GESIS Leibniz Institute for the Social Sciences

gesis training

GESIS Fall Seminar in Computational Social Science 2025
“Causal Machine Learning”

Lecturer:	Marica Valente
Affiliation:	University of Innsbruck
Email:	marica.valente@uibk.ac.at

Date: September 22-26, 2025
Time: 09:00-12:00 & 13:00-16:00 (CEST / UTC+2)

About the Lecturer

Marica Valente is an economist and applied econometrician and is currently an Assistant Professor at the University of Innsbruck since 2022. Previously, she was a Postdoctoral Researcher at ETH Zurich and earned her PhD from the Berlin School of Economics in 2020. Her work bridges empirical analysis and methodological innovation in the fields of environmental economics, labor and migration, conflict economics, and gender economics. She uses and develops advanced econometric and machine learning tools to address causal inference questions.

Course Description

Machine learning (ML) has revolutionized the way we analyze data, making it an essential tool for prediction in a wide range of applications, from forecasting economic trends to assessing environmental risks. However, while ML excels at predicting what is likely to happen, many key questions in science go beyond prediction and require understanding causal relationships—that is, answering "what if" questions about the effects of treatments and policies. This course is designed to equip participants with the fundamental ML techniques for prediction and show how they can be tailored to answer causal questions.

Starting from linear regression, the foundation of many ML models, the course will introduce participants to high-dimensional predictive modeling and the challenges of applying these tools to causal inference. While standard ML techniques are optimized for minimizing prediction errors, they often fail when directly applied to causal questions. For instance, we might use ML to predict tomorrow’s air pollution levels based on weather conditions, but a policymaker needs to know whether restricting traffic would actually reduce pollution—a fundamentally different problem requiring causal analysis.

This course will teach participants how to adapt machine learning methods for causal inference, integrating modern ML algorithms with causal models from econometrics and statistical inference. Through hands-on tutorials in R—the primary software for implementing causal machine learning—participants will work with real-world datasets to apply, compare, and critically evaluate these methods.

Participants will explore the differences between standard causal effect estimation techniques and causal ML approaches, gaining a clear understanding of what each method can deliver differently and the contexts in which they are most effective. Additionally, they will learn to distinguish between predicting observable outcomes and estimating causal effects, developing the necessary skills to bridge the gap between conventional ML tools and rigorous causal analysis.

Beyond technical skills, the course will emphasize critical thinking and the ability to identify meaningful research questions. Participants will have the opportunity to present their research ideas and preliminary findings through optional oral presentations, fostering discussion and feedback.

By the end of the course, participants will not only be proficient in machine learning tools for prediction but will also understand how to adapt them for rigorous causal analysis—a crucial skill for treatment effect evaluation and evidence-based policy decision-making.

Organizational Structure of the Course

The course consists of live online sessions every day, combining lectures on methods and applications with hands-on R tutorials. Lectures will be highly interactive, with dedicated Q&A sessions at the end of each section to engage participants. R tutorials will feature practical exercises using real-world datasets to reinforce key concepts. Participants will have the opportunity to present their ongoing research in the field during dedicated short presentation slots. Those interested in presenting are encouraged to submit a brief, informal summary (e.g., an abstract) of their research topic to marica.valente@uibk.ac.at.

Keywords

Machine Learning, Causal Inference, High-Dimensional Data, Heterogeneous Treatment Effect Estimation, Policy Evaluation, R Software

Target Group

This course is ideal for you if:

- You are a researcher, student or practitioner interested in causal inference methods for evaluating treatment effects and policy interventions.
- You want to estimate personalized treatment effects in social sciences, such as assessing individualized causal effects of policies to optimize targeting.
- You work with or plan to analyze high-dimensional datasets containing a large number of variables and/or observations.
- You seek applications in the social sciences and economics, where data-driven insights can inform decision-making.
- You have an interest in coding in R and applying machine learning methods for causal analysis.
- If you are not a social scientist, you work or plan to work in a field where personalized treatment effect estimation is valuable, such as evaluating the heterogeneous impacts of medical treatments on health.

Course and Learning Objectives

By the end of the course, you will:

- Understand the distinction between statistics, econometrics, and machine learning, and how these fields approach data analysis differently, particularly in high-dimensional settings.
- Develop proficiency in machine learning methods for prediction, including non-parametric (CART, Random Forests) and parametric (LASSO) techniques, and apply them using R.
- Gain a strong foundation in causal inference methods, learning when and how machine learning can be adapted for causal analysis, including Double Machine Learning and orthogonalization techniques.
- Apply standard methods for causal effect estimation and causal machine learning using R, understanding what they can deliver differently and when to use each approach.
- Explore advanced causal inference methods in high-dimensional settings, such as synthetic controls and synthetic differences-in-differences, and understand their application in policy evaluation.
- Learn how to estimate heterogeneous treatment effects, differentiating between Average Treatment Effects (ATE) and Conditional Average Treatment Effects (CATE), and implement causal trees and generalized random forests in R.
- Improve your ability to critically assess and communicate empirical findings, through hands-on exercises, oral presentations, and discussions on the strengths and limitations of machine learning for causal inference.

Course Prerequisites

- Participants should have completed an undergraduate-level introduction to statistics or econometrics.
- The course requires basic knowledge of the linear OLS regression method.
- No previous knowledge of machine learning is required.
- Prior experience with R is not a prerequisite, however, it is strongly recommended. Alternatively, participants with prior experience in Stata or Python might use some of the resources below (see Recommended Literature to Look at in Advance) to ensure they have sufficient proficiency in R to follow the course. If you have little

experience in R or want to refresh your sills, I recommend to familiarize yourself with the software using free online tools, e.g. https://www.datacamp.com/courses/free-introduction-to-r (sign up and start the free course on Introduction to R), https://swirlstats.com/ (learn R, in R). You may also consider taking the online workshop "Introduction to R" that takes place from 25-27 August.

Software Requirements

Participants should have R and RStudio installed on their machines, including the following packages:
rpart, rpart.plot, randomForest, caret, pdp, glmnet, hdm, weights, gplots, dplyr, plm, lmtest, Synth, SCtools, synthdid, tidyverse, grf, fixest, car, haven, spatstat

We recommend using the latest R version 4.4.2 (2024) and RStudio version 2024.12.0.467. If you need to update R, you can run:

```
install.packages("installr")
library(installr)
updateR()
```

Course Contents

- Foundations of Machine Learning and High-Dimensional Statistics
- Machine Learning Methods for Prediction
- Advanced Causal Inference Methods in High-Dimension
- Double Machine Learning for Average Treatment Effects
- Causal Machine Learning for Heterogeneous Treatment Effects
- Practical Applications and Interpretation of Causal Machine Learning in R

Recommended Literature to Look at in Advance

Below are some insightful readings to explore before the course (optional). You may skip technical sections as needed.

• (paper) Varian, H. (2014): Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28(2), pp. 3-28 https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3

• (paper) Athey, S. (2018): The Impact of Machine Learning on Economics. The Economics of Artificial Intelligence: An Agenda. University of Chicago Press https://www.gsb.stanford.edu/faculty-research/publications/impact-machine-learning-economics

For those looking to familiarize themselves and/or freshen up their R skills before the course:

• (R introductory) Stauffer, R., Chimiak-Opoka, J. Rodríguez-R, L. M., Thorsten, S. Zeileis, A. Introduction to Programming with R. https://discdown.org/rprogramming/

• (R technical) Venables, W. N., Smith, D. M. and the R Core Team (2018): An Introduction to R. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Day-to-day Schedule and Literature

Note: The syllabus readings are not mandatory, but reviewing some before each session will be beneficial.

Day 1: Low- vs. High-Dimensional Problems

- Draw differences between Statistics, Econometrics and Machine Learning
- Linear Regression (OLS), Assumptions, and Flexibility: What to do when OLS breaks down?
- The Curse of Dimensionality: Lost in High-Dimensional Spaces
- Non-parametric methods: CART (Classification and Regression Trees)
- R Tutorial: Mortality Predictions with CART

Lectures are based on the following literature:

• (textbook) Giraud, C. (2021). Chapter 1 of "Introduction to High-Dimensional Statistics." Monographs on Statistics and Applied Probability 139. CRC Press. https://www.imo.universite-paris-saclay.fr/~christophe.giraud/Orsay/Bookv3.pdf

• (paper) Frey, S. & Savage, A. & Torgler, B. (2011). "Behavior under Extreme Conditions: The Titanic Disaster." Journal of Economic Perspectives 25(1), pp. 209-22. https://www.aeaweb.org/articles?id=10.1257/jep.25.1.209

Day 2: Machine Learning Methods for Prediction

- Non-parametric methods: Random Forests
- R Tutorial: Mortality Predictions with Random Forests and comparison with CART
- Parametric methods: LASSO and other regression-based methods (e.g., elastic net)
- R Tutorial: Mortality Predictions with LASSO and other regression-based methods

Lectures are based on the following literature:

:• (textbook) James, G., Witten, D., Hastie, T., and R. Tibshirani, R. (2013). Chapter 4, 5, 6, 8 of "An Introduction to Statistical Learning with Applications in R." Springer. https://www.statlearning.com/

• (paper) Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society. Series B 58(1), pp. 267-288. https://www.math.cuhk.edu.hk/~btjin/math6221/p7.pdf

• (paper) Zou, H., Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net", Journal of the Royal Statistical Society. Series B 67(2), pp. 301–320. https://academic.oup.com/jrsssb/article-abstract/67/2/301/7109482

Day 3: Causal Machine Learning for Average Treatment Effects

- ML Methods for Causal Analysis: Review of Causal Inference Methods in High-Dimension
- Double ML Methods: Orthogonalization and Doubly Robust LASSO
- R Tutorial: Policy Evaluation with Standard OLS (Difference-in-Differences)
- R Tutorial: Policy Evaluation with Doubly Robust ML

Lectures are based on the following literature:

• (paper) Belloni, A., Chernozhukov, V., and Hansen, C. (2014). "High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics." Journal of Economic Perspectives 28(2), pp. 29-50. https://www.aeaweb.org/articles?id=10.1257/jep.28.2.29

• (paper) Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters." The Econometrics Journal, 21: C1-C68. https://academic.oup.com/ectj/article/21/1/C1/5056401

• (R package) Github "hdm: High-Dimensional Metrics." https://github.com/MartinSpindler/hdm

• (paper) Valente, M. (2023). "Policy Evaluation of Waste Pricing Programs Using Heterogeneous Causal Effect Estimation" Journal of Environmental Economics and Management 117, 102755. https://www.sciencedirect.com/science/article/pii/S0095069622001085?via%3Dihub

Day 4: Other Causal Inference Methods in High-Dimension

- Synthetic Controls in Low- vs. High-Dimension
- Synthetic Differences-in-Differences
- R Tutorial: Policy Evaluation with Synthetic Counterfactuals in Low- and High-Dimension
- The Flow of Averages: Why Uncovering Heterogeneities Matters
- Oral Presentations by Interested Participants

Lectures are based on the following literature:

• (paper) Abadie, A. (2020). "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." Journal of Economic Literature 59(2), pp. 391-425. https://pubs.aeaweb.org/doi/pdfplus/10.1257/jel.20191450

• (paper) Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W., Wager, S. (2021). "Synthetic Difference-in-Differences." American Economic Review 111(12), pp. 4088-4118. https://www.aeaweb.org/articles?id=10.1257/aer.20190159

• (book/essay) Ross, T. (2016). "The End of Average: How to Succeed in a World That Values Sameness." Harper Collins. https://en.wikipedia.org/wiki/The_End_of_Average

• (paper) Bueno, M., Valente, M. (2019). "The Effects of Pricing Waste Generation: A Synthetic Control Approach", Journal of Environmental Economics and Management 96, pp. 274-285. https://www.sciencedirect.com/science/article/abs/pii/S0095069618304169?dgcid=author

Day 5: Causal Machine Learning for Heterogeneous Treatment Effects

- Average Treatment Effects (ATE) vs. Conditional Average Treatment Effects (CATE)
- Causal Trees, Causal Forests, and Generalized Random Forests
- R Tutorial: Causal Forest Estimation of the Heterogeneous Effects of Climate Change on Agriculture
- Methods for Analyzing the Heterogeneity of Conditional Average Treatment Effects
- R Tutorial: How to Disentangle the Key Sources of Treatment Effect Heterogeneity

Lectures are based on the following literature:

• (paper) Wager, S., and Athey, S. (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." Journal of the American Statistical Association 113(523), pp. 1228-1242. https://www.tandfonline.com/doi/abs/10.1080/01621459.2017.1319839?journalCode=uasa20

• (paper) Athey, S., Tibshirani, J., and Wager, S. (2019). "Generalized Random Forests." Annals of Statistics 47(2), pp. 1148-1178. https://faculty.ist.psu.edu/vhonavar/Courses/causality/GRF.pdf

• (paper) Athey, S., Wager, S. (2019). "Estimating Treatment Effects with Causal Forests: An Application." Observational Studies 5, pp. 36-51. https://par.nsf.gov/servlets/purl/10311714

• (paper) Valente, M. (2025). "Machine Learning Insights on Crop Yield Responses to Climate Change." SSRN Working Paper. http://dx.doi.org/10.2139/ssrn.4848983

• (R package) Github "grf: Generalized Random Forests." https://grf-labs.github.io/grf/