Counterfactual Reasoning

Simon Ferreira simon.ferreira@sorbonne-universite.fr

L'Institut Pierre Louis d'Epidémiologie et de Santé Publique, Inserm, Sorbonne Université

June, 2026

Summary

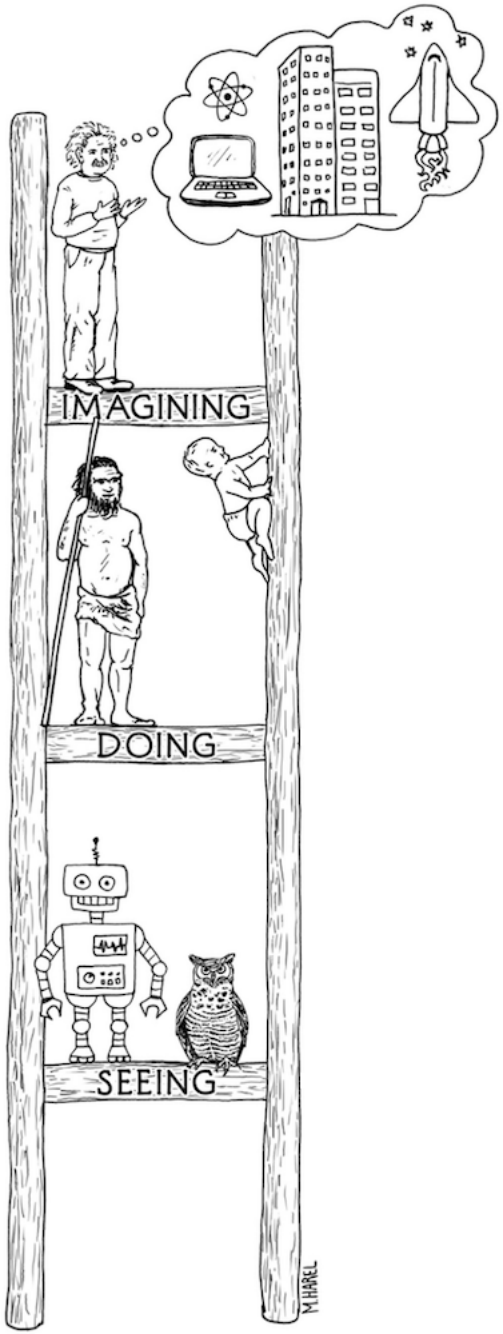
Inserm

iPLesp

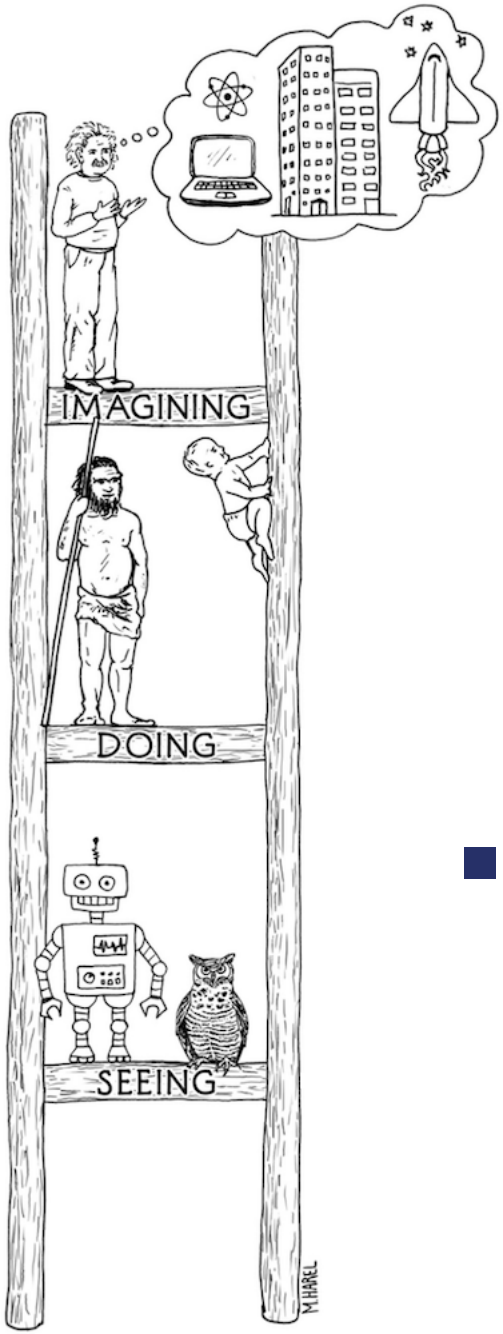
SANTÉ SORBONNE UNIVERSITÉ

1

The Causal Hierarchy

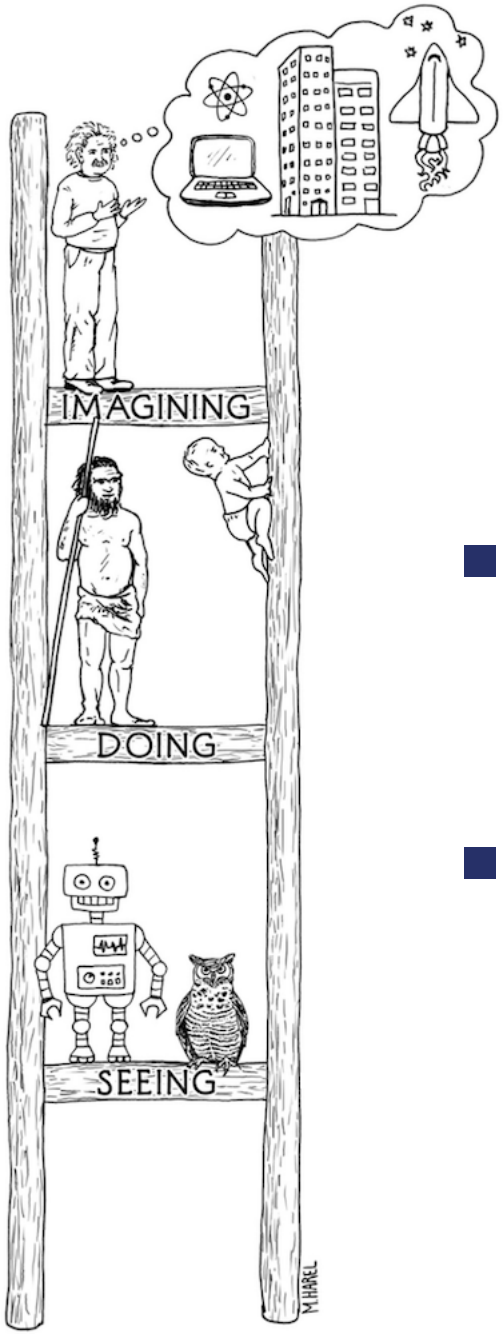


Pearl 2009



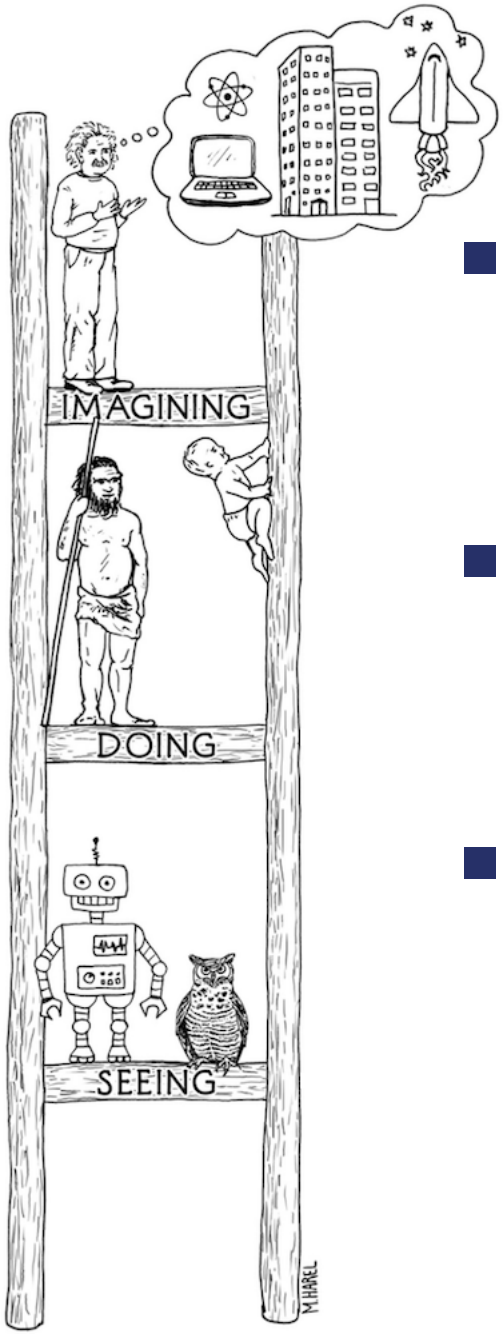
- Associations
  - What if I see…?
  - observational data
  - Tools: statistics, correlation

Pearl 2009



- Interventions
  - Questions: What if I do…? How?
  - interventional data
  - Tools: RCTs, DAGs, do-calculus
- Associations
  - What if I see…?
  - observational data
  - Tools: statistics, correlation

Pearl 2009



- Counterfactuals
  - Questions: What if I had done…? Why?
  - Generating process (SCM)
  - Tools: Twin Networks, SWIGs, AMWN
- Interventions
  - Questions: What if I do…? How?
  - interventional data
  - Tools: RCTs, DAGs, do-calculus
- Associations
  - What if I see…?
  - observational data
  - Tools: statistics, correlation

Pearl 2009

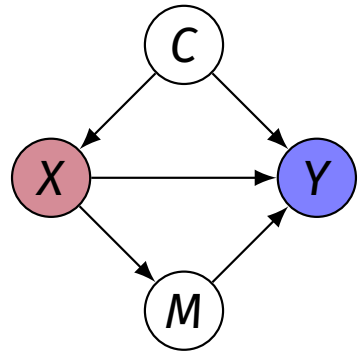
$$\xi_C, \xi_X, \xi_M, \xi_Y$$

$$C := \xi_C$$

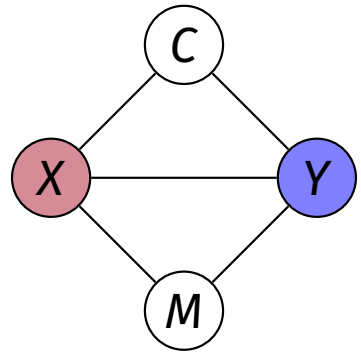
$$X := f_X(C, \xi_X)$$

$$M := f_M(X, \xi_M)$$

$$Y := f_Y(C, X, M, \xi_Y)$$



$$\xi_C, \xi_X, \xi_M, \xi_Y$$
$$C := \xi_C$$
$$X := f_X(C, \xi_X)$$
$$M := f_M(X, \xi_M)$$
$$Y := f_Y(C, X, M, \xi_Y)$$



- Associations:
  $\Pr(Y = y \mid X = x)$

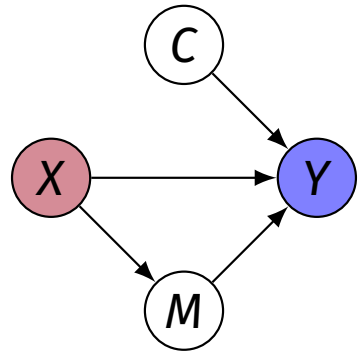
$$\xi_C, \xi_X, \xi_M, \xi_Y$$

$$C := \xi_C$$

$$X := x$$

$$M := f_M(X, \xi_M)$$

$$Y := f_Y(C, X, M, \xi_Y)$$



- Interventions:
  $\Pr(Y = y \mid \mathrm{do}(X = x))$
- Associations:
  $\Pr(Y = y \mid X = x)$

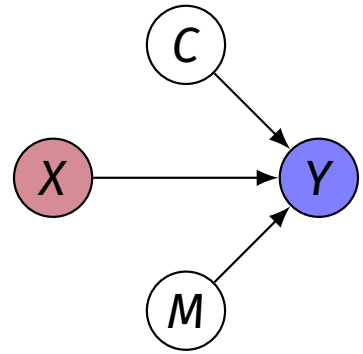
$$\xi_C, \xi_X, \xi_M, \xi_Y$$

$$C := \xi_C$$

$$X := x$$

$$M := f_M(x', \xi_M)$$

$$Y := f_Y(C, X, M, \xi_Y)$$



- Counterfactuals:
  $\Pr\left(Y = y \mid \text{do}\left(X = x\right), \text{do}\left(M = m_{\text{do}(X=x')}\right)\right)$
- Interventions:
  $\Pr\left(Y = y \mid \text{do}\left(X = x\right)\right)$
- Associations:
  $\Pr\left(Y = y \mid X = x\right)$

$$
\begin{aligned}
&\xi_C, \xi_X, \xi_M, \xi_Y \\
C &:= \xi_C \\
X &:= f_X(C, \xi_X) \\
M &:= f_M(X, \xi_M) \\
Y &:= f_Y(C, X, M, \xi_Y)
\end{aligned}
$$

$$
\begin{aligned}
&\xi_C, \xi_X, \xi_M, \xi_Y \\
C &:= \xi_C \\
X_1 &:= x' \\
M &:= f_M(X_1, \xi_M) \\
X_2 &:= x \\
Y &:= f_Y(C, X_2, M, \xi_Y)
\end{aligned}
$$

Counterfactuals:

$$\Pr\left(Y = y \mid \mathrm{do}\left(X = x\right), \mathrm{do}\left(M = m_{\mathrm{do}(X=x')}\right)\right)$$

Inserm

iPLesp

SANTÉ SORBONNE UNIVERSITÉ

2

Examples of Counterfactuals

Examples of counterfactuals

- Effect of the treatment on the treated:
  $\mathbb{E}\left(Y_{\text{do}(X=1)} \mid X = 1\right) - \mathbb{E}\left(Y_{\text{do}(X=0)} \mid X = 1\right)$

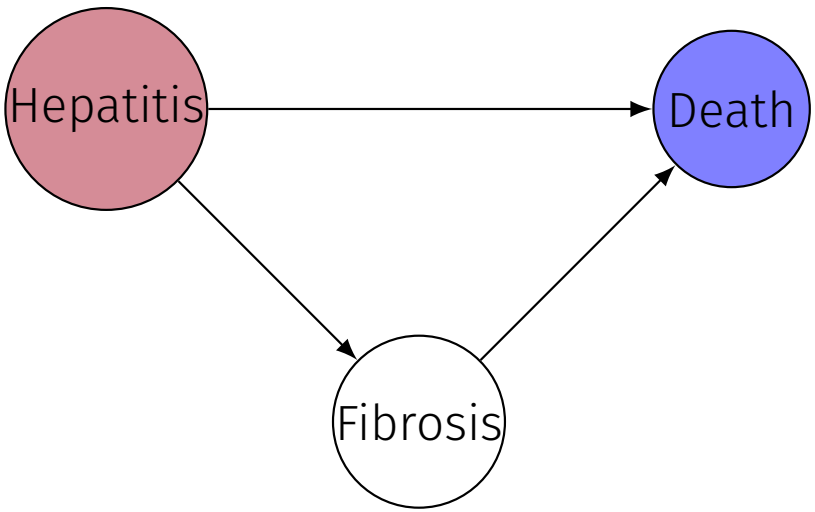
Examples of counterfactuals

- Effect of the treatment on the treated:
  $\mathbb{E}\left(Y_{\text{do}(X=1)} \mid X = 1\right) - \mathbb{E}\left(Y_{\text{do}(X=0)} \mid X = 1\right)$
- Individual effect: "Bob did not take the treatment and died. Would he have lived had he taken the treatment?"

Examples of counterfactuals

- Effect of the treatment on the treated:
  $\mathbb{E}\left(Y_{\text{do}(X=1)} \mid X = 1\right) - \mathbb{E}\left(Y_{\text{do}(X=0)} \mid X = 1\right)$
- Individual effect: "Bob did not take the treatment and died. Would he have lived had he taken the treatment?"
- Natural Direct Effect

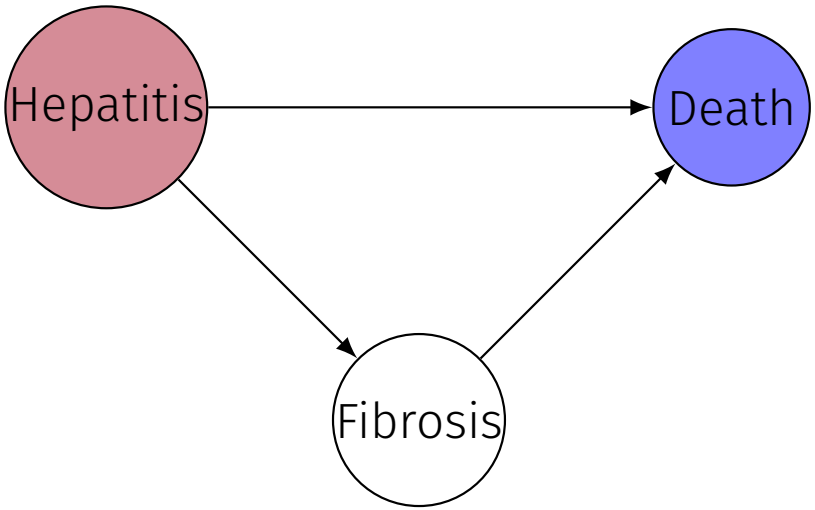
The Natural Direct Effect



Direct Effect: How will a change in the exposure (Hepatitis) modify the outcome (Death) if the mediator (Fibrosis) is kept constant?

Robins and Greenland 1992; Pearl 2001

The Natural Direct Effect



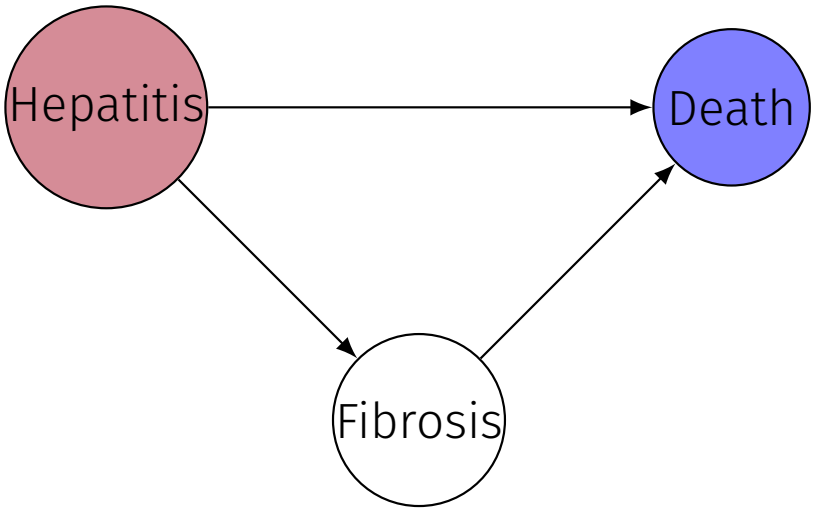
Direct Effect: How will a change in the exposure (Hepatitis) modify the outcome (Death) if the mediator (Fibrosis) is kept constant?

$$CDE(D, H_{0\to 1}, F = f) = \mathbb{E}\left(D \mid \text{do}\left(H = 1\right), \text{do}\left(F = f\right)\right) - \mathbb{E}\left(D \mid \text{do}\left(H = 0\right), \text{do}\left(F = f\right)\right)$$

This is called the controlled direct effect.

Robins and Greenland 1992; Pearl 2001

The Natural Direct Effect



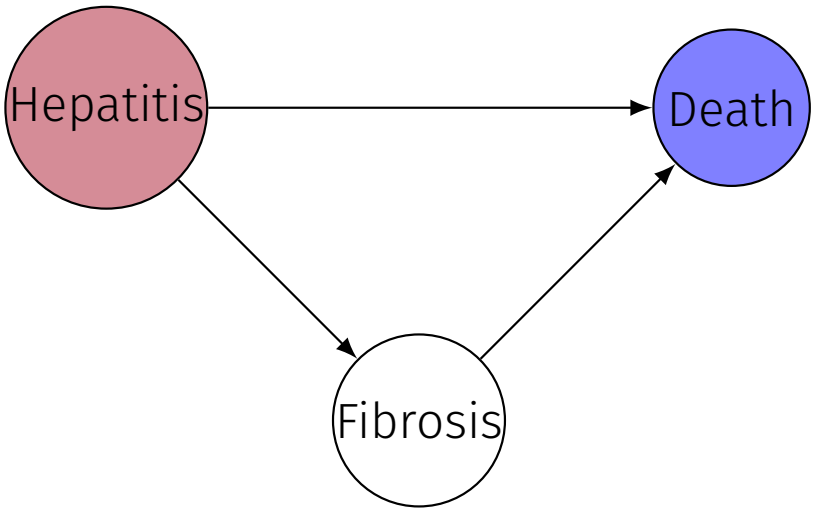
Direct Effect: How will a change in the exposure (Hepatitis) modify the outcome (Death) if the mediator (Fibrosis) is kept constant?

$$CDE(D, H_{0\to 1}, F = f) = \mathbb{E}\left(D \mid \text{do}\left(H = 1\right), \text{do}\left(F = f\right)\right) - \mathbb{E}\left(D \mid \text{do}\left(H = 0\right), \text{do}\left(F = f\right)\right)$$

This is called the controlled direct effect.
Problem: This definition depends on the value of f. Thus there exists as many values of CDE as F has possible values.

Robins and Greenland 1992; Pearl 2001

The Natural Direct Effect

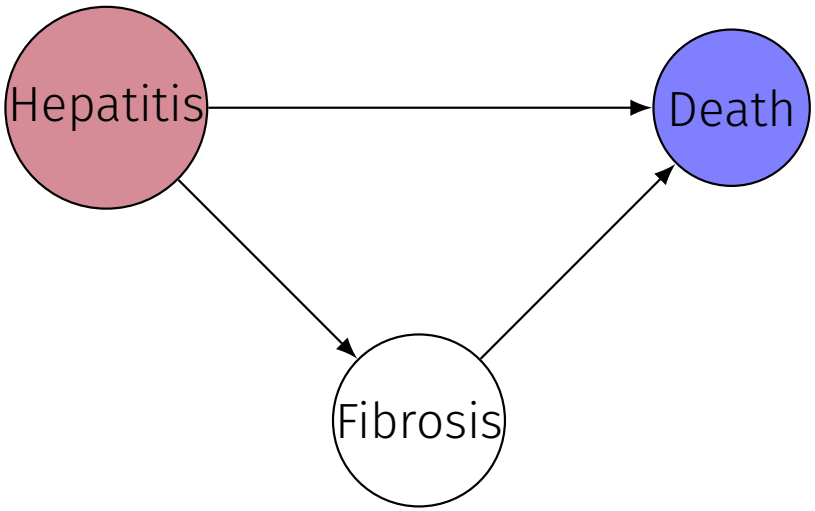


Direct Effect: How will a change in the exposure (Hepatitis) modify the outcome (Death) if the mediator (Fibrosis) is kept constant?

What is the natural value for the mediator Fibrosis?

Robins and Greenland 1992; Pearl 2001

The Natural Direct Effect



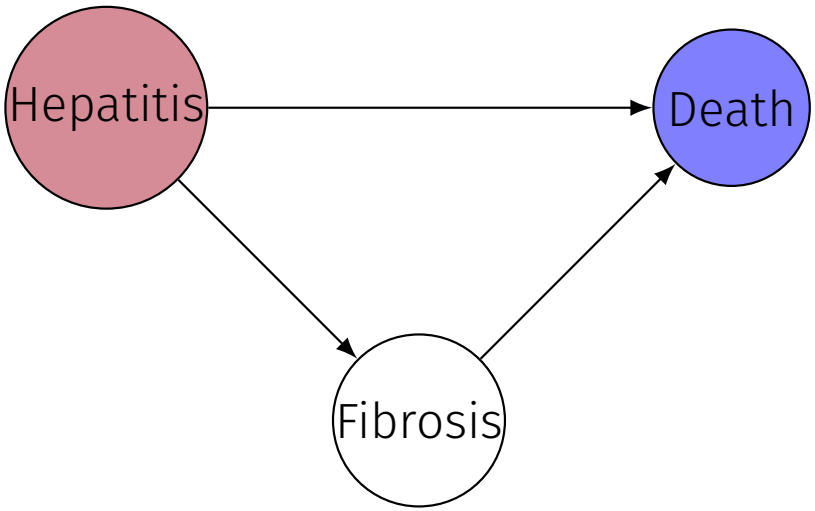
Direct Effect: How will a change in the exposure (Hepatitis) modify the outcome (Death) if the mediator (Fibrosis) is kept constant?

What is the natural value for the mediator Fibrosis?

The natural value for Fibrosis is the value it would have taken had we not modified the exposure (Hepatitis)

Robins and Greenland 1992; Pearl 2001

The Natural Direct Effect



Direct Effect: How will a change in the exposure (Hepatitis) modify the outcome (Death) if the mediator (Fibrosis) is kept constant?

What is the natural value for the mediator Fibrosis?

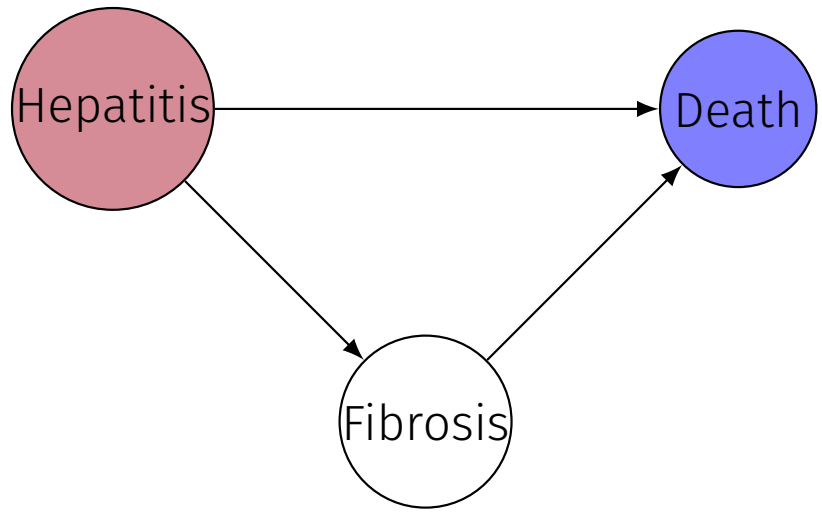
The natural value for Fibrosis is the value it would have taken had we not modified the exposure (Hepatitis)

$$NDE(D, H_{0\to 1}, F) = \mathbb{E}\left(D \mid \text{do}\left(H = 1\right), \text{do}\left(F = f_{\text{do}(H=0)}\right)\right) - \mathbb{E}\left(D \mid \text{do}\left(H = 0\right), \text{do}\left(F = f_{\text{do}(H=0)}\right)\right)$$

Robins and Greenland 1992; Pearl 2001

Toy Example

$$\xi_H, \xi_F, \xi_D \leftarrow \mathcal{U}(0,1)$$
$$H := \xi_H$$
$$F := H \vee \xi_F$$
$$D := (H \vee F) \wedge \xi_D$$



Toy Example

observational Data:

Noise			Variables		
ξ_H	ξ_F	ξ_D	H	F	D
0	0	0	0	0	0
0	0	1	0	0	0
0	1	0	0	1	0
0	1	1	0	1	1
1	0	0	1	1	0
1	0	1	1	1	1
1	1	0	1	1	0
1	1	1	1	1	1

$$\mathbb{E}(D \mid H = 1, F = 0) - \mathbb{E}(D \mid H = 0, F = 0)$$
$$= 0 - 0 = 0$$

Toy Example

interventional Data:

Noise			do $(H = 0, F = 0)$			do $(H = 1, F = 0)$		
ξ_H	ξ_F	ξ_D	H	F	D	H	F	D
0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	1	0	1
0	1	0	0	0	0	1	0	0
0	1	1	0	0	0	1	0	1
1	0	0	0	0	0	1	0	0
1	0	1	0	0	0	1	0	1
1	1	0	0	0	0	1	0	0
1	1	1	0	0	0	1	0	1

Controlled direct effect (CDE)

$$\mathbb{E}(D \mid \text{do}(H = 1, F = 0)) - \mathbb{E}(D \mid \text{do}(H = 0, F = 0))$$

$$= \frac{1}{2} - 0 = \frac{1}{2}$$

Toy Example

Couterfactual Data:

Noise			$\text{do}\left(H=0, F=f_{\text{do}(H=0)}\right)$			$\text{do}\left(H=1, F=f_{\text{do}(H=0)}\right)$		
ξ_H	ξ_F	ξ_D	H	F	D	H	F	D
0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	1	0	1
0	1	0	0	1	0	1	1	0
0	1	1	0	1	1	1	1	1
1	0	0	0	0	0	1	0	0
1	0	1	0	0	0	1	0	1
1	1	0	0	1	0	1	1	0
1	1	1	0	1	1	1	1	1

Natural direct effect:

$$\mathbb{E}\left(D \mid \text{do}\left(H=1, F=f_{\text{do}(H=0)}\right)\right) - \mathbb{E}\left(D \mid \text{do}\left(H=0, F=f_{\text{do}(H=0)}\right)\right)$$
$$= \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Inserm

iPLesp

SANTÉ SORBONNE UNIVERSITÉ

3

Twin Networks

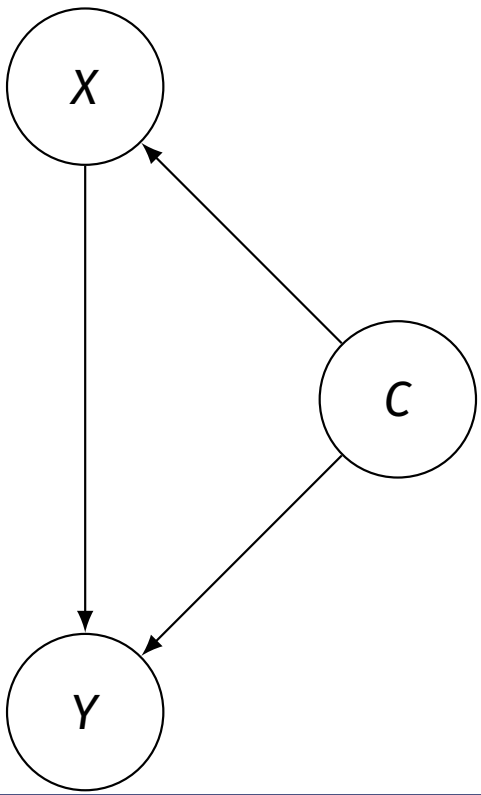
Twin Networks

- One graph for the actual world

Balke and Pearl 1994

Twin Networks

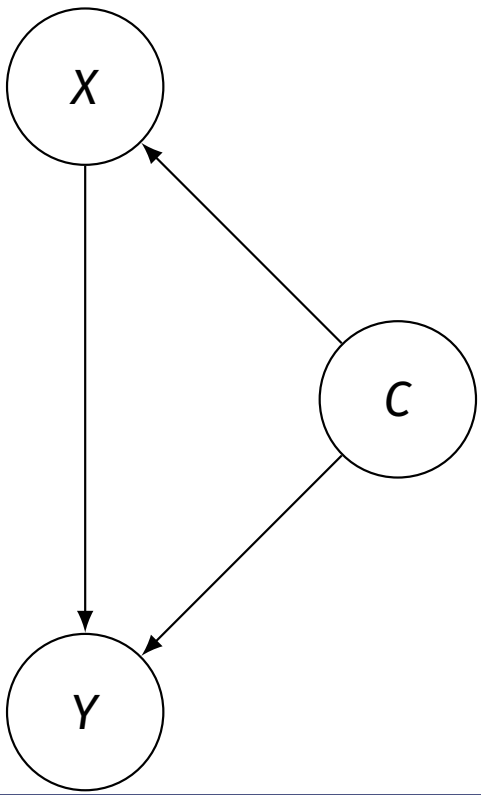
- One graph for the actual world



Balke and Pearl 1994

Twin Networks

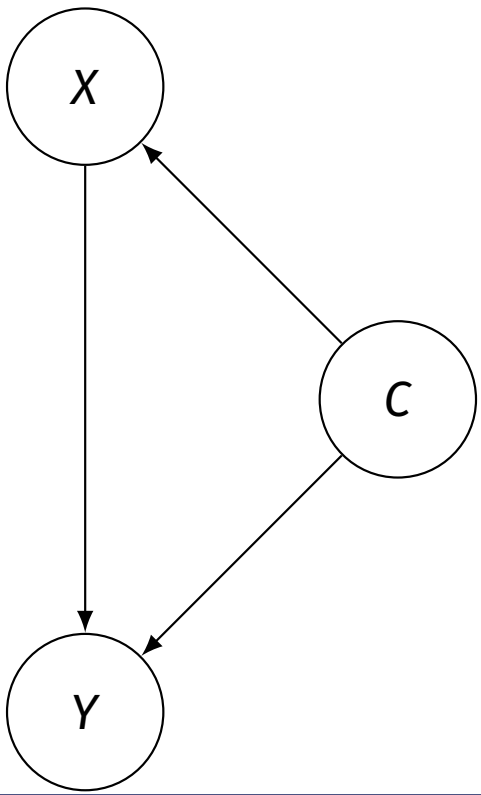
- One graph for the actual world
- One graph for the counterfactual world

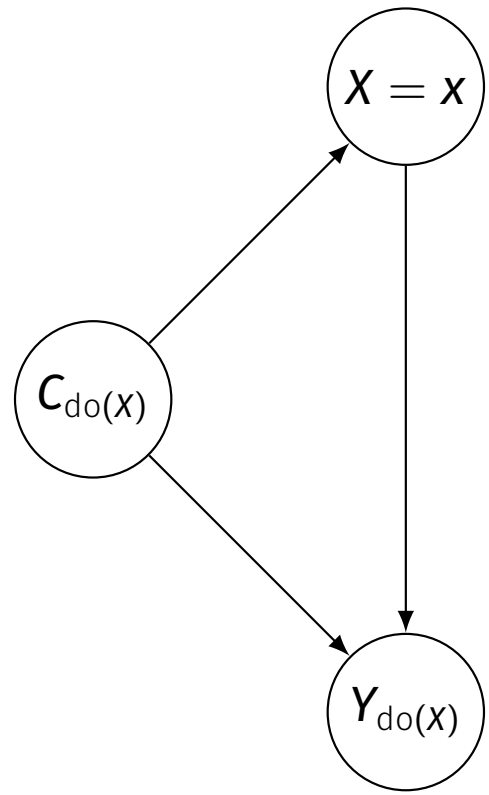


Balke and Pearl 1994

Twin Networks

- One graph for the actual world
- One graph for the counterfactual world

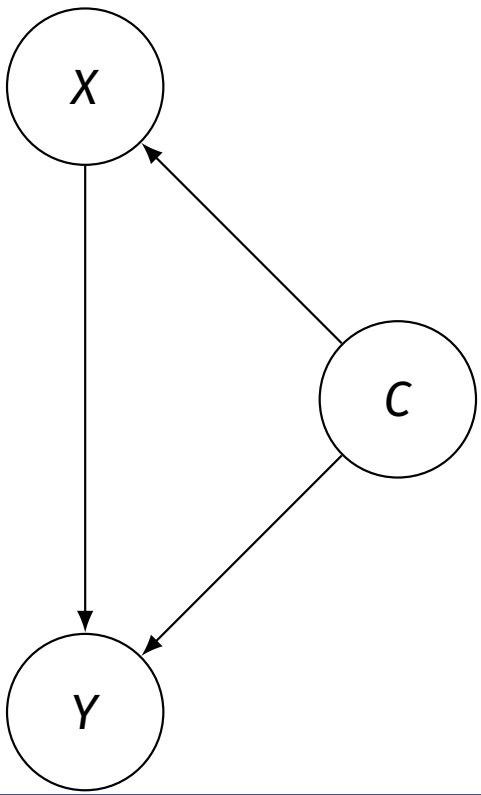


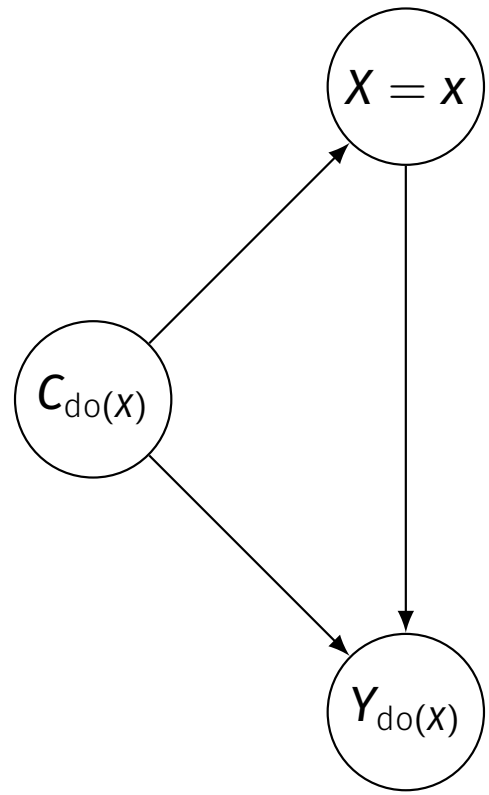


Balke and Pearl 1994

Twin Networks

- One graph for the actual world
- One graph for the counterfactual world
- Common noise variables

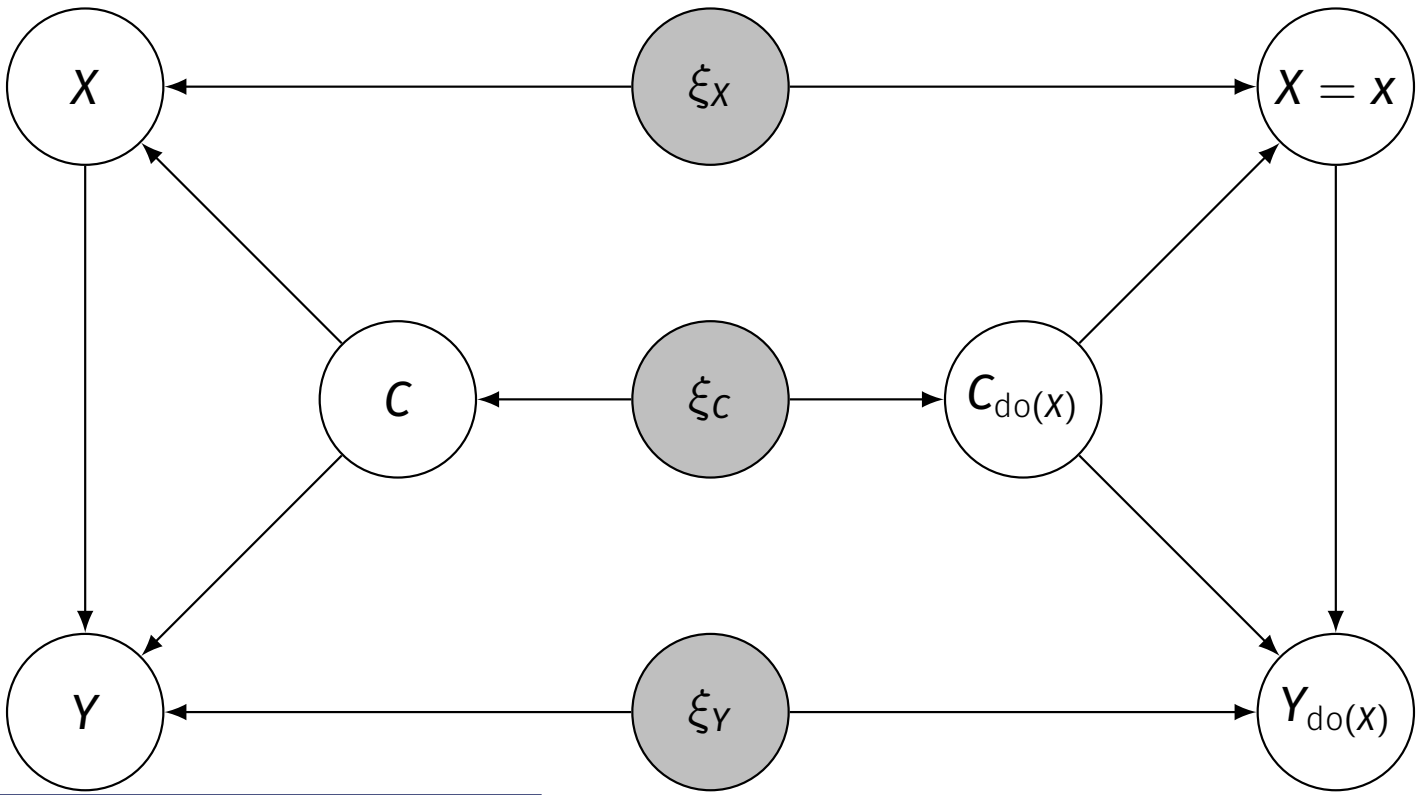




Balke and Pearl 1994

Twin Networks

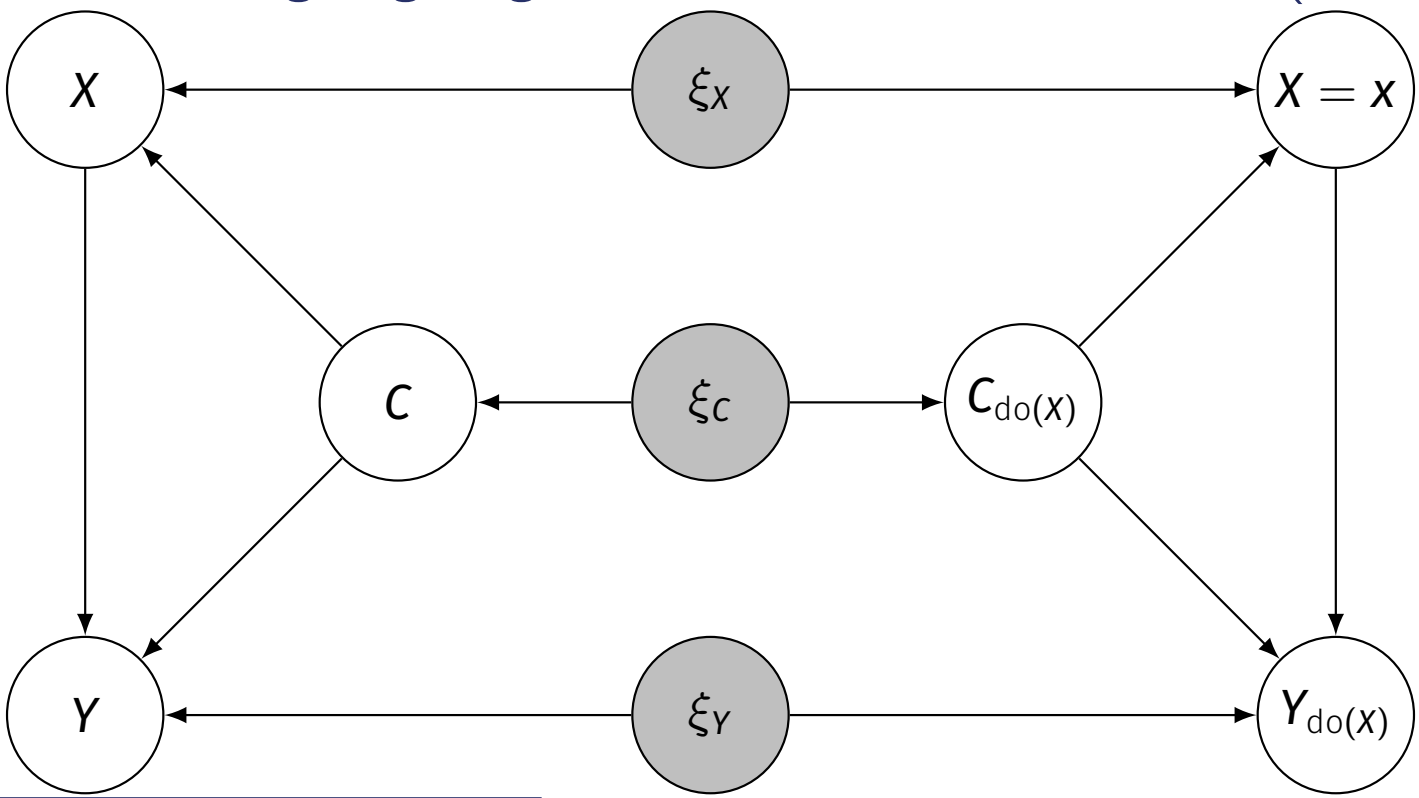
- One graph for the actual world
- One graph for the counterfactual world
- Common noise variables



Balke and Pearl 1994

Twin Networks

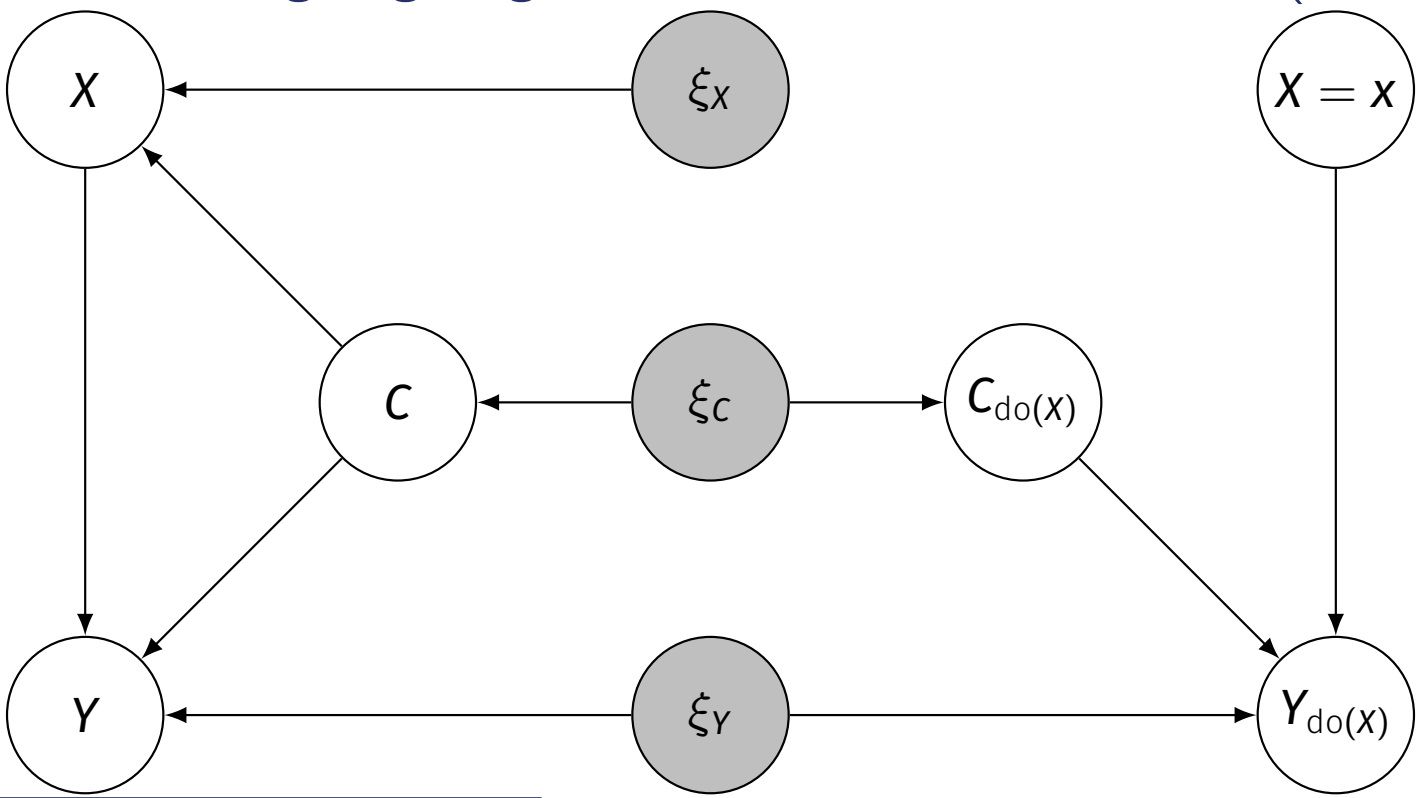
- One graph for the actual world
- One graph for the counterfactual world
- Common noise variables
- Remove edges going in the intervened node (do $(X = x)$)



Balke and Pearl 1994

Twin Networks

- One graph for the actual world
- One graph for the counterfactual world
- Common noise variables
- Remove edges going in the intervened node (do $(X = x)$)



Balke and Pearl 1994

Twin Networks

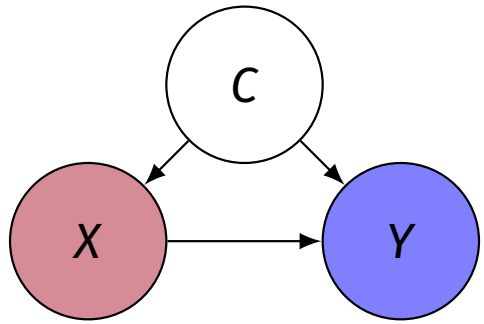
What can Twin Networks be used for?

Twin Networks

What can Twin Networks be used for?

SCM framework (Pearl)

- Assume the graph
- Backdoor criterion: C blocks all backdoor paths from X to Y.

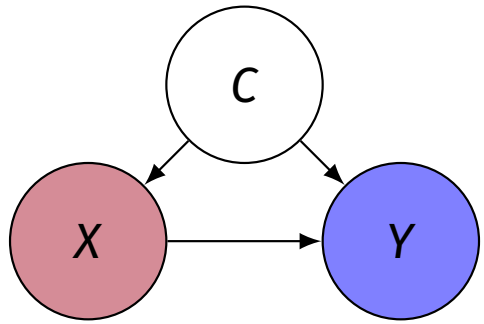


Twin Networks

What can Twin Networks be used for?

SCM framework (Pearl)

- Assume the graph
- Backdoor criterion: C blocks all backdoor paths from X to Y.



Potential Outcome framework (Robins)

- Assume conditional exchangeability

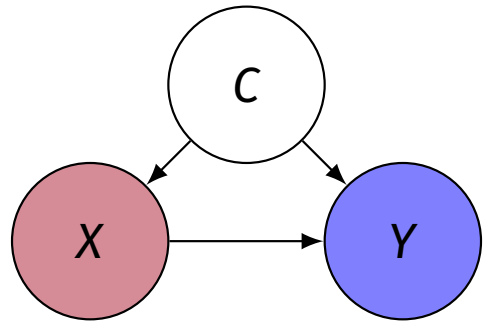
$$Y_{do(X)} \perp\!\!\!\perp_{Pr} X \mid C$$

Twin Networks

What can Twin Networks be used for?

SCM framework (Pearl)

- Assume the graph
- Backdoor criterion: C blocks all backdoor paths from X to Y.



Potential Outcome framework (Robins)

- Assume conditional exchangeability

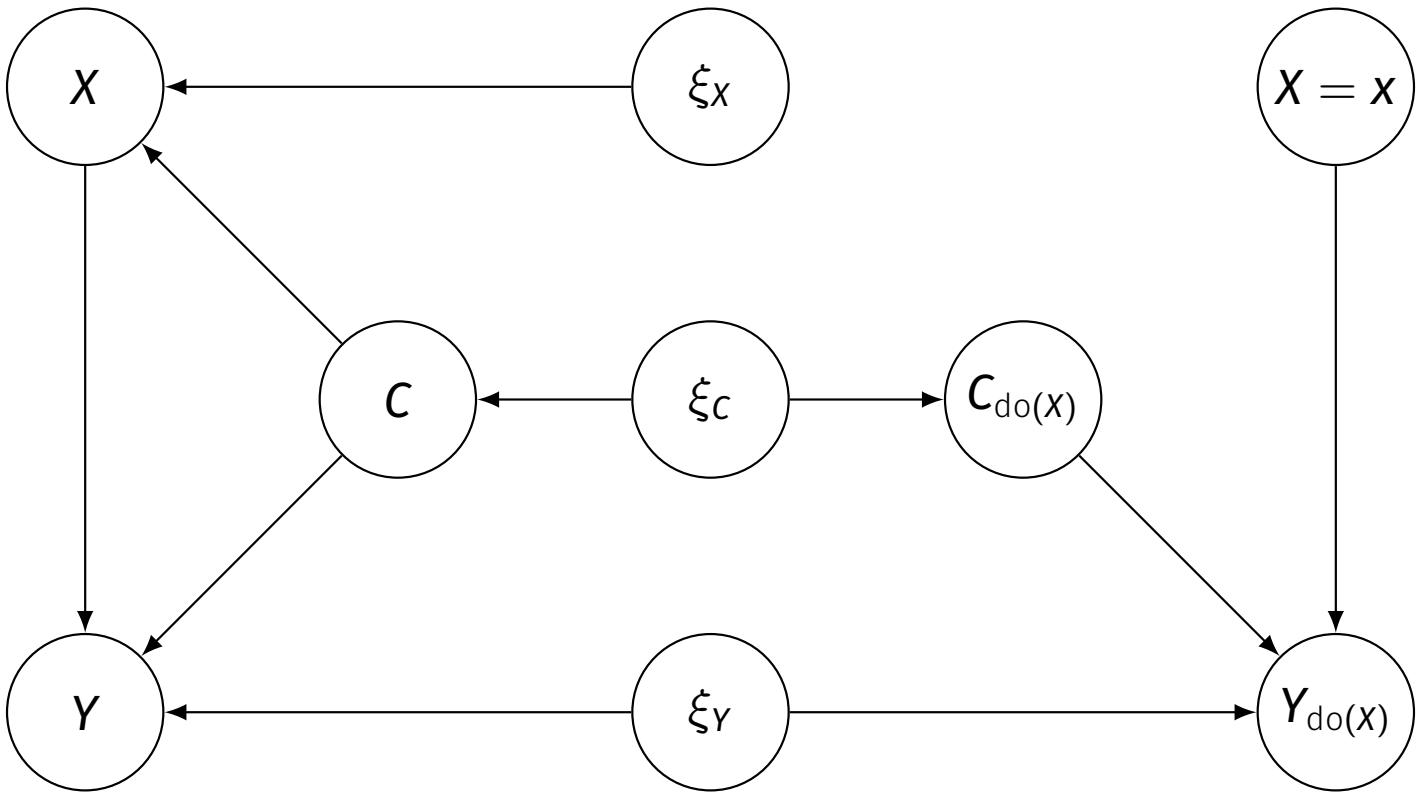
$$Y_{do(X)} \perp\!\!\!\perp_{Pr} X \mid C$$

Average Treatment Effect:

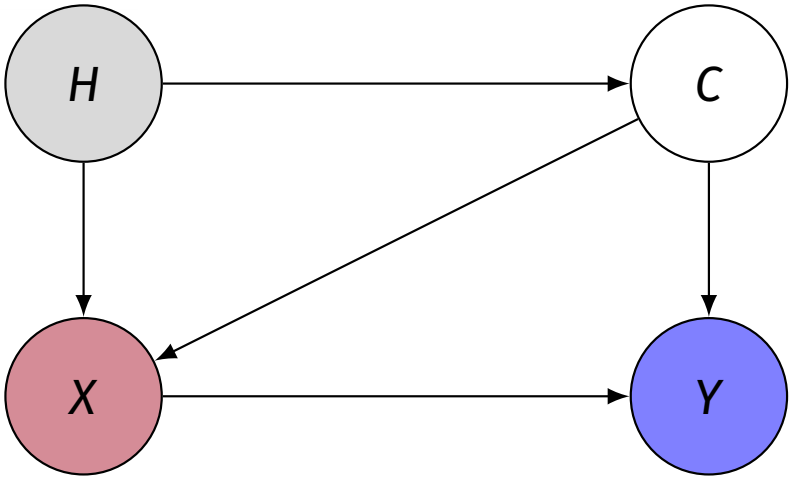
$$\sum_c \left(\mathbb{E}\left(Y \mid X = 1, C = c\right) - \mathbb{E}\left(Y \mid X = 0, C = c\right)\right) \Pr\left(C = c\right)$$

Twin Networks

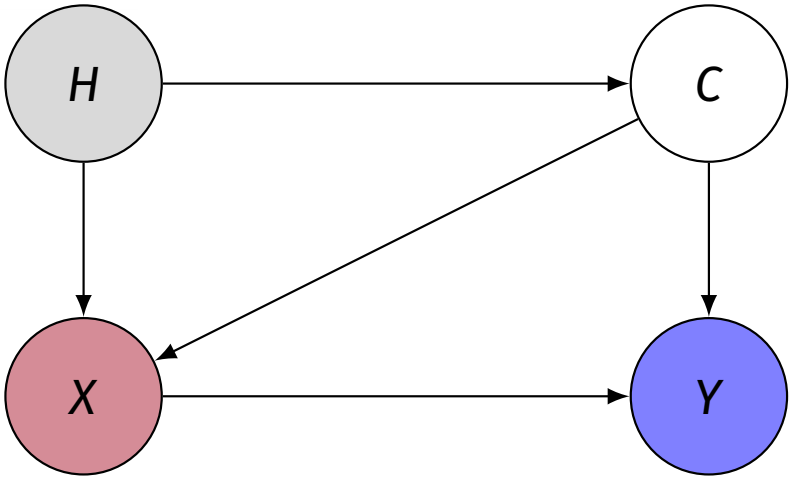
Conditional Exchangeability: $Y_{do(X)} \perp\!\!\!\perp_d X \mid C$?



Twin Networks: an example of non-completeness

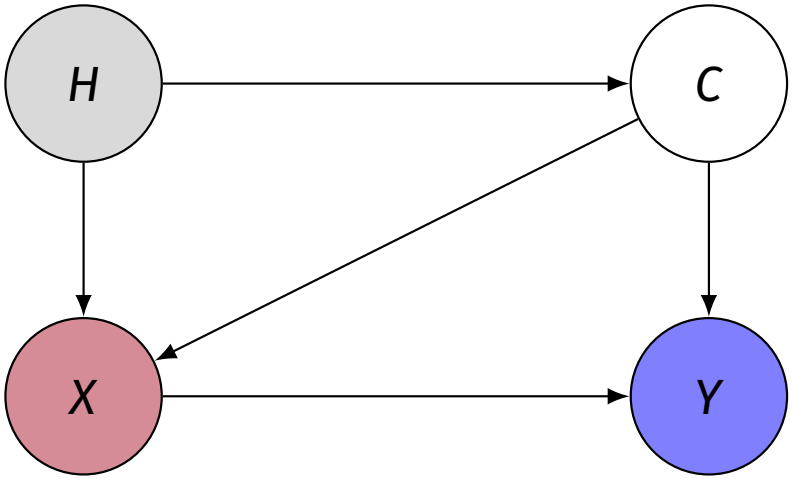


Twin Networks: an example of non-completeness

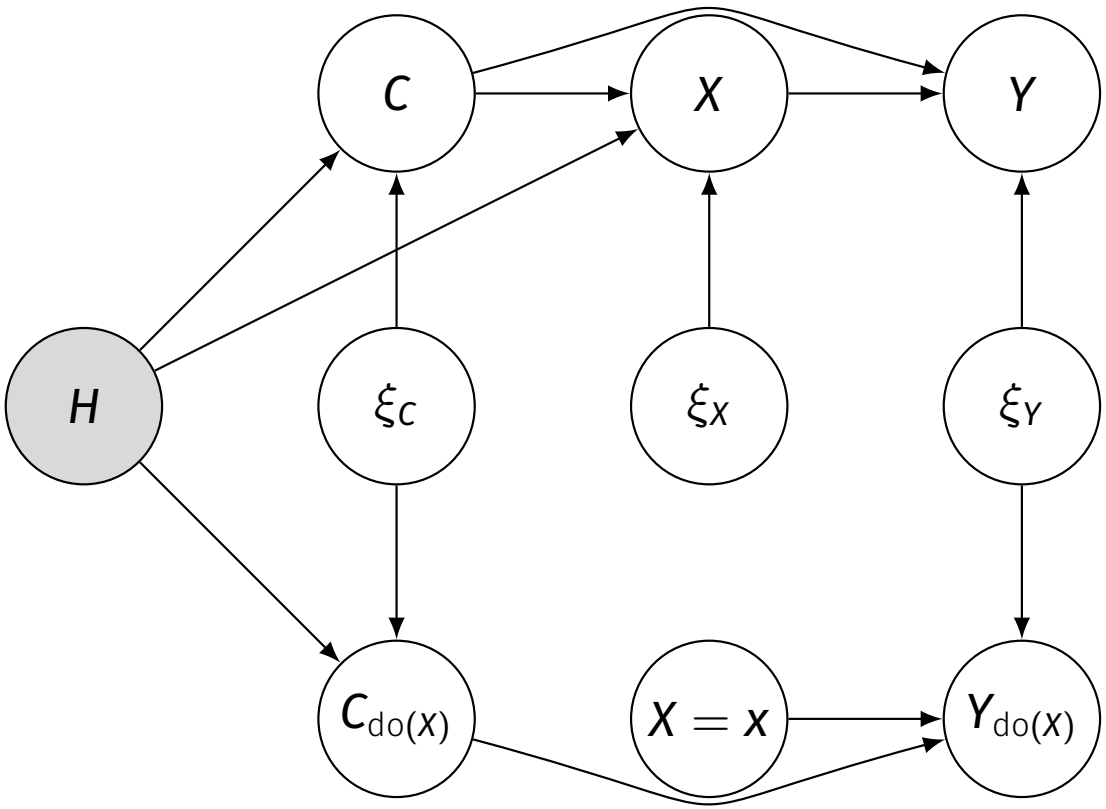


Does C verify the backdoor criterion for X,Y?

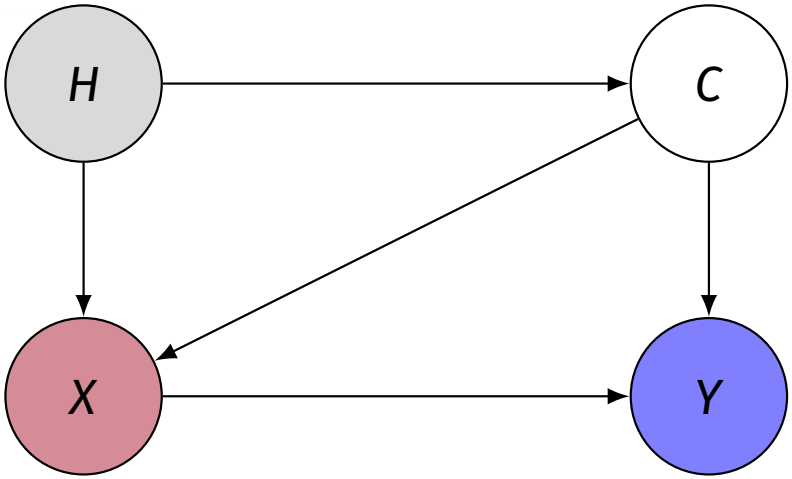
Twin Networks: an example of non-completeness



Does C verify the backdoor criterion for X,Y?

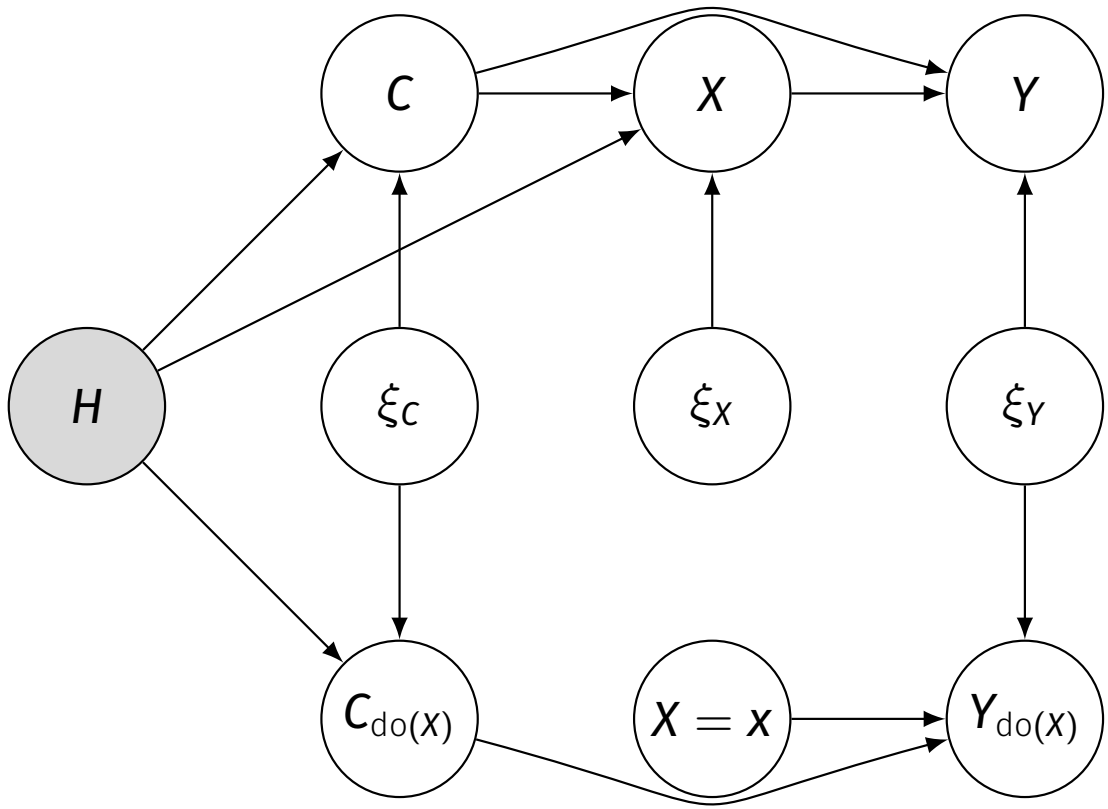


Twin Networks: an example of non-completeness

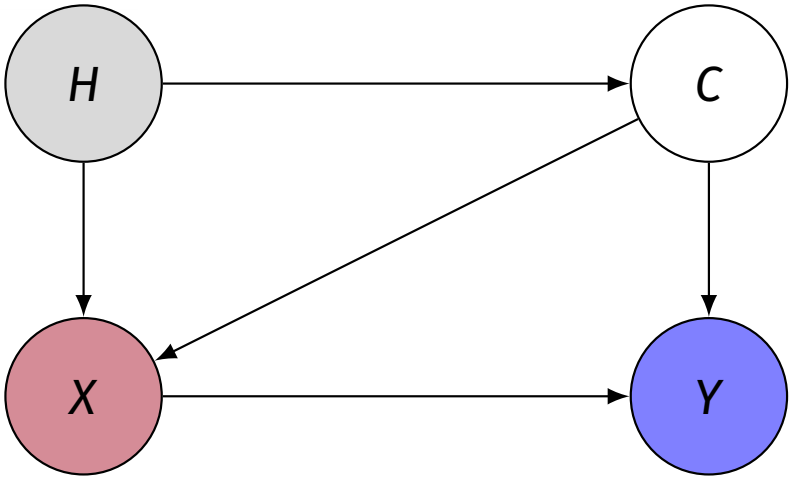


Does C verify the backdoor criterion for X,Y?

$Y_{do(X)} \perp\!\!\!\perp_d X \mid C$?

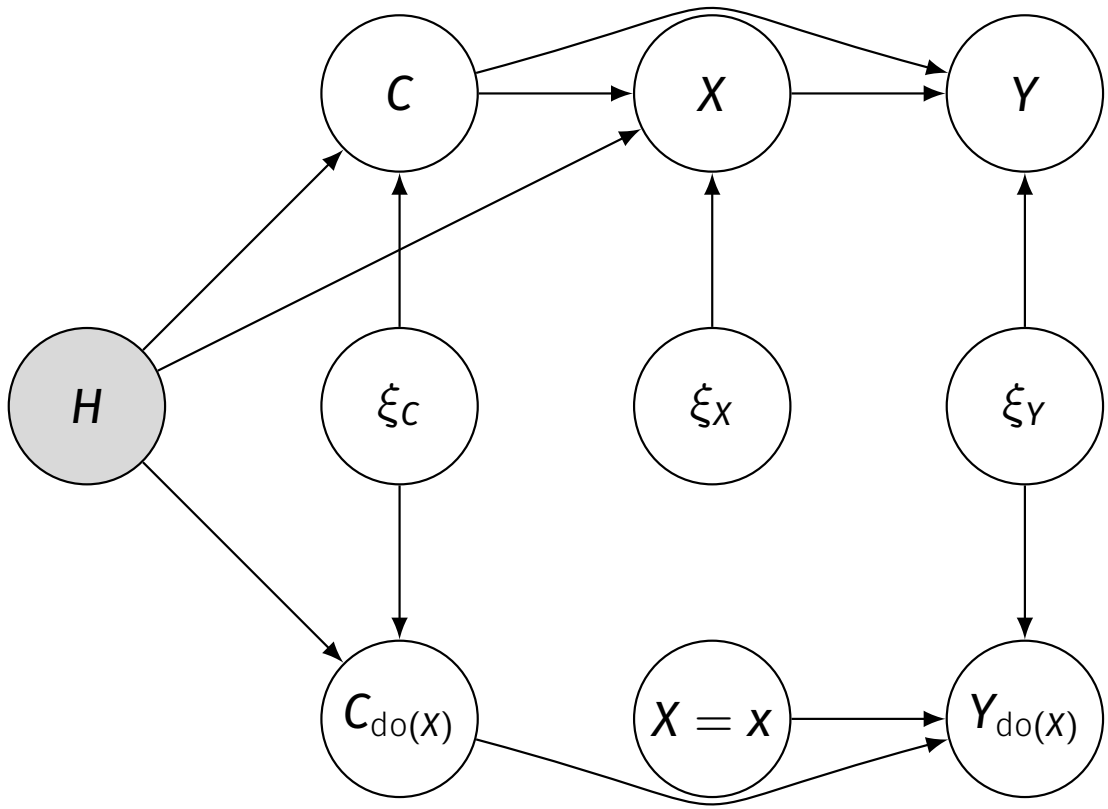


Twin Networks: an example of non-completeness

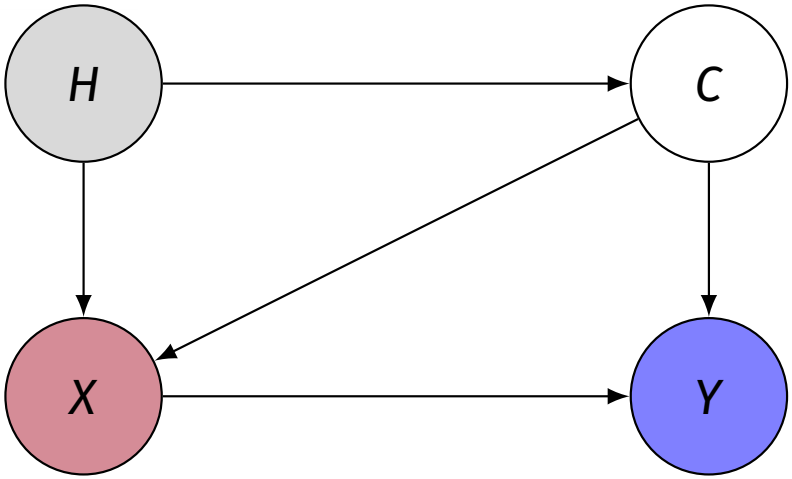


Does C verify the backdoor criterion for X,Y?

$Y_{do(X)} \perp\!\!\!\perp_d X \mid C$?
No!

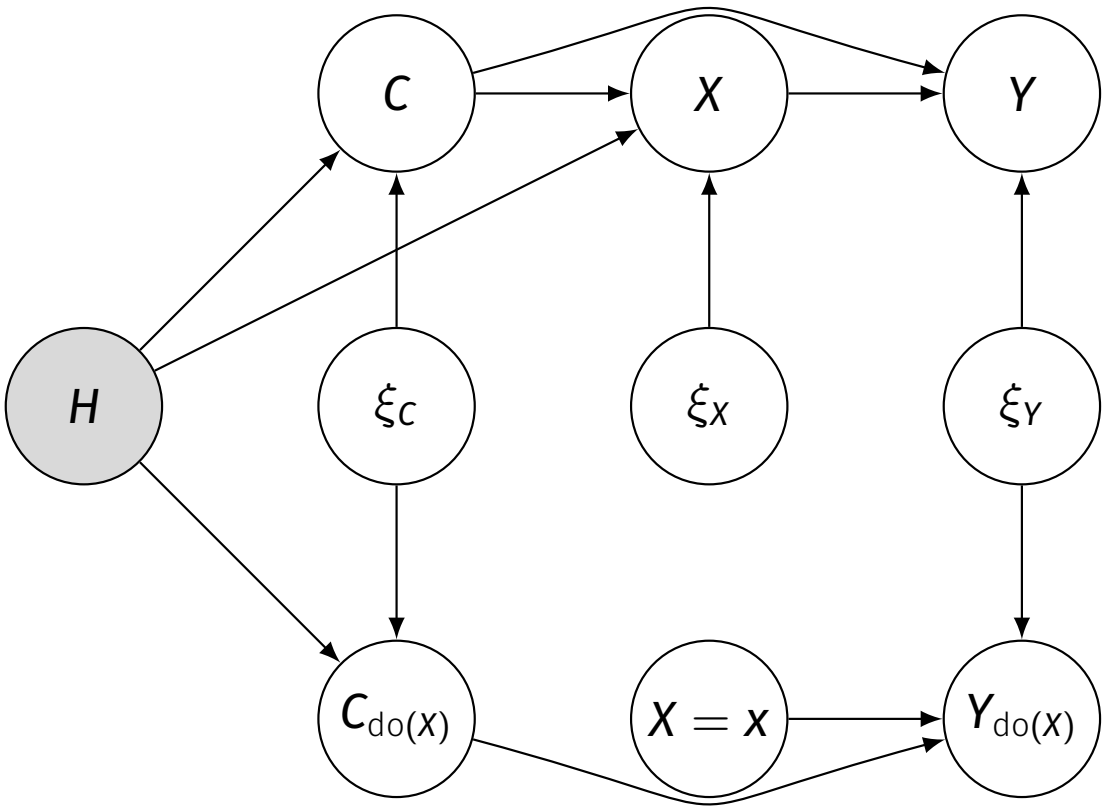


Twin Networks: an example of non-completeness

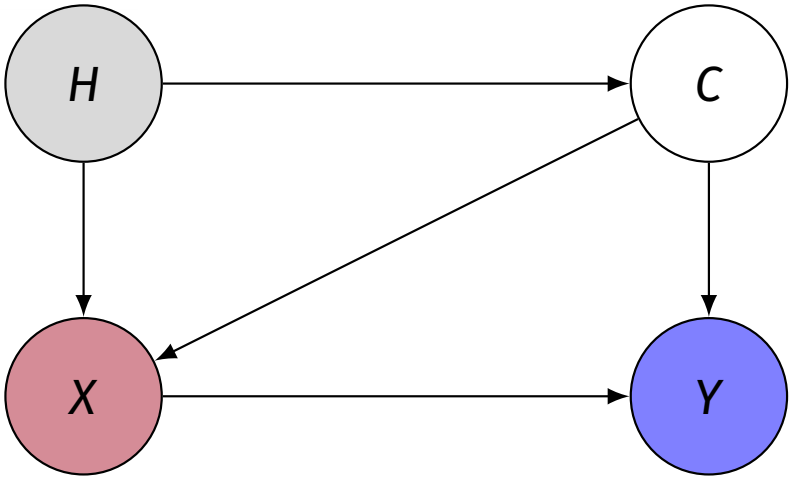


Does C verify the backdoor criterion for X,Y?

$Y_{do(x)} \perp\!\!\!\perp_d X \mid C$?
No! $Y_{do(x)} \perp\!\!\!\perp_d X \mid C, C_{do(x)}$?

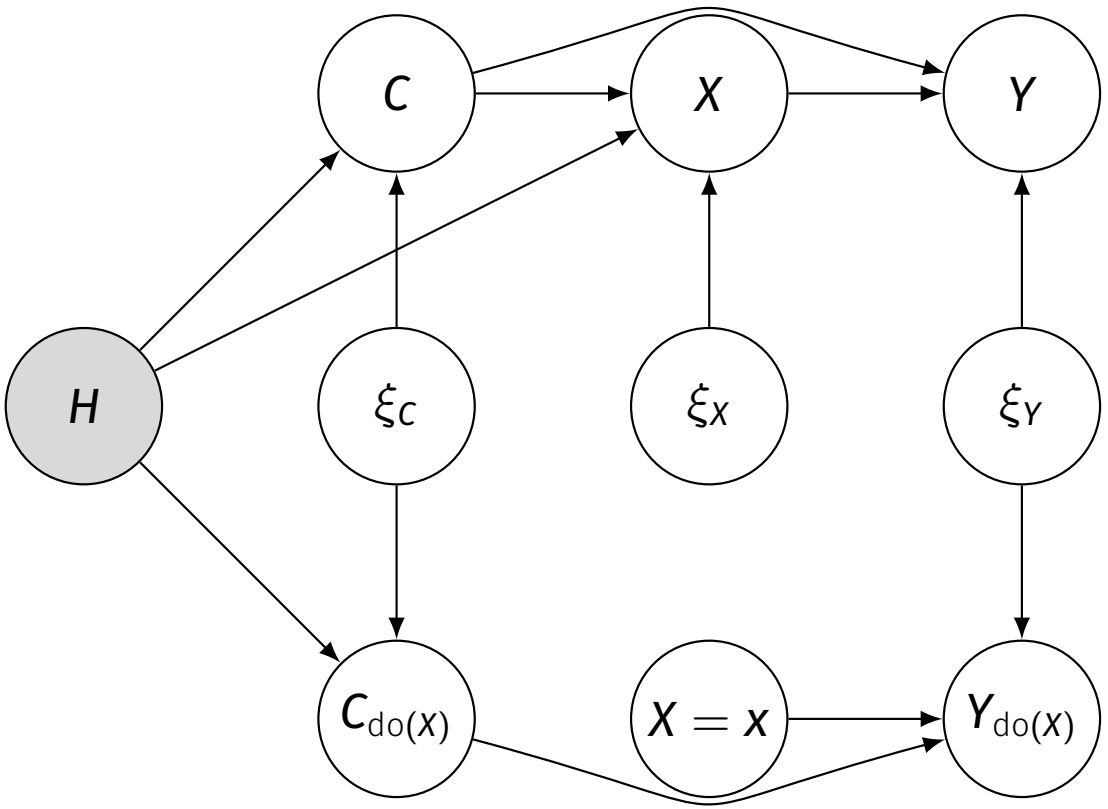


Twin Networks: an example of non-completeness



Does C verify the backdoor criterion for X,Y?

$Y_{do(x)} \perp\!\!\!\perp_d X \mid C$?
No! $Y_{do(x)} \perp\!\!\!\perp_d X \mid C, C_{do(x)}$?
Yes!



Inserm

iPLesp

SANTÉ SORBONNE UNIVERSITÉ

4

Single World Intervention Graphs (SWIGs)

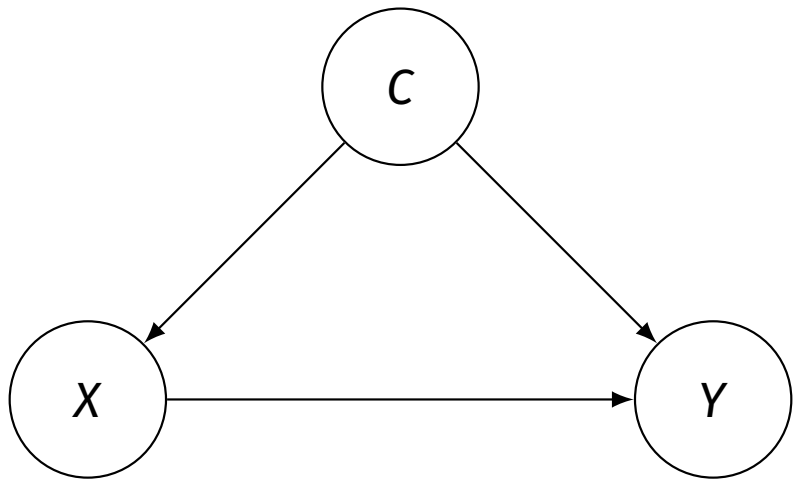
SWIGs

- One graph for both worlds

Richardson and Robins 2013

SWIGs

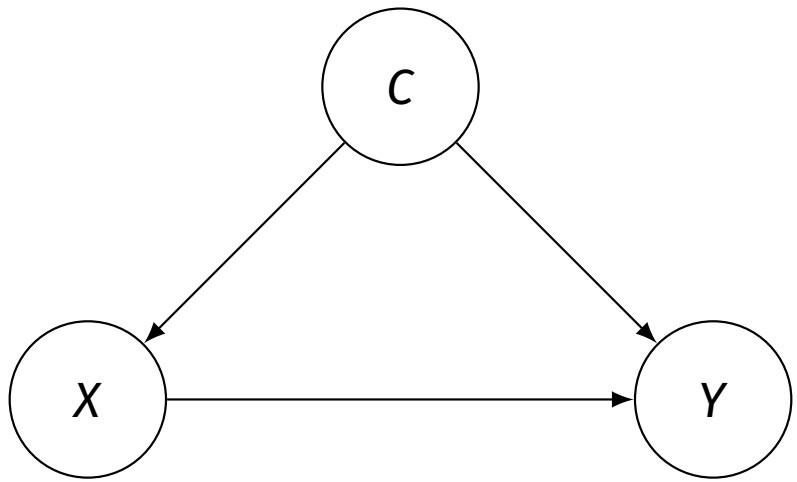
- One graph for both worlds



Richardson and Robins 2013

SWIGs

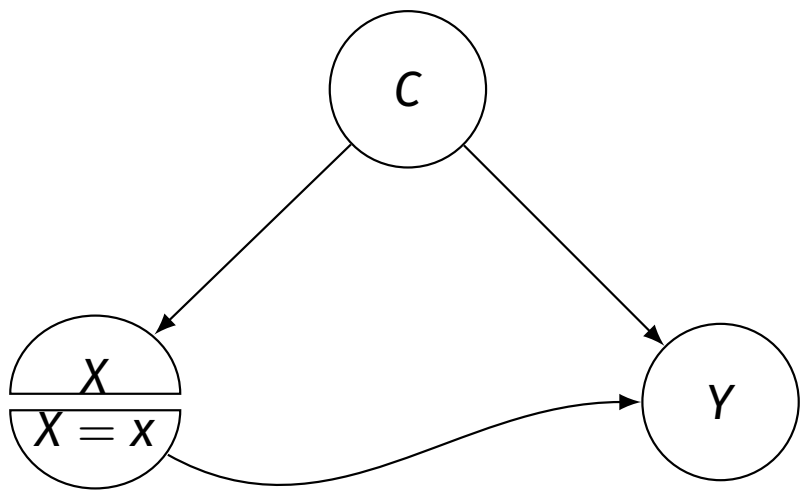
- One graph for both worlds
- Divide the intervened nodes (do $(X = x)$)



Richardson and Robins 2013

SWIGs

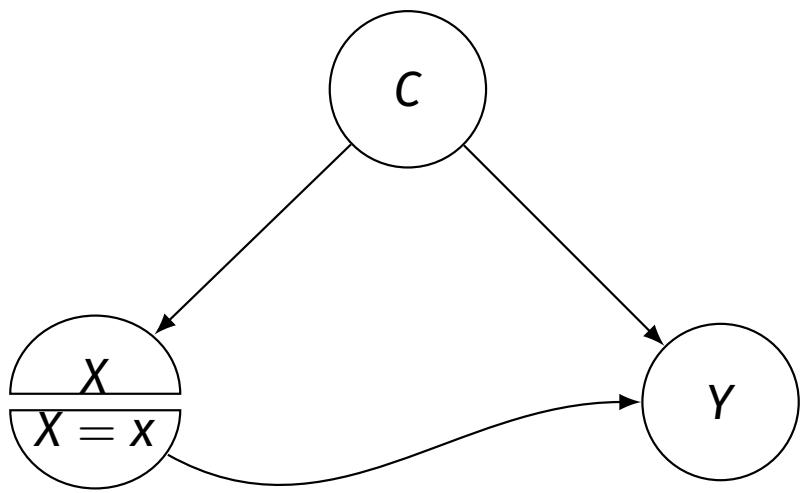
- One graph for both worlds
- Divide the intervened nodes (do $(X = x)$)



Richardson and Robins 2013

SWIGs

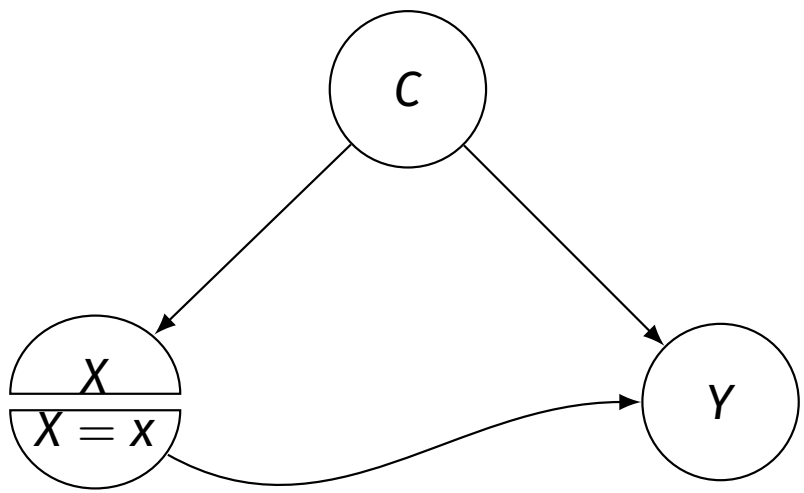
- One graph for both worlds
- Divide the intervened nodes (do $(X = x)$)
- Notice the edges



Richardson and Robins 2013

SWIGs

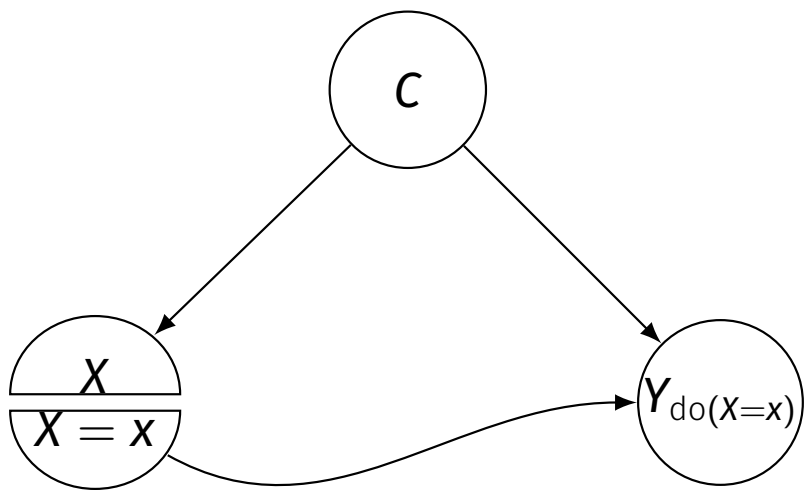
- One graph for both worlds
- Divide the intervened nodes (do $(X = x)$)
- Notice the edges
- Relabel nodes descendant of interventions



Richardson and Robins 2013

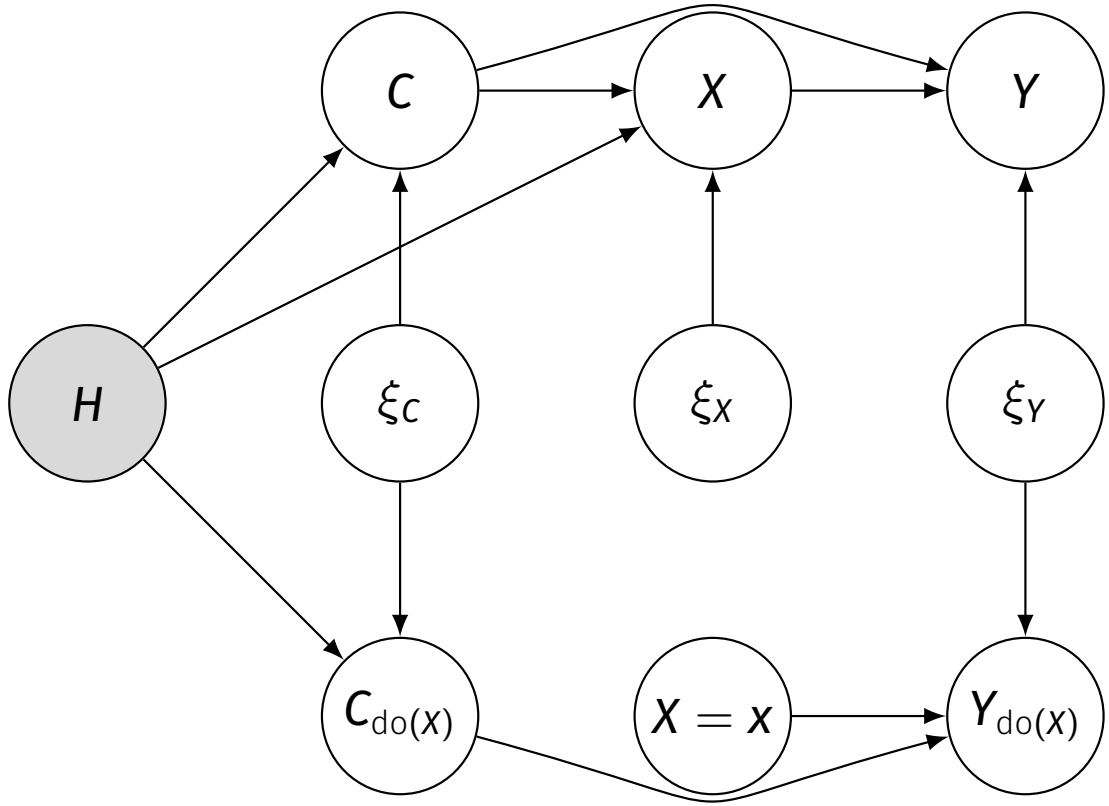
SWIGs

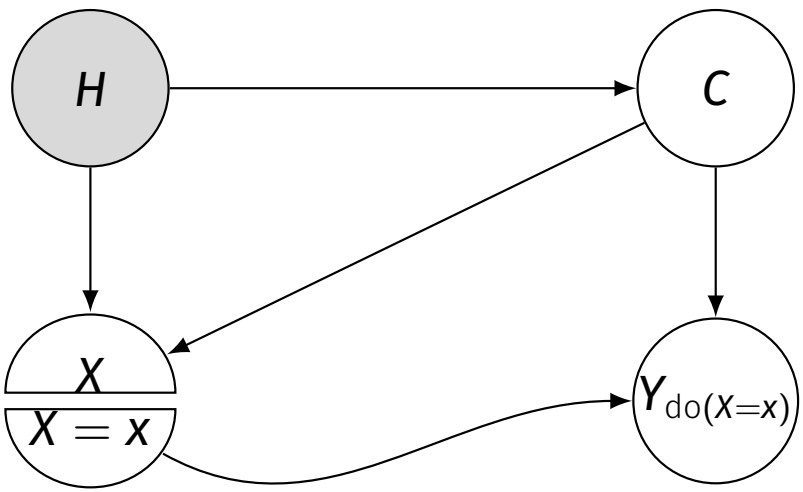
- One graph for both worlds
- Divide the intervened nodes (do $(X = x)$)
- Notice the edges
- Relabel nodes descendant of interventions



Richardson and Robins 2013

SWIGs





$Y_{do(x)} \perp\!\!\!\perp_d X \mid C$?

DO-CALCULUS

The do-calculus is sound
(and complete for interventional queries)

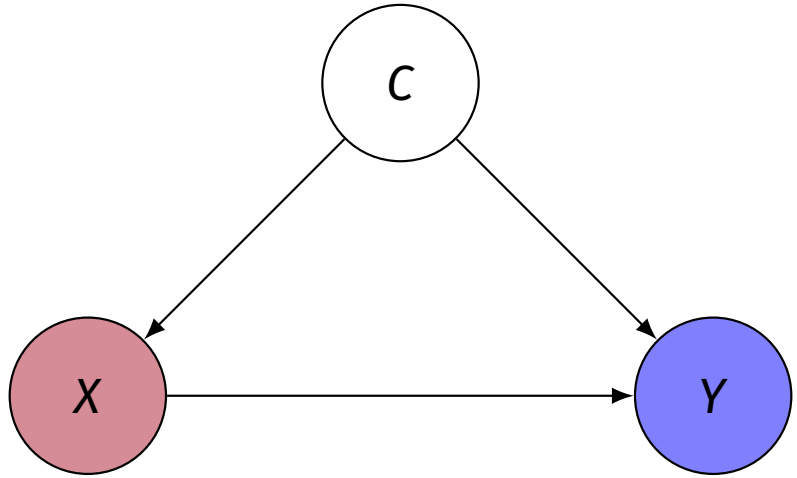
Rule 1 $\Pr(y \mid \mathrm{do}(x), z, w) = \Pr(y \mid \mathrm{do}(x), w)$
if $(Y \perp\!\!\!\perp_d Z \mid X, W)_{\mathcal{G}_{\overline{X}}}$

Rule 2 $\Pr(y \mid \mathrm{do}(x, z), w) = \Pr(y \mid \mathrm{do}((x)), z, w)$
if $(Y \perp\!\!\!\perp_d Z \mid X, W)_{\mathcal{G}_{\overline{X}\underline{Z}}}$

Rule 3 $\Pr(y \mid \mathrm{do}(x, z), w) = \Pr(y \mid \mathrm{do}(x), w)$
if $(Y \perp\!\!\!\perp_d Z \mid X, W)_{\mathcal{G}_{\overline{X}\overline{Z(W)}}}$

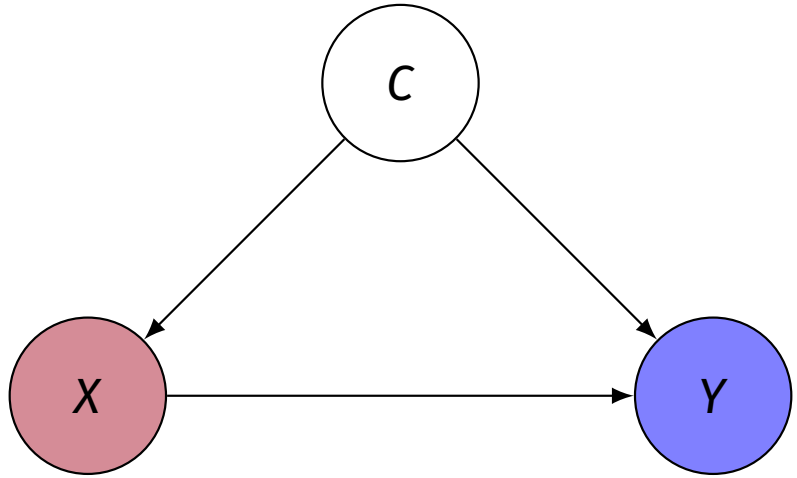
Pearl 1995

The backdoor: an application of the do-calculus



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

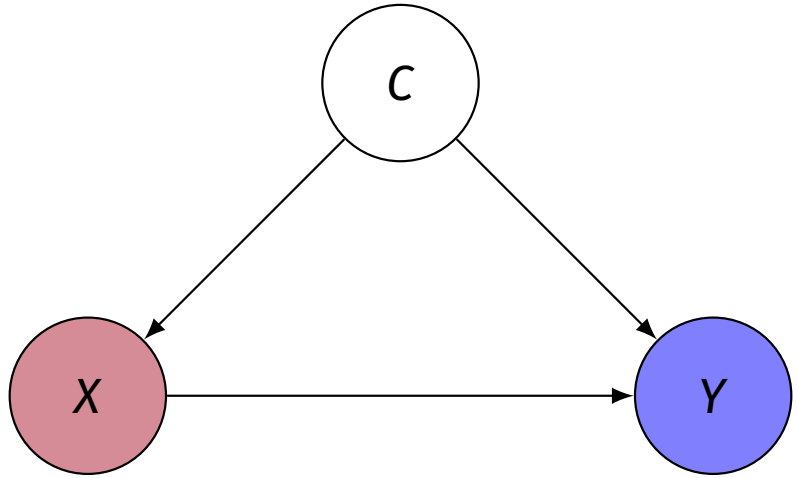
The backdoor: an application of the do-calculus



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

$\Pr(Y \mid \mathrm{do}(X))$

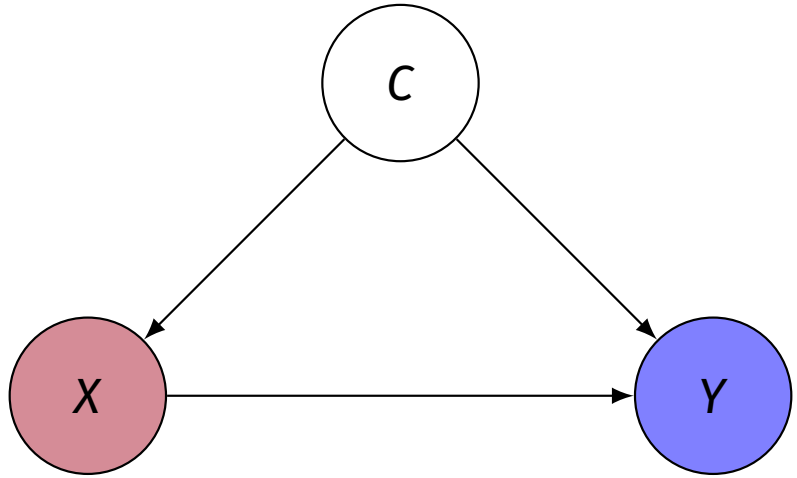
The backdoor: an application of the do-calculus



$$\text{ATE: } \mathbb{E}\left(Y \mid \text{do}\left(X = x_1\right)\right) - \mathbb{E}\left(Y \mid \text{do}\left(X = x_0\right)\right)$$

$$\Pr\left(Y \mid \text{do}\left(X\right)\right) = \sum_c \Pr\left(Y \mid \text{do}\left(X\right), C = c\right) \Pr\left(C = c \mid \text{do}\left(X\right)\right)$$

The backdoor: an application of the do-calculus

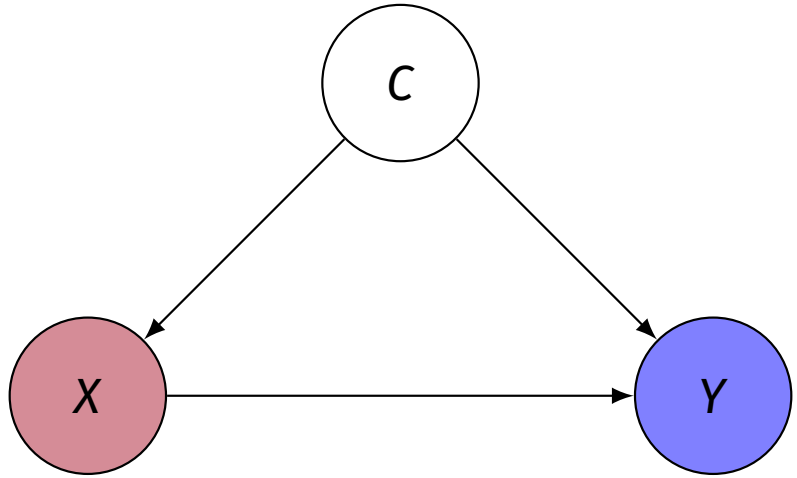


ATE: $\mathbb{E}(Y \mid \mathrm{do}(X = x_1)) - \mathbb{E}(Y \mid \mathrm{do}(X = x_0))$

$$\Pr(Y \mid \mathrm{do}(X)) = \sum_c \Pr(Y \mid \mathrm{do}(X), C = c) \Pr(C = c \mid \mathrm{do}(X))$$

$$\text{(Rule 2)} = \sum_c \Pr(Y \mid X, C = c) \Pr(C = c \mid \mathrm{do}(X))$$

The backdoor: an application of the do-calculus



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

$$\Pr(Y \mid \mathrm{do}(X)) = \sum_c \Pr(Y \mid \mathrm{do}(X), C = c) \Pr(C = c \mid \mathrm{do}(X))$$

$$\text{(Rule 2)} = \sum_c \Pr(Y \mid X, C = c) \Pr(C = c \mid \mathrm{do}(X))$$

$$\text{(Rule 3)} = \sum_c \Pr(Y \mid X, C = c) \Pr(C = c)$$

PO-CALCULUS

Rule 1 $\Pr\left(\mathbb{Y}_{do(x)} \mid \mathbb{Z}_{do(x)}, \mathbb{W}_{do(x)}\right) = \Pr\left(\mathbb{Y}_{do(x)} \mid \mathbb{W}_{do(x)}\right)$

if $\left(\mathbb{Y}_{do(x)} \perp\!\!\!\perp_d \mathbb{Z}_{do(x)} \mid \mathbb{W}_{do(x)}\right)_{\mathcal{G}_{do(x)}}$

Rule 2 $\Pr\left(\mathbb{Y}_{do(x,z)} \mid \mathbb{W}_{do(x,z)}\right) = \Pr\left(\mathbb{Y}_{do(x)} \mid \mathbb{Z}_{do(x)}, \mathbb{W}_{do(x)}\right)$

if $\left(\mathbb{Y}_{do(x,z)} \perp\!\!\!\perp_d \mathbb{Z}_{do(x,z)} \mid \mathbb{W}_{do(x,z)}\right)_{\mathcal{G}_{do(x,z)}}$

Rule 3 $\Pr\left(\mathbb{Y}_{do(x,z)} \mid \mathbb{W}_{do(x,z)}\right) = \Pr\left(\mathbb{Y}_{do(x)} \mid \mathbb{W}_{do(x)}\right)$

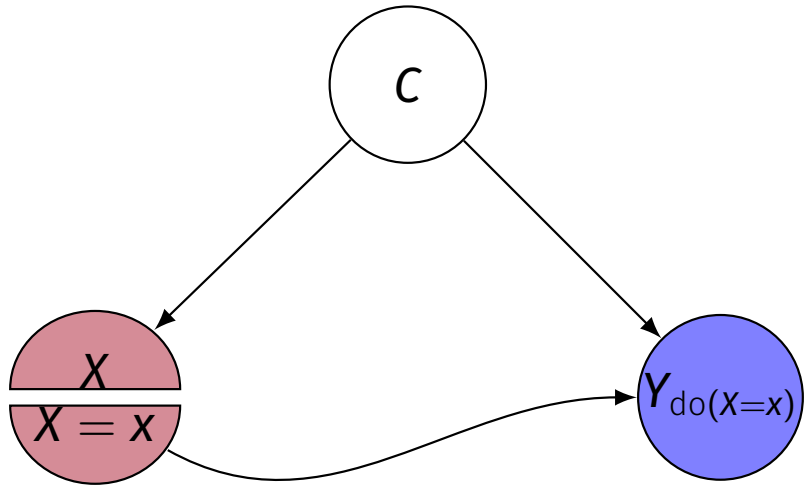
if $\left(\mathbb{Y}_{do(x,z_1)}, \mathbb{W}_{do(x,z_1)} \perp\!\!\!\perp_d \mathbb{Z}_1\right) \mathcal{G}_{do(x,z_1)}$

and $\left(\mathbb{Y}_{do(x,z_1)} \perp\!\!\!\perp_d \mathbb{Z}_{2do(x,z_1)} \mid \mathbb{W}_{do(x,z_1)}\right)_{\mathcal{G}_{do(x,z_1)}}$

where $Z_1 = Z \setminus An\left(W, \mathcal{G}_{do(x)}\right)$ and $Z_2 = Z \cap An\left(W, \mathcal{G}_{do(x)}\right)$

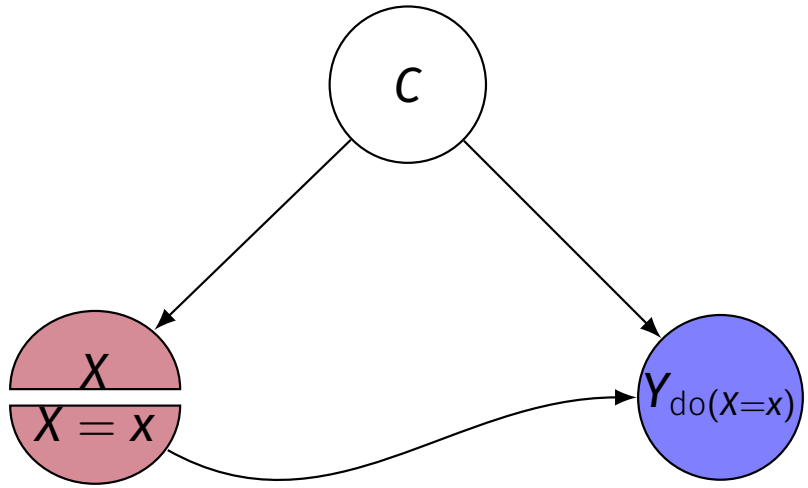
Malinsky, Shpitser, and Richardson 2019

The backdoor: an application of the po-calculus



$$\text{ATE: } \mathbb{E}\left(Y_{\text{do}(X=x_1)}\right) - \mathbb{E}\left(Y_{\text{do}(X=x_0)}\right)$$

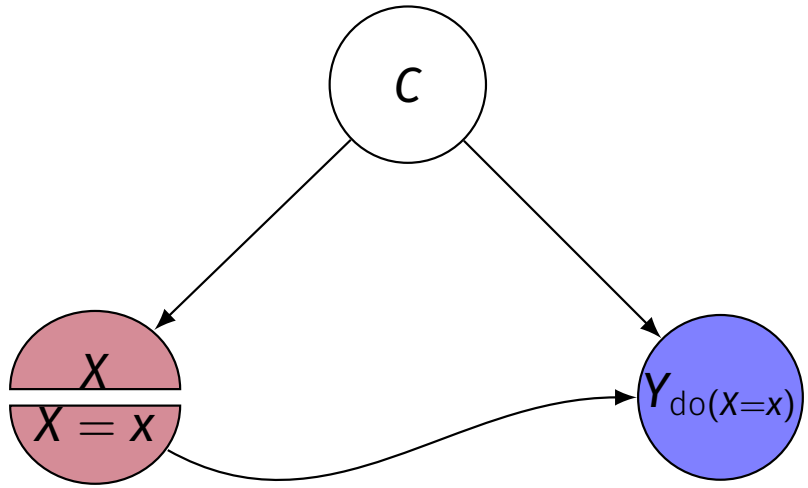
The backdoor: an application of the po-calculus



$$\text{ATE: } \mathbb{E}\left(Y_{\text{do}(X=x_1)}\right) - \mathbb{E}\left(Y_{\text{do}(X=x_0)}\right)$$

$$\Pr\left(Y_{\text{do}(X)}\right)$$

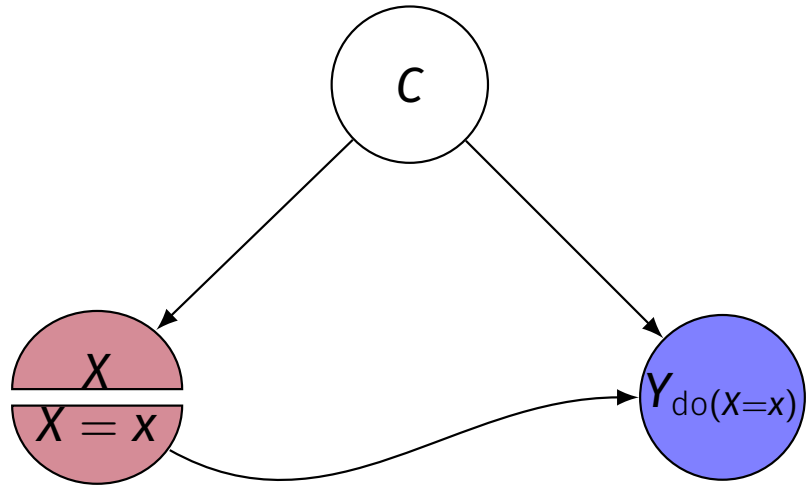
The backdoor: an application of the po-calculus



$$\text{ATE: } \mathbb{E}\left(Y_{\text{do}(X=x_1)}\right) - \mathbb{E}\left(Y_{\text{do}(X=x_0)}\right)$$

$$\Pr\left(Y_{\text{do}(x)}\right) = \sum_c \Pr\left(Y_{\text{do}(x)} \mid C_{\text{do}(x)} = c\right) \Pr\left(C_{\text{do}(x)} = c\right)$$

The backdoor: an application of the po-calculus

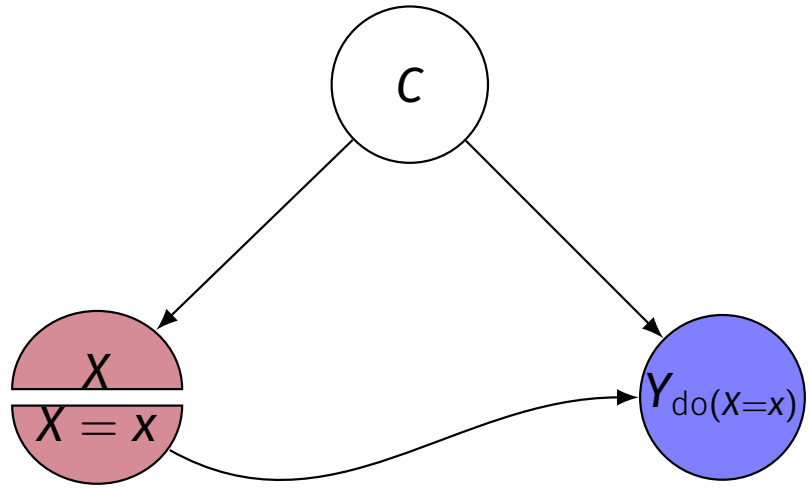


$$\text{ATE: } \mathbb{E}\left(Y_{\text{do}(X=x_1)}\right) - \mathbb{E}\left(Y_{\text{do}(X=x_0)}\right)$$

$$\Pr\left(Y_{\text{do}(X)}\right) = \sum_c \Pr\left(Y_{\text{do}(X)} \mid C_{\text{do}(X)} = c\right) \Pr\left(C_{\text{do}(X)} = c\right)$$

$$\text{(Rule 2)} = \sum_c \Pr\left(Y \mid X, C = c\right) \Pr\left(C_{\text{do}(X)} = c\right)$$

The backdoor: an application of the po-calculus



$$\text{ATE: } \mathbb{E}\left(Y_{do(X=x_1)}\right) - \mathbb{E}\left(Y_{do(X=x_0)}\right)$$

$$\Pr\left(Y_{do(X)}\right) = \sum_c \Pr\left(Y_{do(X)} \mid C_{do(X)} = c\right) \Pr\left(C_{do(X)} = c\right)$$

$$\text{(Rule 2)} = \sum_c \Pr\left(Y \mid X, C = c\right) \Pr\left(C_{do(X)} = c\right)$$

$$\text{(Rule 3)} = \sum_c \Pr\left(Y \mid X, C = c\right) \Pr\left(C = c\right)$$

Limitation of SWIGs

SWIGs: Single World Interventional Graphs

Limitation of SWIGs

SWIGs: Single World Interventional Graphs

Limitation of SWIGs

SWIGs: Single World Interventional Graphs

What is the effect of in the population that would die if they do not take it?

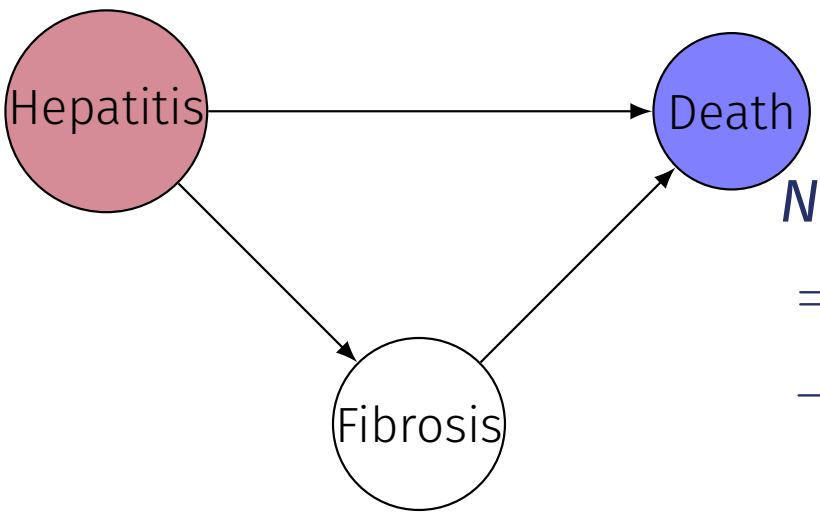
$$\mathbb{E}\left(Y_{do(X=1)} \mid Y_{do(X=0)} = 0\right)$$

Limitation of SWIGs

SWIGs: Single World Interventional Graphs

What is the effect of in the population that would die if they do not take it?

$$\mathbb{E}\left(Y_{\text{do}(X=1)} \mid Y_{\text{do}(X=0)} = 0\right)$$



$$\begin{aligned} &NDE(D, H_{0\to 1}, F) \\ &= \mathbb{E}\left(D \mid \text{do}\left(H = 1, F = f_{\text{do}(H=0)}\right)\right) \\ &- \mathbb{E}\left(D \mid \text{do}\left(H = 0, F = f_{\text{do}(H=0)}\right)\right) \end{aligned}$$

Inserm

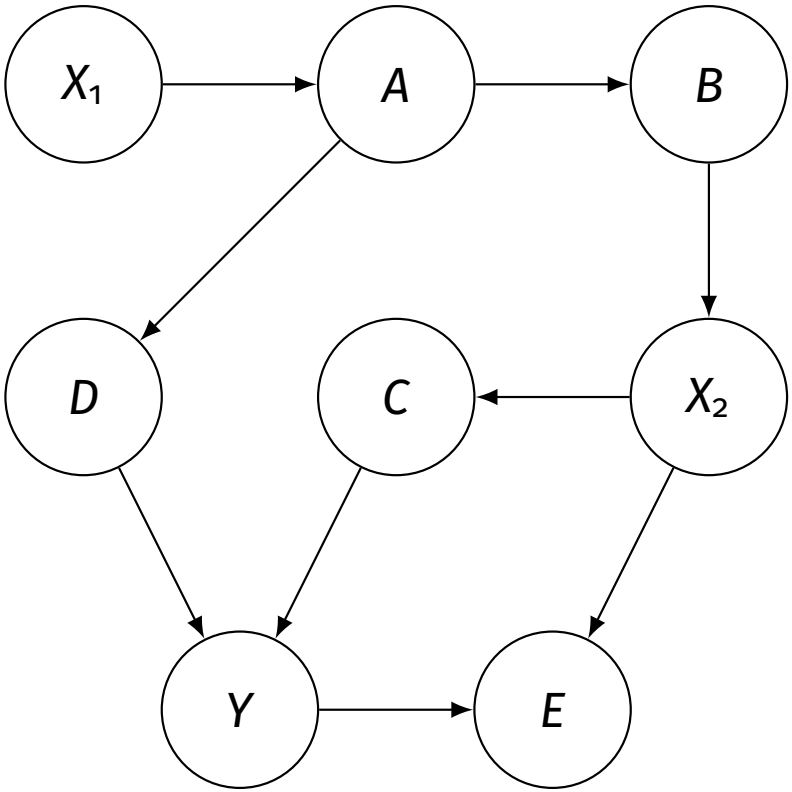
iPLesp

SANTÉ SORBONNE UNIVERSITÉ

5

Ancestral Multi World Network (AMWN)

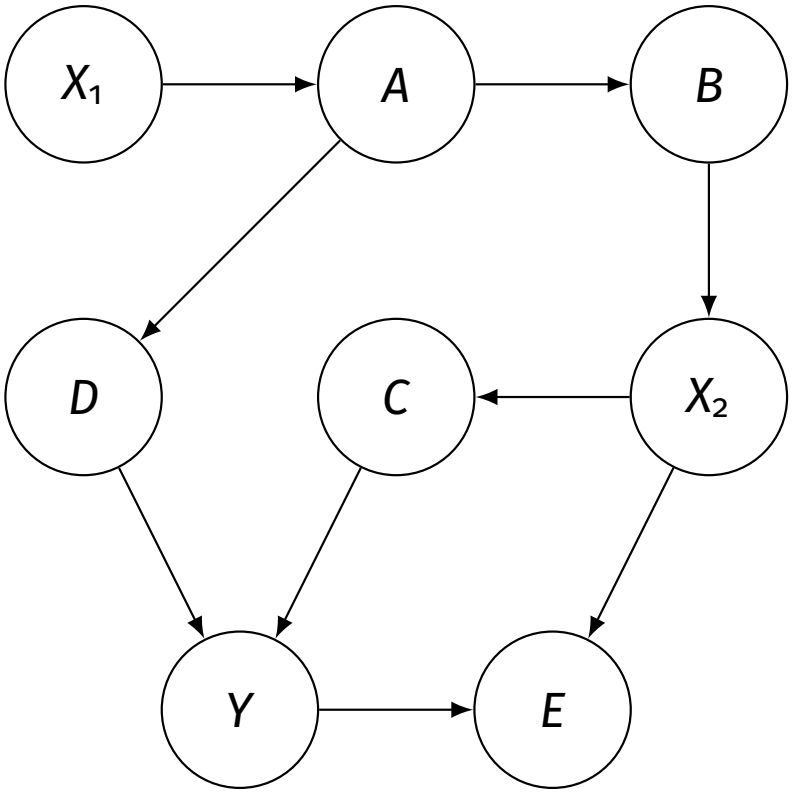
Ancestor of a Counterfactual Variable



$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

Correa, Lee, and Bareinboim 2021

Ancestor of a Counterfactual Variable

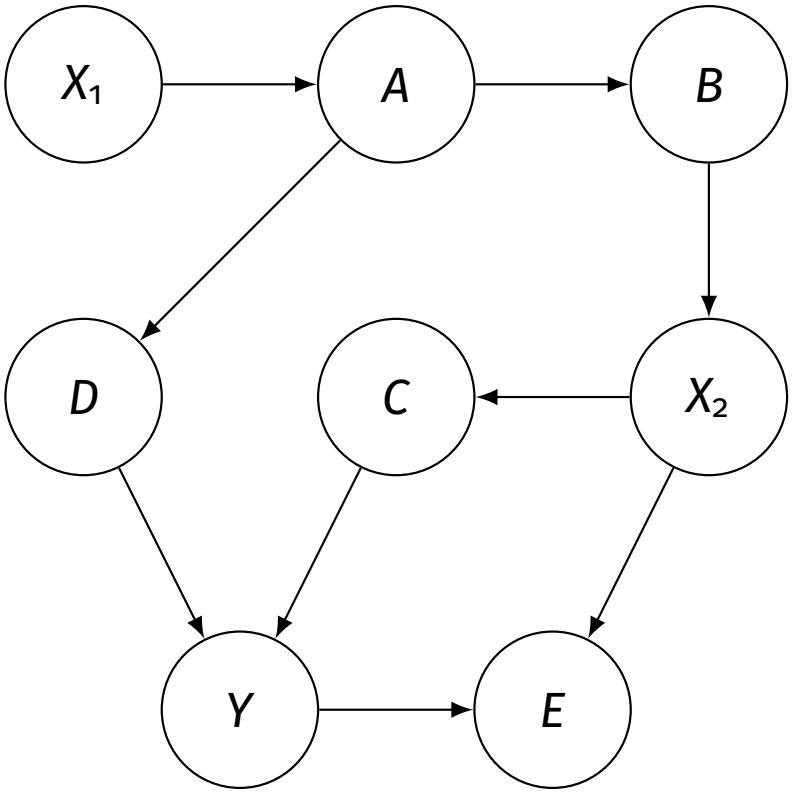


$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

$An(Y_{do(X_1,X_2)})$?

Correa, Lee, and Bareinboim 2021

Ancestor of a Counterfactual Variable



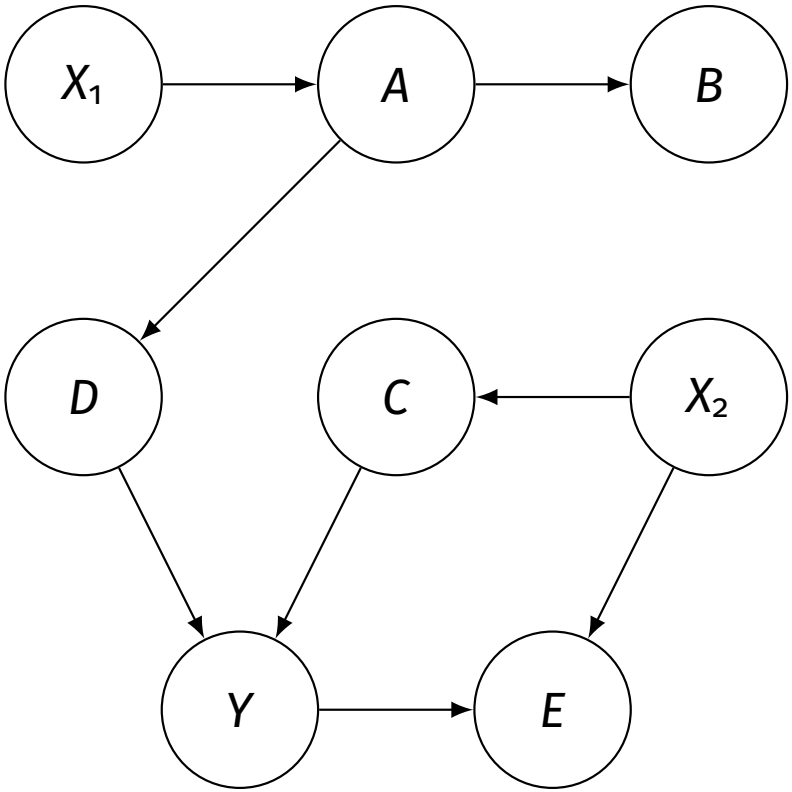
$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

$An(Y_{do(X_1,X_2)})$?

- $G_{\overline{X_1,X_2}}$

Correa, Lee, and Bareinboim 2021

Ancestor of a Counterfactual Variable



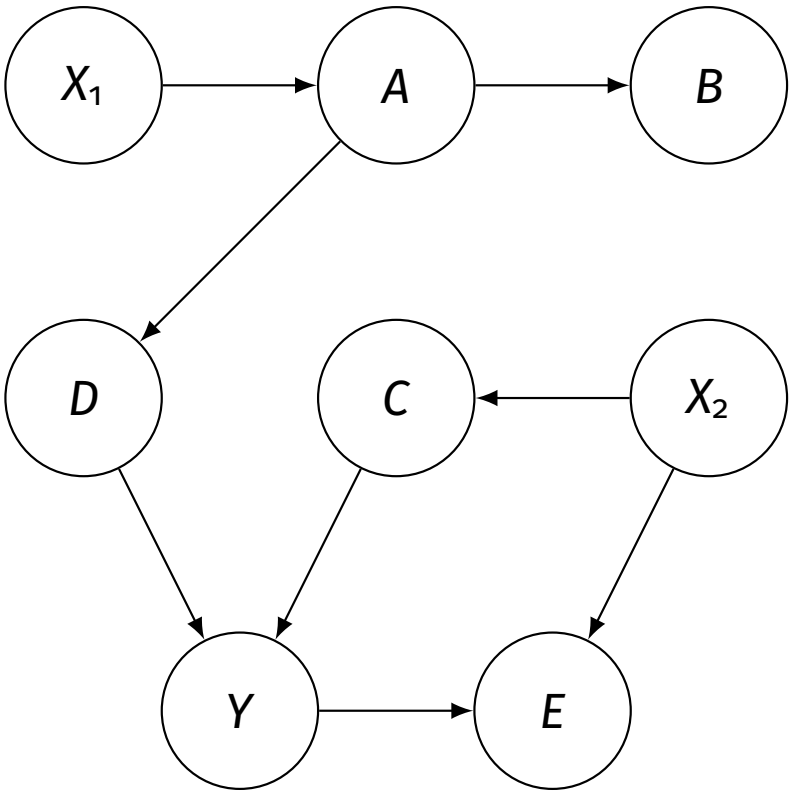
$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

$An(Y_{do(X_1,X_2)})$?

- $G_{\overline{X_1,X_2}}$

Correa, Lee, and Bareinboim 2021

Ancestor of a Counterfactual Variable



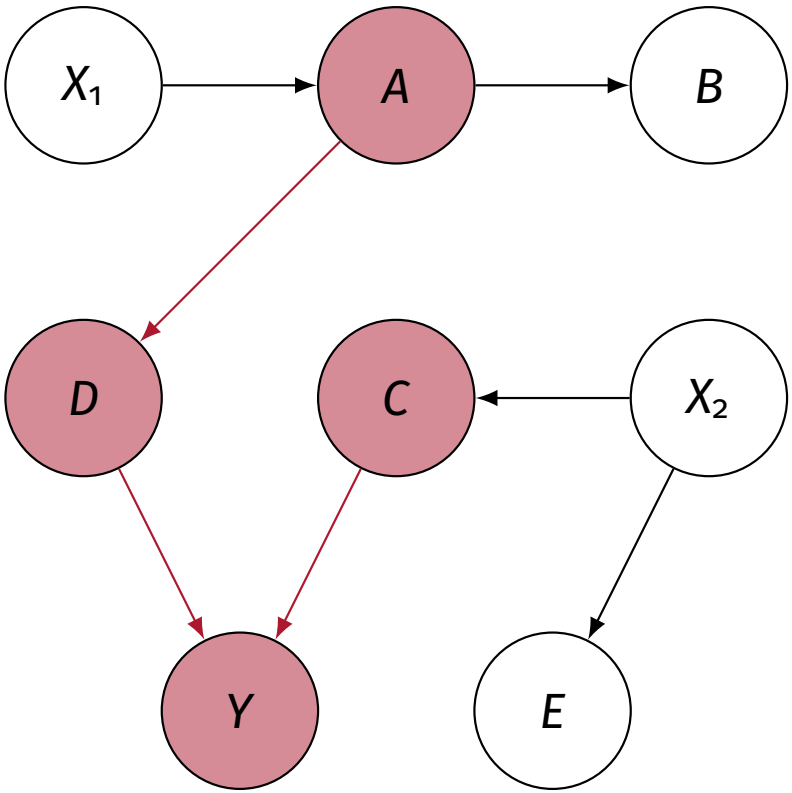
$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

$An(Y_{do(X_1,X_2)})$?

$\cdot\ G_{\overline{X_1,X_2}}$ $\cdot\ An\left(Y, G_{\overline{X_1,X_2}}\right) \setminus \{X_1, X_2\}$

Correa, Lee, and Bareinboim 2021

Ancestor of a Counterfactual Variable



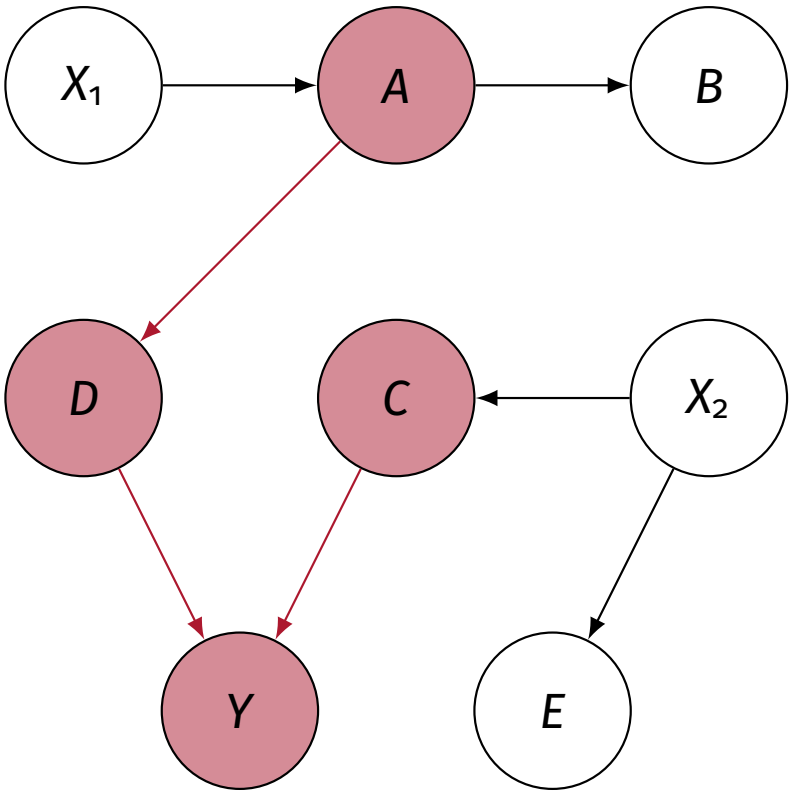
$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

$An(Y_{do(X_1,X_2)})$?

$\cdot\ G_{\overline{X_1,X_2}}$ $\cdot\ An\left(Y, G_{\overline{X_1,X_2}}\right) \setminus \{X_1, X_2\}$

Correa, Lee, and Bareinboim 2021

Ancestor of a Counterfactual Variable



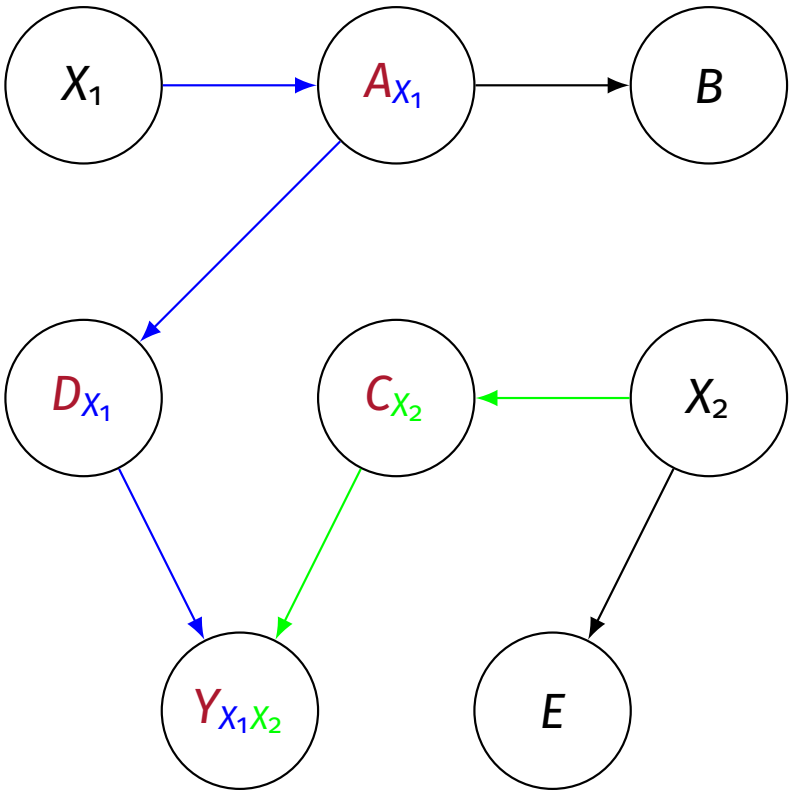
$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

$$An(Y_{do(x_1,x_2)})?$$

$$\cdot\, G_{\overline{X_1,X_2}} \quad \cdot\, An\left(Y, G_{\overline{X_1,X_2}}\right) \setminus \{X_1, X_2\} \quad \cdot\, \{x_1, x_2\} \cap An\left(W, G_{\overline{X_1,X_2}}\right)$$

Correa, Lee, and Bareinboim 2021

Ancestor of a Counterfactual Variable



$$An(\mathbb{Y}_{do(\mathbb{x})}) = \{W_{do(\mathbb{z})} \mid W \in An\left(Y, G_{\overline{\mathbb{X}}}\right) \setminus \mathbb{X}, \mathbb{z} = \mathbb{x} \cap An\left(W, G_{\overline{\mathbb{X}}}\right)\}$$

$$An(Y_{do(x_1,x_2)})?$$

$$\cdot\, G_{\overline{X_1,X_2}} \quad \cdot\, An\left(Y, G_{\overline{X_1,X_2}}\right) \setminus \{X_1, X_2\} \quad \cdot\, \{x_1, x_2\} \cap An\left(W, G_{\overline{X_1,X_2}}\right)$$

Correa, Lee, and Bareinboim 2021

AMWN

For a query on counterfactual variables:

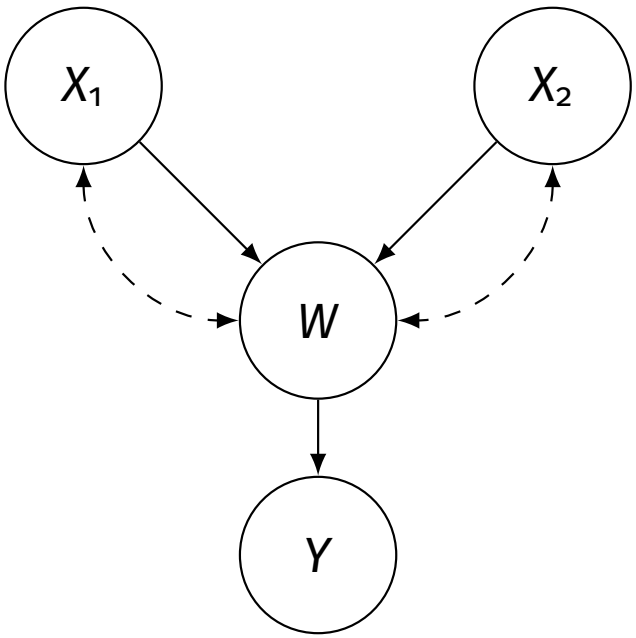
- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables

Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables

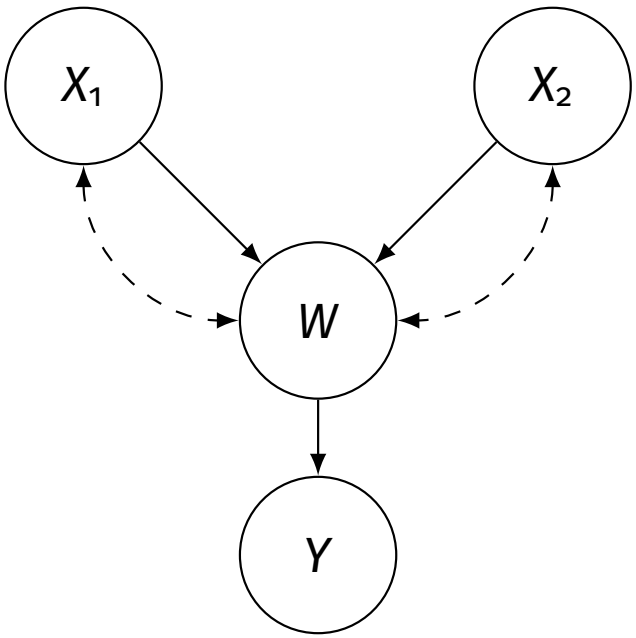


Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables



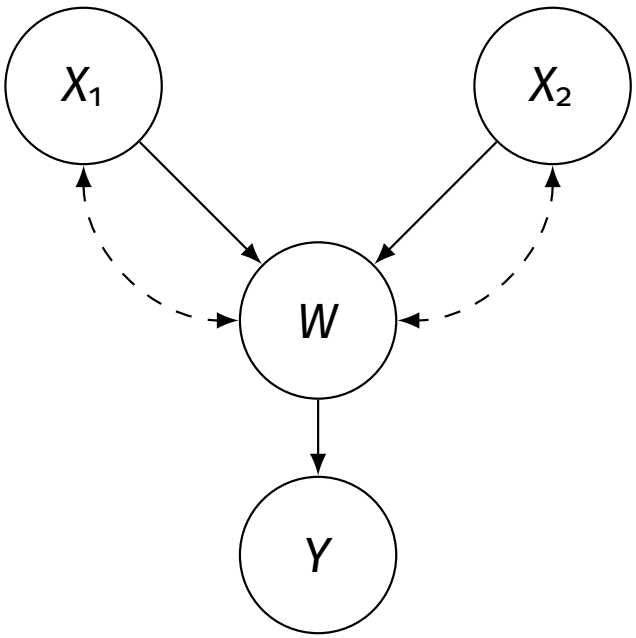
$$Y_{\text{do}(X_1)} \not\perp\!\!\!\perp_d X_2 \mid W_{\text{do}(X_1)}, W_{\text{do}(X_1,X_2)}$$

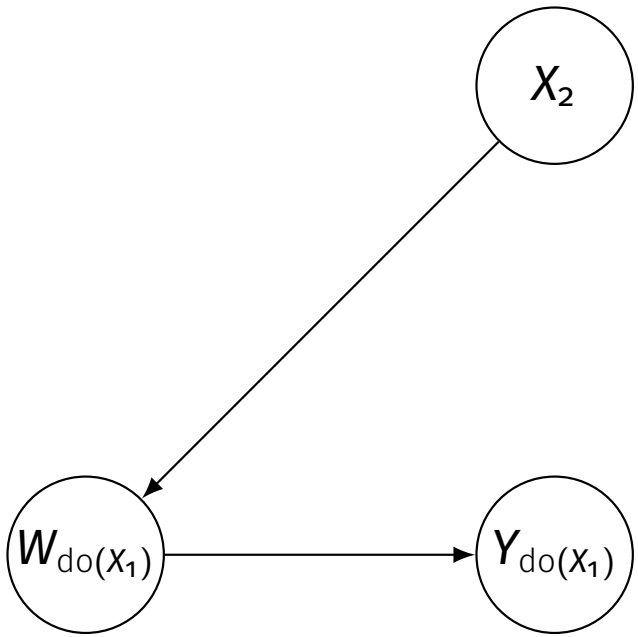
Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables





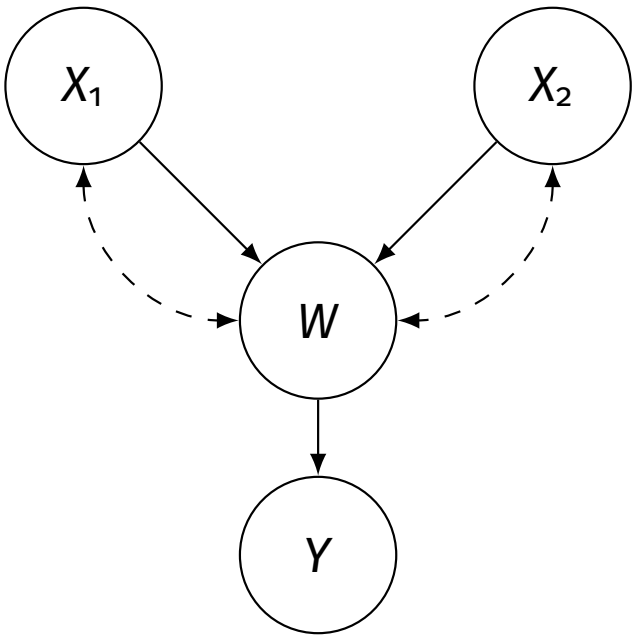
$$Y_{do(X_1)} \perp\!\!\!\perp_d X_2 \mid W_{do(X_1)}, W_{do(X_1,X_2)}$$

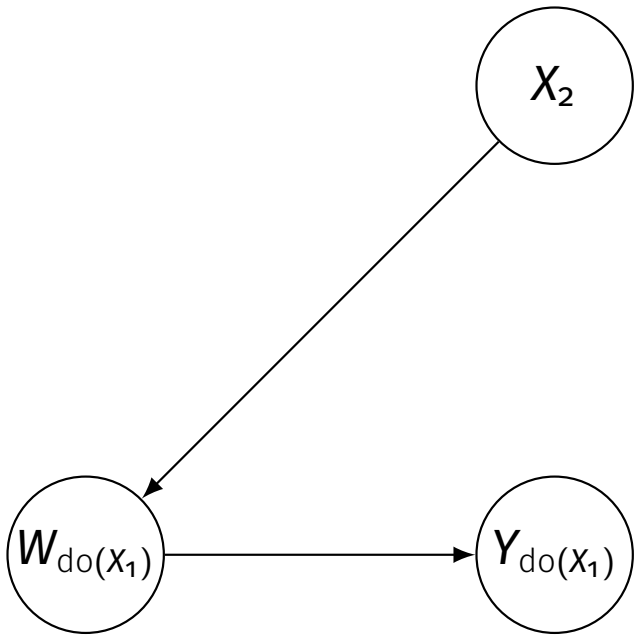
Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables





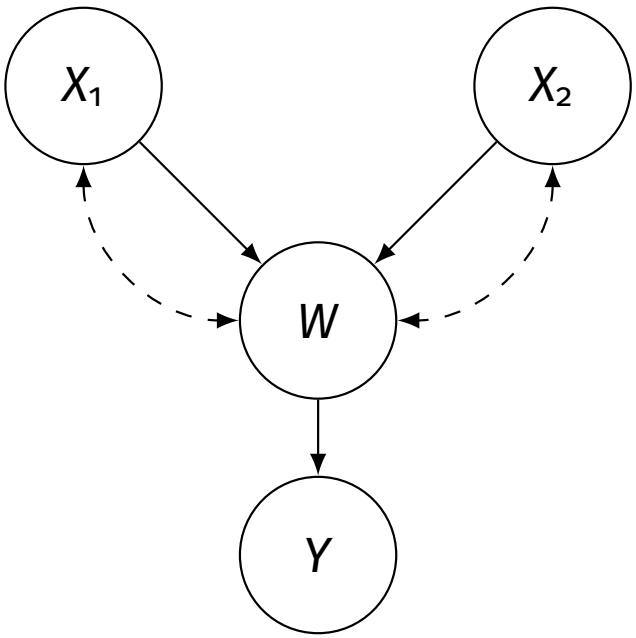
$$Y_{do(X_1)} \perp\!\!\!\perp_d X_2 \mid W_{do(X_1)}, W_{do(X_1,X_2)}$$

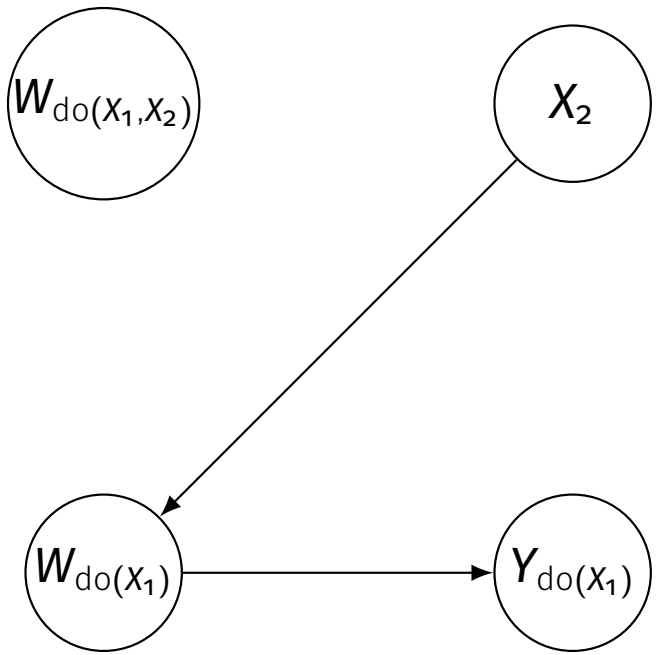
Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables





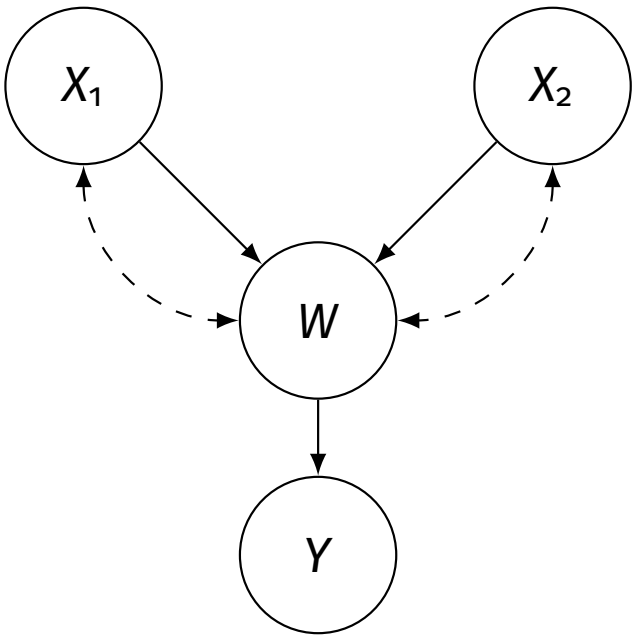
$$Y_{do(X_1)} \perp\!\!\!\perp_d X_2 \mid W_{do(X_1)}, W_{do(X_1,X_2)}$$

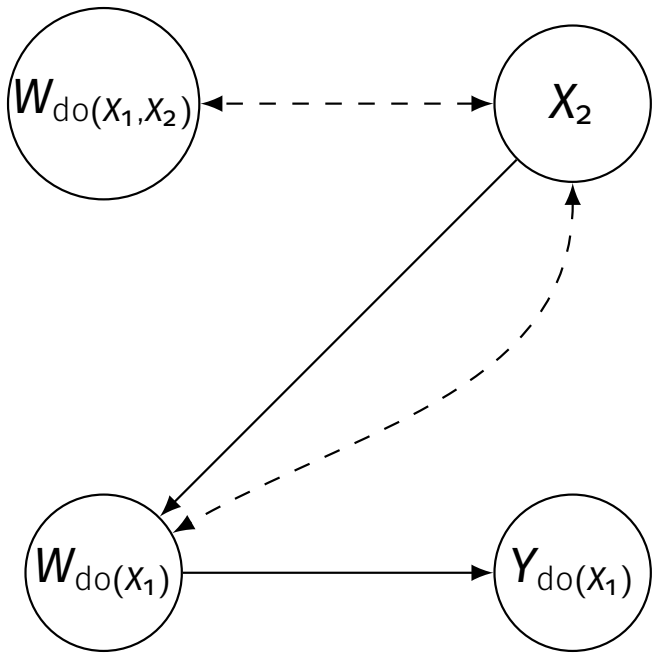
Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables





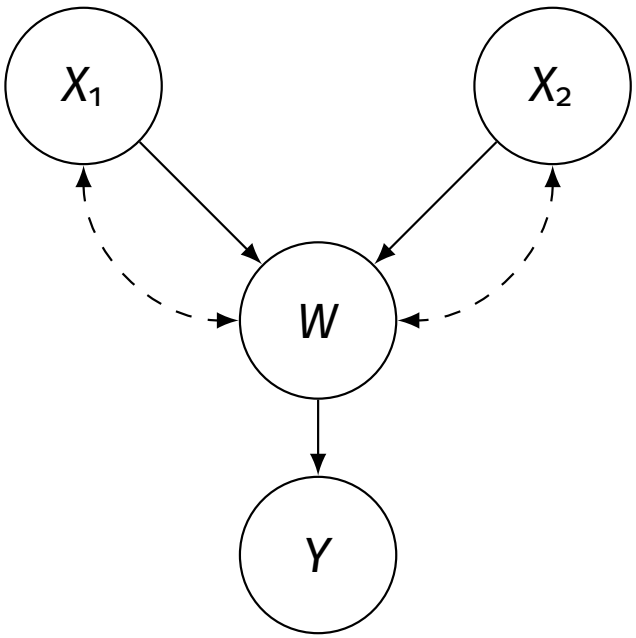
$$Y_{do(X_1)} \perp\!\!\!\perp_d X_2 \mid W_{do(X_1)}, W_{do(X_1,X_2)}$$

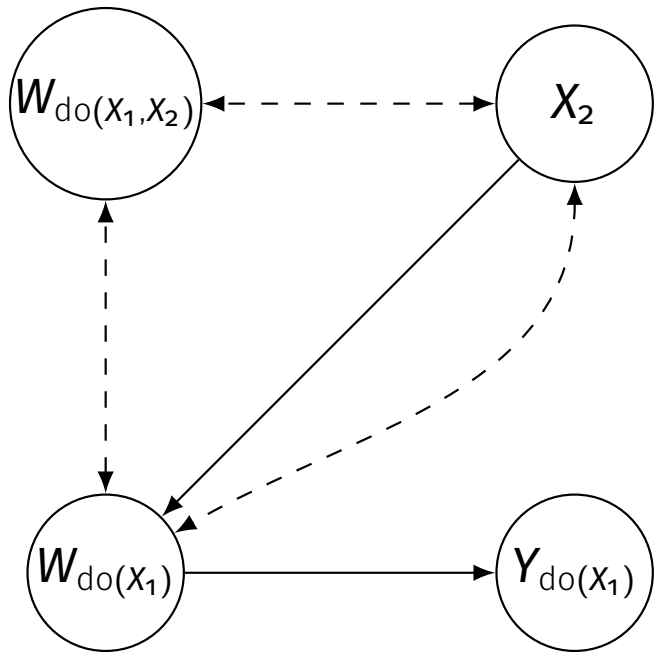
Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables





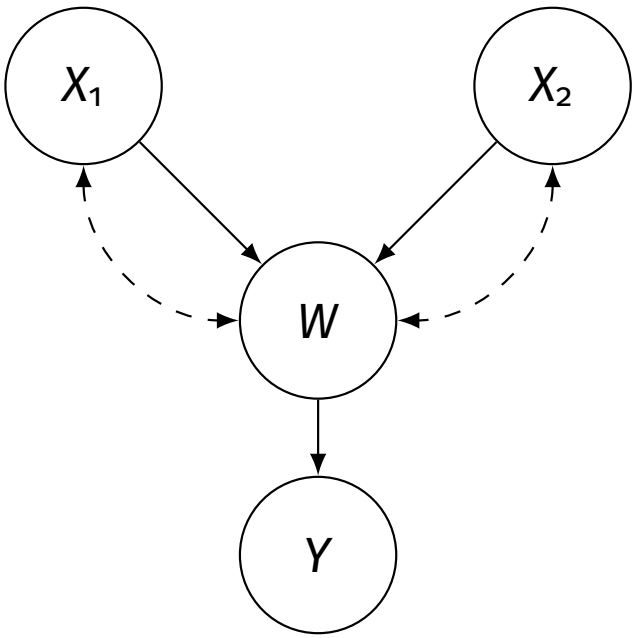
$$Y_{do(X_1)} \perp\!\!\!\perp_d X_2 \mid W_{do(X_1)}, W_{do(X_1,X_2)}$$

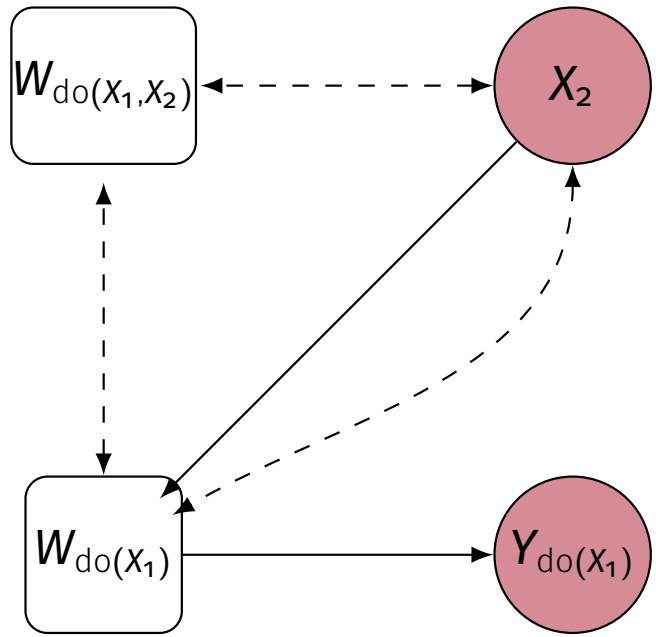
Correa and Bareinboim 2025

AMWN

For a query on counterfactual variables:

- Add all ancestors of the counterfactual variables
- Add edges witnessing the ancestrality
- Add the hidden variables
- Add hidden confounders between same variables





$$Y_{do(X_1)} \perp\!\!\!\perp_d X_2 \mid W_{do(X_1)}, W_{do(X_1,X_2)}$$

Correa and Bareinboim 2025

Counterfactual(Ctf)-Calculus

The ctf-calculus is sound and complete for counterfactual queries.

Rule 1 $\Pr(y_{T_* x}, x_{T_*}, w_*) = \Pr(Y_{T_*}, x_{T_*}, w_*)$

Rule 2 $\Pr(y_r \mid x_t, w_*) = \Pr(y_r \mid w_*)$

$$\text{if } (Y_r \perp\!\!\!\perp_d X_t \mid W_*)_{\mathcal{G}_A}$$

Rule 3 $\Pr(Y_{xz}, w_*) = \Pr(y_z, w_*)$

$$\text{if } X \cap An\left(Y, \mathcal{G}_{\overline{Z}}\right) = \emptyset$$

Correa and Bareinboim 2025

Counterfactual(Ctf)-Calculus

The ctf-calculus is sound and complete for counterfactual queries.

Rule 1 $\Pr\left(y_{T_* x}, x_{T_*}, w_*\right) = \Pr\left(Y_{T_*}, x_{T_*}, w_*\right)$

Rule 2 $\Pr\left(y_r \mid x_t, w_*\right) = \Pr\left(y_r \mid w_*\right)$

$$\text{if } \left(Y_r \perp\!\!\!\perp_d X_t \mid W_*\right)_{\mathcal{G}_A}$$

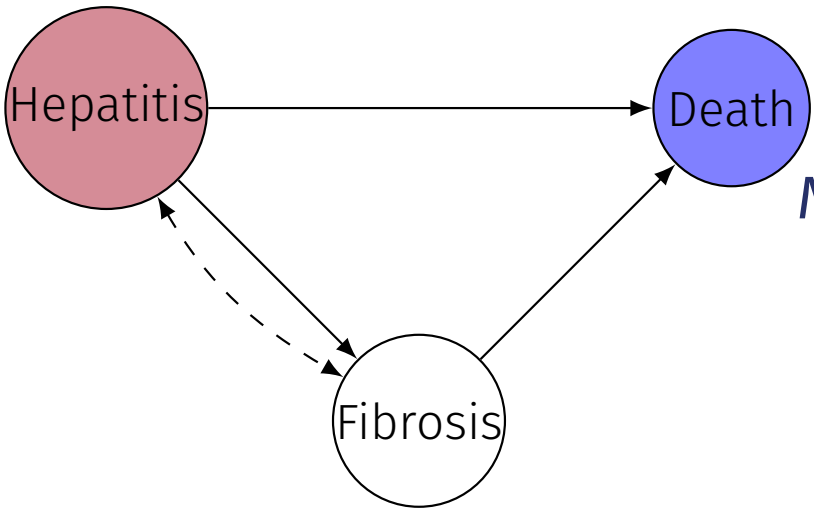
Rule 3 $\Pr\left(Y_{xz}, w_*\right) = \Pr\left(y_z, w_*\right)$

$$\text{if } X \cap An\left(Y, \mathcal{G}_{\overline{Z}}\right) = \emptyset$$

It transforms counterfactual queries into interventional queries. You might still need to use the do-calculus aftewards to transform the interventional queries into observational queries.

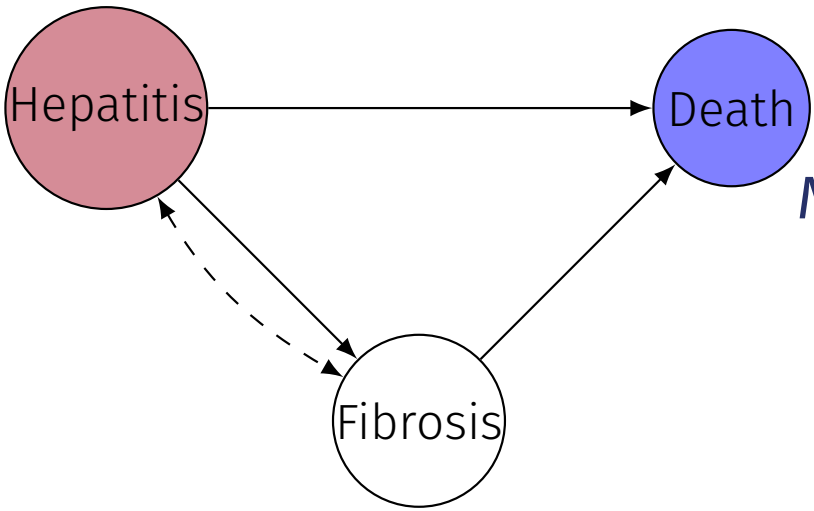
Correa and Bareinboim 2025

The Natural Direct Effect



$$NDE(D, H_{0\to 1}, F)$$
$$= \mathbb{E}\left(D \mid \text{do}\left(H = 1, F = f_{\text{do}(H=0)}\right)\right)$$
$$- \mathbb{E}\left(D \mid \text{do}\left(H = 0, F = f_{\text{do}(H=0)}\right)\right)$$

The Natural Direct Effect



$$NDE(D, H_{0\to 1}, F)$$
$$= \mathbb{E}\left(D \mid \text{do}\left(H = 1, F = f_{\text{do}(H=0)}\right)\right)$$
$$- \mathbb{E}\left(D \mid \text{do}\left(H = 0, F = f_{\text{do}(H=0)}\right)\right)$$

How do you identify
$\Pr\left(D \mid \text{do}\left(H = 1, F = f_{\text{do}(H=0)}\right)\right) = \Pr\left(d_{h' F_h}\right)$?

CTF-CALCULUS: EXAMPLE OF CALCULATIONS

$$\Pr\left(d_{h'F_h}\right)$$

CTF-CALCULUS: EXAMPLE OF CALCULATIONS

$$\Pr\left(d_{h'F_h}\right) = \sum_f \Pr\left(d_{h'F_h}, f\right)$$

CTF-CALCULUS: EXAMPLE OF CALCULATIONS

$$\Pr\left(d_{h'F_h}\right) = \sum_f \Pr\left(d_{h'F_h}, f\right)$$

$$(R1) = \sum_f \Pr\left(d_{h'f}, f_h\right)$$

CTF-CALCULUS: EXAMPLE OF CALCULATIONS

$$\begin{aligned}
\Pr\left(d_{h'F_h}\right) &= \sum_f \Pr\left(d_{h'F_h}, f\right) \\
(R1) &= \sum_f \Pr\left(d_{h'f}, f_h\right) \\
&= \sum_f \Pr\left(d_{h'f} \mid f_h\right) \Pr\left(f_h\right)
\end{aligned}$$

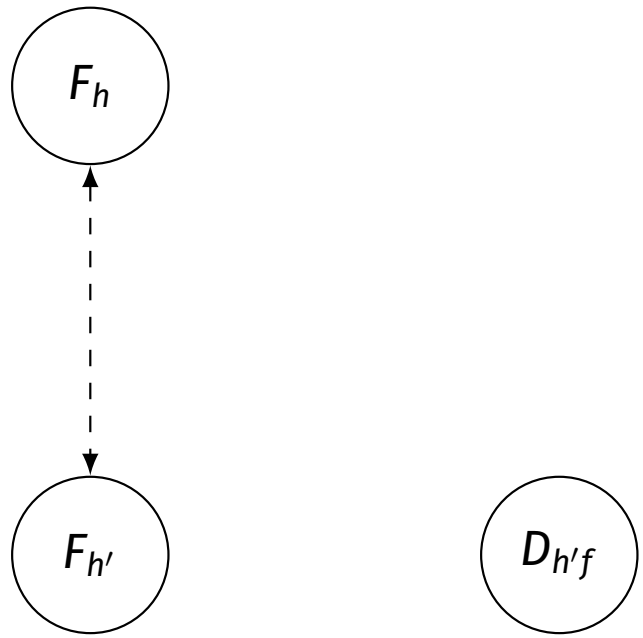
CTF-CALCULUS: EXAMPLE OF CALCULATIONS

$$\Pr\left(d_{h'F_h}\right) = \sum_f \Pr\left(d_{h'F_h}, f\right)$$

$$(R1) = \sum_f \Pr\left(d_{h'f}, f_h\right)$$

$$= \sum_f \Pr\left(d_{h'f} \mid f_h\right) \Pr\left(f_h\right)$$

$$(R2) = \sum_f \Pr\left(d_{h'f} \mid f_{h'}\right) \Pr\left(f_h\right)$$



$$\left(D_{h'f} \perp\!\!\!\perp_d F_h, F'_h\right)_{\mathcal{G}_A}$$

CTF-CALCULUS: EXAMPLE OF CALCULATIONS

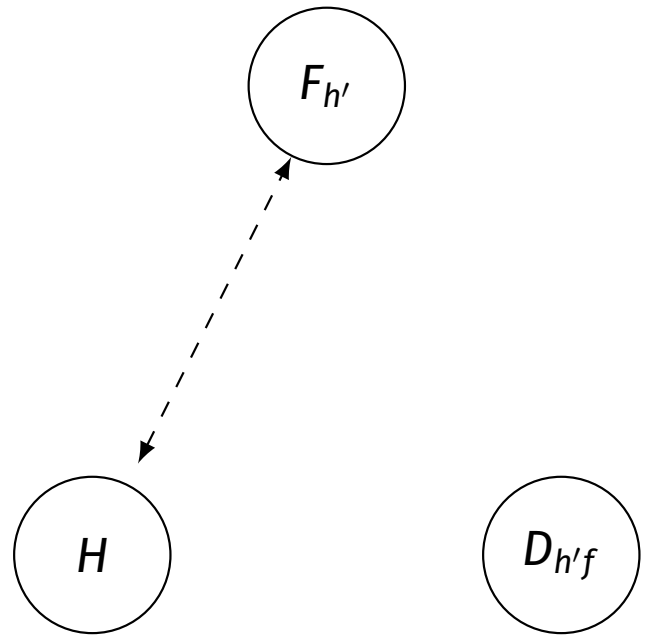
$$\Pr\left(d_{h'F_h}\right) = \sum_f \Pr\left(d_{h'F_h}, f\right)$$

$$(R1) = \sum_f \Pr\left(d_{h'f}, f_h\right)$$

$$= \sum_f \Pr\left(d_{h'f} \mid f_h\right) \Pr\left(f_h\right)$$

$$(R2) = \sum_f \Pr\left(d_{h'f} \mid f_{h'}\right) \Pr\left(f_h\right)$$

$$(R2) = \sum_f \Pr\left(d_{h'f} \mid f_{h'}h'\right) \Pr\left(f_h\right)$$



$$\left(D_{h'f} \perp\!\!\!\perp_d H \mid F_{h'}\right)_{\mathcal{G}_A}$$

CTF-CALCULUS: EXAMPLE OF CALCULATIONS

$$
\begin{aligned}
\Pr\left(d_{h'F_h}\right) &= \sum_f \Pr\left(d_{h'F_h}, f\right) \\
(R1) &= \sum_f \Pr\left(d_{h'f}, f_h\right) \\
&= \sum_f \Pr\left(d_{h'f} \mid f_h\right) \Pr\left(f_h\right) \\
(R2) &= \sum_f \Pr\left(d_{h'f} \mid f_{h'}\right) \Pr\left(f_h\right) \\
(R2) &= \sum_f \Pr\left(d_{h'f} \mid f_{h'} h'\right) \Pr\left(f_h\right) \\
(R1) &= \sum_f \Pr\left(d_{h'f} \mid f, h'\right) \Pr\left(f_h\right)
\end{aligned}
$$

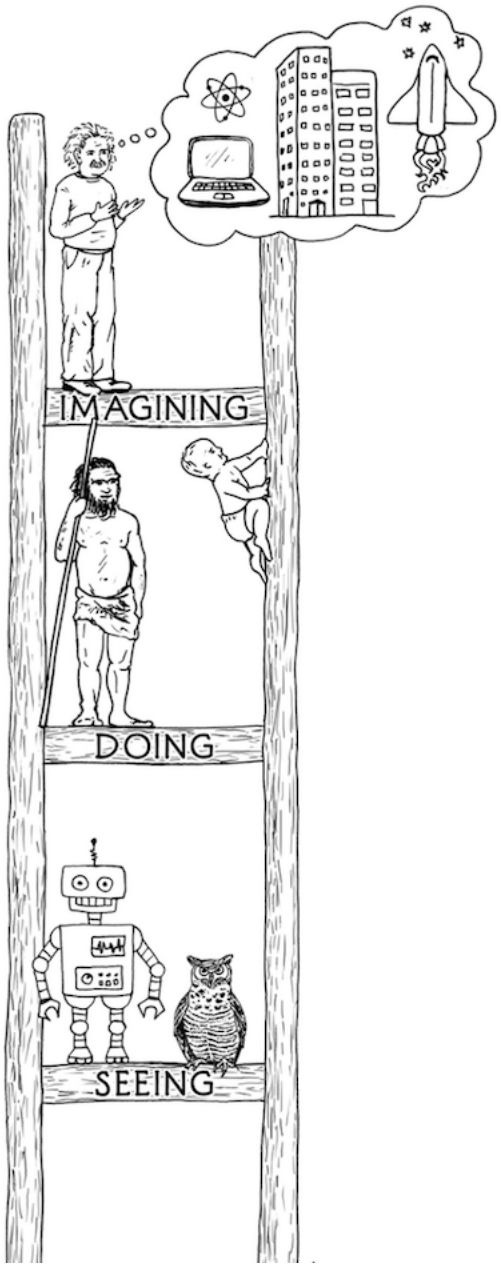
CTF-CALCULUS: EXAMPLE OF CALCULATIONS

$$
\begin{aligned}
\Pr\left(d_{h'F_h}\right) &= \sum_f \Pr\left(d_{h'F_h}, f\right) \\
(R1) &= \sum_f \Pr\left(d_{h'f}, f_h\right) \\
&= \sum_f \Pr\left(d_{h'f} \mid f_h\right) \Pr\left(f_h\right) \\
(R2) &= \sum_f \Pr\left(d_{h'f} \mid f_{h'}\right) \Pr\left(f_h\right) \\
(R2) &= \sum_f \Pr\left(d_{h'f} \mid f_{h'} h'\right) \Pr\left(f_h\right) \\
(R1) &= \sum_f \Pr\left(d_{h'f} \mid f, h'\right) \Pr\left(f_h\right) \\
(R1) &= \sum_f \Pr\left(d \mid f, h'\right) \Pr\left(f_h\right)
\end{aligned}
$$

6

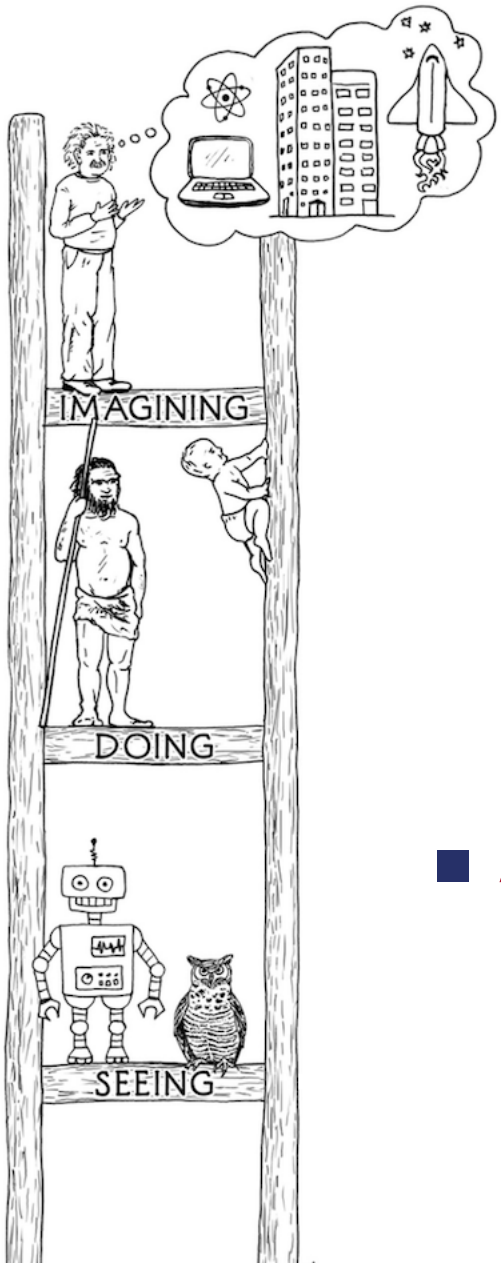
Conclusion

Conclusion



No assumptions

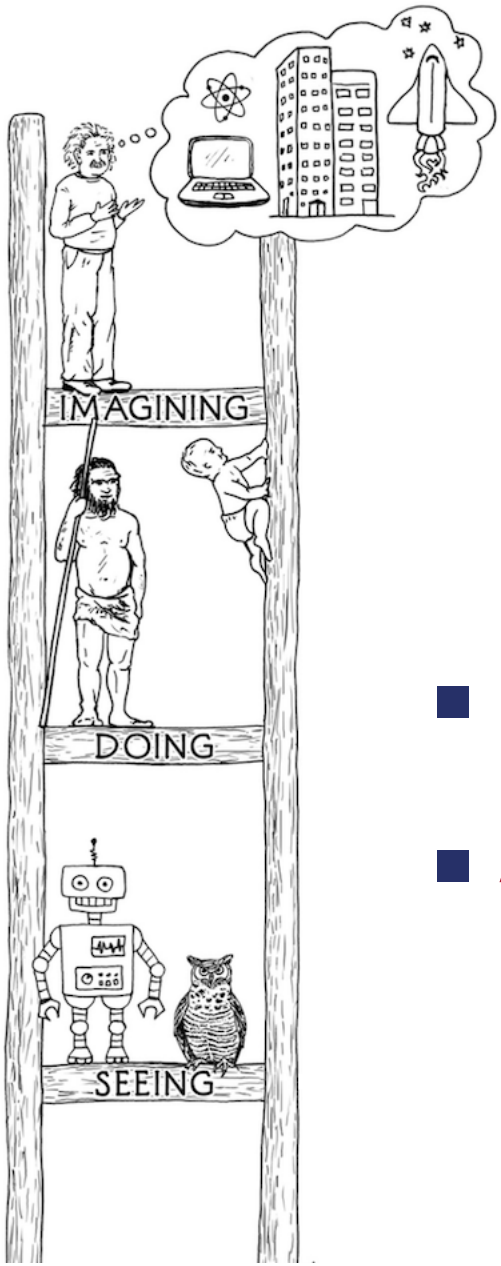
Conclusion



No assumptions

- Associations
  - observational data

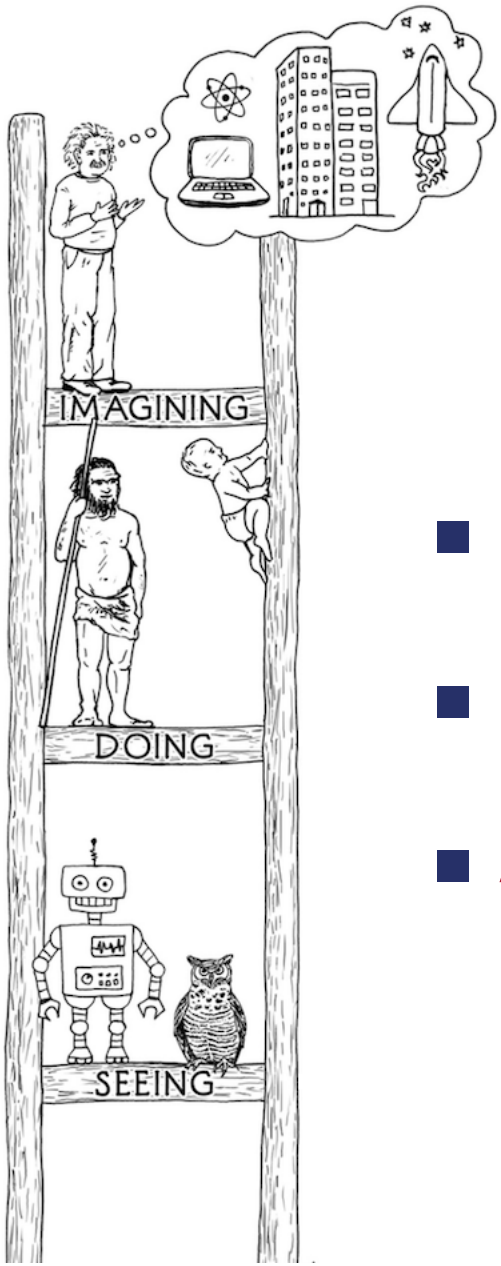
Conclusion



No assumptions

- Interventions
  - interventional data
- Associations
  - observational data

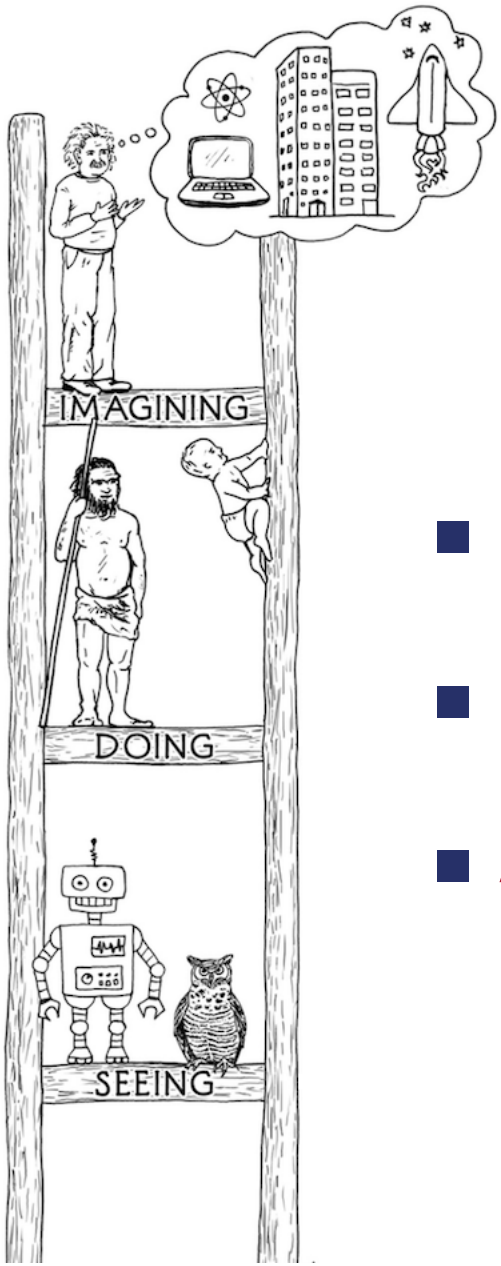
Conclusion



No assumptions

- Counterfactuals
  - Generating process
- Interventions
  - interventional data
- Associations
  - observational data

Conclusion

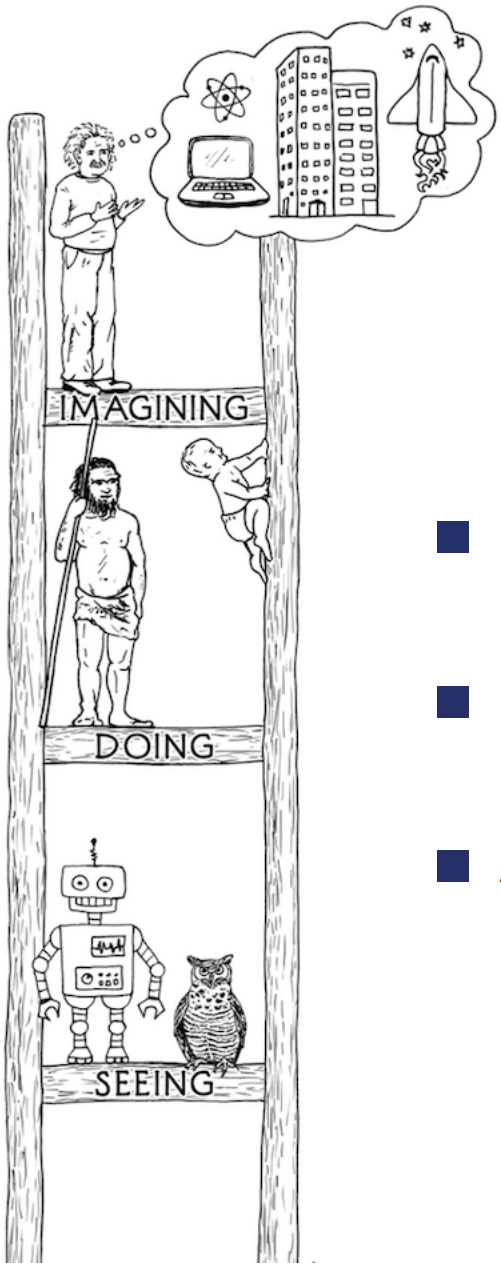


No assumptions

Assumptions: DAG

- Counterfactuals
  - Generating process
- Interventions
  - interventional data
- Associations
  - observational data

Conclusion



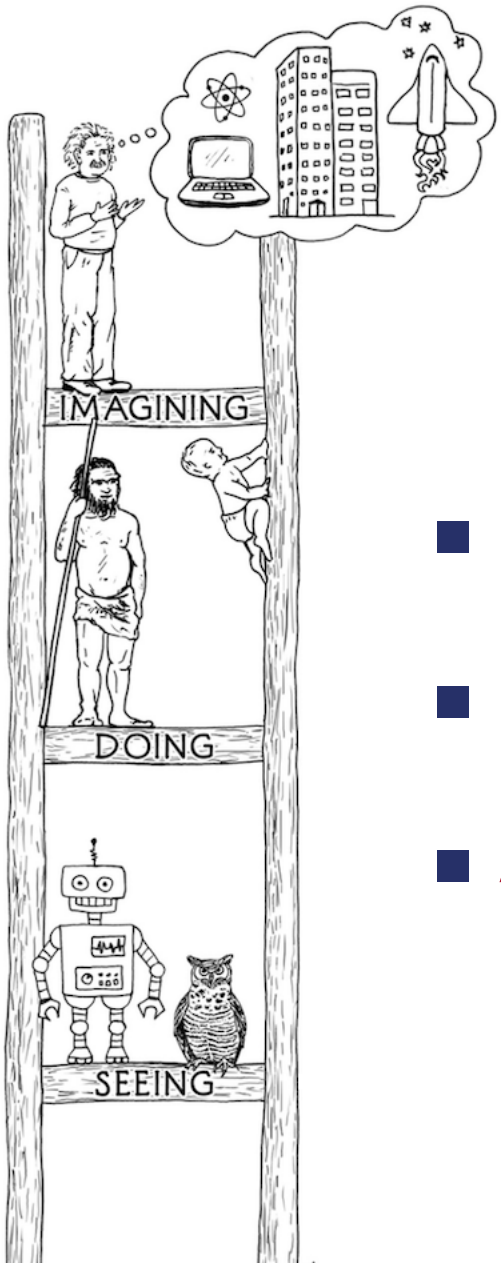
No assumptions

- Counterfactuals
  - Generating process
- Interventions
  - interventional data
- Associations
  - observational data

Assumptions: DAG

do-calculus

Conclusion



No assumptions

- Counterfactuals
  - Generating process
- Interventions
  - interventional data
- Associations
  - observational data

Assumptions: DAG

ctf-calculus

do-calculus

THANK YOU FOR YOUR ATTENTION!

ANY QUESTIONS?

References I

- Balke, Alexander and Judea Pearl (1994). "Probabilistic evaluation of counterfactual queries". In: Proceedings of the Twelfth AAAI National Conference on Artificial AAAI'94. Seattle, Washington: AAAI Press, 230–237.
- Correa, Juan D and Elias Bareinboim (2025). "Counterfactual graphical models: Constraints and inference". In: Forty-second International Conference on Machine Learning.
- Correa, Juan D., Sanghack Lee, and Elias Bareinboim (2021). "Nested counterfactual identification from arbitrary surrogate experiments". In: Proceedings of the 35th International Conference on Neural Inforr NIPS '21. Red Hook, NY, USA: Curran Associates Inc. ISBN: 9781713845393.

References II

- Malinsky, Daniel, Ilya Shpitser, and Thomas Richardson (2019). “A potential outcomes calculus for identifying conditional path-specific effects”. In: The 22nd International Conference on Artificial Intelligence and St PMLR, pp. 3080–3088.
- Pearl, Judea (Dec. 1995). “Causal diagrams for empirical research”. In: Biometrika 82.4, pp. 669–688. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.669.
- — (2001). “Direct and Indirect Effects”. In: Probabilistic and Causal Inference.
- — (2009). Causality: Models, Reasoning, and Inference. Cambridge University Press.
- Richardson, Thomas S and James M Robins (2013). “Single world intervention graphs: a primer”. In: Second UAI workshop on causal structure learning, Bellevue, Wash

References III

Robins, James M and Sander Greenland (1992). “Identifiability and exchangeability for direct and indirect effects”. In: Epidemiology 3.2, pp. 143–155.