UNIVERSITAT DE BARCELONA

MSc in Fundamental Principles of Data Science

Ethical Data Science

Bias and Discrimination II: Causality

Jordi Vitrià

2020-2021

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

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

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, again, a significant difference.

Such reversals are sometimes 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.

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

Therefore, one explanation for the data we see is that **women chose to apply to more competitive departments, hence getting rejected 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

It is difficult to debate this conclusion on the basis of the available data alone. The question of discrimination, however, is far from resolved.

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

There are several possible reasons.

- Perhaps less competitive departments, such as engineering schools, 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

It is difficult to debate this conclusion on the basis of the available data [illegible]on, however, is far from resolv[illegible]

We can ask why [illegible] departments in the first place.

There are sever[illegible]

- Perhaps less c[illegible]ineering schools, were [illegible]his may have been a genera[illegible]university.
- Perhaps some[illegible]or treatment of women that w[illegible]
- Perhaps the de[illegible]a manner that discouraged women from applying.

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.

Causality

"We can distinguish two kinds of intentionality. Primary intentionality is the repetition of causes that worked in the past. This is the intentionality of adaptation by natural selection and of conditioned reflexes. Past effects are anticipated to occur again. Secondary intentionality is choice of action after simulation of possible choices and their effects. Simulated effects are anticipated to occur when the action is performed. Secondary intentionality requires imagination, an ability to "hold in mind" and evaluate virtual outcomes. Primary intentionality is "primary" in the sense that anticipation evolved before imagination."

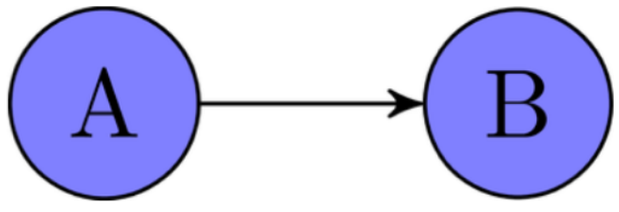
Fragment from: Haig, David; Dennett, Daniel C.;. "Selfish Genes, Social Selves, and the Meanings of Life".

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 ***correlated*** 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 3 different features in this dataset, (X, Y, Z).

Which can of questions can we answer from this dataset?

Association vs. Causality

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 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 Inference is using data to predict certain features of the world <u>if the world had been different</u>. 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, analytics, and expert **causal knowledge**.

> For example, estimate the *income* that would have been observed if **<u>all individuals</u>** had *race=1* vs. if they had *race=2* or *race= 3*.

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

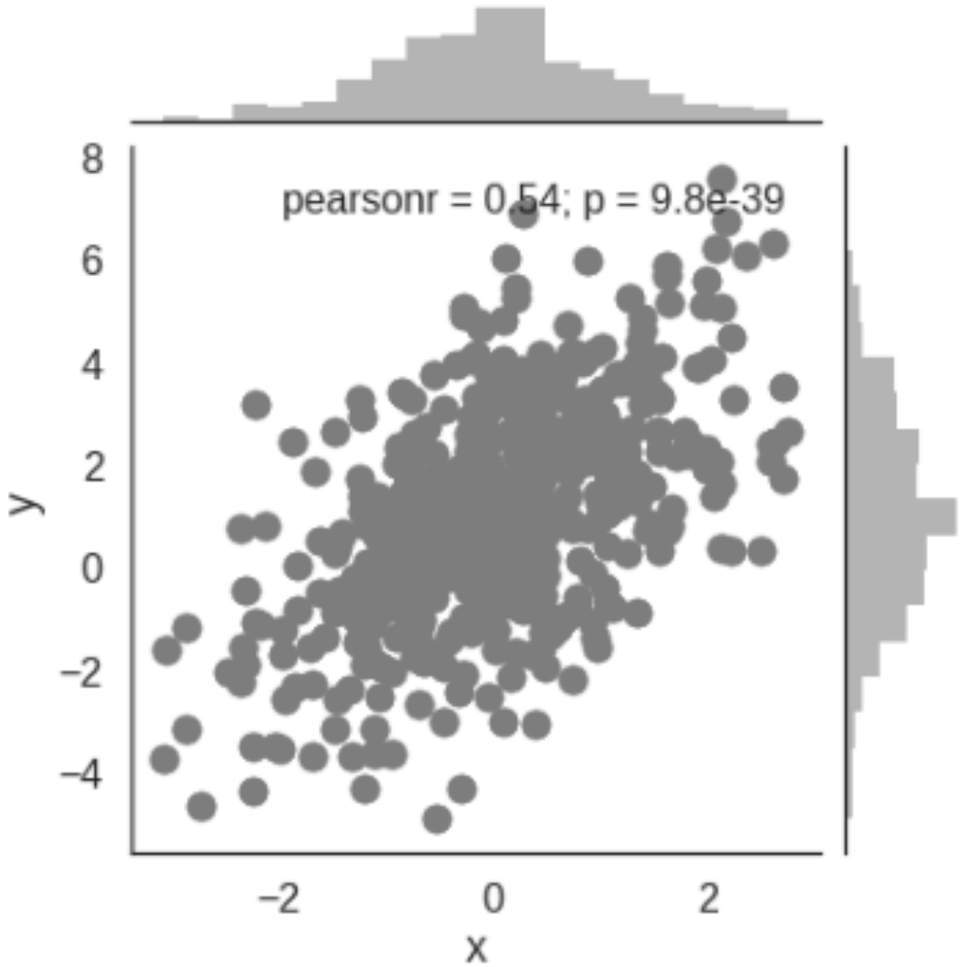
  For example, estimate the income that would have been observed if **all individuals** had race=1 vs. if they had race=2 or race= 3.

- **interventional**, based on $p(Y \mid do(X)) \neq p(Y \mid X)$:
  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)

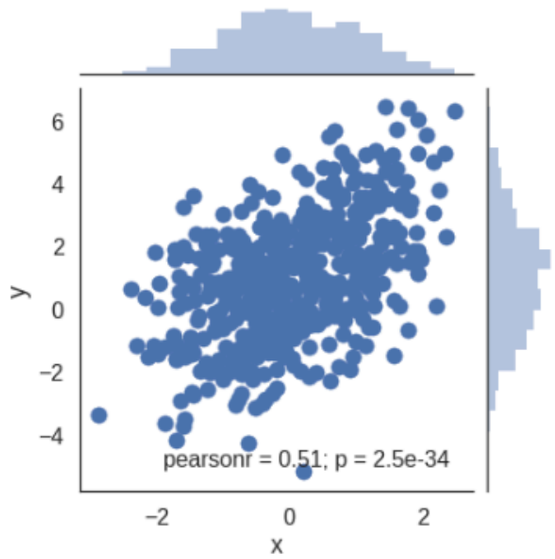
Association vs. Causality

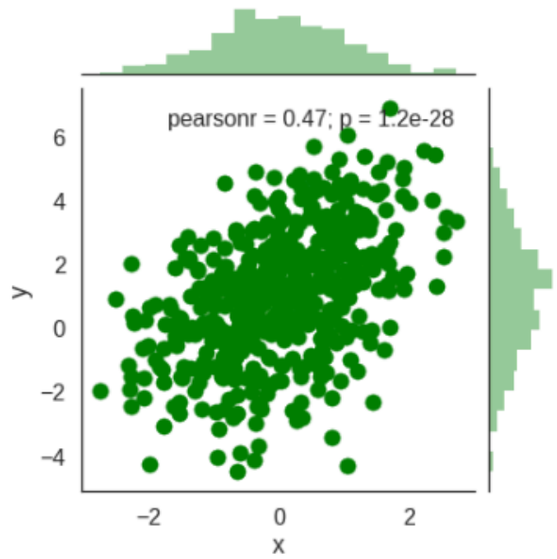
Given $p(X, Y)$, **there are several generative models** that are compatible with $p(X, Y)$:

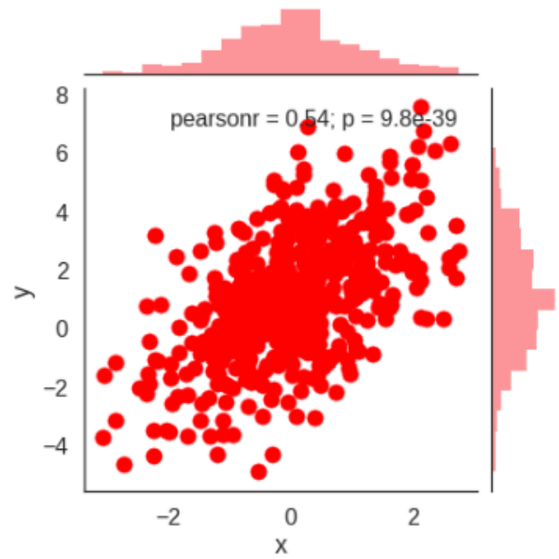
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
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.

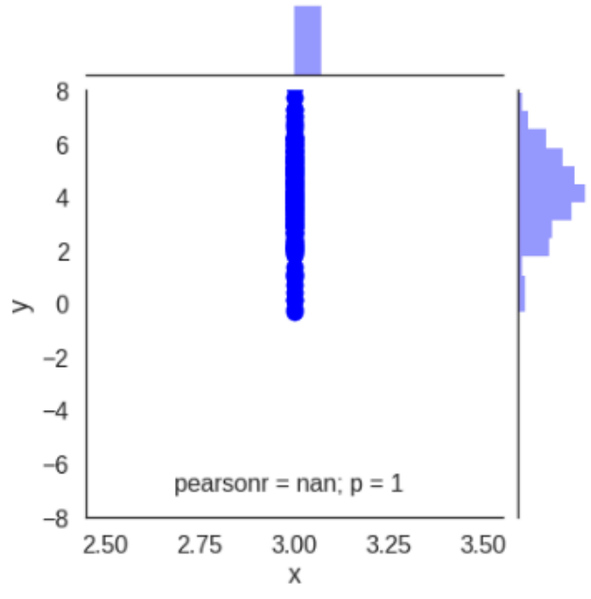
Association vs. Causality

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

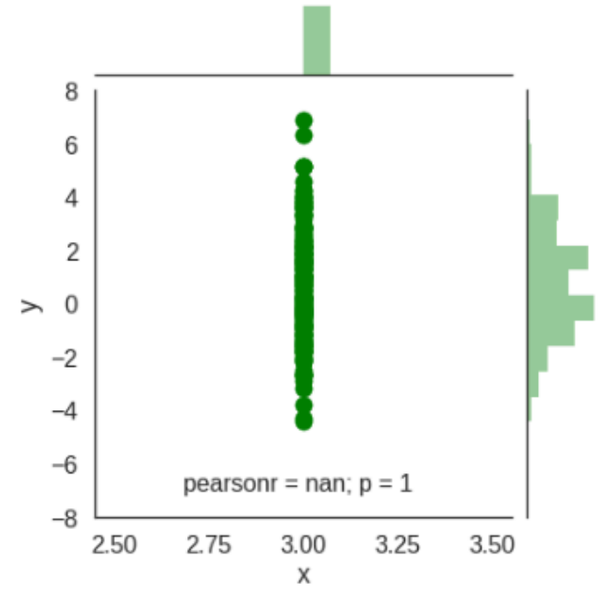
```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```

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

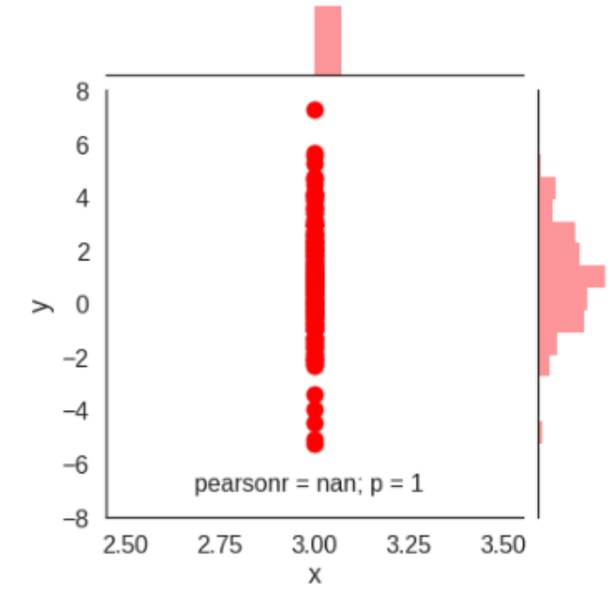
```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```



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



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



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

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

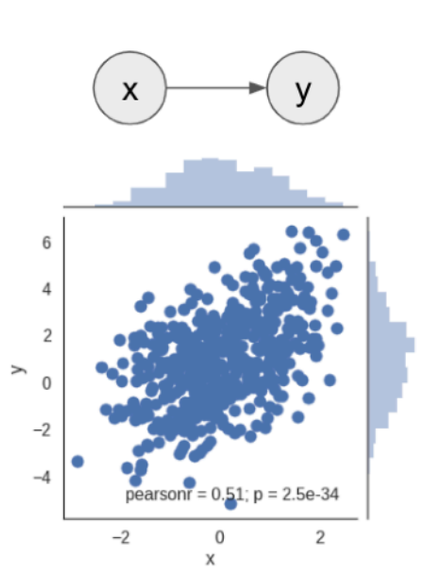
An intervention can be understood as a **modification of the generative model of the data, producing a different probability distribution** $p(do(X), Y, Z)$.

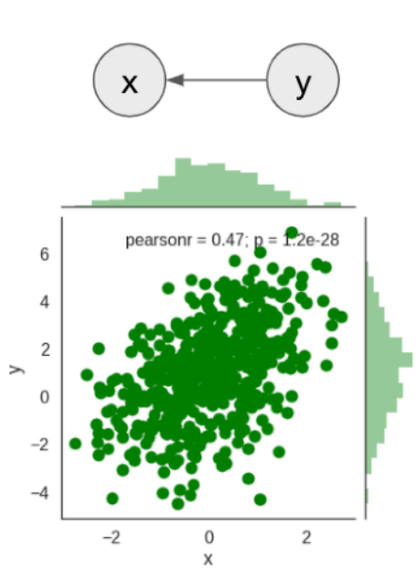
The resulting distribution depends on the original model:

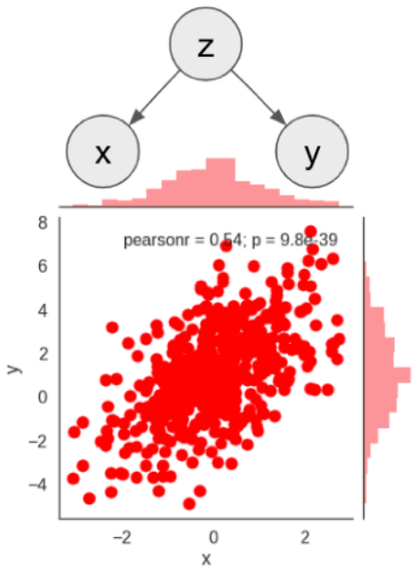
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
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**.

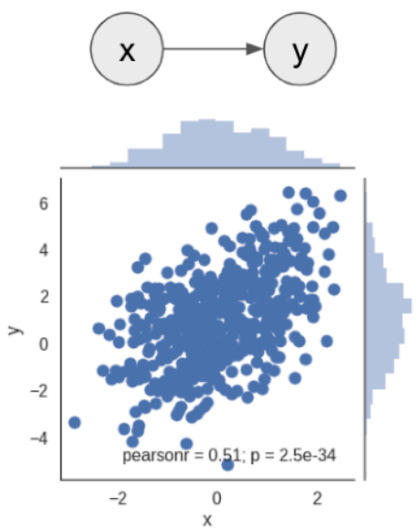
Association vs. Causality

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:

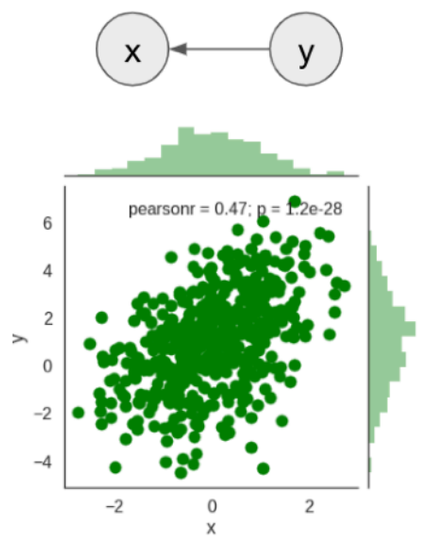
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

X is the cause of Y



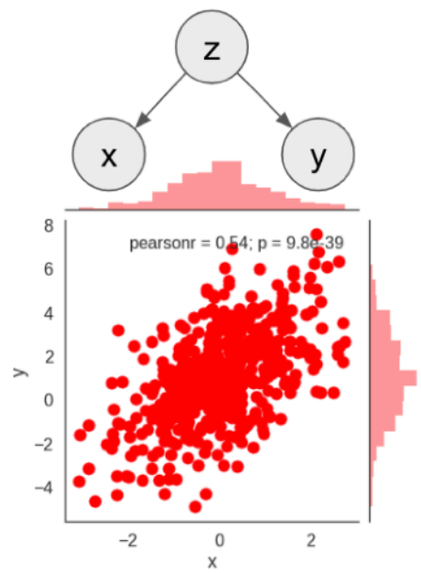

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

Y is the cause of X




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

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



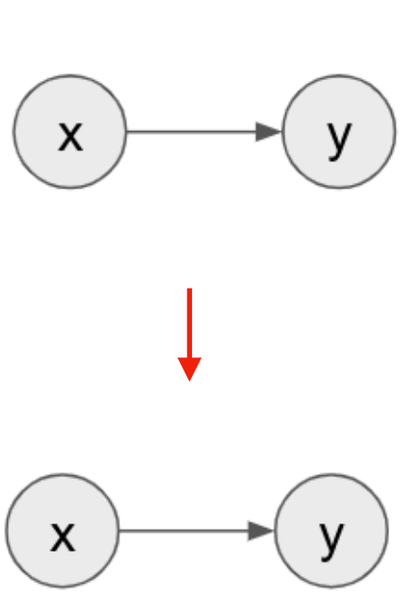
https://www.inference.vc/untitled/

Directed Acyclic Graphs (DAG).

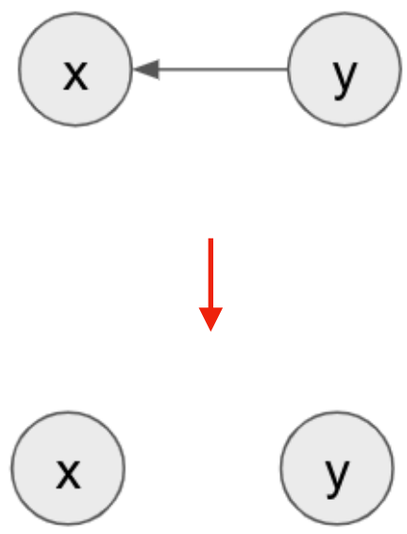
No assumptions about the exact form of the functional relationships are needed. The only requirement is that causal relationships are **acyclic**.

Association vs. Causality

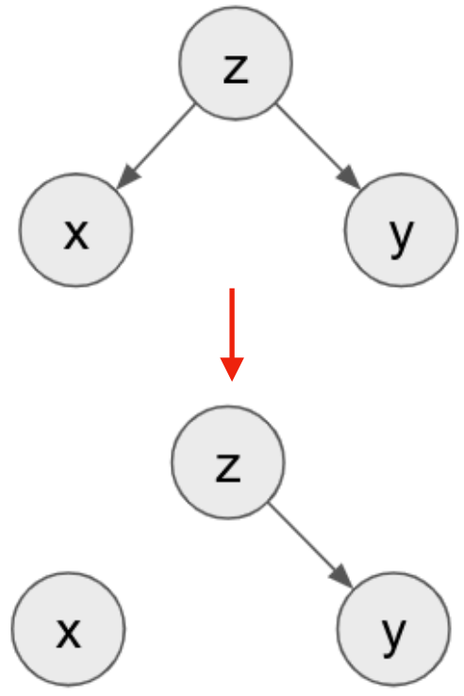
Graphically, to simulate the effect of an intervention, you **mutilate** the graph by **removing all edges that point into the variable on which the intervention is applied**, in this case X.



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



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

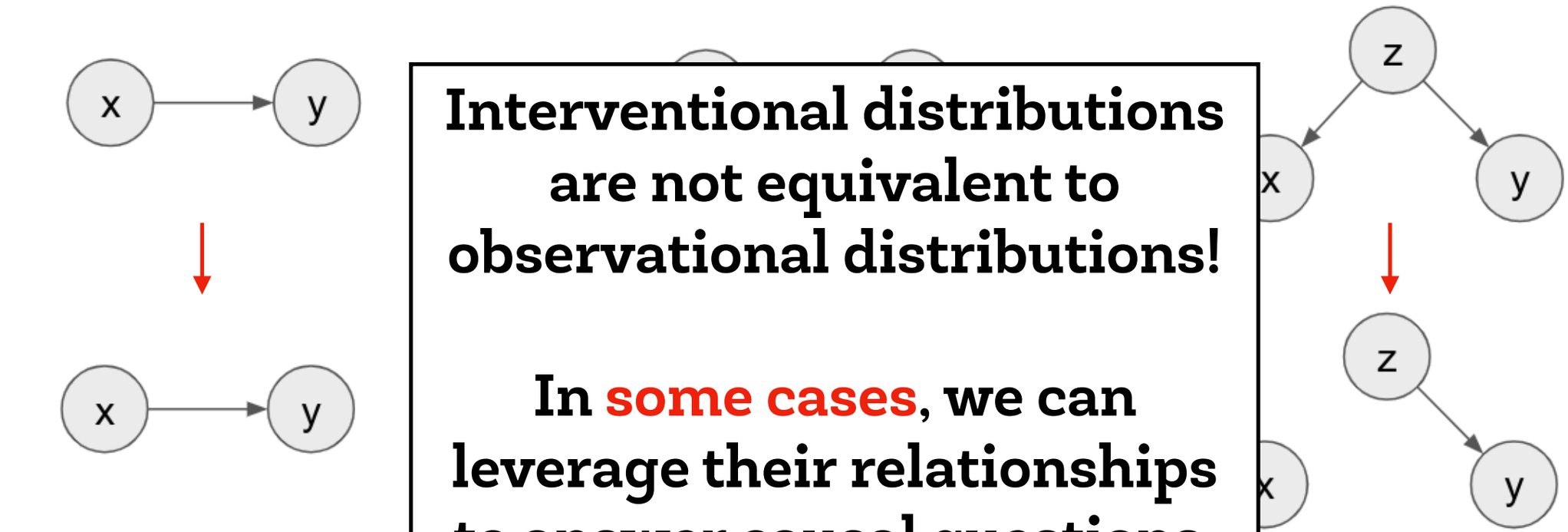


$p(Y \mid do(X) = p(Y)$

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!

In some cases, we can leverage their relationships to answer causal questions.

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

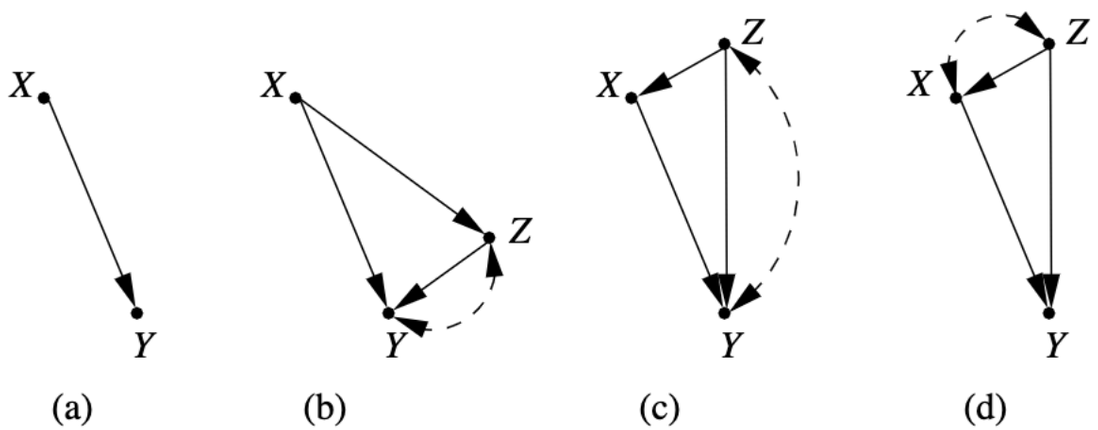
Just by only looking at the causal diagram, we are now able to predict how the scripts are going to behave under the intervention $X = 3$.

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

Causal Inference and do-calculus

There are two ways to assess the existence of a **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 **ethical** or **economical** reasons)

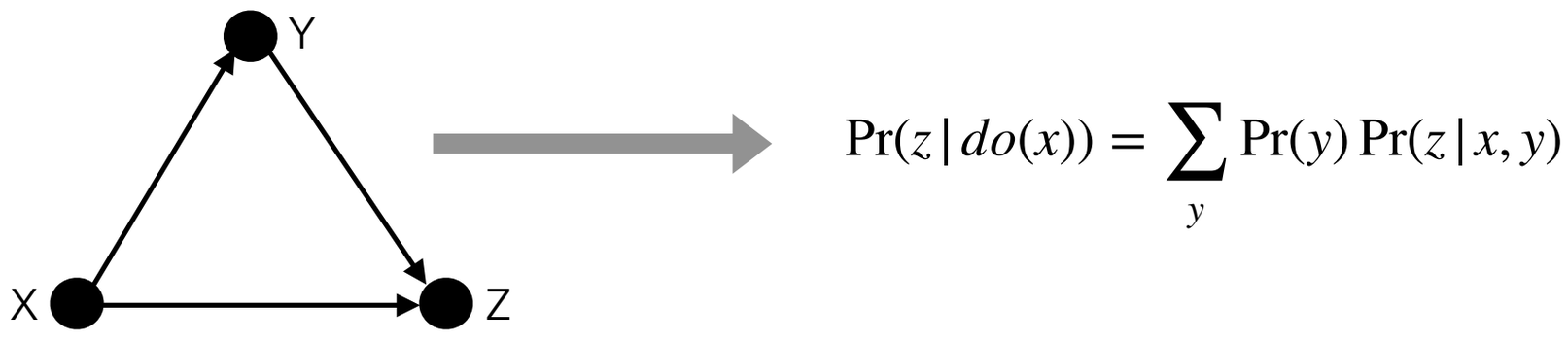
Causal Inference and do-calculus

There are two ways to assess the existence of a **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 $p(X, Y, Z)$

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.

Example:



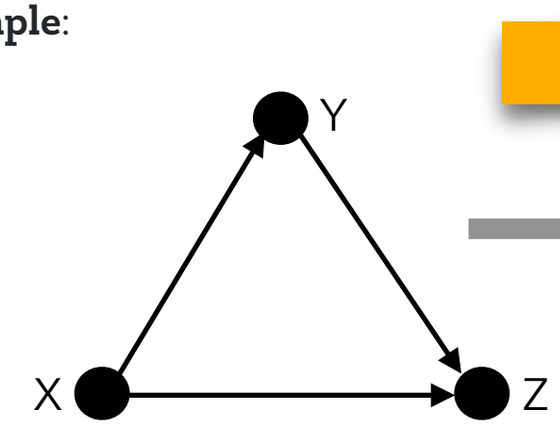
Causal Inference and do-calculus

There are two ways to assess the existence of a **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 $p(X, Y, Z)$

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.

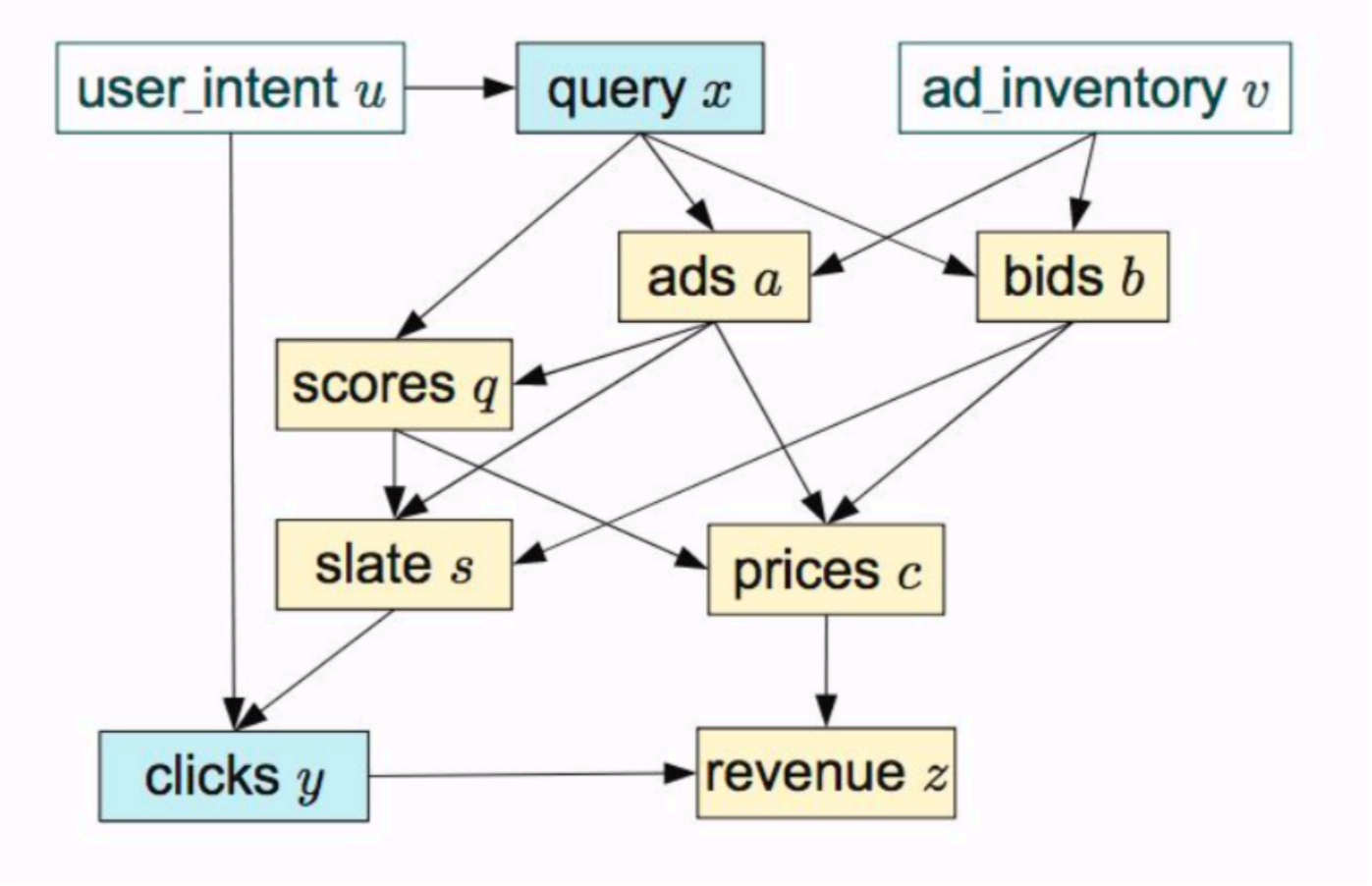
Example:



Only depends on the causal graph!

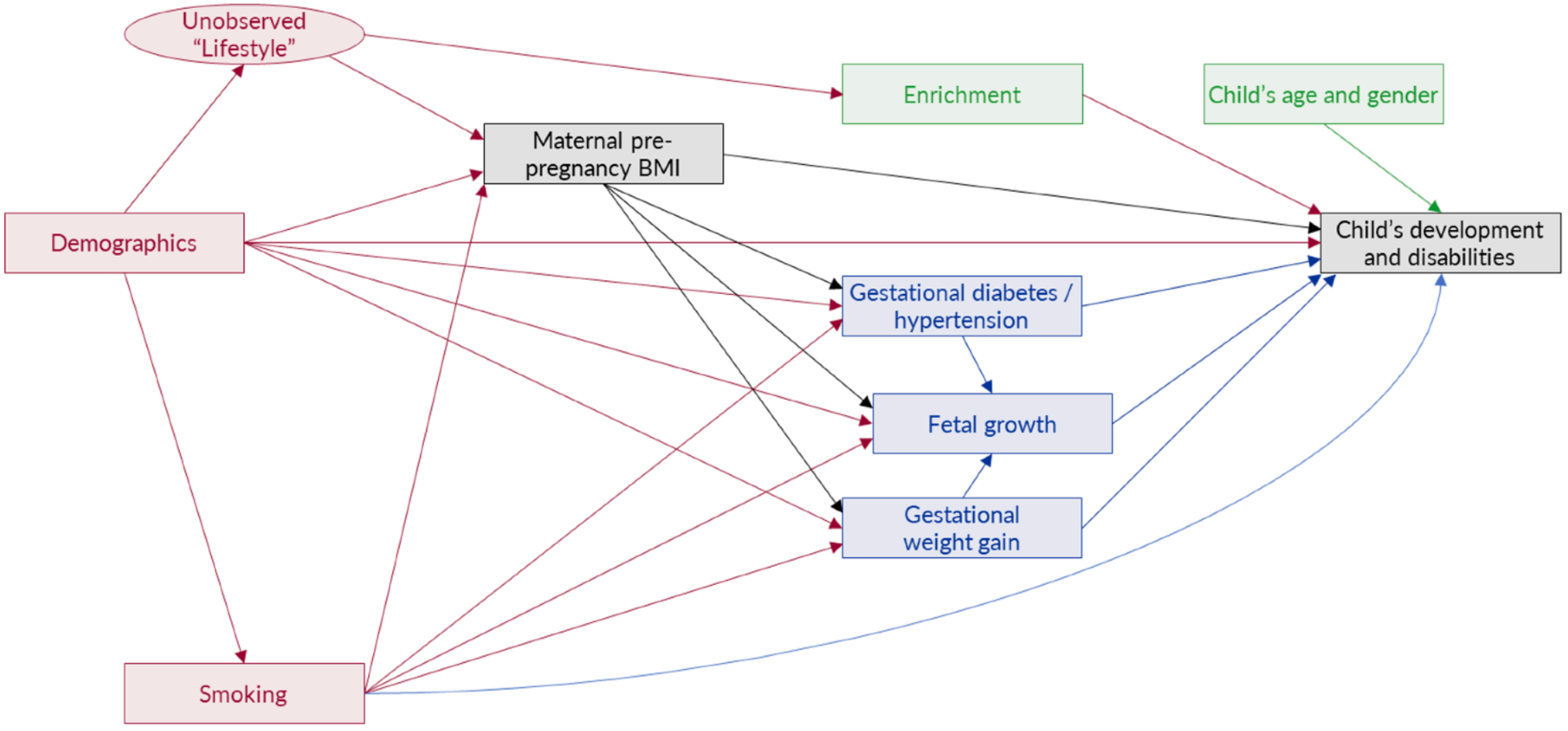
$$\Pr(z \mid do(x)) = \sum_{y} \Pr(y) \Pr(z \mid x, y)$$

Causal Inference and do-calculus



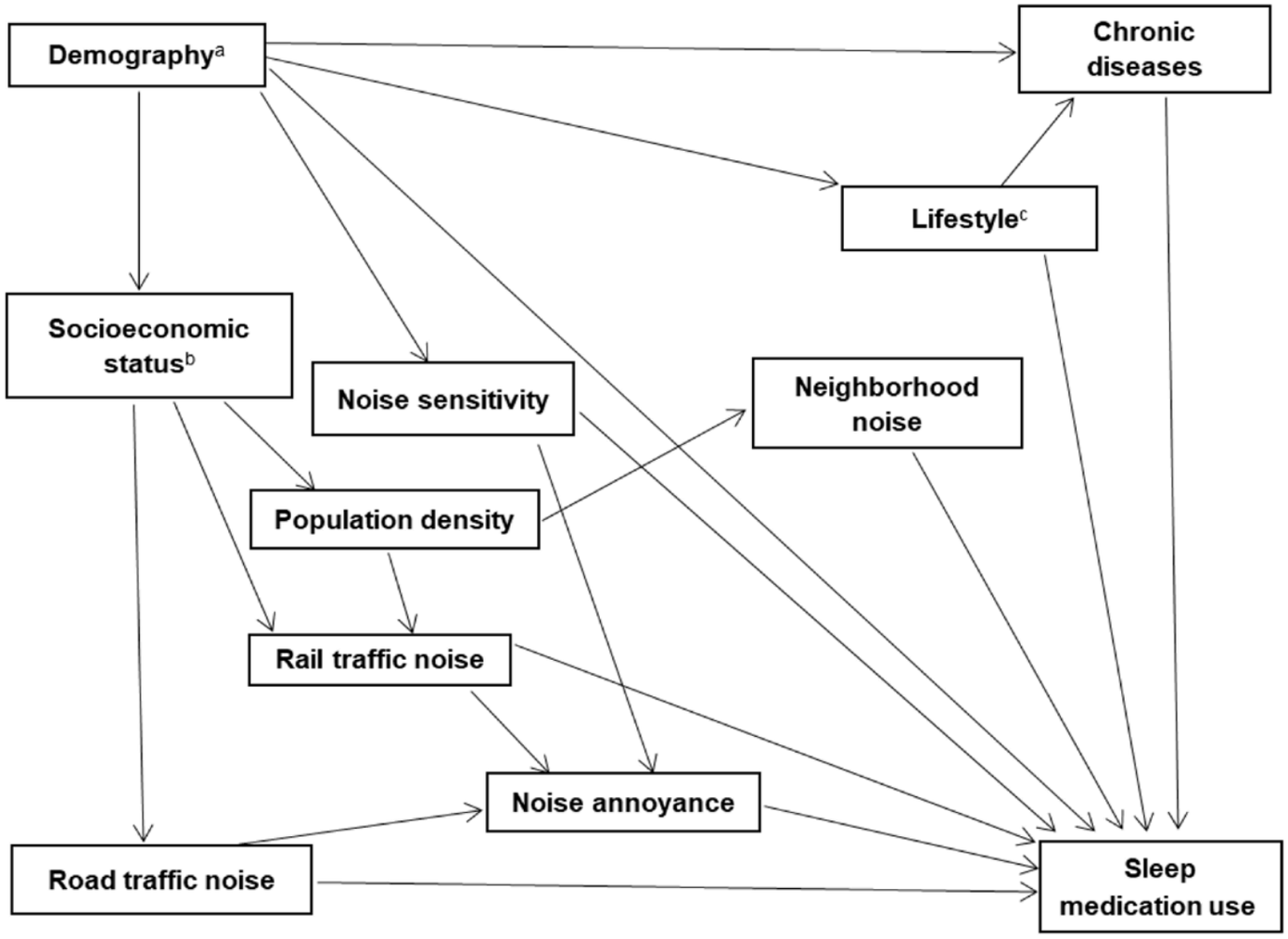
Bottou, Léon, et al. "Counterfactual reasoning and learning systems: the example of computational advertising." *The Journal of Machine Learning Research* 14.1 (2013): 3207-3260.

Causal Inference and do-calculus



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 and do-calculus



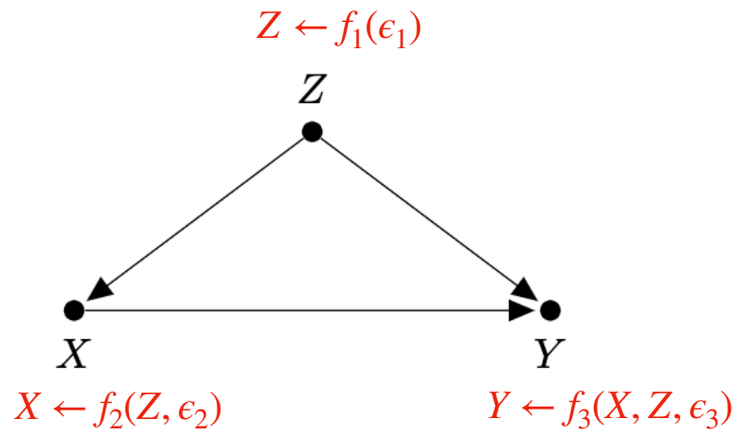
REPRODUCED UNDER CC-BY 4.0 LICENSE FROM: Evandt J, Oftedal B, Krog NH, et al. Road traffic noise and registry based use of sleep medication. *Environ Health.* 2017;16(1):110. DOI: 10.1186/s12940-017-0330-5 (Figure S1). Freely available at: https://www.doi.org/10.1186/s12940-017-0330-5

Causal Inference and SCM

The **causal diagram** can be seen as a representation of an underlying **structural causal model:**

Directed Acyclic Graphs (DAG):
No assumptions about the exact form of the functional relationships, as well as the distribution of background factors, are needed. The only requirement is that causal relationships are acyclic.

	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



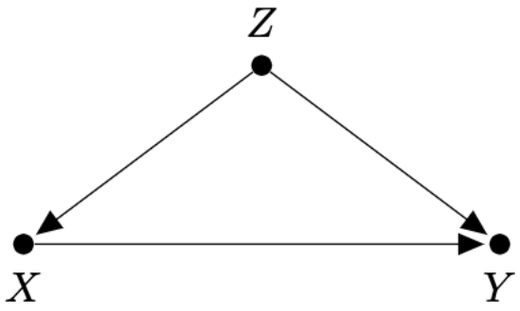
ϵ_i are **exogenous background factors** represented by an arbitrary noise distribution.

They exert an influence on the endogenous variables in the model. Because they are unobserved from the standpoint of the analyst, this renders the model stochastic with a probability distribution over the set of **endogenous variables**.

Causal Inference and SCM

We can estimate f_i from data by using predictive methods.

Actions can now be defined as interventions on variables in the model.



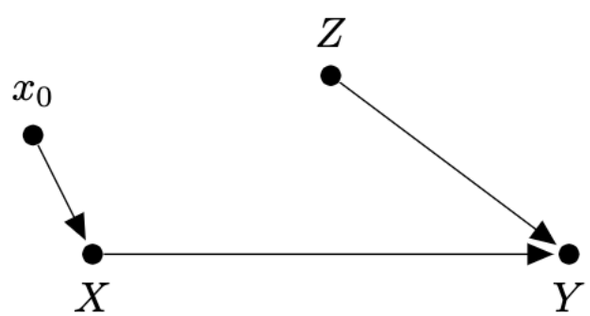
$$Z \leftarrow f_1(\epsilon_1)$$
$$X \leftarrow f_2(Z, \epsilon_2)$$
$$Y \leftarrow f_3(X, Z, \epsilon_3)$$

Structural Causal 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



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

For example, compute **causal effects**:

$$p(Y \mid do(X = x_1) - p(Y \mid do(X = x_0)$$

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€.

Counterfactuals

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.

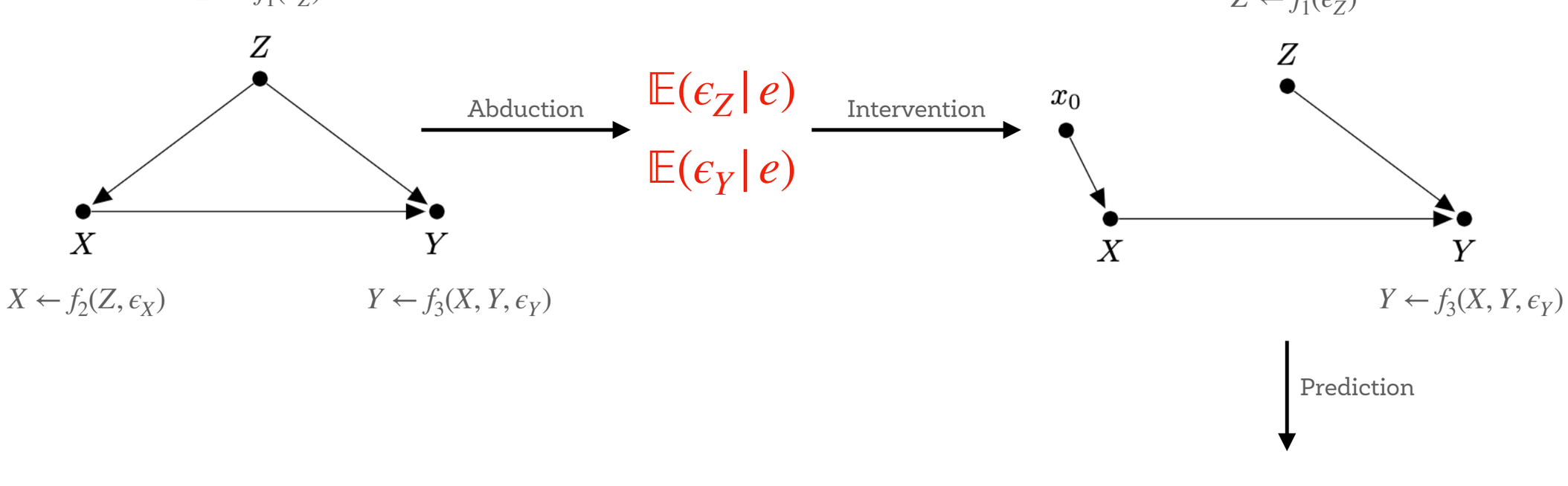
Identifiable counterfactuals can be computed as a three-step process:

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

Counterfactuals

e

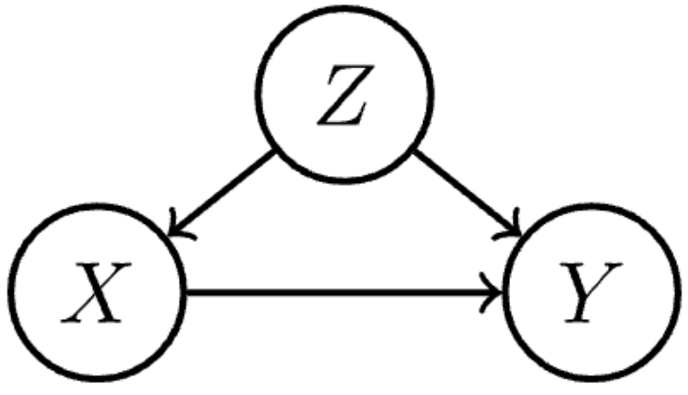
R1001	M	1	70	50	1	12	1.73
R1002	M	2	72	100	2	20	4.53
R1003	F	1	55	250	1	16	2.99
R1004	M	2	65	20	2	16	1.13
R1005	F	1	60	10	3	12	3.81
R1006	M	1	68	30	1	9	4.76
R1007	F	5	66	25	2	21	2.01



Causal Graph Structures

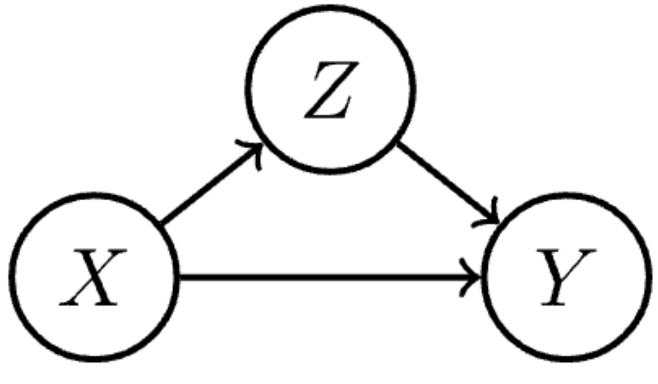
A **fork** is a node Z in a graph that has outgoing edges to two other variables X and Y.

Put differently, the node Z is a common cause of X and Y.



Z has a **confounding** effect: There is a positive association between X and Y (not due to X→Y) because of Z.

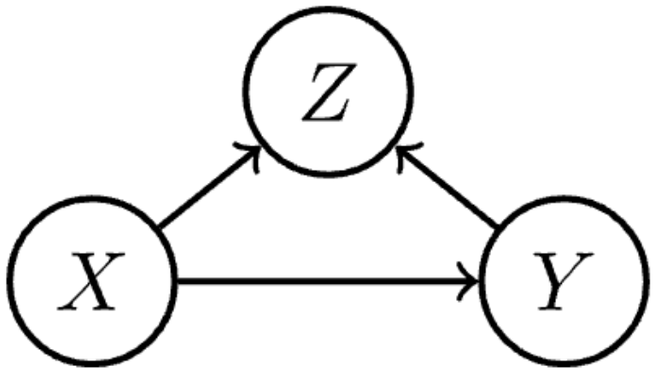
Causal Graph Structures



In this case, the path X→Z→Y contributes to the total effect of X on Y. It's a causal path and thus one of the ways in which X causally influences Y. That's why Z is not a confounder. We call Z a **mediator** instead.

The notion of a mediator is particularly relevant to the topic of **discrimination analysis**.

Causal Graph Structures



This is a **collider**. Colliders aren't confounders. In fact, in the above graph, X and Y are unconfounded, meaning that we can replace do-statements by conditional probabilities.

However, something interesting happens **when we condition** on a collider. The conditioning step can create association between X and Y, a phenomenon called *explaining away*.

Causal Graph Structures

When we speak of the **causal effect** **of a variable** X **on another variable** Y we refer to **all the ways** in which setting X to any possible value x affects the distribution of Y.

Often we think of X as a binary treatment variable and are interested in a quantity such as:

$$\mathbb{E}_{do(X=1)}[Y] - \mathbb{E}_{do(X=0)}[Y]$$

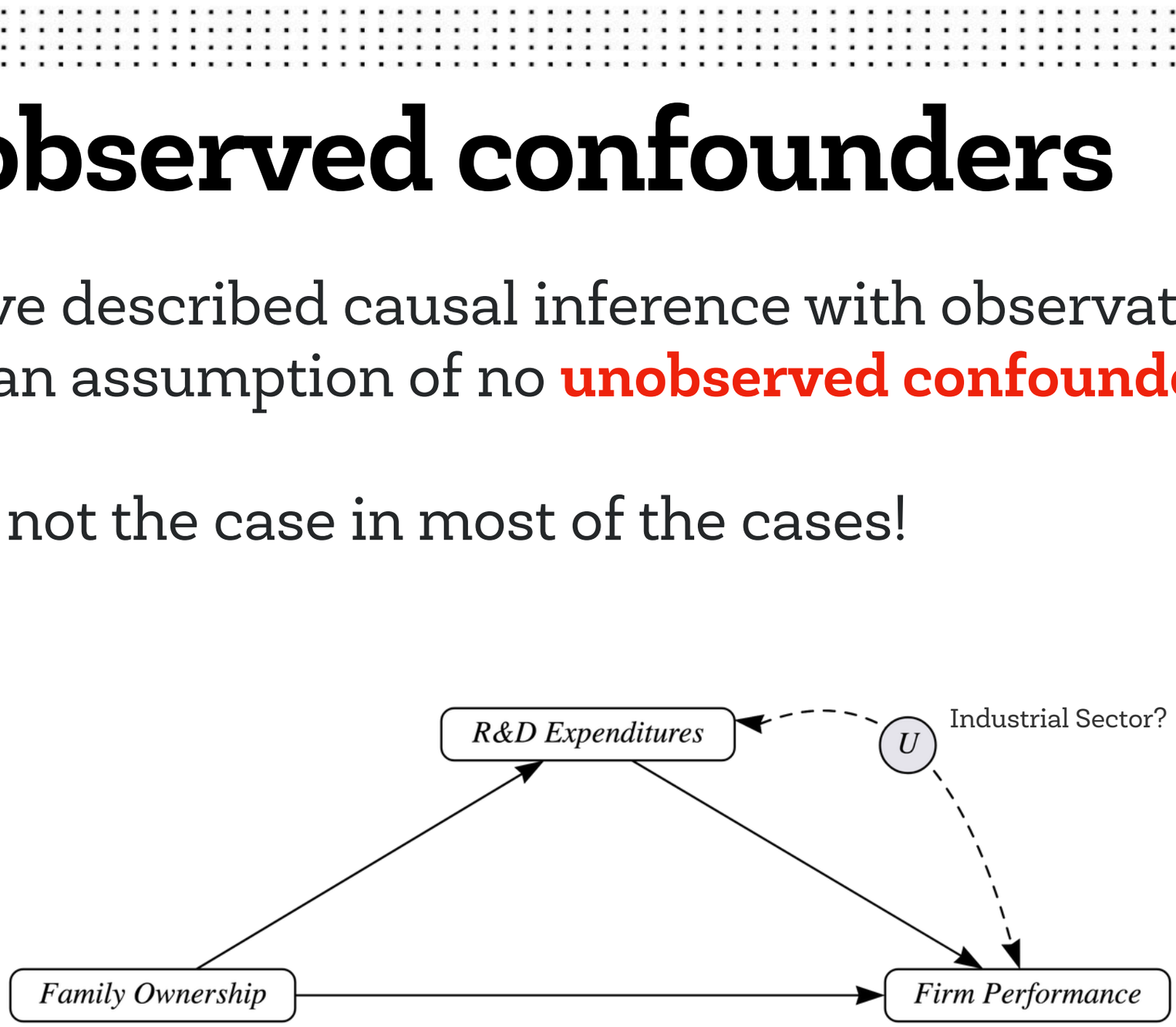
Causal effects are **population quantities**. They refer to effects averaged over the whole population. Often the effect of treatment varies greatly from one individual or group of individuals to another. Such treatment effects are called *heterogeneous*.

This quantity is called the **average treatment effect**.

Unobserved confounders

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!



https://p-hunermund.com/2018/08/27/why-you-shouldnt-control-for-post-treatment-variables-in-your-regression/

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

Causal Discrimination Analysis

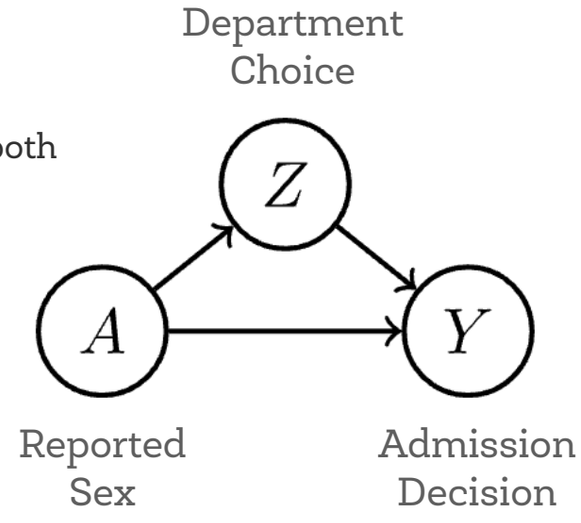
Graphical Discrimination Analysis

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 $A \rightarrow Y$ and $Z \rightarrow Y$, because both features are available to the institution when making the admissions decision.



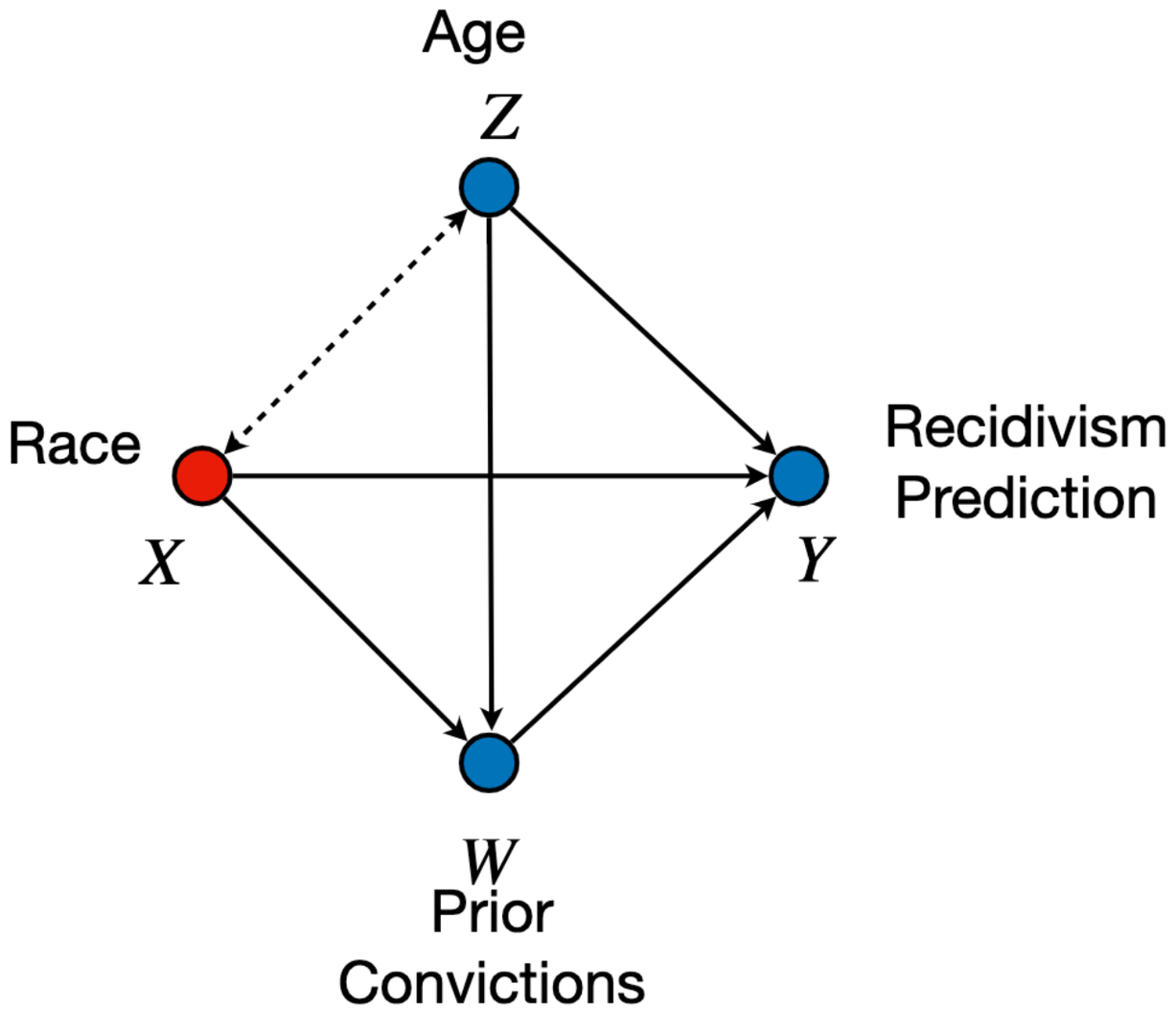
2. A and Z are not statistically independent. We can see from the table that several departments have a statistically significant gender bias among applicants. This means we need to include either the arrow $A \rightarrow Z$ or $Z \rightarrow A$. Which one?

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

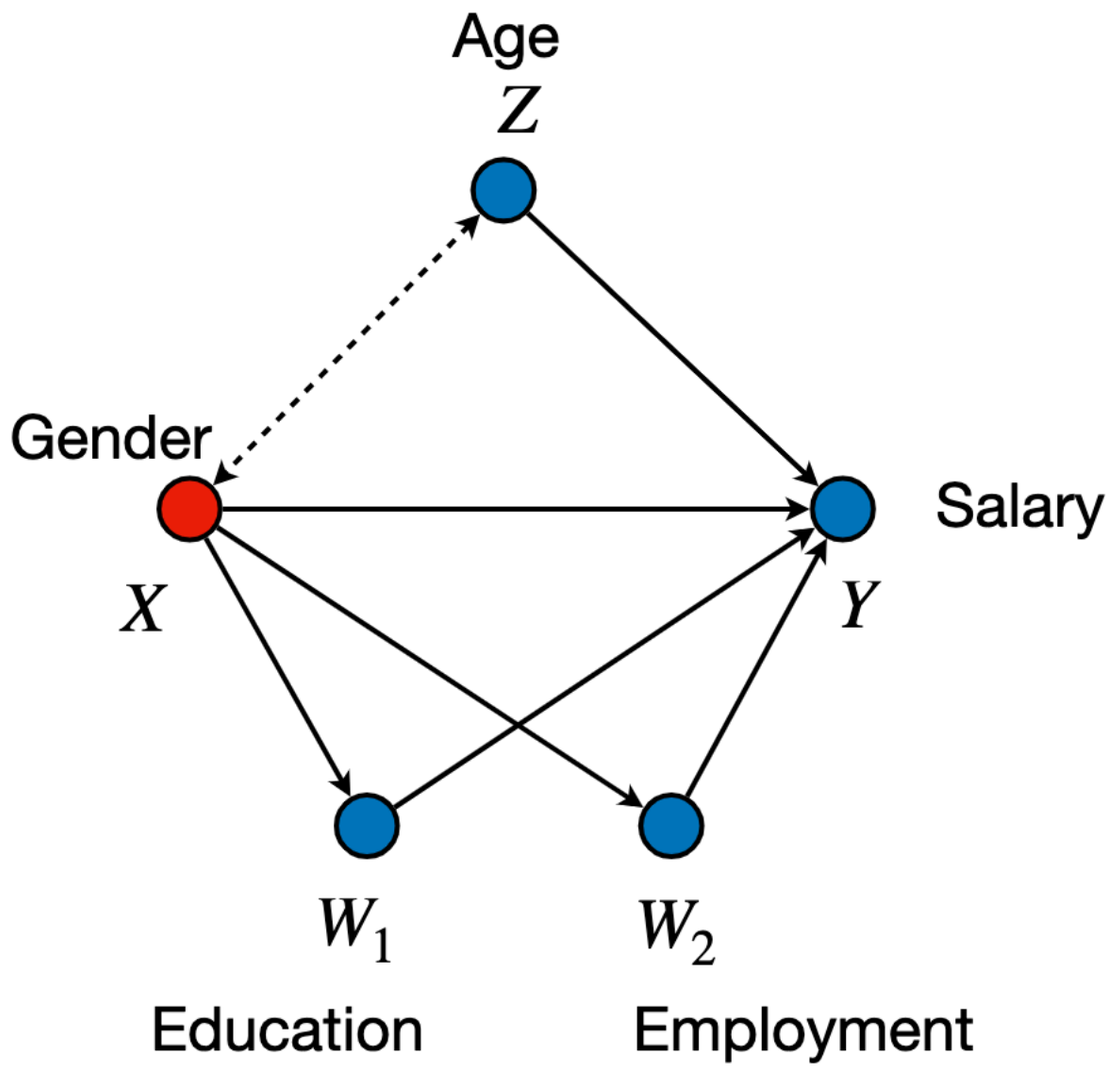


https://fairness.causalai.net/

Graphical Discrimination Analysis

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).



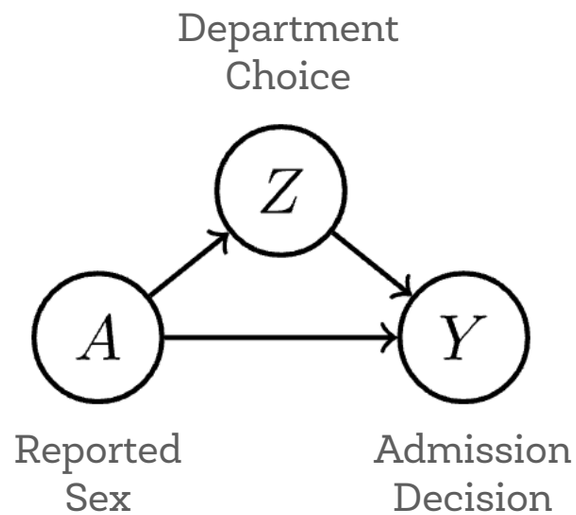
https://fairness.causalai.net/

Graphical Discrimination Analysis

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

- There appears to be **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 department choice 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.

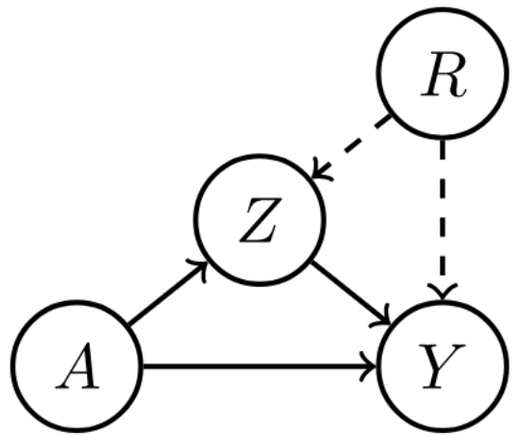
$$\Pr(Y \mid do(A) = a) = \sum_z \Pr(Z = z) \Pr(Y \mid A = a, Z = z)$$

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

Graphical Discrimination Analysis

Direct Effects

A problem would arise if department choice and admissions outcome were confounded by another variable, such as, state of residence R.

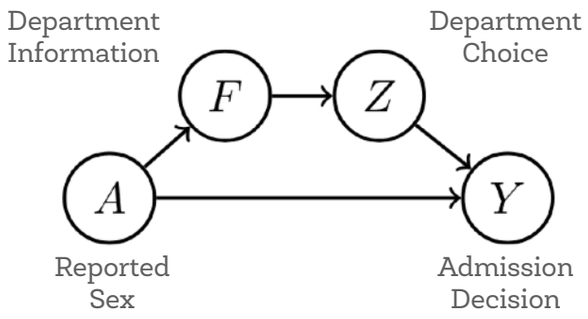


Department choice Z is now a collider between A and R. Conditioning on a collider opens the backdoor path $A \rightarrow Z \leftarrow R \rightarrow Y$. **In this graph, conditioning on department choice does *not* give us the desired direct effect**.

Graphical Discrimination Analysis

Indirect Effects

The direct effect of a protected variable on a decision is a poor measure of discrimination on its own as 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.

Graphical Discrimination Analysis

Measuring causal effects is informative about the role of protected features in algorithmic decision making.

Is Berkeley admission process using a protected attribute?

But, what about the causal role of protected features in specific decisions/cases? What about fairness?

Is Berkeley admission process a fair system?
Was its decision about me a fair decision?

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, 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),

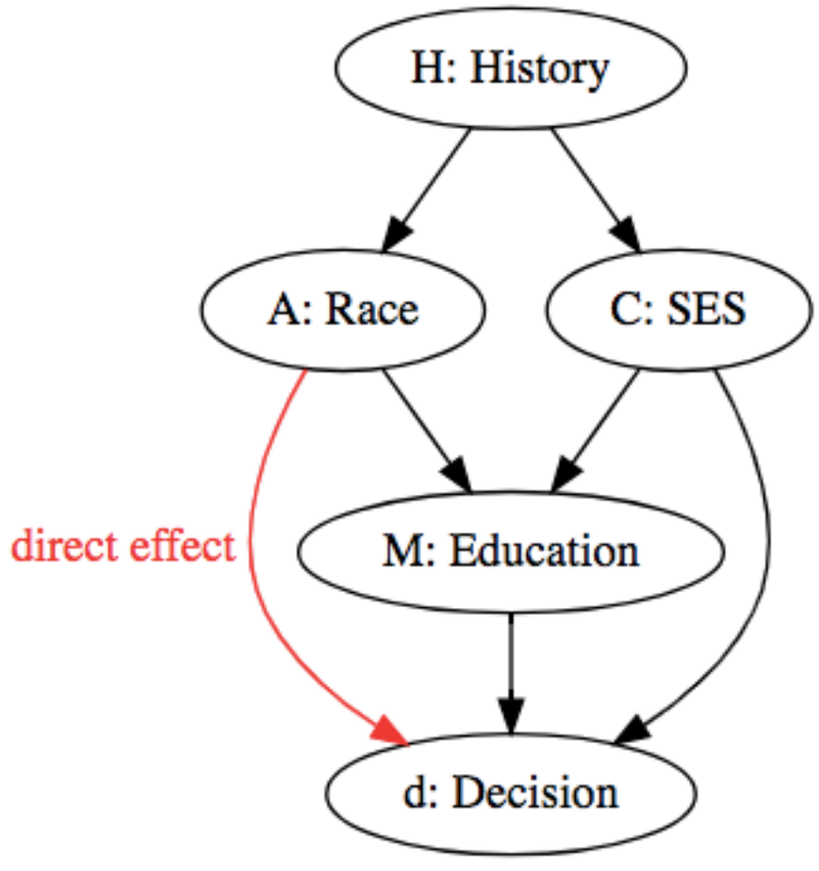
$$\mathbb{E}[Pr(Y^*|A^* = \texttt{non\_black}, X)] = \mathbb{E}[Pr(Y^*|A^* = \texttt{black}, X)]$$

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

These causal queries measure the **total effect** of race on the decision.

Counterfactual discrimination analysis

We can measure **non direct** causal pathways effects by defining different counterfactuals.



Race and *SES* (Socio-economical status at birth) affect the decision through mediating variables such as ***education***

https://shiraamitchell.github.io/fairness/

Counterfactual discrimination analysis

Let $Pr(Y|A = \texttt{non\_black}, \texttt{M}(\texttt{black}))$ be the decision if the applicant had not been black, but education remained the same.

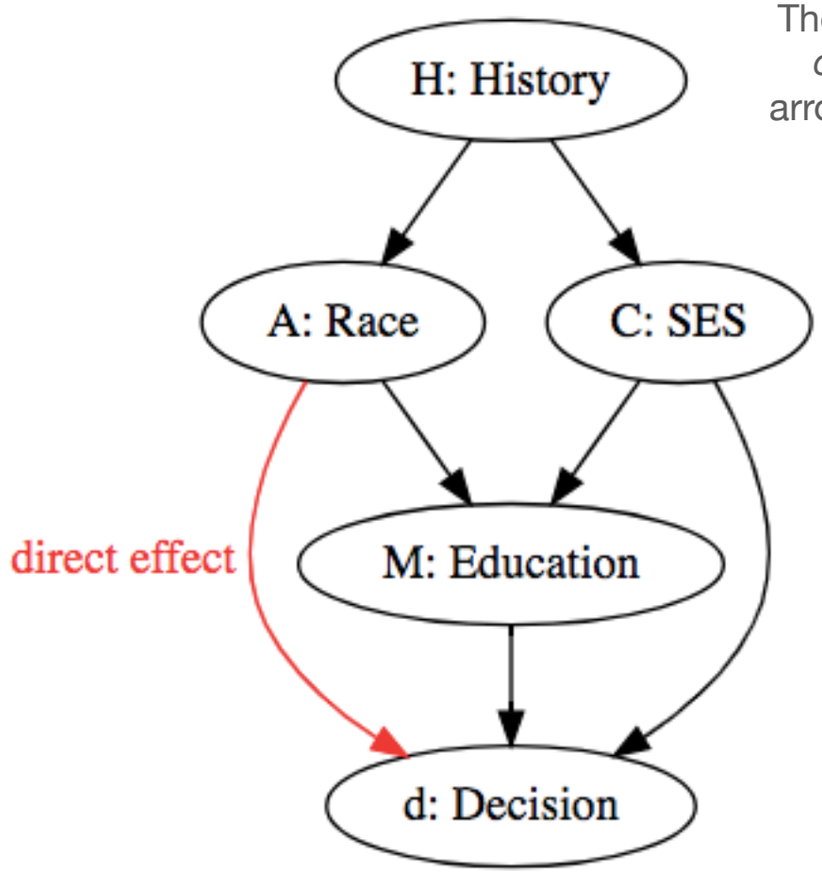
No Direct Effect Fairness

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

This counterfactual assumes that there is no direct effect of race on hiring, but race is allowed to affect decision through education.

Counterfactual discrimination analysis

WARNING: Causal paths do not exhaust possible explanations when fairness criteria are not met. **Back-door paths** can explain it!



The path between *race* and the *decision* that starts with an arrow going **into** race is called a **backdoor-path**.

A person's race does not cause to be born in a low income family, **but expect an association between *race* and *decision* that is not causal.**

If we have an intuition that such a path are unfair, we would requiere additional fairness criteria.