Causal Inference

Mariyam Khan

Department of Informatics
University of Bergen

November 16, 2022

Summary

Part 1
Motivating Example: Simpson's Paradox

Motivating Example: Simpson's Paradox

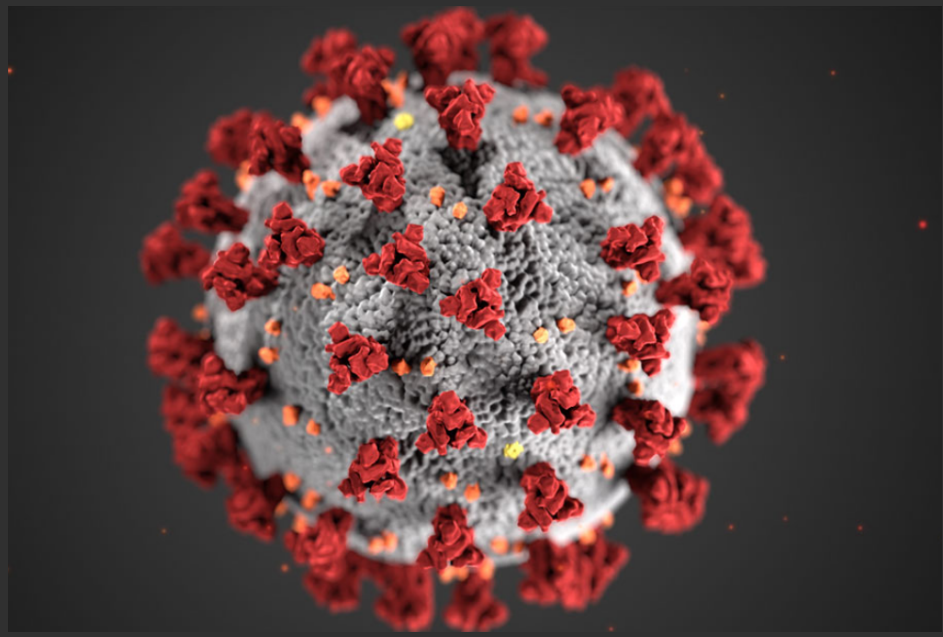
"What treatment should you choose?"

New Disease: COVID-27

Treatment T: A (0) and B (1)

Condition C: mild (0) or severe (1)

Outcome Y: alive (0) or dead (1)



[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Motivating Example: Simpson's Paradox

Simpson's Paradox: Mortality rate table

Treatment	Total
A	16% (240/1500)
B	19% (105/550)

$$\mathbb{E}(Y|T)$$

[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Motivating Example: Simpson's Paradox

Simpson's Paradox: Mortality rate table

		Condition	
Treatment	**Total**	**Mild**	**Severe**
A	16% (240/1500)	15% (210/1400)	30% (30/100)
B	19% (105/550)	10% (5/50)	20% (100/500)

$$\mathbb{E}(Y|T) \qquad \mathbb{E}(Y|T, C = 0) \qquad \mathbb{E}(Y|T, C = 1)$$

[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Motivating Example: Simpson's Paradox

Simpson's Paradox: Mortality rate table

		Condition	
T	**Total**	**Mild**	**Severe**
A	16% (240/1500)	15% (210/1400)	30% (30/100)
B	19% (105/550)	10% (5/50)	20% (100/500)
	$\mathbb{E}(Y\|T)$	$\mathbb{E}(Y\|T,C{=}0)$	$\mathbb{E}(Y\|T,C{=}1)$

$\frac{1400}{1500}(0.15)+\frac{100}{1500}(0.3)=0.16$

$\frac{50}{550}(0.1)+\frac{500}{550}(0.2)=0.19$

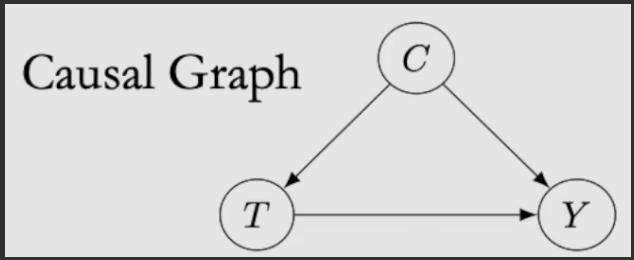
Which treatment would you choose?

[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Motivating Example: Simpson's Paradox

Simpson's Paradox: Scenario 1 (treatment B)

T	Total	Mild	Severe
A	16% (240/1500)	15% (210/1400)	30% (30/100)
B	19% (105/550)	10% (5/50)	20% (100/500)



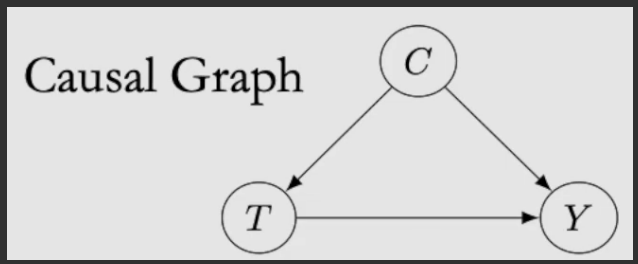
Treatment B is better than Treatment A in this scenario.

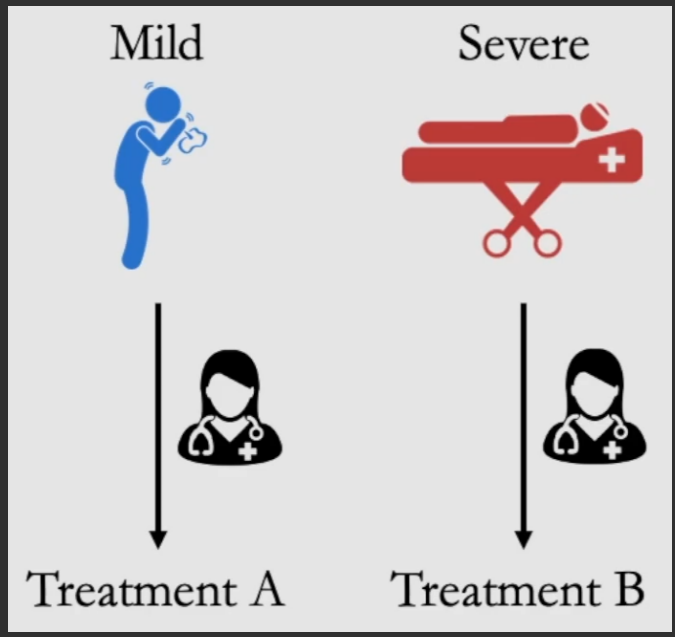
[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Motivating Example: Simpson's Paradox

Simpson's Paradox: Scenario 1 (treatment B)

T	Total	Mild	Severe
A	16% (240/1500)	15% (210/1400)	30% (30/100)
B	19% (105/550)	10% (5/50)	20% (100/500)





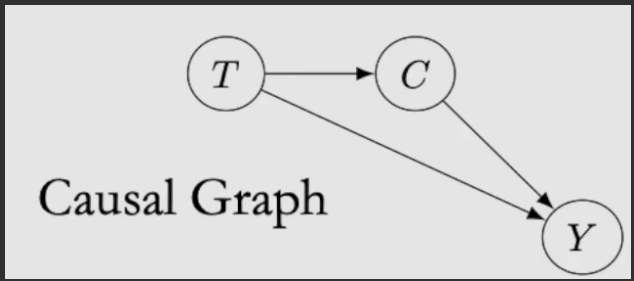
Treatment B is better than Treatment A in this scenario.

[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Motivating Example: Simpson's Paradox

Simpson's Paradox: Scenario 2 (treatment A)

T	Total	Mild	Severe
A	16% (240/1500)	15% (210/1400)	30% (30/100)
B	19% (105/550)	10% (5/50)	20% (100/500)



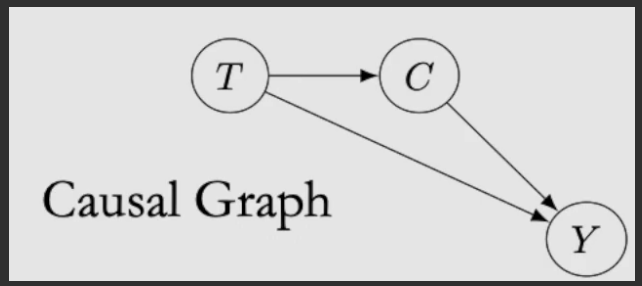
Treatment A is better than Treatment B in this scenario.

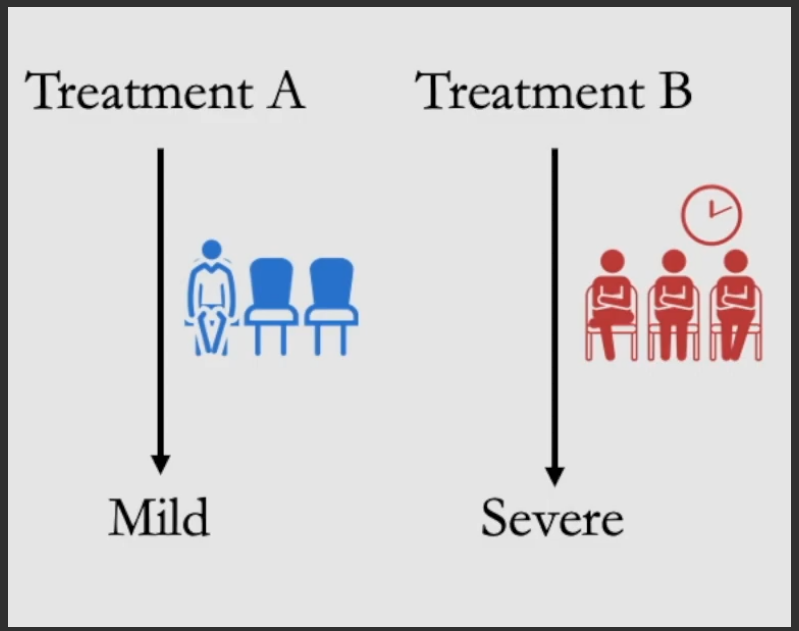
[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Motivating Example: Simpson's Paradox

Simpson's Paradox: Scenario 2 (treatment A)

T	Total	Mild	Severe
A	16% (240/1500)	15% (210/1400)	30% (30/100)
B	19% (105/550)	10% (5/50)	20% (100/500)





Treatment A is better than Treatment B in this scenario.

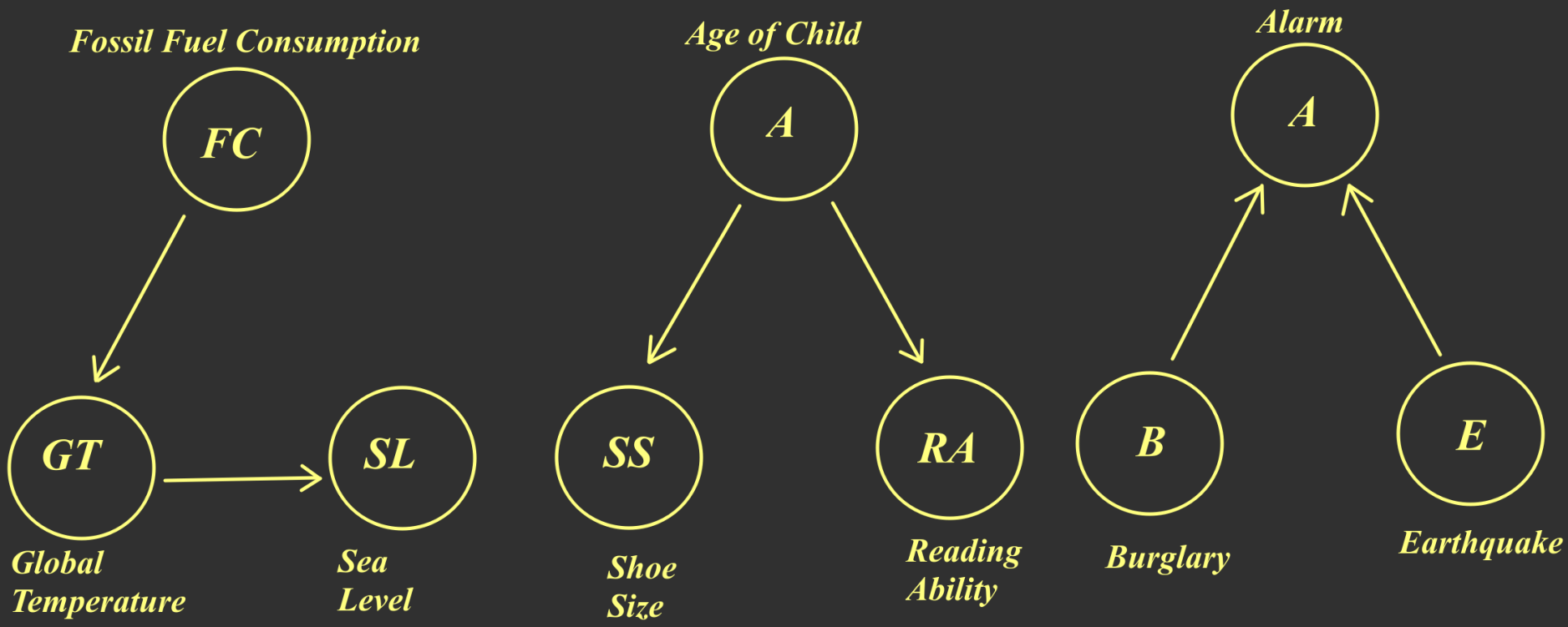
[1] Example taken from Introduction to Causal Inference, Brady Neal [2].

Discussion Task 1

Consider the following variable sets. Draw causal graphs between the variables:

→ Sea level, Fossil fuel consumption, Global temperature
→ Age of Child, Shoe size, Reading Ability
→ Alarm in your house, Burglary, Earthquake

Discussion Task 1



Recap: Conditional independence

Let X, Y and Z be random variables.

- X and Y are **(marginally) independent**, $X \perp Y$, if $P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y.$
- X and Y are **conditionally independent** given Z, $X \perp Y|Z$, if $\forall x, y, z$ with $P(Z = z) > 0$, $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z).$
- If two random variables are not (conditionally) independent, they are (conditionally) dependent.

Recap: Conditional independence

Let X, Y and Z be random variables.

- X and Y are **(marginally) independent**, $X \perp Y$, if $P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y.$
- X and Y are **conditionally independent** given Z, $X \perp Y|Z$, if $\forall x, y, z$ with $P(Z = z) > 0$, $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z).$
- If two random variables are not (conditionally) independent, they are (conditionally) dependent.

Recap: Conditional independence

Let X, Y and Z be random variables.

- X and Y are **(marginally) independent**, $X \perp Y$, if $P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y.$
- X and Y are **conditionally independent** given Z, $X \perp Y|Z$, if $\forall x, y, z$ with $P(Z = z) > 0$, $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z).$
- If two random variables are not (conditionally) independent, they are (conditionally) dependent.

Basic Causal Graphs

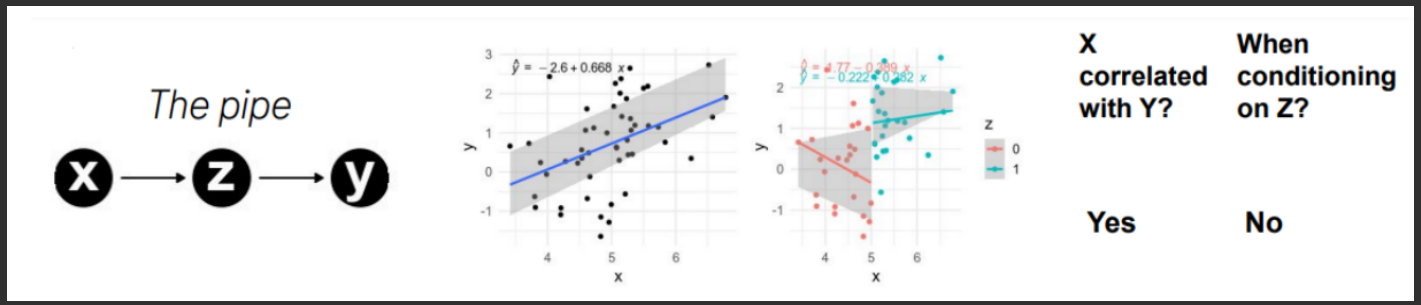


Figure: X is Gaussian with mean of 5 and standard deviation, Z takes the value of 0 if $x < 5$ and 1 otherwise, and Y is Gaussian with mean defined by $2 * z$ and standard deviation.

- Fossil fuel consumption $\rightarrow$Global temperature $\rightarrow$Sea level
- **X and Y are marginally correlated, and they are independent conditional on Z.**

[2] Example, figure and plots taken from Causal inference in drug discovery and development, Tom Michoel and Jitao David Zhang [1].

Basic Causal Graphs

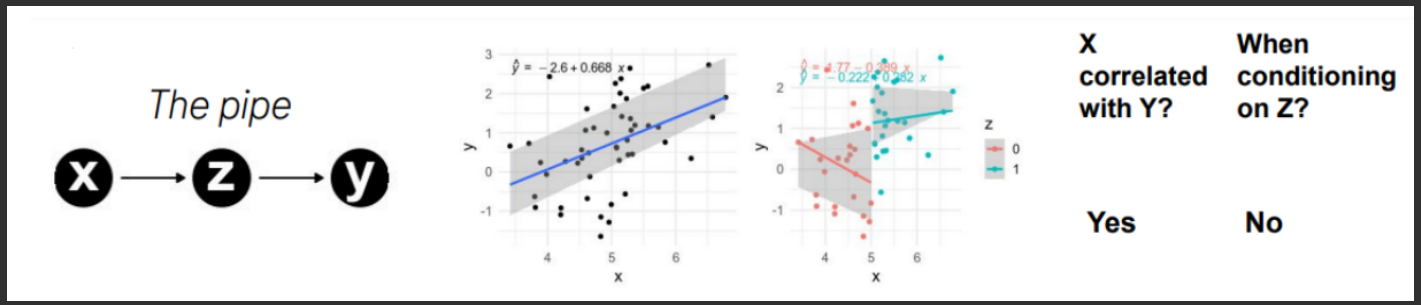


Figure: X is Gaussian with mean of 5 and standard deviation, Z takes the value of 0 if $x < 5$ and 1 otherwise, and Y is Gaussian with mean defined by $2 * z$ and standard deviation.

- Fossil fuel consumption $\rightarrow$Global temperature $\rightarrow$Sea level
- **X and Y are marginally correlated, and they are independent conditional on Z.**

[2] Example, figure and plots taken from Causal inference in drug discovery and development, Tom Michoel and Jitao David Zhang [1].

Basic Causal Graphs

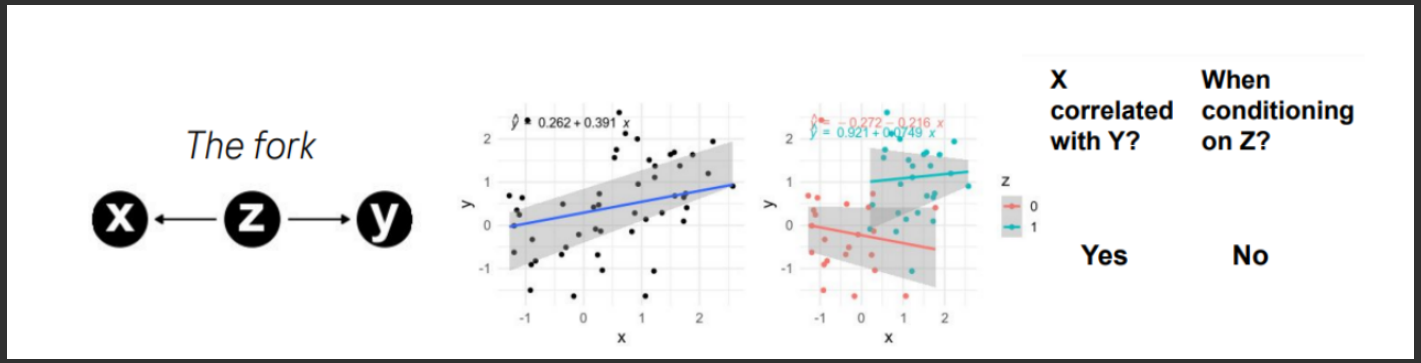


Figure: Z is Bernoulli with a probability of success of 0.5. Both X and Y are Gaussian with mean defined by z and standard deviation.

- Shoe size ←Age of child →Reading ability
  Children with larger shoes tend to read at a higher level. But giving a child larger shoes won't make him read better...
  - **X and Y are marginally correlated, and they become independent conditional on Z.**

[2] Example, figure and plots taken from Causal inference in drug discovery and development, Tom Michoel and Jitao David Zhang [1].

Basic Causal Graphs

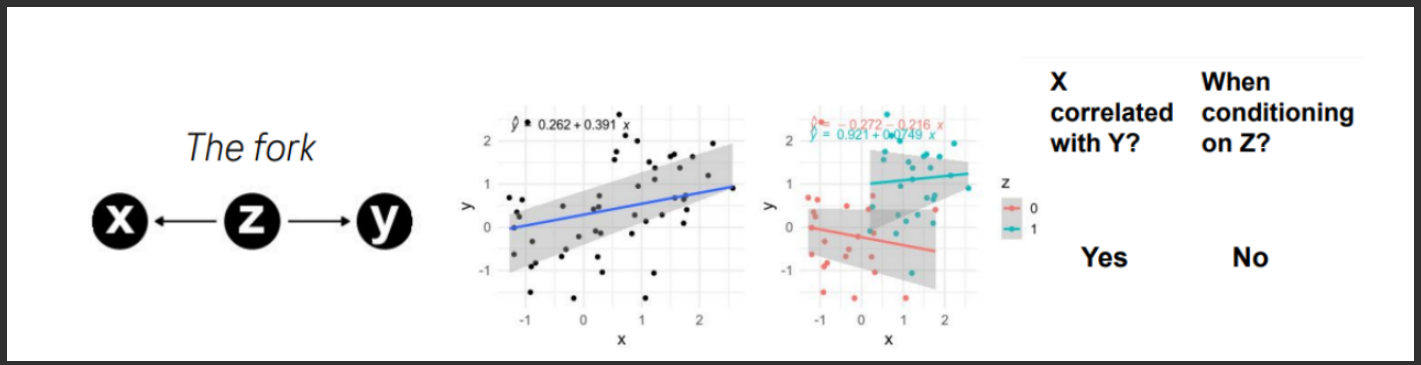


Figure: Z is Bernoulli with a probability of success of 0.5. Both X and Y are Gaussian with mean defined by z and standard deviation.

- Shoe size ←Age of child →Reading ability
  Children with larger shoes tend to read at a higher level. But giving a child larger shoes won't make him read better...
- **X and Y are marginally correlated, and they become independent conditional on Z.**

[2] Example, figure and plots taken from Causal inference in drug discovery and development, Tom Michoel and Jitao David Zhang [1].

Question

Earthquakes and burglaries are independent events that will cause an alarm to go off. Suppose you hear an alarm. How does hearing on the radio that there's an earthquake change your beliefs?

It increases the probability of burglary

It decreases the probability of burglary

It does not change the probability of burglary

Basic Causal Graphs

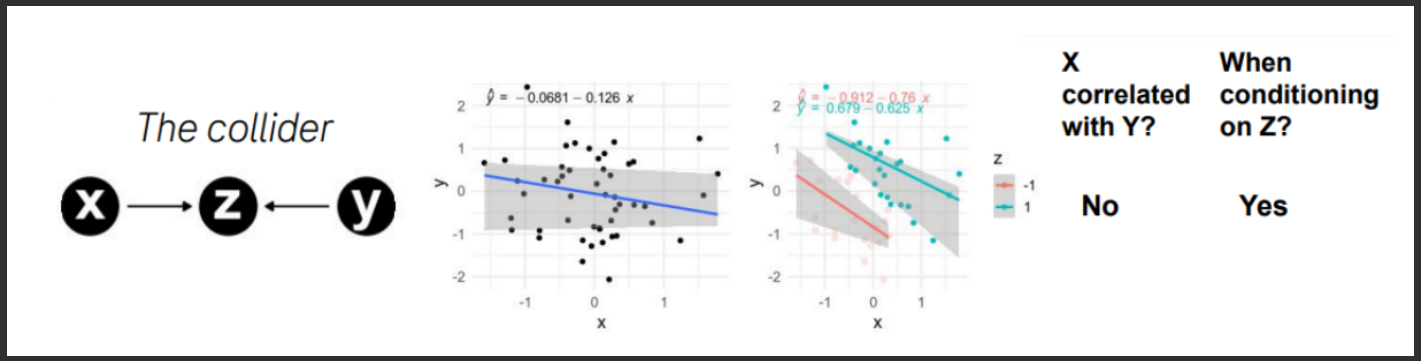


Figure: Both X and Y are Gaussian with 0 mean and standard deviation. The value of Z is 1 if $x + y > 0$ and -1 otherwise.

- Earthquake $\rightarrow$Alarm $\leftarrow$Burglary
- **In contrast to pipes and forks, X and Y are marginally independent but become correlated conditional on Z.**

[2] Example, figure and plots taken from Causal inference in drug discovery and development, Tom Michoel and Jitao David Zhang [1].

Basic Causal Graphs

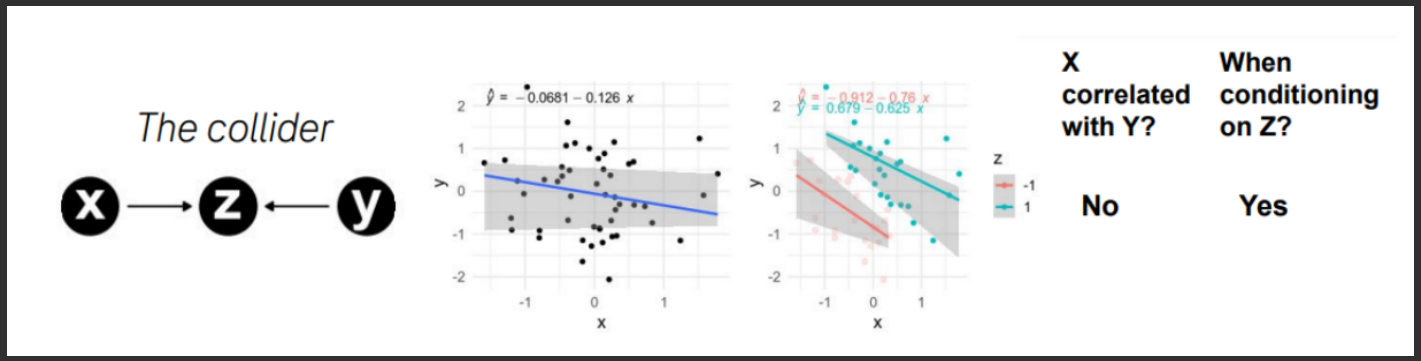


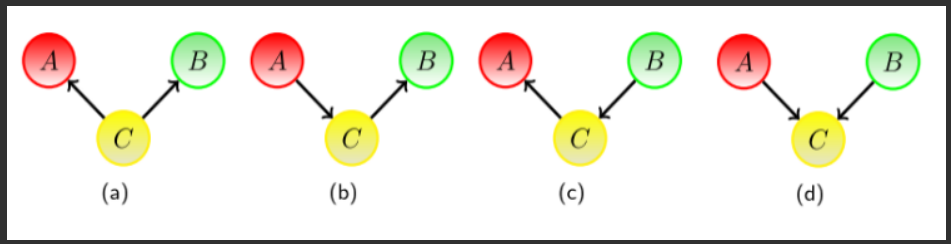
Figure: Both X and Y are Gaussian with 0 mean and standard deviation. The value of Z is 1 if $x + y > 0$ and -1 otherwise.

- Earthquake $\rightarrow$Alarm $\leftarrow$Burglary
- **In contrast to pipes and forks, X and Y are marginally independent but become correlated conditional on Z.**

[2] Example, figure and plots taken from Causal inference in drug discovery and development, Tom Michoel and Jitao David Zhang [1].

Markov equivalence

Two graphs are **Markov equivalent**, if they entail the same conditional independencies and can be used for representing exactly the same set of probability distributions.



Discussion Task 2

Which graphs are Markov Equivalent?

Part 2
Bayesian Networks

Bayesian Networks

What are Bayesian Networks?

A probabilistic graphical model representing probabilistic relationships between random variables.

How do Bayesian Networks help us?

A BN specifies a joint distribution in a structured form.

Bayesian Networks

What are Bayesian Networks?

A probabilistic graphical model representing probabilistic relationships between random variables.

How do Bayesian Networks help us?

A BN specifies a joint distribution in a structured form.

Bayesian Networks

What are Bayesian Networks?

A probabilistic graphical model representing probabilistic relationships between random variables.

How do Bayesian Networks help us?

A BN specifies a joint distribution in a structured form.

Bayesian Networks

What are Bayesian Networks?

A probabilistic graphical model representing probabilistic relationships between random variables.

How do Bayesian Networks help us?

A BN specifies a joint distribution in a structured form.

Chain Rule in probability

From definition of conditional probability, we have,

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

Can be generalized to multiple events,

$$P(A_1, A_2, A_3) = P(A_1|A_2, A_3)P(A_2|A_3)P(A_3)$$

Generally,

$$P(A_1, A_2, ..., A_k) = \prod_{i=1}^{k} P(A_i|A_{i+1}, A_{i+2}, ..., A_k)$$

$\rightarrow$ This factorization holds for any ordering of the variables.

$\rightarrow$ This is the chain rule for probabilities.

Chain Rule in probability

From definition of conditional probability, we have,

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

Can be generalized to multiple events,

$$P(A_1, A_2, A_3) = P(A_1|A_2, A_3)P(A_2|A_3)P(A_3)$$

Generally,

$$P(A_1, A_2, ..., A_k) = \prod_{i=1}^{k} P(A_i|A_{i+1}, A_{i+2}, ..., A_k)$$

$\rightarrow$ This factorization holds for any ordering of the variables.

$\rightarrow$ This is the chain rule for probabilities.

Chain Rule in probability

From definition of conditional probability, we have,

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

Can be generalized to multiple events,

$$P(A_1, A_2, A_3) = P(A_1|A_2, A_3)P(A_2|A_3)P(A_3)$$

Generally,

$$P(A_1, A_2, ..., A_k) = \prod_{i=1}^{k} P(A_i|A_{i+1}, A_{i+2}, ..., A_k)$$

$\rightarrow$ This factorization holds for any ordering of the variables.

$\rightarrow$ This is the chain rule for probabilities.

Chain rule and conditional independence

If X and Y are **conditionally independent** given Z, $X \perp Y|Z$ then:

$$
\begin{aligned}
P(X|Y,Z) =& \frac{P(X,Y,Z)}{P(Y,Z)} \\
=& \frac{P(X,Y|Z)P(Z)}{P(Y|Z)P(Z)} \\
=& \frac{P(X|Z)P(Y|Z)P(Z)}{P(Y|Z)P(Z)} \\
=& P(X|Z)
\end{aligned}
$$

Graphs as Joint Distribution Factorizations

You have 4 binary random variables :

Cloudy (C)
Sprinkler(S)
Rain(R)
Wet grass(W)

Discussion Task 3

Use the chain rule of probabilities to factorize the joint distribution.
How many values do you need to store to specify $P(W, S, R, C)$?

Graphs as Joint Distribution Factorizations

You have 4 binary random variables :

Cloudy (C)
Sprinkler(S)
Rain(R)
Wet grass(W)

Use the chain rule of probabilities:
$P(C, S, R, W) =$
$P(W|S, R, C)P(S, R, C)$
Repeatedly applying this idea,
$P(S, R, C) =$
$P(S|R, C)P(R|C)P(C)$
Order does not matter.

How many values do you need to store to specify $p(W|S, R, C)$?

$$P(W = 0|S = 0, R = 0, C = 0)$$
$$P(W = 0|S = 0, R = 0, C = 1)$$
$$\vdots$$
$$P(W = 0|S = 1, R = 1, C = 1)$$

Need $2^n = 8$ parameters.
$(\rightarrow P(A^c|C) = 1 - P(A|C))$
Similarly for entire $P(C, S, R, W)$,
$2^n - 1 = 15$ parameters.

Graphs as Joint Distribution Factorizations

You have 4 binary random variables :

Cloudy (C)
Sprinkler(S)
Rain(R)
Wet grass(W)

Use the chain rule of probabilities:
$P(C, S, R, W) =$
$P(W|S, R, C)P(S, R, C)$
Repeatedly applying this idea,
$P(S, R, C) =$
$P(S|R, C)P(R|C)P(C)$
Order does not matter.

How many values do you need to store to specify $p(W|S, R, C)$?

$$P(W = 0|S = 0, R = 0, C = 0)$$
$$P(W = 0|S = 0, R = 0, C = 1)$$
$$\vdots$$
$$P(W = 0|S = 1, R = 1, C = 1)$$

Need $2^n = 8$ parameters.
$(\rightarrow P(A^c|C) = 1 - P(A|C))$
Similarly for entire $P(C, S, R, W)$,
$2^n - 1 = 15$ parameters.

Discussion Task 4

For the 4 binary random variables :

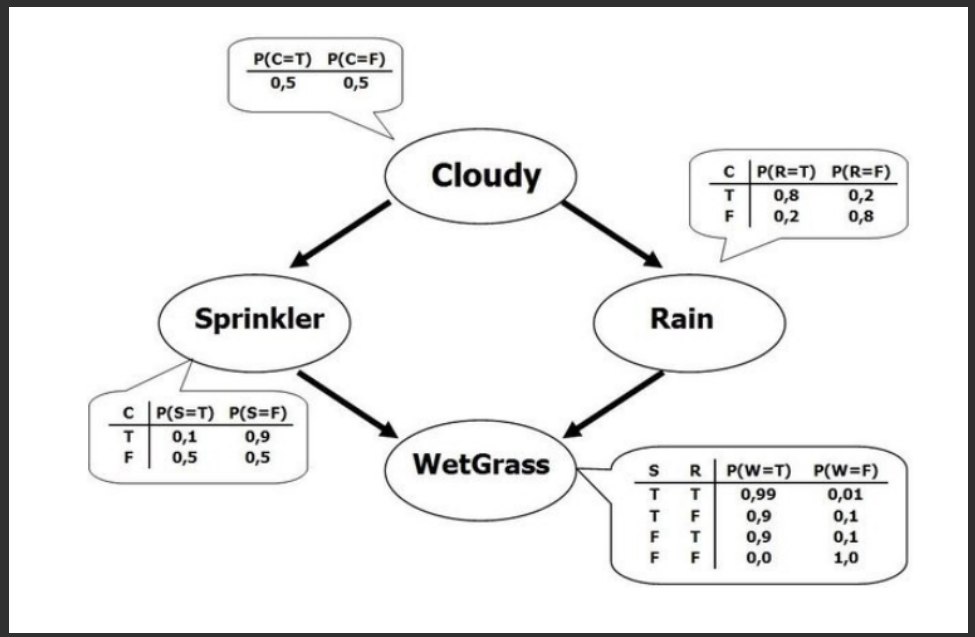
Cloudy (C)
Sprinkler(S)
Rain(R)
Wet grass(W)

1. Draw the causal graph for these variables.
2. How many parameters do you need to specify the joint distribution using this causal graph?

Order matters!

Encoding (conditional) independence relationships using graphs:

$$P(C, S, R, W) = P(W|S, R, C)P(S|R, C)P(R|C)P(C)$$
$$P(W, S, R, C) = P(W|S, R)P(S|C)P(R|C)P(C)$$



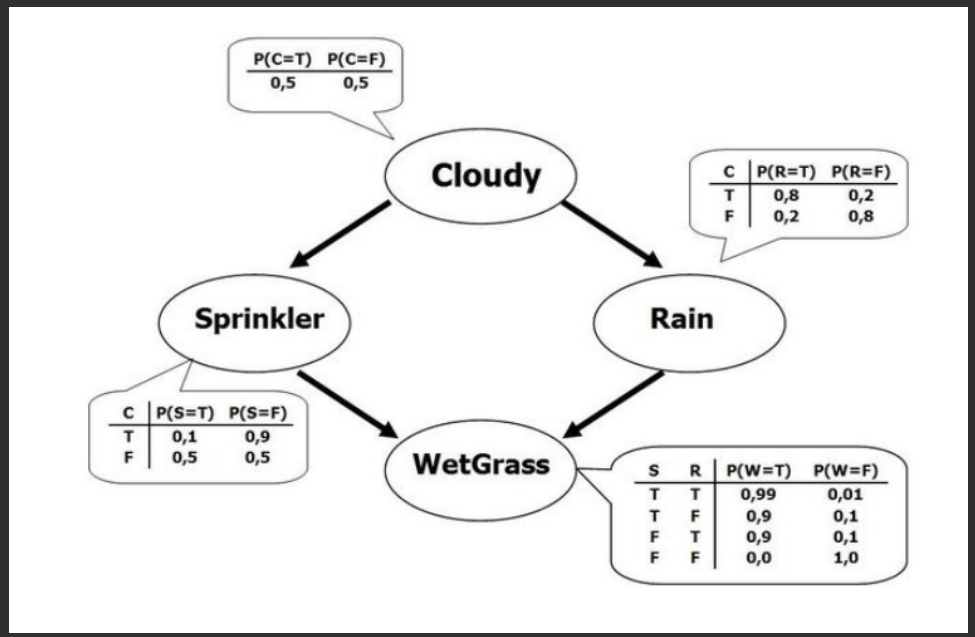
How many values do you need to store to specify $P(W, S, R, C)$?

Need 9 parameters!

Order matters!

Encoding (conditional) independence relationships using graphs:

$$P(C, S, R, W) = P(W|S, R, C)P(S|R, C)P(R|C)P(C)$$
$$P(W, S, R, C) = P(W|S, R)P(S|C)P(R|C)P(C)$$



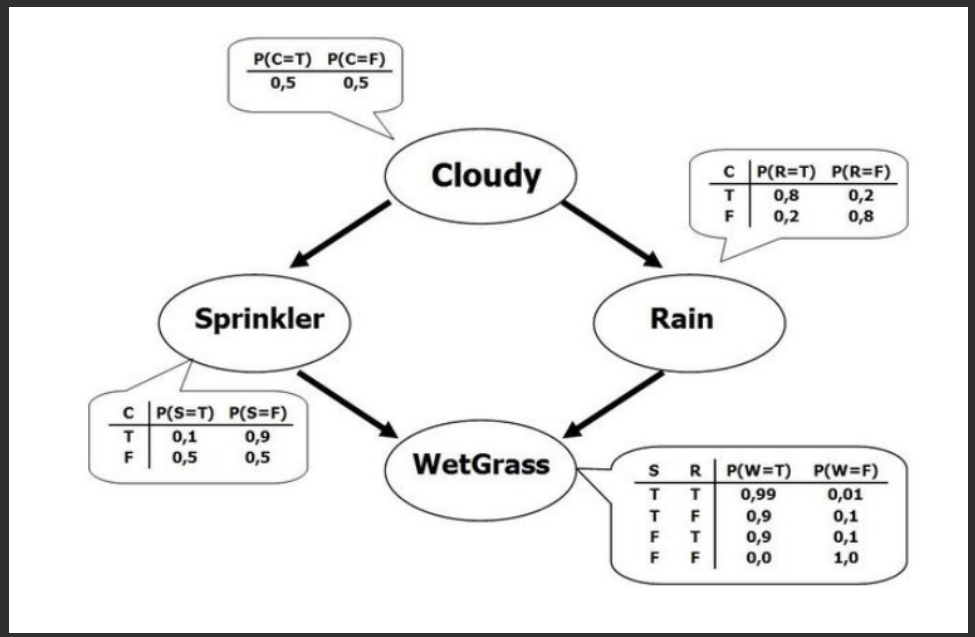
How many values do you need to store to specify $P(W, S, R, C)$?

Need 9 parameters!

Order matters!

Encoding (conditional) independence relationships using graphs:

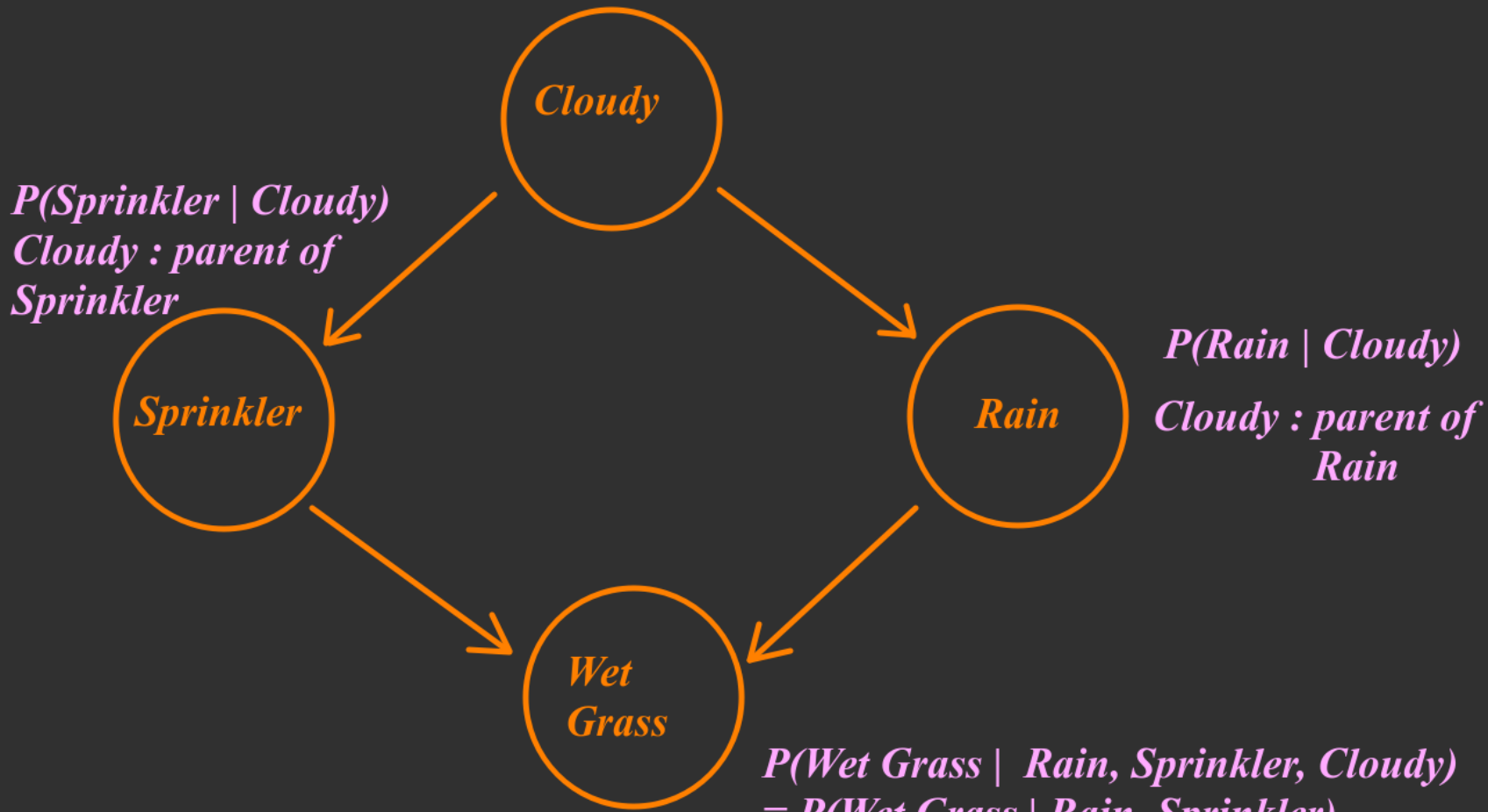
$$P(C, S, R, W) = P(W|S, R, C)P(S|R, C)P(R|C)P(C)$$
$$P(W, S, R, C) = P(W|S, R)P(S|C)P(R|C)P(C)$$



How many values do you need to store to specify $P(W, S, R, C)$?

Need 9 parameters!

Order matters!



Markov Condition

Definition: Bayesian network

Let $X = (X_1, ..., X_n)$ be random variables. A Bayesian network is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i | parents(X_i)$$

Compactness of Bayesian Network

Assume we have d binary variables.

1. To specify a probability table, one needs $2^d - 1$, that is, $O(2^d)$ parameters.
2. If the distribution factorizes according to a DAG where each node has at most k parents, one needs at most $O(d2^k)$ parameters.

For example, if $d = 30$ and $k = 5$, a Bayesian network would have less than 960 parameters while a probability table would require over a billion parameters.

Compactness of Bayesian Network

Assume we have d binary variables.
1. To specify a probability table, one needs $2^d - 1$, that is, $O(2^d)$ parameters.
2. If the distribution factorizes according to a DAG where each node has at most k parents, one needs at most $O(d2^k)$ parameters.

For example, if $d = 30$ and $k = 5$, a Bayesian network would have less than 960 parameters while a probability table would require over a billion parameters.

Compactness of Bayesian Network

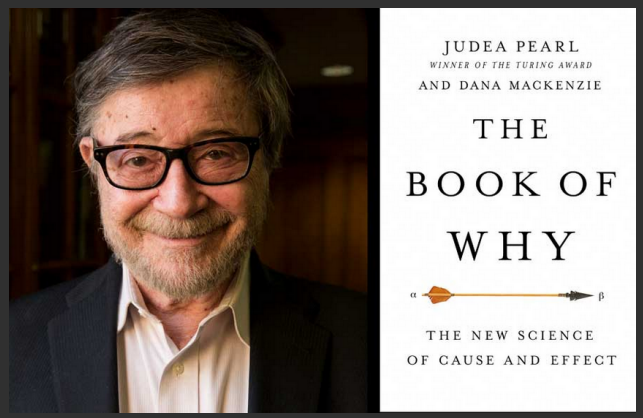
Assume we have d binary variables.
1. To specify a probability table, one needs $2^d - 1$, that is, $O(2^d)$ parameters.
2. If the distribution factorizes according to a DAG where each node has at most k parents, one needs at most $O(d2^k)$ parameters.

For example, if $d = 30$ and $k = 5$, a Bayesian network would have less than 960 parameters while a probability table would require over a billion parameters.

Part 3
Causal Inference

Causal Bayesian Networks

What can a causal reasoner do?



Causal Bayesian Networks

Association

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see...?*
How are two variables related?

EXAMPLES: How likely is a customer who bought toothpaste to also buy dental floss?

[3] Example taken from The book of Why, Dana Mackenzie and Judea Pearl [4].

Causal Bayesian Networks

What is an intervention?

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do...?*
What would happen if I do X?

EXAMPLES: What will happen to our floss sales if we double the price of toothpaste?

[3] Example taken from The book of Why, Dana Mackenzie and Judea Pearl [4].

Causal Bayesian Networks

What are counterfactuals?

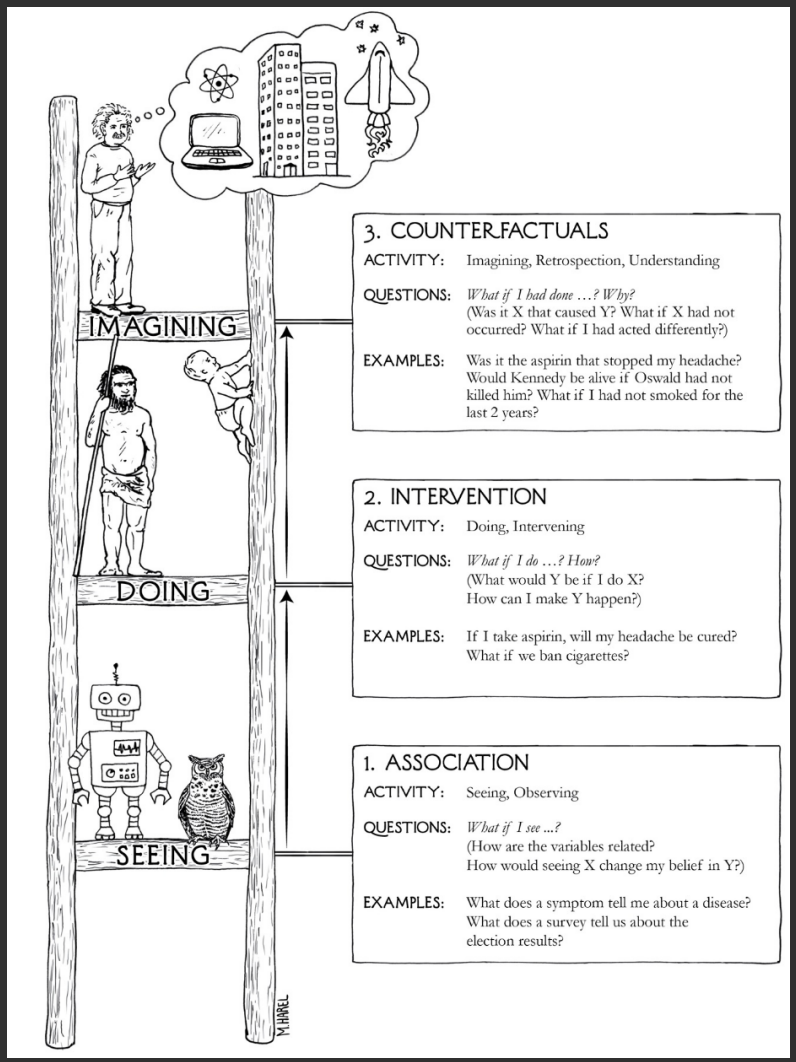
ACTIVITY: Imagining, Retrospection

QUESTIONS: *What if I had done...?*
What if I had acted differently?

EXAMPLES: What is the probability that a customer who bought toothpaste would still have bought it if we had doubled the price?

[3] Example taken from The book of Why, Dana Mackenzie and Judea Pearl [4].

Pearl's Ladder of Causation

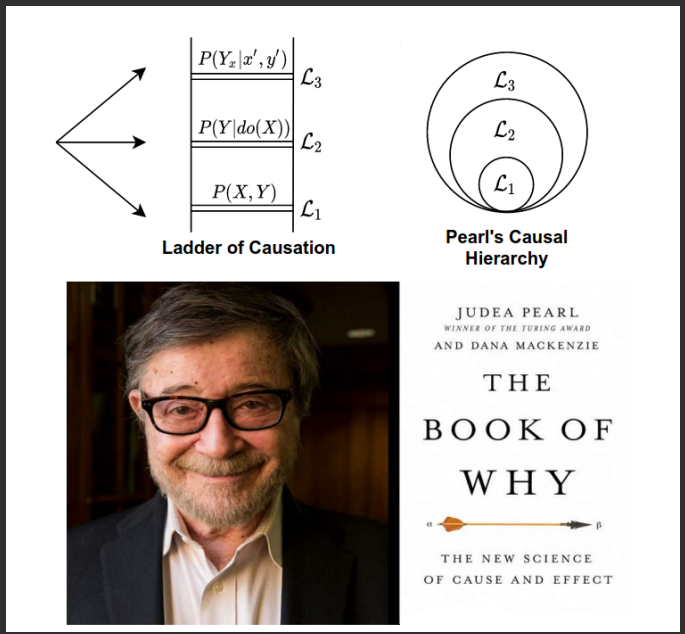


"What can a causal reasoner do?"

$\rightarrow \mathcal{L}_1$ *Associations*

$\rightarrow \mathcal{L}_2$ *Interventions*

$\rightarrow \mathcal{L}_3$ *Counterfactuals*



[3] Figure taken from The book of Why, Dana Mackenzie and Judea Pearl [4].

How do we intervene?

Interventions and counterfactuals are defined through a mathematical operator called $do(x)$*:*

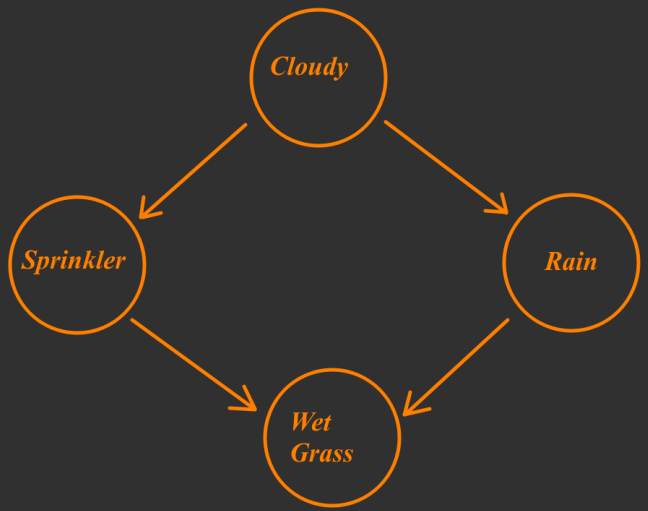
→***Simulates physical interventions by deleting certain functions from the model, replacing them with a constant*** $X = x$***, while keeping the rest of the model unchanged.***

→The Do-Calculus Revisited, Judea Pearl.

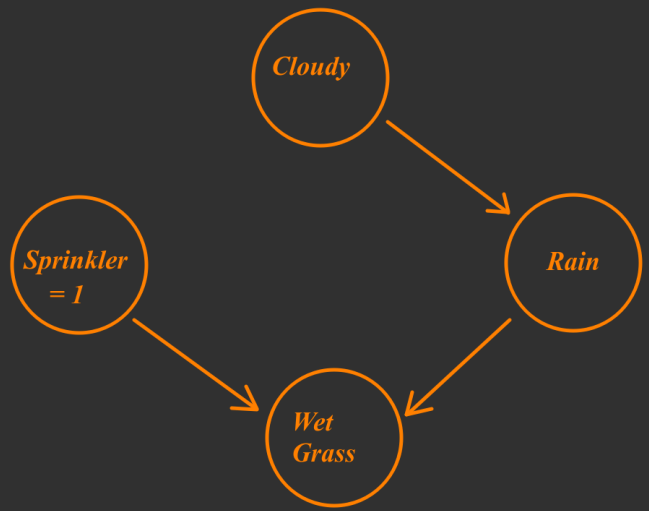
[4] Definition taken from The Do-Calculus Revisited ,Judea Pearl [3].

Difference between *seeing* and *doing*

$P(WetGrass|Sprinkler)$:
observation

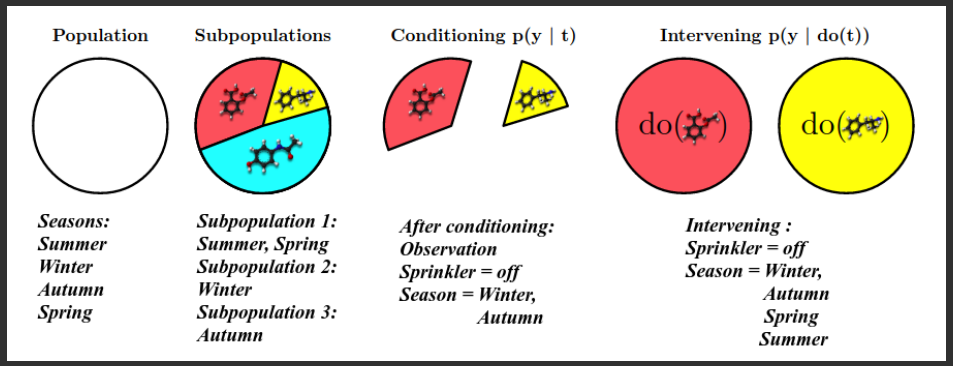


$P(WetGrass|do(Sprinkler))$:
intervention



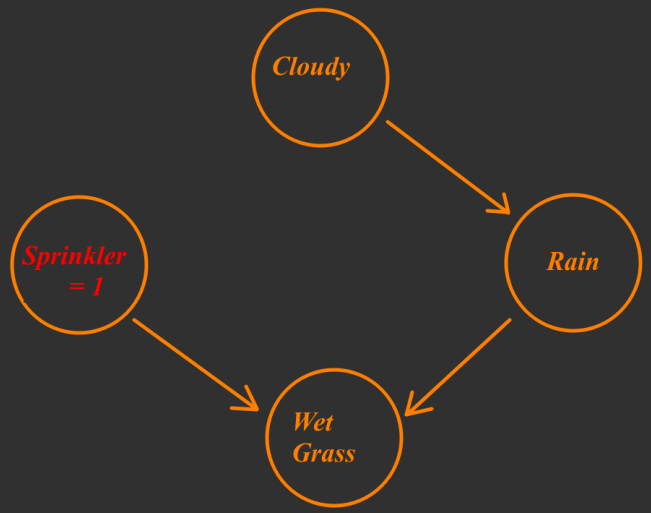
Do operator and Interventions

Conditioning versus intervening



Sprinkler: Intervention $P(y|do(Sprinkler))$ refers to the scenario in which independent of season, sprinkler is turned off versus the conditional distribution $P(y|Sprinkler)$ restricting our focus to the sub-population of seasons that likely lead to turning sprinkler off.

Do operator and Interventions



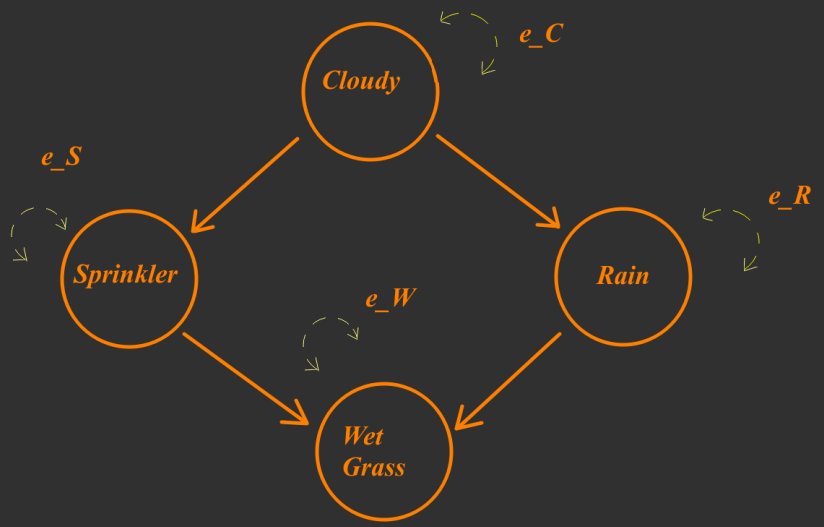
$$P(C, R, W|do(S = On)) = P(C)P(R|C)\underbrace{P(S|C)}_{=1}P(W|S, R)$$

$$\neq P(C, R, W|S)$$

SCM's

A Structural Causal Model consist of two components:

A **Causal Graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a *directed acyclic graph* (DAG) that describes causal effects between variables, where $\mathcal{V}$ is node set and $\mathcal{E}$ is the edge set.



Structural Equations: Given a causal graph, specify the causal effects signified by the directed edges:

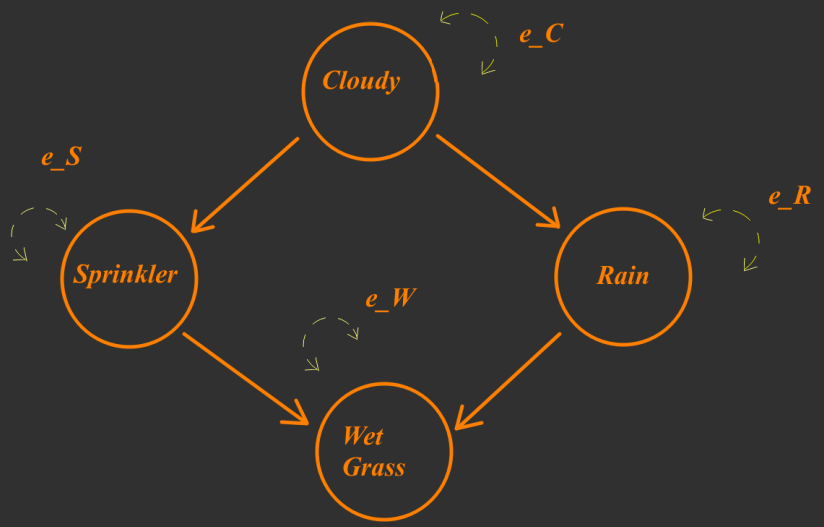
$$c = f_c(e_C), s = f_s(c, e_S)$$
$$r = f_r(c, e_R), w = f_w(r, s, e_W)$$

e_C ,e_S, e_R and e_W denote the "noise" of the observed variables.

SCM's

A Structural Causal Model consist of two components:

A **Causal Graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a *directed acyclic graph* (DAG) that describes causal effects between variables, where $\mathcal{V}$ is node set and $\mathcal{E}$ is the edge set.



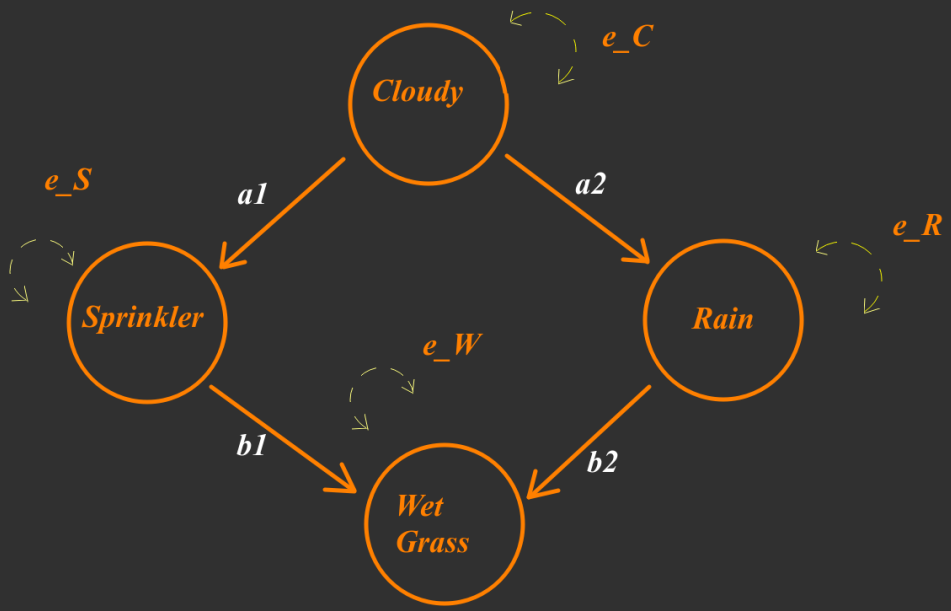
Structural Equations: Given a causal graph, specify the causal effects signified by the directed edges:

$$c = f_c(e_C), s = f_s(c, e_S)$$
$$r = f_r(c, e_R), w = f_w(r, s, e_W)$$

e_C ,e_S, e_R and e_W denote the "noise" of the observed variables.

SCM example

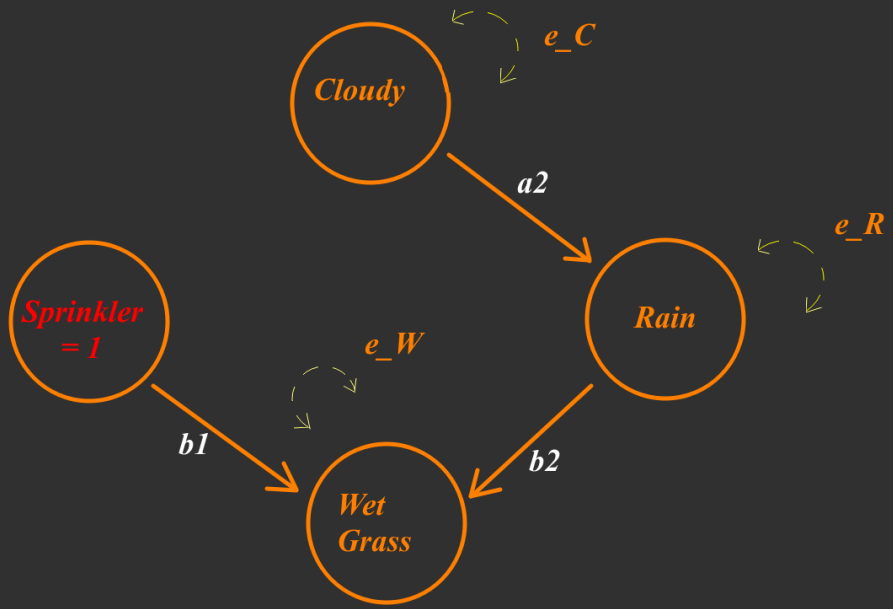
For the SCM $\mathcal{M} := (\mathcal{S}, P(e))$, this is the DAG $\mathcal{G}$.



For the SCM $\mathcal{M} := (\mathcal{S}, P(e))$, this is the structural assignments $\mathcal{S}$.

$$C = e_C$$
$$R = a_2 C + e_R$$
$$S = a_1 C + e_S$$
$$W = b_1 S + b_2 R + e_W$$

SCM after intervention



Modified SCM

$$\tilde{\mathcal{M}} = \mathcal{M}_{x;do(Sprinkler=1)} = (\tilde{\mathcal{S}}, P(e))$$

$$C = e_C$$

$$R = a_2 C + e_R$$

$$W = b_1(1) + b_2 R + e_W$$

Counterfactual inference

Intervention:
Substitute structural assignments with the intervention's value.

Interventional distribution:
Noise terms consist of the prior distribution $P(e)$

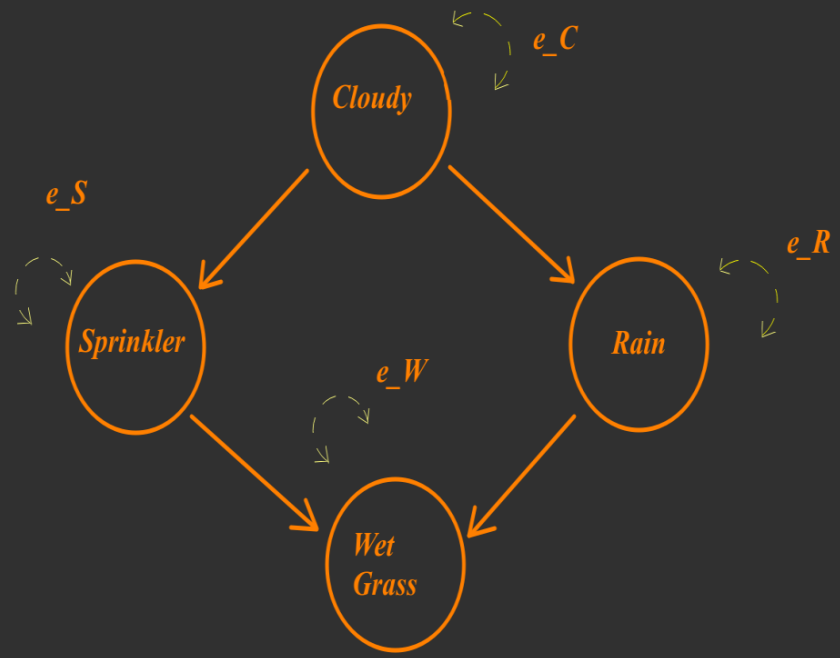
Counterfactual:
Substitute structural assignments with the intervention's value.

Counterfactual distribution:
Noise terms consist of the posterior distribution $P(e|data)$ (which incorporates our knowledge of what already happened).

Conterfactual Inference

Original SCM with noise variables e_C, e_R, e_S, e_W.

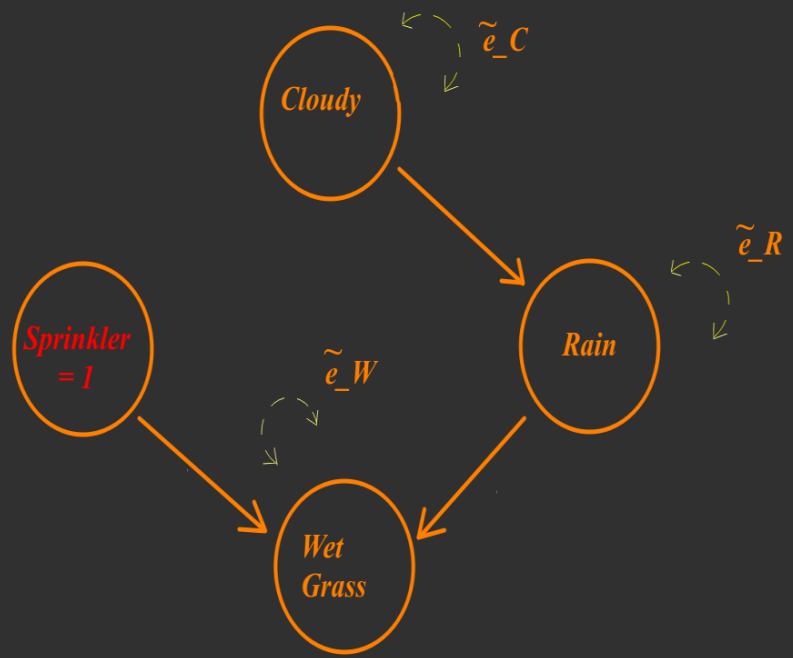
$$C = e_C$$
$$R = a_2 C + e_R$$
$$S = a_1 C + e_S$$
$$W = b_1 S + b_2 R + e_W$$



Modified SCM:

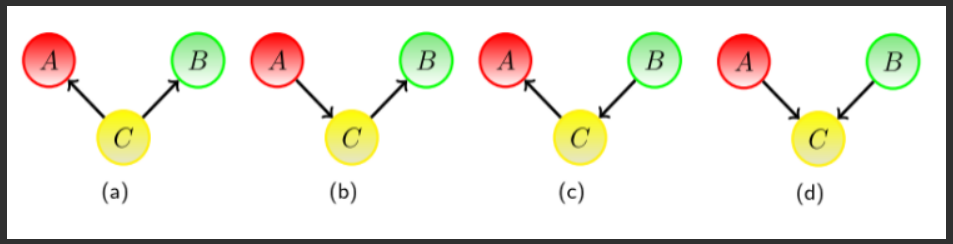
$$C = \tilde{e}_C$$
$$R = a_2 C + \tilde{e}_R$$
$$W = b_1 (1) + b_2 R + \tilde{e}_W$$

where $p(\tilde{e}_i) = p(e_i|x)$, x is observed C, R, W



Difference between BN's and causal BN's

Consider the following graphs from before:



Discussion Task 3

How are BN's and causal graphs different (discuss using these graphs)?

Part 4
Causal Fairness

Why Causality matters for Fairness?

- "To establish a disparate-treatment claim, a plaintiff must prove that impact was the "but-for" cause of the employer's adverse decision." $\rightarrow$US Supreme Court, 2008
- "If the plaintiff cannot show a causal connection between the Department's policy and a disparate impact—that should result in dismissal of this case."

Why Causality matters for Fairness?

- "To establish a disparate-treatment claim, a plaintiff must prove that impact was the "but-for" cause of the employer's adverse decision." $\rightarrow$US Supreme Court, 2008
- "If the plaintiff cannot show a causal connection between the Department's policy and a disparate impact—that should result in dismissal of this case."

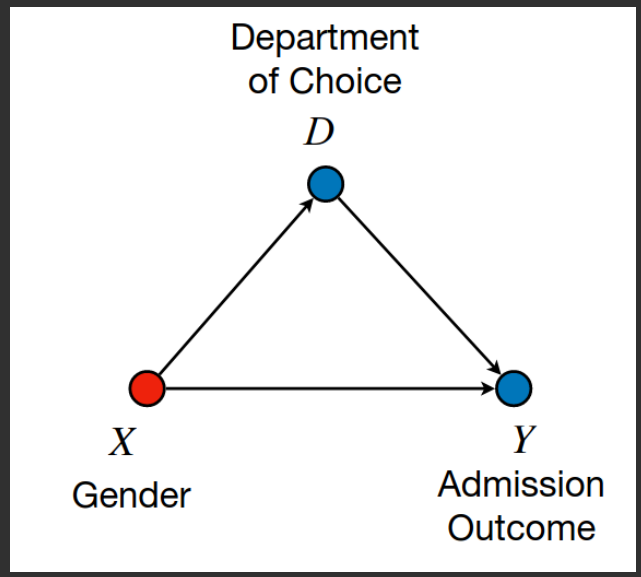
Examples

Example 1 (Berkeley admission)

Students apply for university admission (Y), and choose specific departments to which they wish to join ($D = 0$ for sciences, $D = 1$ for arts and humanities). For the purpose of discrimination monitoring, gender is also recorded ($X = 0$ for male, $X = 1$ for female).

SCM M*

$$X \leftarrow f_X(U_X)$$
$$D \leftarrow f_D(X, U_D)$$
$$Y \leftarrow f_Y(X, D, U_Y)$$
$$P(U_X, U_D, U_Y)$$



Truth Unobserved

[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

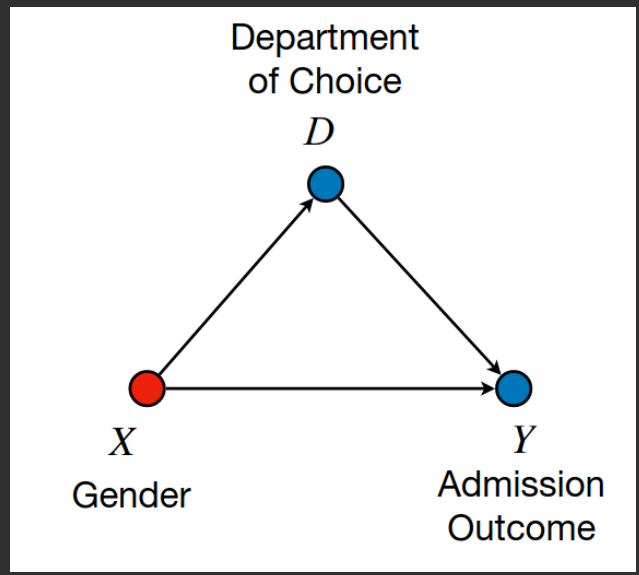
Examples

Example 1 (Berkeley admission)

Students apply for university admission (Y), and choose specific departments to which they wish to join ($D = 0$ for sciences, $D = 1$ for arts and humanities). For the purpose of discrimination monitoring, gender is also recorded ($X = 0$ for male, $X = 1$ for female).

$$X \leftarrow Bernoulli(0.5)$$
$$D \leftarrow Bernoulli(0.5 + \lambda X)$$
$$Y \leftarrow Bernoulli(0.1 + \alpha X + \beta D)$$

Truth Unobserved



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Examples

Example 1 (Berkeley admission)

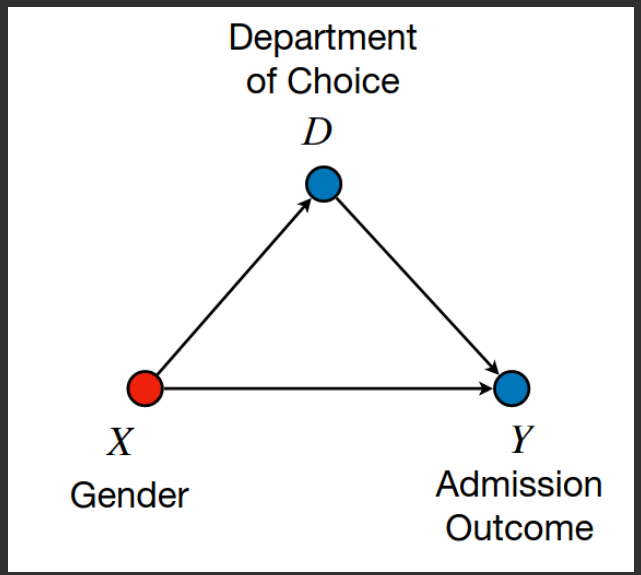
Students apply for university admission (Y), and choose specific departments to which they wish to join ($D = 0$ for sciences, $D = 1$ for arts and humanities). For the purpose of discrimination monitoring, gender is also recorded ($X = 0$ for male, $X = 1$ for female).

Data Analysis reveals:

$$TV_{x_0,x_1}(Y) = \mathbf{E}[Y|x_1] - \mathbf{E}[Y|x_0] < 0$$

Female Applicant predicted to have lower probability of admission than males.

Is this enough to conclude that female applicants were discriminated against?



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Examples

Example 1 (Berkeley admission)

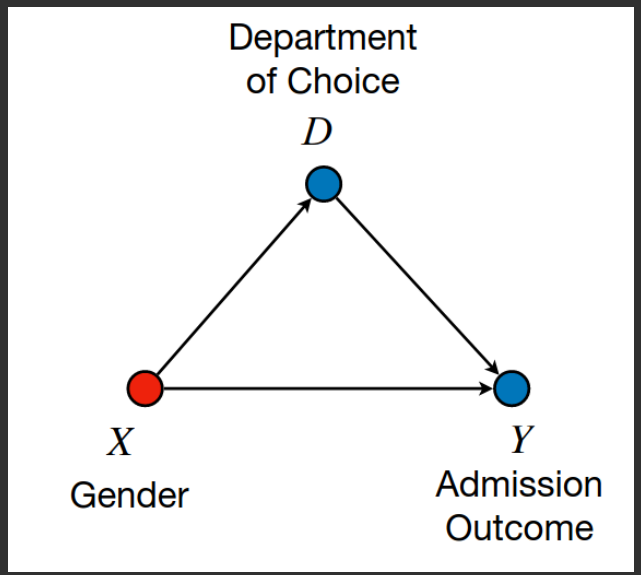
Students apply for university admission (Y), and choose specific departments to which they wish to join ($D = 0$ for sciences, $D = 1$ for arts and humanities). For the purpose of discrimination monitoring, gender is also recorded ($X = 0$ for male, $X = 1$ for female).

Data Analysis reveals:

$$TV_{x_0,x_1}(Y) = \mathbf{E}[Y|x_1] - \mathbf{E}[Y|x_0] < 0$$

Female Applicant predicted to have lower probability of admission than males.

Is this enough to conclude that female applicants were discriminated against?



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Examples

Example 1 (Berkeley admission)

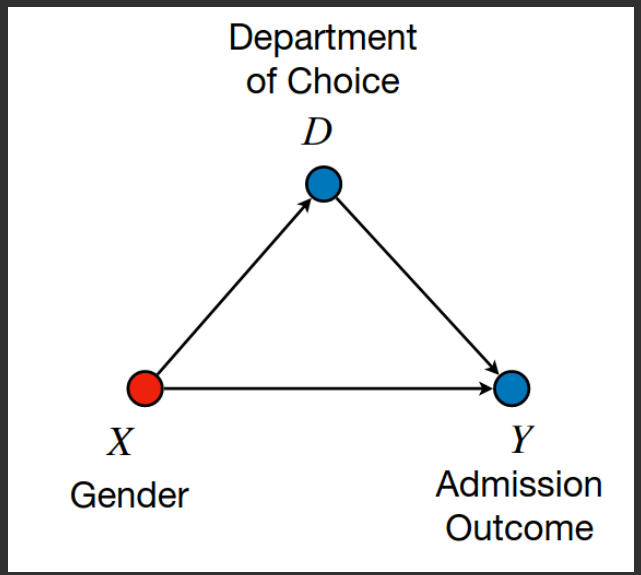
Students apply for university admission (Y), and choose specific departments to which they wish to join ($D = 0$ for sciences, $D = 1$ for arts and humanities). For the purpose of discrimination monitoring, gender is also recorded ($X = 0$ for male, $X = 1$ for female).

Data Analysis reveals:

$$TV_{x_0,x_1}(Y) = \mathbf{E}[Y|x_1] - \mathbf{E}[Y|x_0] < 0$$

Female Applicant predicted to have lower probability of admission than males.

Is this enough to conclude that female applicants were discriminated against?



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Records from the largest departments

Acceptance rates were relatively high in the largest departments, and they were higher for women than for men...

Department	**Men**		**Women**	
	Applicants	Admitted	Applicants	Admitted
	825	62%	108	82%
	560	63%	25	68%
	325	37%	593	34%
	417	33%	375	35%
	191	28%	393	24%
	373	6%	341	7%

Records from the largest departments

..but women were severely underrepresented in the applicant pools.

Department	**Men**		**Women**	
	Applicants	Admitted	Applicants	Admitted
	825	62%	108	82%
	560	63%	25	68%
	325	37%	593	34%
	417	33%	375	35%
	191	28%	393	24%
	373	6%	341	7%

Discussion Task 5

Should direct effect be the only measure a discrimination?

Think of ways in which discrimination can be defined using this causal graph. What are the arguments for and against seeing only the direct effect as a measure of the discrimination?

Discussion Task 5

Should direct effect be the only measure a discrimination?

Think of ways in which discrimination can be defined using this causal graph. What are the arguments for and against seeing only the direct effect as a measure of the discrimination?

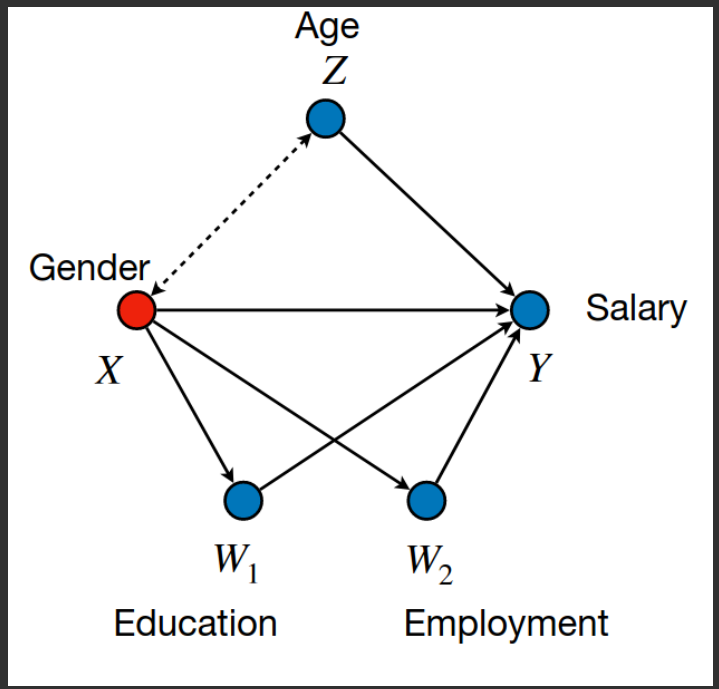
Examples

Example 2 (Government Census)

The US census data records a person's yearly salary (in tens of thousands of \$). The census also records age (Z), gender ($X = 0$ for male, $X = 1$ for female), education level (W_1) and employment status (W_2, 10 job types).

SCM M*

$$
\begin{aligned}
X &\leftarrow f_X(U_X) \\
Z &\leftarrow f_Z(U_Z) \\
W_1 &\leftarrow f_{W_1}(X, U_{W_1}) \\
W_2 &\leftarrow f_{W_2}(X, U_{W_2}) \\
Y &\leftarrow f_Y(X, Z, W_1, W_2, U_Y)
\end{aligned}
$$



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Examples

Example 2 (Government Census)

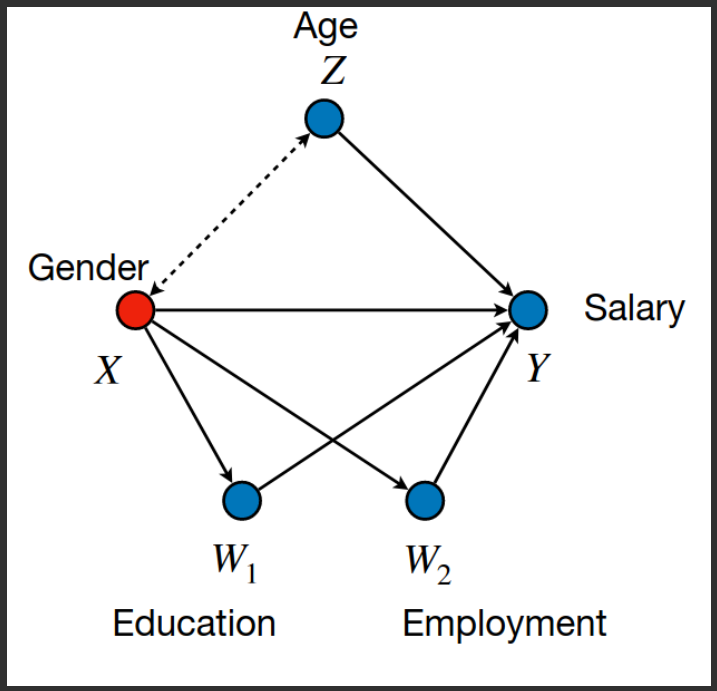
The US census data records a person's yearly salary (in tens of thousands of \$). The census also records age (Z), gender ($X = 0$ for male,$X = 1$ for female), education level (W_1) and employment status (W_2, 10 job types).

Data Analysis reveals:

$$TV_{x_0,x_1}(Y) = \mathbf{E}[Y|x_1] - \mathbf{E}[Y|x_0] < 0$$

Female employee has lower chances of high income than male employee.

Is this enough to conclude that females are systematically discriminated again in companies?



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Examples

Example 2 (Government Census)

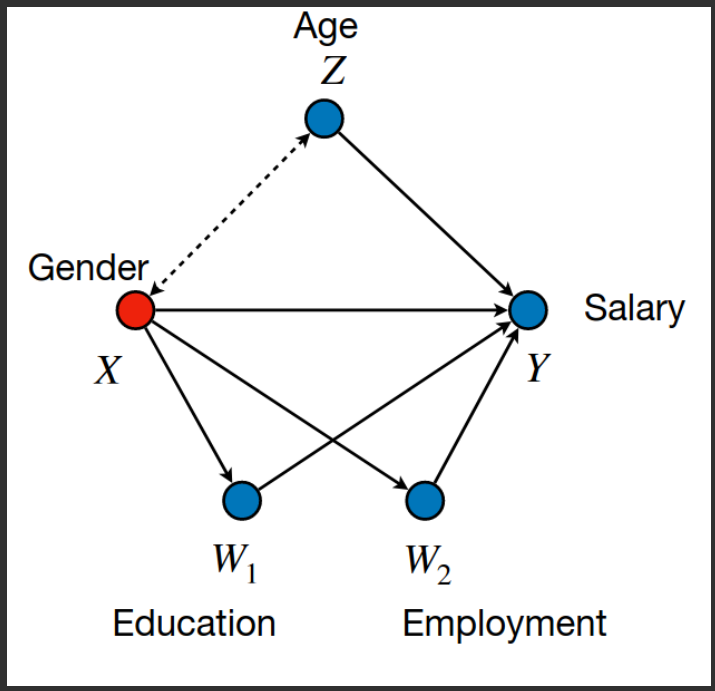
The US census data records a person's yearly salary (in tens of thousands of \$). The census also records age (Z), gender ($X = 0$ for male,$X = 1$ for female), education level (W_1) and employment status (W_2, 10 job types).

Data Analysis reveals:

$$TV_{x_0,x_1}(Y) = \mathbf{E}[Y|x_1] - \mathbf{E}[Y|x_0] < 0$$

Female employee has lower chances of high income than male employee.

Is this enough to conclude that females are systematically discriminated again in companies?



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Examples

Example 2 (Government Census)

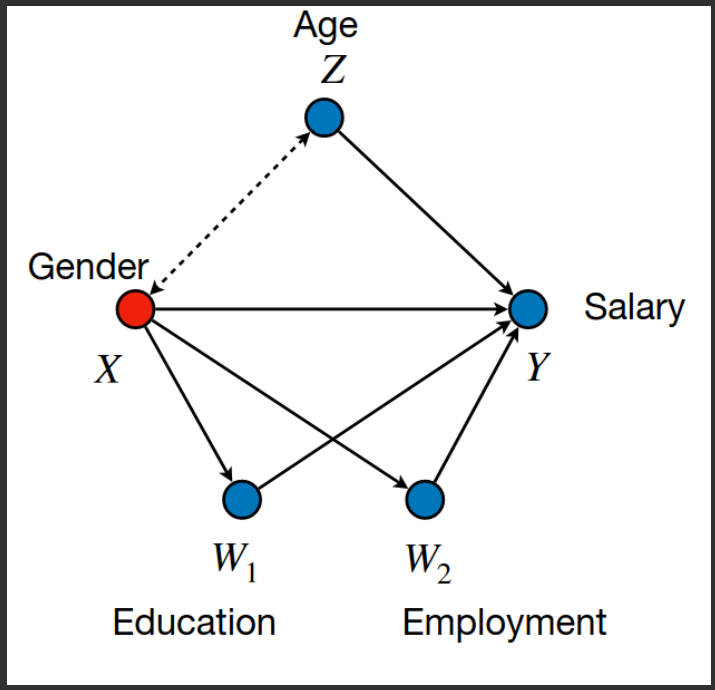
The US census data records a person's yearly salary (in tens of thousands of \$). The census also records age (Z), gender ($X = 0$ for male,$X = 1$ for female), education level (W_1) and employment status (W_2, 10 job types).

Data Analysis reveals:

$$TV_{x_0,x_1}(Y) = \mathbf{E}[Y|x_1] - \mathbf{E}[Y|x_0] < 0$$

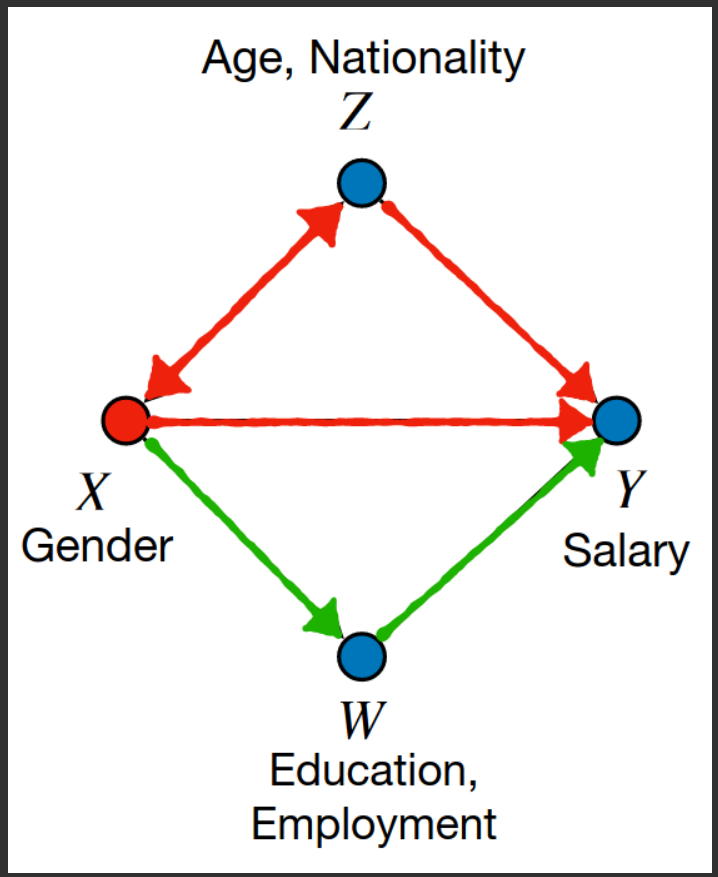
Female employee has lower chances of high income than male employee.

Is this enough to conclude that females are systematically discriminated again in companies?



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Examples



$$TV = E[Y|male] - E[Y|female] > 0.$$

How could the observed disparity be explained?

(1) The salary decision is based on employee's gender: $X \to Y$

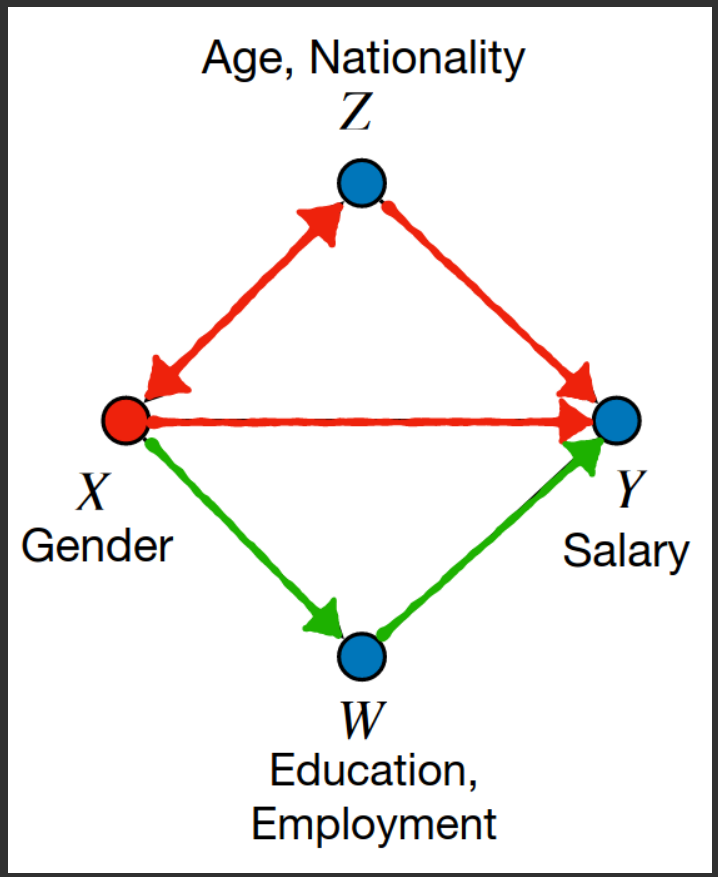
(2) Decisions were based on education or employment: $X \to W \to Y$

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \to Y$

As a legal argument, the jury may be okay with Y's variations due to education, but not okay with the variations due to gender or age.

[5] Figure and examples from Causal Fairness Analysis, Drago Plecko andElias Bareinboim [5].

Examples



$TV = E[Y|male] - E[Y|female] > 0.$

How could the observed disparity be explained?

(1) The salary decision is based on employee's gender: $X \to Y$

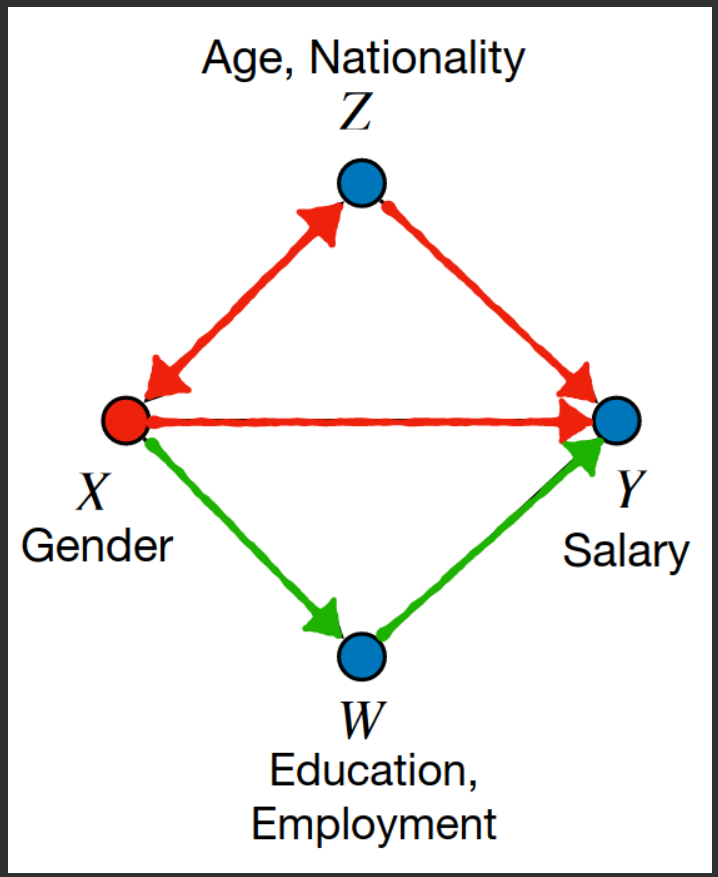
(2) Decisions were based on education or employment: $X \to W \to Y$

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \to Y$

As a legal argument, the jury may be okay with Y's variations due to education, but not okay with the variations due to gender or age.

[5] Figure and examples from Causal Fairness Analysis, Drago Plecko andElias Bareinboim [5].

Examples



$TV = E[Y|male] - E[Y|female] > 0.$

How could the observed disparity be explained?

(1) The salary decision is based on employee's gender: $X \rightarrow Y$

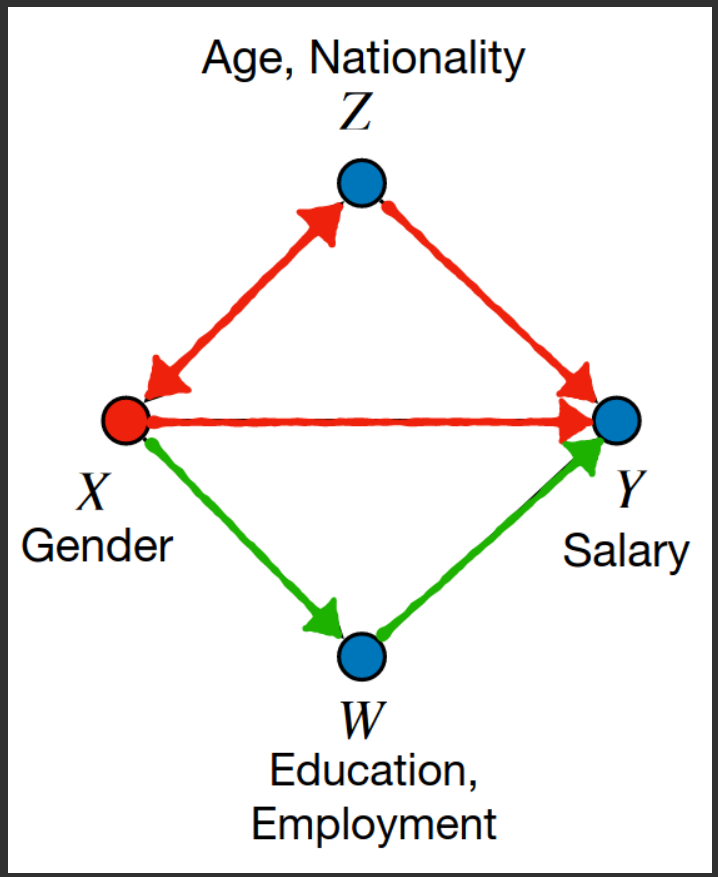
(2) Decisions were based on education or employment: $X \rightarrow W \rightarrow Y$

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \rightarrow Y$

As a legal argument, the jury may be okay with Y's variations due to education, but not okay with the variations due to gender or age.

[5] Figure and examples from Causal Fairness Analysis, Drago Plecko andElias Bareinboim [5].

Examples

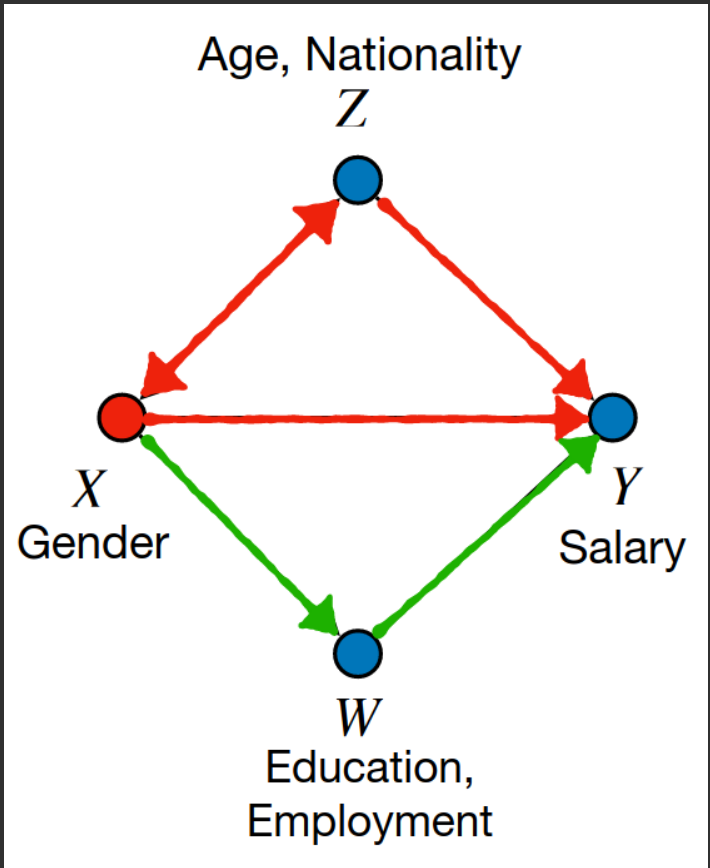


$TV = E[Y|male] - E[Y|female] > 0.$
How could the observed disparity be explained?
(1) The salary decision is based on employee's gender: $X \rightarrow Y$
(2) Decisions were based on education or employment: $X \rightarrow W \rightarrow Y$
(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \rightarrow Y$

As a legal argument, the jury may be okay with Y's variations due to education, but not okay with the variations due to gender or age.

[5] Figure and examples from Causal Fairness Analysis, Drago Plecko andElias Bareinboim [5].

Examples



$$TV = E[Y|male] - E[Y|female] > 0.$$

How could the observed disparity be explained?

(1) The salary decision is based on employee's gender: $X \to Y$

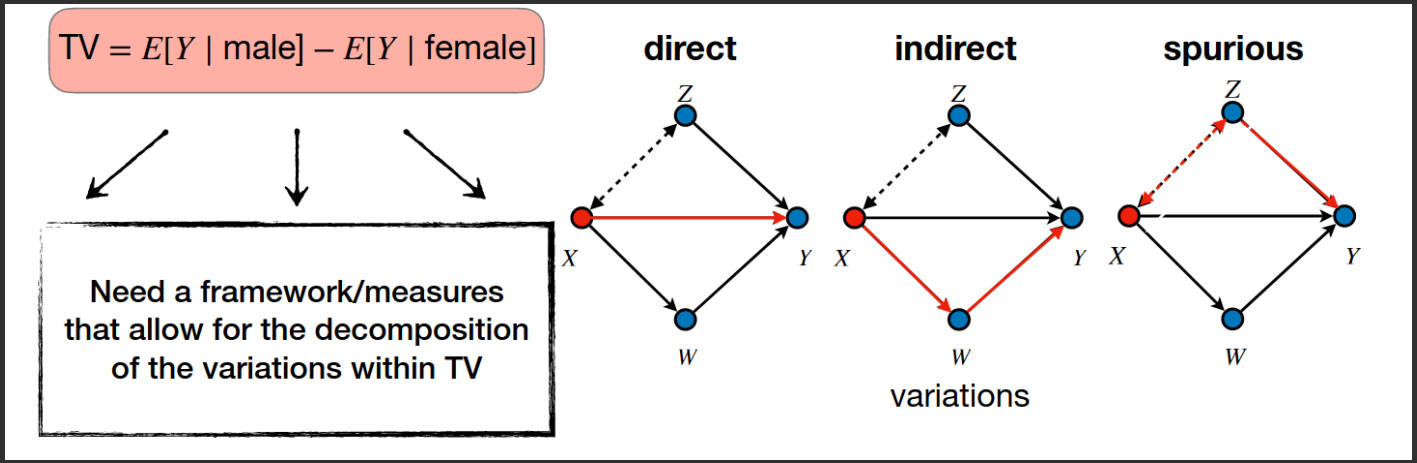
(2) Decisions were based on education or employment: $X \to W \to Y$

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \to Y$

As a legal argument, the jury may be okay with Y's variations due to education, but not okay with the variations due to gender or age.

[5] Figure and examples from Causal Fairness Analysis, Drago Plecko andElias Bareinboim [5].

Examples

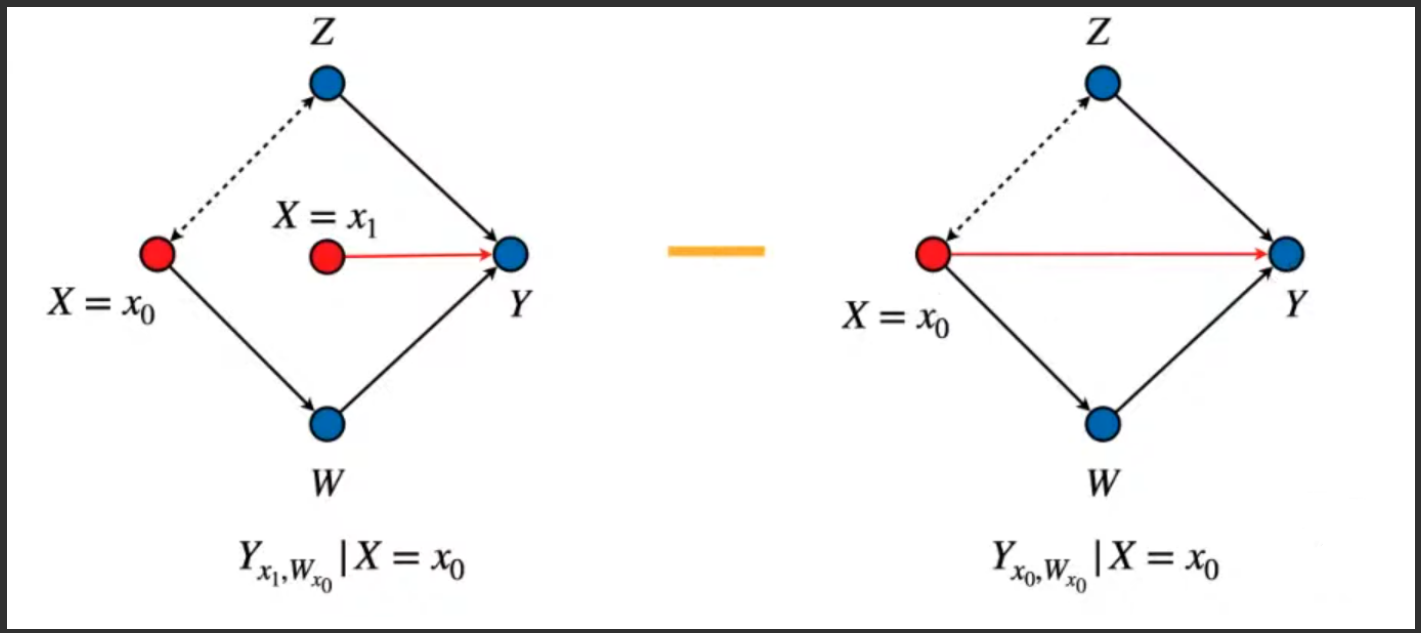


[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Counterfactual Direct Effect

For a male employee $(X = x_0)$, how would his salary (Y) change had he been a female $(X = x_1)$, while keeping the age, nationality, education, employment status unchanged (i.e., at the natural level $X = x_0$)?

$$DE_{x_0,x_1}(y|x) = E[y_{x_1,W_{x_0}}|x] - E[y_{x_0,W_{x_0}}|x]$$

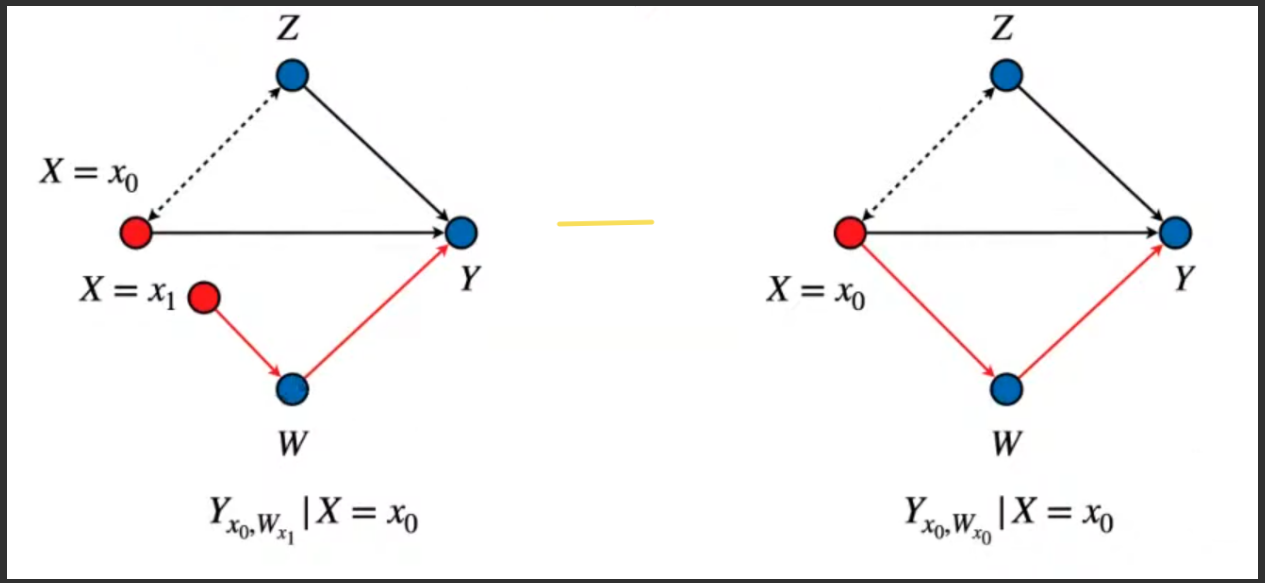


[5] Figure and examples from Causal Fairness Analysis, Drago Plecko andElias Bareinboim [5].

Counterfactual Indirect Effect

For a male employee $(X = x_0)$, how would his salary (Y) change had his education and employment status had been that of a female person $X = x_0$, while keep age, nationality and gender unchanged (i.e., at the natural level $X = x_0$)?

$$IE_{x_0,x_1}(y|x) = E[y_{x_0,W_{x_1}}|x] - E[y_{x_0,W_{x_0}}|x]$$

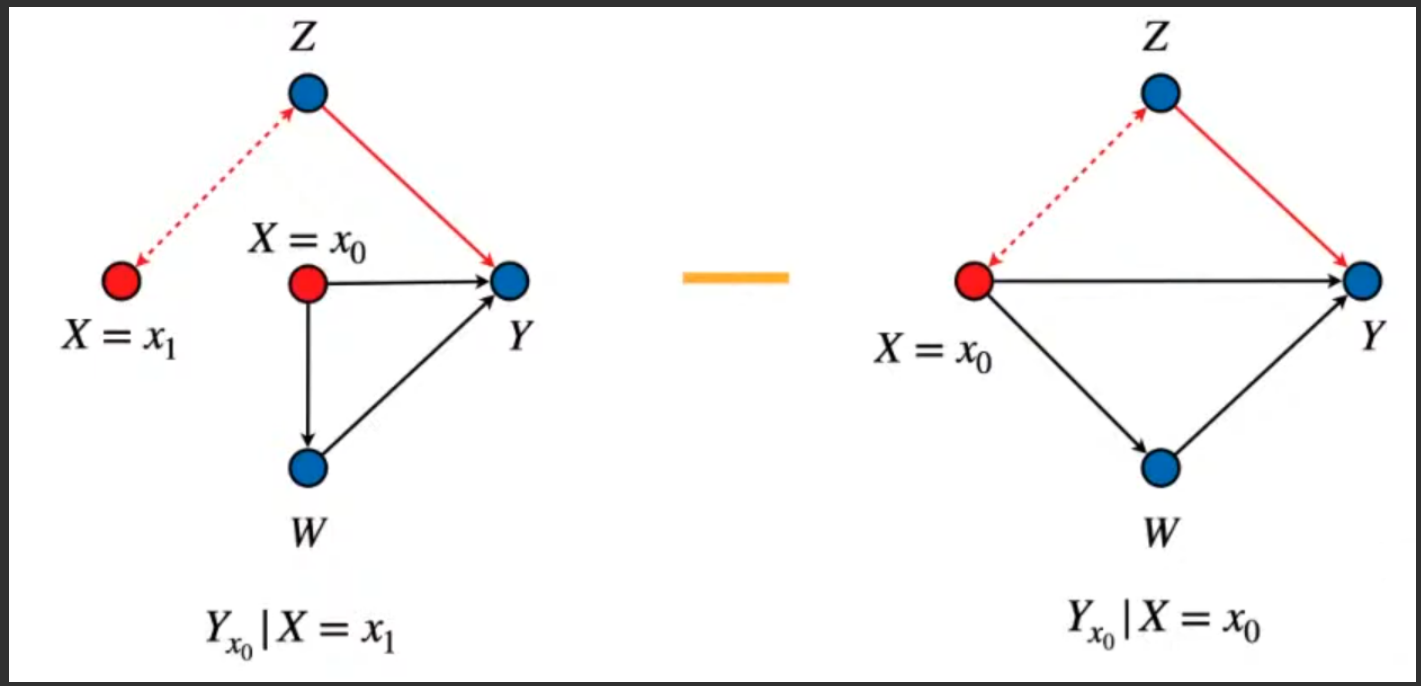


[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Counterfactual Spurious Effect

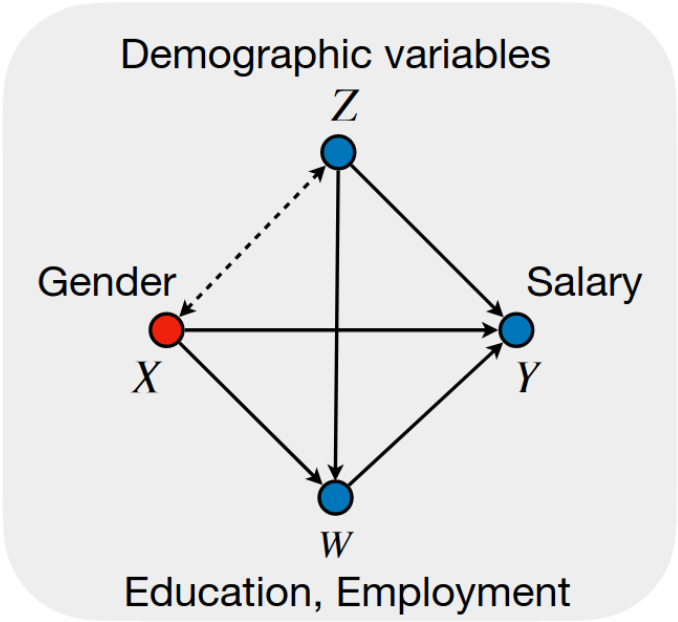
How would an individuals salary (Y) change if their gender is set to male (or female) by intervention, compared to observing their salary as male (female)?

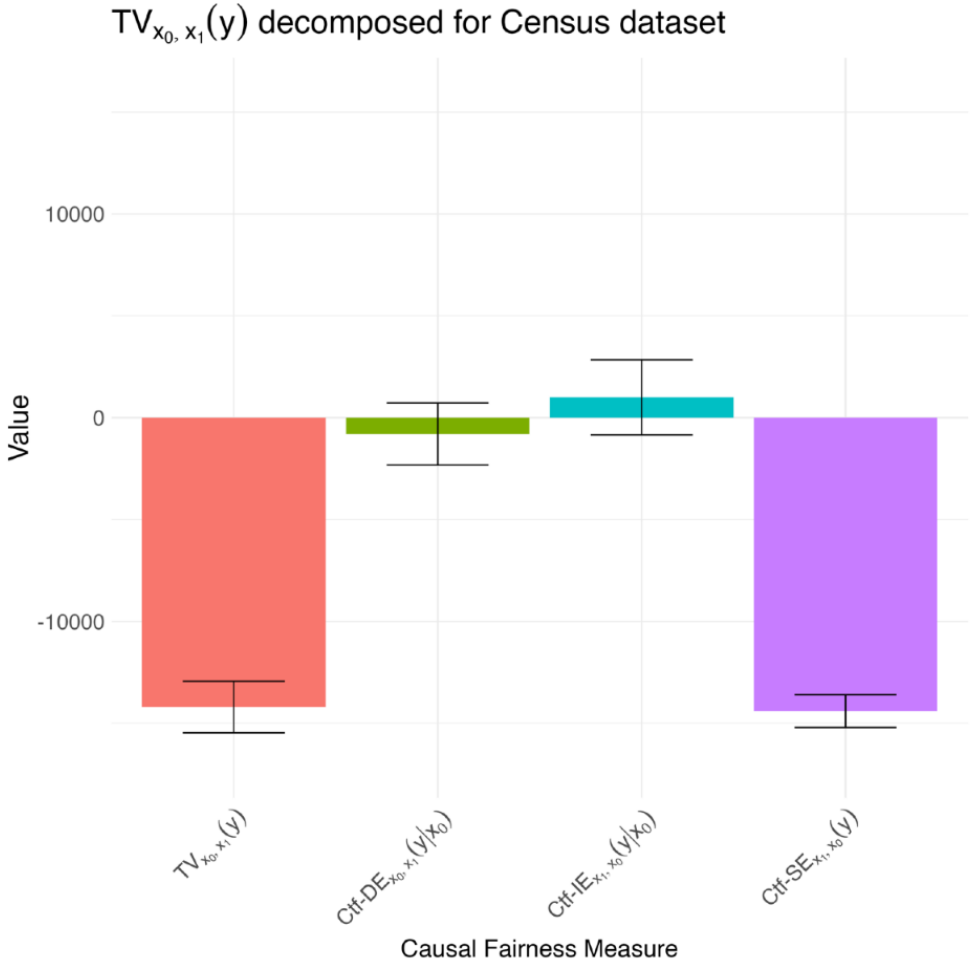
$$SE_{x_0,x_1}(y|x) = E[y_{x_0}|x_1] - E[y_{x_0}|x_0]$$



[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Census 2018 Dataset





- Observed disparity:
  $TV_{x_0,x_1}(y) = \$14{,}000/\text{year}$

[5] Figure and examples from Causal Fairness Analysis, Drago Plecko and Elias Bareinboim [5].

Summary and Conclusion

- We saw association is not causation. Some basic causal graphs which can help in causal reasoning.
- We covered Bayesian nets which are a good starting point to causal graphs.
- We formally defined causal graphs and introduced intervention and counterfactual reasoning.
  - Fairness.
    Evaluate the fairness of prediction models dependent on sensitive attributes of interest.
  - Invariant Feature mapping.
    Uses intervention samples and tries to find causal parents of the target.
  - Causal Generative Modeling.
    Generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables.
  - Causal Reinforcement Learning.

Summary and Conclusion

- We saw association is not causation. Some basic causal graphs which can help in causal reasoning.
- We covered Bayesian nets which are a good starting point to causal graphs.
- We formally defined causal graphs and introduced intervention and counterfactual reasoning.
  - Fairness.
    Evaluate the fairness of prediction models dependent on sensitive attributes of interest.
  - Invariant Feature mapping.
    Uses intervention samples and tries to find causal parents of the target.
  - Causal Generative Modeling.
    Generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables.
  - Causal Reinforcement Learning.

Summary and Conclusion

- We saw association is not causation. Some basic causal graphs which can help in causal reasoning.
- We covered Bayesian nets which are a good starting point to causal graphs.
- We formally defined causal graphs and introduced intervention and counterfactual reasoning.
  - Fairness.
    Evaluate the fairness of prediction models dependent on sensitive attributes of interest.
  - Invariant Feature mapping.
    Uses intervention samples and tries to find causal parents of the target.
  - Causal Generative Modeling.
    Generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables.
  - Causal Reinforcement Learning.

Summary and Conclusion

- We saw association is not causation. Some basic causal graphs which can help in causal reasoning.
- We covered Bayesian nets which are a good starting point to causal graphs.
- We formally defined causal graphs and introduced intervention and counterfactual reasoning.
  - Fairness.
    Evaluate the fairness of prediction models dependent on sensitive attributes of interest.
  - Invariant Feature mapping.
    Uses intervention samples and tries to find causal parents of the target.
  - Causal Generative Modeling.
    Generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables.
  - Causal Reinforcement Learning.

Summary and Conclusion

- We saw association is not causation. Some basic causal graphs which can help in causal reasoning.
- We covered Bayesian nets which are a good starting point to causal graphs.
- We formally defined causal graphs and introduced intervention and counterfactual reasoning.
  - Fairness.
    Evaluate the fairness of prediction models dependent on sensitive attributes of interest.
  - Invariant Feature mapping.
    Uses intervention samples and tries to find causal parents of the target.
  - Causal Generative Modeling.
    Generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables.
  - Causal Reinforcement Learning.

Summary and Conclusion

- We saw association is not causation. Some basic causal graphs which can help in causal reasoning.
- We covered Bayesian nets which are a good starting point to causal graphs.
- We formally defined causal graphs and introduced intervention and counterfactual reasoning.
  - Fairness.
    Evaluate the fairness of prediction models dependent on sensitive attributes of interest.
  - Invariant Feature mapping.
    Uses intervention samples and tries to find causal parents of the target.
  - Causal Generative Modeling.
    Generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables.
  - Causal Reinforcement Learning.

Summary and Conclusion

- We saw association is not causation. Some basic causal graphs which can help in causal reasoning.
- We covered Bayesian nets which are a good starting point to causal graphs.
- We formally defined causal graphs and introduced intervention and counterfactual reasoning.
  - Fairness.
    Evaluate the fairness of prediction models dependent on sensitive attributes of interest.
  - Invariant Feature mapping.
    Uses intervention samples and tries to find causal parents of the target.
  - Causal Generative Modeling.
    Generate counterfactual samples that consider causal dependencies between the attributes of interest and the other generative variables.
  - Causal Reinforcement Learning.

The End