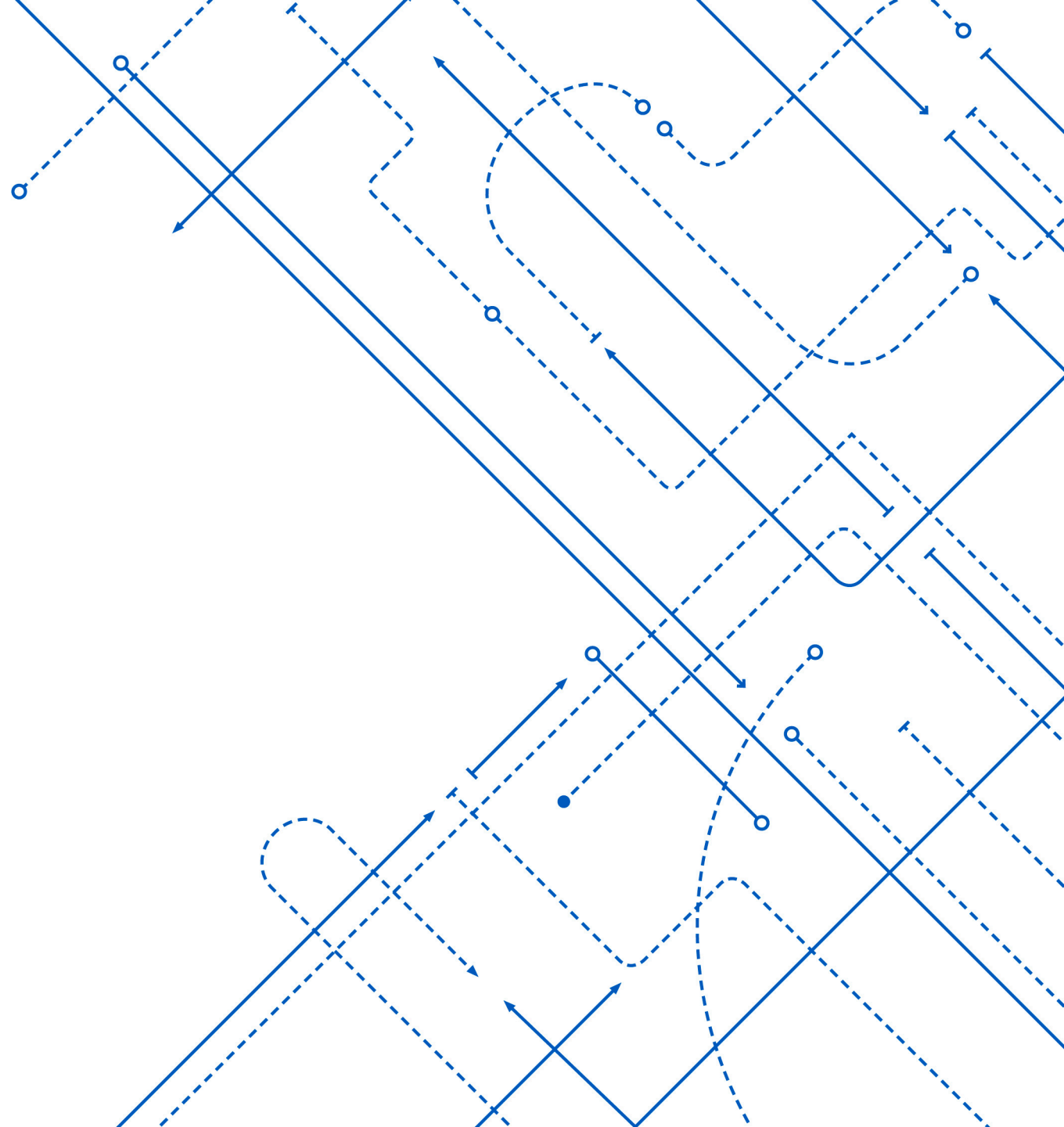
Measuring White Supremacy

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences



Attendance: Abigail Z. Jacobs

azjacobs at umich.edu
CV | Google Scholar

ABIGAIL Z. JACOBS

Assistant Professor of Information, School of Information
Assistant Professor of Complex Systems, College of Literature, Science, and the Arts
University of Michigan

Measurement and Fairness

Abigail Z. Jacobs
azjacobs@umich.edu
University of Michigan

Hanna Wallach
hanna@dirichlet.net
Microsoft Research

Projects

- Teams seem to have lost points in my batch on
  - Superficial, or no, inclusion of prior work
  - Superficial inclusion of history teammate
- **You can address the prior work part on Wednesday**
- Talks on Wednesday will be timed, strictly
  - You will be evaluated on what you present
  - Practice

Today

- Why measure?
- What exactly are we trying to measure when we are "measuring white supremacy"?
- How do we get the data we need?
- What analytical tools do we need?

Why do we measure?

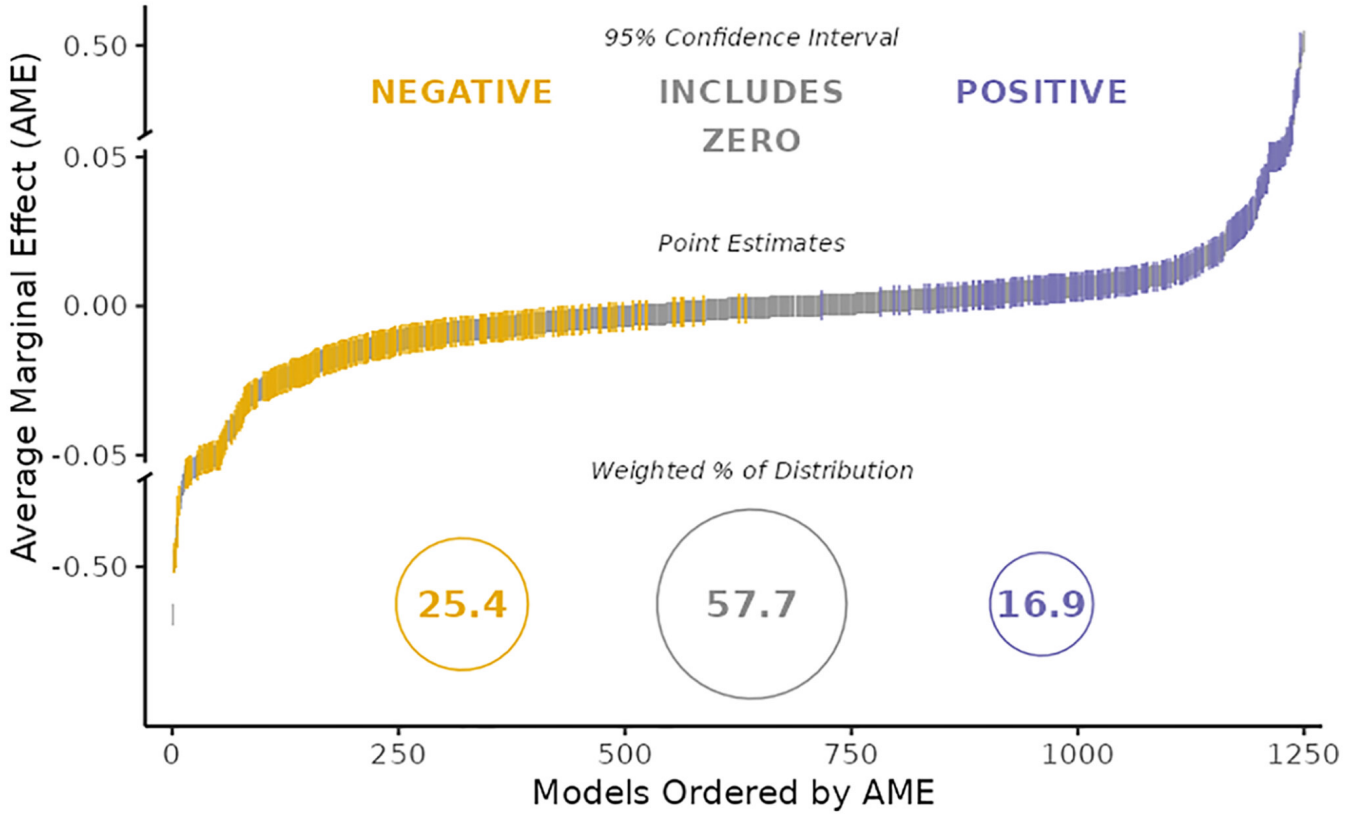
- What is the need for measurement? Think a minute, give me a concrete example of where *any kind* of measurement is useful

Why do we measure?

- We measure to **know** or to **show**
- Know
  - So we can change/improve/understand
  - How big is the problem?
  - Did what I tried have any impact?
- Show
  - To convince other people of things
  - See?! Look how big this problem is!
  - See?! This problem impacts all of us!

But measurement is *hard*

Researchers' expertise, prior beliefs, and expectations barely predict the wide variation in research outcomes. More than 95% of the total variance in numerical results remains unexplained even after qualitative coding of all identifiable decisions in each team's workflow



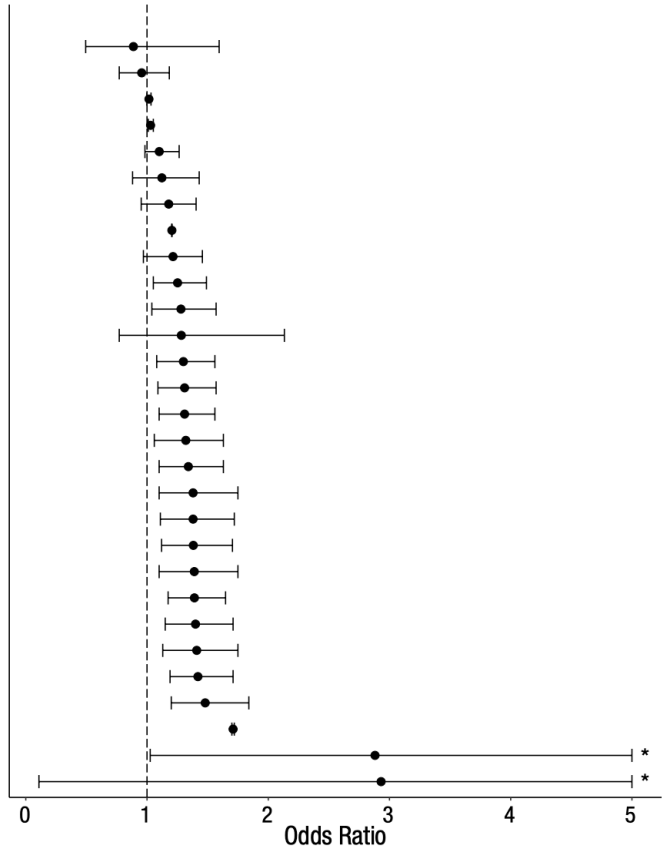
https://www.pnas.org/doi/full/10.1073/pnas.2203150119

But measurement is *hard*

Abstract

Twenty-nine teams involving 61 analysts used the same data set to address the same research question: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from 0.89 to 2.93 (*Mdn* = 1.31) in odds-ratio units. Twenty teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship. Overall, the 29 different analyses used 21 unique combinations of covariates. Neither analysts' prior beliefs about the effect of interest nor their level of expertise readily explained the variation in the outcomes of the analyses. Peer ratings of the quality of the analyses also did not account for the variability. These findings suggest that significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions. Crowdsourcing data analysis, a strategy in which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how defensible, yet subjective, analytic choices influence research results.

https://journals.sagepub.com/doi/pdf/10.1177/2515245917747646



And there are other challenges to measurement

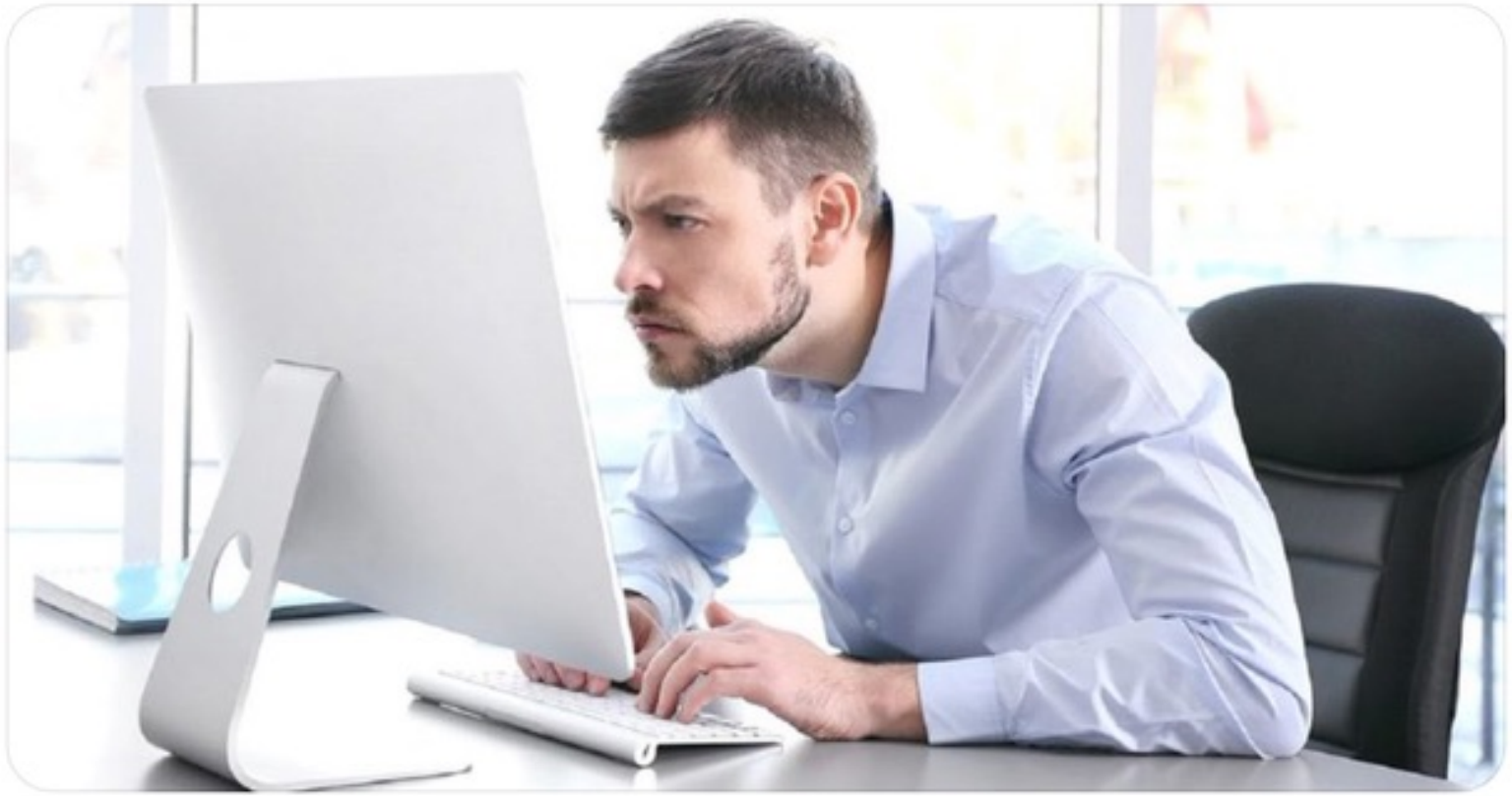
Bottom line

- Measurement is necessary
- (Good) Measurement is hard
  - It takes a convincing *base* of evidence to provide a set of measurements that, together, convince us of a truth
  - Even with those sets of measurements, there's people that don't want to believe

amy

@ahsurelook

lecturers choosing which meme from 2006 should go into their slides



nny_joseph

Today

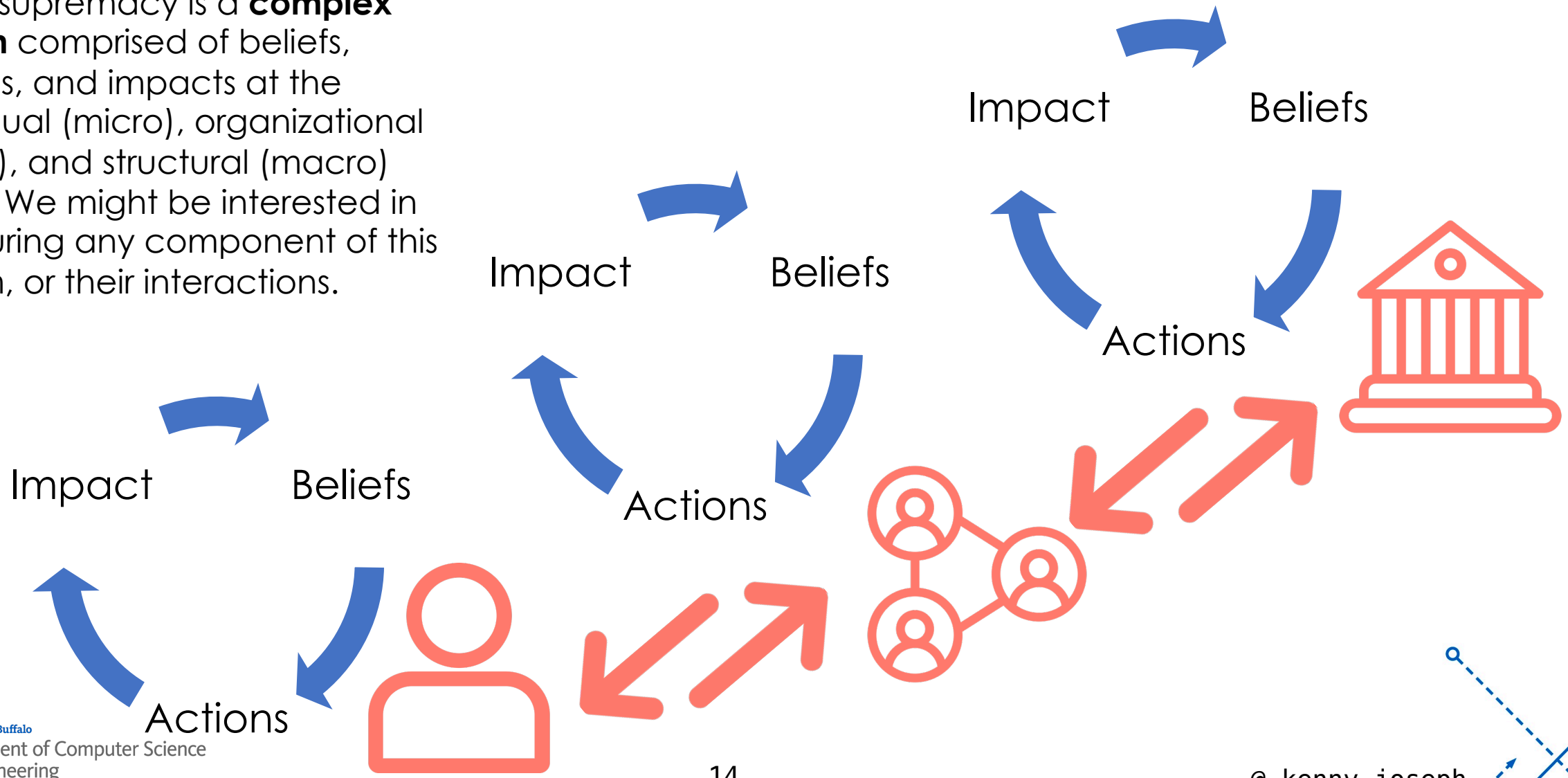
- Why measure?
- What exactly are we trying to measure when we are "measuring white supremacy"?
- How do we get the data we need?
- What analytical tools do we need?

Well, what do you think?

What exactly are we trying to measure when we are "measuring white supremacy"?

What we are trying to measure

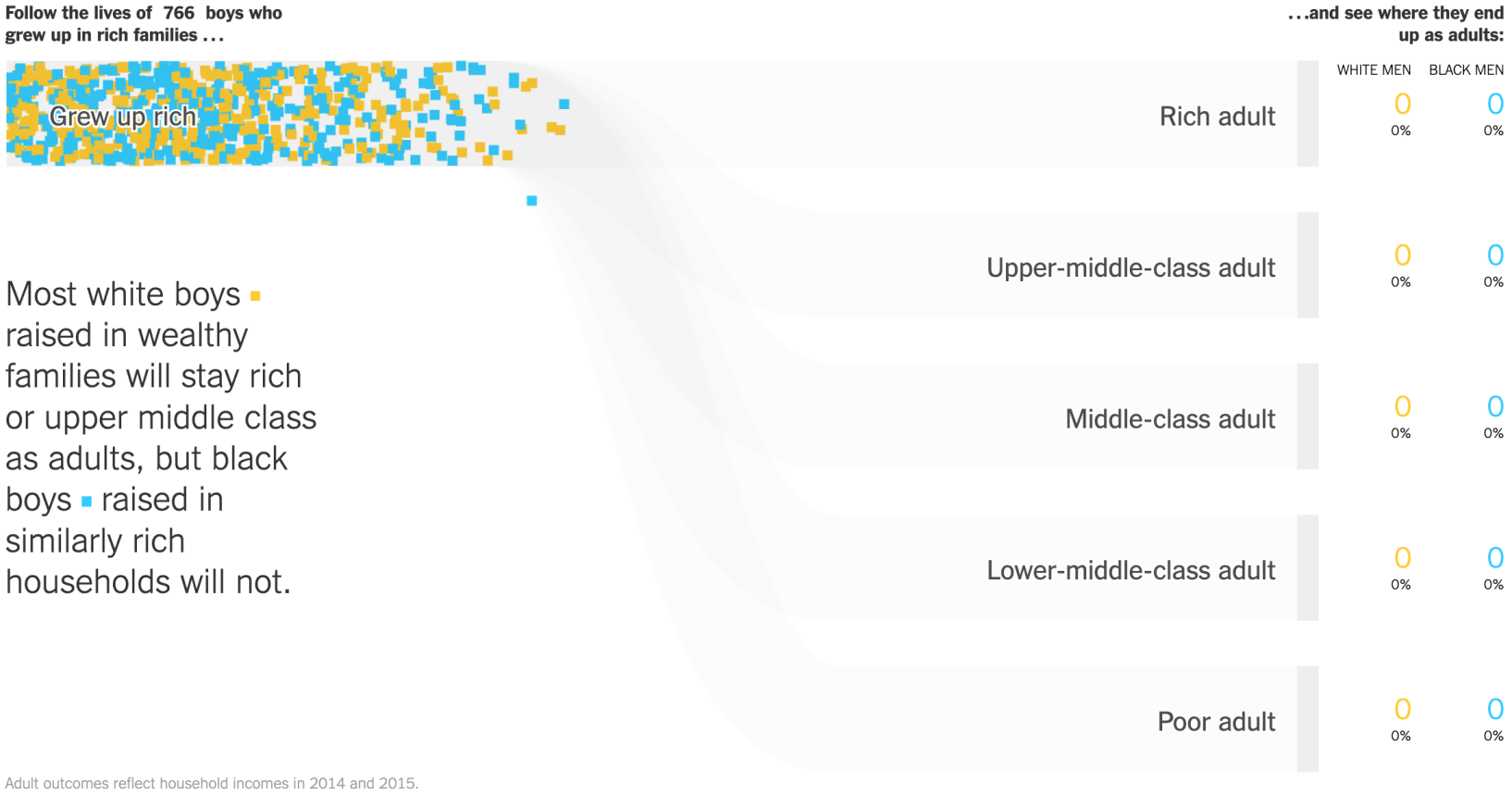
White supremacy is a **complex system** comprised of beliefs, actions, and impacts at the individual (micro), organizational (meso), and structural (macro) levels. We might be interested in measuring any component of this system, or their interactions.



Some examples – Individual Beliefs

- In truth, we must get considerably more specific than "beliefs"
  - A *stereotype* is a cognitive association between two ways of categorizing people. Example?
  - An *attitude* is a valence judgement of a particular *issue*. Example?
  - A *prejudice* is a valenced judgement of a particular "type of person." Example?
  - A *value* is "a moral or ethical commitment to something as being right or wrong, good or bad, moral or immoral, important or unimportant"
  - A *belief* is a generalized assumption about the way the world works. Example: Essentialism vs. Constructivism
- All slightly different! Why? (Precision and Independent creation of ideas)

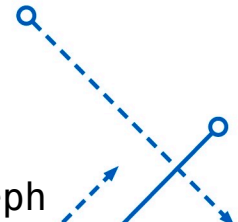
Some examples – Macro-level effects



https://www.nytimes.com/interactive/2018/03/19/upshot/race-class-white-and-black-men.html

A side note – "measuring" with simulation

- https://ncase.me/polygons/



Bottom line

- There are many, many things we can measure
  - "Beliefs" – What a person/org/policy "thinks"
  - Actions – What they do
  - Effects – What the outcomes of a collective of beliefs and actions are
- We can measure these at various *levels*: micro, meso, macro
- We have to be clear about:
  - What, precisely, we are (and are not) measuring **(good operationalization)**
  - Who else has measured this

Today

- Why measure?
- What exactly are we trying to measure when we are "measuring white supremacy"?
- **How do we get the data we need?**
- What analytical tools do we need?

Some things to keep in mind for coming exercise

- Possible data sources
  - Administrative data
  - Experiments
  - Interviews
  - Surveys
  - Social media
    - APIs
    - Data Brokers
  - Web crawling
  - … others?
- There is a field of study for each of these… i.e. we know that there are good and bad ways to do each of them

Exercise

- Come up with a thing you think is worth measuring
  - State whether it is a belief, action, or impact
  - State what level you are measuring it
- **(Fake) Bonus points if this relates to your project!**
- Now, think about *where you could get the data to measure this*. Answer the following:
  - Who has that data right now?
  - If not you, how would you get it? Would you have to pay? How much? Or annotate data? How long would that take?
  - Is it *ethical* to collect this data? Could you do it ethically? Would that impact the quality of your measurement?

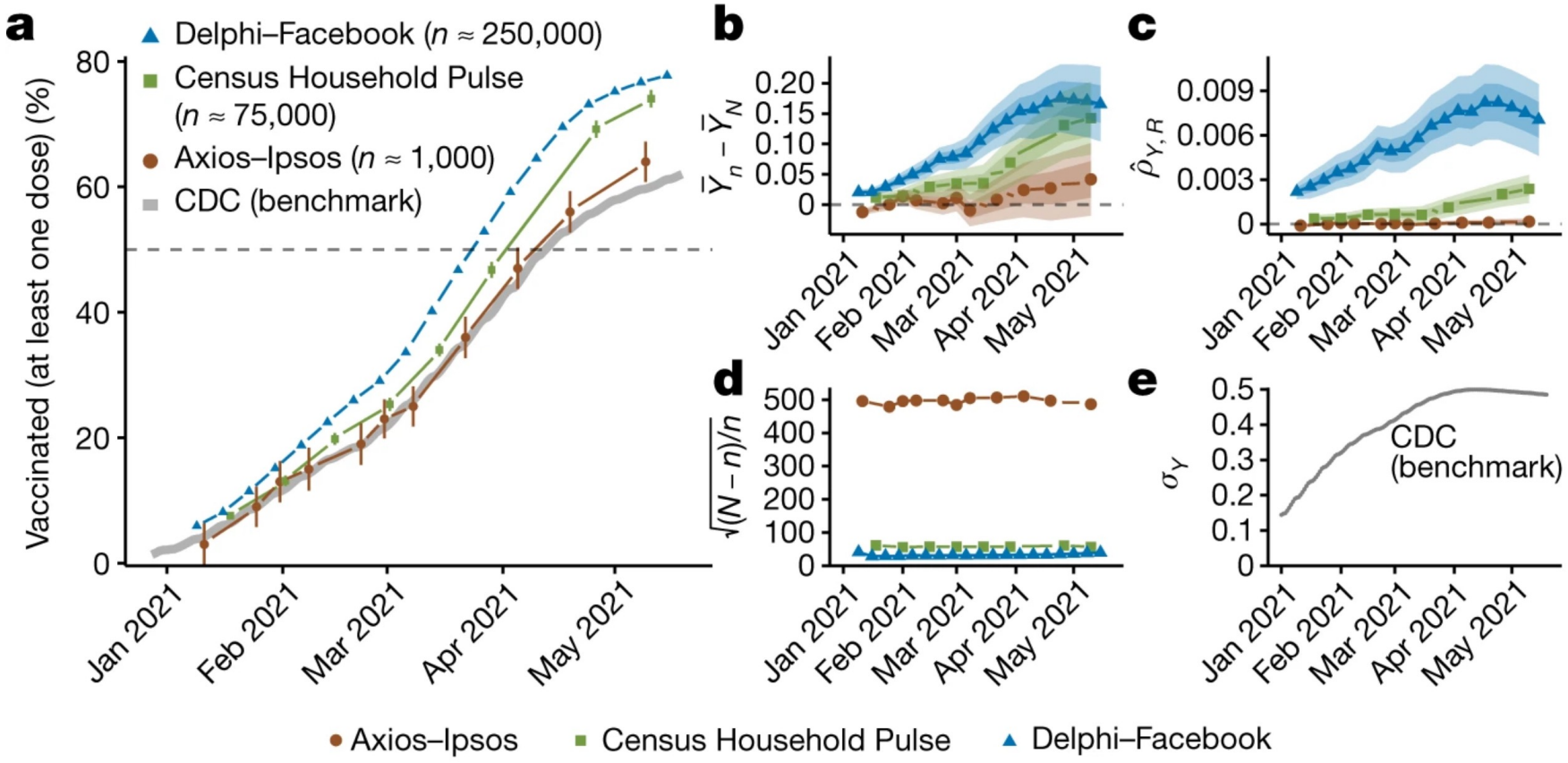
Data as a Discussion Between You and the Participants

- Desirability Bias
- Representativeness
- Which question are you asking?
- What biases exist in social media data?
- What to do with individual responses?

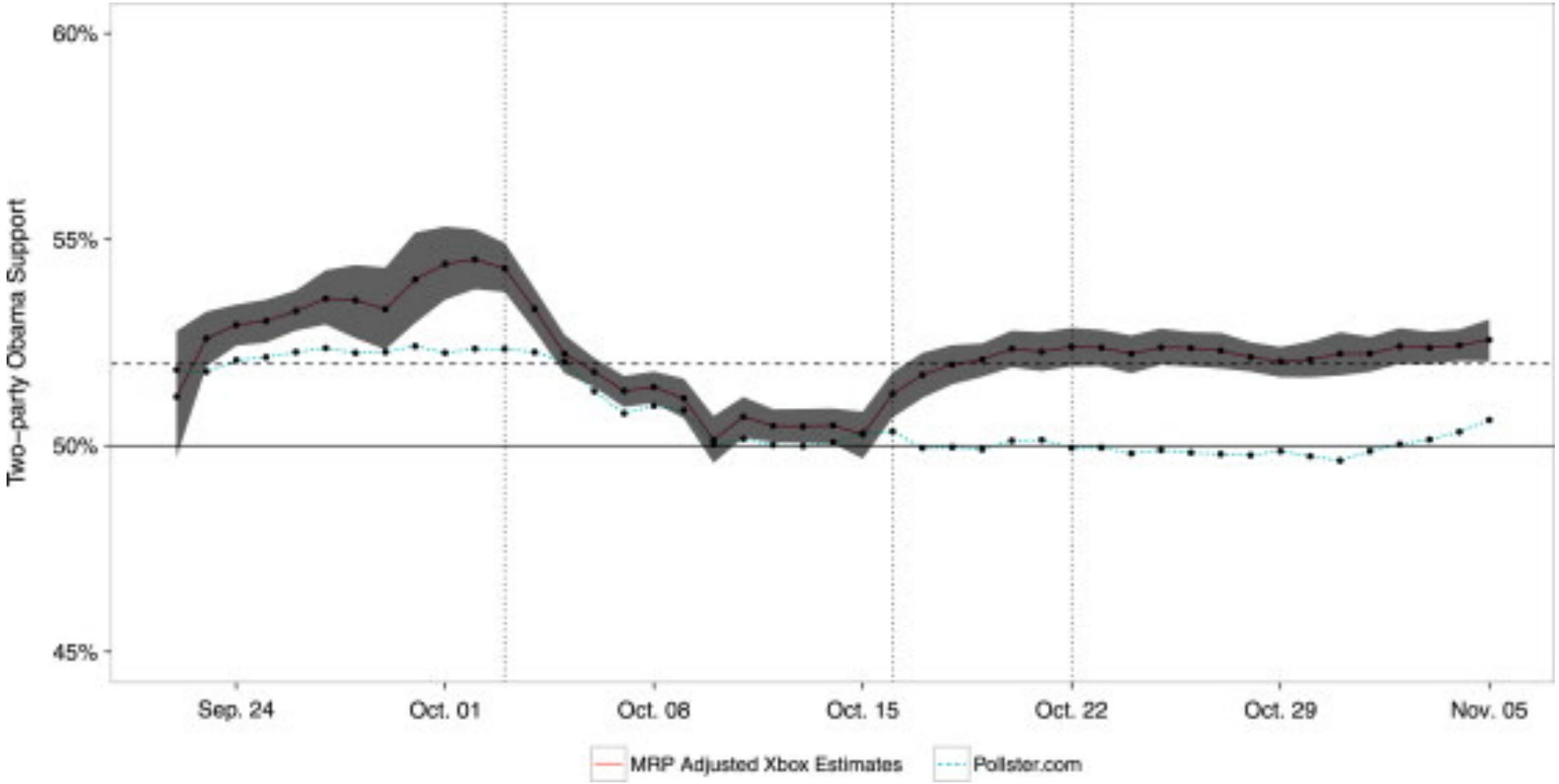
But bigger is better, right? Right?!

Fig 1: Errors in estimates of vaccine uptake.

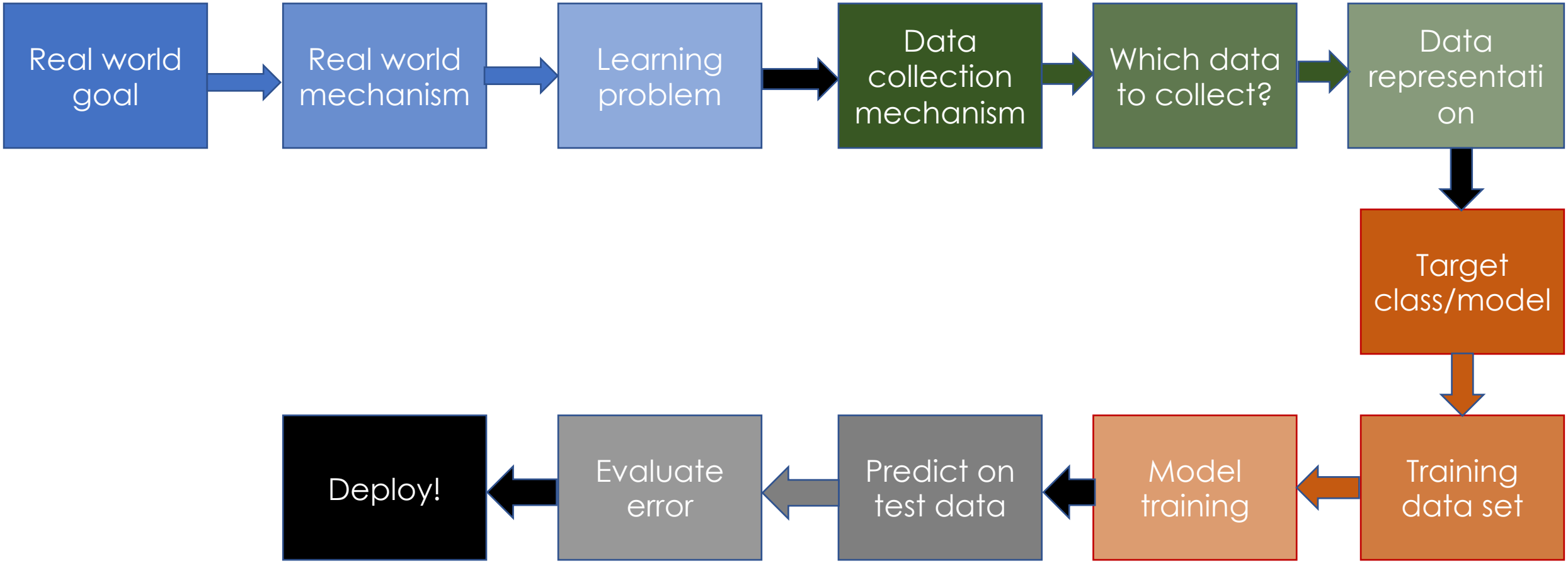
From: Unrepresentative big surveys significantly overestimated US vaccine uptake



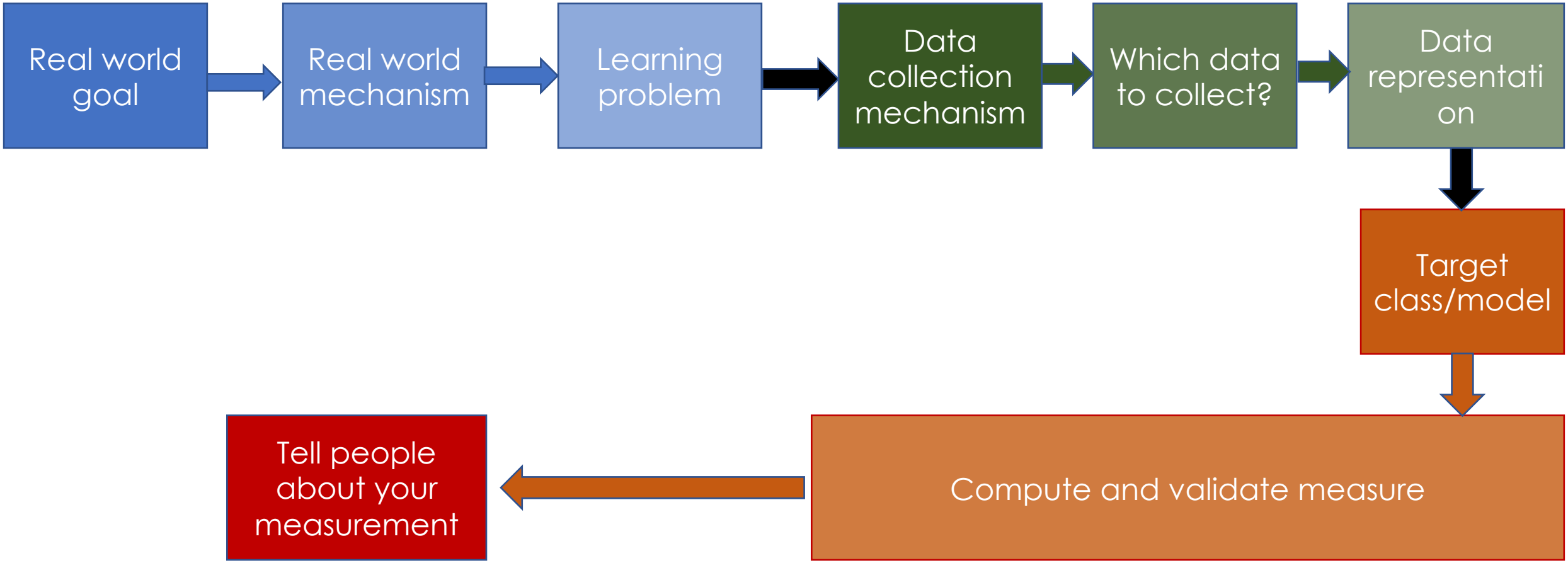
All hope is not lost



ML pipeline


Real world goal
Real world mechanism
Learning problem
Data collection mechanism
Which data to collect?
Data representation
Target class/model
Training data set
Model training
Predict on test data
Evaluate error
Deploy!

Big reveal: the ML pipeline is a measurement pipeline with different analytical tools



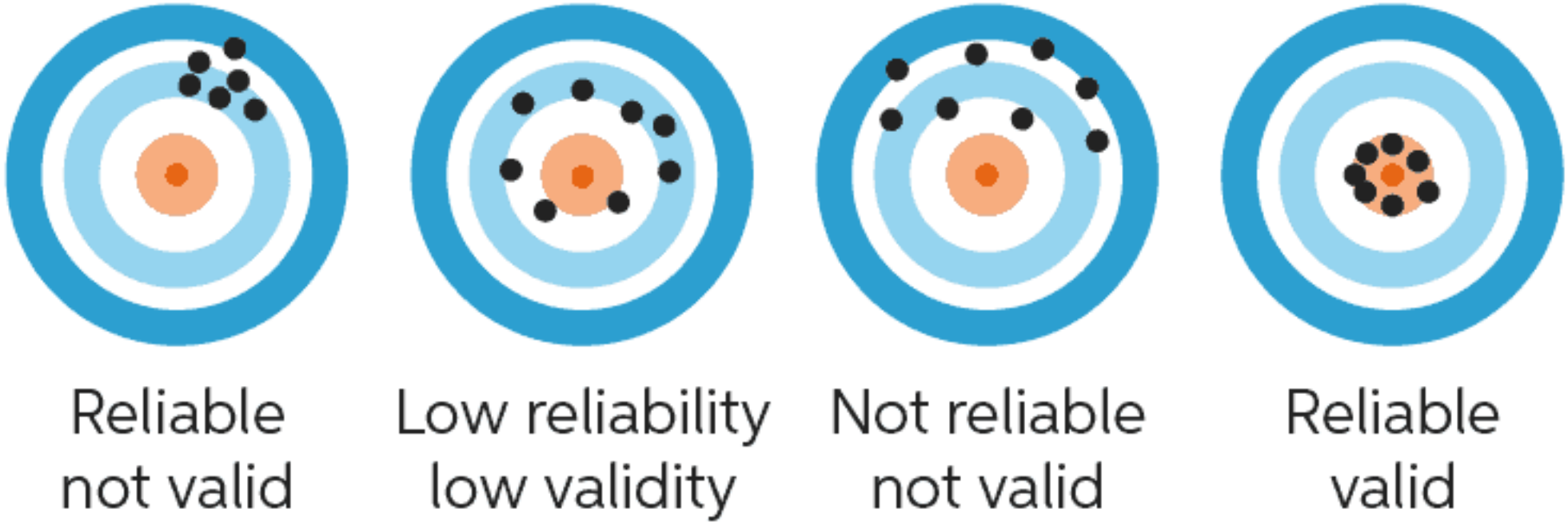
Today

- Why measure?
- What exactly are we trying to measure when we are "measuring white supremacy"?
- How do we get the data we need?
- What analytical tools do we need?

Tools for measurement

- (Aside: Machine learning can be (and often is!) a tool for measurement)
- Two kinds of tools:
  - Tools to *do the measuring*
  - Tools to *make sure the measure is* ***valid and reliable***

Validity and Reliability



https://www.chegg.com/writing/guides/research/reliability-vs-validity/

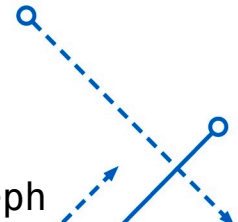
… lol chegg

(Some) Types of Validity

- **Face validity-** passes the "sniff test"
- **Content validity** - do you actually measure the entirety of what you think you're measuring?
- **Convergent validity** – aligns OK with other things that measure the same concept
- **Discriminant validity** – doesn't measure what it is not supposed to be measuring
- **Predictive validity** – predicts things we expect it would predict
- **Consequential validity** – what are the impacts of us measuring this thing?

Exercise

- I say I am going to measure criminality based on a model trained to differentiate **mugshots** from **profile pictures on LinkedIn**
- Pick two kinds of validity and assess the validity of this measure



Linking to Fairness

Measurement and Fairness

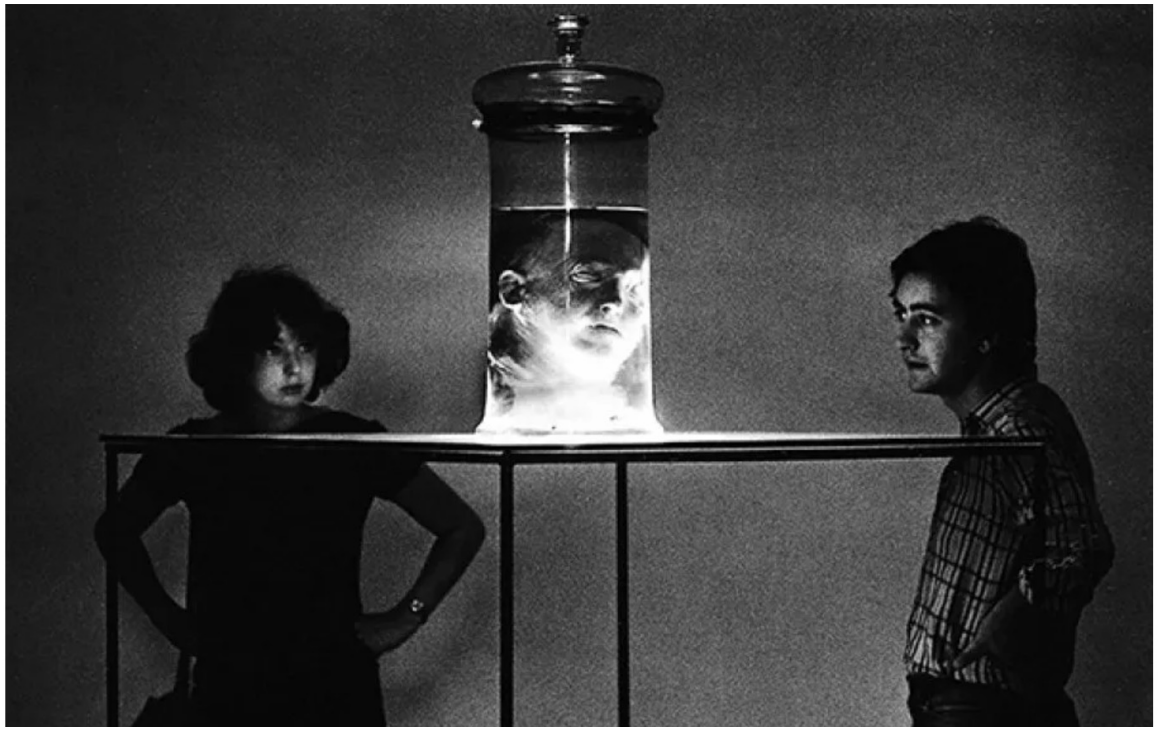
Abigail Z. Jacobs
azjacobs@umich.edu
University of Michigan

Hanna Wallach
hanna@dirichlet.net
Microsoft Research

Linking to "Bias in AI"

Physiognomy's New Clothes

by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov



https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Tools for measuring

- There are many, many ways to measure
- We'll go over three in the next few weeks:
  - Causal inference
    1. Experimental: (Algorithmic) Audits
    2. Observational: Statistical models for causal inference
  3. Measuring beliefs using text data (NLP)
- Examples of things we won't cover:
  - Social network analysis
  - Image analysis
  - Simulation / Theoretical modeling

If time

- A review of basic statistics