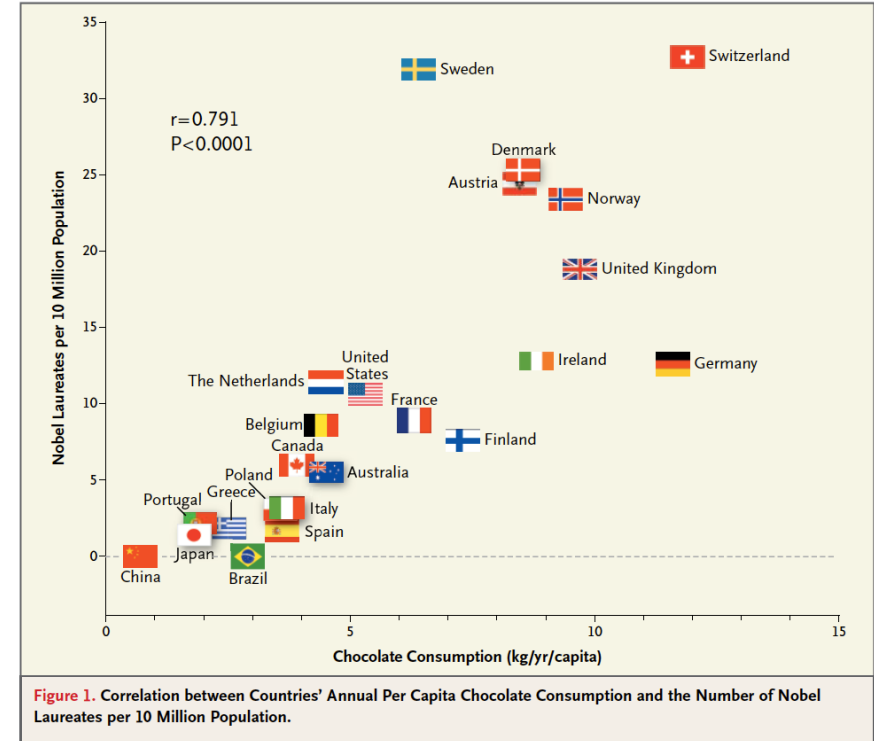# Introduction to Causal Inference
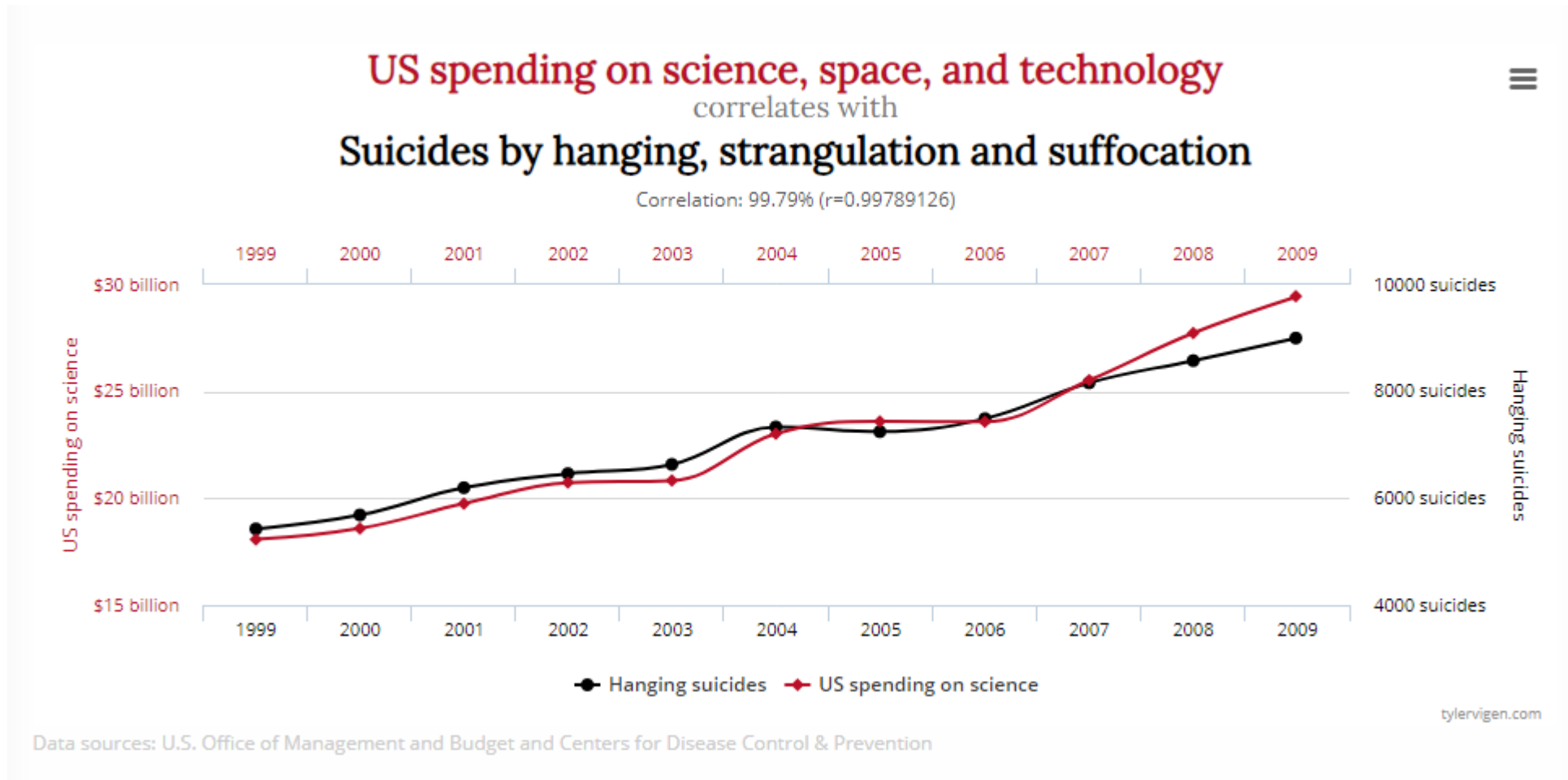
Hrushikesh Loya

What is Causal Inference ?

- Finding the effect of any particular treatment/policy/intervention

- **Example:**
  - Effect of smoking on Lung cancer ?
  - Effect of chocolate consumption on academic performance ?
  - Effect of mandatory face covering on COVID cases ?

Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Correlation doesn't imply Causation

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
Correlation: 99.79% (r=0.99789126)

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Check out: Spurious Correlations (http://www.tylervigen.com/spurious-correlations)

How to capture causality ?

1. Randomized Experiments:
   ➢ Advantages: Deals with all confounders
   ➢ Disadvantages: Costly, time-consuming, ethical issues, etc.

2. From Observational Data:
   ➢ Advantages: Cheap, fast, feasible
   ➢ Disadvantages: Might not deal with all confounders, power limitations

Notations

Structure causal model (SCM) equations:

- $X := N_x$
- $Y := -6X + N_Y$

- $N_x, N_y \sim N(0,1)$

$$(X, Y) \sim$$

Bayesian Network:

- Assume: Directed Acyclic Graph

X — Altitude

Y — Temperature

Modelling interventions

Structure causal model (SCM) equations:

- $\cancel{X := N_x}$ $X := 3$
- $Y := -6X + N_Y$

- $N_x, N_y \sim N(0,1)$

$(X, Y) \sim$

Bayesian Network:

- Assume: Directed Acyclic Graph

Altitude   Temperature
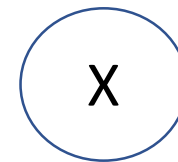
Modelling interventions

Structure causal model (SCM) equations:

- $X := N_x$
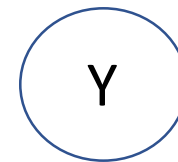- ~~$Y := -6X + N_y$~~ $Y := N(2,2)$

- $N_x, N_y \sim N(0,1)$

$$(X, Y) \sim$$

Bayesian Network:

- Assume: Directed Acyclic Graph

X   Y

Altitude   Temperature

Modelling interventions

Structure causal model (SCM) equations:
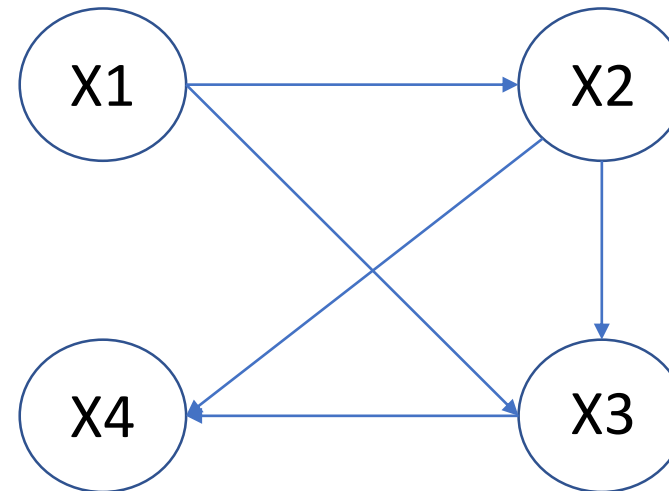
Bayesian Network:

- $X_1 := N_{x_1}$
- $X_2 := f_1(X_1, N_{x_2})$
- $X_3 := f_2(X_1, X_2, N_{x_3})$
- $X_4 := f_3(X_2, X_3, N_{x_4})$

Modelling interventions

Structure causal model (SCM) equations:

- $X_1 := N_{x_1}$
- $X_2 := f_1(X_1, N_{x_2})$
- $X_3 := f_2(X_1, X_2, N_{x_3})$
- $X_4 := f_3(X_2, X_3, N_{x_4})$

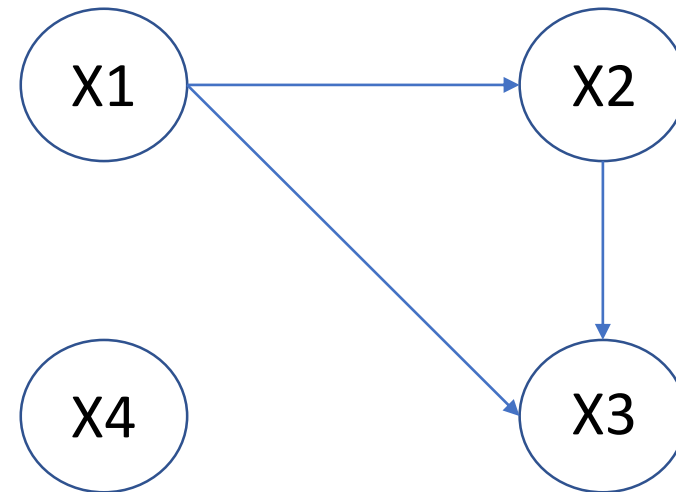**Bayesian Network:**

- Directed
- Acyclic

Modelling interventions

Call the intervention as: $P_{do(X_4:=0)} \neq P(.\,|X_4 = 0)$

Structure causal model (SCM) equations:

Bayesian Network:

- $X_1 := N_{x_1}$
- $X_2 := f_1\left(X_1, N_{x_2}\right)$
- $X_3 := f_2\left(X_1, X_2, N_{x_3}\right)$
- $\cancel{X_4 := f_3(X_2, X_3, N_{x_4})}$ $X_4 := 0$

Valid adjustment set

- Given:

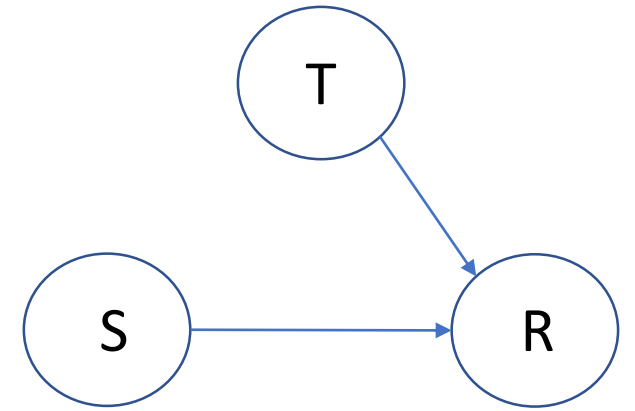|	Treatment A	Treatment B
Small Stones (300)	80/100	190/200
Large Stones (300)	165/250	20/50
	245/350 = 70%	210/250 = 84%

- Calculate $P(R \mid T = A) = \dfrac{245}{350} = 0.7$

- What about $P_{do(T := A)}(R) = ?$

- Hint: $P(s)$ and $P(R \mid T, S)$ are invariants
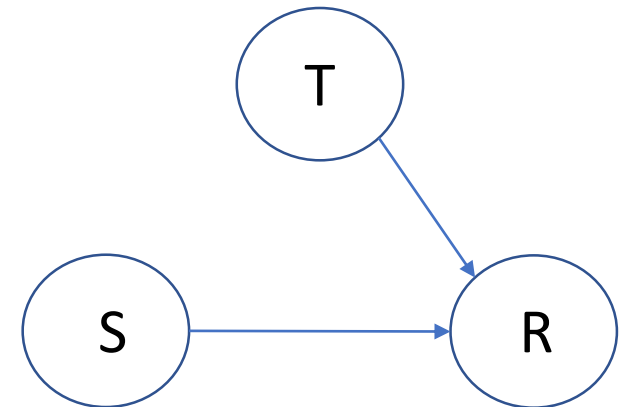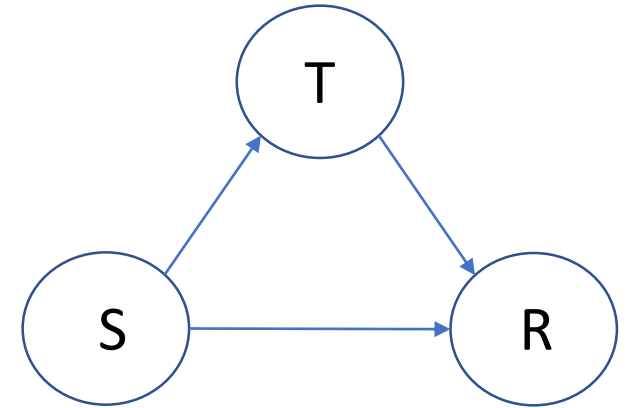
Valid adjustment set

- What about $P_{do(T:=A)}(R) = ?$
- Hint: $P(s)$ and $P(R \mid T, S)$ are invariants

$$P_{do(T:=A)}(R = 1) = \sum_{S} P_{do(T:=A)}(R = 1, T = A, S = s)$$

$$= \sum_{S} P_{do(T:=A)}(R = 1 \mid T = A, S = s) \, P_{do(T:=A)}(S = s, T = A)$$

$$= \sum_{S} P_{do(T:=A)}(R = 1 \mid T = A, S = s) \, P_{do(T:=A)}(S = s)$$

$$= \sum_{S} P(R = 1 \mid T = A, S = s) \, P(S = s)$$

$$= 0.8 * 0.5 + 0.66 * 0.5$$
$$= 0.73 > 0.7$$

Valid adjustment set

Definition: A valid adjustment set for (X, Y) is the set of covariates Z such that:

$$P_{do(X:=x)}(y) = \sum_z P(y \mid x, z)P(z) \neq P(y \mid x)$$
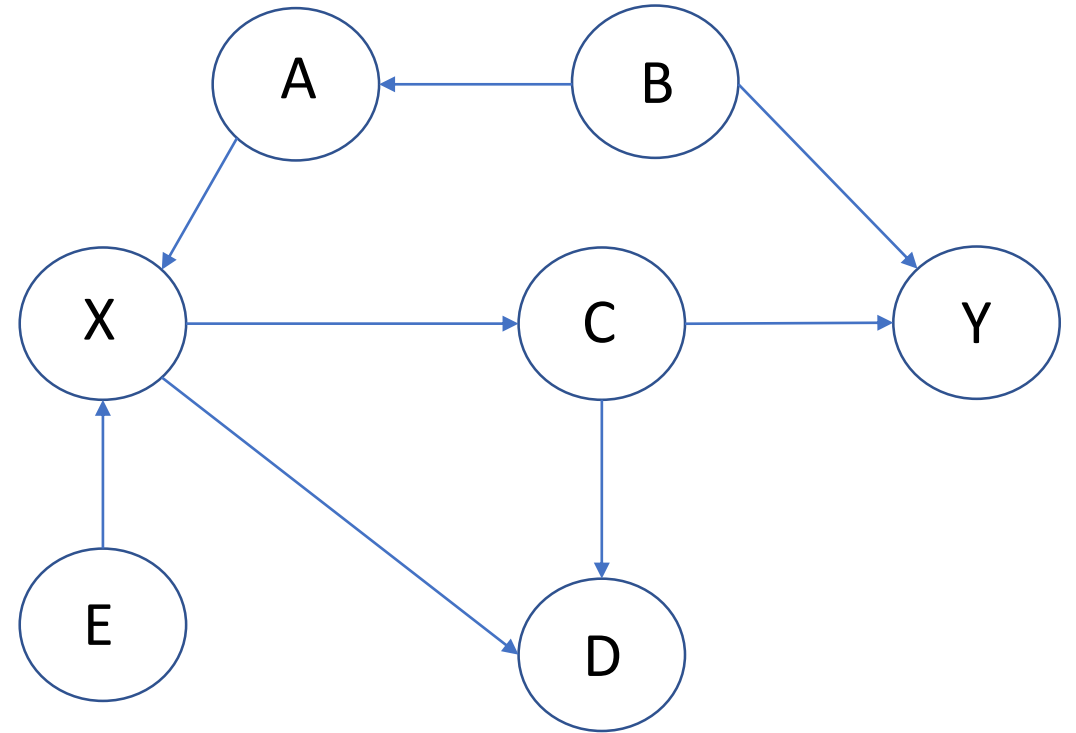
Parent Adjustment:

Assume $Y \notin PA(X)$. Then:

$PA(X)$ is a valid adjustment set for (X, Y)

TLDR: In order to get the causal effect perform your linear regression with the confounders

Valid adjustment set

- Valid adjustment sets:
1. Parent Adjustment: {A, E}
2. {B}
3. ...

- Just including elements from valid adjustment set while linear regression

- What if we have hidden random variables ?

Instrumental variables

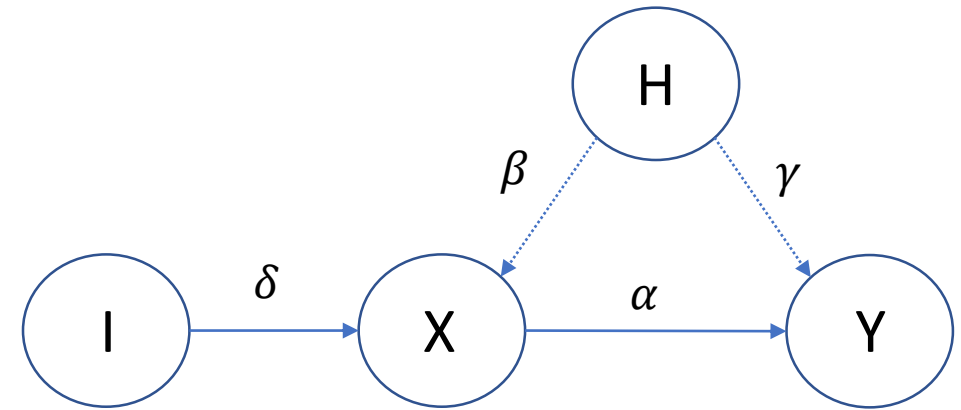- Introduce instrumental RV which **causes X** but **isn't associated with H or Y**

- We want to estimate α

$$Y = \alpha X + \gamma H + N_y \qquad (1)$$
$$X = \beta H + \delta I + N_x \qquad (2)$$

Plugging (2) in (1) we get:
$$Y = (\alpha\beta + \gamma)H + \textcolor{red}{\alpha\delta I} + \alpha N_x + N_y \qquad (3)$$

Step 1: Fit (I, X) to get δ

Step 2: Use the fitted values δI to fit (δI, Y) to get α

Possible problems ??

Mendelian Randomization

- Uses genetic information as the instrument
- E.g.: CHRNA1 Expression to detect causal relationship between smoking and lung cancer

- Challenges:
  - **Low statistical Power:** As h2 is low
  - **Population stratification:** Transferability problem
  - **Pleiotropy:** Za affects A and B

Additional reading: Smith & Hemani, Human Molecular Genetics 2014
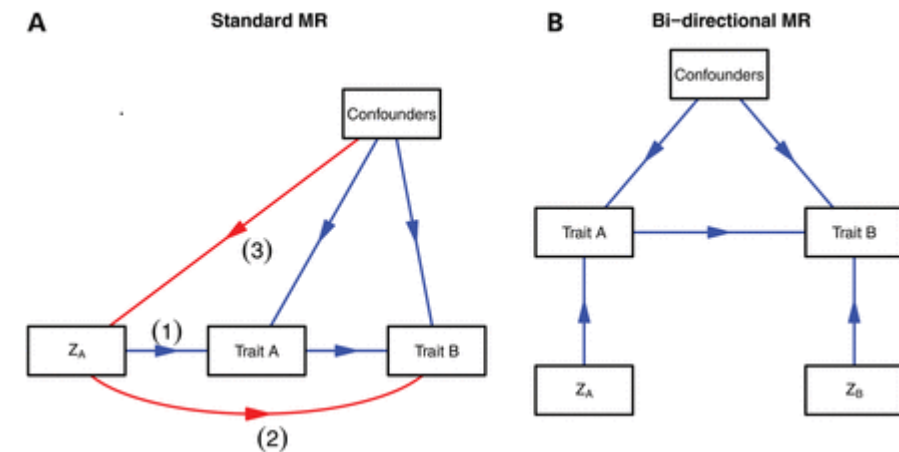
Independence based methods

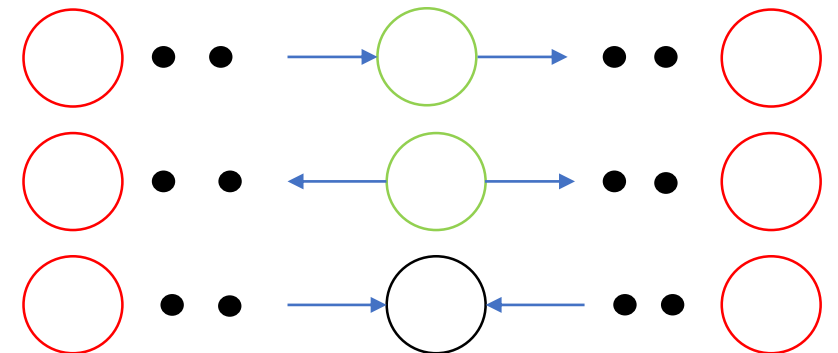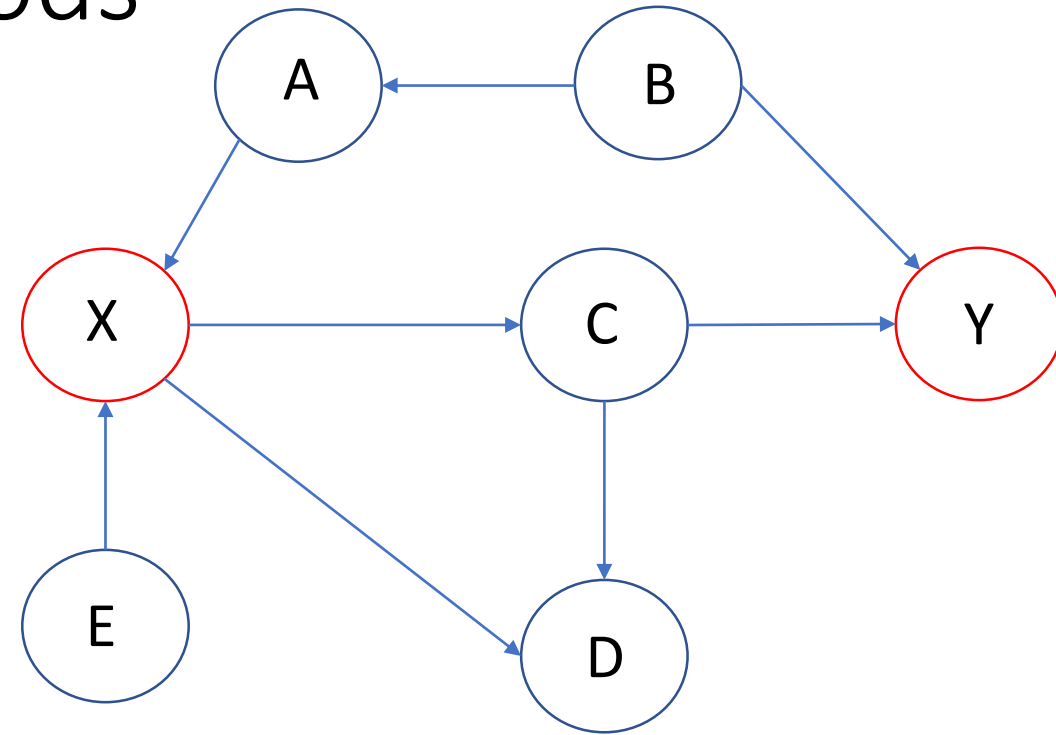- What if we don't know the causal graph ?
  Can we learn that ?

Reichenbach's common cause principle:

If $X \not\!\perp Y$ then:
  - X "causes" Y
  - Y "causes" X
  - There is a hidden common "cause"
  - Combination of above

d-separation: X_i and X_j are d-separated by S if all paths between X_i and X_j are **blocked** by S
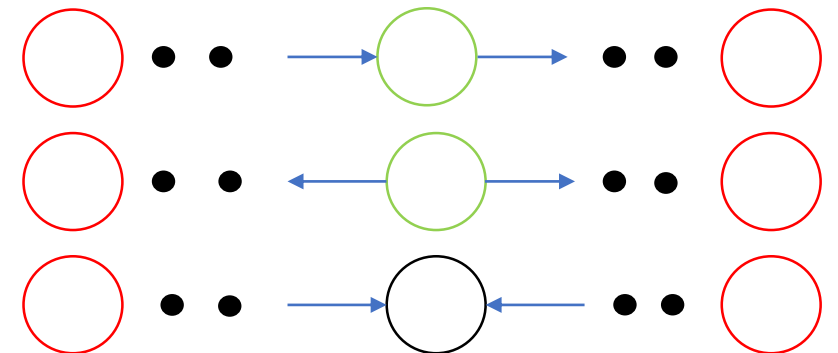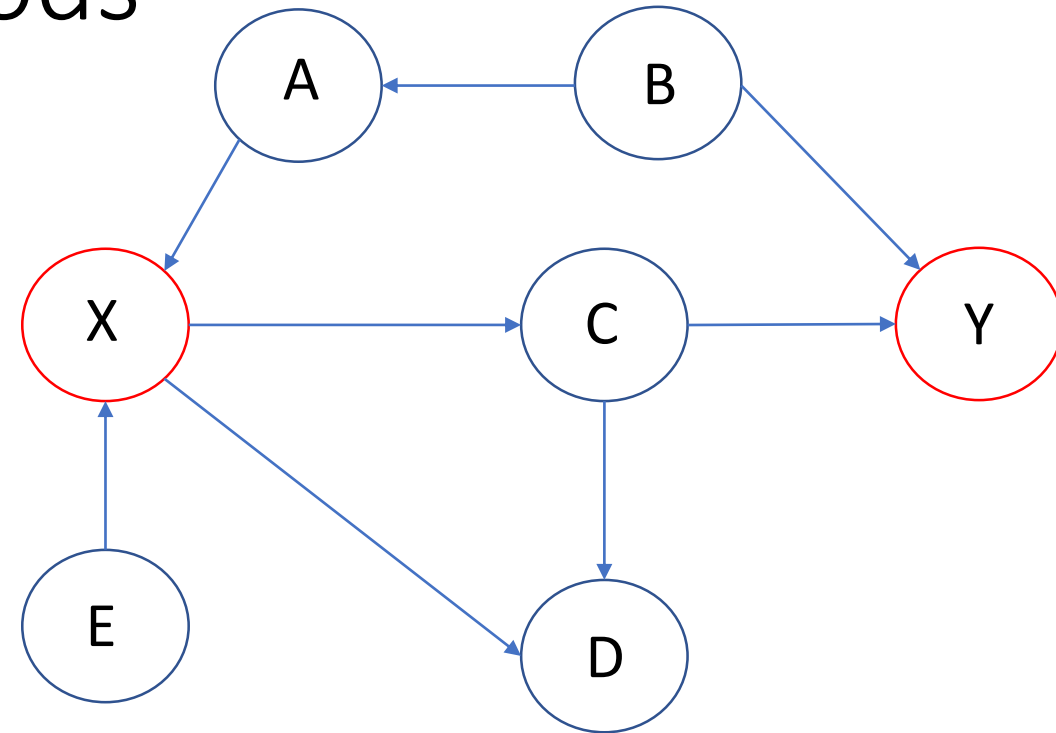
X and Y are d-separated by {C, A, E}

Independence based methods

d-separation: X_i and X_j are d-separated by S if all paths between X_i and X_j are **blocked** by S

Given P is Markov and Faithful w.r.t G then:

$$X_i \text{ and } X_j \text{ are d-separated by S in G}$$
$$\leftrightarrow X \perp Y | S$$

Additional reading: IC (Pearl, 2009); PC, FCI (Spirtes et al., 2000)

Invariant Causal Prediction

- What if we have observational (genotype-phenotype) and perturbation (gene knockout) data ?

- We have two experimental setting
  - e = 1: observational data
  - e = 2: data from unspecified intervention

- Assumption: Invariance across experiments (SEM doesn't change)

- Proposition: If $S^* = PA_y$ then for all $e \in E$: $P(Y^e | X_{S^*}^e = x)$ remains same

- Idea: Find all S^* which satisfy the invariance condition

$$S(\mathcal{E}) = \bigcap \{S; \; H_{0,S}(\mathcal{E}) \text{ holds}\},$$

Invariant Causal Prediction

What about speed ? Worst-case complexity $O(2^n)$

1. Start with smaller sets, progress to larger if all previous small subsets are rejected
2. Early stopping: Stop if two disjoint subsets are not rejected
3. Initialization: Only consider non-zero regression parameters
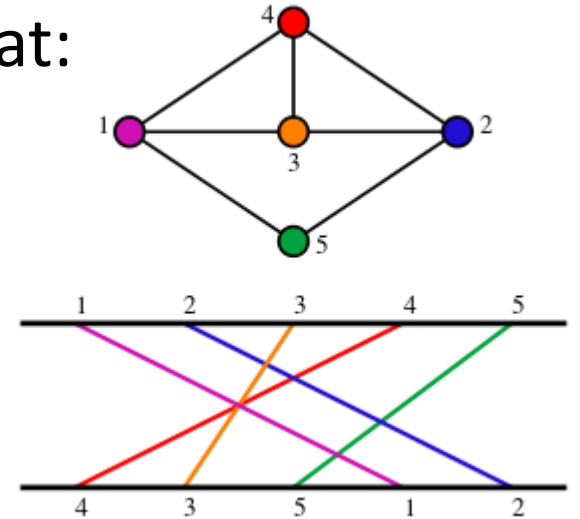
Table 2.   Timing comparisons in minutes

Method	No. genes		
	50	500	5,000
ICP	0.233	2.64	27.7
hiddenICP	0.012	0.12	1.4
pc	0.004	0.10	2.4
rfci	0.004	0.12	3.6
ges	0.002	0.80	1,002.4
gies	0.010	4.06	842.8
Regression	0.069	0.70	7.5

Additional reading: Meinshausen, PNAS 2016
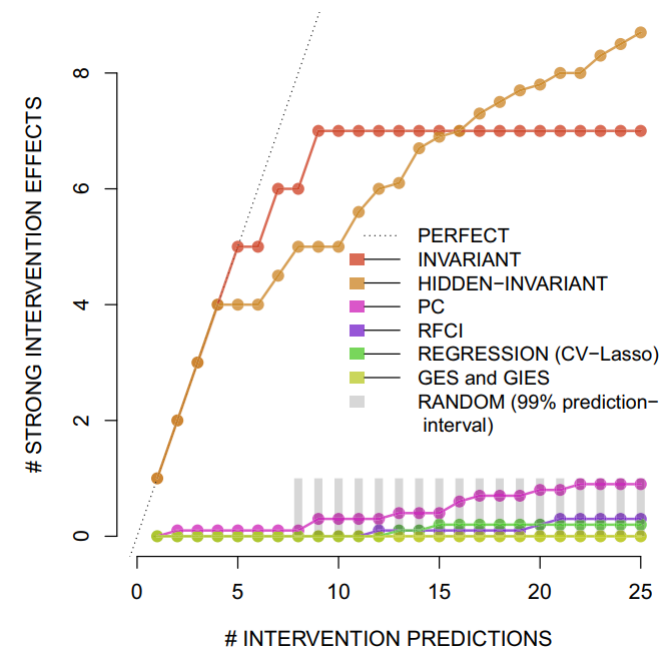
Greedy SP

Given a permutation π we can construct a graph G such that:

- G has n nodes $v_1, v_2, \ldots v_n$ and,

- There is a edge $v_i v_j$ for which i < j and $\pi(i) > \pi(j)$

Sparsest Permutation Algorithm (Raskutti, Stat 2018):

1. Construct the DAG, given the permutation

$$(\pi(j), \pi(k)) \in E_\pi \iff j < k \text{ and } X_{\pi(j)} \not\perp X_{\pi(k)} \mid X_{\{\pi(1), \pi(2), \ldots, \pi(k-1)\} \setminus \{\pi(j)\}} \text{ in } \mathbb{P}.$$

2. Select the π that yields the smallest number of edges

Additional reading: Wang et al., NeurIPS 2017