



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



COLLECTING DATA FROM THE WEB: FOUNDATIONS OF WEB SCRAPING

Dr Aybuke Atalay & Ponrawee Prasertsom



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

Workshop Overview

Session 1 (code-free) [30 October 2025]

Legal and Ethical Issues in Web Scraping
Using Selector Gadget
Exploring HTML Structures

Session 2 (R & Python) [6 November 2025]

Scraping data from a static website using R /
Python



www.cdcs.ed.ac.uk



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

PROGRAMME OF TODAY

14:00– 15:00:

- Overview of UoE Ethics Policies
- Web Scraping basics and Terms of Services
- 'Decision Tree' for Data Collection
- Data privacy, data ethics and the law
- Exercises

15:00 – 16:00:

- Web basics: How the web actually works (GET request, response, status code, etc.)
- HTML, CSS basics: HTML tags, CSS property/values
- Selector gadgets
- Exercises



www.cdcs.ed.ac.uk

STATIC & DYNAMIC WEBPAGES

Static pages:

- Typically small, displaying a limited amount of content
- Look the same for all users
- Built in scripting languages (e.g. HTML, CSS, JavaScript)
- Blogs, GitHub Pages sites, CV page on personal website

Dynamic pages:

- Can change based on users' data, device, and behaviour
- Contain interactive content
- Built in content management systems using a combination of scripting and server-side languages
- Vulnerable to more security risks
- Amazon, Netflix, BBC News



WHY SCRAPE THE WEB?

To collect data—social media, public records, government data

To expand, update, or complete datasets

To examine public discourse about a topic (comments, replies)

To analyse the relationship between online and offline behaviour



HOW TO SCRAPE THE WEB?

1) Application Programme Interface (API)

- Request data from sites on their own terms.
- Standard approach for social media pre-2023, but has become more complicated


2) Scraping and Crawling HTML/XML

- Generate a list of URLs from which to extract information, selecting relevant sections embedded in HTML, and downloading content

Other approaches: Optical Character Recognition (OCR), private agreement, purchase



IF YOU WANT TO LEARN ABOUT API...


 THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

[Home](#) / [Events](#) / Using API for Research

USING API FOR RESEARCH

26 NOV 2025, 14:00 – 16:00 3 DEC 2025, 14:00 – 16:00

[BOOK NOW](#)



www.cdcs.ed.ac.uk

WHEN TO SCRAPE THE WEB

There are multiple considerations into the decision of scraping:

1. Is content copyrighted?
2. Does scraping infringe IP rights?
3. Are there 'terms of use' or 'terms of services' for the platform
 - Is the use non-commercial?
 - If so, is the website hosted in the UK?*

*A 2014 UK law and 2019 EU directive gives exemption for non-commercial research



TERMS OF SERVICE (TOS)

Also called 'Terms of Use'

The 'fine print' users agree to when accessing a particular platform or site

Most people accept ToS without reading them

Violating ToS can lead to a range of consequences:

- rate limiting
- Account deletion
- Possible legal action
- For researchers, a breach of ethics



IMPORTANT CONSIDERATIONS ABOUT 'LEGALITY'

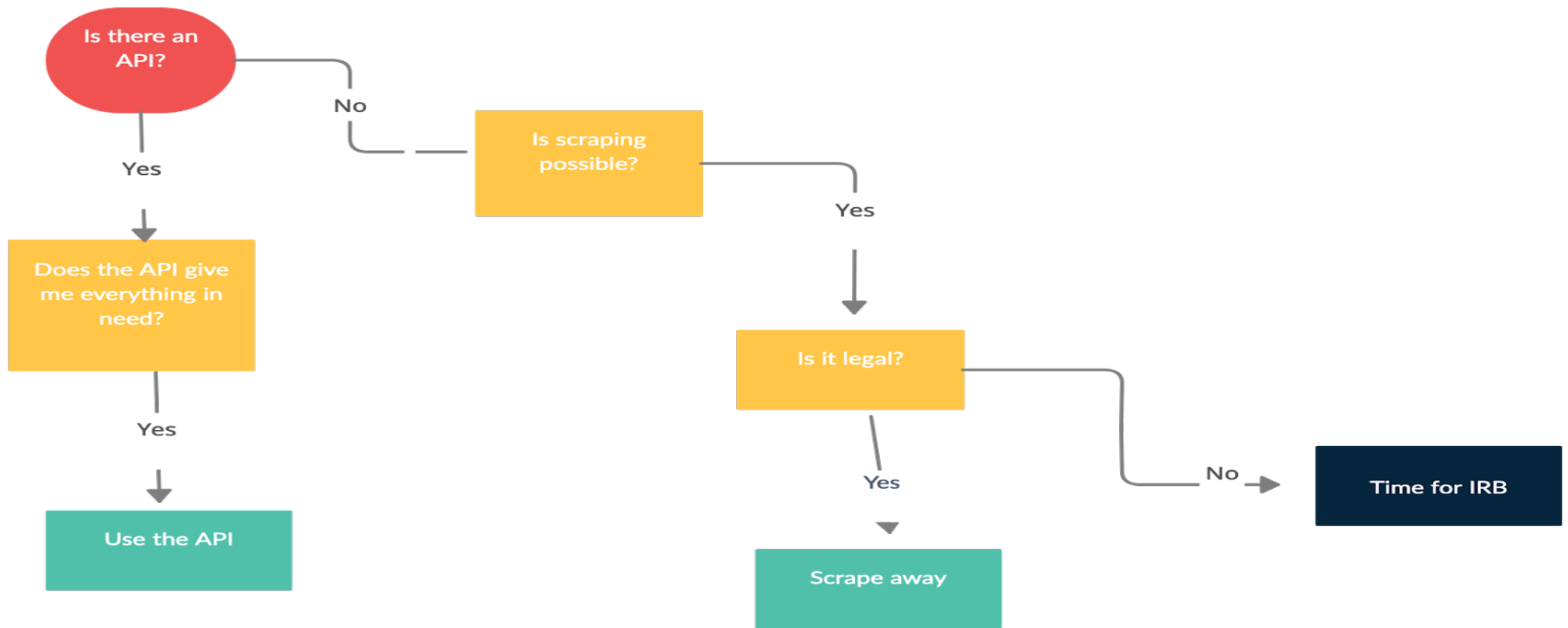
Existing policies have gray areas and are incomplete for some use cases
Best practices differ by country, discipline, sector, and interpretation of laws and policies.

Illegal web scraping:

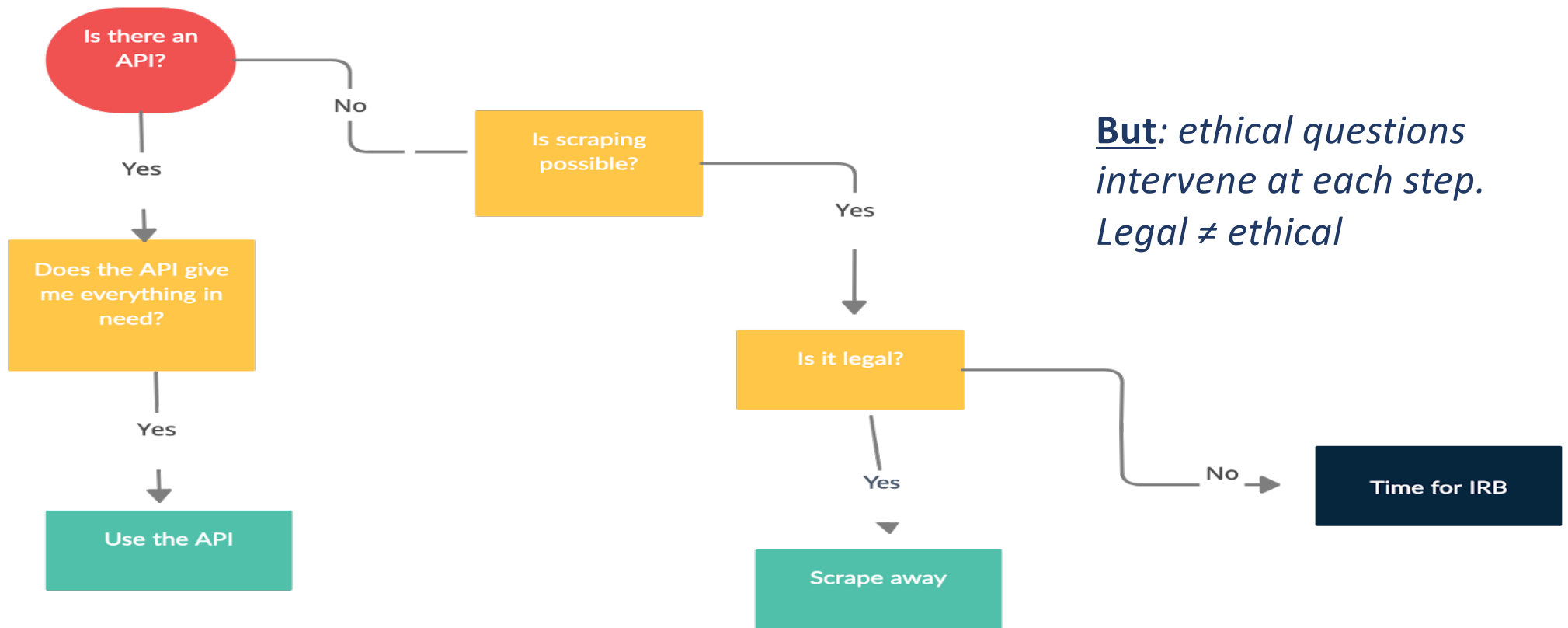
- Collects sensitive or private data
- Violates copyright or intellectual property law
- Extracts personally identifying information
- Sells data collected through webscraping



THE 'DECISION' TREE



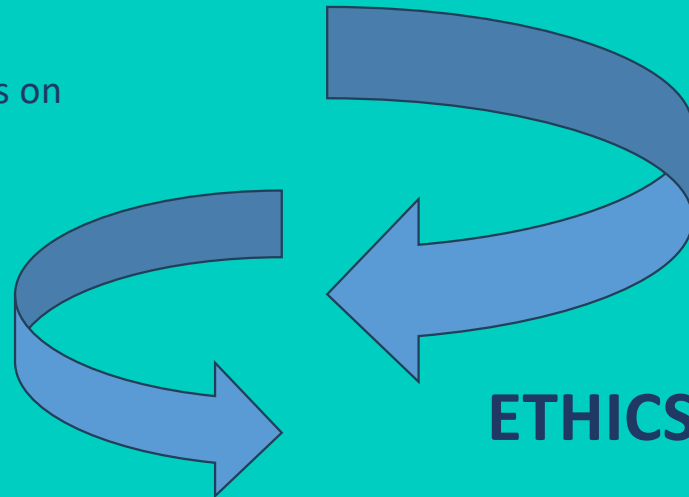
THE 'DECISION' TREE



LEGAL VS ETHICAL

LAW

Legal constraints placed by platforms on accessing content
E.g., not to use scrapers/crawlers



ETHICS

Ethical protection of user privacy/violations of contextual integrity

E.g., not to use scrapers/crawlers

Protection of minors/vulnerable groups



HOW TO SCRAPE 'ETHICALLY'

University Research Ethics Policy

Safeguard the interests and well-being' of all involved with or impacted by a research project

Ethical research principles:

- Beneficence and non-maleficence
- Integrity, openness and transparency
- Dignity and respect
- Responsibility and accountability
- Equality, diversity and inclusion

Must be integrated into all stages of a project's design, including its impact

"All research carried out by members of the School is are subject to ethical review. The ethical review process is designed to support researchers in managing risks associated with their research, ensuring the highest professional standards in designing, conducting and disseminating research." (SPS)



www.cdcs.ed.ac.uk

HOW TO SCRAPE 'ETHICALLY'

University Research Misconduct Policy

Research misconduct includes:

- ✓ Misuse of personal data
- ✓ Lack of informed consent from participants
- ✓ Breach of confidentiality
- ✓ Noncompliance with legal and ethical requirements
- ✓ Breach of duty of care
 - Disclosing participants' identity
 - Exposing personal or sensitive data
 - Improper conduct



HOW TO SCRAPE 'ETHICALLY'

Political Communication, 35:665–668, 2018
Copyright © 2018 Taylor & Francis Group, LLC
ISSN: 1058-4609 print / 1091-7675 online
DOI: <https://doi.org/10.1080/10584609.2018.1477506>



The Forum

Computational Research in the Post-API Age

DEEN FREELON

Keywords API, computational, Facebook, Twitter, social media

“By employing TOS- compliant methods, you are respecting the business prerogatives of the company that created the platform you are studying, but you may or may not be respecting the dignity and privacy of the platform’s users”



www.cdcs.ed.ac.uk

HOW TO SCRAPE 'ETHICALLY'



“I don’t think researchers should not be automatically bound by such terms-of-service agreements. Ideally, if researchers violate terms- of-service agreements, they should explain their decision openly... as suggested by transparency-based accountability. But this openness may expose researchers to added legal risk; in the United States, for example, the Computer Fraud and Abuse Act may make it illegal to violate terms-of-service agreements...”

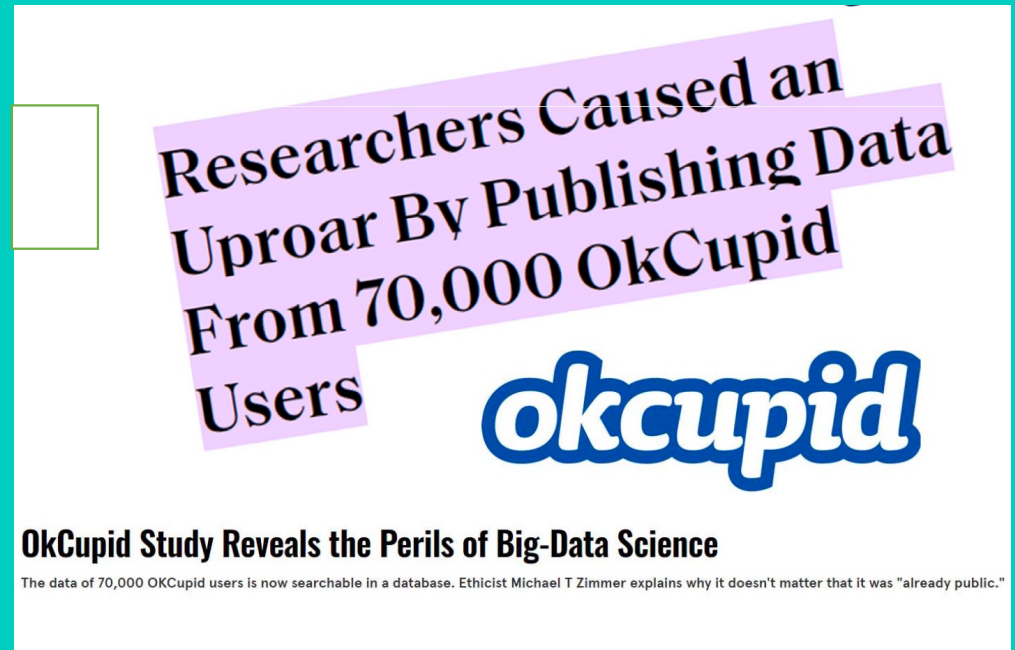


www.cdcs.ed.ac.uk

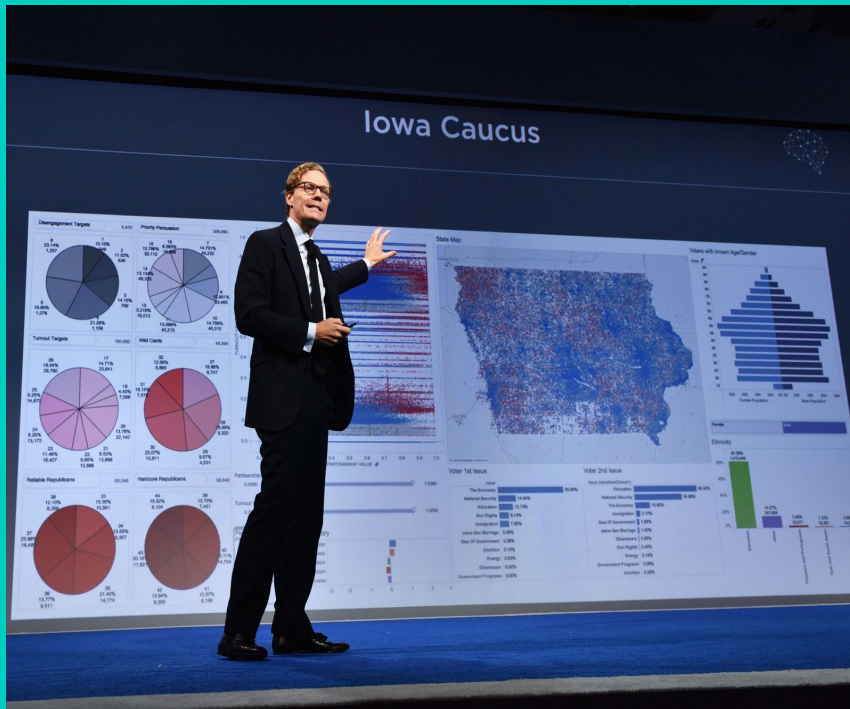
THE DANISH OKCUPID STUDY

In 2016, Danish researchers analysing the dating site OkCupid shared data on the Open Science Framework

- Their dataset included users' personal information such as their names, ages, gender, and responses to questions
- Open science values include transparency and sharing to ensure reproducibility in research design
- However, open research is not always ethical (and vice-versa)



THE CAMBRIDGE ANALYTICA SCANDAL



- Prior to 2016 U.S. Elections, a consulting firm called Cambridge Analytica collected personal data from 87 million Facebook profiles without users' knowledge or consent
- Created psychological profiles of users based on aspects of their personal data such as their location, gender, interests, and age
- Political campaigns in the US used the data to show targeted ads based on these profiles
- In 2018, a whistleblower reported the data breach
- Investigation led to Facebook restricting its API and third-party data access



EXERCISE 1: UNDERSTANDING TOS

Choose 1-2 platforms from the list below and read the ToS:

- 1- WhatsApp
- 2- GoodReads
- 3- Amazon
- 4- Twitter/X
- 5- Wikipedia

ToS can be notoriously difficult to read. Give it a go on your own, but if you're really stuck, the 'Terms of Service; Didn't Read' project is a good resource: <https://tosdr.org/>

Answer the following questions for the site of your choice

Data extraction—What are the platform's policies on web scraping? Does the platform have an API? If not, how can users acquire its data?

Data collection—What information can the platform collect from users, and what can they do with it? Who else can see users' data?

Data storage & deletion—How long does the platform keep personal data? If users delete their account, what happens?



EXERCISE 2: MAKING DATA ETHICS DECISIONS

Read each example and discuss whether the proposed method(s) comply with the principles of research & data ethics. If not, how might the methodology be improved?

1) A professor is examining the role of emotion in how readers form opinions about books. She plans to analyse book review data scraped from Goodreads using the platform's API.

2) A team of researchers are mapping the popularity of cycling in different areas of Edinburgh. They have received permission from the moderator of a private Facebook group for Edinburgh residents to extract all posts and comments contributed since 2019.

3) A lecturer has received UKRI funding to analyse Twitter posts about climate change. He has budgeted £25k for a six-month subscription to Twitter's API, which he will use to collect the data.

4) A postdoc is tracking health outcomes in people diagnosed with COVID-19. 1,000 patients have agreed to donate their health records, which the postdoc will anonymise and analyse at the aggregate level.

