Using Text Embeddings for Causal Inference

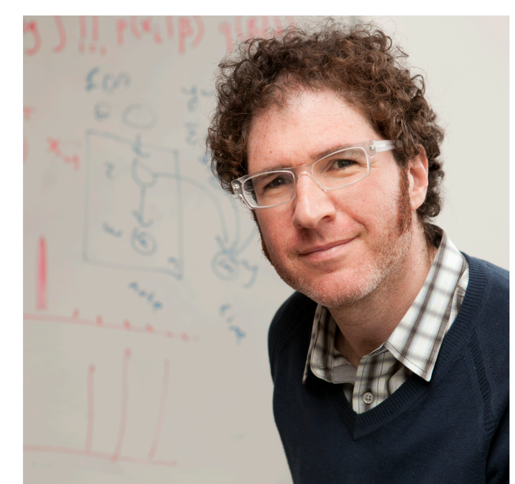
Dhanya Sridhar

Joint work with Victor Veitch and David Blei

Columbia University

New Directions in Analyzing Text as Data

Oct. 4, 2019



Example 1: Effect of Theorems

Does including a theorem in my paper cause it to get accepted?

```
\begin{theorem}
...
\end{theorem}
```

Example 1: Effect of Theorems

Does including a theorem in my paper cause it to get accepted?

\begin{theorem}
...
\end{theorem}

→

...
\begin{theorem}
\end{theorem}
...

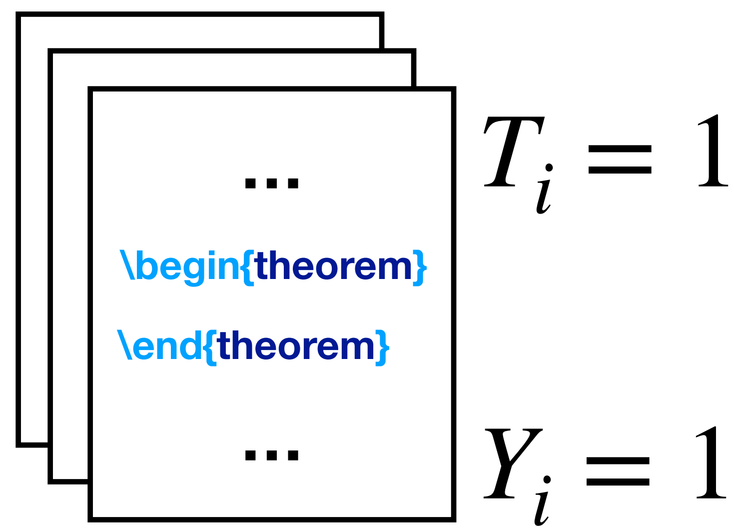
$T_i = 1$

$Y_i = 1$

Dataset of papers with theorem inclusion (T) and paper acceptance (Y)

Naive Estimation Strategy

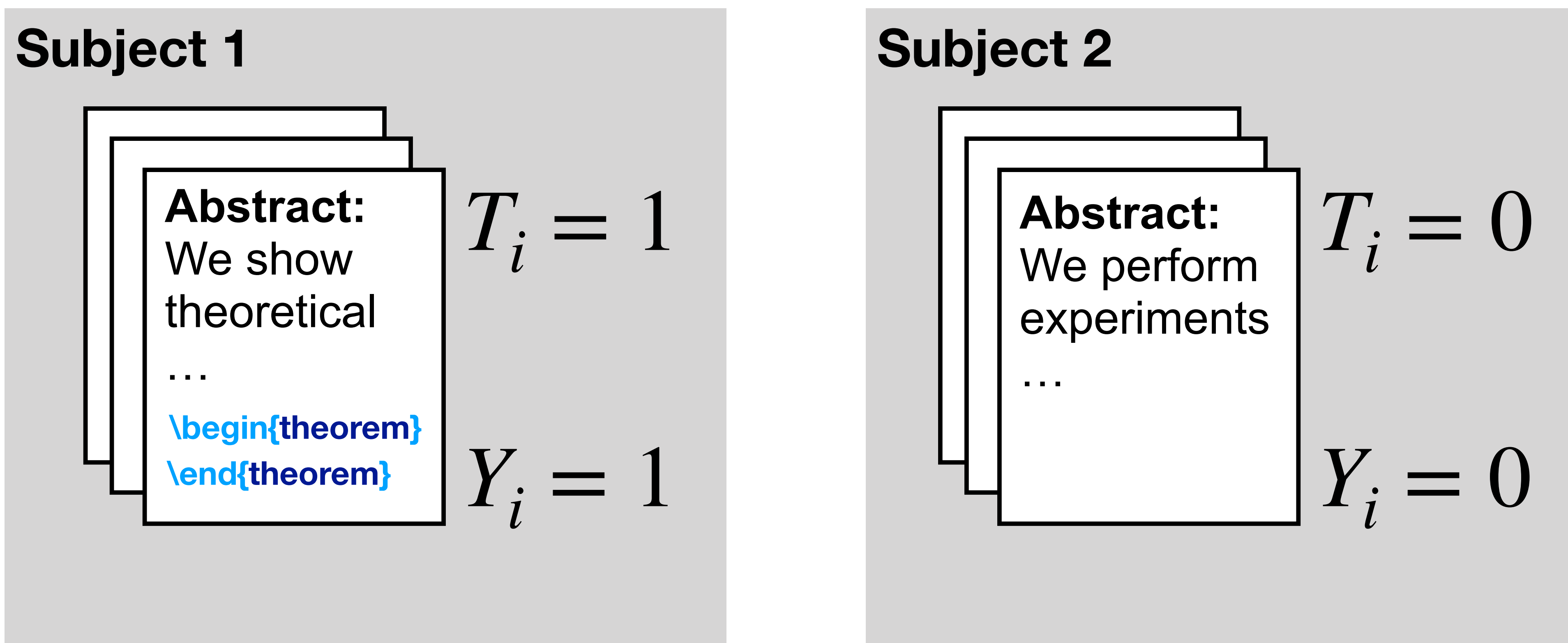
Estimate effect as: $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$



Mean difference in acceptance rates for theorem-having and not theorem-having papers

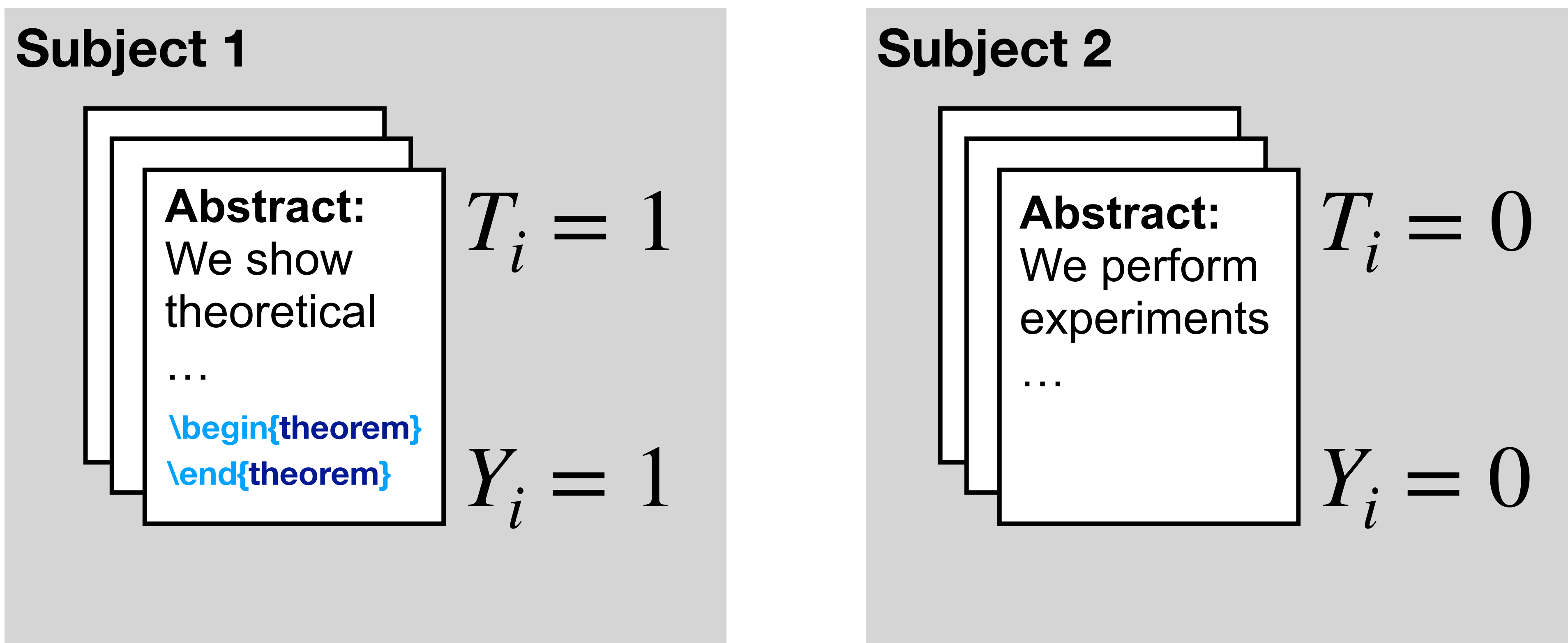
Naive Estimation Strategy

Different paper subjects call for theorems, and also have higher or lower acceptance rates



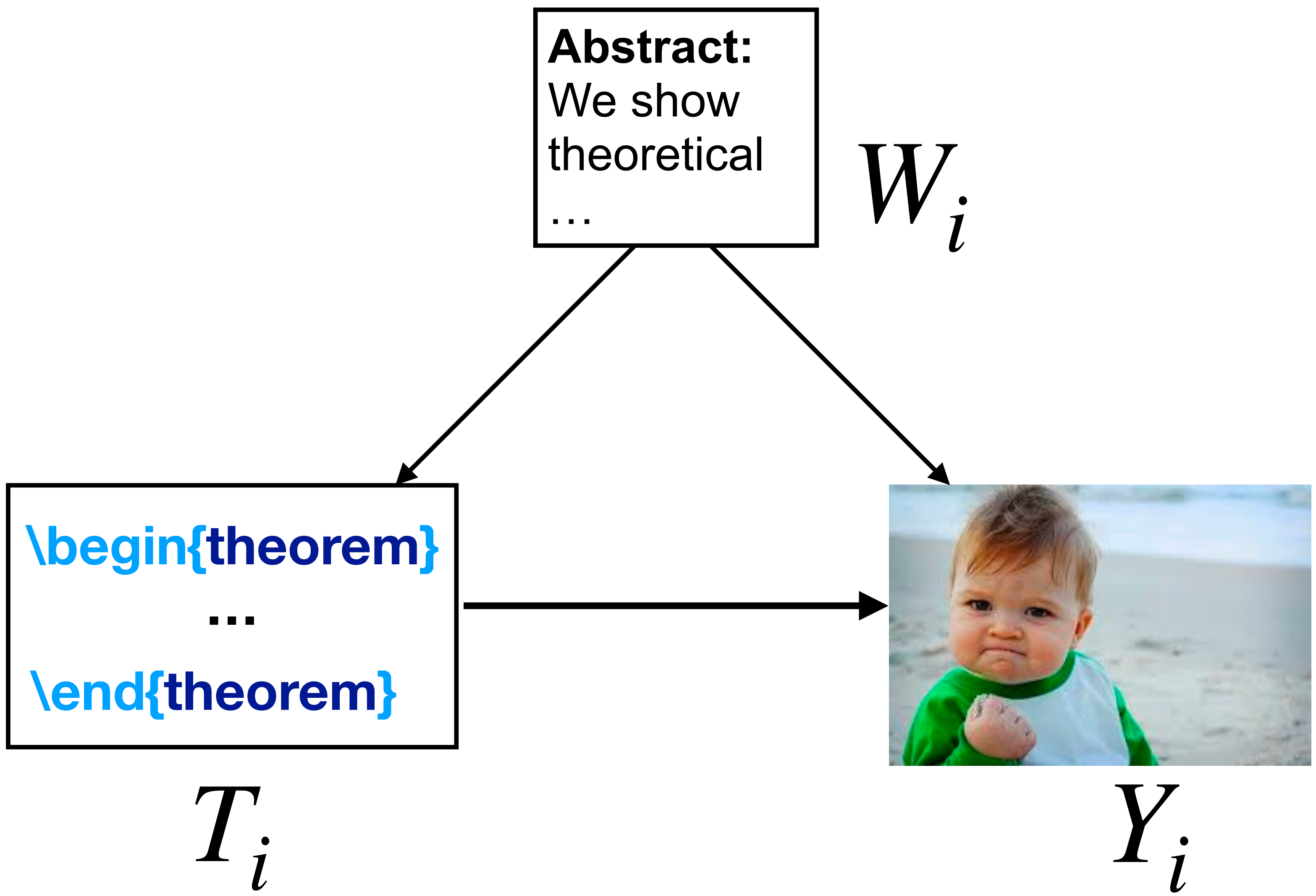
Challenge of Observational Data

$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] \neq \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 1]$$

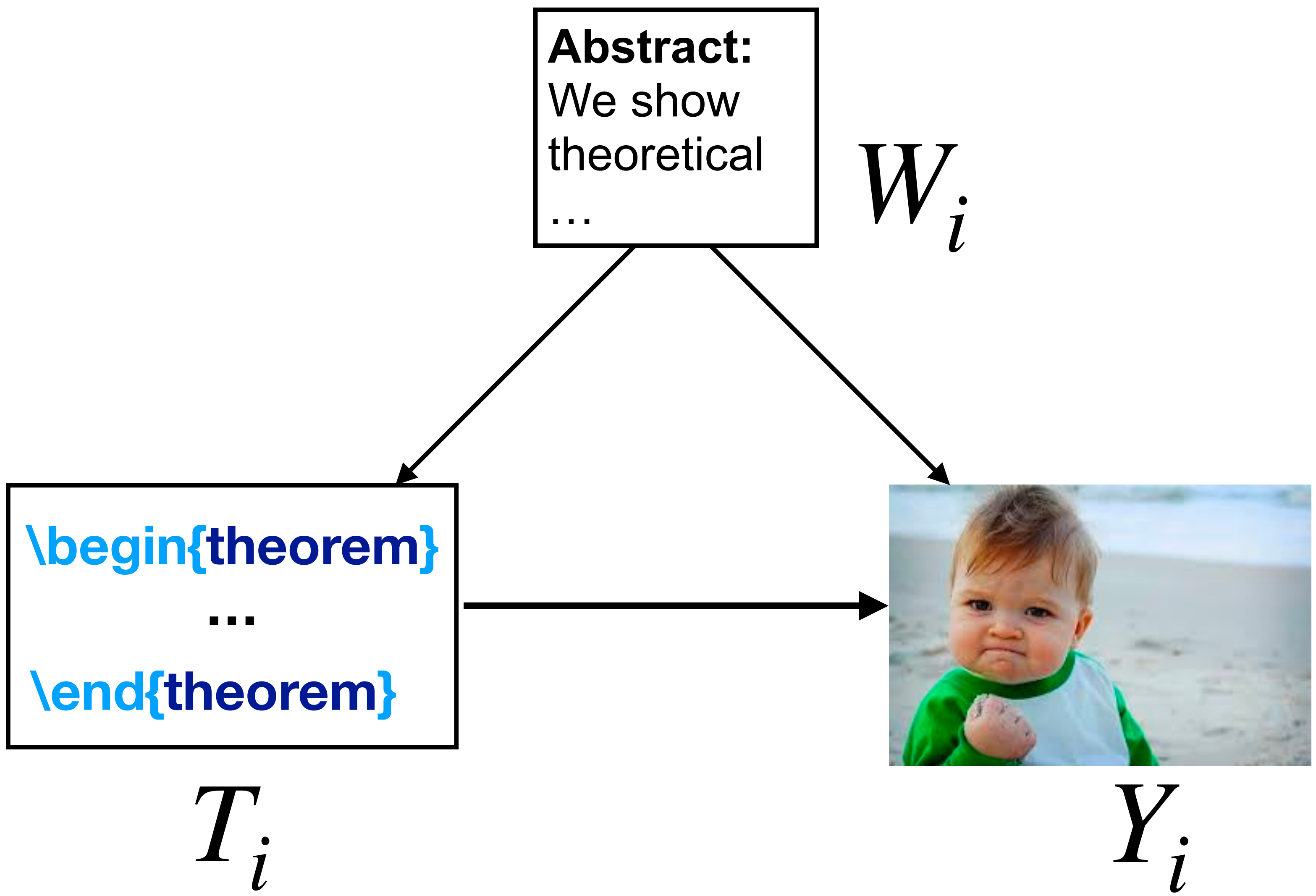


Conditioning and intervening are not the same

Causal Graphical Model

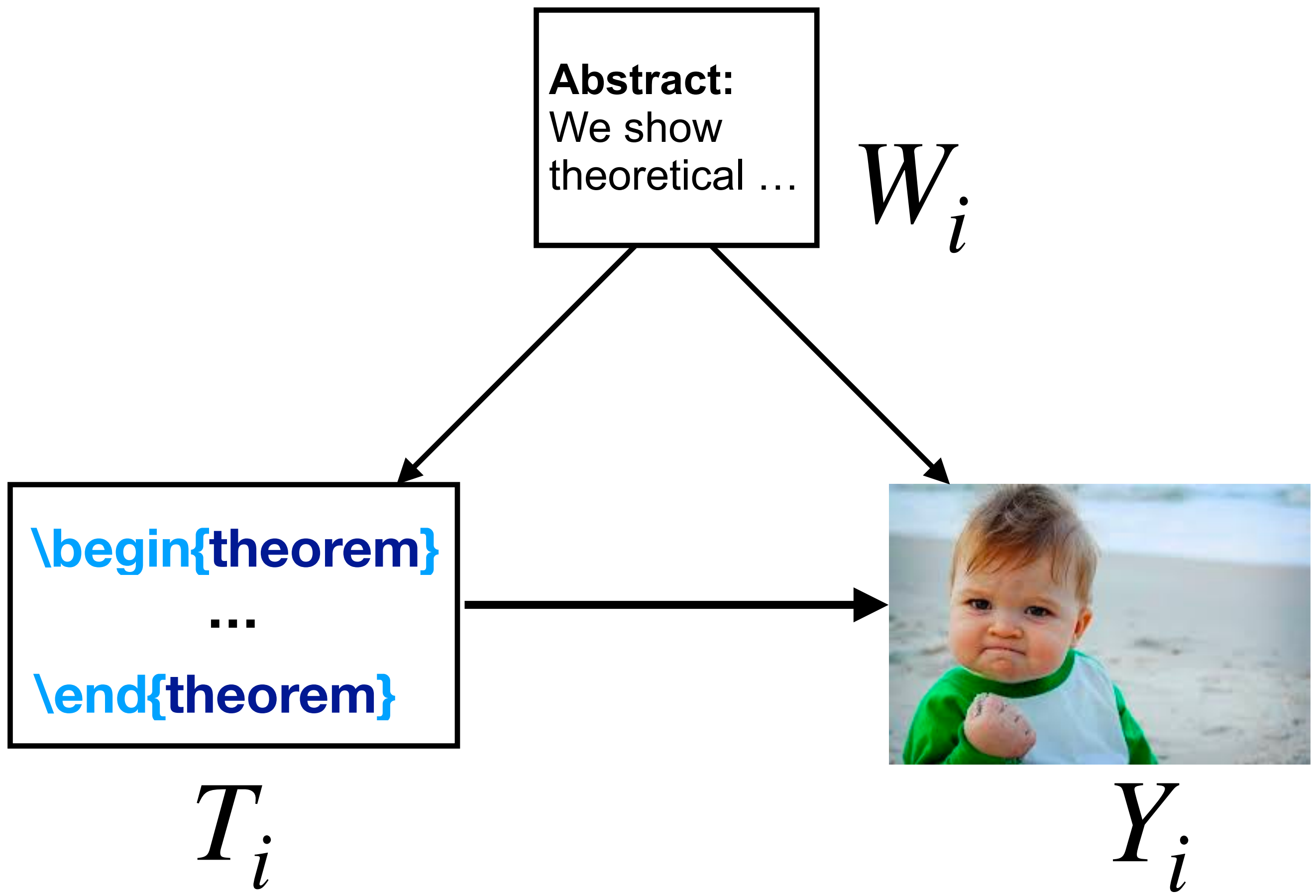


Solution: Backdoor Adjustment



$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] = \mathbb{E}_{\mathrm{W}}[\mathbb{E}[\mathrm{Y} \,|\, \mathrm{T} = 1, \mathrm{W}]]$$

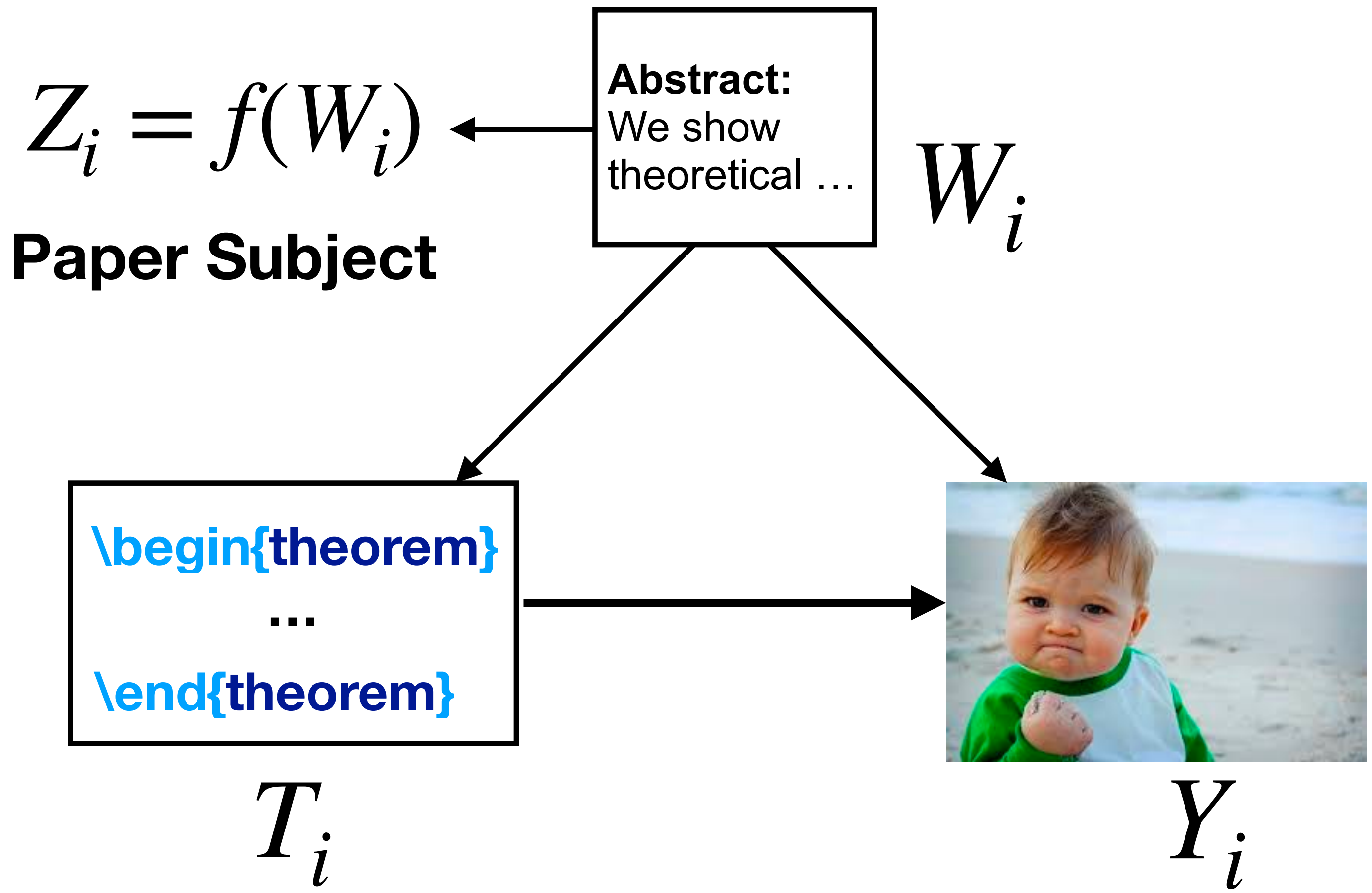
High-dimensional Data



$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] = \mathbb{E}_{\mathrm{W}}[\mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 1, \mathrm{W}]]$$

High-dimensional!

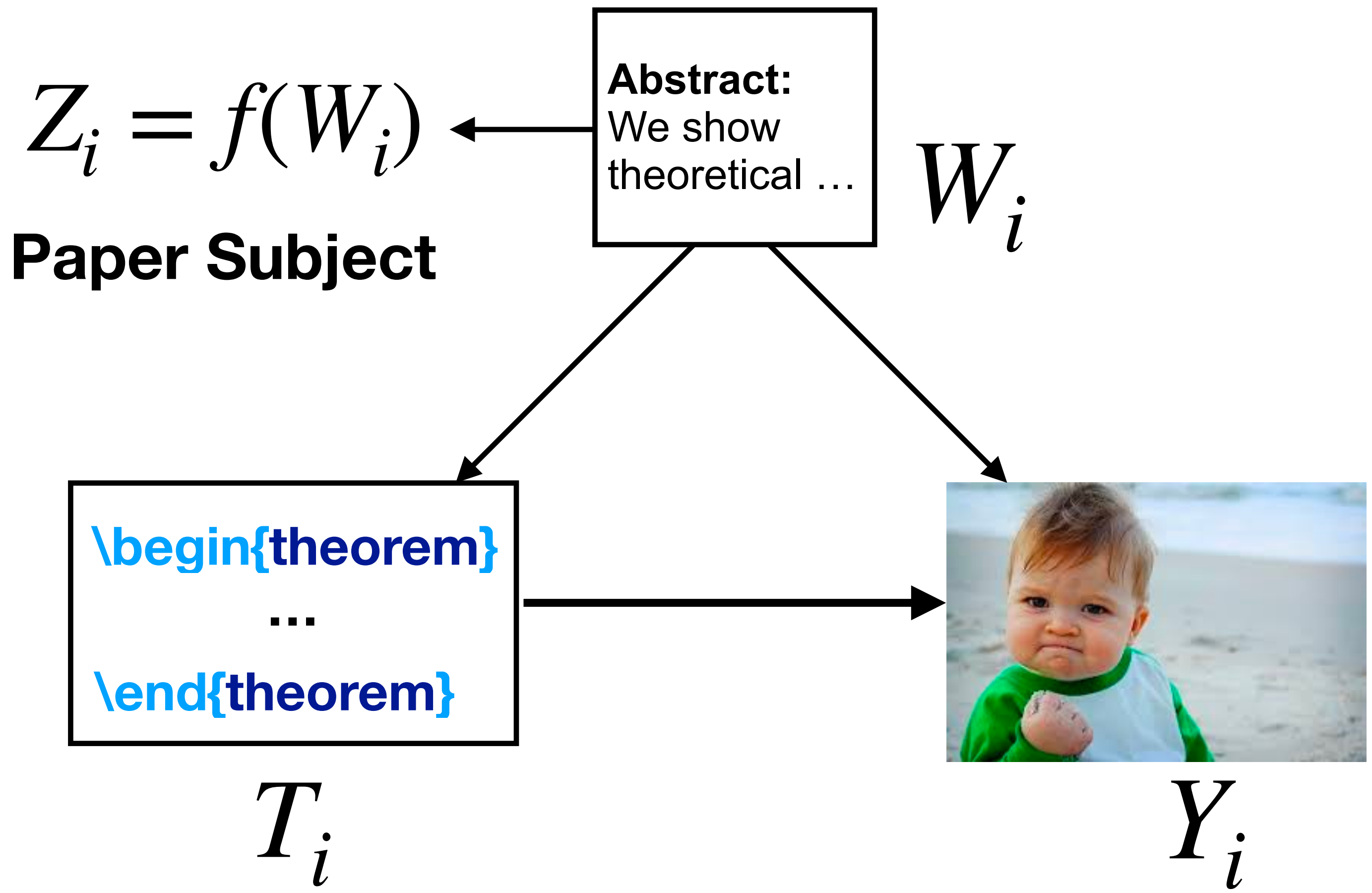
Solution: Dimensionality Reduction



$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] = \mathbb{E}_Z[\mathbb{E}[\mathrm{Y} \mid \mathrm{T} = 1, \mathrm{W}]]$$

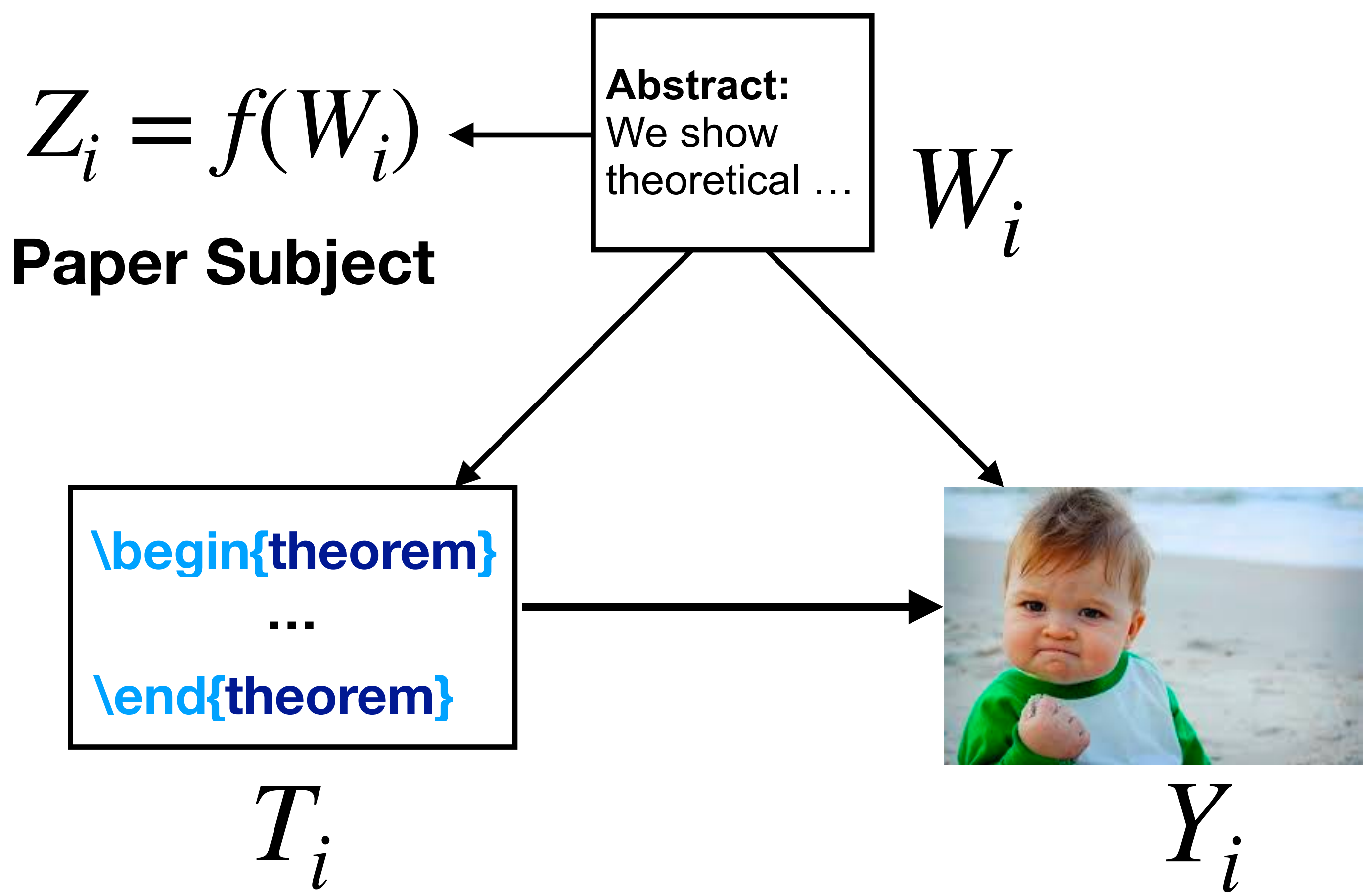
Insight: confounding variable is a low-dimensional representation of words

Why not topic modeling?



One option: fit generative model of abstract text, e.g., LDA

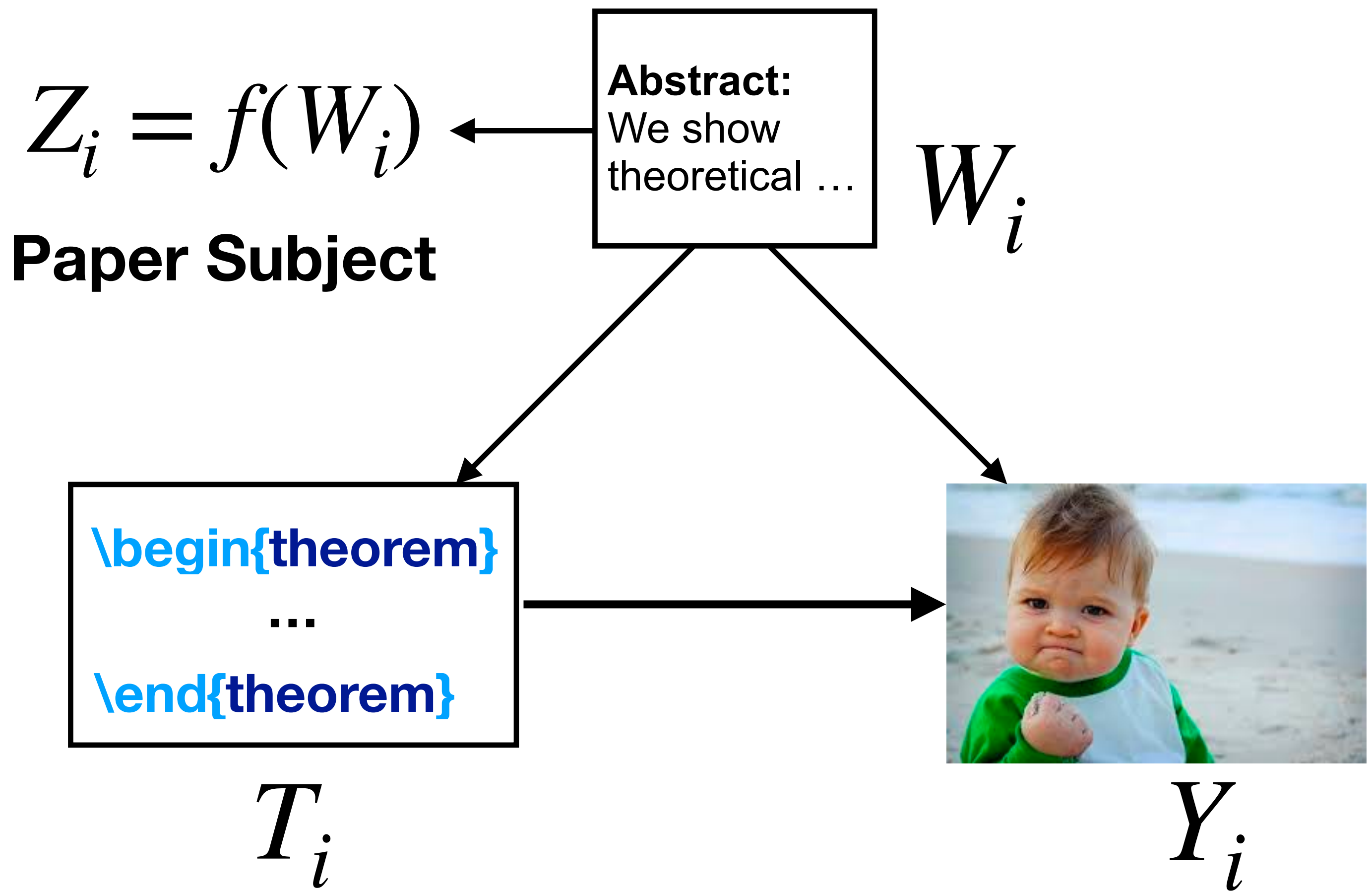
Why not topic modeling?



One option: fit generative model of abstract text, e.g., LDA

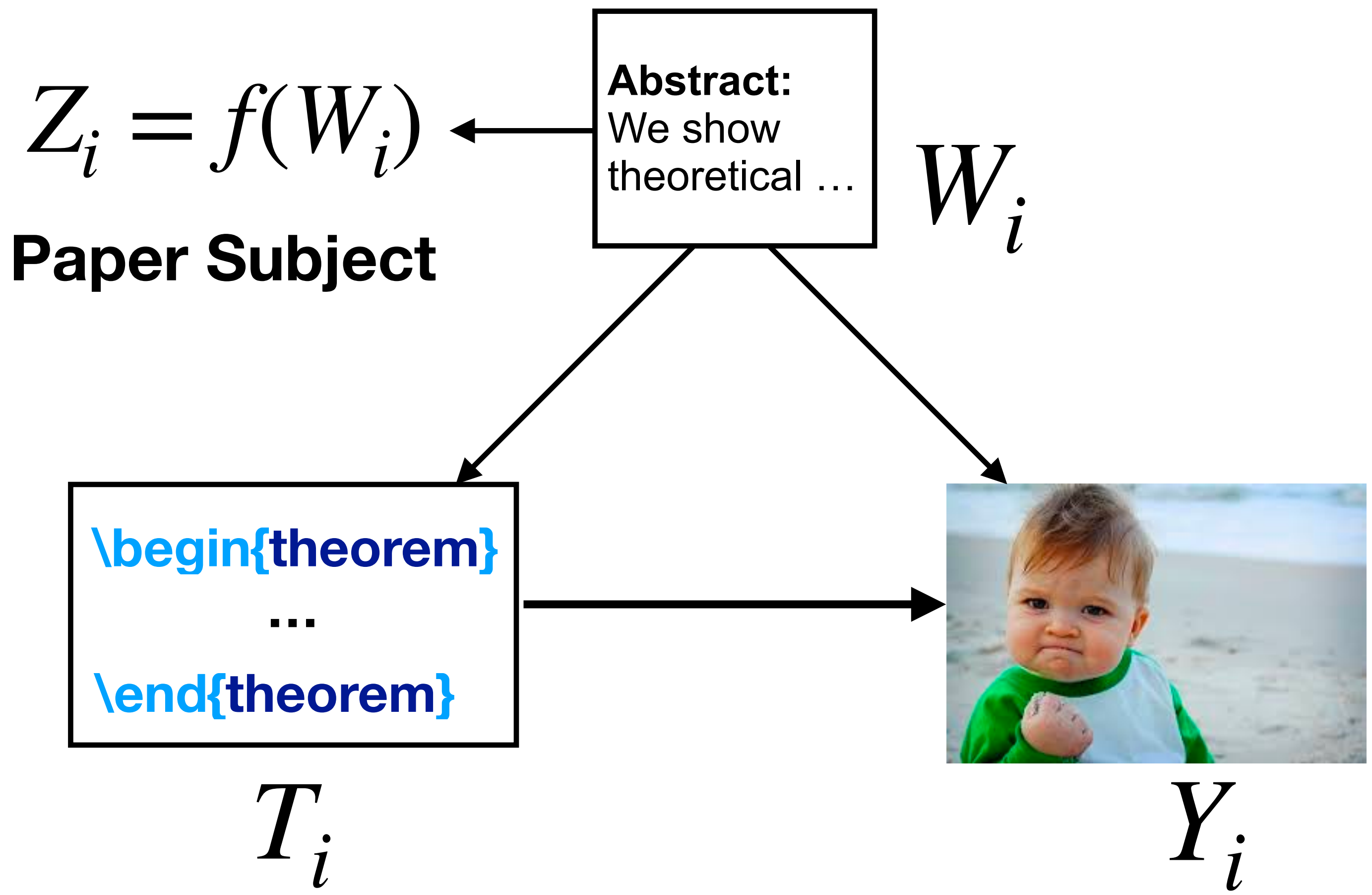
But do we really need full data generating distribution?

Main Ideas



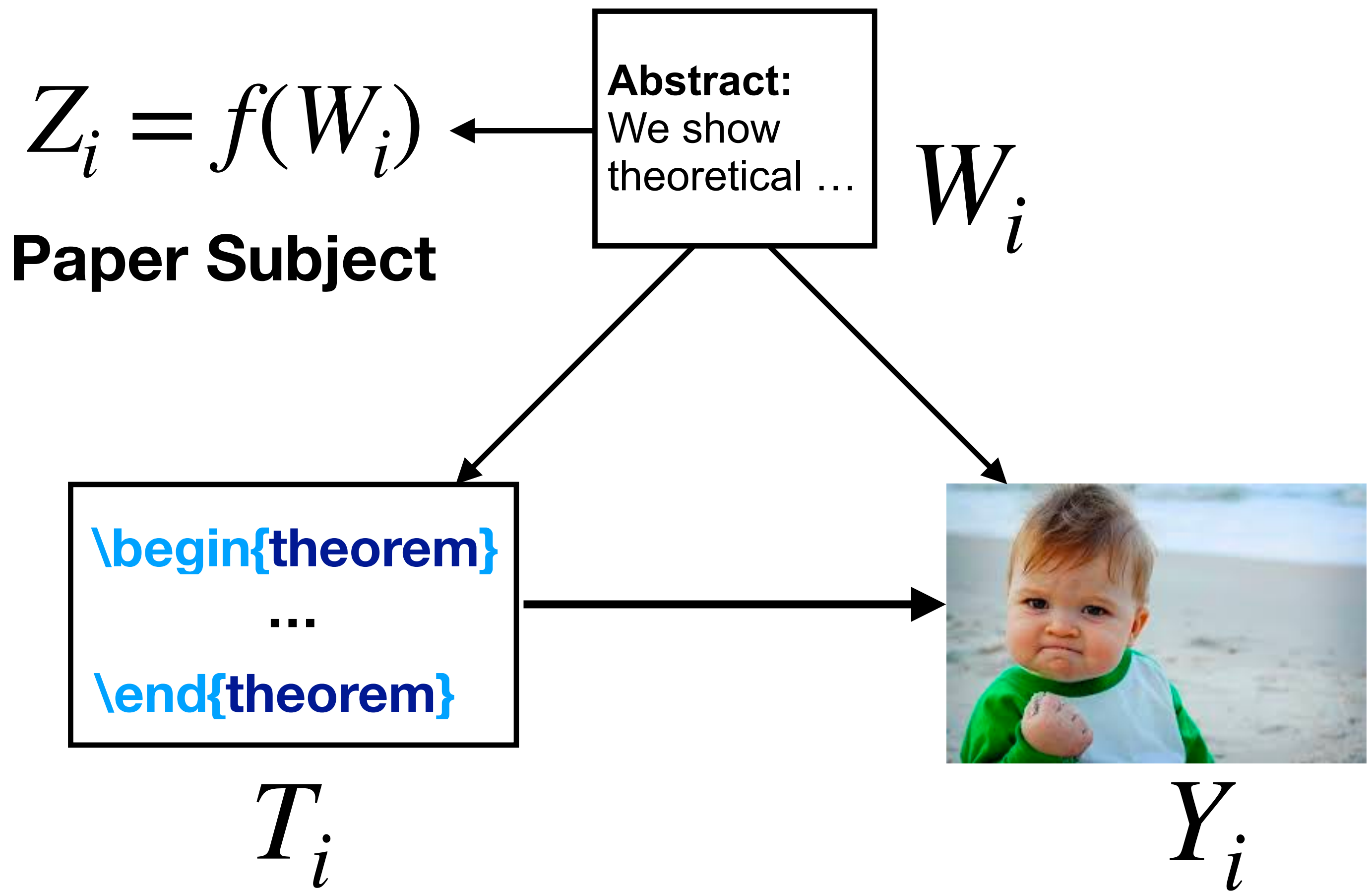
1. Neural language models produce embeddings that work well for supervised problems.

Main Ideas



1. Neural language models produce embeddings that work well for supervised problems.
2. Out-of-the-box, embeddings may not suffice for causal adjustment.

Main Ideas



1. Neural language models produce embeddings that work well for supervised problems.
2. Out-of-the-box, embeddings may not suffice for causal adjustment.
3. **Insight:** the part of text which carries information about treatment and outcome is all that matters.

Adapting Embeddings for Causal Inference

$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] = \mathbb{E}_{\mathrm{Z}}[\mathbb{E}[\mathrm{Y} \,|\, \mathrm{T} = 1, \mathrm{Z}]]$$

$$= \frac{1}{n} \sum_i \mathbb{E}[Y_i \,|\, T_i = 1, f(W_i)]$$

$$= \frac{1}{n} \sum_i Q(T_i, f(W_i))$$

Want mapping of words to minimize error on predicting outcomes given treatment

Adapting Embeddings for Causal Inference

$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] = \mathbb{E}_{\mathrm{Z}}[\mathbb{E}[\mathrm{Y} \,|\, \mathrm{T} = 1, \mathrm{Z}]]$$

$$= \frac{1}{n} \sum_i \mathbb{E}[Y_i \,|\, T_i = 1, f(W_i)]$$

$$= \frac{1}{n} \sum_i Q(T_i, f(W_i))$$

Learn embedding $\lambda = f(W)$ **to predict conditional outcomes**

Want mapping of words to minimize error on predicting outcomes given treatment

Adapting Embeddings for Causal Inference

$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] = \mathbb{E}_{\mathrm{Z}}[\mathbb{E}[\mathrm{Y} \,|\, \mathrm{T} = 1, \mathrm{Z}]]$$

$$= \frac{1}{n} \sum_i \mathbb{E}[Y_i \,|\, T_i = 1, f(W_i)]$$

$$= \frac{1}{n} \sum_i Q(T_i, f(W_i))$$

Estimators with better statistical efficiency use propensity score:

$$P(T = 1 \,|\, \lambda = f(W)) = g(\lambda)$$

Adapting Embeddings for Causal Inference

$$\mathbb{E}[Y; \mathrm{do}(\mathrm{T} = 1)] = \mathbb{E}_{\mathrm{Z}}[\mathbb{E}[\mathrm{Y} \,|\, \mathrm{T} = 1, \mathrm{Z}]]$$

$$= \frac{1}{n} \sum_i \mathbb{E}[Y_i \,|\, T_i = 1, f(W_i)]$$

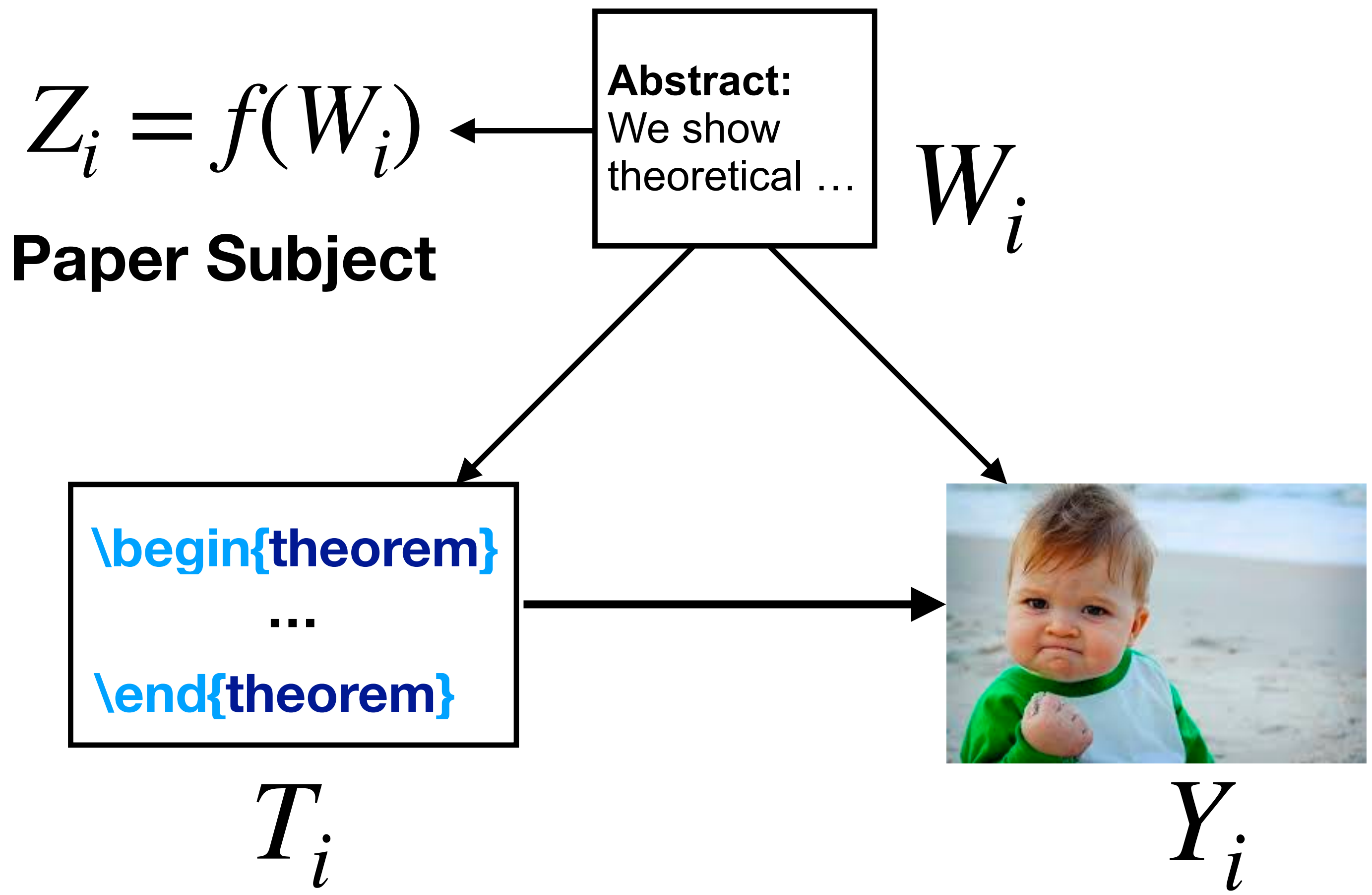
$$= \frac{1}{n} \sum_i Q(T_i, f(W_i))$$

Estimators with better statistical efficiency use propensity score:

$$P(T = 1 \,|\, \lambda = f(W)) = g(\lambda)$$

Learn embedding λ to predict conditional outcomes and propensity scores

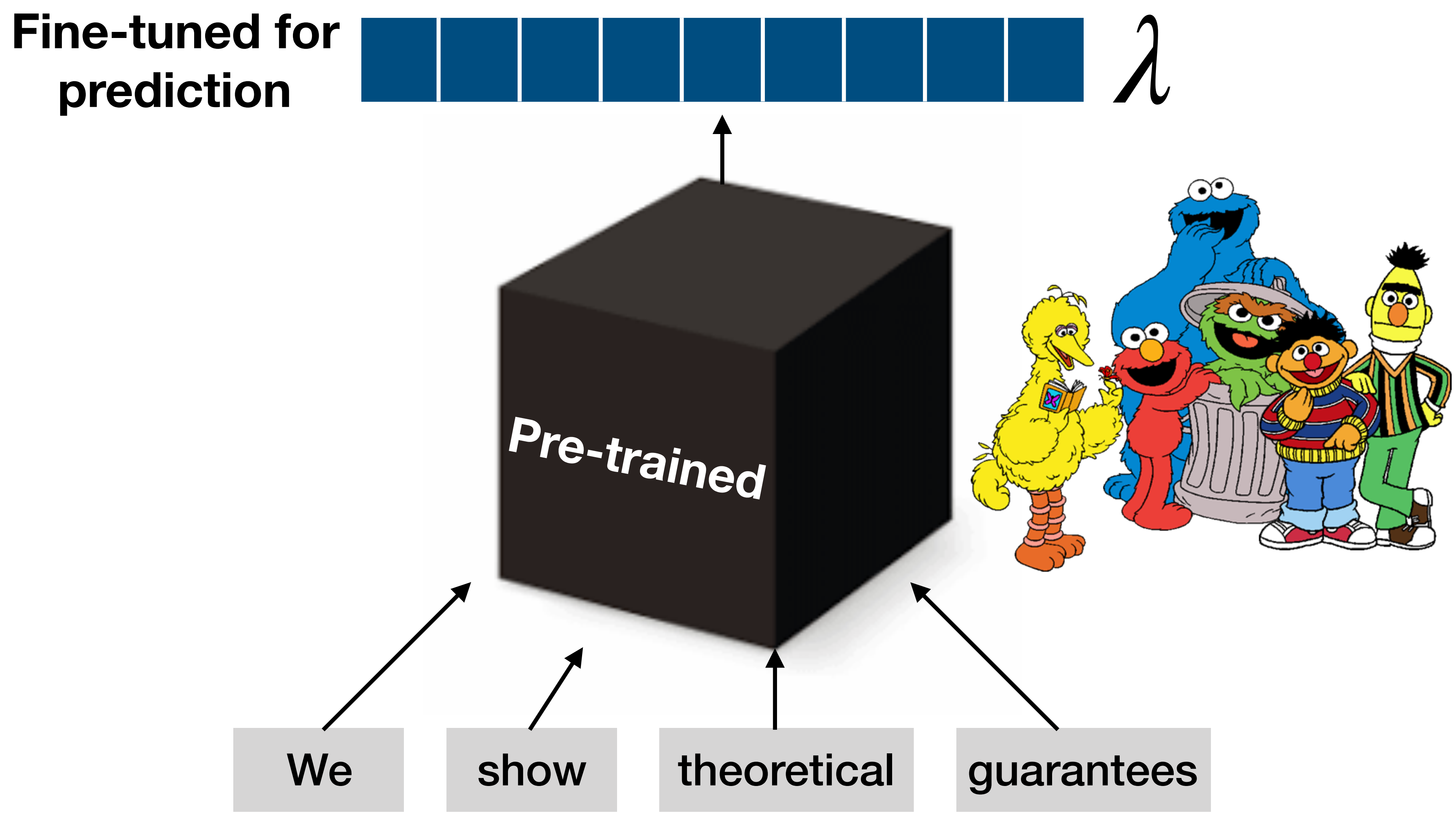
Main Ideas



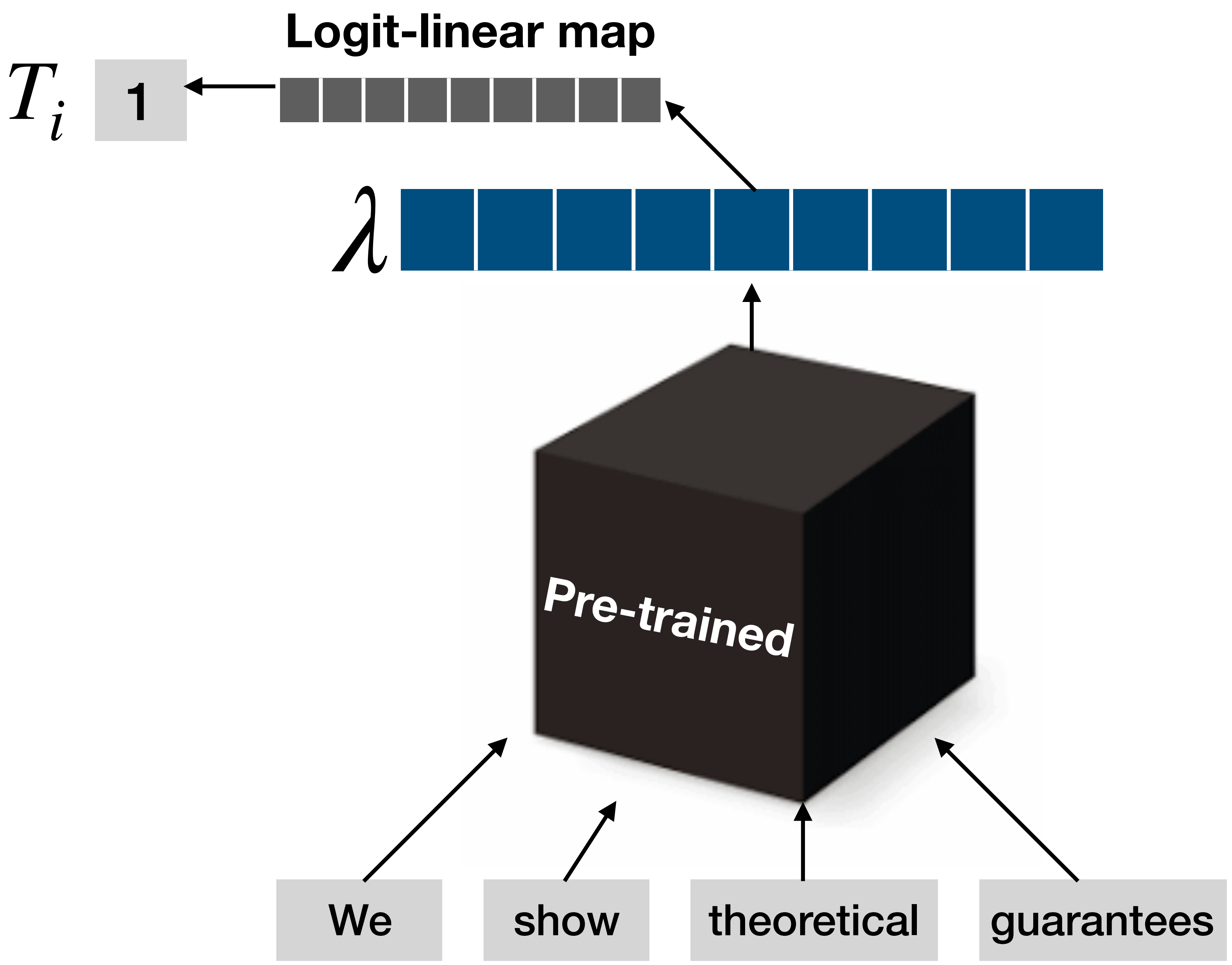
1. **Neural language models produce embeddings that work well for supervised problems.**
2. Out of the box, embeddings may not suffice for causal adjustment
3. **Insight:** the part of text which carries information about treatment and outcome is all that matters

Standard BERT

Transformer model that produces a task-specific embedding given a sequence of tokens, e.g., abstract

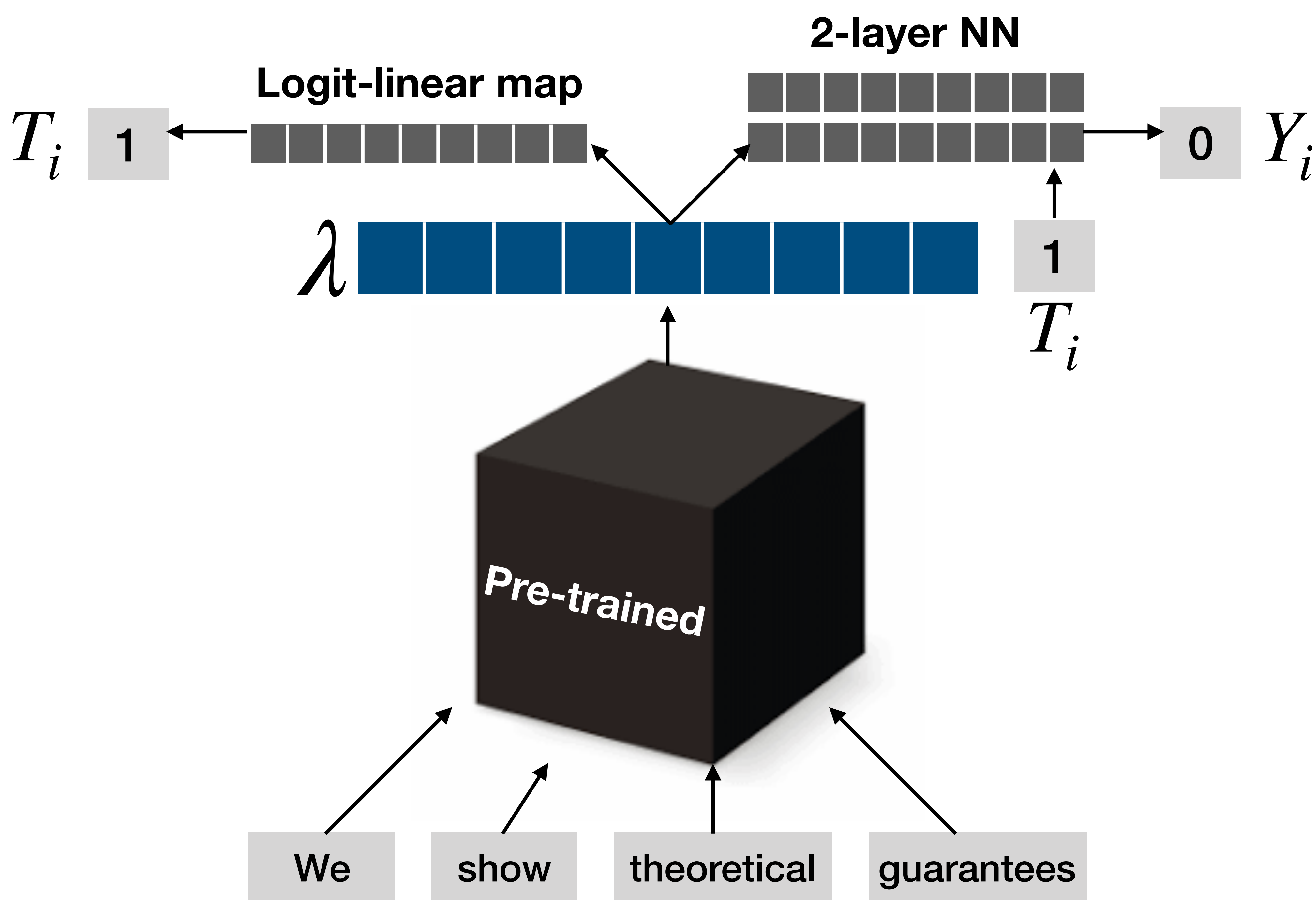


Causal BERT



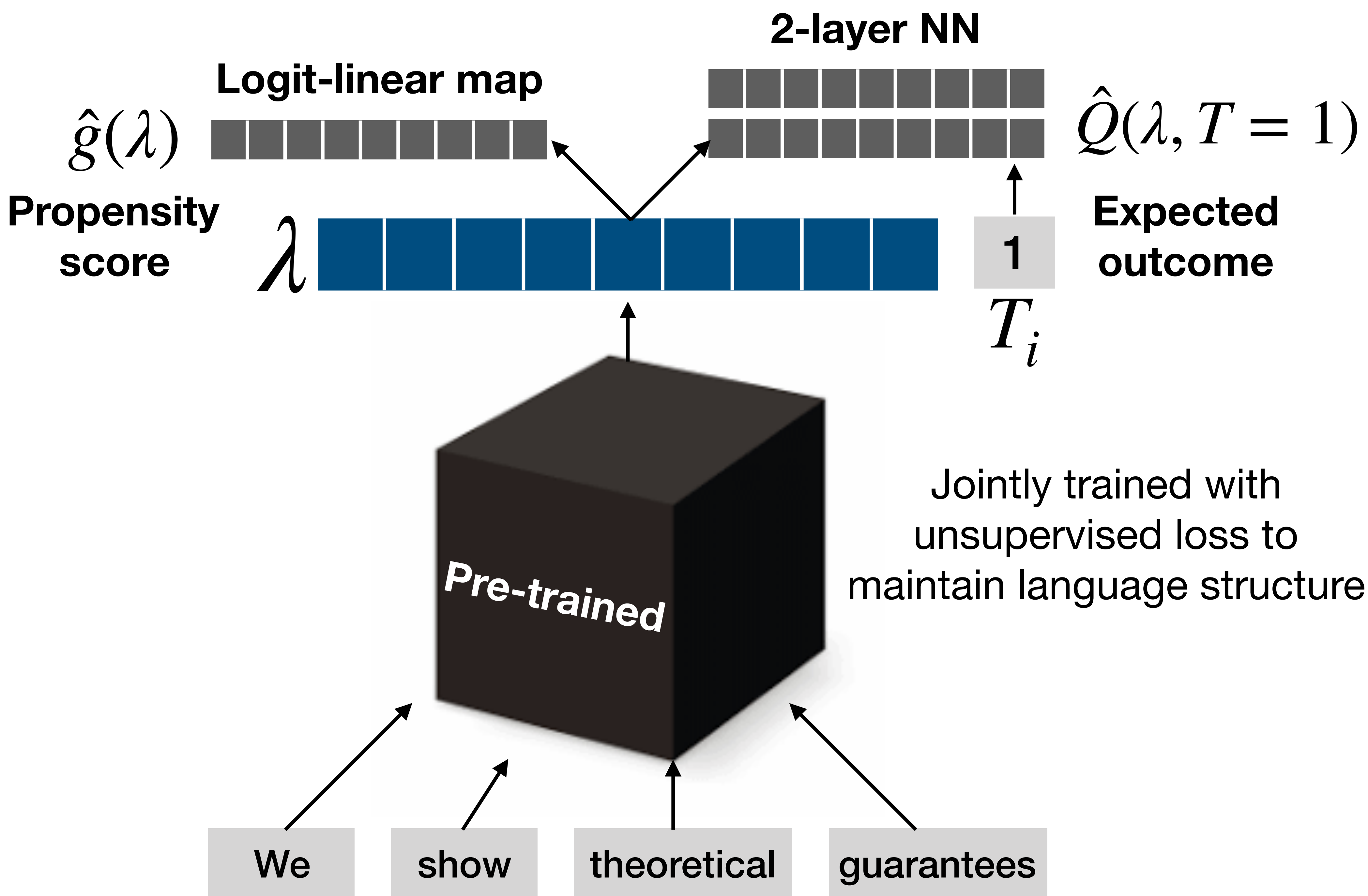
Logit-linear map
T_i
1
λ
Pre-trained
We
show
theoretical
guarantees

Causal BERT



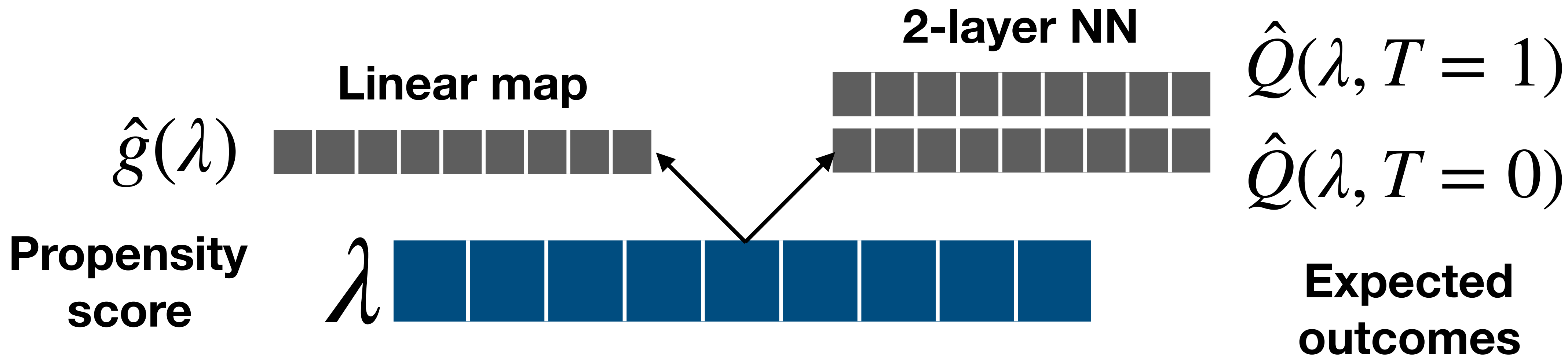
2-layer NN
Logit-linear map
T_i
1
0
Y_i
λ
1
T_i
Pre-trained
We
show
theoretical
guarantees

Causal BERT

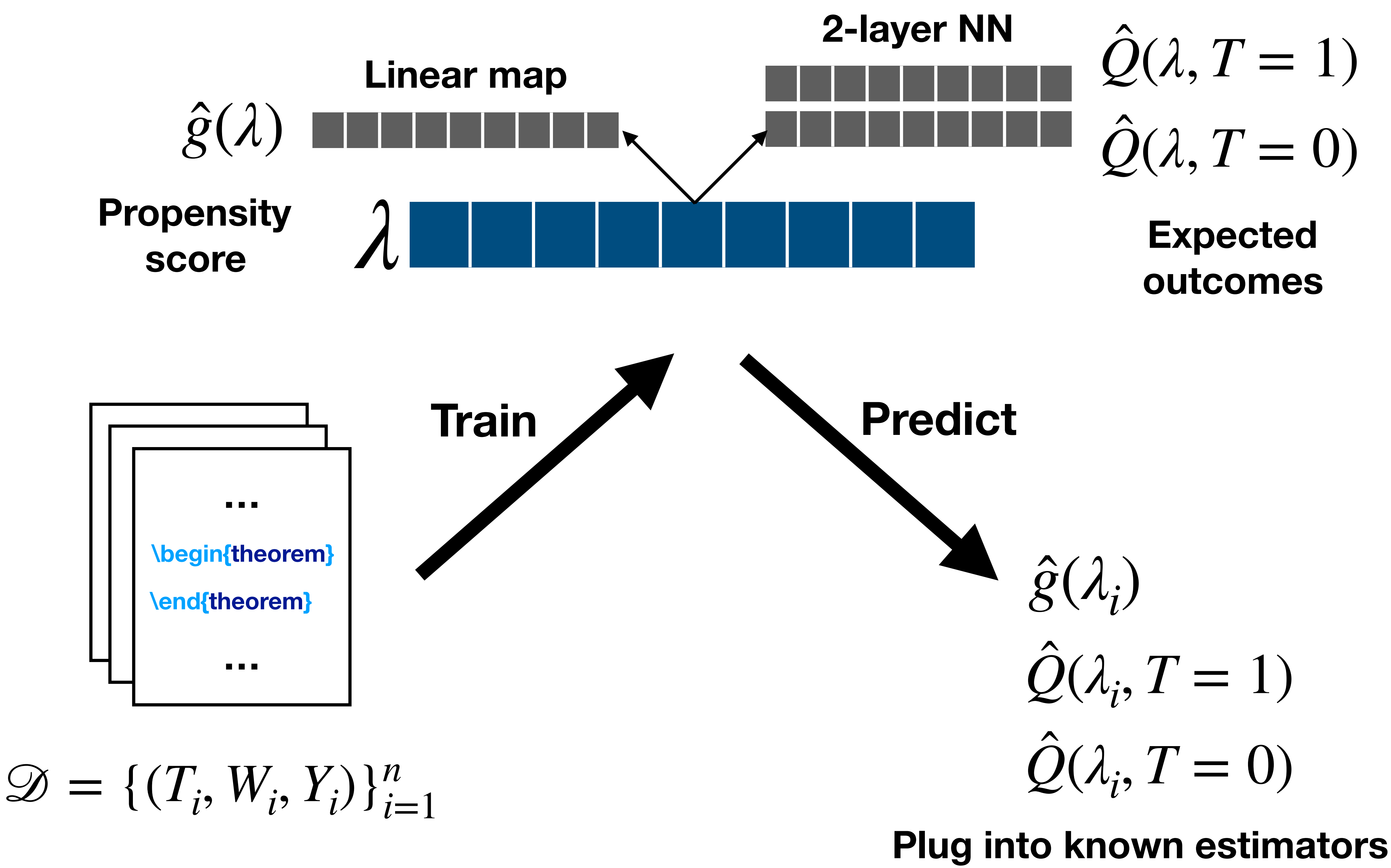


2-layer NN
Logit-linear map
$\hat{g}(\lambda)$
$\hat{Q}(\lambda, T = 1)$
Propensity score
λ
1
T_i
Expected outcome
Pre-trained
Jointly trained with unsupervised loss to maintain language structure
We
show
theoretical
guarantees

Causal Estimation

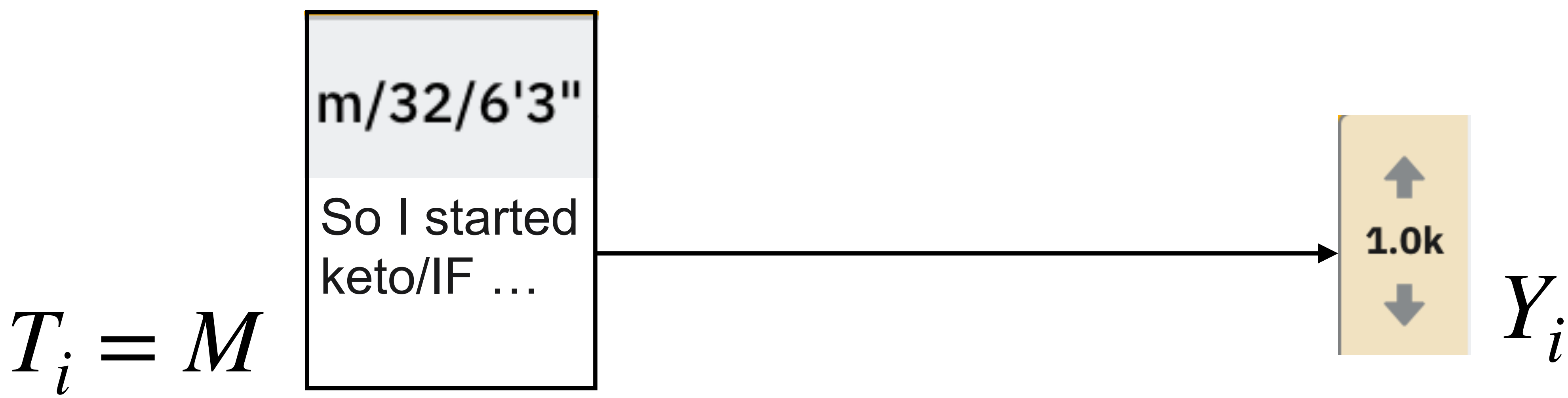


Causal Estimation



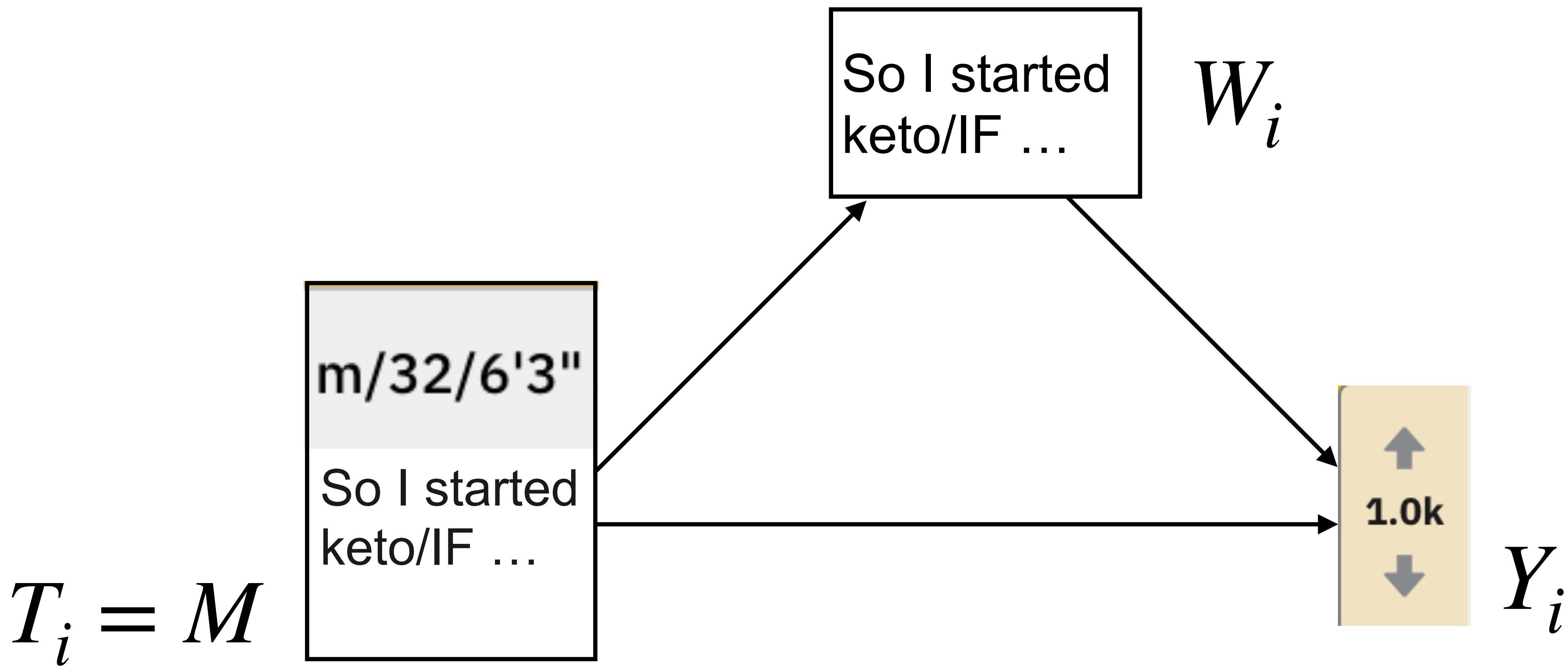
Example 2: Direct Effects

Does labeling a Reddit post with gender directly affect its popularity?



Example 2: Direct Effects

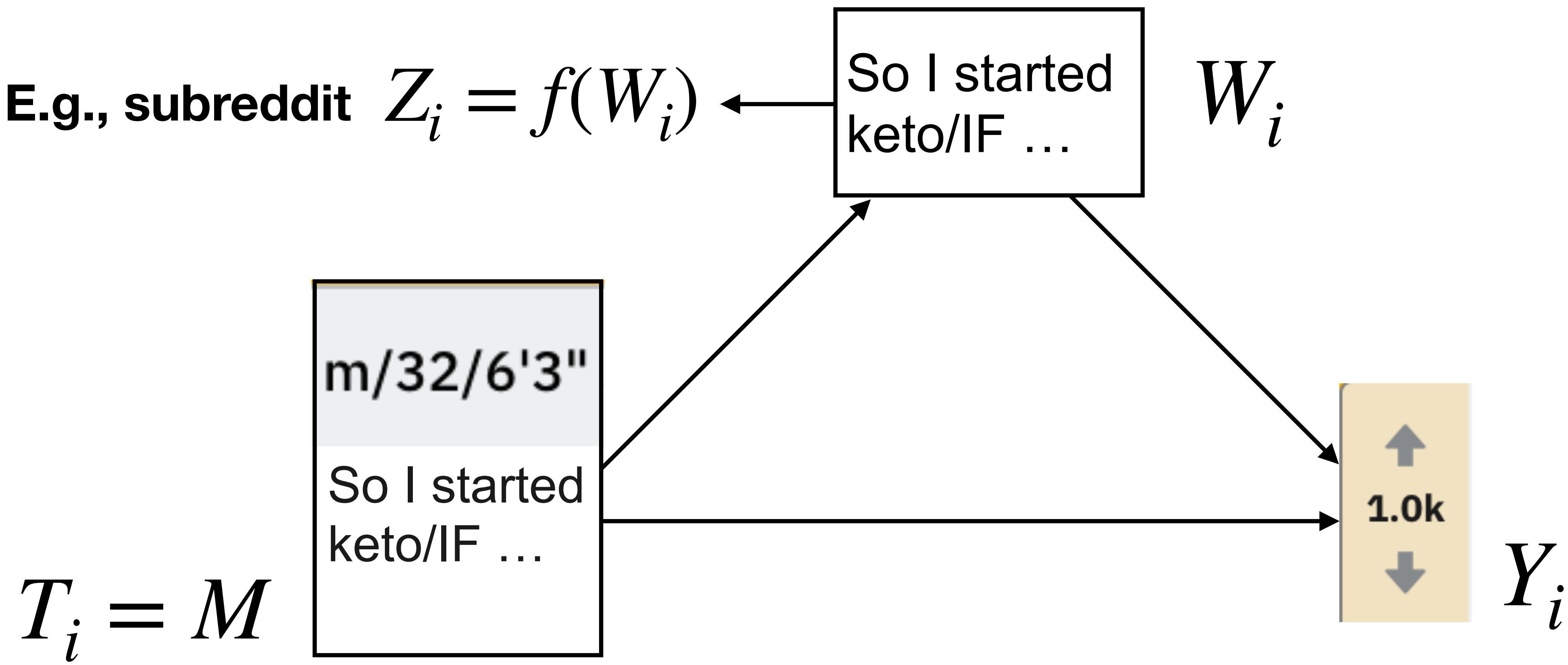
Does labeling a Reddit post with gender directly affect its popularity?



Want to estimate direct effect after accounting for effect mediated by variations in text

Example 2: Direct Effects

Does labeling a Reddit post with gender directly affect its popularity?



Estimator of direct effect also involves propensity score and expected outcomes

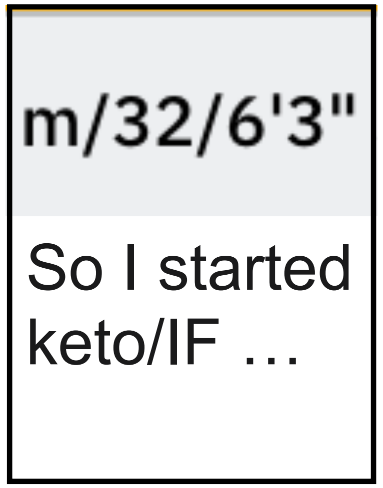
Does Causal BERT work?

How do we evaluate this method?

Average treatment effects

\begin{theorem}
...
\end{theorem}

Natural direct effects



No available ground truth causal effects!

Does Causal BERT work?

How do we evaluate this method?
Strategy: simulate only outcomes

Average treatment effects

\begin{theorem}
...
\end{theorem}

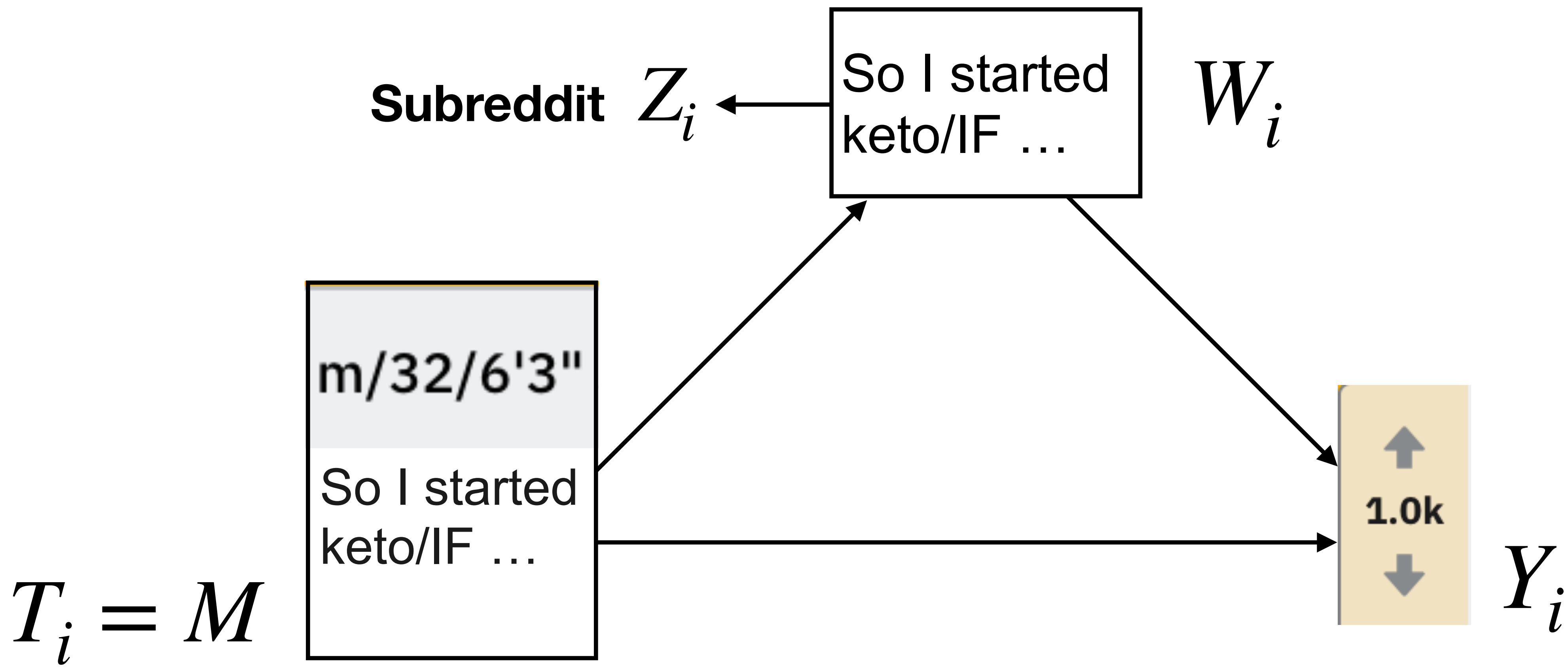
Natural direct effects

m/32/6'3"

So I started keto/IF …

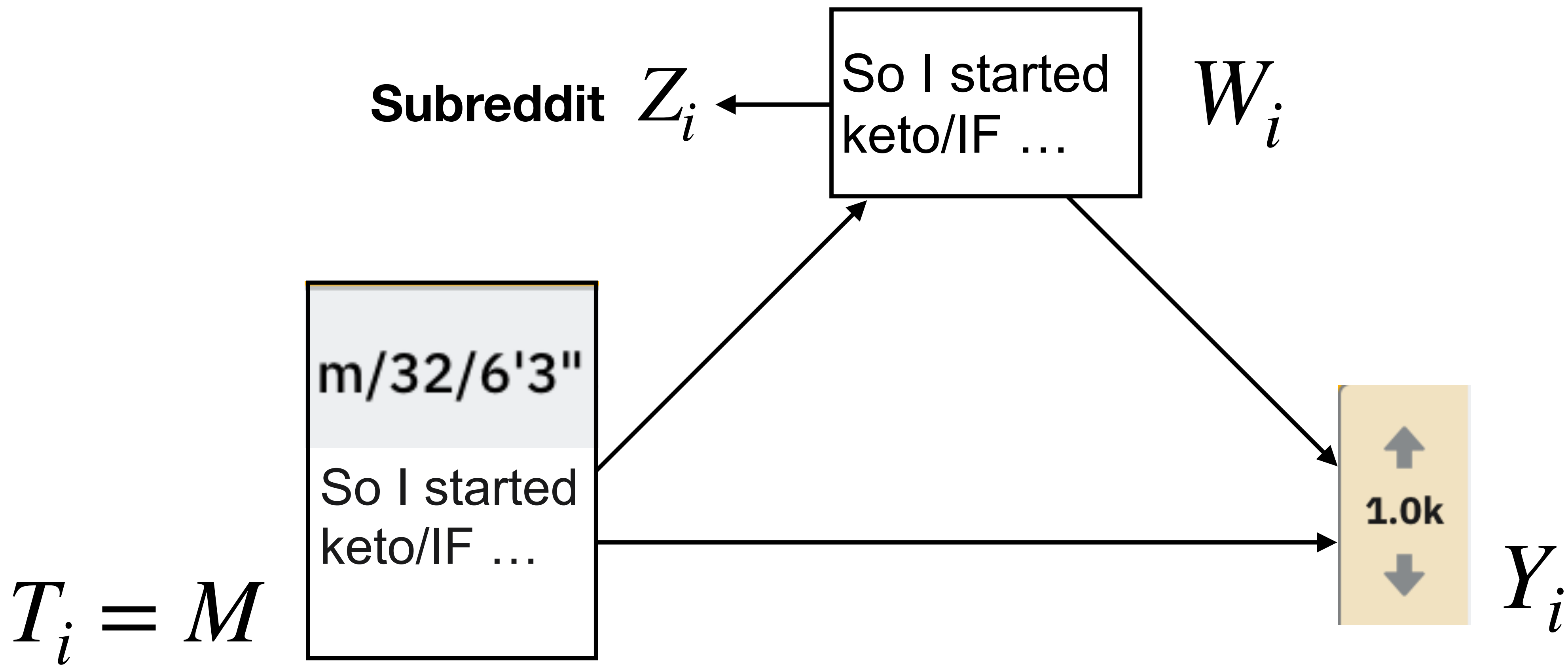
No available ground truth causal effects!

Example 2: Direct Effects



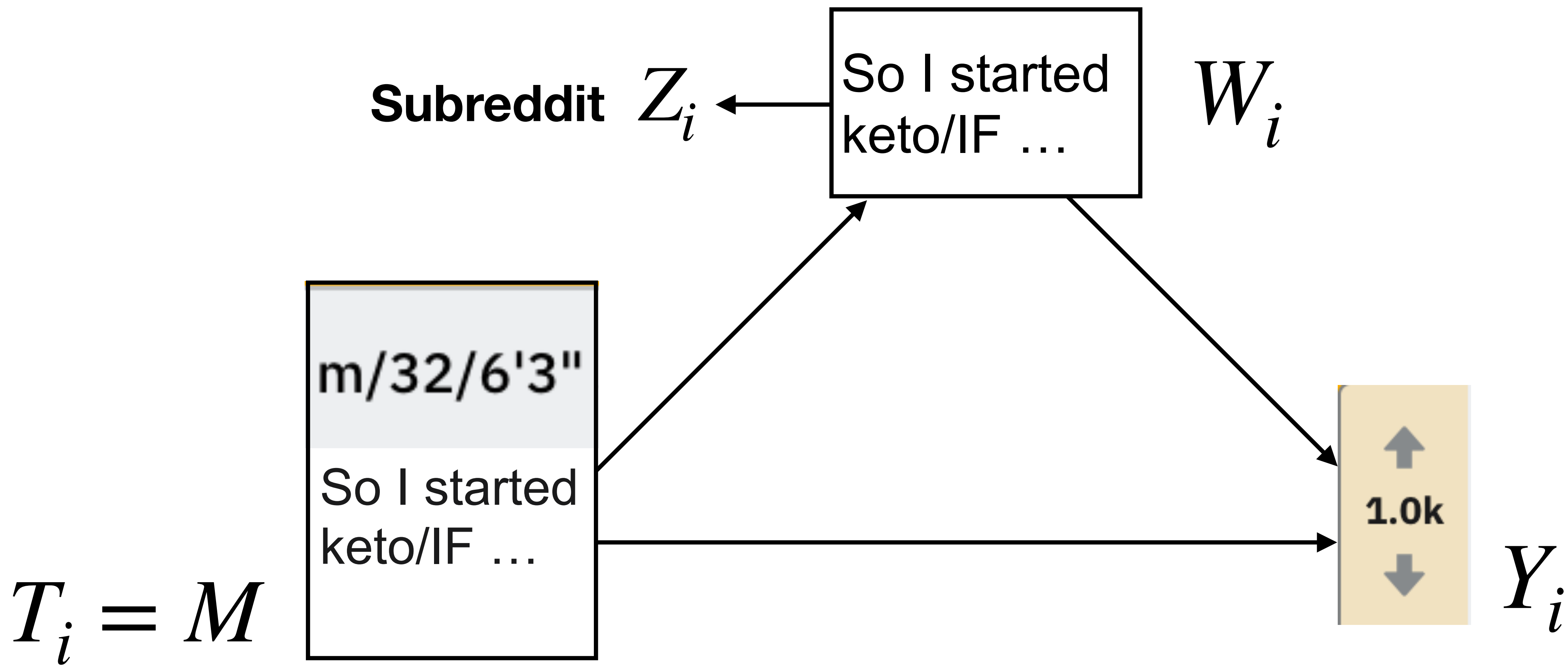
Identify known covariate which text encodes and varies between genders, e.g., subreddit

Example 2: Direct Effects



Simulate outcomes in a way that uses both the treatment and subreddit information

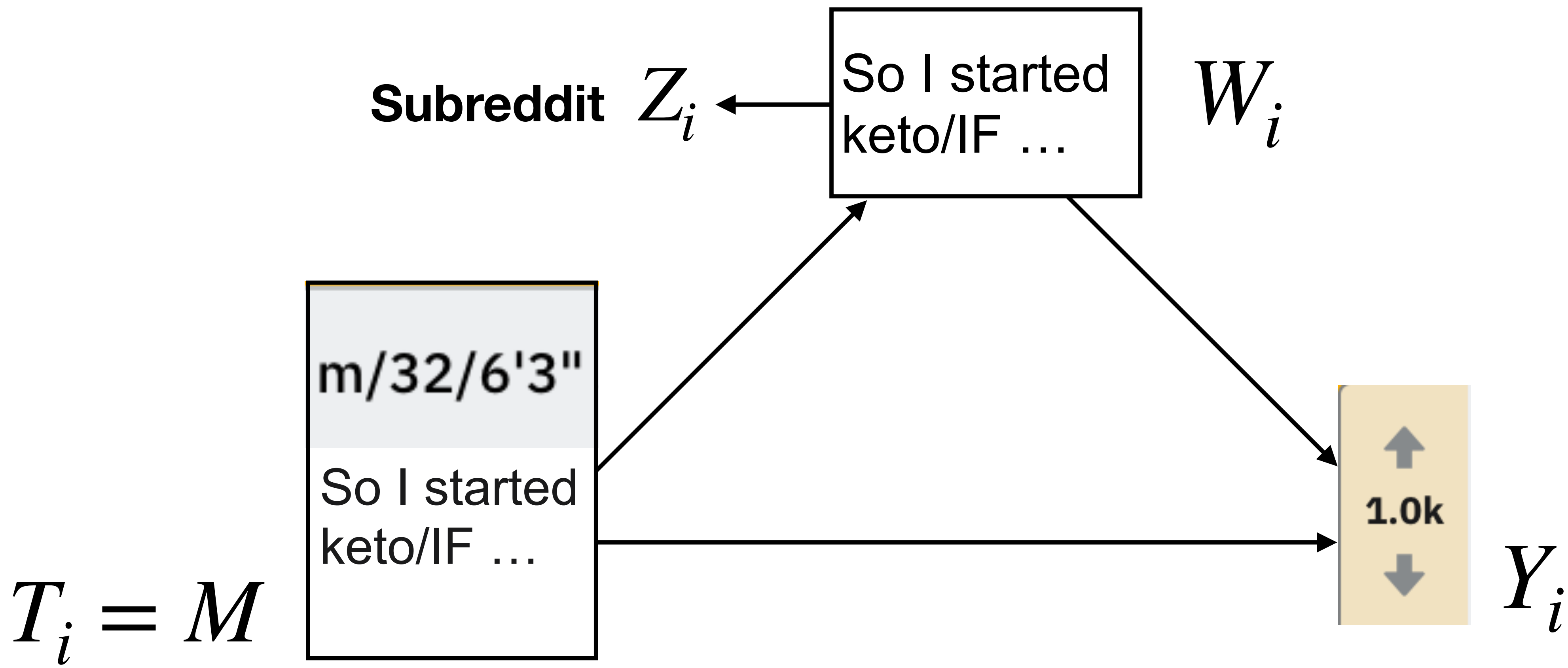
Example 2: Direct Effects



$$Y_i = T_i + \beta_1(\pi(Z_i) - 0.5) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0,\gamma)$$

Treatment effect = 1

Example 2: Direct Effects

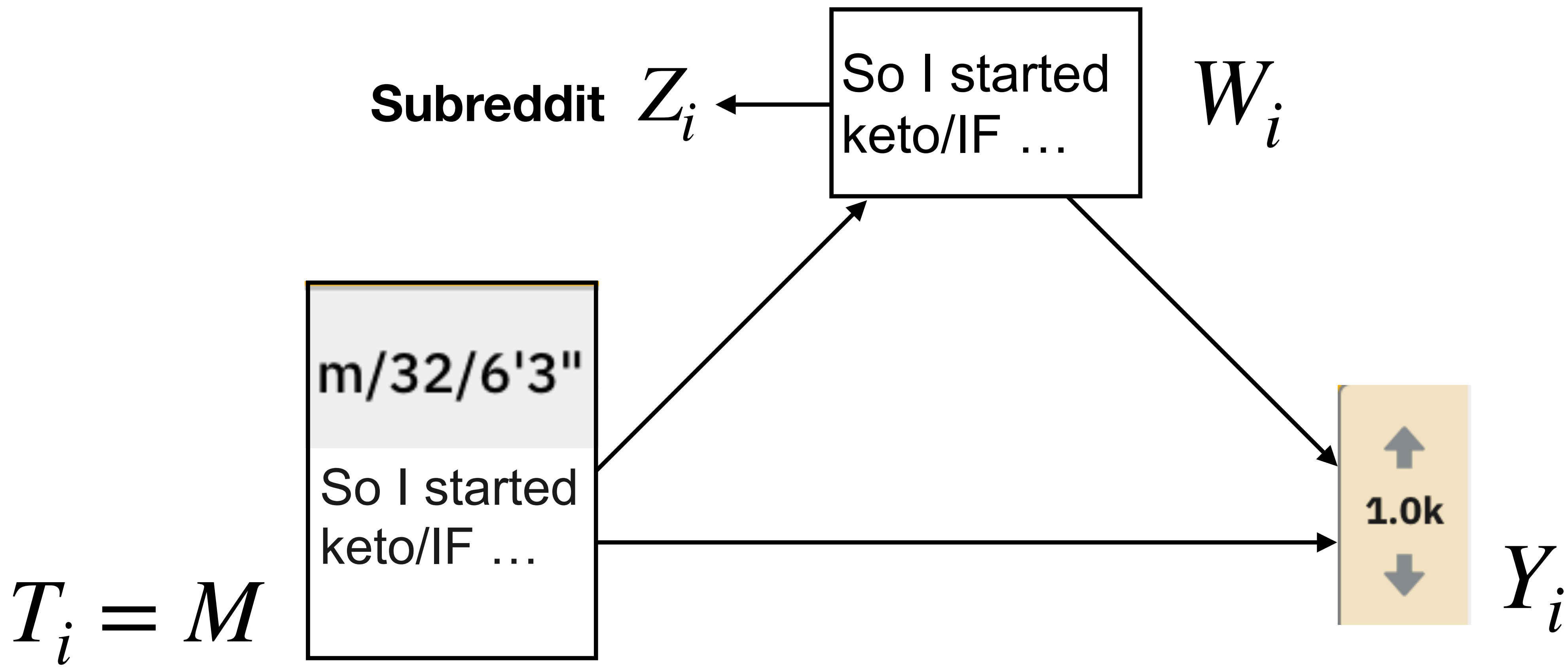


$$Y_i = T_i + \beta_1(\pi(Z_i) - 0.5) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0,\gamma)$$

Treatment effect = 1

Proportion of M in subreddit

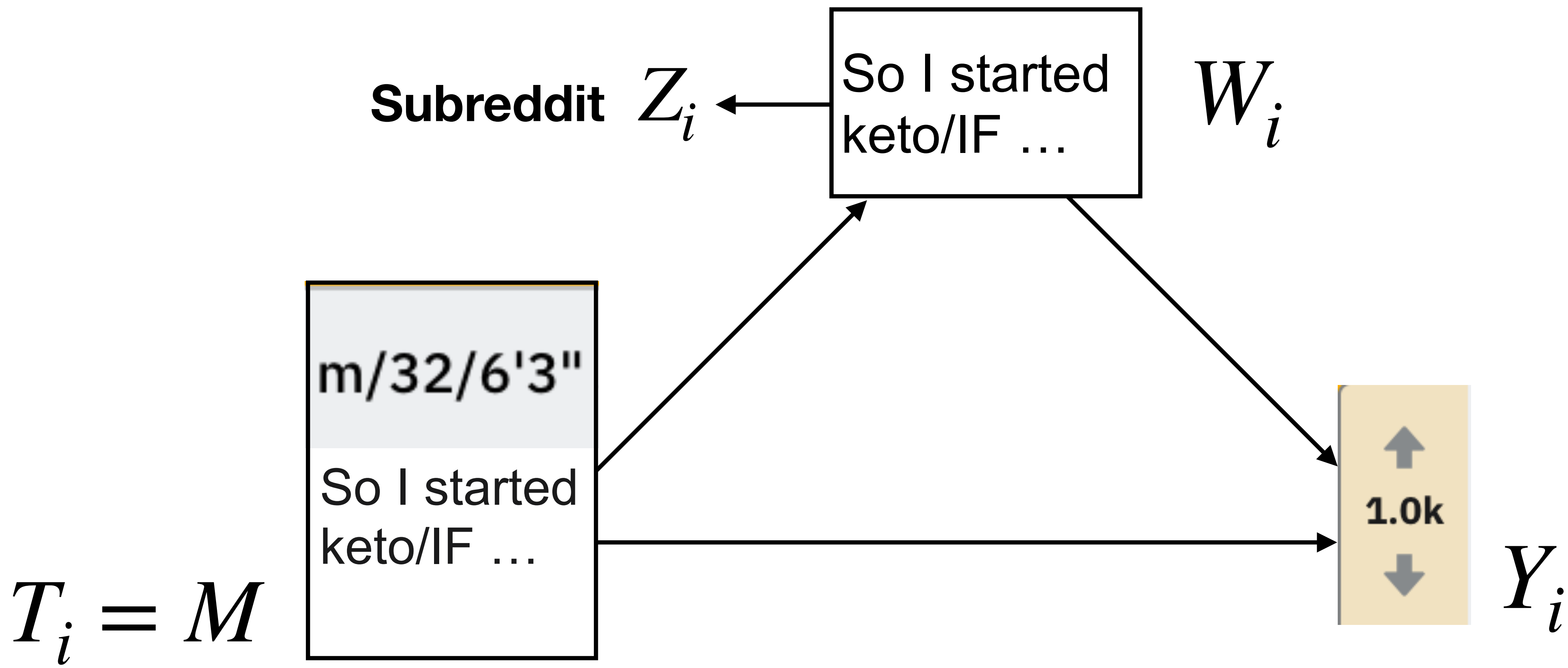
Example 2: Direct Effects



$$Y_i = T_i + \beta_1(\pi(Z_i) - 0.5) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0,\gamma)$$

Posits that subreddits that men typically post in have more popular posts

Example 2: Direct Effects



$$Y_i = T_i + \beta_1(\pi(Z_i) - 0.5) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0,\gamma)$$

Strength of indirect effect

Simulation Studies

Data:

1) **PeerRead:** arXiv papers (cs.cl, cs.lg, or cs.ai) with accept decision, theorem inclusion and buzzy title (‘deep’, ‘neural’, ‘embed’ or ‘adversarial net’)
2) **Reddit:** top-level comments from subreddits with gender labels and upvotes

Simulation Studies

Data:

1) **PeerRead:** arXiv papers (cs.cl, cs.lg, or cs.ai) with accept decision, theorem inclusion and buzzy title (‘deep’, ‘neural’, ‘embed’ or ‘adversarial net’)
2) **Reddit:** top-level comments from subreddits with gender labels and upvotes

Comparisons

1) **BOW:** expected outcomes and propensity score models fit with BOW features
2) **LDA:** models fit with each document’s inferred topic proportions

Reddit Simulation

Reddit: top-level comments from subreddits with gender labels and upvotes

Across two estimators of treatment effect

Noise:	$\gamma = 1.0$			$\gamma = 4.0$		
Confounding:	Low	Med.	High	Low	Med.	High
Ground truth	1.00	1.00	1.00	1.00	1.00	1.00
Unadjusted	1.03	1.24	3.48	0.99	1.22	3.51
Words $\hat{\beta}^{\text{plugin}}$	1.01	1.17	2.69	1.04	1.16	2.63
Words $\hat{\beta}^{\text{TMLE}}$	1.02	1.18	2.71	1.04	1.17	2.65
LDA $\hat{\beta}^{\text{plugin}}$	1.01	1.20	2.95	1.02	1.19	2.91
LDA $\hat{\beta}^{\text{TMLE}}$	1.01	1.20	2.96	1.02	1.19	2.91
$\hat{\beta}^{\text{plugin}}$	0.96	1.05	1.24	0.83	0.63	1.31
$\hat{\beta}^{\text{TMLE}}$	0.98	1.05	1.58	0.95	1.00	1.51

PeerRead Simulation

PeerRead: arXiv papers (cs.cl, cs.lg, or cs.ai) with accept decision, theorem metadata and buzzy title

Confounding:	Low	Med.	High
Ground truth	0.06	0.05	0.03
Unadjusted	0.08	0.15	0.16
Words $\hat{\psi}^{Q}$	0.07	0.13	0.15
Words $\hat{\psi}^{\mathrm{TMLE}}$	0.07	0.13	0.15
LDA $\hat{\psi}^{Q}$	0.06	0.06	0.06
LDA $\hat{\psi}^{\mathrm{TMLE}}$	0.06	0.06	0.06
$\hat{\psi}^{Q}$	0.07	0.06	−0.01
$\hat{\psi}^{\mathrm{TMLE}}$	0.06	0.07	0.04

Example 1: Effect of Theorems

Does including a theorem in my paper cause it to get accepted?

```
\begin{theorem}
...
\end{theorem}
```

	buzzy	theorem
Unadjusted	0.08 ± 0.01	0.21 ± 0.01
$\hat{\psi}^{Q}$	0.01 ± 0.03	0.03 ± 0.03
$\hat{\psi}^{\mathrm{TMLE}}$	0.06 ± 0.04	0.10 ± 0.03

On PeerRead

Conclusions

1. Adapted black-box embedding method, e.g., BERT, to obtain embeddings that can be used to make valid causal inferences.

2. Using metadata like subreddit and buzzy title, which text encodes, we simulated outcomes that are affected by confounders or mediators.

3. Empirical studies suggested that Causal BERT embedding best captures the information in text that’s needed for adjustment.

Conclusions

1. Adapted black-box embedding method, e.g., BERT, to obtain embeddings that can be used to make valid causal inferences.

2. Using metadata like subreddit and buzzy title, which text encodes, we simulated outcomes that are affected by confounders or mediators.

3. Empirical studies suggested that Causal BERT embedding best captures the information in text that's needed for adjustment.

Code and data: github.com/blei-lab/causal-text-embeddings

Contact: {vveitch, dhanya.sridhar}@columbia.edu