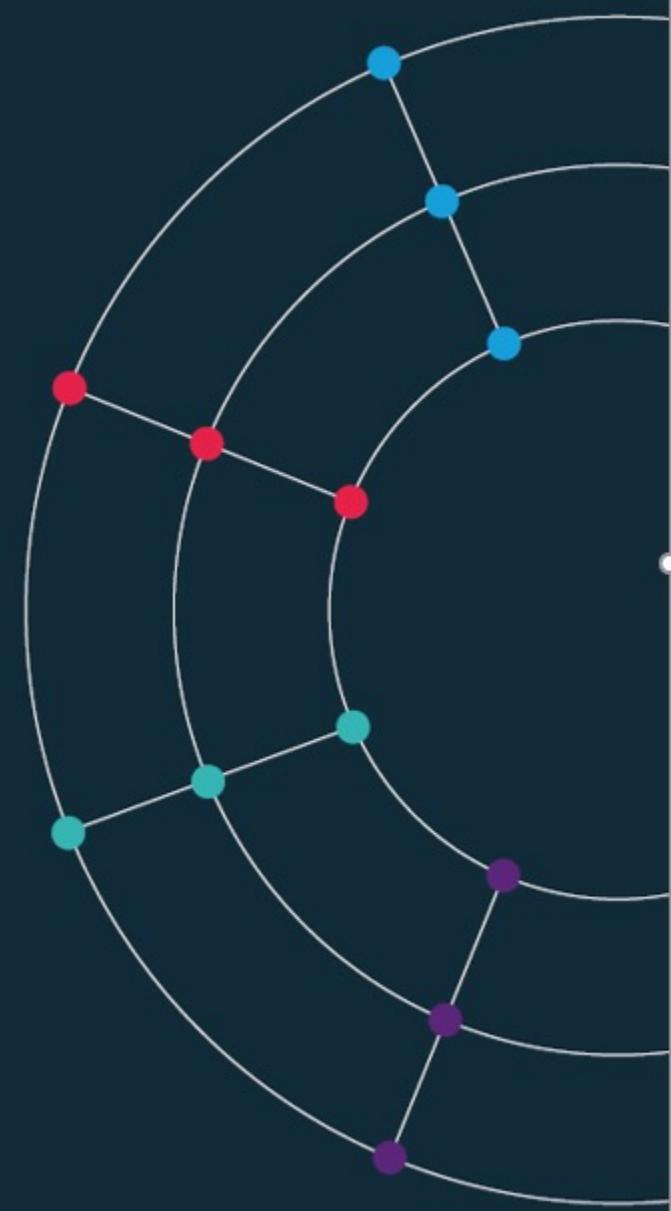


Session 5.

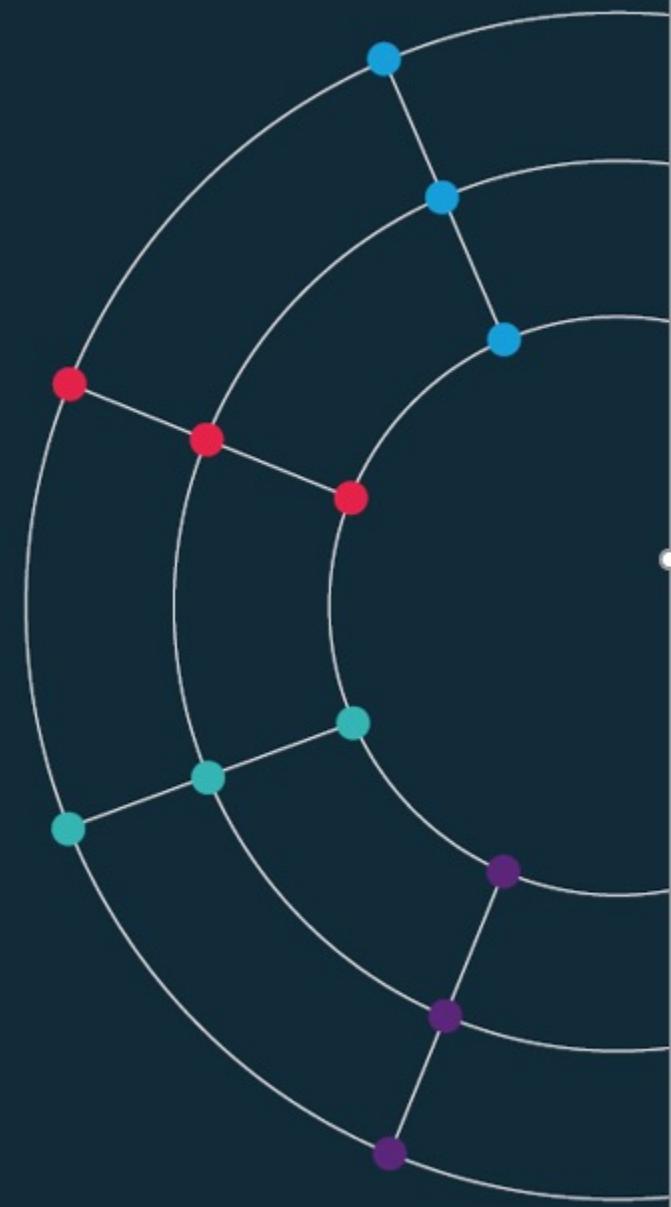
Data scraping



Session 5.

Data scraping

Scraping the HTML source



Data, so far...

Today we have used data from

	Country Name	Country Code	Year	GDP pc
1	United Kingdom	GBR	01/01/1990	17091.3051
2	Korea, Rep.	KOR	01/01/1990	8355.33277
4	United States	USA	01/01/1990	23888.6
5	United Kingdom	GBR	01/01/1991	17420.4212
6	Korea, Rep.	KOR	01/01/1991	9474.6426
7	United States	USA	01/01/1991	24342.2589
8	United Kingdom	GBR	01/01/1992	17840.551
9	Korea, Rep.	KOR	01/01/1992	10184.8557
10	United States	USA	01/01/1992	25418.9908
11	United Kingdom	GBR	01/01/1993	18673.3485
12	Korea, Rep.	KOR	01/01/1993	11030.7119
13	United States	USA	01/01/1993	26387.2937
14	United Kingdom	GBR	01/01/1994	19755.2551
15	Korea, Rep.	KOR	01/01/1994	12187.255
16	United States	USA	01/01/1994	27694.8534
17	United Kingdom	GBR	01/01/1995	20595.7082
18	Korea, Rep.	KOR	01/01/1995	13502.5827
19	United States	USA	01/01/1995	28690.8757
20	United Kingdom	GBR	01/01/1996	21946.1081
21	Korea, Rep.	KOR	01/01/1996	14694.0962
22	United States	USA	01/01/1996	29947.7127

Structured Files
(e.g. Excel, CSV, JSON)

The image shows two screenshots of data retrieval platforms. On the left, the 'Developer Hub' of the Office for National Statistics (ONS) is displayed, featuring an 'Introduction' section with links to rate limiting, the API tour, and Census 2021 observations. It also includes a 'Data Hub.' section with a search bar and social media links. On the right, the 'FRED API' website is shown, which is described as a web service for developers. It features a 'Data Explorer' section where users can search for data by category and series, and a detailed 'API Documentation' section with various endpoints and examples.

APIs
(e.g. ONS, ECO, FRED)

But what do we do when the data we want isn't available?

Data, so far...

What if we want

Overview of G7 summits					
#	Date	Host	Host leader	Location held	Notes
1st	November 1975	France	Valéry Giscard d'Estaing	Château de Rambouillet, Yvelines	The first and last G6 summit.
2nd	27–28 June 1976	United States	Gerald R. Ford	Dorado, Puerto Rico ^[74]	Also called "Rambouillet II"; Canada joined the group, forming the G7 ^[74]
3rd	7–8 May 1977	United Kingdom	James Callaghan	London, England	The President of the European Commission was invited to join the annual G7 summits.
4th	16–17 July 1978	West Germany	Helmut Schmidt	Bonn, North Rhine-Westphalia	
5th	28–29 June 1979	Japan	Masayoshi Ohira	Tokyo	
6th	22–23 June 1980	Italy	Francesco Cossiga	Venice, Veneto	Prime Minister Ohira died in office on 12 June; Foreign Minister Saburo Okita led the delegation that represented Japan.
7th	20–21 July 1981	Canada	Pierre E. Trudeau	Montebello, Quebec	
8th	4–6 June 1982	France	François Mitterrand	Versailles, Yvelines	
9th	28–30 May 1983	United States	Ronald Reagan	Williamsburg, Virginia	
10th	7–9 June 1984	United Kingdom	Margaret Thatcher	London, England	
11th	2–4 May 1985	Germany	Helmut Kohl	Bonn, North Rhine-Westphalia	
12th	4–6 May 1986	Japan	Yasuhiro Nakasone	Tokyo	
13th	8–10 June 1987	Italy	Ambrolio Fanfani	Venice, Veneto	
14th	19–21 June 1988	Canada	Brian Mulroney	Toronto, Ontario	
15th	14–16 July 1989	France	François Mitterrand	Paris, Paris	FATF was formed
16th	9–11 July 1990	United States	George H. W. Bush	Houston, Texas	
17th	15–17 July 1991	United Kingdom	John Major	London, England	
18th	6–8 July 1992	Germany	Helmut Kohl	Munich, Bavaria	The first G7 summit in reunified Germany.
19th	7–9 July 1993	Japan	Kichi Miyazawa	Tokyo	
20th	8–10 July 1994	Italy	Silvio Berlusconi	Naples, Campania	
21st	15–17 June 1995	Canada	Jean Chrétien	Halifax, Nova Scotia	
22nd	27–29 June 1996	France	Jacques Chirac	Lyon, Rhône	The first summit to debut International organizations, namely the International Monetary Fund, International

Data from Wikipedia

The Economics Observatory is a platform for questions and answers about the economy. It features a Data Hub section with various infographics and articles. One article discusses labour market power in the UK, another explores plastics in marine ecosystems, and a third looks at youth custody issues.

News and Media

The Tesco website displays a grid of various cookie products available for purchase. The categories include Biscuits & Cereal Bars, Shortbread, Cereal Bars & On the Go Snack Bars, Continental Biscuits, Chocolate Biscuits & Jaffa Cakes, and Chocolate Biscuit Bars & Mini Biscuits. Each product listing shows the name, price, and a 'Buy' button.

Prices from Supermarkets

Scraping.

- The **automated** extraction of data from websites
- Scraping the HTML source. **Easy. Automated.**
- Scraping static HTML pages.
 - A bit more **difficult**. **Can be automated.**
 - CAPTCHA. Impersonating a **human** user. **Zombie browser.**
- Scraping HTML pages generated on-the-fly with JavaScript.
 - **Hard.** Only **zombie browser** works, and only in some cases.

Parsing HTML.

We started today by writing HTML

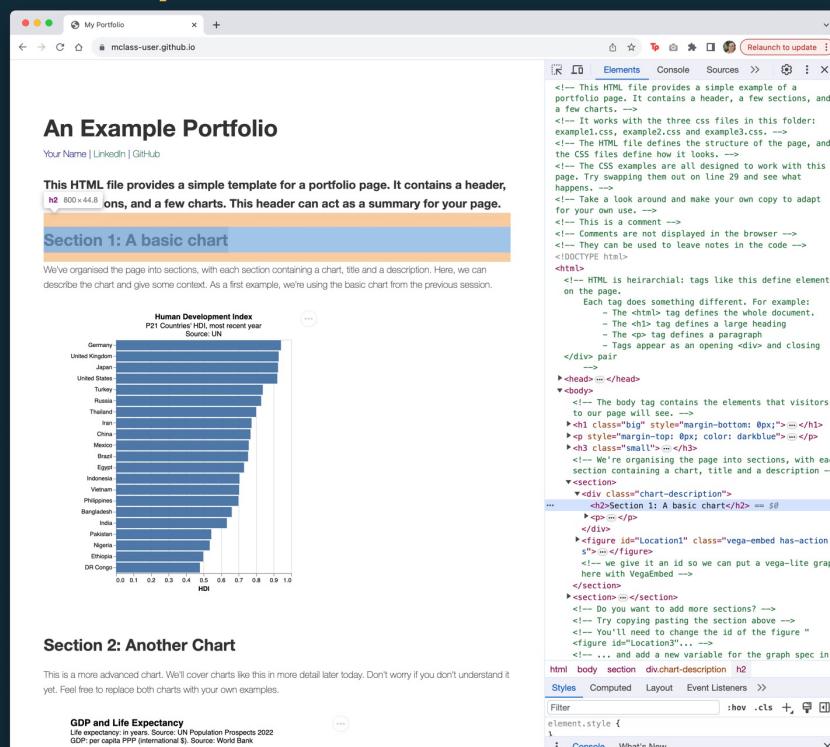
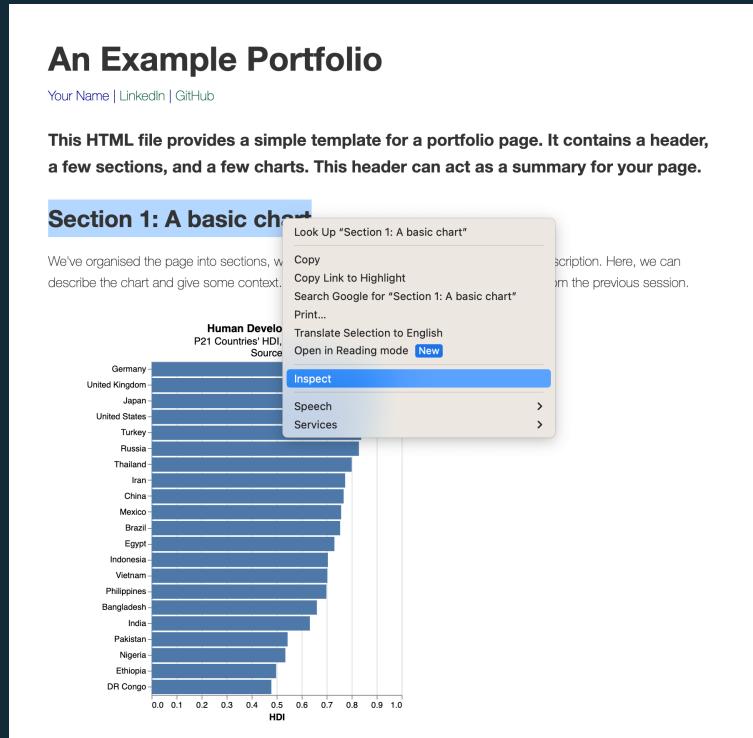
```
1 <html>
2 <head>
3   <title>My Portfolio</title>
4 </head>
5 <body>
6   <h1> My Portfolio</h1>
7   <p> A collection of masterclass-work </p>
8 </body>
9 </html>
10
```

```
<a href="https://www.economicsobservatory.com/
could-a-new-policy-institution-help-solve-the-uks-productivity-probl
em">
  <div>
    <h3 class="home__blocks-item-title">Could a new policy
      institution help solve the UK's productivity problem?
    </h3>
    <div class="home__blocks-item-teaser display">
      <p>Comparatively sluggish productivity growth is one of
        the UK's biggest policy challenges. Past strategies
        to solve the problem have lacked commitment and
        proper evaluation. One solution could be to
        establish a
        growth and productivity institution to coordinate
        policies over the long term.</p>
    </div>
  </div>
</a>
```

We'll end it by reading HTML

Using Inspect-Element.

Our most important tool is ‘inspect-element’



In Chrome, Right-click or Ctrl. Click and select ‘Inspect’

Parsing HTML.

- (Almost) all the data displayed on websites is found in the HTML
- We can **extract** data by searching the HTML
- **Everything is defined in the HTML, we just have to find it**

Overview.

1. Determine how data is defined in HTML (**inspect element**)
2. Parse the HTML (**Beautiful Soup**)
3. Clean the data (**Pandas**)
4. Visualise (**Vega-lite**)

An Example.

For example, we can scrape ECO headlines and tag-lines by ‘parsing’ the HTML

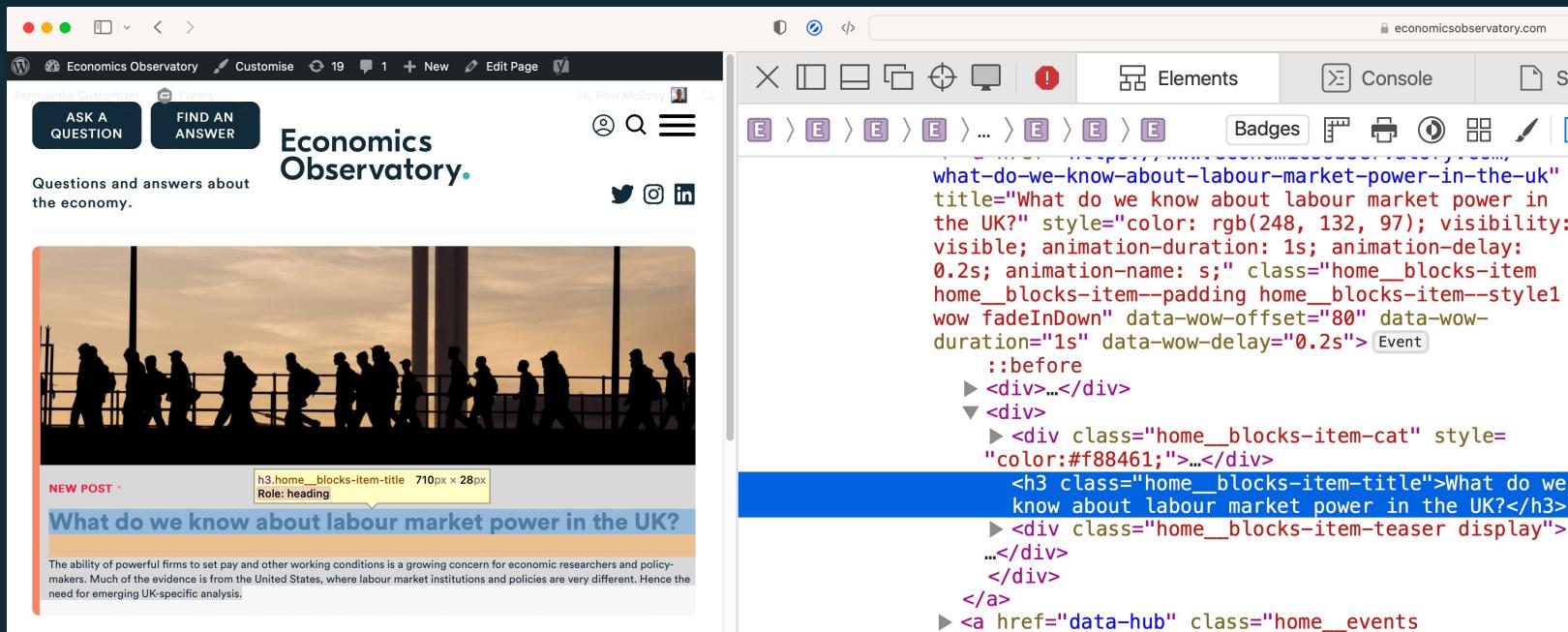
```
1386 </div>
1387 </div>
1388 </div>
1389 </div>
1390 </div>
1391 </div>
1392 </div>
1393 </div>
1394 <main id="content" class="home_main">
1395 <section class="home_blocks">
1396 <div class="container">
1397 <div class="home_blocks-grid-wrap">
1398 <div class="home_blocks-grid">
1399 <a href="https://www.economicsobservatory.com/what-do-we-know-about-labour-market-power-in-the-uk"
1400 title="What do we know about labour market power in the UK?" style="color: #F88461;">
1401 <div class="home_blocks-item home_blocks-item--style-wfow fadInDown"
1402 data-wow-offset="0px" data-wow-duration="1s" data-wow-delay="0.2s">
1403 <div>
1404 <div class="home_blocks-item-image" style="color: #F88461;">
1405 <img alt="Thumbnail image for the article 'What do we know about labour market power in the UK?'">
1406 <div class="image" style="border-color: #F88461; background-image: url('https://www.economicsobservatory.com/wp-content/uploads/2023/02/1Stock-'
1407 <div>
1408 <div class="home_blocks-item-cat" style="color: #F88461;">
1409 <span>New Post</span>
1410 </div>
1411 <div>
1412 <h3 class="home_blocks-item-title">What do we know about labour market power in the UK?
1413 <h3>
1414 <div class="home_blocks-item-teaser display">
1415 <p>The ability of powerful firms to set pay and other working conditions is a growing concern for economic researchers and policy-makers. Much of the evidence is from the United States, where labour market institutions and policies are very different. Hence the need for emerging UK-specific analysis.</p>
1416 </div>
1417 </div>
1418 </div>
1419 <a href="data-hub" class="home_events home_blocks-item--styleB">
1420 <div id="circles3"></div>
1421 <script>
1422 var animationData = { "v": "5:10,0.", "fr": 30, "ip": 0, "op": 363, "u": 795, "h": 275, "nm": "Addation_23", "ddd": 0, "assets": [
1423 "https://www.economicsobservatory.com/wp-content/themes/economics/assets/images/icon-addition.svg", "https://www.economicsobservatory.com/wp-content/themes/economics/assets/images/icon-circles3.svg" ];
1424 var params = {
1425 "fr": 30, "ip": 0, "op": 363, "u": 795, "h": 275, "nm": "Addation_23", "ddd": 0, "assets": [
1426 "https://www.economicsobservatory.com/wp-content/themes/economics/assets/images/icon-addition.svg", "https://www.economicsobservatory.com/wp-content/themes/economics/assets/images/icon-circles3.svg" ],
1427 "render": "svg", "loop": true };
1428 </script>
```



The screenshot shows the homepage of the Economics Observatory. At the top, there are two dark blue buttons: "ASK A QUESTION" and "FIND AN ANSWER". The URL "economicsobservatory.com" is in the address bar. The main title "Economics Observatory." is centered above a large image of people walking on a bridge at sunset. Below the image, a "NEW POST" section features the title "What do we know about labour market power in the UK?". A descriptive paragraph follows, mentioning the ability of powerful firms to set pay and other working conditions as a growing concern. To the right, there's a "Data Hub" section with "EXPLORE", "CREATE", and "SHARE" buttons, and a "SCHOOLS, UNIVERSITIES & TRAINING" section with a question about reducing gender gaps in mathematics education. Another section titled "LESSONS FROM HISTORY" shows a building and a question about minority treatment in the UK judicial system.

Using Inspect-Element.

We determine how the target data is defined using ‘inspect-element’



We see titles have a class “home_blocks-item-title”

Parse the HTML

We'll use a Python module, **BeautifulSoup**, to interpret the HTML.

For example, we can look for every title by searching for:

class "home_blocks-item-title"

with

```
soup.find_all(class_="home_blocks-item-title")
```

Code-along.

A more advanced scraper

In this bonus practical session, we will use [Google Colab](#) to use scrape data from the Economics Observatory website using Python. Again, we can also embed a chart displaying the scraped data into your website, using [VS Code](#) and [GitHub](#).

- Work through the following guided notebook: “[s5_Scraping.ipynb](#)” (open in Google Colab)

Learn more.

- In this session we have tried basic data scraping with `BeautifulSoup`
- There's still much more to learn
 - Choose your own projects (how can you make your job easier?)
 - Try bigger projects (scrape 100 pages, not just 1)
 - Try advanced tools (e.g. `Selenium`)

Learn more, responsibly.

- **Rate-limiting:** avoid making too many requests at once
- **Ethics:** Ensure your scraping activities do not harm the website's operation
- **Data Privacy:** Be mindful of personal data collection. Comply with relevant data protection laws (like GDPR).

