# Web Scraping and Web Crawling

Sjur Løne Nilsen

July 10th, 2023

University of Bonn

# What this lecture will cover

- What web scraping and web crawling are, and an example of how we can use it.
- Utilizing Seleniums automation capabilities to navigate through web pages, extract data, and perform actions like clicking buttons. This will allow you to collect specific information from websites in an automated manner.

# Crawling and Scraping the Web

# Web Crawling vs. Web Scraping

## Web Crawling

Web crawling is an *automated* process where software, known as a **crawler or spider**, systematically browses the internet by following links to discover and collect information from web pages **given a set of instructions**

# Web Crawling vs. Web Scraping

## Web Crawling

Web crawling is an *automated* process where software, known as a **crawler or spider**, systematically browses the internet by following links to discover and collect information from web pages **given a set of instructions**

## Web Scraping

Web scraping is the *automated* extraction of data from websites. It involves using software tools or scripts **to parse the HTML structure of web pages**, target specific elements, and retrieve desired information.

# What do you need to engage in web scraping? (in our case)

**Python**
Obivoulsy. We will use the Selenium framework in Python.

**Browser**
A internet browser to access the web. We will use Google Chrome.

**Web Driver**
Some tool to automatically interact with a browser.

**Targets**
Meaning some kind of information that you either want to collect or interact with.

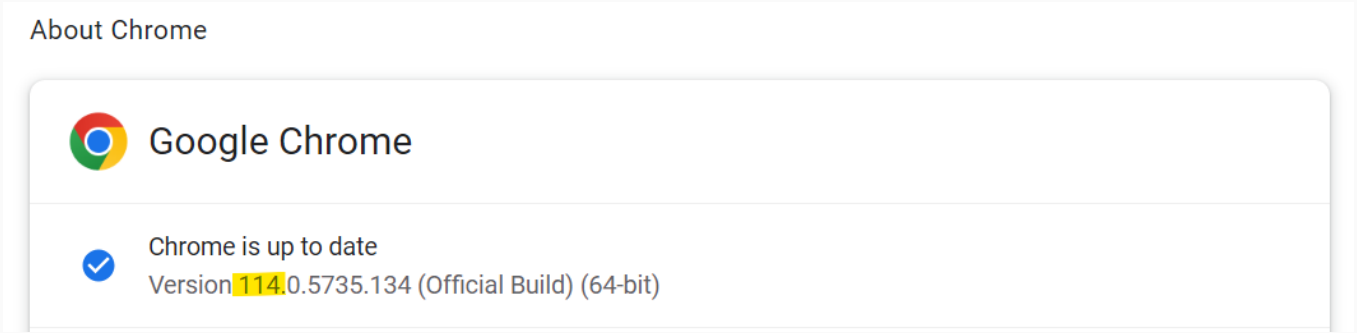
# TASK 1: Collecting our ingredients (10 minutes)

1. Open this link (ChromeDriver) or google ChromeDriver, and find Downloads in the menu.
2. Check your version of Chrome by clicking the three dots in the upper right corner of your Google Chrome.

   `Customize and control Google Chrome`
   — `Help`
   —— `About Google Chrome`

   Consider updating if possible, as it will make it easier to find the corresponding driver.

4. Download the corresponding version of ChromeDriver. This version should match the first three numbers in both the driver and Chrome.




5. Locate the zip folder you just downloaded, open it, and copy the file called `ChromeDriver.exe`.
6. Paste it in somewhere where that you can easily access the path!
   I chose: `C:/Program Files (x86)/chromedriver.exe`

# Let's open a Jupyter Notebook in VS code