KDD 2020 Tutorial

Machine Learning for Causal Inference

Sheng Li[1], Liuyi Yao[2], Yaliang Li[3], Zhixuan Chu[1], Jing Gao[2]

[1] University of Georgia, Athens, GA
[2] University at Buffalo, Buffalo, NY
[3] Alibaba Group, Bellevue, WA

- **A Survey** on Causal Inference. https://arxiv.org/abs/2002.02770

Outline

- **Overview**
- Causal Inference: Background and Challenges
- Classical Causal Inference Methods
- Subspace Learning for Causal Inference
- Deep Representation Learning for Causal Inference
- Applications and Potential Directions
- Conclusions

Causality

- *Causality* is also referred to as “causation”, or “cause and effect”
- Causality has been extensively discussed in many fields, such as statistics, philosophy, psychology, economics, education, and health care.

Figure: https://www.pinterest.com/pin/195906652514487365/

Correlation and Causation

- **Correlation does not imply causation**
- For two correlated events A and B, the possible relations might be: (1) A causes B, (2) B causes A, (3) A and B are consequences of a common cause, but do not cause each other, etc.
- Example of (3): As ice cream sales increase, the rate of drowning deaths increases sharply. The two events are correlated. However, increasing ice cream consumption and drowning deaths may not have causal relationships.

Causal Inference and Causal Discovery

- *Causal discovery* is identifying causal relationships from large quantities of data through computational methods
- *Causal inference* is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect
  - Experimental Study (e.g., randomized controlled trials)
  - Observational Study (e.g., potential outcome framework)

Experimental Study

- **Experimental Study**
  - Randomized Controlled Trial (RCT)
  - Assignment of control/treated is random
  - Study the effect of treatment (e.g. design A/B) to the outcome (e.g. conversion)
  - Gold-standard for studying causal relationships
  - Expensive and time-consuming

Observational Study

- **Observational Study**
  - Unlike RCTs, treatment assignment in observational study is NOT random
  - Approaches: graphical causal models, potential outcome framework
  - Simple, efficient
  - Potential to leverage *Big Data*

About This Tutorial

- Causal inference is an active research area with many research topics, this tutorial mainly focuses on the potential outcome framework in observational study
- Machine learning could potentially assist causal inference at different stages. In this tutorial, we focus on how to design representation learning approaches for causal inference

About this Tutorial

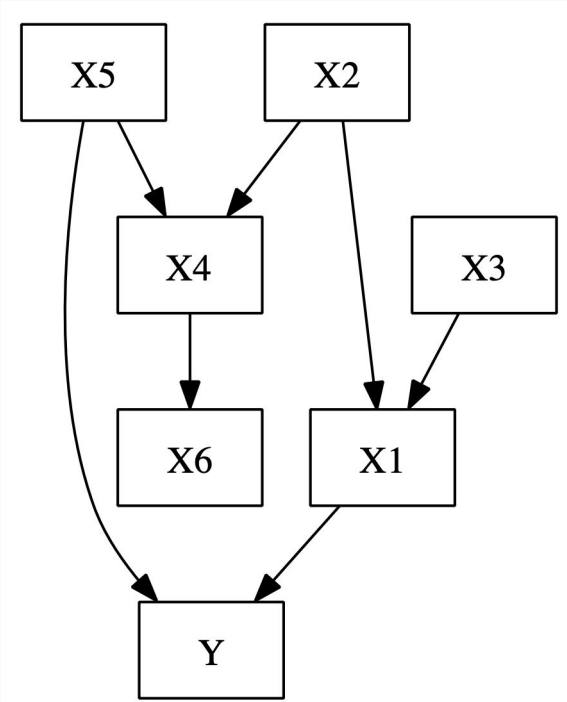
- **Schedule**
  - 8:30 AM - 8:45 AM: Overview (Gao)
  - 8:45 AM - 9:05 AM: Causal Inference: Background and Challenges (S. Li)
  - 9:05 AM - 9:35 AM: Classical Causal Inference Methods (Chu)
  - 9:35 AM - 10:00 AM: Subspace Learning for Causal Inference (S. Li)
  - 10:00 AM - 10:30 AM: Coffee Break
  - 10:30 AM - 11:20 AM: Deep Learning for Causal Inference (Yao)
  - 11:20 AM - 11:40 AM: Applications and Potential Directions (Y. Li)
  - 11:40 AM - 11:45 AM: Conclusions (Y. Li)
- **Website**: http://kdd2020tutorial.thumedialab.com/
- **A Survey** on Causal Inference (02/2020): https://arxiv.org/abs/2002.02770

Outline

- ❏ Overview
- ❏ **Causal Inference: Background and Challenges**
- ❏ Classical Causal Inference Methods
- ❏ Subspace Learning for Causal Inference
- ❏ Deep Representation Learning for Causal Inference
- ❏ Applications and Potential Directions
- ❏ Conclusions

Graphical Causal Models

- Causal graphs are probabilistic graphical models to encode assumptions about the data-generating process *[Pearl, 2009]*
- **D-separation** allows determining whether the causal structure implies that two sets of variables are independent given a third set
- Other related approaches: structural equation modeling (SEM)



An example of DAG

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.

Potential Outcome Framework (1)

- Also known as Rubin causal model (RCM), or Neyman–Rubin causal model
- **Unit**: A unit is the atomic research object in the causal study
- **Treatment**: An action that applies to a unit

  In the binary treatment case (i.e., $W = 0$ or 1), *treated* group contains units received treatment $W = 1$, while *control* group contains units received treatment $W = 0$
- **Outcome**: response of units after treatment/control, denoted as Y
- **Treatment Effect (or Causal Effects)**: The change of outcome when applying the different treatments on the units

Potential Outcome Framework (2)

- **Potential Outcome**: For each unit-treatment pair, the outcome of that treatment when applied on that unit is the potential outcome. $Y(W=w)$
- **Observed Outcome**: Outcome of treatment that is actually applied. In binary case, $Y^F = Y(W = w)$
- **Counterfactual Outcome**: Potential outcome of the treatments that the unit had not taken. In binary case, $Y^{CF} = Y(W = 1 - w)$
- A unit can only take one treatment. Thus, counterfactual outcomes are not observed, leading to the well-known "*missing data*" problem

[Holland, 1986] P. Holland, "Statistics and causal inference." *Journal of the American Statistical Association*, 81.396 (1986): 945-960.

Potential Outcome Framework (3)

- **Treatment Effects** can be defined at the population, treated group, subgroup and individual levels
- *Population Level*: Average Treatment Effect (**ATE**)

$$\text{ATE} = \mathbb{E}[\mathbf{Y}(W=1) - \mathbf{Y}(W=0)]$$

- *Treated group*: Average Treatment Effect on the Treated Group (**ATT**)

$$\text{ATT} = \mathbb{E}[\mathbf{Y}(W=1)|\mathbf{W}=1] - \mathbb{E}[\mathbf{Y}(W=0)|\mathbf{W}=1]$$

- *Subgroup*: Conditional Average Treatment Effect (**CATE**)

$$\text{CATE} = \mathbb{E}[\mathbf{Y}(W=1)|X=x] - \mathbb{E}[\mathbf{Y}(W=0)|X=x]$$

- *Individual*: Individual Treatment Effect (**ITE**)

$$\text{ITE}_i = Y_i(W=1) - Y_i(W=0)$$

An Illustrative Example

- **Task:** Evaluate the treatment effects of several different medications for one disease, by exploiting the observational data, such as the electronic health records (EHR)
- **Observational data** may include: (1) demographic information of patients, (2) specific medication with the specific dosage taken by patients, and (3) the outcome of medical tests
- **Units**: patients
- **Treatments**: different medications
- **Outcome**: recovery, blood test results, or others

Assumptions

- **Assumption 1**: Stable Unit Treatment Value Assumption (**SUTVA**)

  The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

- This assumption emphasizes that:
  - **Independence of each units**, i.e., there are no interactions between units. In our example, one patient's outcome will not affect other patients' outcomes
  - **Single version for each treatment**. For instance, one medicine with different dosages are different treatments under the SUTVA assumption

Assumptions

- **<u>Assumption 2</u>: Ignorability**

  *Given the background variable, **X**, treatment assignment W is independent of the potential outcomes, i.e.,* $W \perp\!\!\!\perp Y(W=0), Y(W=1)|X$

- Following our example, this assumption implies that:
  - If two patients have the same background variable ***X***, their potential outcome should be the **same** whatever the treatment assignment is.
  - Analogously, if two patients have the same background variable value, their ***treatment assignment mechanism*** should be same whatever the value of potential outcomes they have

Assumptions

- **<u>Assumption 3</u>: Positivity**

  *For any set of values of **X**, treatment assignment is not deterministic:*

$$P(W = w|X = X) > 0 \quad \forall w \text{ and } x.$$

- If treatment assignment for some values of X is deterministic, the outcomes of at least one treatment could never be observed. It would be unable and meaningless to estimate causal effects
- It implies “common support” or “overlap” of treated and control groups
- The ignorability and the positivity assumptions together are also called ***Strong Ignorability*** or ***Strongly Ignorable Treatment Assignment***

A Naive Solution

- The core problem in causal inference is: how to estimate the average potential treated/control outcomes over a specific group?
- One naive solution is to calculate the difference between the average treated and control outcomes, i.e.,

$$\hat{\mathrm{ATE}} = \frac{1}{N_T}\sum_{i=1}^{N_T} Y_i^F - \frac{1}{N_C}\sum_{i=1}^{N_C} Y_j^F$$

- However, this solution is not reasonable due to the existence of confounders

Confounders

- **Confounders**: Variables that affect both treatment assignment and outcome
- In the medical example, **age** is a confounder
  - Age would affect the recovery rate
  - Age would also affect the treatment choice

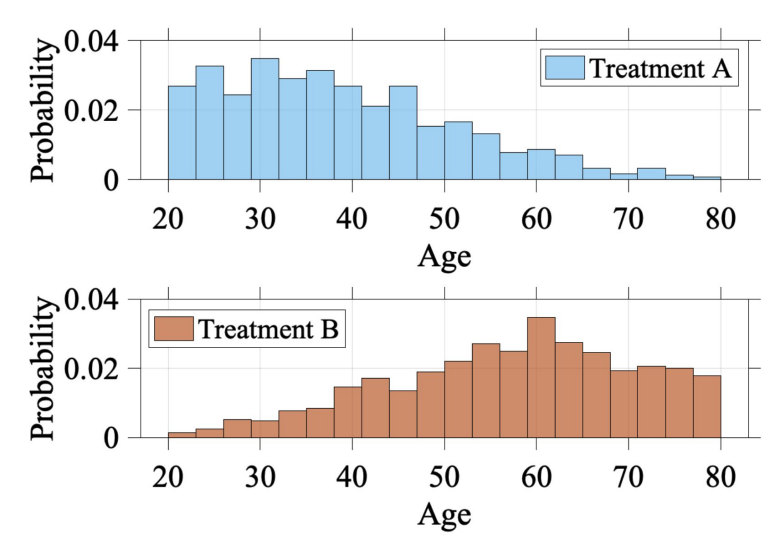
Recovery Rate / Treatment / Age	Treatment A	Treatment B
Young	234/270 = 87%	81/87 = **92%**
Older	55/80 = 69%	192/263 = **73%**
Overall	289/350 = **83%**	273/350 = 78%

Table 1. An example to show the spurious effect of confounder variable *Age*.

Simpson's paradox due to confounder

Selection Bias

- **Selection Bias**: The distribution of the observed group is not representative to the group we are interested in
- Confounder variables affect units' treatment choices, which leads to the selection bias
- Selection bias makes counterfactual outcome estimation more difficult



Discussions

- Estimating treatment effects from observational data
- Potential outcome framework
- Main challenges in causal inference from observational data
  - How to handle *confounders*?
  - How to deal with *selection bias*?

Outline

- ❏ Overview
- ❏ Causal Inference: Background and Challenges
- ❏ **Classical Causal Inference Methods**
- ❏ Subspace Learning for Causal Inference
- ❏ Deep Representation Learning for Causal Inference
- ❏ Applications and Potential Directions
- ❏ Conclusions

Classical Causal Inference Methods

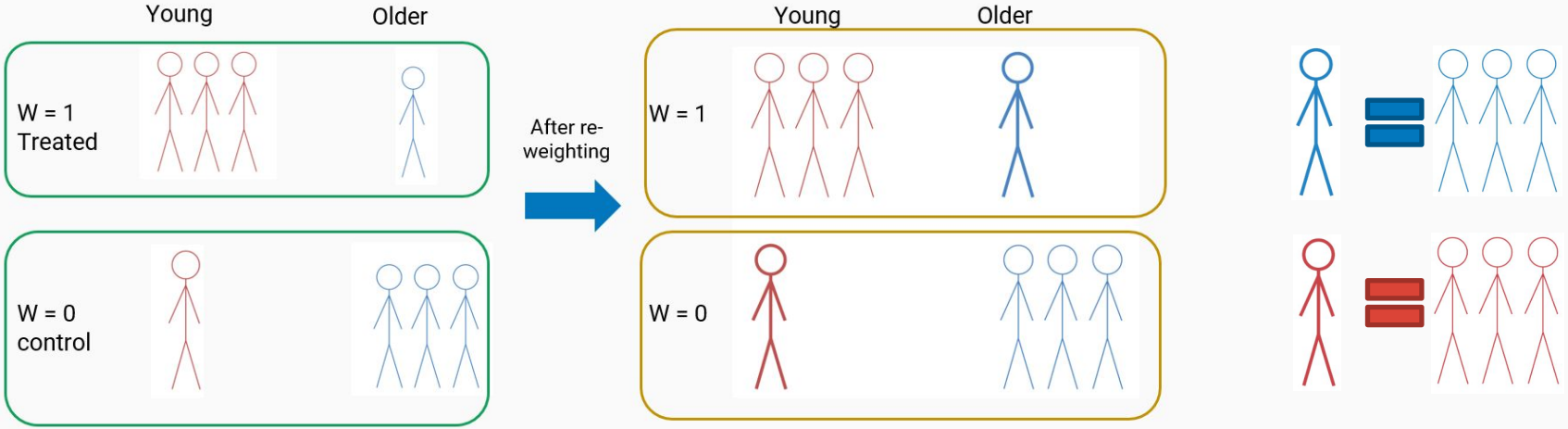
- Categorization of Methods
  - Re-weighting methods
  - Matching methods
  - Tree-based methods
  - Others
    - Stratification methods
    - Multitask Learning Methods
    - Meta-Learning Methods

Re-weighting Methods

- **Challenge of Selection Bias**: due to different distributions of treated and control groups
- Sample re-weighting is a common way to overcome the selection bias problem
- **Idea**: By *assigning appropriate weight to each sample* in the observation dataset, a pseudo-population is created on which the distributions of the treated group and control group are similar

Sample Re-weighting Methods

- Intuitive example: **Age** (Young/older) as the confounder
  - Young people: 75% chance of receiving treatment
  - Older people: only a 25% chance of receiving treatment



http://www.rebeccabarter.com/blog/2017-07-05-ip-weighting/

Balancing Score

- **Balancing Score**: Balancing score $b(\boldsymbol{X})$ is a general weighting score, which is the function of covariates $\boldsymbol{X}$ satisfying: W ⫫ $\boldsymbol{x}$ | $b(\boldsymbol{x})$.
- One representative balancing score: Propensity Score
- **Propensity Score**: Conditional probability of assignment to a treatment given a vector of observed covariates

$$e(x) = Pr(W = 1|X = x)$$

Inverse propensity weighting (IPW)

- The weight assigned for each unit is:

$$r = \frac{W}{e(x)} + \frac{1-W}{1-e(x)}$$

where W is the treatment and $e(x)$ is the propensity score

- After re-weighting, the IPW estimator of ATE is defined as:

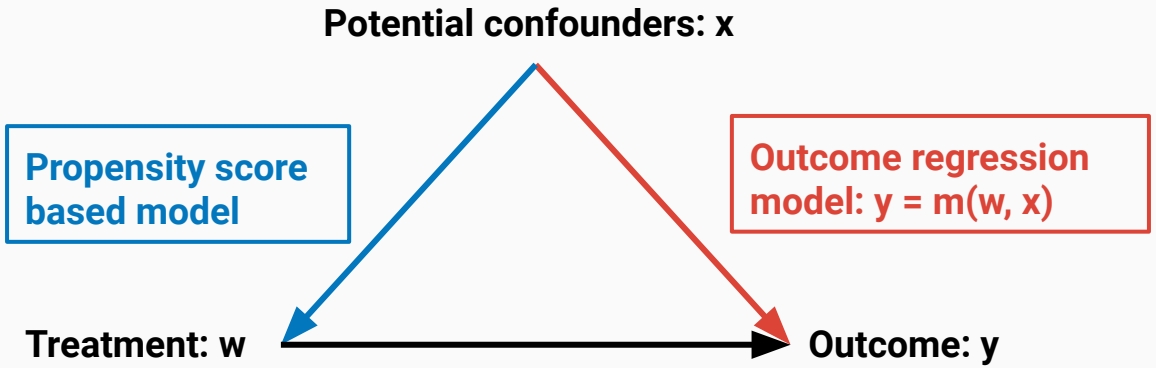
$$\hat{\mathrm{ATE}}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i^F}{\hat{e}(x)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i^F}{1-\hat{e}(x)}$$

  - Theoretical results show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates.
  - However, IPW highly relies on the correctness of propensity scores

J. Robins, A. Rotnitzky, and L. Zhao. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association* 89.427 (1994): 846-866.

Doubly Robust Estimator (DR)

- It combines the propensity score weighting with the outcome regression
- Unbiased when one of the propensity score or outcome regression is correct



J. Robins, A. Rotnitzky, and L. Zhao. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association* 89.427 (1994): 846-866.

Doubly Robust Estimator (DR)

$$\hat{\mathrm{ATE}}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{W_i Y_i^F}{\hat{e}(x_i)} - \frac{W_i - \hat{e}(x_i)}{\hat{e}(x_i)}\hat{m}(1, x_i)\right] - \left[\frac{(1-W_i)Y_i^F}{1-\hat{e}(x_i)} - \frac{W_i - \hat{e}(x_i)}{1-\hat{e}(x_i)}\hat{m}(0, x_i)\right]\right\}$$

IPW **augmented** **IPW** **augmented**

Observed outcome

Outcome from regression model

$$\hat{\mathrm{ATE}}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{W_i Y_i^F}{\hat{e}(x_i)} - \frac{W_i - \hat{e}(x_i)}{\hat{e}(x_i)}\hat{m}(1, x_i)\right] - \left[\frac{(1-W_i)Y_i^F}{1-\hat{e}(x_i)} - \frac{W_i - \hat{e}(x_i)}{1-\hat{e}(x_i)}\hat{m}(0, x_i)\right]\right\}$$

Estimation of treated outcome **Estimation of control outcome**

Covariate balancing propensity score (CBPS)

- As propensity score serves as both the probability of being treated and covariate balancing score, CBPS exploits this dual characteristics
- CBPS estimate the propensity score by solving the following problem:

$$\mathbb{E}\left[\frac{W_i\tilde{x}_i}{e(x_i;\beta)} - \frac{(1-W_i)\tilde{x}_i}{1-e(x_i;\beta)}\right] = 0$$

K. Imai and M. Ratkovic. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263 2014.

Classical Causal Inference Methods

- Categorization of Methods
  - Re-weighting methods
  - **Matching methods**
  - Tree-based methods
  - Others
    - Stratification methods
    - Multitask Learning Methods
    - Meta-Learning Methods

Matching

- Matching methods estimate the counterfactuals and meanwhile reduce the estimation bias brought by the confounders
- Potential outcomes of the i-th unit estimated by matching are:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 1; \end{cases}$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1; \end{cases}$$

Where $J(i)$ is the matched neighbors of unit i in the opposite treatment group

Distance Metrics for Matching

- *Original Data Space*
  - Euclidean distance
  - Mahalanobis distance
- *Transformed Feature Space*
  - Propensity score based transformation
  - Other transformations (Prognosis scores)

Propensity Score Matching

- Propensity scores denote conditional probability of assignment to a particular treatment given a vector of observed covariates.

$$e(x) = Pr(W = 1|X = x)$$

- Based on propensity scores, the distance between two units is

$$D(\mathbf{x}_i, \mathbf{x}_j) = |e_i - e_j|$$

- Alternatively, linear propensity score based distance metric

$$D(\mathbf{x}_i, \mathbf{x}_j) = |\text{logit}(e_i) - \text{logit}(e_j)|$$

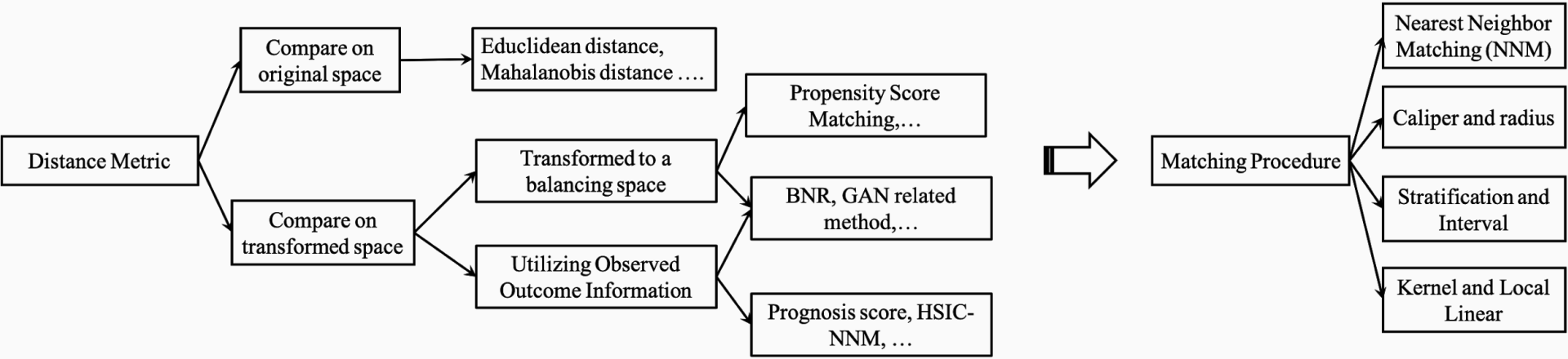
P. Rosenbaum, and D. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.

Matching procedure

- **Choosing Matching Algorithm**
  - Nearest Neighbors
  - Caliper
  - Stratification
  - Kernels
- **Variable Selection**
  - The more, the better?
  - Post-treatment variables
  - Instrumental variables
  - Irrelevant variables

Matching

❑ Summary of Matching Methods



Classical Causal Inference Methods

- Categorization of Methods
  - Re-weighting methods
  - Matching methods
  - **Tree-based methods**
  - Others
    - Stratification methods
    - Multitask Learning Methods
    - Meta-Learning Methods

Tree-based Methods

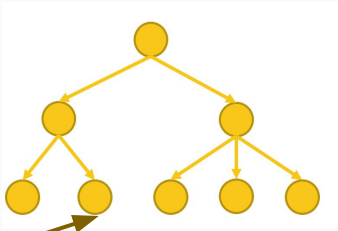
- It is a predictive modeling approach based on decision tree
- Decision tree is a non-parametric supervised learning method
- Classification trees
- Regression trees

Classification And Regression Trees

- **Classification And Regression Trees (CART)**
  - Recursively partition the data space
  - Fit a simple prediction model for each partition
  - Represent every partitioning as a decision tree
- Leaf specific effect:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

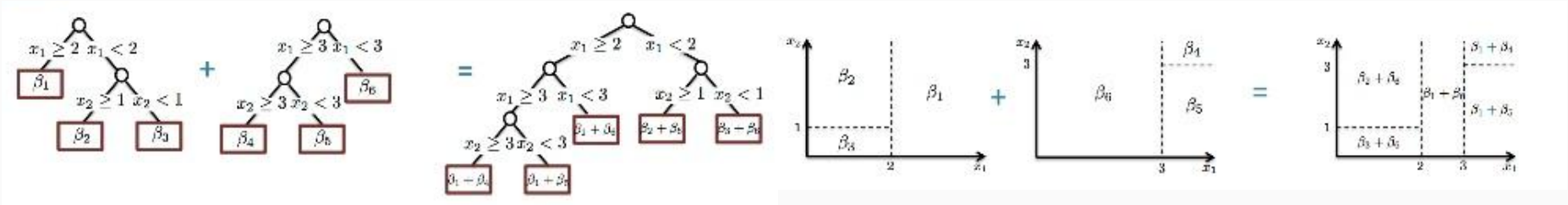
A specific leaf node



$$\tau(x \mid \Pi) \equiv \quad \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

Bayesian Additive Regression Trees

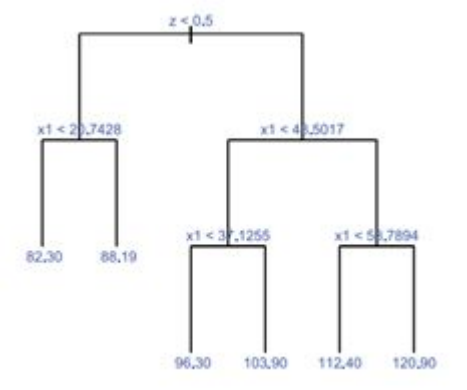
- In CART, a tree is built up until a splitting tolerance is reached. There is only one tree, and it is grown and pruned as needed.
- BART is an ensemble of trees, so it is more comparable to random forests.
- Nonparametric Bayesian regression model



Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20.1 (2011): 217-240.
https://www.slideshare.net/SAMSI_Info/mums-bayesian-fiducial-and-frequentist-conference-multiscale-analysis-of-bart-veronika-rockova-april-29-2019

BART to estimate CATE

- BART is designed to estimate a model for the outcome $Y = f(z, x) + \epsilon$
- Let T denote a binary tree
- Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$, μ_j represents the mean response of the subgroup of units that fall in that node.
- $Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon$



Causal forest

- Single tree is noisy -> using forest
- Forests = nearest neighbor methods + adaptive neighborhood metric
  - k-nearest neighbors: seek the k closest points to x according to some prespecified distance measure
  - tree-based methods: closeness is defined with respect to a decision tree, and the closest points to x are those that fall in the same leaf

S. Wager, and S. Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113.523 (2018): 1228-1242.

Advantages of Causal forest

- The data determine which dimensions are most important to consider in selecting nearest neighbors
- Allow for data driven feature selection while maintaining the benefits of classical methods, that is, asymptotically normal and unbiased point estimates with valid confidence intervals
- Extended to multiple treatments

Classical Causal Inference Methods

- Categorization of Methods
  - Re-weighting methods
  - Matching methods
  - Tree-based methods
  - **Others**
    - Stratification methods
    - Multitask Learning Methods
    - Meta-Learning Methods

Stratification

- **Stratification** adjusts the selection bias by splitting the entire group into subgroups, where within each subgroup, the treated group and the control group are similar under some measurements
- Stratification is also named as ***subclassification*** or ***blocking***
- ATE for stratification is estimated as

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \left[\bar{Y}_t(j) - \bar{Y}_c(j)\right]$$

Multitask Learning Methods

- Treated and control group share some common features
- But they also have their own peculiarities
- Therefore, causal inference can be conceptualized as a multitask learning problem:
  - shared layers for treated group and control group together
  - specific layers for treated group and control group separately

Meta-Learning Methods

- Above methods control the confounders and estimate CATE simultaneously
- Meta-learning algorithms separate them into two steps
- Procedures:

  (1) Estimate the conditional mean outcome $E[Y|X = x]$, and the prediction model learned in this step is the base learner.

  (2) Derive the CATE estimator based on the difference of results obtained from step (1)
- Existing meta-learning methods: T-learner, S-learner, X-learner, U-learner and R-learner

Outline

- ❏ Overview
- ❏ Causal Inference: Background and Challenges
- ❏ Classical Causal Inference Methods
- ❏ **Subspace Learning for Causal Inference**
- ❏ Deep Representation Learning for Causal Inference
- ❏ Applications and Potential Directions
- ❏ Conclusions

Why ML is helpful for Causal Inference?

- **Machine Learning**
  - Various learning tasks, e.g., regression, classification, clustering
  - Various settings: multi-view, multi-task, transfer learning, etc.
  - Feature learning by both shallow and deep models
- **Connections between Causal Inference and Machine Learning**
  - Counterfactual inference is considered as a regression problem
  - Matching in representation space
  - Covariate shift and group balancing

Subspace Learning

- **Goal:** Learning low-dimensional subspaces for dimensionality reduction
- Representative subspace learning methods include: principal component analysis (PCA), locality preserving projections (LPP), canonical correlation analysis (CCA), etc.

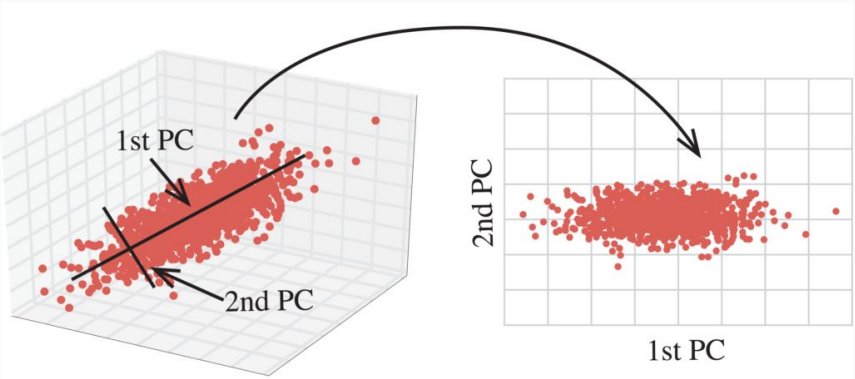


Figure: https://medium.com/@TheDataGyan/dimensionality-reduction-with-pca-and-t-sne-in-r-2715683819

Subspace Learning for Causal Inference

- **Motivation**: Matching in the original data space is simple and flexible, but it could be misled by variables that do not affect the outcome. To address this issue, matching could be performed in subspaces instead.
- **Methods**
  - Random Subspaces
  - Informative Subspace
  - Balanced and Nonlinear Subspace

Nearest Neighbor Matching (NNM)

- For a treated unit *i*, nearest neighbor matching (NNM) finds its nearest neighbor in control group in terms of covariates.
- NNM usually uses metrics such as Euclidean distance and Mahalanobis distance.
- NNM has difficulty in dealing with a large number of covariates. Also, bias of NNM increases with with the dimensionality of data at a rate $O(N^{-1/d})$ [Abadie and Imbens, 2006]

A. Abadie and G. Imbens. "Large sample properties of matching estimators for average treatment effects." Econometrica 74.1 (2006): 235-267.

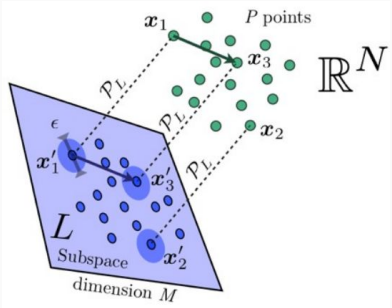
NNM with Random Subspaces

- **Motivation**
  - Dimension reduction, to soften the dependence of bias to dimension
  - Linear projection, to deal with ‘big data’
- *Johnson-Lindenstrauss (JL) Lemma*

Project data to a randomly generated subspace while preserving original distances between points *[Johnson and Lindenstrauss, 1984]*

Johnson-Lindenstrauss (JL) lemma. For any $0 < \epsilon < 1/2$ and $x_1, \cdots, x_N \in \mathbb{R}^d$, there exists a mapping $f : \mathbb{R}^d \to \mathbb{R}^k$, with $k = O(\epsilon^{-2} \log N)$, such that

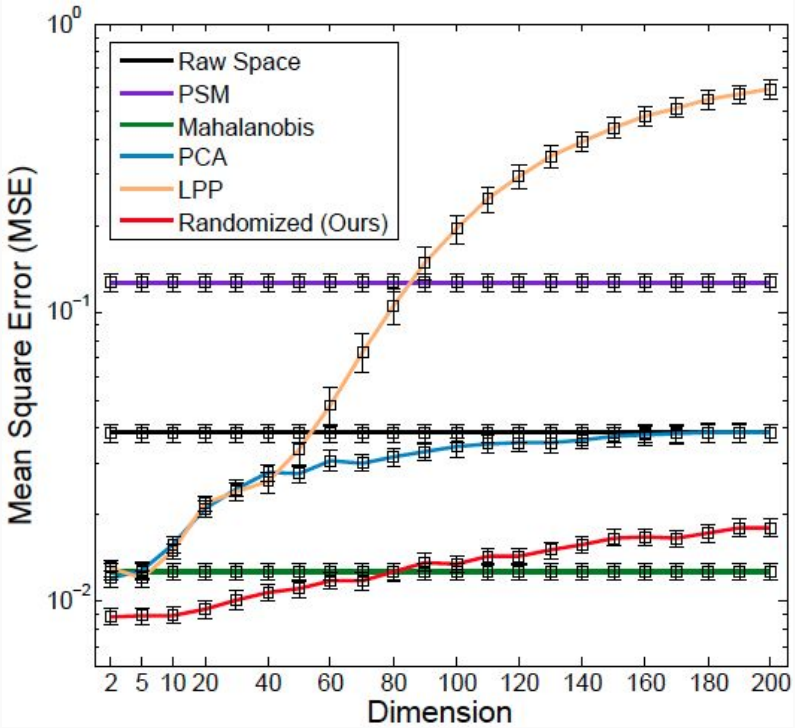
$$\forall i, j \ (1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2.$$



S. Li, N. Vlassis, J. Kawale, Y. Fu. “Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns”. *IJCAI 2016.*

NNM with Random Subspaces

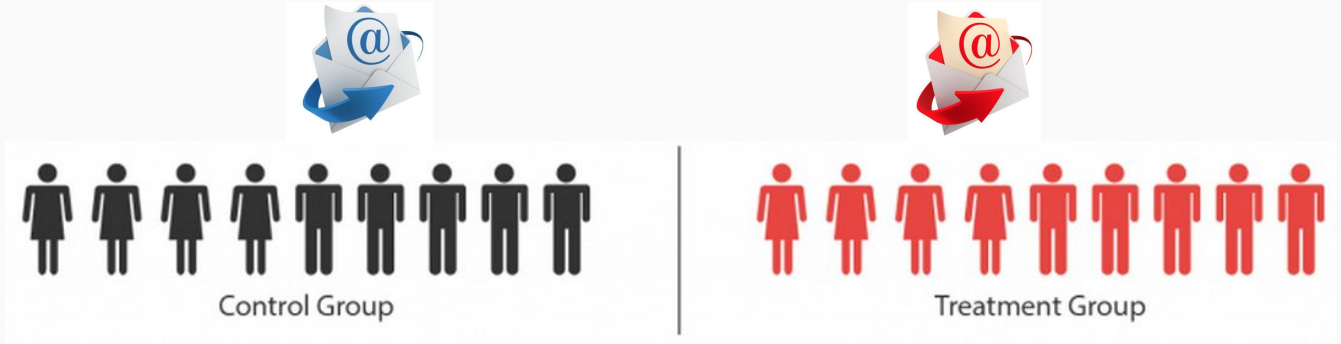
- **Experiments on Synthetic Dataset**
  - 1,000 samples, 200 features
  - True outcomes are determined by a set of basis functions
  - Simulated outcomes are drawn from a normal distribution
  - **Ground Truth** of ATT is 1
  - Metric: average of mean square error (MSE) over 1,000 simulations



S. Li, N. Vlassis, J. Kawale, Y. Fu. "Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns". *IJCAI 2016.*

NNM with Random Subspaces

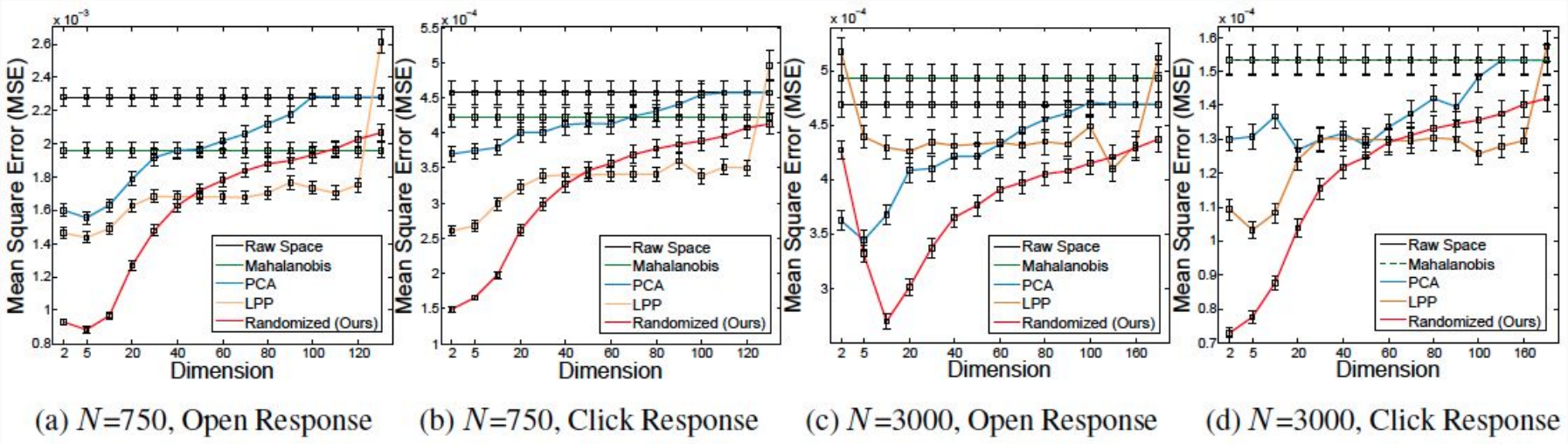
- **Experiments on Marketing Dataset**
  - Email Campaigns: sending two types of promotional emails to two groups of customers separately
  - 1.2 million units in control group, and 0.8 million units in treated group. 209 dimensional features. Outcome: open or click emails



S. Li, N. Vlassis, J. Kawale, Y. Fu. “Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns”. *IJCAI 2016.*

NNM with Random Subspaces

- **Experiments on Marketing Dataset: Semi-synthetic Settings**
  - Generate a pseudo-treated population from the control group
  - True causal effect is 0



JL bound: log(3000)=8

Informative Subspace Learning

- Hilbert-Schmidt Independence Criterion (HSIC) based NNM
- HSIC-NNM learns two linear projections for control outcome estimation task and treated outcome estimation task separately
- It maximizes nonlinear dependency between the projected subspace and the outcome by

$$M_w = \max_{M_w} \mathrm{HSIC}(\mathbf{X}_w M_w, Y_w^F) - \mathcal{R}(M_w)$$

where $X_w M_w$ is the transformed subspace, Y_w^F is the observed control/ treated outcome, and R denotes the regularization term

Y. Chang, J. Dy. “Informative Subspace Learning for Counterfactual Inference”. *AAAI 2017*.

Informative Subspace Learning

- Experiments on IHDP Dataset
- Source: Collected by Infant Health and Development Program Treatment group
- Samples: 24 Covariates; 747 samples (608 control units and 139 treated units)
- Outcomes: Simulated using covariates and treatment information

	PEHE	$\mathcal{E}_{ATE}$	$\mathcal{E}_{ATT}$
MDM	7.2 ± 3.2 **(5)**	4.1 ± 0.7 **(6)**	3.2 ± 1.4 **(3)**
PSM	3.6 ± 1.4 **(2)**	0.8 ± 0.6 **(3)**	3.3 ± 1.6 **(3)**
DR-RLP	7.3 ± 3.1 **(5)**	4.1 ± 0.7 **(6)**	3.2 ± 1.4 **(3)**
LASSO	6.4 ± 3.7 **(4)**	1.6 ± 0.6 **(4)**	3.5 ± 2.1 **(3)**
BART	5.2 ± 2.9 **(3)**	1.9 ± 1.3 **(5)**	-0.1 ± 0.3 **(1)**
CausalForest	5.1 ± 2.6 **(2)**	0.4 ± 0.4 **(2)**	0.2 ± 1.0 **(2)**
NNM-HSIC	1.7 ± 0.5 **(1)**	0.1 ± 0.1 **(1)**	0.3 ± 0.3 **(2)**

Y. Chang, J. Dy. “Informative Subspace Learning for Counterfactual Inference”. *AAAI 2017*.

Nonlinear and Balanced Subspace Learning

- **Challenges**
  - Bias increases with the dimension of data
  - Complex & unbalanced distributions of high-dimensional covariates
- **Our Solution**
  - Convert counterfactual prediction to a multi-class classification problem with pseudo labels
  - Ordinal scatter discrepancy criterion to extract nonlinear representations
  - Maximum mean discrepancy criterion to learn balanced representations

S. Li, and Y. Fu. “Matching on balanced nonlinear representations for treatment effects estimation”. *NIPS 2017*.

Nonlinear and Balanced Subspace Learning

- **Objective Function**

$$\begin{aligned} \arg\max_{P} \quad & F(P, \Phi(X), Y_c) - \beta \mathrm{Dist}(\Psi(X_C), \Psi(X_T)) \\ & = \mathrm{tr}(P^\top (K_I - \alpha K_W) P) - \beta \mathrm{tr}(P^\top K L K P) \\ s.t. \quad & P^\top P = \mathrm{I}, \end{aligned}$$

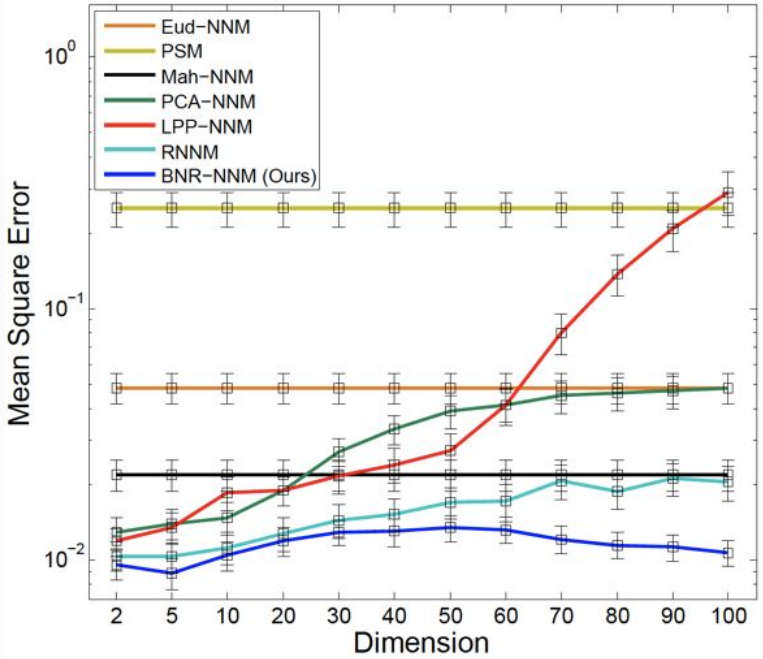
P is the learned nonlinear projection

- It can be solved with a closed-form solution

S. Li, and Y. Fu. “Matching on balanced nonlinear representations for treatment effects estimation”. *NIPS 2017*.

Nonlinear and Balanced Subspace Learning

- **Experiments on Synthetic Dataset**
  - 1,000 samples, 200 features
  - True outcomes are determined by a set of basis functions
  - Simulated outcomes are drawn from a normal distribution
  - Ground Truth of ATT is 1
  - Metric: average of mean square error (MSE) over 1,000 simulations



S. Li, and Y. Fu. “Matching on balanced nonlinear representations for treatment effects estimation”. *NIPS 2017*.

Nonlinear and Balanced Subspace Learning

- **Experiments on IHDP Dataset**
- Source: Collected by the Infant Health and Development Program Treatment group: all children with non-white mothers.
- Samples: 24 Covariates; 747 samples (608 control units and 139 treated units)
- Outcomes: Simulated using covariates and treatment information
- Metric: Error in ATT as evaluation metric

Table 1: Results on IHDP dataset.

Method	ε_{ATT}
Eu-NNM	0.18±0.06
Mah-NNM	0.31±0.12
PSM	0.26±0.08
PCA-NNM	0.19±0.11
LPP-NNM	0.25±0.13
RNNM	**0.16±0.07**
BNR-NNM	**0.16±0.06**

S. Li, and Y. Fu. "Matching on balanced nonlinear representations for treatment effects estimation". *NIPS 2017*.

Nonlinear and Balanced Subspace Learning

- **Experiments on LaLonda Dataset**
- Source: Collected by a randomized study of a job training program
- Treatment group: 297 units who participated in the training program
- Control group: 2,915 units from surveys and other studies
- Outcome: earnings in 1978 Ground truth of ATT is $886 with a standard error of $448

Table 2: Results on LaLonde dataset. BIAS (%) is the bias in percentage of the true effect.

Method	ATT	SD	BIAS (%)
Ground Truth	886	488	N/A
Eu-NNM	-565.9	592.8	164%
Mah-NNM	-67.9	526.1	108%
PSM	-947.6	567.9	201%
PCA-NNM	-499.8	592.5	156%
LPP-NNM	-457.1	581.2	152%
RNNM	-557.6	584.9	163%
CBPS	423.3	1295.2	52%
DNN	742.0	N/A	16%
BNR-NNM	783.6	546.3	**12%**

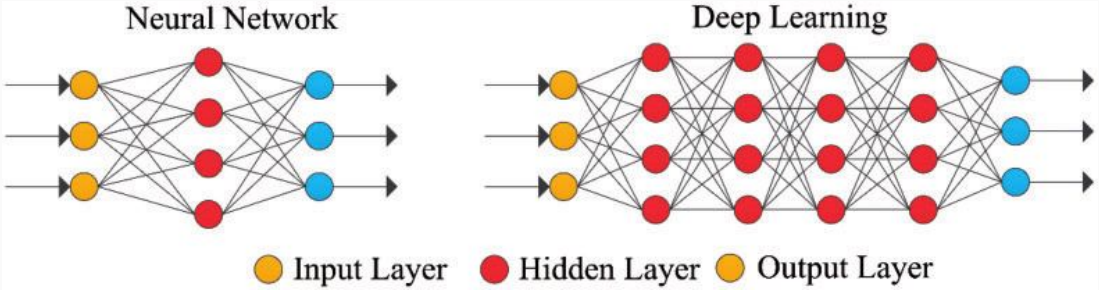
Discussions

- **Subspace Learning for Causal Inference**
  - (+) Most methods are highly efficient owing to their closed-form solutions
  - (-) Subspace learning methods usually have strong assumptions on underlying data distributions
  - (-) They are usually combined with Matching estimators, but are not capable of estimating counterfactuals directly

Outline

- ❏ Overview
- ❏ Causal Inference: Background and Challenges
- ❏ Classical Causal Inference Methods
- ❏ Subspace Learning for Causal Inference
- ❏ **Deep Representation Learning for Causal Inference**
- ❏ Applications and Potential Directions
- ❏ Conclusions

Deep Representation Learning



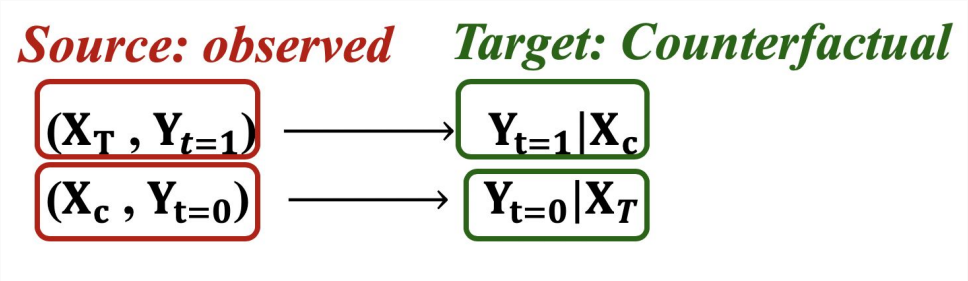
- Deep learning architecture is composed of an input layer, hidden layers, and an output layer
  - The output of each intermediate layer can be viewed as a representation of the original input data
  - Ability to deliver high-quality features and enhanced learning performance
  - Examples: Feed forward NN, CNN, Auto Encoder, VAE, GAN, etc.

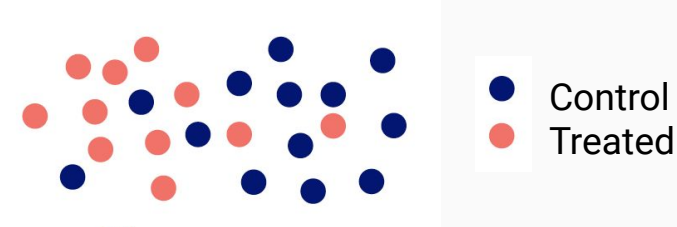
Deep Representation Learning for Causal Inference

- ❏ Balanced representation learning
- ❏ Local similarity preserving based methods
- ❏ Deep Generative model based methods

Balanced Representation Learning

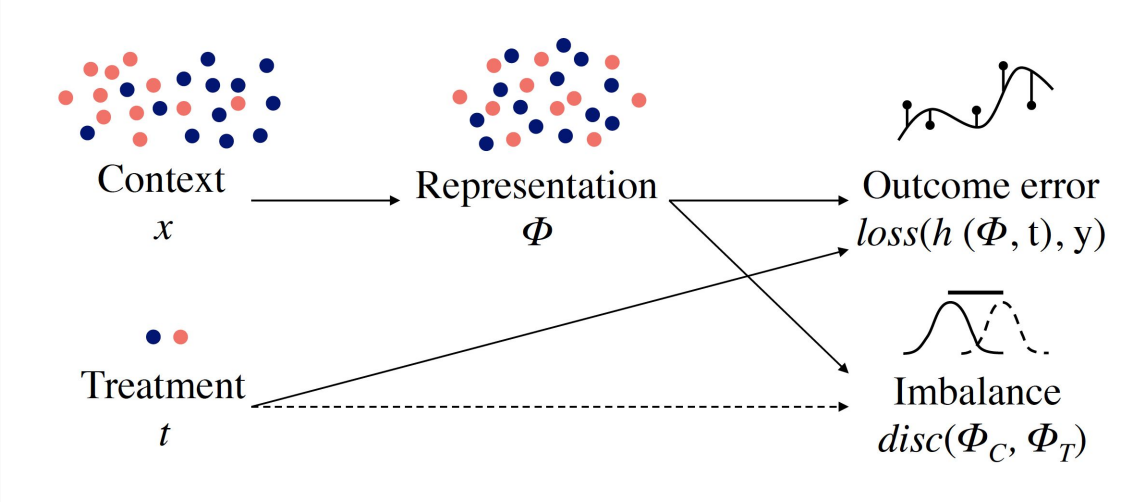
- Motivation
  - Counterfactual inference <-> Domain adaptation





Balanced Representation Learning

- Balancing the two groups in the latent space



Representation Learning For Causal Inference

- Theoretical background

The expected error of estimating ITE:

representation learner

outcome predictor

$$\mathbb{E}_x\left[error\left(\widehat{ITE}^{\Phi,h}(x)\right)\right] \leq 2\mathbb{E}_{x,t}\left[error\left(\hat{Y}_t^{\Phi,h}(x)\right)\right] + dist(p_\Phi^{treated}, p_\Phi^{control})$$

- $\widehat{ITE}^{\Phi,h}(x) = \hat{Y}_{t=1}^{\Phi,h}(x) - \hat{Y}_{t=0}^{\Phi,h}(x)$
- $ITE(x) = \frac{1}{|\{i: x_i = x\}|}\sum_{\{i: x_i = x\}} ITE_i$

Johansson, Fredrik, et al. "Learning representations for counterfactual inference." International conference on machine learning. 2016.

Representation Learning For Causal Inference

- Theoretical background

The expected error of estimating ITE:

Expected supervised learning generalization error

representation learner

outcome predictor

$$\mathbb{E}_x\left[error\left(\widehat{ITE}^{\Phi,h}(x)\right)\right] \leq 2\mathbb{E}_{x,t}\left[error\left(\hat{Y}_t^{\Phi,h}(x)\right)\right] + dist(p_\Phi^{treated}, p_\Phi^{control})$$

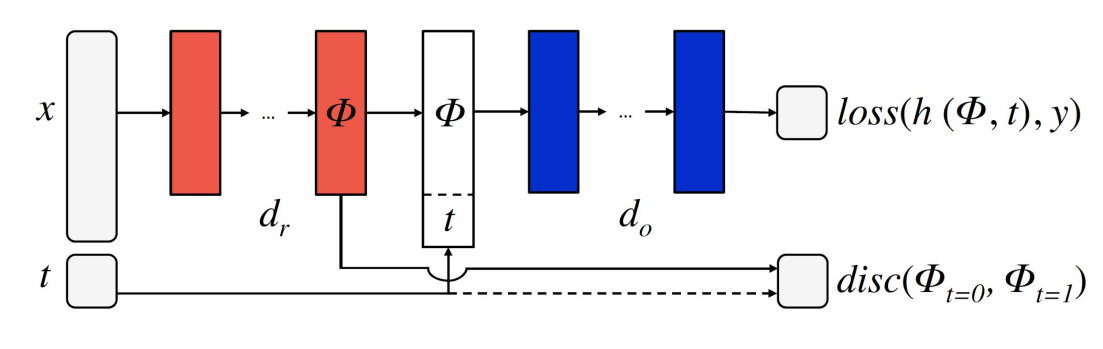
Distance between the learned representations

- $\widehat{ITE}^{\Phi,h}(x) = \hat{Y}_{t=1}^{\Phi,h}(x) - \hat{Y}_{t=0}^{\Phi,h}(x)$
- $ITE(x) = \frac{1}{|\{i: x_i = x\}|}\sum_{\{i: x_i = x\}} ITE_i$

Johansson, Fredrik, et al. Learning representations for counterfactual inference. International conference on machine learning. 2016.

Balanced Representation Learning

- **BNN/TARNet**



- Objective Function

Factual loss

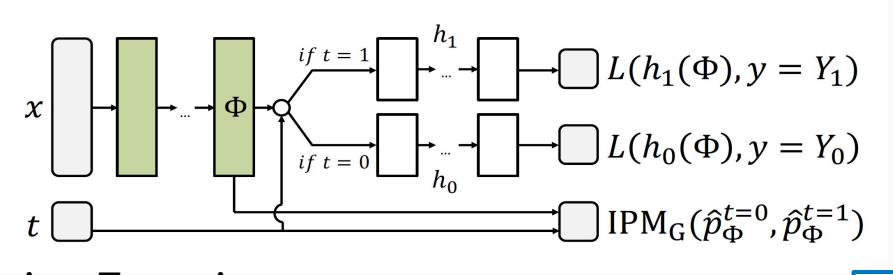
Discrepancy

$$B_{\mathcal{H},\alpha,\gamma}(\Phi, h) = \frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F| + \alpha \operatorname{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF})$$

F. Johansson, U. Shalit, and D. Sontag. "Learning representations for counterfactual inference." *International conference on machine learning*. 2016.

Balanced Representation Learning

- **Counterfactual Regression**



- Objective Function

Factual loss

$$\min_{\substack{h,\Phi \\ \|\Phi\|=1}} \frac{1}{n}\sum_{i=1}^{n} w_i \cdot L\left(h(\Phi(x_i), t_i), y_i\right) + \lambda \cdot \Re(h)$$

Discrepancy

$$+ \alpha \cdot \mathrm{IPM}_G\left(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1}\right)$$

U. Shalit, F. Johansson, and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Balanced Representation Learning

BNN/TARNET, CFR Experiments

- Experiment on IHDP and Jobs dataset (Lalonde dataset)
- Evaluation metric:
  - IHDP dataset:
    - For ITE: Precision in Estimation of Heterogeneous Effect (PEHE)

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_1(x_i) - \hat{y}_0(x_i) - (Y_1(x_i) - Y_0(x_i))\right)^2}$$

    - For ATE: absolute error of ATE
  - Jobs dataset:
    - For ITE: Policy risk. The average loss in value when treating according to the policy implied by an ITE estimator.
    - For ATT: absolute error of ATT

Balanced Representation Learning

- Within-sample: Estimate the ITE of the units whose outcome of one treatment is observed
  - Training + Validation sets

Within-sample

	IHDP		JOBS	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}
OLS/LR-1	$5.8 \pm .3$	$.73 \pm .04$	$.22 \pm .0$	$.01 \pm .00$
OLS/LR-2	$2.4 \pm .1$	$.14 \pm .01$	$.21 \pm .0$	$.01 \pm .01$
BLR	$5.8 \pm .3$	$.72 \pm .04$	$.22 \pm .0$	$.01 \pm .01$
k-NN	$2.1 \pm .1$	$.14 \pm .01$	$.02 \pm .0$	$.21 \pm .01$
TMLE	$5.0 \pm .2$	$.30 \pm .01$	$.22 \pm .0$	$.02 \pm .01$
BART	$2.1 \pm .1$	$.23 \pm .01$	$.23 \pm .0$	$.02 \pm .00$
RAND.FOR.	$4.2 \pm .2$	$.73 \pm .05$	$.23 \pm .0$	$.03 \pm .01$
CAUS.FOR.	$3.8 \pm .2$	$.18 \pm .01$	$.19 \pm .0$	$.03 \pm .01$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$.20 \pm .0$	$.04 \pm .01$
TARNET	$.88 \pm .0$	$.26 \pm .01$	$.17 \pm .0$	$.05 \pm .02$
CFR MMD	$.73 \pm .0$	$.30 \pm .01$	$.18 \pm .0$	$.04 \pm .01$
CFR WASS	$.71 \pm .0$	$.25 \pm .01$	$.17 \pm .0$	$.04 \pm .01$

Balanced Representation Learning

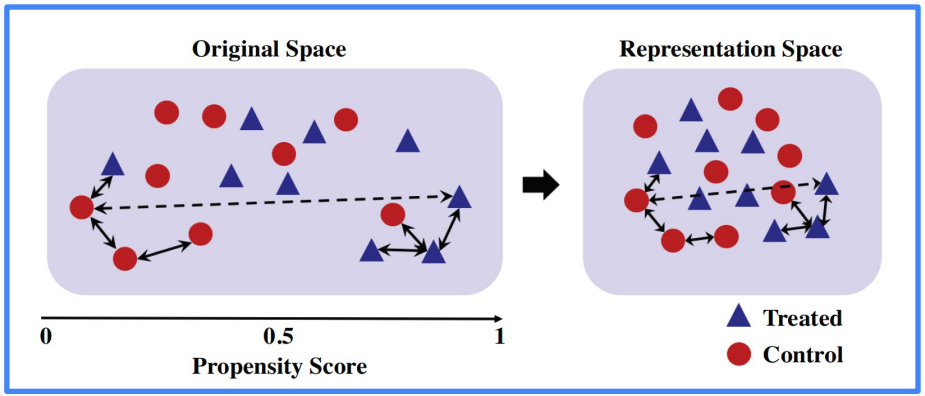
- Out-of-sample: Estimate the ITE of the units with no observed outcome
  - Test set
- Example: a new patient arrives and the goal is to select the best possible treatment

Out-of-sample

	IHDP		JOBS	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{POL}	ϵ_{ATT}
OLS/LR-1	$5.8 \pm .3$	$.94 \pm .06$	$.23 \pm .0$	$.08 \pm .04$
OLS/LR-2	$2.5 \pm .1$	$.31 \pm .02$	$.24 \pm .0$	$.08 \pm .03$
BLR	$5.8 \pm .3$	$.93 \pm .05$	$.25 \pm .0$	$.08 \pm .03$
k-NN	$4.1 \pm .2$	$.79 \pm .05$	$.26 \pm .0$	$.13 \pm .05$
BART	$2.3 \pm .1$	$.34 \pm .02$	$.25 \pm .0$	$.08 \pm .03$
RAND.FOR.	$6.6 \pm .3$	$.96 \pm .06$	$.28 \pm .0$	$.09 \pm .04$
CAUS.FOR.	$3.8 \pm .2$	$.40 \pm .03$	$.20 \pm .0$	$.07 \pm .03$
BNN	$2.1 \pm .1$	$.42 \pm .03$	$.24 \pm .0$	$.09 \pm .04$
TARNET	$.95 \pm .0$	$.28 \pm .01$	$.21 \pm .0$	$.11 \pm .04$
CFR MMD	$.78 \pm .0$	$.31 \pm .01$	$.21 \pm .0$	$.08 \pm .03$
CFR WASS	$.76 \pm .0$	$.27 \pm .01$	$.21 \pm .0$	$.09 \pm .03$

Local Similarity Preserving based Method

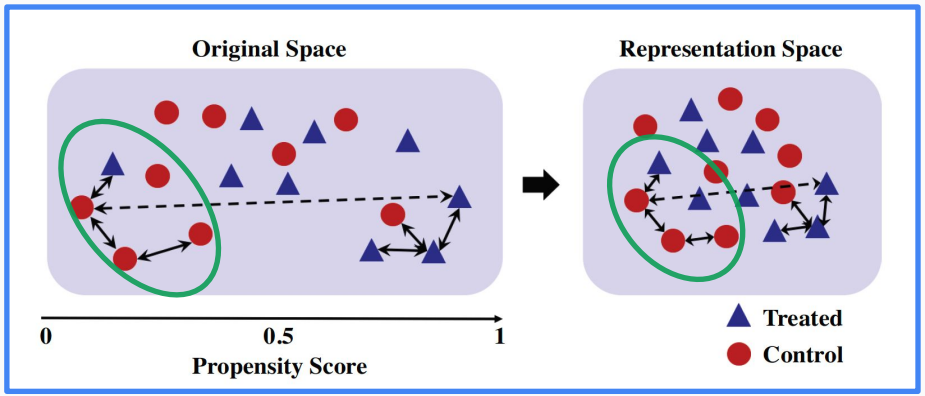
- **Motivation**: The latent space should encode:
  - The distribution in latent space is **balanced.**
  - The **similarity order** information in X
  - For different data points, the strength of similarity should be different.



Liuyi Yao, et al. Representation learning for treatment effect estimation from observational data. NeurIPS 2018.

Local Similarity Preserving based Method

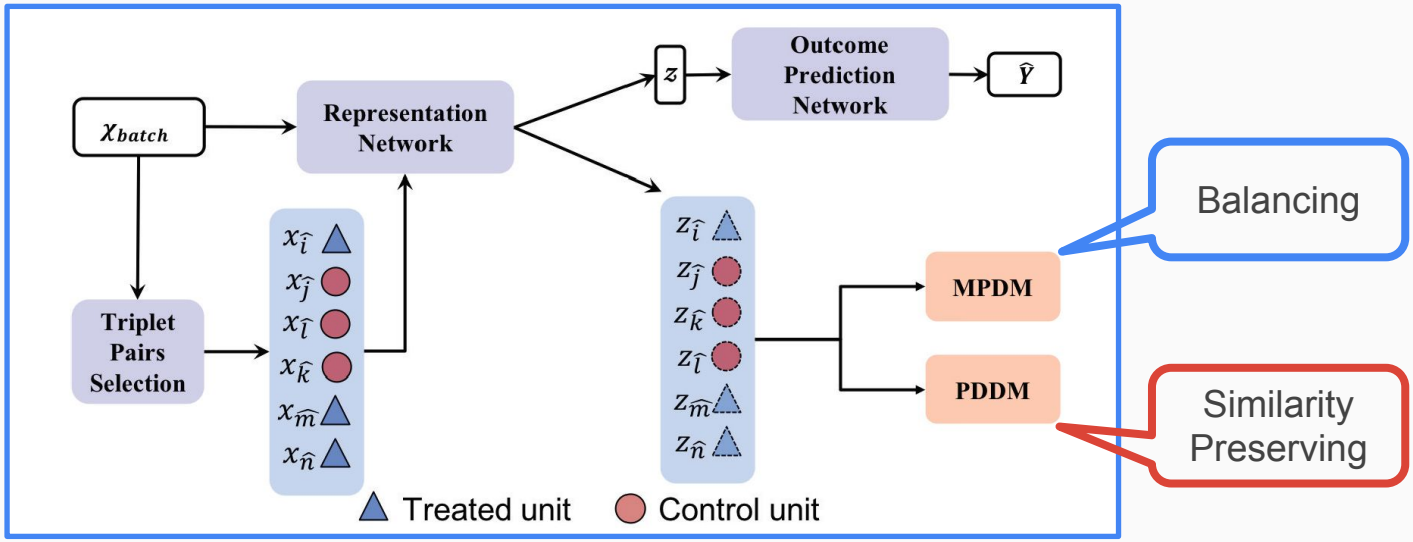
- **Motivation**: The latent space should encode:
  - The distribution in latent space is **balanced.**
  - The **similarity order** information in X
  - For different data points, the strength of similarity should be different.



Liuyi Yao, et al. Representation learning for treatment effect estimation from observational data. NeurIPS 2018.

Local Similarity Preserving based Method: SITE

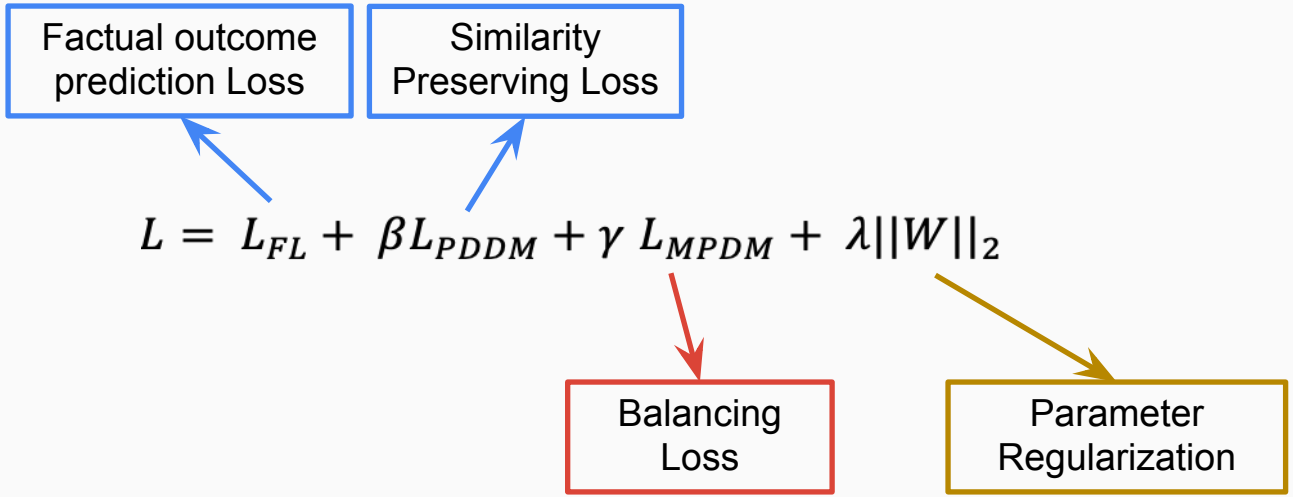
- Idea: Using triplet loss to preserve the local similarity



Liuyi Yao, et al. “Representation learning for treatment effect estimation from observational data”. NeurIPS 2018.

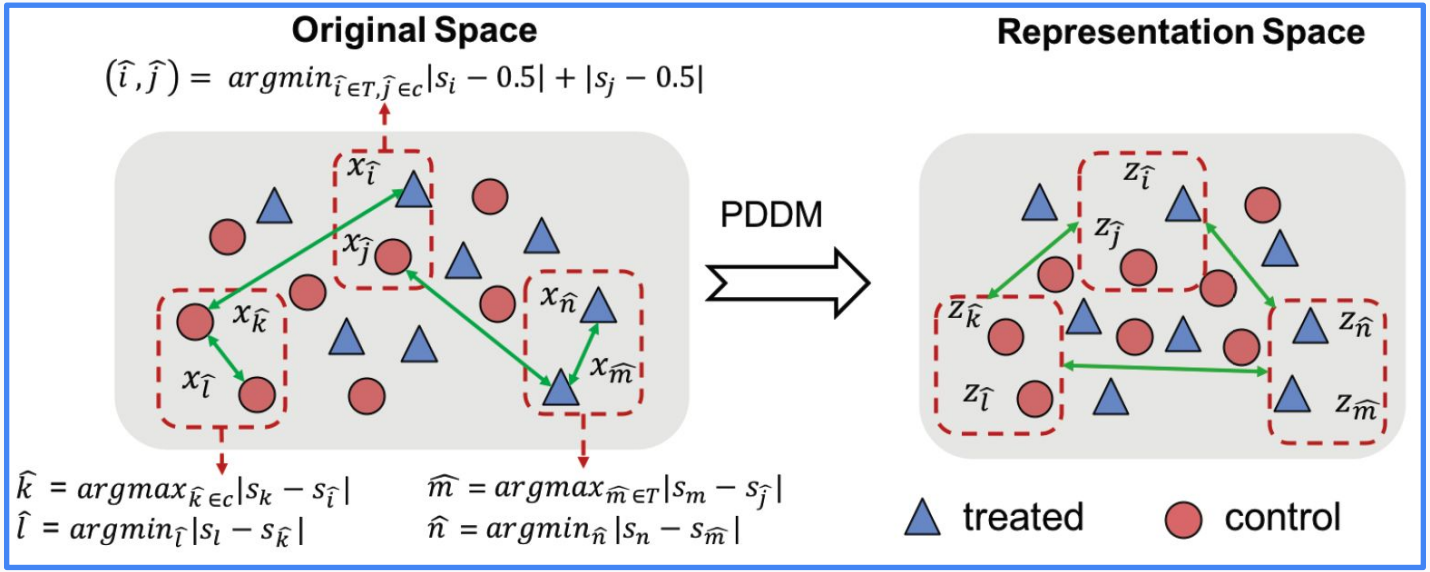
Local Similarity Preserving based Method: SITE

- Objective Function:



SITE: Triplet Pair Selection & PDDM

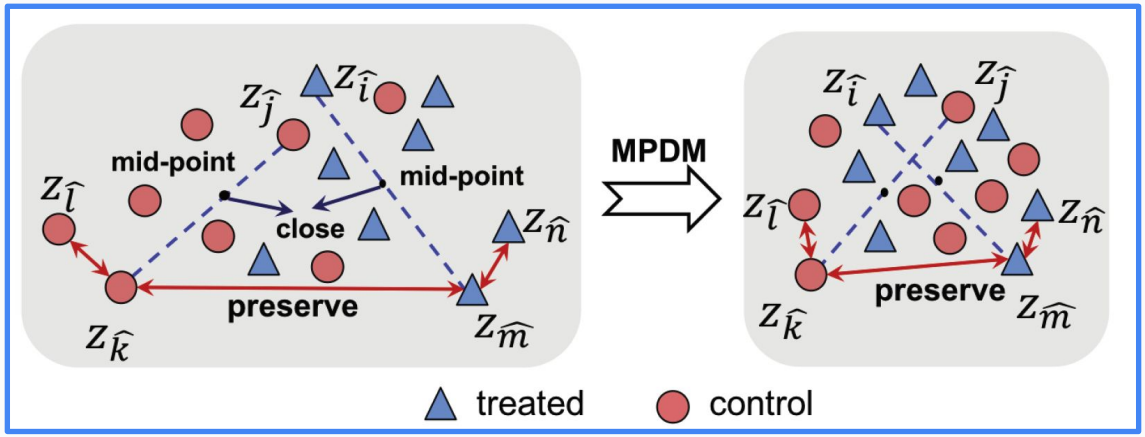
- Triplet pair selection



- s_i is the propensity score, which is the probability that a unit is in the treated group

SITE: MPDM

- Middle Point Distance Minimization (MPDM):
  - MPDM makes two mid-points close to each other
  - MPDM balances the distribution in the latent space



Balanced

SITE: Experiments

The lower, the better.

Table 1: Performance comparison on IHDP and Jobs Dataset.

	IHDP ($\sqrt{\epsilon_{\text{PEHE}}}$)		Jobs ($\mathcal{R}_{\text{pol}}$)	
Method	Within-sample	Out-of-sample	Within-sample	Out-of-sample
OLS/LR$_1$	10.761 ± 4.350	7.345 ± 2.914	0.310 ± 0.017	0.279 ± 0.067
OLS/LR$_2$	10.280 ± 3.794	5.245 ± 0.986	0.228 ± 0.012	0.733 ± 0.103
HSIC-NNM [5]	2.439 ± 0.445	2.401 ± 0.367	0.291 ± 0.019	0.311 ± 0.069
PSM [27]	7.188 ± 2.679	7.290 ± 3.389	0.292 ± 0.019	0.307 ± 0.053
k-NN [8]	4.432 ± 2.345	4.303 ± 2.077	0.230 ± 0.016	0.262 ± 0.038
Causal Forest [33]	4.732 ± 2.974	4.095 ± 2.528	0.232 ± 0.018	0.224 ± 0.034
BNN [18]	3.827 ± 2.044	4.874 ± 2.850	0.232 ± 0.008	0.240 ± 0.012
TARNet [30]	0.729 ± 0.088	1.342 ± 0.597	0.228 ± 0.004	0.234 ± 0.012
CFR-MMD [30]	0.663 ± 0.068	1.202 ± 0.550	$\mathbf{0.213 \pm 0.006}$	0.231 ± 0.009
CFR-WASS [30]	0.649 ± 0.089	1.152 ± 0.527	0.225 ± 0.004	0.225 ± 0.010
SITE (Ours)	$\mathbf{0.604 \pm 0.093}$	$\mathbf{0.656 \pm 0.108}$	0.224 ± 0.004	$\mathbf{0.219 \pm 0.009}$

Experiment on IHDP and Jobs dataset

SITE: Experiments

Experiment on Twins datasets:

- Dataset:
  - Source: the data of twins birth in the USA between 1989-1991
  - Samples: Total 11,400 twins pairs with 30 features relating to the parents, the pregnancy and the birth
  - Treated group: heavier twin;
  - Control group: light twin
  - Outcome: 1-year mortality

- Evaluation Metric:
  - AUC on outcome estimation

Method	**Twins** (AUC)	
	Within-sample	Out-of-sample
OLS/LR_1	0.660 ± 0.005	0.500 ± 0.028
OLS/LR_2	0.660 ± 0.004	0.500 ± 0.016
HSIC-NNM [5]	0.762 ± 0.011	0.501 ± 0.017
PSM [27]	0.500 ± 0.003	0.506 ± 0.011
k-NN [8]	0.609 ± 0.010	0.492 ± 0.012
BNN [18]	0.690 ± 0.008	0.676 ± 0.008
TARNet [30]	0.849 ± 0.002	0.840 ± 0.006
CFR-MMD [30]	0.852 ± 0.001	0.840 ± 0.006
CFR-WASS [30]	0.850 ± 0.002	0.842 ± 0.005
SITE (Ours)	$\mathbf{0.862 \pm 0.002}$	$\mathbf{0.853 \pm 0.006}$

Local Similarity Preserving based Method: ACE

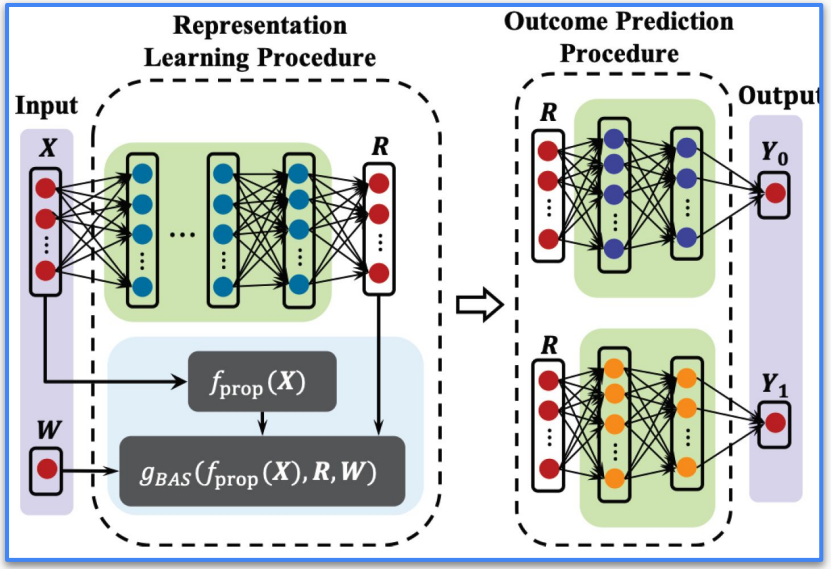
- An improvement of SITE:
  - SITE only considers the similarity of extreme cases
  - SITE requires that the underlying data are spherically distributed when calculating the group distance
- ACE
  - Goal:
    - preserve the fine-grained similarity information
    - Obtain balanced distribution in the latent representation

Liuyi Yao, et al. "ACE: Adaptively Similarity-Preserved Representation Learning for Individual Treatment Effect Estimation." *ICDM* 2019.

ACE: Methodology

Approach: Imposing Balancing & Adaptive-similarity preserving regularization (BAS) on the representation ***R***:

- ❏ Balancing: control/treated group distance minimization in the representation space
- ❏ Adaptive Similarity Preserving
  - ❏ Explores all the pairwise similarity
  - ❏ Adaptively preserves the important similarity information



ACE: Methodology

Balancing & Adaptive-similarity preserving regularization (BAS) regularization

$$g_{BAS}(f_{prop}(\mathbf{X}), \mathbf{R}, \mathbf{W}) = \alpha\mathcal{L}_d + \gamma\mathcal{L}_s,$$

- $\mathcal{L}_d$: Group Distance between treated/control group in the representation space

$$\mathcal{L}_d = IPM(\mathbf{R}_{I_c}, \mathbf{R}_{I_t})$$

- $\mathcal{L}_s$: Adaptive Similarity Preserving loss → Measure the similarity loss after representation learning by KL-divergence:

$$\mathcal{L}_s(\mathcal{P}, \mathcal{Q}) = -\sum_{i,j} \mathcal{P}_{i,j} \log \frac{\mathcal{Q}_{i,j}}{\mathcal{P}_{i,j}}$$

Similarity between i and j in the original space

$$\mathcal{P}_{i,j} = \frac{\exp(S(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{k \neq l} \exp(S(\mathbf{x}_k, \mathbf{x}_l))}$$

- S(. , .) *adaptive* similarity measure score based on propensity score

Similarity preserving strength

$$S(\mathbf{x}_i, \mathbf{x}_j) = 0.75\left|\frac{f_{prop}(\mathbf{x}_i) + f_{prop}(\mathbf{x}_j)}{2} - 0.5\right| - 0.5\left|f_{prop}(\mathbf{x}_i) - f_{prop}(\mathbf{x}_j)\right| + 0.5$$

relative distance within the pair

Similarity between i and j in the latent space

$$\mathcal{Q}_{i,j} = \frac{\exp\left(-\|\mathbf{R}_i - \mathbf{R}_j\|^2\right)}{\sum_{k \neq l} \exp(-\|\mathbf{R}_k - \mathbf{R}_l\|^2)}$$

ACE: Experiment

Experiment on IHDP, Jobs and Twins datasets

The lower, the better.

	IHDP ($\mathcal{E}_{\text{PEHE}}$)		Jobs ($\mathcal{R}_{\text{pol}}$)		Twins ($\hat{\mathcal{E}}_{\text{PEHE}}$)	
Method	Within-Sample	Out-of-Sample	Within-Sample	Out-of-Sample	Within-Sample	Out-of-Sample
OLS/LR$_1$	10.761 ± 4.350	7.345 ± 2.914	0.297 ± 0.010	0.307 ± 0.084	0.308 ± 0.001	0.309 ± 0.012
OLS/LR$_2$	10.280 ± 3.794	5.245 ± 0.986	0.295 ± 0.006	0.297 ± 0.084	0.313 ± 0.002	0.312 ± 0.020
HSIC-NNM	2.439 ± 0.445	2.401 ± 0.367	0.291 ± 0.019	0.311 ± 0.069	0.602 ± 0.010	0.606 ± 0.028
PSM	7.188 ± 2.679	7.290 ± 3.389	0.292 ± 0.019	0.307 ± 0.053	0.607 ± 0.015	0.597 ± 0.021
k-NN	4.432 ± 2.345	4.303 ± 2.077	0.230 ± 0.016	0.262 ± 0.038	0.534 ± 0.008	0.573 ± 0.022
C. Forest	4.732 ± 2.974	4.095 ± 2.528	0.232 ± 0.018	0.224 ± 0.034	$\mathbf{0.306 \pm 0.000}$	0.305 ± 0.003
BNN	3.827 ± 2.044	4.874 ± 2.850	0.232 ± 0.008	0.240 ± 0.012	0.307 ± 0.001	0.309 ± 0.004
TARNet	0.729 ± 0.088	1.342 ± 0.597	0.228 ± 0.004	0.234 ± 0.012	0.314 ± 0.001	0.313 ± 0.002
CFR-MMD	0.663 ± 0.068	1.202 ± 0.550	$\mathbf{0.213 \pm 0.006}$	0.231 ± 0.009	0.312 ± 0.001	0.316 ± 0.003
CFR-WASS	0.649 ± 0.089	1.152 ± 0.527	0.225 ± 0.004	0.225 ± 0.010	0.308 ± 0.001	0.309 ± 0.003
SITE	0.604 ± 0.093	0.656 ± 0.108	0.224 ± 0.004	0.219 ± 0.009	0.309 ± 0.002	0.311 ± 0.004
ACE (Ours)	$\mathbf{0.489 \pm 0.046}$	$\mathbf{0.541 \pm 0.061}$	0.216 ± 0.005	$\mathbf{0.215 \pm 0.009}$	$\mathbf{0.306 \pm 0.000}$	$\mathbf{0.301 \pm 0.002}$

Causal Inference with Text Covariates

Treatment W	Outcome Y	Confounder X
Platform 	User experience 	News 
Item 	Click/Conversion 	product reviews 

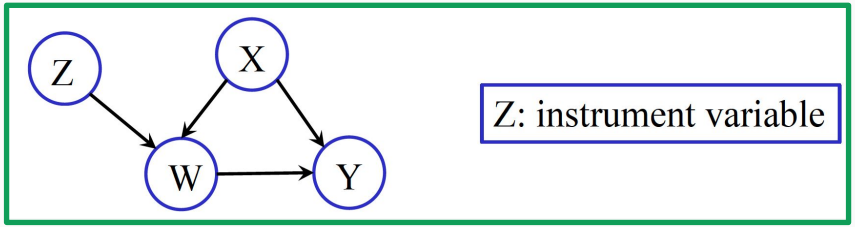
Text type data

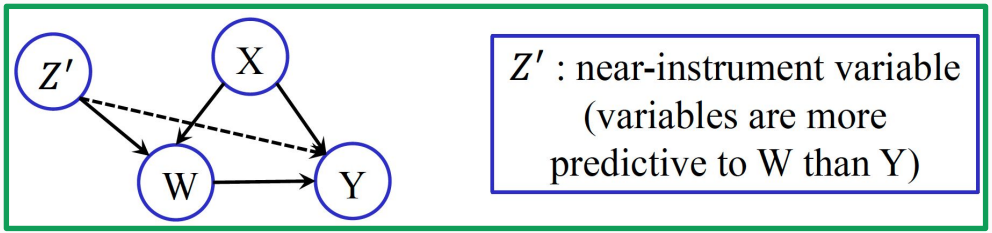
Matching method:

- Counterfactual control outcome:
  - For a treated unit *i*, nearest neighbor matching (NNM) finds its nearest neighbor in control group in terms of covariates.

Causal Inference with Text Covariates

- Challenge: **Hard to filter out near-instrument variables in text covariates case!**
  - Near-instrument variables:
    - very predictive to the treatment assignment might not be that predictive to the outcome.



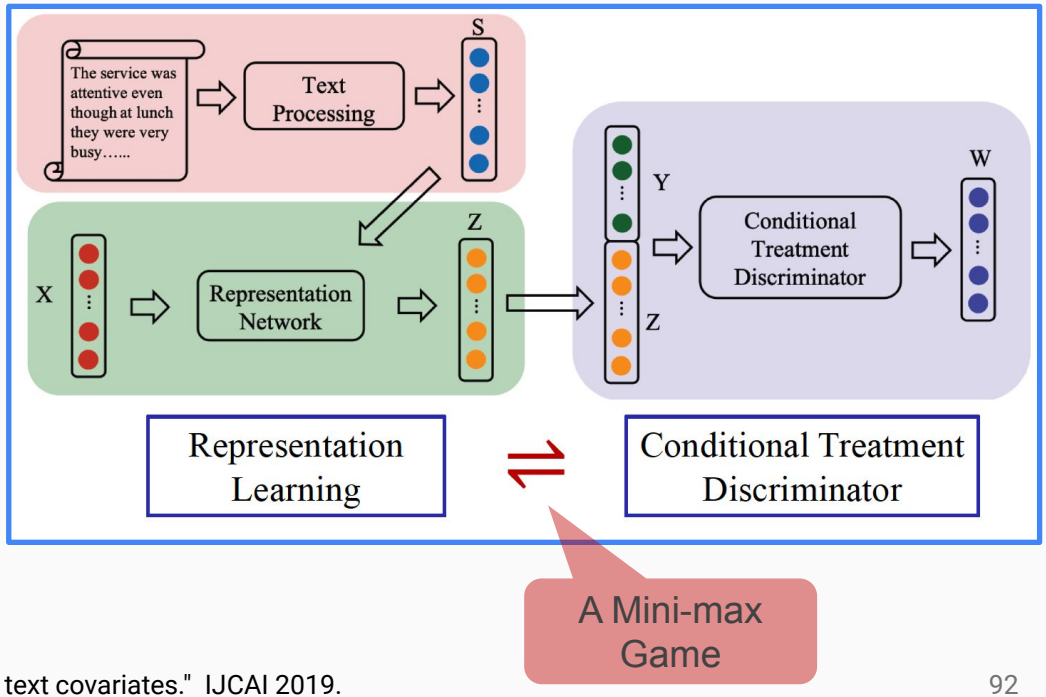


  - Existing works [Pearl, 2012; Wooldridge, 2016] have shown:
    - Matching method: Conditioning on the **near-instrument variables** tends to **amplify the bias** in the analysis of treatment effects.

[Pearl, 2012] Judea Pearl. On a class of bias-amplifying variables that endanger effect estimates. ArXiv preprint arXiv:1203.3503, 2012.
[Wooldridge, 2016] Jeffrey M Wooldridge. Should instrumental variables be used as matching variables? Research in Economics, 70(2):232–237, 2016.

Proposed Method: CTAM

- **Conditional Treatment Discriminator**
  - Input: representation Z & the potential outcome Y
  - Output: the treatment that the unit received. (0 or 1)
- **Mini-max Game**
  - The representation learner aims to fool the Conditional Treatment Discriminator.
  - Filter out the information related to the near-instrument variables.



Liuyi Yao, et al. "On the estimation of treatment effect with text covariates." IJCAI 2019.

CTAM: Experiment

Experiment on News datasets:

- Dataset:
  - Source: NY Times corpus
  - Samples: 5000 news items
  - Treated group: viewed on mobile
  - Control group: viewed on desktop
  - Outcome: reader experience

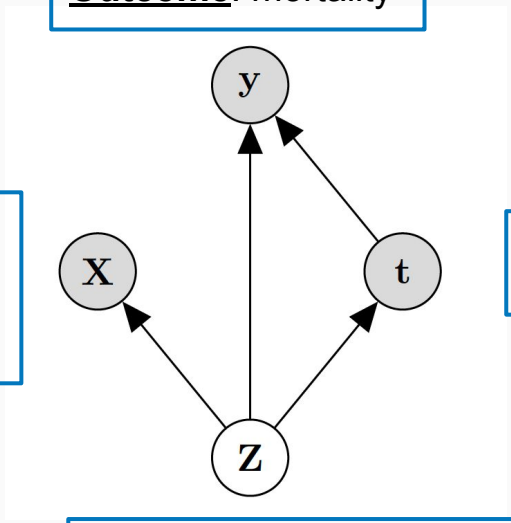
	PEHE	ϵ_{ATE}	ϵ_{ATT}
LASSO	$3.47 \pm 1.26^*$	$0.88 \pm 0.33^*$	$1.75 \pm 0.73^*$
BART	$4.10 \pm 1.27^*$	$1.98 \pm 1.36^*$	$2.87 \pm 1.44^*$
CF	2.69 ± 0.98	$1.88 \pm 0.53^*$	$2.20 \pm 0.80^*$
MDM	$3.29 \pm 0.80^*$	$0.64 \pm 0.61^*$	$0.74 \pm 0.57^*$
PSM	$2.69 \pm 0.33*$	$0.21 \pm 0.14^*$	$0.15 \pm 0.11^*$
DR-RLP	$4.03 \pm 1.44^*$	$0.85 \pm 0.57^*$	0.10 ± 0.05
STM	2.29 ± 0.41	$0.20 \pm 0.15^*$	$\mathbf{0.07 \pm 0.04}$
NNM-HSIC	$4.25 \pm 1.21^*$	$0.83 \pm 0.71^*$	0.12 ± 0.11
CTAM (Ours)	$\mathbf{2.06 \pm 0.03}$	$\mathbf{0.08 \pm 0.01}$	0.09 ± 0.01

The lower, the better.

VAE for Causal Inference

- Hidden confounder:

Outcome: mortality



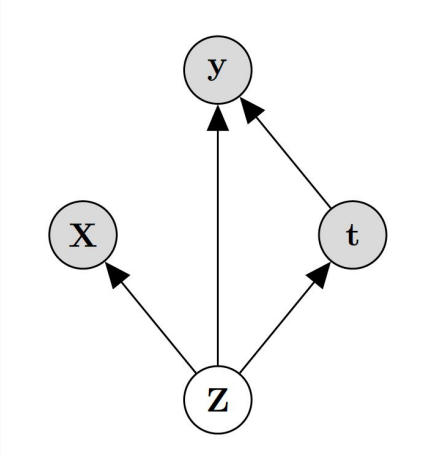
noisy proxies of Z: income in the last year and place of residence

Treatment: medication

Hidden confounder: socio-economic status

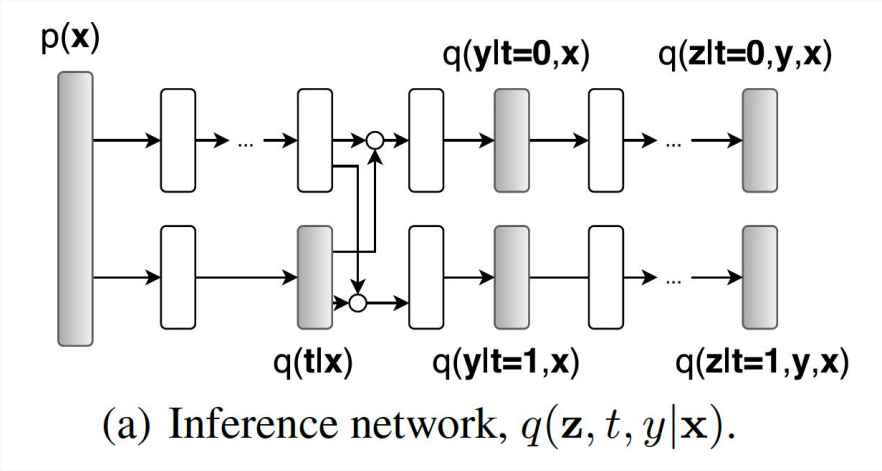
VAE for Causal Inference

- **CEVAE**
  - Requirements: many proxies are available
  - Estimation of a latent-variable model using VAE
    - Discover the hidden confounders
    - Infer how the hidden confounders affect treatment and outcome
  - Advantages:
    - Weaker assumptions about the data generating process and the structure of the hidden confounders

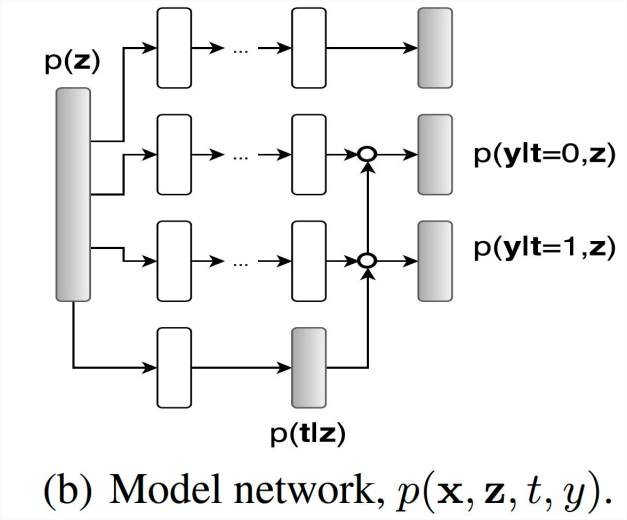


VAE for Causal Inference

- **CEVAE**:



(a) Inference network, $q(\mathbf{z}, t, y|\mathbf{x})$.



(b) Model network, $p(\mathbf{x}, \mathbf{z}, t, y)$.

Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." *Advances in Neural Information Processing Systems*. 2017.

VAE for Causal Inference

- Experiment on IHDP and Jobs datasets:

IHDP

Method	$\sqrt{\epsilon_{\text{PEHE}}^{\text{within-s.}}}$	$\epsilon_{\text{ATE}}^{\text{within-s.}}$	$\sqrt{\epsilon_{\text{PEHE}}^{\text{out-of-s.}}}$	$\epsilon_{\text{ATE}}^{\text{out-of-s.}}$
OLS-1	5.8±.3	.73±.04	5.8±.3	.94±.06
OLS-2	2.4±.1	.14±.01	2.5±.1	.31±.02
BLR	5.8±.3	.72±.04	5.8±.3	.93±.05
k-NN	2.1±.1	.14±.01	4.1±.2	.79±.05
TMLE	5.0±.2	.30±.01	-	-
BART	2.1±.1	.23±.01	2.3±.1	.34±.02
RF	4.2±.2	.73±.05	6.6±.3	.96±.06
CF	3.8±.2	.18±.01	3.8±.2	.40±.03
BNN	2.2±.1	.37±.03	2.1±.1	.42±.03
CFRW	.71±.0	.25±.01	.76±.0	.27±.01
CEVAE	2.7±.1	.34±.01	2.6±.1	.46±.02

Jobs

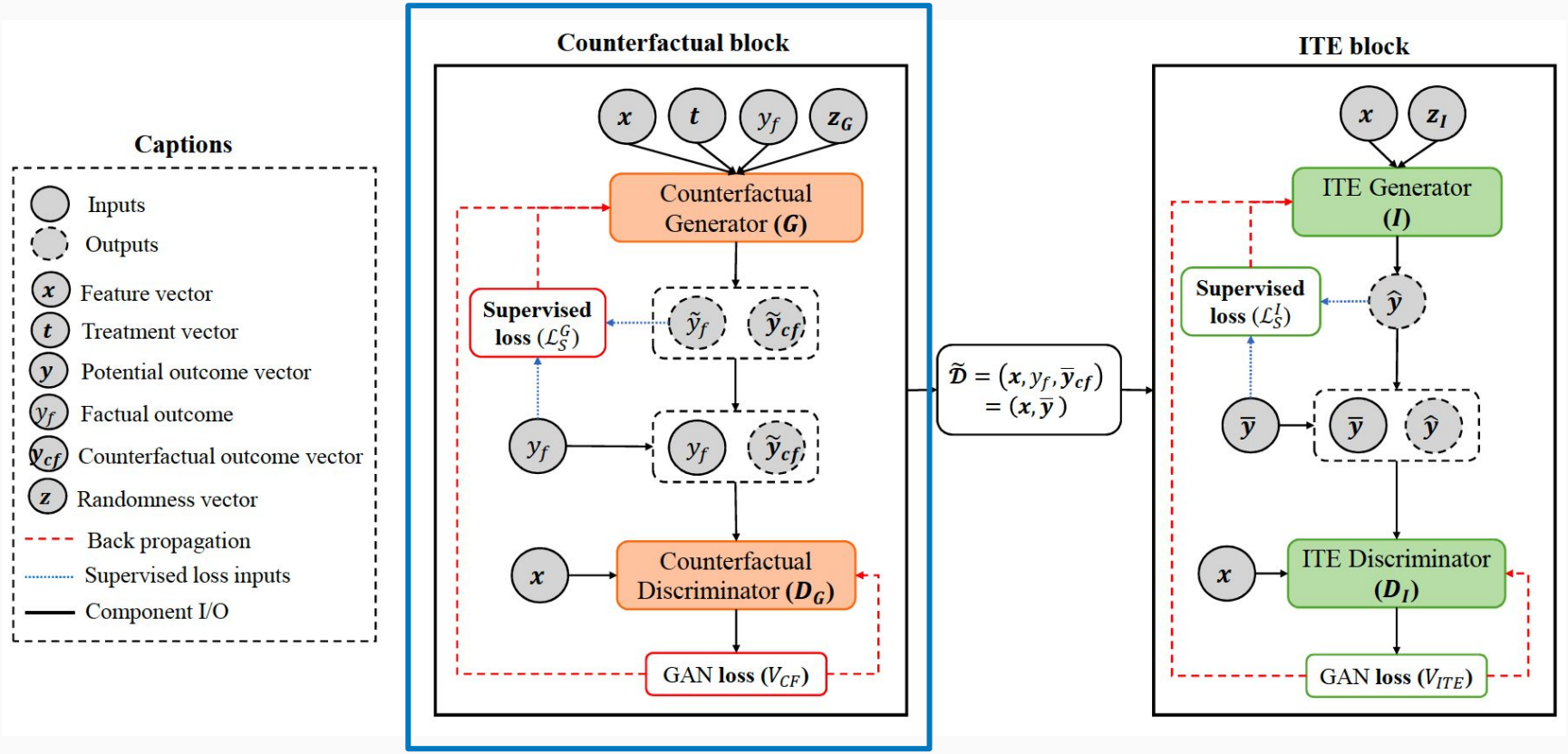
Method	$R_{pol}^{\text{within-s.}}$	$\epsilon_{\text{ATT}}^{\text{within-s.}}$	$R_{pol}^{\text{out-of-s.}}$	$\epsilon_{\text{ATT}}^{\text{out-of-s.}}$
LR-1	.22±.0	.01±.00	.23±.0	.08±.04
LR-2	.21±.0	.01±.01	.24±.0	.08±.03
BLR	.22±.0	.01±.01	.25±.0	.08±.03
k-NN	.02±.0	.21±.01	.26±.0	.13±.05
TMLE	.22±.0	.02±.01	-	-
BART	.23±.0	.02±.00	.25±.0	.08±.03
RF	.23±.0	.03±.01	.28±.0	.09±.04
CF	.19±.0	.03±.01	.20±.0	.07±.03
BNN	.20±.0	.04±.01	.24±.0	.09±.04
CFRW	.17±.0	.04±.01	.21±.0	.09±.03
CEVAE	.15±.0	.02±.01	.26±.0	.03±.01

GAN for Causal Inference: GANITE

- **Motivation**
  - Factual outcome -> observed labels
    Counterfactual outcome -> missing labels
  - Capture the uncertainty in the counterfactual distributions by attempting to learn them using a GAN.
- **Framework**: a combination of two GANs
  - Counterfactual Block:
    - Input: the data with missing labels
    - Goal: estimate the counterfactual outcome
    - Output: the complete data
  - ITE Block
    - A standard GAN
    - Input: the complete data from counterfactual block

Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. "GANITE: Estimation of individualized treatment effects using generative adversarial nets." (2018).

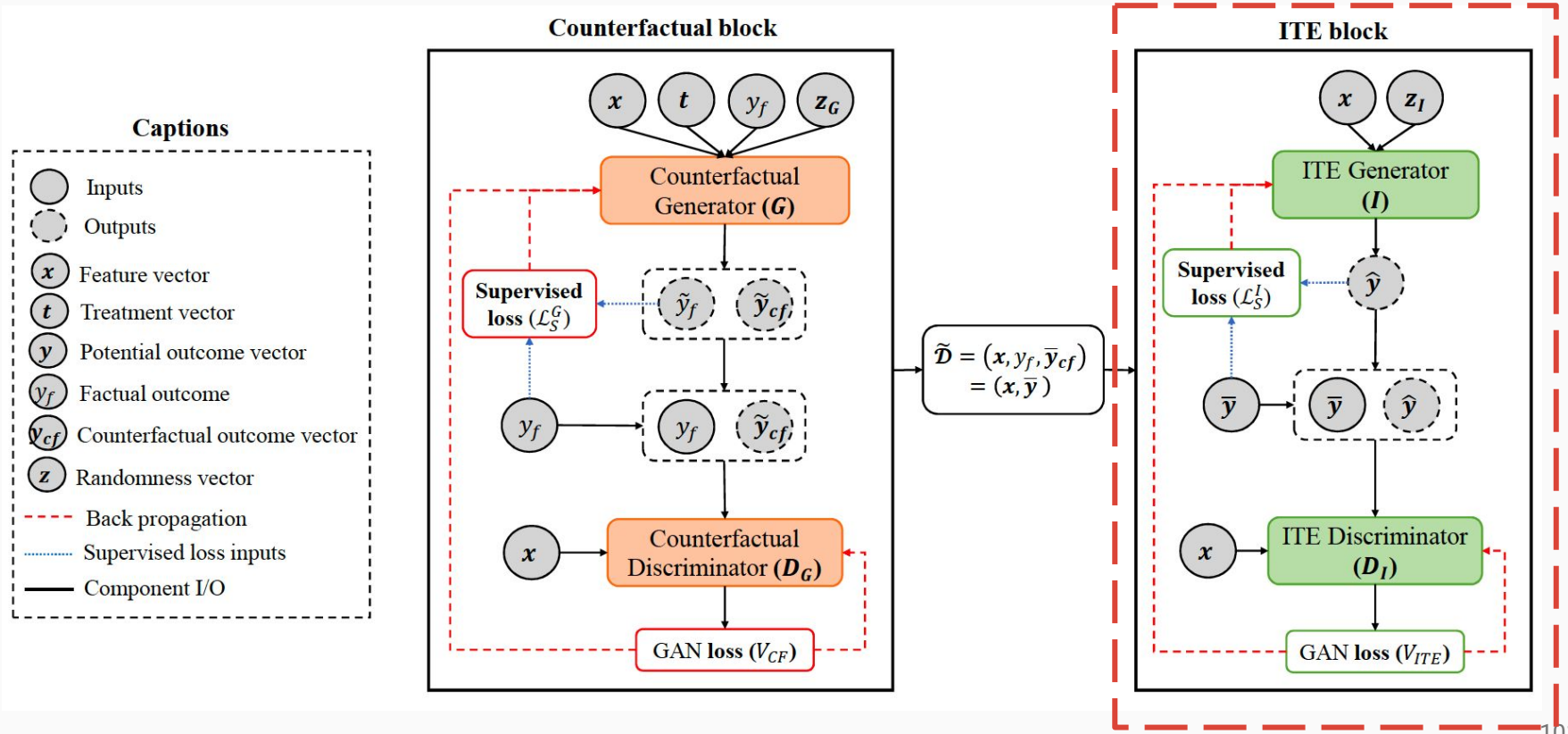
GANITE: Counterfactual Block



GANITE: Counterfactual Block

- **Counterfactual Block**:
  - Counterfactual Generator: generate counterfactual outcomes
  - Counterfactual Discriminator: distinguish which is the factual outcome given the combined vector of factual outcome and generated counterfactual outcome
- **Output of counterfactual Block:**
  - Complete dataset (X, T, factual outcome, generated counterfactual outcome)

GANITE: ITE block



GANITE: ITE block

ITE Block:

- ITE Generator: inferring the potential outcomes of the individual based on the complete dataset in a supervised way
- ITE discriminator:
  - A Standard conditional GAN discriminator.
  - Distinguish the potential outcomes from the complete dataset or the ITE generator

GANITE: Experiment

Experiment on datasets with **binary treatment**.

- Datasets: IHDP, Twins, and Jobs dataset

Methods	Datasets (Mean ± Std)					
	IHDP ($\sqrt{\epsilon_{PEHE}}$)		Twins ($\sqrt{\hat{\epsilon}_{PEHE}}$)		Jobs ($\mathcal{R}_{pol}(\pi)$)	
	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample
GANITE	1.9 ± .4	2.4 ± .4	**.289 ± .005**	**.297 ± .016**	**.13 ± .01**	**.14 ± .01**
OLS/LR_1	5.8 ± .3*	5.8 ± .3*	.319 ± .001*	.318 ± .007	.22 ± .00*	.23 ± .02*
OLS/LR_2	2.4 ± .1	2.5 ± .1	.320 ± .002	.320 ± .003*	.21 ± .00*	.24 ± .01*
BLR	5.8 ± .3*	5.8 ± .3*	.312 ± .003*	.323 ± .018	.22 ± .01*	.25 ± .02*
k-NN	2.1 ± .1	4.1 ± .2*	.333 ± .001*	.345 ± .007*	.02 ± .00	.26 ± .02*
BART	2.1 ± .1	2.3 ± .1	.347 ± .009*	.338 ± .016	.23 ± .00*	.25 ± .02*
R Forest	4.2 ± .2*	6.6 ± .3*	.306 ± .002*	.321 ± .005	.23 ± .01*	.28 ± .02*
C Forest	3.8 ± .2*	3.8 ± .2*	.366 ± .003*	.316 ± .011	.19 ± .00*	.20 ± .02*
BNN	2.2 ± .1	2.1 ± .1	.325 ± .003*	.321 ± .018	.20 ± .01*	.24 ± .02*
TARNET	.88 ± .02	.95 ± .02	.317 ± .005*	.315 ± .003	.17 ± .01*	.21 ± .01*
CFR_{WASS}	.71 ± .02	**.76 ± .02**	.315 ± .007*	.313 ± .008	.17 ± .01*	.21 ± .01*
CMGP	**.65 ± .44**	.77 ± .11	.320 ± .002*	.319 ± .008	.22 ± .03*	.24 ± .05

GANITE: Experiment

Experiment on datasets with **multiple treatments**.

- Datasets: constructed from original Twins dataset
- 4 treatments:
  t = 1: lower weight, female sex
  t = 2: lower weight, male sex
  t = 3: higher weight, female sex
  t = 4: higher weight, male sex
- Evaluation metric:
  mean-squared error on outcomes:

$$\text{MSE}_y = \frac{1}{N \times |\mathcal{T}_i|} \sum_{i=1}^{N} \sum_{t \in \mathcal{T}_i} \left(y_t(x_i) - \hat{y}_t(x_i) \right)^2$$

Methods	Metric: MSE_y			
	In Sample	Gain (%)	Out Sample	Gain (%)
GANITE	**.0427 ± .0161**	(-)	**.0723 ± .0183**	(-)
OLS/LR_1	.0855 ± .0096	50.1%	.0871 ± .0142	17.0%
OLS/LR_2	.0857 ± .0099	50.2%	.0883 ± .0147	18.1%
BLR	.0996 ± .0081*	57.1%	.1017 ± .0127	28.9%
KNN	.0930 ± .0101*	54.1%	.1008 ± .0236	28.3%
BART	.1097 ± .0084*	61.1%	.1037 ± .0283	30.3%
R Forest	.0442 ± .0069	3.4%	.0927 ± .0138	22.0%
C Forest	.1607 ± .0014*	73.4%	.1665 ± .0035*	56.6%
BNN	.0602 ± .0102	29.1%	.1031 ± .0145	29.9%
TARNET	.0854 ± .0091	50.0%	.0879 ± .0030	17.7%
CFR_{WASS}	.0896 ± .0036*	52.3%	.0894 ± .0057	19.1%
CMGP	.0844 ± .0073*	49.4%	.0793 ± .0191	8.3%

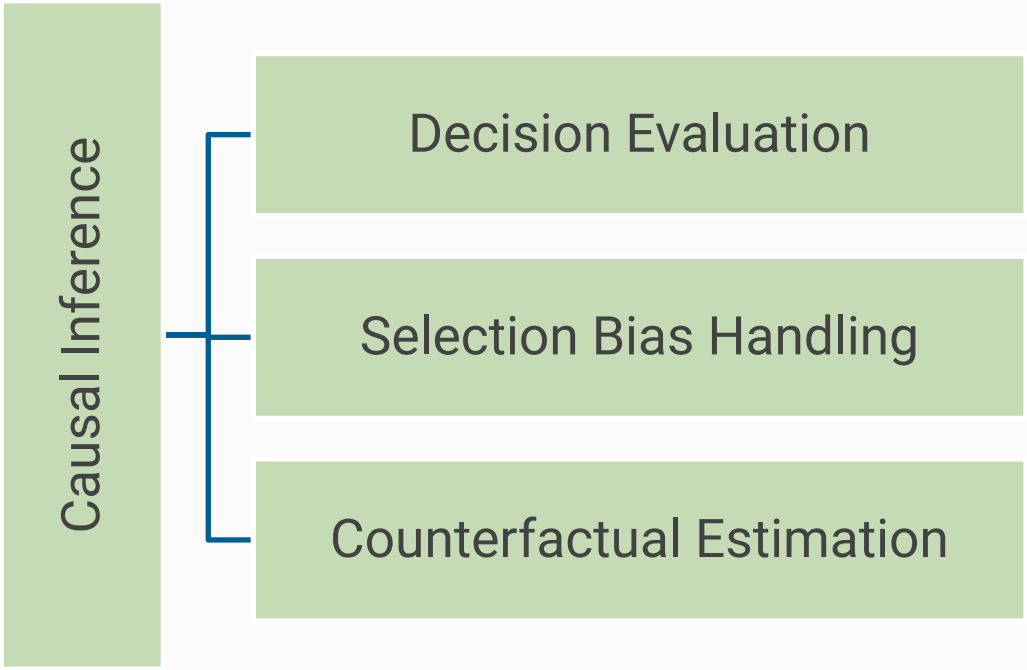
Summary

- **Deep Representation Learning for Causal Inference**
  - Balanced representation learning
  - Capture the uncertainty in the distribution of counterfactual/unobserved confounders
  - Drawbacks:
    - Some methods lack theoretical guarantee
    - Require large samples to train

Outline

- ❑ Overview
- ❑ Causal Inference: Background and Challenges
- ❑ Classical Causal Inference Methods
- ❑ Subspace Learning for Causal Inference
- ❑ Deep Representation Learning for Causal Inference
- ❑ **Applications and Potential Directions**
- ❑ Conclusions

Applications



Decision Evaluation: Online Advertising

Will the ad attract user clicks?

Will a campaign increase sales?

Randomized experiments such as A/B testing?

Time-consuming and Expensive

Estimating the ad effect from observational data!

Decision Evaluation: Online Advertising

- Online Advertising as Causal Inference:

  Estimating the ad effect from observational data

Observational data → Logged feedback records under **current** advertising system's policy

Treatment W

Ads

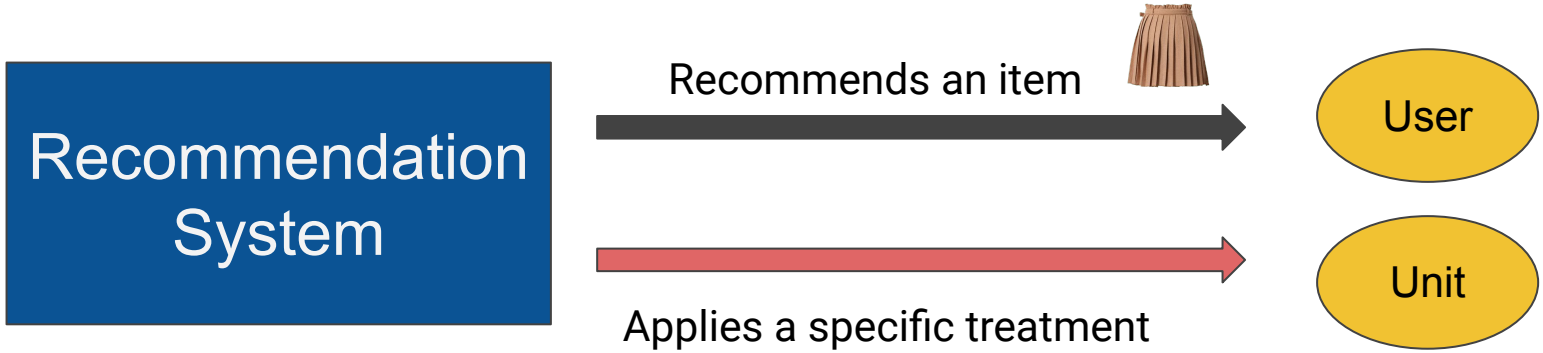
Outcome Y

Click

Variable X

Ad content

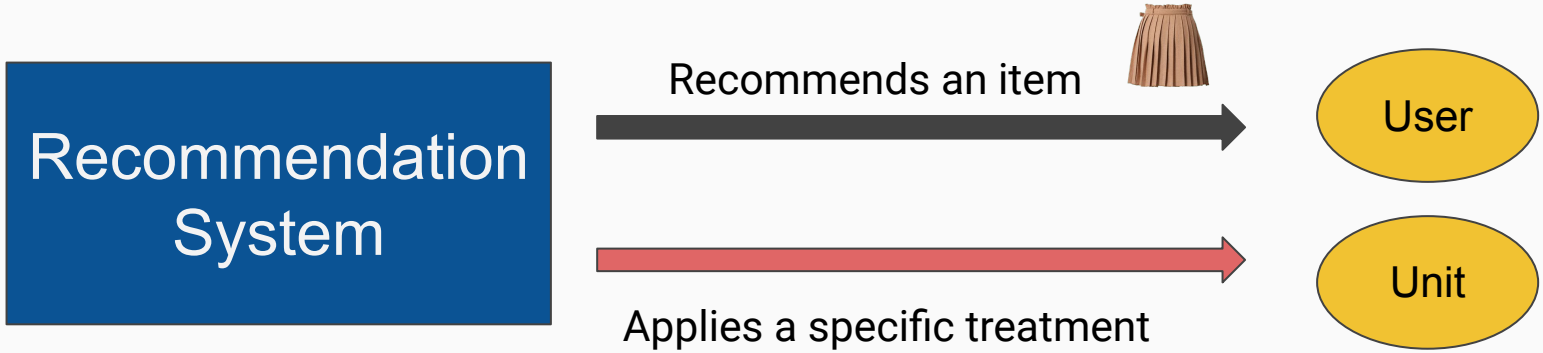
Selection Bias Handling: Recommendation

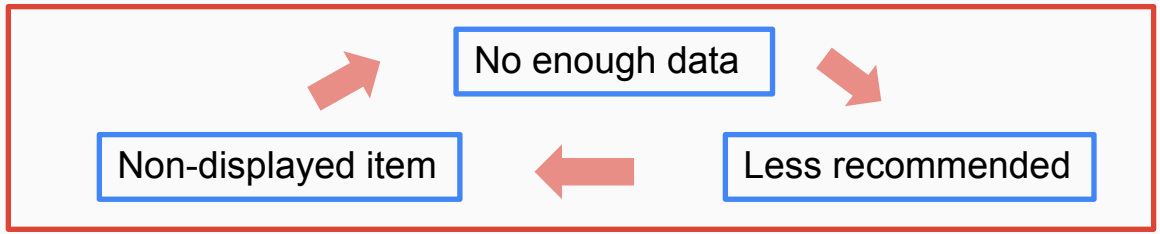


Selection bias:

- Users tend to rate the items that they like:
  - The horror movie ratings are mostly made by horror movie fans and less by romantics movie fans.
- The records in the datasets are **not representative** of the whole population.

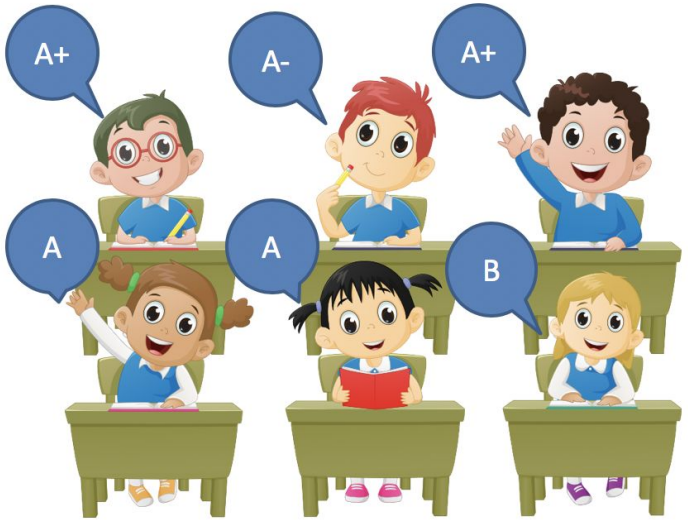
Selection Bias Handling: Recommendation





Counterfactual Estimation: Education

What would happen if the teacher adopted another teaching method?



Teachers can find the best teaching method for each individual!

Potential Direction: Perspective of Treatment

- **Similar treatments**:

  Finding neighbors that have similar treatments.

- **Multiple treatments**:

  Each treatment has different levels.

- **Continuous treatments:**

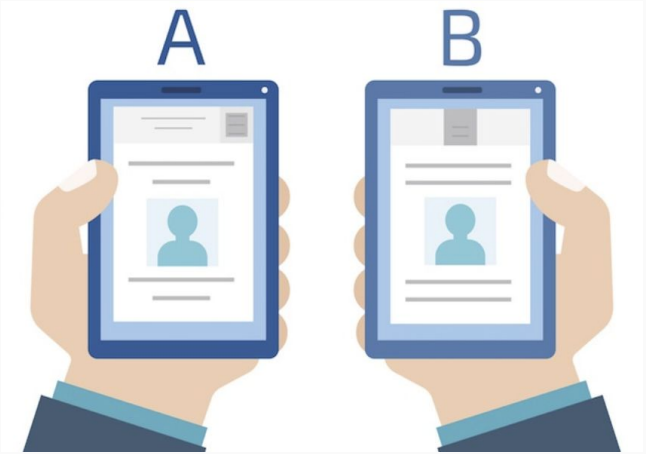
  Treatment can take values from a continuous range.

- **Causal interaction**:

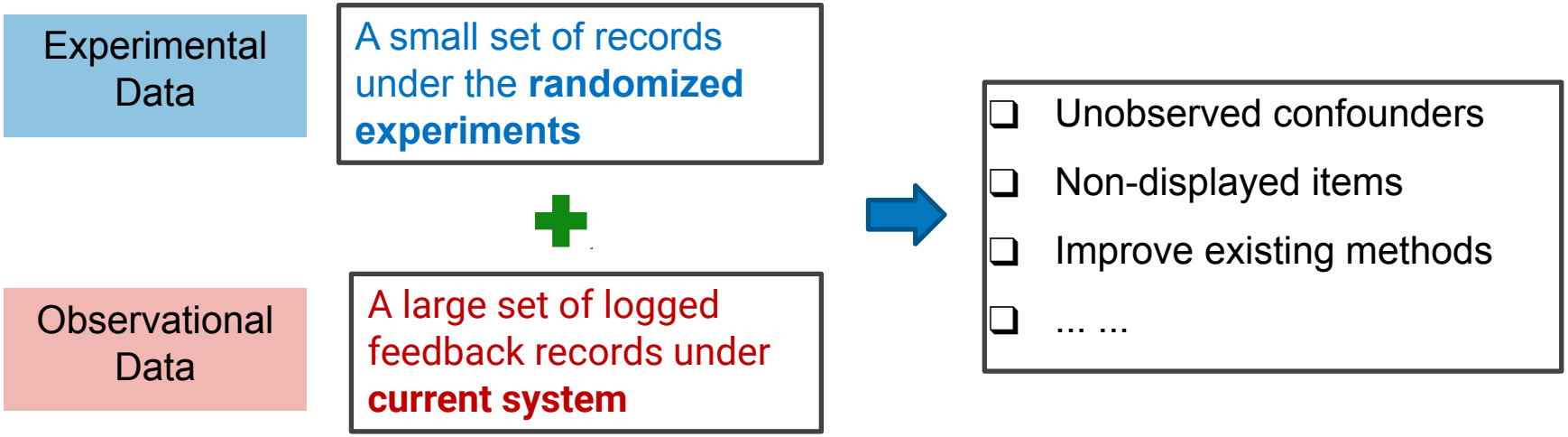
  Identifying the effect of combinations of treatments.

Potential Direction: Evaluation

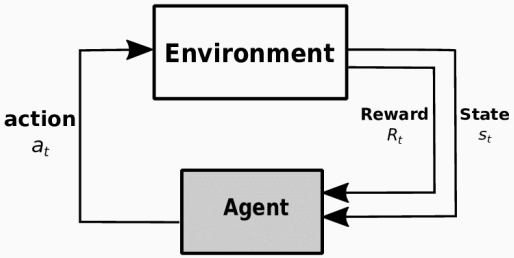
For real-world applications, how can we evaluate the performance of different causal inference methods?



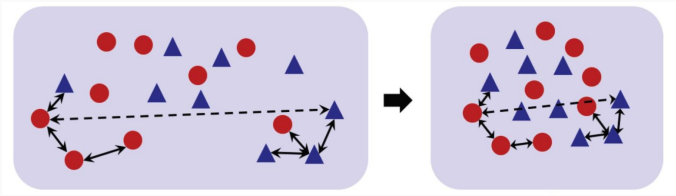
Potential Direction: Data Fusion



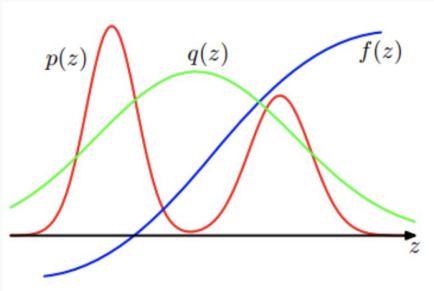
Potential Direction: Connecting other areas



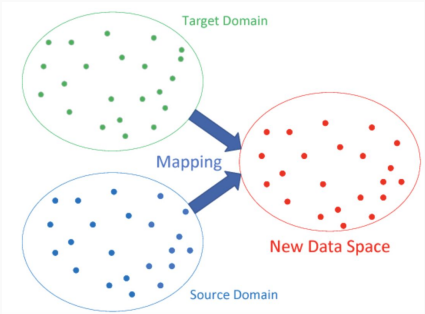
Reinforcement Learning



Representation Learning based Causal Inference



Statistical Sampling



Transfer Learning

Outline

- ❏ Overview
- ❏ Causal Inference: Background and Challenges
- ❏ Classical Causal Inference Methods
- ❏ Subspace Learning for Causal Inference
- ❏ Deep Representation Learning for Causal Inference
- ❏ Applications and Potential Directions
- ❏ **Conclusions**

Conclusions

- Challenges in causal inference
- Connections between machine learning problems and causal inference problems
- Subspace learning approaches for causal inference
- Deep representation learning approaches for causal inference
- Real-world applications

Thank you!

- **Website**: http://kdd2020tutorial.thumedialab.com/
- **A Survey on Causal Inference** (Feb., 2020): https://arxiv.org/abs/2002.02770