Causal Machine Learning with DoubleML

Prof. Dr. Martin Spindler, Universität Hamburg & Economic AI

ARIC Lunch Seminar

19.9.2023

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Causal Inference based on ML

Part I: Introduction to Causal Machine Learning

DoubleML

ECONOMIC AI

Causal Machine Learning

Why? 〉〉 **Causal Machine Learning** 〈〈 AI

Causal Modeling

- Learning causal relationships
- Going beyond correlations
- Pioneers: Pearl, Rubin, Imbens (Nobel Prize 2021)

Machine Learning

- Learning complex patterns in data
- Correlation based
- Good at forecasting / prediction

Predictive vs. Causal ML

ECONOMIC AI

Predictive ML	Causal ML
How can we build a good prediction rule, $\boldsymbol{f}(\boldsymbol{X})$, that uses features $\boldsymbol{X}$ to predict? **Example:** Customer Churn *"How well can we predict whether customers churn?"*	What is the causal effect of a treatment $\boldsymbol{D}$ on an outcome $\boldsymbol{Y}$? **Example:** Customer Retention *"Why do customer churn?“* *“How can we retain customers?”*

Causal Inference: Overview

ECONOMIC AI

Resource Allocation

Customer Journey

Dynamic Pricing

Causal Understanding Is Key for "Real Intelligence"

Customer Churn

Targeted Marketing

Returns in Retail

Methods

ECONOMIC AI

A/B-Testing / Experimentation

- Getting more popular with tech companies
- Control for covariates to improve precision
- Heterogenous treatment effects

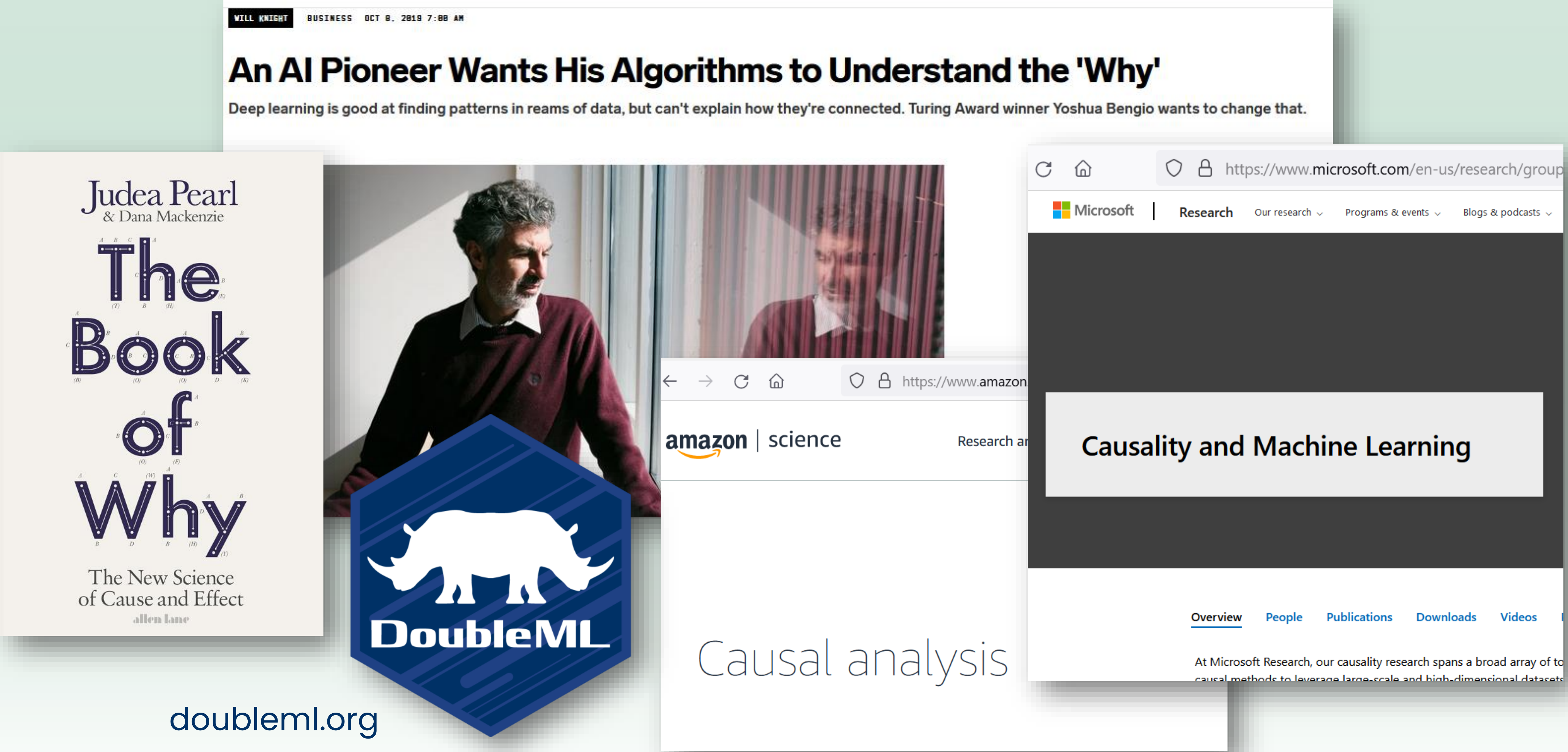
Structural Economic Models

- Allow for policy evaluation
- Based on economic principles of rational behavior, incentives, etc.
- Allow for competition, strategic effects, etc.

Hybrid Methods: Instrumental Variables

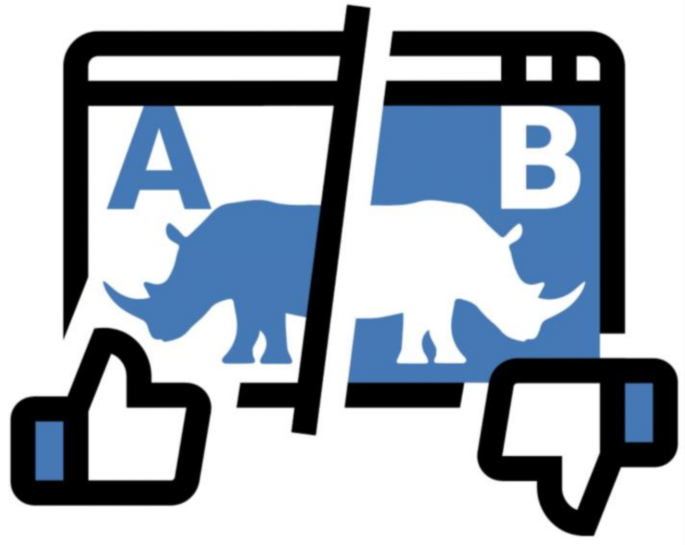
- Invented in economics but has become popular more broadly
- Used in settings when Randomized Trials are not feasible and/or new policies policy predictions are needed
- Can handle large dimensions with solid statistical properties

Causal Machine Learning



doubleml.org

Application: Randomized Experiments



- **General:** What is the effect of a certain variable D on a relevant outcome variable Y?
- **Randomized experiments** are a direct way to estimate such effects (assuming they are conducted properly)

Challenges in practice:

1. No (pure) A/B-testing / experiments possible → observational data
2. A/B test suffers from low power
3. Heterogenous treatment effects

Solution with DoubleML

1. **Observational study**: Include control variables X which may also impact the variables Y or D
2. Include covariates X that help to predict the outcome Y using ML methods
3. Detection of complex treatment effect patterns

A/B Testing Powered by AI

More precise estimation with ML & AI

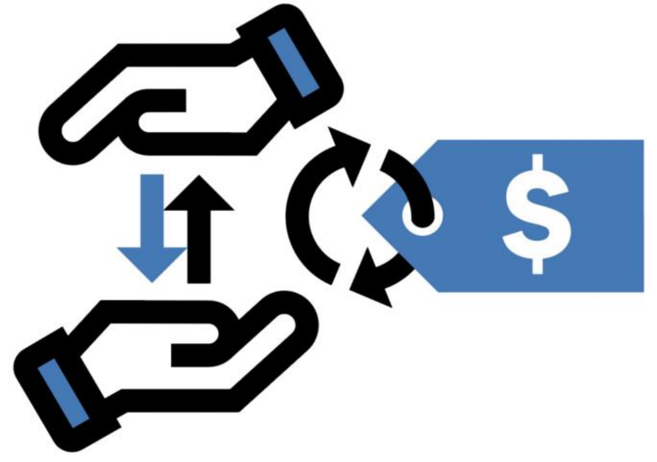
Heterogenous treatment effects & policy optimization

- „Personalized medicine“
- „Personalized marketing“
- „Dynamic Pricing“

Adaptive Experiments & Reinforcement Learning

Example: Price Elasticity of Demand

Price Elasticity of Demand: How does the price impact sales?



- Absolute change in price (EUR 100) and the resulting absolute change in sales (10 million units) can be difficult to interpret

- **Price elasticity of demand**: Percentage change in quantity demanded $\boldsymbol{D}$ when there is a one percent increase in price $\boldsymbol{P}$

$$E_d = \frac{\Delta Q/Q}{\Delta P/P} = \frac{-10/200}{100/1000} = \frac{-0.05}{0.1} = -0.5$$

Econometric model for estimating the price elasticity $\boldsymbol{\theta_0}$:

$$\log(Q) = \alpha + \theta_0 \log(P) + X'\beta + \varepsilon,$$

where the vector of controls $\boldsymbol{X}$ can be very **high-dimensional**

DoubleML - Models

ECONOMIC AI

Implemented Extensions

- Different standard causal models (PLM, IRM, IV)
- Simultaneous Inference for Multiple Treatments
- Clustered Standard Errors
- Group Average Treatment Effects (GATEs)
- Conditional Average Treatment Effects (CATEs)
- (Local) Quantile Treatment Effects (QTEs)
- Effects on Conditional Value at Risk (CVaR)

DoubleML – Models

ECONOMIC AI

Planned Extensions

- DoubleML for difference-in-differences models
- AutoDML
- Sensitivity analysis for omitted variable bias
- Support for unstructured data
- Copula models

Technical Resources

ECONOMIC AI

Available Resources

- UAI Tutorial, 2022, available online
- Online Causal Inference Seminar (OCIS), Stanford, 2023
- useR tutorial 2021
- Chamberlain Seminar presentation, 2022, online
- Online documentation (doubleml.org)

Part II: Use Case

Development of
Marketing Promotion Strategy
For a Software Company

ECONOMIC AI

Development of Marketing Promotion Strategy

ECONOMIC AI

Initial Situation

- Client: **Company** that **develops and sells enterprise software** to other businesses
- Company used **two types of promotional activities** in the past to increase sales
  - **Free training** for users of the software
  - **Vouchers** to get a discount on purchases
- **Goal**: Company wants to know
  - **Do the promotions actually increase sales?**
  - **Which customer should receive what incentive?**

- **Gold standard**: **Experiment** to test effectiveness

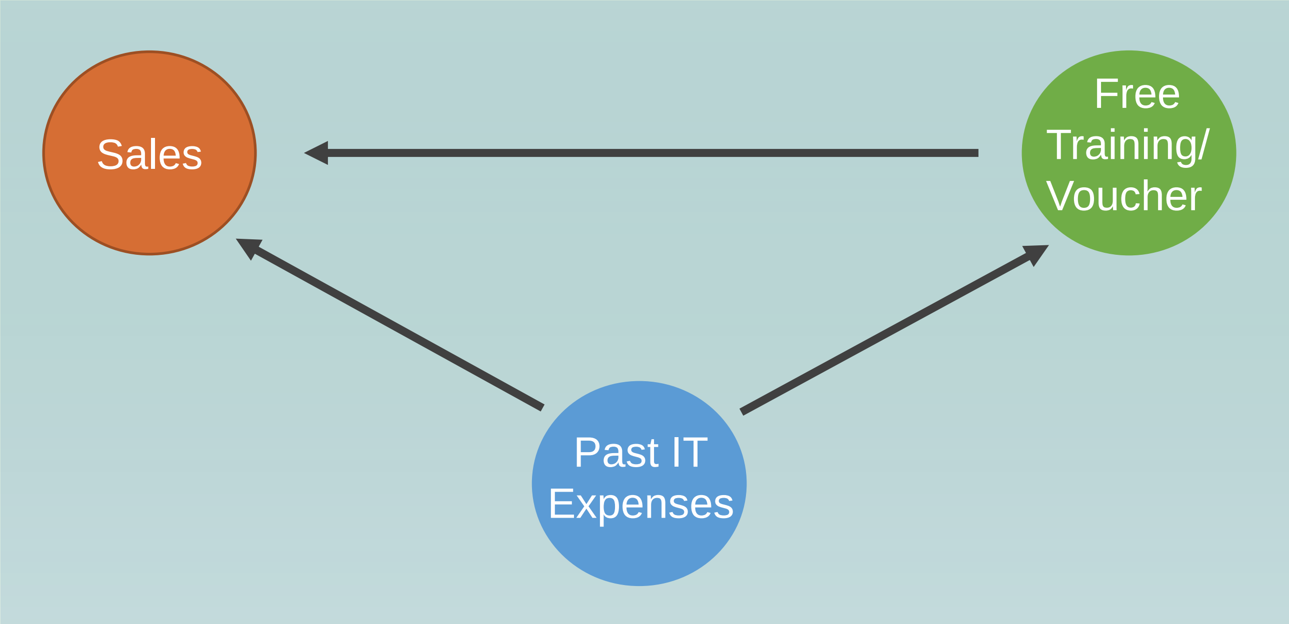
  Company **did not want to do this**

- But: **Data from past transactions** and **promotions available**

  Use Double Machine Learning to estimate effects and derive optimal promotion strategy

Confounding Factors Are Present

ECONOMIC AI



Confounding factors are present

- Customers with **large IT expenses in past received more often incentives** & **correlation between past IT expenses and present sales**
- Potentially other unknown confounders

→ **Need to control for confounders** to get “correct” estimates!

Goal: Estimate causal effect of promotional activities on sales

DML to Estimate Effects of Promotional Activities

ECONOMIC AI

```python
import pandas as pd
import doubleml as dml
from xgboost import XGBRegressor

# Initialize DML data
data_dml = dml.DoubleMLData(
    data,
    y_col='sales',
    d_cols=treatment_vars,
    x_cols=features)

# Instantiate DML model
dml_plr = dml.DoubleMLPLR(
    data_dml,
    ml_l=XGBRegressor(),
    ml_m=XGBRegressor())

# Fit model
dml_plr.fit()
```

$$\boldsymbol{Y = D'\theta_0 + g_0(X) + \epsilon}$$

Y: Sales

D: Promotional activities (& interactions with past revenue)

- Free training & voucher
- Allow effect of promotions to vary by past revenue (treatment effect heterogeneity!)

X: Potential confounding variables

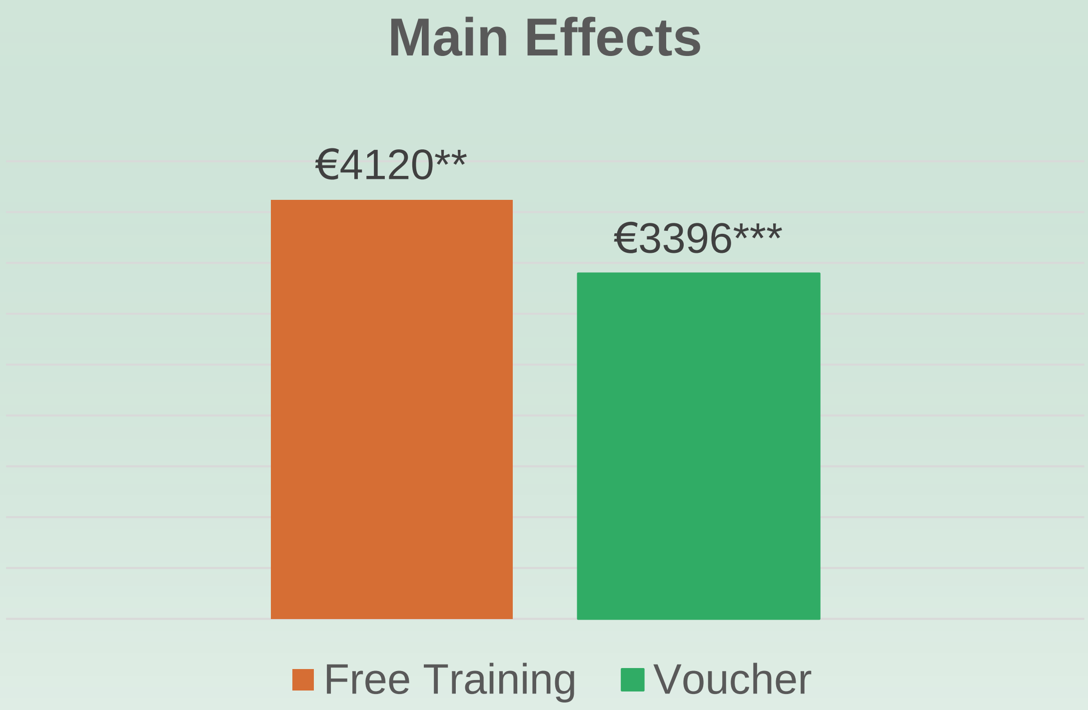
- Large set of customer characteristics, including past IT expenses, # of employees, active worldwide indicator, industry, …

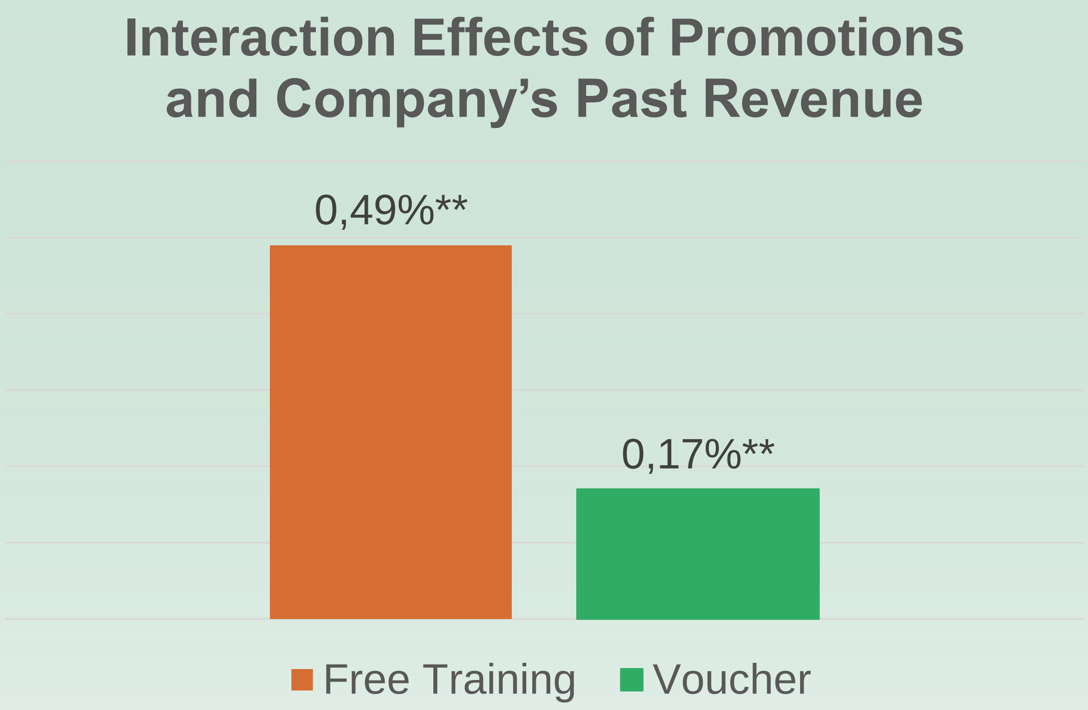
We use a boosted trees algorithm to estimate $g_0(X)$

→ **Fully flexible functional form**

→ **Do not need to worry which variables in X are actually relevant**

Estimated Effects of Promotional Activities





Both types of promotional activities have a positive effect on sales

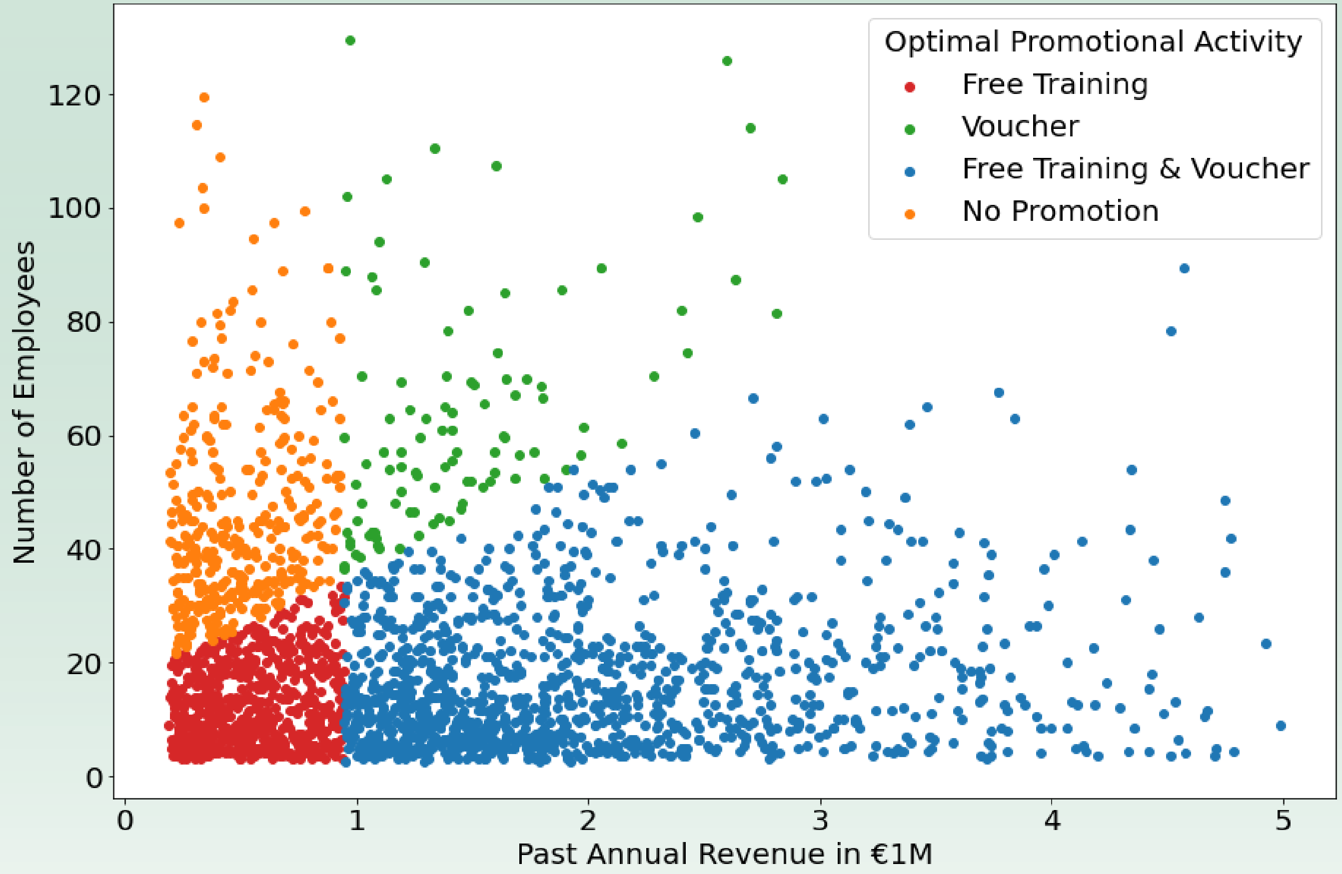
- Free training increases sales by €4120 + 0.49% of past revenue
- Voucher increases sales by €3396 + 0.17% of past revenue

Due to confidentiality reasons, the data and estimates are not from the actual project. However, the insights are qualitatively the same.

Significant at ** 5% level, *** 1% level

Optimal Marketing Promotion Strategy

ECONOMIC AI



What promotional activity should a customer receive?

1. Calculate increase in sales for each type of promotional activity, using effect estimates
2. Subtract costs from sales gains
3. **Choose promotion type that yields highest net gain in sales**

Optimal strategy increases sales by 7.8% vs. 3.7% under strategy actually observed in data (net of promotion costs)

Results & Outlook

- **Company recently started to use the developed strategy** to decide how to target customers

- Sales figures since strategy adoption indicate that **strategy yields indeed gains in sales**

- Next steps
  - **Develop additional strategies** for other promotional activities, customer segments, and markets

More Use Cases for DoubleML

Dynamic Pricing

A/B Testing

Resource Allocation

Personalized Marketing

Customer Retention

… and much more

Further use cases available upon request!

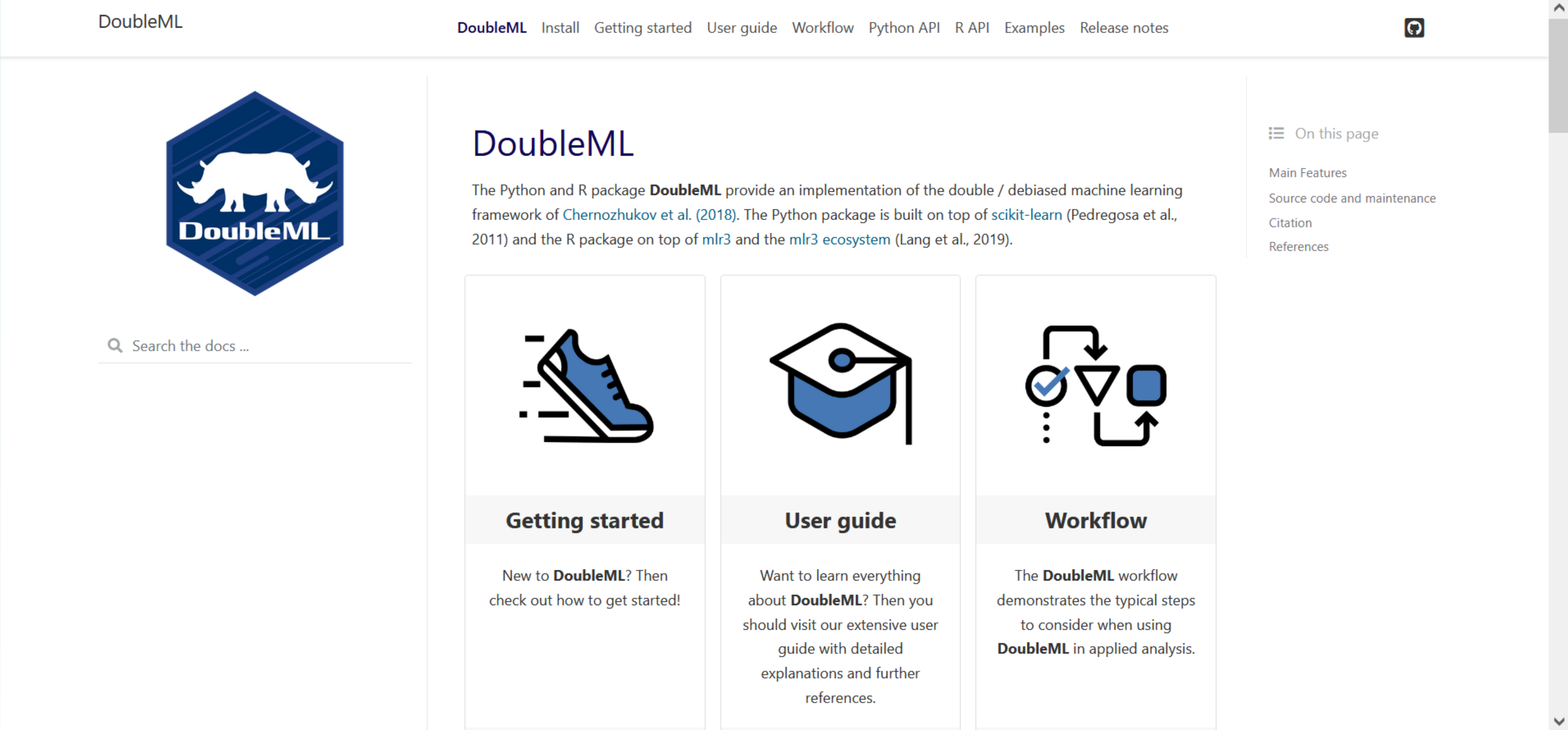
References, Resources & Trainings

Prof. Dr. Martin Spindler

Universität Hamburg &
Economic AI

Online Resources

ECONOMIC AI



DoubleML

DoubleML Install Getting started User guide Workflow Python API R API Examples Release notes

Search the docs ...

DoubleML

The Python and R package **DoubleML** provide an implementation of the double / debiased machine learning framework of Chernozhukov et al. (2018). The Python package is built on top of scikit-learn (Pedregosa et al., 2011) and the R package on top of mlr3 and the mlr3 ecosystem (Lang et al., 2019).

On this page

- Main Features
- Source code and maintenance
- Citation
- References

Getting started

New to **DoubleML**? Then check out how to get started!

User guide

Want to learn everything about **DoubleML**? Then you should visit our extensive user guide with detailed explanations and further references.

Workflow

The **DoubleML** workflow demonstrates the typical steps to consider when using **DoubleML** in applied analysis.

References - Theory

ECONOMIC AI

Double Machine Learning Approach

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duo, E., Hansen, C., Newey, W. and Robins, J. (2018), Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21: C1-C68, doi:10.1111/ectj.12097.

- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (forthcoming), Applied Causal Inference Powered by ML and AI.

DoubleML Package for Python and R

- Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2021), DoubleML - An Object-Oriented Implementation of Double Machine Learning in R, arXiv:2103.09603.

- Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2022), DoubleML - An Object-Oriented Implementation of Double Machine Learning in Python, Journal of Machine Learning Research, 23(53): 1-6, https://www.jmlr.org/papers/v23/21-0862.html.

References – Applications

ECONOMIC AI

Dynamic Pricing mit Künstlicher Intelligenz

Fallstudie aus dem Ride-Sharing-Markt

Big Data stellt Unternehmen vor die Herausforderung, Daten zur Weiterentwicklung des Geschäftsmodells zu verwenden und dabei auf modernste ökonomische und statistische Methoden zu setzen. Damit Unternehmensentscheidungen langfristig zum Geschäftserfolg beitragen, kommt der Kausalität eine herausragende Rolle zu.

Ye Luo, Prof. Dr. Martin Spindler, Philipp Bach

Digital Finance
https://doi.org/10.1007/s42521-021-00046-2

ORIGINAL ARTICLE

Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls

Helmut Wasserbacher[1] · Martin Spindler[2]

Received: 27 May 2021 / Accepted: 17 November 2021

Abstract
This article is an introduction to machine learning for financial forecasting, planning and analysis (FP&A). Machine learning appears well suited to support FP&A with the highly automated extraction of information from large amounts of data. However, because most traditional machine learning techniques focus on forecasting (prediction), we discuss the particular care that must be taken to avoid the pitfalls of using them for planning and resource allocation (causal inference). While the naive application of machine learning usually fails in this context, the recently developed double machine learning framework can address causal questions of interest. We review the current literature on machine learning in FP&A and illustrate in a simulation study how machine learning can be used for both forecasting and planning. We also investigate how forecasting and planning improve as the number of data points increases.

Keywords Financial planning · Machine learning · Forecasting · Causal machine learning · Big data · Double machine learning

JEL Classification Primary G17 · G31 · C53 · C55

The views and opinions expressed in this document are those of the first author, and do not necessarily reflect the official policy or position of Novartis or any of its officers.

Martin Spindler
martin.spindler@uni-hamburg.de

Helmut Wasserbacher
helmut.wasserbacher@novartis.com

[1] Novartis International AG, Novartis Campus, 4002 Basel, Switzerland

[2] Hamburg Business School, University of Hamburg, Moorweidenstr. 18, 20148 Hamburg, Germany

Published online: 16 December 2021

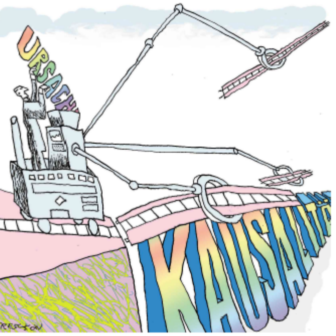
Springer

DER BETRIEBSWIRT

Korrelationen müssen auch kausal sein

Große Datenmengen sind nicht per se von Wert, nur die richtige Analyse von Daten und das Ziehen der richtigen Schlussfolgerungen machen Daten wertvoll.

Von Martin Spindler, Victor Chernozhukov und Ye Luo



Vor der KI kommt die digitale Automatisierung

„Technologie befreit den Menschen vor fehleranfälligen Tätigkeiten“

Use Case: Production Optimization

ECONOMIC AI

https://proceedings.mlr.press/v218/schacht23a.html   Suchen

PMLR **Proceedings of Machine Learning Research**

Volume 218   JMLR   DMLR   TMLR   MLOSS   FAQ   Submission Format   RSS

[edit]

Causally Learning an Optimal Rework Policy

Oliver Schacht, Sven Klaassen, Philipp Schwarz, Martin Spindler, Daniel Grunbaum, Sebastian Imhof *Proceedings of The KDD'23 Workshop on Causal Discovery, Prediction and Decision*, PMLR 218:3-24, 2023.

Abstract

Short Course on Causal Machine Learning

ECONOMIC AI

DoubleML

trainings.doubleml.org

Europe & Asia Edition
Oct 18 & 19, 2023

Short Course:
Causal Machine Learning with DoubleML

A hands-on workshop on causal/double machine learning for data scientists of all technical levels

2-day **Intensive training in Causal ML with DoubleML (virtual):**

- **Instructors:** Prof. Dr. Martin Spindler, Dr. Bach, Dr. Klaassen, Dr. Heinrich Koegel
- **Dates:** Oct 18 & 19, 2023
- **Time:** 10 am –5pm (CEST)
- **Fee:** 950€ (+VAT) regular, 495€ (+VAT) academics
- **Registration:** https://doubleml-training-oct-2023.eventbrite.de

About the Course

- Introduction to Causal Machine Learning with DoubleML for Python
- From basics of Causal ML to advanced topics (heterogeneous treatment effects, difference-in-differences, sensitivity analysis, ...)
- Taught by members of the DoubleML developer team
- Hands-on sessions with examples from pricing, targeted marketing and A/B testing
- Dedicated Q&A sessions
- Small classes (at most 7 participants in hands-on sessions)
- All materials will be shared with participants

Syllabus

Day 1

- Introduction to Causality/Causal Machine Learning
- Basics of Double Machine Learning
- Workflow and Hands-on Examples

Day 2

- Advanced Topics: Overview Causal Models and Extensions
- Senstivity Analyis, Diff-in-Diff
- Hands-on-Examples and Outlook

Contact
trainings@economicai.com
ECONOMIC AI
www.economicai.com
Economic AI GmbH
Nürnberger Str. 262 A
93059 Regensburg
Germany

DoubleML

traini

Short Course:
Causal Machine Learning with Do

A hands-on workshop on causal/double machine data scientists of all technical levels

Two courses coming up!
1) 18 & 19 Oct 2023
2) 14 & 15 Nov 2023

Use Cases and Hands-on Examples:

A/B Testing

Uplift Modelling

Dynamic Pricing

Resource Allocation

Clinical Trials

Production optimization

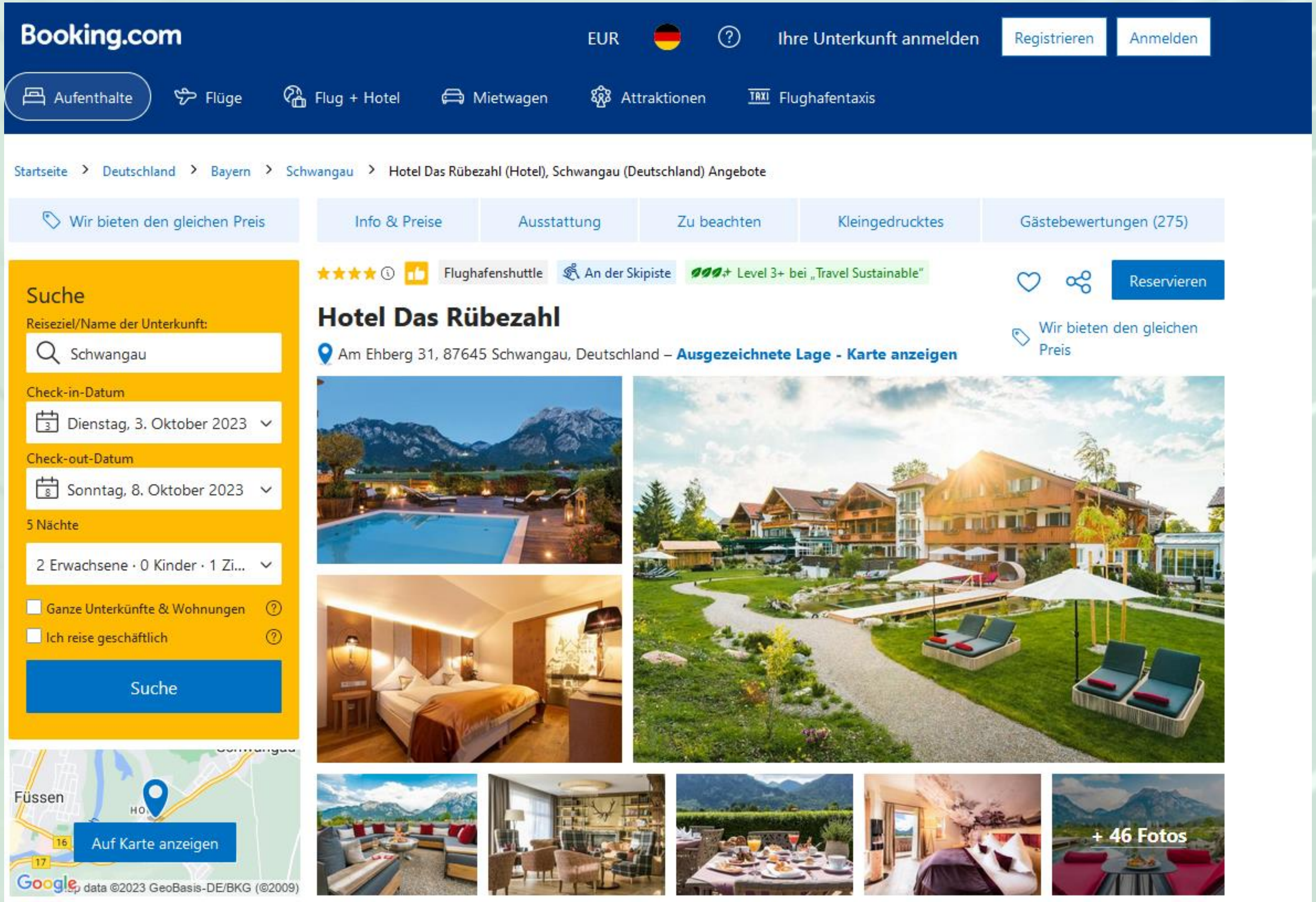
Registration and More Details

Registration (eventbrite)

trainings.doubleml.org

Contact
trainings@economicai.com
ECONOMIC AI
www.economicai.com
Economic AI GmbH
Nürnberger Str. 262 A
93059 Regensburg
Germany

Outlook – DoubleML Deep



Thank You for Your Attention

ECONOMIC AI

Economic AI – The Experts for Causal AI

Software – Modelling – Trainings

BOSTON · HAMBURG · HONG KONG · MUNICH

www.economicai.com

Prof. Dr. Martin Spindler

spindler@economicai.com

martin.spindler@uni-hamburg.de

linkedin.com/in/martin-spindler

For cooperation and use cases feel free to reach out to us!

Double ML: Causal Inference based on ML

More on Theory …

Prof. Dr. Martin Spindler

Universität Hamburg &
Economic AI

DoubleML

ECONOMIC AI

The Key Ingredients of DML

ECONOMIC AI

1. Neyman Orthogonality

The inference is based on a score function $\boldsymbol{\psi}(\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{\eta})$ that satisfies

$$\mathbb{E}[\boldsymbol{\psi}(\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{\eta})] = \mathbf{0}$$

Where $\boldsymbol{W} \coloneqq (\boldsymbol{Y},\boldsymbol{D},\boldsymbol{X},\boldsymbol{Z})$ and with $\boldsymbol{\theta}_0$ being the unique solution that obeys the **Neyman orthogonality condition**

$$\boldsymbol{\partial}_{\boldsymbol{\eta}}\, \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{W};\boldsymbol{\theta},\boldsymbol{\eta})]\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} = \mathbf{0}$$

- For many models the Neyman orthogonal score functions are linear in $\boldsymbol{\theta}$

$$\psi(W;\theta,\eta) = \psi_a(W;\eta)\theta + \psi_b(W;\eta)$$

- The estimator $\tilde{\boldsymbol{\theta}}_0$ then takes the form

$$\tilde{\theta}_0 = -(\mathbb{E}_N[\psi_a(W;\eta)])^{-1}\mathbb{E}_N[\psi_b(W;\eta)]$$

PLR example: Orthogonality by including the first-stage regression, i.e., the regression relationship of the treatment variable $\boldsymbol{D}$ and the regressors $\boldsymbol{X}$

Orthogonal score function $\psi(\cdot) = (Y - E[Y|X] - \theta(D - E[D|X]))(D - E[D|X])$

Neyman Orthogonality

The two strategies rely on very different moment conditions for identifying and estimating $\boldsymbol{\theta_0}$

$$\mathbb{E}[\psi(W, \theta_0, \eta_0)] = 0$$

Naive approach

$$\psi(W, \theta_0, \eta) = (Y - D\theta_0 - g_0(X))D$$

Regression adjustment score

$$\eta = g(X),$$
$$\eta_0 = g_0(X),$$

FWL partialling out

$$\psi(W, \theta_0, \eta_0) = ((Y - E[Y|X]) - (D - E[D|X])\theta_0)(D - E[D|X])$$

Neyman-orthogonal score (Frisch-Waugh-Lovell)

$$\eta = (g(X), m(X)),$$
$$\eta_0 = (g_0(X), m_0(X)) = (\mathbb{E}[Y \mid X], \mathbb{E}[D \mid X])$$

Both estimators solve the empirical analog of the moment conditions:

$$\frac{1}{n}\sum_{i=1}^{n} \psi(W_i, \theta, \hat{\eta}_0) = 0,$$

where instead of unknown nuisance functions we plug-in their ML-based (hold-out) estimators

The Key Ingredients of DML

ECONOMIC AI

2. High-Quality Machine Learning Estimators

The nuisance parameters are estimated with high-quality (fast-enough converging) machine learning methods.

- Different structural assumptions on $\boldsymbol{\eta_0}$ lead to the use of different machine-learning tools for estimating $\boldsymbol{\eta_0}$ (Chernozhukov et al., 2018, Chapter 3)

3. Sample Splitting

To avoid the biases arising from overfitting, a form of sample splitting is used at the stage of producing the estimator of the main parameter $\boldsymbol{\theta_0}$.

- Cross-fitting performs well empirically (efficiency gain by switching roles)

Double ML: Causal Inference based on ML

Part II: Double Machine Learning in Practice

Prof. Dr. Martin Spindler

Universität Hamburg & Economic AI

DoubleML

ECONOMIC AI

Overview

DoubleML provides a general implementation of the Double Machine Learning approach by Chernozhukov et al. (2018) in Python and R

There are also other open-source libraries available for causal machine learning:

- **CausalML** (uber, *https://github.com/uber/causalml*, Chen et al., 2020) - variety of causal ML learners, i.a. with focus on uplift modeling, CATEs and IATEs
- **EconML** (microsoft research, *https://github.com/microsoft/EconML*, Battocchi et al., 2021) – various causal estimators based on machine learning, among others based on double machine learning approach
- …

CausalML and **EconML** have a focus on heterogeneity of treatment effects from their start on

DoubleML focuses on implementing the DML approach and its extensions (example: heterogeneity)
- Object-orientated implementation based on orthogonal score
- Extendibility and flexibility

Building Principles

ECONOMIC AI

Key ingredient and implementation

- **Orthogonal Score**
  - Object-oriented implementation
  - Exploit common structure being centered around a (linear) score function $\boldsymbol{\psi}(\cdot)$
- **High-quality ML**
  - State-of-the-art ML prediction and tuning methods
  - Provided by `scikit-learn` and `scikit-learn`-like learners
- **Sample Splitting**
  - General implementation of sample splitting

Why an Object-Orientated Implementation?

ECONOMIC AI

Given the components $\psi^a(\cdot)$ & $\psi^b(\cdot)$ of a linear Neyman orthogonal score function $\psi(\cdot)$, a general implementation is possible for

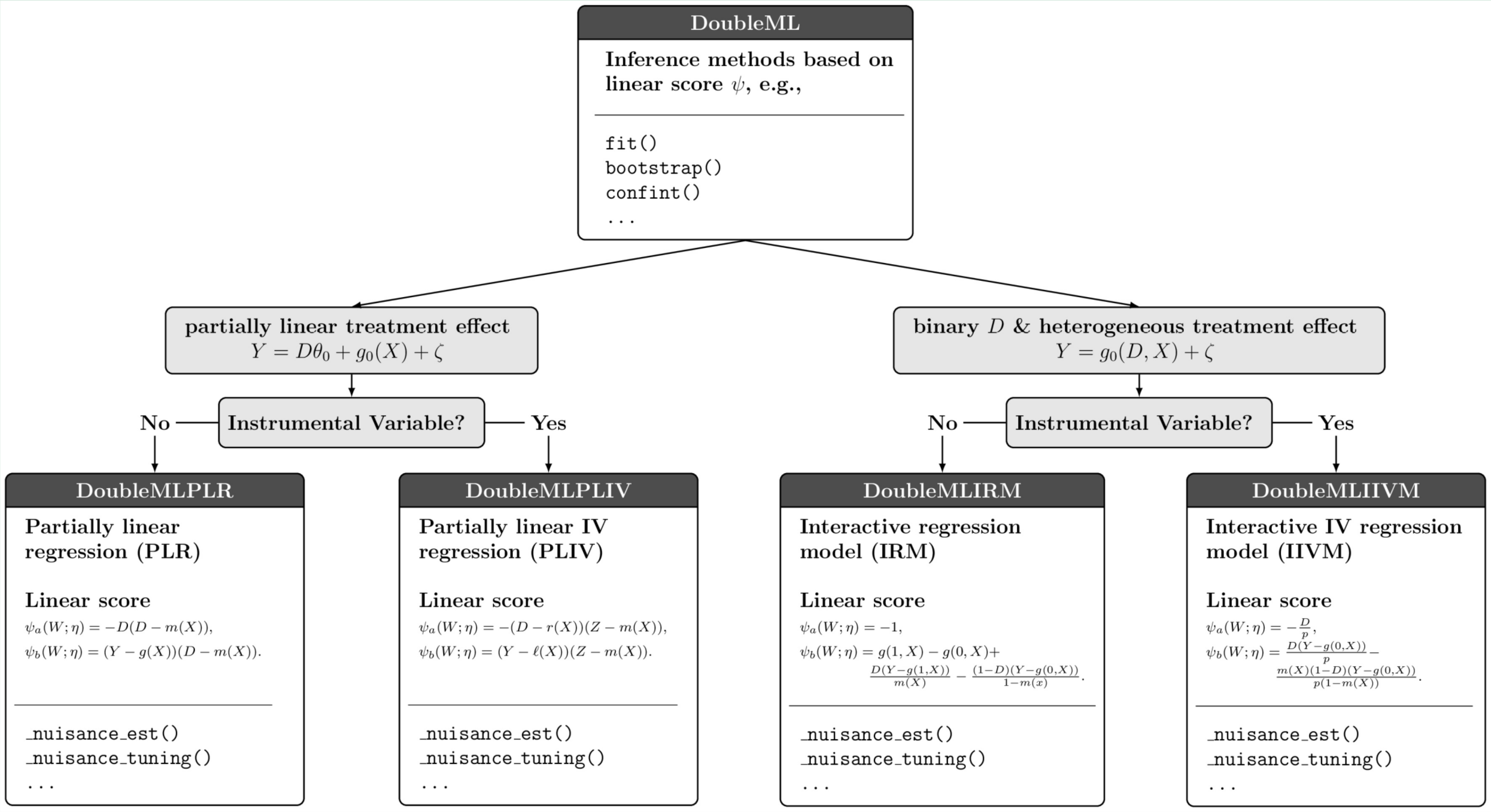
- The estimation of the **orthogonal parameters**
- The computation of the **score** $\psi(W;\theta,\eta)$
- The estimation of **standard errors**
- The computation of **confidence intervals**
- A **multiplier bootstrap** procedure for simultaneous inference

- The **sample splitting** can be implemented in general as well
  - ➢ Implemented in the **abstract base class** `DoubleML`

- The **score components** and the estimation of the **nuisance models** have to be implemented **model-specifically**
  - ➢ Implemented in **model-specific classes** inherited from `DoubleML`

Class Structure and Causal Models

ECONOMIC AI



Advantages of the Object-Orientation

DoubleML gives the user a **high flexibility** with regard to the specification of DML models:

- Choice of ML methods for approximating the nuisance functions
- Different resampling schemes (repeated cross-fitting)
- DML algorithms DML1 and DML2
- Different Neyman orthogonal score functions

DoubleML can be **easily extended:**

- New model classes with appropriate Neyman orthogonal score function can be inherited from DoubleML
- The package features **callables** as score functions which makes it easy to extend existing model classes
- The resampling schemes are customizable in a flexible way

Installation

Install the latest release via pip or conda, see **installation guide**

```
pip install -U DoubleML
```

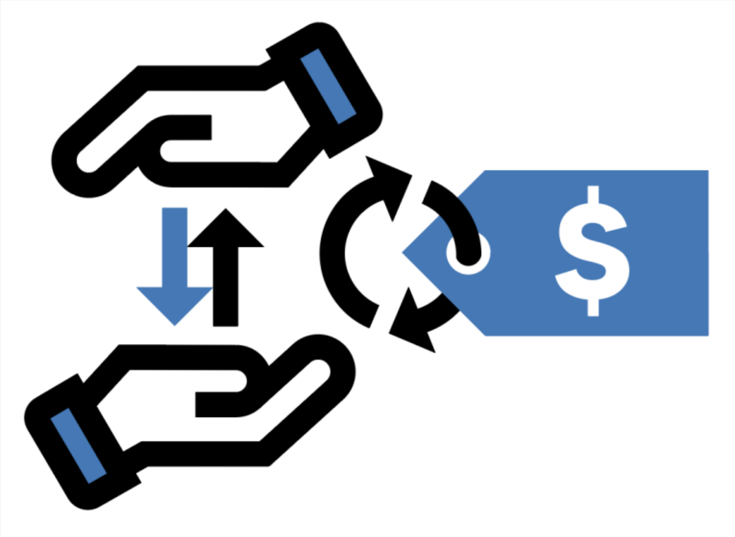
```
conda install -c conda-forge doubleml
```

Install development version from GitHub https://github.com/DoubleML/doubleml-for-py

See the Getting Started page of the tutorial website for more information on prerequisites.

Data Example: Demand Estimation

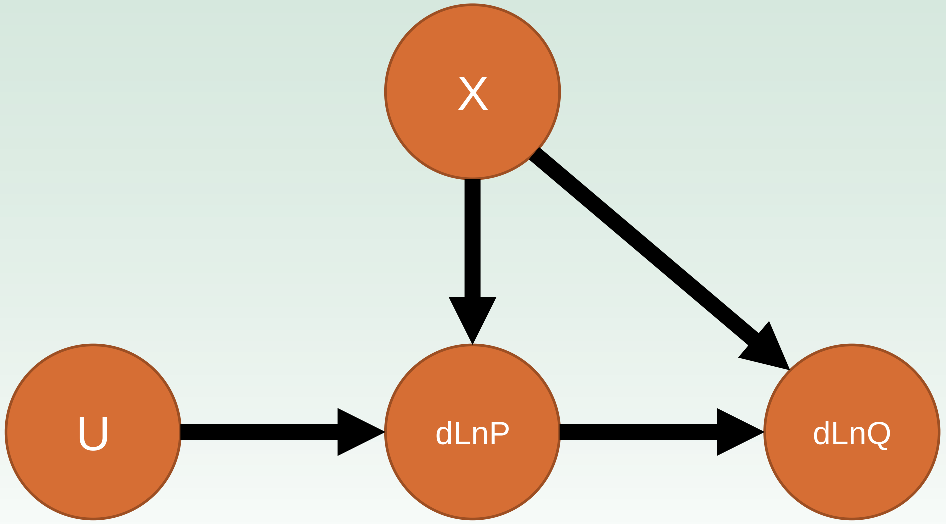
ECONOMIC AI



Data Source:

- Data example based on a blogpost by Lars Roemheld (Roemheld, 2021)
- Original real data set publicly available via kaggle, preprocessing notebook available online

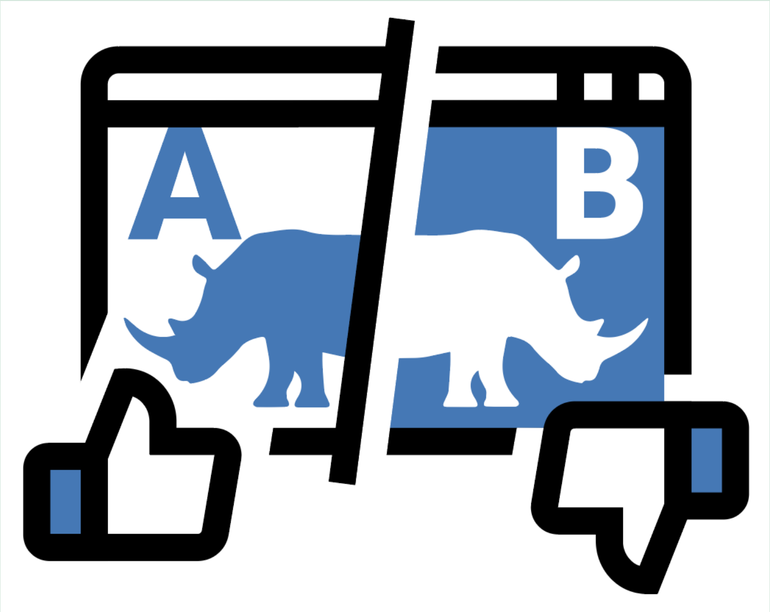
Causal Diagram (DAG):



Causal Problem:

- **Price elasticity of demand**: What is the **effect** of a **price change**, $\boldsymbol{dLnP}$, on **demanded quantity**, $\boldsymbol{dLnQ}$?
- **Observational study**: Flexibly adjust for confounding variables $\boldsymbol{X}$, e.g. product characteristics

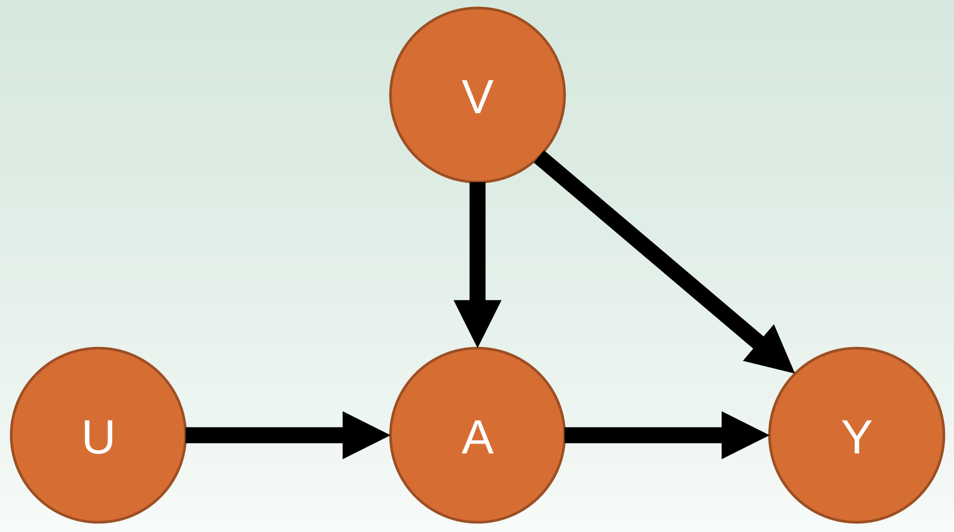
Data Example: A/B Testing



Data Source:

- Data example based on a randomly chosen DGP created for the 2019 ACIC Data Challenge.

Causal Diagram (DAG):



Causal Problem:

- **Online shop**: What is the effect of a **new ad design** A on sales Y (in \$100)?
- **Observational study**: Necessary to adjust for confounding variables V

Online Resources

ECONOMIC AI

- The notebook is organized according to the DoubleML Workflow
- Extensive User Guide available via docs.doubleml.org
- Documentation for the Python API available via https://docs.doubleml.org/stable/api/api.html
- Paper for the Python package available from JMLR or arxiv