

# Causal Inference in Machine Learning



**Ricardo Silva**

Department of Statistical Science and  
Centre for Computational Statistics and Machine Learning

[ricardo@stats.ucl.ac.uk](mailto:ricardo@stats.ucl.ac.uk)

---

Part I:  
Did you have breakfast today?

---

---

Researchers reviewed 47 nutrition studies and concluded that **children and adolescents who ate breakfast had better mental function** and better school attendance records than those who did not.

**They suggested several possible reasons.** For example, eating breakfast may modulate short-term metabolic responses to fasting, cause changes in neurotransmitter concentrations or simply eliminate the distracting physiological effects of hunger.

# Spurious causality (?)

## ■ Eating makes you faithful

- *Will he cheat? How to tell. Ladies, you probably think that it's just in his nature. He can't help it - he HAS to cheat. But here's the sad truth: **you're not feeding him enough**. If you're worried your guy might cheat, try checking out his waistline. A new study says the size of his belly may reveal whether he'll stray.* [www.match.com/magazine/article/4273/Will-He-Cheat?-How-To-Tell](http://www.match.com/magazine/article/4273/Will-He-Cheat?-How-To-Tell)

## ■ Relaxing makes you die

- *In a prospective cohort study of thousands of employees who worked at Shell Oil, the investigators found that **embarking on the Golden Years at age 55 doubled the risk for death before reaching age 65**, compared with those who toiled beyond age 60.*

<http://www.medpagetoday.com/PrimaryCare/PreventiveCare/1980>

# What is a cause, after all?

- A causes B:

$$P(B \mid A \text{ is manipulated to } a_1) \neq P(B \mid A \text{ is manipulated to } a_2)$$

- Next the concept of an *external agent*
- Examples of manipulations:
  - Medical interventions (treatments)
  - Public policies (tax cuts for the rich)
  - Private policies (50% off! Everything must go!)
- A manipulation (intervention, policy, treatment, etc.) changes the data generation mechanism. It sets a *new regime*



# But what exactly is a manipulation?

- Some intervention  $T$  on  $A$  can only be “effective” if  $T$  is a cause of  $A$
- ??!??
- Don't be afraid of circularities
  - Or come up with something better, if you can

**Bart:** What is "the mind"? Is it just a system of impulses or is it...something tangible?

**Homer:** Relax. What is mind? No matter. What is matter? Never mind.



**Simpsons, The (1987)**

---

# An axiomatic system

- When you can't define something, axiomatize it:
    - From points to lines and beyond
  - We will describe languages that have causal concepts as primitives
  - The goal: use such languages to
    - Express causal assumptions
    - Compute answers to causal queries that are entailed by such assumptions
-

---

## Causal queries: hypothetical causation vs. counterfactual causation

- I have a headache. If I take an aspirin now, will it go away?
  - I had a headache, but it passed. Was it because I took an aspirin two hours ago? Had I not taken such an aspirin, would I still have a headache?
-



---

# Prediction vs. explanation

- The first case is a typical “predictive” question
    - You are calculating the effect of a hypothetical intervention
    - Pretty much within decision theory
      - Think well before offering the 50% discount!
  - The second case is a typical “explanatory” question
    - You are calculating the effect of a counterfactual intervention
      - Have things been different...
    - Ex.: law
  - What about scientific/medical explanation?
-

---

# Prediction vs. explanation

- This talk will focus solely on prediction
  - Explanation is fascinating, but too messy, and not particularly useful (at least as far as Science goes)...
-

---

# Preparing axioms: Seeing vs. doing

- Observe again the notation

$$P(B \mid A \text{ is manipulated to } a_1)$$

- Why not...

$$P(B \mid A = a_1)$$

...?

---

---

# Seeing vs. doing: an example

- The reading in a barometer is useful to predict rain

$$P(\text{rain} \mid \text{barometer reading} = \text{high}) > \\ P(\text{rain} \mid \text{barometer reading} = \text{low})$$

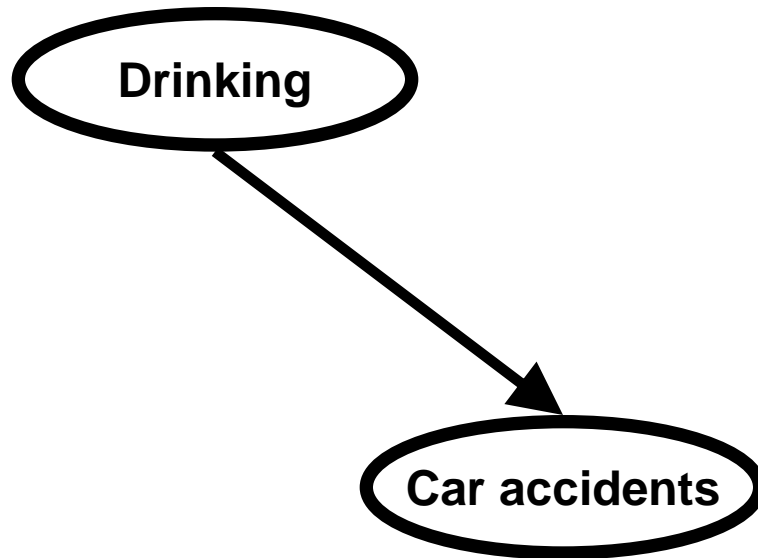
- But hacking a barometer won't cause rain

$$P(\text{rain} \mid \text{barometer hacked to high}) = \\ P(\text{rain} \mid \text{barometer hacked to low})$$

- (Sometimes this is called intervening vs. conditioning. You should see this as a way of indexing regimes.)
-

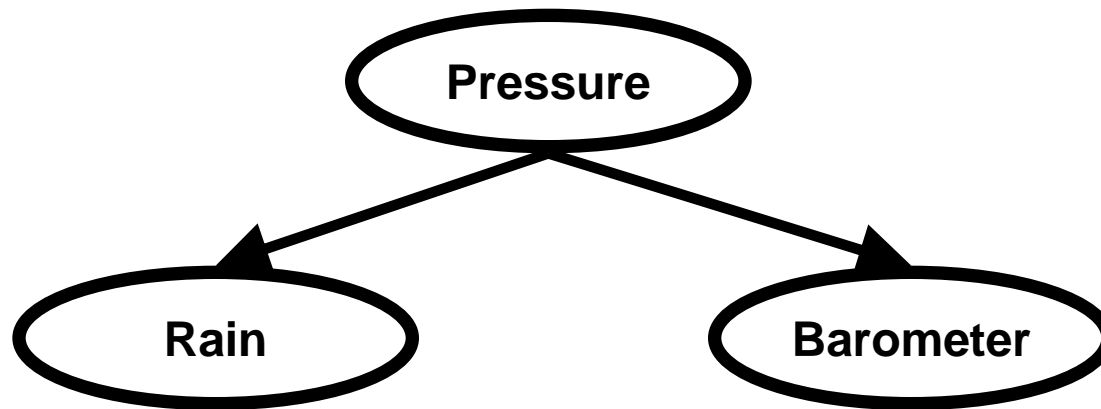
# Why is seeing different from doing?

- Issue #1: directionality



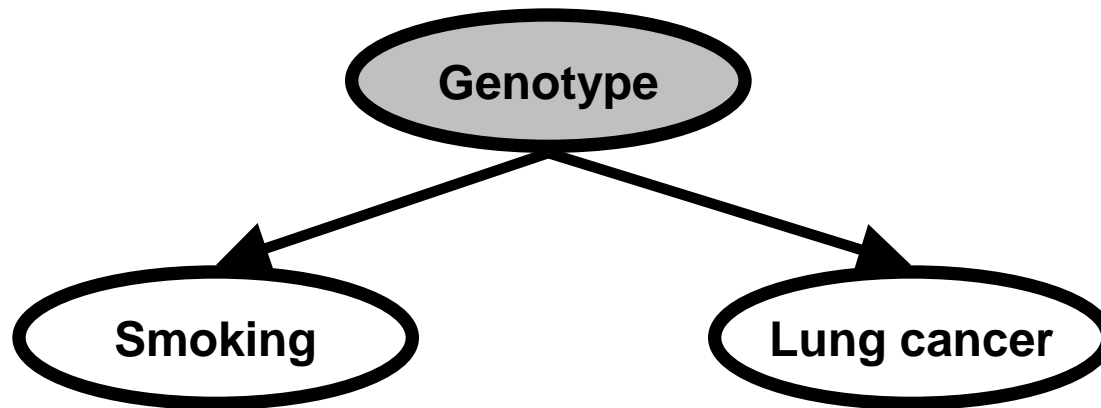
# Why is seeing different from doing?

- Issue #2: confounding (i.e., common causes)



# Why is seeing different from doing?

- *Most important lesson:* unmeasured confounding (i.e., hidden common causes) is perhaps the most complicating factor of all



- (but see also: measurement error and sampling selection bias)

# The *do* operator (Pearl's notation)

- A shorter notation
- $P(A \mid B = b)$ : the probability of  $A$  being true given an observation of  $B = b$ 
  - That is, no external intervention
  - This is sometimes called the distribution under the *natural state* of  $A$
- $P(A \mid \text{do}(B = b))$ : the probability of  $A$  given an intervention that sets  $B$  to  $b$ 
  - $P(A \mid \text{do}(B))$ : some shorter notation for  $\text{do}(B) = \text{true}$



---

# Different do's

- $P(A \mid \text{do}(B), C)$ 
    - Intervening on B, seeing C
  - $P(A \mid \text{do}(B), \text{do}(C))$ 
    - Multiple interventions
  - $P(A \mid \text{do}(P(B) = P'))$ 
    - A change on the distribution of B (not only a point mass distribution)
-

# Causal models

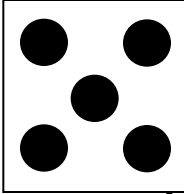
- A causal model is defined by a set of  $P(A_1, A_2, \dots, A_N \mid \text{do}(B_1), \text{do}(B_2), \dots, \text{do}(B_M), B_{M+1}, B_{M+2}, \dots, B_O)$
- How to estimate this? Which data can I use?
- The Radical Empiricist says:

Every *do* is a change of regime. Anything can happen. In general, there is no logical connection between states!

*Every different set of do's specify a brave new World.*  
**(or does it?)**

# Learning causal models

- The gold standard\*: randomized experiments

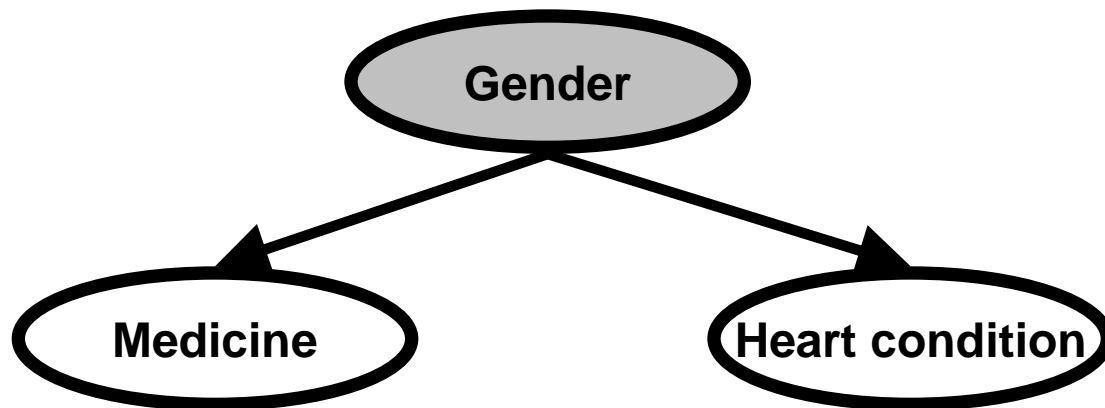


<b>Treatment</b>	<b>Patient <i>ID</i></b>	<b>Age</b>	<b>Heart Condition</b>
Medicine	1	32	+
Medicine	2	41	+
Placebo	3	40	0
Placebo	4	37	0
Medicine	5	36	0
...	...	...	...

\*and a recipe for knighthood

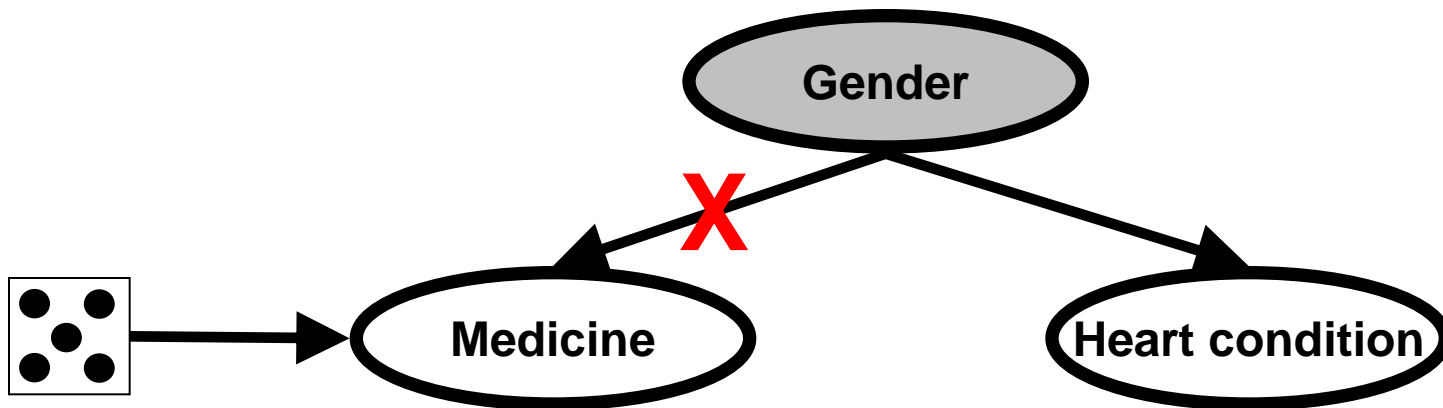
# The role of randomization

- Breaking the hidden common causes
- Example: gender may cause both self-selection of treatment, and heart condition



# The role of randomization

- The randomized assignment overrides the original causal mechanisms



- Notice: placebo is a surrogate for no-treatment
- With blind/double-blind assignments, its role is to avoid psychological effects

---

# Causal models

- A causal model is defined by a set of  $P(A_1, A_2, \dots, A_N \mid \text{do}(B_1), \text{do}(B_2), \dots, \text{do}(B_M), B_{M+1}, B_{M+2}, \dots, B_O)$
  - Do I always have to perform an experiment?
-

---

# Observational studies

- The art and science of inferring causation without experiments
  - This can only be accomplished if extras assumptions are added
  - Most notable case: inferring the link between smoking and lung cancer
  - This tutorial will focus on observational studies
-

---

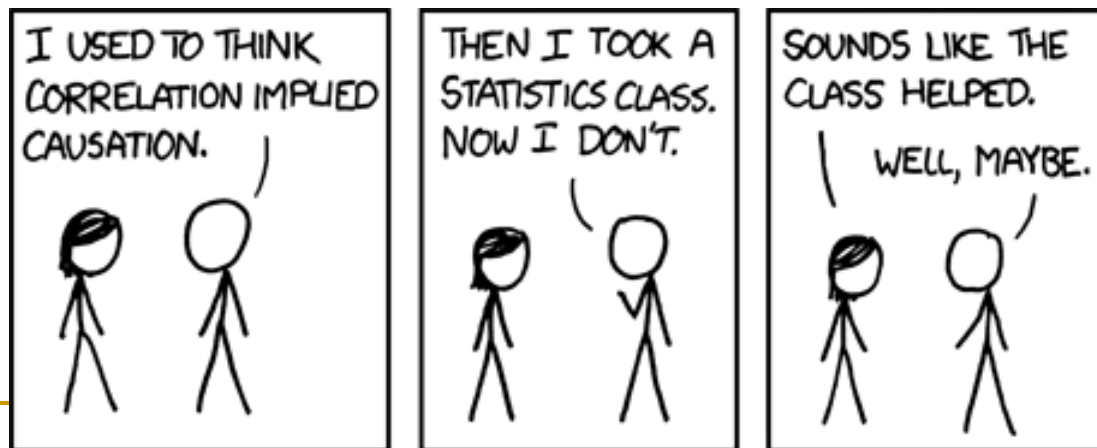
# Observational studies

- If you can do a randomized experiment, you should do it
  - Observational studies have important roles, though:
    - When experiments are impossible for unethical/practical reasons
      - The case for smoking/lung cancer link
    - When there are many experiments to perform
      - A type of exploratory data analysis/active learning tool
        - E.g., biological systems
-



# Observational studies

- It *is* certainly true that correlation is not causation
  - And as statisticians know, it may well be the case correlation-hat is not even correlation
- But it is also lazy to stop there



# Observational studies

Copyrighted Material

# FREAKONOMICS

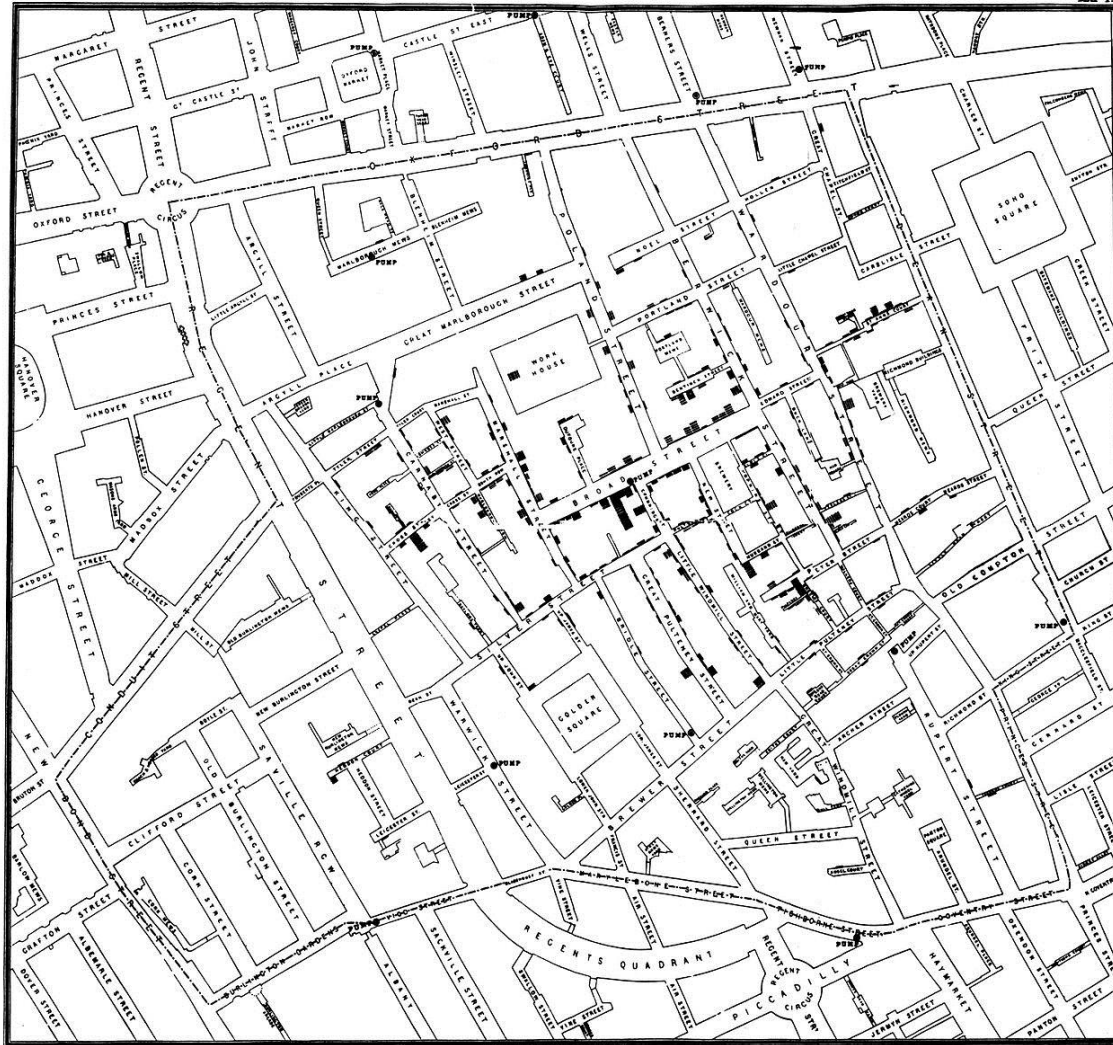
**A ROGUE ECONOMIST EXPLORES  
THE HIDDEN SIDE OF EVERYTHING**

**“Prepare to be dazzled.”**

— Malcolm Gladwell, author of *The Tipping Point* and *Blink*



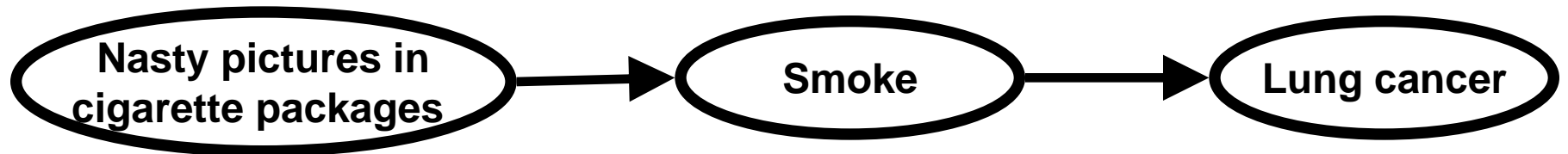
# John Snow's Soho



All image sources:  
Wikipedia

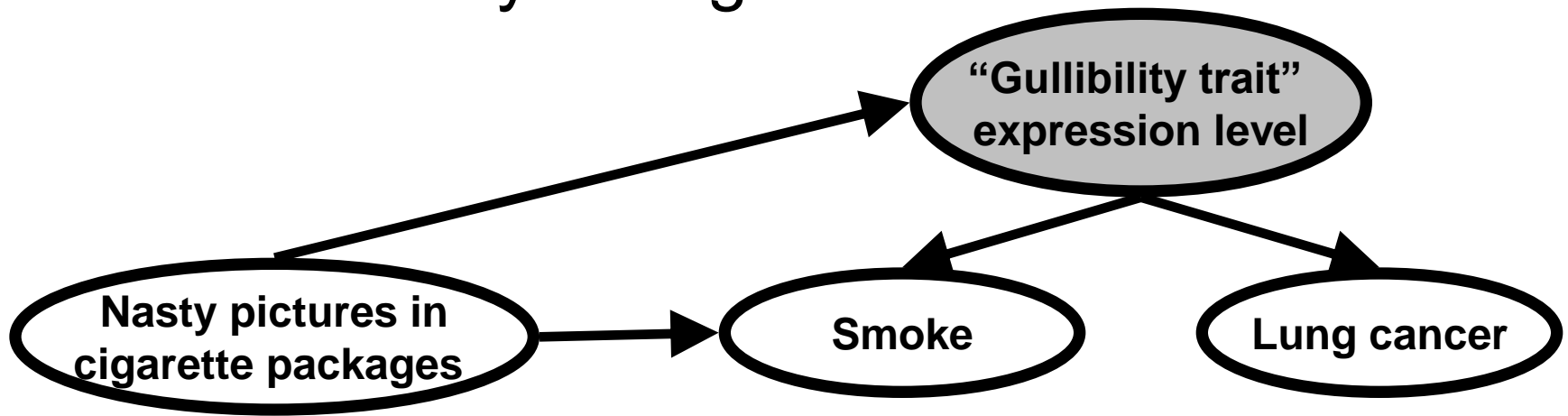
# Observational studies

- But in the end, don't we always have a testable condition?



# Observational studies

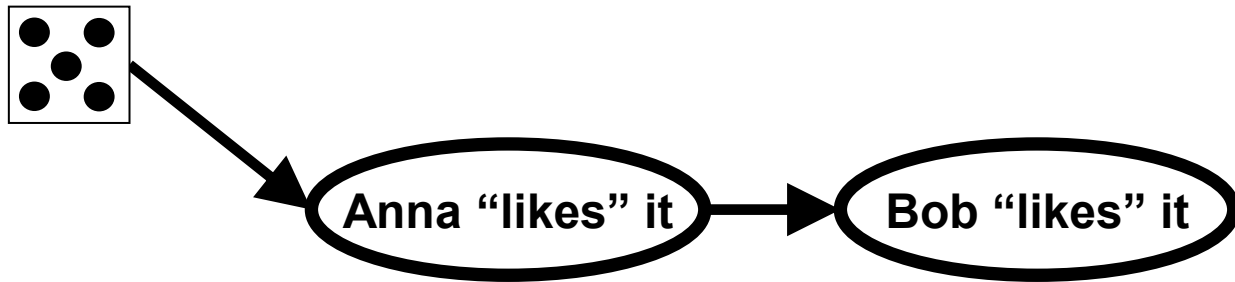
- Appropriate interventions are much more subtle than you might think...



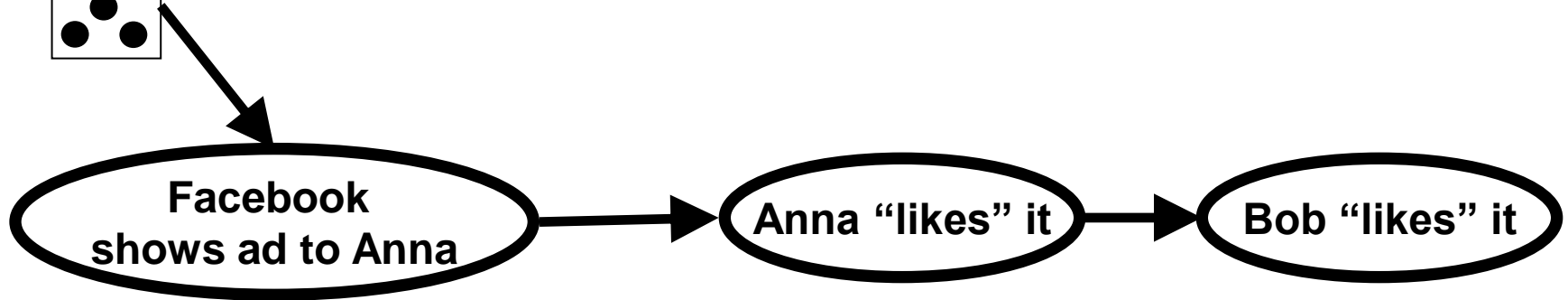
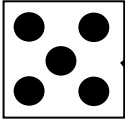
Smoke  $\perp\!\!\!\perp$  Lung cancer | do(Smoke)

# (Sort of) Observational studies

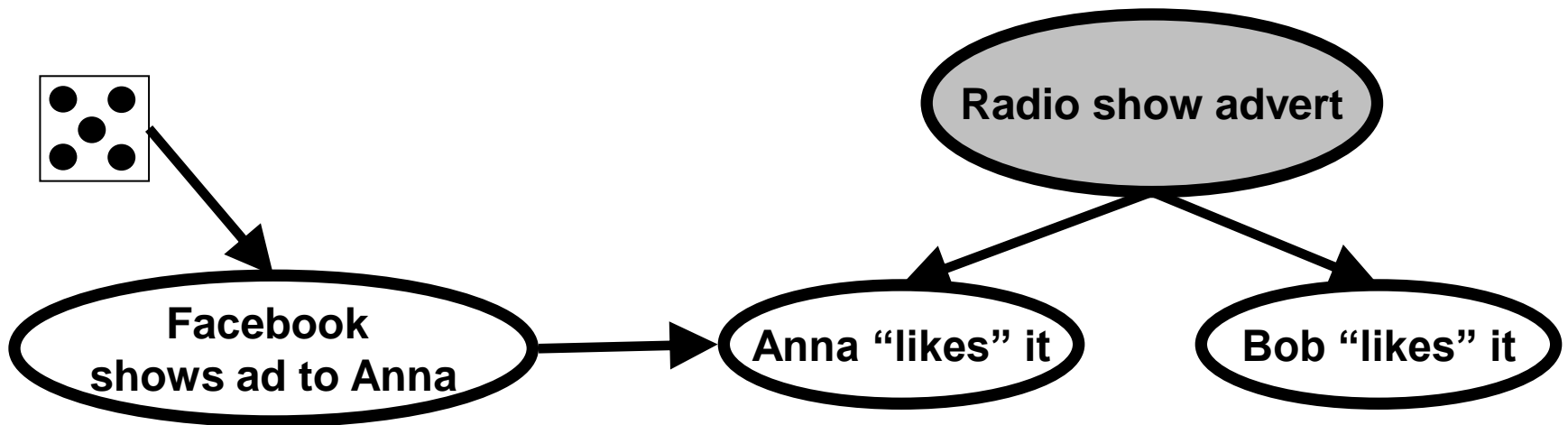
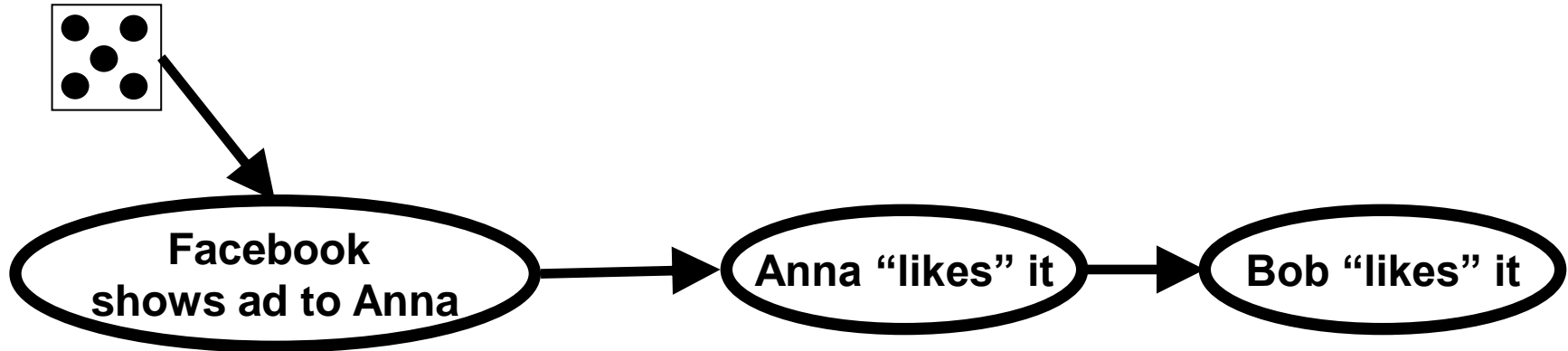
- But I'm Facebook and I have 1 googol-pounds of money for experiments. I'm covered, right?



# (Sort of) Observational studies

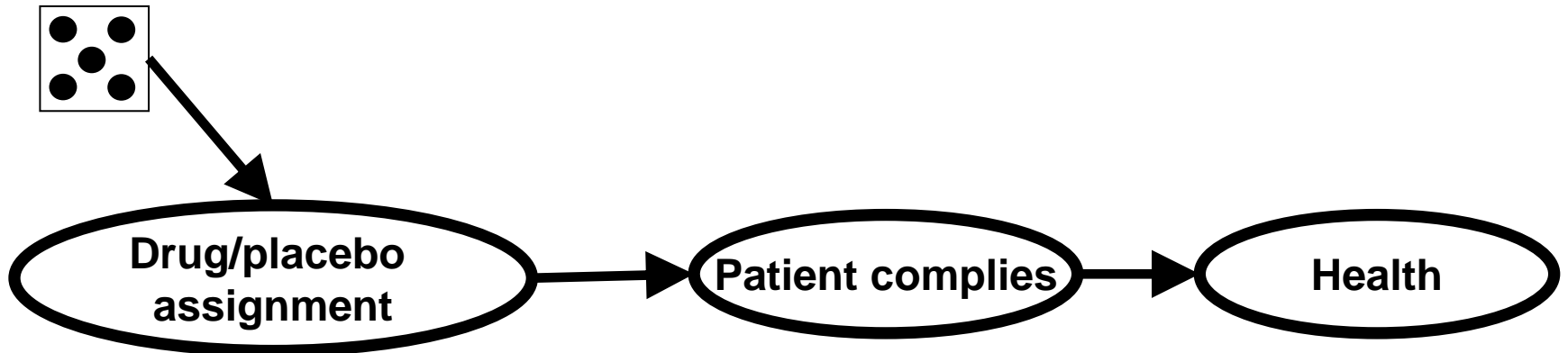


# (Sort of) Observational studies

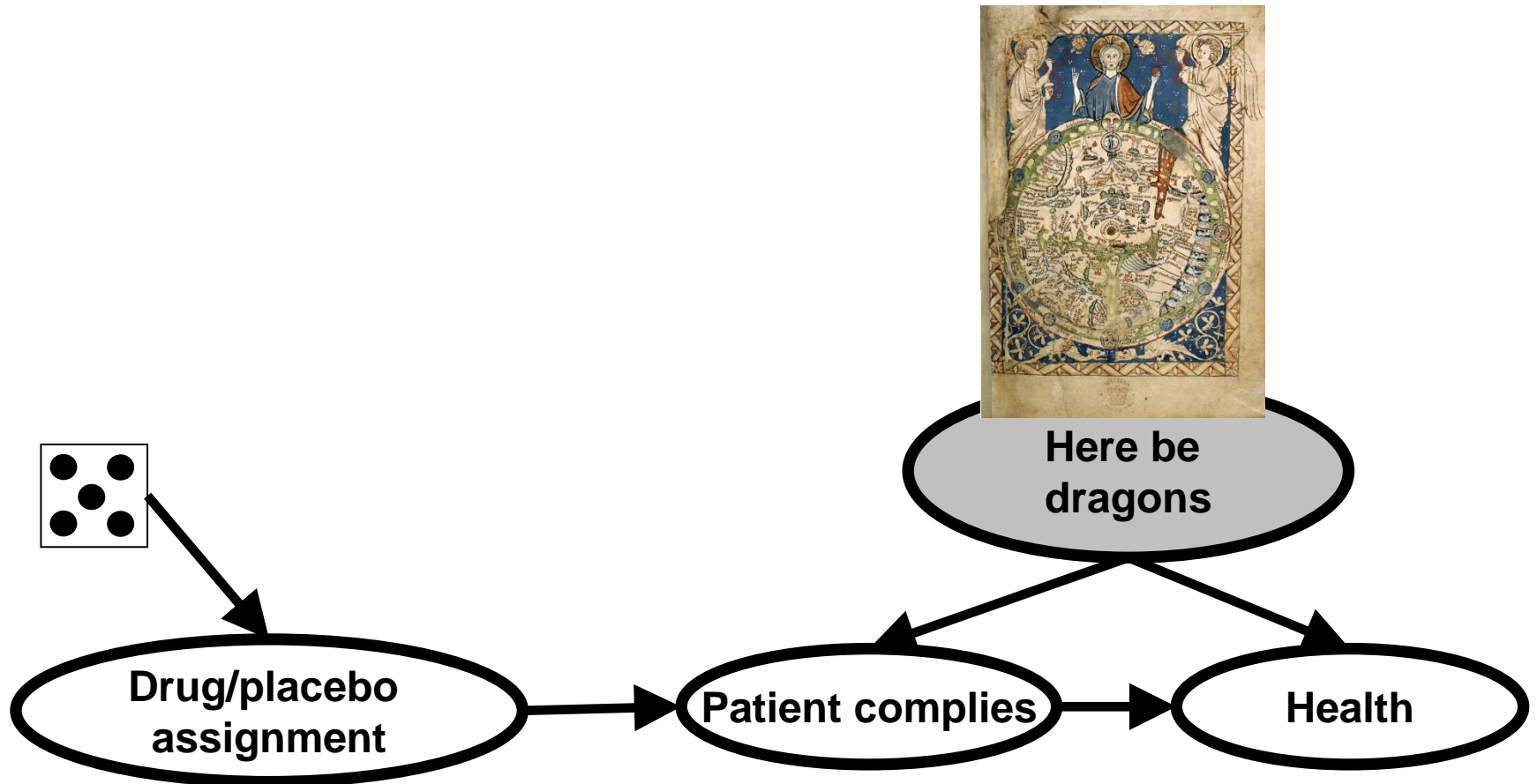




# (Sort of) Observational studies



# (Sort of) Observational studies



---

# Observational studies: starting from natural state models

- How are full joint/conditional distributions specified?

$$P(A_1, A_2, \dots, A_N \mid B_1, B_2, \dots, B_M, B_{M+1}, B_{M+2}, \dots, B_O)$$

- There is a notion of modularity in the natural state. Why wouldn't we have some *stable modularity across "Worlds"*?
-

---

# Definitions and axioms of causal modularity: DAGs

- = Directed acyclic graphs
  - Start with a “reference system”, a set of events/random variables  $V$
  - Each element of  $V$  is a vertex in causal graph  $G$
  - $A$  causes  $B$  in causal graph  $G$  only if  $A$  is an ancestor of  $B$
  - DAGs with such an assumption are causal graphs
-

---

# Definitions and axioms of causal modularity

- $A$  is a *direct cause* of  $B$  wrt  $V$  if and only if  $A$  causes  $B$  for some choice of intervention in  $V \setminus \{A, B\}$
- “ $A$  is a direct cause of  $B$ ” implies the edge



---

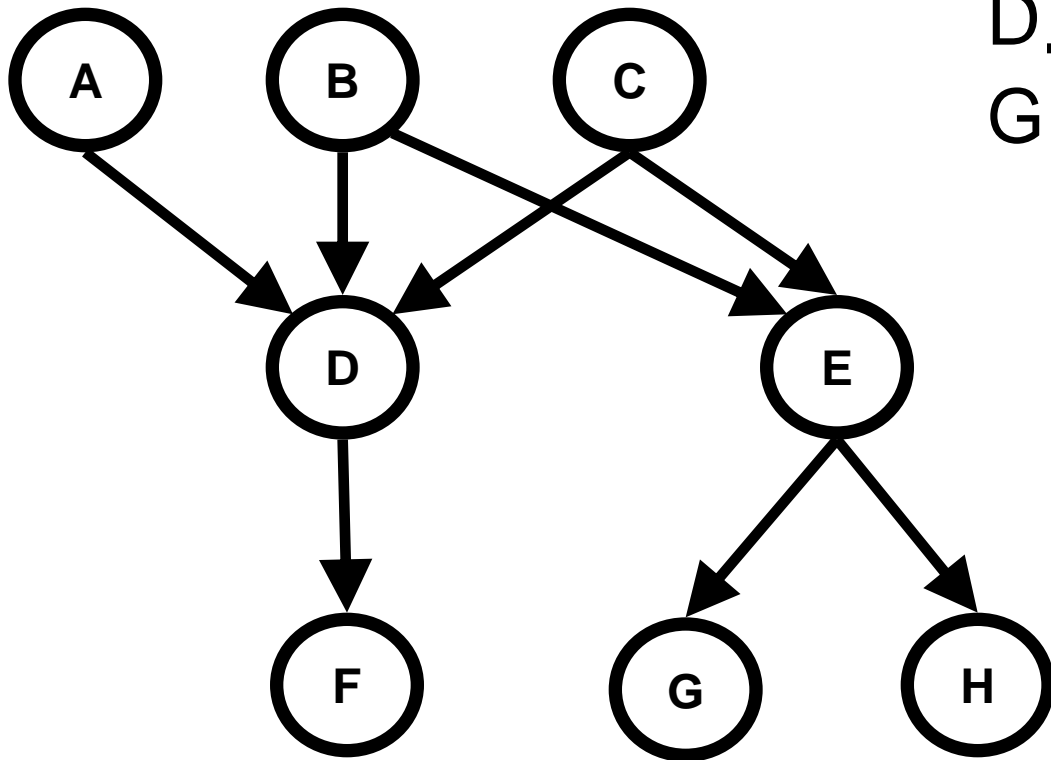
# The Causal Markov Condition

- Let  $G$  be a DAG representing a causal system over  $V$ , and  $P$  a distribution over  $V$
- $(G, P)$  satisfy the Causal Markov Condition if and only if:

**$A \perp\!\!\!\perp \{\text{All of its (non-parental) non-descendants}\} \mid A\text{'s parents}$**

where  $A$ 's parents are its direct causes in  $G$

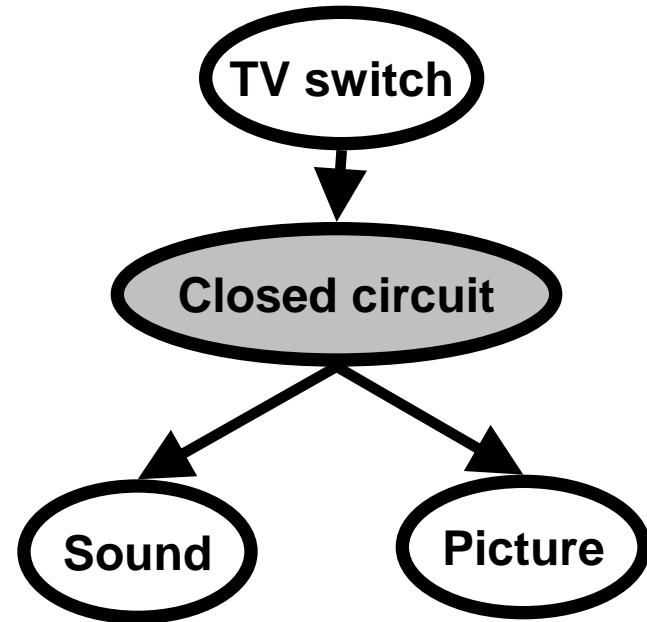
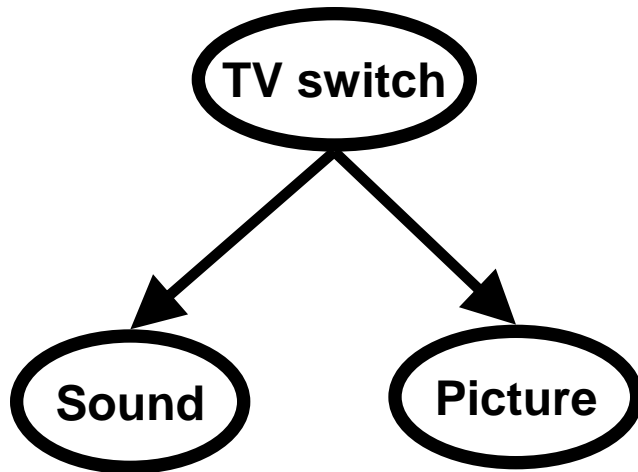
# The Causal Markov Condition



$D \perp\!\!\!\perp \{E, G, H\} \mid \{A, B, C\}$   
 $G \perp\!\!\!\perp \text{everybody else} \mid E$

# Limitations of the Causal Markov condition?

“The Interactive Fork”



$P(\text{Picture} \mid \text{Switch}) < P(\text{Picture} \mid \text{Switch}, \text{Sound})$   
*Where did the independence go?*



---

# Causal models, revisited

- Instead of an exhaustive “table of interventional distributions”:
    - $G = (V, E)$ , a causal graph with vertices  $V$  and edges  $E$
    - $P(\theta)$ , a probability over the “natural state” of  $V$ , parameterized by  $\theta$
    - $(G, \theta)$  is a causal model if pair  $(G, P)$  satisfies the Causal Markov condition
    - We will show how to compute the effect of interventions
-

---

# To summarize: what's different?

- As you probably know, DAG models can be non-causal
- What makes



causal?

Answer: **because I said so!**

---

---

# To summarize

- *A causal graph is a way of encoding causal assumptions*
  - *Graphical models allow for the evaluation of the consequences of said assumptions*
  - Typical criticism:
    - “this does not advance the ‘understanding’ of causality”
  - However, it is sufficient for predictions
  - And no useful non-equivalent alternatives are offered
-

# Example of axioms in action: Simpson's paradox

Combined	$E$	$\neg E$		Recovery Rate
drug ( $C$ )	20	20	40	50%
no-drug ( $\neg C$ )	16	24	40	40%
	36	44	80	

Males	$E$	$\neg E$		Recovery Rate
drug ( $C$ )	18	12	30	60%
no-drug ( $\neg C$ )	7	3	10	70%
	25	15	40	

Females	$E$	$\neg E$		Recovery Rate
drug ( $C$ )	2	8	10	20%
no-drug ( $\neg C$ )	9	21	30	30%
	11	29	40	

The "paradox":

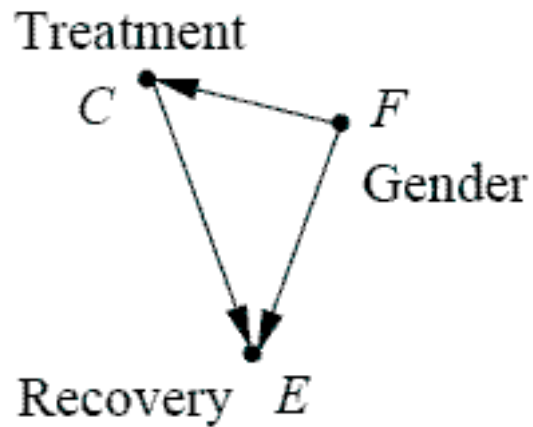
$$P(E \mid F, C) < P(E \mid F, \sim C)$$

$$P(E \mid \sim F, C) < P(E \mid \sim F, \sim C)$$

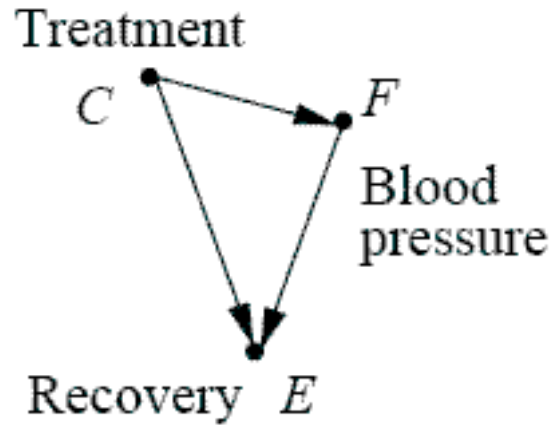
$$P(E \mid C) > P(E \mid \sim C)$$

Which table to use?  
(i.e., condition on gender or not?)

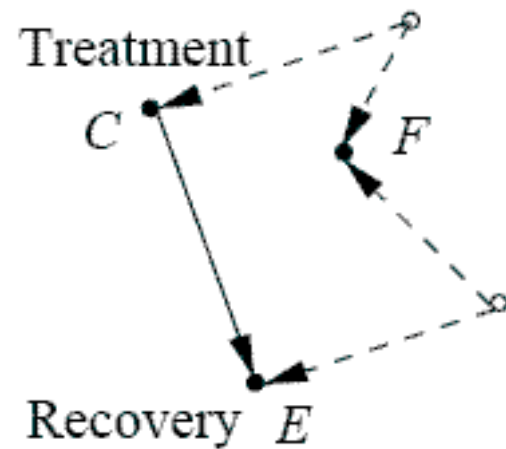
# To condition or not to condition: some possible causal graphs



(a)



(b)



(c)

# Dissolving a “paradox” using the *do* operator

- Let our population have some subpopulations
  - Say,  $F$  and  $\sim F$
- Let our treatment  $C$  not cause changes in the distribution of the subpopulations
  - $P(F \mid \text{do}(C)) = P(F \mid \text{do}(\sim C)) = P(F)$
- Then for outcome  $E$  it is impossible that we have, simultaneously,
  - $P(E \mid \text{do}(C), F) < P(E \mid \text{do}(\sim C), F)$
  - $P(E \mid \text{do}(C), \sim F) < P(E \mid \text{do}(\sim C), \sim F)$
  - $P(E \mid \text{do}(C)) > P(E \mid \text{do}(\sim C))$

# Proof

$$\begin{aligned}P(E|do(C)) &= P(E|do(C), F)P(F|do(C)) \\ &\quad + P(E|do(C), \neg F)P(\neg F|do(C)) \\ &= P(E|do(C), F)P(F) + P(E|do(C), \neg F)P(\neg F).\end{aligned}$$

$$\begin{aligned}P(E|do(\neg C)) &= P(E|do(\neg C), F)P(F) \\ &\quad + P(E|do(\neg C), \neg F)P(\neg F)\end{aligned}$$

---

$$P(E|do(C)) < P(E|do(\neg C)),$$

---

---

# Part II:

## Predictions with observational data

---



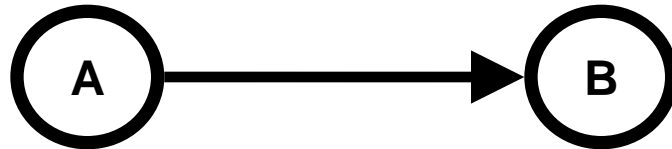
---

# Goals and methods

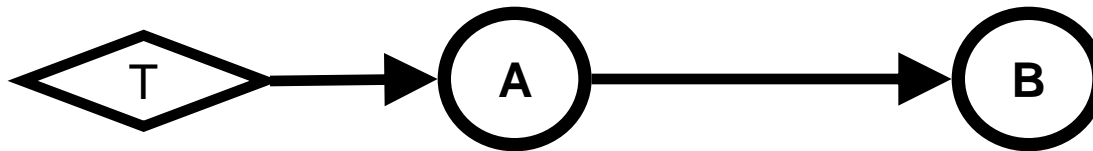
- Given: a causal graph, observational data
- Task: estimate  $P(E \mid \text{do}(C))$
- Approach:
  - Perform a series of modifications on  $P(E \mid \text{do}(C))$ , as allowed by the causal assumptions, until no *do* operators appear
  - Estimate quantity using observational data
  - That is, reduce the causal query to a probabilistic query

# The trivial case

- Graph:



- A representation of a  $do(A)$  intervention



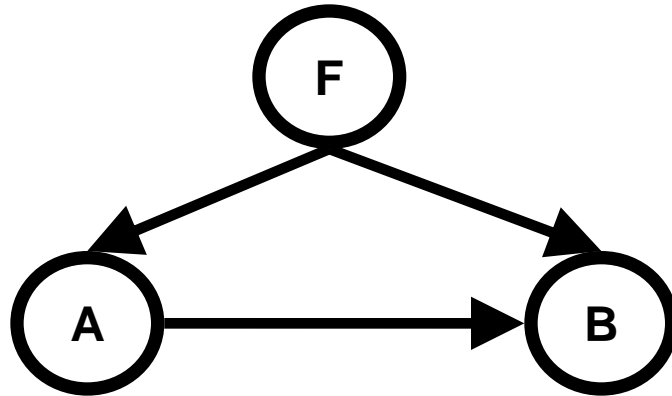
---

# The trivial case

- B is independent of T given A
    - $P(B \mid \text{do}(A)) = P(B \mid A, T) = P(B \mid A)$
  - Term on the right is identifiable from observational data
    - *do-free*
  - That is,  $P(B \mid \text{do}(A))$  can be estimated as  $P(B \mid A)$
-

# A less trivial case

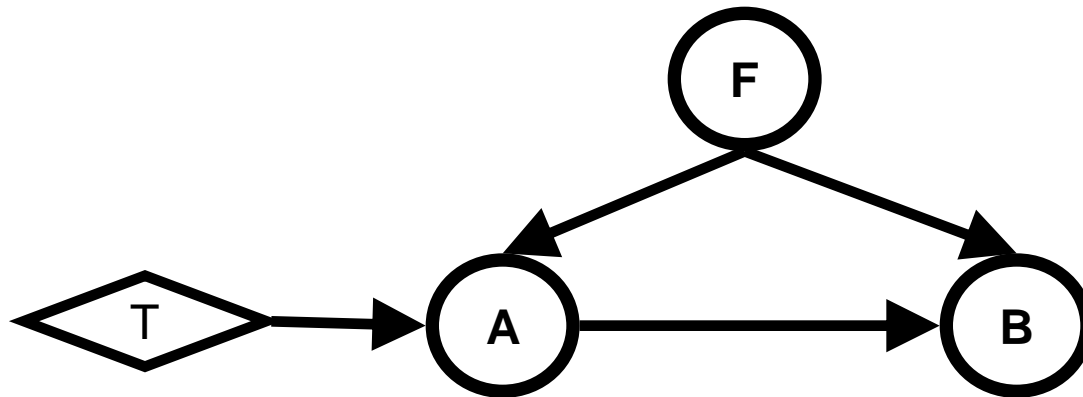
- Knowledge:



- Query:  $P(B \mid \text{do}(A))$

# A less trivial case

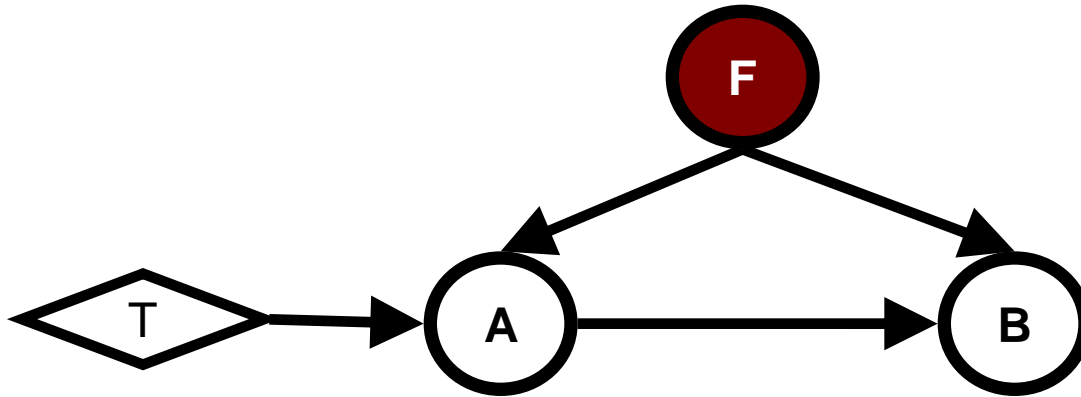
- With intervention



- B and T are not independent given A anymore...

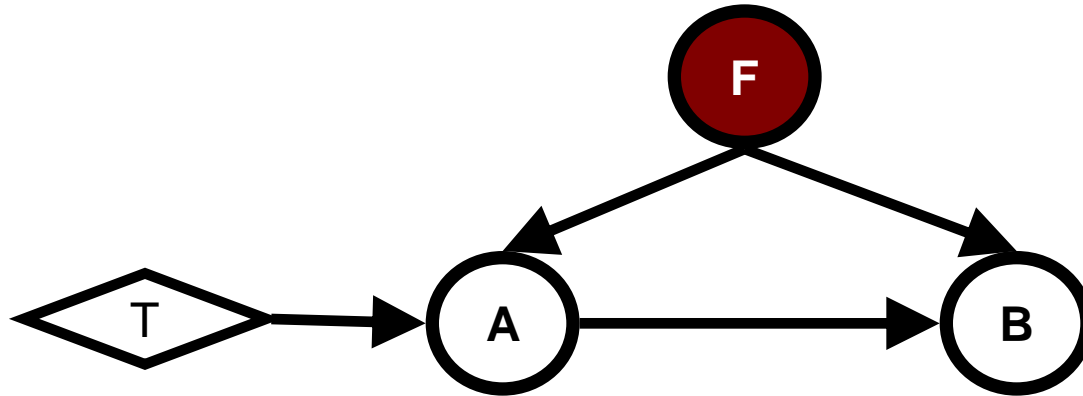
# A less trivial case

- Solution: conditioning



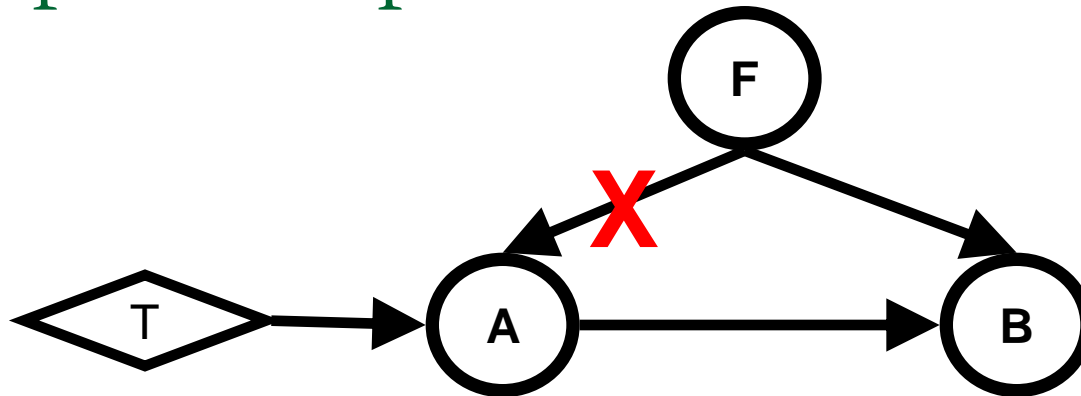
- Now, B is independent of T given A and F

# A less trivial case



$$\begin{aligned} P(B \mid \text{do}(A)) &= \\ &P(B \mid \text{do}(A), F)P(F \mid \text{do}(A)) + \\ &P(B \mid \text{do}(A), \sim F)P(\sim F \mid \text{do}(A)) = \text{“F-independent” intervention} \\ &P(B \mid A, F, T)P(F) + P(B \mid A, \sim F, T)P(\sim F) = \\ &P(B \mid A, F)P(F) + P(B \mid A, \sim F)P(\sim F) \end{aligned}$$

# Simplified operation for independent point interventions

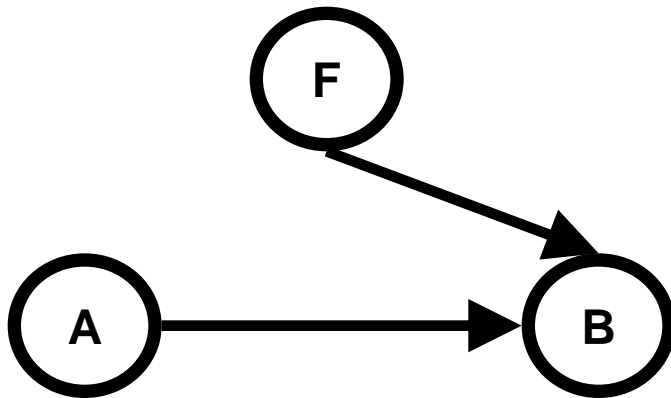


Before intervention:

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

After intervention:

$$\begin{aligned} P(A, B, F | \text{do}(A)) &= P(B | A, F)P(A \text{X} F)P(F) \\ &= P(B | A, F) \delta(A = \text{true})P(F) \end{aligned}$$



A “mechanism substitution” system

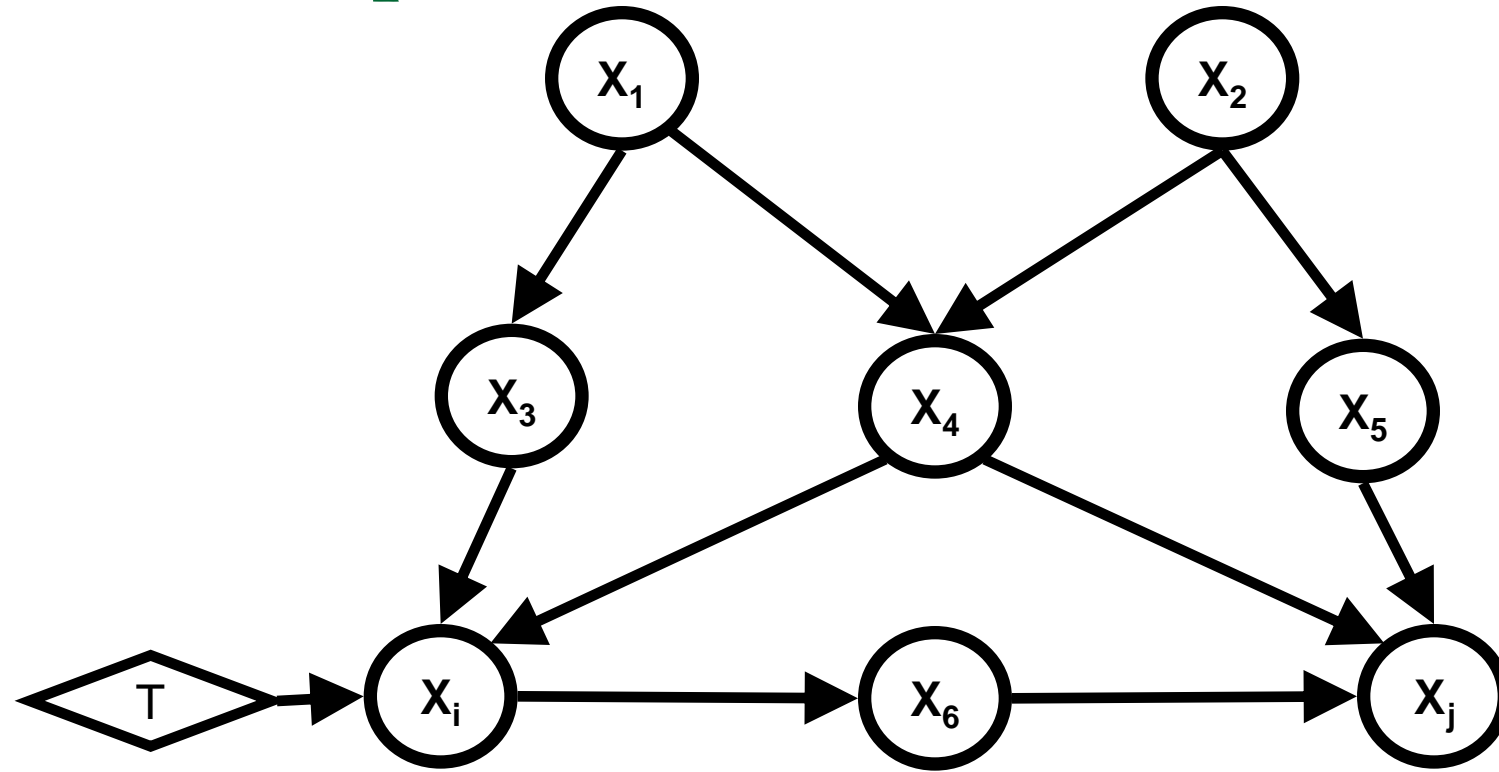


---

# Those “back-doors”...

- Any common ancestor of A and B in the graph is a confounder
  - Confounders originate “back-door” paths that need to be blocked by conditioning
-

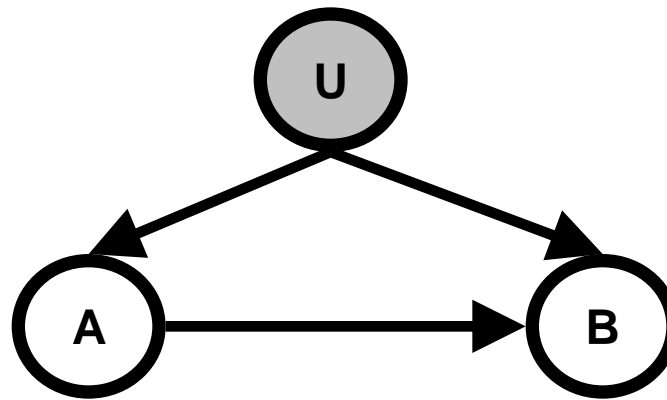
# Example



- In general, one should condition on and marginalize minimal sets, since this reduces statistical variability

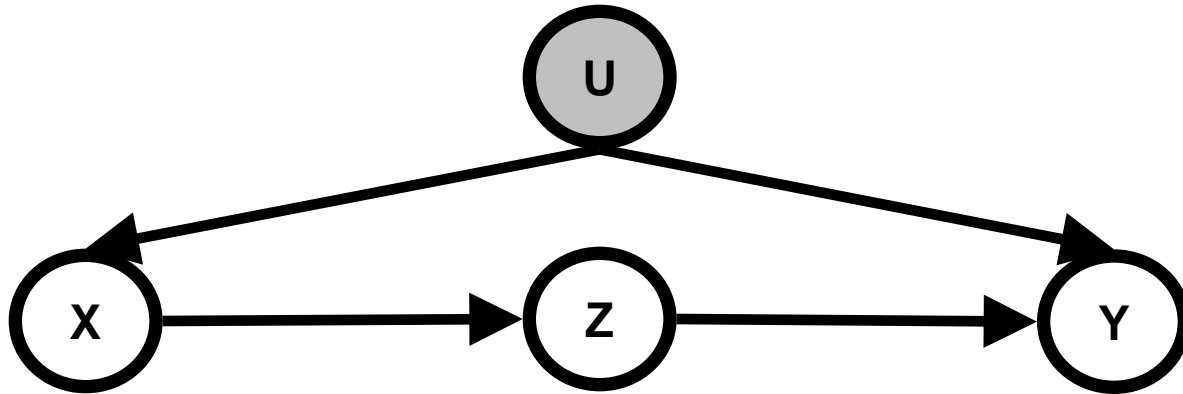
# Unobserved confounding

- If some variables are hidden, then there is no data for conditioning



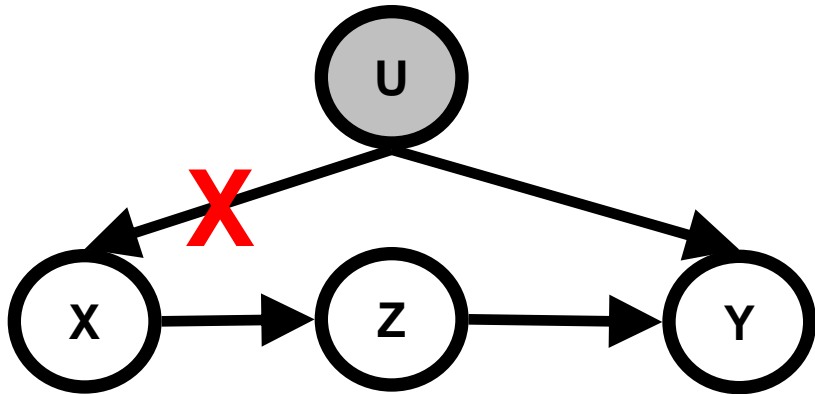
- Ultimately, some questions cannot be answered
  - without extra assumptions
- But there are other methods beside back-door adjustment

# The front-door criterion



- Interestingly enough,  $P(Y \mid \text{do}(X))$  is identifiable in this case
  - Even though we will be conditioning on a variable  $Z$  that is in the causal path!

# The front-door criterion

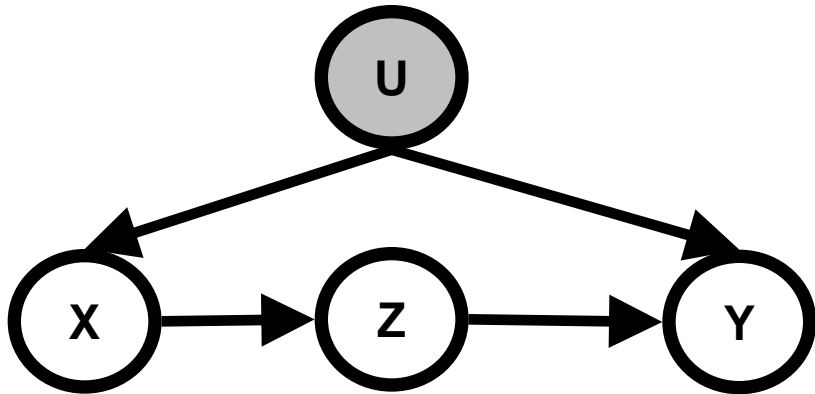


$$P(X, Y, Z, U) = P(U)P(X | U)P(Z | X)P(Y | Z, U)$$

$$P(Y, Z, U | \text{do}(X)) = P(Y | Z, U) P(Z | X)P(U)$$

$$P(Y | \text{do}(X)) = \sum_z P(Z | X) \sum_u P(Y | Z, U)P(U)$$

# The front-door criterion



$$P(U | X) = P(U | Z, X)$$

$$P(Y | Z, U) = P(Y | X, Z, U)$$

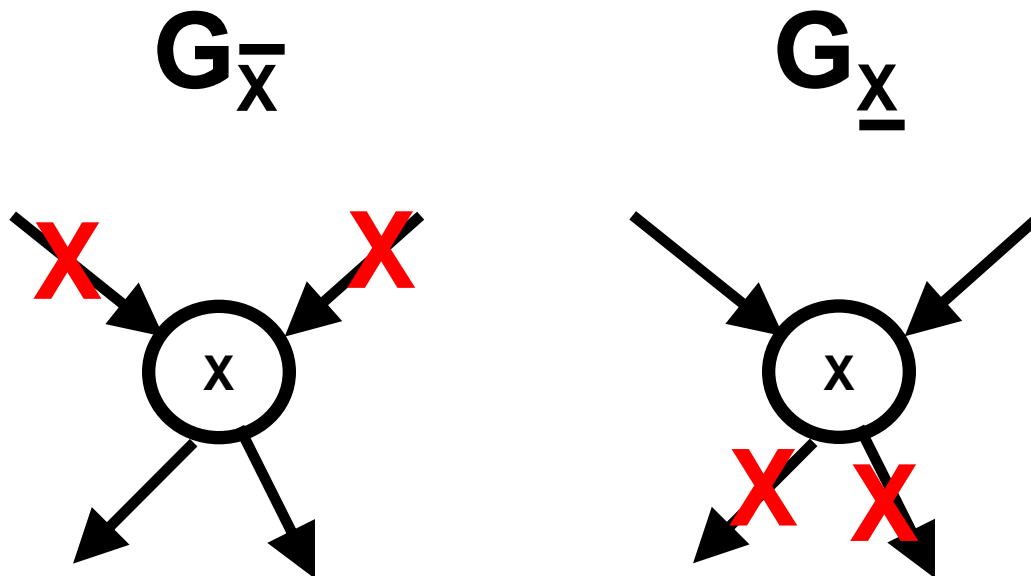
$$\sum_u P(Y | Z, U)P(U) = \sum_x \sum_u P(Y | X, Z, U)P(U | X)P(X)$$

$$= \sum_x \sum_u P(Y | X, Z, U)P(U | X, Z)P(X)$$

$$= \sum_x P(Y | X, Z)P(X) \quad \text{U free!}$$

# A calculus of interventions

- Back-door and front-door criteria combined result in a set of reduction rules
- Notation:



# Examples of *do-calculus* inference rules

- Insertion/deletion of observations:

$$P(Y \mid \text{do}(X), Z, W) = P(Y \mid \text{do}(X), W), \text{ if } (Y \perp\!\!\!\perp Z \mid X, W) \text{ in } \mathbf{G}_{\bar{X}}$$

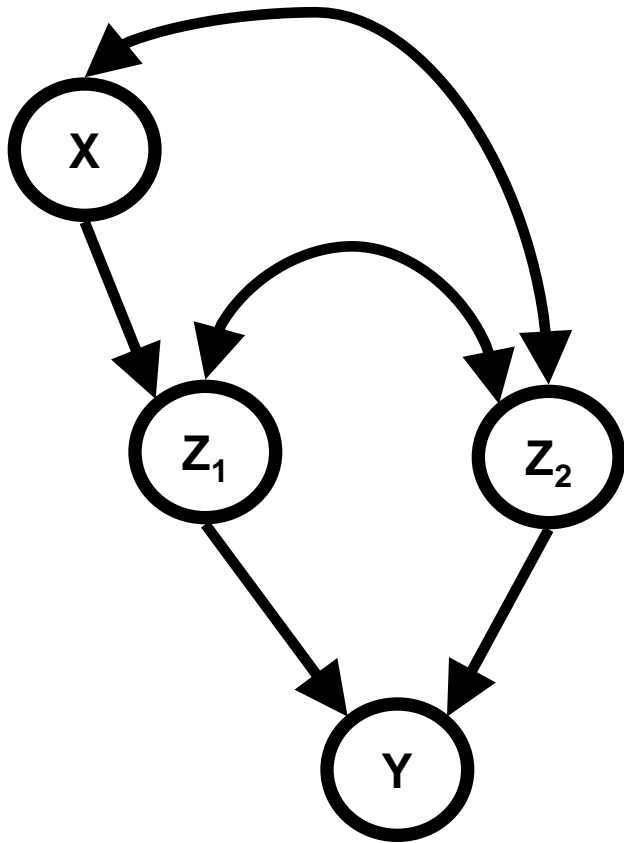
- Action/observation exchange:

$$P(Y \mid \text{do}(X), \text{do}(Z), W) = P(Y \mid \text{do}(X), Z, W), \text{ if } (Y \perp\!\!\!\perp Z \mid X, W) \text{ in } \mathbf{G}_{\bar{X}\bar{Z}}$$

- Sound and complete algorithms that use these rules exist (Huang and Valtorta, 2006)



# A more complex example...



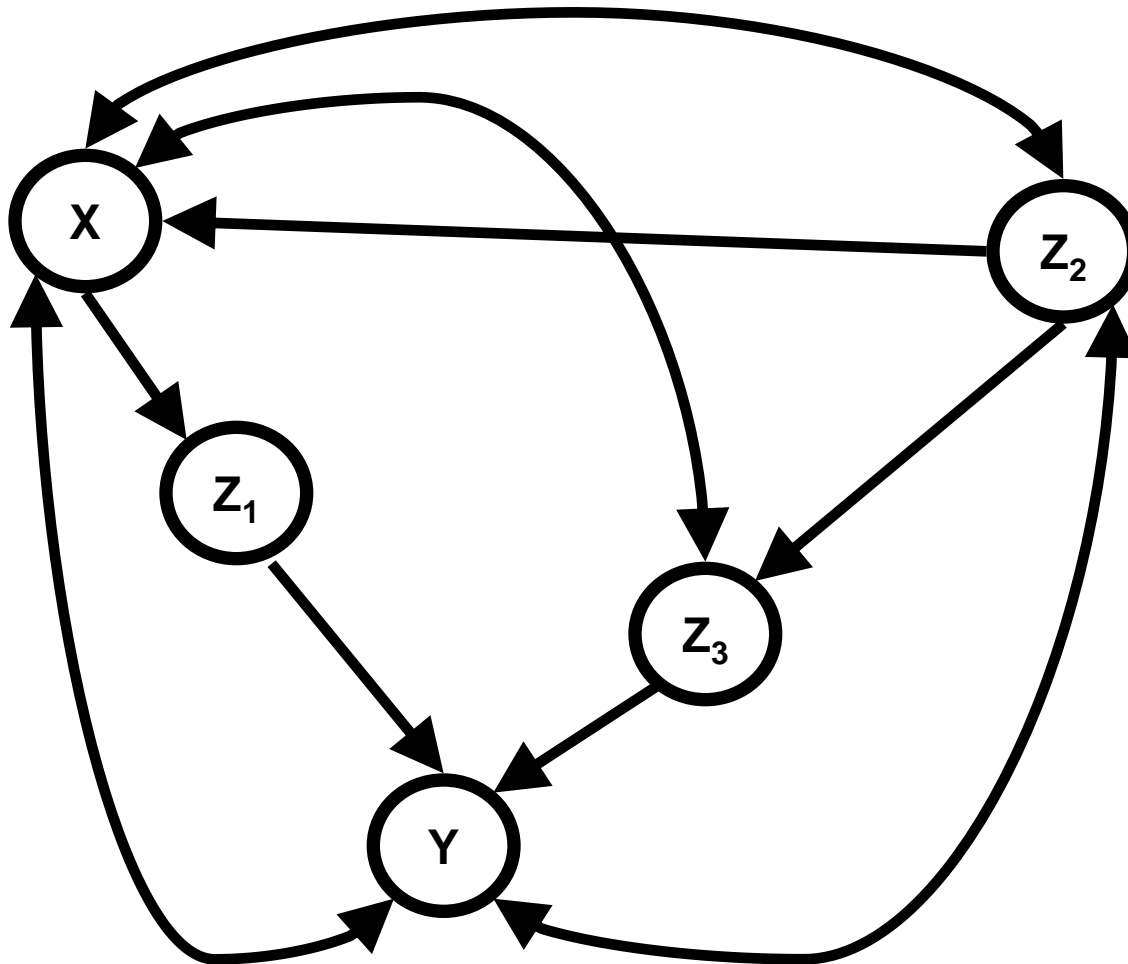
$$P(Y \mid \text{do}(X), \text{do}(Z_2)) = \sum_{z_1} P(Y \mid Z_1, \text{do}(X), \text{do}(Z_2)) P(Z_1 \mid \text{do}(X), \text{do}(Z_2))$$

(Now, Rule 2, for interchanging observation/intervention)

$$= \sum_{z_1} P(Y \mid Z_1, X, Z_2) P(Z_1 \mid X)$$

**Notice:  $P(Y \mid \text{do}(X))$  is NOT identifiable!**

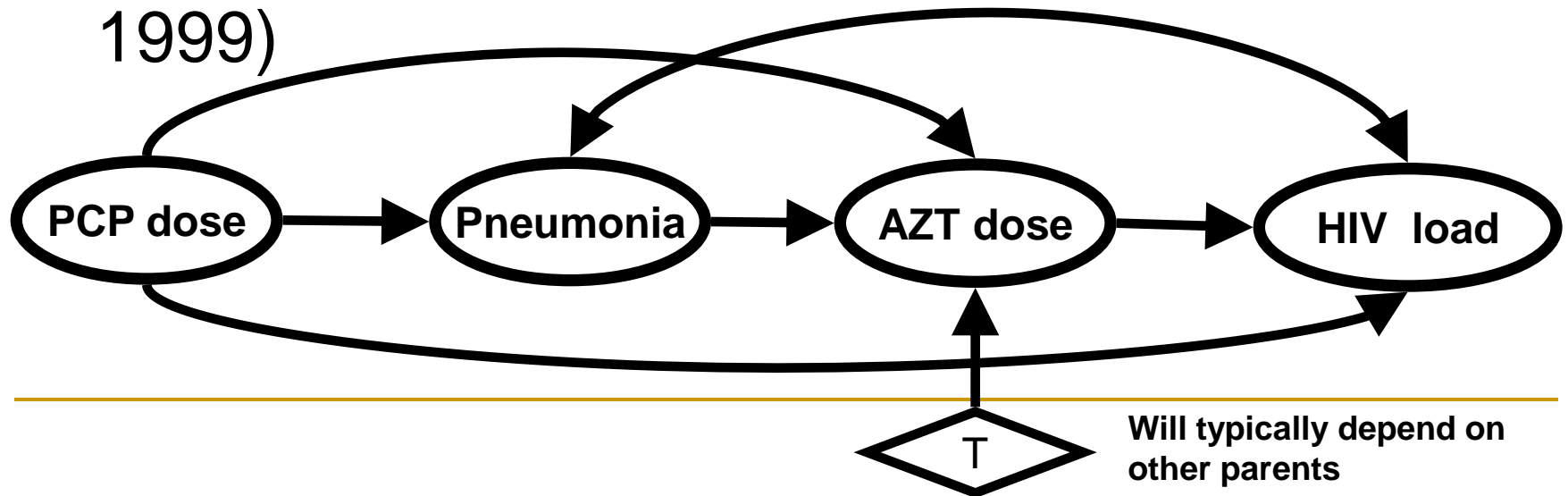
... and even more complex examples



**$P(Y \mid \text{do}(X))$  is identifiable  
(I'll leave it as an exercise)**

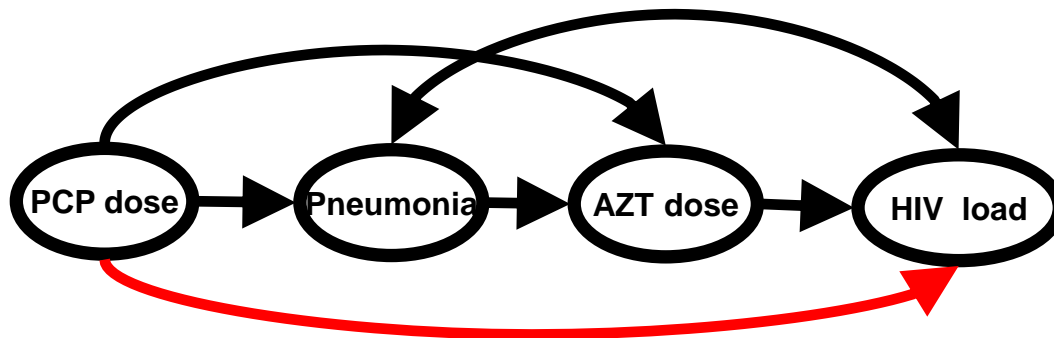
# Planning

- Sequential decision problems:
  - More than one intervention, at different times
  - Intervention at one time depends on previous interventions and outcomes
- Example: sequential AIDS treatment (Robins, 1999)



# Total and direct effects

- A definition of causal effect: ACE
  - $ACE(x, x', Y) = E(Y \mid \text{do}(X = x')) - E(Y \mid \text{do}(X = x))$



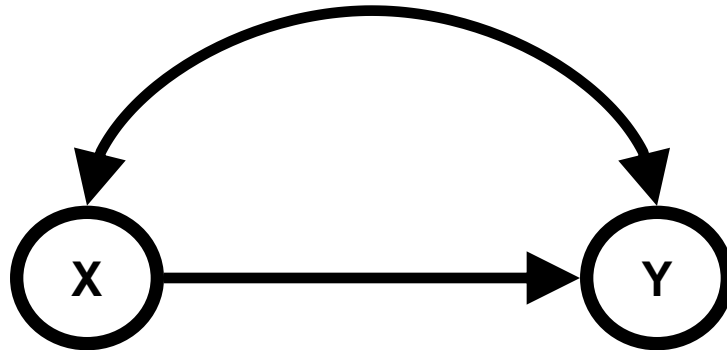
- Controlled direct effects in terms of  $\text{do}(\cdot)$ :
  - $DE_a(\text{pcp}_1, \text{pcp}_2, \text{HIV}) =$   
 $E(\text{HIV} \mid \text{do}(\text{AZT}) = a, \text{do}(\text{PCP} = \text{pcp}_1))$   
 $- E(\text{HIV} \mid \text{do}(\text{AZT}) = a, \text{do}(\text{PCP} = \text{pcp}_2))$

# Standardized and natural direct effects

- Controlling intermediate variables can also be done in a randomized way
  - E.g., controlled according to the age of the patient
- This notion is known as standardized effect
- Natural direct effects:
  - Intermediate variables arise from natural state
  - E.g., adjusting for intermediate psychological effects by using placebos

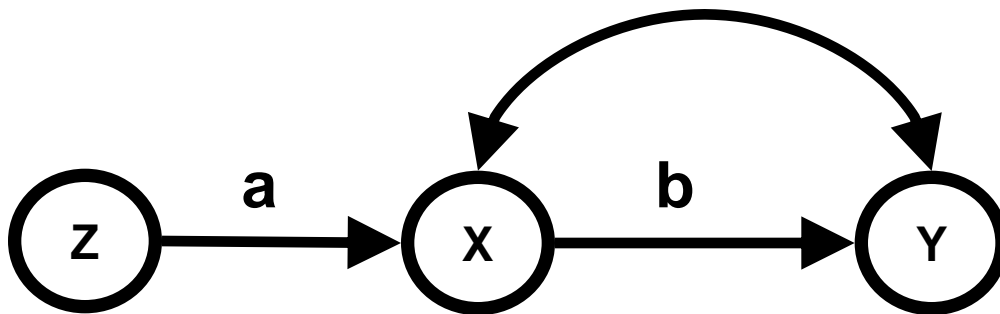
# Dealing with unidentifiability

- We saw techniques that identify causal effects, if possible
- What if it is not possible?
- The dreaded “bow-pattern”:



# Instrumental variables

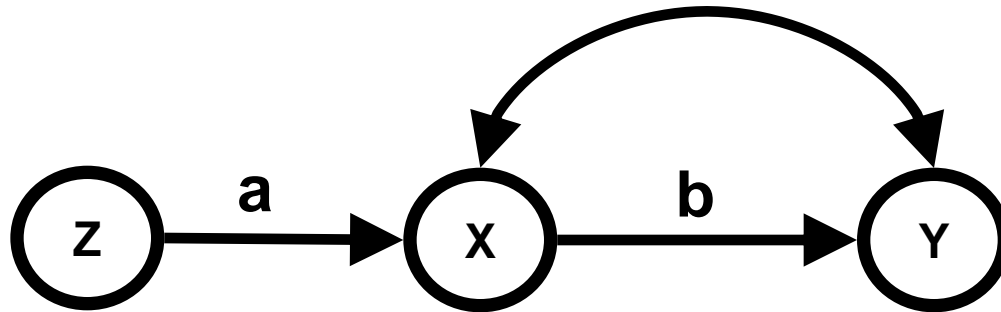
- One solution: explore parametric assumptions and other variables
- Classical case: the linear instrumental variable



$$X = aZ + \varepsilon_X$$
$$Y = bX + \varepsilon_Y$$

$$\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$$

# Instrumental variables



- Let  $Z$  be a standard Gaussian:
  - $\sigma_{YZ} = ab$ ,  $\sigma_{XZ} = a$
  - That is,  $b = \sigma_{YZ} / \sigma_{XZ}$
- Bounds can be generated for non-linear systems
  - Advertising: see my incoming NIPS paper for an example and references



---

# Bayesian analysis of confounding

- Priors over confounding factors
- Buyer Beware: priors have to have a convincing empirical basis
  - not a small issue
- Example: epidemiological studies of occupational hazards
  - Are industrial sand workers more likely to suffer from lung cancer?
    - Since if so, they should receive compensations

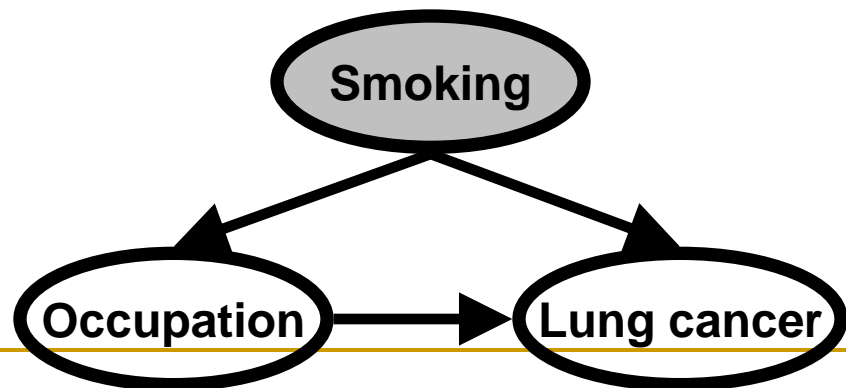
---

# Bayesian analysis of confounding

- Evidence for:
  - Observational evidence of higher proportion of cancer incidence in said population
  - Exposure to silica is likely to damage lungs
- Evidence against:
  - Blue-collar workers tend to smoke more than general population

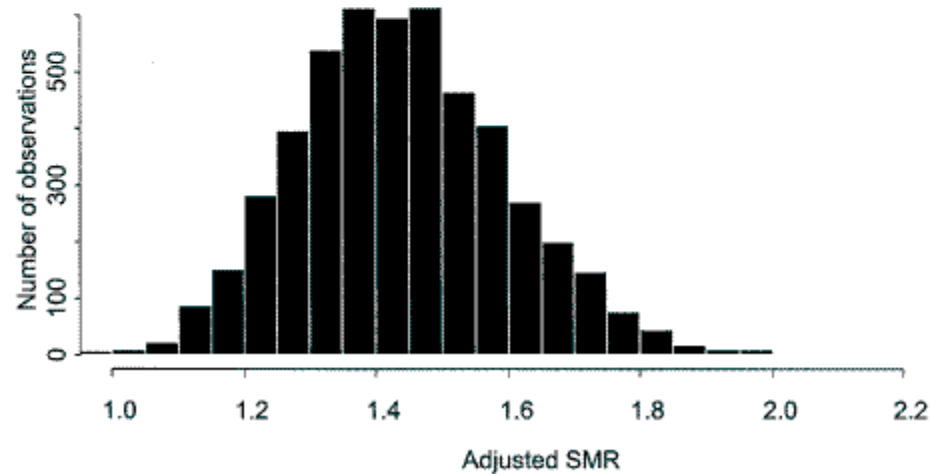
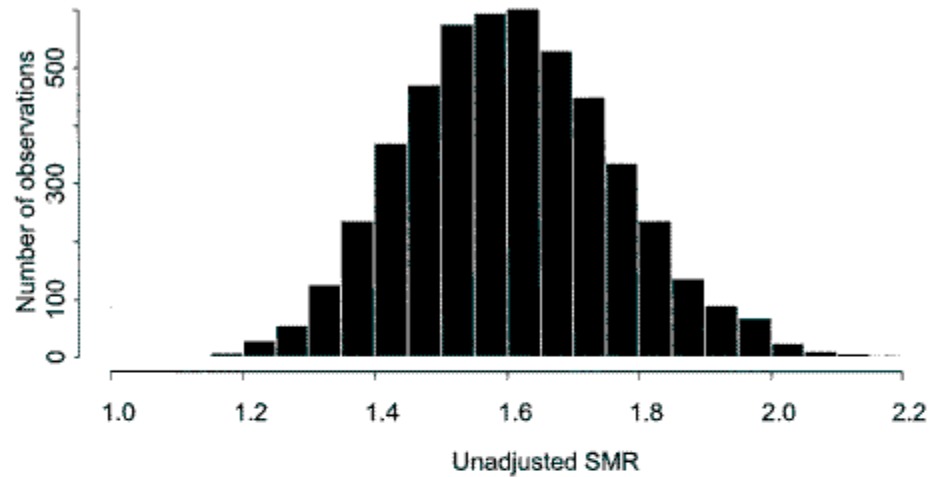
# Quantitative study

- Sample of 4,626 U.S. workers, 1950s-1996
  - Smoking not recorded: becomes unmeasured confounder
  - Prior: empirical priors pulled from population in general
    - Assumes relations between subpopulations are analogous



(Steenland and Greenland, 2004)

# Quantitative study



(Steenland and Greenland, 2004)

---

# Part III: Learning causal structure

---

---

# From association to causation

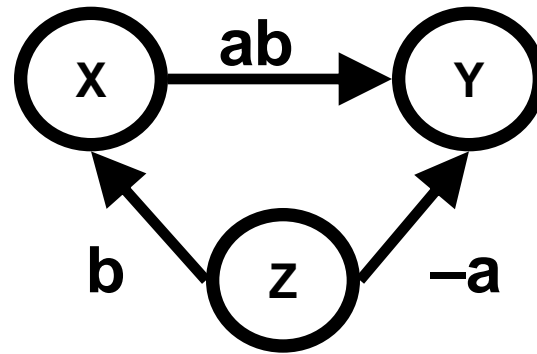
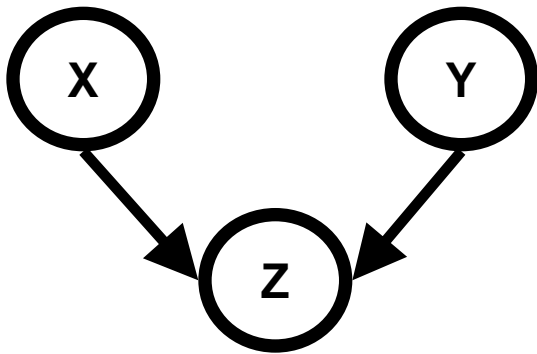
- We require a causal model to compute predictions
  - Where do you get the model?
    - Standard answer: prior knowledge
  - Yet one of the goals is to use observational data
  - Can observational data be used to infer a causal model?
    - or at least parts of it?
-

---

# From association to causation

- This will require going beyond the Causal Markov condition...
  - independence in the causal graph  $\Rightarrow$  independence in probability
- ...into the Faithfulness Condition
  - independence in the causal graph  $\Leftrightarrow$  independence in probability
- Notice: semiparametric constraints also relevant, but not discussed here

# Why do we need the Faithfulness Condition?



$X \perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y$
<del><math>X \perp\!\!\!\perp Y \mid Z</math></del>	<del><math>X \perp\!\!\!\perp Y \mid Z</math></del>

Graph

Distribution

<del><math>X \perp\!\!\!\perp Y</math></del>	$X \perp\!\!\!\perp Y$
<del><math>X \perp\!\!\!\perp Y \mid Z</math></del>	<del><math>X \perp\!\!\!\perp Y \mid Z</math></del>

Graph

Distribution



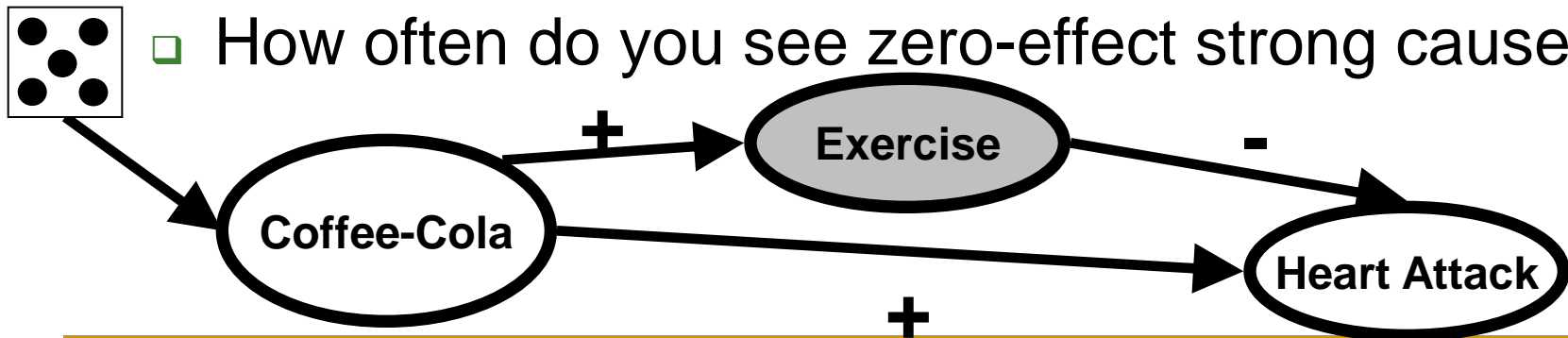
---

# Why would we accept the Faithfulness Condition?

- Many statisticians don't
    - Putting the Radical Empiricist hat: “anything goes”
    - Yet many of these don't see much of a problem with the Causal Markov condition
  
  - But then unfaithful distributions are equivalent to accidental cancellations between paths
    - How likely is that?
-

# Arguments for Faithfulness

- The measure-theoretical argument :
  - probability one in multinomial and Gaussian families (Spirtes et al., 2000)
- The experimental analysis argument:
  - Not spared of faithfulness issues (in a less dramatic sense)
  - How often do you see zero-effect strong causes?



# Arguments against Faithfulness (serious and non-serious ones)

- In practice, one only needs a distribution “close” to unfaithful for things to fail
    - Honest concern: this is possible on any sample size
  - The anti-model argument:
    - “there is no such a thing as independence”
    - but accepting an independence from data is also a matter of prior. There is no such a thing called “prior-free” learning
      - What exactly does “failing to reject a null hypothesis” mean?
      - All models are null hypotheses. Mankind’s knowledge (i.e. model) of the Universe is one big null hypothesis.
  - The Luddite argument:
    - “Never trust a machine to do a man’s job”
    - This is no excuse: competing models are out there and you ought to know of their existence
-

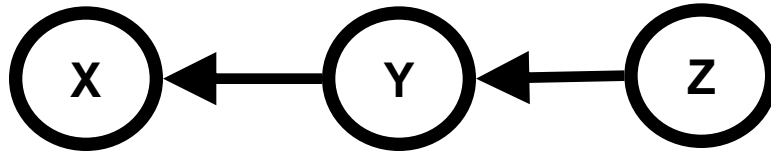
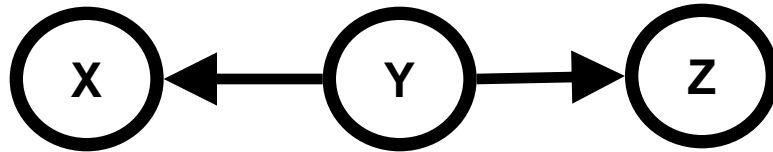
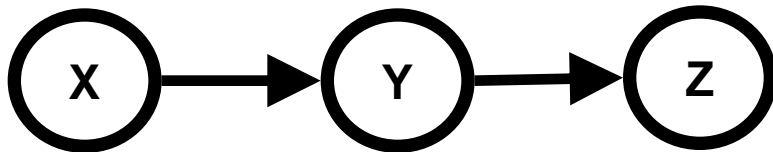
---

# In practice

- There is plenty of justification for deriving what data + faithfulness entail
    - Other models can explain the data. Never trust blindly an “expert” model
      - *Fear of competition for pet-theory can be a hidden reason against “automatic” causality discovery*
    - No reason why use a single model: e.g. sample graphs from posterior
    - No reason to throw skepticism away
    - No reason to forget the GIGO principle
  - Prior knowledge can (and should) always be added
-

# Algorithms: principles

- Markov equivalence classes:
  - Limitations on what can be identifiable with conditional independence constraints



$$X \perp\!\!\!\perp Z \mid Y$$

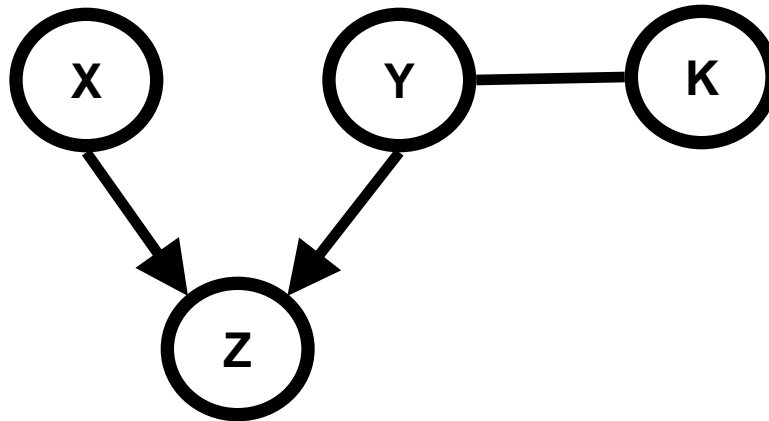
---

# Algorithms: principles

- The goal:
    - Learn a Markov equivalence class
    - Some predictions still identifiable (Spirtes et al., 2000)
    - A few pieces of prior knowledge (e.g., time order) can greatly improve identifiability results
    - Provides a roadmap for experimental analysis
    - Side note: Markov equivalence class is not the only one
-

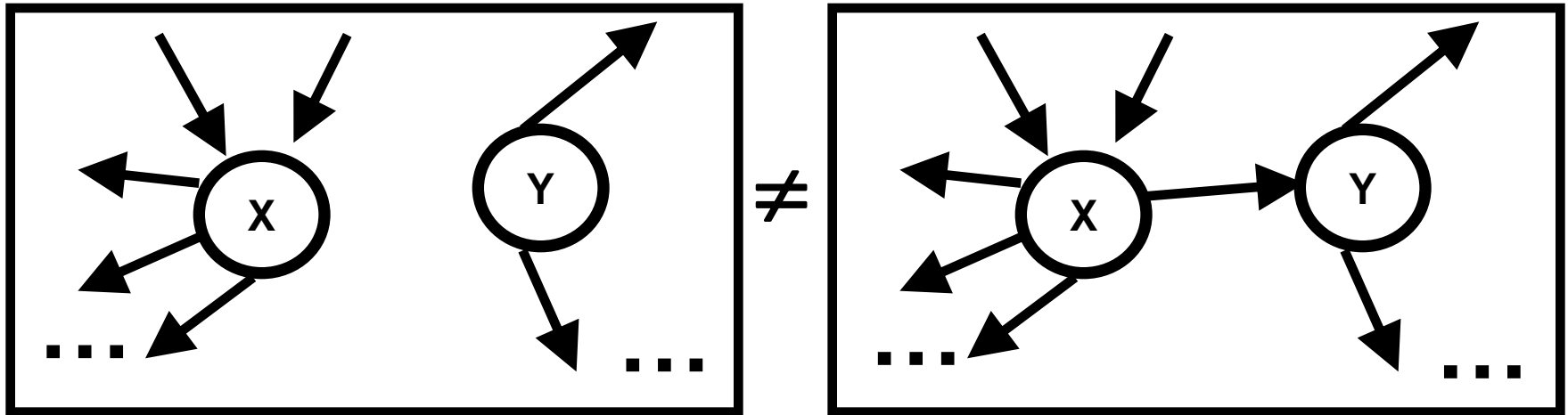
# Initial case: no hidden common causes

- Little motivation for that, but easier to explain
- “Pattern”: a graphical representation of equivalence classes



# More on equivalence classes

- Adjacencies are always the same in all members of a Markov equivalence class

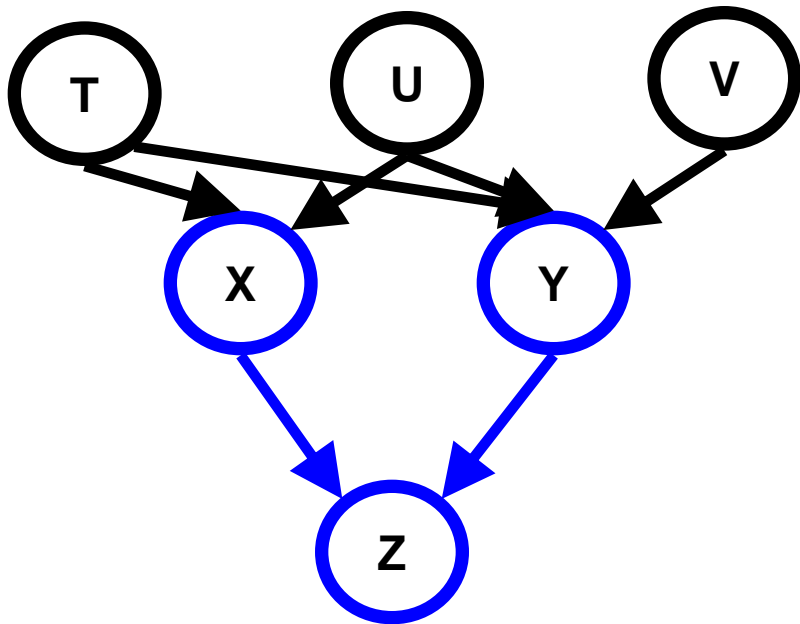


**Never equivalent, since on the left we have  
 $X \perp\!\!\!\perp Y \mid \text{some set } S$**

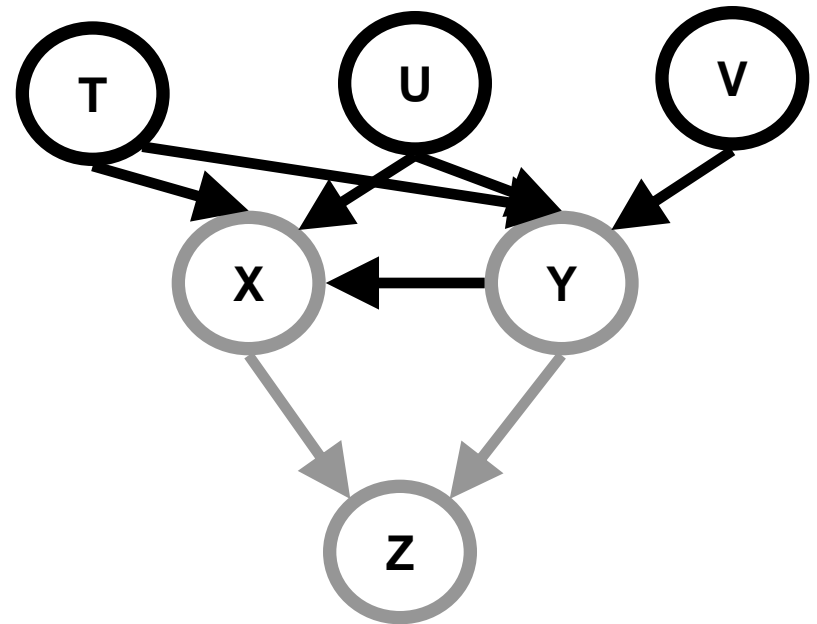


# More on equivalence classes

- Unshielded colliders: always identifiable



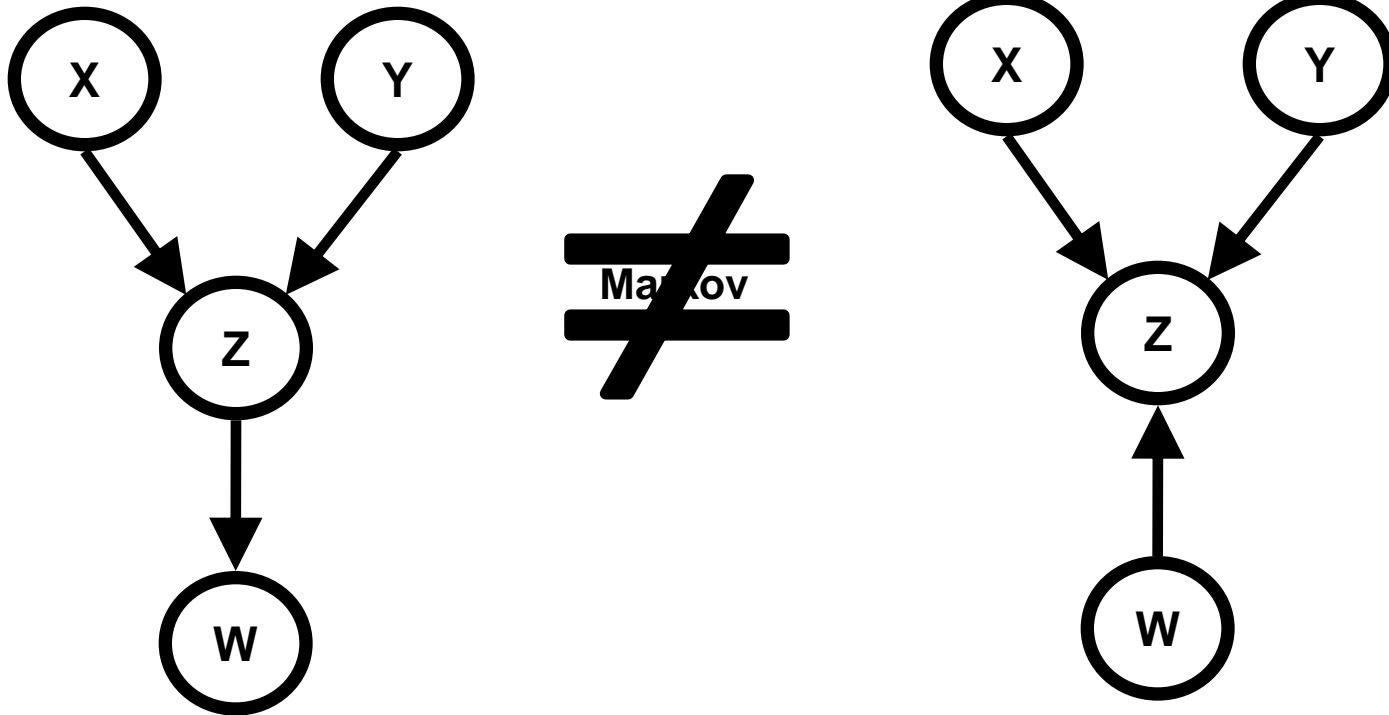
**Unshielded collider**



**Not a unshielded collider**

# More on equivalence classes

- “Propagating” unshielded colliders



**Why? Different unshielded colliders**

---

# Algorithms: two main families

- Piecewise (constraint-satisfaction) algorithms
    - Evaluate each conditional independence statement individually, put pieces together
  - Global (score-based) algorithms
    - Evaluate “all” models that entail different conditional independencies, pick the “best”
      - “Best” in a statistical sense
      - “All” in a computationally convenient sense
  - Two endpoints of a same continuum
-

---

# A constraint-satisfaction algorithm: the PC algorithm

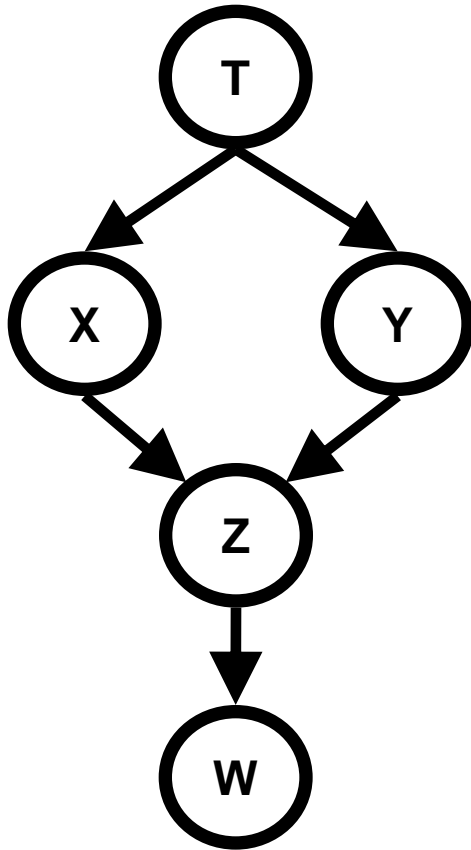
- Start by testing marginal independencies
    - Is  $X_1$  independent of  $X_2$ ?
    - Is  $X_1$  independent of  $X_3$ ?
    - ...
    - Is  $X_{N-1}$  independent of  $X_N$ ?
  - Such tests are usually frequentist hypothesis tests of independence
    - Not essential: could be Bayes factors too
-

---

# The PC algorithm

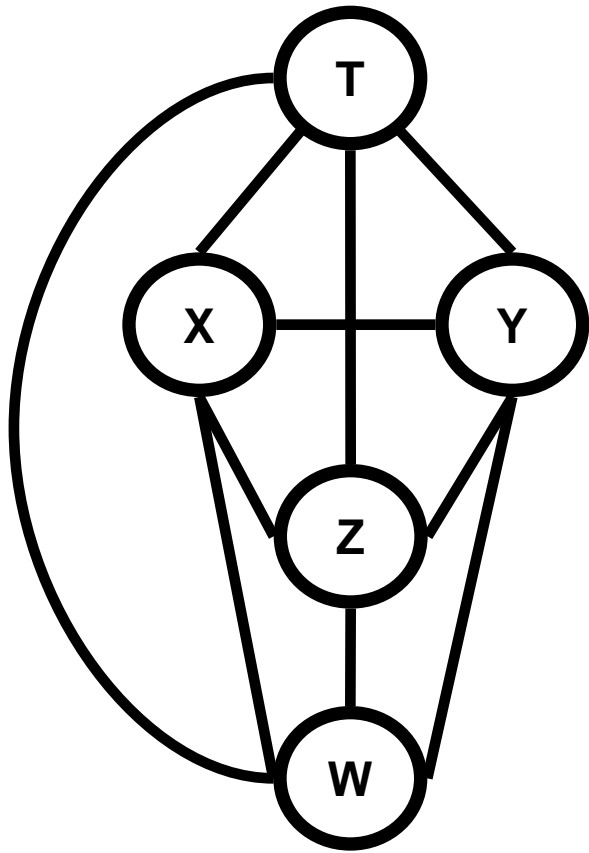
- Next step: conditional independencies tests of “size” 1
    - Is  $X_1$  independent of  $X_2$  given  $X_3$ ?
    - Is  $X_1$  independent of  $X_2$  given  $X_4$ ?
    - ...
    - (In practice only a few of these tests are performed, as we will illustrate)
  - Continue then with tests of size 2, 3, ... etc. until no tests of a given size pass
  - Orient edges according to which tests passed
-

# The PC algorithm: illustration

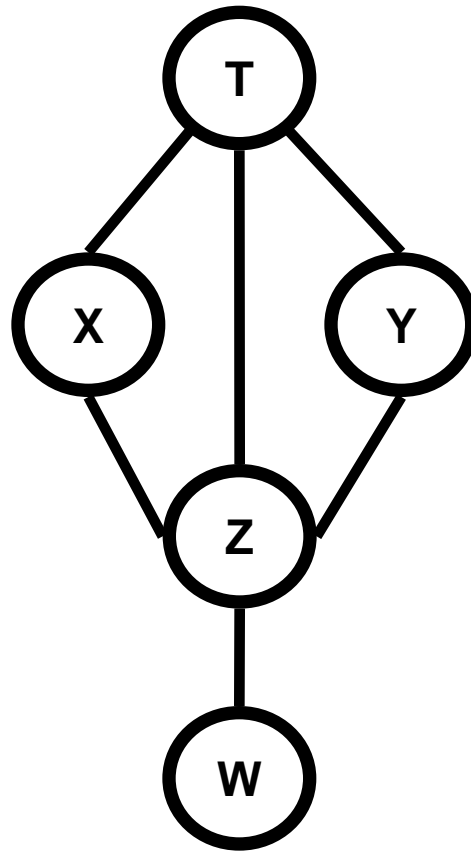


- Assume the model on the left is the real model
- Observable: samples from the observational distribution
- Goal: recover the pattern (equivalence class representation)

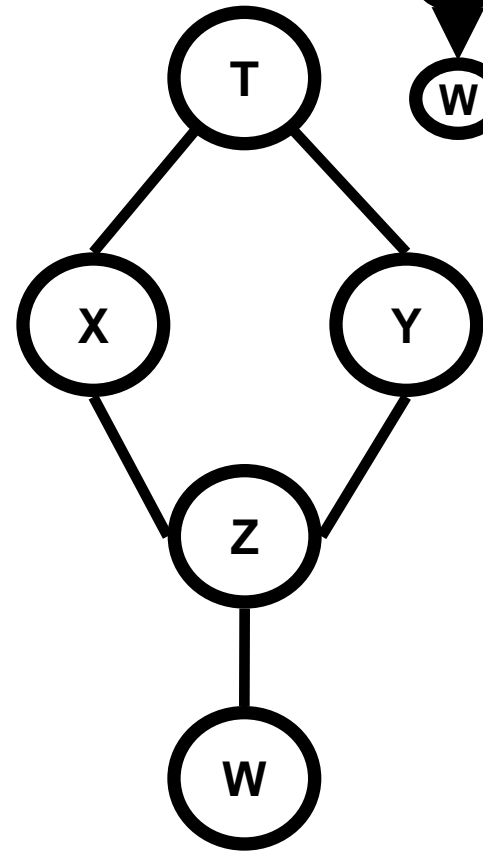
# PC, Step 1: find adjacencies



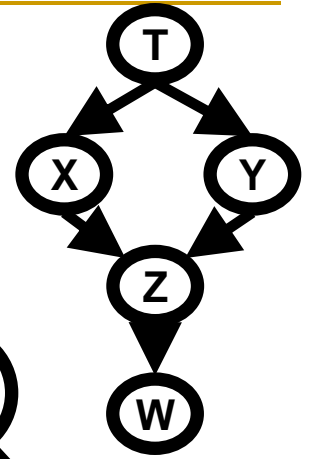
**Start**



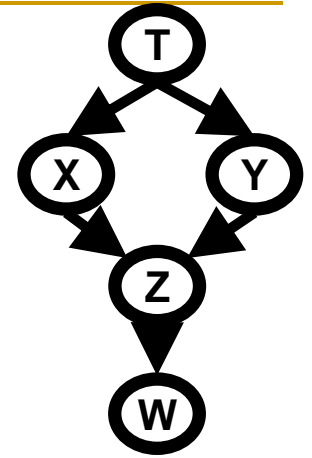
**Size 1**



**Size 2**



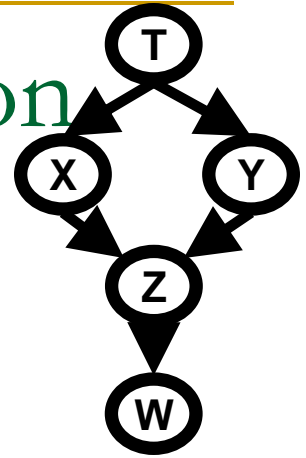
# PC, Step 2: collider orientation



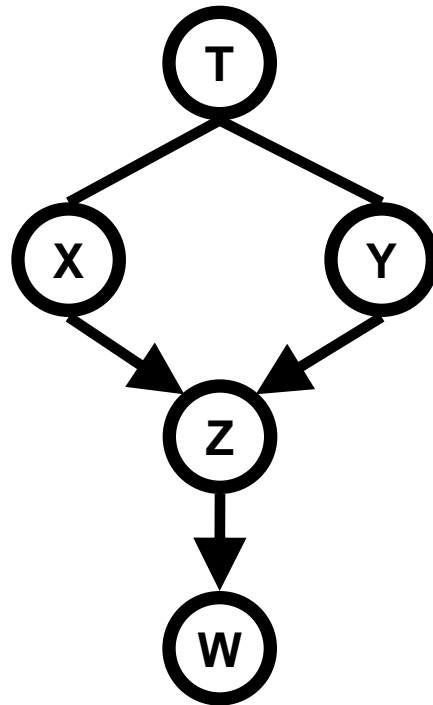
- X and Y are independent given T
  - Therefore,  $X \rightarrow T \leftarrow Y$  is not possible
  - At the same time,
    - $X \leftarrow Z \leftarrow Y$
    - $X \rightarrow Z \rightarrow Y$
    - $X \leftarrow Z \rightarrow Y$are not possible, or otherwise X and Y would not be independent given T
  - Therefore, it has to be the case that  $X \rightarrow Z \leftarrow Y$
- Check all unshielded triples



# PC, Step 3: orientation propagation



- Since  $X \rightarrow Z \text{ --- } W$  is not a collider, only option left is  $X \rightarrow Z \rightarrow W$
- Pattern:



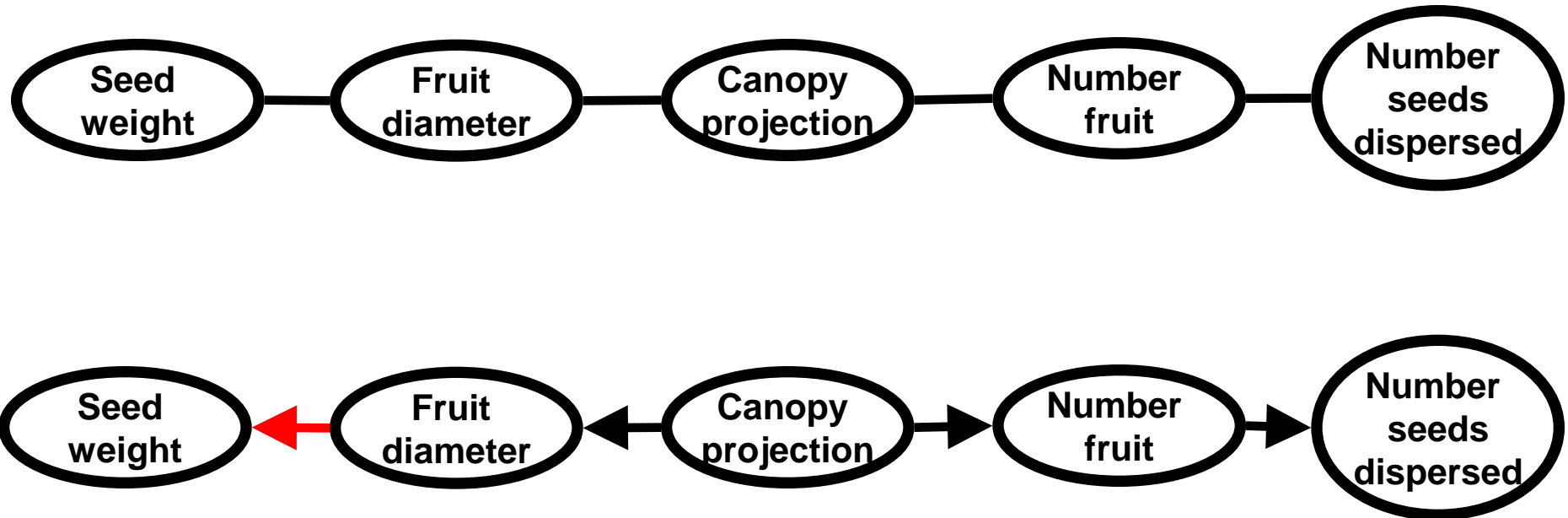
---

# Advantages and shortcomings

- **Fast**
    - Only submodels are compared
    - Prunes search space very effectively
  - **Consistent**
    - On the limit on infinite data
  - **But brittle**
    - Only submodels are compared: very prone to statistical mistakes
    - Doesn't enforce global constraint of acyclicity
      - Might generate graphs with cycles
        - (which is actually good and bad)
-

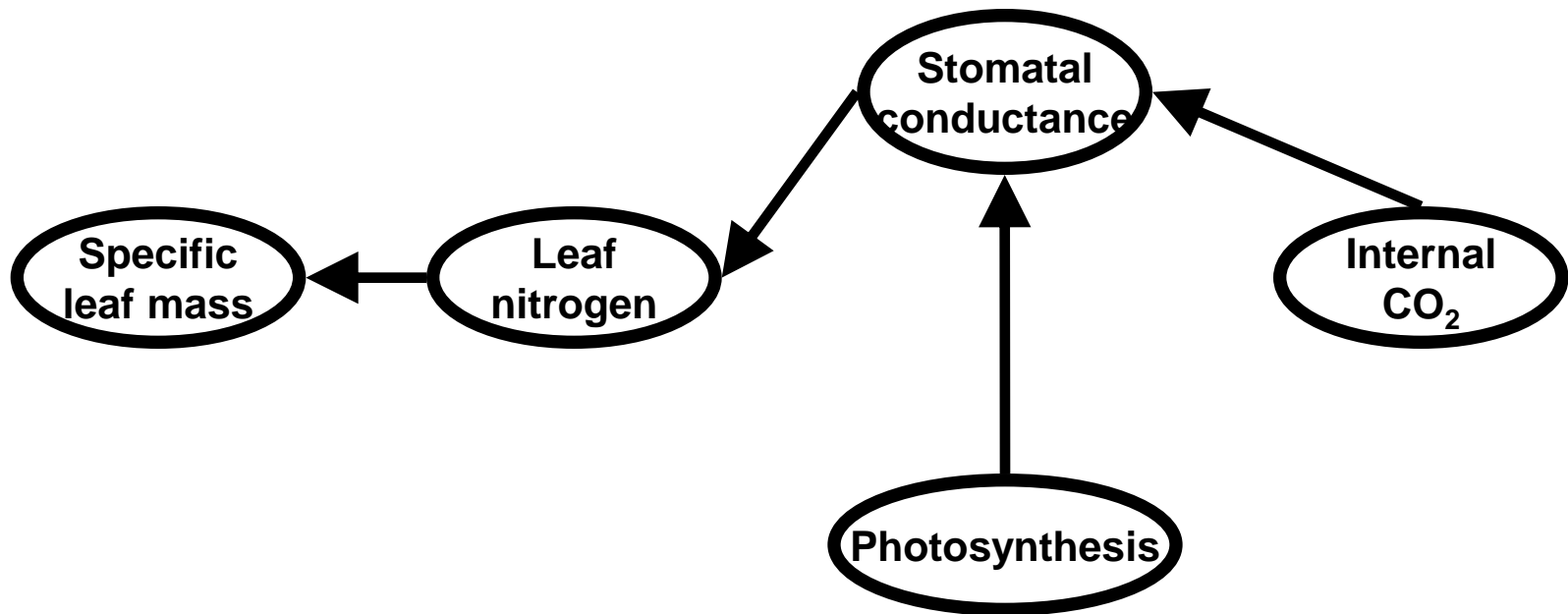
# Simple application: evolutionary biology

- Using a variation of PC + bootstrapping in biological domain:



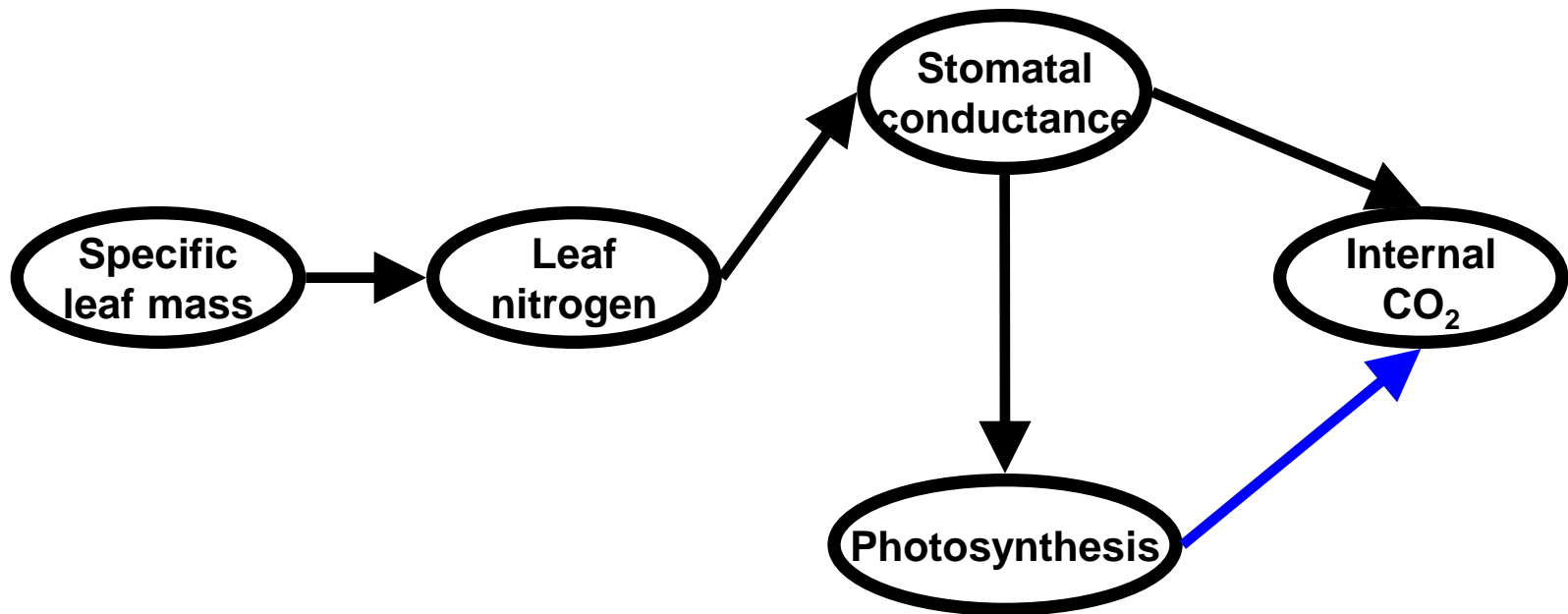
# Simple application: botanic

- Very small sample size (35):



# Simple application: botanic

- Forcing blue edge by background knowledge

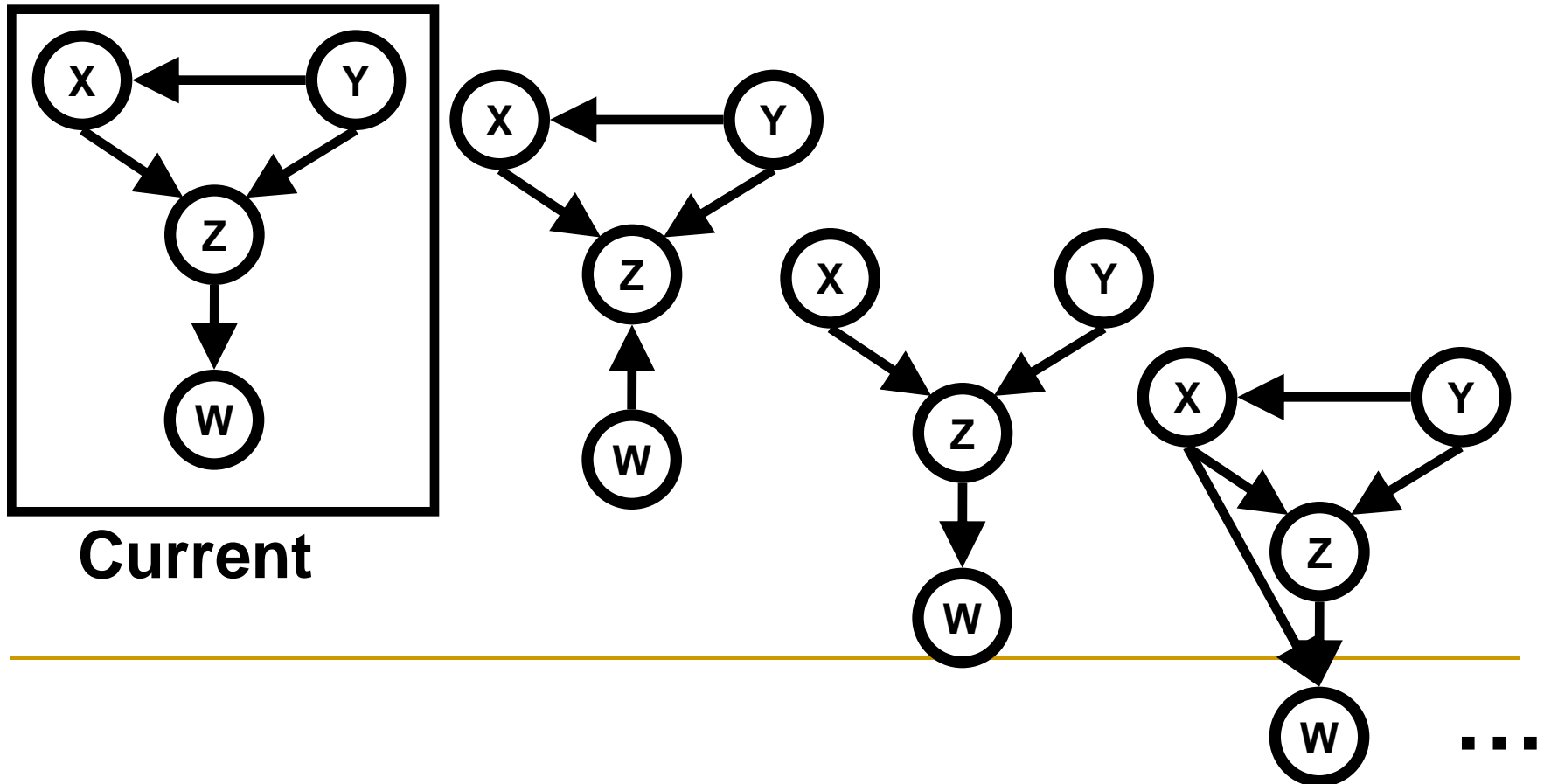


# Global methods for structure learning

- Compares whole graphs against whole graphs
- Typical comparison criterion (*score function*): posterior distribution
  - $P(G_1 \mid \text{Data}) > P(G_2 \mid \text{Data})$ , or the opposite?
- Classical algorithms: greedy search
  - Compares nested models: one model differs from the other by an adjacency
  - Some algorithms search over DAGs, others over patterns

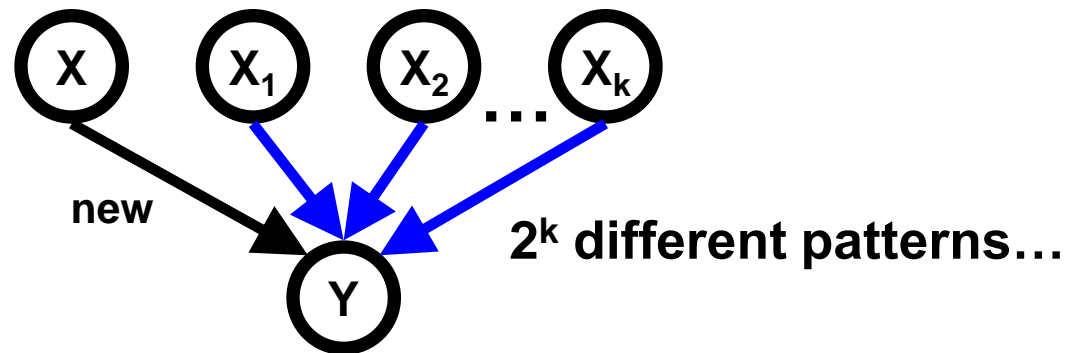
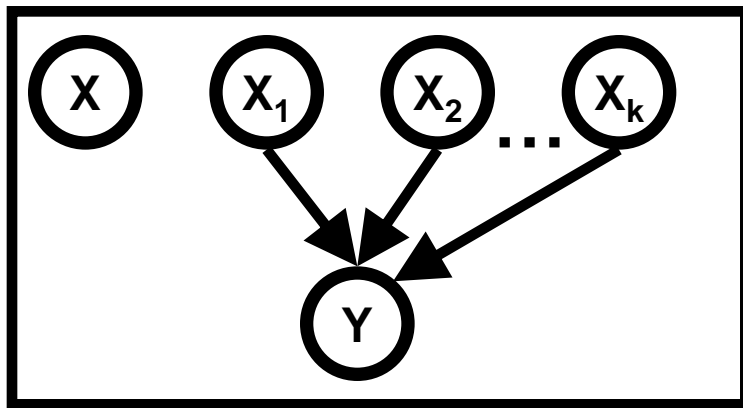
# Greedy search over DAGs

- From the current point, evaluate all edge insertions, deletions and reversals



# Greedy search over patterns

- Evaluate all patterns that differ by one adjacency from the current one
- Unlike DAG-search, consistent (starting point doesn't matter)
- But the problem is NP-hard...





# Combining observational and experimental data

- Model selection scores are usually decomposable:
  - Remember DAG factorization:

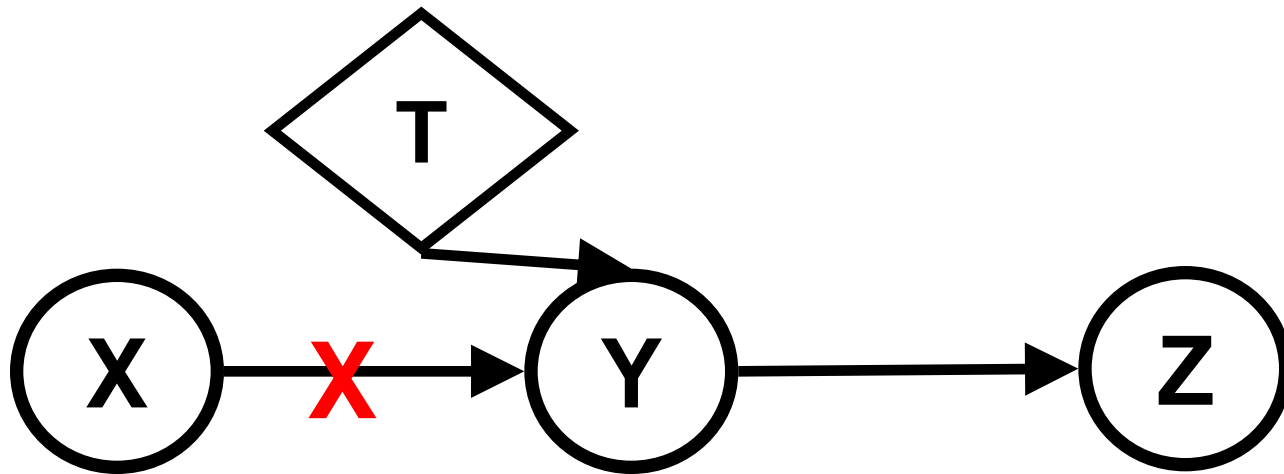
$$\prod_i P(X_i \mid \text{Parents}(X_i))$$

- Score factorization (such as log-posterior):

$$\text{Score}(\mathbf{G}) = \sum_i \mathbf{S}(X_i, \text{Parents}(X_i))$$

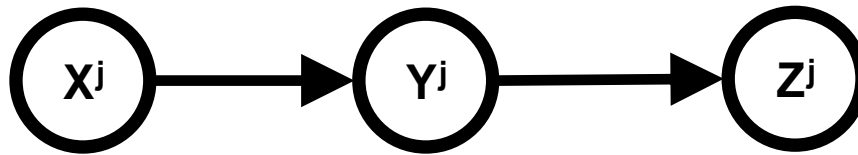
# Combining observational and experimental data

- Experimental data follows from a local probability substitution
- Apply the “mechanism substitution” principle:



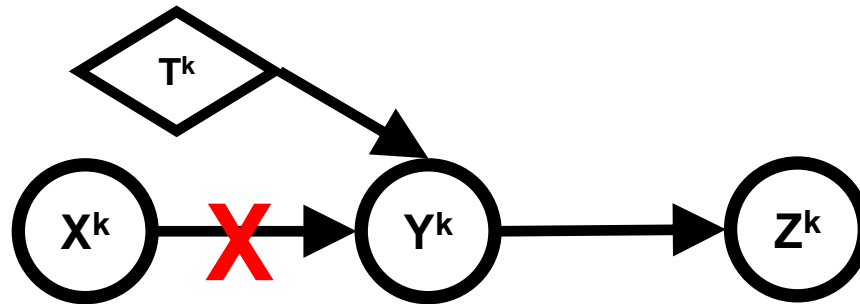
# Combining observational and experimental data

- For data point  $j$ , natural state:



$$\text{Score}(G; j) = \log P(X^j) + \log P(Y^j | X^j) + \log P(Z^j | Y^j)$$

- For data point  $k$ , random intervention on  $Y$



$$\text{Score}(G; k) = \log P(X^k) + \log P(Y^k | T^k) + \log P(Z^k | Y^k)$$

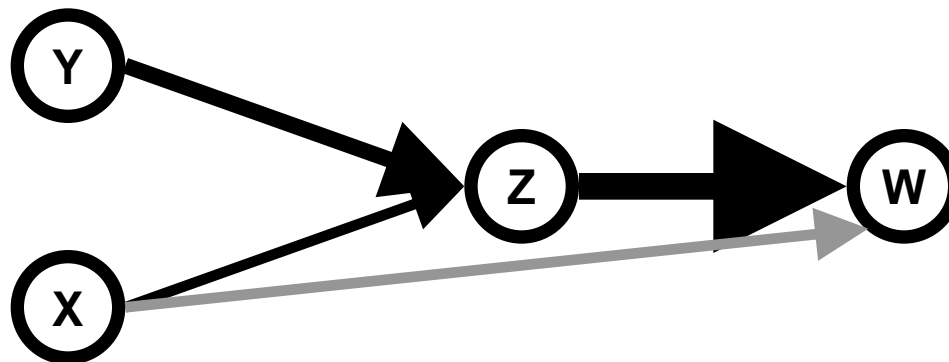
e.g.,  $\text{Score}(G; k) = \log P(X^k) + \log 1/2 + \log P(Z^k | Y^k)$

# Computing structure posteriors

- Notice: greedy algorithms typically return the maximum a posteriori (MAP) graph
  - Or some local maxima of the posterior
- Posterior distributions
  - Practical impossibility for whole graphs
    - MCMC methods should be seen as stochastic search methods, mixing by the end of the universe
    - Still: 2 graphs are more useful than 1
  - Doable for (really) small subgraphs: edges, short paths (Friedman and Koller, 2000)

# Computing structure posteriors: a practical approach

- Generate a few high probability graphs
  - E.g.: use (stochastic) beam-search instead of greedy search
- Compute and plot marginal edge posteriors



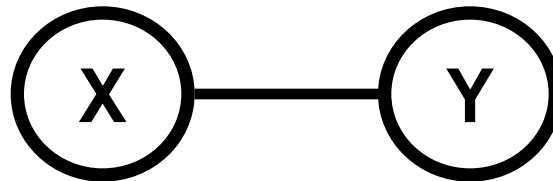
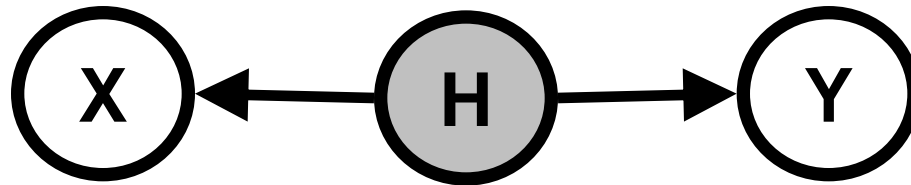
---

# A word of warning

- Uniform consistency: impossible with faithfulness only (Robins et al., 2003)
    - Considering the case with unmeasured confounding
  - Rigorously speaking, standard Bayesian posteriors reflect independence models, not causal models
  - There is an implicit assumption that the distribution is not “close” to unfaithfulness
    - A lot of work has yet to be done to formalize this (Zhang and Spirtes, 2003)
-

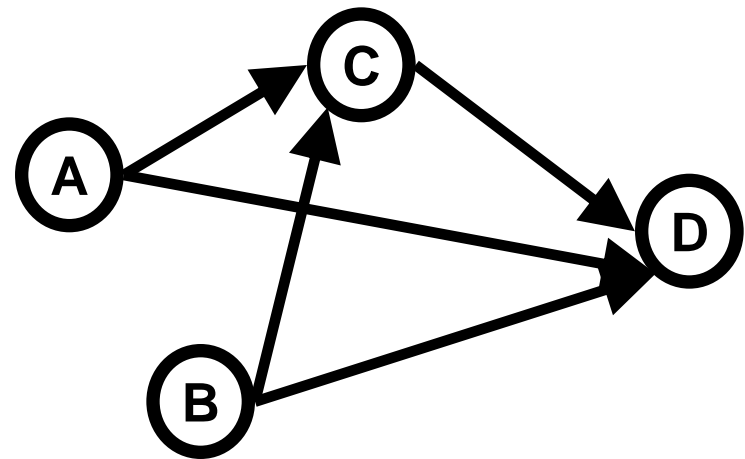
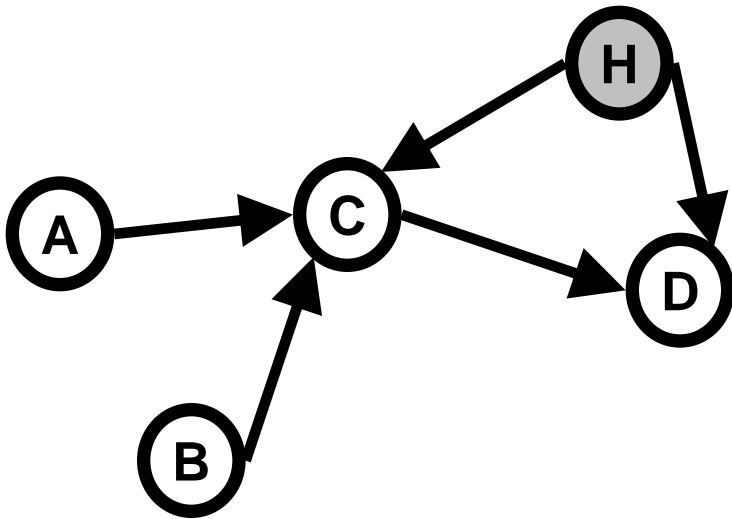
# Methods robust to hidden common causes

- What happens to these algorithms when there are hidden common causes?



# Methods robust to hidden common causes

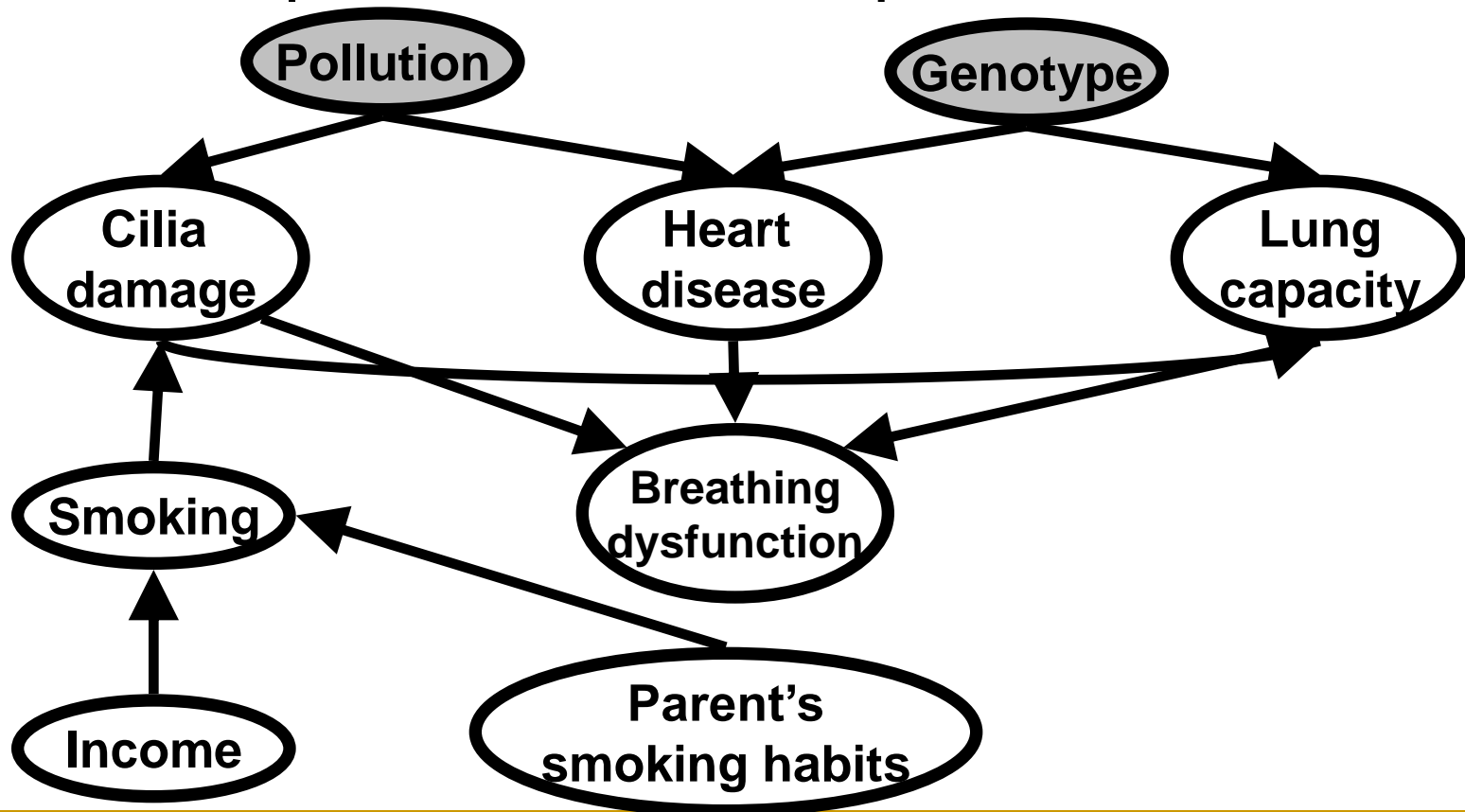
- Even if directionality is correct:
  - they don't tell you correct direct effects
  - which directions are unconfounded





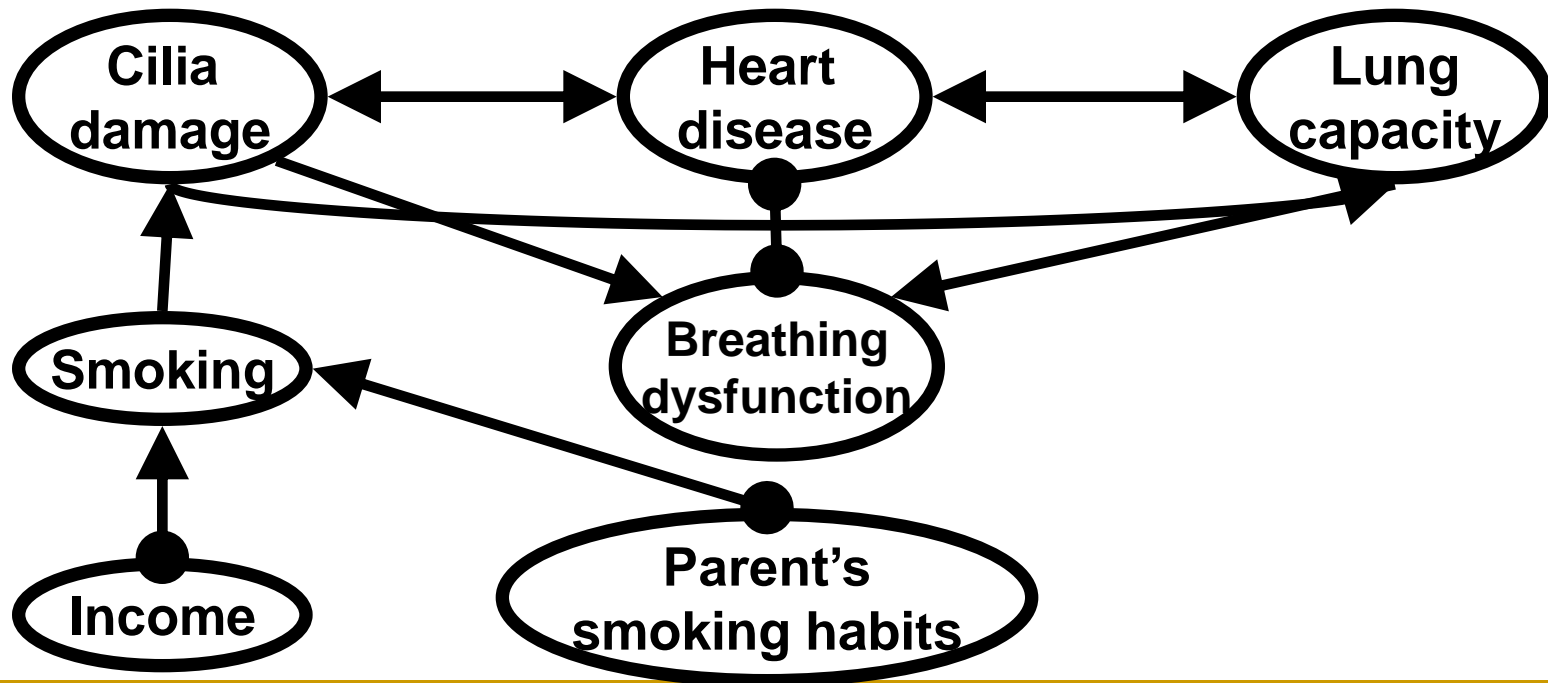
# Partial ancestral graphs (PAGs)

- New representation of equivalence classes



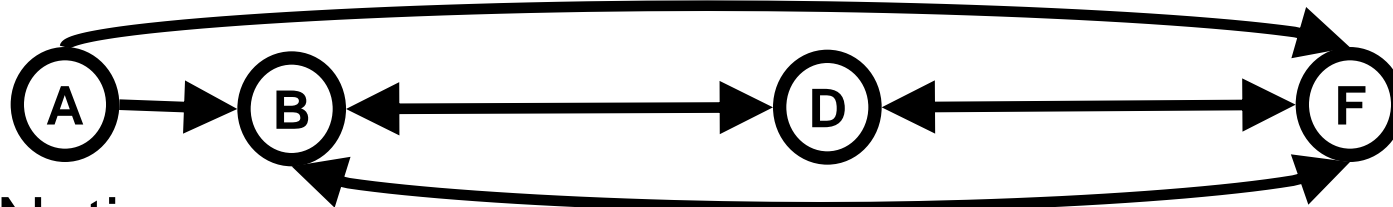
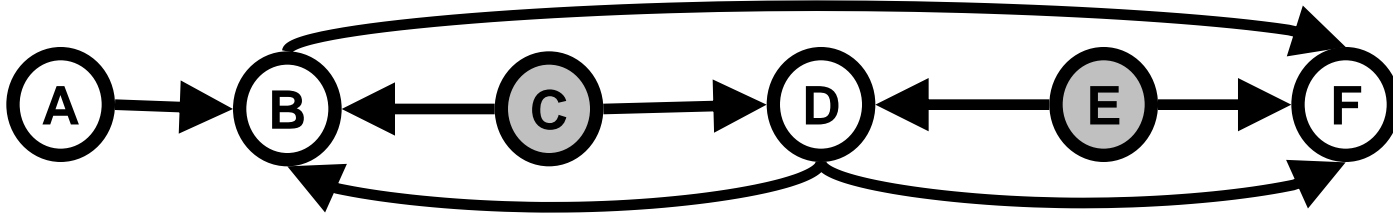
# Partial ancestral graphs (PAGs)

- Type of edge:  $\longrightarrow$   $\longleftrightarrow$   $\bullet \longrightarrow$   $\bullet \text{---} \bullet$



# Discovery algorithms

- Discovers and partially orients *inducing paths*:
  - Sequences of edges between nodes that can't be blocked



- Notice
  - Can't tell if A is a direct or indirect cause of F
  - Can't tell if B is a cause of F

---

# Algorithms

- The “Fast” Causal Inference algorithm (FCI, Spirtes et al., 2000):
    - “Fast” because it has a clever way of avoiding exhaustive search (e.g., as in Pearl, 2000)
  - Sound *and* complete algorithms are fairly recent: Zhang, 2005
  - Bayesian algorithms are largely underdeveloped
    - Discrete model parameterization still a challenge
-

---

# Conclusion

---

---

# Summary and other practical issues

- There is no magic:
    - It's assumptions + data + inference systems
    - Emphasis on assumptions
  - Still not many empirical studies
    - Requires expertise, ultimately requires experiments for validation
    - Lots of work in fixed back-door designs
      - Graphical models not that useful (more so in longitudinal studies)
-

---

# The future

- Biological systems might be a great domain
    - That's how it all started after all (Wright, 1921)
    - High-dimensional: make default back-door adjustments dull
      - Lots of direct and indirect effects of interest
    - Domains of testable assumptions
      - Observational studies with graphical models can be a great aid for experimental design
    - But beware of all sampling issues: measurement error, small samples, dynamical systems, etc.
-

---

# What I haven't talked about

- Dynamical systems (“continuous-time” models)
  - Other models for (Bayesian) analysis of confounding
    - Structural equations, mixed graphs et al.
    - Potential outcomes (Rosenbaum, 2002)
  - Detailed discovery algorithms
    - Including latent variable models/non-independence constraints
  - Active learning
  - Measurement error, sampling selection bias
  - Lack of overlap under conditioning
  - Formalizing non-ideal interventions
    - Non-compliance, etc.
-



---

Thank you

---

---

# Textbooks

- Glymour, C. and Cooper, G. (1999). *Computation, Causation and Discovery*. MIT Press.
  - Pearl, J. (2000). *Causality*. Cambridge University Press.
  - Rosenbaum, P. (2002). *Observational Studies*. Springer.
  - Spirtes, P, Glymour, C. and Scheines, R. (2000). *Causation, Prediction and Search*. MIT Press.
-

# Other references

- Brito, C. and Pearl, J. (2002) “Generalized instrumental variables.” UAI 2002
- Cooper, G. and Yoo, C. (1999). “Causal inference from a mixture of experimental and observational data”. UAI 1999.
- Robins, J. (1999). “Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models.” In *Computation, Causation and Discovery*.
- Robins, J., Scheines, R., Spirtes, P. and Wasserman, L. (2003). “Uniform convergence in causal inference.” *Biometrika*.
- Shipley, B. (1999). “Exploring hypothesis space: examples from organismal biology”. *Computation, Causation and Discovery*, MIT Press.
- Steenland, K. and Greenland, S. (2004). “Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer”. *American Journal of Epidemiology* 160 (4).
- Wright, S. (1921). “Cause and correlation.” *Journal of Agricultural Research*
- Huang, Y. and Valorta, M. (2006). "Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm." AAAI-06.
- Zhang, J. and Spirtes, P. (2003). “Strong Faithfulness and Uniform Consistency in Causal Inference”. UAI 2003.
- Zhang, J. (2005). PhD Thesis, Department of Philosophy, Carnegie Mellon University.

# To know more

- A short article by me:
  - <http://www.homepages.ucl.ac.uk/~ucgtrbd/papers/causality.pdf>
- Hernan and Robins' incoming textbook
  - <http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>
- Pearl's "Causality"
- Spirtes/Glymour/Scheine's "Causation, Prediction and Search"
- Morgan and Winship's "Counterfactuals and Causal Inference" (2<sup>nd</sup> edition out this weekend)