Text and Causal Estimation

Text as causal confounders and mediators

Katherine A. Keith

March 28, 2022

Georgia Tech, CS 6471: Computational Social Science

How does COVID-19 vaccination affect the severity of COVID-19 (when contracted)?

This is a causal question!

Intervention (Treatment)

How does *COVID-19 vaccination* affect the severity of COVID-19 (when contracted)?

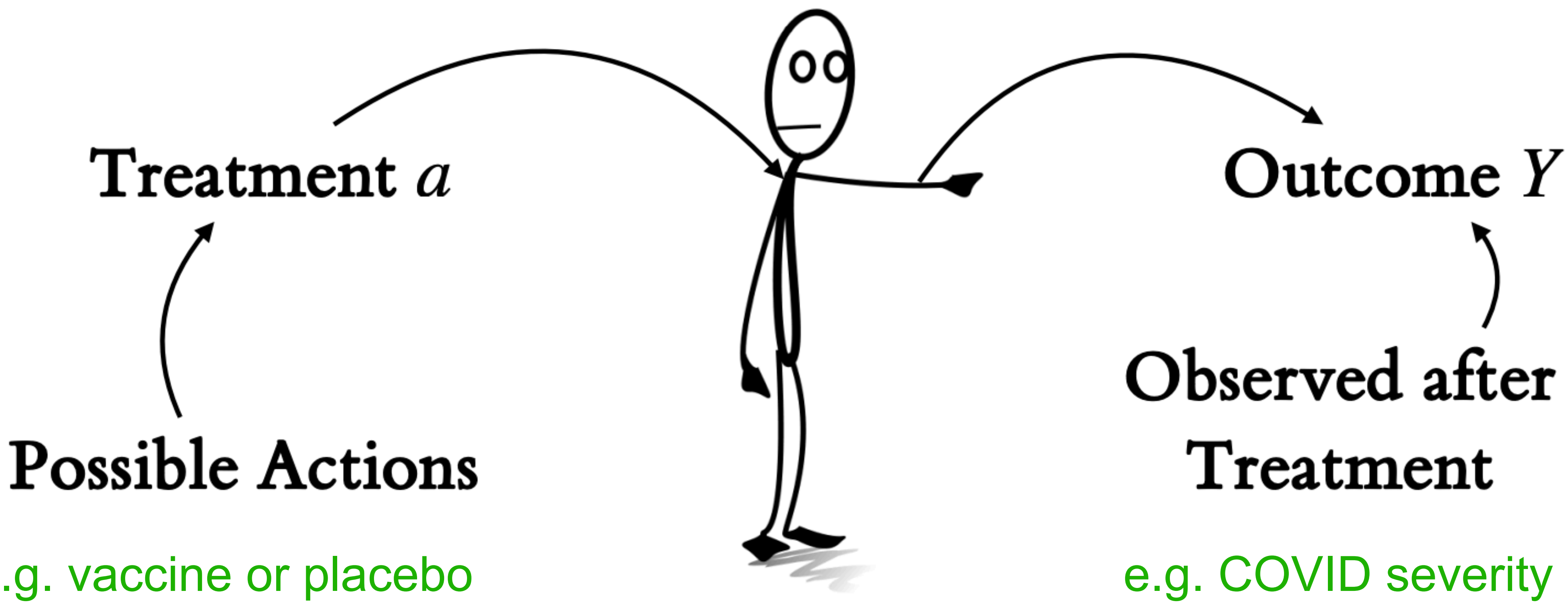
This is a causal question!

Intervention (Treatment)

How does *COVID-19 vaccination* affect the *severity* of COVID-19 (when contracted)?

Outcome

General causal set-up



e.g. vaccine or placebo

e.g. COVID severity

Slide credit: Emaad Manzoor

Causal Estimand: Individual Treatment Effect (ITE)

$Y^{a=1}$

Outcome had individual been vaccinated

$Y^{a=0}$

Outcome had individual been given placebo

Potential outcomes notation

(Rubin, 1974)

Slide credit: Emaad Manzoor

Causal Estimand: Individual Treatment Effect (ITE)

$Y^{a=1}$	$Y^{a=0}$
Outcome had individual been vaccinated	Outcome had individual been given placebo

Individual Treatment Effect (ITE)	$Y^{a=1} - Y^{a=0}$

Slide credit: Emaad Manzoor

ITEs can not be "identified"

"Cannot be measured from observable data"

$Y^{a=1} = Y$ if vaccinated $\longrightarrow$ $Y^{a=0} = ?$

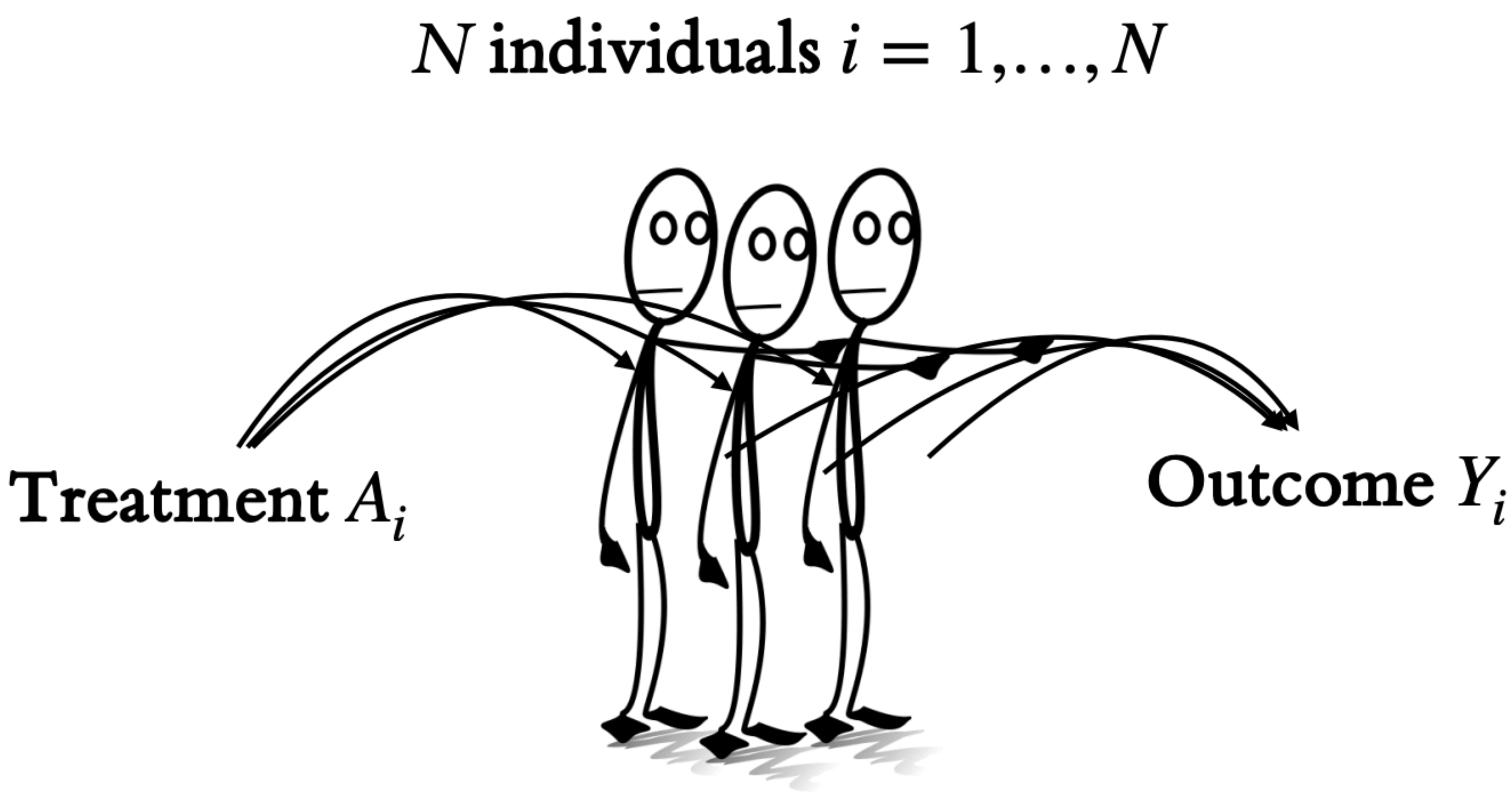
$Y^{a=0} = Y$ if placebo $\longrightarrow$ $Y^{a=1} = ?$

$Y^{a=1}$ and $Y^{a=0}$ not observable simultaneously

Slide credit: Emaad Manzoor

Causal Estimand: Average Treatment Effect (ATE)

Changing our estimand to be at the ***population*** level



"Fundamental problem of causal inference" (Holland 1986)

Slide credit: Emaad Manzoor

Causal Estimand: Average Treatment Effect (ATE)

Changing our estimand to be at the ***population*** level

$Y_i^{A_i=1}$

Outcome had i been vaccinated

$Y_i^{A_i=0}$

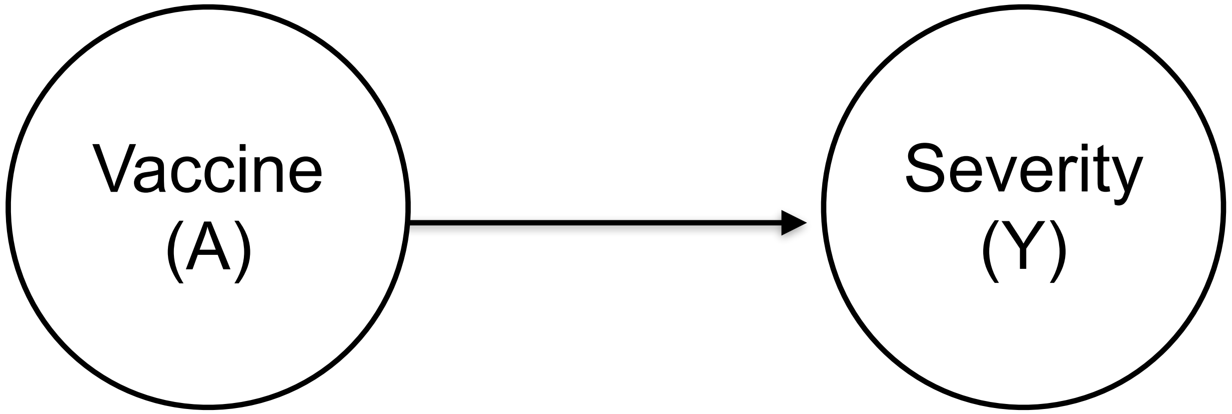
Outcome had i been given placebo

Average Treatment Effect (ATE)	$\mathbb{E}[Y_i^{A_i=1}] - \mathbb{E}[Y_i^{A_i=0}]$

Slide credit: Emaad Manzoor

Naive estimation approach to ATE

Compare outcomes of vaccine takers to non-vaccine takers

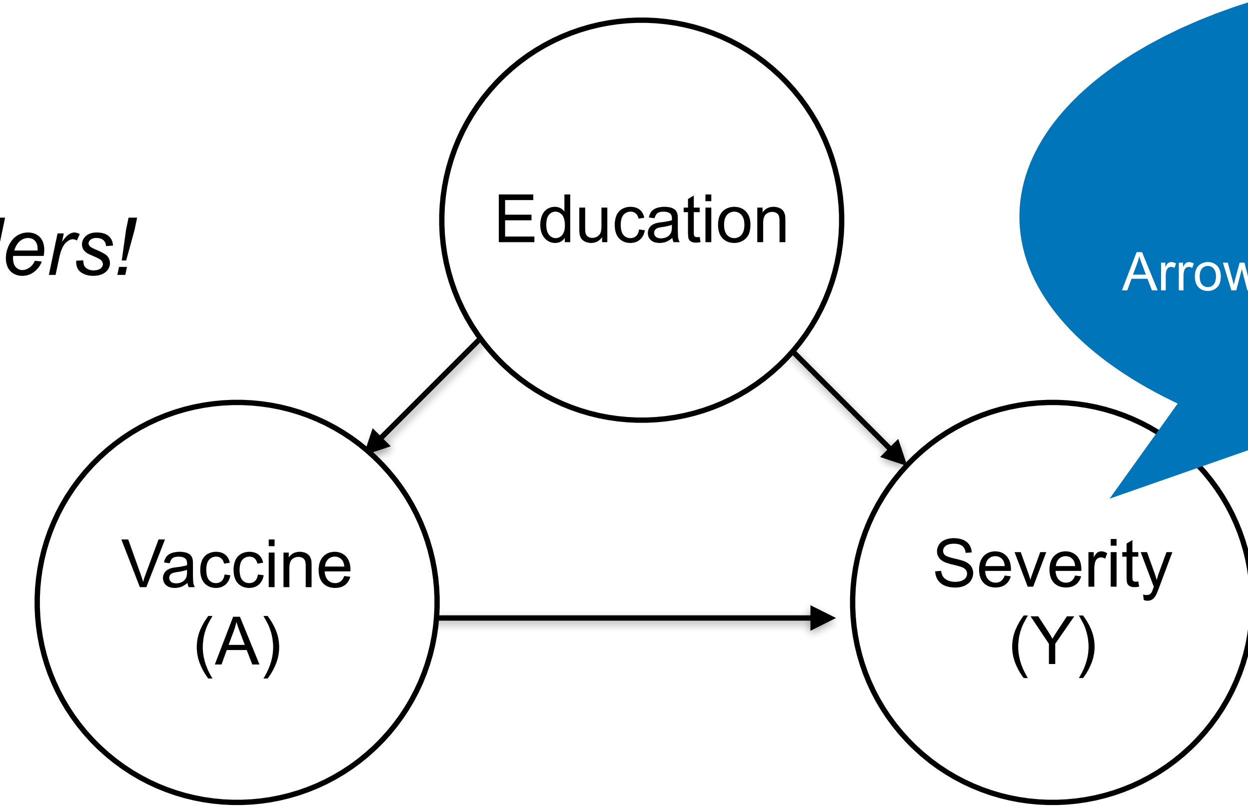


$$E[Y_i | A_i = 1] - E[Y_i | A_i = 0]$$

Naive estimation approach to ATE

Compare outcomes of vaccine takers to non-vaccine takers

Issue:
Confounders!

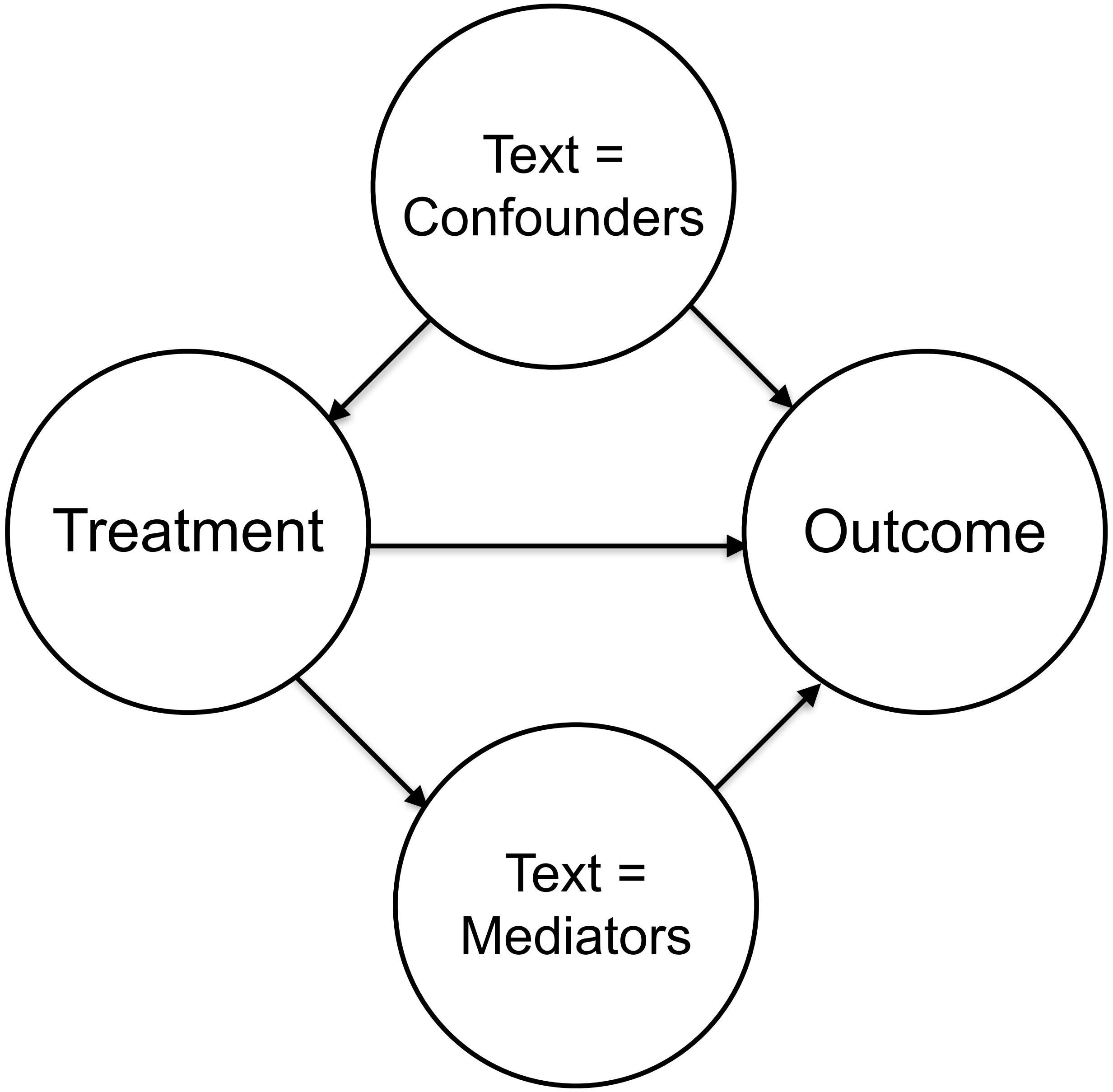


Causal graph
Nodes = variables
Arrows = causal dependencies

$$E[Y_i \,|\, A_i = 1] - E[Y_i \,|\, A_i = 0]$$

With confounders, this **does not** equal the ATE

Scope of this talk



Lecture Outline and Learning Objectives

1. Causal estimation, in general
   A. What is causal estimation and how does it differ from association and prediction?
   B. What are the challenges with causal estimation with text?

2. Text as causal confounders
   A. For observational data, how does one use back-door adjustment for text as a confounder?

3. Text as causal mediators
   A. For observational data, how does one estimate the natural direct and indirect causal effects with text as a mediator?

Lecture Outline and Learning Objectives

1. Causal estimation, in general
   A. What is causal estimation and how does it differ from association and prediction?
   B. What are the challenges with causal estimation with text?

2. Text as causal confounders
   A. For observational data, how does one use back-door adjustment for text as a confounder?

3. Text as causal mediators
   A. For observational data, how does one estimate the natural direct and indirect causal effects with text as a mediator?

Pearl's "Causal Hierarchy"

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?

Pearl's "Causal Hierarchy"

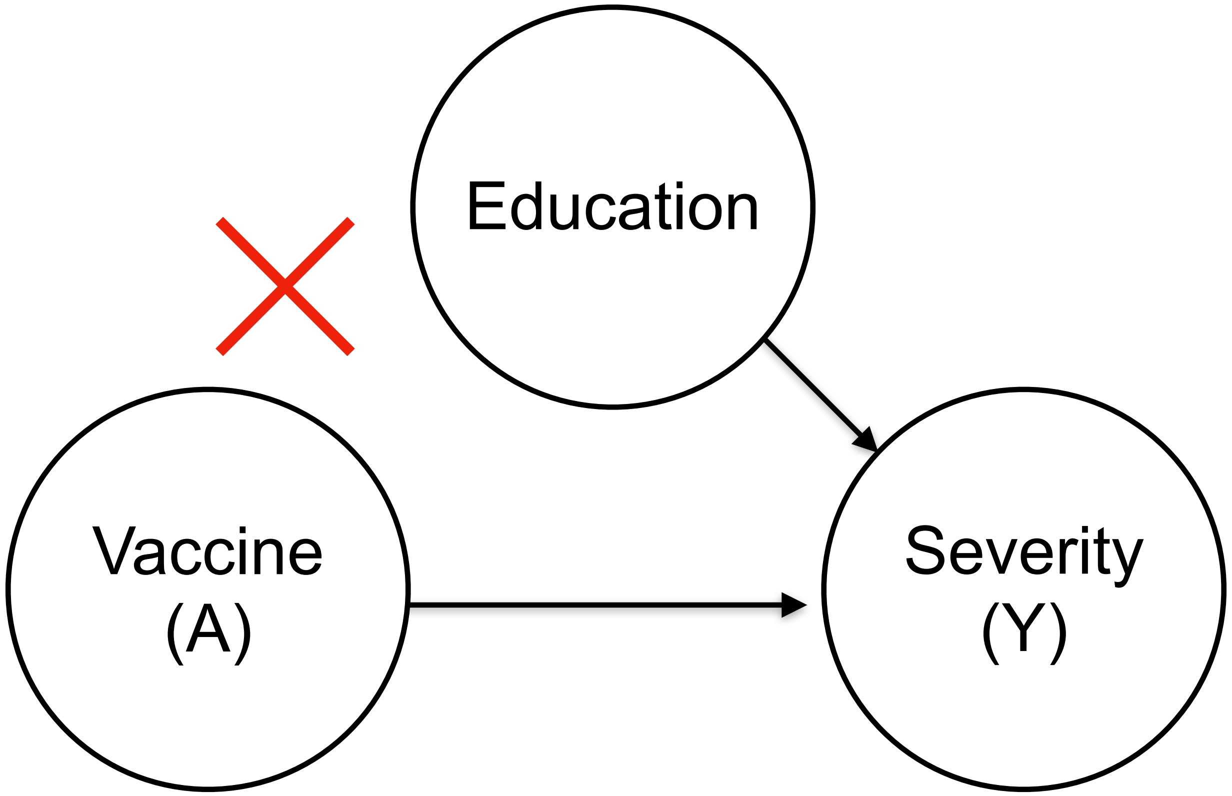
Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?

Pearl's "Causal Hierarchy"

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing *X* change my belief inY?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing, Intervening	What if? What if I do *X*?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it *X* that caused *Y*? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past two years?

Association -> intervention

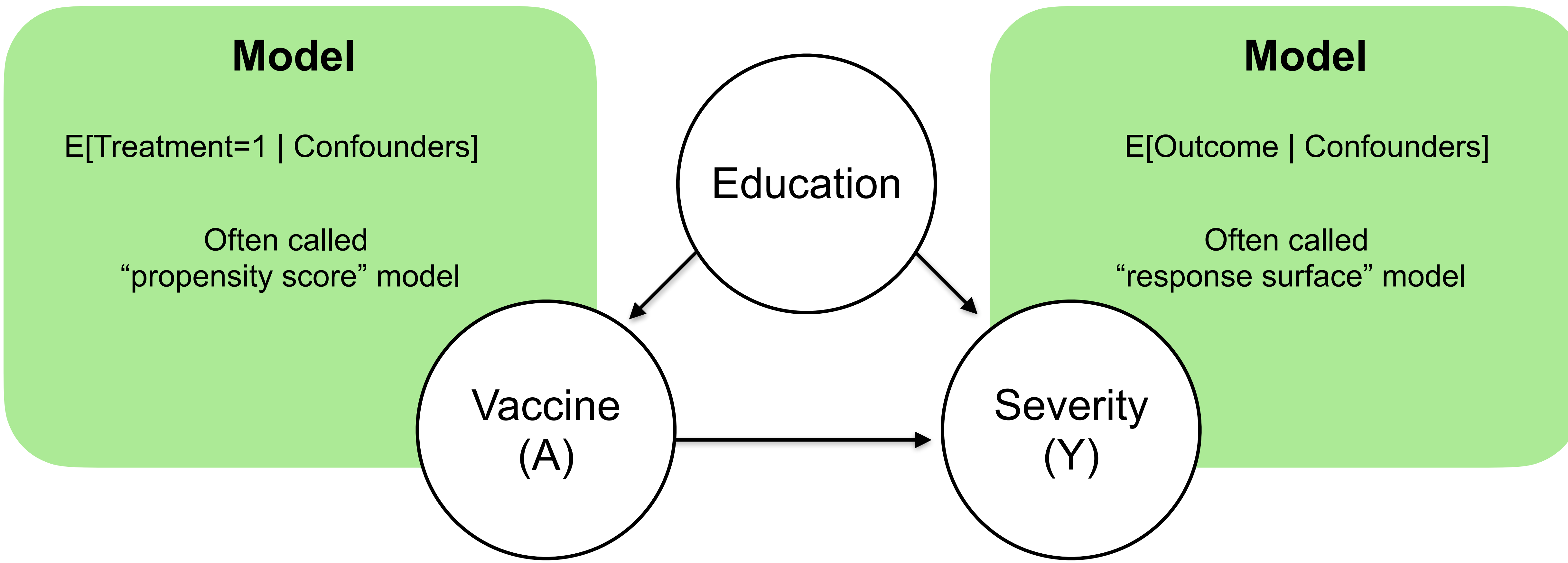
Gold standard for causal inference: run a **randomized control trial (RCT)** in which treatment is randomly assigned



This random assignment breaks the dependence between confounders and treatment

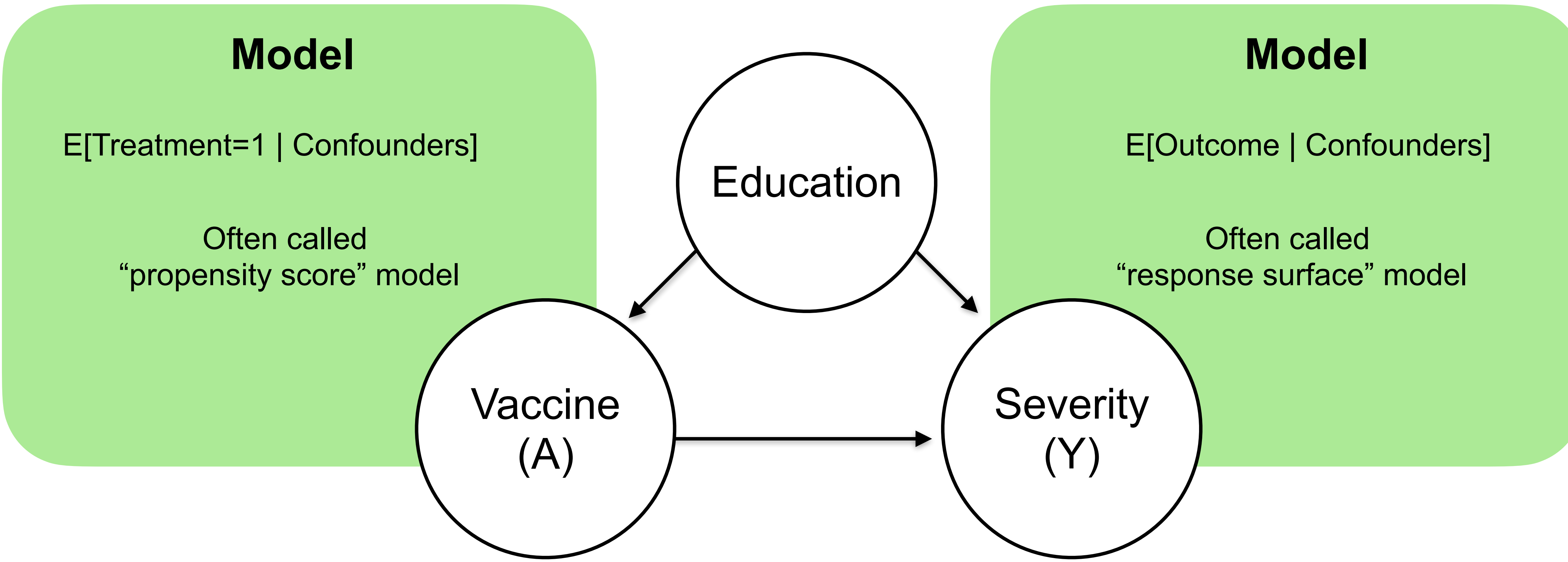
Causal estimation w/o intervention

We’ll need to adjust for the confounders



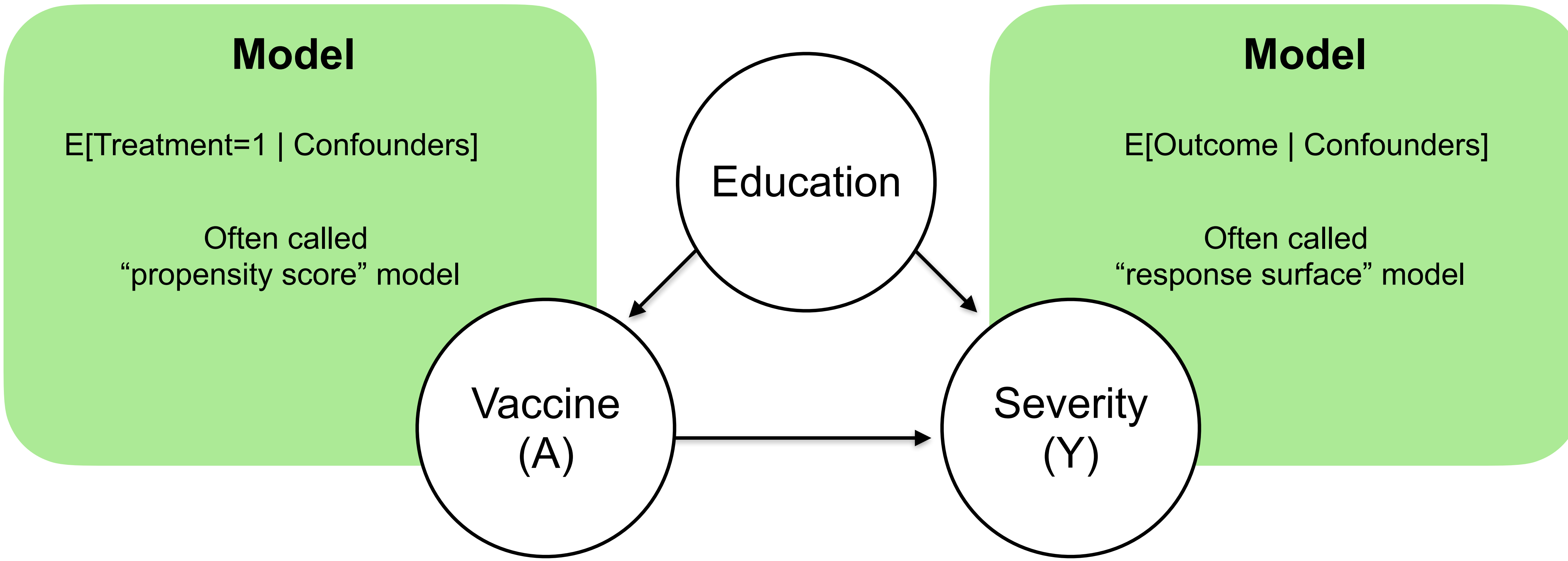
Differences between prediction vs. causal estimation models

We'll need to adjust for the confounders



Differences between prediction vs. causal estimation models

We'll need to adjust for the confounders



- ***Prediction:*** We would like to be able to perfectly separate classes (T=1, T=0) given features (confounders)
- ***Causal estimation:*** We require the assumption of *overlap:* given a set of confounders, we need to have *both* treated and untreated individuals (counterfactuals exist = not classes not perfectly separable)

- ***Prediction:*** Regression with confounders as features
- ***Causal estimation:*** Outcomes are actually *potential outcomes* Y(T=1) and Y(T=0). Need to model both.

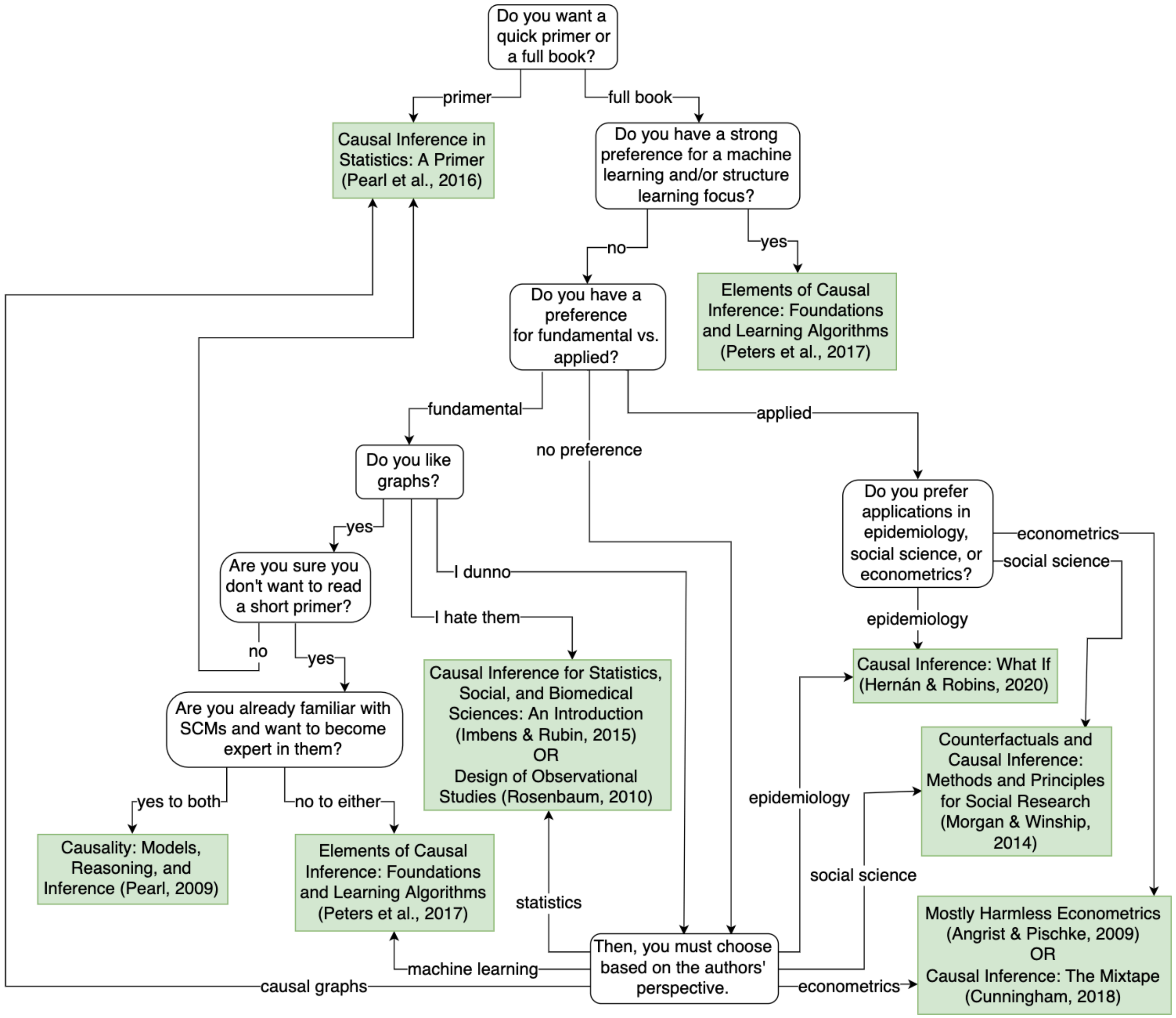
Lecture Outline and Learning Objectives

1. Causal estimation, in general
   A. What is causal estimation and how does it differ from association and prediction?
   B. What are the challenges with causal estimation with text?

2. Text as causal confounders
   A. For observational data, how does one use back-door adjustment for text as a confounder?

3. Text as causal mediators
   A. For observational data, how does one estimate the natural direct and indirect causal effects with text as a mediator?

Many different theories and formalisms for causal inference



Brady Neal flowchart of causal textbooks:
https://www.bradyneal.com/which-causal-inference-book

Challenges of causal estimation + text

Causal estimation = hard, currently being developed, “black-boxy”

(particularly w/ machine learning for high dimensional variables)

+

NLP = hard, currently being developed, “black-boxy”

(particularly representation learning, deep learning)

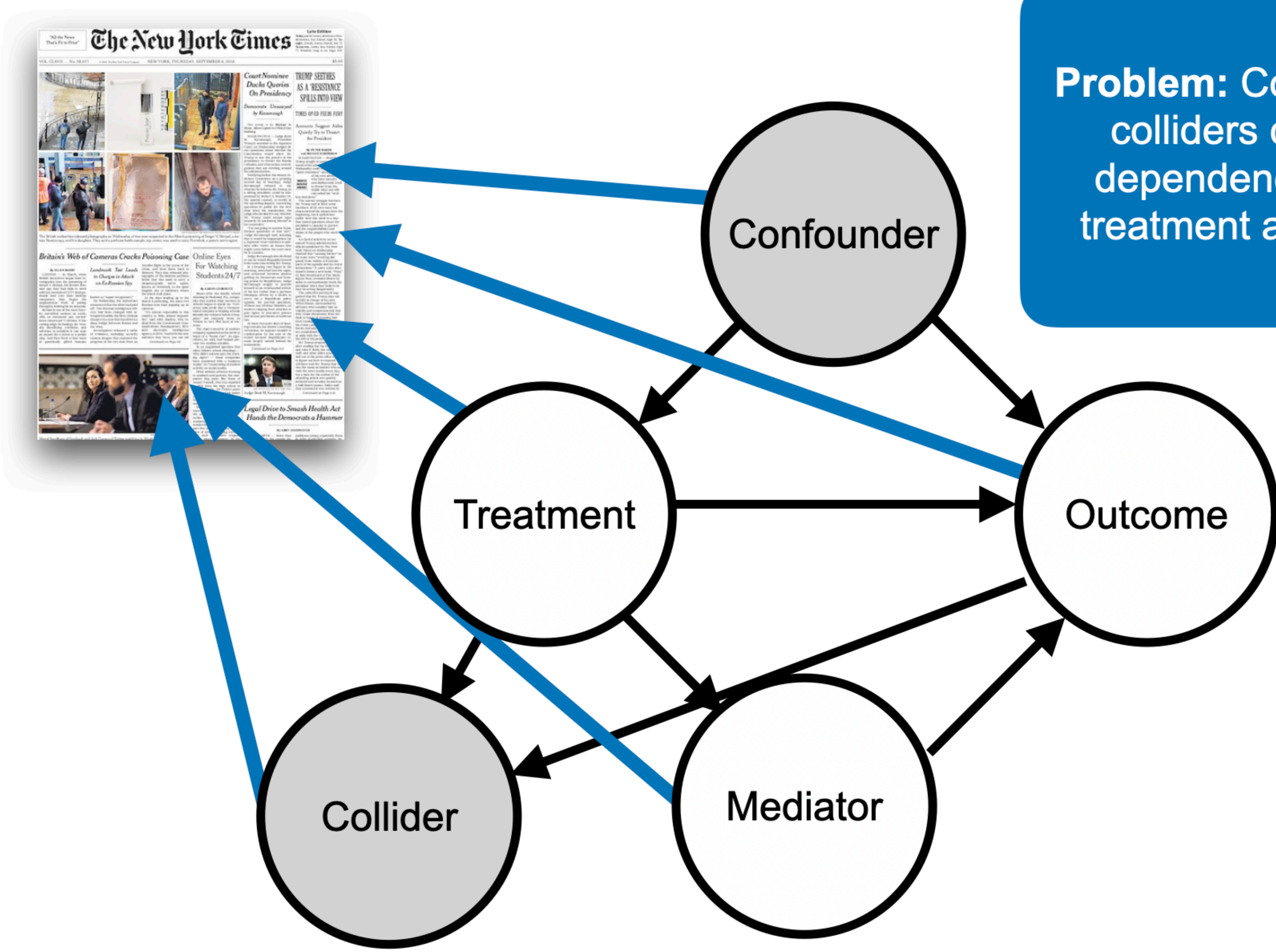
=

HARD!

Challenges of causal estimation + text

> “A classical causal inference question is how does smoking cause cancer. But with text, it's like we have a picture of a cigarette butt outside a person’s house and sometimes we hear them cough and we have to figure out that they smoke from that.”
>
> —Aron Culotta

Challenges of causal estimation + text



Problem: Conditioning on colliders can induce dependence between treatment and outcome

Lecture Outline and Learning Objectives

1. Causal estimation, in general
   A. What is causal estimation and how does it differ from association and prediction?
   B. What are the challenges with causal estimation with text?

2. Text as causal confounders
   A. For observational data, how does one use back-door adjustment for text as a confounder?

3. Text as causal mediators
   A. For observational data, how does one estimate the natural direct and indirect causal effects with text as a mediator?

Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates

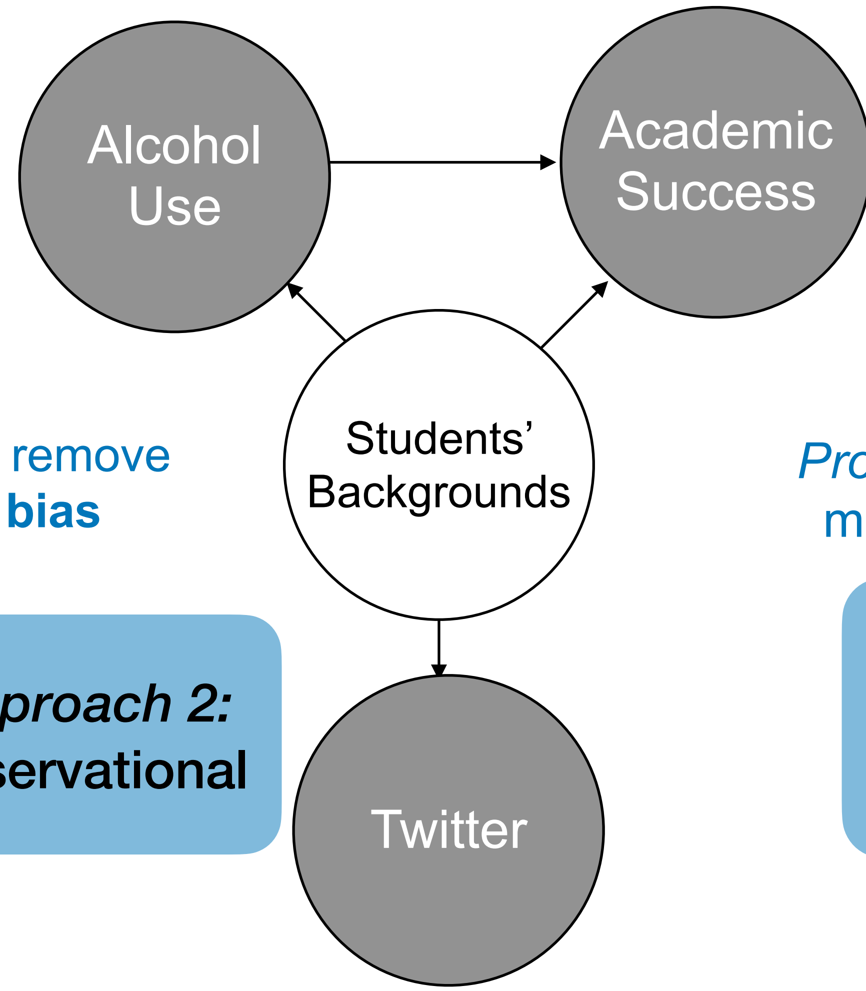


Katherine A. Keith, David Jensen, and Brendan O'Connor

ACL 2020

UMass Amherst

Causal question: For college students, what is the effect of alcohol use on academic success?



Problem: need to remove **confounding bias**

Approach 1: **Intervention**

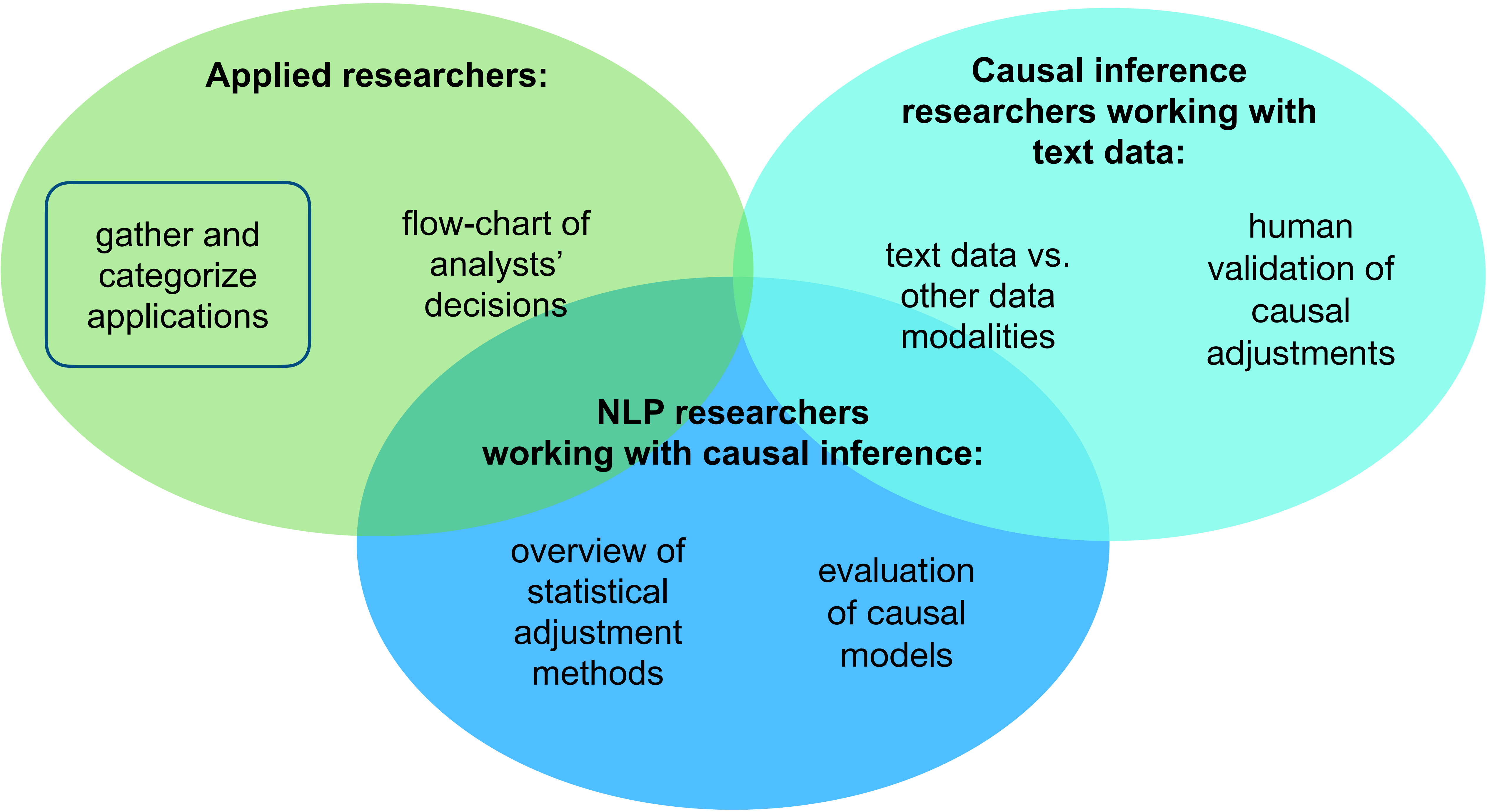
Approach 2: **observational**

Problem: cannot directly measure confounders

One approach: **Use text as a surrogate for**

(Kiciman et al. Using longitudinal social media analysis to understand the effects of early college alcohol use. ICWSM, 2020)

Our contributions for using text to adjust for causal confounding



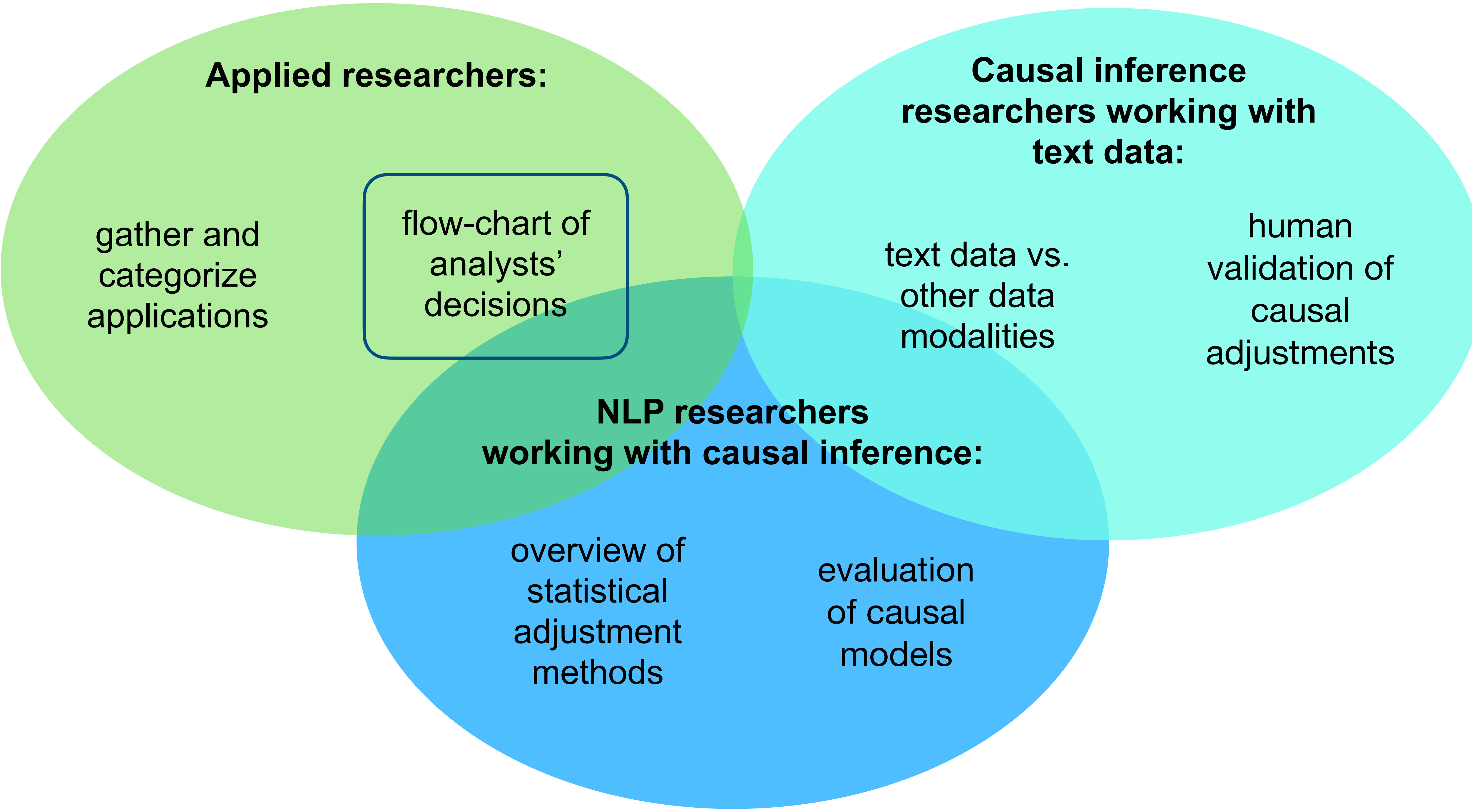
We gather and categorize applications under a common schema

Paper	Treatment	Outcome(s)	Confounder	Text data	Text rep.	Adjustment method
Johansson et al. (2016)	Viewing device (mobile or desktop)	Reader's experience	News content	News	Word counts	Causal-driven rep. learning
De Choudhury et al. (2016)	Word use in mental health community	User transitions to post in suicide community	Previous text written in a forum	Social media (Reddit)	Word counts	Stratified propensity score matching
De Choudhury and Kiciman (2017)	Language of comments	User transitions to post in suicide community	User's previous posts and comments received	Social media (Reddit)	Unigrams and bigrams	Stratified propensity score matching
Falavarjani et al. (2017)	Exercise (Foursquare checkins)	Shift in topical interest on Twitter	Pre-treatment topical interest shift	Social media (Twitter, Foursquare)	Topic models	Matching
Olteanu et al. (2017)	Current word use	Future word use	Past word use	Social media (Twitter)	Top unigrams and bigrams	Stratified propensity score matching
Pham and Shen (2017)	Group vs. individual loan requests	Time until borrowers get funded	Loan description	Microloans (Kiva)	Pre-trained embeddings + neural networks	A-IPTW, TMLE
Kiciman et al. (2018)	Alcohol mentions	College success (e.g. study habits, risky behaviors, emotions)	Previous posts	Social media (Twitter)	Word counts	Stratified propensity score matching
Sridhar et al. (2018)	Exercise	Mood	Mood triggers	Users' text on mood logging apps	Word counts	Propensity score matching
Saha et al. (2019)	Self-reported usage of psychiatric medication	Mood, cognition, depression, anxiety, psychosis, and suicidal ideation	Users' previous posts	Social media (Twitter)	Word counts + lexicons + supervised classifiers	Stratified propensity score matching
Sridhar and Getoor (2019)	Tone of replies	Changes in sentiment	Speaker's political ideology	Debate transcripts	Topic models + lexicons	Regression adjustment, IPTW, A-IPTW
Veitch et al. (2019)	Presence of a theorem	Rate of acceptance	Subject of the article	Scientific articles	BERT	Causal-driven rep. learning + Regression adjustment, TMLE
Roberts et al. (2020)	Perceived gender of author	Number of citations	Content of article	International Relations articles	Topic models + propensity score	Coarsened exact matching
Roberts et al. (2020)	Censorship	Subsequent censorship and posting rate	Content of posts	Social media (Weibo)	Topic models + propensity score	Coarsened exact matching

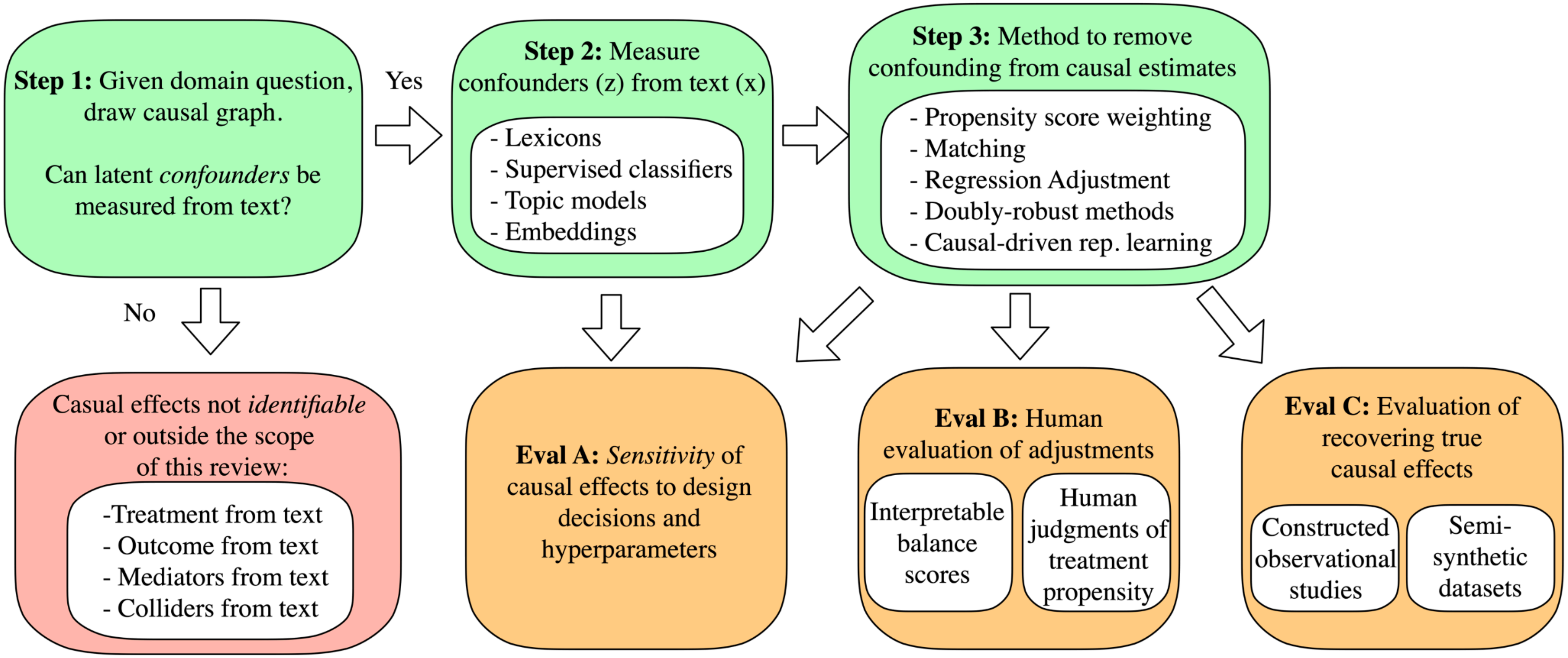
<u>Scattered publication venues:</u> ICML, IJCAI, ICWSM, CHI, CSCW, AJPS

over 8 different causal adjustment methods

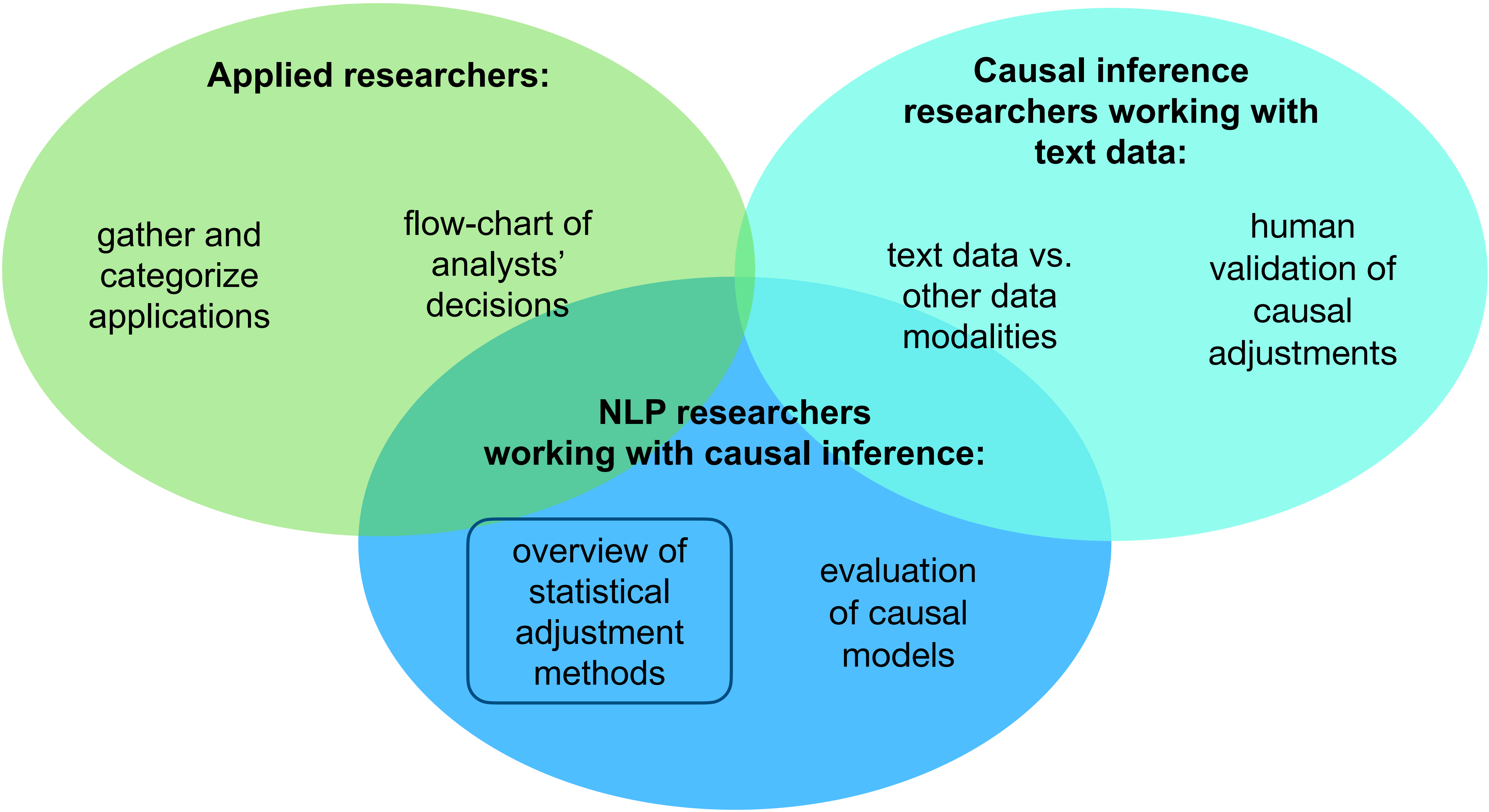
Our contributions for using text to adjust for causal confounding



Flow chart of analysts' decisions



Our contributions for using text to adjust for causal confounding



Adjustment method: matching

1. Define matching criterion: (1) text representation, (2) distance metric, (3) matching algorithm

Example: (1) BERT embeddings, (2) cosine sim. > 0.8, (3) 1-to-many matches

2. Estimate counterfactuals from matches

$$\hat{y}_i(k) = \begin{cases} y_i & \text{if } t_i = k \\ \frac{1}{|\mathcal{M}_i|} \sum_{j \in \mathcal{M}_i} y_j & \text{if } t_i \neq k \end{cases}$$

3. Plug-in matching estimators

$$\hat{\tau}_{\text{match}} = \frac{1}{n} \sum_{i}^{n} \Big(\hat{y}_i(1) - \hat{y}_i(0) \Big)$$

Adjustment method: outcome regression

1. Fit a supervised model on expected outcomes

$$q(t, z) \equiv \mathbb{E}(Y \mid T = t, Z = z)$$

2. Use the learned outcome to predict counterfactuals

$$\hat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i}^{n} \left(\hat{q}(1, z_i) - \hat{q}(0, z_i) \right)$$

Adjustment method (newer): causally-driven representation learning

Encoding layers = BERT
Veitch et al. 2021

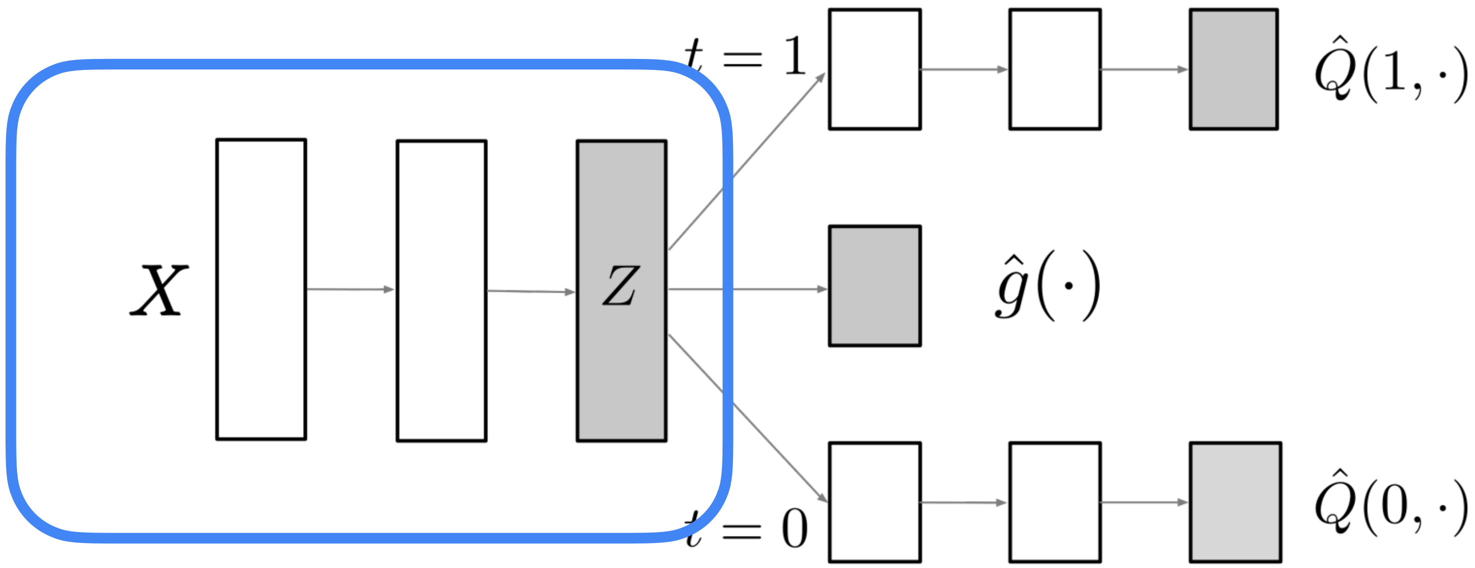
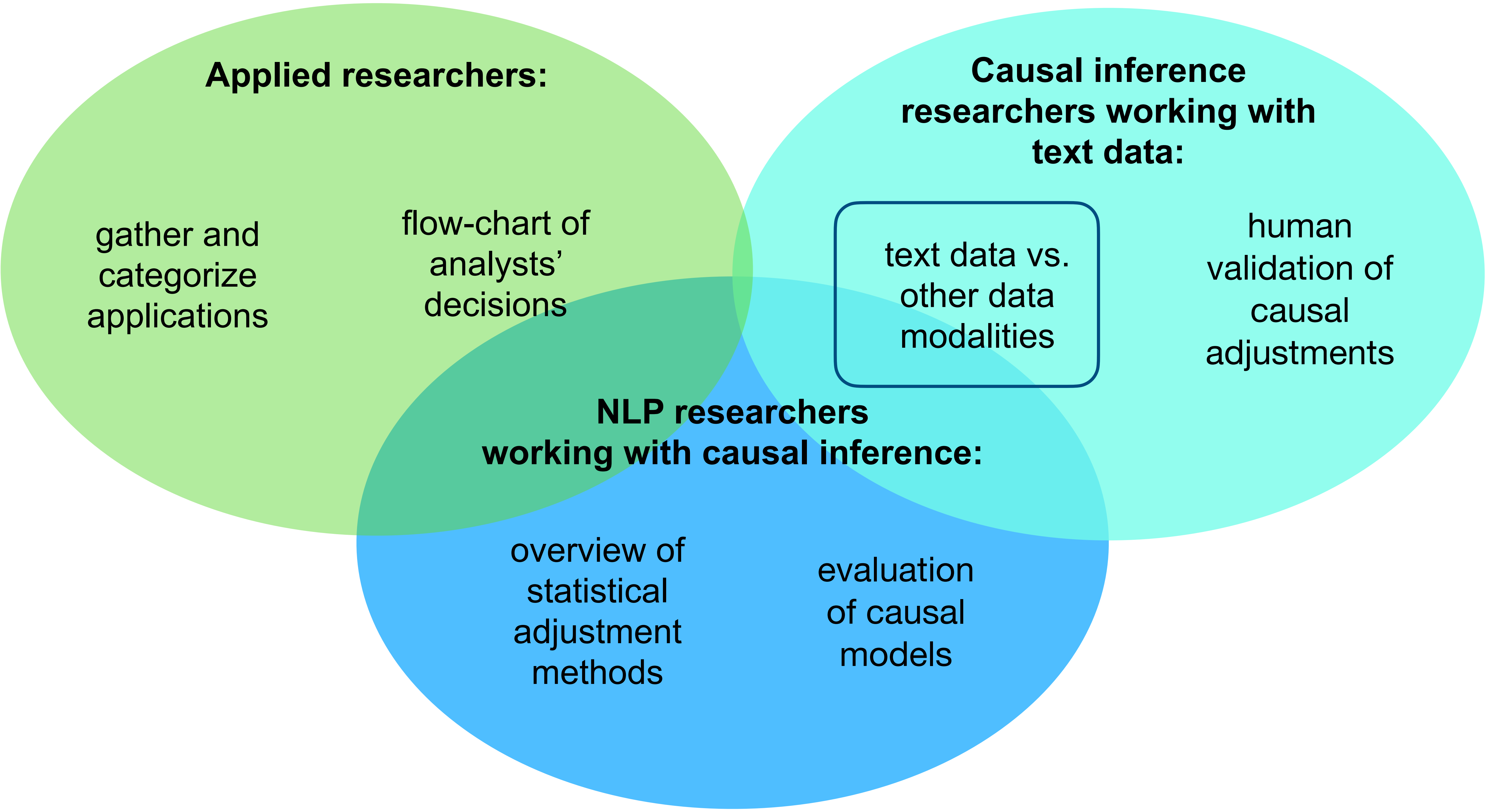


Figure 1: Dragonnet architecture.

Dragonnet: *Shi et al. Adapting Neural Networks for the Estimation of Treatment Effects. NeurIPS, 2019.*
CausalBERT: *Veitch et al. Adapting text embeddings for causal inference. UAI, 2020.*

Our contributions for using text to adjust for causal confounding



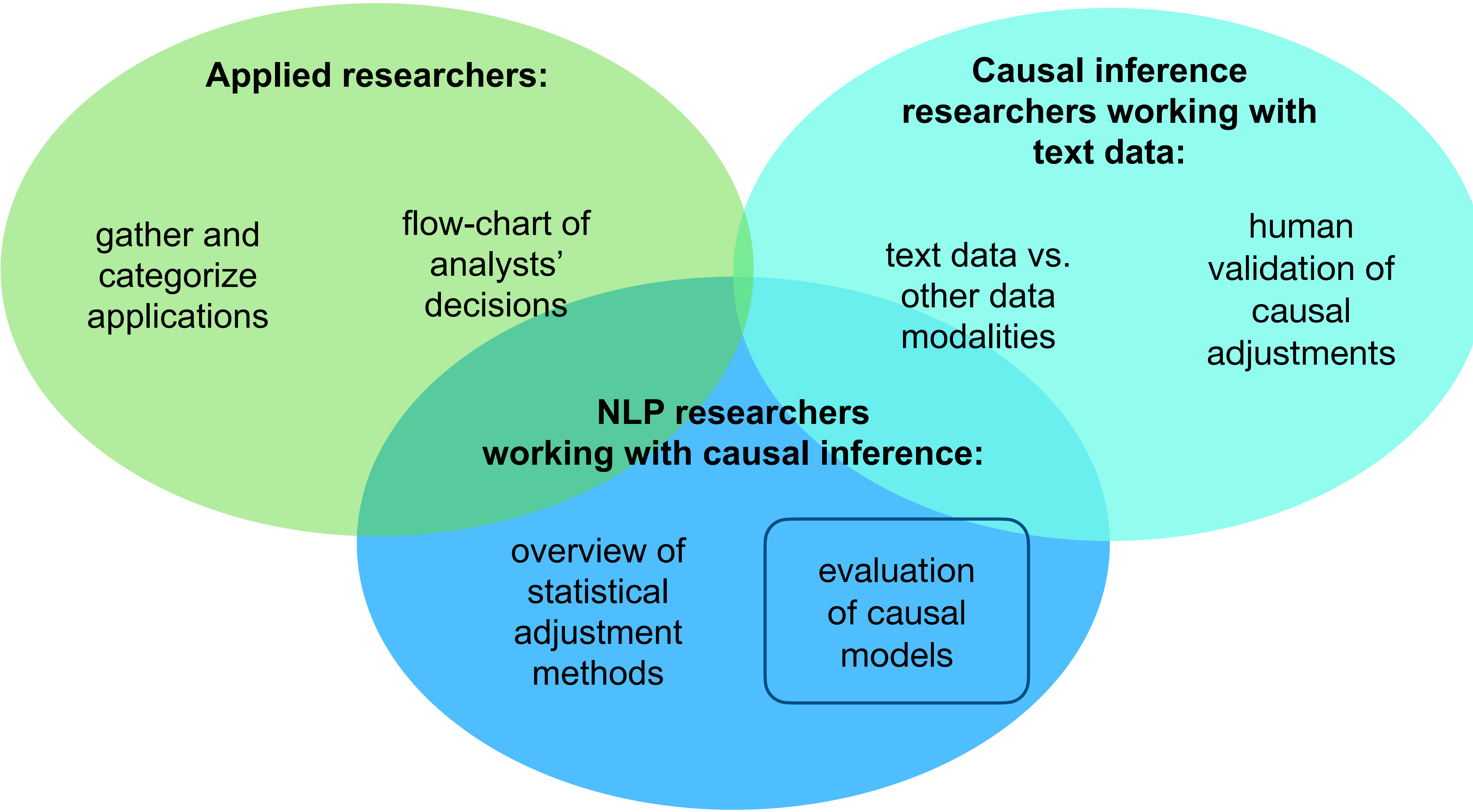
Text-specific open problems for causal inference

- When does text **encode multiple variables** simultaneously (e.g. confounders and colliders) and how do we adjust in these settings?

- Since text is high-dimensional, violation of the assumption of **overlap**, $0<Pr(T|X)<1$ *for all* X, is very likely. When does this occur and what do we do?

- Because text is interpretable, how can one systematically use **human judgements** to evaluate intermediate causal adjustment steps?

Our contributions for using text to adjust for causal confounding



Open problem: evaluating text-based causal methods

Problem Type	Evaluation
Predictive	Predictive performance (e.g. accuracy) on a held-out test set
Causal	Estimated vs. true causal effects

Difficult to obtain!

Open problem: evaluating text-based causal methods

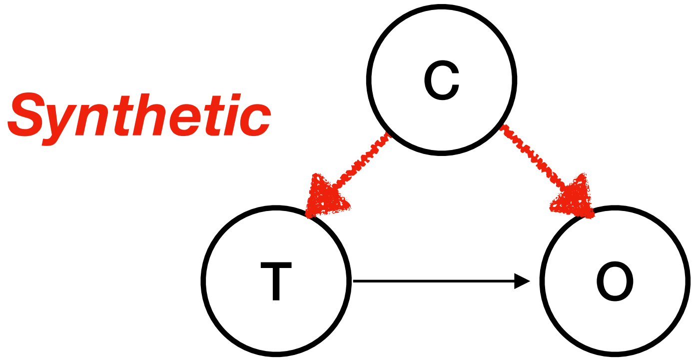
(A) Constructed observational studies

Randomized

Non-randomized

In other social sciences:
(LaLonde (1986); Shadish et al. (2008); Glynn and Kashin (2013))

(B) Semi-synthetic datasets



With **text** to remove confounding:
(Johansson et al. 2016; Veitch et al. 2019; Roberts et al. 2020)

Our contributions for using text to adjust for causal confounding

Applied researchers:

gather and categorize applications

flow-chart of analysts' decisions

Causal inference researchers working with text data:

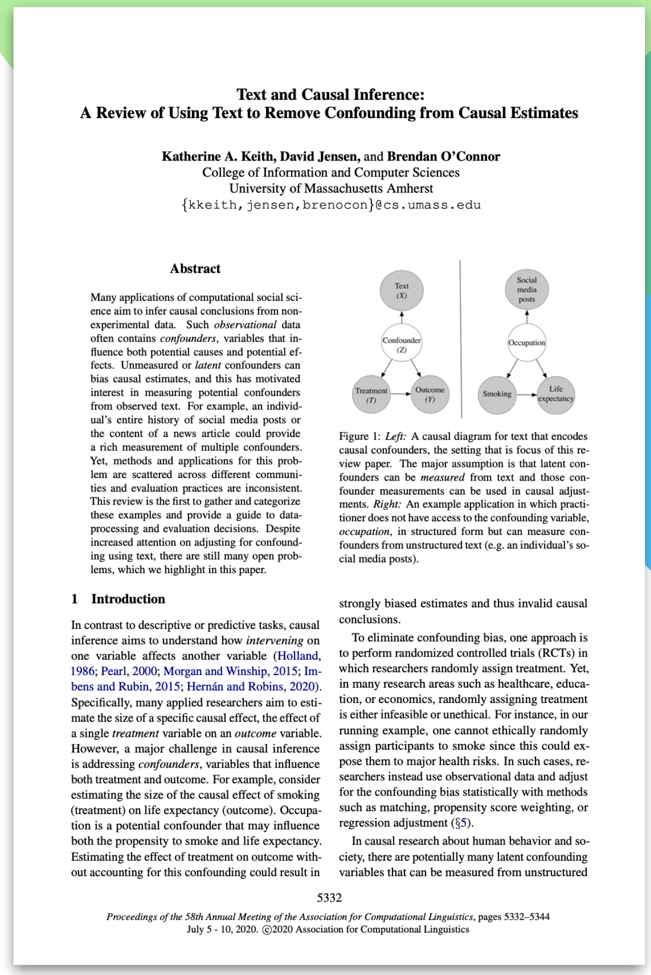
text data vs. other data modalities

human validation of causal adjustments

NLP researchers working with causal inference:

overview of statistical adjustment methods

evaluation of causal models



Text and Causal Inference:
A Review of Using Text to Remove Confounding from Causal Estimates

Katherine A. Keith, David Jensen, and **Brendan O'Connor**
College of Information and Computer Sciences
University of Massachusetts Amherst
{kkeith, jensen, brenocon}@cs.umass.edu

Abstract

Many applications of computational social science aim to infer causal conclusions from non-experimental data. Such *observational* data often contains *confounders*, variables that influence both potential causes and potential effects. Unmeasured or *latent* confounders can bias causal estimates, and this has motivated interest in measuring potential confounders from observed text. For example, an individual's entire history of social media posts or the content of a news article could provide a rich measurement of multiple confounders. Yet, methods and applications for this problem are scattered across different communities and evaluation practices are inconsistent. This review is the first to gather and categorize these examples and provide a guide to data-processing and evaluation decisions. Despite increased attention on adjusting for confounding using text, there are still many open problems, which we highlight in this paper.

Figure 1: *Left:* A causal diagram for text that encodes causal confounders, the setting that is focus of this review paper. The major assumption is that latent confounders can be *measured* from text and those confounder measurements can be used in causal adjustments. *Right:* An example application in which practitioner does not have access to the confounding variable, *occupation*, in structured form but can measure confounders from unstructured text (e.g. an individual's social media posts).

1 Introduction

In contrast to descriptive or predictive tasks, causal inference aims to understand how *intervening* on one variable affects another variable (Holland, 1986; Pearl, 2000; Morgan and Winship, 2015; Imbens and Rubin, 2015; Hernán and Robins, 2020). Specifically, many applied researchers aim to estimate the size of a specific causal effect, the effect of a single *treatment* variable on an *outcome* variable. However, a major challenge in causal inference is addressing *confounders*, variables that influence both treatment and outcome. For example, consider estimating the size of the causal effect of smoking (treatment) on life expectancy (outcome). Occupation is a potential confounder that may influence both the propensity to smoke and life expectancy. Estimating the effect of treatment on outcome without accounting for this confounding could result in strongly biased estimates and thus invalid causal conclusions.

To eliminate confounding bias, one approach is to perform randomized controlled trials (RCTs) in which researchers randomly assign treatment. Yet, in many research areas such as healthcare, education, or economics, randomly assigning treatment is either infeasible or unethical. For instance, in our running example, one cannot ethically randomly assign participants to smoke since this could expose them to major health risks. In such cases, researchers instead use observational data and adjust for the confounding bias statistically with methods such as matching, propensity score weighting, or regression adjustment (§5).

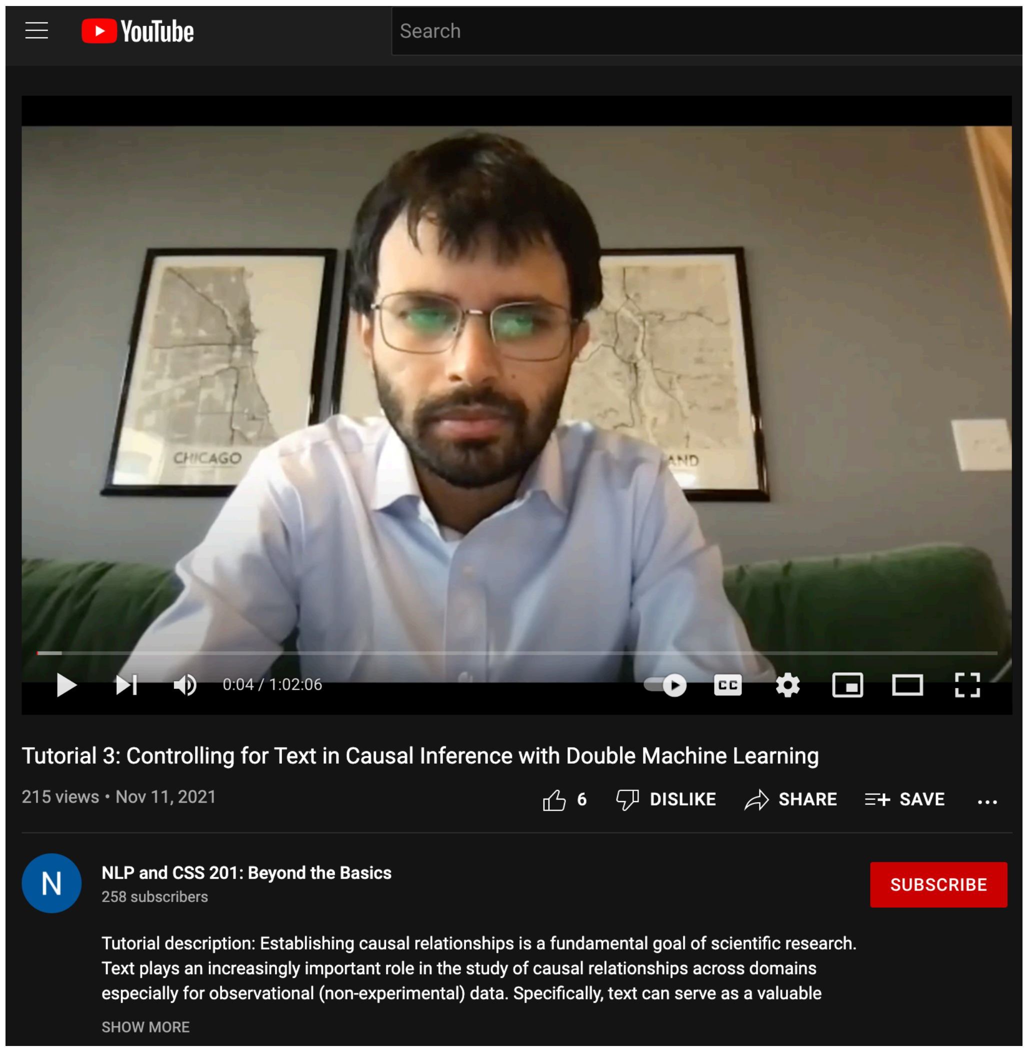
In causal research about human behavior and society, there are potentially many latent confounding variables that can be measured from unstructured

5332

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5332–5344
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

Check-out our paper for more details!

Other resources for text as confounder



Emaad Manzoor's tutorial

Controlling for Text in Causal Inference with Double Machine Learning	Establishing causal relationships is a fundamental goal of scientific research. Text plays an increasingly important role in the study of causal relationships across domains especially for observational (non-experimental) data. Specifically, text can serve as a valuable "control" to eliminate the effects of variables that threaten the validity of the causal inference process. But how does one control for text, an unstructured and nebulous quantity? In this tutorial, we will learn about bias from confounding, motivation for using text as a proxy for confounders, apply a "double machine learning" framework that uses text to remove confounding bias, and compare this framework with non-causal text dimensionality reduction alternatives such as topic modeling.	Emaad Manzoor	Code; Video; Slides

https://nlp-css-201-tutorials.github.io/nlp-css-201-tutorials/

Lecture Outline and Learning Objectives

1. Causal estimation, in general
   A. What is causal estimation and how does it differ from association and prediction?
   B. What are the challenges with causal estimation with text?

2. Text as causal confounders
   A. For observational data, how does one use back-door adjustment for text as a confounder?

3. Text as causal mediators
   A. For observational data, how does one estimate the natural direct and indirect causal effects with text as a mediator?

Text as Causal Mediators: Research Design for Causal Estimates of Differential Treatment of Social Groups via Language Aspects

Katherine A. Keith, Douglas Rice, and Brendan O'Connor

CI+NLP Workshop, EMNLP 2021

UMass Amherst

Bias in interruptions during U.S. Supreme Court oral arguments

Q: Why do some justices interrupt female advocates more than male advocates?

(Patton & Smith, "Lawyer, Interrupted: Gender Bias in Oral Arguments at the U.S. Supreme Court," *Journal of Law and Courts*, 2017)

(Jacobi and Schweers. "Justice, interrupted: The effect of gender, ideology, and seniority at Supreme Court oral arguments." Va. L. Rev, 2017)

Importance of interruptions as causal outcome

- Interruptions => status reinforcement (Mendelberg et al., 2014)
- Justices' oral argument behavior <=> case outcomes (Johnson et al., 2006)
- Timely and relevant

The New York Times

SIDEBAR

Supreme Court Tries to Tame Unruly Oral Arguments

The court, which is hearing major cases on abortion and guns, has revised its procedures to make sure that all justices are heard.

Chief Justice John G. Roberts Jr. made changes to the court after a 2017 study showed female justices were disproportionately interrupted by male colleagues and lawyers.
Stefani Reynolds for The New York Times

By Adam Liptak

Nov. 1, 2021

WASHINGTON — Justices Sonia Sotomayor and Clarence Thomas

Bias in interruptions during U.S. Supreme Court oral arguments

Q: Why do some justices interrupt female advocates more than male advocates?

Legal analysts

Explanation 1: Implicit gender bias

Explanation 2: Women are "less effective" advocates

- Different types of clients with weaker legal arguments
- Decreased quality of the argument
- Manner of speaking

(Patton & Smith, "Lawyer, Interrupted: Gender Bias in Oral Arguments at the U.S. Supreme Court," *Journal of Law and Courts*, 2017)

(Jacobi and Schweers. "Justice, interrupted: The effect of gender, ideology, and seniority at Supreme Court oral arguments." Va. L. Rev, 2017)

Example

Lozano v. Montoya Alvarez (2013)

Audio Source: Oyez

Ann O'Connell Adams (advocate):


(Photo Credit: LinkedIn)

Well—

Antonin Scalia (justice):

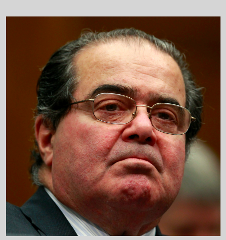

(Photo Credit: Brookings Institute)

I mean, it seems to me it just makes that article impossible to apply consistently country to country.

Ann O'Connell Adams (advocate):

No, I don't think so. And—and, the other signatories have—have almost all, I mean I think the Hong Kong court does say that it doesn't have discretion, but [...] the other courts of signatory countries that have interpreted Article 12 have all found a discretion, whether it be in Article 12 or in Article 8.—

Antonin Scalia (justice):



Have they exercised it? Have they exercised it, that discretion which they say is there?

Example

Lozano v. Montoya Alvarez (2013)

Ann O'Connell Adams (advocate):

(Photo Credit: LinkedIn)

Well—

Interruption

Antonin Scalia (justice):

(Photo Credit: Brookings Institute)

I mean, it seems to me it just makes that article impossible to apply consistently country to country.

Hedging

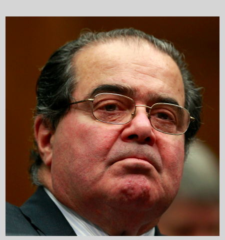
Speech Disfluencies

Ann O'Connell Adams (advocate):

No, I don't think so. And—and, the other signatories have—have almost all, I mean I think the Hong Kong court does say that it doesn't have discretion, but […] the other courts of signatory countries that have interpreted Article 12 have all found a discretion, whether it be in Article 12 or in Article 8.—

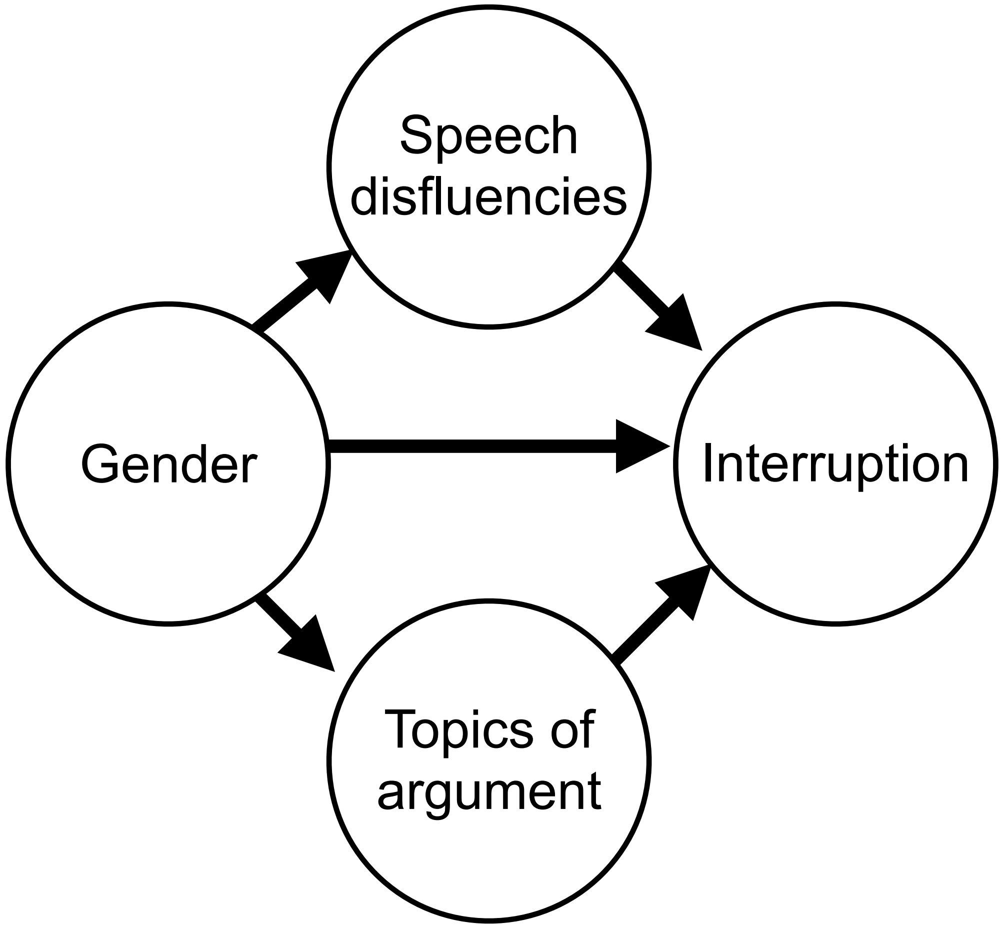
Interruption

Antonin Scalia (justice):

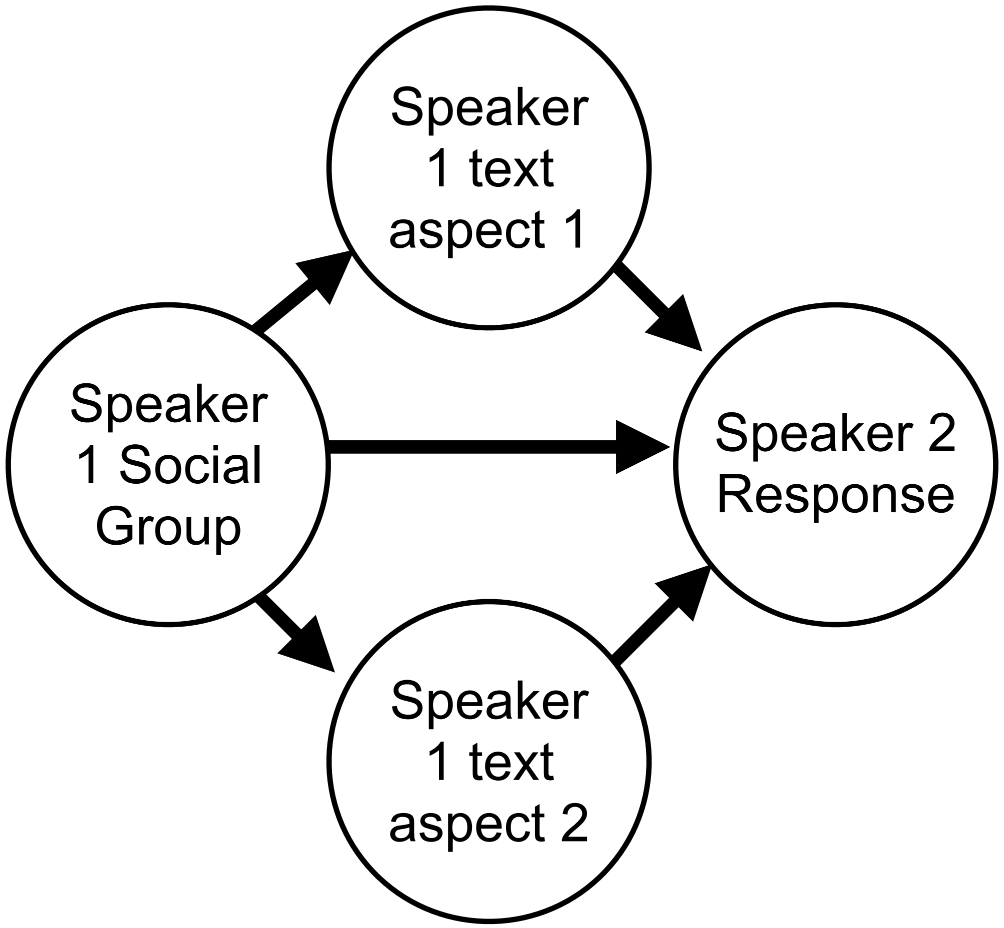


Have they exercised it? Have they exercised it, that discretion which they say is there?

Causal DAG, U.S. Supreme Court



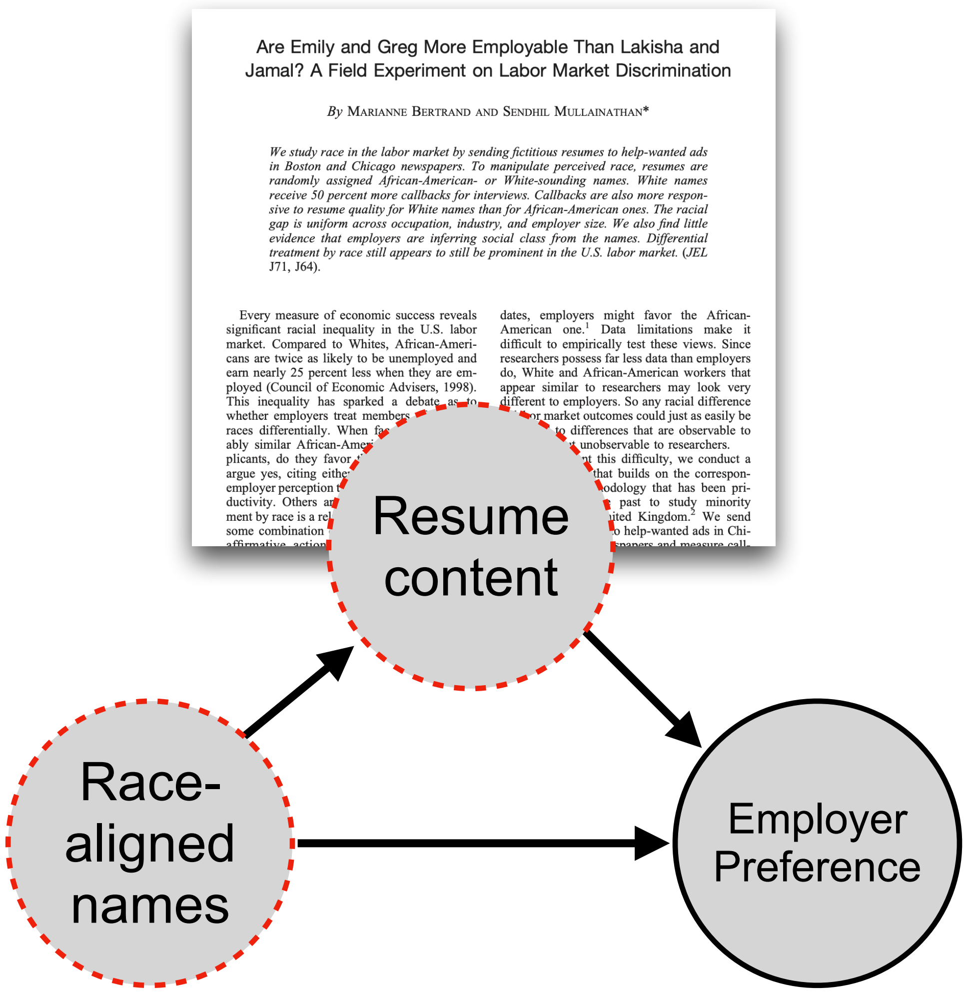
Causal DAG, General Framework



Causal experiments (audit studies) with bias + text

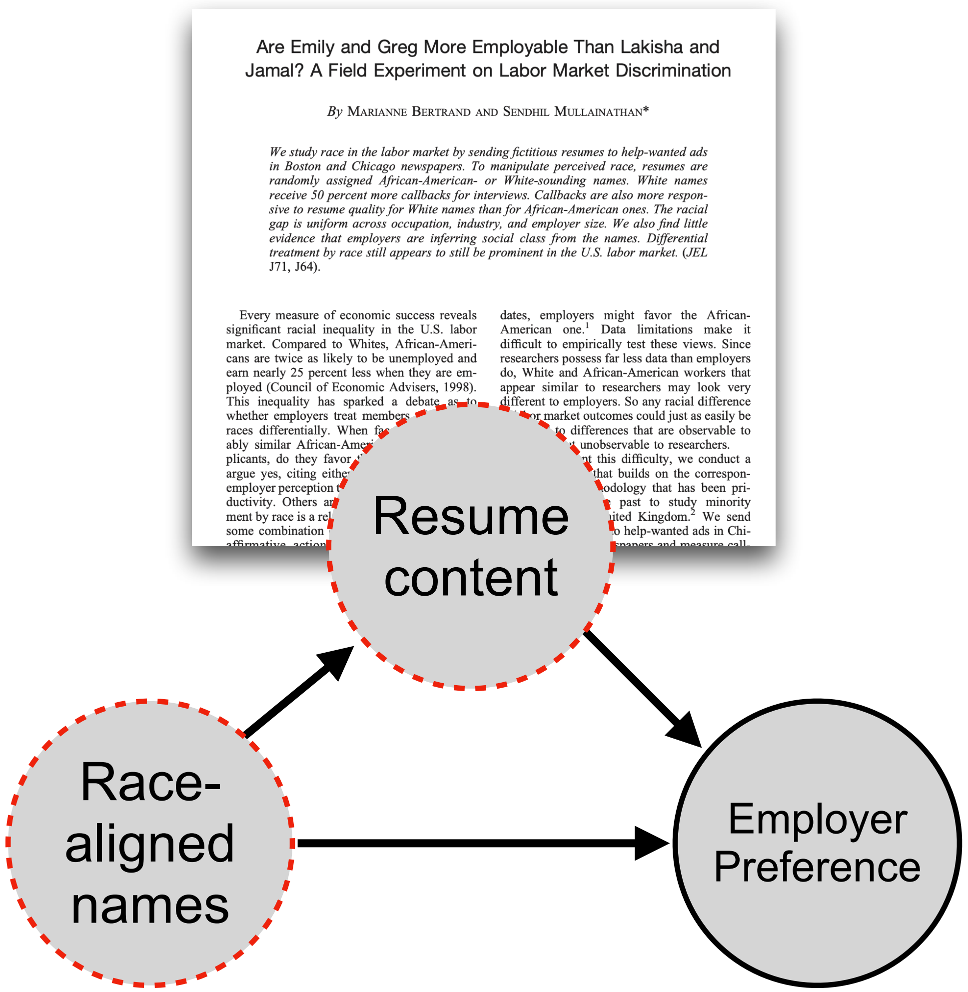
Causal experiments (audit studies) with bias + text

Resumes

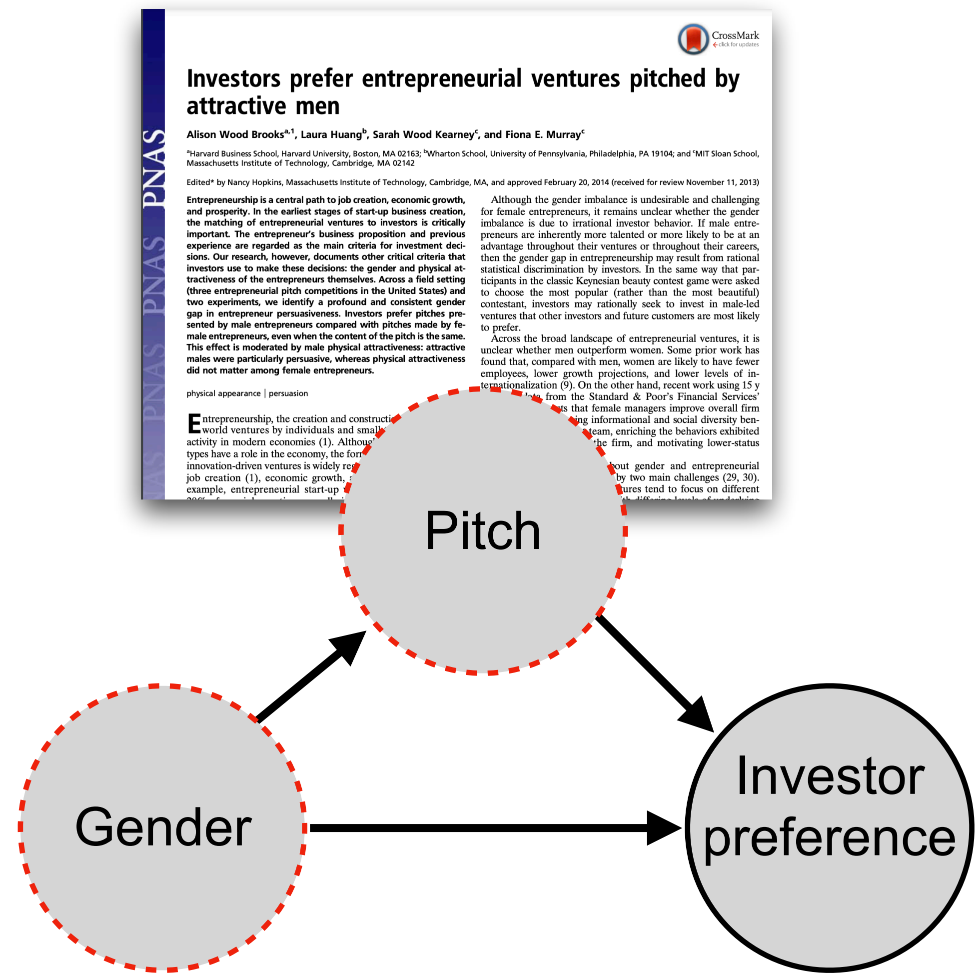


Causal experiments (audit studies) with bias + text

Resumes

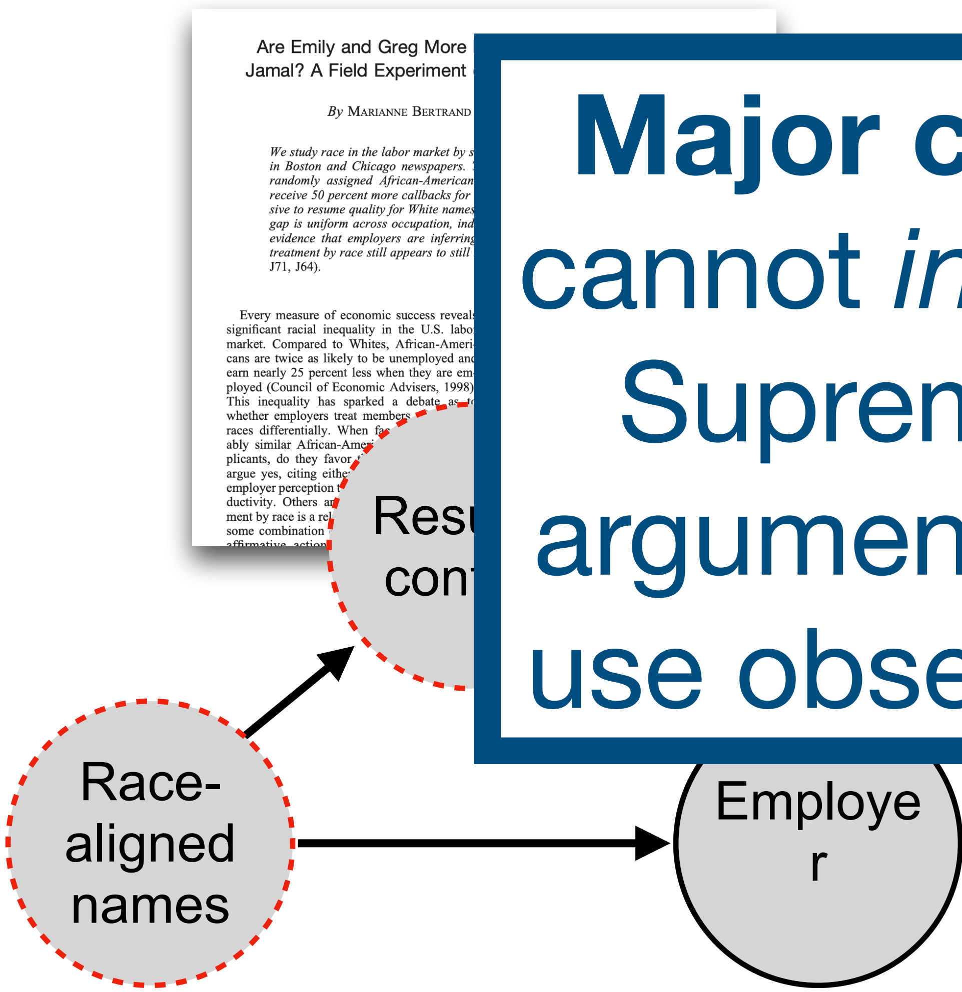


Entrepreneurial pitches

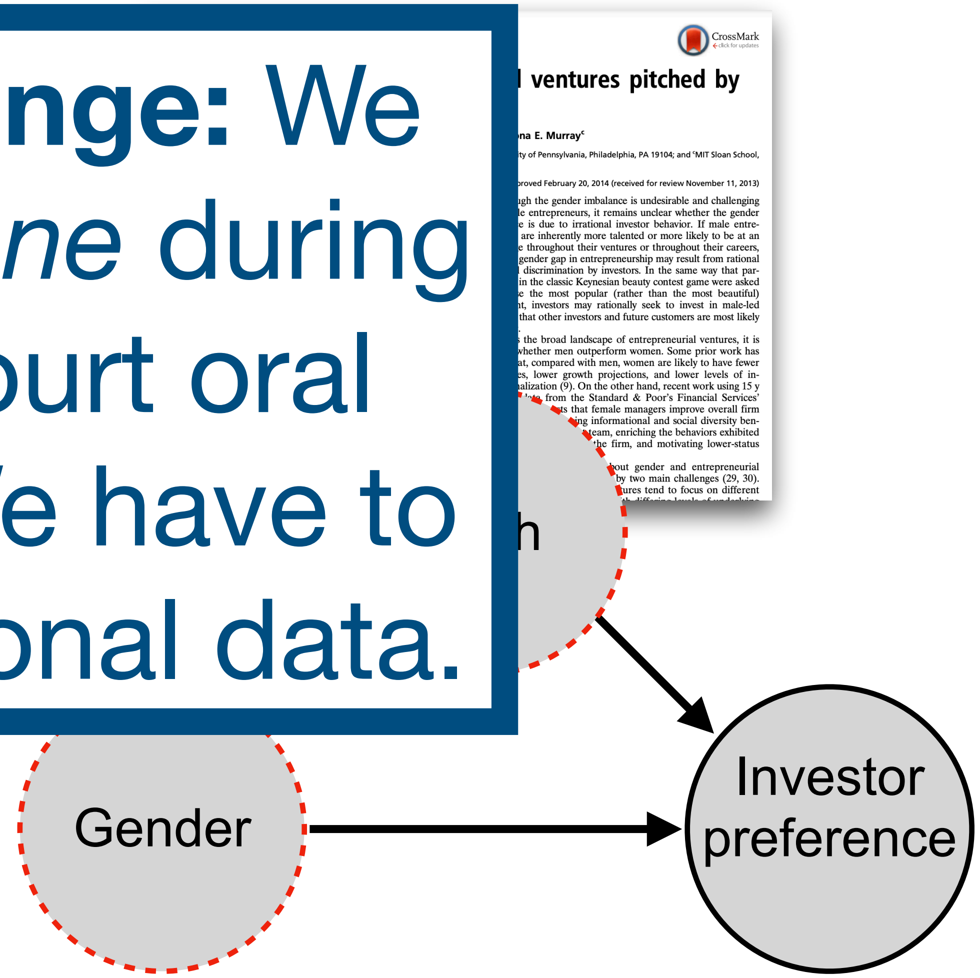


Causal experiments (audit studies) with bias + text

Resumes



Entrepreneurial pitches

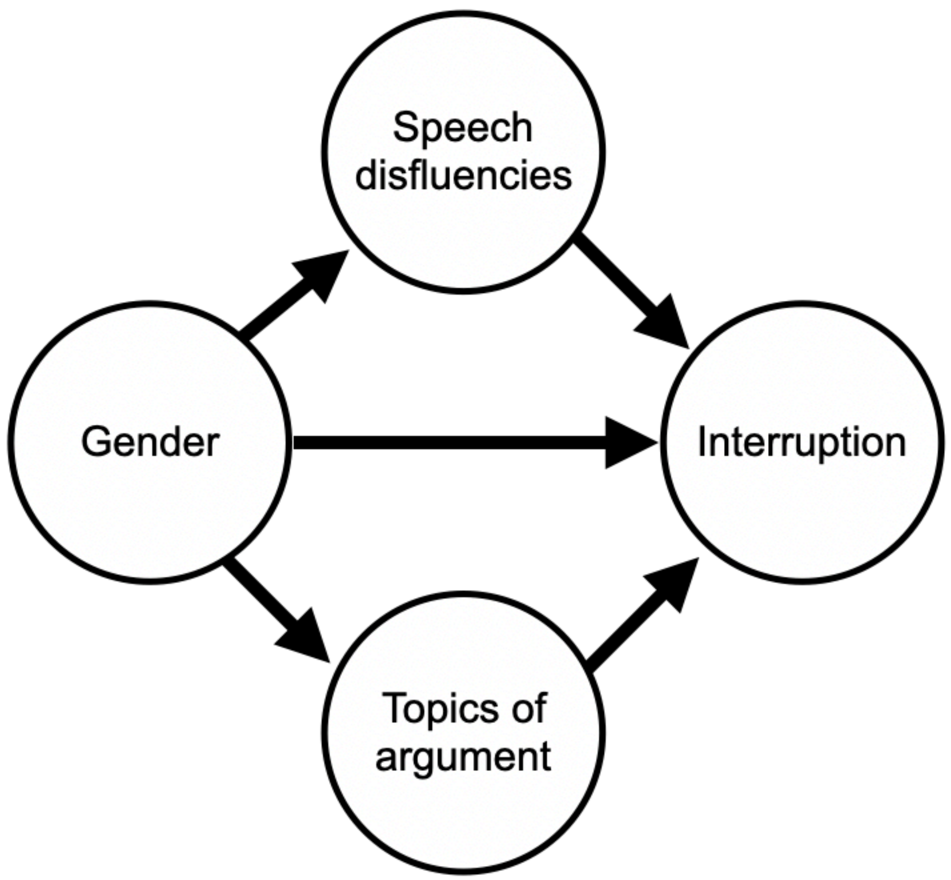


Major challenge: We cannot *intervene* during Supreme Court oral arguments. We have to use observational data.

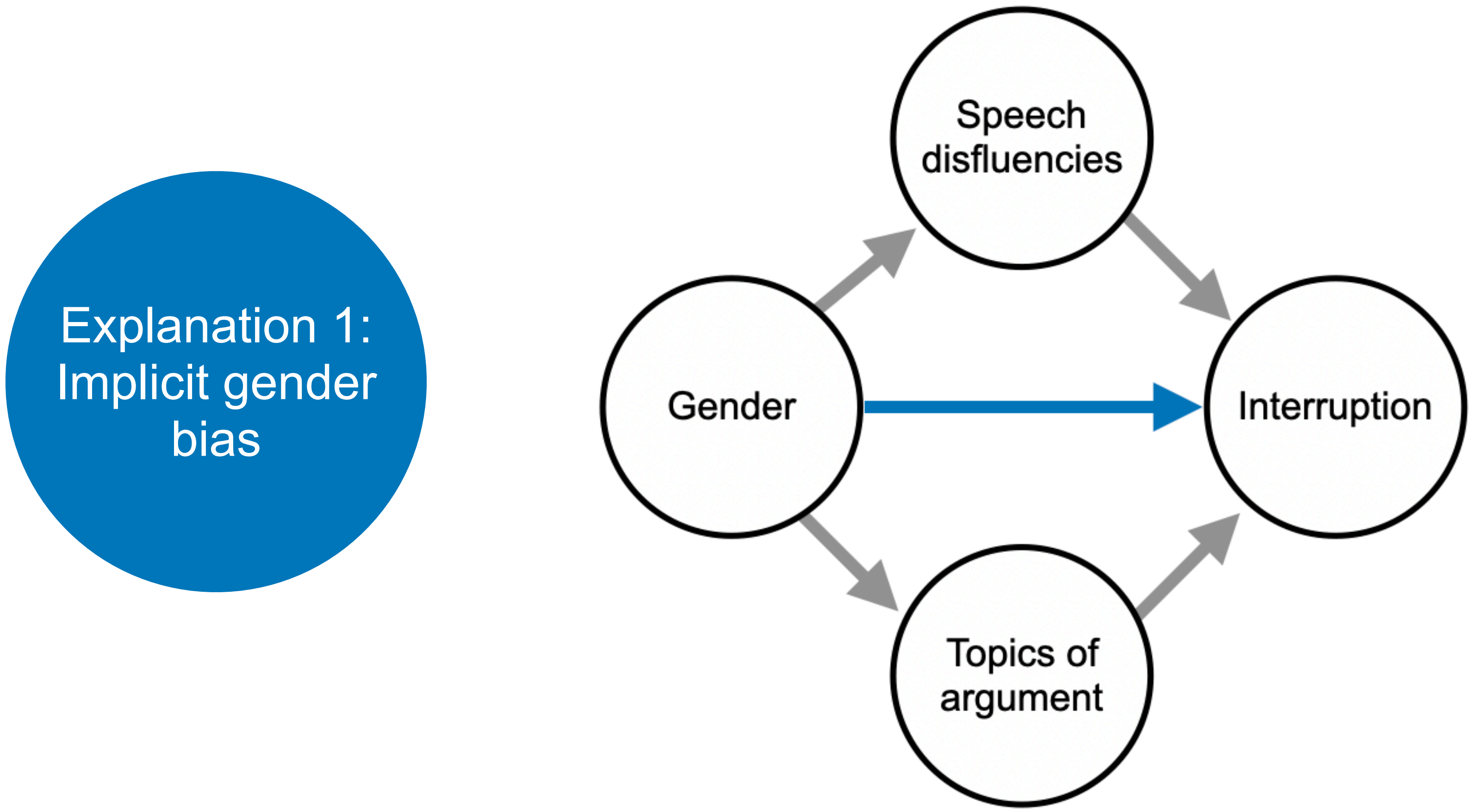
Contributions & past work context

- Intentionally focusing on a thoughtful **causal design** before we obtain empirical results
    - "Design trumps analysis" (Rubin, 2008)
    - We will only every have observational data for the U.S. Supreme Court
- We use **causal mediation analysis** towards the goal of splitting the total effect into the portion of the effect that goes through language mediators and the portion that does not
    - General causal mediation analysis: (Pearl, 2001; Imai et al., 2010; VanderWeele, 2016)
    - Other text and mediation work: (Tierney & Volfovsky, 2021)
- Illustrate the **challenges** conceptualizing and operationalizing causal variables
    - Criticisms of claiming "gender" or "race" as a causal treatments (Sen & Wasow, 2016; Hu & Kohler-Hausmann, 2020)
    - Difficult to choose which language aspects to choose as mediators (e.g. Pryzant et al., 2021 with text as treatment)

Causal DAG, U.S. Supreme Court

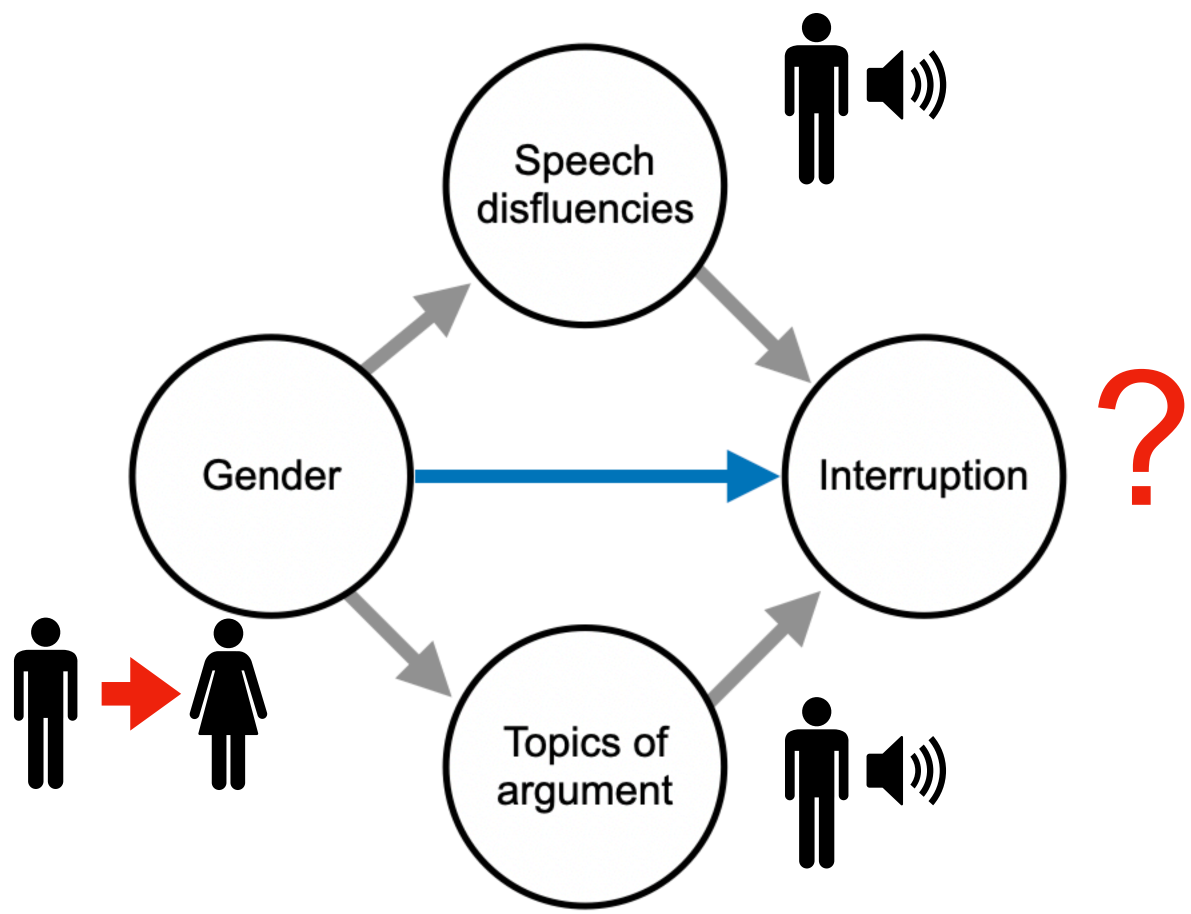


Explanation 1 corresponds to the direct path



Explanation 1 corresponds to the direct path

Explanation 1: Implicit gender bias



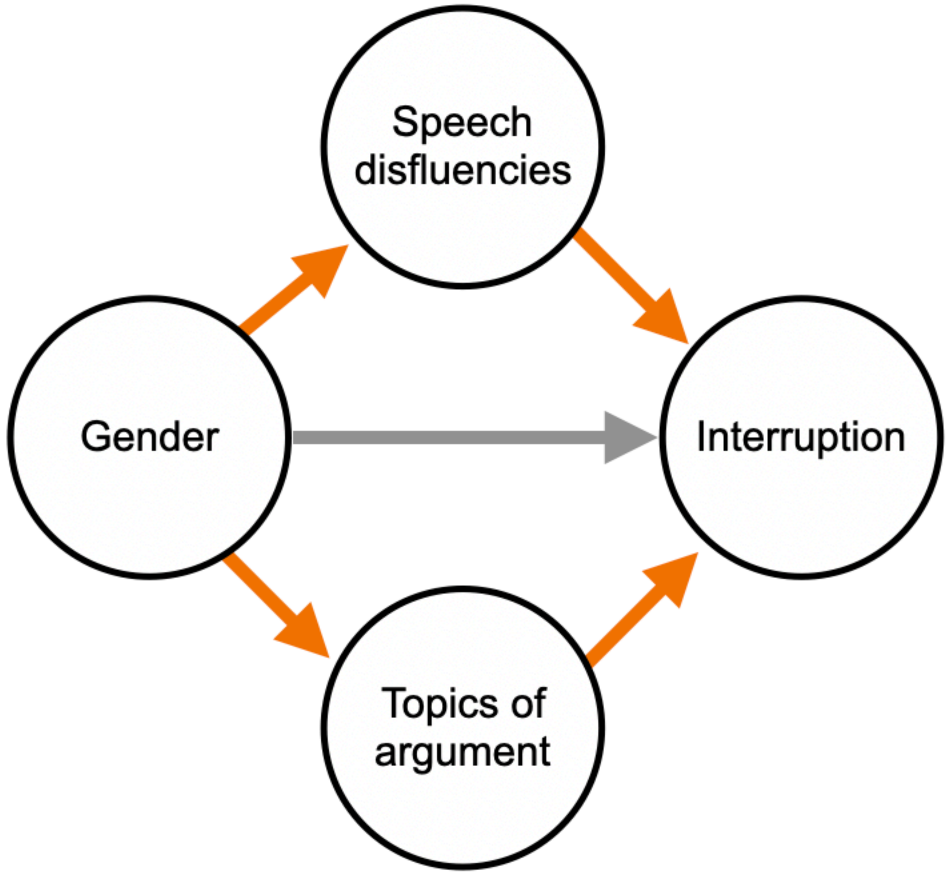
Natural direct effect (NDE)

How would a justice's interruptions of an advocate change if

- the signal of the advocate's gender the justice received flipped from male to female
- but the advocate still used language typical of a male advocate?

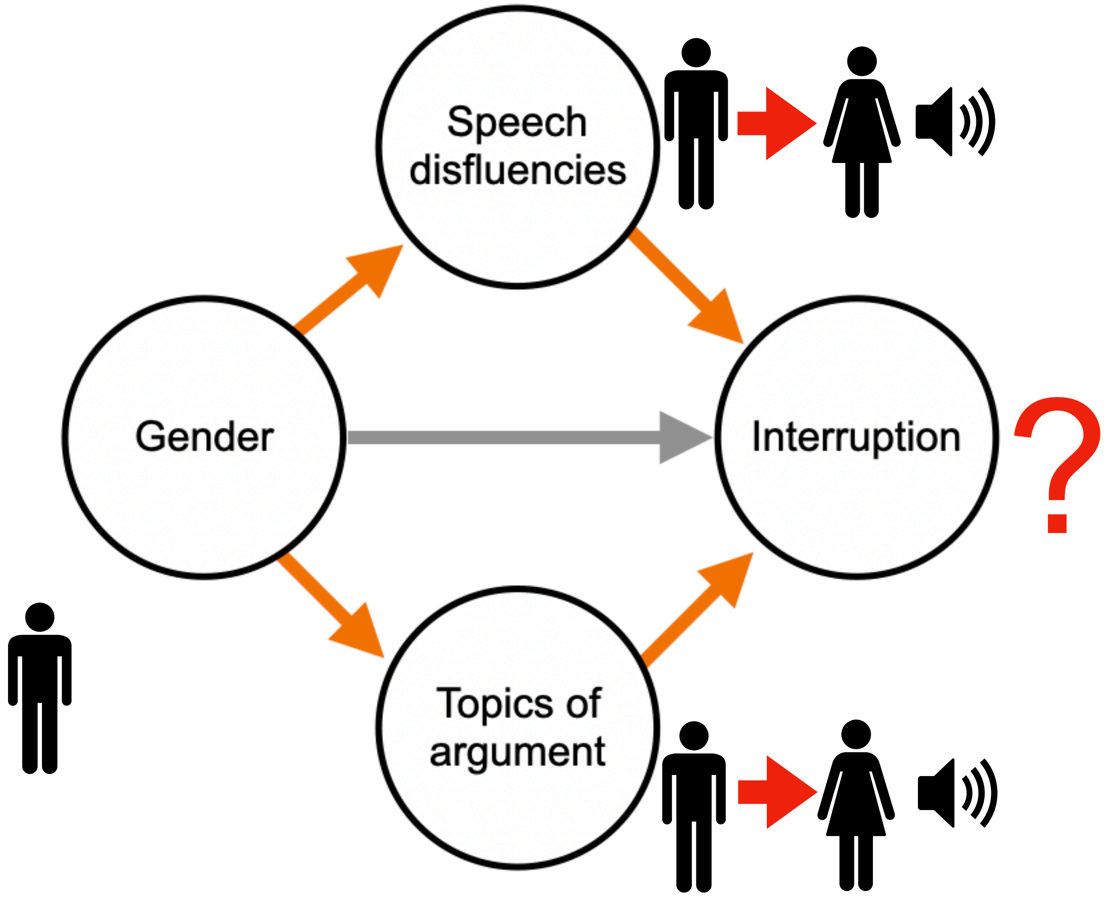
Explanation 2 corresponds to paths through mediators

Explanation 2:
Women are “less effective” advocates



Explanation 2 corresponds to paths through mediators

Explanation 2: Women are "less effective" advocates



Natural indirect effect (NIE)

How would a justice's interruptions of an advocate change if

- a male advocate used language typical of a female advocate
- but the signal of the advocate's gender the justice received remained male?

Identification

(1) Sequential ignorability

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x$$

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid \{T_i = t, X_i = x\}$$

(2) Mediator Independence

$$\forall j, j' : \; M_i^j(t) \perp\!\!\!\perp M_i^{j'}(t) \mid \{T_i = t, X_i = x\}$$

Based on Imai et al. 2010 and Pearl et al. 2016

Estimation

Natural **direct** effect

(e.g. gender->interruption)

$$\text{SA-NDE}^j = \frac{1}{N}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}^j}\Big(\hat{f}^j(Y|M_i^j = m, T_i = 1, X_i = x) - \hat{f}^j(Y|M_i^j = m, T_i = 0, X_i = x)\Big)\hat{g}^j(m|T_i = 0, X_i = x)$$

Natural **indirect** effect

(e.g. gender->mediators-> interruption)

$$\text{SA-NIE}^j = \frac{1}{N}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}^j}\hat{f}^j(Y|M_i^j = m, T_i = 0, X_i = x)\Big(\hat{g}^j(m|T_i = 1, X_i = x) - \hat{g}^j(m|T_i = 0, X_i = x)\Big)$$

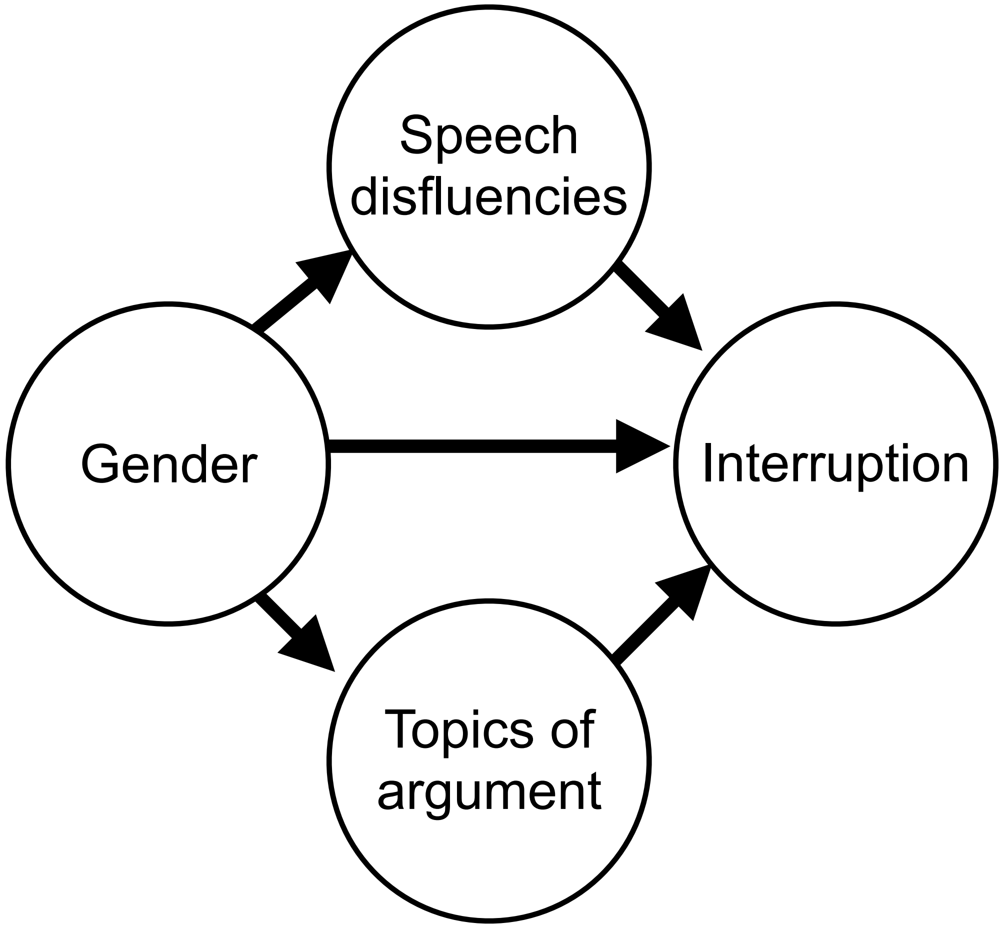
f-function
Models the outcome (y) given treatment (t) and confounders (x), and mediators (m)

g-function
Models the mediators (g) given treatment (t) and confounders (x)

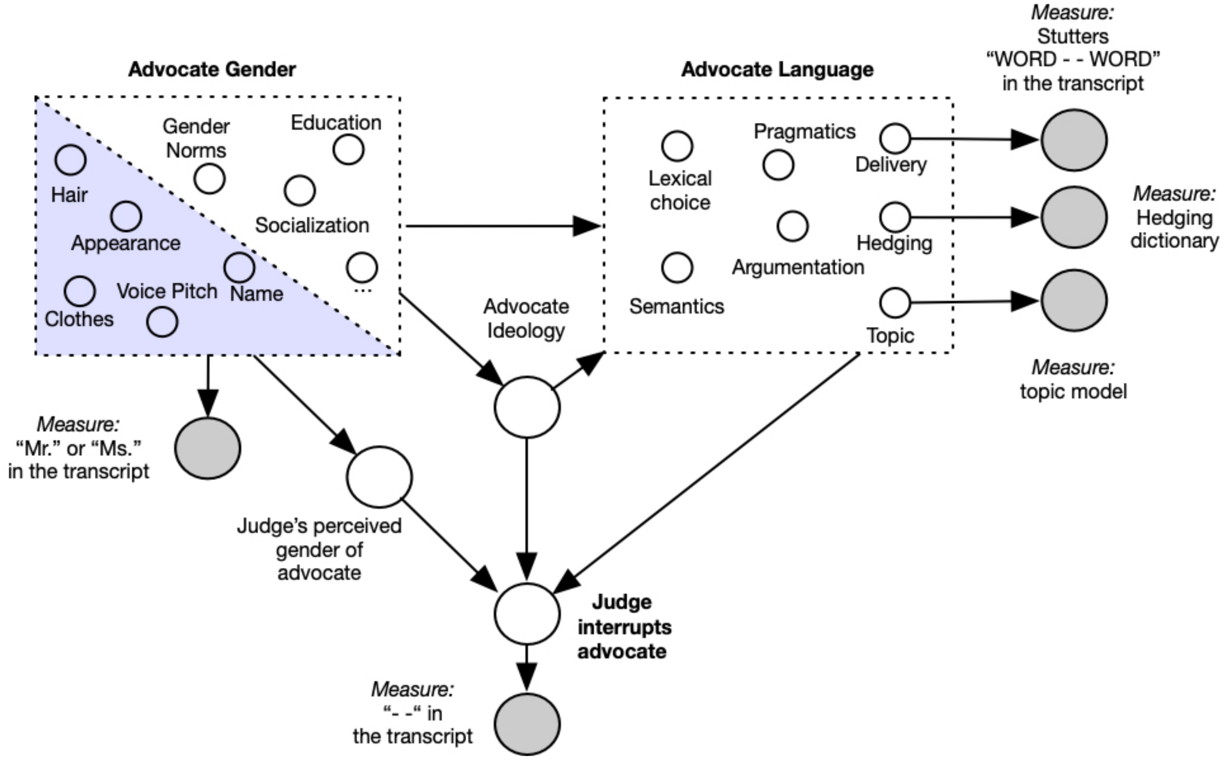
Based on Imai et al. 2010 and Pearl et al. 2016

Limitations of simple assumptions

Simplified DAG



Expanded (more realistic) DAG

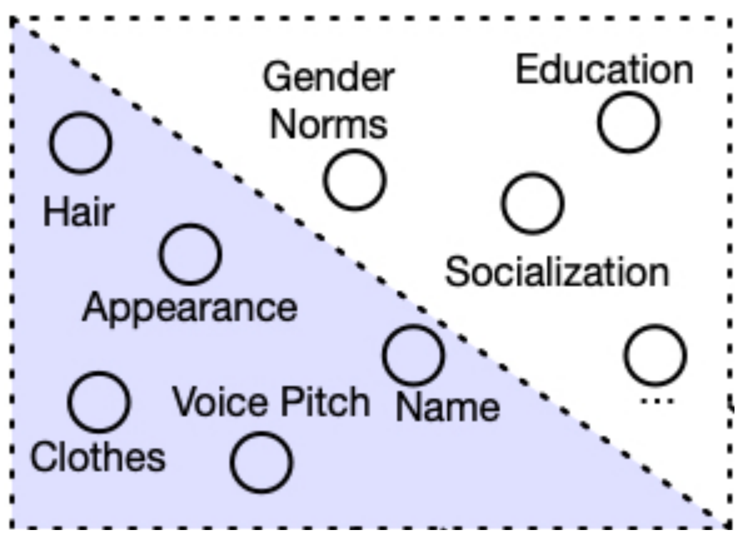


Gender as a causal “treatment”

Treatment options

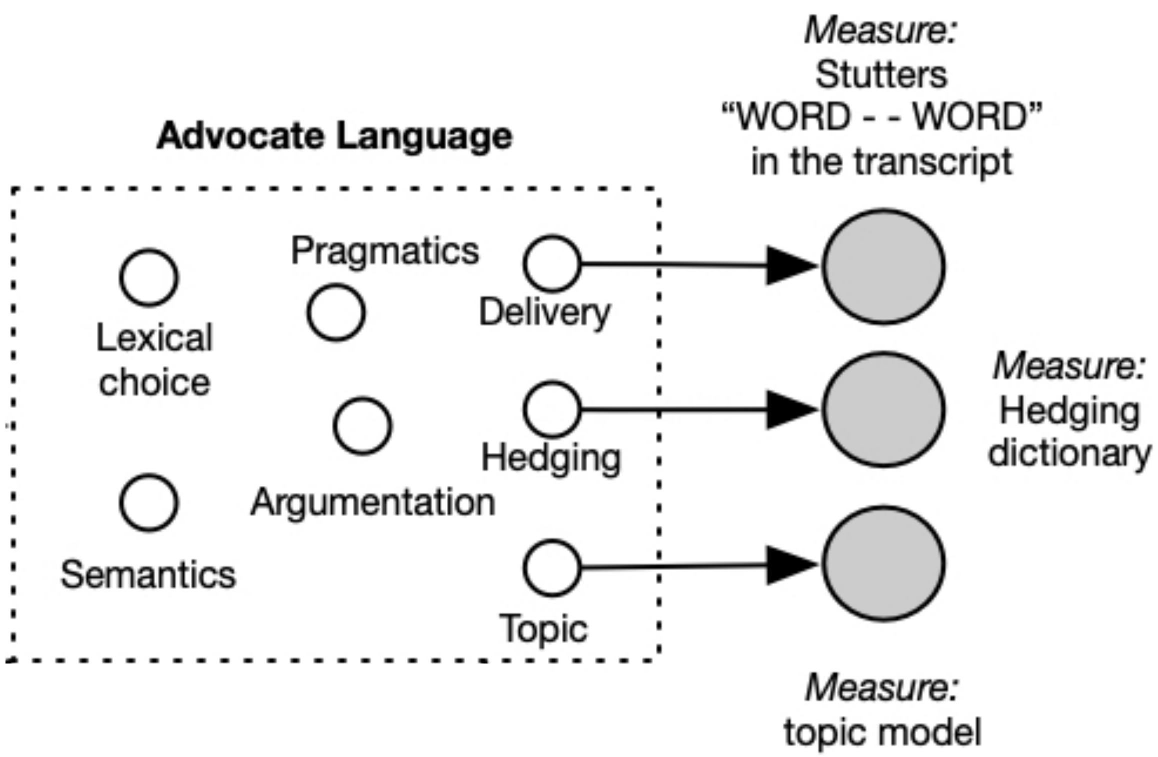
1. Do judges interrupt at different rates based on an advocate’s *gender*?
2. Based on an advocate’s *biological sex assigned at birth*?
3. An advocate’s *perceived gender*?
4. An advocate’s *gender signal*?
5. **An advocate’s *gender signal* as defined by (hypothetical) manipulations of the advocate’s clothes, hair, name, and voice pitch?**
6. An advocate’s *gender signal* by (hypothetical) manipulations of their entire physical appearance, facial features, name, and voice pitch?
7. An advocate’s *gender signal* by setting their physical appearance, facial features, name, and voice pitch to specific values (e.g. all facial features set to that of the same 40-year-old, white female and clothes set to a black blazer and pants).

Advocate Gender



Building from Sen and Wasow (2016);
Hu and Kohler-Hausmann (2020)

Operationalizing language as a causal mediator



Recommendations

- Hypothetical manipulations
- Causally independent mediators
- Substantive theory
- Measurement error

Next steps

- Empirical estimates from real data
- Address causal dependence between temporal utterances
- Analyze between-judge and between-court temporal estimates

Lecture Outline and Learning Objectives

1. Causal estimation, in general
   A. What is causal estimation and how does it differ from association and prediction?
   B. What are the challenges with causal estimation with text?

2. Text as causal confounders
   A. For observational data, how does one use back-door adjustment for text as a confounder?

3. Text as causal mediators
   A. For observational data, how does one estimate the natural direct and indirect causal effects with text as a mediator?

Thanks! Questions?