LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN LMU MUNICH SCHOOL OF MANAGEMENT INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

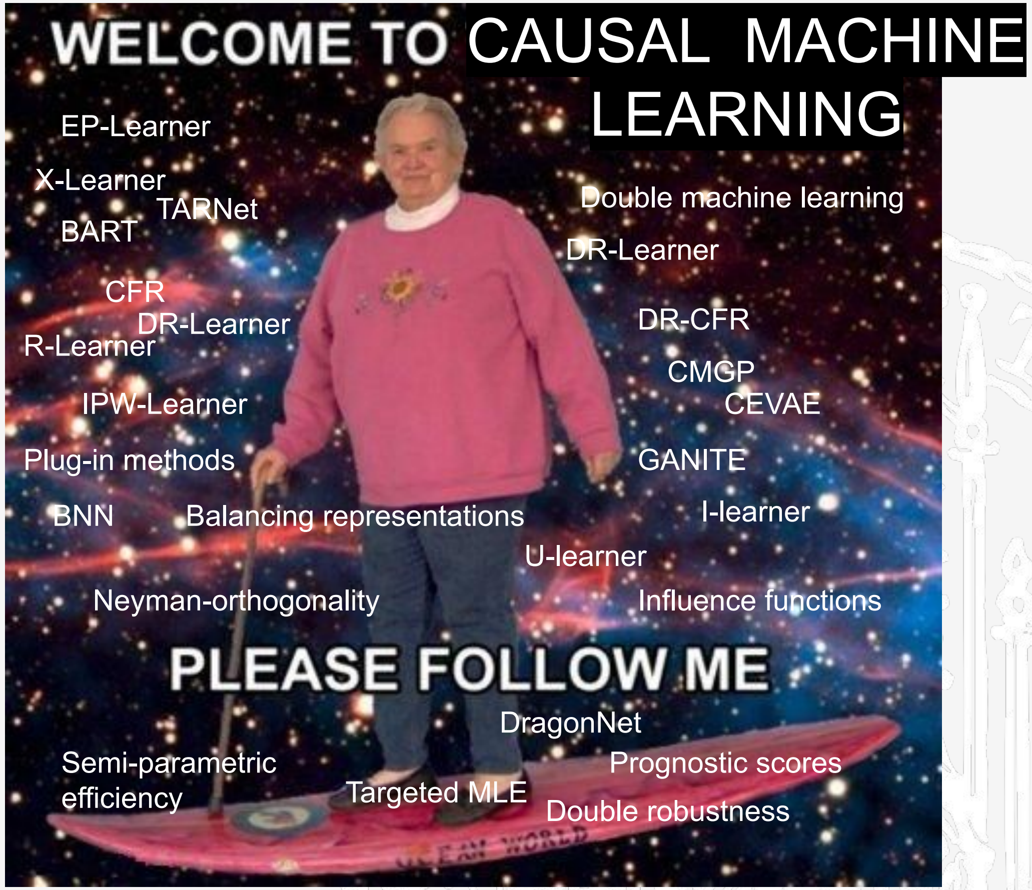
Munich Center for Machine Learning

Tutorial: Causal ML for treatment effect estimation

Valentyn Melnychuk

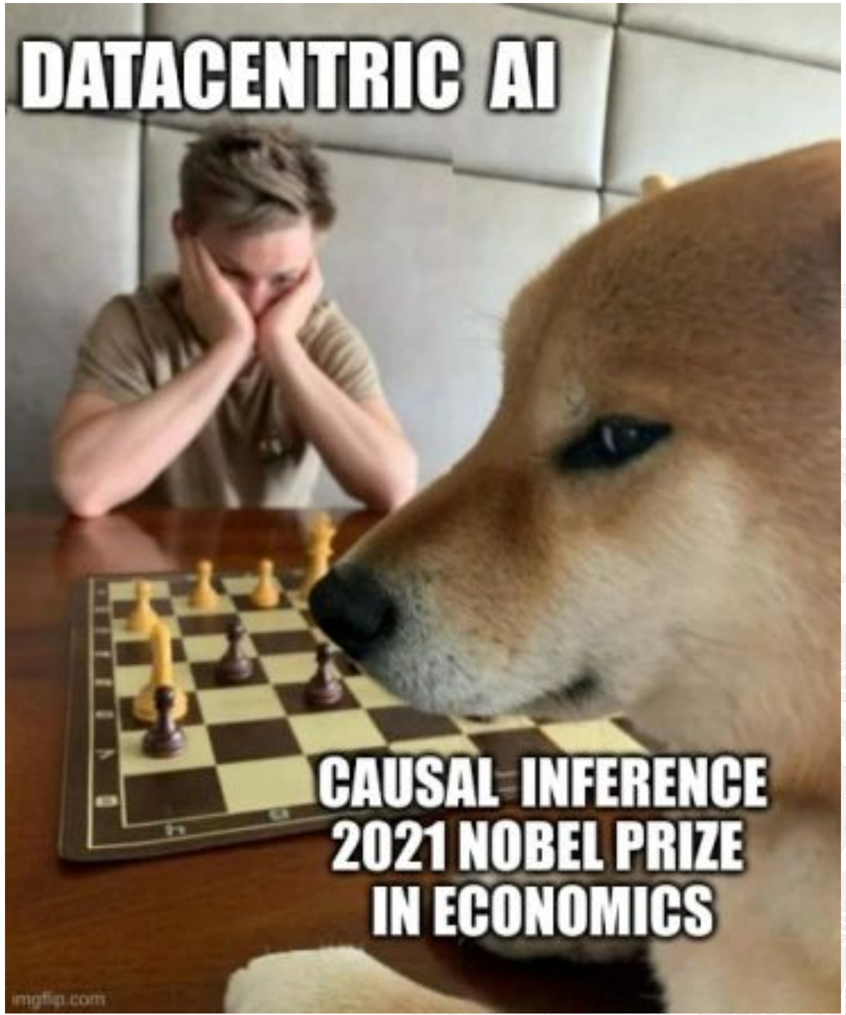
3rd Munich Workshop on Causal Machine Learning

Institute of AI in Management, LMU Munich



Introduction

- Causal Machine Learning
- Treatment effect estimation from observational data
- 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

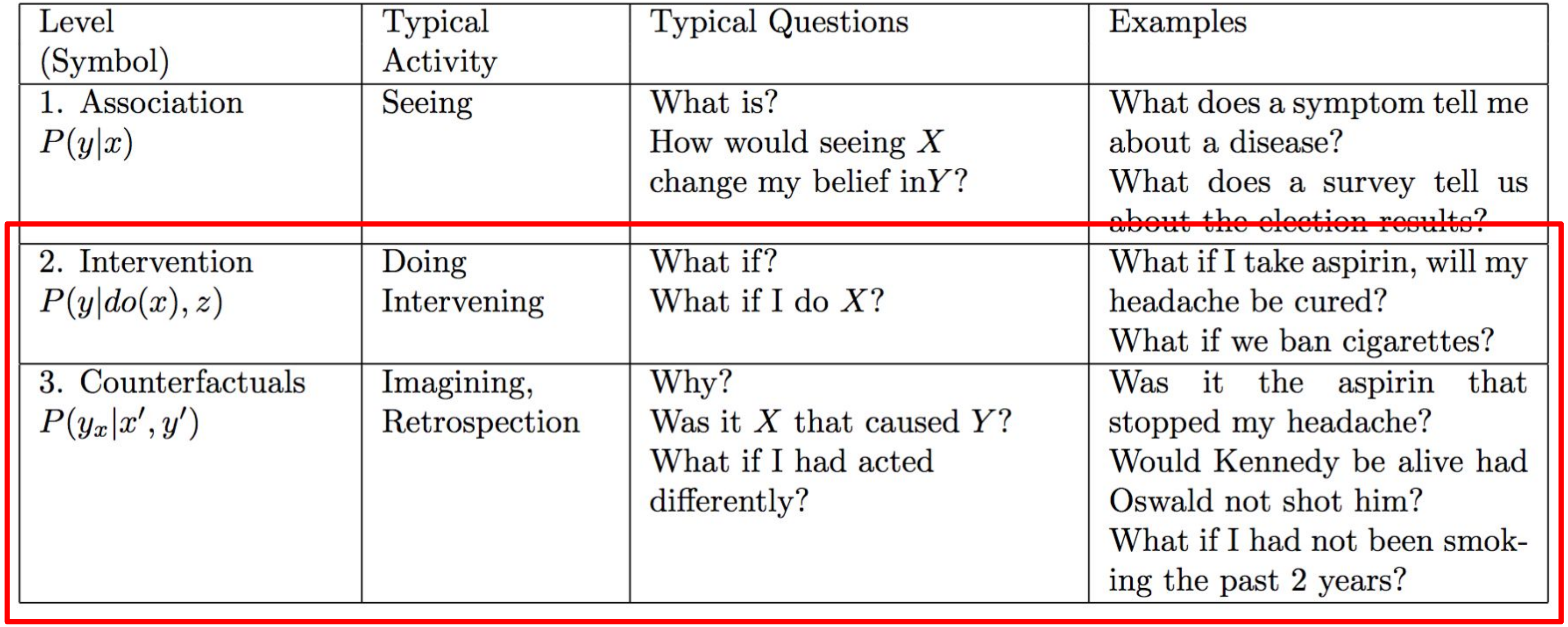
- Treatment effect estimation
- Counterfactual inference
- Causal discovery
- …

←

ML / DL tools

Introduction: Treatment effect estimation from observational data

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



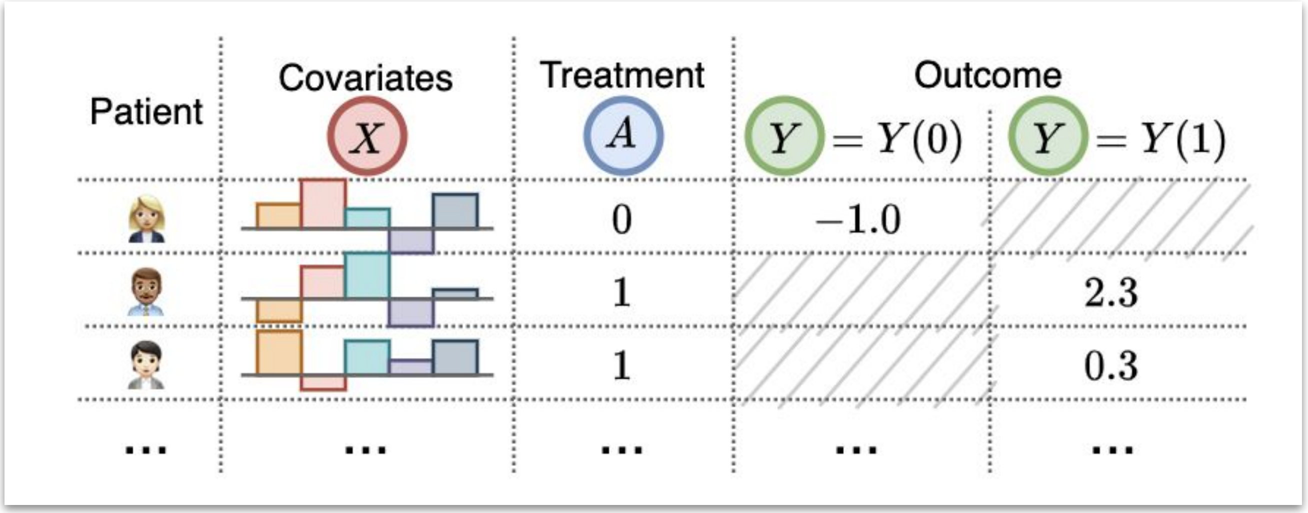
Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? 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?

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

Introduction: Problem formulation

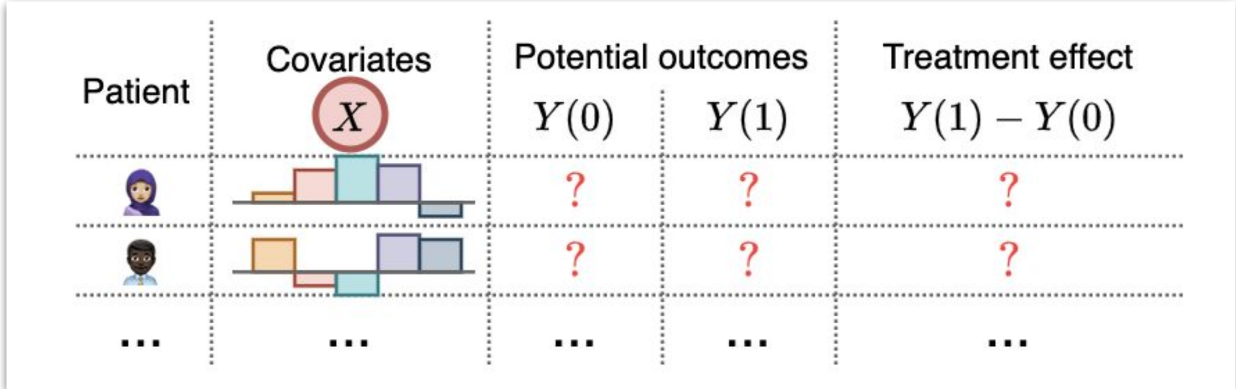
- Given i.i.d. observational dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^{n} \sim \mathbb{P}(X, A, Y)$

X covariates
A (binary) treatments
Y continuous (factual) outcomes



Patient	Covariates X	Treatment A	Outcome $Y = Y(0)$	Outcome $Y = Y(1)$
👩🏼‍💼		0	-1.0	▨
👨🏽‍💼		1	▨	2.3
👨🏻‍💼		1	▨	0.3
...	...	...	...	...

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

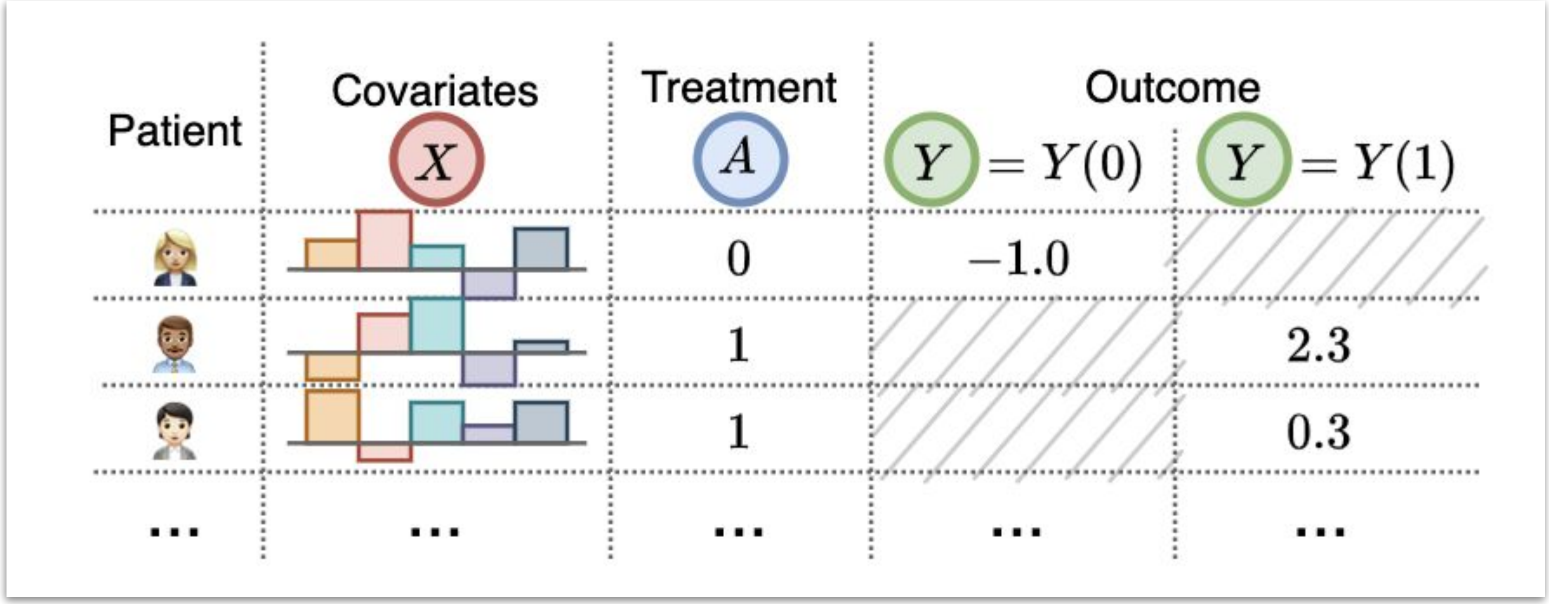


Patient	Covariates X	Potential outcomes $Y(0)$	Potential outcomes $Y(1)$	Treatment effect $Y(1) - Y(0)$
🧕🏽		?	?	?
👨🏿‍💼		?	?	?
...	...	...	...	...

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 -> **selection bias**

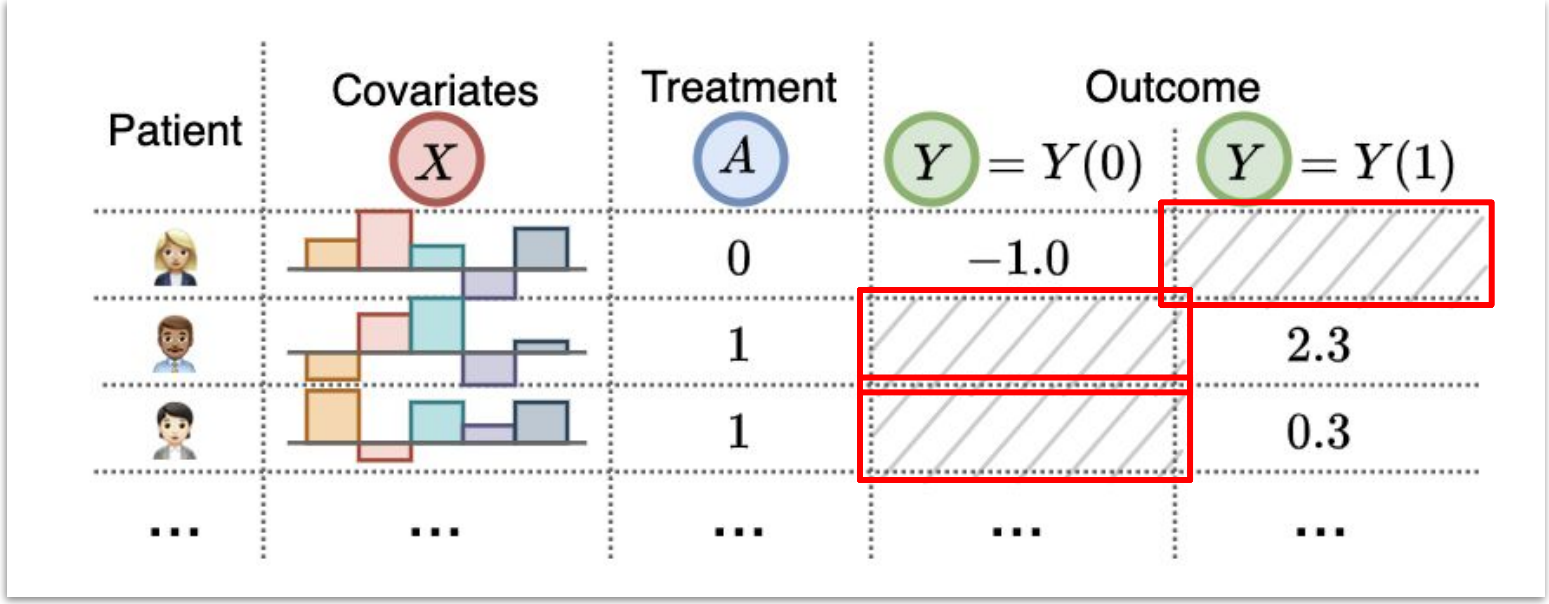


Patient	Covariates	Treatment	Outcome	
	X	A	$Y = Y(0)$	$Y = Y(1)$
👩🏼‍💼		0	-1.0	▨
👨🏽‍💼		1	▨	2.3
🧑🏻‍💼		1	▨	0.3
...	...	...	...	...

Introduction: Fundamental problem of causal inference

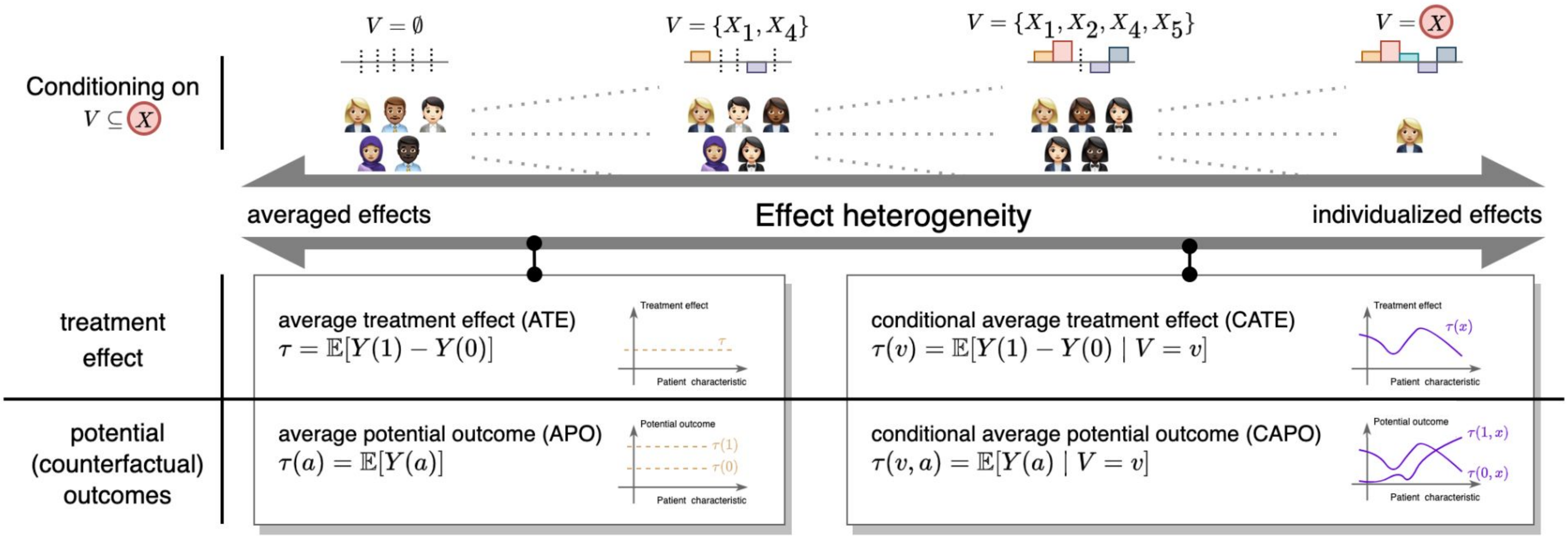
- **Both** potential outcomes (factual and counterfactual) are never observed for any individual -> treatment effects are never observed

- Potential outcomes are only observed for parts of the population -> **selection bias**



Patient	Covariates	Treatment	Outcome	
	X	A	$Y = Y(0)$	$Y = Y(1)$
👩‍💼		0	-1.0	▨
👨🏾‍💼		1	▨	2.3
👨🏻‍💼		1	▨	0.3
...	...	...	...	...

Introduction: Spectrum of causal estimands



Conditioning on $V \subseteq X$

$V = \emptyset$

$V = \{X_1, X_4\}$

$V = \{X_1, X_2, X_4, X_5\}$

$V = X$

averaged effects

Effect heterogeneity

individualized effects

treatment effect

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

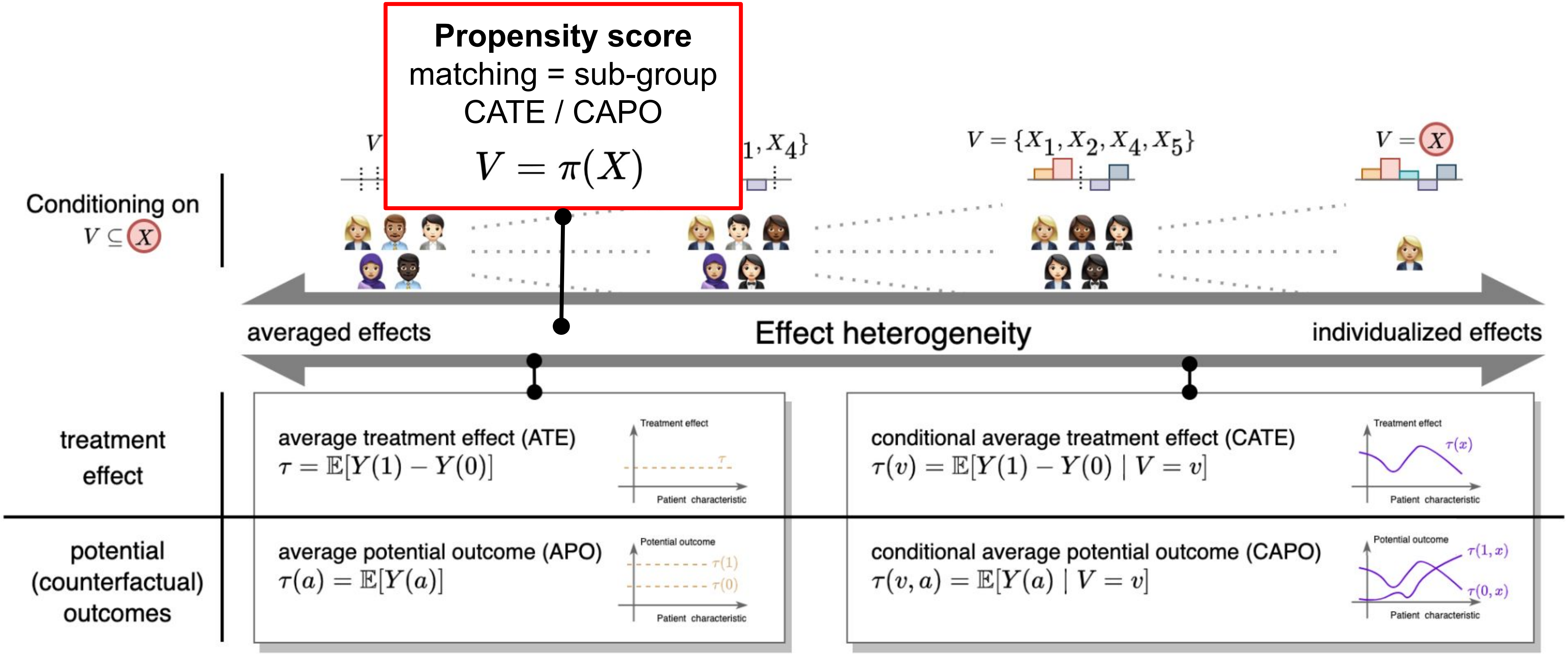
conditional average treatment effect (CATE)
$\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$

potential (counterfactual) outcomes

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

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

Introduction: Spectrum of causal estimands



Propensity score matching = sub-group CATE / CAPO

$$V = \pi(X)$$

Conditioning on $V \subseteq X$

averaged effects

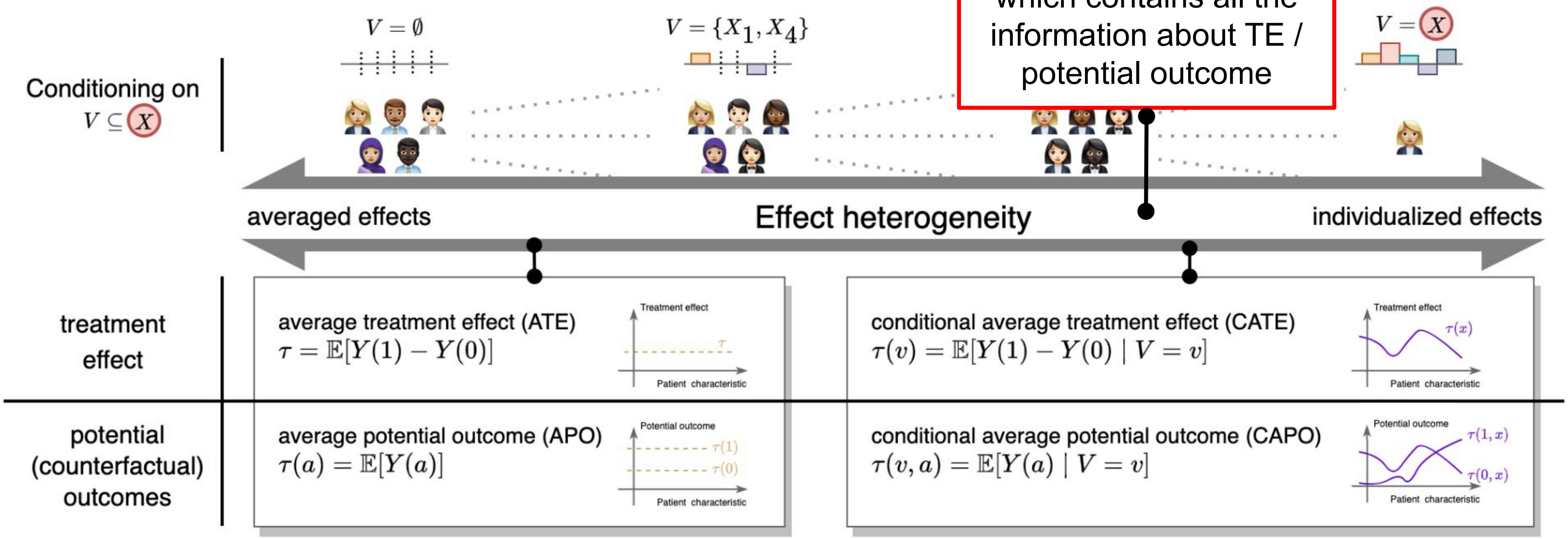
Effect heterogeneity

individualized effects

treatment effect	average treatment effect (ATE) $\tau = \mathbb{E}[Y(1) - Y(0)]$	conditional average treatment effect (CATE) $\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$
potential (counterfactual) outcomes	average potential outcome (APO) $\tau(a) = \mathbb{E}[Y(a)]$	conditional average potential outcome (CAPO) $\tau(v, a) = \mathbb{E}[Y(a) \mid V = v]$

Introduction: Spectrum of causal estimands

Prognostic score = minimal conditioning set, which contains all the information about TE / potential outcome



Conditioning on $V \subseteq X$

$V = \emptyset$

$V = \{X_1, X_4\}$

$V = X$

averaged effects

Effect heterogeneity

individualized effects

treatment effect

potential (counterfactual) outcomes

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

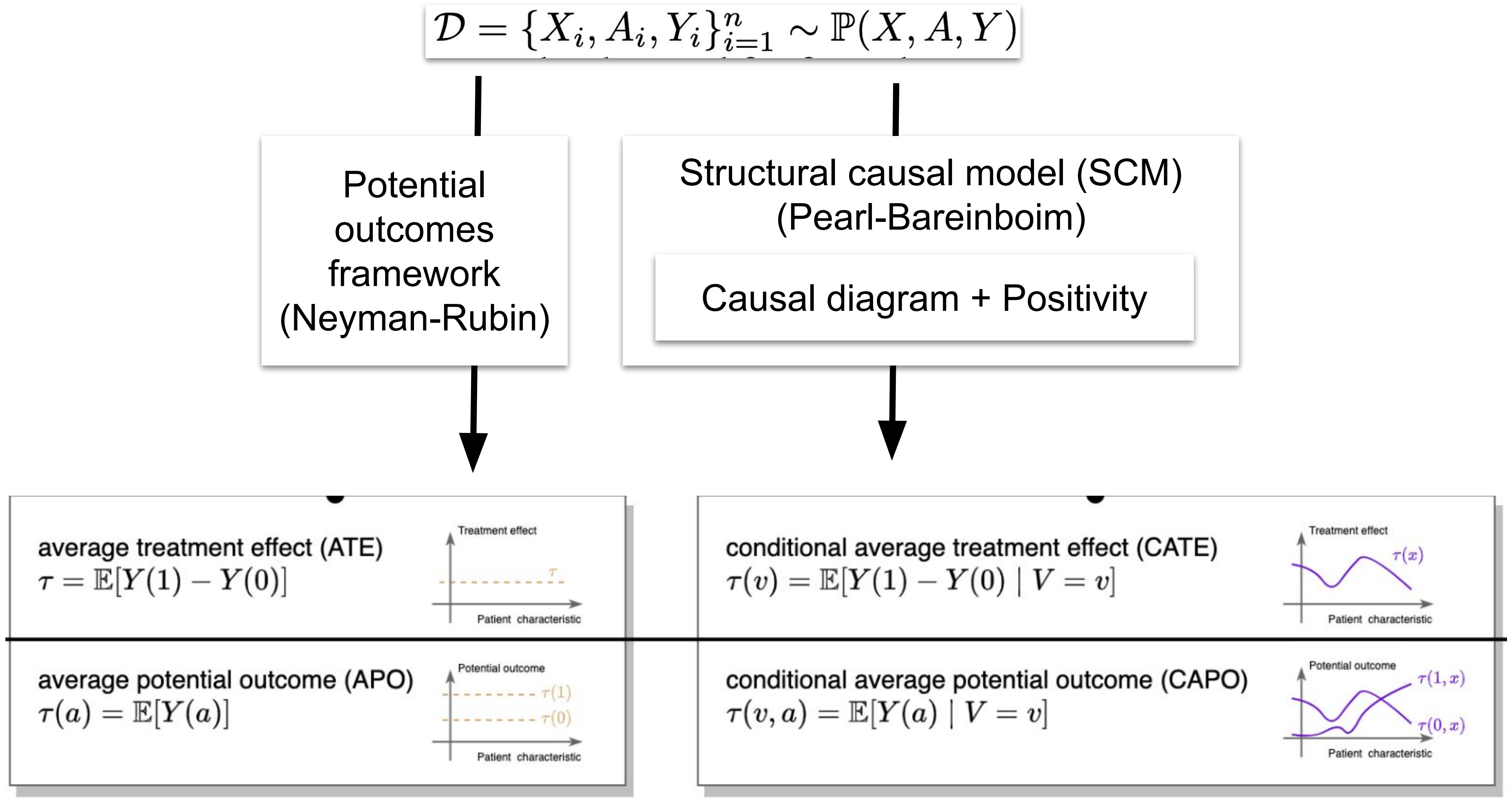
This keeps happening. How heavy are cats?

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



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

Potential outcomes framework (Neyman-Rubin)

Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

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

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

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

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

Causal assumptions: Frameworks

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

=

More general

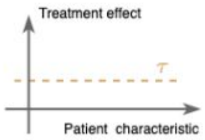
(i) Consistency
(ii) Positivity (Overlap)
(iii) Exchangeability
(Ignorability)

Potential outcomes framework (Neyman-Rubin)

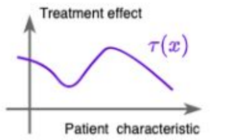
Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

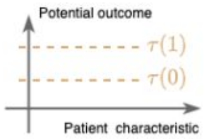
average treatment effect (ATE)
$\tau = \mathbb{E}[Y(1) - Y(0)]$



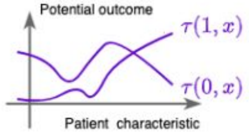
conditional average treatment effect (CATE)
$\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$



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: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$	✅ (but curse of dimensionality kicks in)
(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: $A \perp\!\!\!\perp Y[a] \mid X \text{ for all } a.$	❌ (but we can speculate about plausibility with sensitivity models)

Causal assumptions: Potential outcomes framework (Neyman-Rubin)

Identifiability with potential outcomes framework

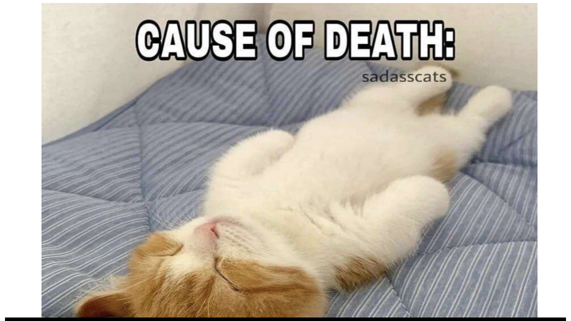
Given Assumptions (i) - (iii), **causal quantities** are identifiable from observational data via

- back-door (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 weighting (IPW):
  - ○ 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)
- Due to the curse of dimensionality problem becomes harder to estimate
- When adjusting for a post-treatment covariate, we induce bias -> **kitty dies**



Post-treatment covariate adjustment

Causal assumptions: Frameworks

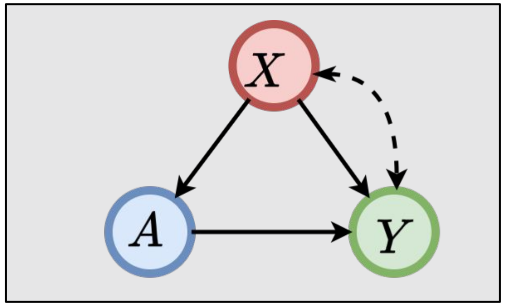
$$\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n \sim \mathbb{P}(X, A, Y)$$

Potential outcomes framework (Neyman-Rubin)

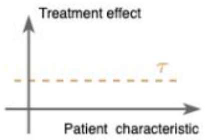
Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

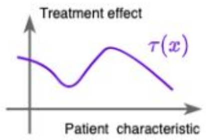
Assumptions can be related to the structural knowledge



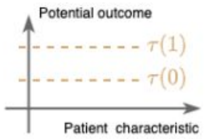
average treatment effect (ATE)
$\tau = \mathbb{E}[Y(1) - Y(0)]$



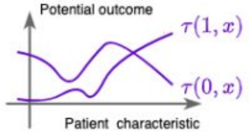
conditional average treatment effect (CATE)
$\tau(v) = \mathbb{E}[Y(1) - Y(0) \mid V = v]$



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



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



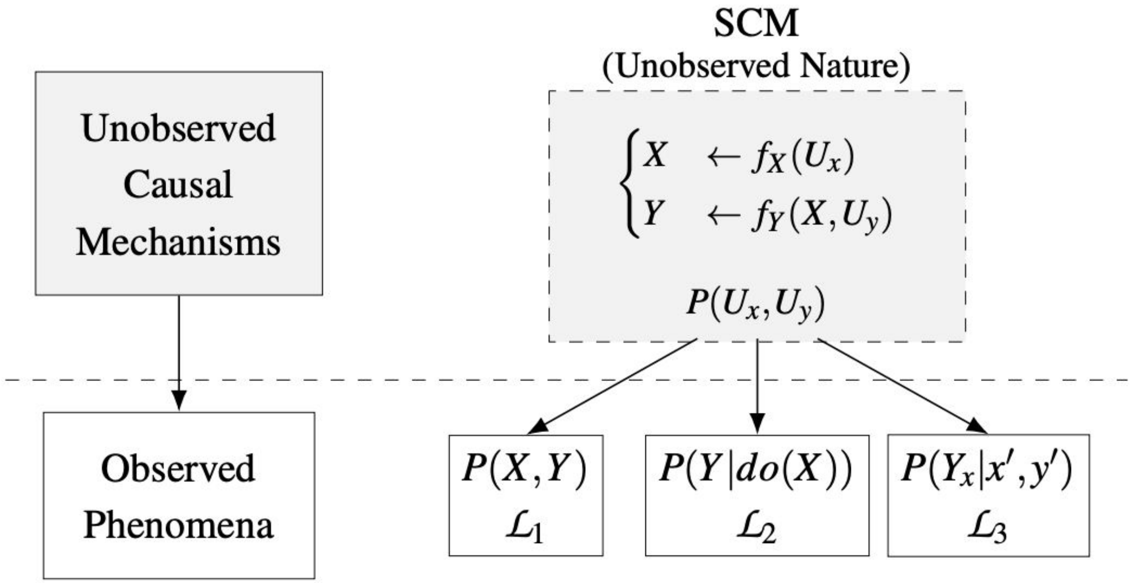
Causal assumptions: Structural causal model (SCM)

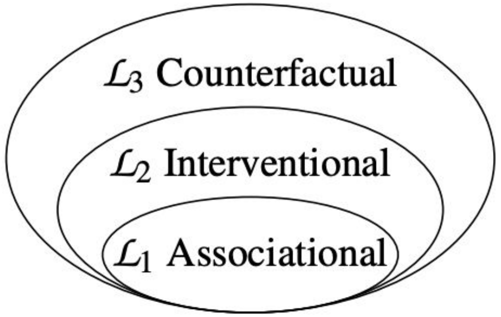
- **Informal**: Assuming a SCM = knowing the full nature of the data generating process
- SCM = {observed variables, hidden variables, functional assignments for every observed covariate, probability distribution for hidden variables}

Verifiable with infinite observational data?

❌

SCM


SCM
(Unobserved Nature)
Unobserved Causal Mechanisms
$\begin{cases} X \leftarrow f_X(U_x) \\ Y \leftarrow f_Y(X,U_y) \end{cases}$
$P(U_x,U_y)$
Observed Phenomena
$P(X,Y)$ $\mathcal{L}_1$
$P(Y|do(X))$ $\mathcal{L}_2$
$P(Y_x|x',y')$ $\mathcal{L}_3$

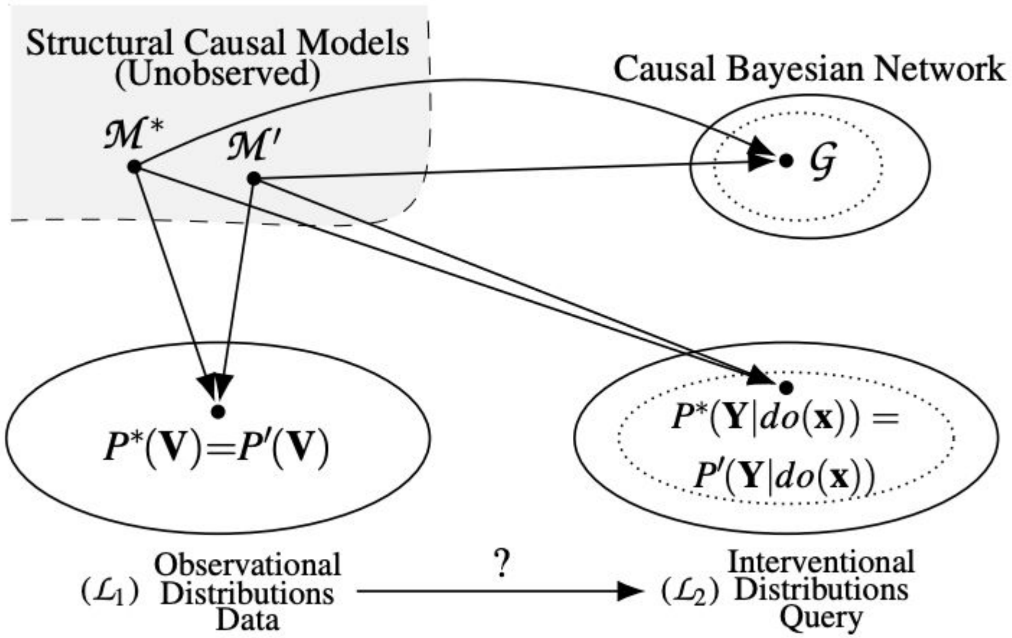

$\mathcal{L}_3$ Counterfactual
$\mathcal{L}_2$ Interventional
$\mathcal{L}_1$ Associational

- 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.

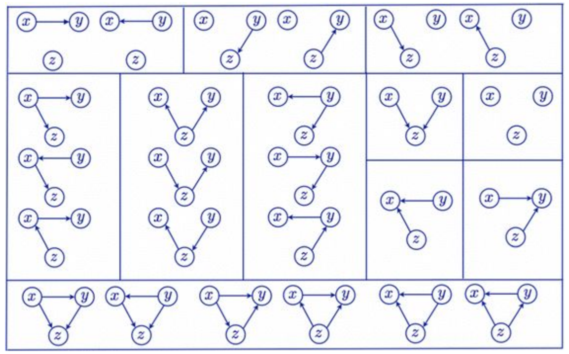
Causal diagram


Structural Causal Models
(Unobserved)
$\mathcal{M}^*$
$\mathcal{M}'$
Causal Bayesian Network
$\mathcal{G}$
$P^*(\mathbf{V})=P'(\mathbf{V})$
$P^*(\mathbf{Y}|do(\mathbf{x})) =$
$P'(\mathbf{Y}|do(\mathbf{x}))$
$(\mathcal{L}_1)$ Observational Distributions Data
?
$(\mathcal{L}_2)$ Interventional Distributions Query

Verifiable with infinite observational data?

❌

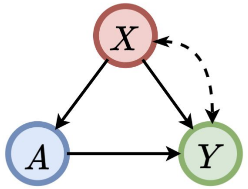
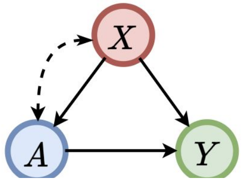
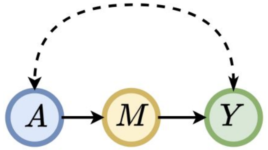
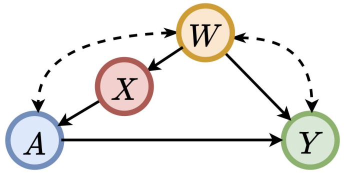
(only Markov equivalence class is identifiable, for Markovian diagrams)



Causal assumptions: Causal diagram

Identifiability with causal diagrams

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

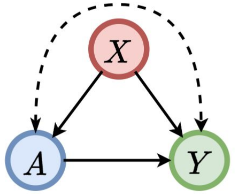
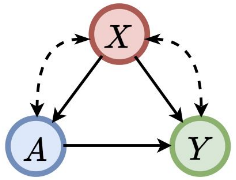
Query:	Causal diagram:	ID:	Formula:
CATE / CAPO		✅	- back-door adjustment - propensity reweighting
CATE / CAPO		✅	- back-door adjustment - propensity reweighting
ATE / APO		✅	- front-door adjustment
ATE / APO		✅	- napkin formula

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

Causal assumptions: Causal diagram

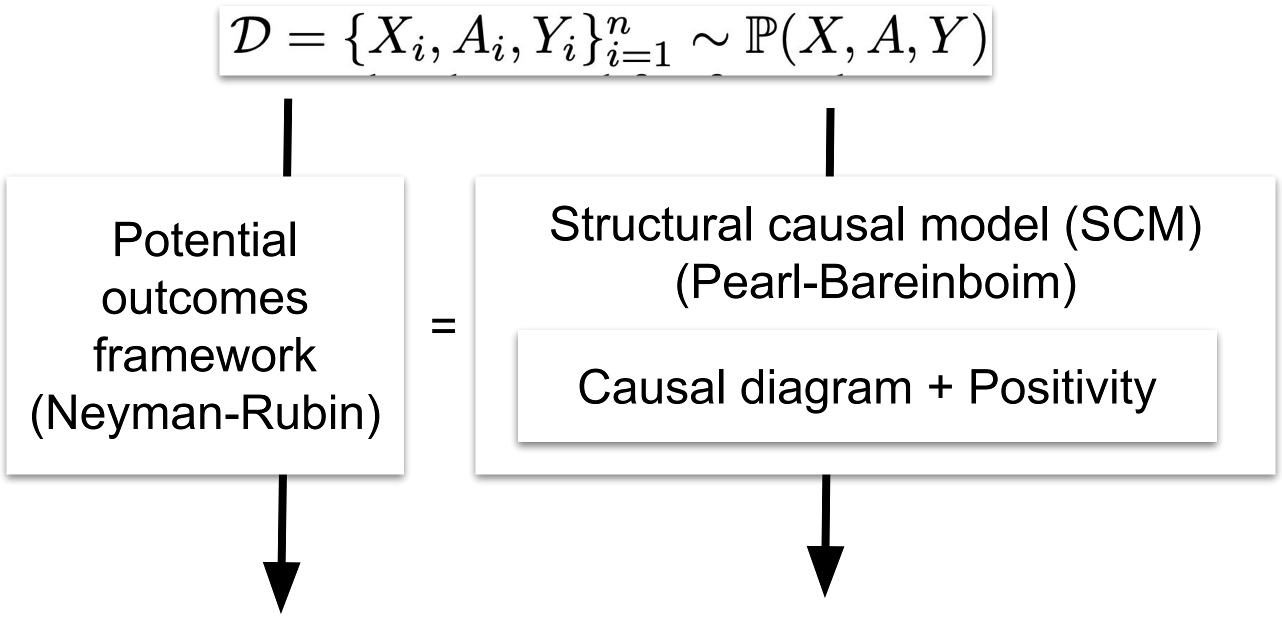
- Sound and complete **identifiability algorithms** (using do-calculus) exist for L2 and L3 causal quantities, e.g.,

Identifiability with causal diagrams

Query:	Causal diagram:		ID:	Formula:
CATE / CAPO		→	❌	(Hidden Confounding)
CATE / CAPO		→	❌	(Butterfly-bias)

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

Causal assumptions: Frameworks



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

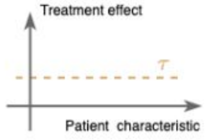
Potential outcomes framework (Neyman-Rubin)

=

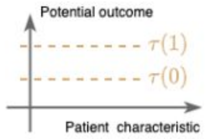
Structural causal model (SCM) (Pearl-Bareinboim)

Causal diagram + Positivity

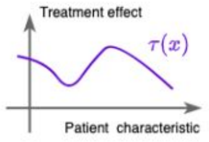
average treatment effect (ATE)
$\tau = \mathbb{E}[Y(1) - Y(0)]$



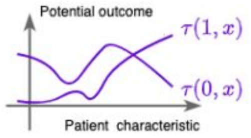
average potential outcome (APO)
$\tau(a) = \mathbb{E}[Y(a)]$



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



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



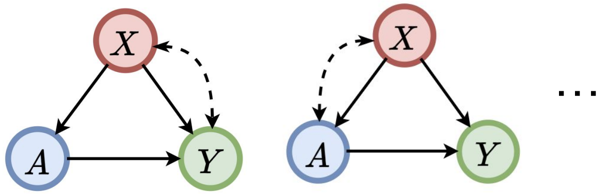
Causal assumptions: Equivalence of the frameworks

Equivalence of assumptions

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

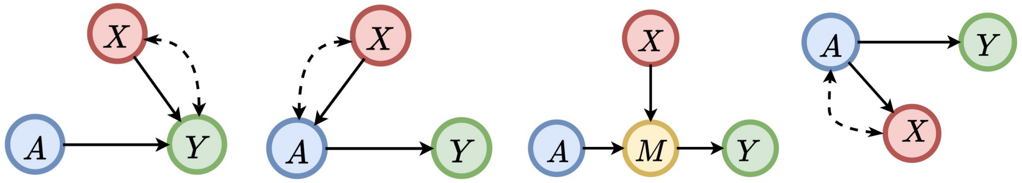
(i) Causal diagrams, where:

- back-door adjustment for X should be applied



(i) Consistency
(iii) Ignorability

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



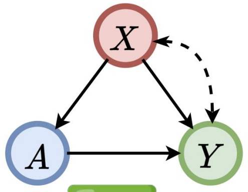
(ii) Positivity

(ii) Positivity

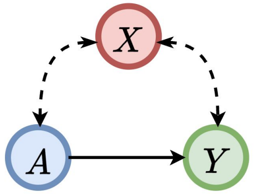
Causal assumptions: Equivalence of the frameworks

Choosing covariates (revisited)

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

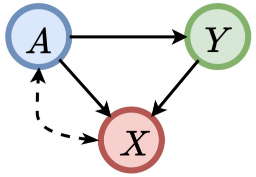


✅

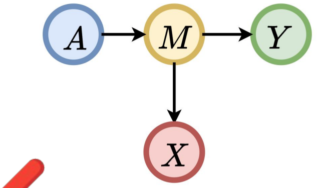


❌ (M-bias)

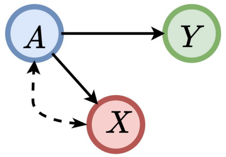
- Most of the post-treatment covariate adjustments lead to the **death of a kitty**



❌ (selection bias)

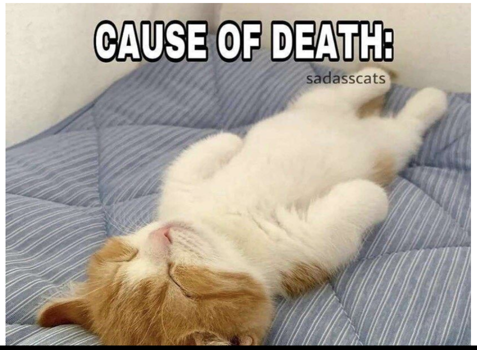


❌ (overcontrol bias)



✅

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



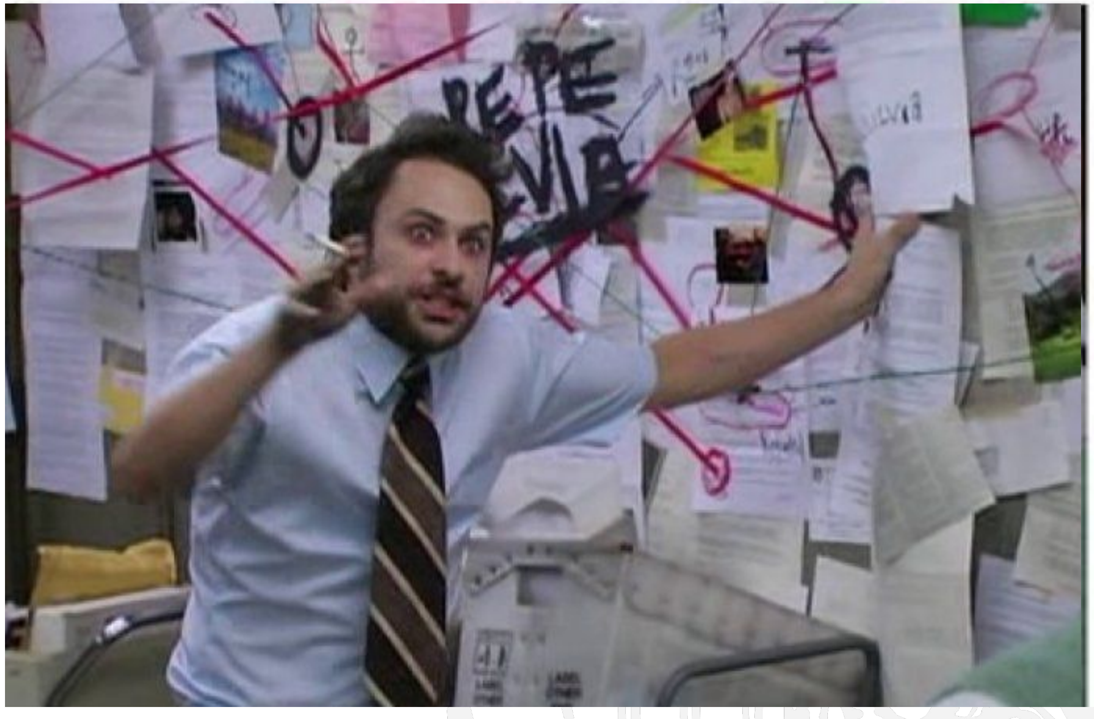
(Most of the) post-treatment covariate adjustments or M-bias

ML and estimation

- Big picture
- Plug-in (one-step) learners
- Issues of plug-in estimation
- 1. “What about the sub-group treatment effects?”
  - Pseudo-outcomes vs custom residualized loss
  - Two-step learners
  - Plug-in (one-step) vs two-step learners
- 2. How to regularize tau(x)?
- 3. “What is better, adjustment or IPW?”
- 4. “Can we do data-driven model selection?”
- 5. “How to address the selection bias?”
- 6. “Can we incorporate inductive biases for nuisance functions estimation?”
- 7. “Can we do end-to-end learning?”

Nobody:

Me explaining all the causal inference methods:



ML and estimation: Big picture

CATE estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:

Pseudo-outcome regression:

- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:

- R-learner (DML)
- U-learner
- EP-learner
- …

Plug-in (one-step) learners:

- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:

- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- Causal Forest

Two-step models:

- GANITE

ATE / APO estimation: estimating a parameter

Sample averaging of pseudo-outcomes:

- IPW estimator
- RA estimator
- A-IPW estimator

Loss-based (TMLE):

- DragonNet

ML and estimation: Big picture

CAPO estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:

Pseudo-outcome regression:

- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:

- IPW-learner
- DR-learner
- i-learner
- …

Plug-in (one-step) learners:

- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:

- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- Causal Forest

Two-step models:

- GANITE

ATE / APO estimation: estimating a parameter

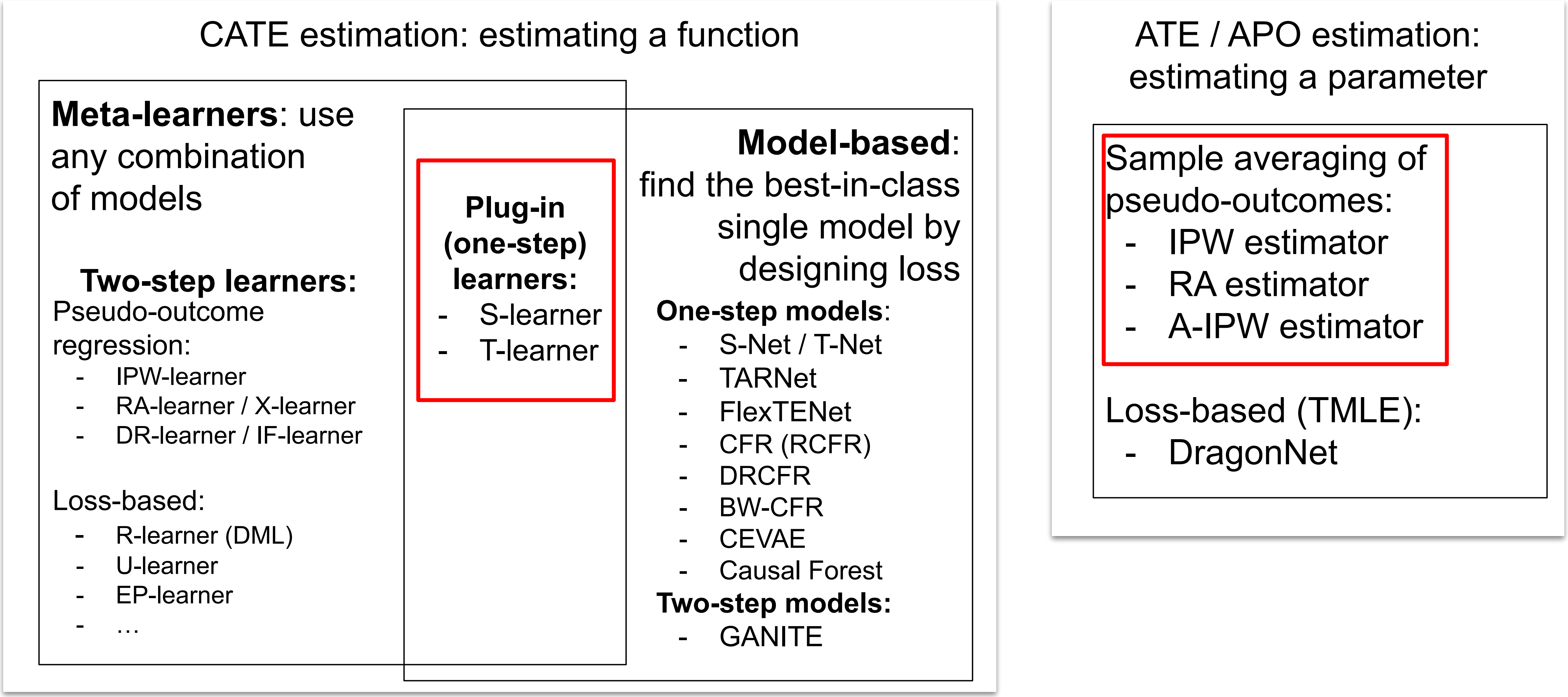
Sample averaging of pseudo-outcomes:

- IPW estimator
- RA estimator
- A-IPW estimator

Loss-based (TMLE):

- DragonNet

ML and estimation: One-step learners



CATE estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:
Pseudo-outcome regression:

- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:

- R-learner (DML)
- U-learner
- EP-learner
- …

Plug-in (one-step) learners:

- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:

- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- CEVAE
- Causal Forest

Two-step models:

- GANITE

ATE / APO estimation: estimating a parameter

Sample averaging of pseudo-outcomes:

- IPW estimator
- RA estimator
- A-IPW estimator

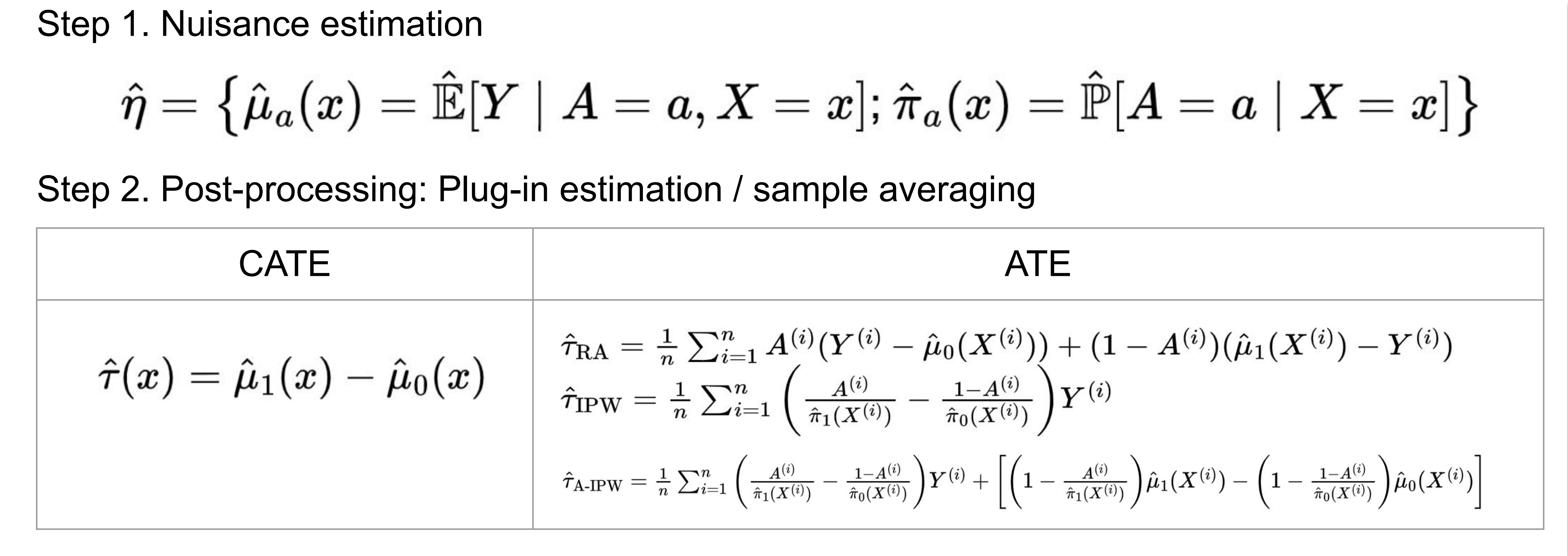
Loss-based (TMLE):

- DragonNet

ML and estimation: Plug-in (one-step) learners

Plug-in (one-step) learners

- With infinite 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
$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$	$\hat{\tau}_{\text{RA}} = \frac{1}{n}\sum_{i=1}^{n} A^{(i)}(Y^{(i)} - \hat{\mu}_0(X^{(i)})) + (1 - A^{(i)})(\hat{\mu}_1(X^{(i)}) - Y^{(i)})$ $\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A^{(i)}}{\hat{\pi}_1(X^{(i)})} - \frac{1-A^{(i)}}{\hat{\pi}_0(X^{(i)})}\right)Y^{(i)}$ $\hat{\tau}_{\text{A-IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A^{(i)}}{\hat{\pi}_1(X^{(i)})} - \frac{1-A^{(i)}}{\hat{\pi}_0(X^{(i)})}\right)Y^{(i)} + \left[\left(1 - \frac{A^{(i)}}{\hat{\pi}_1(X^{(i)})}\right)\hat{\mu}_1(X^{(i)}) - \left(1 - \frac{1-A^{(i)}}{\hat{\pi}_0(X^{(i)})}\right)\hat{\mu}_0(X^{(i)})\right]$

- 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 learners in finite-sample

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 IPW? Can we do even better (e.g., more efficient, more robust) in estimating CATE / ATE?
4. Can we do data-driven model selection?
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 selection bias?
6. Can we incorporate inductive biases for nuisance functions?
7. Can we do end-to-end learning?

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, the fundamental problem of causal inference

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):

Sub-group treatment effects

  - ○ **CATE** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\Big(\boxed{\big(Y[1] - Y[0]} - \hat{\tau}(V)\big)\Big)^2$ never observed
  - ○ **CAPO** $\mathcal{L}(\hat{\tau}) = \mathbb{E}\Big(\boxed{\big(Y[a]} - \hat{\tau}(V, a)\big)\Big)^2$ sometimes observed

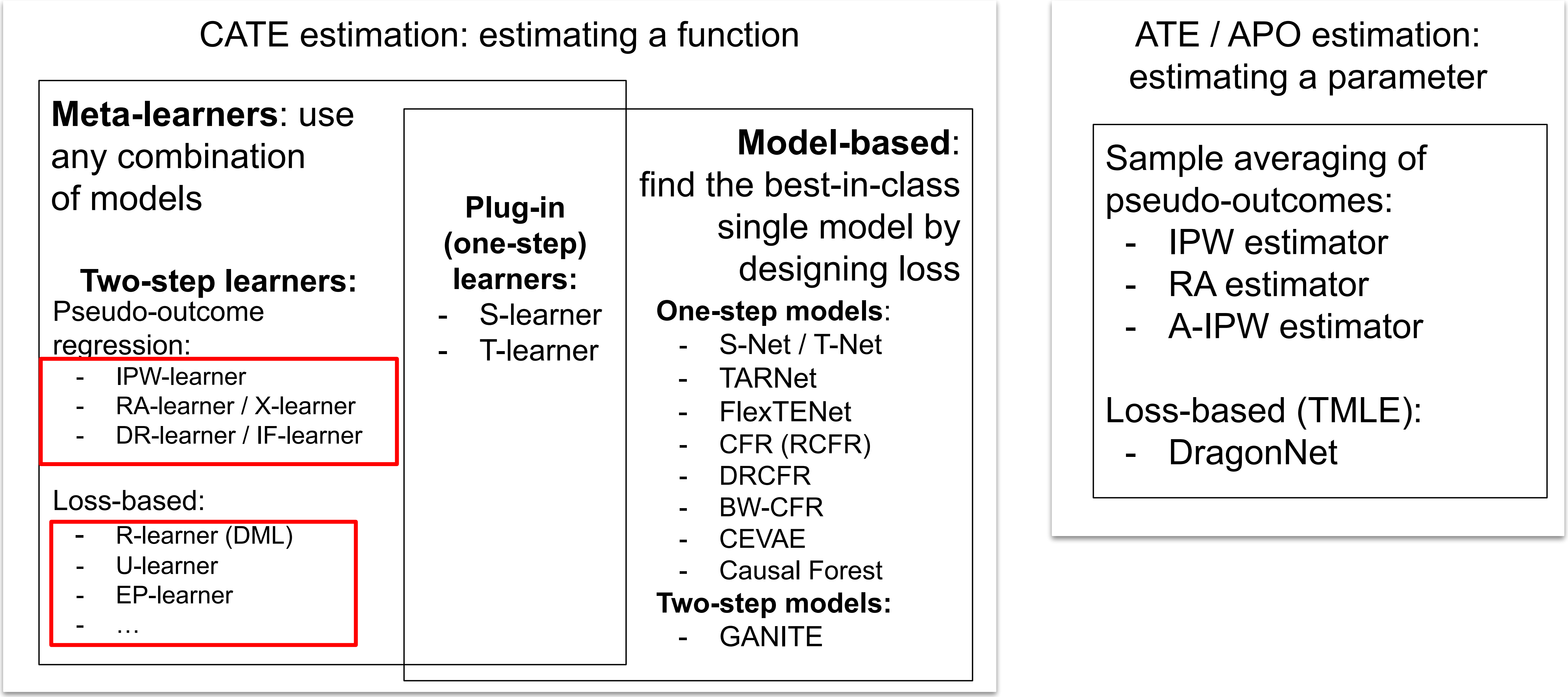
- But, the fundamental problem of causal inference

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

ML and estimation: Two-step learners



CATE estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:

Pseudo-outcome regression:

- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:

- R-learner (DML)
- U-learner
- EP-learner
- …

Plug-in (one-step) learners:

- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:

- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- CEVAE
- Causal Forest

Two-step models:

- GANITE

ATE / APO estimation: estimating a parameter

Sample averaging of pseudo-outcomes:

- IPW estimator
- RA estimator
- A-IPW estimator

Loss-based (TMLE):

- DragonNet

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

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

Sub-group treatment effects

- ATE = Sub-group treatment effect with $V = \emptyset \quad (V \subseteq X)$
  Sample averaging = Regression with intercept only

- **Idea 1**: create **pseudo-outcomes** $\tilde{Y}_{\hat{\eta}}$ with the main property $\mathbb{E}(\tilde{Y}_\eta \mid V = v) = \tau(v)$

$$\tilde{Y}_{\text{RA},\hat{\eta}} = A(Y - \hat{\mu}_0(X)) + (1 - A)(\hat{\mu}_1(X) - Y)$$
$$\tilde{Y}_{\text{IPW},\hat{\eta}} = \left(\frac{A}{\hat{\pi}_1(X)} - \frac{1-A}{\hat{\pi}_0(X)}\right)Y$$
$$\tilde{Y}_{\text{DR},\hat{\eta}} = \left(\frac{A}{\hat{\pi}_1(X)} - \frac{1-A}{\hat{\pi}_0(X)}\right)Y + \left[\left(1 - \frac{A}{\hat{\pi}_1(X)}\right)\hat{\mu}_1(X) - \left(1 - \frac{1-A}{\hat{\pi}_0(X)}\right)\hat{\mu}_0(X)\right]$$

- We regress on 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. “What about the sub-group treatment effects?”

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

Sub-group treatment effects

- **Idea 2**: use nuisance parameters to design a **loss**, so that CATE are well estimated, for example with Robinson decomposition:

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

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

- Then the custom **residuals loss** is following:

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

ML and estimation: Pseudo-outcomes vs custom residualized loss

Pseudo-outcomes vs custom residualized loss

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

Nuisance parameters	Pseudo-outcome based	Loss-based
Estimated	$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$	$\mathcal{L}(\hat{\tau}, \hat{\eta}) = \mathbb{E}\Big((Y - \mu(\hat{X})) - (A - \hat{\pi}_1(X))\hat{\tau}(V)\Big)^2$
Ground-truth	?	?

ML and estimation: Pseudo-outcomes vs custom residualized loss

Pseudo-outcomes vs custom residualized loss

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

Nuisance parameters	Pseudo-outcome based	Loss-based
Estimated	$\mathcal{L}(\hat{\tau},\hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$	$\mathcal{L}(\hat{\tau},\hat{\eta}) = \mathbb{E}\Big((Y - \mu(\hat{X})) - (A - \hat{\pi}_1(X))\hat{\tau}(V)\Big)^2$
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: Pseudo-outcomes vs custom residualized loss

- If we would use ground-truth nuisance parameters, the losses aim at the ground truth CATE or weighted CATE

Pseudo-outcomes vs custom residualized loss

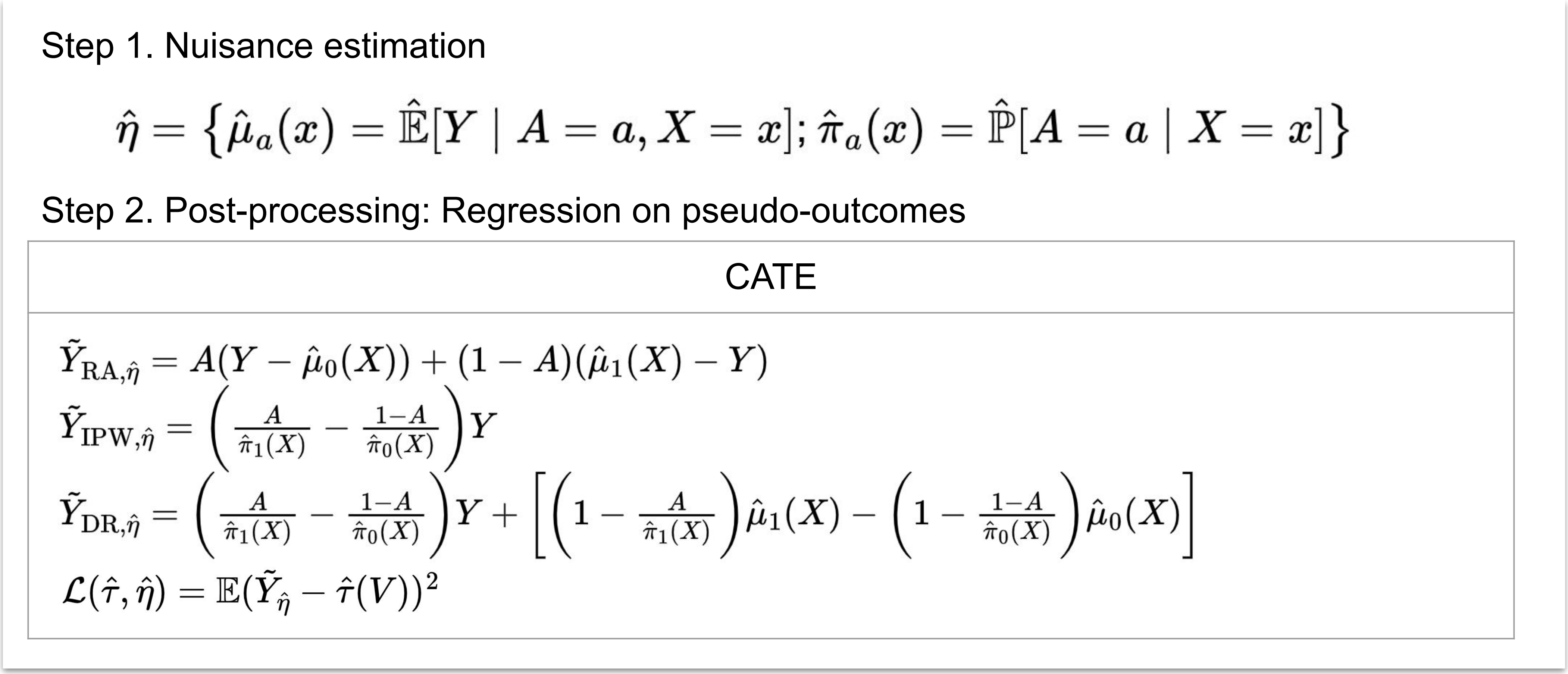
Nuisance parameters	Pseudo-outcome based	Loss-based
Estimated	$\mathcal{L}(\hat{\tau},\hat{\eta}) = \mathbb{E}(\tilde{Y}_{\hat{\eta}} - \hat{\tau}(V))^2$	$\mathcal{L}(\hat{\tau},\hat{\eta}) = \mathbb{E}\Big((Y - \mu(\hat{X})) - (A - \hat{\pi}_1(X))\hat{\tau}(V)\Big)^2$
Ground-truth	$\mathcal{L}(\hat{\tau},\eta) =$ $\mathbb{E}\big((Y(1) - Y(0)) - \hat{\tau}(V)\big)^2$	$\boxed{\mathcal{L}(\hat{\tau},\eta) = \mathbb{E}\big(\pi_1(X)\pi_0(X)\big(\tau(V) - \hat{\tau}(V)\big)\big)^2}$

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

ML and estimation: Two-step learners

- Two-step learners, based on pseudo-adjust are, **IPW-learner**, **RA-learner / X-learner,** and doubly-robust (**DR)-learner / influence-function (IF-learner)**

Two-step 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

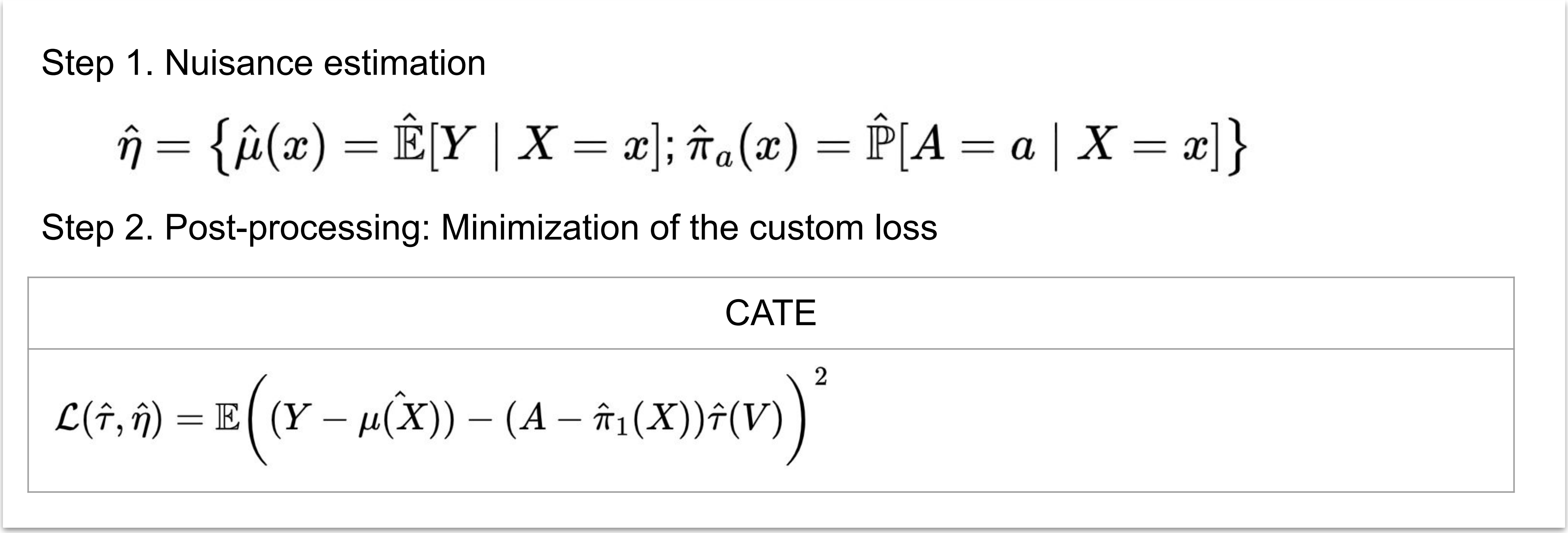
CATE
$\tilde{Y}_{\mathrm{RA},\hat{\eta}} = A(Y - \hat{\mu}_0(X)) + (1 - A)(\hat{\mu}_1(X) - Y)$ $\tilde{Y}_{\mathrm{IPW},\hat{\eta}} = \left(\frac{A}{\hat{\pi}_1(X)} - \frac{1-A}{\hat{\pi}_0(X)}\right)Y$ $\tilde{Y}_{\mathrm{DR},\hat{\eta}} = \left(\frac{A}{\hat{\pi}_1(X)} - \frac{1-A}{\hat{\pi}_0(X)}\right)Y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

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

ML and estimation: Two-step learners

- Other alternative is **residualized (R)-learner**:

Two-step learners



Step 1. Nuisance estimation

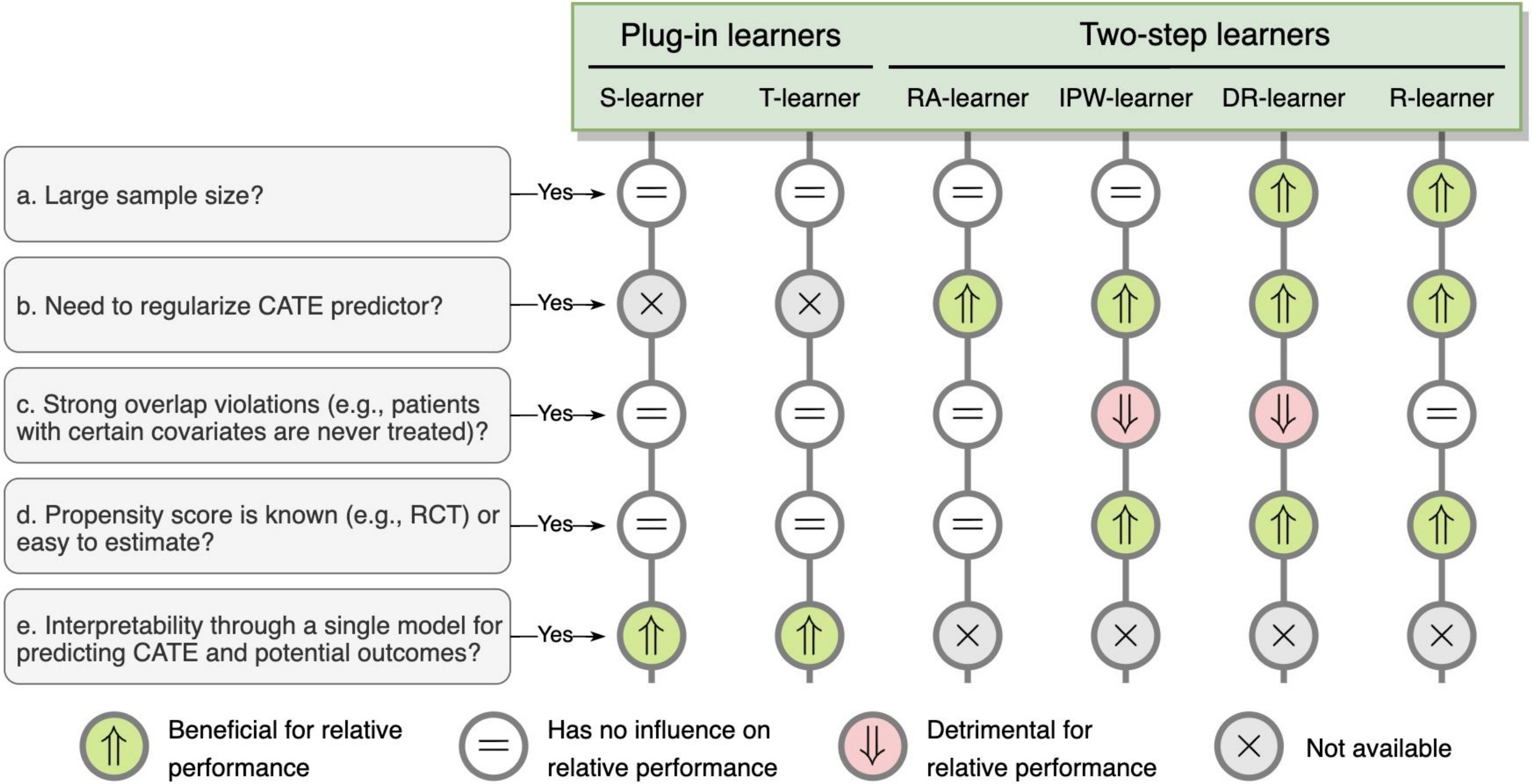
$$\hat{\eta} = \left\{\hat{\mu}(x) = \hat{\mathbb{E}}[Y \mid X = x]; \hat{\pi}_a(x) = \hat{\mathbb{P}}[A = a \mid X = x]\right\}$$

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

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

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

ML and estimation: Plug-in (one-step) vs two-step learners



		Plug-in learners		Two-step learners			
		S-learner	T-learner	RA-learner	IPW-learner	DR-learner	R-learner
a. Large sample size?	Yes →	=	=	=	=	⇑	⇑
b. Need to regularize CATE predictor?	Yes →	×	×	⇑	⇑	⇑	⇑
c. Strong overlap violations (e.g., patients with certain covariates are never treated)?	Yes →	=	=	=	⇓	⇓	=
d. Propensity score is known (e.g., RCT) or easy to estimate?	Yes →	=	=	=	⇑	⇑	⇑
e. Interpretability through a single model for predicting CATE and potential outcomes?	Yes →	⇑	⇑	×	×	×	×

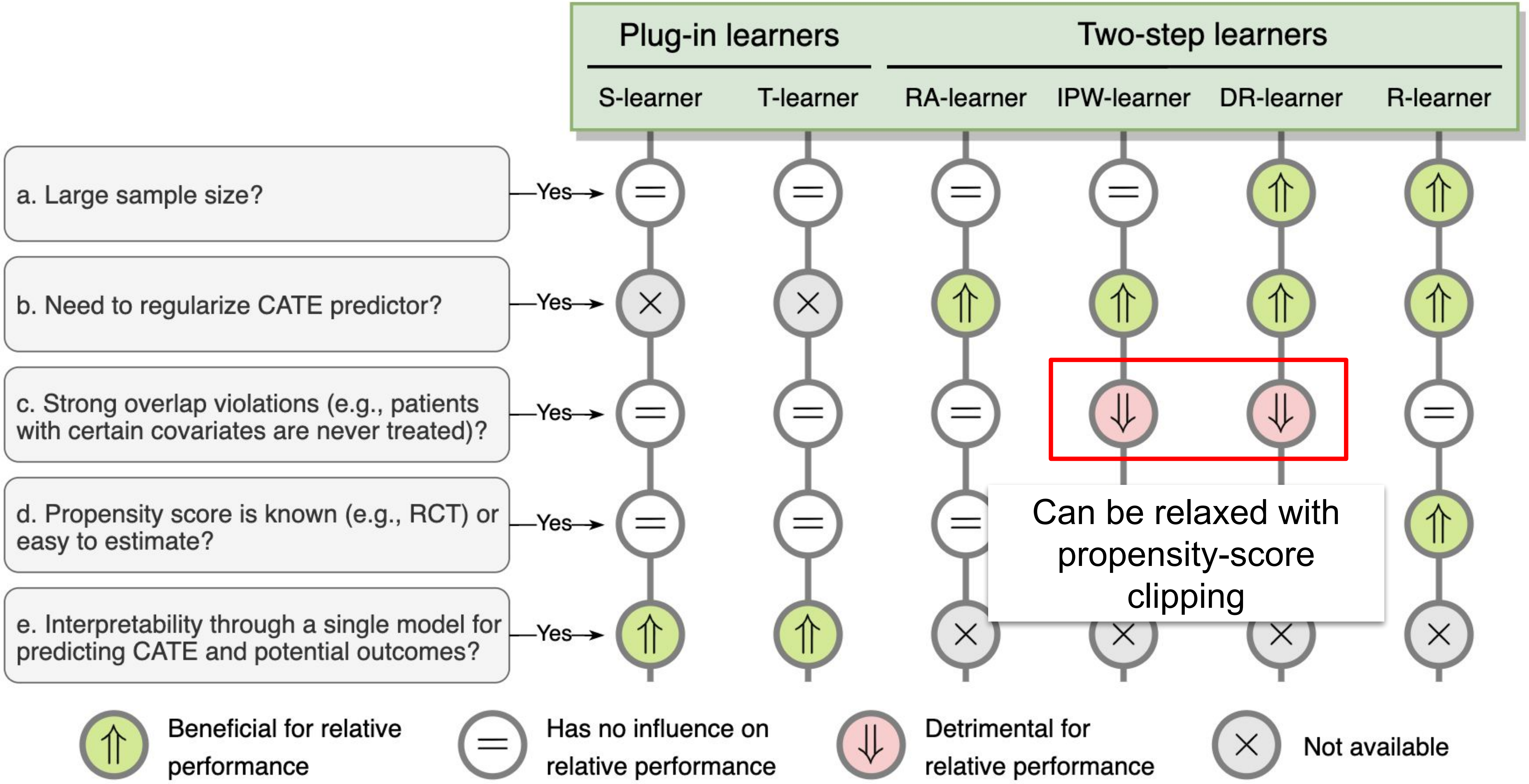
⇑ Beneficial for relative performance

= Has no influence on relative performance

⇓ Detrimental for relative performance

× Not available

ML and estimation: Plug-in (one-step) vs two-step learners



	Plug-in learners		Two-step learners			
	S-learner	T-learner	RA-learner	IPW-learner	DR-learner	R-learner
a. Large sample size? —Yes→	=	=	=	=	⇑	⇑
b. Need to regularize CATE predictor? —Yes→	×	×	⇑	⇑	⇑	⇑
c. Strong overlap violations (e.g., patients with certain covariates are never treated)? —Yes→	=	=	=	⇓	⇓	=
d. Propensity score is known (e.g., RCT) or easy to estimate? —Yes→	=	=	=			⇑
e. Interpretability through a single model for predicting CATE and potential outcomes? —Yes→	⇑	⇑	×	×	×	×

Can be relaxed with propensity-score clipping

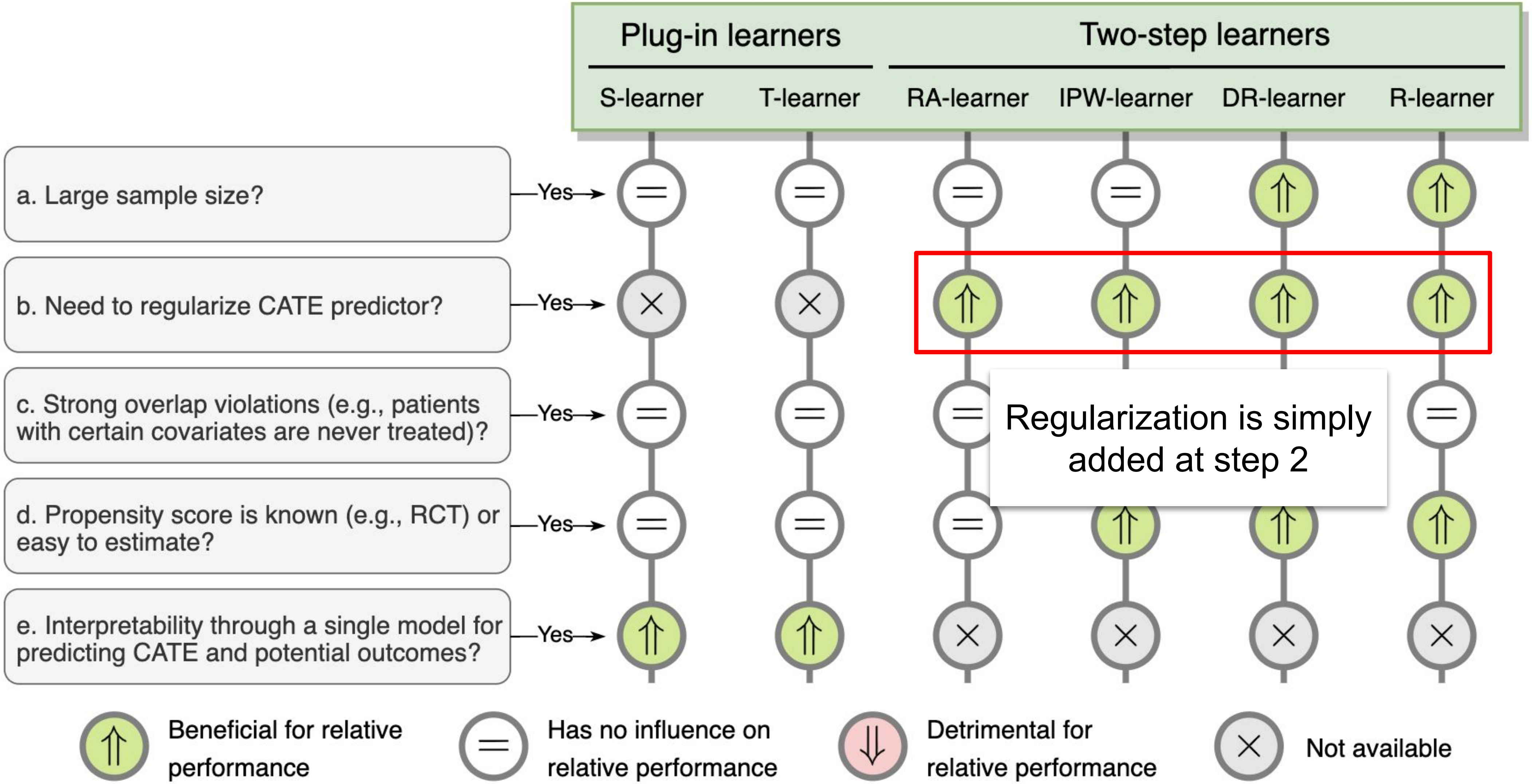
⇑ Beneficial for relative performance

= Has no influence on relative performance

⇓ Detrimental for relative performance

× Not available

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



		Plug-in learners		Two-step learners			
		S-learner	T-learner	RA-learner	IPW-learner	DR-learner	R-learner
a. Large sample size?	Yes →	=	=	=	=	⇑	⇑
b. Need to regularize CATE predictor?	Yes →	×	×	⇑	⇑	⇑	⇑
c. Strong overlap violations (e.g., patients with certain covariates are never treated)?	Yes →	=	=	=			=
d. Propensity score is known (e.g., RCT) or easy to estimate?	Yes →	=	=	=	⇑	⇑	⇑
e. Interpretability through a single model for predicting CATE and potential outcomes?	Yes →	⇑	⇑	×	×	×	×

Regularization is simply added at step 2

⇑ Beneficial for relative performance

= Has no influence on relative performance

⇓ Detrimental for relative performance

× Not available

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

Asymptotically speaking:

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

Finite dimensional estimands

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

- 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. “What is better, adjustment or IPW?”

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 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: Neyman orthogonal methods

CATE estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:

Pseudo-outcome regression:

- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:

- R-learner (DML)
- U-learner
- EP-learner
- …

Plug-in (one-step) learners:

- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:

- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- CEVAE
- Causal Forest

Two-step models:

- GANITE

ATE / APO estimation: estimating a parameter

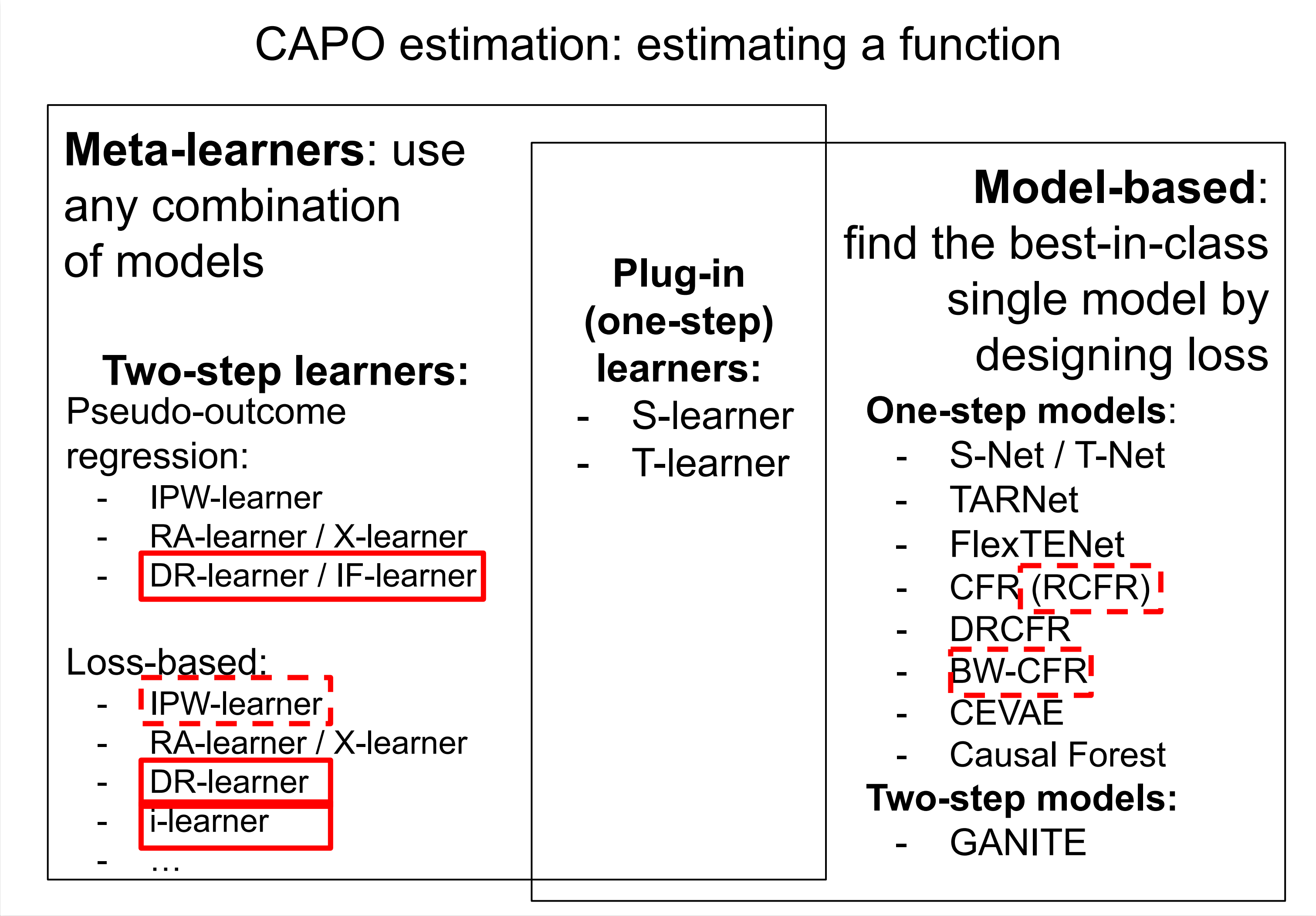
Sample averaging of pseudo-outcomes:

- IPW estimator
- RA estimator
- A-IPW estimator

Loss-based (TMLE):

- DragonNet

ML and estimation: Neyman orthogonal methods



CAPO estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:
Pseudo-outcome regression:
- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:
- IPW-learner
- RA-learner / X-learner
- DR-learner
- i-learner
- …

Plug-in (one-step) learners:
- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:
- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- CEVAE
- Causal Forest

Two-step models:
- GANITE

ATE / APO estimation: estimating a parameter

Sample averaging of pseudo-outcomes:
- IPW estimator
- RA estimator
- A-IPW estimator

Loss-based (TMLE):
- DragonNet

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

Best asymptotically does not mean best in low-sample!

"No Free Lunch" :(

Best approach in low-sample regime

ML and estimation: 4. “Can we do data-driven model selection?”

Best asymptotically does not mean best in low-sample!

"No Free Lunch" :(

Best approach in low-sample regime

+

Now, we don’t even have **data-driven model selection criteria**, but only heuristics (Curth & van der Schaar, 2023)

ML and estimation: 4. “Can we do data-driven model selection?”

Best asymptotically does not mean best in low-sample!

"No Free Lunch" :(

Best approach in low-sample regime

+

Now, we don’t even have **data-driven model selection criteria**, but only heuristics (Curth & van der Schaar, 2023)

ML and estimation: 4. “Can we do data-driven model selection?”

Best asymptotically does not mean best in low-sample!

"No Free Lunch" :(

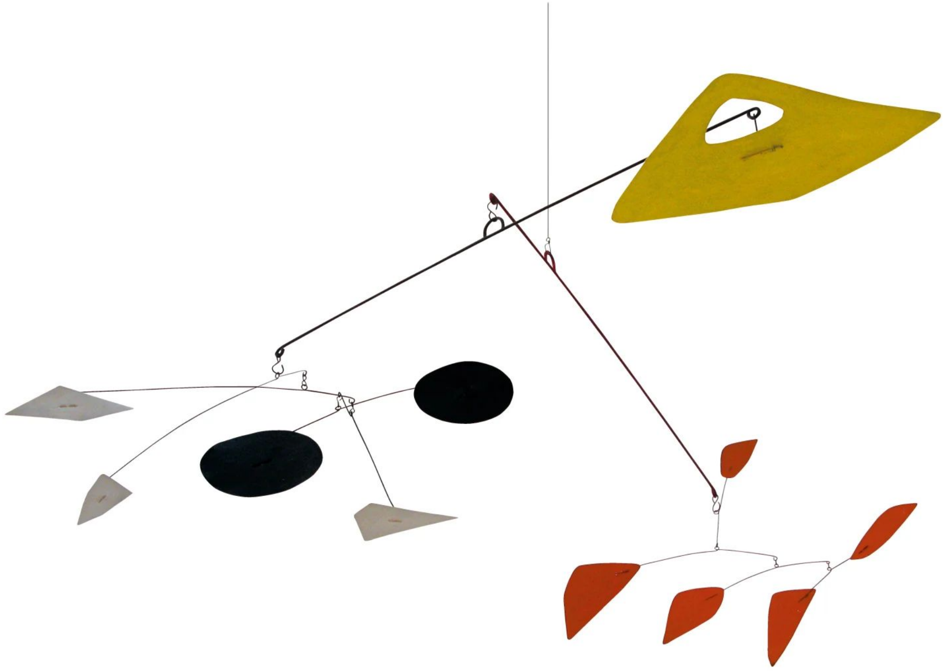
Best approach in low-sample regime

Possible solution: employ RCT (L2) data (with sub-group level counterfactuals)

ML and estimation: 5. “How to address the selection bias?”

Should we do something?

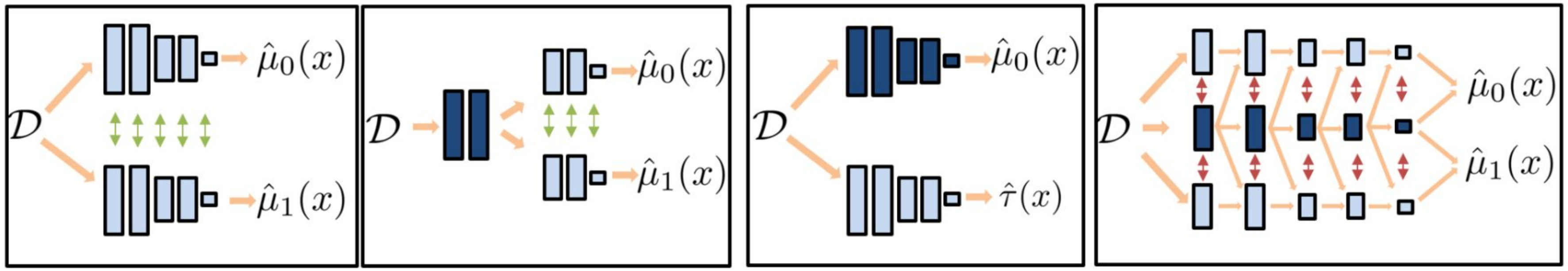
- Selection bias matters in low-sample regime, e.g. $\hat{\mu}_a(x)$ overfits on the factual data with high propensity
- Thus, plug-in (one-step) learners are sub-optimal in a sense, that they don’t use all the data
- Two-step learners act like ‘regularizers’ on the first stage output, acting on the overfitted models
- But by using two-step learners, we introduce more parameters to estimate and need to do sample-splitting



Alexander Calder - Untitled

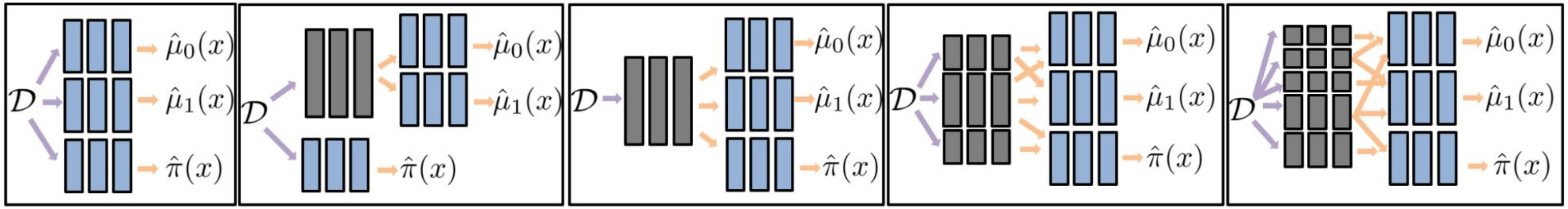
ML and estimation: 6. “Can we incorporate inductive biases for nuisance functions estimation?”

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



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

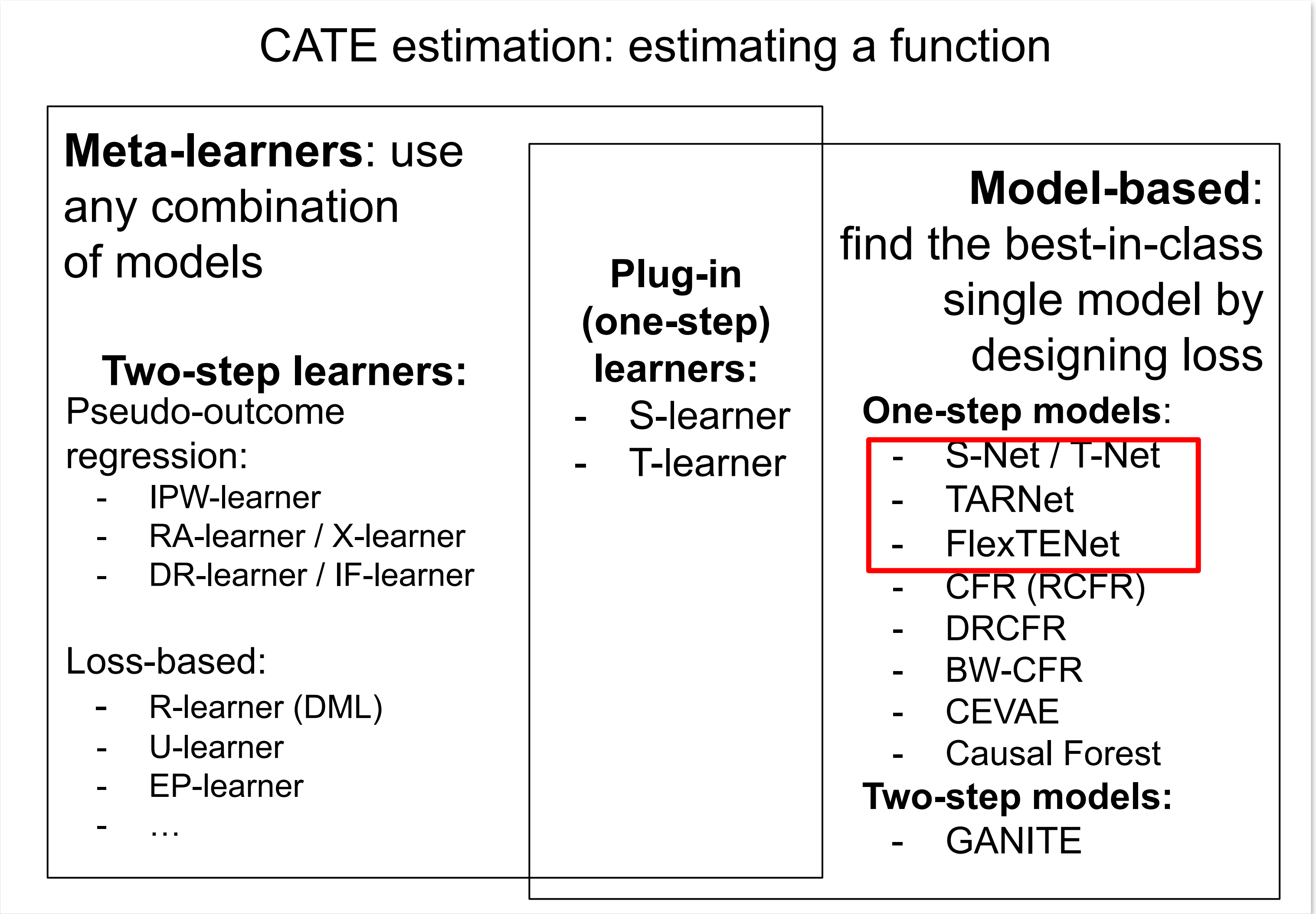
Sharing representations for all the nuisance functions



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: Addressing selection bias



CATE estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:

Pseudo-outcome regression:

- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:

- R-learner (DML)
- U-learner
- EP-learner
- …

Plug-in (one-step) learners:

- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:

- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- CEVAE
- Causal Forest

Two-step models:

- GANITE

ATE / APO estimation: estimating a parameter

Sample averaging of pseudo-outcomes:

- IPW estimator
- RA estimator
- A-IPW estimator

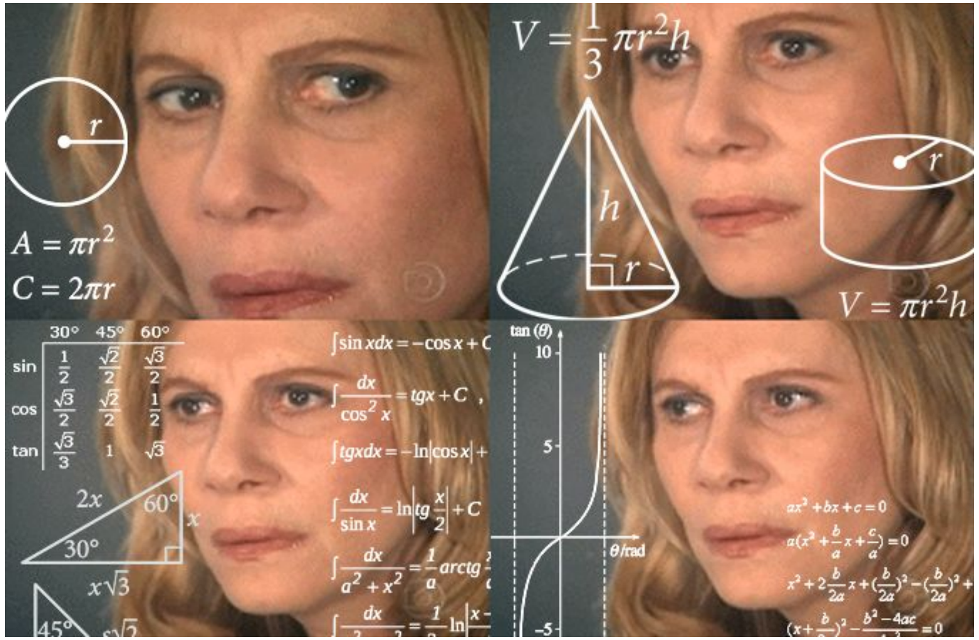
Loss-based (TMLE):

- DragonNet

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? (We hope it can be)

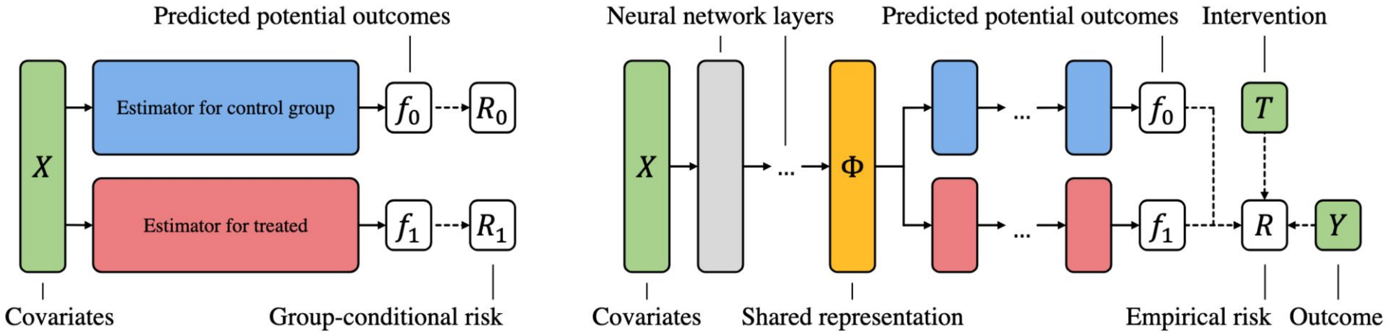
ML and estimation: 7. “Can we do end-to-end learning?”

Representation learning for CATE estimation

- We want to design a loss to find best-in-class model to estimate CATE.
- **Idea**: employ representation learning to map the covariates to a lower-dimensional space and reduce variance of CATE estimation:

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

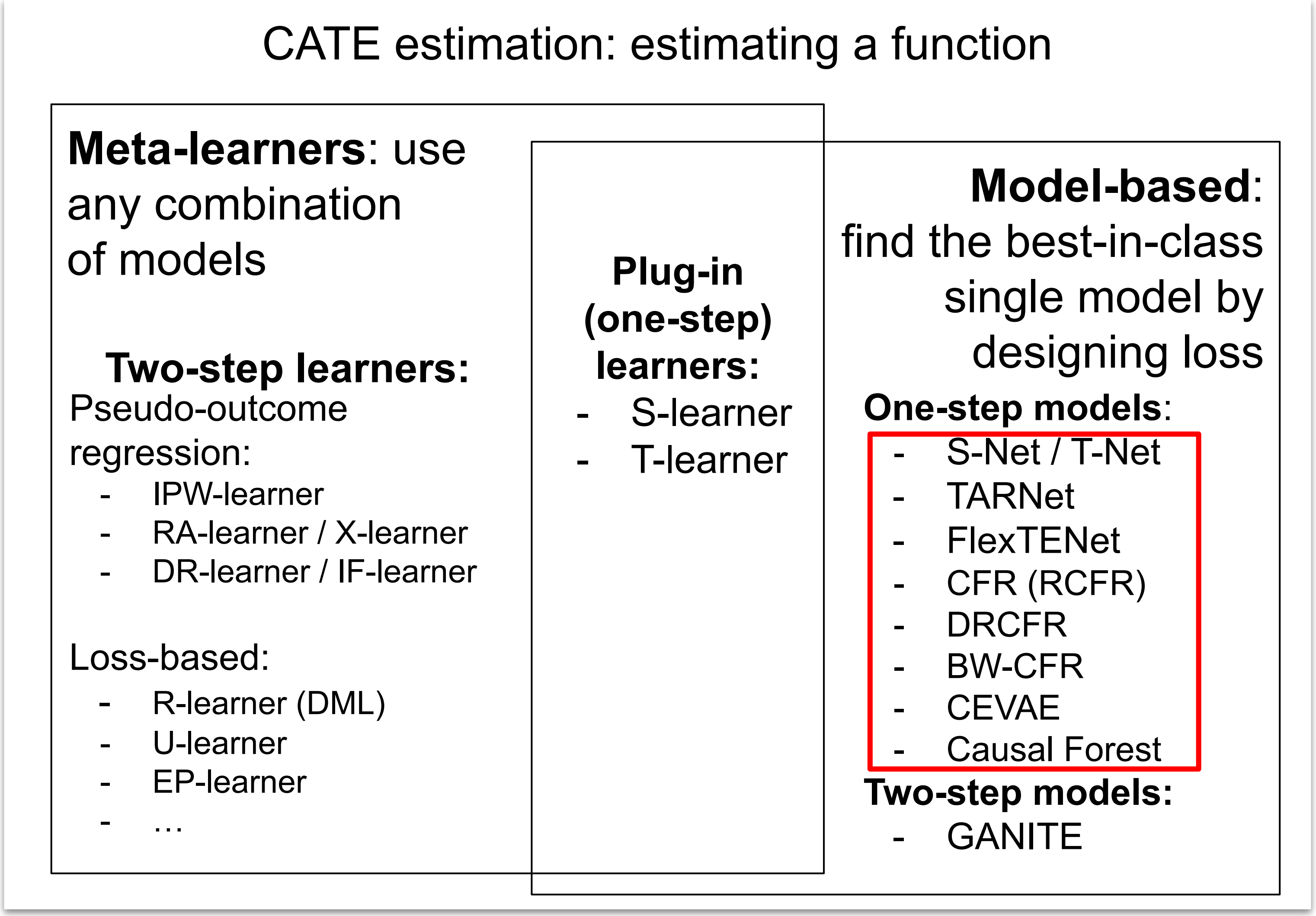
- Holy grail: **prognostic score**, namely minimal sufficient information in covariates for CATE estimation.
- Most common implementation, neural-network based approach, e.g., TARNet:



(a) T-learner

(b) TARNet (Shalit et al., 2017)

ML and estimation: End-to-end learning methods



CATE estimation: estimating a function

Meta-learners: use any combination of models

Two-step learners:

Pseudo-outcome regression:

- IPW-learner
- RA-learner / X-learner
- DR-learner / IF-learner

Loss-based:

- R-learner (DML)
- U-learner
- EP-learner
- …

Plug-in (one-step) learners:

- S-learner
- T-learner

Model-based: find the best-in-class single model by designing loss

One-step models:

- S-Net / T-Net
- TARNet
- FlexTENet
- CFR (RCFR)
- DRCFR
- BW-CFR
- CEVAE
- Causal Forest

Two-step models:

- GANITE

ATE / APO estimation: estimating a parameter

Sample averaging of pseudo-outcomes:

- IPW estimator
- RA estimator
- A-IPW estimator

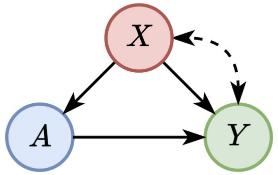
Loss-based (TMLE):

- DragonNet

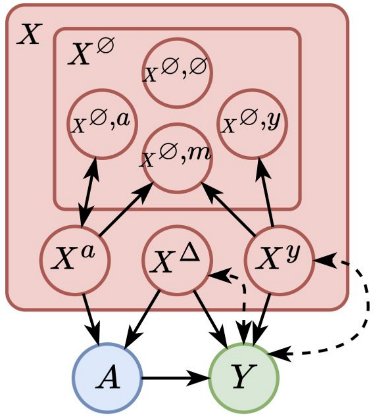
ML and estimation: Representation learning for CATE

Prognostic scores

- For identifying prognostic score, we would need to know the structure inside of X, namely, what are the ground-truth confounders, instruments, and noise:



Original causal diagram



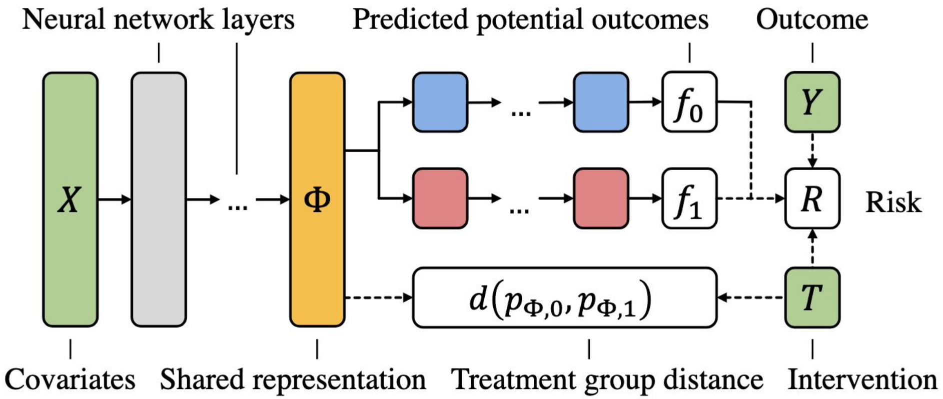
Clustered causal diagram

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

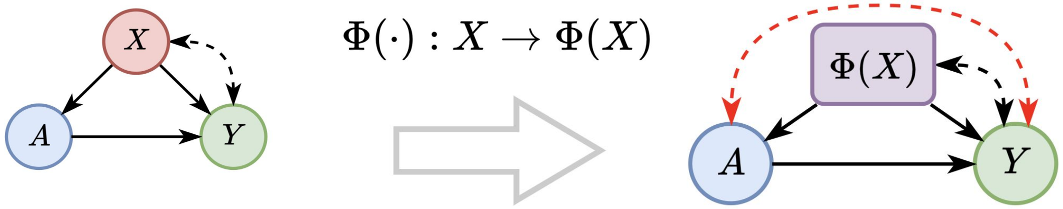
ML and estimation: Representation learning for CATE

Balanced representations

- (Shalit et al. 2017) proposed to enforce treatment balancing on top of the **invertible** representations with Counterfactual Regression (CFR):



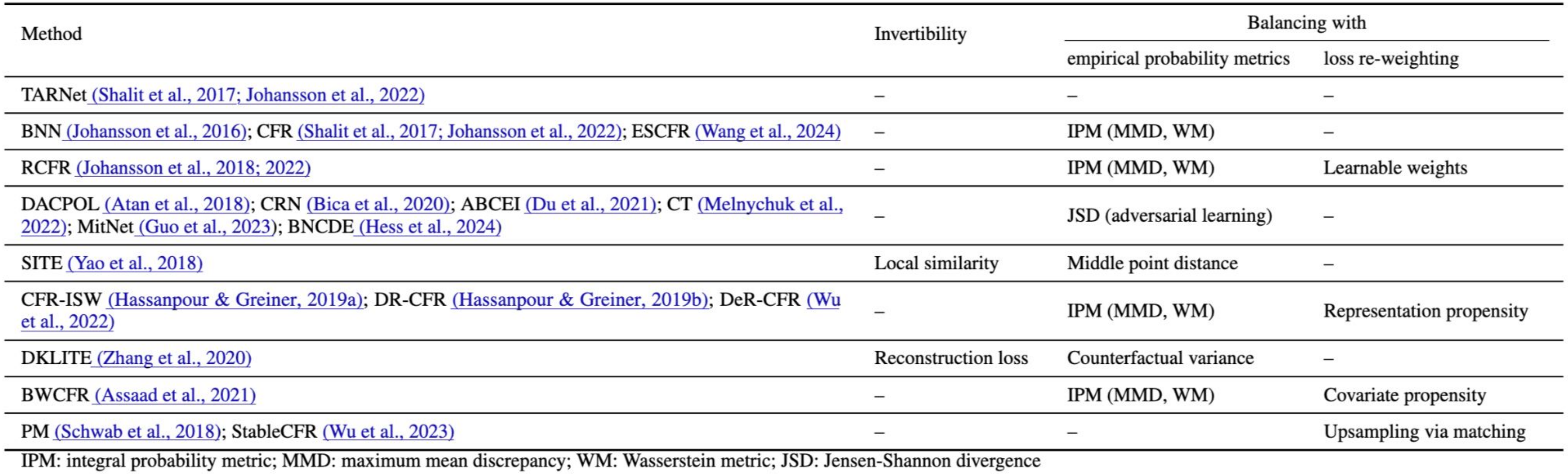
- It was shown, that we can improve the counterfactual generalization risk (= address selection bias).
- We can also build CFR with low-dimensional (=non-invertible) representations, but then we can induce the confounding bias (Melnychuk et al. 2023).



ML and estimation: Representation learning for CATE

Post-CFR papers

- After CFR, the whole bunch of methods were proposed (which is not really helpful tbh):



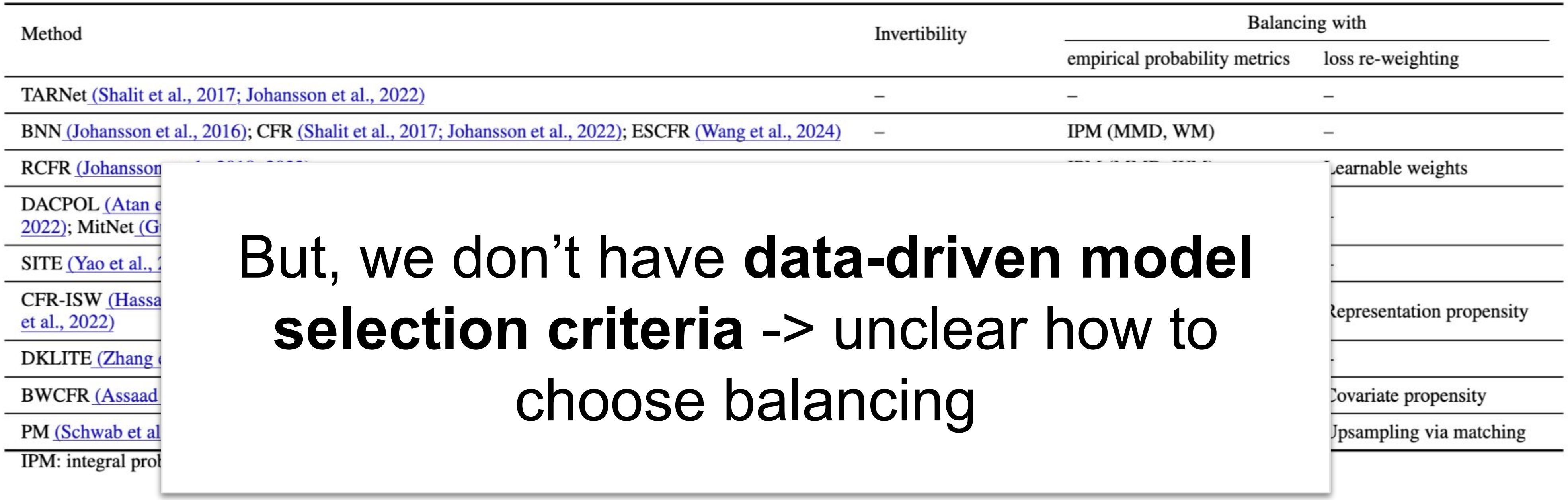
Method	Invertibility	Balancing with	
		empirical probability metrics	loss re-weighting
TARNet (Shalit et al., 2017; Johansson et al., 2022)	–	–	–
BNN (Johansson et al., 2016); CFR (Shalit et al., 2017; Johansson et al., 2022); ESCFR (Wang et al., 2024)	–	IPM (MMD, WM)	–
RCFR (Johansson et al., 2018; 2022)	–	IPM (MMD, WM)	Learnable weights
DACPOL (Atan et al., 2018); CRN (Bica et al., 2020); ABCEI (Du et al., 2021); CT (Melnychuk et al., 2022); MitNet (Guo et al., 2023); BNCDE (Hess et al., 2024)	–	JSD (adversarial learning)	–
SITE (Yao et al., 2018)	Local similarity	Middle point distance	–
CFR-ISW (Hassanpour & Greiner, 2019a); DR-CFR (Hassanpour & Greiner, 2019b); DeR-CFR (Wu et al., 2022)	–	IPM (MMD, WM)	Representation propensity
DKLITE (Zhang et al., 2020)	Reconstruction loss	Counterfactual variance	–
BWCFR (Assaad et al., 2021)	–	IPM (MMD, WM)	Covariate propensity
PM (Schwab et al., 2018); StableCFR (Wu et al., 2023)	–	–	Upsampling via matching

IPM: integral probability metric; MMD: maximum mean discrepancy; WM: Wasserstein metric; JSD: Jensen-Shannon divergence

- If representations are low-dimensional, then they might contain **confounding bias** -> but this might be fine, we just consider it as a part of the **statistical bias-variance trade-off**

ML and estimation: Representation learning for CATE

- After CFR, the whole bunch of methods were proposed (which is not really helpful tbh):



Method	Invertibility	Balancing with	
		empirical probability metrics	loss re-weighting
TARNet (Shalit et al., 2017; Johansson et al., 2022)	–	–	–
BNN (Johansson et al., 2016); CFR (Shalit et al., 2017; Johansson et al., 2022); ESCFR (Wang et al., 2024)	–	IPM (MMD, WM)	–
RCFR (Johansson			Learnable weights
DACPOL (Atan e 2022); MitNet (G			–
SITE (Yao et al.,			–
CFR-ISW (Hassa et al., 2022)			Representation propensity
DKLITE (Zhang			–
BWCFR (Assaad			Covariate propensity
PM (Schwab et al			Upsampling via matching

IPM: integral prob

But, we don't have **data-driven model selection criteria** -> unclear how to choose balancing

Post-CFR papers

- If representations are low-dimensional, then they might contain **confounding bias** -> but this might be fine, we just consider it as a part of the **statistical bias-variance trade-off**

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN LMU MUNICH SCHOOL OF MANAGEMENT INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

Extensions

Extensions: New challenges

Uncertainty of TEs / POs

- Epistemic uncertainty was studied for CATE / CAPO
- Aleatoric uncertainty for POs (Melnychuk et al. 2023), TEs (submitted to NeurIPS 2024)
- Total uncertainty for CATE and CAPO with conformal prediction

Hidden confounding

- Marginal sensitivity model, general sensitivity model (Frauen et al. 2023), B-learner
- Instrumental variables regression
- Proxy variables

Time-varying potential outcomes

- LSTMs / Transformer-based models
- Irregular sampling times / continuous time

Explainability Interpretability

- Explainability/interpretability of two-step learners

Thank you for your attention!

Main message: CATE estimation is very different from regular ML predictive modelling

Questions?

🙏