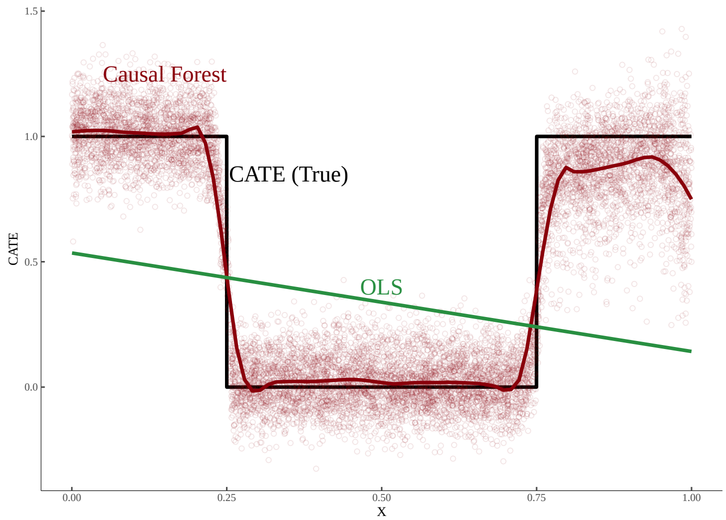
Causal Machine Learning for Observational and Experimental Research

ICPSR Summer Program in Quantitative Methods of Social Research
4–8 August, 2025 (9am–6pm EST)
Zoom



Instructors

Sam Fuller
Postdoctoral Fellow
Center for American Political Studies (CAPS)
Harvard University
sfuller@fas.harvard.edu

Jack T. Rametta
PhD Candidate
Methodology and American Politics
University of California, Davis
jtrametta@ucdavis.edu

Course Description

Machine learning is increasingly being applied at various stages in the social science workflow, and particularly for causal inference. This course provides a comprehensive overview for social scientists looking to incorporate machine learning methods into design-oriented, causal analyses, a broad set of methods that we call *Causal Machine Learning* (*CML*). Specifically, the course walks students through a flexible, data-driven approach to research design and analysis that represents a marked improvement over standard methods for applications such as treatment effect estimation (e.g., ATEs, CATEs, etc.), balance testing, and attrition detection. The topics covered include doubly-robust machine learning (DRML); interpretable ML; variable importance analysis; applications to observational data, experiments, and designed-based cases (e.g., difference-in-differences, regression-discontinuity, etc.); and various CML methods such as Causal Forest. Throughout, we compare the utility of causal machine learning approaches to traditional methods and provide practical guidance for evaluating validity, enhancing performance, and presenting results. Importantly, we provide detailed guidance on how to apply CML methods in a robust, principled way. Overall, this course provides an approachable guide for social scientists who wish to apply cutting-edge machine learning methods to enhance the quality and resolution of their causal research. All of our analyses will be conducted and demonstrated in R.

Required Software

`R` is available for download from CRAN (Comprehensive R Archive Network): `https://cran.r-project.org/`. Whatever flavor of integrated developer environment (IDE) you prefer is fine for this course but you will need one. We generally recommend `RStudio` for beginners and intermediate users. For advanced users, you might also consider `VSCode`, or the more recent, AI-focused IDE `Cursor`. There are many required `R` packages that will be provided in an installation script separately.

While `Python` is not strictly required, it is recommended for some additional/bonus content we may cover in PyTorch (time permitting both EconML and TabPFN).

Course Materials Course materials (slides, code, and materials) will be available on the course Canvas page.

Tentative Schedule

This schedule is subject to change:

- **Day One: Causal Inference & Causal Machine Learning**

  Causal inference refresher: estimands, designed-based inference, classical inference methods

  ⇒*Cunningham Chapter 4.0–4.1, Wager 1, & Chernozhukov et al. 2.1*

  Causal/doubly-robust machine learning framework

  ⇒*Wager 3, & Chernozhukov et al. 4, 10*

- **Day Two: Integrating Causal Inference & Classical Machine Learning**

  Classical machine learning refresher: decision trees, random forests, gradient boosting

  ⇒*Hastie et al. 9.2, 10.1, 15*

  Exploiting classical ML for inference tasks

  ⇒*Molnar 8.1*

- **Day Three: Causal Machine Learning for Observational Data (Selection on Observables)**

  Propensity score estimation, overlap weighting estimators

  ⇒*Wager 2, 7 & Chernozhukov et al. 5*

  Integrating ML into standard causal inference designs (e.g., diff-in-diff, RDD, etc.)

  ⇒*Wager Ch. 8, 13, Chernozhukov et al. 12, 16, & 17*

- **Day Four: Causal Machine Learning for Experimental Data**

  CML for conditional average treatment effect estimation and inference

⇒*Wager Ch. 4, Chernozhukov et al. 15*

CML for balance and attrition testing

⇒*Rametta & Fuller (2025)*

- **Day Five: Causal Machine Learning Frontier & Extensions**

  Conformal inference for CML uncertainty estimation

  ⇒*Samii's Conformal Tutorial*

  Integrating unsupervised learning and CML

  CML and research design

  Bonus `Python` implementations (if time permits)

  ⇒*tabPFN & econML*

Texts

In addition to our own materials, we will rely heavily on selections from Wager (2024) and Chernozkuhov et al. (2024) for this course as well as a smattering of other published works. These are excellent resources and, remarkably, all are available for free online.

1. Wager, Stefan. 2024. *Causal Inference: A Statistical Learning Approach.*
   `https://web.stanford.edu/~swager/causal_inf_book.pdf`

2. Chernozhukov, Hansen, Kallus, Spindler, and Syrgkanis. 2024. *Applied Causal Inference Powered by ML and AI.*
   `https://causalml-book.org/assets/chapters/CausalML_book.pdf`

3. Cunningham, Scott. 2021. *Causal Inference: The Mixtape.*
   `https://mixtape.scunning.com/`

4. Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. 2017. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.*
   `https://hastie.su.domains/ElemStatLearn/`

5. Molnar, Christoph. 2024. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable.*
   `https://christophm.github.io/interpretable-ml-book/`

6. Alves, Matheus Facure. 2022. *Causal Inference for The Brave and True.*
   `https://matheusfacure.github.io/python-causality-handbook/landing-page.html`