# Controlling for Text in Causal Inference

Emaad Manzoor

emaadmanzoor.com

ICWSM-22

# Emaad Manzoor

Carnegie Mellon University → THE UNIVERSITY of WISCONSIN MADISON → CORNELL UNIVERSITY FOUNDED A.D. 1865

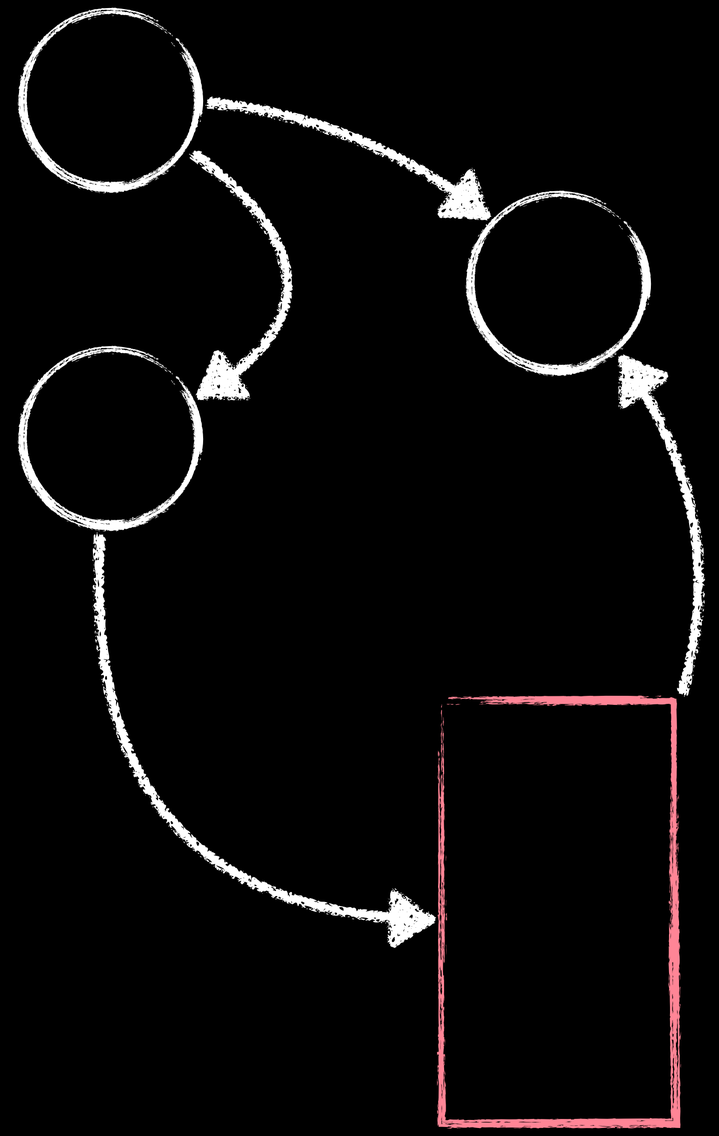
2016 — 2021 | 2021 — 2023 | 2023 —

emaadmanzoor.com/causal-inference-and-nlp-icwsm2022/

# Empirically Establishing Causality

*with*

# Unstructured Text as a Control



# Learning Outcomes

In this tutorial, you will learn how to:

1. Formalize casual questions as causal estimands
2. Understand the process of causal identification
3. Estimate causal effects with text as a control

# What is Causal Inference?

Using statistics & data to quantify the strength and existence of causal relationships

*Causal Inference in Statistics: An Overview. Pearl (2009).*

# Interdisciplinary Challenge

Vote

↑

Call

Click

↑

Ad

Buy

↑

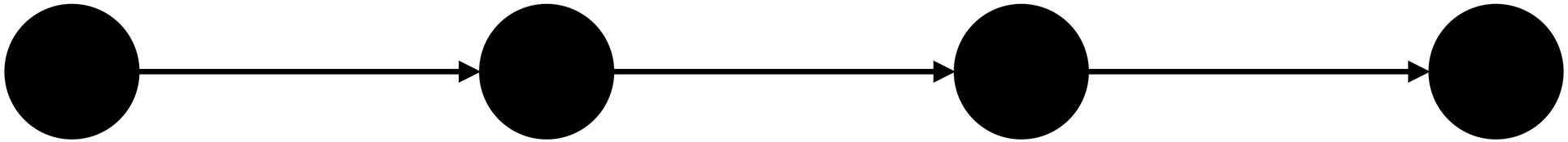
Email

Mask

↑

Pay

# The Process of Causal Inference



Causal Research Question

Causal Estimand

Causal Identification

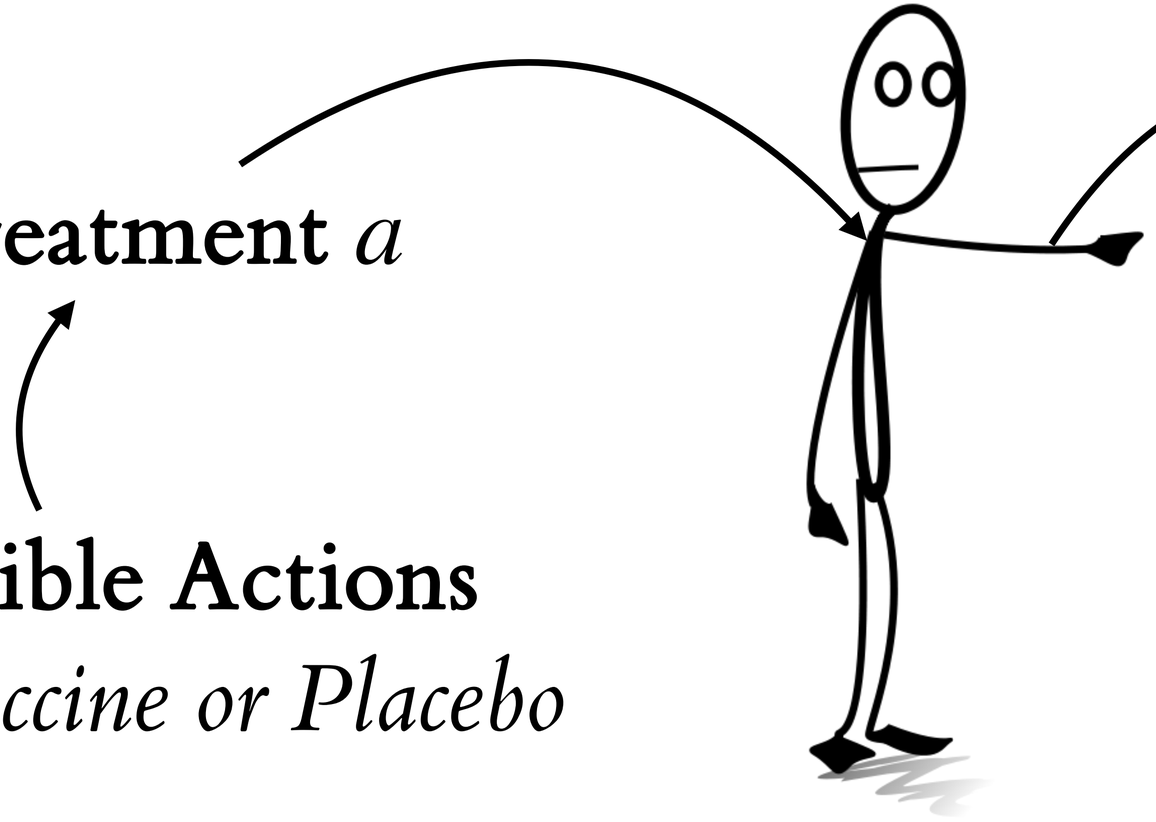
Estimation & Inference

# Causal Research Questions

- There are innumerable possible "effects of X on Y" questions in the world, only some are **meaningful**
- **Meaningful** causal research questions are typically motivated by social or economic **theory** — empirical studies are "tests of theory" (similar to physics)
- Many such tests → chances of **false discovery**: causal research questions need to be **motivated**

# Causal Estimand

## Formal definition of "target of causal inference"



Treatment $a$

Possible Actions

*Eg. Vaccine or Placebo*

Outcome $Y$

Observed after Treatment

*Eg. Positive or Negative*

# Example Estimand: ITE

$Y^{a=1}$

Outcome had individual been vaccinated

$Y^{a=0}$

Outcome had individual been given placebo

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

| | $Y^{a=0}$ | $Y^{a=1}$ |
|---|---|---|
| Rheia | 0 | 1 |
| Kronos | 1 | 0 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 0 | 0 |
| Poseidon | 1 | 0 |
| Hera | 0 | 0 |
| Zeus | 0 | 1 |
| Artemis | 1 | 1 |
| Apollo | 1 | 0 |
| Leto | 0 | 1 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 0 | 1 |
| Aphrodite | 0 | 1 |
| Cyclope | 0 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

# ITEs Cannot Be "Identified"

Cannot be "expressed as a function of observable data" (without making strong assumptions)

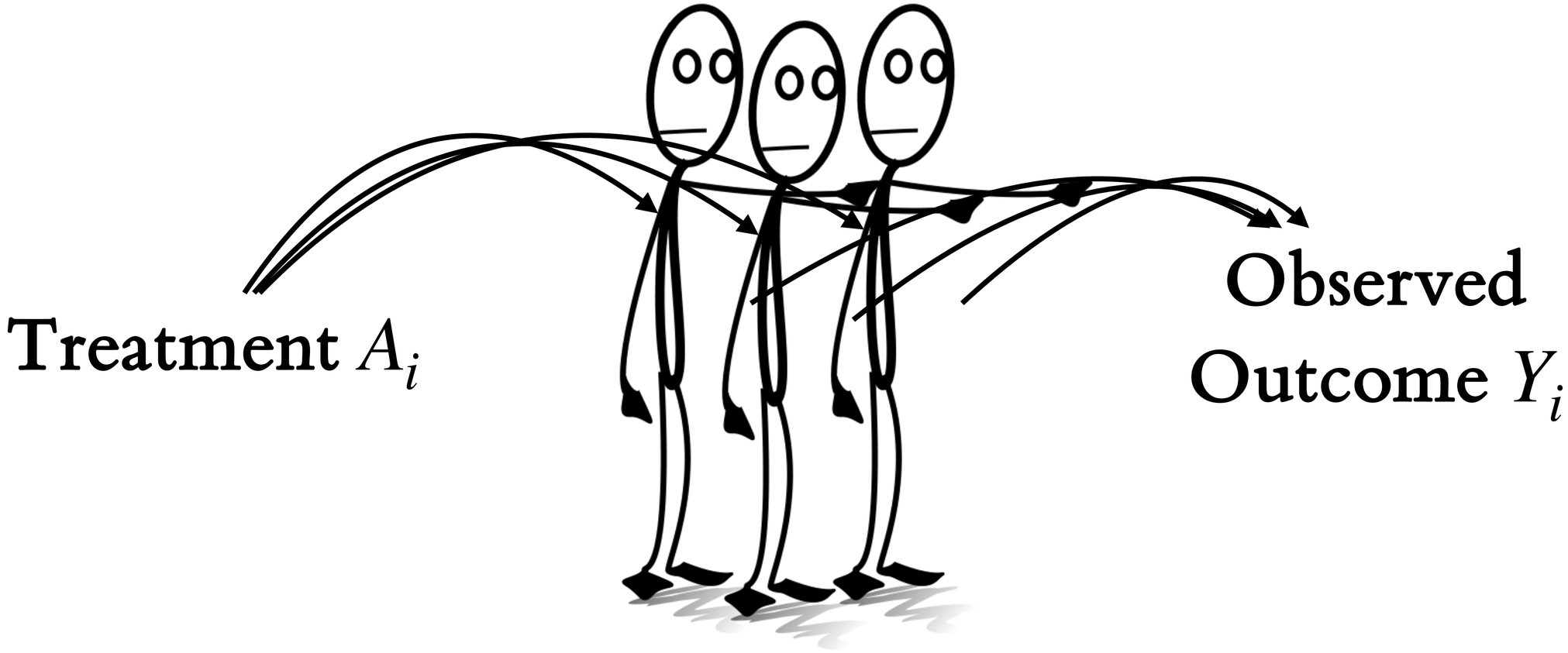
$Y^{a=1} = Y$ if vaccinated $\longrightarrow$ $Y^{a=0} = ?$

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

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

# Example Estimand: ATE

$N$ individuals $i = 1, \ldots, N$



# Example Estimand: ATE

$Y_i^{A_i=1}$

Outcome had $i$ been vaccinated

$Y_i^{A_i=0}$

Outcome had $i$ been given placebo

| | $Y_i^{a=0}$ | $Y_i^{a=1}$ |
|---|---|---|
| Rheia | 0 | 1 |
| Kronos | 1 | 0 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 0 | 0 |
| Poseidon | 1 | 0 |
| Hera | 0 | 0 |
| Zeus | 0 | 1 |
| Artemis | 1 | 1 |
| Apollo | 1 | 0 |
| Leto | 0 | 1 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 0 | 1 |
| Aphrodite | 0 | 1 |
| Cyclope | 0 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

**Average Treatment Effect (ATE)**

$$\mathbb{E}[Y_i^{A_i=1}] - \mathbb{E}[Y_i^{A_i=0}]$$

Most common causal estimand

# Causal Identification

**Expressing causal estimands in terms of observable (not counterfactual) quantities**

Requires making identification assumptions — a good "identification strategy" minimizes the assumptions required

When is ATE $= \mathbb{E}[Y_i^{A_i=1}] - \mathbb{E}[Y_i^{A_i=0}] =$ $E[Y_i|A_i = 1] - E[Y_i|A_i = 0]$?

| | $A_i$ | $Y_i$ |
|---|---|---|
| Rheia | 0 | 0 |
| Kronos | 0 | 1 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 1 | 0 |
| Poseidon | 1 | 0 |
| Hera | 1 | 0 |
| Zeus | 1 | 1 |
| Artemis | 0 | 1 |
| Apollo | 0 | 1 |
| Leto | 0 | 0 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 1 | 1 |
| Aphrodite | 1 | 1 |
| Cyclope | 1 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

# Identifying the ATE

**Assumption 1:** $Y_i^{A_i=a} = Y_i$ **if** $A_i = a$

| | $A$ | $Y$ |
|---|---|---|
| Rheia | 0 | 0 |
| Kronos | 0 | 1 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 1 | 0 |
| Poseidon | 1 | 0 |
| Hera | 1 | 0 |
| Zeus | 1 | 1 |
| Artemis | 0 | 1 |
| Apollo | 0 | 1 |
| Leto | 0 | 0 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 1 | 1 |
| Aphrodite | 1 | 1 |
| Cyclope | 1 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

# Identifying the ATE

**Assumption 1:** $Y_i^{A_i=a} = Y_i$ **if** $A_i = a$

Violated when there is more than one "version" of the treatment (for example, if $A_i = 1$ implies vaccination by Moderna *or* Pfizer)

$Y_i^{A_i=1} = M$ **for Moderna**
$Y_i^{A_i=1} = P$ **for Pfizer**

# Identifying the ATE

**Assumption 1:** $Y_i^{A_i=a} = Y_i$ **if** $A_i = a$

Violated when there is "interference" or "spillover" (for example, if vaccinating individual $i$ makes individual $j$ get the vaccine)

$$Y_i^{A_i=0} = P \textbf{ when } A_j = 0$$
$$Y_i^{A_i=0} = Q \textbf{ when } A_j = 1$$

# Identifying the ATE

**Assumption 1:** $Y_i^{A_i=a} = Y_i$ **if** $A_i = a$

| | $A$ | $Y$ | $Y^0$ | $Y^1$ |
|---|---|---|---|---|
| Rheia | 0 | 0 | 0 | ? |
| Kronos | 0 | 1 | 1 | ? |
| Demeter | 0 | 0 | 0 | ? |
| Hades | 0 | 0 | 0 | ? |
| Hestia | 1 | 0 | ? | 0 |
| Poseidon | 1 | 0 | ? | 0 |
| Hera | 1 | 0 | ? | 0 |
| Zeus | 1 | 1 | ? | 1 |
| Artemis | 0 | 1 | 1 | ? |
| Apollo | 0 | 1 | 1 | ? |
| Leto | 0 | 0 | 0 | ? |
| Ares | 1 | 1 | ? | 1 |
| Athena | 1 | 1 | ? | 1 |
| Hephaestus | 1 | 1 | ? | 1 |
| Aphrodite | 1 | 1 | ? | 1 |
| Cyclope | 1 | 1 | ? | 1 |
| Persephone | 1 | 1 | ? | 1 |
| Hermes | 1 | 0 | ? | 0 |
| Hebe | 1 | 0 | ? | 0 |
| Dionysus | 1 | 0 | ? | 0 |

# Identifying the ATE

**Assumption 1:** $Y_i^{A_i=a} = Y_i$ if $A_i = a$

**Assumption 2:** $Y_i^a \perp A_i$

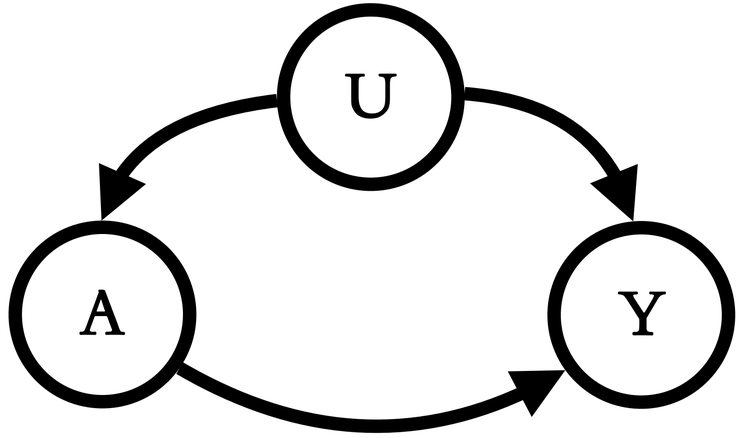
| | $A$ | $Y$ | $Y^0$ | $Y^1$ |
|---|---|---|---|---|
| Rheia | 0 | 0 | 0 | ? |
| Kronos | 0 | 1 | 1 | ? |
| Demeter | 0 | 0 | 0 | ? |
| Hades | 0 | 0 | 0 | ? |
| Hestia | 1 | 0 | ? | 0 |
| Poseidon | 1 | 0 | ? | 0 |
| Hera | 1 | 0 | ? | 0 |
| Zeus | 1 | 1 | ? | 1 |
| Artemis | 0 | 1 | 1 | ? |
| Apollo | 0 | 1 | 1 | ? |
| Leto | 0 | 0 | 0 | ? |
| Ares | 1 | 1 | ? | 1 |
| Athena | 1 | 1 | ? | 1 |
| Hephaestus | 1 | 1 | ? | 1 |
| Aphrodite | 1 | 1 | ? | 1 |
| Cyclope | 1 | 1 | ? | 1 |
| Persephone | 1 | 1 | ? | 1 |
| Hermes | 1 | 0 | ? | 0 |
| Hebe | 1 | 0 | ? | 0 |
| Dionysus | 1 | 0 | ? | 0 |

# Identifying the ATE

**Assumption 1:** $Y_i^{A_i=a} = Y_i$ **if** $A_i = a$

**Assumption 2:** $Y_i^a \perp A_i$

In general, $Y_i^a \perp A_i$ is formally assessed under assumptions using causal directed acyclic graphs and do-calculus *(Pearl, 2008)*

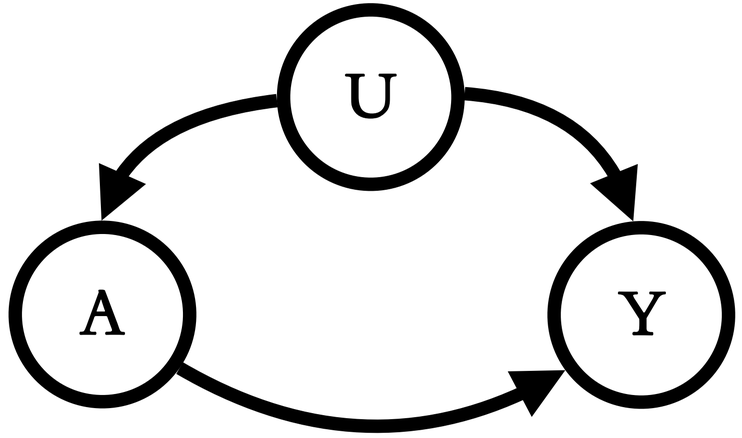



Causal effect of $A$ on $Y$ is identified if pathways between $U$ and $A$ or $U$ and $Y$ are blocked

# Identifying the ATE

**Assumption 1:** $Y_i^{A_i=a} = Y_i$ **if** $A_i = a$

**Assumption 2:** $Y_i^a \perp A_i$




**Identification Proof:**

$$\text{ATE} = \mathbb{E}[Y_i^{A_i=1}] - \mathbb{E}[Y_i^{A_i=0}]$$

$$= \mathbb{E}[Y_i^{A_i=1} | A_i = 1] - \mathbb{E}[Y_i^{A_i=0} | A_i = 0] \text{ (assumption 2)}$$

$$= \mathbb{E}[Y_i | A_i = 1] - \mathbb{E}[Y_i | A_i = 0] \text{ (assumption 1)}$$

# Randomized Experiments (RCTs)

Randomly assign individuals to treatment actions

$Y_i^a \perp A_i$ by design if treatment is randomized

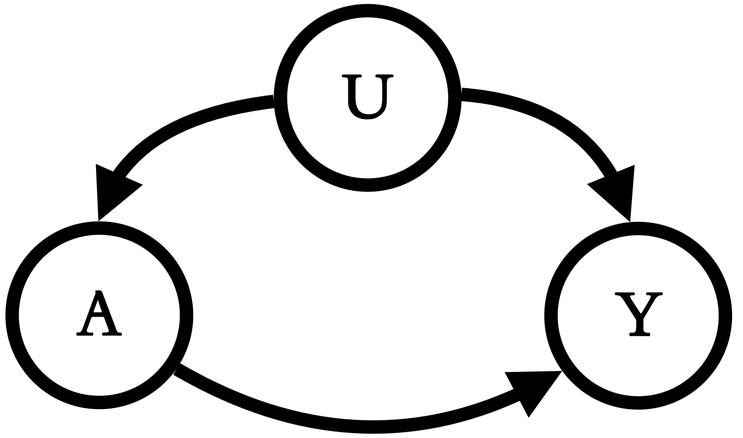
(and there is no attrition / selection bias)

**Issues with randomized experiments:**
Ethicality, feasibility, cost, generalizability

# Causality with Observational Data

**Strategies to argue for $Y_i^a \perp A_i$:**

1. Control for observed confounders
2. Block causal pathways between unobserved confounders and treatment/outcome
3. Find natural or quasi-experiments to reduce the assumptions required

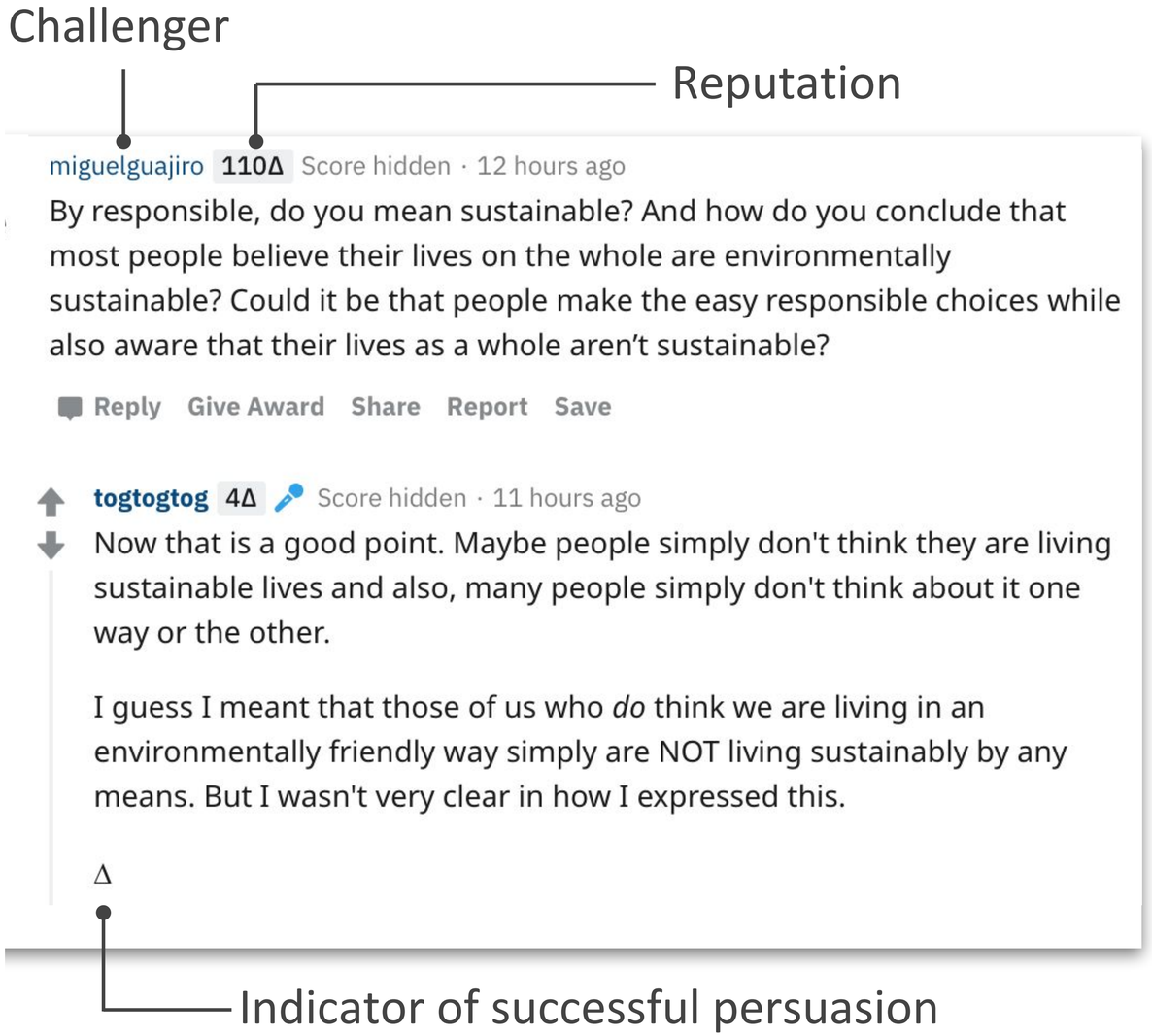



Causal effect of $A$ on $Y$ is identified if pathways between $U$ and $A$ or $U$ and $Y$ are blocked

# Text for Causal Identification

**Does visible status make you more persuasive in online conversations?**

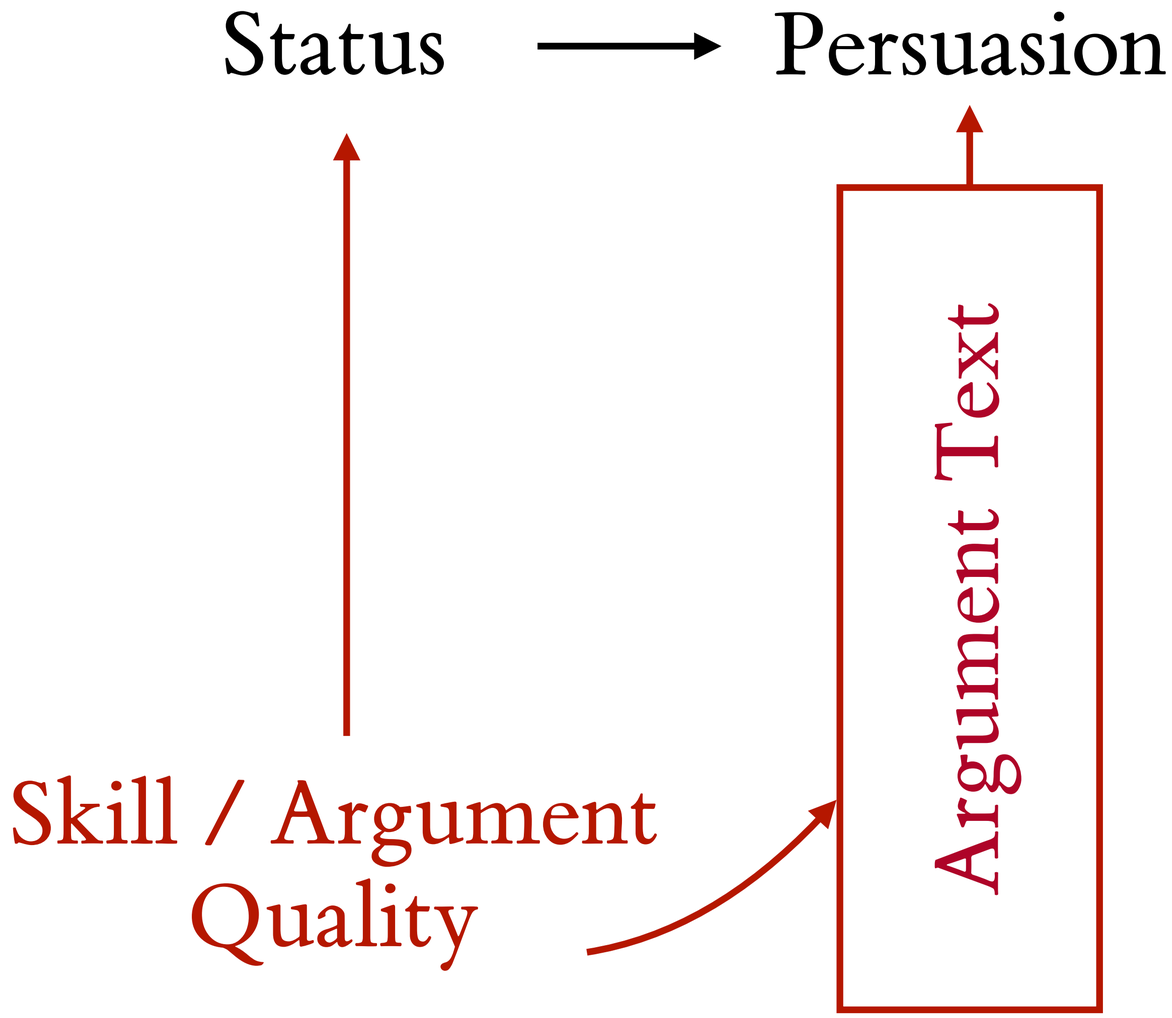
Based on "Influence via Ethos: On the Persuasive Power of Reputation in Deliberation Online" *(arxiv.org/abs/2006.00707)*



# Text for Causal Identification

**Does visible status make you more persuasive in online conversations?**

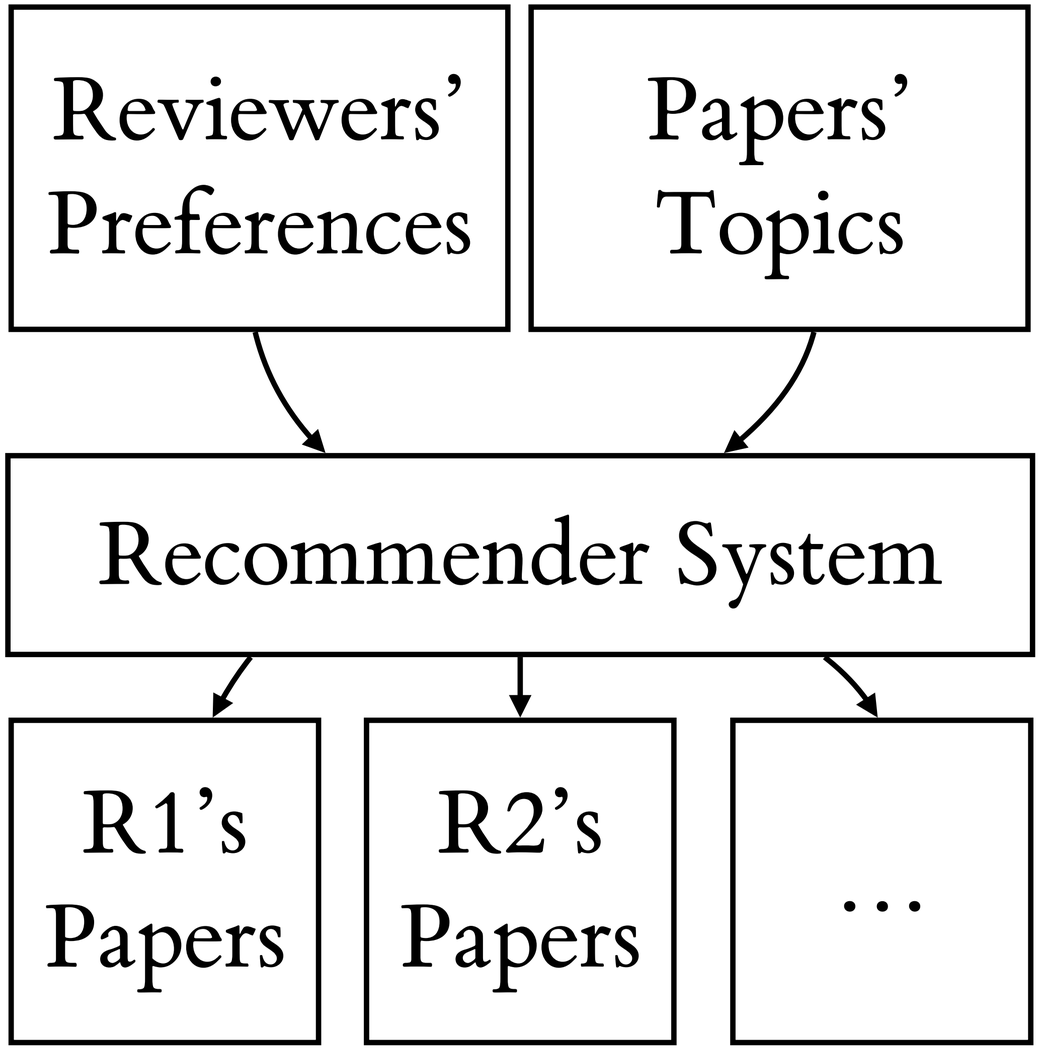
Based on "Influence via Ethos: On the Persuasive Power of Reputation in Deliberation Online" *(arxiv.org/abs/2006.00707)*



# Text for Causal Identification

**Would having a theorem improve a paper's rating?**

**Setting**: Recommender system provides a small paper list to each reviewer based on reviewer preferences and the paper text

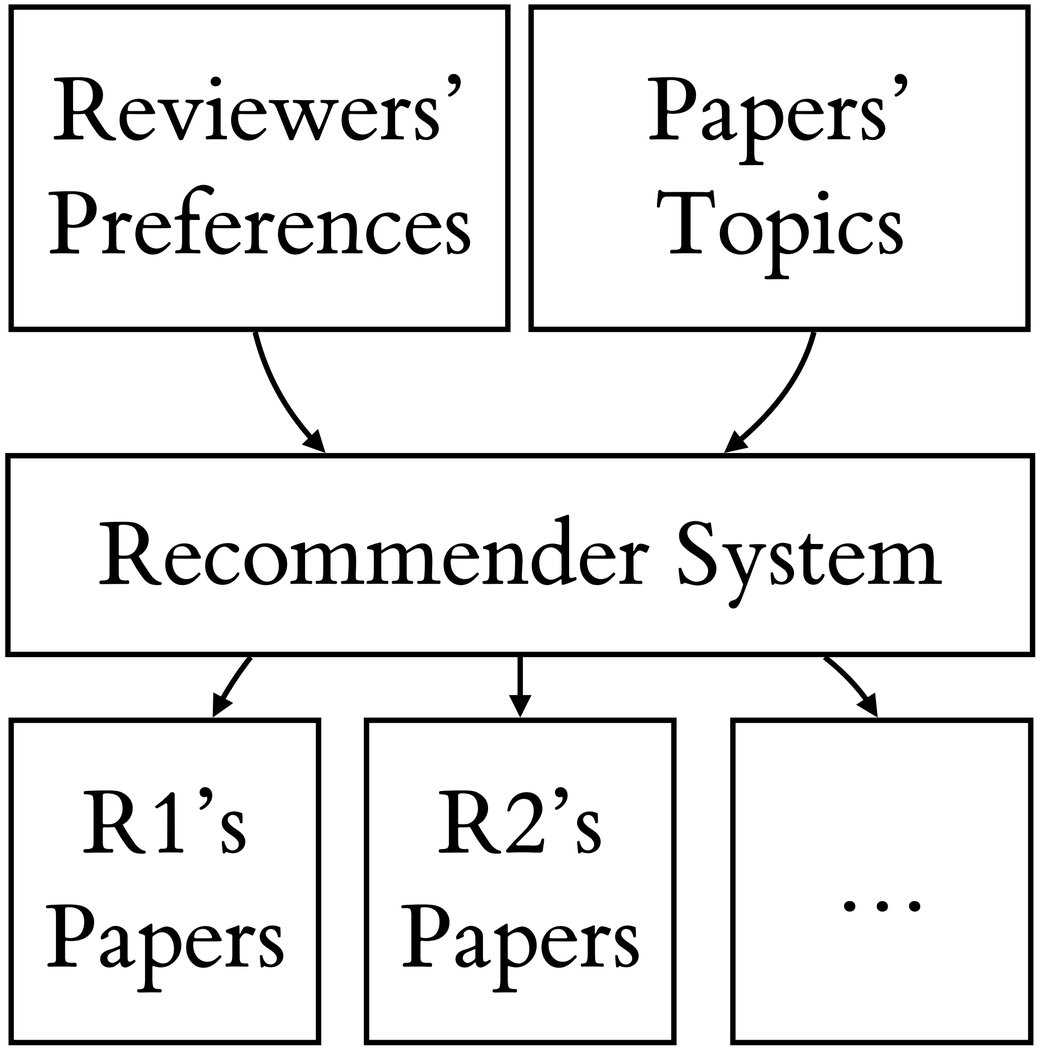


# Text for Causal Identification

**Would having a theorem improve a paper's rating?**

**Treatment** ($A_j = 0,1$): Paper $j$ has theorem ($A_j = 1$) or not

**Outcome** ($r_{ij} = 1,\ldots,5$): Reviewer $i$'s rating for paper $j$

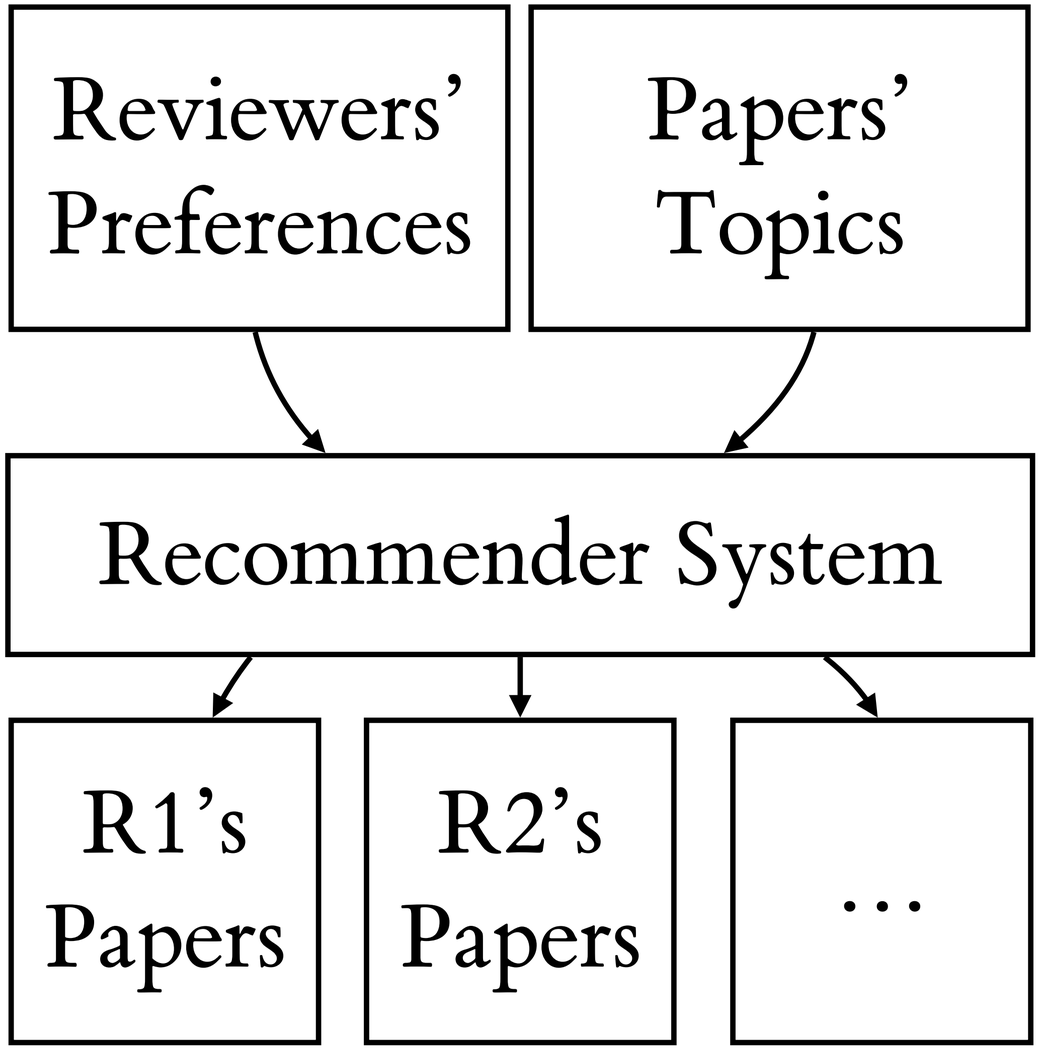


# Text for Causal Identification

**Would having a theorem improve a paper's rating?**

**Target Estimand**: ATE for each reviewer $i$, over all papers $j$

$$\mathbf{ATE_i} = E[r_{ij}^{A_j=1}] - E[r_{ij}^{A_j=0}]$$



# Text for Causal Identification

**Would having a theorem improve a paper's rating?**

**Is treatment assigned randomly?**

**No.** For each reviewer, some papers more likely to be recommended than others

$$E[r_{ij}^{A_j=a}] \neq E[r_{ij} \mid A_j = a]$$

# Text for Causal Identification

**Would having a theorem improve a paper's rating?**

For a given reviewer, if I fix the research topic, any paper is equally likely to be recommended (random)

$$E[r_{ij}^{A_j=a} \mid \text{Topic}_j] =$$

$$E[r_{ij} \mid A_j = a, \text{Topic}_j]$$

**Conditional Randomization**

# Text for Causal Identification

**Would having a theorem improve a paper's rating?**

Since each paper's topic can be fully inferred from its text, I can simply control for each paper's text

$$E[r_{ij}^{A_j=a} \mid \text{Text}_j] =$$

$$E[r_{ij} \mid A_j = a, \text{Text}_j]$$

**Conditional Randomization**

# The Estimation Challenge

## Text is inherently unstructured, high dimensional

- **Several ad-hoc ways to structure text and reduce its dimensionality**: Topic modeling (LDA, NMF), document embeddings, hand-coding features
- **Key issue 1**: No guarantee confounders are retained
- **Key issue 2**: Brittle (which representation is the best?)
- **Key issue 3**: Inference is generally invalid

# The Estimation Challenge

## Common approach: Fixed $g(.)$ (eg. topics)

**Mechanics:**

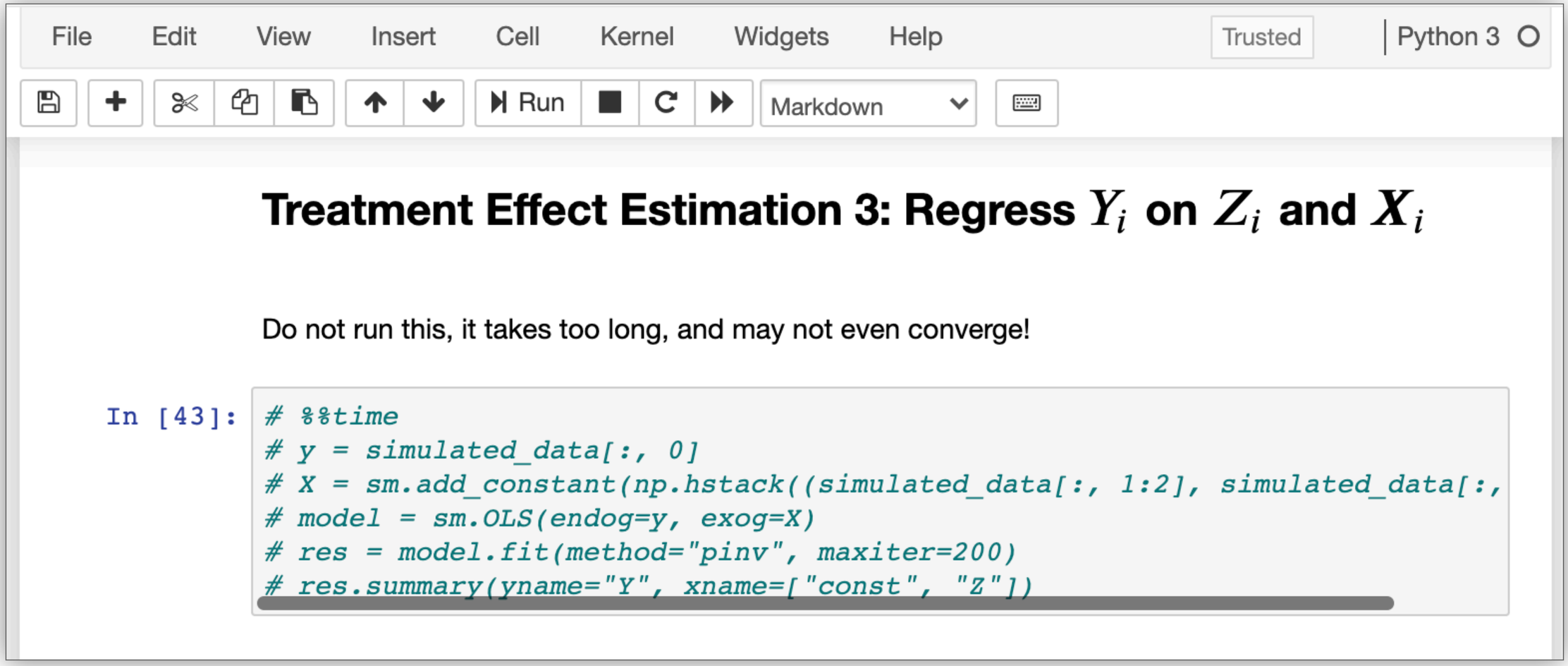
1. Apply $g(.)$ to text to obtain text covariates
2. Regress $Y$ on $T$ and text covariates

$$Y = \theta T + g(\text{text}) + \epsilon$$

$\theta$: Treatment Effect

$g(\text{text})$: Fixed function of text

# Demo: Controlling for Words



# Demo: Control for Topics

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

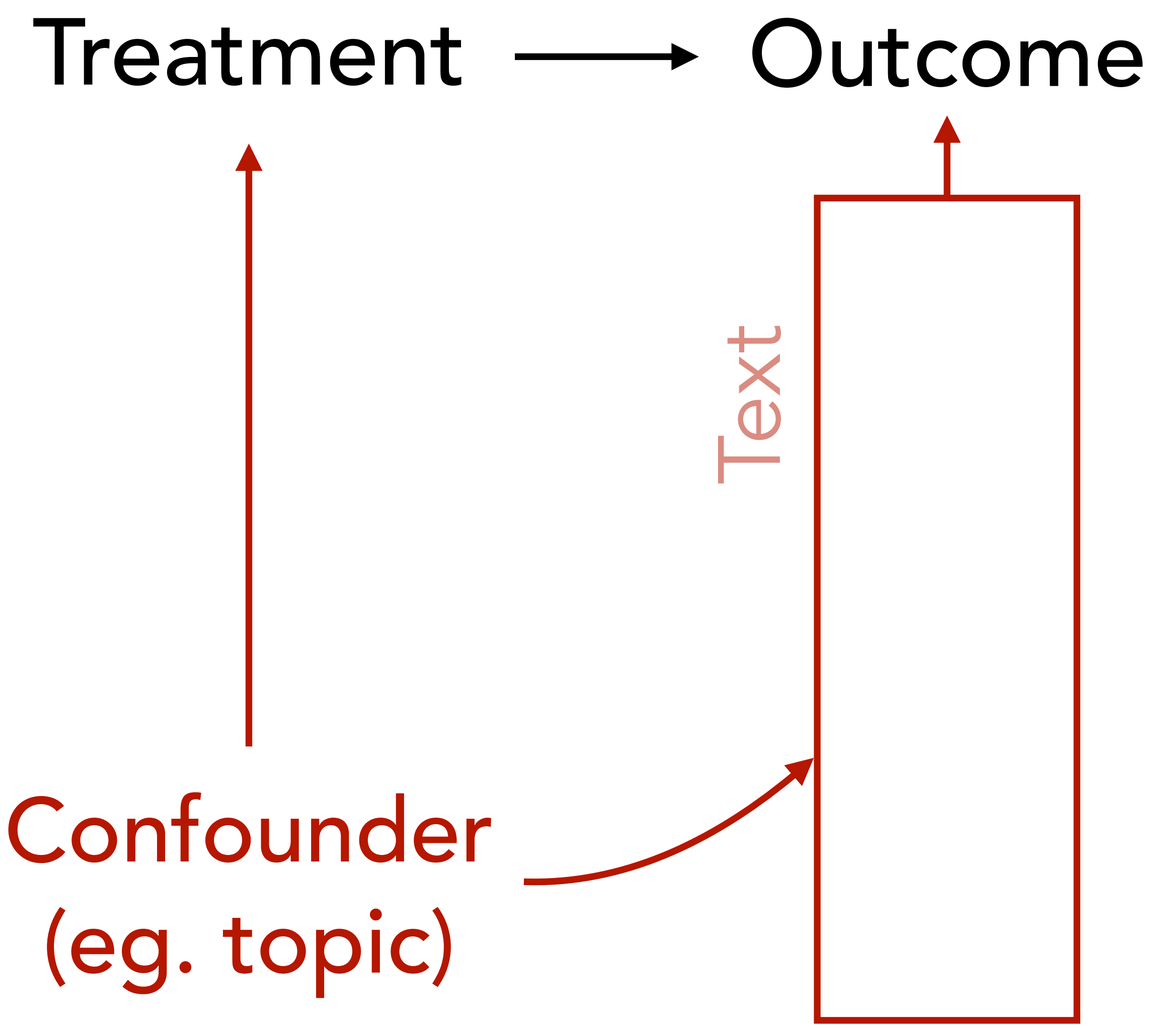
## Treatment Effect Estimation 4: Regress $Y_i$ on $Z_i$ and Document-Topic Weights

Play around with the number of topics.

In [77]:
```
%time
def print_top_words(model, feature_names, n_top_words):
    for topic_idx, topic in enumerate(model.components_):
        message = "Topic #%d: " % topic_idx
        message += " ".join([feature_names[i]
                             for i in topic.argsort()[:-n_top_words - 1:-1]])
        print(message)
    print()

umtopics = 50
mf = NMF(n_components=numtopics).fit(tfidf_vectors)
```

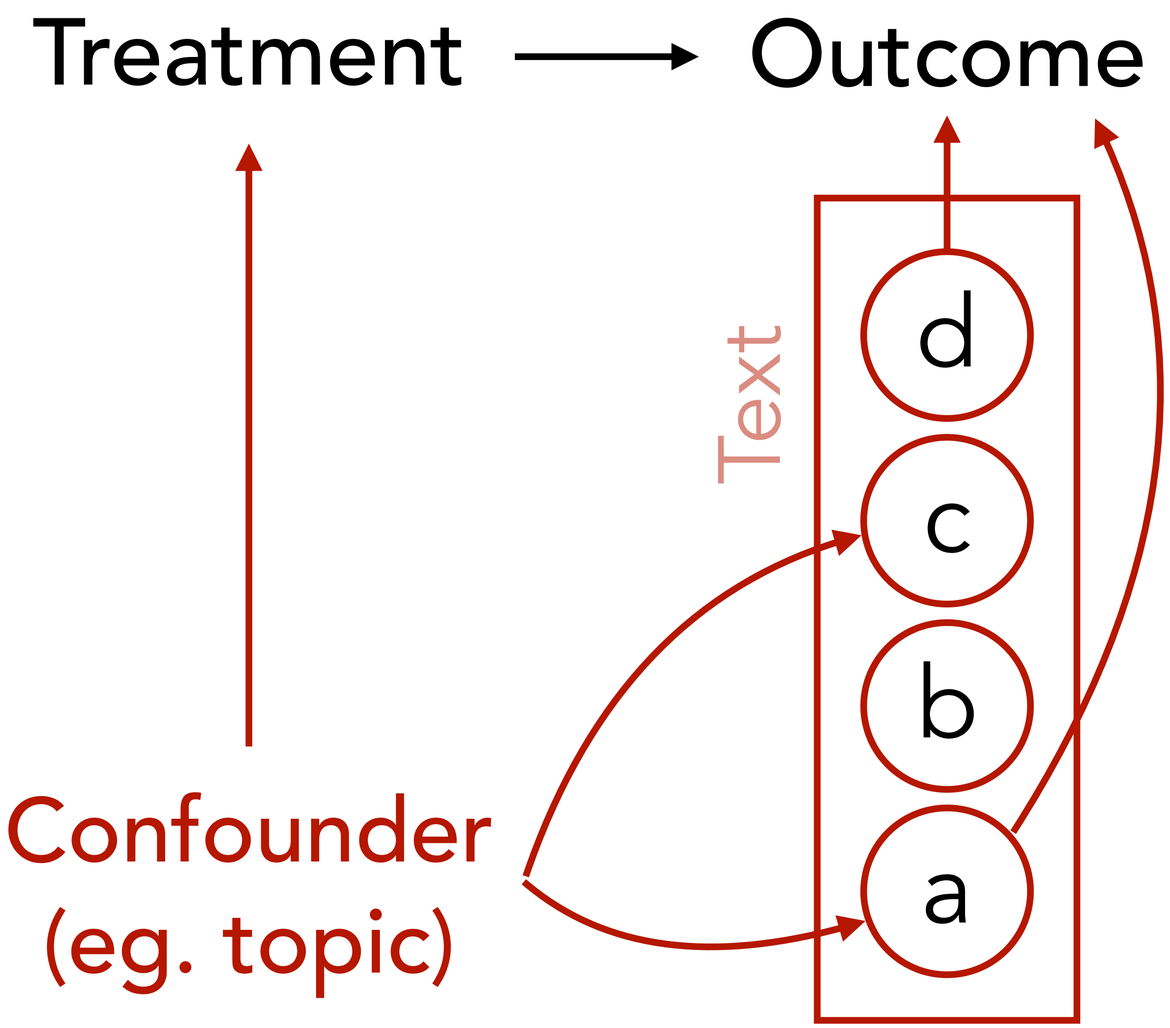
# An Alternate Approach




**Directed Acyclic Graph:** Arrows represent possible causality, no arrow represents no causality

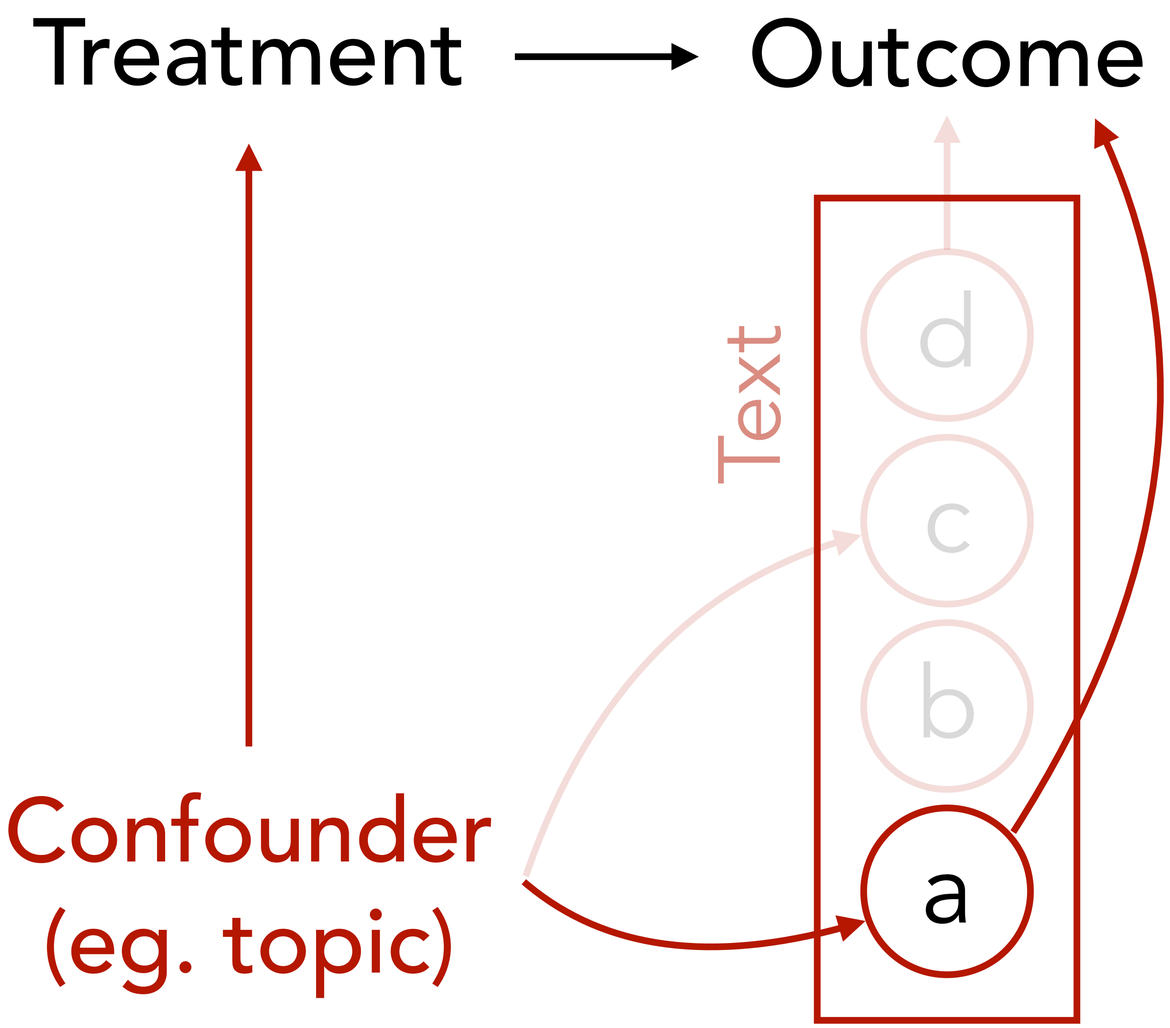
**Recall:** Confounder is common cause of treatment and outcome

# An Alternate Approach




Can view text as 4 logical components

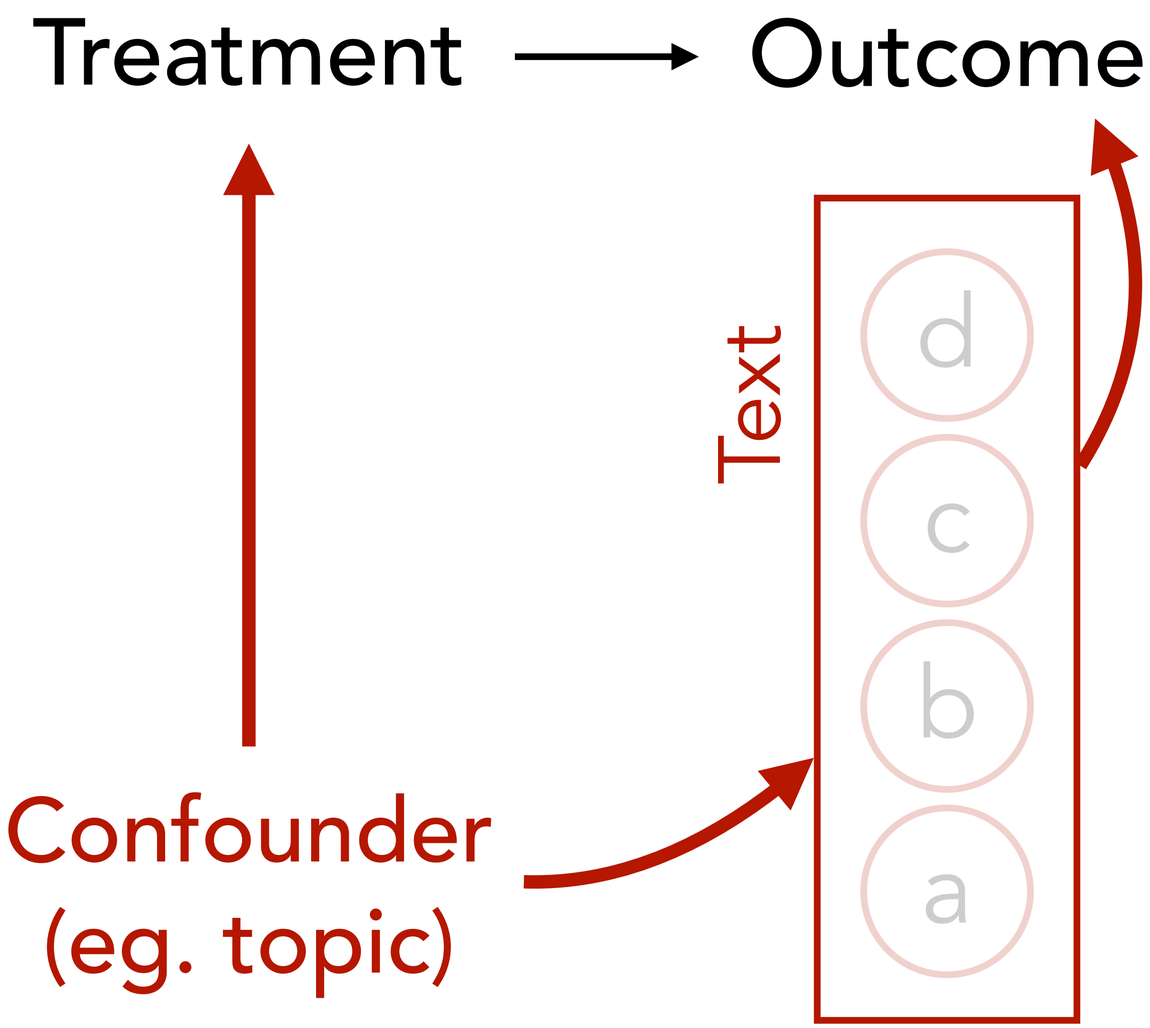
# An Alternate Approach




**Only need to somehow find and control for component *a***

Needle in a haystack

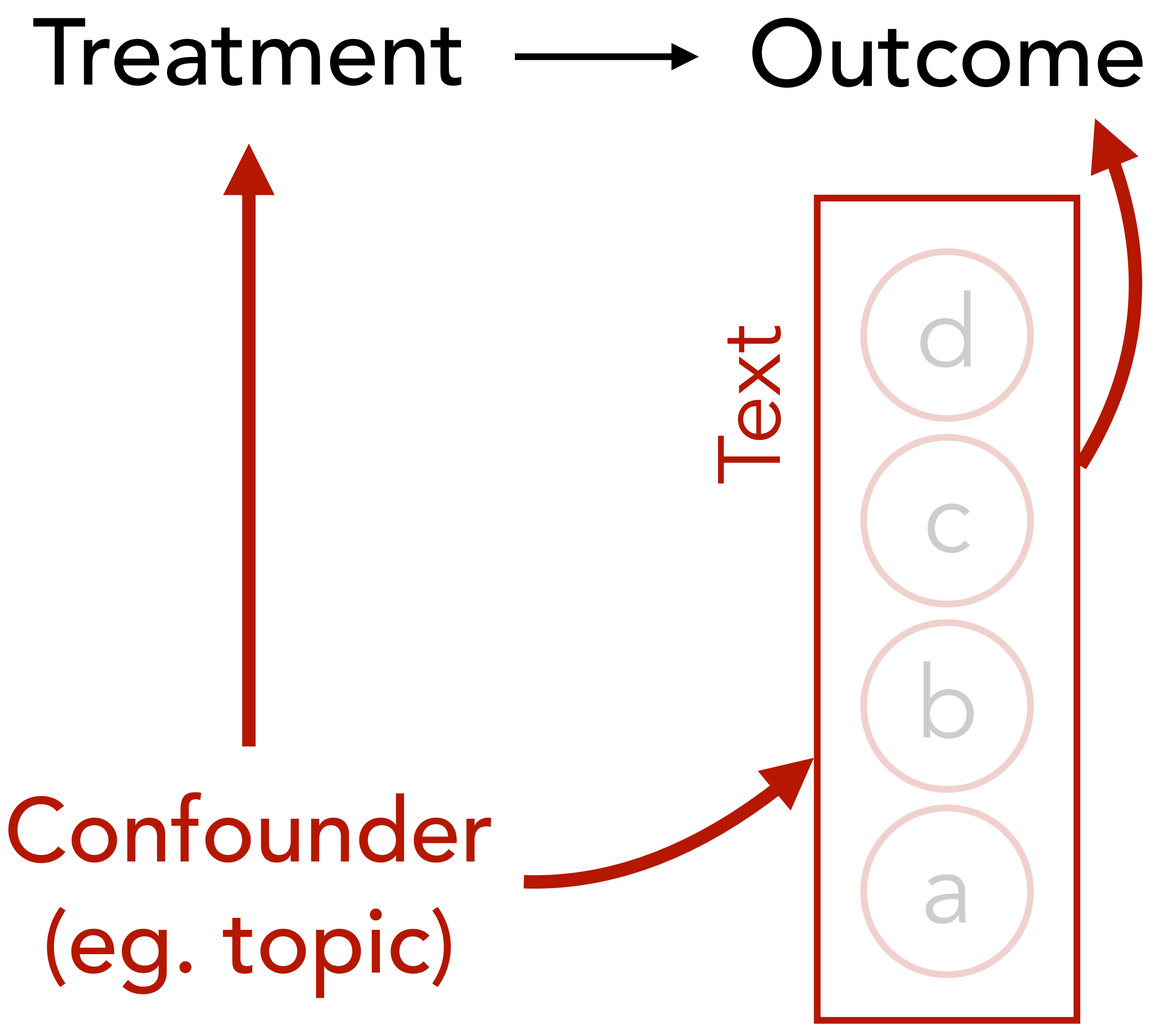
# An Alternate Approach




**Alternative to finding this needle without using dimensionality reduction**

Measure and combine correlation between text, treatment, and outcome

# An Alternate Approach




**Measuring correlations**

How well can I predict the treatment status / outcome value from the text?

# Double ML Mechanics

**Unknown $g(.)$ jointly estimated with $\hat{\theta}$**

**Mechanics** *(Robinson, 1988)*:

1. Measure prediction errors
   $\tilde{Y} = Y - m_1(\text{text})$,
   $\tilde{T} = T - m_2(\text{text})$
2. Regress $\tilde{Y}$ on $\tilde{T}$

$$Y = \theta T + g(\text{text}) + \epsilon$$

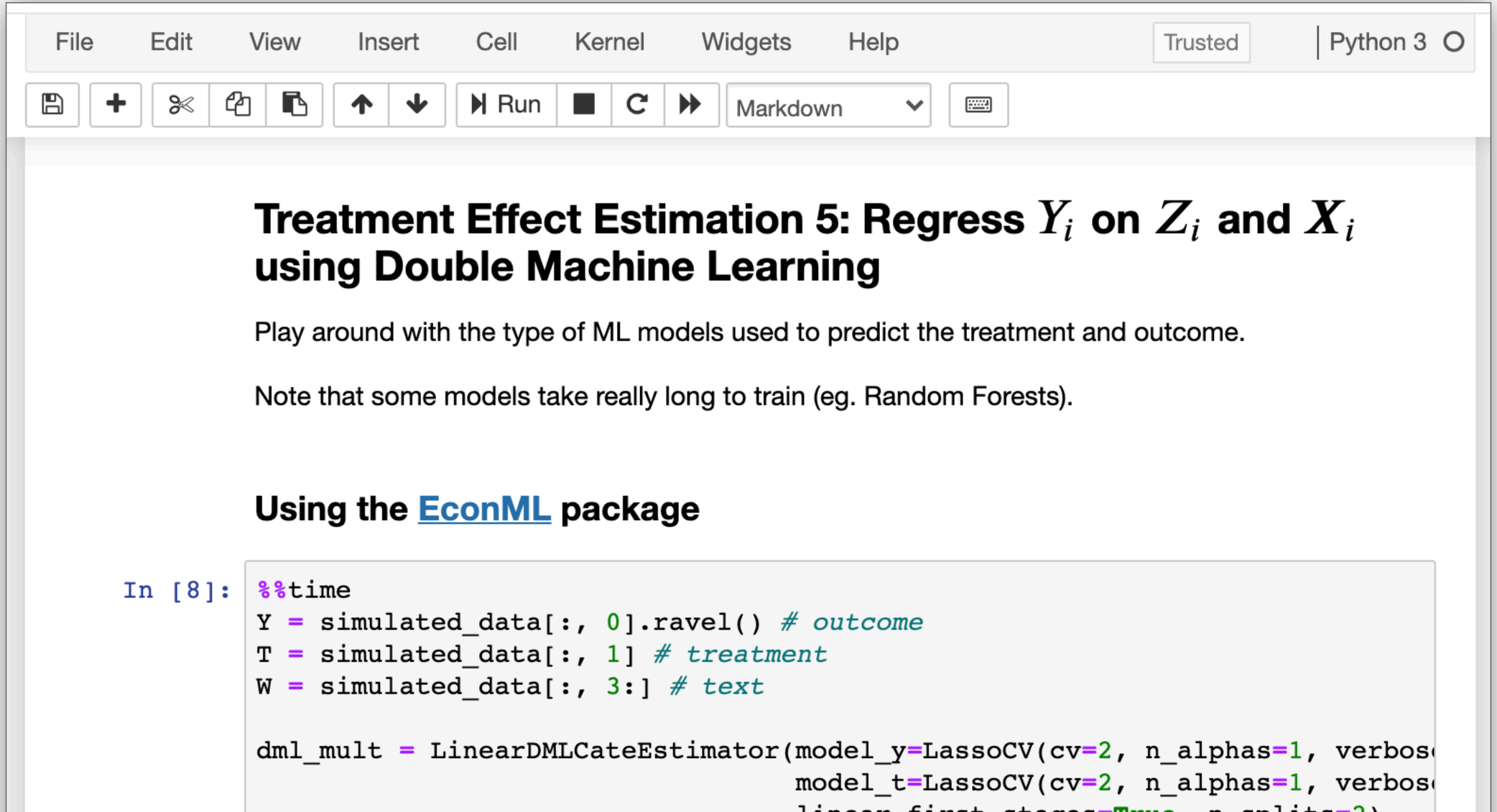
Treatment Effect

Unknown function of text

# Double ML Theory

- **Issue with** *(Robinson, 1988)*: $\sqrt{n}$ consistency of $\hat{\theta}$ requires $m_1, m_2$ to be sieve estimators — poor models of text
- **Double machine learning (DML)**: $\sqrt{n}$ consistency possible if $m_1, m_2$ are regularized ML models trained on held-out data — great for text! (must converge at $n^{-1/4}$ or better)
- Valid asymptotic confidence intervals
- General recipe (extensible beyond partially-linear models)

# Double ML + Text Demo



# Biases Eliminated by Double ML

## Regularization bias and overfitting bias

$$Y = D\hat{\theta}_0 + \hat{g}_0(X) + \hat{U}$$

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{(\frac{1}{n}\sum_{i\in I} D_i^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i\in I} D_i U_i}_{:=a}$$

$$+ \underbrace{(\frac{1}{n}\sum_{i\in I} D_i^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i\in I} D_i (g_0(X_i) - \hat{g}_0(X_i))}_{:=b}$$

$g_0$ is an ML method — $b$ goes to 0 too slowly!

# Biases Eliminated by Double ML

**Eliminating Regularization Bias**: Partialling-out procedure as we saw earlier, also called orthogonalization

$$b^* = (E[V^2])^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} \underbrace{(\hat{m}_0(X_i) - m_0(X_i))}_{\hat{m}_0 \text{ estimation error}} \underbrace{(\hat{g}_0(X_i) - g_0(X_i))}_{\hat{g}_0 \text{ estimation error}}$$

Bias after orthogonalization is the product of 2 errors — goes to zero more quickly (also source of $n^{-1/4}$ convergence rate requirement)

# Biases Eliminated by Double ML

**Eliminating Overfitting Bias:** Cross-fitting

- Split sample into train and estimation subsets
- Train ML models on the train subset
- Impute prediction errors on the estimation subset
- Estimate causal effect on the estimation subset
- Repeat after flipping subsets, average estimates

# Alternative Approaches

1. **Causal Forests**: Restricted to tree-structured models, double ML permits using neural networks
2. **Causal BERT, DragonNet, etc.**: Do not have consistency guarantees, ways to do inference
3. **Targeted learning / TMLE (van der Laan et al)**: Better finite sample properties

# Next Steps

- **Survey (preprint)**: *Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond*
- **Survey (ACL '20)**: *Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates*
- **Preprint using double ML to control for text**: *On the Persuasive Power of Reputation in Deliberation Online*