Econ 148 -- Data Science for Economists -- Summer 2023

Unit Value: 4 units

Tamer Çetin

Lecture 1 - Introduction

Course Description

• Data Science for Economists is a unique field in data analysis.
• While it refers to core Data Science analysis on one hand, it introduces a still-emerging interdisciplinary approach to applied data analysis on the other hand.
• This approach has been built where Machine Learning algorithms and economic applications meet.
• For that reason, in this course we will basically cover four main sections in data analysis: data exploration, regression analysis, prediction, and causal analysis.
• The purpose is to gain a newly developing and strong analytical tool in applied data analysis.
• To this aim, we will examine how to analyze all types of datasets using both Data Science and Machine Learning-based Econometrics since over 80% of data today is only applicable to Data Science-based analysis, but not the traditional econometric models.
• On the other hand, causal analysis still matters since Data Science-based applied analysis focuses on the prediction of expected value of conditional y on conditioning x while causal inference econometrics is interested in the effect of x on y.
• In order to provide more reliable results in data analysis, Data Science for Economists will help students uncover such relationships between y and x.
• After completing this course, students will:

- better understand how to generate and prepare real-world data for applied analysis using extremely useful Data Science techniques,
- be using Data Science tools from exploratory data analysis to Machine Learning algorithms and their usage in Econometric models in both predictive and causal analyses.

Course Description

- **Prerequisites**

  DATA C8\COMPSCI C8\INFO C8\STAT C8

  or

  Stat 20 + familiarity with Python recommended

- **Lectures**
  - There will be 10 hours of lecture per week - 4x2.5-hour lectures for six weeks.
  - When appropriate these lectures will include data explorations and live coding (mostly on Thursdays).
  - Use of math will be informative and illustrative.

Like this

$$y^E = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta_0 + \sum_{j=1}^{p} \beta_j - x_j$$

$$\hat{y}_j = \hat{f}(x_j)$$

$$L(e_j) = e_j^2 = (\hat{y}_j - y_j)^2$$

$$MSE = \frac{1}{j}\sum_{j=1}^{j}(\hat{y}_j - y_j)^2$$

$$= \left(\frac{1}{j}\sum_{j=1}^{j}(\hat{y}_j - y_j)\right)^2 + \frac{1}{j}\sum_{j=1}^{j}(y_j - \bar{\hat{y}})^2$$

Course Description: Assignments

- Data Science is about analyzing real-world data sets, and so a series of projects involving real data are a required part of the course.
- Students may work with a single partner on each individual project/homework, and you are allowed to work with a partner in your data/code challenge section.
- Each student must submit each homework independently, but you are allowed to discuss problems with other folks.
  - *Biweekly Individual Project/Homework:* Four biweekly homework/project will be done using Python, when necessary, individually.
  - *Final:* The final exam might be a final project that would be carried out as a group project based on a reproduction analysis of a published journal article. We will talk about that!

Course Description: Projects

- Project 1 – Selected data project - import, clean, describe, and explore data
  - Data Preparation
  - Data Exploration
  - Feature Engineering
  - Missing Values, Outliers, and Sparsity
  - EDA for Numerical and Categorical Variables
  - Visualizing Distributions, Frequencies, and Probabilities
- Project 2 - Selected data prediction project – EDA and Linear Regression Prediction
-
- Project 3 - Code to Cleaning to Visualization to Outputs in Classification and Clustering Models
- Project 4 – Selected Topics

Course Description: Grading

- Biweekly Individual Homework (4): 40% - These will be due on the last class day of the week at 5pm a week after they are released. Basically, you will have one week to submit your homework after you receive it.
- Midterms (2): 40% - Midterms will be on the last class days of the second and fourth weeks.
- Final: 20% - Final will be on the last class day of the last week.
- **Late Policy**
  - Students are allowed to submit data/code challenge and homework late for a 50% penalty until the next Thursday at 5 pm after the predetermined submission deadline for homework.

Disabled Students Policy

- UC Berkeley is committed to creating a learning environment that meets the needs of its diverse student body including students with disabilities. If you anticipate or experience any barriers to learning in this course, please feel welcome to discuss your concerns with me.
- If you have a disability, or think you may have a disability, you can work with the Disabled Students' Program (DSP) to request an official accommodation. The Disabled Students' Program (DSP) is the campus office responsible for authorizing disability-related academic accommodations in cooperation with the students themselves and their instructors. You can find more information about DSP, including contact information and the application process here: dsp.berkeley.edu. If you have already been approved for accommodations through DSP, please meet with me so we can develop an implementation plan together.
- Students who need academic accommodations or have questions about their accommodations should contact DSP, located at 260 César Chávez Student Center. Students may call 642-0518 (voice), 642-6376 (TTY), or e-mail dsp@berkelely.edu(link sends e-mail)(link sends e-mail).

Learning outcomes

- Learning outcomes will be:

  a hands-on knowledge of how datasets are created and explored using the most common data analysis tools,

  and an understanding of how data can be used to answer economic research or business questions using the appropriate data analysis techniques.
- This course is based on the elements of the data science life cycle; from data exploration and feature engineering to formulating questions, visualization, and modeling.
- By looking at a variety of datasets, students will be exposed to a broad range of Data Science applications to economic questions.
- Students will get a combination of skills and tools needed to be a successful data scientist, researcher, analyst, and/or (applied) economist.

How to Make it Work?

- We will analyse data with the tools, methods, and skills needed to answer data-focused, real-life questions; to carry out empirical analysis; and to visualize and interpret results to support better decisions in business, economics, and public policy.
- Data wrangling and exploration, regression analysis, machine learning, and causal analysis are comprehensively covered, as well as when, why, and how the methods work, and how they relate to each other.
- As the most effective way to communicate data analysis, running case studies play a central role in this class.
- Each case starts with an industry-relevant question and answers it by using real-world data and applying the tools and methods covered in the lectures.

Materials & Resources

- **Books**
- Bekes, G. and G. Kezdi, 2021. *Data Analysis for Business, Economics, and Policy,* Cambridge University Press.
- Cleff, T., 2014. *Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel,* Springer.
- Hastie, T., R. Tibshirani, and J. Friedman, 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Second Ed., Springer.
- James, G., D. Witten, T. Hastie, and R. Tibshirani, 2021. *An Introduction to Statistical Learning: with Applications in R*, Second Ed., Springer.
- Kuhn, M. and K. Johnson, 2013. *Applied Predictive Modeling*, Springer.
- Vanderplas, J., 2017. *Python for Data Analysis*, O'Reilly.
- **Suggested Readings**
- Athey, S. 2021. 'The Impact of Machine Learning on Economics', Eds. A. Agrawal, J. Gans, and A. Goldfarb, *The Economics of Artificial Intelligence,* Ch. 21, University of Chicago Press, 507-552.
- Breiman, L., 2001. 'Statistical Modeling: The Two Cultures', *Statistical Science*, 16(3), 199-231.
- Ludwig, J, and S. Mullainathan, 2021. 'Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System', *Journal of Economic Perspectives*, 35(4), 71–96.
- Mullainathan, S. and J. Spiess, 2017. 'Machine Learning: An Applied Econometric Approach', *Journal of Economic Perspectives*, 31(2), 87–106.
- Wager, S. and S. Athey, 2019. 'Machine Learning Methods That Economists Should Know About', *Annual Review of Economics,* 11, 685-725.
-

Materials & Resources

- **Further Readings**

- Chan, F. and L. Matyas, 2022. *Econometrics with Machine Learning*, Springer.
- Semenova, V. and V. Chernozhukov, 2021. 'Debiased Machine Learning for Conditional Average Treatment Effects and Other Causal Functions', *The Econometrics Journal*, 24(2), 264-289.
- Semenova, V., 2023. 'Debiased Machine Learning of Set-Identified Linear Models'. *Journal of Econometrics*,
- Wager, S. and S Athey, 2018. 'Estimation and Inference of Heterogeneous Treatment Effects using Random Forests', *Journal of the American Statistical Association,* 113, 1228-1242.

Course Outline

- SECTION I: INTRODUCTION TO DATA SCIENCE: DATA EXPLORATION
  - Technology for Data Science
  - Origins of Data
  - Big Data, Statistical Inference, and External Validity
  - Data Wrangling: Preparing Raw Data for Analysis
  - Exploratory Data Analysis and Feature Engineering
  - Model Selection, Evaluation, and Validation
  - Regularization, Parameters, and Hyperparameters

Course Outline

- SECTION II: REGRESSION
  - Testing Hypotheses
  - Comparison and Correlation
  - Introduction to Regression
  - Linear Regression in ML vs Linear Regression in Econometrics (with application in Python)
  - Time Series Forecasting (with application in Python)
  - Predicting Probabilities
  - Logistic Regression in ML vs Logistic Regression in Econometrics (with application in Python)

Course Outline

- SECTION III: PREDICTION: ML ALGORITHMS
  - Supervised and Unsupervised Models
  - Prediction
  - Regression
  - Classification
  - Clustering
  - Experimental Design, Reinforcement Learning, and Multi-Armed Bandits

Course Outline

- SECTION IV: CAUSALITY IN DATA ANALYSIS
  - Data Science and Machine Learning
  - Data Science and Economics
  - Econometrics and ML
  - Causality, Intervention, and Variation
  - The Setup: Intervention, Treatment, Subjects, and Outcomes
  - Potential Outcomes Framework
  - Causal Maps (DAGs) to Uncover Causal Structure
  - Controlled Experiments
  - Randomized Experiments
- Guest Talks!

Course Outline

- DATA EXPLORATION
- REGRESSION ANALYSIS
- PREDICTION
- CAUSAL ANALYSIS

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests -- Stefan Wager and Susan Athey, *Journal of the American Statistical Association,* 2018, 113, 523, 1228-1242

- **Abstract:** Many scientific and engineering challenges—ranging from personalized medicine to customized marketing recommendations—require an understanding of ***treatment effect heterogeneity***. In this article, we develop ***a nonparametric causal forest*** for ***estimating*** heterogeneous treatment effects that extends Breiman's widely used ***random forest algorithm***. In the ***potential outcomes framework*** with ***unconfoundedness***, we show that causal forests are pointwise consistent for the true treatment effect and have an ***asymptotically Gaussian*** and ***centered sampling distribution***. We also discuss a practical method for constructing ***asymptotic confidence intervals*** for the true treatment effect that are centered at the causal forest estimates. Our theoretical results rely on a generic Gaussian theory for a large family of random forest algorithms. To our knowledge, this is the first set of results that allows any type of random forest, including ***classification and regression forests***, to be used for provably ***valid statistical inference***. In ***experiments***, we find causal forests to be substantially more powerful than classical methods based on ***nearest-neighbor matching***, especially in the ***presence of irrelevant covariates***.

Big Data, Data Science, and Machine Learning

- Data Science: Google what data science is, it is a big confusion!
- Why Big Confusion?
  - Interdisciplinary interactions
  - Overlaps
  - Blurry borders
  - Lack of background in statistics
- Simple Definition of Data Science
  - The transformation of raw data-based information to measurable inference
  - Measurable means predictable in data science!
- Two components of data science
  - Data
  - Prediction

Data

- What is data?
  - Almost everything!
    - Being born, eating, drinking, studying, purchasing, selling, driving, liking, hating, loving, traveling, ageing, and dying!
- When is it data?
  - When it reports the past, describes the current, and predicts the future:
    - Quantifying: Formalization of information to interpretable results!

Data

- What to do using data
  - Basically, gathering information about the events
    - To formalize the past
    - To describe the current
    - To predict the future
  - Detecting anomalies
  - Predicting the events in the future
  - Revealing the reasons for the current/past events
- This is basically what data science is!

A General Workflow for Analysis in Data Science

- Generating Data
  - Collection and storage of data
- Data Preparation
  - Structuring and organizing data
- Data Visualization or Exploratory Data Analysis (EDA)
  - Inferential statistics
- Modeling data
  - Experimentation vs Prediction vs Forecast vs Estimation

A General Workflow for Analysis in Data Science

- Storing data
  - Location
    - where to store data
      - Big data ----> Impossible to store it in single computer -----> companies own storages such as ‘clusters’ or ‘servers’ or cloud storages such as Microsoft Azure, Amazon Web Services (AWS), and Google Could
  - Data type
    - Need to know what kind of data we need to store: Different data types require different storage solutions
    - Document (qualitative) database storage for unstructured data such as text, email, videos, pictures, web pages, and social media
    - Relational (quantitative) database storage for structured datasets such as tabular formats

Data type	Storage type	Query language
Unstructured (text/qualitative)	Document database	Non-SQL
Structured (tabular/quantitative)	Relational database	SQL

  - Extracting data
    - How to retrieve data from the storage

A General Workflow for Analysis in Data Science

- Data preparation
- Why?
  - Real-life data is messy and dirty leading to
    - Errors
    - Deceptive results
    - Bias in ML algorithms
- How?
  - Make data readable
  - Remove outliers
  - Drop/impute missing values
  - Handle categorical and numerical variables separately

EDA in General: Goals of EDA

- Size of data
- Properties of features
- Properties of target variable
- Underlying patterns for feature engineering

EDA in General

- **Univariate non-graphical**
  - This is simplest form of data analysis, where the data being analyzed consists of just one variable.
  - Since it's a single variable, it doesn't deal with causes or relationships.
  - The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- **Univariate graphical**
  - Non-graphical methods don't provide a full picture of the data.
  - Graphical methods are therefore required.
  - Common types of univariate graphics include:
    - Stem-and-leaf plots, which show all data values and the shape of the distribution.
    - Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
    - Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.

EDA in General

- **Multivariate nongraphical**
  - Multivariate data arises from more than one variable.
  - Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
- **Multivariate graphical**
  - Multivariate data uses graphics to display relationships between two or more sets of data.
  - The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

EDA: Common Visualization Tools

- Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.
- Multivariate chart, which is a graphical representation of the relationships between factors and a response.
- Run chart, which is a line graph of data plotted over time.
- Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.
- Heat map, which is a graphical representation of data where values are depicted by color.

Data sampling

- Expectation is a population concept.
- In practice, data usually come in the form of samples and rarely consist of an entire population.
- We therefore use samples to make inferences about the population.
- For example, the sample conditional expectation function is used to learn about the population.

Roles in Data Science

- Data Engineer
- Data Analyst
- Data Scientist
- Machine Learning Scientist/Engineer
    - Researcher
    - Research Scientist
    - Applied Research Scientist
    - Applied ML Engineer

Roles in Data Science

- Data engineer
  - Controls the flow of data by
    - Constructing data pipelines (reliable)
    - Storage systems (reproducible)
  - Focused on data generation mostly
  - Proficient in SQL to collect, store, and organize data
  - Common to use one of Java, Scala, or Python to process data.
  - Uses cloud computing to handle big data

Roles in Data Science

- Data Analyst
  - Reveals the *present* via data by exploring data and creating visualizations and dashboards
    - Clean data
    - Explore data
    - Visualize data
  - Less programming and statistics experience needed
  - Data preparation and exploration and visualization in workflow of data analysis
  - Uses SQL to query data
  - As data engineers build and configure SQL storage solutions, data analyst uses existing databases to retrieve and aggregate data relevant to their analysis
  - Data analysts use spreadsheets to perform simple analyses on small quantities of data
  - Analysts also use Business Intelligence, or BI Tools, such as Tableau, Power BI, or Looker, to create dashboards and share their analyses
  - More advanced data analysts may be comfortable with Python or R for cleaning and analyzing data

Roles in Data Science

- Data scientist
  - Strong background in statistics to find new insights from data rather than solely describing data
  - They also use traditional machine learning for prediction and forecasting
  - Within the workflow, they focus on the last three stages: data preparation and exploration and visualization, and experimentation and prediction
  - They must be proficient in one of Python or R
  - Using these programs, they use popular data science libraries, such as pandas and matplotlib to contain reusable code for common data science tasks

Roles in Data Science

- Machine learning scientist/engineer
  - Similar to data scientists with more specialization in modeling using ML algorithms
  - The most exciting role in data science because they reveal the truth by testing hypotheses
  - For instance;
    - Revelation of the future earnings in stock markets using predictive models
    - Or, revelation of masked identifiers in data using unsupervised ML algorithms
  - This role includes the last three stages with a strong focus on prediction in the workflow of data analysis
  - Uses either Python or R to create their predictive models

Technology

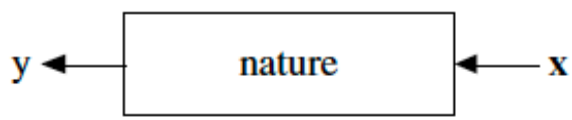
- Anaconda Navigator ---> Jupyter Notebook ---> Python
- Alternatively datahub.berkeley.edu ---> select lab
  - R, Stata, and SQL
  - Code will be provided for all the analysis in Stata and R if you are interested in those packages
- GitHub
- BCourses

Big Picture

- Statistics
- Data Science
- Machine Learning
- Economics
- Econometrics

Data Science in Statistics (Breiman, 2001)

- Statistics starts with data.
- Think of the data as being generated by a black box in which a vector of input variables x (independent variables) go in one side, and on the other side the response variables y come out.
- Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



Data Science in Statistics (Breiman, 2001)

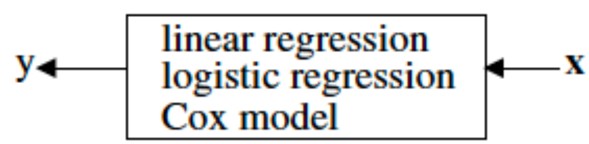
- There are two goals in analyzing data:
  - Information
    - To extract some information about how nature is associating response variables to input variables
  - Prediction
    - To be able to predict what responses are going to be to future input variables
  - Estimation???

Data Science in Statistics (Breiman, 2001)

- Keep in mind the following:
  - Data Science from the Perspective of Statistics
    - Data Modelling
      - Assuming data are generated by a given stochastic data model
    - Algorithmic Modelling
      - Treating data as unknown

Data Science in Statistics (Breiman, 2001)

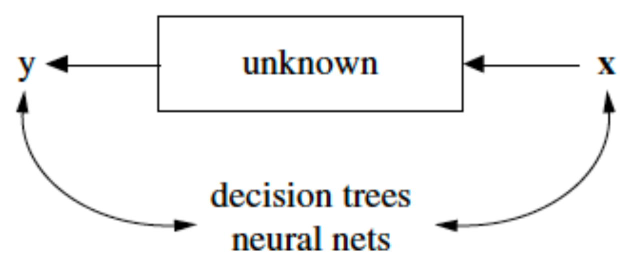
- Data Modeling Culture
  - The analysis in this culture starts with assuming a stochastic data model for inside of the black box.
    - For example, a common data model is that data are generated by independent draws from
      - $response\ variables\ =\ f(predictor\ variables, random\ noise, parameters)$
  - The values of parameters are estimated from data and the model then is used for information and/or prediction.
  - Thus, the black box is filled in like this:



  - Model validation: Using goodness-of-fit tests and residual examination such as R^2 and P *values.*
  - Estimated culture population: 98% of all statisticians.

Data Science in Statistics (Breiman, 2001)

- Algorithmic Modeling Culture
  - The analysis in this culture considers inside of the box complex and unknown.
  - Their approach is to find a function $f(x)$ — an algorithm that operates on x to predict the response y.
  - Their black box looks like this:



  - Model validation: Measured by predictive accuracy (scores)

Data Science in Statistics (Breiman, 2001)

- Traditionally, statistical community has been committed to the exclusive use of data models.
- This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.
- Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics.
- It has been used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets.
- If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Data Science in Statistics (Breiman, 2001)

- Note that the main difference between data science and statistics still refers to the same difference between econometrics/economics and data science!!
- However, statistics community has by and large accepted Machine Learning (ML) revolution that Breiman refers to as the algorithm modeling culture, and many textbooks in statistics discuss ML methods alongside more traditional statistical methods.
- However, the adoption of these methods in economics has been slower, they are now beginning to be widely used in empirical work and are the topic of a rapidly increasing methodological literature (Athey and Imbens, 2019).

Econometrics and Machine Learning (Athey and Imbens, 2019)

- In this sense, economic/econometric research also should use data science or ML;
  1. If our goal as a field in economics is to use data to solve problems,
  2. If researchers in economics want to communicate effectively with researchers in other fields where these methods are routinely being adopted.
- Why has the acceptance of ML methods been so much slower in economics compared to the broader statistics community?
  - Economics journals emphasize the use of methods with formal properties of a type that many of the ML methods do not naturally deliver
  - The adaptation of causal inference in business questions in industry
  - Data is mostly not usable for economic applications

Econometrics and Machine Learning (Athey and Imbens, 2019)

- The focus of the traditional approach in econometrics is to specify a target, an estimand, that is a functional of a joint distribution of data.
- Given a random sample from the population of interest the parameter of interest and the parameters are estimated by finding the parameter values that best fit the full sample, using an objective function such as the sum of squared errors, or the likelihood function.
- The focus is on the quality of the estimators of the target, traditionally measured through large sample efficiency.
- Researchers typically report point estimates and standard errors.

Econometrics and Machine Learning

- In contrast, in the ML literature the focus is typically on developing algorithms.
- The goal for the algorithms is typically to make predictions about some variables given others, or classify units on the basis of limited information.
- Supervised machine learning algorithms seek functions that predict well out of sample.

Econometrics and Machine Learning

- ML uses terminology for concepts that have well established labels in the older literatures.
- In the context of a regression model the sample used to estimate the parameters is often referred to as the training sample.
- Instead of estimating the model, it is being trained.
- Prediction problems are divided into
    - supervised learning problems where we observe both the predictors/features x_i and the outcome y_i,
    - and unsupervised learning problems where we only observe the x_i and try to group them into clusters or otherwise estimate their joint distribution.
- Unordered discrete response problems are generally referred to as classification problems.

Econometrics and Machine Learning

- Many economic applications, instead, revolve around parameter estimation: produce good estimates of parameters β that underlie the relationship between y and x.
- It is important to recognize that machine learning algorithms are not built for this purpose.
- x: Covariates, estimators, predictors, independent variables, features, attributes, or descriptors used as input for the prediction equation.
- y: Outcome, dependent variable, target, class, response refer to the outcome event or quantity that is being predicted.

Econometrics and Machine Learning

- Main difference between econometrics and ML:
  - In Econometrics, the goal is to understand what happens (to y) if I hold everything $(x_n - x_1)$ fixed and change one variable (x_1) (***estimation***).
  - In ML, the goal is to understand what is the future values of y *(**prediction**).*
    - *What is the best prediction of y?*
  - In ML, we do that to minimize the mean squared error in a new dataset in which we only have x variables observable.
  - Before that, I have dataset with y and x to build a model and after getting new x, we can predict what y will be using the estimated model earlier.

Econometrics and Machine Learning

- In ML, ML algorithms work well because you can accurately evaluate model performance without making a lot of additional assumptions (non-parametric)
  - All you have to see how the model fits since the only goal is to fit
  - You do not need to argue additional assumptions
  - Maybe only other algorithms with better performance including less bias or less errors
- ML starts in a non-functional form

Limitations of ML and Econometrics (Athen, 2022)

- ML society argues that ML is everything!
- A ML person could ask you if economists even use data!
- You will see some ML studies arguing that they are answering questions or solving problems but actually they do not!
  - Every question is answerable or every problem is solvable!
  - You just need good algorithms!
  - Basically we have a great hammer and everything is nail!
- The main reason for this deceptive understanding in ML world is that they think that predictive models are strong enough to answer questions.
- So they basically ignore whether it is:
  - causality
  - equilibrium
  - And/or feedback effects

Limitations of ML and Econometrics (Athen, 2022)

- Is it completely wrong? NO!
- Actually, it is true that because as a general-purpose technology predictive models is extremely applicable to the real-world business problems.
- It is also applicable to economics not only economic problems but economics itself as a discipline in terms of scientific research.
- But still that does not mean ML is the best/only tool to answer such questions.
- This is why there is big divergence between economics/econometrics society and ML world.

Limitations of ML and Econometrics (Athen, 2022)

- Also, it is not entirely true that all ML scientists or engineers only rely on standard predictive models.
- Even it is minority in the field, there are some concerns and efforts about interpretability, robustness, and fairness of the models and results.
- But, again, ML projects commonly ignore such questions.

Limitations of ML and Econometrics (Athen, 2022)

- Econometrics is BLUE!
  - OLS estimator is Best Linear Unbiased Estimator
- Econometrics starts with a functional form
  - Economics tells us the best estimation function in model identification
- In the existence of high dimensional data, there is no useful mechanisms to handle dimensionality curse