LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

HUMAN TECHNOPOLE

Tutorial: Causal ML for treatment effect estimation

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

19.03.2026 | Hosted by Dr. Florian Jug
Internal Seminar @ Computational Biology Research Centre
Human Technopole, Milano

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

- Our team:
  - Prof. Dr. Stefan Feuerriegel

  - 2 co-directors (soon to graduate)


  - 8 PhD students

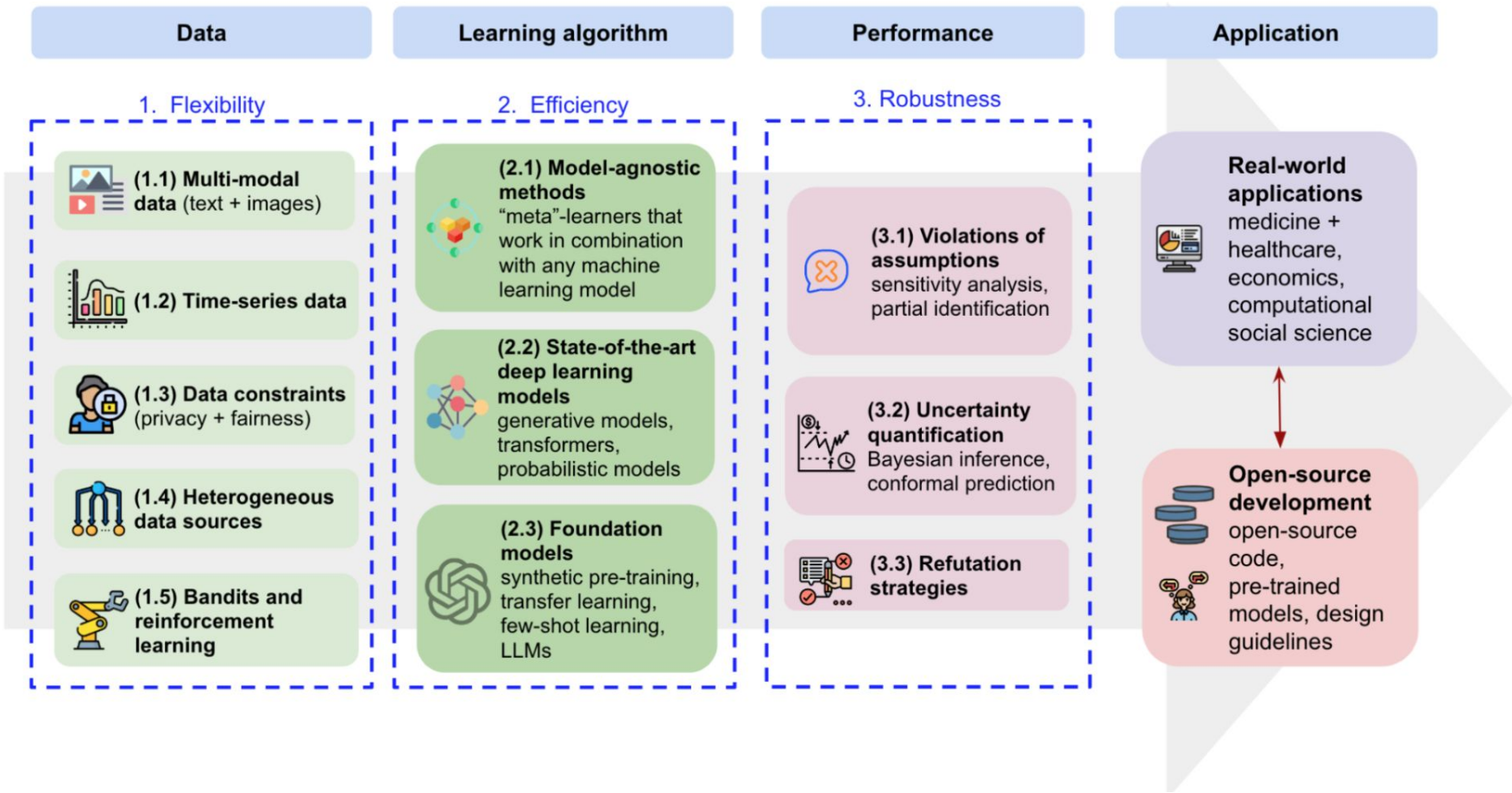







About our group: Causal ML Lab @ Institute of AI in Management

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



Introduction

- Causal Machine Learning
- Treatment effect estimation from observational data
- Motivational example
- Problem formulation
- Fundamental problem of causal inference
- Spectrum of causal estimands

https://www.smbc-comics.com/comic/autism-and-vaccines

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: Treatment effect estimation from observational data

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

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Pearl's ladder of causation

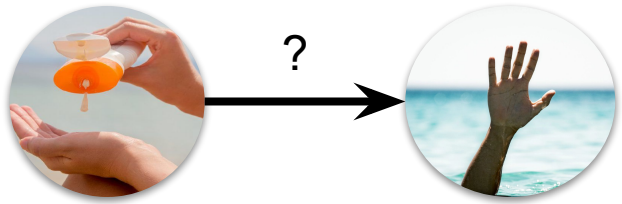
- Gold standard, Randomized controlled trials (RCTs), are expensive / unethical
- Abundance of the observational data
- Recent advances in ML/DL provide many tools

Introduction: Motivational example

- Why is answering interventional/counterfactual questions from observational data challenging?

- Consider the following study:

Observational data:
  - sunscreen usage (binary treatment)
  - number of drowning-related deaths (outcome)



Aim: effect of sunscreen on the chance of drowning

- **Evidence**: The higher the usage of sunscreen -> the higher is the chance of drowning
- **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:

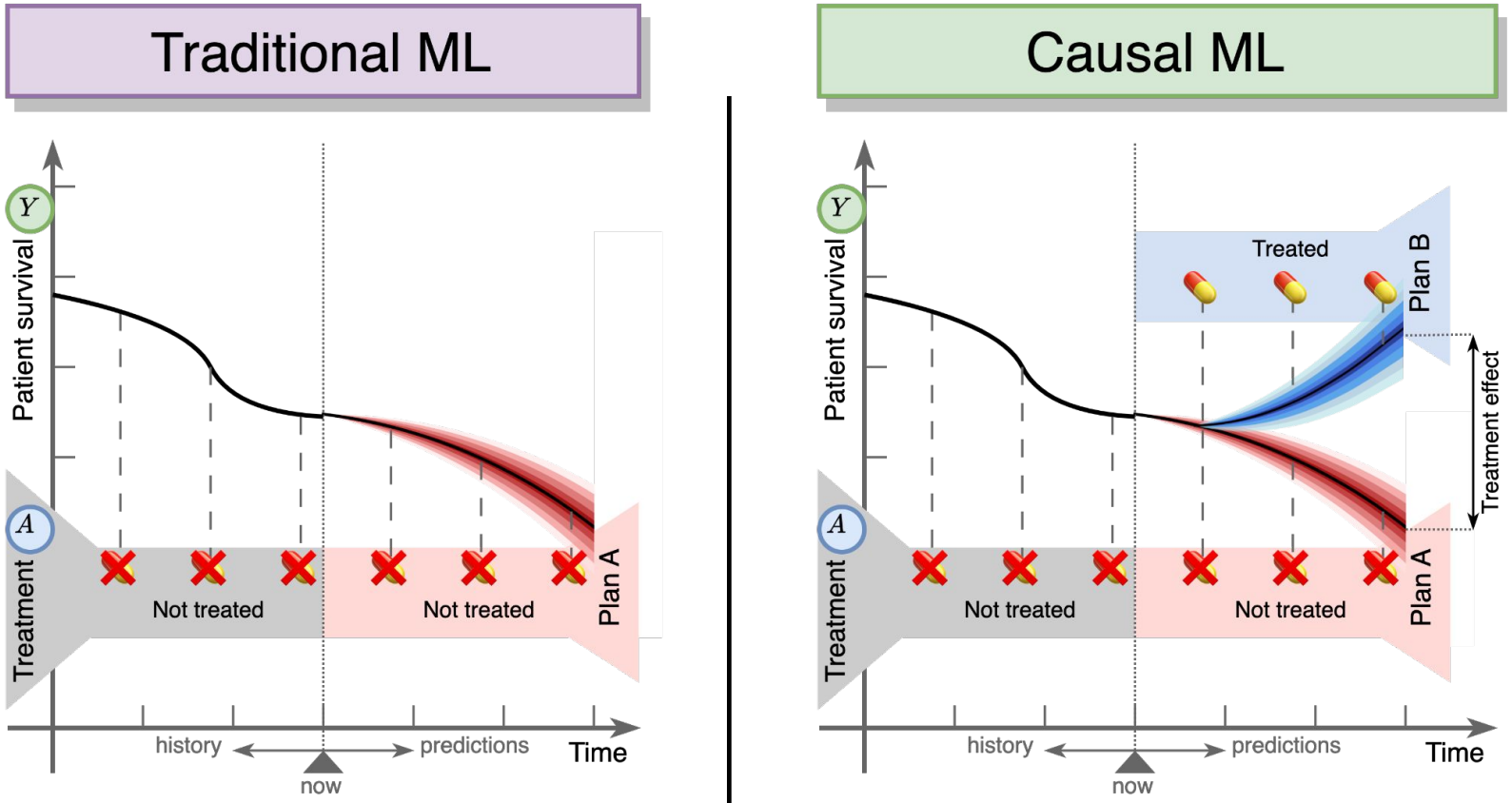
Observational data:
- sunscreen usage (binary treatment)
- number of drowning-related deaths (outcome)
- **intensity of sunlight (covariates)**

Aim: effect of sunscreen on the chance of drowning

- **Evidence**: No association between sunscreen usage and chance of drowning in each group of sunlight
- **Comparing with the previous slide:** Intensity of sunlight is a confounder

Introduction: Motivational example

- Causal ML allows to move from diagnostics (predictive inference) to **therapeutics** (prescriptive inference)

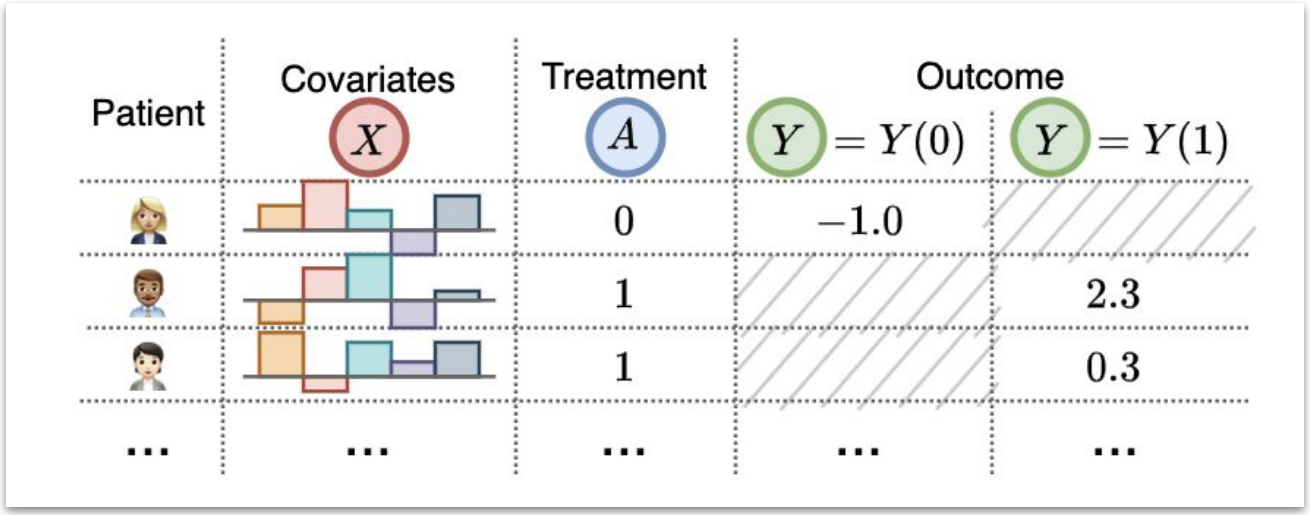


Feuerriegel, Stefan, et al. "Causal machine learning for predicting treatment outcomes." Nature Medicine 30.4 (2024): 958-968.

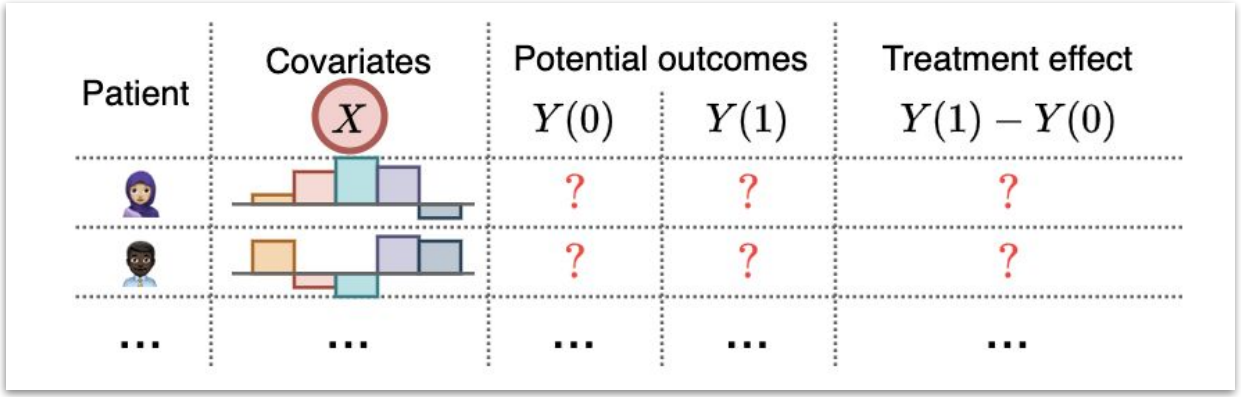
Introduction: Problem formulation

- 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
A (binary) treatments
Y continuous (factual) outcomes



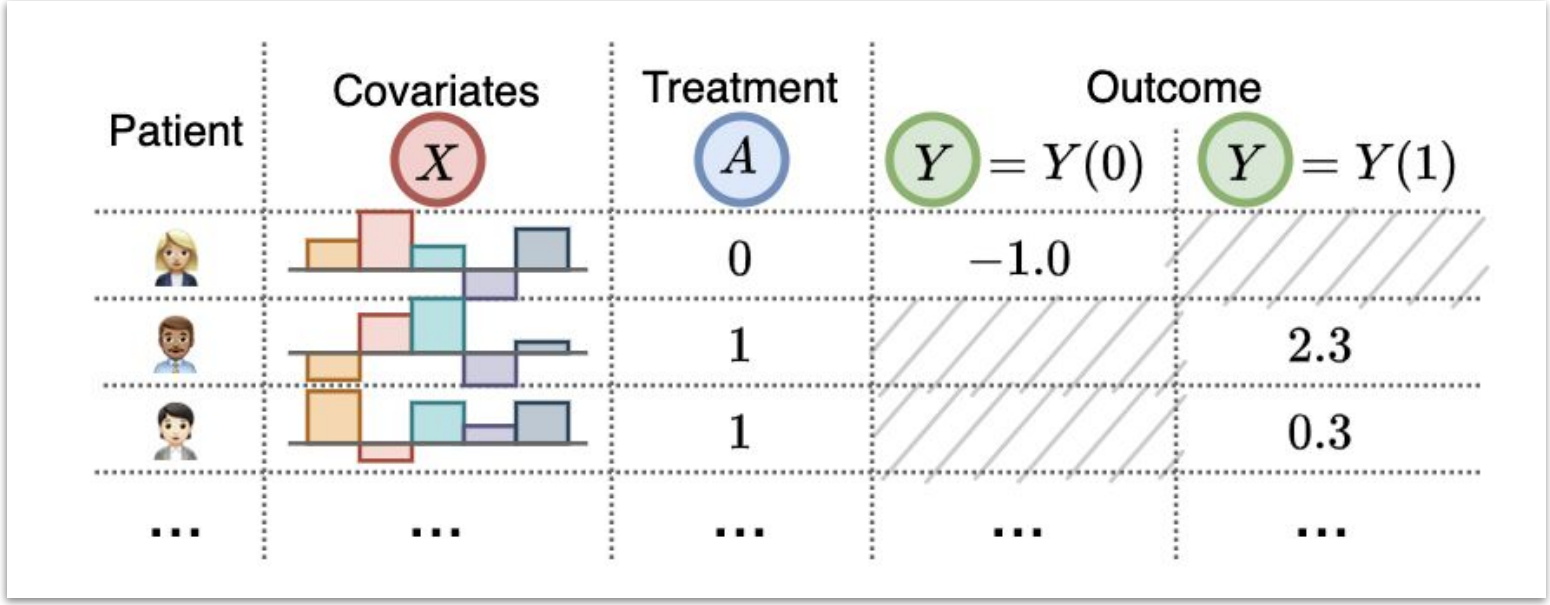
- We want to predict:
  - **treatment effects** $Y[1] - Y[0]$
  - **counterfactual (potential) outcomes** $Y[0]$ $Y[1]$



Introduction: Fundamental problem of causal inference

- **Both** potential outcomes (factual and counterfactual) are never observed for any individual -> treatment effects are never observed

- Potential outcomes are only observed for parts of the population -> **covariate shift**

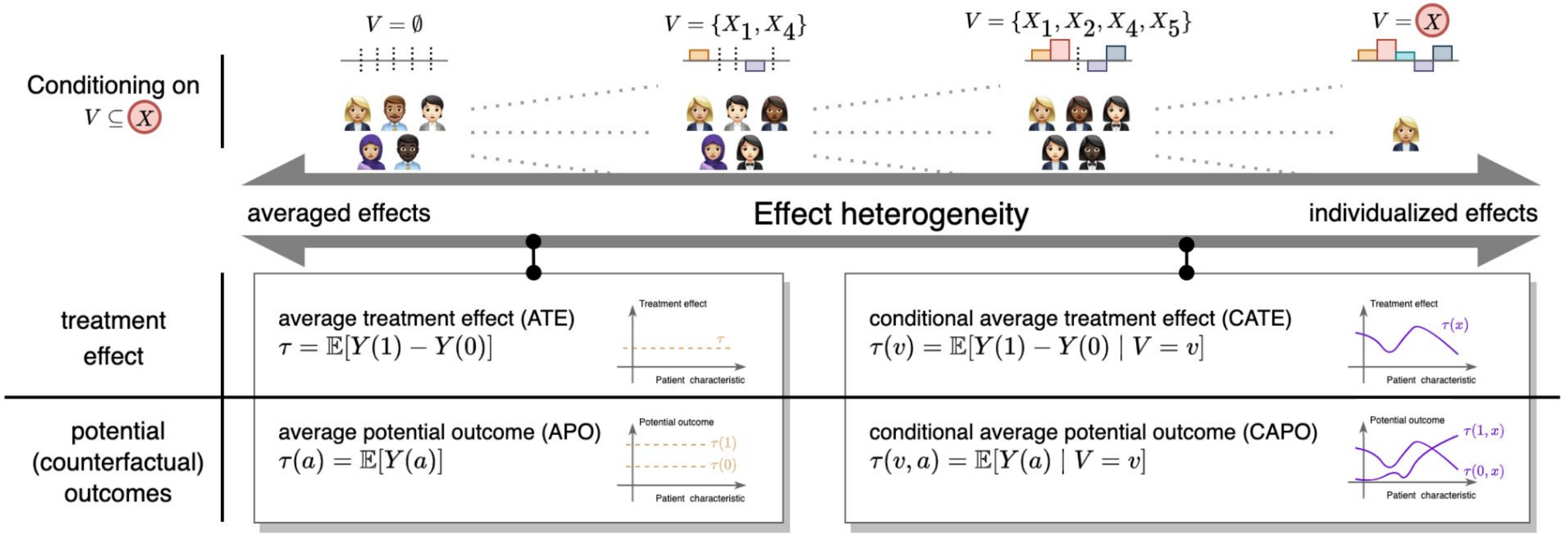


Introduction: Fundamental problem of causal inference

- **Both** potential outcomes (factual and counterfactual) are never observed for any individual -> treatment effects are never observed

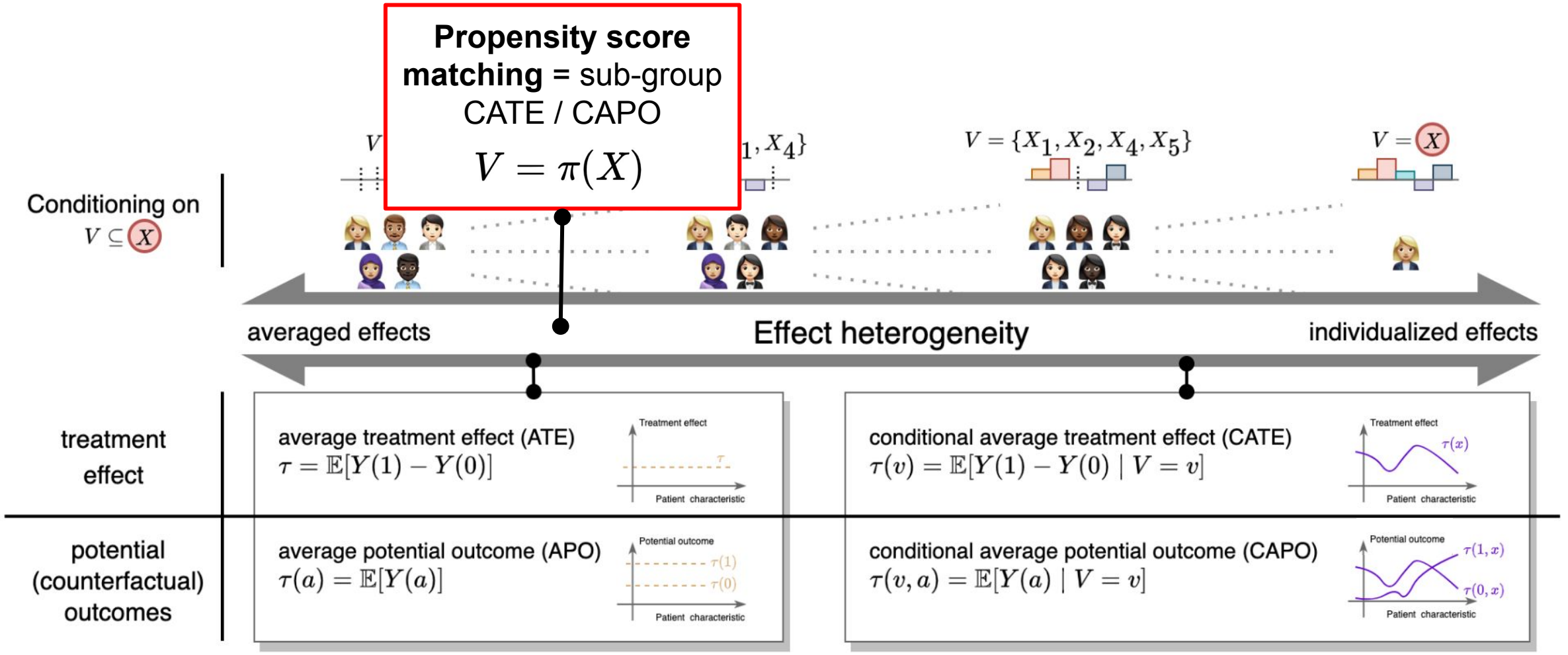
- Potential outcomes are only observed for parts of the population -> **covariate shift**

Introduction: Spectrum of causal estimands



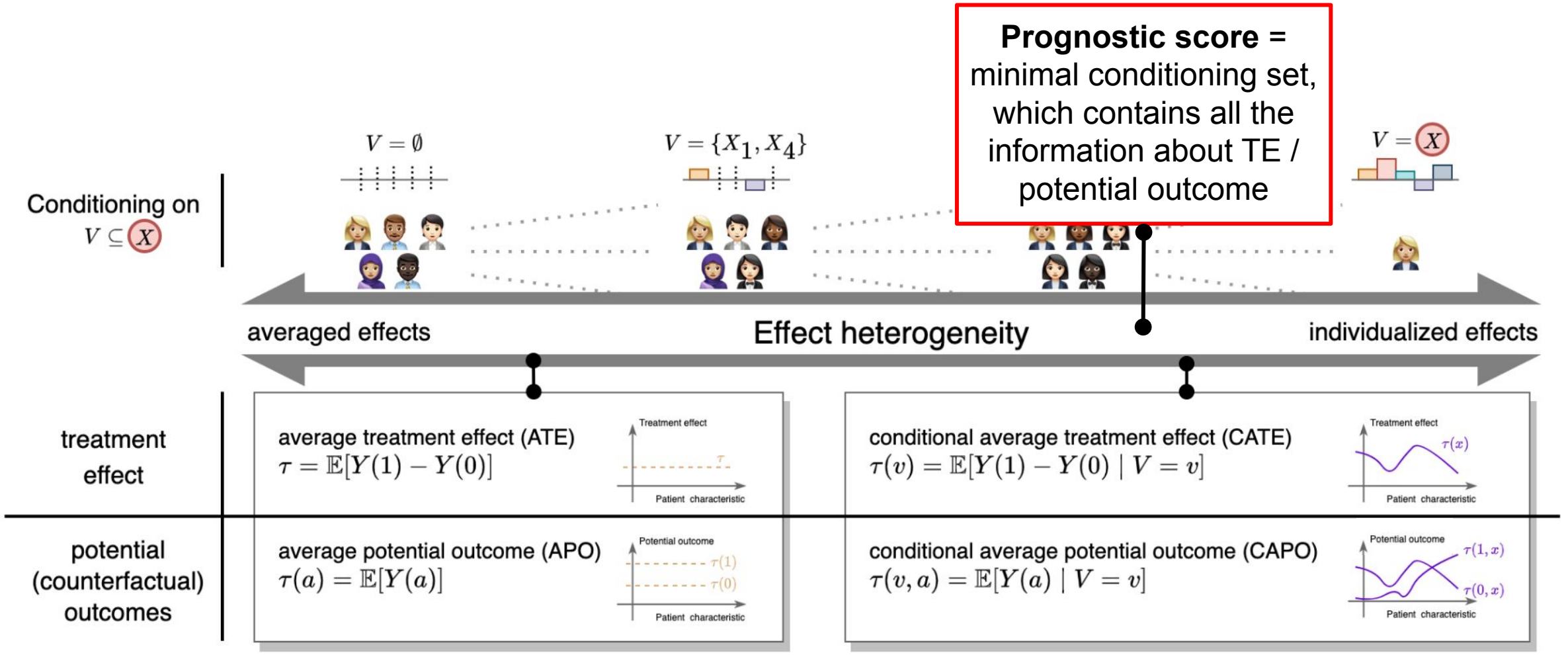
Feuerriegel, Stefan, et al. "Causal machine learning for predicting treatment outcomes." Nature Medicine 30.4 (2024): 958-968.

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.

Causal assumptions

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

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

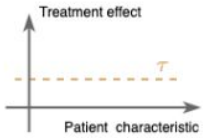
$$\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^{n} \sim \mathbb{P}(X, A, Y)$$

Potential outcomes framework (Neyman-Rubin)

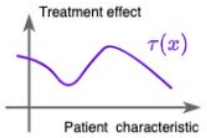
Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

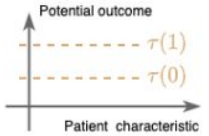
average treatment effect (ATE)
$\tau = \mathbb{E}[Y(1) - Y(0)]$



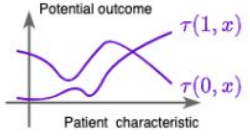
conditional average treatment effect (CATE)
$\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$



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: Frameworks

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

Abstract minimal assumptions

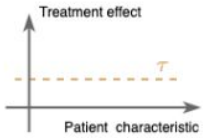
(i) Consistency
(ii) Positivity (Overlap)
(iii) Exchangeability
(Ignorability)

Potential outcomes framework (Neyman-Rubin)

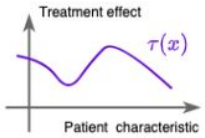
Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

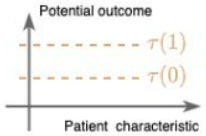
average treatment effect (ATE)
$\tau = \mathbb{E}[Y(1) - Y(0)]$



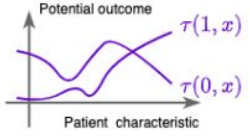
conditional average treatment effect (CATE)
$\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$



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 real, patient-individual, and (sometimes) observed
- If $A = a$ is a treatment for some patient, 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 / Exchangeability

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

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

		Verifiable with infinite observational data?
(i) Consistency	• **Informal**: Potential outcomes are real, patient-individual, and (sometimes) observed • If $A = a$ is a treatment for some patient, 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: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$	✅ (!beware of the curse of dimensionality)
(iii) Ignorability / Unconfoundedness / Exchangeability	• **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 we can speculate about plausibility with sensitivity models)

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

Identifiability with potential outcomes framework

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

- back-door adjustment / regression adjustment (RA))
  - CATE $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x] = \mu_1(x) - \mu_0(x)$
  - ATE $\tau = \mathbb{E}\big[\mathbb{E}[Y \mid A = 1, X] - \mathbb{E}[Y \mid A = 0, X]\big] = \mathbb{E}[\mu_1(X) - \mu_0(X)]$
  - CAPO $\tau(x, a) = \mathbb{E}[Y(a) \mid X = x] = \mathbb{E}[Y \mid A = a, X = x] = \mu_a(x)$
  - APO $\tau(a) = \mathbb{E}\big[\mathbb{E}[Y \mid a, X]\big] = \mathbb{E}[\mu_a(X)]$
- inverse propensity of treatment weighting (IPTW):
  - CATE $\tau(x) = \mathbb{E}\left[\left(\frac{A}{\pi_1(X)} - \frac{1-A}{1-\pi_1(X)}\right)Y | X = x\right]$
  - ATE $\tau = \mathbb{E}\left[\left(\frac{A}{\pi_1(X)} - \frac{1-A}{1-\pi_1(X)}\right)Y\right]$
  - CAPO $\tau(x, a) = \mathbb{E}\left[\frac{1(A=a)}{\pi_a(X)}Y | X = x\right]$
  - APO $\tau(a) = \mathbb{E}\left[\frac{1(A=a)}{\pi_a(X)}Y\right]$

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

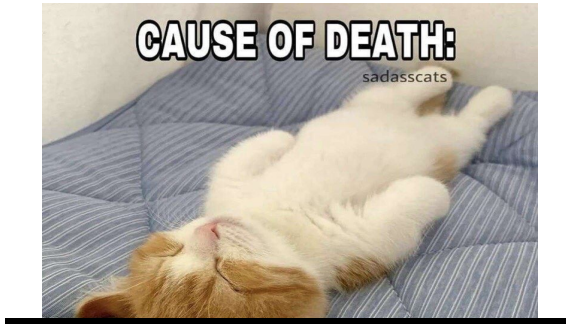
Choosing covariates

- According to econometricians: **All the pre-treatment covariates are fine**.
  - ground-truth confounders (A <- X -> Y)
  - instruments (A <- X)
  - background noise (X / X -> Y)
- However, with more covariates, causal estimation becomes harder

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

Choosing covariates

- According to econometricians: **All the pre-treatment covariates are fine**.
  - ground-truth confounders (A <- X -> Y)
  - instruments (A <- X)
  - background noise (X / X -> Y)
- However, with more covariates, causal estimation becomes harder
- When adjusting for a post-treatment covariate, we are inducing a bias -> a **kitten dies**



Post-treatment covariate adjustment

Causal assumptions: Frameworks

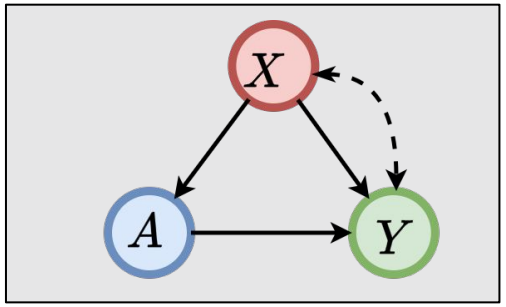
$$\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n \sim \mathbb{P}(X, A, Y)$$

Potential outcomes framework (Neyman-Rubin)

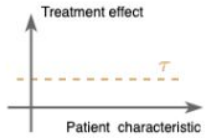
Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

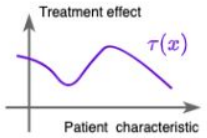
Assumptions can be related to the structural knowledge



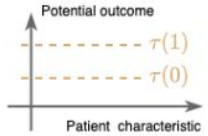
average treatment effect (ATE)
$\tau = \mathbb{E}[Y(1) - Y(0)]$



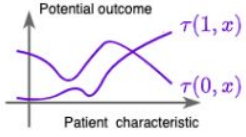
conditional average treatment effect (CATE)
$\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$



average potential outcome (APO)
$\tau(a) = \mathbb{E}[Y(a)]$



conditional average potential outcome (CAPO)
$\tau(v, a) = \mathbb{E}[Y(a) \mid V = v]$

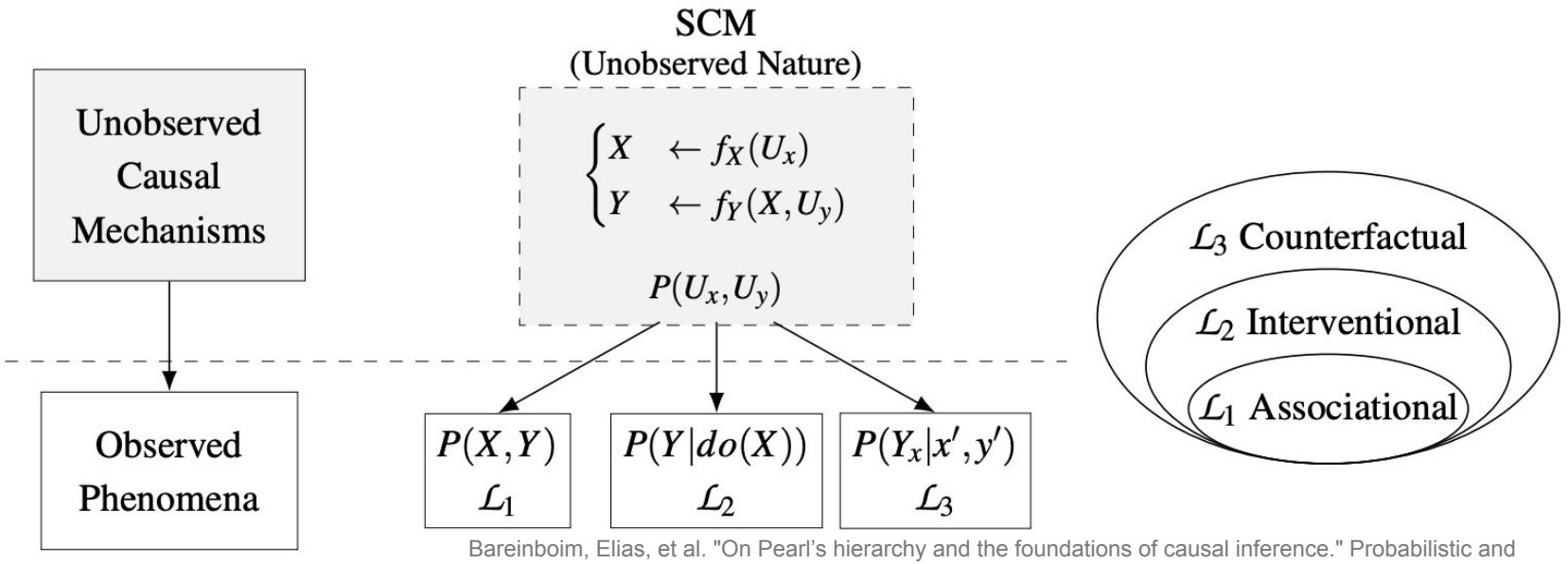


Causal assumptions: Structural causal model (SCM)

Verifiable with infinite observational data?

- **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}

SCM



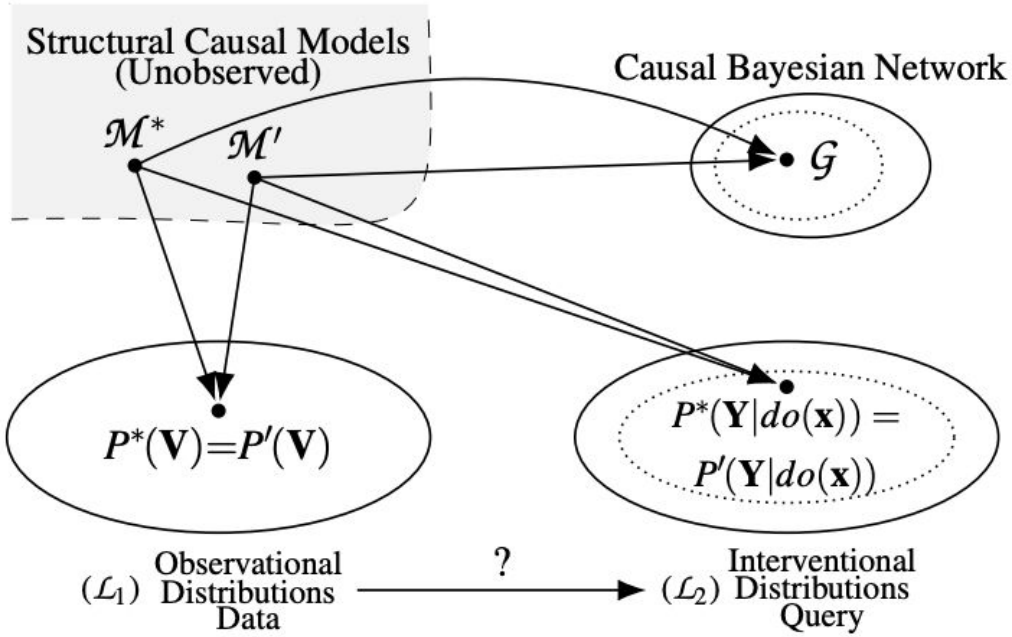
Bareinboim, Elias, et al. "On Pearl's hierarchy and the foundations of causal inference." Probabilistic and causal inference: the works of Judea Pearl. 2022. 507-556.

- All the L1, L2, L3 queries can inferred with the probability calculus, including, **CATE/ATE** and **CAPO/APO** -> unnecessary strong assumption

Causal assumptions: Causal diagram

- **Informal**: Causal diagram (Causal DAG, Causal Bayesian network) encodes **structural constraints** of an SCM: **conditional dependencies / independencies** for L1 and L2 distributions
- Every SCM induces a causal diagram. Every causal diagram encompasses a class of SCMs.

Causal diagram

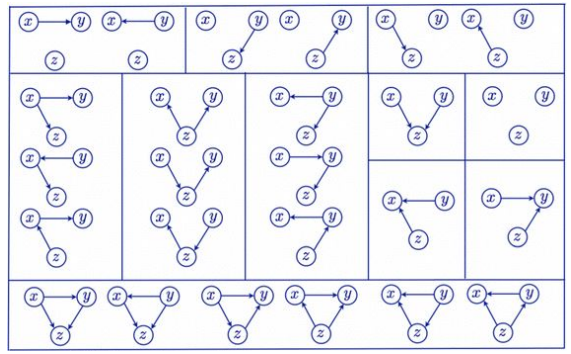


Bareinboim, Elias, et al. "On Pearl's hierarchy and the foundations of causal inference." Probabilistic and causal inference: the works of Judea Pearl. 2022. 507-556.

Verifiable with infinite observational data?

❌

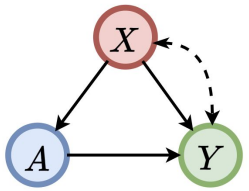
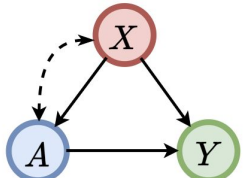
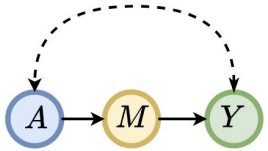
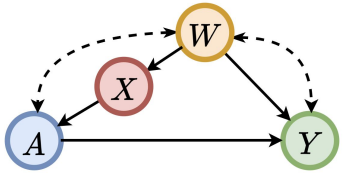
(only Markov equivalence class is identifiable, for Markovian diagrams)



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

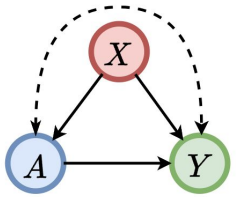
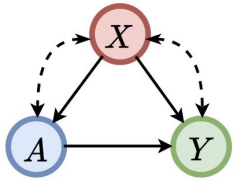
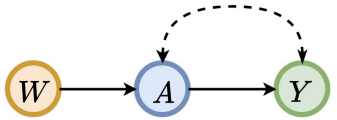
Query:	Causal diagram:	ID:	Formula:
CATE / CAPO		✅	- back-door adjustment - propensity reweighting
CATE / CAPO		✅	- back-door adjustment - propensity reweighting
ATE / APO		✅	- front-door adjustment
ATE / APO		✅	- napkin formula

- 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 / CAPO		❌	(Hidden Confounding)
CATE / CAPO		❌	(Butterfly-bias)
ATE / APO		❌	(Hidden Confounding)
	+ functional assumption on Y	✅	- Instrumental variable - Mendelian randomization

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

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)

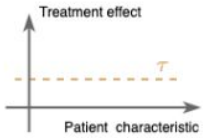
=

Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

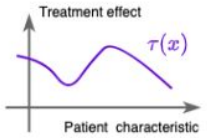
average treatment effect (ATE)

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



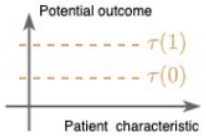
conditional average treatment effect (CATE)

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



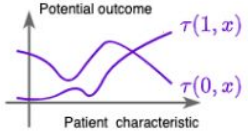
average potential outcome (APO)

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



conditional average potential outcome (CAPO)

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



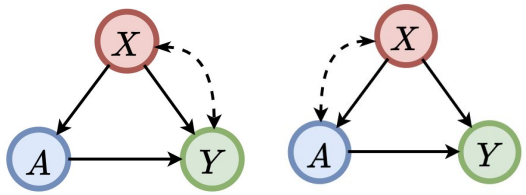
Causal assumptions: Equivalence of the frameworks

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

Equivalence of assumptions

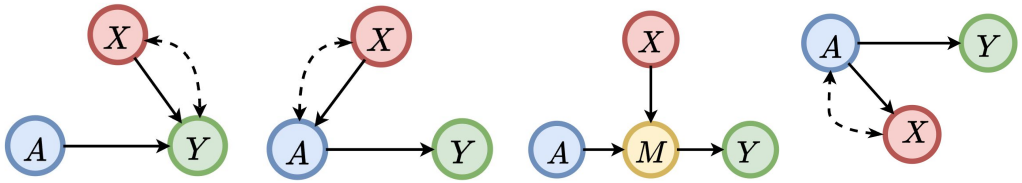
(i) Causal diagrams, where:

- back-door adjustment for X should be applied



(i) Consistency
(iii) Ignorability

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



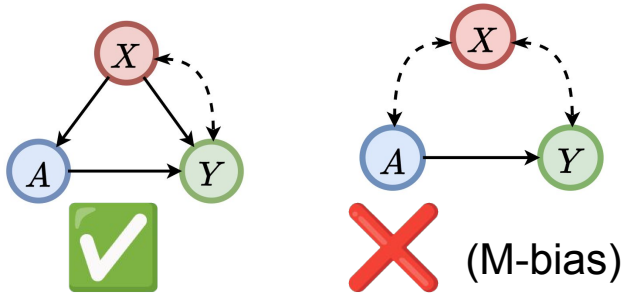
(ii) Positivity

(ii) Positivity

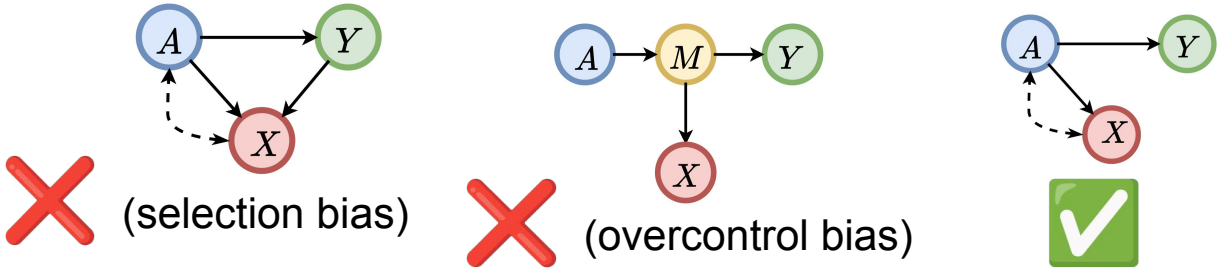
Causal assumptions: Equivalence of the frameworks

Choosing covariates (revisited)

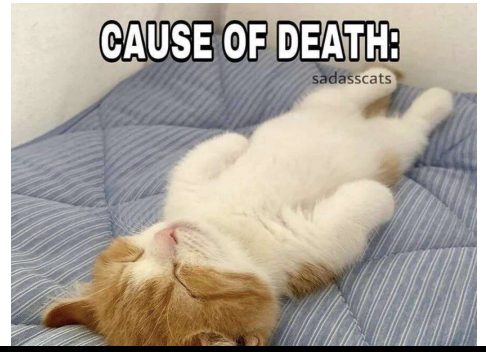
- **Almost all** pre-treatment covariates are fine except for (rarely) variables, that can induce **M-bias**



- Most of the post-treatment covariate adjustments lead to the **death of a kitten**



- See (Cinelli et al. 2022) for details.



(Most of the) post-treatment covariate adjustments or M-bias

Causal assumptions: Case study

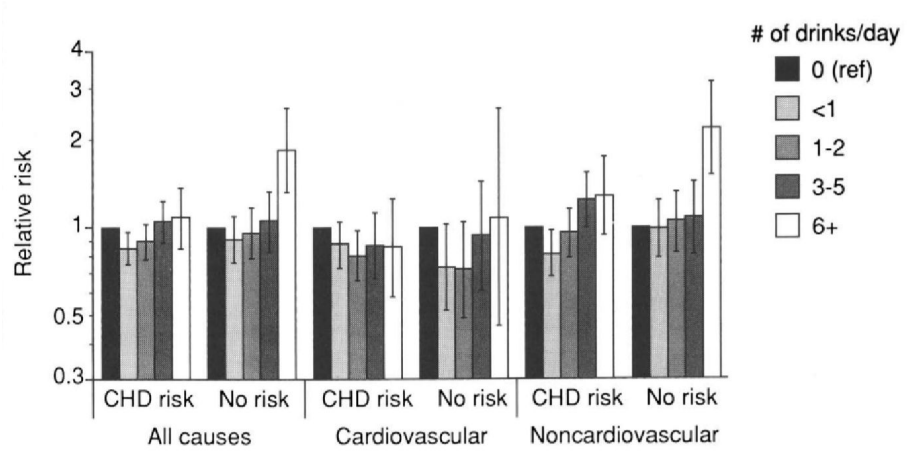
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

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 re-

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

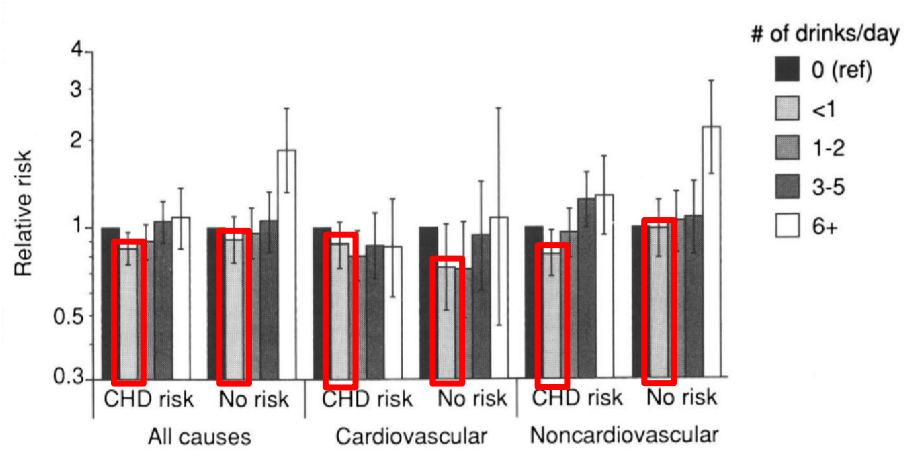
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

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 re-

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

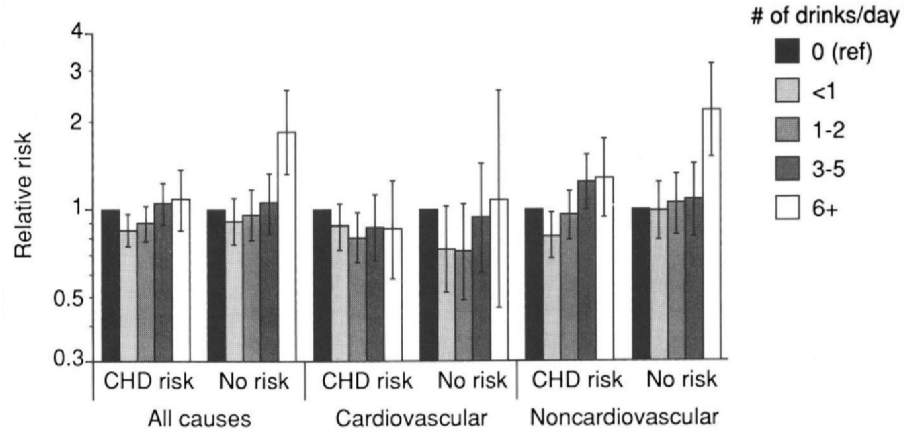
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

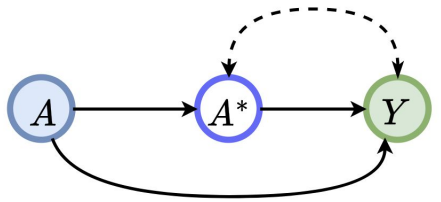
Observational data:

- 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)



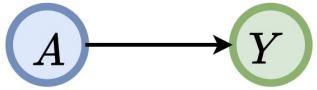


Causal assumptions: Case study

- 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

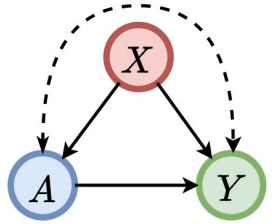
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

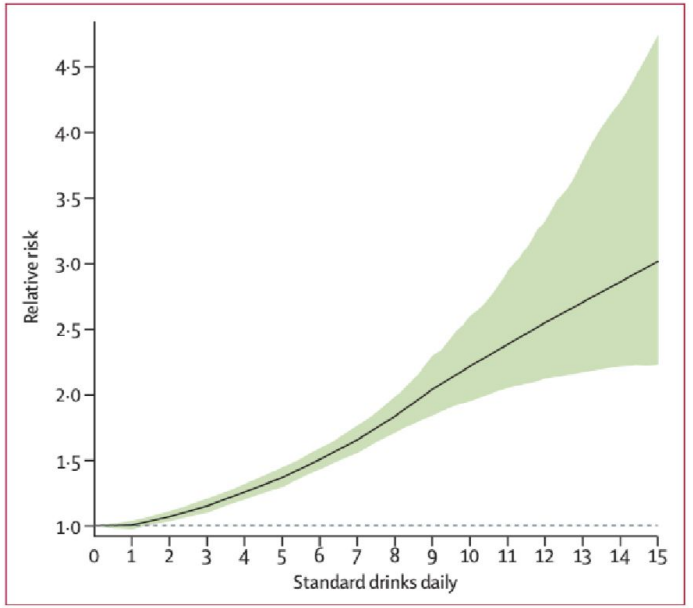
Observational data:

- 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

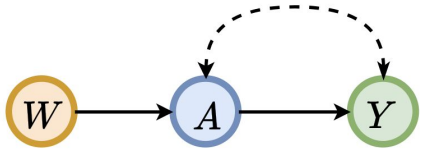
- Covariates selection is very tricky for observational studies
- Consider studies of effect of alcohol consumption on general health

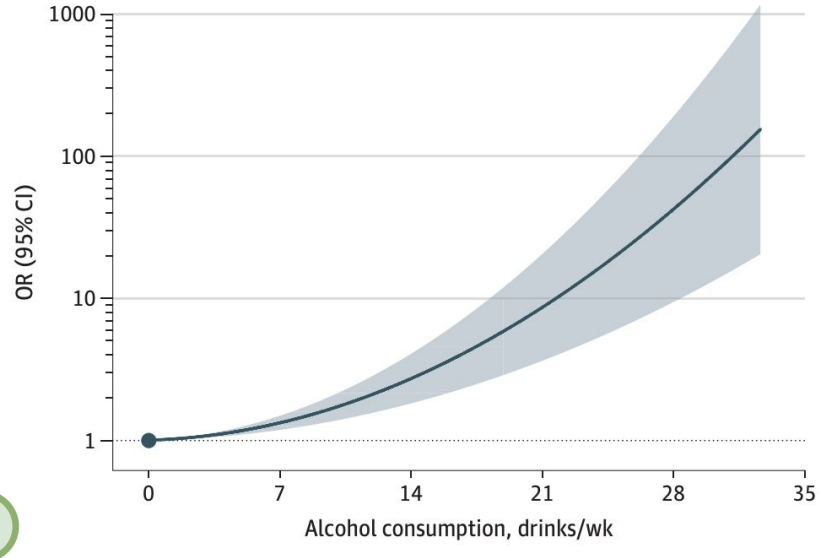
Observational data:

- 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






JAMA Network | Open

Original Investigation | Cardiology

Association of Habitual Alcohol Intake With Risk of Cardiovascular Disease

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

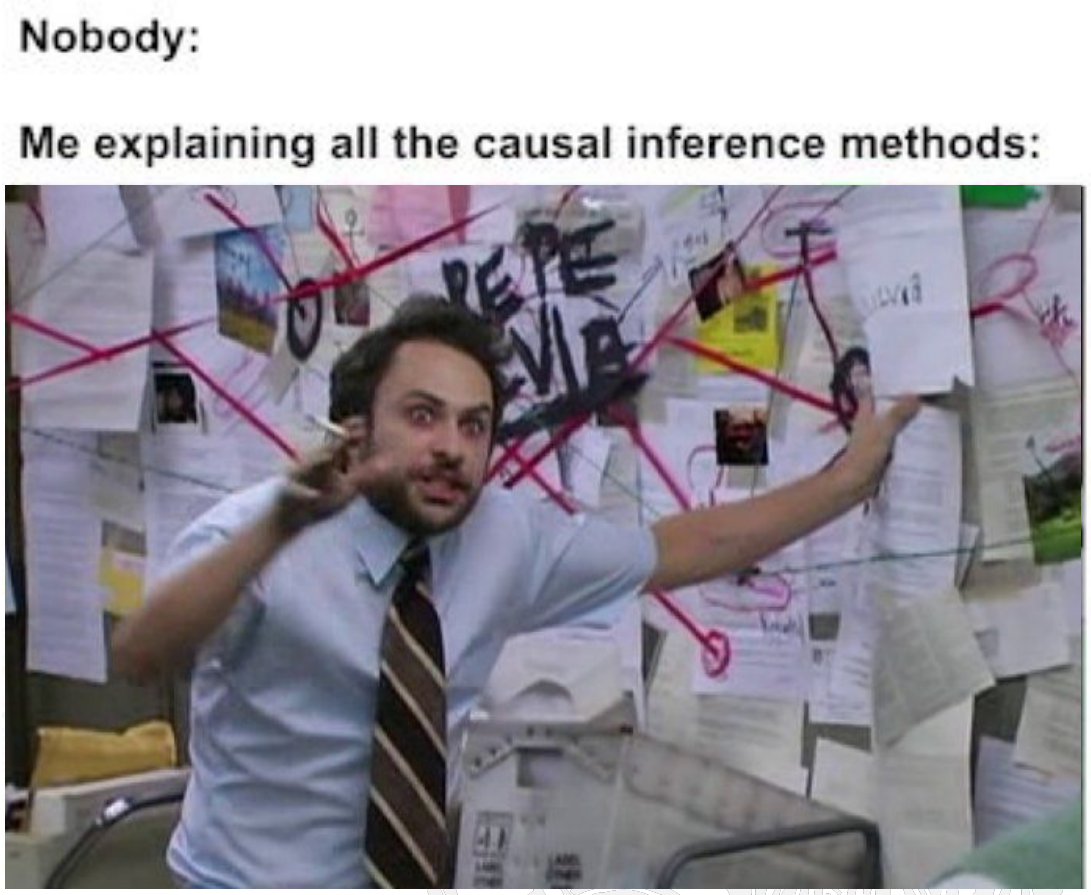
LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

HUMAN TECHNOPOLE

ML and estimation

- Big picture
- Plug-in estimators / single-stage plug-in learners
- Issues of plug-in estimation
- 1. “What about the sub-group treatment effects?”
- 2. “How to regularize $\hat{\tau}(x)$?”
- 3. “What is better, adjustment or IPTW?”
- 4. “Can we do model selection / hyperparameter tuning?”
- 5. “How to address the covariate shift?”
- 6. “Can we incorporate inductive biases for nuisance functions estimation?”
- 7. “How can representation learning help?”



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, BNN, Causal Forest)
- T-learner (T-Net, TARNet, DragonNet, CFR & variants)
- Mixed approaches (FlexTENet)

Debiased learners*:
- EP-learner

Two-stage learners:

Plug-in learners:
- IPW-learner
- RA-learner / X-learner (GANITE)
- 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

CAPO 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, BNN)
- T-learner (T-Net, TARNet, DragonNet, CFR & variants)
- Mixed approaches (FlexTENet)

Debiased learners*:
- i-learner

Two-stage learners:

Plug-in learners:
- IPW-learner
- RA-learner / X-learner (GANITE)

Debiased learners:
- DR-learner

APO estimation: estimating a parameter

Plug-in estimators:
- Plug-in estimator
- RA estimator
- IPTW estimator

Debiased estimators:
- A-IPTW estimator
- TMLE estimator

ML and estimation: Plug-in estimators / single-stage plug-in learners

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, BNN, Causal Forest)
- T-learner (T-Net, TARNet, DragonNet, CFR & variants)
- Mixed approaches (FlexTENet)

Debiased learners*:
- EP-learner

Two-stage learners:

Plug-in learners:
- IPW-learner
- RA-learner / X-learner (GANITE)
- 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: Plug-in estimators / single-stage plug-in learners

Plug-in estimators – 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

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

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

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

- We can learn nuisance functions either as a joint Single model (**S-learner**) or as a Two separate models (**T-learner**).

ML and estimation: Issues of plug-in estimation

Issues of plug-in estimators / plug-in learners

Problem solved? **NO!**

1. What about the sub-group treatment effects (we still need to adjust for the full X)?
2. How to regularize $\hat{\tau}(x)$?
3. What is better, adjustment or IPTW? Can we do even better (e.g., more efficient, more robust) in estimating CATE / ATE?
4. Can we do model selection / hyperparameter tuning?
5. $\hat{\mu}_a(x)$ can only be well estimated for some parts of the population, e.g., only in treated group. How to address the covariate shift?
6. Can we incorporate inductive biases for nuisance functions?
7. How can representation learning help?

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):
  - **CATE** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\big((Y[1] - Y[0] - \hat{\tau}(V)\big)^2$
  - **CAPO** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\big((Y[a] - \hat{\tau}(V, a)\big)^2$
- But, we face the fundamental problem of causal inference

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):
  - **CATE** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\big((Y[1] - Y[0] - \hat{\tau}(V)\big)^2$ never observed
  - **CAPO** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\big((Y[a] - \hat{\tau}(V, a)\big)^2$ sometimes observed
- But, we face the fundamental problem of causal inference

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):
  - **CATE** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\big((Y[1] - Y[0] - \hat{\tau}(V)\big)^2$
  - **CAPO** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\big((Y[a] - \hat{\tau}(V, a)\big)^2$
- But, we face the fundamental problem of causal inference
- **Idea:** machine learning with the nuisance functions
  - **CATE** $\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E}\big(\tau(X) - \hat{\tau}(V)\big)^2$
  - **CAPO** $\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E}\big(\tau(X, a) - \hat{\tau}(V, a)\big)^2$ $\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E}\big(\frac{1(A=a)}{\pi_a(X)}(Y - \hat{\tau}(V, a)\big)^2$

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

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, BNN, Causal Forest)
- T-learner (T-Net, TARNet, CFR & variants)
- Mixed approaches (FlexTENet)

Debiased learners*:
- EP-learner

Two-stage learners:

Plug-in learners:
- IPW-learner
- RA-learner / X-learner (GANITE)
- 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 (DragonNet)

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

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

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** $\tilde{Y}_{\hat{\eta}}$: with the main property $\mathbb{E}(\tilde{Y}_{\eta} \mid V = v) = \tau(v)$

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

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

- 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**

Two-stage plug-in learners

Step 1. Nuisance estimation

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

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

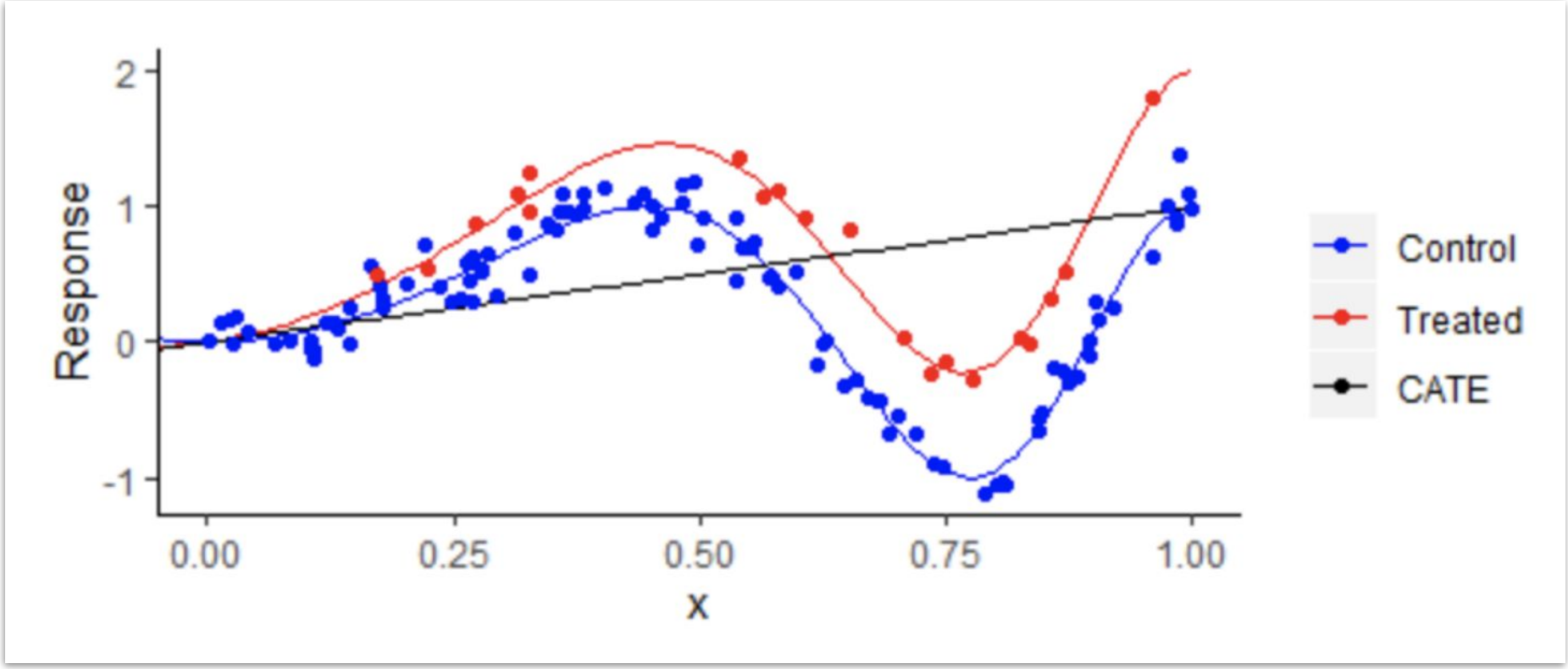
CATE
$\tilde{Y}_{\mathrm{RA},\hat{\eta}} = A(Y - \hat{\mu}_0(X)) + (1 - A)(\hat{\mu}_1(X) - Y)$ $\tilde{Y}_{\mathrm{IPTW},\hat{\eta}} = \left(\frac{A}{\hat{\pi}_1(X)} - \frac{1-A}{\hat{\pi}_0(X)} \right) Y$ $\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$

- Sample splitting needed, if too flexible models are chosen!

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

How to regularize a target model?

- What if the ground-truth CATE is a simpler function than each of the CAPOs?



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

How to regularize a target model?

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

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

- 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 -> “sub-group” CATE
  - large values of regularization hyperparameter -> ATE

ML and estimation: 3. “What is better, adjustment or IPTW?”

Optimality of the estimator / learner

- We ask ourselves, what is an optimal estimator / learner?
- Ideally, it should:
  - omit the error from estimated nuisance function as much as possible
  - have the lowest asymptotic variance (= statistical efficiency)

ML and estimation: ✅ 3. Debiased estimators / learners

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, BNN, Causal Forest)
- T-learner (T-Net, TARNet, DragonNet, CFR & variants)
- Mixed approaches (FlexTENet)

Debiased learners*:
- EP-learner

Two-stage learners:

Plug-in learners:
- IPW-learner
- RA-learner / X-learner (GANITE)
- U-learner

Debiased learners:
- DR-learner
- R-learner (DML)

ATE estimation: estimating a parameter

Plug-in estimators:
- IPTW estimator
- RA estimator

Debiased estimators:
- A-IPTW estimator
- TMLE estimator

ML and estimation: ✅ 3. Debiased estimators

Finite dimensional estimands

Asymptotically speaking:

- **ATE** are finite-dimensional estimands
- **Efficient estimation** is properly defined is a semi-parametric sense (lowest variance estimator from all the local parametric sub-models). Therein, the theory of influence functions can be used.
- **A-IPTW estimator** is efficient is a combination of both adjustment and IPTW:

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

- A-IPW estimators are **doubly-robust**: if at least one of the nuisance parameters are consistently estimated - the ATE is consistently estimated
- Alternatives: TMLE estimator (efficient), A-IPTW estimator with clipped propensities (biased, but reduces variance).

ML and estimation: ✅ 3. Debiased learners

- Analogously, we can build debiased two-step learners: doubly-robust (**DR)-learner**

Step 1. Nuisance estimation

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

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

DR-learner

CATE
$\tilde{Y}_{\mathrm{DR},\hat{\eta}} = \left(\frac{A-\hat{\pi}_1(X)}{\hat{\pi}_1(X)\hat{\pi}_0(X)}\right)\left(Y - \hat{\mu}_A(X)\right) + \hat{\mu}_1(X) - \hat{\mu}_0(X)$
$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$

- Sample splitting needed, if too flexible models are chosen!

ML and estimation: ✅ 3. Debiased learners

- DR-learner may suffer from high variance with low overlap, as we divide by the propensity score

- **Alternative idea**: use a so-called Robinson decomposition of the potential outcomes:

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

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

R-learner

- Then a custom **residual loss** is as follows:

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

- This yields **R-learner**

ML and estimation: ✅ 3. Debiased learners

- DR-learner may suffer from high variance with low overlap, as we divide by the propensity score
- Other alternative is **residualized (R)-learner** (uses a special Robinson's decomposition)

R-learner

Step 1. Nuisance estimation

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

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

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

- Sample splitting needed, if too flexible models are chosen!

ML and estimation: ✅ 3. Debiased learners

- If we would use ground-truth nuisance parameters, it turns out that the losses target at the **CATE** and the **overlap-weighted CATE**

DR-learner vs. R-learner

Nuisance parameters	DR-learner	R-learner
Estimated via debiasing	$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\text{DR},\hat{\eta}} - \hat{\tau}(V))^2$	$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}\Big((Y - \hat{\mu}(X)) - (A - \hat{\pi}_1(X))\hat{\tau}(V) \Big)^2$
Ground-truth	$\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E}\big((\tau(V) - \hat{\tau}(V)\big)^2$	$\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E}\big(\pi_1(X)\pi_0(X)\big(\tau(V) - \hat{\tau}(V)\big)\big)^2$

ML and estimation: ✅ 3. Debiased learners

- If we would use ground-truth nuisance parameters, it turns out that the losses target at the **CATE** and the **overlap-weighted CATE**

DR-learner vs. R-learner

Nuisance parameters	DR-learner	R-learner
Estimated via debiasing	$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\mathrm{DR},\hat{\eta}} - \hat{\tau}(V))^2$	$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}\Big((Y - \hat{\mu}(X)) - (A - \hat{\pi}_1(X))\hat{\tau}(V) \Big)^2$
Ground-truth	$\mathcal{L}(\hat{\tau}, \eta) =$ $\mathbb{E}\big((\tau(V) - \hat{\tau}(V)\big)^2$	$\mathcal{L}(\hat{\tau}, \eta) = \mathbb{E}\big(\pi_1(X)\pi_0(X)\big(\tau(V) - \hat{\tau}(V)\big)\big)^2$

- Overlap weighted CATE estimation: only focusing on patients, where decisions were uncertain. For many applications this may be more useful than usual CATE
- Minimization of the two losses gives **different results**, if ground-truth CATE is not in the model class for $\hat{\tau}(x)$, or when doing sub-group CATE

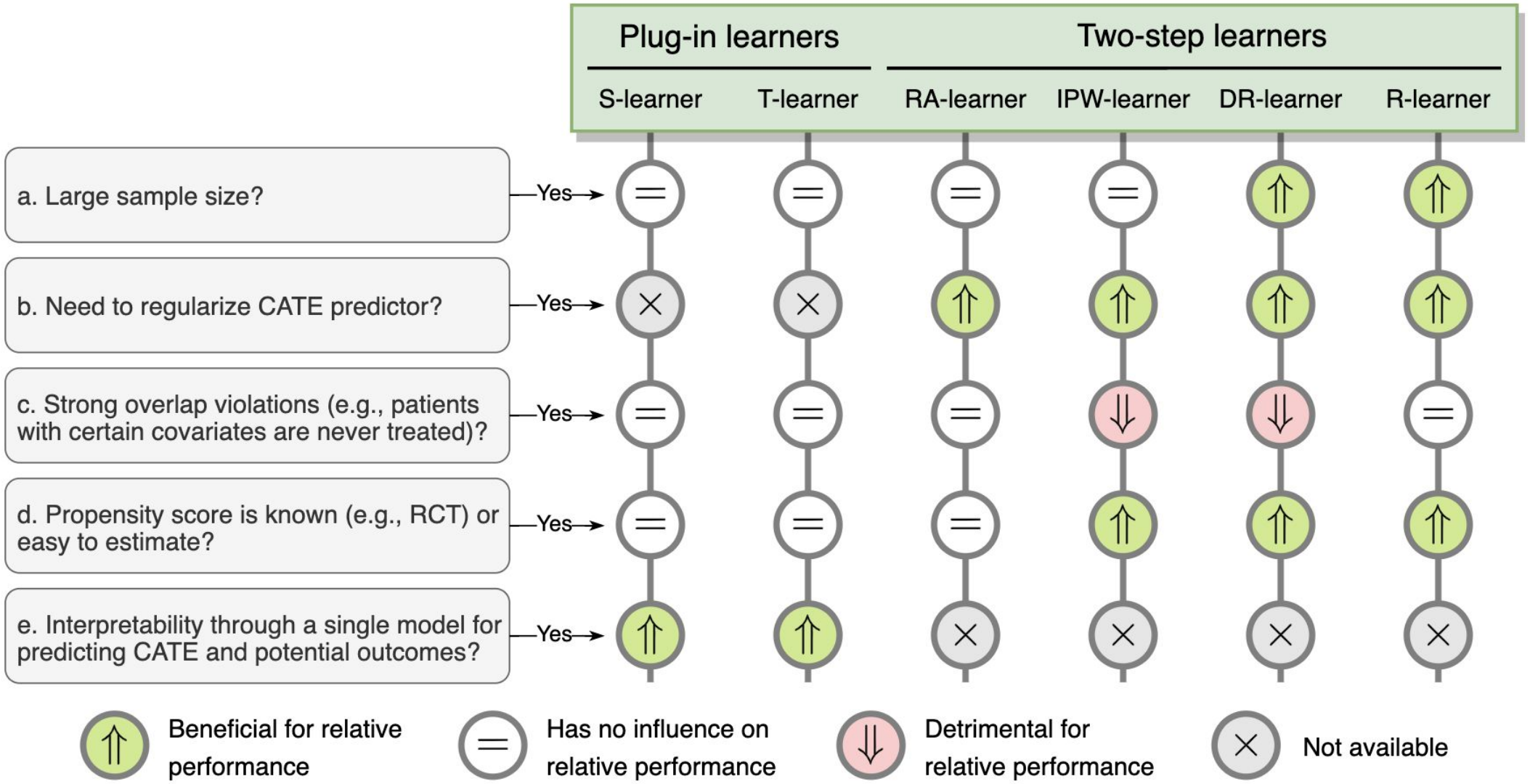
ML and estimation: ✅ 3. Debiased learners

Infinite dimensional estimands

Asymptotically speaking:

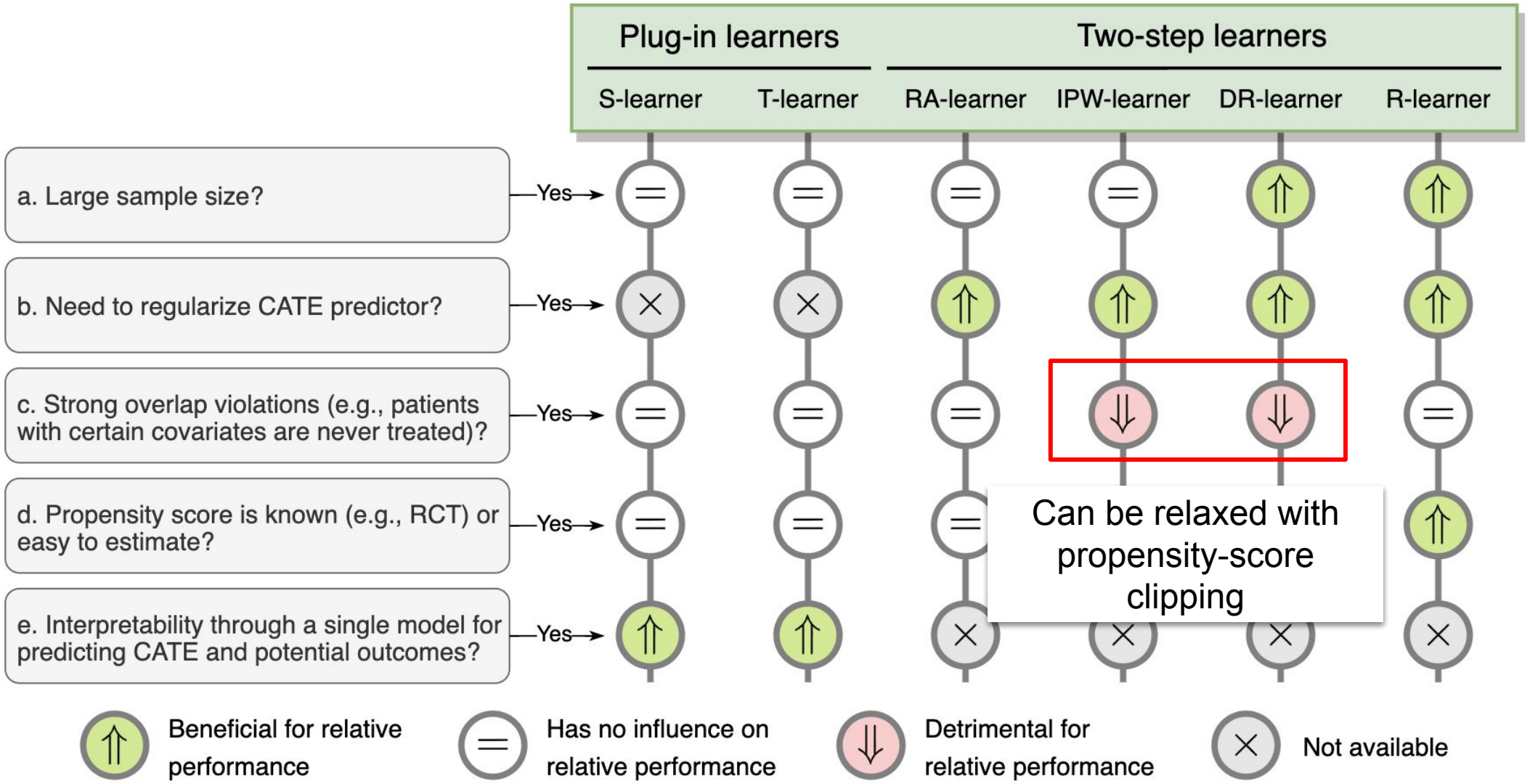
- **CATE** are functions, thus, infinite-dimensional estimands
- **No** notion of efficient estimation, but there is **Neyman orthogonality** of a loss:
  - loss is a finite-dimensional estimand
  - so we can **efficiently estimate the loss**
  - **Informally**: it says that the estimation of CATE procedures that are at most minimally affected by the estimation of nuisance parameters -> small errors in the estimated nuisance parameters have only small impact on the estimation of the target function.
- **DR- and R-learners** are Neyman orthogonal
- For CATE, Neyman orthogonality also implies **two double-robustnesses**:
  - model double-robustness (at least one nuisance is estimated consistently -> CATE is estimated consistently)
  - rate double-robustness (convergence speed is the same of the fastest convergence of the nuisance functions)

ML and estimation: Meta-learners summary for questions 1-3



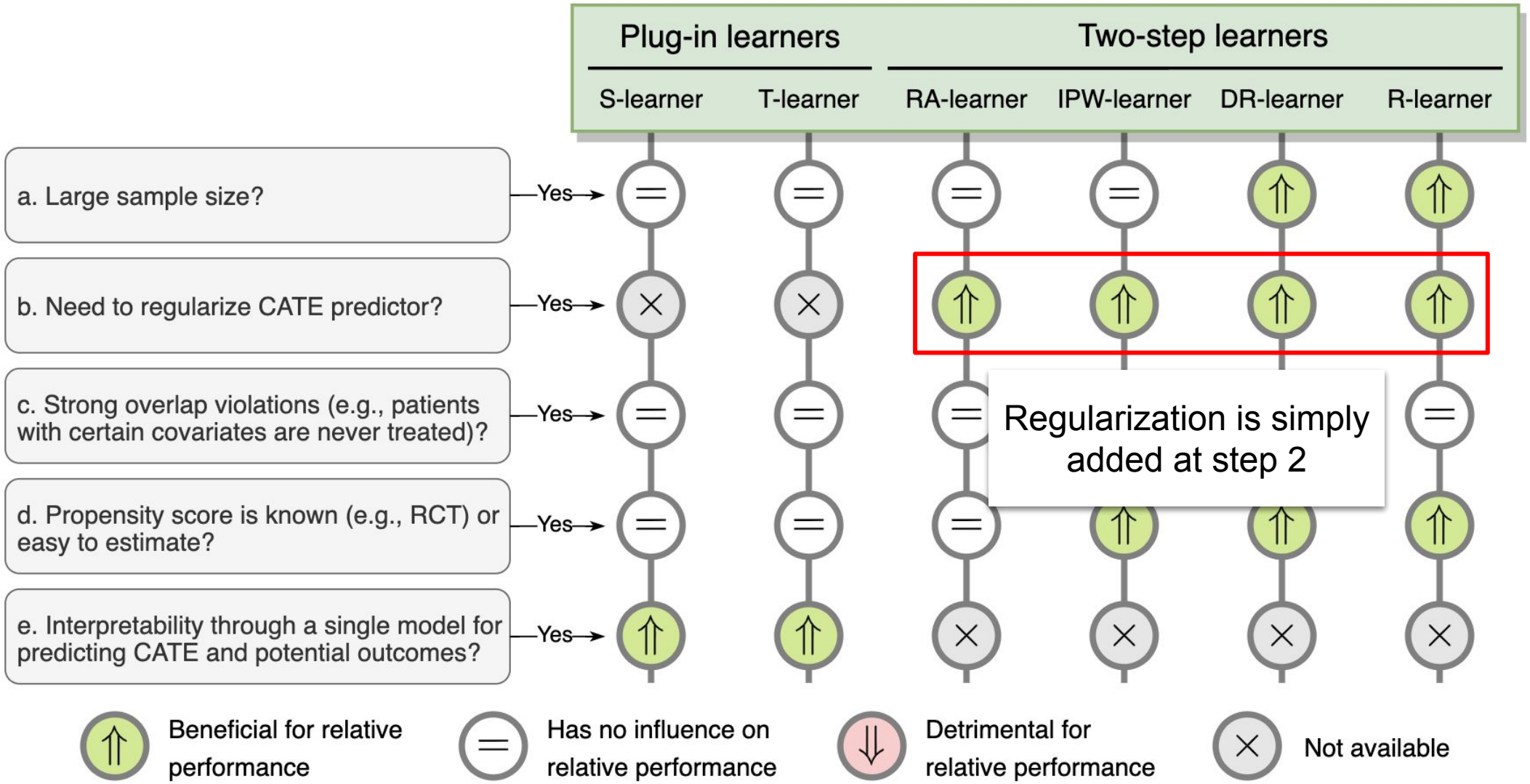
Feuerriegel, Stefan, et al. "Causal machine learning for predicting treatment outcomes." Nature Medicine 30.4 (2024): 958-968.

ML and estimation: Meta-learners summary for questions 1-3



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

ML and estimation: Meta-learners summary for questions 1-3



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

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

Best asymptotically does not mean best in low-sample!

"No Free Lunch" :(

Best approach in low-sample regime

ML and estimation: ℹ️ 4. Only heuristics are available

Best asymptotically does not mean best in low-sample!

"No Free Lunch" :(

Best approach in low-sample regime

+

- Now, we don't have **loss-based model selection criteria** (we always need estimated nuisance functions)
- Only heuristics are available (Rothenhäusler 2020, Curth & van der Schaar, 2023)

ML and estimation: ℹ️ 4. Only heuristics are available

Best asymptotically does not mean best in low-sample!

"No Free Lunch" :(

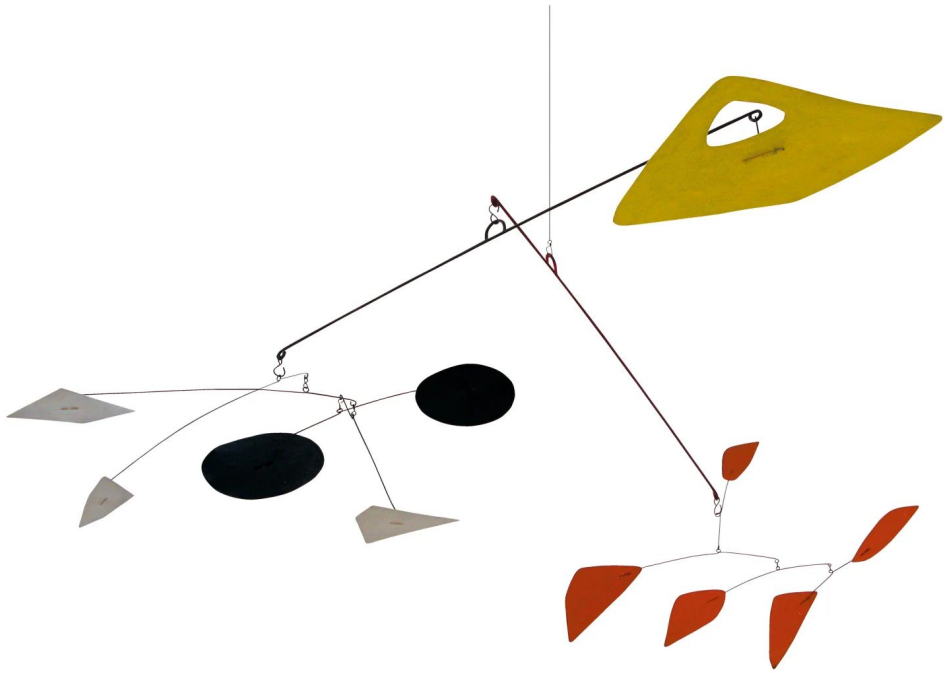
Best approach in low-sample regime

- **Possible solution**: employ RCTs (L2) data (with known propensity score)
- However, the ground-truth in this case has high variance

ML and estimation: 5. “How to address the covariate shift?”

Should we do something?

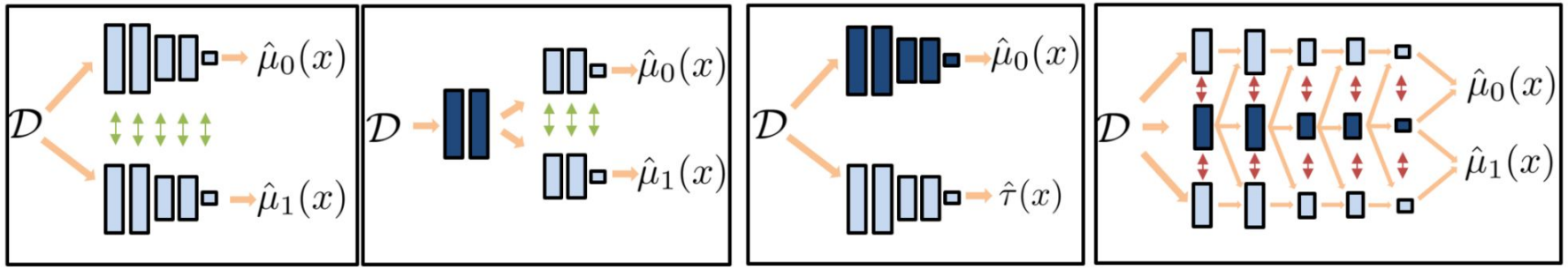
- Selection bias matters in low-sample regime, e.g. $\hat{\mu}_a(x)$ overfits on the factual data with high propensity
- Thus, single-stage plug-in learners are sub-optimal in a sense, that they don’t use all the information from data
- Debiased two-stage learners act like ‘regularizers’ on the first stage output, acting on the overfitted models
- But by using two-step learners, we introduce more parameters to estimate / need to choose hyperparameters
- Debiased learners may suffer from large inverse propensity scores (DR-learner)



Alexander Calder - Untitled

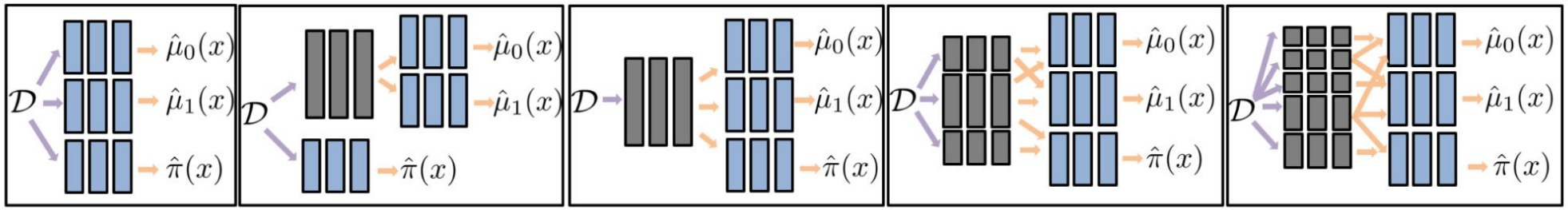
ML and estimation: 6. “Can we incorporate inductive biases for nuisance functions estimation?”

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



(1) Regularization for TNet (left) and TARNet (right) (2) Reparametrization (3) FlexTENet

Sharing representations for all the nuisance functions



TNet TARNet (SNet-1) DragonNet (SNet-2) DR-CFR (SNet-3) SNet

See (Curth & van der Schaar, 2021a; Curth & van der Schaar, 2021b)

ML and estimation: ✅ 6. Incorporating inductive biases

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, BNN, Causal Forest)
- T-learner (T-Net, TARNet, DragonNet, CFR & variants)
- Mixed approaches (FlexTENet)

Debiased learners*:
- EP-learner

Two-stage learners:

Plug-in learners:
- IPW-learner
- RA-learner / X-learner (GANITE)
- U-learner

Debiased learners:
- DR-learner
- R-learner (DML)

ATE estimation: estimating a parameter

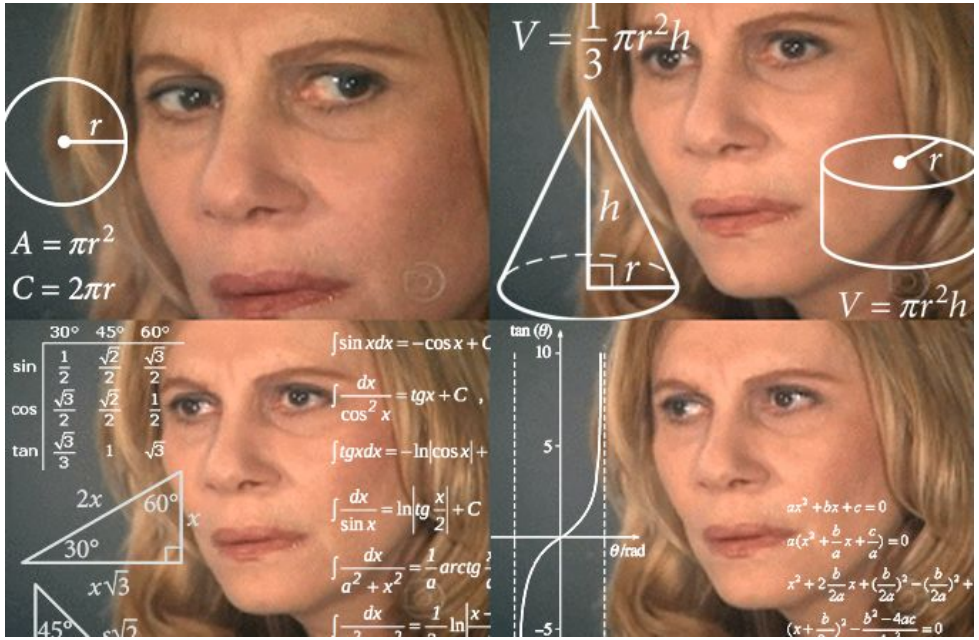
Plug-in estimators:
- IPTW estimator
- RA estimator

Debiased estimators:
- A-IPTW estimator
- TMLE estimator

ML and estimation: 6. “Can we incorporate inductive biases for nuisance functions estimation?”

We can design ML models, which incorporate inductive biases, but we cannot validate/select them in a data-driven way.

Dilemma of the model selection

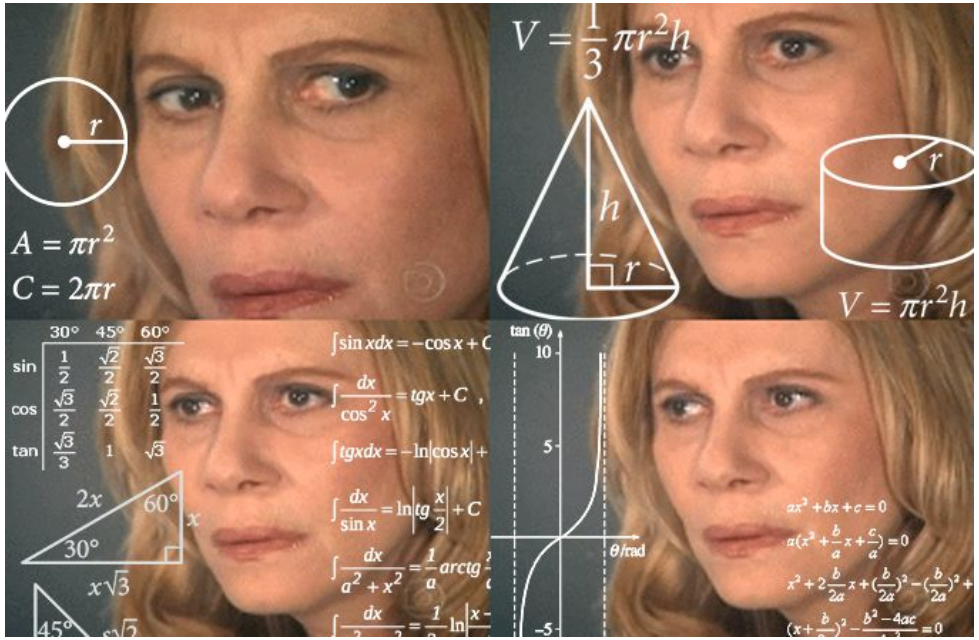


Is deep-learning even useful in this case?

ML and estimation: 6. “Can we incorporate inductive biases for nuisance functions estimation?”

We can design ML models, which incorporate inductive biases, but we cannot validate/select them in a data-driven way.

Dilemma of the model selection

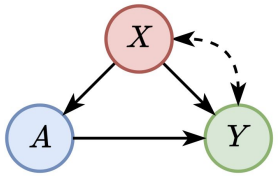


Is deep-learning even useful in this case? -> Sometimes!

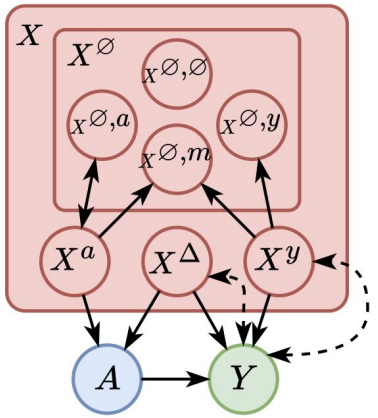
ML and estimation: 7. “How can representation learning help?”

- Is deep-learning even useful in this case? -> Yes, if ground-truth confounders lie on low-dimensional manifold (= additional assumption)
- Holy grail: **prognostic score**, namely minimal sufficient information in covariates for CATE estimation
- For identifying prognostic score, we would need to know the structure inside of X, namely, what are the ground-truth confounders, instruments, and noise:

Prognostic scores



Original causal diagram

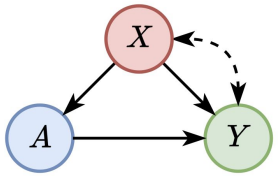


- But to do that perfectly, we have to learn an original full CATE (which makes the prognostic score obsolete)

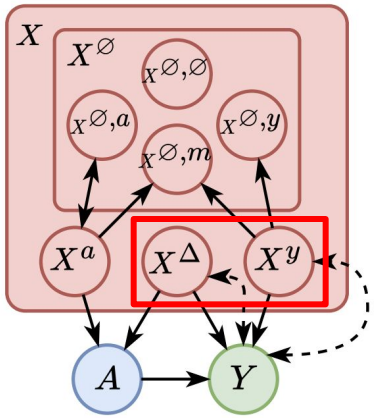
ML and estimation: 7. “How can representation learning help?”

- Is deep-learning even useful in this case? -> Yes, if ground-truth confounders lie on low-dimensional manifold (= additional assumption)
- Holy grail: **prognostic score**, namely minimal sufficient information in covariates for CATE estimation
- For identifying prognostic score, we would need to know the structure inside of X, namely, what are the ground-truth confounders, instruments, and noise:

Prognostic scores



Original causal diagram



- But to do that perfectly, we have to learn an original full CATE (which makes the prognostic score obsolete)

ML and estimation: ✅ 7. Representation learning for CATE estimation

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, BNN, Causal Forest)
- T-learner (T-Net, TARNet, DragonNet, CFR & variants)
- Mixed approaches (FlexTENet)

Debiased learners*:
- EP-learner

Two-stage learners:

Plug-in learners:
- IPW-learner
- RA-learner / X-learner (GANITE)
- U-learner

Debiased learners:
- DR-learner
- R-learner (DML)

ATE estimation: estimating a parameter

Plug-in estimators:
- IPTW estimator
- RA estimator

Debiased estimators:
- A-IPTW estimator
- TMLE estimator

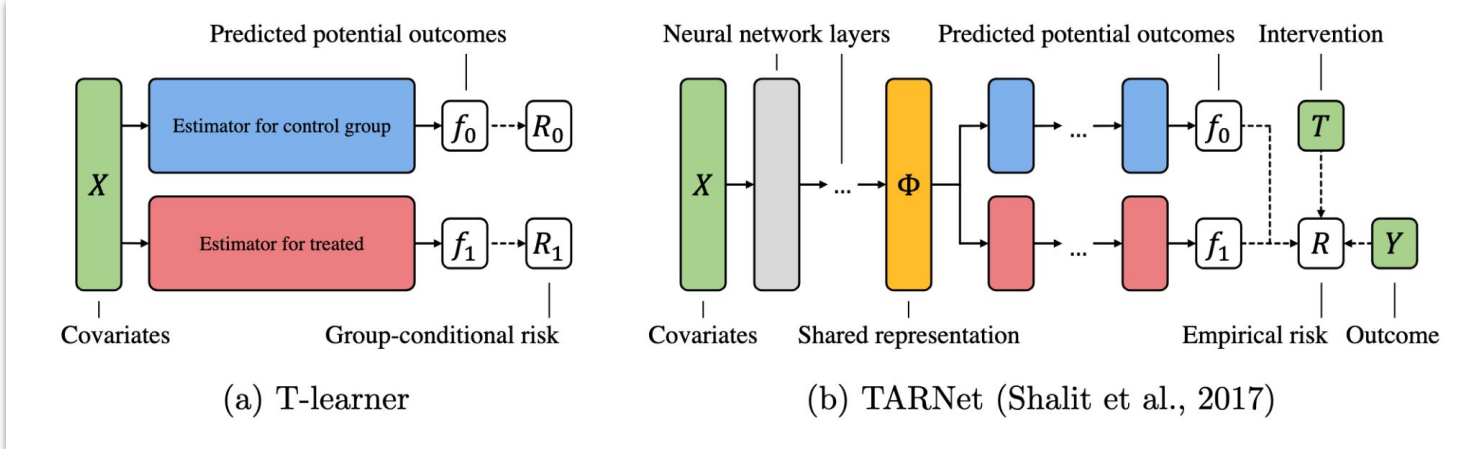
ML and estimation: ✅ 7. Representation learning for CATE estimation

Representation learning for CATE estimation

- **Idea**: employ representation learning to map the covariates to a lower-dimensional space and reduce variance of CATE estimation:

$$\Phi(\cdot) : X \to \Phi(X)$$

- Most common implementation, neural-network based approach (e.g., TARNet):

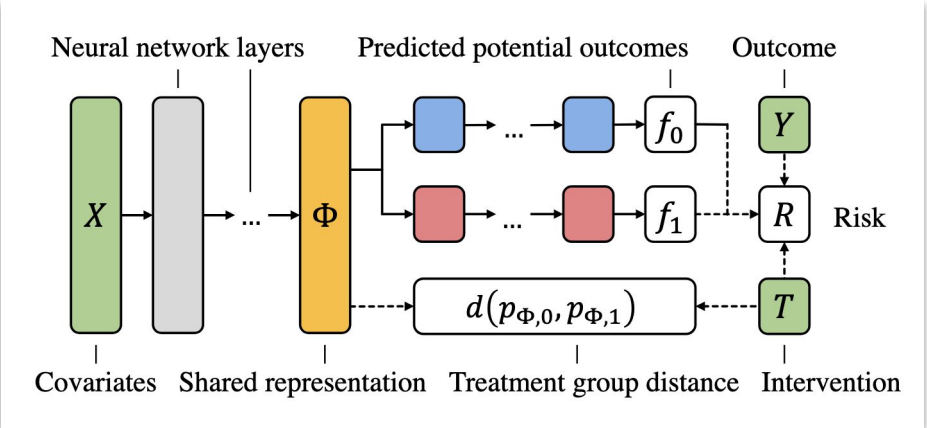


Johansson, Fredrik D., et al. "Generalization bounds and representation learning for estimation of potential outcomes and causal effects." Journal of Machine Learning Research 23.166 (2022): 1-50.

ML and estimation: ✅ 7. Representation learning for CATE estimation

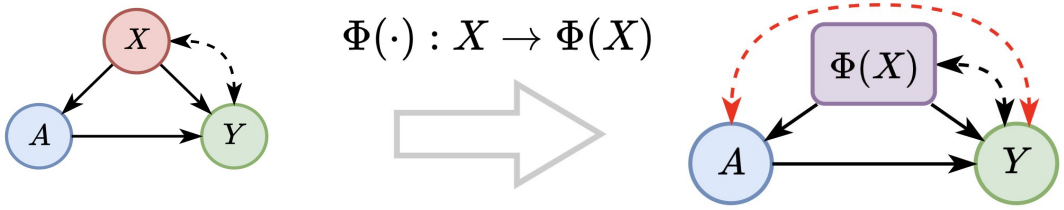
Representation learning for CATE estimation

- Some methods suggested balancing the representation wrt. treatment indicator (e.g., BNN, CFR) to further decrease the variance.



Johansson, Fredrik D., et al. "Generalization bounds and representation learning for estimation of potential outcomes and causal effects." Journal of Machine Learning Research 23.166 (2022): 1-50.

- However, this may lead to the **representation-induced confounding bias** (Melnychuk et. al., 2024): so balancing is not recommended



- Still, learned representations can generally improve Neyman-orthogonal learners and vice-versa (Melnychuk et. al., 2026)

Extensions

Extensions: Ongoing / open challenges

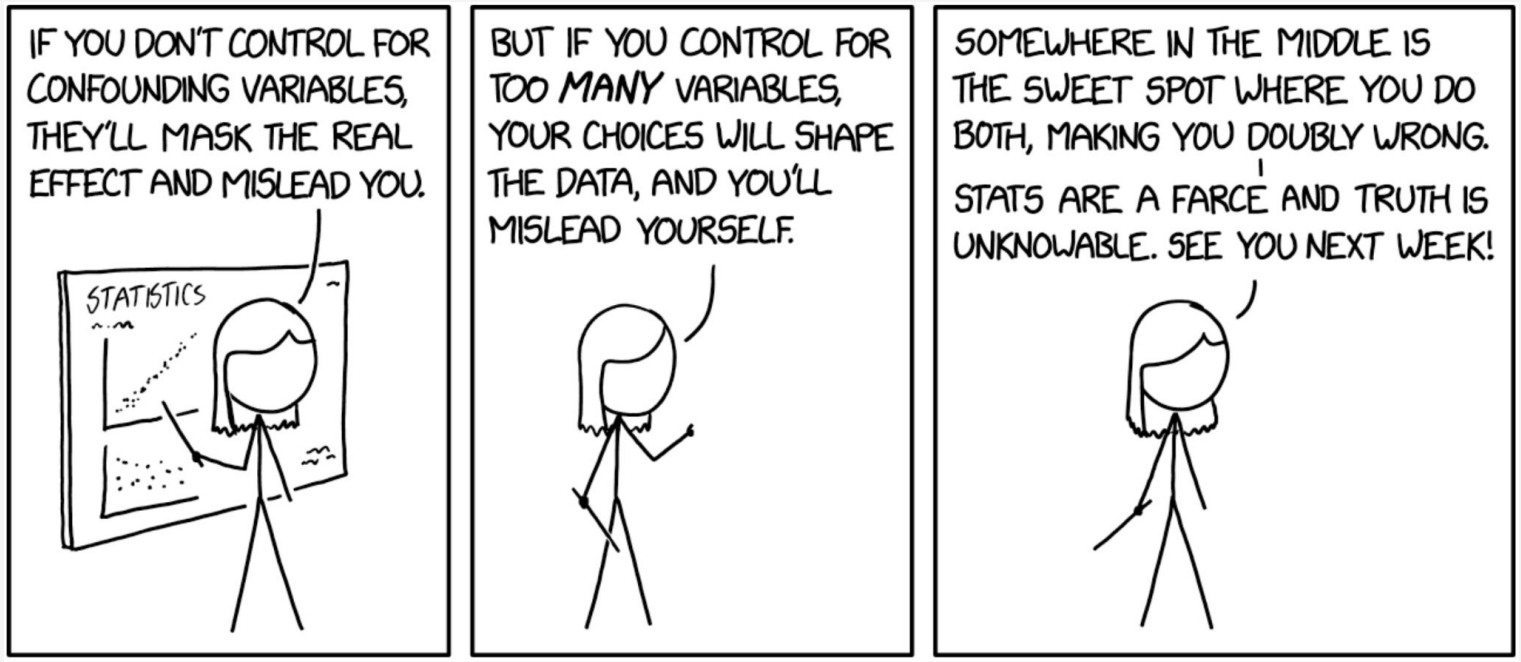
Uncertainty of TEs / POs	• Epistemic (estimation) uncertainty for CATE / CAPO • Aleatoric uncertainty for POs (Melnychuk et al. 2023), TEs (Melnychuk et al., 2024) • Total uncertainty for CATE and CAPOs with conformal prediction • Bayesian causal inference (Melnychuk et al. 2026, Javurek et al. 2026)
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	• Explainability / interpretability for causal inference (ongoing work)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

HUMAN TECHNOPOLE



https://xkcd.com/2560/

🙏 Thank you for your attention!

Main message: causal ML is very different from regular ML

My web-page

Causal ML Lab page