Causal effects of language aspects

Dhanya Sridhar

NLP+CSS Tutorial
April 15, 2022

Pryzant, Card, Jurafsky, Veitch, **Sridhar.** 2021. *Causal effects of linguistic properties.* In NAACL.

arxiv.org/abs/2010.12919 **github.com/rpryzant/causal-text**

Email: dhanya.sridhar@mila.quebec

Motivating example

An official website of the United States government

Consumer Financial Protection Bureau

Product

Checking or savings account

Sub-product: Checking account

Issue

Problem caused by your funds being low

Sub-issue: Overdrafts and overdraft fees

Consumer consent to publish narrative

Consent provided

Timely response?

Yes

Motivating example

An official website of the United States government

Consumer Financial Protection Bureau

Product

Checking or savings account

Sub-product: Checking account

Issue

Problem caused by your funds being low

Sub-issue: Overdrafts and overdraft fees

Consumer consent to publish narrative

Consent provided

Timely response?

Yes

IF YOU'RE GOING TO CREATE BANKS. FIX THE XXXX OBVIOUS ABUSE OF POWER. DO YOUR JOB. THE FACT THAT OVERDRAFT FEES EXIST PROVE HOW MUCH YOU XXXX XXXX AND SHOULD NEVER HAVE HAD OVERSIGHT. I WOULD CREATE A NEW BANK TO FIX THIS PROBLEM, IF YOU LET ME. YOU ONL.Y LET THE XXXX EVIL PEOPLE START BANKS.

Motivating example

An official website of the United States government

Consumer Financial Protection Bureau

Product

Checking or savings account

Sub-product: Checking account

Issue

Problem caused by your funds being low

Sub-issue: Overdrafts and overdraft fees

Consumer consent to publish narrative

Consent provided

Timely response?

Yes

IF YOU'RE GOING TO CREATE BANKS. FIX THE XXXX OBVIOUS ABUSE OF POWER. DO YOUR JOB. THE FACT THAT OVERDRAFT FEES EXIST PROVE HOW MUCH YOU XXXX XXXX AND SHOULD NEVER HAVE HAD OVERSIGHT. I WOULD CREATE A NEW BANK TO FIX THIS PROBLEM, IF YOU LET ME. YOU ONL.Y LET THE XXXX EVIL PEOPLE START BANKS.

Does the politeness of the complaint affect response time?

Politeness is latent

We only observe texts and outcomes!

$\mathcal{D} = ($

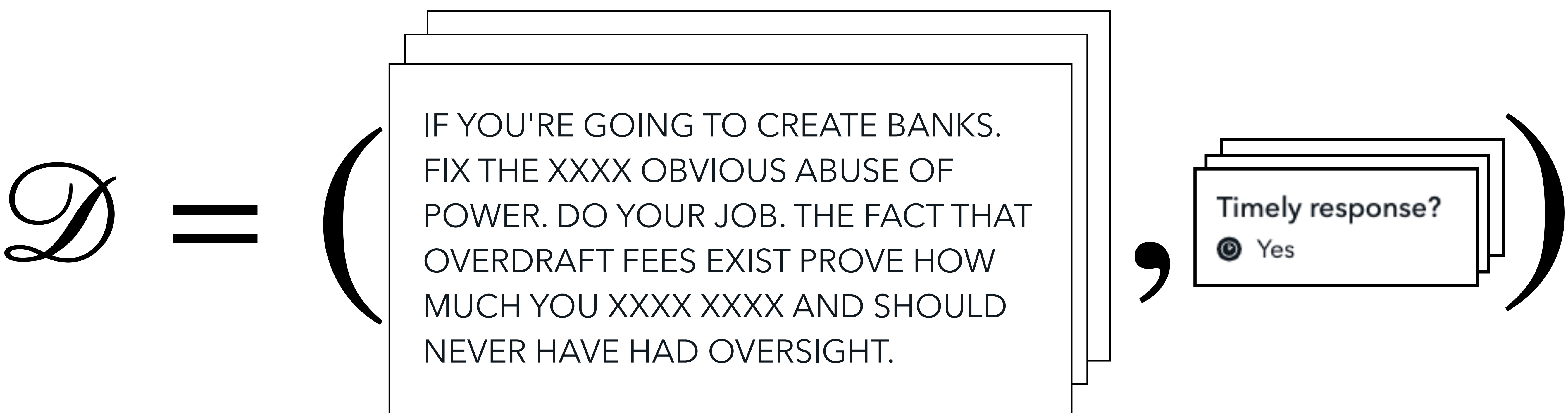
IF YOU'RE GOING TO CREATE BANKS. FIX THE XXXX OBVIOUS ABUSE OF POWER. DO YOUR JOB. THE FACT THAT OVERDRAFT FEES EXIST PROVE HOW MUCH YOU XXXX XXXX AND SHOULD NEVER HAVE HAD OVERSIGHT.

$,$

$)$

Noisy prediction of politeness

We can infer politeness from texts using a lexicon or classifier.



$$T_{\text{proxy}} = f(\text{text})$$

This talk

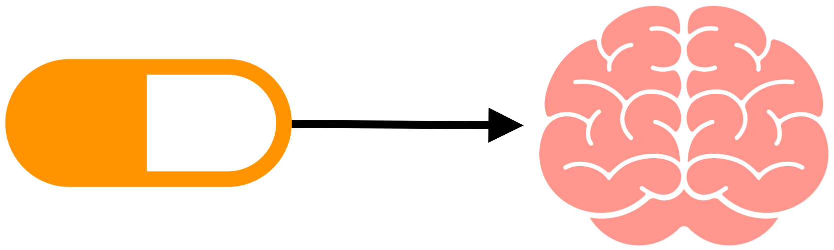
- ❑ How do we formalize the effect of politeness on response times?
- ❑ Is it possible to recover the effect with a proxy of politeness?
- ❑ If it’s possible, how can we do it?

This talk

- ☐ $\boxed{\text{How do we formalize the effect of politeness on response times?}}$
- ☐ Is it possible to recover the effect with a proxy of politeness?
- ☐ If it’s possible, how can we do it?

Example: Clinical setting

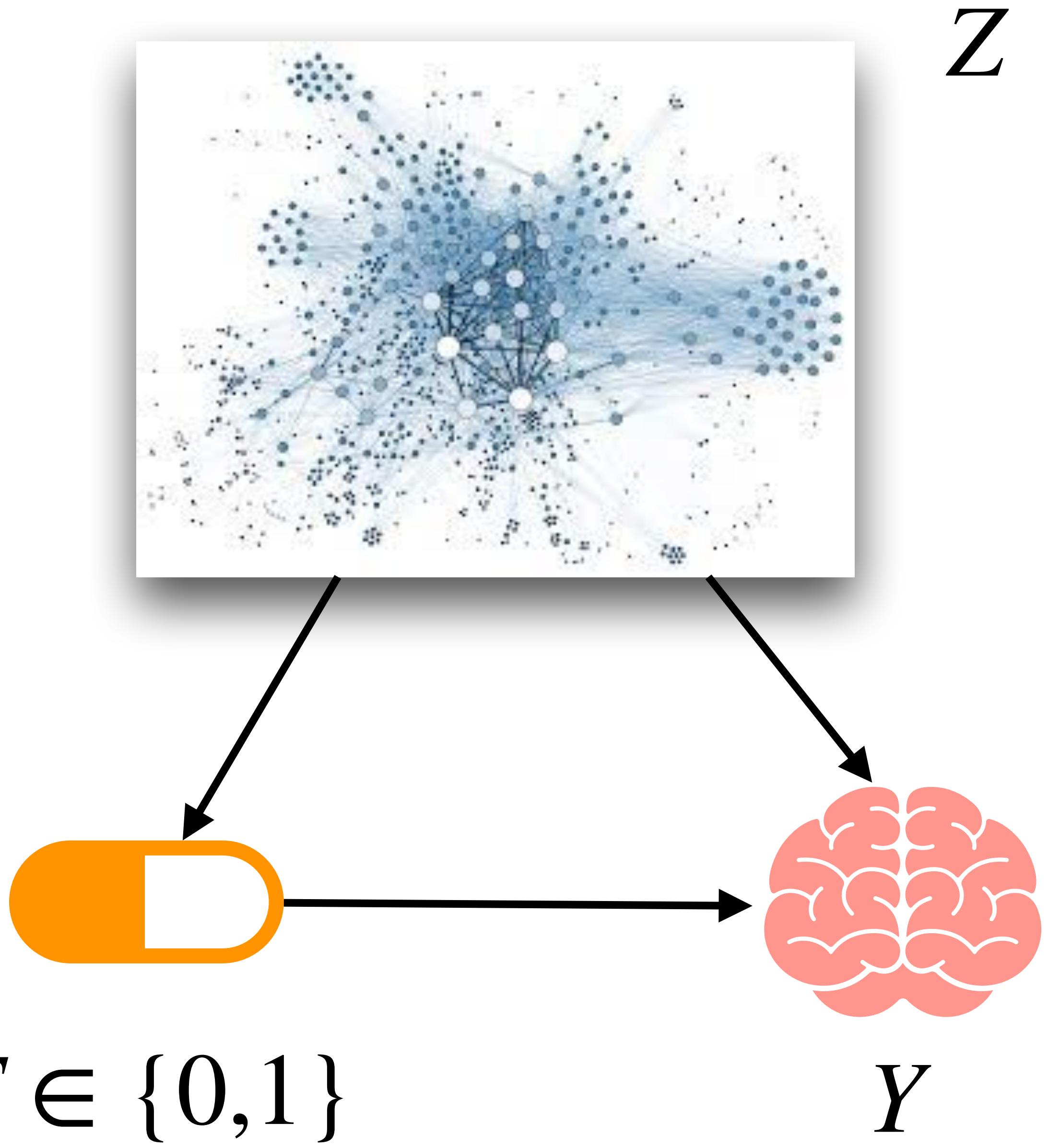
Does an antidepressant have an effect on reported depression levels?



$T \in \{0,1\}$ $\quad$ Y

- This is a causal question.
- A variable X causes Y if manipulating X "changes Y."
- We need a formalism for "manipulating X" and "changes to Y."

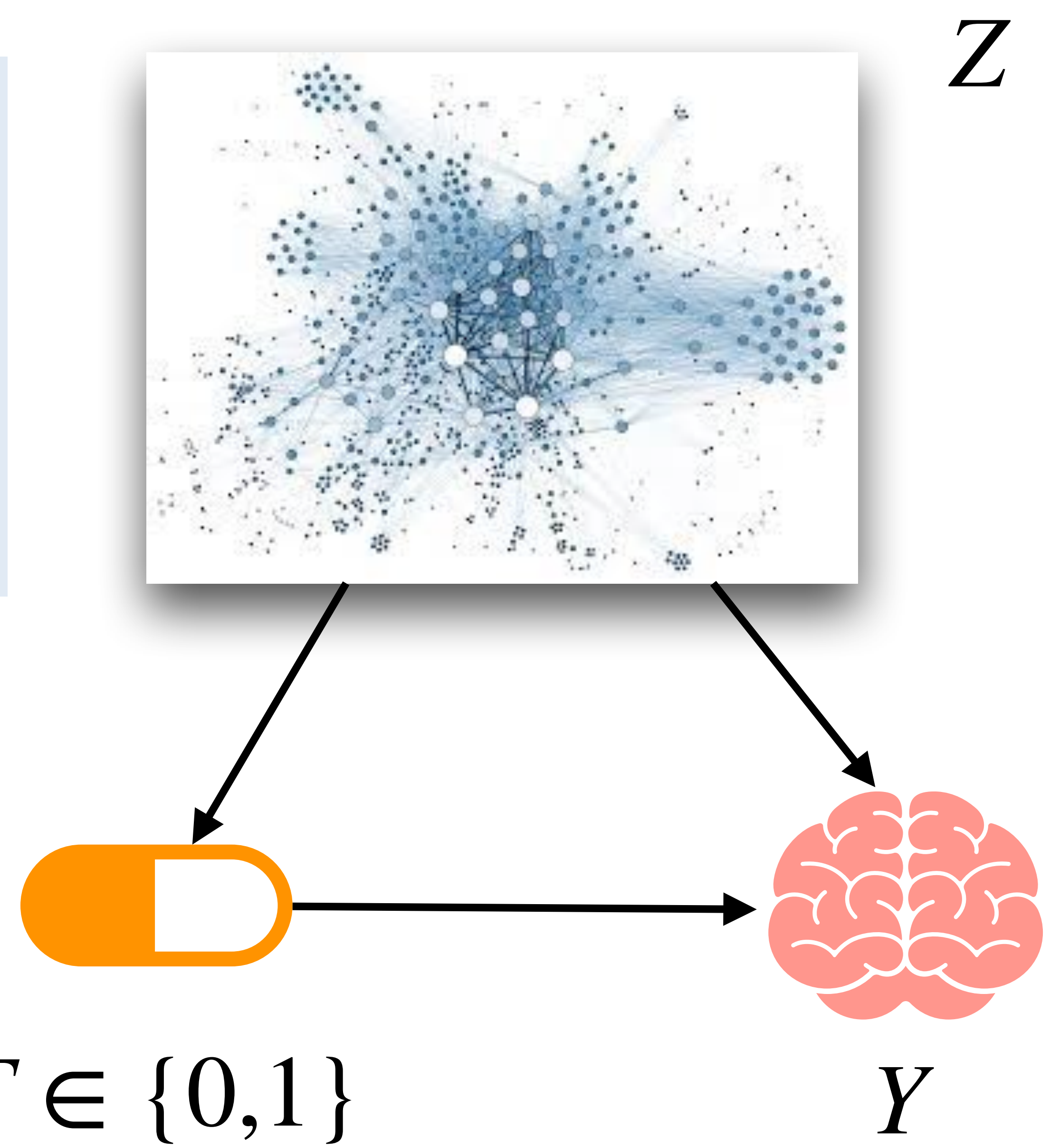
The data generating process



The data generating process

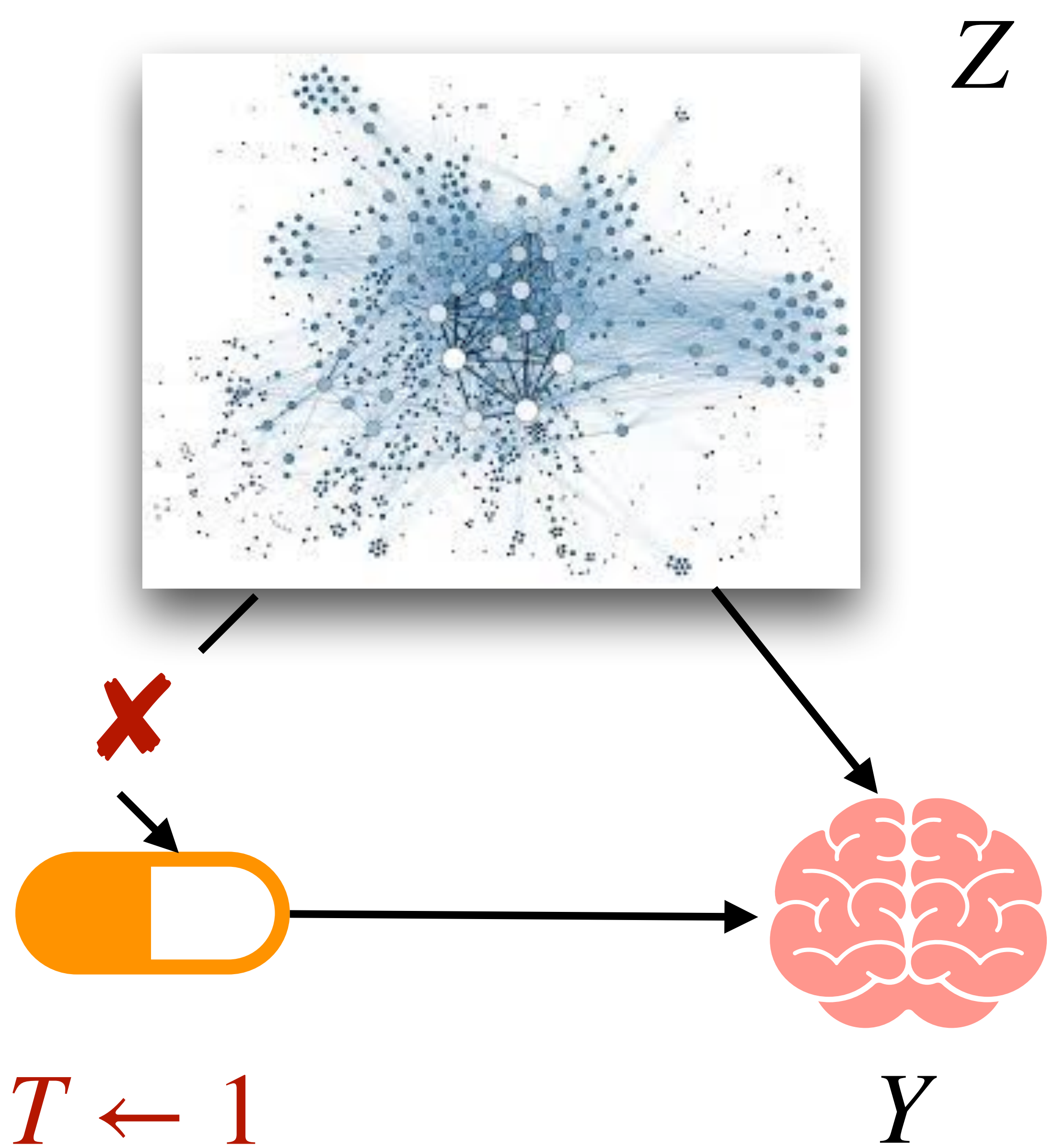
Nature:

$$P(T, Z, Y)$$



Intervening on the system

"Intervene and administer the antidepressant to everyone."

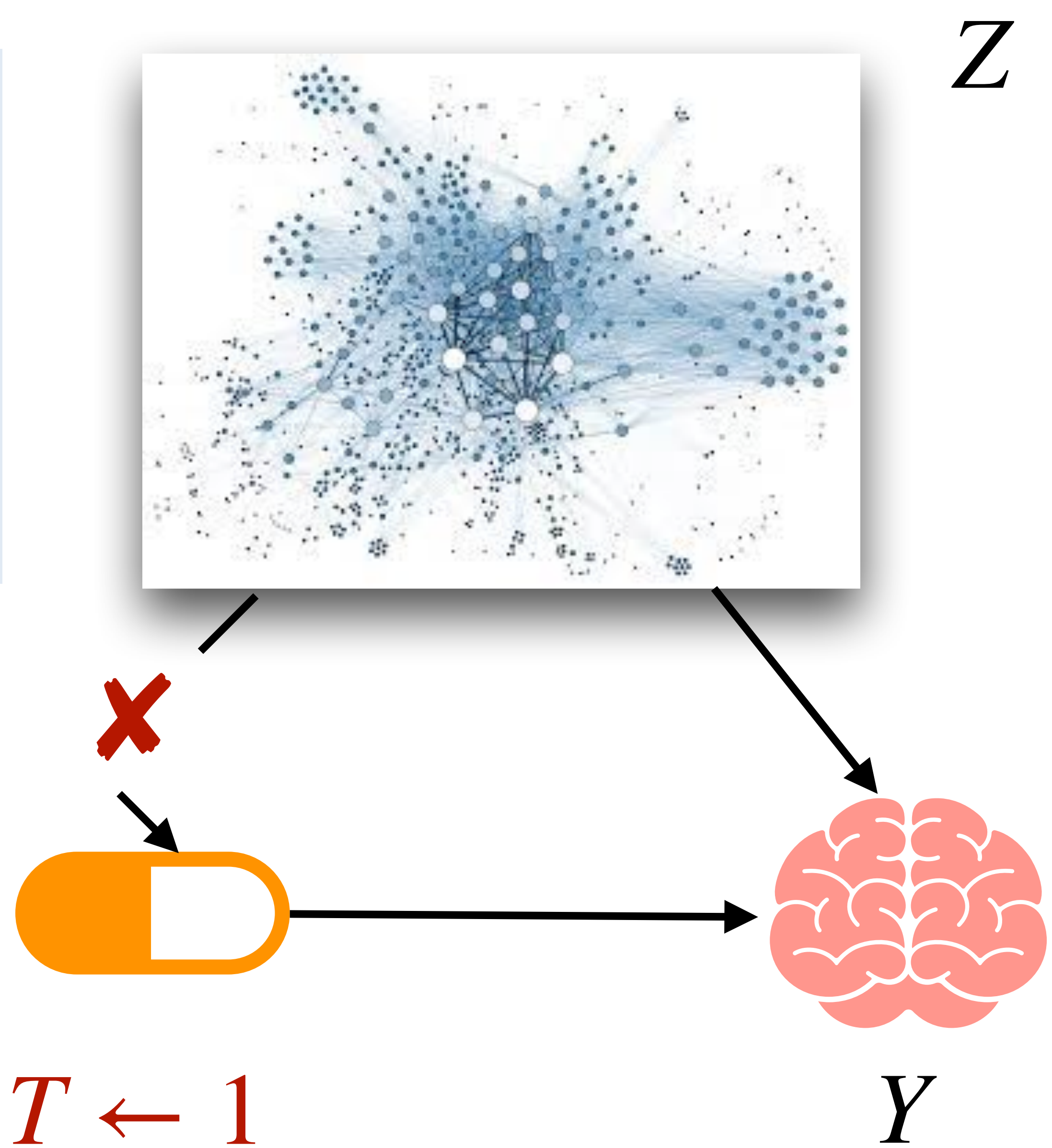


Intervening on the system

"Intervene and administer the antidepressant to everyone."

Intervention:

$$P(Y; \mathrm{do}(T = 1))$$

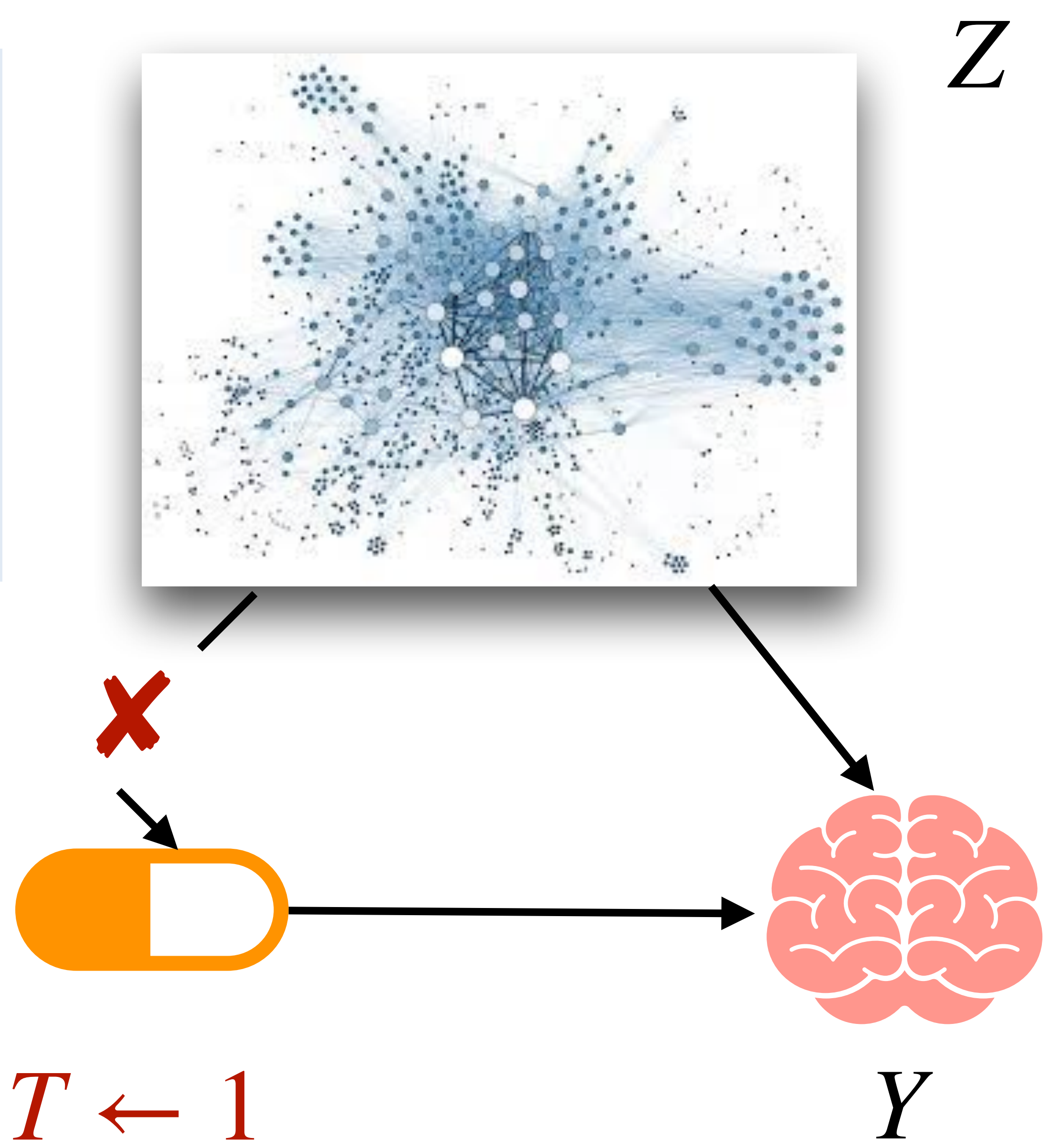


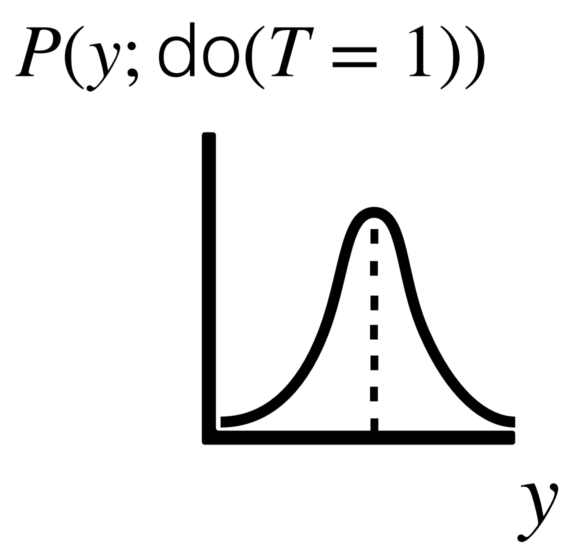
Intervening on the system

This is nothing more than a new distribution we'd like to sample from.

Intervention:

$$P(Y; \mathrm{do}(T = 1))$$





Intervening on the system

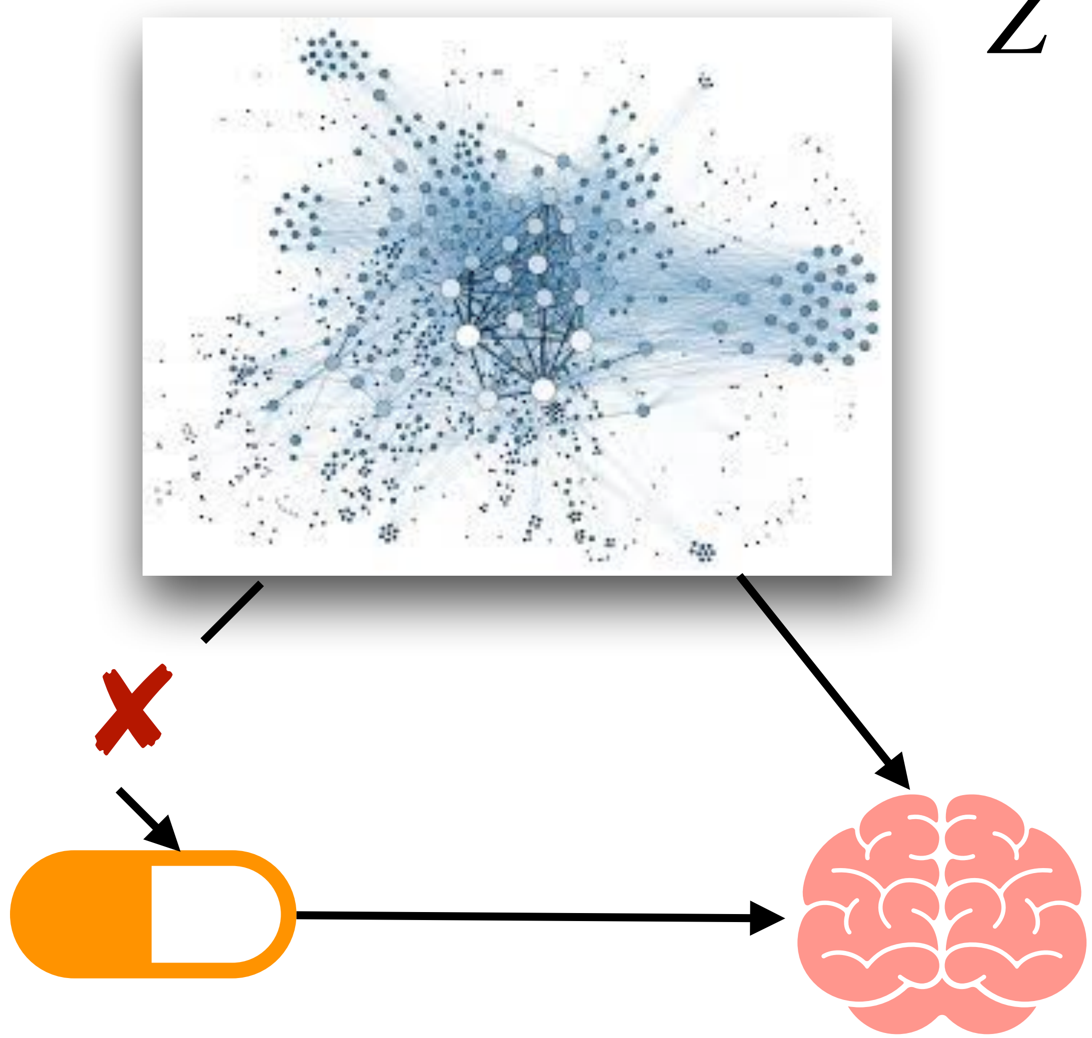
This is nothing more than a new distribution we'd like to sample from.

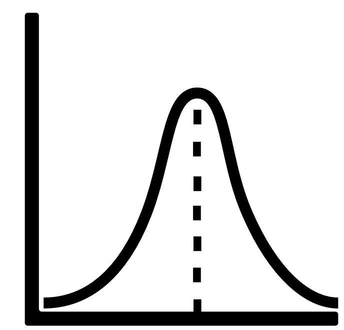
Intervention:

$$P(Y; \mathrm{do}(T = 1))$$

Potential outcome:

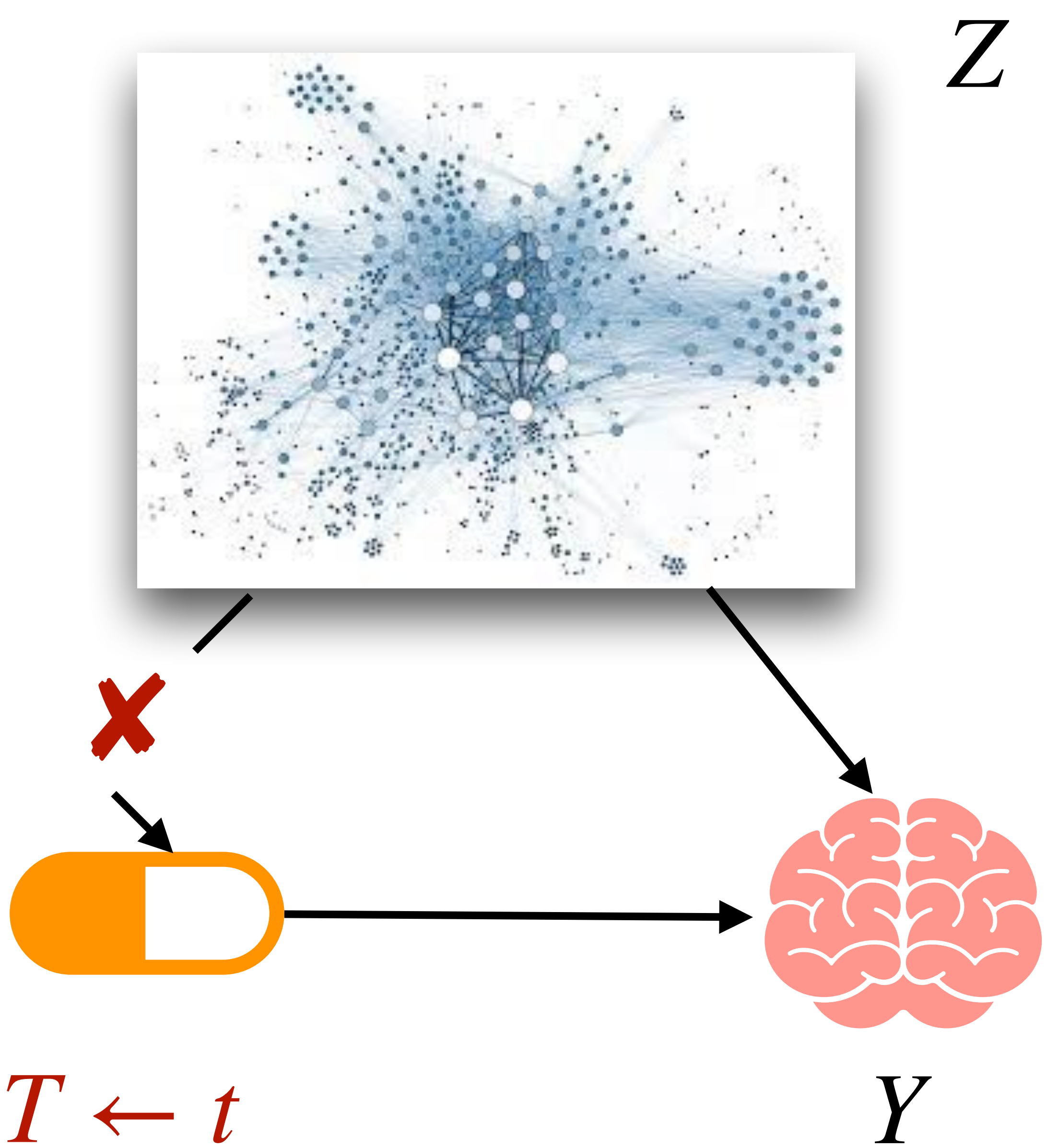
$$Y(1) \sim P$$





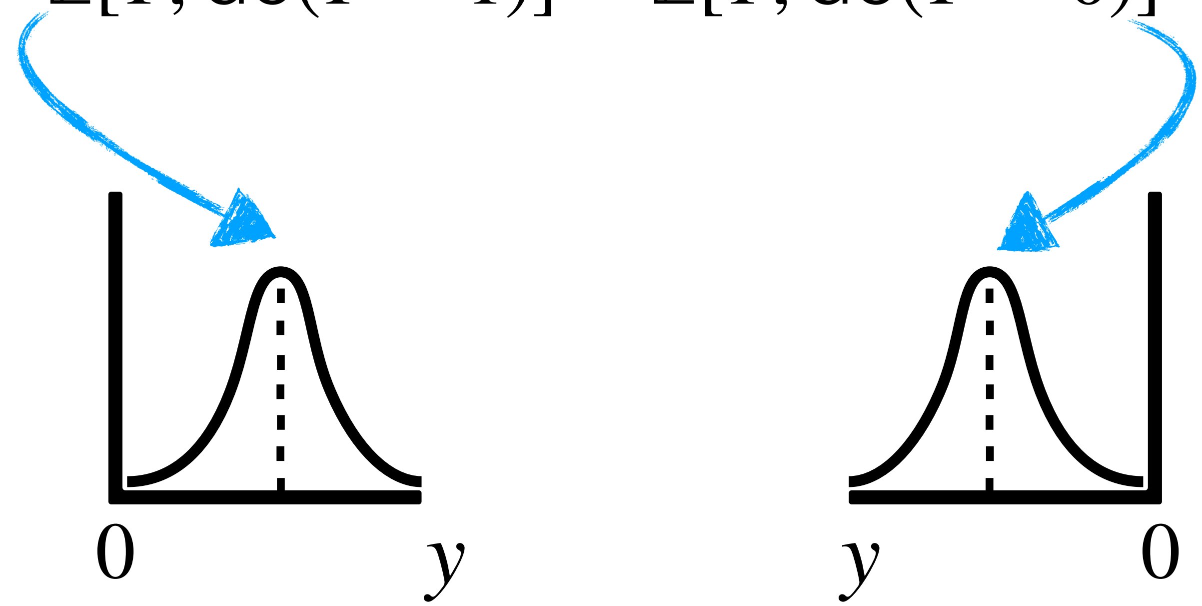
Causal effects as contrasts

What is the causal effect of the antidepressant on depression levels?



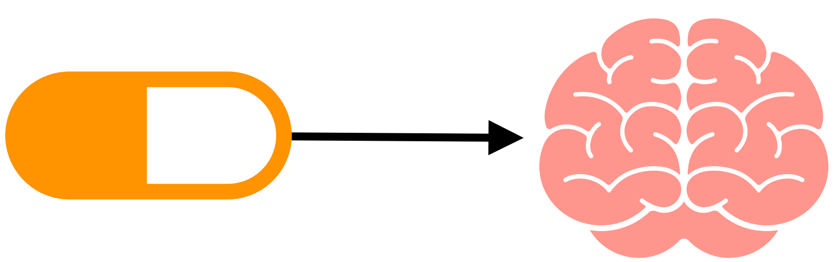
Average treatment effect (ATE)

$$\mathbb{E}[Y; \mathrm{do}(T = 1)] - \mathbb{E}[Y; \mathrm{do}(T = 0)]$$



Example: Clinical setting

Does an antidepressant have an effect on reported depression levels?



$T \in \{0,1\}$ $\quad$ Y

- This is a causal question.
- A variable X causes Y if manipulating X "changes Y."
- We need a formalism for "manipulating X" and "changes to Y."

Using interventions to define text effects can be ambiguous!

Motivating example

An official website of the United States government

Consumer Financial Protection Bureau

Product

Checking or savings account

Sub-product: Checking account

Issue

Problem caused by your funds being low

Sub-issue: Overdrafts and overdraft fees

Consumer consent to publish narrative

Consent provided

Timely response?

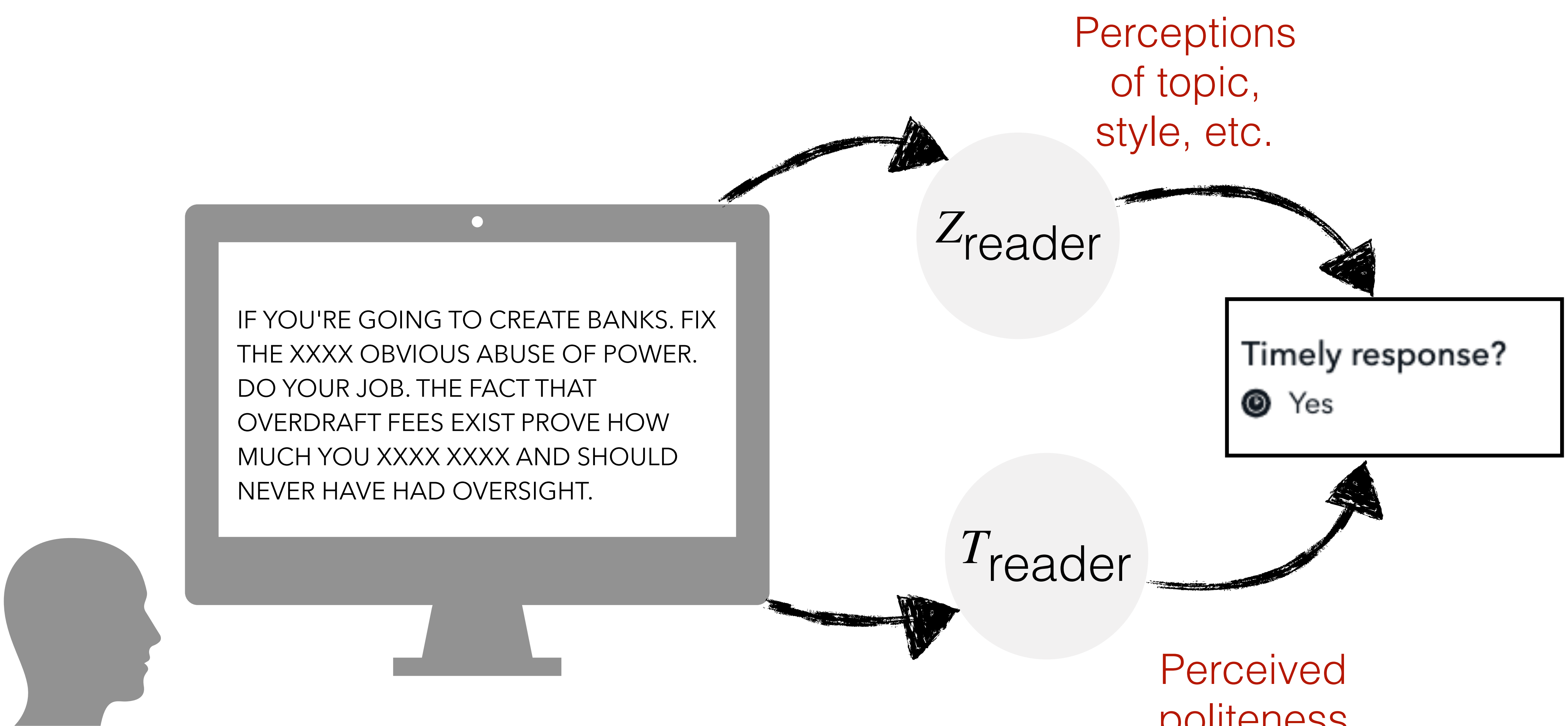
Yes

IF YOU'RE GOING TO CREATE BANKS. FIX THE XXXX OBVIOUS ABUSE OF POWER. DO YOUR JOB. THE FACT THAT OVERDRAFT FEES EXIST PROVE HOW MUCH YOU XXXX XXXX AND SHOULD NEVER HAVE HAD OVERSIGHT. I WOULD CREATE A NEW BANK TO FIX THIS PROBLEM, IF YOU LET ME. YOU ONL.Y LET THE XXXX EVIL PEOPLE START BANKS.

Does the politeness of the complaint affect response time?

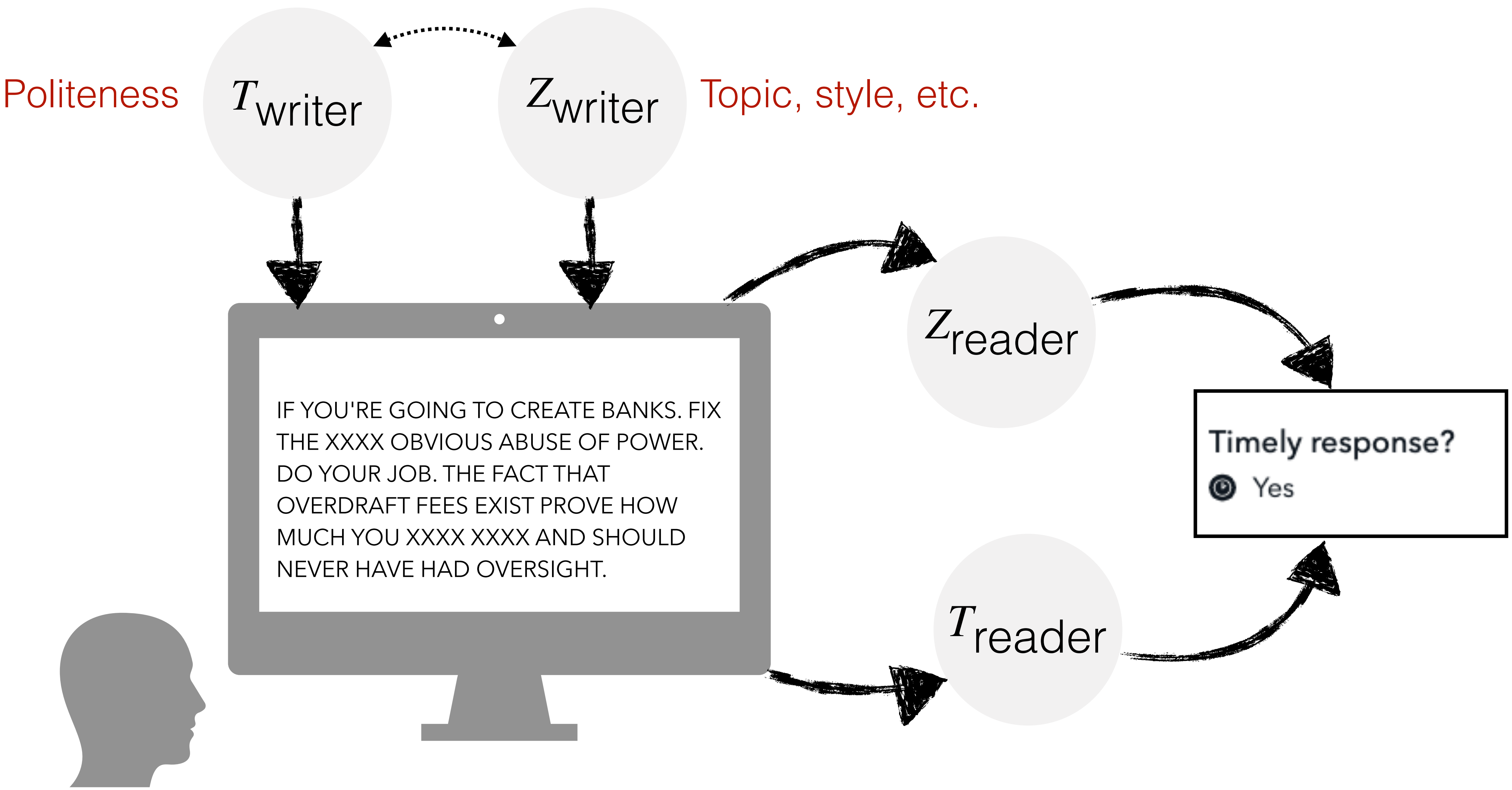
Assumed causal model

Reader responds based on perceived politeness and other properties of the text.



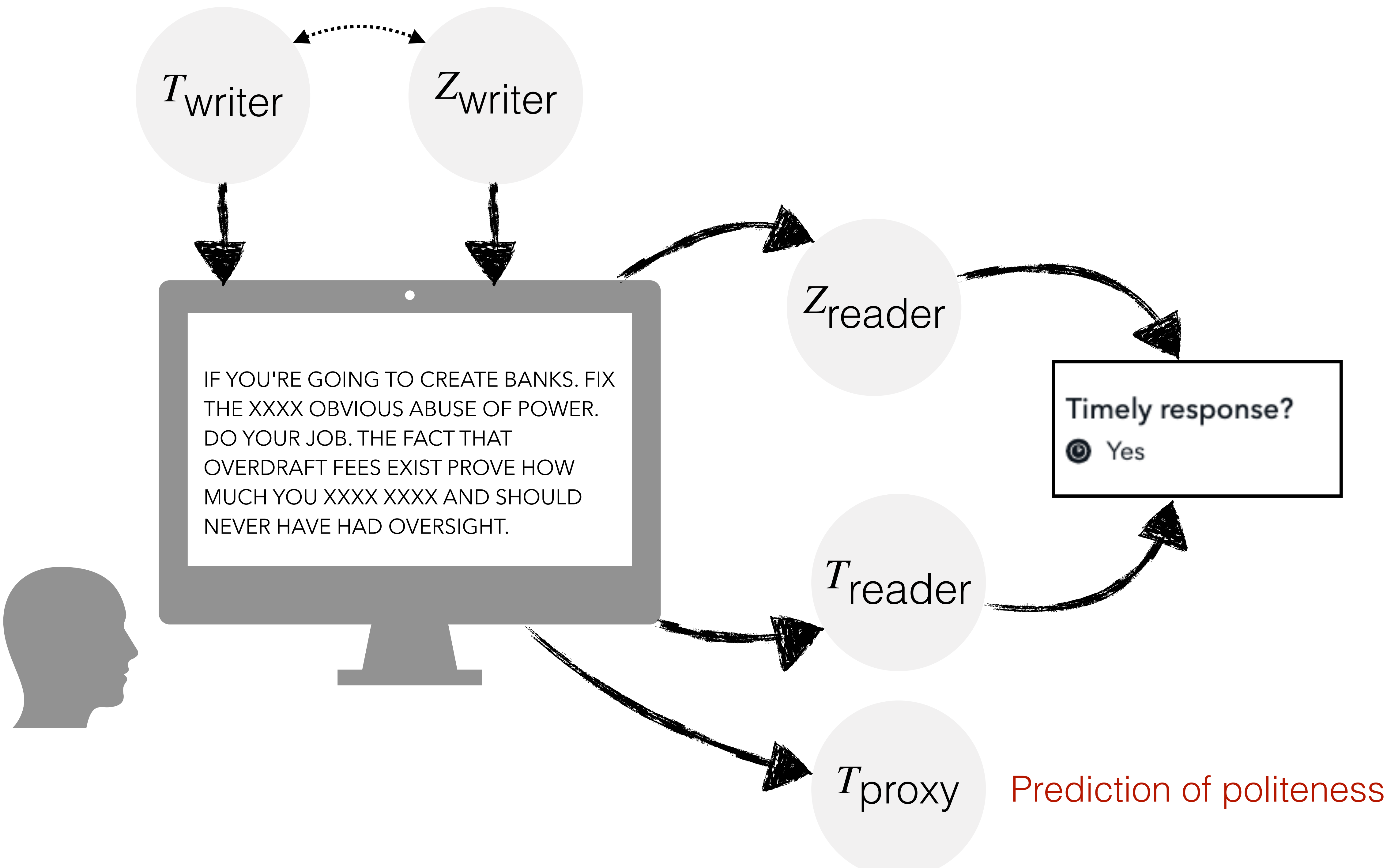
Assumed causal model

The writer controls the generative process of text.



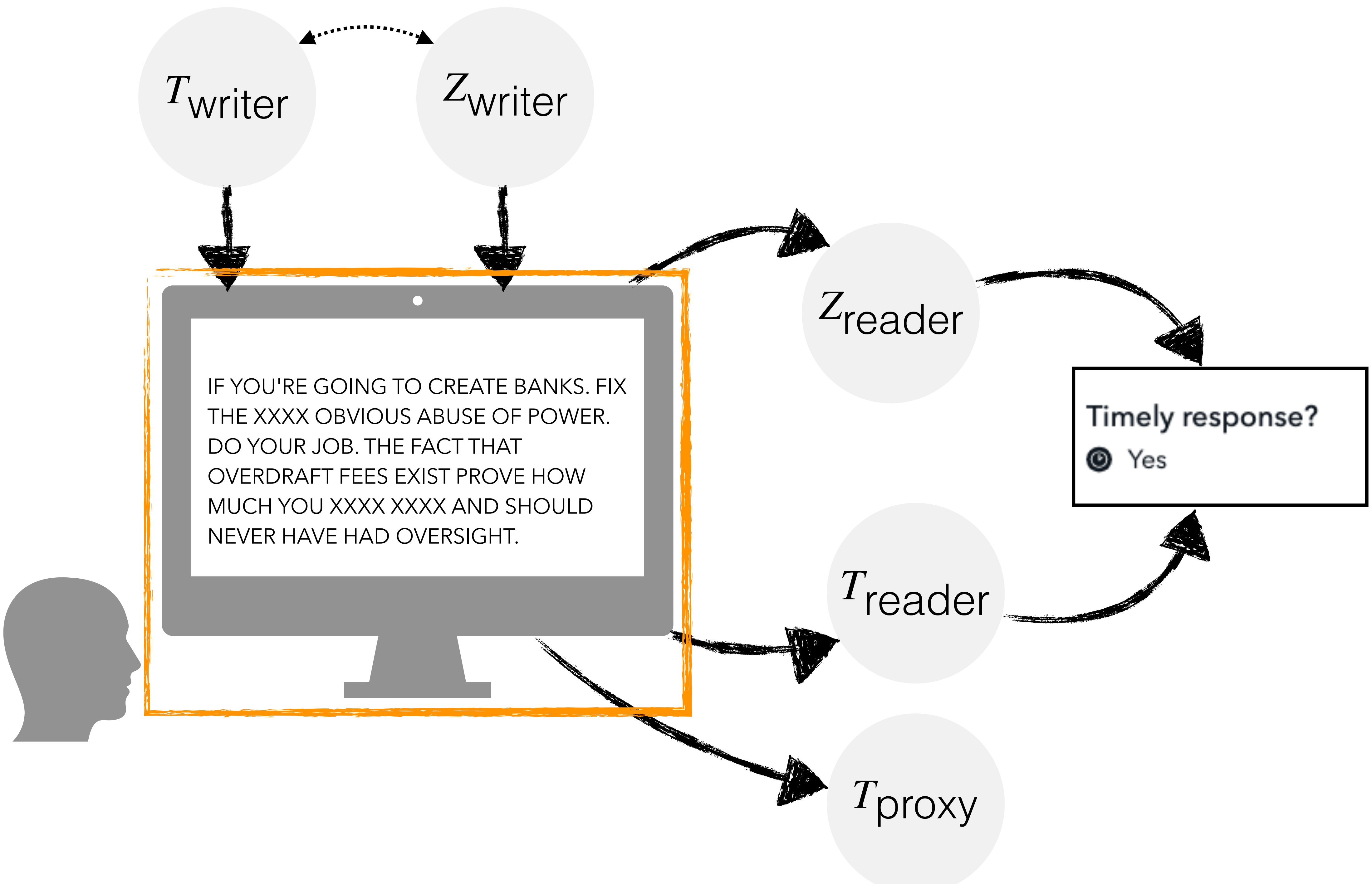
Assumed causal model

We use separately trained classifiers or lexicons to produce a noisy treatment.



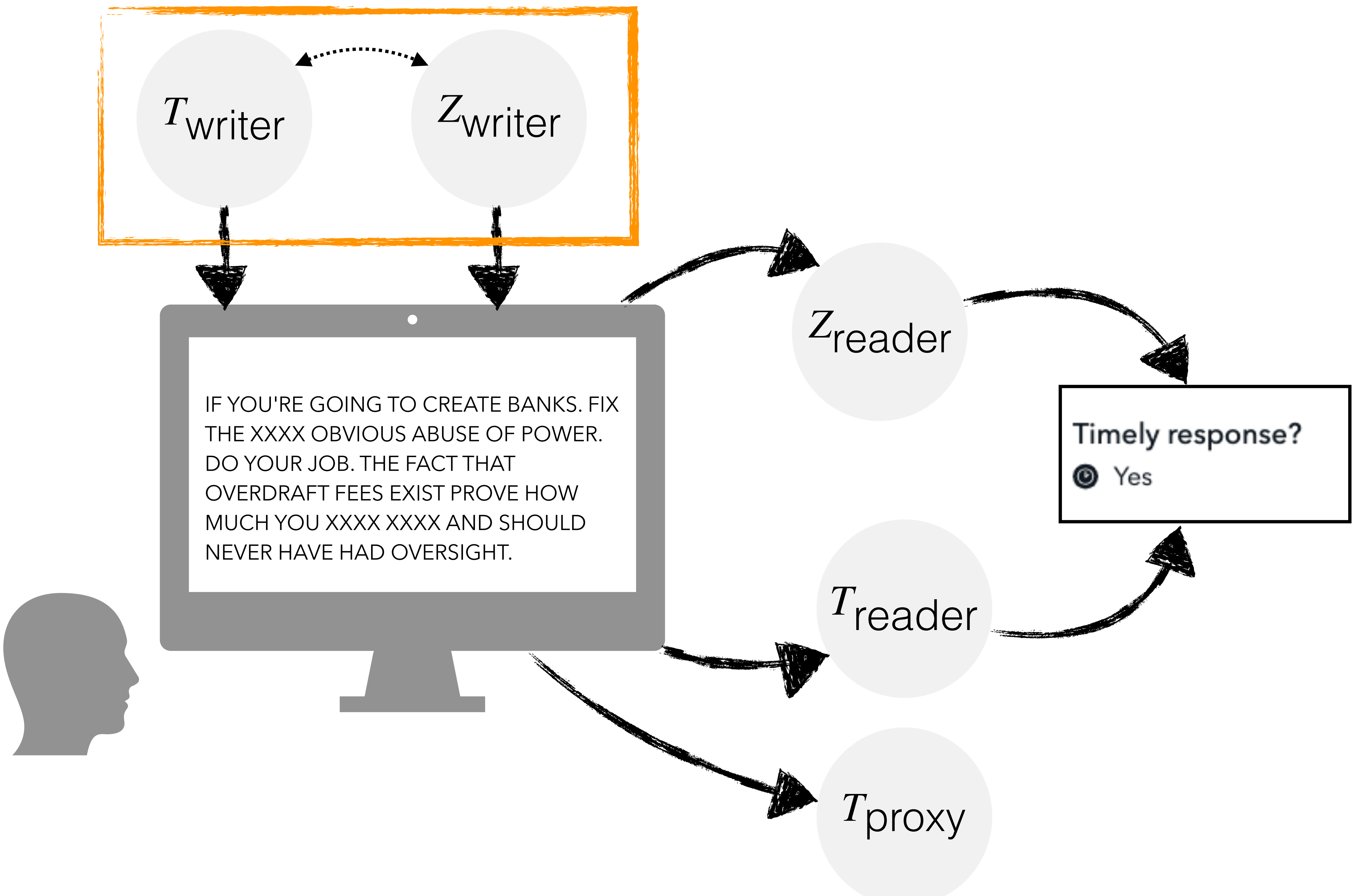
Assumed causal model

First, notice a symmetry between T_{reader} and T_{proxy}.



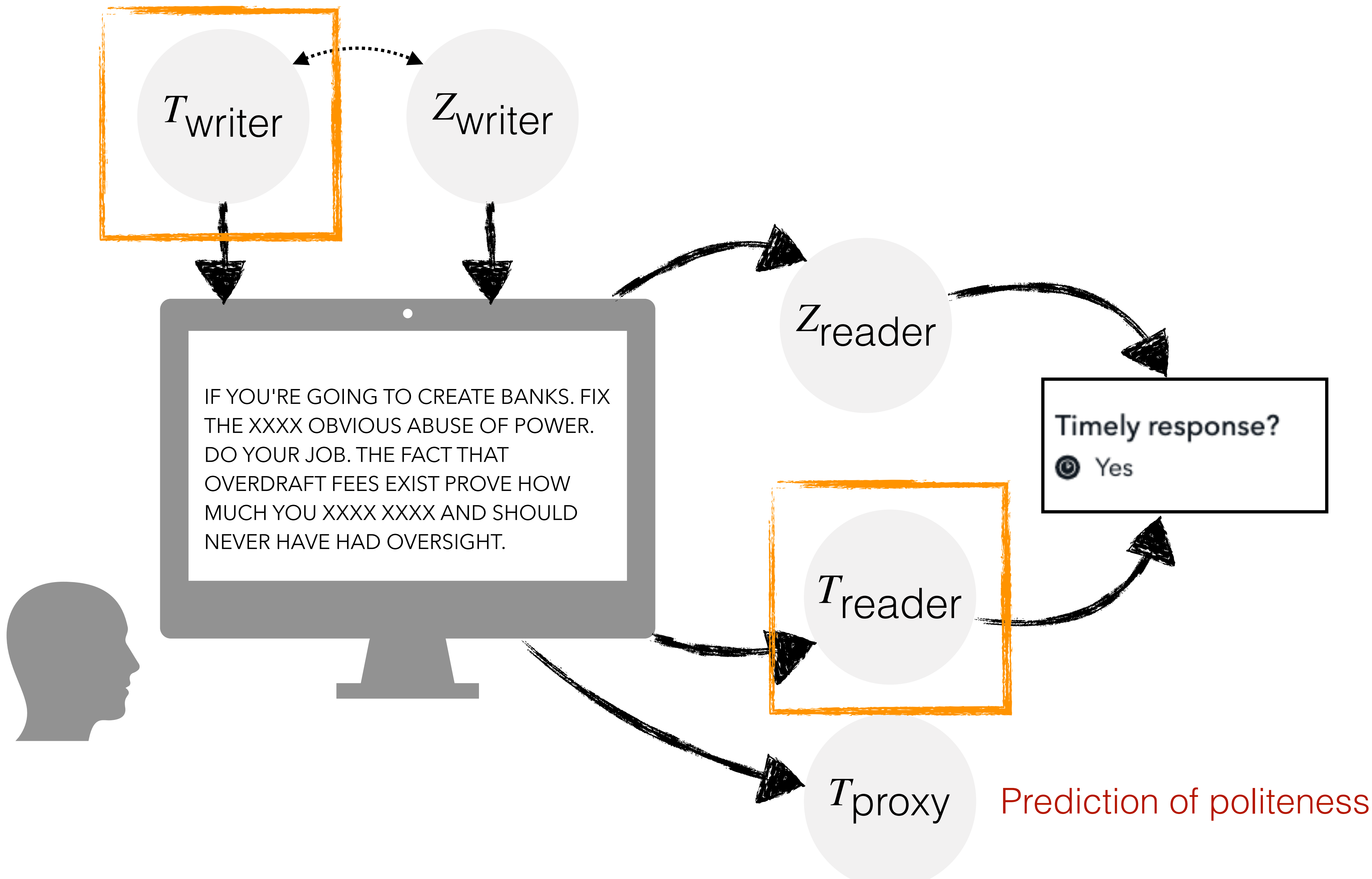
Assumed causal model

Text doesn't screen the writer's intended politeness from the outcome.



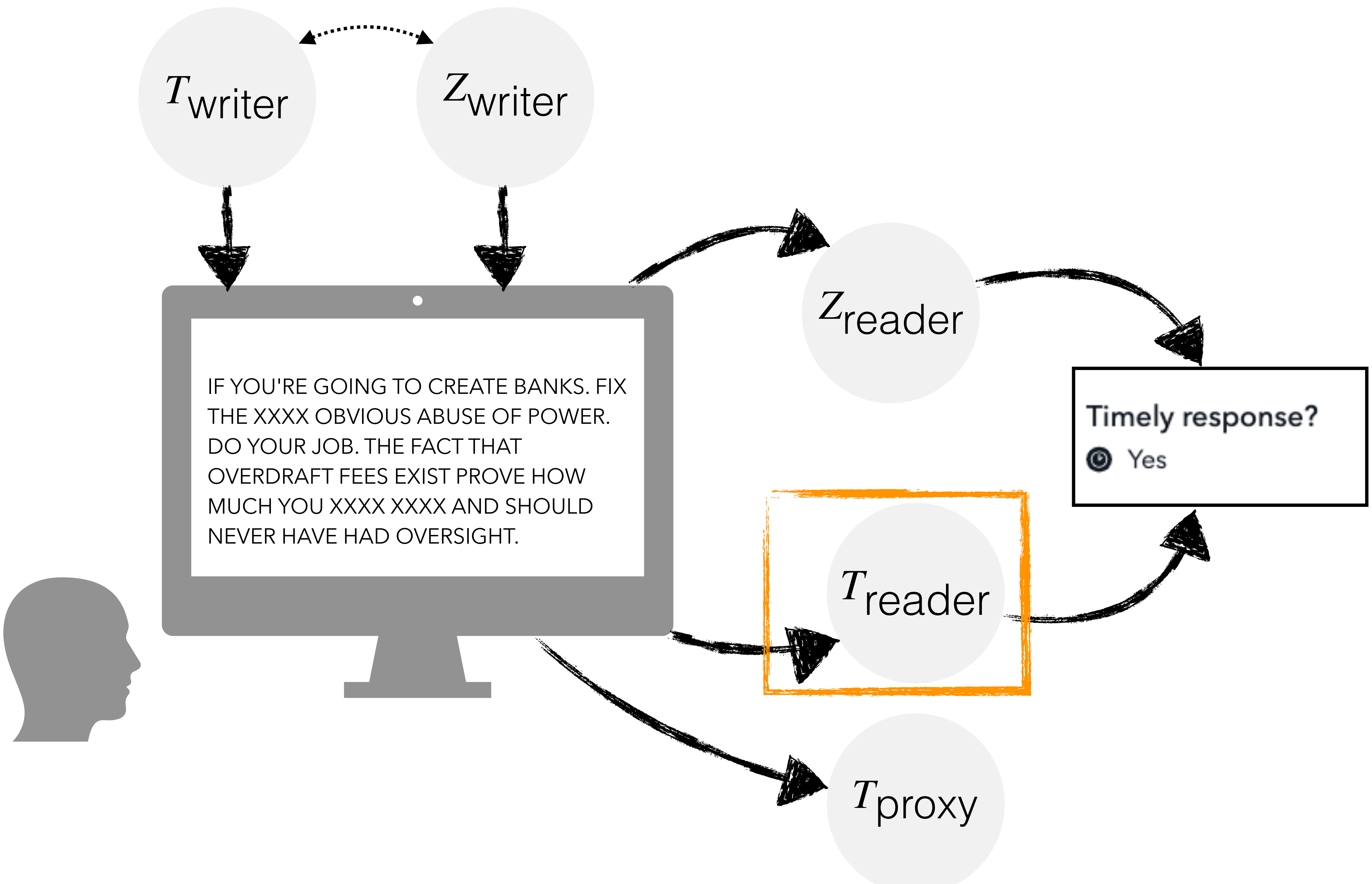
Assumed causal model

From which agent's perspective should we consider interventions?



Defining the causal question

$$\beta = \mathbb{E}[Y; \mathrm{do}(T_{\text{reader}} = 1)] - \mathbb{E}[Y; \mathrm{do}(T_{\text{reader}} = 0)]$$



This talk

- [x] How do we formalize the effect of politeness on response times?
- [ ] Is it possible to recover the effect with a proxy of politeness?
- [ ] If it's possible, how can we do it?

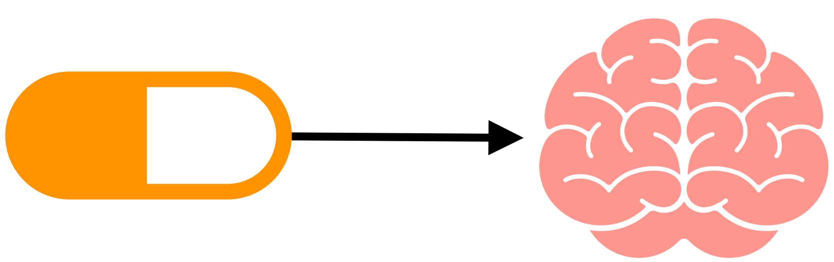
This talk

- [x] How do we formalize the effect of politeness on response times?
- [ ] Is it possible to recover the effect with a proxy of politeness?
- [ ] If it's possible, how can we do it?

So far, we haven't talked about making causal inferences!

Example: Clinical setting

Does an antidepressant have an effect on reported depression levels?

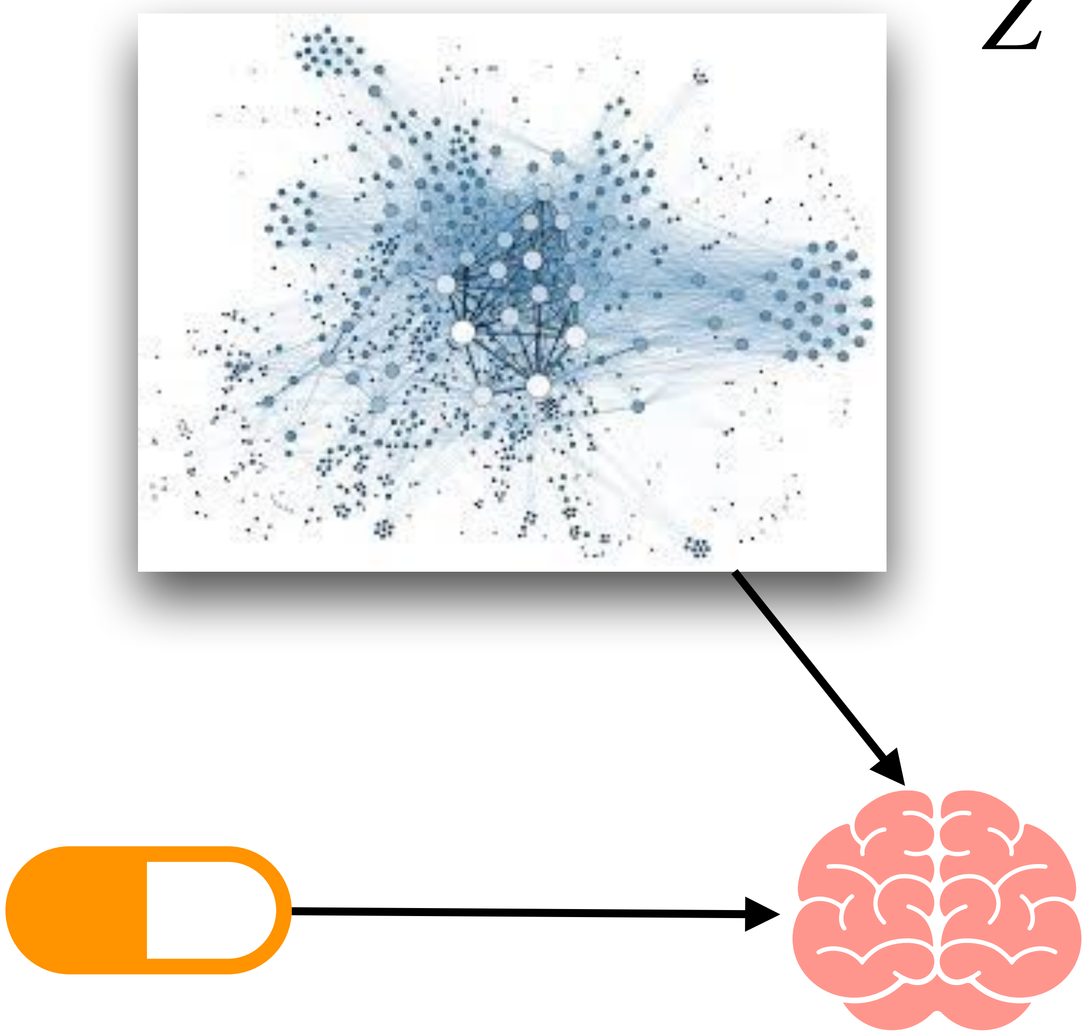


$T \in \{0,1\}$ $\quad$ Y

- This is a causal question.
- A variable X causes Y if manipulating X "changes Y."
- We need a formalism for "manipulating X" and "changes to Y."

When is causal inference possible?

Randomized control trials (RCTs)



Z

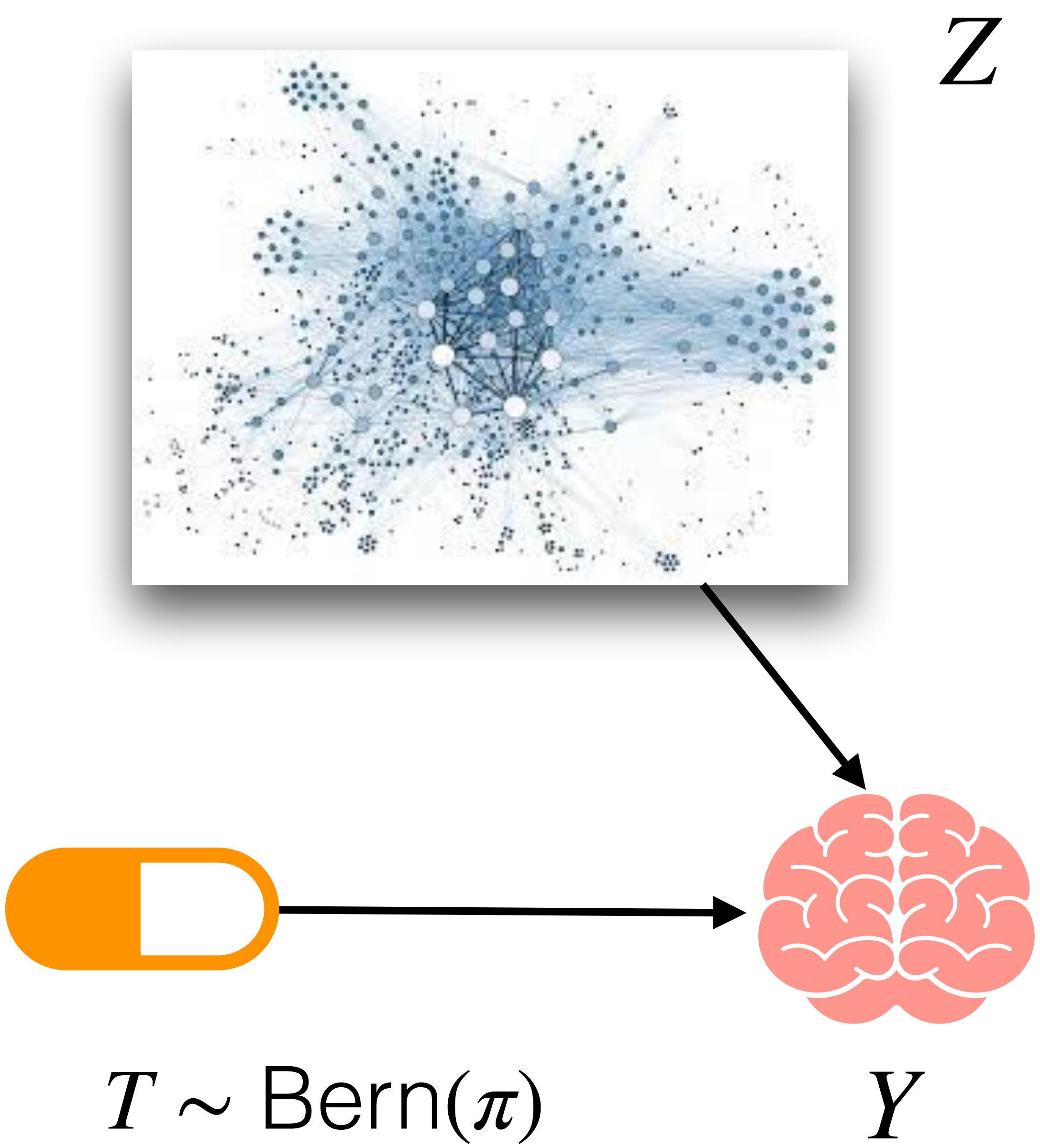
$T \sim \text{Bern}(\pi)$

Y

Average treatment effect (ATE)

$$\mathbb{E}[Y; \text{do}(T=1)] - \mathbb{E}[Y; \text{do}(T=0)]$$

Randomized control trials (RCTs)



Average treatment effect (ATE)

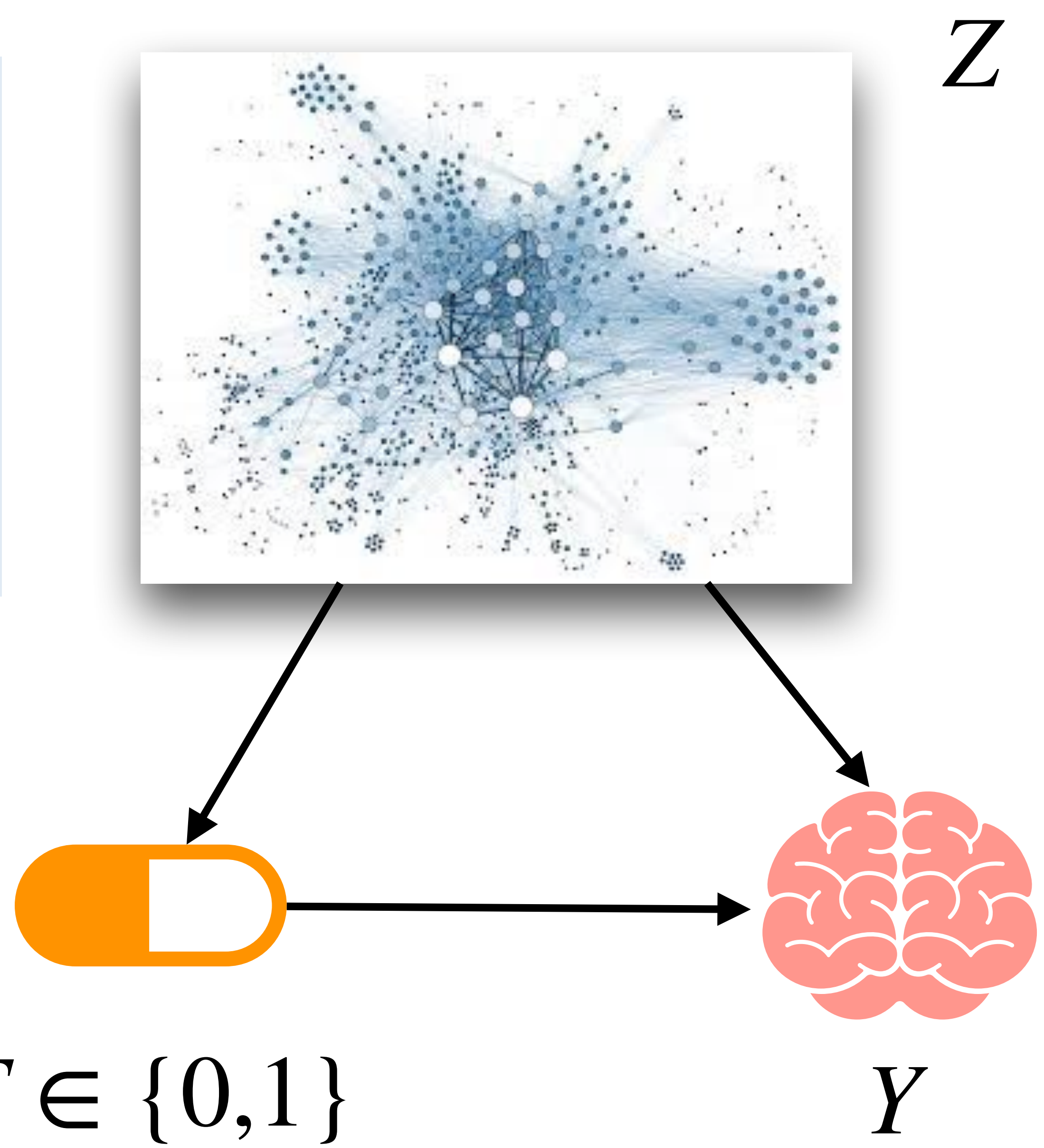
$$\mathbb{E}[Y; \mathrm{do}(T = 1)] - \mathbb{E}[Y; \mathrm{do}(T = 0)]$$

$$= \mathbb{E}[Y \,|\, T = 1] - \mathbb{E}[Y \,|\, T = 0]$$

Observational setting

Nature:

$$P(T, Z, Y)$$



Observational setting

Treatment	Not Depressed
1	0
1	0
1	1
1	1

Treatment	Not Depressed
0	1
0	0
0	0
0	1

$$\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = 0$$

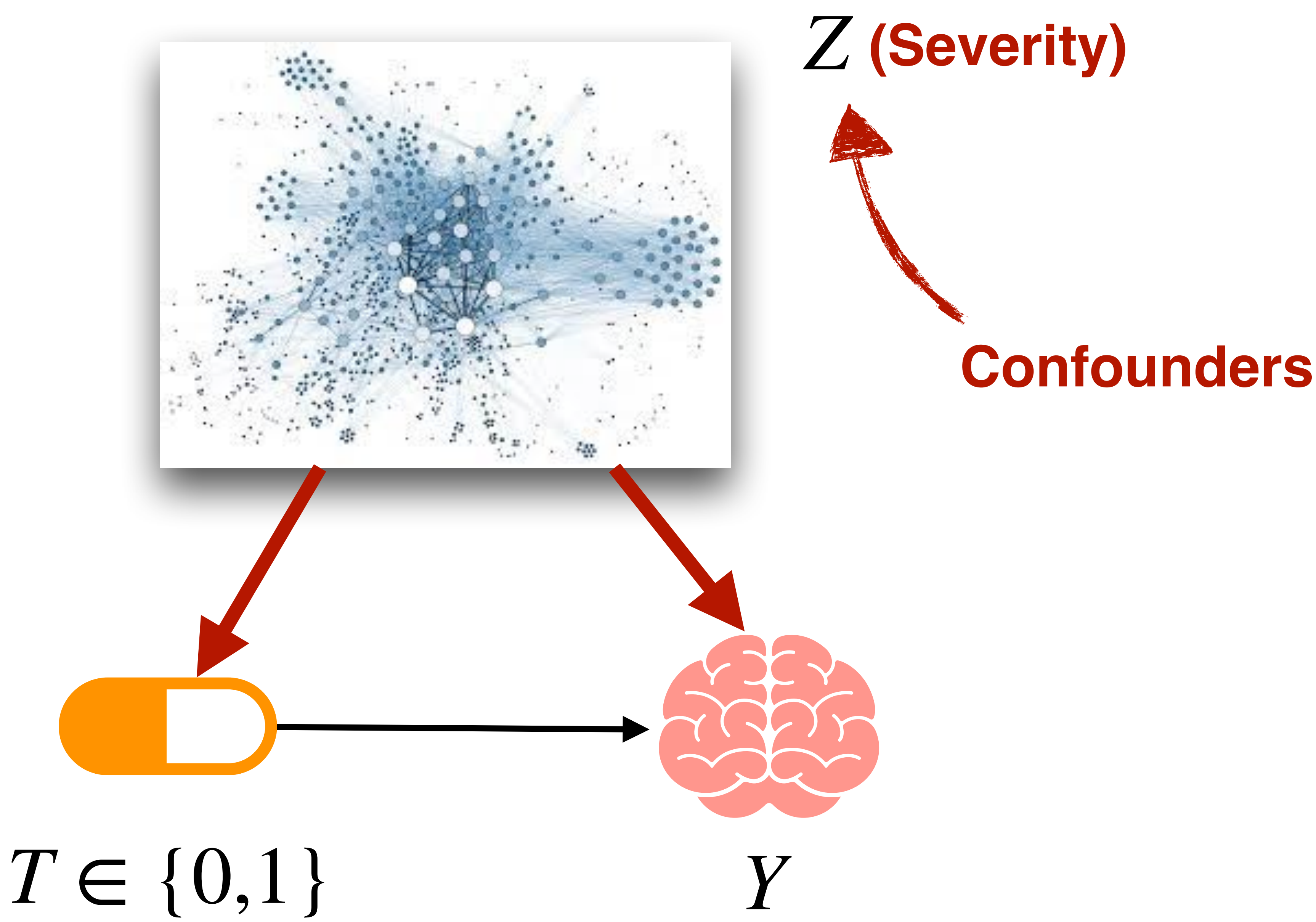
What happened?

Observational setting

Treatment	Not Depressed	Severity
1	0	1
1	0	1
1	1	0
1	1	0

Treatment	Not Depressed	Severity
0	1	0
0	0	1
0	0	0
0	1	0

Observational setting



Causal adjustment

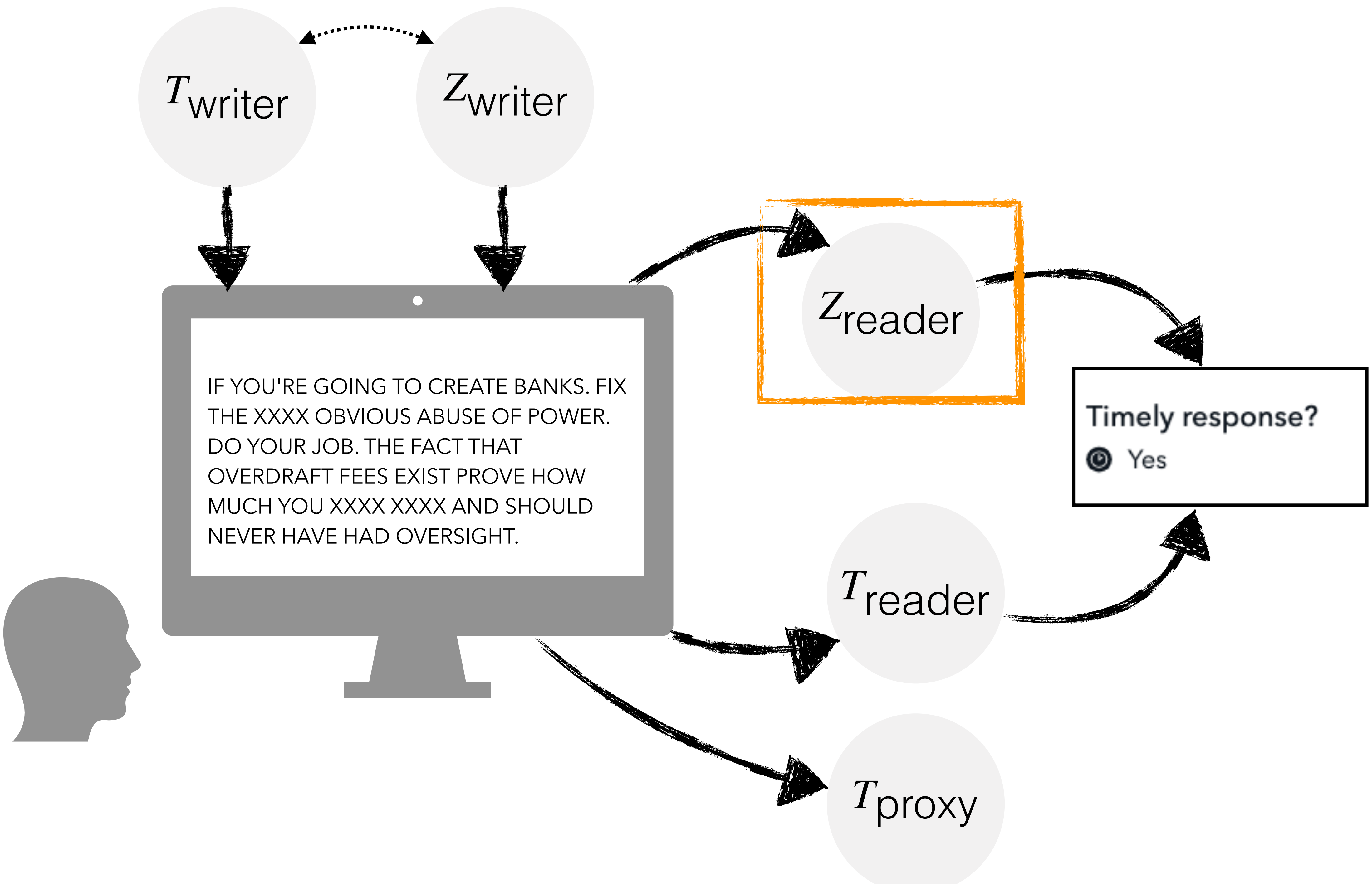
Treatment	Not Depressed	Severity
1	0	1
1	0	1
1	1	0
1	1	0

Treatment	Not Depressed	Severity
0	1	0
0	0	1
0	0	0
0	1	0

Informally: Group by confounders. In each group, calculate expected differences in outcome. Average across groups.

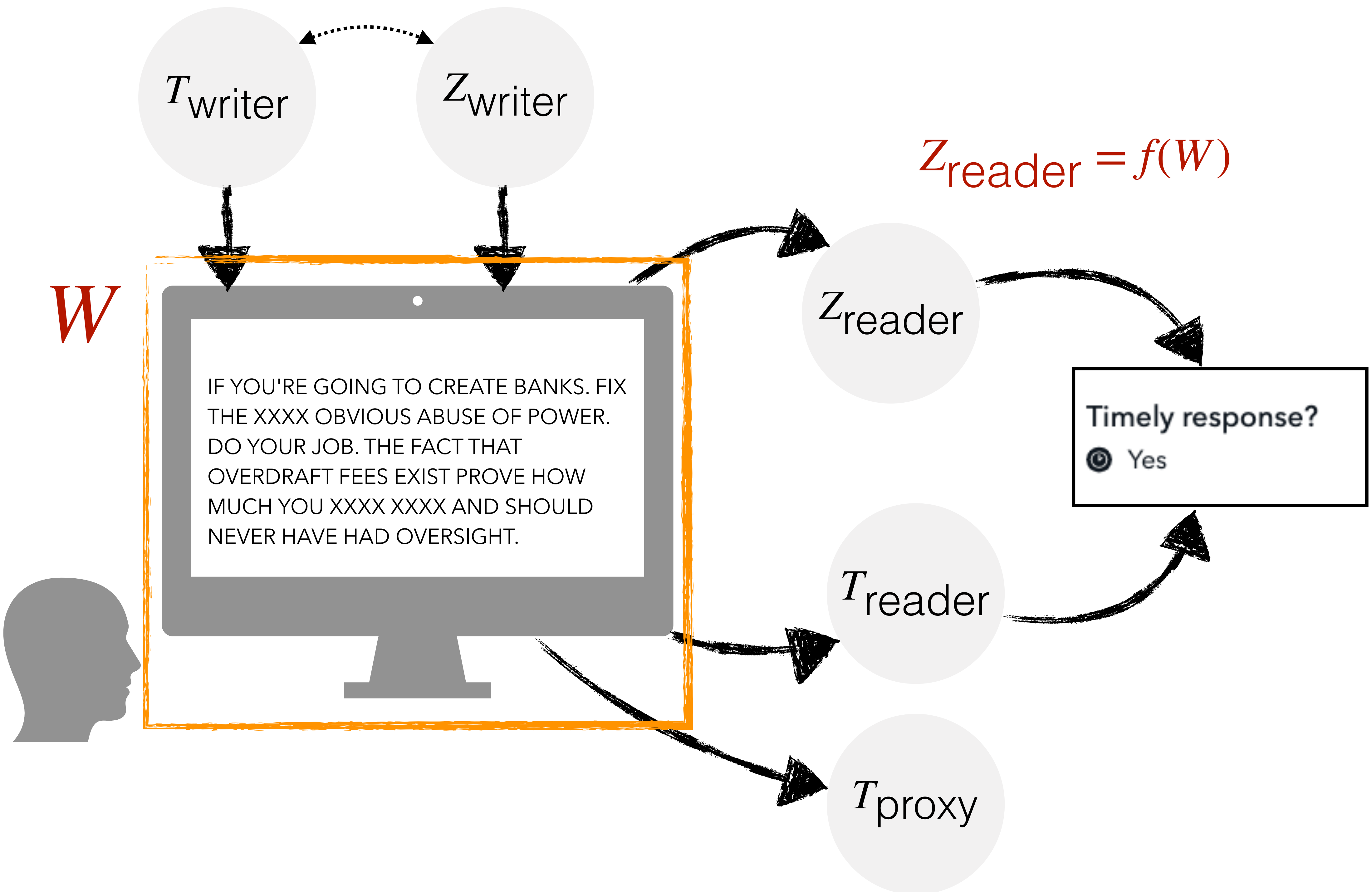
Causal assumptions

Other perceived properties are confounding, e.g., complaints about banks are perceived as impolite and receive slower responses.



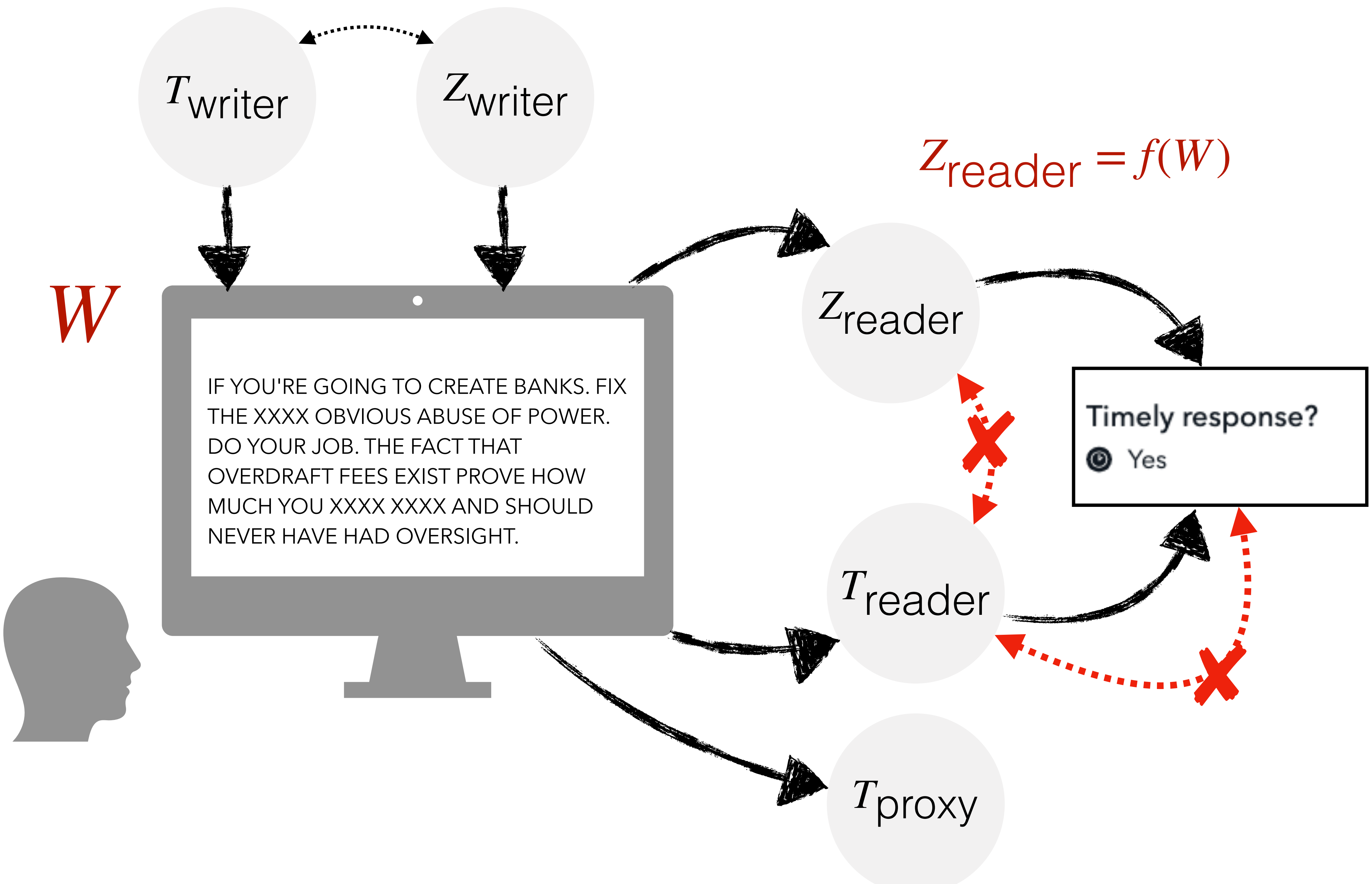
Causal assumptions

Text suffices to adjust for confounding.

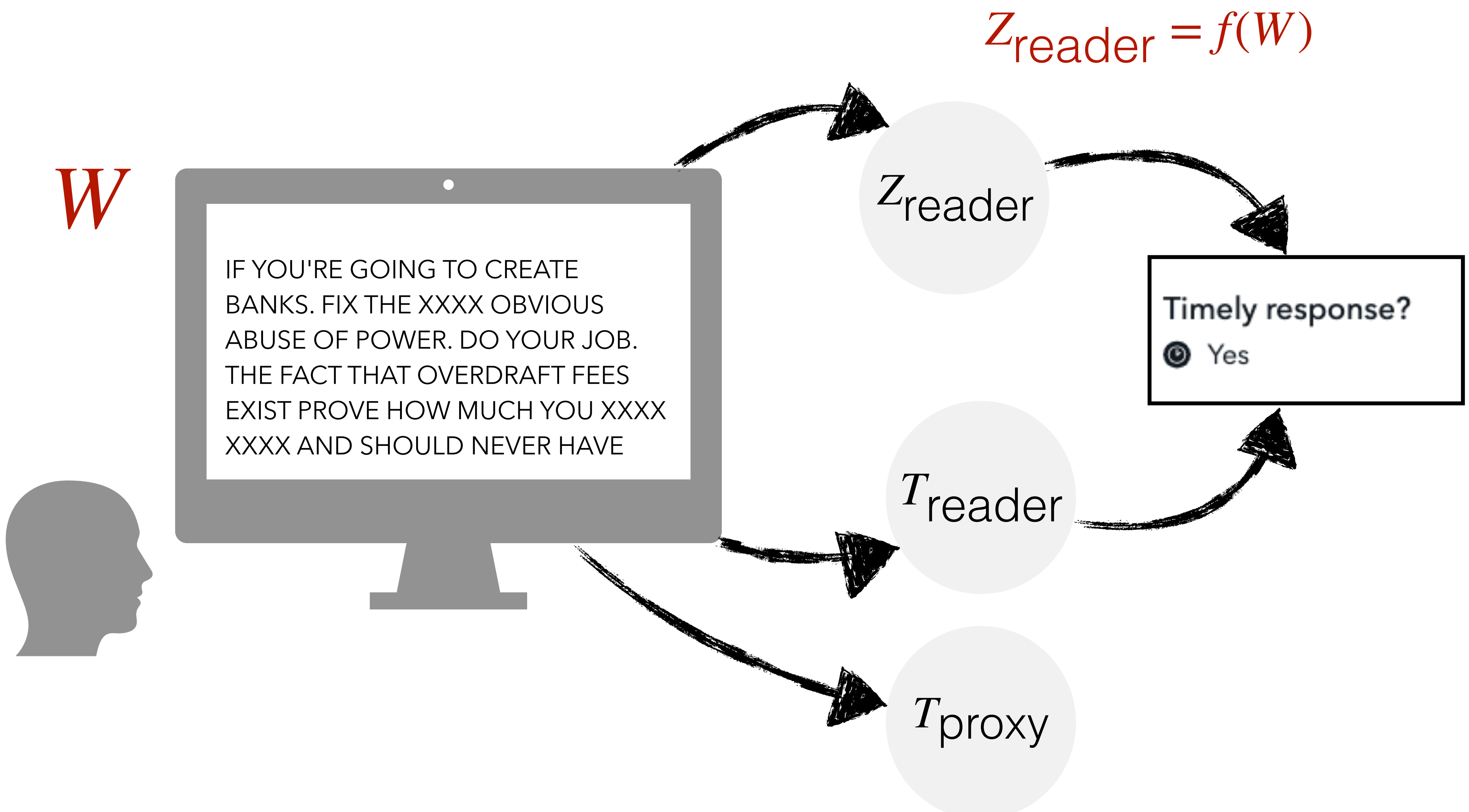


Reflections on causal assumptions

Let's consider what's being ruled out by these assumptions.

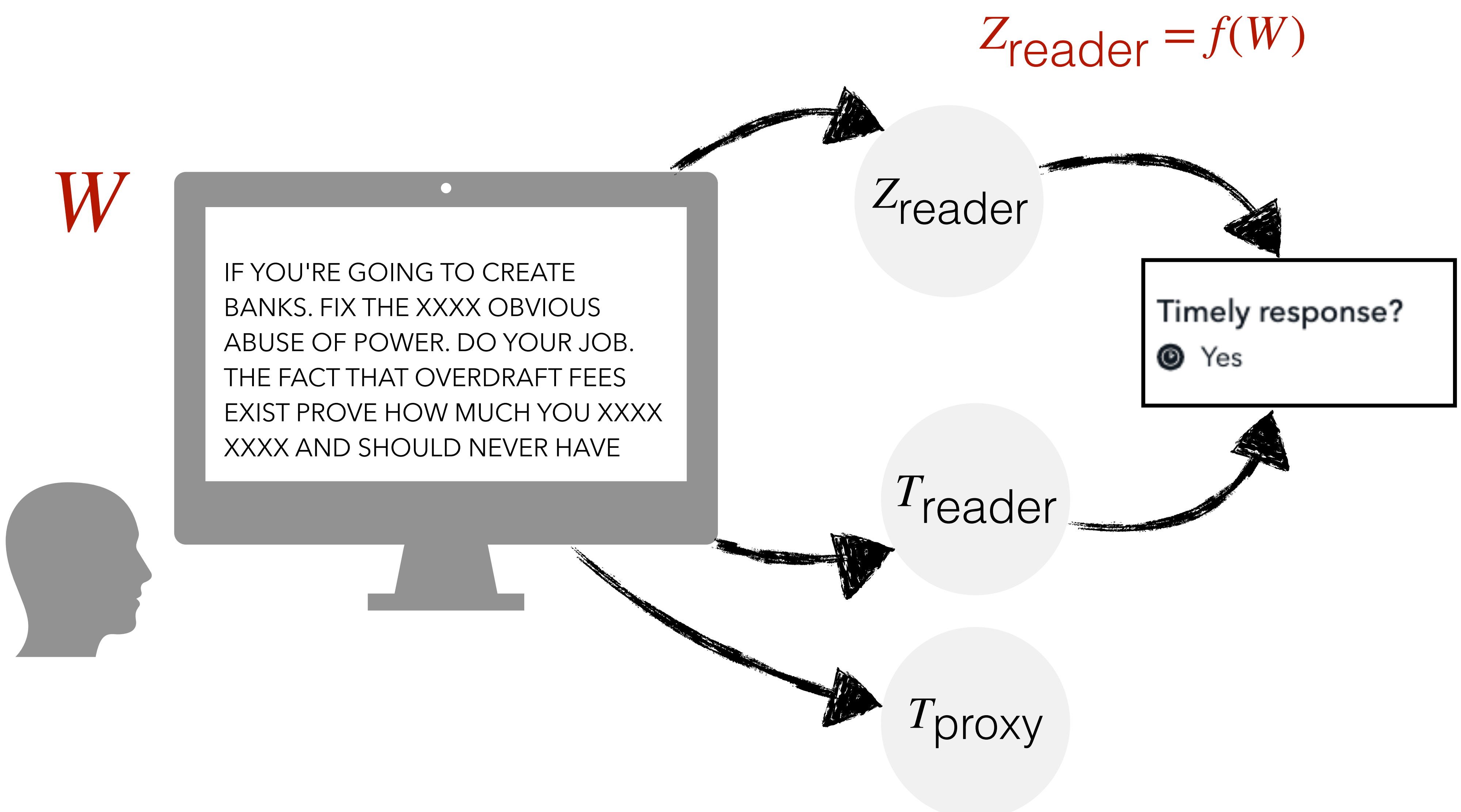


Ground truth causal effect



$$\beta = \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 1)] - \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 0)]$$

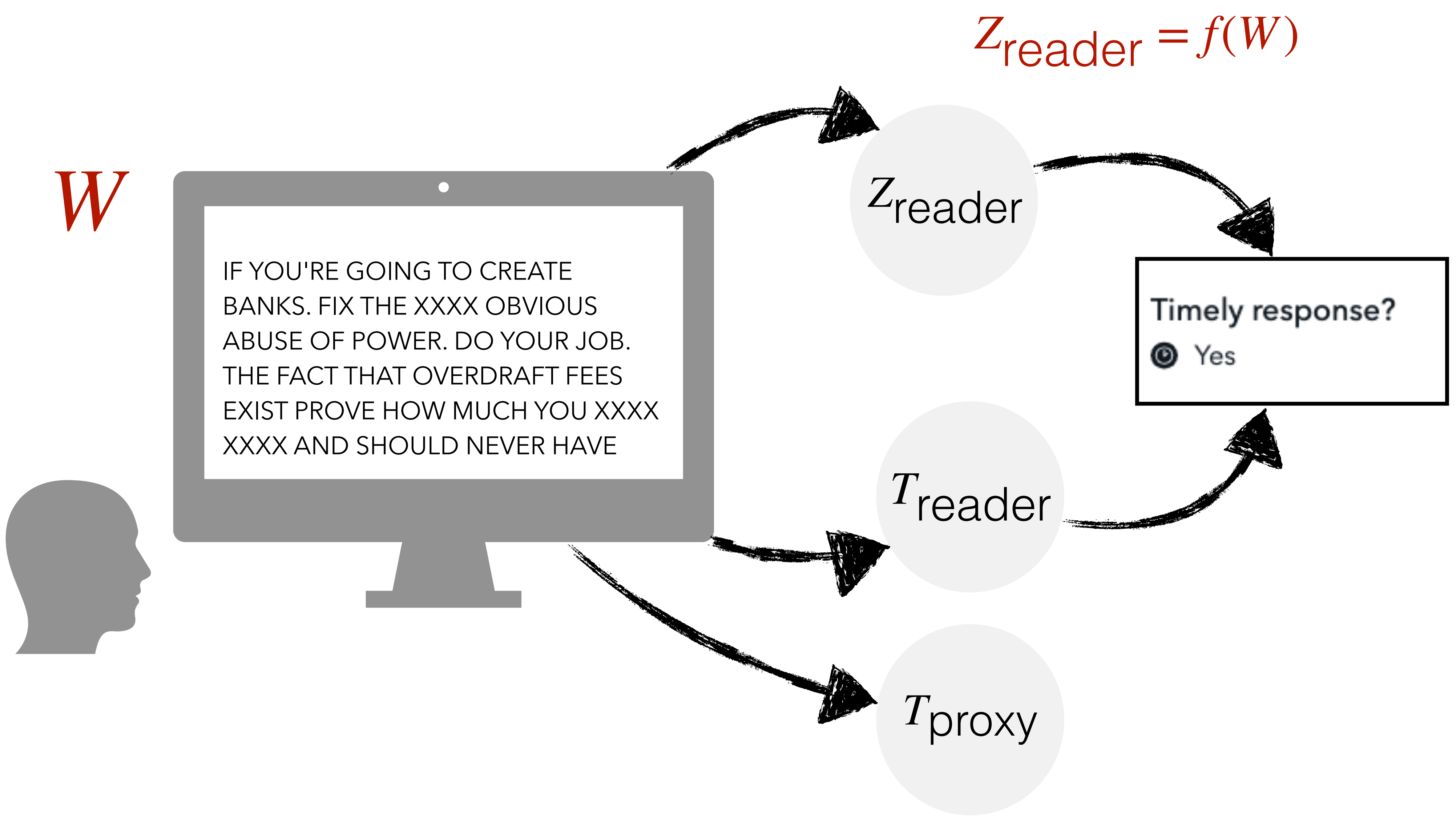
Ground truth causal effect



$$\beta = \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 1)] - \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 0)]$$

$$= \mathbb{E}_W[\mathbb{E}[Y \mid T_{\text{reader}} = 1, W] - \mathbb{E}[Y \mid T_{\text{reader}} = 0, W]]$$

Ground truth causal effect

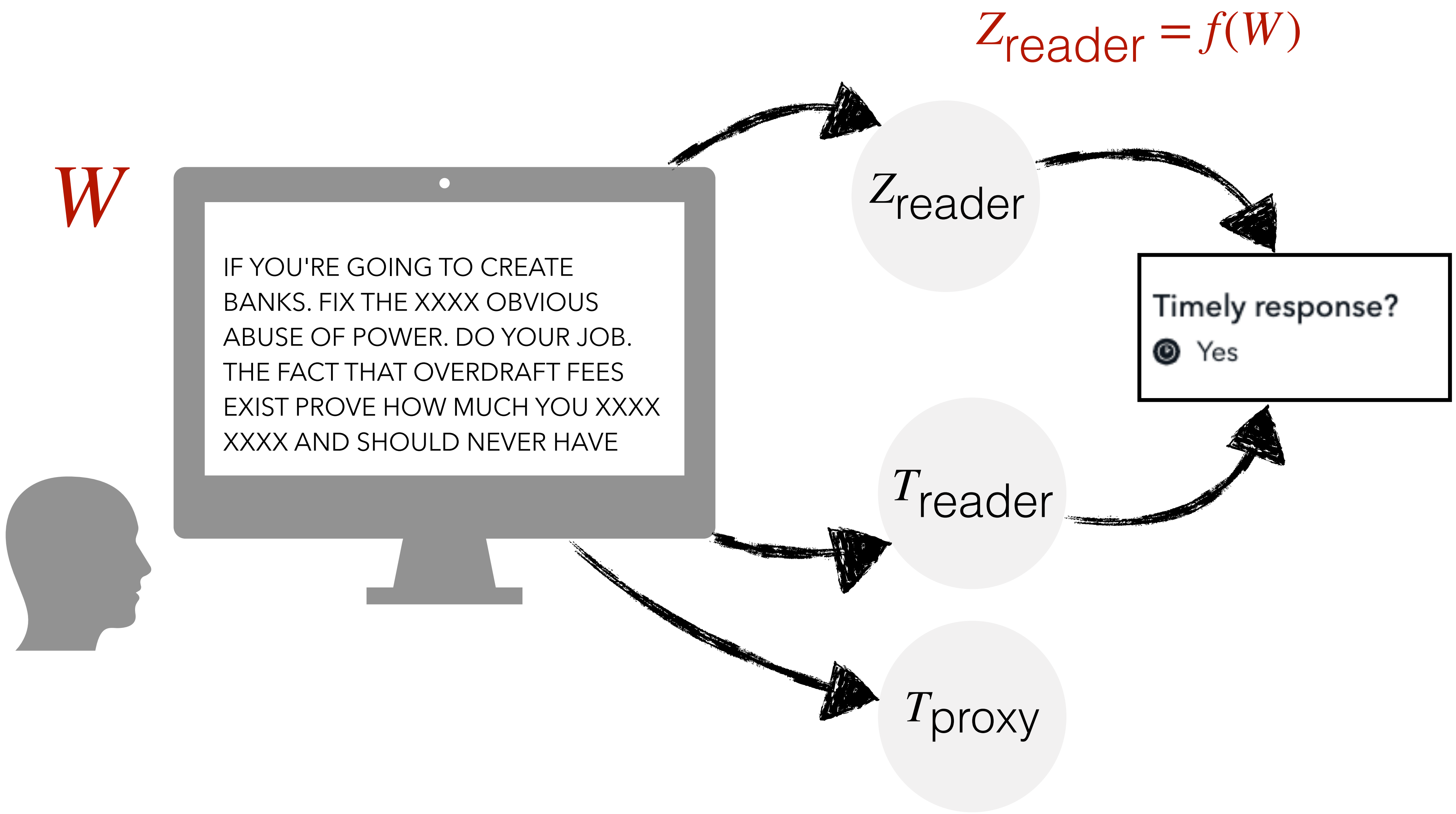


$$\beta = \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 1)] - \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 0)]$$

$$= \mathbb{E}_W[\mathbb{E}[Y \mid T_{\text{reader}} = 1, W] - \mathbb{E}[Y \mid T_{\text{reader}} = 0, W]]$$

$$= \mathbb{E}_W[\mathbb{E}[Y \mid T_{\text{reader}} = 1, Z_{\text{reader}}] - \mathbb{E}[Y \mid T_{\text{reader}} = 0, Z_{\text{reader}}]]$$

Ground truth causal effect



$$\beta = \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 1)] - \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 0)]$$

$$= \mathbb{E}_W[\mathbb{E}[Y \mid T_{\text{reader}} = 1, W] - \mathbb{E}[Y \mid T_{\text{reader}} = 0, W]]$$

$$= \mathbb{E}_W[\mathbb{E}[Y \mid \boxed{T_{\text{reader}} = 1, Z_{\text{reader}}}] - \mathbb{E}[Y \mid \boxed{T_{\text{reader}} = 0, Z_{\text{reader}}}]]$$

Example of a “bad” confounder

An official website of the United States government

cfpb Consumer Financial Protection Bureau

Product

Checking or savings account

Sub-product: Checking account

Issue

Problem caused by your funds being low

Sub-issue: Overdrafts and overdraft fees

Consumer consent to publish narrative

Consent provided

Timely response?

Yes

IF YOU'RE GOING TO CREATE BANKS. FIX THE XXXX OBVIOUS ABUSE OF POWER. DO YOUR JOB. THE FACT THAT OVERDRAFT FEES EXIST PROVE HOW MUCH YOU XXXX XXXX AND SHOULD NEVER HAVE HAD OVERSIGHT. I WOULD CREATE A NEW BANK TO FIX THIS PROBLEM, IF YOU LET ME. YOU ONL.Y LET THE XXXX EVIL PEOPLE START BANKS.

Z

Example of a “nice” confounder

An official website of the United States government

cfpb Consumer Financial Protection Bureau

Product

Checking or savings account

Sub-product: Checking account

Issue

Problem caused by your funds being low

Sub-issue: Overdrafts and overdraft fees

Consumer consent to publish narrative

Consent provided

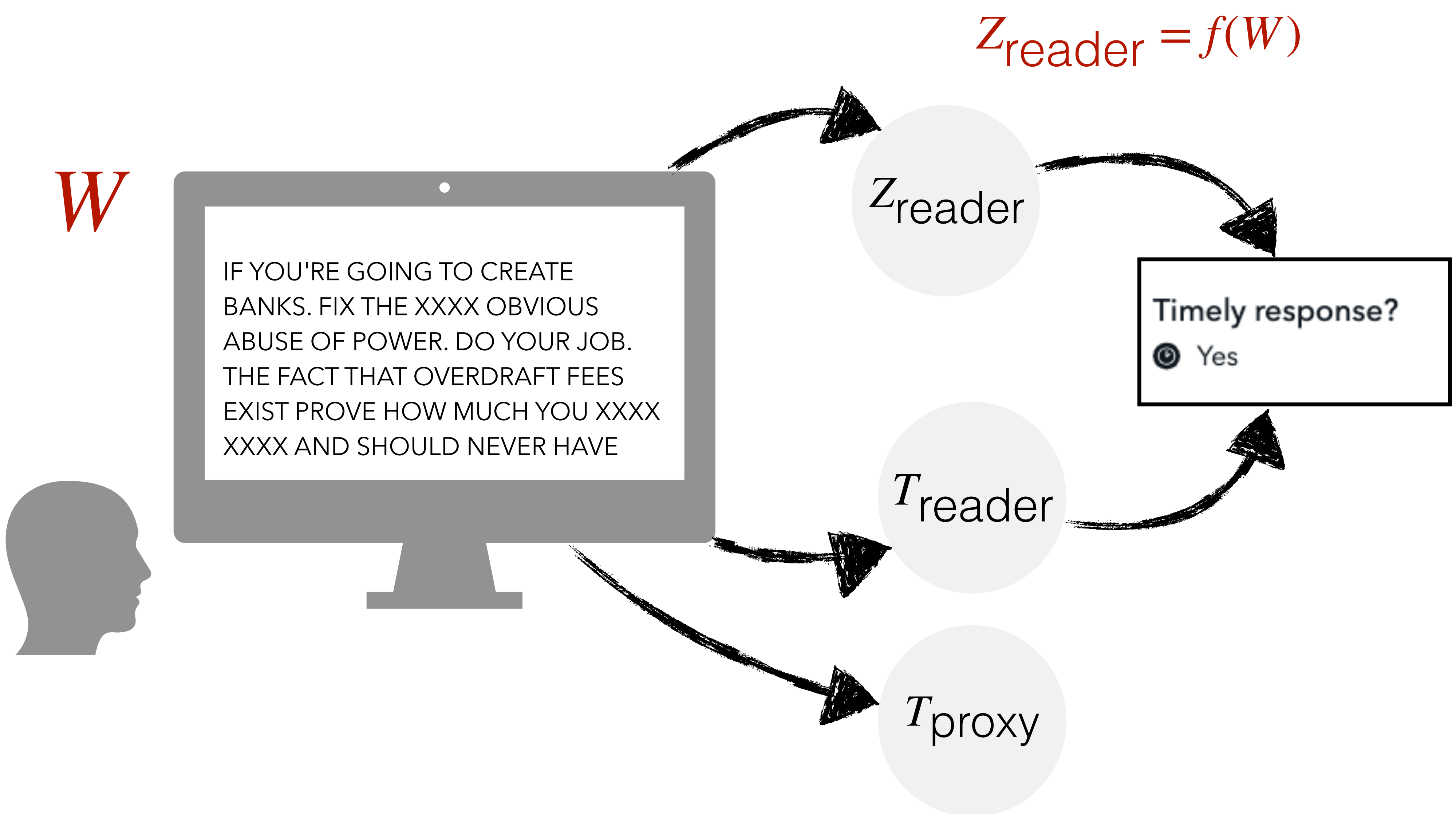
Timely response?

Yes

IF YOU'RE GOING TO CREATE BANKS. FIX THE XXXX OBVIOUS ABUSE OF POWER. DO YOUR JOB. THE FACT THAT OVERDRAFT FEES EXIST PROVE HOW MUCH YOU XXXX XXXX AND SHOULD NEVER HAVE HAD OVERSIGHT. I WOULD CREATE A NEW BANK TO FIX THIS PROBLEM, IF YOU LET ME. YOU ONL.Y LET THE XXXX EVIL PEOPLE START BANKS.

Z

Ground truth causal effect



$$\beta = \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 1)] - \mathbb{E}[Y; \text{do}(T_{\text{reader}} = 0)]$$

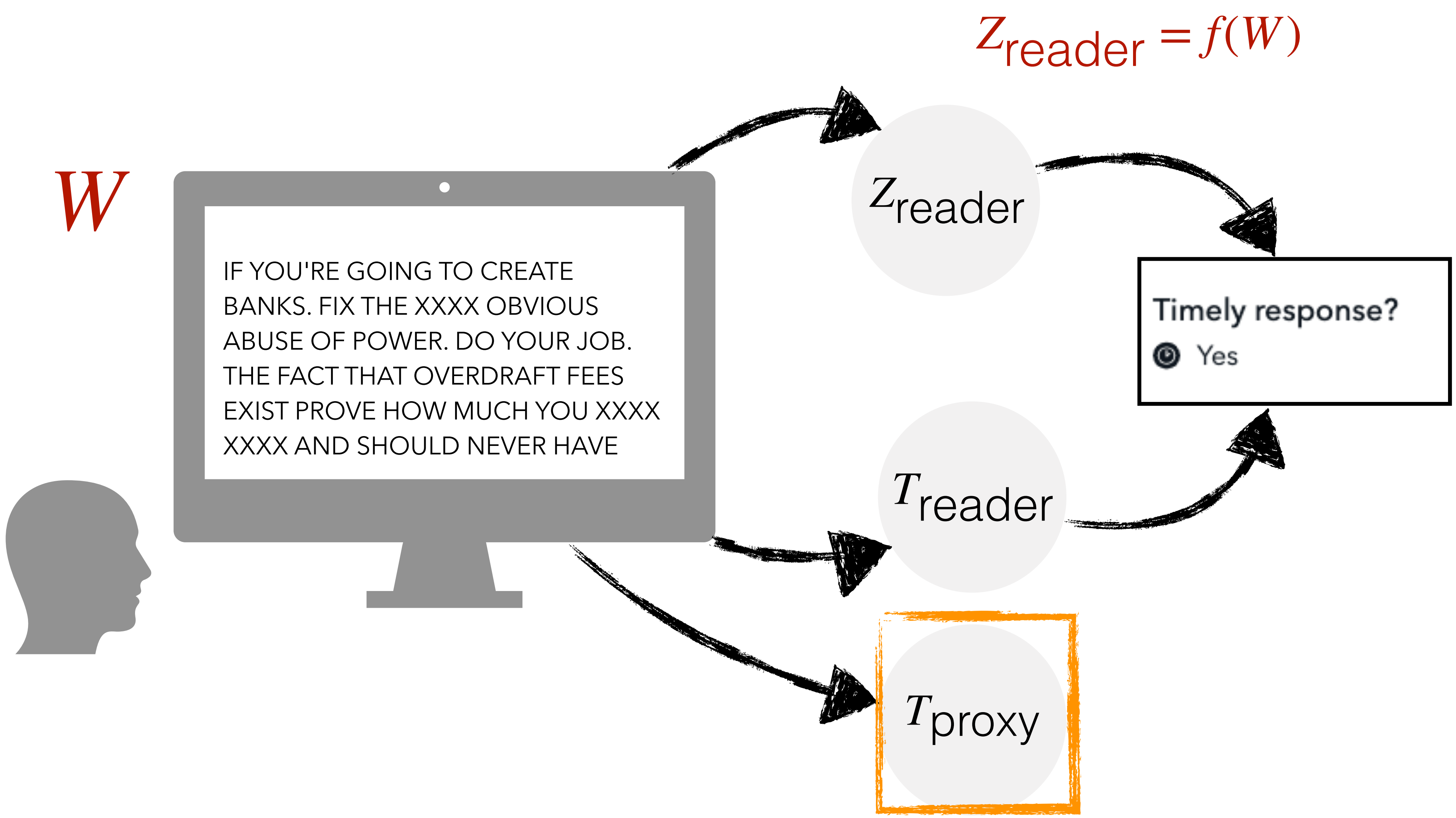
$$= \mathbb{E}_W[\mathbb{E}[Y \mid T_{\text{reader}} = 1, W] - \mathbb{E}[Y \mid T_{\text{reader}} = 0, W]]$$

$$= \mathbb{E}_W[\mathbb{E}[Y \mid \boxed{T_{\text{reader}} = 1, Z_{\text{reader}}}] - \mathbb{E}[Y \mid \boxed{T_{\text{reader}} = 0, Z_{\text{reader}}}]]$$

This talk

- [x] What does “the effect of politeness on response times” mean?
- [ ] Is it possible to recover the effect with a proxy of politeness?
- [ ] If it’s possible, how can we do it?

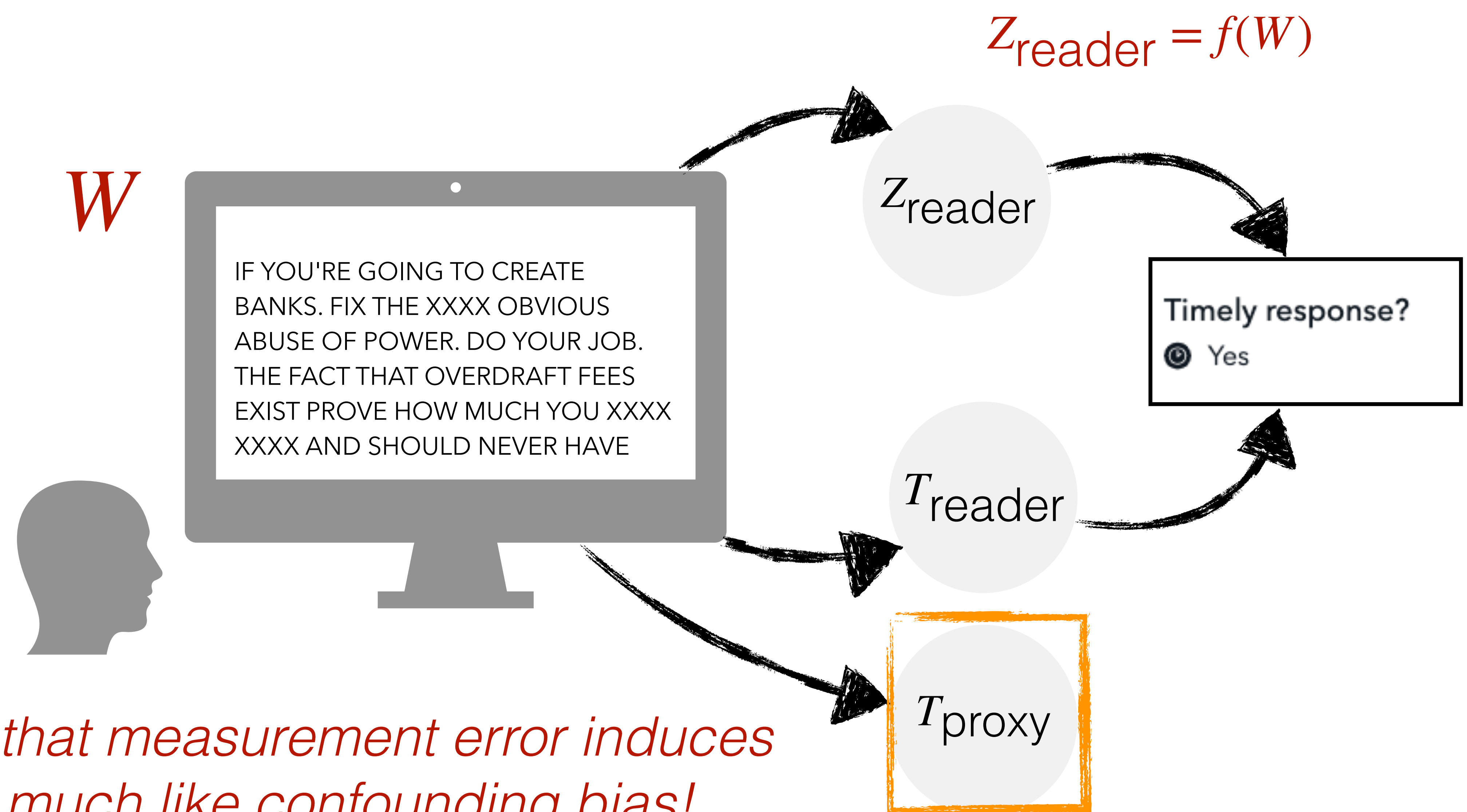
Noisy effect due to proxy



Proposal:

$$\mathbb{E}[Y \mid T_{\text{proxy}} = 1] - \mathbb{E}[Y \mid T_{\text{proxy}} = 0]$$

Noisy effect due to proxy

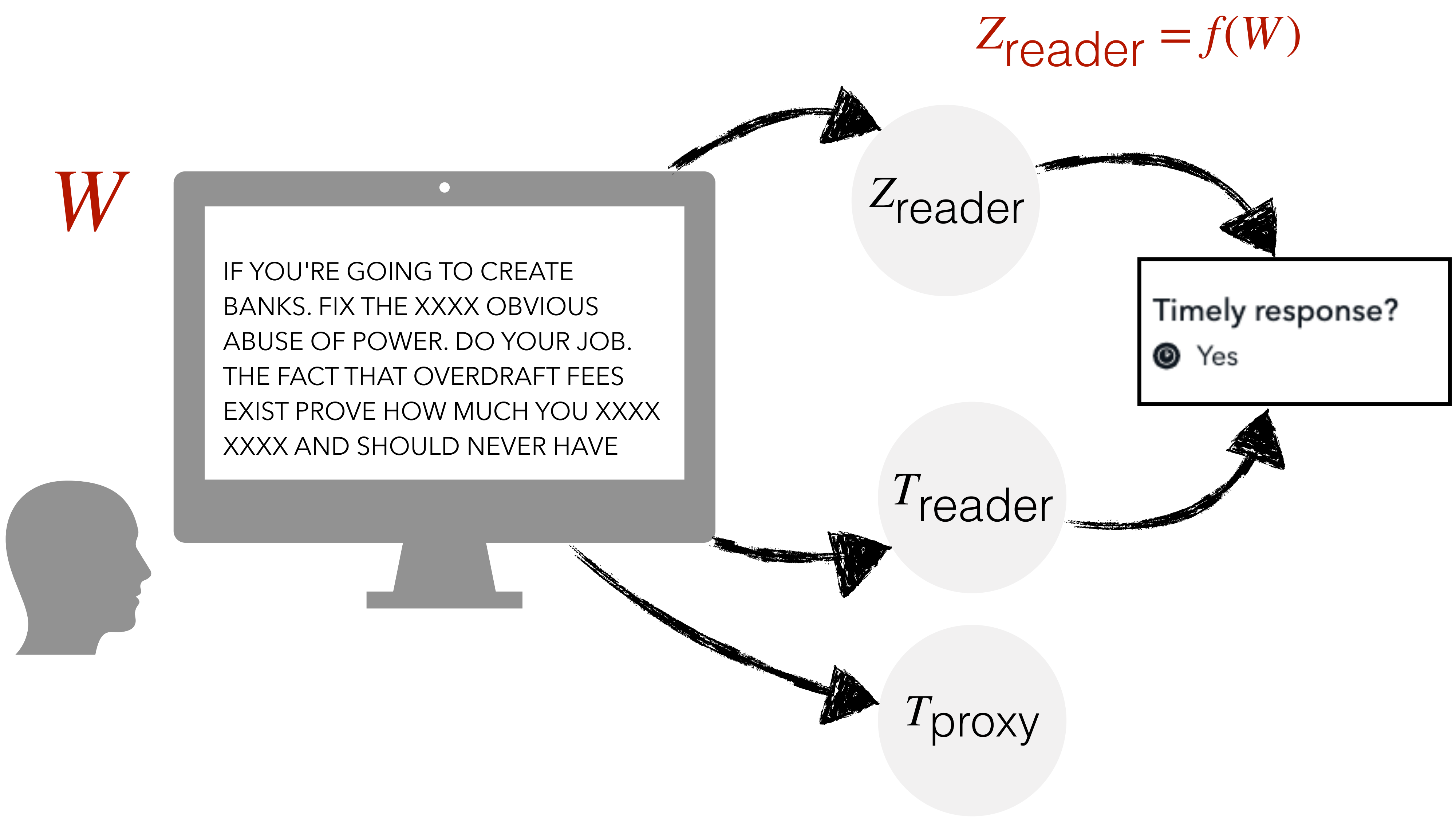


Turns out that measurement error induces a bias much like confounding bias!

Proposal:

$$\mathbb{E}[Y \mid T_{\text{proxy}} = 1] - \mathbb{E}[Y \mid T_{\text{proxy}} = 0]$$

Noisy effect due to proxy



Adjustment with proxy:

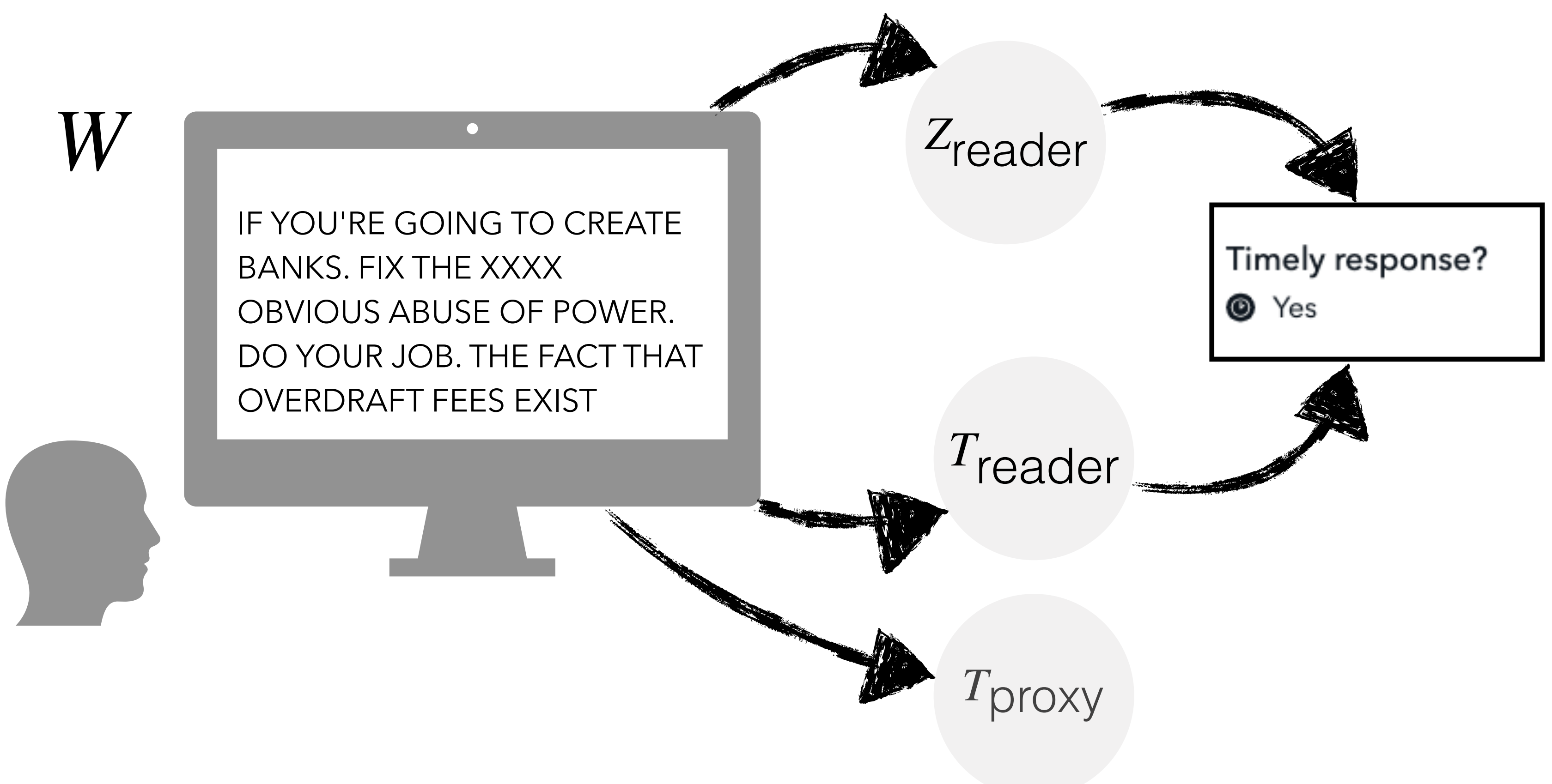
$$\tilde{\beta} = \mathbb{E}_W[\mathbb{E}[Y \mid \boxed{T_{\text{proxy}} = 1}, Z_{\text{reader}}] - \mathbb{E}[Y \mid \boxed{T_{\text{proxy}} = 0}, Z_{\text{reader}}]]$$

Theorem: bias attenuation

$$\tilde{\beta} = \beta - \mathbb{E}_W[\boxed{(\mathbb{E}[Y \mid T = 1, Z] - \mathbb{E}[Y \mid T = 0, Z])}\epsilon_0 + \epsilon_1]$$

$$\epsilon_0 = P(\tilde{T} = 0 \mid \hat{T} = 1, \tilde{Z}); \quad \epsilon_1 = P(\tilde{T} = 1 \mid \hat{T} = 0, \tilde{Z})$$

Terms related to proxy mis-measurement

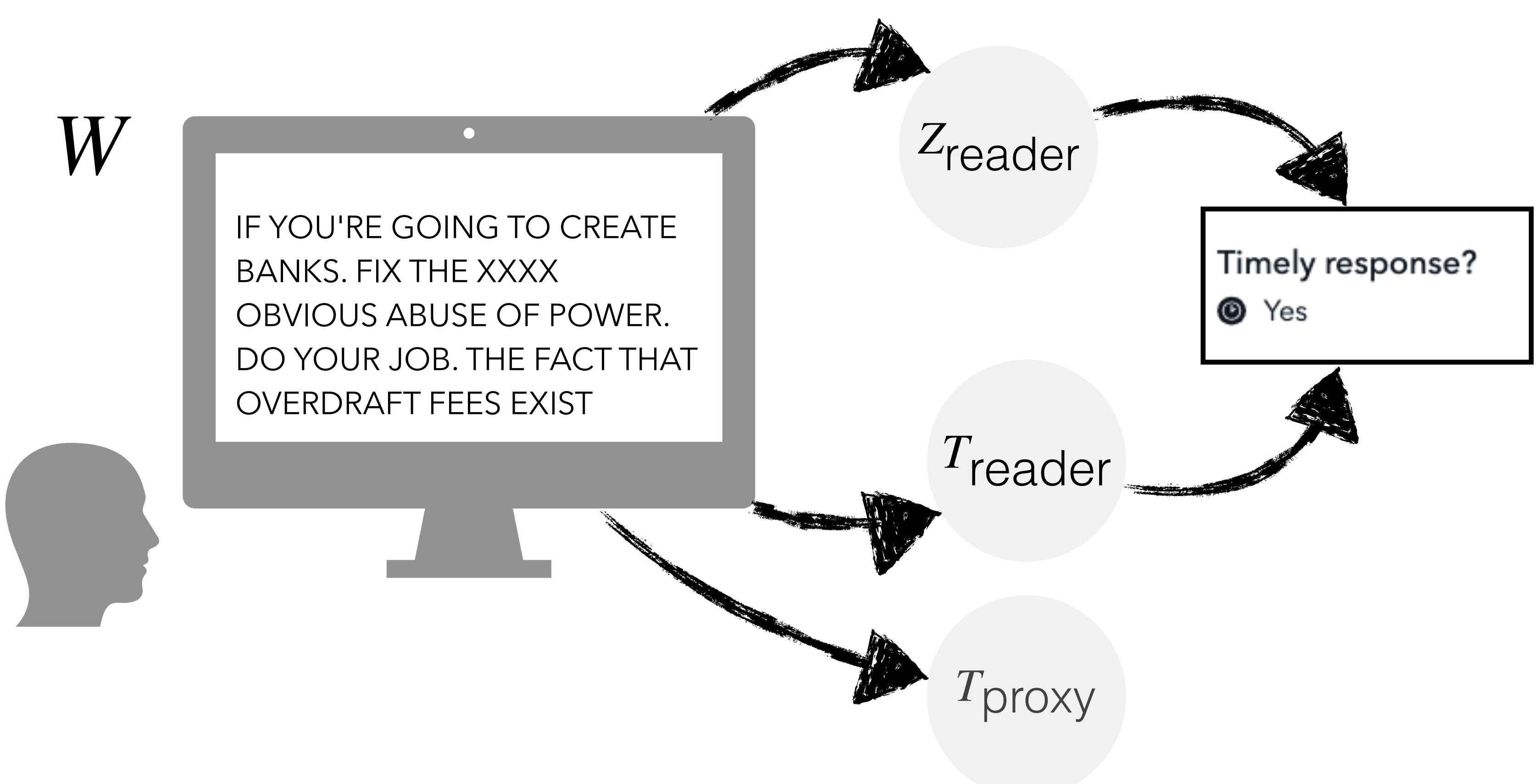


Theorem: bias attenuation

$$\tilde{\beta} = \beta - \mathbb{E}_W[(\mathbb{E}[Y \mid T = 1, Z] - \mathbb{E}[Y \mid T = 0, Z])\boxed{\epsilon_0 + \epsilon_1}]$$

$$\epsilon_0 = P(\tilde{T} = 0 \mid \hat{T} = 1, \tilde{Z}); \quad \epsilon_1 = P(\tilde{T} = 1 \mid \hat{T} = 0, \tilde{Z})$$

Terms related to proxy mis-measurement



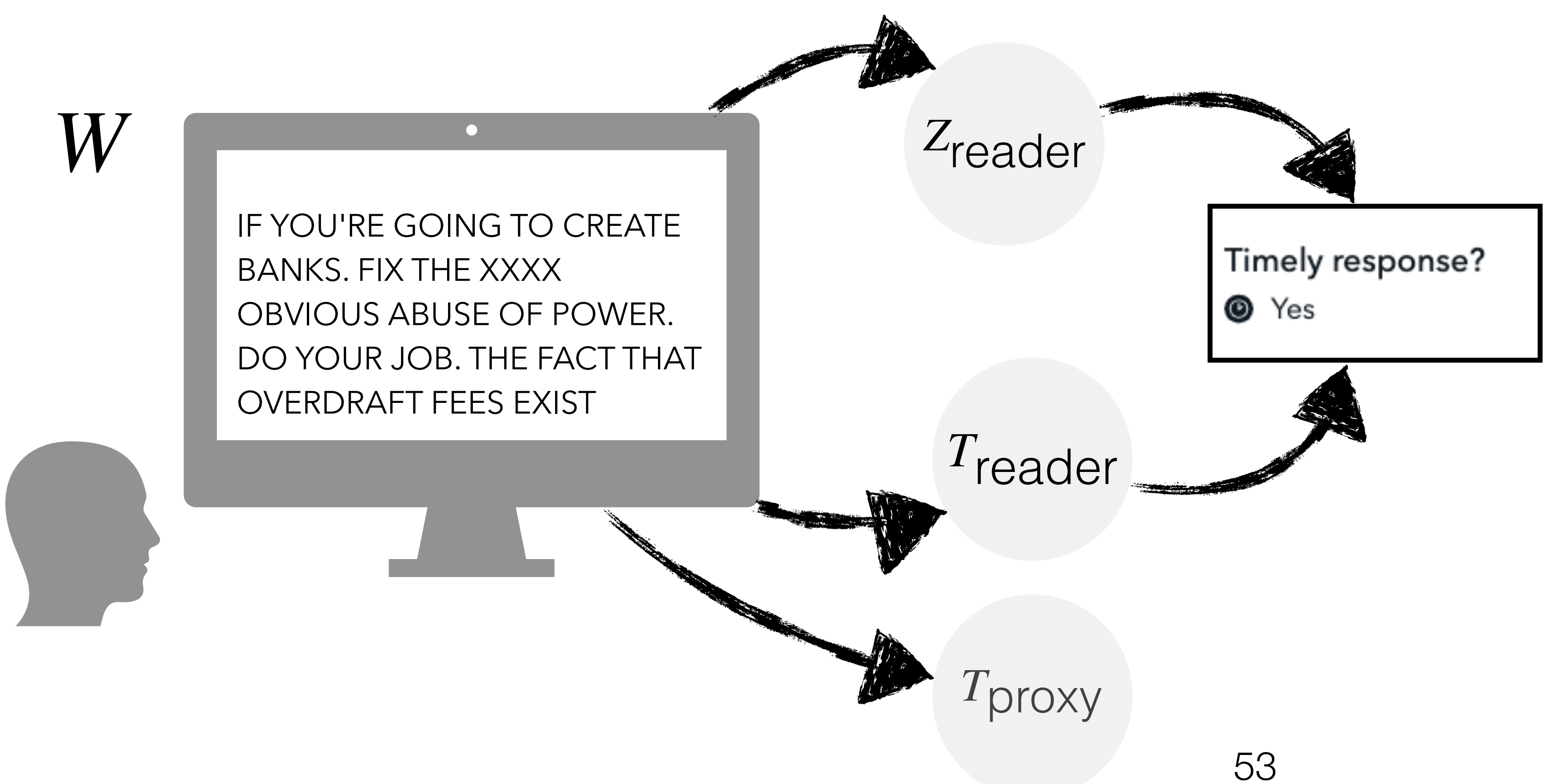
Theorem: bias attenuation

Bias term is smaller in size than true effect.

$$\tilde{\beta} = \beta - \mathbb{E}_W[\boxed{(\mathbb{E}[Y \mid T = 1, Z] - \mathbb{E}[Y \mid T = 0, Z])\epsilon_0 + \epsilon_1}]$$

$$\epsilon_0 = P(\tilde{T} = 0 \mid \hat{T} = 1, \tilde{Z}); \quad \epsilon_1 = P(\tilde{T} = 1 \mid \hat{T} = 0, \tilde{Z})$$

Terms related to proxy mis-measurement



Two lessons:

1. Estimate and adjust for Z.
2. Minimize proxy error.

This talk

- [x] How do we formalize the effect of politeness on response times?
- [ ] Is it possible to recover the effect with a proxy of politeness?
- [ ] If it's possible, how can we do it?

Related work

Fully correct for proxy noise.

Proxy variables in causal inference

- Proxy treatment from text [Wood-Doughty *et al.* (2018]
- Proxies of unobserved confounders [Kuroki and Pearl (2014)]

Causal effects of text

- Randomized experiments [Fong and Grimmer (2016), Grimmer and Fong (2020)]
- Latent variable models [Egami et al. (2018), Sridhar and Getoor (2019)]

Related work

Fully correct for proxy noise.

Proxy variables in causal inference

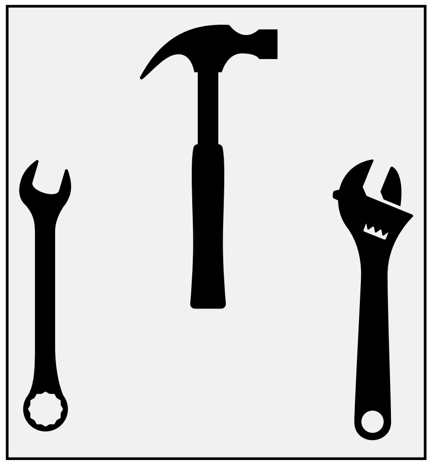
- Proxy treatment from text [Wood-Doughty *et al.* (2018]
- Proxies of unobserved confounders [Kuroki and Pearl (2014)]

Causal effects of text

- Randomized experiments [Fong and Grimmer (2016), Grimmer and Fong (2020)]
- Latent variable models [Egami et al. (2018), Sridhar and Getoor (2019)]

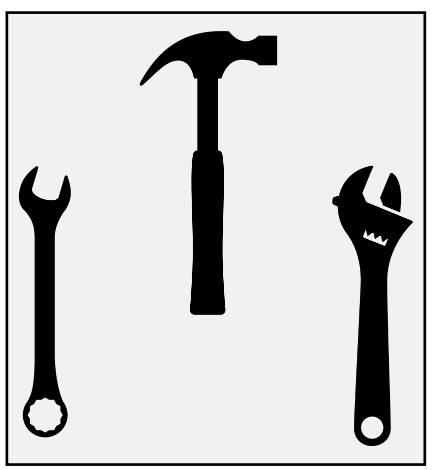
Requires randomized experiments and strong assumptions.

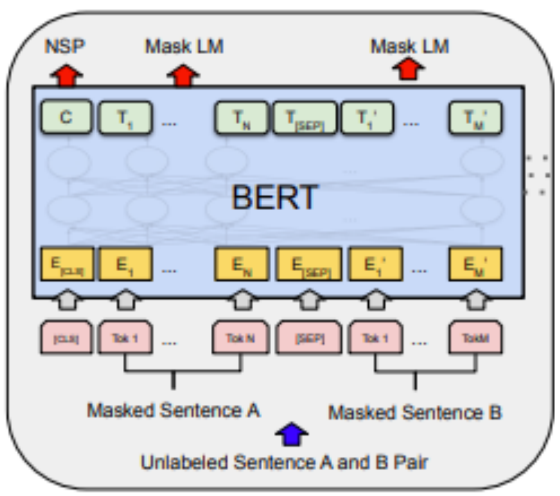
Technical insight



Word	Anger
outraged	0.964
brutality	0.959
satanic	0.828
hate	0.828
violence	0.742

Classifiers and lexicons: Substitute perceived politeness with a prediction, i.e., proxy.





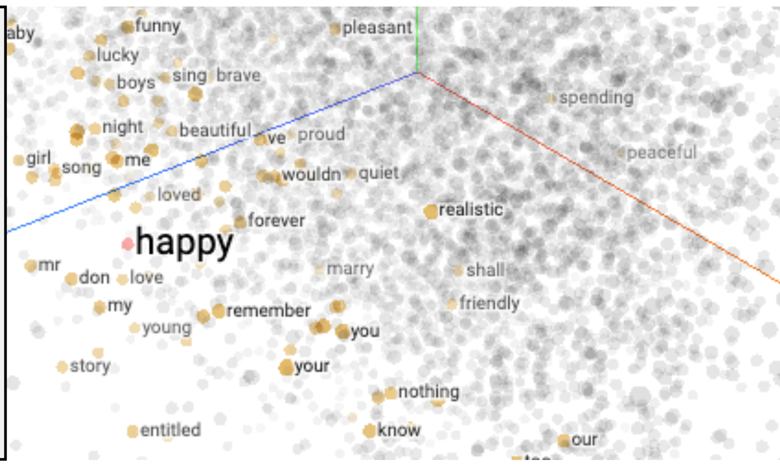
Feature extractors: Pre-trained models such as BERT can be fine-tuned to extract task-relevant features.

Devlin *et al.* (2018), Sanh *et al.* (2019)

We'll appeal to flexible NLP methods.

Technical insight

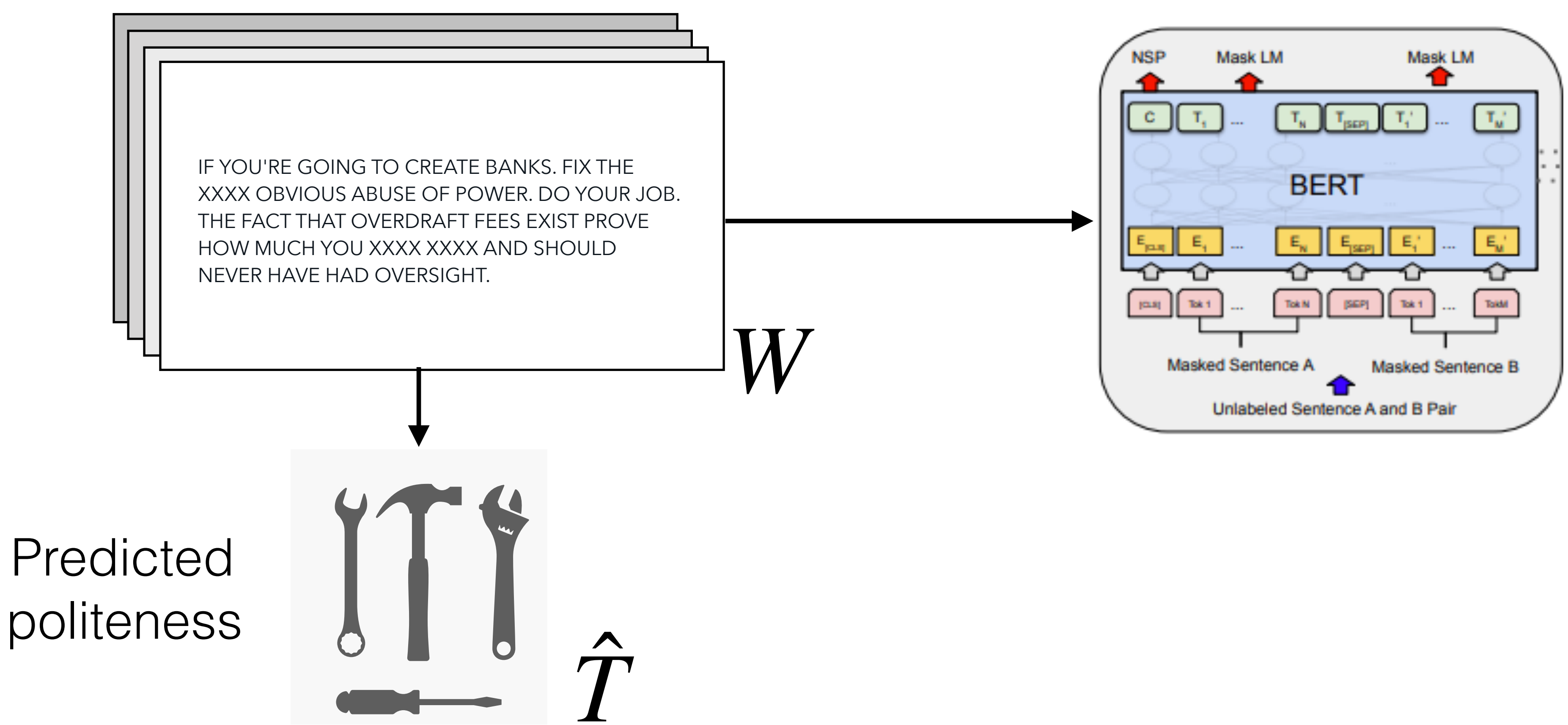
Goal: Represent text for causal inference.



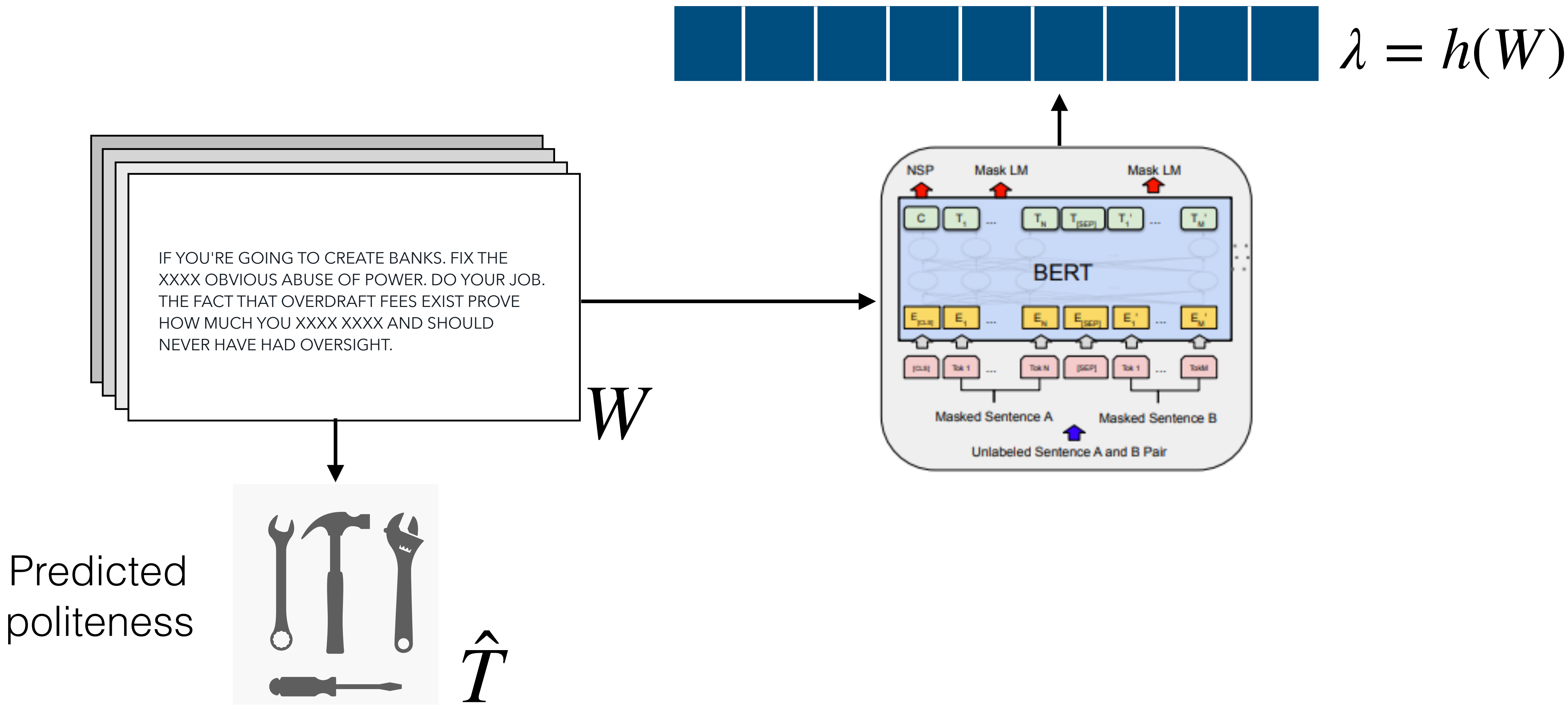
$$\beta = \mathbb{E}_Z\left[\boxed{\mathbb{E}[Y \mid T = 1, Z]} - \boxed{\mathbb{E}[Y \mid T = 0, Z]}\right]$$

Supervised dimensionality reduction:
Learn embeddings that predict well.

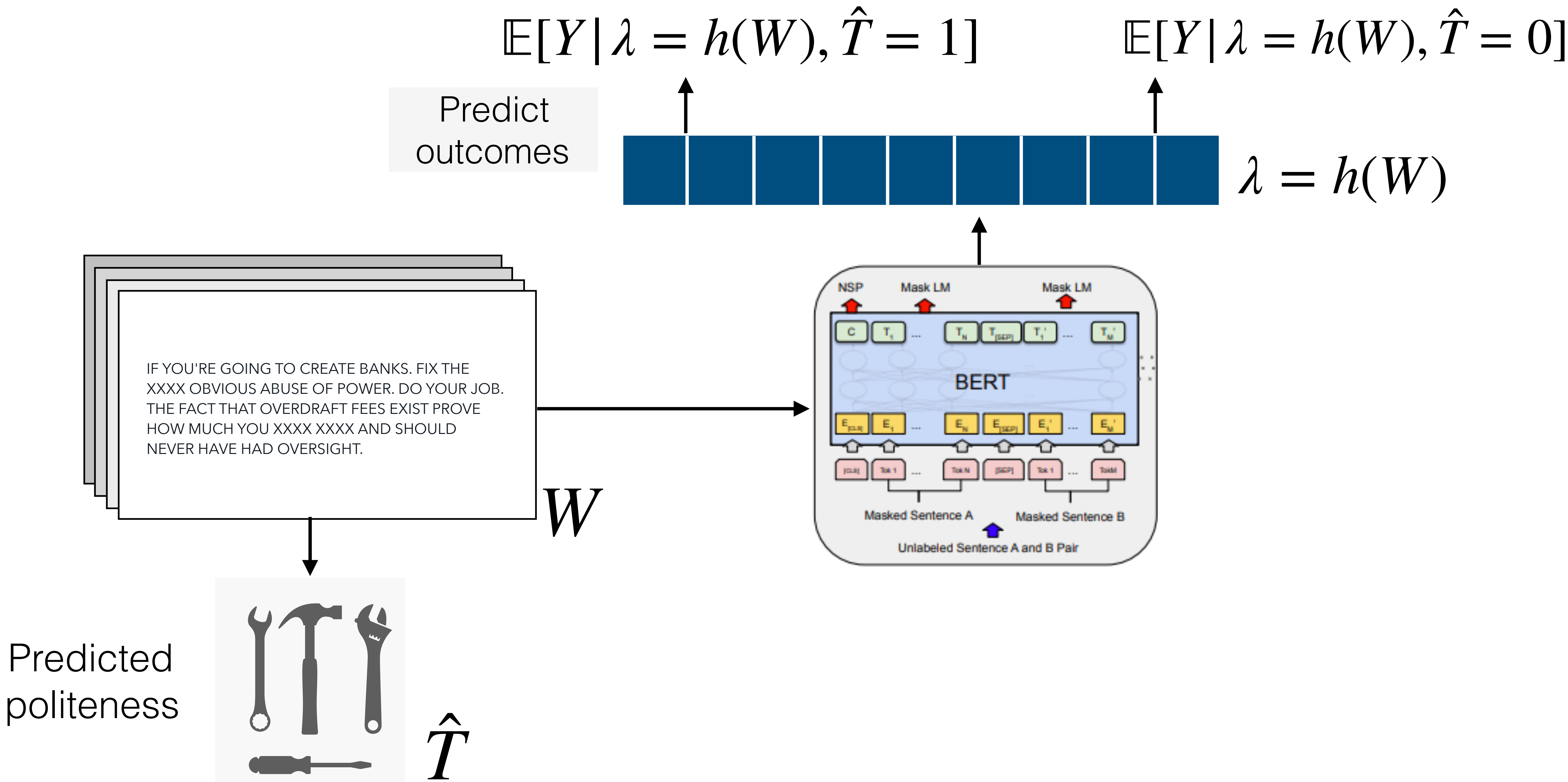
The TextCause algorithm



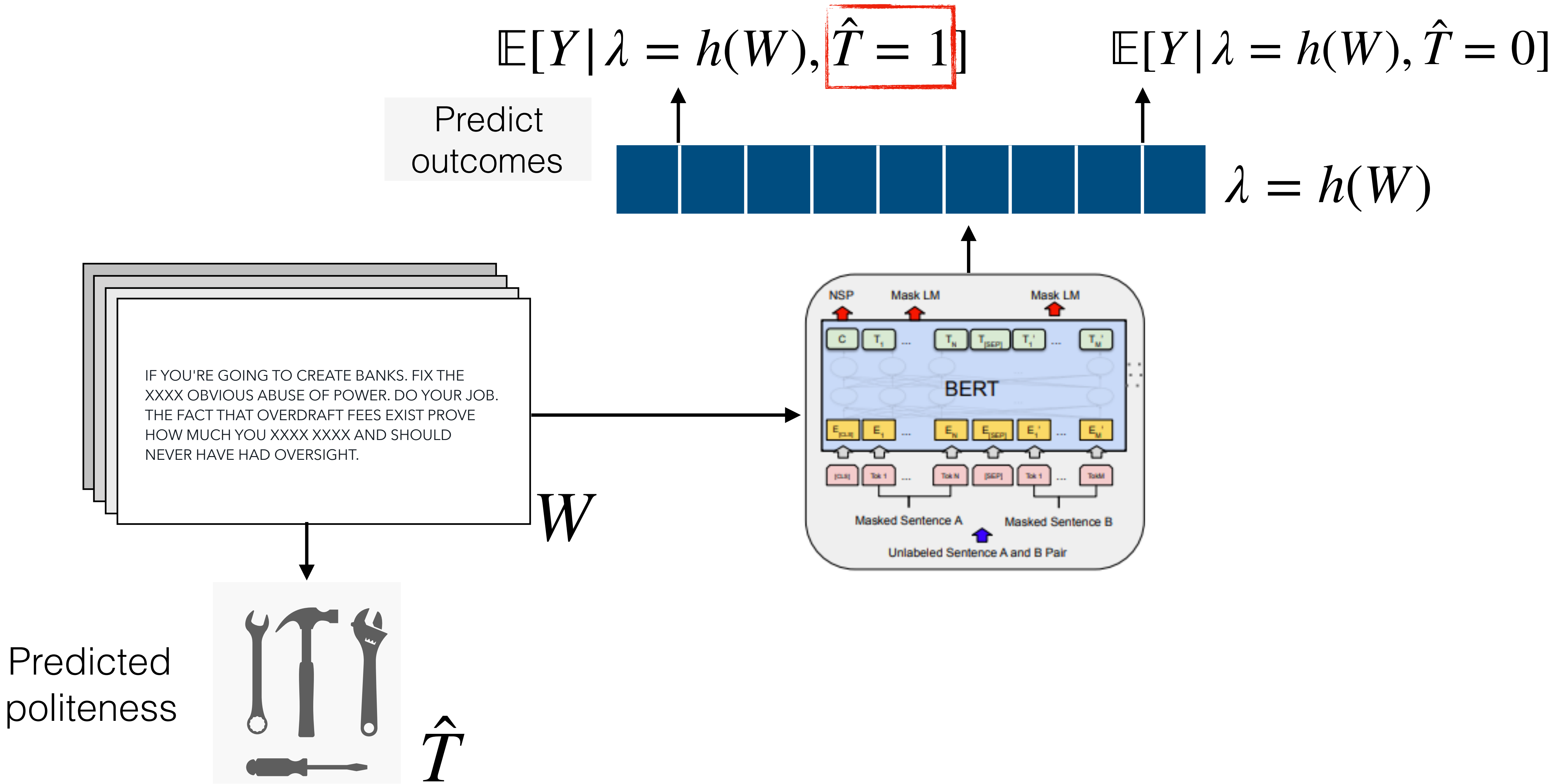
The TextCause algorithm



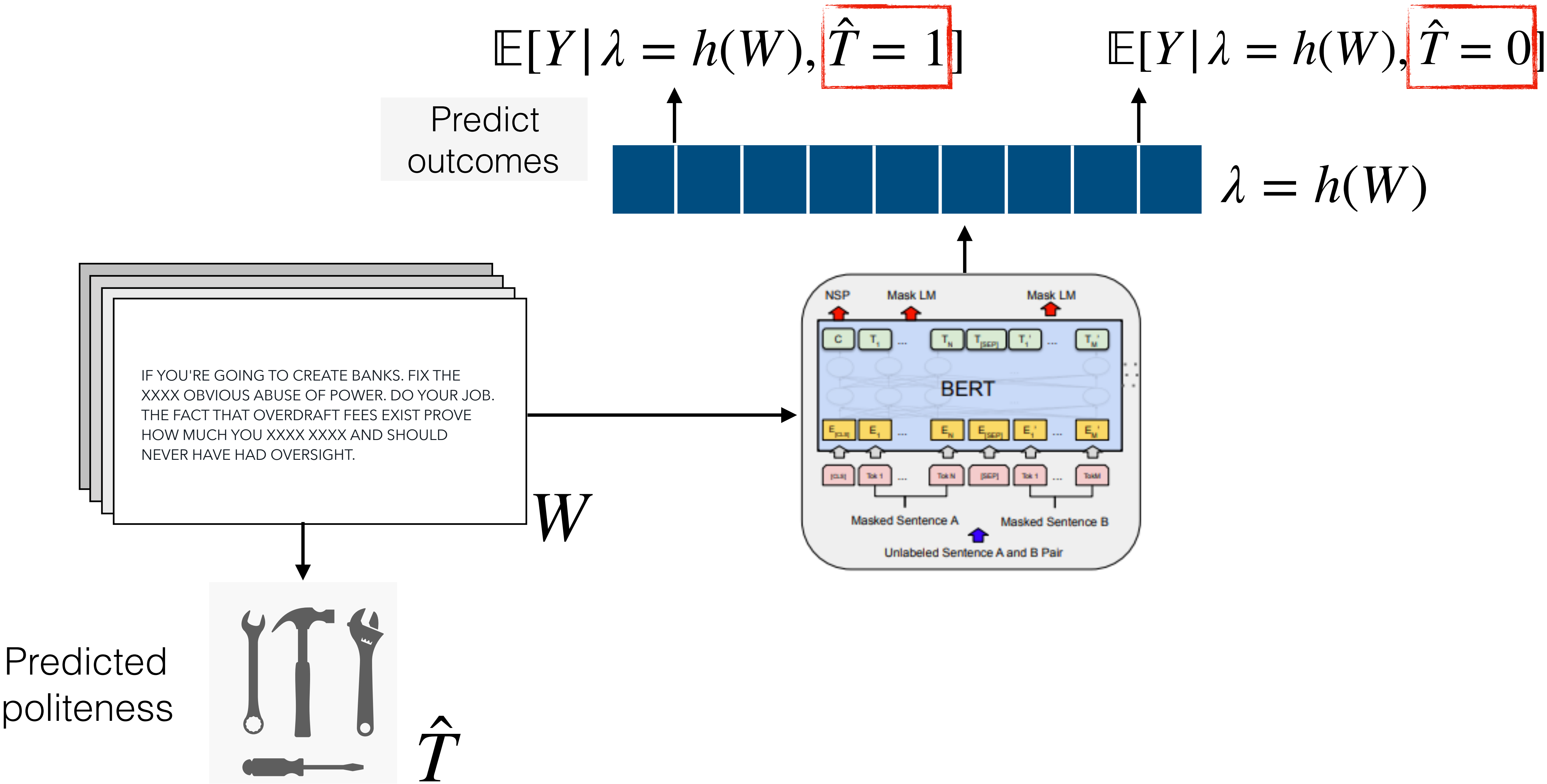
The TextCause algorithm



The TextCause algorithm



The TextCause algorithm



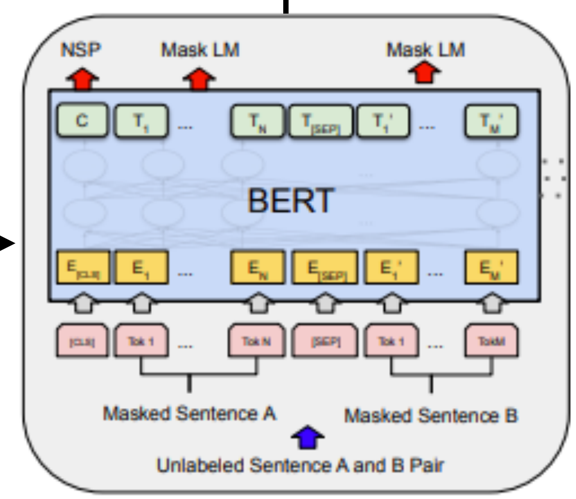
The TextCause algorithm

$\mathbb{E}[Y \mid \lambda = h(W), \boxed{\hat{T} = 1}]$

$\mathbb{E}[Y \mid \lambda = h(W), \boxed{\hat{T} = 0}]$

Predict outcomes

$\lambda = h(W)$



IF YOU'RE GOING TO CREATE BANKS. FIX THE XXXX OBVIOUS ABUSE OF POWER. DO YOUR JOB. THE FACT THAT OVERDRAFT FEES EXIST PROVE HOW MUCH YOU XXXX XXXX AND SHOULD NEVER HAVE HAD OVERSIGHT.

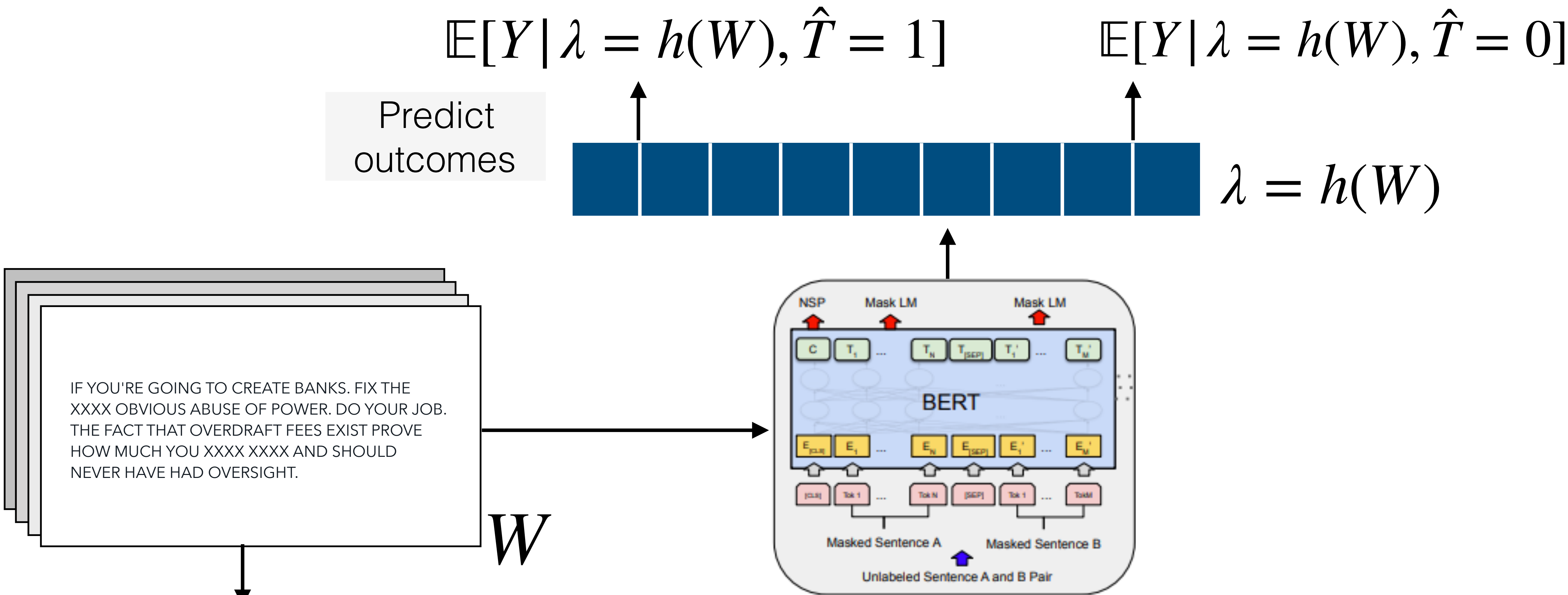
W

Predicted politeness

$\hat{T}$

Encourage embeddings to learn properties other than politeness.

The TextCause algorithm



Predicted politeness

$\hat{T}$

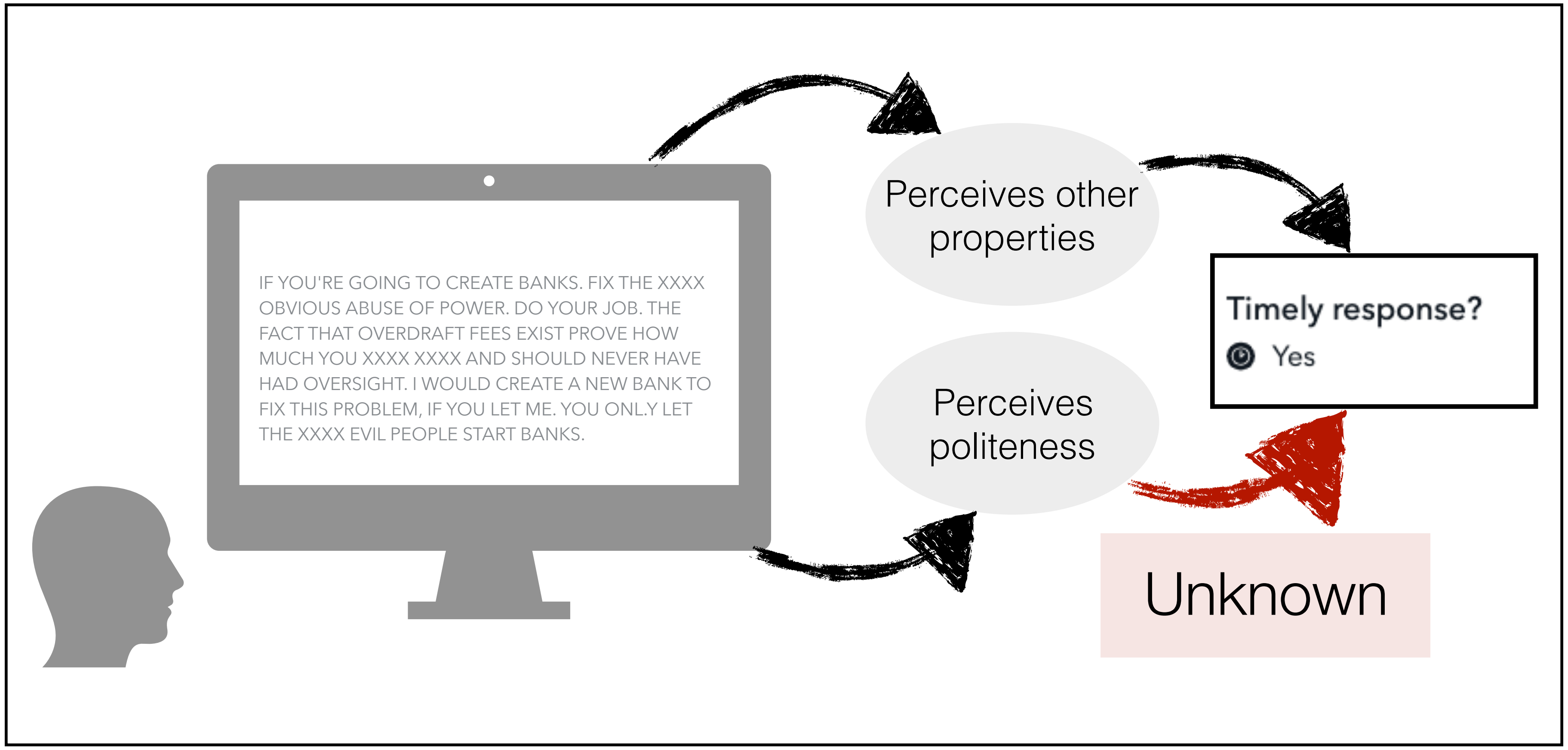
Use predictions to estimate effects.

This talk

- ☑ What does “the effect of politeness on response times” mean?
- ☑ Is it possible to recover the effect with a proxy of politeness?
- ☑ If it’s possible, how can we do it?

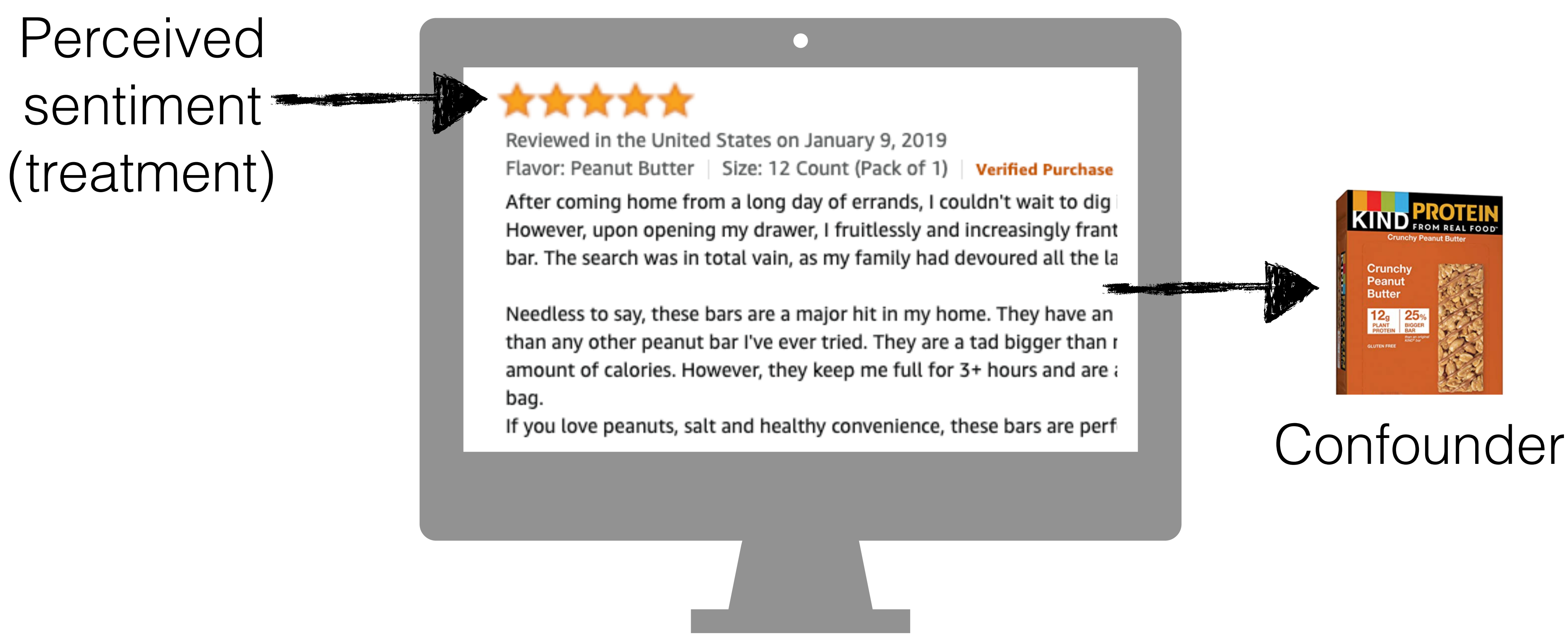
How do we evaluate?

Problem: no ground truth causal effects of documents against which to evaluate estimation error.



Example with Amazon reviews

Example: Amazon reviews, product rating and product categories.



Example with Amazon reviews

Simulate click as log-linear function of rating and product category.

Experimental evaluation

Increase confounding due to product category and evaluate methods.

Demo time!

arxiv.org/abs/2010.12919 **github.com/rpryzant/causal-text**

Recap

1. Abundant text data presents an opportunity to extract more information about people.

2. Defining causal questions can be challenging but articulating causal structure helps.

3. In particular, exploit domain knowledge to derive adjustment or other results. Here, we exploited writer/reader asymmetry.

4. Lots of opportunities for new research designs (exploiting random variation), new estimation methods, and new evaluations.