Science Before Statistics

CAUSAL INFERENCE

Richard McElreath MPI-EVA Leipzig

Horoscope of Prince Iskandar, grandson of Tamerlane, born 25 April 1384

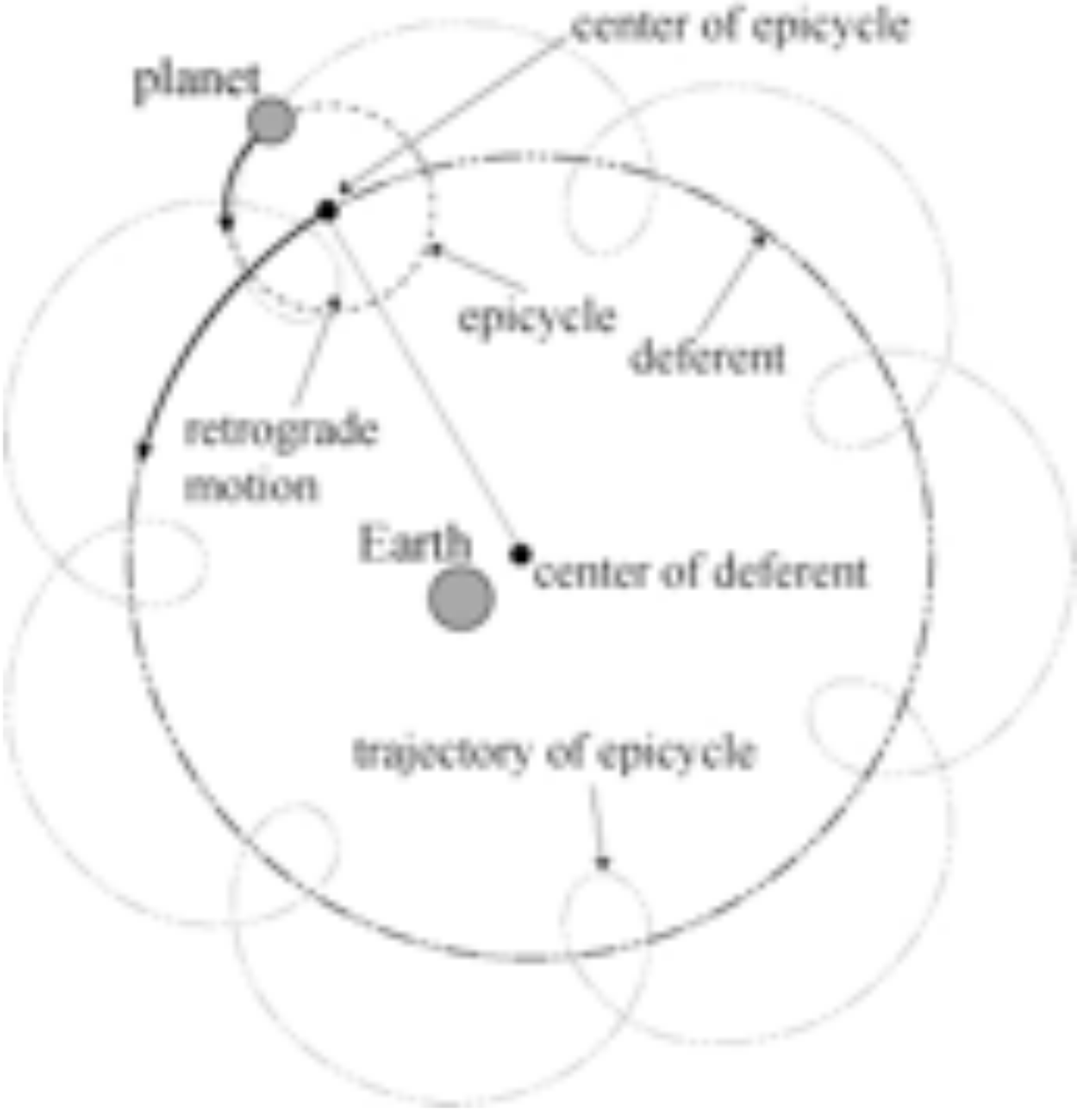
Horoscopes

- Tell me your birthday, and I'll tell you how your life will unfold
- Tell me the data, and I'll tell you how to extract a significant result
- Both are superstition, false prophecy
- Prophecy must be vague, in order to sound correct

Mars and Saturn, wanderers

Horoscopes

- Models can be accurate without being correct
- The data themselves do not contain information about causes
- It is not possible to reliably learn causes from data alone

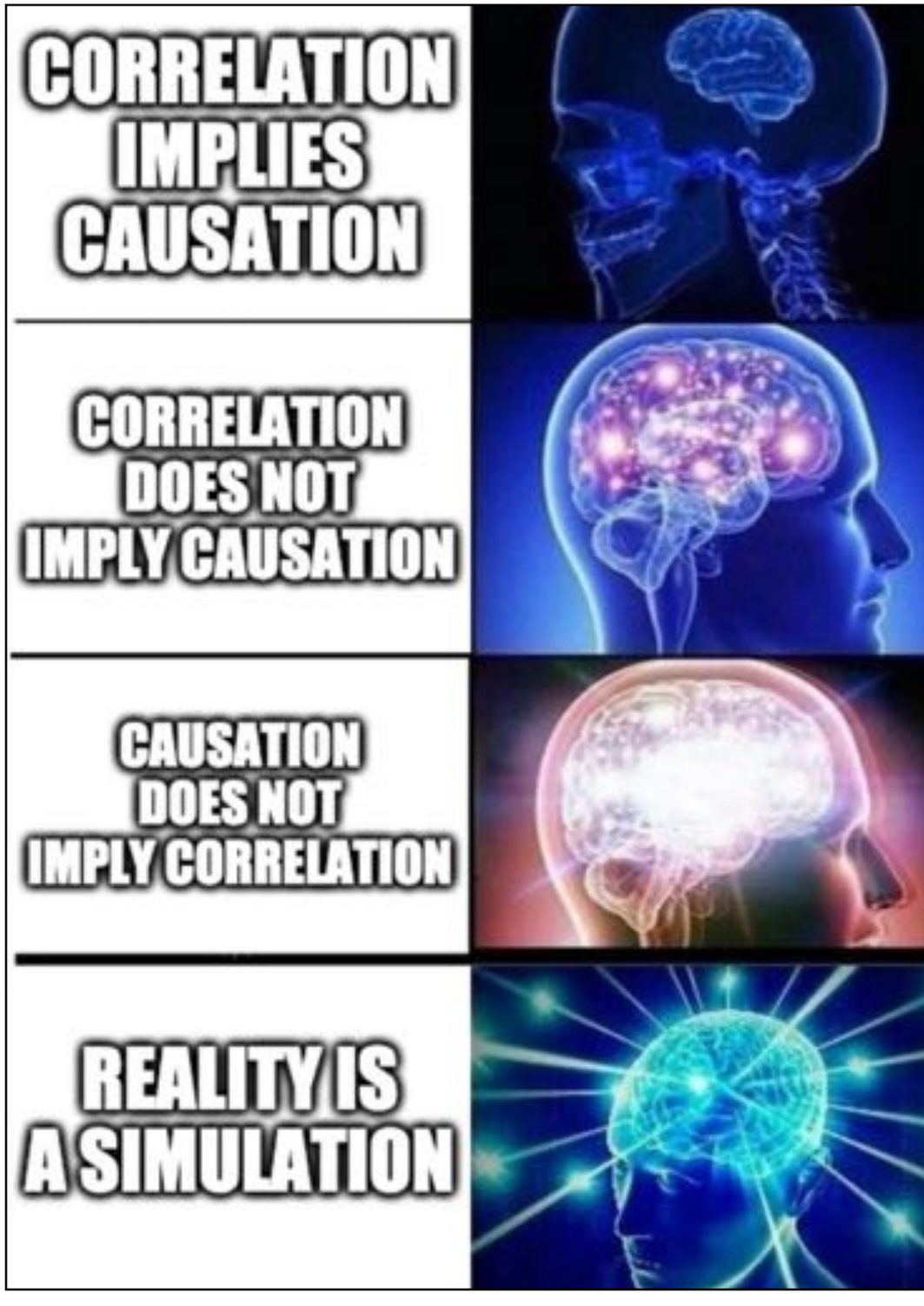



BAYES
FREQUENTISM


CAUSAL
INFERENCE

What is causal inference?

- Causal inference is more than learning an **association**.
- Two concepts:
  - Causal inference is **prediction** of **intervention**
  - Causal inference is **missing data imputation**



Causal prediction

Knowing a **cause** means being able to predict the consequences of an **intervention**.

What if I do this?

Causal imputation

Knowing a **cause** means being able to construct unobserved **counterfactual outcomes**.

What if I had done something else?

Experiments are no refuge

- Why do experiments work? When do they work?
- Should you test for balance?
- What if treatment were imperfect?
- Should you control for anything? Everything?
- What is the causal effect in the target population?
- Answers depend upon causal assumptions.

Description is no refuge

- Descriptive research requires causal inference
  - To do more than describe a **sample**, need causal assumptions
  - Sampling bias, stratified sampling, post-stratification, missing data, measurement error
  - We should not compare samples but **populations**

Today's Agenda

- Part 1: **Causal Salad**
  You must unlearn what you have learned

- Part 2: **Causal Design**
  Drawing our assumptions

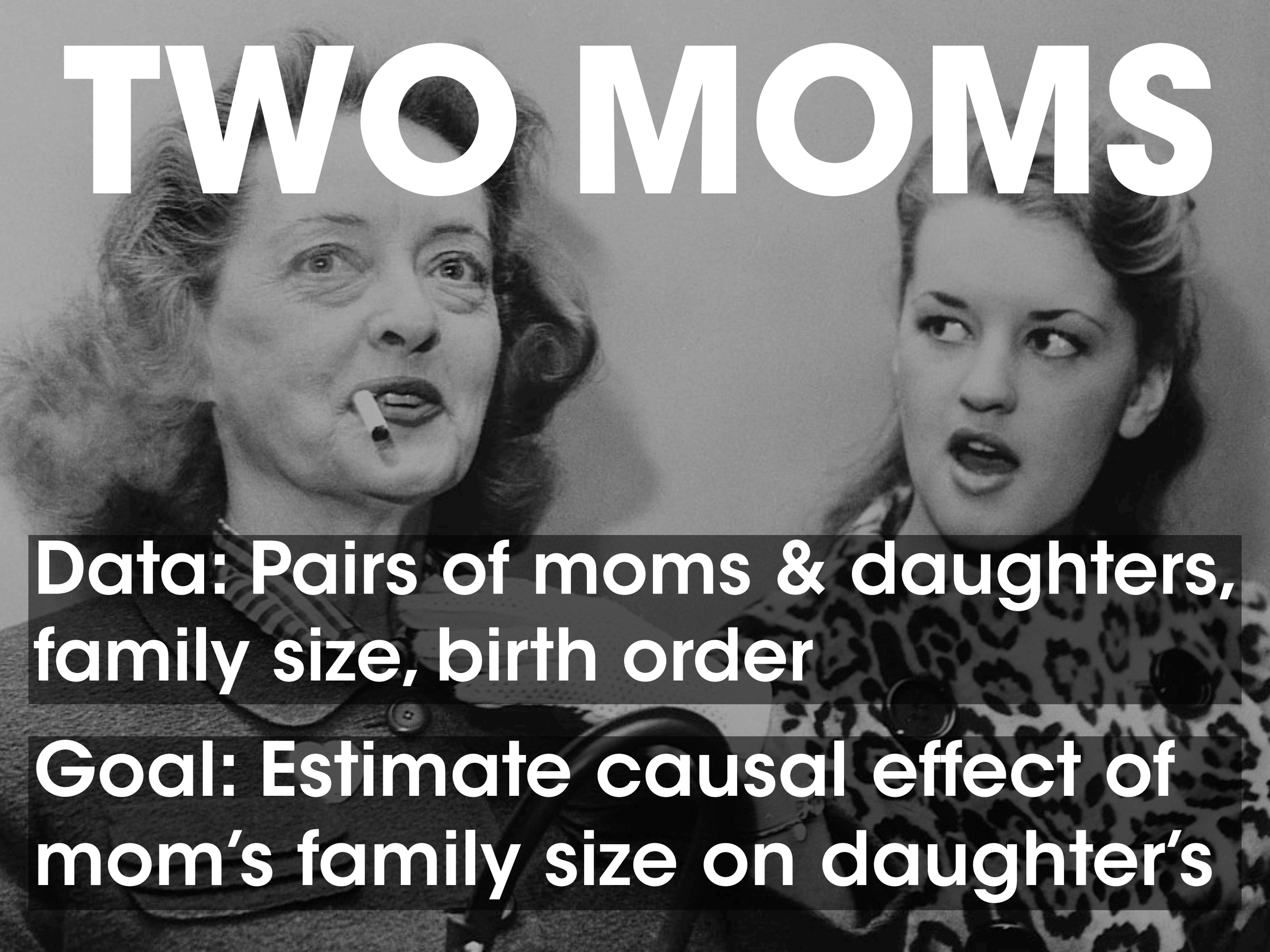
- Part 3: **Full Luxury Bayesian Inference**
  Doing the computations

TWO MOMS

TWO MOMS

Data: Pairs of moms & daughters, family size, birth order

TWO MOMS



Data: Pairs of moms & daughters, family size, birth order

Goal: Estimate causal effect of mom's family size on daughter's

PEER BIAS

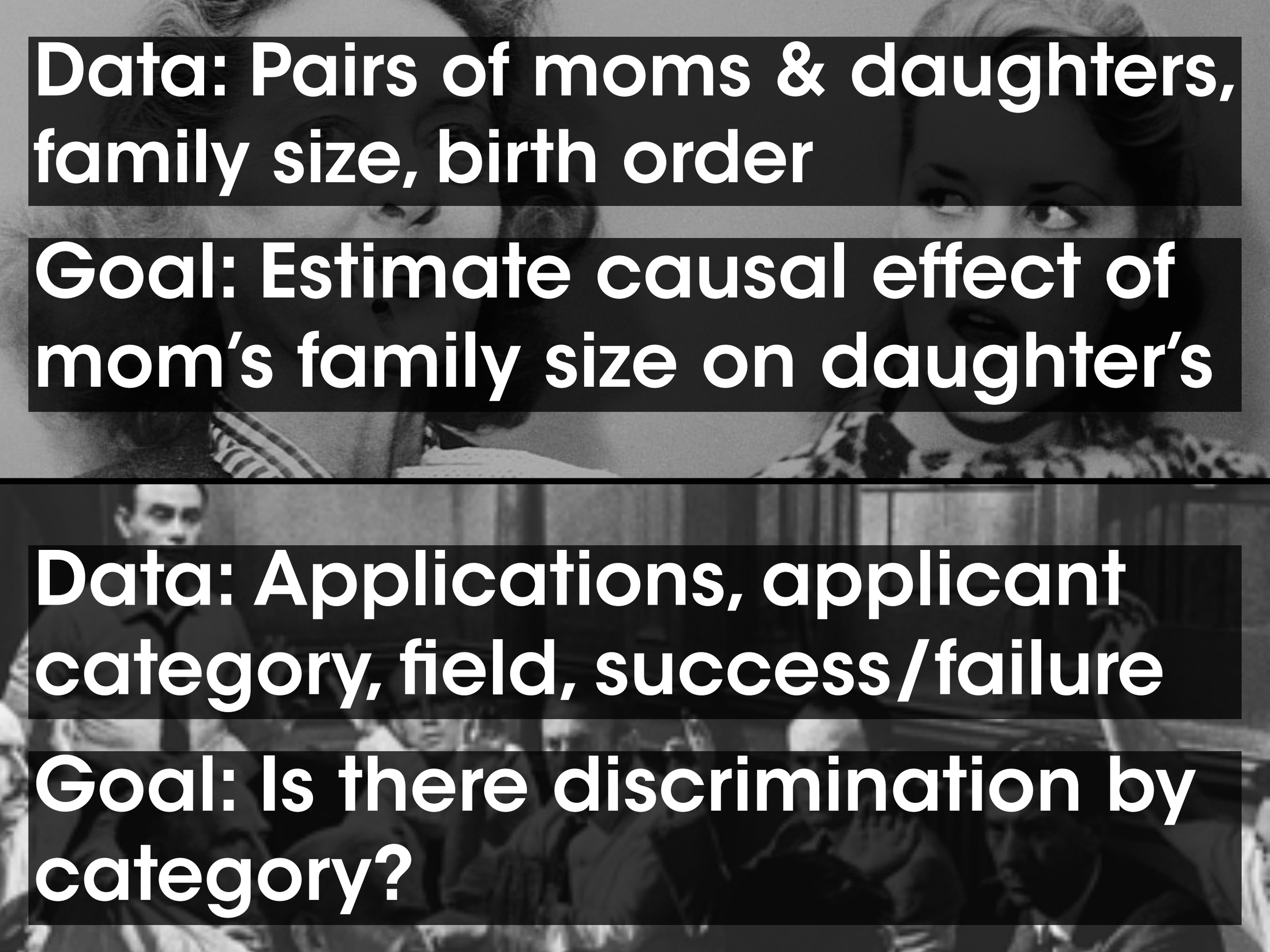
PEER BIAS

Data: Applications, applicant category, field, success/failure

PEER BIAS

Data: Applications, applicant category, field, success/failure

Goal: Is there discrimination by category?


Data: Pairs of moms & daughters, family size, birth order
Goal: Estimate causal effect of mom's family size on daughter's
Data: Applications, applicant category, field, success/failure
Goal: Is there discrimination by category?

CAUSAL SALAD

Robots are cause-blind

x

"panda"
57.7% confidence

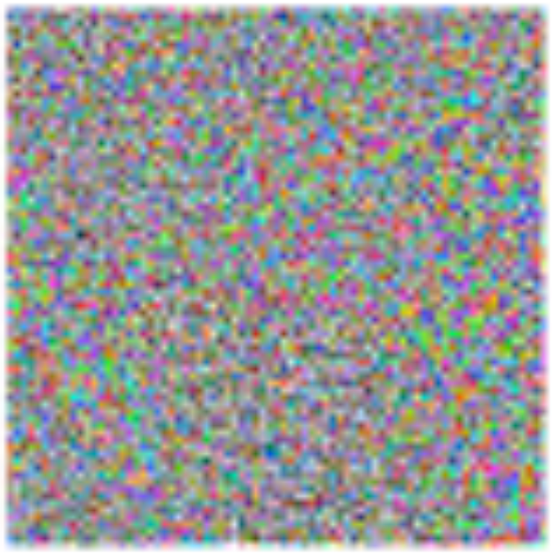
Goodfellow et al 2014 Explaining and Harnessing Adversarial Examples

Robots are cause-blind

x

"panda"
57.7% confidence

$+ .007 \times$

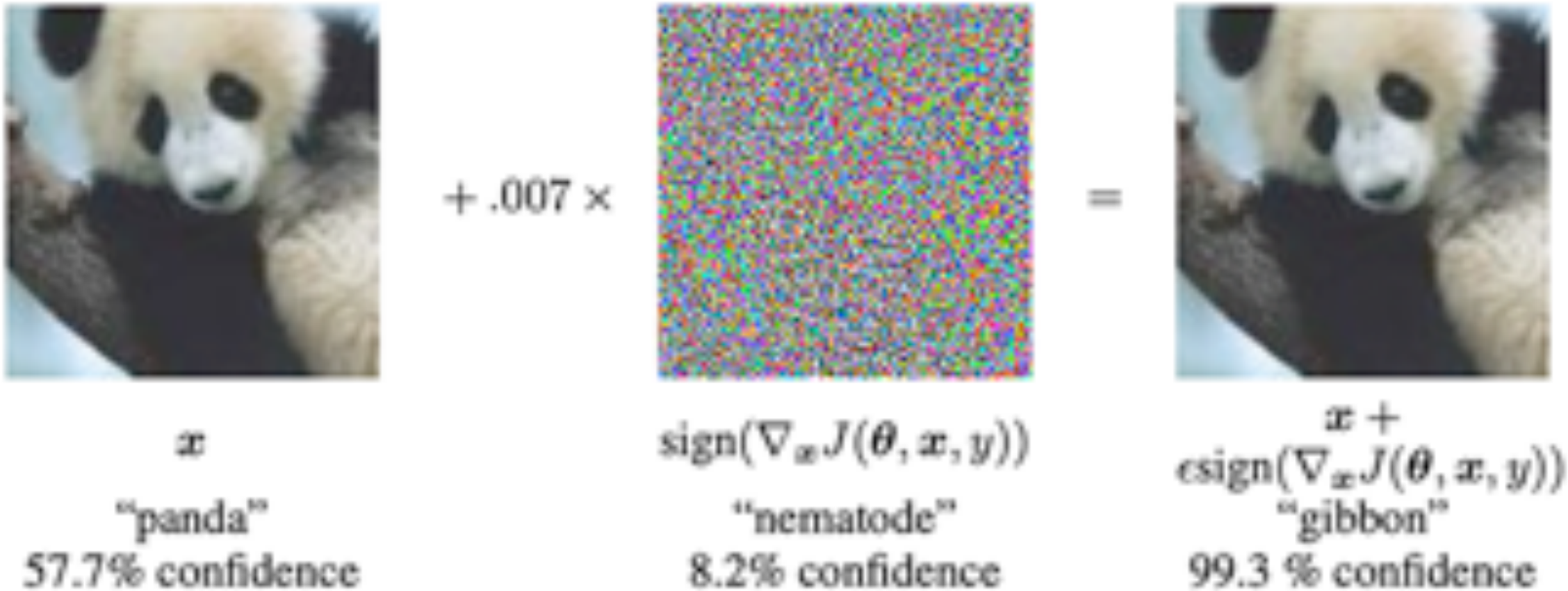


$\text{sign}(\nabla_x J(\theta, x, y))$

"nematode"
8.2% confidence

Goodfellow et al 2014 Explaining and Harnessing Adversarial Examples

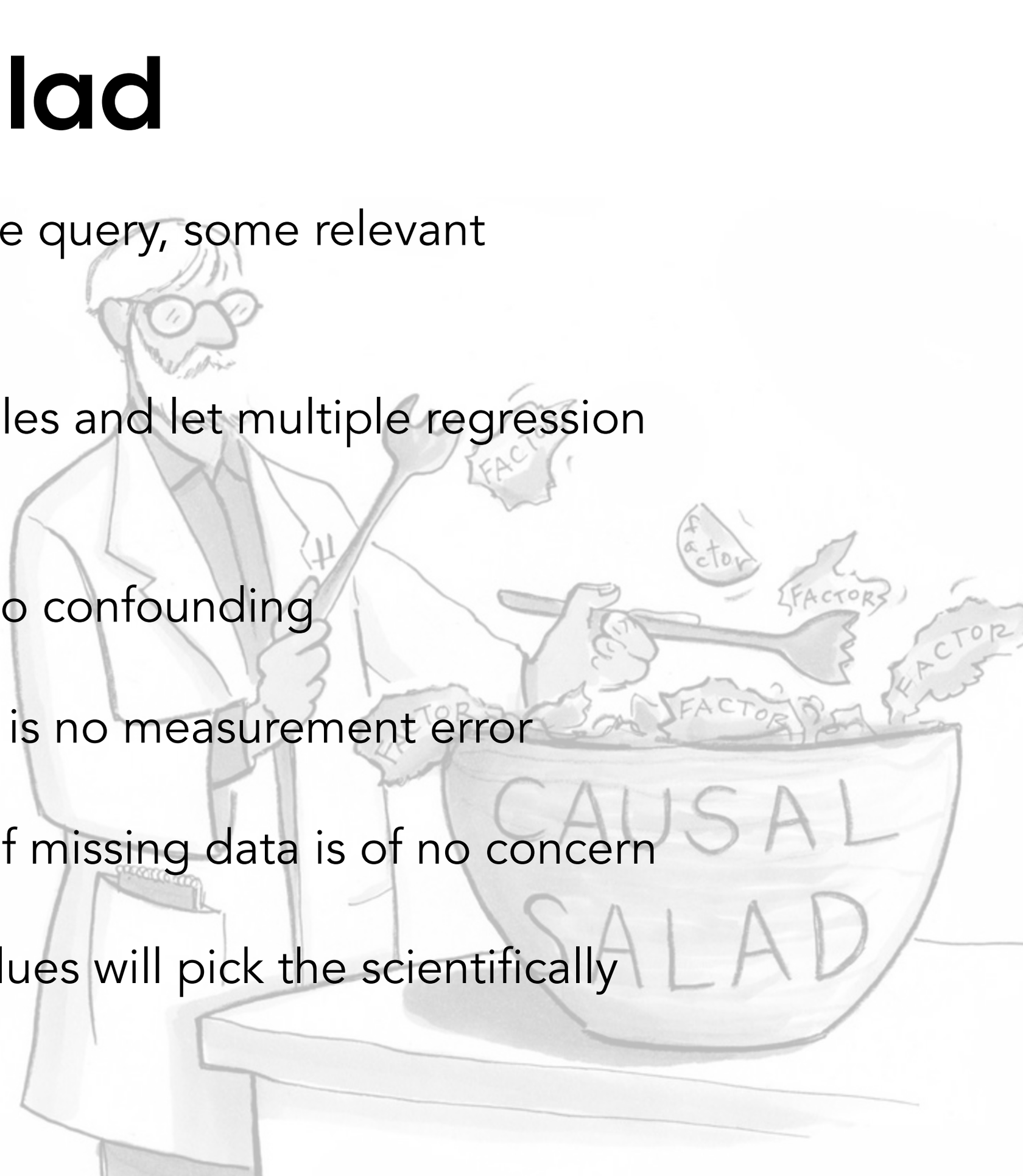
Robots are cause-blind



Goodfellow et al 2014 Explaining and Harnessing Adversarial Examples

Causal Salad

- Ingredients: Vague query, some relevant variables
- Add all the variables and let multiple regression sort it out
- Pretend there is no confounding
- Pretend there are is no measurement error
- Pretend pattern of missing data is of no concern
- Pretend AIC/p-values will pick the scientifically correct model



Multiverse analysis

- Multiverse analysis: "performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of **reasonable** scenarios"

- No surprise that results are sensitive to model structure and data processing

- Conditions that make a scenario "reasonable" are causal, unstated

- Does not prevent causal salad

Table 1. Processing choices

1. Assessment of fertility (F)—high vs low.
   (a) F1: high = cycle days 7–14; low = cycle days 17–25
   (b) F2: high = cycle days 6–14; low = cycle days 17–27
   (c) F3: high = cycle days 9–17; low = cycle days 18–25
   (d) F4: high = cycle days 8–14; low = cycle days 1–7 and 15–28
   (e) F5: high = cycle days 9–17; low = cycle days 1–8 and 18–28
2. Next menstrual onset (NMO)
   (a) NMO1: reported start date previous menstrual onset + computed cycle length
   (b) NMO2: reported start date previous menstrual onset + reported cycle length
   (c) NMO3: reported estimate of next menstrual onset
3. Assessment of relationship status (R) (single vs relationship)
   (a) R1: single = response options 1 and 2; relationship = response options 3 and 4
   (b) R2: single = response option 1; relationship = response options 2, 3, and 4
   (c) R3: single = response option 1; relationship = response options 3 and 4
4. Exclusion of women based on cycle length (ECL)
   (a) ECL1: no exclusion based on cycle length
   (b) ECL2: exclusion of participants with computed cycle length greater than 25 or less than 35 days
   (c) ECL3: exclusion of participants with reported cycle length greater than 25 or less than 35 days
5. Exclusion of women based on certainty ratings of start dates of two previous menstrual periods (EC)
   (a) EC1: no exclusion based on certainty ratings
   (b) EC2: exclusion of participants who are not certain about at least one start date (i.e., sure less than 6)

Steegen et al 2016 Increasing Transparency Through a Multiverse Analysis

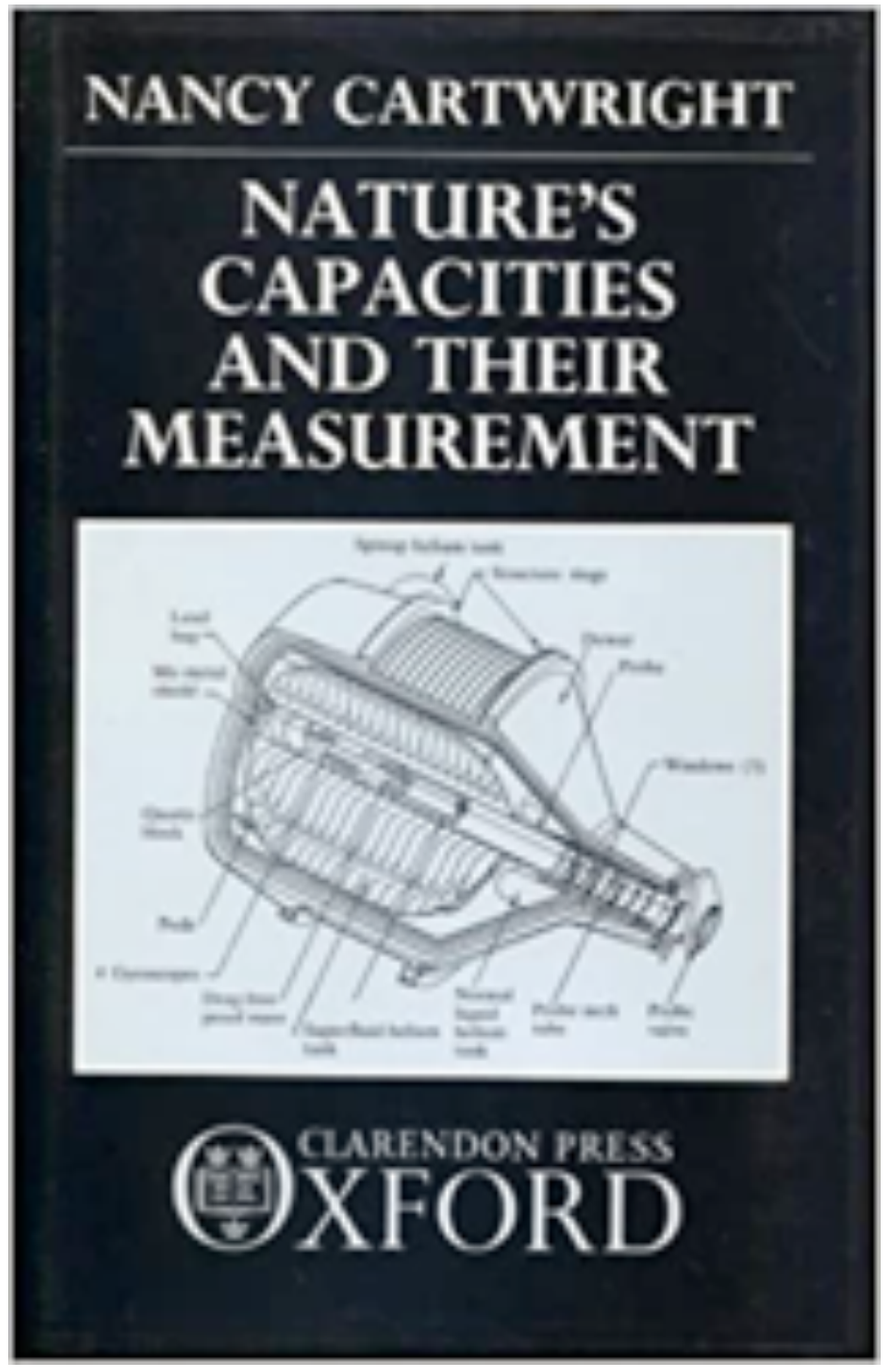
No causes in; No causes out

- Statistical models alone insufficient;
  They do not contain causal information

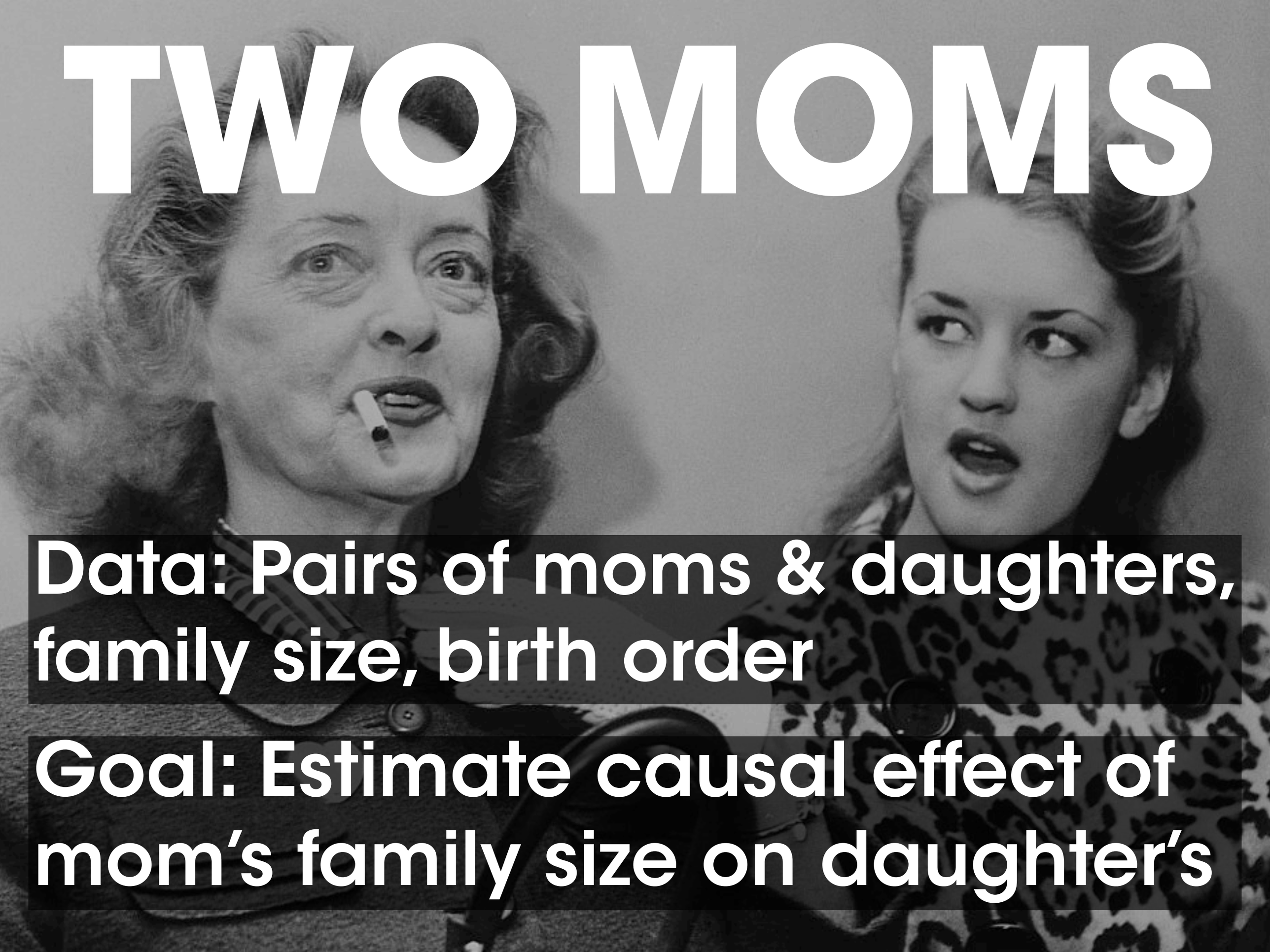
  - Multiple regression does not distinguish causes from confounds

  - *p*-values are not causal statements

  - AIC etc are purely predictive



TWO MOMS



Data: Pairs of moms & daughters, family size, birth order

Goal: Estimate causal effect of mom's family size on daughter's

Two Moms & a Regression

Variables:

Family sizes **M** and **D**

Birth orders **B1** (mom's) and **B2** (daughter's)

How would you construct a regression to estimate influence of **M** on **D**?

Two Moms & a Regression

- Key question: Should you add **B1** and **B2** to the model?

  M ~ D

  M ~ D + B1

  M ~ D + B2

  M ~ D + B1 + B2

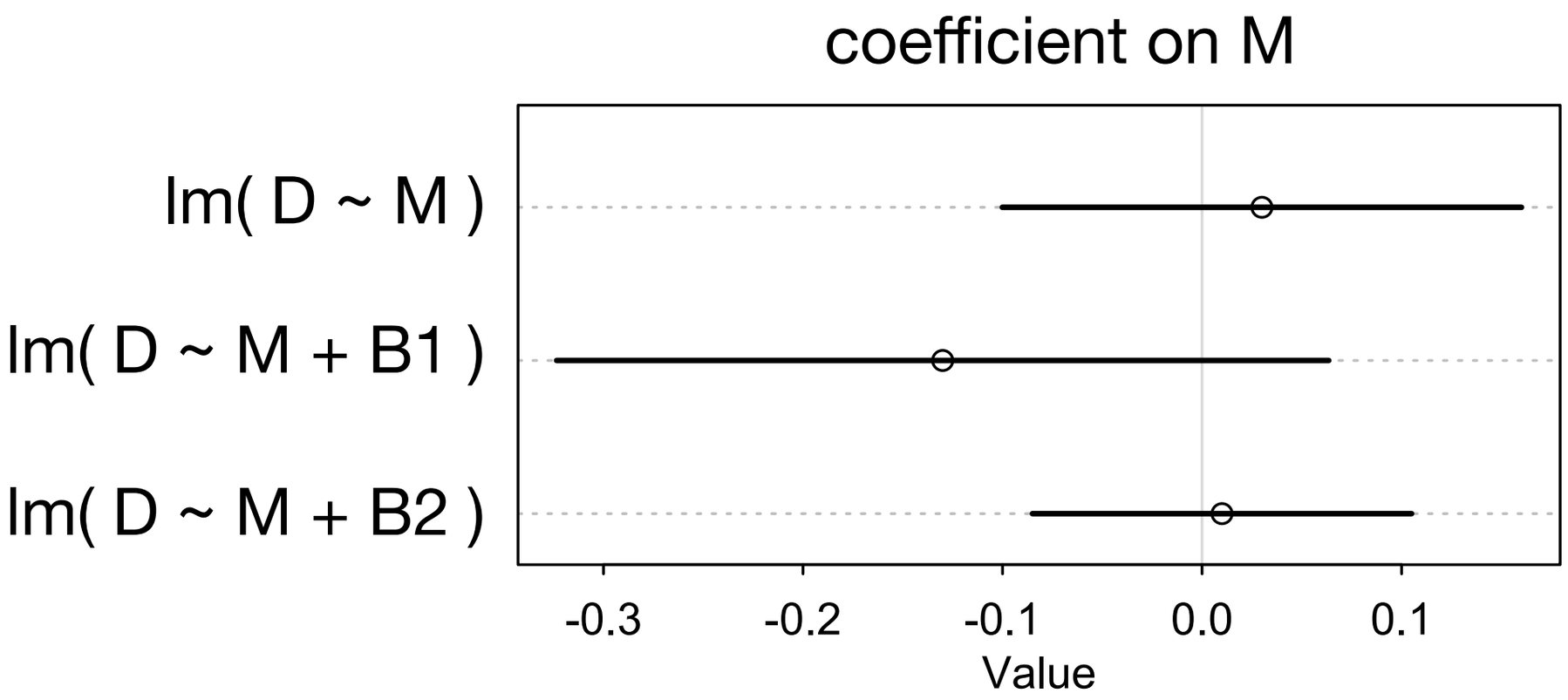
  M ~ D*B1*B2

Two Moms & a Regression

- To know the right thing to do, need a simulation
- Assume that mom's family size has ZERO influence on daughter's
- Assume first borns have higher fertility
- Simulate 200 mother-daughter pairs

Code for examples: github.com/rmcelreath/causal_salad_2021

Two Moms & a Regression



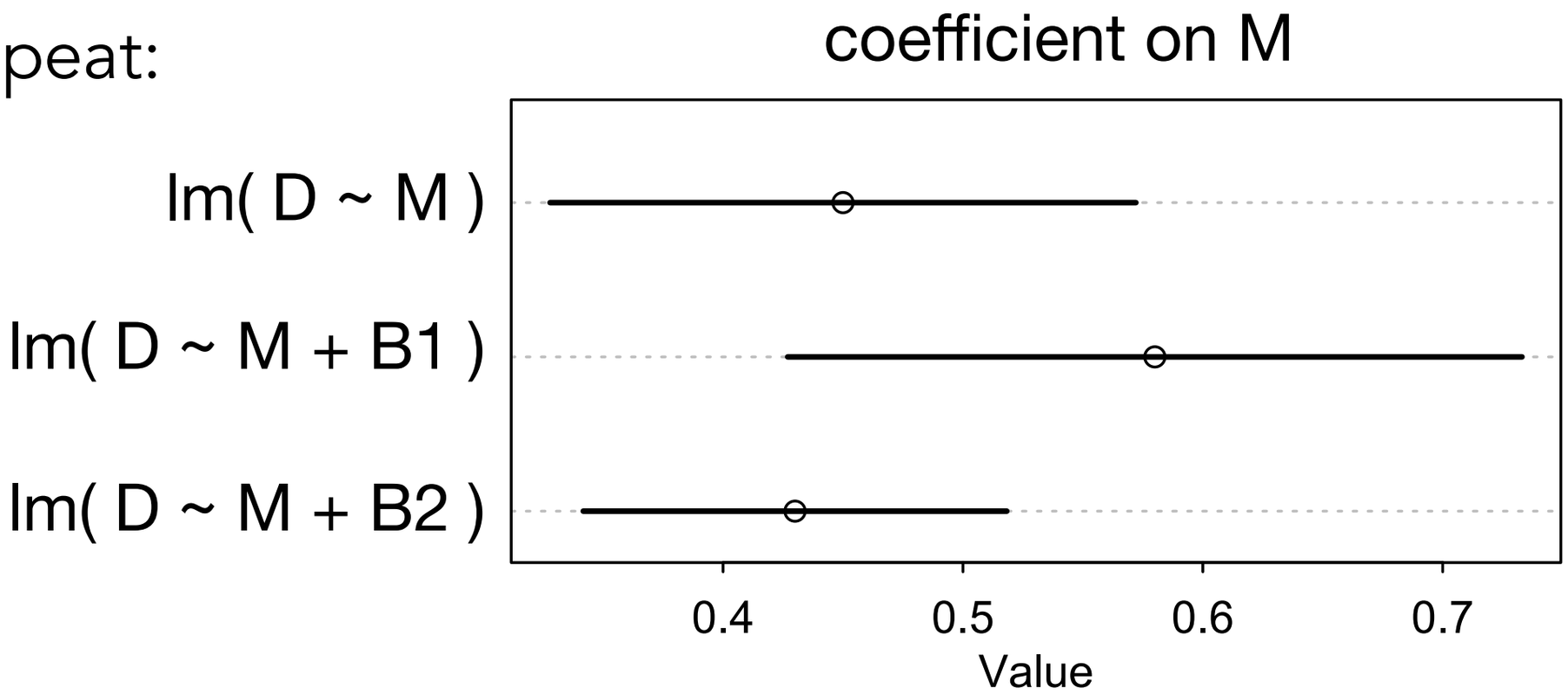
Just an example, but representative of average result.

Adding **B1** hurts inference, adding **B2** helps. Why?

Two Moms & a Confound

- Now suppose there is an unobserved confound:
  A common cause of **M** and **D** such as wealth or education

- Repeat:



When there is a confound, **B1** exaggerates the bias

Two Moms & an Info Criterion

- Predictive criteria like AIC do not help

  AIC(lm(D ~ M)) : 764
  AIC(lm(D ~ M + B1)) : 757
  AIC(lm(D ~ M + B2)) : 635
  AIC(lm(D ~ M + B1 + B2)) : 629

- Why? A model can make good predictions without correct causal structure. Association is not causation.

- Don't ask about *p*-values — they aren't even predictive criteria

Two Moms Summary

- Assuming birth order influences family size:
  - Including mom's birth order (**B1**) hurts inference
  - Including daughter's birth order (**B2**) helps
  - AIC etc select models that include **B1**
  - When **M** and **D** are confounded, adding **B1** exaggerates the confound

PEER BIAS

Data: Applications, applicant category, field, success/failure

Goal: Is there discrimination by category?

Measuring Peer Bias

Variables:

Applicant category **X**

Field/department **E**

Outcome **Y**

How would you construct a regression to estimate influence of **X** on **Y**?

Peer Bias

- Long history of attempts to address this question.
- Key finding: Including field/department in the model changes the result.
- But what does this change mean?

1975

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Determining whether discrimination because of sex or ethnic identity is being practiced against persons seeking passage from one social status or locus to another is an important problem in our society today. It is legally important and morally important. It is also often quite difficult. This article is an exploration of some of the issues of measurement and assessment involved in one example of the general problem, by means of which we hope to shed some light on the difficulties. We will proceed in a straightforward and indeed naive way, even though we know how misleading an unsophisticated approach to the problem is. We do this because we think it quite likely that other persons interested in questions of bias might proceed in just the same way, and careful exposure of the mistakes in our discovery procedure may be instructive.

Data and Assumptions

The particular body of data chosen for examination here consists of applications for admission to graduate study at the University of California, Berkeley, for the fall 1973 quarter. In the admissions cycle for that quarter, the Graduate Division at Berkeley received approximately 15,000 applications, some of which were later withdrawn or transferred to a different proposed entry quarter by the applicants. Of the applications finally remaining for the fall 1973 cycle 12,763 were sufficiently complete to permit a decision to admit or to deny admission. The question we wish to pursue is whether the decision to admit or to deny was influenced by the sex of the applicant. We cannot know with any certainty the influences on the evaluators in the Graduate Admissions Office, or on the faculty reviewing committees, or on any other administrative personnel participating in the chain of actions that led to a decision on an individual application. We can, however, say that if the admissions decision and the sex of the applicant are statistically associated in the results of a series of applications, we may judge that bias existed, and we may then seek to find whether discrimination existed. By "bias" we mean here a pattern of association between a particular decision and a particular sex of applicant, of sufficient strength to make us confident that it is unlikely to be the result of chance alone. By "discrimination" we mean the exercise of decision influenced by the sex of the applicant when that is immaterial to the qualifications for entry.

The simplest approach (which we shall call approach A) is to examine the aggregate data for the campus. This approach would surely be taken by many persons interested in whether bias in admissions exists on any campus. Table 1 gives the data for all 12,763 applications to the 101 graduate departments and interdepartmental graduate majors to which application was made for fall 1973 (we shall refer to them all as departments). There were 8442 male applicants and 4321 female applicants. About 44 percent of the males and about 35 percent of the females were admitted. Just this

Dr. Bickel is professor of statistics, Dr. Hammel is professor of anthropology and associ-

2015

Gender contributes to personal research funding success in The Netherlands

Romy van der Lee[1] and Naomi Ellemers[2]

Department of Social and Organizational Psychology, Institute of Psychology, Leiden University, 2300 RB Leiden, The Netherlands

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved August 19, 2015 (received for review May 26, 2015)

We examined the application and review materials of three calls (n = 2,823) of a prestigious grant for personal research funding in a national full population of early career scientists awarded by the Netherlands Organization for Scientific Research (NWO). Results showed evidence of gender bias in application evaluations and success rates, as well as in language use in instructions and evaluation sheets. Male applicants received significantly more competitive "quality of researcher" evaluations (but not "quality of proposal" evaluations) and had significantly higher application success rates than female applicants. Gender disparities were most prevalent in scientific disciplines with the highest number of applications and with equal gender distribution among the applicants (i.e., life sciences and social sciences). Moreover, content analyses of the instructional and evaluation materials revealed the use of gendered language favoring male applicants. Overall, our data reveal a 4% "loss" of women during the grant review procedure, and illustrate the perpetuation of the funding gap, which contributes to the underrepresentation of women in academia.

gender bias | research funding | success rates | academia | STEM

Women are still underrepresented in academia today. Despite various attempts to promote gender equality (e.g., affirmative action initiatives, quotas), female scientists are less likely to get offered tenure, are judged to be less competent, receive less payment and research facilities, and are less likely to be awarded research grants compared with male scientists (1–3). Over time, this type of bias accumulates and contributes to the attrition of women from associated with male traits and considered necessary for academic career success (20). Moreover, women still earn on average 18% less than their male colleagues for the same work with similar responsibilities (3). Although the salary gap seems to narrow for early career researchers, women in top academic positions are still substantially underpaid compared with men. Finally, across different career phases, success rates for female scientists applying for research funding tend to be lower than for male scientists (3, 21, 22). Even when overall success rates for men and women are equal, women receive less research funding than men, and are less often listed as principal investigators (23–25). Closing the funding gap is of particular importance, because this may help retain women in academia and foster the closing of other gaps by facilitating negotiations about salaries, research facilities, and promotion opportunities.

Current Study

To investigate the possibility of a funding gap, we examined a national full population of early career researchers who applied for a prestigious personal grant between 2010 and 2012 (Innovational Research Incentives Scheme Veni; n = 2,823, with 42.1% female applicants) awarded by the Netherlands Organization for Scientific Research (NWO). The NWO made available anonymized data from their archives for the purpose of this study, and approved publication of this research.

Our focus in this study was twofold. First, we tested for applicant gender differences in success rates and application evaluations. In

Dutch research funding, gender bias, and Simpson's paradox

Based on, among other criteria, three consecutive years of grant applications to the "Veni programme" of the Netherlands Organization for Scientific Research (NWO), van der Lee and Ellemers (1) conclude that these data "provide compelling evidence of gender bias in personal grant applications to obtain research funding." One of the main results this claim is based upon is that of the 1,635 applications by males, 17.7% were successful, whereas of the 1,188 applications by females, only 14.4% were successful. When applying the χ^2 test to the data, the authors found a P value of 0.045 (1). This conclusion is based on the application of an inappropriate statistical procedure, and is therefore questionable, due to the so-called "Simpson's paradox."

Simpson's paradox dates back to Simpson (2) and gained familiarity after a report by Bickel et al. (3). In essence, Simpson's paradox states that an apparently significant relation between two variables in a contingency table can be due to a joint dependency on a third variable.

In table S1 of the article by van der Lee and Ellemers (1), a breakdown of grant applications per research discipline is presented. The proportion of female applicants varies from 11.8% visualizes these data and immediately shows a clear negative relation between the proportion of female applicants and the total success rate. It turns out that women tend to apply more often to competitive disciplines, such as health sciences and social sciences, whereas men apply more often to less competitive disciplines, such as physics and chemical sciences. In four of the nine disciplines, women have a higher success rate than men, and in five of the nine disciplines, men have a higher success rate than women. When taking into account that multiple comparisons are performed, the gender bias is significant (at the $\alpha = 0.05$ level) for none of the disciplines. Thus, when taking into account the spurious correlation, the statistical significance of the relation vanishes.

As a consequence, the conclusion of van der Lee and Ellemers (1) could be severely influenced by this issue, at least for the (large) part where this claim is based on the supposed P value of 0.045. The authors perform other tests on various aspects of the proposal review process, but the data provided in the paper are insufficient to check whether these tests suffer from Simpson's paradox as well. Because these tests study related issues, it is not unlikely that Dutch science funding. This scrutiny includes studying data from other years and programs. Interestingly, on the aggregate level, men received relatively more grants than women in eight of the 14 Veni calls, with the difference, again, not being statistically significant (4).

ACKNOWLEDGMENTS. This contribution has benefitted from communications with Daniël Lakens, Jelte Wicherts, and Ivo Krausz.

Casper J. Albers[1]

Heymans Institute for Psychological Research, University of Groningen, 9712 TS Groningen, The Netherlands

1 van der Lee R, Ellemers N (2015) Gender contributes to personal research funding success in The Netherlands. *Proc Natl Acad Sci USA* 112(40):12349–12353.

2 Simpson EH (1951) The interpretation of interaction in contingency tables. *J R Stat Soc, B* 13:238–241.

3 Bickel PJ, Hammel EA, O'Connell JWO (1975) Sex bias in graduate admissions: Data from berkeley. *Science* 187(4175): 398–404.

4 Albers CJ (2015) NWO, Gender Bias and Simpson's Paradox. Available at blog.casperalbers.nl/science/nwo-gender-bias-and-simpsons-paradox/. Accessed November 23, 2015.

An Empirical Analysis of Racial Differences in Police Use of Force

Roland G. Fryer Jr.

Harvard University and National Bureau of Economic Research

This paper explores racial differences in police use of force. On nonlethal uses of force, blacks and Hispanics are more than 50 percent more likely to experience some form of force in interactions with police. Adding controls that account for important context and civilian behavior reduces, but cannot fully explain, these disparities. On the most extreme use of force—officer-involved shootings—we find no racial differences either in the raw data or when contextual factors are taken into account. We argue that the patterns in the data are consistent with a model in which police officers are utility maximizers, a fraction of whom have a preference for discrimination, who incur relatively high expected costs of officer-involved shootings.

> We can never be satisfied as long as the Negro is the victim of the unspeakable horrors of police brutality. (Martin Luther King Jr., August 28, 1963)

I. Introduction

From "Bloody Sunday" on the Edmund Pettus Bridge to the public beatings of Rodney King, Bryant Allen, and Freddie Helms, the relationship

This work has benefited greatly from discussions and debate with Chief William Evans, Chief Charles McClelland, Chief Martha Montalvo, Sergeant Stephen Morrison, Jon Murad, Lynn Overmann, Chief Bud Riley, and Chief Scott Thomson. I am grateful to David Card, Kerwin Charles, Christian Dustmann, Michael Greenstone, James Heckman, Richard Holden, Lawrence Katz, Steven Levitt, Jens Ludwig, Glenn Loury, Kevin Murphy, Derek Neal, John Overdeck, Jesse Shapiro, Andrei Shleifer, Jorg Spenkuch, Max Stone, John Van Reenan, Christopher Winship, and seminar participants at Brown University, University of Chicago, London

American Political Science Review (2020) 114, 3, 619–637
doi:10.1017/S0003055420000039

Administrative Records Mask Racially Biased Policing

DEAN KNOX *Princeton University*
WILL LOWE *Hertie School of Governance*
JONATHAN MUMMOLO *Princeton University*

Researchers often lack the necessary data to credibly estimate racial discrimination in policing. In particular, police administrative records lack information on civilians police observe but do not investigate. In this article, we show that if police racially discriminate when choosing whom to investigate, analyses using administrative records to estimate racial discrimination in police behavior are statistically biased, and many quantities of interest are unidentified—even among investigated individuals—absent strong and untestable assumptions. Using principal stratification in a causal mediation framework, we derive the exact form of the statistical bias that results from traditional estimation. We develop a bias-correction procedure and nonparametric sharp bounds for race effects, replicate published findings, and show the traditional estimator can severely underestimate levels of racially biased policing or mask discrimination entirely. We conclude by outlining a general and feasible design for future studies that is robust to this inferential snare.

Concern over racial bias in policing, and the public availability of large administrative data sets documenting police–civilian interactions, have prompted a raft of studies attempting to quantify the effect of civilian race on law enforcement behavior. These studies consider a range of outcomes including ticketing, stop duration, searches, and the use of force (e.g., Antonovics and Knight 2009; Fryer 2019; Ridgeway 2006; Nix et al. 2017). Most research in this area attempts to adjust for omitted variables that may correlate with suspect race and the outcome of interest. In contrast, this study addresses a more fundamental problem that remains even if the vexing issue of omitted variable bias is solved: the inevitable statistical bias that results from studying racial discrimination using records that are themselves the product of racial discrimination (Angrist and Pischke 2008; Elwert and Winship 2014; Rosenbaum 1984). We show that when there is any racial discrimination in the decision to detain civilians—a decision that determines which encounters appear in police administrative data at all—then estimates of the effect of civilian race on subsequent police behavior are biased absent additional data and/or strong and untestable assumptions.

This study makes several contributions. We clarify the causal estimands of interest in the study of racially discriminatory policing—quantities that many studies appear to be targeting, but are rarely made explicit—and show that the conventional approach fails to recover any known causal quantity in reasonable settings. Next, we highlight implicit and highly implausible assumptions in prior work and derive the statistical bias when they are violated. We proceed to develop informative nonparametric sharp bounds for the range of possible race effects, apply these in a reanalysis and extension of a prominent article on police use of force (Fryer 2019), and present bias-corrected results that suggest this and similar studies drastically underestimate the level of racial bias in police–civilian interactions. Finally, we outline strategies for future data collection and research design that can mitigate these threats to inference. These are discussed in the context of a detailed and feasible proposed study of racial bias in traffic stops.

As we show in this article, the difficulty of estimating racial bias using police records stems from a thorny combination of mediation (Hernán, Hernández-Diáz, and Robins 2004; Imai et al. 2011; Pearl 2001; Robins, Hernán, and Brumback 2000; VanderWeele 2009) and selection (Heckman 1979; Lee 2009): the effect of civilian race on the outcome of a police encounter is mediated by whether the civilian is stopped by police, but analysts only have data for one level of the mediator—that is, data on stopped individuals. Because of this, police records do not contain a representative sample of all individuals that police observe, but rather only those civilian encounters which escalated to the point of triggering a reporting requirement. If a civilian's race affects whether officers choose to stop that civilian (Gelman, Fagan, and Kiss 2007; Glaser 2014), then analyzing administrative police records amounts to conditioning on a variable that is itself affected by suspect race, namely, whether a suspect appears in the data at all. This could occur if officers have a higher

Dean Knox, Assistant Professor of Politics, Princeton University, dcknox@princeton.edu.
Will Lowe, Senior Research Scientist, Hertie School of Governance, lowe@hertie-school.org.
Jonathan Mummolo, Assistant Professor of Politics and Public Affairs, Princeton University, jmummolo@princeton.edu.

We thank Matt Blackwell, Chuck Cameron, Tom Clark, Scott Cunningham, Lauren Davenport, Naoki Egami, Jeffrey Fagan, Avi Feller, Adam Glynn, Phillip Atiba Goff, Justin Grimmer, Andy Hall, Anna Harvey, Dan Hopkins, Matias Iaryczower, Kosuke Imai, Damon Jones, Dorothy Kronick, Shiro Kuriwaki, Neil Malhotra, Moritz Marbach, Nolan McCarty, Cyrus Samii, Maya Sen, Tara Slough, Rocio Titiunik, Tyler VanderWeele, Vesla Weaver, and Sean Westwood for helpful feedback. We thank Michael Pomirchy for research assistance. Replication files are available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/KFQOCV.

Received: April 26, 2019; revised: October 10, 2019; accepted: January 8, 2020.

Peer Bias

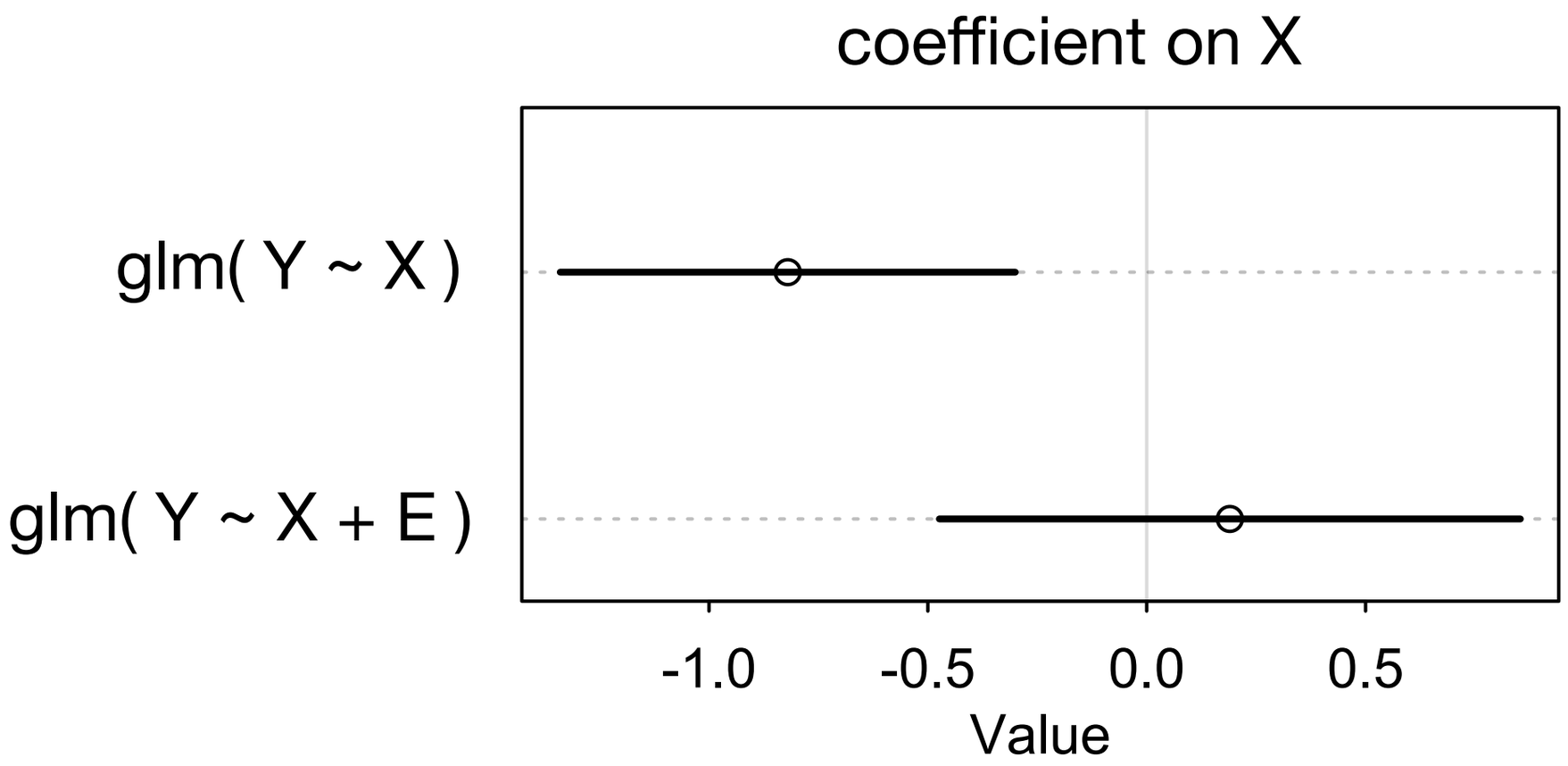
- Simulation example:
  500 applications in two subjects (**E**)
  Subjects vary by average acceptance rate
  Category **X** is NOT a target of discrimination
  Outcome **Y** accepted/rejected

- Consider two regressions:

  Y ~ X
  Y ~ X + E

Code for examples: github.com/rmcelreath/causal_salad_2021

Peer Bias



Why does **E** make false evidence of discrimination vanish?

Peer Bias Again

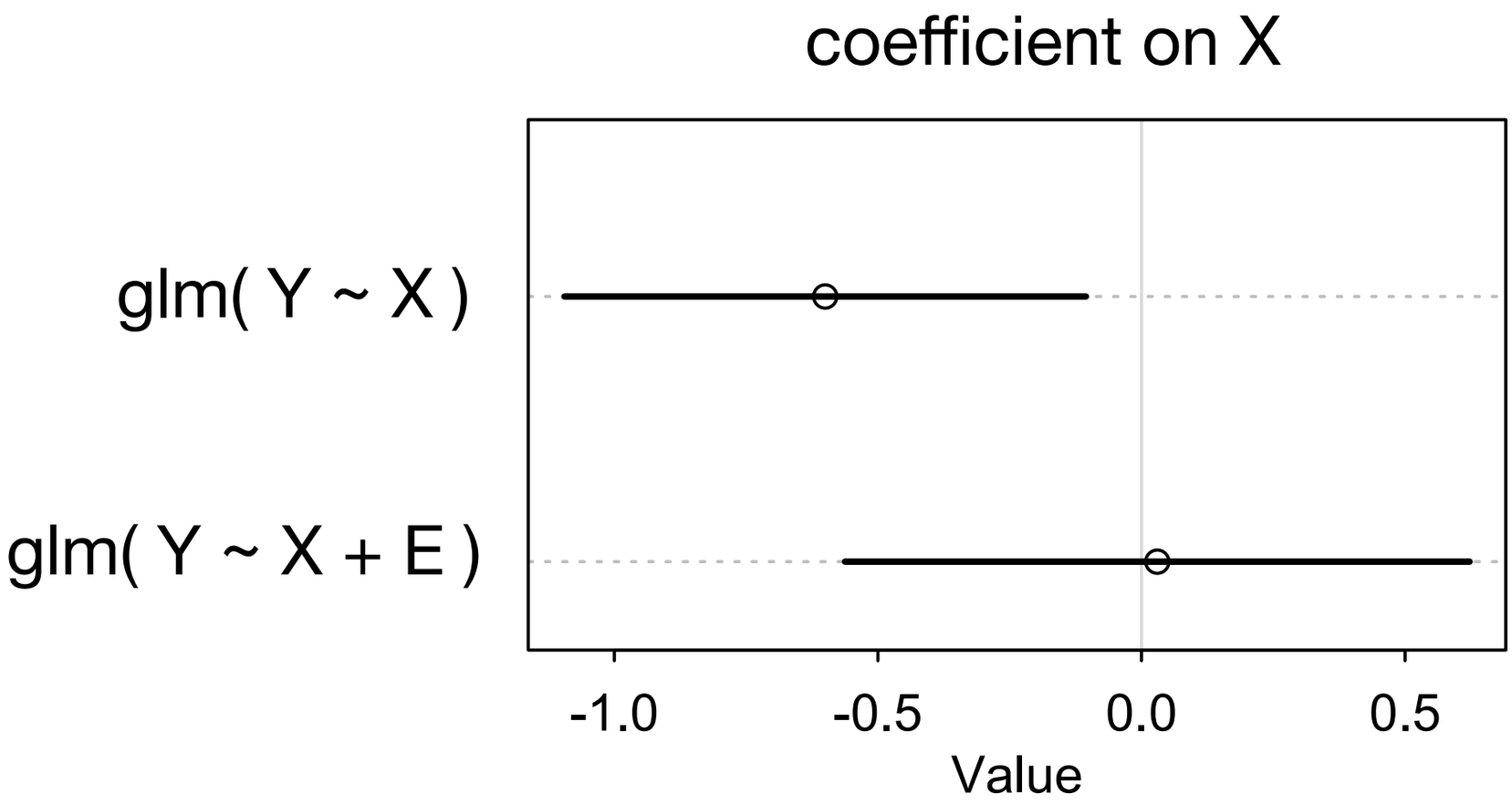
- ANOTHER Simulation example:
  500 applications in two subjects (**E**)
  Subjects vary by average acceptance rate
  Category **X** target of discrimination in both subjects
  Outcome **Y** accepted/rejected

- Consider two regressions:

  Y ~ X
  Y ~ X + E

Code for examples: github.com/rmcelreath/causal_salad_2021

Peer Bias Again



Why does **E** make discrimination vanish, even though there is discrimination in the simulation?

Peer Bias Summary

- Assuming no discrimination against **X**
  - Model without field (**E**) shows discrimination
  - Model with **E** finds no evidence of it
- Assuming there is discrimination based on category **X**
  - Model without field (**E**) finds lower success for **X**
  - Model with **E** finds no evidence of lower success
- In real data, we don’t know the truth
- Change of coefficient could be revelation or deception