



VHB ProDok – Machine Learning – Block III

## L.2: Causal Machine Learning

Stefan Lessmann

# Agenda



## **Introduction**

Correlation and causality

## **A Primer on Causal Inference**

Potential outcome framework, causal parameters, assumptions

## **Estimation of Conditional Average Treatment Effects**

Intuition, T-/S-/X-Learner, transformed outcomes, and beyond

## **Evaluation of Causal ML Models**

Preliminaries, uplift model evaluation

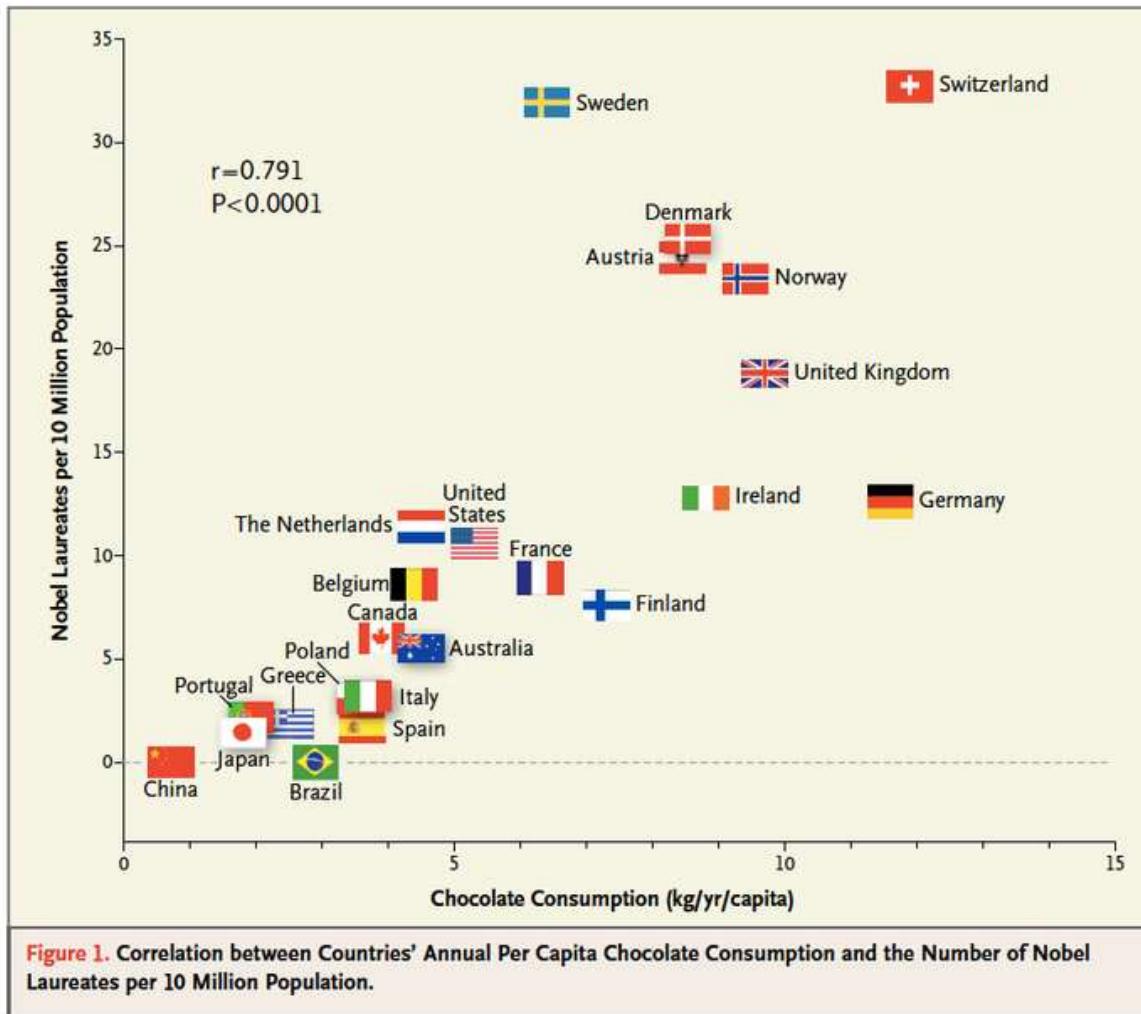
## **Summary**



# Introduction

## Causality and correlation

# Causality and Correlation Anecdotes



Source: <https://www.businessinsider.com/chocolate-consumption-vs-nobel-prizes-2014-4>

## Causality and Correlation Anecdotes

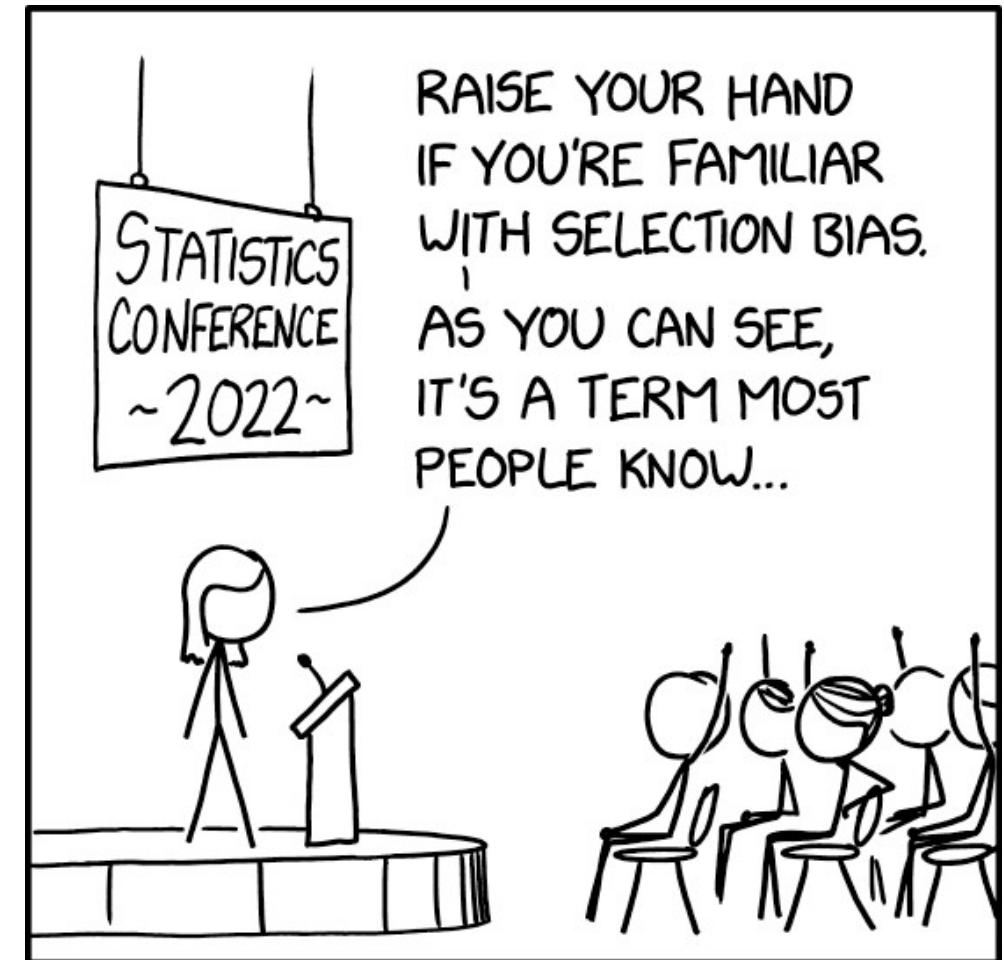
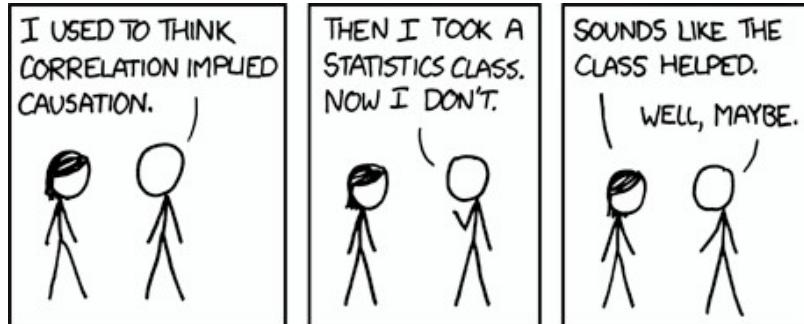


Image sources: <https://languagelog.ldc.upenn.edu/nll/?p=54685>; <https://plotlygraphs.medium.com/spurious-correlations-56752fcffb69>; <https://marketoonist.com/2014/04/big-data-analytics.html>

## Check Your Understanding

Business use case concerning the hotel industry

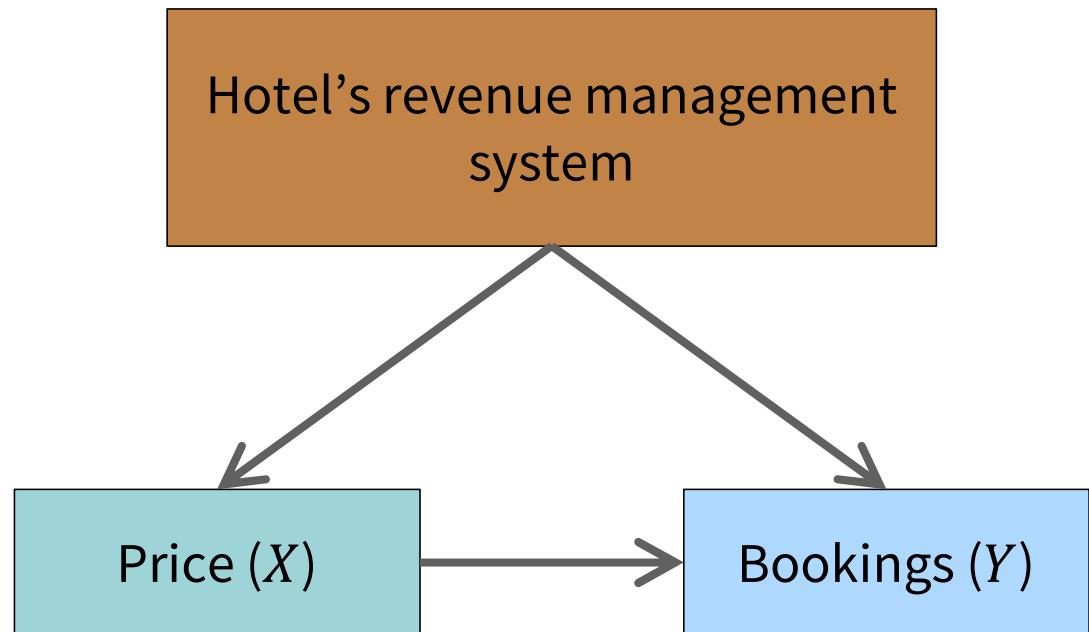
- You are the manager of *JustAsHome*, a five-star hotel in central Berlin
- You use machine learning to predict bookings from room rates using observed data from the last 6 years
- What dependency between prices and rates is the ML model likely to discover in your data?





## Causal Graph of the Hotel Industry Use Case

Past data is likely to show a positive correlation between prices and bookings. An ML model would capture this association and predict higher bookings (i.e., higher demand) for higher prices. The reason for this **spurious correlation** between prices ( $X$ ) and bookings ( $Y$ ) is that hotels increase prices in times of higher demand. This practice is known as revenue management. The causal graph identifies the revenue management system as a **confounder**, a factor that correlates with both, the features and the outcome.



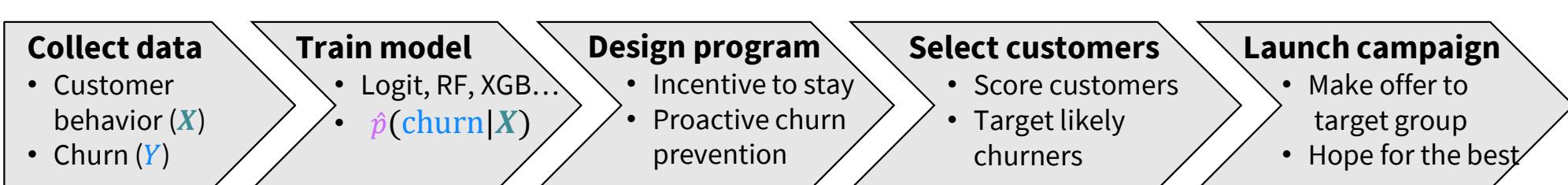
## Another Illustration of the Matter:

Predict customer churn to target retention offers

### ■ Objective: Prevent customer defection

- Cost of customer acquisition
- Value of long-term customers
  - Spend more
  - Transact more often
  - Less vulnerable to competitor offers
  - Positive word-of-mouth
  - ...

### ■ Churn modelling process



## Another Illustration of the Matter:

Predict customer churn to target retention offers

### ■ Objective: Prevent customer defection

- Cost of customer acquisition
- Value of long-term customers
  - Spend more
  - Transact more often
  - Less vulnerable to competitor offers
  - Positive word-of-mouth
  - ...

### ■ Churn modelling process

#### Collect data

- Customer behavior ( $\mathbf{X}$ )
- Churn ( $\mathbf{Y}$ )

#### Train model

- Logit, RF, XGB...
- $\hat{p}(\text{churn}|\mathbf{X})$

#### Design program

- Incentive to stay
- Proactive churn prevention

#### Select customers

- Score customers
- Target likely churners

#### Launch campaign

- Make offer to target group
- Hope for the best



# Treatment-Response Matrix with Four Stereotypes of Clients

## ■ Response/Conversion

- Any action of the client a marketer strives to trigger
- Stay, buy, give data, click, like, share, recommend, ...

## ■ Treatment

- Any form of marketing stimuli
- Bonus, email, phone call, coupon, catalog, ...

## ■ Efficient resource allocation

- Treat persuadables
- Do not treat any other client

## ■ Uplift modeling

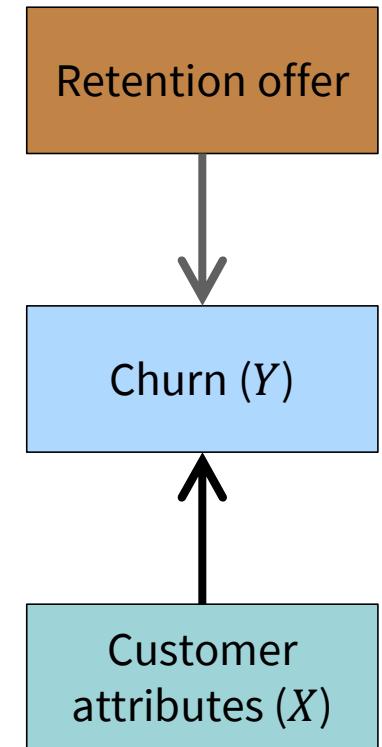
- Predict who is a persuadable
- Require data from a control group
- Challenging since group membership is **latent**

Respond  
when  
**TREATED**

|    |     | No                                 | Yes             |
|----|-----|------------------------------------|-----------------|
|    |     | Respond when<br><b>NOT TREATED</b> |                 |
| No | Yes | Persuadables                       | Sure Things     |
|    | No  | Lost causes                        | Do-not-disturbs |

# Causal Graph of the Churn Modeling Use Case

Customer attributes affect churn. This is what the prediction model learns. The offer also influences churn. This is the causal effect we care about. Being trained on past data, the churn model ignores previous offers made to customers in the past under some (nonrandom) targeting policy. For example, high risk customers may have been contracted in the past and persuaded to stay by a retention offer. The data will then include high risk clients that did not churn. This problem is a **collider bias**, a specific form of **selection bias**, that also causes spurious correlations in the observed data. This bias arises from conditioning on a variable (churn) that is influenced by multiple independent causes.

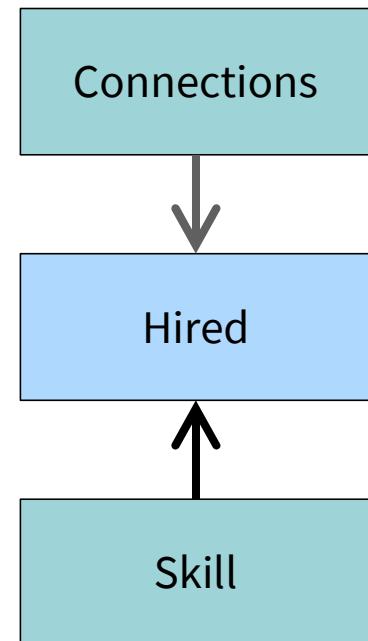


## Another, Illustrative Example of Collider Bias

■ **Definition:** Collider bias occurs when conditioning on a common effect (a "collider") of two independent causes creates a spurious correlation between them.

■ **Example: Collider bias in hiring decisions**

- Imagine a company hires based on either skill or connections
- If we only look at hired candidates, then skill and connections become negatively correlated (because if a candidate was hired without strong skill, they must have had good connections.)
- In reality, skill and connections are independent (or perhaps positively related)
- Conditioning on the hiring decision (a collider) creates a false correlation.
- This bias arises from conditioning on a variable that is influenced by multiple independent causes.
- By conditioning, we mean that we perform the analysis only among hired people and, in this sense, condition on the variable Hired



# When Prediction is Useful (and when it is not)

Athey (2017) and Varian (2014)

- **Supervised ML is about prediction: map feature values to estimates of the target**
- **Accurate prediction is useful in many application settings**

- May solve a problem: image classification, sentiment analysis, fraud screening, ...
- May provide insight into a problem (e.g., associations): predictive maintenance, buyer behavior, ...

- **Prediction models do not answer causal questions**

- Decision-making often involves cause-effect relationships
- In the churn example, the decision whom to target required knowledge of the treatment's effect
- In the hotel example, understanding the true effect of prices on demand would require accounting for other factors that affect both, prices and demand
- In both examples, the available data to train an ML model was not representative of the true cause-effect relationship due to spurious correlation
- Supervised ML can never work (well) when the training data is not representative



# A Primer on Causal Inference

Potential outcome framework, causal parameters, assumptions

## Fundamental Problem of Causal Inference

You can never observe all relevant outcomes



# Different Perspectives and Cultures in Data-Oriented Disciplines

## Athey and Imbens (2019)

### Machine learning

#### ■ Focus on prediction

- Map from covariates to outcomes
- Cope with **high dimensionality**

#### ■ Practices to test predictive quality

- Cross-validation
- Model selection (e.g., regularization)

#### ■ Few assumptions

- Independent observations
- Stability of the joint distribution of  $(Y, X)$

#### ■ More data-driven

### Econometrics

#### ■ Understand structural properties

- Estimate parameters of interest
- Cause-effect relationships are one example

#### ■ Apply linear models to all data

- Explanatory modeling
- Data is typically **low-dimensional**

#### ■ Key interests

- Unbiasedness of parameters
- Efficiency and convergence rates

#### ■ Many assumption

#### ■ More theory-driven

# A Primer on Causal Inference

## The potential outcome or counterfactual framework

■ We assume a **superpopulation**  $\mathcal{D}$  of which a realization of  $N$  independent random

Variables  $(Y_i, X_i, D_i) \sim \mathcal{D}$  are given as training data, with

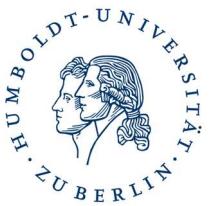
- $m$ -dimensional feature vector  $X_i \in \mathbb{R}^m$
- outcome variable  $Y_i \in \mathbb{R}$
- treatment assignment indicator  $D_i \in \{0,1\}$

■ The realized (observed) outcome for unit  $i$  is

$$Y_i = Y_i(D_i) = \begin{cases} Y_i^{(0)} & \text{if } D_i = 0 \\ Y_i^{(1)} & \text{if } D_i = 1 \end{cases} = Y_i^{(1)}D_i + (1 - D_i)Y_i^{(0)}$$

■ Where  $Y_i^{(0)}$  and  $Y_i^{(1)}$  are the **potential outcomes** (Imbens & Rubin, 2015)

■ Fundamental problem of causal inference: for an individual unit, we observe either  $Y_i^{(0)}$  or  $Y_i^{(1)}$  and never the counterfactual outcome



# A Primer on Causal Inference

## Basic identity of causal inference

$$\begin{aligned}\text{Treatment effect} &= \text{Outcome for treated} - \text{Outcome for untreated} \\ &= [\text{Outcome for treated} - \text{Outcome for treated if not treated}] + \\ &\quad [\text{Outcome for treated if not treated} - \text{Outcome for untreated}] \\ &= [\text{Impact of treatment on treated} + \text{selection bias}]\end{aligned}$$

Comparison of the actual outcome (what happens to the treated) compared with the counterfactual (what would have happened if they had not been treated) is instrumental to causal inference.

# A Primer on Causal Inference

## Parameters of interest (selection)

### ■ Conditional average treatment effect (CATE)

$$\tau^{CATE} = \tau(x) = \mathbb{E}\left(Y_i^{(1)} - Y_i^{(0)} | X_i = x\right)$$

### ■ Average treatment effect (ATE)

$$\tau^{ATE} = \mathbb{E}\left(Y_i^{(1)} - Y_i^{(0)}\right) = \mathbb{E}(\tau(x))$$

### ■ Average treatment effect on the treated (ATT)

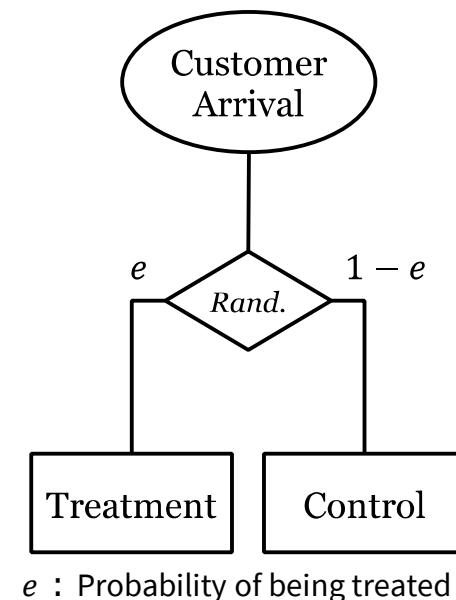
$$\tau^{AT} = \mathbb{E}\left(Y_i^{(1)} - Y_i^{(0)} | D_i = 1\right)$$

# A Primer on Causal Inference

## Study Design and Data Collection

### ■ Randomized controlled trial (RCT)

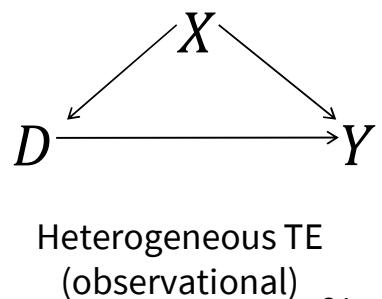
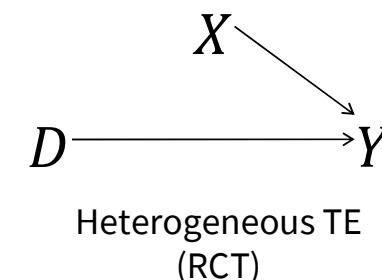
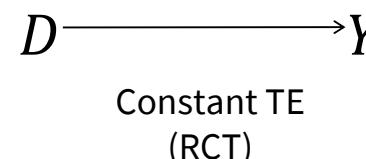
- Groups of subjects who receive different **treatments**
- Assignment into groups is **random**
- Widely known from clinical trials and A/B testing in marketing
- Methodologically the gold standard, but...
  - Experiments are costly
  - Ethical concerns due to withholding treatment



$e$  : Probability of being treated

### ■ Observational studies

- Treatment assignment is **non-random**
- Risk of bias due to confounders
- Risk of bias due to colliders
- Other forms of selection bias (e.g., self-selection)



## Assumptions

Ensure identifiability of causal parameters

### ■ Stable Unit Treatment Value Assumption (SUTVA)

$$Y_i = Y_i^{(0)} + D_i (Y_i^{(1)} - Y_i^{(0)})$$

- No interference between observations (e.g., network effects) or hidden variation in treatment
- Often taken for granted w/o careful examination
- Could even be violated in an RCT

### ■ Ignorability Assumption

$$(Y_i^{(1)}, Y_i^{(0)}) \perp\!\!\!\perp D_i | X_i$$

- Potential outcomes are independent of treatment conditional on covariates
- Idea is that we can control for the effect of covariates (i.e., observed confounders)
- Also called selection on observables, uncounfoundedness, or conditional independence

### ■ Overlap assumption

$$0 < P(D = 1 | X = x) < 1 \quad \forall x \in \text{supp}(X)$$

- We need people in both, the treatment and control group for estimation
- Boundaries of, e.g., 5% to 95% or 10% to 90% typically sufficient in practice

# Estimating Causal Effects in a Randomized Controlled Trial

## Gold standard of causal inference

$$\begin{aligned}
 \tau^{ATE} &= \mathbb{E}(\tau_i) \\
 &= \mathbb{E}\left(Y_i^{(1)} - Y_i^{(0)}\right) \\
 &= \mathbb{E}\left(Y_i^{(1)}\right) - \mathbb{E}\left(Y_i^{(0)}\right) \\
 &= \mathbb{E}\left(Y_i^{(1)}|D_i = 1\right) - \mathbb{E}\left(Y_i^{(0)}|D_i = 0\right) \\
 &= \mathbb{E}(Y_i | D_i = 1) - \mathbb{E}(Y_i | D_i = 0)
 \end{aligned}$$

**In an RCT, we can replace potential outcomes with observed outcomes.**

due to linearity in expectations

due to random assignment

due to ignorability assumption

# Estimating Causal Effects in a Randomized Controlled Trial

## Gold standard of causal inference

$$\begin{aligned}
 \tau^{ATE} &= \mathbb{E}(\tau_i) \\
 &= \mathbb{E}\left(Y_i^{(1)} - Y_i^{(0)}\right) \\
 &= \mathbb{E}\left(Y_i^{(1)}\right) - \mathbb{E}\left(Y_i^{(0)}\right) \\
 &= \mathbb{E}\left(Y_i^{(1)}|D_i = 1\right) - \mathbb{E}\left(Y_i^{(0)}|D_i = 0\right) \\
 &= \mathbb{E}(Y_i | D_i = 1) - \mathbb{E}(Y_i | D_i = 0)
 \end{aligned}$$

**In an RCT, we can replace potential outcomes with observed outcomes.**

due to linearity in expectations

due to random assignment

due to ignorability assumption

Treatment effect = **Outcome for treated** – **Outcome for untreated**

$$\begin{aligned}
 &= [\text{Outcome for treated} - \text{Outcome for treated if not treated}] + \\
 &= [\text{Outcome for treated if not treated} - \text{Outcome for untreated}] \\
 &= [\text{Impact of treatment on treated} + \text{selection bias}]
 \end{aligned}$$

**An RCT does not suffer from a selection bias (e.g., self-selection).**

# Observational Studies w/o Random Treatment Assignment

$$\begin{aligned}
 \tau^{ATE} &= \mathbb{E}(\tau_i) \\
 &= \mathbb{E}\left(Y_i^{(1)} - Y_i^{(0)}\right) \\
 &= \mathbb{E}\left(Y_i^{(1)}\right) - \mathbb{E}\left(Y_i^{(0)}\right) \\
 &\neq \mathbb{E}\left(Y_i^{(1)}|D_i = 1\right) - \mathbb{E}\left(Y_i^{(0)}|D_i = 0\right)
 \end{aligned}$$

due to linearity in expectations

Due to observed or unobserved covariates

$$\begin{aligned}
 \text{Treatment effect} &= \text{Outcome for treated} - \text{Outcome for untreated} \\
 &= [\text{Outcome for treated} - \text{Outcome for treated if not treated}] + \\
 &= [\text{Outcome for treated if not treated} - \text{Outcome for untreated}] \\
 &= [\text{Impact of treatment on treated} + \boxed{\text{selection bias}}]
 \end{aligned}$$



# Estimation of Conditional Average Treatment Effects

## Intuition, T-/S-/X-Learner, transformed outcomes, and beyond

# Some Definition and Notation

## ■ Conditional treatment probability or **propensity score** $e(x) := P(D_i = 1|X_i = x)$

## ■ Conditional response functions

- Conditional expectation of the outcome  $\mu(x) = \mathbb{E}(Y_i|X_i = x)$
- Conditional expectation of the outcome in one treatment arm  $\mu(d, x) = \mathbb{E}(Y_i|D_i = d, X_i = x)$

## ■ Propensity score & conditional response functions also called **nuisance parameter**

- Can be estimated from data  $(Y_i, X_i, D_i), i = 1, \dots, N$
- We can use a classification model to obtain an estimate of the propensity score  $\hat{e}(x)$
- Depending on  $Y$ , we can use regression or classification to estimate the conditional expectation  $\hat{\mu}$

## ■ Let $M$ denote some ML algorithm (neural network, tree-learner, etc.)

- Notation for training a model to predict  $Y$  from  $X$ :  $M(Y \sim X)$
- For example, ML-based estimate of the propensity score:  $\hat{e}(x) = M(D \sim X)$

**Recall TE = Impact of treatment on treated + selection bias**

Predict outcomes for treatment group as if they had not been treated

## ■ Outcome models

- Estimate conditional outcomes among control units  $\mu(0, x) = \mathbb{E}(Y_i | D_i = 0, X_i = x)$
- Use the estimate  $\hat{\mu}(0, x)$  to predict outcomes for treatment group units
- But treatment / control group are different in terms of observables

## ■ Weighting and matching approaches

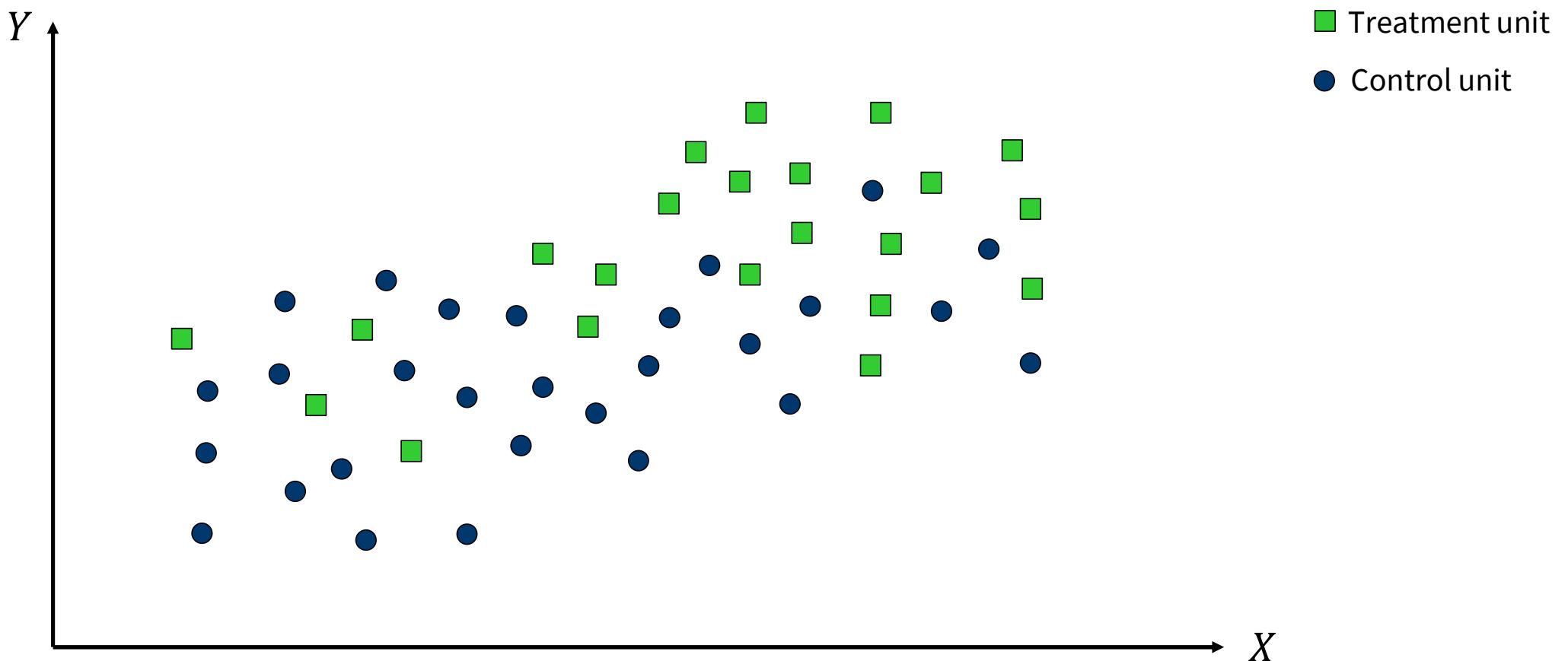
- Assignment is random conditional on  $X$  (due to ignorability assumption)
- Estimate propensity score  $e(x) = \mathbb{E}(D_i = 1 | X_i = x)$
- Use that estimate  $\hat{e}(x)$  to reweight control group to look like treatment group in terms of  $X$

## ■ Doubly robustness

- The above approaches fail if models of propensity score or outcome are incorrect
- Combine both techniques
- Doubly robust estimators “work” as long as one of the two models is correctly specified

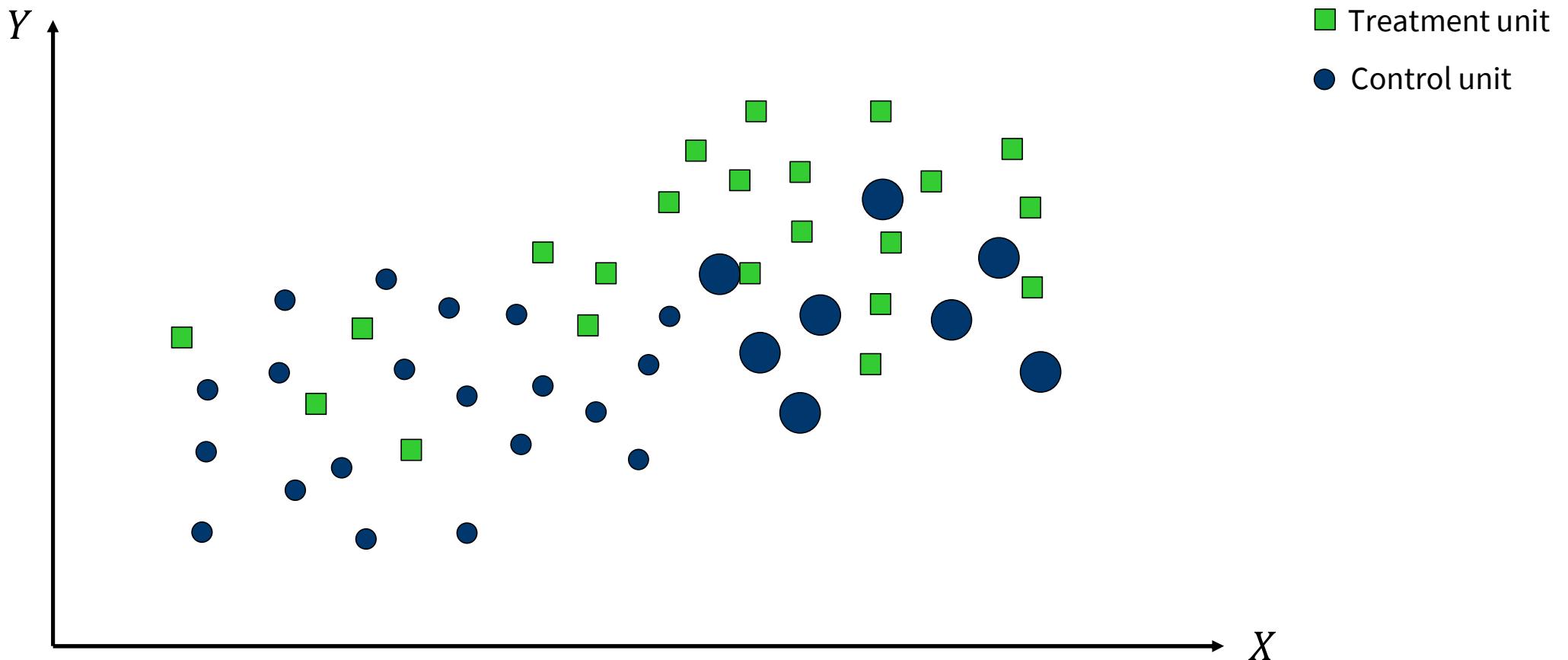
## Intuition

Treated units have higher  $X$  on average



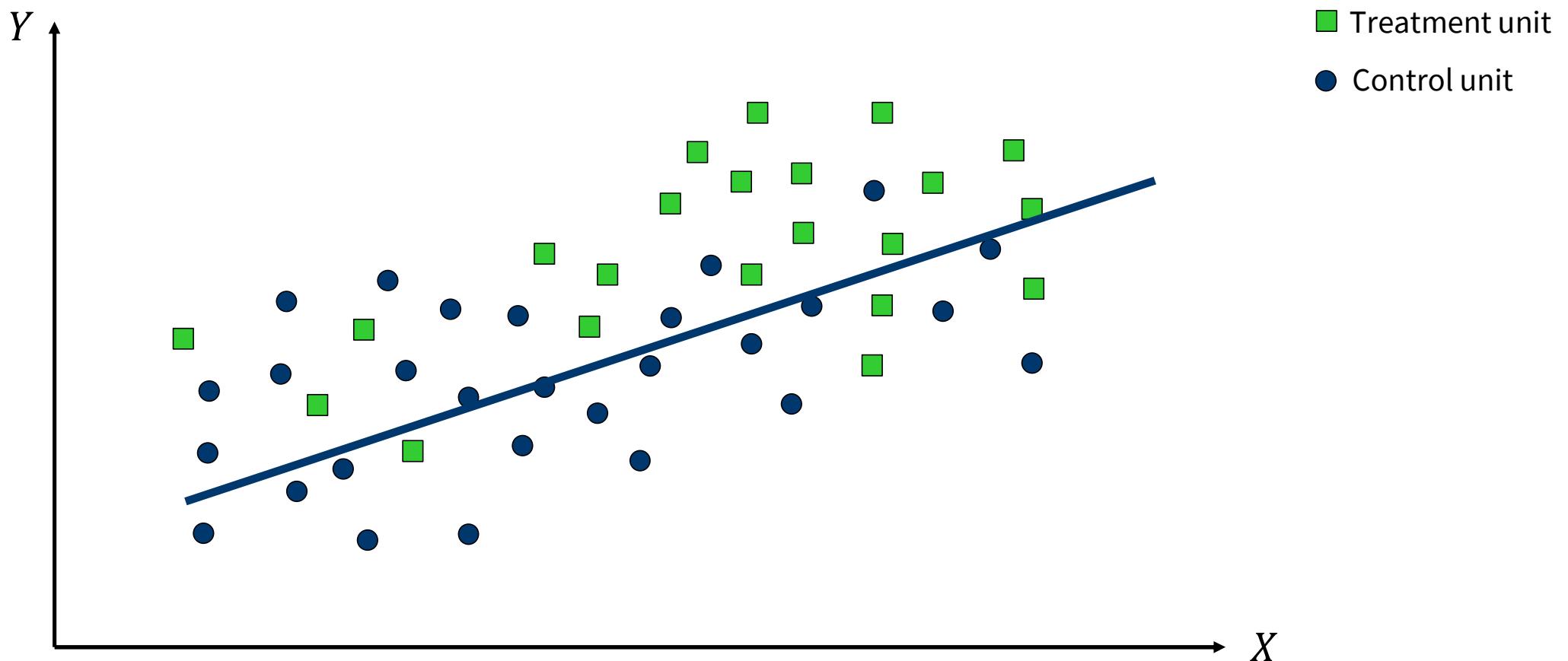
# Intuition

Reweighting control units with high  $X$  to adjust for the difference



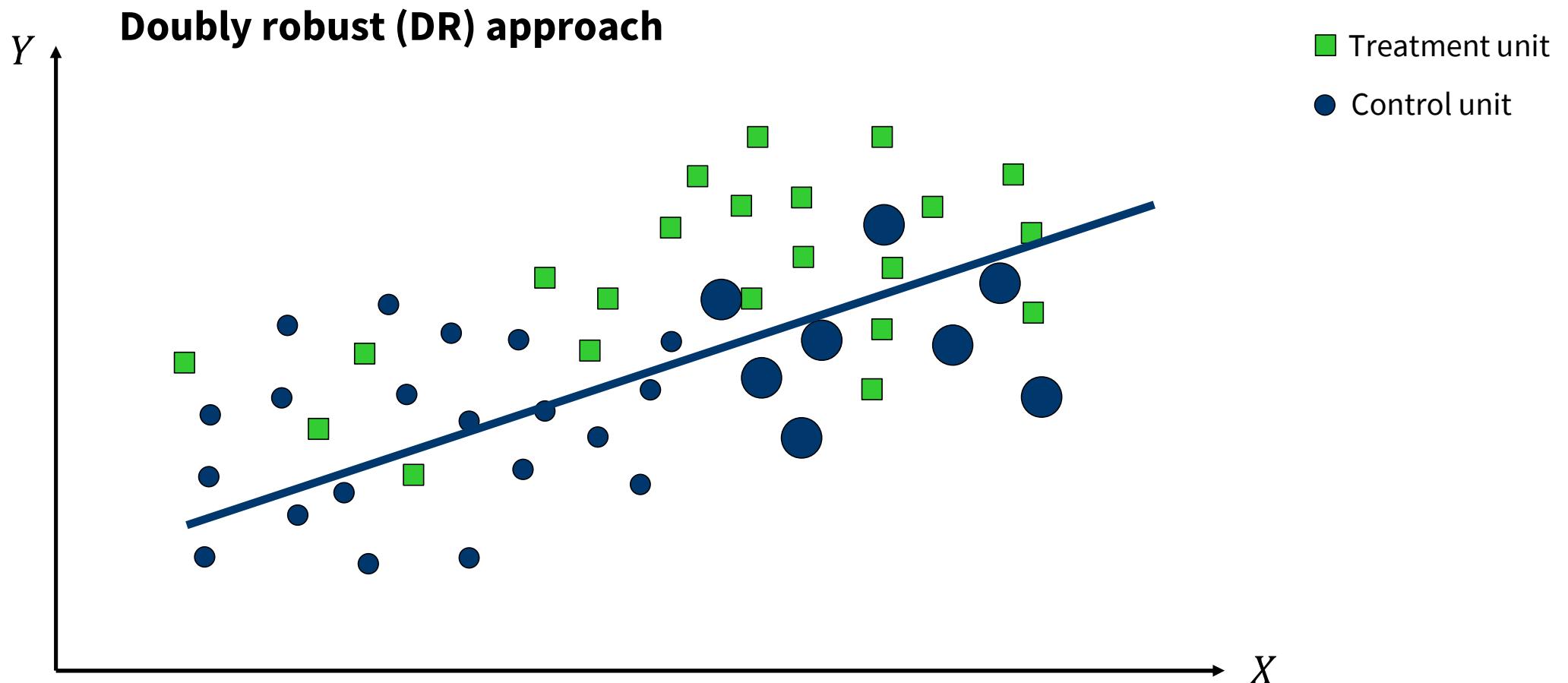
# Intuition

Outcome models adjust for differences in  $X$



## Intuition

Reweighting control units with high  $X$  AND using outcome model to adjust



## Estimating CATE: Two-Model Approach or T-Learner

Künzel et al. (2019), Knaus et al. (2018), Athey & Imbens (2015)

### ■ Two-Step approach

- Estimate conditional response functions among treated and untreated units separately
- Estimate CATE as the difference of the two estimates

### ■ More formally

- Control response function  $\mu_0 = \mu(0, x) = \mathbb{E}(Y_i | X_i = x, D_i = 0)$
- $\hat{\mu}_0 = M_1(Y_i^0 \sim X_i^0)$  where  $(Y_i^0, X_i^0) = \{(Y_i, X_i) | D_i = 0\}$
- Treatment response function  $\mu_1 = \mu(1, x) = \mathbb{E}(Y_i | X_i = x, D_i = 1)$
- $\hat{\mu}_1 = M_2(Y_i^1 \sim X_i^1)$  where  $(Y_i^1, X_i^1) = \{(Y_i, X_i) | D_i = 1\}$
- Estimate CATE by  $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

### ■ Caveats

- High variance of the estimator due to individual optimization of the models for outcome prediction
- See Künzel et al. (2019) for an illustration and empirical results

## Estimating CATE: Single-Model Approach or S-Learner

Künzel et al. (2019), Knaus et al. (2018), Athey & Imbens (2015)

### ■ Estimate response conditional on treatment and features

- Include treatment indicator among the other features
- Train one model on the resulting data set

### ■ Estimate CATE as in the T-learner approach

### ■ More formally

- Estimate conditional response  $\mu(d, x) = \mathbb{E}(Y_i | D_i = d, X_i = x)$
- $\hat{\mu} = M_1(Y_i \sim X'_i)$  where  $X' = X \cup D$
- Estimate CATE by  $\hat{\tau}(x) = \hat{\mu}(1, x) - \hat{\mu}(0, x)$

# Estimating CATE: The X-Learner

Künzel et al. (2019)

## ■ Estimate conditional response functions $\mu(d, x) = \mathbb{E}(Y_i | D_i = d, X_i = x)$

- Treatment outcome estimator  $\hat{\mu}_1 = M_2(Y^1 \sim X^1)$
- Control outcome estimator  $\hat{\mu}_0 = M_1(Y^0 \sim X^0)$

## ■ Compute imputed treatment effects $\tilde{\tau}_i^D$

- For units in the treatment group:  $\tilde{\tau}_i^1 := Y_i^{(1)} - \hat{\mu}_0(X_i)$
- For units in the control group  $\tilde{\tau}_i^0 := \hat{\mu}_1(X_i) - Y_i^{(0)}$

## ■ Use imputed treatment effects as targets in another set of ML models

- Treatment group model  $\hat{t}_1 = M_3(\tilde{\tau}_i^1 \sim X^1)$
- Control group model  $\hat{t}_0 = M_4(\tilde{\tau}_i^0 \sim X^0)$

## ■ Estimate CATE as weighted average $\hat{\tau}(x) = g(x)\hat{t}_0(x) + (1 - g(x))\hat{t}_1(x)$

- Where  $g \in [0,1]$  is a weighting function
- Künzel et al. (2019) recommend using the propensity score

## Estimating CATE: Modified Outcome Methods

Knaus et al. (2018), Athey & Imbens (2015)

### ■ Apply standard ML algorithm to predict a transformed outcome

- Idea is to transform the outcome such that  $\mathbb{E}(Y_i^*|X_i = x) = \tau(x)$  with  $Y_i^*$  the transformed outcome
- Appealing since equivalent to standard ML → predict a target from features

### ■ Inverse probability weighting (IPW)

$$Y_{i,IPW}^* = Y_i \frac{D_i - e(X_i)}{e(X_i)(1 - e(X_i))}$$

### ■ Doubly robust estimator

$$Y_{i,DR}^* = \mu(1, X_i) - \mu(0, X_i) + \frac{D_i(Y_i - \mu(1, X_i))}{e(X_i)} - \frac{(1 - D_i)(Y_i - \mu(0, X_i))}{(1 - e(X_i))}$$

### ■ In practice, the nuisance parameters $\mu(D_i, X_i)$ and $e(X_i)$ are unknown

- Step 1: estimate nuisance parameters and plug them into the above equations to calculate  $Y_i^*$
- Step 2:  $\hat{\tau}(x) = M_1(Y_i^* \sim X_i)$

# Estimating CATE: Nascent Field of Research

Many more interesting approaches

## ■ Random forest-based approaches

- A causal tree splits nodes to maximizes outcome differences between treatment and control units
- Generalizations to forests (Athey et al., 2019; Oprescu et al., 2019; Wager & Athey, 2018)

## ■ Causal boosting (Powers et al. 2018)

## ■ Causal KNN (Hitsch & Misra, 2018)

## ■ Bayesian approaches

- Generalizations of Bayesian additive regression trees (BART)
- (Hahn et al., 2019; Hill, 2011)

## ■ Neural network-based approaches

- Residual networks, multi-task learning architectures, ...
- (Atan et al., 2018; Farrell et al., 2021; Hartford et al., 2017; Küngel et al., 2018)

## Estimating CATE: Concluding remarks

### ■ Main motivation for understanding treatment effect heterogeneity

- Develop optimal policies
- Assignment functions that map from the observable covariates to treatment policy assignments
- Simple policy rule: obtain an estimate of  $\hat{\tau}$  and assign treatment whenever  $\hat{\tau}(x) > 0$

### ■ Can we use ML to learn the policy?

- Yes we can!
- That is one of the key use cases of reinforcement learning

### ■ So maybe we need another course on reinforcement learning

Reinforcement Learning with Online Interactions



Offline Reinforcement Learning





# Evaluation of Causal ML Models

## Preliminaries, uplift model evaluation

## Preliminaries

The evaluation of causal (ML) models is challenging

### ■ Focus of the evaluation

- Quality of model-based estimates
- Conditional outcomes versus causal parameters

### ■ Consider standard performance indicators (e.g., MSE)

- Simple comparison of predictions to actuals **per unit**
- Impossible due to fundamental problem of causal inference
  - One unit (e.g., customer) cannot be part of both, treatment and control group
  - Counterfactual outcome is unobservable

$$MSE_{CATE} = \frac{1}{n} \sum_{i=1}^n (\tau_i(X_i) - \hat{\tau}_i(X_i))^2, \text{ with } \tau_i(X_i) = \mathbb{E}(Y_i^{(1)} - Y_i^{(0)} | X_i)$$

## Approaches Toward Evaluation

No perfect solution

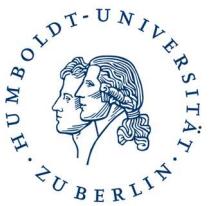
### ■ Simulation study

- Generate synthetic data (e.g., Knaus et al. 2021)
- Treatment effect and potential outcomes are known. All problems solved!
- Very useful but probably not enough in applied research

### ■ Use transformed outcome for performance evaluation

- Recall that  $\mathbb{E}(Y_i^* | X_i = x) = \tau(x)$
- Plausible to calculate, e.g., MSE, as  $MSE_{TO} = \frac{1}{n} \sum_{i=1}^n (Y_i^* - \hat{\tau}_i(X_i))^2$

### ■ Application-specific evaluation strategies



## Uplift Model Evaluation

Modeling goal is to decide who should receive treatment

### ■ Two types of comparisons needed

- Predictions to actuals
- Treatment to control group

### ■ Done for **segments** (as opposed to cases)

### ■ Rank subject by model-estimated CATE $\hat{\tau}(x)$

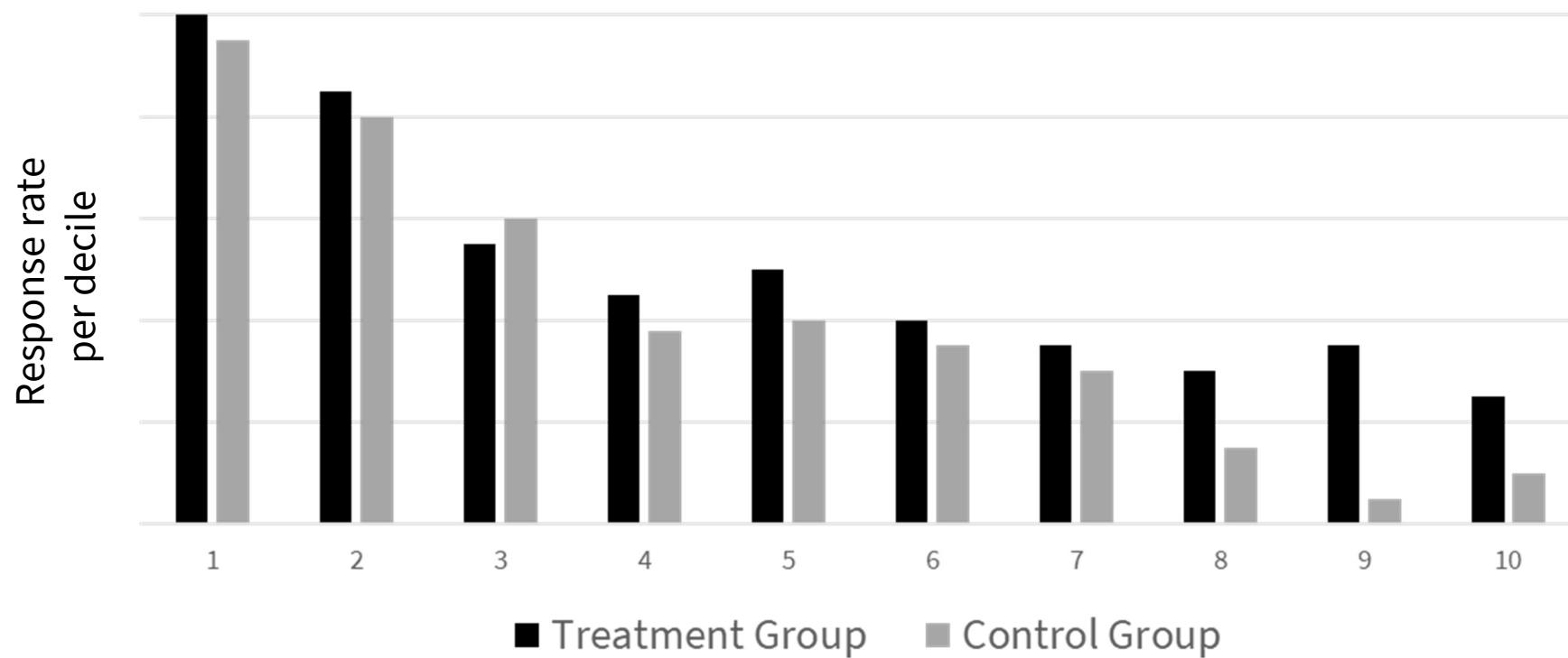
- Analysis per segment (e.g., decile) mimics target marketing setting
- Essentially difference between treatment / control group lift curves

### ■ Train-test set splitting or cross-validation still important

- Every fold contains cases from both, the treatment and control group
- This allows calculating a fold-wise uplift score

# Uplift Model Evaluation

## Uplift curve



# Uplift Model Evaluation

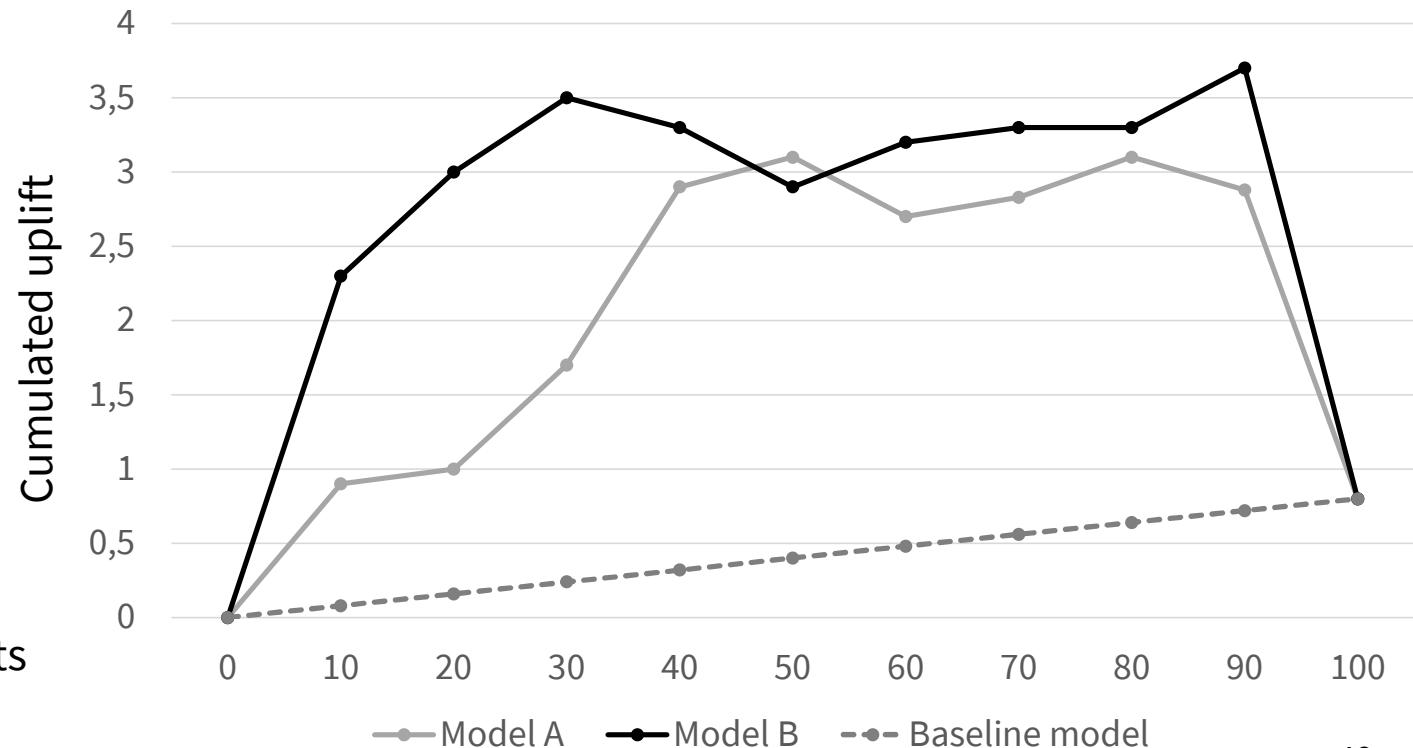
Cumulative uplift or Qini curve (Radcliffe, 2007)

- Adaption of a gain chart for uplift models
- Test set cases sorted in descending order by model estimates (i.e., CATE)
- Calculate cumulated uplift

- Treatment group response
- Control group response
- Absolute or relative uplift
- May need adjustment if the no. of treatment/control group cases differ

## ■ Interpretation

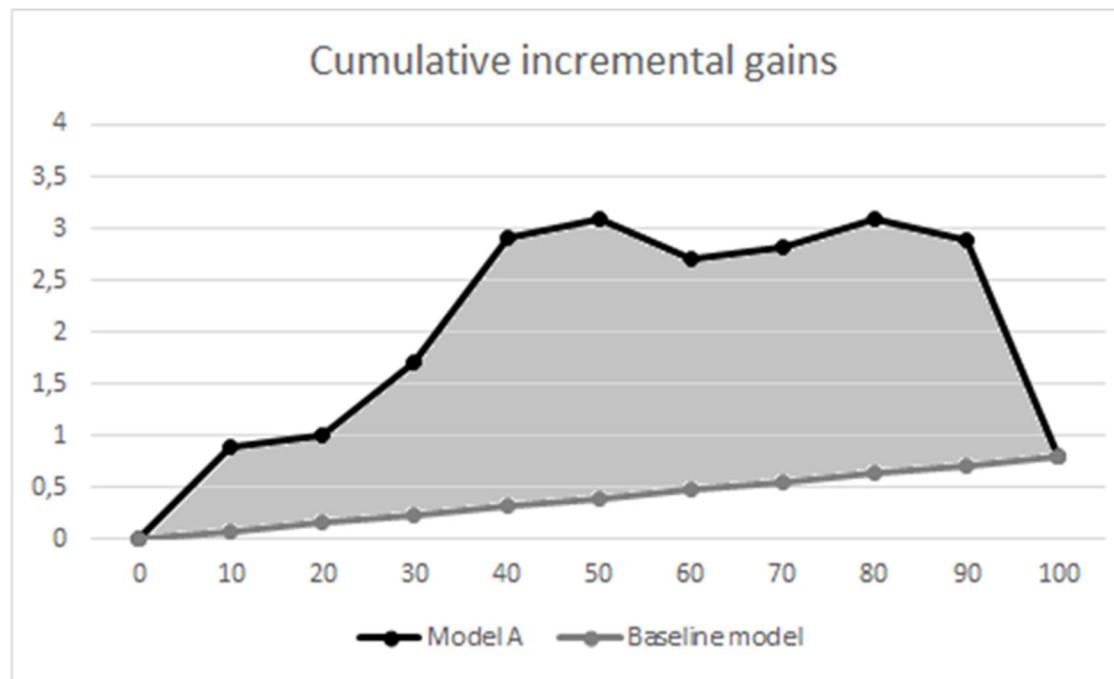
- Net benefit of the treatment
- When targeting n % of the clients



## Qini Coefficient

Summarizes the Qini curve in a single number

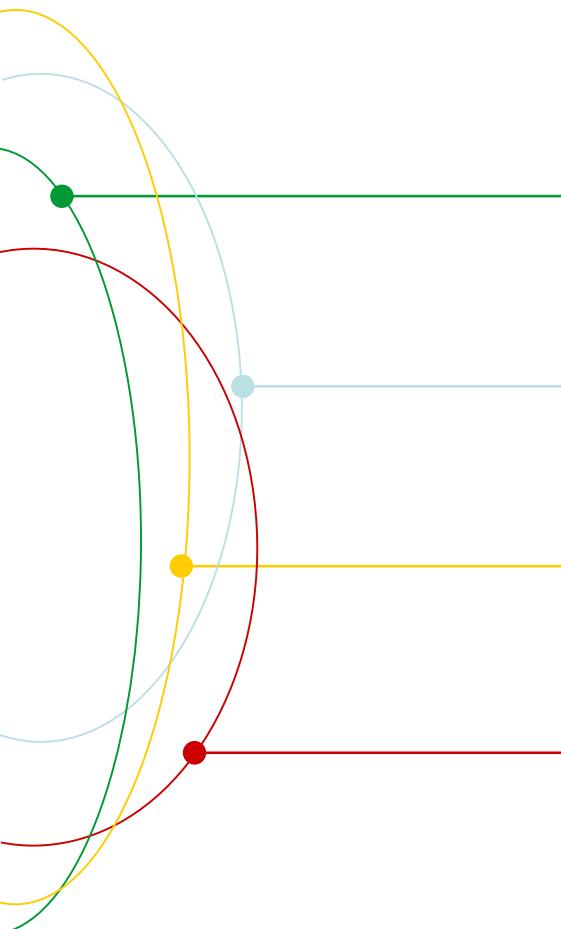
- Scalar measures of model performance useful for comparisons
- Qini-coefficient defined as the area under the model's Qini curve minus that of the random baseline model





# Summary

# Summary



## Learning goals

- Potential outcome framework
- Causal ML for treatment effect estimation
- Uplift models as specific use case of causal ML



## Findings

- Heterogeneous TE as causal parameter
- High relevance for ECON (e.g., resource allocation)
- Data gathering crucial and challenging
- Use ML to estimate nuisance functions
- Generic model evaluation is difficult. Solutions exists for specific applications (e.g. Uplift)



## What next

- Self-study: uplift models for EC analytics
- Discussion of the Assignment

# Literature



- Atan, O., Jordon, J., & van der Schaar, M. (2018). *Deep-Treat: Learning Optimal Personalized Treatments from Observational Data using Neural Networks*. Association for the Advancement of Artificial Intelligence.
- Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324), 483-485.
- Athey, S., & Imbens, G. (2019). Machine Learning Methods Economists Should Know About. *CoRR*, arXiv:1903.10075v1.
- Athey, S., & Imbens, G. W. (2015). Machine Learning for Estimating Heterogeneous Causal Effects. Stanford University, Graduate School of Business. [Retrieved from <https://EconPapers.repec.org/RePEc:ecl:stabus:3350>]
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148-1178.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
- Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econmtrica* 82(1), 181-213.
- Hahn, P. R., Murray, J. S., & Carvalho, C. (2019). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Archive Preprint*, arXiv:1706.09523v4.
- Hartford, J. S., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). *Deep IV: A Flexible Approach for Counterfactual Prediction*. Proceedings of the 34th International Conference on Machine Learning (ICML'2017), Sydney, Australia.
- Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1), 217-240.
- Hitsch, G. J., & Misra, S. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *SSRN*. [Retrieved from: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3111957](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3111957) ]
- Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

## Literature



- Kane, K., Lo, V. S., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, 2(4), 218-238.
- Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *The Econometrics Journal*, 24(1), 134-161.
- Küngel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10), 4156-4165.
- Küngel, S. R., Stadie, B. C., Vemuri, N., Ramakrishnan, V., Sekhon, J. S., & Abbeel, P. (2018). Transfer Learning for Estimating Causal Effects using Neural Networks. *ArXiv preprint*, arXiv:1808.07804v1.
- Lo, V. S. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2), 78-86.
- Oprescu, M., Syrgkanis, V., & Wu, Z. S. (2019). Orthogonal Random Forest for Causal Inference. In K. Chaudhuri & R. Salakhutdinov (Eds.). *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 4932-4941.
- Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11), 1767-1787.
- Radcliffe, N. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal*, 3, 14-21.
- Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 322-331.
- Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3-28.
- Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27), 7310-7315.
- Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

# Thank you for your attention!

Stefan Lessmann

Chair of Information Systems  
School of Business and Economics  
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742  
Fax. +49.30.2093.5741

[stefan.lessmann@hu-berlin.de](mailto:stefan.lessmann@hu-berlin.de)  
<http://bit.ly/hu-wi>

[www.hu-berlin.de](http://www.hu-berlin.de)

