Google Developer Student Clubs

# Web Scraping Workshop

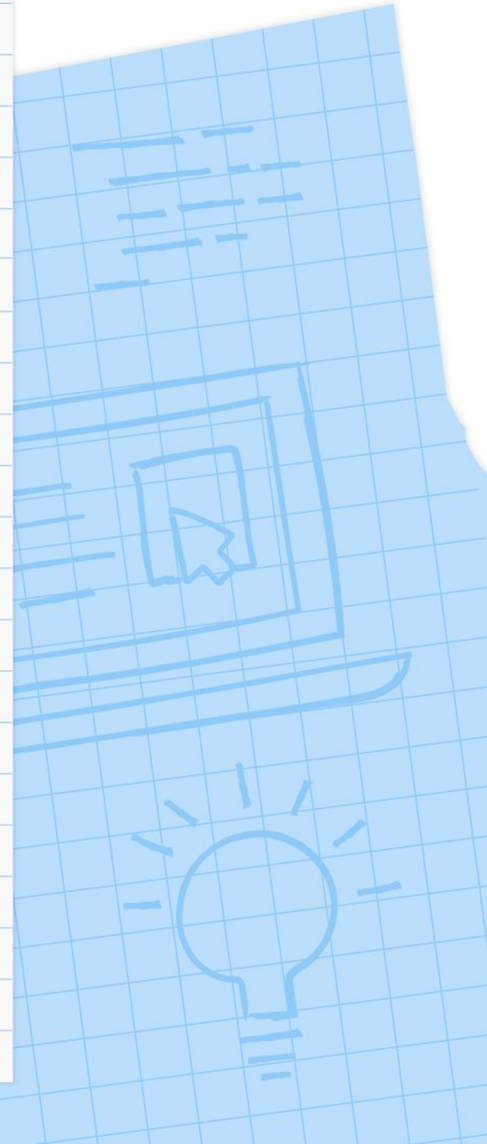
## Using Beautiful Soup Package for Python

Daniyal Bokhari

# What is Web Scraping?

## The textbook definition

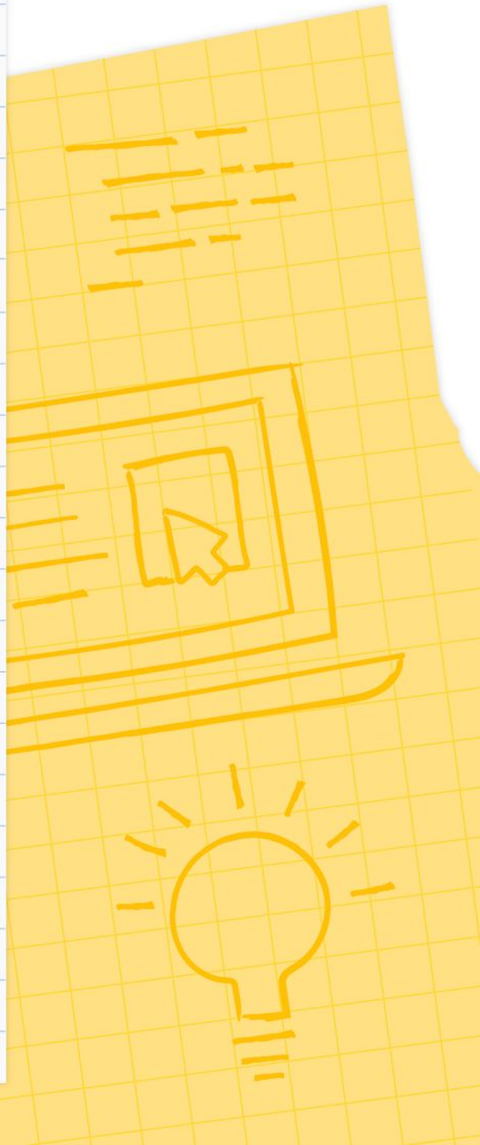
- Web scraping is the act of extracting data from website
- Typically performed using automated tools to reduce time spent on collecting data from websites
- Once data is extracted from websites, it can be parsed to access specific pieces of information relevant to your search



# What is it used for?

## Some common uses

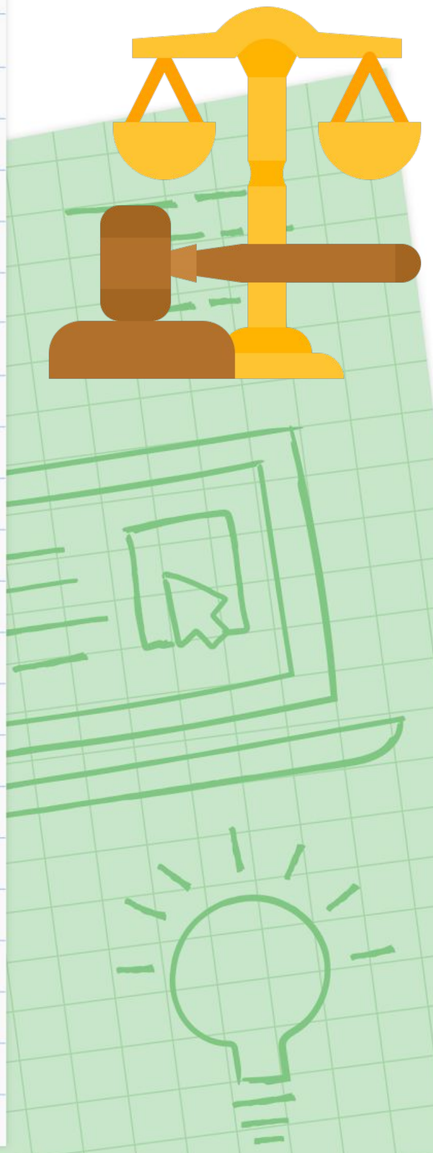
- Industry statistics/insights
- By consumers to compare prices across different stores
- Analyze stock or cryptocurrency prices
- Academic research



# When should it not be used?

## The legality of it

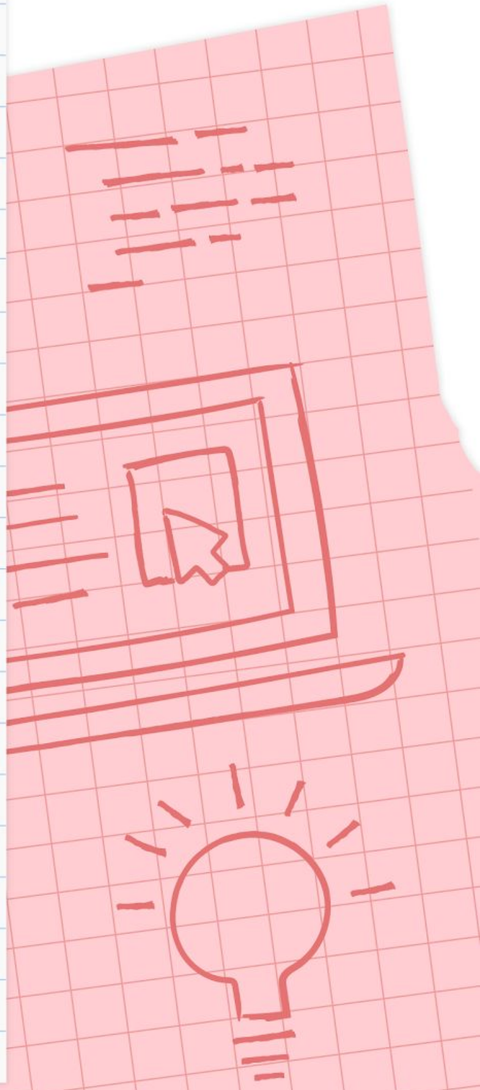
- Web scraping is generally considered to be legal when it's used to find publicly available information
- Should not be used to collect personal information, data that the poster did not intend to be publicly available
- Should not be used to access data that requires you to create an account or login



# BeautifulSoup

## A Python library for web scraping

- Beautiful Soup is a Python library that helps in web scraping by providing useful methods to parse through the HTML code of websites
- Creates python objects out of elements of the HTML code allowing you to easily search and traverse through the contents of websites



# Beautiful Soup vs Selenium


## Which is better?

- Selenium refers to multiple different open-source projects, the Selenium API can be used to control browsers
- Much more complex than Beautiful Soup, can interact with websites by pressing buttons and filling out fields
- Used for automation and testing
- Can also be used to perform web scraping
- Beautiful Soup is useful for simple projects, can extract information faster since it doesn't load up all of the web page content and JavaScript

# Choice of Parser

## Which one should you use?

- In order to use Beautiful Soup, you must specify a parser
- Parsers take the HTML code we receive from a website and construct an object which we can use in Python to navigate and search through this code
- Python has a built-in HTML parser called html.parser
- We will be using lxml, an external parser, as it is faster at parsing HTML than html.parser and works for more versions of Python

# Live Demo

## The best way to learn is to practice

- You will need to install Beautiful Soup, lxml, and the requests library

```
Run these in a terminal:

pip install beautifulsoup4
pip install lxml
pip install requests
```

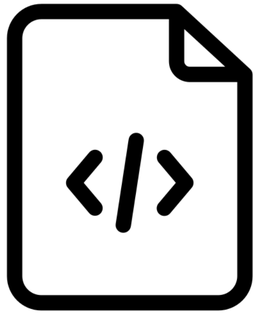
```
Alternate commands:

pip3 install <module>
Python3 -m pip install <module>
Python3 -m pip3 install <module>
```

- You can use your favourite IDE/text editor (Pycharm, VS Code, Vim, etc) to follow along

# Where can I learn more about this?

## Many resources

- The documentation is a great place to look at all available functions
- Online tutorials, YouTube videos are also a great place to learn more about web scraping



Beautiful Soup 4 documentation: https://www.crummy.com/software/BeautifulSoup/bs4/doc/