Data Management

Introduction

Malka Guillot

HEC Liège | ECON2306

Table of contents

Welcome

Introduction: Who are we?

Teaching assistant	Lecturer
Michel Coppee	Malka Guillot
michel.coppee@uliege.be	mguillot@uliege.be

Bât. N1 Economie (bureau 33a)
rue Louvrex 14
4000 Liège
Belgique

Who am I?

PhD in economics from the Paris School of Economics

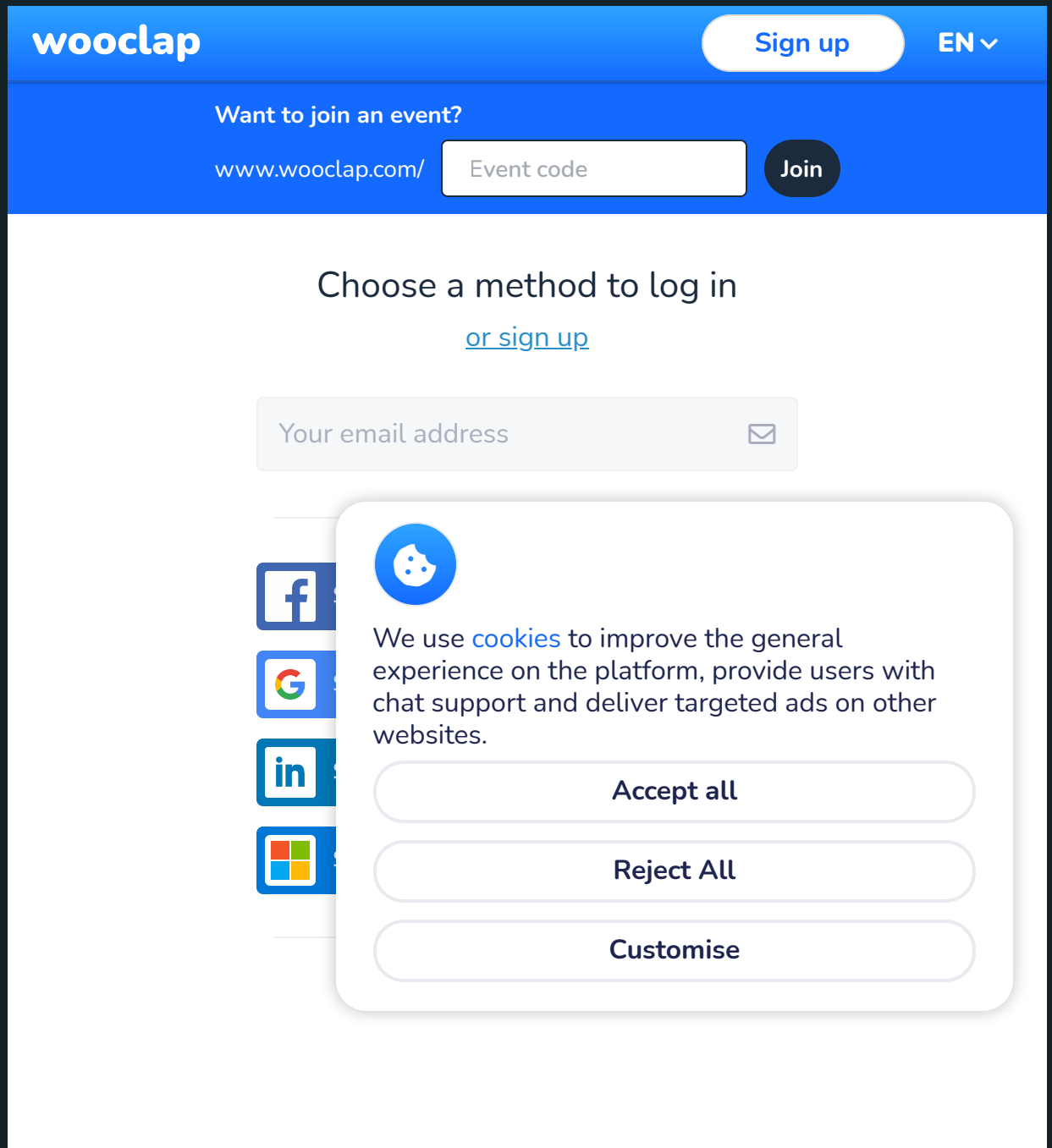
Postdoc at ETH

Assistant professor in applied micro economics at HEC Liège

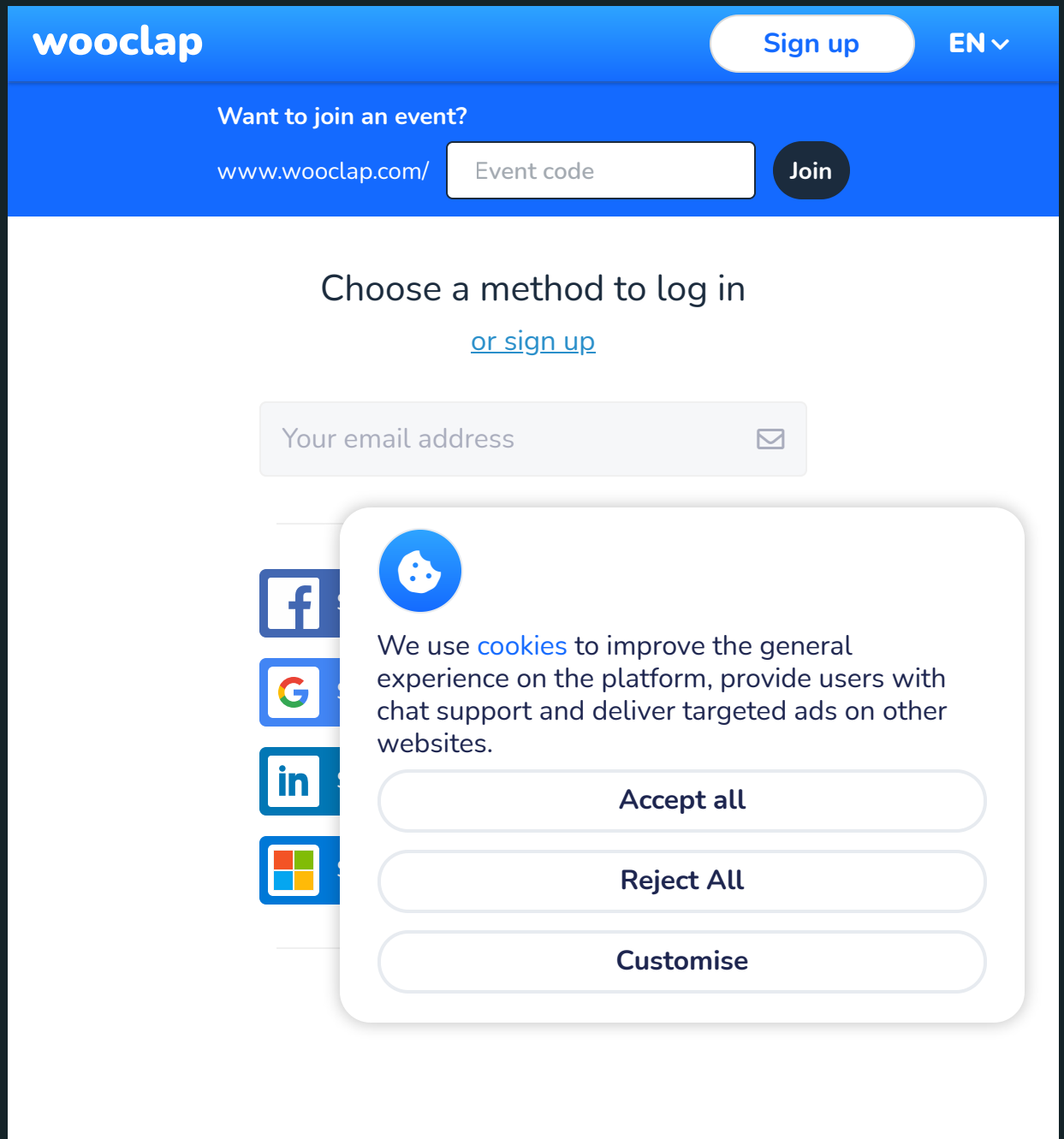
Interested in **public economics** questions: inequality and taxation

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

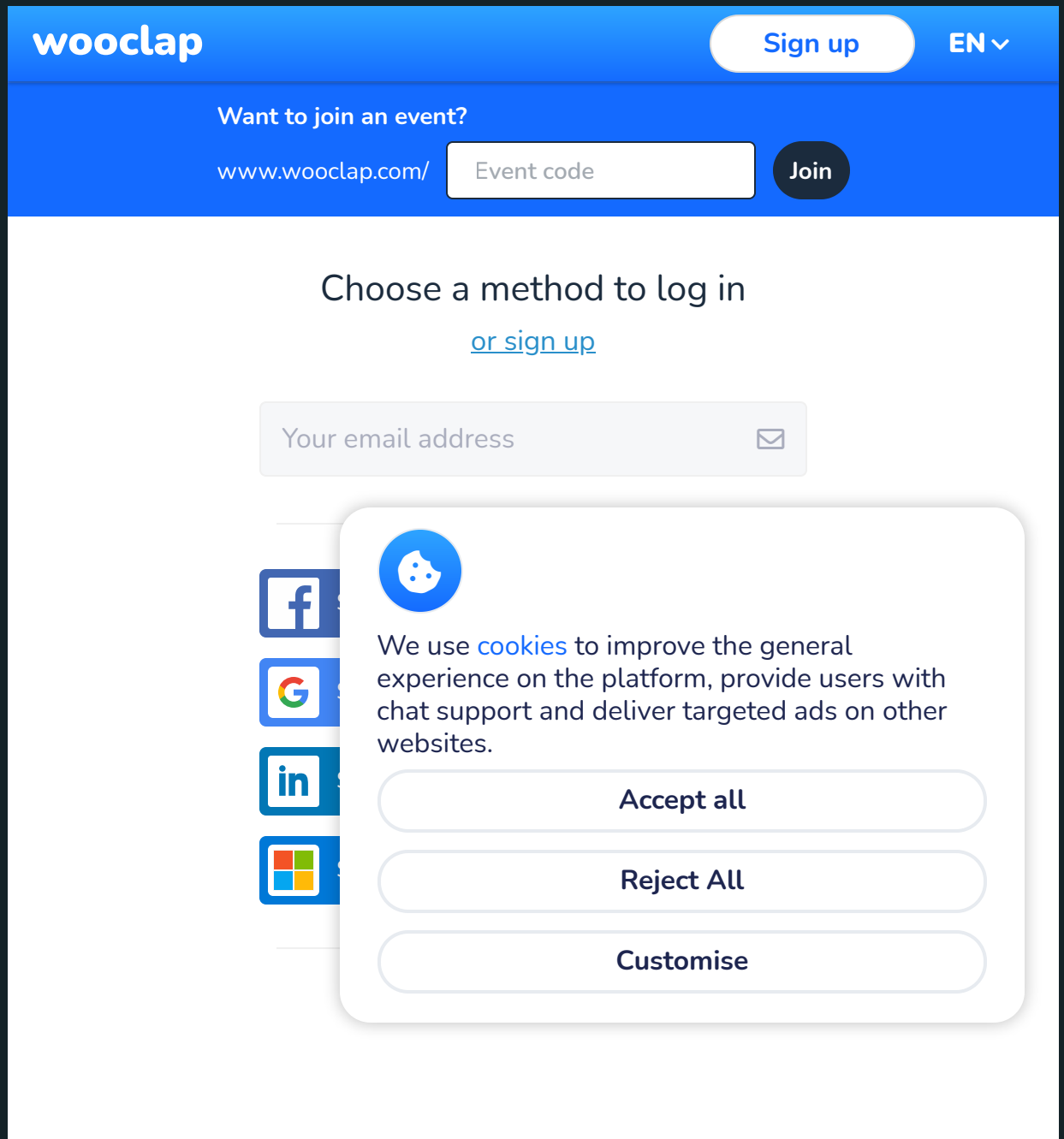
Wooclap


wooclap
Sign up
EN
Want to join an event?
www.wooclap.com/
Event code
Join
Choose a method to log in
or sign up
Your email address
We use cookies to improve the general experience on the platform, provide users with chat support and deliver targeted ads on other websites.
Accept all
Reject All
Customise

Introduction: Who are you ?



What do you expect to learn during the class?



Data management: why, how?

What is data management about?

All processes, tools, and techniques that have to do with **working with data** :

- Data management plan
- Research data archiving
- Metadata :
  - = structured information that describes, explains, locates, and otherwise represents something else [data].

→ Allows data to be found and interpreted

- *Bottom line*: data should be valid, shared and contextualized within (research) communities

The Data Management Plan (DMP)

Supports Transparency and openness, by indicating:

- how data will be made discoverable, accessible, and reusable

Important in the context of open science / governemnts:

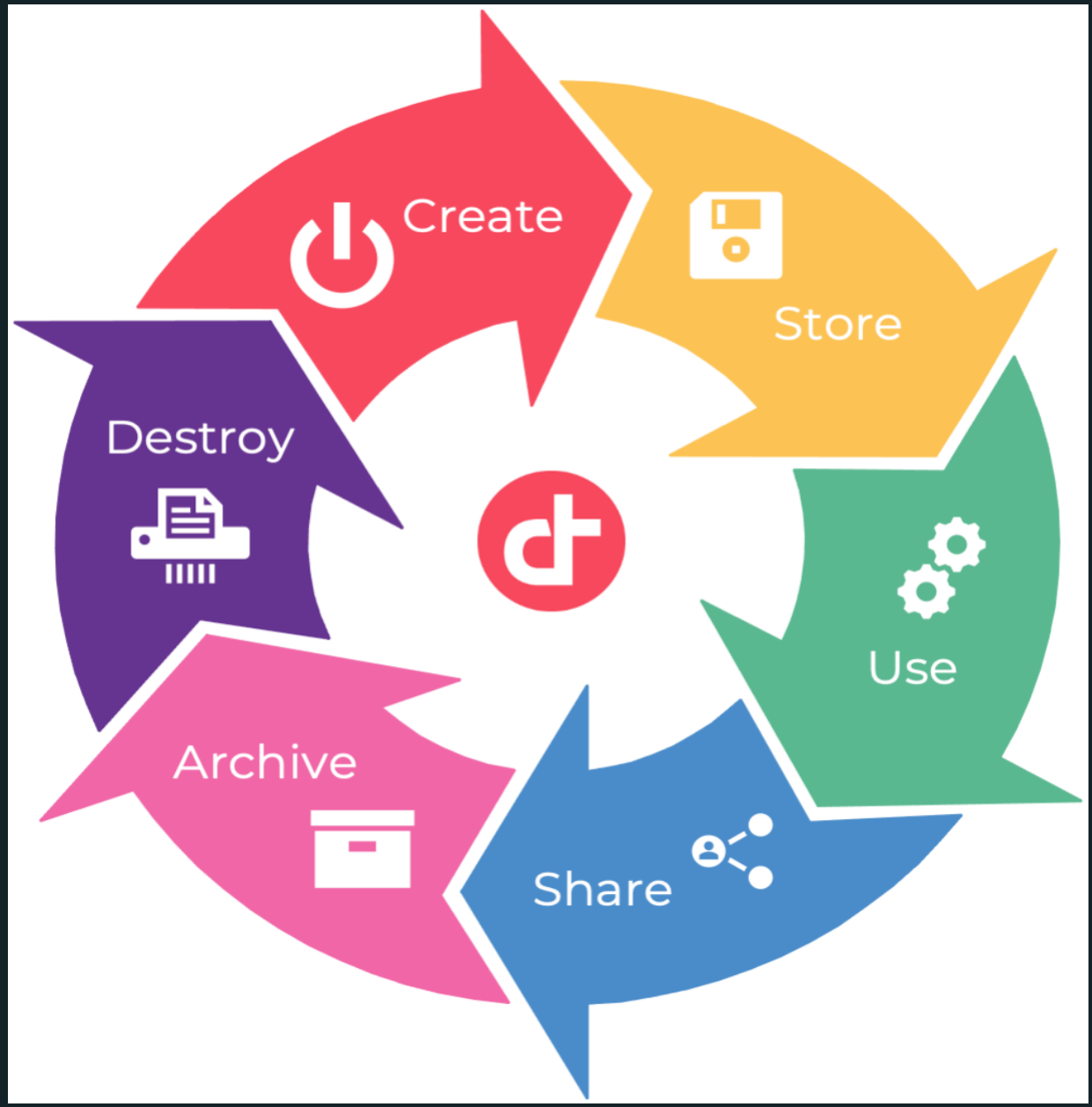
- So that **public investments** are transferable

But also in the context of a firm:

- Long-term investments are key for sustainability

Document that helps you manage the data lifecylce

The data lifecycle



This class: from acquisition of data to data analysis

The class focuses concepts & skills related to the management of data, that are central for the **exploitation** of data.

Goals:

- Equip you with the standard datascience toolkit.
- Put it to work on a real-world project.

Backbone of the class

1. The **skills**:
    - Data collection
    - Data cleaning & operation:
        - Pipelines
    - Data vizualisation
2. The **tools**:
    - python
    - git
3. The **concepts**:
    - *Project management*: documenting, sharing & managing code
    - *Reproducibility*

Public targeted: **anyone using data for projects**. For academics or non academics.

- For research
- For firms

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 research or a field (economics, political science).
  - In python.
- *It is not*:
  - **Computer science**. We're not coding up models from scratch.
  - **Mathematical statistics**. We're not deriving the functions by hand.

(Big) data

Revolution in the use of data

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

⇒ New avenues in:

- research
- policy analysis
- business (costumer services)

New possibilities: exciting!

Examples of business applications

- Decision making:
  - What judges can be replaced by robots?
  - Using algorithms to help diagnose cancer / propose the most effective treatment
- Growth hacking:
  - Identify markets where the investments have the highest returns
- Forecasting:
  - Predict sales

What is (big) data?

Expert Survey (UC Berkeley, 2014)

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

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

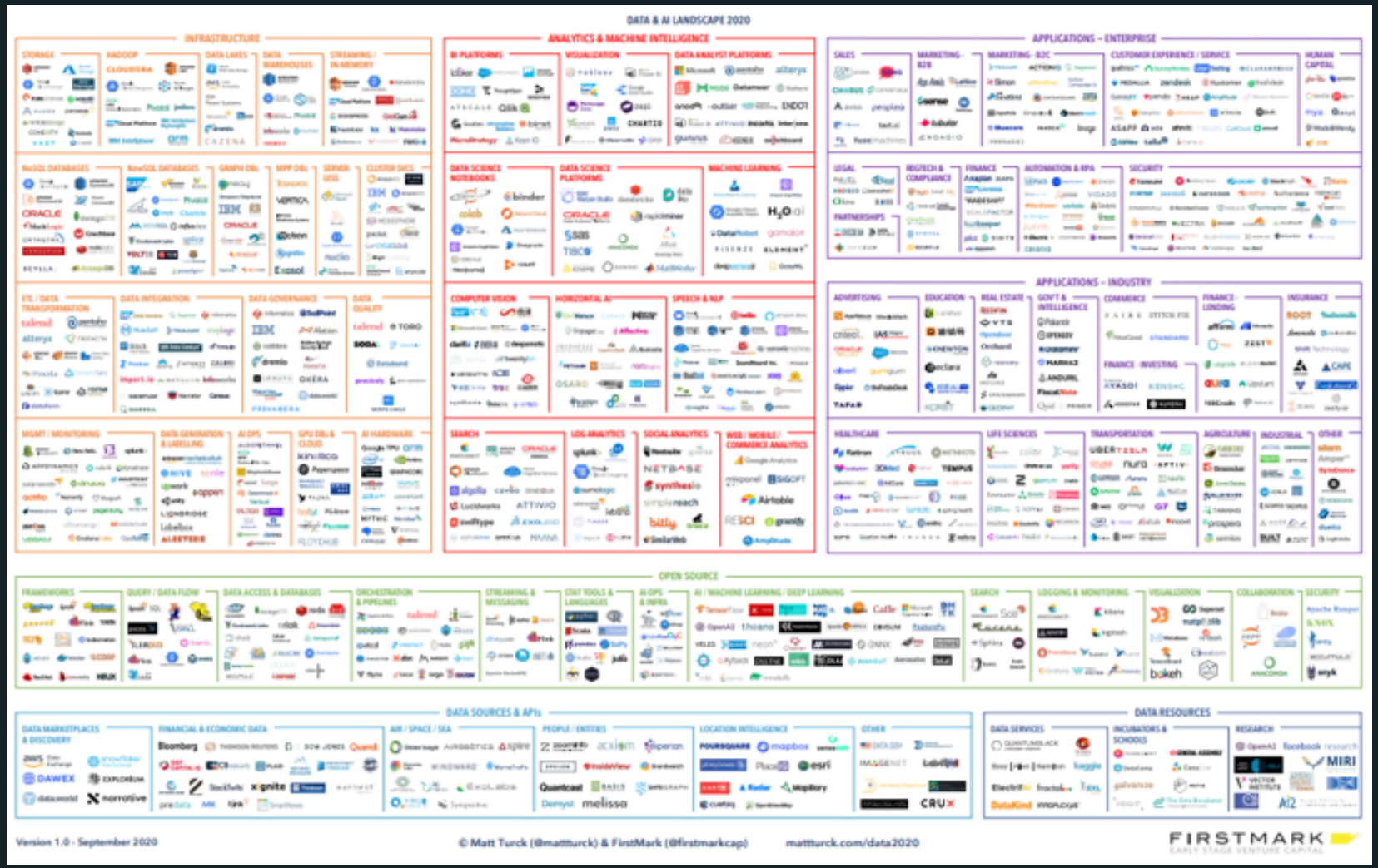
Big data in the Social sciences

- 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 challenging
  - 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 applications

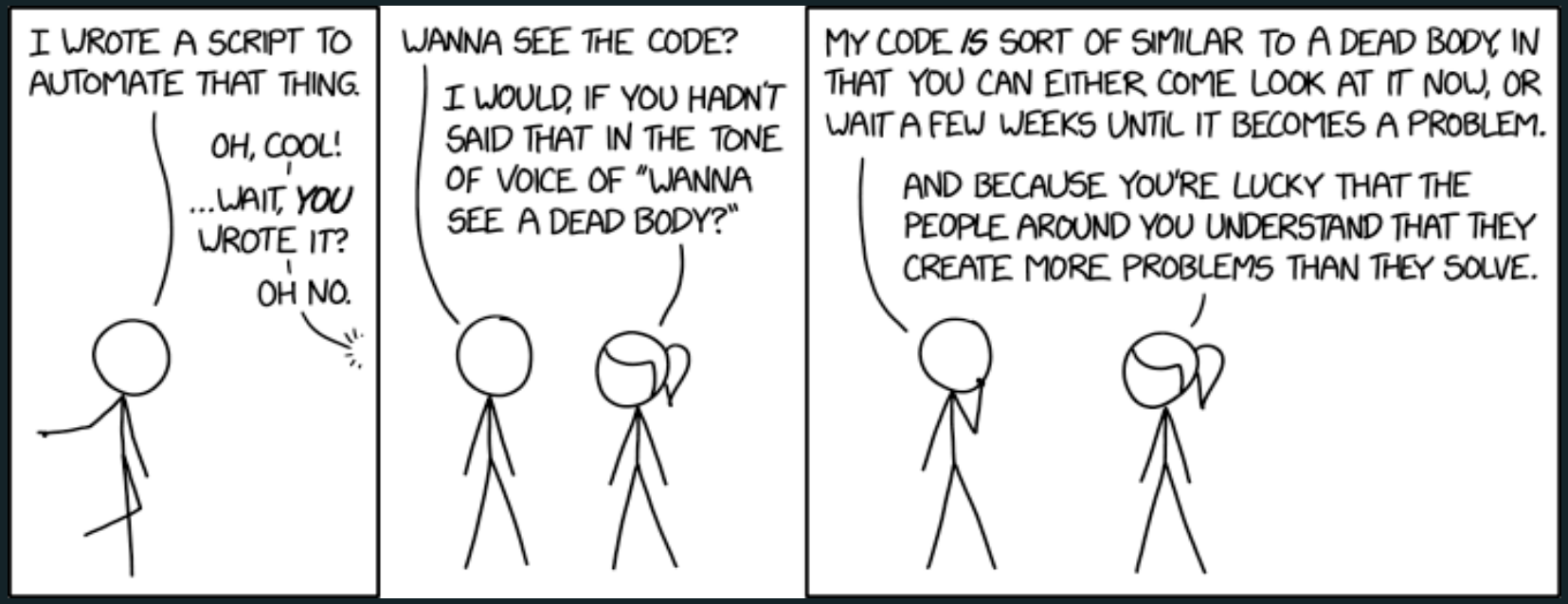
Big data ecosystem



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

Managing a project with data

The importance of good coding practices



Source: xkcd 2138

Readability of the code

The **Pep 8** convention: Style guide for python code

→ makes it easier (possible) to understand a code of someone else (= you + 2 day!)

- **Naming**
  - Variables: underscores & small letters snake_case
  - Constants: underscores & capital letters
  - Classes CapitalizedCase
- **Code layout**
  - Blank lines
  - Maximum line length & line breaking
- **Comments**
  - Should be useful (explain code) but not obvious
  - Not on o code line
  - Documentation Strings (Using docsstrings) -> mainly for functions

Reproducibility principle

The results of the project should be *reproducible* by someone else in the future:

- this is a basic scientific principle... but too often forgotten

WANTED:

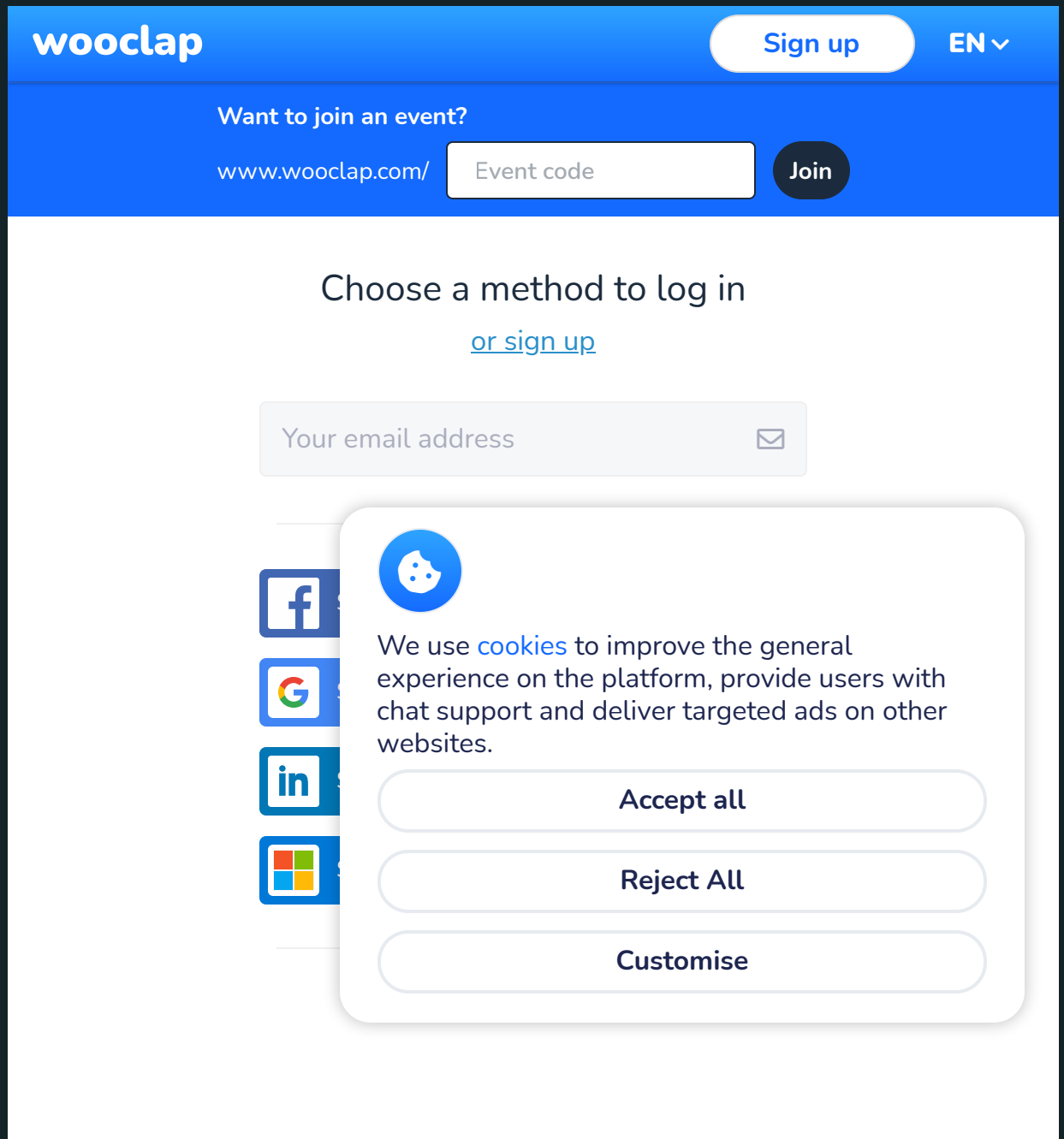
- maintaining a single master file of the data
- version control of the code
- Readme of the project
- document the code (« comments ») & the data (« metadata »)
- controlled coding environment

Next lecture

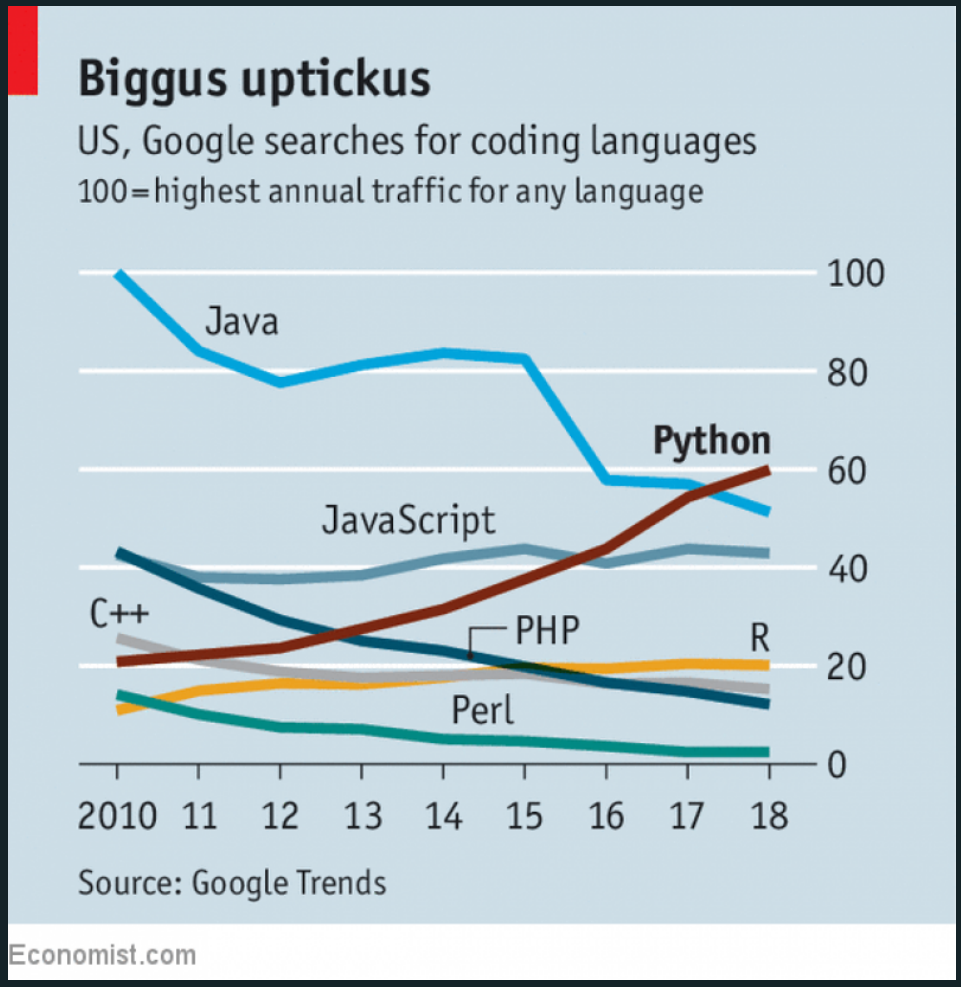
→ **The course project satisfy by the reproducibility principle**

Tools

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?

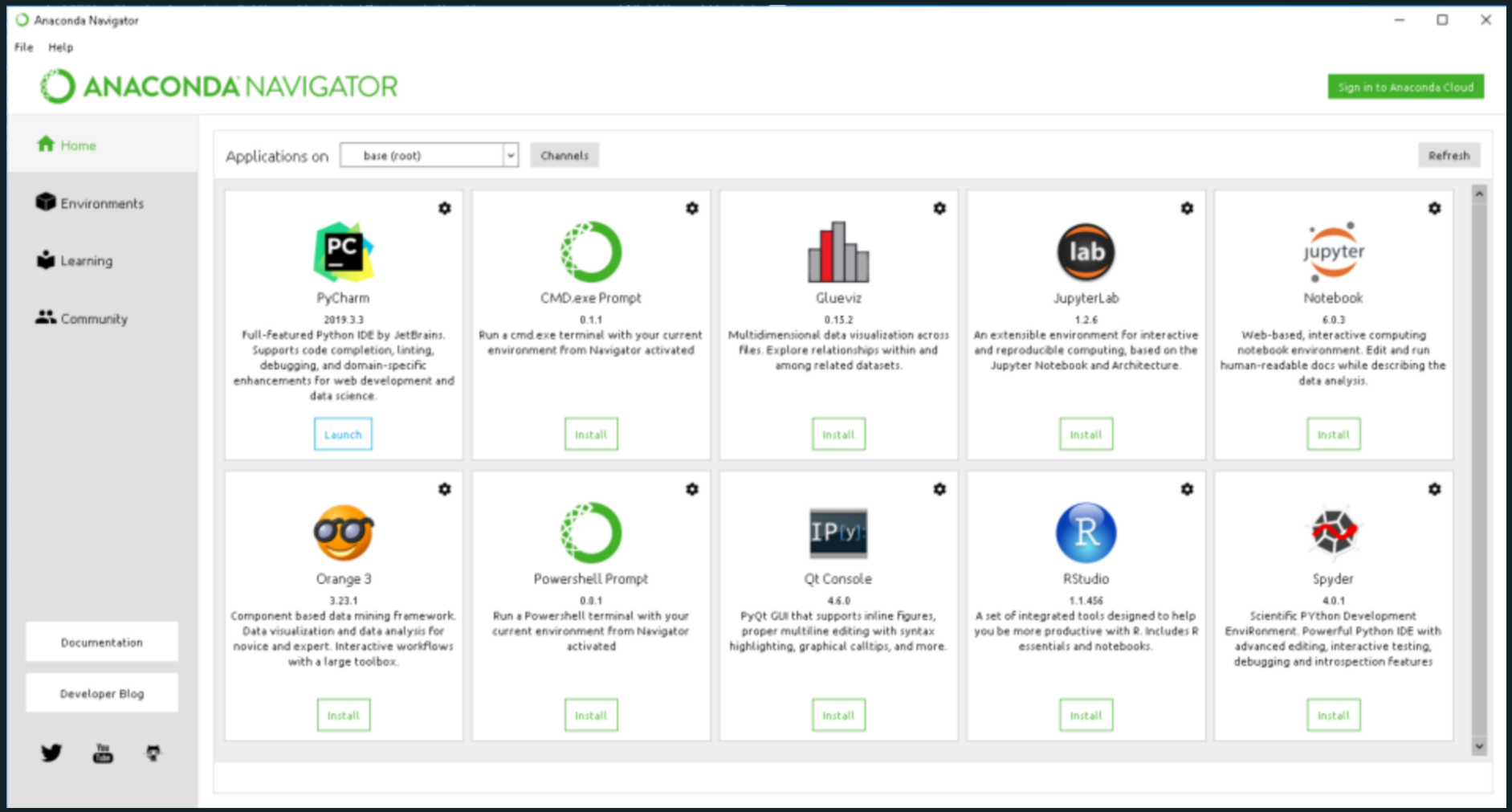
- General-purpose language
  - One of the core languages of scientific computing
- Elegant syntax
- Many useful libraries:
  - Data manipulation: Pandas
  - Machine learning: scikit-learn
  - Statistics: statsmodels
  - Natural Language Procession nltk
- Also path dependency: the language I know the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



Spyder & Jupyter notebook are two development environments from the Anaconda set up.

Main python packages

Task	Package
Webscraping	beautiful soup
Data management	
Visualisation	
Web application	
Machine Learning	
Natural language processing	NLTK &

This class: overview & logistics

How does the class work? Spirit

Sessions are designed to be interactive

- mix of live *coding* & *exercises*
- we want to get you comfortable using your computing environment to solve problems
  - bring your laptop!
  - we expect you have completed the installation guide and have all software installed.
  - ask questions!

How does the class work? Details

- **Lectures**: 3 hours / week
  - 2 hour theory
  - 1 hour practice:
    - coding exercises
  - sometimes the frontier between theory and practice will be fuzzy.
- **Every week**
  - Thursdays
    - Theory: 9:00-10:25 (with a 10 minute break)
    - Practice: 10:35-12:00
  - Where? N1a 220 (2/20) [Liège centre - Louvrex]
  - Dates: 10.02.; 17.02.; 24.02.; 03.03.; 10.03.; 17.03.; 24.03.; 31.04.; 28.04.; 05.05.; 12.05.; 19.05.

Online Course Materials

- Syllabus
- lola:
  - Course announcement and forum
  - Giving back homerwork
- 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

[Evaluation Policy]

- **Homeworks**:
  - should be given back as jupyter notebooks in PDF format on lola.
  - $3\text{h} \times 5$ **=15%**
- **Participation in class & presentations = 5% bonus**:
- **Course project = 85%**

The homeworks are simple exercises designed to help students to “get their hands in the data & code”.

[Course project] Objectives

- The **basics**:
  - End-to-end data project using Python
    - From collection to vizualisation
  - Group project (2 people; 3 of odd no. of students)
- Use what you learn in this course to **solve a non-trivial real-world question/problem** using a graphical analysis
  - Code must be in split into meaningful sub-files
  - Solution must be submitted using GitHub
  - Web application, that should be deployed online

[Course project] Web application deployed online???

→ Some examples in various sector:

- Finance:
  - The Yield Curve
- Health
  - Opioid epidemic in the US
- Transportation:
  - Uber rides
- Energy consumption
- https://xkcd-data.herokuapp.com/
- Research project

→ Be creative, have fun!

What about you?

1 minute to think about a potential field of application.

- Present yourself
- Specify 1 or 2 domain of interest with possible data analysis
  - Can be academic: green finance, agile management
  - or not: sport, important topic

[Course project] Requirements

- Data:
  - Original data collection
- Analysis :
  - 2 tables and 2 Figures (using different commands)
- Deployment:
  - The main output should be a dash page that you develop on Herokuapp
- Submission format:
  - Invite @malkaguillot and @MichelCop to collaborate on your GitHub repository by the due date.

[Course project] Evaluation: 85% =

- Project management = 15%
  - reproducibility, github, readme
- Project relevance = 10%
  - Does the project respond to an interesting/important question?
- Quality of the visualisation = 20%
  - Choice of the graphical representations & colors
- Technical dimension = 15%
  - Is the project using advanced tools/techniques?
- Oral presentations = 25%
  - ML1: Project idea & scrapping methodology = 5%
  - ML2: Visualisation plan = 5%
  - ML3: Final presentation = 15%

Course Communication

- Us → you
  - Course communication will be done through lola's forum
- You → us
  - We will be available
    - During the breaks, after the class.
    - Michel Copée can answer questions about lectures, notebooks, assignments, and projects
  - **Personal question**:
    - face-to-face interaction > email
  - **General interest question**:
    - forum > email

References?

No general texbook. Specific references will be given when corresponding subjects are tackled.

- Introduction to python, pandas, plotting
- Stackoverflow: all the answers are there, but you have to ask the right question.

Epilogue: for next week

Python

- See installation guide on lola
- Install Anaconda, try out to run python in a Jupyter notebook and spyder
- Wait for next week's introduction by Michel !
  - 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