Introduction to Causal Inference

Cheng Wan @ KC Seminar 2017

Material

- ICML 2016 Tutorial: Causal Inference for Observational Studies [Shalit et al. 2016]
- Causal inference in statistics: An overview [Pearl 2010]
- Learning Representations for Counterfactual Inference [Johansson et al. 2016]
- Causal Inference for Recommendation [Liang et al. 2016]
- Counterfactual Risk Minimization: Learning from Logged Bandit Feedback [Swaminathan et al. 2015]
- A Crash Course in Causality: Inferring Causal Effects from Observational Data [Coursera]

Outline

- Definition
- Methods
- Misc

Outline

- Definition
- Methods
- Misc

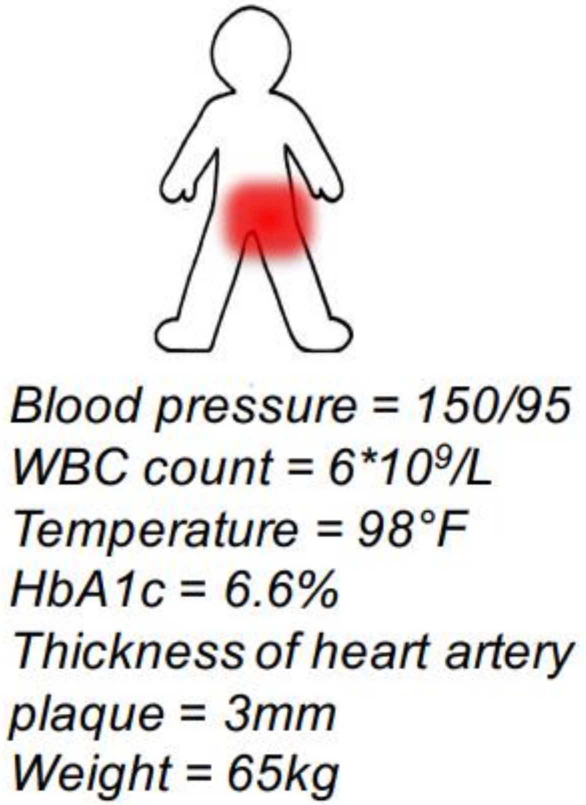
Definition

- **Causal inference** is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect. (Wikipedia)

Example 1: Precision Medicine

- Which treatment is the best for me?
  - Treatment A or treatment B?
- Current situation
  - Clinical trials
  - Doctor’s knowledge & intuition

- **Individualized Treatment Effect (ITE)**



Example 2: Job Training

- Should the government fund a job training program?
- Current situation
  - Past education and employment data

- **Average Treatment Effect (ATE)**

Typical Causal Inference Problem

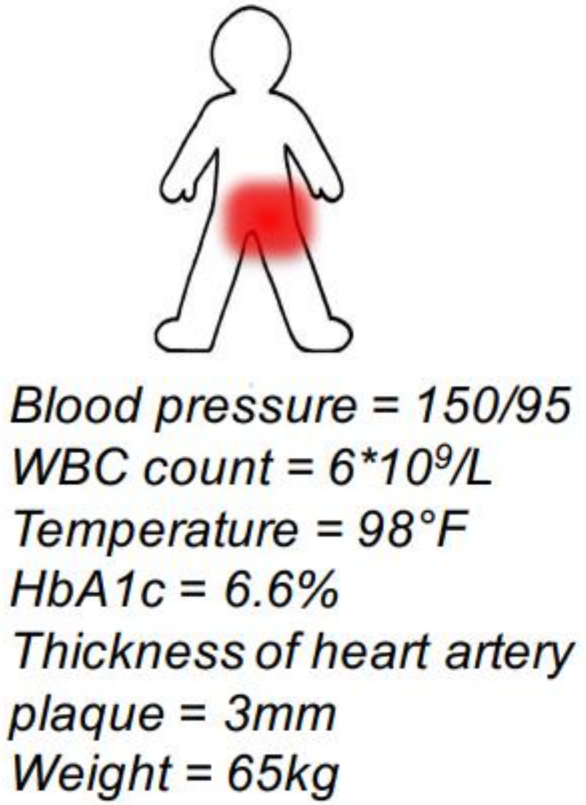
- Causal inference is the process of drawing a conclusion about a causal connection based on the conditions of **the occurrence of an effect**.

- One interesting parameter (**treatment**)
  - Input is denoted as (feature, treatment)

Challenge: Potential Confounder

- Which treatment is the best for me?
  - Cure rate of A: 80%
  - Cure rate of B: 90%
  - Treatment B is better than treatment A?

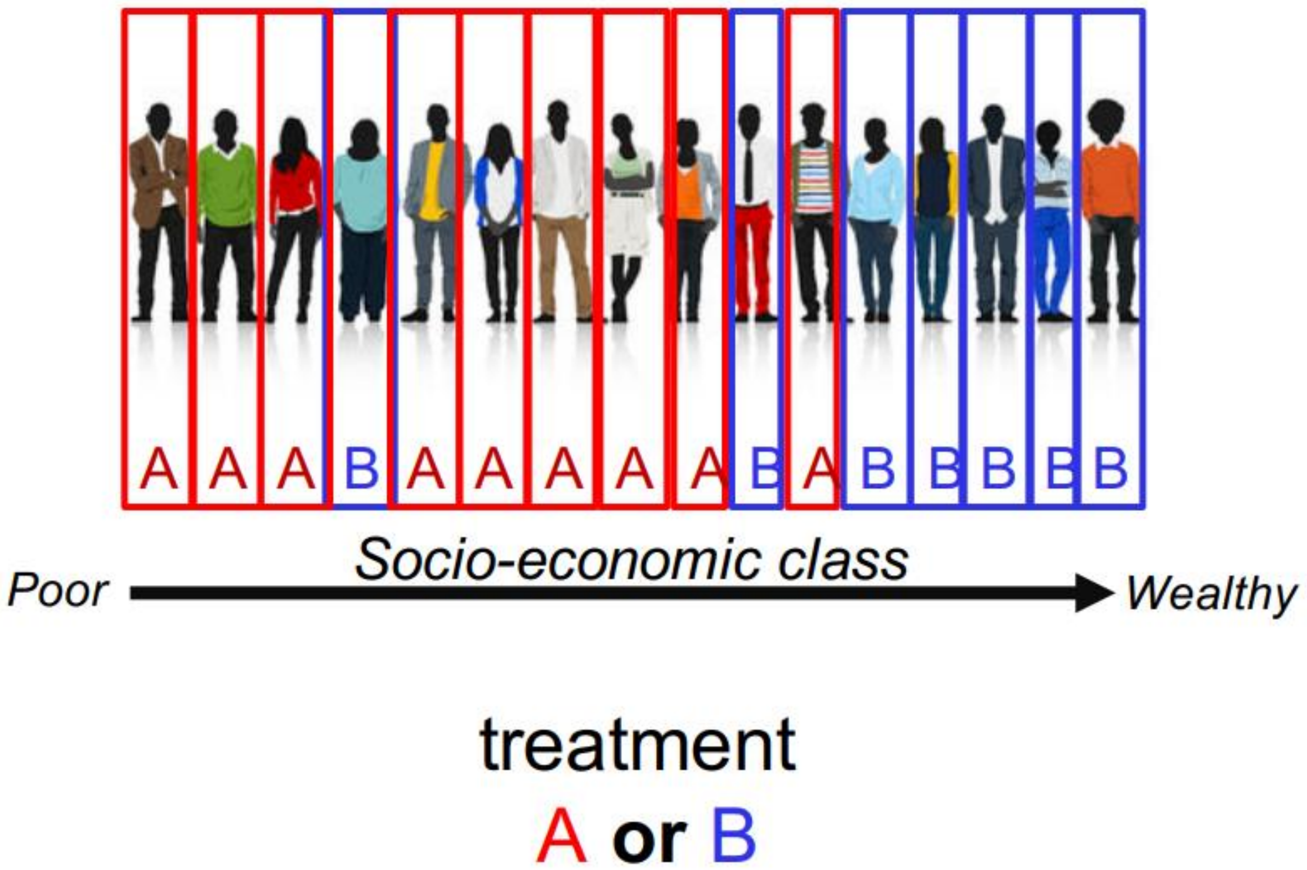
- **Potential confounder**: wealth, policy



Challenge: Potential Confounder

- Should the government fund a job training program?
- Current situation
  - Past education and employment data
- **Potential confounder**: motivation

Challenge: Non-uniform Distribution



Outline

- Definition
- Methods
- Misc

Two Approaches

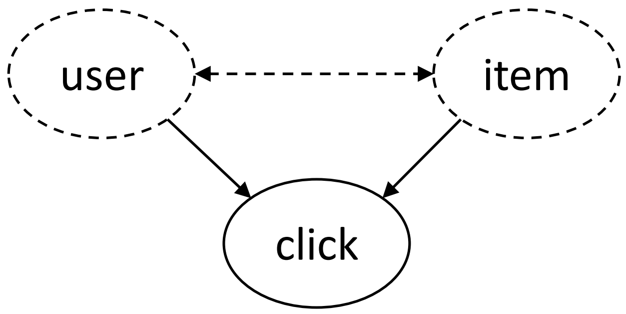
- Estimating Data
- Dealing with biased Data

Estimating Data: Example

- Company X wants to improve its recommendation system.
- Collected data: *(user, item, click)*

Estimating Data: Example

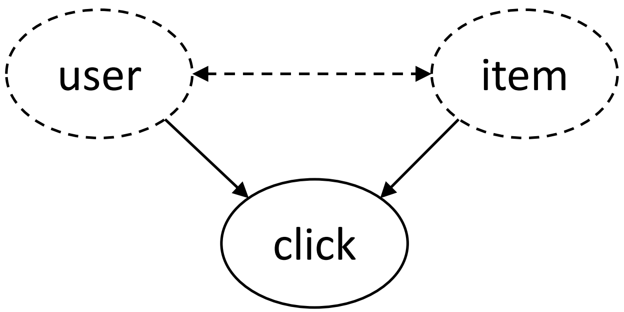
- Company X wants to improve its recommendation system.
- Collected data: *(user, item, click)*



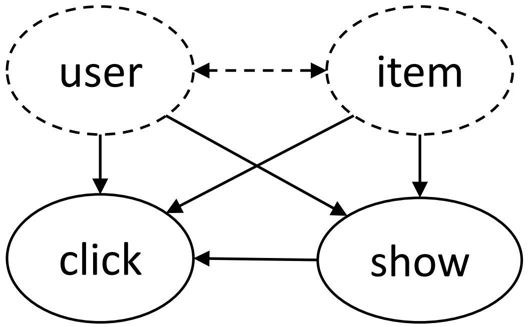
Assumption

Estimating Data: Example

- Company X wants to improve its recommendation system.
- Collected data: *(user, item, click, show)*



Assumption

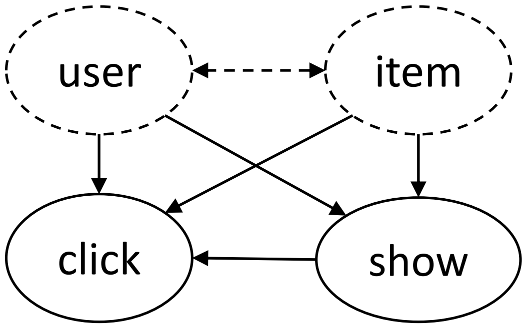


Real

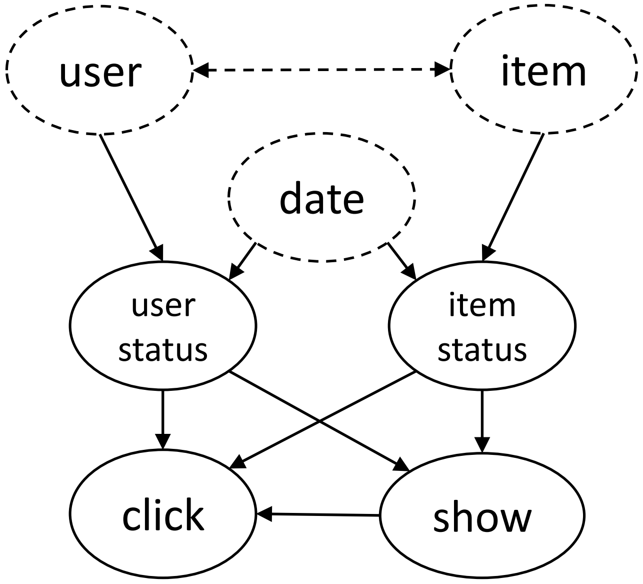
[Liang et al. 2016]

Estimating Data: Example

- Company X wants to improve its recommendation system.
- Collected data: *(user, item, click, show)*



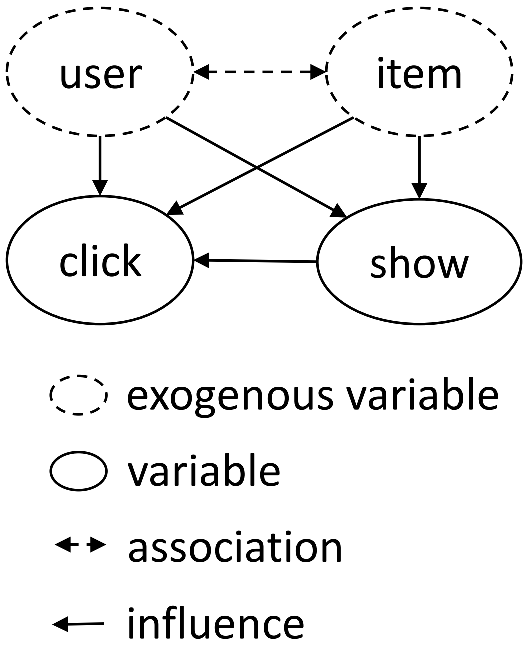
Assumption



Real

Estimating Data: Causal Graphs

- Causal Graphs: graphical models used to encode assumptions about the data-generating process
  - Causal assumptions are encoded in the **missing** links
  - Testing assumptions using *d-separation*
  - A great way of formalizing when is correct causal inference *possible* in face of unmeasured confounders



[Wright 1921, Engle 1983, Pearl 1988]

Dealing with Biased Data: Terminology

- ***Treatment***: (binary) indicator
- ***Treated***: units who received treatment=1
- ***Control***: units who received treatment=0
- ***Factual***: the set of observed units with their respective treatment assignment
- ***Counterfactual***: the factual set with flipped treatment assignment

Dealing with Biased Data: Symbol

- x_i: features (context)
- $Y_0(x_i), Y_1(x_i)$: potential outcomes for control and treated
- t_i: treatment assignment
- Observed factual outcome and unobserved counterfactual outcome:
  - $y_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$
  - $y_i^{CF} = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$
- $ITE(x_i) = \mathbb{E}[Y_1 | x_i] - \mathbb{E}[Y_0 | x_i]$
- $ATE = \mathbb{E}_{x \sim p(x)}[ITE(x)]$

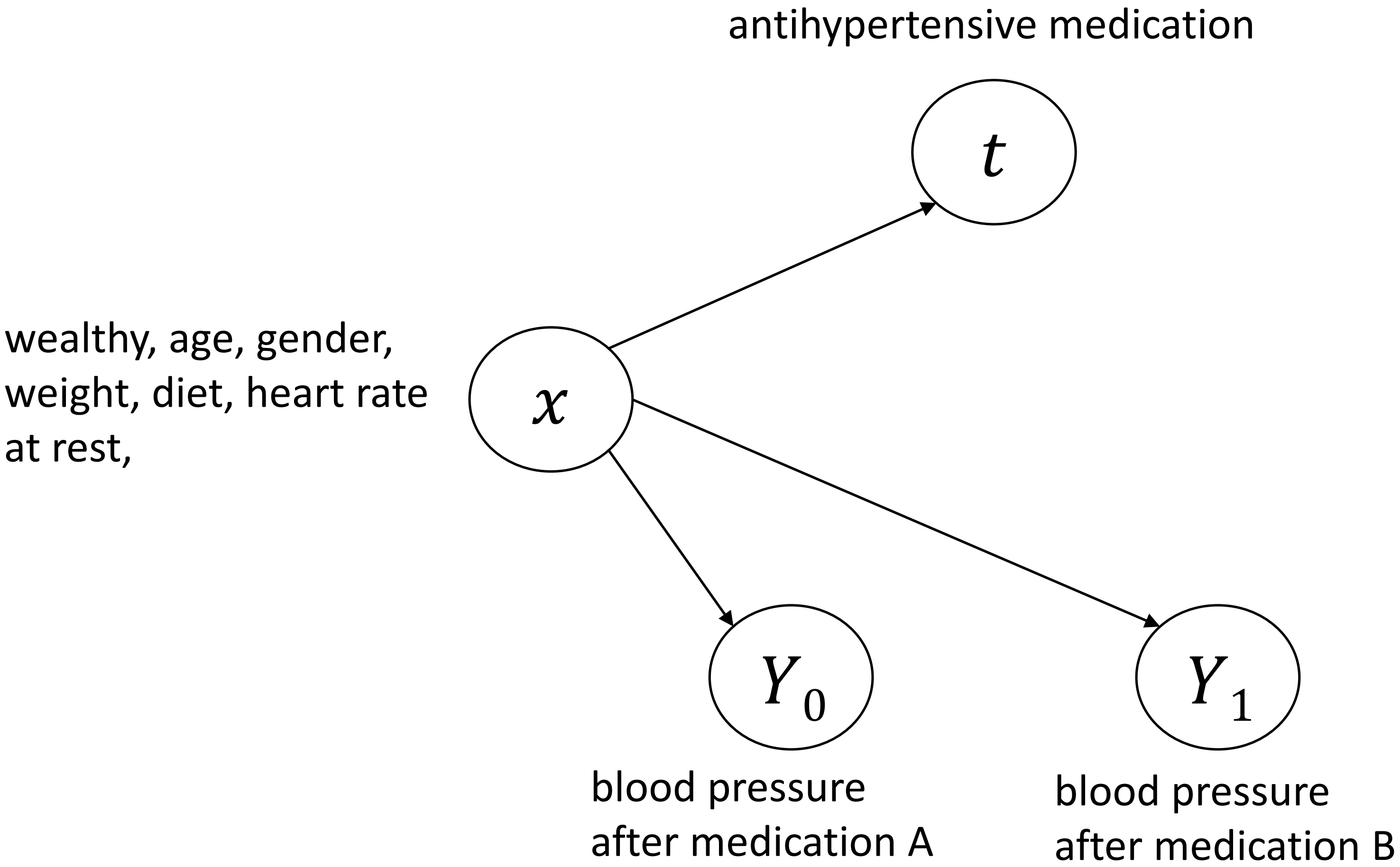
Target

Assumption: no unmeasured confounders

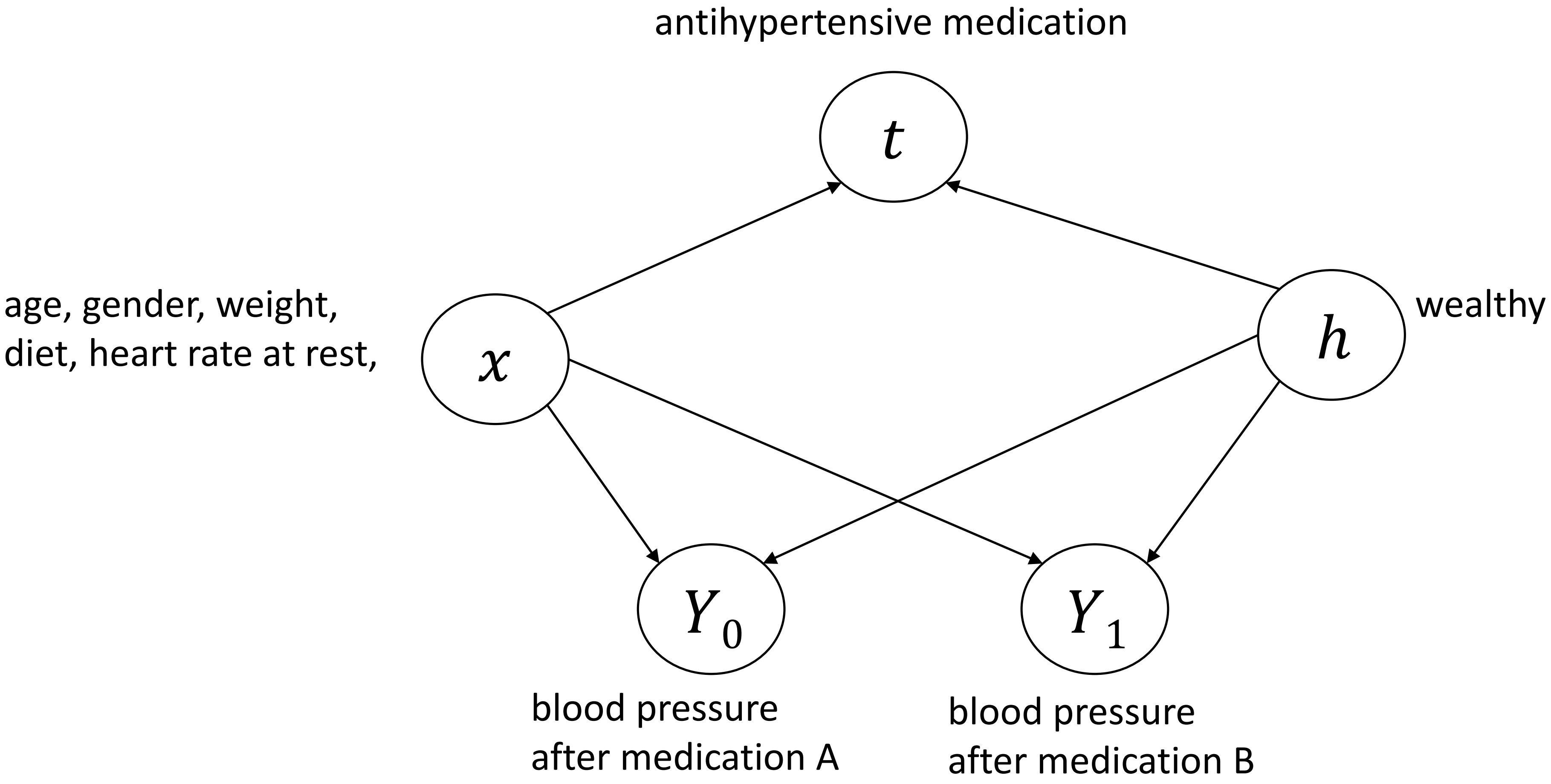
$$(Y_0, Y_1) \perp\!\!\!\perp t | x$$

Ignorability

Ignorability



No Ignorability

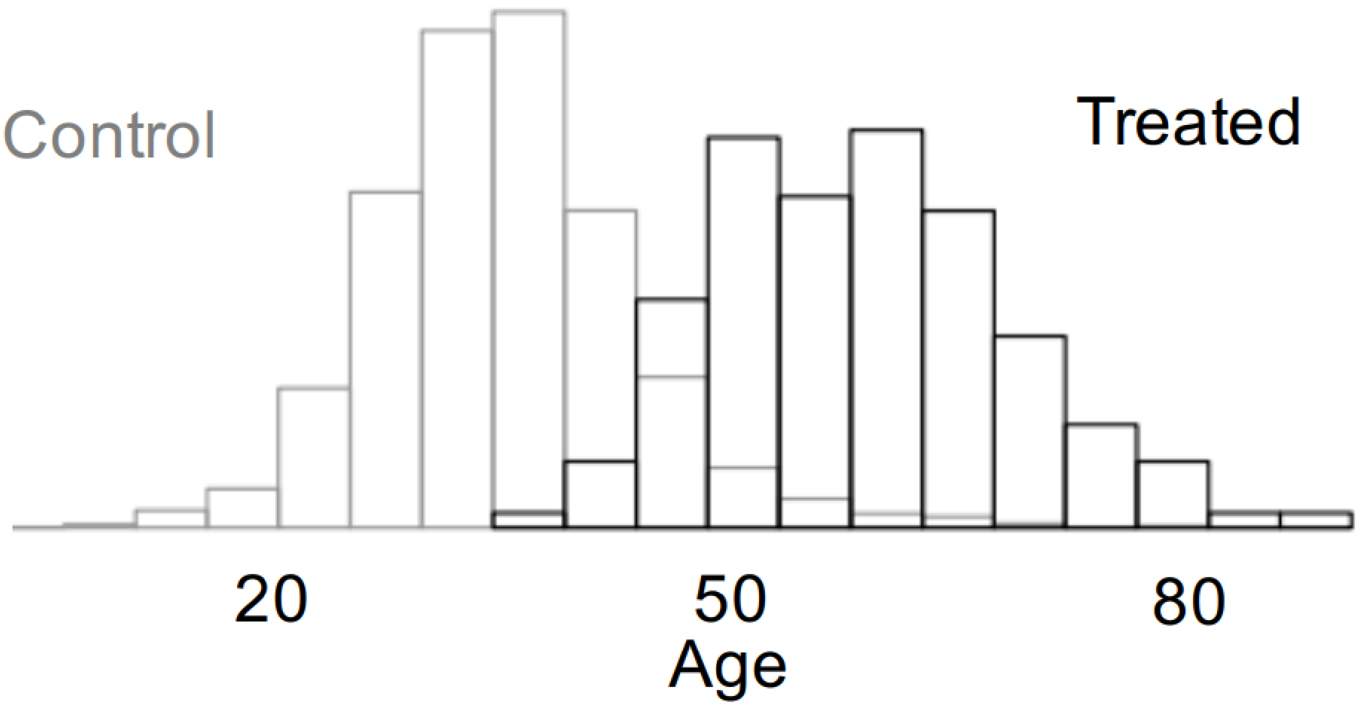


Assumption: common support

$$p(T = t|X = x) > 0 \; \forall t, x$$

Overlap

Overlap



[Gelman et al. 2007]

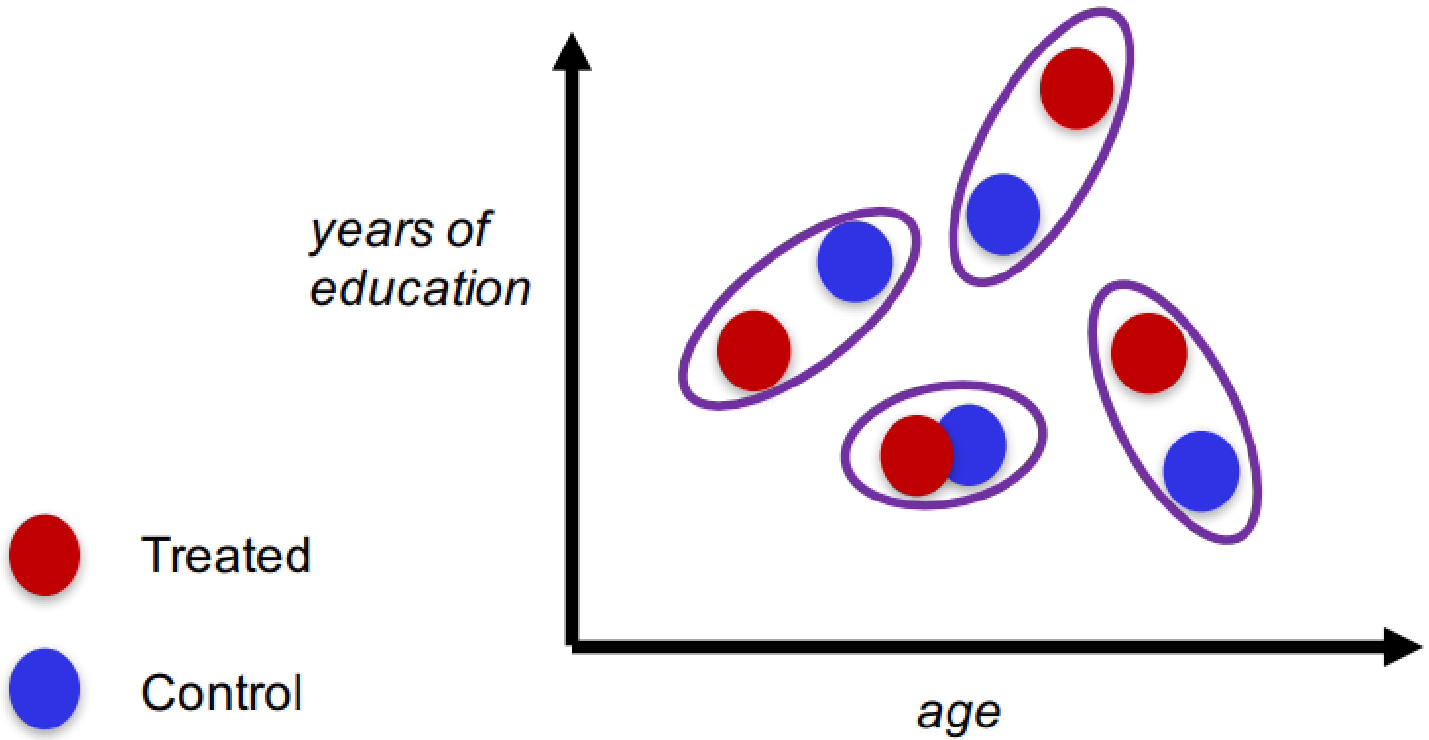
Methods: Overview

- Matching
- Covariate adjustment
- Propensity score
- Double robustness
- Causal forests
- Learning balanced representation
- Counterfactual risk minimization

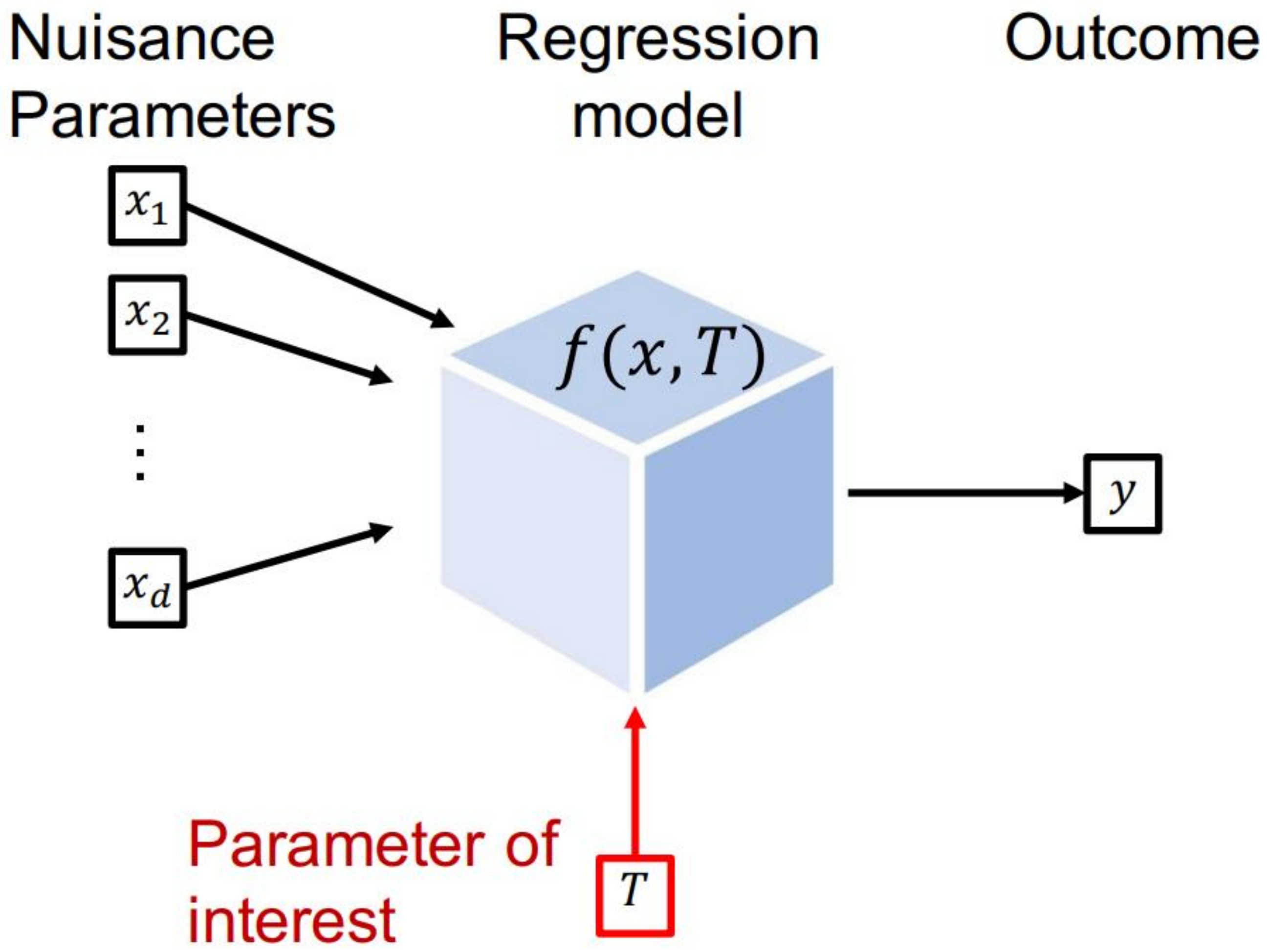
ML approach

Matching

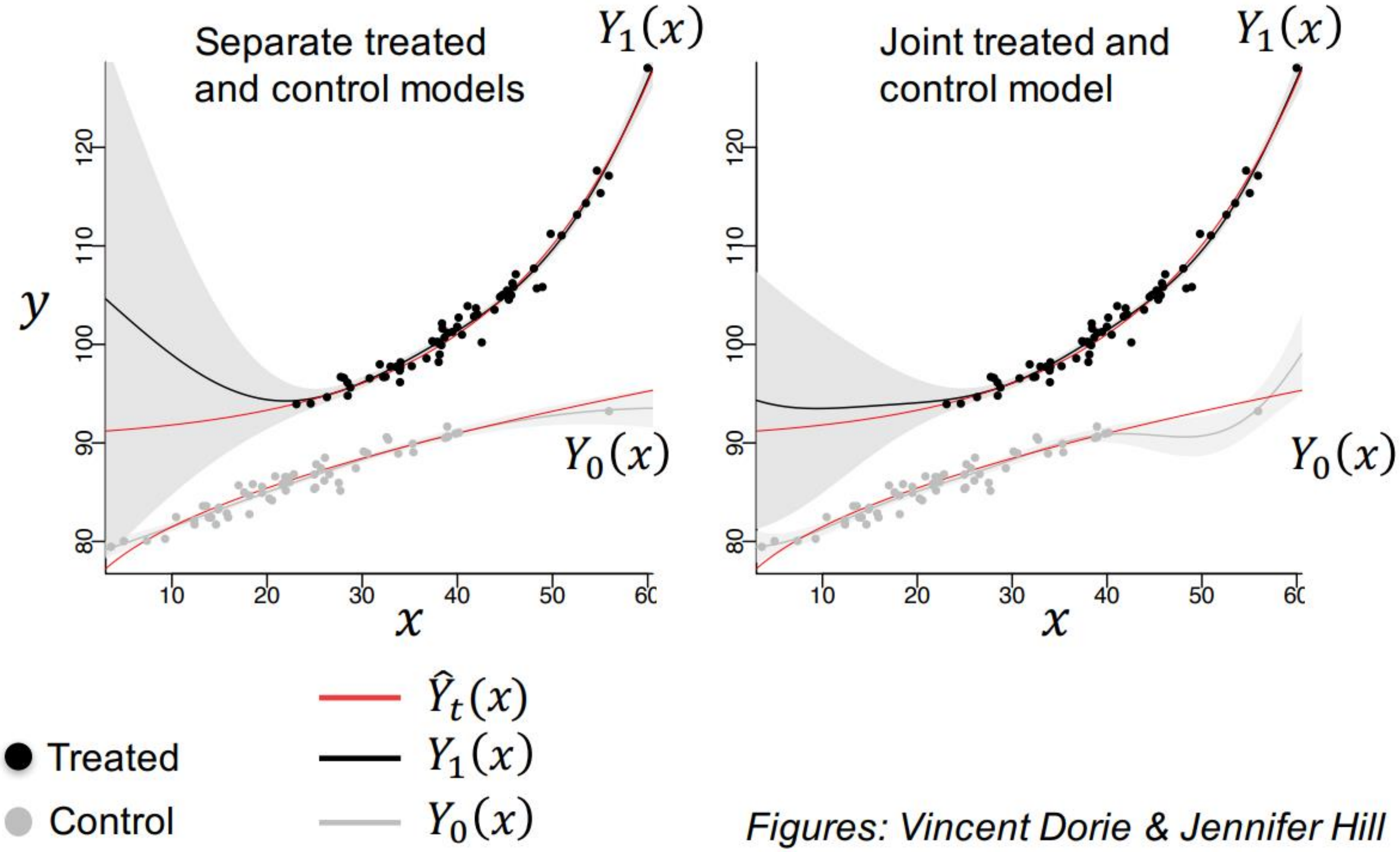
- Match to nearest neighbor from opposite group



Covariate Adjustment

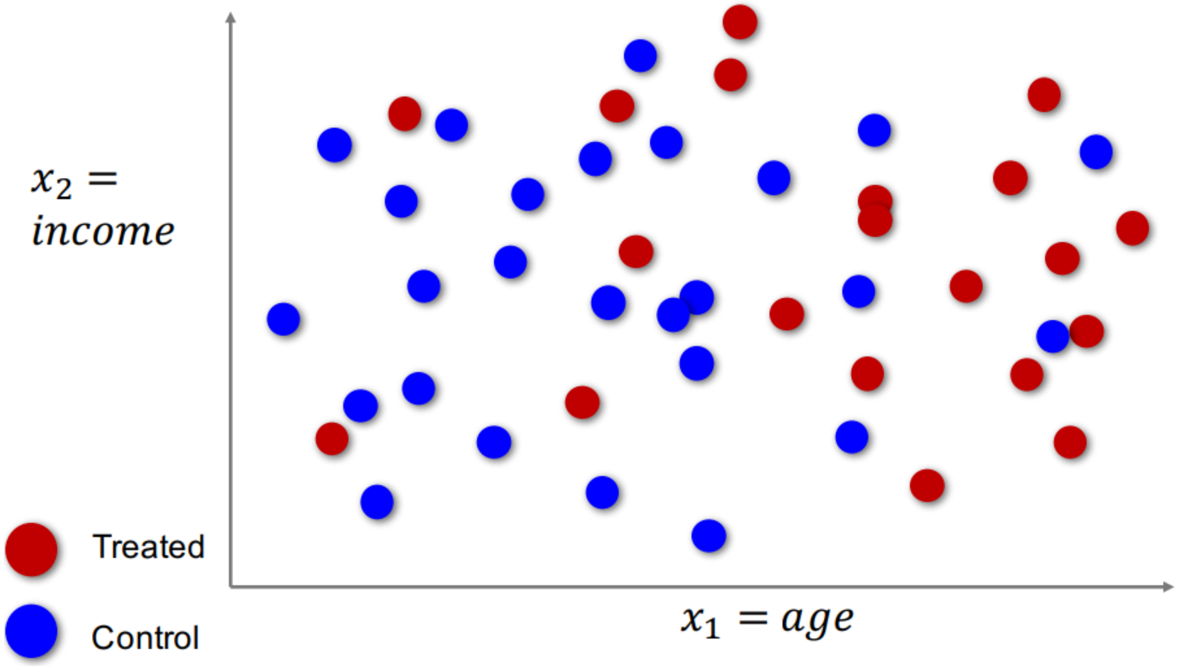


Covariate Adjustment



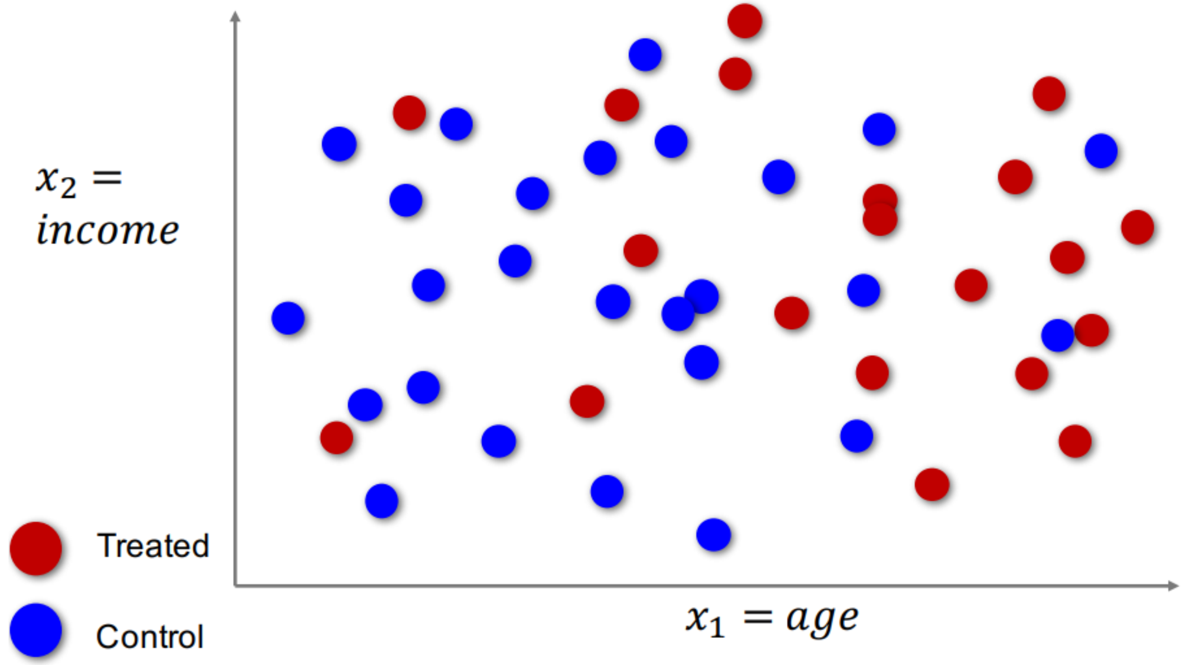
Figures: Vincent Dorie & Jennifer Hill

Propensity Score



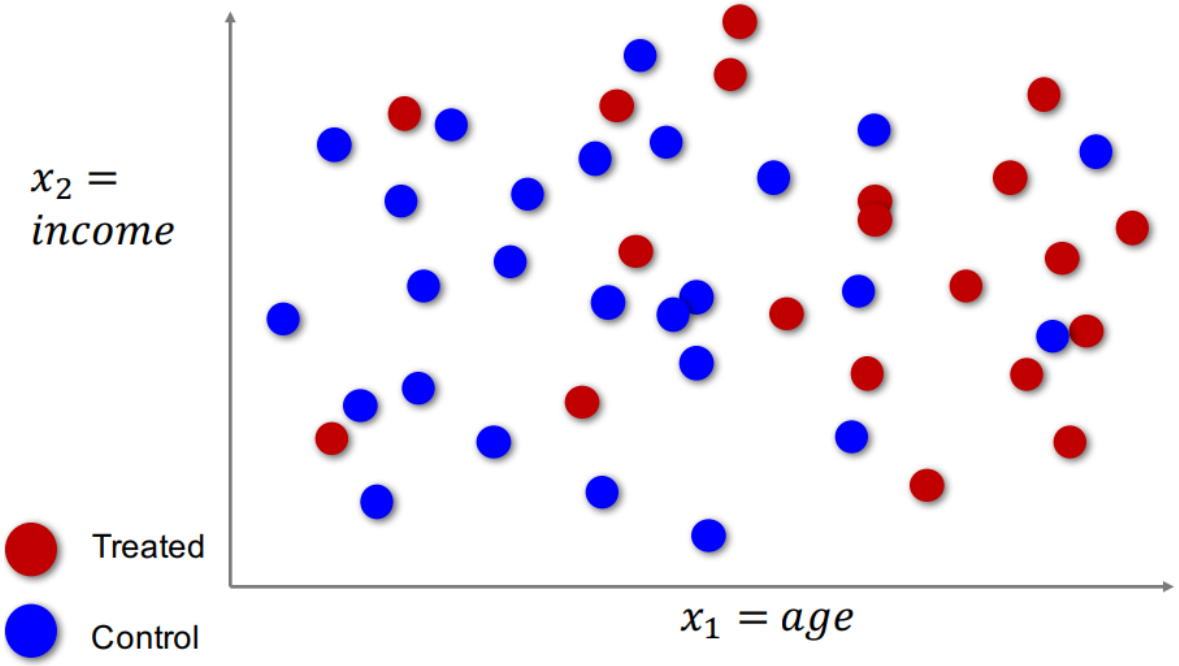
$$p(x|t = 0) \neq p(x|t = 1)$$

Propensity Score



$$p(x|t=0) \cdot w_0(x) = p(x|t=1) \cdot w_1(x)$$

Propensity Score



$$w_i(x) = \frac{p(t = i)}{p(t = i|x)}$$ ⇨ propensity score

Double Robustness

- Combining covariate adjustment and propensity score
  - an estimator which is unbiased if at least one of the models is well-specified

$$\mathbb{E}_{x\sim p(x)}[Y_1] = \frac{1}{n}\sum_{i=1}^{n}\frac{t_i y_i}{\pi_1(x_i)} \qquad \text{(propensity score)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(t_i y_i + (1-t_i)m_1(x_i)) \qquad \text{(covariate adjustment)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_i y_i - (t_i - \pi_1(x_i))m_1(x_i)}{\pi_1(x_i)}\right) \qquad \text{(double robustness)}$$

Causal Forest

Maximize variance of $\widehat{ITE}$

$$\hat{ITE}(x) = \frac{1}{\#\text{treat in leaf}(x)} \sum_{\substack{i \in \text{leaf}(x) \\ t_i = 1}} y_i - \frac{1}{\#\text{control in leaf}(x)} \sum_{\substack{i \in \text{leaf}(x) \\ t_i = 0}} y_i$$

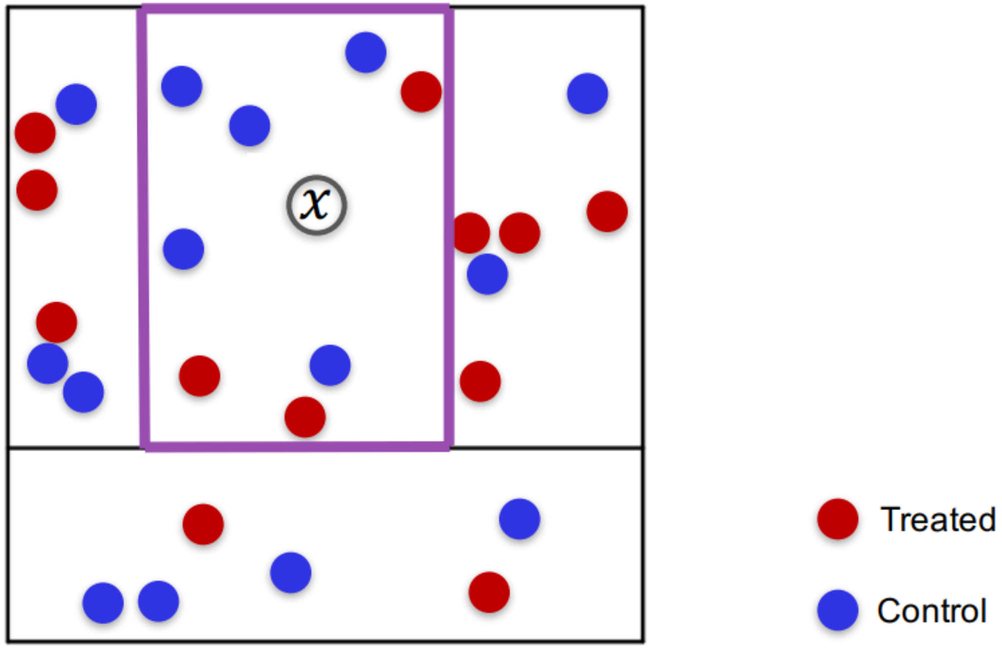
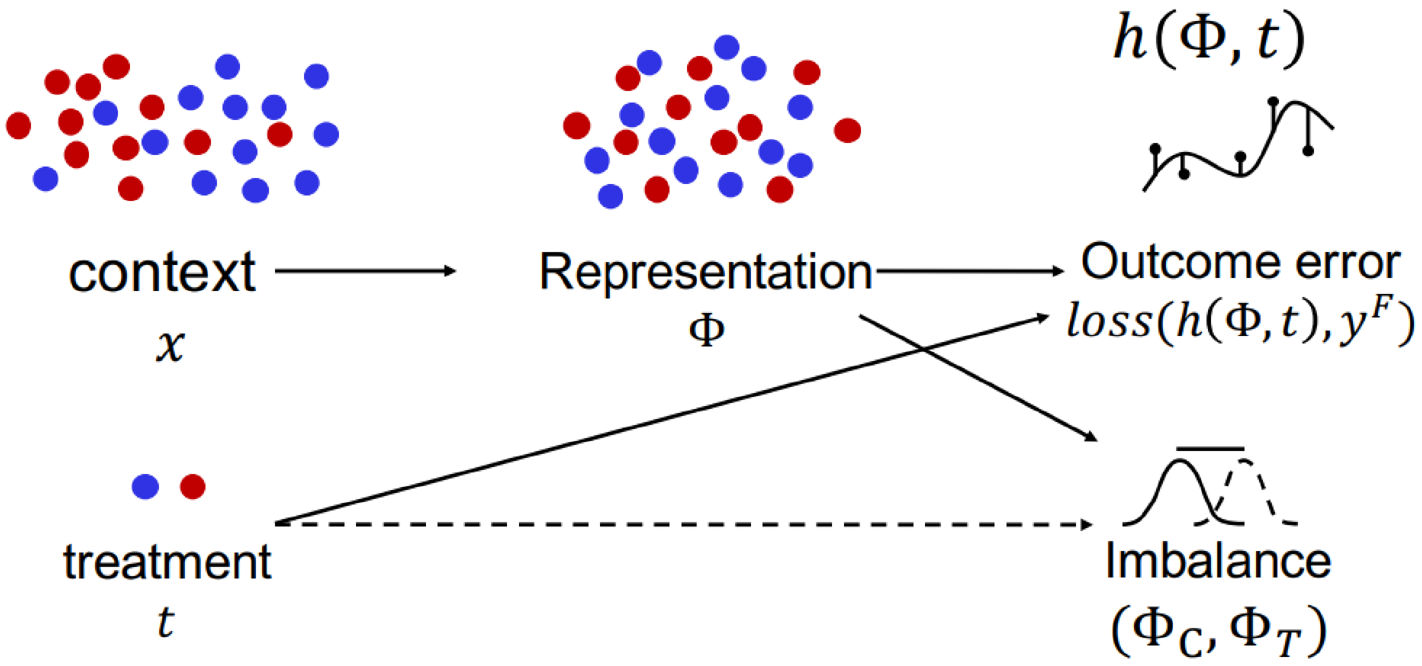


Figure:
Stefan Wager

[Wager et al. 2015]

Learning Balanced Representation

- "… prevent the learner from using 'unreliable' aspects of the data …"



[Johansson et al. 2016]

Counterfactual Risk Minimization

- Company X wants to improve its recommendation system.
  - Old distribution: $h_0(y|x_i)$
  - New distribution: $h(y|x_i)$
- Collected data: *(user, item, loss, prob)* $\Rightarrow (x_i, y_i, \delta_i, p_i)$

$$R(h) = \mathbb{E}_{x\sim p(x)}\mathbb{E}_{x\sim h(y|x)}[\delta(x,y)]$$

$$= \mathbb{E}_{x\sim p(x)}\mathbb{E}_{x\sim h_0(y|x)}\left[\delta(x,y)\frac{h(y|x)}{h_0(y|x)}\right]$$

[Swaminathan et al. 2015]

Counterfactual Risk Minimization

- Company X wants to improve its recommendation system.
  - Old distribution: $h_0(y|x_i)$
  - New distribution: $h(y|x_i)$
- Collected data: *(user, item, loss, prob)* $\Rightarrow (x_i, y_i, \delta_i, p_i)$

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{h(\mathrm{y}_i|x_i)}{p_i}$$

[Swaminathan et al. 2015]

Counterfactual Risk Minimization

- Degenerate results
- Unbounded variance
- Generalization error

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{h(\mathrm{y}_i|x_i)}{p_i}$$

[Swaminathan et al. 2015]

Counterfactual Risk Minimization

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{h(y_i|x_i)}{p_i}$$

- Degenerate results

  $\delta_i \in [-1,0]$

- Unbounded variance

  $$\hat{R}^M(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \min\left\{M, \frac{h(y_i \mid x_i)}{p_i}\right\}$$

- Generalization error

  $$\hat{R}^M(h) + \lambda\sqrt{\frac{\boldsymbol{Var}_h(u)}{n}}$$

[Swaminathan et al. 2015]

Counterfactual Risk Minimization

POEM Training Objective:

$$w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \, \overline{u_w} + \lambda \sqrt{\frac{\boldsymbol{Var}_w(u)}{n}},$$

$$u_w{}^i \equiv \delta_i \min\{M, \frac{\exp(w \cdot \phi(x_i, y_i))}{p_i \cdot \mathbb{Z}(x_i)}\}, \;\; \overline{u_w} \equiv \sum_{i=1}^{n} u_w{}^i / n,$$

$$\boldsymbol{Var}_w(u) \equiv \sum_{i=1}^{n} (u_w{}^i - \overline{u_w})^2 / (n-1).$$

[Swaminathan et al. 2015]

Counterfactual Risk Minimization

Proposition 1. *For any w_0,*

$$\sqrt{\boldsymbol{Var}_w(u)} \leq A_{w_0} \sum_{i=1}^{n} {u_w}^i + B_{w_0} \sum_{i=1}^{n} \{{u_w}^i\}^2 + C_{w_0}$$

$$= Q(w; w_0).$$

$$A_{w_0} \equiv -\overline{u_{w_0}} / \{(n-1)\sqrt{\boldsymbol{Var}_{w_0}(u)}\},$$

$$B_{w_0} \equiv 1/\{2(n-1)\sqrt{\boldsymbol{Var}_{w_0}(u)}\},$$

$$C_{w_0} \equiv \frac{n\{\overline{u_{w_0}}\}^2}{2(n-1)\sqrt{\boldsymbol{Var}_{w_0}(u)}} + \frac{\sqrt{\boldsymbol{Var}_{w_0}(u)}}{2}.$$

[Swaminathan et al. 2015]

Outline

- Definition
- Methods
- Misc

Question

- How to leverage cross validation?
  - Artificially introducing imbalance?

- Difference with Reinforcement Learning
  - ~~Off-policy~~ Offline policy

Misc

Education [edit]

Graduate courses on causal inference have been introduced to the curriculum of many schools.

- Saint Louis University, College of Public Health & Social Justice
- Carnegie Mellon University, Department of Philosophy
- Harvard University, School of Public Health
- Johns Hopkins University, Department of Computer Science
- Karolinska Institutet, Department of Medical Epidemiology and Biostatistics
- McGill University, Department of Epidemiology, Biostatistics and Occupational Health
- Northwestern University, Department of Sociology and Kellogg School of Management
- University of Pittsburgh, Department of Psychology in Education
- University of Groningen, Department of Statistics & Measurement Theory
- University of California, Los Angeles, Department of Epidemiology and Department of Computer Science
- University of California, Berkeley, School of Public Health
- University of Copenhagen, Department of Public Health
- University of Pennsylvania, Department of Biostatistics and Epidemiology
- The University of British Columbia, School of Population and Public Health
- Vanderbilt University, Department of Leadership, Policy, and Organizations
- Stevens Institute of Technology, Department of Computer Science [10]
- University of North Carolina at Chapel Hill, Department of Biostatistics [11]

Misc

- http://causality.cs.ucla.edu/blog/