MPhil in Economics and Data Science:
DS 300: Causal Inference and Machine Learning

Faculty of Economics
University of Cambridge

Lent 2026

Dr. M. J. Weeks
Associate Professor of Economics

Overview

The course will focus upon topics at the intersection of machine learning and econometrics, covering a mix of theory and applications. In making the distinction between models which are used to solve a prediction problem and models which are used to estimate some form of causal effect, we demonstrate how empirical strategies such as unconfoundedness, instrumental variables, and difference-in-difference can be used alongside machine learning methods for prediction.

Using Breiman's (2001) notion of two cultures in the use of statistical modelling, the course begins with a review of the fundamental differences between machine learning and econometrics.

> *There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.*
>
> Breiman [2001], p199.

The tension between parametric and nonparametric approaches reflects fundamental disciplinary differences. Econometricians prioritise parametric models with explicit functional forms to ensure interpretable parameters and structural understanding of economic relationships. Machine learning practitioners prioritise nonparametric flexibility, treating the data-generating mechanism as unknown to maximize external validity and generalisation performance. Modern causal machine learning confronts the challenge of reconciling these competing objectives - achieving both structural understanding and robust generalisation

Causal tasks - whether parametric or nonparametric - ask fundamentally different questions: what happens to an outcome Y if an element of X changes? For causal inference, we need variation that isolates the causal effect of interest, ensuring internal validity - that observed associations reflect genuine causal relationships. For example, a model that predicts ice cream sales across locations using price, temperature, and store characteristics tells us nothing about the causal effect of a price change, because price itself may reflect local demand conditions that also drive sales. Causal machine learning

methods, such as Double ML and causal forests, seek to ensure external validity through techniques like sample splitting and cross fitting.

During the course we review three literatures

- Causal estimation and inference using alternative methods for identification.
- High dimensional statistics for trading bias and variance
- Machine Learning algorithms for prediction problems

Points of Departure

In covering three broad literatures we provide a number of points of departure which provide entry points for participants with little background in causal machine learning. These are:

➢ Conditional Expectation Function

The Conditional Expectation Function $E[Y|X]$ serves as the fundamental mathematical object connecting prediction and causal inference. As the minimum mean squared error predictor of Y given X, the CEF provides the optimal solution to prediction problems regardless of whether we approach it through parametric assumptions or algorithmic flexibility.

We begin with the familiar linear regression model where $E[Y|X] = X'\beta$ provides the best linear approximation to the CEF. This parametric approach offers interpretability, well-understood statistical properties, and computational tractability.

However, modern applications increasingly demand that we move beyond linearity. We explore this through nonparametric methods that can flexibly capture complex, nonlinear relationships while maintaining the CEFs role as the bridge between prediction accuracy and causal identification. This sets the stage for understanding how machine learning methods can be adapted from pure prediction tasks to serve causal inference objectives.

➢ High Dimensional Methods in Statistics

A high-dimensional setting arises naturally in two ways: either we observe many covariates directly (hundreds of demographic variables, thousands of product features), or we wish to consider many transformations and interactions of a smaller set of base covariates to capture nonlinearities.

High-dimensional methods like the LASSO address this challenge by adding penalty terms to the estimation criterion, trading off model complexity against fit. However, regularisation methods that makes these methods work for prediction introduces bias that can invalidate causal inference - a problem that motivates the development of "double-debiased" machine learning methods that separate the tasks of model selection and parameter estimation.

➢ The Frisch-Waugh-Lovell (FWL) Theorem: Low Dimensionsal Case

In a world where the analyst cares about a specific causal parameter, the Frisch-Waugh-Lovell (FWL) theorem provides a fundamental insight: we can transform a multivariate regression problem into a bivariate one through the process of "partialing out."

For d the variable of interest, X represents confounding variables, and τ is our target causal parameter, the FWL theorem reveals that an esimate of τ can be obtained by first removing (partialing out) the linear influence of X from both Y and d, then regressing the residualised Y on the residualised d. This two-step procedure yields exactly the same estimate as the full multivariate regression, demonstrating that causal estimation fundamentally involves isolating variation in treatment that is orthogonal to confounders.

➢ Parametric versus Nonparametric Methods

The choice between parametric and nonparametric methods represents a fundamental decision about how we approach estimating the CEF. Parametric methods assume a specific functional form - typically $E[Y|X] = X\beta$ - offering interpretability and established tools for statistical inference.

Nonparametric methods take a different approach by allowing the data to reveal the shape of the CEF. Consider the covariate space $\mathcal{X}$ - the set of all possible values our features can take. Rather than imposing a global linear structure, nonparametric methods partition $\mathcal{X}$ into local regions and estimate conditional expectations within each region. This can be as simple as computing average outcomes within regions defined by covariate values, or as sophisticated as using kernels, trees, or neural networks to approximate $E[Y|X] = f(X)$ without prespecifying the form of $f(\cdot)$.

➢ Treatment Effects

We briefly review the treatment effects literature including alternative parametric estimators and identification assumptions. This will provide a useful reference point when considering the use of nonparametric methods based on ML algorithms.

➢ Machine Learning Methods for Prediction

We make reference to the two broad types of machine learning methods for prediction, making the link to nonparametric regression and nearest neighbours. We then consider a number of fundamental building blocks, starting with error decomposition in terms of bias and variance, the role of sample splitting, and regularization as a means to avoid overfitting.

The Frisch-Waugh-Lovell Theorem: Nonparametric Case

This is where the FWL theorem reveals its modern relevance. When we apply regularisation methods like LASSO to select which variables from a high-dimensional X to include, we cannot simply run a penalised regression of Y on (d, X) and read off the coefficient on d; the regularisation bias contaminates our causal estimate. The automatic isolation of variation that makes FWL work in the linear case breaks down under regularisation because the penalty term does not respect orthogonality required for causal identification; it shrinks all coefficients toward zero based on predictive considerations rather than preserving the specific variation needed to identify τ.

What we can do is manually implement the FWL logic: first use regularisation to select the relationship between X and Y, then between X and d, and finally estimate τ from the residualised regression. This "double selection" or "double machine learning" approach maintains the conceptual clarity of FWL while adapting it to a high-dimensional setting.

The key to understanding modern causal machine learning lies in recognising that the points of departure are fundamentally interconnected. The CEF provides the unifying mathematical framework that links prediction and causal inference. The parametric versus nonparametric choice determines how we estimate this CEF, whether through restrictive but interpretable linear models or flexible but complex algorithmic approaches. High-dimensional methods become necessary when the covariate space is too rich for traditional approaches, whether because we observe many features directly or because capturing the true nonlinear CEF requires considering many transformations and interactions.

Applications include

Causal Forest Estimation of Heterogeneous Household Response

Time-Of-Use Electricity Pricing Schemes

The Impact of Institutions on Growth

Finance - including models for predicting financial crises and explaining post–earnings-announcement drift.

Software

Throughout the course we will use Stata, R and `Python` to provide examples of ML methods.

We demonstrate the use of R and Python for both Random Forests and Generalised Random Forests.

Computing Resources

Google Colaboratory (Colab) is a cloud-based Jupyter notebook computing environment that requires no setup. Colab is a free tool offered by Google Research that allows users to write and execute Python code in their web browsers. Colab is based on Jupyter open source and allows you to create and share computation files without having to download or install anything.

Since Google Colab is built on top of vanilla Jupyter Notebook, which is built on top of Python kernel, Colab provides a large number of pre-installed machine learning libraries such as Keras, TensorFlow, and PyTorch.

Whilst the Anaconda platform allows you to use the computing resources of your local machine, Colab provides access to powerful cloud-based computing resources, which is machine-independent. Colab also provides, for a fee, access to GPUs and TPUs, which can be useful for running computationally intensive tasks.

Stata

Although it is generally recognised that both R and `Python` have greater overall functionality, Stata plays a central role in implementing cutting-edge machine learning methods for causal inference. The course leverages several key Stata packages and capabilities:

Regularized Regression and Variable Selection

- `LASSOPACK` - Comprehensive lasso, ridge, and elastic net estimation with cross-validation
- `PDSLASSO` - Specialized for causal inference in structural models, facilitating control variable and instrument selection from large sets of variables

Both packages are available through `ssc install package`.[1]

Double Machine Learning Framework

- `DDML` - Implementation of double/debiased machine learning methods with the same functionality as the R equivalent
- Supports five different model types: binary or continuous treatment variables, endogeneity, high-dimensional controls and/or instrumental variables
- Integration with `pystacked` for ensemble methods and stacked generalization via Python's scikit-learn
- Short-stacking capabilities for optimal weighting of multiple learners based on out-of-sample performance

Stata-Python Integration for Machine Learning

Two specialized Stata commands `r_ml_stata_cv` and `c_ml_stata_cv` make use of the Stata/Python integration interface. Developed by Cerulli (2022), these commands provide wrapping Python functions for fitting popular ML methods in both regression and classification settings.

Key Readings

[1] J. Freidman, T. Hastie, R. Tibshirani. *The Elements of Statistical Learning.* Springer, 2009.

[2] G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, 2013.

[3] M. Huber. *Impact Evaluation and Causal Machine Learning with Applications in R. MIT Press*, 2023.

[4] V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, V. Syrgkanis. *Applied Causal Inference Powered by ML and AI. Online*, 2024.

[5] J. D. Angrist, J.-S. Pischke *Mostly Harmless Econometrics.* Princeton University Press, 2008.

[6] J. Wooldridge *Econometric Analysis of Cross-Section and Panel Data.* MIT Press, 2010.

[1] For the latest stable versions, see the Stata Lasso page: `https://statalasso.github.io/`

Material will be taken from the following list of sessions. Given time constraints some of the material may be summarised

Sessions

- Session 1 Introduction
- Session 2 The Best Predictor and The Conditional Expectation Function
- Session 3 Estimation and Inference for Causal Effects
- Session 4 High Dimensional Methods for Linear Models
- Session 5 Applications of Regularised Regression for Linear Models
- Session 6 Double Machine Learning
- Session 7 Treatment Effects and Double Robust Estimators
- Session 8 Random Forests
- Session 9 The Architecture of Causal Trees and Generalised Random Forests
- Session 10 Generalised Causal Forests
- Session 10b Testing for Heterogeneity
- Session 11 An Introduction to Generative AI and Large Language Models (time permitting)

Session 1 Introduction

1. High-level overview of Causal Machine Learning and ai
2. Points of Departure
    - The Conditional Expectation Function and Linear Estimators
    - High Dimensional Methods in Statistics
    - Frisch-Waugh-Lovell Theorem
    - Treatment Effects and Causal Inference
    - Machine Learning Methods for Prediction
3. Machine Learning: The Vernacular
4. Applications
5. Computing

Session 2 The Best Predictor and The Conditional Expectation Function

Course Notes: *Prediction and Evaluation*

1. The Conditional Expectation (cef) Function
2. Best Predictor Problem
3. Linear Regression, Nonlinear in Variables
4. Parametric and Non-parametric models
5. Local Estimation, Kernel Estimators and Decision Trees
6. Departing From The Linear cef: Wages and Education

Session 3 Estimation and Inference for Causal Effects

1. Causal Effects: Beyond Prediction and Attribution
    - Schooling, Ability and Wages
    - An Introduction to Treatment Effect Model
    - Anticipating Heterogeneous Causal Effects
    - Weak Instruments in Nonparametric Settings
    - Binary Response and Endogeneity

Session 4 High Dimensional Methods for Linear Models

1. High Dimensional Methods
2. Least absolute Shrinkage and Selection (lasso)
    i Choosing λ
    ii Causal Inference in High-Dimensions
    iii lasso For Treatment Models
    iv Double lasso
    Double Section and Orthogonalisation

Session 5 Applications of Regularised Regression for Linear Models

Session 6 Double Machine Learning

1. The Partial Linear Model (plm)
2. Regularisation Bias Revisited
3. Orthogonalisation and Sample Fitting
4. A Decomposition of Bias
5. Variations on the plm
6. Stata Practical: The Impact of Institutions on Growth

Session 7: Treatment Effects and Double Robust Etimators

1. Treatment Effects: ate and cate
2. Identification Strategies
3. Estimators
4. Regression Adjustment
5. Inverse propensity score weighting
6. aipw and Double Robust Estimators
7. Double Robust Estimators Meet Machine Learning
8. Application:
    - Job Training on Earnings

Session 8 Random Forests

1. Machine Learning and Decision Trees
    i Machine Learning: Terminology and Concepts
    ii An Overview of Regression Trees
    iii The Bias-Variance Tradeoff revisited
    iv Training, Testing and Cross Validation
    v Regularization: Variance reduction and Ensemble Learning
2. Practical: Machine Learning for Prediction

Session 9: The Architecture of Causal Trees and Generalised Random Forests

1. Why not use Off-the Shelf Methods for Prediction?
2. From Random to Causal Forests
3. Infeasible mse
4. Sample Splitting
5. Honest Estimation
6. From Hard Partitioning to Forest-Based Weights

Session 10 Generalised Random Forests

1. Generalised Random Forests
2. Application:
    - Instrumental Variables Random Forests

      The Impact of Fertility on Labour Supply

Session 10b Testing for Heterogeneity

1. Chernozhukov et al. *Generic Machine Learning Inference on Heterogenous Treatment Effects*
2. Application:
    - The Impact of tou Tariffs on Energy Demand

Session 11: An Introduction to Generative ai and Large Language Models (Time Permitting)

1. Concepts and Definitions

- What is AI
- What is Generative AI
- Generative AI as a Co-pilot
- What are Large Language Models
- The Technology Stack

2. Language and Semantic Understanding for Machines
   - Natural Language Processing
   - Context and the Context Window

   Transformer: Attention is all you need!
   - Transformer architecture
   - Generating with Claude

3. Another Perspective: Generation and Retrieval
   Search versus Reason Engines
4. A Comparison of LLMs
5. Types of Learning

END

Applications

Session 1: *Overview*

- Lasso Low Dimensions
  Code: `DS300_Topic_1_Sim_LowDim_PostSingle.ipynb` $\surd$
- Predict house prices
  Code: `DS300_SingleTree_R.ipynb` $\surd$
  Code: `Table1_R.ipynb`
  Data: `ahs2011forjep.rdata`

Session 2: *Best Linear Predictor and the* CEF

- Wages and Gender
  Code: `DS300_Topic_2_WagesGender.ipynb` $\surd$ closer look
- Frisch-Waugh-Lovell Theorem
  Code: `DS300_fwl_Theorem.do` $\surd$
  Code: `DS300 Debiasing-with-Orthogonalization_LinearQs_A.ipynb` X
- Birth weight and smoking
  Data: `BWGHT.dta`, `Bwght.csv` X

Session 3a: *Prediction Estimation and Attribution* [**Not 2026**]

- Surface Plus Noise Models
  `DS300_Topic3a_RegSurf.ipynb` $\surd$
  Regression Tree Surface
  `DS300_Topic_3a_RTreeSurface.ipynb` $\surd$
- Predict house prices
  Code: `DS300_SingleTree_R.ipynb` $\surd$ ANACONDA
  Code: `Table1_R.ipynb`
  Data: `ahs2011forjep.rdata`

Session 3: *Causal Effects: Beyond Prediction and Attribution*

- Children and Their Parents Labor Supply
  Code: `Ds300_Endog_BinRespBvps` $\surd$
  Data: `pums80m.dta`
- The Demand for Cigarettes [**Not Instats**]
  Code: `Ds300_Topic 3b_IV` - R
  Code: `Ds300_Topic 3b_IV` - STATA
  Data: `cig_data.dat` and `xls`

Session 4: *High Dimensional Methods for Linear Models*

- Wages and Gender: Lasso

  Code:`DS300_Topic_5_WagesGender_Lasso.ipynb` CHECK

  Code:`DS300_Topic_5_WagesGenderL.ipynb` X

- Double (Linear) Lasso

  Code:`DS300_Topic_5_DoubleL.ipynb` BUG?

Session 5: *Applications of Regularised Regression*

- Predicting Boston house prices

  Code: `Ds300_Lasso_HousePrices_New` $\surd$

  Data: `Ahrens.dta`

- Institutions and Growth - Lasso for Linear model

  Code: `Ds300_Lasso_Institutions` $\surd$

  Data: `AJR.dta`

  Data: `https://raw.githubusercontent.com/statalasso/statalasso.github.io/master/dta/housing.csv`

Session x: *Treatment Effects* [**Not Instats**]

- Impact of Job Training on Earnings

  Simple Regression Adjustment and IPW estimators

  Code: `Cate_JobTtrain.do`

  Data: `jtrain2.dta`

  Session 6: *Double Machine Learning*

  Double-debiased ML in Stata

  - Double Debiased Ensemble Estimators - Stata

    Code: `DS300_ddmlStacked.do`

  - Double-debiased ML - R

    Code: `DS300_DoubleDebiased_Institutions.ipynb` BUG SE?

  Session 7: *Regression Trees and Forests*

  - Predict house prices using Random Forests

    Code: `DS300_Comp_Session_RF_HousePrice_v2.do` $\surd$

    Data: `tnewhouse_Modified.csv` - web version in code

Session 8b: *Generalised Random Forests*

Heterogeneous treatment effects using DDML

Time of Use Tariffs and Smart Meter Data

- Fertility and Labour supply

  Code: `DS300_grf_maternal_employment_run2.ipynb` √

  Code: `DS300_grf_maternal_employment_run2.r`

- Impact of Microcedit

  Crepon et al.. 's (2015) study on the effects of microcredits

  Code:: `TestHet_GenML_Morocco.ipynb`

Session 11: *Applications to Finance* [**Not Instats**]

- Credit Card Default

  `Credit_Card_Default_FullModel.do`, `Credit_Card_Tune.do` (**Review**)

- Forecasting Financial Crises with Classification Tree Ensembles and many predictors (r code for replication)

  Spotting the Danger Zone

- Explaining the Post-Earnings Announcement Drift (r code for replication) **No Code**

- The Impact of Digital Footprints on Credit Markets and Financial Inclusion

- Determinants of corporate cash holdings: An application of a robust variable selection technique

Session 12: *Machine Learning and Classification* [**Not Instats**]

- Fertility and Labour supply (See Session 9)

  Children and Their Parents Labor Supply: Evidence from Exogenous Variation in Family Size – Angrist and Evans (1998)

  `Prob_Bvp_2sls.do`, data: `pums80m.dta`

  See Juypter Notebook R file: `grf_maternal_employment_run2`

- Credit Card Default

  `Credit_Card_Default_FullModel.do`, `Credit_Card_Tune.do`

- Surviving the Titanic

  `titanic.do`, `titanic-data-science-solutions.ipynb`

Session 12: *Natural Language Processing* [**Not Instats**]

- Sentiment Analysis: `Sentiment Analysis - Predicting Tesla Stock Price with Article Headlines.ipynb`

- Central Bank Communication: `Finbert.ipynb`

- ChatGPT as a LLM - within Colab

Readings

Machine Learning: Overview

[1] L. Breiman, J. Freidman, R. Olshen, C. Stone. *Classification and Regression Trees.* Klein-Verlag, 1990.

[2] L. Breiman, J. Freidman, R. Olshen, C. Stone. *Random Forests.* Machine Learning, 45(1):5-32 1990.

[3] *Random Forests.* `https://en.wikipedia.org/wiki/Random_forest.`

[4] *Training, Validation, and Test sets.* `https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets`

[5] J. Freidman, T. Hastie, R. Tibshirani. *The Elements of Statistical Learning.* Springer, 2009.

[6] G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, 2013.

[7] S. Russell, P. Norvig Artificial Intelligence: A Modern Approach *3rd edition, 2009*

Machine Learning: Software

[1] G. Cerulli *Machine Learning using Stata/Python.* The Stata Journal (2022) 22, Number 4, pp. 772–810.

[2] G. Cerulli *Improving econometric prediction by machine learning. Applied Economics Letters*, 28, 16, 1419-1425,

Nonparametrics

[1] E. Nadaraya *On Estimating Regression.* Theory of Probability and Its Applications 9(1): 141-2

[2] G. Watson *Smooth regression analysis. Sankhyā: The Indian Journal of Statistics,Series A 26 (4): 359-72.*

Neural Networks

[1] H. Xu,K. Kinfu,A. W. Levine *When are Deep Networks really better than Decision Forests at small sample sizes, and how?* arXiv:2108.13637, 2021.

Machine Learning and Econometrics

[1] L. Breiman *Statistical Modeling: The Two Cultures* Statistical Science, Vol. 16, No. 3. pp. 199-215

[2] S. Athey *The Impact of Machine Learning on Economics.* in, The Economics of Artificial Intelligence: An Agenda, 2018. National Bureau of Economic Research. See `http://bit.ly/2EENtvy` S. Athey,G. Imbens *Machine Learning Methods Economists Should Know About.* Working Paper, 2019, Graduate School of Business, Stanford University.

[3] S. Mullainathan, J. Spiess. *Machine Learning: An Applied Econometric Approach* Journal of Economic Perspectives vol. 31, 2017, pp. 87-106.

[4] A. Belloni, V. Chernozhukov, C. Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives, 28(2):29-50, 2014(a)*

Policy Problems

[1] Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015), *Prediction policy Problems*. American Economic Review, 105(5): 491-495.

[2] Andini, M., Ciani, C., Blasio, G. and D'Ignazio, A. (2018). *Effective Policy Targeting Machine Learning* `https://voxeu.org/article/effective-policy-targeting-machine-learning`

[3] Andini, M., Ciani, C., Blasio, G., D'Ignazio, A. and Paladini, A. (2018). *Machine learning in the service of policy targeting: the case of public credit guarantees.* Banca d'Italia Temi di discussione.

[4] Chalfin, A, O. Danieli, Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., and S. Mullainathan (2016), *Productivity and selection of human capital with machine learning*, American Economic Review, 106(5): 124-127.

[5] Athey, S. (2017). *Beyond prediction: Using big data for policy problems*. Science 03, Vol. 355, pp. 483-485

Treatment Effect Models

[1] Angrist, J.D. and J.-S. Pischke, (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

[2] Joshua D. Angrist (2004) *Treatment Effect Heterogeneity in Theory and Practice*. The Economic Journal, Vol. 114.

[3] LaLonde, R.J. 1986. *Evaluating the econometric evaulations of training programs with experimental data*. American Economic Review, Vol.76, No.4, pp. 604-620.

[4] Dehejia, R., and Wahba, S. (1999). *Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs*, Journal of the American Statistical Association, Vol. 94, No. 448, pp. 1053-1062.

[5] Blundell, R.W., and M. Costa Dias (2002): *Alternative Approaches to Evaluation in Empirical Microeconomics.* Portuguese Economic Journal, 1, 91-115. `http://cemmap.ifs.org.uk/wps/cwp0210.pdf`

[6] Imbens, G., and J. Wooldridge (2009): *Recent Developments in the Econometrics of Program Evaluation.* Journal of Economic Literature, 47(1): pp. 5-86.

[7] Athey, S., and G. Imbens (2017). *The State of Applied Econometrics: Causality and Policy Evaluation.* Journal of Economic Perspectives, 31 (2): 3-32.

[8] Chib, S. and B.H. Hamilton, (2000), Bayesian analysis of cross-section and clustered data treatment models, *Journal of Econometrics*, 97(1), 25-50

[9] Munkin, M.K. and P.K. Trivedi, (2003), Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare, *Journal of Econometrics*, 114(2), 197-220

[10] Li, M, and J. Tobias, (200*), Bayesian analysis of Treatment Effects in an Ordered Potential Outcomes Mode1, Advances in Econometris,

LASSO

[1] Efron, B., Hastie, T., Johnstone, I. and R. Tibshirani, (2004). Least angle regression (with discussion), *Annals of Statistics* 32(2): 407- 499

[2] Friedman, Jerome; Hastie, Trevor; Tibshirani, Robert (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition (Springer Series in Statistics) (Kindle Locations 13024-13026). Springer - A. Kindle Edition.

[3] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pages 267-288).

[4] Hofmarcher, P., Cuaresma, J., Grun, B., and K. Hornik (2015). Last Night a Shrinkage Saved My Life: Economic Growth, Model Uncertainty and Correlated Regressors, *Journal of Forecasting*, Vol. 34, pages 133-144

[5] Park, T., and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

[6] O'Hara, R. B., & Sillanpaa, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1), 85-117.

[7] Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2), 369-411.

[8] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273-282.

[9] Belloni, A., Chernozhukov, V. and C. Hansen (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects, Journal of Economic Perspectives, Vol. 28, no. 2 pp. 29-50,

[10] Belloni, A., Chernozhukov, V. and C. Hansen (2014). Inference on Treatment Effects after Selection amongst High- Dimensional Controls. Review of Economic Studies ...

[11] Imai, K., and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443-470.

Machine Learning for Causal Inference

[1] S. Athey, G. Imbens. Recursive partitioning for heterogeneous causal effects *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[2] E. O'Neill, M. Weeks. Causal Tree Estimation of Heterogeneous Household Response to Time-Of-Use Electricity Pricing Schemes arXiv:1810.09179v1, 2018.

[3] S. Athey, G. Imbens, Y. Kong, V. Ramachandra. An Introduction to Recursive Partitioning for Heterogeneous Causal Effects Estimation Using causalTree package. https://github.com/susanathey/causalTree/blob/master/briefintro.pdf, 2016.

[4] S. Athey, S. Wager. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017.

[5] Y. Lin, J. Yongho. Random forests and adaptive nearest neighbors *Technical Report No. 1055. University of Wisconsin, 2002* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.9168

[6] S. Athey, G. Imbens, S. Wager(201x). *Estimating Average Treatment Effects: Supplementary Analyses*

[7] V. Chernozhukov Double/debiased machine learning for treatment and structural parameters *The Econometrics Journal*, 21 (1) pp. C1-C68.

[8] A. Belloni, V. Chernozhukov, C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies, 81(2):608-650, 2014(b).*

[9] A. Belloni, V. Chernozhukov, I. Fernandez-Val,C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica, 85(1): 233-298, 2017.*

[10] V. Chernozhukov, I. Fernandez-Val, M. Demirer, E. Duflo. Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. *arXiv preprint arXiv:1712.04802.*

[11] T. Christiansen, M. Weeks. Heterogeneous Impacts of Microcredit Expansions: An Analysis of Three Randomised Experiments Using Machine Learning.' *Cambridge Working Papers in Economics, 2020*

[12] L. Breiman Random Forests *Machine Learning, Vol. 45, pp.5-32*

[13] *Random Forests.* https://en.wikipedia.org/wiki/Random_forest.

[14] *Training, Validation, and Test sets.* https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets

Natural Language Process for Economists

Ellingsen, J., Larsen, V. H., & Thorsrud, L. A. (2022). News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics*, 37(1), 63-81.

Zhai, S., & Zhang, Z. (2022). Read the News, Not the Books: Forecasting Firms' Long-term Financial Performance via Deep Text Mining. ACM *Transactions on Management Information Systems* (TMIS).

Lutz, B., Pröllochs, N., & Neumann, D. (2022). Are longer reviews always more helpful? Disentangling the interplay between review length and line of argumentation. *Journal of Business Research*, 144, 888-901.

Mansouri, S., & Momtaz, P. P. (2022). Financing sustainable entrepreneurship: ESG measurement, valuation, and performance. *Journal of Business Venturing*, 37(6), 106258.

Ferrario, B., & Stantcheva, S. (2022, May). Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions. In AEA *Papers and Proceedings* (Vol. 112, pp. 163-69).

Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: machine learning vs. dictionary methods. Management Science, 68(7), 5514-5532.

Caldara, Dario and Matteo Iacoviello (2022). Measuring Geopolitical Risk, *International Finance Discussion Papers* 1222r1. Washington: Board of Governors of the Federal Reserve System, `https://doi.org/10.17016/IFDP.2022.1222r1`

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022). Making text count: economic forecasting using newspaper text. *Journal of Applied Econometrics*, 37(5), 896-919.

Maximilian Ahrens and Michael McMahon. 2021. Extracting Economic Signals from Central Bank Speeches. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 93–114, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Davis, S. J., Hansen, S., & Seminario-Amez, C. (2020). Firm-level risk exposures and stock returns in the wake of COVID-19 (No. w27867). *National Bureau of Economic Research.*

Kong, N., Dulleck, U., Jaffe, A. B., Sun, S., & Vajjala, S. (2020). Linguistic metrics for patent disclosure: Evidence from university versus corporate patents (No. w27803). *National Bureau of Economic Research.*

Baylis, P. (2020). Temperature and temperament: Evidence from Twitter. Journal of Public Economics, 184, 104161.

Ros, R., van Erp, M., Rijpma, A., & Zijdeman, R. (2020). Mining Wages in Nineteenth-Century Job Advertisements. The Application of Language Resources and Language Technology to study Economic and Social Inequality. In *Proceedings of the Workshop about Language Resources for the* SSH Cloud (pp. 27-32).

Moreno-Ortiz, A., Fernandez-Cruz, J., & Hernández, C. P. C. (2020). Design and Evaluation of SentiEcon: a fine-grained Economic/Financial Sentiment Lexicon from a Corpus of Business News. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5065-5072).

Masson, C., & Paroubek, P. (2020). NLP analytics in finance with dore: A french 250m tokens corpus of corporate annual reports. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 2261-2267).

Qin, Y., & Yang, Y. (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 390-401).

Zamani, M., & Schwartz, H. A. (2017, April). Using twitter language to predict the real estate market. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers (pp. 28-33).

Lefever, E., & Hoste, V. (2016, May). A classification-based approach to economic event detection in dutch news text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (pp. 330-335).

Jelveh, Z., Kogut, B., & Naidu, S. (2014, October). Detecting latent ideology in expert text: Evidence from academic papers in economics. In P*roceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1804-1809).

Takala, P., Malo, P., Sinha, A., & Ahlgren, O. (2014, May). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. In LREC (Vol. 2014, pp. 2152-2157).

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3), 221-247.

Ghose, A., Ipeirotis, P., & Sundararajan, A. (2007, June). Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistic*s (pp. 416-423).

Brekke, M., Innselset, K., Kristiansen, M., & Ovsthus, K. (2006, May). Automatic Term Extraction from Knowledge Bank of Economics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC'06).