

DR-VIDAL - Doubly Robust Variational Information theoretic Deep Adversarial Learning for Counterfactual Prediction and Treatment Effect Estimation

S90

Shantanu Ghosh,
University of Pittsburgh

Zheng Feng, Jiang Bian, Kevin Butler, Mattia Prosperi
University of Florida

Nancy is having fever



I am having a mild fever. Shall I take medicine or not?

Nancy is having fever

Features

Age: 32

Sex: F

Blood group: A+

Race: Asian

Blood Sugar: High

Temperature: 100° F

....



I am having a mild
fever. Shall I take
medicine or not?

Nancy is having fever

Features

Age: 32

Sex: F

Blood group: A+

Race: Asian

Blood Sugar: High

Temperature: 100° F

....



I am having a mild fever. Shall I take medicine or not?

Medication A
Control
 $T=0$

Temperature = ?
 Y_0

Medication B
Treated
 $T=1$

Temperature = ?
 Y_1

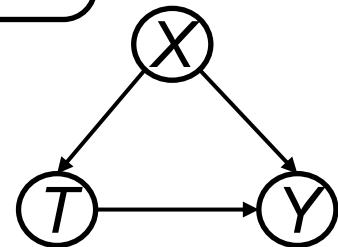
Nancy is having fever

Features

Age: 32
Sex: F
Blood group: A+
Race: Asian
Blood Sugar: High
Temperature: 100° F
....



I am having a mild fever. Shall I take medicine or not?



Medication A
Control
 $T=0$

Temperature = ?
 Y_0

Medication B
Treated
 $T=1$

Temperature = ?
 Y_1

Nancy is having fever

Features

Age: 32

Sex: F

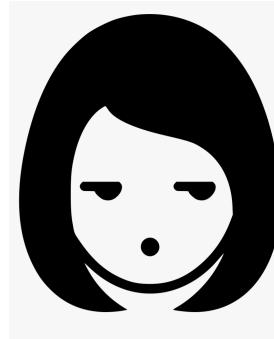
Blood group: A+

Race: Asian

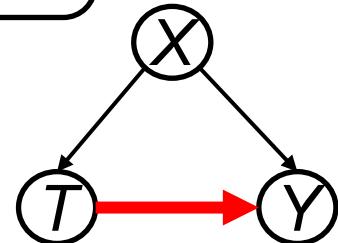
Blood Sugar: High

Temperature: 100° F

....

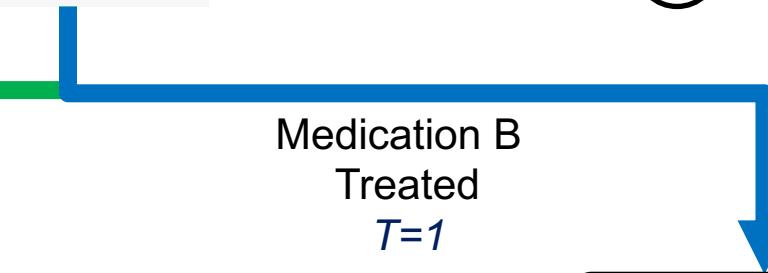


I am having a mild fever. Shall I take medicine or not?



Medication A
Control
 $T=0$

Temperature = ?
 Y_0



Medication B
Treated
 $T=1$

Temperature = ?
 Y_1

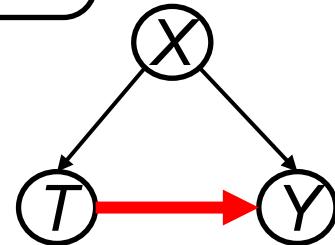
Fundamentally challenging problem

Features

Age: 32
Sex: F
Blood group: A+
Race: Asian
Blood Sugar: High
Temperature: 100° F
....



I am having a mild fever. Shall I take medicine or not?



Medication A
Control
 $T=0$

Temperature = ?
 Y_0

Medication B
Treated
 $T=1$

Temperature = ?
 Y_1

Only one outcome is observed

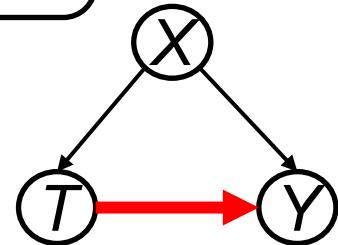
Fundamentally challenging problem

Features

Age: 32
Sex: F
Blood group: A+
Race: Asian
Blood Sugar: High
Temperature: 100° F
....



I am having a mild fever. Shall I take medicine or not?



Counterfactual

Medication A
Control
 $T=0$

Temperature = ?
 Y_0

Medication B
Treated
 $T=1$

Temperature = ?
 Y_1

Only one outcome is observed

Causal Inference is challenging due to counterfactuals



Solutions

1. Randomized control trial
2. Propensity score matching
3. Inverse probability weighting

Sibbald et al. [BMJ, 1998]
Tian et al.[2018]
Austin et al. [2011]

Causal Inference is challenging due to counterfactuals



Solutions

1. Randomized control trial
2. Propensity score matching
3. Inverse probability weighting

Drawbacks

1. Unethical, not feasible
2. Rest are mostly linear

Sibbald et al. [BMJ, 1998]
Tian et al.[2018]
Austin et al. [2011]

Causal Inference is challenging due to counterfactuals



Solutions

1. Randomized control trial
2. Propensity score matching
3. Inverse probability weighting

Drawbacks

1. Unethical, not feasible
2. Rest are mostly linear

Growth in deep learning

Deep Counterfactual Networks with Propensity-Dropout

Ahmed M. Alaa,¹ Michael Weisz,² Mihaela van der Schaar^{1,2,3}

Causal Effect Inference with Deep Latent-Variable Models

GANITE: ESTIMATION OF INDIVIDUALIZED TREATMENT EFFECTS USING GENERATIVE ADVERSARIAL NETS

Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit, Fredrik D. Johansson, David Sontag

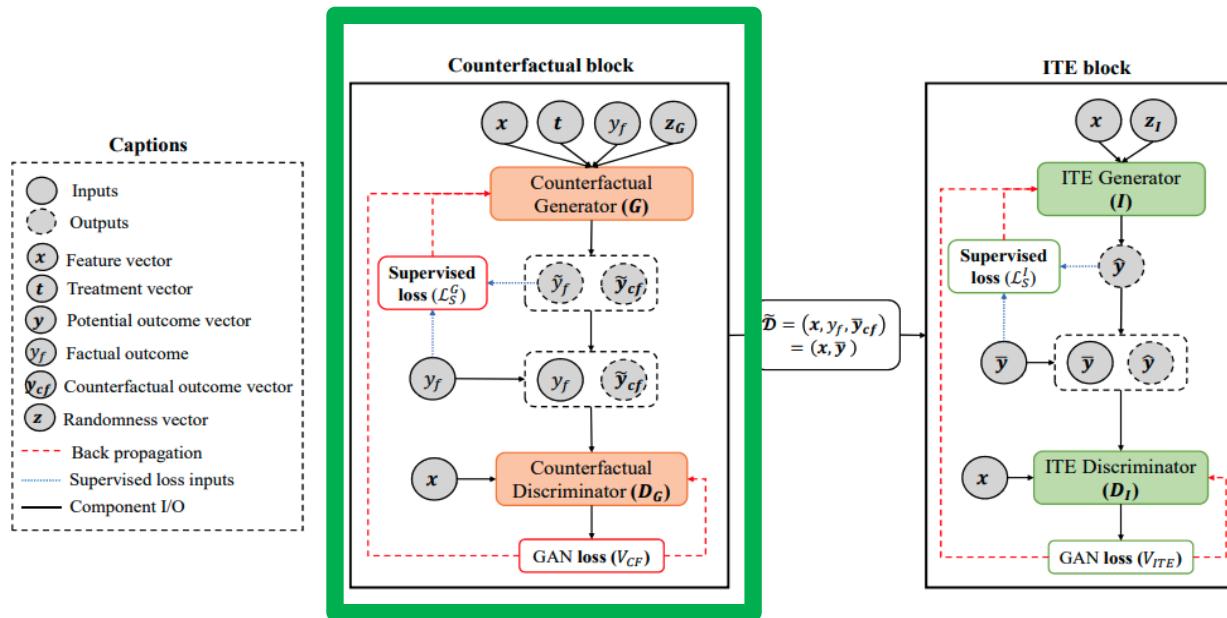
Sibbald et al. [BMJ, 1998]

Tian et al.[2018]

Austin et al. [2011]

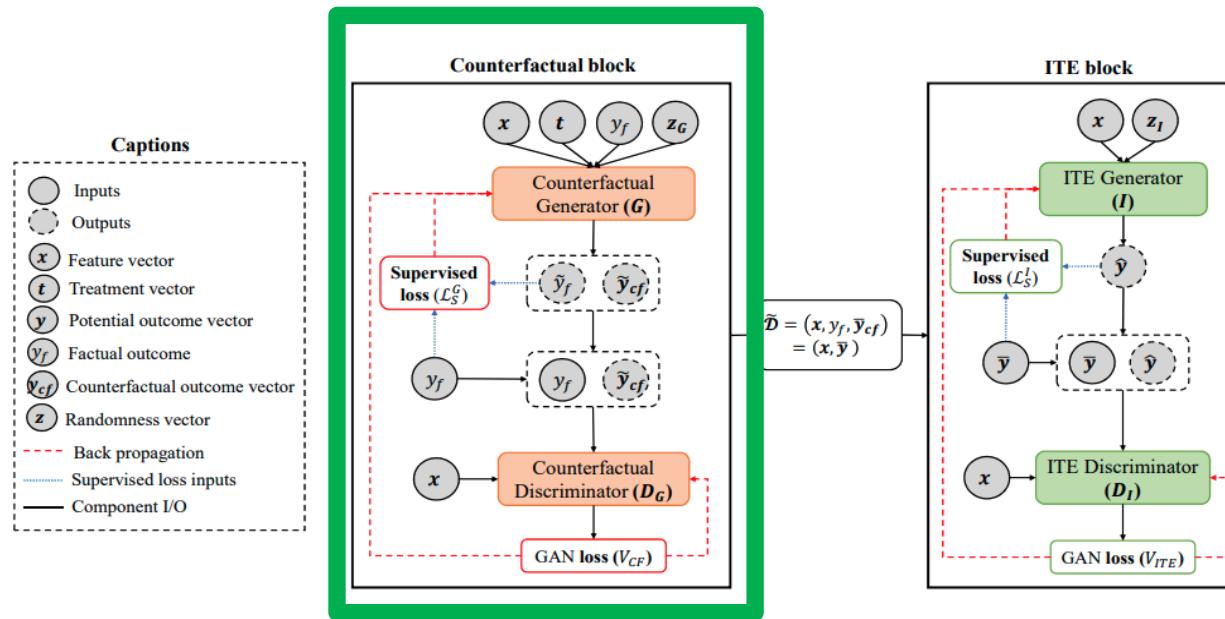
The are plenty more....

Estimation of counterfactuals using deep learning



Yoon et al. [ICLR, 2018]

Prior research motivates to use latent encoding than data



Dumoulin et al. [ICLR, 2017]
Yu et al. [ICJAI, 2019]

Our contribution



1. Incorporate of an underlying causal structure to approximate the data generation process.
2. Infer the latent variables, responsible for generating data using a VAE.
3. Generate counterfactual outcomes using a GAN with variational information theoretic regularization.
4. Estimate the individual treatment effect by minimizing the factual and counterfactual outcomes both.
5. Utilize doubly robust regularizer for faster convergence.

Problem formulation

Used potential outcome framework by Rubin.

The complete tuple $\{X_i, T_i, Y_i\}$, for $i=1\dots N$

Y_i^0 and Y_i^1 are the potential outcomes for treatment $T_i=0$ and $T_i=1$

The ITE for the subject i with covariates $X_i = \mathbf{x}$ is defined as,

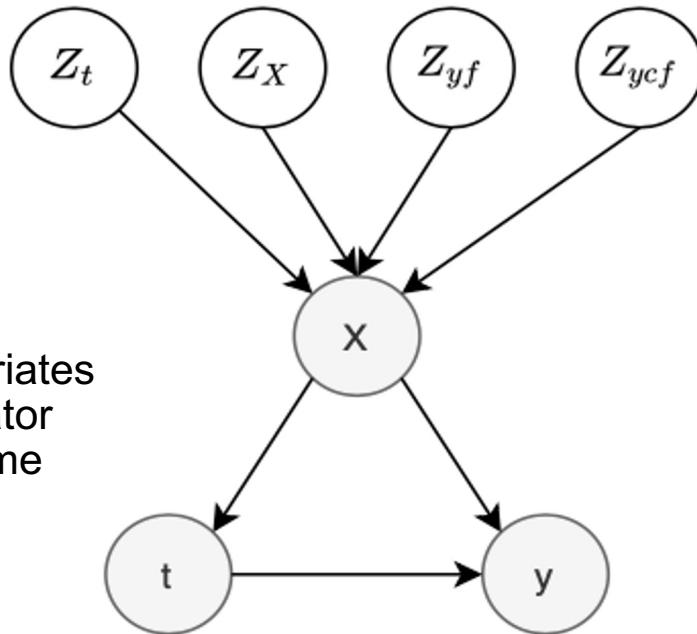
$$\tau(\mathbf{x}) = \mathbb{E}[Y_i^1 - Y_i^0 | X_i = \mathbf{x}]$$

Assumption:

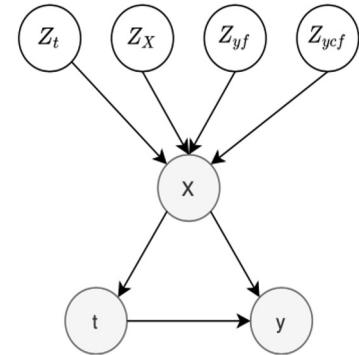
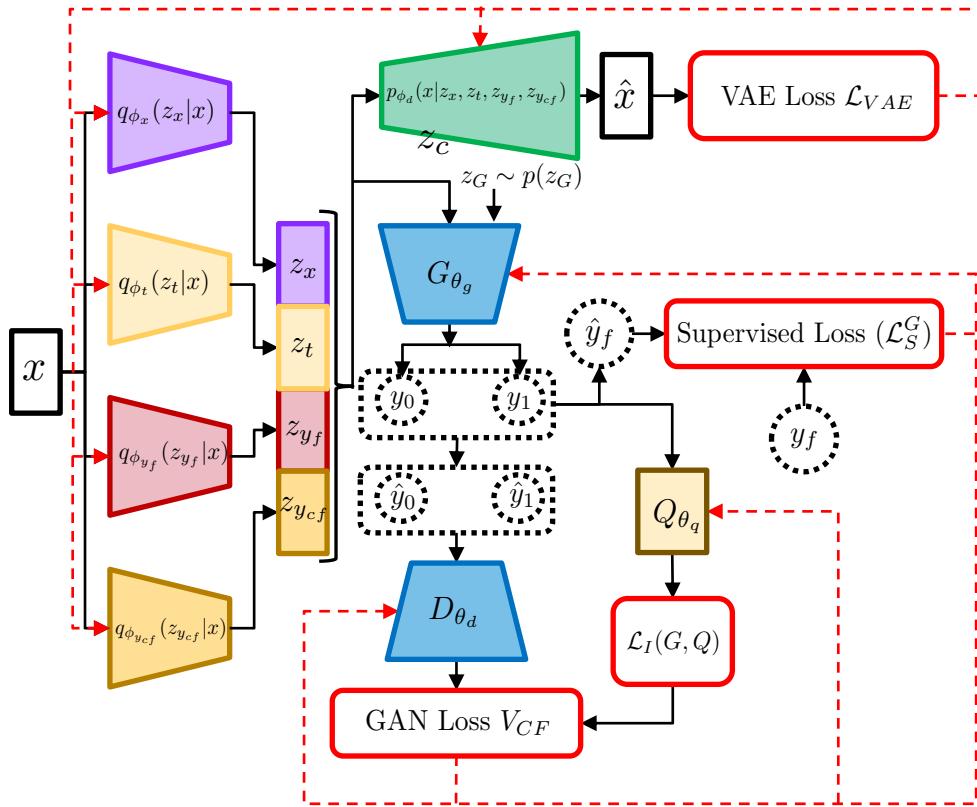
Followed strongly ignorable treatment assignment assumption (SITA), defined as, $\{Y^1, Y^0\} \perp T | X$

Contribution #1: Causal structure

- Z: Latent space
- X: Observed covariates
- t: Treatment indicator
- Y: Potential outcome



Contribution #2: Infer the latent variables



$$z_x \sim q_{\phi_x}(z_x|x)$$

$$z_t \sim q_{\phi_t}(z_t|x)$$

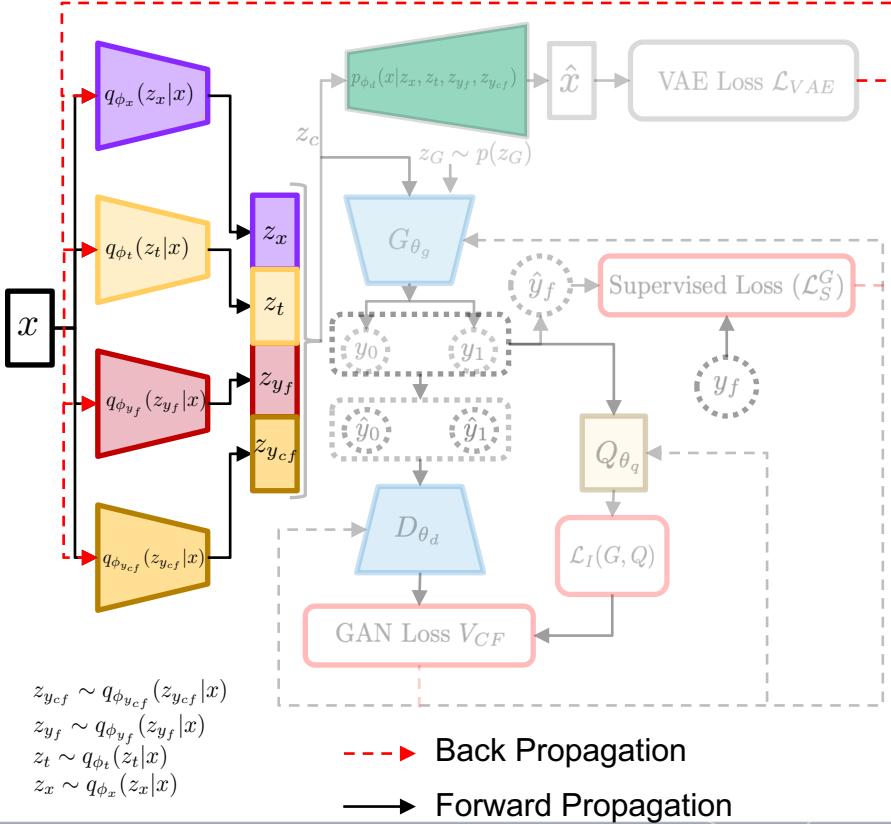
$$z_{y_f} \sim q_{\phi_{y_f}}(z_{y_f}|x)$$

$$z_{y_cf} \sim q_{\phi_{y_cf}}(z_{y_cf}|x)$$

→ Back Propagation

→ Forward Propagation

Contribution #2: Infer the latent variables



All the latent factors - z , are assumed to have a prior gaussian distributions defined as,

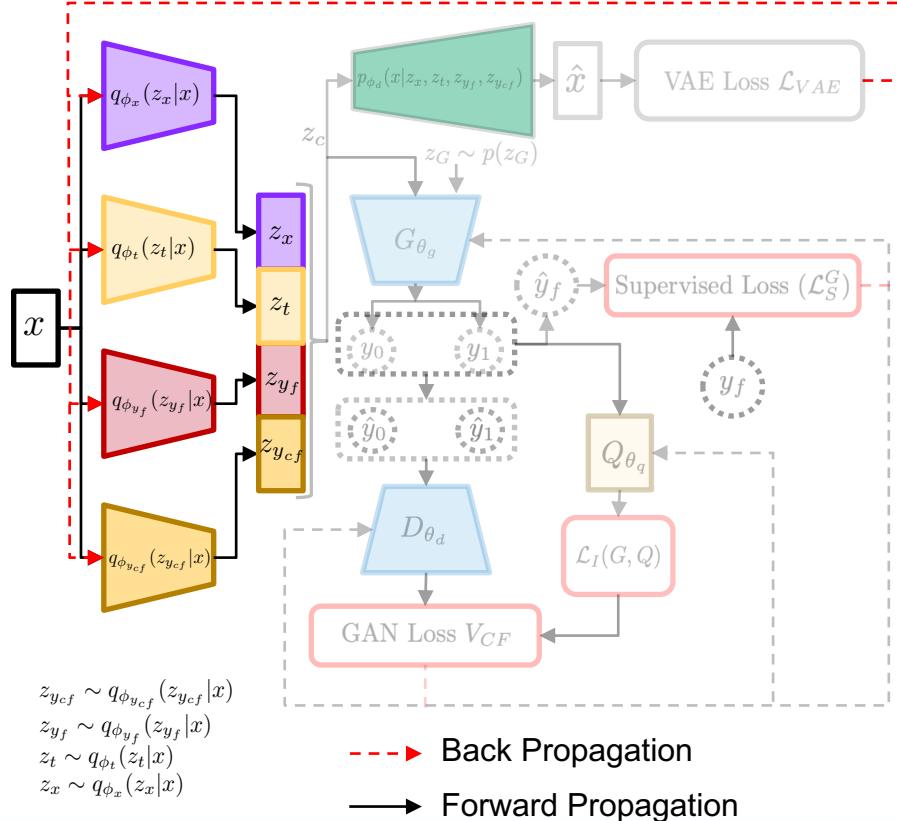
$$p(\mathbf{z}_x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(z_{xi}|0, 1)$$

$$p(\mathbf{z}_t) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(z_{ti}|0, 1)$$

$$p(\mathbf{z}_{yf}) = \prod_{i=1}^{D_{z_{yf}}} \mathcal{N}(z_{yf_i}|0, 1)$$

$$p(\mathbf{z}_{ycf}) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(z_{ycf_i}|0, 1)$$

Contribution #2: Infer the latent variables



All the latent factors - z , are assumed to have a prior gaussian distributions defined as,

$$p(\mathbf{z}_x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(z_{x_i}|0, 1) \quad p(\mathbf{z}_t) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(z_{t_i}|0, 1) \quad p(\mathbf{z}_{y_f}) = \prod_{i=1}^{D_{z_{y_f}}} \mathcal{N}(z_{y_f i}|0, 1) \quad p(\mathbf{z}_{ycf}) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(z_{ycf i}|0, 1)$$

The variational posteriors of the inference model is defined as,

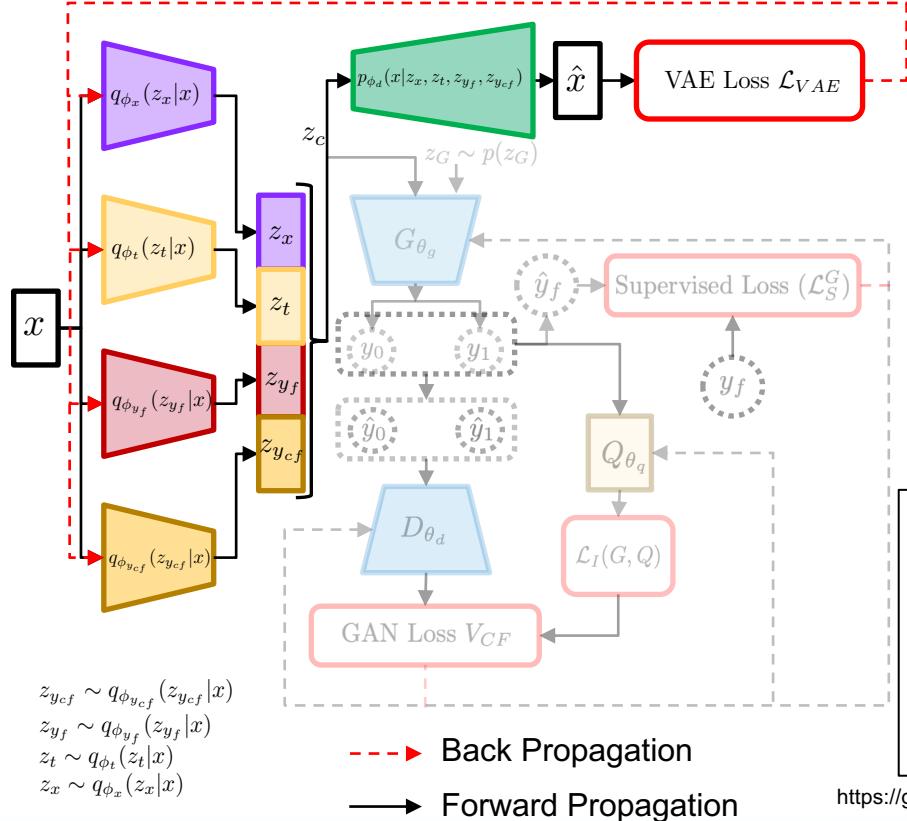
$$q_{\phi_x}(\mathbf{z}_x|\mathbf{x}) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(\mu = \hat{\mu}_x, \sigma^2 = \hat{\sigma}_x^2)$$

$$q_{\phi_t}(\mathbf{z}_t|\mathbf{x}) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(\mu = \hat{\mu}_t, \sigma^2 = \hat{\sigma}_t^2)$$

$$q_{\phi_{y_f}}(\mathbf{z}_{y_f}|\mathbf{x}) = \prod_{i=1}^{D_{z_{y_f}}} \mathcal{N}(\mu = \hat{\mu}_{y_f}, \sigma^2 = \hat{\sigma}_{y_f}^2)$$

$$q_{\phi_{ycf}}(\mathbf{z}_{ycf}|\mathbf{x}) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(\mu = \hat{\mu}_{ycf}, \sigma^2 = \hat{\sigma}_{ycf}^2)$$

Contribution #2: Infer the latent variables



All the latent factors - z , are assumed to have a prior gaussian distributions defined as,

$$p(\mathbf{z}_x) = \prod_{i=1}^{D_{zx}} \mathcal{N}(z_{xi}|0, 1) \quad p(\mathbf{z}_t) = \prod_{i=1}^{D_{zt}} \mathcal{N}(z_{ti}|0, 1) \quad p(\mathbf{z}_{yf}) = \prod_{i=1}^{D_{zyf}} \mathcal{N}(z_{yfi}|0, 1) \quad p(\mathbf{z}_{ycf}) = \prod_{i=1}^{D_{zycf}} \mathcal{N}(z_{ycfi}|0, 1)$$

The variational posteriors of the inference model is defined as,

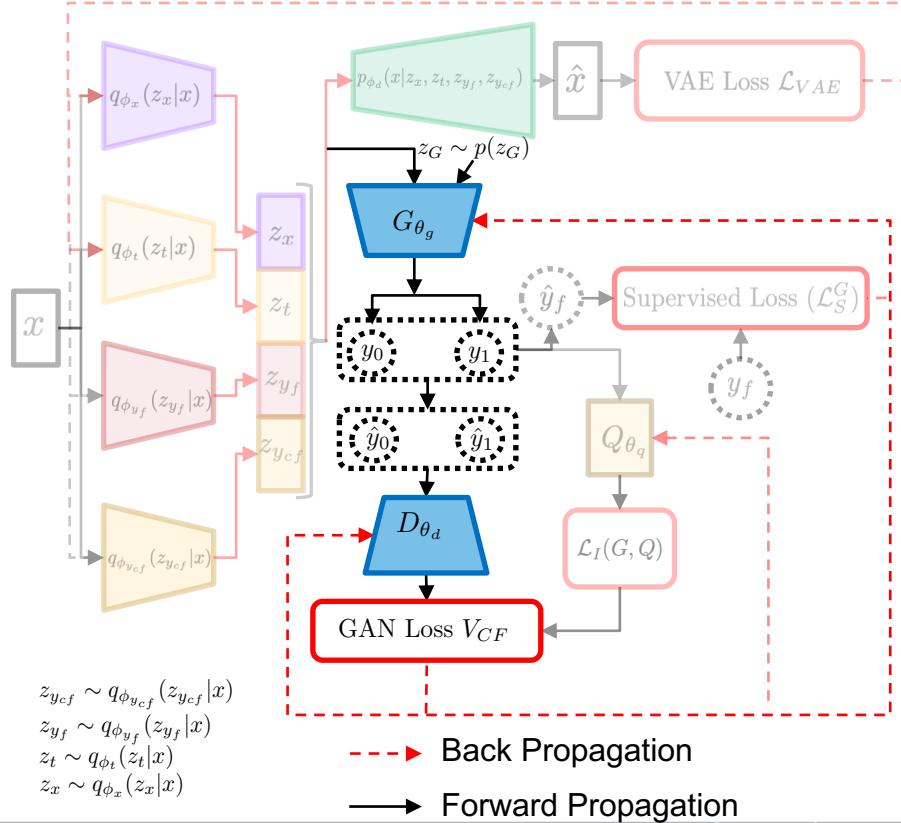
$$\begin{aligned} q_{\phi_x}(\mathbf{z}_x|\mathbf{x}) &= \prod_{i=1}^{D_{zx}} \mathcal{N}(\mu = \hat{\mu}_x, \sigma^2 = \hat{\sigma}_x^2) \\ q_{\phi_t}(\mathbf{z}_t|\mathbf{x}) &= \prod_{i=1}^{D_{zt}} \mathcal{N}(\mu = \hat{\mu}_t, \sigma^2 = \hat{\sigma}_t^2) \\ q_{\phi_{yf}}(\mathbf{z}_{yf}|\mathbf{x}) &= \prod_{i=1}^{D_{zyf}} \mathcal{N}(\mu = \hat{\mu}_{yf}, \sigma^2 = \hat{\sigma}_{yf}^2) \\ q_{\phi_{ycf}}(\mathbf{z}_{ycf}|\mathbf{x}) &= \prod_{i=1}^{D_{zycf}} \mathcal{N}(\mu = \hat{\mu}_{ycf}, \sigma^2 = \hat{\sigma}_{ycf}^2) \end{aligned}$$

The overall ELBO function to optimize

$$\begin{aligned} \mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; \mathbf{x}, \mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf}) \\ = \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}} [\log p_{\phi_d}(\mathbf{x}|\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf})] \\ - KL(q_{\phi_x}(\mathbf{z}_x|\mathbf{x})||p_{\phi_d}(\mathbf{z}_x)) - KL(q_{\phi_t}(\mathbf{z}_t|\mathbf{x})||p_{\phi_d}(\mathbf{z}_t)) \\ - KL(q_{\phi_{yf}}(\mathbf{z}_{yf}|\mathbf{x})||p_{\phi_d}(\mathbf{z}_{yf})) - KL(q_{\phi_{ycf}}(\mathbf{z}_{ycf}|\mathbf{x})||p_{\phi_d}(\mathbf{z}_{ycf})) \end{aligned}$$

https://github.com/Shantanu48114860/DR-VIDAL-AMIA-22/blob/main/DR_VIDAL_AMIA-Supp.pdf

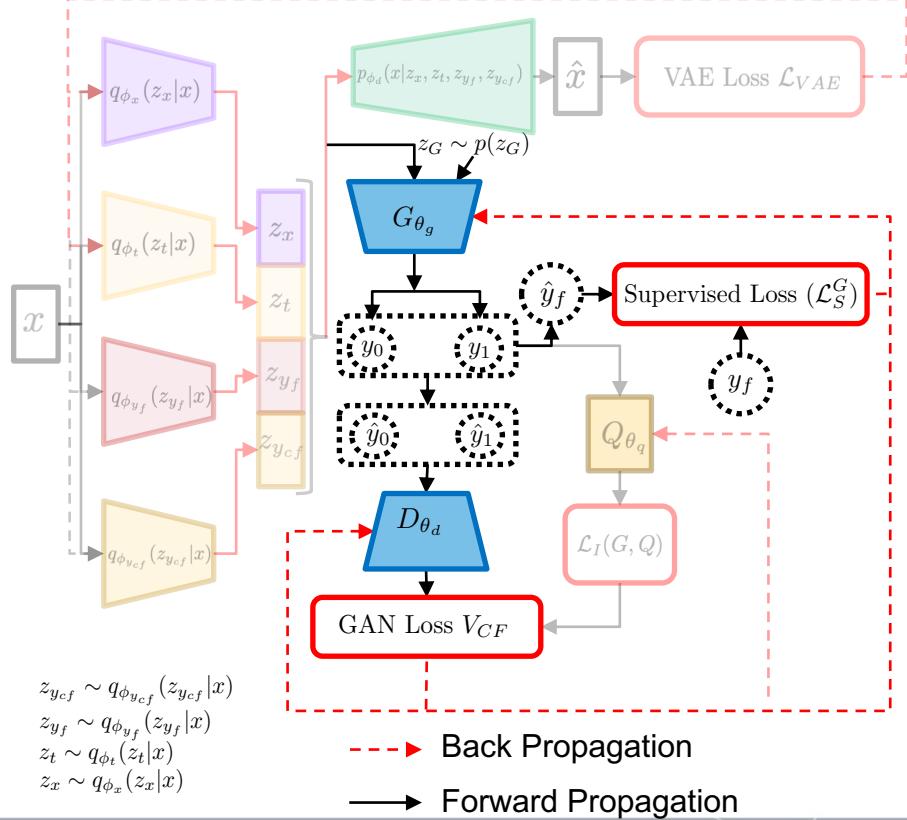
Contribution #3: Generate counterfactuals



Following GANITE, the optimization function of the GAN block,

$$V_{GAN}(G, D) = \mathbb{E}_{\mathbf{x}, \mathbf{z}_G, \mathbf{z}_c} [t^T \log D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)) + (1-t)^T \log(1 - D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)))]$$

Contribution #3: Generate counterfactuals



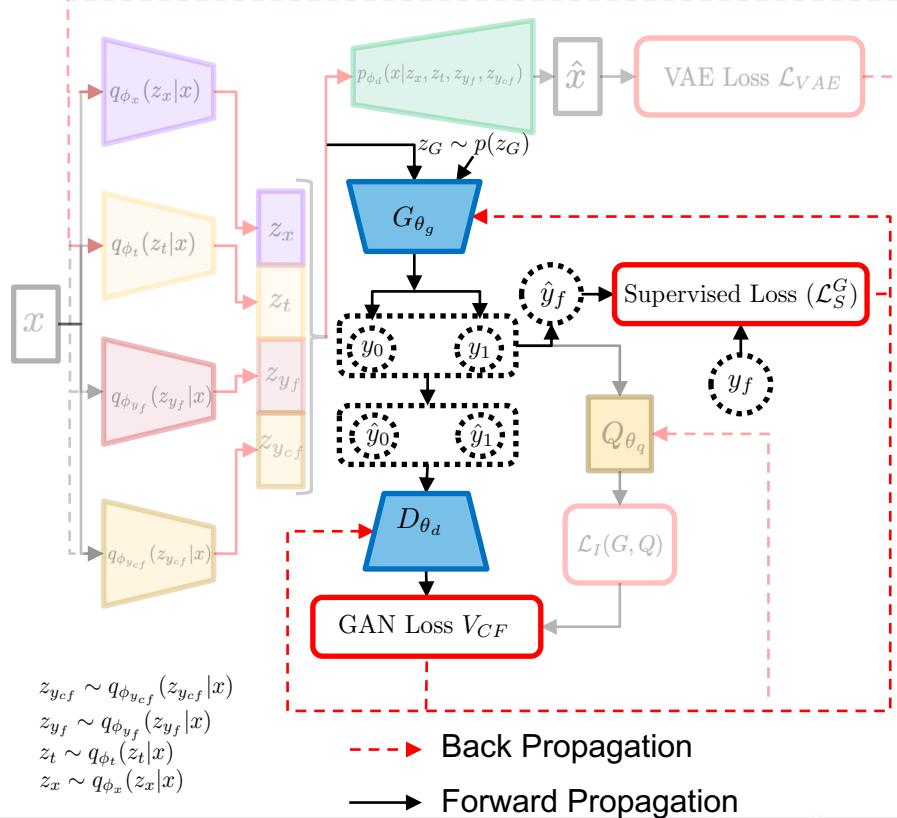
Following GANITE, the optimization function of the GAN block,

$$V_{GAN}(G, D) = \mathbb{E}_{\mathbf{x}, \mathbf{z}_G, \mathbf{z}_c} [t^T \log D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)) + (1-t)^T \log (1 - D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)))]$$

The supervised loss is defined as,

$$\mathcal{L}_S^G(y_f, \hat{y}_f) = \frac{1}{n} \sum_{i=1}^n (y_f(i) - \hat{y}_f(i))^2$$

Contribution #3: Generate counterfactuals



Following GANITE, the optimization function of the GAN block,

$$V_{GAN}(G, D) = \mathbb{E}_{\mathbf{x}, \mathbf{z}_G, \mathbf{z}_c} [t^T \log D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)) + (1-t)^T \log (1 - D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)))]$$

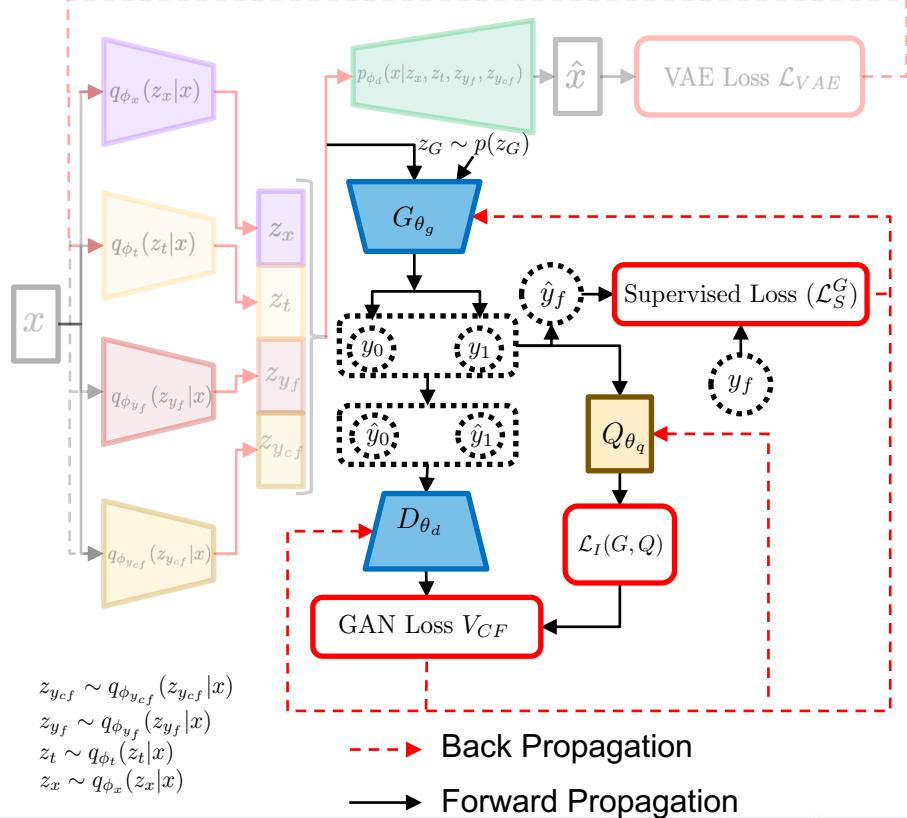
The supervised loss is defined as,

$$\mathcal{L}_S^G(y_f, \hat{y}_f) = \frac{1}{n} \sum_{i=1}^n (y_f(i) - \hat{y}_f(i))^2$$

The complete loss of the counterfactual GAN block is defined as,

$$V_{CF}(G, D) = V_{GAN}(G, D) + \gamma \mathcal{L}_S^G(y_f, \hat{y}_f)$$

Contribution #3: Generate counterfactuals



Following GANITE, the optimization function of the GAN block,

$$V_{GAN}(G, D) = \mathbb{E}_{\mathbf{x}, \mathbf{z}_G, \mathbf{z}_c} [t^T \log D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)) + (1-t)^T \log (1 - D(\mathbf{x}, G(\mathbf{z}_G, \mathbf{z}_c)))]$$

The supervised loss is defined as,

$$\mathcal{L}_S^G(y_f, \hat{y}_f) = \frac{1}{n} \sum_{i=1}^n (y_f(i) - \hat{y}_f(i))^2$$

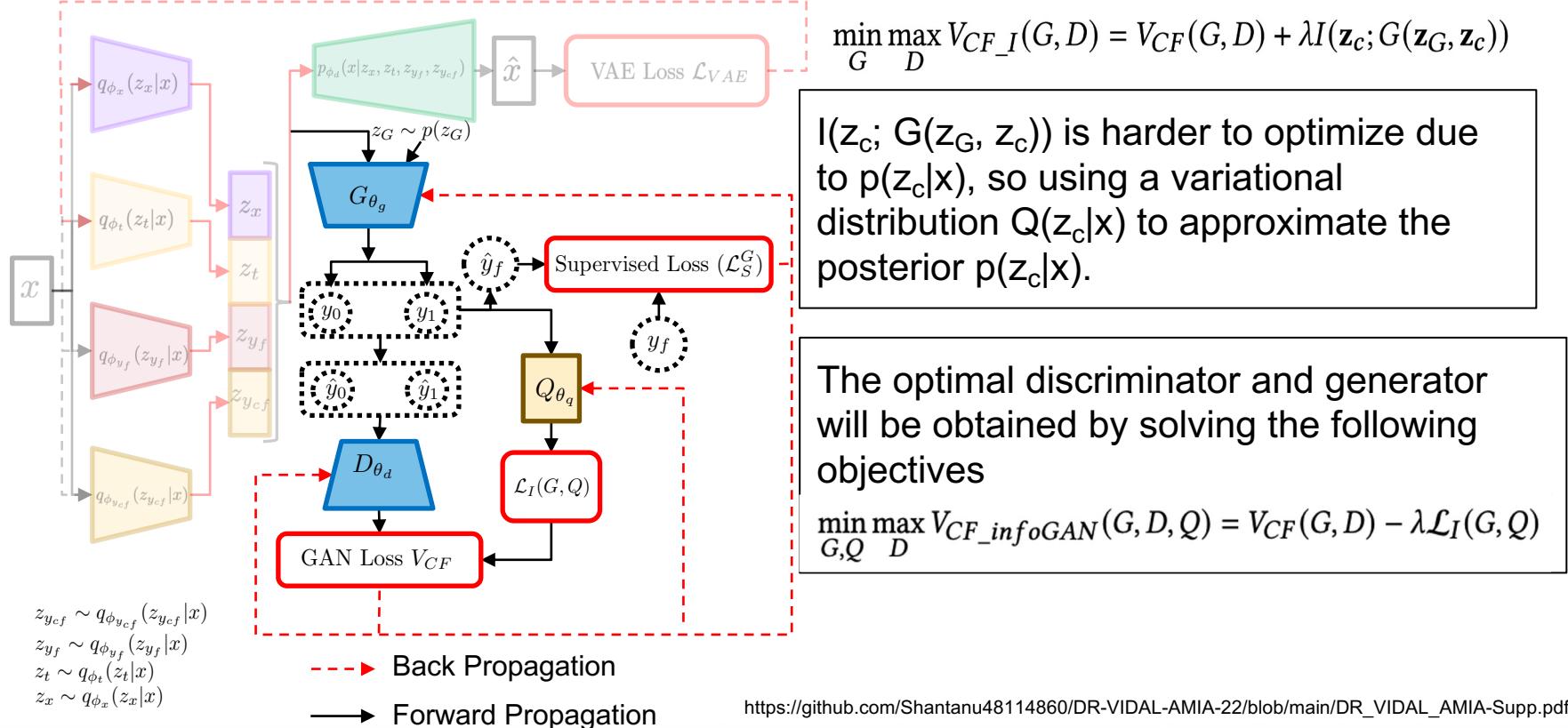
The complete loss of the counterfactual GAN block is defined as,

$$V_{CF}(G, D) = V_{GAN}(G, D) + \gamma \mathcal{L}_S^G(y_f, \hat{y}_f)$$

Maximize $I(z_c; G(z_G, z_c))$ to solve the optimization function

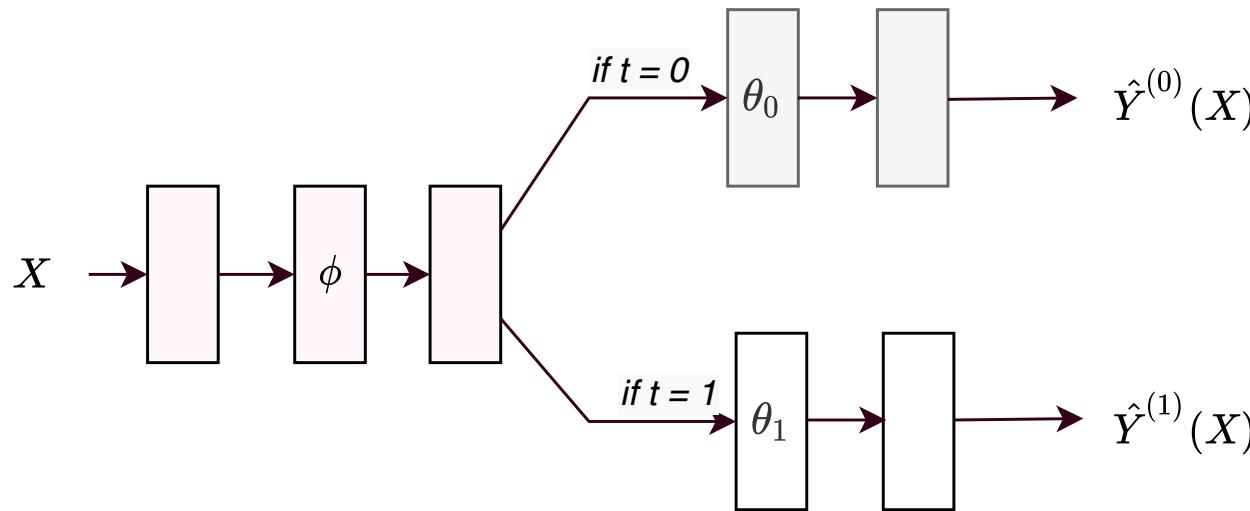
$$\min_G \max_D V_{CF_I}(G, D) = V_{CF}(G, D) + \lambda I(z_c; G(z_G, z_c))$$

Contribution #3: Generate counterfactuals



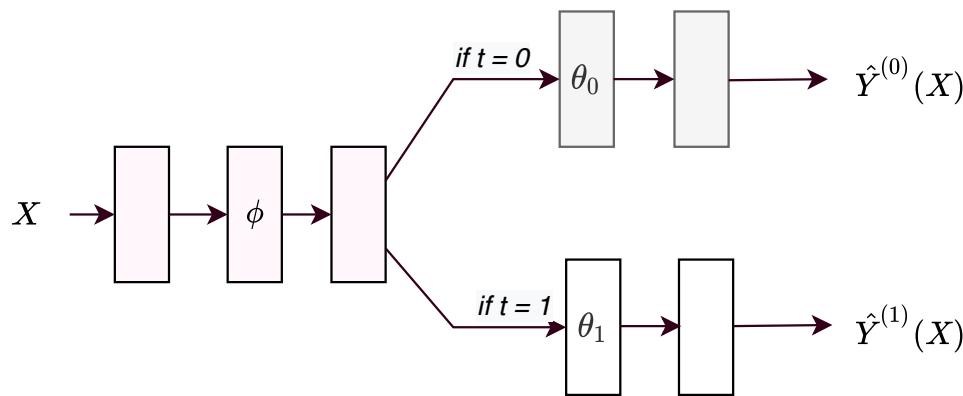
https://github.com/Shantanu48114860/DR-VIDAL-AMIA-22/blob/main/DR_VIDAL_AMIA-Supp.pdf

Contribution #4: Estimate ITE by minimizing the factuals and counterfactuals



Shalit et al. [ICML, 2017]

Contribution #4: Estimate ITE by minimizing the factuals and counterfactuals



The factual and the counterfactual outcomes were estimated as,

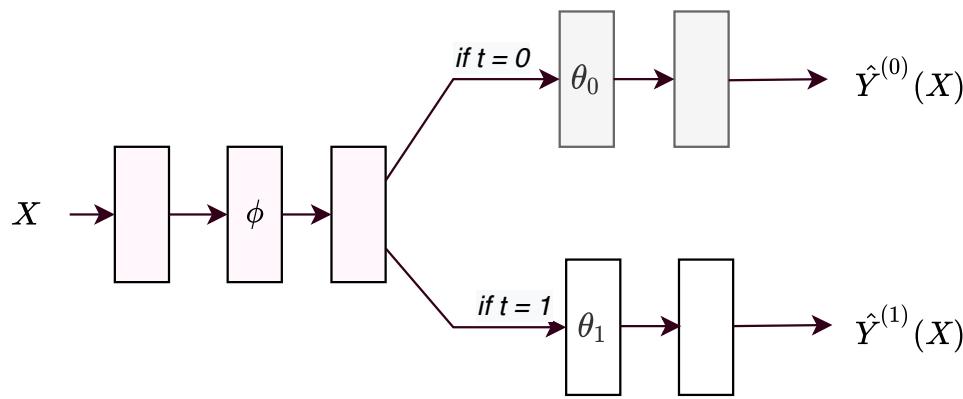
$$\hat{y}_i^{(0)} = f_{\theta_0}(f_{\phi}(\mathbf{x}_i)) \quad \text{if } t_i = 0$$

$$\hat{y}_i^{(1)} = f_{\theta_1}(f_{\phi}(\mathbf{x}_i)) \quad \text{if } t_i = 1$$

$$\hat{y}_f^{(i)} = t_i \hat{y}_i^{(1)} + (1 - t_i) \hat{y}_i^{(0)}$$

$$\hat{y}_{cf}^{(i)} = (1 - t_i) \hat{y}_i^{(1)} + t_i \hat{y}_i^{(0)}$$

Contribution #4: Estimate ITE by minimizing the factuals and counterfactuals



The factual and the counterfactual outcomes were estimated as,

$$\hat{y}_i^{(0)} = f_{\theta_0}(f_{\phi}(\mathbf{x}_i)) \quad \text{if } t_i = 0$$

$$\hat{y}_i^{(1)} = f_{\theta_1}(f_{\phi}(\mathbf{x}_i)) \quad \text{if } t_i = 1$$

$$\hat{y}_f^{(i)} = t_i \hat{y}_i^{(1)} + (1 - t_i) \hat{y}_i^{(0)}$$

$$\hat{y}_{cf}^{(i)} = (1 - t_i) \hat{y}_i^{(1)} + t_i \hat{y}_i^{(0)}$$

$$\mathcal{L}_i^p(\theta_1, \theta_0, \phi) = (\hat{y}_f^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf}^{(i)} - y_{cf}^{(i)})^2$$

Shalit et al. [ICML, 2017]

Contribution #5: Doubly robust ITE Estimation



Using propensity score, the doubly robust estimation of causal effect is defined as,

$$\hat{\delta}_{DR} = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i t_i - (t_i - \pi(x_i))\mu(x_i, t_i)}{\pi(x_i)} - \frac{y_i(1 - t_i) - (t_i - \pi(x_i))\mu(x_i, t_i)}{1 - \pi(x_i)} \right]$$

where,

$$\mu(x, t) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \cdots + \hat{\alpha}_n x_n + \hat{\delta}t$$

Jonsson et al. [American Journal of epidemiology, 2011]

Contribution #5: Doubly robust ITE Estimation



Using propensity score, the doubly robust estimation of causal effect is defined as,

$$\hat{\delta}_{DR} = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i t_i - (t_i - \pi(x_i))\mu(x_i, t_i)}{\pi(x_i)} - \frac{y_i(1 - t_i) - (t_i - \pi(x_i))\mu(x_i, t_i)}{1 - \pi(x_i)} \right]$$

where,

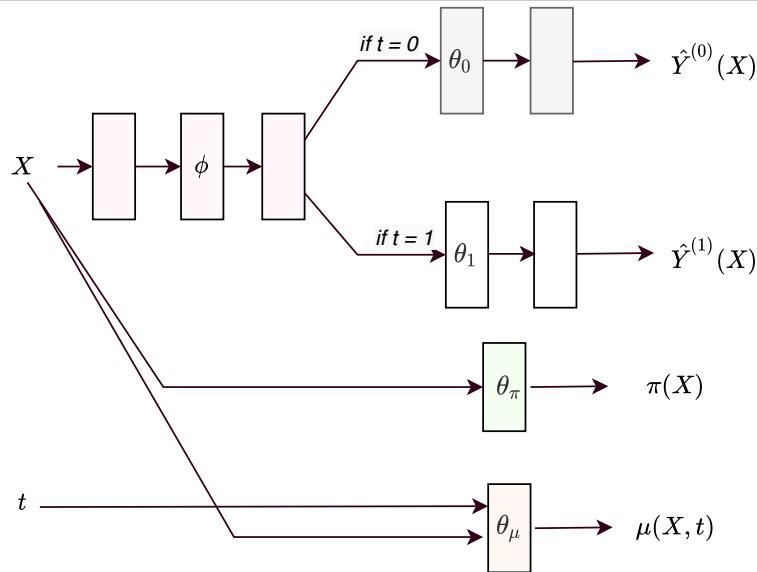
$$\mu(x, t) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \cdots + \hat{\alpha}_n x_n + \hat{\delta}t$$

The propensity score is defined as,

$$\pi(\mathbf{x}) = P(T = 1 | \mathbf{X} = \mathbf{x}).$$

Jonsson et al. [American Journal of epidemiology, 2011]

Contribution #5: Doubly robust ITE Estimation



Using propensity score, the doubly robust estimation of causal effect is defined as,

$$\hat{\delta}_{DR} = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i t_i - (t_i - \pi(x_i))\mu(x_i, t_i)}{\pi(x_i)} - \frac{y_i(1 - t_i) - (t_i - \pi(x_i))\mu(x_i, t_i)}{1 - \pi(x_i)} \right]$$

where,

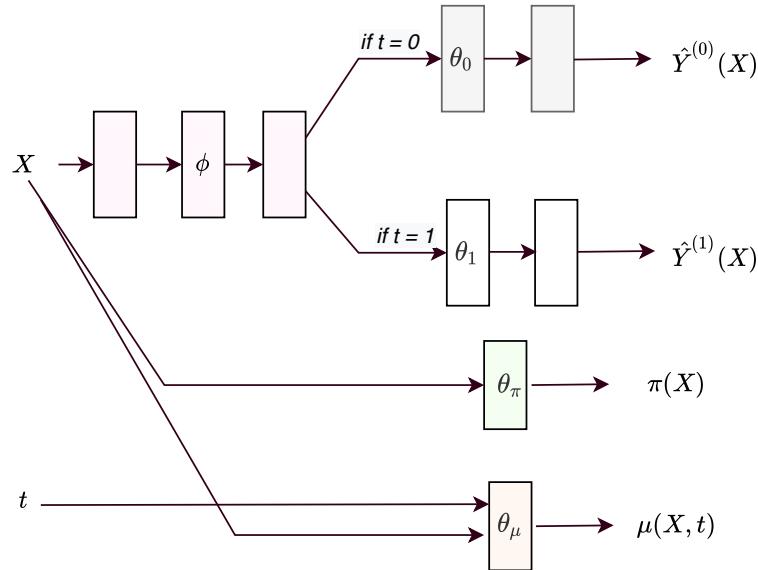
$$\mu(x, t) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \cdots + \hat{\alpha}_n x_n + \hat{\delta}t$$

The propensity score is defined as,

$$\pi(\mathbf{x}) = P(T = 1 | \mathbf{X} = \mathbf{x}).$$

Jonsson et al. [American Journal of epidemiology, 2011]

Contribution #5: Doubly robust ITE Estimation



Using propensity score, the doubly robust estimation of causal effect is defined as,

$$\hat{\delta}_{DR} = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i t_i - (t_i - \pi(x_i))\mu(x_i, t_i)}{\pi(x_i)} - \frac{y_i(1 - t_i) - (t_i - \pi(x_i))\mu(x_i, t_i)}{1 - \pi(x_i)} \right]$$

where,

$$\mu(x, t) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \dots + \hat{\alpha}_n x_n + \hat{\delta}t$$

The predicted loss to be optimized as,

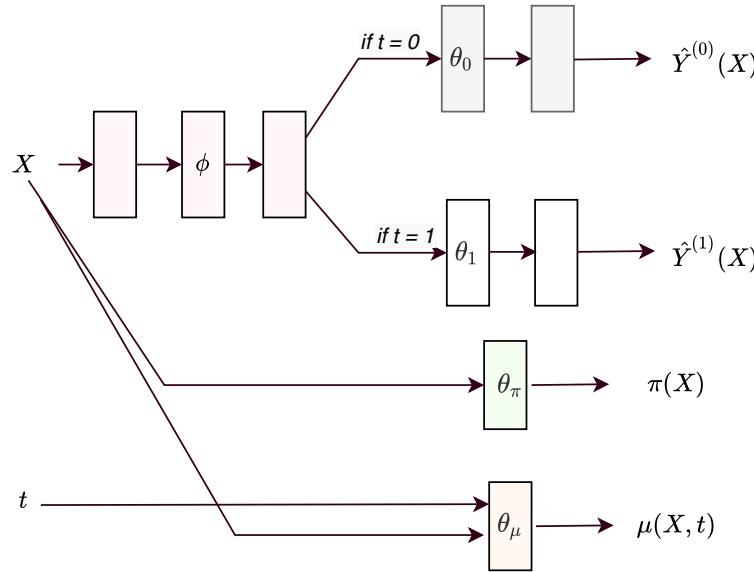
$$\begin{aligned} \mathcal{L}_i^p(\theta_1, \theta_0, \phi) &= (\hat{y}_f^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf}^{(i)} - y_{cf}^{(i)})^2 \\ &\quad + \alpha \text{BinaryCrossEntropy}(\pi(x_i), t_i) \end{aligned}$$

The propensity score is defined as,

$$\pi(\mathbf{x}) = P(T = 1 | \mathbf{X} = \mathbf{x}).$$

Jonsson et al. [American Journal of epidemiology, 2011]

Contribution #5: Doubly robust ITE Estimation

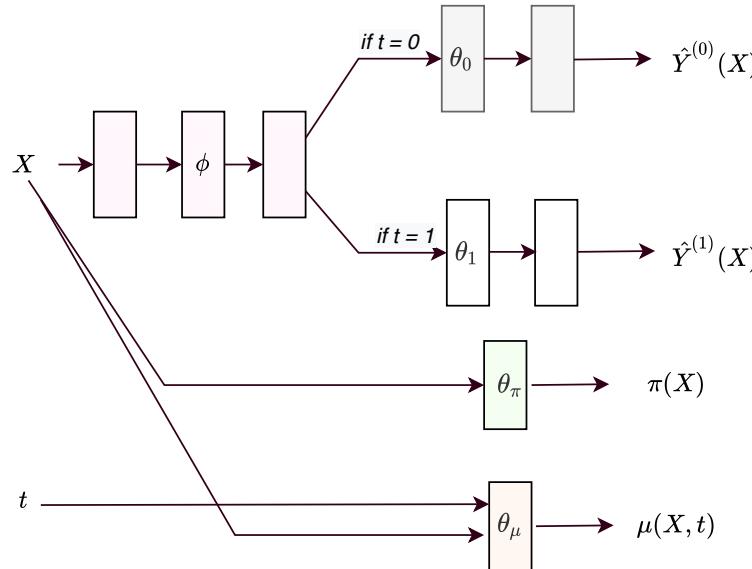


The factual and the counterfactual doubly robust outcomes were estimated as,

$$\begin{aligned}\hat{y}_{f_{DR}}^{(i)} &= t_i \left[\frac{t_i \hat{y}_i^{(1)} - (t_i - \pi(\mathbf{x}_i)\mu(\mathbf{x}_i, t_i))}{\pi(\mathbf{x}_i)} \right] \\ &\quad + (1 - t_i) \left[\frac{(1 - t_i) \hat{y}_i^{(0)} - (t_i - \pi(\mathbf{x}_i)\mu(\mathbf{x}_i, t_i))}{1 - \pi(\mathbf{x}_i)} \right] \\ \hat{y}_{cf_{DR}}^{(i)} &= (1 - t_i) \left[\frac{(1 - t_i) \hat{y}_i^{(1)} - (t_i - \pi(\mathbf{x}_i)\mu(\mathbf{x}_i, t_i))}{\pi(\mathbf{x}_i)} \right] \\ &\quad + t_i \left[\frac{t_i \hat{y}_i^{(0)} - (t_i - \pi(\mathbf{x}_i)\mu(\mathbf{x}_i, t_i))}{1 - \pi(\mathbf{x}_i)} \right]\end{aligned}$$

Jonsson et al. [American Journal of epidemiology, 2011]

Contribution #5: Doubly robust ITE Estimation



The factual and the counterfactual doubly robust outcomes were estimated as,

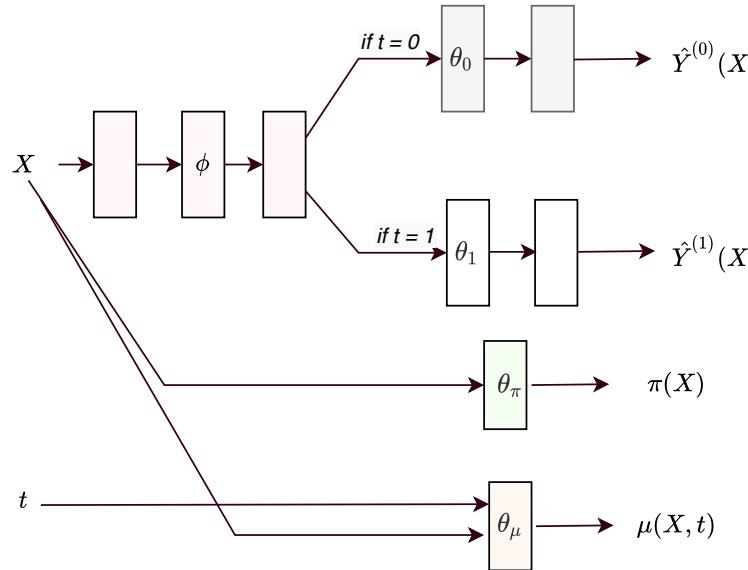
$$\begin{aligned}\hat{y}_{f_{DR}}^{(i)} &= t_i \left[\frac{t_i \hat{y}_i^{(1)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{\pi(\mathbf{x}_i)} \right] \\ &\quad + (1 - t_i) \left[\frac{(1 - t_i)\hat{y}_i^{(0)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{1 - \pi(\mathbf{x}_i)} \right] \\ \hat{y}_{cf_{DR}}^{(i)} &= (1 - t_i) \left[\frac{(1 - t_i)\hat{y}_i^{(1)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{\pi(\mathbf{x}_i)} \right] \\ &\quad + t_i \left[\frac{t_i \hat{y}_i^{(0)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{1 - \pi(\mathbf{x}_i)} \right]\end{aligned}$$

The doubly robust loss is optimized as ,

$$\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi) = (\hat{y}_{f_{DR}}^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf_{DR}}^{(i)} - y_{cf}^{(i)})^2$$

Jonsson et al. [American Journal of epidemiology, 2011]

Contribution #5: Doubly robust ITE Estimation



The complete loss to estimate ITE to be optimized as,

$$\mathcal{L}^{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi) = \frac{1}{n} \sum_{i=1}^n \left(\mathcal{L}_i^p + \beta \mathcal{L}_i^{DR} \right)$$

The factual and the counterfactual doubly robust outcomes were estimated as,

$$\begin{aligned}\hat{y}_{f_{DR}}^{(i)} &= t_i \left[\frac{t_i \hat{y}_i^{(1)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{\pi(\mathbf{x}_i)} \right] \\ &\quad + (1 - t_i) \left[\frac{(1 - t_i)\hat{y}_i^{(0)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{1 - \pi(\mathbf{x}_i)} \right] \\ \hat{y}_{cf_{DR}}^{(i)} &= (1 - t_i) \left[\frac{(1 - t_i)\hat{y}_i^{(1)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{\pi(\mathbf{x}_i)} \right] \\ &\quad + t_i \left[\frac{t_i \hat{y}_i^{(0)} - (t_i - \pi(\mathbf{x}_i))\mu(\mathbf{x}_i, t_i))}{1 - \pi(\mathbf{x}_i)} \right]\end{aligned}$$

The doubly robust loss is optimized as ,

$$\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi) = (\hat{y}_{f_{DR}}^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf_{DR}}^{(i)} - y_{cf}^{(i)})^2$$

Jonsson et al. [American Journal of epidemiology, 2011]

Performance metrics

$$ATT = \frac{1}{|T_1 \cap E|} \sum_{x_i \in T_1 \cap E} Y_1(x_i) - \frac{1}{|T_0 \cap E|} \sum_{x_i \in C \cap E} Y_0(x_i)$$

$$\epsilon_{ATT} = \left| ATT - \frac{1}{|T_1 \cap E|} \sum_{x_i \in T_1 \cap E} \hat{Y}_1(x_i) - \hat{Y}_0(x_i) \right|$$

where T_1 , T_0 and E are the subsets corresponding to treated, controlled samples, and randomized controlled trials, respectively.

$$\epsilon_{PEHE} = \frac{1}{N} \sum_{n=0}^N \left(\mathbb{E}_{y_j(n) \sim \mu_j(n)} [y_1(n) - y_0(n)] - [\hat{y}_1(n) - \hat{y}_0(n)] \right)^2$$

$$\epsilon_{ATE} = \left\| \frac{1}{N} \sum_{n=0}^N \mathbb{E}_{y(n) \sim \mu(n)} [y(n)] - \frac{1}{N} \sum_{n=0}^N \hat{y}(n) \right\|_2^2$$

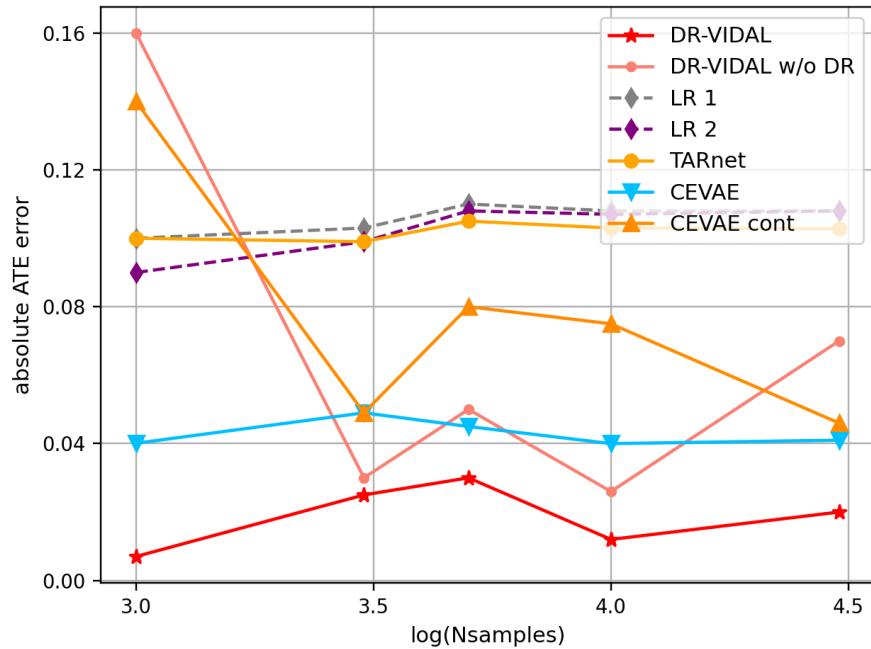
$$R_{pol}(\pi) = \frac{1}{N} \sum_{n=0}^N \left[1 - \left(\sum_{i=1}^k \left[\frac{1}{|\Pi_i \cap T_i \cap E|} \sum_{x(n) \in \Pi_i \cap T_i \cap E} y_i(n) \times \frac{|\Pi_n \cap E|}{|E|} \right] \right) \right]$$

where $\pi_i = \{\mathbf{x}(n) : i = \arg \max \hat{\mathbf{y}}\}$,
 $T_i = \{\mathbf{x}(n) : t_i(n) = 1\}$, and E is the randomized sample.

Results – Synthetic datasets 1

$$\begin{aligned} \mathbf{z}_i &\sim \text{Bern}(0.5); & \mathbf{x}_i | \mathbf{z}_i &\sim \mathcal{N}(\mathbf{z}_i, \sigma_5^2 \mathbf{z}_i + \sigma_3^2 (1 - \mathbf{z}_i)) \\ t_i | \mathbf{z}_i &\sim \text{Bern}(0.75\mathbf{z}_i + 0.25(1 - \mathbf{z}_i)) \\ \mathbf{y}_i | i, \mathbf{z}_i &\sim \text{Bern}(\text{Sigmoid}(3(\mathbf{z}_i + 2(2t_i - 1)))) \end{aligned}$$

Comparison of the performance (ATE) of DR- VIDAL vs. all other models on samples from the generative process of Synthetic dataset 1 - sample size {1000, 3000, 5000, 10000, 30000}

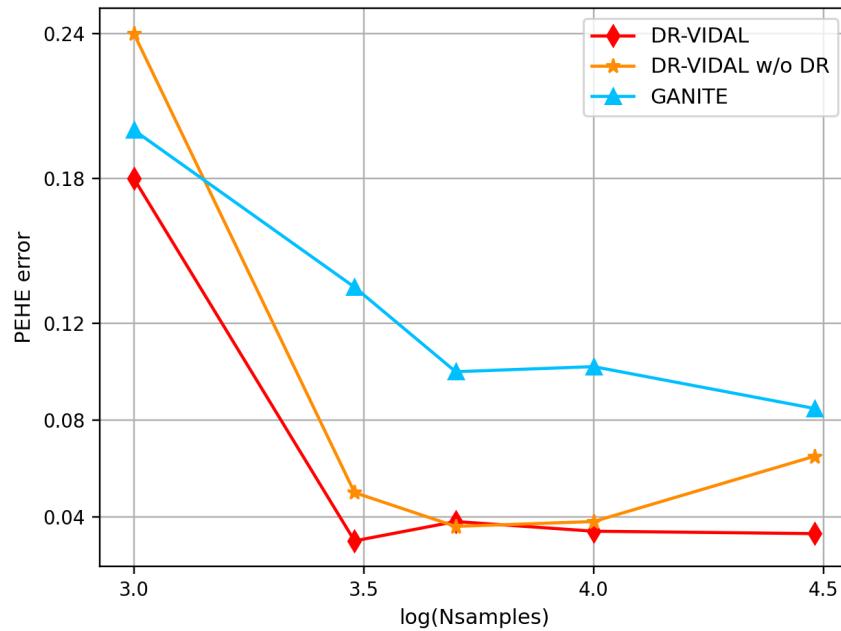


Louizos et al. [Neurips, 2017]

Results – Synthetic datasets 2

Performance comparison (PEHE) of GANITE vs. DR-VIDAL

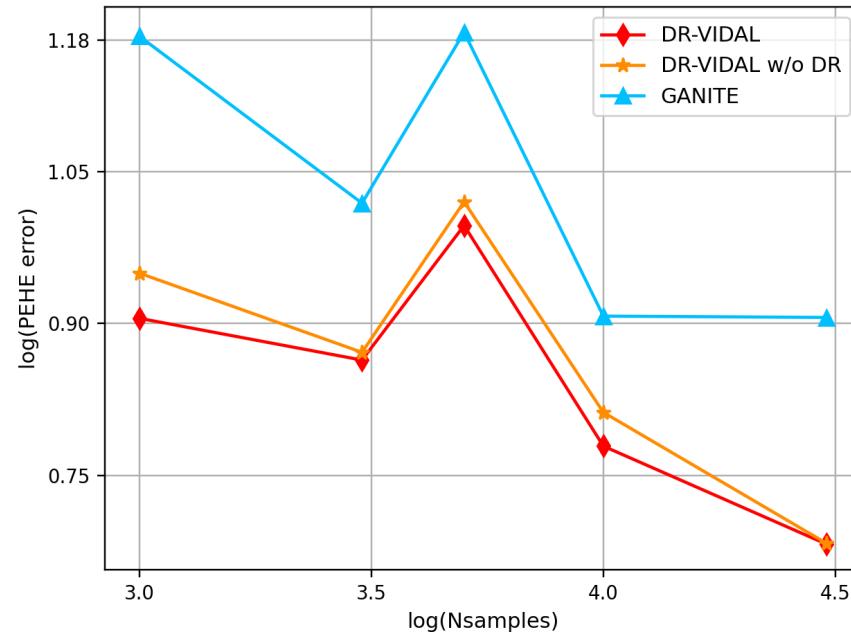
$\mathbf{z}_x \sim Bern(0.5); \quad \mathbf{z}_t \sim Bern(0.5)$
 $\mathbf{z}_{yf} \sim Bern(0.5); \quad \mathbf{z}_{ycf} \sim Bern(0.5)$
 $\mathbf{x}_x | \mathbf{z}_x \sim \mathcal{N}(\mathbf{z}_x, 5(\mathbf{z}_x) + 3(1 - \mathbf{z}_x))$
 $\mathbf{x}_t | \mathbf{z}_t \sim \mathcal{N}(\mathbf{z}_t, 2(\mathbf{z}_t) + 0.5(1 - \mathbf{z}_t))$
 $\mathbf{x}_{yf} | \mathbf{z}_{yf} \sim \mathcal{N}(\mathbf{z}_{yf}, 10(\mathbf{z}_{yf}) + 6(1 - \mathbf{z}_{yf}))$
 $\mathbf{x}_{ycf} | \mathbf{z}_{ycf} \sim \mathcal{N}(\mathbf{z}_{ycf}, 10(\mathbf{z}_{ycf}) + 6(1 - \mathbf{z}_{ycf}))$
 $\mathbf{w}_t^T \sim \mathcal{U}((-0.1, 0.1)^{10 \times 1}); \quad \mathbf{n}_t \sim \mathcal{N}(0, 0.1)$
 $\mathbf{w}_y^T \sim \mathcal{U}((-1, 1)^{10 \times 2}); \quad \mathbf{n}_y \sim \mathcal{N}(0^{2 \times 1}, 0.1 \mathbf{I}^{2 \times 2})$
 $t|x \sim Bern(Sigmoid(\mathbf{w}_t^T \mathbf{x} + \mathbf{n}_t)); \quad \mathbf{y}|x \sim \mathbf{w}_y^T \mathbf{x} + \mathbf{n}_y$



Results – Synthetic datasets 2

Performance comparison (PEHE) of GANITE vs. DR-VIDAL

$\mathbf{z}_x \sim Bern(0.5); \quad \mathbf{z}_t \sim Bern(0.5)$
 $\mathbf{z}_{yf} \sim Bern(0.5); \quad \mathbf{z}_{ycf} \sim Bern(0.5)$
 $\mathbf{x}_x | \mathbf{z}_x \sim \mathcal{N}(\mathbf{z}_x, 5(\mathbf{z}_x) + 3(1 - \mathbf{z}_x))$
 $\mathbf{x}_t | \mathbf{z}_t \sim \mathcal{N}(\mathbf{z}_t, 2(\mathbf{z}_t) + 0.5(1 - \mathbf{z}_t))$
 $\mathbf{x}_{yf} | \mathbf{z}_{yf} \sim \mathcal{N}(\mathbf{z}_{yf}, 10(\mathbf{z}_{yf}) + 6(1 - \mathbf{z}_{yf}))$
 $\mathbf{x}_{ycf} | \mathbf{z}_{ycf} \sim \mathcal{N}(\mathbf{z}_{ycf}, 10(\mathbf{z}_{ycf}) + 6(1 - \mathbf{z}_{ycf}))$
 $\mathbf{w}_t^T \sim \mathcal{U}((-0.1, 0.1)^{10 \times 1}); \quad \mathbf{n}_t \sim \mathcal{N}(0, 0.1)$
 $\mathbf{w}_y^T \sim \mathcal{U}((-1, 1)^{10 \times 2}); \quad \mathbf{n}_y \sim \mathcal{N}(0^{2 \times 1}, 0.1 \mathbf{I}^{2 \times 2})$
 $t|x \sim Bern(Sigmoid(\mathbf{w}_t^T \mathbf{x} + \mathbf{n}_t)); \quad \mathbf{y}|x \sim \mathbf{w}_y^T \mathbf{x} + \mathbf{n}_y$



Results – Real world datasets

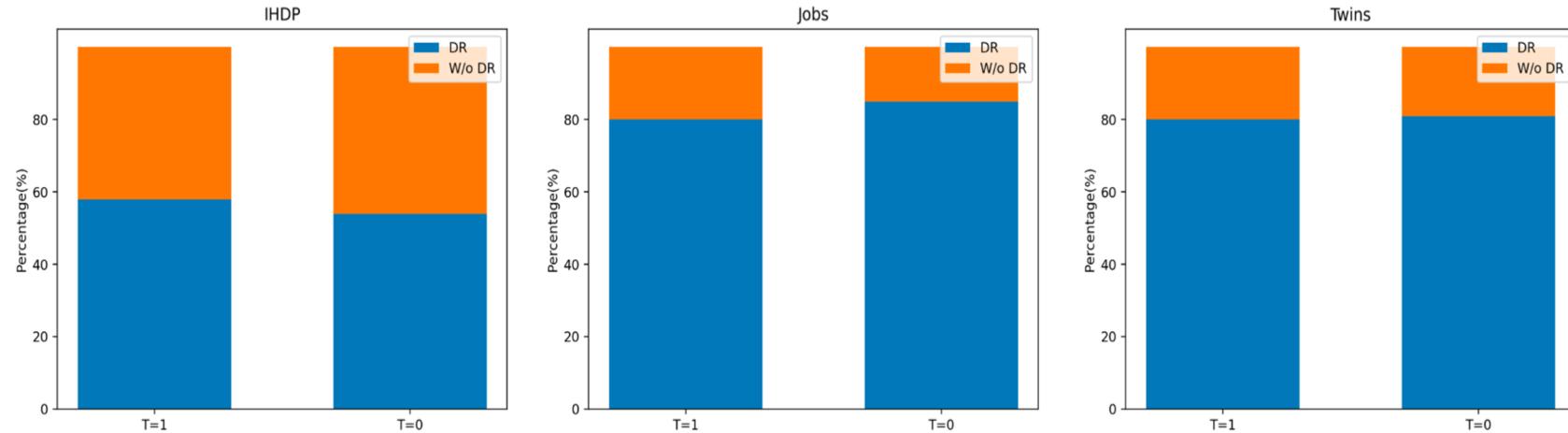


- IHDP
- Jobs
- Twins

Results – Ablation study

	IHDP $\sqrt{\epsilon_{PEHE}^{out-of-s}}$	Jobs $R_{Pol}^{out-of-s}$	Twins $\sqrt{\epsilon_{PEHE}^{out-of-s}}$
DR-VIDAL	0.62 ± 0.06	0.102 ± 0.01	0.318 ± 0.008
DR-VIDAL (w/o DR loss)	0.85 ± 0.06	0.110 ± 0.01	0.324 ± 0.007
DR-VIDAL (w/o Info loss)	0.67 ± 0.04	0.109 ± 0.01	0.318 ± 0.012
DR-VIDAL (w/o DR + Info loss)	0.81 ± 0.05	0.113 ± 0.01	0.326 ± 0.008

Results – correct classification of factual outcomes



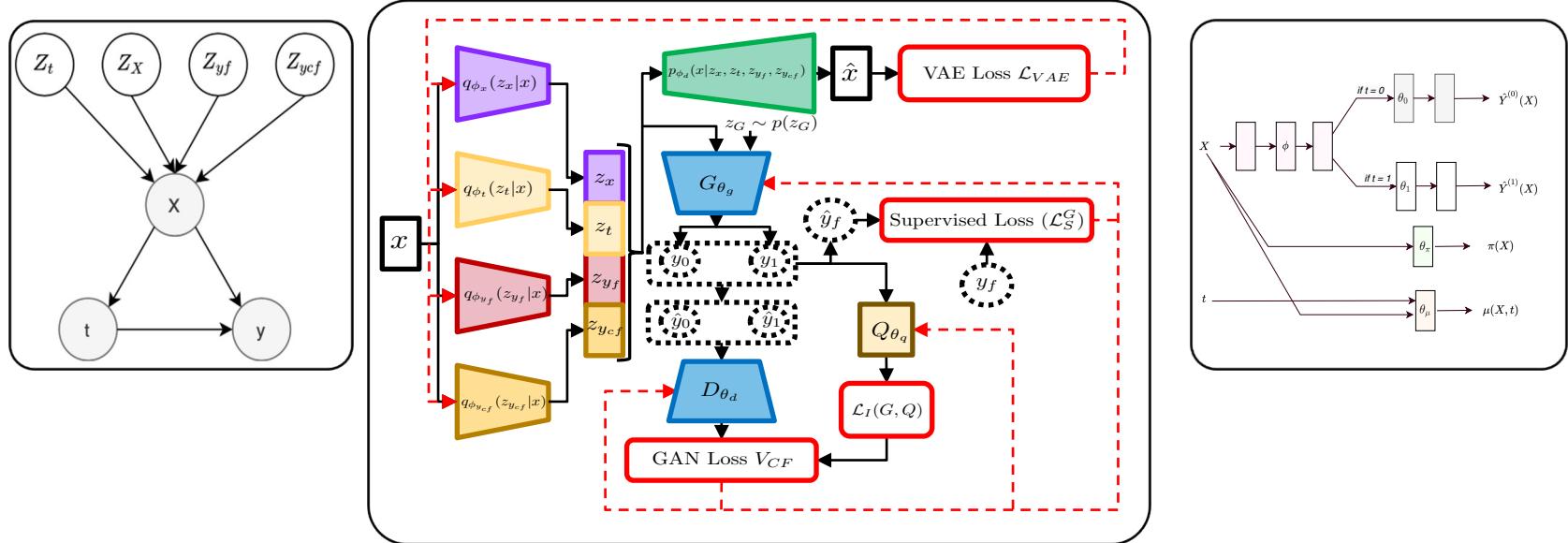
Results – PEHE and policy risk (mean \pm st.dev)

	IHDP($\sqrt{\epsilon_{PEHE}}$)		Twins($\sqrt{\epsilon_{PEHE}}$)		Jobs(R_{Pol})	
	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample
OLS/LR1	$5.8 \pm 0.3^*$	$5.8 \pm 0.3^*$	0.318 ± 0.007	$0.319 \pm 0.005^*$	$0.23 \pm 0.02^*$	$0.22 \pm 0.00^*$
OLS/LR2	$2.5 \pm 0.1^*$	$2.4 \pm 0.1^*$	$0.320 \pm 0.003^*$	$0.320 \pm 0.001^*$	$0.24 \pm 0.01^*$	$0.21 \pm 0.00^*$
BLR	$5.8 \pm 0.3^*$	$5.8 \pm 0.3^*$	$0.323 \pm 0.018^*$	$0.312 \pm 0.002^*$	$0.25 \pm 0.02^*$	$0.22 \pm 0.01^*$
k-NN	$4.1 \pm 0.2^*$	$2.1 \pm 0.1^*$	$0.345 \pm 0.007^*$	$0.333 \pm 0.003^*$	$0.26 \pm 0.02^*$	$0.02 \pm 0.00^*$
BART	$2.3 \pm 0.1^*$	$2.1 \pm 0.2^*$	$0.338 \pm 0.016^*$	$0.347 \pm 0.009^*$	$0.25 \pm 0.00^*$	$0.23 \pm 0.02^*$
R Forest	$6.6 \pm 0.3^*$	$4.2 \pm 0.2^*$	$0.321 \pm 0.005^*$	0.306 ± 0.002	$0.28 \pm 0.02^*$	$0.23 \pm 0.01^*$
C Forest	$3.8 \pm 0.2^*$	$3.8 \pm 0.2^*$	0.316 ± 0.011	$0.366 \pm 0.003^*$	$0.20 \pm 0.02^*$	$0.19 \pm 0.00^*$
BNN	$2.1 \pm 0.1^*$	$2.2 \pm 0.1^*$	$0.321 \pm 0.018^*$	$0.325 \pm 0.003^*$	$0.24 \pm 0.02^*$	$0.20 \pm 0.01^*$
TARNET (Tensor- Flow)	$0.95 \pm 0.02^*$	$0.88 \pm 0.02^*$	0.315 ± 0.003	0.317 ± 0.007	$0.21 \pm 0.01^*$	$0.17 \pm 0.01^*$
TARNeT (Pytorch)	$1.10 \pm 0.02^*$	-	-	-	$0.29 \pm 0.06^*$	-
CFRW _{ASS}	$0.76 + 0.0^*$	$0.71 + 0.0^*$	$0.313 + 0.008$	$0.315 + 0.007$	$0.21 + 0.01^*$	$0.17 + 0.01^*$
GANITE	$2.4 \pm 0.4^*$	$1.9 \pm 0.4^*$	0.297 ± 0.05	0.289 ± 0.005	$0.14 \pm 0.01^*$	$0.13 \pm 0.01^*$
CEVAE	$2.6 \pm 0.1^*$	$2.7 \pm 0.1^*$	n.r	n.r	$0.26 \pm 0.0^*$	$0.15 \pm 0.0^*$
DR- VIDAL	0.69 ± 0.06	0.69 ± 0.05	0.318 ± 0.008	0.317 ± 0.002	0.10 ± 0.01	0.09 ± 0.005

Results – ATE and ATT (mean \pm st.dev)

	IHDP(ϵ_{ATE})		Twins(ϵ_{ATE})		Jobs(ϵ_{ATT})	
	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample
OLS/LR1	0.94 \pm 0.06	0.73 \pm 0.04	0.0069 \pm 0.0056	0.0038 \pm 0.0025	0.08 \pm 0.04	0.01 \pm 0.00
OLS/LR2	0.31 \pm 0.02	0.14 \pm 0.01	0.0070 \pm 0.0025	0.0039 \pm 0.0025	0.08 \pm 0.03	0.01 \pm 0.01
BLR	0.93 \pm 0.05	0.72 \pm 0.04	0.0334 \pm 0.0092	0.0057 \pm 0.0036	0.08 \pm 0.03	0.01 \pm 0.011
k-NN	0.90 \pm 0.05	0.14 \pm 0.01	0.0051 \pm 0.0039	0.0028 \pm 0.0021	0.13 \pm 0.05	0.21 \pm 0.01
BART	0.34 \pm 0.02	0.23 \pm 0.01	0.1265 \pm 0.0234	0.1206 \pm 0.0236	0.08 \pm 0.03	0.02 \pm 0.00
R Forest	0.96 \pm 0.06	0.73 \pm 0.05	0.0080 \pm 0.0051	0.0049 \pm 0.0034	0.09 \pm 0.04	0.03 \pm 0.01
C Forest	0.40 \pm 0.03	0.18 \pm 0.01	0.0335 \pm 0.0083	0.0286 \pm 0.0035	0.07 \pm 0.03	0.03 \pm 0.01
BNN	0.42 \pm 0.03	0.37 \pm 0.03	0.0203 \pm 0.0071	0.0056 \pm 0.0032	0.09 \pm 0.04	0.03 \pm 0.01
TARNET	0.28 \pm 0.01	0.26 \pm 0.01	0.0151 \pm 0.0018	0.0108 \pm 0.0017	0.09 \pm 0.04	0.03 \pm 0.01
CEP	0.27 \pm 0.01	0.25 \pm 0.01	0.0024 \pm 0.0020	0.0119 \pm 0.0016	0.09 \pm 0.03	0.04 \pm 0.01
GANITE	0.49 \pm 0.05	0.43 \pm 0.05	0.0089 \pm 0.0075	0.0058 \pm 0.0017	0.06 \pm 0.03	0.01 \pm 0.01
CEVAE	0.46 \pm 0.02	0.34 \pm 0.01	n.r	n.r	0.03 \pm 0.01	0.02 \pm 0.01
DR-VIDAL	0.49 \pm 0.06	0.49 \pm 0.07	0.0111 \pm 0.0137	0.0102 \pm 0.0128	0.05 \pm 0.02	0.04 \pm 0.03

Conclusion



1. Beats the performance of previous generative models on synthetic datasets
2. Comparable performance on real-world datasets.

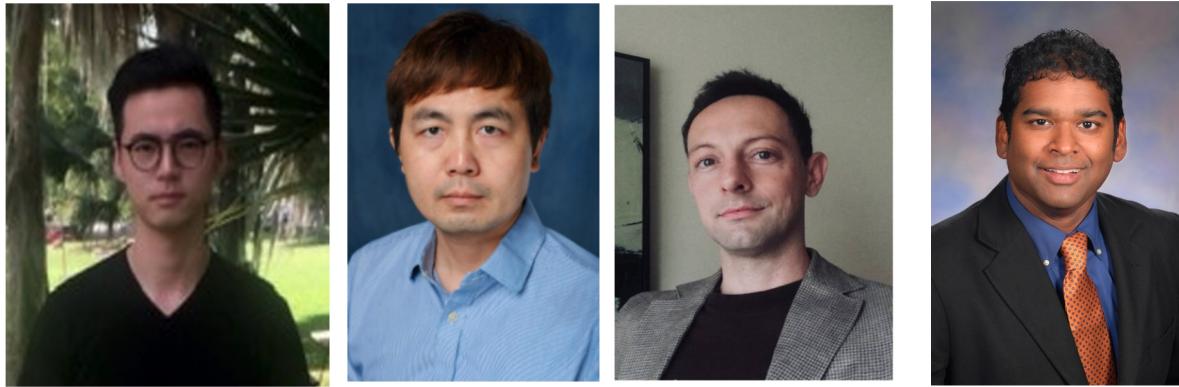
Code: <https://github.com/Shantanu48114860/DR-VIDAL-AMIA-22/>

Future directions



1. Causal Graph is too simple.
2. How to enforce strict disentanglement?
3. Instead of Variational Information Maximization, what about Variational Information Bottleneck Layer?
4. TARNET, DRAGONNET, DCN-PD, ITE block as a downstream model instead of the Doubly robust treatment estimator.
6. More realistic data from News-8 or MIMIC-II(EBB).

Acknowledgement



UF | Data Intelligence Systems Lab
Department of Epidemiology
UNIVERSITY of FLORIDA



Thank you!

Email me at: shg121@pitt.edu