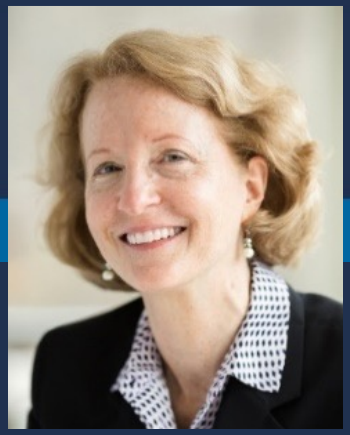
JOURNAL of

Marketing

June 23, 2022



Co-Host: Christine Moorman
Editor-in-Chief
Duke University

AMA | AMERICAN MARKETING ASSOCIATION

Webinar

JOURNAL of
Marketing

The *Journal of Marketing* develops and disseminates knowledge about real-world marketing questions relevant to scholars, educators, managers, policy makers, consumers, and other societal stakeholders around the world.

AMA | AMERICAN MARKETING ASSOCIATION

Fields of Gold: Scraping Web Data for Marketing Insights

Johannes Boegershausen

Erasmus University,
The Netherlands

Hannes Datta

Tilburg University,
The Netherlands

Abhishek Borah

INSEAD,
France

Andrew Stephen

Oxford University,
The UK

Journal of Marketing

Agenda

- **Introduction**
  - Web data in academic marketing research & how to extract it
  - Pathways for creating new marketing knowledge
- **Methodological framework**
  - Managing the idiosyncratic legal, technical and validity challenges of web data
  - Focus on three key stages: source selection, design, extraction
- **Food for thought & conclusion**
  - Key insights
  - Exploiting new fields of gold
- **Q&A**

Web data in academic marketing research

INTRODUCTION

Enormous & diverse data for marketing research

7:11
hours

time spent online per day by the average American consumer

85%

proportion of US consumers that use the Internet every single day

yelp **~ 244m** reviews

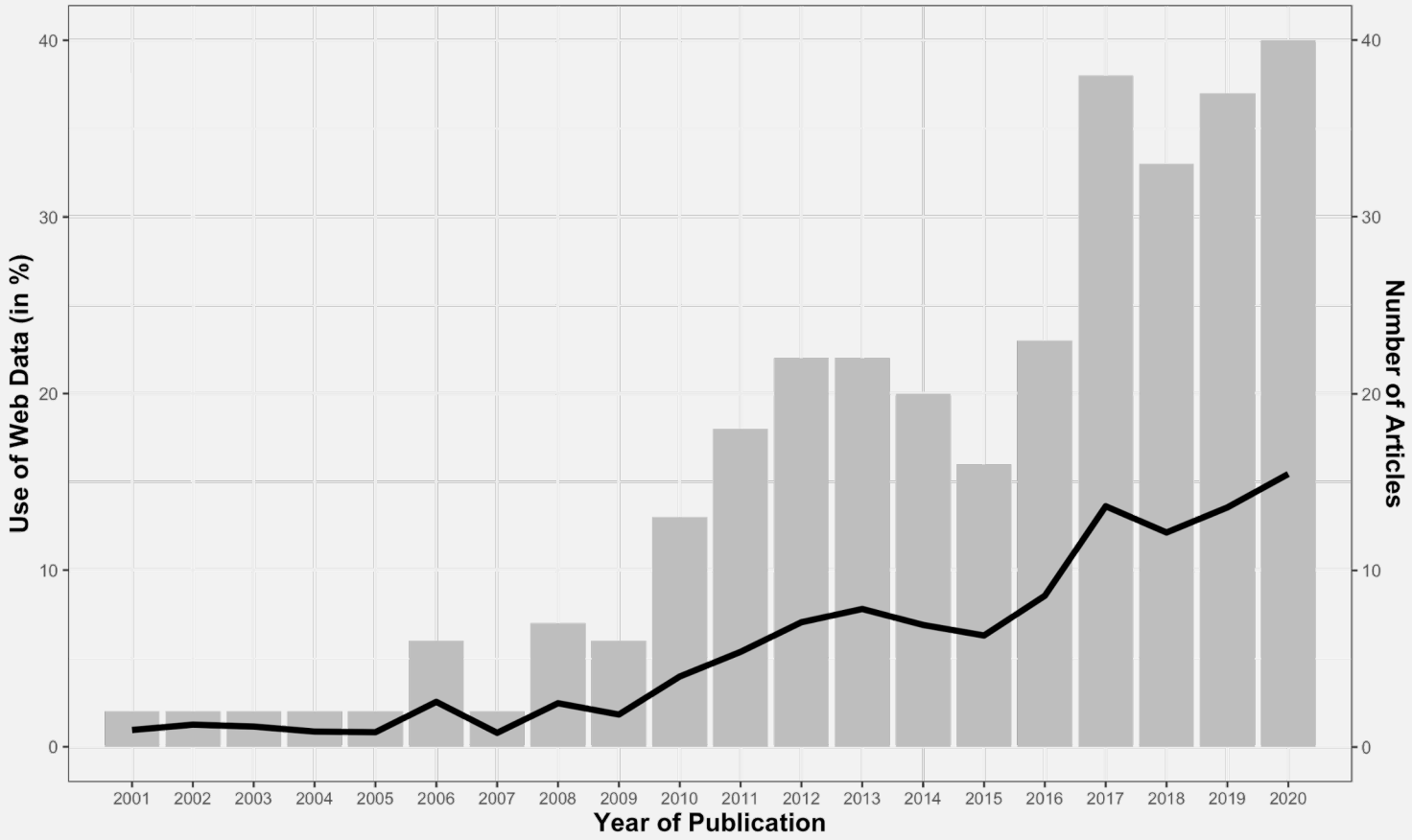
tripadvisor **> 1b** reviews & opinions

500m/day

funded with KICK STARTER **556K** projects

based on available company and market research statistics in May 2022

Increasing usage of web data in marketing research



Extracting web data at scale via…

Web Scraping

… the process of developing software to automatically collect information **displayed in a web browser**

EXAMPLE SOURCES

yelp amazon tripadvisor

Example articles:
Chevalier & Mayzlin (2006); Ludwig et al. (2013)

API

Application Programming Interface

… allow programmatic access to the **internal databases or algorithms of data providers**

EXAMPLE SOURCES

Spotify for Developers

Google Cloud Vision API

amazon Product Advertising API

Twitter API @twitterapi

Example articles:
Tellis et al. (2019); Toubia & Stephen (2013)

Highly versatile data collection technique

Pathway ①

Studying new phenomena

e.g., Zervas et al. (2017); Datta et al. (2018)

Pathway ②

Boosting ecological value

e.g., Du et al. (2015); Ludwig et al. (2013)

Pathway ③

Facilitating methodological advancement

e.g., Netzer et al. (2012); Liu et al. (2020)

Pathway ④

Improving measurement

e.g., Li et al. (2017); Datta et al. (2022)

Source: Boegershausen, Datta, Borah, and Stephen (2022)

Collecting web data can be challenging

- Data generation process is often opaque
- Highly dynamic and unstable environment
- Mostly poorly or undocumented measures
- Cannot be "downloaded" → needs to be generated through automated browsing
- *Numerous idiosyncratic pitfalls (more on that later)*

Single vs. multisource?

Algorithmic biases?

Extraction frequency?

How to sample?

Which software to use?

Keep the raw (HTML, JSON) data?

Single vs. multisource?

Algorithmic biases?

Extraction frequency?

Existing guidance limited

- Focus on technicalities (and not validity)
- Unclear how to deal with or mitigate legal concerns
- Scope of methodological guidance rather narrow

How to sample?

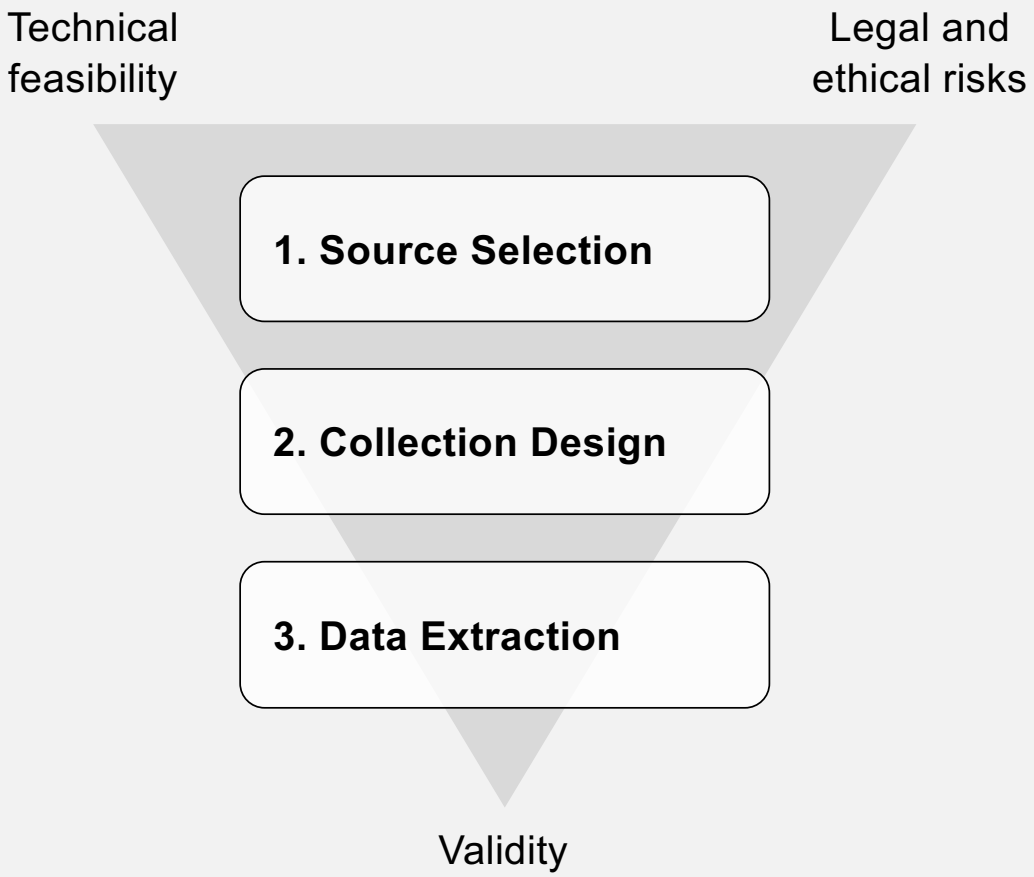
Which software to use?

Keep the raw (HTML, JSON) data?

Managing the idiosyncratic legal, technical and validity challenges of web data

METHODOLOGICAL FRAMEWORK

Methodological framework

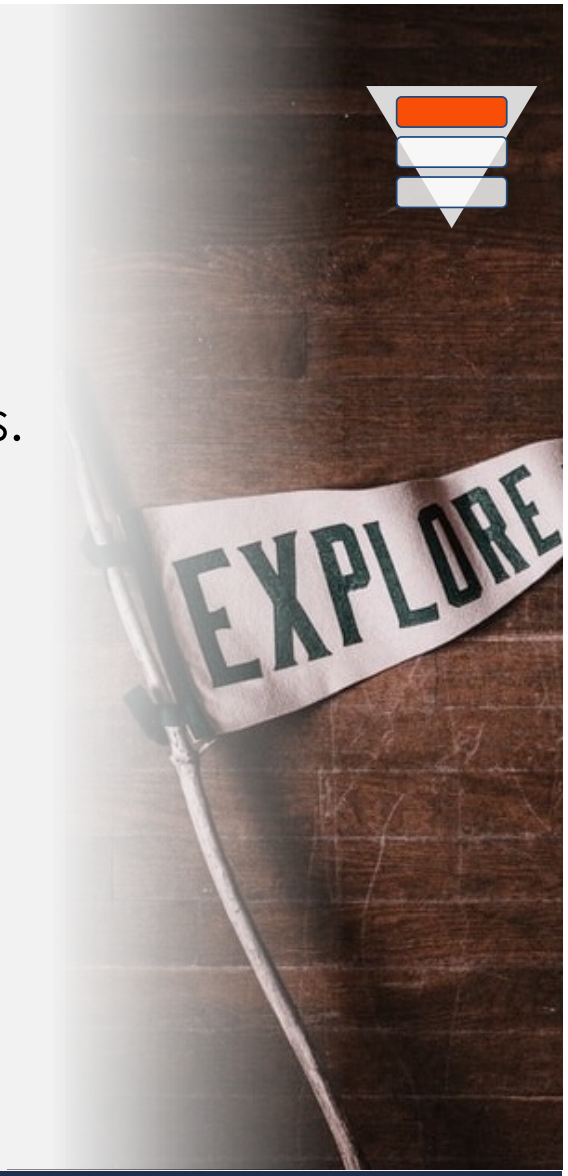


Source: Boegershausen, Datta, Borah, and Stephen (2022)

Source selection: challenges

- Access to near-to infinite number of potential sources without traditional gatekeepers. Different forms of access.
- But sources vary vastly in terms of quality, stability, and retrievability.

→ Might prompt researchers to primarily consider dominant or familiar platforms only.



Source selection: recommendations I

- **Explore the universe** of potential web sources
  - Broaden geographic search criteria (e.g., non-Western)
  - Identify adjacent data sources (e.g., Google Trend's "related search queries")
  - Expand search to non-primary data providers (e.g., aggregators like SocialBlade)

Source selection: example

yelp

tripadvisor

amazon

Source selection: example

yelp

SKYTRAX

tripadvisor

healthgrades.

How America finds a doctor.

amazon

IMDb

Review us on... facebook

Source selection: example

yelp

SKYTRAX

bol.com
de winkel van ons allemaal

tripadvisor

healthgrades
How America finds a doctor.

wayfair.co.uk

amazon

IMDb

Review us on... facebook

Source selection: recommendations II

- Explore the universe of potential web sources
  - Broaden geographic search criteria (e.g., non-Western)
  - Identify adjacent data sources (e.g., Google Trend's "related search queries")
  - Expand search to non-primary data providers (i.e., aggregators, databases)
- Consider **alternatives to web scraping**
  - Expand search by explicitly including terms such as "API" or "dataset download"
  - APIs? How does the data compare to data that could be scraped?

Recommender Systems and Personalization Datasets

Julian McAuley, UCSD

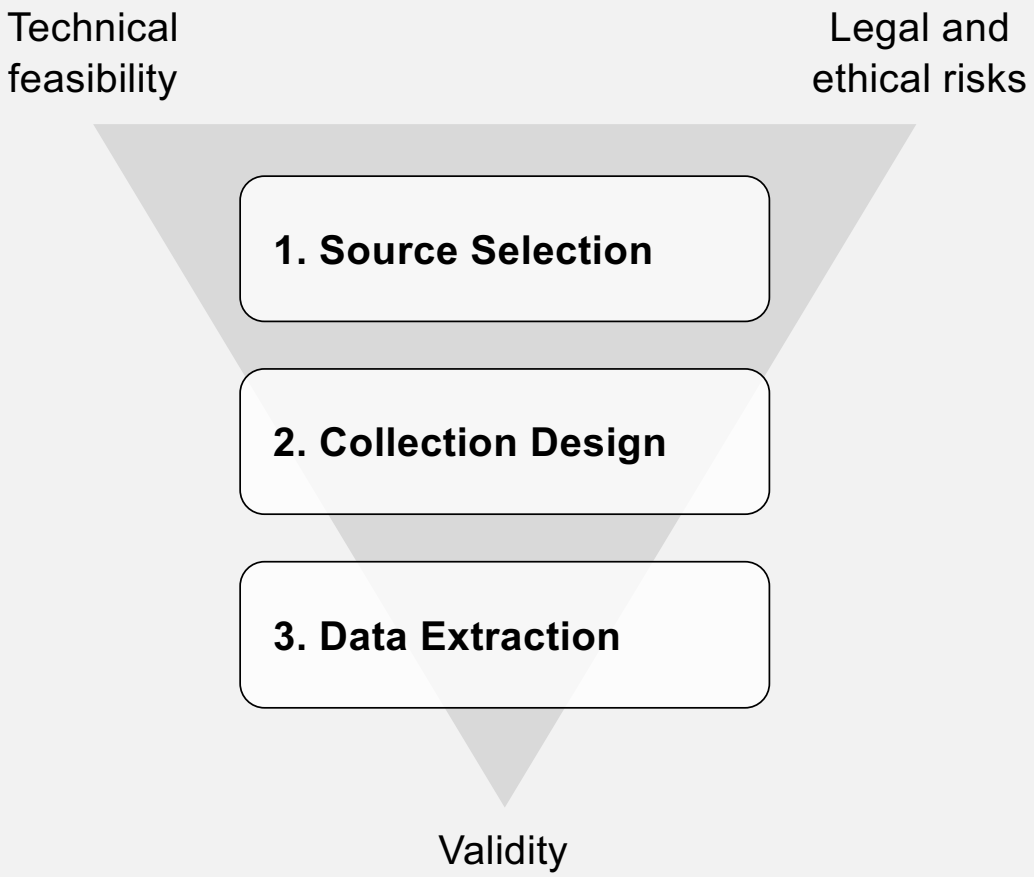
yelp Dataset

kaggle

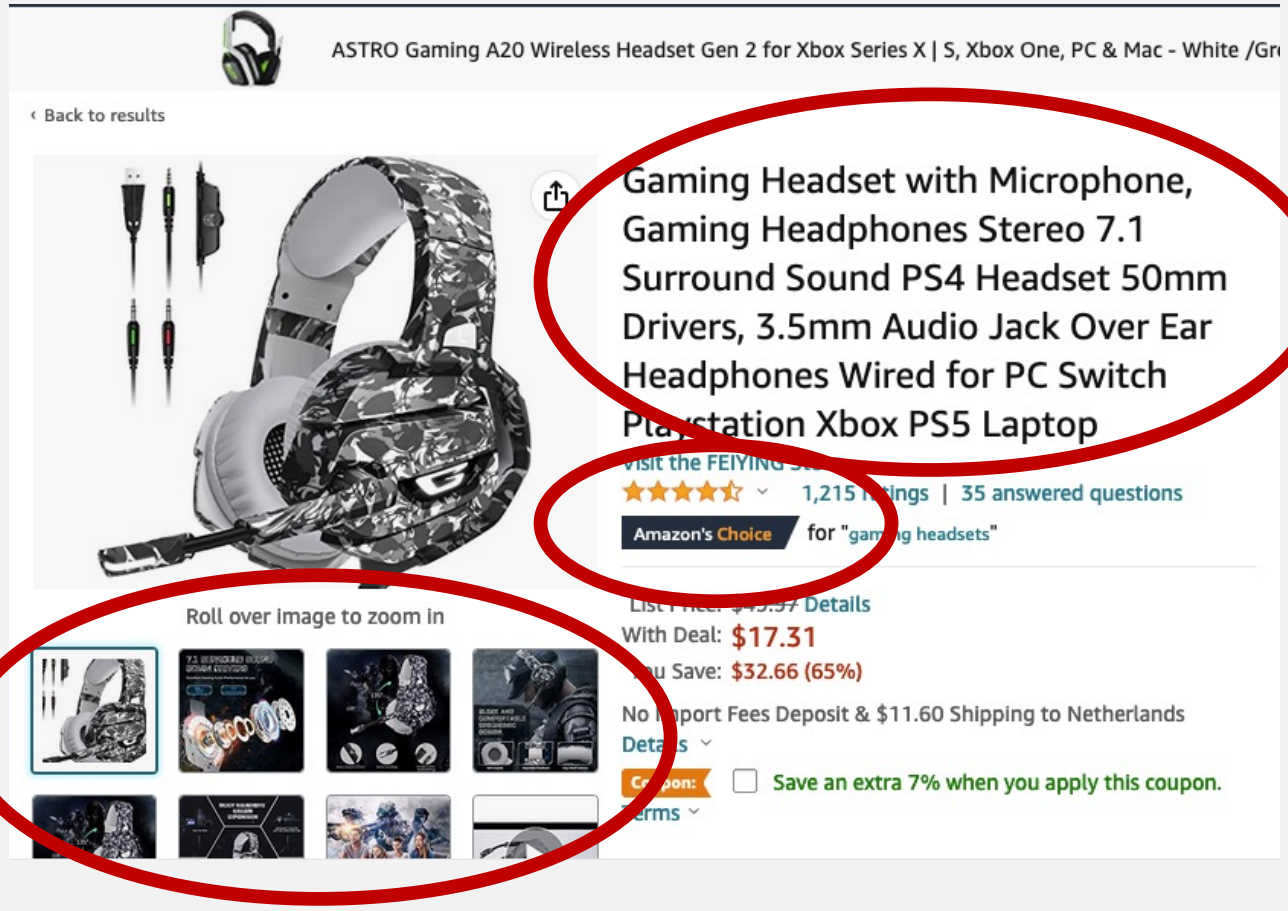
Source selection: recommendations III

- Explore the universe of potential web sources
  - Broaden geographic search criteria (e.g., non-Western)
  - Identify adjacent data sources (e.g., Google Trend's "related search queries")
  - Expand search to non-primary data providers (i.e., aggregators, databases)
- Consider alternatives to web scraping
  - Expand search by explicitly including terms such as "API" or "dataset download"
  - APIs? How does the data compare to data that could be scraped?
- **Map the data context**
  - Screen blogs, press releases, a source's software "changelogs,", …
  - Understand changes to the data-generating process (e.g., archive.org)
  - Algorithms present? Visit source using different devices/times, inspect source code

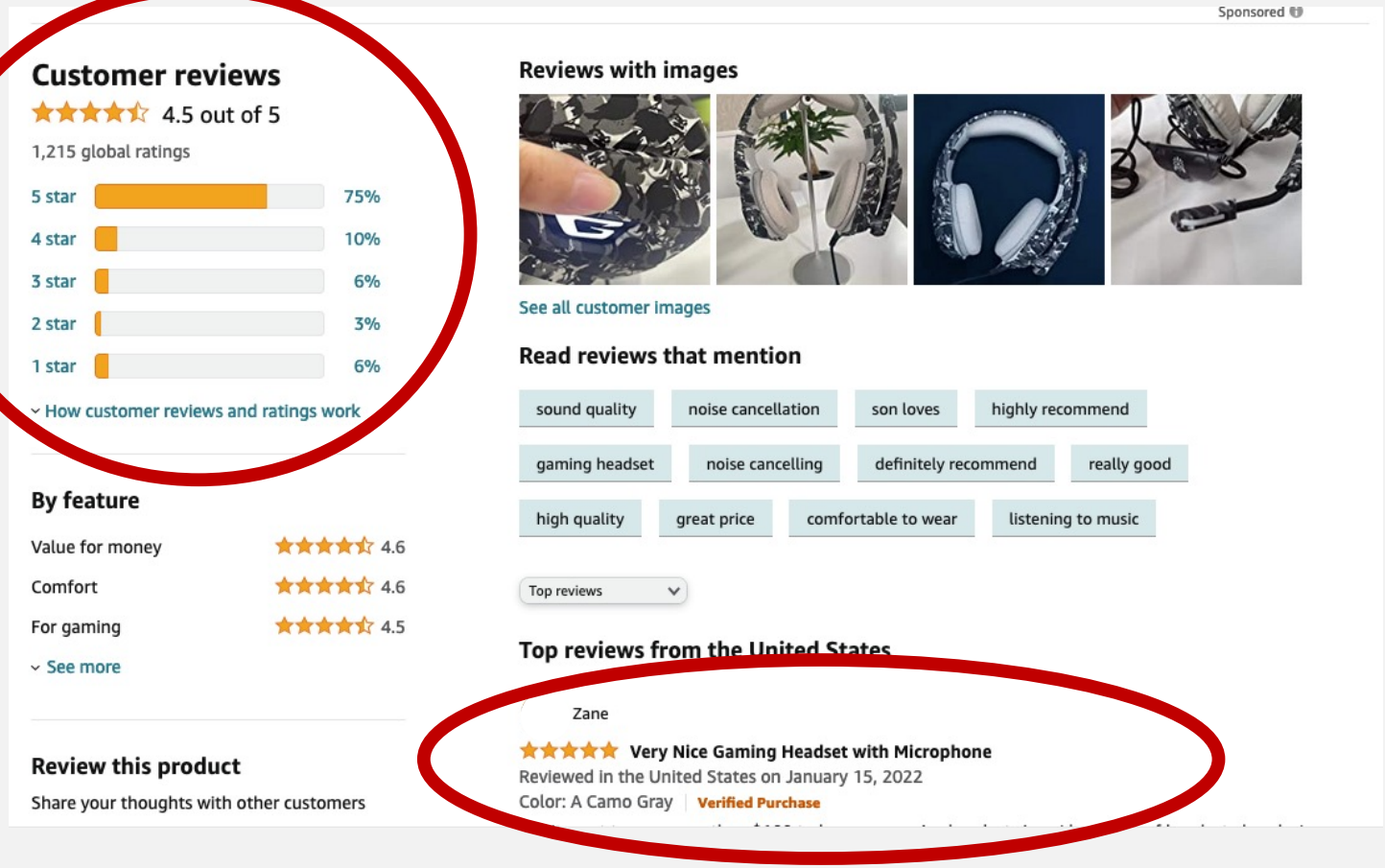
Designing the data collection



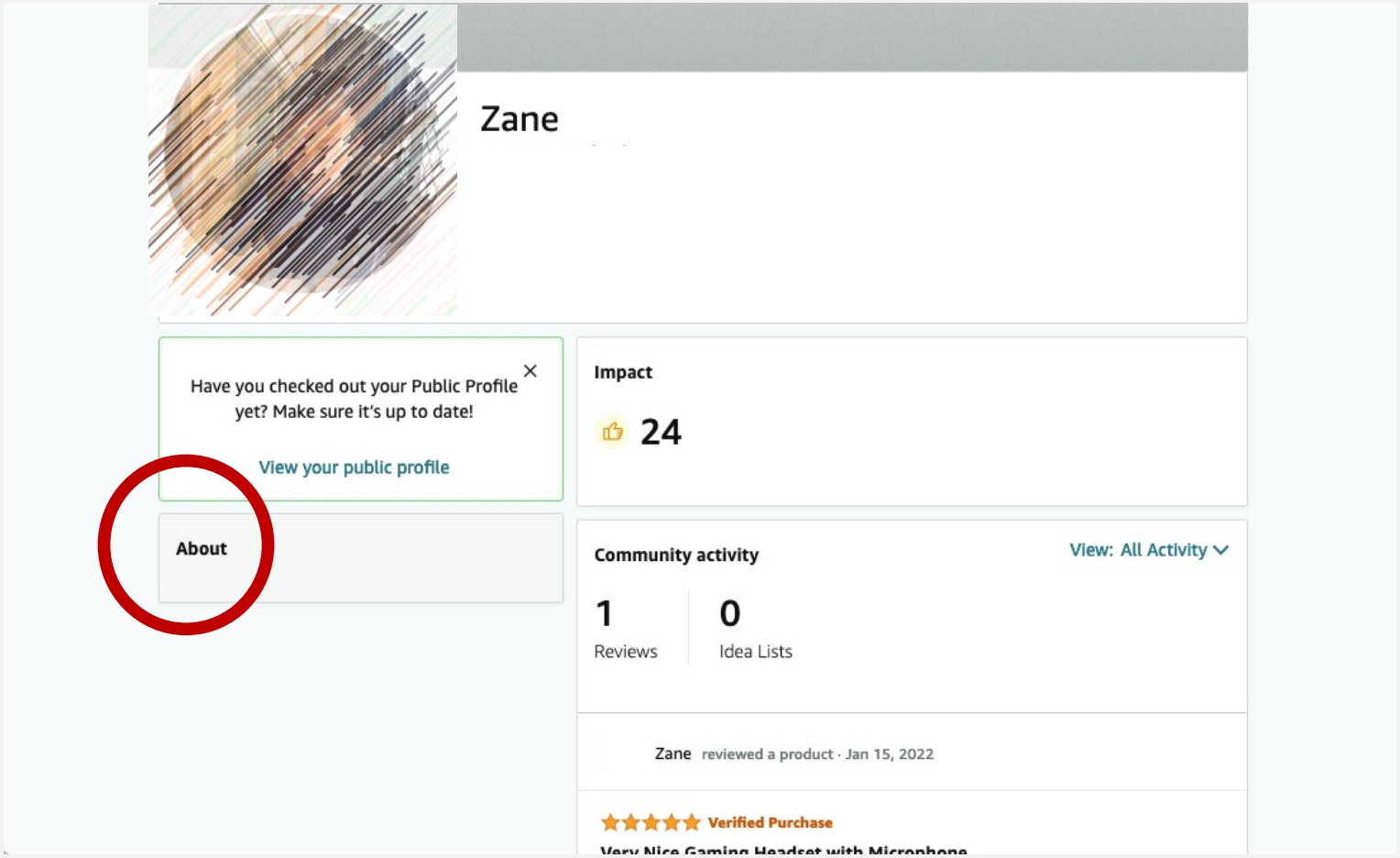
Which information to extract? Example



Which information to extract? Example



Which information to extract? Example



Which information to extract? Example

Validity implications

- Is information subject to algorithmic biases or missing data?
  Delete cookies & check?
- Are there significant changes to the data-generating process?
  Archive.org
- Is meta data required to make sense of variables?
  Save timestamps/IP addresses

Legal/ethical risks

- Publicly accessible vs. login? Consent to ToS? Implicit or explicit?
  Focus on public pages
- Personal or sensitive information?
  Anonymize while collecting
- Overlap original intent of posting & research question / scientific justification?
  Formulate scientific justification

Technical feasibility

- All information extractable?
  Build prototype
- Limits to iterating through pages?
  Check last page, try a few in-between

How to sample? Challenges & considerations

- How to capture the entire population (or a sample) of…?
  - Internal pages (e.g., bestseller, category, search page)
  - Externally available lists?
- Sampling frames (might) create different datasets or even induce systematic biases
- Which sample size is technically feasible?

At what frequency to extract data? Challenges

- Validate "data" assumptions early on
  - Configuration (e.g., "data is historically available")
  - Data-generating process (e.g., "website hasn't changed")
  - Characteristics (e.g., measurement is clear; use of interpolation)

- Examples
  - Archival versus "live" data → discover fake reviews
  - Gains from capturing information more than once? → build longitudinal data set
  - Balance sample size and extraction frequency → sufficient power?

How to process data during the extraction?

- Web data is “messy”
- BUT “on-the-fly” processing can create significant threats to validity

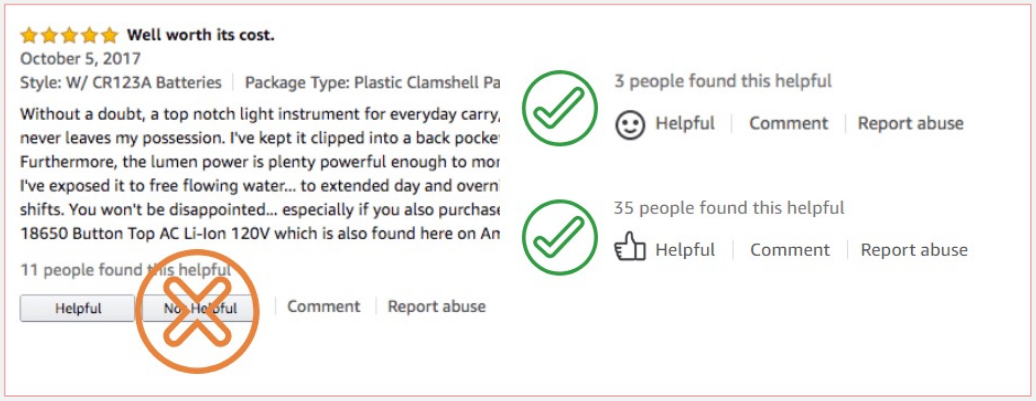
→ **<u>Keep the raw data whenever possible</u>**

How to process data during the extraction?

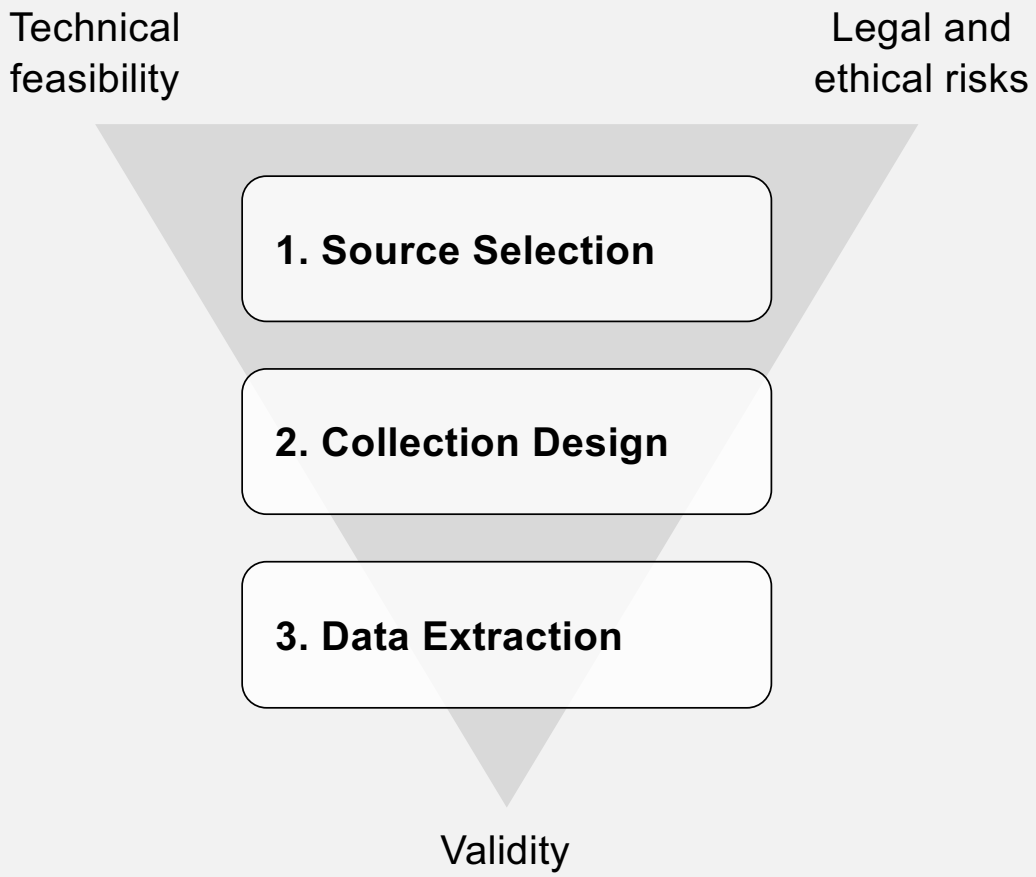
- Web data is “messy”
- BUT “on-the-fly” processing can create significant threats to validity

→Keep the raw data whenever possible

- **Opportunity: “stumbling” into natural experiments**



Data extraction



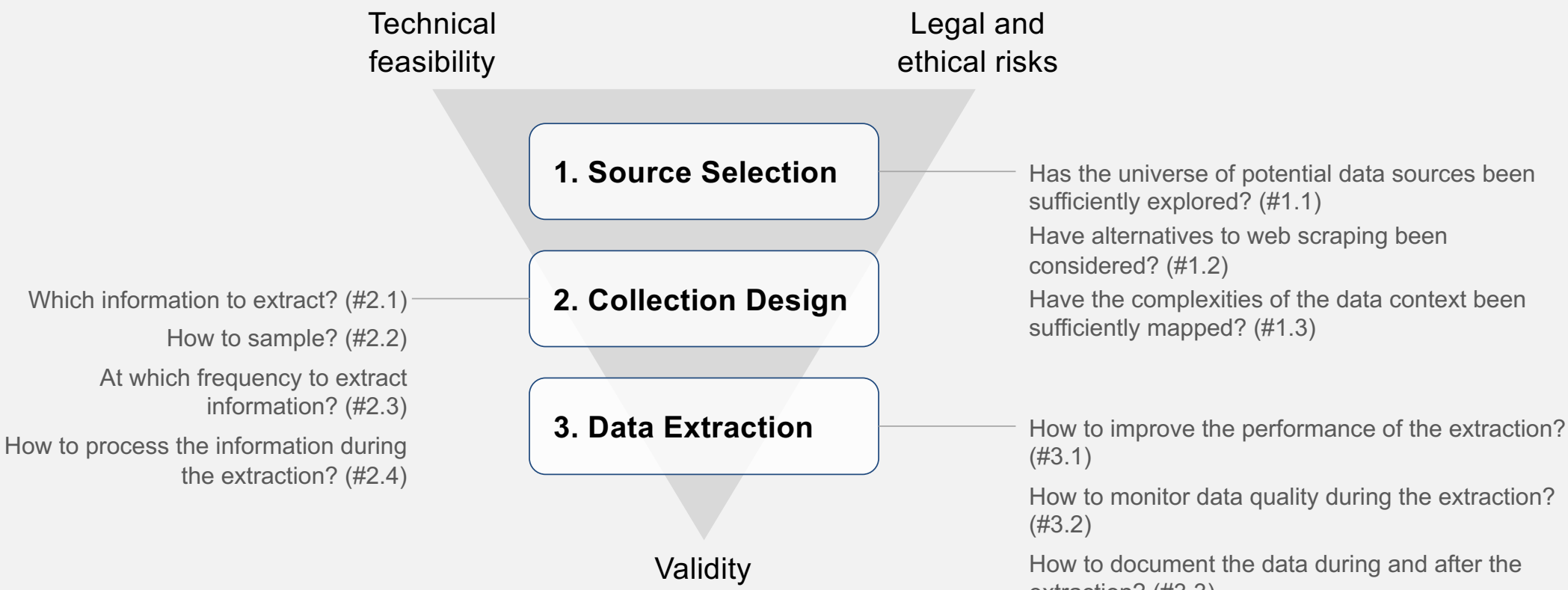
Data extraction

- How to **improve** the performance of the data extraction?
  - Keep the collection running for some time – does it continue to work?
  - Log the (timestamped) URLs of scraped pages and visualize performance over an extended period.

- How to **monitor** data quality during the extraction?
  - Collect and report metadata to diagnose issues in real-time

- How to **document** the data **during** and **after** the extraction?
  - Nobody, except you, knows how the data was generated!
  - Start early! Logbook. Collect information around the focal source(s).

Methodological framework: summary



Source: Boegershausen, Datta, Borah, and Stephen (2022)

Our paper helps reasoning through design challenges…

TABLE 3
CHALLENGES AND SOLUTIONS IN DESIGNING WEB DATA COLLECTIONS

Challenge #2.1: Which information to extract from which pages?

Validity Challenges [V]	Legal/Ethical Challenges [L]	Technical Challenges [T]	Solutions and Best Practices
• Which information is necessary to justify construct?operationalization and allow analysis? • Which metadata might enhance internal and external validity? • Is information subject to algorithmic biases or missing data? • Are there significant changes to the data-generating process?	• Is all of the required information publicly accessible, or is a login required? • Does the data contain personal or sensitive information, and can subjects be identified? • Is there a sufficient scientific justification for using the data? • How large is the overlap between the research objective and the original intent of subjects disclosing the data?	• Is all information extractable? • Are there any limits to iterating through pages or endpoints? • Does the extraction software obtain information reliably?	• Explore different types of pages to detect unique vs. identical information [V] • Explore whether alternative ways to browse/navigate the site (e.g., URLs, clicking, scrolling, logging in to the site) provides different or reveals new information [T] • Explore how extraction methods (e.g., "headless" HTTP requests vs. simulated browsing, different user agents, screen width, login status, use of different packages) affect information display [V, T] • Assess the accuracy of timestamps (e.g., time zones) [V] • Save screenshots of pages that describe the calculation of metrics [V] • Explore (temporarily available) information in the source code of a website using the browser's "inspect" tools [V] • Assess the presence of technical roadblocks (e.g., captchas) [T] • Assess how data was generated historically at the source (e.g., via archive.org) [V] • Explore limits to iterating through pages [T] • Obtain information from various sources to reduce dependency on data provider [L] • If possible, opt-out of firm-administered experiments or block cookies; alternatively, identify relevant metadata that can be used to control for the presence of algorithms [V]

Challenge #2.2: How to sample?

Validity Challenges [V]	Legal/Ethical Challenges [L]	Technical Challenges [T]	Solutions and Best Practices
• Is the sample size sufficient to effectively inform the research question? • To which population does the sample generalize? • Is the sampling frame corresponding to the research objective (e.g., randomness)? • How prevalent is panel attrition?	• Does the data represent an excessive portion relative to all data available? • Can the data be obtained in similar forms elsewhere, or is the research question only answerable with the targeted data? • Are some of the sampled units (potentially) vulnerable?	• Is the required sample size technically feasible? • Can external information (e.g., IDs) be consistently matched to the data?	• Assess characteristics of the population (e.g., using secondary sources) [V] • Explore options to sample directly from the source (e.g., from different pages, randomization through filtering/searching, obtaining usernames from forums, see also Neuendorf 2017 and Humphreys and Wang 2018) [V] • Choose lists or pages that are not affected by algorithmic influence [V] • Refresh sample (or use multiple types of sampled units) to assess the stability of sample and counterbalance panel attrition [V] • Discard units from the sample to prevent data collection from subjects falling under prohibitive national and supranational legislation (e.g., GDPR) [L] • Explore external sources to inform the sampling frame [V], or facilitate linkage [T] • Assess the efficiency of different navigation paths and their impact on sample size [T] • Pseudo-anonymize or discard sensitive or personal information [L] • Ensure no excessive amount of data (e.g., data on all users) is collected (absolute volume, relative volume) [L] • Re-examine alternative sources to improve justification of data extraction [L]

TABLE 3
CHALLENGES AND SOLUTIONS IN DESIGNING WEB DATA COLLECTIONS [CONTINUED]

Challenge #2.3: At which frequency to extract the data?

Validity Challenges [V]	Legal/Ethical Challenges [L]	Technical Challenges [T]	Solutions and Best Practices
• Is the extraction frequency in sync with the studied phenomena? • Is the refresh rate of the source sufficient? • Is the data (thought to be archival) really archival? • Is the information consistently available across all periods of interest? • Does the order and frequency in which information is retrieved induce bias?	• Does the extraction frequency pose an excessive load on the source? • Does collecting more data at higher frequencies make the data more sensitive?	• Does the desired extraction frequency pose new technical hurdles? • How can the stability of data collection be guaranteed, and different collection batches be distinguished?	• Explore the gains in collecting data multiple times rather than once (e.g., in a "live" data collection) [V] • Adhere to best practices in setting the extracting frequency (e.g., 5 requests per second for APIs, 1 request per 2 seconds for web scraping) [L, T] • Experiment with technical parameters (e.g., number of computers) to balance technically feasible sample size and desired frequency of data extraction [T] • Formulate, test, and refine data source theory (Landers et al. 2016) [V] • Reinspect the robots.txt file to avoid exceeding retrieval limits for selected pages [T] • Consider randomizing extraction order for sampled units over time [V] • Consider (cost) implications for storage and computation time [T] • Consider getting in touch with the data provider if the targeted data set is infeasible to extract via web scraping or APIs [T, L] • Devise a schedule for the automatic extraction of the data (e.g., using Windows Task Manager or Cron) [T, V]

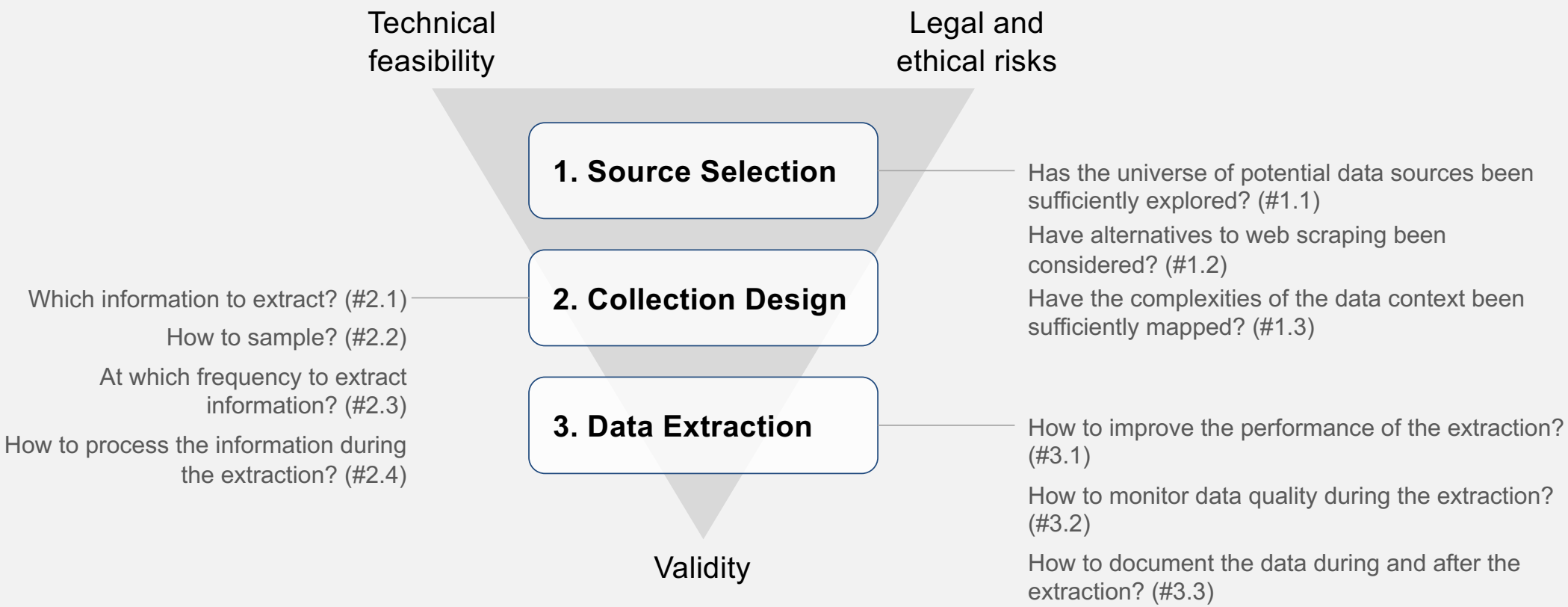
Challenge #2.4: How to process the data during the collection?

Validity Challenges [V]	Legal/Ethical Challenges [L]	Technical Challenges [T]	Solutions and Best Practices
• Could erroneous processing lead to unexpected data loss? • Could there be any significant scientific value in retaining the raw data?	• Is the collected data in conflict with prohibitive laws (e.g., GDPR)? • Is the collected data sufficiently secured from unauthorized access? • Is anonymization or pseudonymization required?	• Which storage facilities to use to accommodate the expected data (size, location, format, encoding) • Is normalization necessary?	• Retain raw data (e.g., HTML pages, JSON responses) whenever possible [V, T] • Always parse some minimal amount of data (e.g., timestamps) to facilitate monitoring checks in real-time [V, T] • Remove sensitive and personal information on the fly; if personal or sensitive information is strictly required to meet the research objective, consider pseudo-anonymizing (potentially via third parties) [L] • Verify data storage during collection meets legal requirements for potentially sensitive or personal data [L] • Ensure proper encoding of (non-English) characters, retain correct digit separators and correct data format

IMPORTANT: trade-offs are (almost) inevitable

MAKE TRADE-OFFS EXPLICIT IN THE MANUSCRIPT

Questions

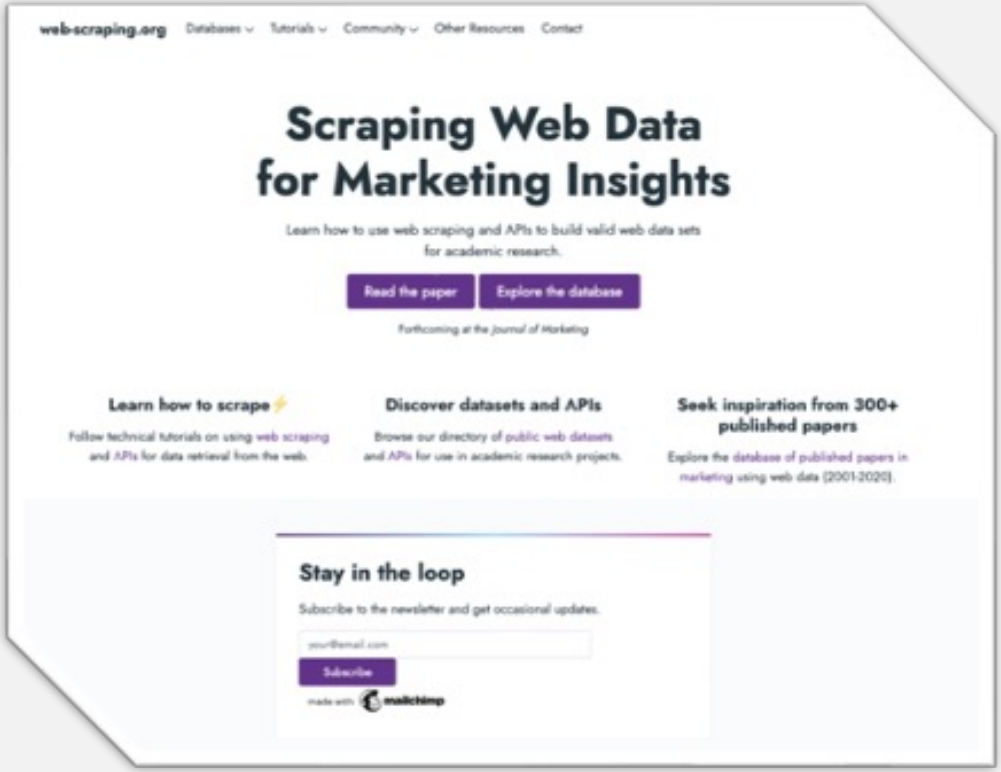


Source: Boegershausen, Datta, Borah, and Stephen (2022)

Key insights & exploiting new fields of gold

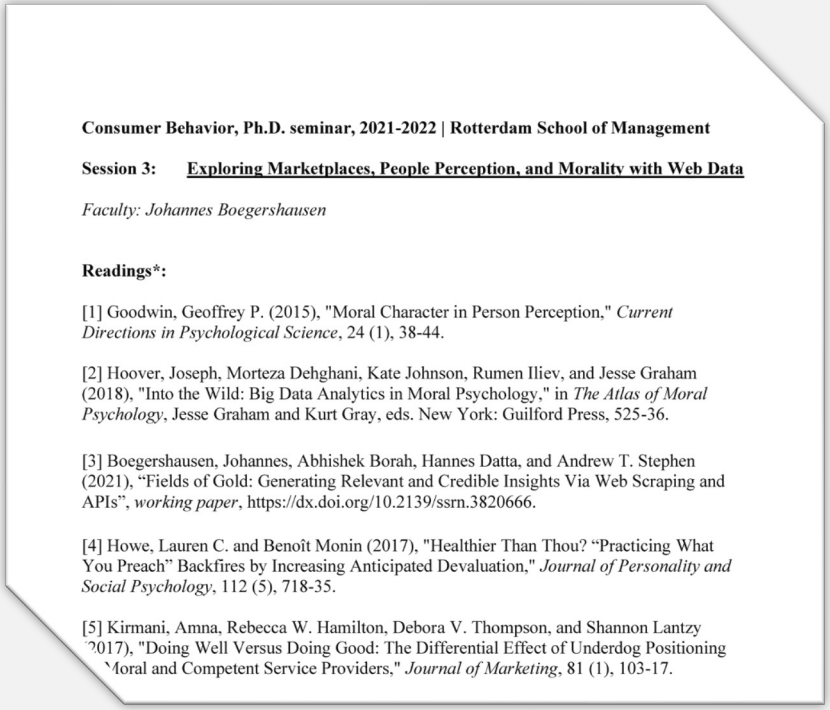
FOOD FOR THOUGHT

Our framework & companion website



- Explore our **database with 300+ published marketing articles** using web data.
- Discover **web datasets & APIs** for your research projects.
- **Tutorials and example code** for collecting web data using web scraping & APIs.

... for teaching



Consumer Behavior, Ph.D. seminar, 2021-2022 | Rotterdam School of Management

Session 3: **Exploring Marketplaces, People Perception, and Morality with Web Data**

Faculty: Johannes Boegershausen

Readings*:

[1] Goodwin, Geoffrey P. (2015), "Moral Character in Person Perception," *Current Directions in Psychological Science*, 24 (1), 38-44.

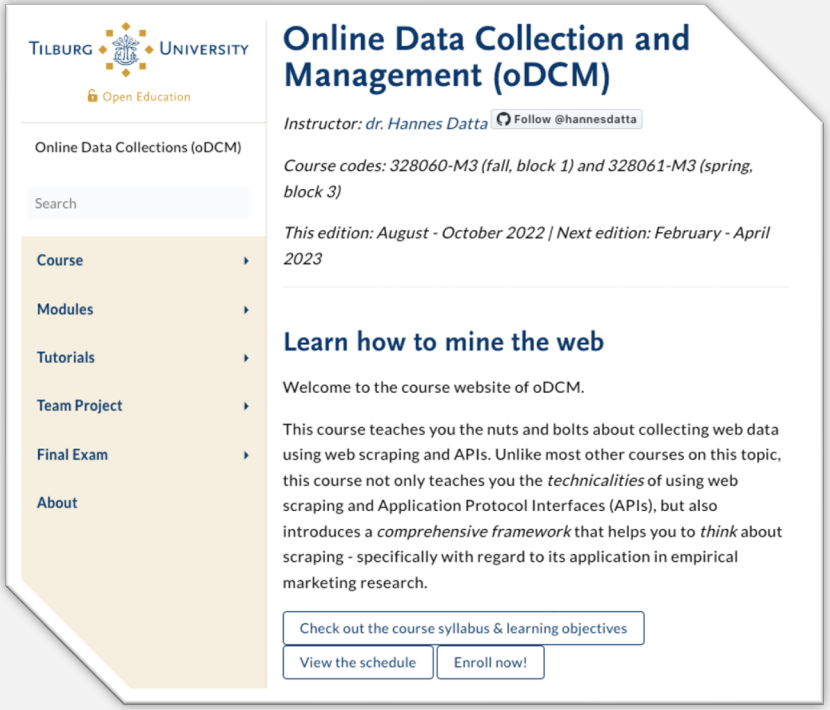
[2] Hoover, Joseph, Morteza Dehghani, Kate Johnson, Rumen Iliev, and Jesse Graham (2018), "Into the Wild: Big Data Analytics in Moral Psychology," in *The Atlas of Moral Psychology*, Jesse Graham and Kurt Gray, eds. New York: Guilford Press, 525-36.

[3] Boegershausen, Johannes, Abhishek Borah, Hannes Datta, and Andrew T. Stephen (2021), "Fields of Gold: Generating Relevant and Credible Insights Via Web Scraping and APIs", *working paper*, https://dx.doi.org/10.2139/ssrn.3820666.

[4] Howe, Lauren C. and Benoît Monin (2017), "Healthier Than Thou? "Practicing What You Preach" Backfires by Increasing Anticipated Devaluation," *Journal of Personality and Social Psychology*, 112 (5), 718-35.

[5] Kirmani, Amna, Rebecca W. Hamilton, Debora V. Thompson, and Shannon Lantzy (2017), "Doing Well Versus Doing Good: The Differential Effect of Underdog Positioning on Moral and Competent Service Providers," *Journal of Marketing*, 81 (1), 103-17.

Ph.D. seminars



TILBURG UNIVERSITY

Open Education

Online Data Collections (oDCM)

Search

Course

Modules

Tutorials

Team Project

Final Exam

About

Online Data Collection and Management (oDCM)

Instructor: dr. Hannes Datta Follow @hannesdatta

Course codes: 328060-M3 (fall, block 1) and 328061-M3 (spring, block 3)

This edition: August - October 2022 | Next edition: February - April 2023

Learn how to mine the web

Welcome to the course website of oDCM.

This course teaches you the nuts and bolts about collecting web data using web scraping and APIs. Unlike most other courses on this topic, this course not only teaches you the *technicalities* of using web scraping and Application Protocol Interfaces (APIs), but also introduces a *comprehensive framework* that helps you to *think* about scraping - specifically with regard to its application in empirical marketing research.

Check out the course syllabus & learning objectives

View the schedule | Enroll now!

method courses

… as an inspiration?

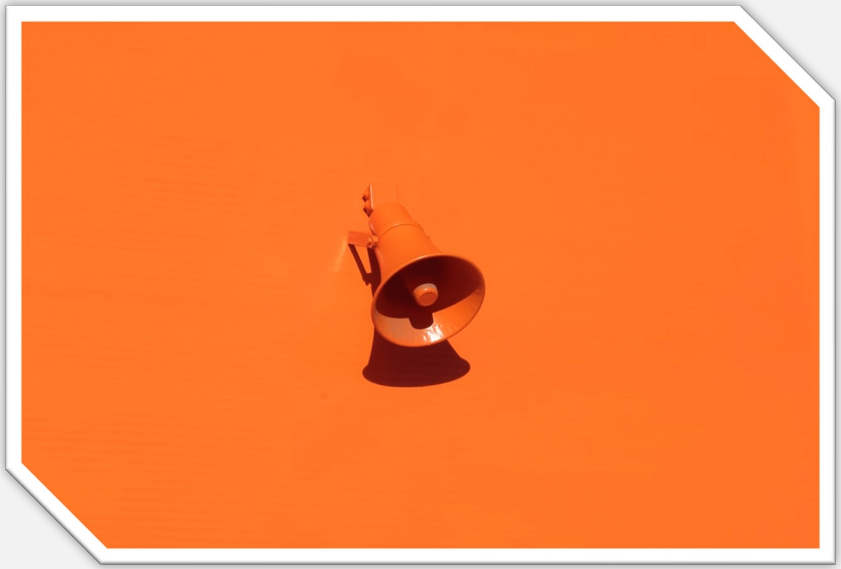
- If you want to learn scraping…

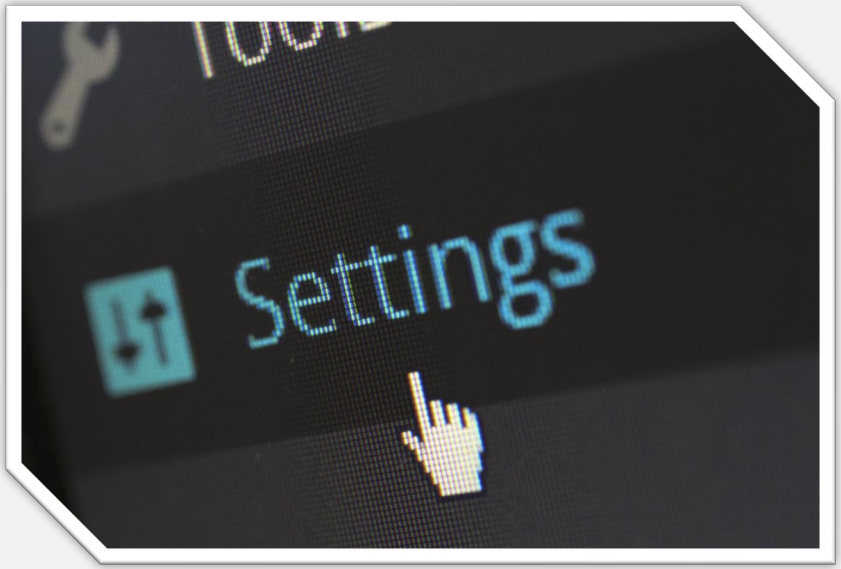
… as an inspiration?

- If you already work on a scraping/API project…

… as an inspiration?

- If you are an expert in working with web data …





Conclusion

- Webinar = primer
- Web (data) is here to stay (and grow)
- Important methodological tool for (early-career) researchers regardless of substantive or methodological focus.
- Many unexploited fields of gold! Potential for innovation!

Thank You &
Time for more questions

SLIDES AVAILABLE AT
https://web-scraping.org

boegershausen@rsm.nl | *@JoBoegershausen*
h.datta@tilburguniversity.edu | *@hannesdatta*
abhishek.borah@insead.edu
andrew.stephen@sbs.ox.ac.uk | *@AndrewTStephen*

https://www.ama.org/journal-of-marketing/

All forthcoming articles

Editorial Teams

Editors
Associate Editors
Editorial Review Board
Advisory Board
Meet *JM* at a Conference

JM Articles

Current Issue
Articles in Advance
Accepted Manuscripts
Go to the Journal Archive
AMA Member Access
Awards

Authors

Editorial Mission
Submission Guidelines
AMA Journal Policies
Manuscript Central
Appeal Policy

New editors began handling new manuscripts on February 1 and will assume full operational control on July 1.

JOURNAL of

Marketing

Hari Sridhar
Editor-in-Chief Designate
Texas A&M University

Cait Lamberton
Editor Designate
University of Pennsylvania

Detelina Marinova
Editor Designate
University of Missouri

Vanitha Swaminathan
Editor Designate
University of Pittsburgh