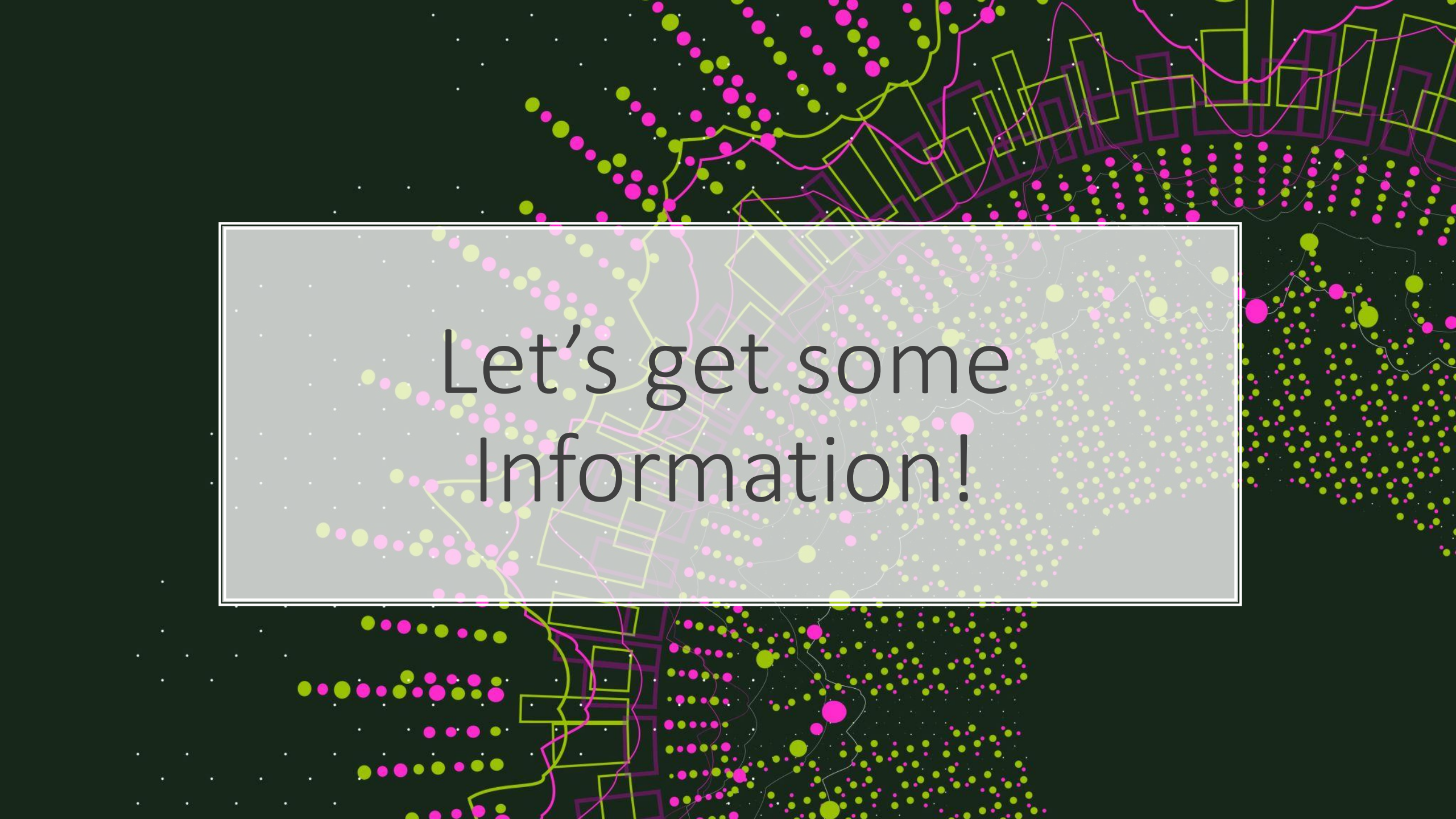
Welcome

Review from last time!

- Data and information
- Examples of data sources
- First Party vs. third party data
- APIs


Let’s get some
Information!

3 minutes: Get the top stocks and changes from the top of this website

https://www.nasdaq.com/

Let’s say I had 10,000 stocks. How can I speed this up?

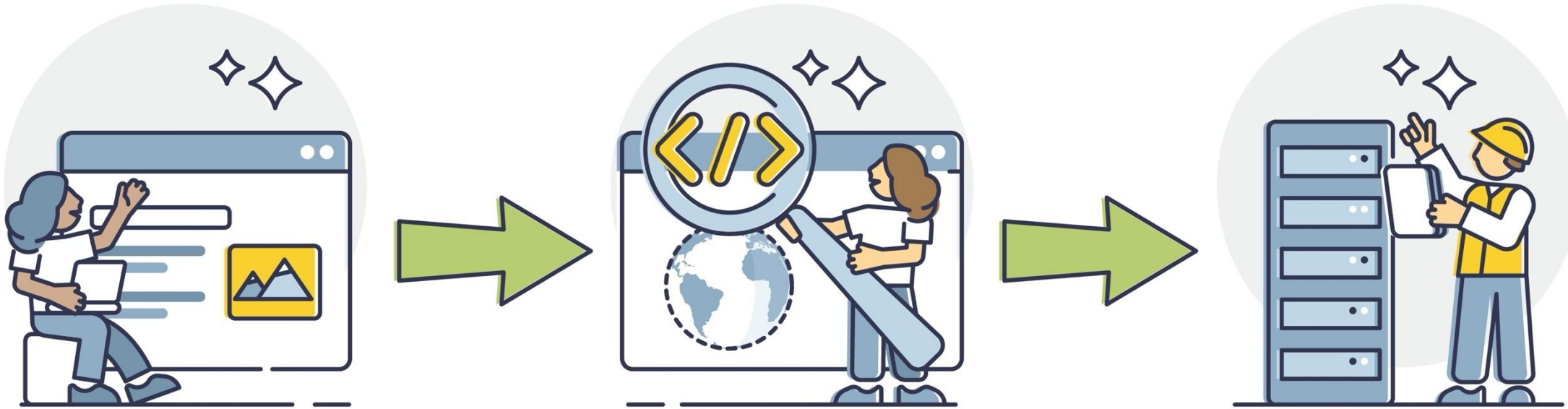
AN API

(IF THERE IS ONE)

HIRE PEOPLE TO DO IT
(MTURK)

WEB SCRAPPING

WEB SCRAPING

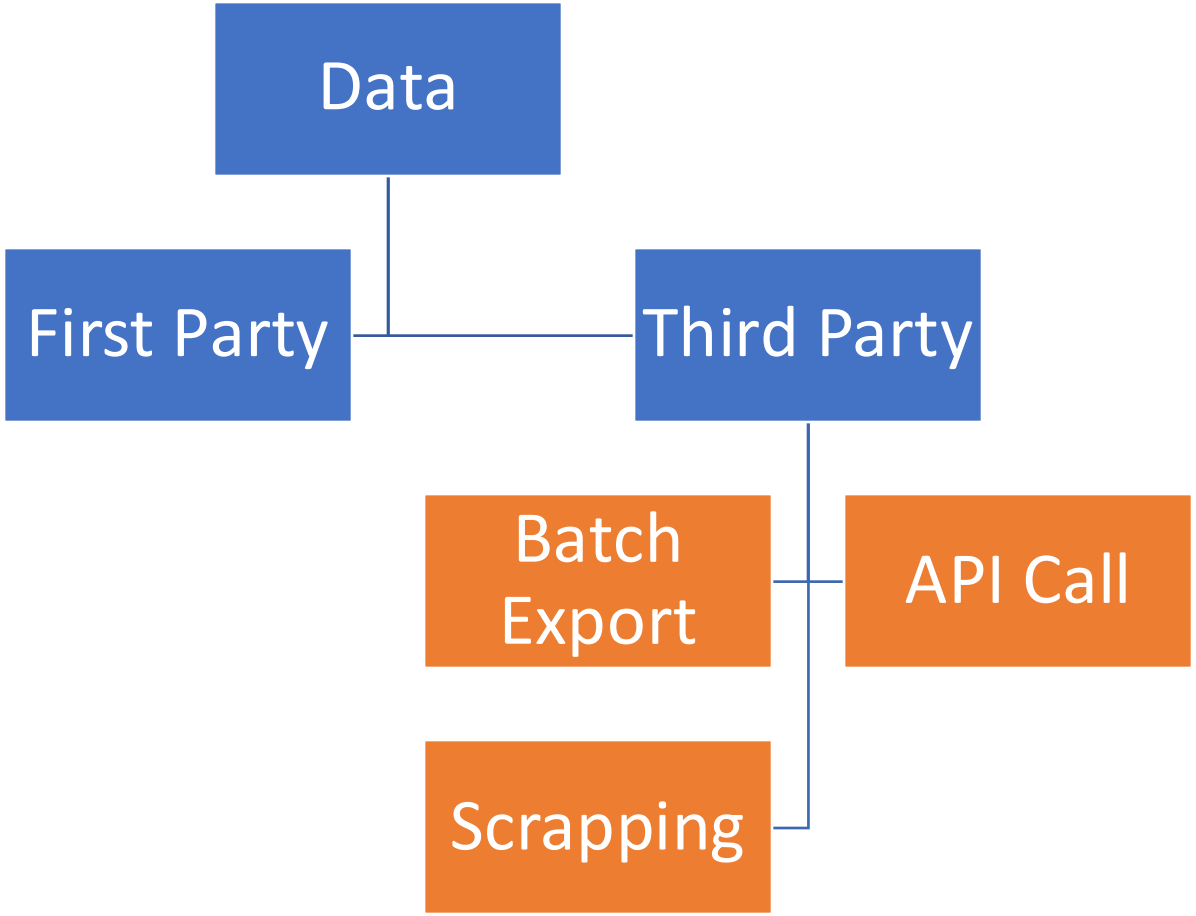


HTML WEBSITES

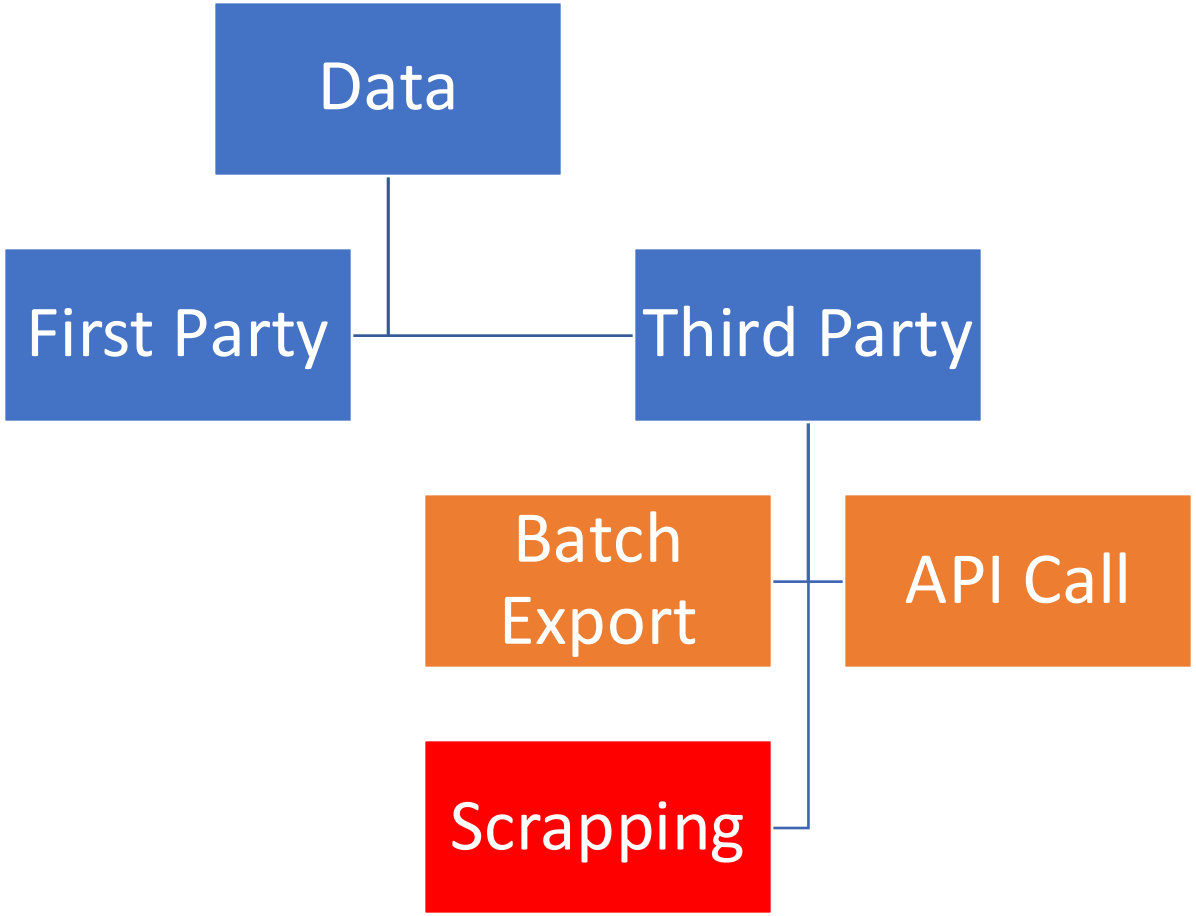
WEB SCRAPING

DATA

Reminder



Reminder



What is Web Scraping

Web scraping, also known as web harvesting or web data extraction, is a process of fetching a web page and extracting data from it

Why webscraping vs. api?

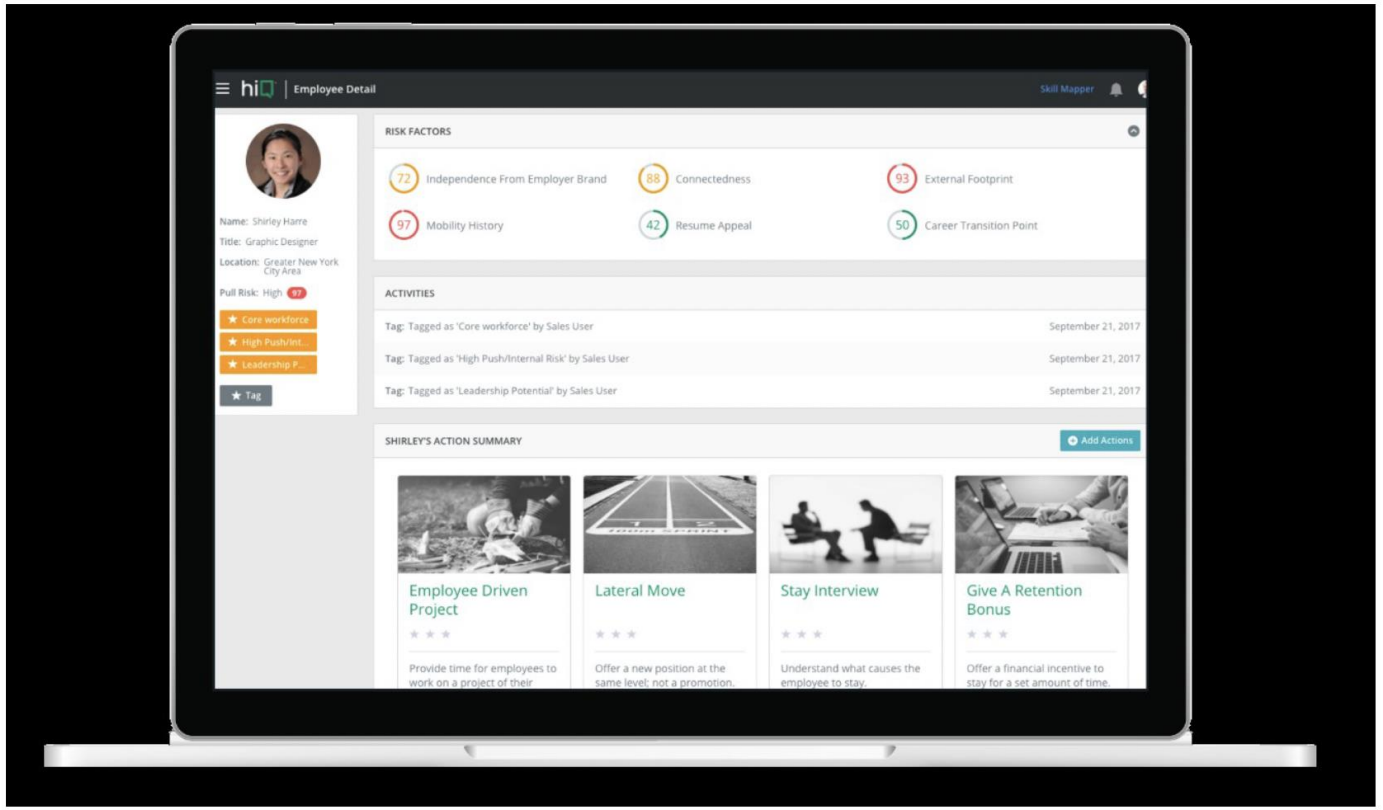
- Web scraping is fragile. It can easily break (stop working) when changes are made to the website.
- Companies often implement technology to try to stop web scraping.
- Sometimes it’s illegal, depending on your use case.
- But sometimes, it is your best option because an API is not available to you!

Examples

- Real Estate: What properties are for sale or for rent?
- Industry Stats and Insights: what are employment trends?
- eCommerce comparison sights: what retailer offers this product least expensively?
- Lead generation: what is the contact information of people who might be interested in my product?
- Social Media Sentiment Analysis: how do people like my product?

HiQ scrapes LinkedIn

Which employees are currently thinking about quitting?



Doesn't exist anymore?

hiQ Labs v. LinkedIn

Article Talk

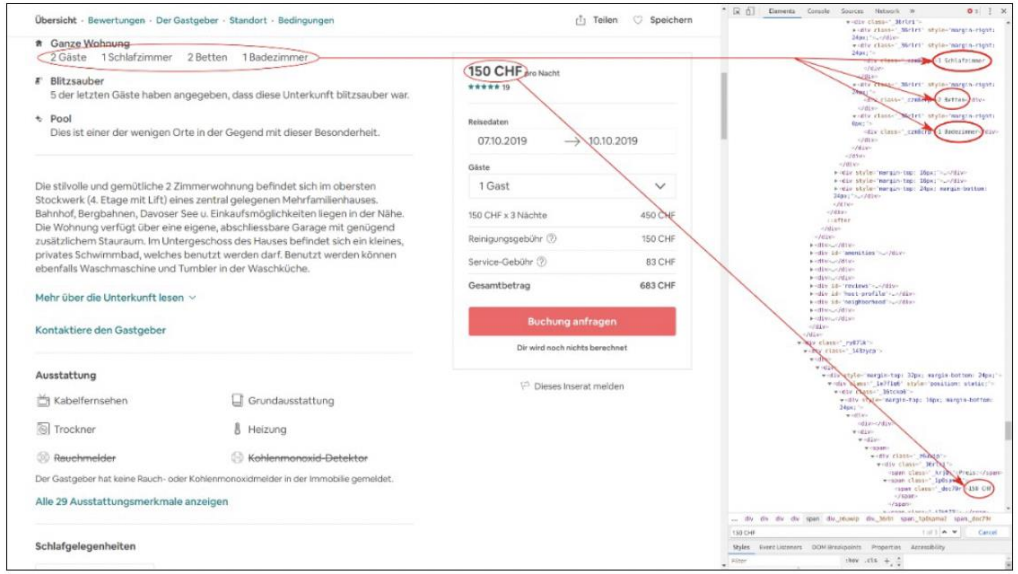
From Wikipedia, the free encyclopedia

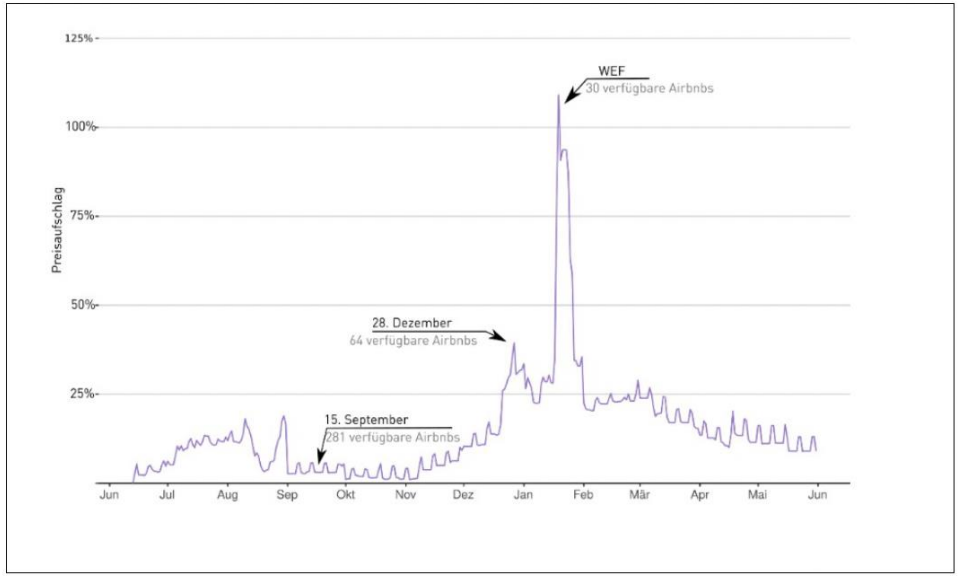
hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985 (9th Cir. 2019), was a United States Ninth Circuit case about web scraping. The 9th Circuit affirmed the district court's preliminary injunction, preventing LinkedIn from denying the plaintiff, hiQ Labs, from accessing LinkedIn's publicly available LinkedIn member profiles. hiQ is a small data analytics company that used automated bots to scrape information from public LinkedIn profiles.

The court ruled that hiQ had the right to do web scraping.[1][2][3] However, the Supreme Court, based on its *Van Buren v. United States* decision,[4] vacated the decision and remanded the case for further review in June 2021. In a second ruling in April 2022 the Ninth Circuit affirmed its decision.[5][6] In November 2022 the U.S. District Court for the Northern District of California ruled that hiQ had breached LinkedIn's User Agreement and a settlement agreement was reached between the two parties.[7]

Datahouse.ch

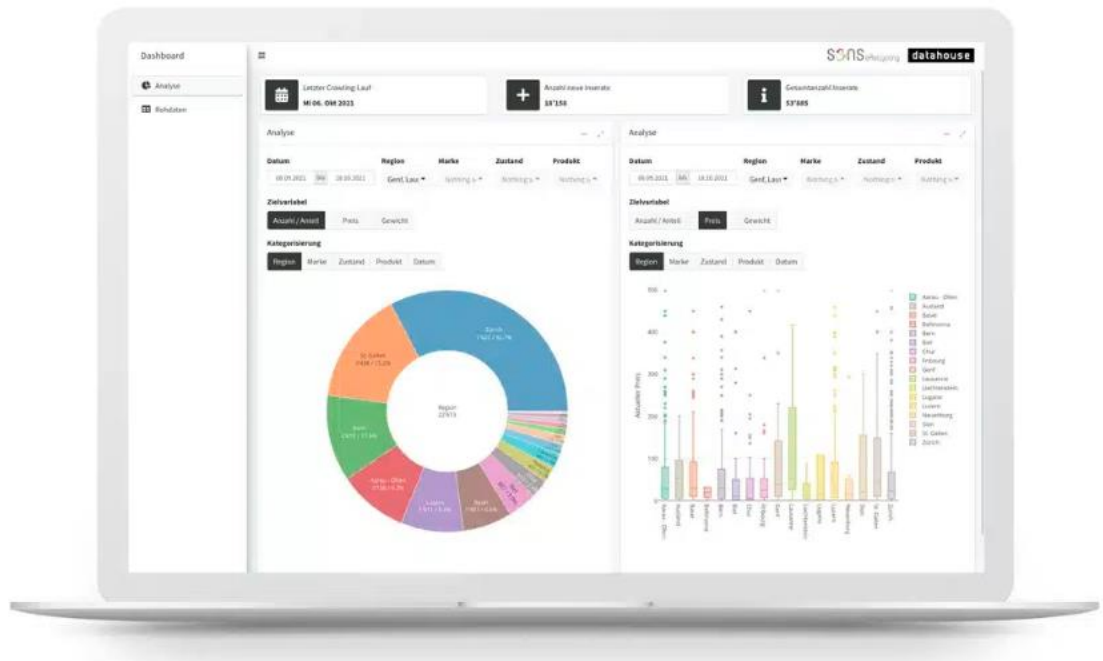
How do hotel prices develop over time?





Web crawling dashboard

October 21, 2021



Investigating the indoor environmental quality of different workplaces through **web-scraping** and text-mining of Glassdoor **reviews**

G Chinazzo - Building Research & Information, 2021 - Taylor & Francis

… on **web-scraping** and text-mining of online job **reviews**. A total of 1,158,706 job **reviews** posted on … Within these **reviews**, 10,593 include complaints about at least one IEQ aspect. The …

☆ Save 99 Cite Cited by 16 Related articles All 3 versions

[PDF] Don't Lie To Me: Integrating Client-Side **Web Scraping** And **Review** Behavior Analysis To Detect Fake **Reviews**

BJ Levinson - 2019 - core.ac.uk

… **review** title length to be very useful signals of **review** manipulation, we did find **review** length to be correlated to **review** manipulation across our datasets, with shorter **reviews** being …

☆ Save 99 Cite Cited by 1 Related articles All 2 versions »

Web scraping for hospitality research: Overview, opportunities, and implications

S Han, CK Anderson - Cornell Hospitality Quarterly, 2021 - journals.sagepub.com

… Then, we explain the sample **web scraping** code for collecting online hotel **reviews** and … complete scripts for scraping **reviews** and prices from major OTAs, **review** sites and hotel brand …

☆ Save 99 Cite Cited by 41 Related articles »

[PDF] Analyzing and Filtering Food Items In Restaurant **Reviews**: Sentiment Analysis and **Web Scraping**

N Luo, C Kwan, Y Sun, F Zhang - … & Information Technology (CS & IT) …, 2020 - csitcp.org

… **reviews** on the popular **review** site known as Yelp. We have created an application that uses **web scraping**, … Through analyzing two different ways to obtain Yelp **reviews** and evaluating …

☆ Save 99 Cite Cited by 1 Related articles All 5 versions »

Web scraping and data science in applied research in communication: A study on online **reviews**

MT de Farias, ACB Angeluci… - Revista Observatório, 2021 - sistemas.uft.edu.br

With the spread of access and use of information through the web and social networks, information retrieval in large volumes of data has become unfeasible by manual methods. In this …

☆ Save 99 Cite Cited by 2 Related articles All 3 versions »

Reviews, Reputation, and Revenue: The Case of Yelp.Com

Harvard Business School NOM Unit Working Paper No. 12-016

41 Pages • Posted: 16 Sep 2011 • Last revised: 16 Mar 2016

Michael Luca

Harvard University - Business School (HBS)

Date Written: March 15, 2016

Abstract

Do online consumer reviews affect restaurant demand? I investigate this question using a novel dataset combining reviews from the website Yelp.com and restaurant data from the Washington State Department of Revenue. Because Yelp prominently displays a restaurant's rounded average rating, I can identify the causal impact of Yelp ratings on demand with a regression discontinuity framework that exploits Yelp's rounding thresholds. I present three findings about the impact of consumer reviews on the restaurant industry: (1) a one-star increase in Yelp rating leads to a 5-9 percent increase in revenue, (2) this effect is driven by independent restaurants; ratings do not affect restaurants with chain affiliation, and (3) chain restaurants have declined in market share as Yelp penetration has increased. This suggests that online consumer reviews substitute for more traditional forms of reputation. I then test whether consumers use these reviews in a way that is consistent with standard learning models. I present two additional findings: (4) consumers do not use all available information and are more responsive to quality changes that are more visible and (5) consumers respond more strongly when a rating contains more information. Consumer response to a restaurant's average rating is affected by the number of reviews and whether the reviewers are certified as "elite" by Yelp, but is unaffected by the size of the reviewers' Yelp friends network.

Individually, take 5 minutes to think about a research/business question that you would like to explore using web scrapping.

Then, we'll take 2 minutes to share that with your neighbor.

Legal and Ethical Considerations

- In 2014/2018, the German Federal Court of Justice (Bundesgerichtshof) ruled that the scraping of publicly available data, including personal data, is generally permissible under German data protection laws
- In the United States, web scraping can be considered legal as long as it does not infringe upon the Computer Fraud and Abuse Act (CFAA), the Digital Millennium Copyright Act (DMCA), or violate any terms of service agreements.
- UK Case: Data scraping is not in itself illegal in the UK, but the data (factual or otherwise) may be subject to copyright. Using it without the copyright owner's permission, especially if you are selling it, could lead to legal action.
- In Canada, the legality of web scraping has not been fully defined. In 2011, a B.C. court sided with a company that accused another website of scraping its content without authorization. … But that website had a specific clause against scraping. With other websites, especially government data, the rules are not so clear.
- The CAN-SPAM Act of 2003 specifically prohibits the practice. Beyond the illegality, however, there are many other reasons to avoid email scraping. … You might have thousands of email addresses in your database, but you do not have the consent of the email owners to receive your emails.
- In Switzerland, scraping is allowed as long as the data is not subject to copyright (see Jusletter).
- With GDPR, one needs consent of individuals if personalized data should be stored.

yelp Support Center

How can we help?

Ask a question... Search

Support Center

Can I copy or "scrape" data from the Yelp site?

No – Yelp does not allow any “scraping” of the site, and does not permit the use of any third party software, including bots, browser plug-ins, or browser extensions (also called "add-ons"), that "scrapes" or copies Yelp reviews, business pages, photos or profile information. Such tools violate our Terms of Service, including many of the restrictions listed specifically in **Section 6(b)**. Please read that section for full details, but to put it simply, you’re not allowed to:

- Exploit the site by taking content for display or sale, even if you’ve modified it
- Scrape or index any portion of the slte through any means, including bots or spiders, or for any purpose
- Record, process, or mine information about users

Any user who uses tools for such purposes is in violation of the Terms of Service – Yelp may restrict or terminate such users’ access to the site, and reserves all rights. Of course, you can share or embed reviews, or use content in other ways expressly authorized by Yelp, and we have a dataset available on our Yelp Dataset Challenge page (subject to certain restrictions).

In short: please don’t scrape our site!

comments or suggestions as we see fit and without any obligation to you.

Misuse of the Services

You also agree not to misuse the Services, for example, by interfering with them or accessing them using a method other than the interface and the instructions that we provide. You agree that you will not work around any technical limitations in the software provided to you as part of the Services, or reverse engineer, decompile or disassemble the software, except and only to the extent that applicable law expressly permits. You may not do any of the following while accessing or using the Services: (i) access, tamper with, or use non-public areas of the Services, our computer systems, or the technical delivery systems of our providers; (ii) probe, scan, or test the vulnerability of any system or network or breach or circumvent any security or authentication measures; (iii) access or search or attempt to access or search the Services by any means (automated or otherwise) other than through our currently available, published interfaces that are provided by us (and only pursuant to the applicable terms and conditions), unless you have been specifically allowed to do so in a separate agreement with us (NOTE: crawling or scraping the Services in any form, for any purpose without our prior written consent is expressly prohibited); (iv) forge any TCP/IP packet header or any part of the header information in any email or posting, or in any way use the Services to send altered, deceptive or false source-identifying information; (v) engage in any conduct that violates our Platform Manipulation and Spam Policy or any other Rules and Policies; or (vi) interfere with, or disrupt, (or attempt to do so), the access of any user, host or network, including, without limitation, sending a virus, overloading, flooding, spamming, mail-bombing the Services, or by scripting the creation of Content in such a manner as to interfere with or create an undue burden on the Services. It is also a violation of these Terms to facilitate or assist others in violating these Terms, including by distributing products or services that enable or encourage violation of these Terms.

Use of Public Data

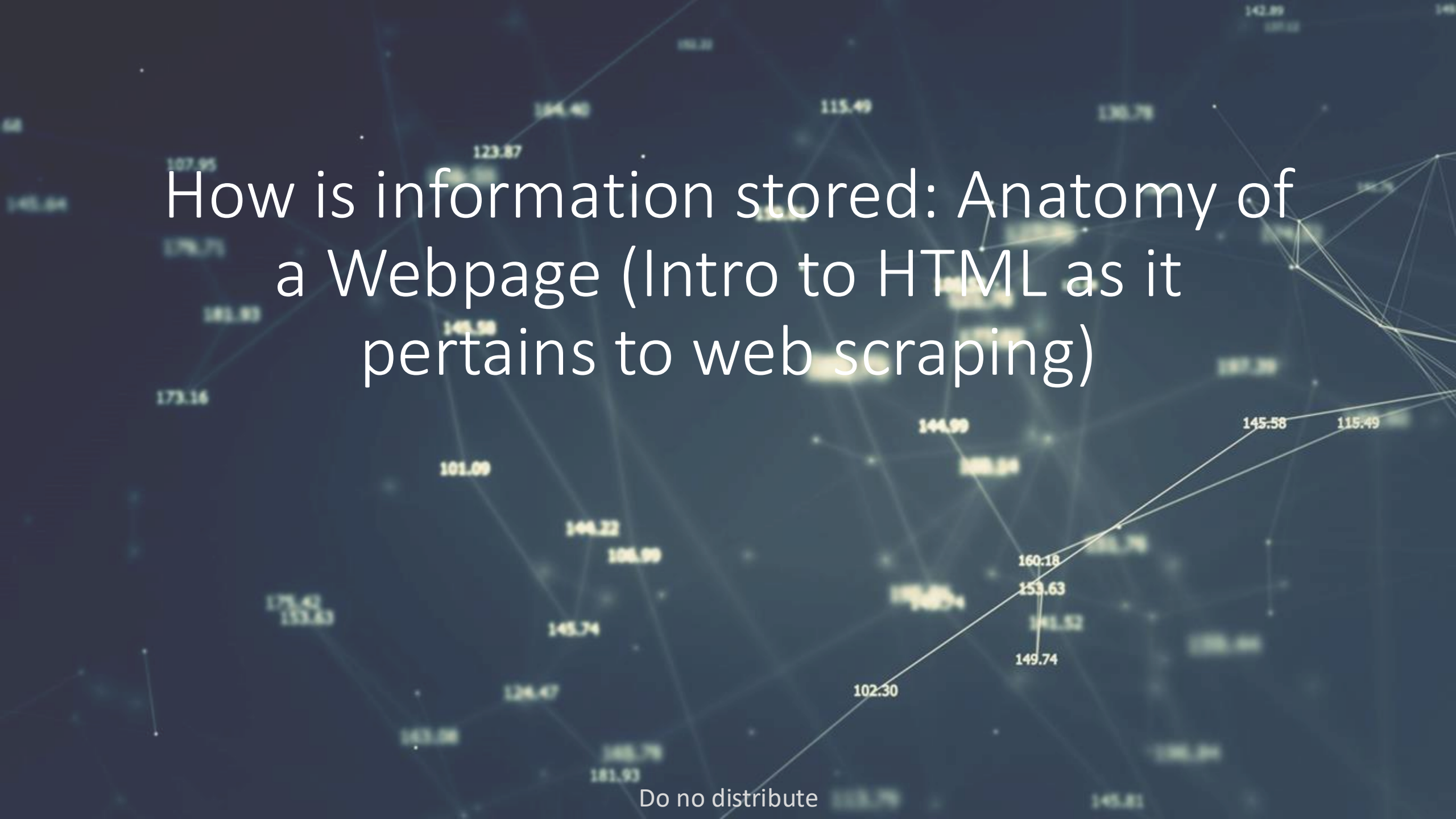
- Often OK for most use cases, but double check with Terms of Use agreement
- Technical considerations: don’t unintentionally cause a denial-of-service attack

Personal, Research, Commercial Use of Non-Public Information

- Always check terms of use service.
- Technical considerations: don’t unintentionally cause a denial-of-service attack
- General Guidance:
  - Personal: Generally OK
  - Research: Often done. Having permission is better.
  - Commercial: Often require permission

What can companies do to you?

- Deny you access
- File lawsuit, especially if you are gaining a commercial advantage on copyrighted material or causing technical problems.
  - Financial consequences
  - Cease and desist



How is information stored: Anatomy of a Webpage (Intro to HTML as it pertains to web scraping)

Intro to HTML

- Elements
- Tags
- Attributes
- Content
- Ids
- Classes
- Parents
- Children

```
<!DOCTYPE html>
<html>
<body>
<div id="main">
  <h1>Introduction to Webscraping</h1>
  <p class="instructions">Welcome to our class</p>
  <span data-content="main">
    <a href="https://www.w3schools.com/html/html_intro.asp">Learning html is fun</a>
  </span>
</div>

</body>
</html>
```

Intro to HTML – Key Components

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Page</title>
  </head>
  <body>
    <h1>Welcome to Web Scraping!</h1>
    <p>This is a paragraph of text.</p>
    <a href="https://example.com">Click here for more</a>
  </body>
</html>
```

Doctype Declaration: defines the document type

HTML Tag: the root of the page

Head Section: metadata like the title and links to CSS/JavaScript

Body Section: The visible content of the webpage

Intro to HTML – Tags, Attributes and Content

Tags
<h1>, <p>, <a>, etc., define the purpose of the content

```
<!DOCTYPE html>
<html>
  <head>
    <title>Example Page</title>
  </head>
  <body>
    <h1>Welcome to Web Scraping!</h1>
    <p>This is a paragraph of text.</p>
    <a href="https://example.com">Click here for more</a>
  </body>
</html>
```

Content
The data displayed to the user

Attributes
Add details to tags (e.g., href specifies the link for <a>)

Intro to HTML – IDs, Classes

```
<!DOCTYPE html>
<html>
<body>
<div id="main">
  <h1>Introduction to Webscraping</h1>
  <p class="instructions">Welcome to our class</p>
  <span data-content="main">
    <a href="https://www.w3schools.com/html/html_intro.asp">Learning html is fun</a>
  </span>
</div>

</body>
</html>
```

Id

Unique to one element on a page

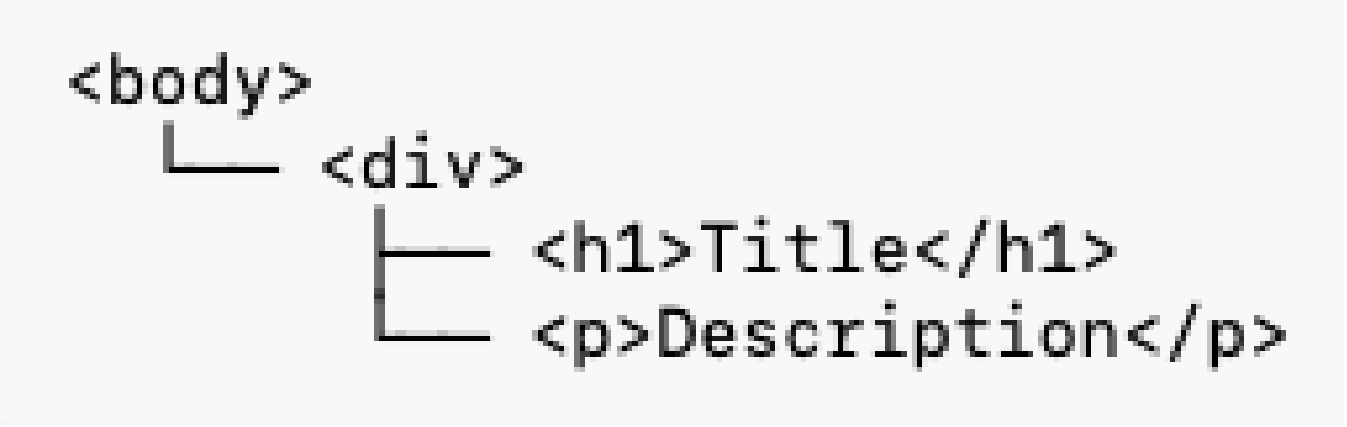
Class

Shared by multiple elements with similar formatting or roles.

Intro to HTML – Parents and Children

The browser reads HTML as a Document Object Model (DOM), which is a tree-like structure.

```
<body>
  <div>
    <h1>Title</h1>
    <p>Description</p>
  </div>
</body>
```



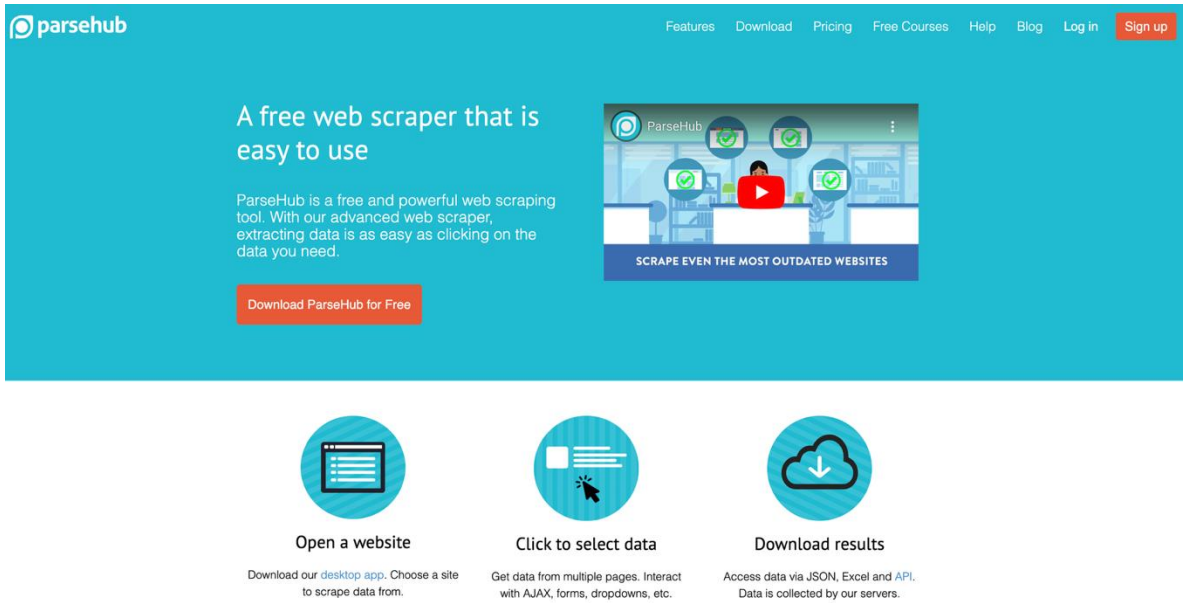
Parent-Child Relationship: <body> is the parent of , and is the parent of <h1> and <p>.

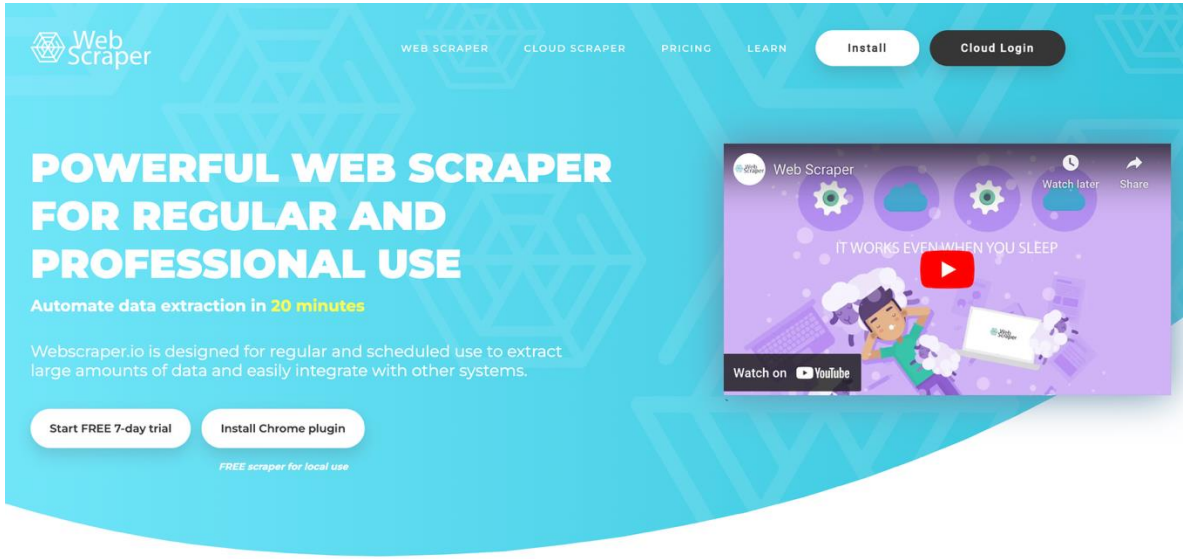
In web scraping, you navigate this tree to target specific nodes (elements).

Part 2

Skills

Commercial Tools To Help





Our goal, do it in R for analysis and visualization

Steps in Web Scraping

What data do you want and why?

How can you identify that data?

Write code to get that data

Save the data

rvest

Rselenium: more advanced

Let’s Dive In

Books to Scrape We love being scraped!

Home / All products

Books

- Travel
- Mystery
- Historical Fiction
- Sequential Art
- Classics
- Philosophy
- Romance
- Womens Fiction
- Fiction
- Childrens
- Religion
- Nonfiction
- Music

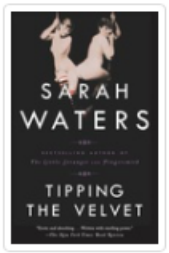
https://www.dropbox.com/scl/fi/kt8zfoj0xtxt705qfonuy/bookstoscrape.R?rlkey=mz16v4dik3d66u2yf9adziszx&st=29s2fhbq&dl=0

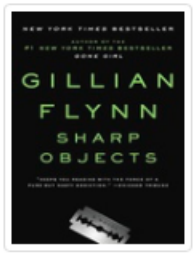
- Paranormal
- Art
- Psychology
- Autobiography
- Parenting
- Adult Fiction
- Humor
- Horror
- History
- Food and Drink
- Christian Fiction
- Business

All products

1000 results - showing **1** to **20**.

Warning! This is a demo website for web scraping purposes. Prices and ratings here were randomly assigned and have no real meaning.





Tipping the Velvet

£53.74

✔ In stock

Add to basket

Soumission

£50.10

✔ In stock

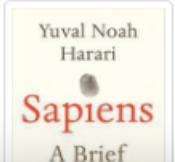
Add to basket

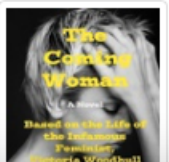
Sharp Objects

£47.82

✔ In stock

Add to basket

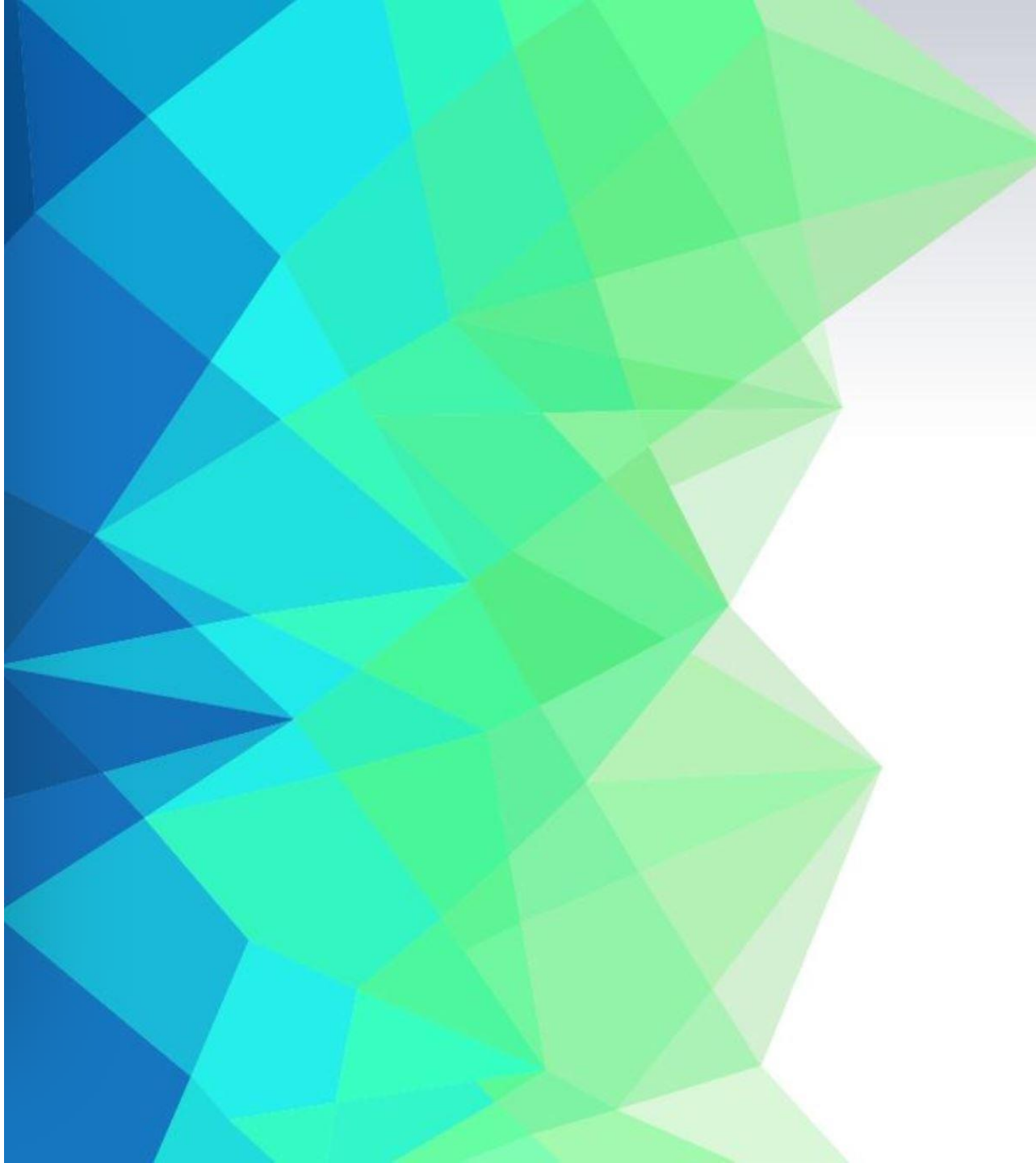




Helpful Skills

- CSS selectors
  - Select elements based on **style attributes** such as class, id, tag, and nesting.
  - Simple, clean, easy for most scraping tasks.

- Xpath
  - Select elements based on their **position and structure in the HTML tree**.
  - More powerful: can move up, down, or sideways in the DOM, and select by index, conditions, or text.



CSS Selectors: what it is

CSS selectors are **patterns used to find elements in HTML** : *„find this element on the page“*

```
<div class="farm">
  <p class="cow">Cow says Moo</p>
</div>
```

Tags: div, p, ul, li
Attributes: id, class
Children / parents / siblings

Selector Type	CSS Syntax	Meaning	Example Use Case
ID Selector	#id	Selects **one specific element** with a unique ID	#dog selects the element with id="dog"
Class Selector	.class	Selects **all elements** that share a class	.cow selects all elements with class="cow"
Descendant Selector	.parent .child	Selects elements inside another element	.farm .cow selects cows **inside** the farm
Attribute Selector	[attr="value"]	Selects elements with a specific attribute value	div[data-my-id="goat"] selects the with that attribute

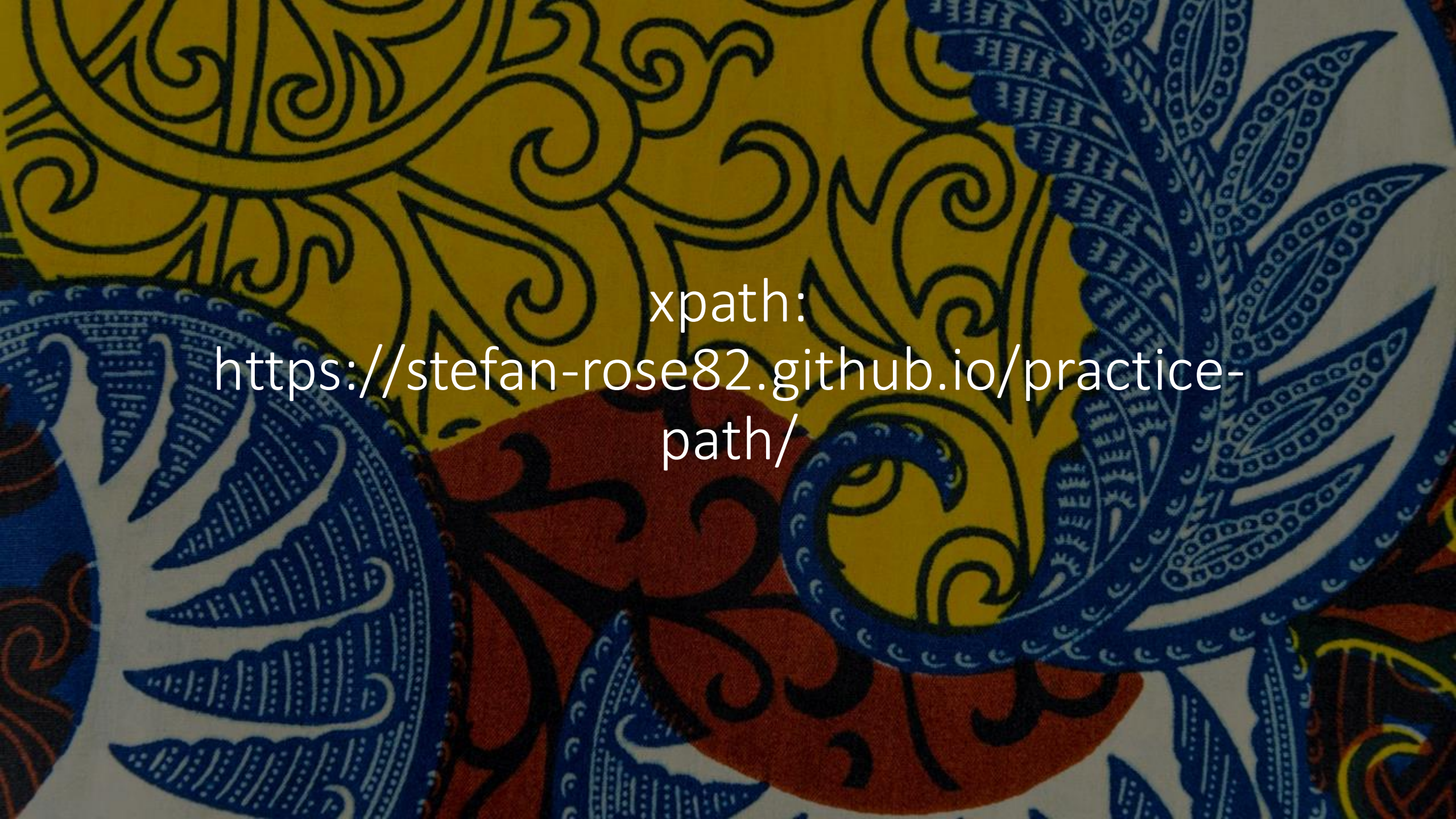
XPath: where it is

XPath selects elements by **location within the DOM tree**: *„go to this part of the tree → find these elements → then pick #1.”*

XPath Concept	XPath Pattern	Meaning	Description
Select by attribute (ID, class, custom)	//*[@attr='value']	Select element by ID, class, or any attribute	Selects elements with a specific attribute, e.g., “Find the element whose class is exactly ‘item’.”
Select by text (contains)	//tag[contains(text(),'Go')]	Match visible text inside an element	Finds elements based on what the user sees on the page (inner text match).
Parent navigation	node/.. (parent)//div/a (child)	Move up or down the HTML hierarchy	Lets you move to parent nodes, useful when starting from a known child.
Position selectors	//tag[1] (first)//tag[last()] (last)	Select elements by index/position	Selects elements by their order, e.g., first or last item in a list.

CSS:

https://stefan-rose82.github.io/practice-css-selector/index.html

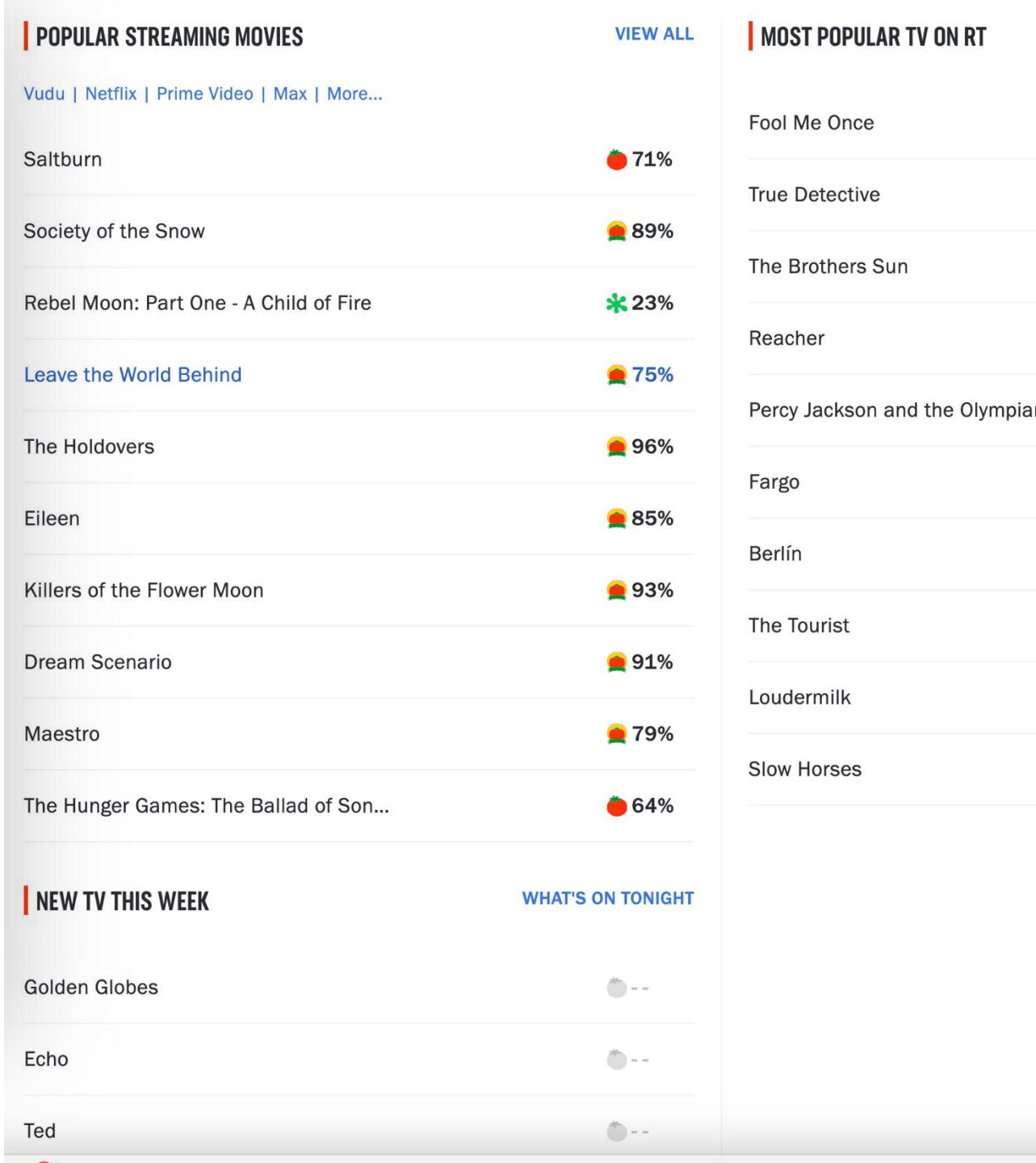

xpath:
https://stefan-rose82.github.io/practice-path/

Practice

Some more practice

With CSS Selector

- Go to this : https://www.rottentomatoes.com/
- Get the name and rating from all of the ”lists”
- A solution: https://www.dropbox.com/scl/fi/vg8iqyhsgm452n339wtfe/rottentomatoes.R?rlkey=zoxzb0war4x26vdg6artrozkm&dl=0



Some more practice

More Practice

- https://www.amazon.de/s?k=tablet
- https://www.cologne-bonn-airport.com/en/flights/departure-arrival.html
- https://www.kleinanzeigen.de/s-autos/c216

Potential Final Exam Questions

- What is web scraping?
- What are the pros and cons of web scraping compared to API calls?
- What are some examples of use cases where web scraping is commonly used?
- What legal and ethical considerations should you consider when web scraping? How does this differ across different contexts (private versus public data; personal use, research use, and commercial use; etc.)
- How does it differ for private use, research use, and commercial use?
- How are targets identified in HTML that scrapers can use?
- What are the steps in web scraping?
- Some high-level questions about R syntax to scrape data.