When and How Should One Use Deep Learning for Causal Effect Inference

Uri Shalit – IEM, Technion

Israeli Statistical Association
May 2018

Causal effect inference questions

- Which medication will make patients better?
- Which economic policy will lower unemployment?
- Which school-curriculum will teach students best?
- 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
- Which school-curriculum will teach students best?
  - Infer from school records
- The effects of **actions** on **outcomes**

Causal inference from observational data - confounding

- Which medication will make patients better?
  - Infer from medical records
  - Maybe younger/richer/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?
- Which school-curriculum will teach students best?
  - Infer from school records
  - Maybe school in higher socio-economic status tend to have curriculum A over B?
- Model the interaction of confounders, actions, and outcomes

Causal inference problems are increasingly

Observational (non-experimental)

Many noisy confounders

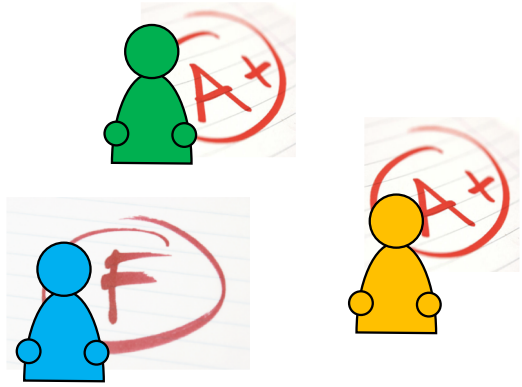
Causal inference problems are becoming

High-dimensional

What are the effects of genetics?

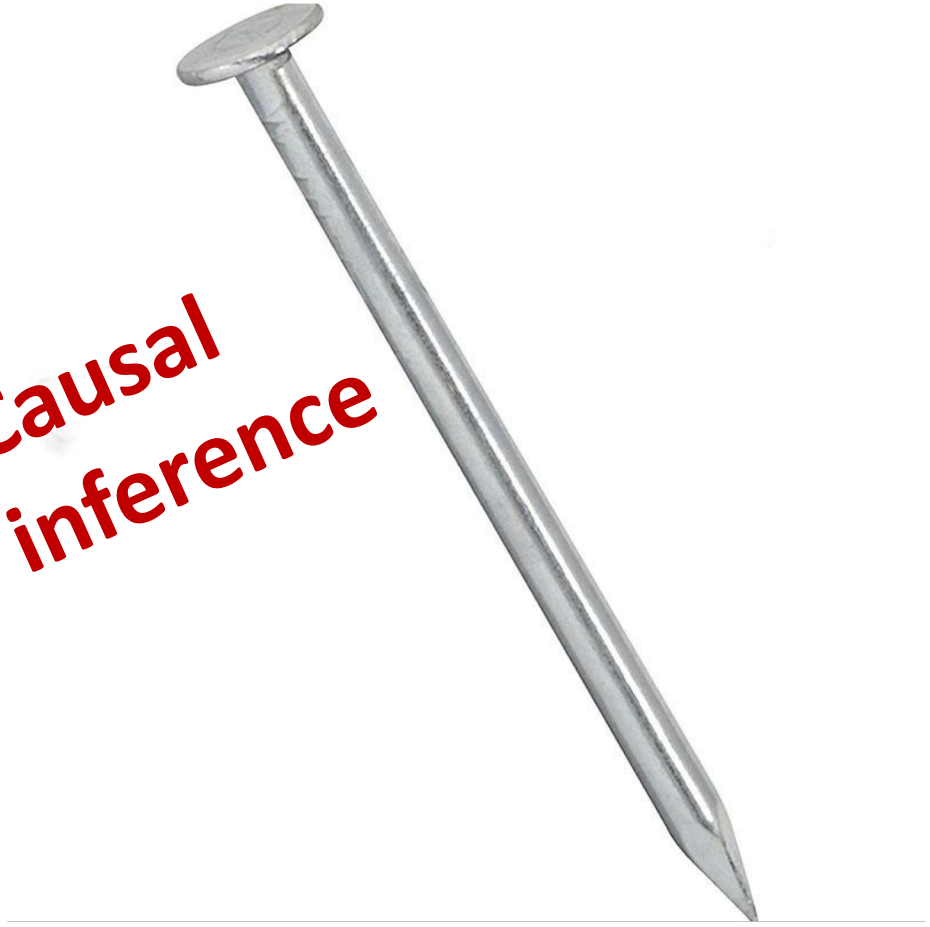
Causal inference problems are becoming

Personalized



How should we teach our students?

Causal effect inference from observational data is a tough nail



Causal effect inference from observational data is a tough nail

- No ground truth, no test set
- Must extrapolate to unseen populations
- Unmeasured confounding

Deep learning is a heavy hammer

Deep learning is a heavy hammer

- Object recognition in images and videos
- Machine translation
- Generative models

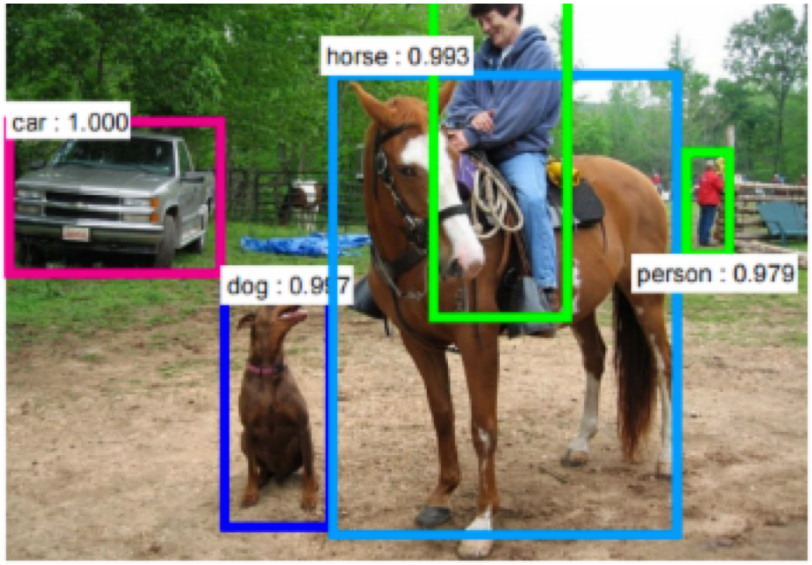
What causal effect inference problems might benefit from Deep Learning?

What is Deep Learning good for?

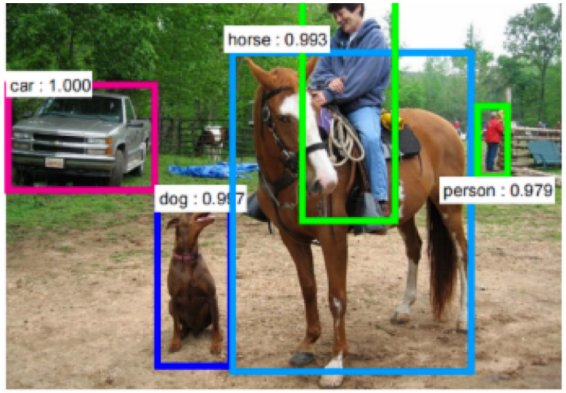
- Modeling high-dimensional relationships

What is Deep Learning good for?

- Modeling high-dimensional relationships
  - DL success: mapping pixels to objects



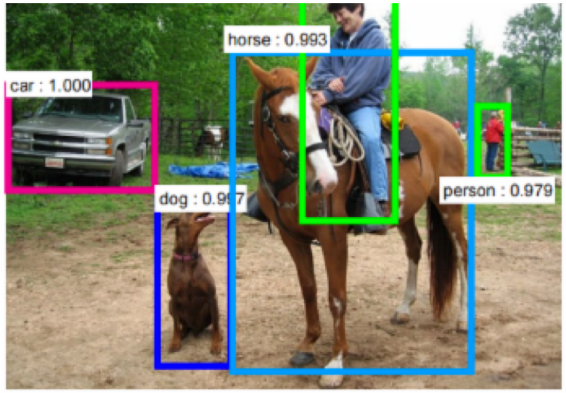
What is Deep Learning good for?



- Modeling high-dimensional relationships
  - DL success: mapping pixels to objects

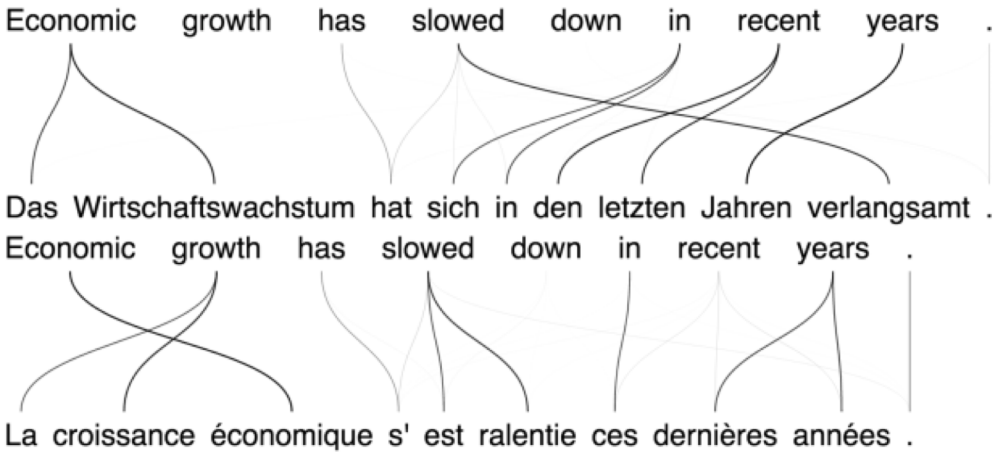
- Optimizing over rich function classes

What is Deep Learning good for?

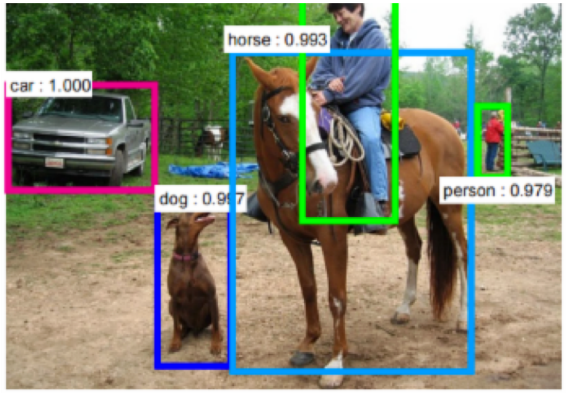


- Modeling high-dimensional relationships
  - DL success: mapping pixels to objects

- Optimizing over rich function classes
  - DL success: mapping English text to French text



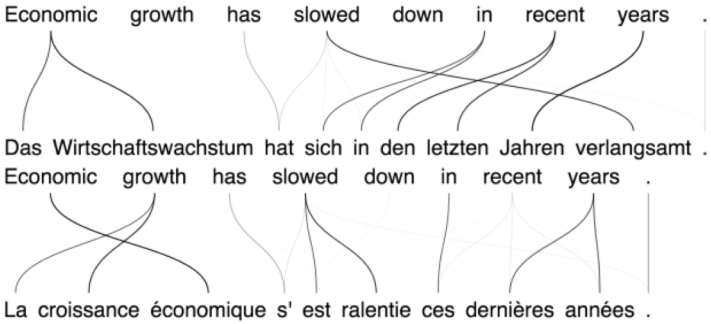
What is Deep Learning good for?



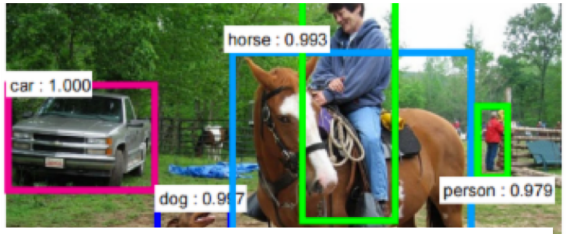
- Modeling high-dimensional relationships
  - DL success: mapping pixels to objects

- Optimizing over rich function classes
  - DL success: mapping English text to French text

- Finding latent structure



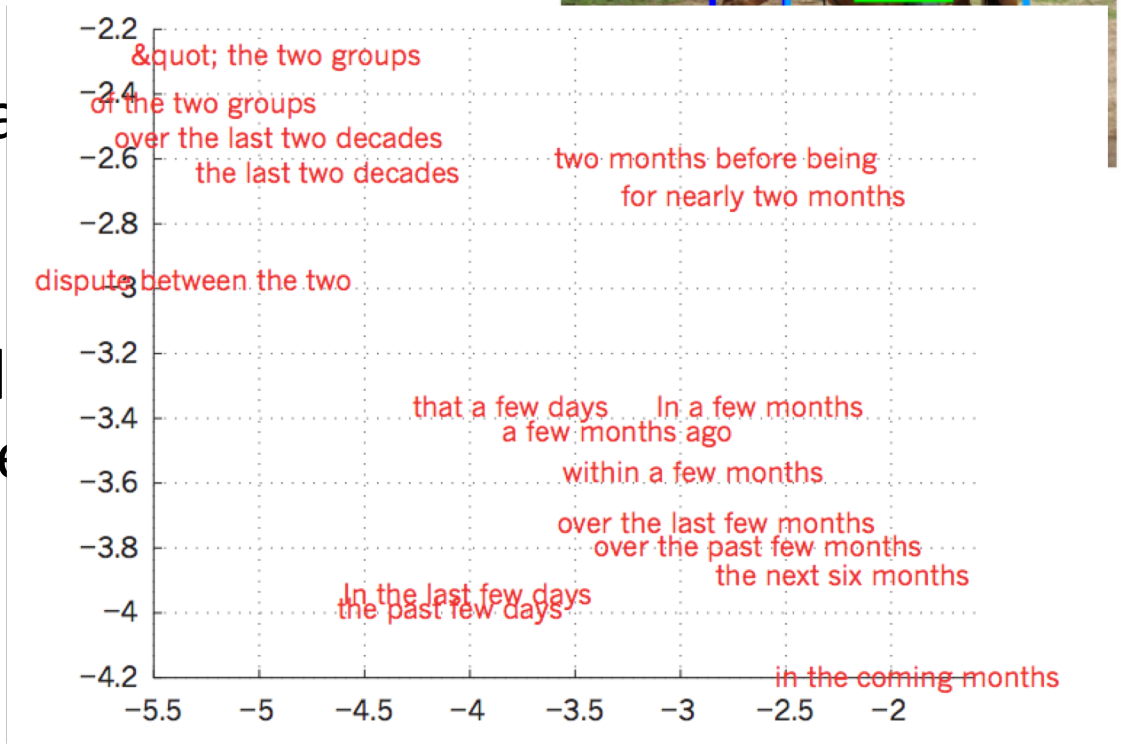
What is Deep Learning good for?



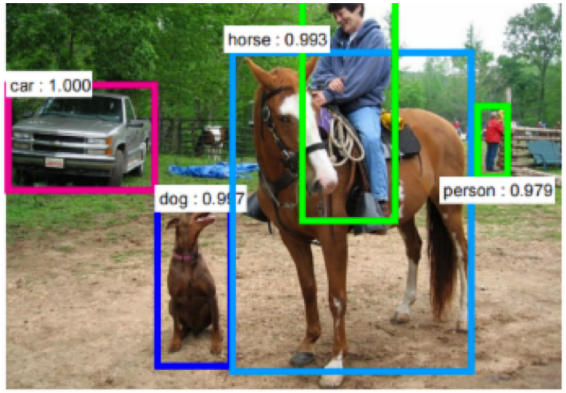
- Modeling high-dimensional rela
  - DL success: mapping pixels to

- Optimizing over rich function cl
  - DL success: mapping English te
  French text

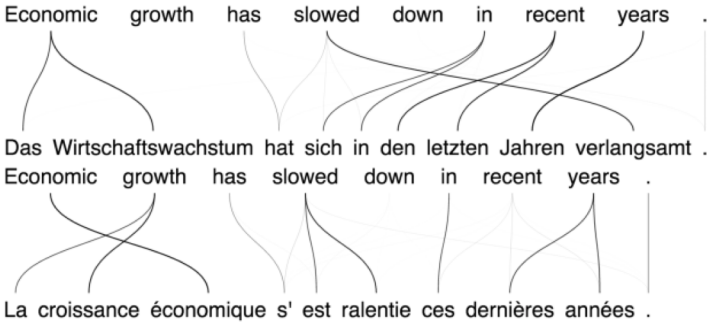
- Finding latent structure
  - DL success: semantic embedding of words



What is Deep Learning good for?

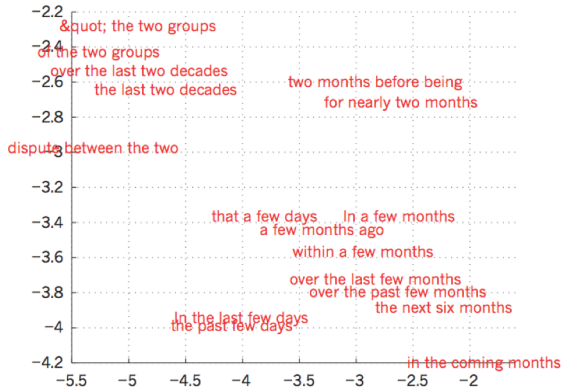


- Modeling high-dimensional relationships
  - DL success: mapping pixels to objects

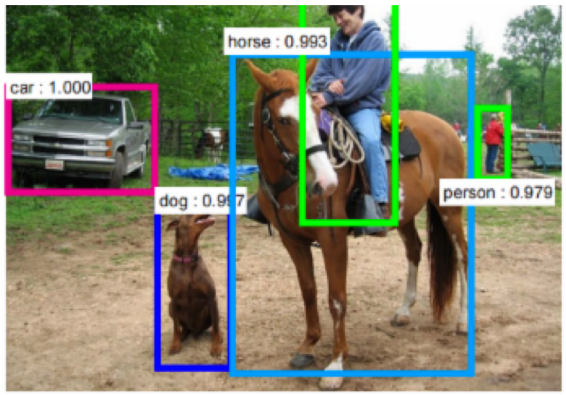


- Optimizing over rich function classes
  - DL success: mapping English text to French text

- Finding latent structure
  - DL success: semantic embedding of words



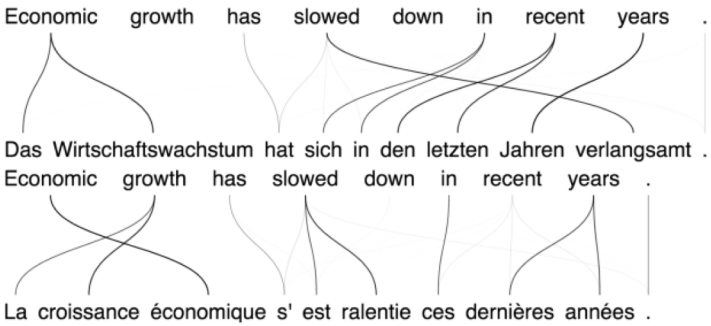
What causal effect inference problems might benefit from Deep Learning?

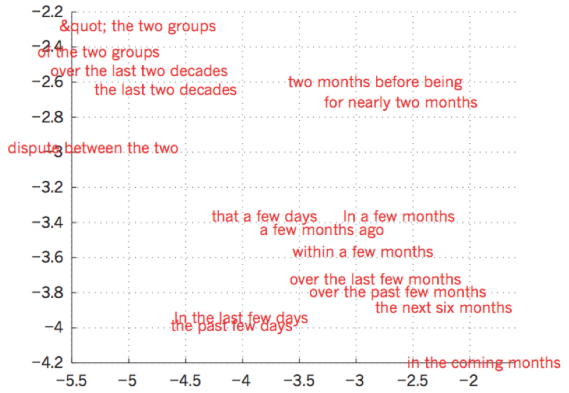


- Modeling high-dimensional relationships
  - Observational dataset are often high-dim

- Optimizing over rich function classes
  - Learning individual-level treatment effects

- Finding latent structure
  - Unmeasured confounders with proxies





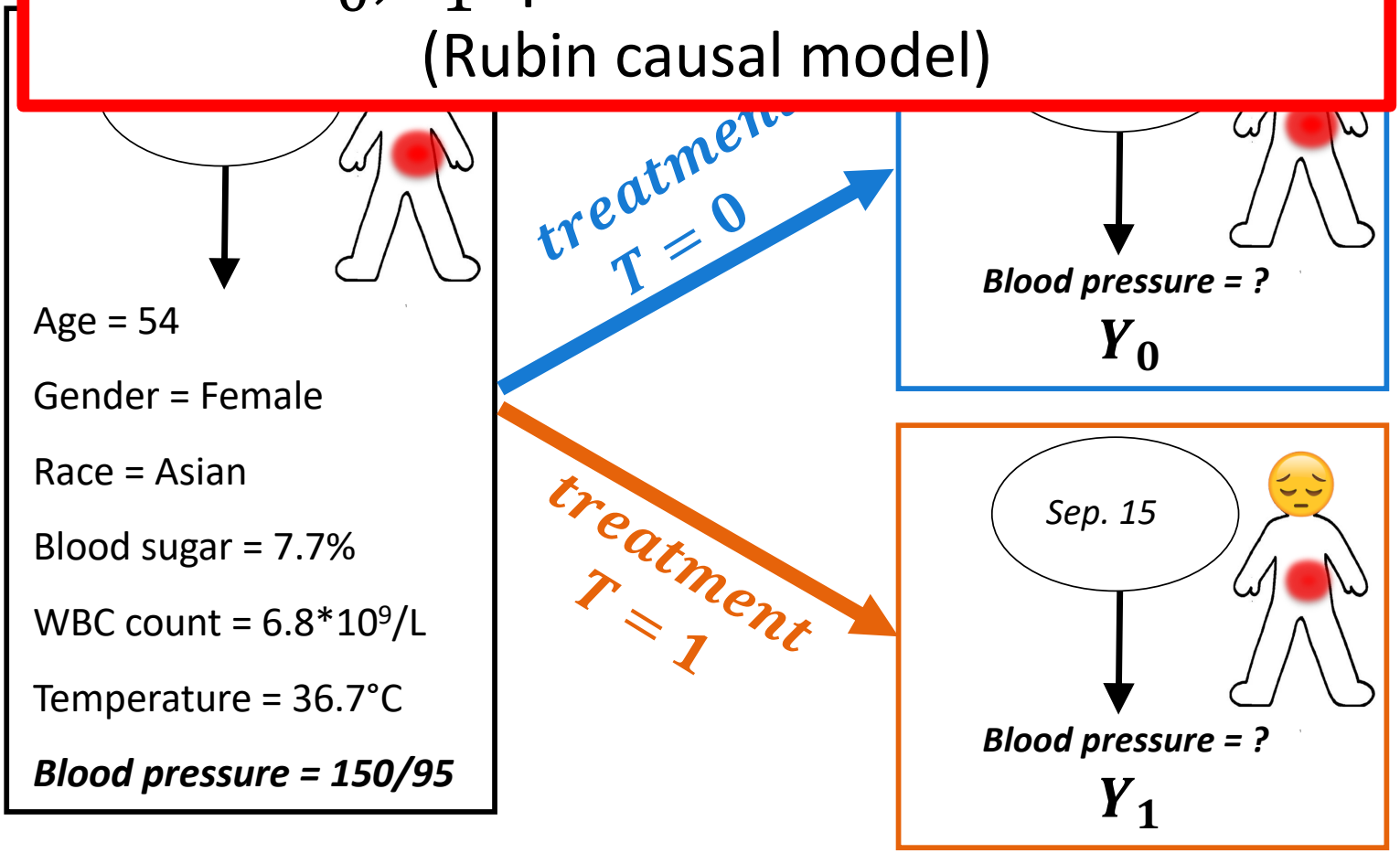
Outline

- **Background**
- **Individual-level treatment effects**
- **Latent, not hidden, confounders**

Outline

- **Background**
- Individual-level treatment effects
- Latent, not hidden, confounders

Y_0, Y_1: potential outcomes
(Rubin causal model)

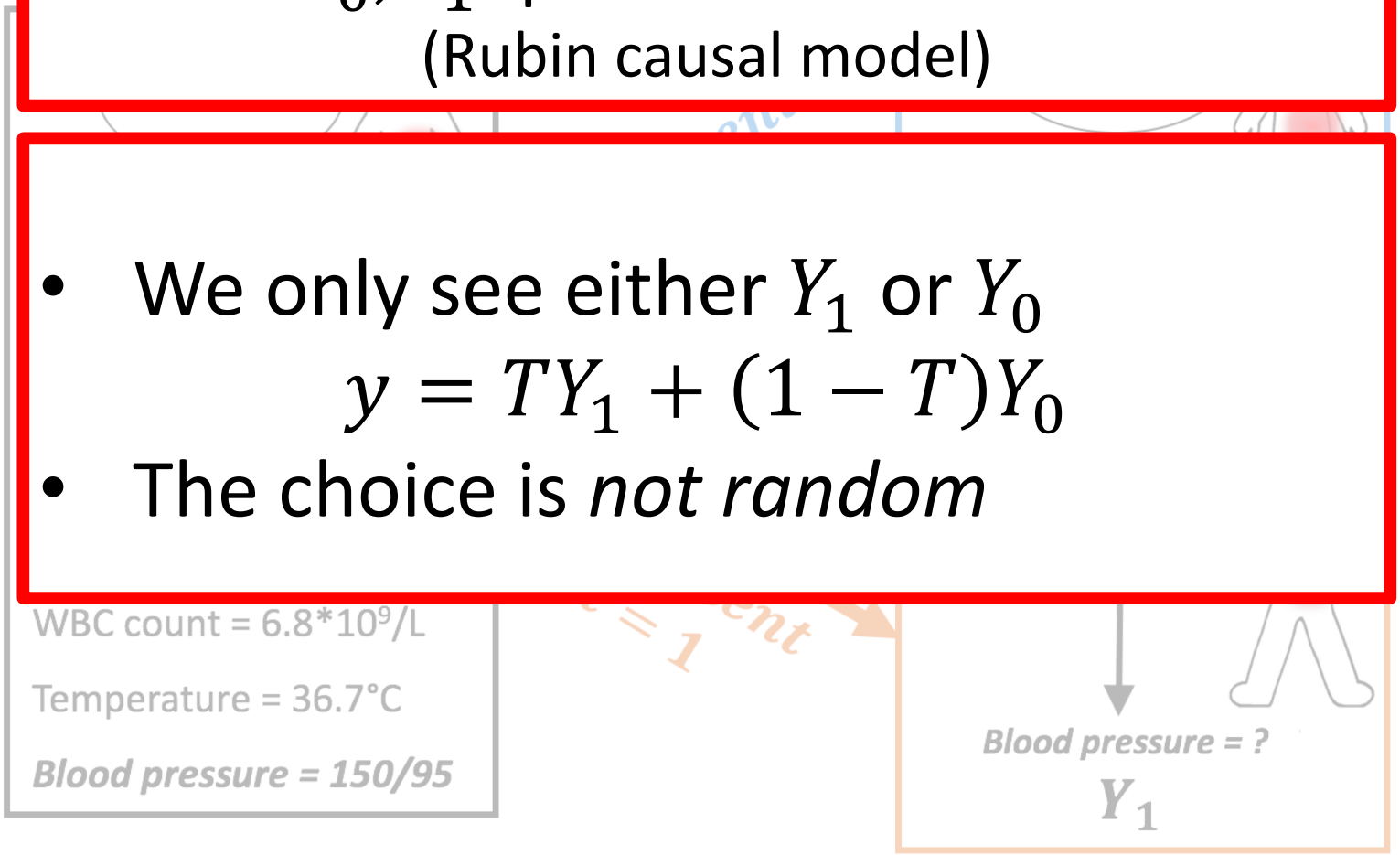


Y_0, Y_1: potential outcomes
(Rubin causal model)

- We only see either Y_1 or Y_0

$$y = TY_1 + (1 - T)Y_0$$

- The choice is *not random*



Y_0, Y_1: potential outcomes
(Rubin causal model)

- We only see either Y_1 or Y_0

$$y = TY_1 + (1 - T)Y_0$$

- The choice is *not random*

Average Treatment Effect

$$ATE \equiv \mathbb{E}[Y_1 - Y_0]$$

When can we hope to obtain causal effects from observational data?

- **Strong ignorability**: "No hidden confounders"

$$Y_1, Y_0 \perp\!\!\!\perp T | X$$

All variables that affect both potential outcomes and *treatment are measured*

- **Overlap**:

$$1 > p(T = 1|X) > 0$$

All treatments have non-zero probability of being observed for every unit

- **Stable Unit Treatment Value Assumption** (SUTVA):

Treatments and outcomes of different subjects are independent

When can we hope to obtain causal effects from observational data?

- **Strong ignorability**: "No hidden confounders"

  All variables that affect both potential outcomes and *treatment are measured*

- **Overlap**:

  All treatments have non-zero probability of being observed for every unit

- **Stable Unit Treatment Value Assumption** (SUTVA):

  Treatments and outcomes of different subjects are independent

- These are all non-trivial!

When can we hope to obtain causal effects from observational data?

- **Strong ignorability**: "No hidden confounders"

  All variables that affect both potential outcomes and *treatm* *measured*

Will be modified later

Outline

- **Background**
- Individual-level treatment effects
- Latent, not hidden, confounders

Outline

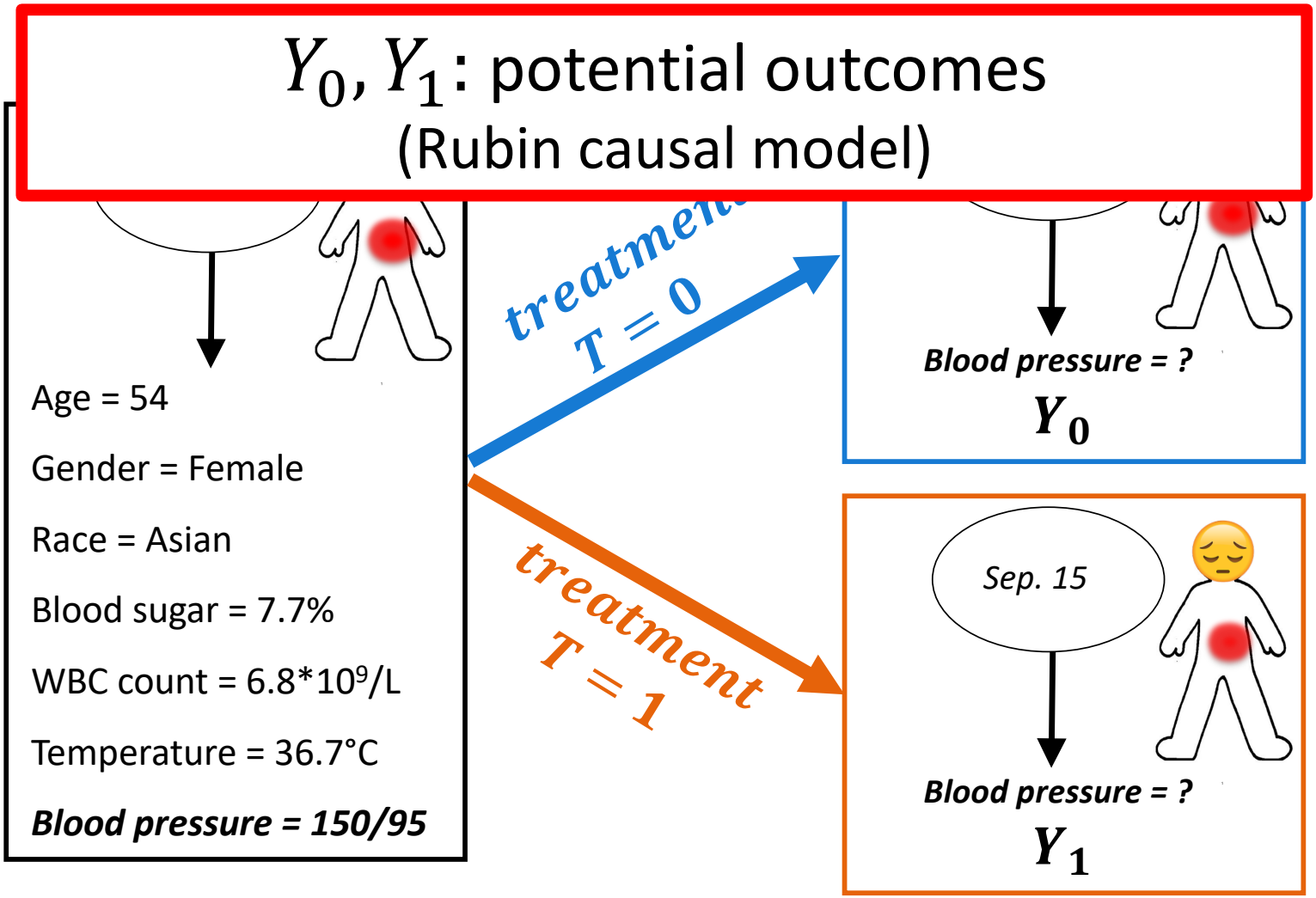
- Background
- **Individual-level treatment effects**
- Latent, not hidden, confounders

Outline

- Background

- **Individual-level treatment effects**
  (i) Johansson, F., Shalit, U., & Sontag, D. (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, F. D., Kallus, N., Shalit, U., & Sontag, D. (2018). Learning Weighted Representations for Generalization Across Designs. *arXiv preprint arXiv:1802.08598*.
- Latent, not hidden, confounders

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

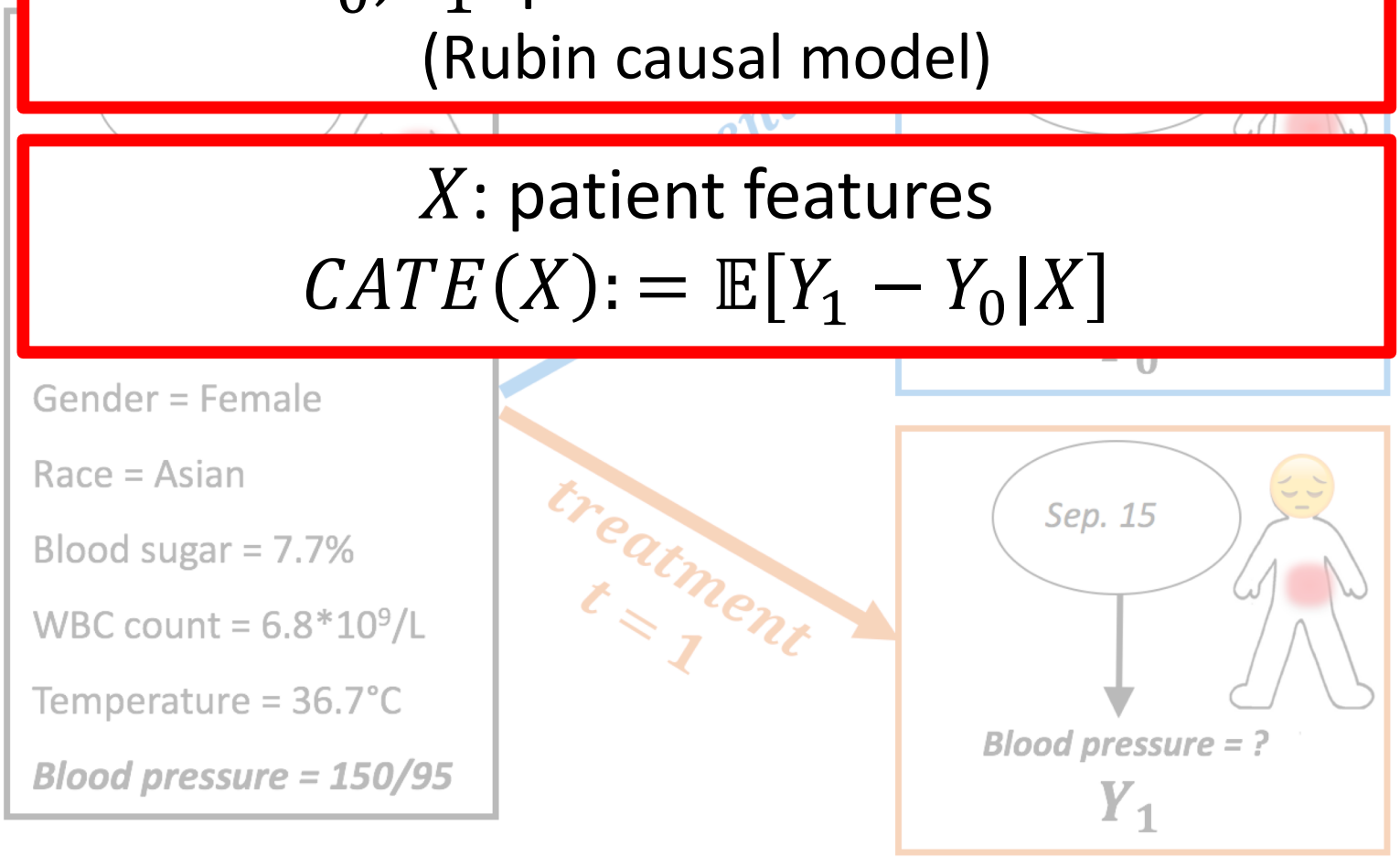


Our goal: Conditional Average Treatment Effect (*CATE*)

Y_0, Y_1: potential outcomes
(Rubin causal model)

X: patient features

$$CATE(X) := \mathbb{E}[Y_1 - Y_0 | X]$$

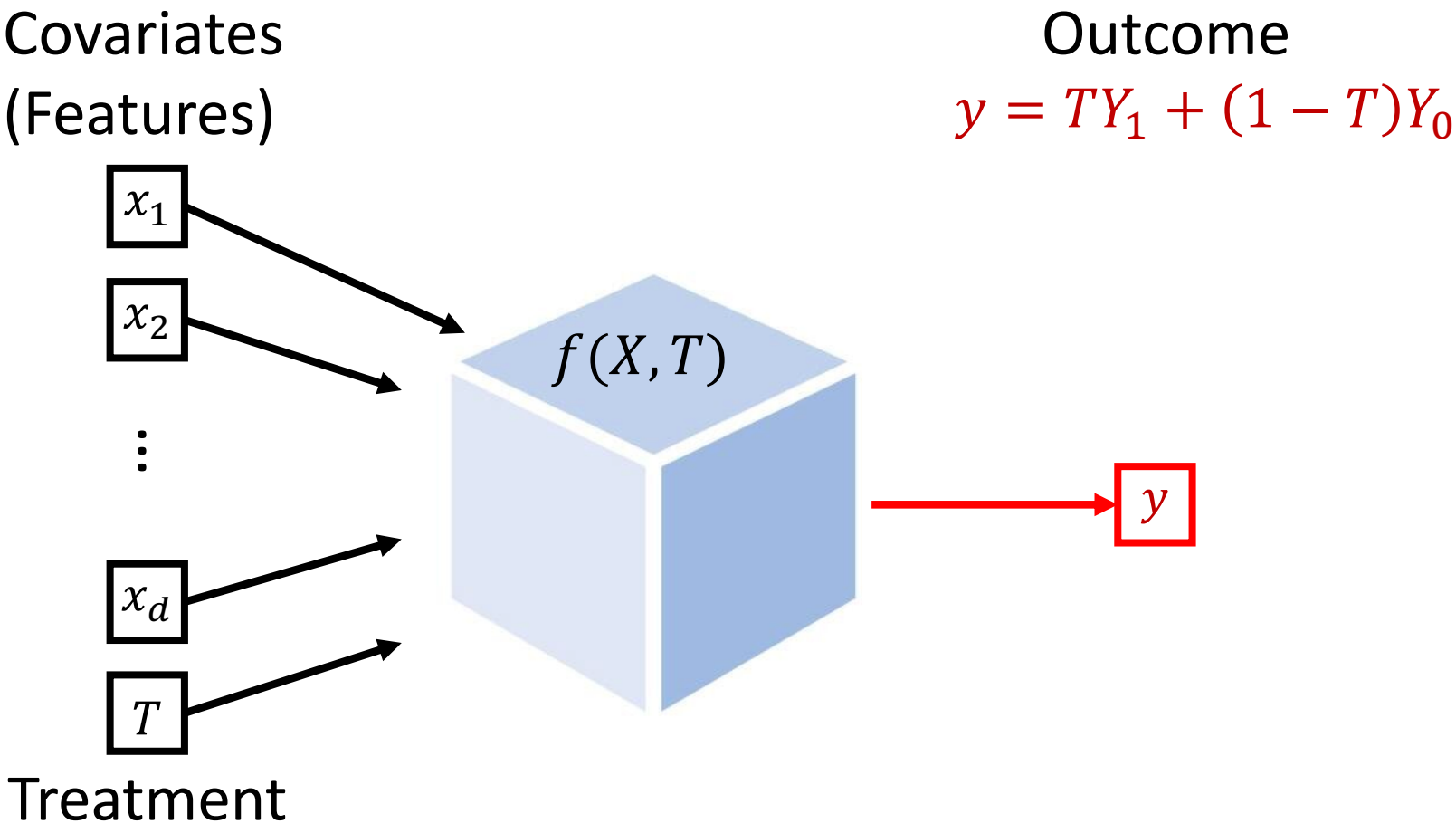


Y_0, Y_1: potential outcomes
(Rubin causal model)

X: patient features

$$CATE(X) := \mathbb{E}[Y_1 - Y_0 | X]$$

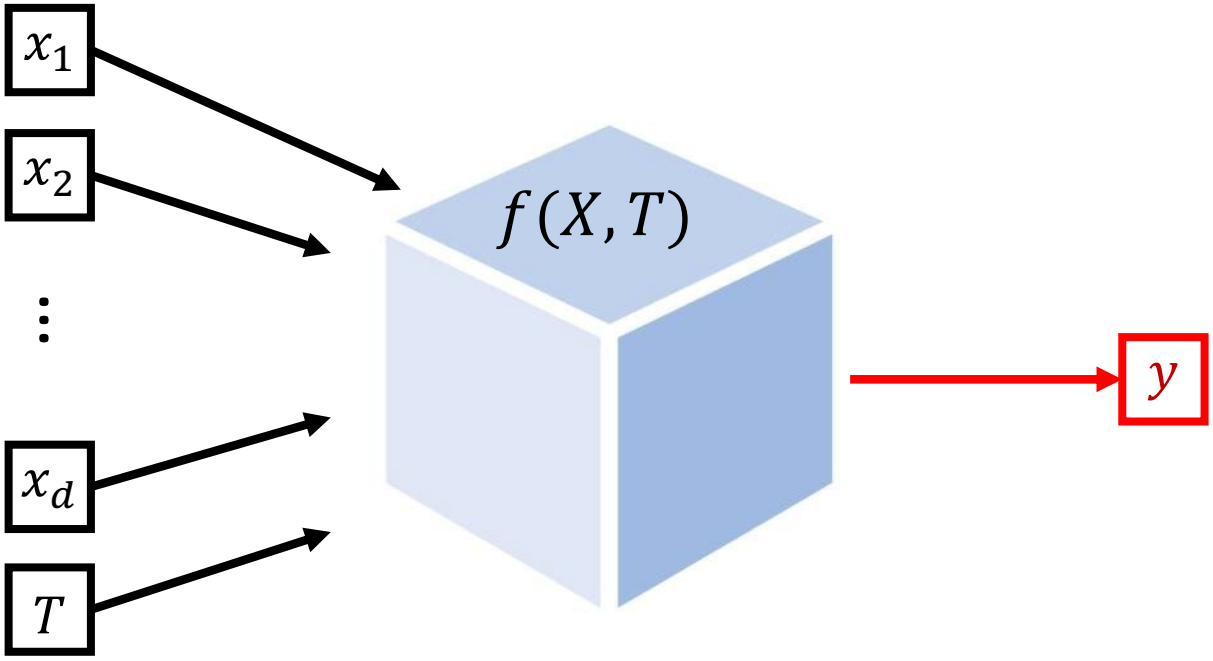
- We never directly observe $CATE$
- We only see either Y_1 or Y_0
- The choice is *not random*
- How to estimate the CATE function?


Covariates
(Features)
x_1
x_2
x_d
T
Treatment
$f(X,T)$
Outcome
$y = TY_1 + (1-T)Y_0$
y

Covariates
(Features)

Outcome

$y = TY_1 + (1 - T)Y_0$



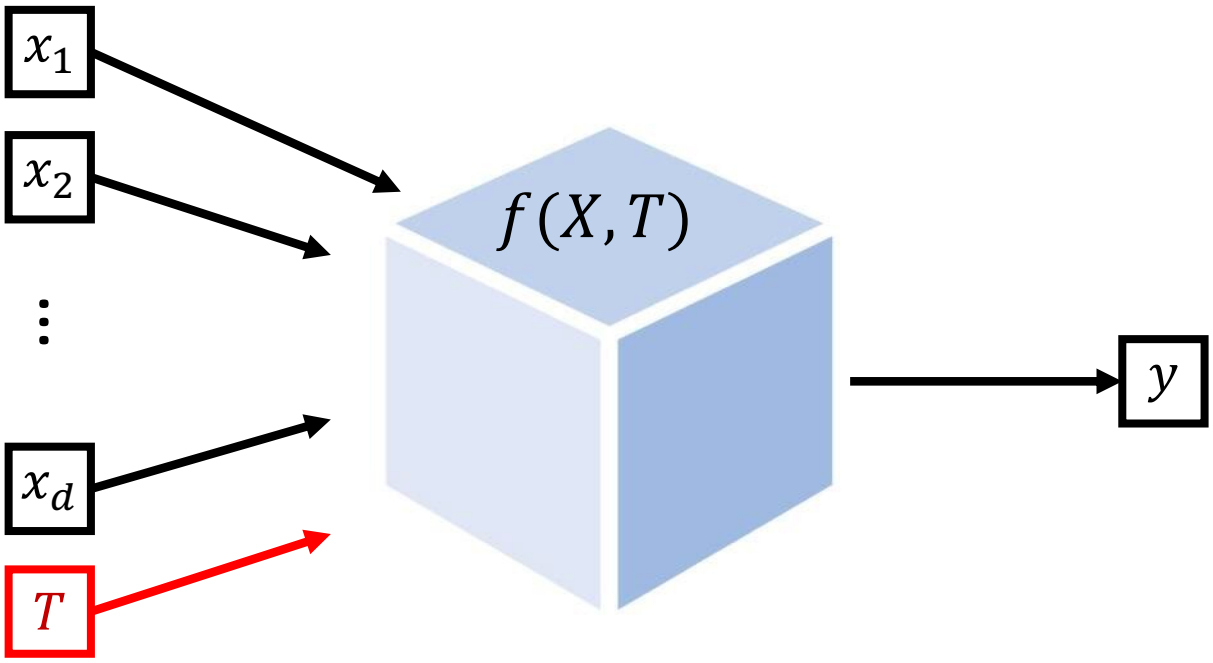
Treatment

Supervised deep learning:
focus on observed outcome y

Covariates
(Features)

Outcome

$y = TY_1 + (1 - T)Y_0$

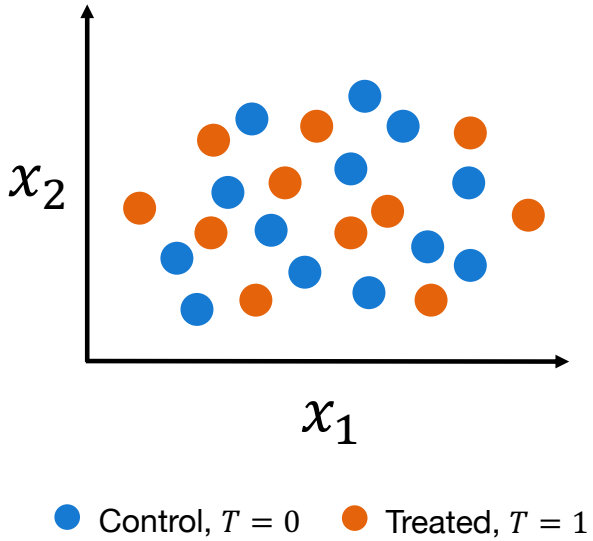


Treatment

Causal inference:
focus on treatment T

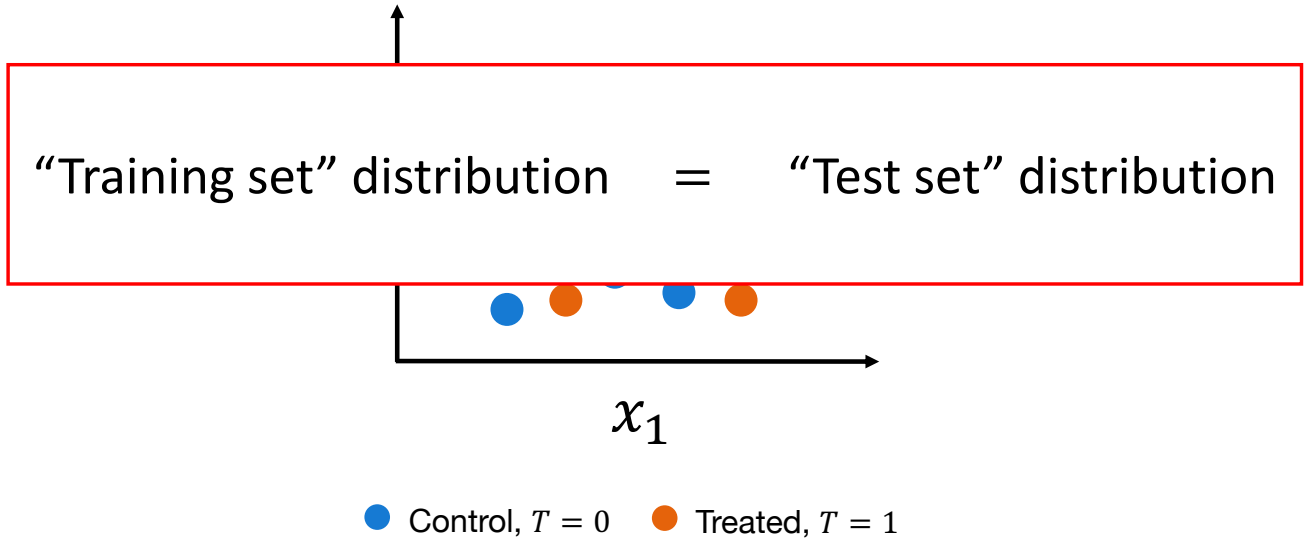
Easier: Randomized Controlled Trials (RCT)

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



Easier: Randomized Controlled Trials (RCT)

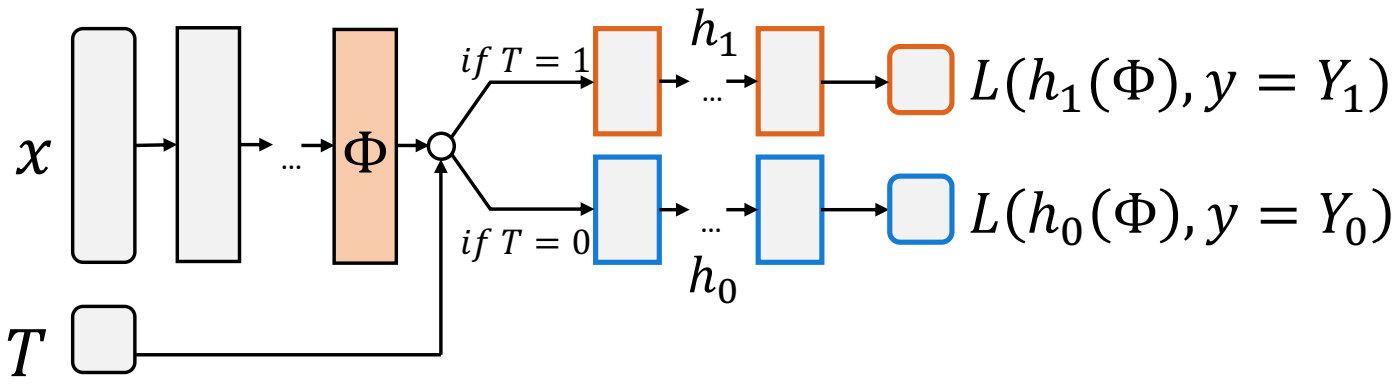
- Treatment is assigned unfirmly at **random:** $p(T = 1 \mid X) = P(T = 1)$
- Here: every dot is a unit, color indicates **observed** treatment
- Predict outcome under **unobserved** treatment



Neural network architecture: TARNet

(Treatment-Agnostic Representation Network)

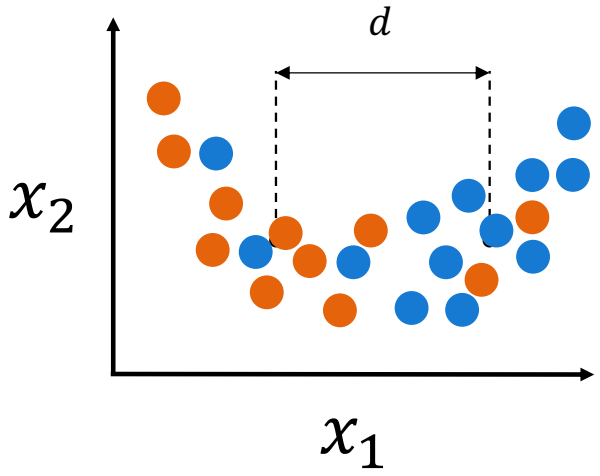
- In randomized control trials, there is no confounding – just do regression!
- New architecture for estimating counterfactuals and CATE



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

Observational studies: test ≠ train

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



Example:
A difference in means

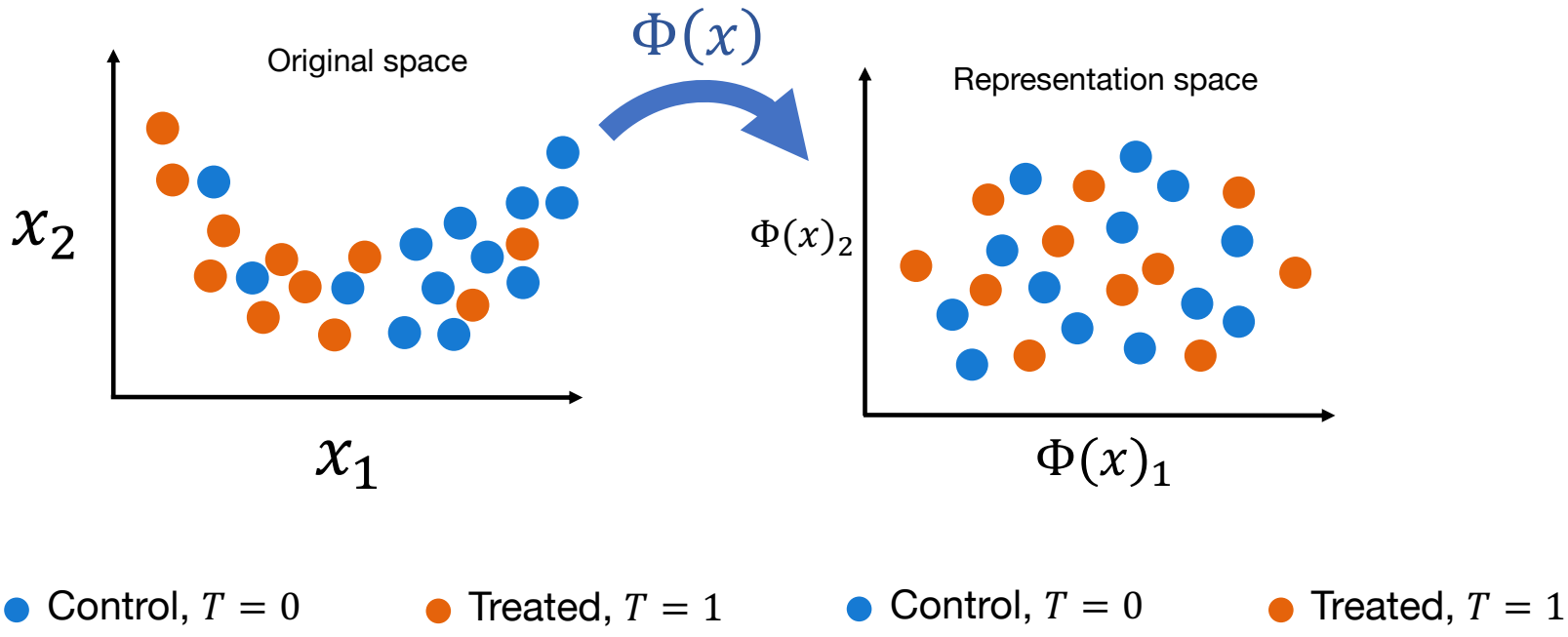
"Treated tend to be younger"

Control, $T = 0$ Treated, $T = 1$

Representation learning

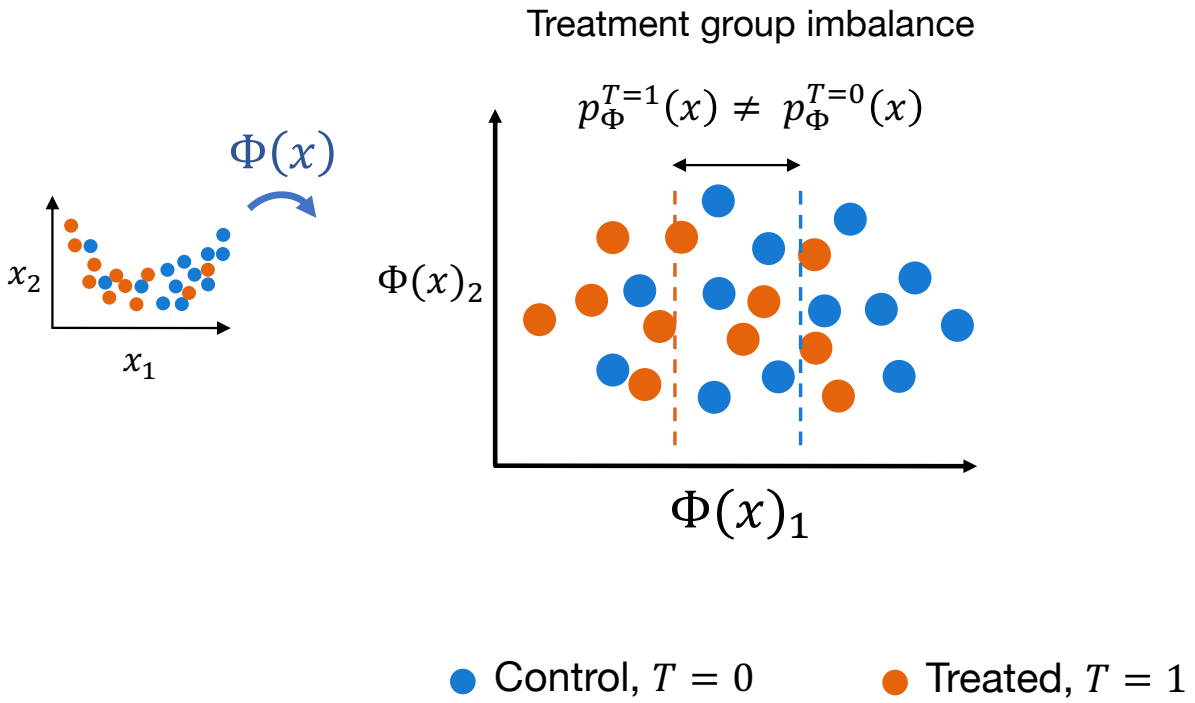
- Learn a representation Φ of the data that makes it more like an RCT
- A shared representation helps identify meaningful interactions
- **Penalize the distributional distance between treatment groups**

  New type of bias-variance tradeoff



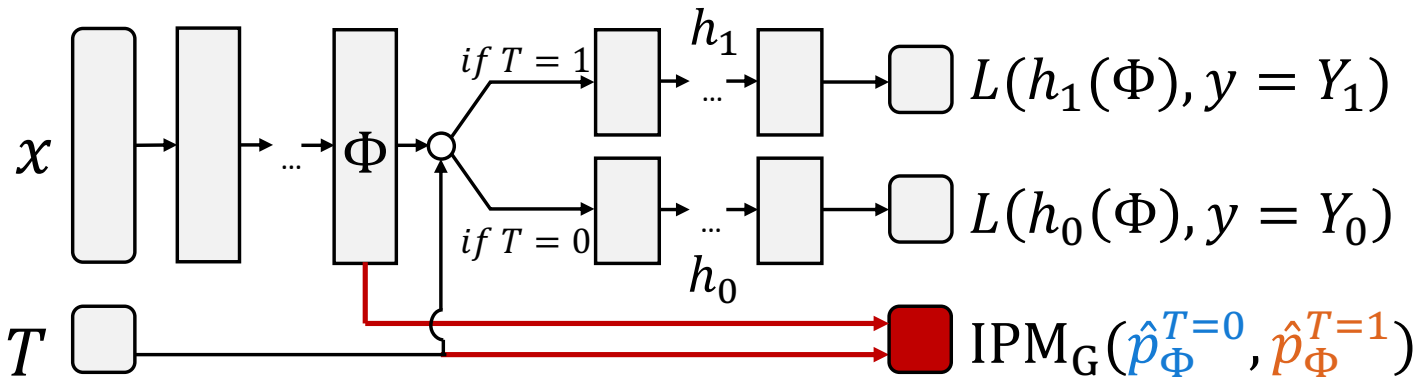
Imbalance in representation space

- 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: $\mathrm{IPM}_G(p_1, p_2) = \sup_{g \in G} \left| \int_S g(s)\big(p_1(s) - p_2(s)\big) ds \right|$

Individual-level treatment effect generalization bound

- Precision in Estimation of Heterogeneous Effects[1]:

$$\epsilon_{\text{CATE}} = \int_x \left(\widehat{\text{CATE}}(x) - \text{CATE}(x)\right)^2 p(x)dx$$

- Factual per-treatment group prediction error

$$\epsilon_F^{T=0} = \int_x \ell_{h,\Phi}(x,0)p^{t=0}(x)dx$$

$$\epsilon_F^{T=1} = \int_x \ell_{h,\Phi}(x,1)p^{t=1}(x)dx$$

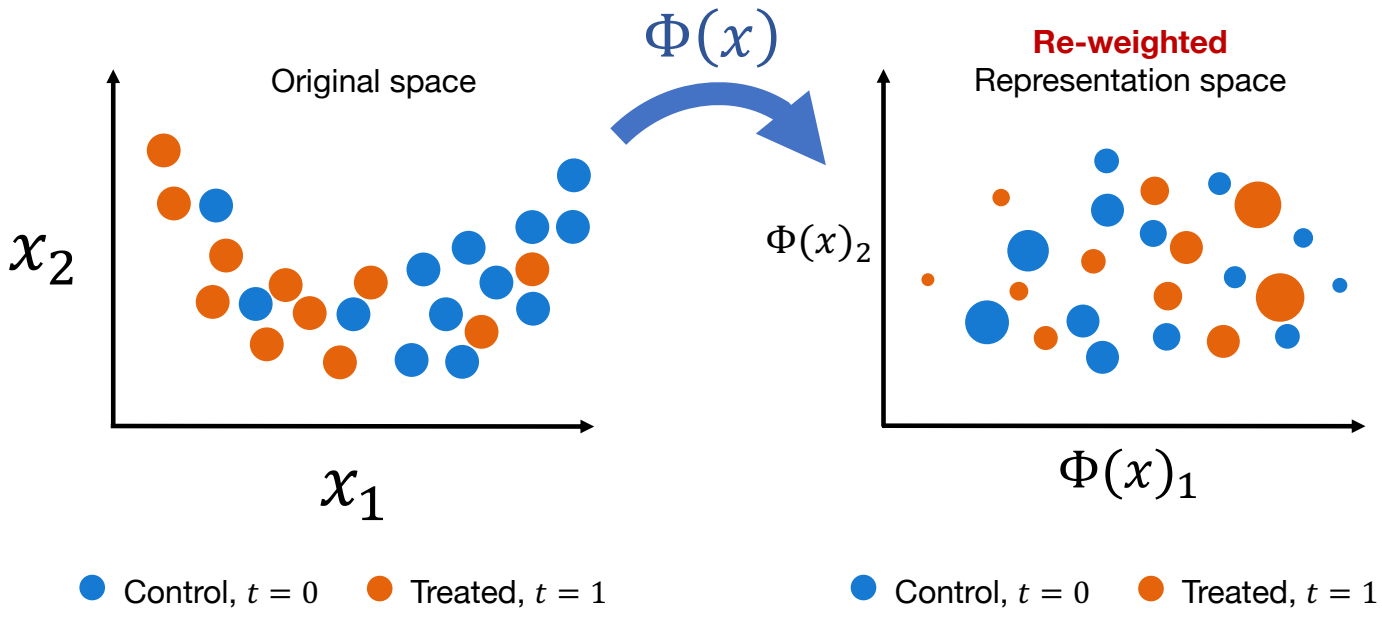
- **Theorem:**

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

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

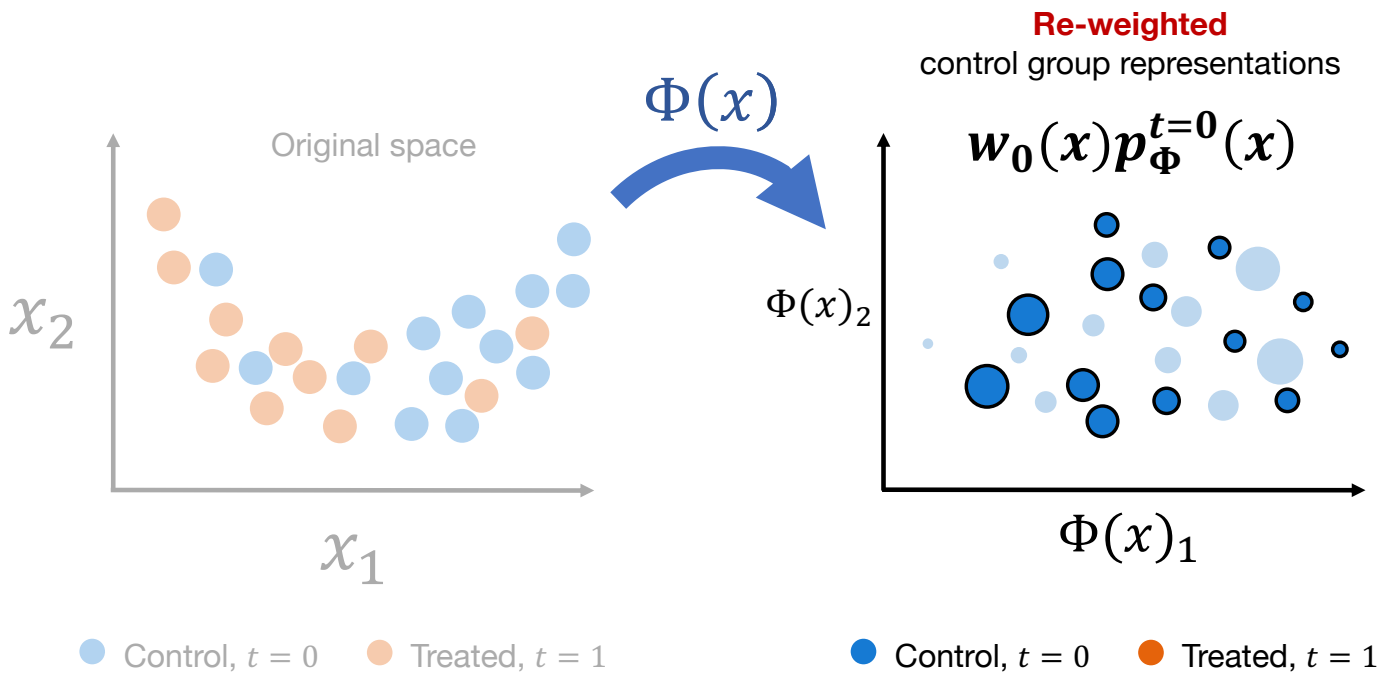
Re-weighted representation learning

- Re-weighted representations for improved treatment group balance



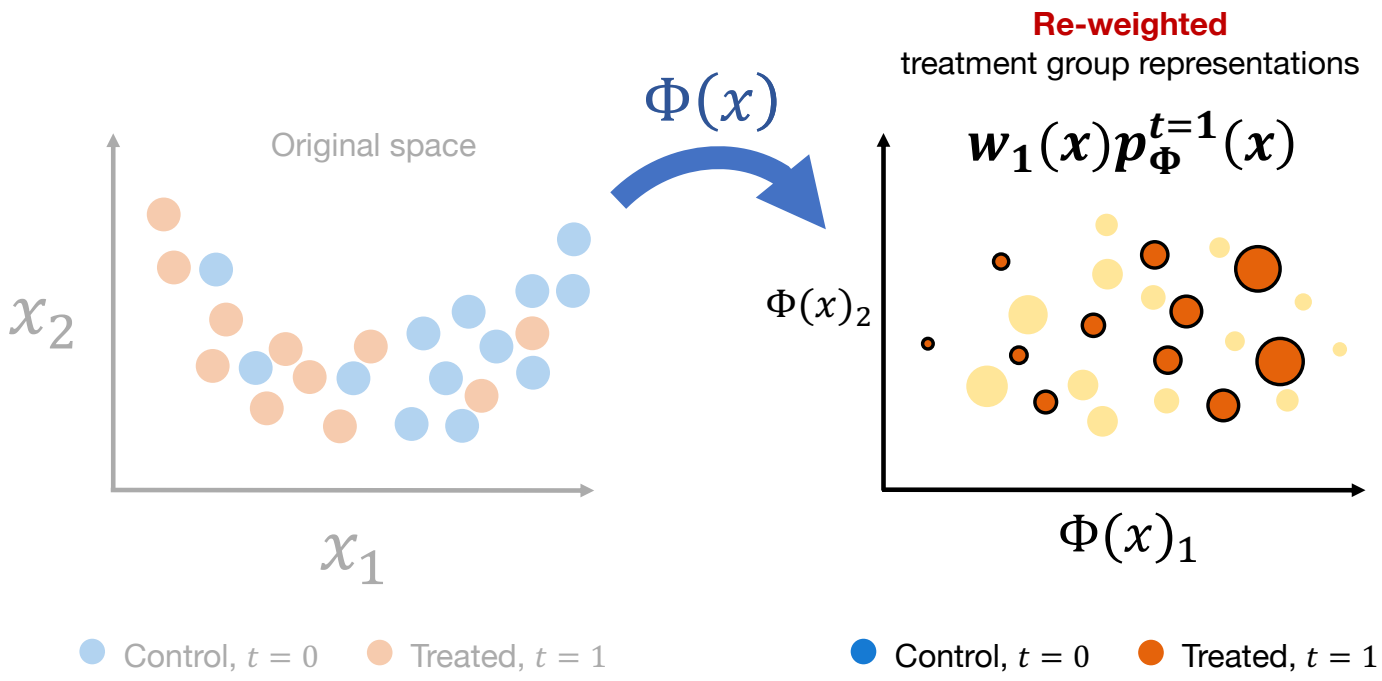
Representation learning

- Re-weighted representations for improved treatment group balance



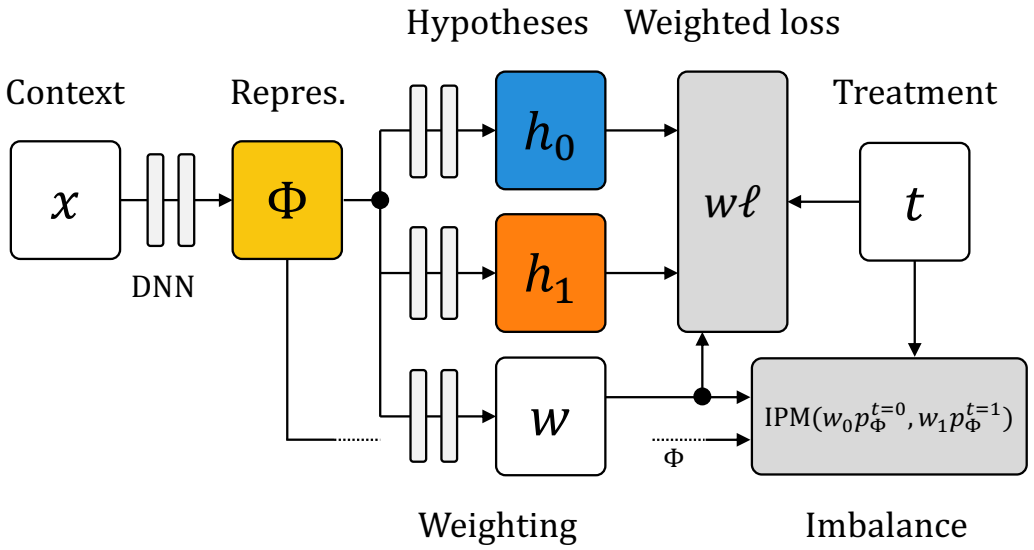
Representation learning

- Re-weighted representations for improved treatment group balance



Trading off accuracy for balance

- Our full architecture learns a representation $\Phi(\mathrm{x})$, 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)

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

- Letting $\Phi(x) = x$, and $w_t(x)$ be inverse propensity weights, we recover classic result

Individual-treatment effect generalization bound

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

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

- 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 ITE error

*Extension to finite samples available

Evaluating 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

Empirical results: IHDP (Hill, 2011)

- IHDP is a widely used benchmark for causal effect estimation
- Original **randomized** study examined the effect of home-visits and high-quality child care on child cognitive test scores
- Feature set contains aspects of the child, mother, pregnancy etc
- The benchmark was made **observational** by removing all non-white mothers from the dataset.
- The outcome was synthesized based on original features and treatment

Empirical results: IHDP

- Results on **held-out** units:

Error in conditional effect | **Error in average effect**

	IHDP	
	$\sqrt{\epsilon_{\text{CATE}}}$	ϵ_{ATE}
OLS/LR$_1$	$5.8 \pm .3$	$.94 \pm .06$
OLS/LR$_2$	$2.5 \pm .1$	$.31 \pm .02$
BLR	$5.8 \pm .3$	$.93 \pm .05$
k-NN	$4.1 \pm .2$	$.79 \pm .05$
TMLE	†	†
BART	$2.3 \pm .1$	$.34 \pm .02$
R.FOR.	$6.6 \pm .3$	$.96 \pm .06$
C.FOR.	$3.8 \pm .2$	$.40 \pm .03$
BNN	$2.1 \pm .1$	$.42 \pm .03$
TARNET	$.95 \pm .02$	$.28 \pm .01$
CFR$_{\text{MMD}}$	$.78 \pm .02$	$.31 \pm .01$
CFR$_{\text{WASS}}$	$.76 \pm .02$	$.27 \pm .01$
RCFR	$.67 \pm .05$	—

Concatenating Φ and T — BNN

Twin-head neural net ($\alpha = 0$) — TARNET

+ IPM regularization — CFR$_{\text{MMD}}$, CFR$_{\text{WASS}}$

+ Re-weighting — RCFR

Empirical results: IHDP

- Results on **held-out** units:

	Error in conditional effect	Error in average effect
	IHDP	
	$\sqrt{\epsilon_{\text{CATE}}}$	ϵ_{ATE}
OLS/LR$_1$	$5.8 \pm .3$	$.94 \pm .06$
OLS/LR$_2$	$2.5 \pm .1$	$.31 \pm .02$
BLR	$5.8 \pm .3$	$.93 \pm .05$
k-NN	$4.1 \pm .2$	$.79 \pm .05$
TMLE	†	†
BART	$2.3 \pm .1$	$.34 \pm .02$
R.For.	$6.6 \pm .3$	$.96 \pm .06$
C.For.	$3.8 \pm .2$	$.40 \pm .03$
Concatenating Φ and T — BNN	$2.1 \pm .1$	$.42 \pm .03$
Twin-head neural net ($\alpha = 0$) — TARNet	$.95 \pm .02$	$.28 \pm .01$
+ IPM regularization — CFR$_{\text{MMD}}$	$.78 \pm .02$	$.31 \pm .01$
+ IPM regularization — CFR$_{\text{WASS}}$	$.76 \pm .02$	$.27 \pm .01$
+ Re-weighting — RCFR	$.67 \pm .05$	—

Empirical results: IHDP

- Results on **held-out** units:

	IHDP	
	$\sqrt{\epsilon_{\text{CATE}}}$	ϵ_{ATE}
OLS/LR$_1$	$5.8 \pm .3$	$.94 \pm .06$
OLS/LR$_2$	$2.5 \pm .1$	$.31 \pm .02$
BLR	$5.8 \pm .3$	$.93 \pm .05$
k-NN	$4.1 \pm .2$	$.79 \pm .05$
TMLE	†	†
BART	$2.3 \pm .1$	$.34 \pm .02$
R.For.	$6.6 \pm .3$	$.96 \pm .06$
C.For.	$3.8 \pm .2$	$.40 \pm .03$
BNN	$2.1 \pm .1$	$.42 \pm .03$
TARNET	$.95 \pm .02$	$.28 \pm .01$
CFR$_{\text{MMD}}$	$.78 \pm .02$	$.31 \pm .01$
CFR$_{\text{WASS}}$	$.76 \pm .02$	$.27 \pm .01$
RCFR	$.67 \pm .05$	—

Concatenating Φ and T

Twin-head neural net ($\alpha = 0$)

+ IPM regularization

+ Re-weighting

Error in conditional effect

Error in average effect

Outline

- Background
- **Individual-level treatment effects**
- Latent, not hidden, confounders

Outline

- Background
- Individual-level treatment effects
- **Latent, not hidden, confounders**

Outline

- Background
- Individual-level treatment effects
- **Latent, not hidden, confounders**
  Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*

When can we hope to obtain causal effects from observational data?

- **Strong ignorability**: "No hidden confounders"

  All variables that affect both potential outcomes and *treatm* *measured*

Will be modified later

High-dimensional confounders

- Big-data gives us many potential confounders
- But what is the true confounder space?

High-dimensional confounders

- Big-data gives us many potential confounders
- But what is the true confounder space?
  - Is the value of my house a confounder?

High-dimensional confounders

- Big-data gives us many potential confounders
- But what is the true confounder space?
  - Is the value of my house a confounder?
  - How about my credit score, that lab test from last year, what car I drive, my internet browsing habits…

Confounders and proxies

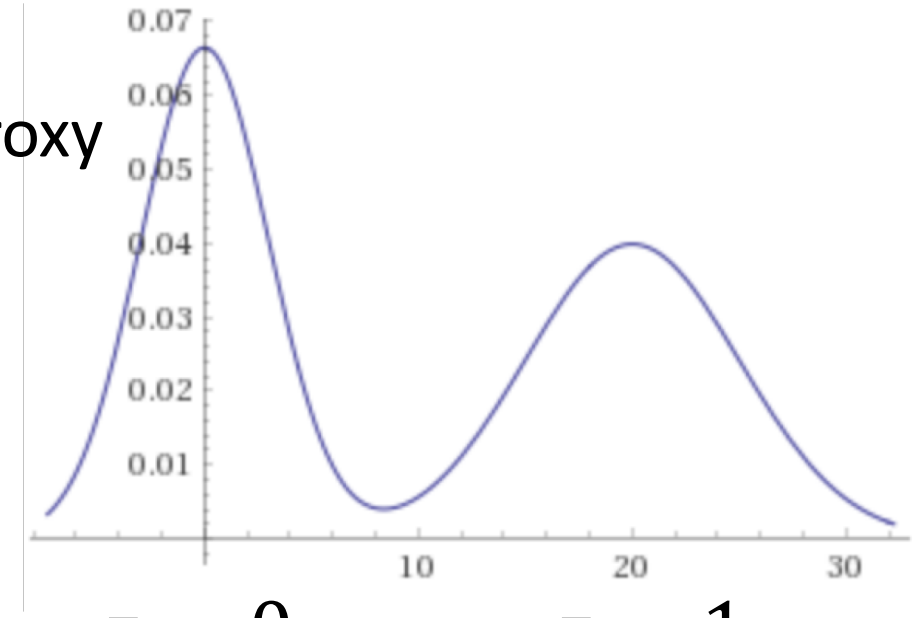
- With “big data” we hope we have information on more confounders
- Example: information about social-economic status
  - House value
  - Job title
  - Education level
  - Credit score
- What is the true confounder here?
  These are often just **proxies**
  which might not affect the treatment or outcome directly

Confounders and proxies

- With “big data” we hope we have information on more confounders
- Example: information about social-economic status
  - House value
  - Job title
  - Education level
  - Credit score
- What is the true confounder here? These are often just proxies
- *Theorem (*Essentially *Wickens, 1972, Frost, 1979)*
  Controlling for noisy proxies of confounders can induce bias in causal estimates, even with infinite samples

observed proxy

x



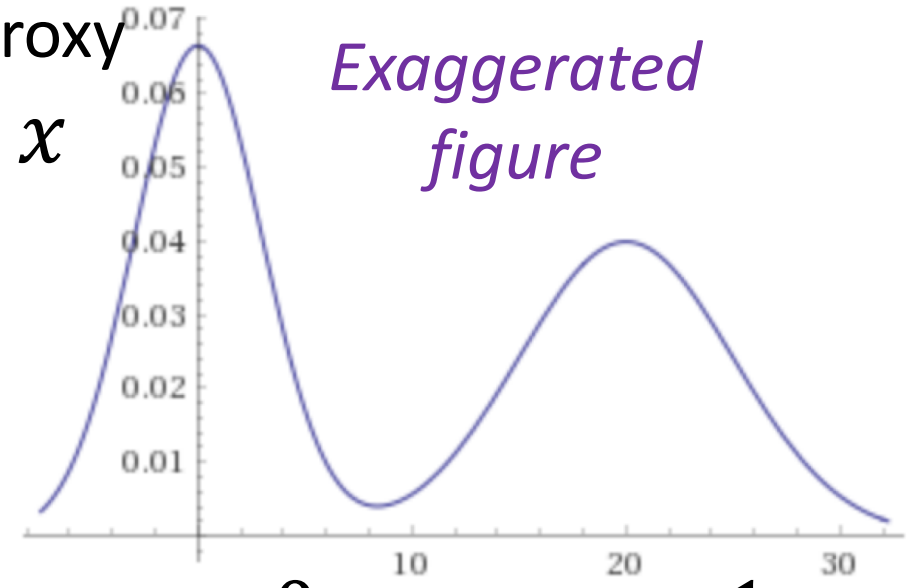
$z = 0$ $z = 1$

latent

confounder

proxy

x



Exaggerated figure

$z = 0$ $z = 1$

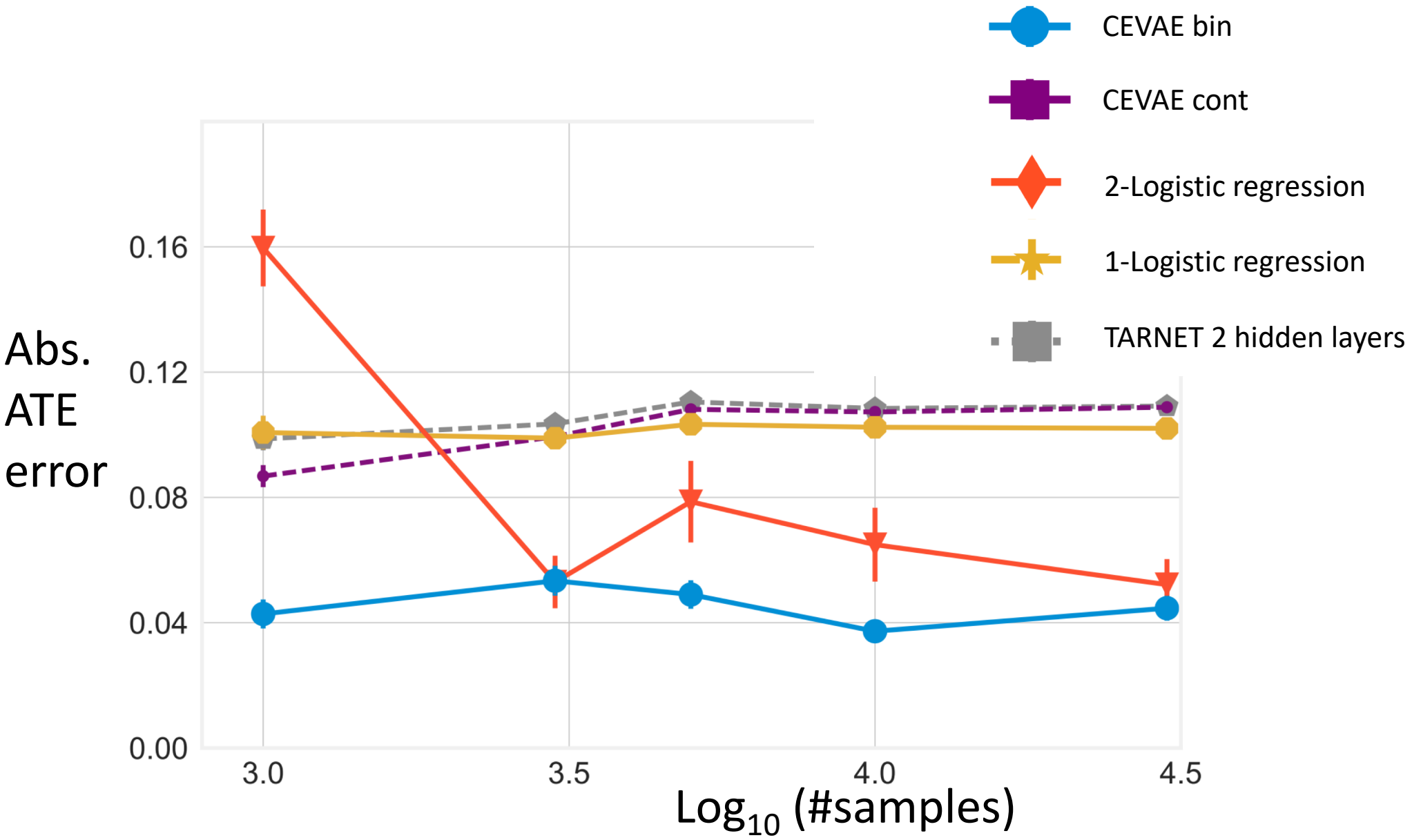
latent confounder

$$z \sim Bern\{0.5\}$$

$$t|z \sim Bern\left\{\frac{3}{4}z + \frac{1}{4}(1-z)\right\}$$

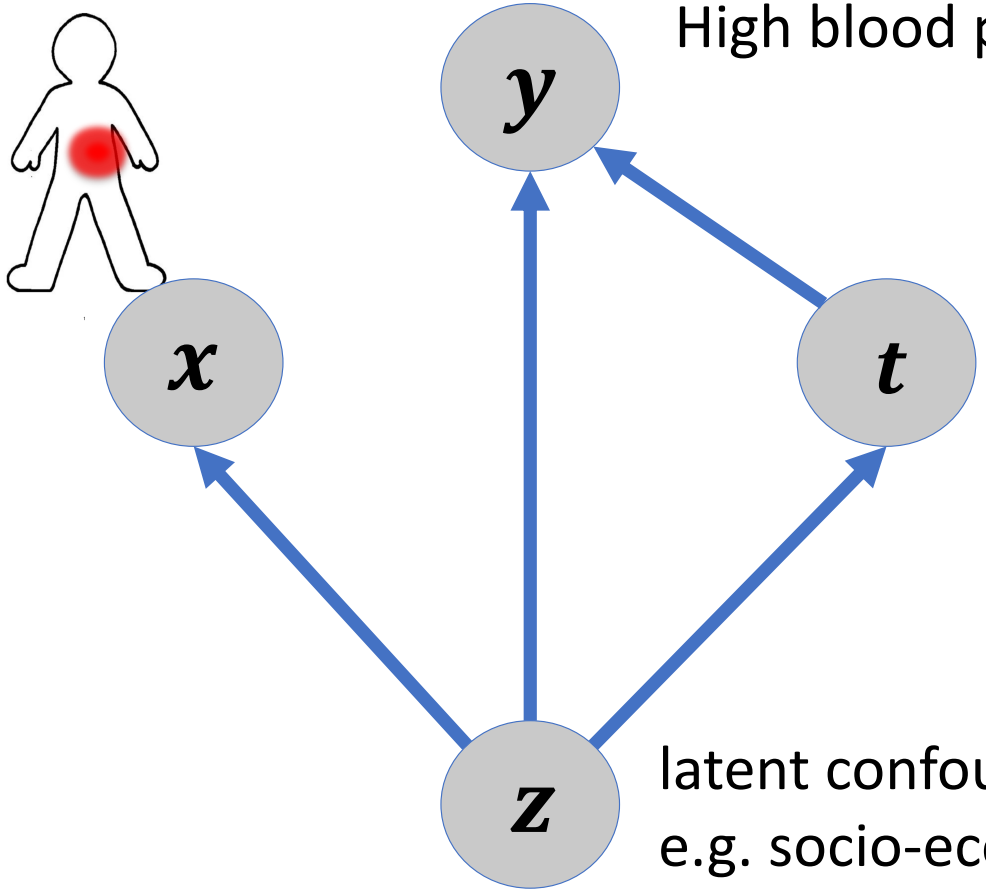
$$Y_1|t,z \sim Bern\{\text{expit}(3(z+2))\}$$

$$Y_0|z \sim Bern\{\text{expit}(3(z-2))\}$$


CEVAE bin
CEVAE cont
2-Logistic regression
1-Logistic regression
TARNET 2 hidden layers
Abs. ATE error
0.16
0.12
0.08
0.04
0.00
3.0
3.5
4.0
4.5
Log10 (#samples)

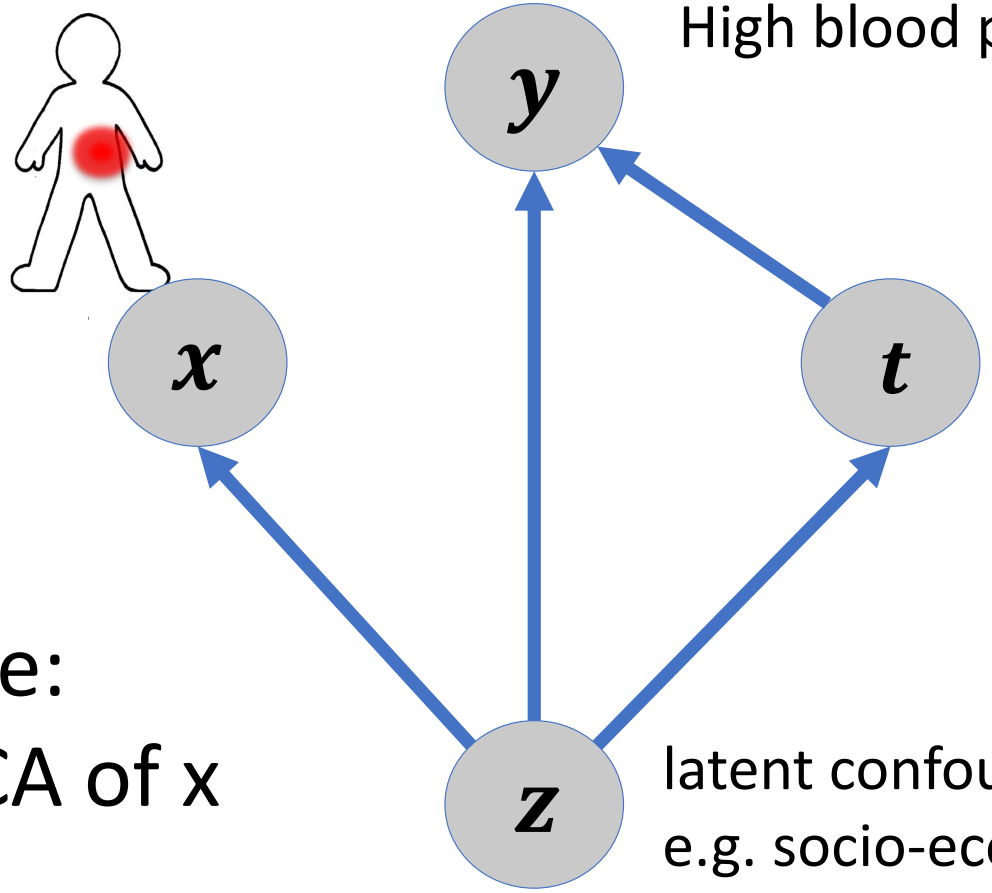
Generative Bayesian modeling of latent confounder space

Anna's measured confounders
e.g. home value, browsing history, credit score …



Generative Bayesian modeling of latent confounder space

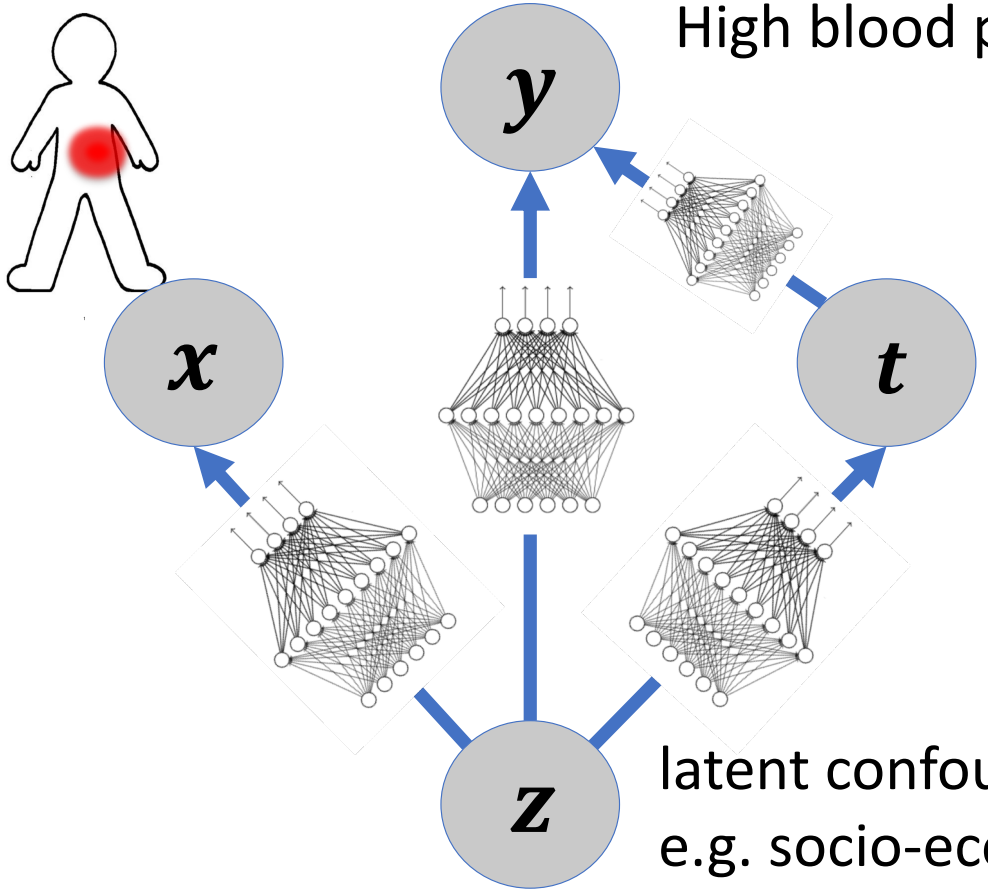
Anna's measured confounders
e.g. home value, browsing history, credit score …



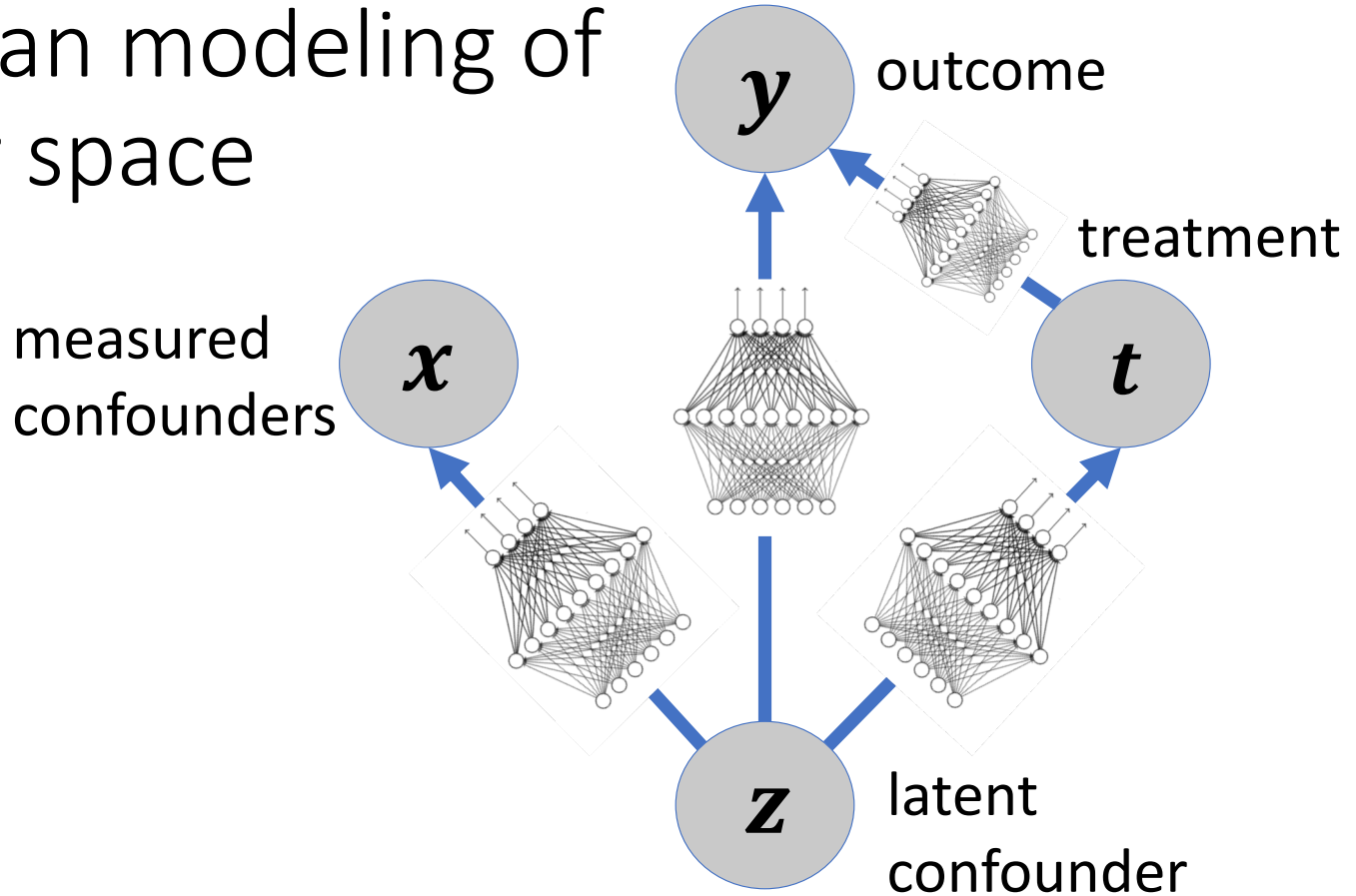
Simple example:
z is low-dim PCA of x

Generative Bayesian modeling of latent confounder space

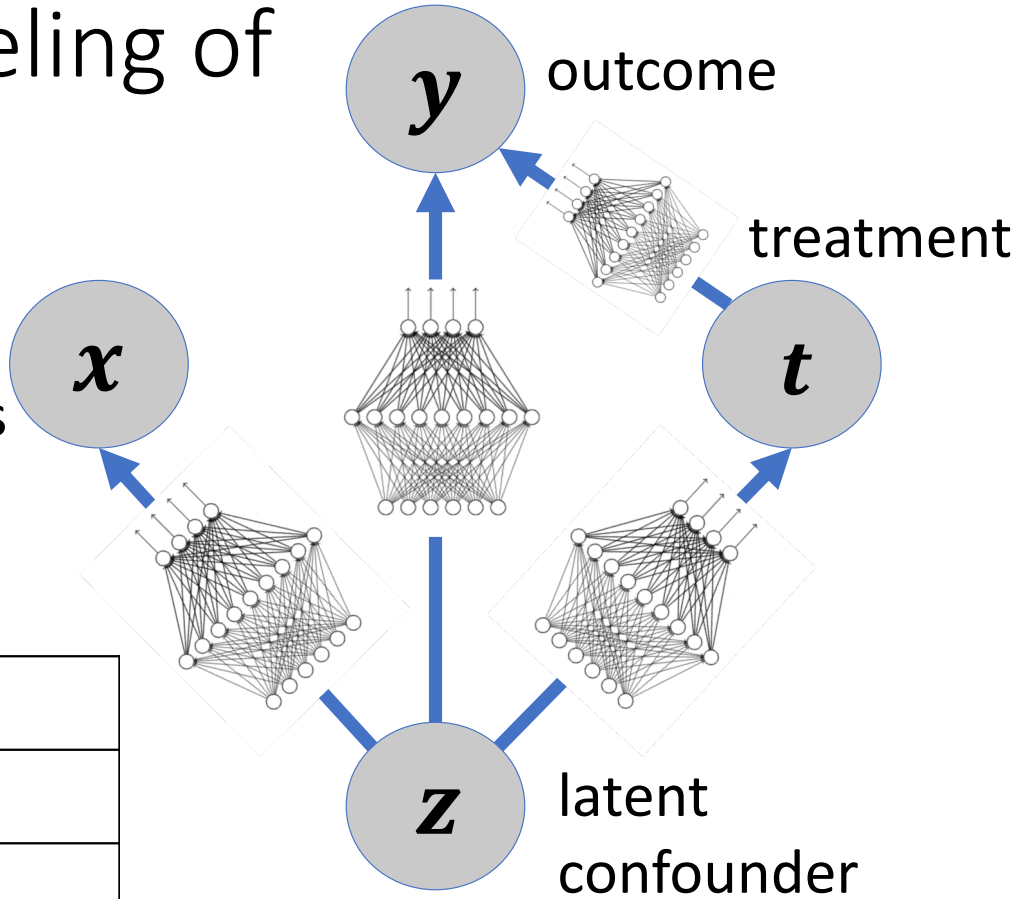
Anna's measured confounders
e.g. home value, browsing history, credit score …



Generative Bayesian modeling of latent confounder space

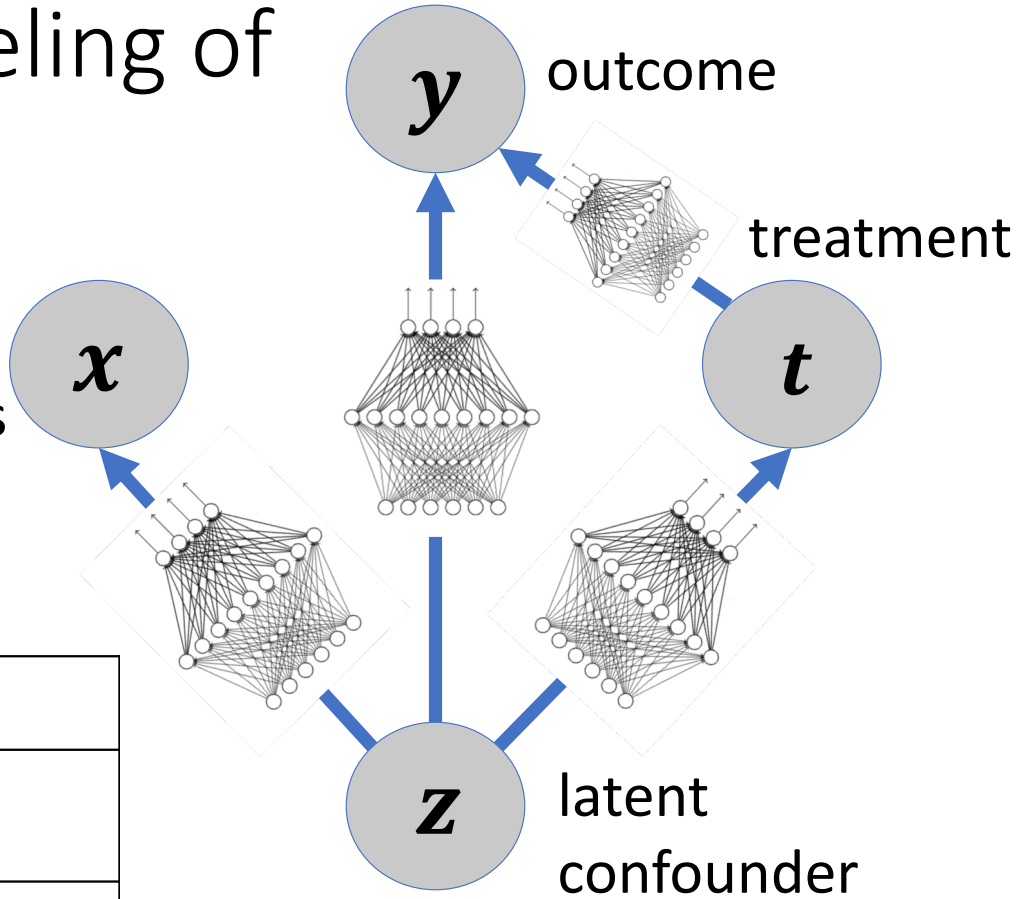


Generative Bayesian modeling of latent confounder space



Latent state:	$z \sim \mathcal{N}(\mu, \Sigma)$

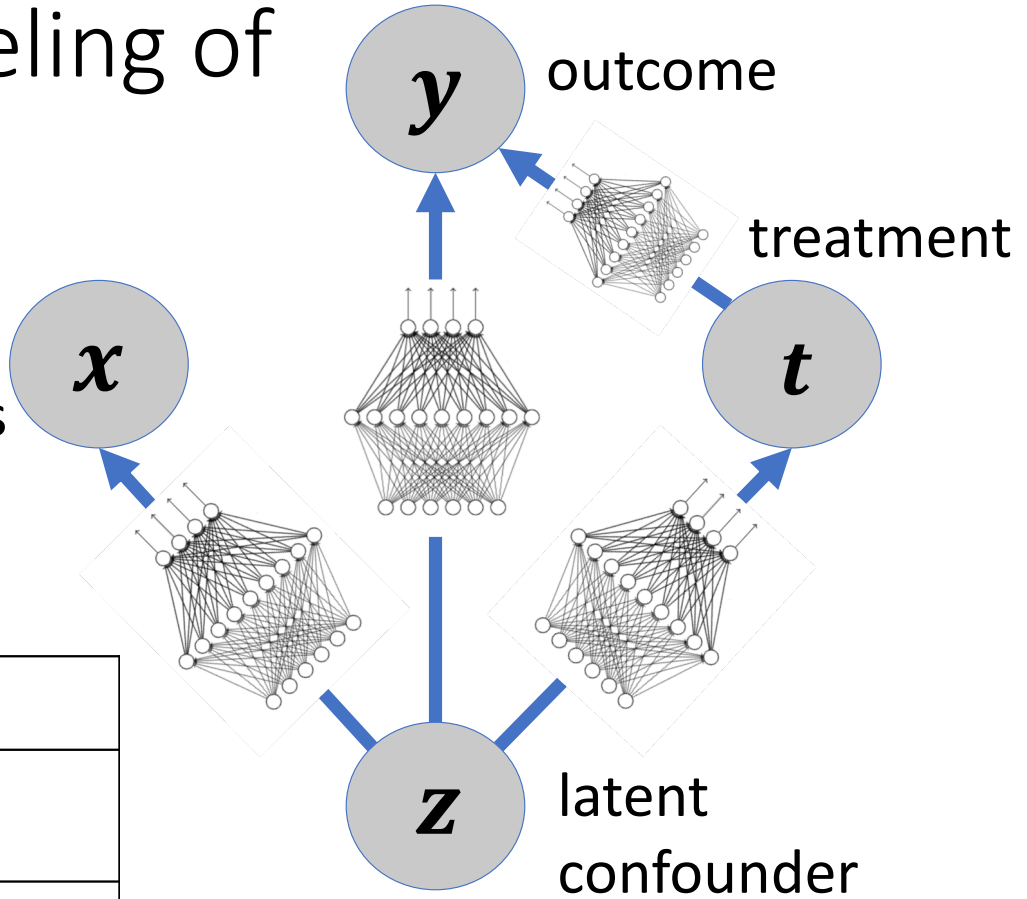
Generative Bayesian modeling of latent confounder space



Latent state:	$z \sim \mathcal{N}(\mu, \Sigma)$
Proxy process:	$x \sim \Pi\left(NN_{\theta_1}(z)\right)$

Π stands for any appropriate distribution: exponential, Poisson, Gamma, multinomial, etc.

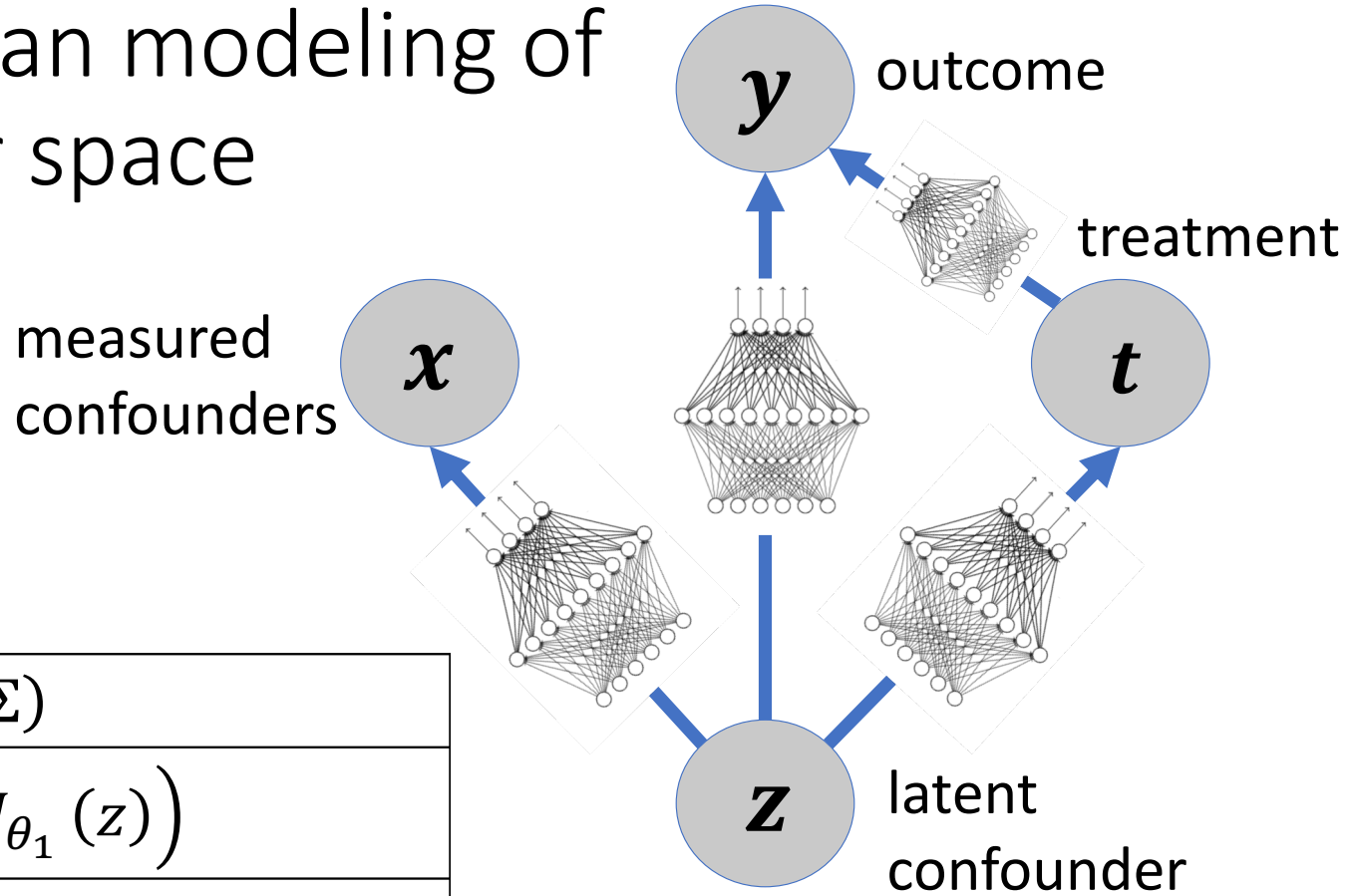
Generative Bayesian modeling of latent confounder space



Latent state:	$z \sim \mathcal{N}(\mu, \Sigma)$
Proxy process:	$x \sim \Pi\left(NN_{\theta_1}(z)\right)$
Treatment process:	$t \sim Bernoulli\left(NN_{\theta_2}(z)\right)$

Π stands for any appropriate distribution: exponential, Poisson, Gamma, multinomial, etc.

Generative Bayesian modeling of latent confounder space



Latent state:	$z \sim \mathcal{N}(\mu, \Sigma)$
Proxy process:	$x \sim \Pi\left(NN_{\theta_1}(z)\right)$
Treatment process:	$t \sim Bernoulli\left(NN_{\theta_2}(z)\right)$
Outcome:	$y \sim \Pi\left(TARNET_{\theta_3}(z,t)\right)$

Π stands for any appropriate distribution: exponential, Poisson, Gamma, multinomial, etc.

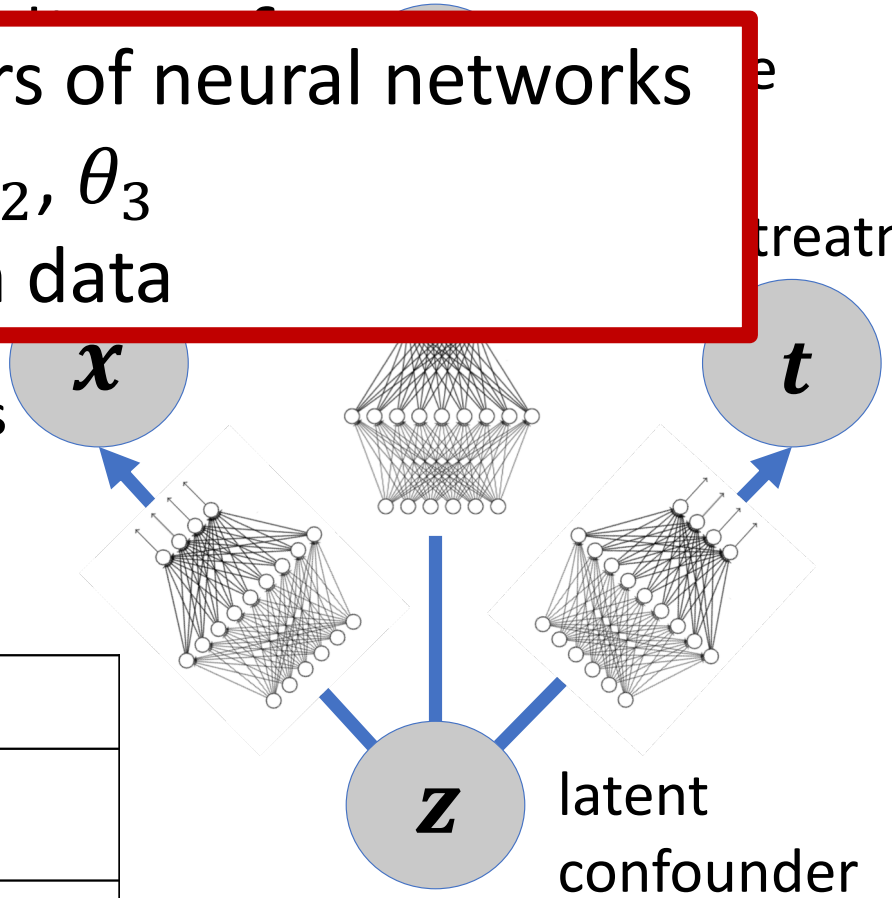
Generat

latent c

Learn the parameters of neural networks

$\theta_1, \theta_2, \theta_3$

from data



Latent state:	$z \sim \mathcal{N}(\mu, \Sigma)$
Proxy process:	$x \sim \Pi\left(NN_{\theta_1}(z)\right)$
Treatment process:	$t \sim Bernoulli\left(NN_{\theta_2}(z)\right)$
Outcome:	$y \sim \Pi\left(TARNET_{\theta_3}(z,t)\right)$

Generat

latent c

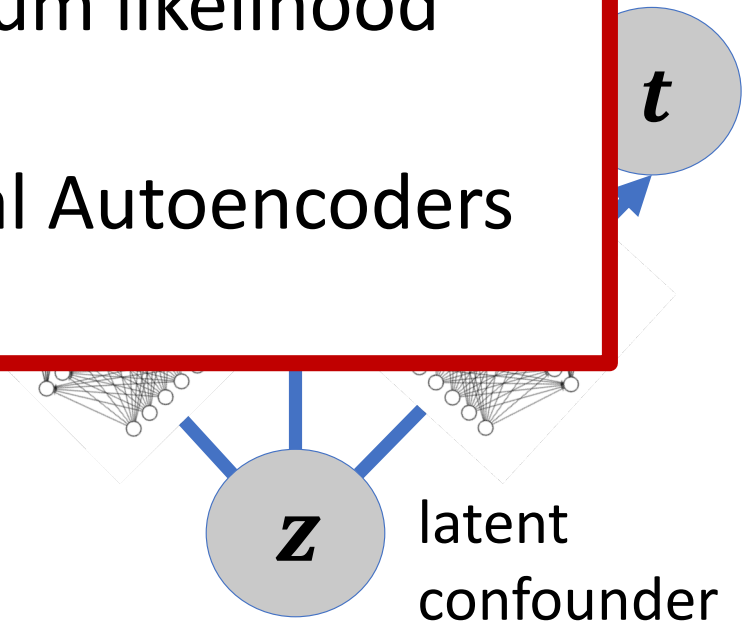
Learn the parameters of neural networks

$\theta_1, \theta_2, \theta_3$

from data via maximum likelihood

Causal Effect Variational Autoencoders

(CEVAE)



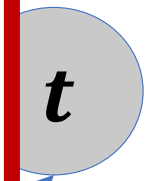
Latent state:	$z \sim \mathcal{N}(\mu, \Sigma)$
Proxy process:	$x \sim \Pi\left(NN_{\theta_1}(z)\right)$
Treatment process:	$t \sim Bernoulli\left(NN_{\theta_2}(z)\right)$
Outcome:	$y \sim \Pi\left(TARNET_{\theta_3}(z, t)\right)$

Generat
latent c

Learn the parameters of neural networks
$\theta_1, \theta_2, \theta_3$
from data via maximum likelihood

Causal Effect Variational Autoencoders
(CEVAE)

Based on Variational Autoencoders,
Kingma & Welling (2014)
Rezende et al. (2014)



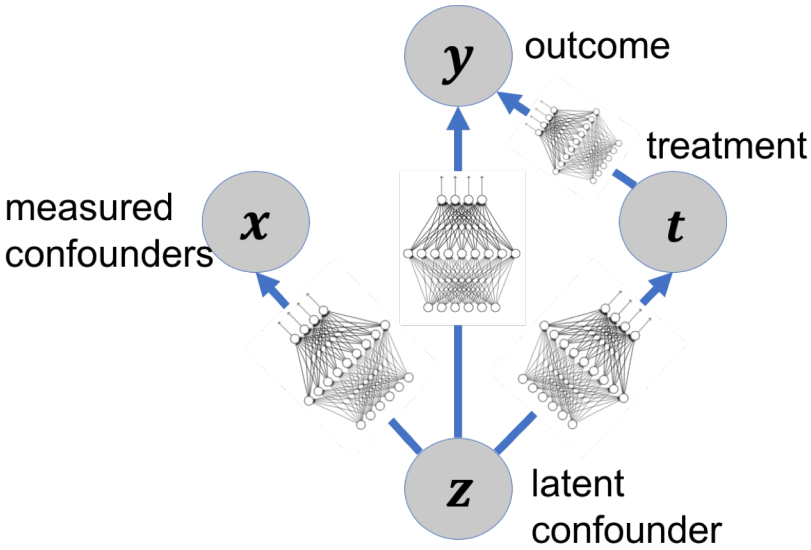
treatment

nder

Latent state:	
Proxy process	
Treatment process:	
Outcome:	$y \sim \Pi\left(TARNET_{\theta_3}(z,t)\right)$

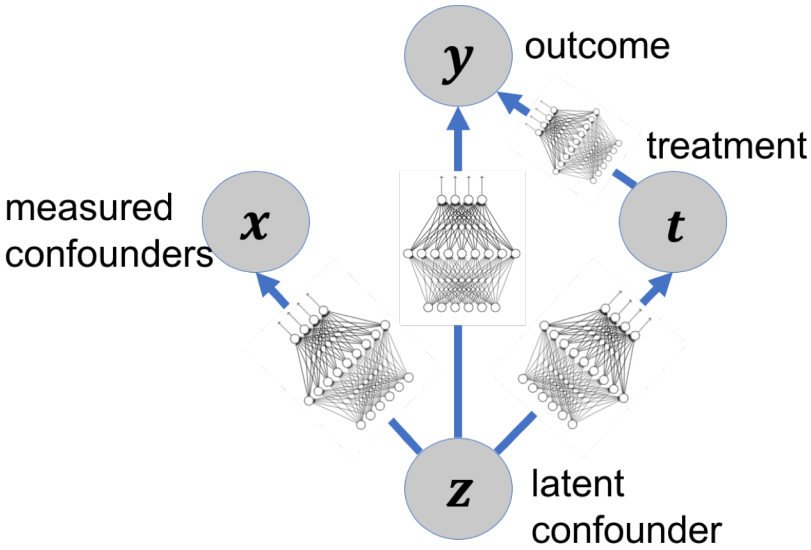
How to use Causal Effect Variation Autoencoder

- Learn parameters of networks for $p_{\theta_1}(x|z), p_{\theta_2}(t|z), p_{\theta_3}(y|t,z)$ from observational data



How to use Causal Effect Variation Autoencoder

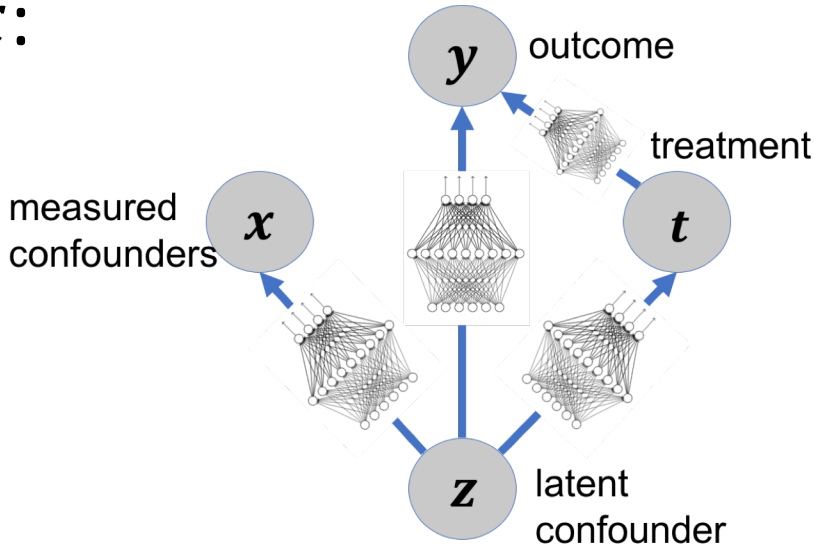
- Learn parameters of networks for $p_{\theta_1}(x|z), p_{\theta_2}(t|z), p_{\theta_3}(y|t,z)$ from observational data
- Learn approximate deep inverse model $q(z|x)$



How to use Causal Effect Variation Autoencoder

- Learn parameters of networks for $p_{\theta_1}(x|z), p_{\theta_2}(t|z), p_{\theta_3}(y|t,z)$ from observational data
- Learn approximate deep inverse model $q(z|x)$
- Given sample (patient) with features x:

1. Sample $\tilde{z}_k \sim q(z|x)$ for k=1,…, 100
2. $\hat{Y}_1 = \frac{1}{100} \sum_{k=1}^{100} p_{\theta_3}(y|t=1, \tilde{z}_k)$
3. $\hat{Y}_0 = \frac{1}{100} \sum_{k=1}^{100} p_{\theta_3}(y|t=0, \tilde{z}_k)$



Experiments with proxies and latent confounders – the TWINS dataset

- Records of 12,000 pairs of same-gender twins born in the US 1989-1991 weighing under 2kg
- Treatment: being born the heavier twin
- Outcome: 1-year mortality
  Y_1 = mortality of heavier twin
  Y_0 = mortality of lighter twin
- Baseline covariates: parents' socio-economic, pregnancy risk factors, birth factors, length of gestation
- $\mathbb{E}[Y_1] = 16.4\%$

$$\mathbb{E}[Y_0] = 18.9\%$$

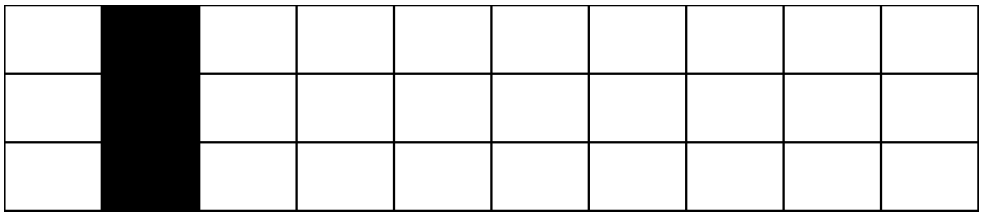
$$\text{Average Treatment Effect} = -2.5\%$$

Experiments with proxies and latent confounders – the TWINS dataset

- Length of gestation highly predictive of mortality
- We remove the gestation length covariate and replace it with increasingly noisy proxies:
  - Original is categorical with 10 categories
  - Encode with 3 replications of one-hot-encoding →30 binary covariates

Experiments with proxies and latent confounders – the TWINS dataset

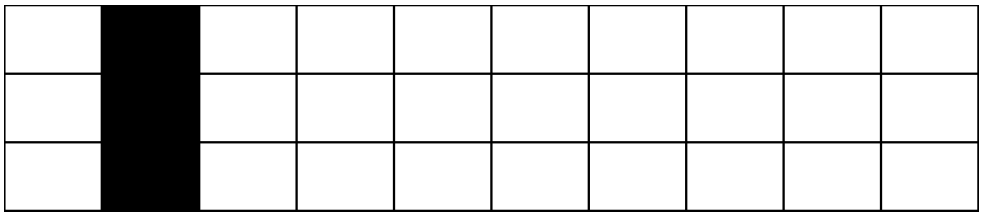
- Length of gestation highly predictive of mortality
- We remove the gestation length covariate and replace it with increasingly noisy proxies:
  - Original is categorical with 10 categories
  - Encode with 3 replications of one-hot-encoding →30 binary covariates

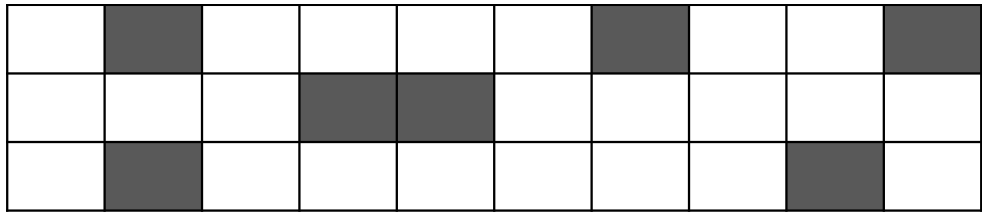


Original encoding

Experiments with proxies and latent confounders – the TWINS dataset

- Length of gestation highly predictive of mortality
- We remove the gestation length covariate and replace it with increasingly noisy proxies:
  - Original is categorical with 10 categories
  - Encode with 3 replications of one-hot-encoding →30 binary covariates
  - Flip each of the 30 covariates at random with probability $p \in \{0.05, \ldots, 0.5\}$



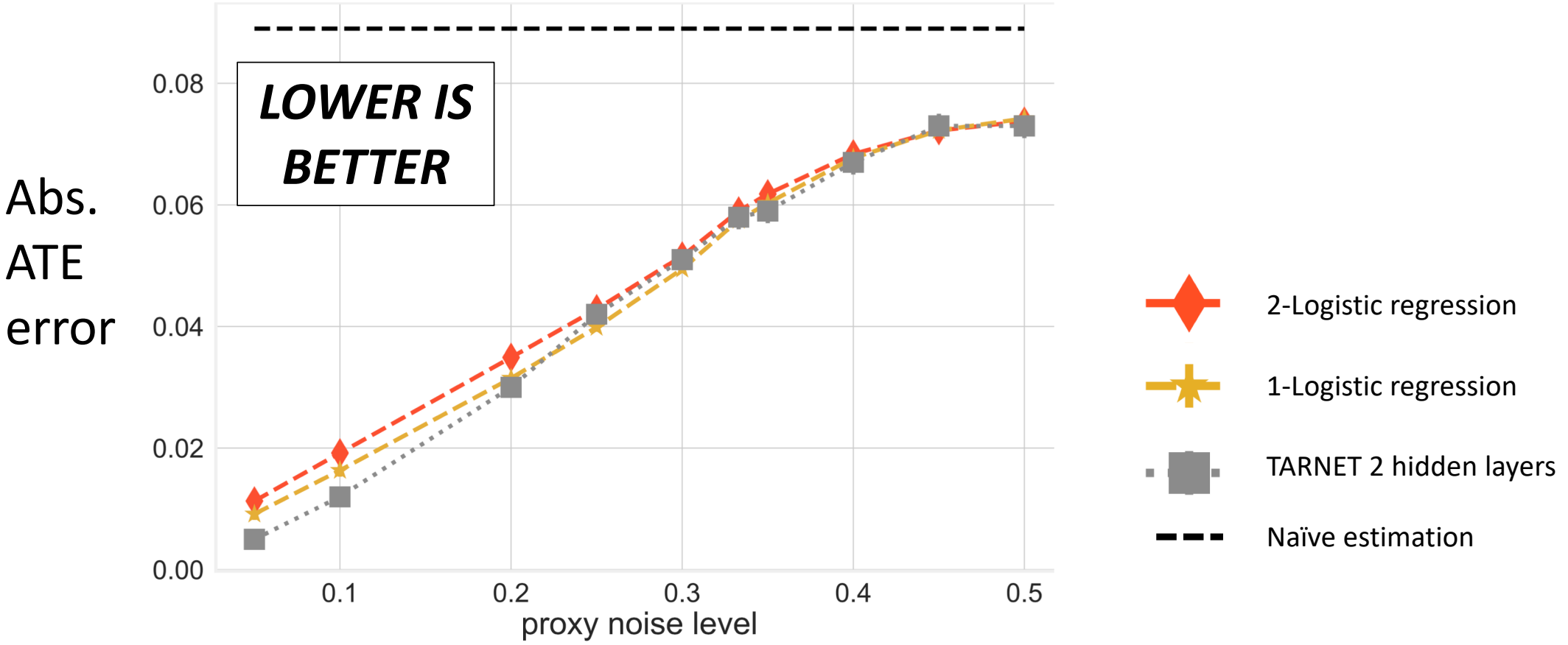


Original encoding → Noisy proxy

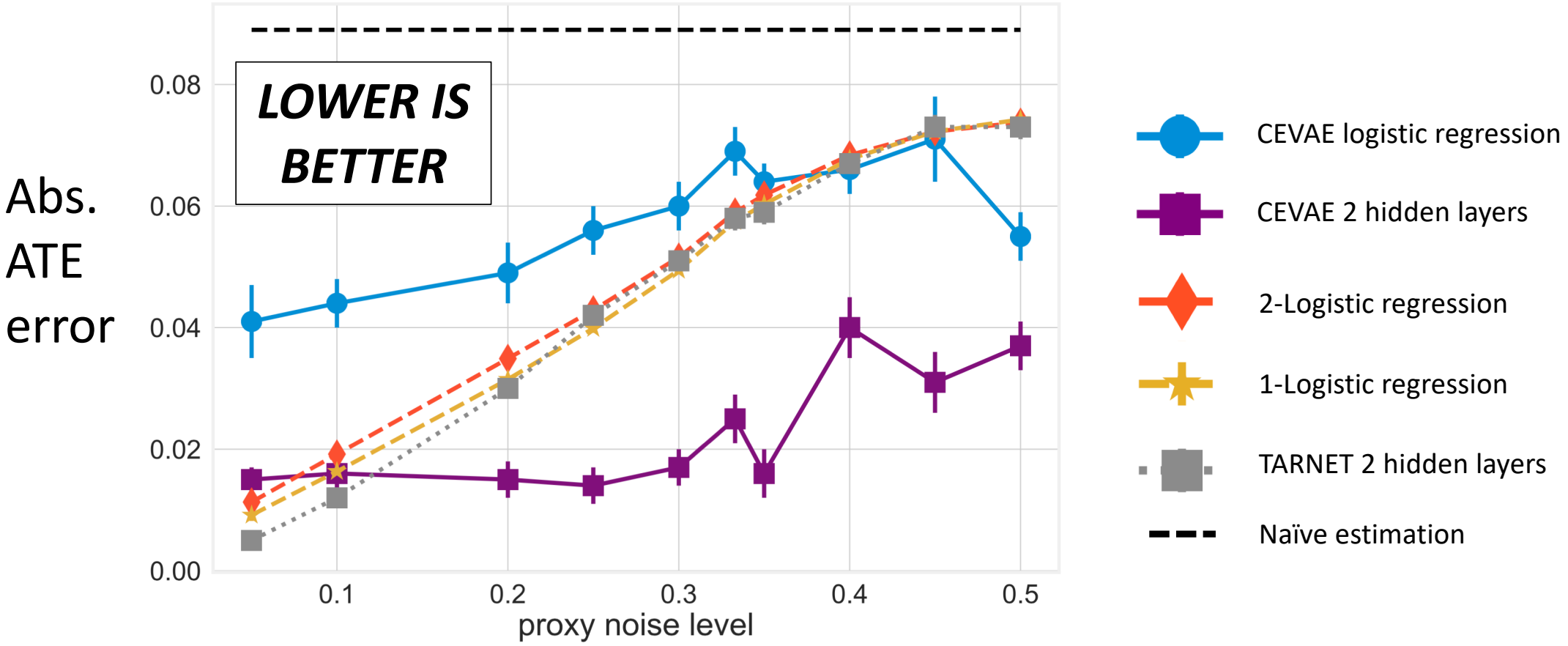
Experiments with proxies and latent confounders – the TWINS dataset

- Length of gestation highly predictive of mortality
- We remove the gestation length covariate and replace it with increasingly noisy proxies:
  - Original is categorical with 10 categories
  - Encode with 3 replications of one-hot-encoding →30 binary covariates
  - Flip each of the 30 covariates at random with probability $p \in \{0.05, \ldots, 0.5\}$
  - Other covariates are "natural" proxies for gestation length, e.g. incompetent cervix risk factor
- Observe only one of each twin-pair, with a biased process based on baseline covariates with emphasis on gestation length
- Tasks:
  - predict ATE
  - predict mortality of unobserved twin

Experiments with proxies and latent confounders – the TWINS dataset

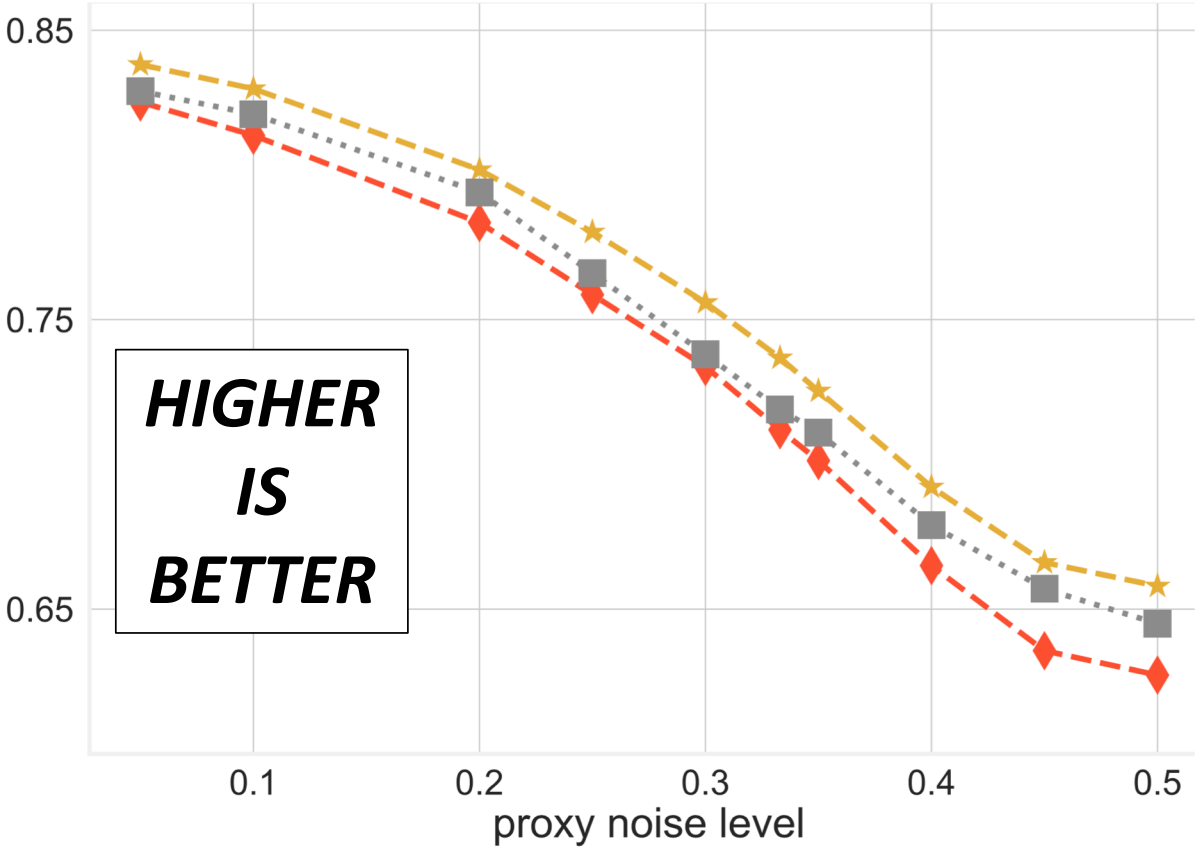


Experiments with proxies and latent confounders – the TWINS dataset



Experiments with proxies and latent confounders – the TWINS dataset

Counter-factual AUC



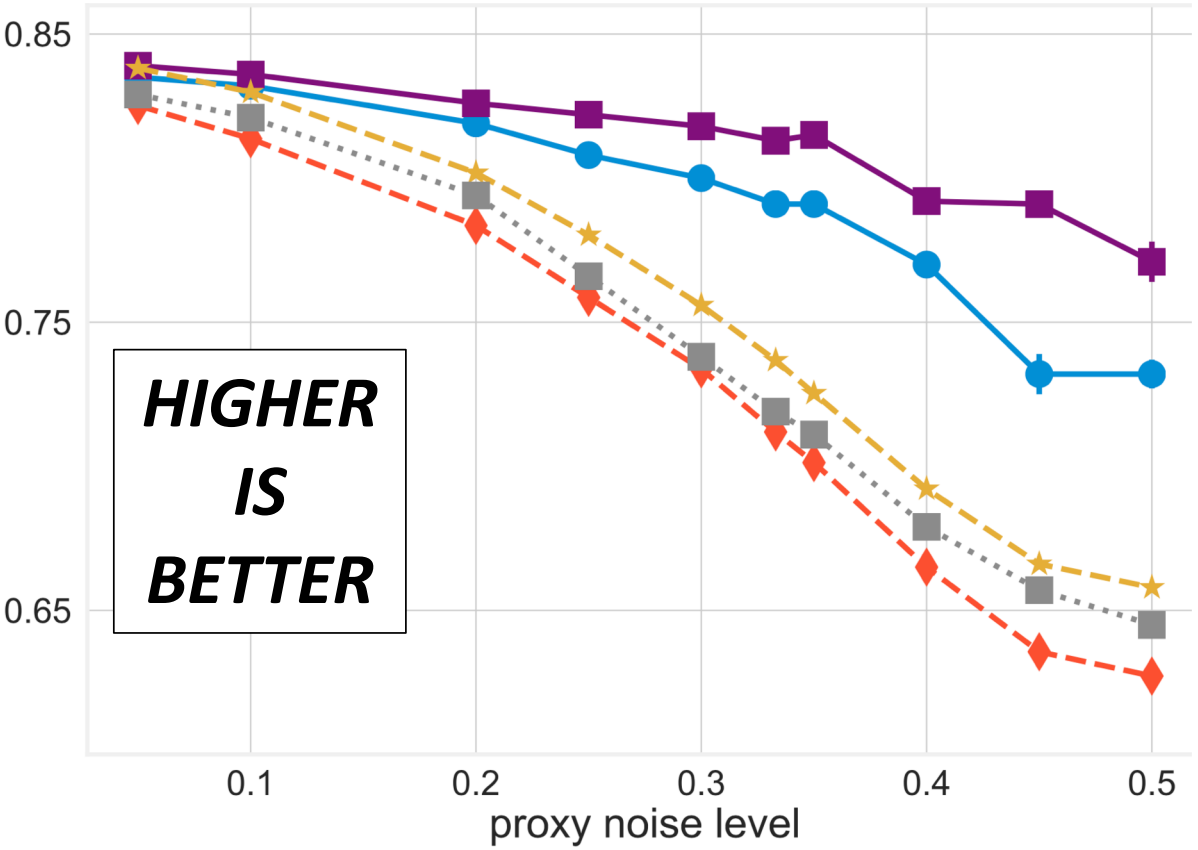
2-Logistic regression

1-Logistic regression

TARNET 2 hidden layers

Experiments with proxies and latent confounders – the TWINS dataset

Counter-factual AUC



CEVAE 2 hidden layers

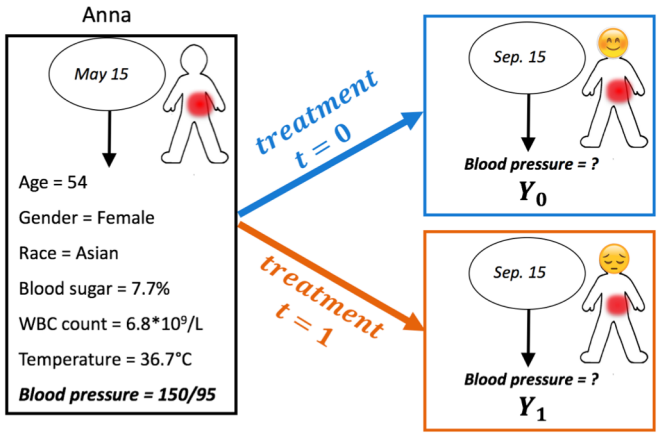
CEVAE logistic regression

2-Logistic regression

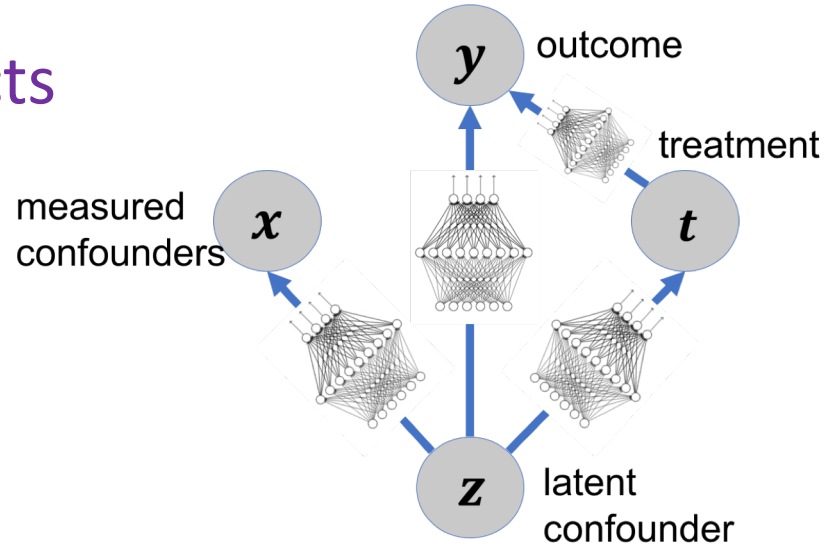
1-Logistic regression

TARNET 2 hidden layers

What causal inference problems might benefit from Deep Learning?



- Modeling high-dimensional relationships
  - Observational dataset are often high-dim
- Optimizing over rich function classes
  - Learning individual-level treatment effects
- Finding latent structure
  - Unmeasured confounders with proxies



More Deep Learning + Causal effect inference work

- *Counterfactual Prediction with Deep Instrumental Variables Networks*
  Hartford, Lewis, Leyton-Brown & Taddy

- *Deep Match: Balancing Deep Covariate Representations*
  Nathan Kallus

- *Matching on Balanced Nonlinear Representations*
  Sheng Li & Yun Fu

Practical advice (with personal bias)

- How to use modern Deep Learning for causal inference is still very much an open problem
- I advise caution

Some practical advice (with personal bias)

When to use

- Many proxies
- Estimating high-dim CATE, individual-level treatment effects
- Image, text, or voice confounders

Some practical advice (with personal bias)

When to use

- Many proxies
- Estimating high-dim CATE, individual-level treatment effects
- Image, text, or voice confounders

When not to use

- Small set of confounders
- Preference for well-established methods
- Deep learning does badly at predicting outcome

Thank you to my collaborators

Fredrik Johansson (MIT)
David Sontag (MIT)
Rahul Krishnan (MIT)

Jennifer Hill (NYU)

Nathan Kallus (Cornell-Tech)

Christos Louizos (UVA)
Joris Mooij (UVA)
Max Welling (UVA)

Rich Zemel (U. Toronto)