Controlling for Text in Causal Inference

NLP+CSS 201, Fall 2021

Emaad Manzoor

emaadmanzoor.com

Outline

1. Causal inference primer
2. Causality from non-experimental data
3. Text as a control
4. Double machine learning
5. Next steps

What is Causal Inference?

Process of (i) establishing, and (ii) quantifying causal relationships empirically (using statistics + data)

Interdisciplinary Challenge

Vote

↑

Call

Click

↑

Ad

Buy

↑

Email

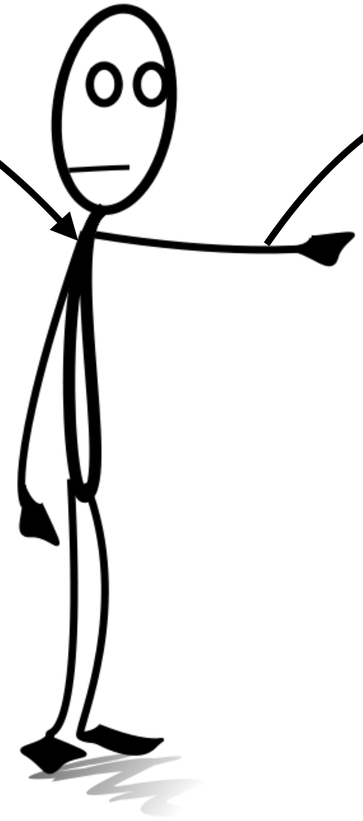
Mask

↑

Pay

What is a Causal Effect?

Typical Setup



Treatment a

Possible Actions

Eg. Vaccine or Placebo

Outcome Y

Observed after Treatment

Eg. Positive or Negative

Causal 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 measured from observable data"

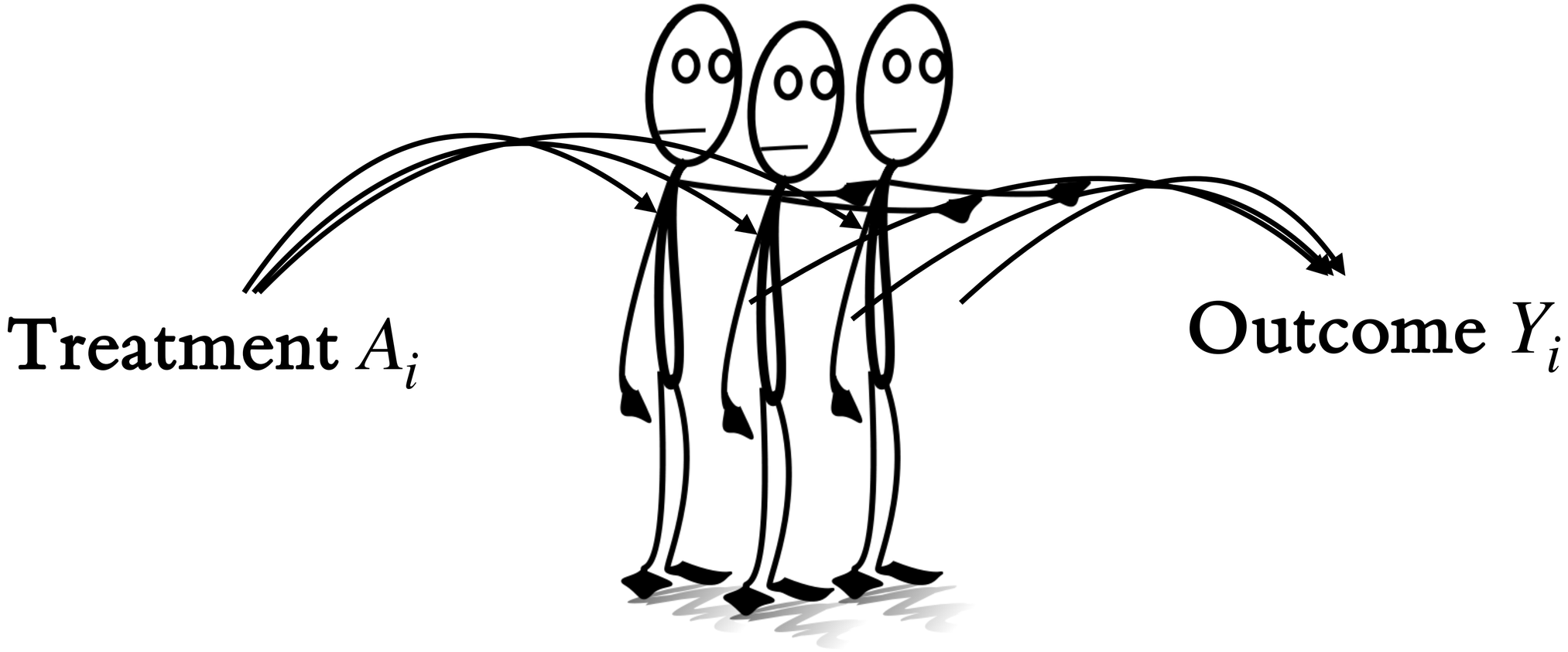
$Y^{a=1}$ = Y if vaccinated ⟶ $Y^{a=0}$ = ?

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

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

Causal Estimand: ATE

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



Causal Estimand: ATE

$Y_i^{A_i=1}$

Outcome had i been vaccinated

$Y_i^{A_i=0}$

Outcome had i been given placebo

	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

Average Treatment Effect (ATE)

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

Is ATE = $E[Y_i|A_i = 1] - E[Y_i|A_i = 0]$**?**

Identifying the ATE

	Y =1 (Patient Died)	Y = 0 (Patient Survived)
A = 1 (Patient Admitted to ICU)	100	25
A = 0 (Patient Not Admitted to ICU)	10	1000

Does ICU admission cause death?

Identifying the ATE

E[Death | Admitted] $= E[Y_i \mid A_i=1]$	75/100 = 75%
E[Death | Not Admitted] $= E[Y_i \mid A_i=0]$	100/1000 = 10%
$E[Y_i \mid A_i=1] - E[Y_i \mid A_i=0]$	**65%**

Is the **ATE** $= E[Y_i^{A_i=1}] - E[Y_i^{A_i=0}]$

$= E[Y_i \mid A_i=1] - E[Y_i \mid A_i=0]$? — **not always**

$E[Y_i = a \mid A_i = b]$ is an Association

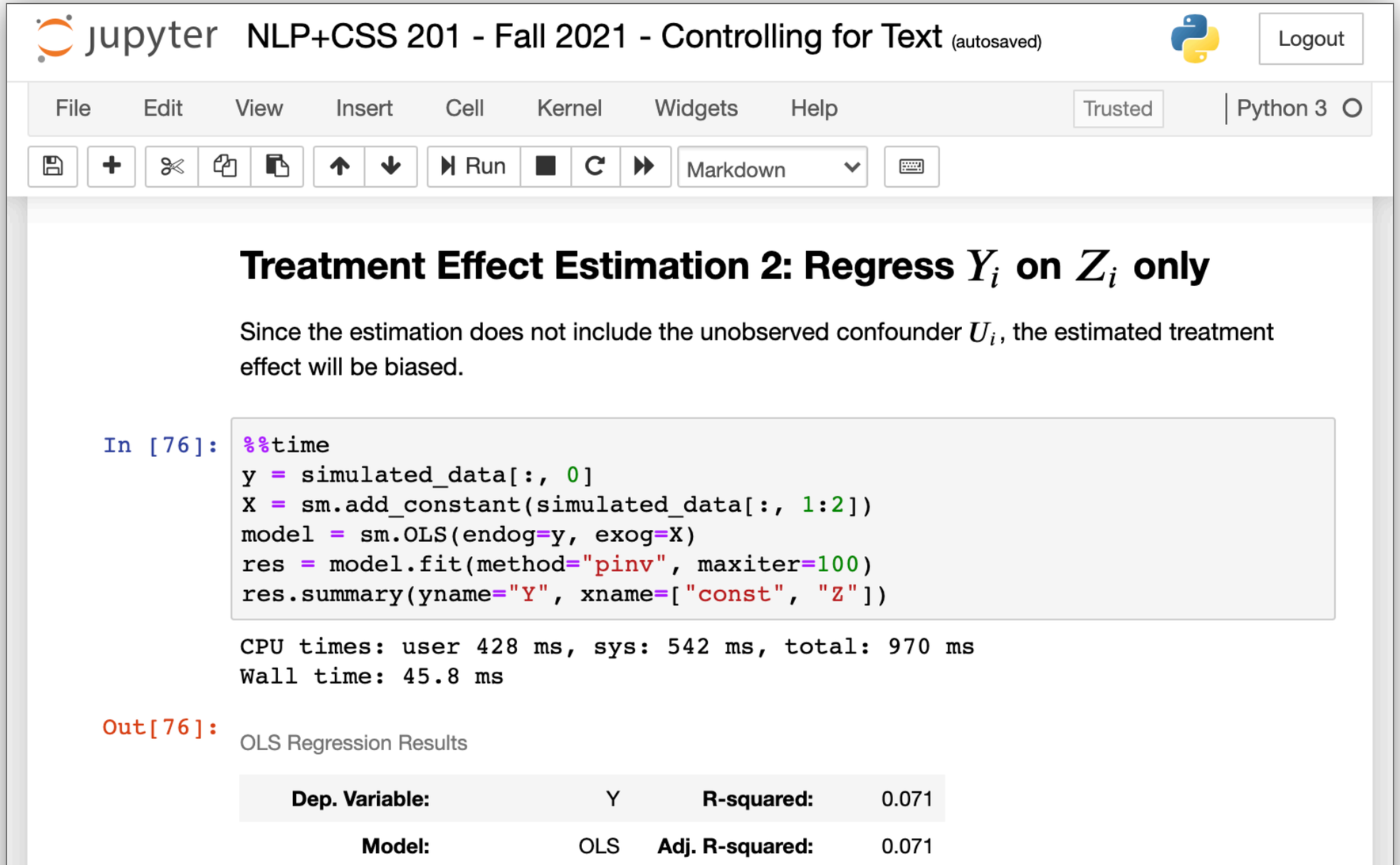
Association ≠ Causation under "Confounding"

Confounders are factors that are common causes of both the treatment the the outcome

Previous example: patient's age

Confounders are often unobserved!

Simulation of Cofounding Bias



The Magic of Randomized Experiments (RCTs)

The only perfect solution to confounding:
Randomly assign individuals to treatment actions

$$E[Y_i^{A_i=a}] = E[Y_i \mid A_i=a] \text{ if } Y_i^a \perp A_i \text{ (randomization)}$$

Issues: Expensive, infeasible, unethical (eg. randomly send patients to ICU)

Dealing with Observational Data

Control for / condition on observed confounders

For example, measure each patient's age, then compute ATE within each age bucket, average all the bucket-specific ATEs to get the overall ATE

Confounders are often unobserved!

Dealing with Observational Data

Use controls + natural /quasi-experiment

Example: Effect of Airbnb certification on booking rates

Popular methods: Regression discontinuity designs, difference-in-differences, instrumental variables

Controls + Randomization

Improve precision (smaller confidence intervals), deal with conditional randomization, endogenous selection bias, etc.

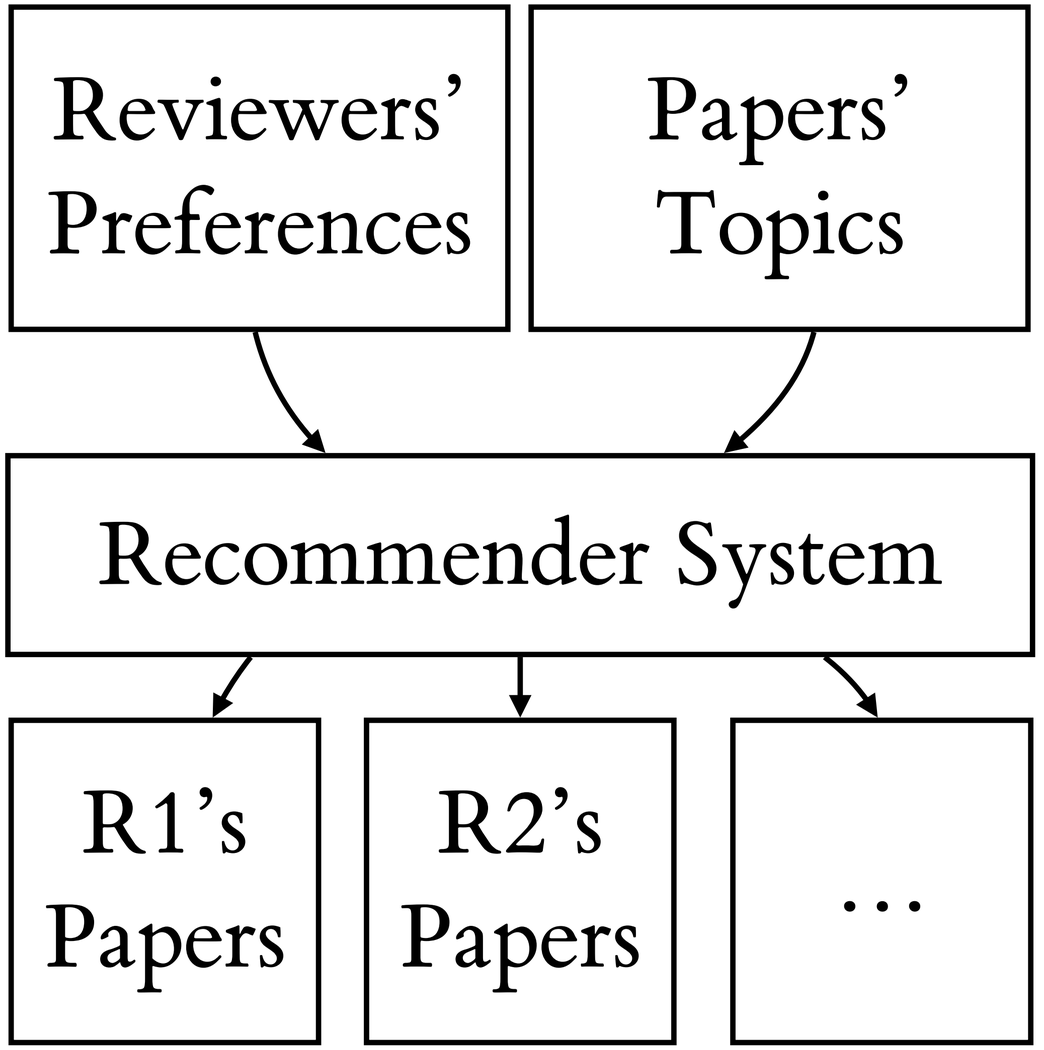
$$E[Y_i^{A_i=a} \mid X_i] = E[Y_i \mid A_i=a, X_i] \text{ if } Y_i^a \perp A_i \mid X_i$$

X_i is a scalar or vector of controls for individual i

Unstructured Text as a Control

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

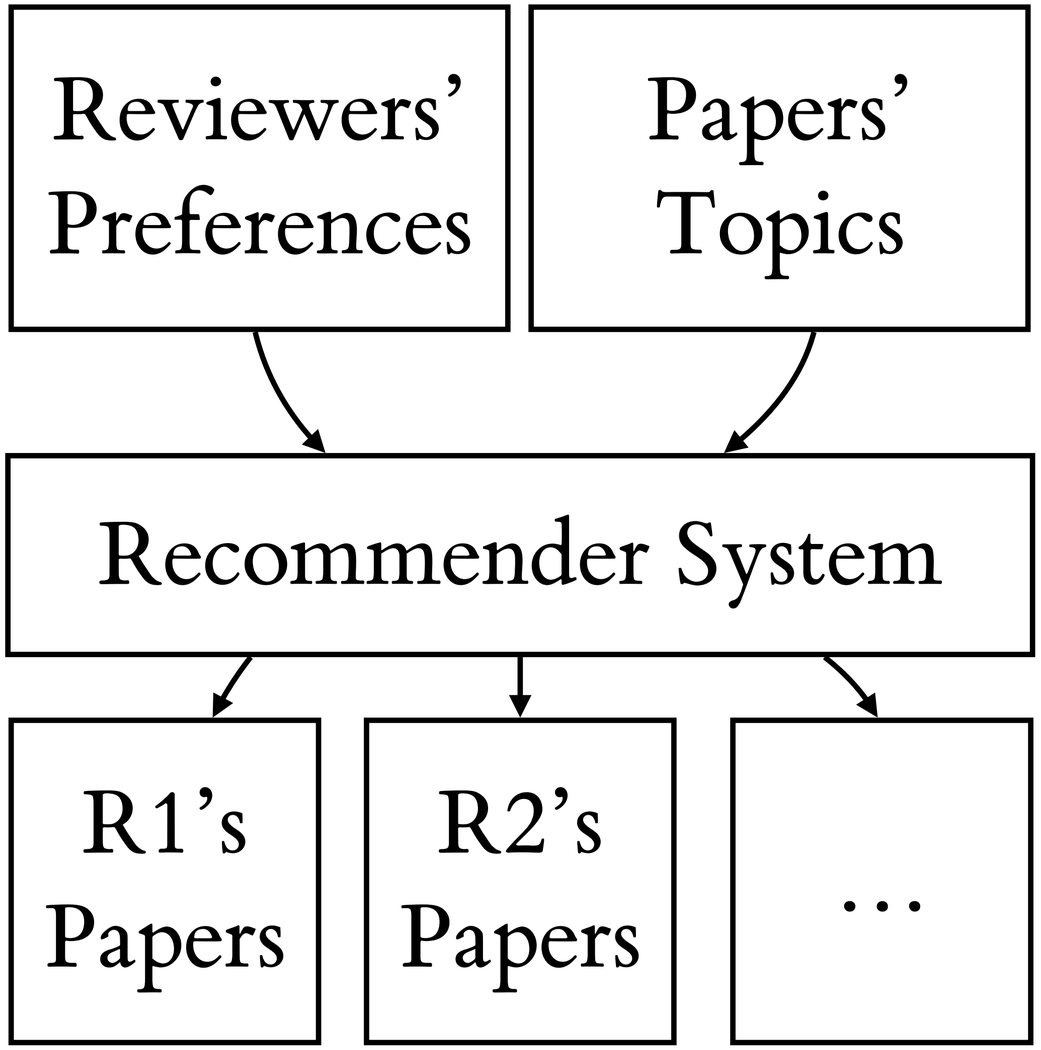


Unstructured Text as a Control

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

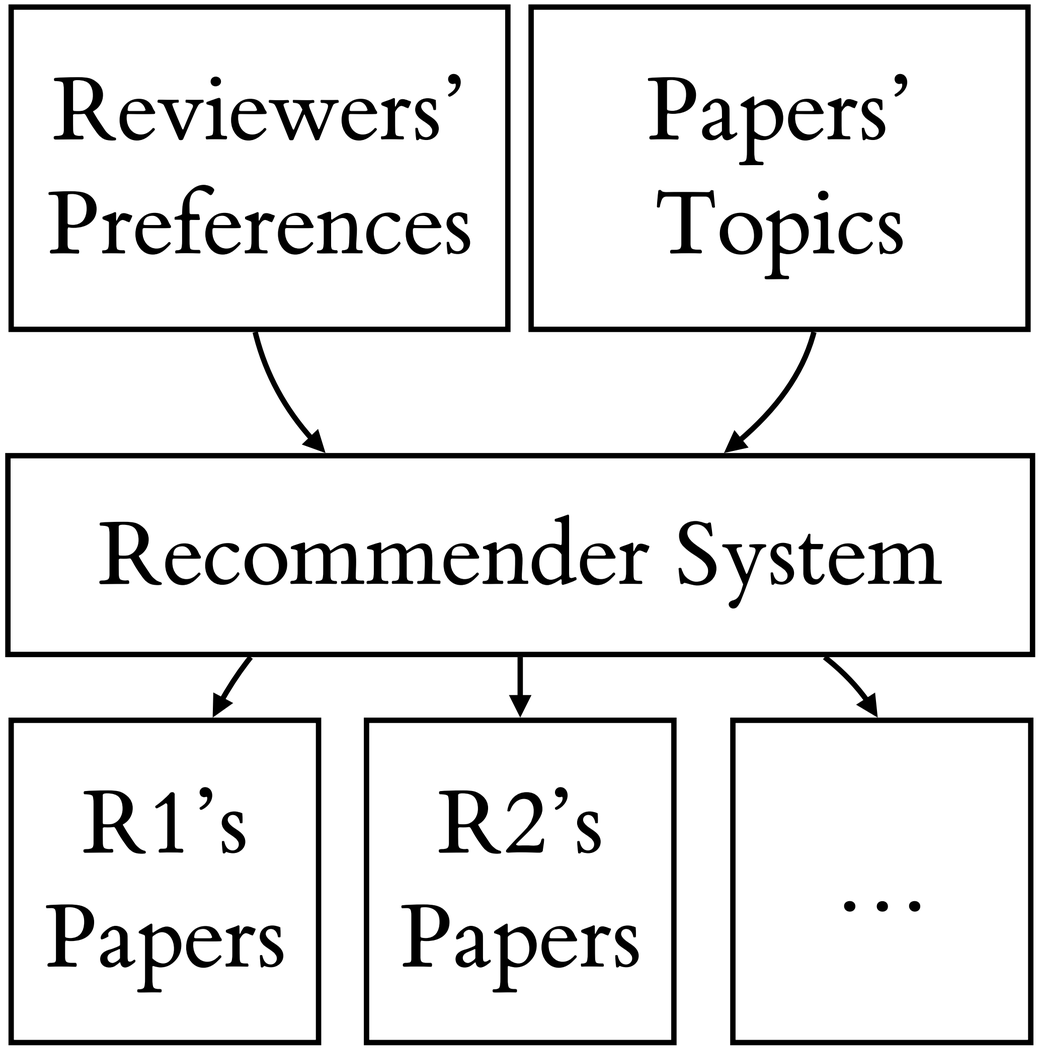


Unstructured Text as a Control

Would having a theorem improve a paper's rating?

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

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



Unstructured Text as a Control

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

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

Unstructured Text as a Control

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

Unstructured Text as a Control

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

Unstructured Text as a Control

More examples:

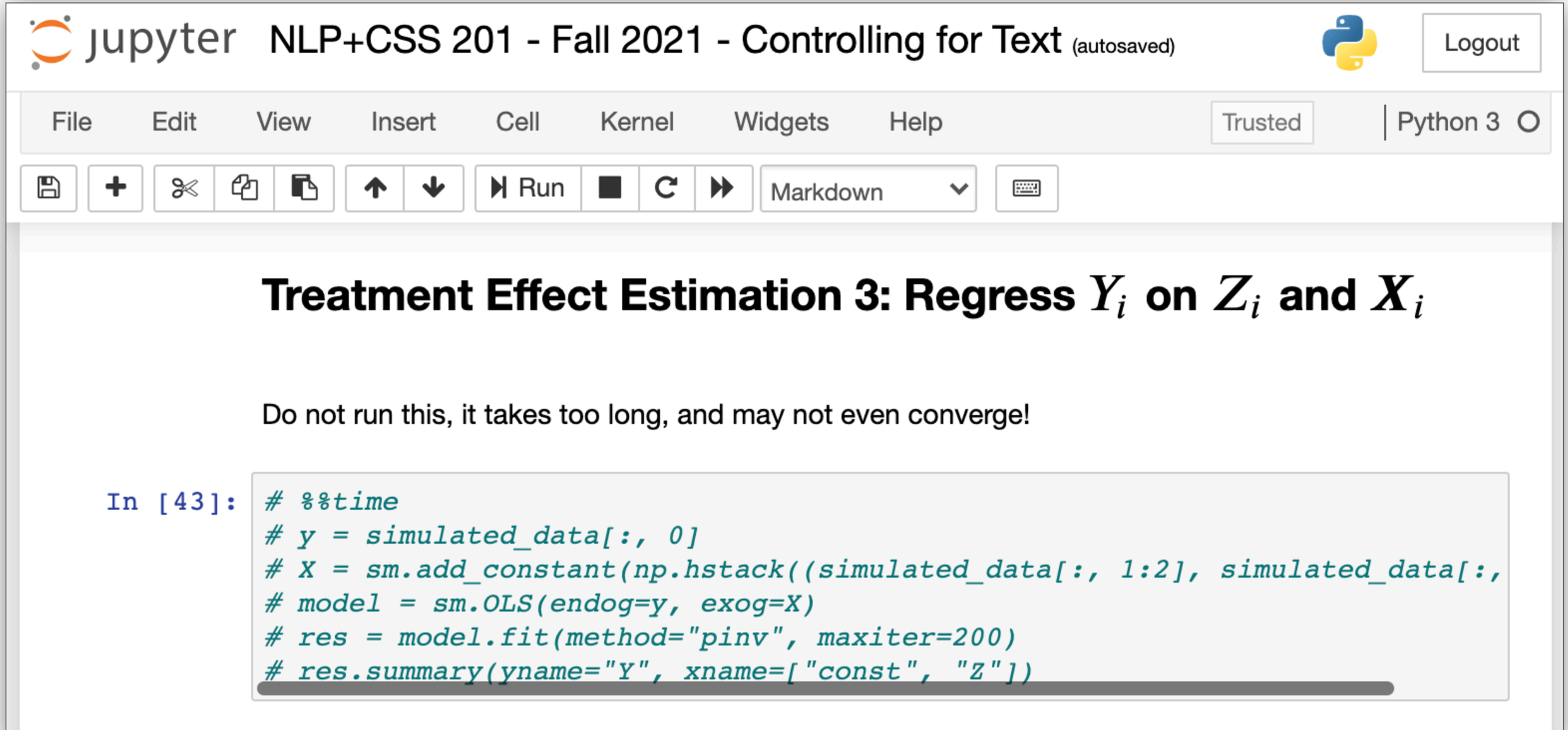
- **Does having higher reputation on a debating website make you more persuasive?** *(https://arxiv.org/abs/2006.00707)* Instrumental variable (quasi-experiment) for reputation — *need to control for argument text*
- *Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates (ACL 2020)*

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

First Attempt: Control for Words



Second Attempt: Control for Topics

jupyter NLP+CSS 201 - Fall 2021 - Controlling for Text (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

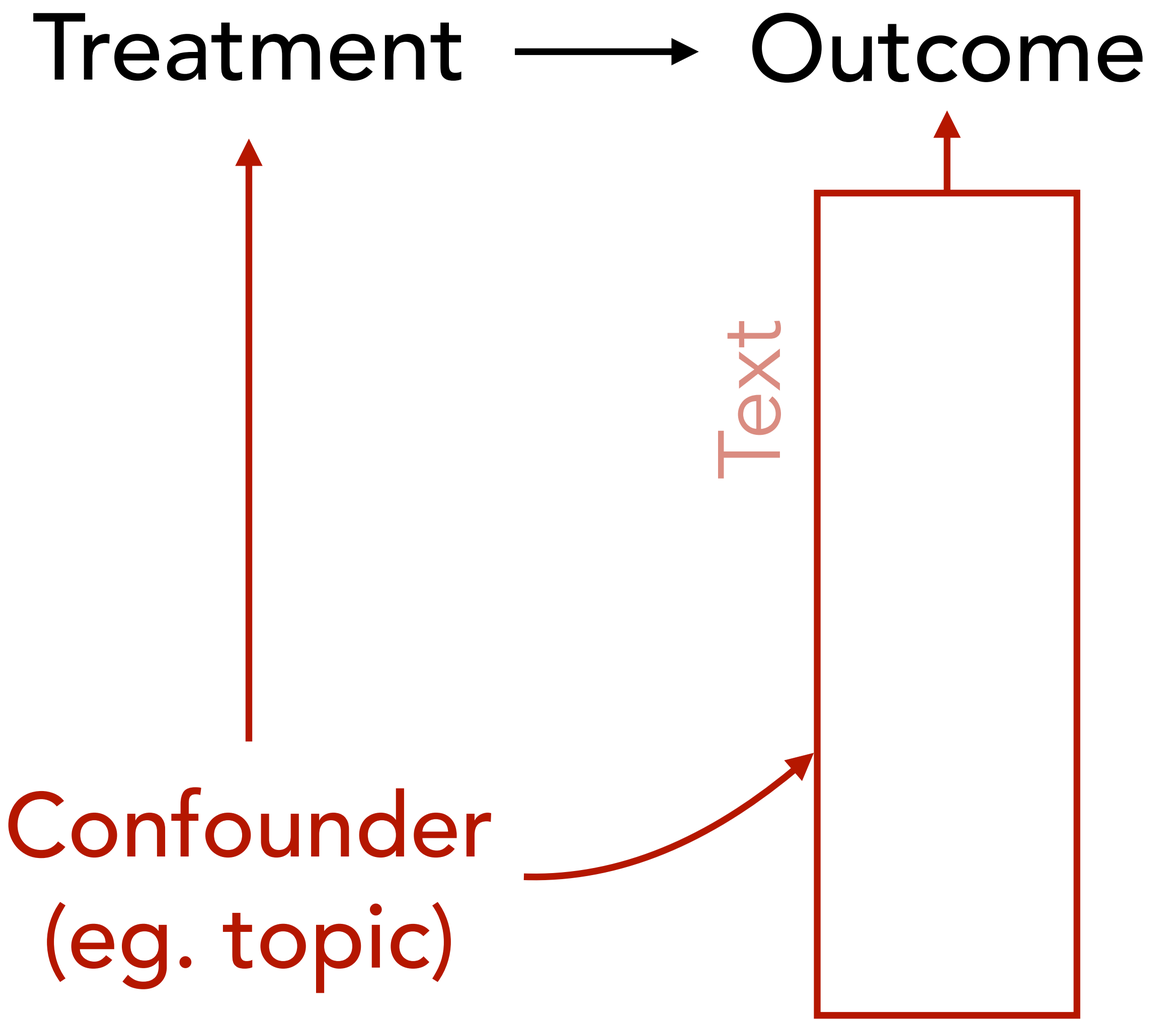
Treatment Effect Estimation 4: Regress Y_i on Z_i and Document-Topic Weights

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)
fidf_feature_names = vectorizer.get_feature_names()
rint_top_words(nmf, tfidf_feature_names, 10)
```

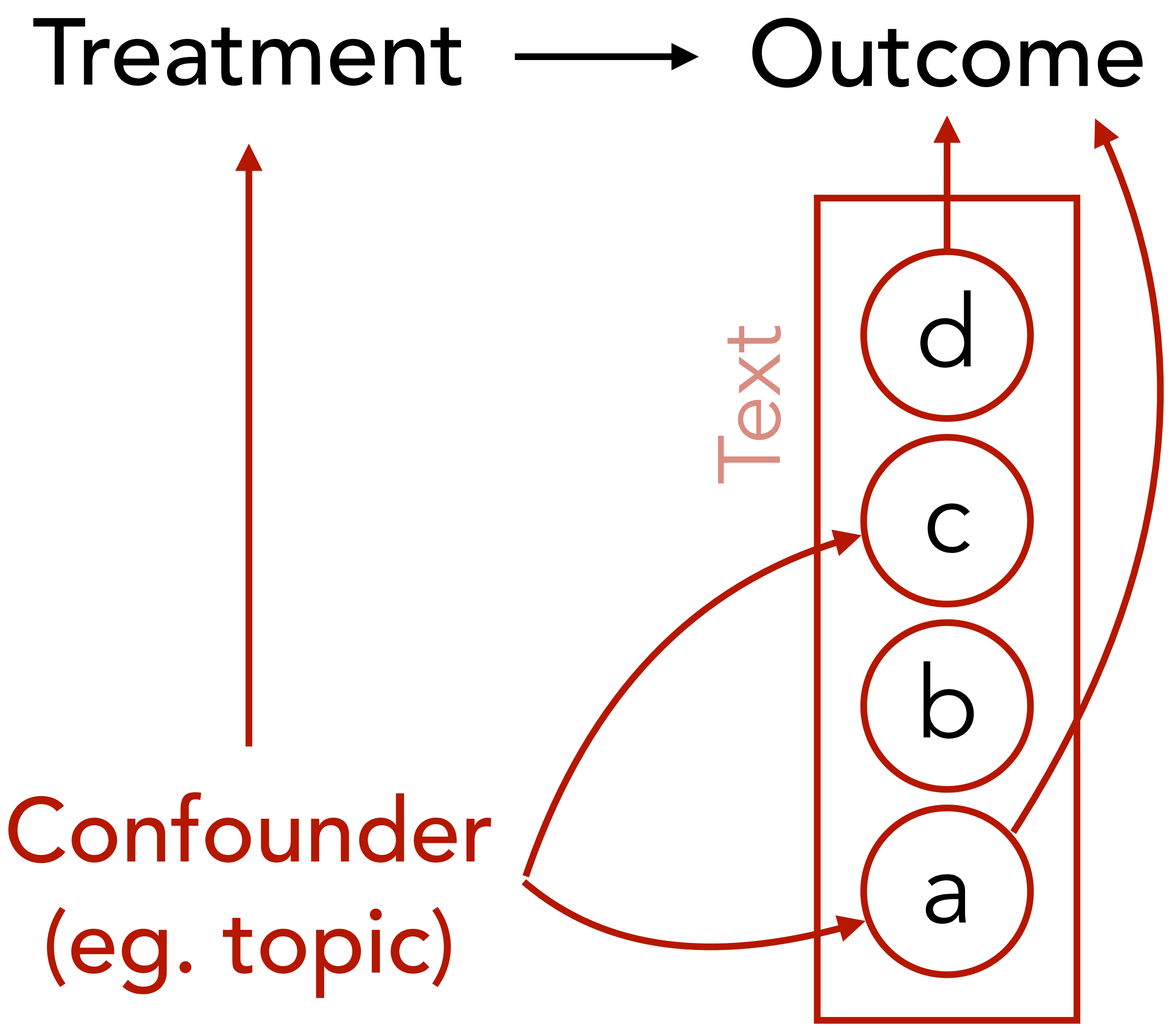
Controlling for Text Intuition



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

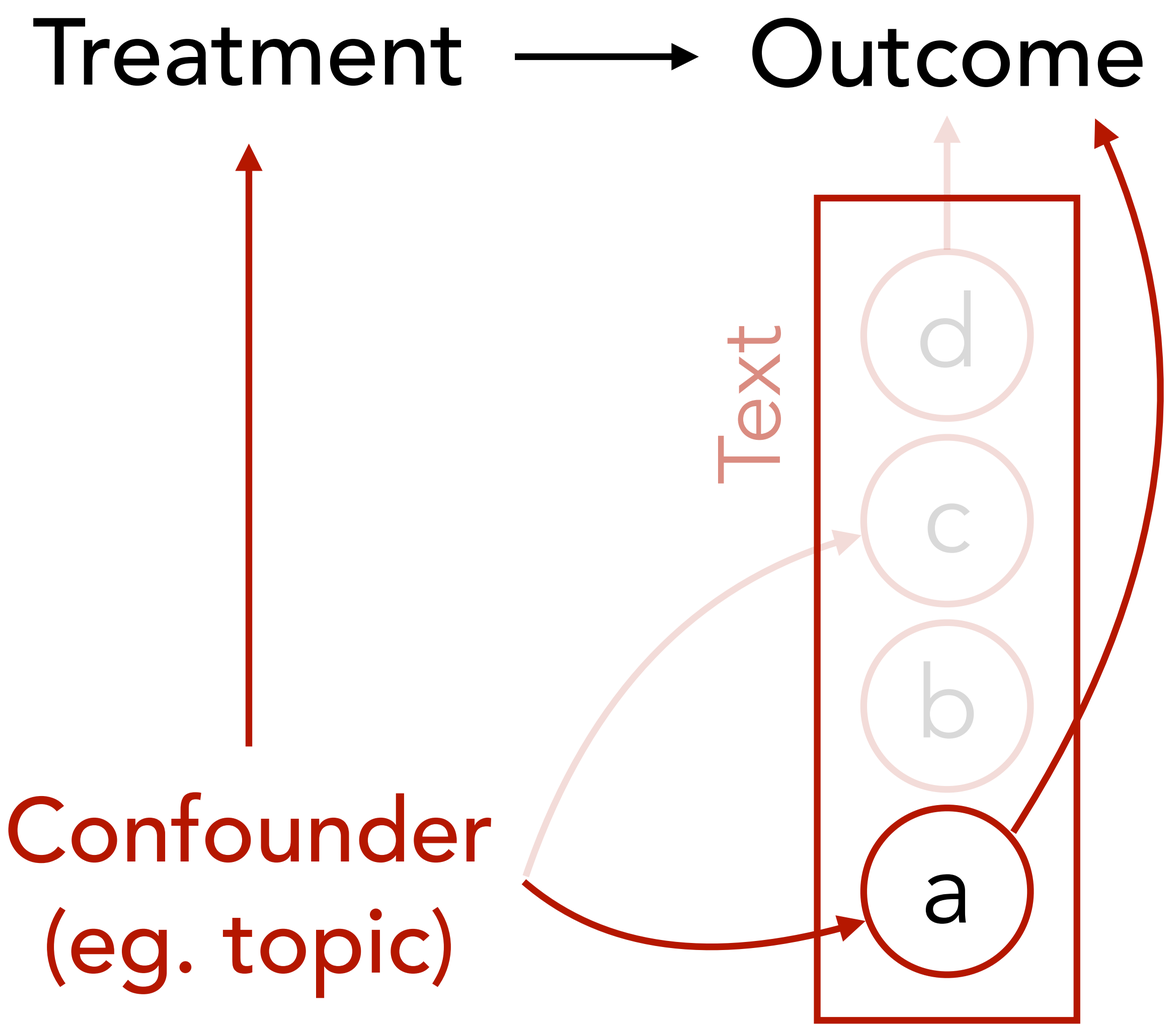
Recall: Confounder is common cause of treatment and outcome

Controlling for Text Intuition



Can view text as 4 logical components

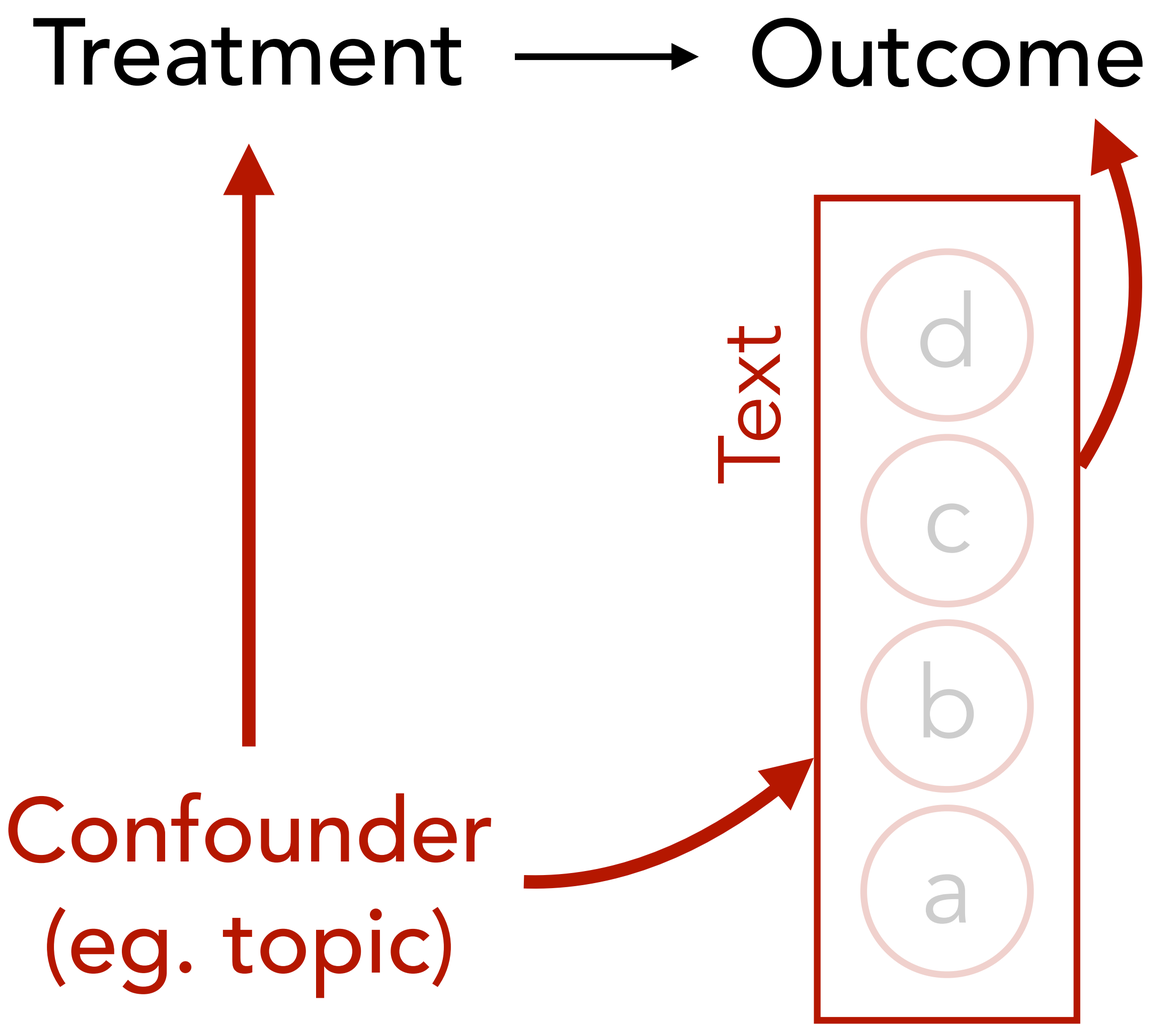
Controlling for Text Intuition



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

Needle in a haystack

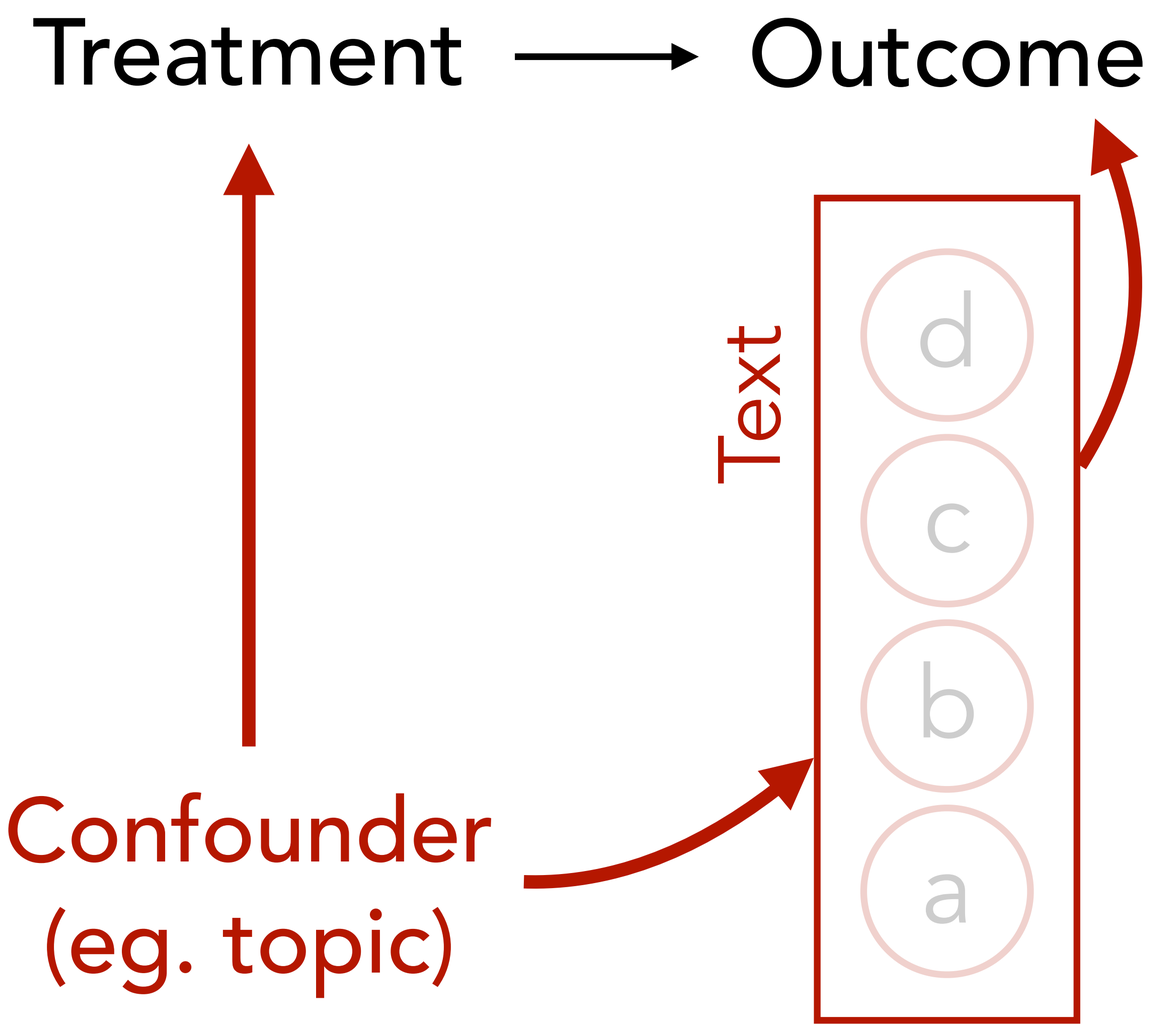
Controlling for Text Intuition



Alternative to finding this needle without using dimensionality reduction

Measure and combine correlation between text, treatment, and outcome

Controlling for Text Intuition



Measuring correlations

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

Double Machine Learning

[Chernozhukov et. al. 2016]

- General "recipe" to perform statistically valid estimation/inference after incorporating ML models
- Neutralizes "regularization bias" and "overfitting bias" that arise from ML model estimation
- Fast $O(\sqrt{n})$ convergence rates despite slowly-converging nonparametric ML models

Double Machine Learning

[Chernozhukov et. al. 2016]

Partially Linear Regression Model

$$r_{ij} = \theta_0 + \theta_1 A_j + f(\mathbf{text}_j) + \epsilon_i$$

$\mathbf{ATE_i}$

Unknown function

Double Machine Learning

[Chernozhukov et. al. 2016]

Special moment condition combines correlations between text, treatment, and outcome

Outcome Treatment

$$\mathbb{E}[(r_{ij} - \mathbb{E}[r_{ij} \mid \text{text}_j] - \beta_1(A_j - \mathbb{E}[A_j \mid \text{text}_j])) \times (A_j - \mathbb{E}[A_j \mid \text{text}_j])] = 0$$

Double Machine Learning

[Chernozhukov et. al. 2016]

Double ML Procedure

1. Construct a Neyman-orthogonal moment condition for the regression equation
2. Empirically solve for the coefficients

In practice: Can be done using a sequence of ordinary-least-squares regressions

Double Machine Learning

[Chernozhukov et. al. 2016]

1. Compute the prediction errors of treatment from text, outcome from text
2. Regress the outcome prediction error on the treatment prediction error

Text prediction models must be trained on a held out subset of the data (called sample-splitting)

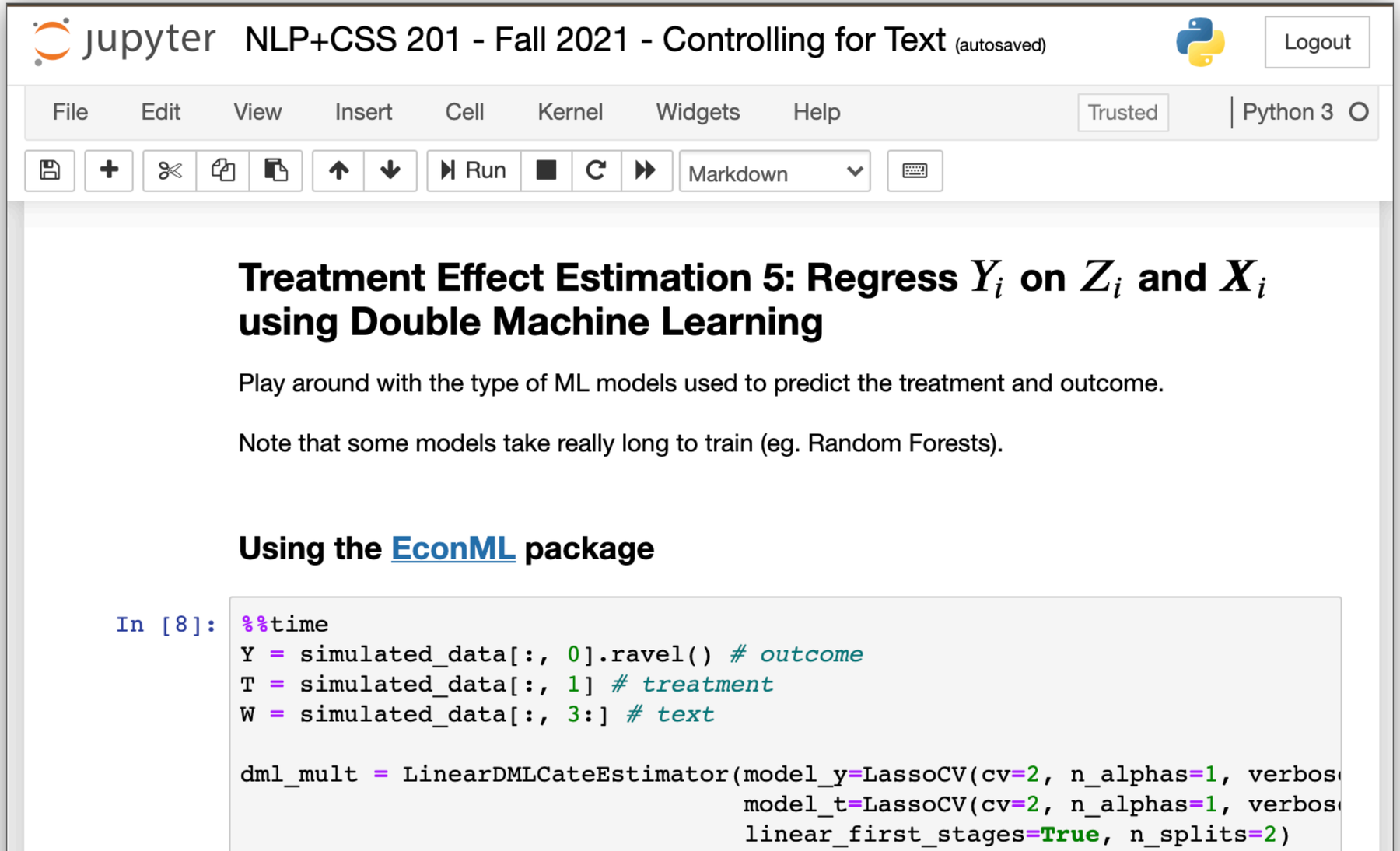
Double Machine Learning

[Chernozhukov et. al. 2016]

Double ML Theoretical Details

Check out Chris Felton's slides: *https://scholar.princeton.edu/sites/default/files/bstewart/files/chern.handout.pdf*

Double ML + Text Examples



What We've Learned Today

1. ITEs, ATEs, randomized experiments
2. Observational data and confounding bias
3. Double machine learning
4. Controlling for text with double machine learning

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 and colleagues)**: Approach worth exploring

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*
- **Recommended Books**: *Causal Inference: What If* **(free)** and *Mostly Harmless Econometrics*