

Causal ML for treatment effect estimation

Valentyn Melnychuk

Institute of AI in Management,
LMU School of Management, LMU Munich

14.10.2026

Input talk @ MZES Social Science Data Lab
University of Mannheim



About our group: Causal ML Lab @ Institute of AI in Management

- Our team:

- Prof. Dr. Stefan Feuerriegel



- 1 director (soon-to-be PostDoc)

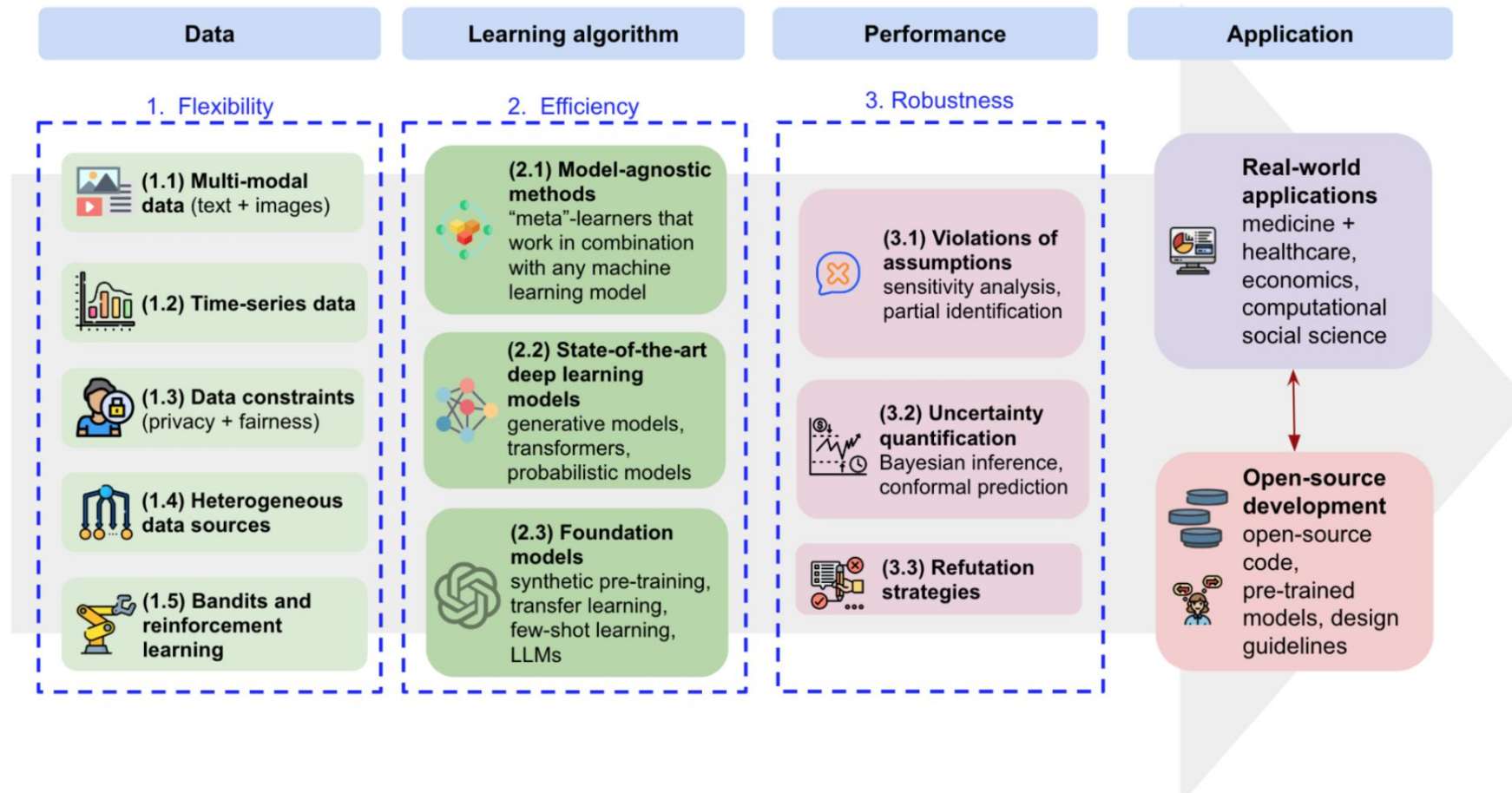


- 8 PhD students



About our group: Causal ML Lab @ Institute of AI in Management

- We develop new machine learning methods for **causal inference** and **decision-making**



About this talk: Causal inference in the Data Lab series

- Earlier sessions at MZES Social Science Data Lab laid some foundations, mostly for **average effects**:

OCT 2022

Erik H. Wang

Matching methods for causal inference with time-series cross-sectional data

Causal setting: difference-in-differences for panel data
Key method: matching estimators

FEB 2024

Asya Magazinnik

Models all the way down

Causal setting: potential outcomes framework

SEP 2024

Achim Ahrens

Robust causal inference using double/debiased machine learning

Causal setting: potential outcomes framework, partially linear model
Key method: double ML, R-learner

FEB 2026

Klara Müller

When events change samples: Disentangling Causal Effects from Compositional Bias in Quasi-Experimental Designs

Causal setting: sensitivity models to unobserved confounding

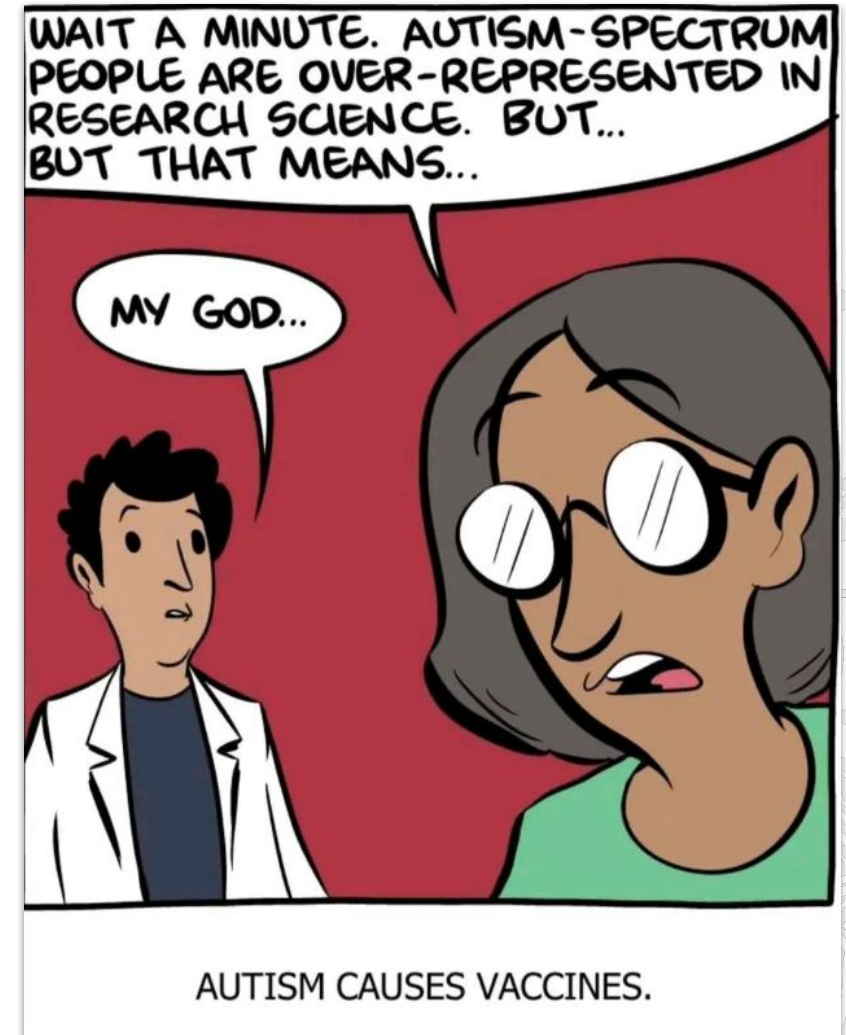
Today: we shift our attention from average to **individualized (heterogeneous) treatment effects**.

About this talk: Outlook

1. **Introduction**: From correlations to causal questions (about 15 min)
2. **Causal assumptions**: What must hold before any ML can be applied (about 20 min)
3. **ML and estimation**: Meta-learners for heterogeneous effects (about 25 min)
4. **Causal ML in practice**: Workflow, software, checklist (about 7 min)
5. **Extensions and open challenges**: Followed by discussion (about 5 + 15 min)

1. Introduction

- Causal machine learning
- Rungs of causal questions
- Motivational example: Campaign contact and turnout
- From prediction to policy
- Problem formulation
- Fundamental problem of causal inference
- Spectrum of causal estimands



Introduction: Causal Machine Learning

Ambiguity of the definition. “Causal Machine Learning” is both:

- causal inference used for machine learning

Causal inference concepts



ML / DL problems

- Explainability
- Fairness
- Algorithmic recourse
- Robustness / domain adaptation
- ...

- machine learning used for causal inference

Causal inference problems

- Treatment effect estimation
- Counterfactual inference
- Causal discovery
- ...



ML / DL tools



Introduction: Causal Machine Learning

Ambiguity of the definition. “Causal Machine Learning” is both:

- causal inference used for machine learning

Causal inference concepts



ML / DL problems

- Explainability
- Fairness
- Algorithmic recourse
- Robustness / domain adaptation
- ...

- machine learning used for causal inference

Causal inference problems

- Treatment effect estimation
- Counterfactual inference
- Causal discovery
- ...



ML / DL tools



Introduction: Rungs of causal questions

- Treatment effect estimation is one of the main **causal inference problems**

Pearl's ladder of causation

Level	Typical activity	Typical questions	Social-science examples
1. Association:	Seeing	What is? How would seeing X change my belief in Y?	What does a pre-election poll tell us about the election result? Are university graduates more likely to vote?
2. Intervention:	Doing, intervening	What if? What if I do X?	What if every registered voter received a get-out-the-vote letter? What if compulsory schooling were one year longer?
3. Counterfactuals:	Imagining, retrospection	Why? Was it X that caused Y? What if I had acted differently?	Would this citizen have voted had she not been canvassed? Would she earn more today had she gone to university?

- **Randomized experiments (L2 data)** are scarce (field, survey, lab). Often expensive, unethical or infeasible: we cannot randomize education, job loss, divorce or migration
- Abundance of **observational data (L1 data)**: surveys, registers, digital trace data
- Recent advances in ML / DL provide many tools for flexible estimation

Introduction: Rungs of causal questions

- Treatment effect estimation is one of the main **causal inference problems**

Pearl's ladder of causation

Level	Typical activity	Typical questions	Social-science examples	Traditional ML
1. Association:	Seeing	What is? How would seeing X change my belief in Y?	What does a pre-election poll tell us about the election result? Are university graduates more likely to vote?	
2. Intervention:	Doing, intervening	What if? What if I do X?	What if every registered voter received a get-out-the-vote letter? What if compulsory schooling were one year longer?	
3. Counterfactuals:	Imagining, retrospection	Why? Was it X that caused Y? What if I had acted differently?	Would this citizen have voted had she not been canvassed? Would she earn more today had she gone to university?	

- **Randomized experiments (L2 data)** are scarce (field, survey, lab). Often expensive, unethical or infeasible: we cannot randomize education, job loss, divorce or migration
- Abundance of **observational data (L1 data)**: surveys, registers, digital trace data
- Recent advances in ML / DL provide many tools for flexible estimation

Introduction: Rungs of causal questions

- Treatment effect estimation is one of the main **causal inference problems**

**Pearl's
ladder of
causation**

Level	Typical activity	Typical questions	Social-science examples
1. Association:	Seeing	What is? How would seeing X change my belief in Y?	What does a pre-election poll tell us about the election result? Are university graduates more
2. Intervention:	Doing, intervening	What if? What if I do X?	What if every registered voter received a get-out-the-vote letter? What if compulsory schooling were one year longer?
3. Counterfactuals:	Imagining, retrospection	Why? Was it X that caused Y? What if I had acted differently?	Would this citizen have voted had she not been canvassed? Would she earn more today had she gone to university?

Causal ML

- Randomized experiments (L2 data)** are scarce (field, survey, lab). Often expensive, unethical or infeasible: we cannot randomize education, job loss, divorce or migration
- Abundance of **observational data (L1 data)**: surveys, registers, digital trace data
- Recent advances in ML / DL provide many tools for flexible estimation

Introduction: Rungs of causal questions

- Treatment effect estimation is one of the main **causal inference problems**

Pearl's ladder of causation

Level	Typical activity	Typical questions	Social-science examples
1. Association:	Seeing	What is? How would seeing X change my belief in Y?	What does a pre-election poll tell us about the election result? Are university graduates more likely to vote?
2. Intervention:	Doing, intervening	What if? What if I do X?	What if every registered voter received a get-out-the-vote letter? What if compulsory schooling were one year longer?
3. Counterfactuals:	Imagining, retrospection	Why? Was it X that caused Y? What if I had acted differently?	Would this citizen have voted had she not been canvassed? Would she earn more today had she gone to university?

Causal ML (today)

- **Randomized experiments (L2 data)** are scarce (field, survey, lab). Often expensive, unethical or infeasible: we cannot randomize education, job loss, divorce or migration
- Abundance of **observational data (L1 data)**: surveys, registers, digital trace data
- Recent advances in ML / DL provide many tools for flexible estimation

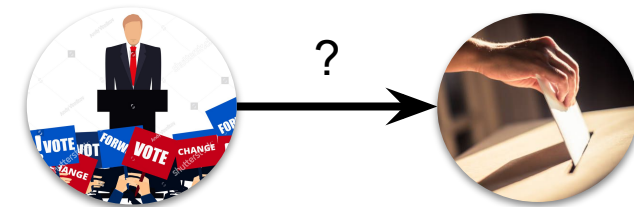
Introduction: Motivational example

- Why is answering interventional/counterfactual questions from observational data challenging?
- Consider the following study:

Observational data from post-election survey:

- contacted by a party campaign before the election (binary treatment)
- voted in the election (outcome)

Aim: effect of campaign contact on the chance of voting



-
- **Evidence:** Respondents who were contacted turn out at much higher rates than those who were not
 - **P.S.** Is there something we didn't account for?

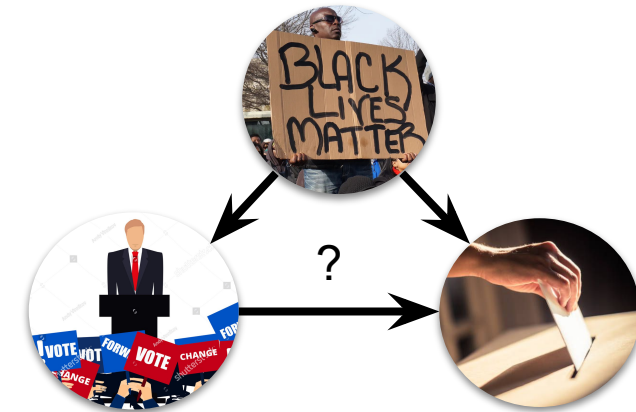
Introduction: Motivational example

- Why is answering interventional/counterfactual questions from observational data challenging?
-> **Hidden confounding**
- Consider the following study:

Observational data from post-election survey:

- contacted by a party campaign before the election (binary treatment)
- voted in the election (outcome)
- **political interest, past turnout, party membership, age (covariates)**

Aim: effect of campaign contact on the chance of voting



-
- **Evidence:** Interest and past turnout drive both contact and turnout, **no direct association** between contact and turnout
 - **Comparing with the previous slide:** After considering pre-treatment covariates, the spurious association disappeared

Introduction: From prediction to policy

- Causal ML allows us to move from **prediction** (questions like: what is the risk of ... ?) to **prescription** (questions like: what works, and for whom?)

Traditional ML: prediction

- **"Umbrella" problem:** Will it rain so that I should take an umbrella? A good forecast is all you need to decide!
- Who is at risk of long-term unemployment? (algorithmic profiling of job seekers)
- Who is likely to vote? (turnout models in polling)
- **Checked against observed outcomes:** accuracy, AUC, test error

Causal ML: prescription

- **"Rain dance" problem:** Does dancing cause rain? You need to know the effect of the action!
- Which job seekers would benefit from which training programme?
- Which citizens would be mobilized by a canvassing visit?
- **Never directly checkable:** individual effects are unobserved

Good prediction ≠ good prescription: the people most likely to remain unemployed are not necessarily the people a programme helps most.

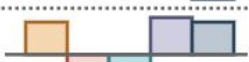
Introduction: Problem formulation

- Given i.i.d. observational dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n \sim \mathbb{P}(X, A, Y)$

- X covariates (e.g., age, education, past turnout, political interest, ...)
- A (binary) treatment (e.g., contacted (1) or (0) not by a campaign)
- Y factual outcome (e.g., voted (1) or (0) not voted)

Patient	Covariates X	Treatment A	Outcome	
			$Y = Y(0)$	$Y = Y(1)$
		0	0	
		1		0
		1		1
...	...	...	...	...

- We want to predict:
 - treatment effects:** $Y(1) - Y(0)$
 - potential outcomes:** $Y(0)$, $Y(1)$

Patient	Covariates X	Potential outcomes		Treatment effect $Y(1) - Y(0)$
		$Y(0)$	$Y(1)$	
		?	?	?
		?	?	?
...	...	...	...	...

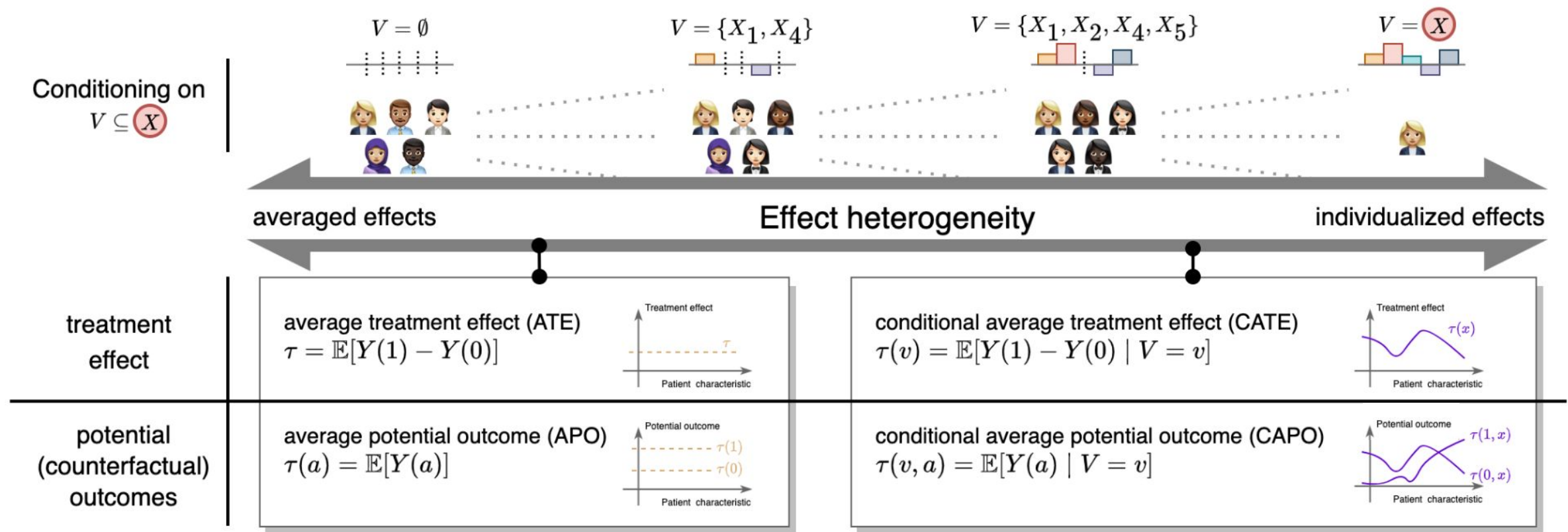
Introduction: Fundamental problem of causal inference

- **Both** potential outcomes (factual and counterfactual) are never observed for any individual → individual treatment effects are never observed
- Potential outcomes are only observed for **parts of the population** → **covariate shift**: we see $Y(1)$ mostly for older, habitual voters (whom campaigns target) and $Y(0)$ mostly for young, disengaged citizens

Patient	Covariates X	Treatment A	Outcome $Y = Y(0)$	Outcome $Y = Y(1)$
		0	0	
		1		0
		1		1
...	...	...	...	...

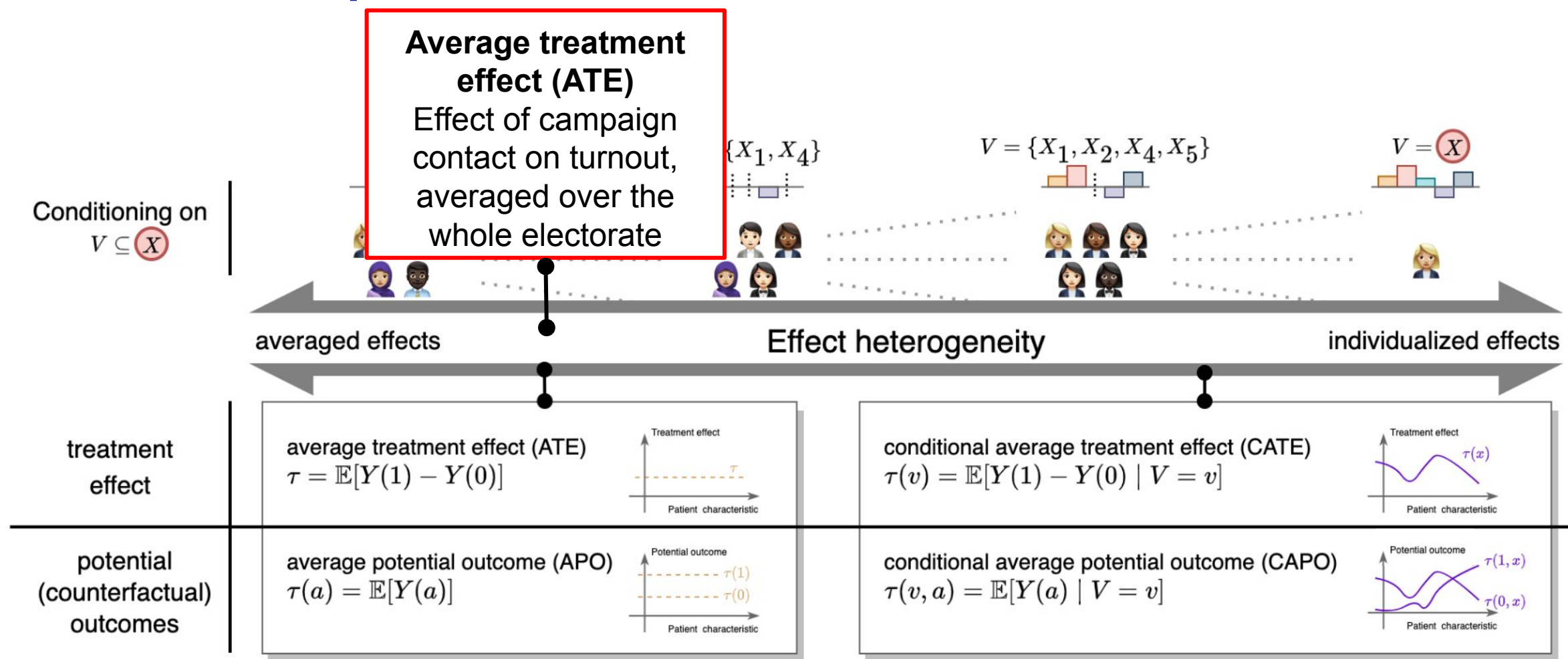
!! Causal inference is fundamentally a **missing-data problem**: half of the potential outcomes are **missing**, and **not at random**.

Introduction: Spectrum of causal estimands



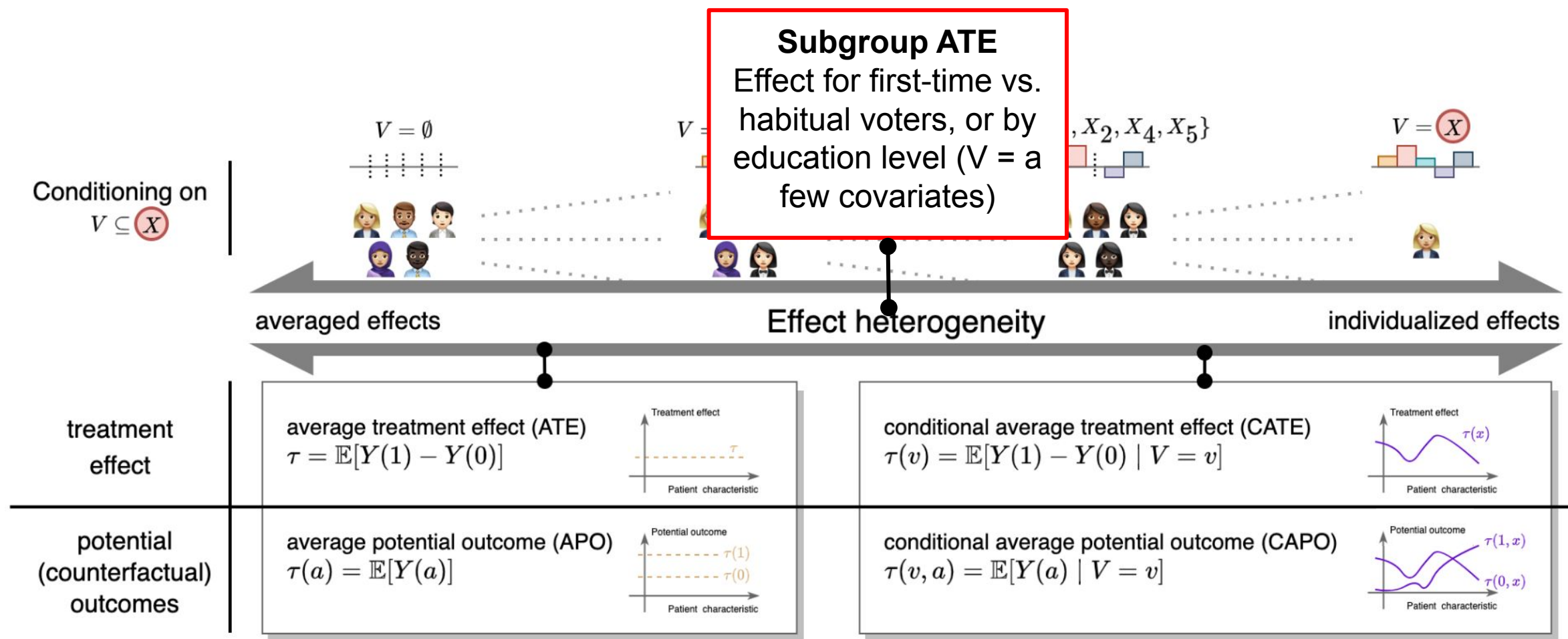
Feuerriegel, Stefan, et al. "Causal machine learning for predicting treatment outcomes." Nature Medicine 30.4 (2024): 958-968.

Introduction: Spectrum of causal estimands



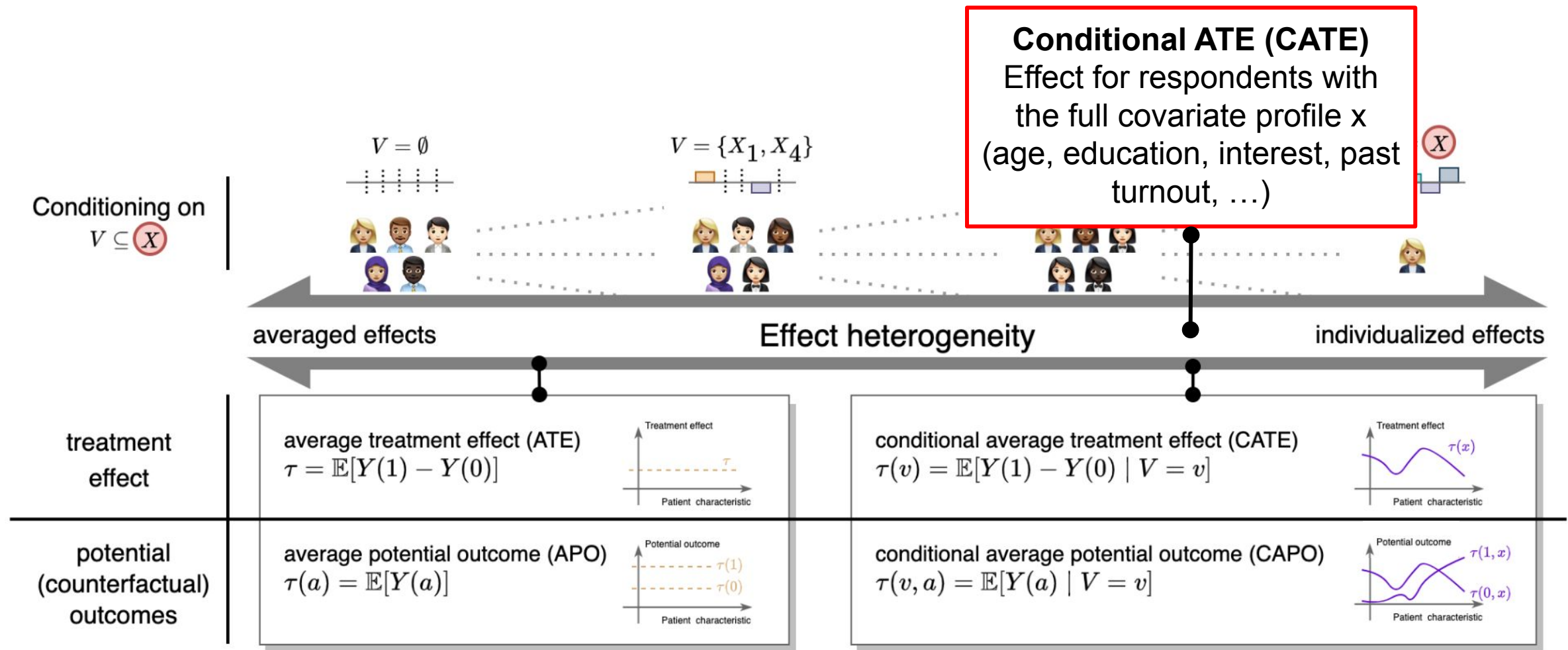
Feuerriegel, Stefan, et al. "Causal machine learning for predicting treatment outcomes." Nature Medicine 30.4 (2024): 958-968.

Introduction: Spectrum of causal estimands



Feuerriegel, Stefan, et al. "Causal machine learning for predicting treatment outcomes." Nature Medicine 30.4 (2024): 958-968.

Introduction: Spectrum of causal estimands



Feuerriegel, Stefan, et al. "Causal machine learning for predicting treatment outcomes." Nature Medicine 30.4 (2024): 958-968.

Introduction: Why effect heterogeneity matters?

Testing theories

- Who benefits most from college?
- Brand & Xie (2010) found evidence of **negative selection**: Those least likely to attend may gain the most.

Brand, Jennie E., and Yu Xie. "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education." *American sociological review* 75.2 (2010): 273-302.

Finding hidden subgroups

- How can we tackle transphobia in a targeted way?
- Door-to-door canvassing was shown to be effective on average (Broockman & Kalla, 2016): An ML meta-learner can further find subgroups where canvassing is the most / least effective

Broockman, David, and Joshua Kalla. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352.6282 (2016): 220-224.

Targeting policies

- How to assign training programs?
- Swiss job-search programs: employment effects differ strongly across unemployed workers, which matters for who should be assigned to which program.

Knaus, Michael C., Michael Lechner, and Anthony Strittmatter. "Heterogeneous employment effects of job search programs: A machine learning approach." *Journal of Human Resources* 57.2 (2022): 597-636.



"ATE = 0" does not mean "no effect for anyone": An average effect can hide opposite effects in different groups

2. Causal assumptions

- Frameworks
- Potential outcomes framework (Neyman-Rubin)
- Structural causal model (SCM)
- Causal diagrams
- Equivalence of the frameworks
- Case study: Effects of alcohol consumption

This keeps happening. How heavy are cats?

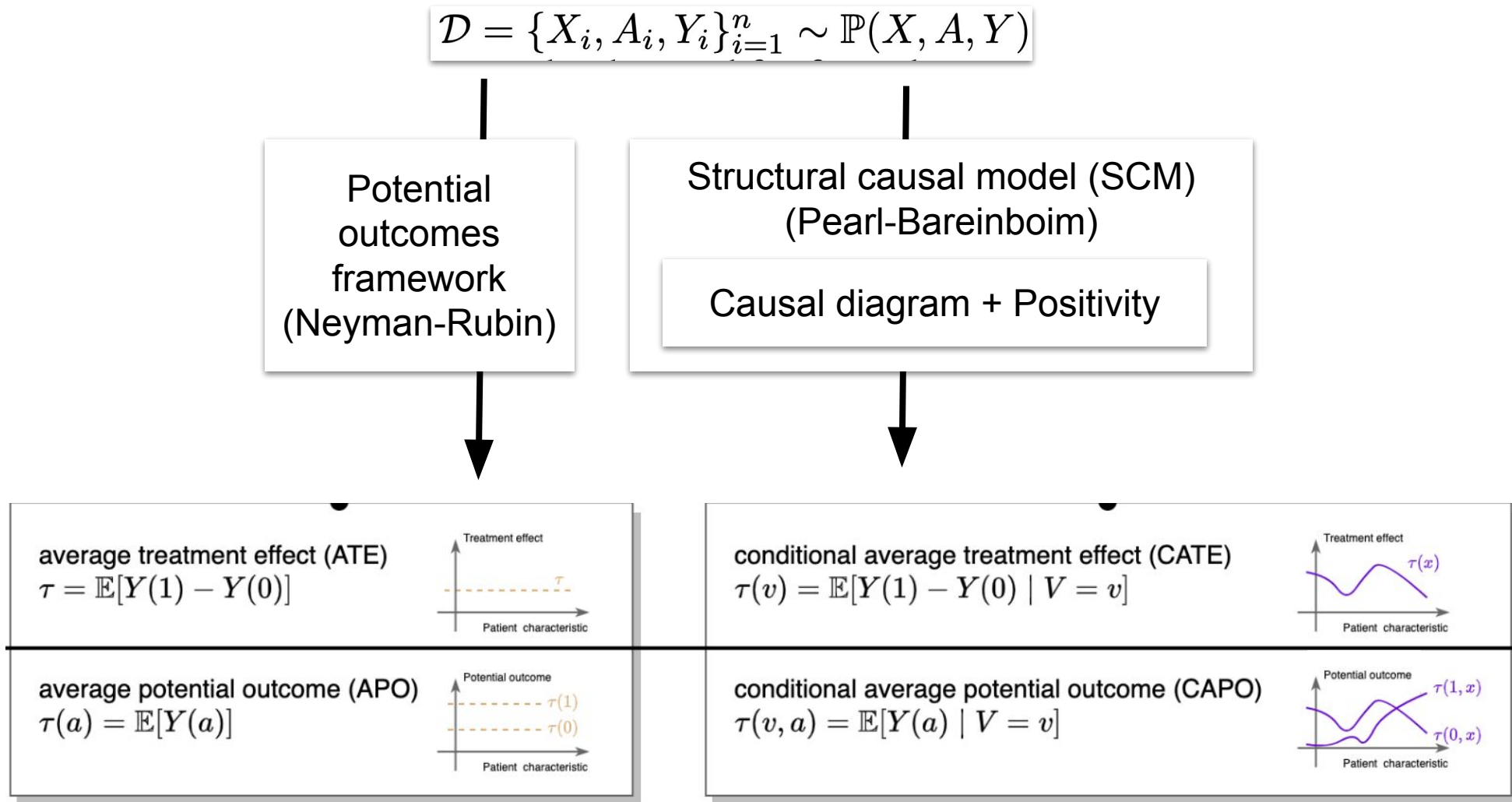


Causal assumptions: Philosophy

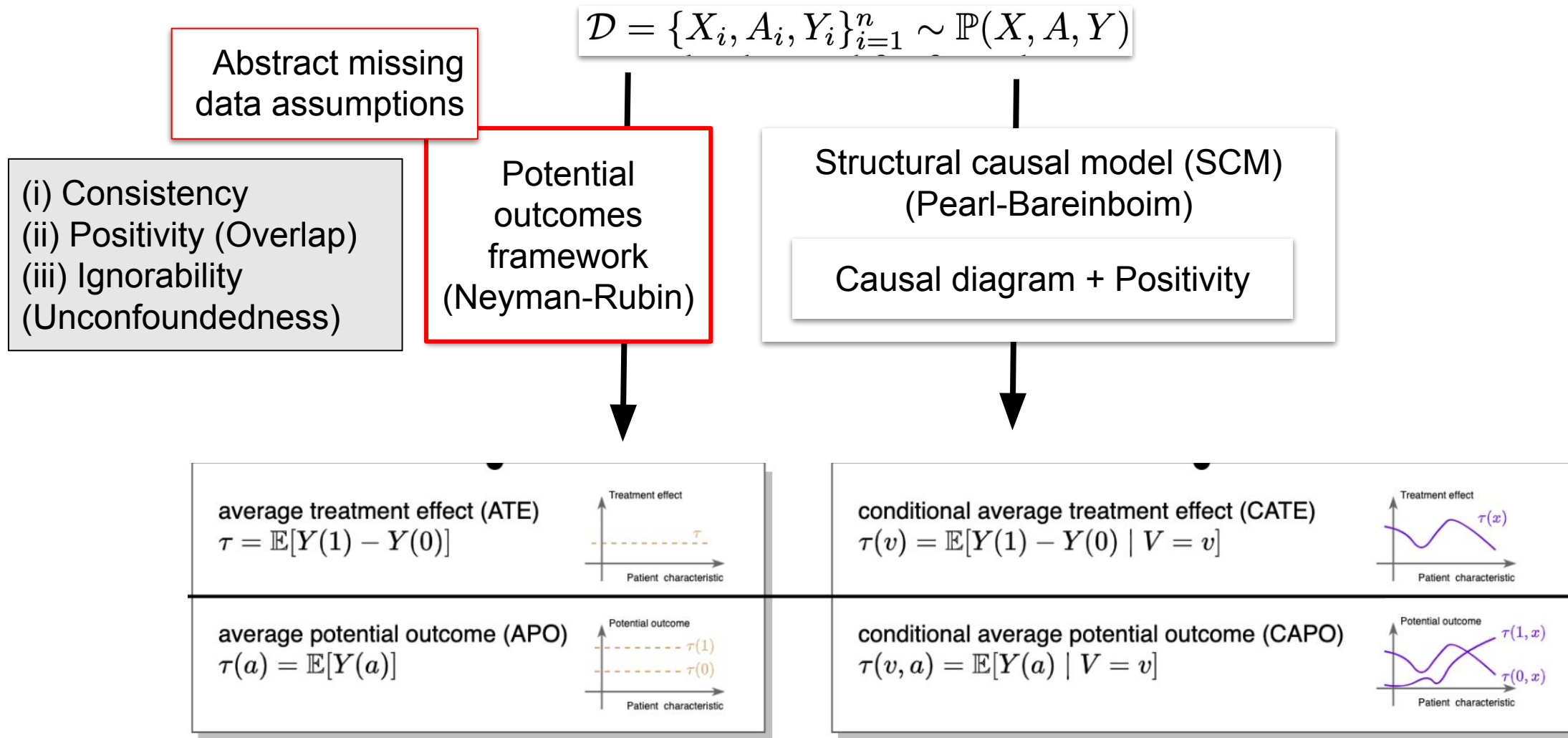
“The credibility of inference decreases with the strength of the assumptions maintained.”

Manski, C. F. (2003). Partial identification of probability distributions, volume 5. Springer.

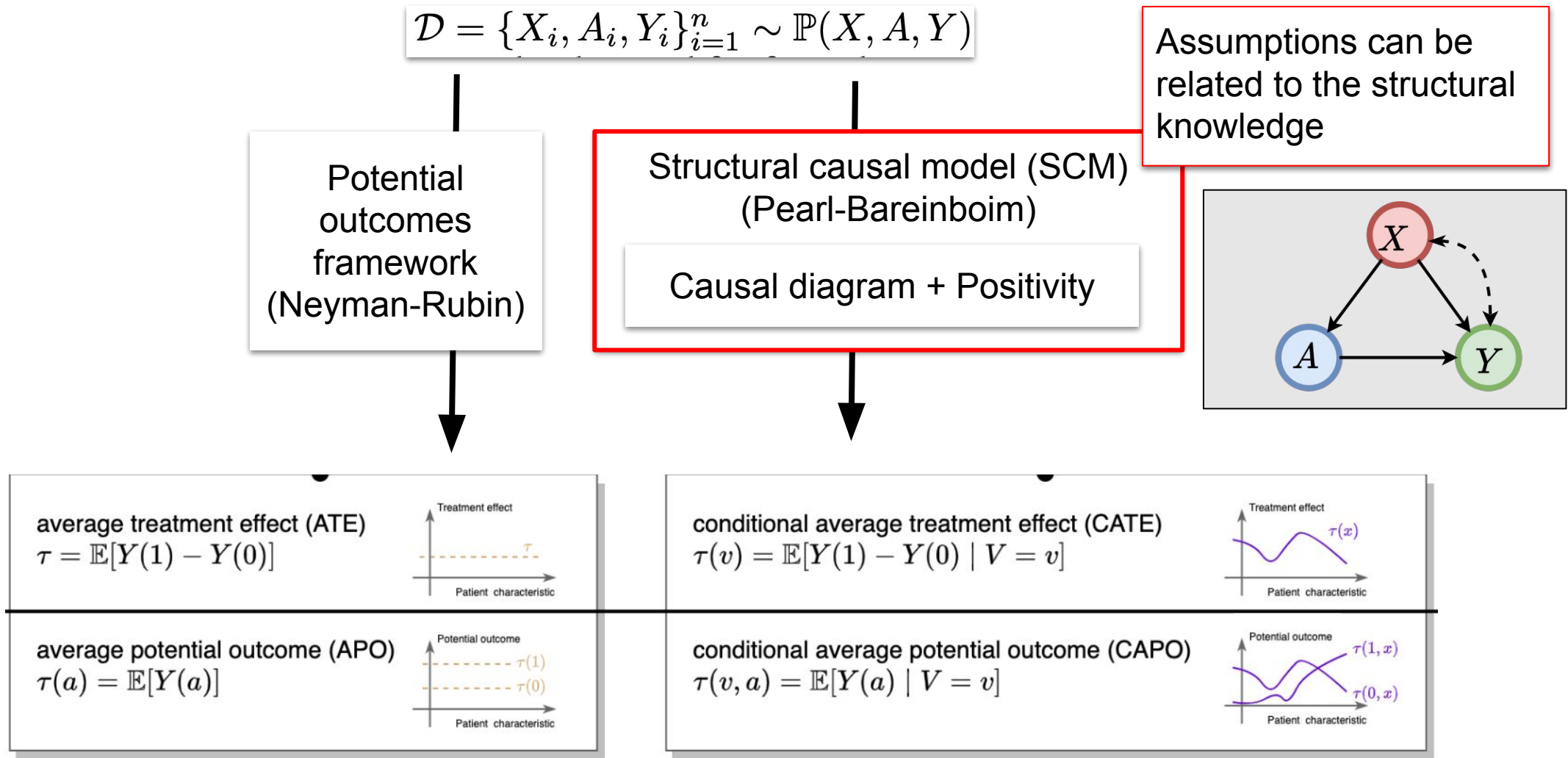
Causal assumptions: Frameworks



Causal assumptions: Frameworks



Causal assumptions: Frameworks



Causal assumptions: Frameworks

$$\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n \sim \mathbb{P}(X, A, Y)$$

Potential outcomes framework
(Neyman-Rubin)

=

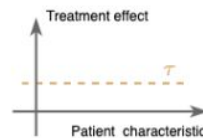
Structural causal model (SCM)
(Pearl-Bareinboim)

Causal diagram + Positivity

Two different languages to describe the same reality

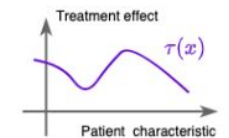
average treatment effect (ATE)

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$



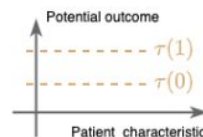
conditional average treatment effect (CATE)

$$\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$$



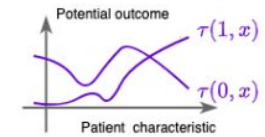
average potential outcome (APO)

$$\tau(a) = \mathbb{E}[Y(a)]$$



conditional average potential outcome (CAPO)

$$\tau(v, a) = \mathbb{E}[Y(a) \mid V = v]$$



Causal assumptions: Potential outcomes framework (Neyman-Rubin)

- (i) **Consistency**

 - **Informal:** Potential outcomes are (sometimes) observed and do not depend on outcomes of others (no interference)
 - If $A = a$, then $Y = Y[a]$
-
- (ii) **Overlap (Positivity)**

 - **Informal:** Both treatments are assigned randomly enough
 - There is always a non-zero probability of receiving/not receiving any treatment, conditioning on the covariates:
 $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$
-
- (iii) **Ignorability (Unconfoundedness)**

 - **Informal:** Confounding issue is resolved, if we condition on enough covariates
 - Current treatment is independent of the potential outcome, conditioning on the covariates:
 $A \perp\!\!\!\perp Y[a] \mid X$ for all a .

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

(i) Consistency	● Informal: Potential outcomes are (sometimes) observed and do not depend on outcomes of others (no interference) ● If $A = a$, then $Y = Y[a]$	Verifiable with observational data? 
(ii) Overlap (Positivity)	● Informal: Both treatments are assigned randomly enough ● There is always a non-zero probability of receiving/not receiving any treatment, conditioning on the covariates: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$	 (by inspecting propensity scores)
(iii) Ignorability (Unconfoundedness)	● Informal: Confounding issue is resolved, if we condition on enough covariates ● Current treatment is independent of the potential outcome, conditioning on the covariates: $A \perp\!\!\!\perp Y[a] \mid X$ for all a.	 (but sensitivity analysis can probe mild violations)

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

Given Assumptions (i) - (iii), **causal quantities** are identifiable from observational data via

- **Regression adjustment (back-door adjustment):** modelling the both conditional outcomes within each covariate profile, then averaging
 - CATE: $\tau(x) = \mu_1(x) - \mu_0(x)$ where $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$
 - ATE: $\tau = \mathbb{E}[\mu_1(X) - \mu_0(X)]$
 - Familiar from: regression with controls, matching, stratification
- **Inverse propensity of treatment weighting (IPTW):** reweight units so that treatment becomes independent of X, using the propensity score $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$
 - CATE: $\tau(x) = \mathbb{E}\left[\left(\frac{A}{\pi_1(X)} - \frac{1-A}{1-\pi_1(X)}\right)Y \mid X = x\right]$
 - ATE: $\tau = \mathbb{E}\left[\left(\frac{A}{\pi_1(X)} - \frac{1-A}{1-\pi_1(X)}\right)Y\right]$
 - Familiar from: survey design weights (Horvitz-Thompson estimator).

Identifiability
with potential
outcomes
framework

$\pi_a(x)$ $\mu_a(x)$ are **nuisance functions**: not of interest in themselves, but needed to get the causal effects.

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

Choosing covariates

- Rule of thumb: We can adjust for (almost) **any pre-treatment covariate**
 - **confounders** ($A \leftarrow X \rightarrow Y$): needed
 - **outcome predictors** ($X \rightarrow Y$): helpful, reduce variance
 - **instruments** ($A \leftarrow X$), **background noise** (X): no bias, but inflate variance
- However, with more covariates, causal estimation becomes harder
- When adjusting for a post-treatment covariate, we are re-inducing a bias:
 - e.g. controlling for "talked about the election with family" when estimating the effect of campaign contact on turnout
 - even a randomized experiment can be **ruined** this way

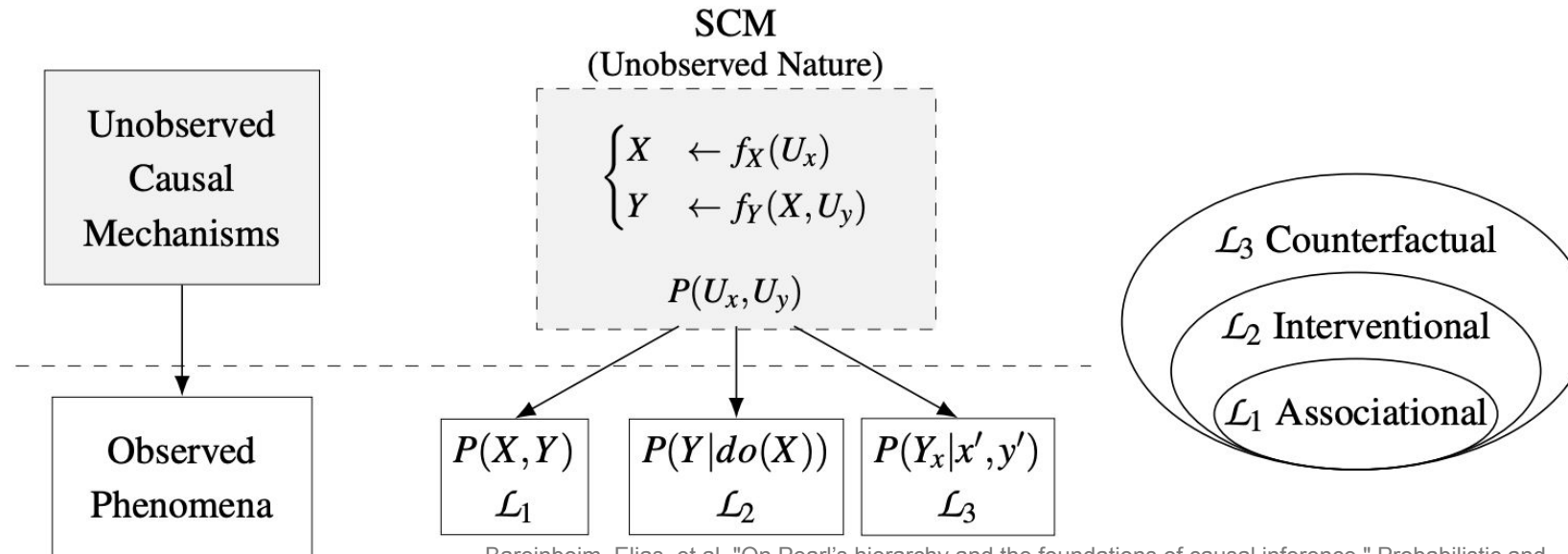
Causal assumptions: Structural causal model (SCM)

- **Informal:** Assuming an SCM = knowing the full nature of the data generating process
- SCM = {observed variables, hidden variables, functional assignments for every observed covariate, probability distribution for hidden variables}

Verifiable with infinite observational data?



SCM



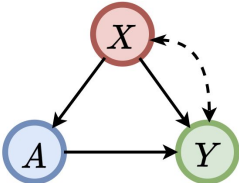
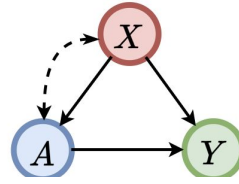
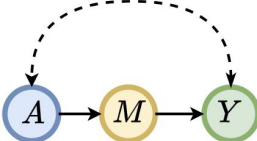
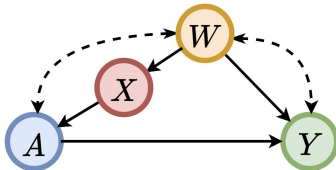
Bareinboim, Elias, et al. "On Pearl's hierarchy and the foundations of causal inference." Probabilistic and causal inference: the works of Judea Pearl. 2022. 507-556.

- All the $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ causal questions can be inferred with the probability calculus, including, **CATE/ATE** and **CAPO/APO** → unnecessary strong assumption

Causal assumptions: Causal diagram

- Sound and complete **identifiability algorithms** (using do-calculus) exist for L2 and L3 causal quantities, e.g.,

Identifiability
with causal
diagrams

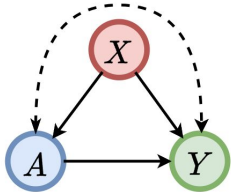
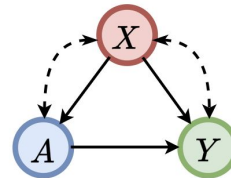
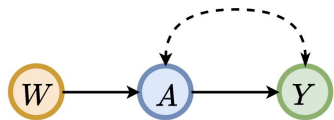
Query:	Causal diagram:	ID:	Formula:
CATE / ATE			- back-door adjustment - IPTW
CATE / ATE			- back-door adjustment - IPTW
ATE			- front-door adjustment
ATE / CATE			- napkin formula

- The theory holds, when covariates are high-dimensional (= **clustered causal diagrams**)

Causal assumptions: Causal diagram

- Sound and complete **identifiability algorithms** (using do-calculus) exist for L2 and L3 causal quantities, e.g.,

Identifiability
with causal
diagrams

Query:	Causal diagram:	ID:	Formula:
CATE / ATE		➔	✗ (Hidden Confounding)
CATE / ATE		➔	✗ (Butterfly-bias)
ATE		➔	✗ (Hidden Confounding)
	+ functional assumption on Y	➔	✓ - Instrumental variable

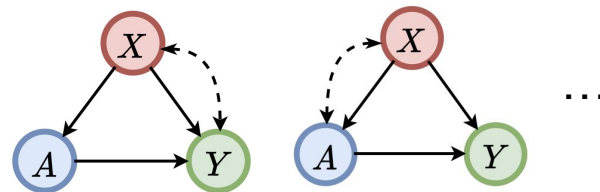
- The theory holds, when covariates are high-dimensional (= **clustered causal diagrams**)

Causal assumptions: Equivalence of the frameworks

- Assumptions of potential outcomes framework are **equivalent** to assuming: (i) causal diagram, to which back-door adjustment can be applied, and (ii) positivity.

(i) Causal diagrams, where:

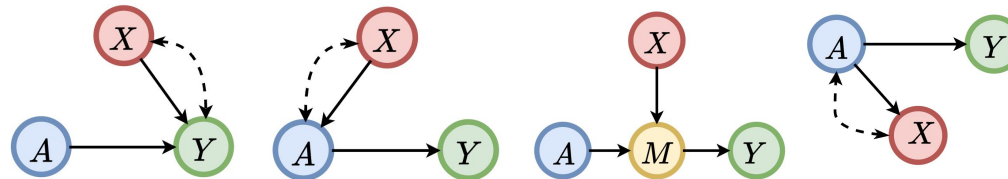
- back-door adjustment for X should be applied



(i) Consistency
(iii) Ignorability

Equivalence of
assumptions

- causal effect is already identifiable and adjustment for X does not create bias



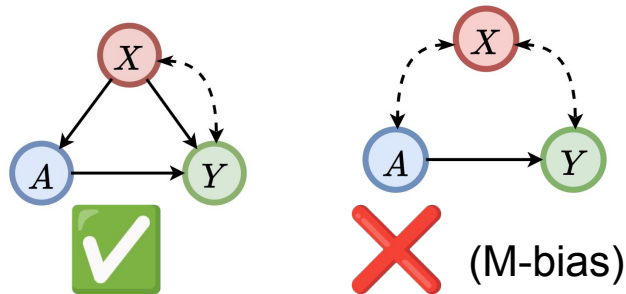
(ii) Positivity



(ii) Positivity

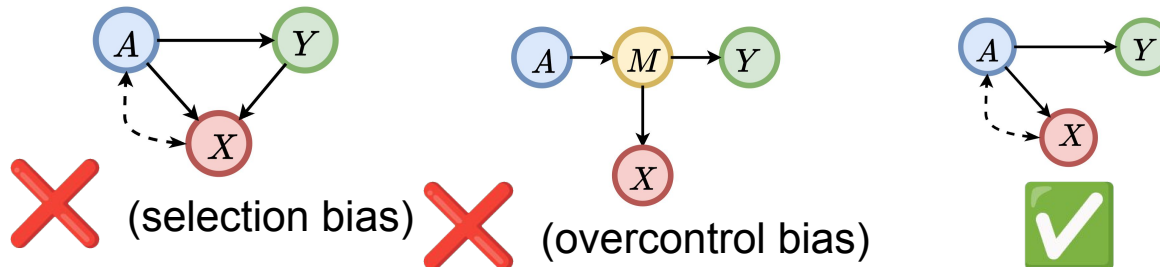
Causal assumptions: Equivalence of the frameworks

- Rule of thumb (revisited): Almost all pre-treatment covariates are fine except for (rarely) variables, that can induce **M-bias**



Choosing covariates (revisited)

- Most of the post-treatment covariate adjustments lead to **re-introducing bias**

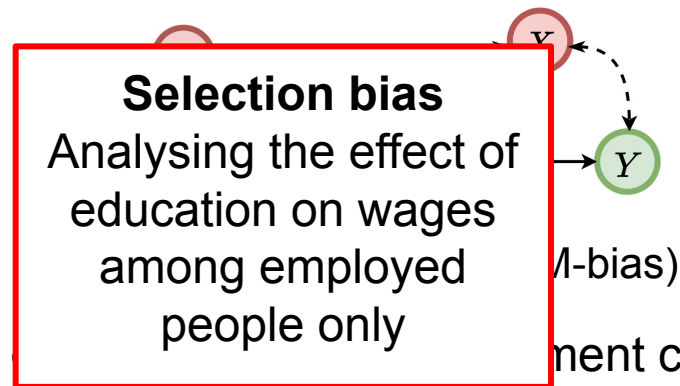


- See ([Cinelli et al. 2022](#)) for details.

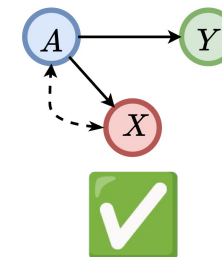
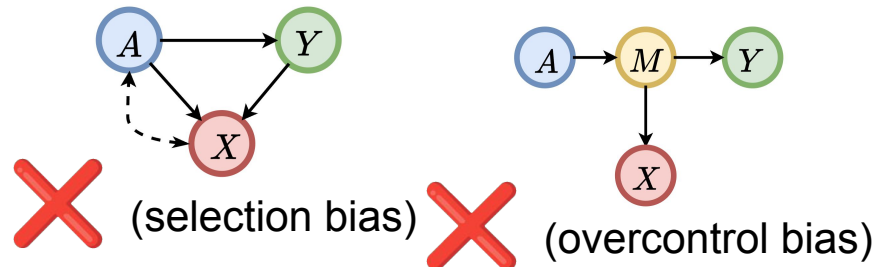
Causal assumptions: Equivalence of the frameworks

- Rule of thumb (revisited): Almost all pre-treatment covariates are fine except for (rarely) variables, that can induce **M-bias**

Choosing
covariates
(revisited)



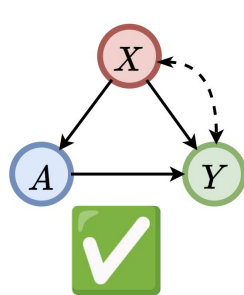
ment covariate adjustments lead to **re-introducing bias**



- See ([Cinelli et al. 2022](#)) for details.

Causal assumptions: Equivalence of the frameworks

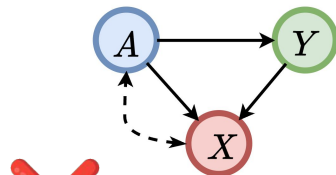
- Rule of thumb (revisited): Almost all pre-treatment covariates are fine except for (rarely) variables, that can induce **M-bias**



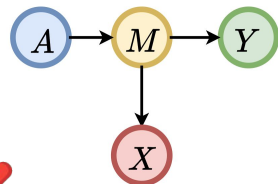
Overcontrol bias
Occupation when estimating the total effect of a university degree on income

- Most of the post-treatment covariates lead to **re-introducing bias**

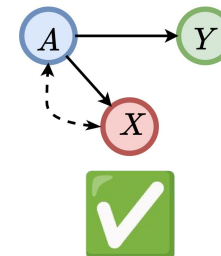
Choosing covariates (revisited)



(selection bias)



(overcontrol bias)



- See ([Cinelli et al. 2022](#)) for details.

Causal assumptions: Beyond unconfoundedness

- Causal diagrams + do-calculus tell us exactly when an effect is identifiable
- When hidden confounding **cannot be ruled out** (= arguably, always), **alternative designs** exist:
 - **Instrumental variables**: Using variables that affect the treatment but not the outcome
 - **Difference-in-differences**: Panel data with parallel trends instead of unconfoundedness
 - **Regression discontinuity**: Based on whether a continuous running variable crosses a specific cutoff threshold
 - **Front-door adjustment**: Effect of smoking on lung cancer through tar deposition in lungs
 - **Sensitivity analysis**: Assuming amount of confounding is bounded, infer the bounds on the causal effects
 - **Partial identification**: Infer the bounds on the causal effects
- ! Still, all the approaches replace one unverifiable assumption with other !
- Causal ML estimators exist for **all of these designs**. A design delivers identification; ML only does the estimation

Causal assumptions: Case study

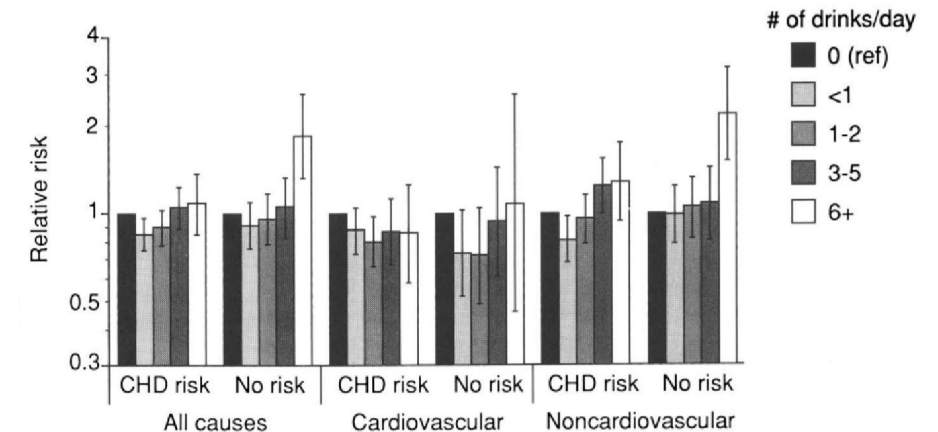
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

Observational data:

- number of alcoholic drinks per day (discrete treatment)
- death after 10 years (outcome)

Aim: APO of alcohol consumption on the risk of death

- **Evidence:** 1-2 drinks/day is the healthiest choice (lowest mortality rate from all the reasons)



Alcohol and Mortality

Arthur L. Klatsky, MD; Mary Anne Armstrong, MA; and Gary D. Friedman, MD

■ **Objective:** To study the relation between alcohol intake and mortality in a large ambulatory population with attention to causes of death and differences related to age, sex, and health status.

Population studies show disparate relations between alcohol intake and various causes of death (1-11). Heavier drinkers have an increased risk for death from sev-

Causal assumptions: Case study

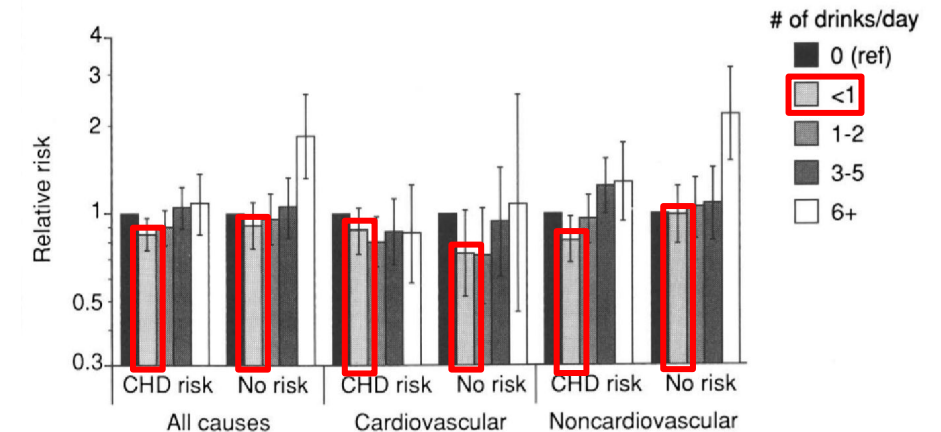
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

Observational data:

- number of alcoholic drinks per day (discrete treatment)
- death after 10 years (outcome)

Aim: APO of alcohol consumption on the risk of death

- **Evidence:** 1-2 drinks/day is the healthiest choice (lowest mortality rate from all the reasons)



Alcohol and Mortality

Arthur L. Klatsky, MD; Mary Anne Armstrong, MA; and Gary D. Friedman, MD

■ **Objective:** To study the relation between alcohol intake and mortality in a large ambulatory population with attention to causes of death and differences related to age, sex, and health status.

Population studies show disparate relations between alcohol intake and various causes of death (1-11). Heavier drinkers have an increased risk for death from sev-

Causal assumptions: Case study

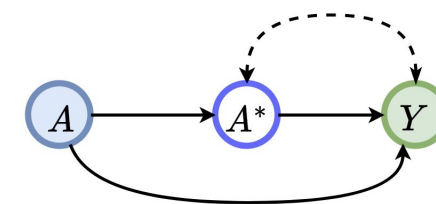
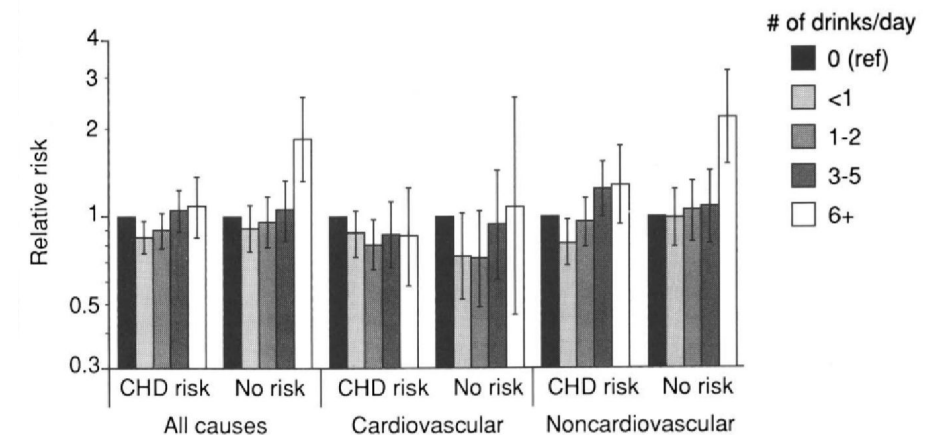
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

Observational data:

- number of alcoholic drinks per day (discrete treatment)
- death after 10 years (outcome)

Aim: APO of alcohol consumption on the risk of death

- **Evidence:** 1-2 drinks/day is the healthiest choice (lowest mortality rate from all the reasons)
- **Problems with the study:**
 - Unobserved confounding (e.g., socio-economic factors, avoidance of alcohol due to health conditions)
 - Violation of consistency (e.g., people were asked whether they drank last month)

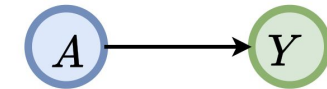


Causal assumptions: Case study

- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

Experimental data?

- number of alcoholic drinks per day (discrete treatment)
- death after 10 years (outcome)



The New York Times

Major Study of Drinking Will Be Shut Down

An investigation at the National Institutes of Health concluded that the \$100 million trial had been tainted by funding appeals to the alcohol industry.

 Share full article    122

Causal assumptions: Case study

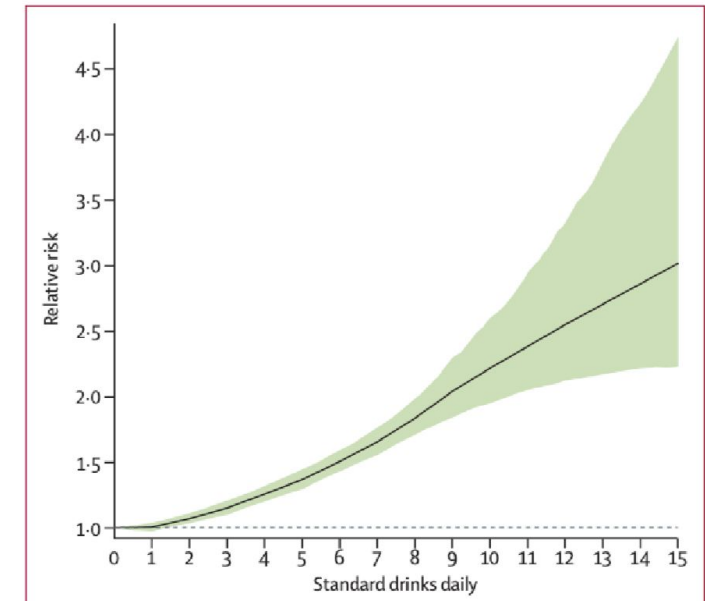
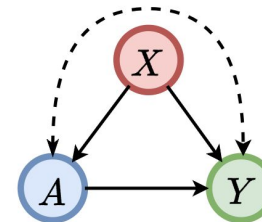
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

Observational data:

- number of alcoholic drinks per day (discrete treatment)
- death after 10 years (outcome)
- socio-economic factors: income, access to healthcare (covariates)

Aim: APO of alcohol consumption on the risk of death

- **Evidence:** no drinks per day are healthy
- **Potential problems with the study:**
 - It is impossible to fully avoid unobserved confounding



Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016

GBD 2016 Alcohol Collaborators*



Causal assumptions: Case study

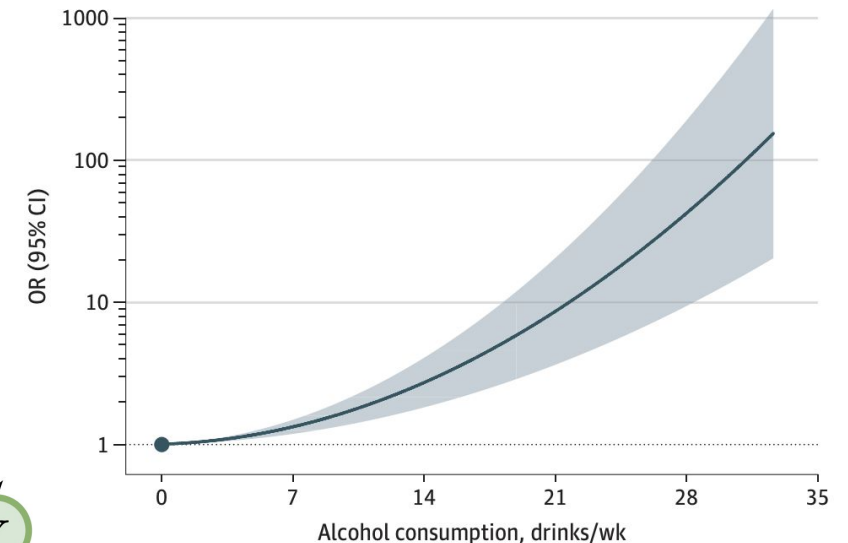
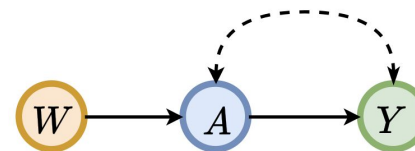
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

Observational data:

- number of alcoholic drinks per day (discrete treatment)
- death after 10 years (outcome)
- genetic information (instrumental variable)

Aim: APO of alcohol consumption on the risk of death

- **Evidence:** no drinks per day are healthy
- **Potential problems with the study:**
 - Strong assumptions about the instrument



Causal assumptions: Case study

- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

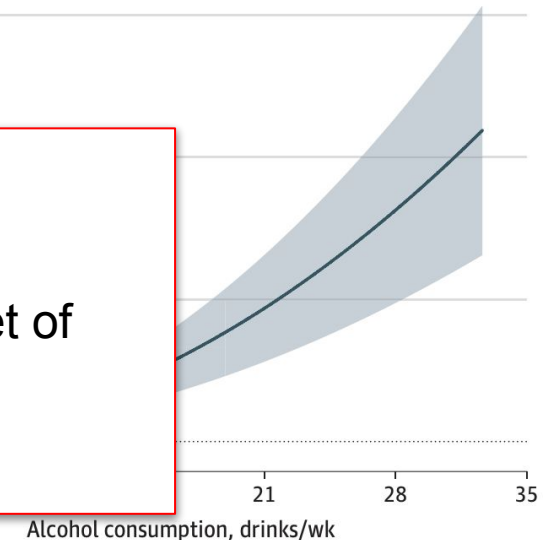
Observational data:

- number of alcoholic drinks per day (discrete treatment)

Aim

Same question, different assumptions, different answers.
Causal ML does not settle this: Once we have committed to a set of assumptions, it only makes estimation flexible.

Strong assumptions about the instrument

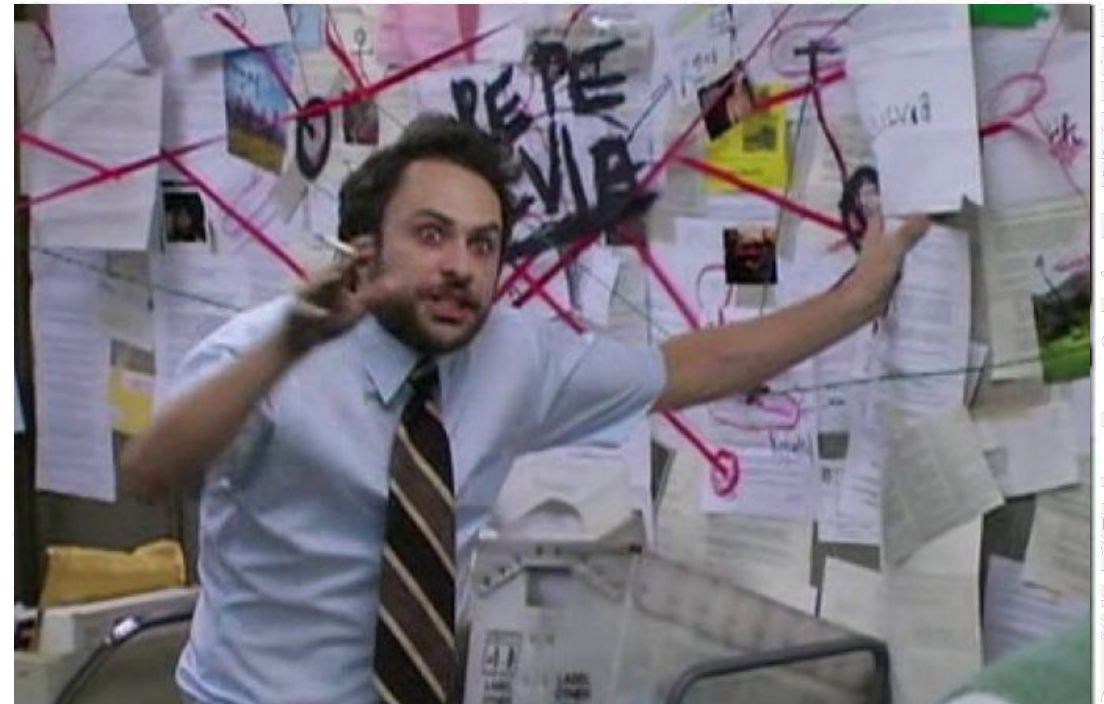


3. ML and estimation

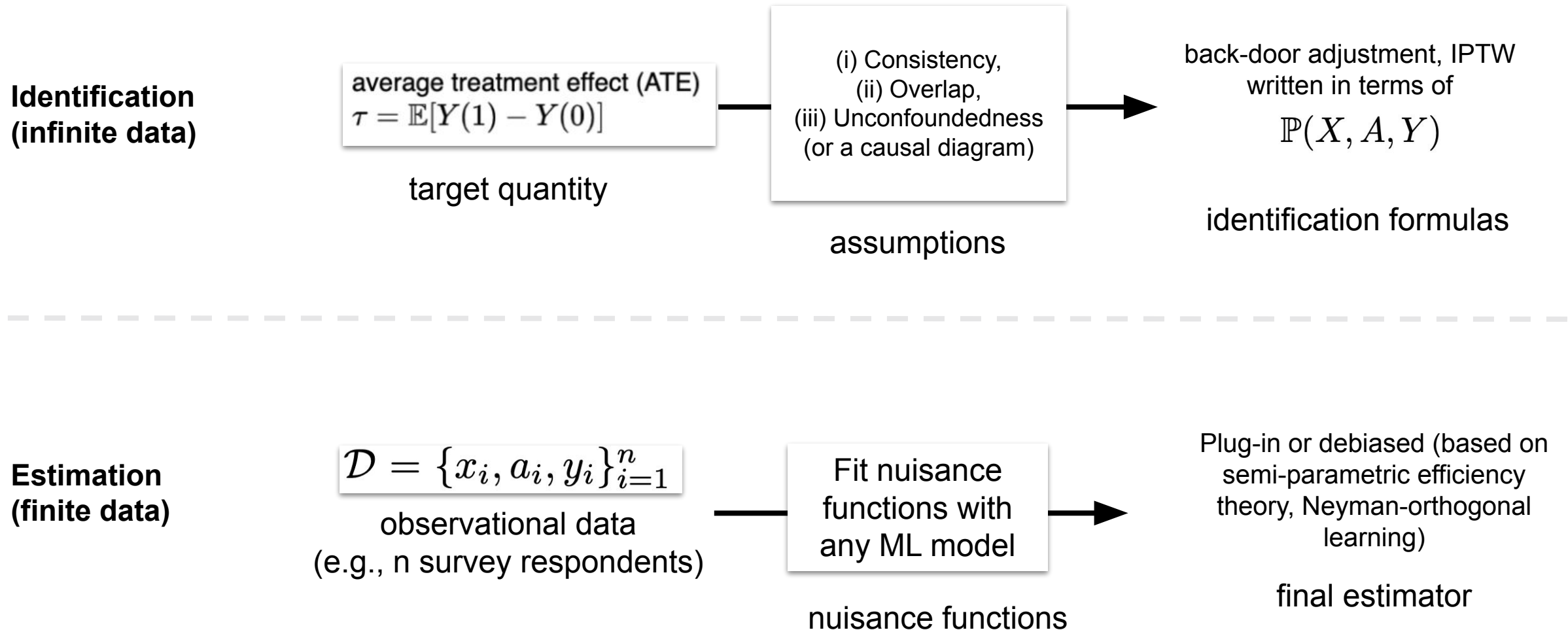
- Identification vs. estimation
- Why not just OLS?
- Big picture of methods
- Plug-in learners and their challenges
- Causal forests
- Neural models
- Two-stage learners and pseudo-outcomes
- Debiased learners: DR- and R-learner
- Model selection and inference in practice
- Cross-fitting

Nobody:

Me explaining all the causal inference methods:



ML and estimation: Identification vs. estimation



ML and estimation: Why not just OLS?

- We might start from a naïve linear model: $Y = \alpha + \tau A + \beta^T X + \varepsilon$
What does the coefficient τ actually estimate?
- **Functional form:** Confounders enter linearly and additively. If, in reality, covariates have non-linear dependencies, the adjustment is incomplete and τ is biased.
- **Hidden weighting:** It can be shown that τ is a weighted average of CATE, with the weights equal in expectation to conditional variance of treatment, $\text{Var}(A | X)$. Hence, it over-represents people whose treatment is most uncertain, not the population
- **Interaction terms:** Interaction terms between X and A are missing
- **High-dimensional covariates:** When X are high-dimensional or unstructured data (e.g., images), linear models fail to represent complex X - Y dependencies

Causal ML keeps the **causal target fixed** (ATE, CATE) and lets the **data** choose the functional form of the nuisance functions

ML and estimation: Big picture

CATE estimation: estimating a function

Meta-learners: use any supervised model of choice for single/two-stage learning

Single-stage learners:

Plug-in learners:

- [S-learner](#) ([S-Net](#), [Causal Forest](#))
- [T-learner](#) ([T-Net](#), [TARNet](#), [DragonNet](#))
- Mixed approaches ([FlexTENet](#))

Debiased learners*:

- [EP-learner](#)

Two-stage learners:

Plug-in learners:

- [IPW-learner](#)
- [RA-learner](#) / [X-learner](#)
- [U-learner](#)

Debiased learners:

- [DR-learner](#)
- [R-learner \(DML\)](#)

ATE estimation: estimating a parameter

Plug-in estimators:

- Plug-in estimator
- RA estimator
- IPTW estimator

Debiased estimators:

- A-IPTW estimator
- TMLE estimator

ML and estimation: Big picture

CATE estimation: estimating a function

Meta-learners: use any supervised model of choice for single/two-stage learning

Single-stage learners:

Plug-in learners:

- [S-learner](#) ([S-Net](#), [Causal Forest](#))
- [T-learner](#) ([T-Net](#), [TARNet](#), [DragonNet](#))
- Mixed approaches ([FlexTENet](#))

Debiased learners*:

- [EP-learner](#)

Two-stage learners:

Plug-in learners:

- [IPW-learner](#)
- [RA-learner](#) / [X-learner](#)
- [U-learner](#)

Debiased learners:

- [DR-learner](#)
- [R-learner \(DML\)](#)

ATE estimation: estimating a parameter

Plug-in estimators:

- Plug-in estimator
- RA estimator
- IPTW estimator

Debiased estimators:

- A-IPTW estimator
- TMLE estimator

Today's focus

ML and estimation: Single-stage plug-in learners

- Given observational data, we just need to estimate **nuisance functions** and
 - plug-in them for CATE
 - take a sample average for ATE

Step 1. Nuisance estimation

$$\hat{\eta} = \left\{ \hat{\mu}_a(x) = \hat{\mathbb{E}}[Y \mid A = a, X = x] \right\}$$

Step 2. Post-processing: Plug-in estimation / sample averaging

CATE	ATE
$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$	$\hat{\tau}_{\text{PI}} = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X^{(i)}) - \hat{\mu}_0(X^{(i)})$

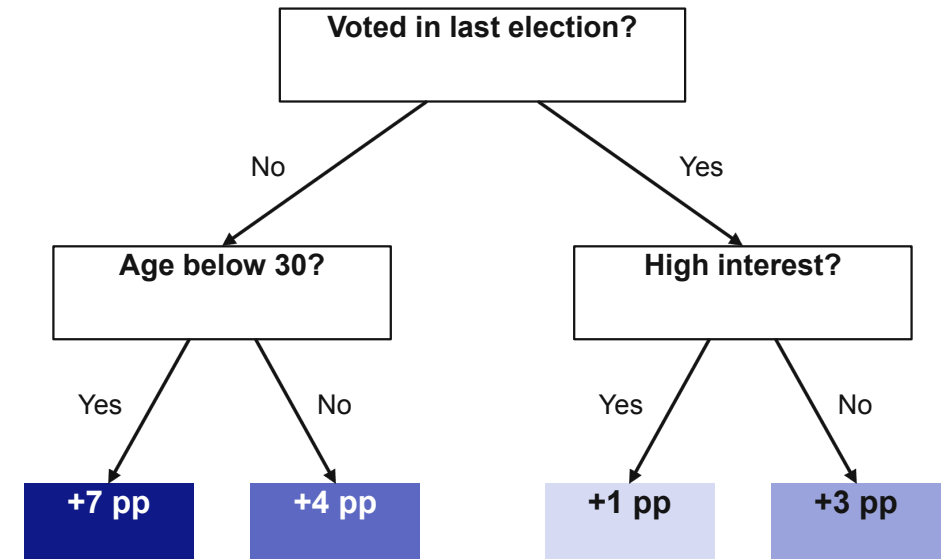
Plug-in
estimators
–
Single-stage
plug-in
learners

- We can learn nuisance functions
 - as a joint “single” model (**S-learner**): e.g., random forest with A as a feature
 - as a “two” separate models (**T-learner**): e.g., two random forests for each A

ML and estimation: Causal forest (specific plug-in learners)

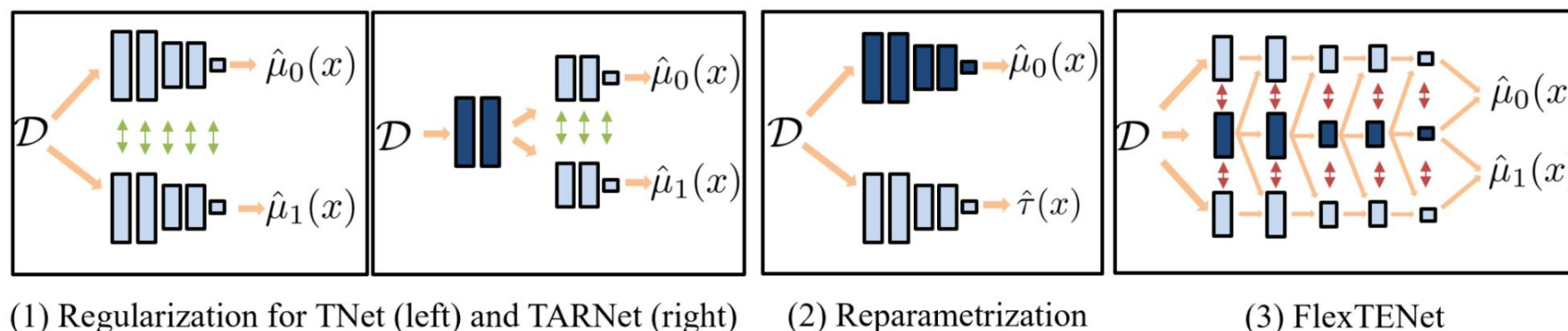
- **Causal tree:** split where the treatment effect differs most between the two groups, not where the outcome differs most. Respondents who share a leaf with person i form a data-driven neighborhood of i
- Each leaf can be seen as an estimator of
$$\hat{\eta} = \left\{ \hat{\mu}_a(x) = \hat{\mathbb{E}}[Y \mid A = a, X = x] \right\}$$
- **Data splitting (honesty):** one half of the sample chooses the splits, the other half estimates the effects in the leaves
- **Causal forest:** grow many trees on random subsamples

Causal forest

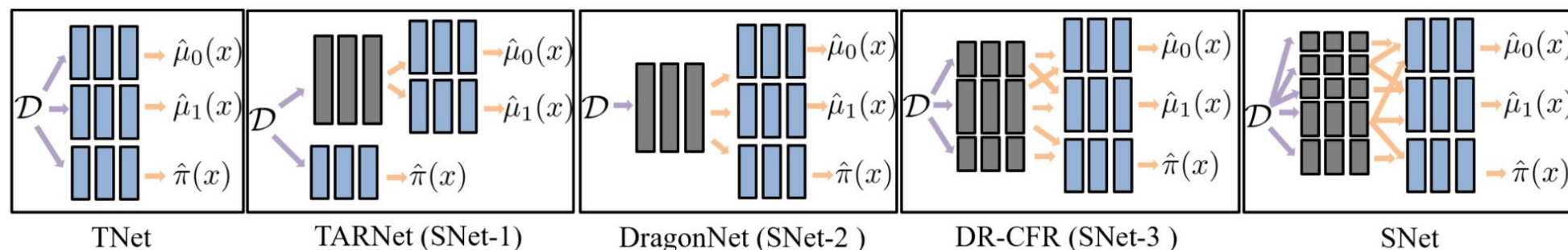


ML and estimation: Neural models (specific plug-in learners)

Sharing
representations
for $\hat{\mu}_a(x)$



Sharing
representations
for all the
nuisance
functions



ML and estimation: Issues of plug-in estimation

Problem solved? **NO!**

Issues of
plug-in
estimators /
plug-in
learners

1. What about the sub-group treatment effects?
2. How to regularize $\hat{\tau}(x)$?
3. $\hat{\mu}_a(x)$ can only be well estimated for some parts of the population, e.g., only in treated group. How to address the covariate shift?
4. What is an optimal meta-learner?
5. Can we do model selection / hyperparameter tuning?

ML and estimation: 1. “What about the sub-group treatment effects?”

- ATE = Sub-group treatment effect with $V = \emptyset$
- What if we want to learn arbitrary $V \subseteq \mathcal{X}$?
- In traditional ML, we would simply do a regression with less features (= minimize MSE):
 - Subgroup ATE: $\mathcal{L}(\hat{\tau}) = \mathbb{E}((Y[1] - Y[0] - \hat{\tau}(V))^2$
- But, we face the fundamental problem of causal inference

**Sub-group
treatment
effects**

ML and estimation: 1. “What about the sub-group treatment effects?”

- ATE = Sub-group treatment effect with $V = \emptyset$
- What if we want to learn arbitrary $V \subseteq X$?
- In traditional ML, we would simply do a regression with less features (= minimize MSE):
 - Subgroup ATE: $\mathcal{L}(\hat{\tau}) = \mathbb{E}((Y[1] - Y[0] - \hat{\tau}(V))^2$ **never observed**
- But, we face the fundamental problem of causal inference

**Sub-group
treatment
effects**

ML and estimation: 1. “What about the sub-group treatment effects?”

Sub-group treatment effects

- ATE = Sub-group treatment effect with $V = \emptyset$
- What if we want to learn arbitrary $V \subseteq X$?
- In traditional ML, we would simply do a regression with less features (= minimize MSE):
 - Subgroup ATE: $\mathcal{L}(\hat{\tau}) = \mathbb{E}((Y[1] - Y[0] - \hat{\tau}(V))^2$
- But, we face the fundamental problem of causal inference
- **Idea:** machine learning with the nuisance functions
 - Subgroup ATE: $\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E}(\tau(X) - \hat{\tau}(V))^2$
(imputing full CATE to learn the subgroup ATE)

ML and estimation: Big picture

CATE estimation: estimating a function

Meta-learners: use any supervised model of choice for single/two-stage learning

Single-stage learners:

Plug-in learners:

- [S-learner](#) ([S-Net](#), [Causal Forest](#))
- [T-learner](#) ([T-Net](#), [TARNet](#), [DragonNet](#))
- Mixed approaches ([FlexTENet](#))

Debiased learners*:

- [EP-learner](#)

Two-stage learners:

Plug-in learners:

- [IPW-learner](#)
- [RA-learner](#) / [X-learner](#)
- [U-learner](#)

Debiased learners:

- [DR-learner](#)
- [R-learner \(DML\)](#)

ATE estimation: estimating a parameter

Plug-in estimators:

- Plug-in estimator
- RA estimator
- IPTW estimator

Debiased estimators:

- A-IPTW estimator
- TMLE estimator

ML and estimation: 1. Two-stage plug-in learners

CATE	ATE
$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$	$\hat{\tau}_{\text{PI}} = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X^{(i)}) - \hat{\mu}_0(X^{(i)})$

Sub-group treatment effects

- ATE = Sub-group treatment effect with $V = \emptyset$ ($V \subseteq X$)
Sample averaging = Regression with intercept only
- **Core idea:** create **pseudo-outcomes** $\tilde{Y}_{\hat{\eta}}$, whose property is $\mathbb{E}(\tilde{Y}_{\hat{\eta}} | V = v) = \tau(v)$

$$\tilde{Y}_{\text{RA}, \hat{\eta}} = A(Y - \hat{\mu}_0(X)) + (1 - A)(\hat{\mu}_1(X) - Y) \quad (\text{Observed outcome minus imputed counterfactual})$$

$$\tilde{Y}_{\text{IPTW}, \hat{\eta}} = \left(\frac{A}{\hat{\pi}_1(X)} - \frac{1-A}{\hat{\pi}_0(X)} \right) Y \quad (\text{Reweighted outcome, as in survey weighting})$$
- We regress them on V with e.g. L2 loss: $\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$

ML and estimation: 1. Two-stage plug-in learners

- Two-step learners, based on pseudo-adjust are, **IPW-learner**, **RA-learner** / **X-learner**

Step 1. Nuisance estimation

$$\hat{\eta} = \{ \hat{\mu}_a(x) = \hat{\mathbb{E}}[Y \mid A = a, X = x]; \hat{\pi}_a(x) = \hat{\mathbb{P}}[A = a \mid X = x] \}$$

Step 2. Post-processing: Regression on pseudo-outcomes

CATE

$$\tilde{Y}_{\text{RA}, \hat{\eta}} = A(Y - \hat{\mu}_0(X)) + (1 - A)(\hat{\mu}_1(X) - Y)$$

$$\tilde{Y}_{\text{IPTW}, \hat{\eta}} = \left(\frac{A}{\hat{\pi}_1(X)} - \frac{1-A}{\hat{\pi}_0(X)} \right) Y$$

$$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$$

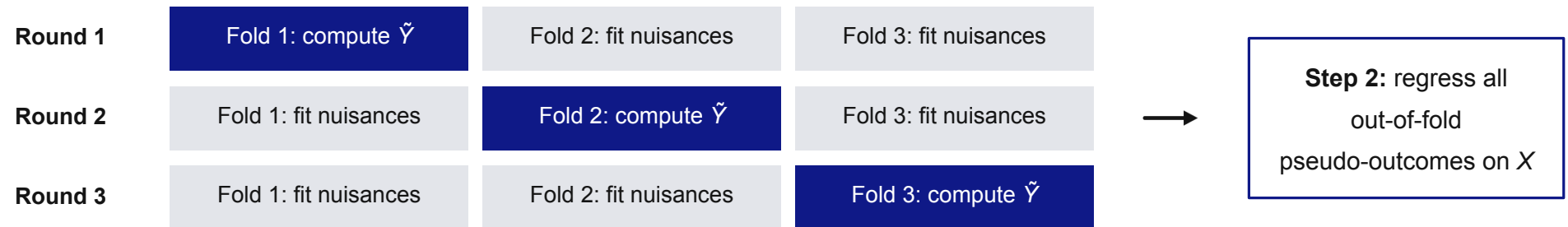
Two-stage
plug-in
learners

- Flexible models in step 1 require sample splitting (cross-fitting)

ML and estimation: Cross-fitting

- Flexible ML models overfit. If pseudo-outcomes are computed on the same data used to fit the nuisance functions, **the overfitting leaks into the effect estimate**

Cross-fitting

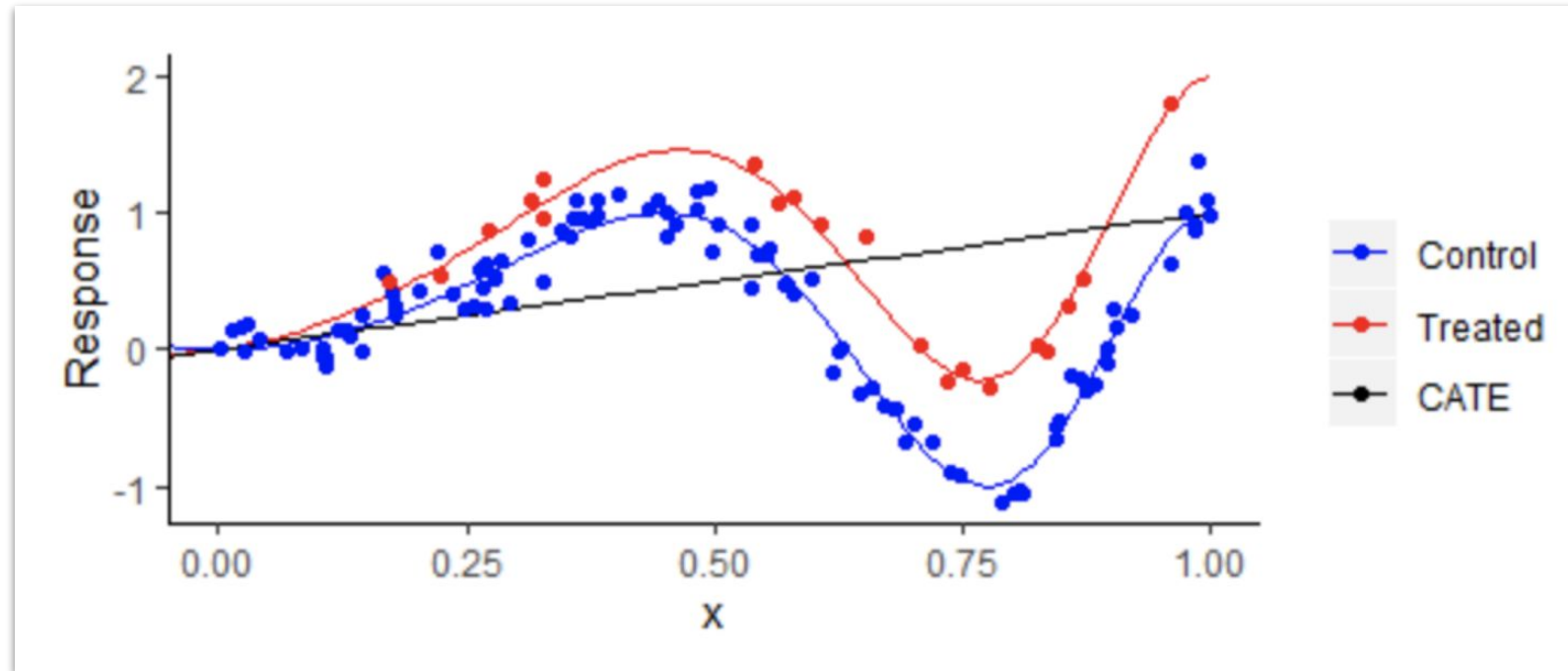


- Every respondent gets a pseudo-outcome from models that **never saw** that respondent
- Built into software packages; causal forests use the related idea of honesty

ML and estimation: 2. “How to regularize $\hat{\tau}(x)$?”

- What if the true CATE is much simpler than each potential outcome? (turnout depends on age in complex ways; contact adds a roughly constant boost):
 - **T-learner fails:** the difference of two separately regularized fits is noisy and too wiggly
 - **S-learner fails:** may ignore the treatment among other inputs if regularized too much

How to
regularize a
target model?



Curth, Alicia, and Mihaela Van der Schaar. "On inductive biases for heterogeneous treatment effect estimation." Advances in Neural Information Processing Systems 34 (2021): 15883-15894.

ML and estimation: 2. Two-stage plug-in learners

- With two-step learners, regularization of the target model becomes **straightforward**:

$$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2 + \Lambda(\hat{\tau})$$

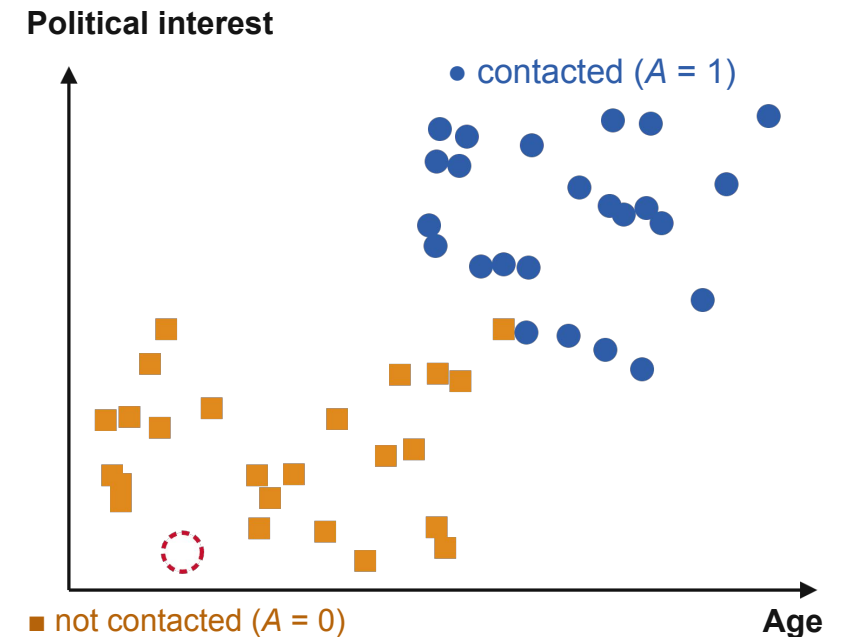
How to
regularize a
target model?

- Can be done **separately** from the nuisance functions
- Regularization acts as a “**dial**” for heterogeneity:
 - small values of regularization hyperparameter → full CATE
 - medium values of regularization hyperparameter → effects driven by key covariates → “sub-group” ATE
 - large values of regularization hyperparameter → regression on the intercept only → ATE

ML and estimation: 3. “How to address the covariate shift?”

Covariate shift problem

- Campaigns contact older, politically interested citizens: $\mu_1(x)$ is learned almost only from them
- For young, disengaged citizens, $\mu_1(x)$ must be **extrapolated**: CATE estimates are least reliable exactly where mobilization might matter most
- The problem becomes worse in the regions where the propensity score is close to 0 or 1. Thus, **low overlap** → covariate shift
- Plug-in learners **do not** use all the information in the data (e.g., the propensity score) to correct for this → inefficient data usage

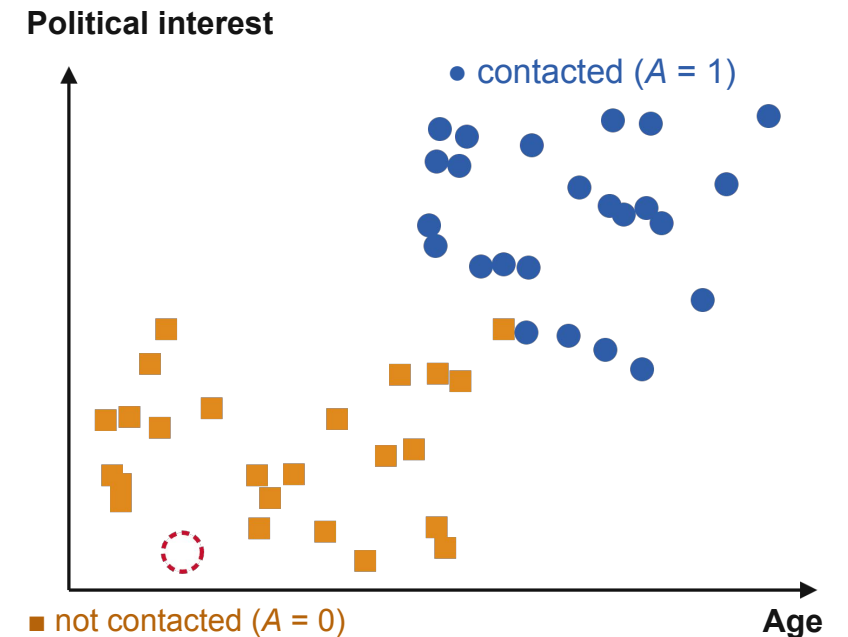


ML and estimation: 3. “How to address the covariate shift?”

Covariate shift problem

- Campaigns contact older, politically interested citizens: $\mu_1(x)$ is learned almost only from them
- For young, disengaged citizens, $\mu_1(x)$ must be **extrapolated**: CATE estimates are least reliable exactly where mobilization might matter most
- The problem becomes worse in the regions where the propensity score is close to 0 or 1. Thus, **low overlap** → covariate shift
- Plug-in learners **do not** use all the information in the data (e.g., the propensity score) to correct for this → inefficient data usage

Potential estimation bias



Inefficient estimation

ML and estimation: Big picture

CATE estimation: estimating a function

Meta-learners: use any supervised model of choice for single/two-stage learning

Single-stage learners:

Plug-in learners:

- [S-learner](#) ([S-Net](#), [Causal Forest](#))
- [T-learner](#) ([T-Net](#), [TARNet](#), [DragonNet](#))
- Mixed approaches ([FlexTENet](#))

Debiased learners*:

- [EP-learner](#)

Two-stage learners:

Plug-in learners:

- [IPW-learner](#)
- [RA-learner](#) / [X-learner](#)
- [U-learner](#)

Debiased learners:

- [DR-learner](#)
- [R-learner](#) ([DML](#))

ATE estimation: estimating a parameter

Plug-in estimators:

- Plug-in estimator
- RA estimator
- IPTW estimator

Debiased estimators:

- A-IPTW estimator
- TMLE estimator

ML and estimation: 3. Debiased learners

- Solution: two-step learner that uses both the propensity score and the outcome model:
doubly-robust learner (DR-learner)

Step 1. Nuisance estimation

$$\hat{\eta} = \{ \hat{\mu}_a(x) = \hat{\mathbb{E}}[Y \mid A = a, X = x]; \hat{\pi}_a(x) = \hat{\mathbb{P}}[A = a \mid X = x] \}$$

Step 2. Post-processing: Regression on pseudo-outcomes

DR-learner

CATE

$$\tilde{Y}_{\text{DR}, \hat{\eta}} = \left(\frac{A - \hat{\pi}_1(X)}{\hat{\pi}_1(X) \hat{\pi}_0(X)} \right) (Y - \hat{\mu}_A(X)) + \hat{\mu}_1(X) - \hat{\mu}_0(X)$$

$$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$$

- Flexible models in step 1 require sample splitting (cross-fitting)

ML and estimation: 3. Debiased learners

- Solution: two-step learner that uses both the propensity score and the outcome model:
doubly-robust learner (DR-learner)

Step 1. Nuisance estimation

$$\hat{\eta} = \{ \hat{\mu}_a(x) = \hat{\mathbb{E}}[Y \mid A = a, X = x]; \hat{\pi}_a(x) = \hat{\mathbb{P}}[A = a \mid X = x] \}$$

Step 2. Post-processing: Regression on pseudo-outcomes

DR-learner

CATE	
$\tilde{Y}_{\text{DR}, \hat{\eta}} = \left(\frac{A - \hat{\pi}_1(X)}{\hat{\pi}_1(X) \hat{\pi}_0(X)} \right) (Y - \hat{\mu}_A(X)) + \hat{\mu}_1(X) - \hat{\mu}_0(X)$	IPTW correction term Regression adjustment
$\mathcal{L}(\hat{\tau}, \hat{\eta})$	

- Flexible models in step 1 require sample splitting (cross-fitting)

ML and estimation: 3. Debiased learners

Benefits:

- Averaging DR pseudo-outcomes yields **augmented ITPW (A-IPTW)** estimator of **ATE**: the basis of semi-parametric inference with valid confidence intervals:

$$\hat{\tau}_{\text{A-IPTW}} = \frac{1}{n} \sum_{i=1}^n \left(\frac{A^{(i)} - \hat{\pi}_1(X^{(i)})}{\hat{\pi}_1(X^{(i)})\hat{\pi}_0(X^{(i)})} \right) \left(Y^{(i)} - \hat{\mu}_{A^{(i)}}(X^{(i)}) \right) + \hat{\mu}_1(X^{(i)}) - \hat{\mu}_0(X^{(i)})$$

Properties of DR-learner

- **Neyman-orthogonality (= first-order debiasing)**: errors in the nuisance functions have zero first-order influence on the target model (e.g., if the error of the outcome model is 0.1, then it contributes to the target model only in the higher order, $0.1^2 = 0.01$)
- **Double robustness**: Consistent if **either** the outcome models **or** the propensity score is estimated well

Weaknesses:

- Dividing by propensity score **explodes the estimation variance** when overlap is low ($\pi(x)$ close to 0 or 1) → trimming / clipping helps

ML and estimation: 3. Debiased learners

- Alternative Neyman-orthogonal learner w/o inverse propensities → **residualized learner (R-learner)** (uses a special Robinson's decomposition):

$$Y - \mu(X) = (A - \pi_1(X))\tau(X) + \varepsilon(A)$$

$$\text{where } \varepsilon(a) = Y(a) - (\mu_0(X) + a\tau(X)), \quad \mathbb{E}(\varepsilon(A) \mid A = a, X = x) = 0, \quad \mu(X) = \mathbb{E}(Y \mid X = x)$$

Step 1. Nuisance estimation

$$\hat{\eta} = \{ \hat{\mu}(x) = \hat{\mathbb{E}}[Y \mid X = x]; \hat{\pi}_a(x) = \hat{\mathbb{P}}[A = a \mid X = x] \}$$

Step 2. Post-processing: Minimization of the custom loss

CATE

$$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E} \left((Y - \hat{\mu}(X)) - (A - \hat{\pi}_1(X))\hat{\tau}(V) \right)^2$$

- Flexible models in step 1 require sample splitting (cross-fitting)

R-learner

ML and estimation: 3. Debiased learners

Benefits:

- Fitting R-learner is equivalent to learning CATE in the **overlapping regions** of the covariate space (fitting intercept only yields **overlap-weighted ATE**):

$$\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E} \left(\pi_1(X) \pi_0(X) \left(\tau(X) - \hat{\tau}(V) \right)^2 \right)$$

Properties of R-learner

- **Neyman-orthogonality (= first-order debiasing)**: errors in the nuisance functions have zero first-order influence on the target model (e.g., if the error of the outcome model is 0.1, then it contributes to the target model only in the higher order, $0.1^2 = 0.01$)
- **Stability with low-overlap**: No division by the propensity score

Weaknesses:

- **No double robustness**: Propensity score needs to be estimated well for consistency
- **Implicit weighting**: Similarly to OLS, performs implicit weighting with overlap weights $\text{Var}(A|X) = \pi_1(X) \pi_0(X)$. Depending on the setting, that is not always what we want

ML and estimation: 3. Debiased learners

Benefits:

- Fitting R-learner is equivalent to learning CATE in the **overlapping regions** of the covariate space (fitting intercept only yields **overlap-weighted ATE**):

$$\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E} \left(\pi_1(X) \pi_0(X) \left(\tau(X) - \hat{\tau}(V) \right)^2 \right)$$

Overlap weights

- Neyman-orthogonality (– first-order debiasing)**: errors in the nuisance functions have zero first-order influence on the target model (e.g., if the error of the outcome model is 0.1, then it contributes to the target model only in the higher order, $0.1^2 = 0.01$)
- Stability with low-overlap**: No division by the propensity score

Weaknesses:

- No double robustness**: Propensity score needs to be estimated well for consistency
- Implicit weighting**: Similarly to OLS, performs implicit weighting with overlap weights $\text{Var}(A|X) = \pi_1(X) \pi_0(X)$. Depending on the setting, that is not always what we want

Properties of
R-learner

ML and estimation: 4. “What is an optimal meta-learner?”

- What is better, RA-, IPTW-, DR-, or R-learner?
- Ideal estimator / learner should:
 - omit the error from estimated nuisance function as much as possible
 - have the lowest asymptotic variance (= statistical efficiency)

**Optimality of
the estimator /
learner**

ML and estimation: 4. “What is an optimal meta-learner?”

- What is better, RA-, IPTW-, DR-, or R-learner?
- Ideal estimator / learner should:
 - omit the error from estimated nuisance function as much as possible
 - have the lowest asymptotic variance (= statistical efficiency)

**Optimality of
the estimator /
learner**

26 Jan 2023

Semiparametric Doubly Robust Targeted Double Machine Learning: A Review ^{*†}

Edward H. Kennedy
Department of Statistics & Data Science
Carnegie Mellon University

edward@stat.cmu.edu

Contents

1	Introduction	2
1.1	Notation	2

ML and estimation: 4. Debiased learners

- What is better, RA-, IPTW-, DR-, or R-learner?
- Ideal estimator / learner should:
 - omit the error from estimated nuisance function as much as possible
 - have the lowest asymptotic variance (= statistical efficiency)
- Debiased learners / estimators (DR-/R-/A-IPTW) were shown to be **asymptotically optimal**:
 - They remove first-order errors of the nuisance functions → Higher-order debiasing drastically increases variance, yet only marginally reduces bias
 - **Local optimality**: no other semi-parametric model has lower asymptotic variance (A-IPTW attains the semi-parametric efficiency bound)
 - **Global optimality**: given the nuisance functions have certain error, A-IPTW is the best possible way to combines the data and the nuisance functions (it archives best mini-max MSE risk)

Optimality of
the estimator /
learner

ML and estimation: 4. “Can we do model selection / hyperparameter tuning?”

- **Best asymptotically does not mean best in finite-sample!** → “no free lunch theorem” but for meta-learners

**Best approach
in low-sample
regime**

ML and estimation: 4. “Can we do model selection / hyperparameter tuning?”

Best approach in low-sample regime

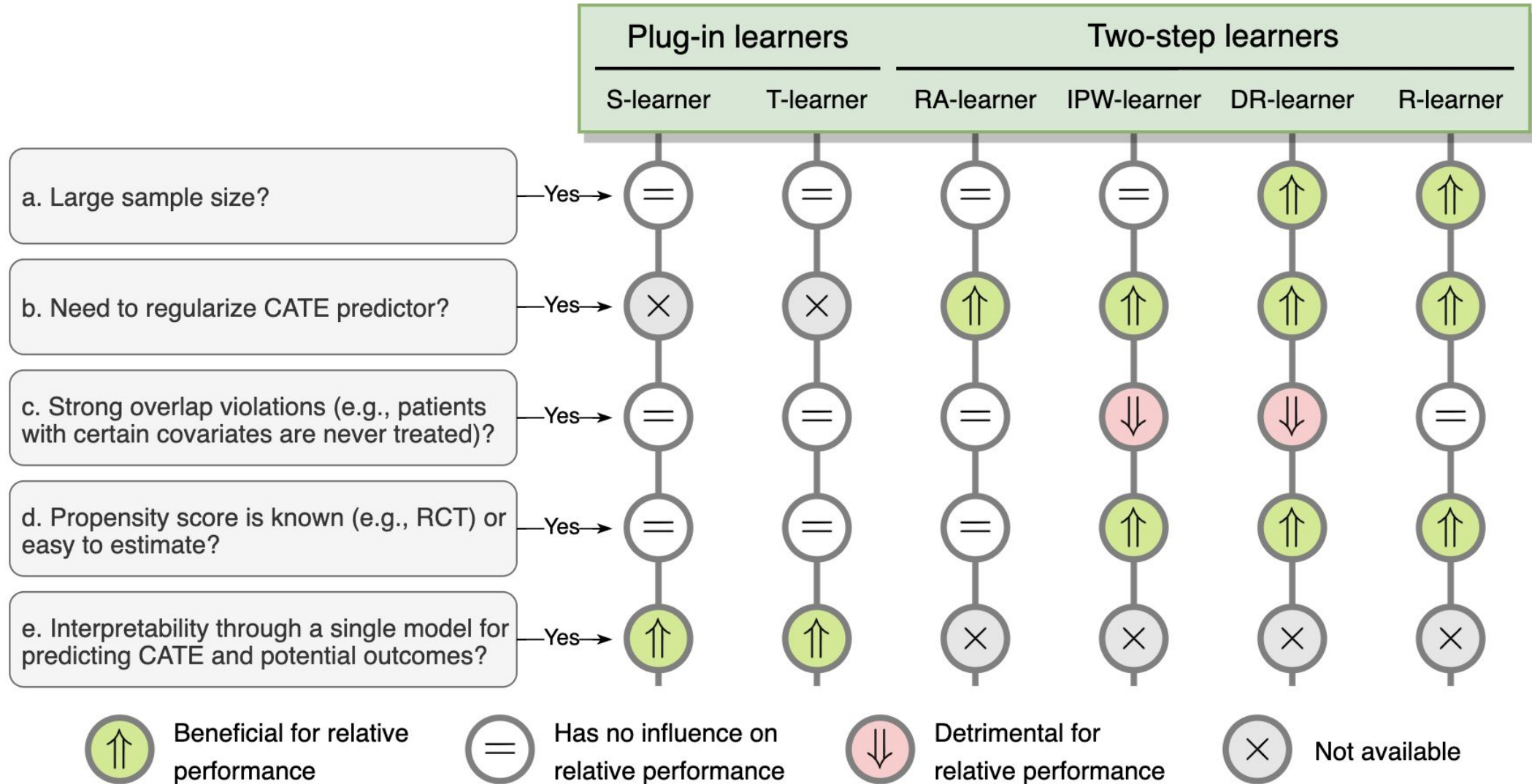
- **Best asymptotically does not mean best in finite-sample!** → “no free lunch theorem” but for meta-learners
- In causal ML, we also **do not** have **nuisance-free** model selection criteria for target models (fundamental problem of causal inference → no ground-truth):
 - evaluating model selection criteria is similar to estimating ATE itself
 - we always need estimated nuisance functions for model selection / hyperparameter tuning → **nuisance errors are irreducible**
 - **congeniality bias**: i.e. DR-loss selection criterion would prefer DR-learners
 - only heuristics are available (Rothenhäusler 2020, Curth & van der Schaar, 2023)
- Ultimate selection model selection criterion: IPTW-loss from the randomized data experiment (i.e., when propensities are known) → **still can be very noisy**

ML and estimation: 4. “Can we do model selection / hyperparameter tuning?”

Practical advice

- Tune nuisance models (here ground-truth is available), calibrate the propensity score
- Compare several debiased learners and target models (**including simple parametric ones**)
- Report how sensitive your conclusions are to these choices. Individual CATE estimates are noisy → Report **aggregated, testable** summaries instead:
 - **Is there heterogeneity at all?** Calibration / best linear predictor test
 - **CATE vs. ATE.** Sort people into quantile groups of predicted CATE; estimate the ATE within each group with debiased estimators
 - **What are those groups?** Compare covariate means across these groups (age, education, past turnout, ...)
 - **Policy values** (value of targeting). Does prioritizing by predicted CATE beat random allocation / allocation wrt. ATE?

ML and estimation: Meta-learners summary

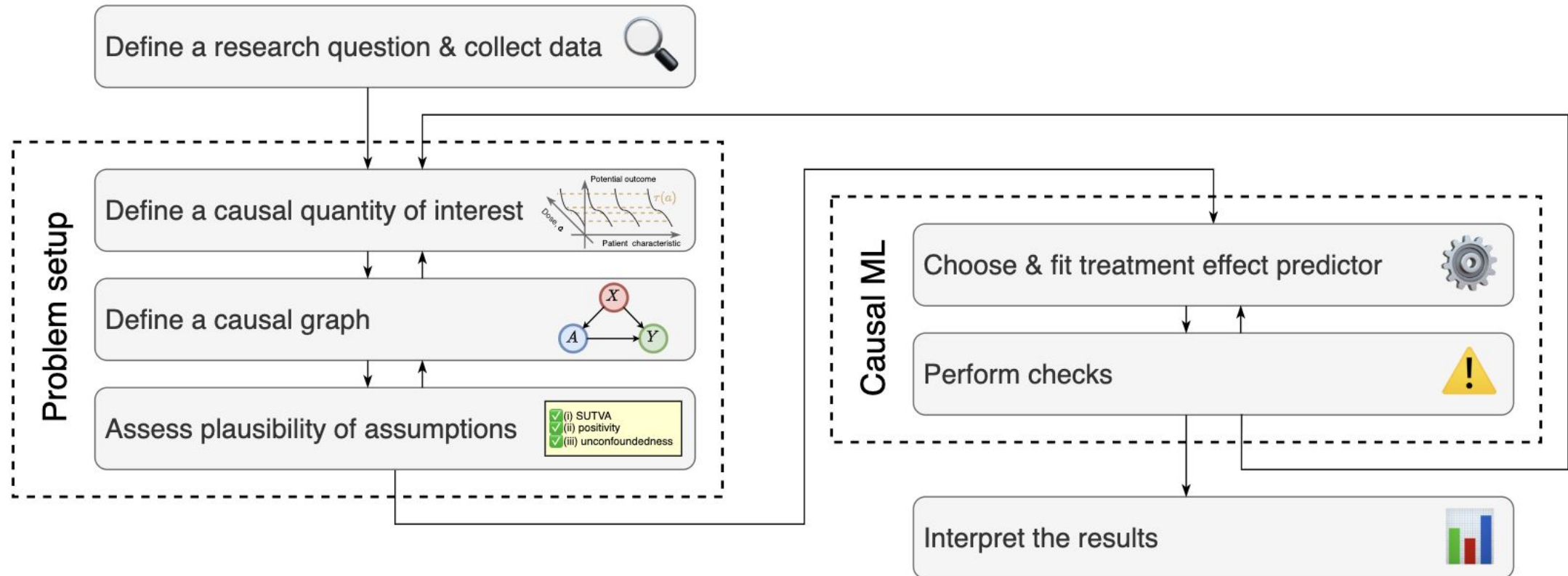


4. Causal ML in practice

- Workflow
- Software
- Checklist



Causal ML in practice: Workflow



- **Problem setup:** Estimand (e.g., CATE of campaign contact on turnout), covariates, overlap check
- **Causal ML:** Debiased learner with cross-fitting; calibration test and sensitivity analysis
- **Interpretation:** Group effects, ATE vs. CATE, policy values; state the assumptions explicitly

Causal ML in practice: Software

Package	Language	What it does
grf	R	Causal forests and instrumental forests: CATE with confidence intervals, calibration test, best linear projection
DoubleML	R, Python	Double/debiased ML for average effects in partially linear, interactive and IV models, with cross-fitting
ddml	Stata	Double/debiased ML with stacking of learners
EconML	Python	DR-, R- and X-learners, causal forests, policy learning
CausalML	Python	Meta-learners and uplift modelling
sensemakr	R, Stata, Python	Sensitivity analysis for omitted variable bias (Cinelli & Hazlett)

Causal ML in practice: Checklist

Before estimation:

1. Define the estimand and the target population: ATE or CATE, for whom?
2. Draw the causal diagram; use pre-treatment covariates only; no mediators or colliders
3. Check overlap: plot propensity scores; trim or redefine the population if needed
4. Pre-register heterogeneity hypotheses where possible

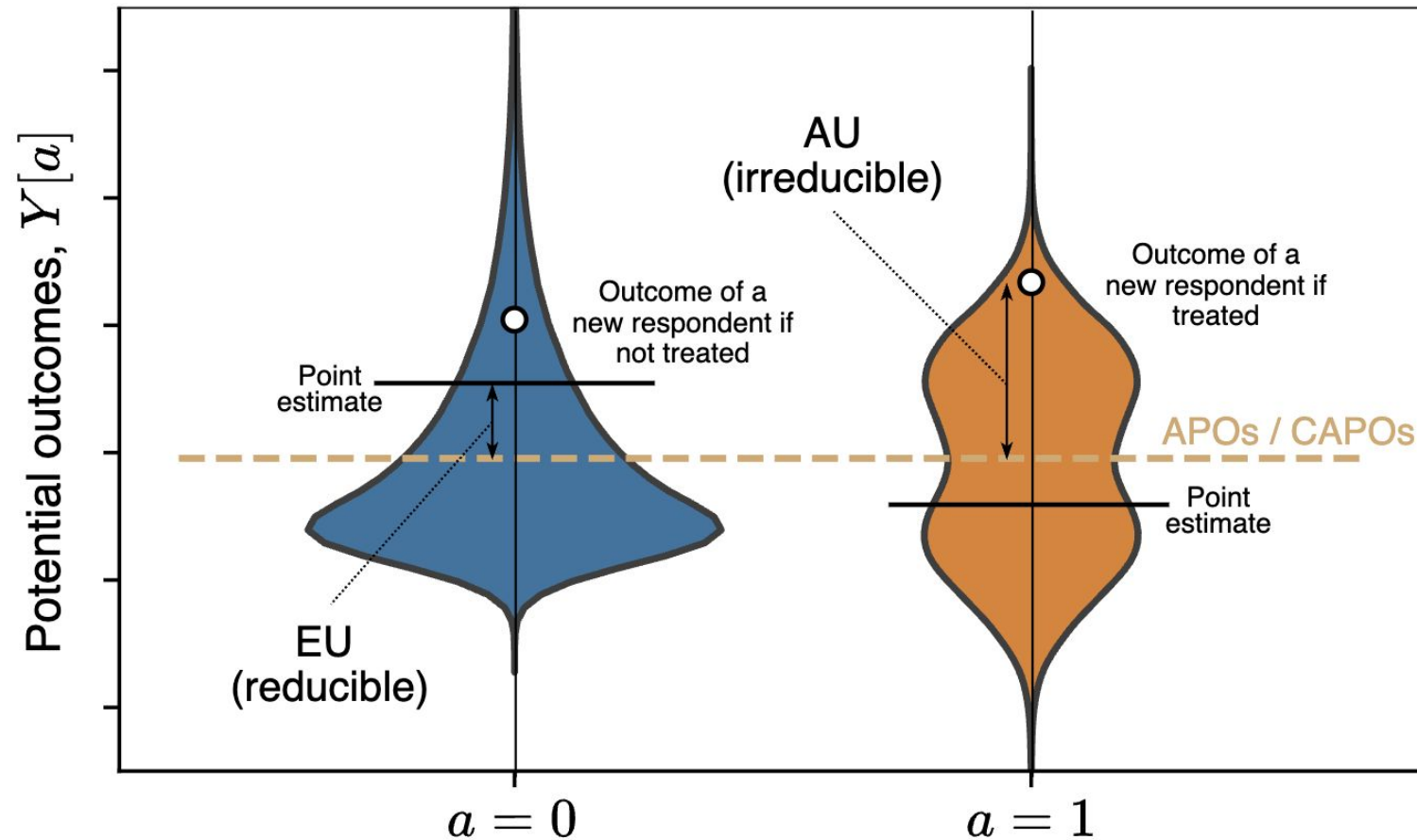
Estimation and reporting:

5. Tune the nuisance models & calibrate the propensity score
6. Use debiased learners with cross-fitting (DR-, R-learner, orthogonal causal forest)
7. Compare several target models, including linear ones
8. Test for heterogeneity; report group effects
9. Run a sensitivity analysis for unobserved confounding

5. Extensions and open challenges

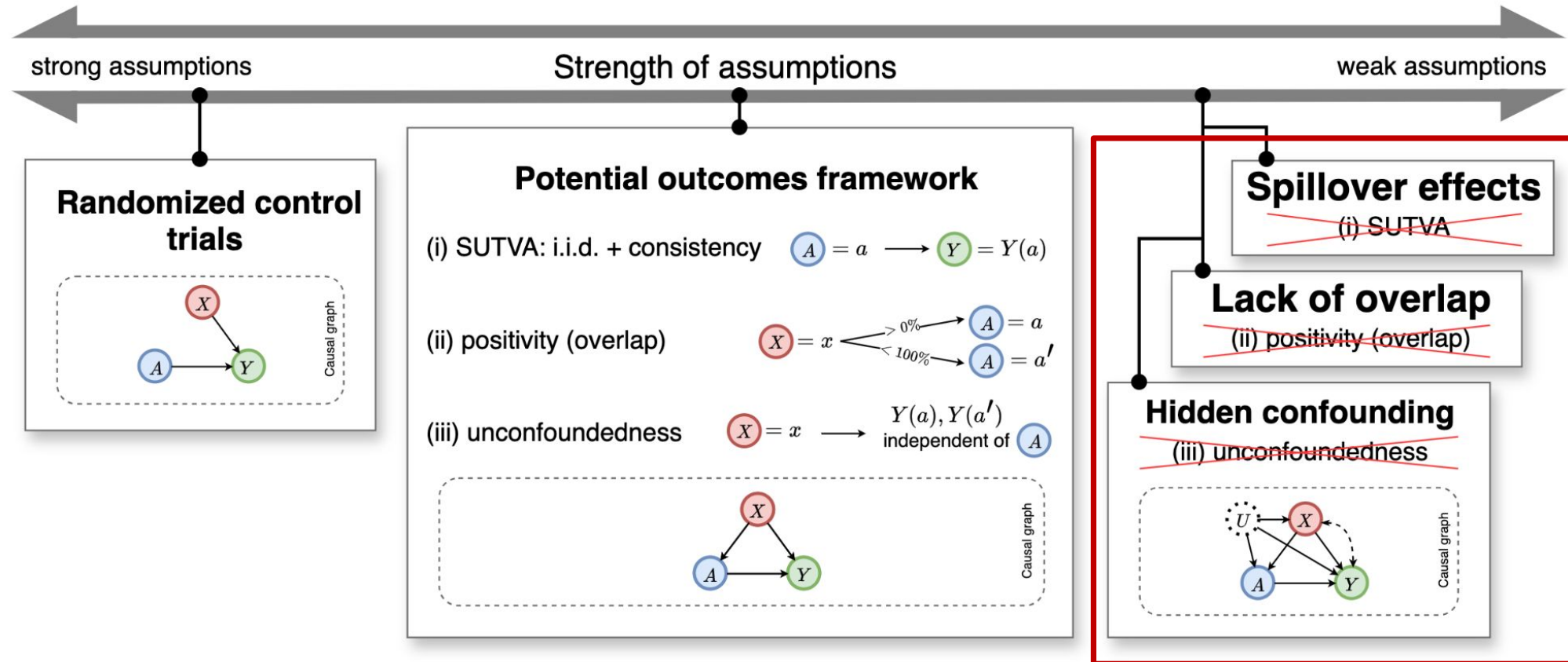


Extensions: Uncertainty quantification



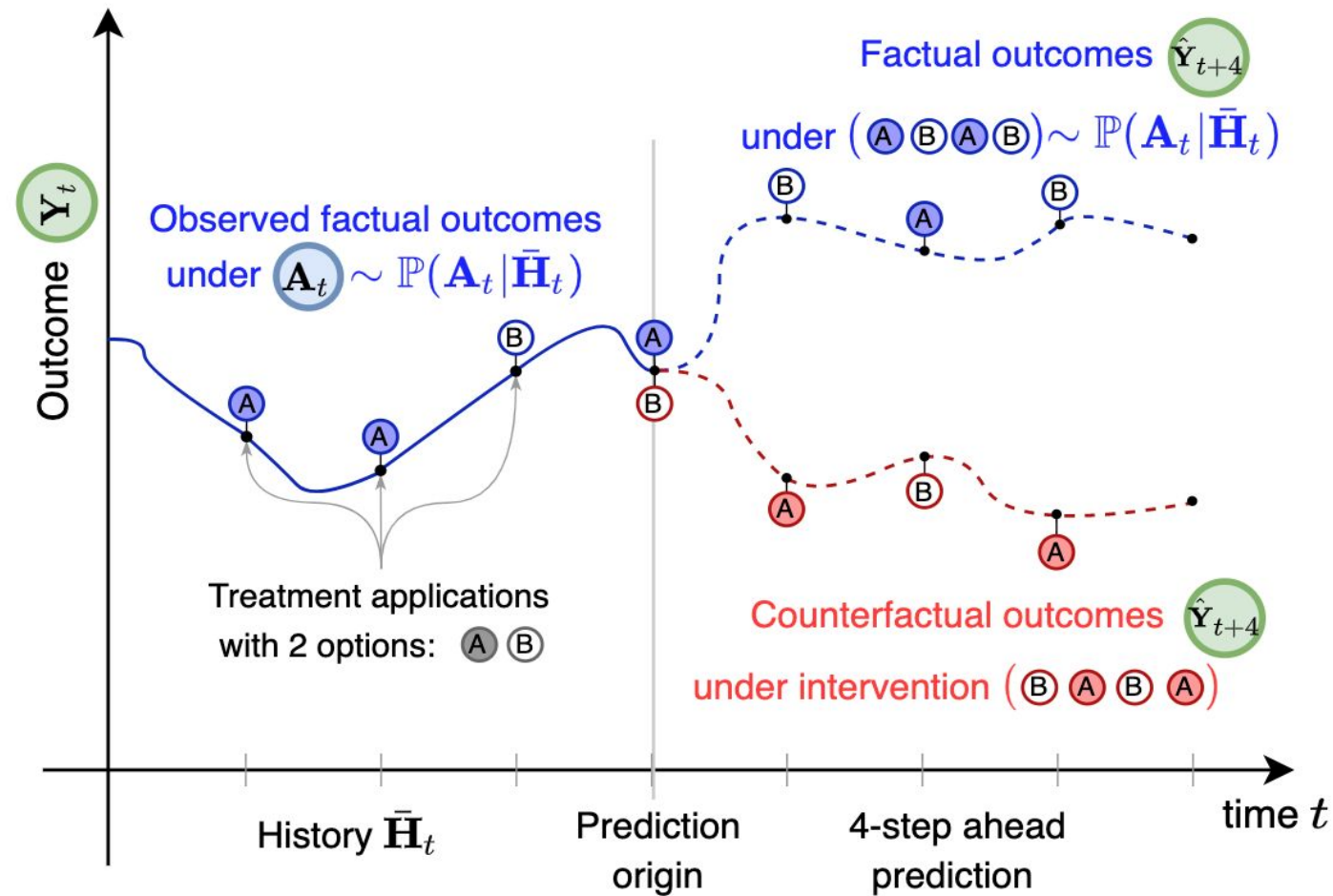
Challenges: finite-sample non-asymptotic uncertainty (e.g., getting finite-sample valid CI for ATE are impossible); randomness of treatment effects (variance of the treatment effect is not identifiable)

Extensions: Hidden confounding / sensitivity models



Challenges: each sensitivity model has its own identification and estimation machinery (no unified software)

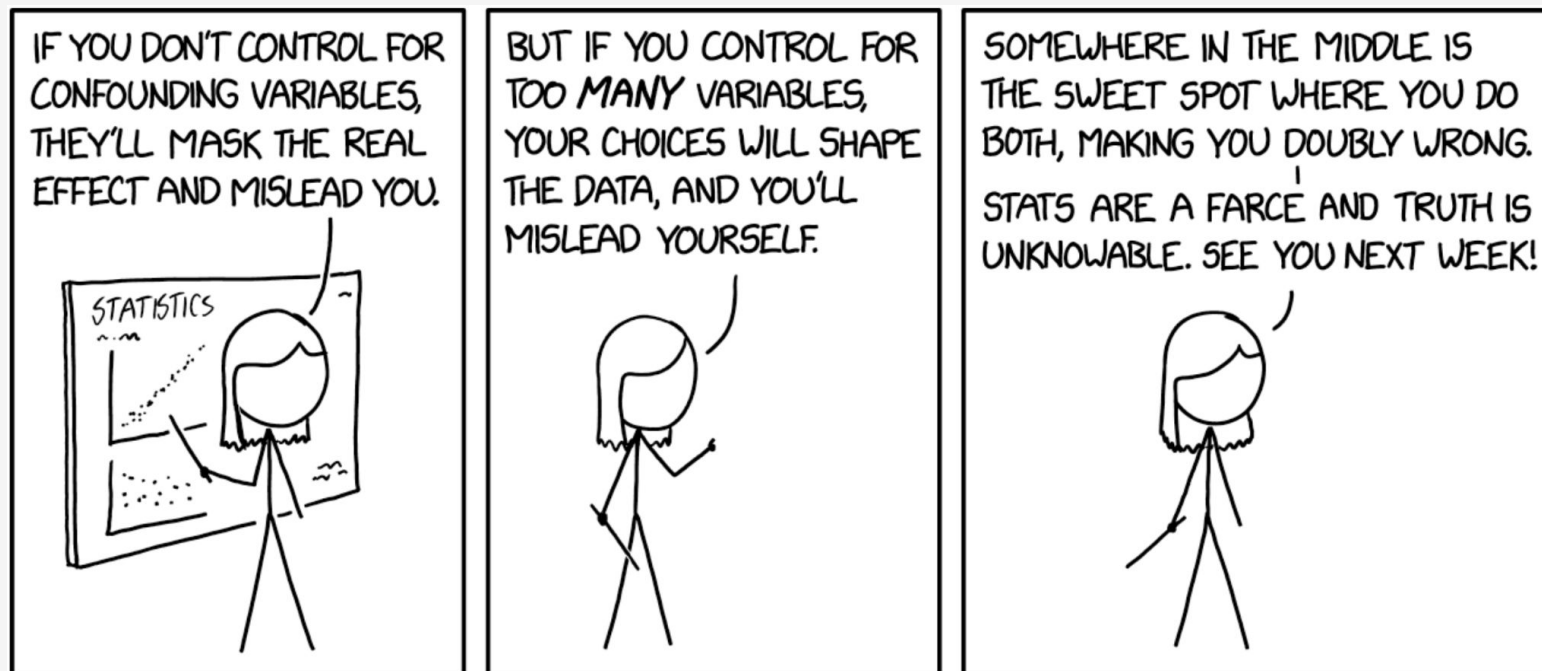
Extensions: Panel data / time-varying potential outcomes



Challenges: complexity of implementation and validation; curse of long horizons (variance of efficient estimator grows exponentially)

Extensions: Works in our lab

Uncertainty quantification	• Epistemic (estimation) uncertainty for CATE / CAPO • Aleatoric uncertainty for POs (Melnychuk et al. 2023), TEs (Melnychuk et al., 2024) • Total uncertainty for CATE and CAPOs with conformal prediction • Bayesian causal inference (Melnychuk et al. 2026, Javurek et al. 2026)
Hidden confounding	• Marginal sensitivity model, general sensitivity model (Frauen et al. 2023), B-learner • Partial identification with instrumental variables (Schweisthal et al. 2025) • Proxy variables
Time-varying potential outcomes	• G-computation (Hess et al. 2026) • Irregular sampling times / continuous time (Hess et al. 2025) • Reinforcement learning (Javurek et al. 2026)
Foundation models	• Causal foundation models (Ma et al. 2026)
Explainability Interpretability	• Measuring the amount of observed confounding with SHAP values (Brockschmidt et al. 2026)



<https://xkcd.com/2560/>



Thank you for your attention!

Main message: causal ML works only if the assumptions hold → study design comes first, machine learning – second.



My web-page



Causal ML Lab page