UNIVERSITAT DE BARCELONA

MSc in Fundamental Principles of Data Science

5

# Ethical Data Science

## Bias and Discrimination II: Causality

**Jordi Vitrià**

# Fairness from a causality perspective

Observational Data

Source: https://fairmlbook.org/

UC Berkeley admissions data from 1973.

| | Men | | Women | |
|---|---|---|---|---|
| Department | Applied | Admitted (%) | Applied | Admitted (%) |
| A | 825 | 62 | 108 | 82 |
| B | 520 | 60 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |

Is this a fair admission process?

# Fairness from a causality perspective

Observational Data

Source: https://fairmlbook.org/

UC Berkeley admissions data from 1973.

| | Men | | Women | |
|---|---|---|---|---|
| Department | Applied | Admitted (%) | Applied | Admitted (%) |
| A | 825 | 62 | ~~108~~ | 82 |
| B | 520 | 60 | ~~25~~ | 68 |
| C | 325 | 37 | ~~593~~ | 34 |
| D | 417 | 33 | ~~375~~ | 35 |
| E | 191 | 28 | ~~393~~ | 24 |
| F | 373 | 6 | ~~341~~ | 7 |

**First Observation:**
four of the six largest departments show a higher acceptance ratio among women, while two show a higher acceptance rate for men.

# Fairness from a causality perspective

Observational Data

Source: https://fairmlbook.org/

UC Berkeley admissions data from 1973.

| | Men | | Women | |
|---|---|---|---|---|
| Department | Applied | Admitted (%) | Applied | Admitted (%) |
| A | 825 | 62 | 108 | 82 |
| B | 520 | 60 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |
| | 44% | 1157/2651 | 30% | 556/1835 |

**Second Observation:**
The acceptance rate across all six departments (aggregate admission decisions) for men is about 44%, while it is only roughly 30% for women, a significant difference.

Such reversals caused by aggregation operations are called ***Simpson's paradox***

# Fairness from a causality perspective

Simpson's paradox causes discomfort to some, because **intuition** suggests that a trend which holds for all subpopulations should also hold at the population level.

1st **Observation**:

What is evident from the data is that **gender influences department choice**. Women and men appear to have different preferences for different fields of study.

# Fairness from a causality perspective

Observational Data

Source: https://fairmlbook.org/

UC Berkeley admissions data from 1973.

| | Men | | Women | |
|---|---|---|---|---|
| Department | Applied | Admitted (%) | Applied | Admitted (%) |
| A | 825 | 62 | 108 | 82 |
| B | 520 | 60 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |

**Observation:**
**gender influences department choice**

# Fairness from a causality perspective

Simpson's paradox causes discomfort to some, because **intuition** suggests that a trend which holds for all subpopulations should also hold at the population level.

2ond **Observation:**

Moreover, different departments have different admission criteria. Some have **lower acceptance rates**, **some higher**.

# Fairness from a causality perspective

Observational Data

Source: https://fairmlbook.org/

UC Berkeley admissions data from 1973.

| | Men | | Women | |
|---|---|---|---|---|
| Department | Applied | Admitted (%) | Applied | Admitted (%) |
| A | 825 | 62 | 108 | 82 |
| B | 520 | 60 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |

**Observation:**

**Some Depts. have lower acceptance rates, some higher**

# Fairness from a causality perspective

Observational Data

Source: https://fairmlbook.org/

UC Berkeley admissions data from 1973.

| | Men | | Women | |
|---|---|---|---|---|
| Department | Applied | Admitted (%) | Applied | Admitted (%) |
| A | 825 | 62 | 108 | 82 |
| B | 520 | 60 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |

Less competitive

More competitive

There is a pattern!

# Fairness from a causality perspective

Therefore, one explanation for the data we see is that **women chose to apply to more competitive departments** (men chose to apply to less competitive departments), **hence getting rejected (accepted) at a higher rate than men**.

# Fairness from a causality perspective

Indeed, this is the conclusion an original study (Bickel, Hammel, O'Connell, and others, "Sex Bias in Graduate Admissions.", 1975) drew:

> *The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.*

In other words, the article concluded that the source of gender bias in admissions was a pipeline problem: **Without any wrongdoing by the departments, women were "*shunted by their socialization*" that happened at an earlier stage in their lives.**

# Fairness from a causality perspective

We can ask why women applied to more competitive departments in the first place.

There are several possible reasons.

- Perhaps less competitive departments were unwelcoming of women at the time. This may have been a general pattern at the time or specific to the university.
- Perhaps some departments had a track record of poor treatment of women that was known to the applicants.
- Perhaps the department advertised the program in a manner that discouraged women from applying.

# Fairness from a causality perspective

We can ask why women applied to more competitive departments i

There are seve

- Perhaps less engineering schools, were me. This may have been a ific to the university.
- Perhaps som of poor treatment of licants.
- Perhaps the m in a manner that discoura

There is no way of knowing what was the case from the data we have. We see that at best the original analysis leads to a number of follow-up questions.

At this point, we have two choices. One is to design a new study and collect more data in a manner that might lead to a more conclusive outcome. The other is to argue over which scenario is more likely based on our beliefs and plausible assumptions about the world.

**Causal inference** is helpful in either case.

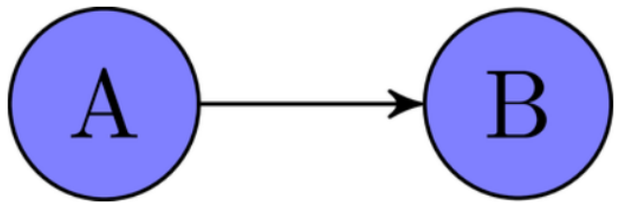
# Causal Inference

# Causality: Intuition

If I randomly picked a person from a population, and found out that she owns a Tesla, I'd find that she was more likely than the average person to have a college degree.

This means that owning a Tesla is ***associated*** with having a college degree, i.e. *knowing if she has a college degree changes the likelihood that she also owns a Tesla.*

**Does this mean that owning a Tesla *causes* people to have a college degree?**



# Getting answers from data

## OBSERVATIONAL
## (passive observation of the world)
## DATASET

| | Sex | Race | Height | Income | Marital Status | Years of Educ. | Liberal-ness |
|---|---|---|---|---|---|---|---|
| R1001 | M | 1 | 70 | 50 | 1 | 12 | 1.73 |
| R1002 | M | 2 | 72 | 100 | 2 | 20 | 4.53 |
| R1003 | F | 1 | 55 | 250 | 1 | 16 | 2.99 |
| R1004 | M | 2 | 65 | 20 | 2 | 16 | 1.13 |
| R1005 | F | 1 | 60 | 10 | 3 | 12 | 3.81 |
| R1006 | M | 1 | 68 | 30 | 1 | 9 | 4.76 |
| R1007 | F | 5 | 66 | 25 | 2 | 21 | 2.01 |
| R1008 | F | 4 | 61 | 43 | 1 | 18 | 1.27 |
| R1009 | M | 1 | 69 | 67 | 1 | 12 | 3.25 |

Let's consider some different features in this dataset, $(X, Y, Z)$.

Which kind of questions can we answer from this dataset?

# Association vs. Causality

**Correlation-based query**

Q1: Which is the *expected income* $Y$ that would have been observed if an individual had $X = x$ and $Z = z$?

**Association** (or prediction) is using data to map some features of the world (the inputs) to other features of the world (the outputs). For example, $\mathbb{E}(Y \mid X, Z)$.

All we need to do prediction is a dataset sampled from $p(X, Y, Z)$ and some inference tools (statistical inference & machine learning).

Mapping observed inputs to observed outputs is a **natural candidate for automated data analysis** ("data speaks for itself") because this task only requires 1) a large dataset with inputs and outputs, 2) an algorithm that establishes a mapping between inputs and outputs, and 3) a metric to assess the performance of the mapping, often based on a gold standard.

# Association vs. Causality

**Causal query about the effect of Race on Income**

Q2: Estimate the *mean income* $Y$ that would have been observed if **all individuals** had $X = x_1$ (*race=1*) vs. if they had $X \neq x_1$ (*race=2,3,4,5*).

**Causal Inference** is using data to predict certain features of the world **if the world had been different**. We cannot get these data by passive observation of the world! The world was different!

Answers to causal questions cannot be derived exclusively from $p(X, Y, Z)$. Answering a causal question (yes, sometimes is possible!) typically requires a combination of data, analytic techniques, and **expert knowledge**.

# Association vs. Causality

| | Sex | Race | Height | Income | Marital Status | Years of Educ. | Liberal-ness |
|---|---|---|---|---|---|---|---|
| R1001 | M | 1 | 70 | 50 | 1 | 12 | 1.73 |
| R1002 | M | 2 | 72 | 100 | 2 | 20 | 4.53 |
| R1003 | F | 1 | 55 | 250 | 1 | 16 | 2.99 |
| R1004 | M | 2 | 65 | 20 | 2 | 16 | 1.13 |
| R1005 | F | 1 | 60 | 10 | 3 | 12 | 3.81 |
| R1006 | M | 1 | 68 | 30 | 1 | 9 | 4.76 |
| R1007 | F | 5 | 66 | 25 | 2 | 21 | 2.01 |
| R1008 | F | 4 | 61 | 43 | 1 | 18 | 1.27 |
| R1009 | M | 1 | 69 | 67 | 1 | 12 | 3.25 |

**Causal effect of Race on Income**

Intervention Intervention

The answer to $\mathbb{E}[Y \mid do(X = x_1)] - E[Y \mid do(X \neq x_1)]$ is not $\mathbb{E}[Y \mid X = x_1] - E[Y \mid X \neq x_1]$

# Association vs. Causality

Let's say we have i.i.d. data sampled from some joint $p(X, Y, Z)$.

Say we are ultimately interested in how variable $Y$ behaves given $X$.
At a high level, one can ask this question in two ways:

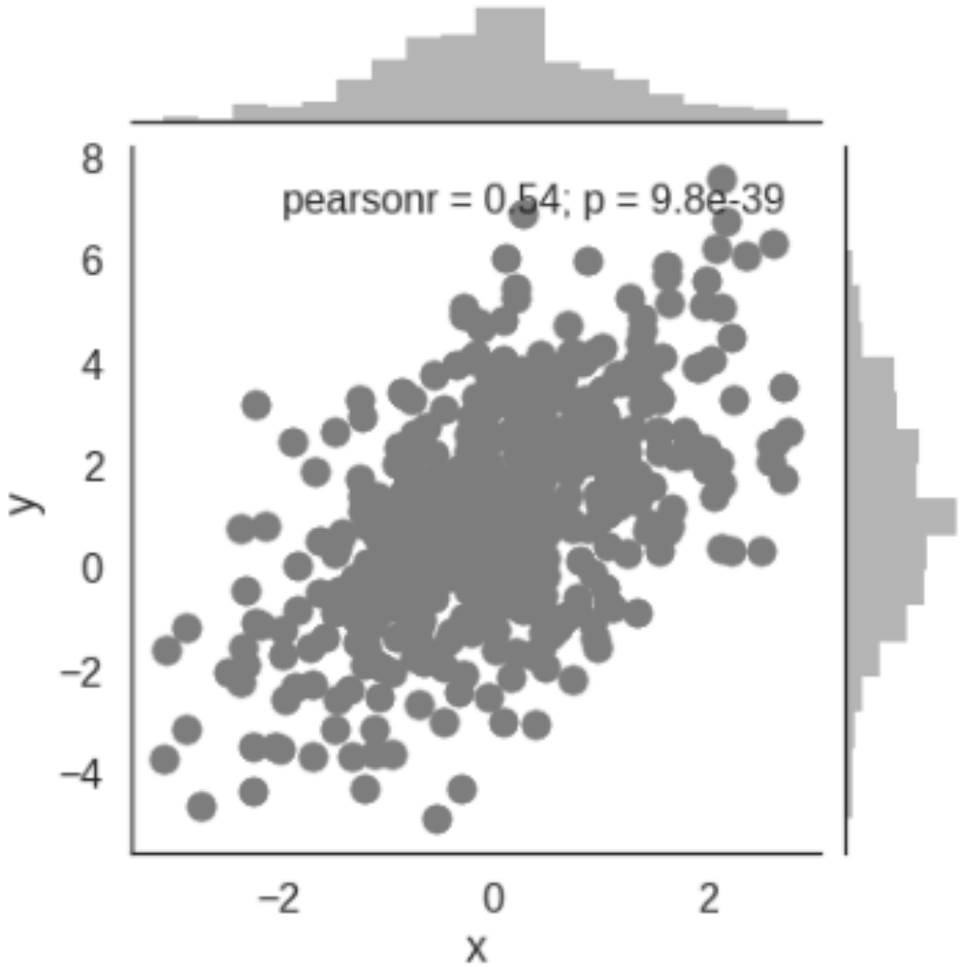
- **observational**, based on $p(Y \mid X, Z)$:
  What is the distribution of $Y$ given that I **observe** variable $X$ takes value $x$?

- **interventional**, based on $p(Y \mid do(X), Z) \neq p(Y \mid X, Z)$:
  What is the distribution of $Y$ if I were to **set** the value of $X$ to $x$.

  This describes the distribution of $Y$ I would observe if I **intervened** in the data generating process by artificially forcing the variable $X$ to take value $x$, but otherwise **simulating the rest of the variables according to the original process that generated the data**.

# Association vs. Causality

In order to understand what is $p(Y \mid do(X))$, let's suppose I have observed $p(X, Y)$.




https://www.inference.vc/untitled/

This is all we need to compute $p(Y \mid X)$. We can give an answer to any associational question.

For example:

- What is the expected value of $Y$ if we observe $X = 3$? (Regression)
- What is the expected MAX/MIN/MEDIAN value of $Y$ if we observe $X = 3$? (Quantile regression)

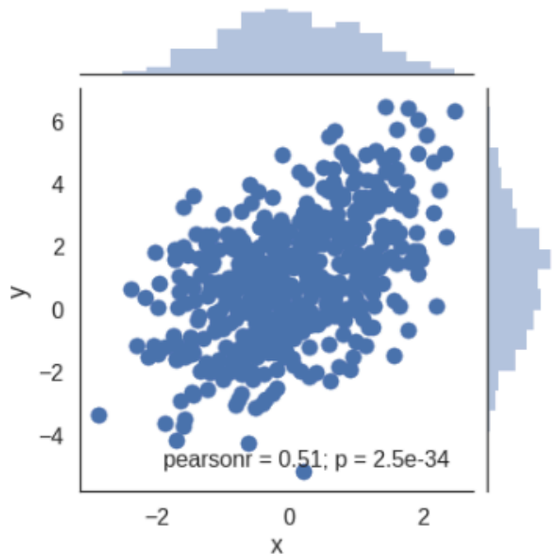
# Association vs. Causality

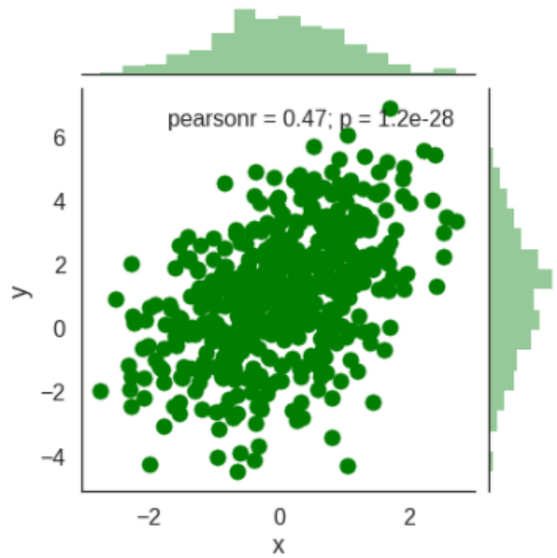
Given $p(X, Y)$, **there are several generative models** that are compatible with $p(X, Y)$:

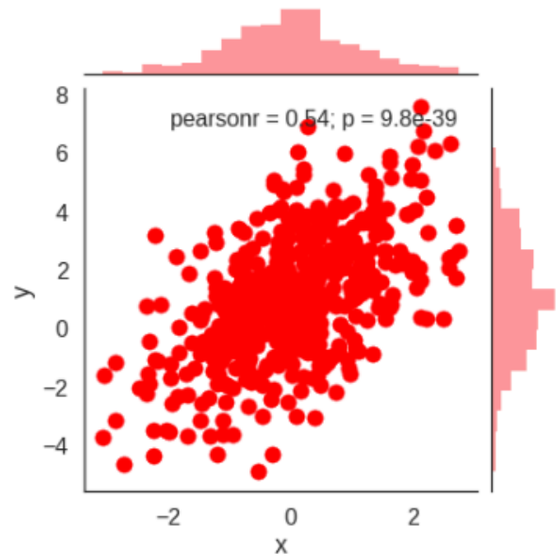
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



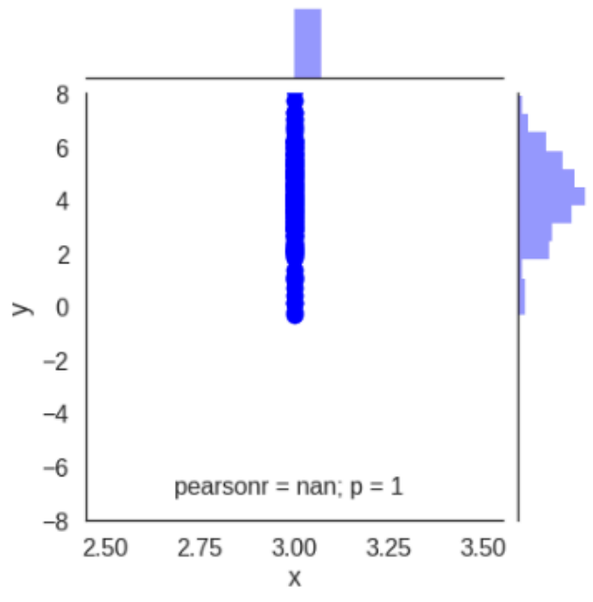

https://www.inference.vc/untitled/

Based on the joint distribution the three scripts are indistinguishable.

# Association vs. Causality

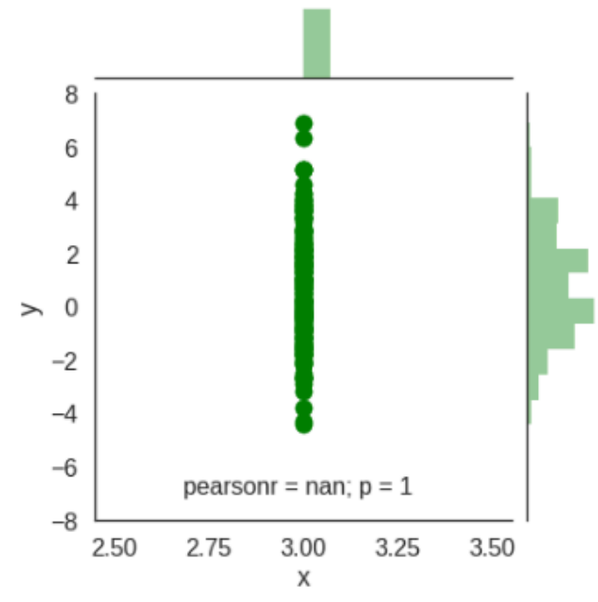
Let's now consider an **intervention** $p(Y \mid do(X = 3))$

```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```



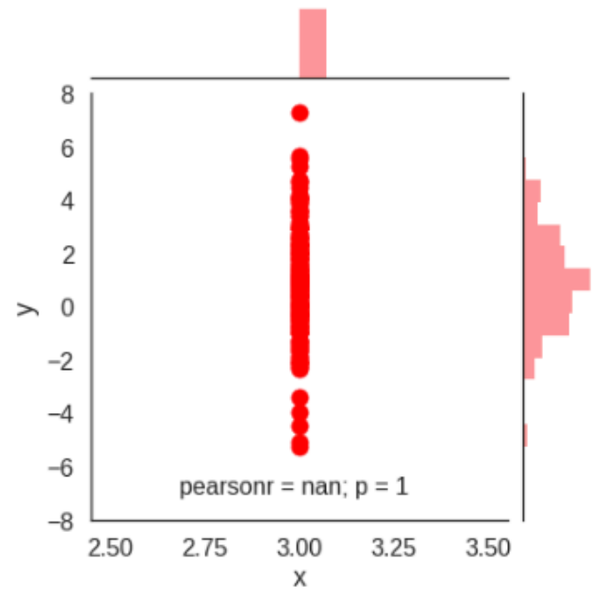

$p(Y \mid do(X)) \neq p(Y \mid X)$

```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```




$p(Y \mid do(X)) = p(Y \mid X)$

```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```




$p(Y \mid do(X)) = p(Y \mid X)$

**The joint distribution of data $p(X, Y, Z)$ alone is insufficient to predict behavior under interventions.**

https://www.inference.vc/untitled/

# Association vs. Causality

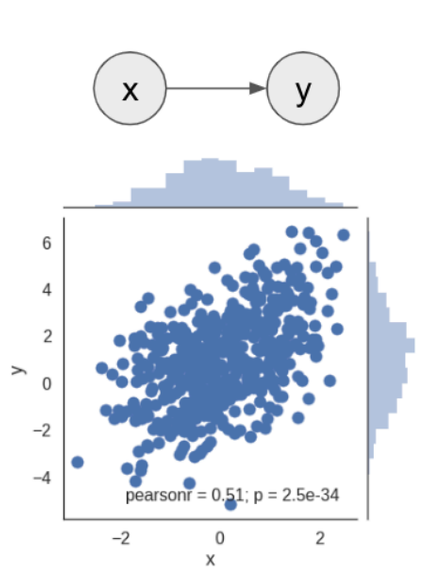
An intervention can be understood as a **modification of the generative model of the data, producing a different probability distribution** $p(do(X), Y, Z)$.

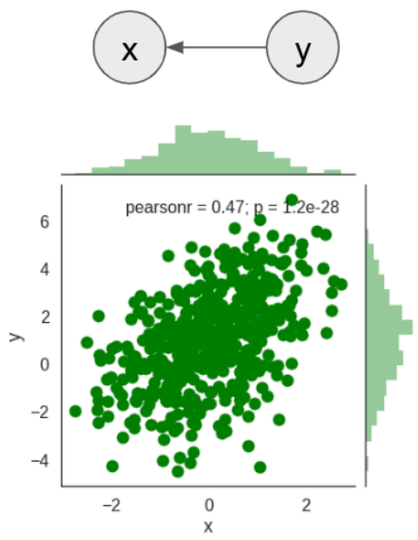
The resulting distribution depends on the original model:

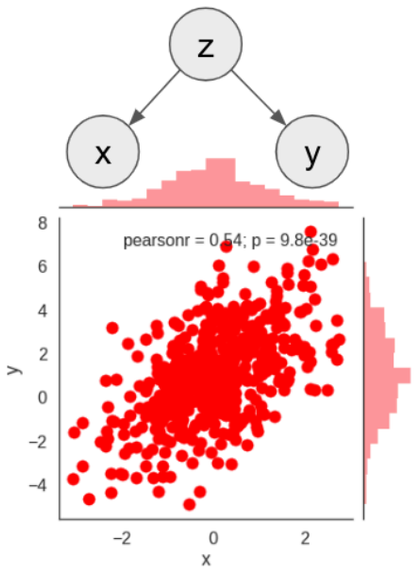
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```




https://www.inference.vc/untitled/

**Directed Acyclic Graphs (DAG).**

No assumptions about the exact form of the functional relationships are needed. The only requirement is that causal relationships are **acyclic**.

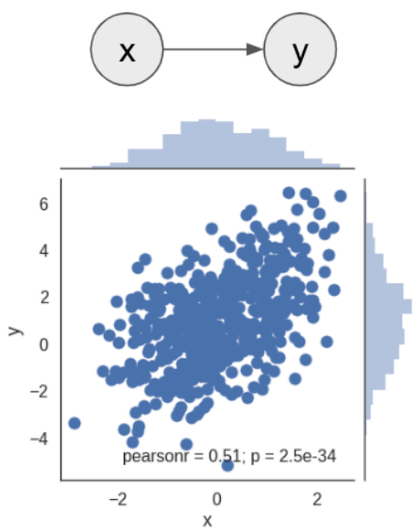
# Association vs. Causality

An intervention can be understood as a **modification of the generative model of the data, producing a different probability distribution** $p(do(X), Y, Z)$.

The resulting distribution depends on the original model:

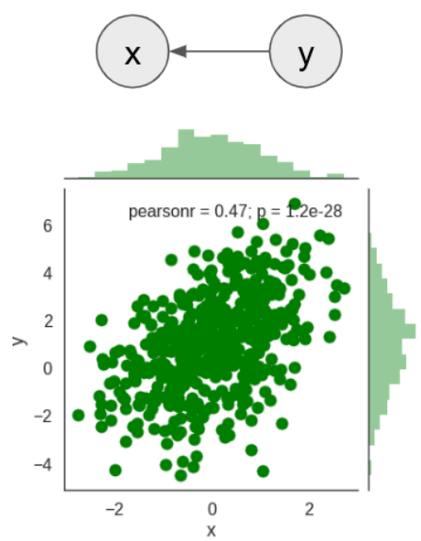
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

$X$ is the cause of $Y$



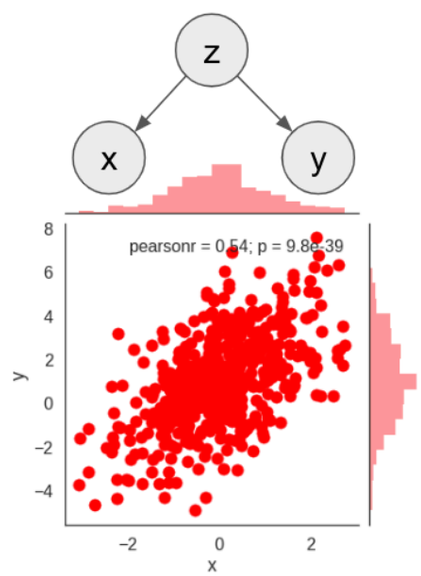

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

$Y$ is the cause of $X$




```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```

$X$ and $Y$ are not causally related (but they are associated!)



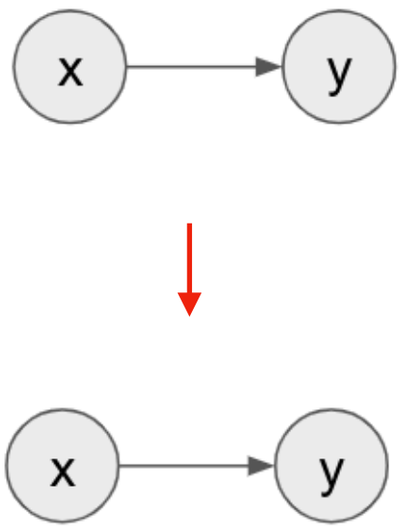

https://www.inference.vc/untitled/

**Directed Acyclic Graphs (DAG).**

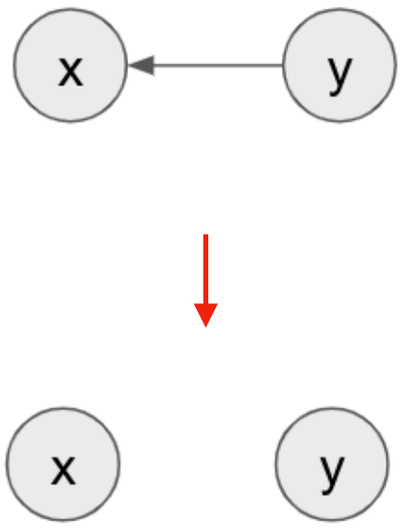
No assumptions about the exact form of the functional relationships are needed. The only requirement is that causal relationships are **acyclic**.

# Association vs. Causality

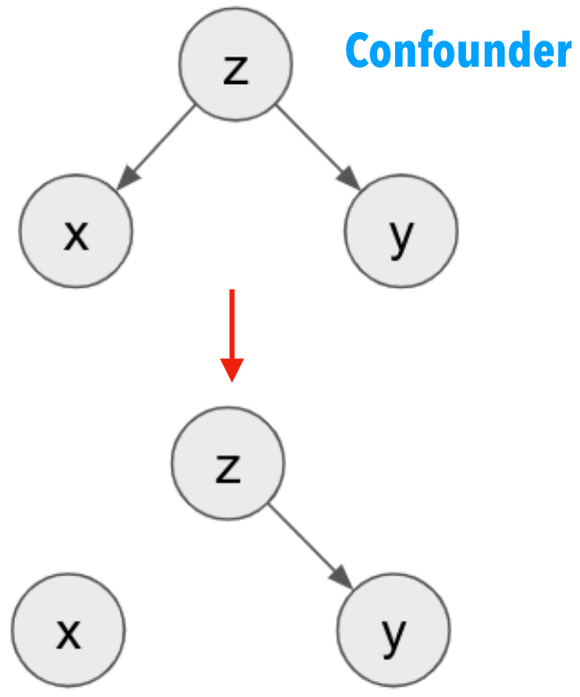
Graphically, to simulate the effect of an intervention, you **mutilate** the graph by **removing all edges that point into the variable on which the intervention is applied**, in this case $X$.




$p(Y \mid do(X) = p(Y \mid X = 3)$




$p(Y \mid do(X) = p(Y)$

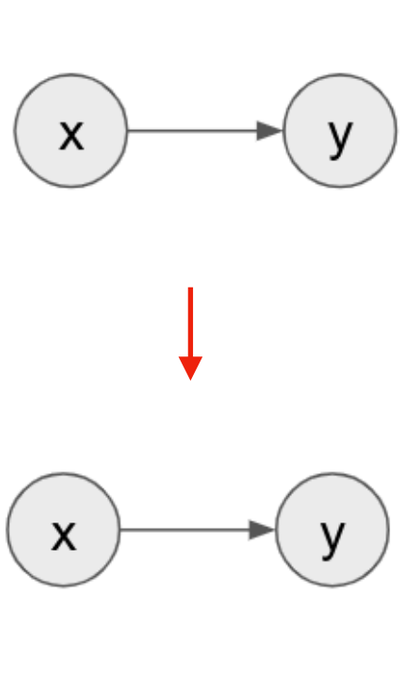



$p(Y \mid do(X) = p(Y)$

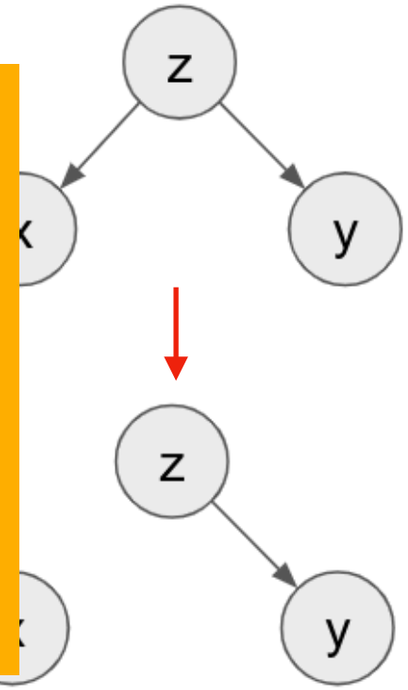
Just by only looking at the causal diagram, we are now able to predict how the scripts are going to behave under the intervention $X = 3$.

# Association vs. Causality

Graphically, to simulate the effect of an intervention, you **mutilate** the graph by removing all edges that point into the variable on which the intervention is applied, in this case $X$.




Interventional distributions are not equivalent to observational distributions!




$p(Y \mid do(X)) = p(Y \mid X = 3)$

$p(Y \mid do(X)) = p(Y)$

$p(Y \mid do(X)) = p(Y)$

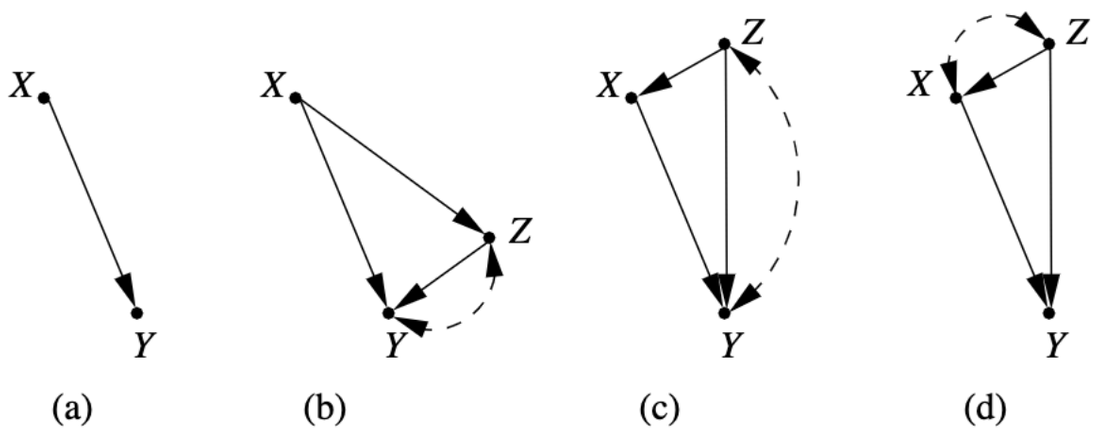
Just by only looking at the causal diagram, we are now able to predict how the scripts are going to behave under the intervention $X = 3$.

# Association vs. Causality

An essential matter in causal inference is that of a query's **identifiability**.

Given a causal query (for example, $p(Y \mid do(X = 3))$) for a certain DAG, we say it is **identifiable** if we can derive an statistical estimand (**only using observational terms**) for this query using the rules of **do-calculus**.

The ***do*-calculus** is an axiomatic system for replacing probability formulas containing the *do* operator with ordinary conditional probabilities. It consists of three axiom schemas that provide **graphical criteria** for when certain substitutions may be made.

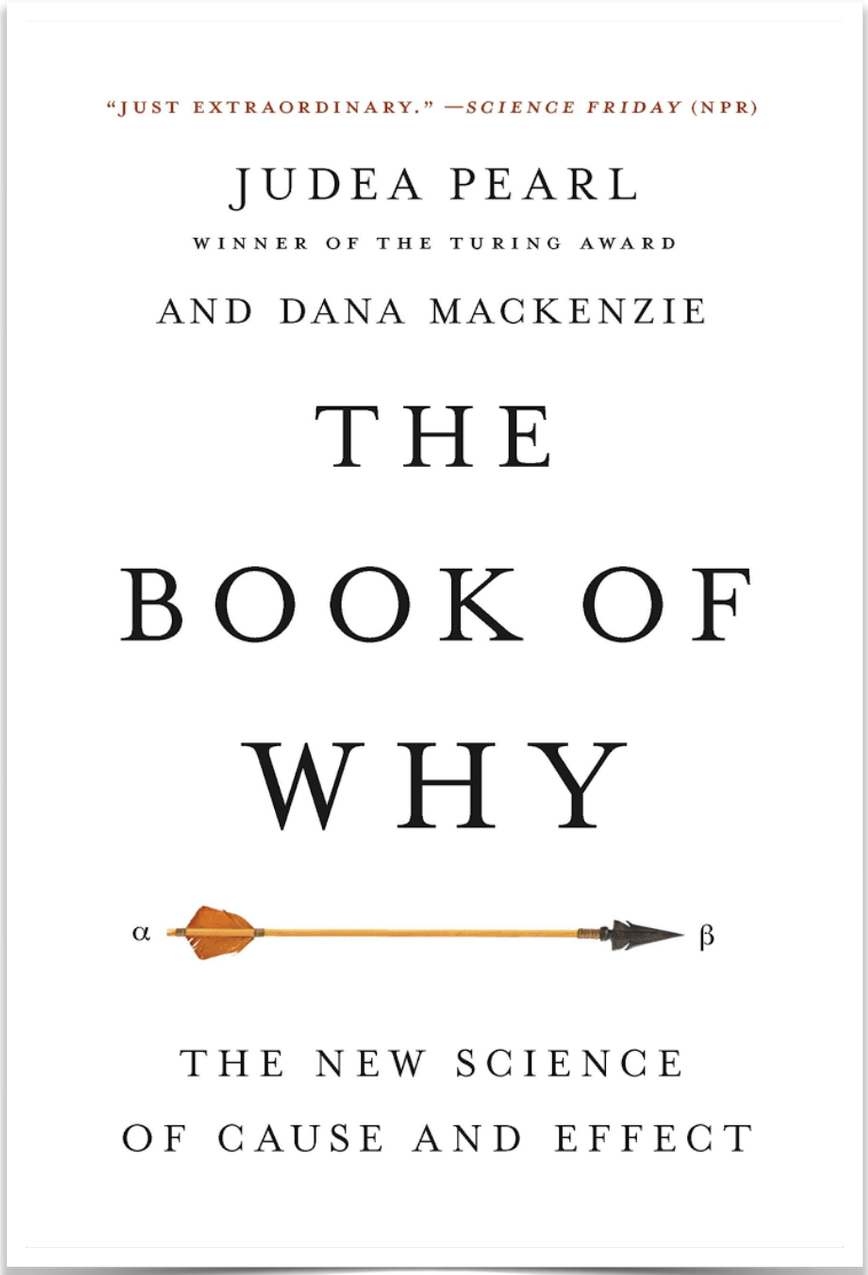


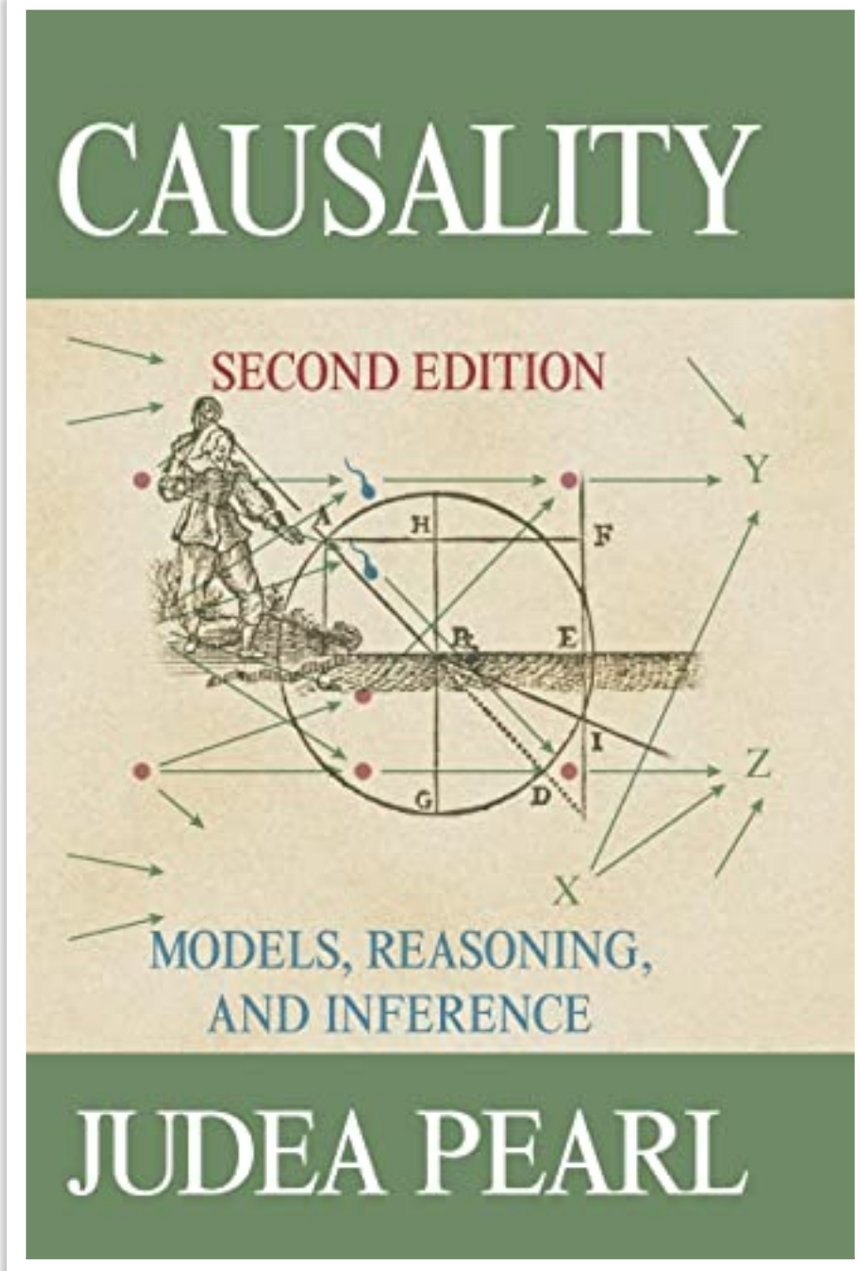

Causal graphs where $P(y|do(\mathbf{x}))$ is identifiable

Source: Complete Identication Methods for Causal Inference, PhD Thesis, University of California. I.Shpitser
https://ftp.cs.ucla.edu/pub/stat_ser/shpitser-thesis.pdf

Dashed lines correspond to **unobserved confounders**, associations produced by unobserved variables.

# Do-calculus


"JUST EXTRAORDINARY." —SCIENCE FRIDAY (NPR)
JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE
THE
BOOK OF
WHY
α
β
THE NEW SCIENCE
OF CAUSE AND EFFECT


CAUSALITY
SECOND EDITION
MODELS, REASONING,
AND INFERENCE
JUDEA PEARL

# Causal Inference and do-calculus

There are two ways to measure the **causal relationship** between two variables, $X$ and $Y$:

1. The easiest way is an **intervention** in the real world: You randomly force $X$ to have different values and you measure $Y$.

   This is what we do in Randomized Clinical Trial (RCT) or in an A/B Test.

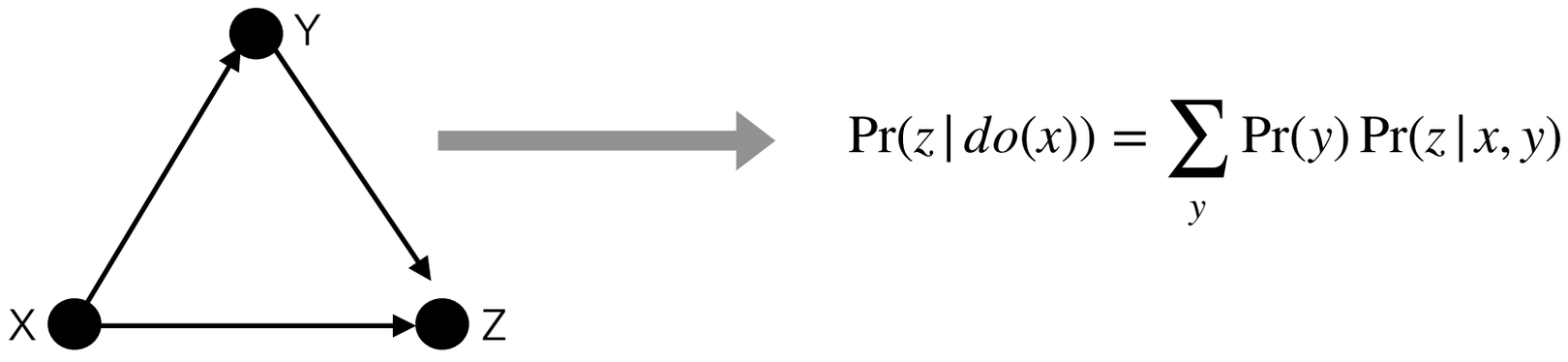
   **This is not always feasible** (because of **practical**, **ethical** or **economical** reasons)

# Causal Inference and do-calculus

There are two ways to measure the **causal relationship** between two variables, $X$ and $Y$:

2. If the query is identifiable, **do-calculus** allows us to massage $p(X, Y, Z)$ until we can express $p(Y \mid do(X))$ in terms of various marginals, conditionals and expectations under $p(X, Y, Z)$

**Example**:

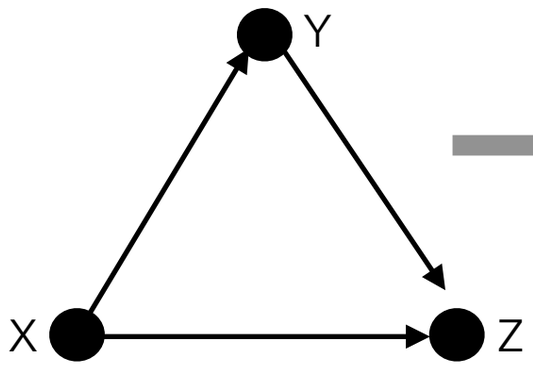


# Causal Inference and do-calculus

There are two ways to measure the **causal relationship** between two variables, $X$ and $Y$:

2. If the query is identifiable, **do-calculus** allows us to massage $p(X, Y, Z)$ until we can express $p(Y \mid do(X))$ in terms of various marginals, conditionals and expectations under $p(X, Y, Z)$

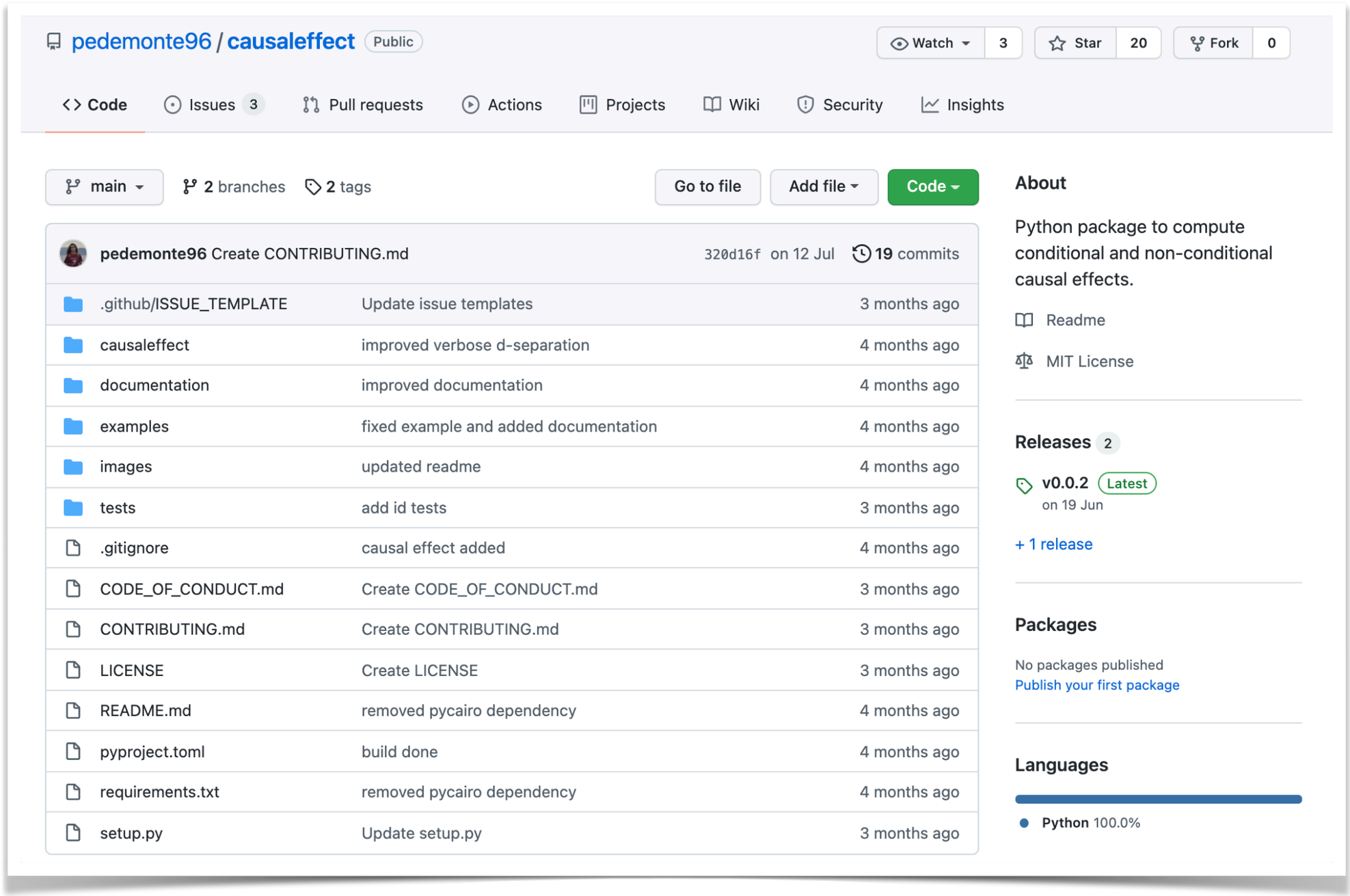
**Example**:



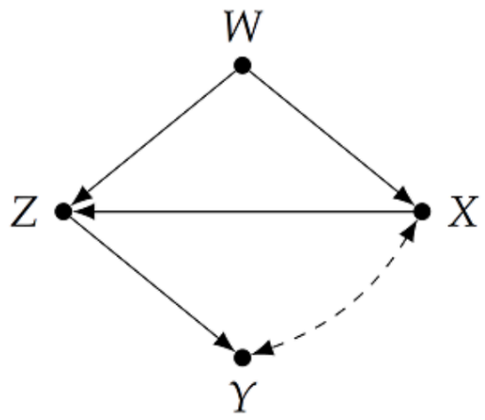

$$\Pr(z \mid do(x)) = \sum_{y} \Pr(y) \Pr(z \mid x, y)$$

Only depends on the causal graph!

# Causal Inference and do-calculus



# Causal Inference and do-calculus




```
import causaleffect
G = causaleffect.createGraph(['X<->Y', 'Z->Y', 'X->Z', 'W->X', 'W->Z'])
causaleffect.plotGraph(G)
```

```
P = causaleffect.ID({'Y'}, {'X'}, G)
P.printLatex()
```

The code above computes the causal effect, and returns a string encoding the distribution in LaTeX notation:

```
'\sum_{w, z}P(w)P(z|w, x)\left(\sum_{x}P(x|w)P(y|w, x, z)\right)'
```

This string, in LaTeX, is

$$\sum_{w,z} P(w)P(z|w,x)\left(\sum_{x} P(x|w)P(y|w,x,z)\right)$$

# Causal Inference Process

**Asking a causal query**

**Gathering data and knowledge**

Expert knowledge!

**Building a causal model**

Causal information is not provided by data.

**Identifying the causal query**

Sometimes it cannot be identified.

**Estimating the causal effect**

**Validating the result**

# Causal Inference Process

From “Causal Inference in AI Education: A Primer”

**Example 3.1.** ***AdBot*** *Consider an online advertising agent attempting to maximizing clickthroughs, with* $X \in \{0, 1\}$ *representing two ads,* $Y \in \{0, 1\}$ *whether or not it was clicked upon, and* $Z \in \{0, 1\}$ *the sex of the viewer. A marketing team collects the following data on purchases following ads shown to focus groups to be used by AdBot:*

| | Ad 0 | Ad 1 |
|---|---|---|
| **Male** | 108/120 (90%) | 340/400 (85%) |
| **Female** | 266/380 (70%) | 65/100 (65%) |
| **Total** | 374/500 (75%) | 405/500 (81%) |

**Table 1.** Clickthroughs in the AdBot setting striated by the ad shown to participants in a focus group, and the sex of the viewer.

**If the sex of a viewer is not know, which ad is the best choice?**

# Causal Inference Process

**Example 3.1.** ***AdBot*** *Consider an online advertising agent attempting to maximizing clickthroughs, with* $X \in \{0, 1\}$ *representing two ads,* $Y \in \{0, 1\}$ *whether or not it was clicked upon, and* $Z \in \{0, 1\}$ *the sex of the viewer. A marketing team collects the following data on purchases following ads shown to focus groups to be used by AdBot:*

| | Ad 0 | Ad 1 |
|---|---|---|
| **Male** | 108/120 (90%) | 340/400 (85%) |
| **Female** | 266/380 (70%) | 65/100 (65%) |
| **Total** | 374/500 (75%) | 405/500 (81%) |

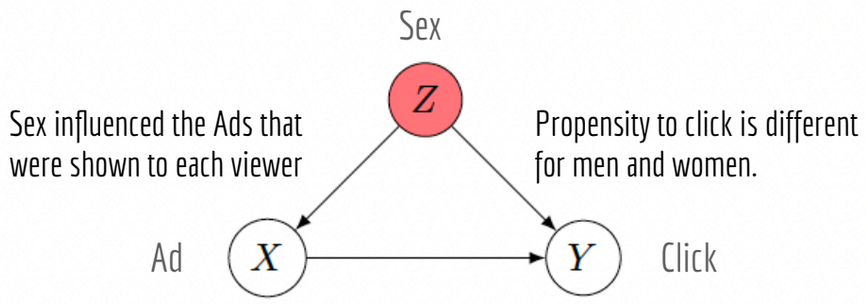
**Table 1.** Clickthroughs in the AdBot setting striated by the ad shown to participants in a focus group, and the sex of the viewer.

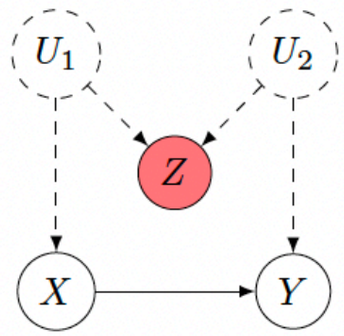
**If the sex of a viewer is not know, which ad is the best choice?**

# Causal Inference Process

**If the sex of a viewer is not know, which ad is the best choice?**

**These are two different causal stories:**



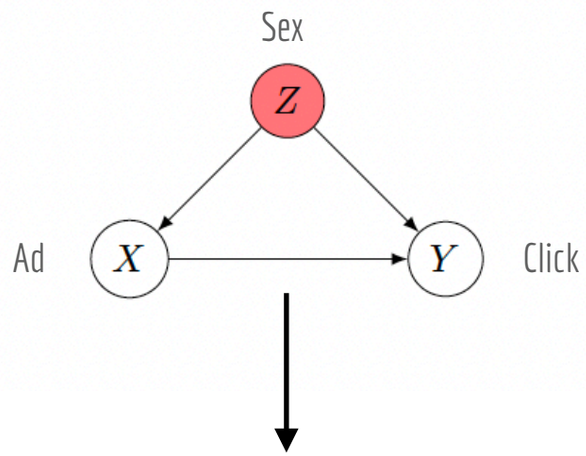



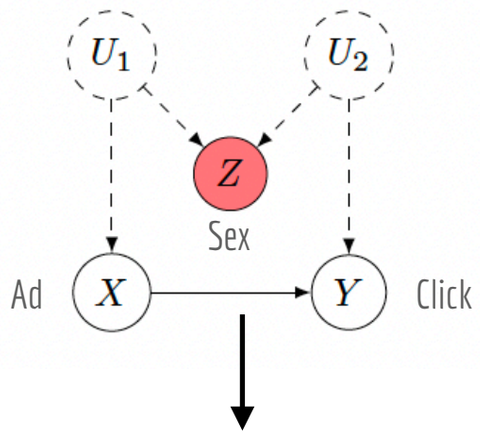

Relevant quantities for answering the question:

$$\Pr(Y \mid do(X_0))?$$
$$\Pr(Y \mid do(X_1))?$$

# Causal Inference Process

**These are two different interventional stories:**




```
1 G = causaleffect.createGraph(['X->Y', 'Z->Y', 'Z->X'])
2 P = causaleffect.ID({'Y'}, {'X'}, G)
```

```
1 G = causaleffect.createGraph(['Z<->Y', 'Z<->X', 'X->Y'])
2 P = causaleffect.ID({'Y'}, {'X'}, G)
```

$$p(Y \mid do(X)) = \sum_{z} P(Y \mid X, Z) P(Z)$$

$$p(Y \mid do(X)) = P(Y \mid X)$$

| | Ad 0 | Ad 1 |
|---|---|---|
| **Male** | 108/120 (90%) | 340/400 (85%) |
| **Female** | 266/380 (70%) | 65/100 (65%) |
| **Total** | 374/500 (75%) | 405/500 (81%) |

# Causal Inference Process

**Example 3.1.** ***AdBot*** *Consider an online advertising agent attempting to maximizing clickthroughs, with* $X \in \{0, 1\}$ *representing two ads,* $Y \in \{0, 1\}$ *whether or not it was clicked upon, and* $Z \in \{0, 1\}$ *the sex of the viewer. A marketing team collects the following data on purchases following ads shown to focus groups to be used by AdBot:*

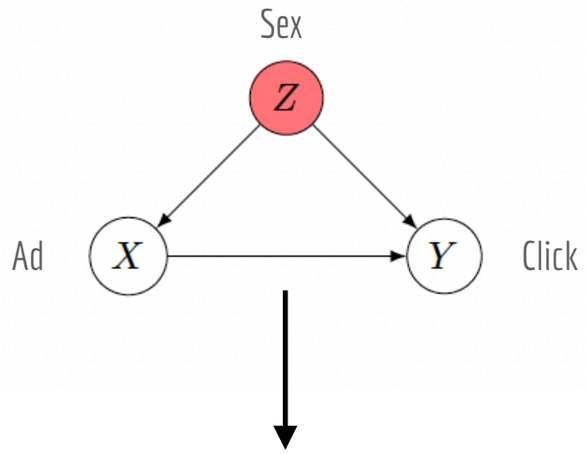
| | Ad 0 | Ad 1 |
|---|---|---|
| **Male** | 108/120 (90%) | 340/400 (85%) |
| **Female** | 266/380 (70%) | 65/100 (65%) |
| **Total** | 374/500 (75%) | 405/500 (81%) |

$$p(Y \mid \text{do}(X)) = \sum_z P(Y \mid X, Z) P(Z)$$

$$p(Y \mid \text{do}(X)) = P(Y \mid X)$$

**Table 1.** Clickthroughs in the AdBot setting striated by the ad shown to participants in a focus group, and the sex of the viewer.

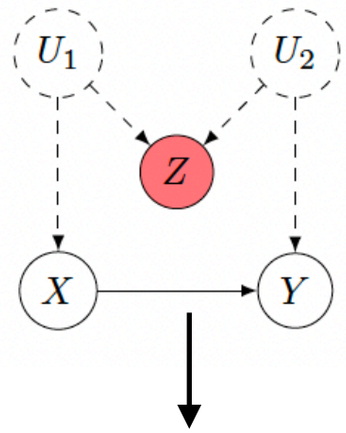
# Causal Inference Process




```
1 G = causaleffect.createGraph(['X->Y', 'Z->Y', 'Z->X'])
2 P = causaleffect.ID({'Y'}, {'X'}, G)
```

$$p(Y \mid \mathrm{do}(X)) = \sum_{z} P(Y \mid X, Z)P(Z)$$

**If this is the generative model of the data, then AdBot should display Ad0.**

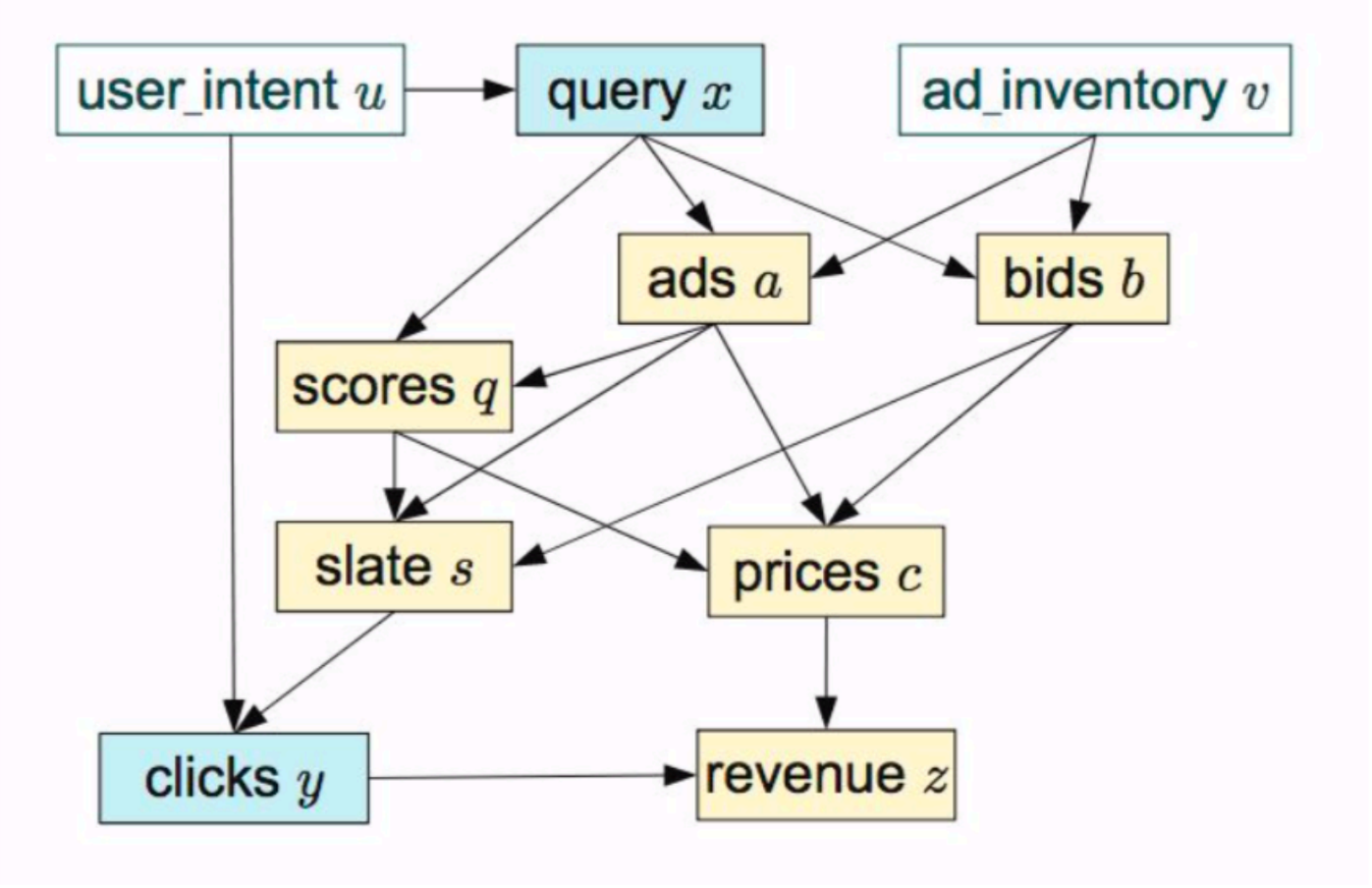



```
1 G = causaleffect.createGraph(['Z<->Y', 'Z<->X', 'X->Y'])
2 P = causaleffect.ID({'Y'}, {'X'}, G)
```

$$p(Y \mid \mathrm{do}(X)) = P(Y \mid X)$$

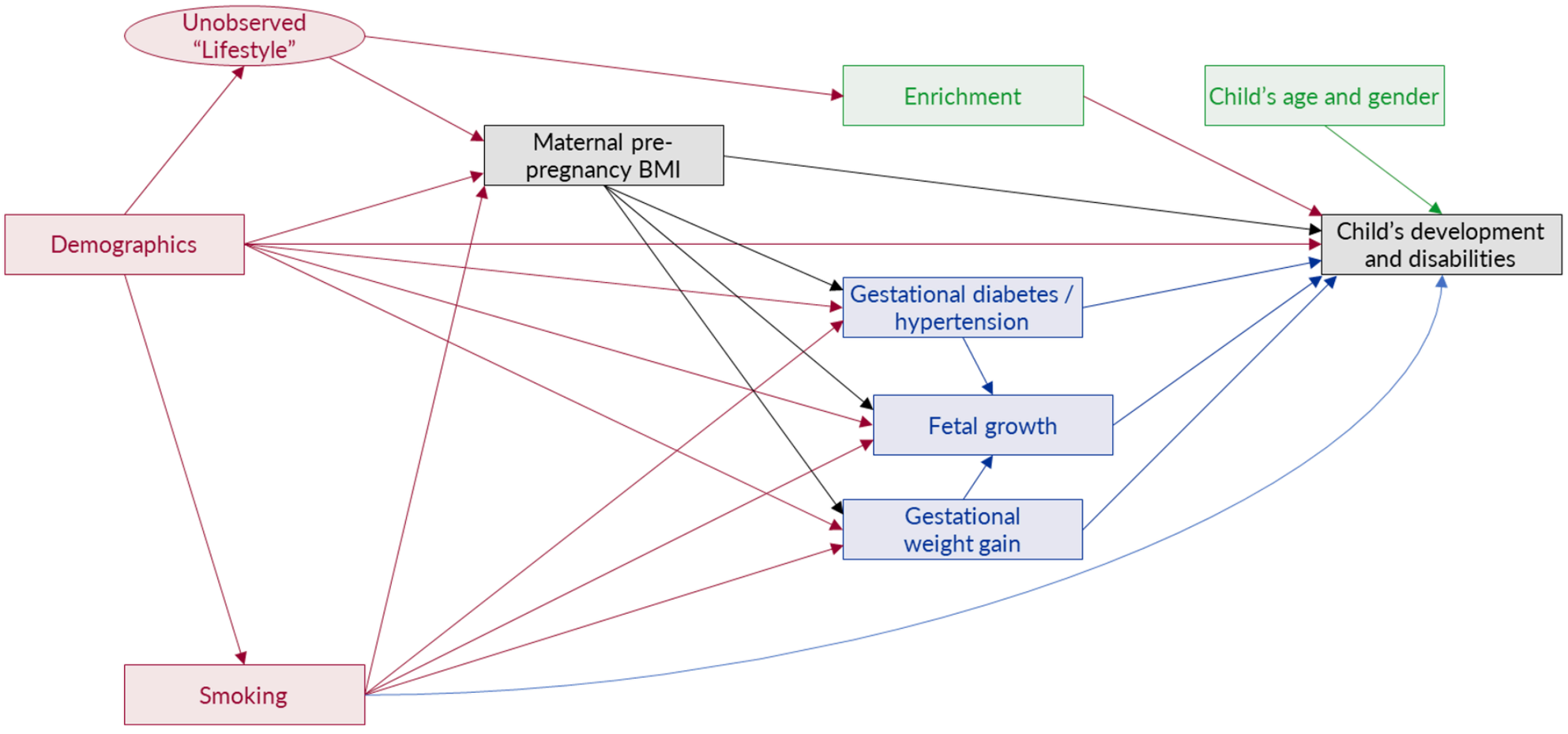
**If this is the generative model of the data, then AdBot should display Ad1.**

# Causal Inference Process



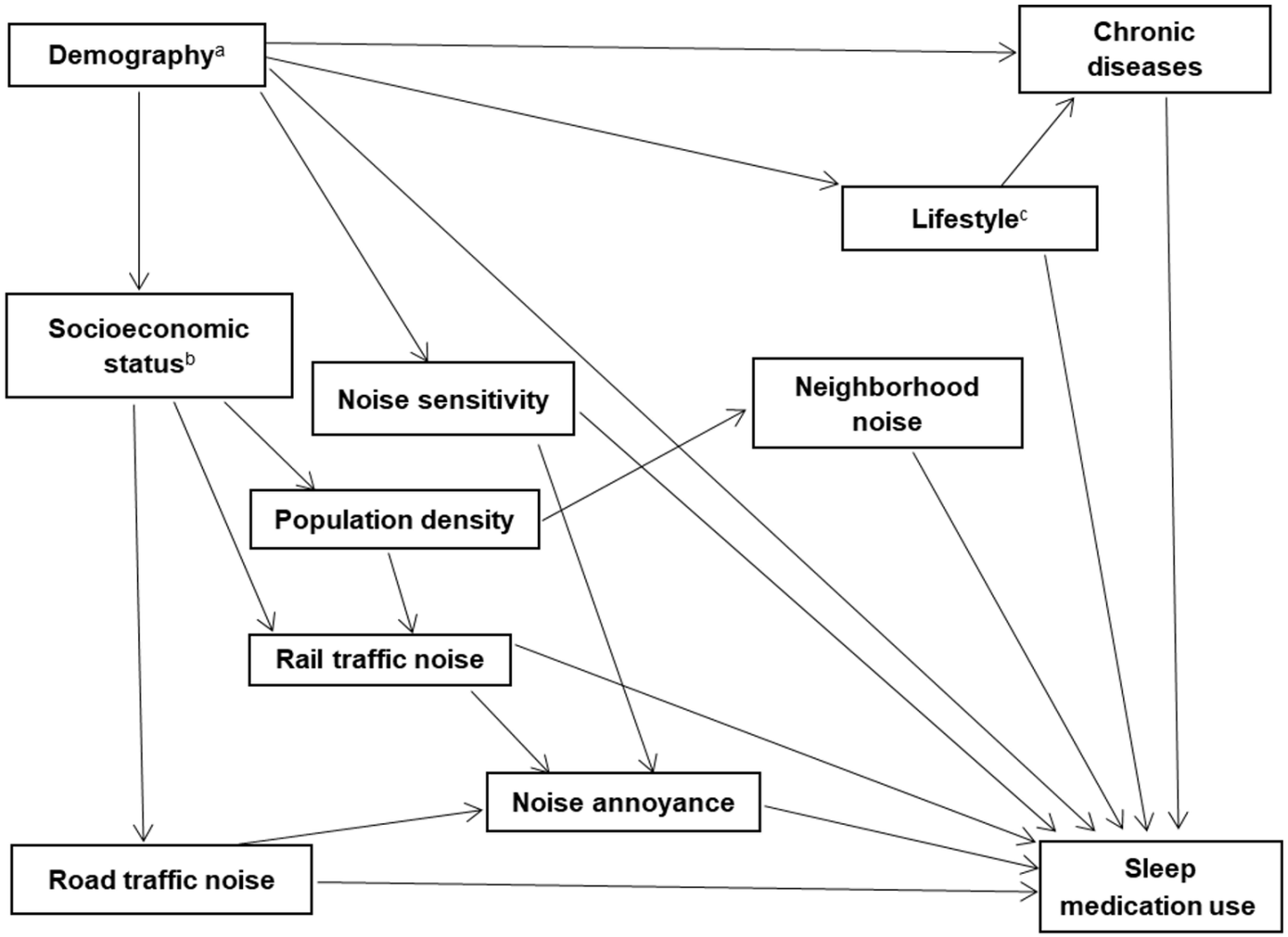

Bottou, Léon, et al. "Counterfactual reasoning and learning systems: the example of computational advertising." *The Journal of Machine Learning Research* 14.1 (2013): 3207-3260.

# Causal Inference Process




**ADAPTED FROM:** Hinkle SN, Sharma AJ, Kim SY, Schieve LA. Maternal prepregnancy weight status and associations with children's development and disabilities at kindergarten. *Int J Obes (Lond)*. 2013;37(10):1344-51. DOI: 10.1038/ijo.2013.128 (Figure 1). Freely available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4407562

# Causal Inference Process




**REPRODUCED UNDER CC-BY 4.0 LICENSE FROM:** Evandt J, Oftedal B, Krog NH, et al. Road traffic noise and registry based use of sleep medication. *Environ Health.* 2017;16(1):110. DOI: 10.1186/s12940-017-0330-5 (Figure S1). Freely available at: https://www.doi.org/10.1186/s12940-017-0330-5

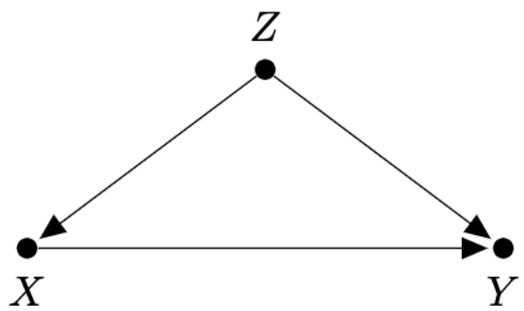
# Causal Inference and SCM

The **causal diagram** can be seen as a representation of an underlying **structural causal model** (generative model).

A structural causal model is comprised of three components:

1. A set of **variables** describing the state of the universe and how it relates to a particular data set we are provided.
2. **Causal model (DAG)**, which describe the causal effect variables have on one another.
3. A **probability distibution** defined over observed variables in the model, describing the likelihood that each variable takes a particular value.

| | Sex | Race | Height | Income | Marital Status | Years of Educ. | Liberal-ness |
|---|---|---|---|---|---|---|---|
| R1001 | M | 1 | 70 | 50 | 1 | 12 | 1.73 |
| R1002 | M | 2 | 72 | 100 | 2 | 20 | 4.53 |
| R1003 | F | 1 | 55 | 250 | 1 | 16 | 2.99 |
| R1004 | M | 2 | 65 | 20 | 2 | 16 | 1.13 |
| R1005 | F | 1 | 60 | 10 | 3 | 12 | 3.81 |
| R1006 | M | 1 | 68 | 30 | 1 | 9 | 4.76 |
| R1007 | F | 5 | 66 | 25 | 2 | 21 | 2.01 |
| R1008 | F | 4 | 61 | 43 | 1 | 18 | 1.27 |
| R1009 | M | 1 | 69 | 67 | 1 | 12 | 3.25 |

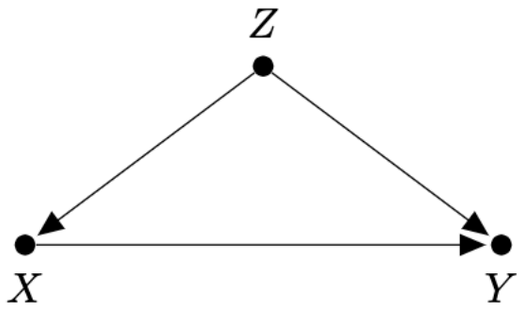



$$P(X, Y, Z) = P(Z)P(X \mid Z)P(Y \mid X, Z)$$

$f_1 \quad f_2 \quad f_3$

# Causal Inference and SCM

## We can estimate $f_i$ from data by using predictive methods.




$$Z \leftarrow f_1(\epsilon_1)$$
$$X \leftarrow f_2(Z, \epsilon_2)$$
$$Y \leftarrow f_3(X, Z, \epsilon_3)$$

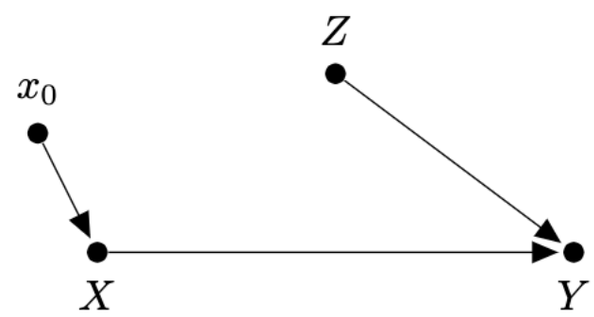
Structural Causal Model

$\epsilon_i$ are **exogenous background factors** represented by an arbitrary noise distribution.

Actions can now be defined as interventions on variables in the model. For example, intervening on $X$ amounts to deleting $f_2$ and setting $X$ to a constant value $x_0$.

$$Z \leftarrow f_1(\epsilon_1)$$
$$X \leftarrow x_0$$
$$Y \leftarrow f_3(X, Z, \epsilon_3)$$

Modified
Structural Causal Model



# Causal Inference and SCM

Given a **certain observational sample** $e = (x_e, y_e, z_e)$ and an intervention $do(X = x_q)$, a **counterfactual** is the result of an hypothetical experiment in the past, what would have happened to the value of variable $Y$ had we intervened on $X$ by assigning value $x_q$.

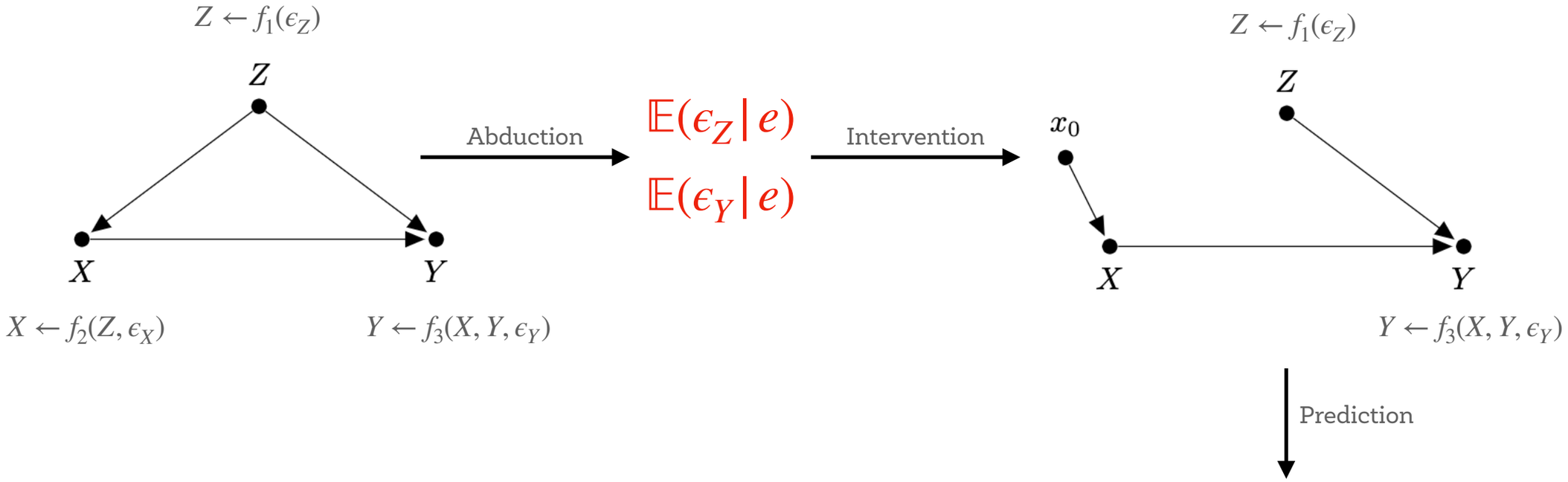
**Identifiable** counterfactuals can be computed as a three-step process by using a SCM:

- **Abduction**: compute the posterior distribution of $\epsilon$ conditioned on $e$.
- **Intervention**: apply the desired intervention $do(X = x)$
- **Prediction**: compute the required prediction in the intervened distribution.

# Causal Inference and SCM

$e$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| R1001 | M | 1 | 70 | 50 | 1 | 12 | 1.73 |
| R1002 | M | 2 | 72 | 100 | 2 | 20 | 4.53 |
| R1003 | F | 1 | 55 | 250 | 1 | 16 | 2.99 |
| R1004 | M | 2 | 65 | 20 | 2 | 16 | 1.13 |
| R1005 | F | 1 | 60 | 10 | 3 | 12 | 3.81 |
| R1006 | M | 1 | 68 | 30 | 1 | 9 | 4.76 |
| R1007 | F | 5 | 66 | 25 | 2 | 21 | 2.01 |



# Causal Inference and SCM

An SCM encodes the intervened distribution, from which we can **sample** and **compute causal queries** (if they are identifiable).

For example, compute **causal effects**:

$$\mathbb{E}((Y \mid do(X = x_1)) - \mathbb{E}(Y \mid do(X = x_0))$$

# Counterfactuals

## Counterfactuals in our life:

Source: https://christophm.github.io/interpretable-ml-book/counterfactual.html

Let's support that we want to rent an apartment and we train a model with real data to predict a price.

After entering all the details about size, location, whether pets are allowed and so on, the model tells us that we can charge 900€.

How could we get (by doing an intervention) 1000€?  We can play with the feature values of the apartment to see how we can improve the value of the apartment!

We find out that the apartment could be rented out for over 1000 Euro, if it were 15 $m^2$ larger. Interesting, but non-actionable knowledge, because we cannot enlarge the apartment.

Finally, by tweaking only the feature values under our control (built-in kitchen yes/no, pets allowed yes/no, type of floor, etc.), we find out that if we allow pets and install windows with better insulation, we can charge 1000€.

# Causal Discrimination Analysis

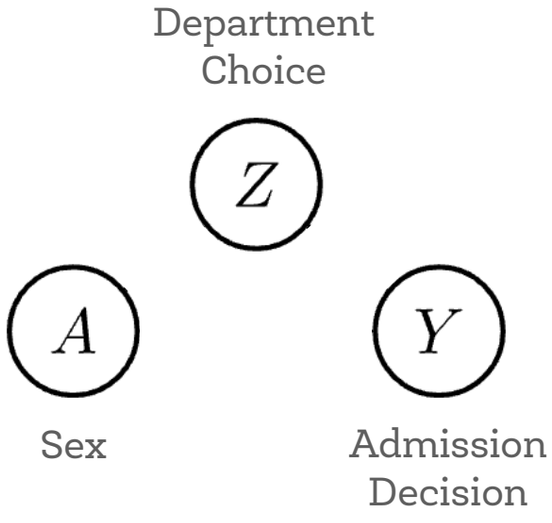
# Graphical Discrimination Analysis

We now explore how we can bring causal graphs to bear on discussions of discrimination.

The first step is to come up with a plausible causal graph consistent with the data that we saw earlier.

BERKELEY ADMISSION

1. It makes sense to draw two arrows (A,Y) and (Z,Y), because bo features are available to the institution when making the admissions decision.



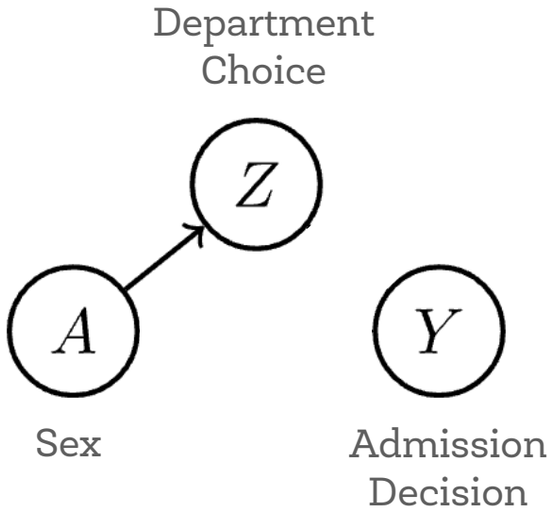

2. A and Z are not statistically independent. We can see from the table that several departments have a statistically significant gender bias among applicants. This means we need to include either the arrow A→Z or Z→A. Which one?

# Graphical Discrimination Analysis

We now explore how we can bring causal graphs to bear on discussions of discrimination.

The first step is to come up with a plausible causal graph consistent with the data that we saw earlier.

BERKELEY ADMISSION



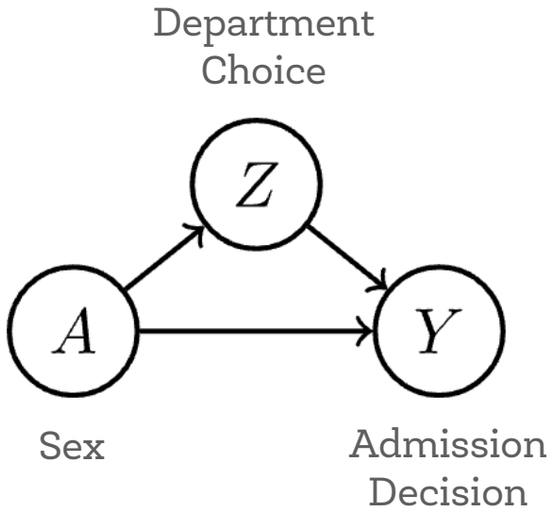

2. A and Z are not statistically independent. We can see from the table that several departments have a statistically significant gender bias among applicants. This means we need to include either the arrow A→Z or Z→A. Which one?

# Graphical Discrimination Analysis

We now explore how we can bring causal graphs to bear on discussions of discrimination.

The first step is to come up with a plausible causal graph consistent with the data that we saw earlier.

BERKELEY ADMISSION

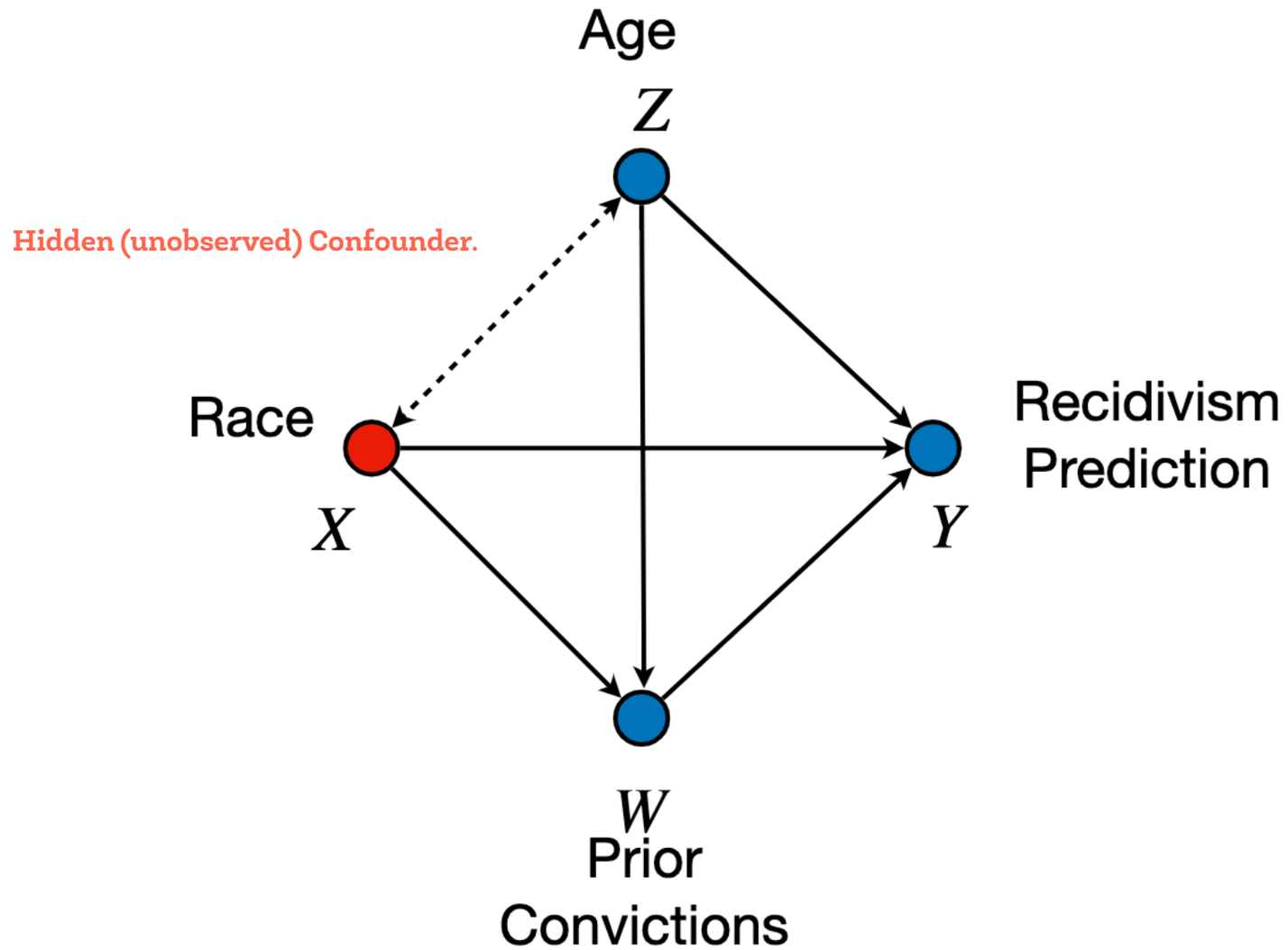



3. To align Bickel's story with our causal graph, we need the variable A to reference whatever ontological entity it is that through this "socialization process" influences intellectual and professional preferences, and hence, department choice.

It is difficult to maintain that this ontological entity coincides with sex as a biological trait. There is no scientific basis to support that the biological trait sex is what determines our intellectual preferences.

# Graphical Discrimination Analysis

## COMPASS PREDICTION

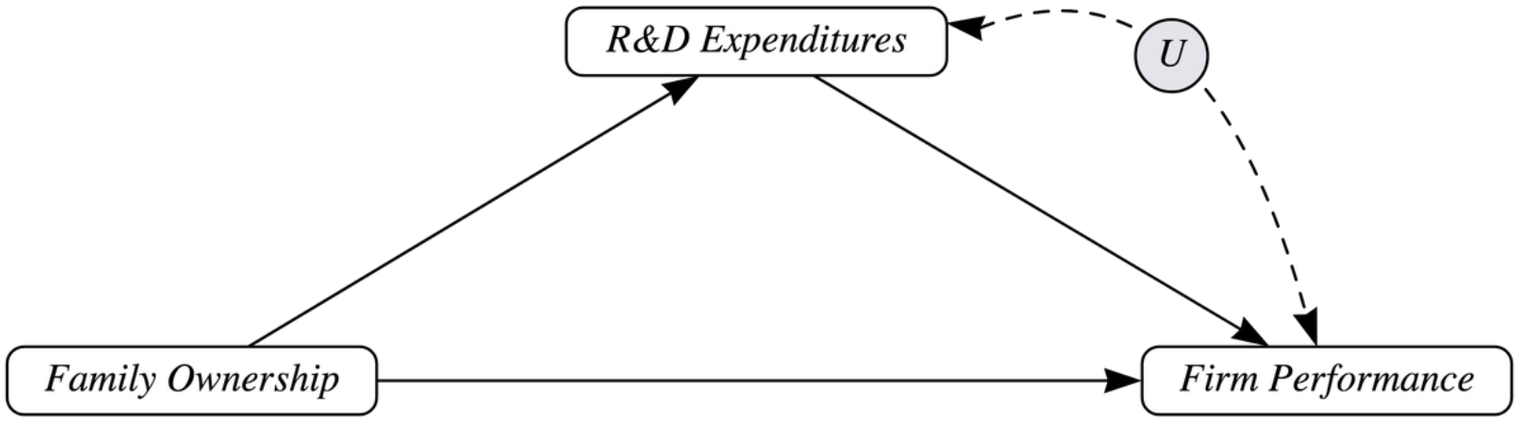



https://fairness.causalai.net/

# Graphical Discrimination Analysis

We have described causal inference with observational data under an assumption of no **unobserved confounders**.

This is not the case in most of the cases!

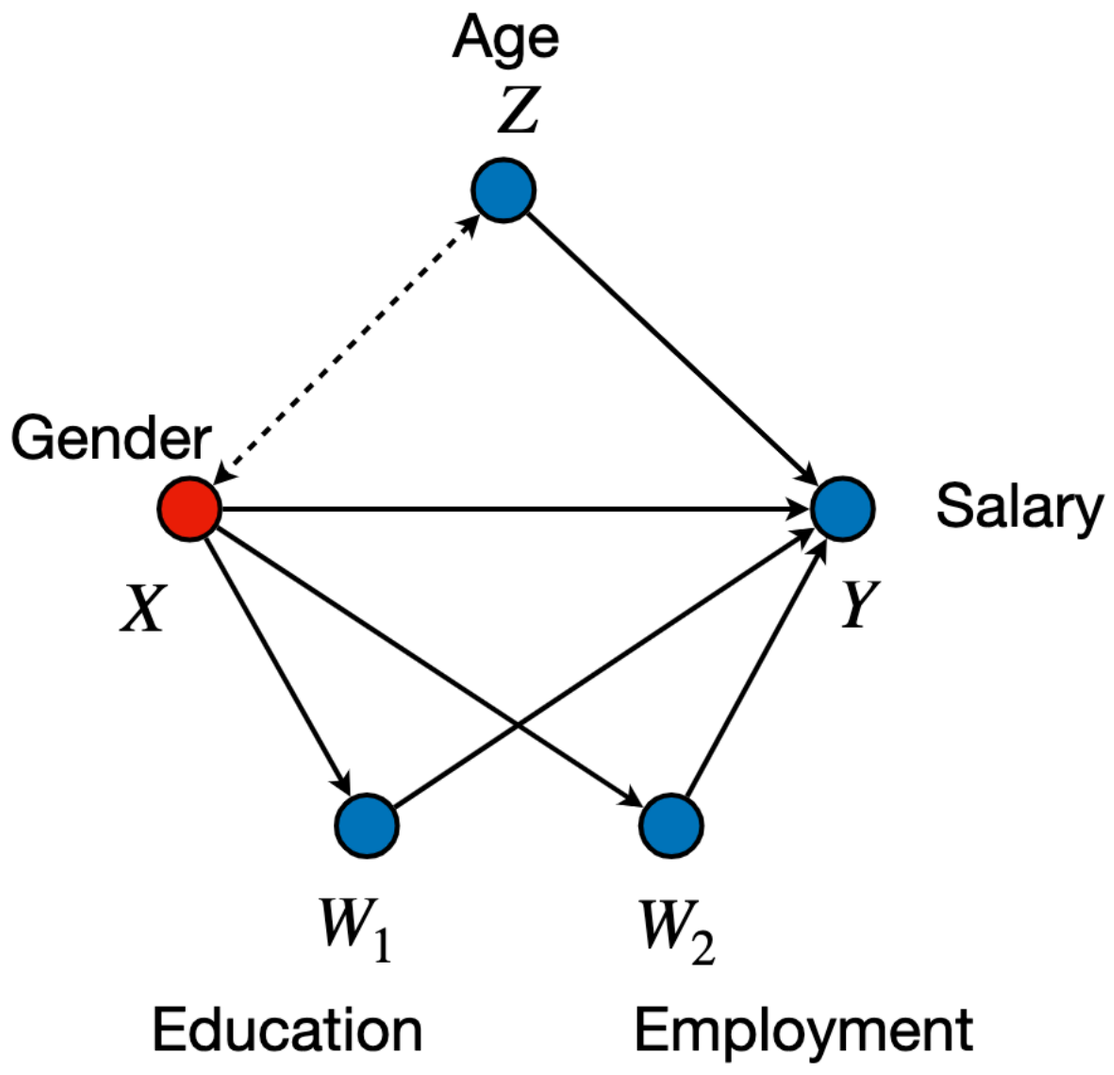



Do-calculus was extended to work with unobserved confounders.
Identifiability issues are harder in the presence of unobserved confounders.

# Graphical Discrimination Analysis

## UCI ADULT

The US census data records whether a person earns more than $50,000/year (Y). The census also records age (Z), gender (X = 0 for male, X = 1 for female), education level ($W_1$) and employment status ($W_2$ with 10 job types).



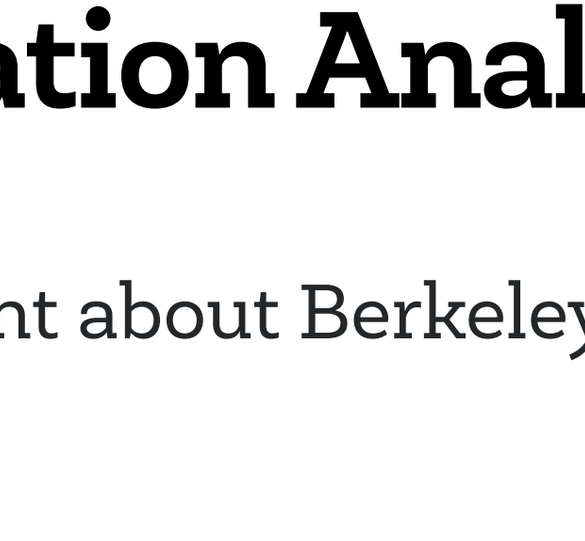

https://fairness.causalai.net/

# Graphical Discrimination Analysis

In causal language, Bickel's argument about Berkeley admissions had two components:

- There is **no direct effect** of sex A on the admissions decision Y that favors men.
- The **indirect effect** of A on Y, that is mediated by Z, should not be counted as evidence of discrimination.

**These are causal concepts that can be measured with data and a causal model!**




https://fairmlbook.org/

# Graphical Discrimination Analysis

## Direct Effects

To measure the direct effect of A on Y we need to disable all paths between A and Y except for the direct link. In our model, we can accomplish this by holding department choice Z constant and evaluating the conditional distribution of Y given A.

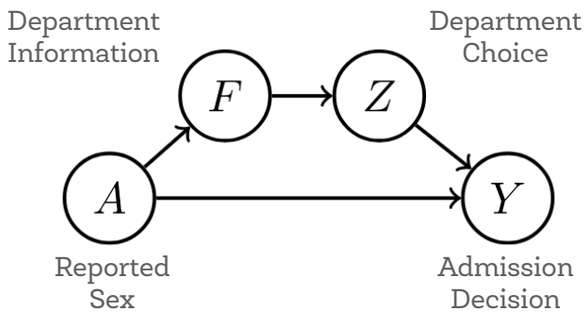
$$\Pr(Y \mid do(A) = a) = \sum_z \Pr(Z = z) \Pr(Y \mid A = a, Z = z)$$

This is the right answer to this Simpson's paradox!

# Graphical Discrimination Analysis

## Indirect Effects

The direct effect of a protected variable on a decision is a measure of discrimination on its own, but it cannot detect any form of *proxy* discrimination.




For example, the department may have advertised the program in a manner that strongly discouraged women from applying.

This indirect path encodes a pattern of discrimination.

We can think of the direct effect as whether or not the decision maker **explicitly** *uses* the attribute in its decision rule. Additionally, we have to carefully discuss what pathways we consider evidence for or against discrimination.

# Counterfactual discrimination analysis

Causal statements can be easily translated into **counterfactuals**:

**Causal Statement**:
*Was I not hired because I was black?*

**Counterfactual Statement**:
*Would I have been hired if I were non-black?*

Then, alternatively, we can state some definitions of fairness in terms of counterfactuals: **Individual Counterfactual Fairness**, **Counterfactual Parity**, **Conditional Counterfactual Parity**.

# Counterfactual discrimination analysis

For every individual $i$ we only see $Pr(Y|A = \texttt{black})$ or $Pr(Y|A = \texttt{non\_black})$ (not both!), but we can consider its counterfactual.

## Individual Counterfactual Fairness (ICF), for individual $i$

$$Pr(Y^*|A^* = \texttt{non\_black}) = Pr(Y|A = \texttt{black})$$

Would the hiring decision have been different if I were $A = \textit{non\_black}$ instead of $A = \textit{black}$?

## Counterfactual Parity (CP),

$$\mathbb{E}[Pr(Y^*|A^* = \texttt{non\_black})] = \mathbb{E}[Pr(Y^*|A^* = \texttt{black})]$$

Would the rates of hiring be different if everyone were *black*?

## Conditional Counterfactual Parity (CCP),

$$\mathbb{E}[Pr(Y^*|A^* = \texttt{non\_black}, X)] = \mathbb{E}[Pr(Y^*|A^* = \texttt{black}, X)]$$

Would the rates of hiring be different if I everyone were black, **conditioned on education**?