Webscraping in FinEco Research: Risks and Opportunities

Cynthia Huang

Econometrics & Business Statistics

Introduction

About Me

- 👩‍🎓 PhD Candidate in Monash EBS, affiliated with
  - NUMBATS (EBS), SoDa Labs (ECON), EmVis (Faculty of IT)
  - Monash Data Futures Institute
- Researching principles and methods for "alternative" data
  - conceptual and practical data provenance tools for harmonised multi-source datasets
  - ***adapting web-scraped retail product & price data for public health research***
  - Statistical properties of alternative data, grammar of graphics
- Open and reproducible research tools, research software design
  - Quarto, git, replication packages
  - NUMBATS Hacky hour, community building

Objectives

🤔 Research opportunities & risks

- Big data, novel data, alternative data
- Operational vs. scientific risk
- Web technologies and web-scraping methods
- Ethical & legal risk

❌ How to code a web-scraper

❌ How to code big-data analysis

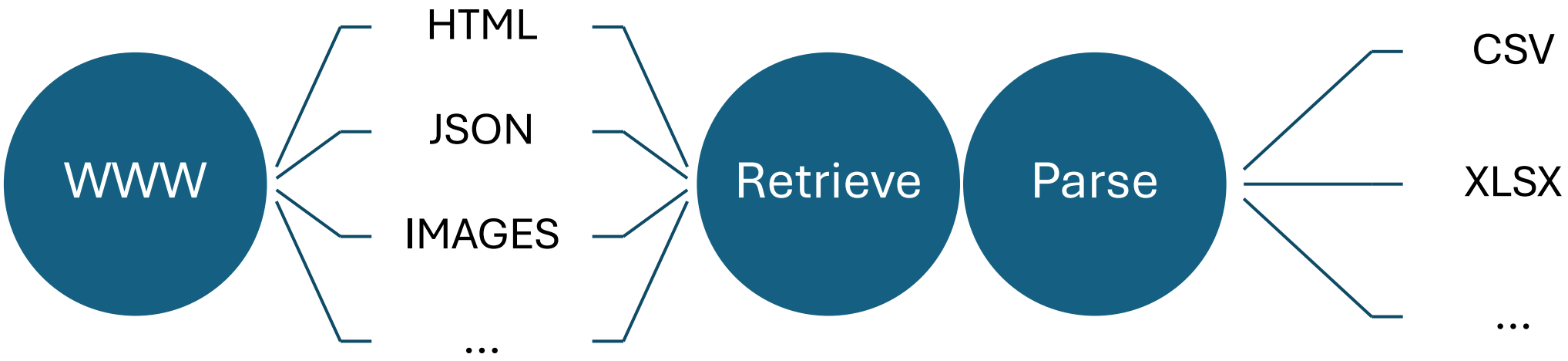
✅ Project level considerations

✅ Resourcing web-scraping projects (skill sets, RA hires etc.)

Web scraping—*verb*.

Cambridge dictionary:

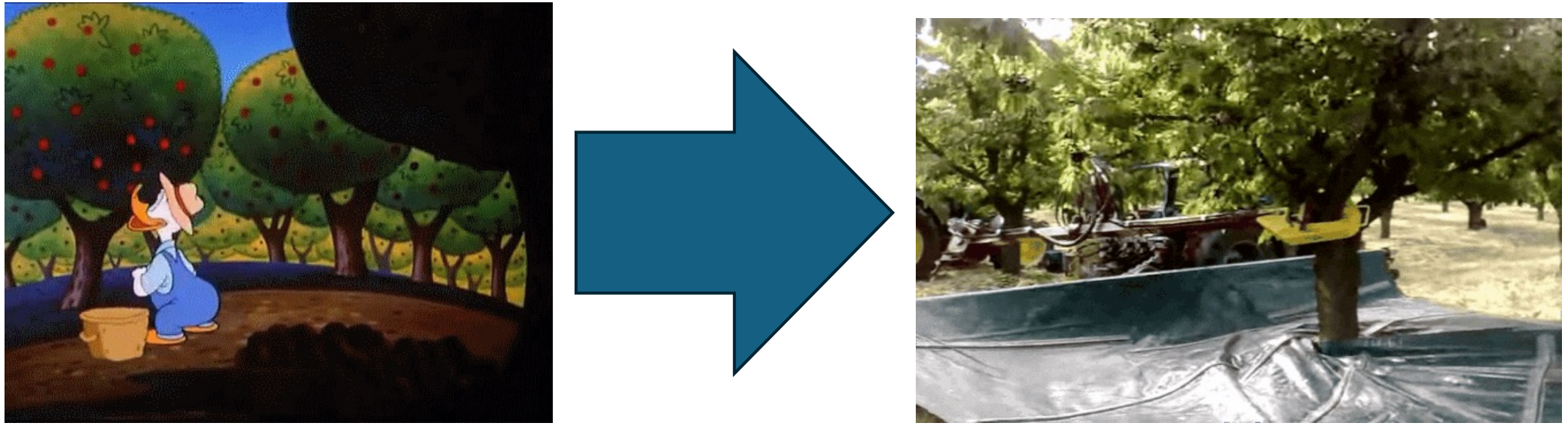
> *the activity of **taking** information from a website or computer screen and **putting** it into an ordered document on a computer*



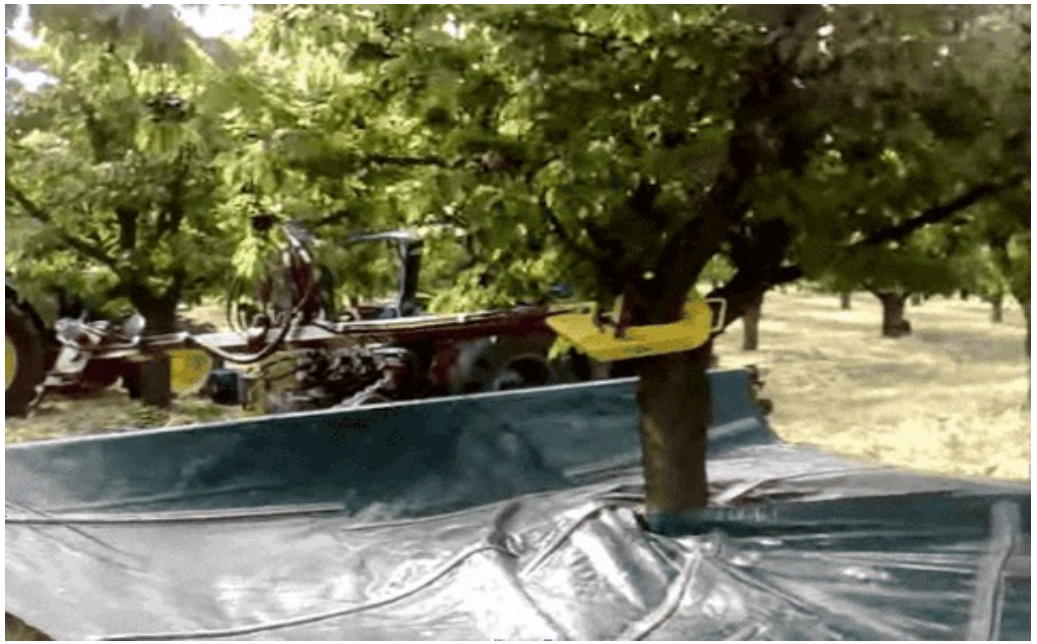
Web scraped data—*noun*.

Cambridge dictionary:

the activity of taking ***information from a website or computer screen*** *and putting it into an ordered document on a computer*



Common properties of web scraped data



- Some degree of messiness
  - Collection, preparation
  - Analysis, statistical properties
- Non-trivial parsing from raw web extraction to analysis-ready data
- Murky data dimensions/quality
  - How many records?
  - How many observational units?
  - Complete/missing observations?

Research Opportunities & Risks

Big, novel, alternative (data) opportunities

- The internet is a rapidly expanding universe of **(collectable*) data**
  - Multiple modalities: text, images, video…
  - Multi-source harmonization: shopping aggregators, knowledge graphs
  - Novel signal sources: social media sentiment, digital economy indicators
- **Lower* cost** than traditionally available data collection methods
  - Digital vs. Analogue access and delivery
  - Automated data extraction vs. Manual data entry
- Volume, Velocity, Variety… **Veracity?**
  - Noise vs. signal for data mining, prediction, casual inference
  - Collecting data from a Complex System vs. Library

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

Research Risk: Traditional Data

Operational Risk

- Availability and Access
- Cost
- Legal and ethical

Scientific Risk

- Suitable analysis methods
- Signal-to-noise ratio (in context)
- Data quality

Research Risk: Web Scraped Data

Operational Risk

- **Cost = Time**
- Availability and access
  - Sampling bias
  - Unstable web interfaces
- Legal and ethical
  - Usage terms

Scientific Risk

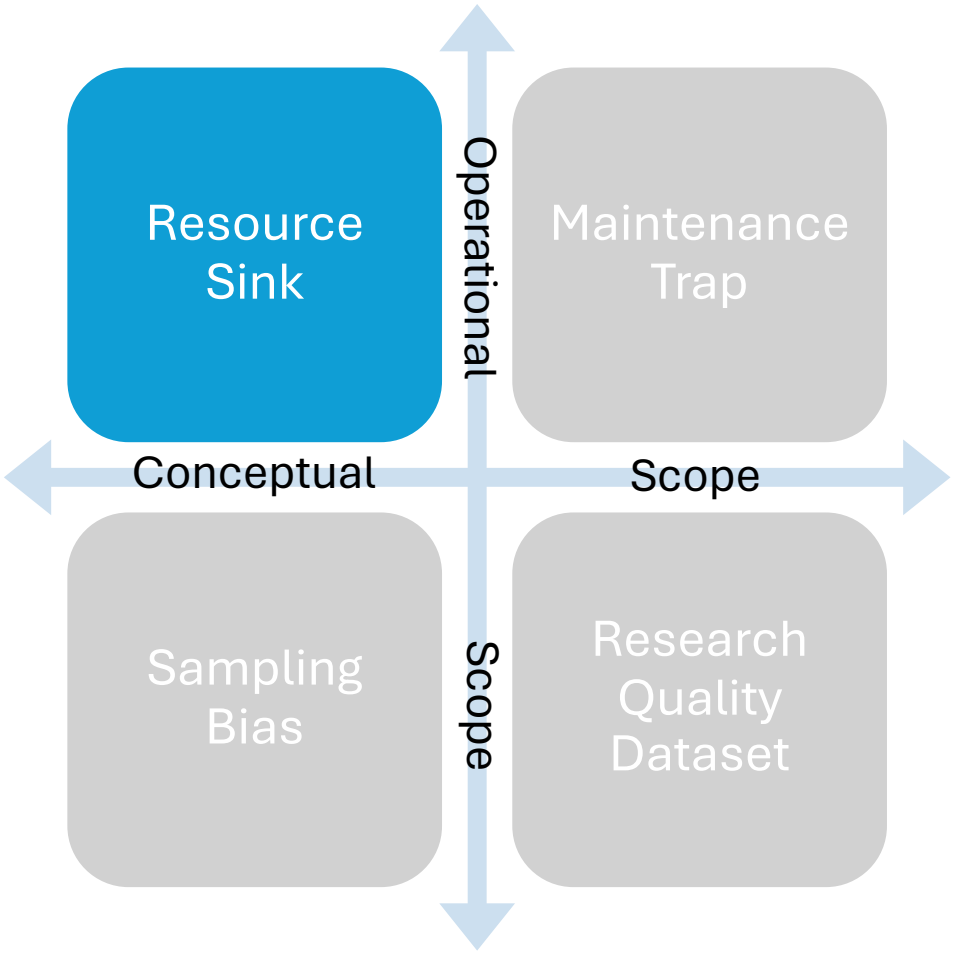
- **Data quality**
  - Completeness
  - Accuracy
- **Signal-to-noise ratio**
  - Big Data Paradox (Bradley et al., 2021)
- Suitable analysis methods

Data Collection Scope

Conceptual Scope

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

Web Scraped Data Risk



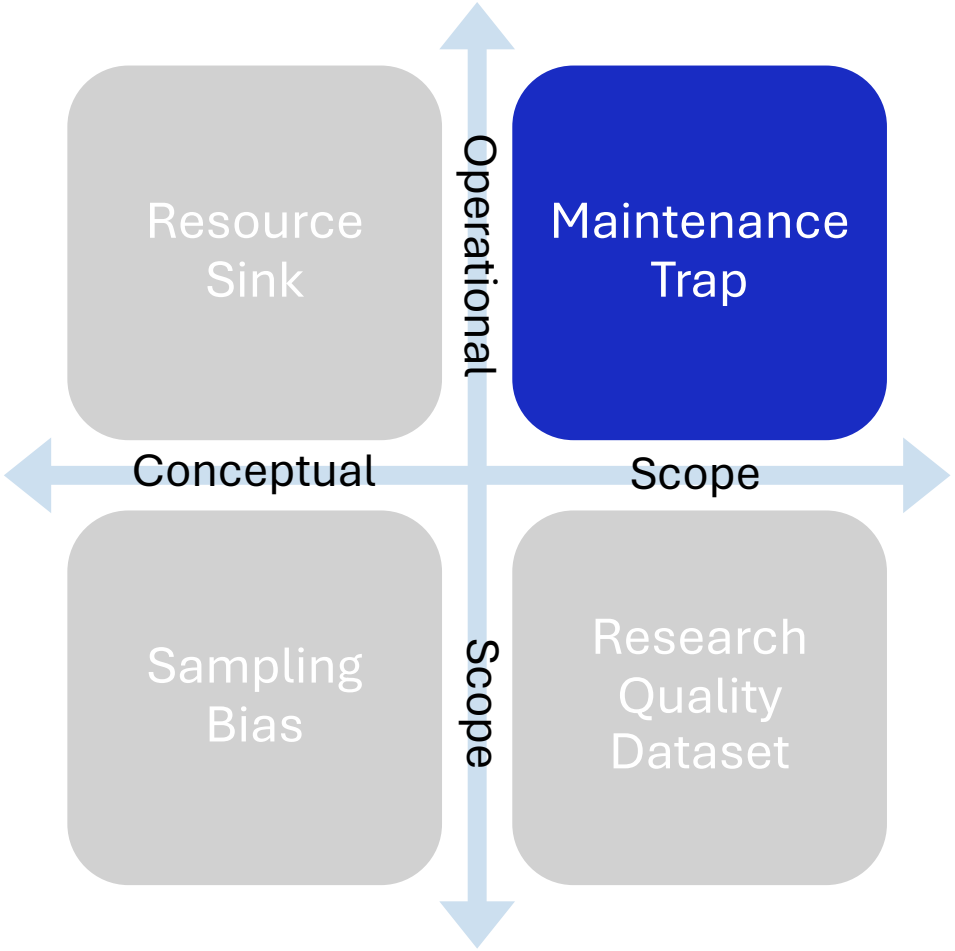
Resource Sink

Example dataset description:

Daily stock prices

- Conceptually and operationally open-ended
- Unclear dataset dimensions
  - Which firms?
  - What time period?

Web Scraped Data Risk



Maintenance Trap

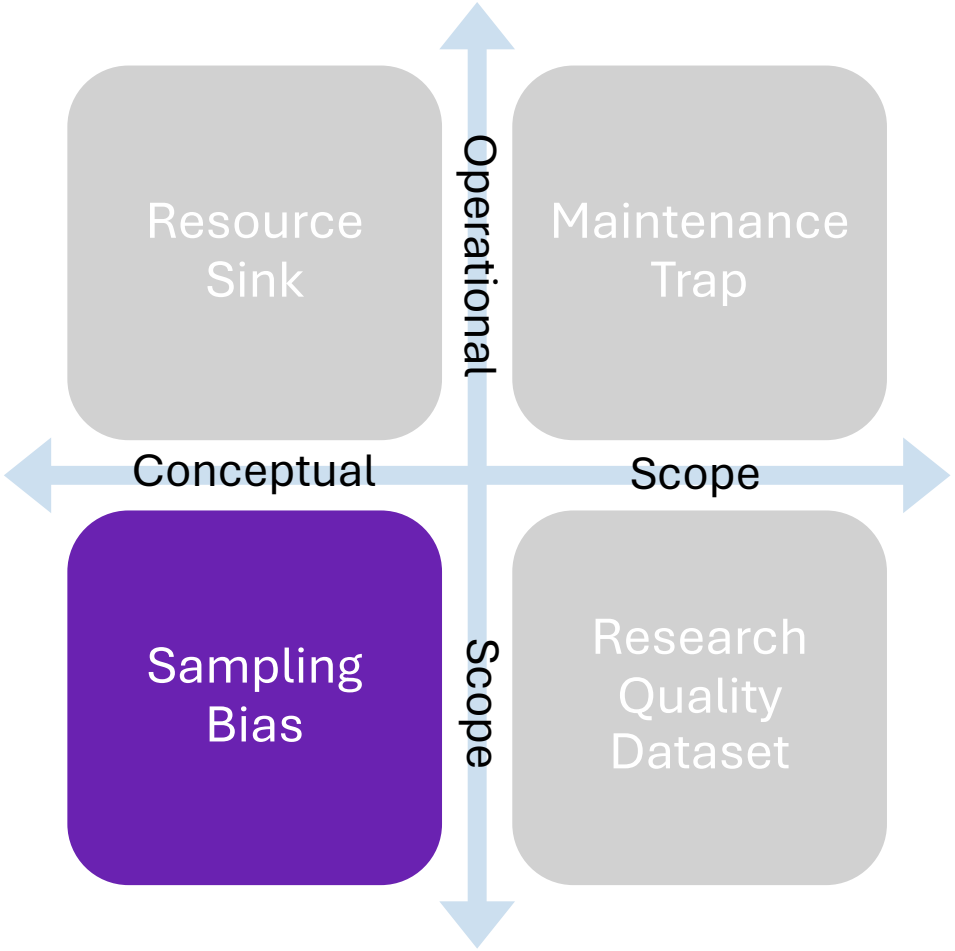
Example dataset description:

Daily prices for alcoholic beverages sold online by retailers with physical stores in Australia

- Conceptually limited, but operationally open-ended
- implies ***ongoing*** collection:
  - Repair "broken" scrapers
  - Upgrade data storage

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives, 30(2), 151–178. https://doi.org/10.1257/jep.30.2.151

Web Scraped Data Risk



Sampling bias

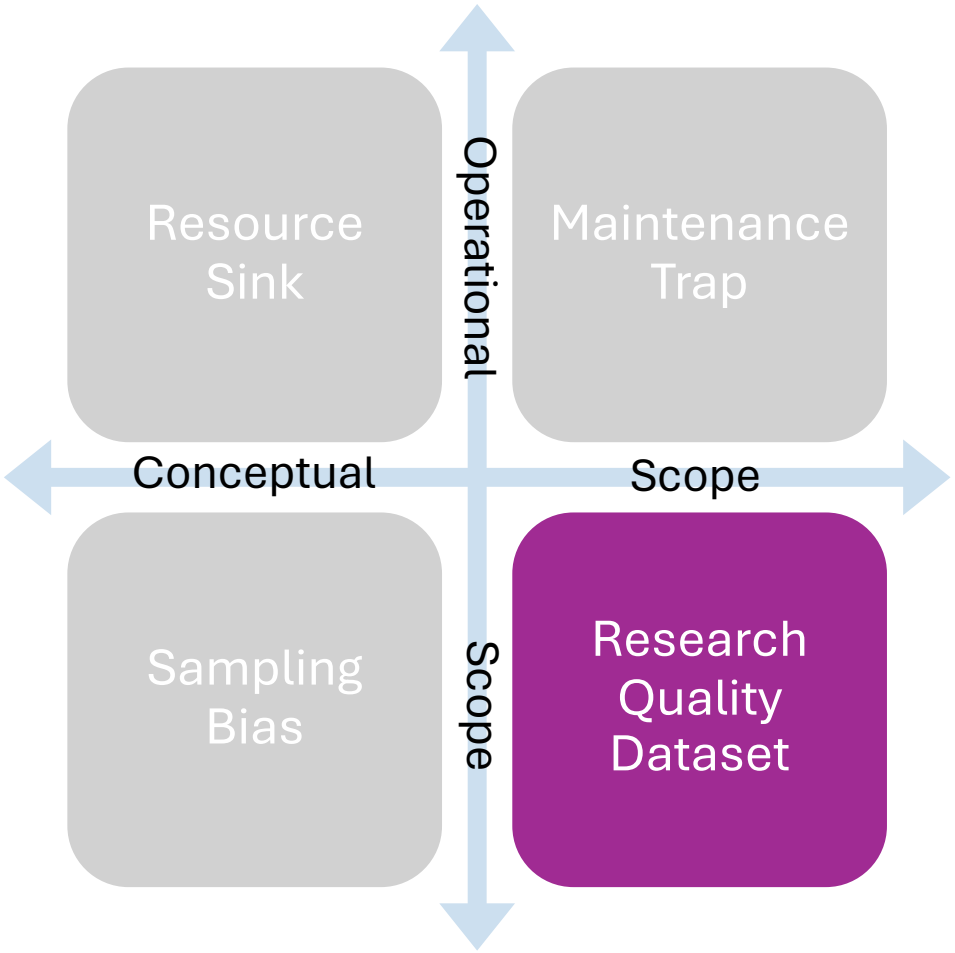
Example dataset description:

Daily prices for alcoholic beverages available online between Jan-Dec 2023

- Operationally limited, but conceptually open-ended
- What population is being sampled from?
  - Index coverage

Foerderer, J. (2023). *Should we trust web-scraped data?* (arXiv:2308.02231). arXiv. http://arxiv.org/abs/2308.02231

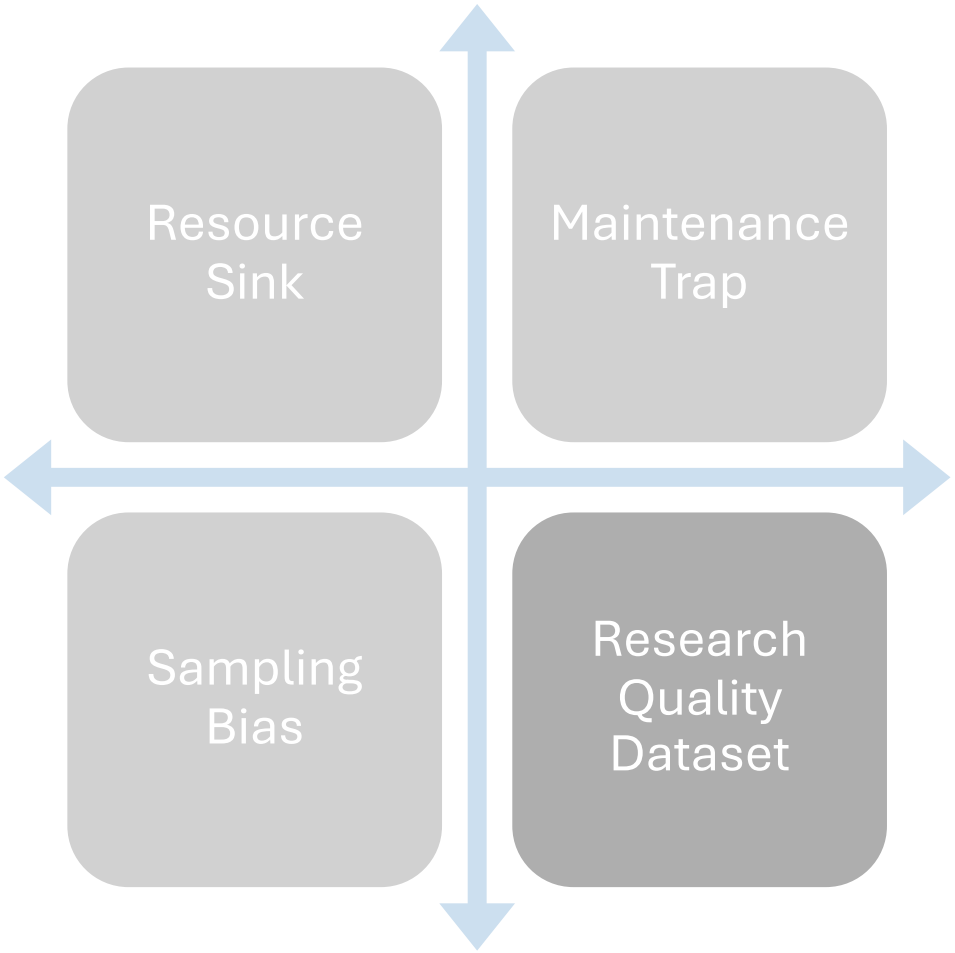
Web Scraped Data Risk

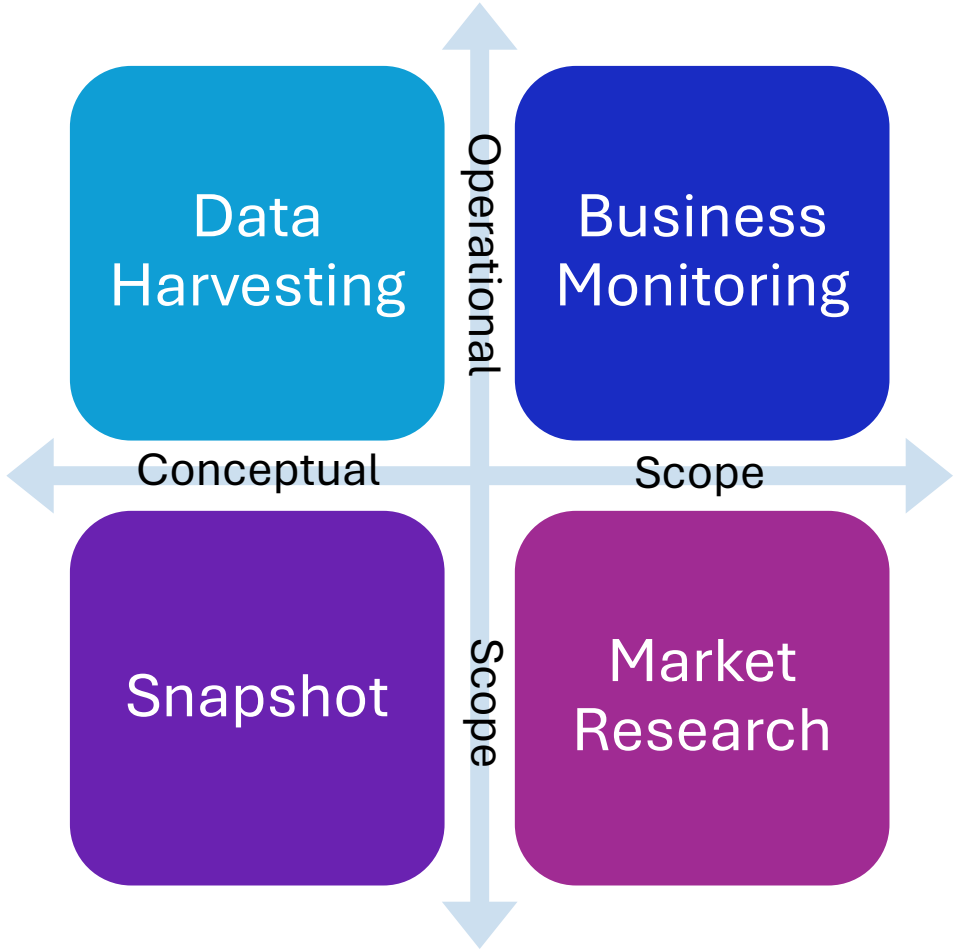


Research Quality Dataset(s)

- Conceptually and operationally limited
- Quality is ***context*** dependent!
- Requires deliberate design decisions
  - Filtering criteria (e.g. top products, largest retailers)
  - Conversion between observation and analysis resolution (e.g. from retailer to brand)

Business vs. Research Needs





Case Study

Adapting web scraped retail price data for public health research

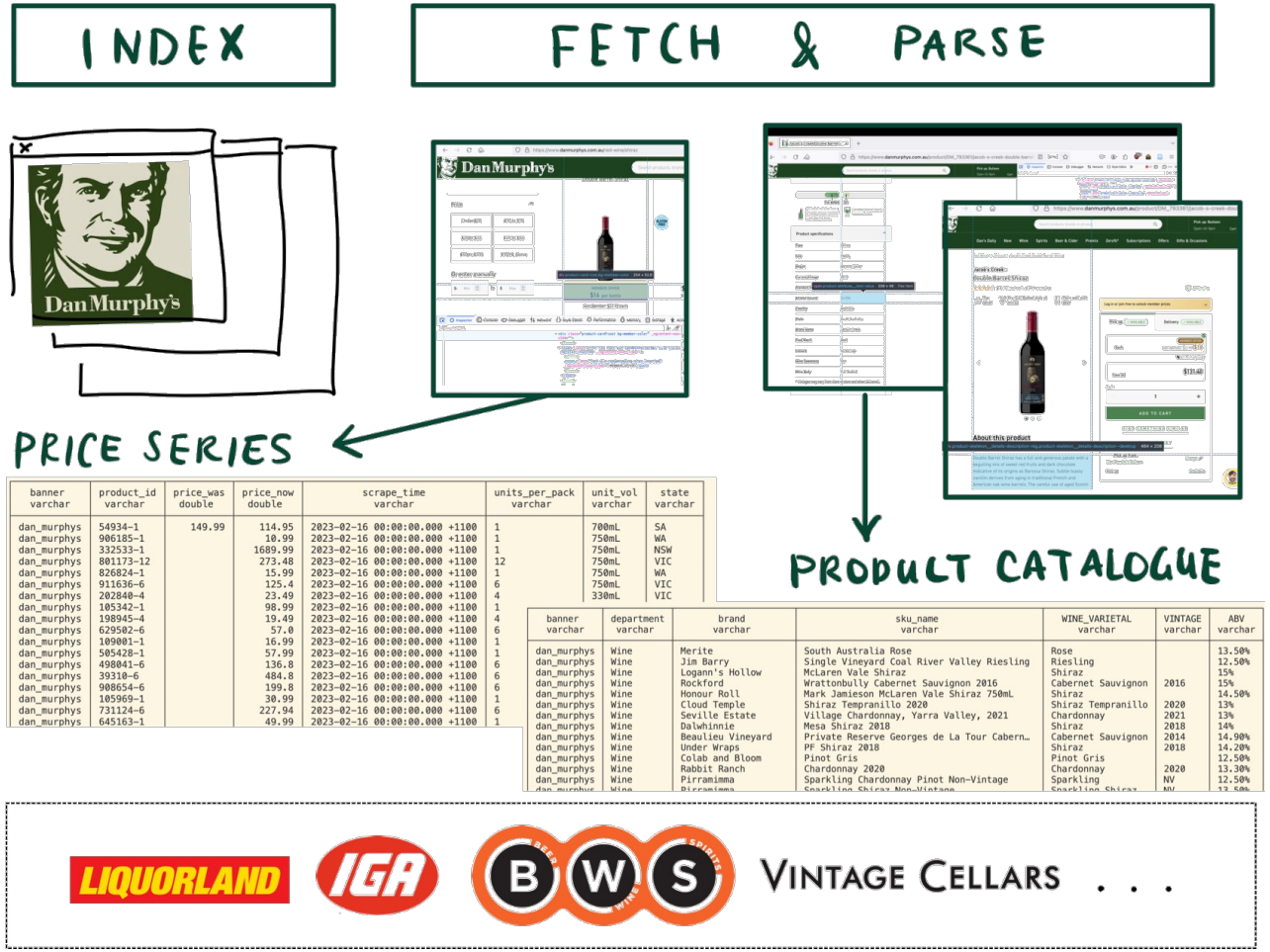
Outsourcing web scraping

Industry Data Partner

MNHS
Health Policy
Research

EBS
Statistics &
Data Science

???
Applied Econ, Causal Modelling

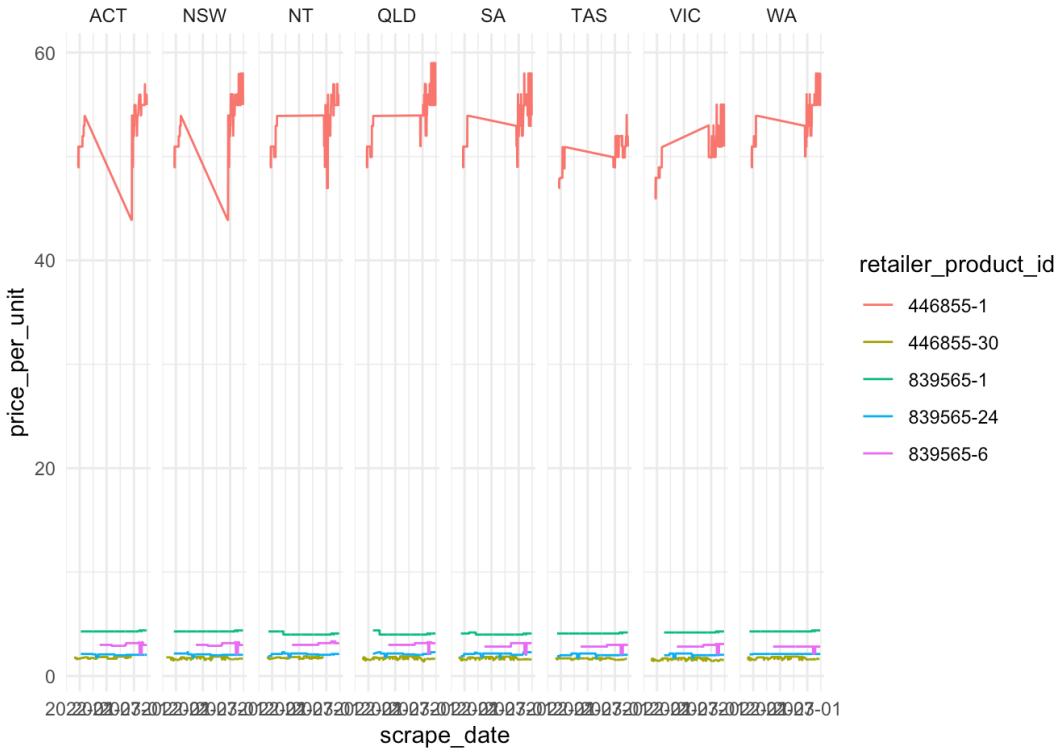


banner varchar	product_id varchar	price_was double	price_now double	scrape_time varchar	units_per_pack varchar	unit_vol varchar	state varchar
dan_murphys	54934-1	149.99	114.95	2023-02-16 00:00:00.000 +1100	1	700mL	SA
dan_murphys	906185-1		10.99	2023-02-16 00:00:00.000 +1100	1	750mL	WA
dan_murphys	332533-1		1689.99	2023-02-16 00:00:00.000 +1100	1	750mL	NSW
dan_murphys	801173-12		273.48	2023-02-16 00:00:00.000 +1100	12	750mL	VIC
dan_murphys	826824-1		15.99	2023-02-16 00:00:00.000 +1100	1	750mL	WA
dan_murphys	911636-6		125.4	2023-02-16 00:00:00.000 +1100	6	750mL	VIC
dan_murphys	202840-4		23.49	2023-02-16 00:00:00.000 +1100	4	330mL	VIC
dan_murphys	105342-1		98.99	2023-02-16 00:00:00.000 +1100	1		
dan_murphys	198945-4		19.49	2023-02-16 00:00:00.000 +1100	4		
dan_murphys	629502-6		57.0	2023-02-16 00:00:00.000 +1100	6		
dan_murphys	109001-1		16.99	2023-02-16 00:00:00.000 +1100	1		
dan_murphys	505428-1		57.99	2023-02-16 00:00:00.000 +1100	1		
dan_murphys	498041-6		136.8	2023-02-16 00:00:00.000 +1100	6		
dan_murphys	39310-6		484.8	2023-02-16 00:00:00.000 +1100	6		
dan_murphys	908654-6		199.8	2023-02-16 00:00:00.000 +1100	6		
dan_murphys	105969-1		30.99	2023-02-16 00:00:00.000 +1100	1		
dan_murphys	731124-6		227.94	2023-02-16 00:00:00.000 +1100	6		
dan_murphys	645163-1		49.99	2023-02-16 00:00:00.000 +1100	1		

banner varchar	department varchar	brand varchar	sku_name varchar	WINE_VARIETAL varchar	VINTAGE varchar	ABV varchar
dan_murphys	Wine	Merite	South Australia Rose	Rose		13.50%
dan_murphys	Wine	Jim Barry	Single Vineyard Coal River Valley Riesling	Riesling		12.50%
dan_murphys	Wine	Logann's Hollow	McLaren Vale Shiraz	Shiraz		15%
dan_murphys	Wine	Rockford	Wrattonbully Cabernet Sauvignon 2016	Cabernet Sauvignon	2016	15%
dan_murphys	Wine	Honour Roll	Mark Jamieson McLaren Vale Shiraz 750mL	Shiraz		14.50%
dan_murphys	Wine	Cloud Temple	Shiraz Tempranillo 2020	Shiraz Tempranillo	2020	13%
dan_murphys	Wine	Seville Estate	Village Chardonnay, Yarra Valley, 2021	Chardonnay	2021	13%
dan_murphys	Wine	Dalwhinnie	Mesa Shiraz 2018	Shiraz	2018	14%
dan_murphys	Wine	Beaulieu Vineyard	Private Reserve Georges de La Tour Cabern…	Cabernet Sauvignon	2014	14.90%
dan_murphys	Wine	Under Wraps	PF Shiraz 2018	Shiraz	2018	14.20%
dan_murphys	Wine	Colab and Bloom	Pinot Gris	Pinot Gris		12.50%
dan_murphys	Wine	Rabbit Ranch	Chardonnay 2020	Chardonnay	2020	13.30%
dan_murphys	Wine	Piramimma	Sparkling Chardonnay Pinot Non-Vintage	Sparkling	NV	12.50%

Data refinements & augmentation

Anomalies in cross-retailer harmonisation



Augmenting product attributes

- Product name
- Brand name
- Manufacturer
- Product category
- Volume per unit
- Alcohol by Volume
- Standard drinks per unit

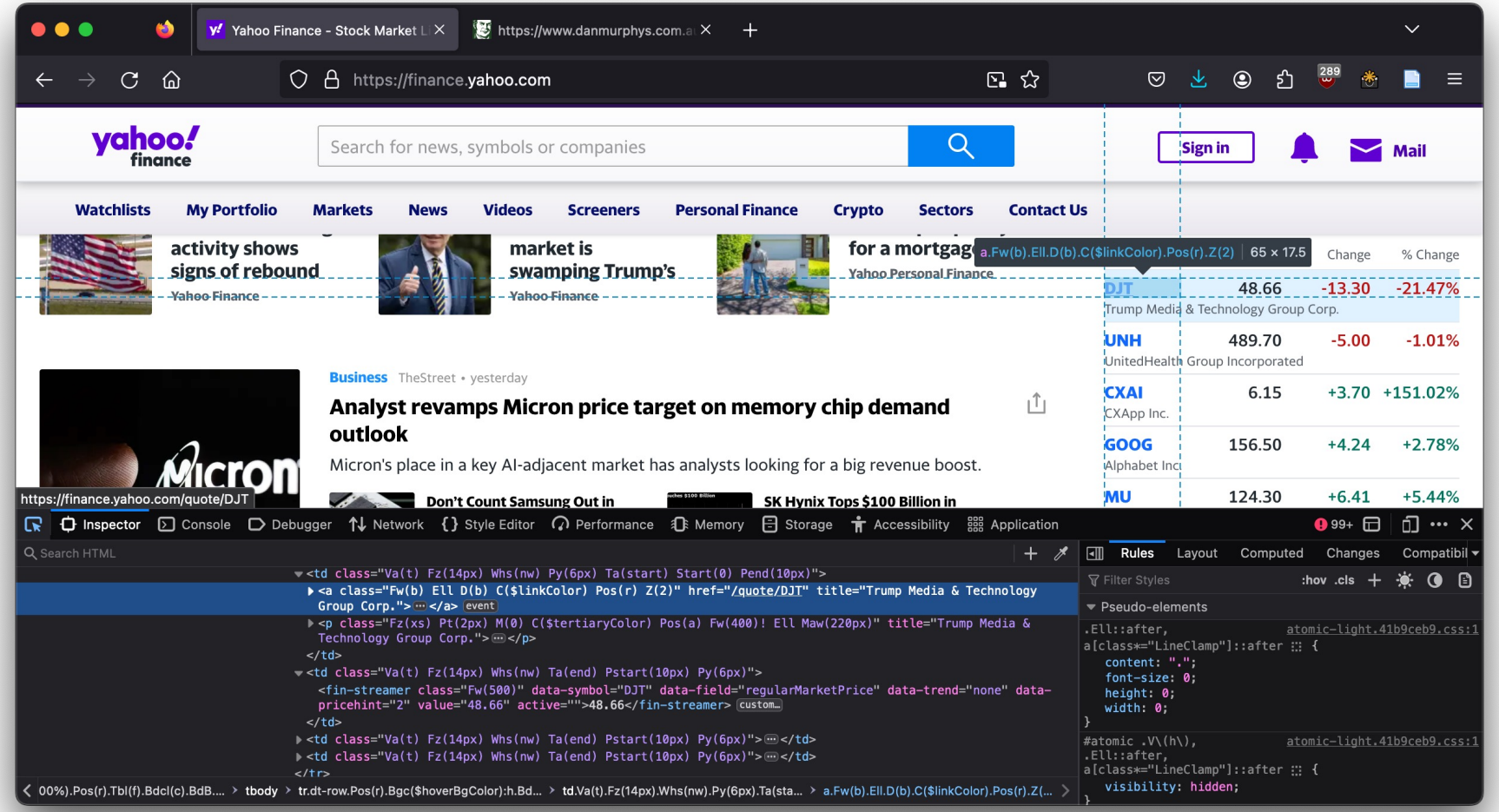
Web technologies, access interfaces, and collection methods

Web scraping: Expectations vs. Reality

Web scraping: Expectations vs. Reality

Live Demo

- **HTML elements**
- API responses



Nigam, H., & Biswas, P. (2021). Web Scraping: From Tools to Related Legislation and Implementation Using Python. In J. S. Raj, A. M. Iliyasu, R. Bestak, & Z. A. Baig (Eds.), *Innovative Data Communication Technologies and Application* (Vol. 59, pp. 149–164). Springer Singapore. https://doi.org/10.1007/978-981-15-9651-3_13

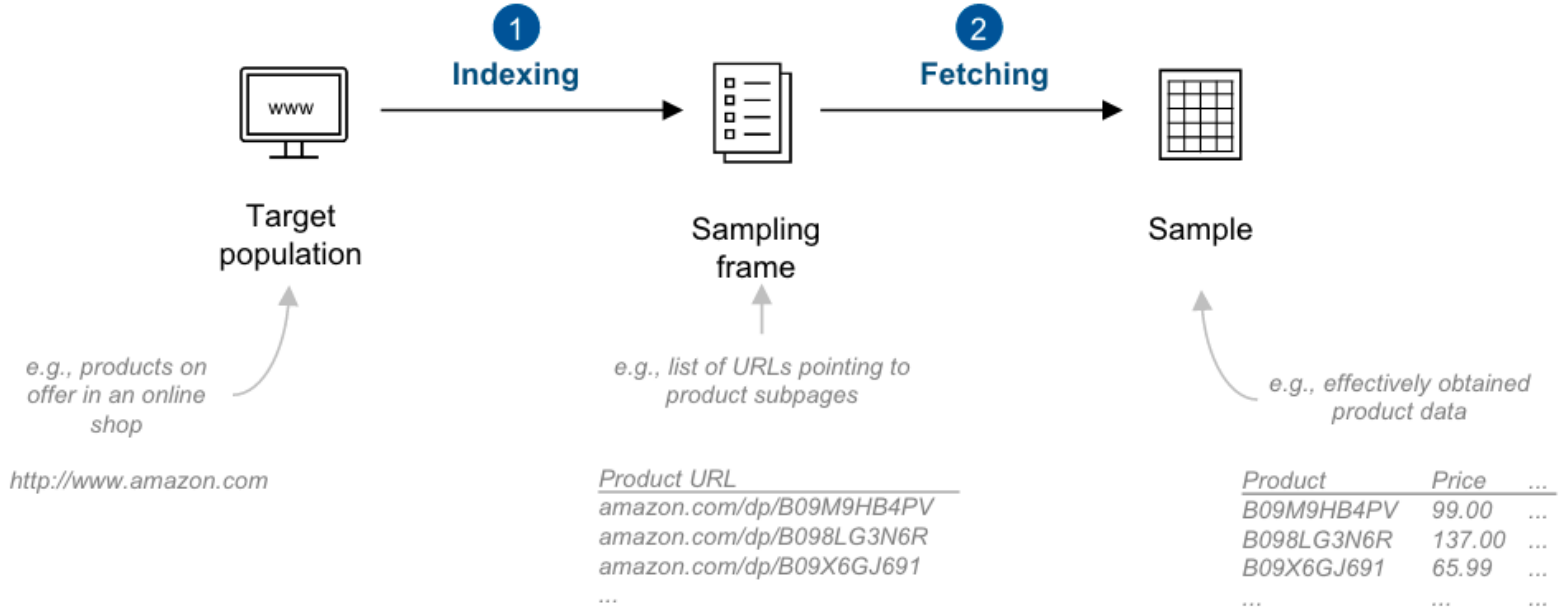
Web technologies

- Webpages and websites
  - Styling: CSS
  - Static content: HTML/XML
  - Dynamic content: Javascript, API queries
- Web Protocols (HTTP/HTTPS)
- Application Programming Interface (APIs)
  - Query parameters
  - Response formats: JSON, CSV…

Web scraping methods: Indexing & Crawling

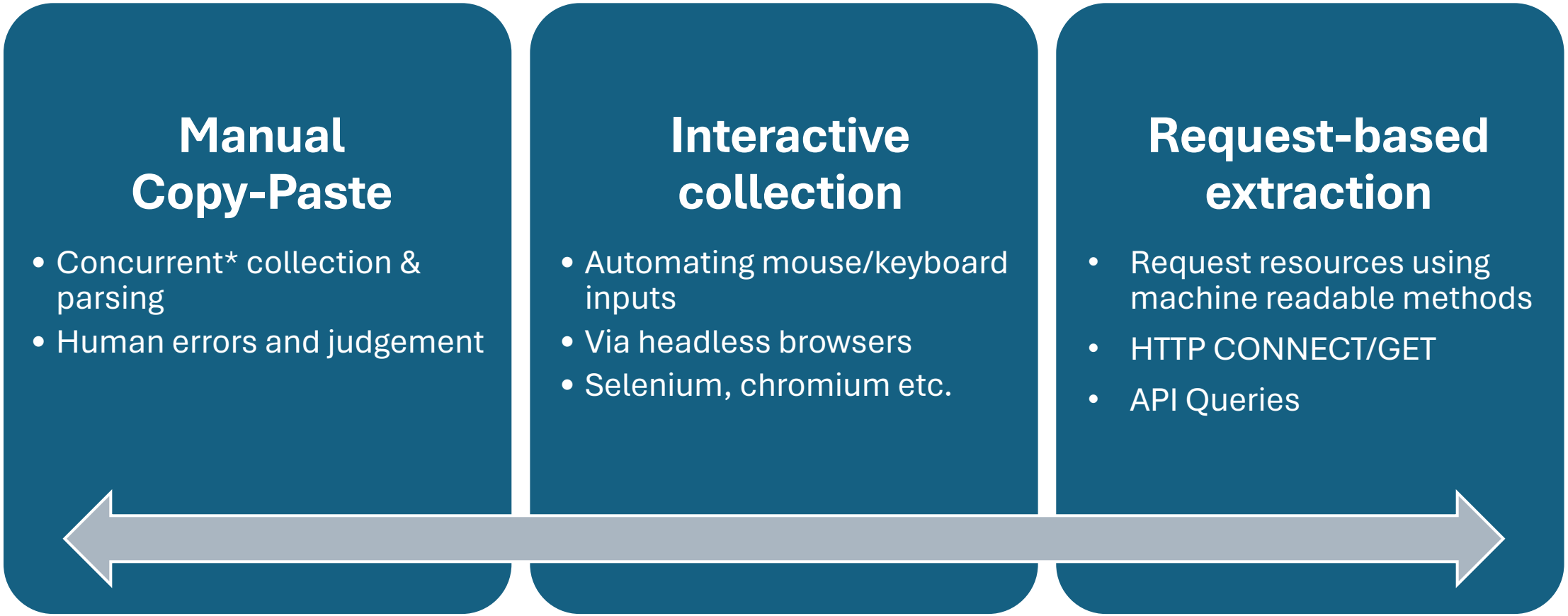
Figure 1: Web Scraping Process: From the Target Population to the Sample.

Web scraping entails two steps, *indexing* and *fetching*. In indexing, the target population is systematically registered. Indexing yields the frame in terms of a register of all units in the population, together with the URLs pointing to each unit. Fetching automatically visits each URL listed in the frame and downloads the resource at which it points, typically an HTML document.

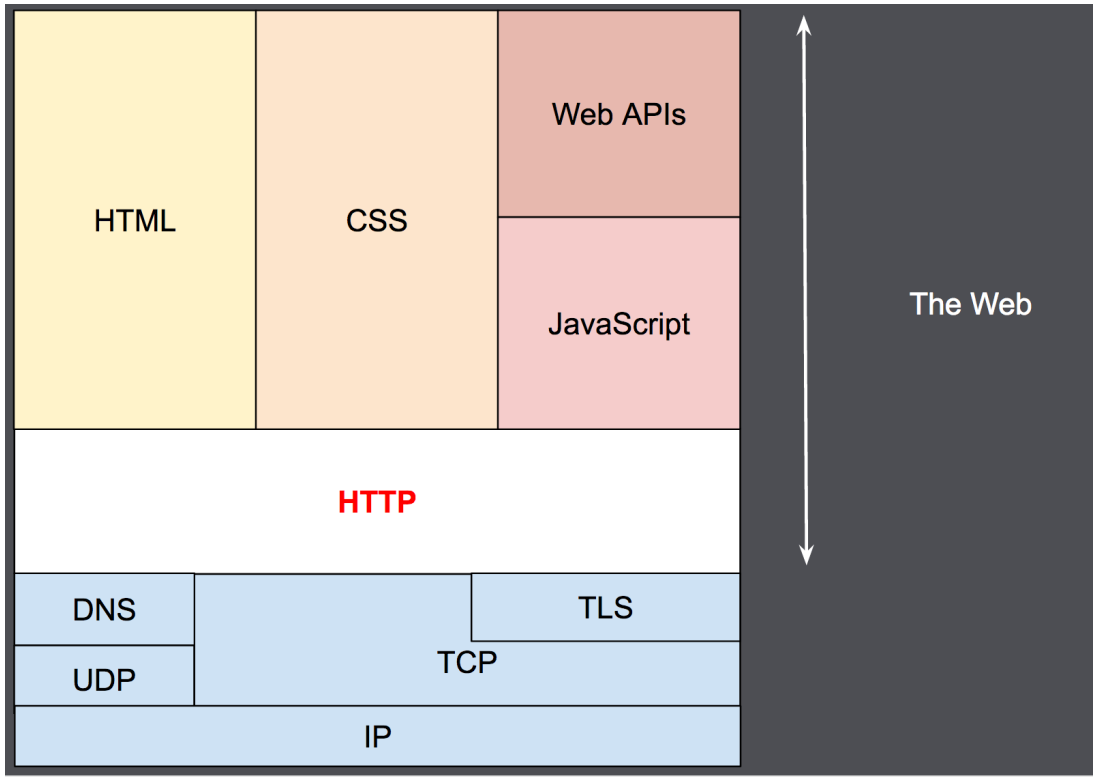


Foerderer, J. (2023). *Should we trust web-scraped data?* (arXiv:2308.02231). arXiv. http://arxiv.org/abs/2308.02231

Web scraping methods: Retrieval



Web scraping methods: Parsing



An overview of HTTP - HTTP | MDN. (2023, December 16). https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview

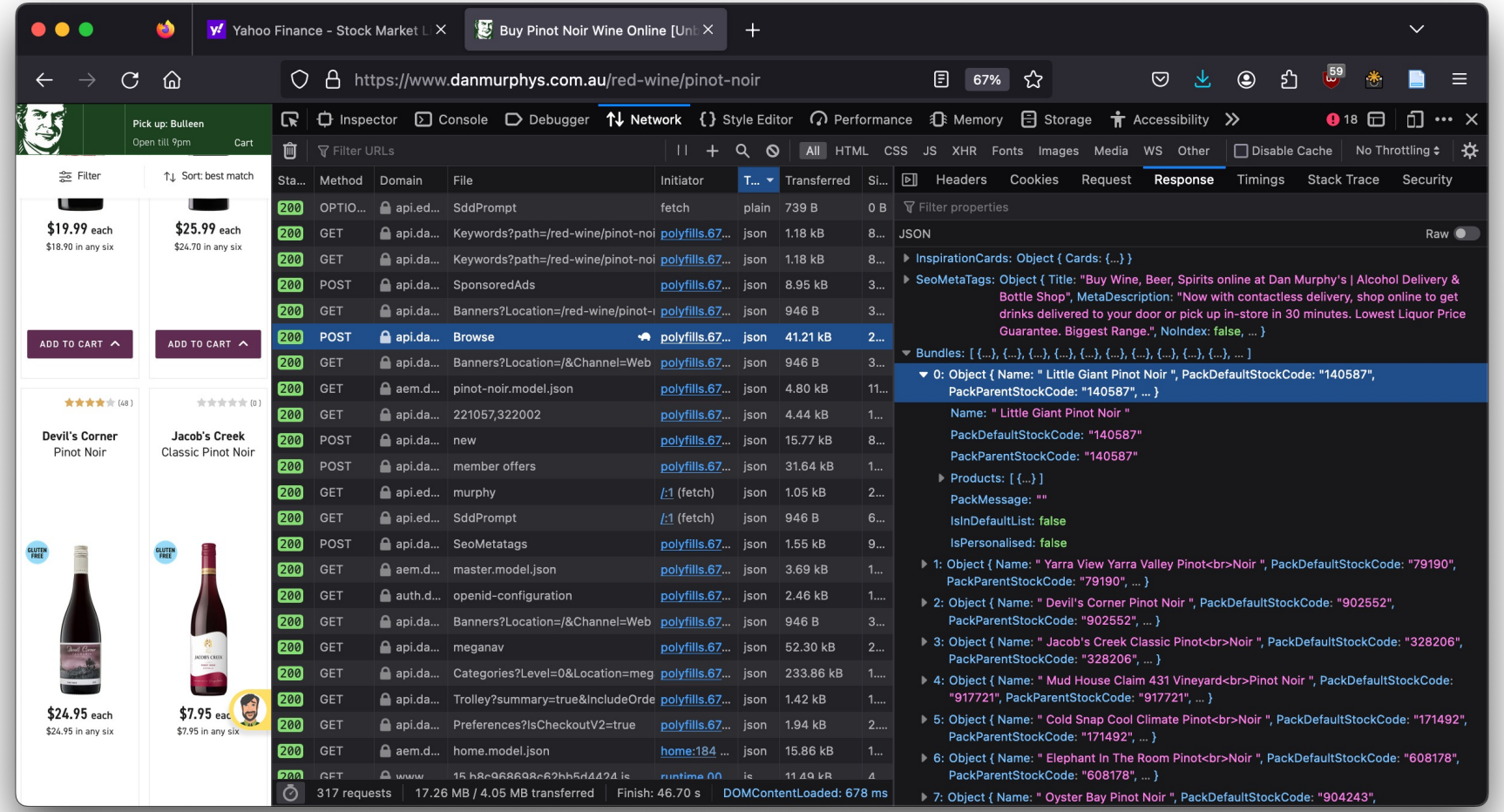
- HTML element extraction:
  - Tables, if you're lucky
  - Often requires detangling style information from metadata and data
- API response parsing:
  - A way for computers to talk to computers without human readability/presentation layer
  - Deciphering JSON keys can be tricky without access to dictionary

Collection Risk by Method / Format

Method	Access Stability	Access Difficulty	Parsing Complexity	Common raw formats
Manual	M	L	L-H???	…
Interactive sessions	L	H	M-H	HTML elements
HTTP page requests	M	L	M-H	HTML elements
API queries	H	L-M	L	JSON, CSV…

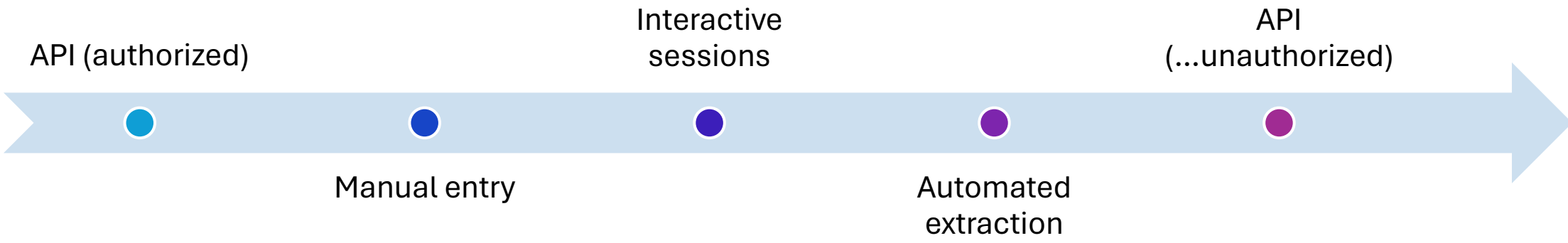
Live Demo

- HTML elements
- **API responses**



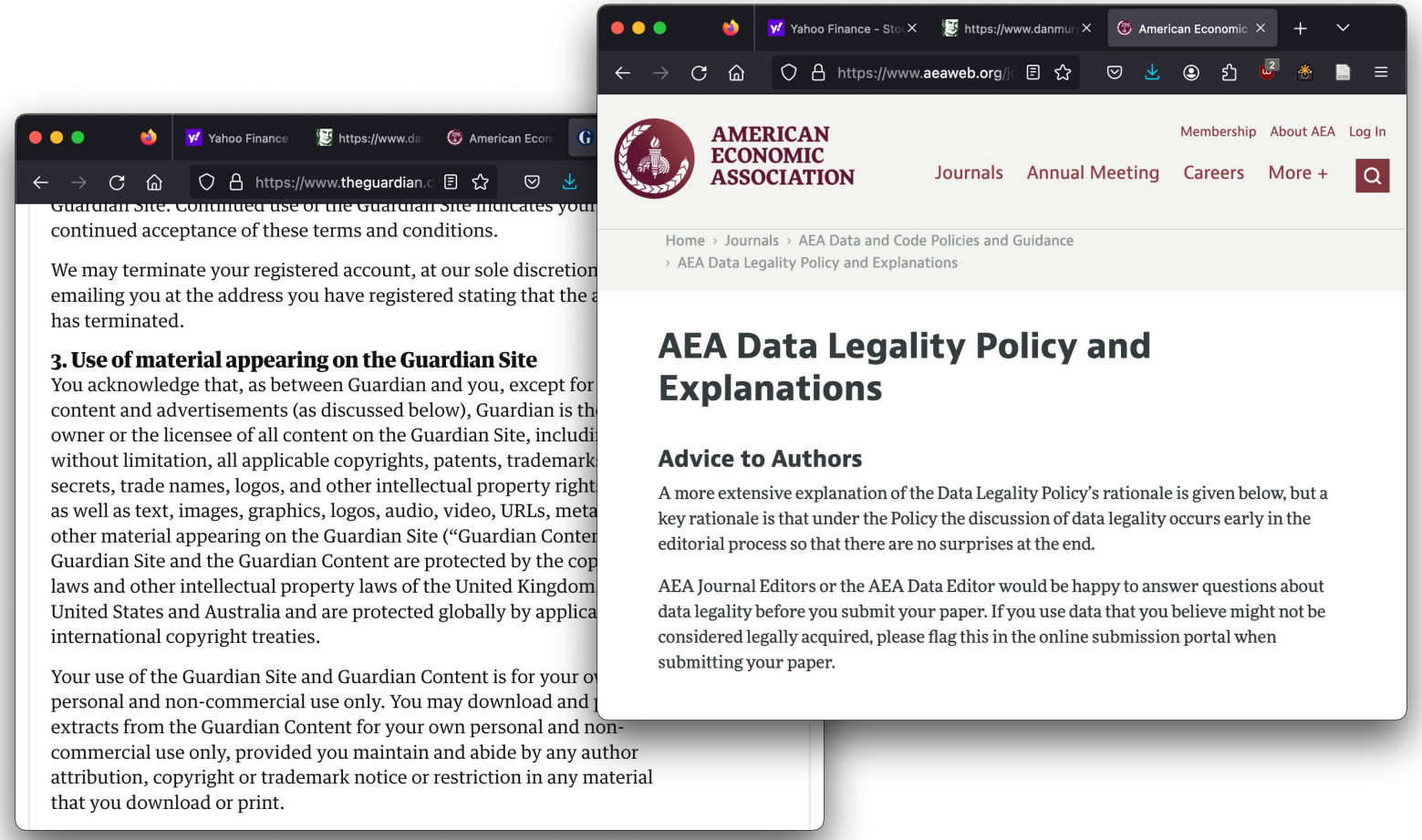
Final considerations

Ethical & Legal Risk: Retrieval



Ethical & Legal Risk: Data Usage

- It depends, check…
  - Website T&Cs
  - Journal policies
  - robots.txt
  - …



Starting a Web Scraped Data Project?

Assess Feasibility & Quality:

- Inspect webpages
- Assess consistency
  - How many different webpage architectures?
- Consider data collection vs. analysis unit/resolution
  - Define ”research quality” in your context
- Consider website and data owner motives
  - Are there any incentives to obfuscate data?

Resource appropriately:

- For in-house hires, consider experience & knowledge of:
  - Retrieval methods,
  - Scraping etiquette,
  - Parsing methods,
  - Data wrangling
- Consider industry partners if available
- Paid services can be suitable for limited data collection

Appendix

Resources

Web-scraping in R

- Extended tutorial by Hadley Wickham:
  - https://github.com/hadley/web-scraping
- Scraping tools:
  - https://rvest.tidyverse.org
- Polite sessions:
  - https://github.com/dmi3kno/polite

Case Study: Resource Sink

- “Although gathering this massive amount of prices was cheaper online than with traditional methods, it required funding that could not be sustained through grants. Thus, in 2011 we started a company called **PriceStats** that now collects the data and produces high-frequency indexes for central banks and financial-sector customers.” (Cavallo and Rigobon, 2016, p. 153)

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives, 30(2), 151–178. https://doi.org/10.1257/jep.30.2.151

AEA Data Legality Policy

- “A particular concern expressed by many researchers is the treatment of computer-assisted acquisition of data (“scraped data”) when such acquisition contravenes terms of use of the data owner. While scraping may contravene terms of use, it may not be illegal. Its legality is not settled under current US law (as of January 2023). Editors will treat papers with scraped data as legally acquired as long as this issue is unsettled when the paper was submitted. Should it in the future become settled law that scraped data is illegal, the AEA will communicate how scraped data will be treated under the Policy."

https://www.aeaweb.org/journals/data/data-legality-policy