
ACQUIRING BIG DATA;
ETHICS & LAW

THIS LECTURE

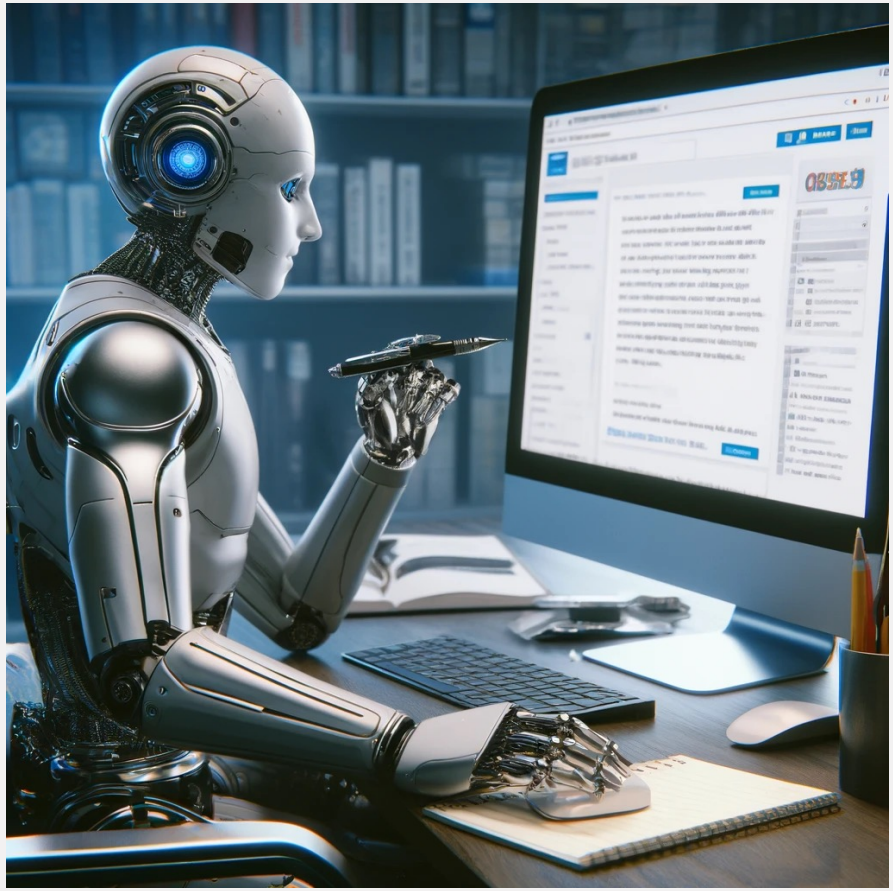
Part 1. Web scraping and APIs

- Important concepts and ideas (before actually doing it on Wednesday)

Part 2. Law & Ethics of Computational Social Science

- What laws are relevant for CSS research? GDPR
- How do you make sure you're an ethical researcher?
- Practical skills: How to write an IRB request, Data Management Plan, Ethics section?


Part 1: The art of web scraping



WHAT IS WEB SCRAPING?

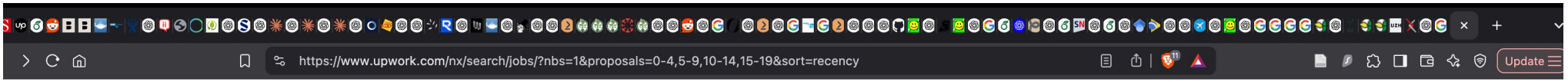
Using a script to pretend to be a user and automatically collect data from websites.

Transform unstructured web data, typically HTML or JavaScript, into structured data that can be stored and analyzed in a format like CSV, JSON, or a pandas dataframe.

EXAMPLE USES OF WEB SCRAPING

Observational data from websites and platform are among the most important data sources for CSS research:

- Collect social media posts from a platform
- Collecting listings and reviews from Airbnb
- Getting job listings from a job portal
- Getting news paper articles
- Get websites and the links between them
- *... other examples?*



upwork Find work Deliver work Manage finances Messages Search Jobs

Search Advanced search

Less than 5 Proposals × 5 to 10 Proposals × 10 to 15 Proposals × 15 to 20 Proposals × Clear filters

Category

Select Categories

Experience level

- Entry Level (8,708)
- Intermediate (80,955)
- Expert (30,695)

Job type

- Hourly (71,998)
  - $ Min /hr $ Max /hr
- Fixed-Price (48,396)
  - Less than $100 (22,611)
  - $100 to $500 (17,285)
  - $500 - $1K (4,041)
  - $1K - $5K (3,682)
  - $5K+ (776)

Save search Saved jobs Sort by: Newest

Posted 59 seconds ago

Looking for Business developer or lead generation expert

Payment unverified 4.1 $5 spent PAK

Hourly: $5.00 - $6.00 - Expert - Est. time: Less than 1 month, Less than 30 hrs/week

Hello everyone, I am looking for a skilled Lead Generation Expert to identify and deliver high-quality, sales-ready leads. My focus is on clients actively seeking to build custom web solutions, and i need someone who can target decision-makers ready to purchase within this niche. Key...

Social Media Lead Generation | Lead Nurturing | Lead Generation Content Creation | Lead Generation Analysis | +6

Posted 1 minute ago

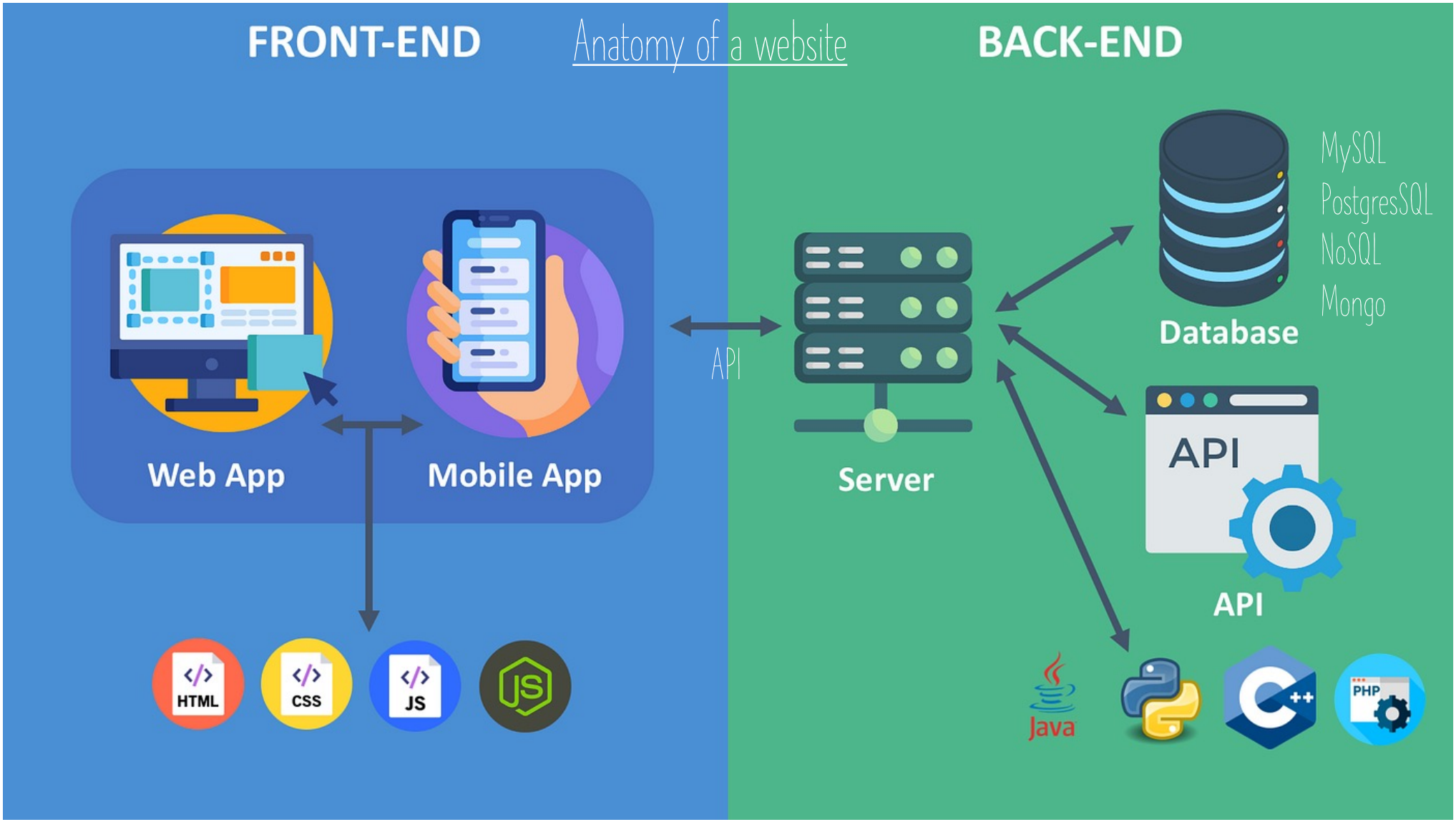
3D Floor Plan Design for Office Space

Payment verified 1.8 $6K+ spent Australia

Fixed price - Intermediate - Est. budget: $200.00

We are looking for a skilled 3D designer to create an accurate floor plan of our office space, including a revised layout featuring furniture placement. The ideal candidate will visualize our requirements and incorporate feedback to produce an appealing and functional design. Experience in...

Interior Design | Architectural Design | SketchUp | Autodesk AutoCAD | 3D Rendering


FRONT-END
Anatomy of a website
BACK-END
Web App
Mobile App
API
Server
Database
MySQL
PostgresSQL
NoSQL
Mongo
API
API
HTML
CSS
JS
Java
PHP

FRONT-END LANGUAGES

This code is being run by your web browser on your computer.

HTML (Hypertext Markup Language): HTML provides the basic structure of the website, using a system of tags and attributes to denote text, links, images, and other content.

CSS (Cascading Style Sheets): CSS is used for styling and visually formatting the HTML elements. It controls layout, colors, fonts, and even some animations.

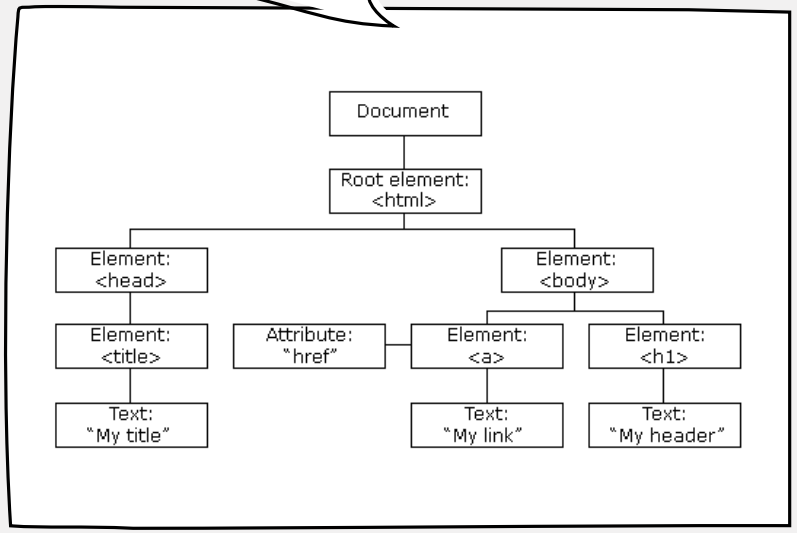
JavaScript: JavaScript adds interactivity to web pages. It's a scripting language that enables dynamic content, control multimedia, animate images, and much more.

HTML USUALLY HAS OUR CONTENT

HTML is a series of elements or tags that structure content, allowing browsers to display text, links, images, and other resources.

- This is a paragraph.
- This is a link.
- This
is a list
- This is an object of the class 'box'

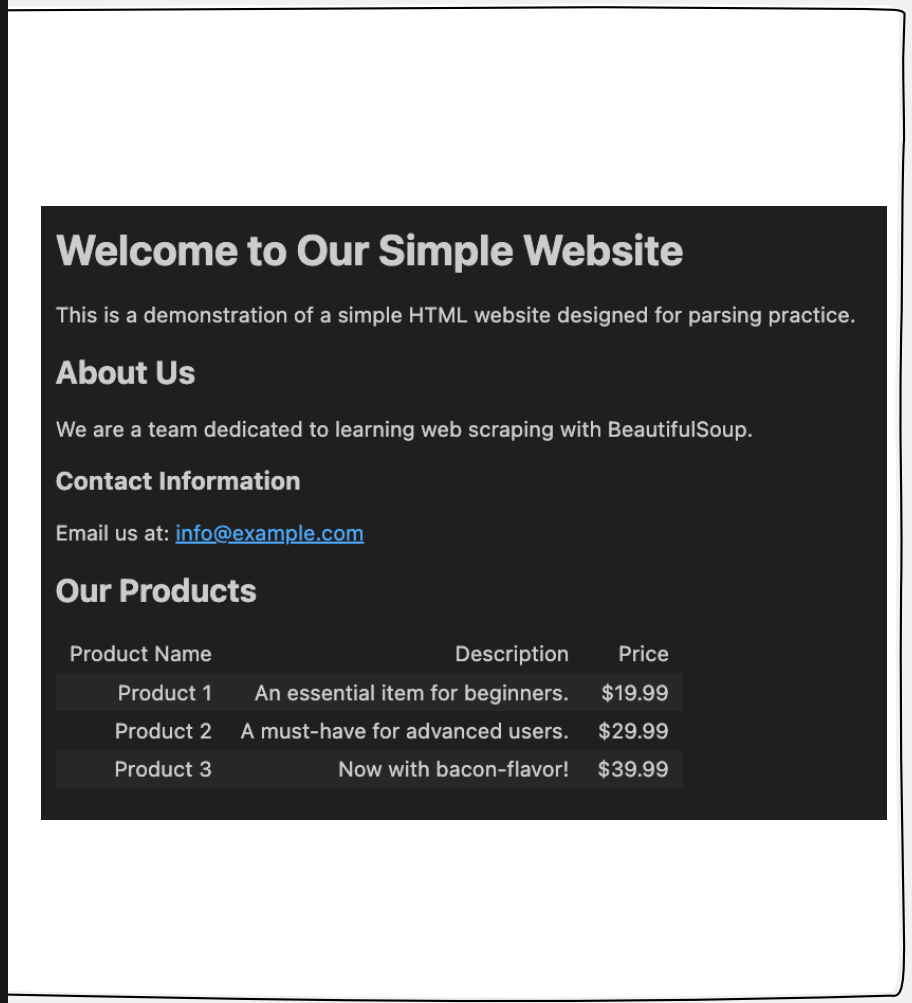
THE DOCUMENT OBJECT MODEL (DOM)



- The DOM is a programming interface for web documents.
- Each element in an HTML document is represented as a node in the DOM tree, making it possible to traverse the hierarchy and access elements programmatically.
- When you use web scraping tools, you're essentially navigating the DOM to access specific parts of the page.

```html
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Simple Website Example</title>
</head>
<body>
<h1>Welcome to Our Simple Website</h1>
<p>This is a demonstration of a simple HTML website designed for parsing practice.</p>
<h2>About Us</h2>
<p>We are a team dedicated to learning web scraping with BeautifulSoup.</p>
<h3>Contact Information</h3>
<p>Email us at: <a href="mailto:info@example.com">info@example.com</a></p>
<h2>Our Products</h2>
<table border="1">
    <tr>
        <th>Product Name</th>
        <th>Description</th>
        <th>Price</th>
    </tr>
    <tr>
        <td>Product 1</td>
        <td>An essential item for beginners.</td>
        <td>$19.99</td>
    </tr>
    <tr>
        <td>Product 2</td>
        <td>A must-have for advanced users.</td>
        <td>$29.99</td>
    </tr>
    <tr>
        <td>Product 3</td>
        <td>Now with bacon-flavor!</td>
        <td>$39.99</td>
    </tr>
</table>

</body>
</html>'''
```



Welcome to Our Simple Website

This is a demonstration of a simple HTML website designed for parsing practice.

About Us

We are a team dedicated to learning web scraping with BeautifulSoup.

Contact Information

Email us at: info@example.com

Our Products

Product Name	Description	Price
Product 1	An essential item for beginners.	$19.99
Product 2	A must-have for advanced users.	$29.99
Product 3	Now with bacon-flavor!	$39.99

Onderwijs Onderzoek Nieuws & Agenda Over de UvA Bibliotheek

RESEARCH MASTER Vergelijk

Social Sciences

To solve complex societal issues, we must first investigate them. In this research master's programme, you'll gain diverse perspectives from various social sciences, while developing advanced research skills. This unique combination of interdisciplinary knowledge and research expertise empowers you to explore topics you care deeply about and contribute effectively to innovative solutions

Language of instruction
English

Mode
Full-time

Location
Roeterseiland campus

🎓 Combination of an interdisciplinary focus and rigorous methodological training

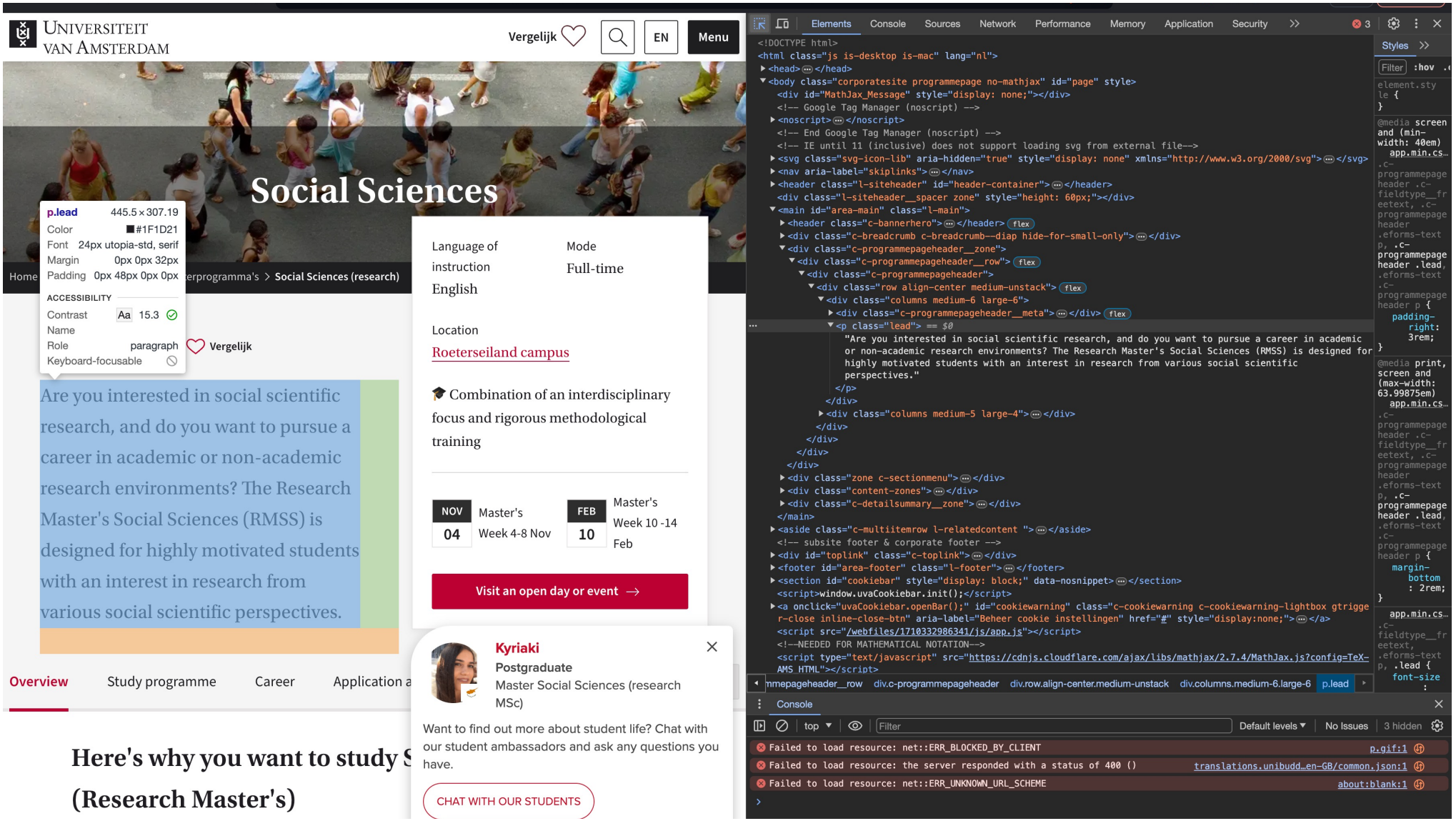
NOV 15 Master's Week: 15-21 Nov

FEB 09 Master's Week 9 -13 Feb

Visit an open day or event →

Receive the programme summary →

Keep me informed →


Universiteit van Amsterdam
Vergelijk
EN
Menu
Social Sciences
p.lead 445.5 × 307.19
Color #1F1D21
Font 24px utopia-std, serif
Margin 0px 0px 32px
Padding 0px 48px 0px 0px
ACCESSIBILITY
Contrast Aa 15.3
Name
Role paragraph
Keyboard-focusable
Home
erprogramma's > Social Sciences (research)
Vergelijk
Are you interested in social scientific research, and do you want to pursue a career in academic or non-academic research environments? The Research Master's Social Sciences (RMSS) is designed for highly motivated students with an interest in research from various social scientific perspectives.
Language of instruction
English
Mode
Full-time
Location
Roeterseiland campus
Combination of an interdisciplinary focus and rigorous methodological training
NOV 04 Master's Week 4-8 Nov
FEB 10 Master's Week 10 -14 Feb
Visit an open day or event
Overview
Study programme
Career
Application a
Here's why you want to study S
(Research Master's)
Kyriaki
Postgraduate
Master Social Sciences (research MSc)
Want to find out more about student life? Chat with our student ambassadors and ask any questions you have.
CHAT WITH OUR STUDENTS
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Styles
Filter
:hov
<p class="lead"> == $0
"Are you interested in social scientific research, and do you want to pursue a career in academic or non-academic research environments? The Research Master's Social Sciences (RMSS) is designed for highly motivated students with an interest in research from various social scientific perspectives."
Console
Filter
Default levels
No Issues
3 hidden
Failed to load resource: net::ERR_BLOCKED_BY_CLIENT
p.gif:1
Failed to load resource: the server responded with a status of 400 ()
translations.unibudd…en-GB/common.json:1
Failed to load resource: net::ERR_UNKNOWN_URL_SCHEME
about:blank:1

PYTHON LIBRARIES FOR SCRAPING (WEDNESDAY!)

We use programming libraries to scrape:

Requests: simple protocol to communicate with the website and get the HTML

Simple, fast, lightweight and best choice for static pages.

BeautifulSoup: used to easily parse the HTML

Selenium: runs an entire web browser and pretends to be a user clicking around

Slow and heavy, but very capable and works with dynamic websites

EXAMPLE: REQUESTS + BEAUTIFULSOUP

```
import requests

url = "https://example.com"

# Send a GET request to the website
response = requests.get(url)

# Check if the request was successful
if response.status_code == 200:
    html_content = response.text
    print("Page content:")
    print(html_content)
else:
    print(f"Failed to retrieve page. Status code: {response.status_code}")


from bs4 import BeautifulSoup

soup = BeautifulSoup(html_content, 'html.parser')

# Example: Get all links on the page
for link in soup.find_all('a'):
    print(link.get('href'))
```

WEB SCRAPING CHALLENGES

Web scraping can be difficult as websites try to prevent scraping.

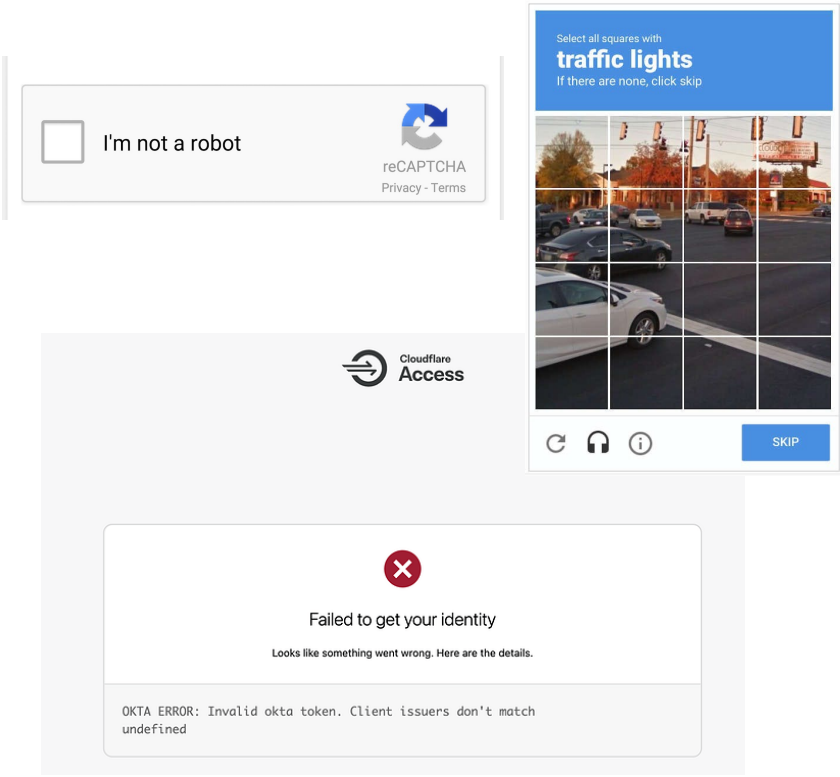
User-agent checks: You have to fake your identity

IP rate limit: Blocks too many calls

Cloudflare: US company. key infrastructure of the internet. Prevents script access to websites.

Captcha: Designed to prevent scraping.

Handling cookies and sessions: Websites can be complex



ROBOTS.TXT

A document that tells scrapers what the website owner is ok with you scraping or not.

Do you have to respect it?

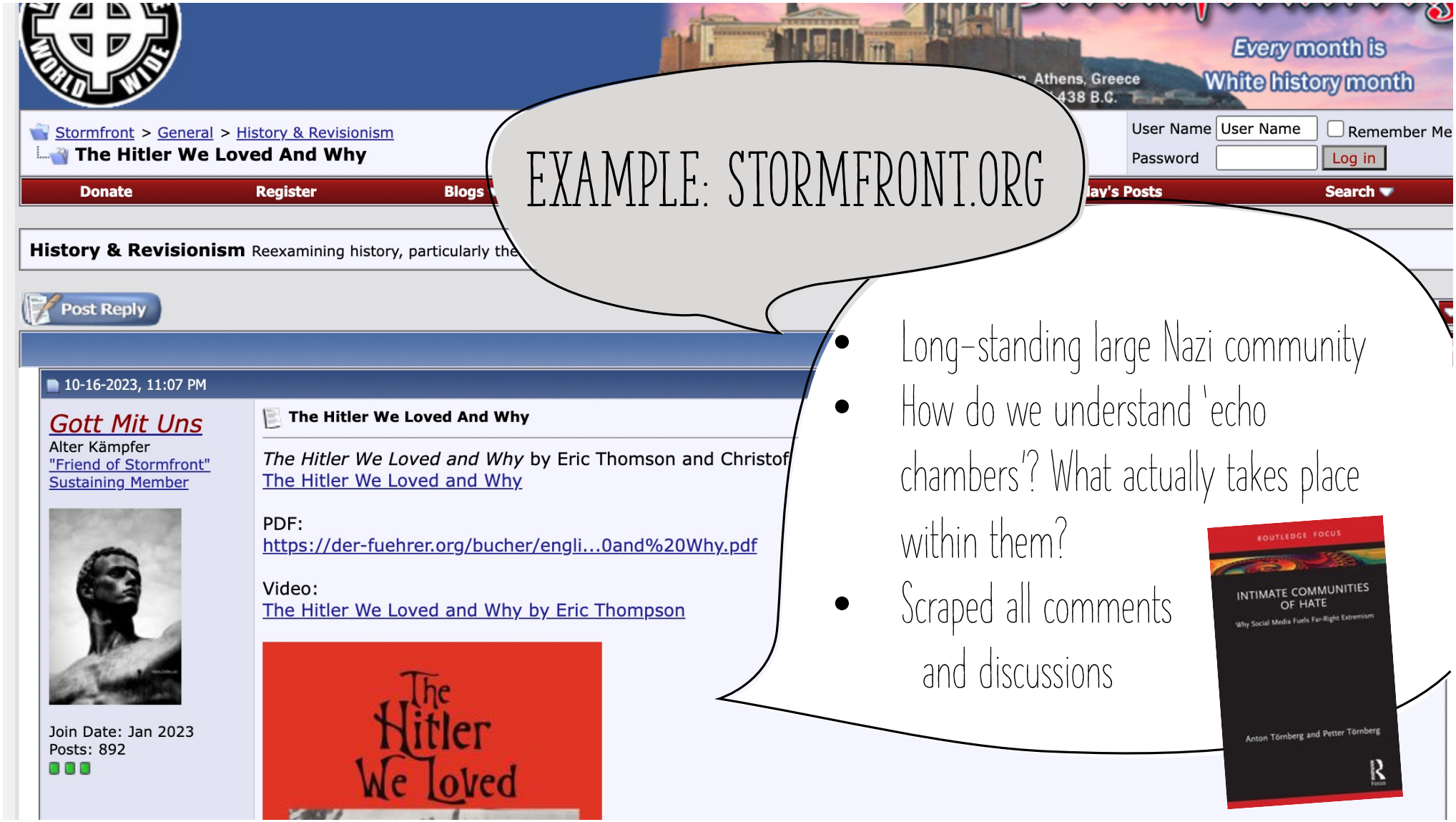
- Legally, no.

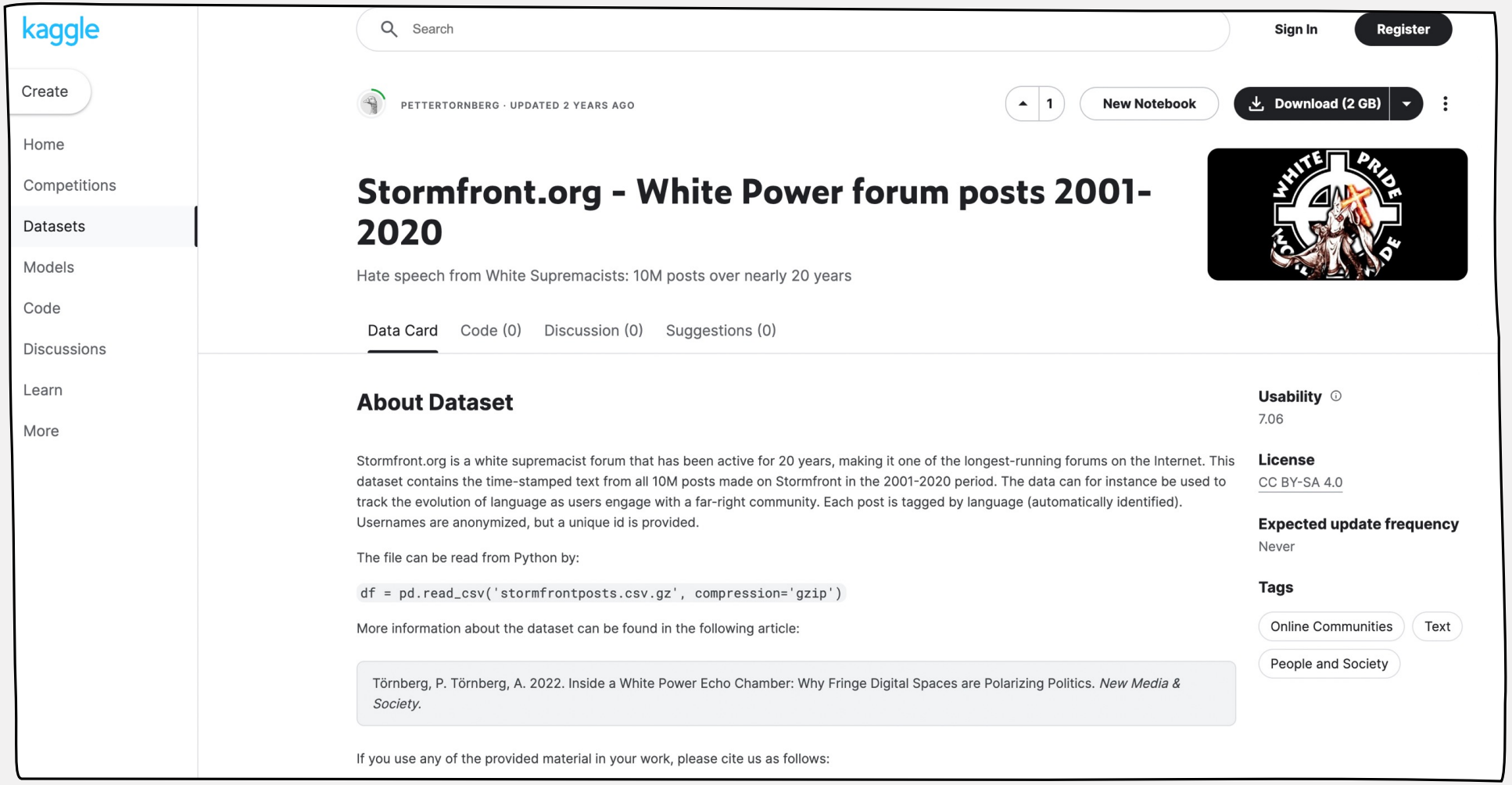
google.com/robots.txt:

```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/static
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*&
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&*&gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
```

SCRAPED DATA MUST BE CLEANED

- Remove duplicates
- Remove rows with missing data
- Fill in missing data
- Remove outliers
- Harmonize data
- Fix data types


Every month is White history month
Athens, Greece 438 B.C.
Stormfront > General > History & Revisionism
The Hitler We Loved And Why
User Name
Remember Me
Password
Log in
Donate
Register
Blogs
Search
History & Revisionism Reexamining history, particularly the
Post Reply
10-16-2023, 11:07 PM
Gott Mit Uns
Alter Kämpfer
"Friend of Stormfront"
Sustaining Member
Join Date: Jan 2023
Posts: 892
The Hitler We Loved And Why
The Hitler We Loved and Why by Eric Thomson and Christof
The Hitler We Loved and Why
PDF:
https://der-fuehrer.org/bucher/engli...0and%20Why.pdf
Video:
The Hitler We Loved and Why by Eric Thompson
The Hitler We Loved
EXAMPLE: STORMFRONT.ORG
Long-standing large Nazi community
How do we understand 'echo chambers'? What actually takes place within them?
Scraped all comments and discussions
ROUTLEDGE FOCUS
INTIMATE COMMUNITIES OF HATE
Why Social Media Fuels Far-Right Extremism
Anton Törnberg and Petter Törnberg


kaggle
Search
Sign In
Register
Create
Home
Competitions
Datasets
Models
Code
Discussions
Learn
More
PETTERTORNBERG · UPDATED 2 YEARS AGO
1
New Notebook
Download (2 GB)
Stormfront.org - White Power forum posts 2001-2020
Hate speech from White Supremacists: 10M posts over nearly 20 years
Data Card
Code (0)
Discussion (0)
Suggestions (0)
About Dataset
Stormfront.org is a white supremacist forum that has been active for 20 years, making it one of the longest-running forums on the Internet. This dataset contains the time-stamped text from all 10M posts made on Stormfront in the 2001-2020 period. The data can for instance be used to track the evolution of language as users engage with a far-right community. Each post is tagged by language (automatically identified). Usernames are anonymized, but a unique id is provided.
The file can be read from Python by:
df = pd.read_csv('stormfrontposts.csv.gz', compression='gzip')
More information about the dataset can be found in the following article:
Törnberg, P. Törnberg, A. 2022. Inside a White Power Echo Chamber: Why Fringe Digital Spaces are Polarizing Politics. New Media & Society.
If you use any of the provided material in your work, please cite us as follows:
Usability
7.06
License
CC BY-SA 4.0
Expected update frequency
Never
Tags
Online Communities
Text
People and Society

API (APPLICATION PROGRAMMING INTERFACE) DATA

- APIs are made to facilitate data exchange between applications.
- APIs send structured data, such as JSON, rather than HTML.
- Many platforms provide public API access, which allows you to get data.
- This data is very easy to parse programmatically:
  parsed_json = json.loads(json_string)

- APIs come with manuals, and often have libraries for making it even easier! E.g. Tweepy for Twitter.
- Research data from Twitter, Reddit, Facebook, etc., generally came from their APIs


```
{
  "longitude": 47.60,
  "latitude": 122.33,
  "forecasts": [
    {
      "date": "2015-09-01",
      "description": "sunny",
      "maxTemp": 22,
      "minTemp": 20,
      "windSpeed": 12,
      "danger": false
    },
    {
      "date": "2015-09-02",
      "description": "overcast",
      "maxTemp": 21,
      "minTemp": 17,
      "windSpeed": 15,
      "danger": false
    },
    {
      "date": "2015-09-03",
      "description": "raining",
      "maxTemp": 20,
      "minTemp": 18,
      "windSpeed": 13,
      "danger": false
    }
  ]
}
```

SOCIAL SCIENCE ANALYSIS AS A SERVICES

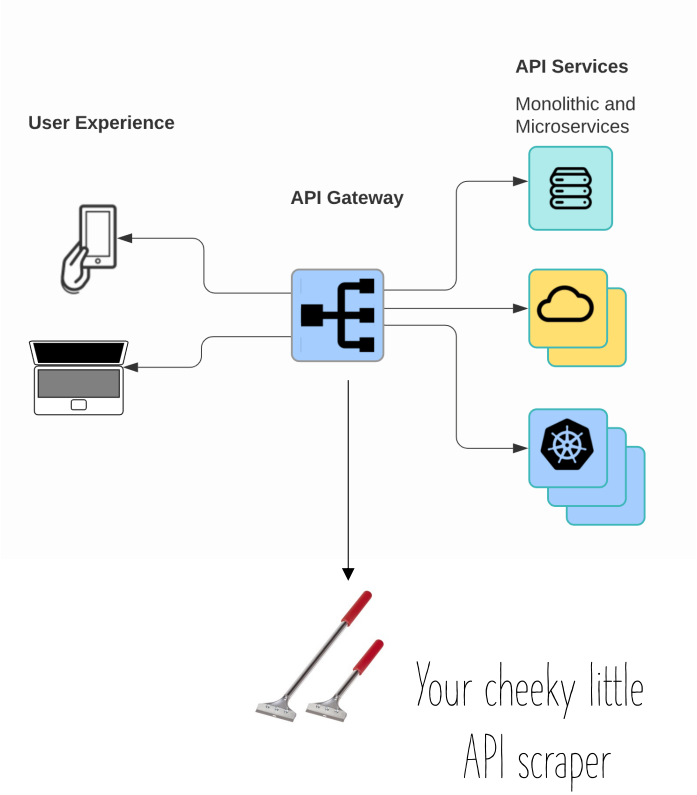
APIs also provide access to services: they allow code to interact with an online service.

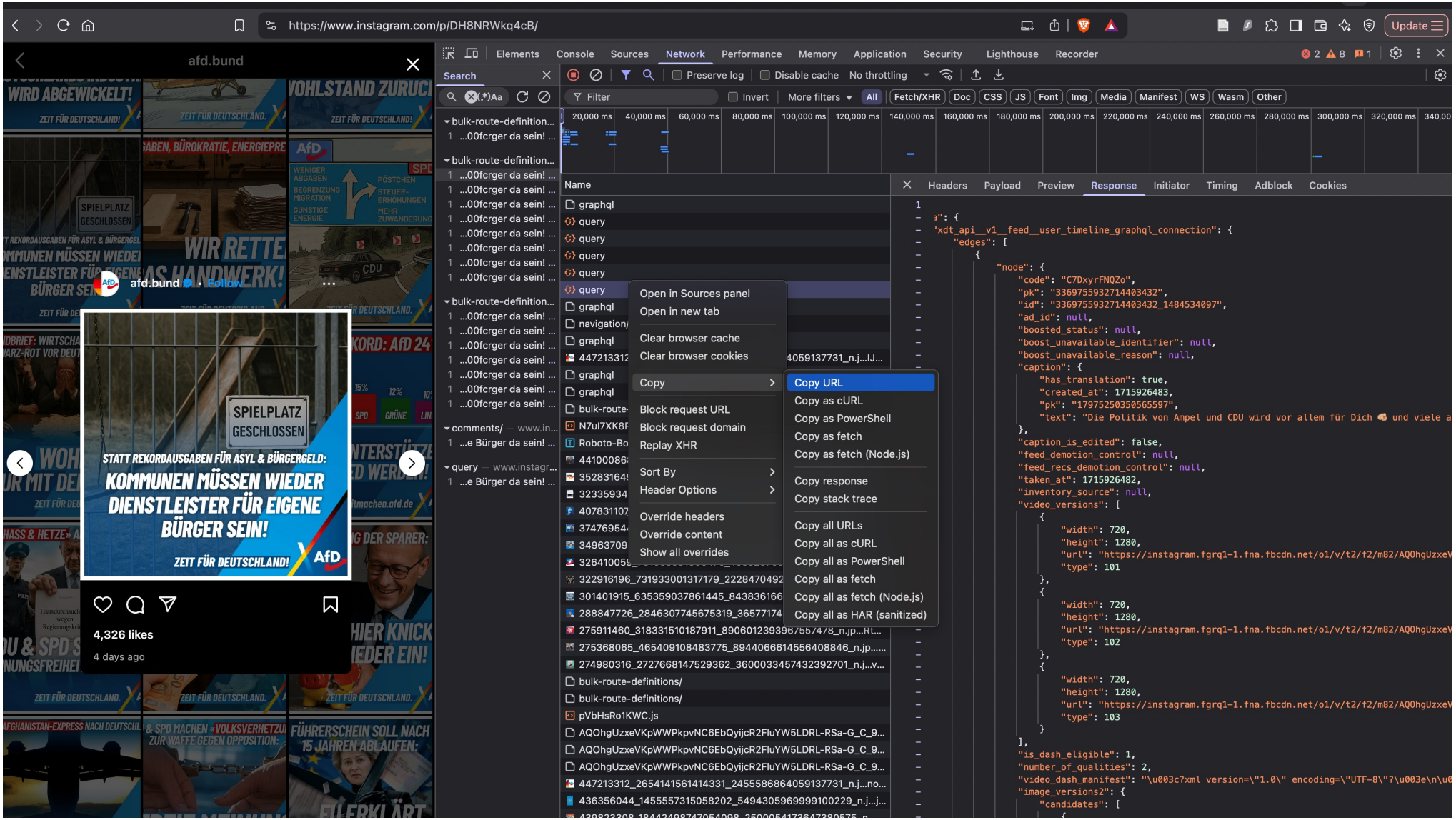
- MTurk is based on an API: humans-as-a-service
- OpenAI ChatGPT is available through API. So are almost all LLMs.
- Many AI, NLP and image analysis tools are available through APIs
- We will use Perspective Toxicity API to analyze how toxic a comment is

GETTING DATA FROM INTERNAL APIS

- Many dynamic websites and apps run on internal APIs
- Can also be a strategy for scraping: direct access to internal APIs by reverse engineering

For instance, Instagram!




https://www.instagram.com/p/DH8NRWkq4cB/
afd.bund
SPIELPLATZ GESCHLOSSEN
STATT REKORDAUSGABEN FÜR ASYL & BÜRGERGELD:
KOMMUNEN MÜSSEN WIEDER DIENSTLEISTER FÜR EIGENE BÜRGER SEIN!
ZEIT FÜR DEUTSCHLAND!
AfD
4,326 likes
4 days ago
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Lighthouse
Recorder
Preserve log
Disable cache
No throttling
Filter
Invert
More filters
All
Fetch/XHR
Doc
CSS
JS
Font
Img
Media
Manifest
WS
Wasm
Other
Name
Headers
Payload
Preview
Response
Initiator
Timing
Adblock
Cookies
graphql
query
Open in Sources panel
Open in new tab
Clear browser cache
Clear browser cookies
Copy
Block request URL
Block request domain
Replay XHR
Sort By
Header Options
Override headers
Override content
Show all overrides
Copy URL
Copy as cURL
Copy as PowerShell
Copy as fetch
Copy as fetch (Node.js)
Copy response
Copy stack trace
Copy all URLs
Copy all as cURL
Copy all as PowerShell
Copy all as fetch
Copy all as fetch (Node.js)
Copy all as HAR (sanitized)
"xdt_api__v1__feed__user_timeline_graphql_connection": {
"edges": [
"node": {
"code": "C7DxyrFNQZo",
"pk": "3369755932714403432",
"id": "3369755932714403432_1484534097",
"caption_is_edited": false,
"taken_at": 1715926482,
"video_versions": [
"is_dash_eligible": 1,
"number_of_qualities": 2,
"image_versions2": {
"candidates": [

```
import requests

cookies = {
    'datr': 'quu8Z4BY8Xtm8hnGltQ2EwJY',
    'csrftoken': 'PgJBHugGiGK8zth9uHItmGz65OzDoLov',
    'ig_did': '90763CE2-2BDC-4B44-A142-217346EE0F52',
    'mid': 'Z_ODdgAEAAEi2PE6pa8Sx2l9vwUi',
    'sessionid': '192593382%3ArqZ5oQ7YmfEd5P%3A6%3AAYeYDI3ZpQdOfEUwlMoLZtZECEx8G6hFeHphybzH0w',
    'ds_user_id': '192593382',
    'rur': '"FRC\\054192593382\\0541775548178:01f7ab0ca5ee20cfb838469cd56a7c796c342baab21299379f4a05de64bb0f6733ed3c5d"',
    'wd': '857x1045',
}

headers = {
    'accept': '*/*',
    'accept-language': 'en-US,en;q=0.8',
    'content-type': 'application/x-www-form-urlencoded',
    'origin': 'https://www.instagram.com',
    'priority': 'u=1, i',
    'referer': 'https://www.instagram.com/',
    'sec-ch-ua': '"Chromium";v="134", "Not:A-Brand";v="24", "Brave";v="134"',
    'sec-ch-ua-full-version-list': '"Chromium";v="134.0.0.0", "Not:A-Brand";v="24.0.0.0", "Brave";v="134.0.0.0"',
    'sec-ch-ua-mobile': '?0',
    'sec-ch-ua-model': '""',
    'sec-ch-ua-platform': '"macOS"',
    'sec-ch-ua-platform-version': '"14.6.1"',
    'sec-fetch-dest': 'empty',
    'sec-fetch-mode': 'cors',
    'sec-fetch-site': 'same-origin',
    'sec-gpc': '1',
    'user-agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/134.0.0.0 Safari/537.36',
    'x-asbd-id': '359341',
    'x-bloks-version-id': '0d99de0d13662a50e0958bcb112dd651f70dea02e1859073ab25f8f2a477de96',
    'x-csrftoken': 'PgJBHugGiGK8zth9uHItmGz65OzDoLov',
    'x-fb-friendly-name': 'PolarisProfilePostsQuery',
    'x-fb-lsd': 'BY3c_od_cnaQhTpnO6wdYr',
    'x-ig-app-id': '936619743392459',
    # 'cookie': 'datr=quu8Z4BY8Xtm8hnGltQ2EwJY; csrftoken=PgJBHugGiGK8zth9uHItmGz65OzDoLov; ig_did=90763CE2-2BDC-4B44-A142-217346EE0F52; mid=Z_ODdgAEAAEi2PE6pa8Sx2l9vwUi;
    sessionid=192593382%3ArqZ5oQ7YmfEd5P%3A6%3AAYeYDI3ZpQdOfEUwlMoLZtZECEx8G6hFeHphybzH0w; ds_user_id=192593382; rur="FRC\\054192593382\\0541775548178:01f7ab0ca5ee20cfb838469cd56a7c796c342baab21299379f4a05de64bb0f6733ed3c5d";
    wd=857x1045',
}

data = {
    'av': '17841400521463009',
    '__d': 'www',
    '__user': '0',
    '__a': '1',
    '__req': '1o',
    '__hs': '20185.HYP:instagram_web_pkg.2.1...1',
    'dpr': '2',
    '__ccg': 'GOOD',
    '__rev': '1021627382',
    '__s': 'io8b45:vxe68p:kwfz00',
    '__hsi': '7490475217897250447',
    '__dyn': '7xeUjG1mxu1syaxG4Vp41twpUnwgU7SbzEdF8aUco2qwJyEiw9-1DwUx60p-0LVE4W0qa321Rw8G11wBz81s8hwGxu786a3a1YwBgao6C0Mo2iyo7u3ifK0zEkxe2GewGw9a3614xm0zK5o4q3y261kx-0ma2-azqwt8d-2u2J08O321LwTwKG1pg2fwxyo6O1FwlA3a3zhA6bwIDyXx
    ui2qiUqwm8jxK2K0P8K9x6',
    '__csr': 'ggMlgssQaOhAJhi8yrsticzOBSPuYZlt4WOGjObCxXnrASApvF7HiGiJ69heT8nAOIF6t2aV96q8AJYBG6VkqEiihXnyXBDiGkxKaIBrFfAAAzmpUgpykiQiQF_BADyFfzkWBWBtpdp9pu8TUKV-LUK5C4d5GbyoF2F4p95ByV8DAxyqimUtiK4EB2qCXyXGczqCw04WHa0vKq7FE1Qp
    EhYE6W1L8ewAgCdgV7PCwoEronyXx21e8EdEy2a2lq
    8m6oze0mmzjG2q2ubJ05Tw30oS1_UKfU5q0gy09140Mw6Q4Ja37xKaEE6B1Sfg2ACoAj81kyV83Hw6ww4fw4xy9no4d122m7y3410wnCq0kV04mw5HvzFx4g1Awp49wHg2axe6Uo8q1uyk1LixC0TUdoaK5m5Q5UdEy01pDtAghxu3y04BU-4omwOx22m32lo0b9m084w',
    '__comet_req': '7',
    'fb_dtsg': 'NAcNt_Yo3qjwkvXeOoBbyshsWjxpv2-O3M7F4cb7sUCahl9VbvnUhYA:17865379441060568:1744012156',
    'jazoest': '26528',
    'lsd': 'BY3c_od_cnaQhTpnO6wdYr',
    '__spin_r': '1021627382',
    '__spin_b': 'trunk',
    '__spin_t': '1744012166',
    '__crn': 'comet.igweb.PolarisFeedRoute',
    'fb_api_caller_class': 'RelayModern',
    'fb_api_req_friendly_name': 'PolarisProfilePostsQuery',
    'variables': '{"data":{"count":12,"include_reel_media_seen_timestamp":true,"include_relationship_info":true,"latest_besties_reel_media":true,"latest_reel_media":true},"username":"afd.bund","__relay_internal__pv__PolarisIsL
    oggedInrelayprovider":true,"__relay_internal__pv__PolarisShareSheetV3relayprovider":true}',
    'server_timestamps': 'true',
    'doc_id': '9750506811647048',
}

response = requests.post('https://www.instagram.com/graphql/query', cookies=cookies, headers=headers, data=data)
```

APIS BUILD INFRASTRUCTURAL POWER

Anne Helmond (Utrecht University) argues that platforms offer APIs to extend their ecosystems beyond their own websites, effectively turning third-party websites and apps into extensions of the platform.

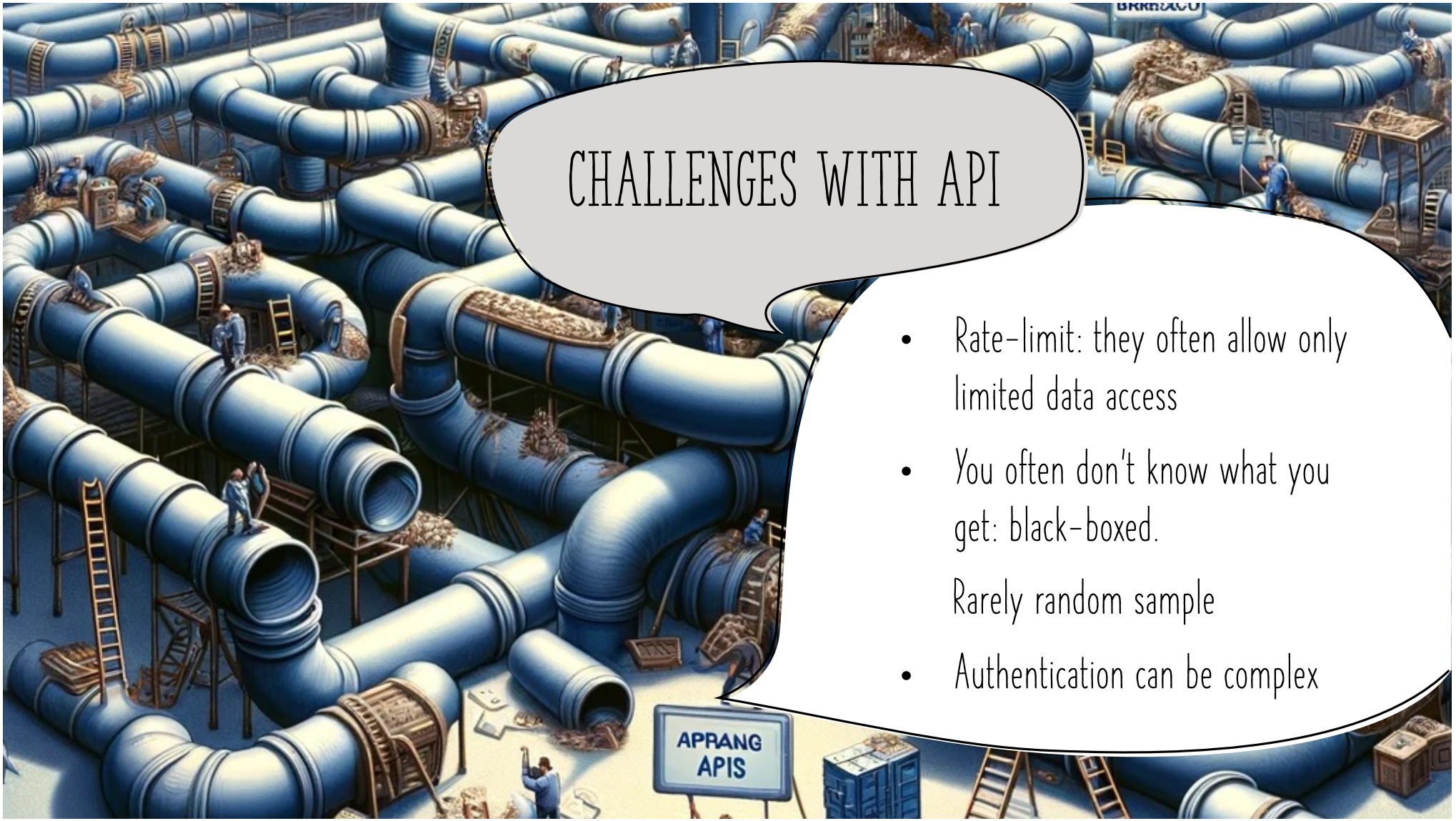
This dependency gives platforms a significant amount of control over the wider platform ecosystem.

Helmond, A. (2015). The platformization of the web: Making web data platform ready. Social media+ society, 1(2)

THE APICALYPSE

- Many APIs have been shutting down
- Instagram, Reddit, Twitter, Facebook...
- In part due to costs, privacy concerns, and in part to LLMs

- But DSA might come to the rescue



CHALLENGES WITH API

- Rate-limit: they often allow only limited data access
- You often don't know what you get: black-boxed.

  Rarely random sample
- Authentication can be complex

RESPONSIBLE DATA COLLECTION PRACTICES

- When scraping or collecting data from an APIs, don't collect too quickly since it can overburden the server. Pause between requests e.g., *time.sleep(0.5)*
- Don't collect data on people that you don't need (data minimization)
- Be mindful of privacy and ethical implications
- Consider anonymizing the data before storing (anonymization)

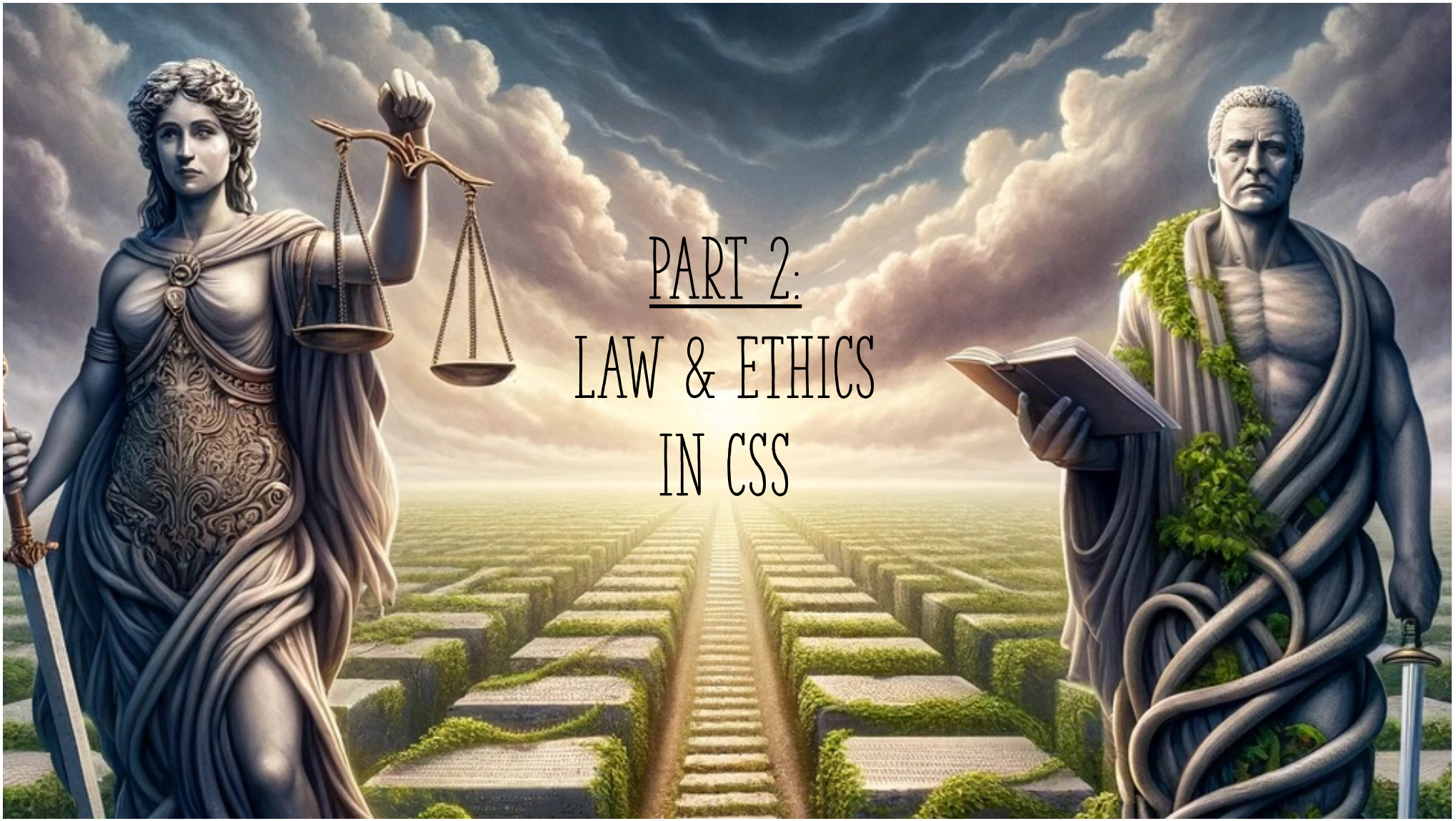
SHOULD YOU SCRAPE AS PART OF YOUR PROJECT?

If your project is about a new and untapped data source, and the novelty in your paper is centered on the data source:

Sure! Go for it!

else:

Probably not.


PART 2:
LAW & ETHICS
IN CSS

What's the relationship between Law and Ethics?

ETHICS != LAW

- There are legal things that are not ethical.
- There are ethical things that are not legal.

- But we usually want to try to be *both* ethical and legal.

Concerns over data privacy and various scandals have meant more regulation

LAW

1. Terms of Service
2. Copyright Infringement
3. EU Database Directive in the EU
4. GDPR

TERMS OF SERVICE & SCRAPING

Can websites contractually limit scraping in their terms of use? *Yes, they can.*

But are those provisions enforceable? *Well, that depends. How was the contract created?*

- Browsewrap agreement: contracts that were concluded simply by visiting a website. Legal theory generally does not accept agreements of this type as valid.
- Clickwrap agreements: require an action by the user. "By pressing Accept, you agree to our Terms and Conditions". Clickwrap agreements are perfectly fine and fair, and the courts will readily enforce them.

Basically: If you need to log in, create a user, or click "I agree", you *have* signed a ToS and need to follow it.

If not, you have not entered a legal agreement, and you can *go wild!*

APIS AND TERMS OF SERVICE

- APIs often have ToS that you need to follow
- (In)famously, Twitter/X does not allow you to share any collected data
- Be mindful of the ToS when using APIs!

Developer terms

Developer Agreement and Policy

X Developer Agreement

Last Updated: November 14, 2023

By clicking “Accept & Subscribe”, continuing to pay the recurring subscription fee for Paid Services, and by otherwise accessing or using any Licensed Material, you agree to the terms of our Developer Agreement. Subscriptions auto-renew until canceled, as described in the Terms. A verified phone number is required to subscribe. If you've subscribed on another platform, manage your subscription through that platform.

COPYRIGHT INFRINGEMENT

Scraping and reproducing content from a website could infringe on copyright laws.

In the EU, the scraping of copyrighted content is permitted (Article 3, 4 of Directive 2019/790 in the Digital Single Market DSM Directive).

- "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes, but is not limited to, patterns, trends and correlations";

Scraping of copyrighted content is only permitted for the purposes of generating information.

For example, you can scrape a webpage to extract data to analyze using natural language analysis.

But you *cannot* scrape news articles and then republish them on your own website.

For scientific research, you can freely scrape almost anything! But do not republish original work, and be mindful of sharing scraped data publicly.

EU DATABASE DIRECTIVE (DIRECTIVE 96/9/EC)

- Databases are protected - which can impact web scraping.
- The maker of a database the right to prevent extraction of a substantial part of the contents of a database. This means that scraping data from databases could potentially infringe on the rights of the database maker.

- But the directive offers exemptions for teaching and scientific research, provided that you cite your source and that the use is for non-commercial purposes.

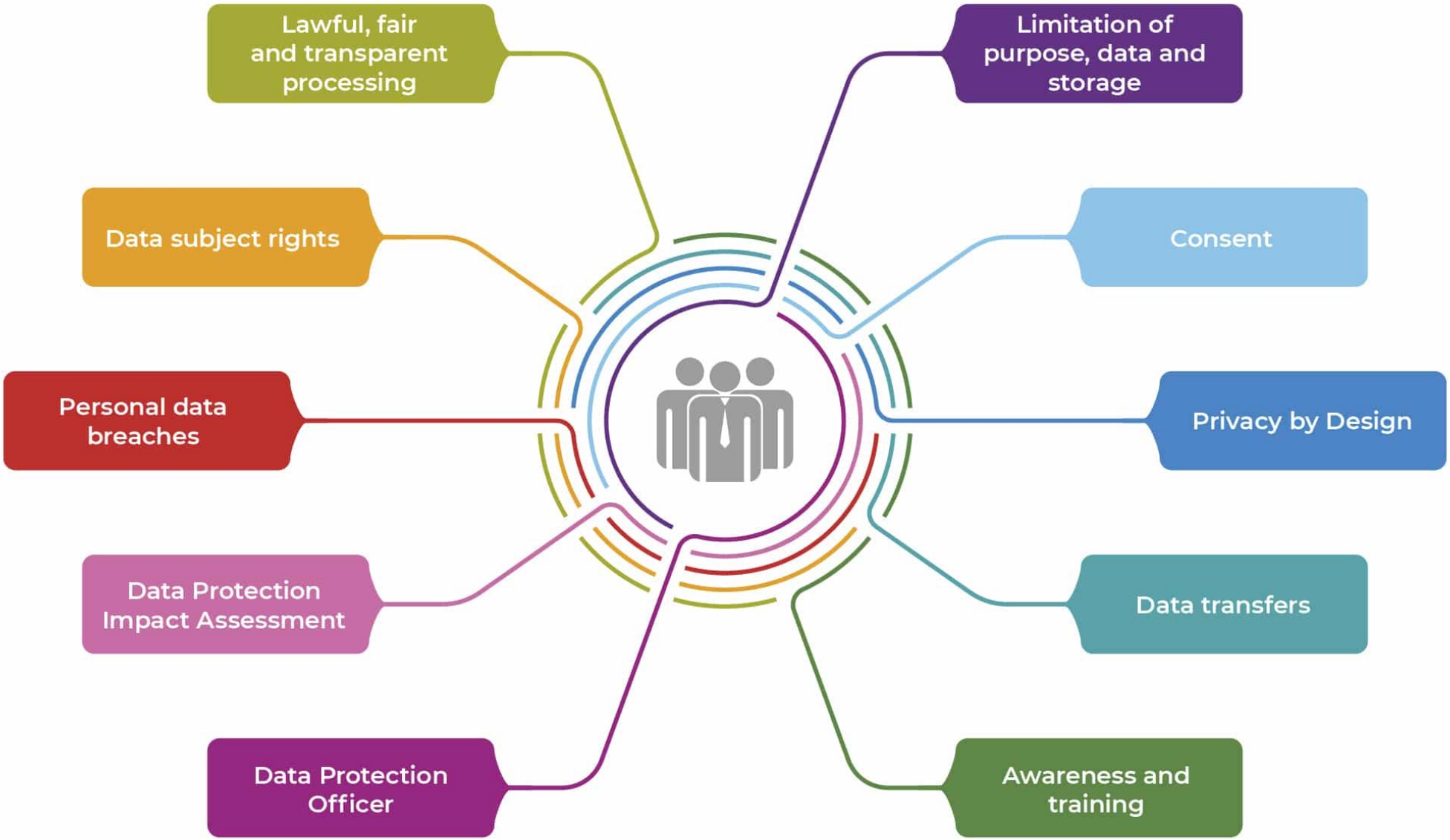
GDPR: GENERAL DATA PROTECTION REGULATION

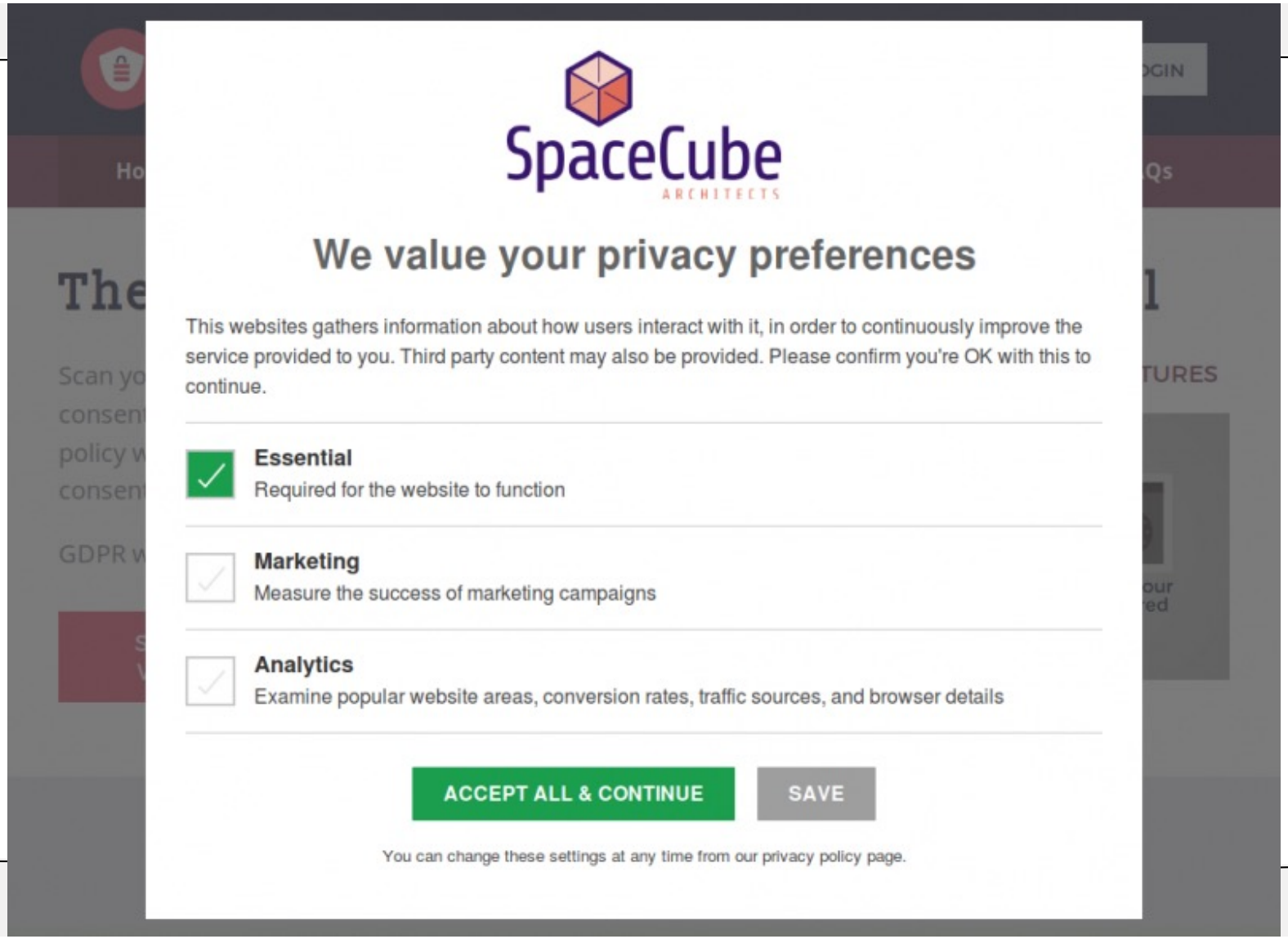
This is the big one. Implemented in 2018.

- The GDPR requires that data collection practices comply with principles like *lawfulness*, *fairness*, and *transparency*.

Researchers must ensure explicit consent for the collection and use of personal data, provide clear information about data usage, and implement robust data protection measures.

The GDPR grants individuals the right to access, correct, and delete their data.
Planning and need to justify the data collected and its use in research = procedures and paperwork!


Lawful, fair and transparent processing
Limitation of purpose, data and storage
Data subject rights
Consent
Personal data breaches
Privacy by Design
Data Protection Impact Assessment
Data transfers
Data Protection Officer
Awareness and training


SpaceCube
ARCHITECTS
We value your privacy preferences
This websites gathers information about how users interact with it, in order to continuously improve the service provided to you. Third party content may also be provided. Please confirm you're OK with this to continue.
Essential
Required for the website to function
Marketing
Measure the success of marketing campaigns
Analytics
Examine popular website areas, conversion rates, traffic sources, and browser details
ACCEPT ALL & CONTINUE
SAVE
You can change these settings at any time from our privacy policy page.

WHAT IS PERSONAL DATA ANYWAY?

GDPR only applies to personal data: "Personal data means any information relating to an identified or identifiable natural person"

If the data is not possible to link to a person, it does not fall under GDPR.

Direct identifying personal data refers to data which contains for example the name of the data subject, identity numbers, telephone numbers, email addresses, postal addresses or bank account numbers, etc.

Indirect identifying personal data refers to data that can be traced back to an individual when combined with other information. Since these data can be traced back to an individual (even though not in a direct way), the data are personal data and should be treated as such.

E.g.. *Name, surname, date of birth, address, social security number, passport number, national ID number, employment information, Contact details, phone number, email address, IP address, Facebook, Twitter, and other network handles, location either by address or GPS, shopping preferences, behavioral data, Video + audio recordings of people and biometric data. Special categories of personal data: sex, gender, and sexual orientation, racial or ethnic origin, religious beliefs, political opinions, medical records*

DE-IDENTIFICATION BY ANONYMIZATION / PSEUDONYMIZATION

If research data are anonymised/pseudonymised, to what extent do they still fall under the GDPR?

- Anonymisation is the process of removing all the direct and indirect information that can link the data to an individual.

After anonymisation, nobody can link the data to an individual anymore, i.e. no key files, pseudonyms, etc. are used. When data is fully anonymised, the data isn't personal data anymore and it doesn't fall in the scope of the GDPR.

- When the data is pseudonymised it is still possible to identify the person. Generally, a key file is used so that at least one person can link the data to an individual.

Even though most people cannot trace the data back the participant, the data is still personal data and therefore falls in the scope of the GDPR.

Re-identification is a constant risk!

DATA RE-IDENTIFICATION

"Anonymous" data isn't always anonymous.

Data can often be re-identified, even if personal identifiers like names or ID numbers are removed, by using other available information (like birth dates, ZIP codes, or purchase history).

Example:
A dataset has anonymous health records with age, ZIP code, and hospital visit dates. Another public dataset (like voter registration) has names, ages, and ZIP codes. Matching the age and ZIP code between the two can reveal identities.

True anonymization is hard. If data can *possibly* be linked back to a person, it falls under GDPR protection.

So, is the text of a post from a social media platform personal data?

PRINCIPLES OF GDPR

Data minimization: only data necessary should be processed, short storage period, limited accessibility

Privacy by design: any actions involving the processing of personal data is done with data protection and privacy in mind.

Privacy by default: all technical and organisational measures are taken to process data with the highest privacy protection.

As a researcher, you need to be proactive about data protection, and evidence (document) the steps you take to meet your obligations and protect people's rights.

GDPR PAPERWORK

1. Data management plan (DMP)
2. Privacy Protection Review (PPR)
3. Data Protection Impact Assessment (DPIA)
4. Processor agreement (PA)
5. Data exchange agreement (DEA)

DATA MANAGEMENT PLAN (DMP)

A Data Management Plan (DMP) is always required under GDPR.

- What data are you going to collect? What type of data or what file formats? How many?
- Where and how will you store your data? How will you provide back-ups?
- How are you going to organise and describe your data?
- Who gets access to your data? When? How are you going to manage access?
- What data will be archived when the project is finished? Where and for how long?
- Will the archived data be made available to others? When? Under what licence?
- Who owns the data? Who is responsible for the management of your data?
- Is there any funding to cover the costs of the implementation of the plan?

PRIVACY PROTECTION REVIEW (PPR)

PPRs are generally not mandated by law but are adopted by organizations as a best practice to ensure comprehensive privacy governance.

UvA requires this for any research projects.

- What type of data are you collecting?
- How will you store it? Anonymize? Share? Etc.

1. During your research AND the pre-research phase (i.e. screening, selection of participants), what (special) personal data relating to subjects are processed? *

For more information see infobox below.

- [ ] In this research no personal data are processed
- [ ] A. Name
- [ ] A. Contact details (e-mail, telephone number, or home address)
- [ ] A. IP-addresses
- [ ] A. Student number
- [ ] A. Citizen service number (BSN)
- [ ] A. Facial images/video and/or voice recordings
- [ ] A. Financial details (account number or creditcard number)
- [ ] A. Copies of passport or other identity documents
- [ ] A. Location data (GPS)
- [ ] A. Username
- [x] B. Gender
- [x] B. Age in years
- [ ] B. Date of Birth
- [ ] B. Nationality
- [ ] B. City or area of residence/postal code
- [ ] B. Unique identifier (e.g. number that can be used to re-identify a research participant)
- [ ] B. Number plates and device numbers (e.g. car, mobile phone IMSI)
- [ ] B. Language Background
- [ ] C. Personal data concerning convictions, criminal offenses or relevant safety precautions
- [ ] C. Personal data that point to ethnic background
- [x] C. Personal data that point to political views
- [ ] C. Personal data that point to religious or philosophical beliefs
- [ ] C. Physical or mental health details
- [ ] C. Data on sexual behaviour or orientation
- [ ] C. Genetic details
- [ ] C. Biometric data for the purpose of uniquely identifying a person
- [ ] C. Trade union
- [ ] C. (increased risk of) Dyslexia and/or language disorders
- [ ] Other (please specify)

DATA PROTECTION IMPACT ASSESSMENT (DPIA)

Required by GDPR for high-risk projects:

- If you're using new technologies (e.g., AI)
- If you're tracking people's location or behavior
- If you're systematically monitoring a publicly accessible place on a large scale
- If you're processing personal data related to "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation"
- If your data processing is used to make automated decisions about people that could have legal (or similarly significant) effects
- If you're processing children's data
- If the data you're processing could result in physical harm to the data subjects if it is leaked

Processor agreement (PA)

You draw up a processor agreement when you ask a third party to process personal data for your research, for example to transcribe audio files or to deliver an instrument for online surveys. This third party solely carries out your requests and does not take any decisions about the data themselves.

Data exchange agreement (DEA)

Sometimes you collaborate with other organisations or universities on a research project and you collectively decide how personal data will be collected, for what purpose and how they will be processed. In that case a data exchange agreement is required to establish who is responsible for what. If you process personal data, it needs to be clear whom a participant can contact with a remark or complaint.

EXAMPLE PROJECT: "TWITTER MOOD PREDICTS THE STOCK MARKET"

by Johan Bollen, Huina Mao, and Xiao-Jun Zeng in 2011.

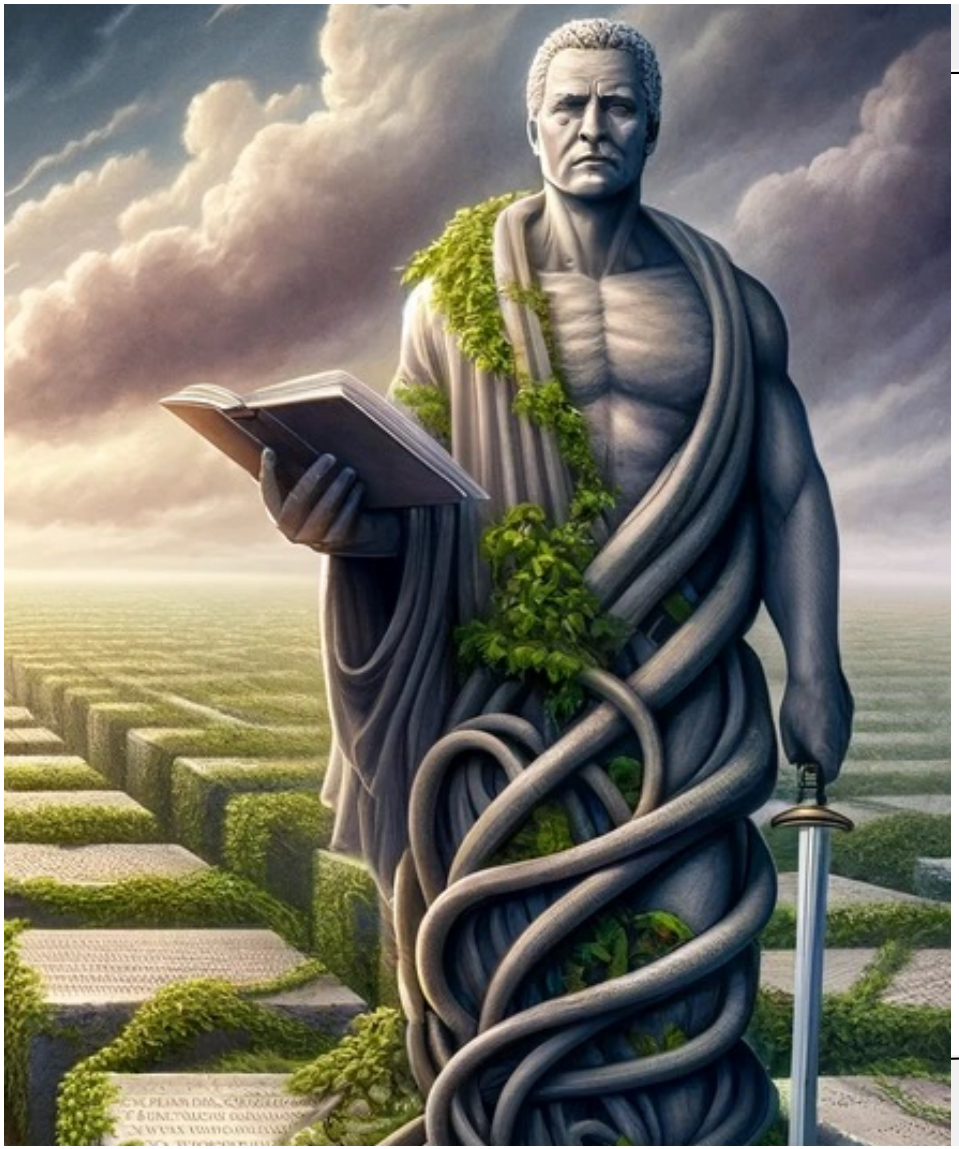
This study utilized Twitter data to analyze the collective mood states of users and examined how these mood states could predict the rise and fall of the Dow Jones Industrial Average.

Data Analysis: The researchers applied text analysis techniques to Twitter feeds to extract mood states using a psychometric instrument known as the Profile of Mood States (POMS). They then correlated these mood states with stock market movements.

Findings: The study found a significant correlation between certain mood states captured through Twitter and the Dow Jones Industrial Average. Leading to the conclusion that social media could forecast economic indicators.

TASK: DISCUSS THROUGH LEGAL CONSIDERATIONS OF THE MOOD STUDY

- What do you need to do to fulfill legal requirements for the study?
- What data should you collect? How store?
- What paperwork do you need for GDPR compliance?



ETHICS

1. Ethical frameworks
2. Ethical pluralism
3. AoIR guidelines
4. ERB application
5. Ethics section in paper

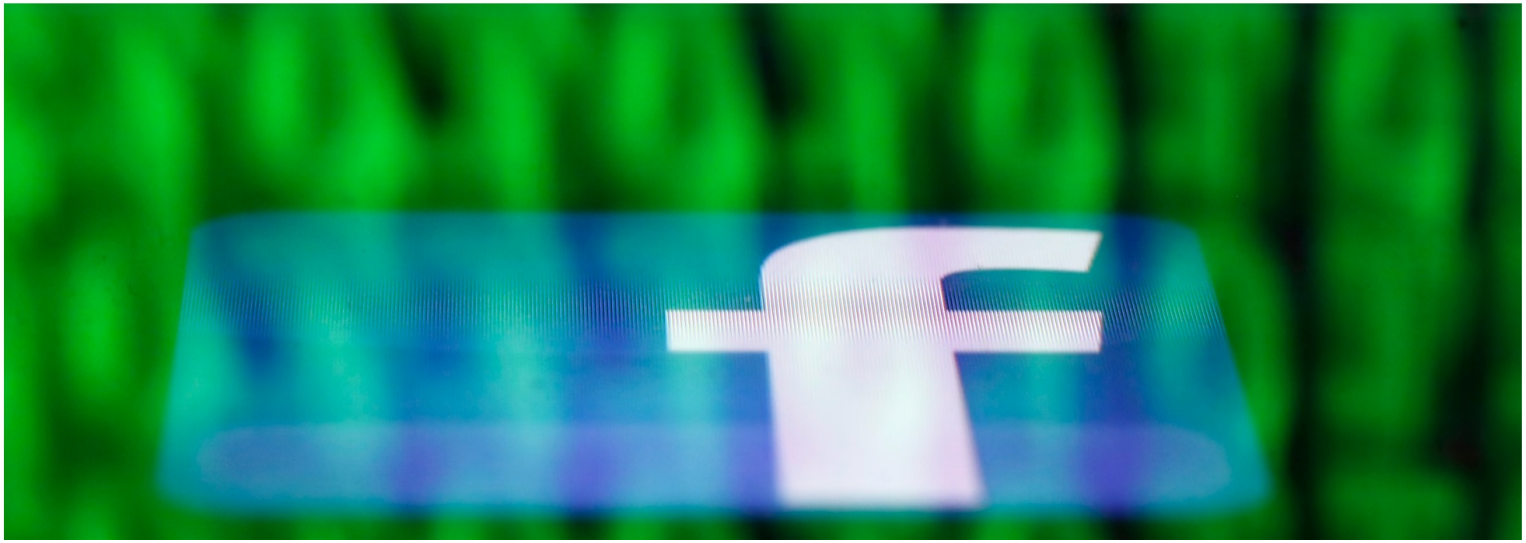
TECHNOLOGY

Everything We Know About Facebook's Secret Mood-Manipulation Experiment

It was probably legal. But was it ethical?

By Robinson Meyer

- 700,000 users were exposed either to negative posts or positive posts
- Measured effects on their own posts
- People got sadder
- Published in PNAS



RESEARCH ETHICS IN CSS

- Social scientists have long had IRB: come from the tradition of experiments.

  Stem from highly unethical experiments in the 1960s (e.g., Stanford Prison Experiment)
- Computer science has little tradition of minding ethics.

CSS is in the process of debating ethics. Still not well-established.

Varies depending on institution and region.

WHAT IS ETHICAL ANYWAY?

We can analyze our ethical dilemmas and challenges through a range of ethical frameworks:

Deontological: the rightness or wrongness of actions does not depend on their consequences but rather on whether they fulfill our duty or follow certain moral rules. Rule-based. Strong in Europe and Scandinavia.

Consequentialism: The outcomes and consequences of actions are what matter. (UK and US).

Utilitarianism: calculate the greater good for the collective and society in general.

Virtue-ethics: Actions are right if they are what a virtuous person would do in the same context.

Feminist ethics: Critically examine and address ethical theories and practices that perpetuate gender inequalities

Ethics of care: Prioritizes maintaining and fostering relationships, empathy, and caring for others, particularly those dependent on or vulnerable to one's actions.

So which one do we use?

DISCUSSION QUESTION: HARMING NAZIS

We generally want to avoid harming our research subjects.

But when studying Stormfront, the question was how fast to collect the data:

fast scraping means getting data in weeks instead of months, but may slow down their website.

But their website fuels neo-Nazi violence and has been linked to 100s of murders.

If we're harming self-proclaimed Nazis, does it make it more okay?

What would different ethical frameworks say?

ETHICAL PLURALISM AND CROSS-CULTURAL AWARENESS

Ethical pluralism: see the issue from multiple frameworks and discuss their competing views.

Accept ambiguity, uncertainty, and disagreement as inevitable.

Cross-cultural awareness: What is the expectations of the person being studied? What ethical framework do they follow?

If you're studying a distant culture, you must understand their views and expectations on ethics.

Research ethics is rarely black and white: there are competing interests that must be weighed

ETHICS GUIDELINES

Key Principles

- Respect for Persons: Emphasizes the importance of respecting the autonomy, privacy, and dignity of individuals and communities involved in or affected by the research.
- Beneficence: Encourages actions that contribute to the welfare of research participants, aiming to maximize benefits and minimize harm.
- Justice: Ensures fair treatment and equitable distribution of the benefits and burdens of research across different groups, avoiding exploitation.

AIOR QUESTIONS TO GUIDE ETHICS

Discussion point: use these questions to think through the Twitter mood project!

1. Have participants been informed and consented? For publicly available data, have you considered the context in which the data was shared and whether the subjects anticipated their data being used for research? What are their expectations of privacy?
2. How does the project ensure the anonymity and privacy of individuals whose data is being analyzed? Are there measures in place to anonymize the data, especially when working with potentially sensitive information? Have you assessed the potential risks of re-identification of anonymized data?
3. Does the research aim to do good, and has it been designed to minimize potential harm to participants and communities involved?
4. How will the results be used, and could the publication of the research findings potentially harm individuals or communities? Have you considered the broader social implications of the research findings and how they contribute to the public good?
5. What measures are in place to secure the data during and after the research process?
6. Does the research involve groups or communities that might be considered vulnerable? If so, what additional protections are in place to safeguard these participants?
7. Are the research methods and ethical considerations clearly documented and accessible for review?

WRITING ETHICS REVIEW APPLICATION - ERB/IRB

- Ethics reviews are often mandatory for CSS research projects
- Offers opportunity to think through the ethics of your research
- The ERB is a floor, not a ceiling.

Task: discuss with your neighbor and go through the ethics review form, thinking of the Mood Stock Market project

WRITING AN ETHICS SECTION IN A PAPER

- Your paper should have an ethics section! Usually under "Methods & data"
- Describe your ethics process. Show that you've considered ethical implications.
- *Make an argument* for your work being ethical based on an existing framework or guidelines.
- *Very brief!*

> We follow the ethical guidelines for Internet research provided by the Association of Internet Researchers (Franzke et al., 2020) and by the British Sociological Association (BSA, 2017). To ensure anonymity for the members of the community, usernames were deleted when collecting the corpus. Since the analytical focus is on broader discursive patterns, it is not possible to identify individual users from the results here presented. The project has been assessed and approved by the Regional Ethical Review Board.

BEFORE WEDNESDAY'S WORKSHOP

Get API keys for Perspective API and YouTube API!

https://perspectiveapi.com/

https://developers.google.com/youtube/v3/getting-started