

# Automated Web Data Collection

Nicole Schwitter | Masterclass University of Warwick

Nicole.Schwitter@uni-Mannheim

# Introductions



- Me
  - Postdoctoral researcher at the Mannheim Centre for Social Science Research
  - Research fellow in the Department of Sociology
  - Nine years experience with web data collection
- You
  - Expectations of the course?
    - [www.menti.com](http://www.menti.com) code: 3765 5229 (or directly: <https://www.menti.com/aljfw68vdqnk>)
  - 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 APIs   Approach and exercises
12:00	Lunch
13:00	Session 3 Web Scraping   Approach and exercises
15:00	Break
15:15	Continue Session 3 // 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-warwick2024>
- Worksheets
  - We use R and Rstudio.
  - Examples are based on free resources which do not require an access key and which do not pose ethical issues.
- 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* (feel free use ChatGPT to ask for Python-translations of my code).
- I use Stata.
  - Oh no. Have a look at R or Python.
- I use any other programming language.
  - There are ways to collect web data, 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 searchable 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>](http://www.mit.edu:8001/afs/sipb/user/mkgray/ht/comprehensive.html)

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 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

# ... Google, is it you?

Q: Setting User-Agent Field?

Subscribe



125239 views



Lawrence Page

to

Jan 7, 1996, 9:00:00 AM



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



# Making web data accessible: Early APIs

**eBay™ Developers Program**

Search  Go

Tips for searching  
Secure Search

Getting Started  
Development Products  
Apply for the Developers Program  
Application Certification  
Developer Support  
News

## News

**eBay Launches New Initiative to Provide Expanded E-Commerce Solutions**

SAN JOSE, Calif., Nov. 20 – eBay®, the world's leading online trading community, today announced a new initiative to provide expanded e-commerce solutions for individuals and small businesses to drive their own businesses. The groundbreaking initiative will initially be rolled out in beta form.

"Our new API has tremendous potential to revolutionize the way people do business on eBay by providing the tools that developers need to create applications based on eBay technology," said eBay President of Product Management and Marketing, David Skok.

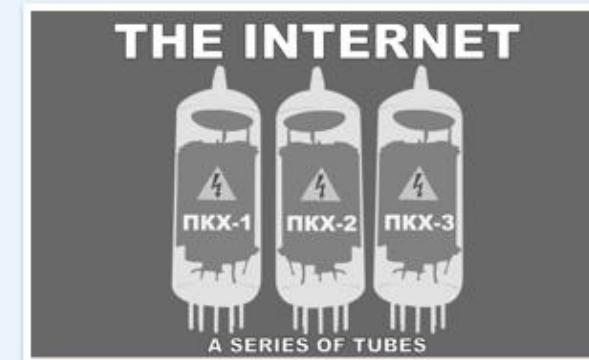
The API will provide three basic benefits. First, it will allow eBay to be fully integrated into existing e-commerce platforms, saving users time and money from the start-up process. For example, a site selling musical instruments through eBay can now offer its customers a one-stop shop for all their needs.

**twitter blog**

## Introducing the Twitter API

Wednesday, September 20, 2006

Some smart folks out there on the Internet have been putting together interesting projects like [this map](#) and [this page](#) without any help from us so we thought it was high time to release an API.



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

# Early (re-)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<sup>1</sup> 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.<sup>2</sup> 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.*



# Current Examples: Web Data in Social Science Research

Article

Right-Wing YouTube

The International Journal of Press/Politics  
1–34

rossMark  
click for updates

Archive About ▾ Submit man

SCIENCE ADVANCES | RESEARCH ARTICLE

CORONAVIRUS

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

Jon Green<sup>1</sup>, Jared Edgerton<sup>1</sup>, Daniel Naftel<sup>1</sup>, Kelsey Shoub<sup>2</sup>, Skyler J. Cranmer<sup>1\*</sup>

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.

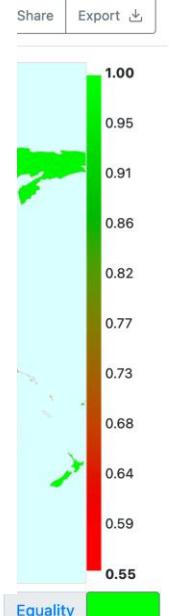
Masoomali Fatehkia

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

### Keywords

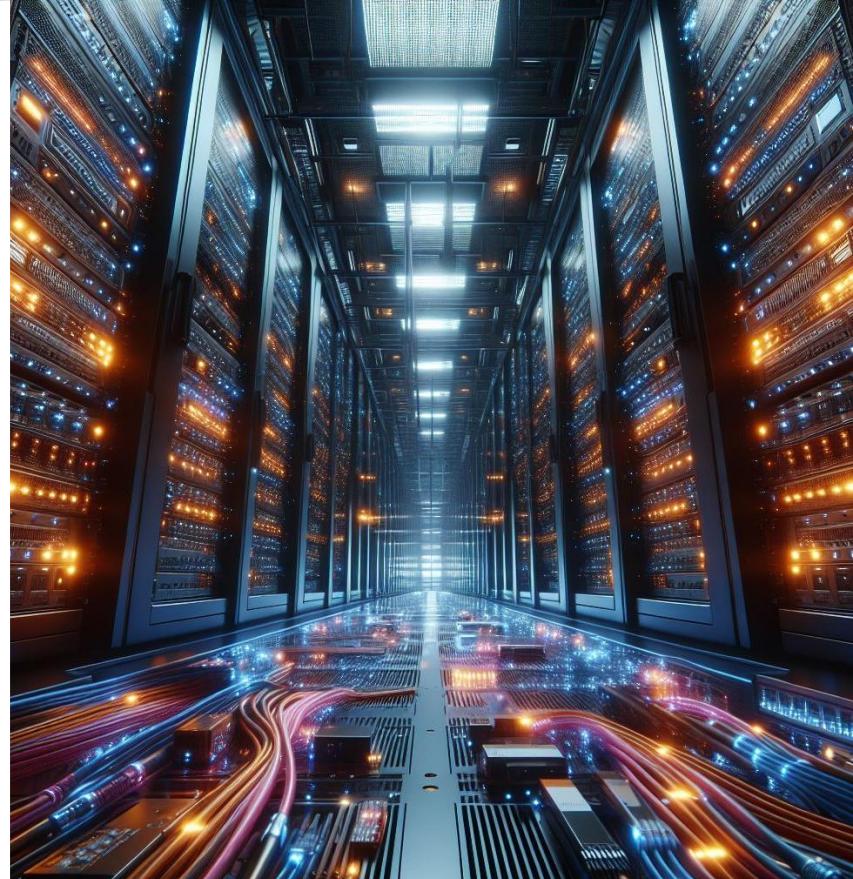
YouTube, radicalization, conservatism, political extremism

Copyright © 2020  
The Authors, some  
rights reserved;  
exclusive licensee  
American Association  
for the Advancement  
of Science. No claim to  
original U.S. Government  
Works. Distributed  
under a Creative  
Commons Attribution  
NonCommercial  
License 4.0 (CC BY-NC).





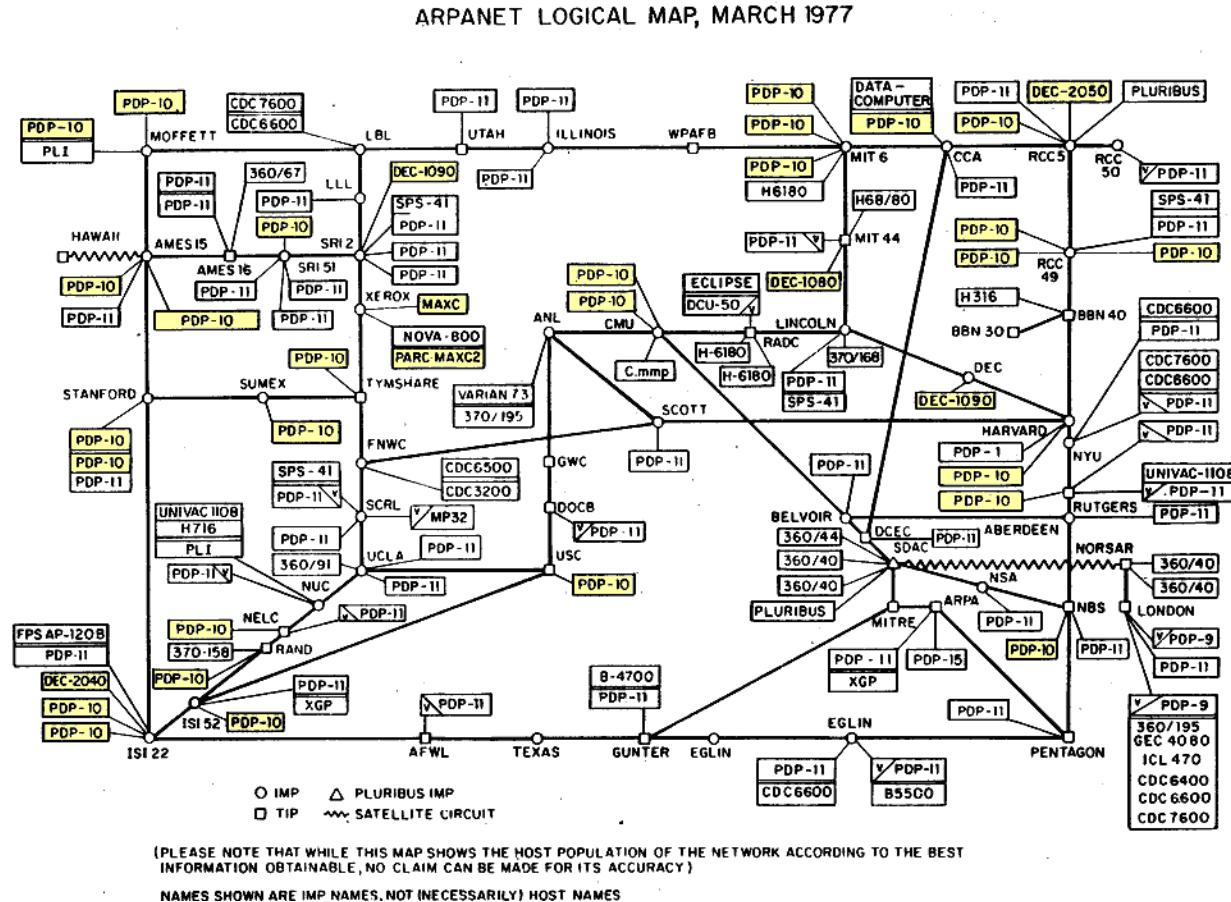
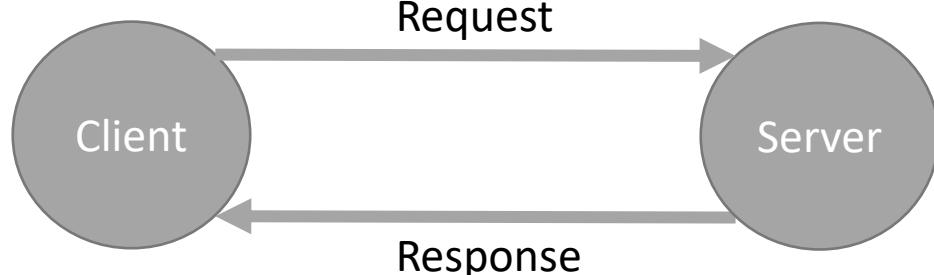
# How do we get the data?



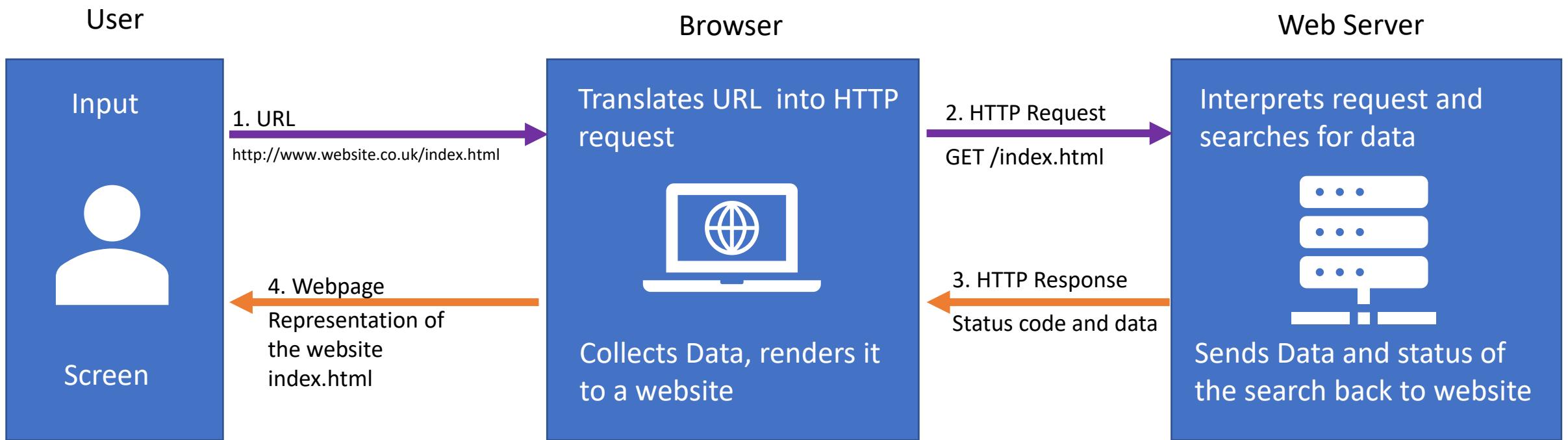
# First... how does the world wide web even work?

# The internet and the web

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

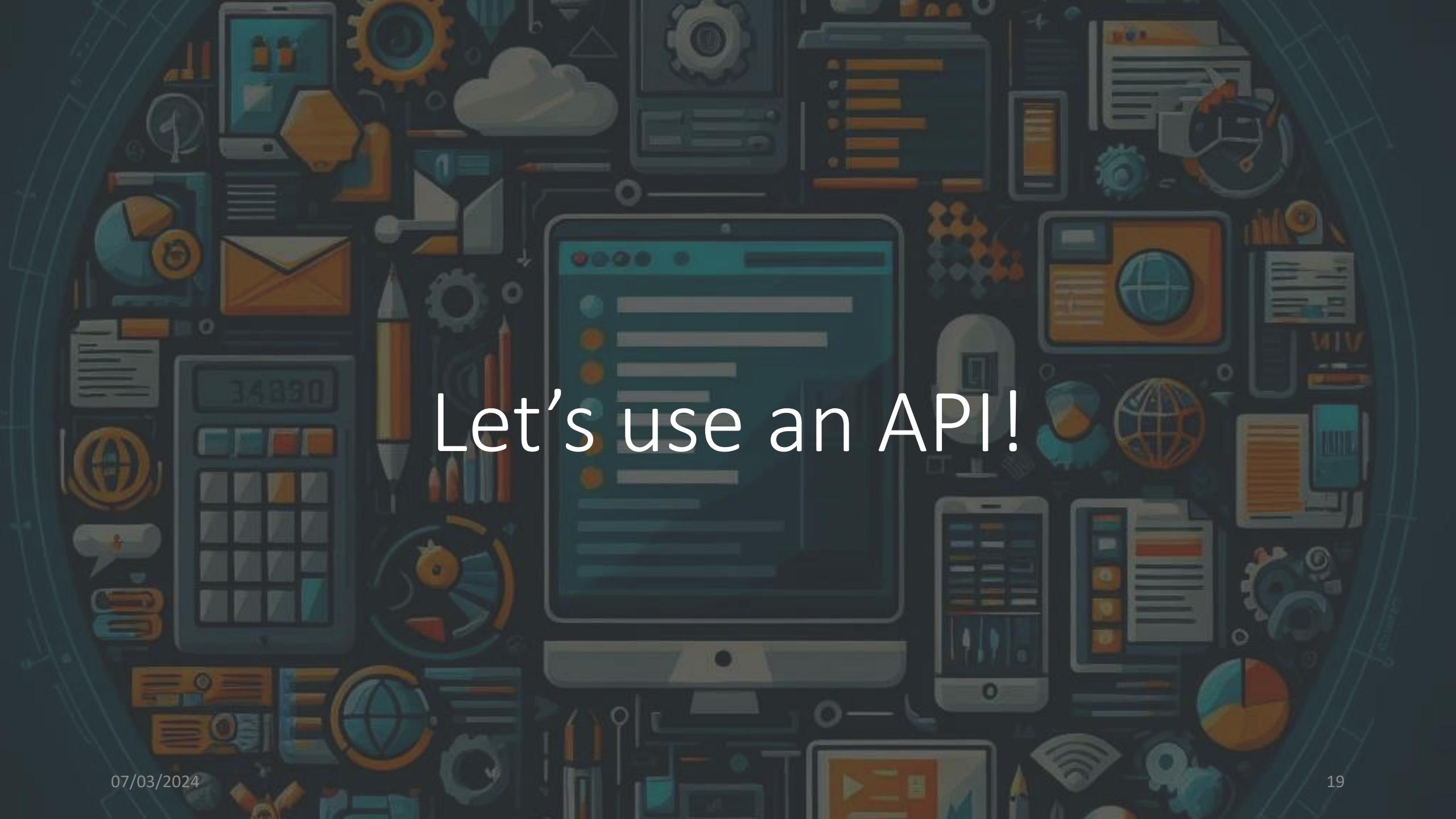


# Communication process (HTTP)



# Getting the data

- Ctrl + c, Ctrl + v from displayed website
  - Tedious, error-prone, slow
  - Unstructured data: Sometimes, it might be your only option!
- APIs
  - Sending your own data requests to the server
  - Structured data
- 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”

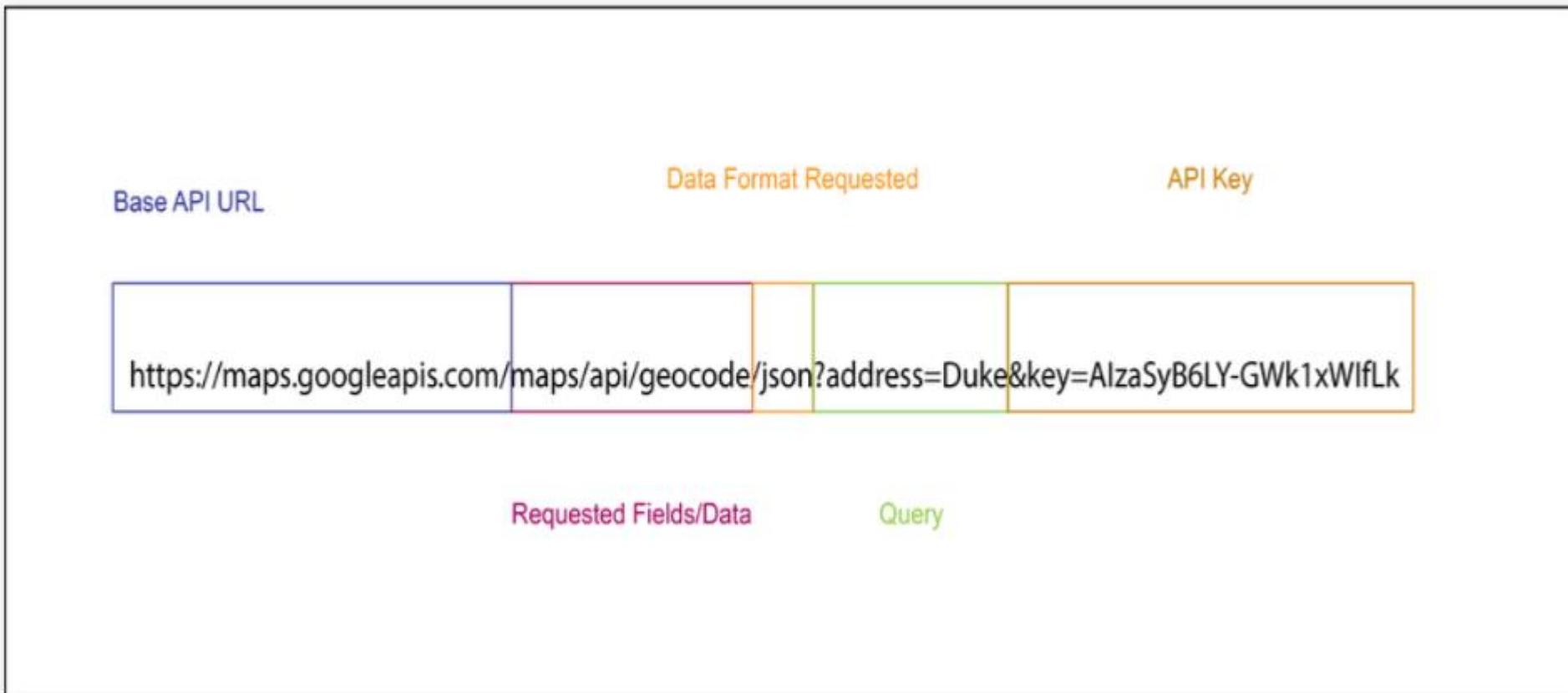


# Let's use an API!

# What is an API?

- Application Programming Interface
- Offer made by internet platform providers
- 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 Chronicling America API to access historic American newspapers. It is free and does not need authentication.
- Documentation: <https://chroniclingamerica.loc.gov/about/api/>

# Let's start with R: Worksheets

- 1\_API
  - Intro to RMarkdown
  - Sending a request
  - Receiving a response

# I want to get data from Twitter/X! And Youtube! And Facebook!

- Great!
- Twitter/X: <https://developer.twitter.com/en/use-cases/do-research/academic-research> (\$\$\$)
- 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>

# Let's web scrape!

# Web scraping

- An API call gave us a structured response.
- When 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.

# Basic web scraping: Worksheets

- `2_HTML.Rmd`
  - Intro to HTML
  - CSS Selectors
- `3_basicWebScraping.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<sup>2</sup>):** 468.0

### United Arab Emirates

**Capital:** Abu Dhabi  
**Population:** 4975593  
**Area (km<sup>2</sup>):** 82880.0

### Afghanistan

**Capital:** Kabul  
**Population:** 29121286  
**Area (km<sup>2</sup>):** 647500.0

### Antigua and Barbuda

**Capital:** St. John's  
**Population:** 86754  
**Area (km<sup>2</sup>):** 443.0

### Anguilla

**Capital:** The Valley  
**Population:** 13254  
**Area (km<sup>2</sup>):** 102.0

### Albania

**Capital:** Tirana  
**Population:** 2986952  
**Area (km<sup>2</sup>):** 28748.0

### Armenia

**Capital:** Yerevan  
**Population:** 2968000  
**Area (km<sup>2</sup>):** 29800.0

### Angola

**Capital:** Luanda  
**Population:** 13068161  
**Area (km<sup>2</sup>):** 1246700.0

### Antarctica

**Capital:** None  
**Population:** 0  
**Area (km<sup>2</sup>):** 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*.<sup>[2][3]</sup> These figures refer to the [de facto](#) population in a country or area as shown in the "estimates" section.

### Contents [hide]

- [1 See also](#)
- [2 Notes](#)
- [3 References](#)
- [4 External links](#)

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



Statistical subregions as defined by the United Nations Statistics Division.<sup>[1]</sup>

Is data from Wikipedia free of biases? Definitely not.

# 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

- 4\_advancedWebScraping.Rmd
  - Extracting attributes
  - Loops
  - RSelenium

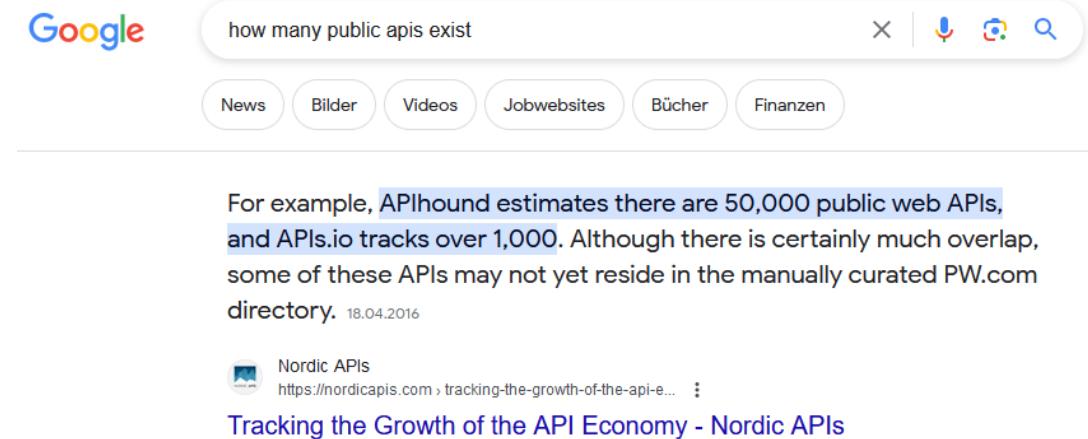
# Scraping vs APIs

- APIs

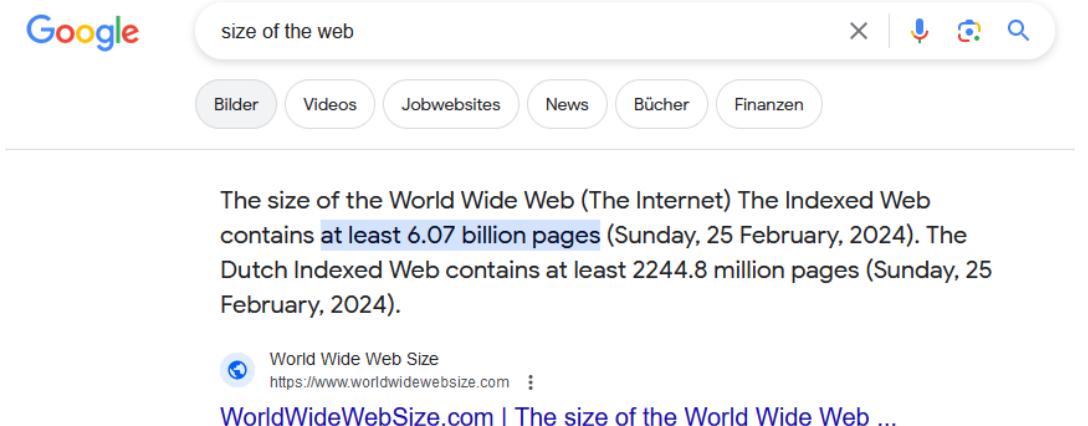
- Extract data from public/non-public and visible/non-visible webpage content.
- Data comes pre-packaged according to specified query.
- Potential APIs to use: growing, but limited

- Scraping

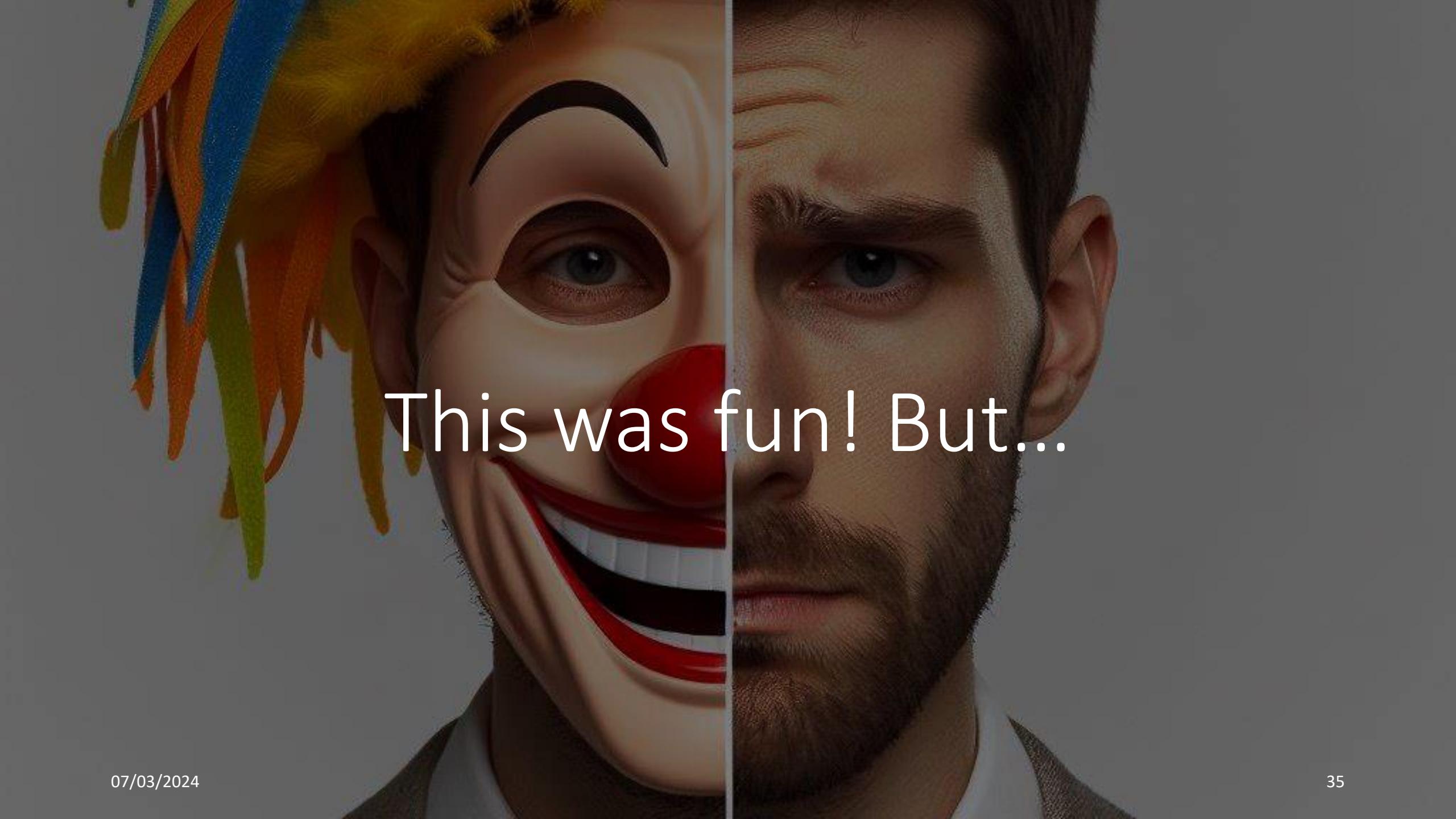
- Extracts data from public/visible webpage content.
- Needs to be reformatted to usable format.
- Potential data sources: universe of webpages in existence: >6bn.



Google search results for "how many public apis exist". The search bar shows the query. Below it are several search results. The first result is a snippet from Nordic APIs' website, which states: "For example, APIhound estimates there are 50,000 public web APIs, and APIs.io tracks over 1,000. Although there is certainly much overlap, some of these APIs may not yet reside in the manually curated PW.com directory." A timestamp indicates the information is from 18.04.2016. Another result link leads to "Tracking the Growth of the API Economy - Nordic APIs".



Google search results for "size of the web". The search bar shows the query. Below it are several search results. The first result is a snippet from WorldWideWebSize.com, stating: "The size of the World Wide Web (The Internet) The Indexed Web contains at least 6.07 billion pages (Sunday, 25 February, 2024). The Dutch Indexed Web contains at least 2244.8 million pages (Sunday, 25 February, 2024)." Another result link leads to "WorldWideWebSize.com | The size of the World Wide Web ...".



This was fun! But...

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

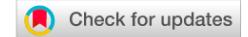
- No.
- Questions to ask:
  - Unit selection: Who/what do we want to study? Over what time frame?
  - Sampling: From where is the data coming? Are there (unknown) biases in the data-generating process?
  - Inference: To what population do our finding relate? To what phenomenon do our data speak?

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

ARTICLES

<https://doi.org/10.1038/s41562-021-01050-7>

nature  
human behaviour



- Unit selection?
- Sampling?
- Inference?

## Individuals with depression express more distorted thinking on social media

Krishna C. Bathina<sup>1</sup>, Marijn ten Thij<sup>1</sup> , Lorenzo Lorenzo-Luaces<sup>1</sup> , Lauren A. Rutter<sup>1</sup> and Johan Bollen<sup>1</sup>

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.

A large pile of wrapped Christmas gifts, including boxes of various sizes wrapped in colorful paper like red with white dots and stripes, and some with gold or silver foil. Many are tied with ribbons in shades of red, green, and blue, and some have small gift tags attached. A small green Christmas tree ornament sits atop one of the larger boxes.

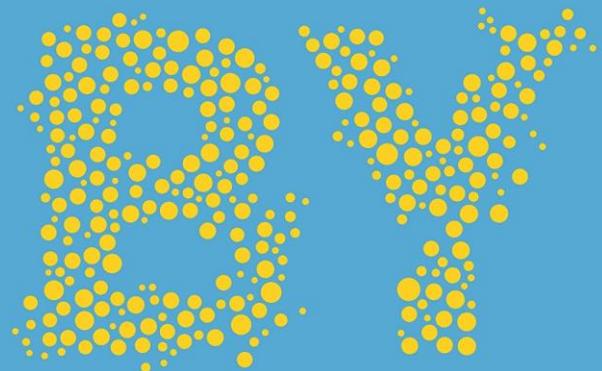
# Wrapping it up

# What have we learnt today?

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



SOCIAL RESEARCH



*in the* DIGITAL AGE



• MATTHEW J. SALGANIK •

# 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.