Big Data for Public Policy

Introduction

Malka Guillot

ETH Zürich | 860-0033-00L

Zoom rules before we begin

- Turn on video and set audio to mute
- Set zoom name to “Full Name, School, Dept/Major”(ex: “Leon Smith, ETH Computer Science”)
- Say “hi” in the chat

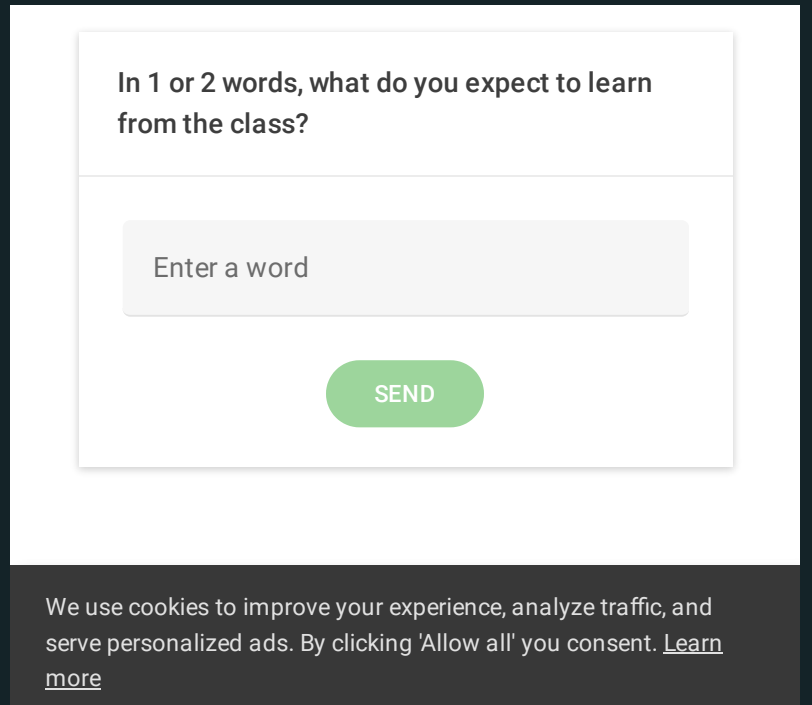
Table of contents

More interaction using

slido

Join at
slido.com
#97107

What do you want to learn during the class?



Prologue:

Machine learning, big data an policy analysis

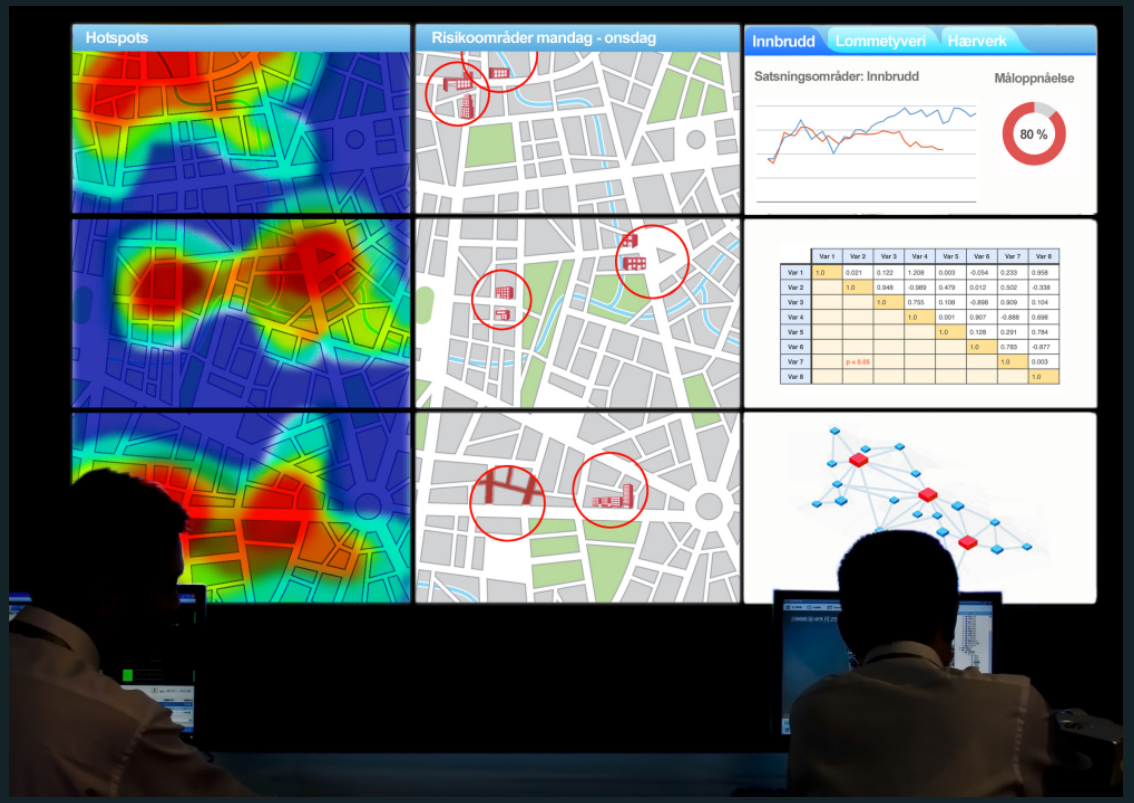
(Big) Data can diagnose (and hopefully help solve) policy problems.

Police discrimination in the US

- **Policy question**:
  - assess racial disparities in policing in the United States
- **Big data**:
  - Analyze a dataset detailing nearly 100 million traffic stops conducted across the country.
- **Methodo**:
  - Use a sunset as a "veil of darkness" masks one’s race
- **Result**:
  - Black drivers were less likely to be stopped after sunset,

(Big) Data can cause (or magnify) problems.

Predictive policing



Predictive policing

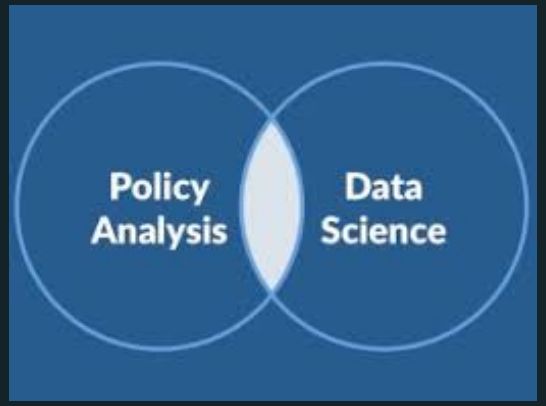
Predictive policing poses discrimination risk, thinktank warns

Machine-learning algorithms could replicate or amplify bias on race, sexuality and age

▲ One officer said human biases including more stop and searches of black men were likely to be introduced into algorithm data sets. Photograph: Carl Court/Getty Images

Welcome

- This course focuses on applications of **big data tools to public policy analysis**



- Goals:
  - Equip you with the standard machine learning toolkit.
  - Put it to work on a real-world policy project.

What this course is, and is not

- It is:
  - Applied and oriented towards practice;
  - General overview of different techniques - what they are and how to use them.
  - Data analysis in general, not restricted to a field (economics, political science).
  - In python.
- It is not:
  - Computer science. We're not coding up models from scratch.

Who am I?

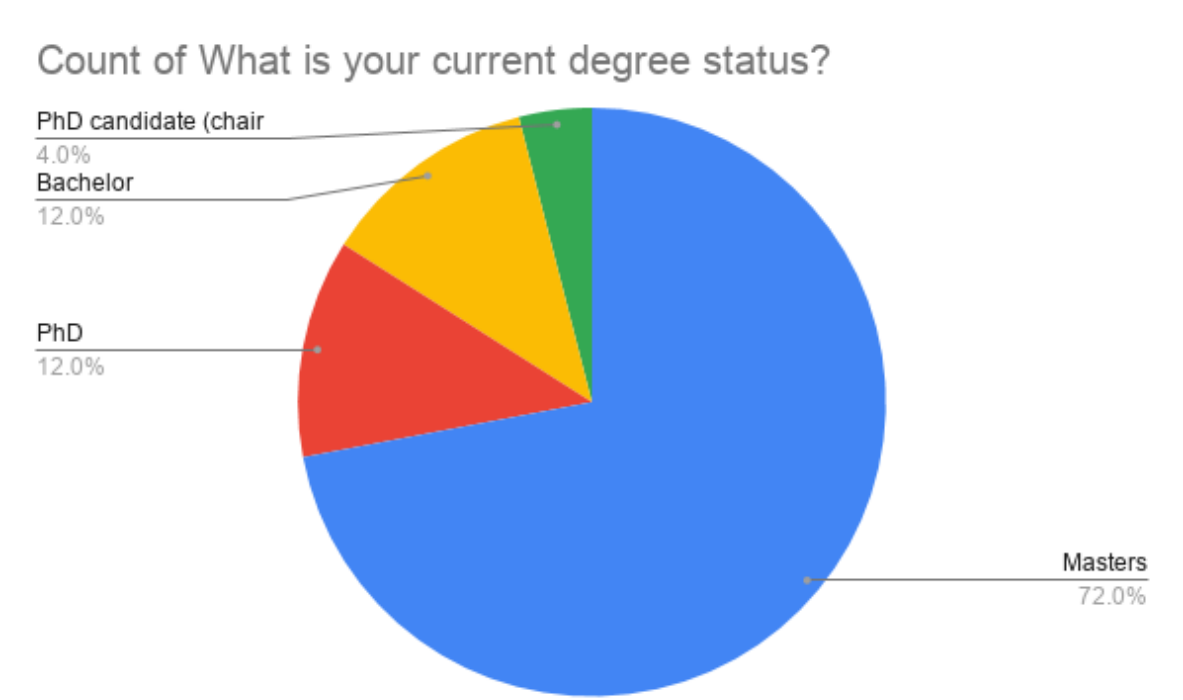
PhD in economics from the Paris School of Economics

Postdoc at ETH

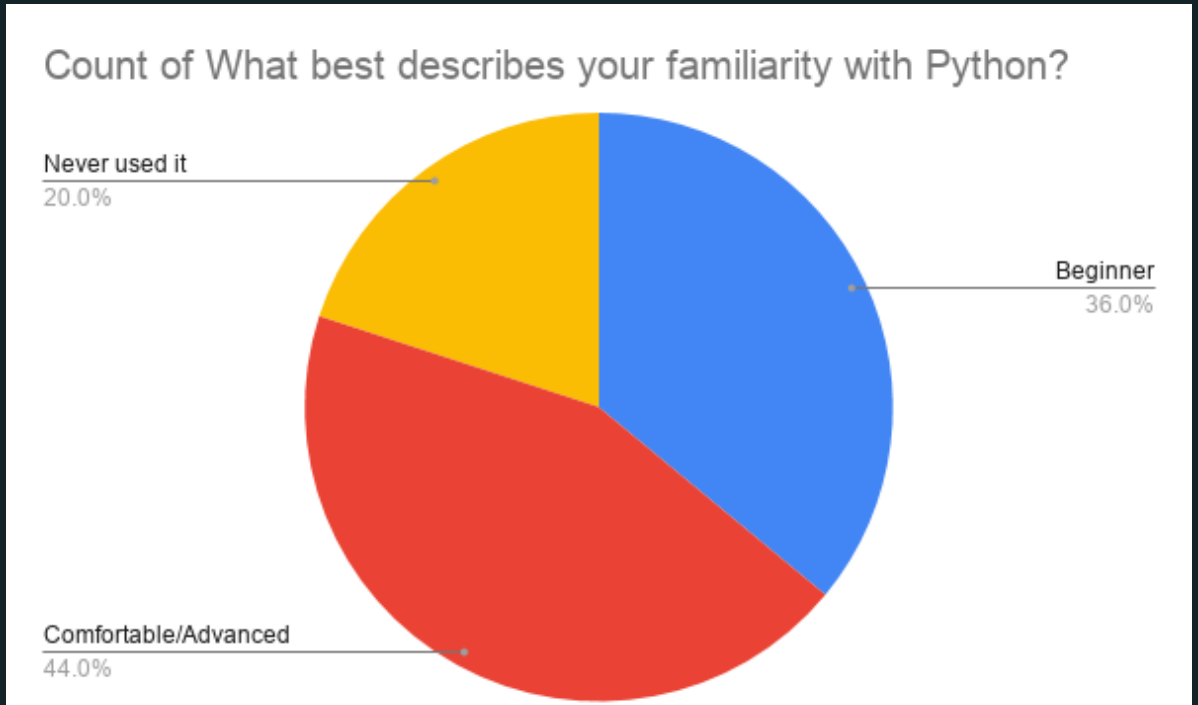
Interested in **public economics** questions: inequality and taxation

Using the standard econometric toolbox + natural language processing + machine learning

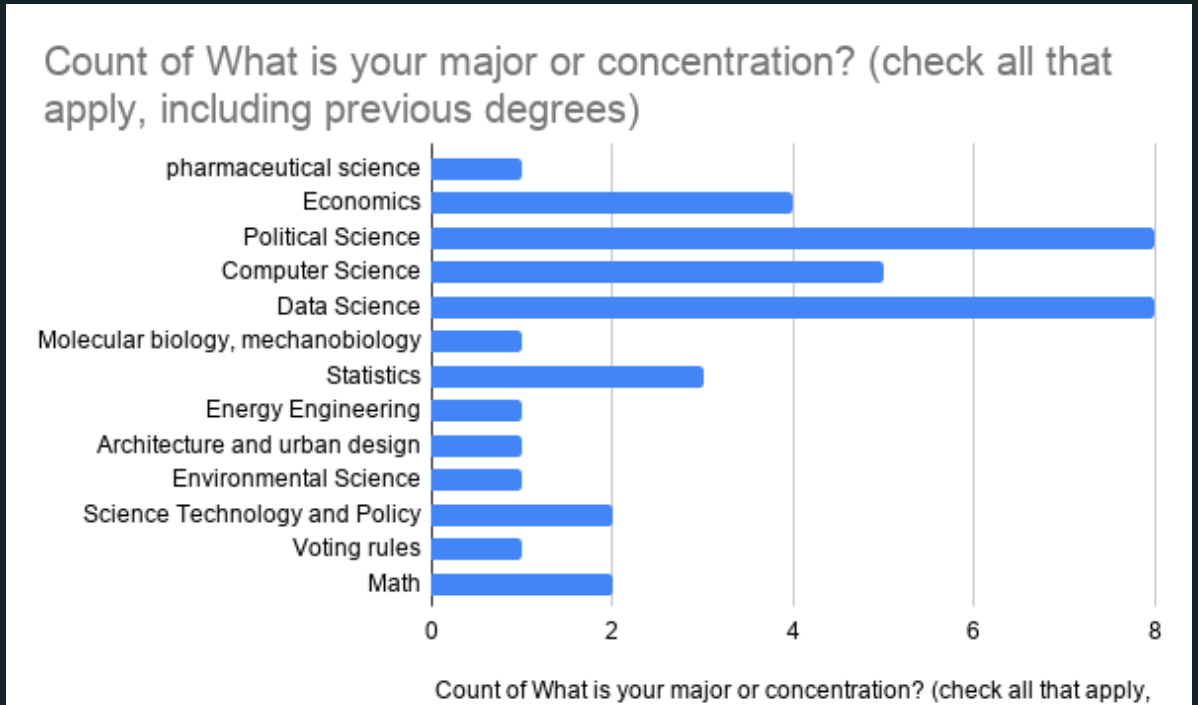
Who are you? Results from pre-class survey



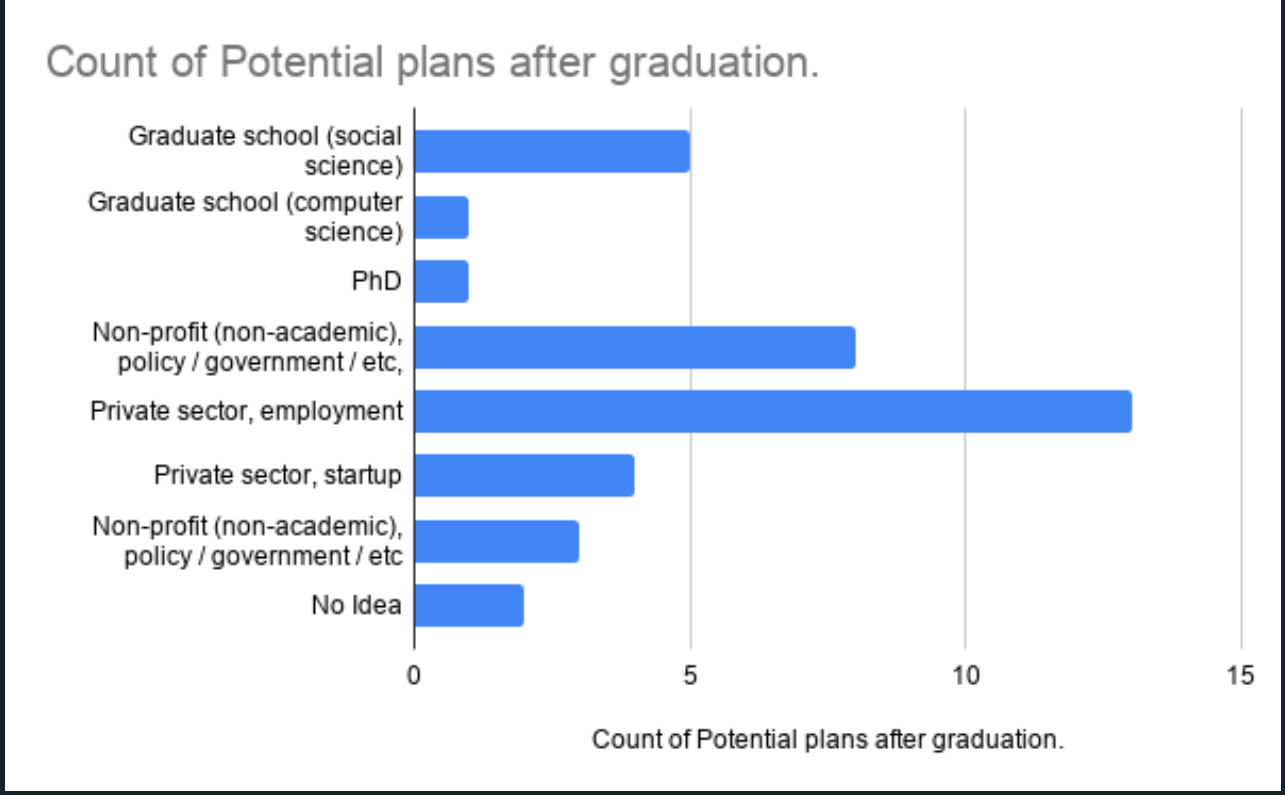
Who are you? Results from pre-class survey



Who are you? Results from pre-class survey



Who are you? Results from pre-class survey



Logistics

How does the class work?

- **Lectures**: 2 hours / week
  - 1 hour theory
  - 1 hour interactive:
    - coding exercise
    - 2 * (15 min students presentations + 10 min of class discussion)
- **Every week**
  - Thursdays 12:15-14 (with a 10 minute break 13-13:10)
  - On zoom: link

Online Course Materials

- Moodle:
  - Course announcement and forum
  - Giving back homerwork
- Syllabus
- Github folder or Github page
  - **Slides**: in html, also available in PDF
    - relying on RevealJS
  - **Coding sessions**: in Jupyter Notebook
    - You can use mybinder in the beginning

Approximative Evaluation Policy

- **Weekly homework**: should be given back as jupyter notebooks in PDF format.
  - $4(hw) * 10 \text{ pts} + 2(hw) * 5 \text{ pts} - 10$ **=40%**
- **Reading =30%**:
  - 1 presentation (2 students) =30%
  - Essay on a paper (1 student) = 30%
- **Participation in class & presentations =5 bonus%**:

Course Communication

- Course communication will be done through eDoz
- I will be available
  - in the zoom 5 minutes early, during the mid-lecture break and after the end of lectures.
  - for 1:1 meetings after the class, just book a 15 minutes slot here.

Online Lecture Norms

- Keep video on **camera on** 📹 [if connection allows]
  - At the beginning/end, when asking questions
  - When discussing papers / coding
- **Visual feedback** 😀 😄 🤩 helps
- Stay muted when not talking
- To make **questions or comments**:
  - In the chat
  - use the“raise hand” function + 🎙

⟶ Your **participation and collaboration** is key for making this a great

Teaching Assistants

Matteo Pinna (matteo.pinna@gess.ethz.ch)

Leo Picard (leo.picard@gess.ethz.ch)

Can answer questions about lectures, notebooks, assignments, and projects

How to reach me?

- **Personal question**: face-to-face interaction > emails
- **General interest question**: forum > email

malka.guillot@gess.ethz.ch

IFW E 44 (Haldeneggsteig 4)
8092 Zürich

General motivation

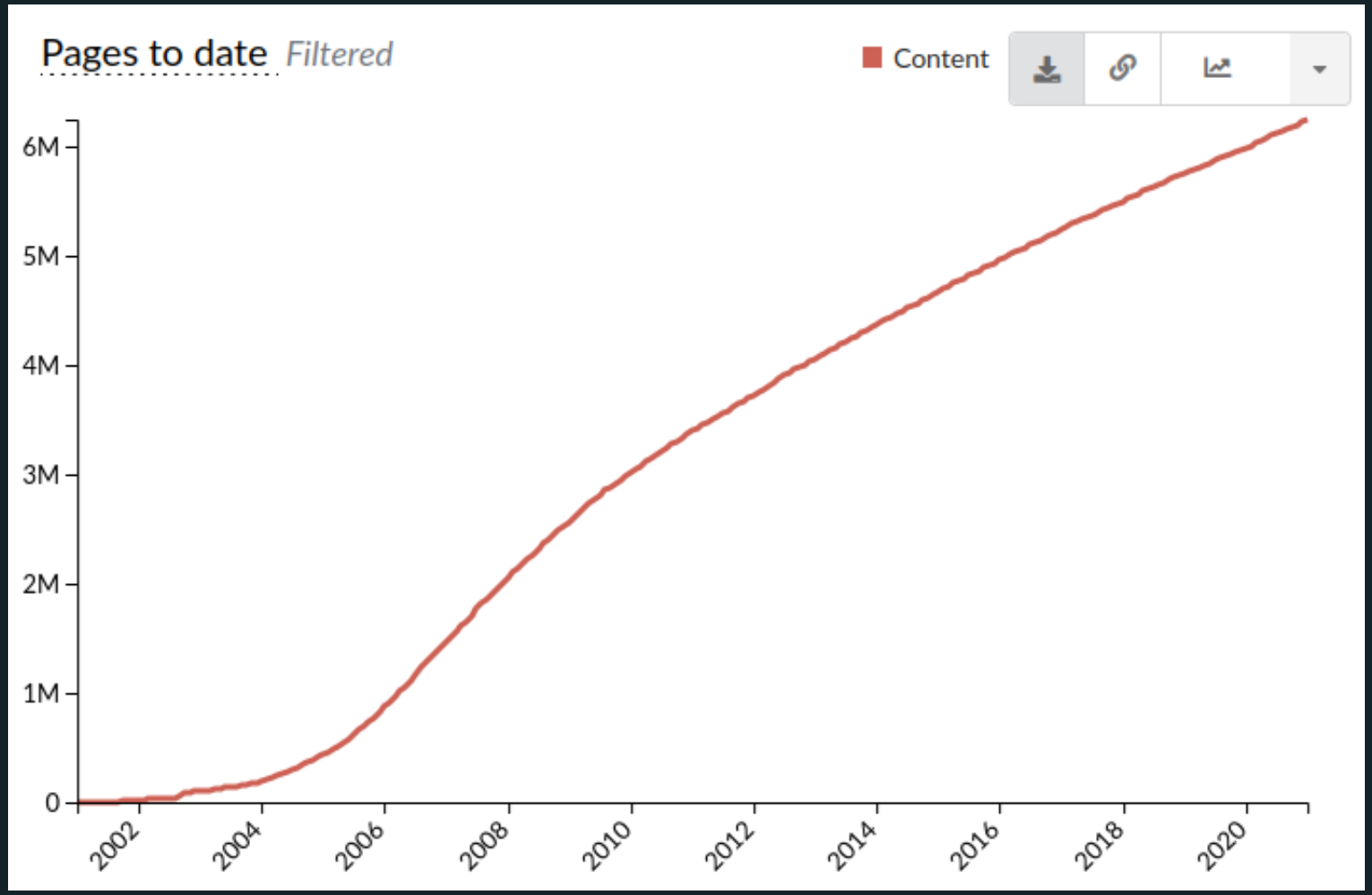
Revolution in policy analysis

- **new datasets** : administrative microdata, digitization of text archives, social media
- **new methods** : causal inference, natural language processing, machine learning

... which contribute to tackle forecasting and public policy evaluation with a new angle

New possibilities: exciting!

of Wikipedia Pages, 2001-2020



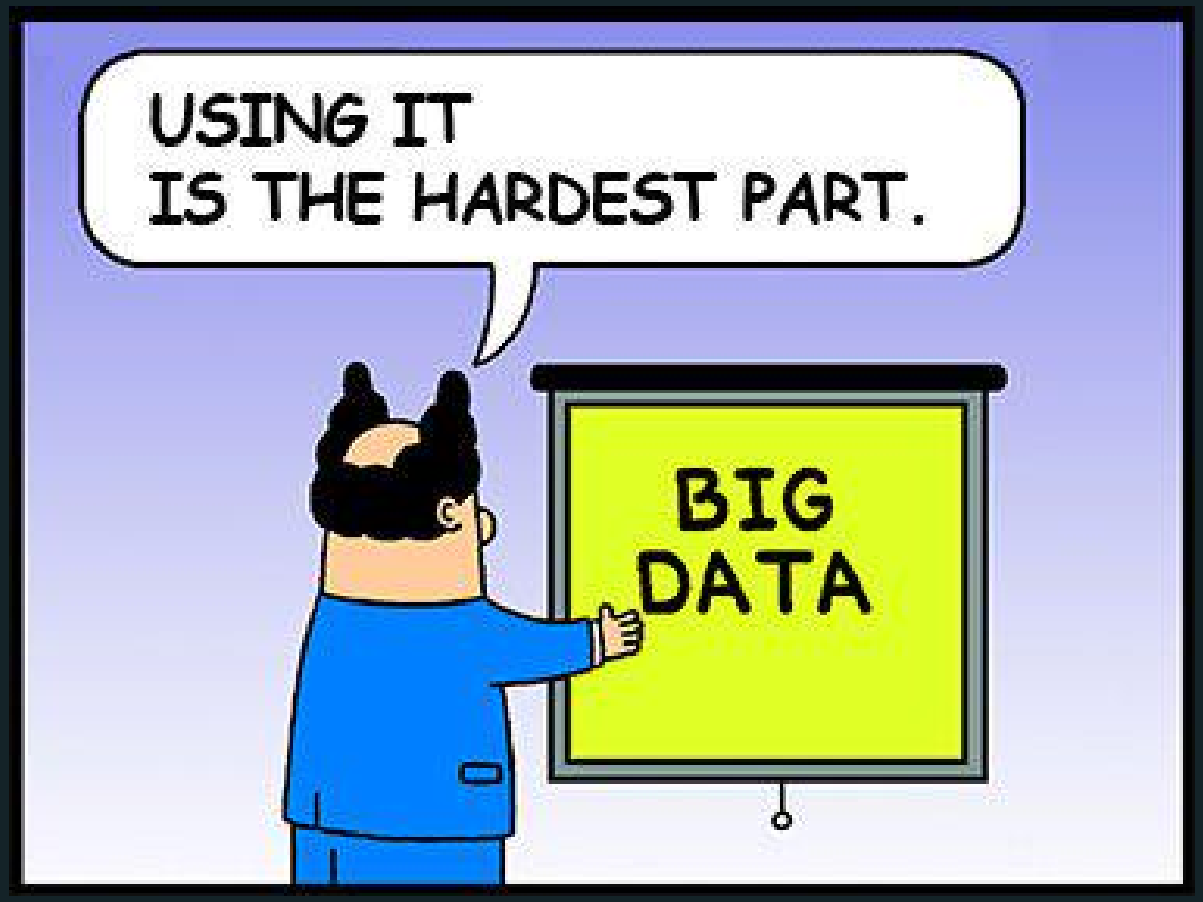
Source: Wikimedia Statistics. The running count of all pages created, excluding pages being redirects.

What is big data?

Expert Survey (UC Berkeley, 2014)

Image by Jennifer Dutcher, source:
https://datascience.berkeley.edu/what-is-big-data

Conclusion



Source: Ingeniero Dilbert

What is big data?

- **Variety** of types/formats of data
  - Structured
  - Unstructured
- **Volume** of data
- **Velocity**: Speed of data flow/stream
- Unusual sources
  - Ready made vs. costummades

→ Use programming and statistics to extract value

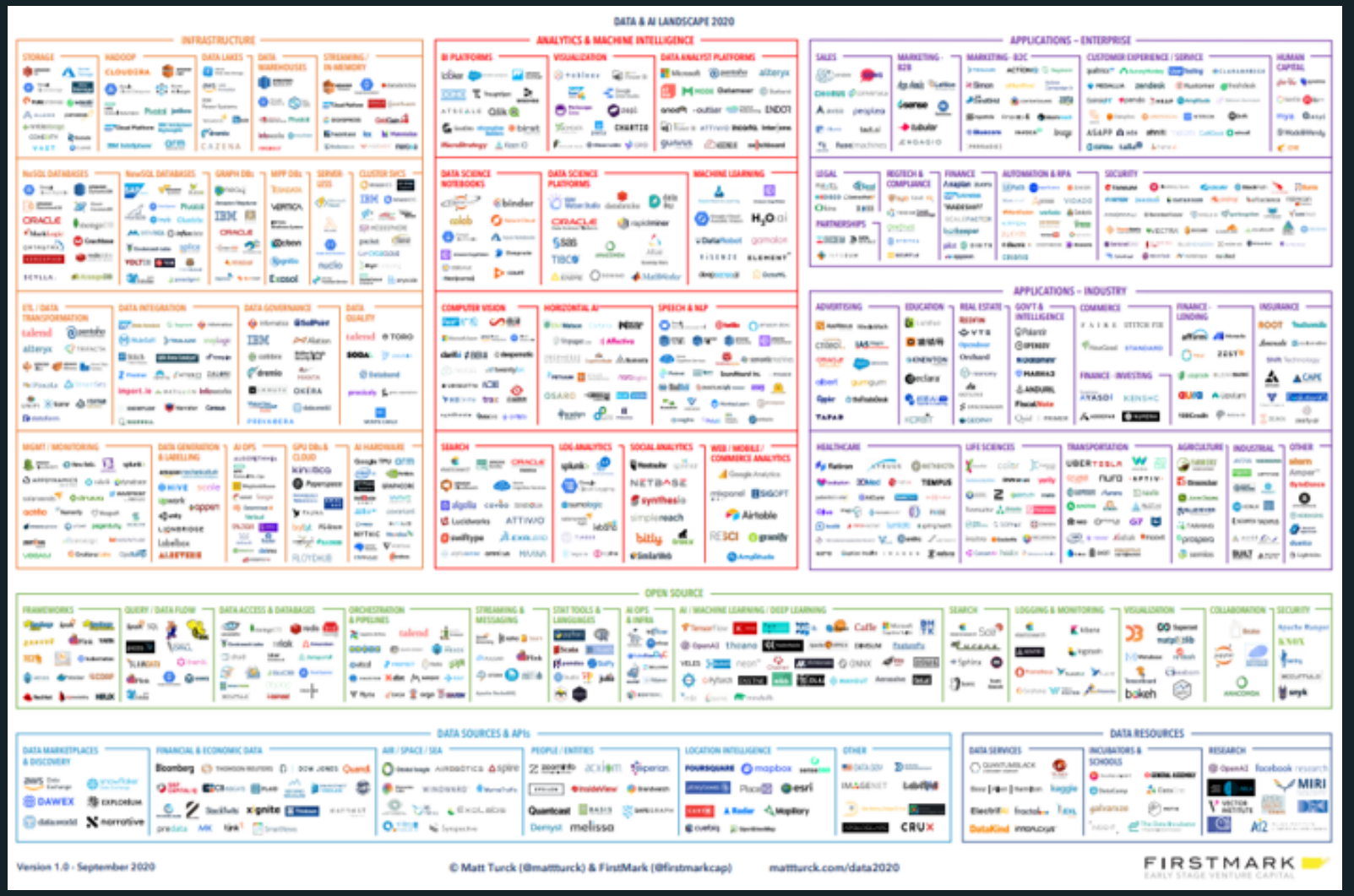
Big data in the Social sciences

- From web applications and digitization of economic and political processes
- **Volume** : can be big, but usually smaller than in natural sciences
- **Variety** and **variability**: often important and challengin
  - Various resources
  - Data generation from 'the real world'
- But usually no streaming applications (**velocity** not that much of an issue)

New tools and methods

- **Data collection**: API, Webscraping
- **Analysis**: text analysis, machine learning
  - Data can be tall (many observations) or **wide/fat** (many regressors) $\Rightarrow$ Machine learning helps to extract the relevant information
- **Visualization**: maps, social networks, web appeals

Big data ecosystem



Source: 'Big Data Landscape (2020)' from http://mattturck.com, high definition image

What is machine learning?

More on this in the statistical learning theory lecture.

Why is it useful to policy analysis?

Empirical policy research (1)

- Standard causal inference framework
- Relying on a **counterfactual** : what happens with and without a policy
- The *art of the counterfactual* intertwine with applied econometrics

  → many **policy applications where causal inference is not central**, or even necessary

(Toy) example

Policy maker facing a drought must decide whether to:

1. Invest in a rain dance to increase the chance of rain

   Causality: do rain dances cause rain?

2. Take an umbrella to work to avoid getting wet on the way home?

   Prediction: is the chance of rain high enough to merit an umbrella?

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z.,2015. Prediction policy problems. American Economic Review, 105(5), pp.491-95.

Conclusion:

Why relying on BD and ML appeals to policy analysis?

1. Not all policy problems are causal inference problems, some require **prediction**

   → ML and BD **supplement** standard econometrics

2. Some data pose **new empirical challenges**

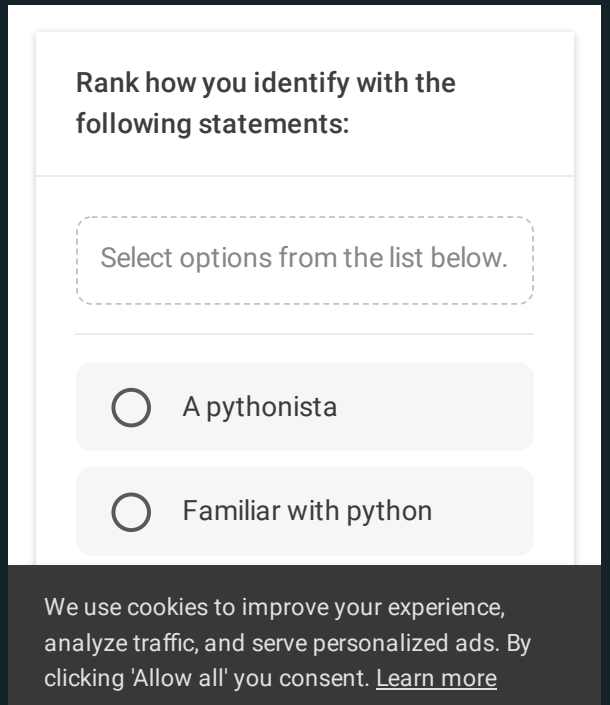
   → ML and BD **complement** standard econometrics

Learning objectives:

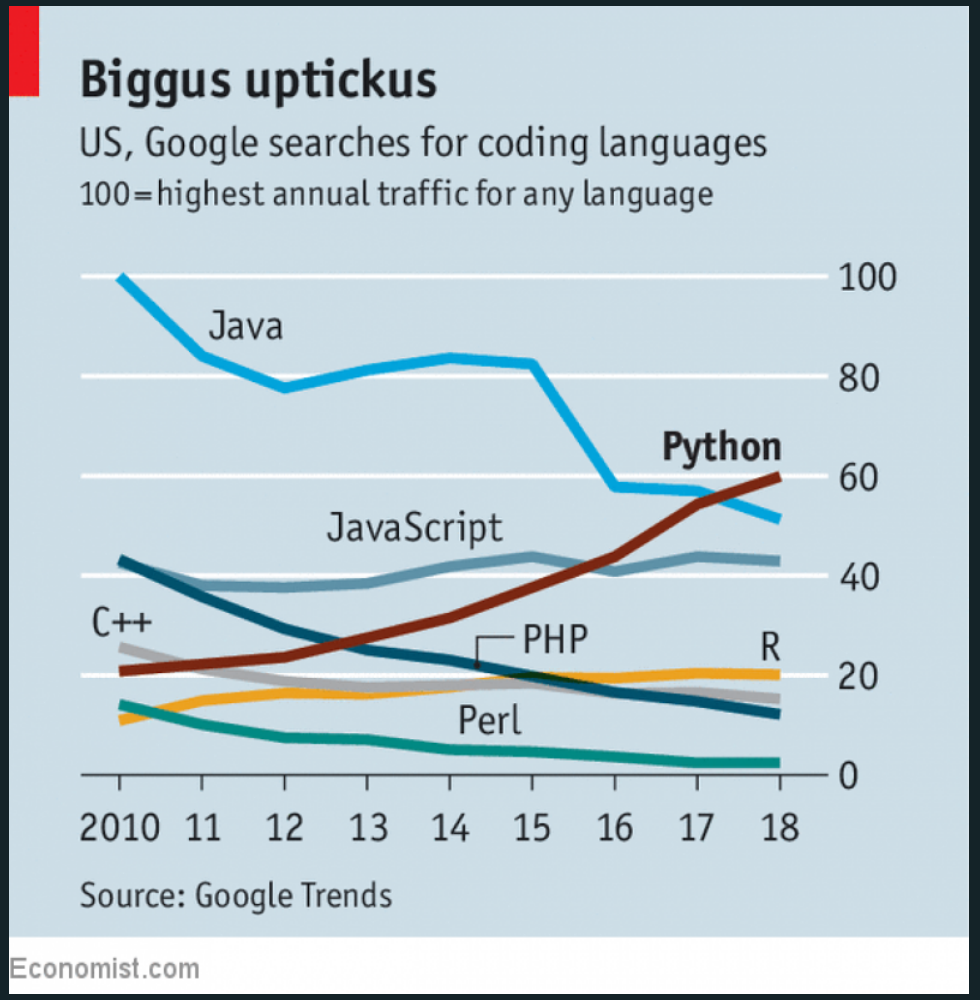
1. Technical skills
   - Introduction data analysis and visualization in python: pandas, web-scraping, API, web-app
   - Programming skills necessary to train and assess the performance of the most popular machine learning algorithms
2. Substantive knowledge
   - Statistical theory underlying common supervised and unsupervised machine learning algorithms.
   - When and how to apply different types of machine learning algorithms to policy issues

Tools and resources

Your programming background



Why Python?


Biggus uptickus
US, Google searches for coding languages
100=highest annual traffic for any language
Java
Python
JavaScript
C++
PHP
R
Perl
100
80
60
40
20
0
2010
11
12
13
14
15
16
17
18
Source: Google Trends
Economist.com

Why Python?

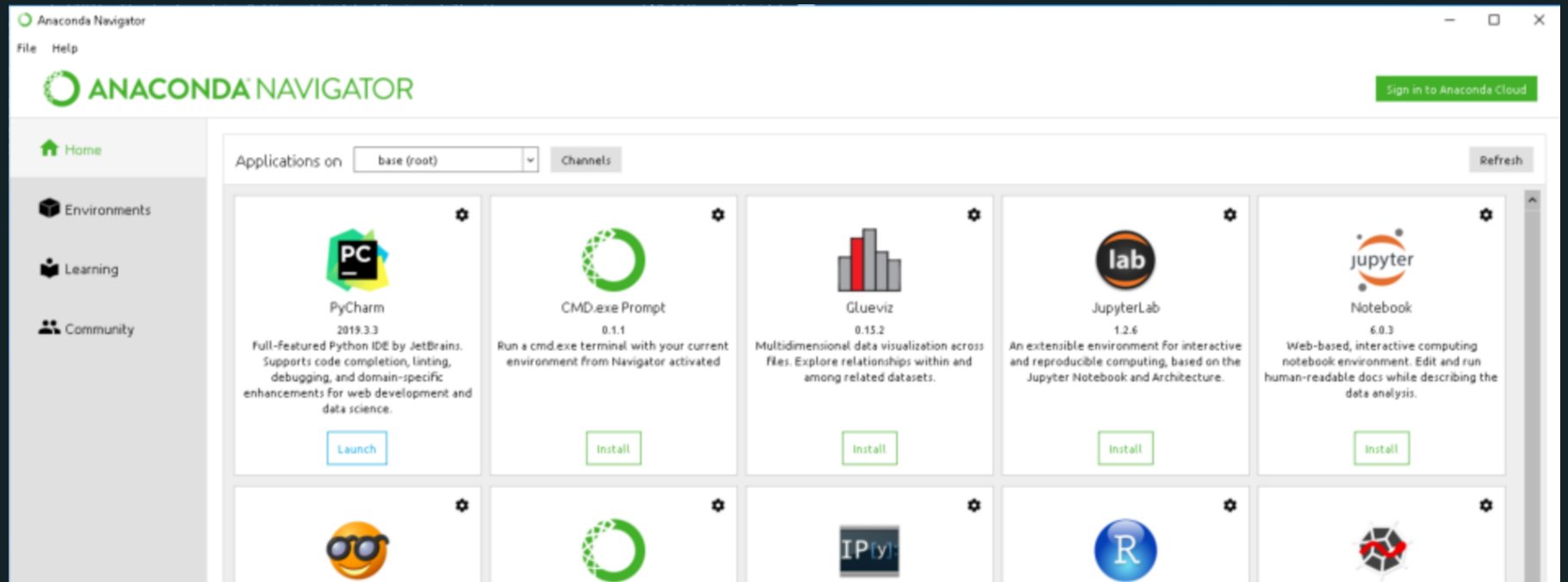
- General-purpose language
  - One of the core languages of scientific computing
- Elegant syntax
- Many useful libraries:
  - Data manipulation: Pandas
  - Machine learning: scikit-learn
  - Statistics: statsmodels
  - Natural Language Procession nltk
- Also path dependency: the language I know the best

Using Python

Anaconda	Jupyter notebook	Spyder
		
a convenient all-in-one install	for homerwork	for longer code

You are welcome to use R instead.

→ Anaconda


Anaconda Navigator
File Help
ANACONDA NAVIGATOR
Sign in to Anaconda Cloud
Home
Environments
Learning
Community
Applications on
base (root)
Channels
Refresh
PyCharm
2019.3.3
Full-featured Python IDE by JetBrains. Supports code completion, linting, debugging, and domain-specific enhancements for web development and data science.
Launch
CMD.exe Prompt
0.1.1
Run a cmd.exe terminal with your current environment from Navigator activated
Install
Glueviz
0.15.2
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Install
JupyterLab
1.2.6
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Install
Notebook
6.0.3
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Install

Course materials are on Github

- Git
  - Git is a distributed version control system.
  - Dropbox + track changes, optimized for codes
- GitHub (≠ Git)
  - = Online hosting platform that provides an array of services built on top of the Git system.
  - Makes life easier

Github is also great for scientific research and for

Why Git and Github?

How to interact with the materials?

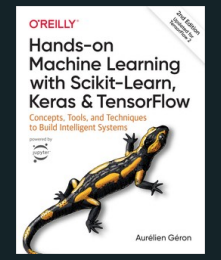
1. **Simple** -> Just use the online GitHub interface to
    - Access the materials
    - Amend the students' presentation signing sheet
2. **Advanced**
    - Download git
    - Create an account on GitHub
    - Go through this simple guide
    - In case it goes wrong: http://ohshitgit.com/

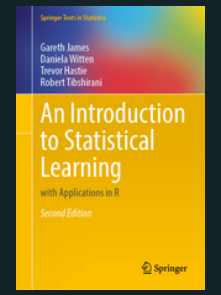
You can use mybinder to launch the notebooks from Github

Main textbook references

Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

James, Witten, Hastie, and Tisbshirani (JWHT), Introduction to statistical learning with applications in R





Other references

Gaillac and L'Hour, Machine Learning for Econometrics.

Course outline

0. Theoretical context

- **W1 & W2**: Statistical learning theory

1. Tools

- **W1**:
  - Overview + tools
  - HW on the basics of python and jupyter notebook
- **W2**: Webscraping and API
- **W13**: Web-app application (dash)

2. Machine Learning

- **W3+5**: Unsupervised ML
- **W4+6**: Supervised ML
- **W7**: Advanced ML
- **W8+10**: Text as data
- **W9**: Advanced ML: Working with time series

3. Causal inference designs

- **W11**: Causal analysis framework
- **W12**: Synthetic control methods

For next week

Python

- Install Anaconda, try out to run python in a Jupyter notebook and spyder
- Basics of python's syntax: Learn Python
  - less Classes and Objects + Modules and Packages.

Troubleshooting

- Use the course forum to share & find answers
- Let's try to make this a **fun collaborative experience** for everyone

Organizing the readings

- Take a slot for a **paper presentation** by:
  - By group of 2
  - Indicate 1st, 2nd and 3rd choice for a presentation
  - You can contact me (*and are encourage to*) if you want to present a paper that is not on the list