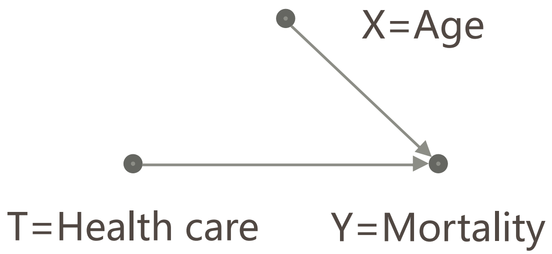
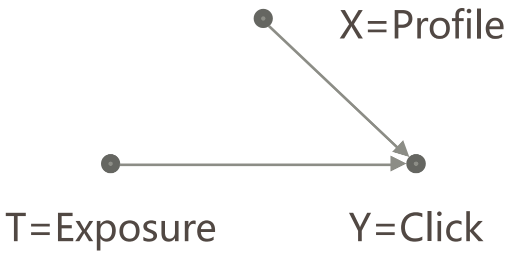
Optimal Transport for Treatment Effect Estimation

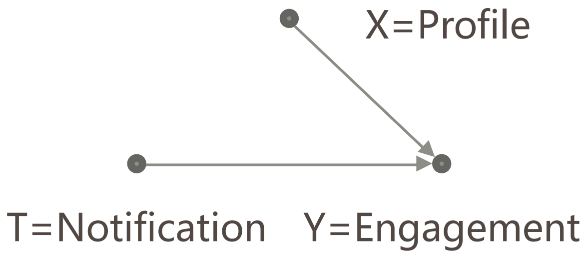
Speaker: Hao Wang

Background

- Treatment effect estimation: estimate the following causal estimands from data:
  - Average Treatment Effect: $\mathrm{ATE} \coloneqq \mathbb{E}[Y_{T=1}] - \mathbb{E}[Y_{T=0}]$
  - **Conditional Average Treatment Effect**: $\mathrm{CATE} \coloneqq \mathbb{E}[Y_{T=1}|X] - \mathbb{E}[Y_{T=0}|X]$



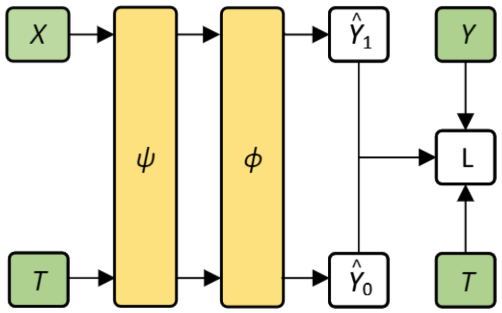




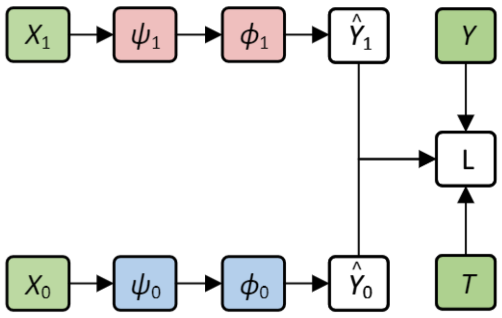
- Missing counterfactual

User	X	T	Y_0	Y_1
1	X1	1	?	click
2	X2	0	click	?
3	X3	0	No click	?
4	X4	1	?	No click

- Solution



(a) Single-learner [1]

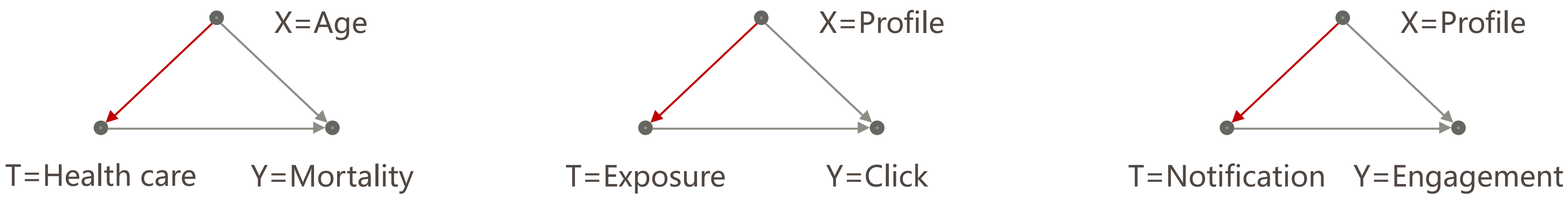


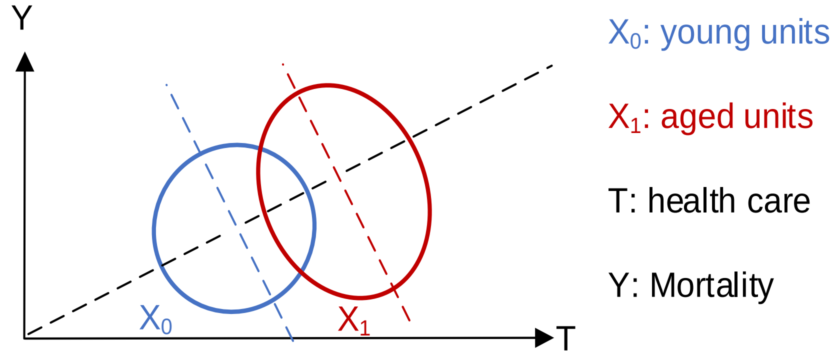
(b) Two-learner [1]

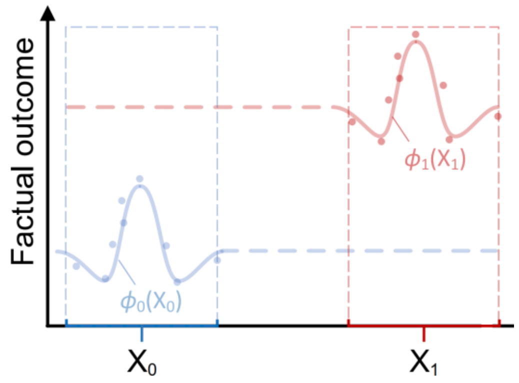
[1] Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." *PNAS*, 2019.

Research problem: selection bias

- Selection bias: the causation T→Y is confounded by the association T ← X → Y
  - It is falsely introduced in data generation process.
  - It manifests as the **discrepancies of covariates** (X) across treatment groups.



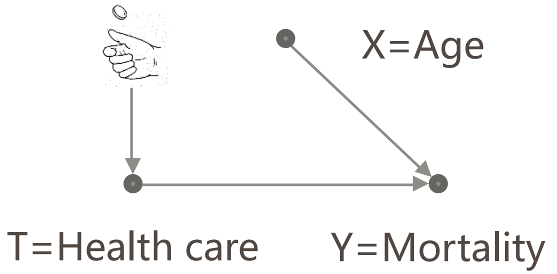


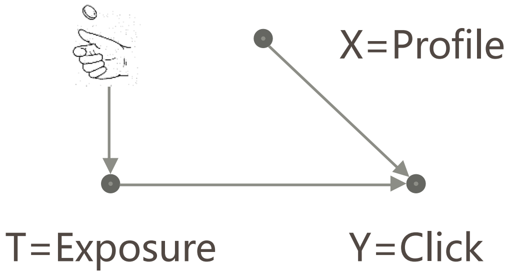


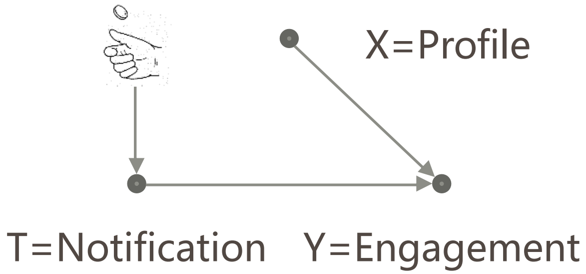
- Confounder —Makes→ Biased data —Train→ Biased estimator

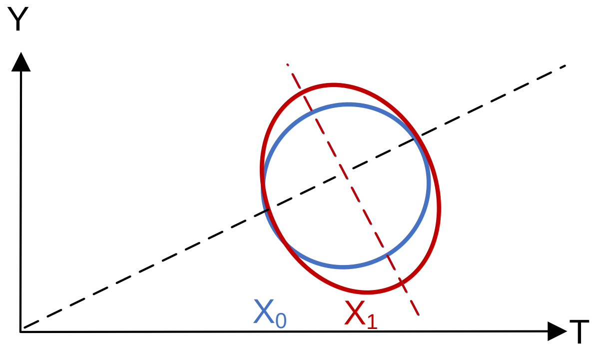
RCT to tackle selection bias

- RCT is a golden approach to eliminate confounding bias. Why?
  - **Randomization** makes covariate balance: $\mathbb{P}(X \mid T=1) = \mathbb{P}(X|T=0), T \perp X$
  - Covariate balance makes association is causation: $\mathbb{P}(Y \mid do(T=t)) = \mathbb{P}(Y \mid T=t)$







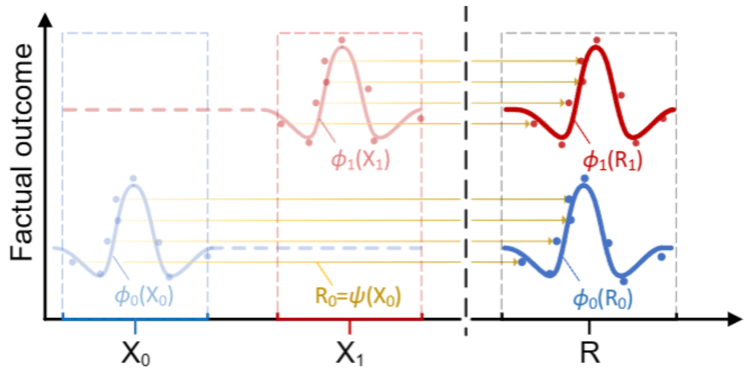


X_0: young units

X_1: aged units

T: health care

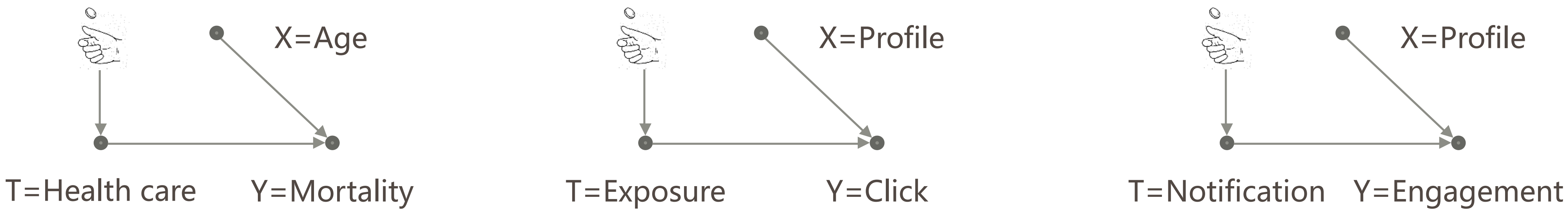
Y: Mortality

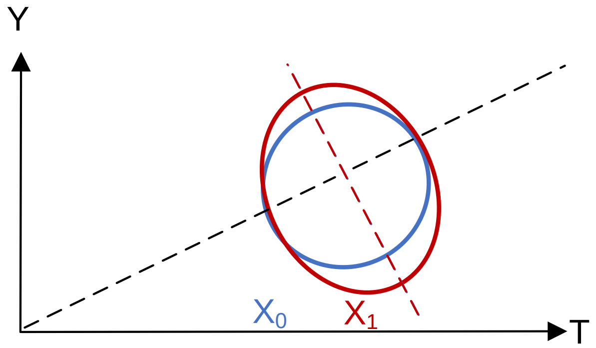


- ~~Confounder~~ → Unbiased data → Unbiased estimator (Train)

Adjustment as an alternative to RCT

- RCT is a golden approach to eliminate confounding bias. Why?
  - ~~Randomization~~ **Adjustment** makes covariate balance: $\mathbb{P}(X \mid T=1) = \mathbb{P}(X|T=0), T \perp X$
  - Covariate balance makes association is causation: $\mathbb{P}(Y \mid do(T=t)) = \mathbb{P}(Y \mid T=t)$



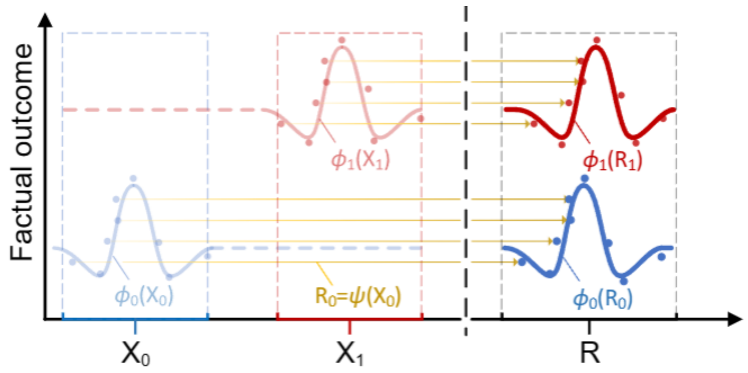


X_0: young units

X_1: aged units

T: health care

Y: Mortality

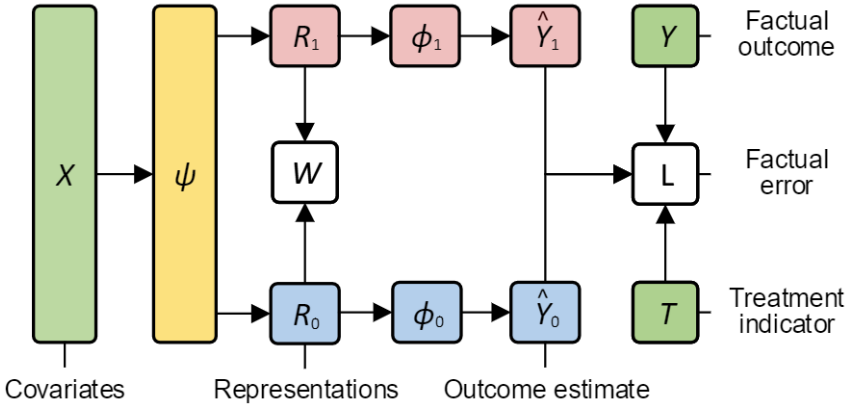


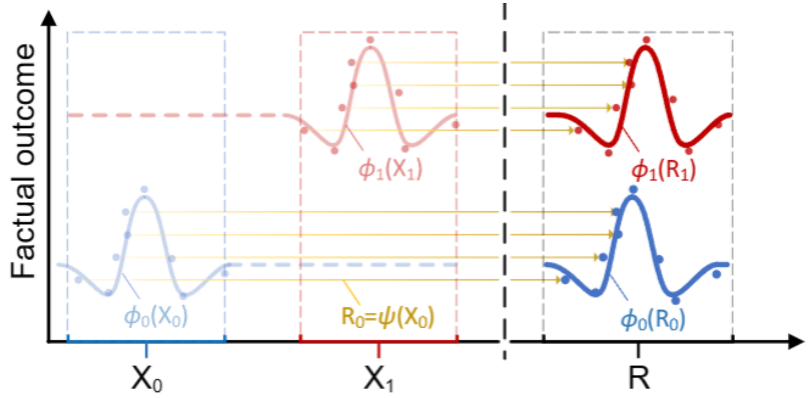
- Confounder → Biased data → Unbiased estimator

Adjustment & Train

Adjustment with CFR

- Goal: generate balanced distribution between different treatment groups.
- CounterFactual Regression [2]: project covariates to a balanced representation space.





Theorem A.1. *Let ψ and ϕ be the maps in Definition 2.2, $\mathcal{F}$ be a predefined sufficiently large function family of ϕ, $\mathrm{IPM}_{\mathcal{F}}$ be the integral probability metric induced by $\mathcal{F}$. Assume there exists a constant $B_\psi > 0$, such that for $t \in \{0,1\}$, $\frac{1}{B_\psi} \cdot l_{\psi,\phi}(x,t) \in \mathcal{F}$ holds. Uri et al. [63] demonstrate:*

$$\epsilon_{\mathrm{PEHE}}(\psi,\phi) \leq 2\left(\epsilon_{\mathrm{F}}^{\mathrm{T}=0}(\psi,\phi) + \epsilon_{\mathrm{F}}^{\mathrm{T}=1}(\psi,\phi) + B_\psi \mathrm{IPM}_{\mathcal{F}}\left(\mathbb{P}_\psi^{\mathrm{T}=1}, \mathbb{P}_\psi^{\mathrm{T}=0}\right) - 2\sigma_Y^2\right), \tag{24}$$

where $\epsilon_{\mathrm{F}}^{\mathrm{T}=0}$ and $\epsilon_{\mathrm{F}}^{\mathrm{T}=1}$ follow Definition A.3, $\mathbb{P}_\psi^{\mathrm{T}=1}(r)$ and $\mathbb{P}_\psi^{\mathrm{T}=0}(x)$ follow Definition A.4.

[2] Shalit, Uri et al. "Estimating individual treatment effect: generalization bounds and algorithms." *ICML*, 2017.

Adjustment with CFR

NEURAL INFORMATION PROCESSING SYSTEMS

- Core of CFR [2]: **accurate calculation of distribution discrepancy**.
  - Inaccurate discrepancy->false update of estimators->biased inference

- Research problem: How to devise discrepancy that can be accurately calculated in the specific context of causal inference?

- Current divergences fail in the situations as follows:

Concerned properties	Wasserstein	f-divergence	GAN-based	MMD	Ours
Free of adversarial training	✓	✓	✗	✓	✓
Non-overlapped supports	✓	✗	✓	✓	✓
Mini-batch sampling effects	✗	✗	✗	✗	✓
Unobserved confounding effects	✗	✗	✗	✗	✓

[2] Shalit, Uri et al. "Estimating individual treatment effect: generalization bounds and algorithms." *ICML*, 2017.

Optimal transport: formulation and application to CFR

- **Optimal Transport (OT):** For empirical distributions α and β with n and m samples, OT aims to find an optimal plan $\boldsymbol{\pi} \in R_{+}^{n\times m}$ that minimizes the transport cost between α and β . Formally, the problem is defined as:

$$W(\alpha,\beta) = \min_{\pi\in\Pi(\alpha,\beta)} \langle \boldsymbol{C}, \boldsymbol{\pi} \rangle, \qquad \Pi(\alpha,\beta) = \{\boldsymbol{\pi} \in R_{+}^{n\times m} \colon \boldsymbol{\pi}\mathbf{1}_{\mathrm{m}} = \boldsymbol{a}, \boldsymbol{\pi}^{T}\mathbf{1}_{\mathrm{n}} = \mathbf{b}\}$$

where $\mathrm{W}(\alpha,\beta)$ is the transport cost, $\boldsymbol{C} \in \mathrm{R}_{+}^{\mathrm{n}\times\mathrm{m}}$ denotes the sample-wise distance between α and β. $\mathbf{1}_{\mathrm{m}}$ and $\mathbf{1}_{\mathrm{n}}$ are column vectors filled with ones. $\boldsymbol{a}$ and $\boldsymbol{b}$ specify the mass of units in α and β.

- We formulate causal inference as an OT problem, where the discrepancy in CFR is computed as the OT cost between the treatment groups.
  - **Unbiased estimator** with theoretical foundations.
  - **Numerical stability** compared with other discrepancy measures (GAN, f-divergence).
  - **Flexibility to incorporate task properties** by editing the transport problem.

Minibatch sampling effect issue with CFR

- Minibatch sampling effect.
  - Minibatch-level outliers, see Fig.2 (b).
  - Minibatch-level outcomeimbalance , see Fig.2 (c).

- Why does it exist?

Theorem 3.1. *Let ψ and ϕ be the representation mapping and factual outcome mapping, respectively; $\hat{\mathbb{W}}_\psi$ be the group discrepancy at a mini-batch level. With the probability of at least $1-\delta$, we have:*

$$\epsilon_{\mathrm{PEHE}}(\psi,\phi) \le 2[\epsilon_{\mathrm{F}}^{T=1}(\psi,\phi) + \epsilon_{\mathrm{F}}^{T=0}(\psi,\phi) + B_\psi \hat{\mathbb{W}}_\psi - 2\sigma_Y^2 + \mathcal{O}(\frac{1}{\delta N})], \quad (7)$$

where $\epsilon_{\mathrm{F}}^{T=1}$ and $\epsilon_{\mathrm{F}}^{T=0}$ are the expected errors of factual outcome estimation, N is the batch size, σ_Y^2 is the variance of outcomes, B_ψ is a constant term, and $\mathcal{O}(\cdot)$ is a sampling complexity term.

- How to solve it?

Definition 3.2. *For empirical distributions α and β with n and m units, respectively, optimal transport with relaxed mass-preserving constraint seeks the transport strategy $\boldsymbol{\pi}$ at the minimum cost:*

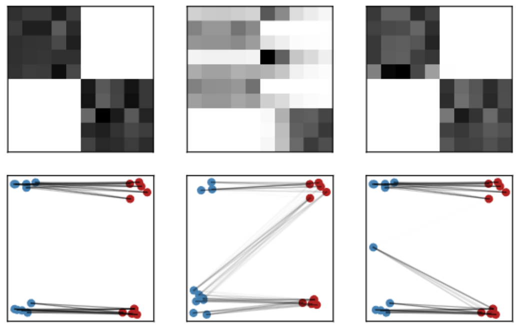
$$\mathbb{W}^{\epsilon,\kappa}(\alpha,\beta) := \langle \mathbf{D},\boldsymbol{\pi}\rangle, \boldsymbol{\pi} := \arg\min_{\boldsymbol{\pi}} \langle \mathbf{D},\boldsymbol{\pi}\rangle - \epsilon \mathrm{H}(\boldsymbol{\pi}) \boxed{+\kappa(\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}\mathbf{1}_m,\mathbf{a}) + \mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}^\top\mathbf{1}_n,\mathbf{b}))} \quad (9)$$

where $\mathbf{D} \in \mathbb{R}_+^{n\times m}$ is the unit-wise distance, and $\mathbf{a}$, $\mathbf{b}$ indicate the mass of units in α and β, respectively.

Corollary 3.1. *For empirical distributions α, β with n and m units, respectively, adding an outlier a' to α and denoting the disturbed distribution as α', we have*

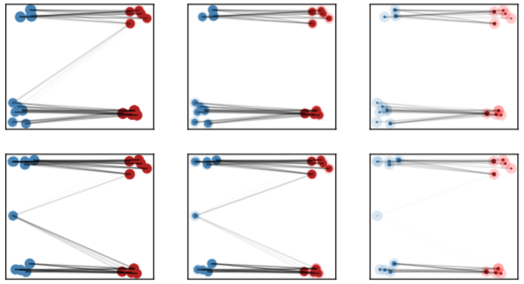
$$\mathbb{W}^{0,\kappa}(\alpha',\beta) - \mathbb{W}^{0,\kappa}(\alpha,\beta) \le 2\kappa(1 - e^{-\sum_{b\in\beta}(a'-b)^2/2\kappa})/(n+1), \quad (10)$$

which is upper bounded by $2\kappa/(n+1)$. $\mathbb{W}^{0,\kappa}$ is the unbalanced discrepancy as per Definition 3.2.



(a) Ideal (b) Imbalance (c) Outlier

Figure 2: Optimal transport plan (upper) and its geometric interpretation (down) in three cases, where the connection strength depicts the transported mass. Different colors (vertical positions) indicate different treatments (outcomes).



(a) $\kappa = 50$ (b) $\kappa = 10$ (c) $\kappa = 2$

Figure 4: Geometric interpretation of OT plan with RMPR under the outcome imbalance (upper) and outlier (down) settings. The dark area indicates the transported mass of a unit, *i.e.*, marginal of the transport matrix π. The light area indicates the total mass.

Unobserved confounding effect issue with CFR

- Effect of unobserved confounders.
  - Invalidate backdoor adjustment.
  - Mislead the update of treatment effect estimator

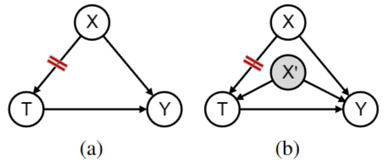


Figure 3: Causal graphs with (a) and w/o (b) the unconfoundedness assumption. The shaded node indicates the hidden confounder X'.

- How to solve it?
  - $$D_{ij}^{\gamma} = \left\|r_i - r_j\right\|^2 + \gamma \underbrace{\left[\left\|y_i^{T=0} - y_j^{T=0}\right\|^2 + \left\|y_i^{T=1} - y_j^{T=1}\right\|^2\right]}_{\textbf{Proximal Factual Outcome Regularizer}}$$

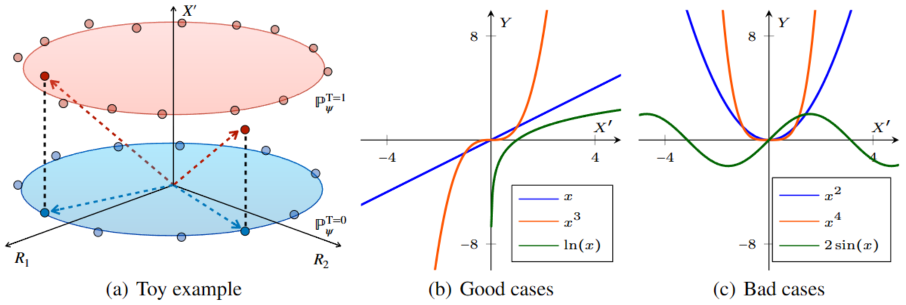


Figure 9: A diagram showing how PFOR works and its limitations. (a) A toy example of PFOR, where R and X' indicate the balanced representations and an unobserved confounder, respectively; scatters indicate the empirical distribution of units in the treated and control groups; for solid scatters with balanced R, the colored dashed line indicates the ground truth outcome $Y = \sqrt{R_1^2 + R_2^2 + X'^2}$ in each group, the black dashed line measures the difference of unobserved X'. (b) Cases that satisfy Assumption D.1, where the the outcome Y is monotone with unobserved X' given observed confounders in R. (c) Cases that violate Assumption D.1, where the Y is non-monotone with X'.

- Limitations.
  - Partial identification of transport strategy given monotonic covariate effect.
  - OT meets partial identification: an interesting topic which warrants further investigation

NEURAL INFORMATION PROCESSING SYSTEMS

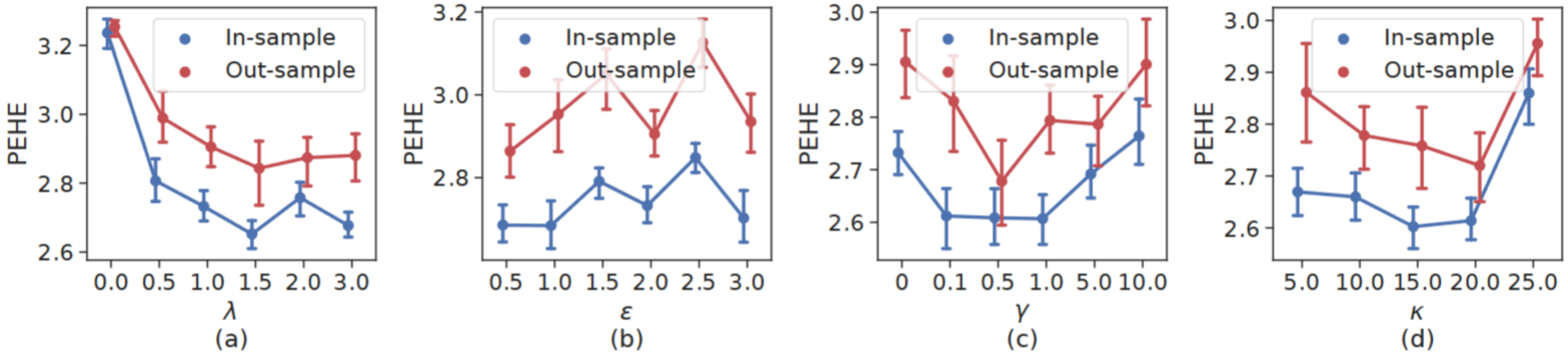
Overall performance

Table 1: Performance (mean±std) on the PEHE and AUUC metrics. "*" marks the baseline estimators that ESCFR outperforms significantly at p-value < 0.05 over paired samples t-test.

Dataset	ACIC (PEHE)		IHDP (PEHE)		ACIC (AUUC)		IHDP (AUUC)	
Model	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample
OLS	3.749±0.080*	4.340±0.117*	3.856±6.018	5.674±9.026	0.843±0.007	0.496±0.017*	0.652±0.050	0.492±0.032*
R.Forest	3.597±0.064*	3.399±0.165*	2.635±3.598	4.671±9.291	0.902±0.016	0.702±0.026*	0.736±0.142	0.661±0.259
S.Learner	3.572±0.269*	3.636±0.254*	1.706±1.600*	3.038±5.319	**0.905±0.041**	0.627±0.014*	0.633±0.183	0.702±0.330
T.Learner	3.429±0.142*	3.566±0.248*	1.567±1.136*	2.730±3.627	0.846±0.019	0.632±0.020*	0.651±0.179	0.707±0.333
TARNet	3.236±0.266*	3.254±0.150*	0.749±0.291	1.788±2.812	0.886±0.046	0.662±0.014*	0.654±0.184	0.711±0.329
C.Forest	3.449±0.101*	3.196±0.177*	4.018±5.602*	4.486±8.677	0.717±0.005*	0.709±0.018*	0.643±0.141	0.695±0.294
k-NN	5.605±0.168*	5.892±0.138*	2.208±2.233*	4.319±7.336	0.892±0.007*	0.507±0.034*	0.725±0.142	0.668±0.299
O.Forest	8.094±4.669*	4.148±2.224*	2.605±2.418*	3.136±5.642	0.744±0.013	0.699±0.022*	0.664±0.157	0.702±0.325
PSM	5.228±0.154*	5.094±0.301*	3.219±4.352*	4.634±8.574	0.884±0.010	0.745±0.021	**0.740±0.149**	0.681±0.253
BNN	3.345±0.233*	3.368±0.176*	0.709±0.330	1.806±2.837	0.882±0.033	0.645±0.013*	0.654±0.184	0.711±0.329
CFR-MMD	3.182±0.174*	3.357±0.321*	0.777±0.327	1.791±2.741	0.871±0.032	0.659±0.017*	0.655±0.183	0.710±0.329
CFR-WASS	3.128±0.263*	3.207±0.169*	0.657±0.673	1.704±3.115	0.873±0.029	0.669±0.018*	0.656±0.187	0.715±0.311
ESCFR	**2.252±0.297**	**2.316±0.613**	**0.502±0.252**	**1.282±2.312**	0.796±0.030	**0.754±0.021**	0.665±0.166	**0.719±0.311**

Ablation & sensitivity studies

			In-sample		Out-sample	
SOT	RMPR	PFOR	PEHE	AUUC	PEHE	AUUC
✗	✗	✗	3.2367±0.2666*	**0.8862+0.0462**	3.2542+0.1505*	0.6624+0.0149*
✓	✗	✗	3.1284±0.2638*	0.8734+0.0291	3.2073+0.1699*	0.6698+0.0187*
✓	✓	✗	2.6459+0.2747*	0.8356+0.0286	2.7688±0.4009	0.7099+0.0157*
✓	✗	✓	2.5705±0.3403*	0.8270±0.0341	2.6330±0.4672	0.7110±0.0287*
✓	✓	✓	**2.2520±0.2975**	0.7968±0.0307	**2.3165+0.6136**	**0.7542±0.0202**



Thanks for your listening

Speaker: Hao Wang

Contact: haohaow@zju.edu.cn