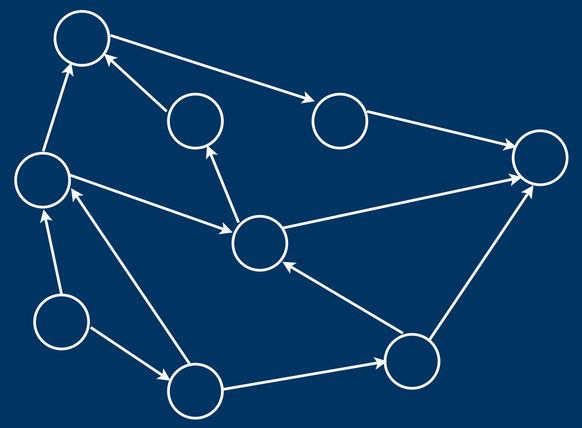


Causation

Frederick Eberhardt

[this slide deck is based on excellent material from many other people, including Peter Spirtes, Thomas Richardson, Daniel Malinsky and Jonas Peters]

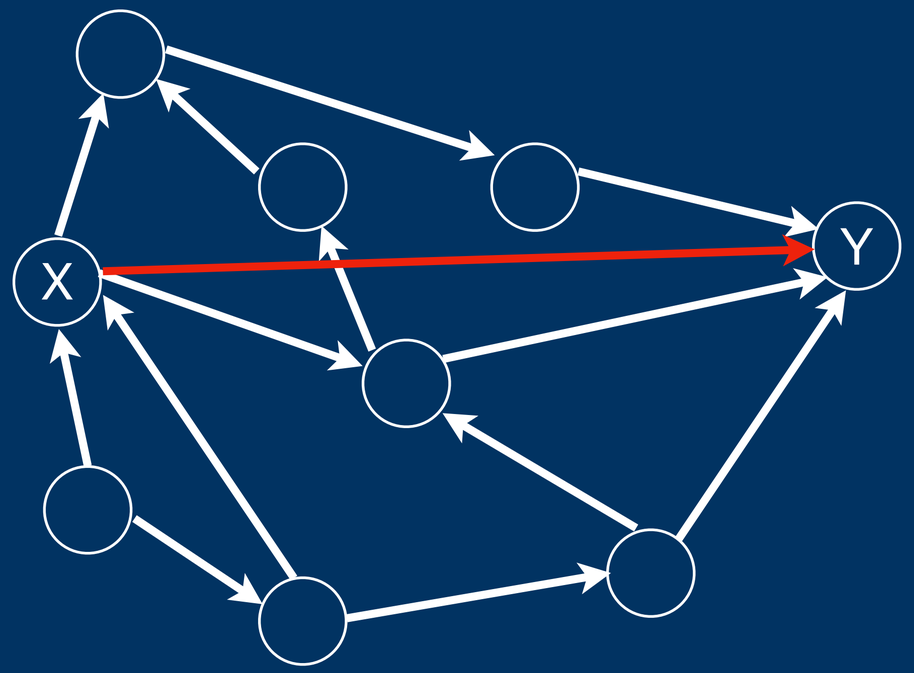
Caltech

Cause and Effect

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

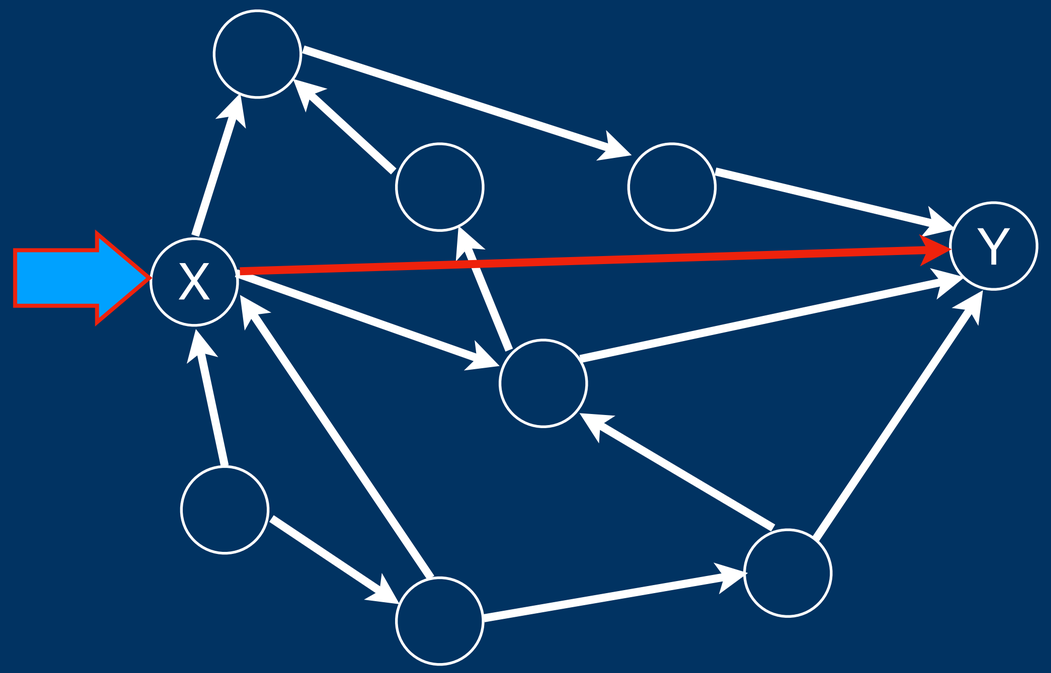
Cause and Effect

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.



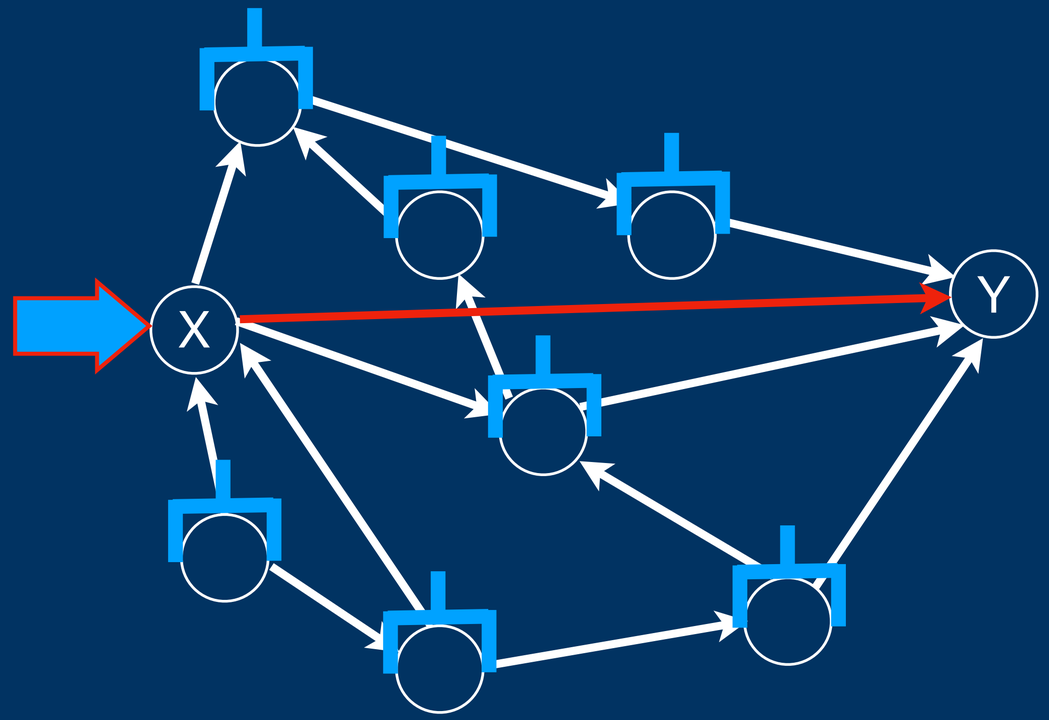
Cause and Effect

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.



Cause and Effect

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.



Cause and Effect

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Potential Outcomes

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	0	1	1	1

We set X to 1 and observed that Y =1

Potential Outcomes

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	0	1	1	1

We set X to 1 and observed that Y =1

If we *had* set X to 1 then Y would have been 1.

If we *had* set X to 0 then Y would have been 0.

Potential Outcomes

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	0	1	1	1

We set X to 1 and observed that Y =1

Consistency: X=x implies Y = Y(x)

If we *had* set X to 1 then Y would have been 1.

If we *had* set X to 0 then Y would have been 0.

Potential Outcomes

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	0	1	1	1

We set X to 1 and observed that Y =1

Consistency: X=x implies Y = Y(x)

If we *had* set X to 1 then Y would have been 1.

If we *had* set X to 0 then Y would have been 0.

➡ Unit treatment effect: $Y_i(x=1) - Y_i(x=0)$

Causal Inference as Missing Data Problem

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	?	1	1	1

We set X to 1 and observed that Y =1

➡ Unit treatment effect: $Y_i(x=1) - Y_i(x=0)$

Causal Inference as Missing Data Problem

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	?	1	1	1
1	0	?	0	0

We set X to 1 and observed that Y =1

Can you experiment on the same unit?

➡ Unit treatment effect: $Y_i(x=1) - Y_i(x=0)$

Causal Inference as Missing Data Problem

- X is a direct cause of Y if there is an intervention on X that results in a change in Y while all other variables have been held fixed by an intervention at some value.

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	?	1	1	1
1	0	?	0	0

We set X to 1 and observed that Y =1

Looks like treatment "helped"!

Can you experiment on the same unit?

➡ Unit treatment effect: $Y_i(x=1) - Y_i(x=0)$

Average Treatment Effect

Unit	Potential outcomes		Actual outcomes	
	Y(x=0)	Y(x=1)	X	Y
1	0	1	1	1
2	0	1	0	0
3	0	0	1	0
4	1	1	1	1
5	1	0	0	1

Average Treatment Effect

Unit	Potential outcomes		Actual outcomes		
	Y(x=0)	Y(x=1)	X	Y	
1	0	1	1	1	Helped
2	0	1	0	0	Helped
3	0	0	1	0	Never recover
4	1	1	1	1	Always recover
5	1	0	0	1	Hurt

Average Treatment Effect

Unit	Potential outcomes		Actual outcomes		
	Y(x=0)	Y(x=1)	X	Y	
1	0	1	1	1	Helped
2	0	1	0	0	Helped
3	0	0	1	0	Never recover
4	1	1	1	1	Always recover
5	1	0	0	1	Hurt

Average Treatment Effect

Unit	Potential outcomes		Actual outcomes		
	Y(x=0)	Y(x=1)	X	Y	
1	?	1	1	1	?
2	0	?	0	0	?
3	?	0	1	0	?
4	?	1	1	1	?
5	1	?	0	1	?

Average Treatment Effect

Unit	Potential outcomes		Actual outcomes		
	Y(x=0)	Y(x=1)	X	Y	
1	?	1	1	1	?
2	0	?	0	0	?
3	?	0	1	0	?
4	?	1	1	1	?
5	1	?	0	1	?

- Average Treatment effect ATE:

$$E_i\,[Y_i(x=1) - Y_i(x=0)] = p(\text{helped}) - p(\text{hurt})$$

Average Treatment Effect

Unit	Potential outcomes		Actual outcomes		
	Y(x=0)	Y(x=1)	X	Y	
1	?	1	1	1	?
2	0	?	0	0	?
3	?	0	1	0	?
4	?	1	1	1	?
5	1	?	0	1	?

- Average Treatment effect ATE:

$$E_i\,[Y_i(x=1) - Y_i(x=0)] = p(\text{helped}) - p(\text{hurt})$$

% improvement if everyone was treated vs. no-one treated.

Average Treatment Effect

Unit	Potential outcomes		Actual outcomes		
	Y(x=0)	Y(x=1)	X	Y	
1	?	1	1	1	?
2	0	?	0	0	?
3	?	0	1	0	?
4	?	1	1	1	?
5	1	?	0	1	?

- Average Treatment effect ATE:

$$E[Y(x=1) - Y(x=0)] = p(\text{helped}) - p(\text{hurt})$$

% improvement if everyone was treated vs. no-one treated.

$$= E[Y(x=1)] - E[Y(x=0)]$$

$$= E[Y(x=1) \mid X=1] - E[Y(x=0) \mid X=0]$$

$$= E[Y|X=1] - E[Y|X=0]$$

by consistency

If X is assigned randomly then X is independent of Y(x=1) and of Y(x=0).

Average Treatment Effect is identified

Unit	Potential outcomes		Actual outcomes		
	Y(x=0)	Y(x=1)	X	Y	
1	?	1	1	1	?
2	0	?	0	0	?
3	?	0	1	0	?
4	?	1	1	1	?
5	1	?	0	1	?

- Average Treatment effect ATE:

$$E[Y(x=1) - Y(x=0)] = p(\text{helped}) - p(\text{hurt})$$

% improvement if everyone was treated vs. no-one treated.

$$= E[Y(x=1)] - E[Y(x=0)]$$

$$= E[Y(x=1) \mid X=1] - E[Y(x=0) \mid X=0]$$

$$= E[Y|X=1] - E[Y|X=0]$$

If X is assigned randomly then X is independent of Y(x=1) and of Y(x=0).

by consistency

➡ ATE is identified from the observed P(Y|do(X))!!

Randomization

Randomizing X specifies the intervention distribution P(Y | do(X)), which characterizes what we mean by X causing Y…

- Average Treatment effect ATE:

$$E[Y(x=1) - Y(x=0)] = p(\text{helped}) - p(\text{hurt})$$

% improvement if everyone was treated vs. no-one treated.

$$= E[Y(x=1)] - E[Y(x=0)]$$

$$= E[Y(x=1) \mid X=1] - E[Y(x=0) \mid X=0]$$

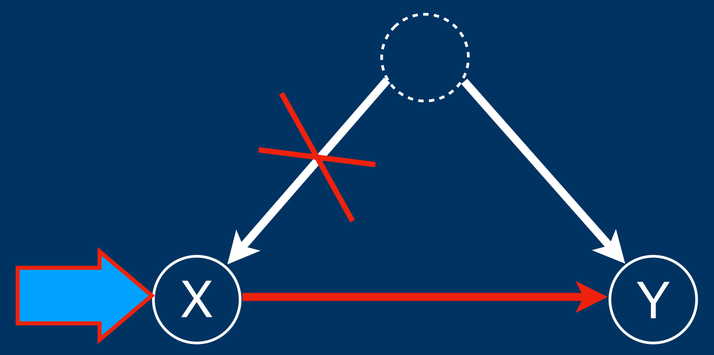
$$= E[Y|X=1] - E[Y|X=0]$$

If X is assigned randomly then X is independent of Y(x=1) and of Y(x=0).

by consistency

➡ ATE is identified from the observed P(Y|do(X))!!

Randomization breaks Confounding



Randomizing X specifies the intervention distribution P(Y | do(X)), which characterizes what we mean by X causing Y…even when there is confounding (due to measured or unmeasured common causes).

- Average Treatment effect ATE:

$$E[Y(x=1) - Y(x=0)] = p(\text{helped}) - p(\text{hurt})$$

% improvement if everyone was treated vs. no-one treated.

$$= E[Y(x=1)] - E[Y(x=0)]$$

$$= E[Y(x=1) \mid X=1] - E[Y(x=0) \mid X=0]$$

$$= E[Y|X=1] - E[Y|X=0]$$

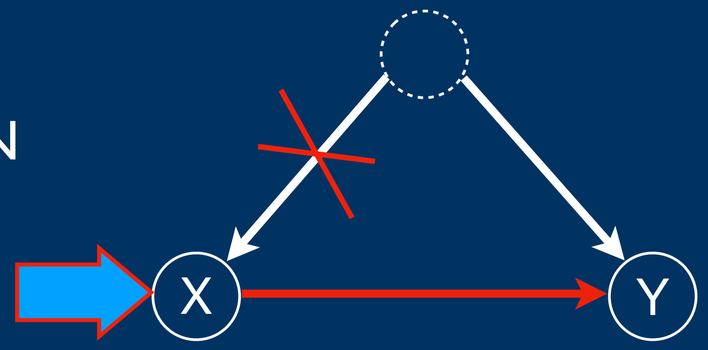
If X is assigned randomly then X is independent of Y(x=1) and of Y(x=0).

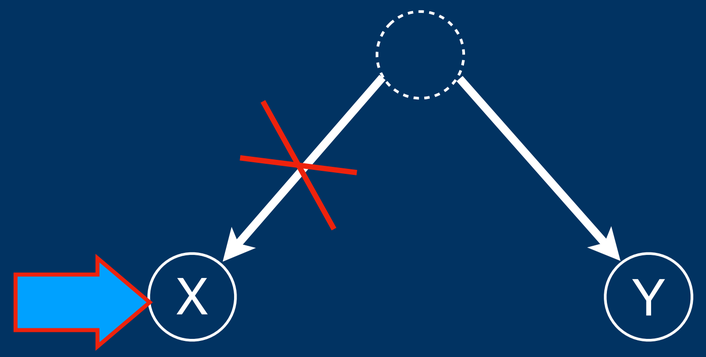
by consistency

➡ ATE is identified from the observed P(Y|do(X))!!

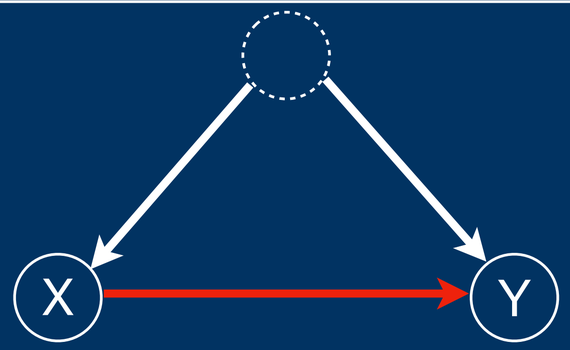
Do-operator

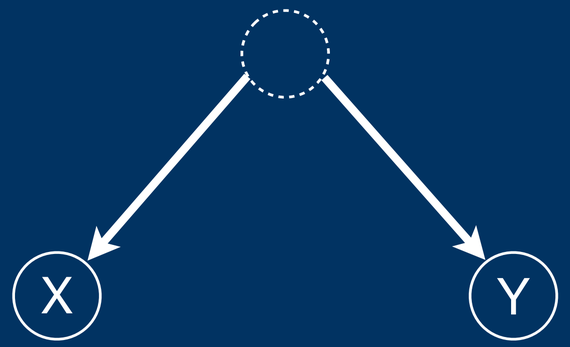
INTERVENTION





OBSERVATION





$P(Y|do(X)) \neq P(Y|X)$

$\neq P(Y)$

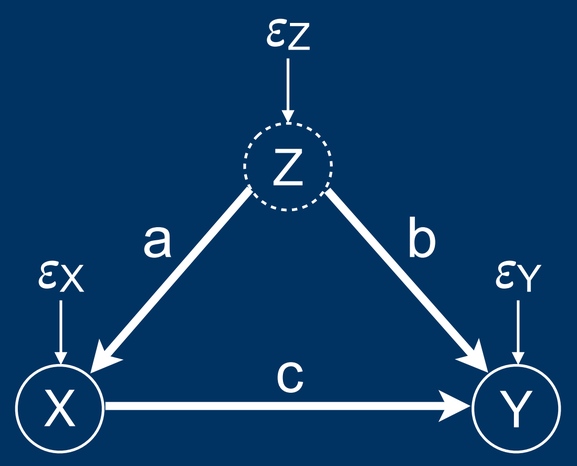
$P(Y|do(X)) = P(Y|X)$

$\neq P(Y)$

$P(Y|do(X)) = P(Y)$

$\neq P(Y|X)$

(Linear Gaussian) Structural Equation Model

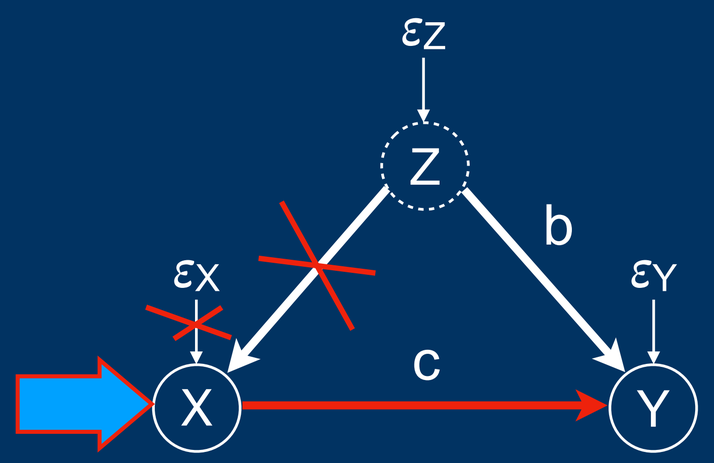


$Z = \varepsilon_Z$ $\quad$ $\varepsilon_Z \sim N(0,1)$

$X = aZ + \varepsilon_X$ $\quad$ $\varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y.$ $\quad$ $\varepsilon_Y \sim N(0,1)$

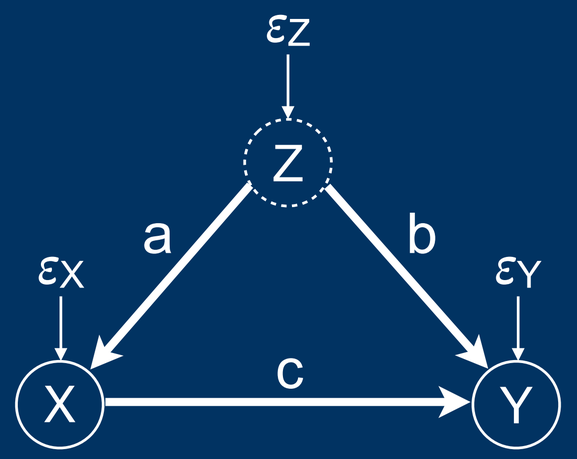
(Linear Gaussian) Structural Equation Model



$Z = \varepsilon_Z$ $\quad$ $\varepsilon_Z \sim N(0,1)$

$X = r$ $\quad$ $r \sim$ intervention distribution

$Y = bZ + cX + \varepsilon_Y$ $\quad$ $\varepsilon_Y \sim N(0,1)$



$$\left.\begin{aligned} Z &= \varepsilon_Z & \varepsilon_Z &\sim N(0,1) \\ X &= aZ + \varepsilon_X & \varepsilon_X &\sim N(0,1) \\ Y &= bZ + cX + \varepsilon_Y & \varepsilon_Y &\sim N(0,1) \end{aligned}\right\} \quad \boldsymbol{\Sigma}_\varepsilon = \begin{bmatrix} \ddots & 0 \\ 0 & \ddots \end{bmatrix}$$

The Blind Patient

$$B = T\varepsilon_B + (1-T)(1-\varepsilon_B) \qquad \varepsilon_B \sim \mathrm{Ber}(0.01)$$

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

This implies that $\varepsilon_B = 1$.

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

This implies that $\varepsilon_B = 1$.

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

COUNTERFACTUAL:

What would have happened if T=0?

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

~~This implies that $\varepsilon_B = 1$.~~

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

COUNTERFACTUAL:

What would have happened if T=0?

B=0; the patient would not have gone blind!

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

~~This implies that $\varepsilon_B = 1$.~~

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

COUNTERFACTUAL:

What would have happened if T=0?

B=0; the patient would not have gone blind!

So did the doctor make an error?

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

This implies that $\varepsilon_B = 1$.

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

COUNTERFACTUAL:

What would have happened if T=0?

B=0; the patient would not have gone blind!

So did the doctor make an error?

Average Treatment Effect: P(B =0 | do(T=1)) vs. P(B =0 | do(T=0))

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

This implies that $\varepsilon_B = 1$.

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

COUNTERFACTUAL:

What would have happened if T=0?

B=0; the patient would not have gone blind!

So did the doctor make an error?

Average Treatment Effect: $P(B=0 \mid do(T=1))$ vs. $P(B=0 \mid do(T=0))$ $= \Sigma_\varepsilon P(B \mid do(T), \varepsilon_B)\, P(\varepsilon_B)$

The Blind Patient

$B = T\varepsilon_B + (1-T)(1-\varepsilon_B)$ $\varepsilon_B \sim Ber(0.01)$

This implies that $\varepsilon_B = 1$.

INTERVENTION on unit: Patient was treated: T=1 Outcome: Patient went blind: B=1

COUNTERFACTUAL:

What would have happened if T=0?

B=0; the patient would not have gone blind!

So did the doctor make an error?

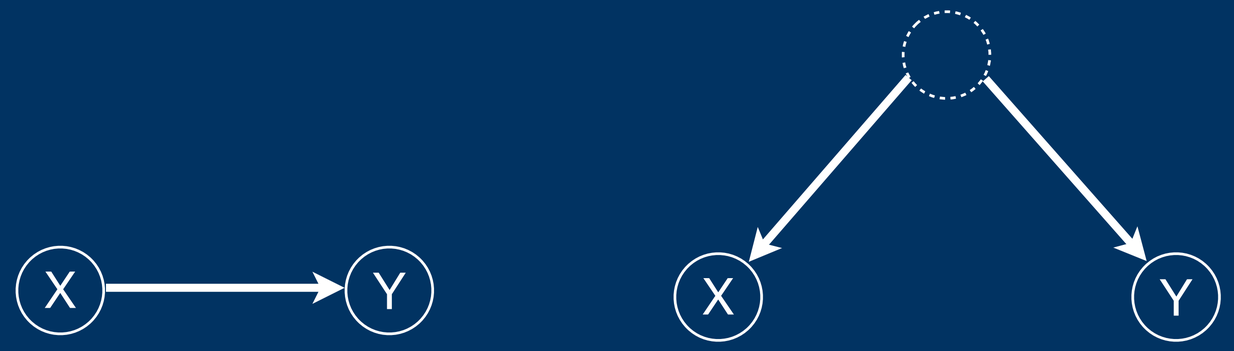
Average Treatment Effect: $P(B=0 \mid do(T=1))$ vs. $P(B=0 \mid do(T=0))$ $= \Sigma_\varepsilon P(B \mid do(T), \varepsilon_B)\, P(\varepsilon_B)$

0.01 vs. 0.99

No! The patient was extremely unlucky.

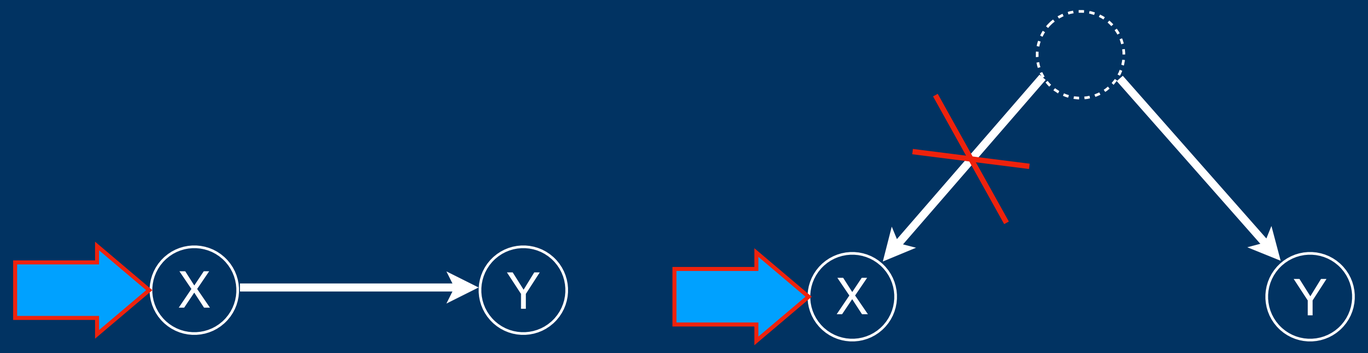
Causality

Prediction



P(Y | X)

Intervention



P(Y | do(X))

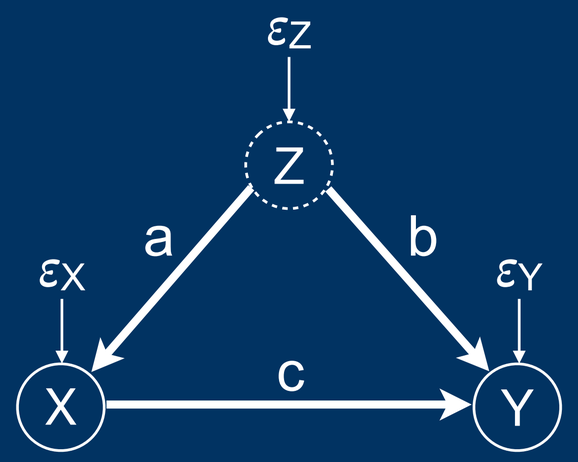
Counterfactual

We set X = 1 and observed Y=1.

What would Y have been if X has been set to 0?

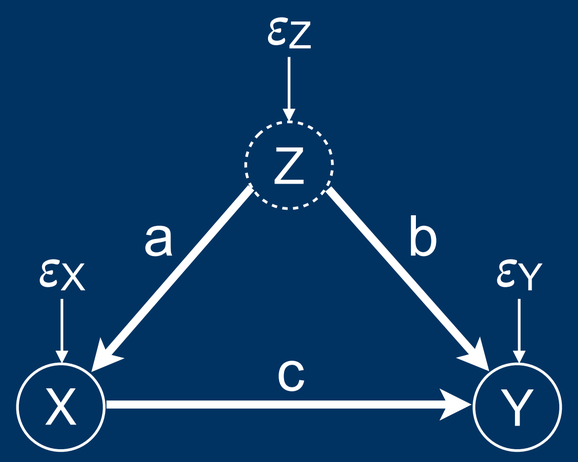
Y(x=0)

Linear Gaussian Model



$$\left.\begin{aligned} Z &= \varepsilon_Z & \varepsilon_Z &\sim N(0,1) \\ X &= aZ + \varepsilon_X & \varepsilon_X &\sim N(0,1) \\ Y &= bZ + cX + \varepsilon_Y. & \varepsilon_Y &\sim N(0,1) \end{aligned}\right\} \quad \boldsymbol{\Sigma}_\varepsilon = \begin{bmatrix} \ddots & 0 \\ 0 & \ddots \end{bmatrix}$$

Matrices of a LinGauss Model



$$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$$

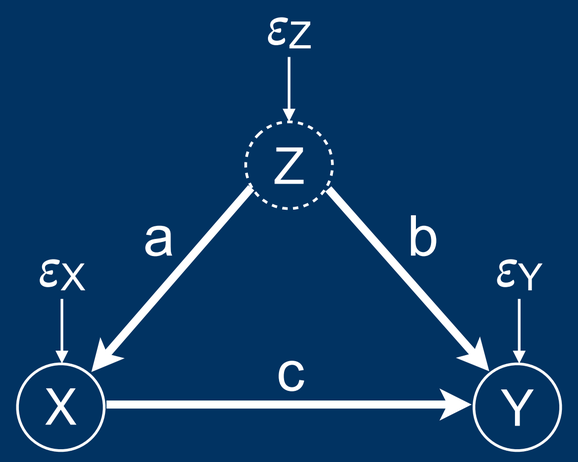
$$X = aZ + \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$$

$$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$$

$$\mathbf{\Sigma}_\varepsilon = \begin{bmatrix} \ddots & 0 \\ 0 & \ddots \end{bmatrix}$$

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 & 0 & a \\ c & 0 & b \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} \epsilon_X \\ \epsilon_Y \\ \epsilon_Z \end{pmatrix}$$

Matrices of a LinGauss Model



$$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$$

$$X = aZ + \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$$

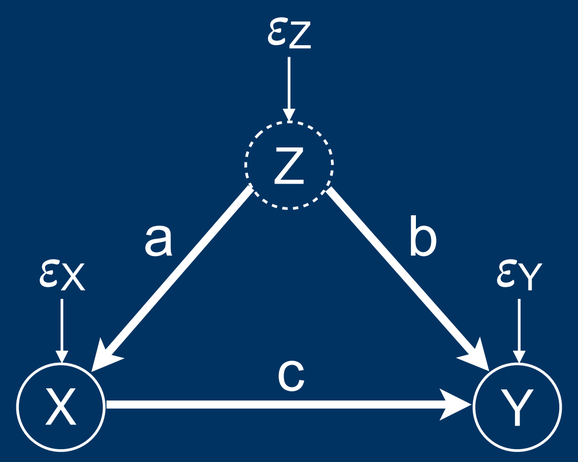
$$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$$

$$\mathbf{\Sigma}_\varepsilon = \begin{bmatrix} \ddots & 0 \\ 0 & \ddots \end{bmatrix}$$

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 & 0 & a \\ c & 0 & b \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} \epsilon_X \\ \epsilon_Y \\ \epsilon_Z \end{pmatrix}$$

Or simply: $\mathbf{x} = \mathbf{B}\mathbf{x} + \epsilon$

Matrices of a LinGauss Model



$$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$$

$$X = aZ + \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$$

$$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$$

$$\mathbf{\Sigma}_\varepsilon = \begin{bmatrix} \ddots & 0 \\ 0 & \ddots \end{bmatrix}$$

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 & 0 & a \\ c & 0 & b \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} \epsilon_X \\ \epsilon_Y \\ \epsilon_Z \end{pmatrix}$$

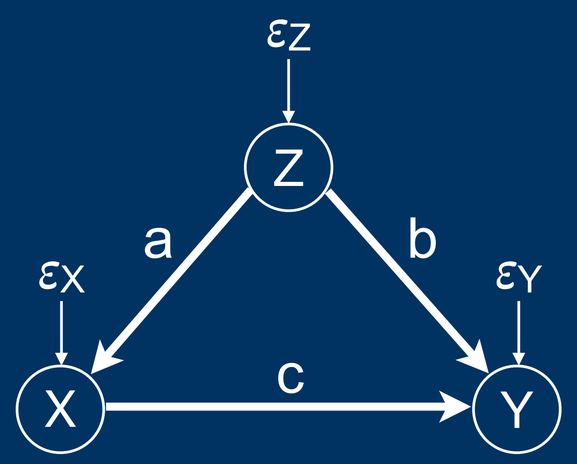
Or simply: $\mathbf{x} = \mathbf{B}\mathbf{x} + \epsilon$

Covariance matrix $cov(\mathbf{x})$ specifies the variances of each variable along the diagonal and the covariation between any pair of variables due to all directed or common cause pathways connecting them.

Correlation matrix $cor(\mathbf{x})$ rescales the covariance matrix such that all variables have variance 1.

Precision matrix $(cov(\mathbf{x}))^{-1}$ specifies the partial covariation between each pair of variables *conditional on all other variables.*

LinGauss as Additive Noise Model



$Z = \varepsilon_Z$ $\quad \varepsilon_Z \sim N(0,1)$

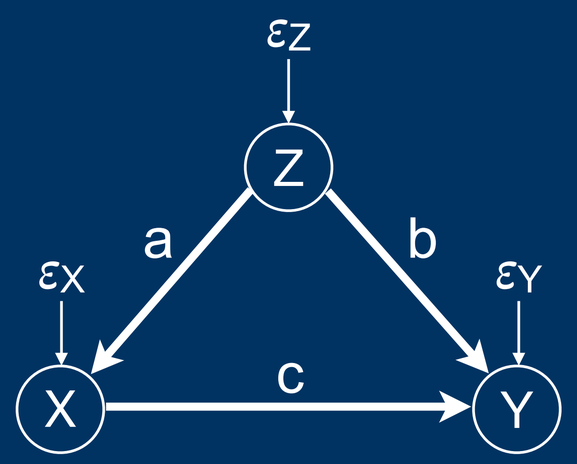
$X = aZ + \varepsilon_X$ $\quad \varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y.$ $\quad \varepsilon_Y \sim N(0,1)$

$$X_i = f(pa(X_i)) + \varepsilon_{X_i}$$

➡ Each variable is a function of its *pa*rents plus an independent additive noise term

LinGauss as Additive Noise Model



$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$

$X = aZ + \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$

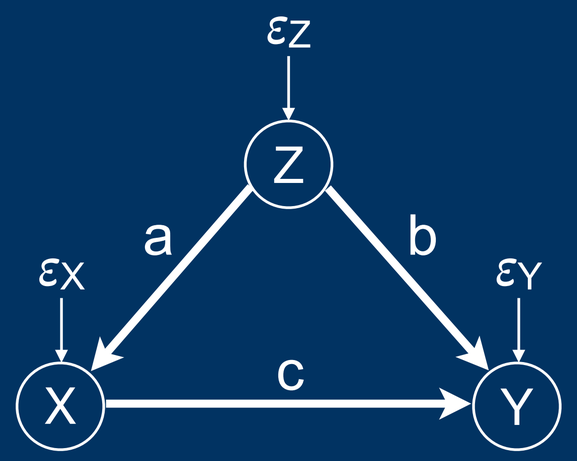
$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$

linear Gaussian additive noise model

$$X_i = f(pa(X_i)) + \varepsilon_{X_i}$$

➡ Each variable is a function of its *pa*rents plus an independent additive noise term

LinGauss as generic BayesNet



$Z = \varepsilon_Z$ $\quad\quad$ $\varepsilon_Z \sim N(0,1)$

$X = aZ + \varepsilon_X$ $\quad\quad$ $\varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y$ $\quad\quad$ $\varepsilon_Y \sim N(0,1)$

linear Gaussian additive noise model

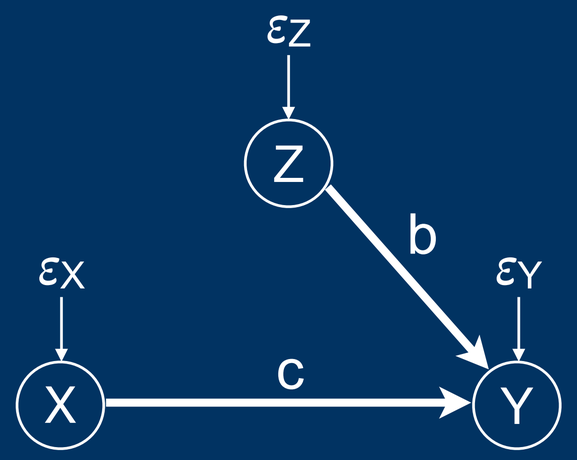
$X_i = f(pa(X_i)) + \varepsilon_{X_i}$

➡ Each variable is a function of its *parents* plus an independent additive noise term

More generally: $$P(\mathbf{V}) = \prod_{V_i \in V} P(V_i \mid pa(V_i))$$

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

A more interesting structure…



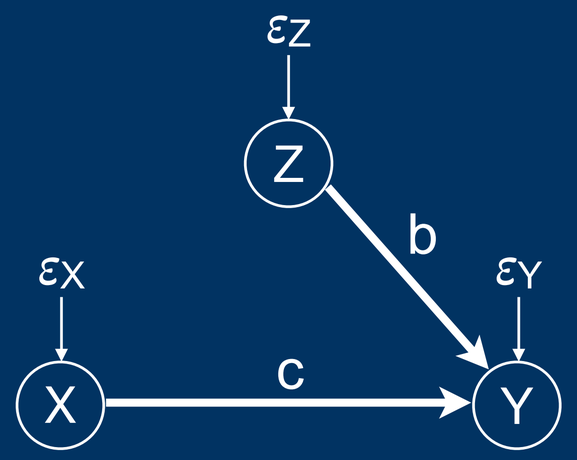
$Z = \varepsilon_Z$ $\quad\quad$ $\varepsilon_Z \sim N(0,1)$

$X = \varepsilon_X$ $\quad\quad$ $\varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y.$ $\quad\quad$ $\varepsilon_Y \sim N(0,1)$

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

A more interesting structure…



$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$

$X = \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$

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

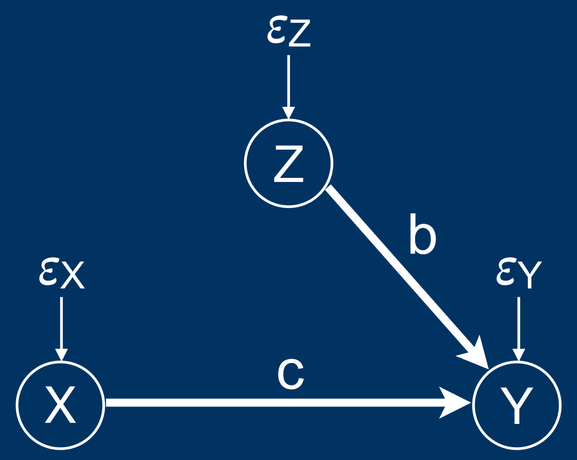
$X \perp\!\!\!\perp Z \qquad X \not\perp\!\!\!\perp Y \qquad Z \not\perp\!\!\!\perp Y$

X is probabilistically independent of Z

X is probabilistically dependent on Z

Y is probabilistically dependent on Z

A more interesting structure…



$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$

$X = \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$

Now: $P(X,Y,Z) = P(Z)P(X)P(Y|X,Z)$

$X \perp\!\!\!\perp Z \qquad X \not\perp\!\!\!\perp Y \qquad Z \not\perp\!\!\!\perp Y$

X is probabilistically independent of Z

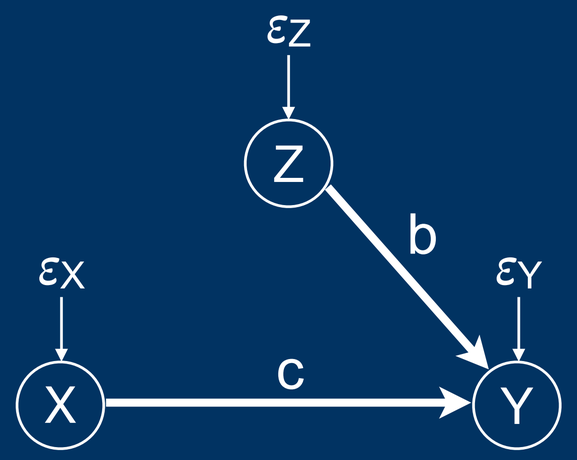
Y is probabilistically dependent on Z

X is probabilistically dependent on Z

In a linear Gaussian model this means:

$cor(X,Z) = 0 \quad cor(X,Y) \neq 0 \quad cor(Y,Z) \neq 0$

A more interesting structure…



$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$

$X = \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$

Now: $P(X,Y,Z) = P(Z)P(X)P(Y|X,Z)$

$X \perp\!\!\!\perp Z \qquad X \not\perp\!\!\!\perp Y \qquad Z \not\perp\!\!\!\perp Y \qquad X \overset{?}{\perp\!\!\!\perp} Z | Y$

X is probabilistically independent of Z

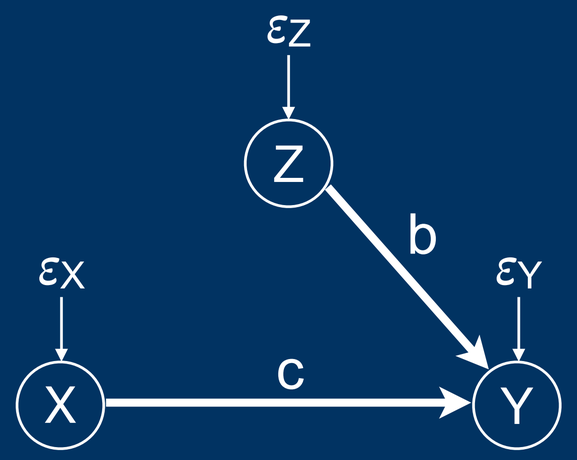
Y is probabilistically dependent on Z

X is probabilistically dependent on Z

In a linear Gaussian model this means:

$cor(X,Z) = 0 \quad cor(X,Y) \neq 0 \quad cor(Y,Z) \neq 0 \qquad cor(X,Z|Y)$

A more interesting structure…



$Z = \varepsilon_Z \qquad \varepsilon_Z \sim N(0,1)$

$X = \varepsilon_X \qquad \varepsilon_X \sim N(0,1)$

$Y = bZ + cX + \varepsilon_Y \qquad \varepsilon_Y \sim N(0,1)$

Now: $P(X,Y,Z) = P(Z)P(X)P(Y|X,Z)$

$X \perp\!\!\!\perp Z \qquad X \not\perp\!\!\!\perp Y \qquad Z \not\perp\!\!\!\perp Y \qquad X \not\perp\!\!\!\perp Z | Y$

X is probabilistically independent of Z

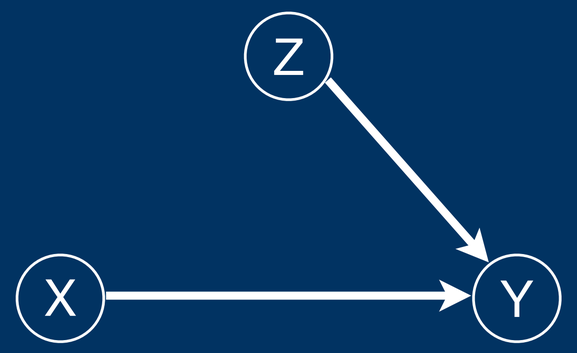
Y is probabilistically dependent on Z

X is probabilistically dependent on Z

In a linear Gaussian model this means:

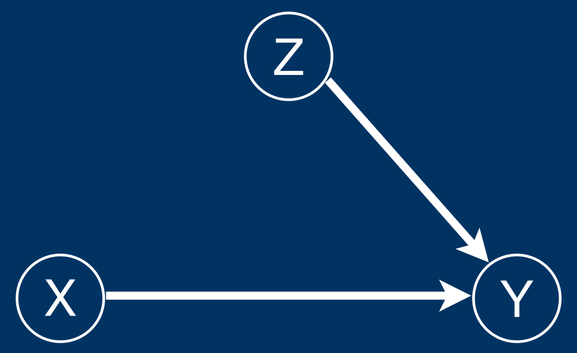
$cor(X,Z) = 0 \quad cor(X,Y) \neq 0 \quad cor(Y,Z) \neq 0 \qquad cor(X,Z|Y) \neq 0$

Unshielded Collider / v-structure

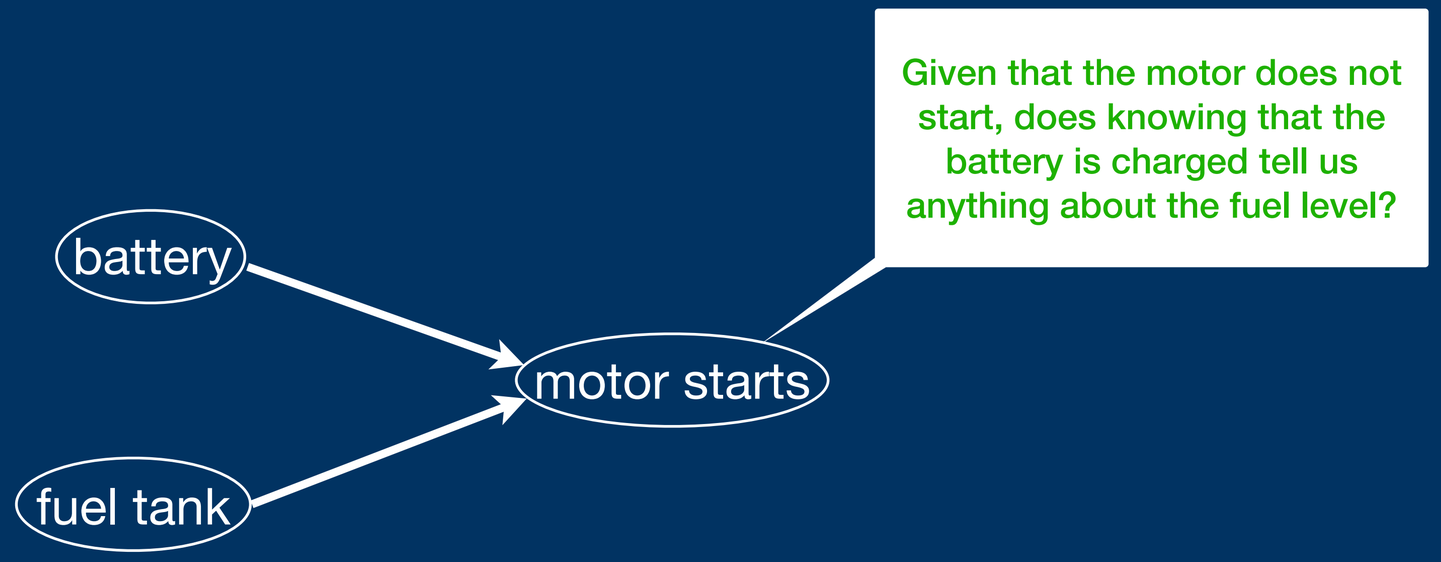


Why $X \not\perp\!\!\!\perp Z \mid Y$?

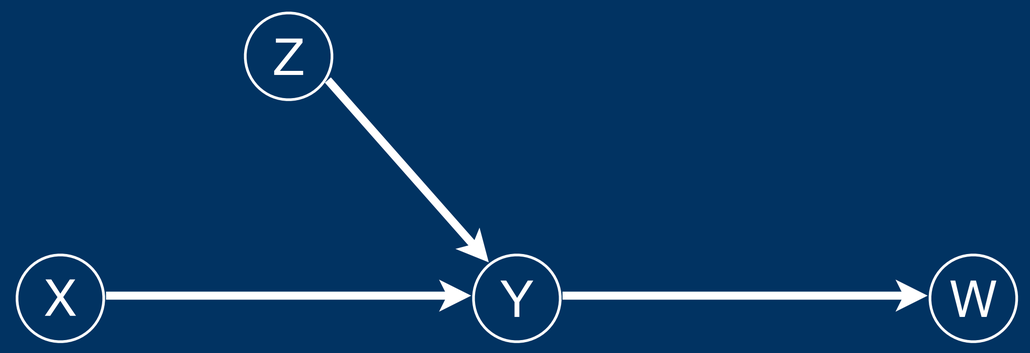
Unshielded Collider / v-structure



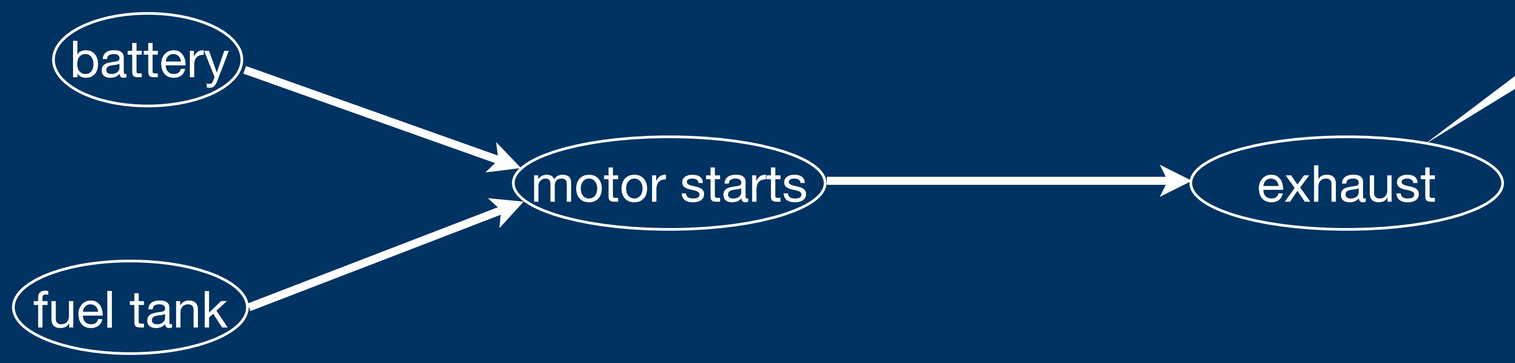
Why $X \not\perp\!\!\!\perp Z \mid Y$?



Unshielded Collider / v-structure

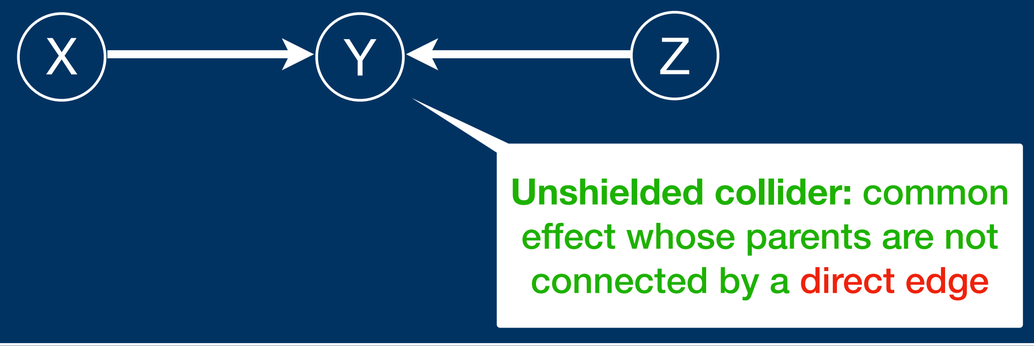


$$X \not\perp\!\!\!\perp Z \mid W \quad !$$

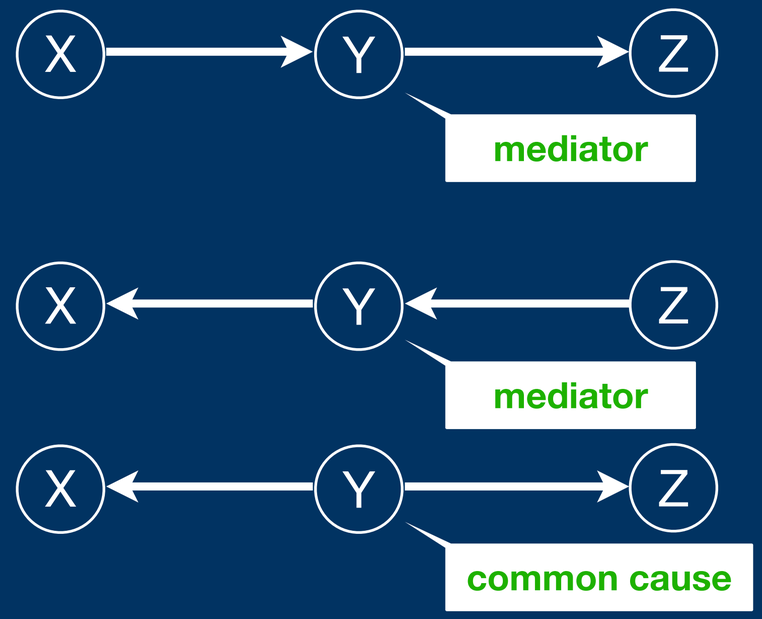


Given that no exhaust is emitted, does knowing that the battery is charged tell us anything about the fuel level?

Unshielded Collider / v-structure

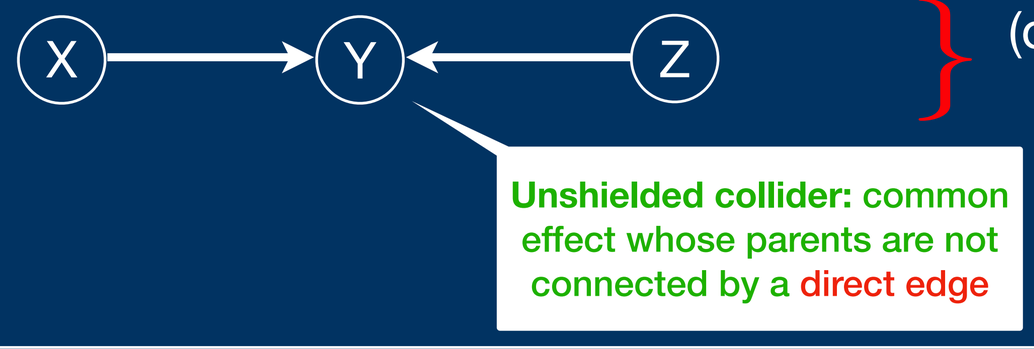


$$X \perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Y$$
$$Y \not\perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Z \mid Y$$



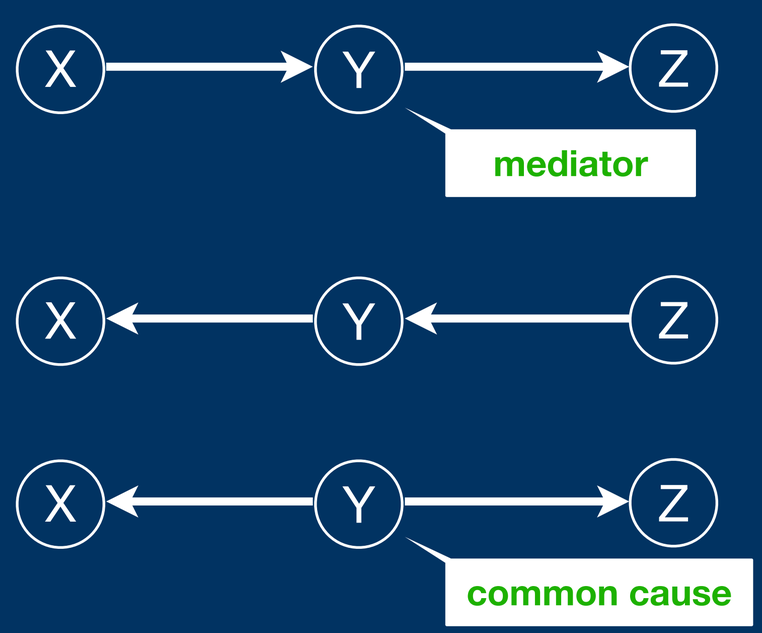
$$X \not\perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Y$$
$$Y \not\perp\!\!\!\perp Z$$
$$X \perp\!\!\!\perp Z \mid Y$$

Unshielded Collider / v-structure



(observational) Markov Equivalence Class

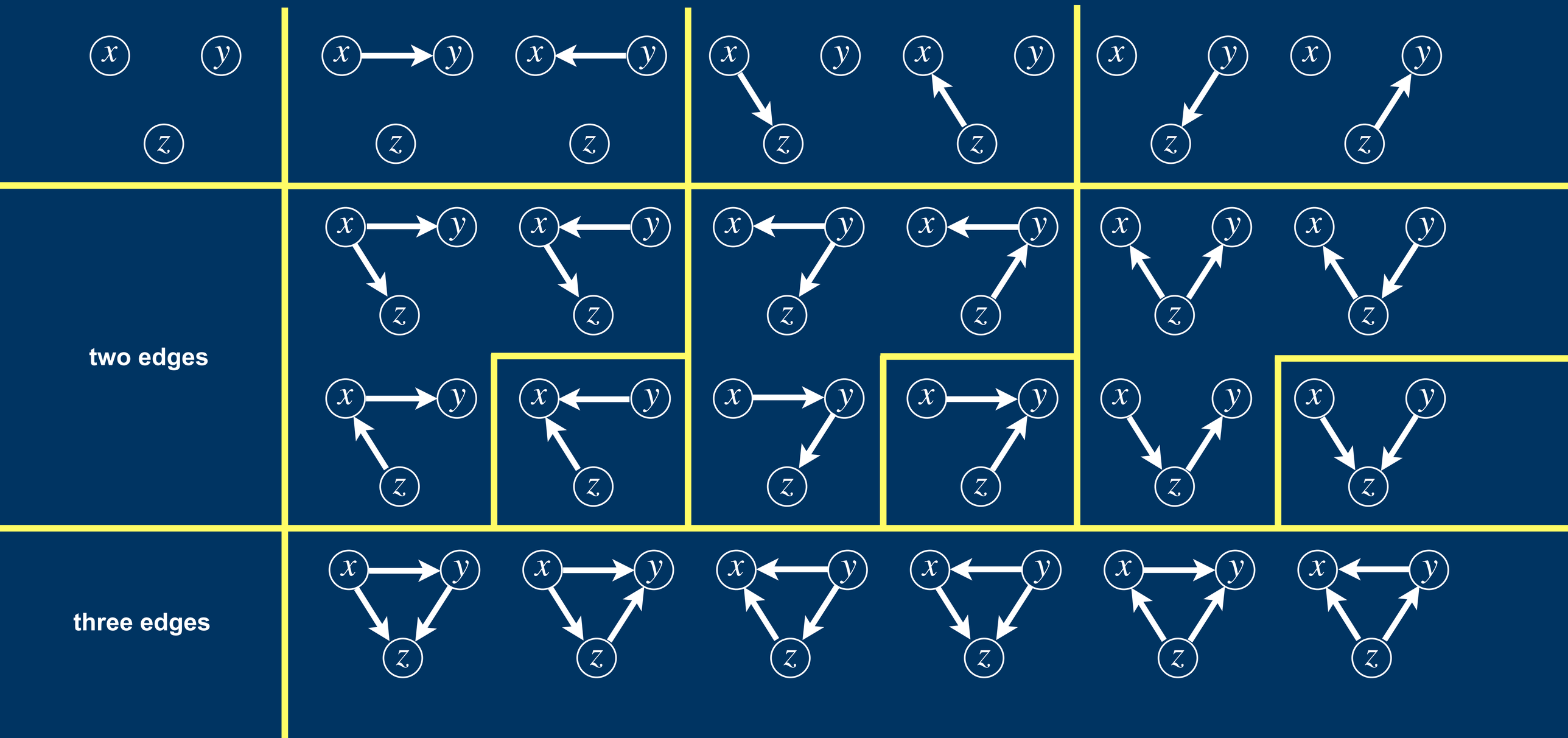
$$X \perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Y$$
$$Y \not\perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Z \mid Y$$



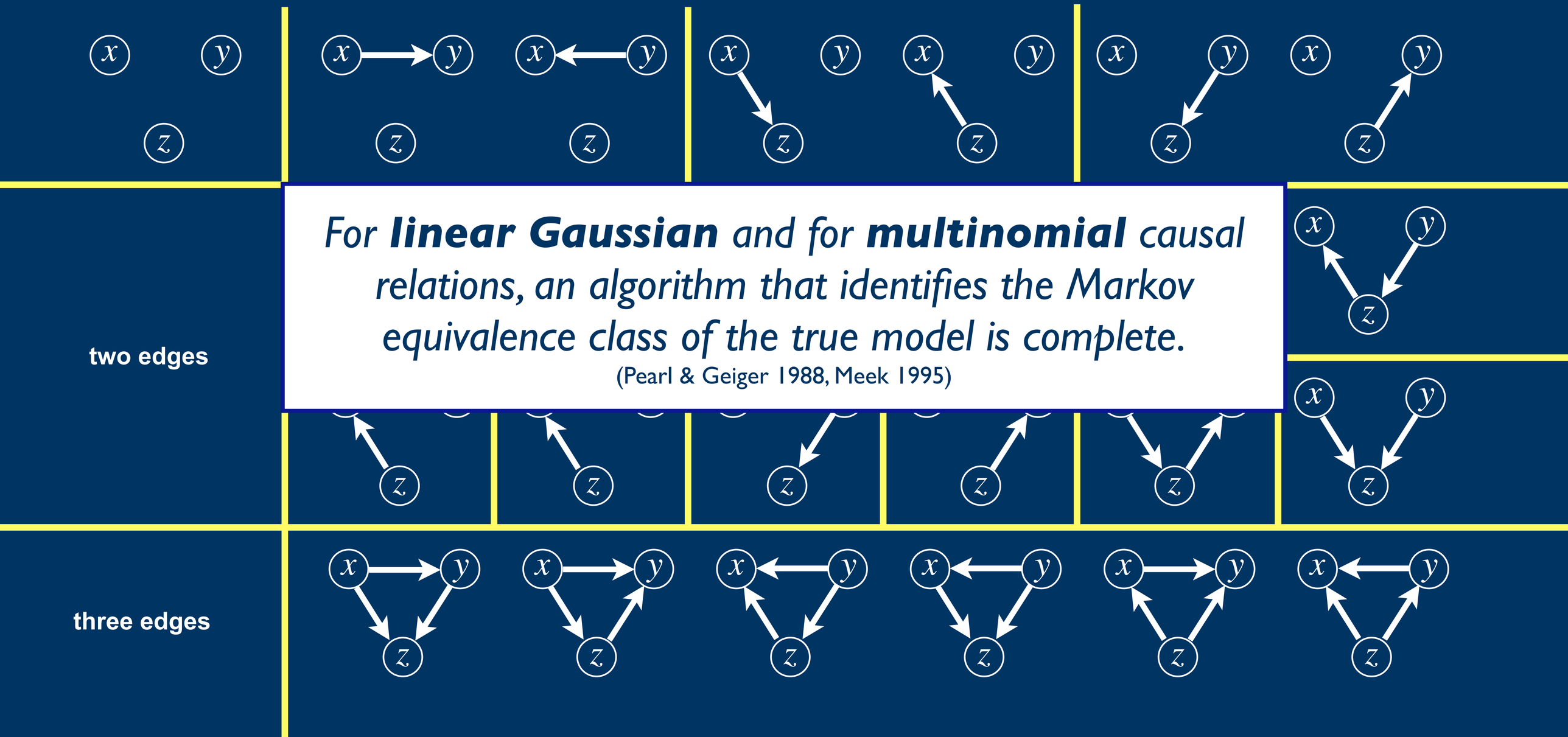
(observational) Markov Equivalence Class

$$X \not\perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Y$$
$$Y \not\perp\!\!\!\perp Z$$
$$X \perp\!\!\!\perp Z \mid Y$$

Equivalence Classes of Causal Models Over Three Variables



Equivalence Classes of Causal Models Over Three Variables



*For **linear Gaussian** and for **multinomial** causal relations, an algorithm that identifies the Markov equivalence class of the true model is complete.*

(Pearl & Geiger 1988, Meek 1995)

Number of Graphs over p variables

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263

http://oeis.org/A003024

Number of Directed Acyclic Graphs (DAGs) over p variables

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263

http://oeis.org/A003024

BUT:
The average number of DAGs *per* Markov equivalence class appears to be bounded by 5, that is, it is independent of the number of variables p!

➡ An algorithm that identifies the Markov equivalence class of the true generating DAG is expected to reduce the candidate graphs to just a handful!

How to identify the data generating graph?

How to identify the data generating graph?

For each possible directed acyclic graph over the variables, test whether its independence structure is consistent with the independence structure observed in the data?

How to identify the data generating graph?

For each possible directed acyclic graph over the variables, test whether its independence structure is consistent with the independence structure observed in the data?

Only need to check this for each Markov equivalence class!

How to identify the data generating graph?

For each possible directed acyclic graph over the variables, test whether its independence structure is consistent with the independence structure observed in the data?

Only need to check this for each Markov equivalence class!

How many are there of those? How do we generate all of them?

PC algorithm (main ideas)

- $X \not\perp\!\!\!\perp Y \ (|\,\mathbf{C})$ implies there is some connection between the variables

- $X \perp\!\!\!\perp Y \ (|\,\mathbf{C})$ implies (at least) that $X \not\rightarrow Y$ and $X \not\leftarrow Y$

- Two variables that do not have a causal arrow between them can be made independent by conditioning on the parent set of one of the variables

PC algorithm (main ideas)

- $X \not\perp\!\!\!\perp Y \ (|\,\mathbf{C})$ implies there is some connection between the variables

  d-connection

  ➡ supported by **Markov Condition**

- $X \perp\!\!\!\perp Y \ (|\,\mathbf{C})$ implies (at least) that $X \not\rightarrow Y$ and $X \not\leftarrow Y$

  ➡ supported by **Faithfulness Condition**

- Two variables that do not have a causal arrow between them can be made independent by conditioning on the parent set of one of the variables

PC algorithm (main ideas)

- $X \not\perp\!\!\!\perp Y \ (|\,\mathbf{C})$ implies there is some connection between the variables

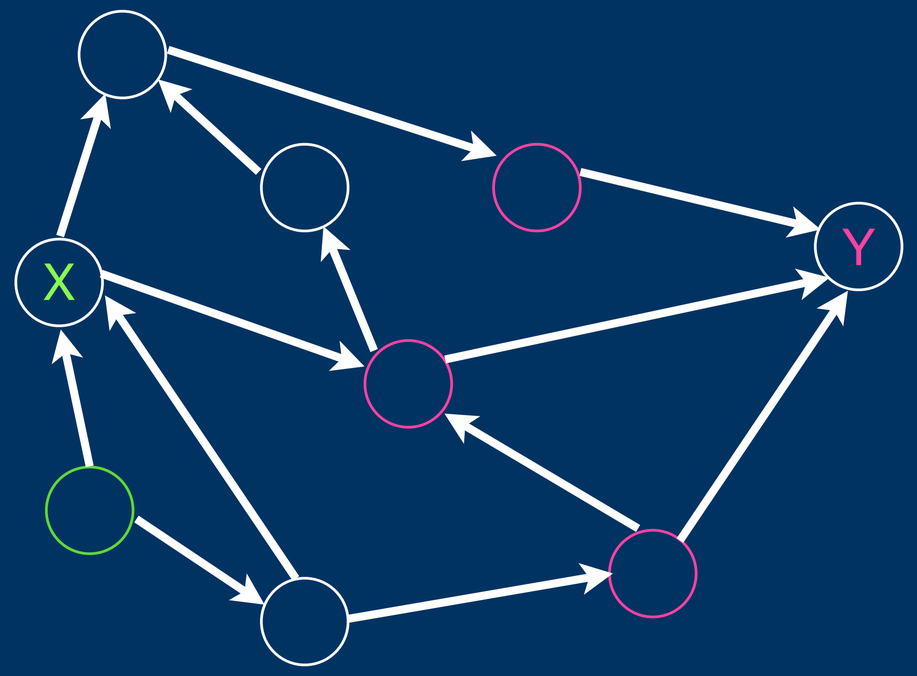
  d-connection

  ➡ supported by **Markov Condition**

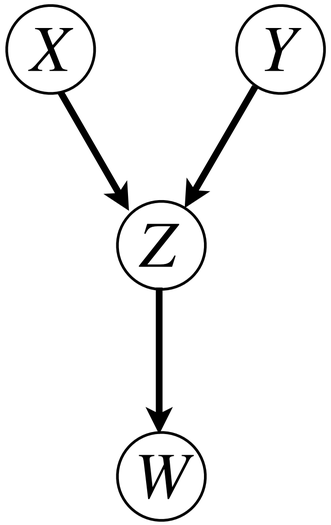
- $X \perp\!\!\!\perp Y \ (|\,\mathbf{C})$ implies (at least) that $X \not\rightarrow Y$ and $X \not\leftarrow Y$

  ➡ supported by **Faithfulness Condition**

- Two variables that do not have a causal arrow between them can be made independent by conditioning on the parent set of one of the variables



Running Example: Y-structure



true causal structure
(unobserved)

unconditional tests

$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Z$

$X \not\perp\!\!\!\perp W$

$Y \not\perp\!\!\!\perp Z$

$Y \not\perp\!\!\!\perp W$

$Z \not\perp\!\!\!\perp W$

conditioning set size 1

$X \not\perp\!\!\!\perp Y|Z$

$X \not\perp\!\!\!\perp Y|W$

$X \not\perp\!\!\!\perp Z|Y$

$X \not\perp\!\!\!\perp Z|W$

$X \not\perp\!\!\!\perp W|Y$

$X \perp\!\!\!\perp W|Z$

$Y \not\perp\!\!\!\perp Z|X$

$Y \not\perp\!\!\!\perp Z|W$

$Y \not\perp\!\!\!\perp W|X$

$Y \perp\!\!\!\perp W|Z$

$Z \not\perp\!\!\!\perp W|X$

$Z \not\perp\!\!\!\perp W|Y$

conditioning set size 2

$X \not\perp\!\!\!\perp Y|Z, W$

$X \not\perp\!\!\!\perp Z|Y, W$

$X \perp\!\!\!\perp W|Y, Z$

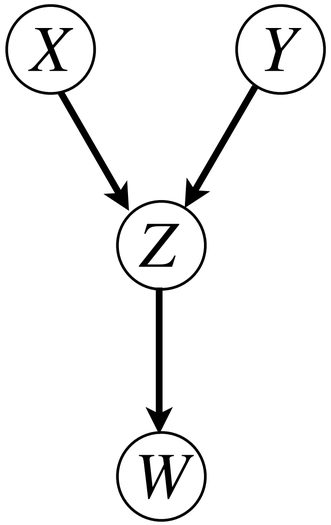
$Y \not\perp\!\!\!\perp Z|X, W$

$Y \perp\!\!\!\perp W|X, Z$

$Z \not\perp\!\!\!\perp W|X, Y$

Total number of possible conditional independence tests $\binom{n}{2} 2^{n-2} = 24$

Running Example: Y-structure



true causal structure
(unobserved)

unconditional tests

$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Z$

$X \not\perp\!\!\!\perp W$

$Y \not\perp\!\!\!\perp Z$

$Y \not\perp\!\!\!\perp W$

$Z \not\perp\!\!\!\perp W$

conditioning set size 1

$X \not\perp\!\!\!\perp Y|Z$

$X \not\perp\!\!\!\perp Y|W$

$X \not\perp\!\!\!\perp Z|Y$

$X \not\perp\!\!\!\perp Z|W$

$X \not\perp\!\!\!\perp W|Y$

$X \perp\!\!\!\perp W|Z$

$Y \not\perp\!\!\!\perp Z|X$

$Y \not\perp\!\!\!\perp Z|W$

$Y \not\perp\!\!\!\perp W|X$

$Y \perp\!\!\!\perp W|Z$

$Z \not\perp\!\!\!\perp W|X$

$Z \not\perp\!\!\!\perp W|Y$

conditioning set size 2

$X \not\perp\!\!\!\perp Y|Z,W$

$X \not\perp\!\!\!\perp Z|Y,W$

$X \perp\!\!\!\perp W|Y,Z$

$Y \not\perp\!\!\!\perp Z|X,W$

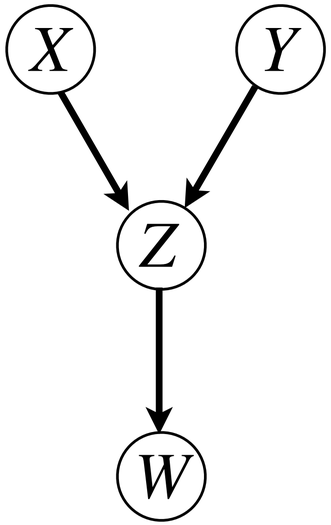
$Y \perp\!\!\!\perp W|X,Z$

$Z \not\perp\!\!\!\perp W|X,Y$

Note: Assume for now that we do not have test errors

Total number of possible conditional independence tests $\binom{n}{2}2^{n-2} = 24$

Running Example: Y-structure



true causal structure
(unobserved)

unconditional tests

$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Z$

$X \not\perp\!\!\!\perp W$

$Y \not\perp\!\!\!\perp Z$

$Y \not\perp\!\!\!\perp W$

$Z \not\perp\!\!\!\perp W$

conditioning set size 1

$X \not\perp\!\!\!\perp Y|Z$

$X \not\perp\!\!\!\perp Y|W$

$X \not\perp\!\!\!\perp Z|Y$

$X \not\perp\!\!\!\perp Z|W$

$X \not\perp\!\!\!\perp W|Y$

$X \perp\!\!\!\perp W|Z$

$Y \not\perp\!\!\!\perp Z|X$

$Y \not\perp\!\!\!\perp Z|W$

$Y \not\perp\!\!\!\perp W|X$

$Y \perp\!\!\!\perp W|Z$

$Z \not\perp\!\!\!\perp W|X$

$Z \not\perp\!\!\!\perp W|Y$

conditioning set size 2

$X \not\perp\!\!\!\perp Y|Z, W$

$X \not\perp\!\!\!\perp Z|Y, W$

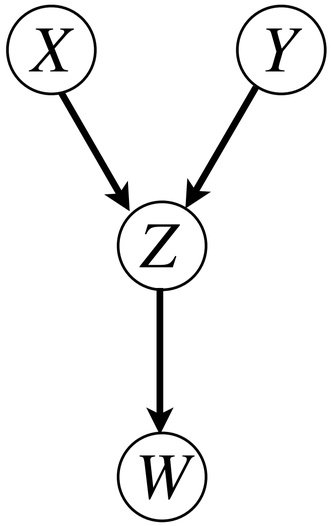
$X \perp\!\!\!\perp W|Y, Z$

$Y \not\perp\!\!\!\perp Z|X, W$

$Y \perp\!\!\!\perp W|X, Z$

$Z \not\perp\!\!\!\perp W|X, Y$

Note: Assume for now that we do not have test errors

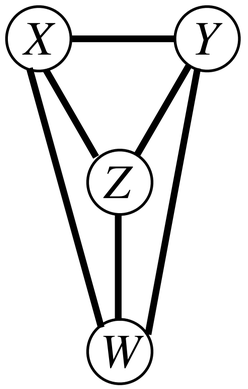


(singleton!) Markov equivalence class

Total number of possible conditional independence tests $\binom{n}{2} 2^{n-2} = 24$

PC algorithm

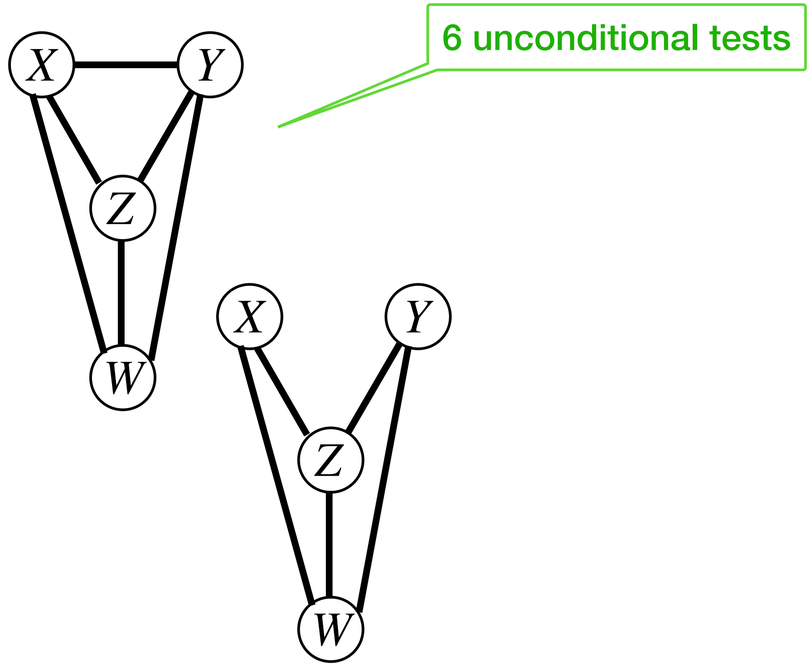
- Start with a complete undirected graph.
- Perform all unconditional tests.



6 unconditional tests

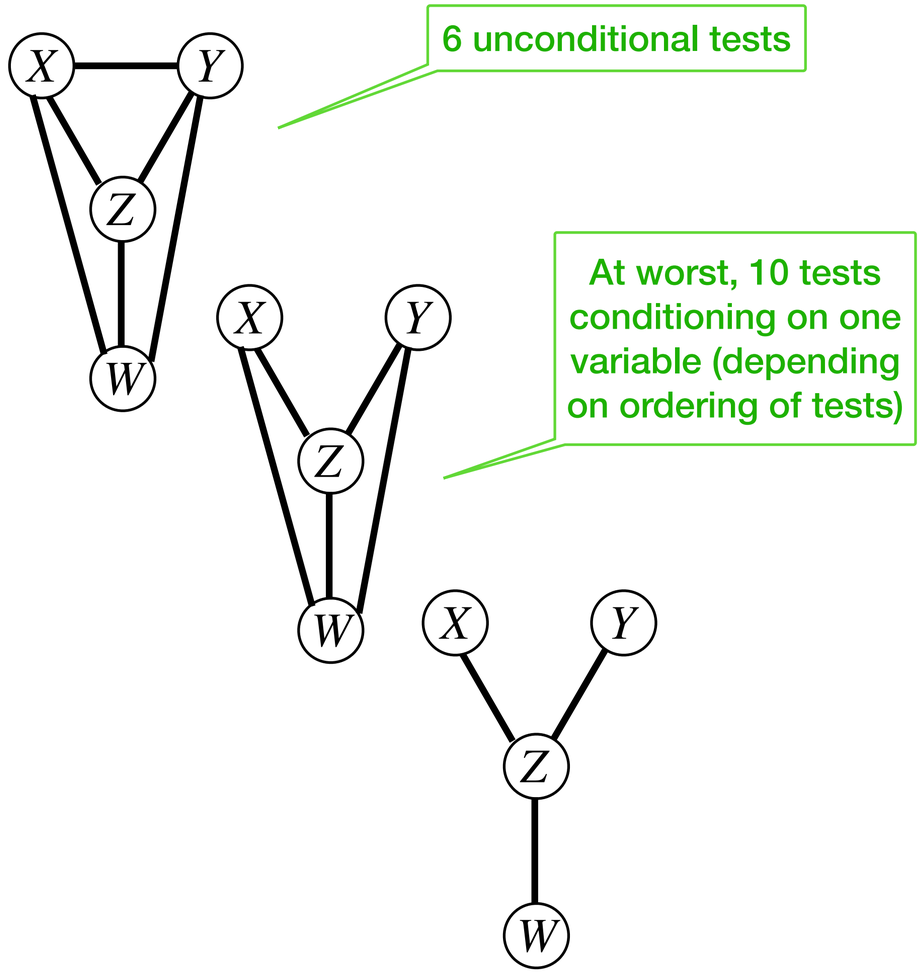
PC algorithm

- Start with a complete undirected graph.
- Perform all unconditional tests.
- Whenever two variables are independent, remove the edge between them.



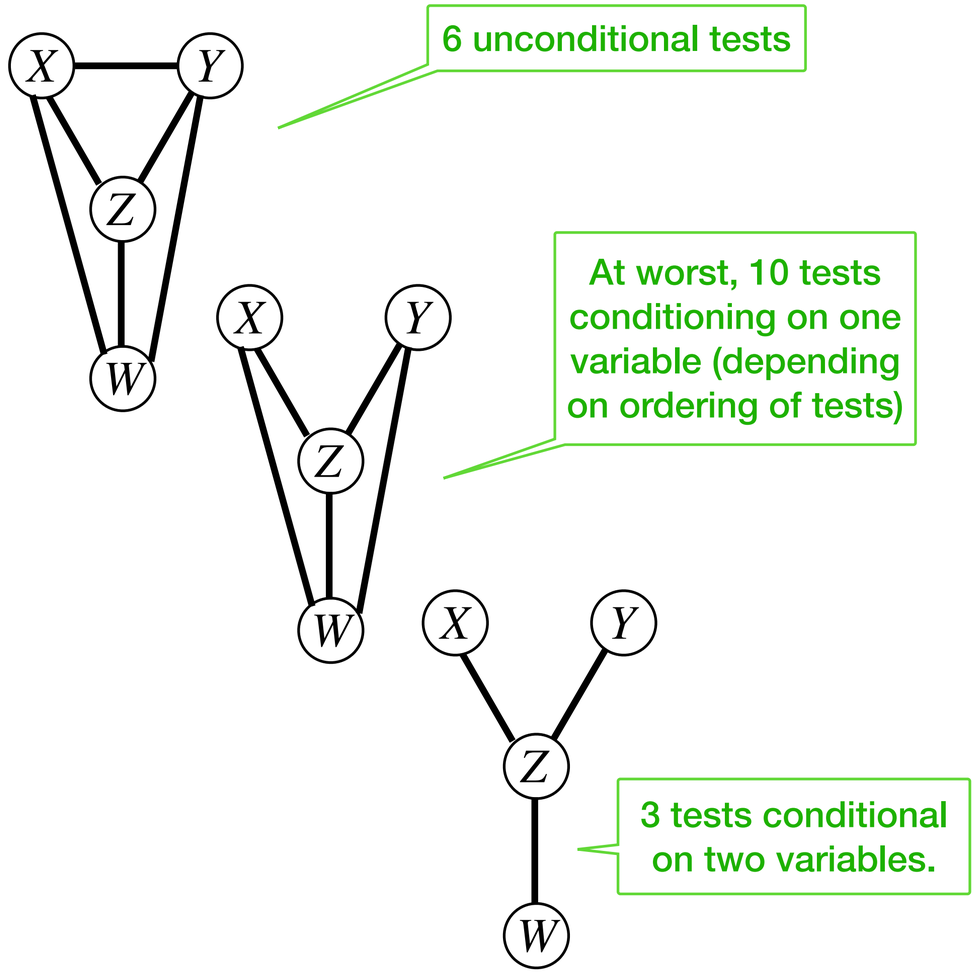
PC algorithm

- Start with a complete undirected graph.
- Perform all unconditional tests.
- Whenever two variables are independent, remove the edge between them.
- For each pair of variables A—B that are still adjacent, and each third variable C adjacent to A or B, test whether $A \perp\!\!\!\perp B \mid C$? If so, remove A—B.



PC algorithm

- Start with a complete undirected graph.
- Perform all unconditional tests.
- Whenever two variables are independent, remove the edge between them.
- For each pair of variables A—B that are still adjacent, and each third variable C adjacent to A or B, test whether $A \perp\!\!\!\perp B \mid C$? If so, remove A—B.
- For each pair of variables A—B that are still adjacent, and each pair of variables C_1, C_2, both adjacent to one of either A or B, test whether $A \perp\!\!\!\perp B \mid C_1, C_2$? If so, remove A—B.

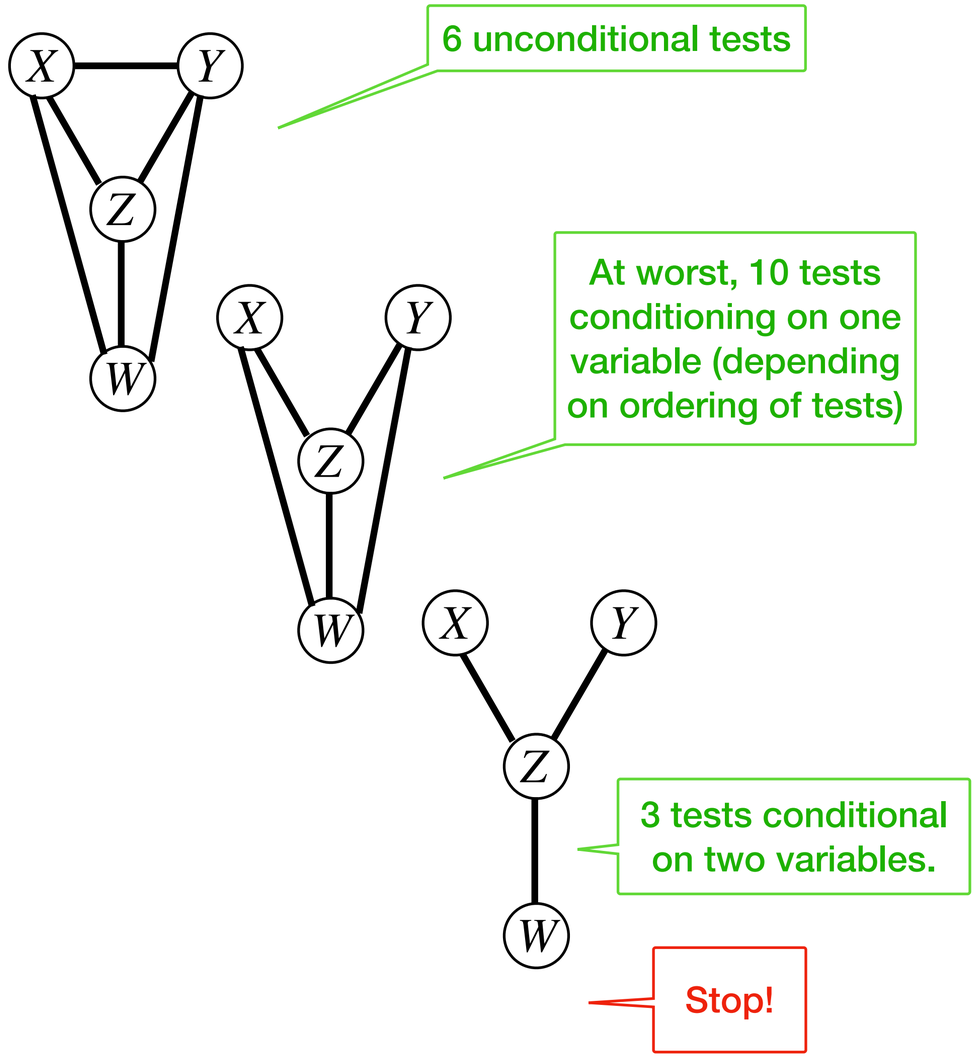


PC algorithm

- Start with a complete undirected graph.
- Perform all unconditional tests.
- Whenever two variables are independent, remove the edge between them.
- For each pair of variables A—B that are still adjacent, and each third variable C adjacent to A or B, test whether $A \perp\!\!\!\perp B \mid C$? If so, remove A—B.
- For each pair of variables A—B that are still adjacent, and each pair of variables C_1, C_2, both adjacent to one of either A or B, test whether $A \perp\!\!\!\perp B \mid C_1, C_2$? If so, remove A—B.
- …continue by iteratively increasing conditioning set size

➡ Stop when

- you run out of adjacent pairs to consider.
- when all conditioning sets of size n-2 have been checked.
- when the conditioning set size exceeds max(neigh(A), neigh(B)).



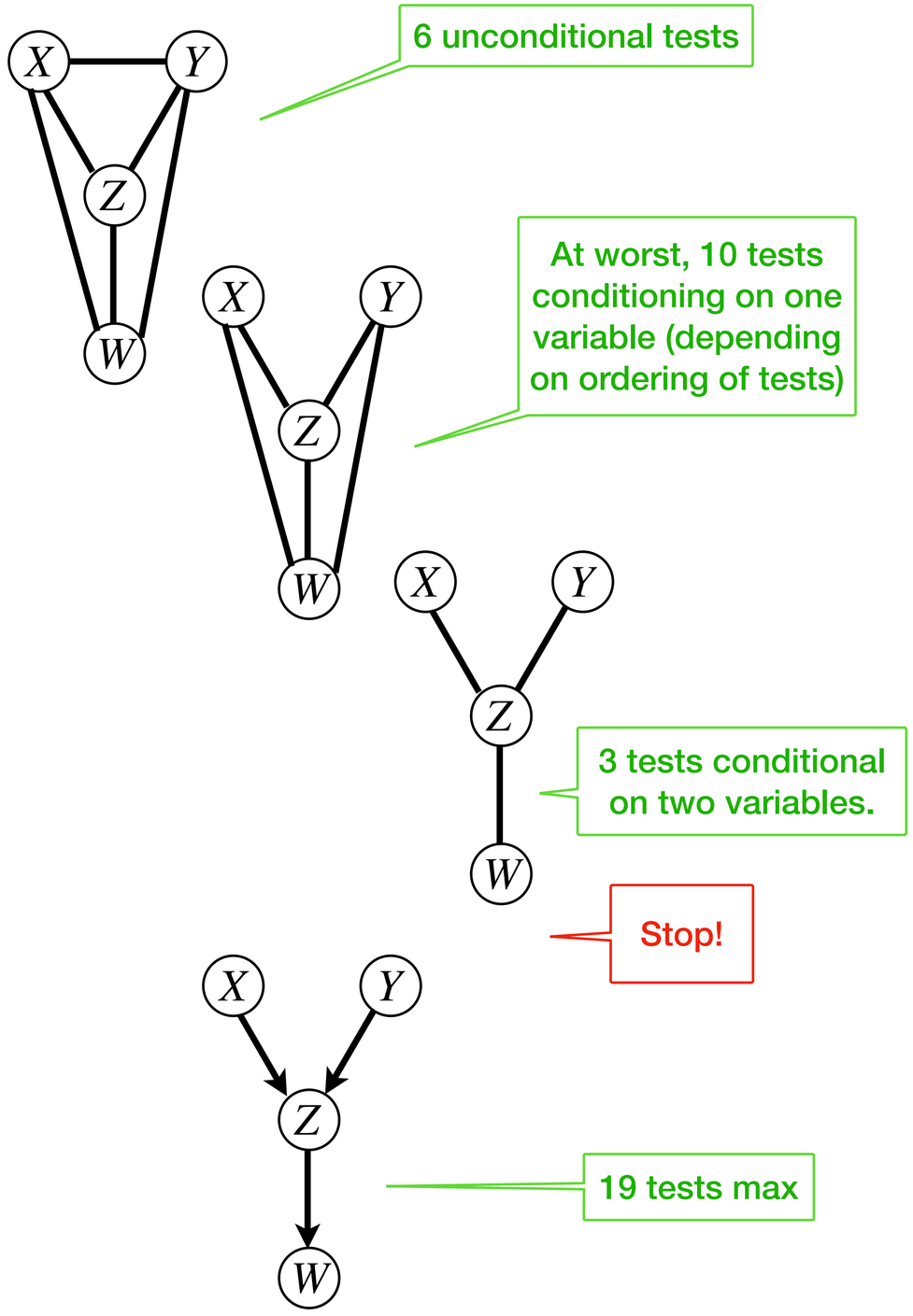
PC algorithm

- Start with a complete undirected graph.
- Perform all unconditional tests.
- Whenever two variables are independent, remove the edge between them.
- For each pair of variables A—B that are still adjacent, and each third variable C adjacent to A or B, test whether $A \perp\!\!\!\perp B \mid C$? If so, remove A—B.
- For each pair of variables A—B that are still adjacent, and each pair of variables C_1, C_2, both adjacent to one of either A or B, test whether $A \perp\!\!\!\perp B \mid C_1, C_2$? If so, remove A—B.
- …continue by iteratively increasing conditioning set size

➡ Stop when

- you run out of adjacent pairs to consider.
- when all conditioning sets of size n-2 have been checked.
- when the conditioning set size exceeds max(neigh(A), neigh(B)).

- Check triples of variables with two edges and orient colliders (and implied orientations, e.g. to avoid cycles)



PC algorithm

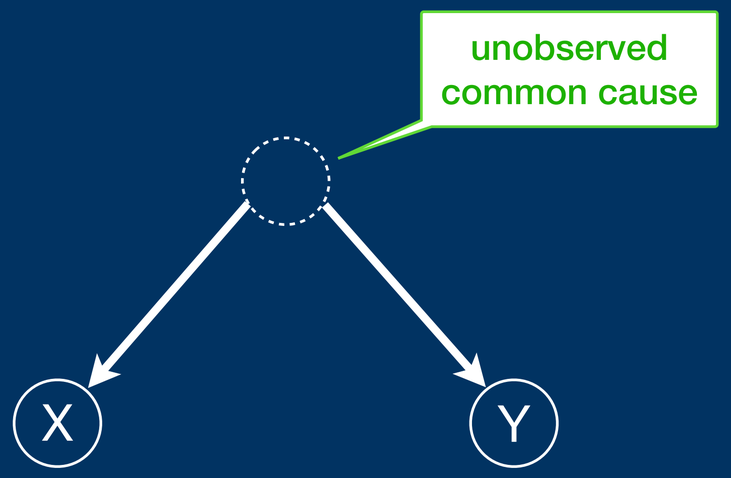
- Correct and complete: given the true (in)dependence relations among the variables, the PC algorithm identifies the Markov equivalence class of the true causal DAG, which is all that can be identified given the assumptions.

- Complexity:
  - Worst case is exponential in the number of variables.
  - Bounded by the largest degree k of a node in the true graph (exponential in k)
  - Use-cases I am familiar with scale to several 1000 variables, but only in sparse networks.

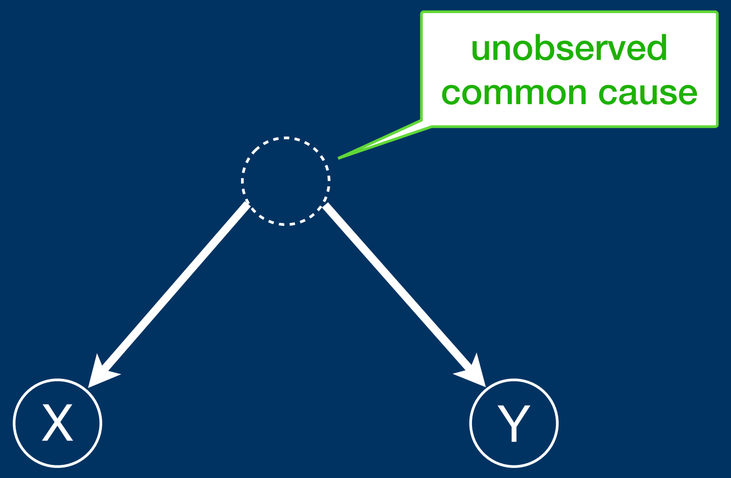
Assumptions for PC algorithm

- Markov Conditon: (conditional) probabilistic dependence implies (conditional) d-connection
- Faithfulness Conditon: (conditional) probabilistic independence implies (conditional) d-separation
- Causal Sufficiency: there are no unmeasured common causes of two or more measured variables.
- Acyclicity: the causal structure contains no cycles
- iid data: the data is independent and identically distributed

Causal INSufficiency

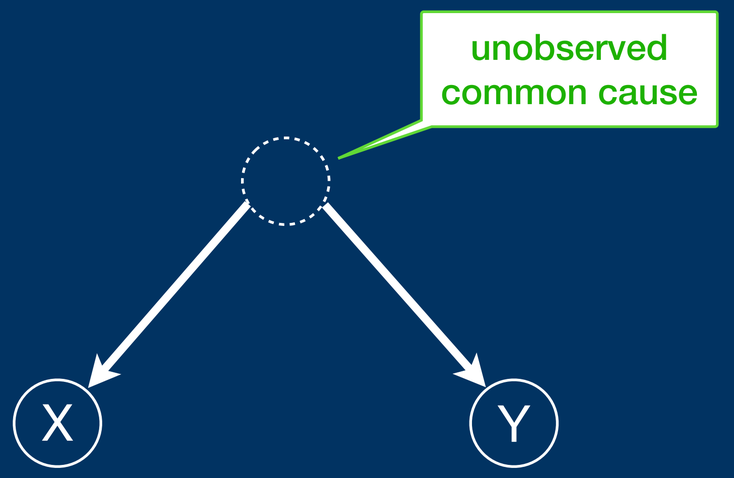


Causal INSufficiency



$$X \not\perp\!\!\!\perp Y$$

Causal INSufficiency

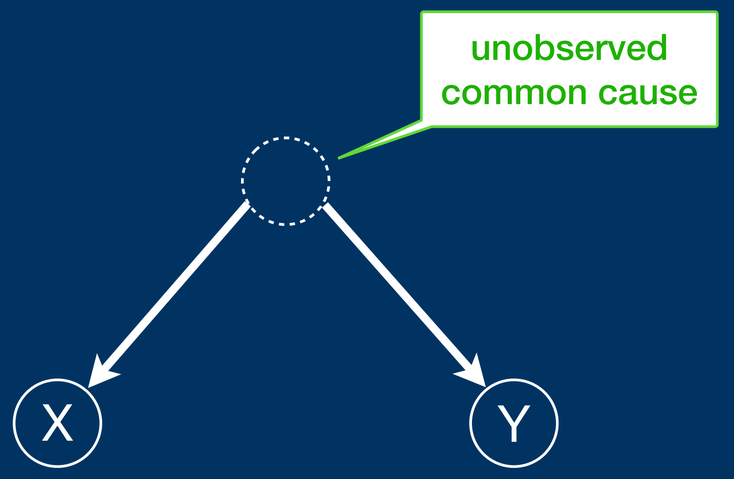


$X \not\perp\!\!\!\perp Y$

➡ PC algorithm would return X—Y

BUT: $X \not\rightarrow Y$ and $X \not\leftarrow Y$

Causal INSufficiency



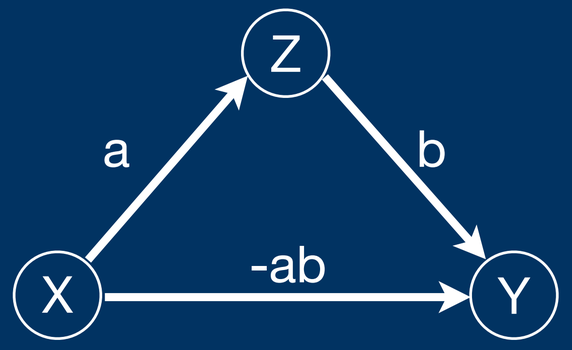
$X \not\perp\!\!\!\perp Y$

➡ PC algorithm would return X—Y

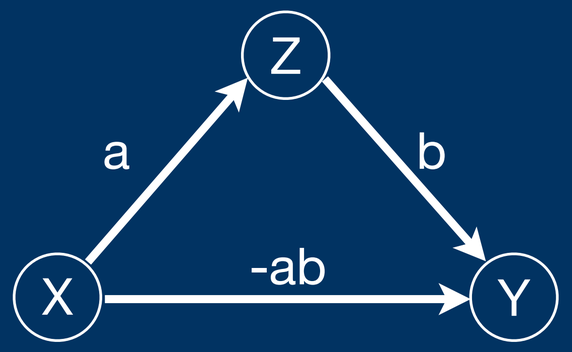
BUT: $X \not\rightarrow Y$ and $X \not\leftarrow Y$

Check out FCI algorithm or other discovery methods that allow for latent variables!

UNFaithfulness



UNFaithfulness

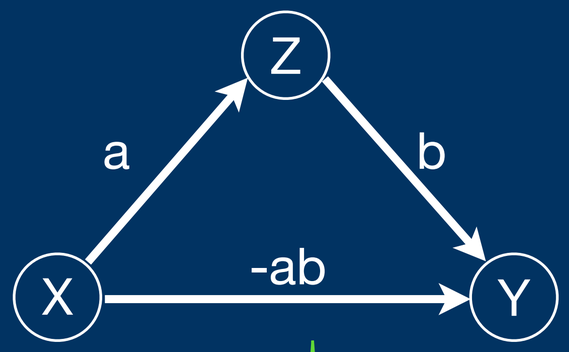


$$X \perp\!\!\!\perp Y$$

➡ PC algorithm would remove X—Y

BUT: $X \rightarrow Y$

UNFaithfulness



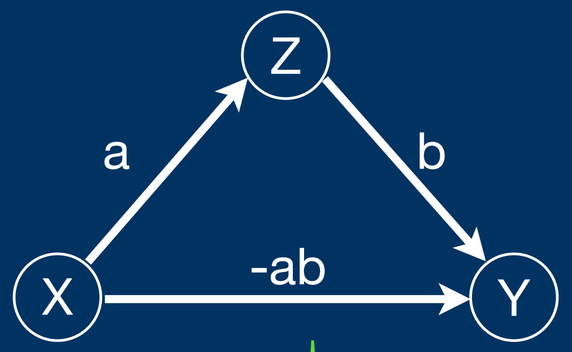
This requires perfect cancelation of the parameters.

$$X \perp\!\!\!\perp Y$$

➡ PC algorithm would remove X—Y

BUT: $X \rightarrow Y$

UNFaithfulness



$$X \perp\!\!\!\perp Y$$

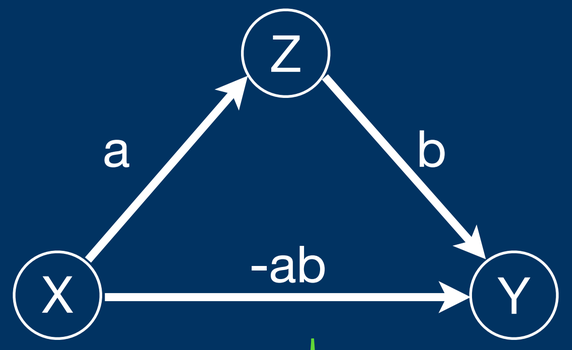
➡ PC algorithm would remove X—Y

BUT: $X \rightarrow Y$

This requires perfect cancelation of the parameters.

Well, but we are always subject to measurement error, so cancelation needn't be perfect.

UNFaithfulness



$$X \perp\!\!\!\perp Y$$

➡ PC algorithm would remove X—Y

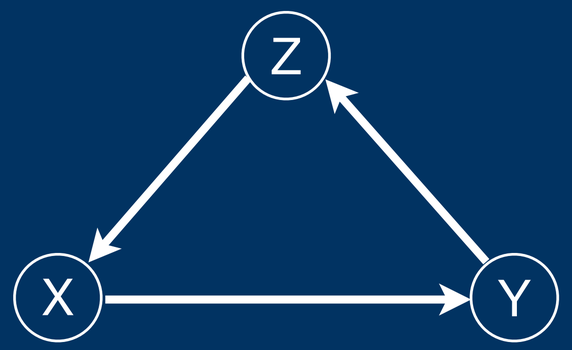
BUT: $X \rightarrow Y$

This requires perfect cancelation of the parameters.

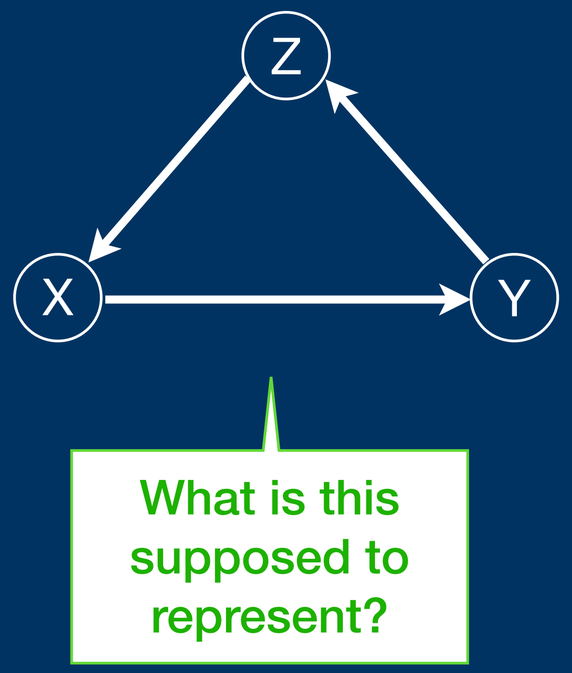
Well, but we are always subject to measurement error, so cancelation needn't be perfect.

Consider methods that weaken faithfulness (though at what price?).

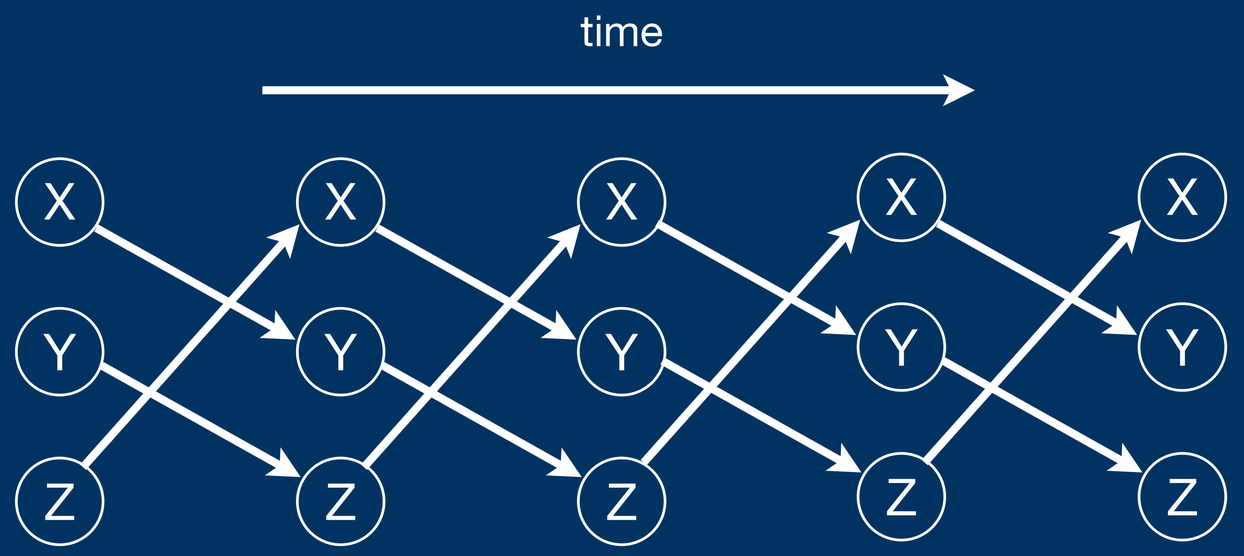
~~A~~Cyclicity

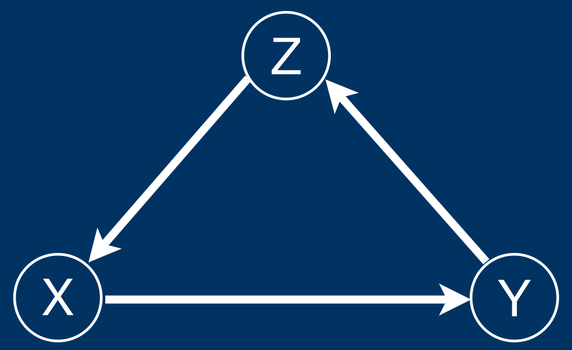


ACyclicity



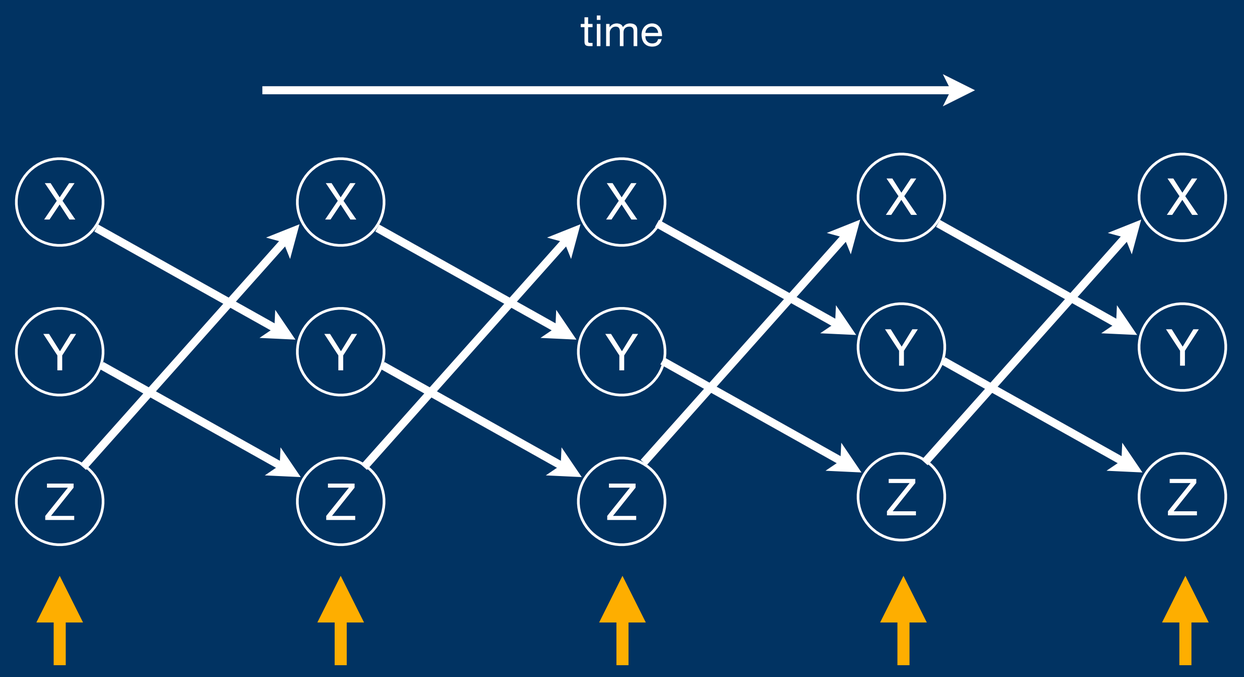
ACyclicity

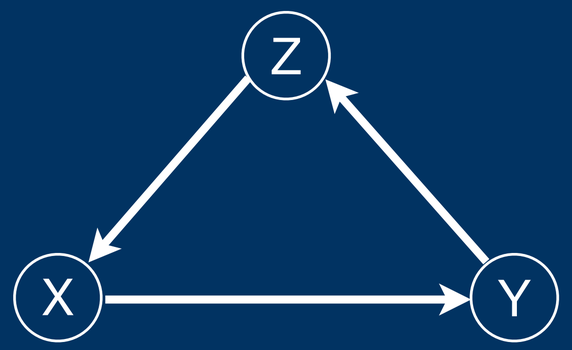




What is this supposed to represent?

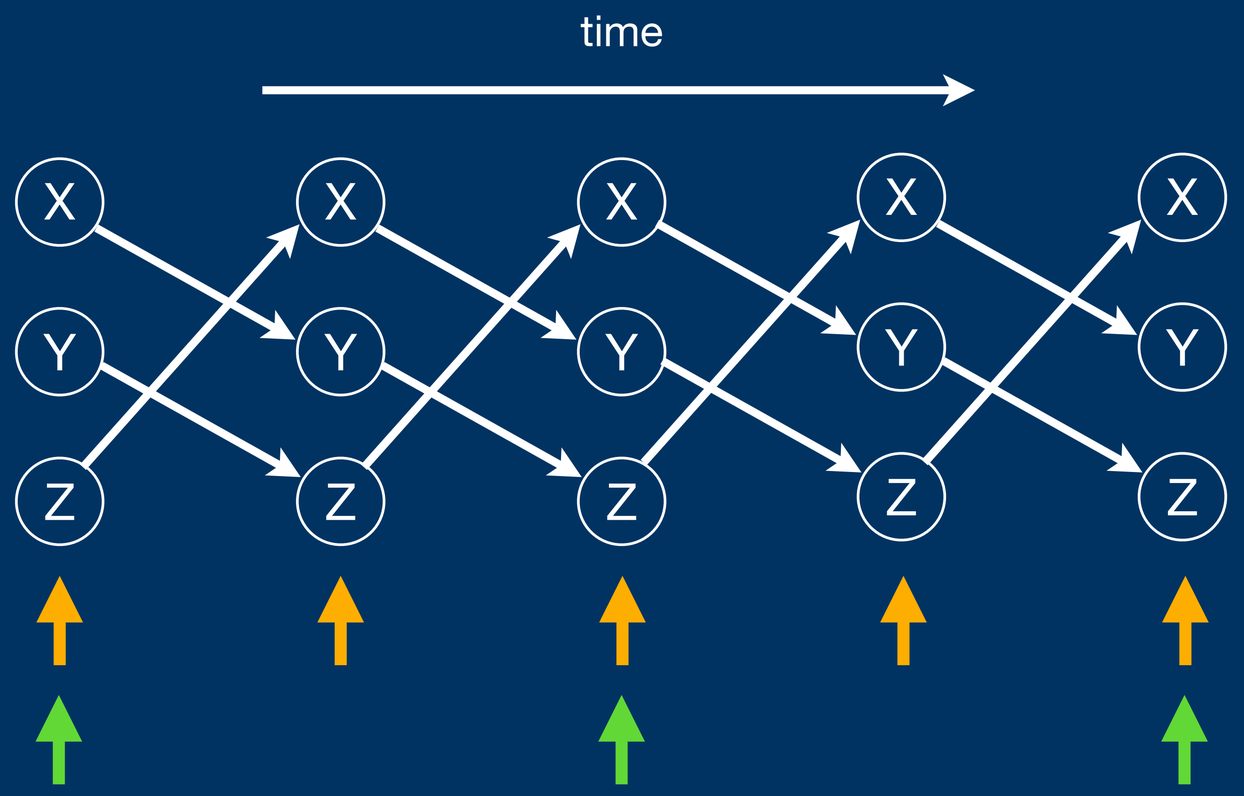
ACyclicity

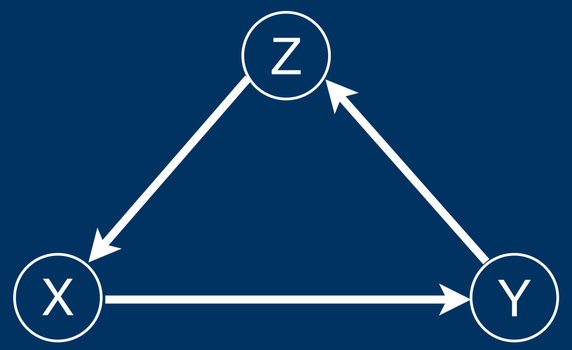




What is this supposed to represent?

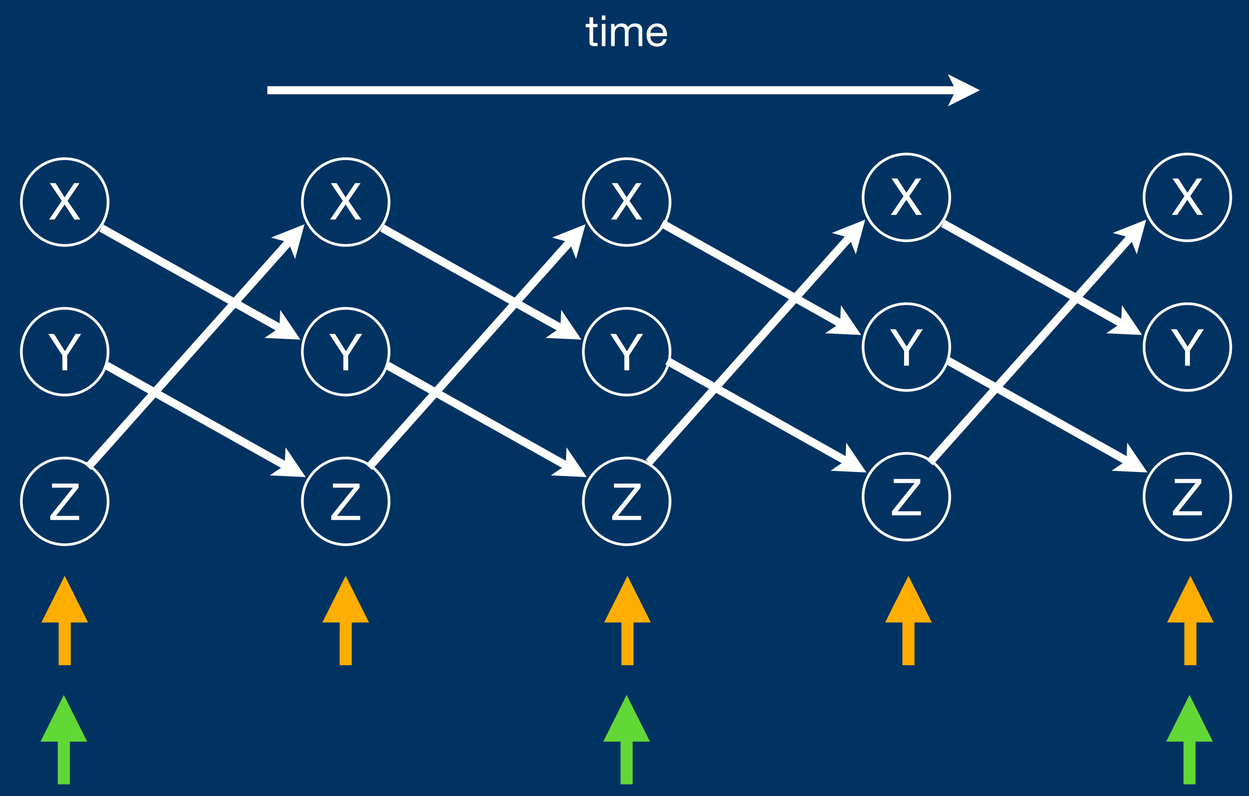
ACyclicity

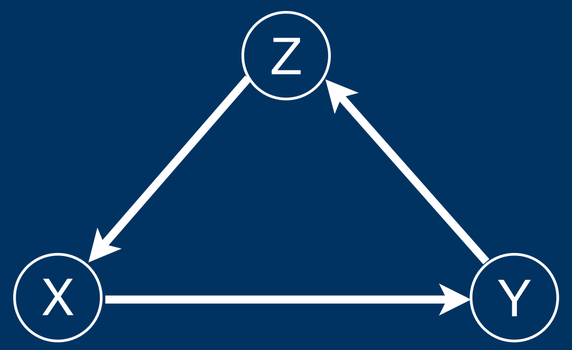




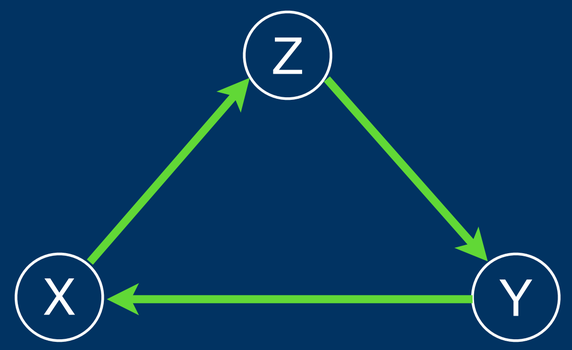
What is this supposed to represent?

ACyclicity



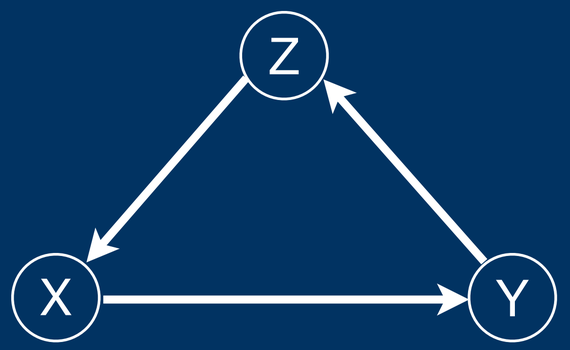


What is this supposed to represent?

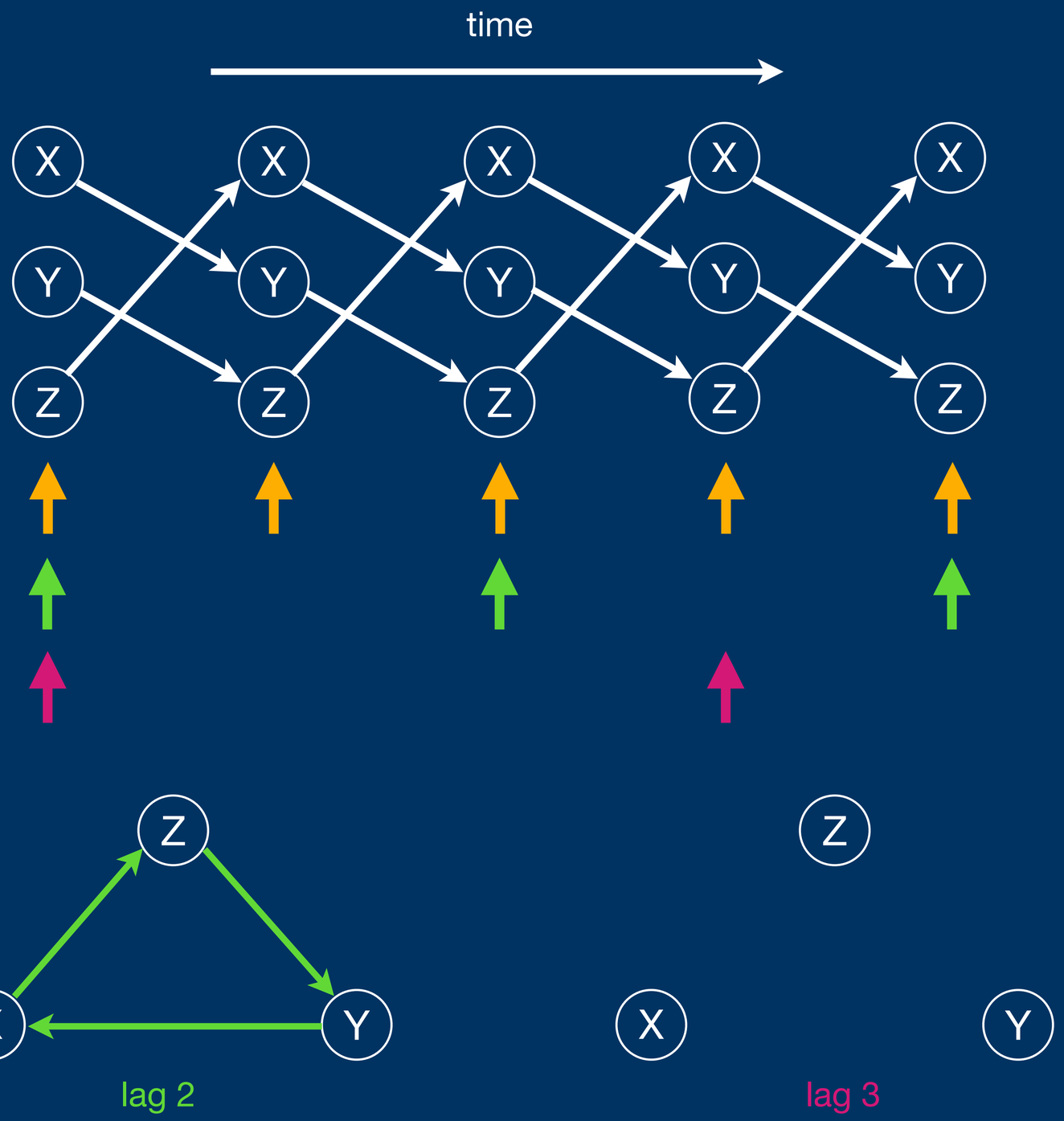


lag 2

ACyclicity



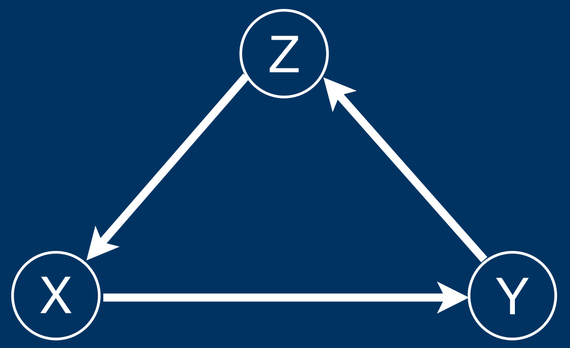
What is this supposed to represent?



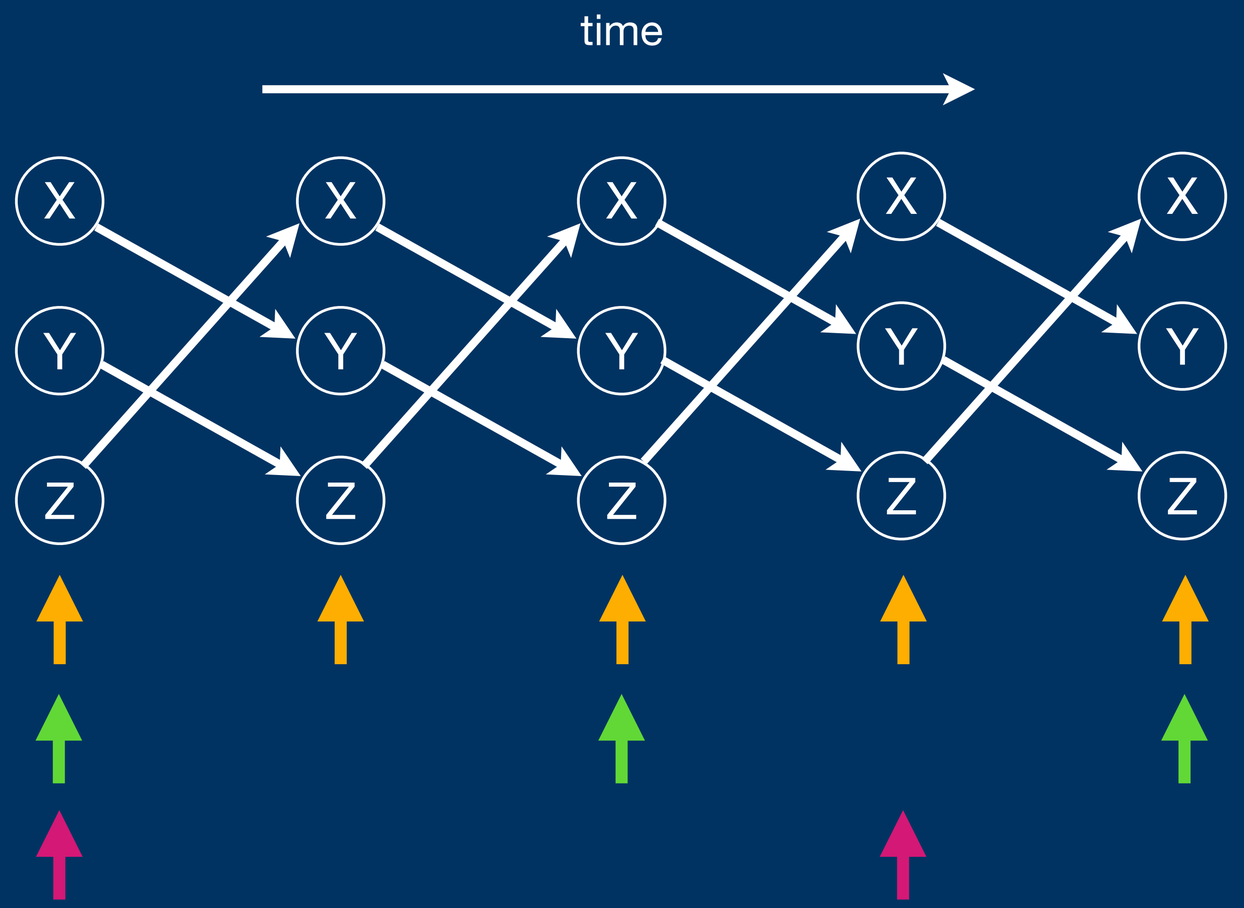
ACyclicity

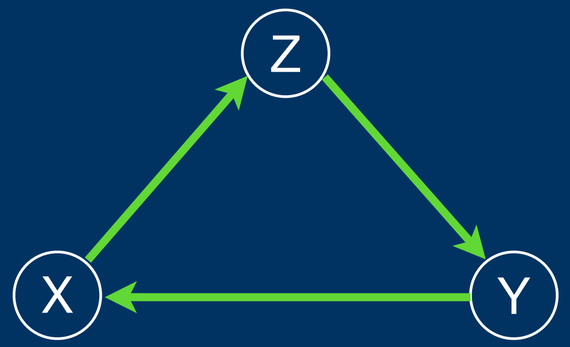
Issues that arise

- ➡ Sampling rate
- ➡ Measurements taken at equilibrium or not?



What is this supposed to represent?





lag 2

lag 3

Violation of the Markov Condition

$X \not\perp\!\!\!\perp Y$ but $X \not\rightarrow Y$ and $X \not\leftarrow Y$, nor is there a common cause.

Violation of the Markov Condition

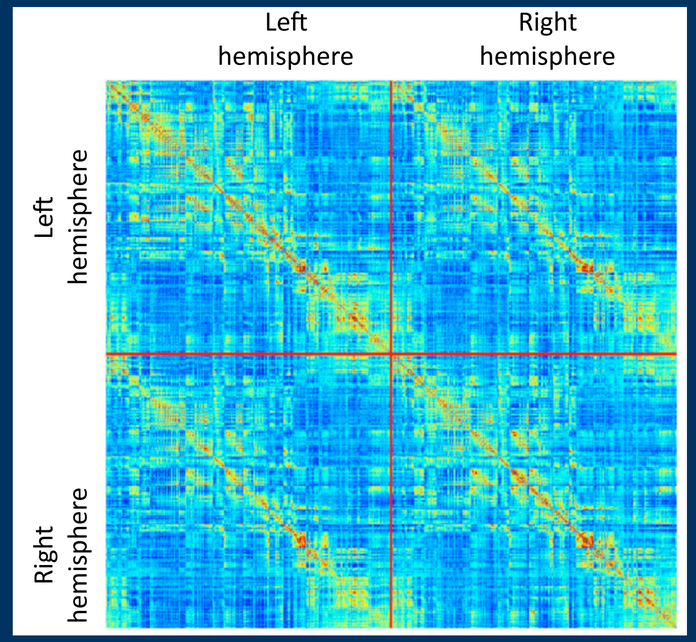
$X \not\perp\!\!\!\perp Y$ but $X \not\to Y$ and $X \not\leftarrow Y$, nor is there a common cause.

Outside of Quantum Mechanics I am not aware of plausible cases.

Plan for the data analysis

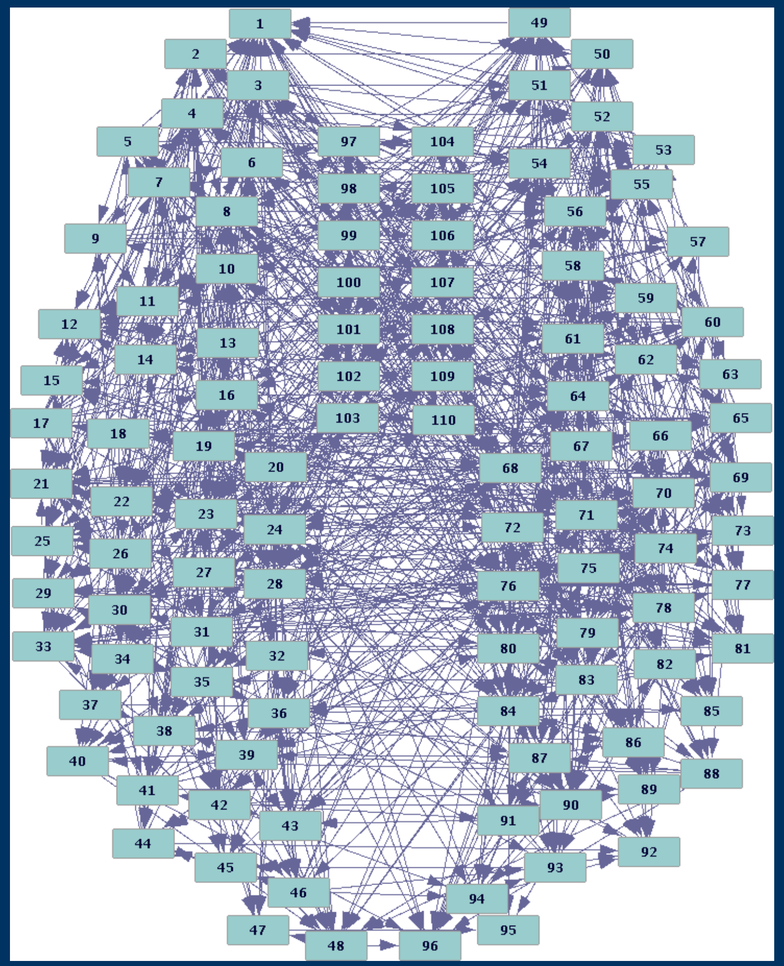
resting state brain

fMRI measure



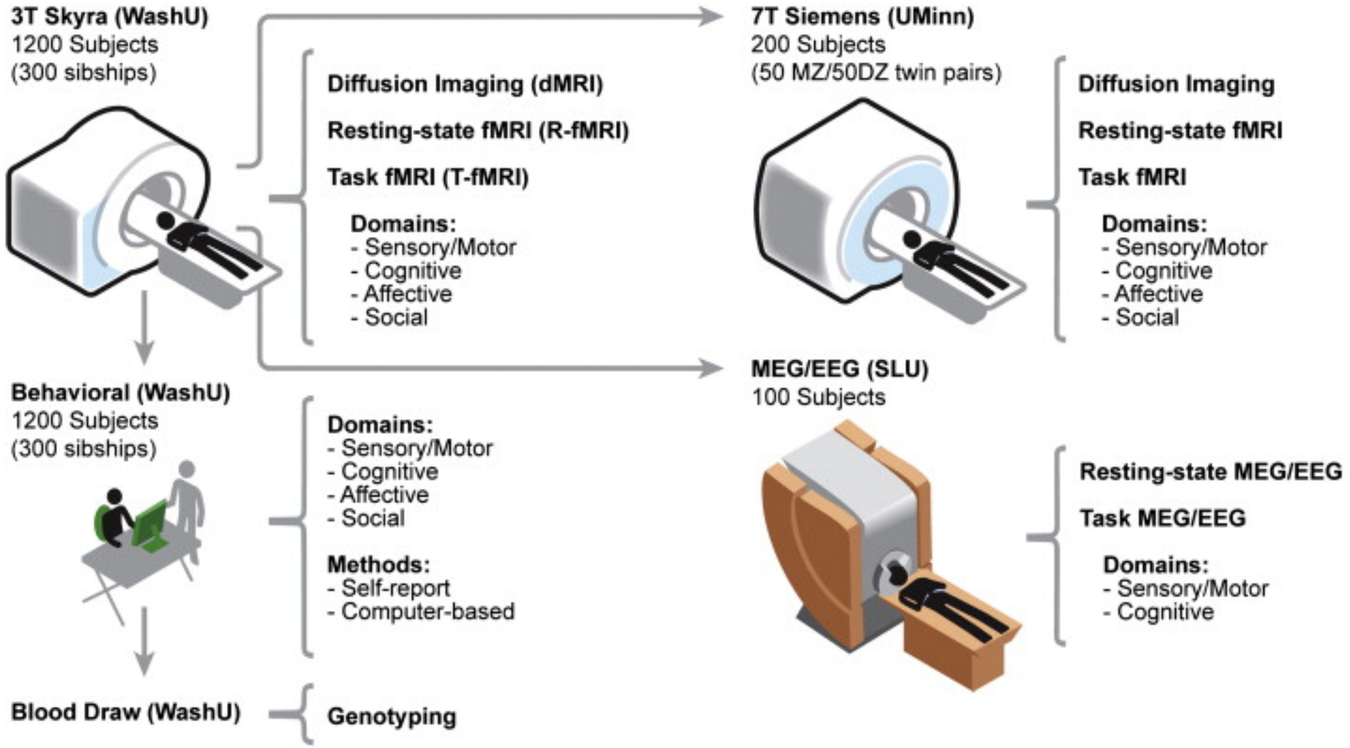
Pearson correlation matrix

Causal discovery



causal graph of brain regions

HUMAN Connectome PROJECT



Final HCP Study (Young Adult) Overall Totals:

1206 Subjects with Behavioral and Demographic Measures

1113 Subjects with 3T MRI (at least structural MRI)

45 Subjects with 3T Retest MRI and Behavioral Measures (46 subjects)

184 Subjects with 7T MRI*

95 Subjects with MEG*

1142 Subjects with Genotyping data (at least 1 gene chip)

HCP Subjects include:

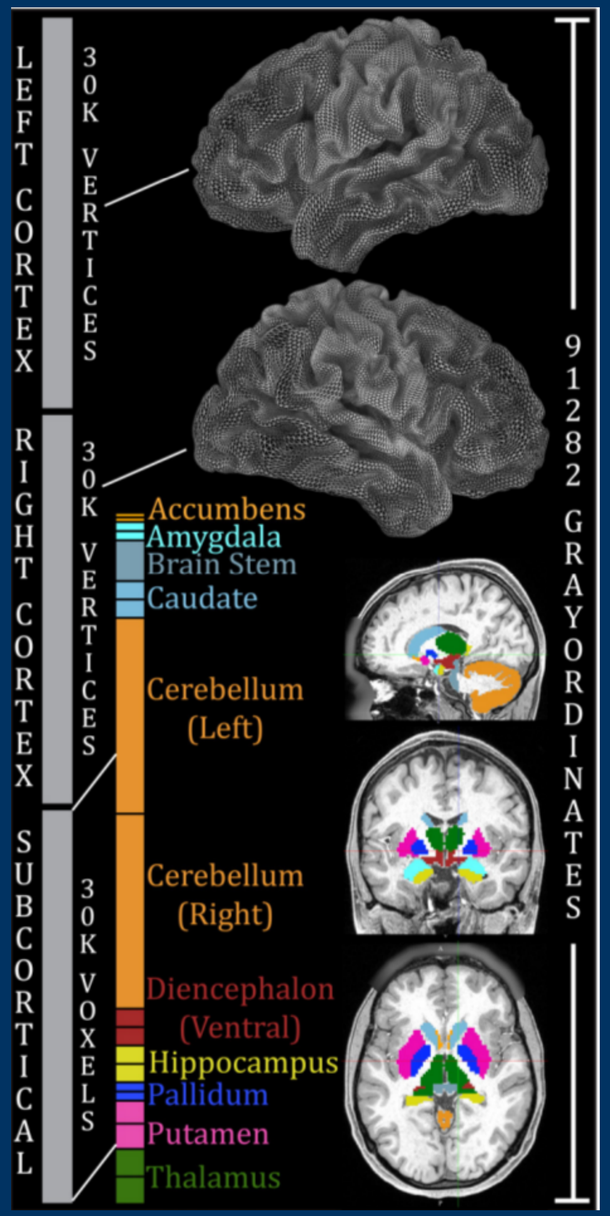
168 Monozygotic Twin pairs** (336 Subjects), including:

105 MZ twin pairs (210 Subjects) with complete 3T image data

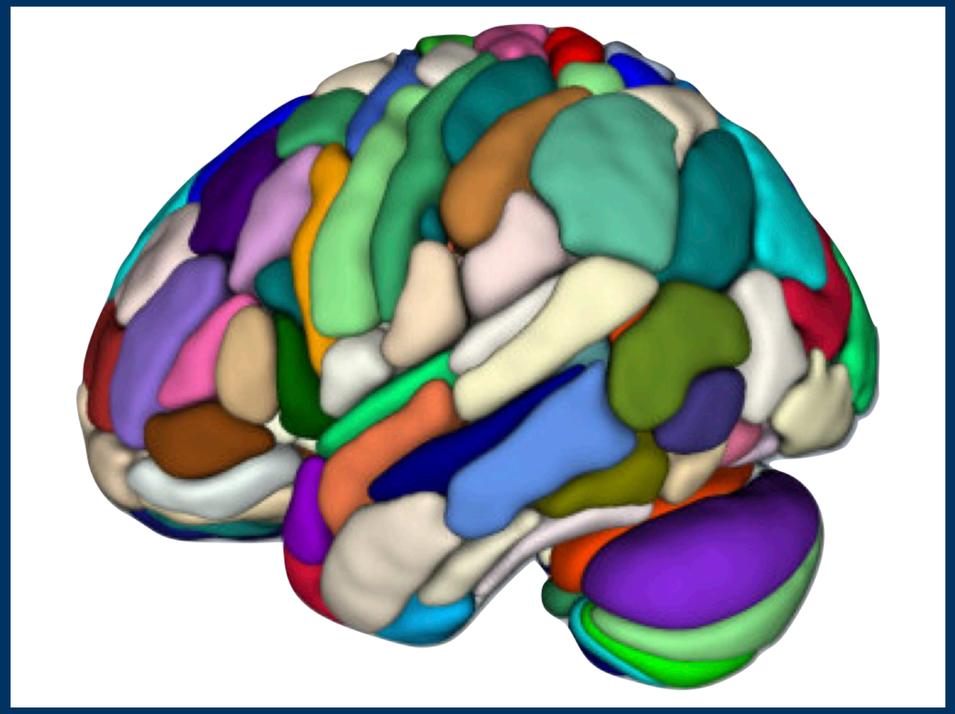
103 Dizygotic Twin Pairs** (224 Subjects), including

58 DZ twin pairs (116 Subjects) with complete 3T image data

HCP data after minimal pre-processing



anatomical parcelation

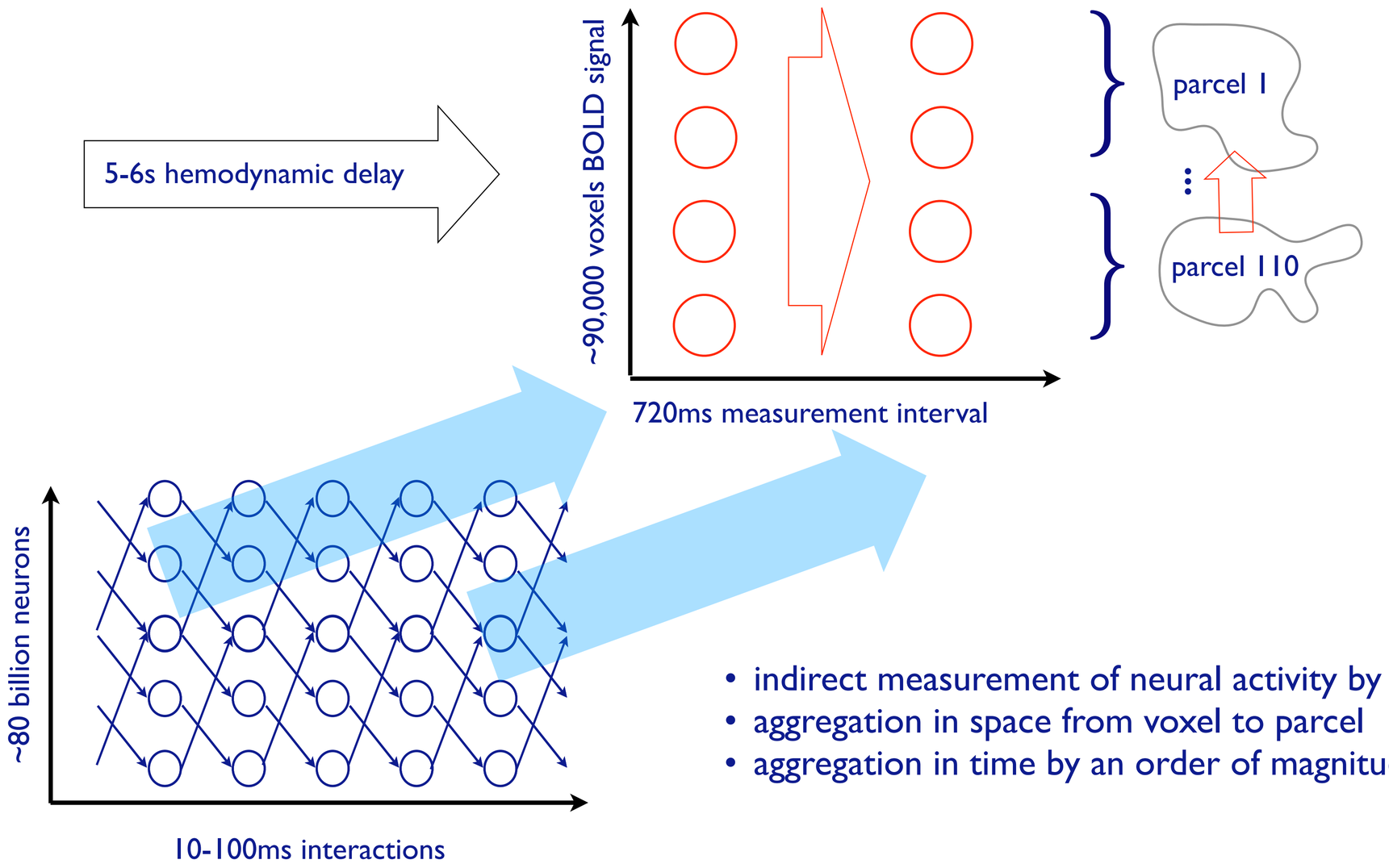


We used the Harvard Oxford Atlas.
(This image is from a different parcellation)

- 877 subjects x
- 4 runs x
- 110 parcels x
- 1200 time points

We sparsely sampled 11 disjoint datasets across the subjects.

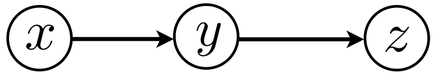
What are we measuring?



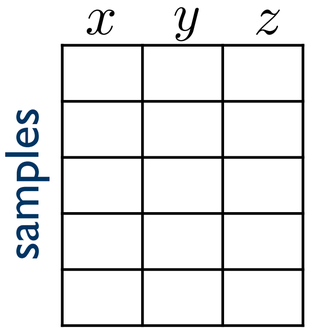
- indirect measurement of neural activity by BOLD signal
- aggregation in space from voxel to parcel
- aggregation in time by an order of magnitude

Causal discovery from observational data

truth
(unknown)

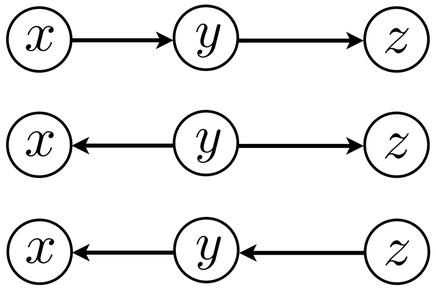


assumptions
- causal Markov
- causal faithfulness
- causal sufficiency
- acyclicity
- …



data sample

inference algorithm



equivalence class of causal structures

Thank you!

Iman Wahle

Frederick Eberhardt (fde@caltech.edu)