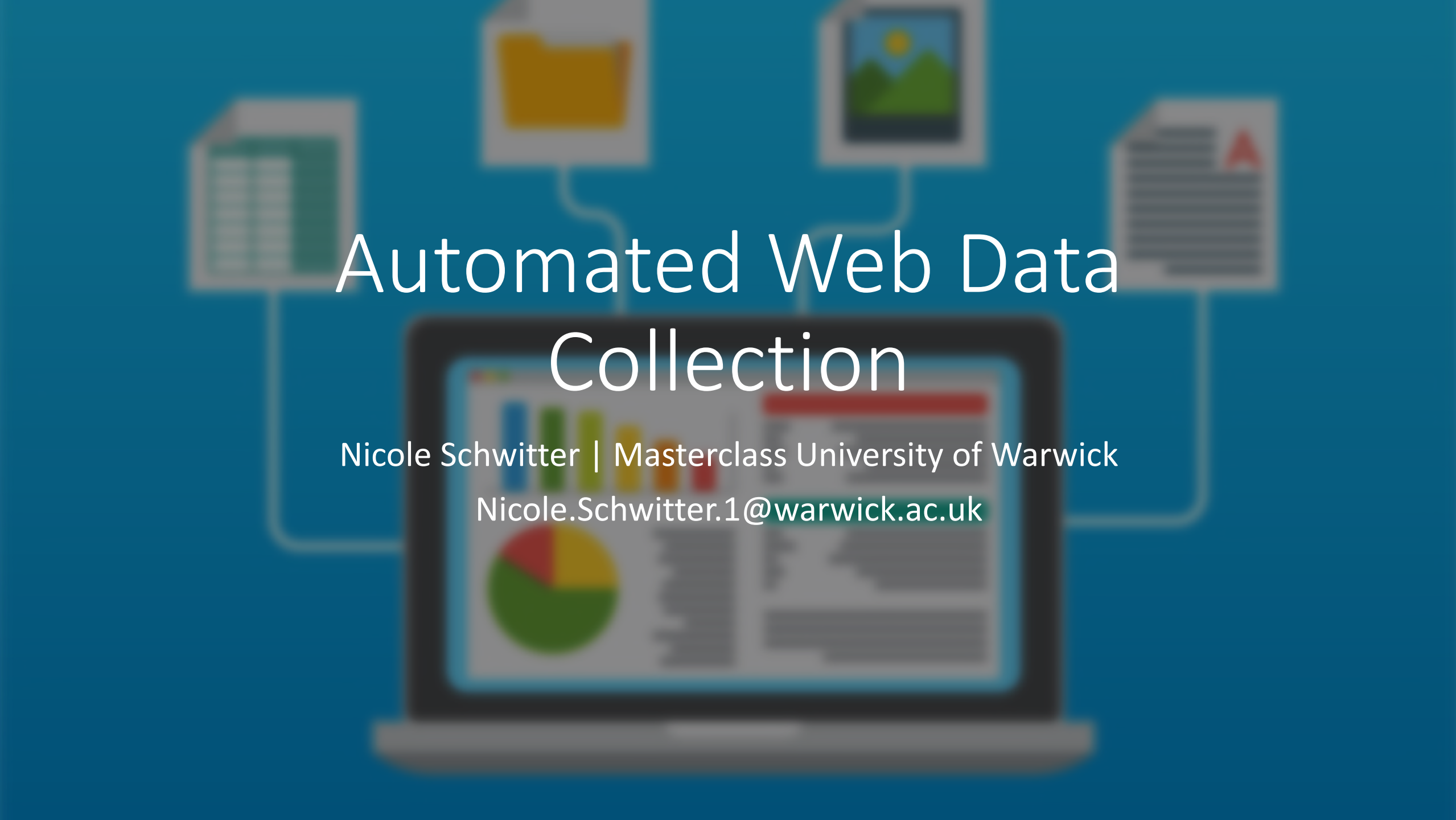


Automated Web Data Collection

Nicole Schwitter | Masterclass University of Warwick

Nicole.Schwitter.1@warwick.ac.uk

Introductions

- Me
  - PhD student in the Department of Sociology
  - Seven years experience with web data collection
  - First time teaching this masterclass

- You
  - Expectations of the course?
    - www.menti.com code: 9820 6964 (or directly: https://www.menti.com/kephommnen)
  - Experience with R? Experience with web data collection?

Agenda Today

9:30	Session 1 Introduction | History, examples, the internet
10:30	Break
10:45	Session 2 Web scraping | Approach and exercises
12:30	Lunch
13:30	Continue Session 2 // Session 3 APIs | Approach and exercises
15:15	Break
15:30	Session 4 The buts | Ethical considerations and terms of services
16:00	Finish

Course material

- Slides and worksheets are available on GitHub: https://github.com/nschwitter/webdata-warwick
- Worksheets
  - We use R and Rstudio.
  - Examples mostly based on Wikipedia: Free resource, few ethical considerations, no access key necessary.

- Part of this material is building upon materials from Theresa Gessler, Christopher Barrie, and Chris Bail.

We will be using R today. Not your preferred language?

- I use Python.
  - Cool! Have a look at the Python libraries *BeautifulSoup* and *Selenium.*

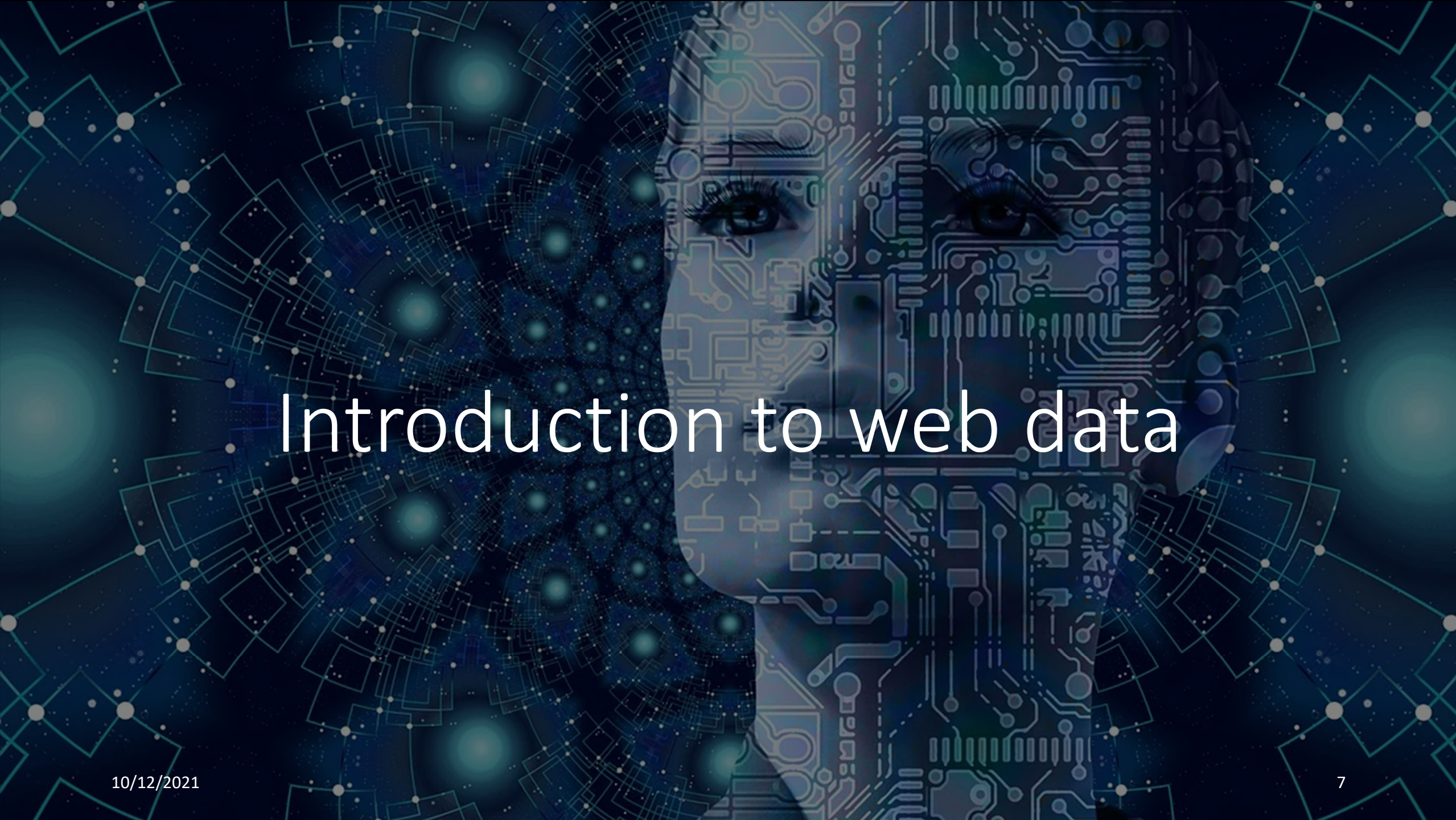
- I use Stata.
  - Oh no. Have a look at R or Python.

- I use any other programming language.
  - There are ways to web scrape, but using a scripting language might just be easier.

(N-)Etiquette and using Zoom today

- Lecture : practice = 1 : 2

- Follow along in RStudio if possible.
- Share screens, interact, help each other, and if stuck, ask!
- Raise hand or just interrupt me.



Introduction to web data

Why are you interested in web data?

You’re not the first wanting web data

Re: searchable index of the web

mkgray@athena.mit.edu

- **Mail folder:** WWW Talk Apr-Jun 1993 Archives
- **Next message:** Ray Stell: "Re: searchable index of the web"
- **Previous message:** Thomas R. Bruce: "Cello Beta 0.6 released"
- **In-reply-to:** joe@athena.mit.edu: "Re: searchable index of the web"
- **Reply:** Ray Stell: "Re: searchable index of the web"

```
From: mkgray@athena.mit.edu
Message-id: <9306301905.AA25385@uranus.MIT.EDU>
To: joe@athena.mit.edu
Cc: www-talk@nxoc01.cern.ch
Subject: Re: searchable index of the web
In-reply-to: Your message of Wed, 30 Jun 93 01:39:29 -0400.
             <9306300539.AA20609@theodore-sturgeon>
Date: Wed, 30 Jun 93 15:05:45 EDT

I have written a perl script that wanders the WWW collecting URLs, keeping
tracking of where it's been and new hosts that it finds.  Eventually,
after hacking up the code to return some slightly more useful information
(currently it just returns URLs), I will produce a searchabe index of this.
There is a complete list of all the sites it has found at

<a href="http://www.mit.edu:8001/afs/sipb/user/mkgray/ht/comprehensive.html">
A complete list of sites found by the W4 (World Wide Web Wanderer)
</a>

I'll announce here when we get this index properly running, however it probably
won't be until sometime in August, as I am going on vacation.  Until then...

                                    Matthew Gray
                                    mkgray@athena.mit.edu

Visit the SIPB WWW Plexus Server.  URL: http://www.mit.edu:8001/
```

Wanderer Results

The web has grown very fast. In fact, the web has grown substantially faster

The rate of the web's growth has been and continues to be exponential, but is
is still under 6 months.

Results Summary			
Month	**# of Web sites**	**% .com sites**	**Hosts* per Web server**
6/93	130	1.5	13,000 (3,846)
12/93	623	4.6	3,475 (963)
6/94	2,738	13.5	1,095 (255)
12/94	10,022	18.3	451 (99)
6/95	23,500	31.3	270 (46)
1/96	100,000	50.0	94 (17)
6/96	230,000 (est)	68.0	41
1/97	650,000 (est)	62.6	NA

* Host in the final column is defined as a listed hostname. The number in par

World Wide Web Wanderer, Matthew Gray, 1993. https://www.mit.edu/~mkgray/net/web-growth-summary.html

… Google, is it you?

Q: Setting User-Agent Field?

Subscribe

125239 views

Lawrence Page Jan 7, 1996, 9:00:00 AM

to

I have a web robot which is a Java app. I need to be able to set the User-Agent field in the HTTP header in order to be a good net citizen (so people know who is accessing their server). Anyone have any ideas?

Right now, Java sends a request that includes something like:

User-Agent: Java/1.0beta2

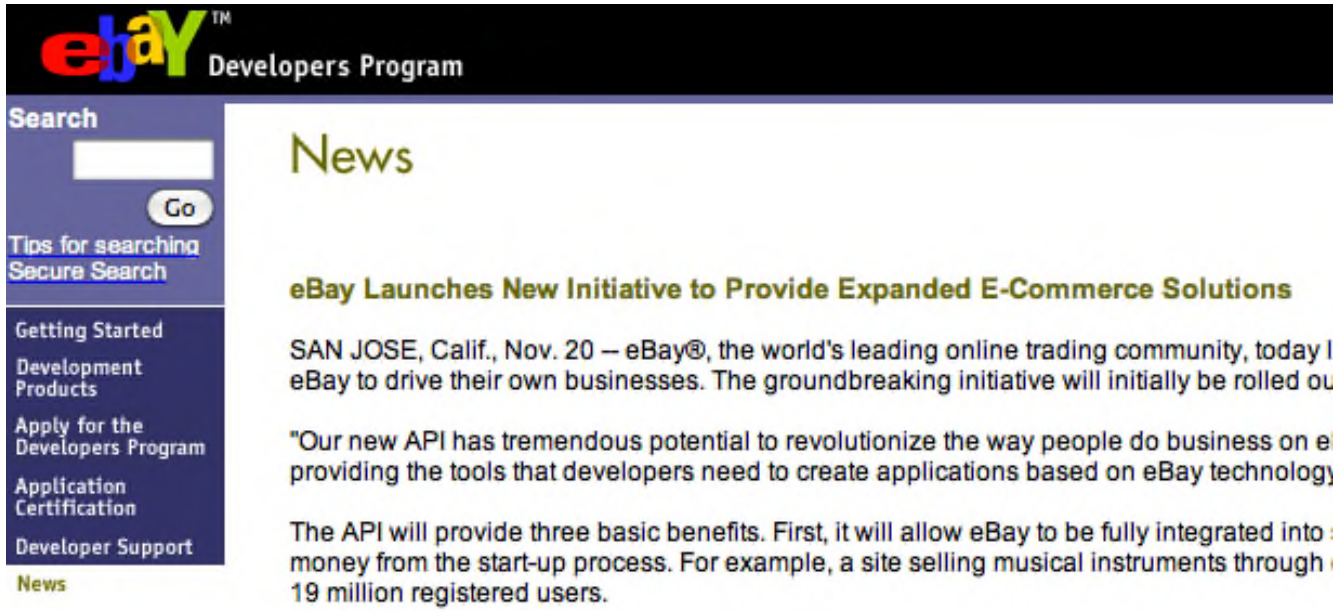
I'd rather not rewrite all the HTTP stuff myself. I tried just searching in the JDK for the Java/1.0beta2 figuring I could just change the string, but I couldn't find it. Perhaps it is stored as a unicode string?

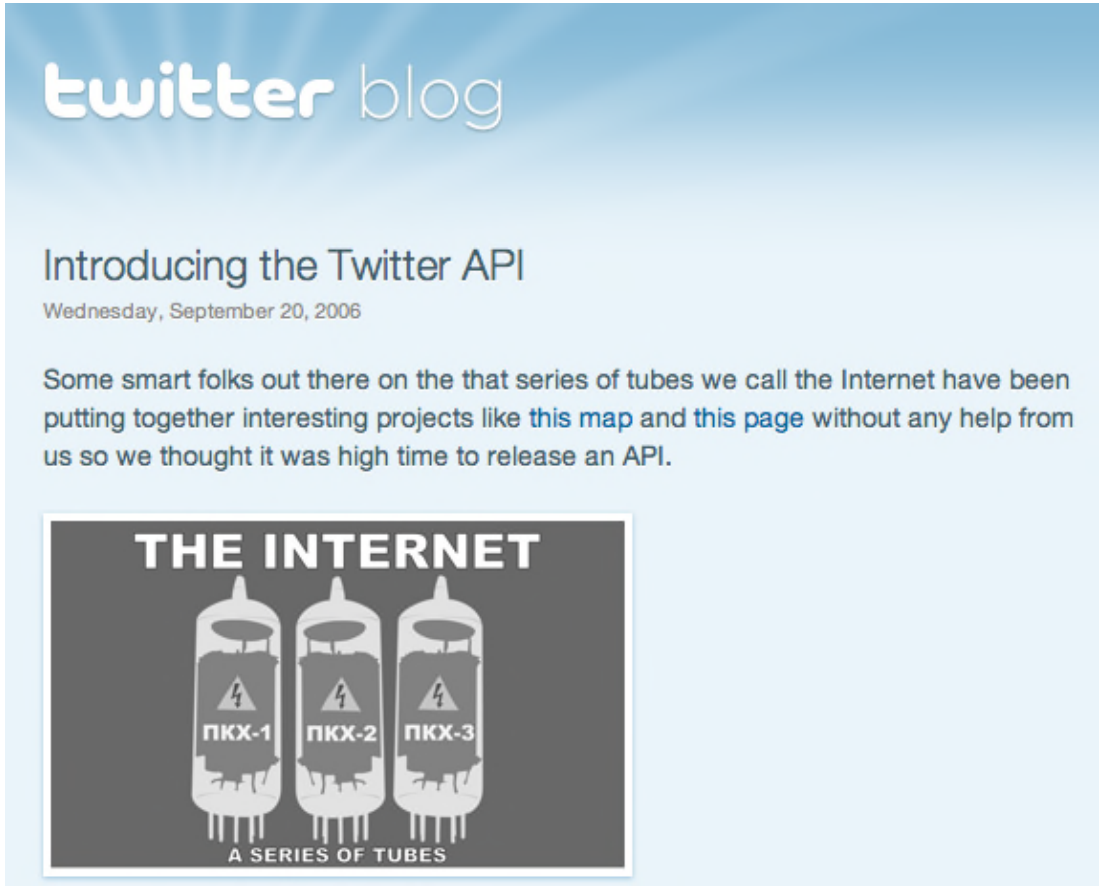
An easy method of setting the User-Agent field should probably be added to Java, so people can properly identify their programs.

Thanks, Larry Page

https://groups.google.com/g/comp.lang.java/c/aSPAJO05LIU/m/ushhUIQQ-ogJ?pli=1

Making web data accessible: Early APIs





https://apievangelist.com/2012/12/20/history-of-apis/

Early uses of web data

Karola Brunner, Financial and Business Mathematician

Automated price collection via the internet

In recent years, e-commerce has become increasingly important. Based on the outlet type weighting[1] of the consumer price index, e-commerce and mail order business accounted for 5.1 % of the entire basket of goods and services for the base year 2010. For some categories of goods, this share is considerably higher.[2] Prices for the relevant categories of goods are therefore increasingly collected online. Moreover, all major retailers now have their own online shops where they often offer products at the same price as in their local stores. The proportion of products for which price data may be collected online is estimated to be much higher than the figure derived purely from the outlet type weighting.

https://www.destatis.de/EN/Methods/WISTAScientificJournal/Downloads/automated-price-collection-brunner-042014.pdf?__blob=publicationFile

Current Examples: Web Data in Social Science Research

Article

The International Journal of Press/Politics
1–34

Right-Wing YouTube:

rossMark
click for updates

Archive About Submit ma

SCIENCE ADVANCES | RESEARCH ARTICLE

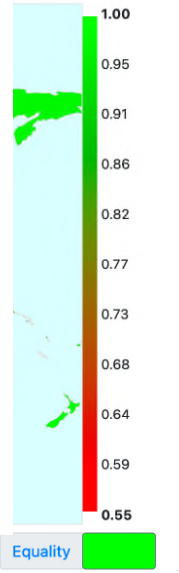
CORONAVIRUS

Elusive consensus: Polarization in elite communication on the COVID-19 pandemic

Jon Green[1], Jared Edgerton[1], Daniel Naftel[1], Kelsey Shoub[2], Skyler J. Cranmer[1]*

Cues sent by political elites are known to influence public attitudes and behavior. Polarization in elite rhetoric may hinder effective responses to public health crises, when accurate information and rapid behavioral change can save lives. We examine polarization in cues sent to the public by current members of the U.S. House and Senate during the onset of the COVID-19 pandemic, measuring polarization as the ability to correctly classify the partisanship of tweets' authors based solely on the text and the dates they were sent. We find that Democrats discussed the crisis more frequently–emphasizing threats to public health and American workers–while Republicans placed greater emphasis on China and businesses. Polarization in elite discussion of the COVID-19 pandemic peaked in mid-February—weeks after the first confirmed case in the United States—and continued into March. These divergent cues correspond with a partisan divide in the public's early reaction to the crisis.

Share Export

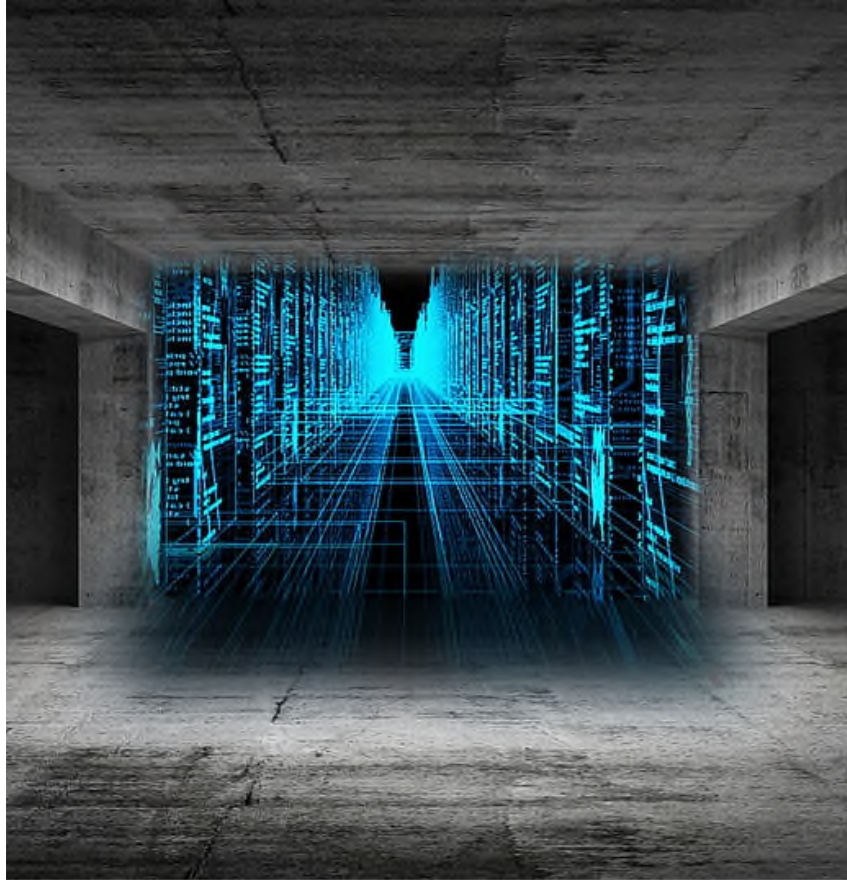


Masoomali Fatehkia

than Heard in Online New
e0148434. doi:10.1371/jo

Keywords
YouTube, radicalization, conservatism, political extremism

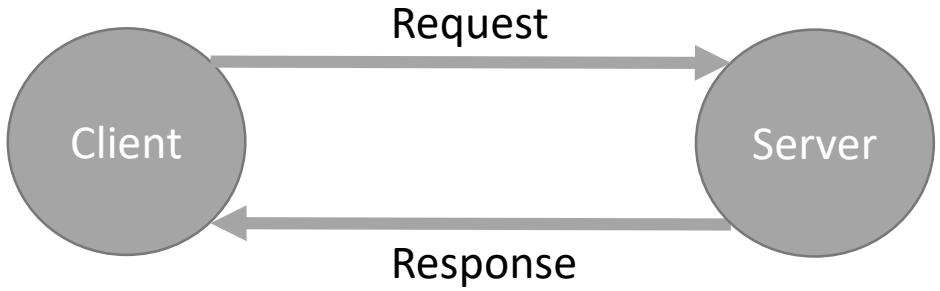
How do we get the data?

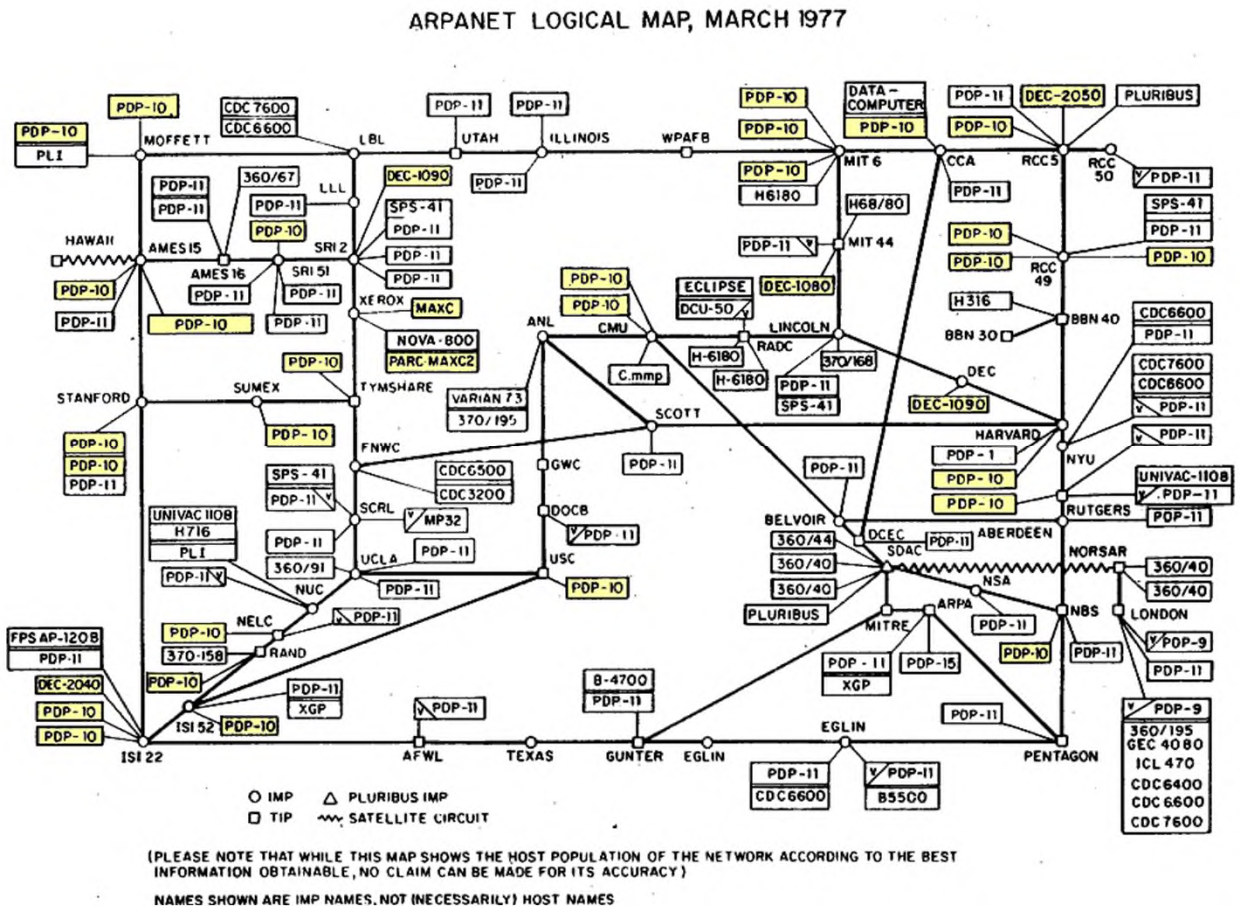


First… how does the world wide web even work?

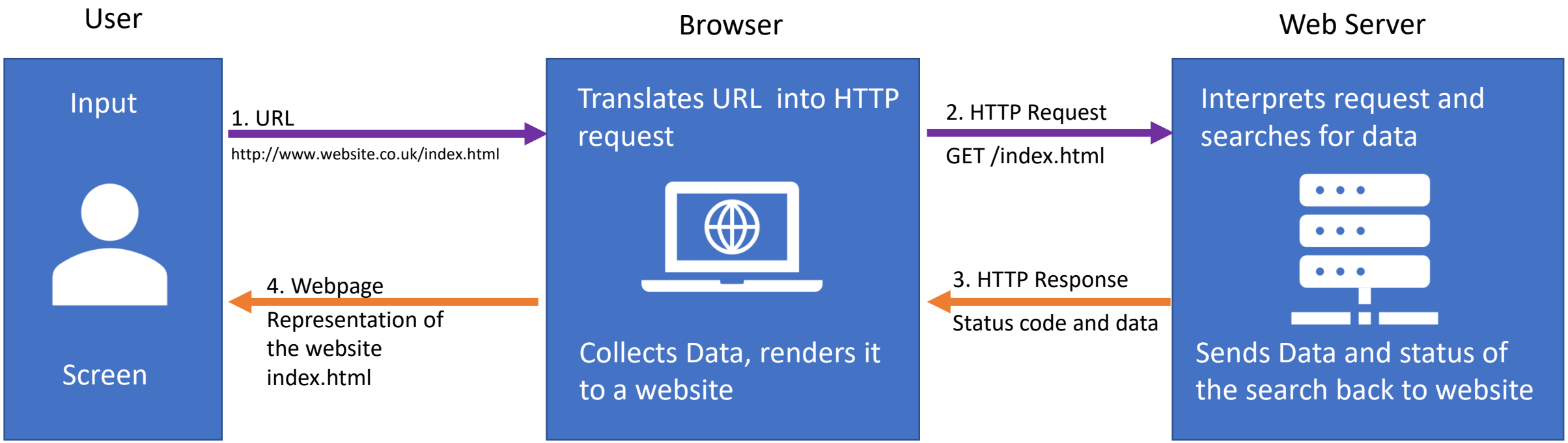
The internet and the web

- The Internet: technical infrastructure, allowing computers to connect
- The web: Application built on the infrastructure



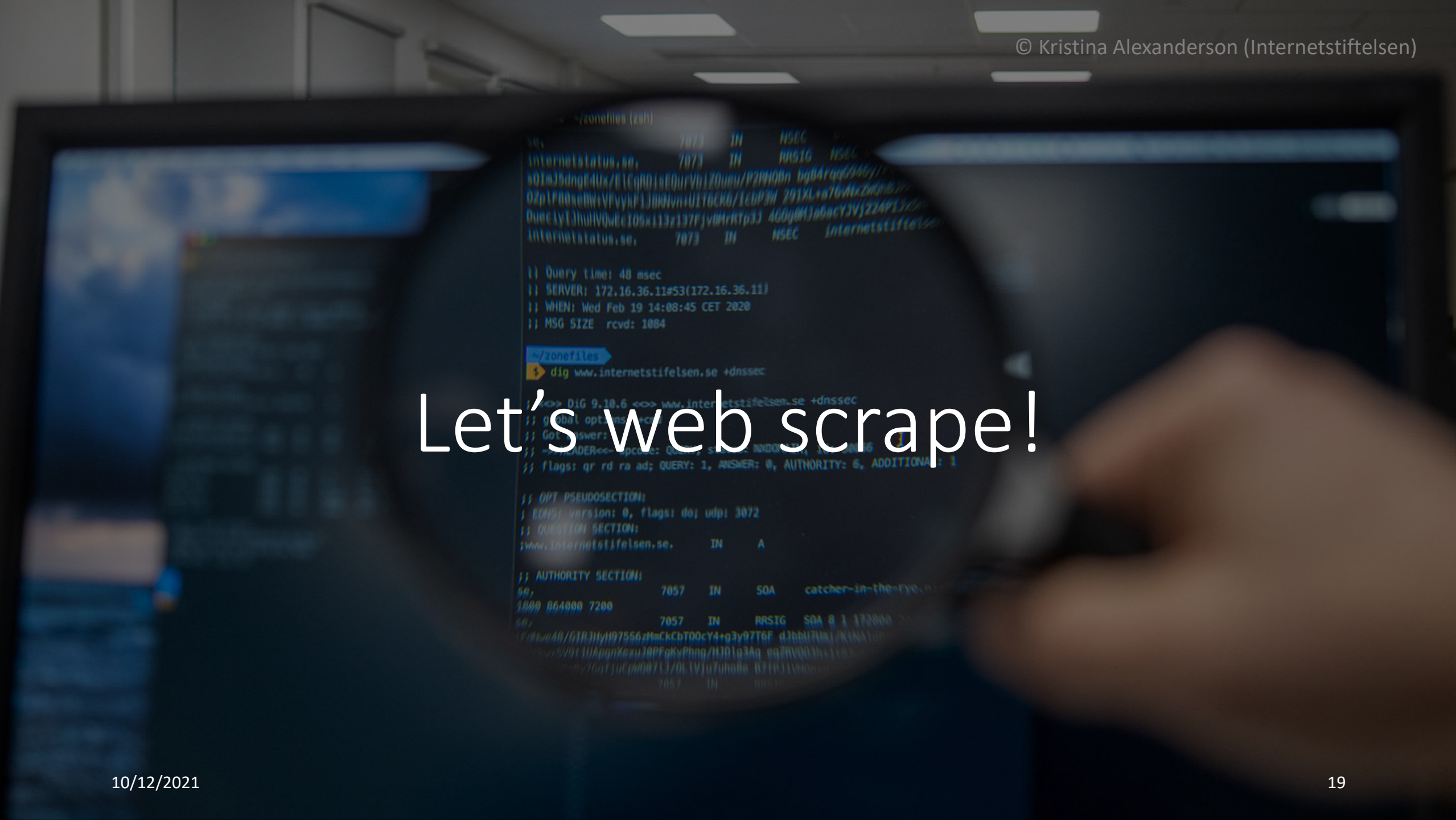


Communication process (HTTP)



Getting the data

- Ctrl + c, Ctrl + v from displayed website
  - Tedious, error-prone, slow
  - Unstructured data: Sometimes, it might be your only option!
- Screen scraping
  - Automated collection of content hosted on webpage
    - Selecting contents of the webpage as accessed by the URL
    - We retrieve the HTML instead of displaying it in a browser.
  - Early origins: “Web crawling”
- APIs
  - Sending your own data requests to the server
  - Structured data



Web scraping

- We retrieve the HTML instead of displaying it in a browser.
- So first: What is HTML?

HTML

- Webpages are written in HyperText Markup Language
  - Assisting technologies: Cascading Stylesheets (CSS), Javascript (JS)
  - Describes structure of web page.
  - Includes cue for the appearance of the document.

- We can see the HTML of any website.
  - How? “View page source” (Right-click)
    - Firefox: Tools → Browser Tools → Page Source
    - Safari: Menu → Preferences → Advanced → Check developer menu → → Develop → Show page source
    - Chrome: Menu → More tools → View Source
    - Specific elements: “Inspect”
  - Often messy

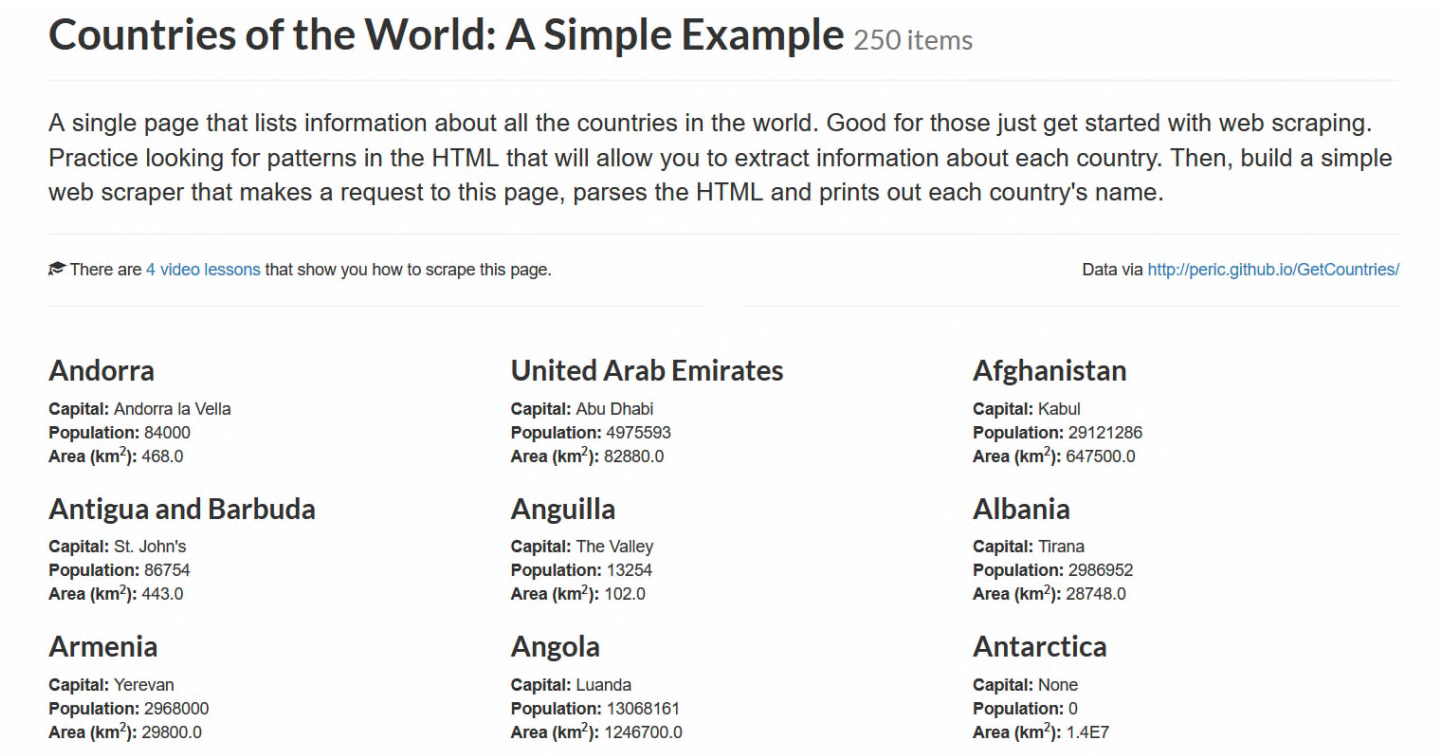
- Do you need to know HTML to web scrape?
  - Only rudimentarily, but it helps.

Let’s start with R: Worksheets

- 1_intro.Rmd
  - Intro to RMarkdown
  - Intro to HTML
  - CSS Selectors
- 2_webscrapingbasic.Rmd
  - Parsing HTML
  - Selecting elements
  - Scraping tables

All countries of the world

- I am researching all countries of the world, now I got a great list!



Countries of the World: A Simple Example 250 items

A single page that lists information about all the countries in the world. Good for those just get started with web scraping. Practice looking for patterns in the HTML that will allow you to extract information about each country. Then, build a simple web scraper that makes a request to this page, parses the HTML and prints out each country's name.

There are 4 video lessons that show you how to scrape this page. Data via http://peric.github.io/GetCountries/

Andorra
Capital: Andorra la Vella
Population: 84000
Area (km^2): 468.0

Antigua and Barbuda
Capital: St. John's
Population: 86754
Area (km^2): 443.0

Armenia
Capital: Yerevan
Population: 2968000
Area (km^2): 29800.0

United Arab Emirates
Capital: Abu Dhabi
Population: 4975593
Area (km^2): 82880.0

Anguilla
Capital: The Valley
Population: 13254
Area (km^2): 102.0

Angola
Capital: Luanda
Population: 13068161
Area (km^2): 1246700.0

Afghanistan
Capital: Kabul
Population: 29121286
Area (km^2): 647500.0

Albania
Capital: Tirana
Population: 2986952
Area (km^2): 28748.0

Antarctica
Capital: None
Population: 0
Area (km^2): 1.4E7

A cautionary tale on using web-scraped data

- Why is this information online?
- When was it published?
- Who published it?
- Where is this information coming from?
- Are there biases?

Let’s use a second source!

List of countries by population (United Nations)

From Wikipedia, the free encyclopedia

This is a list of countries and other inhabited territories of the world by total population, based on estimates published by the United Nations in the 2019 revision of *World Population Prospects*.[2][3] These figures refer to the de facto population in a country or area as shown in the "estimates" section.

Contents [hide]

Statistical subregions as defined by the United Nations Statistics Division.[1]

	Country/Area	UN continental region[4]	UN statistical subregion[4]	Population (1 July 2018)	Population (1 July 2019)	Change
1	China[a]	Asia	Eastern Asia	1,427,647,786	1,433,783,686	+0.43%
2	India	Asia	Southern Asia	1,352,642,280	1,366,417,754	+1.02%
3	United States	Americas	Northern America	327,096,265	329,064,917	+0.60%
4	Indonesia	Asia	South-eastern Asia	267,670,543	270,625,568	+1.10%
5	Pakistan	Asia	Southern Asia	212,228,286	216,565,318	+2.04%
6	Brazil	Americas	South America	209,469,323	211,049,527	+0.75%
7	Nigeria	Africa	Western Africa	195,874,683	200,963,599	+2.60%
8	Bangladesh	Asia	Southern Asia	161,376,708	163,046,161	+1.03%

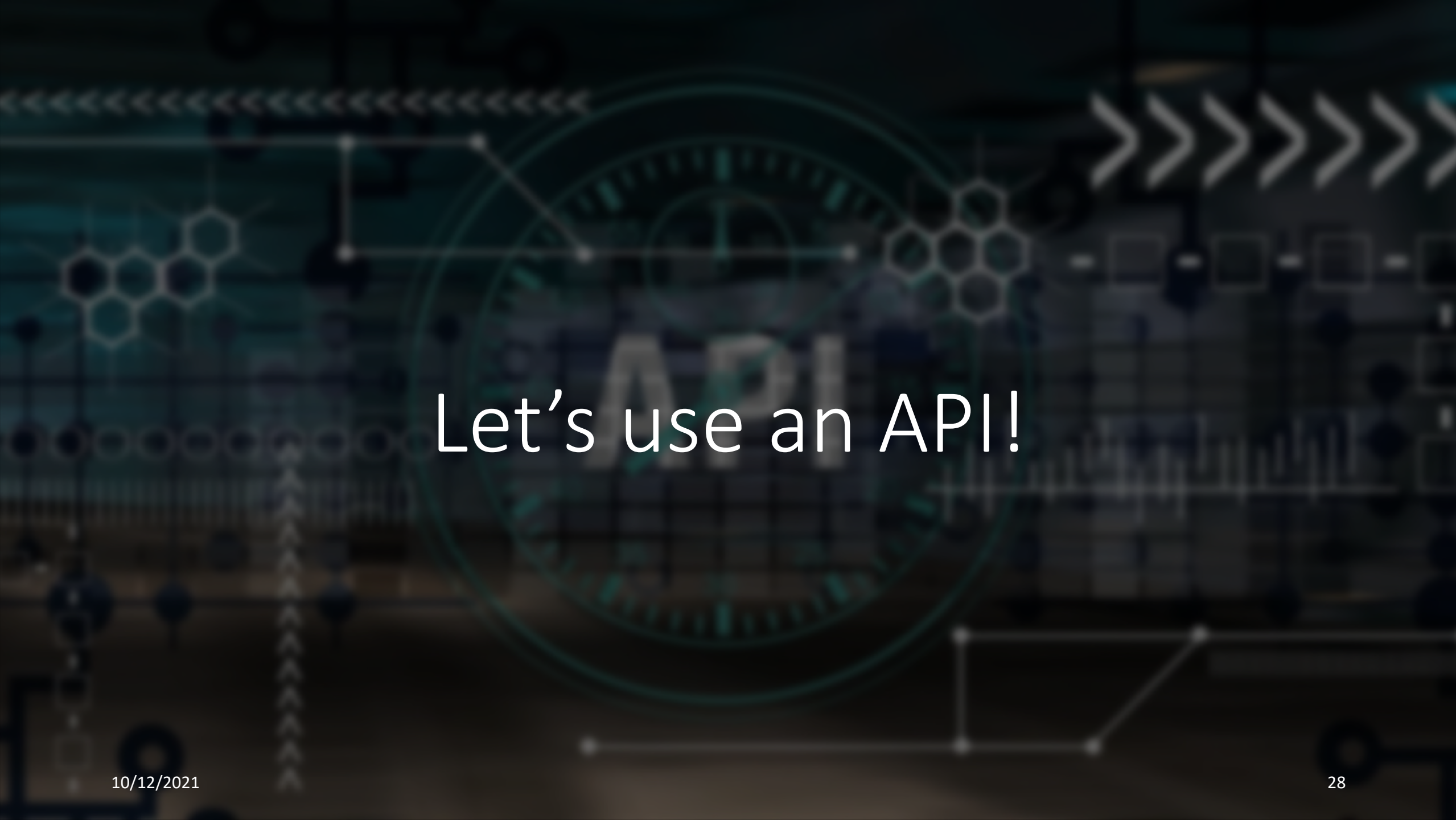
Is data from Wikipedia free of biases? Definitely no.

Being fair to the servers

- When you are web-scraping, you often send many, many requests to servers.
- This might be problematic:
  - In the best case, a server that has too many requests, will queue up the request.
  - In the worst case, the server will crash. You might get blocked.
  - As a web-scraper, you want to be fair.
    - Only send as many requests as you need.
    - Use timeouts to force breaks.

Advanced web scraping: Worksheet

- 3_webscrapingadvanced.Rmd
  - Extracting attributes
  - Loops
  - RSelenium



Let’s use an API!

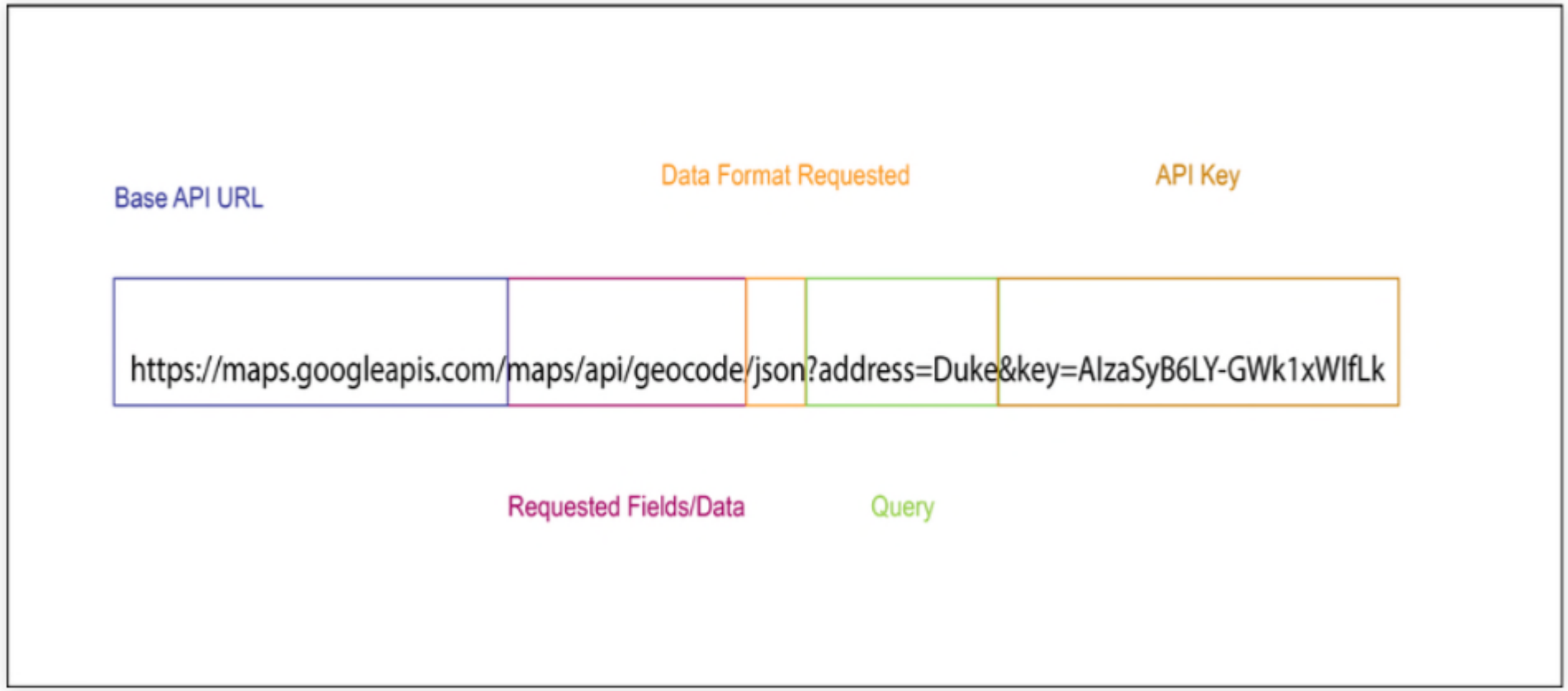
What is an API?

- Application Programming Interface

- Made for programs, not for humans
- First used for developers to extend platform use

- We send a more specific request to the webserver
- … if they allow us to.

Anatomy of an API call



Using an API

- Most APIs require some sort of authentication so you can access it (Oauth or API Key).
- Depending on the provider, getting this authentication can be more or less burdensome.
  - Some just require an email sign-up, others review your application in more detail.
  - And others want your money.
- We will use the Mediawiki API to access Wikipedia. It is free and does not need authentication.

Introduction to APIs: Worksheet

- 4_API
  - Sending a request

I want to get data from Twitter! And Youtube! And Facebook!

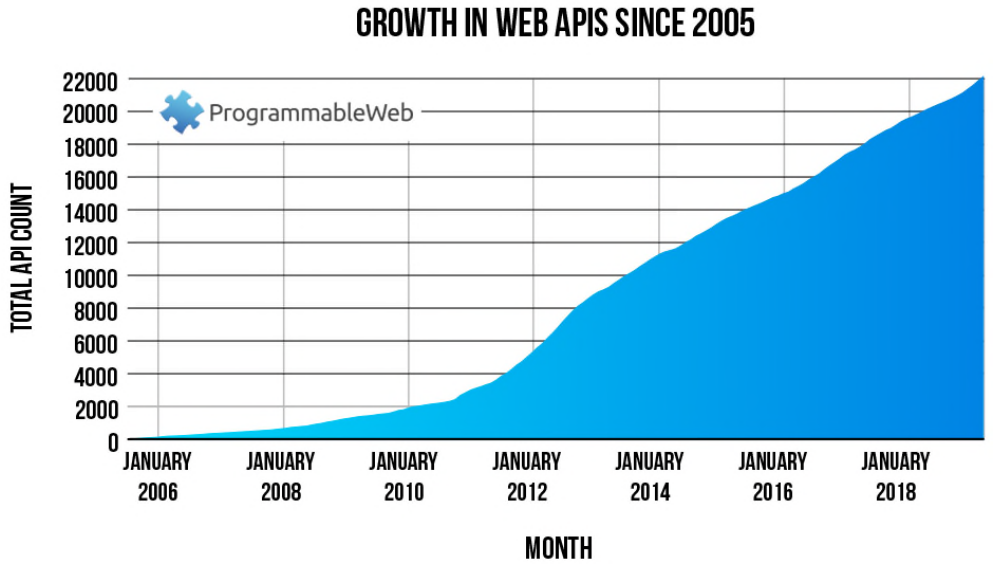
- Great!
- Twitter
  - Free API for academic researchers, but need to sign up for a developer account: https://developer.twitter.com/en/products/twitter-api/academic-research
  - R packages: rtweet, academictwitteR
- Facebook: https://developers.facebook.com/ (public pages)
- Youtube: https://developers.google.com/youtube/v3
- Google Maps: https://developers.google.com/maps ($$$)
- Guardian: https://open-platform.theguardian.com/

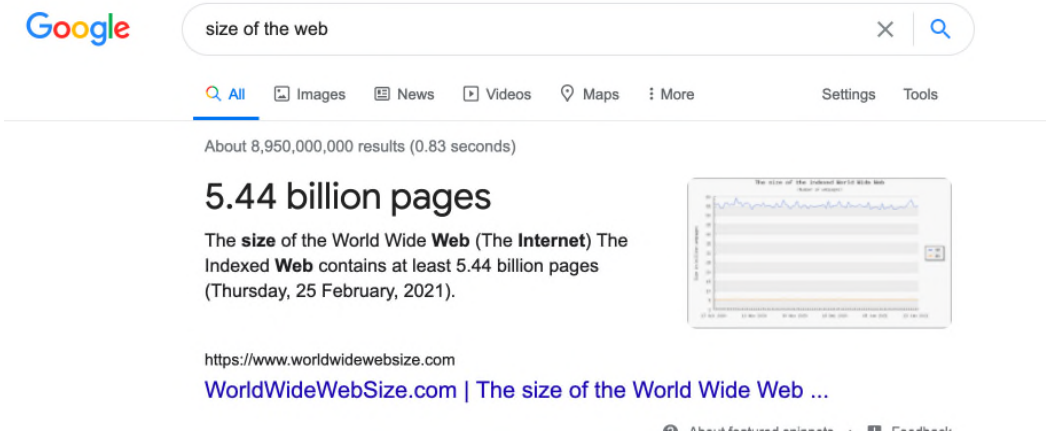
- List of public APIs: https://github.com/public-apis/public-apis

Scraping vs API

- APIs
  - Extract data from public/non-public and visible/non-visible webpage content.
  - Data comes pre-packaged according to specified query.
  - Potential APIs to use: >22k indexed on: https://www.programmableweb.com/api

- Scraping
  - Extracts data from public/visible webpage content.
  - Needs to be reformatted to usable format.
  - Potential data sources: universe of webpages in existence: >5bn.





This was fun! But…

Are web data going to solve all our problems and make every question answerable?

- No.

- Questions to ask:
  - Unit selection: Who/what do we want to study? Over what time frame?
  - Sampling: From where is the data coming? Are there (unknown) biases in the data-generating process?
  - Inference: To what population do our finding relate? To what phenomenon do our data speak?

Individuals with depression express more distorted thinking on social media (Bathina et al. 2021)

- Unit selection?
- Sampling?
- Inference?

ARTICLES
https://doi.org/10.1038/s41562-021-01050-7
nature human behaviour

Check for updates

Individuals with depression express more distorted thinking on social media

Krishna C. Bathina[1], Marijn ten Thij[1], Lorenzo Lorenzo-Luaces[2], Lauren A. Rutter[2] and Johan Bollen[1]✉

Depression is a leading cause of disability worldwide, but is often underdiagnosed and undertreated. Cognitive behavioural therapy holds that individuals with depression exhibit distorted modes of thinking, that is, cognitive distortions, that can negatively affect their emotions and motivation. Here, we show that the language of individuals with a self-reported diagnosis of depression on social media is characterized by higher levels of distorted thinking compared with a random sample. This effect is specific to the distorted nature of the expression and cannot be explained by the presence of specific topics, sentiment or first-person pronouns. This study identifies online language patterns that are indicative of depression-related distorted thinking. We caution that any future applications of this research should carefully consider ethical and data privacy issues.

Can we just collect everything and anything?

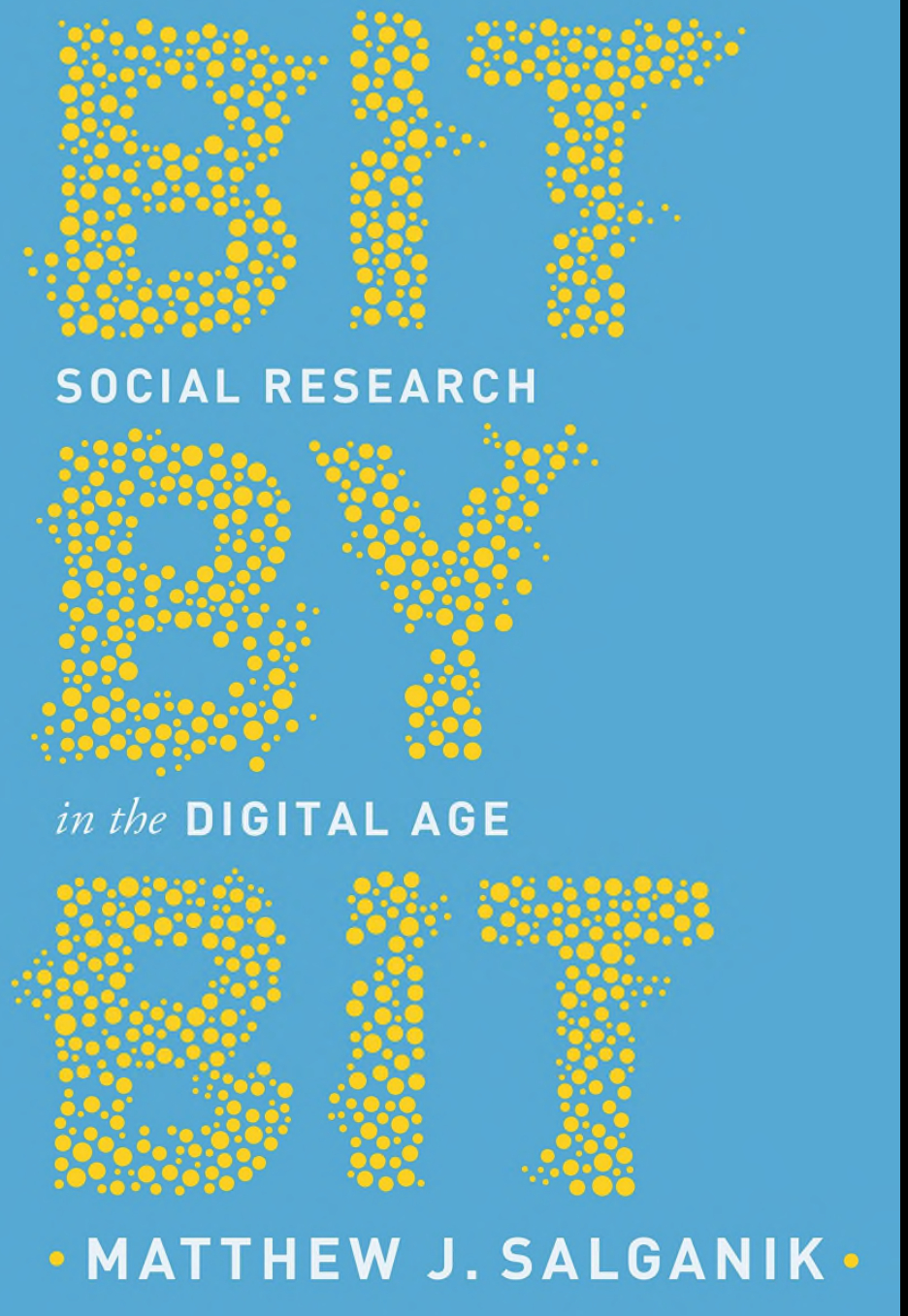
- No.

- Legal constraints placed by platforms (terms of services)
- Ethical protection of users' privacy and contextual integrity
  - Protection of minorities and vulnerable groups
  - Users are not posting on social media to become research observations.

Wrapping it up

What have we learnt today?

- You know the basics about HTML.
- You know how to collect a webpage using `rvest` and how to select elements with CSS selectors.
- You know that you can automate browsers with RSelenium.
- You know how an API works.
- You know that web data needs to be questioned.
- You know that web data comes with ethical and legal challenges.



How to go forward

- Go get your web data!
- But consider the terms of services, the privacy of the users, and your research question. Web-scraping in the social sciences is a means to an end.

- Suggested reading: Matthew Salganik (2019). Bit by Bit: Social Research in the Digital Age. Princeton University Press.