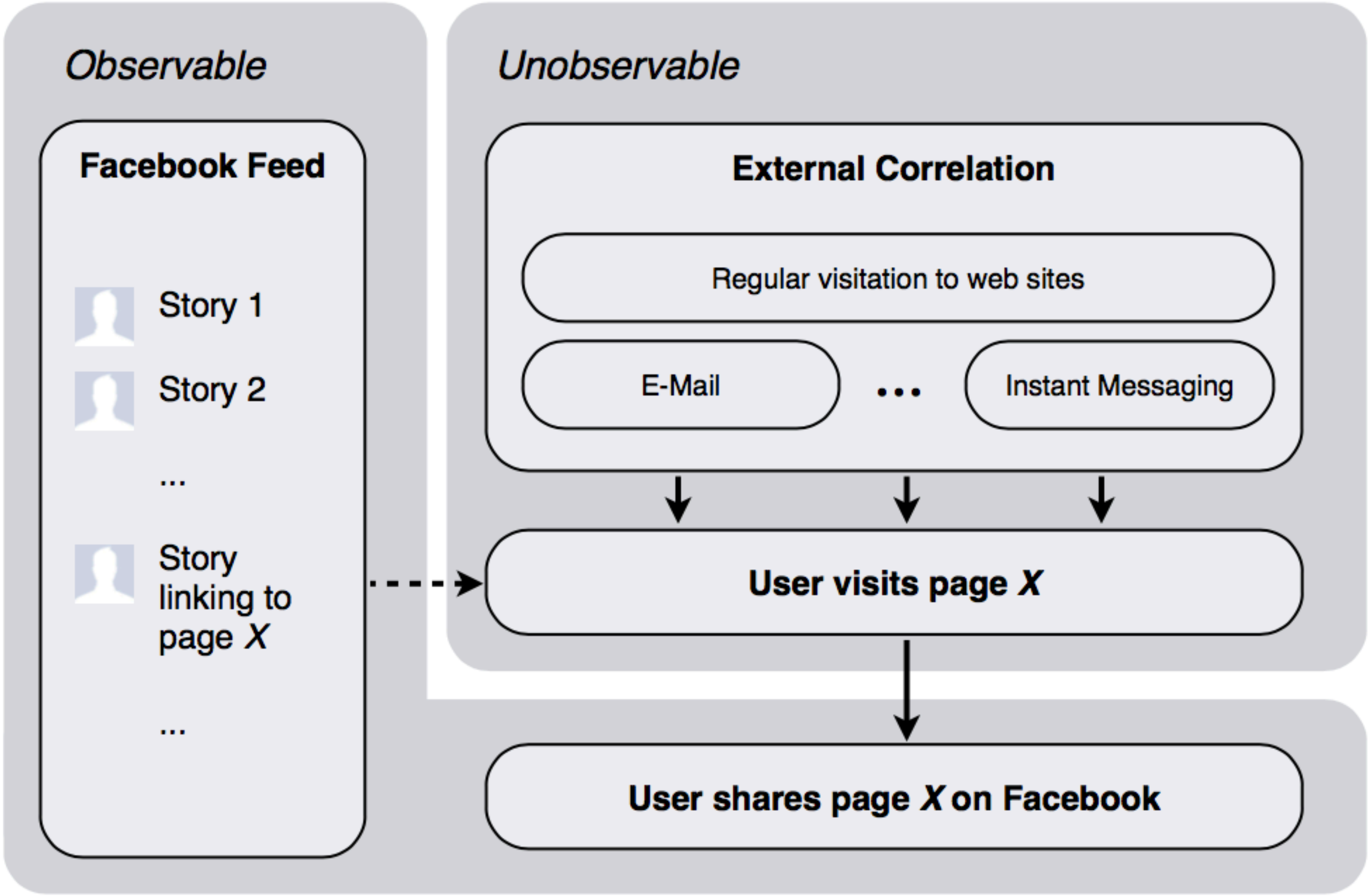
Eytan Bakshy*
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
ebakshy@fb.com

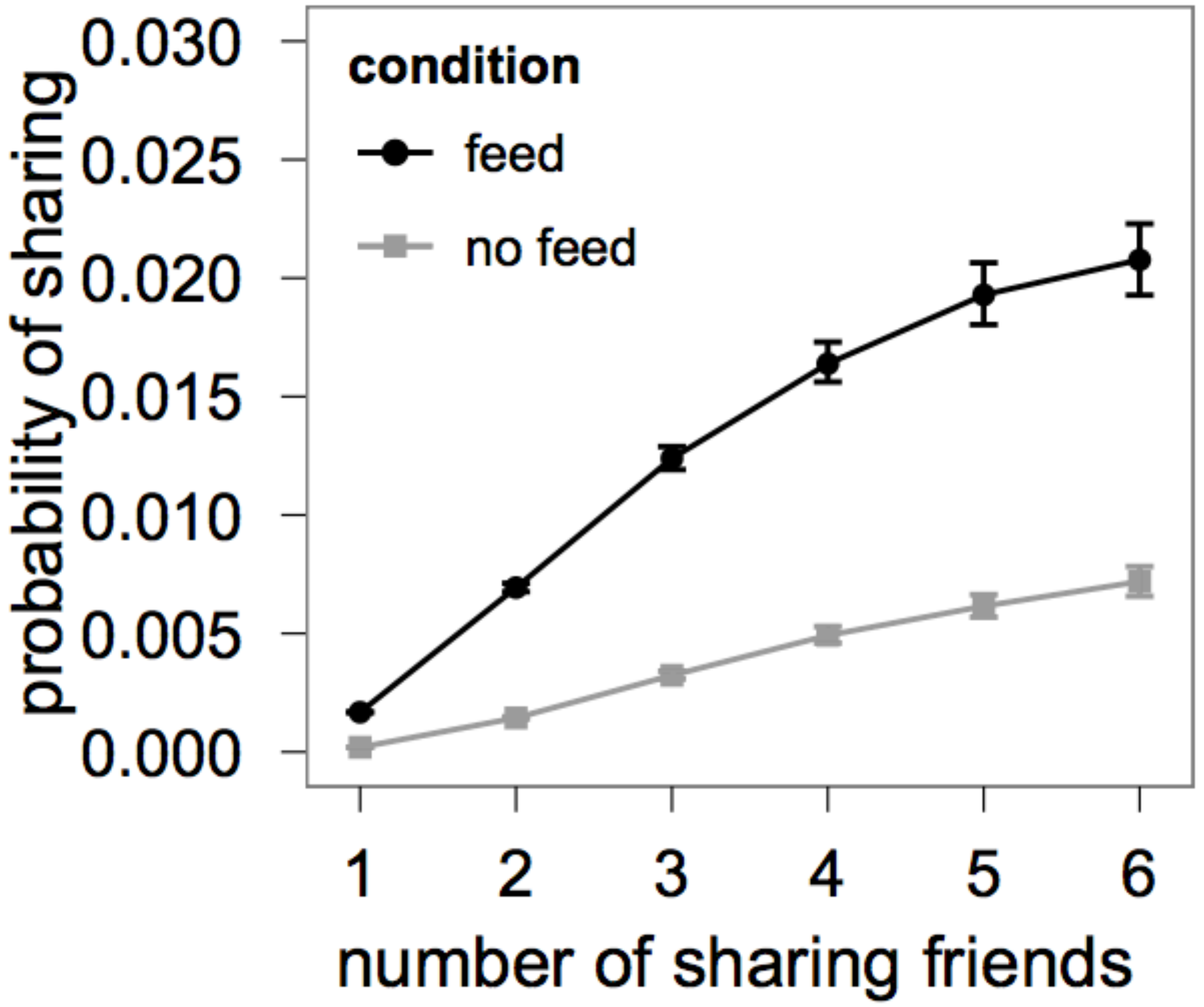
Itamar Rosenn
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
itamar@fb.com

Cameron Marlow
Facebook
1601 Willow Rd.
Menlo Park, CA 94025
cameron@fb.com

Lada Adamic
University of Michigan
105 S. State St.
Ann Arbor, MI 48104
ladamic@umich.edu

How do social networks mediate the information you receive from your friends?



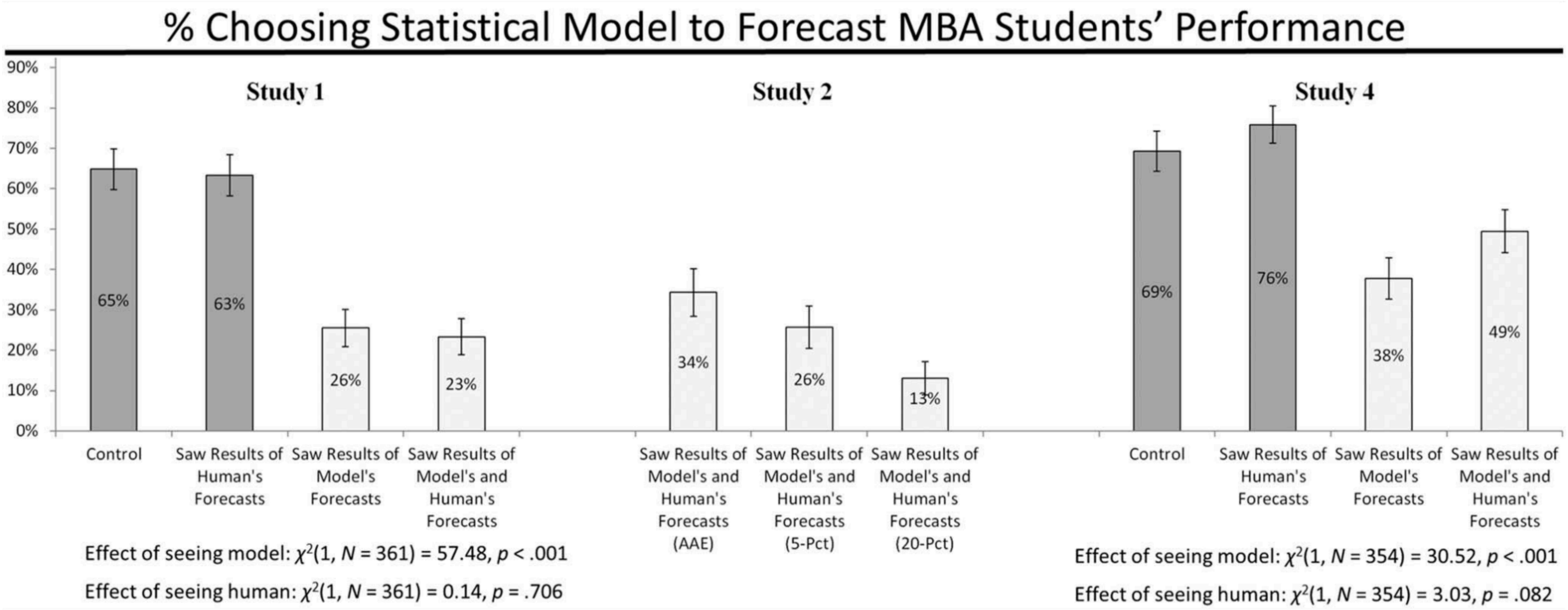


Experiments 1

Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey
University of Pennsylvania

Do people trust algorithms (even when they should)?



Experiments 2

The Welfare Effects of Social Media†

By Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow*

What are the causal effects of social media on time spent online, political polarisation, and well-being?

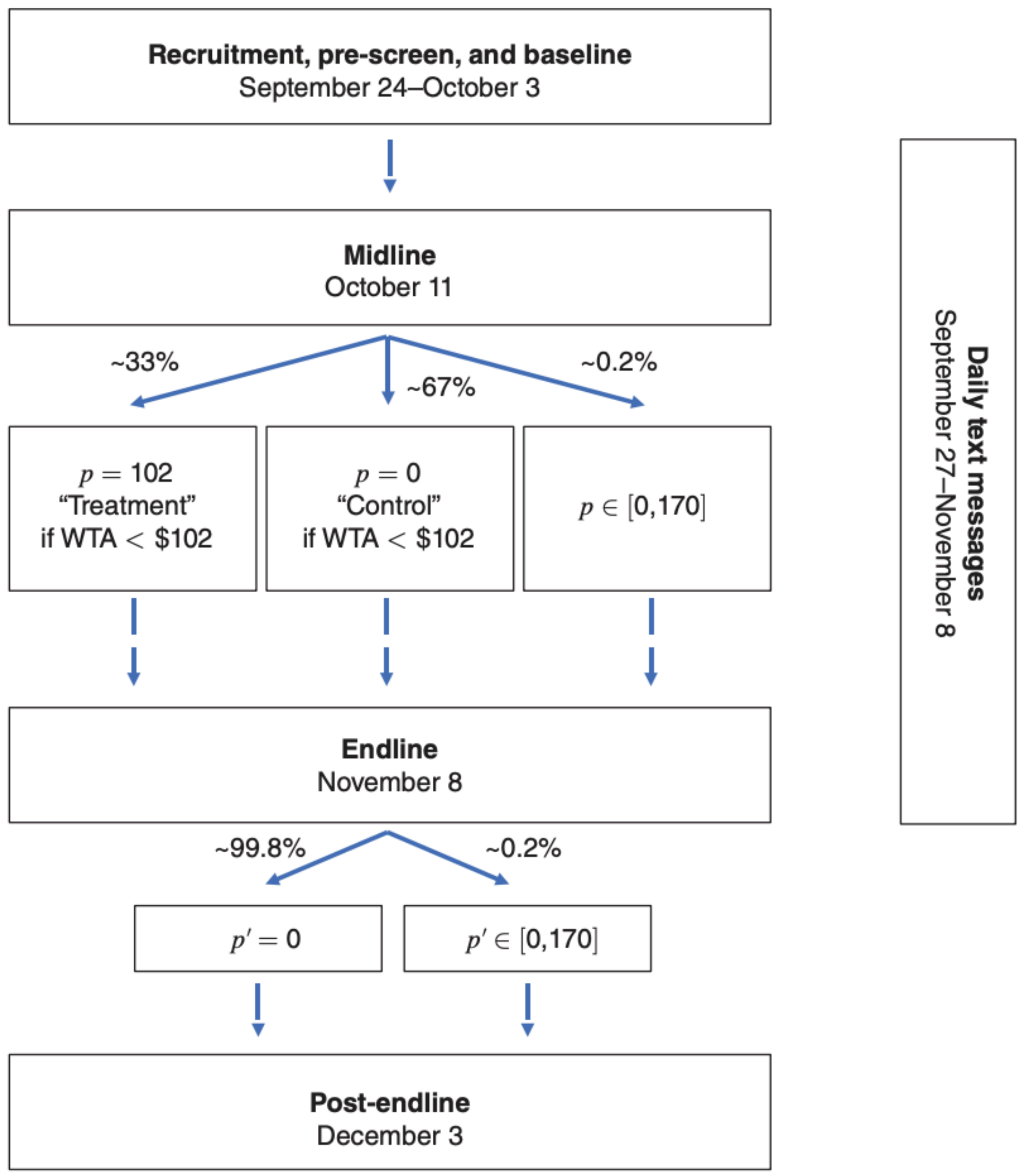


Figure 1. Experimental Design

Table 1—Sample Sizes

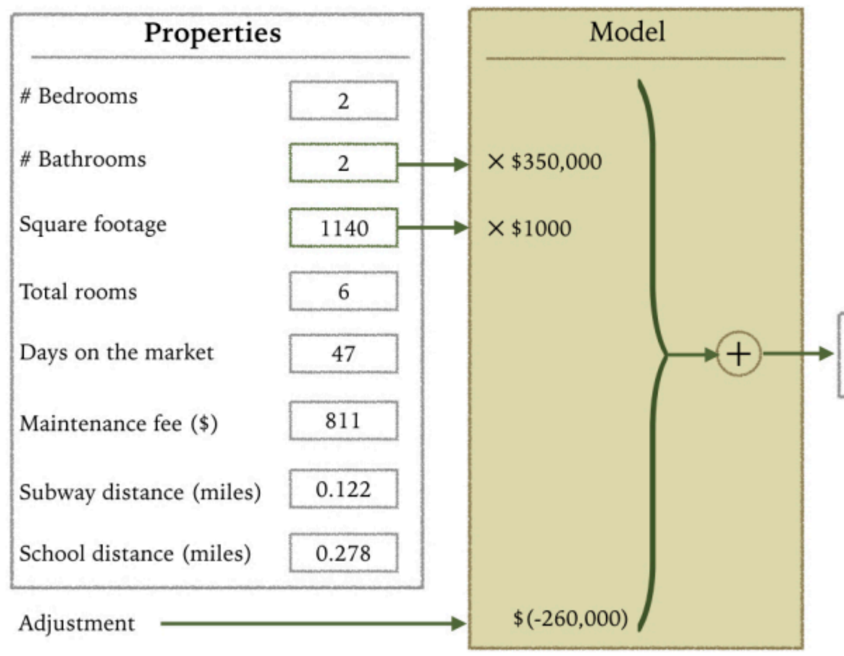
Phase	Sample size
Recruitment and baseline	N = 1,892,191 were shown ads N = 32,201 clicked on ads N = 22,324 completed pre-screen survey N = 20,959 were from United States and born between 1900 and 2000 N = 17,335 had 15 $<$ daily Facebook minutes $\leq$ 600 N = 7,455 consented to participate N = 3,910 finished baseline N = 2,897 had valid baseline and were randomized, of which:
Midline	N = 2,897 began midline N = 2,743 received a price offer, of which: N = 1,661 were in impact evaluation sample
Endline	N = 2,710 began endline N = 2,684 finished endline, of which: N = 1,637 were in impact evaluation sample
Post-endline	N = 2,067 reported Facebook mobile app use, of which: N = 1,219 were in impact evaluation sample

Experiments 2

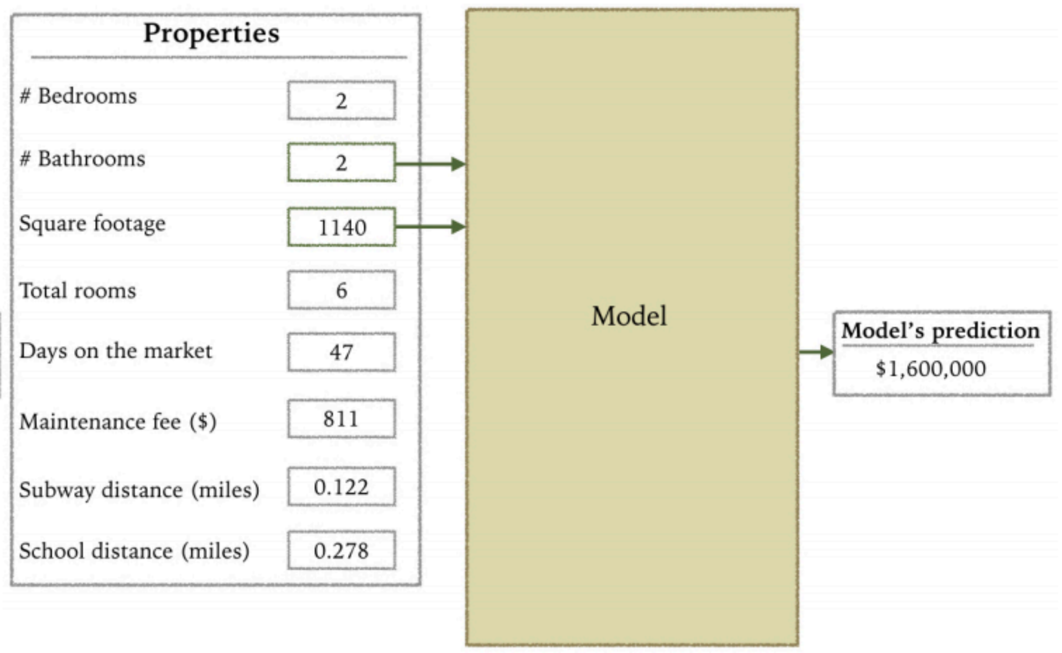
Manipulating and Measuring Model Interpretability

FOROUGH POURSABZI-SANGDEH, Microsoft Research
DANIEL G. GOLDSTEIN, Microsoft Research
JAKE M. HOFMAN, Microsoft Research
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research

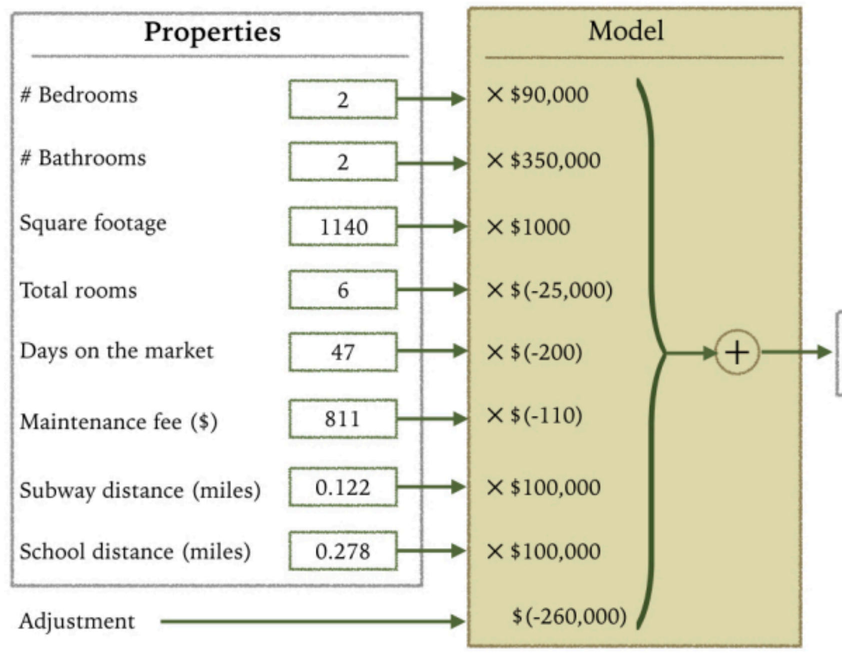
What are the effects of model interpretability on the end users?



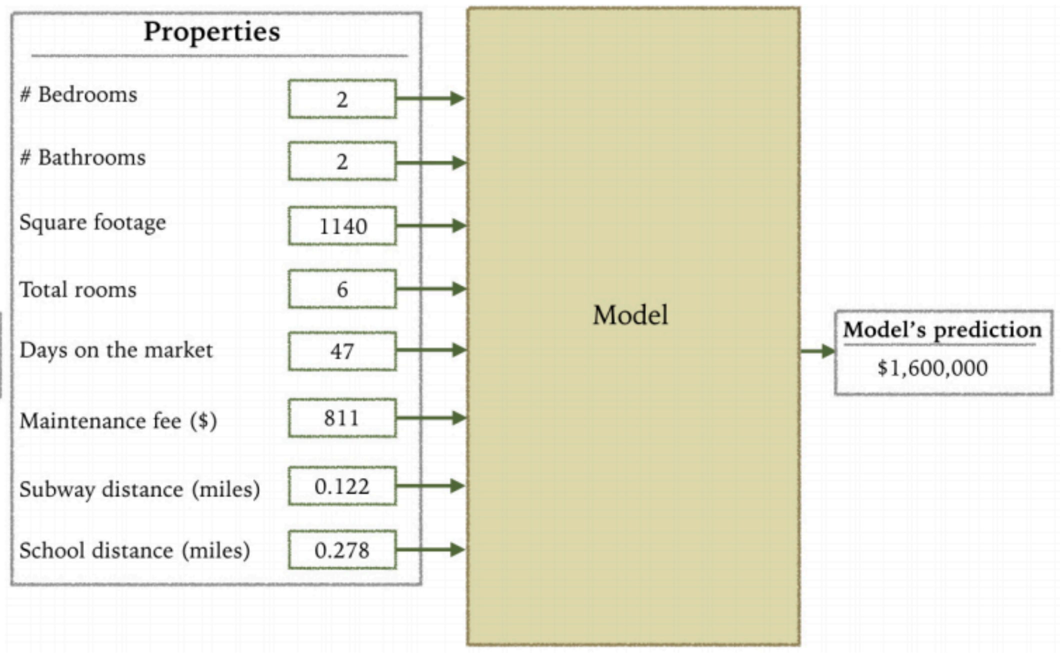
(a) Clear, two-feature condition (CLEAR-2).



(b) Black-box, two-feature condition (BB-2).



(c) Clear, eight-feature condition (CLEAR-8).



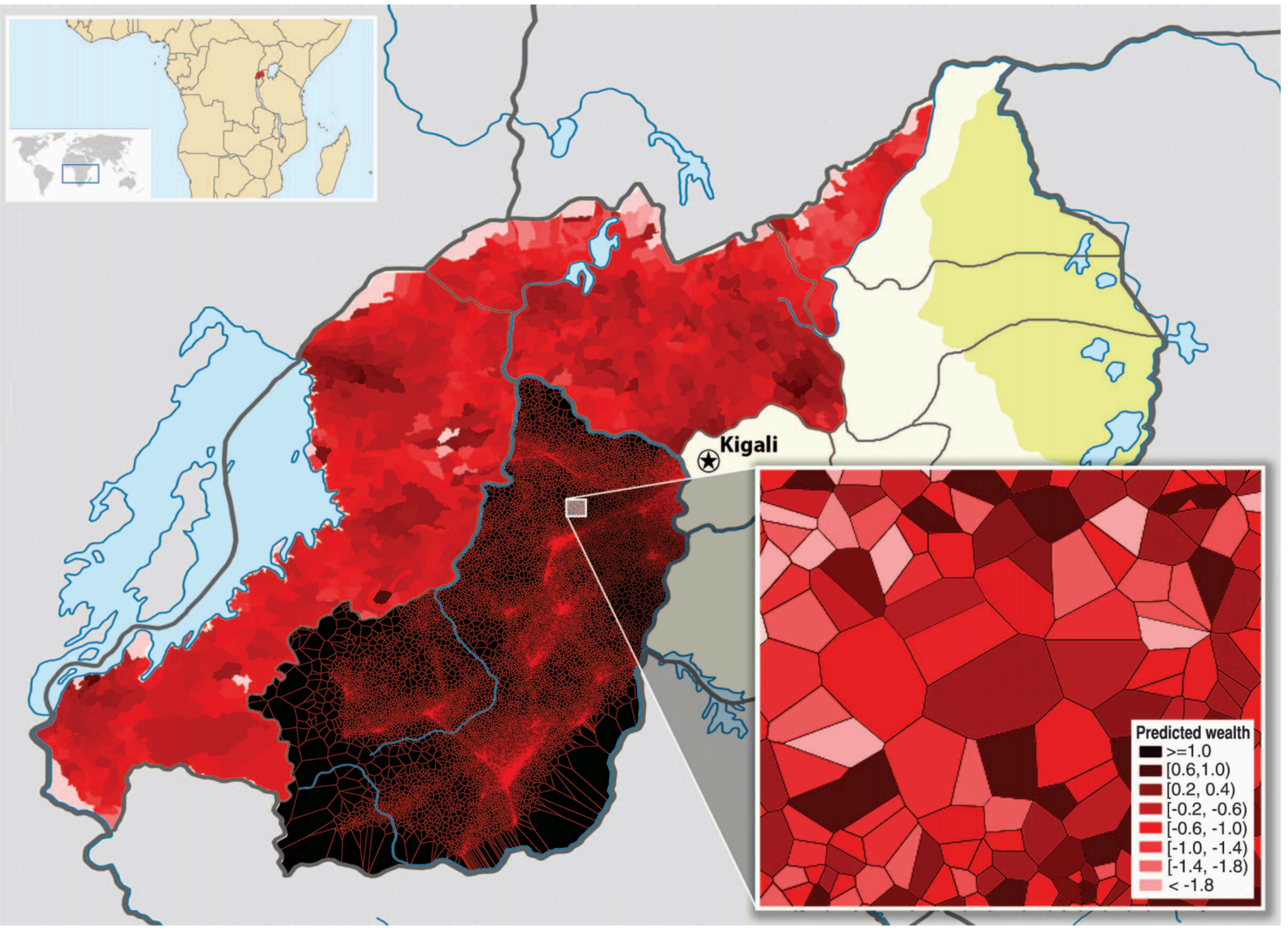
(d) Black-box, eight-feature condition (BB-8).

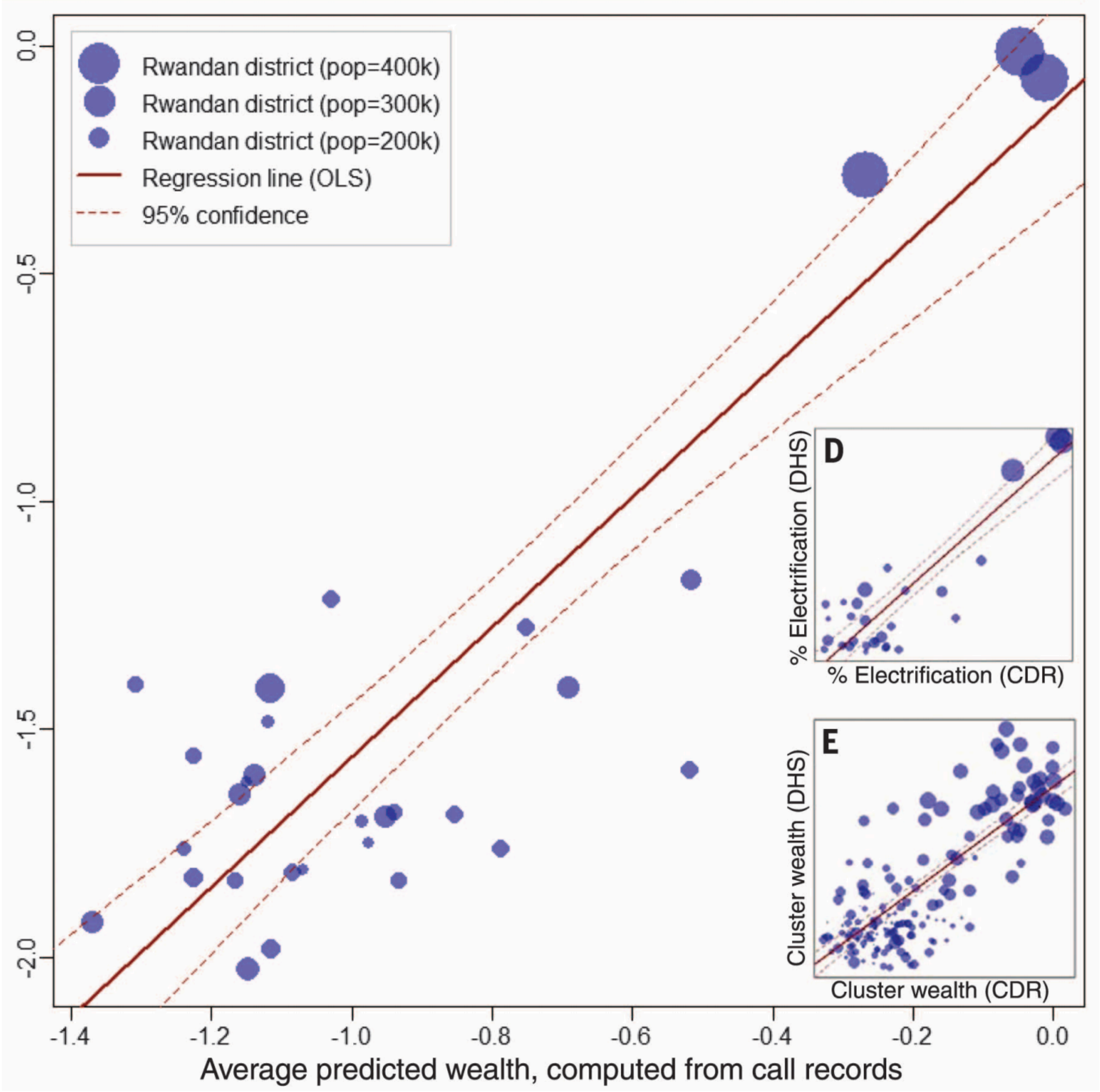
Asking questions

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

Can we amplify surveys with big data to accurately measure important macroscopic quantities?



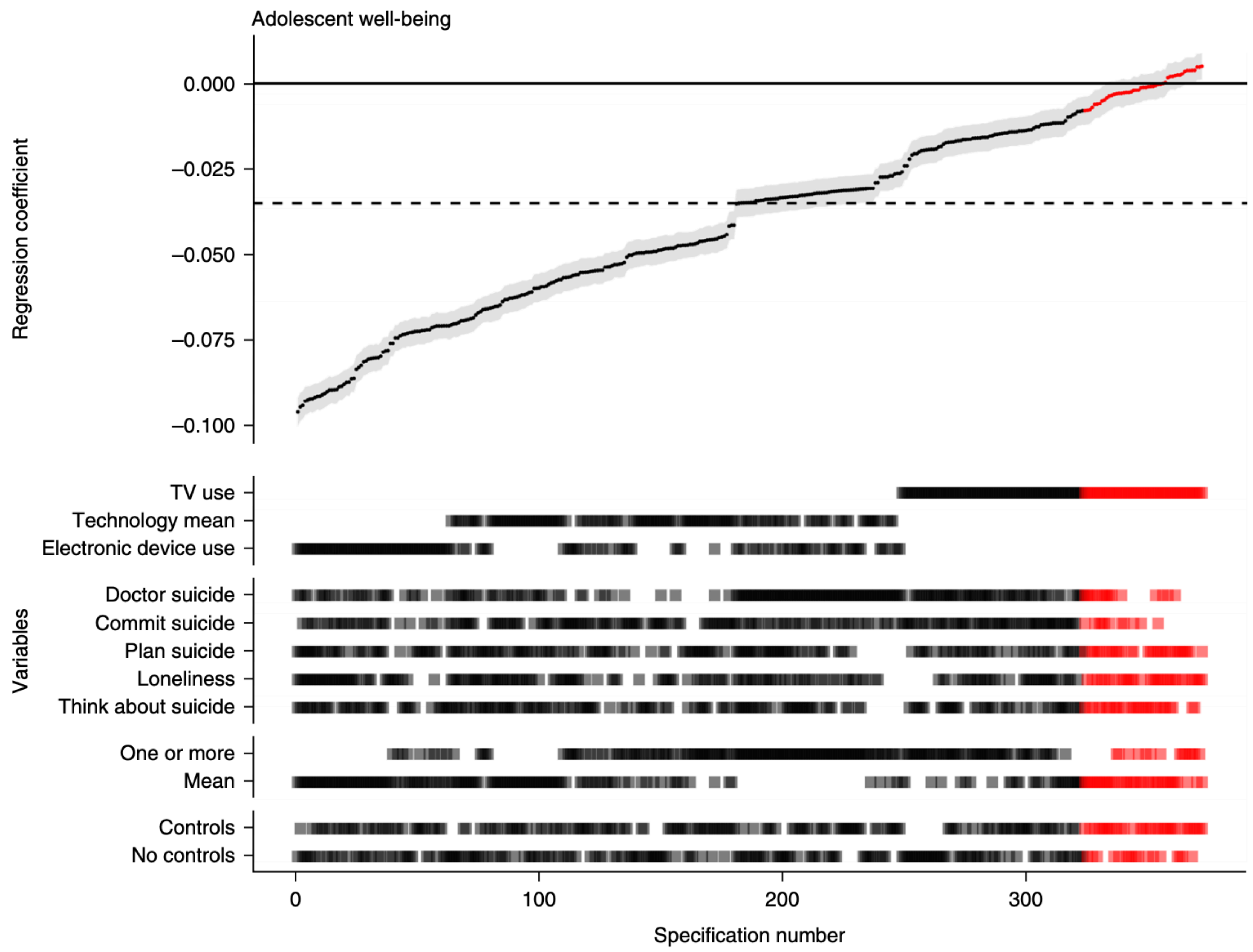


Asking questions

The association between adolescent well-being and digital technology use

Amy Orben[1*] and Andrew K. Przybylski[1,2]

What is the association between adolescent well-being and digital technology use, and how do we properly measure it?



Mass Collaboration

Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data

KENNETH BENOIT *London School of Economics and Trinity College*
DREW CONWAY *New York University*
BENJAMIN E. LAUDERDALE *London School of Economics and Political Science*
MICHAEL LAVER *New York University*
SLAVA MIKHAYLOV *University College London*

What are political entities saying in their manifestos?

FIGURE 1. Hierarchical Coding Scheme for Two Policy Domains with Ordinal Positioning

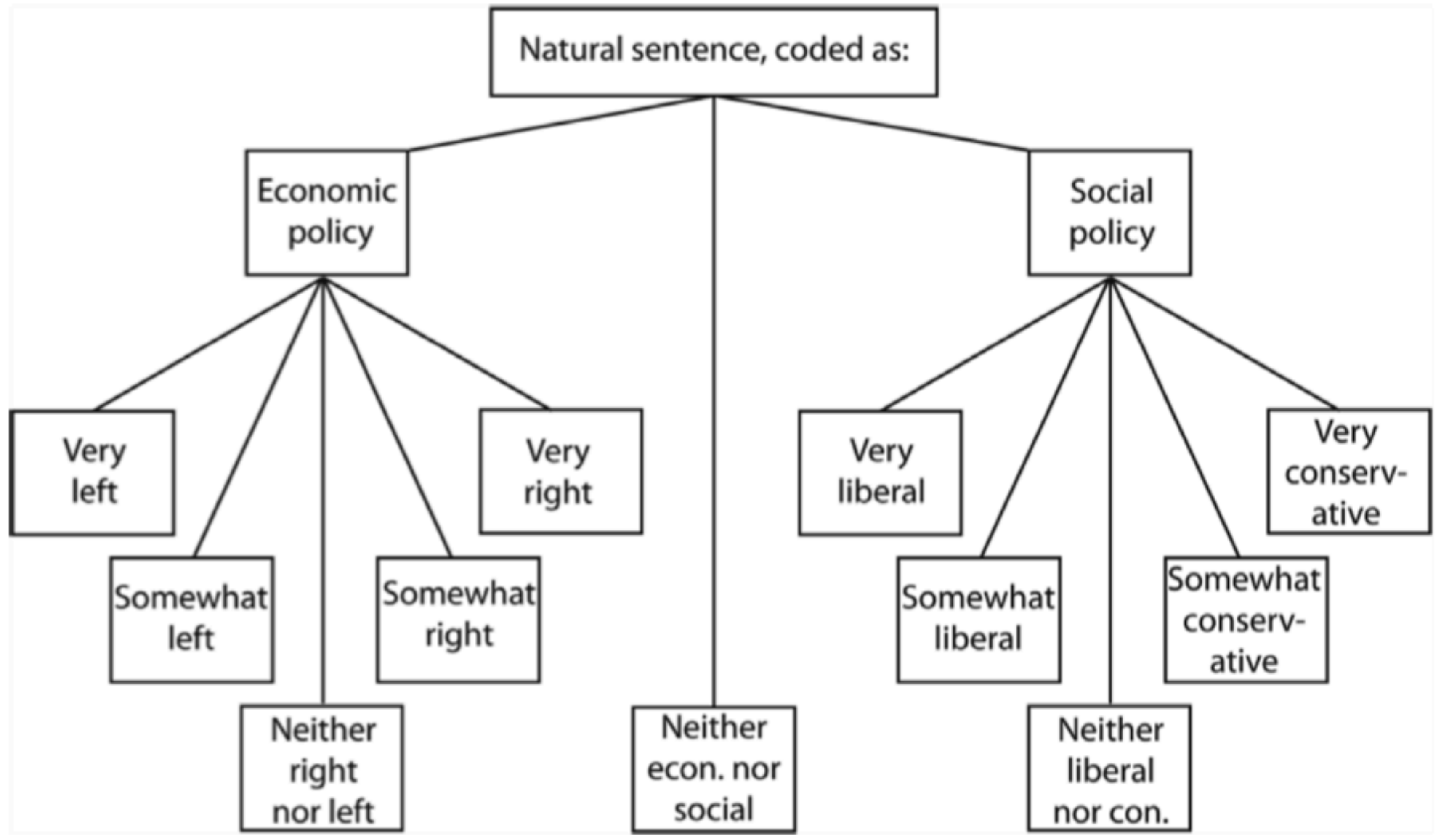
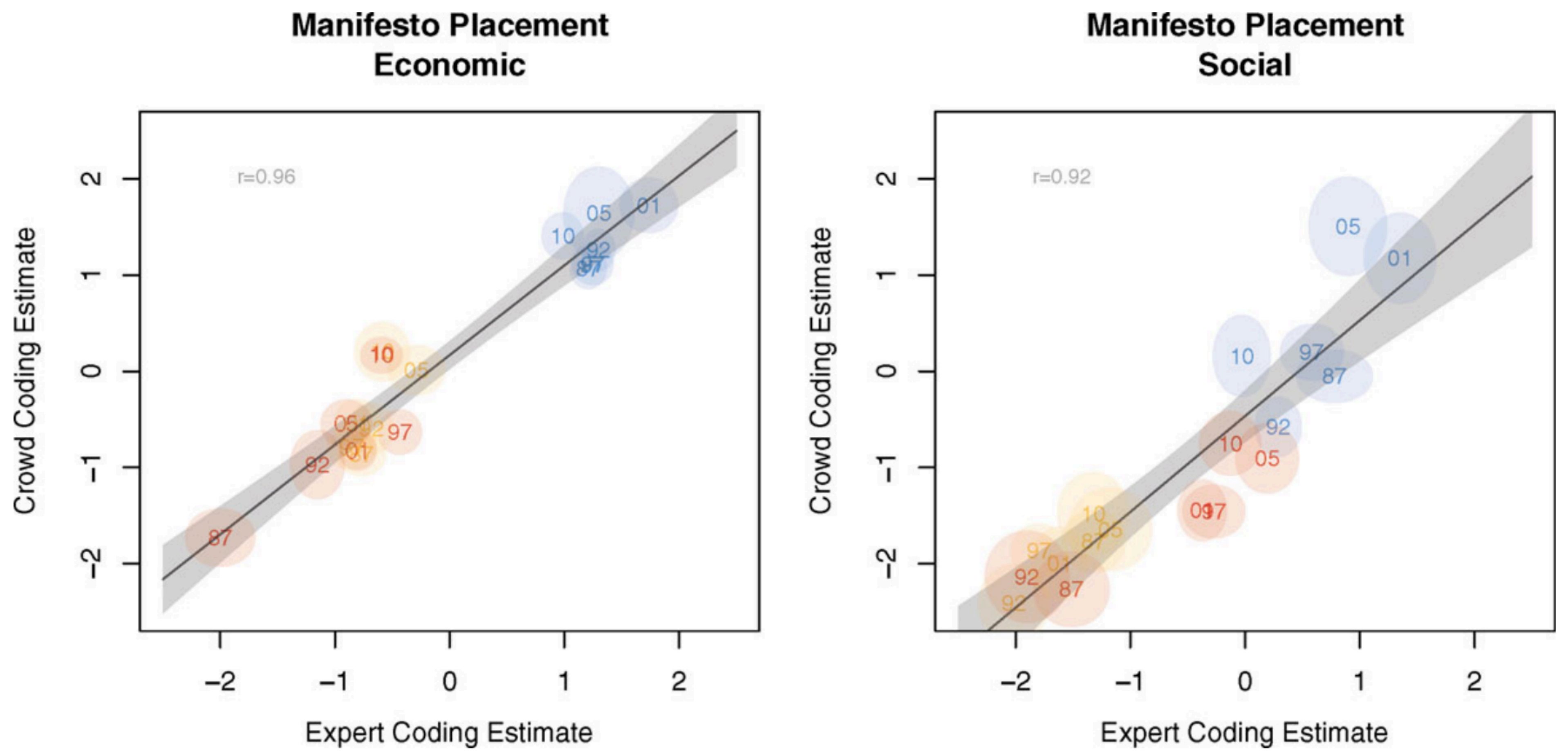


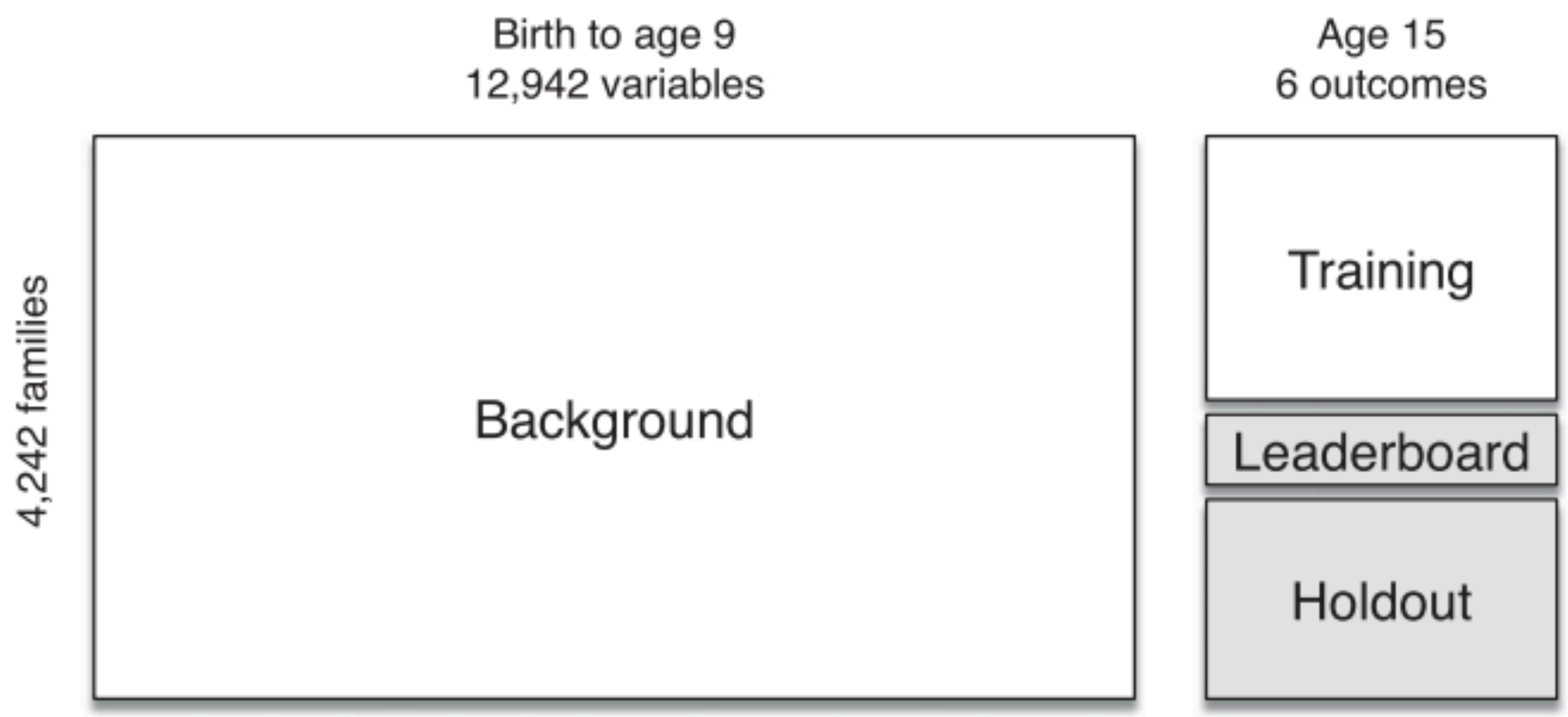
FIGURE 3. Expert and Crowd-sourced Estimates of Economic and Social Policy Positions

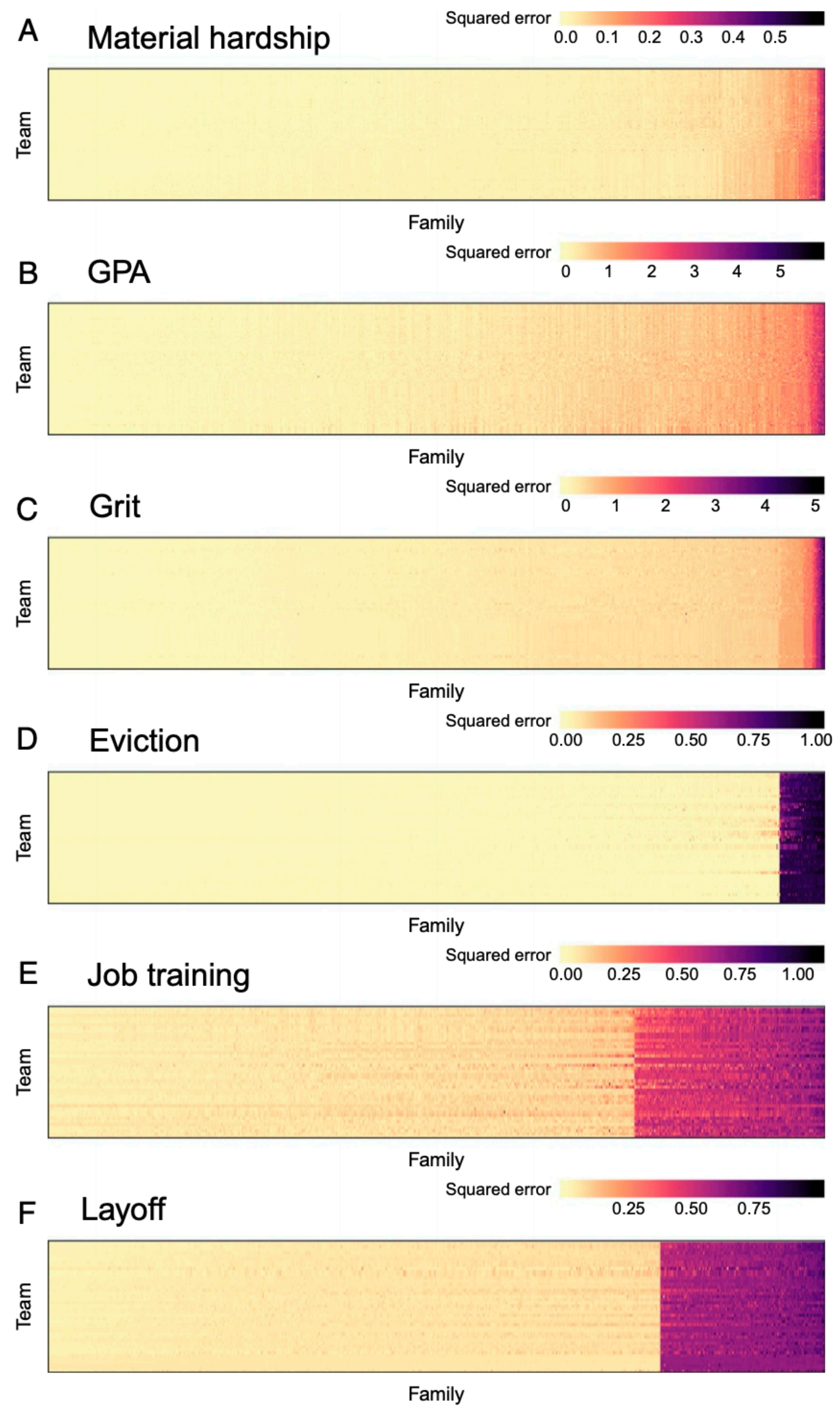


Mass Collaboration

Measuring the predictability of life outcomes with a scientific mass collaboration

How predictable are life outcomes?



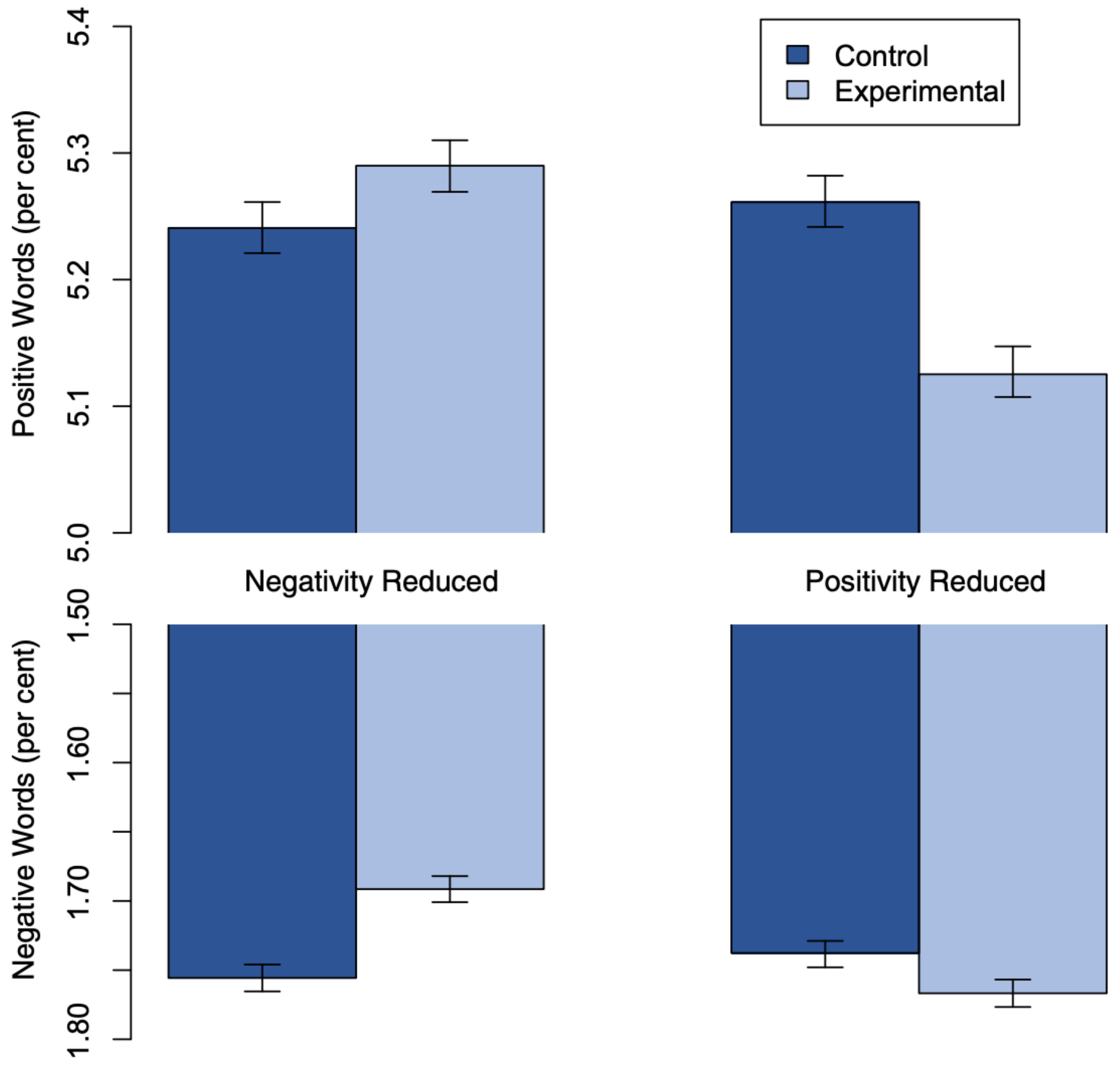


Ethics in computational social science

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

Are emotional states transferred via social networks?



Ethics in computational social science

danah boyd & Kate Crawford

CRITICAL QUESTIONS FOR BIG DATA

Provocations for a cultural, technological, and scholarly phenomenon

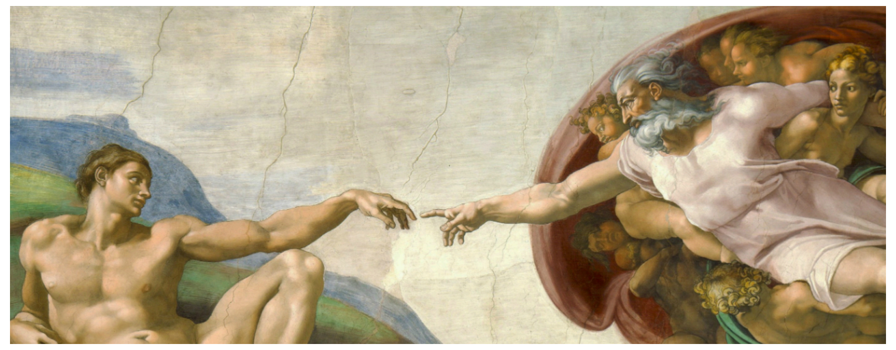
1. **Big Data changes the definition of knowledge**
2. **Claims to objectivity and accuracy are misleading**
3. **Bigger data are not always better data**
4. **Taken out of context, Big Data loses its meaning**
5. **Just because it is accessible does not make it ethical**
6. **Limited access to Big Data creates new digital divides**

Computational social science in 7 easy pieces

Week	Date	Topic	Reviews Due	Textbook Readings
1	1/9	Introduction to computational social science (slides)		Ch. 1
2	1/16	Introduction to computational social science cont'd (slides)		Ch. 1
3	1/23	Observational studies 1	1/22 9:00pm	Ch. 2
4	1/30	Observational studies 2	1/29 9:00pm	Ch. 2
5	2/6	Experiments 1	2/5 9:00pm	Ch. 4
6	2/13	Experiments 2	2/12 9:00pm	Ch. 4
7	2/27	Project proposals		
8	3/5	Asking questions	3/4 9:00pm	Ch. 3
9	3/12	Mass collaboration	3/11 9:00pm	Ch. 5
10	3/19	Ethics in computational social science	3/18 9:00pm	Ch. 6
11	3/26	Project presentations (Part 1)		
12	4/2	Project presentations (Part 2)		

Readymades

Custommades



Logistics

Course grades:

- 35% Project (proposal, presentation, report)
- 25% Reviews (relevance, quality, shows thought)
- 15% Paper Discussion Leading (clarity, organization, discussion provoking)
- 15% Assignments
- 10% Participation (quality not quantity)

Logistics

- Course webpage: http://www.cs.toronto.edu/~ashton/csc2552/

- Due Wednesday at 9pm: Reviews of the two papers we will discuss
- Reviews will be submitted on MarkUs in PDF format
- In-class discussions: 2-3 people will present each paper
  - Who wants to go next week? (fake news! fun!)
  - Present for ~10 minutes, focus on discussion and critical review and questions rather than the material since everyone will have read the paper, discuss for ~40 minutes
  - Come prepared with discussion questions and opinions
- Todo: log in to MarkUs (link will be on course webpage)
- First reviews due next week