

Data Management

Introduction

Malka Guillot

HEC Liège | ECON2306

Table of contents

1. Course objectives

- Data lifecycle
- Goals of data analysis

2. The tools

- python
- git

3. Organisation & logistics

Introduction: Who are we?

Teaching assistant

Michel Coppee

Lecturer

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

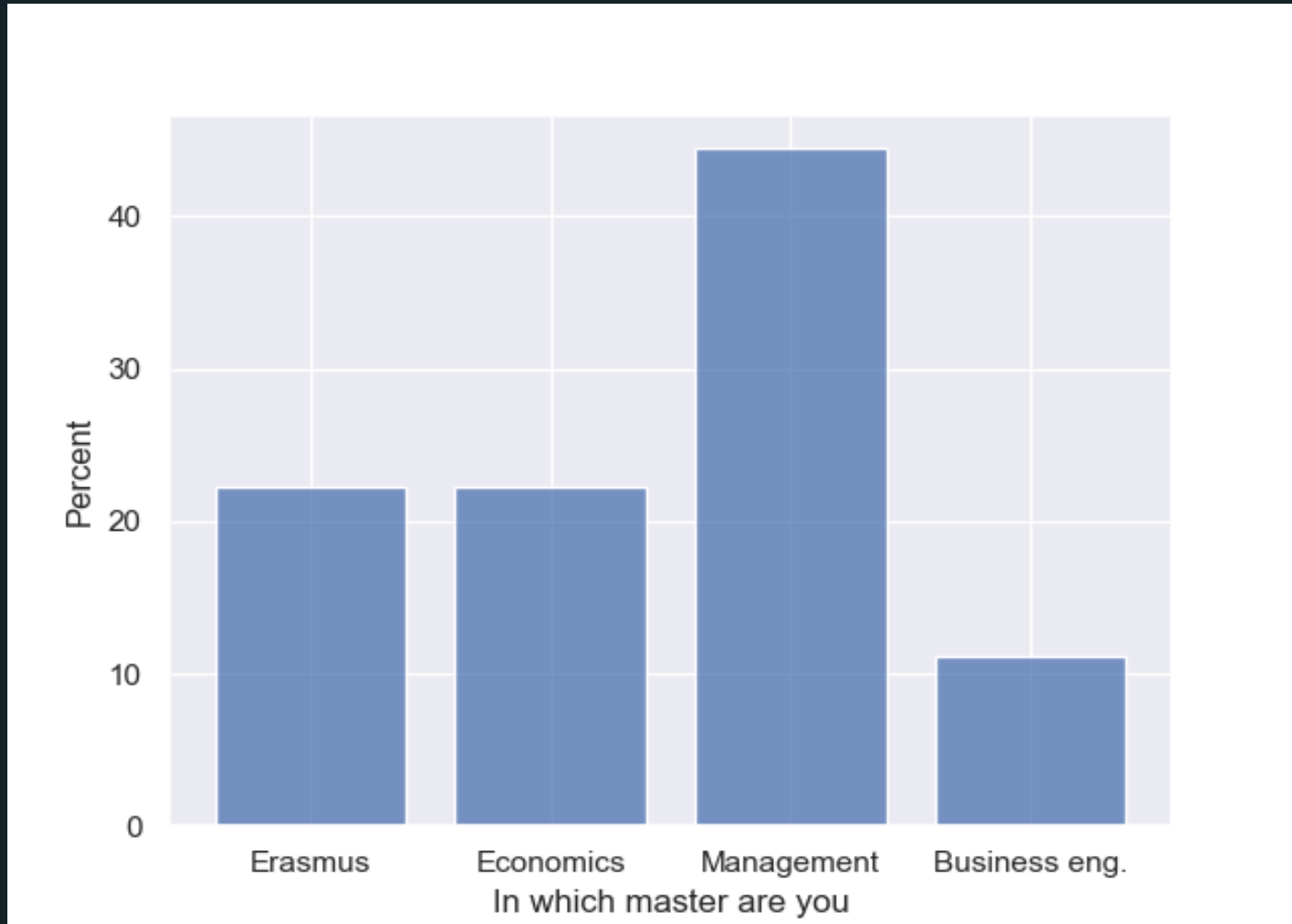
Assistant professor in applied micro economics at HEC Liège

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

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

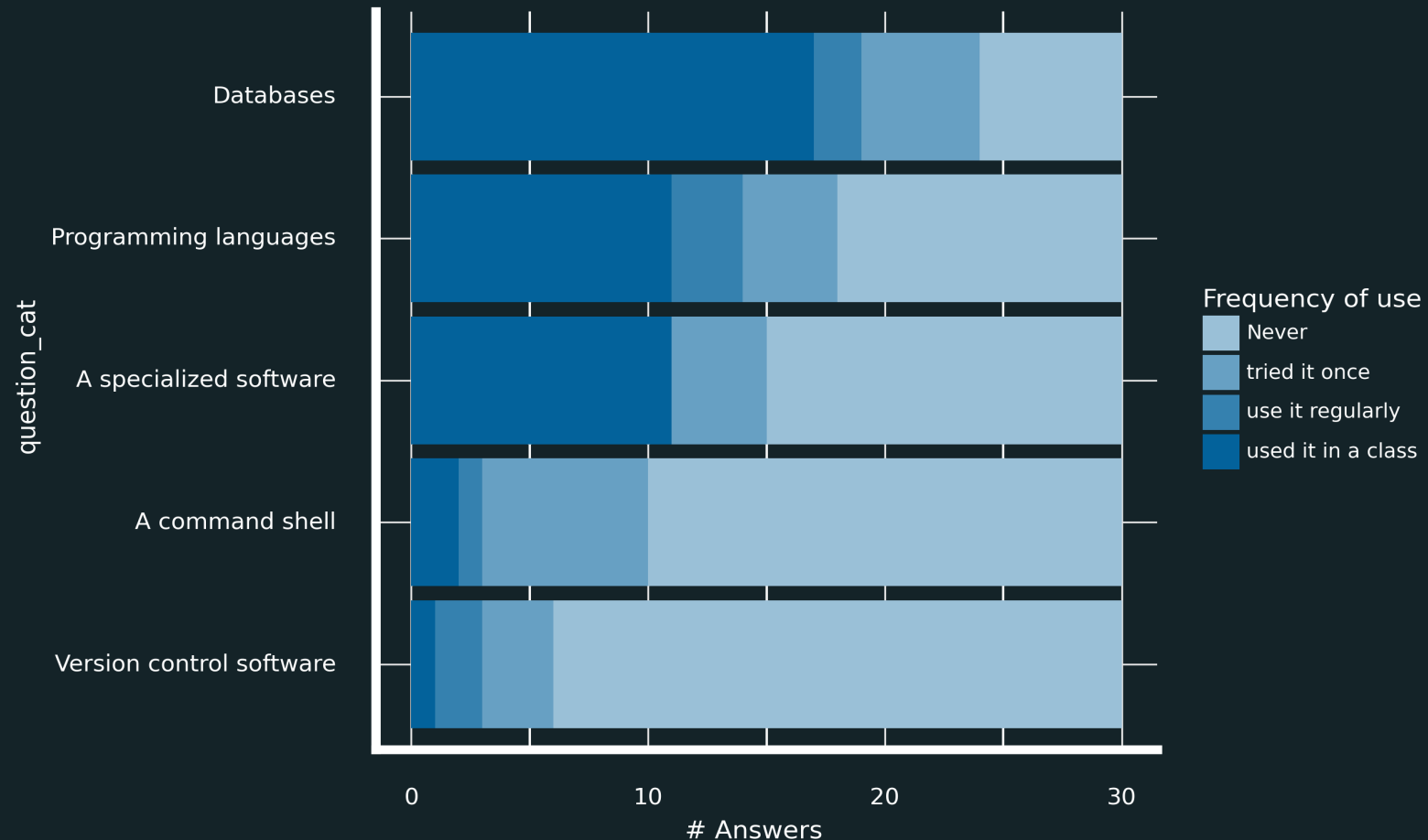


Introduction: Who are you ?



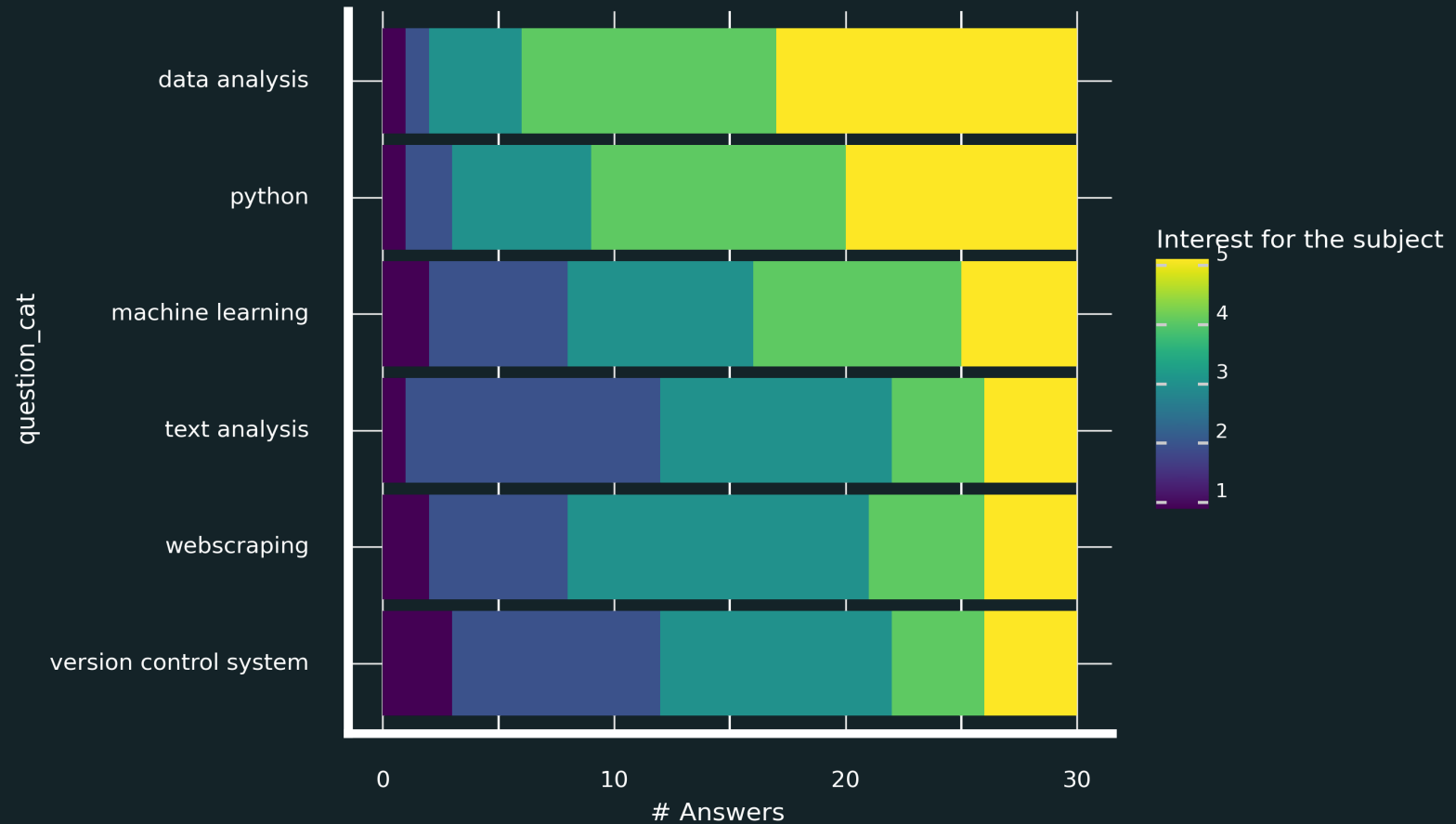
Source: pre-class survey

Your programming experience



Source: pre-class survey

Which part of the class interest you most?



Source: pre-class survey

Why have you chosen this class?



Source: pre-class survey

Course objectives

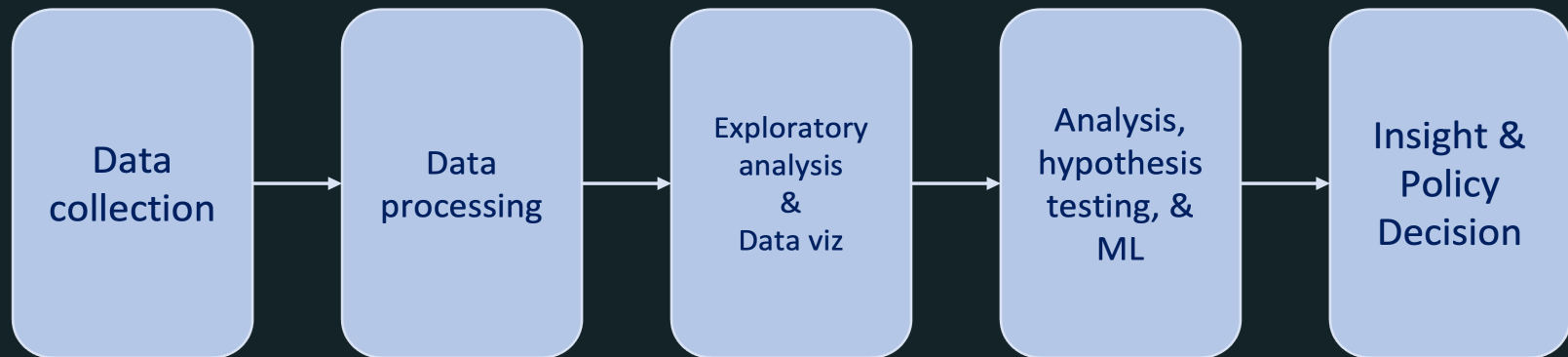
From acquisition of data to data analysis

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

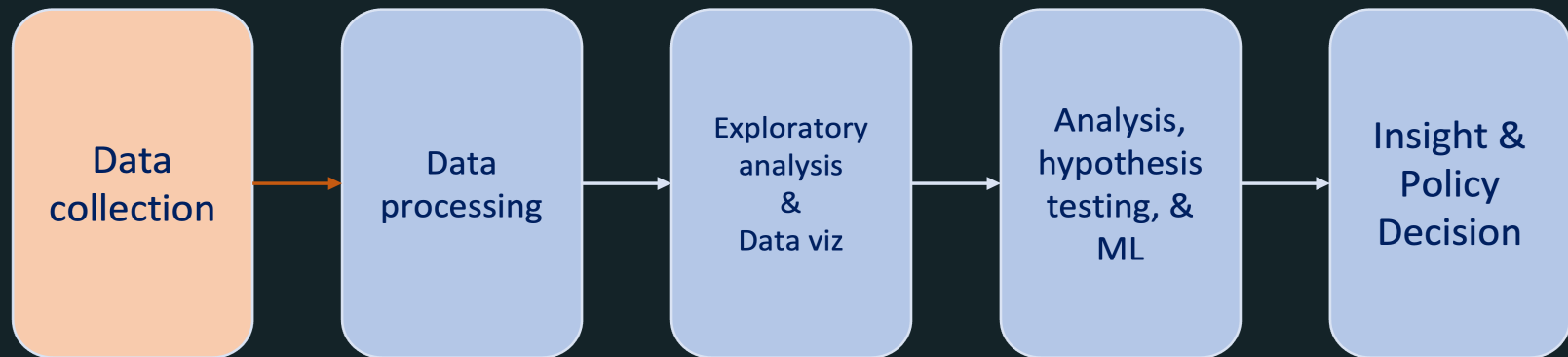
Goals:

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



Data lifecycle



Data lifecycle



Looking for data?

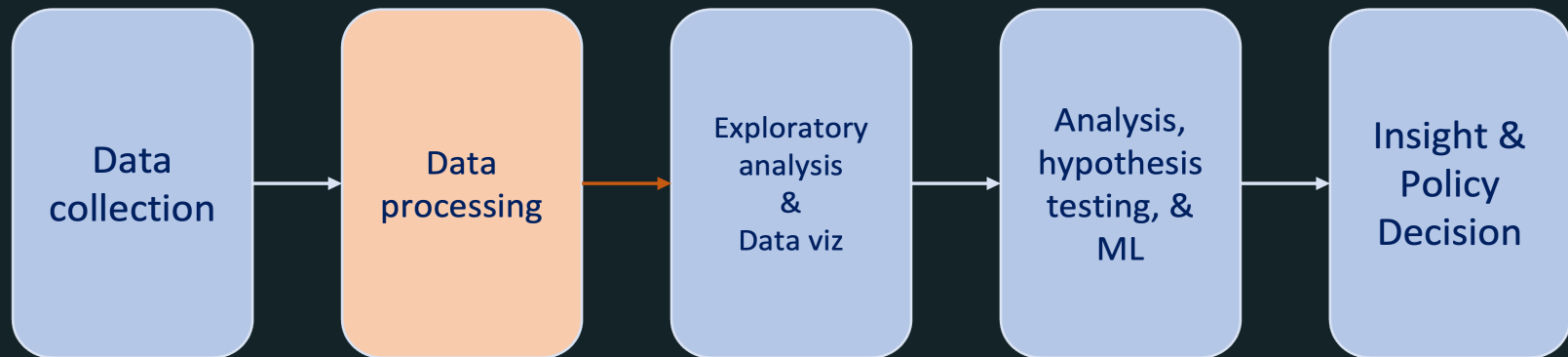
- Search engines:
 - [google datasetsearch](#)
 - <https://fr.statista.com/>
- Institutional repositories
 - [Open data Liège](#)
 - [The official portal for European data](#)
 - [OECD data](#)
 - [Statbel and Open data Belgium](#)
 - [Open data France \(also here\)](#)
- List of resources:
 - [Common-built archive](#)
 - [Personal lists here, here or here](#)
 -   [Registry of Open Data on AWS](#)

You do not find what you are looking for?

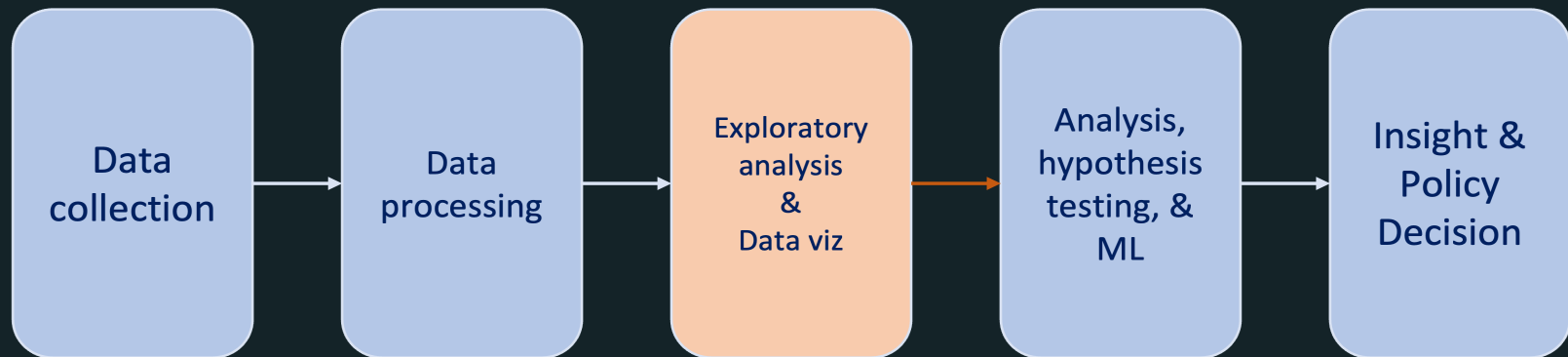
Some solutions:

- Data collection
 - webscraping
 - asking for the data
 - administrative sources
- Implement a survey

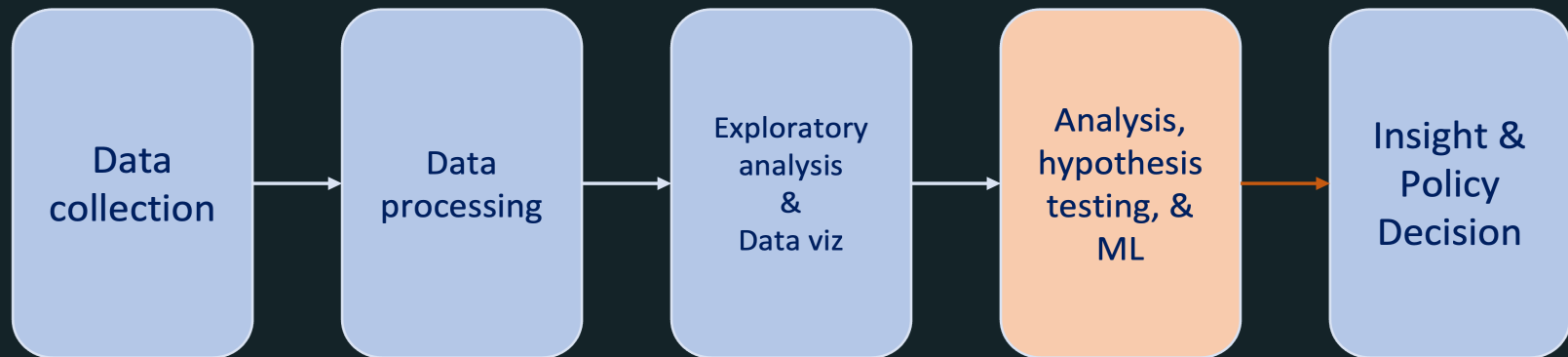
Data lifecycle



Data lifecycle



Data lifecycle



Analysing data

When we analyze data, what is our ultimate goal?

- **Description:** Simply characterize observed patterns in the data.
- **Causation:** Learn about causal relationships: If we change X , how will Y change?
- **Prediction:** Be able to guess the value of one variable from other information.

Analysing data

Examples of a (research) question in each category:

- **Description**: Is wealth inequality increasing faster in the U.S. than in Europe?
- **Causation**: Does Medicaid coverage reduce the risk of bankruptcy?
- **Prediction**: Can nighttime satellite imagery be used as a real-time indicator of GDP?

Discerning which type of goal you have is critical for:

- Choosing methods:
 - Distinct approaches are required to achieve different goals.
- Interpreting results:
 - Mistaking one goal for another can lead your audience to make very bad decisions.

The data generating process

What's shared across all goals of data analysis.

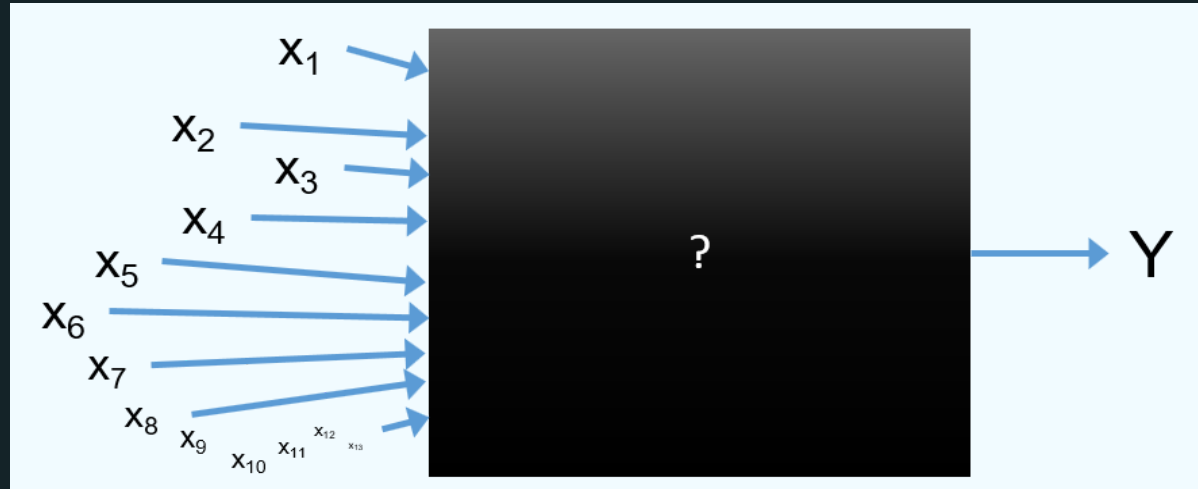
We start with an **outcome variable** called Y . We want to know:

- Where do the values of Y come from?
- How does Y relate to other variables?

The data generating process (DGP) in the real world determines the values of Y

The data generating process

The true DGP is unknowable, a black box:



We interpret the picture with a function:

$$Y = f(x_1, x_2, \dots, x_N)$$

Questions:

- What are all these inputs x_1, x_2, \dots ?
- What is the functional form? (How does the box work?)



Types of goals:

1. **Descriptive**: Document observed relationships between X and Y .
 - Does not not necessarily tell us about the true DGP.
 - Helps us understand facts about the world.
2. **Causal**: Try to understand one piece of how the box works (the true DGP)
 - When you change one factor, how does it change the result?
 - Helps us make decisions about what to do
 - (in policy, business, personal life).
3. **Predictive**: Create your own box to try to match the output.
 - Doesn't matter if it works the same, or if you have the correct inputs.
 - Only matters how closely your box produces the same result.
 - Helps us know what's likely to happen in a new situation.



What methods should we use for each goal?

1. **Descriptive**: Exploratory analysis and regression
 - Covering soon!
2. **Causal**: Econometrics.
 - See your other classes.
3. **Predictive**: Statistical learning / machine learning.
 - Introduction in a few weeks!

Example

How did the under-developed and agricultural Switzerland from the 1850s become such an economically successful country?

A research question in history... where data are central

- Timing of the development of certain economic sectors?
- Geographical distribution of the firms?
- Interaction with public policies ?



Example: REFLEX project

- *Approach:* The Swiss Commercial Registry (1883-)



- *Methods:*
 - Transforming the pdf into structured data using:
 - google vision
 - Open AI
 - deep learning

Example: REFLEX project

- Objectives:

16 dicembre 1974. Impianti elettrici, ecc.
C. Cortinovis, in P o s c h i a v o. Titolare: Carlo Cortinovis, di nazionalità italiana, in Poschiavo. Impianti elettrici e consulenza tecnica. Borgo.

16. Dezember 1974.
ABAG Appartement-Bau AG, in St. Moritz. Liegenschaften (SHAB Nr. 111 vom 14. 5. 1973, S. 1390). Diese Firma wird infolge Verlegung des Sitzes nach Gstaad BE (SHAB Nr. 282 vom 2. 12. 1974, S. 3211) im Handelsregister des Kantons Graubünden von Amtes wegen gelöscht.

16. Dezember 1974.
GBAG, Gaststätten Betriebs A.G. Chur, in Chur (SHAB Nr. 57 vom 10. 3. 1965, S. 750). Cuoni Meier, Mitglied, zeichnet nun kollektiv zu zweien, statt wie bisher einzeln. Neue Verwaltungsräte: Reto Cottinelli, von und in Chur, Präsident; Paul Suter, von Seon AG und Chur, in Chur. Geschäftsführer: Kurt O. Winkler, von Winterthur, in Chur. Die Mitglieder des Verwaltungsrates und der Geschäftsführer zeichnen kollektiv zu zweien.

- 1 Event date
- 2 Firm name
- 3 Firm location (municipality)
- 4 Firm business
- 5 Reference to previous SHAB publication
- 6 Reason for mutation: firm relocation to commercial registry of different canton
- 7 Firm change of legal address
- 8 Reference to previous SHAB publication
- 9 Mutation in registry: firm deletion

- Challenges:

- Correcting for OCR mistakes
- Building a firm identifier
- 3 languages
- ...

Backbone of the class

1. The **skills**:

- Data collection
- Data manipulation:
 - Cleaning, Pipelines, data structure
- Data visualisation
- Data modelling

2. The **tools**:

- python
- git

3. The **concepts**:

- *Project management*: documenting, sharing & managing code
- *Reproducibility*

Public targeted: **anyone using data for projects.**

What this course is, and *is not*

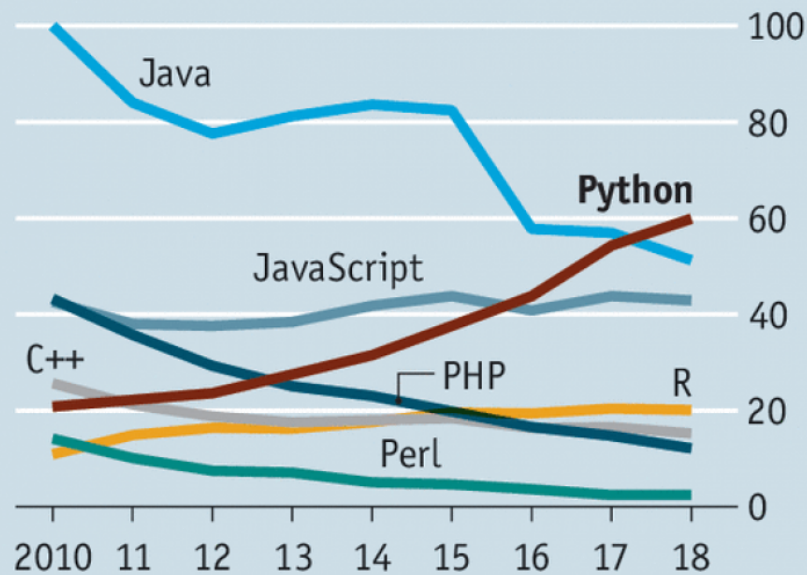
- It is:
 - **Applied** and oriented towards practice;
 - **General** overview of different techniques - what they are and how to use them.
 - **Data analysis** in general, not restricted to a research or a field (economics, political science).
 - In **python**.
- *It is not:*
 - **Computer science**.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

The tools: python & git

Why Python?

Biggus uptickus

US, Google searches for coding languages
100=highest annual traffic for any language



Source: Google Trends

Economist.com

Why choose Python for this class?

- Easy-to-use, wide-spread and popular programming language
- created by Guido Van Rossum (first released it in 1991 as Python 0.9.0.)
- General-purpose language
 - One of the core languages of scientific computing
- Elegant syntax

Python is a *high-level general-purpose* programming language that emphasizes *readability* and *extensibility*.

A high-level programming language

	High-level programming language	Low-level programming language
1.	It is programmer friendly language.	It is a machine friendly language.
2.	High level language is less memory efficient.	Low level language is high memory efficient.
3.	It is easy to understand.	It is tough to understand.
4.	Debugging is easy.	Debugging is complex comparatively.
5.	It is simple to maintain.	It is complex to maintain comparatively.
6.	It is portable .	It is non-portable.
7.	It can run on any platform.	It is machine-dependent.
8.	It needs compiler or interpreter for translation.	It needs assembler for translation.
9.	It is used widely for programming.	It is not commonly used now-a-days in programming.

A general-purpose programming language

- software development (alternative to Java, C, Fortran)
- data analysis
- good for webistes with data applications

... That consequently sports a huge library

- Web interaction
 - HTTP requests (**requests**)
 - Scraping web pages (**Beautifulsoup4**)
 - Performing Browser automations (**Selenium**)
- Data manipulation: **Pandas**
- Data visualisation:
 - Generic visualization **matplotlib**
 - Statistical data visualization **seaborn**
- Machine learning: **scikit-learn**
- Statistics: **statsmodels**
- Natural Language Proceession **nltk**, **spacy**

Using Python

Anaconda

Jupyter notebook

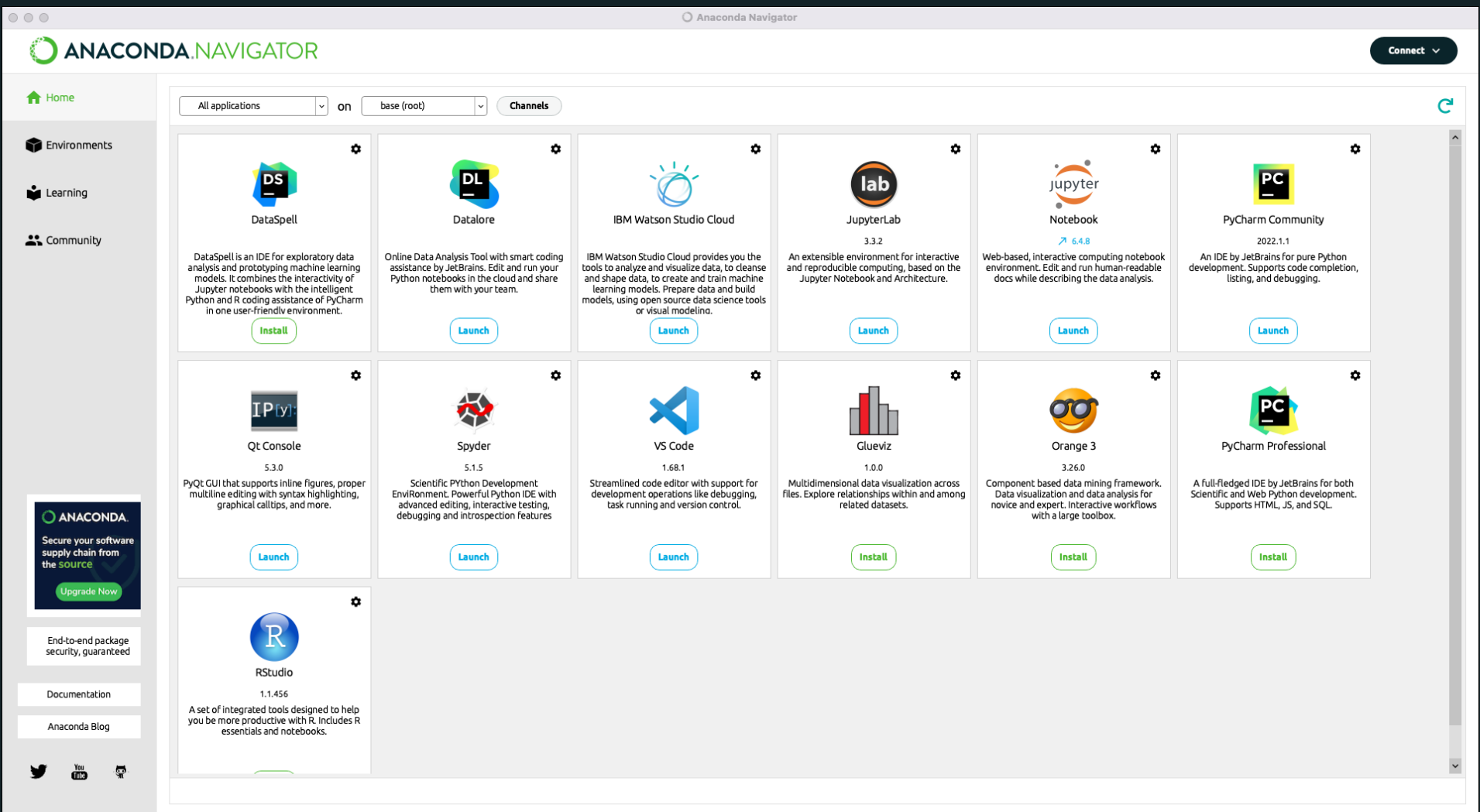
Visual studio code

a convenient all-in-one install

for homework

for longer code

→ Anaconda



Visual studio code & 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 &

Why git?

- we need a way to make version control and collaboration easier
 - *Version control*: managing changes to code, and easily going back and forth between older and new versions of a code base
- tool used in research & industry



Course organisation & 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

- **Mondays:** 1 hour / week
 - 1 hour practice, often an open house session around the class project
- **Wednesdays:** 3 hours / week
 - 2 hours of lecture + practice
 - sometimes the frontier between theory and practice will be fuzzy.
 - 1 hour of "individual work"
 - voluntarily
 - you can work at your pace on the course content and directly ask questions to the professor

Online Course Materials

- Website:
 - with the course architecture and content
 - <https://malkaguillot.github.io/ECON2206-Data-Management/>
- Github repository :
 - <https://github.com/malkaguillot/ECON2206-Data-Management-2023>
 - includes slides and tutorials
 - Can be tested under a temporary computing environment:



- [lola](#):
 - For homeworks
 - Course announcement and forum



[Evaluation Policy]

- 4 problem sets:
 - should be given back as jupyter notebooks in PDF format on lola.
 - =20%
 - The problem sets are simple exercises designed to help students to “get their hands in the data & code”.
- Participation in class = 5%:
- Course project = 75%

cf. details

[Course project] Objectives

- The **basics**:
 - End-to-end data project using Python
 - From collection to modelling through visualisation
 - Should answer an open-question with data
 - 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 data analysis
- **Deepen one aspect** of the course:
 - Three dimensions must be present in the project:
 - data retrieval, visualisation or modelling;
 - To a greater or lesser extent depending on the project.
 - One dimension must go further than the others



[Course project] Example of web application

→ Some examples in various sector:

- Finance:
 - The **Yield Curve**
- Health
 - **Opioid epidemic in the US**
- Transportation:
 - **Uber rides**
- **Energy consumption**
- **Research project**

→ Be creative, have fun!

[Course project] Other example

- Modelling hotel prices ;
- Mapping of cycling practices using Villo!'s data
- Mapping social and economic inequality in Wallonie
- What is the parliament talking about?

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, social inequality
 - or not: sport, important topic

[Course project] Requirements

1. Data retrieval

- Standard: Get some data from the web, data cleaning
- Advanced: Data collection with webscraping

2. Data visualisation

- Standard: Propose at least 4 exhibits (graphs or tables)
- Advanced: More elaborated figures, such as maps, or unsupervised learning representations

3. Data modelling

- Standard:
 - Supervised ML approach: compare at least 2 models or NLP
- Advanced:
 - Both a ML model and some NLP (for example)

[Course project] Requirements


- Submission format:
 - Invite [@malkaguillot](#) and [@MichelCop](#) to collaborate on your GitHub repository by the due date.
- Abide by **good coding practices**:
 - code must be split into meaningful sub-files;
 - code must be documented;
- Requirement for **reproducibility**:
 - Code should replicate
 - Project must be available on GitHub

[Course project] Evaluation: 75% =

Indicative weights

- **Project management = 5% (G)**
 - reproducibility, github, readme
- **Project relevance = 10% (G)**
 - Does the project respond to an interesting/important question?
- **The 3 dimensions of the class = 30% (G)**
 - **Data retrieval**
 - **Data visualisation**
 - **Data modelling**
- **One dimension that goes further = 10% (G)**
- **Oral presentation = 15% (I)**
- **Group feedback = 5% (I)**

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 textbook. Specific references will be given when corresponding subjects are tackled.

- List on the course website
- Online books:
 - Coding for Economists
 - Python for Economics and Business Research introduction to python, pandas, plotting
- Stackoverflow: all the answers are there, but you have to ask the right question.

Troubleshooting

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