



Machine learning and causal inference: a two-way road

Uri Shalit
Technion – Israel Institute of Technology

DATAIA Seminar
Paris, January 2020

What is causality?

A big question!



Extremely short intro to causality (in the context of statistics and learning)

- Aspirin **caused** my headache to disappear
- The car crashed **because** it didn't brake in time
- The students succeeded **because** of the new teacher

Extremely short intro to causality (in the context of statistics and learning)

- Aspirin **caused** my headache to disappear
 - Had I not taken Aspirin, I would still have had the headache
- The car crashed **because** it didn't brake in time
 - Had the car braked in time, it wouldn't have crashed
- The students succeeded **because** of the new teacher
 - Had the students remained with the old teacher, they wouldn't have succeeded

Extremely short intro to causality (in the context of statistics and learning)

- Aspirin **caused** my headache to disappear
 - Had I not taken Aspirin, I would still have had a headache
- The car crashed **because** it didn't brake
 - Had the car braked in time, it wouldn't have crashed
- The students succeeded **because** of the new teacher
 - Had the students remained with the old teacher, they wouldn't have succeeded

counterfactuals

Extremely short intro to causality

(in the context of

- Aspirin **causes**

- Had I not

- The car crash

- Had the

- The student

- Had the student
succeeded

Counterfactuals:
imagine a world
where everything is
the same except
the “cause”

headache

they wouldn't have

Counterfactuals

- Often in terms of imagined interventions
- Never directly observable – we need a causal model
 - “Counterfactual world” is sometimes statistically identical to observed reality, for example in Randomized Controlled Trials

Outline

- ML for causal inference
- Causal inference for ML
 - Off-policy evaluation in a partially observable Markov decision process
 - Robust learning for unsupervised covariate shift

Outline

- **ML for causal inference**
- Causal inference for ML
 - Off-policy evaluation in a partially observable Markov decision process
 - Robust learning for unsupervised covariate shift

Causal *effect* inference questions

- Which medication will make patients better?
- Which economic policy will lower unemployment?
- The effects of **actions** on **outcomes**

Causal effect inference from observational data

- Which medication will make patients better?
 - Infer from medical records
- Which economic policy will lower unemployment?
 - Infer from past economic measurement
- The effects of **actions** on **outcomes**

Causal inference from observational data - **confounding**

- Which medication will make patients better?
 - Infer from medical records
 - Maybe **younger/wealthier/female/...** patients tend to receive medication A over B?
- Which economic policy will lower unemployment?
 - Infer from past economic measurement
 - Maybe policy was enacted in **better past economic times**?

This part based on work with Fredrik Johansson (MIT→Chalmers), Nathan Kallus (Cornell) and David Sontag (MIT)

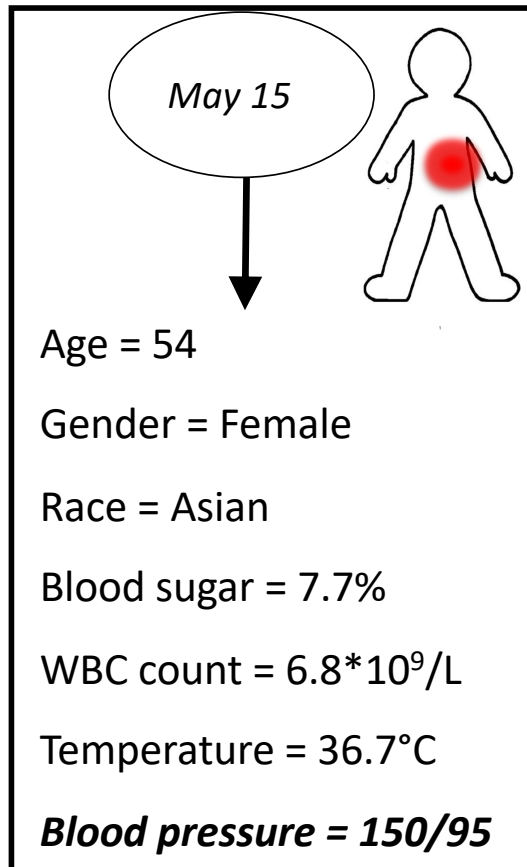
(i) Johansson, S, Sontag, (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning*.

(ii) Shalit, U., Johansson, F., & Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*.

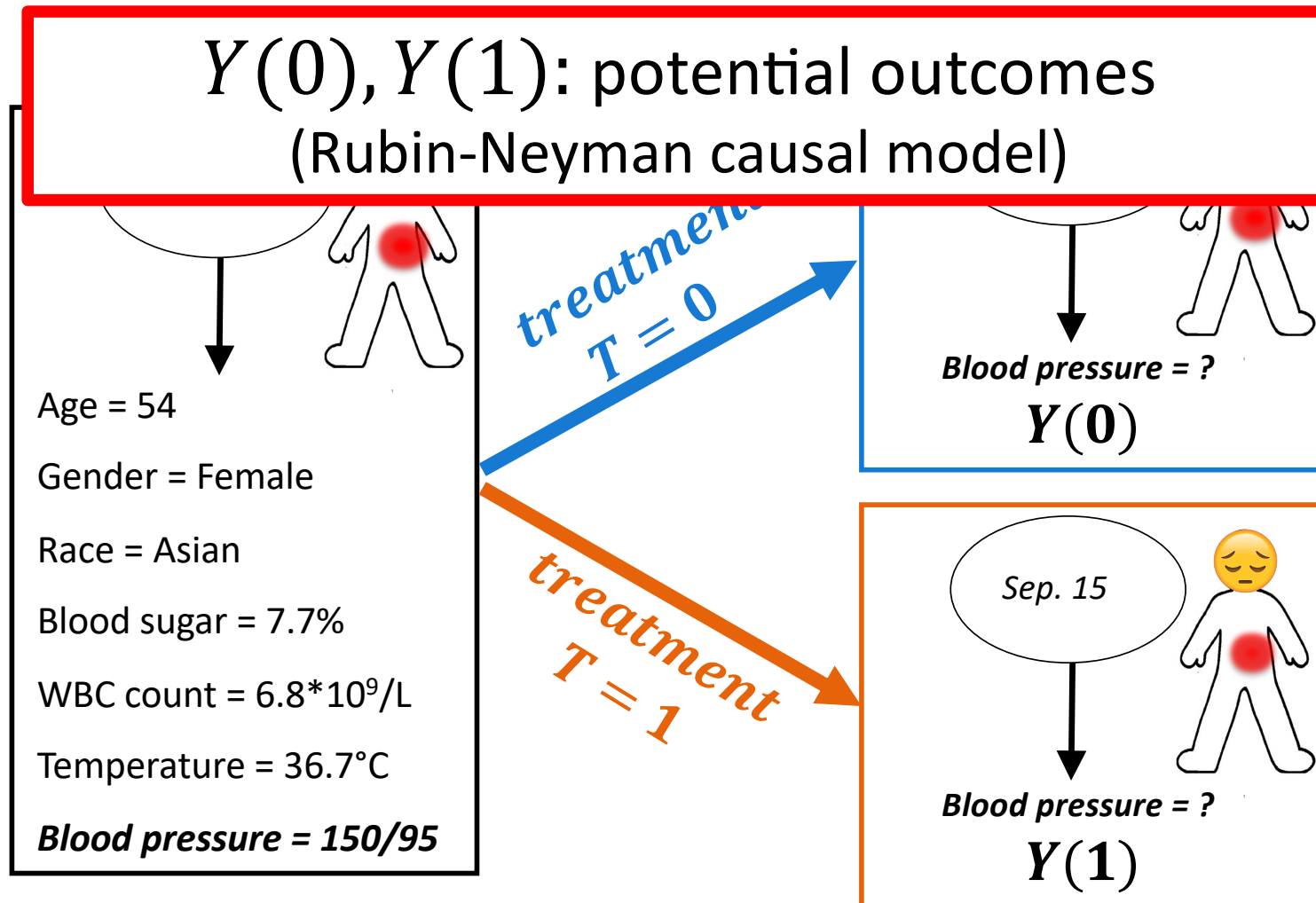
(iii) Johansson, Kallus, S, Sontag, (2020)
Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*.

Our goal: Conditional Average Treatment Effect (*CATE*)

Anna



Our goal: Conditional Average Treatment Effect (CATE)



Our goal: Conditional Average Treatment Effect (*CATE*)

$Y(0), Y(1)$: potential outcomes
(Rubin-Neyman causal model)

X : patient features

$$CATE(X) := \mathbb{E}[Y(1) - Y(0)|X]$$

Gender = Female

Race = Asian

Blood sugar = 7.7%

WBC count = $6.8 \cdot 10^9/L$

Temperature = $36.7^\circ C$

Blood pressure = 150/95

treatment
 $t = 1$

Sep. 15



Blood pressure = ?

Y_1

$Y(0), Y(1)$: potential outcomes
(Rubin-Neyman causal model)

X : patient features

$$CATE(X) := \mathbb{E}[Y(1) - Y(0)|X]$$

- We never directly observe CATE
- We only see either $Y(1)$ or $Y(0)$
- The choice is *not random*
- How to estimate the CATE function?

Blood pressure = 150/95

Y_1

Estimate potential outcomes

- **Outcomes** under treatment and control, $Y(1), Y(0) \in \mathbb{R}$

- **Treatments** $T \in \{0,1\}$, $Y = TY(1) + (1 - T)Y(0)$

- **Confounders** $X \in \mathbb{R}^d$

Only one observed for any one patient!

- **Conditional effect (CATE)** $\tau(X) := \mathbb{E}[Y(1) - Y(0) \mid X]$

Observational datasets: Rheumatoid arthritis

- ▶ Historical records of treatments and outcomes

Patient	X		T	Y
	Age	Prior disease activity	Observed treatment	Disease activity
Anna	54	High	A	High
Calvin	52	High	A	Low
John	48	Low	B	Low
Peter	60	Low	B	High

Observational datasets: Rheumatoid arthritis

► Unobserved **counterfactual** outcomes

Outcomes under
alternative treatments

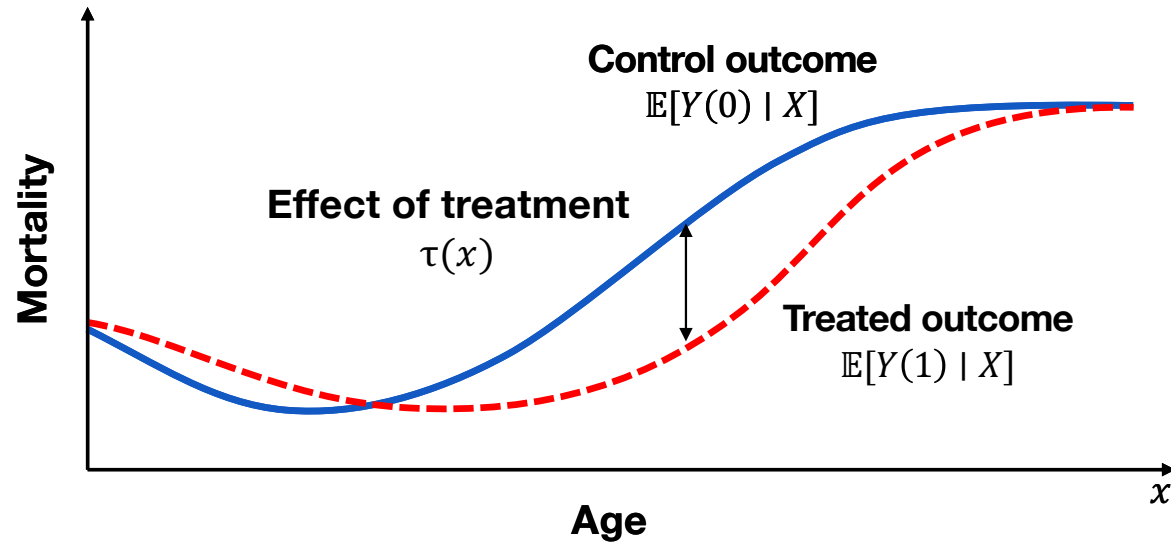
X

$Y(0)$

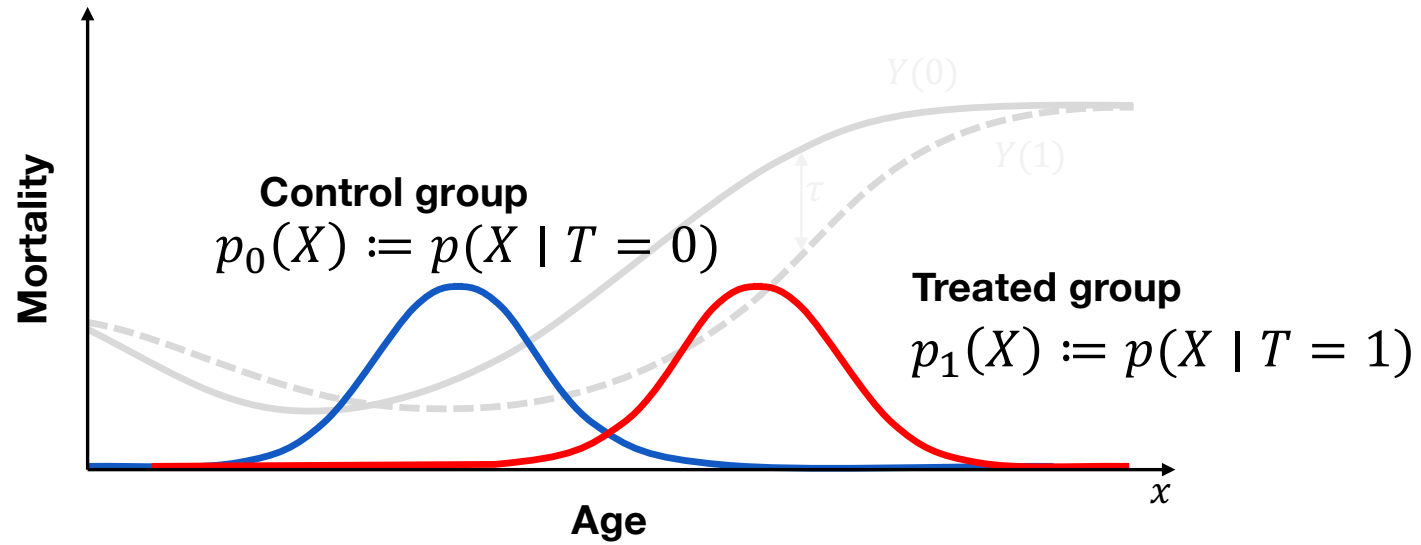
$Y(1)$

Patient	Age	Prior disease activity	Disease activity (A)	Disease activity (B)
Anna	54	High	High	?
Calvin	52	High	Low	?
John	48	Low	?	Low
Peter	60	Low	?	High

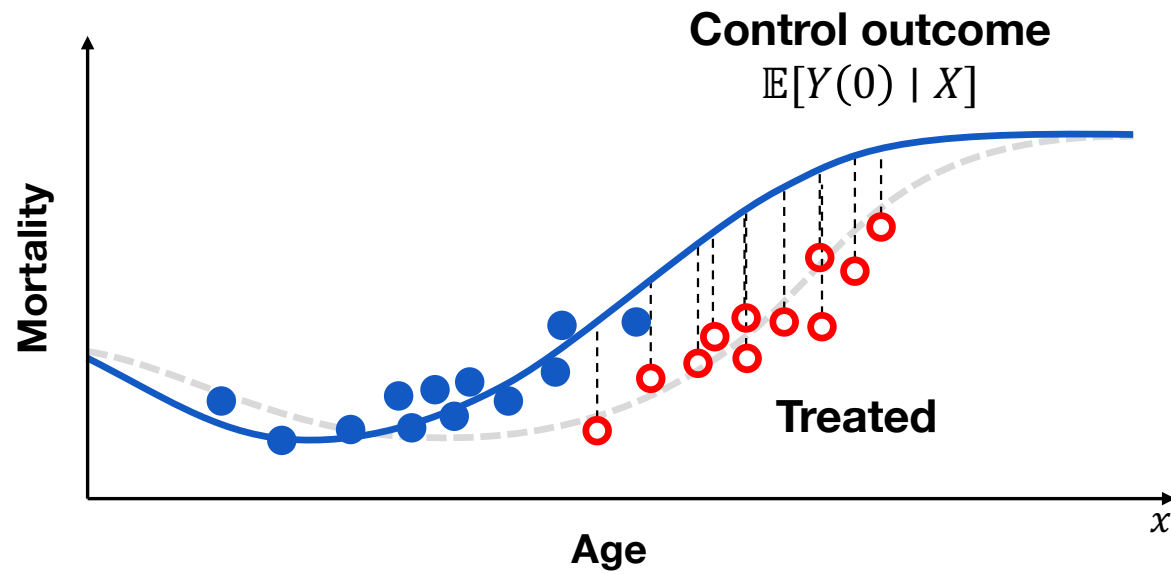
Estimating potential outcomes



Estimating potential outcomes



Estimating counterfactual for treated



Formalizing sufficient assumptions

1. Ignorability (no unmeasured confounders):

“Patients with similar X respond similarly”

$$\forall t : Y(t) \perp T \mid X$$

2. Overlap: “Similar patients with different treatments exist”

$$\forall t, x : p(T = t \mid X = x) > 0$$

3. SUTVA: “No patient-patient interference”

4. Consistency: “We observe $Y(t)$ for patients with $T = t$ ”

-
1. These are strong assumptions that don't always hold
 2. Even when they do, estimation is still challenging

Classical view

- Causal estimation often focused on **parameter estimation**

E.g., assume: $Y = \beta^T X + \theta T + \epsilon$, Goal: find θ !

Observed outcome

Treatment effect

Machine learning view

- Causal estimation often focused on **parameter estimation**

E.g., assume: $Y = \beta^\top X + \theta T + \epsilon$, Goal: find θ !

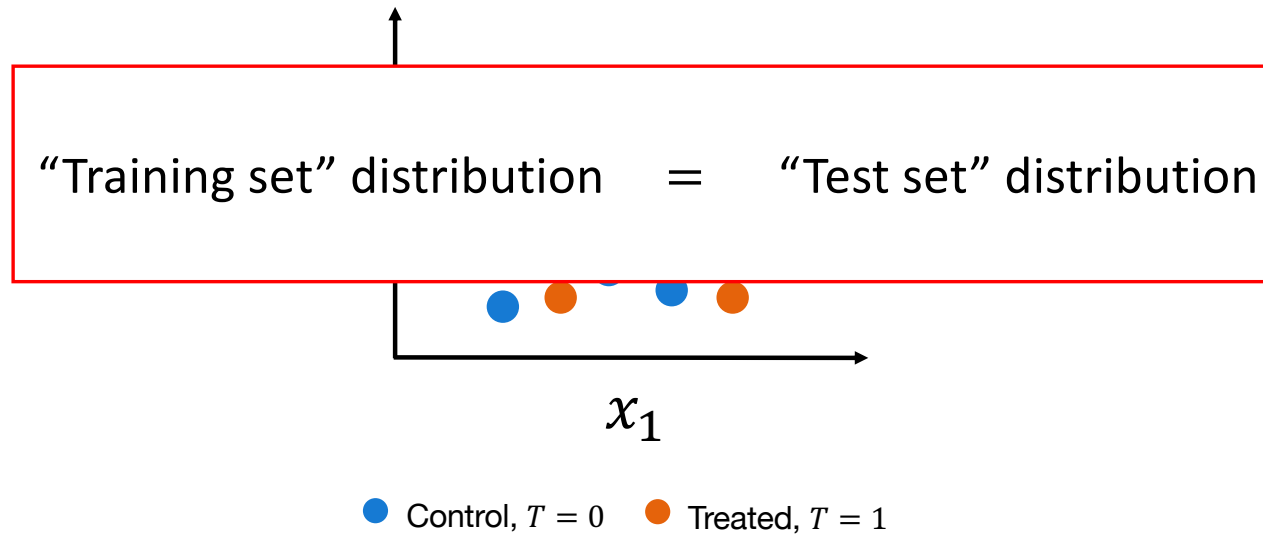
Observed outcome Treatment effect

- **ML view:** Find prediction of $\tau = Y(1) - Y(0)$ with small error $L(\hat{\tau}, \tau)$

$$\hat{\tau}^* = \arg \min_{\hat{\tau} \in \mathcal{T}} \mathbb{E}[L(\hat{\tau}, \tau)] = \arg \min_{\hat{\tau} \in \mathcal{T}} \mathbb{E}[(\hat{\tau}(X) - \tau)^2]$$

Easier: Randomized Controlled Trials (RCT)

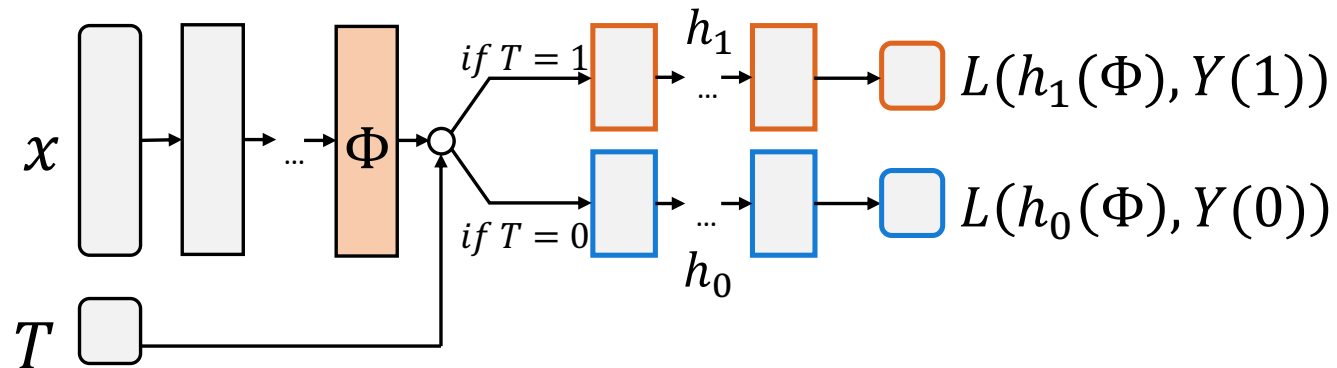
- ▶ Treatment is assigned uniformly at **random**: $p(T = 1 | X) = P(T = 1)$
- ▶ Here: every dot is a unit, color indicates **observed** treatment
- ▶ Predict outcome under **unobserved** treatment



Neural network architecture: TARNet

(Treatment-Agnostic Representation Network)

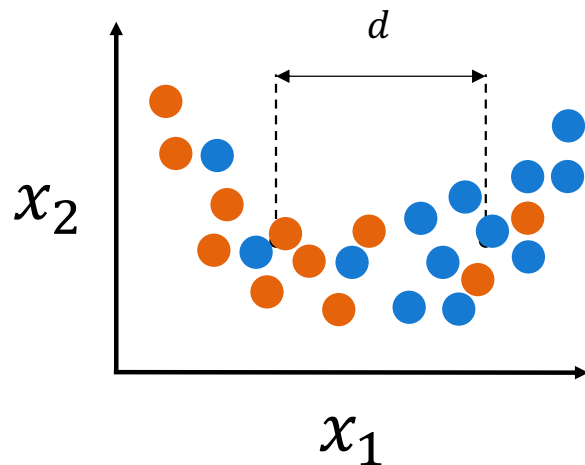
- ▶ In randomized control trials, there is no confounding – just do regression!
- ▶ New architecture for estimating counterfactuals and CATE



- ▶ One “head” per potential outcome – avoids washing away treatment
- ▶ Shared representation layers $\Phi(x)$ for sample efficiency

Observational studies: test \neq train

- ▶ Predict outcome under **unobserved** treatment
- ▶ Treatment is **not** assigned equally at random: $p(T = 1 | X) \neq P(T = 1)$
- ▶ There is a non-negligible difference between treatment group distributions



Example:

A difference in means

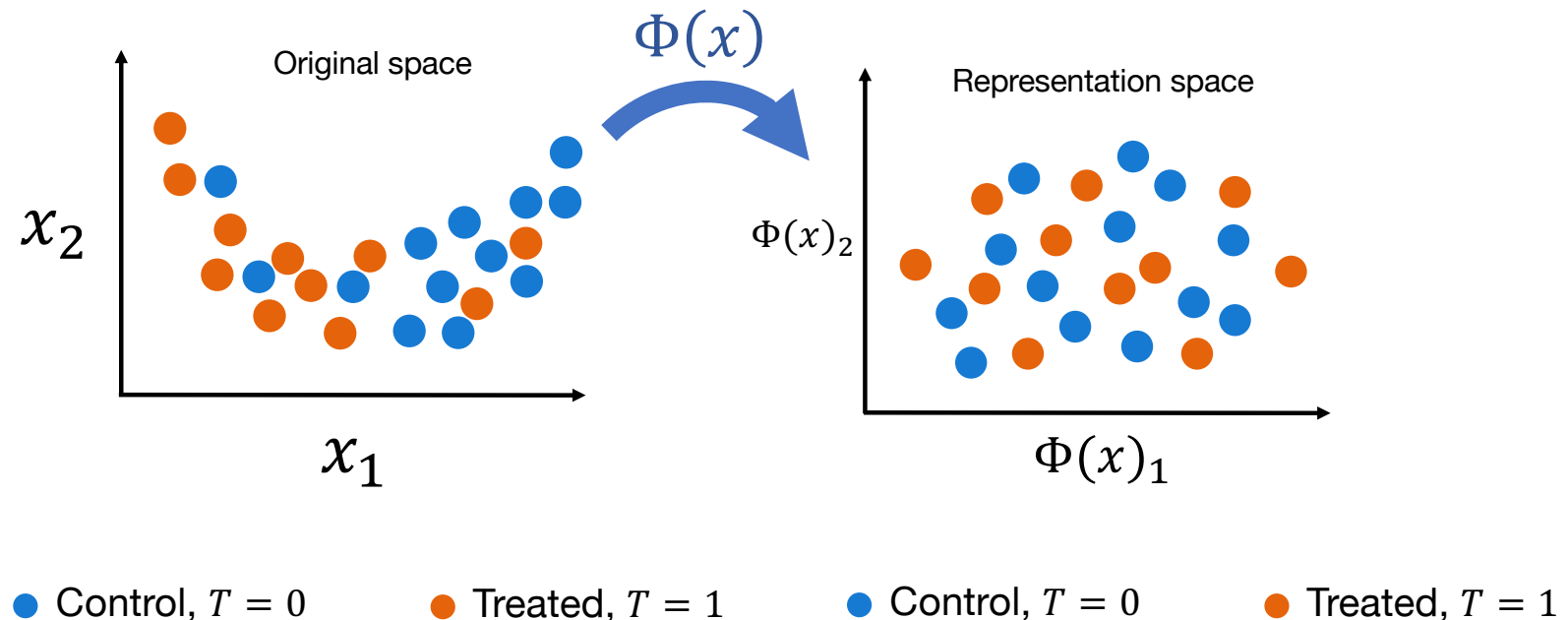
“Treated tend to be younger”

● Control, $T = 0$ ● Treated, $T = 1$

Representation learning

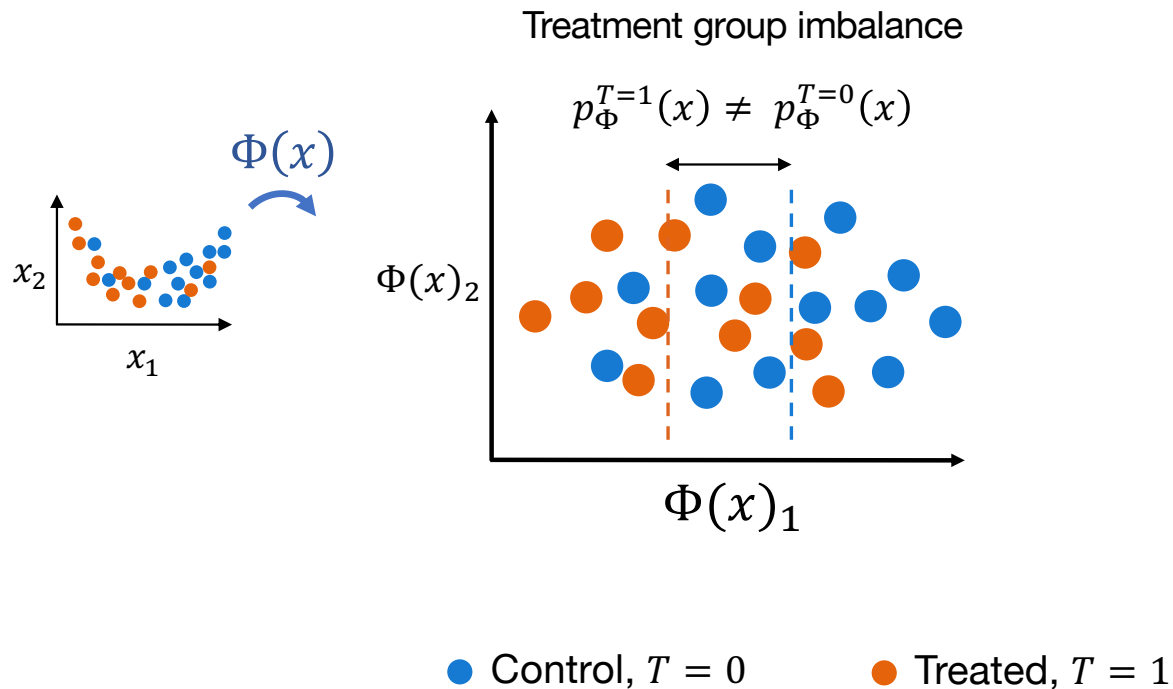
- ▶ Learn a representation Φ of the data that makes it more like an RCT
- ▶ A shared representation helps identify meaningful interactions
- ▶ **Penalize the distributional distance between treatment groups**

New type of bias-variance tradeoff



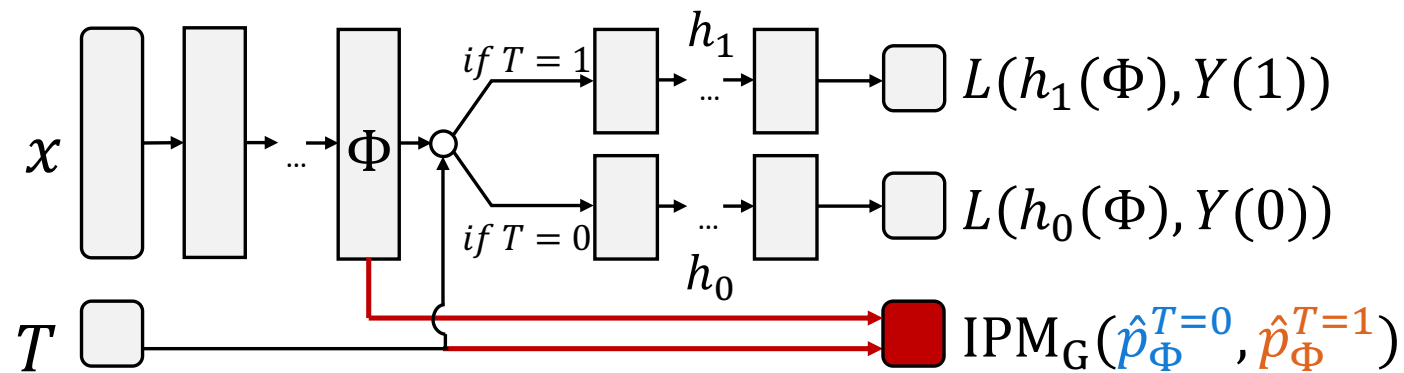
Imbalance in representation space

- We do not want treatment groups to be *identical*



Integral probability metric penalty

- ▶ **Regularizer** to improve counterfactual estimation
- ▶ **Penalize** treatment distributional distance **in representation space**

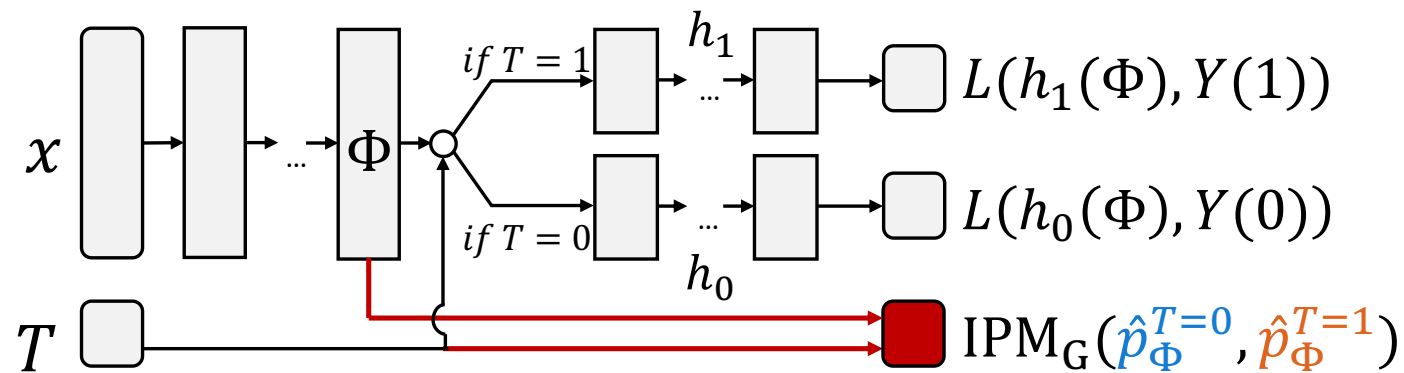


- ▶ Integral Probability Metrics (IPM) such as Wasserstein distance and MMD

With G a function family:
$$\text{IPM}_G(p_0, p_1) = \sup_{g \in G} \left| \int_S g(s)(p_0(s) - p_1(s)) ds \right|$$

Integral probability metric penalty

- ▶ **Regularizer** to improve counterfactual estimation
- ▶ **Penalize** treatment distributional distance **in representation space**



- ▶ Integral Probability Metrics (IPM) such as Wasserstein distance and MMD

With G a
function family:

$$\text{IPM}_G(p_0, p_1) = \sup_{g \in G} \left| \int_{\mathcal{S}} g(s) (p_0(s) - p_1(s)) ds \right|$$

Individual-level treatment effect generalization bound

- ▶ Precision in Estimation of Heterogeneous Effects¹:

- ▶ $\widehat{CATE}_{\Phi, h} = h(\Phi(x), 1) - h(\Phi(x), 0)$

$$\epsilon_{CATE}(\phi, h) = \int_x \left(\widehat{CATE}_{\Phi, h} - CATE(x) \right)^2 p(x) dx$$

- ▶ Factual per-treatment group prediction error

$$\epsilon_F^{T=0} = \int_x \left(\hat{Y}(0) - Y(0) \right)^2 p^{t=0}(x) dx$$

$$\epsilon_F^{T=1} = \int_x \left(\hat{Y}(1) - Y(1) \right)^2 p^{t=1}(x) dx$$

▶ **Theorem 1:**

$$\underbrace{\epsilon_{CATE}(\phi, h)}_{\text{Effect error}} \leq 2 \left(\underbrace{\epsilon_F^{T=0}(\Phi, h) + \epsilon_F^{T=1}(\Phi, h)}_{\text{Prediction error}} + B_{\Phi} \underbrace{\text{IPM}_G(p_{\Phi}^{T=1}, p_{\Phi}^{T=0})}_{\text{Treatment group distance}} \right)$$

¹Hill, *Journal of Computational and Graphical Statistics* 2011

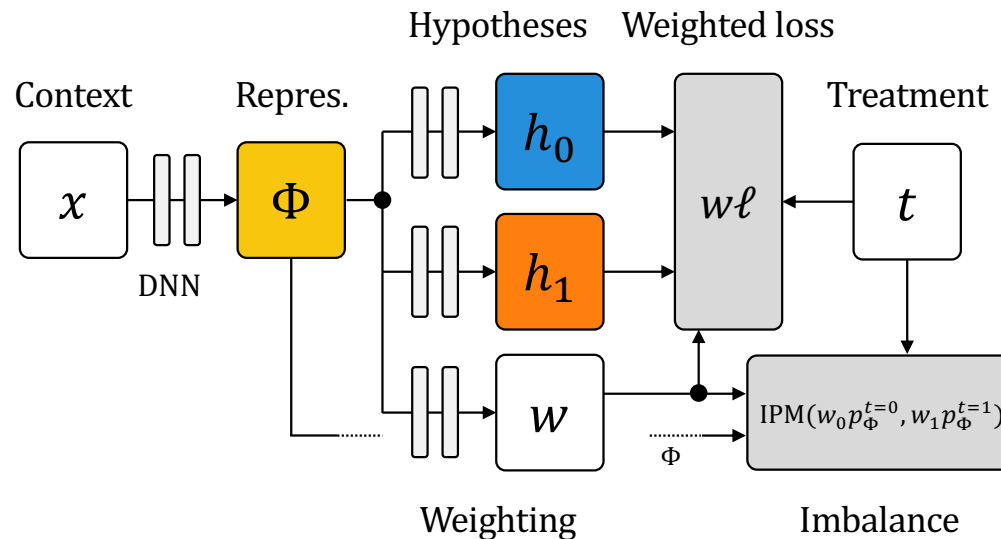
► **Theorem 1:**

$$\epsilon_{\text{CATE}} \leq 2 \left(\underbrace{\epsilon_F^{T=0}(\Phi, h)}_{\text{Effect error}} + \underbrace{\epsilon_F^{T=1}(\Phi, h)}_{\text{Prediction error}} + \underbrace{B_\Phi \text{IPM}_G(p_\Phi^{T=1}, p_\Phi^{T=0})}_{\text{Treatment group distance}} \right)$$

- Problem with Theorem 1:
Too loose when we have overlap + infinite samples
- We should be able to achieve the prediction error itself on either group

Trading off accuracy for balance

- ▶ Our full architecture learns a representation $\Phi(x)$, a re-weighting $w_t(x)$ and hypotheses $h_t(\Phi)$ to trade-off between the re-weighted loss $w\ell$ and imbalance between re-weighted representations



Individual-treatment effect generalization bound

- ▶ **Theorem 2***: (Representation learning)

$$\underbrace{\epsilon_{\text{CATE}}}_{\text{Effect risk}} \leq 2 \sum_{t \in \{0,1\}} \left(\underbrace{\epsilon_t^{w_t}(\Phi, h)}_{\text{Re-weighted factual loss}} + \underbrace{B_\Phi \text{IPM}_G(p_\Phi^{1-t}(x), w_t p_\Phi^t(x))}_{\text{Imbalance of re-weighted representations}} \right)$$

- ▶ Letting $\Phi(x) = x$, and $w_t(x)$ be inverse propensity weights, we recover classic result
- ▶ Minimizing a weighted loss and IPM converge to the representation and hypothesis that minimize CATE error

*Extension to finite samples available

Evaluating Individual Treatment Effect (CATE) Estimates

- ▶ **No ground truth**, similar to off-policy evaluation in reinforcement learning

Evaluating Individual Treatment Effect (CATE) Estimates

- ▶ **No ground truth**, similar to off-policy evaluation in reinforcement learning
- ▶ Requires either:
 - ▶ Knowledge of the true outcome (synthetic)
 - ▶ Knowledge of treatment assignment policy (e.g. a randomized controlled trial)

Evaluating Individual Treatment Effect (CATE) Estimates

- ▶ **No ground truth**, similar to off-policy evaluation in reinforcement learning
- ▶ Requires either:
 - ▶ Knowledge of the true outcome (synthetic)
 - ▶ Knowledge of treatment assignment policy (e.g. a randomized controlled trial)
- ▶ Our framework has proven effective in both settings

IHDP Benchmark¹

- ▶ The Infant Health and Development Program (IHDP)
 - ▶ Studied the effects of home visits and other interventions
- ▶ Real covariates and treatment, synthesized outcome
- ▶ **Overlap** is not satisfied (by design)
- ▶ Used to evaluate MSE in CATE prediction

¹Hill, *JCGS*, 2011

Empirical results

- ▶ BART, Bayesian Additive Regression Trees, are state-of-the-art baselines
- ▶ Standard neural networks competitive
- ▶ Shared representation learning with ERM halves the MSE on IHDP²
- ▶ Minimizing upper bounds on risk, including $d_{\mathcal{H}}$ further reduces the MSE

Method	CATE MSE
BART ¹	2.3 ± 0.1
Neural net	2.0 ± 0.0
Shared rep. ²	1.0 ± 0.0
Shared rep. + invariance ²	0.8 ± 0.0
Shared rep. + invariance + weighting ³	0.7 ± 0.0

¹Hill, *JCGS*, 2011, ²S., Johansson, Sontag. *ICML*, 2017, ³Johansson, Kallus, S., Sontag. *arXiv*, 2018

Intermediate conclusions

- ▶ ML is well understood when test data \approx training data
- ▶ Learning individualized policies from observational data requires going beyond test \approx train
- ▶ Fewer/worse guarantees when assumptions are violated

Outline

- **ML for causal inference**
- Causal inference for ML
 - Off-policy evaluation in a partially observable Markov decision process
 - Robust learning for unsupervised covariate shift

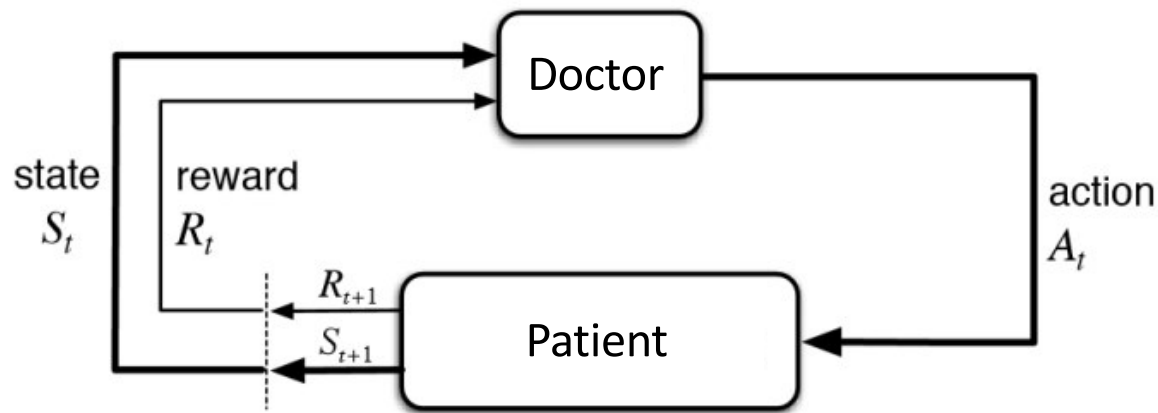
Outline

- ML for causal inference
- **Causal inference for ML**
 - **Off-policy evaluation in a partially observable Markov decision process**
 - Robust learn

“Off-Policy Evaluation in Partially
Observable Environments”,
Tennenholtz, Mannor, S
AAAI 2020

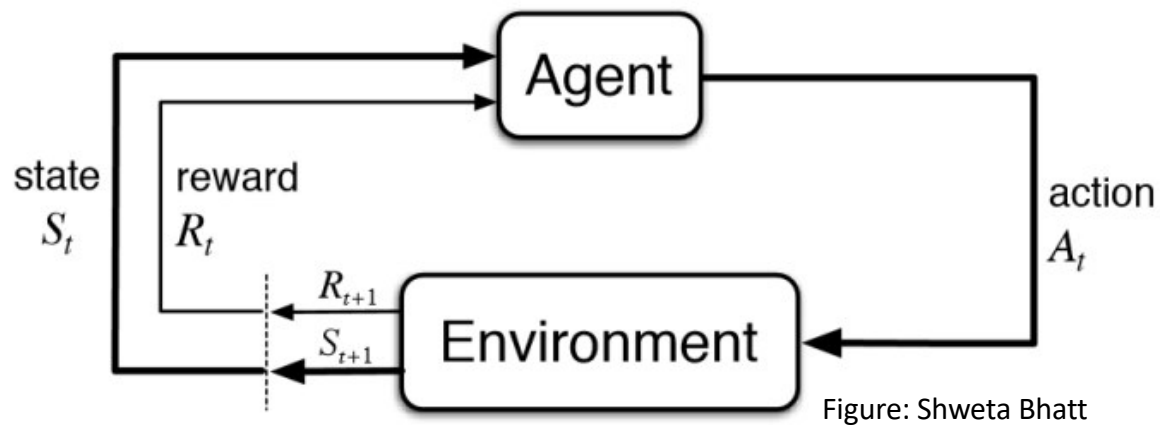
Healthcare with time-varying decisions

- Physicians make ongoing decisions: treat, see change in patients state, modify treatment, and so on



Healthcare with time-varying decisions

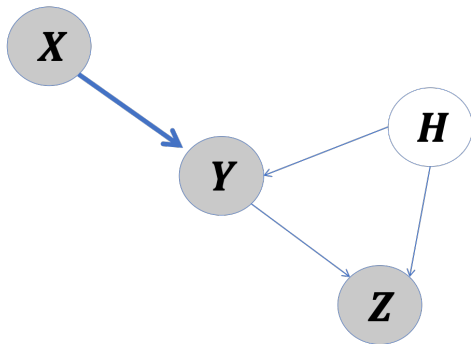
- Maps very well to reinforcement learning paradigm



Reinforcement learning (RL) and causal inference

From causal inference perspective

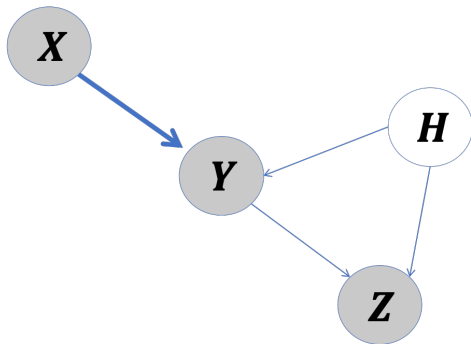
- RL usually assumes we can intervene directly
- → mostly about how to experiment optimally in a dynamic environment



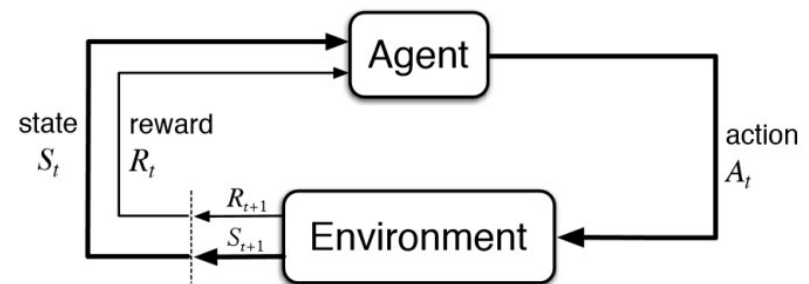
Reinforcement learning (RL) and causal inference

From causal inference perspective

- RL usually assumes we can intervene directly
- → mostly about how to experiment optimally in a dynamic environment



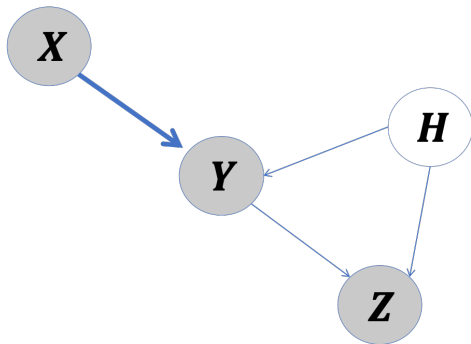
From RL perspective



Reinforcement learning (RL) and causal inference

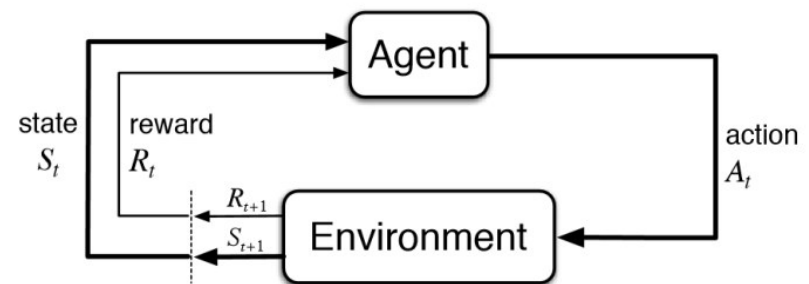
From causal inference perspective

- RL usually assumes we can intervene directly
- → mostly about how to experiment optimally in a dynamic environment



From RL perspective

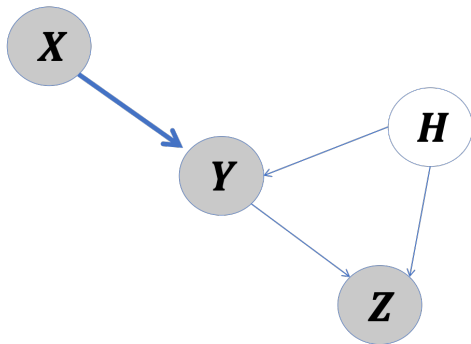
- Causal inference usually deals with cases we cannot intervene directly



Reinforcement learning (RL) and causal inference

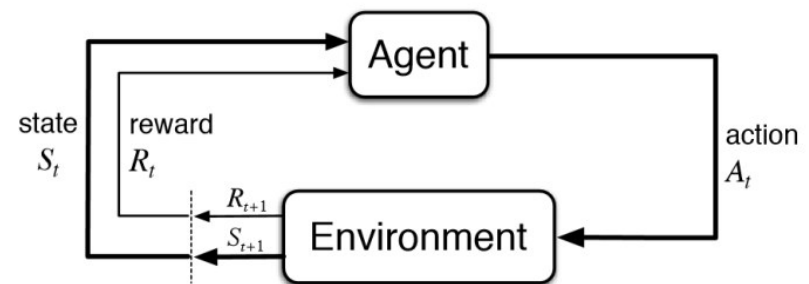
From causal inference perspective

- RL usually assumes we can intervene directly
- → mostly about how to experiment optimally in a dynamic environment



From RL perspective

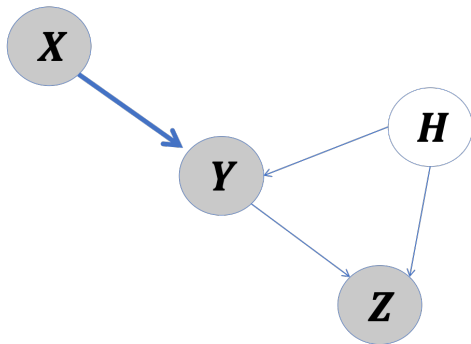
- Causal inference usually deals with cases we cannot intervene directly
- Causal inference usually focuses on single point-in-time actions



Reinforcement learning (RL) and causal inference

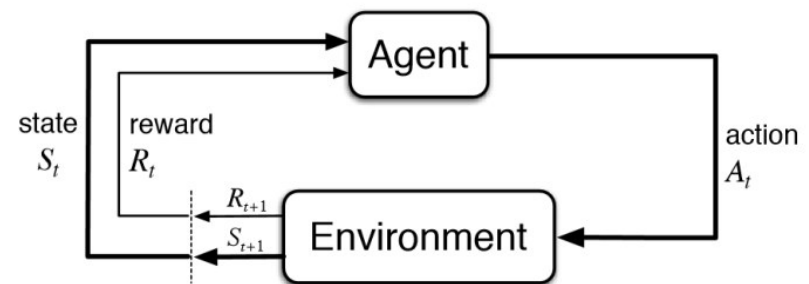
From causal inference perspective

- RL usually assumes we can intervene directly
- → mostly about how to experiment optimally in a dynamic environment



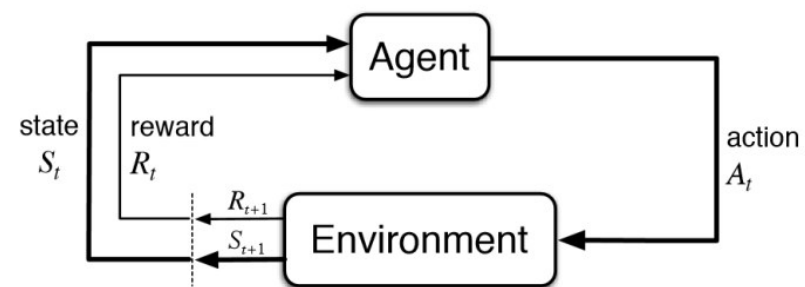
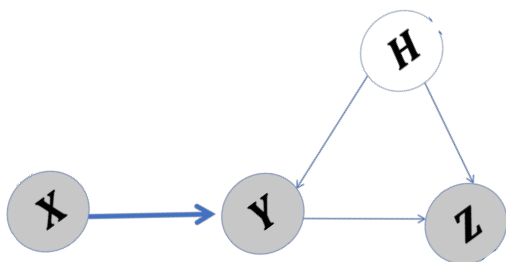
From RL perspective

- Causal inference usually deals with cases we cannot intervene directly
- Causal inference usually focuses on single point-in-time actions
- → mostly about off-policy evaluation of a simple policy such as “treat everyone”



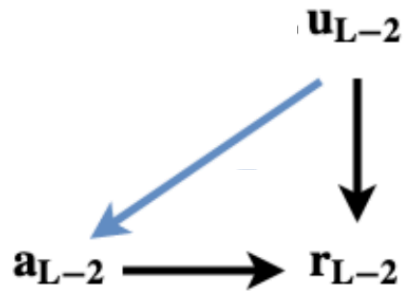
A meeting point of RL and causal inference

- When performing off-policy evaluation of data from
 - i. dynamic environment with ongoing actions
 - ii. while we possibly do not have access to the same data as the agent
- Example: learning from records of physicians treating patients in an intensive care unit (ICU)
- Mistakes were made: applying RL to observational intensive care unit data without considering hidden confounders or overlap (common support / positivity)
(see “Guidelines for Reinforcement Learning in Healthcare” Gottesman et al. 2019)
- In RL nomenclature, hidden confounding can be described by a Partially Observable Markov Decision Process (POMDP)



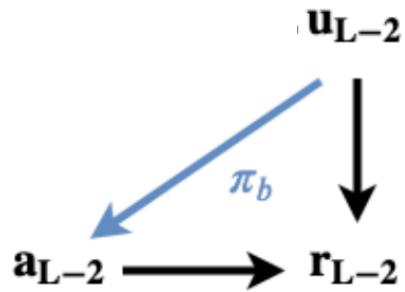
Partially Observable Markov Decision Process (POMDP): some formalism

- 7-tuple $(\mathcal{U}, \mathcal{A}, \mathcal{Z}, \mathcal{P}, \mathcal{O}, r, \gamma)$
- \mathcal{U} - finite state space
- \mathcal{A} - finite action space
- \mathcal{Z} - finite observation space
- $\mathcal{P} : \mathcal{U} \times \mathcal{U} \times \mathcal{A} \mapsto [0, 1]$ - state transition kernel
- $\mathcal{O} : \mathcal{U} \times \mathcal{Z} \mapsto [0, 1]$ - observation function
- $r : \mathcal{U} \times \mathcal{A} \mapsto \mathbb{R}$ - reward function
- $\gamma \in (0, 1)$ - discount factor



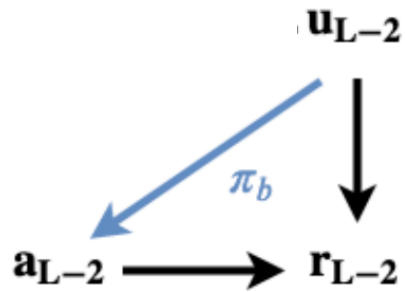
POMDP causal graph

	Causal name	RL name
u_t	confounder (possibly "hidden")	state (possibly "unobserved")
a_t	action, treatment	action
r_t	outcome	reward



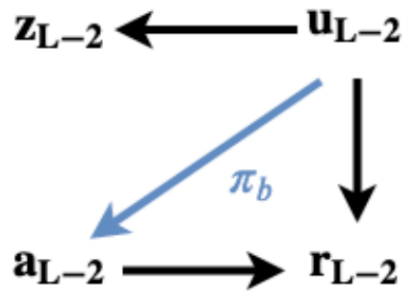
POMDP causal graph

	Causal name	RL name
u_t	confounder (possibly "hidden")	state (possibly "unobserved")
a_t	action, treatment	action
r_t	outcome	reward
π_b	treatment assignment process	behavioral policy



POMDP causal graph

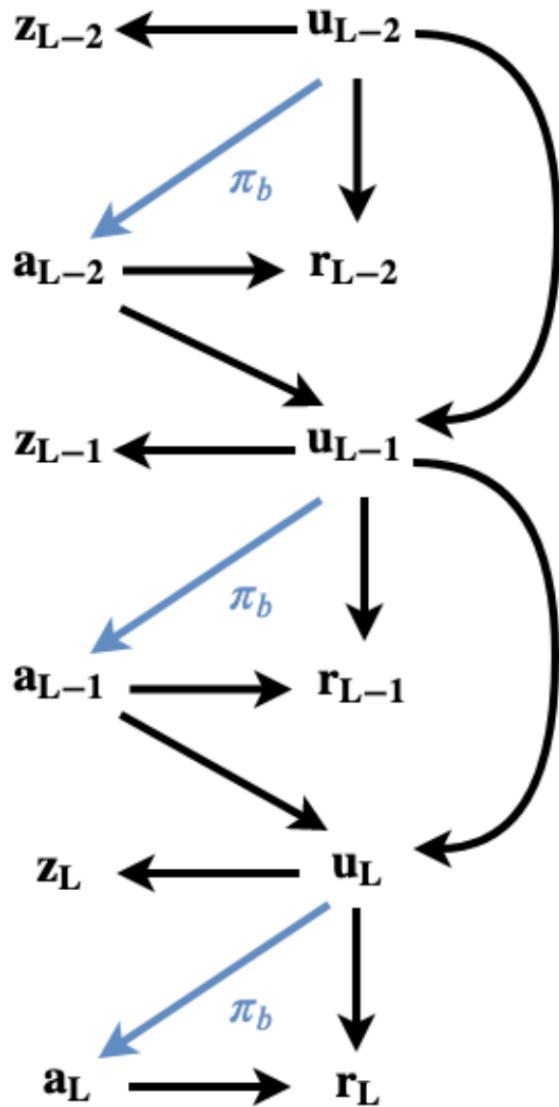
	Causal name	RL name	Example
\mathbf{u}_t	confounder (possibly "hidden")	state (possibly "unobserved")	information available to the doctor
\mathbf{a}_t	action, treatment	action	medications, procedures...
\mathbf{r}_t	outcome	reward	mortality
π_b	treatment assignment process	behavioral policy	the way doctors treat patients



POMDP causal graph

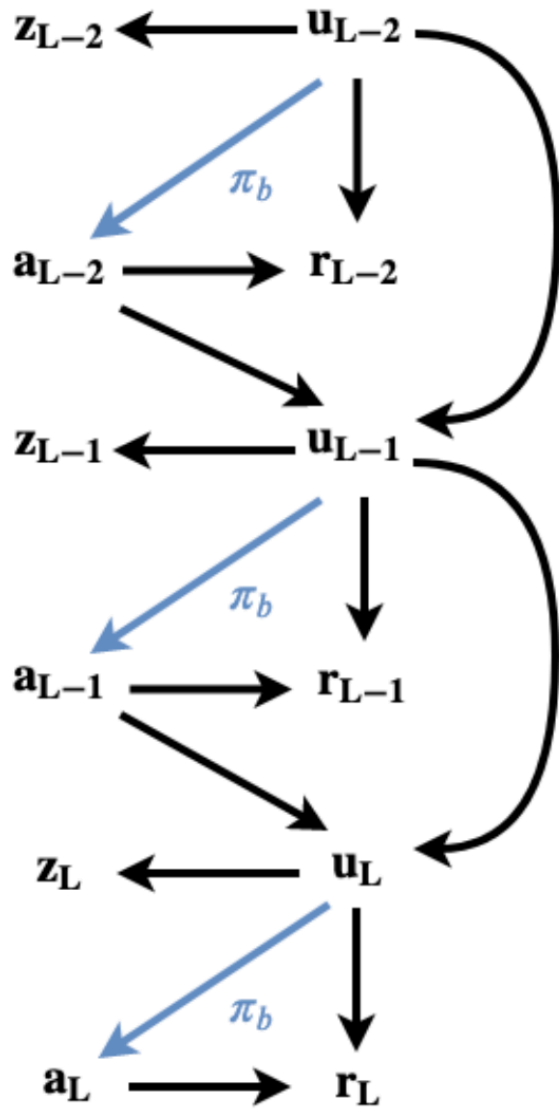
	Causal name	RL name	Example
u_t	confounder (possibly "hidden")	state (possibly "unobserved")	information available to the doctor
a_t	action, treatment	action	medications, procedures...
r_t	outcome	reward	mortality
π_b	treatment assignment process	behavioral policy	the way doctors treat patients
z_t	proxy variable	observation	electronic health record

POMDP causal graph

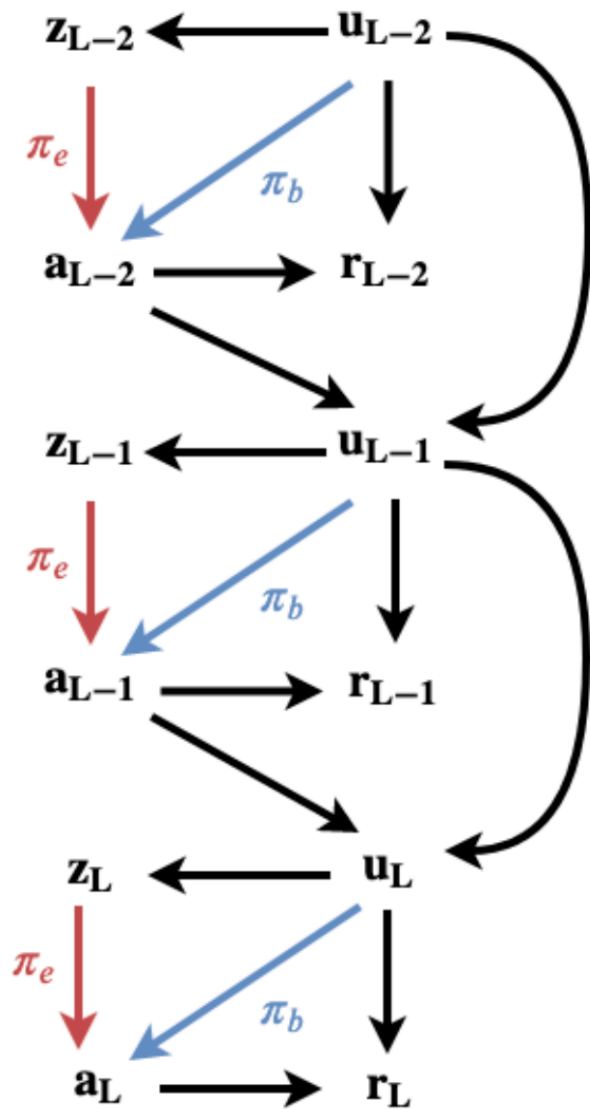


	Causal name	RL name	Example
u_t	confounder (possibly "hidden")	state (possibly "unobserved")	information available to the doctor
a_t	action, treatment	action	medications, procedures...
r_t	outcome	reward	mortality
π_b	treatment assignment process	behavioral policy	the way doctors treat patients
z_t	proxy variable	observation	electronic health record

POMDP causal graph



- Observe data from π_b , with u_t unobserved

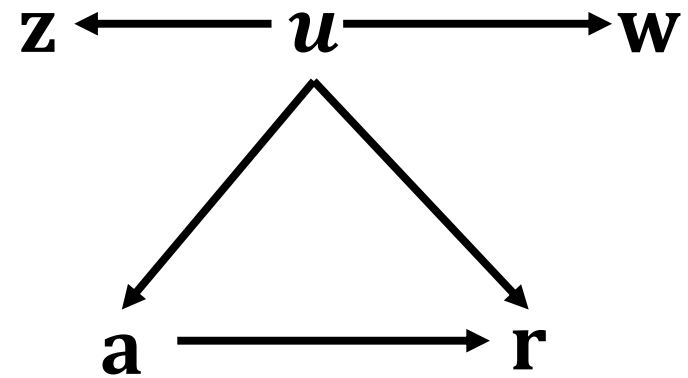


Our goal: evaluate a new policy π_e given data from π_b

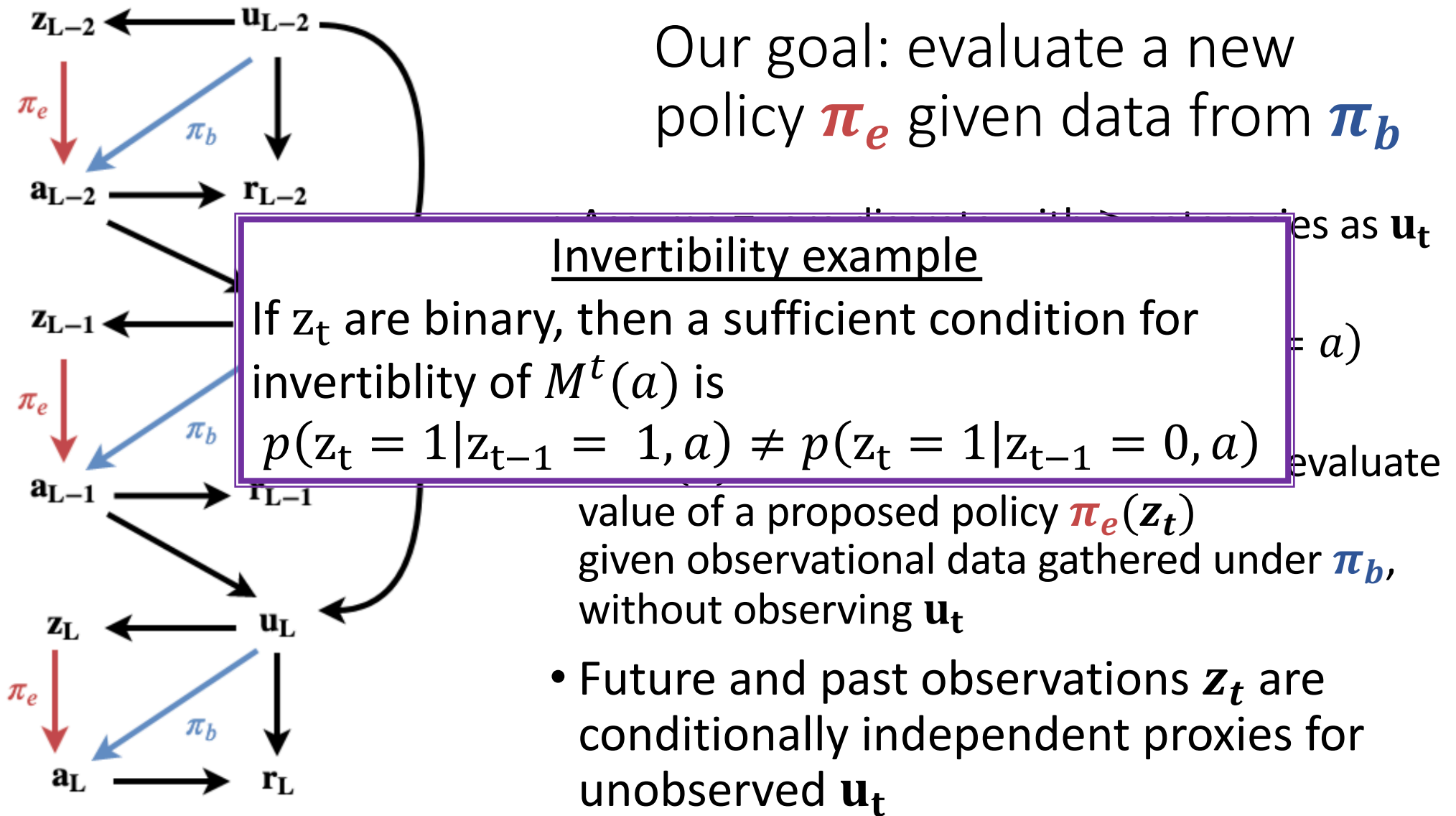
- Observing data from π_b , with \mathbf{u}_t unobserved evaluate a proposed policy $\pi_e(\mathbf{z}_t)$ in terms of policy value (discounted over a finite horizon)
- Without further assumptions: IMPOSSIBLE
- Example: ICU doctors treating sicker patients more aggressively
- Impossible even when conditioning on entire observable history $\{(\mathbf{z}_1, \mathbf{a}_1, \mathbf{r}_1), \dots, (\mathbf{z}_T, \mathbf{a}_T, \mathbf{r}_T)\}$
- Due to hidden confounding by \mathbf{u}_t
- But much harder: confounder \leftrightarrow action dynamics

Proxies and negative controls

- Miao, Geng, & Tchetgen Tchetgen.
“Identifying causal effects with proxy variables of an unmeasured confounder.”
Biometrika (2018)
- Only \mathbf{u} is unobserved
- Goal: identify the causal effect of \mathbf{a} on \mathbf{r}
- $\mathbf{z} \perp\!\!\!\perp \mathbf{w} \mid \mathbf{u}$
- In general: impossible
- New identification condition:
matrices $M_{ij}(a) = p(\mathbf{w} = i \mid \mathbf{z} = j, \mathbf{a} = a)$
are invertible for all a
- Requires \mathbf{w} and \mathbf{z} to be discrete with
as many categories as discrete \mathbf{u}

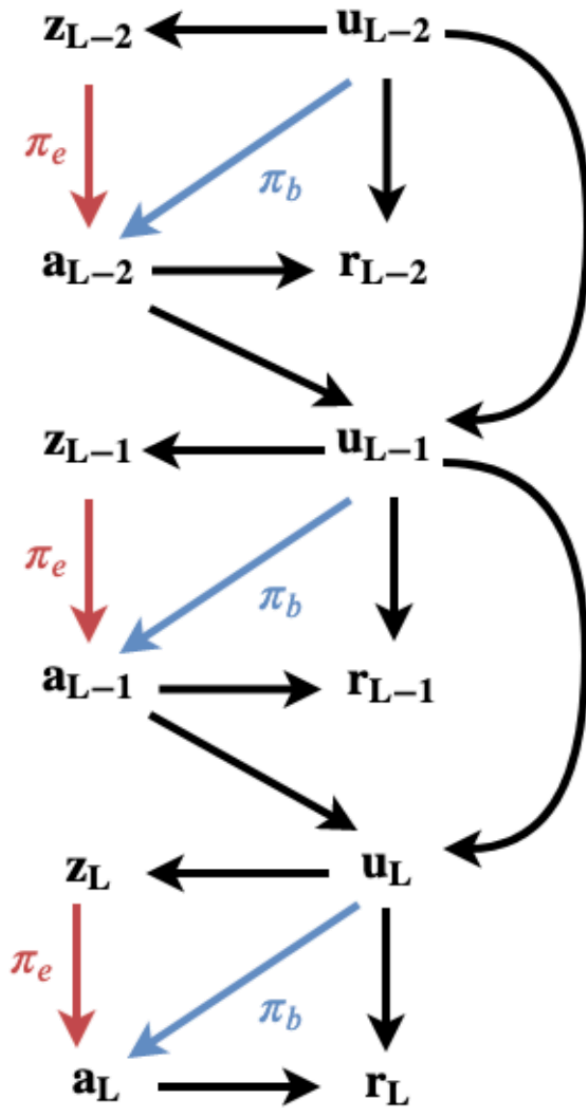


Our goal: evaluate a new policy π_e given data from π_b



Assumptions

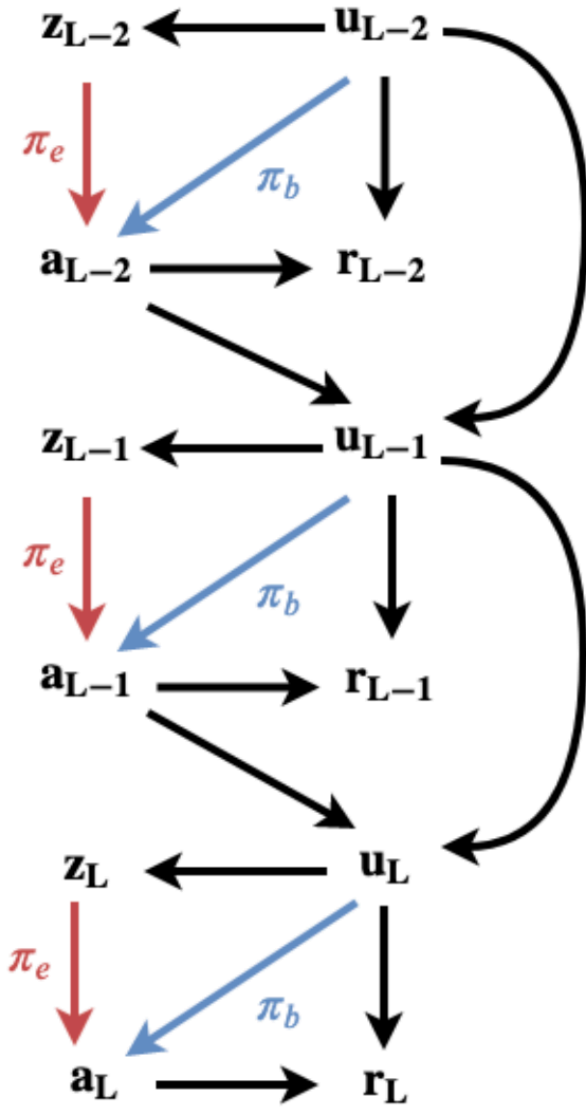
1. Assume \mathbf{z}_t are discrete with \geq categories as \mathbf{u}_t
2. Matrices $M_{ij}^t(a) = p^{\pi_b}(\mathbf{z}_t = i | \mathbf{z}_{t-1} = j, \mathbf{a}_t = a)$ are invertible for all a and t



- Allow off-policy evaluation for class of POMDPs
- No need to measure or even know what is \mathbf{u}_t
- As usual in Causal Inference, some of the assumptions are unverifiable from data

Assumptions

1. Assume \mathbf{z}_t are discrete with \geq categories as \mathbf{u}_t
2. Matrices $M_{ij}^t(a) = p^{\pi_b}(\mathbf{z}_t = i | \mathbf{z}_{t-1} = j, \mathbf{a}_t = a)$ are invertible for all a and t



- Observed sequence $\tau = (z_0, a_0, \dots, z_T, a_T) \in \mathcal{T}_T$
- $N_{ij}^t(a) = p^{\pi_b}(\mathbf{z}_t = i, \mathbf{z}_{t-1} = z_{t-1} | \mathbf{z}_{t-2} = j, \mathbf{a}_{t-1} = a)$
- $W^t(\tau) = M^t(a_t)^{-1} N^t(a_{t-1})$
- $Q_i^0(\tau) = \sum_j M^0(a_0)_{ij}^{-1} p^{\pi_b}(\mathbf{z}_0 = j)$
- $\Omega(\tau) = (\prod_{t=0}^T W^t(\tau)) \cdot Q^0(\tau)$
- $\Lambda_e(\tau) = \prod_{t=0}^T \pi_e(a_t | z_0, a_0, \dots, z_{t-1}, a_{t-1}, z_t)$
- Then:

$$p^{\pi_e}(r_t) = \sum_{\tau \in \mathcal{T}_T} \Lambda_e(\tau) p^{\pi_b}(r_t, z_t | a_t, z_{t-1}) \Omega(\tau)$$

Off-policy POMDP evaluation

- The above evaluation requires estimating the inverses of many conditional probability tables
- Scales poorly statistically
- We introduce another causal model called **decoupled-POMDP**
 - Similar causal graph
 - Significantly reduces the dimensions and improves condition number of the estimated inverse matrices

Off-policy POMDP evaluation

- The above evaluation requires estimating the inverses of many conditional probability tables
- Scales poorly statistically
- We introduce another causal model called **decoupled-POMDP**
 - Similar causal graph
 - Significantly reduces the dimensions and improves condition number of the estimated inverse matrices
- Current challenge: scaling to realistic health data

Outline

- ML for causal inference
- **Causal inference for ML**
 - **Off-policy evaluation in a partially observable Markov decision process**
 - Robust learning for unsupervised covariate shift

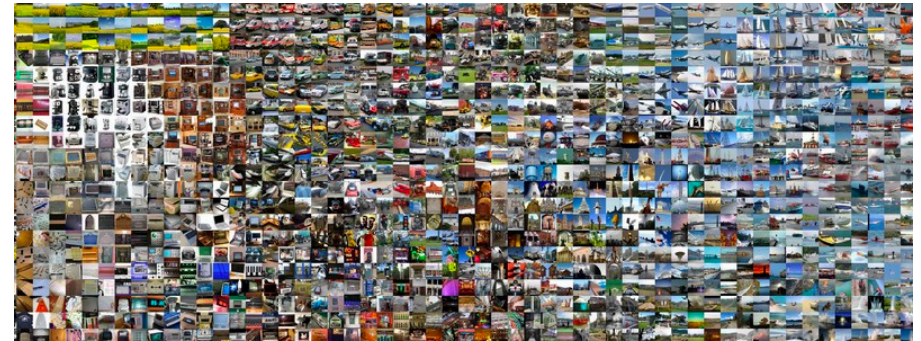
Outline

- ML for causal inference
- **Causal inference for ML**
 - Off-policy evaluation in a partially observable Markov decision process
 - **Robust learning for unsupervised covariate shift**

“Robust learning with the Hilbert-Schmidt independence criterion”,
Greenfeld & S
arXiv:1910.00270

Classic non-causal tasks in machine learning: many success stories

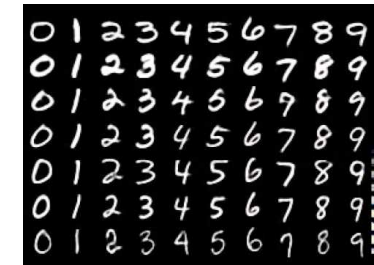
- Classification
 - ImageNet
 - MNIST
 - TIMIT (sound)
 - Sentiment analysis



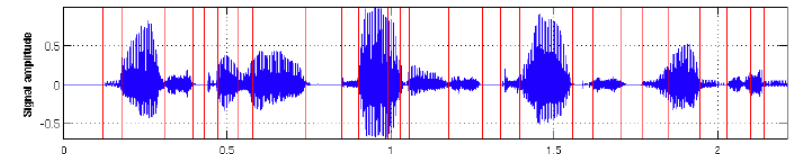
1 really enjoyed using the 2 Canon Ixus in Madrid on March 4. The
3 Panasonic Lumix 4 is a bit disappointing, but the 5 Canon camera is
6 not bad at all. All I want when taking photos is point it and then just press the
7 button. For only 200 dollars, a 8 really fair price, this 9 camera is 0 perfect
for me. Besides, I have had a 1 good customer service 2 experience.
3 John Faraday was 4 very nice!

LEGEND color key

- Sentiment topic
- Positive sentiment text
- Negative sentiment text
- 1 Text and topic link

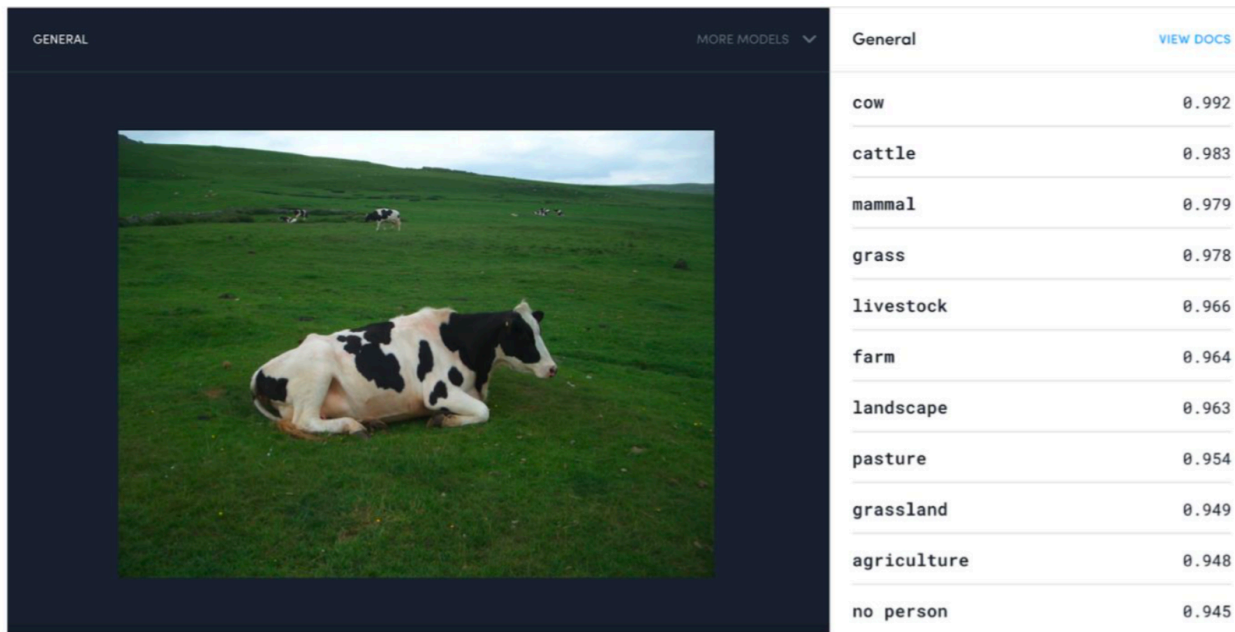


- Prediction
 - Which patients will die?
 - Which users will click?
 - (under current practice)



Failures of ML Classification models

Easy cows



GENERAL MORE MODELS


General [VIEW DOCS](#)

cow	0.992
cattle	0.983
mammal	0.979
grass	0.978
livestock	0.966
farm	0.964
landscape	0.963
pasture	0.954
grassland	0.949
agriculture	0.948
no person	0.945

(From Pietro Perona)

Failures of ML Classification models

Difficult cows



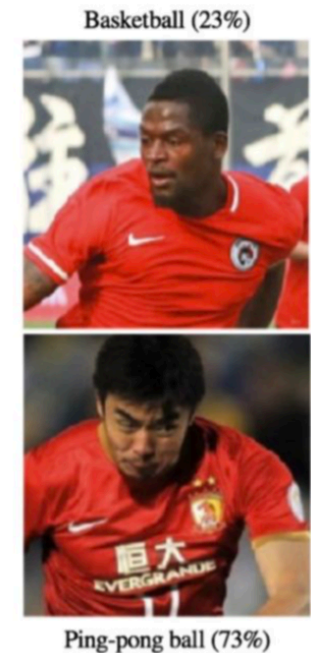
GENERAL MORE MODELS

General [VIEW DOCS](#)

no person	0.991
beach	0.990
water	0.985
sand	0.981
sea	0.980
travel	0.978
seashore	0.972
summer	0.954
sky	0.946
outdoors	0.944
ocean	0.936

(From Pietro Perona)

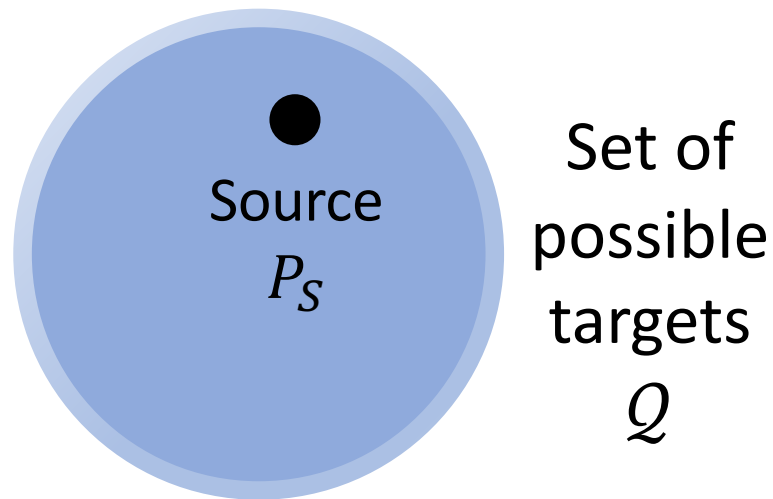
test set \neq train set,
but we know humans succeed here



(Stock and Cisse, 2017)

How to learn models which are **robust** to a-priori unknown changes in test distribution?

- *Source* distribution $P_S(X, Y)$
- Learn model that works well on unknown *Target* distributions $P'(X, Y) \in Q$



How to learn models which are **robust** to a-priori unknown changes in test distribution?

- Source distribution $P_S(X, Y)$
- Learn model that works well on all target distributions $P'(X, Y) \in Q$

- What is Q ?

- We assume **Covariate Shift**:

For all $P'(X, Y) \in Q$,

$$P'(Y|X) = P_S(Y|X)$$

- Further restrictions on Q to follow
- Covariate shift is easy if learning $P_S(Y|X)$ is easy
 - Focus on tasks where it's hard

Unsupervised covariate shift

- A model that works well even when the underlying distribution of instances changes
- Works as long as $P(Y|X)$ is stable
- When does this happen?

Causal mechanisms are stable



Learning with an independence criterion

- X causes Y , structural causal model:

$$Y = f^*(X) + \epsilon, \quad \epsilon \perp\!\!\!\perp X$$

- $f^*(x)$ is the mechanism tying X to Y
- ϵ is independent *additive* noise
- Therefore, $Y - f^*(X) \perp\!\!\!\perp X$
- Mooij, Janzing, Peters & Schölkopf (2009):
Learn structure of causal models by learning functions f such that $Y - f(X)$ is approximately independent of X
- Need a non-parametric measure of independence
- **Hilbert-Schmidt independence criterion, HSIC**

Hilbert-Schmidt independence criterion: HSIC

- Let X, Y be two metric spaces with a joint distribution $P(X, Y)$
- \mathcal{G}_X and \mathcal{G}_Y are reproducing kernel Hilbert spaces on X and Y induced by kernels $K(\cdot, \cdot)$ and $L(\cdot, \cdot)$ respectively
- $HSIC(X, Y)$ measures the degree of dependence between X and Y
- Empirical version: Sample $(x_1, y_1), \dots, (x_n, y_n)$
Denote (some abuse of notation)
 K the $n \times n$ kernel matrix on X , L is $n \times n$ kernel matrix on Y
- $\widehat{HSIC}(X, Y; \mathcal{G}_X, \mathcal{G}_Y) = \frac{1}{(n-1)^2} \text{tr}(KHLH)$

H is a centering matrix, $H_{ij} = \delta_{ij} - \frac{1}{n}$

Learning with HSIC

- Hypothesis class \mathcal{H}
- Classic learning for loss ℓ , e.g. squared loss:

$$\min_{h \in \mathcal{H}} \mathbb{E}[\ell(Y, h(X))]$$

- Learning with HSIC (Mooij et al., 2009):

$$\min_{h \in \mathcal{H}} HSIC(X, Y - h(X); \mathcal{G}_X, \mathcal{G}_Y)$$

Learning with HSIC

- Learning with HSIC (Mooij et al., 2009):

$$\min_{h \in \mathcal{H}} HSIC(X, Y - h(X); \mathcal{G}_X, \mathcal{G}_Y)$$

- Recall: $Y - f^*(X) \perp\!\!\!\perp X$
- If objective equals 0 then $h^*(X) = f^*(x) + b$ for some constant b
- Can learn up to an additive bias term

Learning with HSIC

- Learning with HSIC (Mooij et al., 2009):

$$\min_{h \in \mathcal{H}} HSIC(X, Y - h(X); \mathcal{G}_X, \mathcal{G}_Y)$$

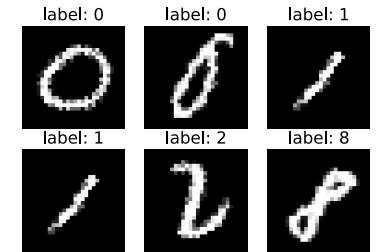
- Differentiable with respect to $h(X)$
- We optimize with SGD using mini-batches to approximate HSIC

Theoretical results

- Learnability: minimizing HSIC-loss over a sample leads to generalization
- Robustness: minimizing HSIC-loss leads to tightly-bounded error in unsupervised covariate shift
 - If density ratio $\frac{P_{target}(x)}{P_{source}(x)}$ is “nice” in the sense of low RKHS norm.

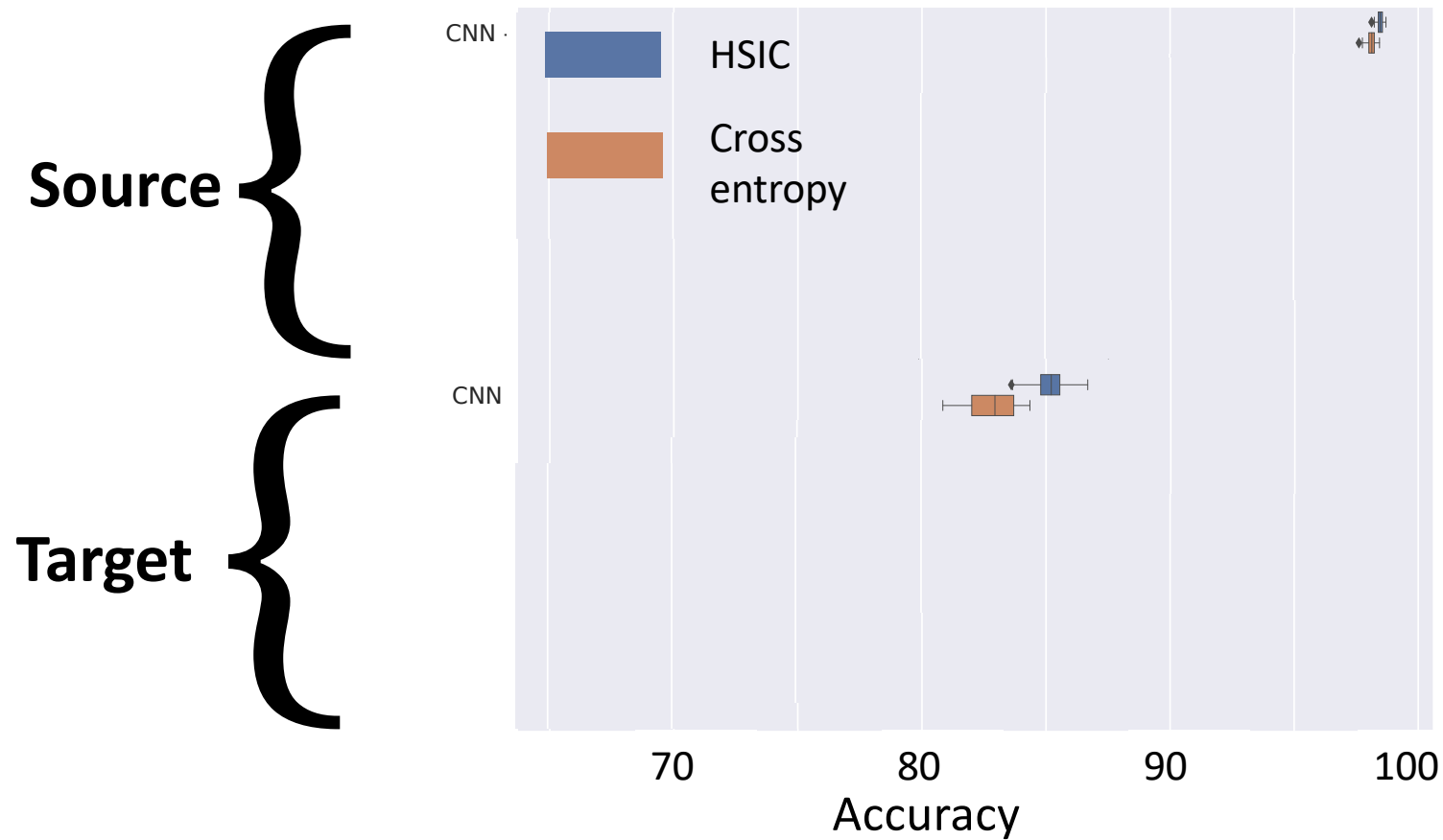
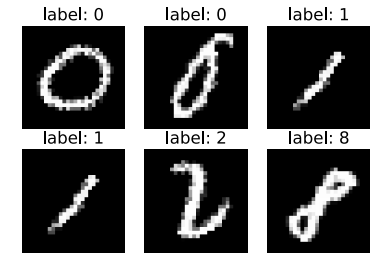
Experiments – rotated MNIST (Heinze-Deml & Meinshausen 2017)

- Train on ordinary MNIST
- Test on MNIST rotated uniformly at random $[-45^\circ, 45^\circ]$



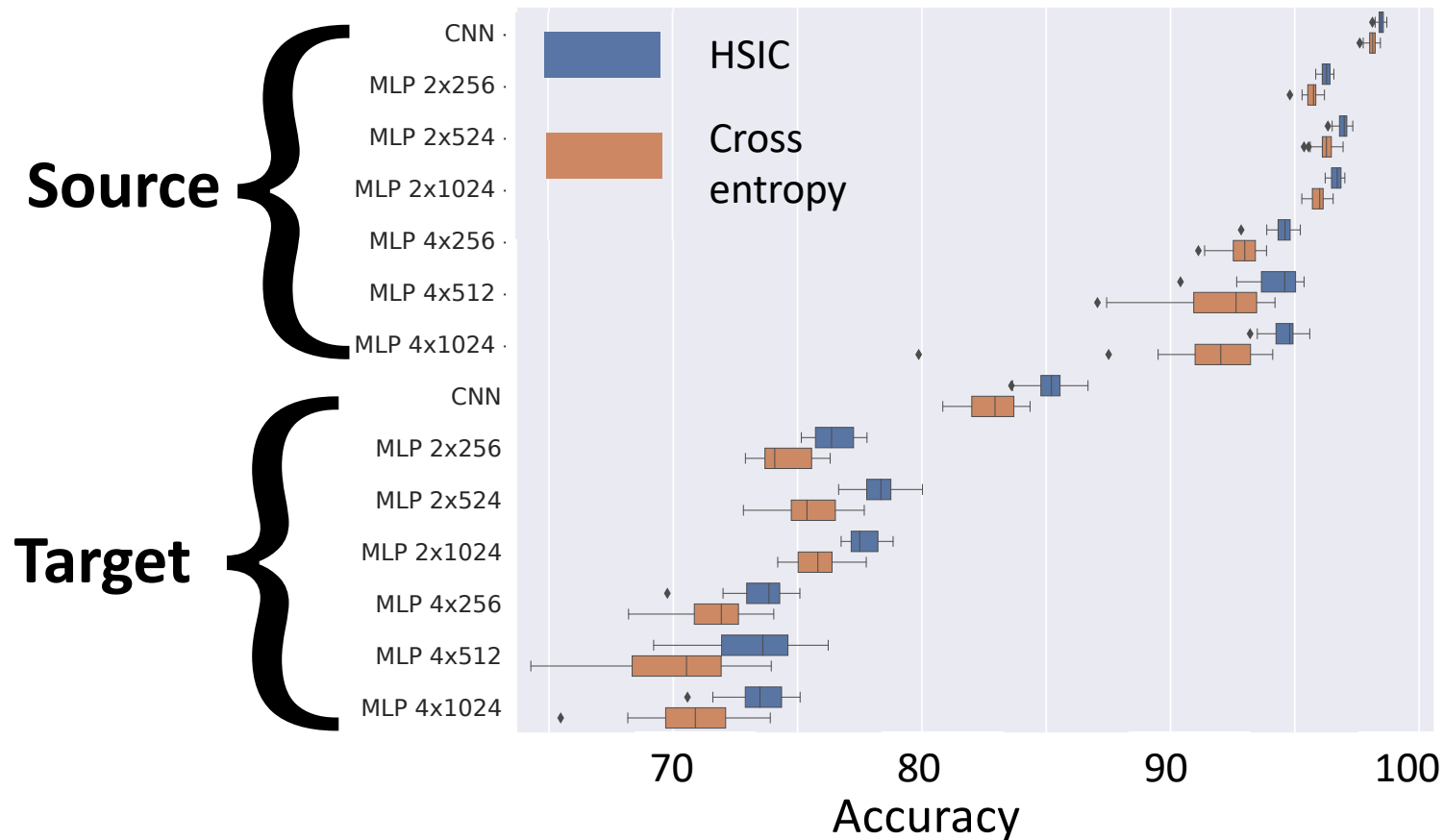
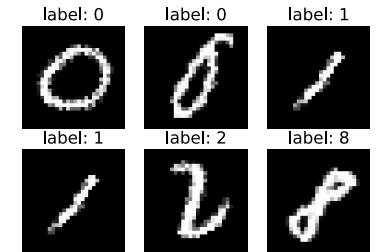
Experiments – rotated MNIST (Heinze-Deml & Meinshausen 2017)

- Train on ordinary MNIST
- Test on MNIST rotated uniformly at random $[-45^\circ, 45^\circ]$



Experiments – rotated MNIST (Heinze-Deml & Meinshausen 2017)

- Train on ordinary MNIST
- Test on MNIST rotated uniformly at random $[-45^\circ, 45^\circ]$

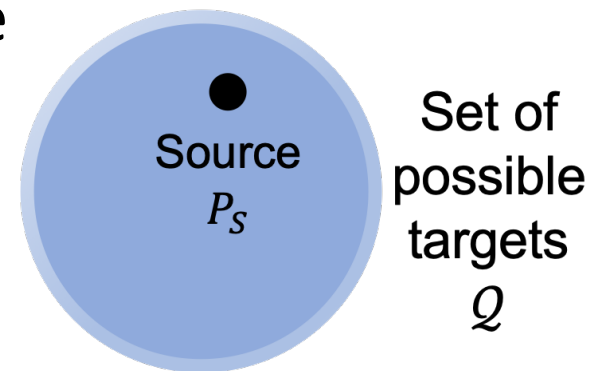
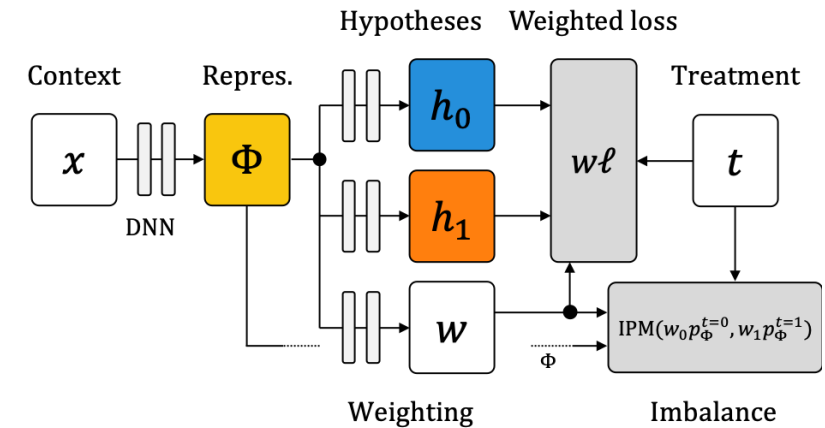


Outline

- ML for causal inference
- Causal inference for ML
 - Off-policy evaluation in a partially observable Markov decision process
 - Robust learning for unsupervised covariate shift

Summary

- Machine learning for causal-inference:
 - Individual-level treatment effects from observational data - robustness to treatment assignments process
- Using recently proposed “negative control” to create first Off-Policy Evaluation scheme for POMDPs, with past and future in the role of the controls
- Learning models robust against unknown covariate shift



Thank you to all my collaborators!

- Fredrik Johansson (Chalmers)
- David Sontag (MIT)
- Nathan Kallus (Cornell-Tech)
- Guy Tennenholtz (Technion)
- Shie Mannor (Technion)
- Daniel Greenfeld (Technion)

Even estimating average effects from observational data is hard!

Do we believe we can estimate *individual-level* effects?



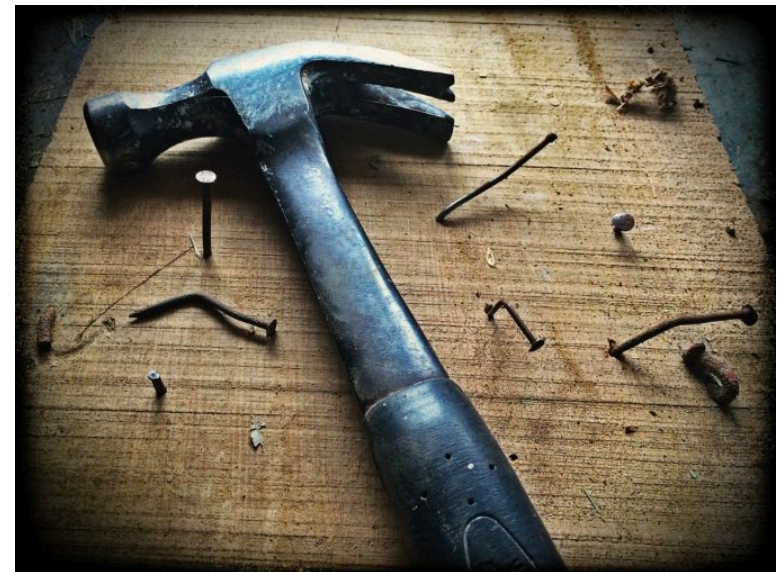
- **Causal identification assumptions:**
- Hidden confounding:
No unmeasured factors that affect both treatment and outcome
- Common support:
 $T = 1$ and $T = 0$ populations should be similar
- Accurate effect estimates:
be able to approximate $\mathbb{E}[Y|x, T = t]$

Even estimating average effects from observational data is hard!

Do we believe we can estimate *individual-level* effects?

- **Causal identification assumptions:**

- Hidden confounding
 - Common support
 - Accurate effect estimates
- We focus on tasks where we hope we can address all three concerns
 - And still be useful
 - Designing for causal identification





You have
condition A.
Treatment
options are
 $T=0$, $T=1$





Obviously,
give $T=0$

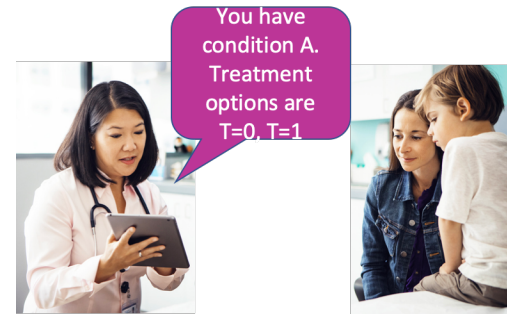
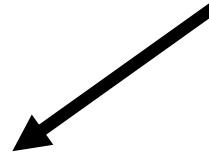
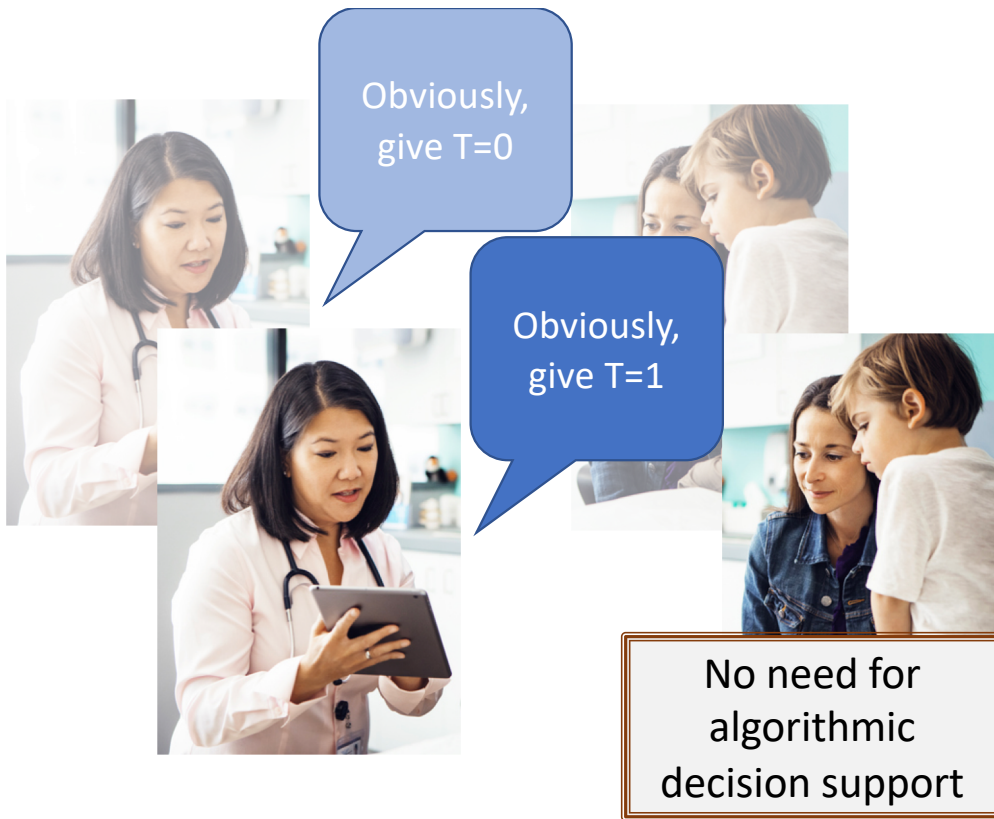


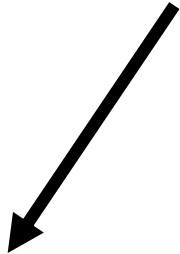
No need for
algorithmic
decision support



You have
condition A.
Treatment
options are
 $T=0$, $T=1$





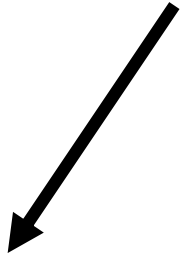


Obviously, give T=1

Obviously, give T=0

I'm not so sure...

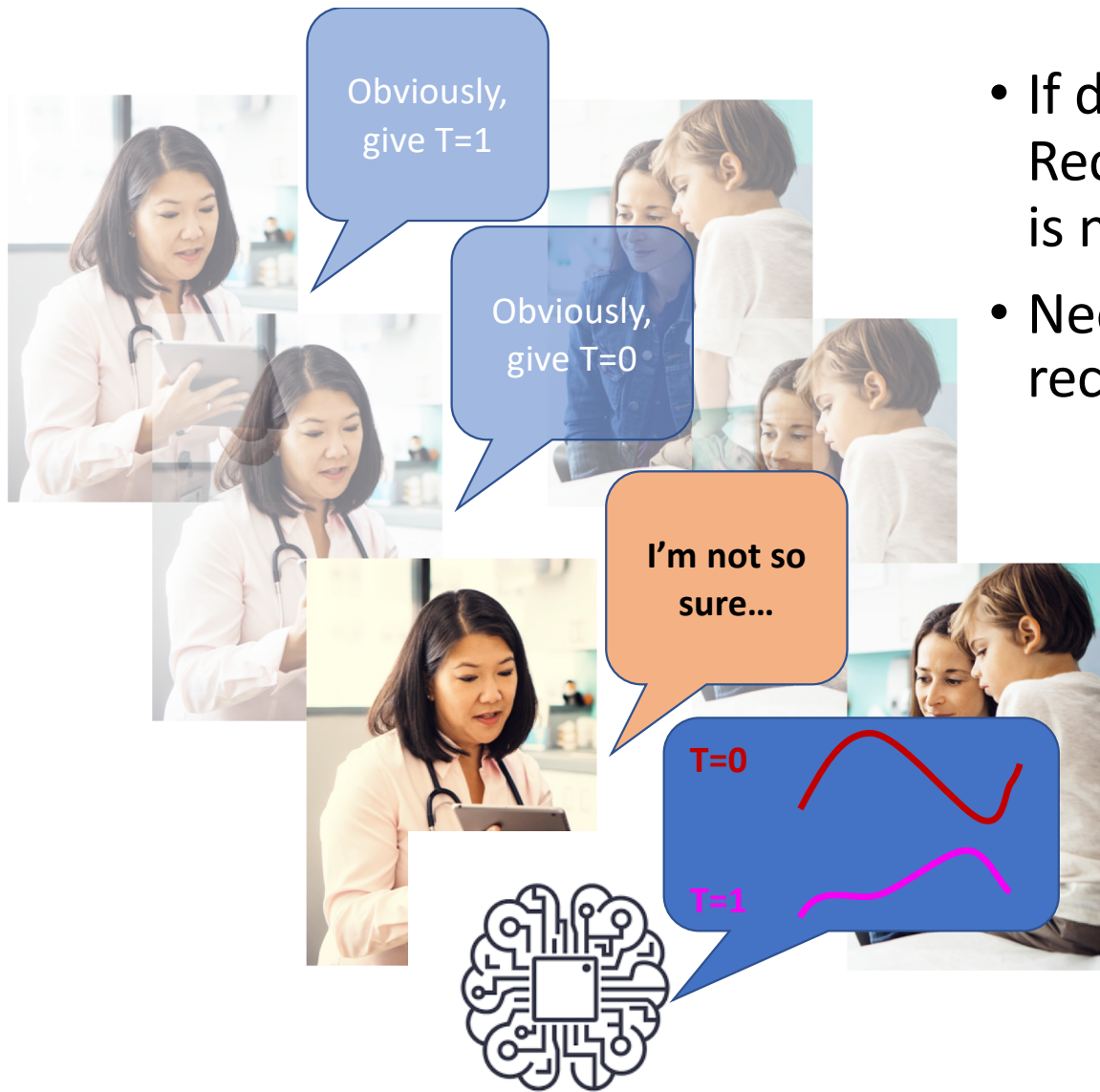
Recommend T=0



You have condition A. Treatment options are T=0, T=1



- If decision could really go either way:
Recommending a suboptimal action
is not as risky



- If decision could really go either way: Recommending a suboptimal action is not as risky
- Need not make explicit recommendation

Estimating average effects is hard!
When do we believe we can estimate individual-level effects?

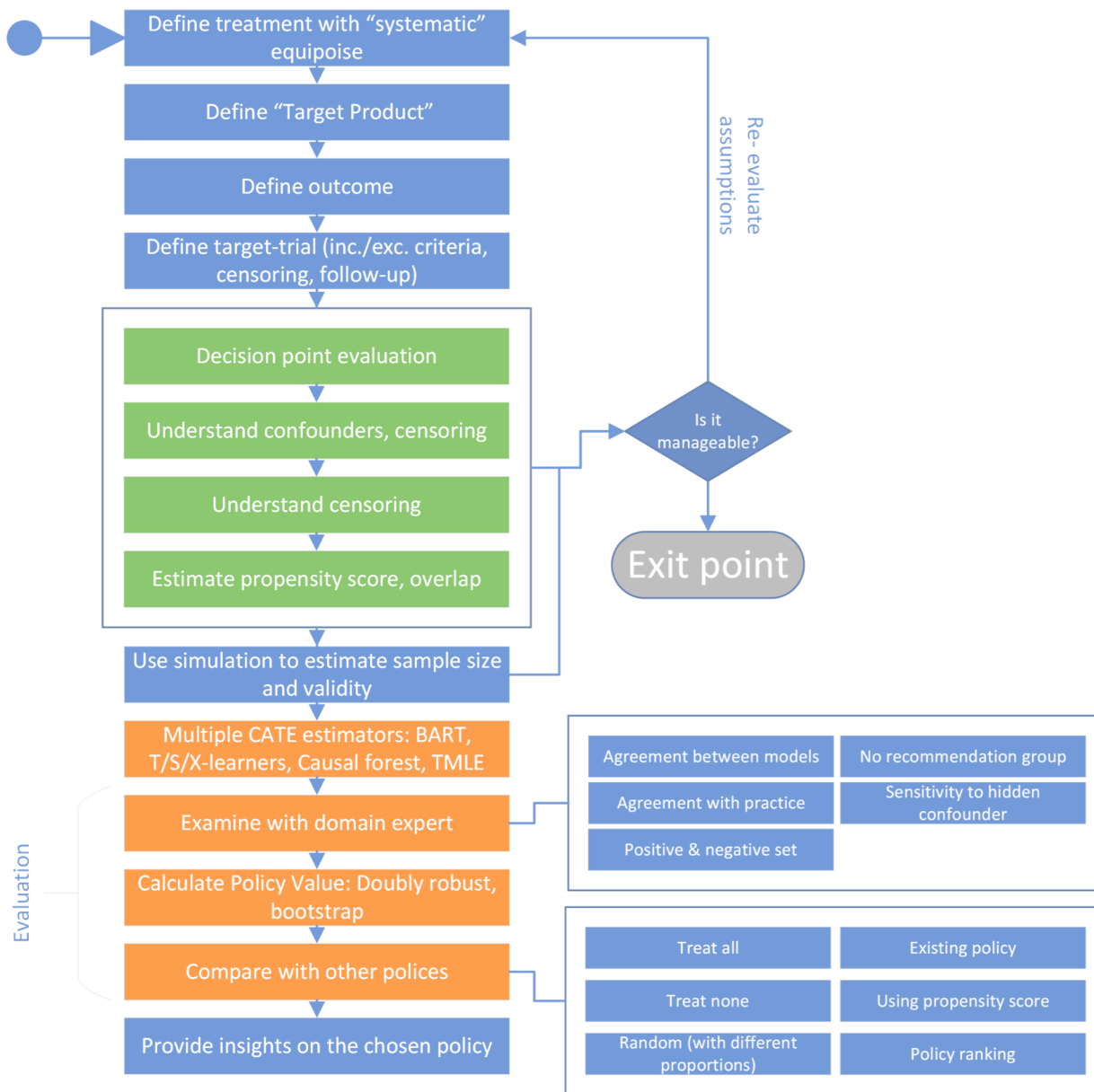
- **Causal identification assumptions:**
- Hidden confounding →
- Common support →
- Accurate effect estimates →



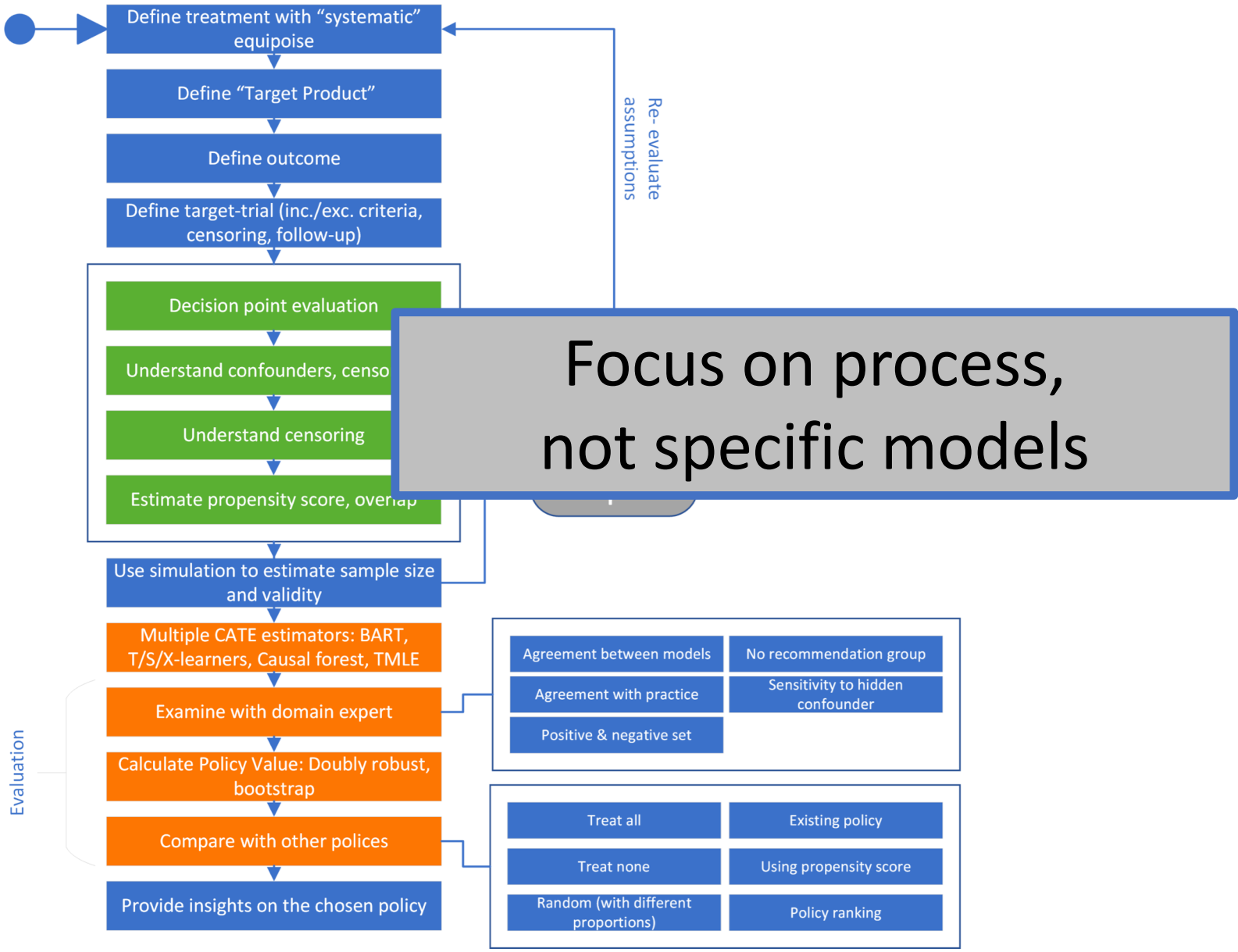
Estimating average effects is hard!
When do we believe we can estimate individual-level effects?

- We don't need to estimate the effects for each patient correctly
- Suffice to give useful recommendation in cases of physician uncertainty
- Physician uncertainty is exactly where we will have more data regarding treatment alternatives for similar patients
- **Include a “we have no recommendation” option**





We are developing a best-practice “pipeline” for decision support models in clinical point-in-time decision support





Preliminary results – study 2

Acute disease treatment

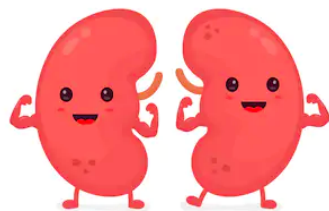
- Investigating the causal effects of diuretics on kidney function in hospitalized acute heart failure patients with kidney injury in Rambam Medical Center
- Physicians tell us:
They have poor guidance how to prescribe diuretics and blood-pressure medications to these patients
- 2157 patients
- More than 200 covariates which are potential confounders:
demographics, lab tests, diagnoses, medications, administrative and more
- Empirically: half of cohort had **increased diuretics**,
half had **decreased diuretics**

Preliminary results – study 2

Acute disease treatment



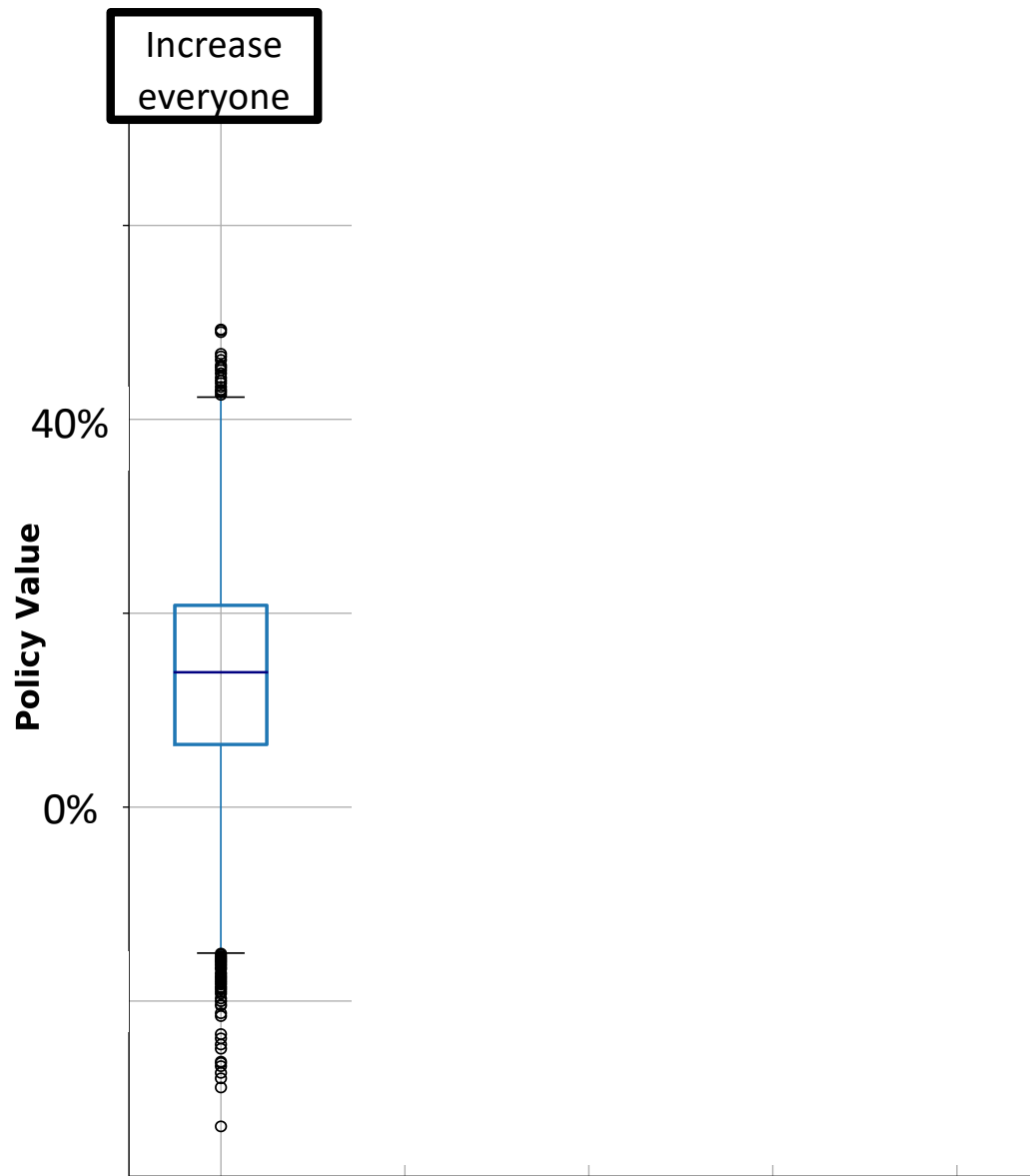
- T=1: “Decrease diuretics”
 - Often improves kidney function
 - Might hurt heart function
- Physicians must balance multiple outcomes
- For now we only examined effect on kidney function



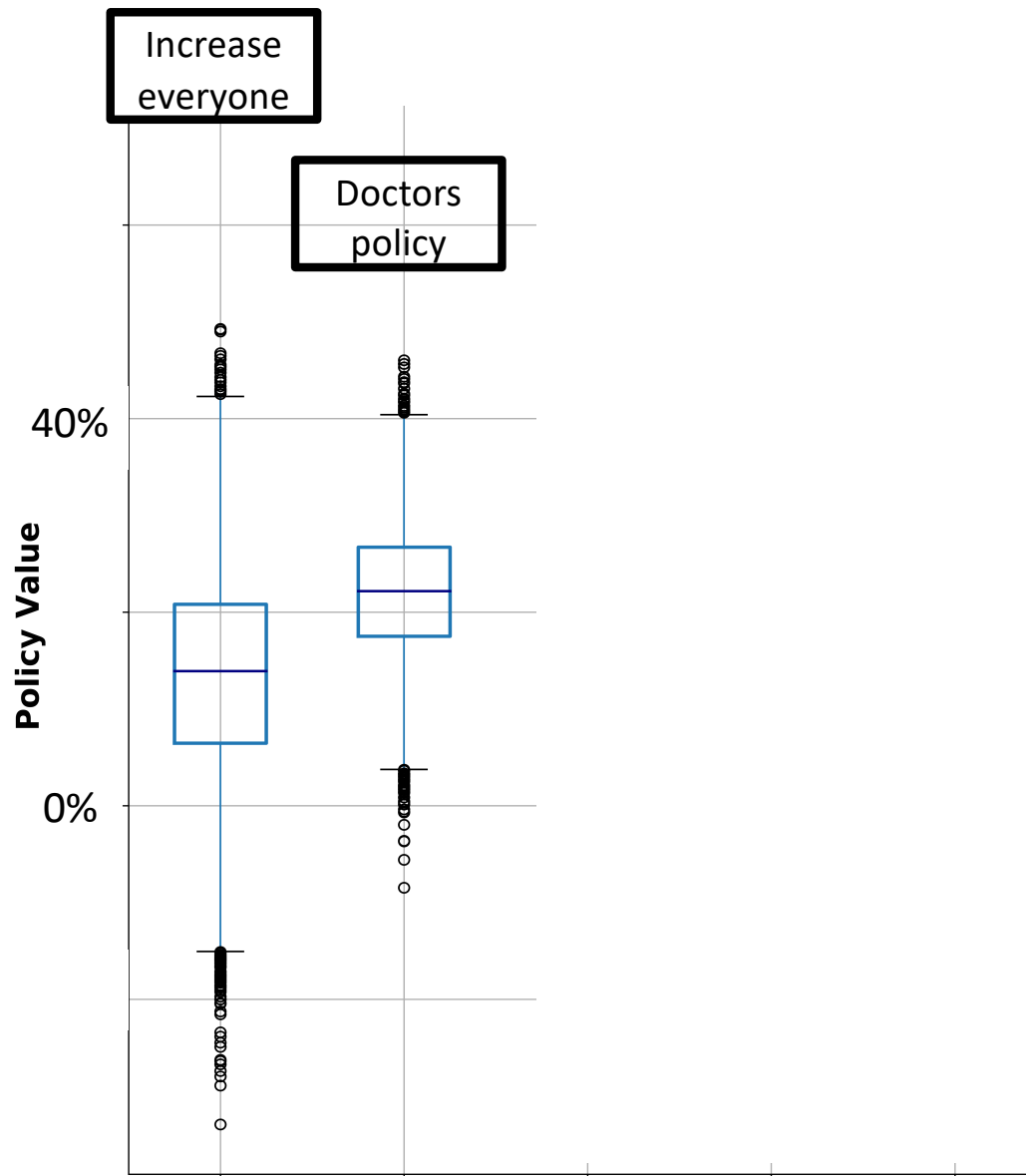
Policy value

- From $\widehat{CATE}(x)$ we can derive a policy recommendation for treatment
- Simple: $\pi(x) = \mathbb{I}(\widehat{CATE}(x) > 0)$
- For any policy π we can estimate its **policy value**:
expected outcome if patients were treated by policy π
- We use Doubly-Robust policy value estimate (Dudík et al. 2011,2014)

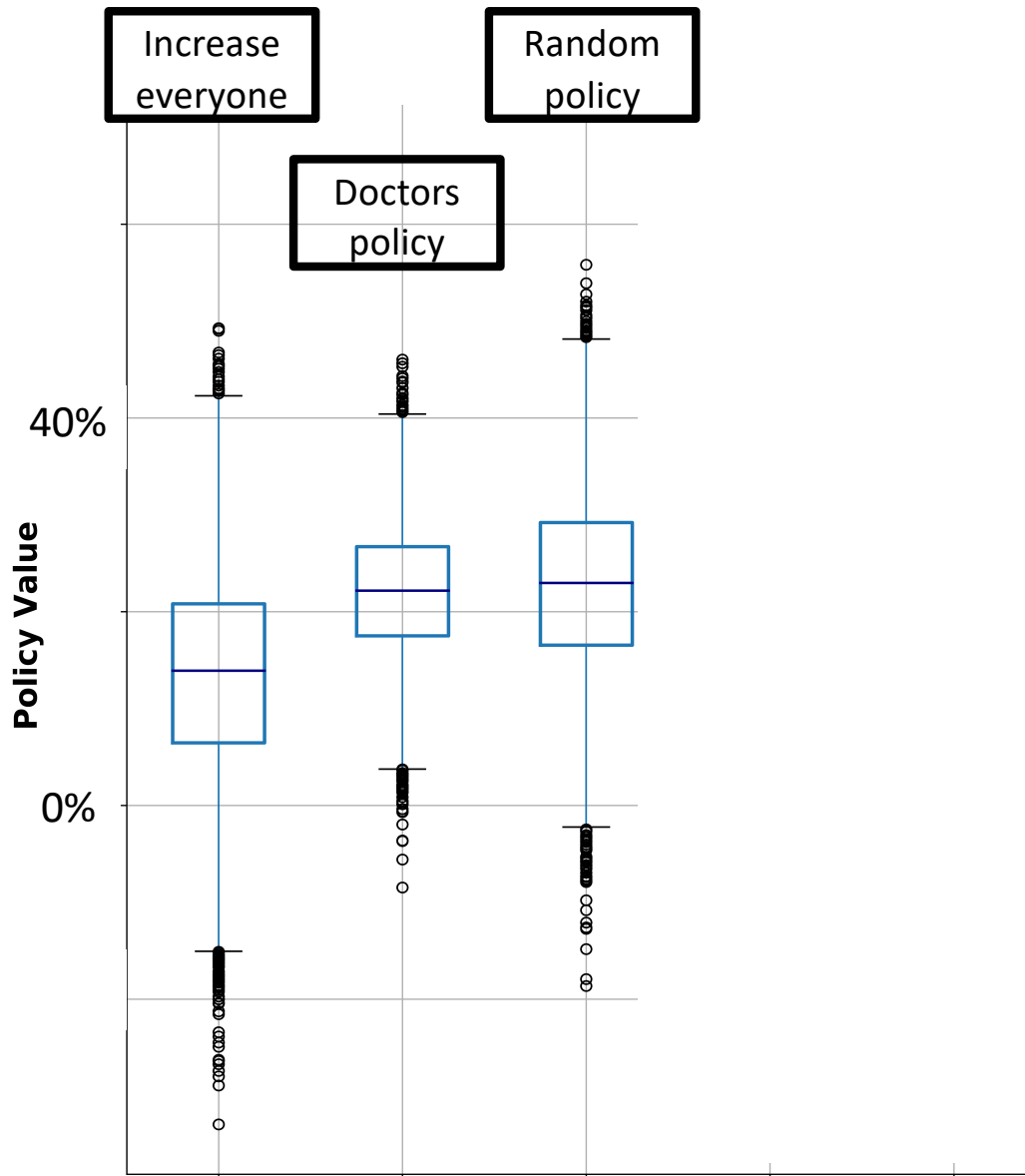
- Increase or decrease diuretics?
- Policy value: % improvement in kidney function (creatinine)
 - 100%: excellent
 - 0%: no improvement
- Recommendations for 461 out of 530 (test set)
- \widehat{CATE} : T-learner XGBoost
- $\pi(x) = \mathbb{I}(\widehat{CATE}(x) > 0)$
- Bootstrap confidence intervals



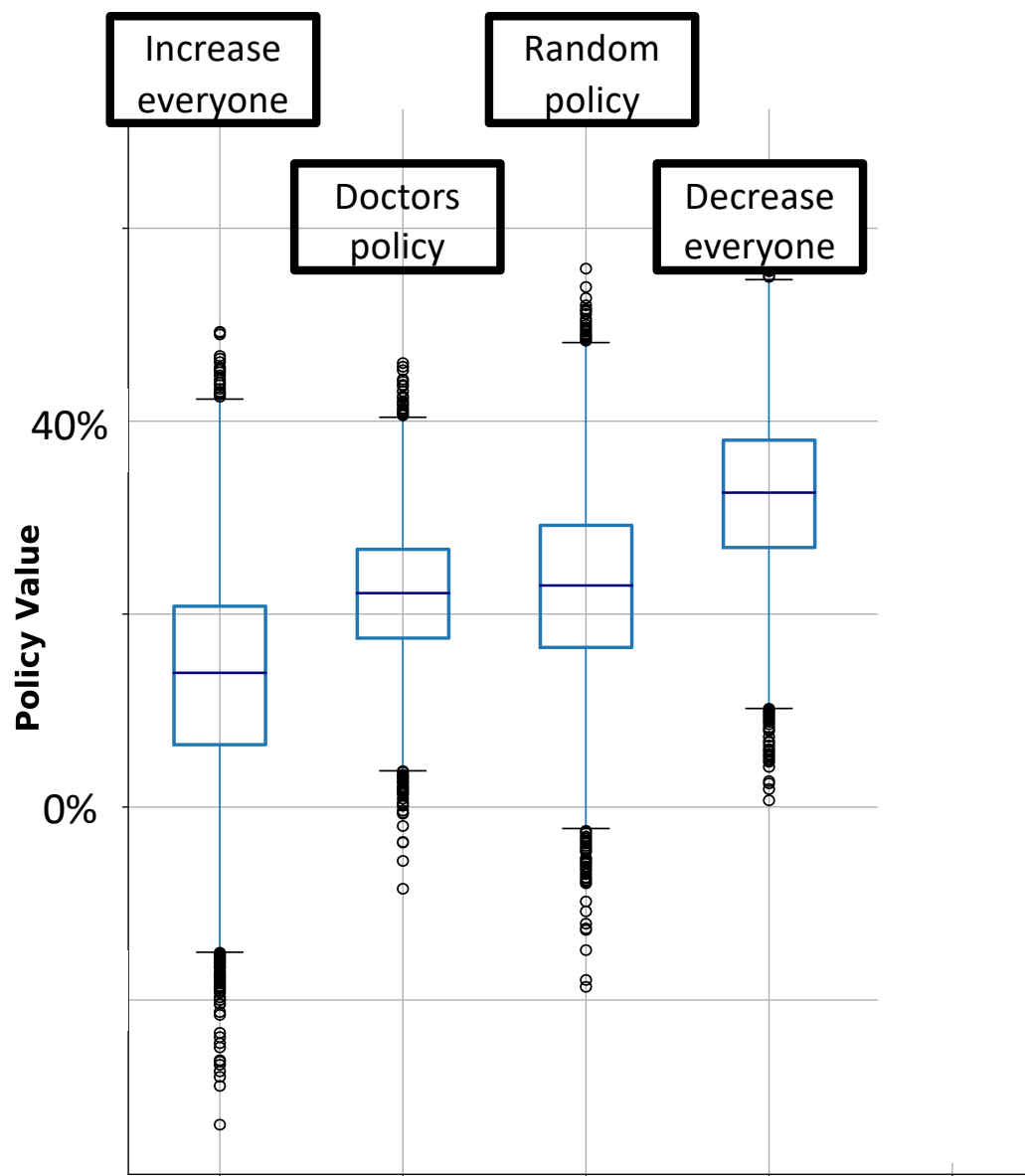
- Increase or decrease diuretics?
- Policy value: % improvement in kidney function (creatinine)
 - 100%: excellent
 - 0%: no improvement
- Recommendations for 461 out of 530 (test set)
- \widehat{CATE} : T-learner XGBoost
- $\pi(x) = \mathbb{I}(\widehat{CATE}(x) > 0)$
- Bootstrap confidence intervals



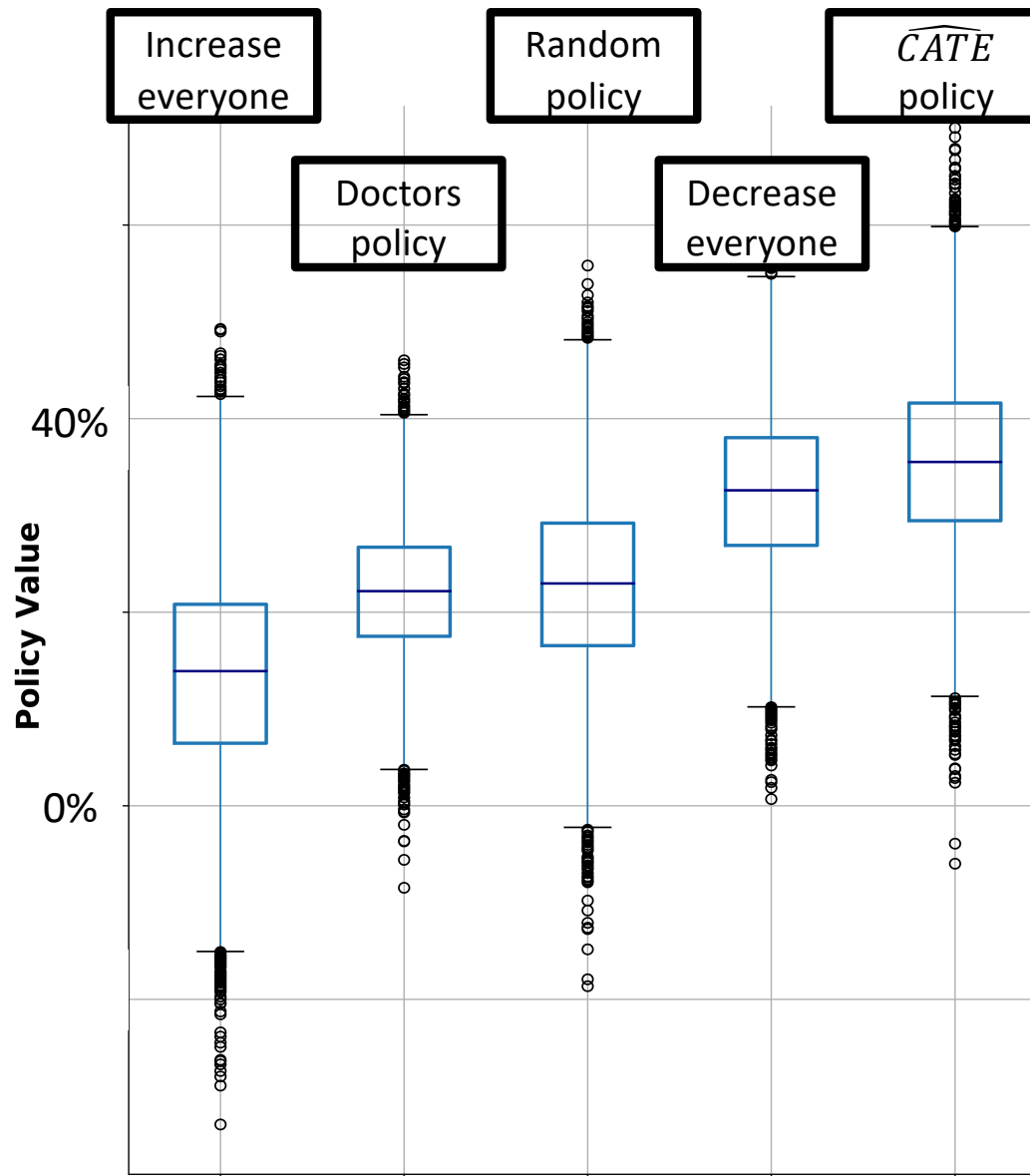
- Increase or decrease diuretics?
- Policy value: % improvement in kidney function (creatinine)
 - 100%: excellent
 - 0%: no improvement
- Recommendations for 461 out of 530 (test set)
- \widehat{CATE} : T-learner XGBoost
- $\pi(x) = \mathbb{I}(\widehat{CATE}(x) > 0)$
- Bootstrap confidence intervals



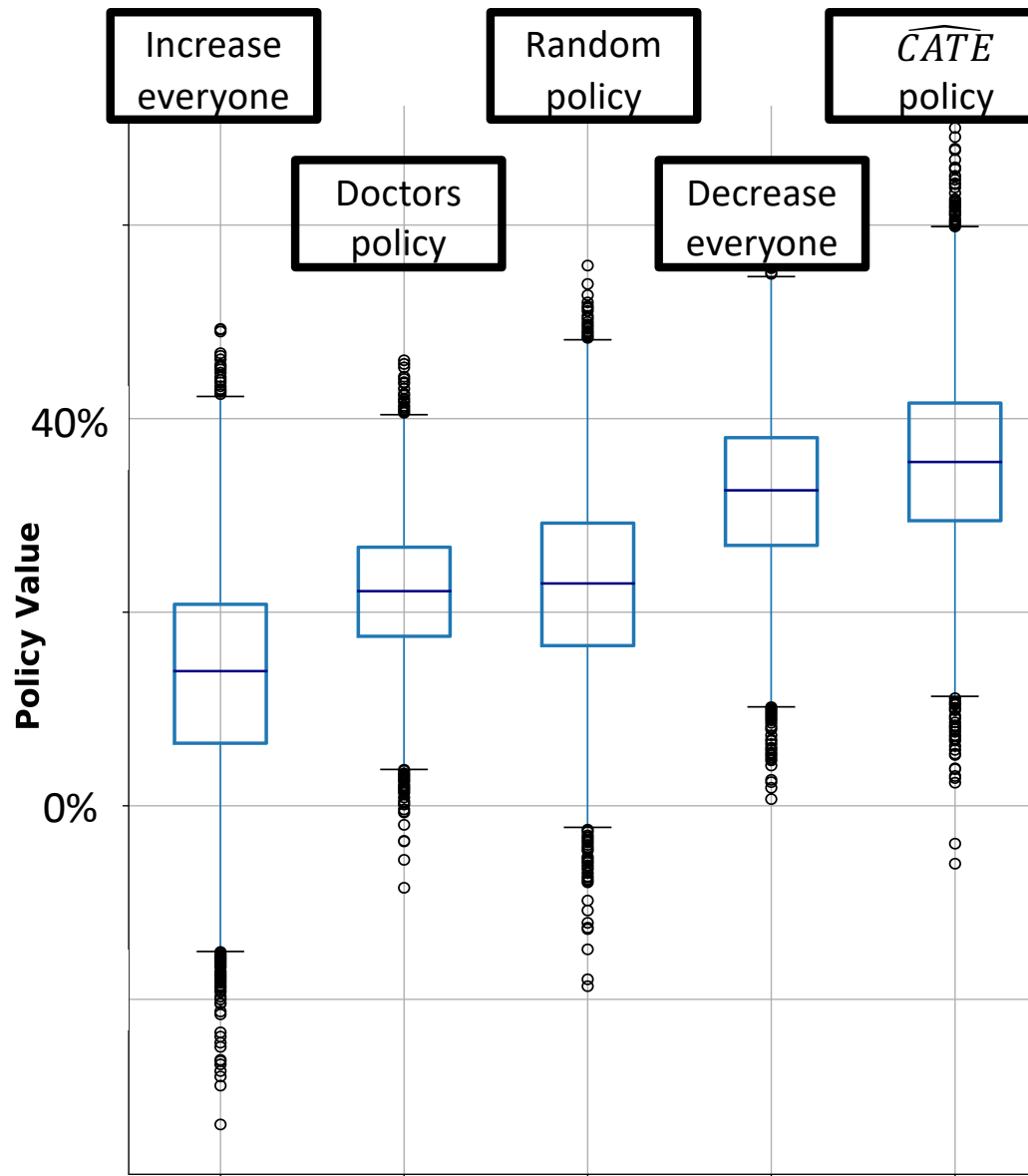
- Increase or decrease diuretics?
- Policy value: % improvement in kidney function (creatinine)
 - 100%: excellent
 - 0%: no improvement
- Recommendations for 461 out of 530 (test set)
- \widehat{CATE} : T-learner XGBoost
- $\pi(x) = \mathbb{I}(\widehat{CATE}(x) > 0)$
- Bootstrap confidence intervals



- Increase or decrease diuretics?
- Policy value: % improvement in kidney function (creatinine)
 - 100%: excellent
 - 0%: no improvement
- Recommendations for 461 out of 530 (test set)
- \widehat{CATE} : T-learner XGBoost
- $\pi(x) = \mathbb{I}(\widehat{CATE}(x) > 0)$
- Bootstrap confidence intervals



- Increase or decrease diuretics?
- Policy value: % improvement in kidney function (creatinine)
 - 100%: excellent
 - 0%: no improvement
- Recommendations for 461 out of 530 (test set)
- \widehat{CATE} : T-learner XGBoost
- $\pi(x) = \mathbb{I}(\widehat{CATE}(x) > 0)$
- Bootstrap confidence intervals



- Our recommendations better than current practice ($p=0.015$)
- Our recommendations have approximately same value as “*decrease diuretics for all patients*”
- Our recommendations **decrease diuretics for only 50% of patients**
- More flexibility with respect to other outcomes
- Effect on other outcomes is work in progress