Causality

A brief introduction to causal inference

Neill D F Campbell

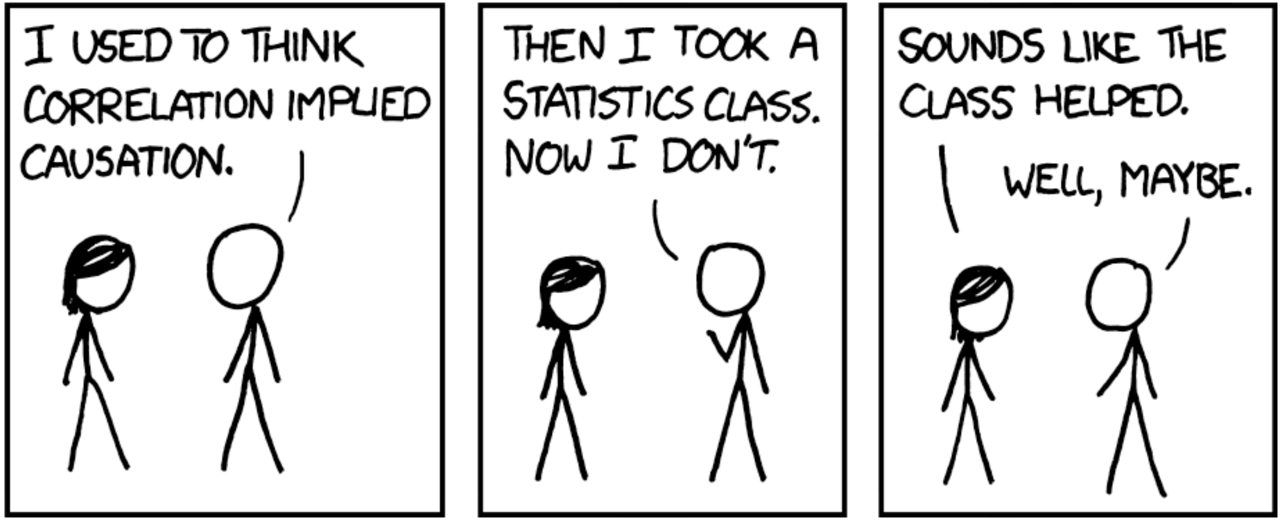
5th Aug 2026

University College London

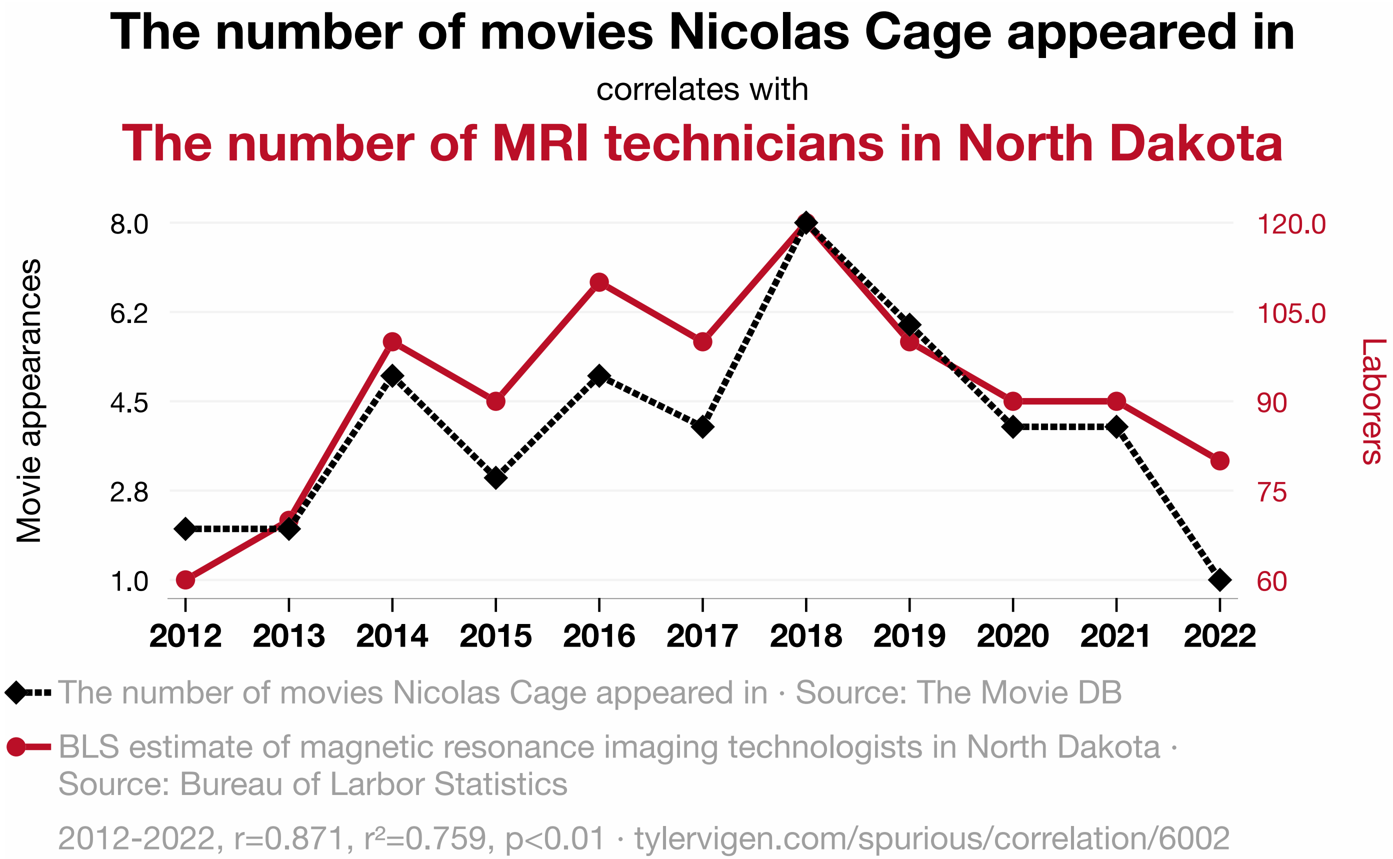
Outline

Correlation vs Causation

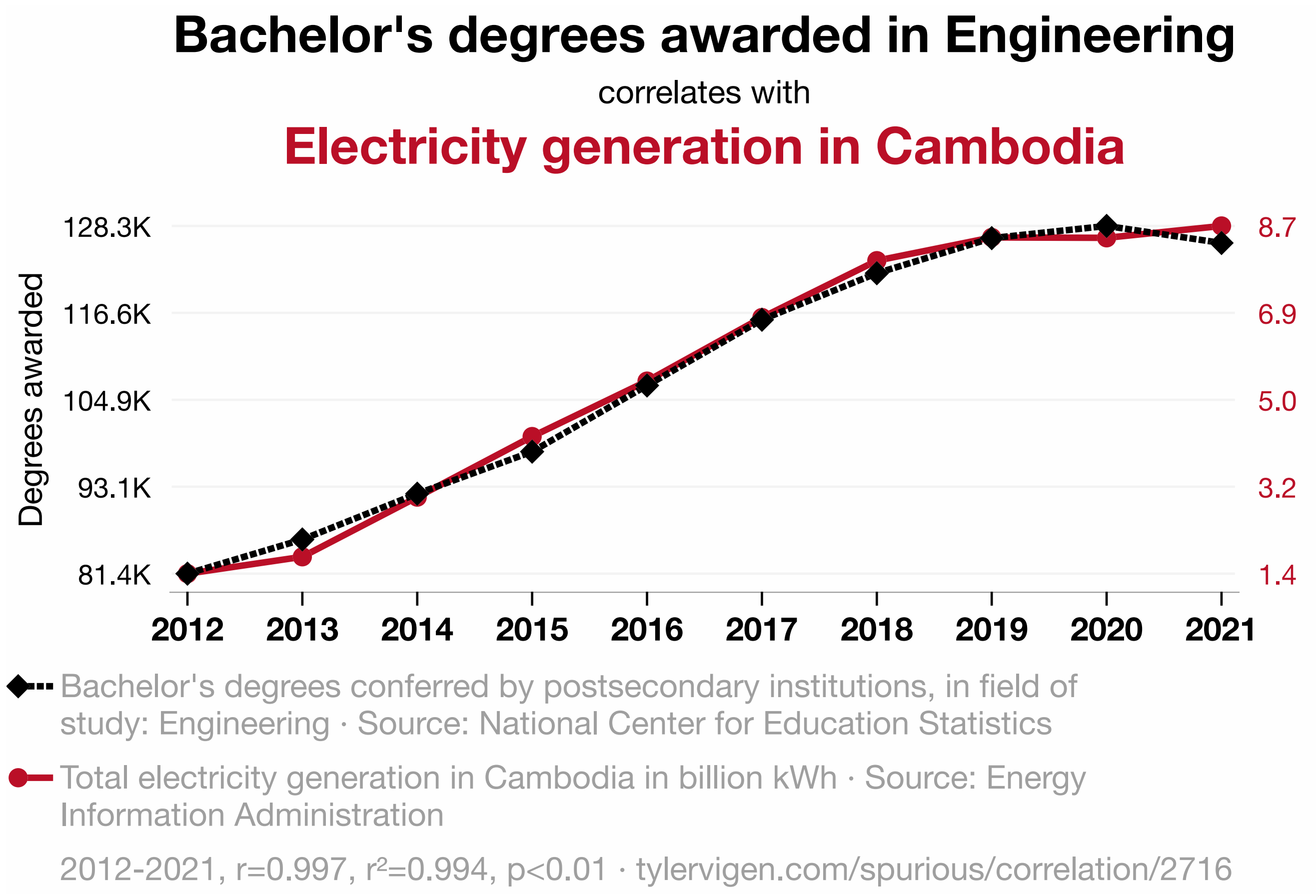
Correlation is not Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Danger, Batman...



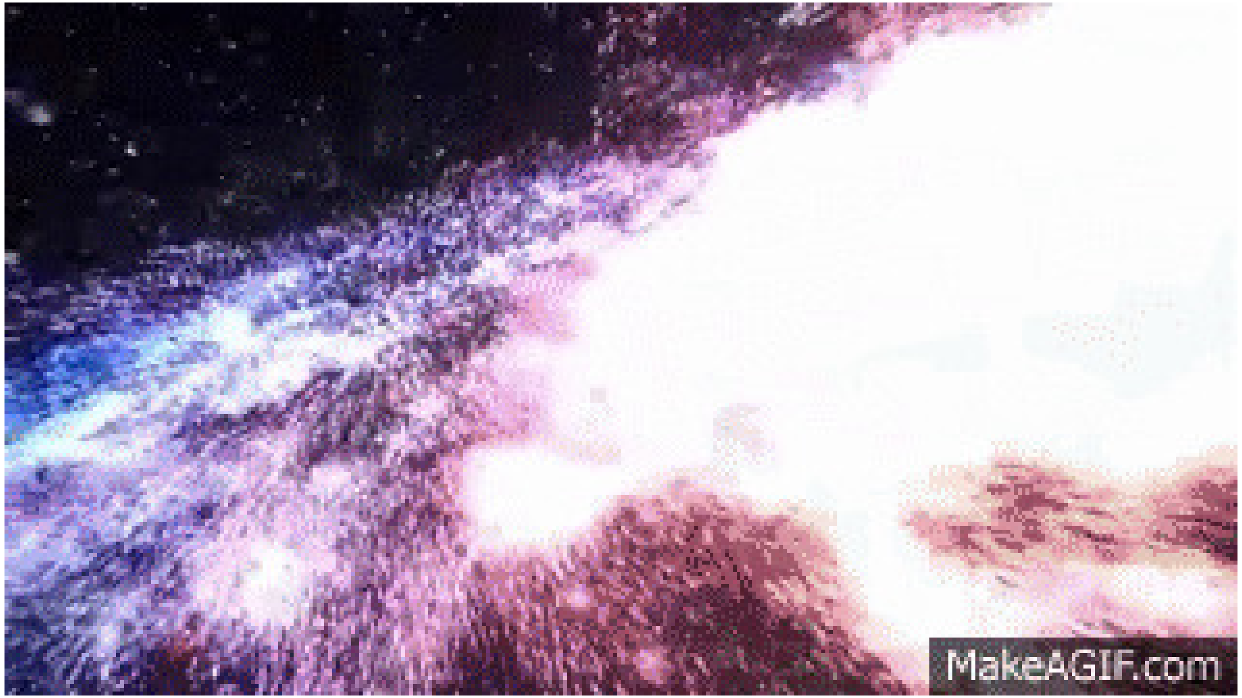
https://tylervigen.com/spurious-correlations



https://tylervigen.com/spurious-correlations

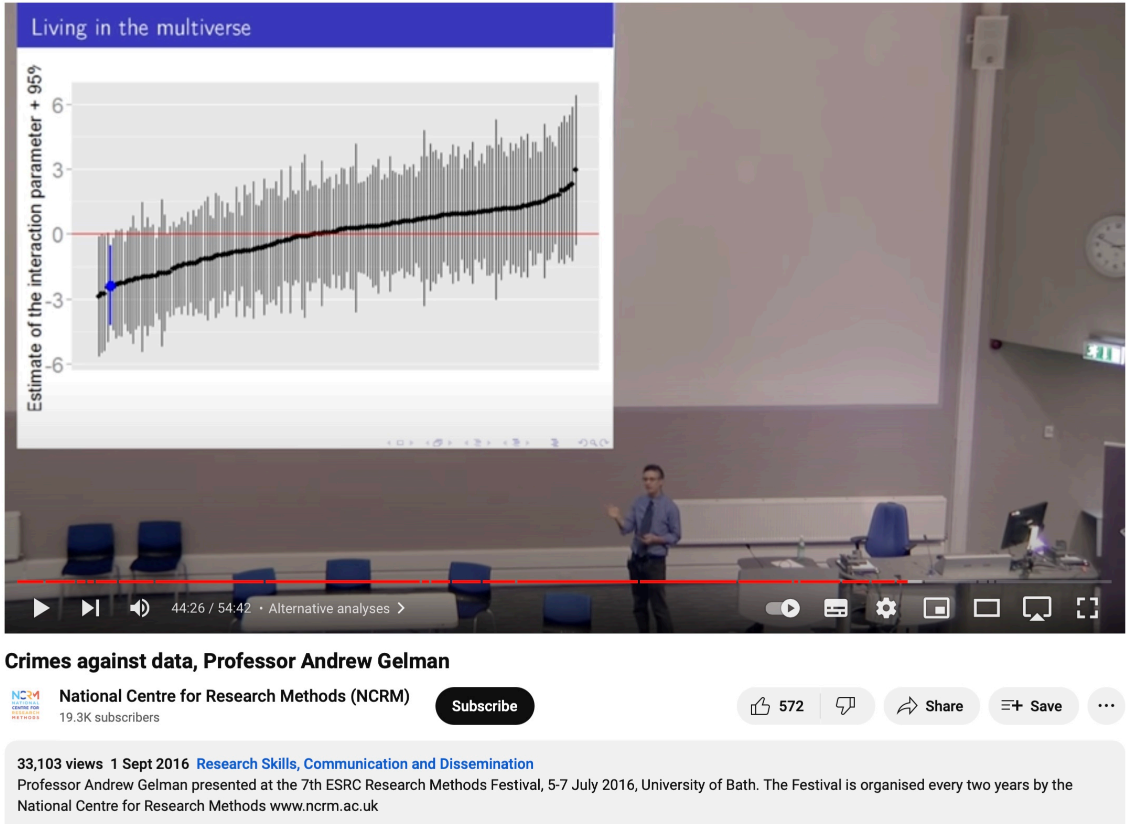
- Do we need causation?
- Is science not just correlation?

Importance simultaneously undervalued and overestimated?



Objectivity...

Objectivity...



Andrew Gelman: “Crimes against Data”

Correlation vs Causation: What's the Difference?

- Well, we all know what the difference is...
  - e.g. atmospheric pressure and the barometer needle reading

Correlation vs Causation: What's the Difference?

- Well, we all know what the difference is...
  - e.g. atmospheric pressure and the barometer needle reading

> **Formal definitions are tricky, but:**
>
> An object is the cause of another ...
>
> "*if the first object had not been, the second never had existed*"
>
> David Hume, *Enquiry Concerning Human Understanding*, 1748

- Well, we all know what the difference is...
  - e.g. atmospheric pressure and the barometer needle reading

> **Formal definitions are tricky, but:**
>
> An object is the cause of another ...
>
> *"if the first object had not been, the second never had existed"*
>
> David Hume, *Enquiry Concerning Human Understanding*, 1748

- Introduces the idea of a **counterfactual**

Counterfactuals

- Real world:

$$\textsc{Action} \longrightarrow \textsc{Outcome}$$

Counterfactuals

- Real world:

$$\text{ACTION} \longrightarrow \text{OUTCOME}$$

- Hypothetical world:

$$\text{COUNTERFACTUAL ACTION} \longrightarrow \text{COUNTERFACTUAL OUTCOME}$$

Counterfactuals

- Real world:

$$\textsc{Action} \longrightarrow \textsc{Outcome}$$

- Hypothetical world:

$$\textsc{Counterfactual Action} \longrightarrow \textsc{Counterfactual Outcome}$$

- Difference in outcome $=$ effect of the action!

Counterfactuals

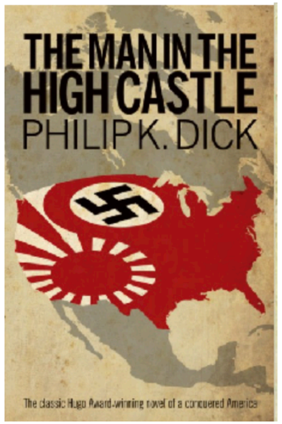
- Real world:

$$\textsc{Action} \longrightarrow \textsc{Outcome}$$

- Hypothetical world:

$$\textsc{Counterfactual Action} \longrightarrow \textsc{Counterfactual Outcome}$$

- Difference in outcome = effect of the action!



So We Are All Done?

Problems with counterfactuals...

Problems with counterfactuals...

- We never get to observe the counterfactual :-(

So We Are All Done?

Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!

So We Are All Done?

Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*

So We Are All Done?

Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties / objections...

Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties / objections...
- Can we approximate the counterfactual?

So We Are All Done?

Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties / objections...
- Can we approximate the counterfactual?
  - Lots of the time in science → the Randomised Control Trial (RCT)!

So We Are All Done?

Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties / objections...
- Can we approximate the counterfactual?
  - Lots of the time in science $\rightarrow$ the Randomised Control Trial (RCT)!
- Exciting question: what if we can't do an RCT?

So We Are All Done?

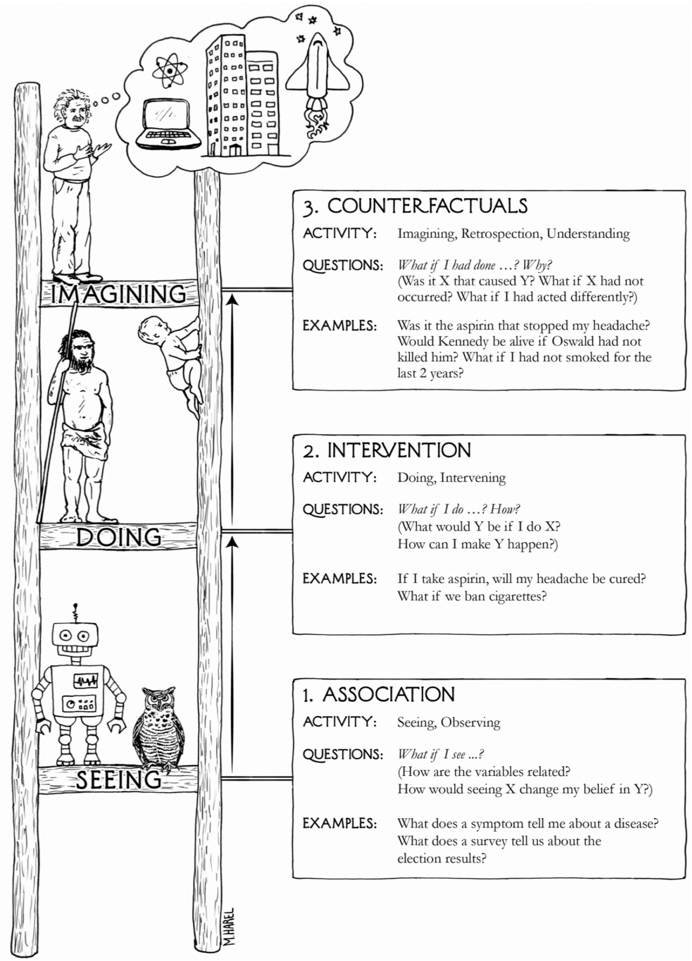
Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties / objections...
- Can we approximate the counterfactual?
  - Lots of the time in science $\rightarrow$ the Randomised Control Trial (RCT)!
- Exciting question: what if we can't do an RCT?
  - Can we use ML to estimate the counterfactuals?

So We Are All Done?

Problems with counterfactuals...

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time, inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties / objections...
- Can we approximate the counterfactual?
  - Lots of the time in science → the Randomised Control Trial (RCT)!
- Exciting question: what if we can't do an RCT?
  - Can we use ML to estimate the counterfactuals? **Possibly!**



[Pearl and Mackenzie 2017]

- Pearl proposes a “ladder of causation”
- Only the bottom rung can be answered by data alone
- Higher levels:
  - Require additional assumptions
  - Cannot be estimated from data alone...
  - Need a model as well!

Bayesian Networks

What Is a Bayesian Network?

- A Bayesian network is a directed acyclic graph (DAG):
  - Nodes represent random variables
  - Edges represent direct probabilistic dependence

What Is a Bayesian Network?

- A Bayesian network is a directed acyclic graph (DAG):
  - Nodes represent random variables
  - Edges represent direct probabilistic dependence
- By the chain rule, any joint distribution factorises, for some ordering of the variables, as

$$p(X_1, \dots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid X_1, \dots, X_{i-1}\right)$$

What Is a Bayesian Network?

- A Bayesian network is a directed acyclic graph (DAG):
  - Nodes represent random variables
  - Edges represent direct probabilistic dependence
- By the chain rule, any joint distribution factorises, for some ordering of the variables, as

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid X_1, \ldots, X_{i-1}\right)$$

- A Bayesian network assumes each variable is conditionally independent of its non-descendants given its parents $\mathrm{pa}(X_i)$ in the graph, so this simplifies to

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid \mathrm{pa}(X_i)\right)$$

What Is a Bayesian Network?

- A Bayesian network is a directed acyclic graph (DAG):
  - Nodes represent random variables
  - Edges represent direct probabilistic dependence
- By the chain rule, any joint distribution factorises, for some ordering of the variables, as

$$p(X_1, \dots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid X_1, \dots, X_{i-1}\right)$$

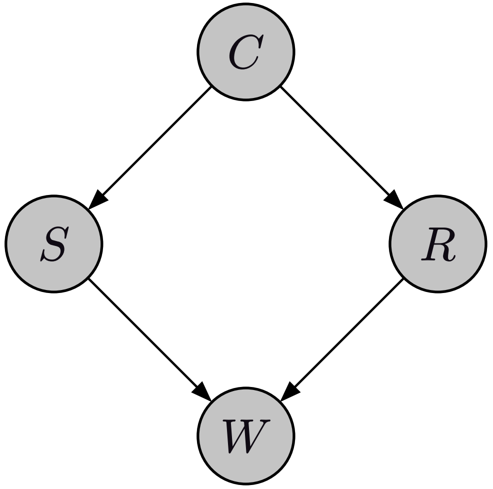
- A Bayesian network assumes each variable is conditionally independent of its non-descendants given its parents $\mathrm{pa}(X_i)$ in the graph, so this simplifies to

$$p(X_1, \dots, X_n) = \prod_{i=1}^{n} p\left(X_i \mid \mathrm{pa}(X_i)\right)$$

- So a Bayesian network encodes a factorisation of the joint distribution.

The absence of edges is the important part!

Consider four binary variables: whether it is cloudy (C), whether the sprinkler is on (S), whether it rains (R), and whether the grass is wet (W).



The graph encodes the factorisation

$$p(C,S,R,W) = p(C)\,p\big(S \mid C\big)\,p\big(R \mid C\big)\,p\big(W \mid S,R\big)$$

Instead of the $2^4 - 1 = 15$ free parameters of a generic joint distribution over four binary variables, this factorisation needs only $1+2+2+4=9$.

- The $C \to S$ edge could just as well have been drawn $S \to C$

From Factorisation to Causation?

- The $C \to S$ edge could just as well have been drawn $S \to C$

$$C \leftarrow S \qquad\qquad C \rightarrow S$$

- Bayes' rule lets us reverse any edge:

$$p\left(C \mid S\right) = \frac{p\left(S \mid C\right) p(C)}{p(S)}$$

From Factorisation to Causation?

- The $C \to S$ edge could just as well have been drawn $S \to C$

- Bayes' rule lets us reverse any edge:

$$p\left(C \mid S\right) = \frac{p\left(S \mid C\right) p(C)}{p(S)}$$

- The factorisation (and the conditional independencies it implies) is unchanged either way

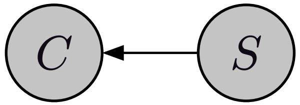
- The $C \to S$ edge could just as well have been drawn $S \to C$

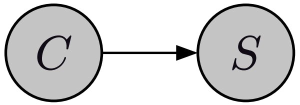
- Bayes' rule lets us reverse any edge:

$$p\left(C \mid S\right) = \frac{p\left(S \mid C\right) p(C)}{p(S)}$$

- The factorisation (and the conditional independencies it implies) is unchanged either way
- So a Bayesian network alone only tells us how a joint distribution factorises

- The $C \to S$ edge could just as well have been drawn $S \to C$



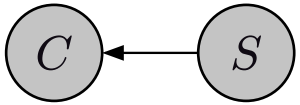


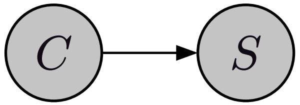
- Bayes' rule lets us reverse any edge:

$$p\left(C \mid S\right) = \frac{p\left(S \mid C\right) p(C)}{p(S)}$$

- The factorisation (and the conditional independencies it implies) is unchanged either way
- So a Bayesian network alone only tells us how a joint distribution factorises
- It does not tell us what causes what!

- The $C \rightarrow S$ edge could just as well have been drawn $S \rightarrow C$





- Bayes' rule lets us reverse any edge:

$$p\left(C \mid S\right) = \frac{p\left(S \mid C\right) p(C)}{p(S)}$$

- The factorisation (and the conditional independencies it implies) is unchanged either way
- So a Bayesian network alone only tells us how a joint distribution factorises
- **It does not tell us what causes what!**

So what extra ingredients turns a DAG into a causal graph?

Illustration: Disease Treatments

Two disease treatments (surgical / non-surgical) for kidney stones

	Positive Outcome
Treatment A	$273/350 = 78\%$
Treatment B	$289/350 = 83\%$

Illustration: Disease Treatments

Two disease treatments (surgical / non-surgical) for kidney stones

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

Illustration: Disease Treatments

Two disease treatments (surgical / non-surgical) for kidney stones

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

What's going on?

Two disease treatments (surgical / non-surgical) for kidney stones

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

What's going on?

Not a fair RCT: uneven allocation of patients

Two disease treatments (surgical / non-surgical) for kidney stones

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

What's going on?

Not a fair RCT: uneven allocation of patients

The stone size S is a confounder: An RCT would remove the confounder:

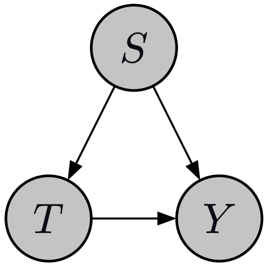
Two disease treatments (surgical / non-surgical) for kidney stones

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

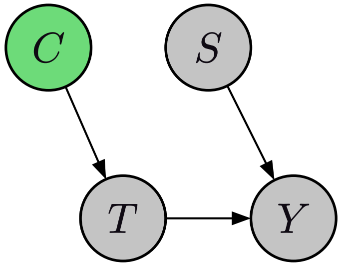
What's going on?

Not a fair RCT: uneven allocation of patients

The stone size S is a confounder:

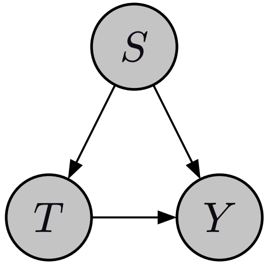


An RCT would remove the confounder:

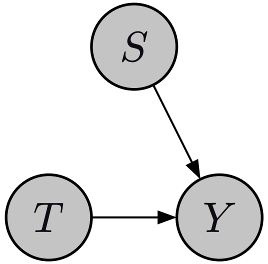


	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

The stone size S is a confounder:



An RCT would remove the confounder:



$p(Y \mid T)$ means "probability of Y given T by observation alone"

$$p(Y \mid T) = \sum_s p(Y \mid S, T)\, p(T \mid S)\, p(S)/p(T)$$

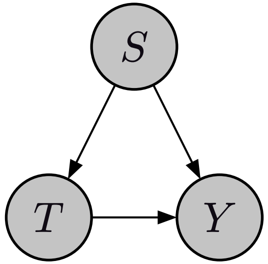
Probability of outcome Y given intervening with treatment T is $p(Y \mid \mathrm{do}(T))$.

$$p(Y \mid \mathrm{do}(T)) = \sum_s p(Y, S \mid \mathrm{do}(T)) = \sum_s p(Y \mid S, \mathrm{do}(T))\, p(S \mid \mathrm{do}(T)) = \sum_s p(Y \mid S, \mathrm{do}(T))\, p(S)$$

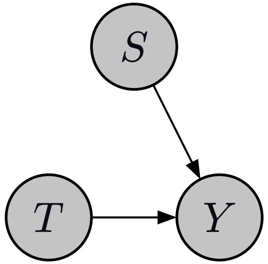
Illustration: Disease Treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

The stone size S is a confounder:



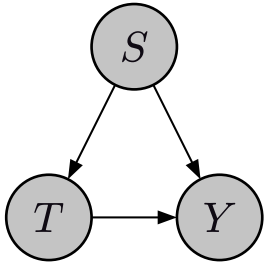
An RCT would remove the confounder:



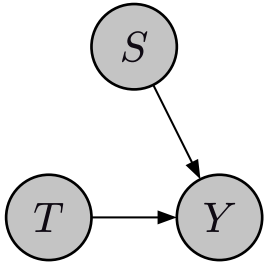
$$p\big(Y \mid \mathrm{do}(T)\big) = \sum_s p\big(Y, S \mid \mathrm{do}(T)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T)\big)\, p(S)$$

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

The stone size S is a confounder:



An RCT would remove the confounder:



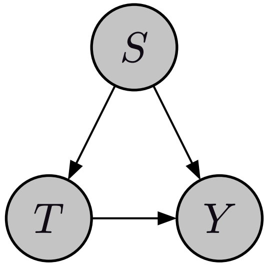
$$p\big(Y \mid \mathrm{do}(T)\big) = \sum_s p\big(Y, S \mid \mathrm{do}(T)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T)\big)\, p(S)$$

$$p\big(Y \mid \mathrm{do}(T = a)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T = a)\big)\, p(S) = 0.93 \times 51\% + 0.73 \times 49\% = 83\%$$

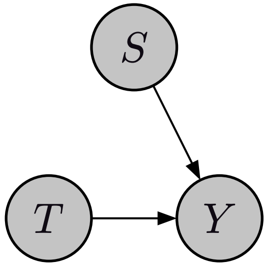
Illustration: Disease Treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

The stone size S is a confounder:



An RCT would remove the confounder:



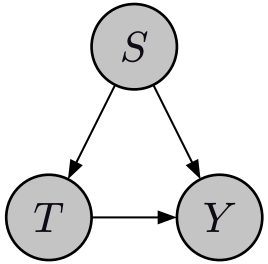
$$p\big(Y \mid \mathrm{do}(T)\big) = \sum_s p\big(Y, S \mid \mathrm{do}(T)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T)\big)\, p(S)$$

$$p\big(Y \mid \mathrm{do}(T=a)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T=a)\big)\, p(S) = 0.93 \times 51\% + 0.73 \times 49\% = 83\%$$

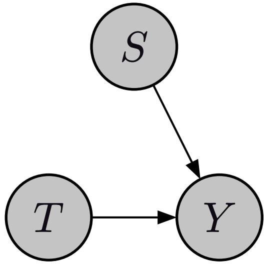
$$p\big(Y \mid \mathrm{do}(T=b)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T=b)\big)\, p(S) = 78\%$$

	Positive Outcome	Small Stones	Large Stones
Treatment A	$273/350 = 78\%$	$81/87 = 93\%$	$192/263 = 73\%$
Treatment B	$289/350 = 83\%$	$234/270 = 87\%$	$55/80 = 69\%$

The stone size S is a confounder:



An RCT would remove the confounder:



$$p\big(Y \mid \mathrm{do}(T)\big) = \sum_s p\big(Y, S \mid \mathrm{do}(T)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T)\big)\, p(S)$$

$$p\big(Y \mid \mathrm{do}(T = a)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T = a)\big)\, p(S) = 0.93 \times 51\% + 0.73 \times 49\% = 83\%$$

$$p\big(Y \mid \mathrm{do}(T = b)\big) = \sum_s p\big(Y \mid S, \mathrm{do}(T = b)\big)\, p(S) = 78\%$$

Adjusting for the confounder recovers the correct causal ranking of the treatments!

Causality Thoughts

- Statistical / probabilistic reasoning alone cannot support causal inference
- Determining the joint probability distribution of variables says nothing about causation
- Causal Inference promises to determine the necessary set of (non-data) assumptions sufficient to make a causal conclusion

Causal Graphs

A causal graph is a DAG where an edge $A \to B$ means A is a direct cause of B

Causal Graphs

A causal graph is a DAG where an edge $A \to B$ means A is a direct cause of B

- Unlike a Bayesian network, this direction is fixed; it cannot be reversed by relabelling or by Bayes' rule

A causal graph is a DAG where an edge $A \to B$ means A is a direct cause of B

- Unlike a Bayesian network, this direction is fixed; it cannot be reversed by relabelling or by Bayes' rule
- Causal graphs encode assumptions about the data-generating process itself, not just how a distribution happens to factorise

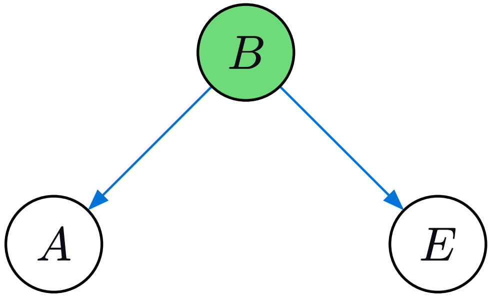
A causal graph is a DAG where an edge $A \to B$ means A is a direct cause of B

- Unlike a Bayesian network, this direction is fixed; it cannot be reversed by relabelling or by Bayes' rule
- Causal graphs encode assumptions about the data-generating process itself, not just how a distribution happens to factorise
- This extra assumption is what lets us answer interventional questions, like $p\big(Y \mid \mathrm{do}(T)\big)$ from the kidney-stone example, that a Bayesian network alone cannot

D-Separation: Forks & Chains

The structure of the paths between two variables determines whether they are dependent.

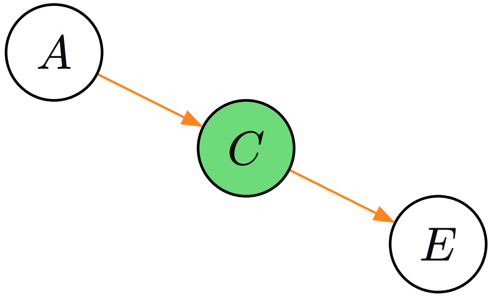
Fork (common cause)



B is a common cause of A and E. Conditioning on B **blocks** the path:

$$A \perp E \mid B$$

Chain (mediation)



C mediates between A and E. Conditioning on C **blocks** the path:

$$A \perp E \mid C$$

In both cases, **without** conditioning on the middle node information flows freely: $A \not\perp E$.

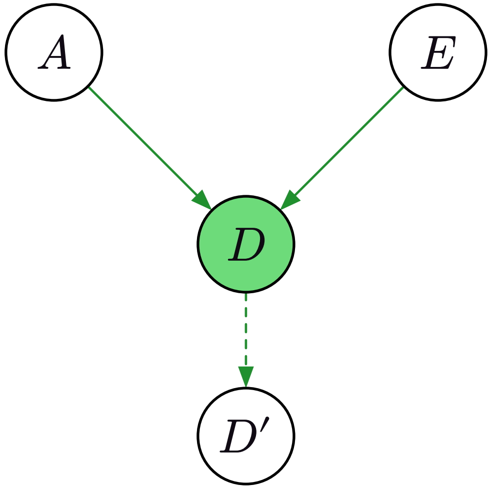
In a collider $A \to D \leftarrow E$, information flow is blocked by default:

$$A \perp E$$

Counterintuitively, conditioning on D (or on a descendant D' of D) opens the path:

$$A \not\perp E \mid D \qquad A \not\perp E \mid D'$$

This is sometimes called explaining away; learning that D occurred, having ruled out A as the cause, makes E more probable as the explanation (and vice versa).



Do-Calculus

Do-calculus evaluates properties on a modified graph with certain edges removed:

Do-calculus evaluates properties on a modified graph with certain edges removed:

- $\mathcal{G}_{\overline{X}}$: remove all edges into X (used for an intervention on X)

Do-calculus evaluates properties on a modified graph with certain edges removed:

- $\mathcal{G}_{\overline{X}}$: remove all edges into X (used for an intervention on X)
- $\mathcal{G}_{\underline{Z}}$: remove all edges out of Z

Do-Calculus

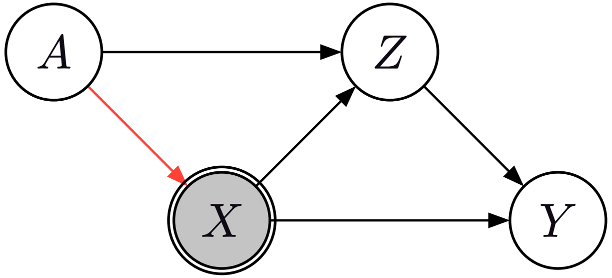
Do-calculus evaluates properties on a modified graph with certain edges removed:

- $\mathcal{G}_{\overline{X}}$: remove all edges into X (used for an intervention on X)
- $\mathcal{G}_{\underline{Z}}$: remove all edges out of Z

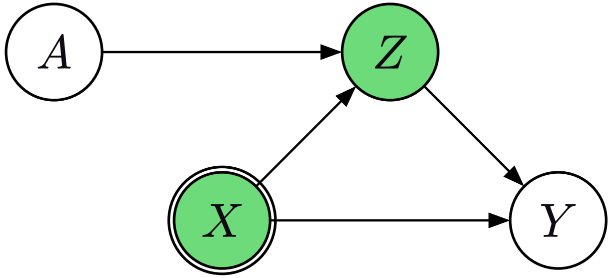
Three transformation rules let us convert between interventional and observational distributions.

Insertion / deletion of an observation. If $Y \perp A \mid \{X, Z\}$ in $\mathcal{G}_{\overline{\overline{X}}}$, we may add or remove the observation A from a conditional distribution:

$$p\big(Y \bigm| Z, A, \mathrm{do}(X)\big) \longleftrightarrow p\big(Y \bigm| Z, \mathrm{do}(X)\big)$$



$\mathcal{G}$

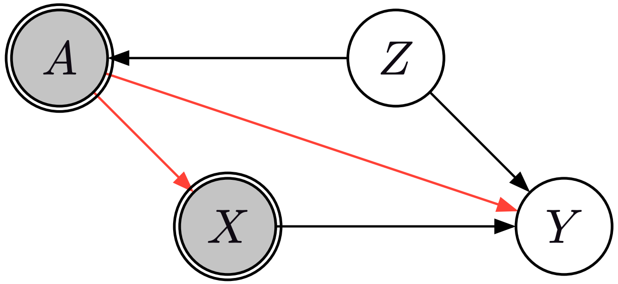


$\mathcal{G}_{\overline{\overline{X}}}$

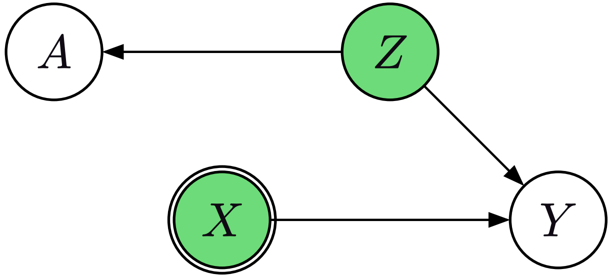
Exchanging an intervention and an observation. If $Y \perp A \mid \{X, Z\}$ in $\mathcal{G}_{\overline{X}\underline{A}}$, we may swap an intervention on A for an observation of A:

$$p\Big(Y \;\Big|\; Z, \mathrm{do}(A), \mathrm{do}(X)\Big) \longleftrightarrow p\Big(Y \;\Big|\; Z, A, \mathrm{do}(X)\Big)$$

This is the most important rule: it lets us infer the effect of an intervention from passive observation, whenever A "behaves as if randomised" once we condition on Z.



$\mathcal{G}$

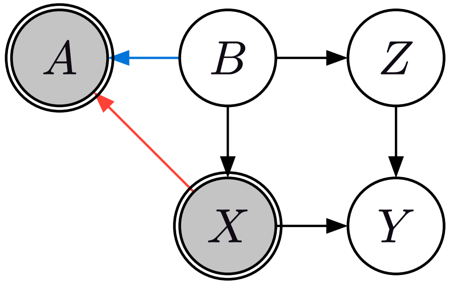


$\mathcal{G}_{\overline{X}\underline{A}}$

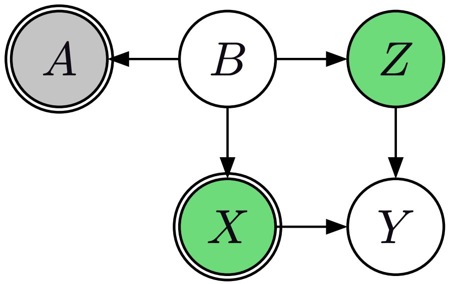
Insertion / deletion of an intervention. If $Y \perp A \mid \{X, Z\}$ in $\mathcal{G}_{\bar{\bar{A}}/\mathrm{pa}(Z)}$, removing all edges into A except those from ancestors of Z, we may add or remove the intervention on A:

$$p\Big(Y \;\Big|\; Z, \mathrm{do}(A), \mathrm{do}(X)\Big) \longleftrightarrow p\Big(Y \;\Big|\; Z, \mathrm{do}(X)\Big)$$

This tells us A has no causal effect on Y in this context, so the intervention can simply be dropped.



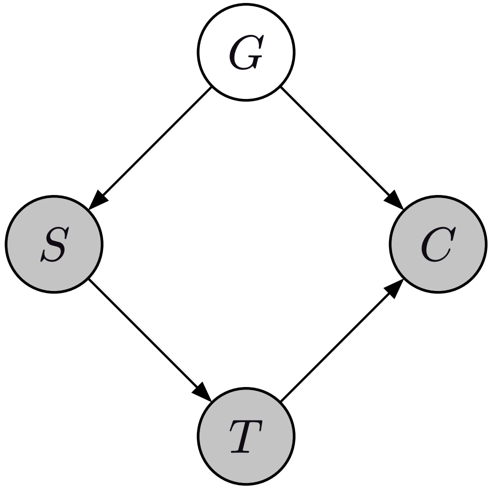
$\mathcal{G}$



$\mathcal{G}_{\bar{\bar{A}}/\mathrm{pa}(Z)}$

Worked Example

We wish to infer the causal effect of smoking on lung cancer complicated by a possible genetic predisposition that affects both smoking and cancer risk. Measuring tar buildup (a mediator) will let us block this confounding path.



$\mathcal{G}$

- S: does the individual smoke? (observed)
- T: is there a tar buildup in the lungs? (observed, mediator)
- C: does the individual have lung cancer? (observed)
- G: genetic predisposition. (latent, confounds S and C)

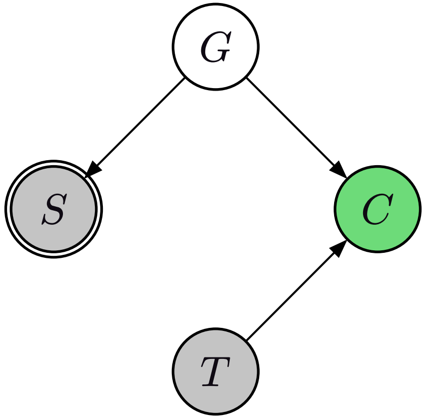
We need the interventional distribution $p\left(C \mid \mathrm{do}(S)\right)$ expressed purely in terms of observational quantities.

$$p\big(C \mid \mathrm{do}(S)\big) = \sum_T p\big(C, T \mid \mathrm{do}(S)\big) = \sum_T p\big(C \mid T, \mathrm{do}(S)\big) p\big(T \mid \mathrm{do}(S)\big) \quad \text{(Marginalisation)}$$

Applying Rule 2 to swap the intervention on S for an observation:

$$p\big(T \mid \mathrm{do}(S)\big) = p\big(T \mid S\big) \qquad \text{(Rule 2: } T \perp S \text{ in } \mathcal{G}_{\underline{S}} \text{, blocked by } C \text{)}$$

$$\implies p\big(C \mid \mathrm{do}(S)\big) = \sum_T p\big(C \mid T, \mathrm{do}(S)\big) p\big(T \mid S\big)$$



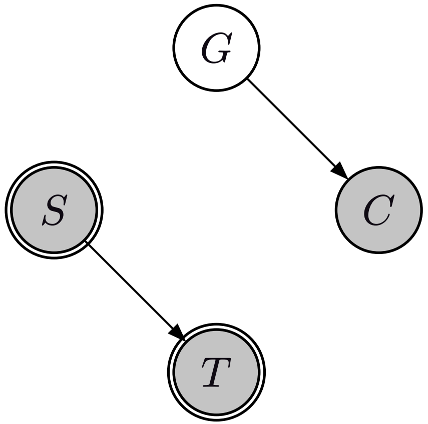
$\mathcal{G}_{\underline{S}}$

Worked Example: Step 2 — Block the Back-Door via Tar

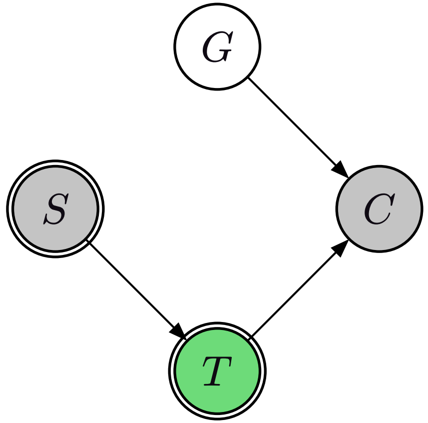
We cannot yet swap $p\big(C \mid T, \mathrm{do}(S)\big)$ directly: G still confounds S and C. Instead we use the mediator T to block that back-door path:

$$
\begin{aligned}
p\big(C \mid T, \mathrm{do}(S)\big) &= p\big(C \mid \mathrm{do}(S), \mathrm{do}(T)\big) \qquad (\text{Rule 2: } \ C \perp T \ \text{ in } \ \mathcal{G}_{\bar{\bar{S}}\underline{\underline{T}}}\ ) \\
&= p\big(C \mid \mathrm{do}(T)\big) \qquad (\text{Rule 3: } \ C \perp S \mid T \ \text{ in } \ \mathcal{G}_{\bar{\bar{S}}}\ )
\end{aligned}
$$

$$
\Longrightarrow p\big(C \mid \mathrm{do}(S)\big) = \sum_T p\big(C \mid \mathrm{do}(T)\big) p\big(T \mid S\big)
$$



$\mathcal{G}_{\bar{\bar{S}}\underline{\underline{T}}}$

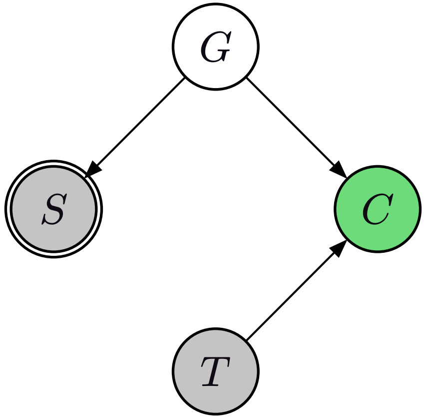


$\mathcal{G}_{\bar{\bar{S}}}$

$$p\big(C \mid \mathrm{do}(T)\big) = \sum_{S'} p\big(C, S' \mid \mathrm{do}(T)\big) = \sum_{S'} p\big(C \mid S', \mathrm{do}(T)\big) p\big(S' \mid \mathrm{do}(T)\big) \quad \text{(Marginalisation)}$$

Applying Rule 3 to drop the (now irrelevant) intervention on T:

$$p\big(S \mid \mathrm{do}(T)\big) = p(S) \qquad \left(\text{Rule 3: } S \perp T \text{ in } \mathcal{G}_{\overline{T}}, \text{ blocked by } C\right)$$



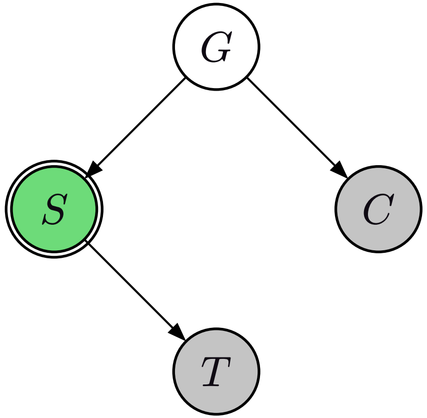
$\mathcal{G}_{\overline{T}}$

Worked Example: Putting It All Together

For the remaining term, Rule 2 lets us swap the intervention on T once we condition on S:

$$p\big(C \mid S, \mathrm{do}(T)\big) = p\big(C \mid S, T\big) \qquad (\text{Rule 2: } \; C \perp T \mid S \; \text{ in } \; \mathcal{G}_{\underline{T}} \;)$$

$$\Longrightarrow p\big(C \mid \mathrm{do}(T)\big) = \sum_{S'} p\big(C \mid S', T\big) p(S')$$



$\mathcal{G}_{\underline{T}}$

Overall, the interventional distribution reduces to purely observational terms:

$$p\big(C \mid \mathrm{do}(S)\big) = \sum_{T} p\big(C \mid \mathrm{do}(T)\big) p\big(T \mid S\big) = \sum_{T} \left(\sum_{S'} p\big(C \mid S', T\big) p(S') \right) p\big(T \mid S\big)$$

Structural Causal Models

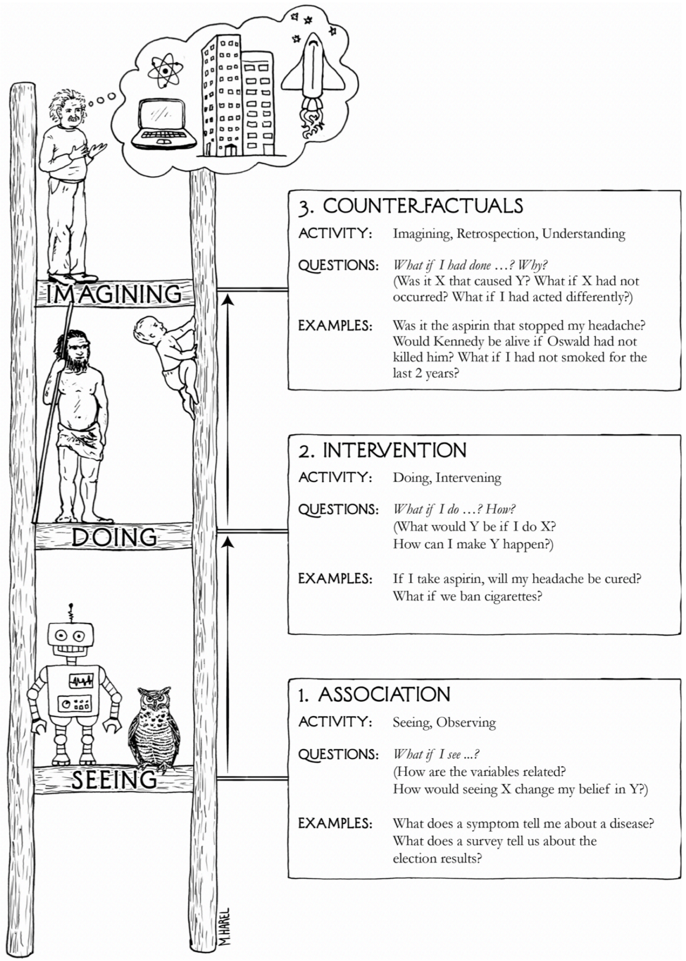
Why SCMs Matter

Causal knowledge cannot be extracted from data alone — you need a model that encodes assumptions about the world's mechanisms. Structural causal models make those assumptions explicit, testable, and mathematically tractable.

They provide a unified framework for:

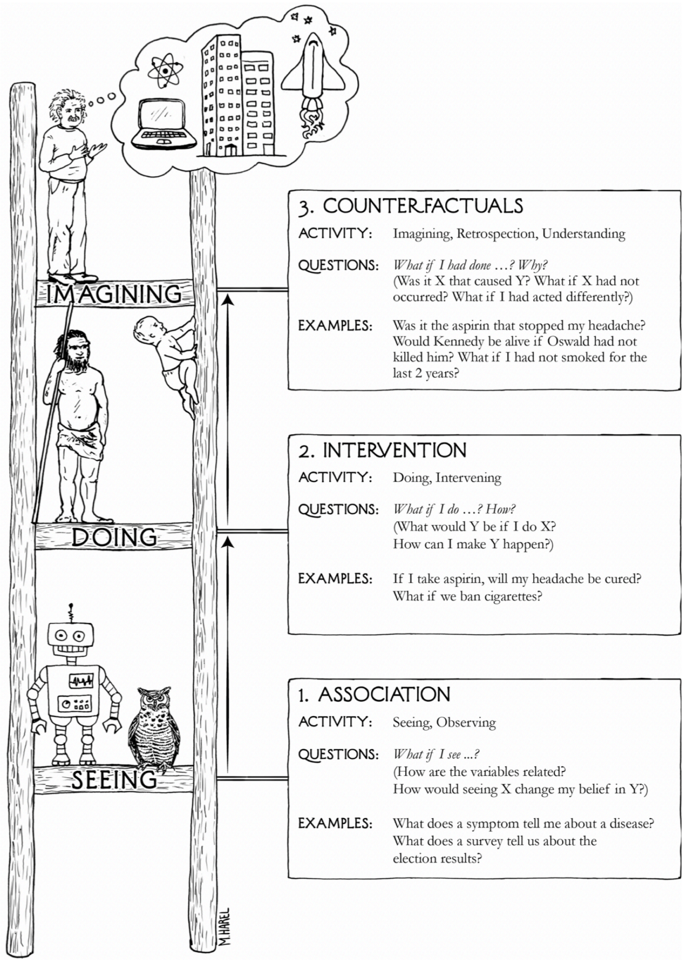
- Causal inference from observational data
- Experimental design — what to randomise or control
- Mediation analysis — direct vs. indirect effects
- Policy evaluation — what happens if we intervene?
- Fairness in ML — defining and detecting discriminatory causation
- Explainable AI — counterfactual explanations

Conclusion:ML for Causation...



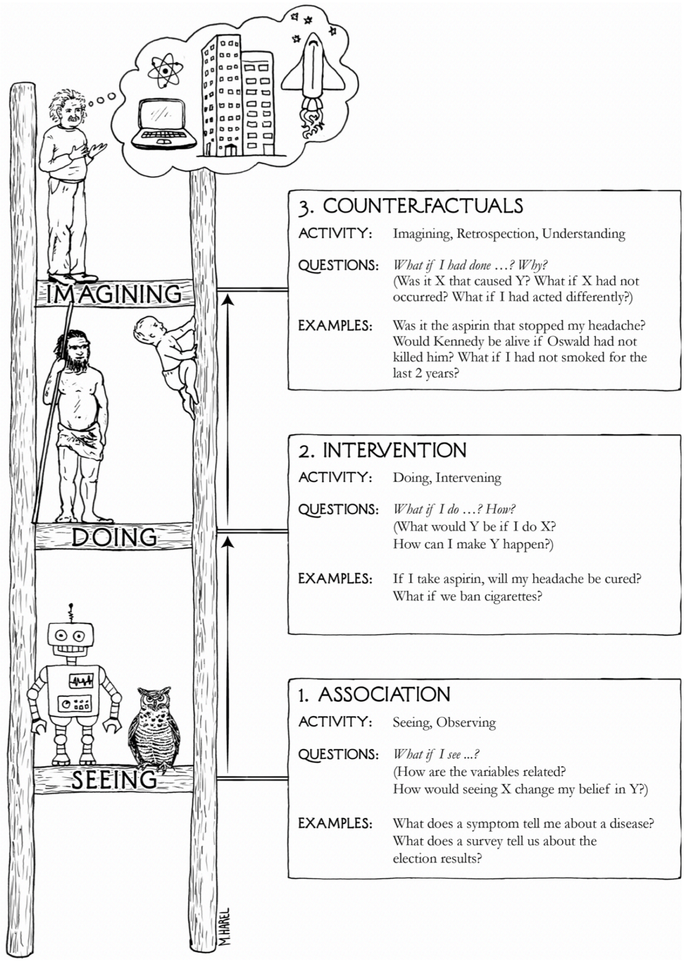
[Pearl and Mackenzie 2017]

- Level 1: Association: correlation, regression, classification
  - Mostly what we are doing in ML today



[Pearl and Mackenzie 2017]

- Level 1: Association: correlation, regression, classification
  - Mostly what we are doing in ML today
- Level 2: Intervention: causal effect, A/B testing, policy
  - Make causal claims at population level from observational data
  - Concerned with the effects of interventions (e.g., “what happens if we do X?”)
  - Models concerned with identifiability of causal effects from observational data



[Pearl and Mackenzie 2017]

- Level 1: Association: correlation, regression, classification
  - Mostly what we are doing in ML today
- Level 2: Intervention: causal effect, A/B testing, policy
  - Make causal claims at population level from observational data
  - Concerned with the effects of interventions (e.g., “what happens if we do X?”)
  - Models concerned with identifiability of causal effects from observational data
- Level 3: Counterfactuals: “what if?”, “why?”, fairness
  - Make causal claims for individuals (from observational data)
  - Can we impute what would have happened to a *specific individual* if they had received a different treatment?

References

- J. Pearl (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.).
- J. Pearl and D. Mackenzie (2017). *The Book of Why: The New Science of Cause and Effect*.

- Thanks to Julian Faraway for the correlation/causation illustrations used earlier in this talk.

Questions?