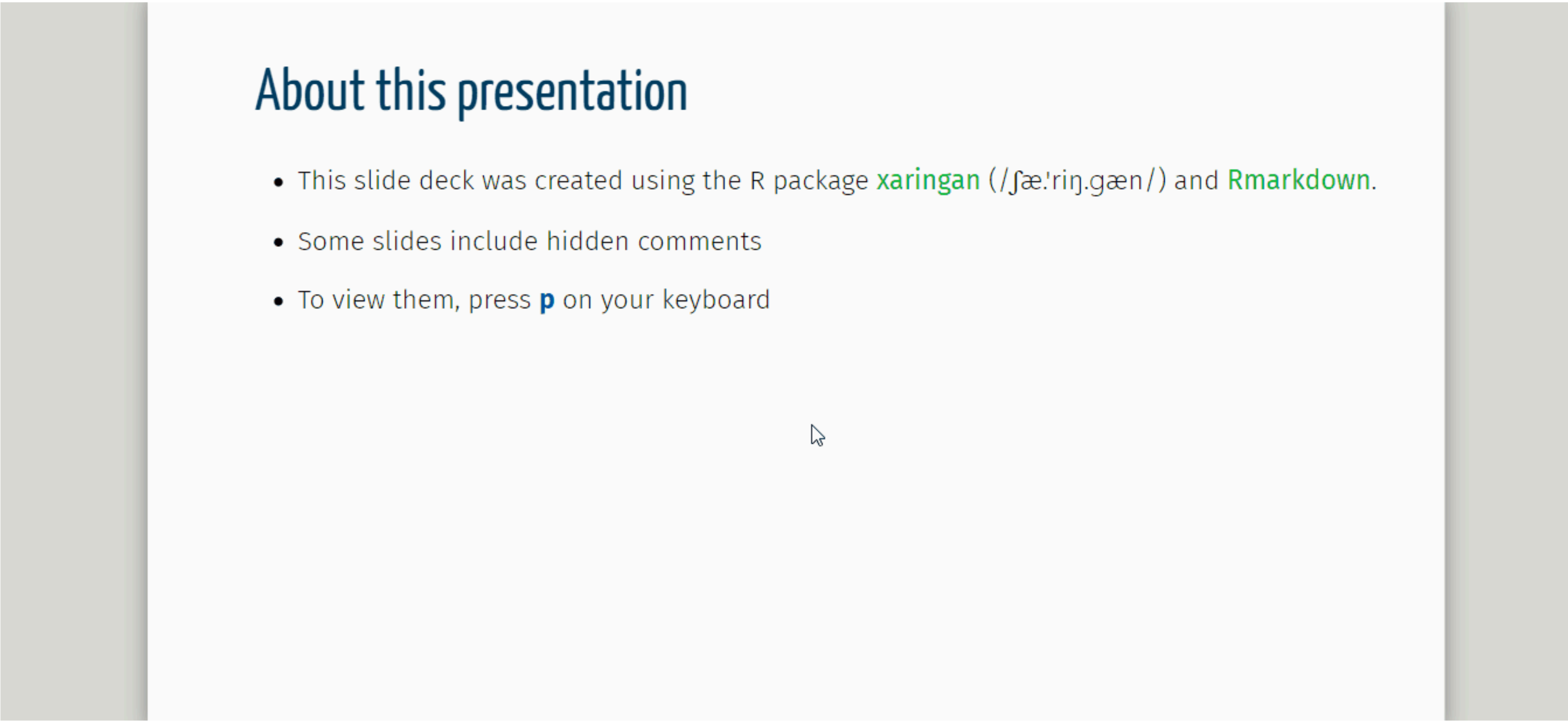
01 - Course Overview

ml4econ, HUJI 2024

Itamar Caspi
May 5, 2024 (updated: 2024-05-05)

An aside: about the structure of these slides

- The course's slide decks are created using the xaringan (/ʃæ.'riŋ.gæn/) R package and Rmarkdown.
- Some slides include hidden comments. To view them, press **p** on your keyboard



About this presentation

- This slide deck was created using the R package xaringan (/ʃæ.'riŋ.gæn/) and Rmarkdown.
- Some slides include hidden comments
- To view them, press **p** on your keyboard

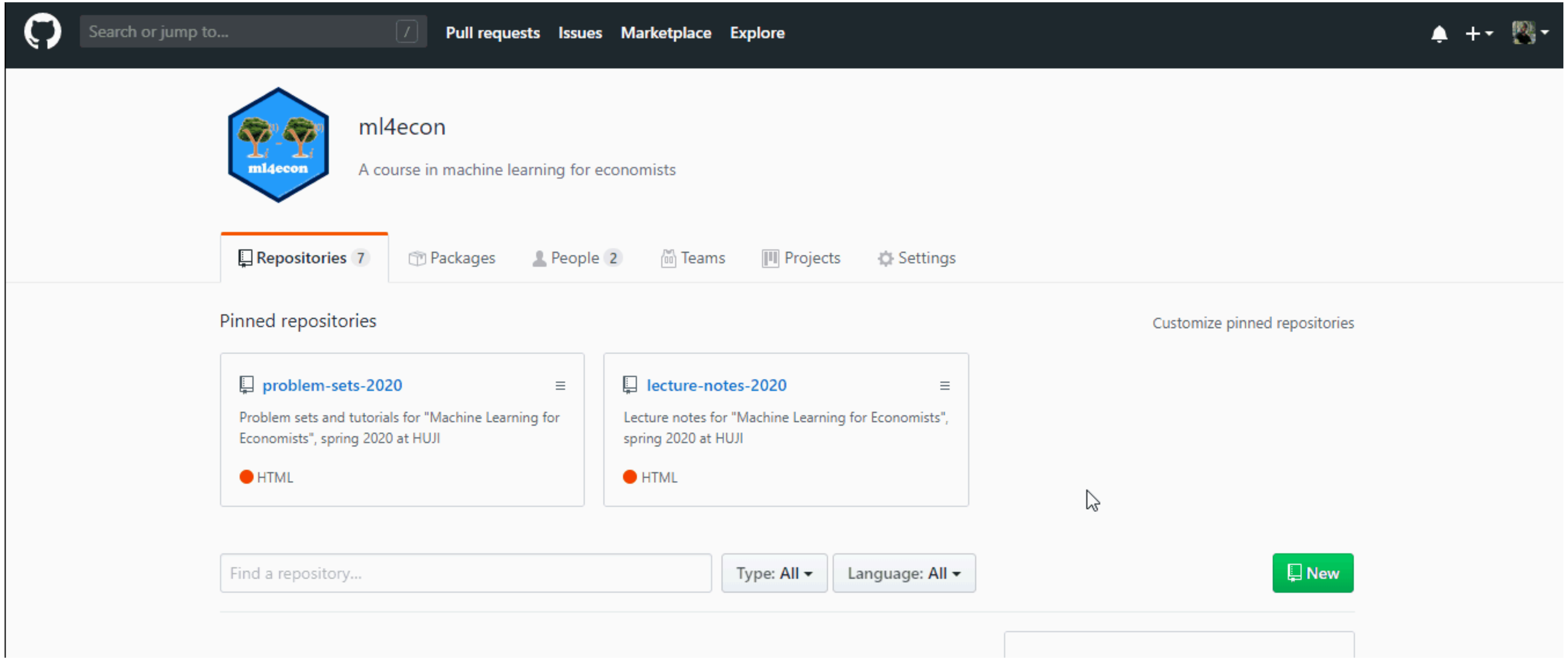
Outline

1. Logistics
2. About the Course
3. To Do List

Logistics

ml4econ GitHub repository

The class's GitHub repository: https://github.com/ml4econ



ml4econ

A course in machine learning for economists

Pinned repositories

problem-sets-2020

Problem sets and tutorials for "Machine Learning for Economists", spring 2020 at HUJI

● HTML

lecture-notes-2020

Lecture notes for "Machine Learning for Economists", spring 2020 at HUJI

● HTML

Posit Cloud workspace

Posit Cloud is a hosted version of RStudio in the cloud that will make it easy for R and Python novices to learn data science and machine learning using R and Python.

People

- **Itamar Caspi**
    - email: caspi.itamar@gmail.com
    - homepage: itamarcaspi.rbind.io
- **Inbar Avni (TA)**
    - email: TBA
- Meeting hours: after class/zoom, on demand.

Feedback

Your continuous feedback is important!

Please feel free to contact us by

- email
- in person
- or open an issue in our discussion forum

About the Course

Prerequisites

- Advanced course in econometrics.
- Some early experience with R (or another programming language) are a plus.

This course is

About

How and when to apply ML methods in economics

- estimate treatment effects.
- prediction policy.
- work with new types of data (e.g., text).

To do that we will need to understand

- what is ML?
- how it relates to stuff you already know?
- how it differs?

Not about

- Cutting-edge ML techniques (e.g., generative AI)

- Computational aspects (e.g., gradient descent)

- Data wrangling (a.k.a. "feature engineering")

- Distributed file systems (e.g., Hadoop, Spark)

Tentative schedule

Week	Topic
1	Course Overview & ML Basics
2	Reproducibility and ML Workflow
3	Regression and Regularization
4	Classification
5	Non-parametrics
6	Unsupervised Learning
7	Text analysis
8	Causal Inference
9	Lasso and Average Treatment Effects
10	Trees and Heterogeneous Treatment Effects
11	Prediction Policy Problems
12	Large Language Models

> **NOTE**: This schedule can (and probably will) go through changes!

Readings on ML for economists

> All materials and lecture notes will be available on the class website.

Please read the following excellent surveys:

- The impact of machine learning on economics Athey (2018)
  In *The Economics of Artificial Intelligence: An Agenda*.
  University of Chicago Press.

- Machine learning: an applied econometric approach Mullainathan and Spiess (2017)
  Journal of Economic Perspectives, 31(2), 87-106.

Readings on ML

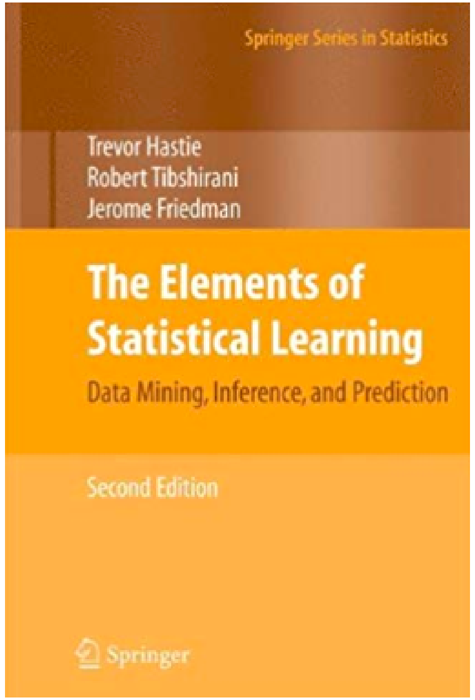
> All materials and lecture notes will be available on the course repo.

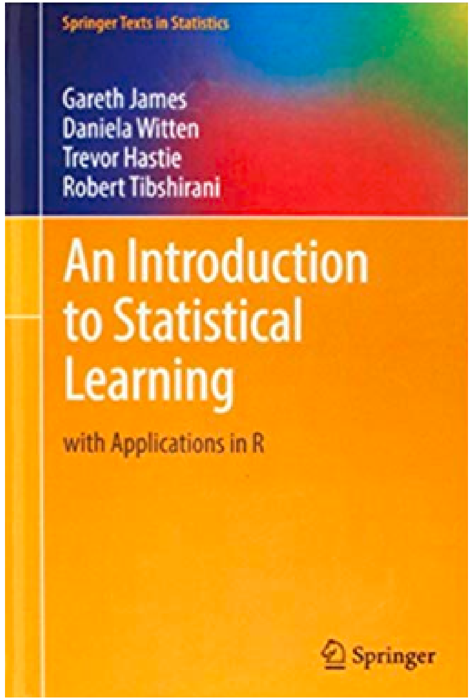
There are **no** required textbooks.

A couple of suggestions:

- An Introduction to Statistical Learning with Applications in R/Python (ISLR), 2 ed.
  James, Hastie, Witten et al. (2013)
  PDF available online

- The Elements of Statistical Learning (ELS)
  Hastie, Tibshirani, and Friedman (2009)
  PDF available online


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer



Textbooks (optional)

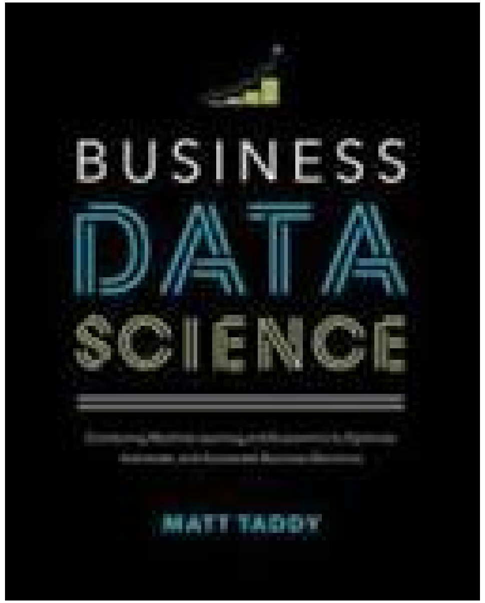
> All materials and lecture notes will be available on the course repo.

There are **no** required textbooks.

A couple of suggestions:

- Business Data Science by Matt Taddy
  No free version available

- Econometrics by Bruce Hansen, Ch. 29
  PDF available online



More resources

Can be found at our GitHub repo:

https://github.com/ml4econ/lecture-notes-2024/blob/master/resources.md

Programming

- Two of the most popular open-source programming languages for data science:
  - R
  - Python
- This course: Mostly R.
- Why R? See presentation notes and the FAQ section of our class website.
- We do encourage you to try out Python. However, I will only be able to provide limited support for Python users. Inbar on the other hand, will be able to provide more support.

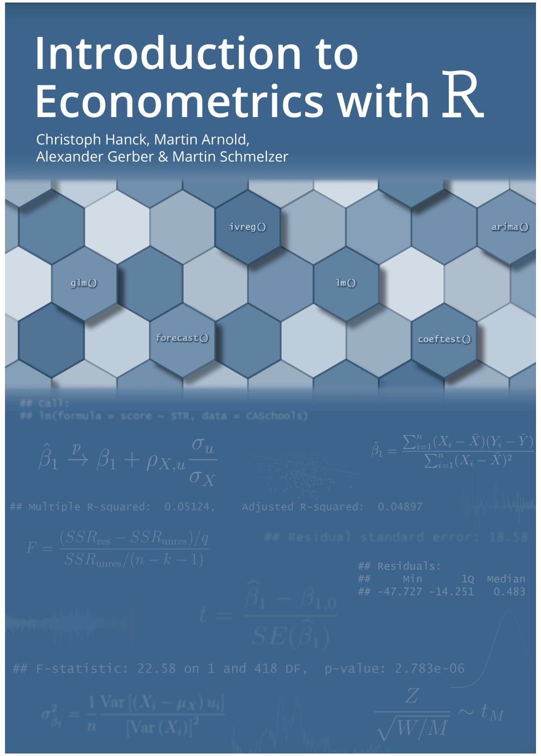
Catching up with R

ChatGPT is a great tool to learn R.

Posit Recipes is a great (old school) resource to learn R.

Econometrics with R

Introduction to Econometrics with R (Hanck, Arnold, Gerber, and Schmelzer)


Introduction to Econometrics with R
Christoph Hanck, Martin Arnold,
Alexander Gerber & Martin Schmelzer
ivreg()
arima()
glm()
lm()
forecast()
coeftest()

Large Language Models (LLMs)

We encourage you to use ChatGPT, Claude, or any other LLM in this course, as it is an **essential skill to acquire**.

It is important you understand the (current) limitations of LLMs:

- Prompt engineering is necessary for quality outcomes.
- Always assume that it is wrong.
- Acknowledge its use in assignments and explain what prompts were used.

Three useful resources:

- Follow @emollick (Ethan Mollick)
- Read "Generative AI for Economic Research: Use Cases and Implications for Economists" by Korinek (2023 JEL).

Share your discoveries with us and your classmates!

Grading

Assignments:

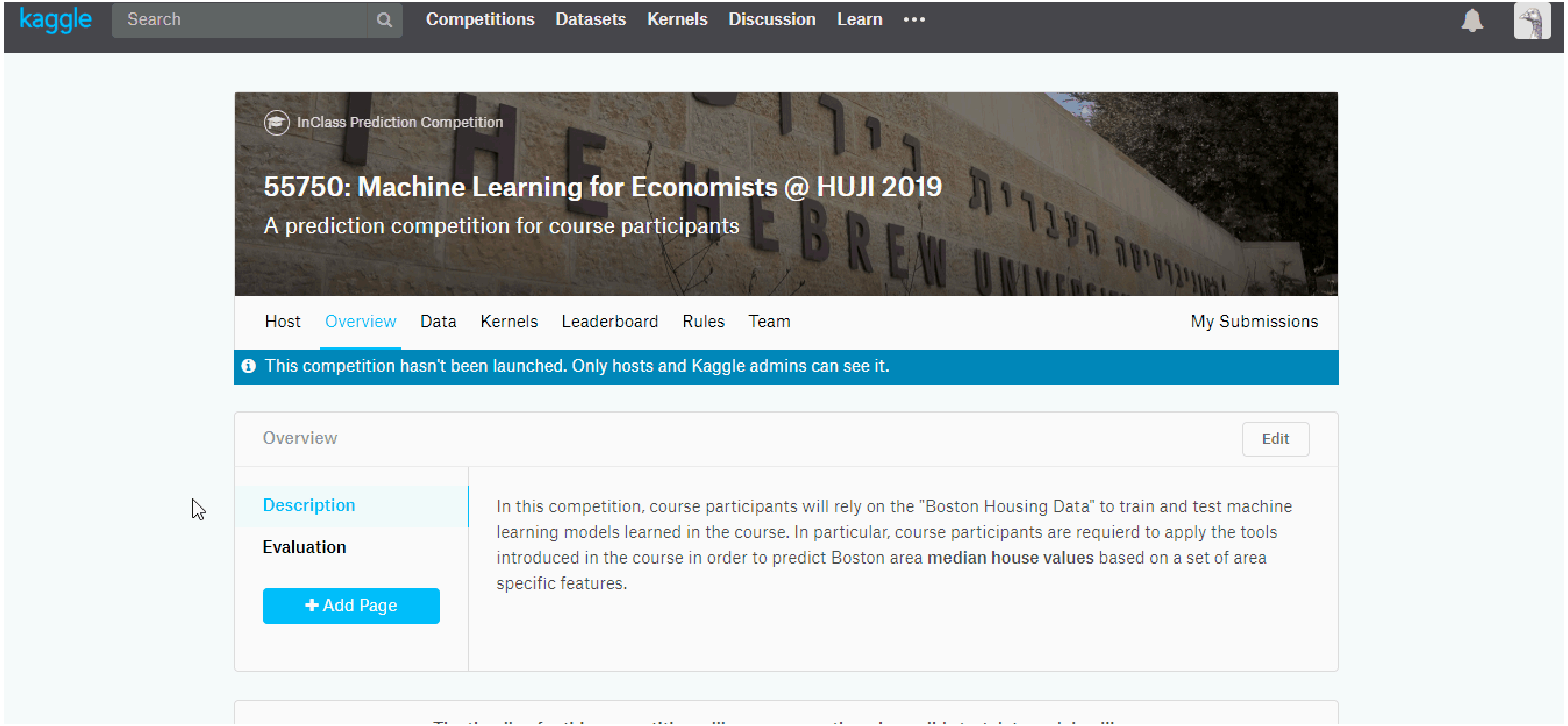
- Submit 4 out of a total of 6 Problem sets.

Two projects:

- Kaggle prediction competition.
- Conduct a replication study based on one of the datasets included in the experimentdatar package, or a paper of your choice.

GRADING: Assingments **20%**, kaggle **30%**, project **50%**.

Kaggle



55750: Machine Learning for Economists @ HUJI 2019

A prediction competition for course participants

This competition hasn't been launched. Only hosts and Kaggle admins can see it.

In this competition, course participants will rely on the "Boston Housing Data" to train and test machine learning models learned in the course. In particular, course participants are requierd to apply the tools introduced in the course in order to predict Boston area **median house values** based on a set of area specific features.

experimentdatar

We will also make use of he `experimentdatar` data package that contains publicly available datasets that were used in Susan Athey and Guido Imbens' course "Machine Learning and Econometrics" (AEA continuing Education, 2018).

- You can install the **development** version from GitHub

```r
# install.packages("devtools")
devtools::install_github("itamarcaspi/experimentdatar")
```

- **EXAMPLE:** Load the `experimentdatar` package and the `social` dataset:

```r
library(experimentdatar)
data(social)
```

- Tips:
  1. Runnig `?social` privides variable definitions.
  2. Running `dataDetails("social")` will open a link to the paper associated with `social`.

To Do List

Homework

- ☑ Download and install Git.
- ☑ Download and install R and RStudio.
- ☑ Create an account on GitHub
- ☑ Download and install GitHub Desktop.

```
slides %>% end()
```

Source code

References

[1] S. Athey. "The impact of machine learning on economics". In: *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2018.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009 .פבר. ISBN: 9780387848570.

[3] G. James, T. Hastie, D. Witten, et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer London, Limited, 2013. ISBN: 9781461471370.

[4] A. Korinek. "Language Models and Cognitive Automation for Economic Research". In: *NBER Working Paper* 30957 (2023).

[5] S. Mullainathan and J. Spiess. "Machine learning: an applied econometric approach". In: *Journal of Economic Perspectives* 31.2 (2017), pp. 87-106.