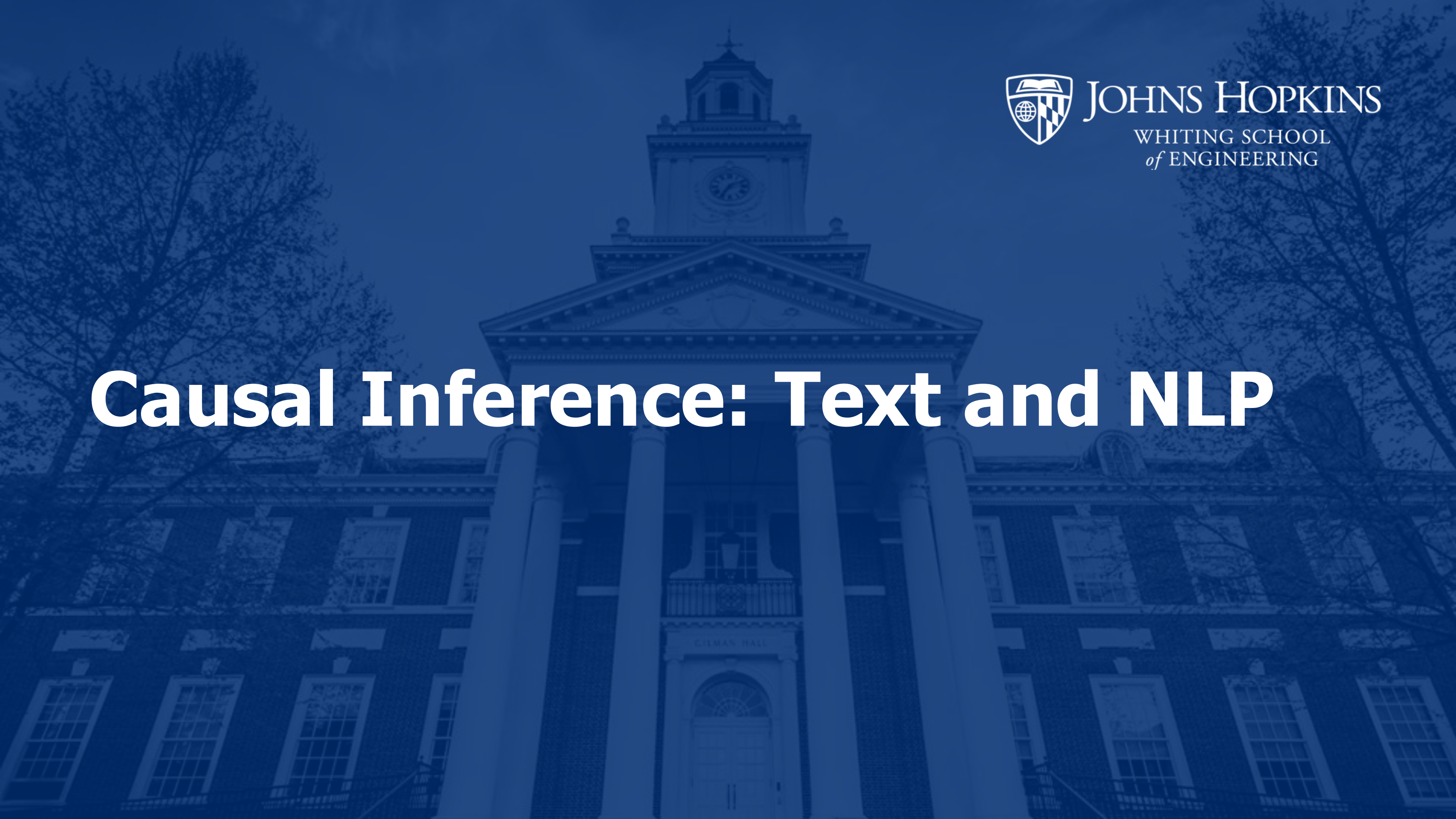
JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Causal Inference: Text and NLP



Logistics

- HW 3 on causal inference has been released
  - Deadline extended to Friday
- Midterm Exam
  - In class next Wednesday
  - Includes all material through Wednesday 3/6 (including Wednesday's guest lecture and homeworks)
  - Sample problems released on Piazza
  - Review session Monday 3/11

Recap

- Methods for adjusting for confounders
  - Regression
  - Matching
  - Propensity scores (matching, weighting, and stratification)

Today:

- Additional notes about when to do adjustments
- Causal inference with text
  - Overview
  - Adjusting for text as confounders (or mediators)
  - Drawing from causal inference to improve NLP models

Some additional notes

Double Machine Learning

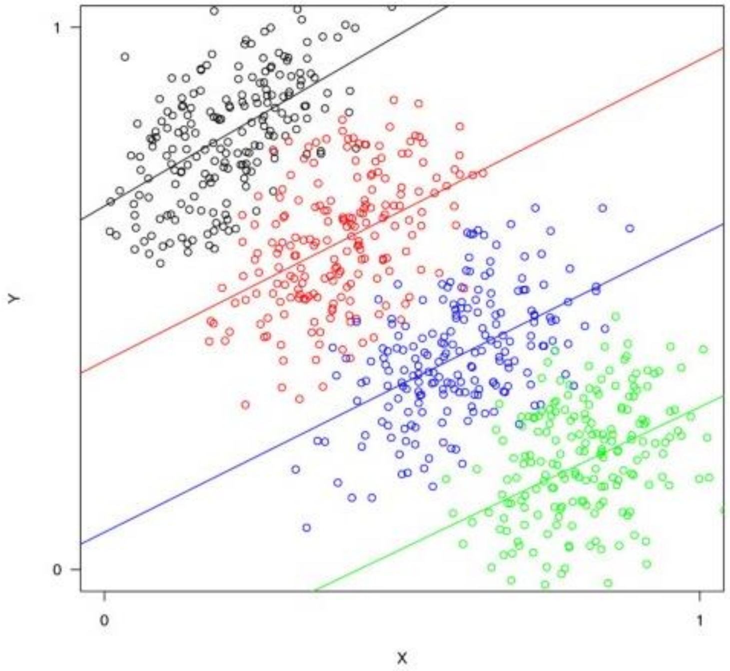
- General framework for estimating causal effects using ML (random forests, lasso or post-lasso, neural nets, boosted regression trees, and various hybrids and ensembles of these methods)

- Available in Python and R packages:
  - https://github.com/DoubleML

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins, Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68, https://doi.org/10.1111/ectj.12097

Mixed Effects Regression Models

- We discussed regression adjustment for confounders
- When data is hierarchical / non-independent we need a better regression model
- E.g. you examine if dosage of medicine affects fevers
- Your data is from hospitals in different countries where underlying health conditions that affect baseline health
- Recall Simpson's Paradox



- Data looks negatively correlated overall
- Subsetting data shows positive correlations

Mixed Effects Regression Models

- We can account for differences across subgroups by allowing subgroups to have different parameters (e.g. different intercepts in linear regression)
- Subgroup is a *random* effect
- Dosage is a *fixed* effect

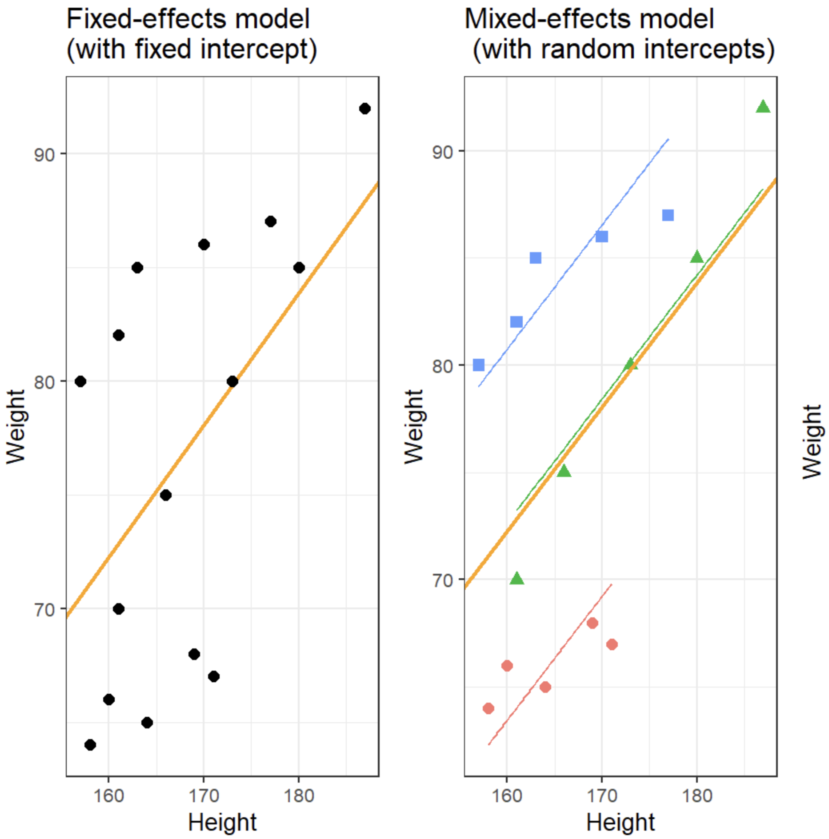
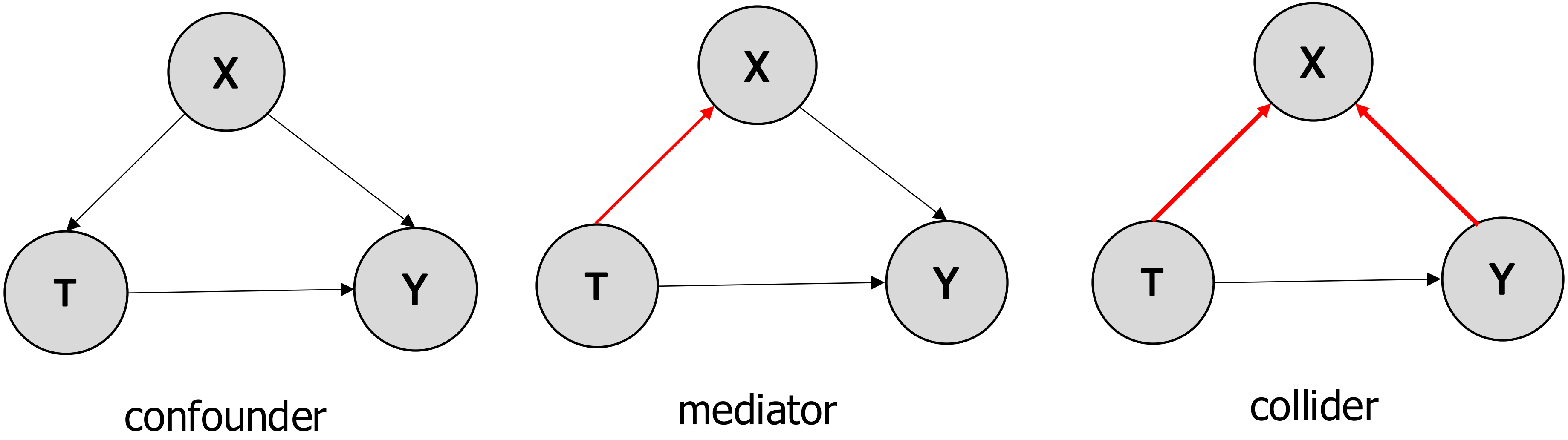
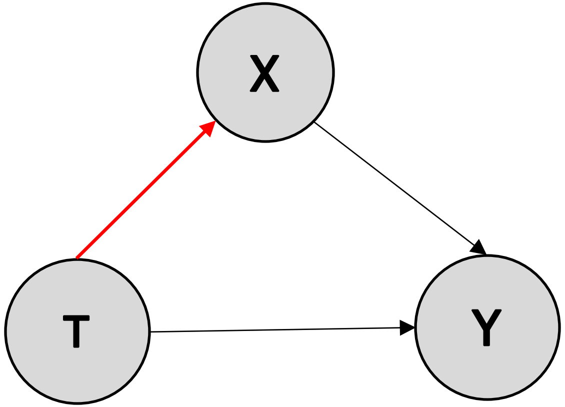


Image credit: https://ladal.edu.au/regression.html

Confounders vs. Mediators vs. Colliders



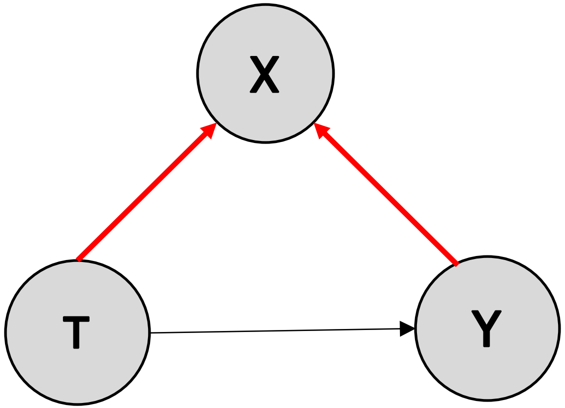
Confounders vs. Mediators vs. Colliders



mediator

- Example:
  - Estimating if gender has an effect on social media likes
  - Gender (T) influences the topic of posts (X)
  - Topic of posts (X) and gender (T) influence number of likes (Y)
- If we adjust for X, we may be removing some of the effect
- We may still choose to adjust for X if we specifically want to capture the direct effect and not the indirect effect
- We may want to separate out direct and indirect effects in a mediation analysis

Confounders vs. Mediators vs. Colliders



collider

- Example:
  - Studying if getting a dog makes people wake up earlier
  - Getting dog (T) influences wake up time (Y) and if you take morning walks (X)
  - People who happen to wake up early (Y) take morning walks too (X)
  - If you condition on X (e.g. restrict data to people who take morning walks), you're selecting for people who wake up early in your control group → you find that having a dog makes you get up later
- If we adjust for X, we are adding bias to our estimator!

Causal Inference in text: Overview

Johns Hopkins
Whiting School of Engineering

What characteristics distinguish text from other data types?

- Text is high dimensional
  - Overfitting, violations of positivity

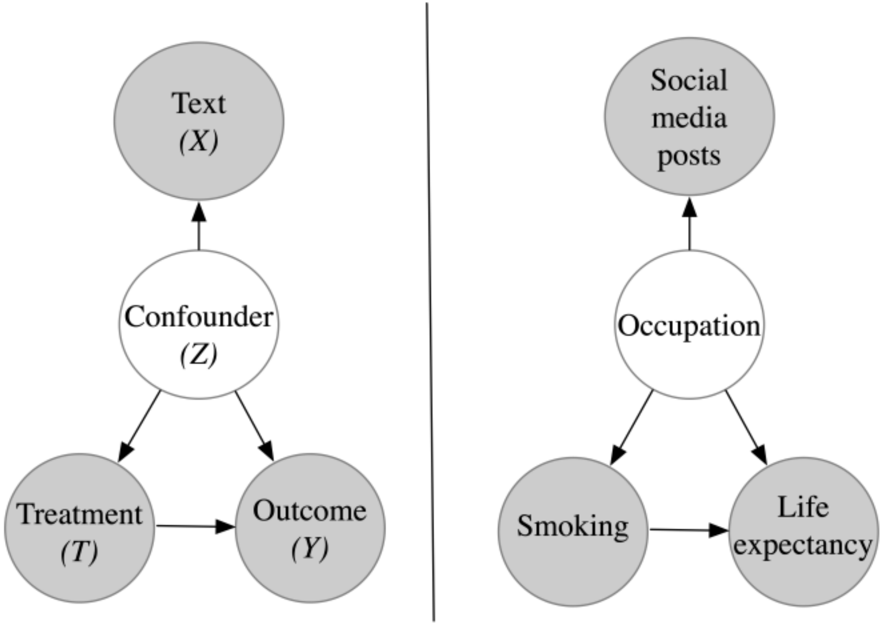
- Compared to other high dimensional data:
  - Text can be read and evaluated by humans
  - Designing meaningful representations of text is an open problem

Keith, Katherine, David Jensen, and Brendan O'Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL. 2020.

Text as confounders

- Text data could either:
- (a) serve as a surrogate for potential confounders
- (b) the language of text itself could be a confounder

Example: the linguistic content of social media posts (confounder) could influence censorship (treatment) and future posting rates (outcome)



Keith, Katherine, David Jensen, and Brendan O'Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL. 2020.

Text as treatment or outcome

- Do Wikipedia articles contain gender bias?
  - Treatment: Perceived gender
  - Outcome: Article text
  - Confounders/Mediators: Perceived characteristics other than gender

- Does a celebrity's social media posts cause them to gain followers?
  - Treatment: The social media posts
  - Outcome: Follower counts
  - Confounders/Mediators: Changes in social media usage, current events

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Adjusting for text as confounders

Two similar approaches

- Topic Inverse Regression Matching
  - Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.
- “Causally sufficient” embeddings
  - Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.

Adjusting for text as confounders: Topic Inverse Regression Matching

- Key ideas:
  - Matching (remember: direct or propensity) is a good approach for adjusting for text as confounder because analysts can manually evaluate the quality of the adjustment by comparing the matched treatment and control text
  - Most use cases what we need to match on are topics (as opposed to sentiment, punctuation, word order, etc). We also may care about individual words
  - We need to match on aspects of the text that are predictive of treatment (definition of confounders)

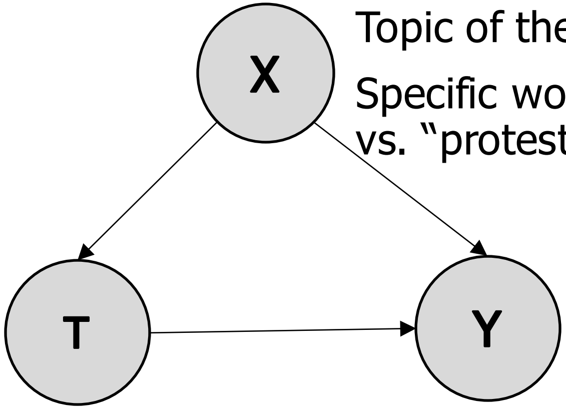
Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.

Example application: Effects of censorship in Chinese social media

- Research questions:
  - 1. "Is censorship completely determined by the text of a particular post, or does censorship become more targeted toward users based on their previous censorship history?"
  - 2. Does having a post censored cause people to post less in the future?

Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.

Example application: Effects of censorship in Chinese social media



Topic of the original post (public gathering)

Specific words in the original post (“parade” vs. “protest”)

Post is censored or not

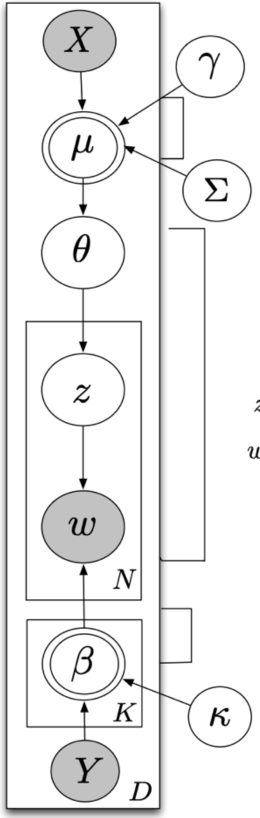
1. Future censorship rate
2. Future posting rate

Topic Inverse Regression Matching using STM

- For bag-of-words representation *W*, define a function *g(W)* to create a low-dimensional estimate that captures topic and word differences that relate to treatment assignment

- Primary model for text representations*: structured topic model (STM)*

- LDA-style topic model that allows flexible inclusion of covariates



Topic Prevalence:

$$\mu_{d,k} = X_d \gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

Language Model:

$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^{k} \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.

Step	Rationale
1. Estimate a structural topic model including the treatment vector as a content covariate.	Reduces dimension of the text
2. Extract each document's topics calculated as though treated (part of $g(\boldsymbol{W})$).	Ensures semantic similarity of matched texts
3. Extract each document's projection onto the treatment variable (part of $g(\boldsymbol{W})$).	Ensures similar treatment probability of matched texts
4. Use a low-dimensional matching method to match on $g(\boldsymbol{W})$ and estimate treatment effects using the matched sample.	Standardizes matching

Example application: Effects of Censorship on Chinese social media

- Research questions:
  - 1. "Is censorship completely determined by the text of a particular post, or does censorship become more targeted toward users based on their previous censorship history?"
  - 2. Does having a post censored cause people to post less in the future?

- Methods:
  - Use TIRM to identify pairs of nearly identical social media posts written by nearly identical users, where one is censored and the other is not
  - Examine subsequent posting and censorship rates of each user

Example application: Effects of Censorship on Chinese social media

- Results:
  - Having a post censored increases the probability of future censorship significantly
  - It does not decrease number of future posts by the censored user

- Conclusions:
  - Option 1: algorithmic targeting of censorship, where social media users are more likely to be censored after censorship because they are flagged by the censors
  - Option 2: social media users may chafe against censorship and respond by posting increasingly sensitive content that is more likely to be censored

A different method: develop “causally sufficient” text embeddings

- Text is high dimensional and data is finite: difficult to fit models directly to text
- Instead, “reduce the text to a low-dimensional representation that suffices for causal identification and enables efficient estimation from finite data.”

- Two key ideas:
  - Supervised dimensionality reduction: we don’t need the full text, causal inference only requires the parts of text that are predictive of the treatment and outcome
  - Efficient language modeling: design representations of text to dispose of “linguistically irrelevant information”, presumed to also be “causally irrelevant”

Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.

General approach: develop “causally sufficient” text embeddings

- Start with a language model (BERT) and modify it to produce 3 outputs:
  - 1) document-level embeddings
  - 2) a map from the embeddings to treatment probability
  - 3) a map from the embeddings to expected outcomes for the treated and untreated
  - [(2) and (3) are small added neural networks on the original model]

- [They also do a variant based on a topic model]

General approach: develop "causally sufficient" text embeddings

- Train model to predict outcome, treatment, and with language-modeling objective (e.g. to learn meaningful text representations)

$$
\begin{aligned}
L(\mathbf{w}_i;\xi,\gamma) =& \left(y_i - \tilde{Q}(t_i,\lambda_i;\gamma)\right)^2 &\longrightarrow \text{Outcome} \\
&+ \text{CrossEnt}\left(t_i,\tilde{g}(\lambda_i;\gamma)\right) &\longrightarrow \text{Treatment} \\
&+ L_{\text{U}}(\mathbf{w}_i;\xi,\gamma). &\longrightarrow \text{Language modeling}
\end{aligned}
$$

- To compute average treatment effect, plug estimated embeddings, propensity scores, and conditional outcomes into a downstream estimator

Evaluation

- Two settings:
  - Peer-reviewed journal articles: Causal effect of including a theorem on paper acceptance.
    - Treatment: the word "theorem" occurs in the paper
    - Confounder: article abstract (subject of the paper)
    - Outcome: accept/reject
  - Effect of gender on Reddit popularity
    - Treatment: "male" label
    - Mediator: Post text (topic or style)
    - Outcome: Popularity score

How can we use this data for *evaluation* rather than analysis?

Evaluations

- Simulated data:
  - Use real confounders and treatments
  - Simulate outcomes (so we know the “true” causal effect)
- Their findings:
  - 1) Yes, language modeling helps recover simulated effects
  - 2) Yes, supervised dimensionality helps
  - 3) Their proposed models C-BERT and C-ATM outperform alternatives

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Drawing from Causal Inference to Improve NLP models

Drawing from Causal Inference to Improve NLP models

- ML in general typically captures associates, not causal effects
- Models are prone to overfitting, exploit spurious correlations in the data
  - E.g. train a model to identify photos of dogs from cats; Model learns that dogs always have collars

→ "DOG"

Drawing from Causal Inference to Improve NLP models

- ML in general typically captures associates, not causal effects
- Models are prone to overfitting, exploit spurious correlations in the data
  - E.g. train a model to identify photos of dogs from cats; Model learns that dogs always have collars

- Maybe by drawing from causal inference we can train models to ignore these spurious correlations, especially for tasks where it's hard to collect good training data
- Case study: drawing from causal inference to detect *subtle gender bias*

Field, Anjalie, and Yulia Tsvetkov. "Unsupervised Discovery of Implicit Gender Bias." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Need to develop new models

- Our goal: detect subtle gender biases like microaggressions, objectifications, and condescension in 2nd-person text

  - “Oh, you work at an office? I bet you’re a secretary”
  - “Total tangent I know, but you’re gorgeous”

- Current classifiers that detect hate speech, offensive language, or negative sentiment cannot detect these comments

- [Note: focus on binary gender]

Naive Approach: Supervised Classification

I like Bob, but you're hot, so kick his butt

Like · Reply ·

Thanks so much **Ma'am**!

Like · Reply ·

I'd vote for you if I lived in **Massachusetts**

Like · Reply ·

...a good way to celebrate **Title IX**, too!

Like · Reply ·

→ amazon mechanical turk

appen →

Supervised Classifier

Naive Approach: Supervised Classification

I like Bob, but you're hot, so kick his butt

Like · Reply ·

Thanks so much **Ma'am**!

Like · Reply ·

I'd vote for you if I lived in **Massachusetts**

Like · Reply ·

...a good way to celebrate **Title IX**, too!

Like · Reply ·

amazon mechanical turk

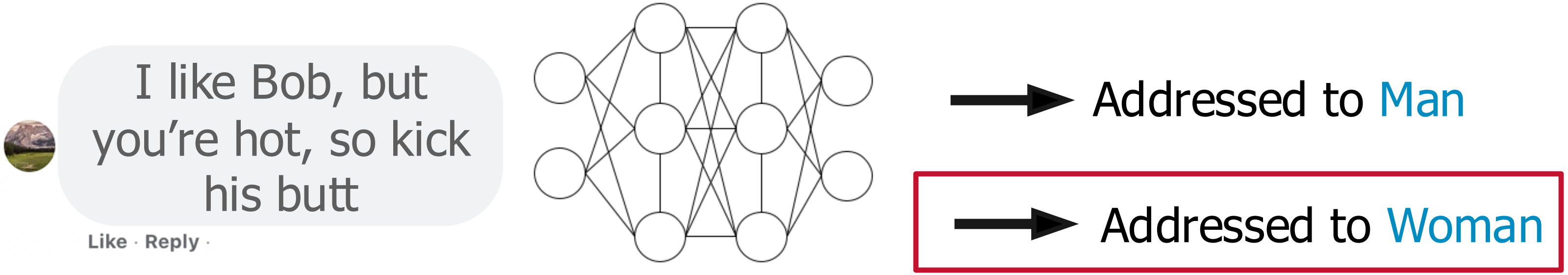
appen

Supervised Classifier

Problem: Biases are *subtle*, *implicit*, and *context-dependent*

Proposed approach: Comments contain gender bias if they are highly predictive of gender

- Train a classifier that predicts the gender of the person the text is addressed to
- If the classifier makes a prediction with high confidence, the text likely contains bias



If a comment is very likely to be addressed to a woman, and is very unlikely to be addressed to a man, it probably contains gender bias.

Challenge: Text main contain *confounds* that are predictive of gender, but not indicative of gender bias

I like Bob, but you're hot, so kick his butt

Like · Reply ·

Thanks so much **Ma'am**!

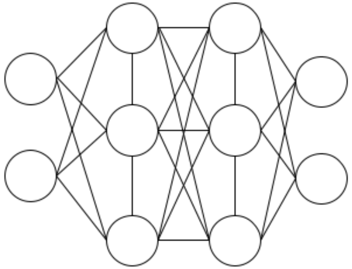
Like · Reply ·

I'd vote for you if I lived in **Massachusetts**

Like · Reply ·

…a good way to celebrate **Title IX**, too!

Like · Reply ·



→ Addressed to Woman

→ Addressed to Woman

→ Addressed to Woman

→ Addressed to Woman

Challenge: Text main contain *confounds* that are predictive of gender, but not indicative of gender bias

- Overtly gendered words
- Preceding context in the conversation
- Traits of people (other than gender) in the conversation

Saturday is the 40th anniversary of **Title IX**…

Like · Reply ·

…a good way to celebrate Title IX, too!

Like · Reply ·

I'd vote for you if I lived in Massachusetts

Like · Reply ·

Bob and I join Bill Hemmer on America's Newsroom to discuss whether or not...

Like · Reply ·

I like Bob, but you're hot, so kick his butt

Like · Reply ·

Thanks so much Ma'am!

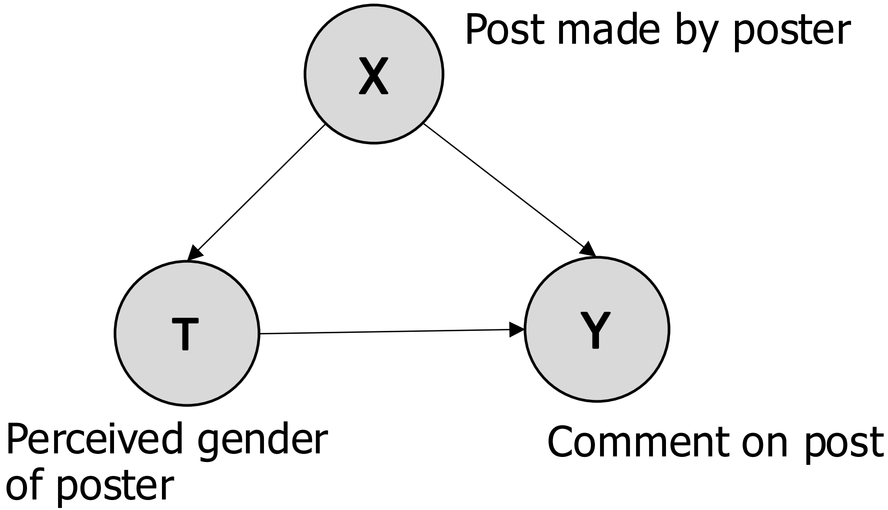
Like · Reply ·

A note on causal set-up

- We’re not really doing causal inference: we are trying to build a classifier to detect microaggressions, not draw conclusions about the state of the world:
  - “confounds”: spurious correlations in our data (not necessarily “confounders”)

- Some of these factors that we don’t want the model to learn are confounding variables

A note on causal set-up



[Note: we have text as an outcome and as a confounder]

Preceding context is an *observed* confounding variables

Writer_Gender: F

Saturday is the 40th anniversary of **Title IX**! I'm celebrating with a Sat morning run - ladies please respond below if you want to join

Like · Reply ·

Wish I could ! Already have plans for a bike ride and breakfast with some awesome ladies - a good way to celebrate **Title IX**, too!

Like · Reply ·

Would love to!

Like · Reply ·

Writer_Gender: M

Any deal with **Iran** — a nation that the United States cut off diplomatic ties with 35 years ago — must protect America's interests at home and abroad.

Like · Reply ·

Iran might be a free, democratic nation today, if not for decades of American interference.

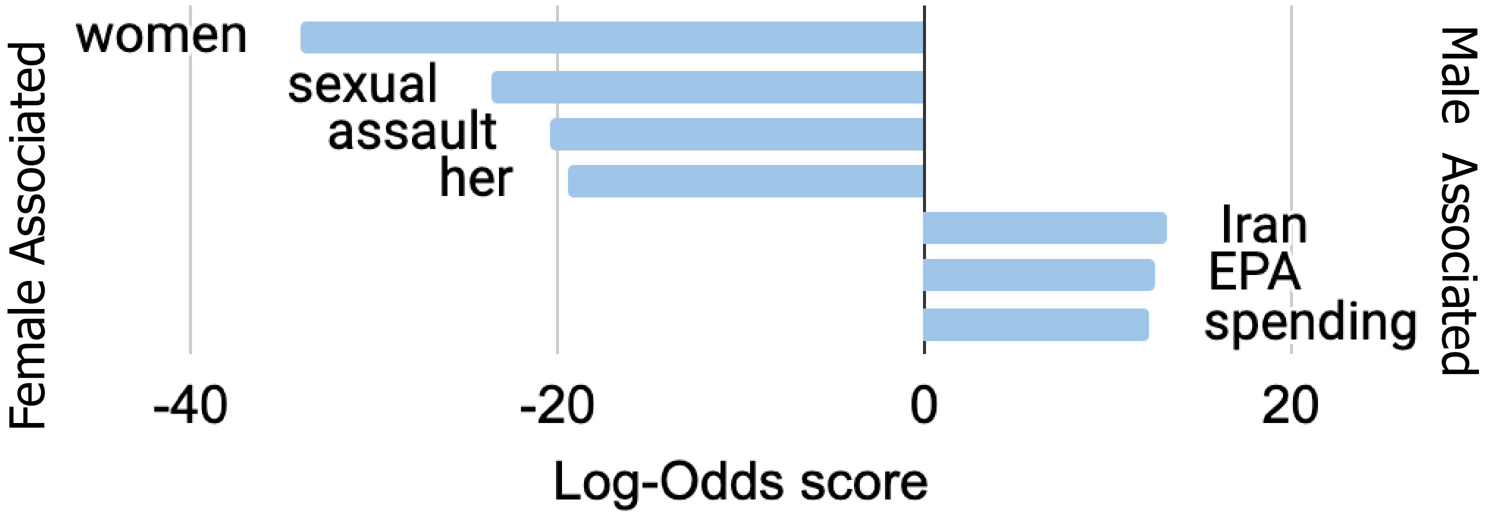
Like · Reply ·

That's for sure! Worst deal he could make! We can't trust **Iran** and America knows it !!!!!

Like · Reply ·

Key problem: Men and women post different content, which is reflected in their replies

Preceding context is an *observed* confounding variables



Propensity matching for *observed* confounding variables

Writer_Gender: F

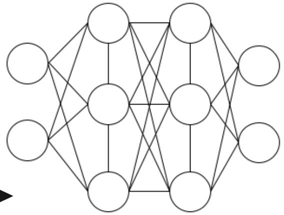
~~Saturday is the 40th anniversary of **Title IX**! I'm celebrating with a Sat morning run - ladies please respond below if you want to join~~

Writer_Gender: M

Any deal with **Iran** — a nation that the United States cut off diplomatic ties with 35 years ago — must protect America's interests at home and abroad.

Writer_Gender: F

My overriding concern is whether or not the agreement is in the national security interest of the United States. **Iran** must be blocked from proceeding any further towards developing a nuclear weapon.



Text classifier to predict WRITER_GENDER

$|e_i - e_l| \geq c \forall l$

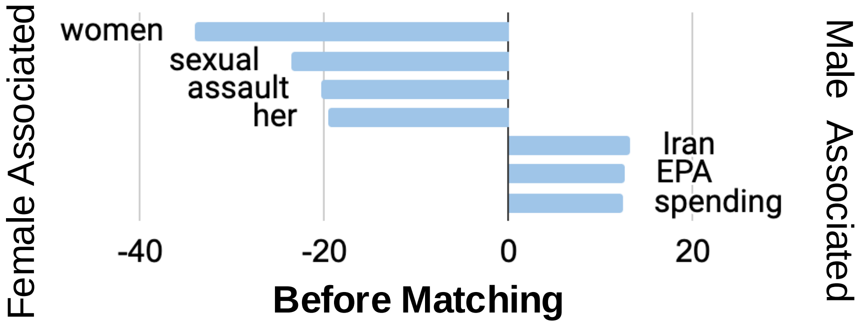
$e_i = P(W.Gender_i = F|Post_i) \approx 0.91$

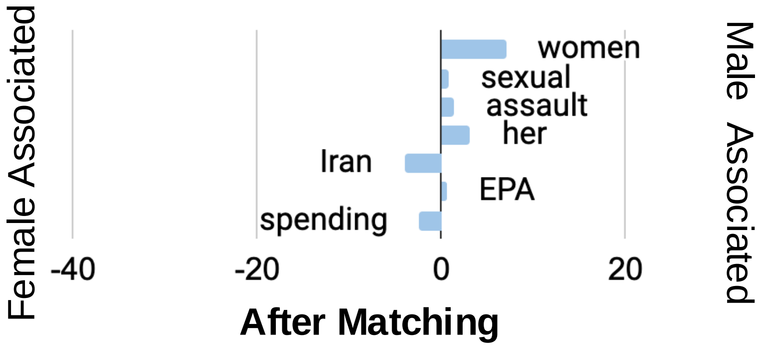
$e_j = P(W.Gender_j = F|Post_j) \approx 0.33$

$e_k = P(W.Gender_k = F|Post_k) \approx 0.32$

$argmin_j|e_k - e_j|$

Propensity matching for *observed* confounding variables

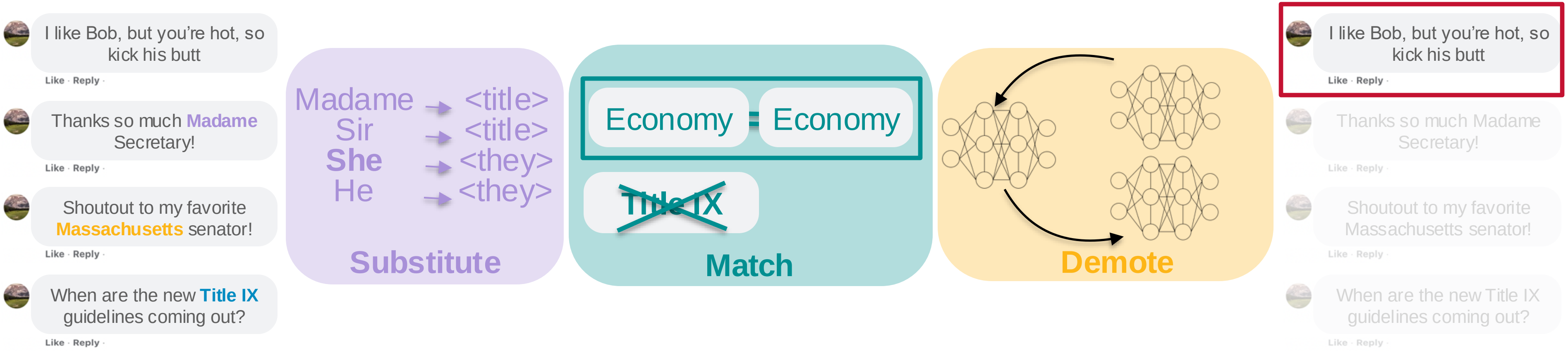




Propensity matching breaks associations between gender and context in the training data

Proposed Model: Comments contain bias if they are highly predictive of gender *despite confound control*

- Substitute overt indicators
- Balance observed confounders through propensity matching
- Demote latent confounders through adversarial training



Self-reported microaggressions

	Public Figs		Politicians	
	F1	**Acc.**	**F1**	**Acc.**
base	61.3	57.3	48.1	64.2
+demotion	**62.2**	57.9	53.7	61.5
+match	38.9	55.9	46.9	50.7
+match+dem.	50.9	57.0	**56.9**	49.9
Random	46.0	49.8	-	-
Class Random	42.1	48.3	-	-

- Models are not trained at all for this task; they are only trained for gender-of-addressee prediction, but they still perform better than chance

Findings: characteristics of bias against women politicians

- Influential words:
  - Competence and domesticity
  - ‘Force’, ‘situation’, ‘spouse’, ‘family’, ‘love’

- Examples:
  - “DINO I hope another real Democrat challenges you next election”
  - “I did not vote for you and have no clue why anyone should have. You do not belong in politics”

Findings: characteristics of bias against women

- Influential words:
  - Appearance and sexualization
  - ‘beautiful’, ‘love’,‘sexo’

- Examples:
  - “Total tangent I know but, you’re gorgeous.”
  - “I like Bob, but you’re hot, so kick his butt.”

Recap

- Overview:
  - Text as confounders, treatment, or outcome
- Text as confounders
  - Topic modeling and language modeling to adjust for text
- Drawing from causal inference to improve NLP models
  - Applying ideas from causal inference to model development
- Next class:
  - Network Analysis

References

- Keith, Katherine, David Jensen, and Brendan O'Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." ACL. 2020.
- Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. "Adjusting for confounding with text matching." American Journal of Political Science 64.4 (2020): 887-903.
- Veitch, Victor, Dhanya Sridhar, and David Blei. "Adapting text embeddings for causal inference." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.
- Field, Anjalie, and Yulia Tsvetkov. "Unsupervised Discovery of Implicit Gender Bias." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.