Automatic Causal Discovery

Richard Scheines
Peter Spirtes, Clark Glymour

Dept. of Philosophy & CALD
Carnegie Mellon

Outline

1. Motivation
2. Representation
3. Discovery
4. Using Regression for Causal Discovery

1. Motivation

Non-experimental Evidence

	Day Care	Aggressivenes
John	A lot	A lot
Mary	None	A little
⋮	⋮	⋮

Typical Predictive Questions

- Can we predict aggressiveness from Day Care
- Can we predict crime rates from abortion rates 20 years ago

Causal Questions:

- Does attending Day Care cause Aggression?
- Does abortion reduce crime?

Causal Estimation

When and how can we use non-experimental data to tell us about the effect of an intervention?

Manipulated Probability $P(Y \mid X \text{ set} = x, \mathbf{Z}=\mathbf{z})$

from

Unmanipulated Probability $P(Y \mid X = x, \mathbf{Z}=\mathbf{z})$

Conditioning vs. Intervening

$$P(Y \mid X = x_1) \text{ vs. } P(Y \mid X \text{ set= } x_1)$$

$\Rightarrow$ Stained Teeth Slides

2. Representation

1. Representing causal structure, and connecting it to probability
2. Modeling Interventions

Causation & Association

X and Y are associated iff

$\exists x_1 \neq x_2 \; P(Y \mid X = x_1) \neq P(Y \mid X = x_2)$

X is a cause of Y iff

$\exists x_1 \neq x_2 \; P(Y \mid X \text{ set} = x_1) \neq P(Y \mid X \text{ set} = x_2)$

Direct Causation

X is a direct cause of Y relative to **S**, iff

$$\exists \mathbf{z}, x_1 \neq x_2 \ P(Y \mid X \text{ set= } x_1, \mathbf{Z} \text{ set= } \mathbf{z})$$

$$\neq P(Y \mid X \text{ set= } x_2, \mathbf{Z} \text{ set= } \mathbf{z})$$

where **Z** = **S** - {X,Y}

$$X \longrightarrow Y$$

Association

X and Y are associated iff

$\exists x_1 \neq x_2 \; P(Y \mid X = x_1) \neq P(Y \mid X = x_2)$

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

X and Y are independent iff

X and Y are not associated

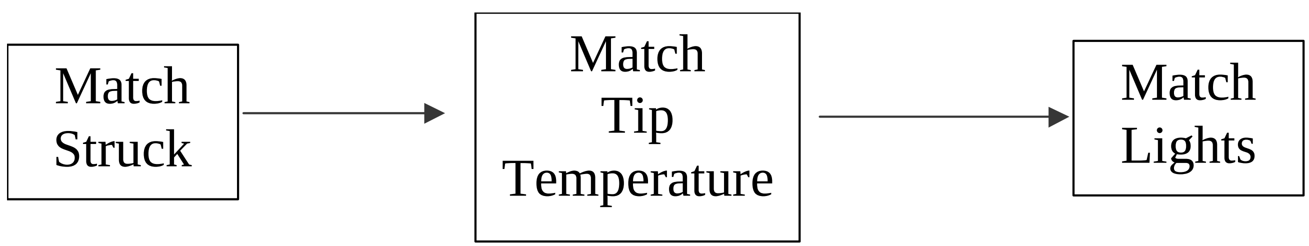
$$X \perp\!\!\!\perp Y$$

Causal Graphs

Causal Graph G = {**V,E**}

Each edge X $\rightarrow$ Y represents a direct causal claim:

X is a direct cause of Y relative to **V**



Modeling Ideal Interventions

Ideal Interventions (on a variable X):

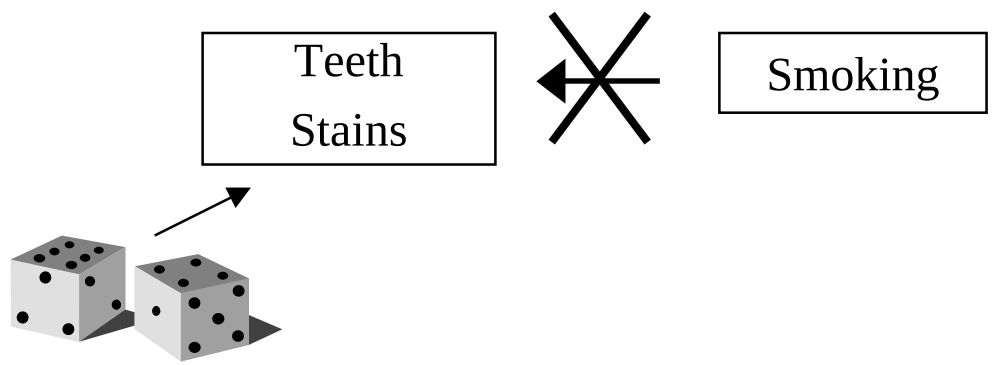
- Completely *determine* the value or distribution of a variable X

- Directly Target only X
  (no "fat hand")
  E.g., Variables: Confidence, Athletic Performance
  Intervention 1: hypnosis for confidence
  Intervention 2: anti-anxiety drug (also muscle relaxer)

Modeling Ideal Interventions

Interventions on the Effect

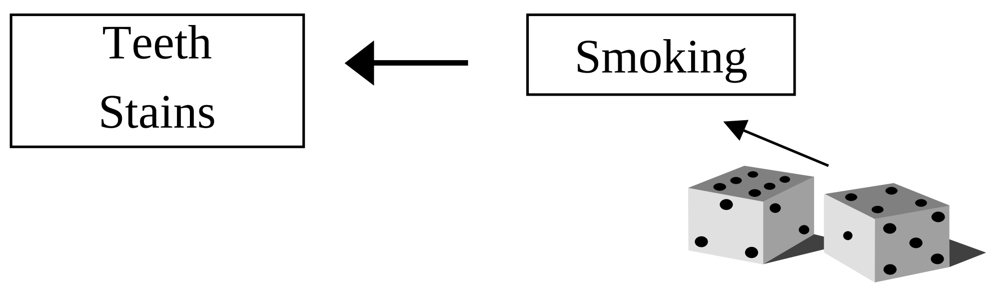
[redacted]experimental System



Modeling Ideal Interventions

Interventions on the Cause

[redacted]-experimental System



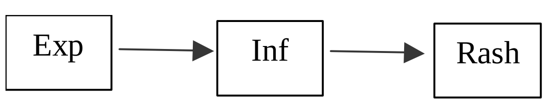
Interventions & Causal Graphs

- Model an ideal intervention by adding an “intervention” variable outside the original system
- Erase all arrows pointing into the variable intervened upon

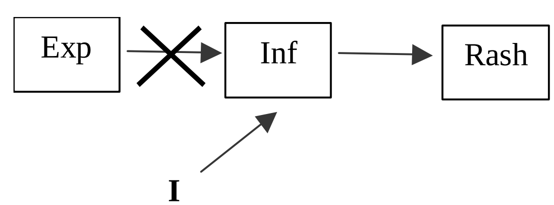
---

Intervene to change Inf

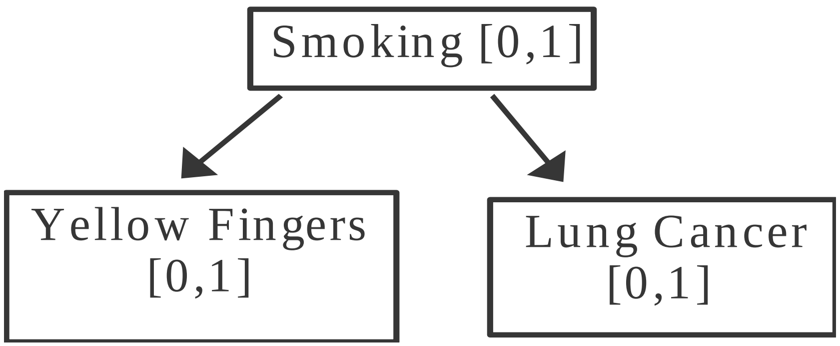
Pre-intervention graph



Post-intervention graph?



Calculating the Effect of Interventions

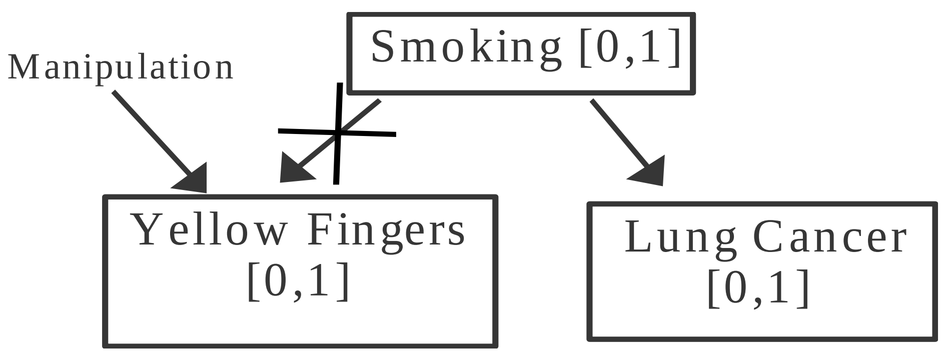


$$P(YF,S,L) = P(S)\,P(YF|S)\,P(L|S)$$

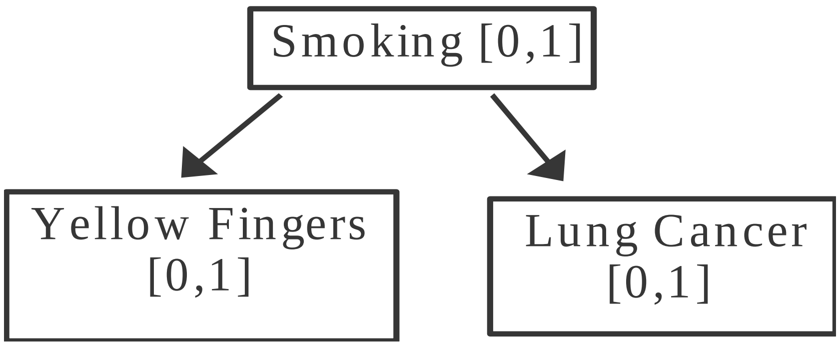
Replace pre-manipulation causes

with manipulation

$$P(YF,S,L)_m = P(S)\,P(YF|Manip)\,P(L|S)$$



Calculating the Effect of Interventions



$P(YF,S,L) = P(S)\,P(YF|S)\,P(L|S)$

Probability Calculus

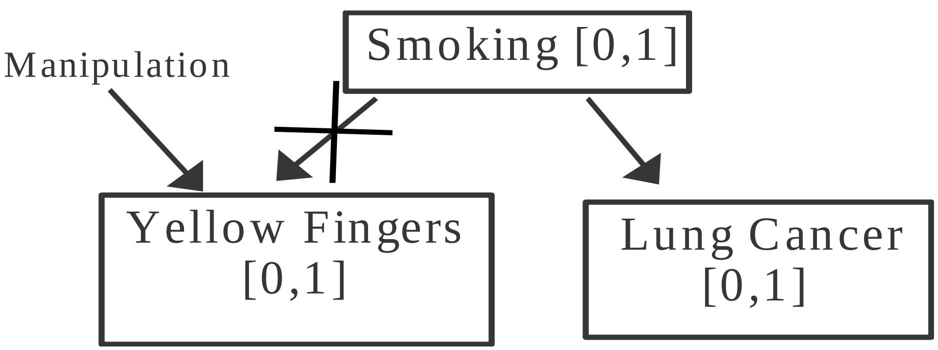
$P(L|YF)$

$\neq$

$P(YF,S,L) = P(S)\,P(YF|Manip)\,P(L|S)$

Probability Calculus

$P(L|\text{ YF set by Manip})$



The Markov Condition

Causal Structure

Markov Condition

Statistical Predictions

Causal Graphs

Independence

$X \rightarrow Y \rightarrow Z$

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

i.e.,

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

Causal Markov Axiom

In a Causal Graph G, each variable V is

independent of its non-effects,
conditional on its direct causes

in every probability distribution that G can parameterize (generate)

Causal Graphs $\Rightarrow$ Independence

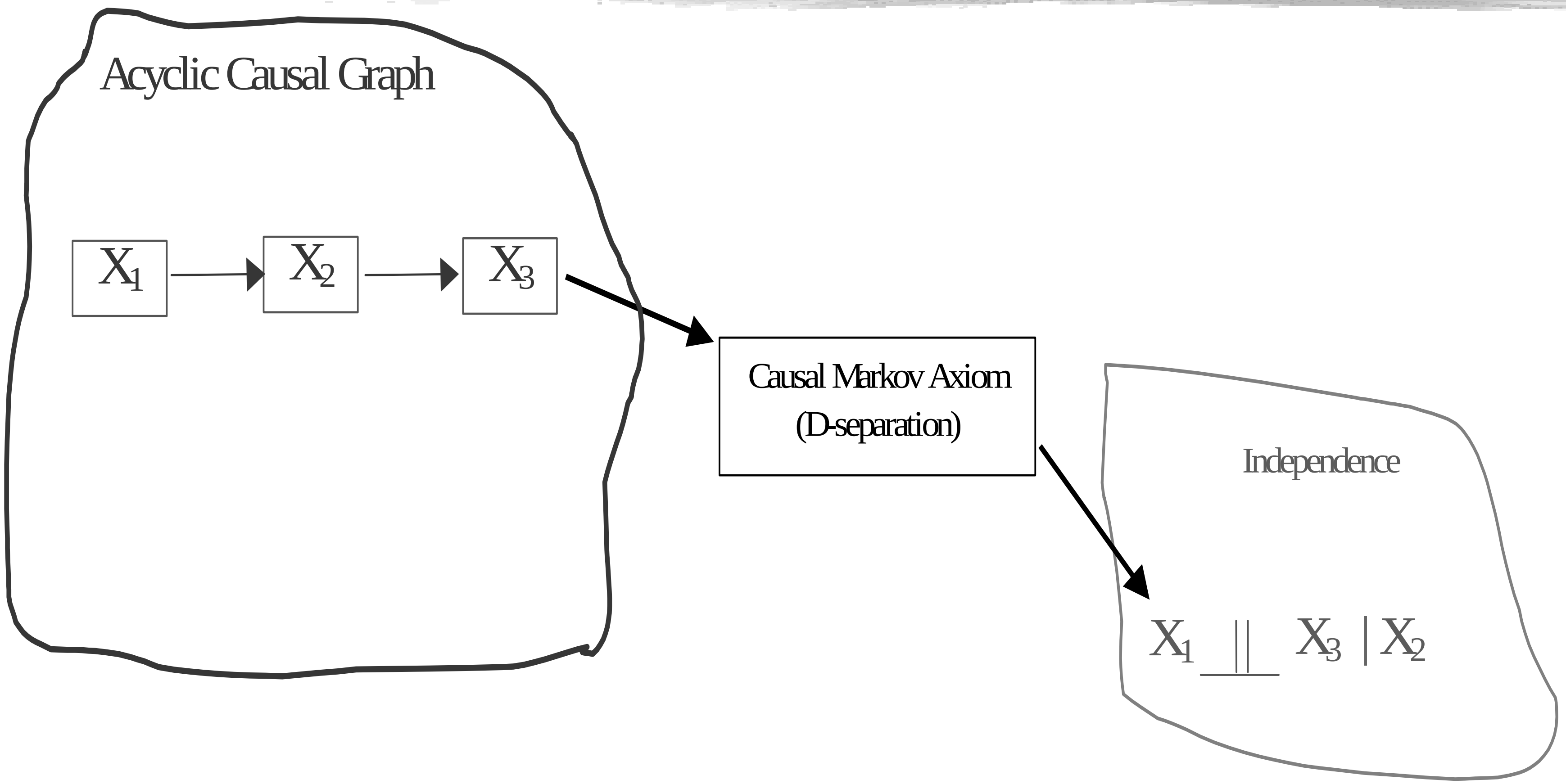
Acyclic causal graphs:

d-separation $\Leftrightarrow$ Causal Markov axiom

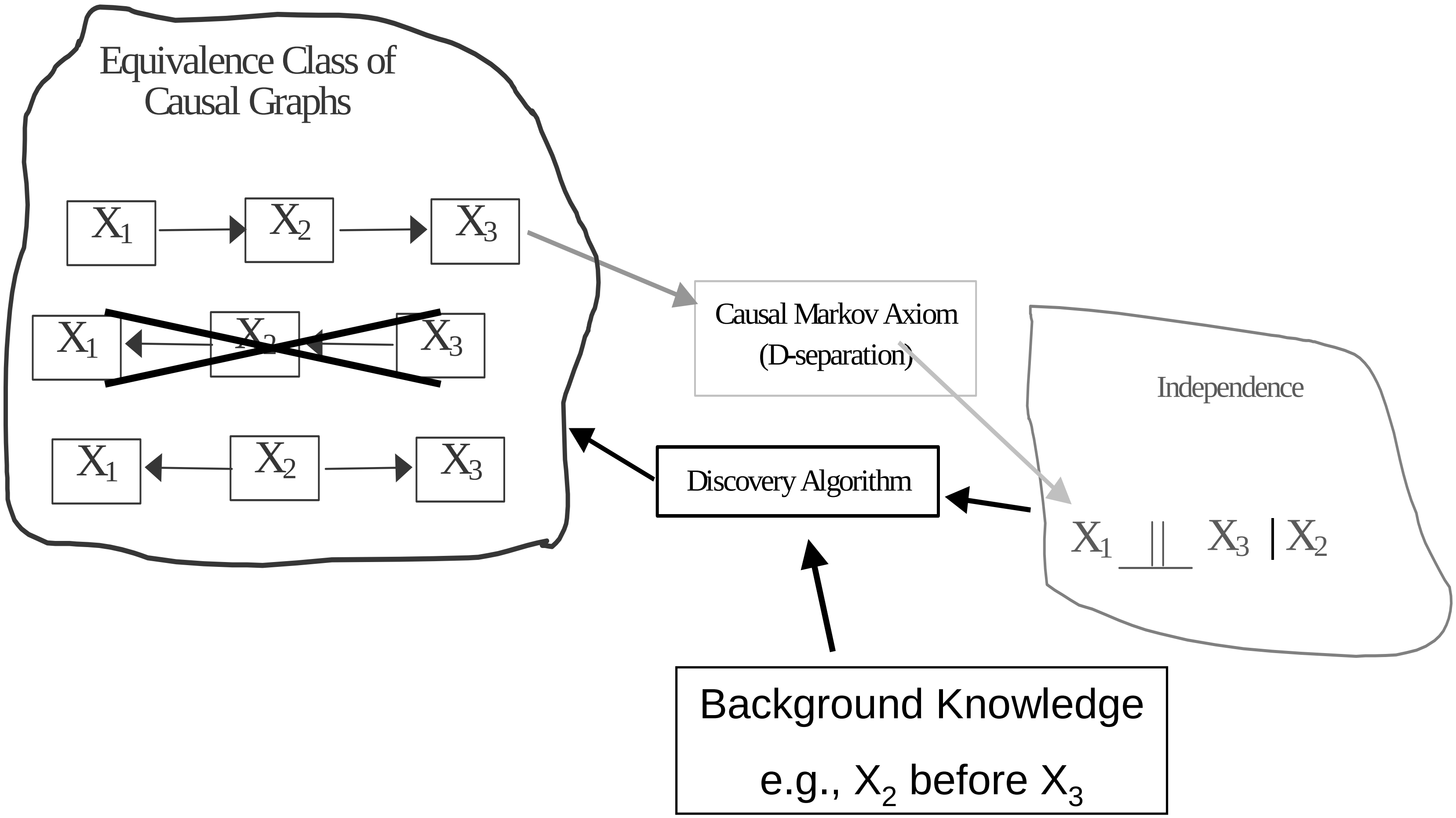
Cyclic Causal graphs:

- Linear structural equation models : d-separation, *not* Causal Markov
- For some discrete variable models: d-separation, not Causal Markov
- Non-linear cyclic SEMs : neither

Causal Structure ⇒ Statistical Data



Causal Discovery
Statistical Data $\Rightarrow$ Causal Structure



Equivalence Classes

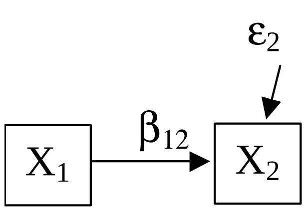
- D-separation equivalence
- D-separation equivalence over a set **O**
- Distributional equivalence
- Distributional equivalence over a set **O**

Two causal models M_1 and M_2 are distributionally equivalent iff for any parameterization θ_1 of M_1, there is a parameterization θ_2 of M_2 such that $M_1(\theta_1) = M_2(\theta_2)$, and vice versa.

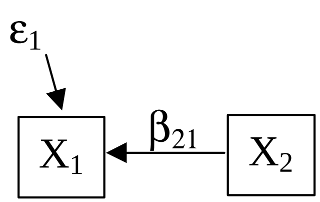
Equivalence Classes

For example, interpreted as SEM models

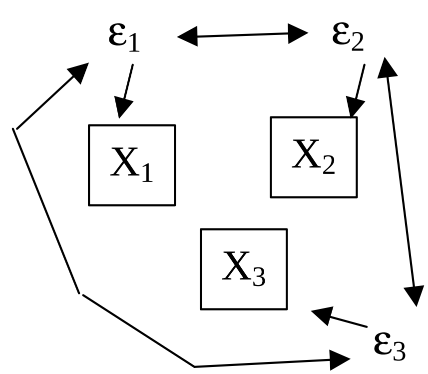
M_1 and M_2 : d-separation equivalent & distributionally equivalent
M_3 and M_4 : d-separation equivalent & *not* distributionally equivalent



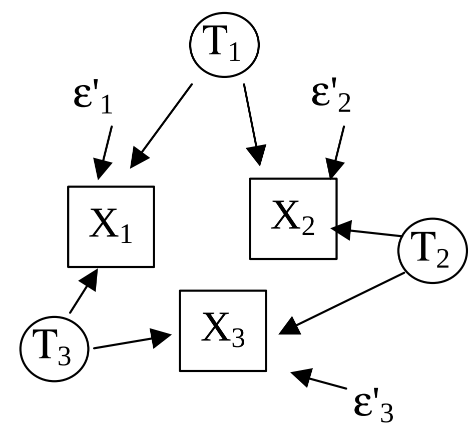
M_1



M_2



M_3



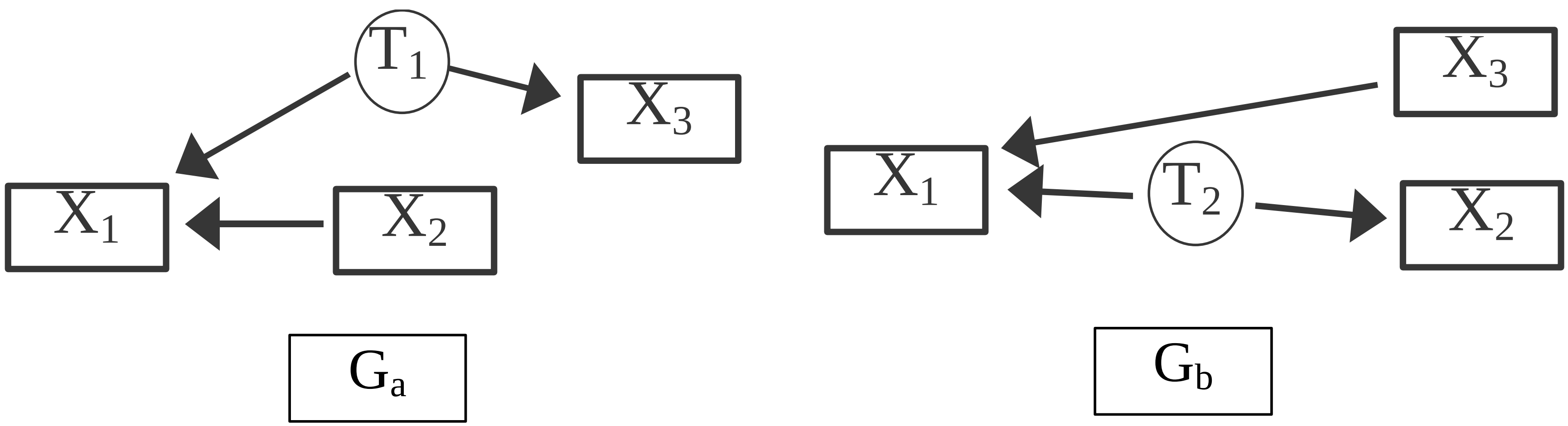
M_4

D-separation Equivalence Over a set X

Let **X** = $\{X_1, X_2, X_3\}$, then G_a and G_b

1) are not d-separation equivalent, but

2) are d-separation equivalent over **X**



D-separation Equivalence

D-separation Equivalence Theorem (Verma and Pearl, 1988)

Two acyclic graphs over the same set of variables are d-separation equivalent iff they have:

- the same adjacencies
- the same unshielded colliders

Representations of D-separation Equivalence Classes

We want the representations to:

- Characterize the Independence Relations Entailed by the Equivalence Class

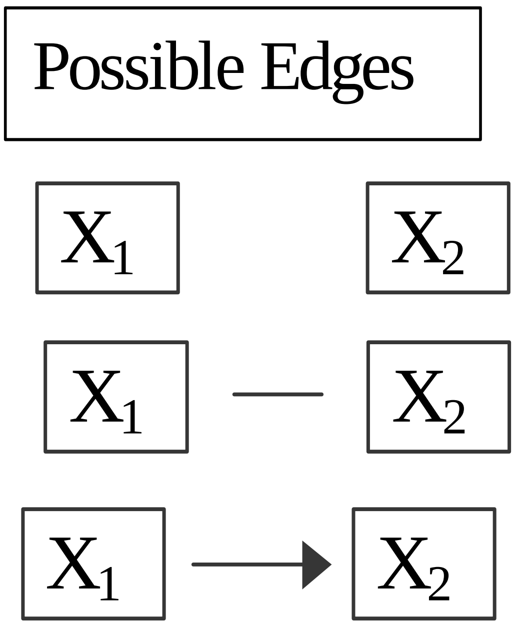
- Represent causal features that are shared by every member of the equivalence class

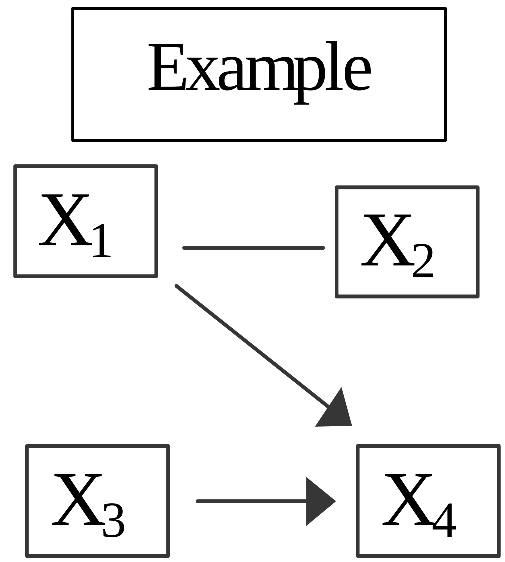
Patterns & PAGs

- Patterns (Verma and Pearl, 1990): graphical representation of an acyclic d-separation equivalence
  - no latent variables.

- PAGs: (Richardson 1994) graphical representation of an equivalence class including *latent variable models* and *sample selection bias* that are d-separation equivalent over a set of measured variables **X**

Patterns



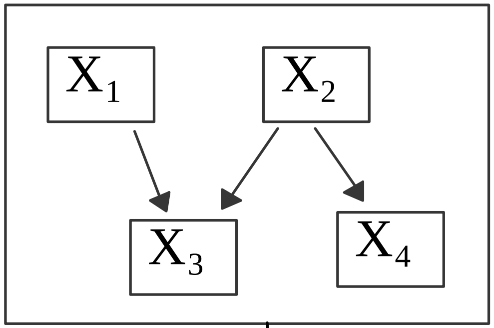


Patterns: What the Edges Mean

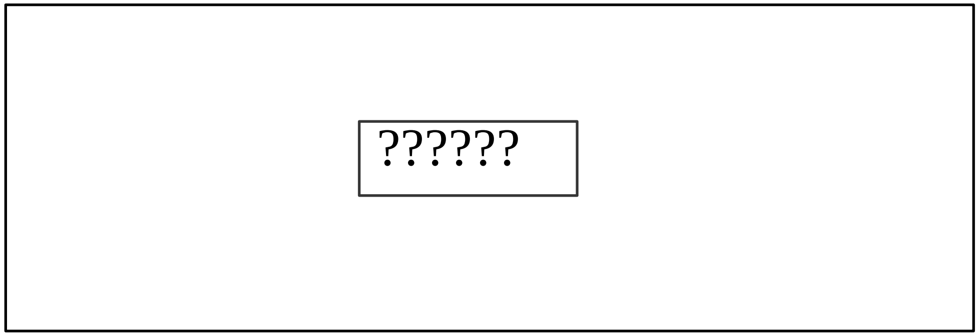
$X_1 \qquad X_2$	X_1 and X_2 are not adjacent in any member of the equivalence class
$X_1 \longrightarrow X_2$	$X_1 \rightarrow X_2$ (X_1 is a cause of X_2) in every member of the equivalence class.
$X_1 \;—\; X_2$	$X_1 \rightarrow X_2$ in some members of the equivalence class, and $X_2 \rightarrow X_1$ in others.

Patterns

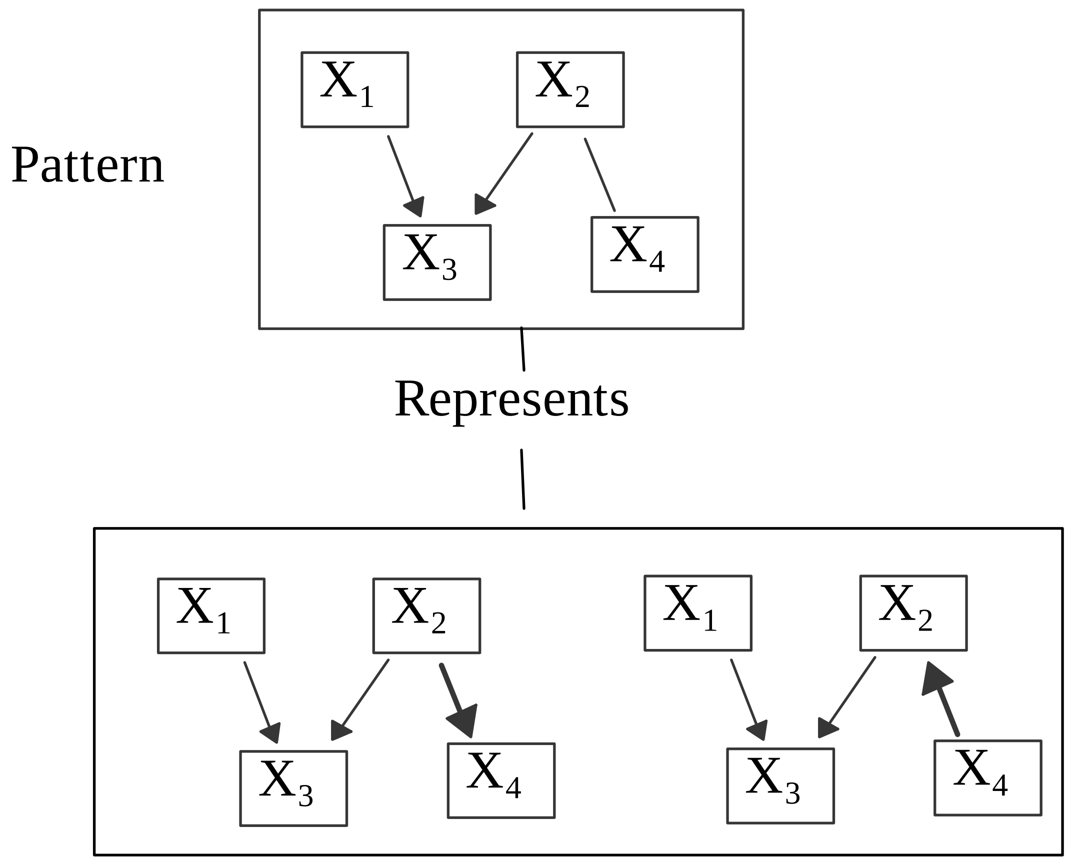
DAG



D-separation Equivalence Class

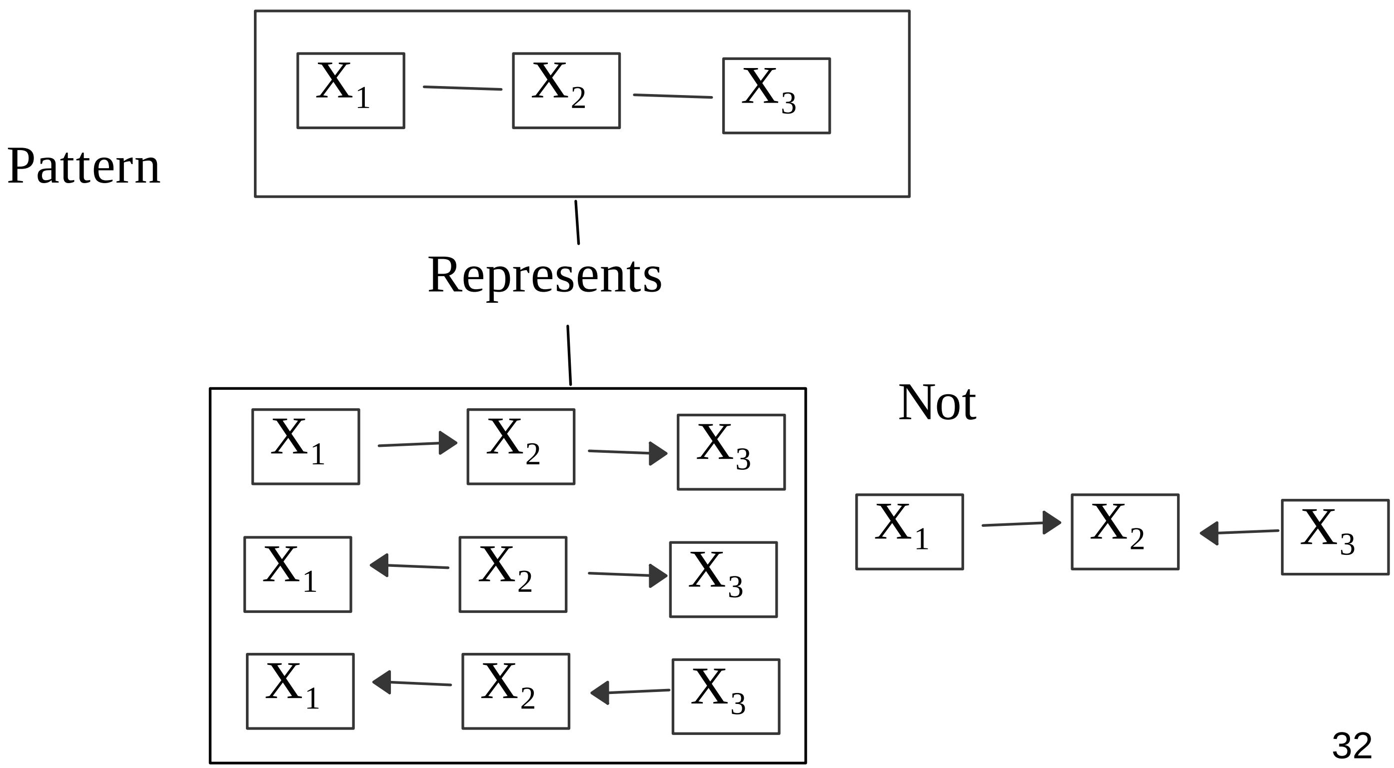


Patterns



Patterns

Not all boolean combinations of orientations of unoriented pattern adjacencies occur in the equivalence class.



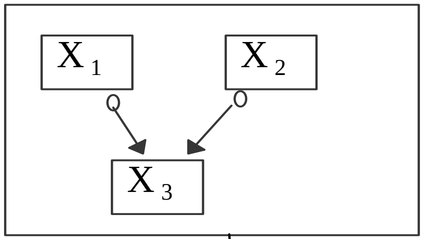
PAGs: Partial Ancestral Graphs

What PAG edges mean.

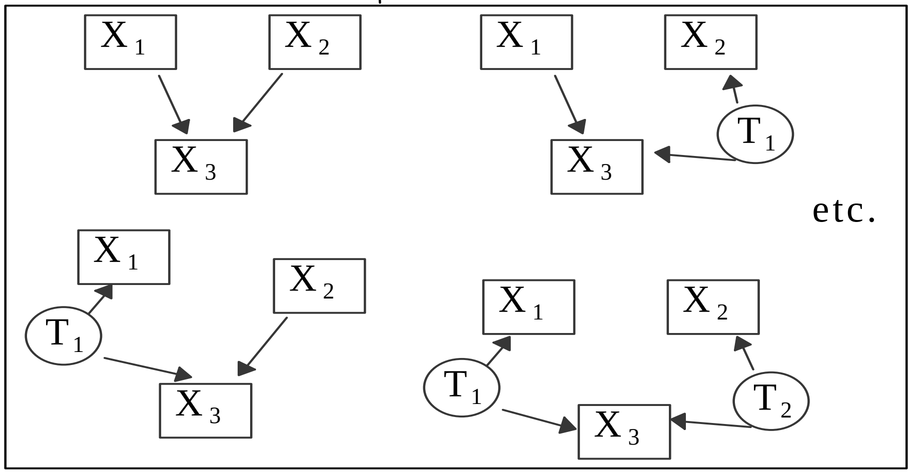
X_1 X_2	X_1 and X_2 are not adjacent
X_1 o→ X_2	X_2 is not an ancestor of X_1
X_1 o—o X_2	No set d-separates X_2 and X_1
X_1 → X_2	X_1 is a cause of X_2
X_1 ↔ X_2	There is a latent common cause of X_1 and X_2

PAGs: Partial Ancestral Graph

PAG



Represents



Search Difficulties

- The number of graphs is super-exponential in the number of observed variables (if there are no hidden variables) or infinite (if there are hidden variables)
- Because some graphs are equivalent, can only predict those effects that are the same for every member of equivalence class
  - Can resolve this problem by outputting equivalence classes

What Isn’t Possible

- Given just data, and the Causal Markov and Causal Faithfulness Assumptions:
  - Can’t get probability of an effect being within a given range without assuming a prior distribution over the graphs and parameters

What Is Possible

- Given just data, and the Causal Markov and Causal Faithfulness Assumptions:
  - There are procedures which are asymptotically correct in predicting effects (or saying “don’t know”)

Overview of Search Methods

- Constraint Based Searches
  - TETRAD
- Scoring Searches
  - Scores: BIC, AIC, etc.
  - Search: Hill Climb, Genetic Alg., Simulated Annealing
  - Very difficult to extend to latent variable models

Heckerman, Meek and Cooper (1999). “A Bayesian Approach to Causal Discovery” chp. 4 in Computation, Causation, and Discovery, ed. by Glymour and Cooper, MIT Press, pp. 141-166

Constraint-based Search

- Construct graph that most closely implies conditional independence relations found in sample
- Doesn't allow for comparing how much better one model is than another
- It is important not to test all of the possible conditional independence relations due to speed and accuracy considerations – FCI search selects subset of independence relations to test

Constraint-based Search

- Can trade off informativeness versus speed, without affecting correctness
- Can be applied to distributions where tests of conditional independence are known, but scores aren't
- Can be applied to hidden variable models (and selection bias models)
- Is asymptotically correct

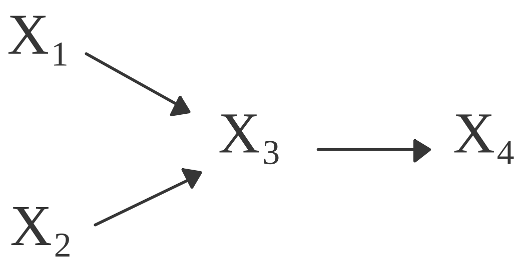
Search for Patterns

Adjacency:

•X and Y are <u>adjacent</u> if they are dependent conditional on *all* subsets that don't include X and Y

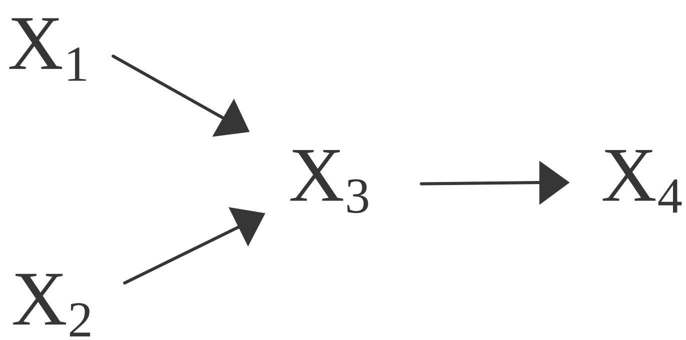
•X and Y are <u>not adjacent</u> if they are independent conditional on <u>any</u> subset that doesn't include X and Y

Search



Independencies entailed???

Search



Independencies entailed

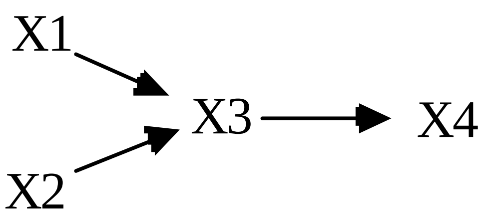
$X_1 \perp\!\!\!\perp X_2$

$X_1 \perp\!\!\!\perp X_4 \mid X_3$

$X_2 \perp\!\!\!\perp X_4 \mid X_3$

Search: Adjacency

Causal Graph



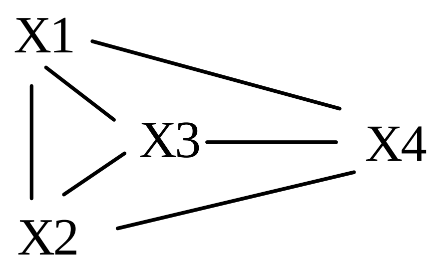
Independendcies

$X1 \perp\!\!\!\perp X2$

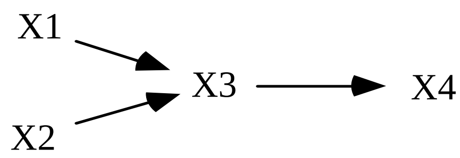
$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

Begin with:



Causal Graph



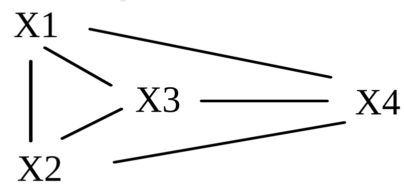
Independencies

$X1 \perp\!\!\!\perp X2$

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

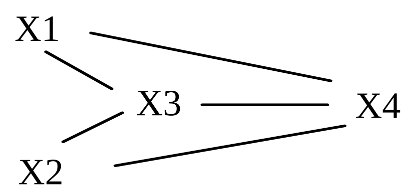
$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

Begin with:



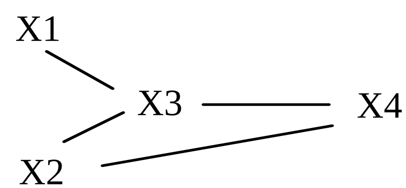
From

$X1 \perp\!\!\!\perp X2$



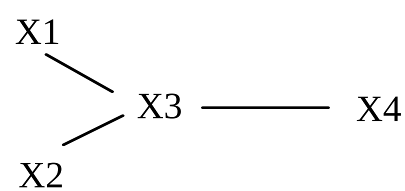
From

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$



From

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$



Search: Orientation in Patterns

Before OrientationY Unshielded

$$X - Y - Z$$

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

Collider

$$X \rightarrow Y \leftarrow Z$$

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

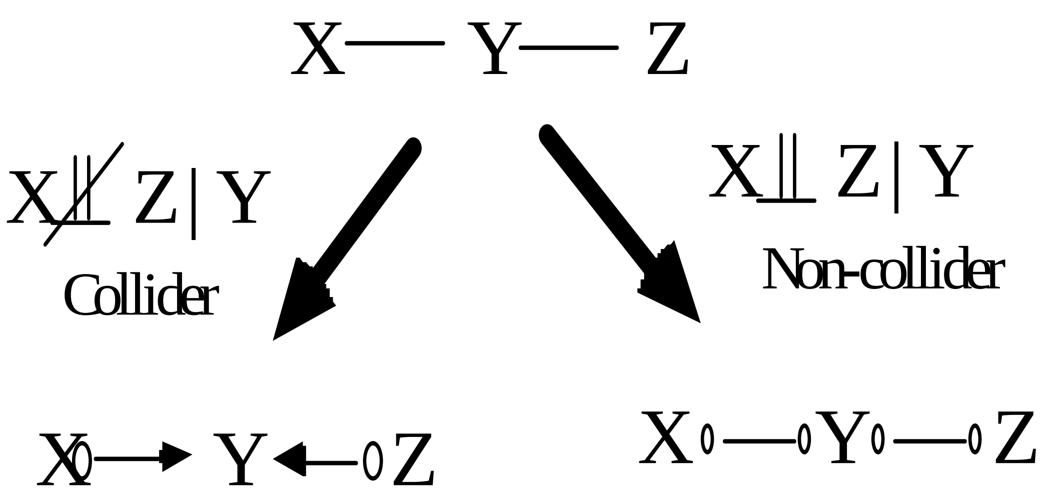
Non-collider

$$X - Y - Z$$

$$\boxed{\begin{array}{c} X \leftarrow Y \rightarrow Z \\ X \leftarrow Y \leftarrow Z \\ X \rightarrow Y \rightarrow Z \end{array}}$$

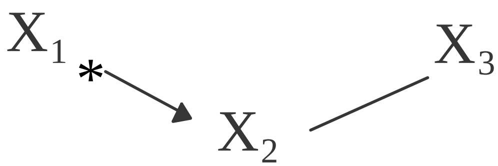
Search: Orientation in PAGs

Y Unshielded



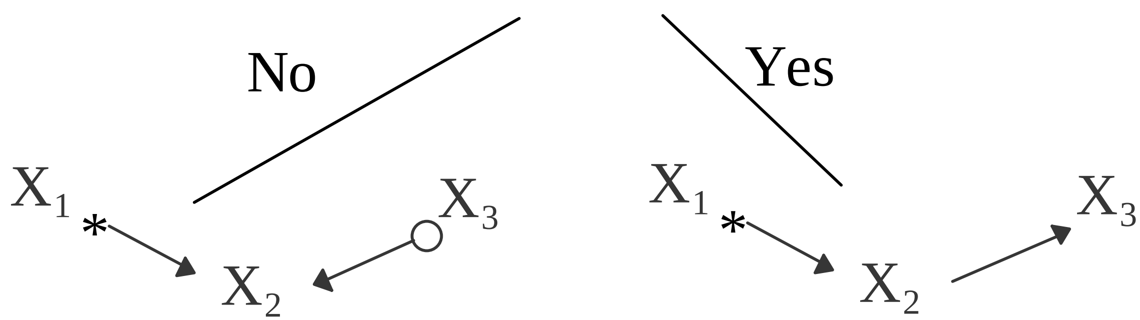
Orientation: Away from Collider

Test Conditions



1) X_1 - X_2 adjacent, and into X_2.
2) X_2 - X_3 adjacent, and unoriented.
3) X_1 - X_3 not adjacent

Test $X_1 \perp\!\!\!\perp X_3 \mid X_2$



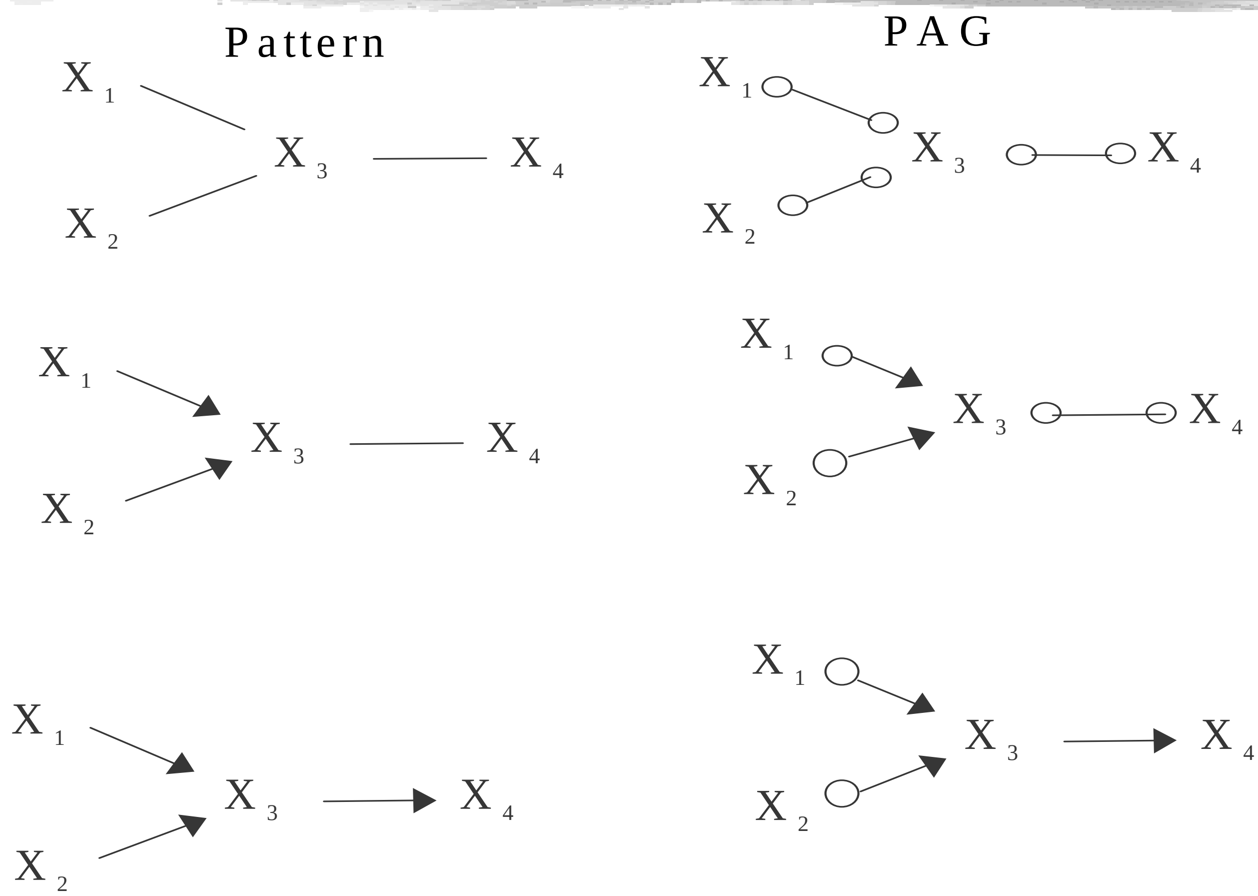
Search: Orientation

After Orientation Phase

$X1 \perp\!\!\!\perp X2$

$X1 \perp\!\!\!\perp X4 \mid X3$

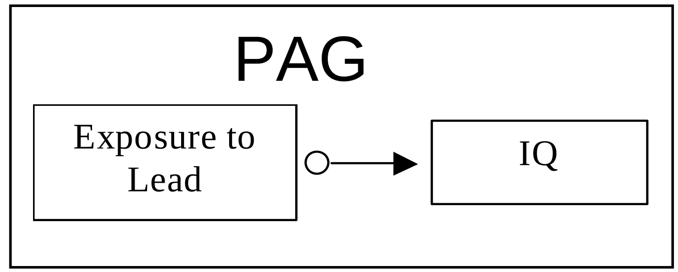
$X2 \perp\!\!\!\perp X4 \mid X3$

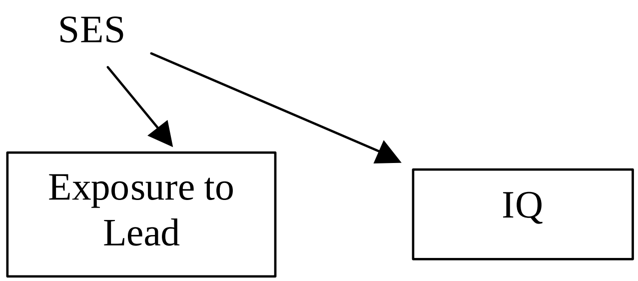


Knowing when we know enough to calculate the effect of Interventions

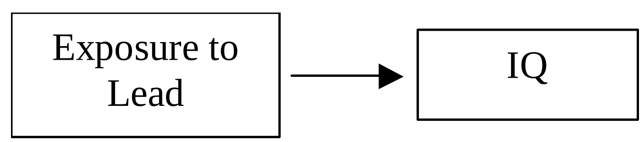
Observation: IQ $\not\perp\!\!\!\perp$ Lead

Background Knowledge: Lead prior to IQ





P(IQ | Lead) $\neq$ P(IQ | Lead set=)



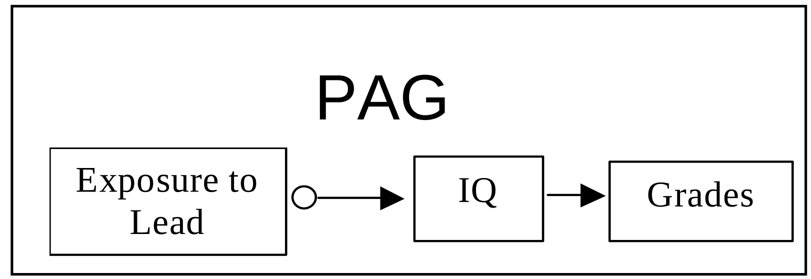
P(IQ | Lead) = P(IQ | Lead set=)

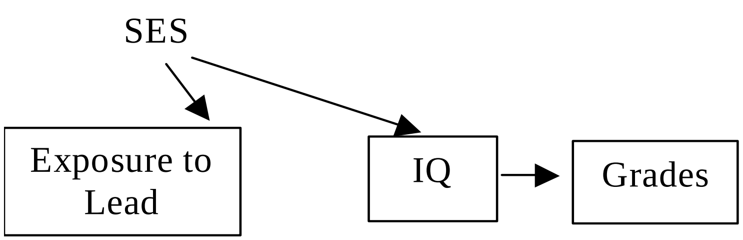
Knowing when we know enough to calculate the effect of Interventions

Observation: All pairs associated

Lead _||_ Grades | IQ

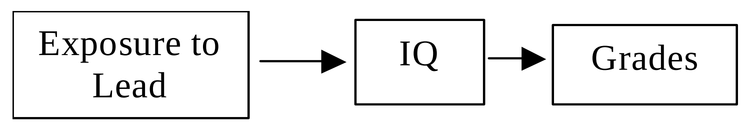
Background Knowledge Lead prior to IQ prior to Grades





P(IQ | Lead) $\neq$ P(IQ | Lead set=)

P(Grades | IQ) = P(Grades | IQ set=)



P(IQ | Lead) = P(IQ | Lead set=)

P(Grades | IQ) = P(Grades | IQ set=)

Knowing when we know enough to calculate the effect of Interventions

- Causal graph known
- Features of causal graph known
    - Prediction algorithm (SGS - 1993)
        - Data tell us when we know enough – i.e., we know when we don't know

4. Problems with Using Regession for Causal Inference

Regression to estimate Causal Influence

- Let **V** = {**X**,Y,**T**}, where

  -measured vars: **X** = $\{X_1, X_2, ..., X_n\}$

  -latent common causes of pairs in **X** U Y: **T** = $\{T_1, ..., T_k\}$

- Let the true causal model over **V** be a Structural Equation Model in which each V $\in$ **V** is a linear combination of its direct causes and independent, Gaussian noise.

Regression to estimate Causal Influence

- Consider the regression equation:

  $$Y = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots b_nX_n$$

- Let the OLS regression estimate b_i be the *estimated* causal influence of X_i on Y.

- That is, holding **X**/Xi experimentally constant, b_i is an estimate of the change in E(Y) that results from an intervention that changes Xi by 1 unit.

- Let the *real* Causal Influence $X_i \rightarrow Y = \beta_i$

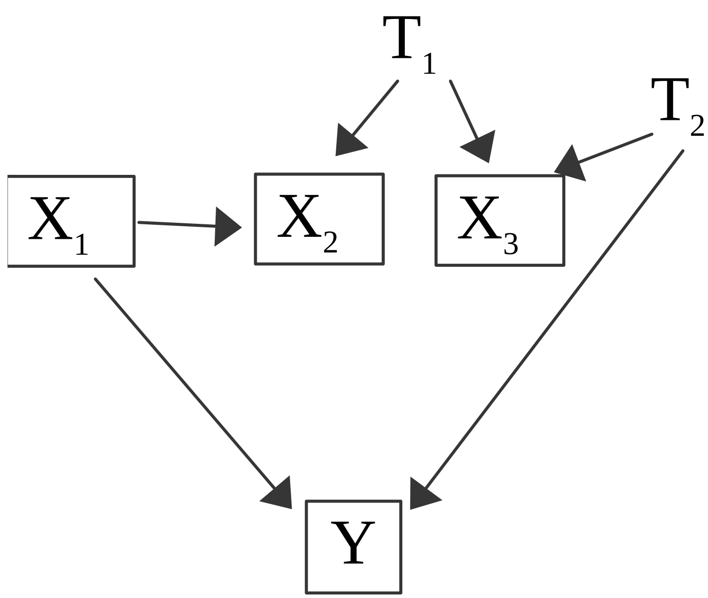
- When is the OLS estimate b_i an unbiased estimate of the the *real* Causal Influence $X_i \rightarrow Y = \beta_i$?

Regression vs. PAGs to estimate Qualitative Causal Influence

- $b_i = 0$ $\Leftrightarrow$ Xi _||_ Y | **X**/Xi
- Xi - Y *not* adjacent
  in PAG over **X** U Y $\Leftrightarrow$ $\exists$**S** $\subseteq$ **X**/Xi, Xi _||_ Y | **S**
- So for any SEM over **V** in which
  - Xi _||_ Y | **X**/Xi and
  - $\exists$**S** $\subset$ **X**/Xi, Xi _||_ Y | **S**

PAG is superior to regression wrt errors of commission

Regression Example

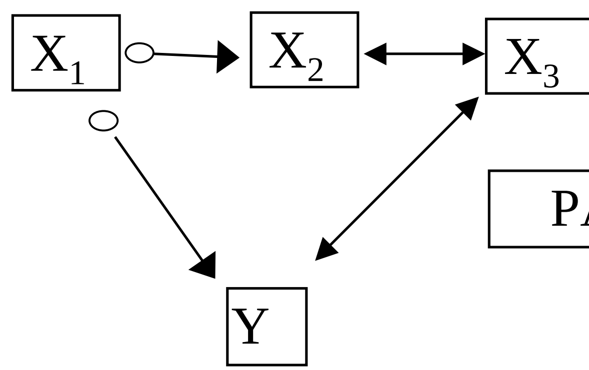


True Model

$b_1 \neq 0$ ✓

$b_2 \neq 0$ X

$b_3 \neq 0$ X



PAG

Regression Bias

If

- X_i is d-separated from Y conditional on $\mathbf{X}/X_i$ in the true graph after removing $X_i \rightarrow Y$, and
- $\mathbf{X}$ contains no descendant of Y, then:

b_i is an unbiased estimate of β_i

Regression Bias Theorem

If **T** = ∅, and **X** prior to Y, then

b_i is an unbiased estimate of β_i

Tetrad 4 Demo

www.phil.cmu.edu/projects/tetrad

Applications

- *Genetic Regulatory Networks*
- *Pneumonia*
- *Photosynthesis*
- *Lead - IQ*
- *College Retention*
- *Corn Exports*
- *Rock Classification*
- *Spartina Grass*
- *College Plans*
- *Political Exclusion*
- *Satellite Calibration*
- *Naval Readiness*

MS or Phd Projects

- Extending the Class of Models Covered
- New Search Strategies
- Time Series Models (Genetic Regulatory Networks)
- Controlled Randomized Trials vs. Observations Studies

Projects: Extending the Class of Models Covered

1) Feedback systems
2) Feedback systems with latents
3) Conservation, or equilibrium systems
4) Parameterizing discrete latent variable models

Projects: Search Strategies

1) Genetic Algorithms, Simulated Annealing
2) Automatic Discretization
3) Scoring Searches among Latent Variable Models
4) Latent Clustering & Scale Construction

References

- *Causation, Prediction, and Search*, 2nd Edition, (2001), by P. Spirtes, C. Glymour, and R. Scheines (MIT Press)

- *Causality: Models, Reasoning, and Inference*, (2000), Judea Pearl, Cambridge Univ. Press

- *Computation, Causation, & Discovery* (1999), edited by C. Glymour and G. Cooper, MIT Press

- *Causality in Crisis?*, (1997) V. McKim and S. Turner (eds.), Univ. of Notre Dame Press.

- *TETRAD IV:* www.phil.cmu.edu/tetrad

- Web Course on Causal and Statistical Reasoning : www.phil.cmu.edu/projects/csr/