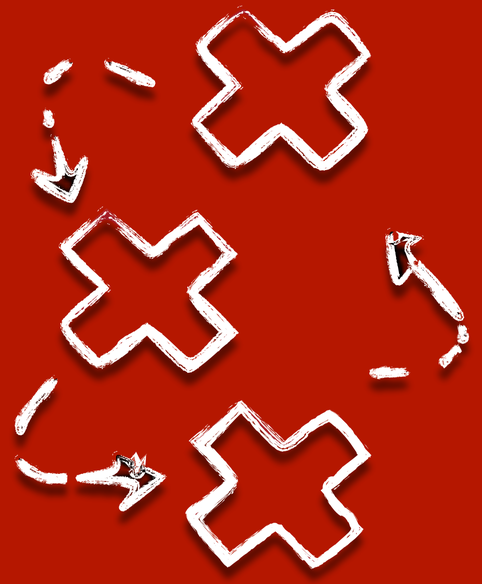


Causal Data Science

Lecture 12.2: Recap of the course

Lecturer: Sara Magliacane

UvA - Spring 2022

Summary of the course

6/02/2023	Introduction
9/02/2023	Probability recap
13/02/2023	Graphical models, d-separation
16/02/2023	Causal graphs, Interventions, SCMs
20/02/2023	Covariate adjustment: backdoor criterion
23/02/2023	Covariate Frontdoor criterion, Instrumental variables
27/02/2023	Counterfactuals, potential outcomes, estimating causal effects 1
2/03/2023	Estimating causal effects 2 (matching, IPW)
6/03/2023	Constraint based structure learning
9/03/2023	Score based structure learning, restricted models
13/03/2023	Do-calculus, transportability, Joint Causal Inference
16/03/2023	Causality-inspired ML, recap of the course

Background on causal graphs

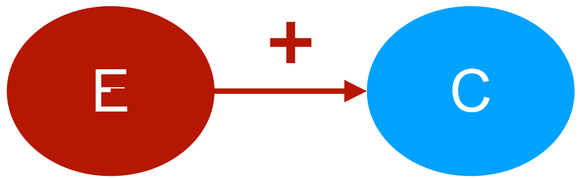
We know the causal graph, how do we estimate causal effects?

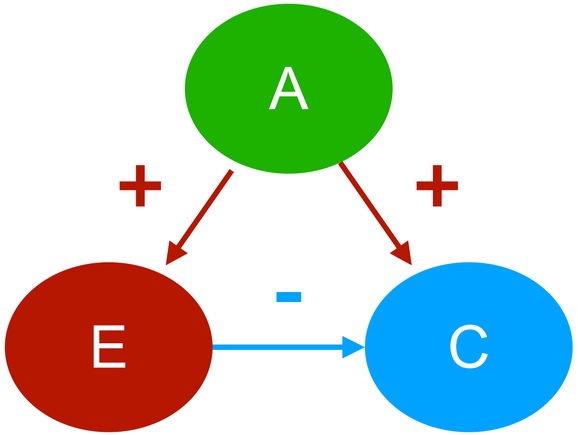
What happens if the graph is unknown?

Cutting edge research

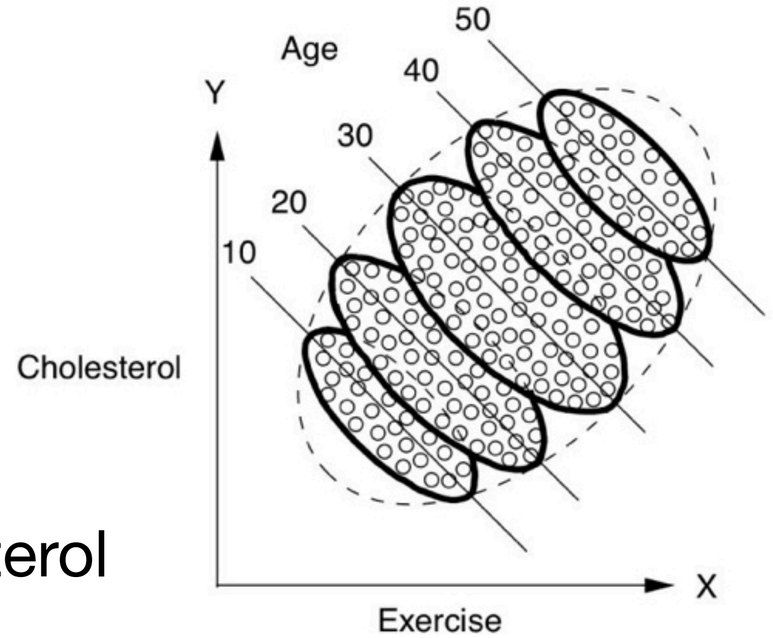
Motivation: Simpson's paradox - confounding

Let's assume we have **observational data** (e.g. data collected by hospitals)





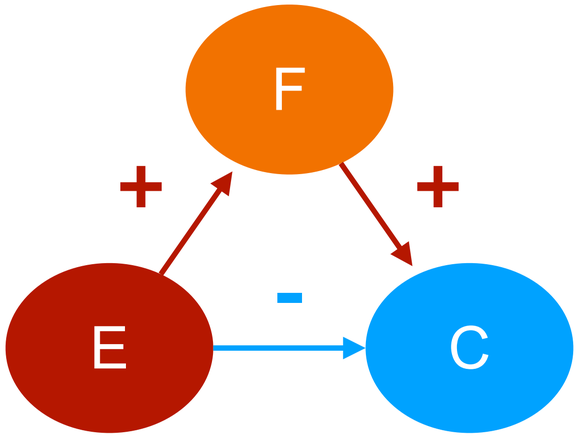
Exercise **decreases** cholesterol



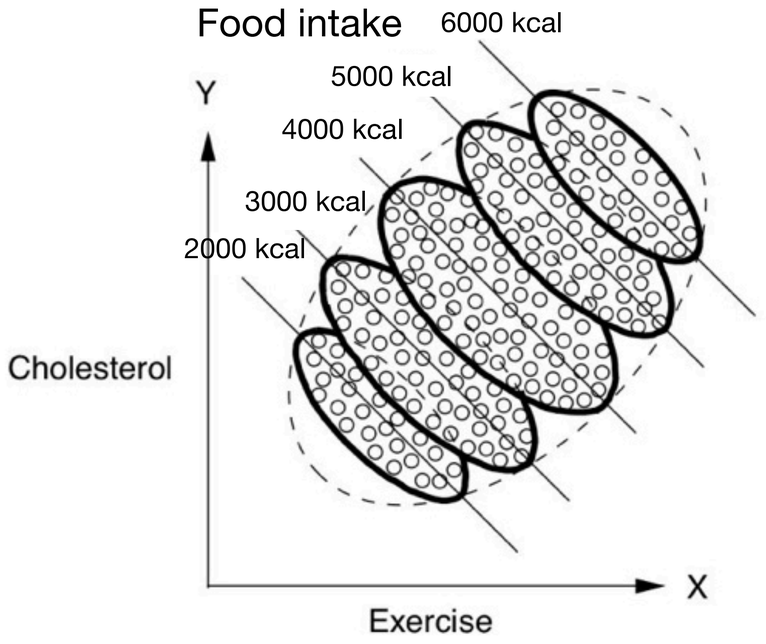
From the Book of Why [Pearl 2018]

Motivation: Shall we always just control on everything?

In alternative universe:



Exercise **increases** cholesterol



Takeaway: The covariates to adjust for can be different based on the graph.

Edited based on the Book of Why [Pearl 2018]

Bayesian networks (BNs)

- We have a set of random variables $X_1, \ldots, X_p$ with joint $p(X_1, \ldots, X_p)$
- We have a DAG G, s.t. **each random variable X_i** is represented by **node i**

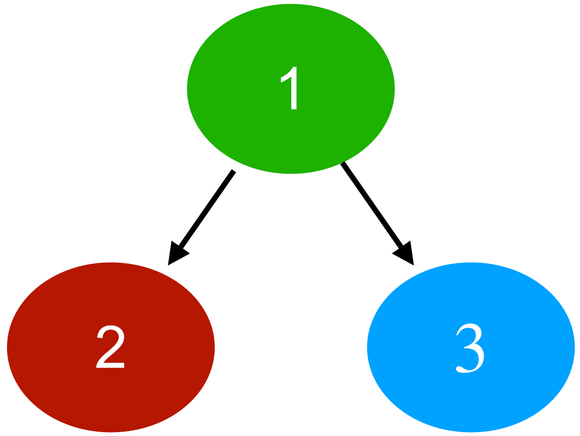
- We then say $p(X_1, \ldots, X_p)$ **factorizes over G** if

$$p(X_1, \ldots, X_p) = \prod_{i \in V} p(X_i \mid \mathbf{X}_{\mathrm{pa}(i)})$$

They can help simplify the factorisation

We can easily read conditional independences

They can represent causal models



D-separation (summary)

- A **path** between **i and j** is **blocked by** $\mathbf{A} \subseteq \mathbf{V}$ at least one condition holds:
  - There is a *non-collider* on the path that is in $\mathbf{A}$**, or**
  - There is a *collider* k on the path, but $k \notin \mathbf{A}$ and $\mathrm{Desc}(k) \cap \mathbf{A} = \varnothing$
- Otherwise it is **active**
- Nodes **i and j** is **d-separated by** A if **all paths** between i, j are **blocked**
  - We denote d-separation as $i \perp j \,|\, \mathbf{A}$

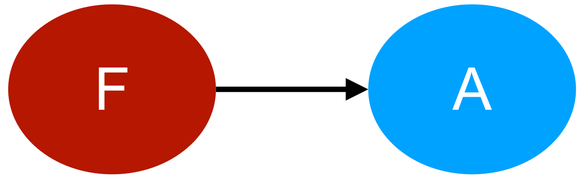
We mostly assumed:

$$\mathbf{A} \perp \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

BNs vs causal BNs - example

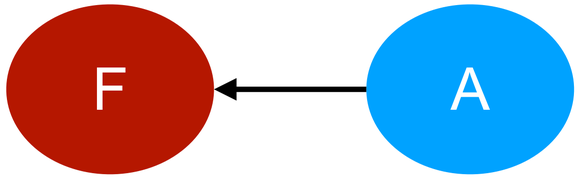
- Fire (F) and Alarm (A) with $p(F, A)$ and $A \not\perp\!\!\!\perp F$ can be factorized as:

$p(F, A) = p(F)\, p(A|F)$



CAUSAL
(lighting a fire triggers alarm)

$p(F, A) = p(A)\, p(F|A)$



NOT-CAUSAL
(triggering alarm does not light a fire)

Causal Bayesian networks

- We introduced a new operator that can represent a **hypothetical intervention** on the whole population, i.e. a perturbation of the system:

$$\mathrm{do}(X = x) \quad \ldots \text{ force X to the value x for all samples}$$

- If $(G = (\mathbf{V}, \mathbf{E}), p)$ is a Bayesian network and if for all $\mathbf{W} \subset \mathbf{V}$:

$$p(X_{\mathbf{V}} \,|\, \mathrm{do}(X_{\mathbf{W}} = x_{\mathbf{W}})) = \prod_{i \in \bar{V} \setminus \bar{W}} p(X_i \,|\, X_{Pa(i)}) \cdot \mathbb{1}(X_{\bar{w}} = x_{\bar{w}})$$

then (G, p) is a **causal Bayesian network**

Parents are now direct causes

Structural causal models (SCMs)

- Let (G, p) be a **causal** Bayesian network
- We can write each variable X_i for $i \in \mathbf{V}$ as a **function of its parents** in G and a **noise term** ϵ_i in a **structural equation:**

$$X_i \leftarrow h_i(X_{\mathrm{Pa}(i)}, \epsilon_i)$$

often linear often Gaussian

- We assume all noises are **independent of each other** $\forall i \neq j : \epsilon_i \perp\!\!\!\perp \epsilon_j$

Summary of the course

6/02/2023	Introduction
9/02/2023	Probability recap
13/02/2023	Graphical models, d-separation
16/02/2023	Causal graphs, Interventions, SCMs
20/02/2023	Covariate adjustment: backdoor criterion
23/02/2023	Covariate Frontdoor criterion, Instrumental variables
27/02/2023	Counterfactuals, potential outcomes, estimating causal effects 1
2/03/2023	Estimating causal effects 2 (matching, IPW)
6/03/2023	Constraint based structure learning
9/03/2023	Score based structure learning, restricted models
13/03/2023	Do-calculus, transportability, Joint Causal Inference
16/03/2023	Causality-inspired ML, recap of the course

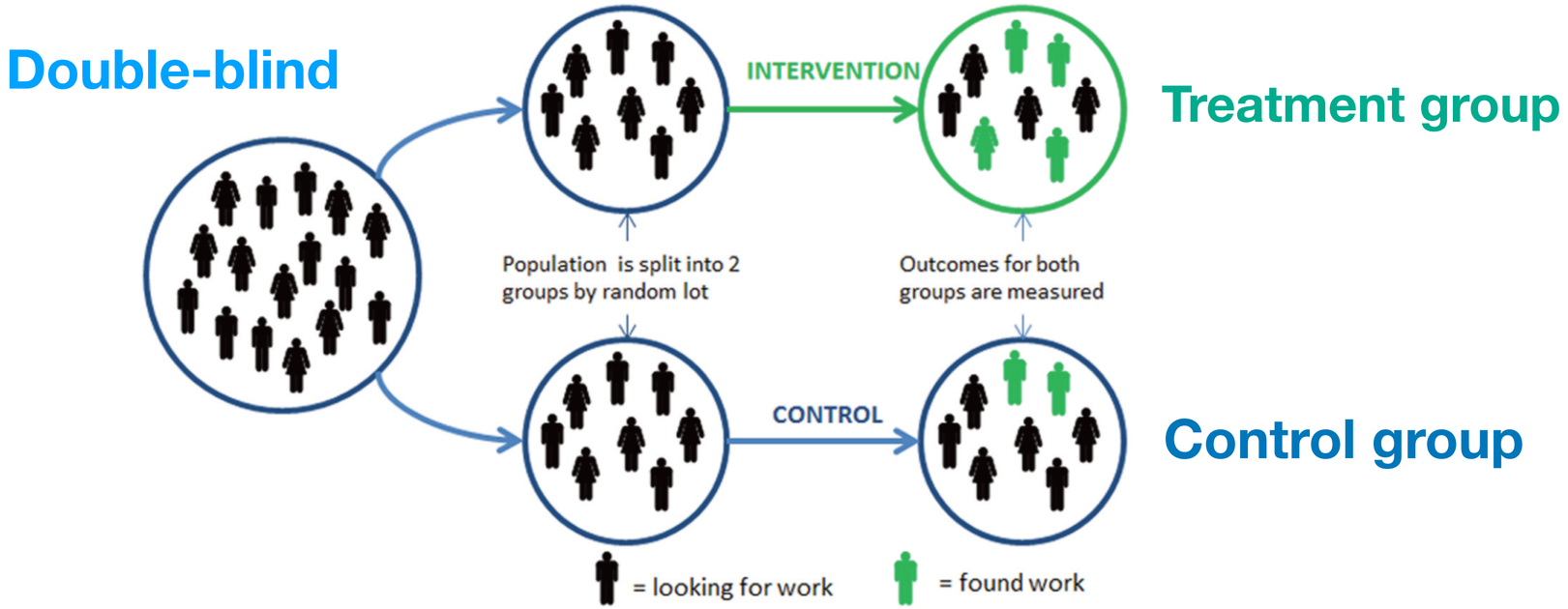
We know the causal graph, how do we estimate causal effects?

$$p(x_j \mid \mathrm{do}(x_i))?$$

EFFECT OF TREATMENT ON OUTCOME

Estimating causal effects: special case RCT

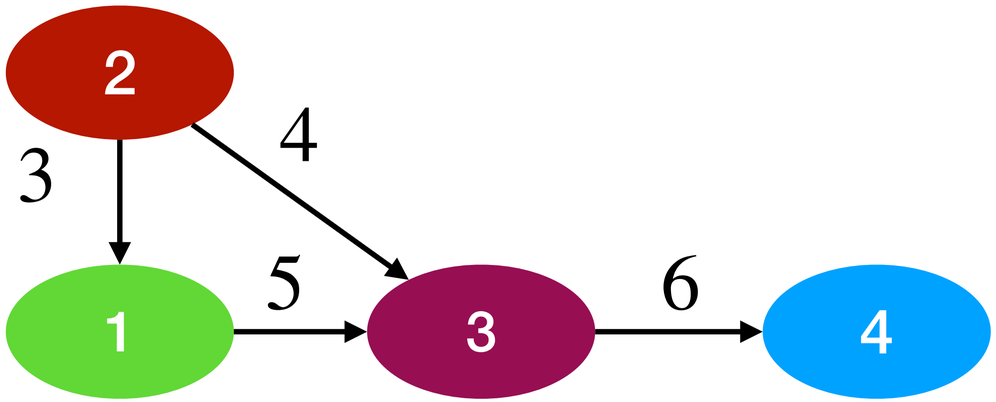
- Randomised Controlled Trials (RCT): intervening on the treatment



https://www.researchgate.net/publication/256031307_Test_Learn_Adapt_Developing_Public_Policy_with_Randomised_Controlled_Trials

Estimating causal effects: special case linear SCMs

- In a linear SCM we estimate the **total average causal effect** of X_i on X_j :
  - For each **directed path from X_i to X_j**, multiply the edge weights
  - Sum the weights from all paths



$$E[X_4 \mid X_2 = 1] - E[X_4 \mid X_2 = 0]$$

$$= 3 \cdot 5 \cdot 6 + 4 \cdot 6 = 114$$

Identification strategies for causal effects

- Given a causal graph G, an **identification strategy** is a formula to estimate an interventional distribution from a combination of observational ones
- **Backdoor criterion, Adjustment criterion**

$$p(x_j \mid \mathrm{do}(x_i)) = \int_{x_{\mathbf{Z}}} p(x_j \mid x_i, x_{\mathbf{Z}}) p(x_{\mathbf{Z}}) dx_{\mathbf{Z}}$$

- **Frontdoor criterion**

$$p(x_j \mid \mathrm{do}(x_i')) = \int_{x_{\mathbf{M}}} p(x_{\mathbf{M}} \mid x_i') \int_{x_i} p(x_j \mid x_{\mathbf{M}}, x_i') p(x_i) dx_i$$

- **Instrumental variables**

Backdoor criterion [Pearl 2009]

- Given a CBN (G, p) with $G = (\mathbf{V}, \mathbf{E})$, a set $\mathbf{Z} \subseteq \mathbf{V} \backslash \{i, j\}$ satisfies the **backdoor criterion** for estimating the causal effect of X_i on X_j with $i \neq j$:
  - $\mathbf{Z}$ does **not contain any descendant of i,** $\mathrm{Desc}(i) \cap \mathbf{Z} = \varnothing$, **and**
  - $\mathbf{Z}$ blocks all **backdoor paths** from i to j (all paths that start with an arrow into $i \leftarrow \ldots \ j$)

The backdoor criterion finds **some (not necessarily all)** valid adjustment sets

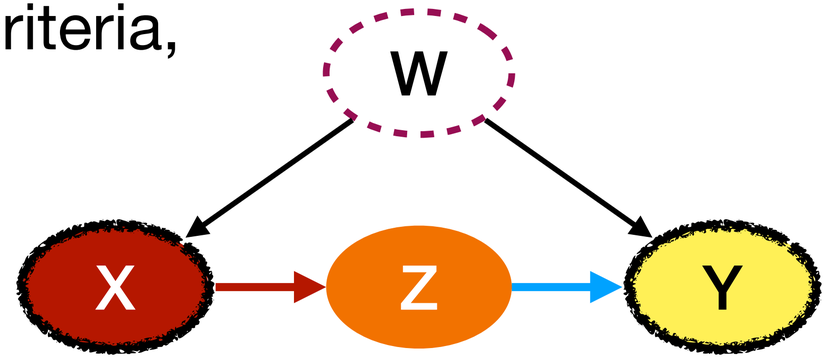
Complete: Adjustment criterion [Shpitser et al, Perković et al]

- Given a CBN (G, p), a set $\mathbf{Z} \subseteq \mathbf{V} \backslash \{i, j\}$ satisfies the **adjustment criterion** for estimating the causal effect of X_i on X_j with $i \neq j$:
  1. $\mathbf{Z}$ does not contain any descendant of **nodes $r \neq i$ on a directed path from i to j**
  2. $\mathbf{Z}$ blocks all paths from i to j that are **not directed paths from i to j**

The adjustment criteria finds **all valid adjustment sets** (but there are other sets that allow identification of total causal effects - e.g. frontdoor criterion)

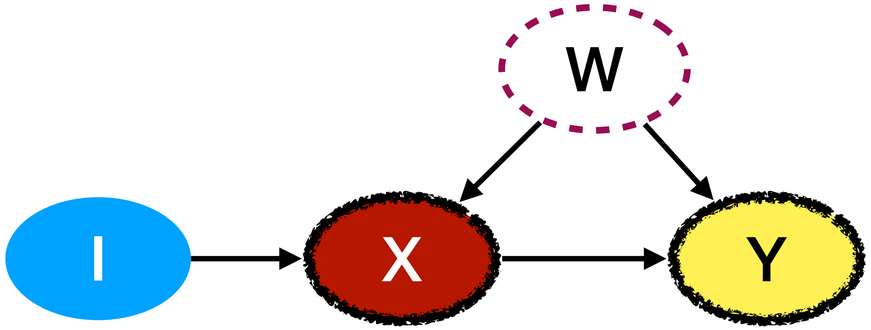
Frontdoor criterion example

- We cannot use the backdoor/adjustment criteria,
- because W is unobserved



- Frontdoor criterion idea:
  1. Find all **mediator variables** $\mathbf{M}$ on the directed paths between X and Y
  2. Estimate effect of X on $\mathbf{M}$ (no unblocked backdoors from $i \leftarrow \ldots \mathbf{M}$)
  3. Estimate effect of $\mathbf{M}$ on Y (i blocks all backdoor paths from $\mathbf{M} \leftarrow \ldots\ j$)
  4. Combine the two effects

Instrumental variables



- We want to estimate the effect of X on Y
- We cannot use the backdoor/adjustment criteria, because W is unobserved
- We cannot use frontdoor because there is no mediator
- We can exploit the **instrumental variable (IV)** I
  - $I \rightarrow X$, but $I \nrightarrow Y$ directly, $I \perp\!\!\!\perp W$

$$\beta = \frac{\text{Cov}(I, Y)}{\text{Cov}(I, X)}$$

(or 2SLS)

Counterfactuals

- Up to now we have been discussing how to estimate the **interventional distribution** $P(X_j \mid \mathrm{do}(X_i))$

- But this does not tell us what would have happen to individuals under different interventions that the ones that were actually performed
  - For example: a patient was treated and they recovered, what would have happened if they were not treated? (**Retrospectively**)

- **Assumption**: the noise variables stay the same

Unit-level counterfactuals for linear SCMs

- Linear SCM S with observed variables $(X_1, \ldots, X_p)$ and noises $(\epsilon_{X_1}, \ldots, \epsilon_{X_p})$
- We can compute counterfactuals for $\mathrm{do}(X_j)$ and unit i with $(x_1^i, \ldots, x_p^i)$:

1. **Abduction:** reconstruct the noise variable values for i using S: $(\hat{\epsilon}_{X_1}^i, \ldots, \hat{\epsilon}_{X_p}^i)$
2. **Action:** If $x_j^i = 0$ in the original data, change the equation for i to $x_j^i \leftarrow 1$,
   else if $x_j^i = 1$, change it to $x_j^i \leftarrow 0$ (the counterfactual assignment)
3. **Prediction:** Recompute $(\hat{x}_1^i, \ldots, \hat{x}_p^i)$ using S and $(\hat{\epsilon}_{X_1}^i, \ldots, \hat{\epsilon}_{X_p}^i)$

Issue: We cannot use **unit-level counterfactuals** to **falsify a wrong causal model**

Unit-level causal effects vs average causal effects

- **Unit-level causal effect:** $Y_i(t=1) - Y_i(t=0)$

- **Fundamental problem of causal inference:** we cannot observe a factual and a counterfactual outcome for each unit.

  - In general the treated population and the untreated population are composed by individuals that are not exactly the same.
    - **SUTVA, consistency, ignorability, positivity**

- **But:** we can estimate the effect from data at a population level

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)]$$

Estimation method: Matching

- We want to estimate the average treatment effect on observational data:

$$\text{ATE} = \mathbb{E}[Y(t=1) - Y(t=0)] = \mathbb{E}[Y \,|\, \text{do}(T=1)] - \mathbb{E}[Y \,|\, \text{do}(T=0)]$$

- **Intuition:** find the most similar couple of patients in terms of covariates $\mathbf{X}$, such that one is in the treatment and the other in the control group
  - Assumption: $\mathbf{X}$ satisfy the backdoor criterion

- **Goal:** discard unmatched units, so we have the same number of units with the same combination of values for $\mathbf{X}$ in treatment and control **(balancing)**

Estimation method: Propensity score matching (PSM)

- **Assumptions**: binary treatment T, $\mathbf{X}$ is valid adjustment set
- **Propensity score:** the probability of getting assigned the treatment

$$\pi := P(T = 1 \mid \mathbf{X} = x)$$

- We then do **matching on propensity scores**

- π encodes all information of $\mathbf{X}$ that is useful for T, i.e. $T \perp\!\!\!\perp \mathbf{X} \mid \pi$
  - If $\mathbf{X}$ has a lot of covariates, it might easier to match for since it's a number
  - π is estimated from data, e.g. with **logistic regression**

Estimation method: Inverse probability weighting (IPW)

- **Inverse probability (of treatment) weighting:** weight by inverse of probability of treatment **received**:
  - For treated $T = 1$: weight by the inverse of $\pi = P(T = 1 \mid \mathbf{X})$
  - For untreated $T = 0$: weight by the inverse of $1 - \pi = P(T = 0 \mid \mathbf{X})$

$$\hat{\mathbb{E}}(Y(t = 1)) = \frac{1}{n} \sum_{i=1}^{n} Y_i \cdot 1\{T = 1\} \cdot \frac{1}{P(T = 1 \mid X_i)}$$

$$\hat{\mathbb{E}}(Y(t = 0)) = \frac{1}{n} \sum_{i=1}^{n} Y_i \cdot 1\{T = 0\} \cdot \frac{1}{P(T = 0 \mid X_i)}$$

Summary of the course

6/02/2023	Introduction
9/02/2023	Probability recap
13/02/2023	Graphical models, d-separation
16/02/2023	Causal graphs, Interventions, SCMs
20/02/2023	Covariate adjustment: backdoor criterion
23/02/2023	Covariate Frontdoor criterion, Instrumental variables
27/02/2023	Counterfactuals, potential outcomes, estimating causal effects 1
2/03/2023	Estimating causal effects 2 (matching, IPW)
6/03/2023	Constraint based structure learning
9/03/2023	Score based structure learning, restricted models
13/03/2023	Do-calculus, transportability, Joint Causal Inference
16/03/2023	Causality-inspired ML, recap of the course

What happens if the graph is unknown?

Causal discovery overview

Constraint-based causal discovery

- Conditional independence tests
- Observational data
- Output: MEC
- SGS, PC

Score-based causal discovery

- Penalised likelihood
- Observational data
- Output: MEC
- GES

Restricted models

- Nonlinear additive noise, Linear Non-Gaussianity
- Observational data
- Output: DAG
- RESIT, LINGAM

Interventional causal discovery / causal invariance

- Observational and Interventional data
- Output: parents of Y, I-MEC
- ICP, JCI

SGS algorithm (Spirtes, Glymour, Scheines)

- Assuming P is Markov and faithful to an unknown graph G

- We can estimate a CPDAG from samples of P in three steps:
  1. Determine the **skeleton**
  2. Determine the **v-structures**
  3. Direct as many remaining edges as possible

- **Note**: the directed parts of the CPDAG will agree with G, but some parts might stay undirected

PC algorithm (Peter Spirtes, Clark Glymour)

- Assuming P is Markov and faithful to an unknown graph G

- We can estimate a CPDAG from samples of P in three steps:
  1. Determine the **skeleton in an optimised way**
  2. Determine the **v-structures**
  3. Direct as many remaining edges as possible

Score-based causal discovery

- **Score-based causal discovery:** find the graph that maximises a **score** $S(G, D)$ (fit of graph G on data D)

- Typically we use **BIC (Bayesian information criterion)**

$$BIC(D, G) := 2 \cdot \log\, p(D \mid G, \theta^{MLE}) - \log(n) \cdot \#\text{parameters}$$

  - **Score equivalence:** all DAGs in a MEC get the same score

  - **Decomposable**, **Local consistency**

Greedy Equivalence Search (GES)

1. Start with empty CPDAG
2. Add edges one by one until local maxima in BIC
3. Remove edges one by one until local maxima in BIC

Phase 1 neighbours ε^+:

Given a a starting equivalence class ε, another class ε' is in the neighbours ε^+ if there exists a DAG $G \in \varepsilon$, such that adding an edge to G results in $G' \in \varepsilon'$

Phase 2 neighbours ε^-: same with removing an edge

Summary of the course

6/02/2023	Introduction
9/02/2023	Probability recap
13/02/2023	Graphical models, d-separation
16/02/2023	Causal graphs, Interventions, SCMs
20/02/2023	Covariate adjustment: backdoor criterion
23/02/2023	Covariate Frontdoor criterion, Instrumental variables
27/02/2023	Counterfactuals, potential outcomes, estimating causal effects 1
2/03/2023	Estimating causal effects 2 (matching, IPW)
6/03/2023	Constraint based structure learning
9/03/2023	Score based structure learning, restricted models
13/03/2023	Do-calculus, transportability, Joint Causal Inference
16/03/2023	Causality-inspired ML, recap of the course

Cutting edge research

Do-calculus (complete identification strategy)

- For disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{W} \subseteq \mathbf{V}$:
- Rule 1: insertion/deletion of observations

$$\mathbf{A} \perp_d \mathbf{B} \,|\, \mathbf{C}, \mathrm{do}(\mathbf{W}) \implies P(X_{\mathbf{A}} \,|\, X_{\mathbf{B}}, X_{\mathbf{C}}, \mathrm{do}(X_{\mathbf{W}})) = P(X_{\mathbf{A}} \,|\, X_{\mathbf{C}}, \mathrm{do}(X_{\mathbf{W}}))$$

- Rule 2: action/observation exchange

$$\mathbf{A} \perp_d I_{\mathbf{B}} \,|\, \mathbf{B}, \mathbf{C}, \mathrm{do}(\mathbf{W}) \implies P(X_{\mathbf{A}} \,|\, \mathrm{do}(X_{\mathbf{B}}), X_{\mathbf{C}}, \mathrm{do}(X_{\mathbf{W}})) = P(X_{\mathbf{A}} \,|\, X_{\mathbf{B}}, X_{\mathbf{C}}, \mathrm{do}(X_{\mathbf{W}}))$$

- Rule 3: insertion/deletion of actions

$$\mathbf{A} \perp_d I_{\mathbf{B}} \,|\, \mathbf{C}, \mathrm{do}(\mathbf{W}) \implies P(X_{\mathbf{A}} \,|\, \mathrm{do}(X_{\mathbf{B}}), X_{\mathbf{C}}, \mathrm{do}(X_{\mathbf{W}})) = P(X_{\mathbf{A}} \,|\, X_{\mathbf{C}}, \mathrm{do}(X_{\mathbf{W}}))$$

Transportability [Bareinboim and Pearl 2016]

- How to combine the data from different **observational** and **experimental** conditions, each conducted on a different population, to estimate a causal effect on a target population?

- Given the true causal graph in the target setting and **the selection diagrams showing the differences in the other settings**, one can find an estimated by applying **do-calculus**

https://www.pnas.org/doi/10.1073/pnas.1510507113

Causality + machine learning (non-exhaustive list)

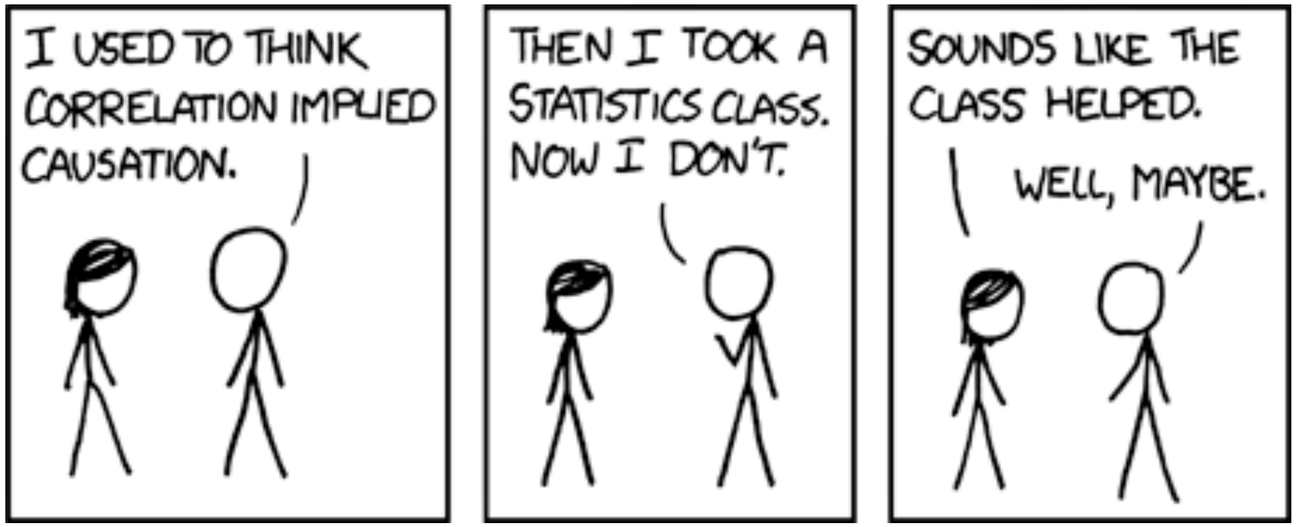
1. Machine learning (ML) helps causality
   - Causal discovery - learning causal graphs from data
   - Causal effect estimation - matching, weighting, double ML
   - (Causal) representation learning

2. Causality (in the most general definition) helps machine learning
   - Robustness, Transfer learning
   - Reinforcement Learning
   - Bias mitigation, fairness

https://arxiv.org/pdf/1705.08821.pdf, https://arxiv.org/pdf/1802.05664.pdf, https://arxiv.org/pdf/1605.03661.pdf, https://crl.causalai.net/, https://www.youtube.com/watch?v=Obuu3w809CI&ab_channel=ConnorJerzak and many many others

Questions??

- If you have any follow-up question use the **Discussions tab in Canvas**



https://xkcd.com/552/