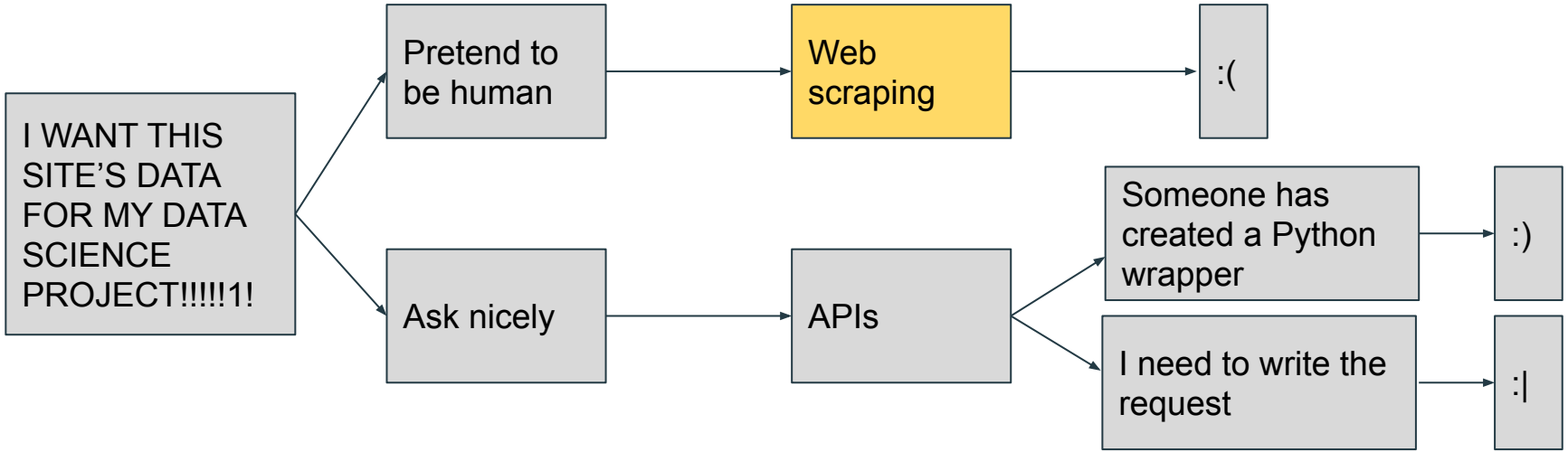
Web Scraping

Find unique data through immense suffering

Two ways to get data from the internet

...and how (un)happy they'll make you



The rewards

Applications of web scraping

- **Market & product research** (brands, products, prices, reviews… of your competitors!)
- **Lead generation** (find people that might be interested in your product!)
- **Keyword research** (what is people looking for?)
- **“People Analytics”** (HR/Talent research: job descriptions, skills, offers…)
- **Science** (stuff that’s not present in official records or traditional datasets: subletting apartments, drug use, people’s language & beheavior…)
- **Stock Analysis** (using data in the web to forecast)
- **Media** Gathering & Analysis
- Lots of cool personal projects: what websites do you like?

Is it legal?

- It depends on the site, on what you do with the scraped data.
- As legal or as illegal as downloading data manually from the internet: there might be copyright issues.
- You probably want good legal advice if you have to make money with it.
- Scraping public government web sites & scraping for personal use is generally safe.

Some websites have a file called “robots.txt” stating whether or not you should scrape them. That does not have legal implications, it’s more an ethical thing.

The tools

Beautiful Soup

- User friendly
- Popular (lots of tutorials, stackoverflow posts...)
- Slow

Scrapy

- Fast
- Requires no dependencies
- Not user friendly (set up pains)

Selenium

- Versatile (also works for testing)
- Can scrape JavaScript content
- Difficult
- Slow

HTML

All these things are HTML tags

```
<html>
    <body>
        <div>
            <p>Hello class!</p>
            <p>Wave into the cam if you already know html</p>
        </div>
        <p>Wow you are so knowledgeable</p>
    </body>
</html>
```

Tags can have attributes

This tag has 2 attributes

```
<div id="unique-id" class="some-class">
        ...div element contents...
</div>
<a href="https://www.reddit.com">
        Don't click this link if you want to stay productive
</a>
```

The href attribute contains a link

HTML tags reference: https://www.w3schools.com/TAGS/default.ASP

Chrome dev tools

On Google Chrome, press CTRL + SHIFT + I

BeautifulSoup

```
import requests
from bs4 import BeautifulSoup

# download the webpage
url = “https://www.the-website-you-wanna-scrape.com”
page = requests.get(url)

# parse the html
soup = BeautifulSoup(page.content, 'html.parser')

# print the formatted html
print(soup.prettify())
```

How to search content

```
# using HTML tags
soup.find_all(“div”, class_=“class-name”)
```

```
# using CSS Selectors
soup.select("body div.class-name a")
```

https://flukeout.github.io/