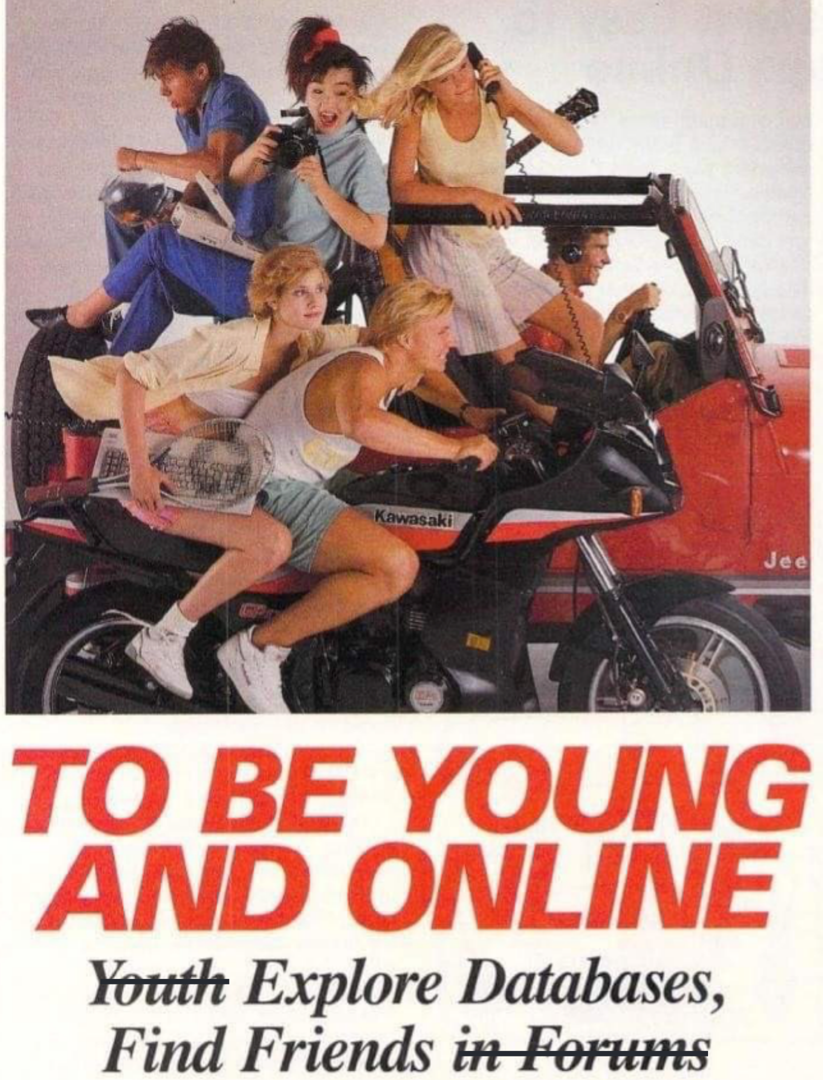
UNIVERSITÄT
LEIPZIG

Forschungsseminar CSS – Scraping the web

GWZ H2 1.15, 28.10.2025

Felix Lennert, M.Sc.



OUTLINE

- Recap
- Scraping – what do we have to bear in mind?
    - Is this legal?
    - How do we protect people’s privacy?

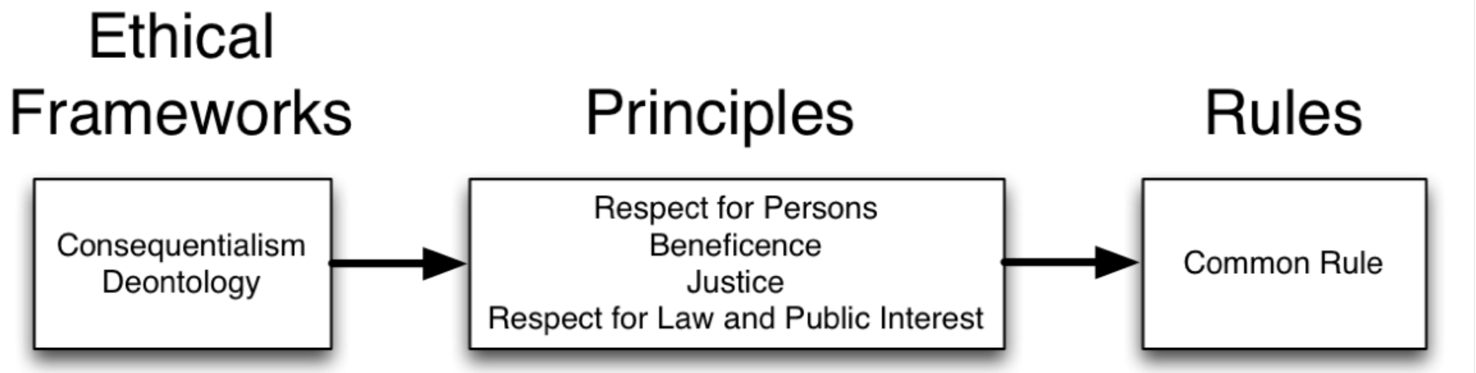
(Today’s slides are inspired by Étienne Ollion’s slides for SICSS-Paris 2023, find all the materials here)

SCRAPING

- Scraping is describing the process of acquiring data from the world wide web
- Multiple ways of doing this exist – in descending order from most to least favorable/ convenient:
  - Data dumps – companies give out significant chunks of their data (e.g., Pushshift for Reddit data)
  - APIs (Application Programing Interfaces) – companies provide you a *structured* way of getting their data (e.g., Spotify, New York Times, etc.)
  - Screen-scraping
    - We write a program to grab raw content
    - Sometimes: we write a program to simulate a browser *and then* grab raw content
- Overview of packages

ETHICAL ASPECTS

- THIS IS NOT LEGAL ADVICE
- In Europe, there is the GDPR (General Data Protection Regulation)
- In the US, there is the case of *hiQ v. LinkedIn* – bottom line: scraping publicly accessible data is legal
- Matt Salganik has something to say about this as well (chapter 6 of *Bit By Bit*)



Ethical Frameworks

Consequentialism
Deontology

Principles

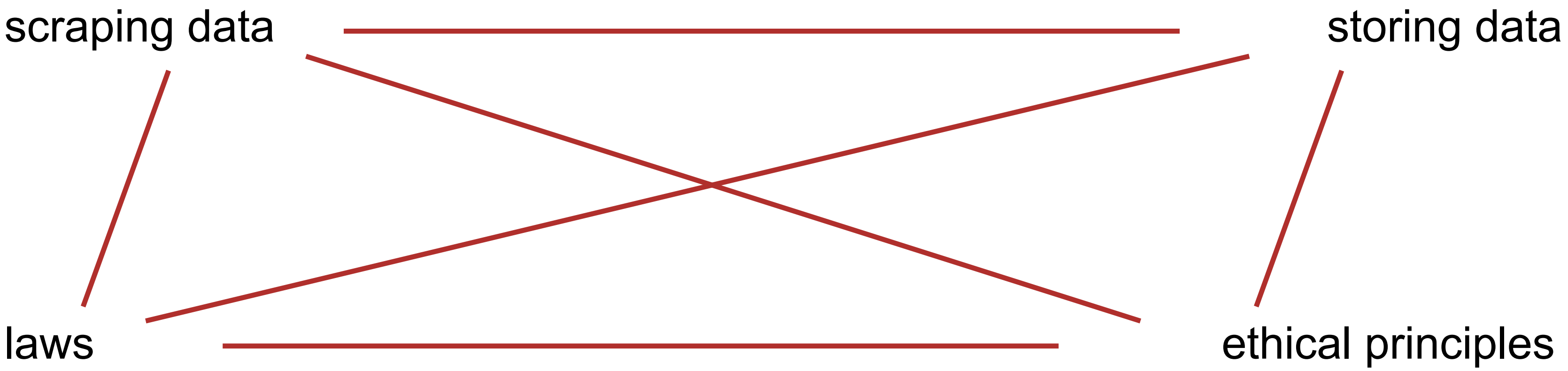
Respect for Persons
Beneficence
Justice
Respect for Law and Public Interest

Rules

Common Rule

ETHICAL ASPECTS

- Different aspects:



SCRAPING X LAWS

- In Europe: scraping a page is often illegal (see *robots.txt*, for instance)
    - due to: copyright, infringement of data base rights, etc.
- How is this enforced? – depends
    - websites might block users that are too active (i.e., robotic behavior)
    - websites might do nothing
    - websites might offer an API with only “legal” data
    - or they might go after you…
- However: we have an exception for research (directive)

SCRAPING X LAWS

- Things to bear in mind – be nice
    - is there an API?
    - maybe introduce yourself in user agent
    - add pauses to not overwhelm the servers
    - download only once, test code thoroughly to ensure this

SCRAPING X LAWS

- We have an exception for research (directive)
  - we can access everything that we are allowed to (of course)
  - some conditions for things that might not really be allowed but that we can access anyways:
    - things we purchased
    - things we subscribe to (e.g., newspaper archives the university pays for)
    - content that's available on the internet (and this is where I realize I'm not an expert)
- Important: data protection is key

SCRAPING X LAWS

- We have an exception for research (directive)
- Important: data protection is key
  - store it securely
  - don’t pass it on to third parties

DATA STORAGE X LAWS (GDPR)

Personal data: is any information that can lead (directly or with other information) to the identification of a living natural person (GDPR 4, Recital 27).

- Name, e-mail address, IP-address
- username or handle – and everything (tweets, forum posts, etc.) connected to this username!
- Images (of a face) and voice recordings
- Contextual information that can lead to identifying a person: ”I am a woman, a sociologist, working at place X. I am also originally from Italy, and…”

⟹ can be processed for “specified, explicit and legitimate purposes” (GDPR 5)

DATA STORAGE X LAWS (GDPR)

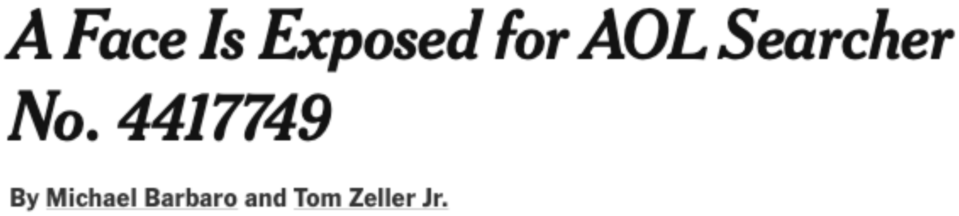
Identifying information are also:

- Racial or ethnic origin
- Political opinions and religious or philosophical beliefs
- Trade union membership
- Health, sex life, and sexual orientation
- Genetic and biometric data that can lead to the identification of a living person (but not images in general).

⟹ Processing personal data is allowed “if…”, processing special categories is prohibited “unless…”

What could go wrong – an example:

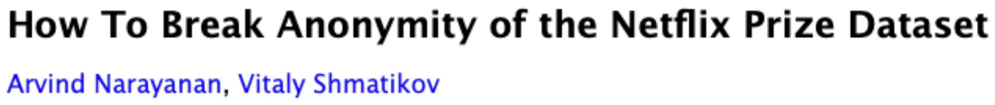
AOL releases 657,000 users’ search queries – did not anonymize the data properly – NYT reporters were able to link queries to people



A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

“And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”
It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her three dogs.” (Barbaro/Zeller 2006)

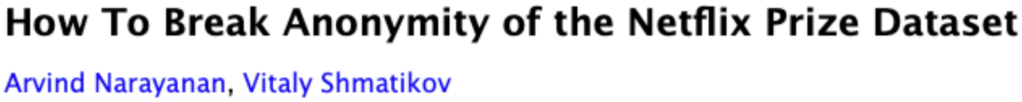


How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

“First, we can immediately find his political orientation based on his strong opinions about ‘Power and Terror: Noam Chomsky in Our Times’ and 'Fahrenheit 9/11.’ Strong guesses about his religious views can be made based on his ratings on ‘Jesus of Nazareth’ and ‘The Gospel of John.’ He did not like ‘Super Size Me’ at all; perhaps this implies something about his physical size? Both items that we found with predominantly gay themes, ‘Bent’ and ‘Queer as folk’ were rated one star out of five. He is a cultish follower of ‘Mystery Science Theater 3000.’ This is far from all we found about this one person, but having made our point, we will spare the reader further lurid details.” (Narayanan/ Shmatikov 2008: 16)

Scraping the web | Law and ethics – how not to



How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

They had information on a user’s movie ratings on Netflix and the date they rated the movie

- they matched it with the IMDb, looking for users who rated the same movie within a range of two weeks from the rating on Netflix in quite the same manner
- political orientation, religious views, etc. can then be extracted from their view on related movies

DATA STORAGE X LAWS (GDPR)

Processing personal data is allowed “if…”:

- Consent
- Fulfilling a contract or legal obligation
- Protecting the vital interests of the data subject
- **Carrying out a task in the public interest** *(this is us, we are doing this)*
- The legitimate interests of the controller (e.g., landlord)

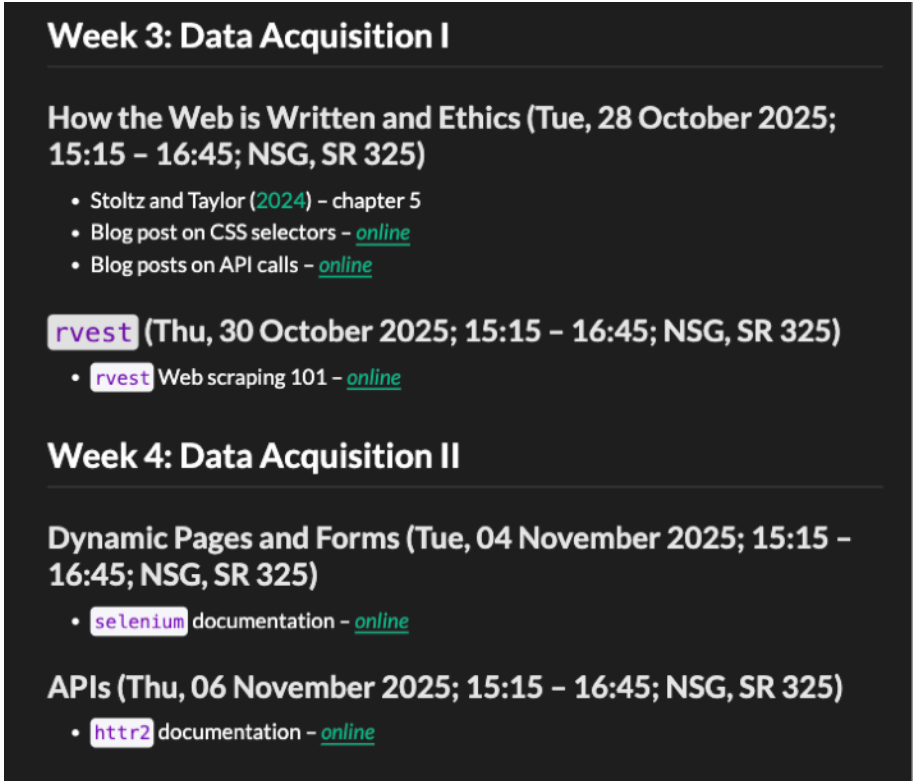
FFS, WHAT CAN WE DO THEN? (ETHICAL PRINCIPLES)

THIS IS NOT LEGAL ADVICE, more like a best practice

- Store data adequately (encrypt your hard drive, no Google Drive, etc.)
- Protect names/anonymize properly (erase names, use for instance a lookup table)

- Good question to ask oneself: “am I affecting my subjects; can anyone find them based on my data”
- Data being public is not an adequate excuse – no informed consent
- Data release: replicability is important, but so is data protection

NEXT



Week 3: Data Acquisition I

How the Web is Written and Ethics (Tue, 28 October 2025; 15:15 – 16:45; NSG, SR 325)

- Stoltz and Taylor (2024) – chapter 5
- Blog post on CSS selectors – *online*
- Blog posts on API calls – *online*

`rvest` (Thu, 30 October 2025; 15:15 – 16:45; NSG, SR 325)

- `rvest` Web scraping 101 – *online*

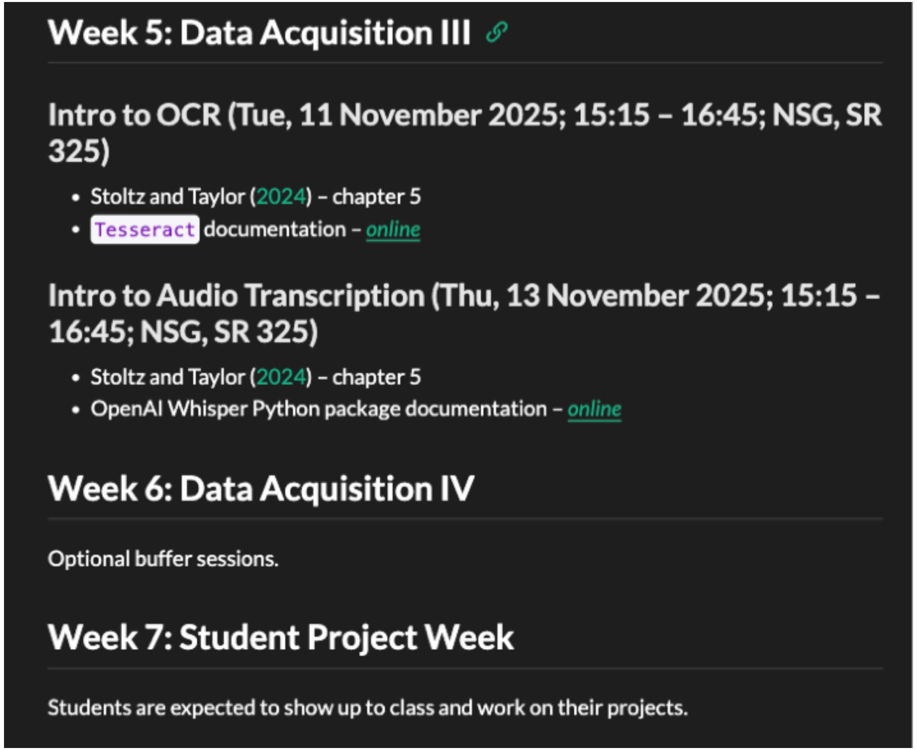
Week 4: Data Acquisition II

Dynamic Pages and Forms (Tue, 04 November 2025; 15:15 – 16:45; NSG, SR 325)

- `selenium` documentation – *online*

APIs (Thu, 06 November 2025; 15:15 – 16:45; NSG, SR 325)

- `httr2` documentation – *online*



Week 5: Data Acquisition III

Intro to OCR (Tue, 11 November 2025; 15:15 – 16:45; NSG, SR 325)

- Stoltz and Taylor (2024) – chapter 5
- `Tesseract` documentation – *online*

Intro to Audio Transcription (Thu, 13 November 2025; 15:15 – 16:45; NSG, SR 325)

- Stoltz and Taylor (2024) – chapter 5
- OpenAI Whisper Python package documentation – *online*

Week 6: Data Acquisition IV

Optional buffer sessions.

Week 7: Student Project Week

Students are expected to show up to class and work on their projects.

UNIVERSITÄT
LEIPZIG

MERCI!

Felix Lennert
Institut für Soziologie

felix.lennert@uni-leipzig.de
www.uni-leipzig.de