Web Scraping

ECON 4810

Why Web Scraping?

- Many websites do not provide a way to download data.
- Even if they do, you might want to automate the process because many files are needed, or the files are constantly updated.

Possible Scenarios

Downloading Files

Scraping Static Pages

Scraping Dynamic Pages

Python Solutions

urllib.request.urlretrieve

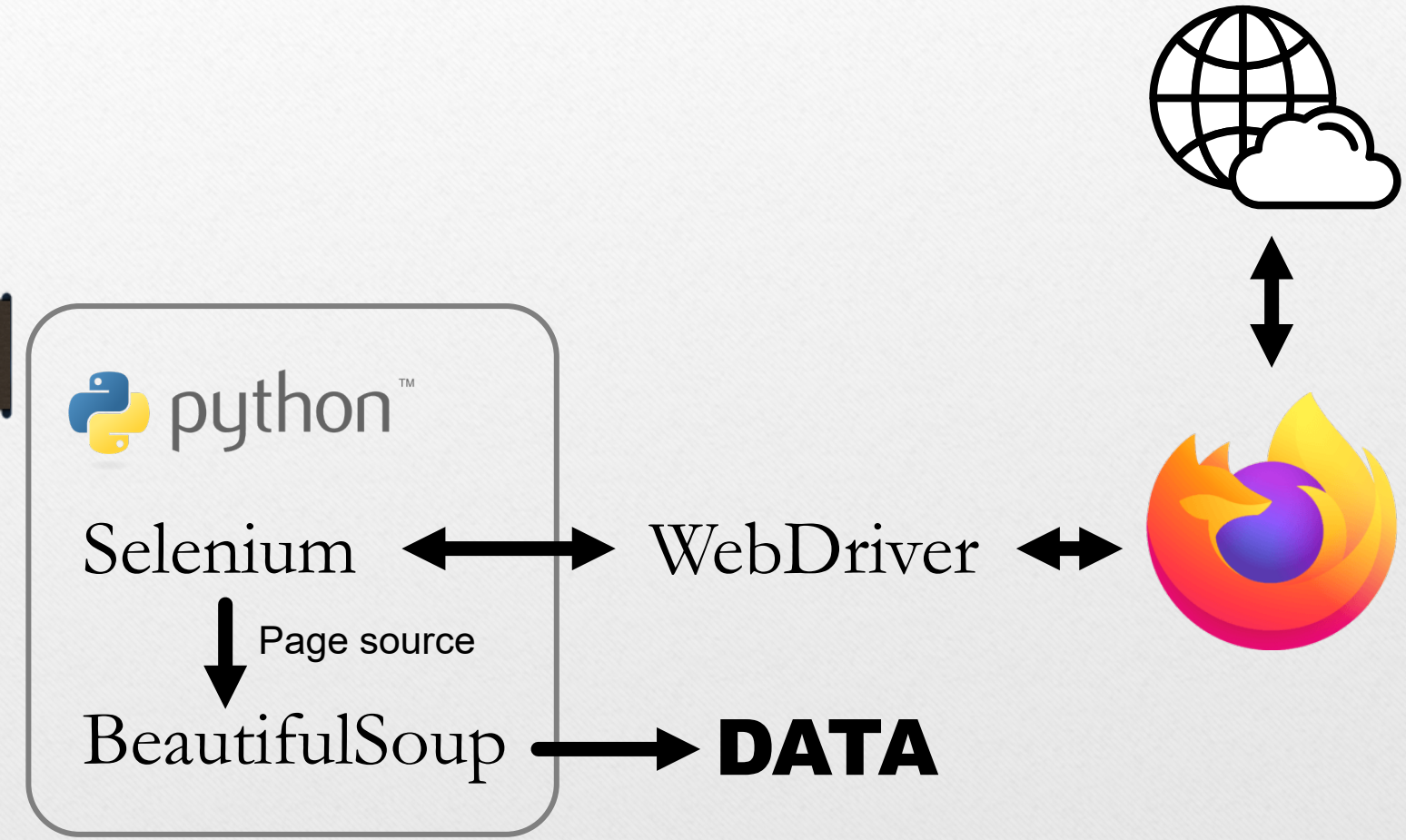
requests
+ beautifulsoup

selenium + browser +
beautifulsoup

How to Tell if a Webpage is Dynamic

- If the site's content can change without needing a page refresh, it is certainly dynamic.
- Even if the content does not change, it might still be generated dynamically. Best to check page source.

Scraping a Dynamic Page



Scraping Regularly

- You could run a Python script that never stops, scraping data regularly in the script.
- …but if the script crashes, you will have to be there to restart it in time.
- Use operating system's built-in tool to run a script periodically
  - Windows: **task scheduler**
  - Linux: **crontab**