Web Scraping

AMA Marketing Strategy Consortium 2024
Cologne, 25 May 2024

Johannes Boegershausen
Erasmus University Rotterdam

Hannes Datta
Tilburg University

Abhishek Borah
INSEAD

Andrew Stephen
Oxford University

Jonne Guyt
University of Amsterdam

RSM Erasmus University · INSEAD The Business School for the World® · Tilburg University · University of Oxford Saïd Business School

Supported by MSI Marketing Science Institute

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
  - Integrating web data into conventional marketing research projects
- **Food for thought & conclusion**
  - Key insights
  - Exploiting new fields of gold
- **Q&A**

Introductory disclaimer

- By any means, we are really not the first (marketing) scholars to gather web data via scraping, APIs, etc.,
  - but we have used this in our own work + reviewed such research (extensively)
  - we have published two methodological papers about collecting web data at scale (Boegershausen, Borah, Datta, and Stephen 2022 ; Guyt, Datta, and Boegershausen 2024)
- There is no boilerplate template for gathering web data for academic research.
- When you feel that I am going too fast, please slow me down.
- **This is NOT a session about the technical details of automatically extracting web data. Why not?**
  - Content extraction is no longer the main problem (e.g., lower barriers to entry, higher technical feasibility)
  - Rather about helping you develop a **design mindset** for collecting web data at scale.
- This is **designed to be an interactive session**, so we might not get through all materials, but I will share extended slides and supporting docs (see also www.web-scraping.org)

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 **~ 265m** reviews

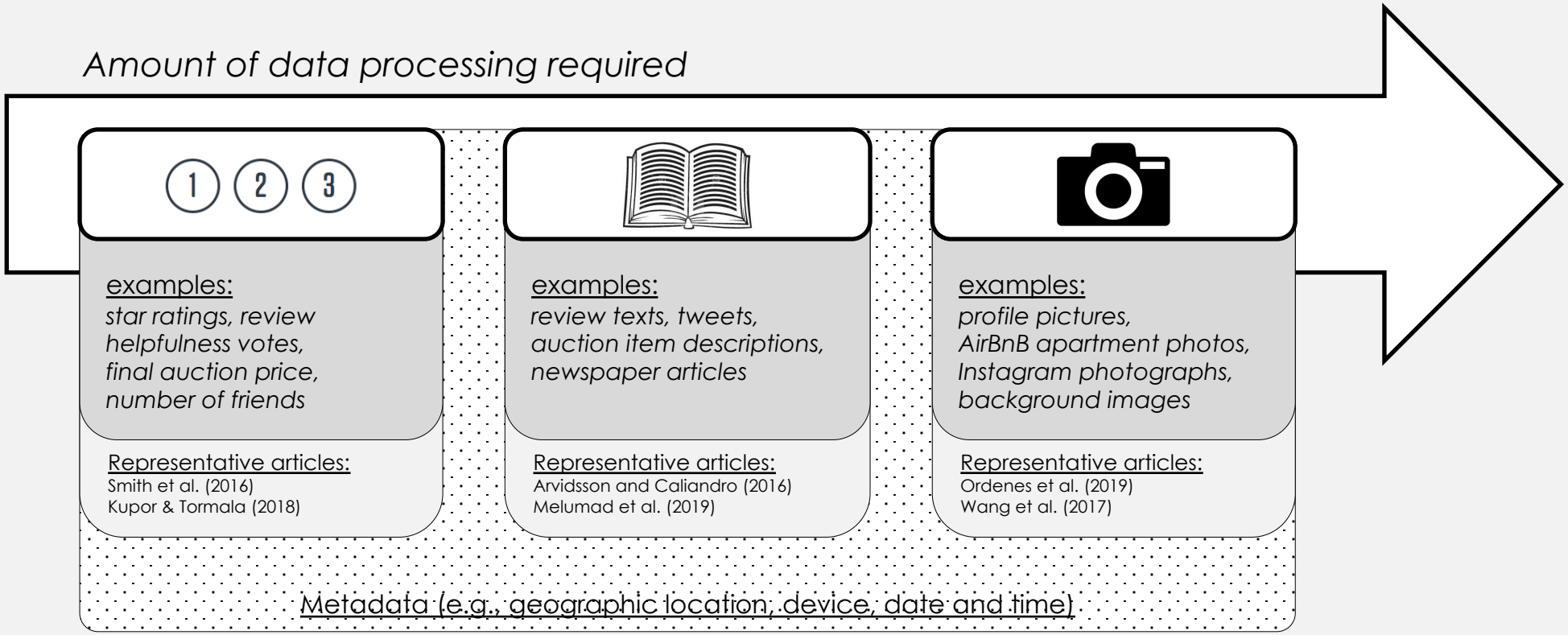
tripadvisor **> 1b** reviews & opinions

500m/day

funded with KICK STARTER **590K** projects

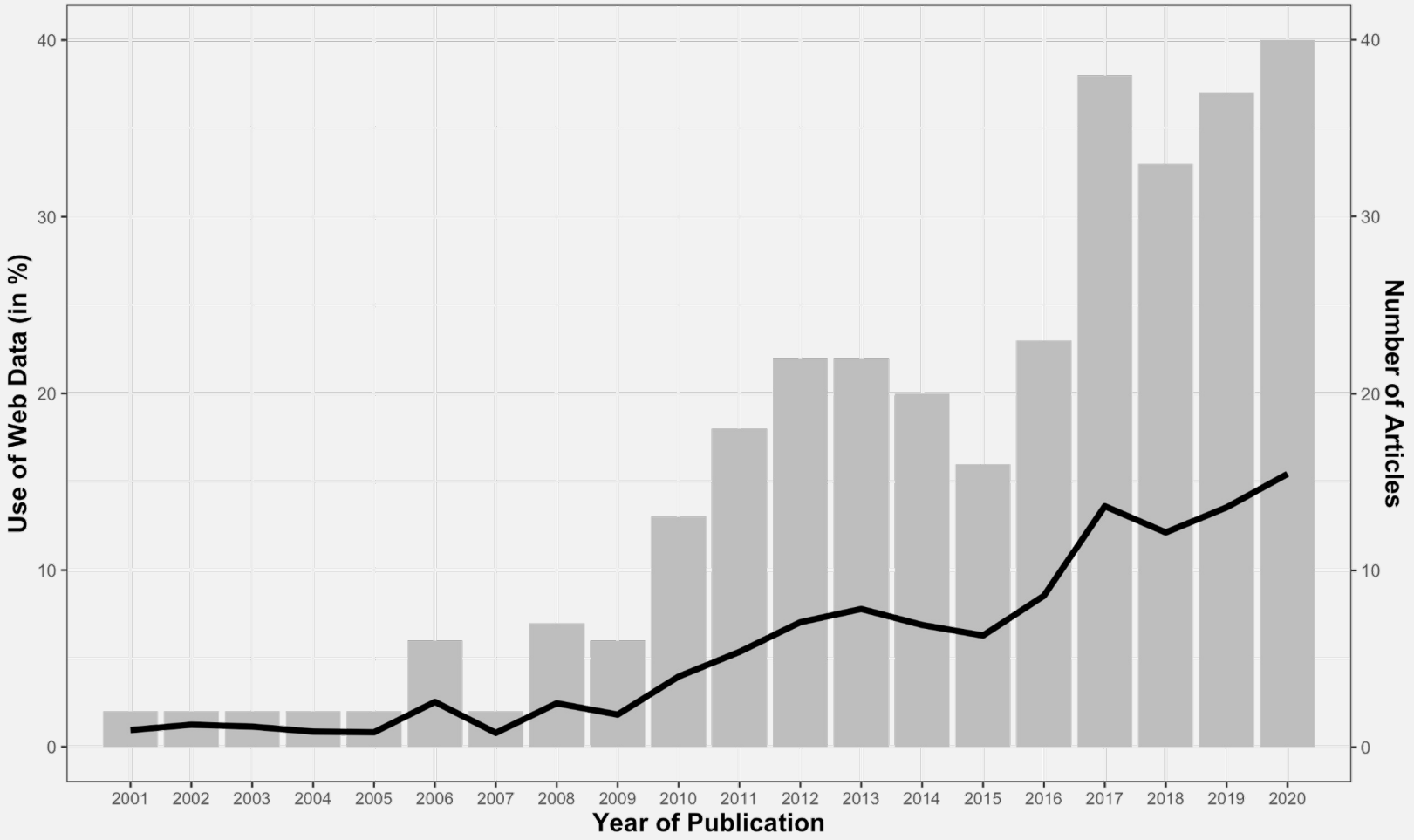
Stats last updated on 14 June 2023

Diverse types of data available



Source: Boegershausen, Borah, and Stephen (2020)

Increasing usage of web data in marketing research



+ 51 (2021)

+ 59 (2022)

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

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)

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)

Web scraping vs. APIs

- Web scraping captures any information displayable on publicly accessible websites
  - Automate interaction with the website (e.g., clicking, scrolling)
  - "Save" information that is displayed in a browser

- Application Programming Interfaces (APIs) make available at scale the official data sources of a firm
  - Data access is authorized, but only when explicitly made available
  - Data is already “structured”
  - Often involves a subscription or partnership

Highly versatile data collection technique

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

Highly versatile data collection technique

TEMU

SHEIN

TikTok

weedmaps

Source: Guyt, Datta, and Boegershausen (2024)

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)

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

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)

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

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
- Integration with other archival data
- *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?

How to sample?

Existing guidance limited

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

Algorithmic biases?

Which software to use?

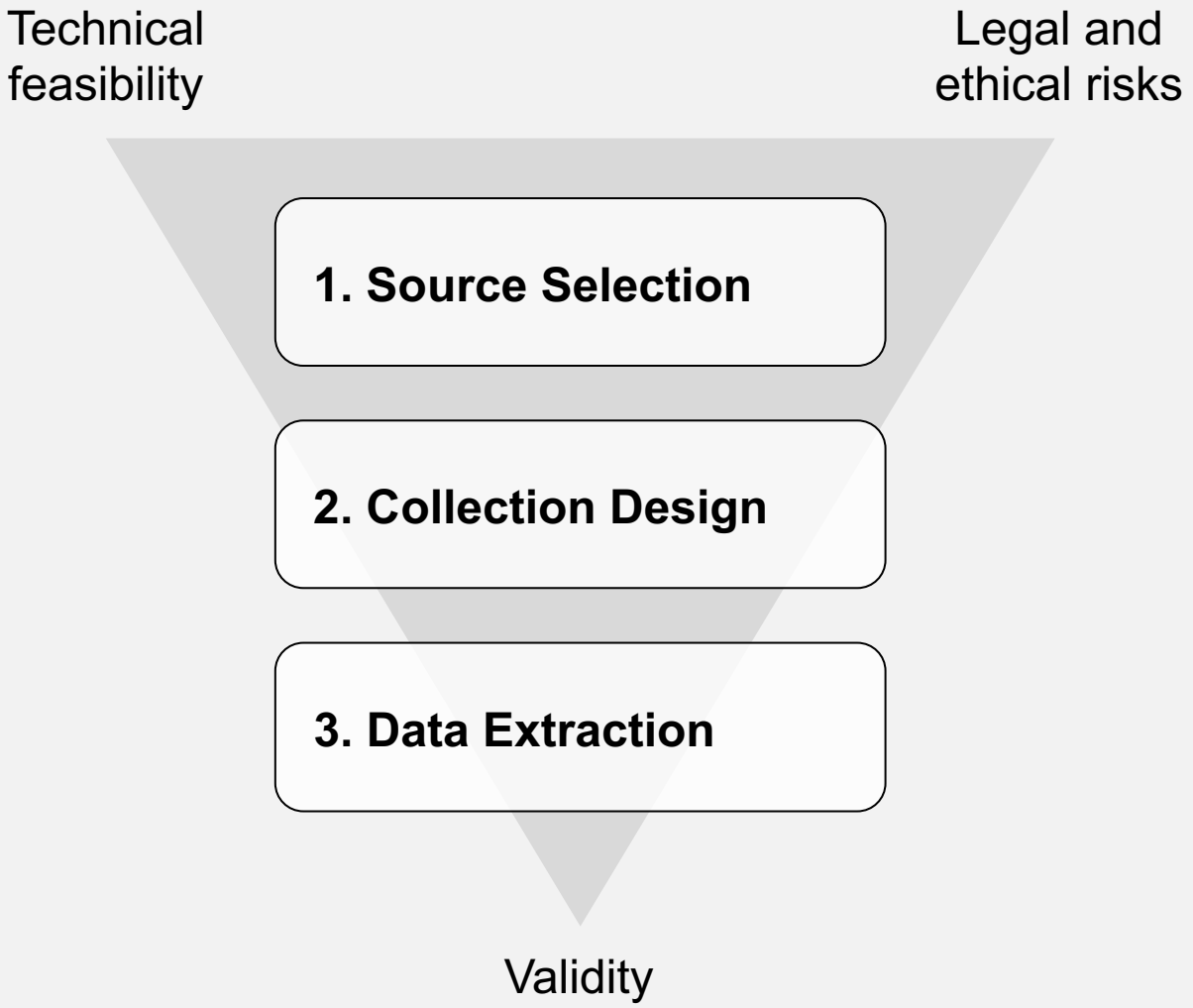
Extraction frequency?

Keep the raw (HTML, JSON) data?

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

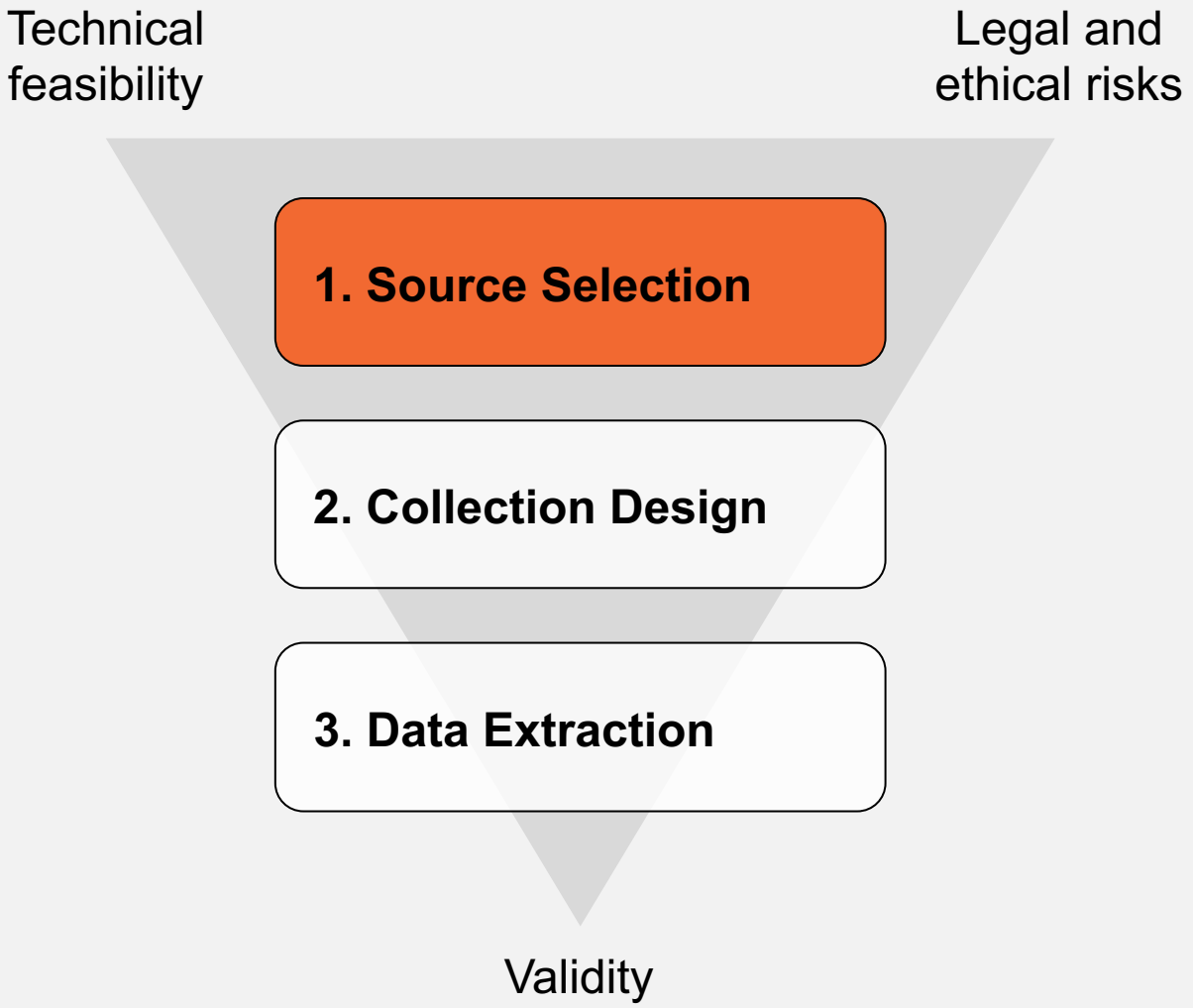
METHODOLOGICAL FRAMEWORK

Methodological framework



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

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

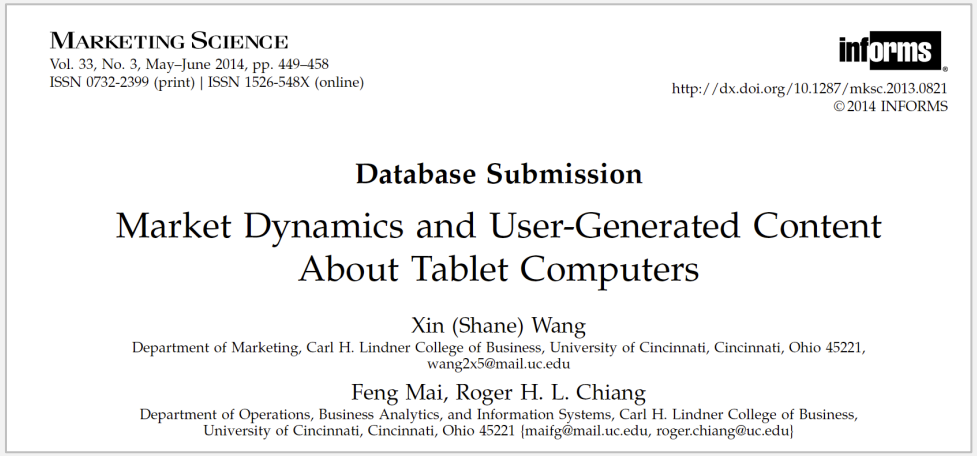
IMDb

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?

Source selection: recommendations II



MARKETING SCIENCE
Vol. 33, No. 3, May–June 2014, pp. 449–458
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

informs

http://dx.doi.org/10.1287/mksc.2013.0821

Database Submission

Market Dynamics and User-Generated Content About Tablet Computers

Xin (Shane) Wang
Department of Marketing, Carl H. Lindner College of Business, University of Cincinnati, Cincinnati, Ohio 45221, wang2x5@mail.uc.edu

Feng Mai, Roger H. L. Chiang
Department of Operations, Business Analytics, and Information Systems, Carl H. Lindner College of Business, University of Cincinnati, Cincinnati, Ohio 45221 {maifg@mail.uc.edu, roger.chiang@uc.edu}

https://osf.io/y2fh4/

- Kickstarter
  - https://webrobots.io/kickstarter-datasets/
  - https://www.icpsr.umich.edu/web/NADAC/studies/38050
- Indiegogo
  - https://webrobots.io/indiegogo-dataset/
- Common Crawl
  - https://commoncrawl.org/

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

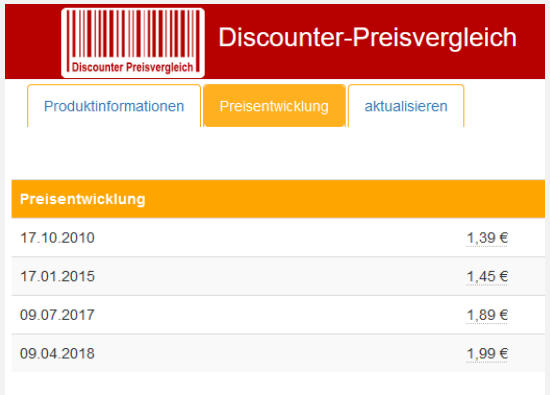
Source selection: combining datasets

- **How can we integrate web data into research projects relying on other archival datasets (e.g., NielsenIQ or GfK data)?**
  - Two key challenges:
    - a) temporal alignment
    - b) matching

Source: Guyt, Datta, and Boegershausen (2024)

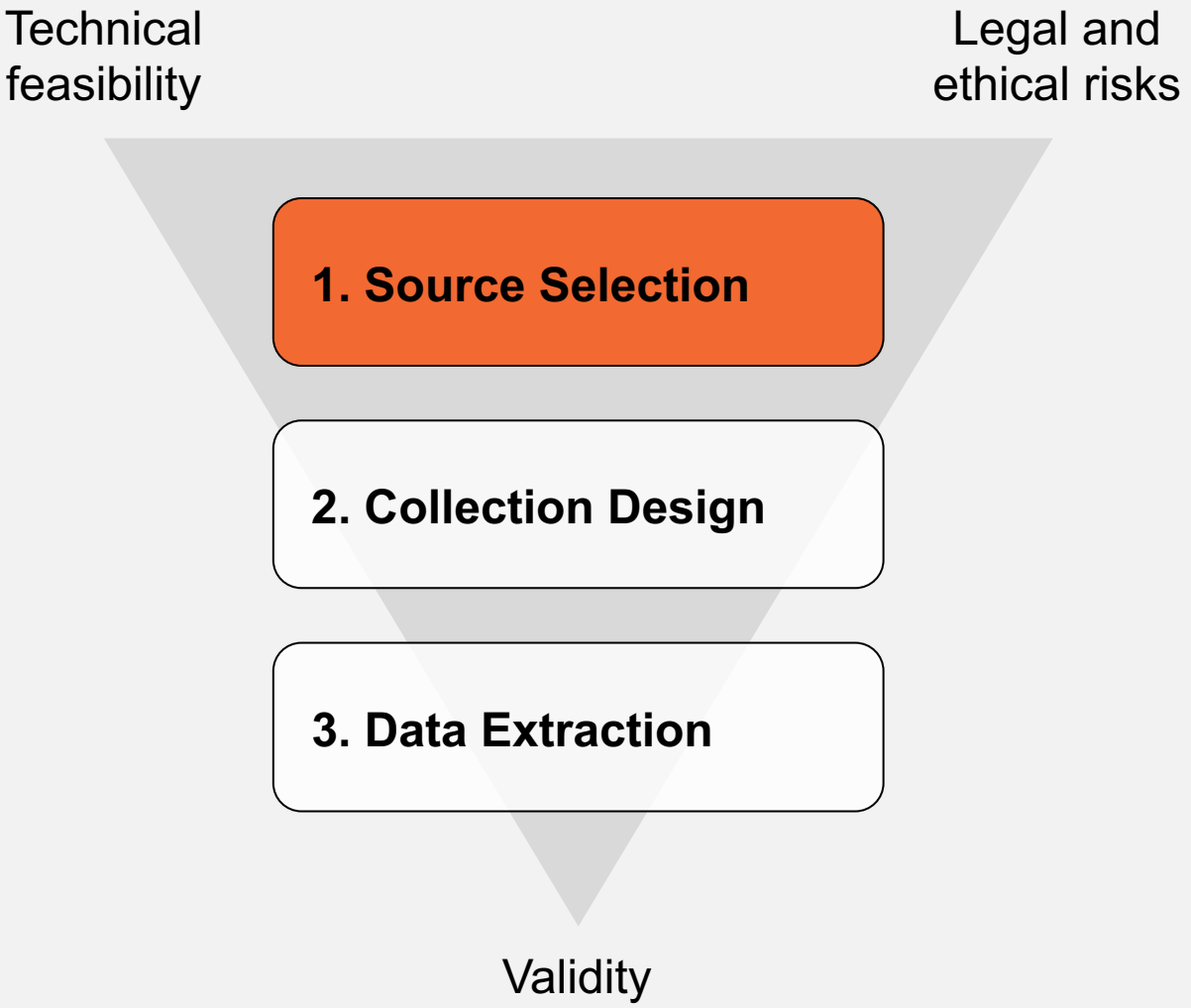
Source selection: combining datasets

- How can we integrate web data into research projects relying on other archival datasets (e.g., NielsenIQ or GfK data)?
  - Two key challenges:
    - a) temporal alignment
    - b) matching
  - **Potential solutions via source selection**
    - Surveying and extracting from aggregators (e.g., https://heisse-preise.io/)
    - Leveraging archival versions of web data (i.e., https://web.archive.org/)
    - Start scraping well ahead of the analysis

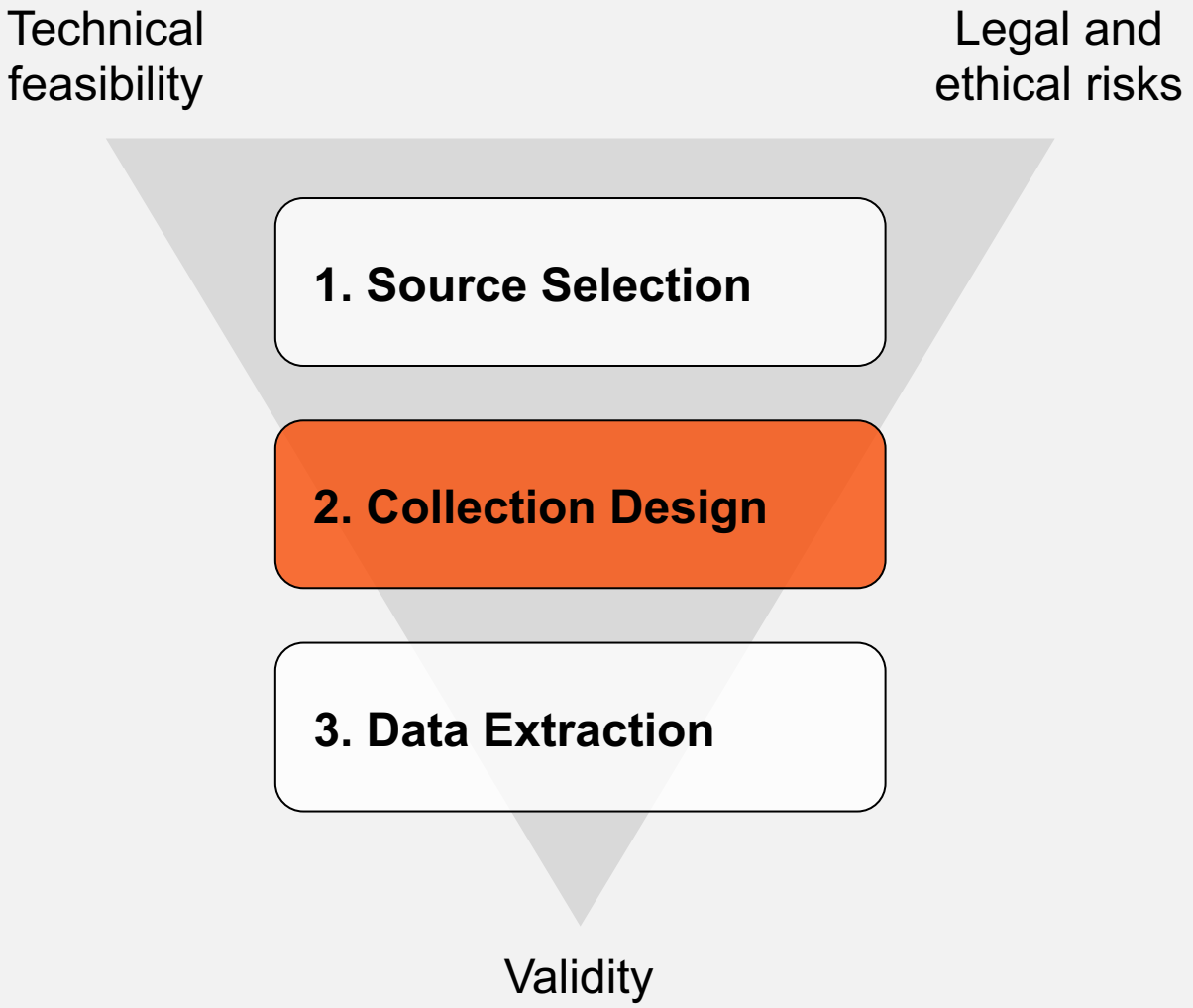


Source: Guyt, Datta, and Boegershausen (2024)

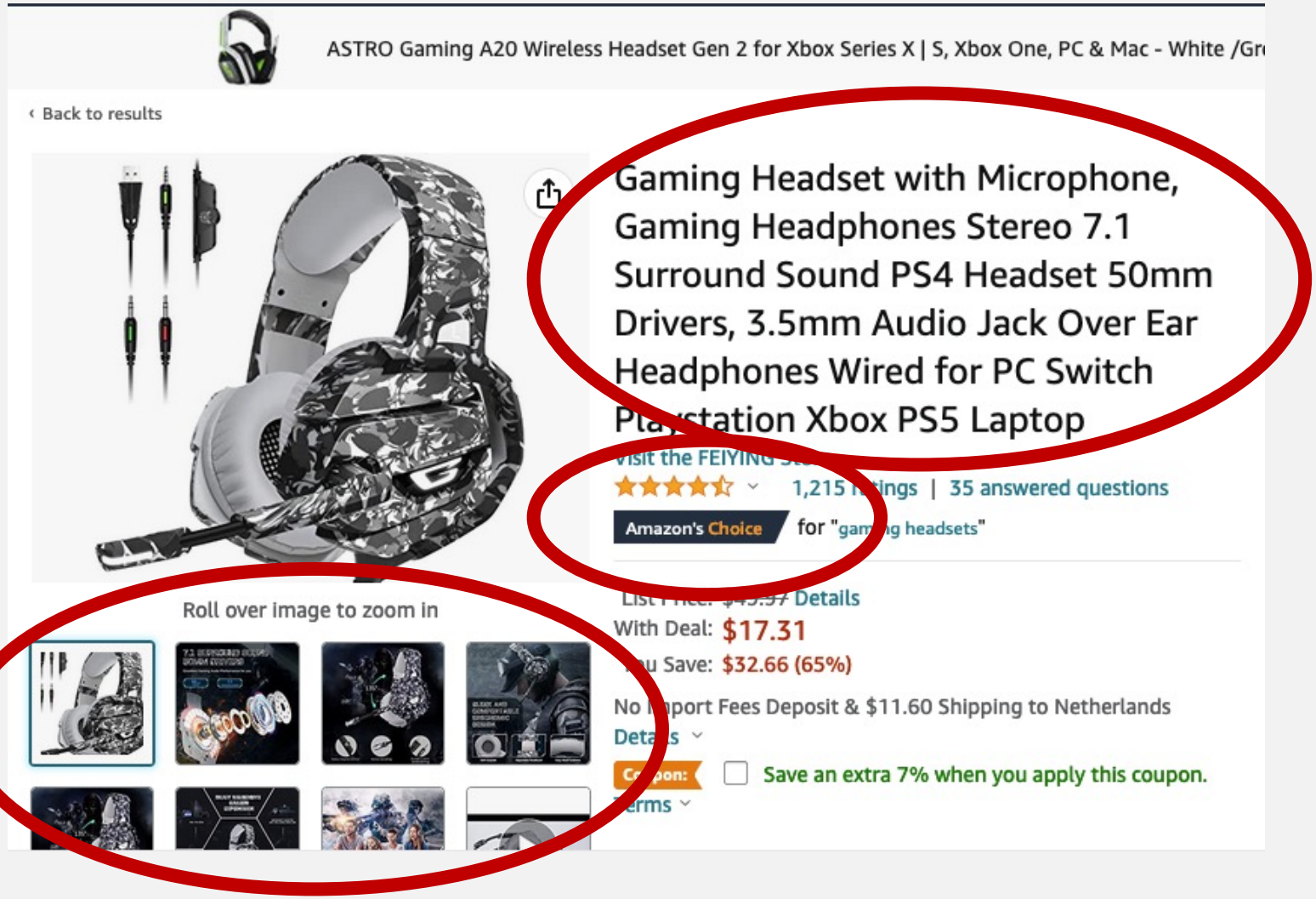
Questions?


Technical feasibility
Legal and ethical risks
1. Source Selection
2. Collection Design
3. Data Extraction
Validity

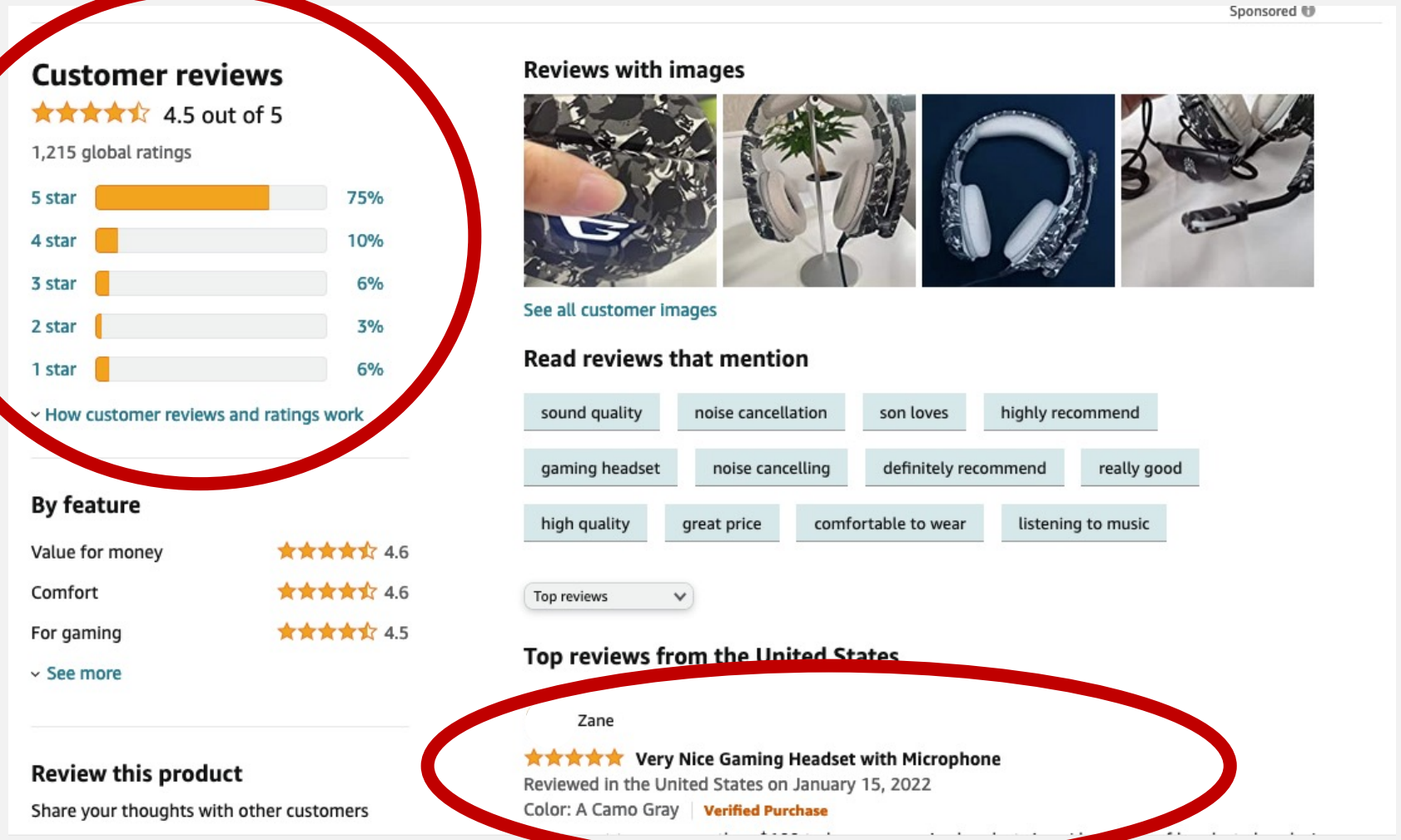
Designing the data collection



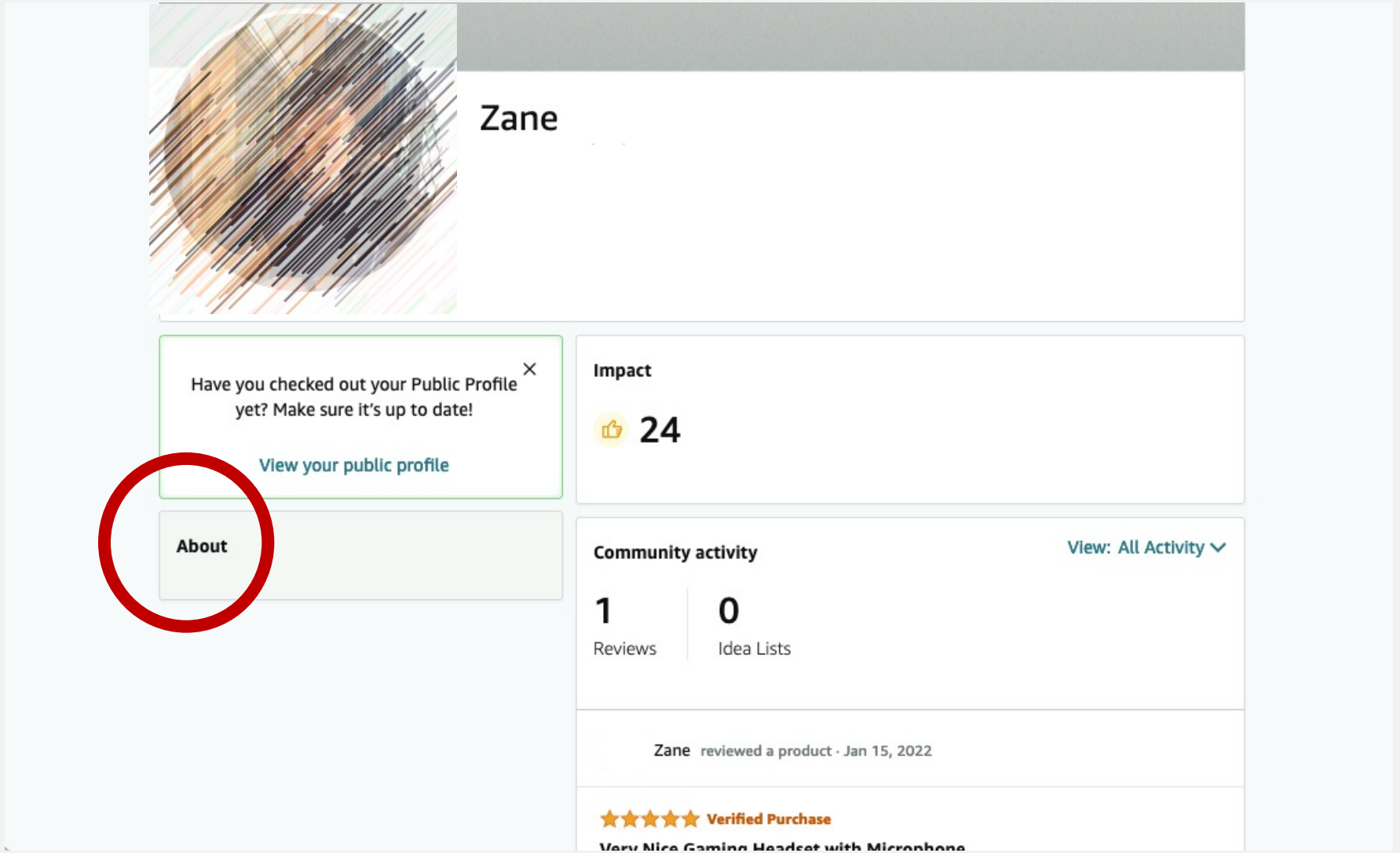
Which information to extract? Example



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

Technical feasibility

Which information to extract? Example

Validity implications

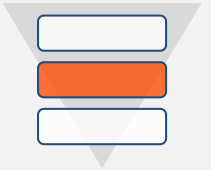
- Is information subject to algorithmic biases or missing data?
  Delete cookies & check?
- Are there significant changes to the data-generating process?
  Archive.org
- Is meta data required to make sense of variables?
  Save timestamps/IP addresses

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

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

Which information to extract? Philosophy

Targeted vs. Comprehensive

Table C1
Targeted vs. comprehensive web data collections.

	Targeted web data collection	Comprehensive web data collection
Description	Identifying elements of interest on the website and collecting those exhaustively (e.g., pricing of all products of a retailer) Focused research projects with clear research question(s) or predictions	Identifying areas of interest and collecting a large body of data from the landing page and X-degree links (e.g., retailer's landing page and pages linked on the landing page) Broad or explorative research projects or agendas where research questions are emerging

Source: Guyt, Datta, and Boegershausen (2024)

Which information to extract? Philosophy

Targeted vs. Comprehensive

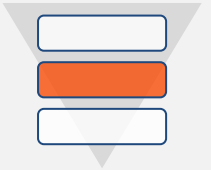
Table C1
Targeted vs. comprehensive web data collections.

	Targeted web data collection	Comprehensive web data collection
Description	Identifying elements of interest on the website and collecting those exhaustively (e.g., pricing of all products of a retailer)	Identifying areas of interest and collecting a large body of data from the landing page and X-degree links (e.g., retailer's landing page and pages linked on the landing page)
	Focused research projects with clear research question(s) or predictions	Broad or explorative research projects or agendas where research questions are emerging
Data coverage	High (allows for collecting more depth)	Low - Moderate (allows for breadth but not depth)
Resource intensity (e.g., computing infrastructure, storage)	Low (only elements of interest)	Moderate – High (includes many and potentially large files, e.g., pictures)[1]
Flexibility to change data collection plan	Low (only if no alternative data is needed)	Moderate – High (can exploit naturally occuring unanticipated policy changes)
Robustness to environmental changes (e.g., website layout changes)	Low (elements or layout may change)	High (code is not prone to changes in web format)
Ease of matching (e.g., based on EAN)	Low (no additional characteristics for matching)	Moderate (additional data may provide matching identifiers)
Ability to create additional control variables (e.g., in review process)	N/A (no additional data collected)	Moderate – High (additional data may contain applicable information)

Notes:
[1] A trade-off that researchers can make is to exclude saving images, vastly reducing the disk space required but foregoing the possibility of visual analytics at a later point.

Source: Guyt, Datta, and Boegershausen (2024)

How to process data during the extraction?



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

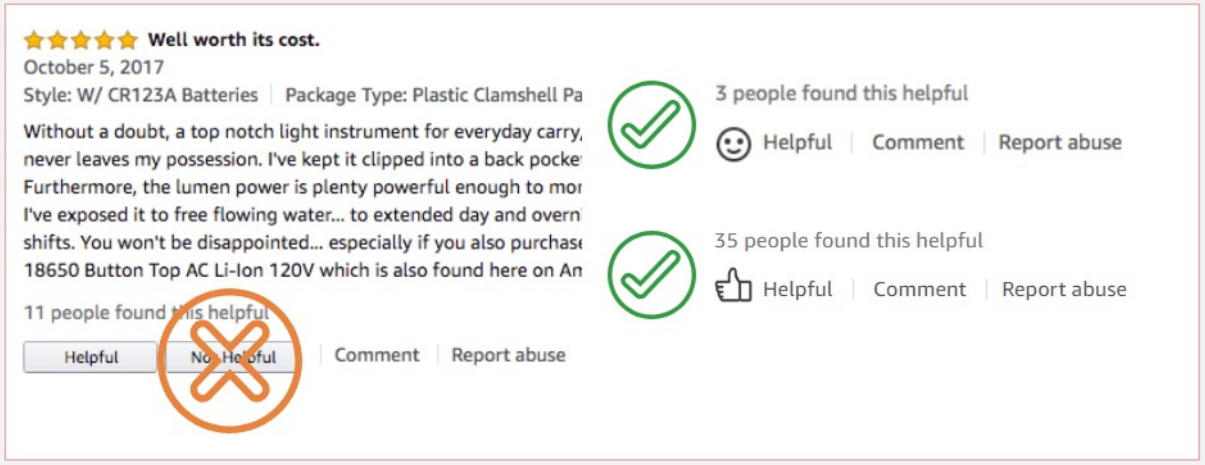
→ **Keep the raw data whenever possible**

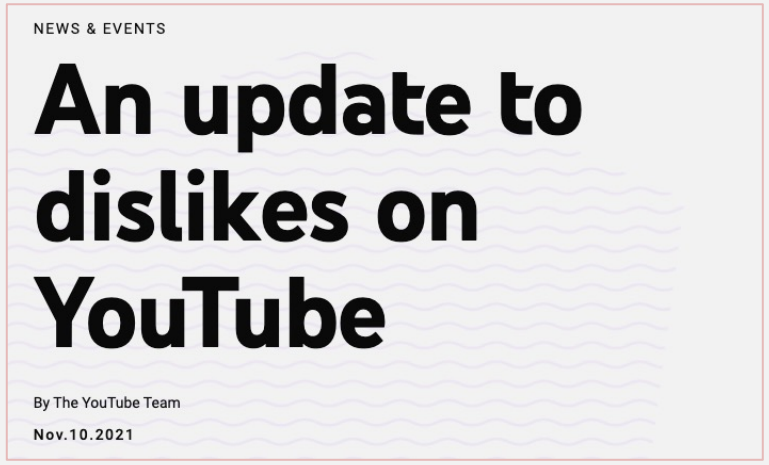
How to process data during the extraction?

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

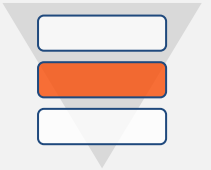
→ Keep the raw data whenever possible

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





How to sample? Challenges & considerations



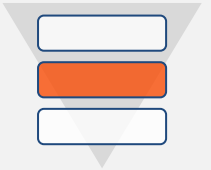
- How to capture the entire population (or a sample) of…?
  - Internal pages (e.g., bestseller, category, search page)
  - Externally available lists?
- Sampling frames (might) create different datasets or even induce systematic biases
- Which sample size is technically feasible?

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?

At what frequency to extract data? Challenges



- What are **your essential assumptions** about the configuration, data-generating process, and characteristics of the data to test predictions?

  Recursive process of *formulating a* ***"data source theory"*** outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

At what frequency to extract data? Example

- What are **your essential assumptions** about the configuration, data-generating process, and characteristics of the data to test predictions?

  Recursive process of *formulating a* ***"data source theory"*** outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

- Case study:
  Prediction: # friends on Yelp → usage of emotional language in reviews (+)
  Sample: all reviews of the 5 most reviewed Japanese restaurants in 5 US cities (NYC, LA, SF, CHI, DC)

User A
(scraped today)

300 friends
437 reviews
775 photos
Elite '2019

User A's review in our dataset
(scraped today)

Sushi House
$$ · Japanese, Sushi Bars

1/26/2014

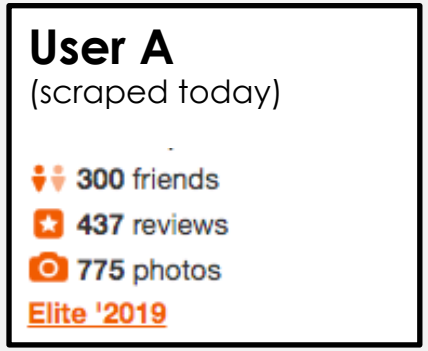
Any issues here?

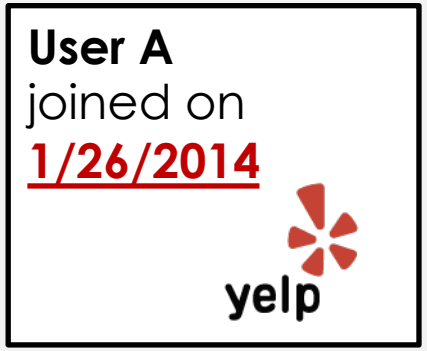
At what frequency to extract data? Example

- What are **your essential assumptions** about the configuration, data-generating process, and characteristics of the data to test predictions?

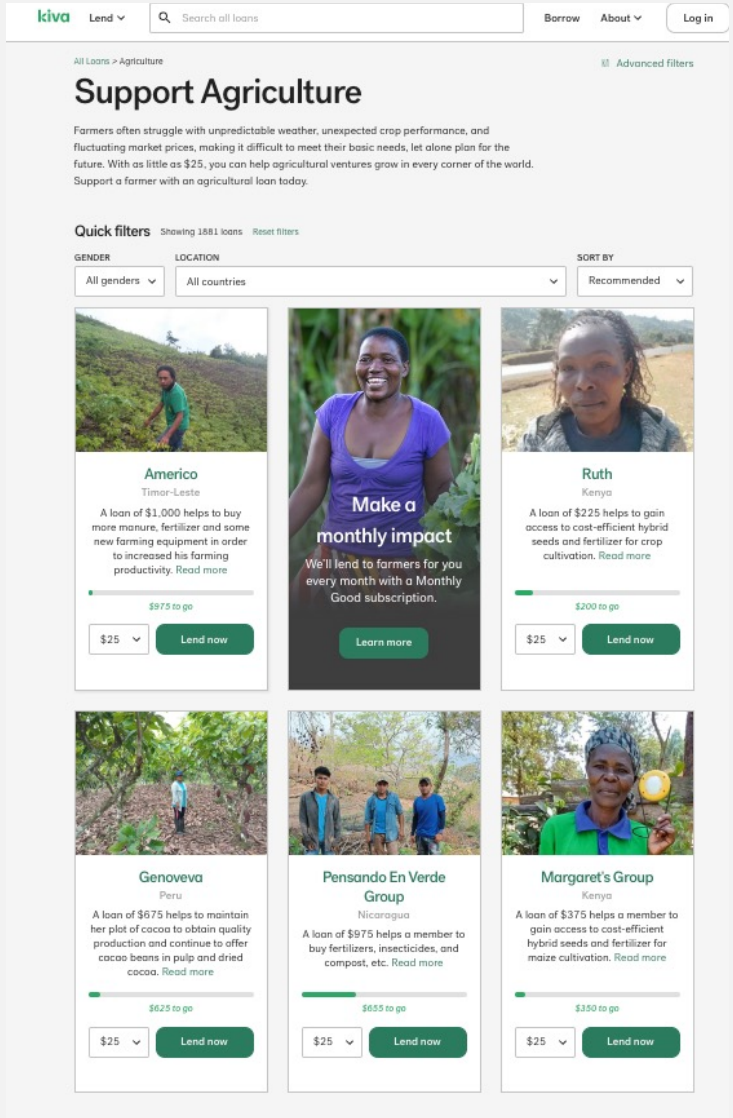
  Recursive process of *formulating a* ***"data source theory"*** outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

- Case study:
  Prediction: # friends on Yelp → usage of emotional language in reviews (+)
  Sample: all reviews of the 5 most reviewed Japanese restaurants in 5 US cities (NYC, LA, SF, CHI, DC)





Collection design: Kiva



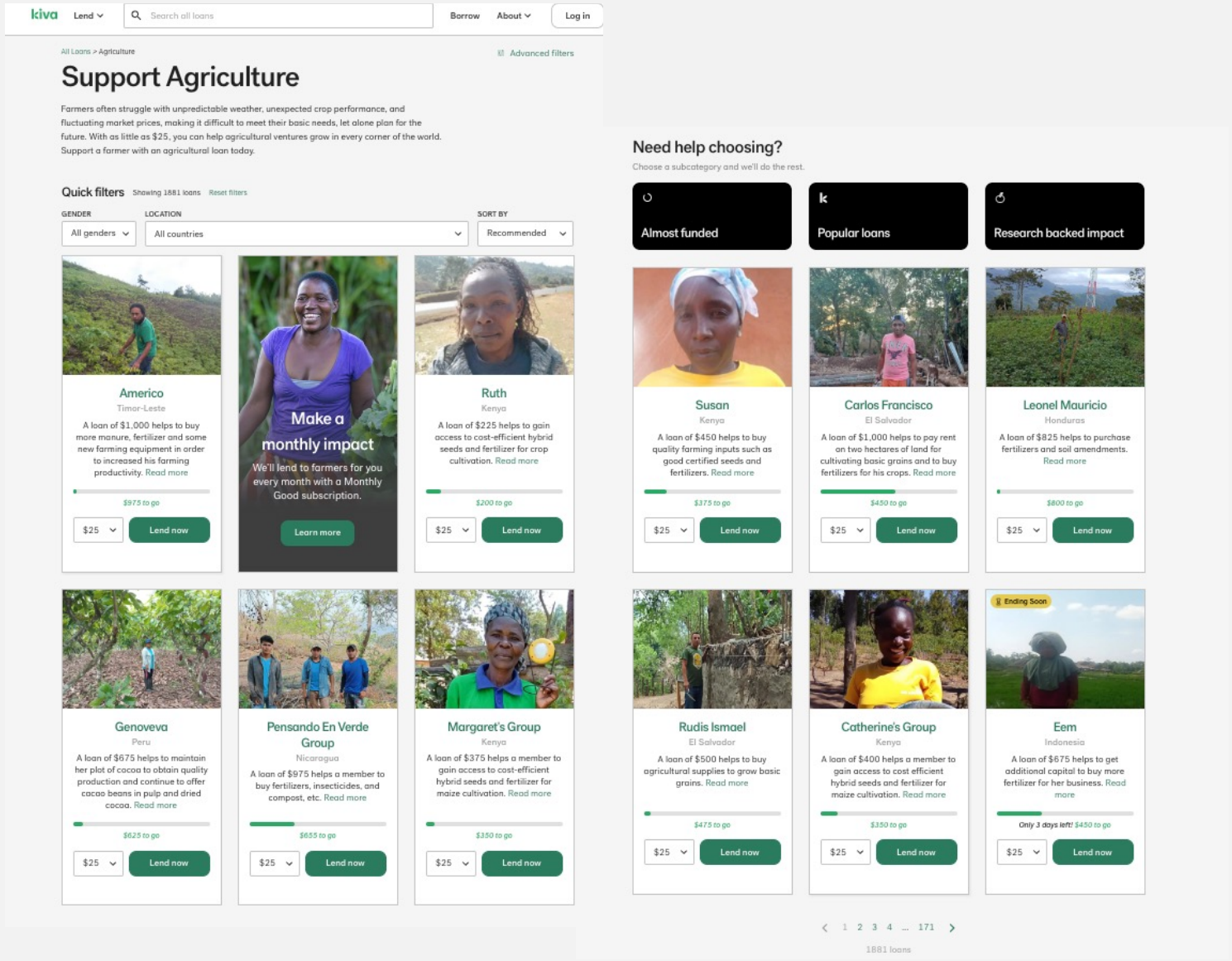
kiva

Research question:

How does distance from the project goal for a charitable project affect donation likelihood?

→ check the GraphiQL:
https://api.kivaws.org/graphql

Collection design: Kiva



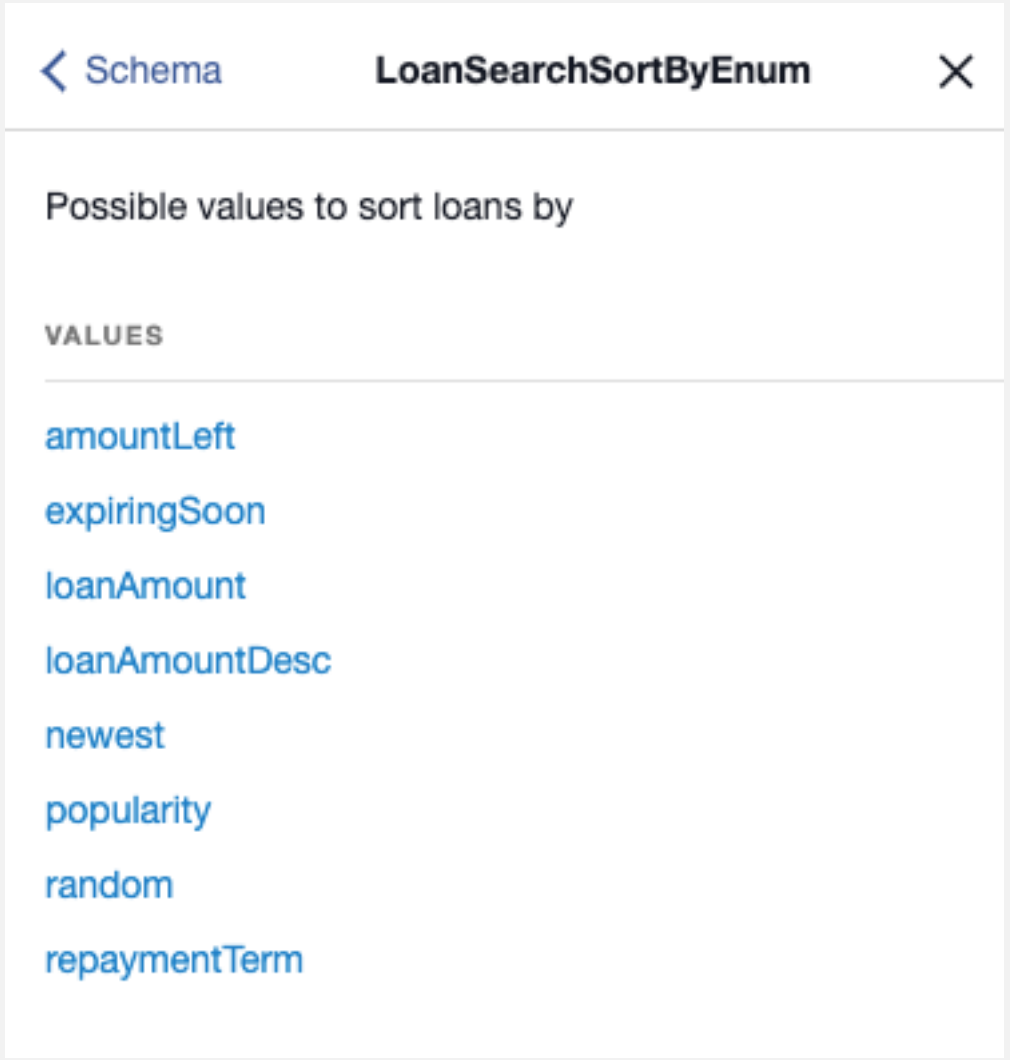
kiva

Research question:

How does distance from the project goal for a charitable project affect donation likelihood?

→ check the GraphiQL: https://api.kivaws.org/graphql

Collection design: *algorithmic interference*



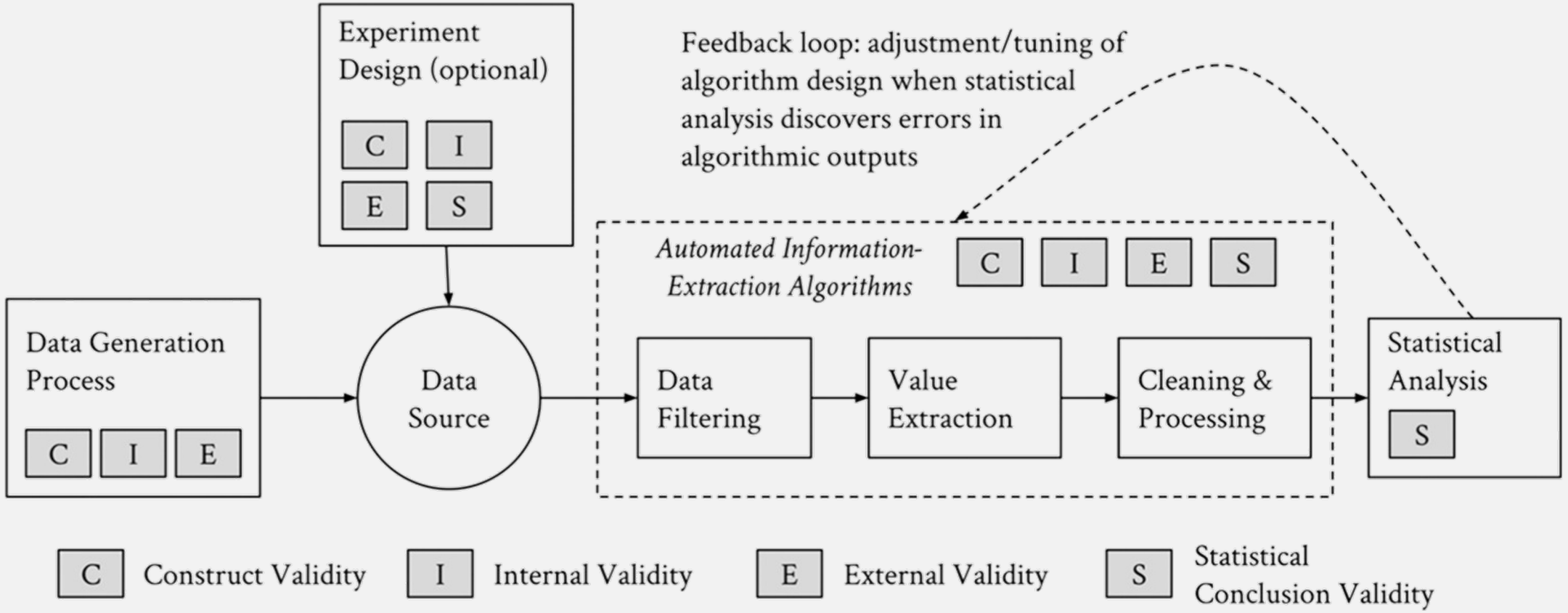
kiva

- The algorithm used by Kiva rewards projects that receive funds by **improving their presentation order**

- **Potential solution:** **capture presentation order via the API and control for it**

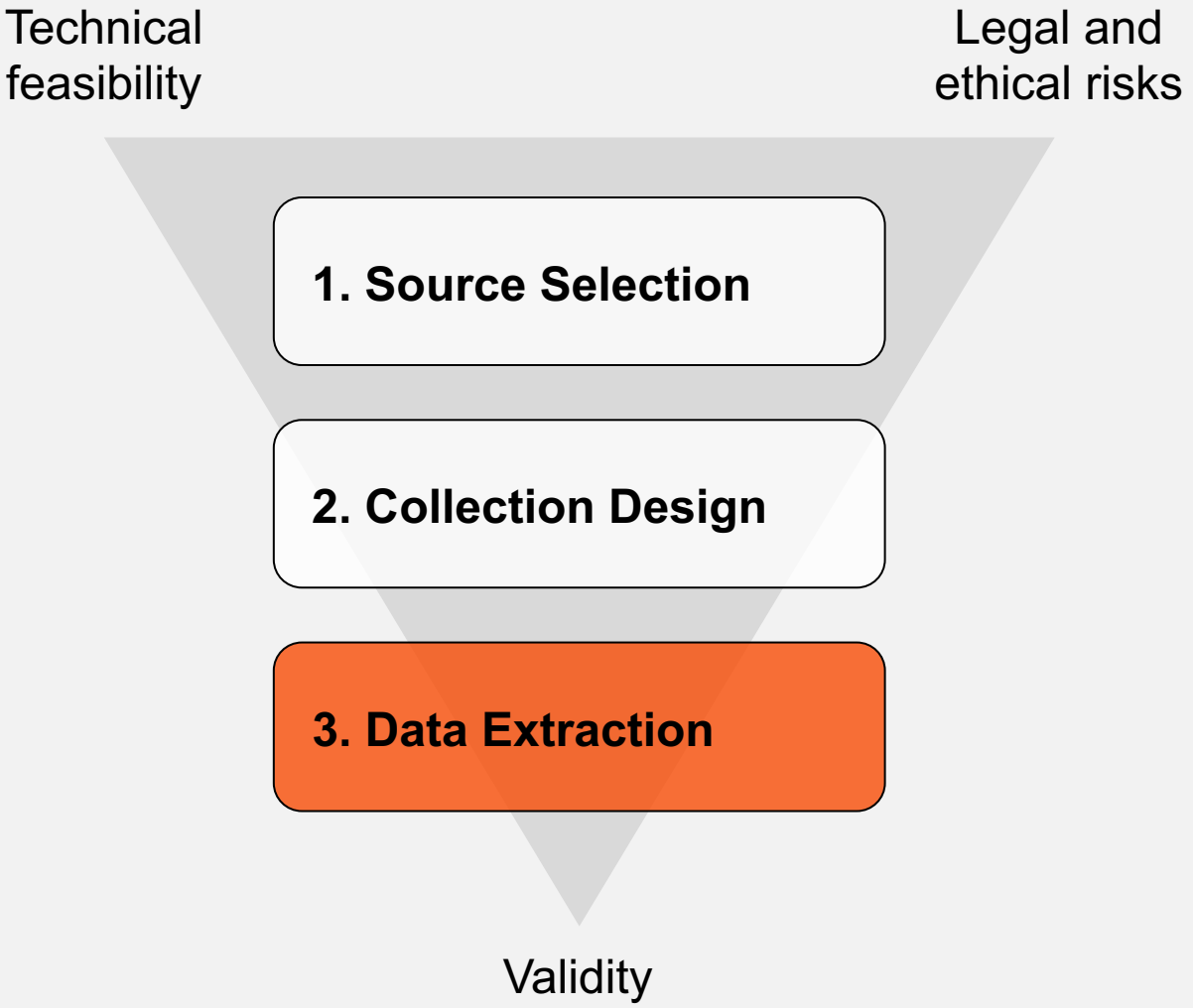
- *"popularity"*

Data-generating mechanism



Source: Xu et al. (2020)

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

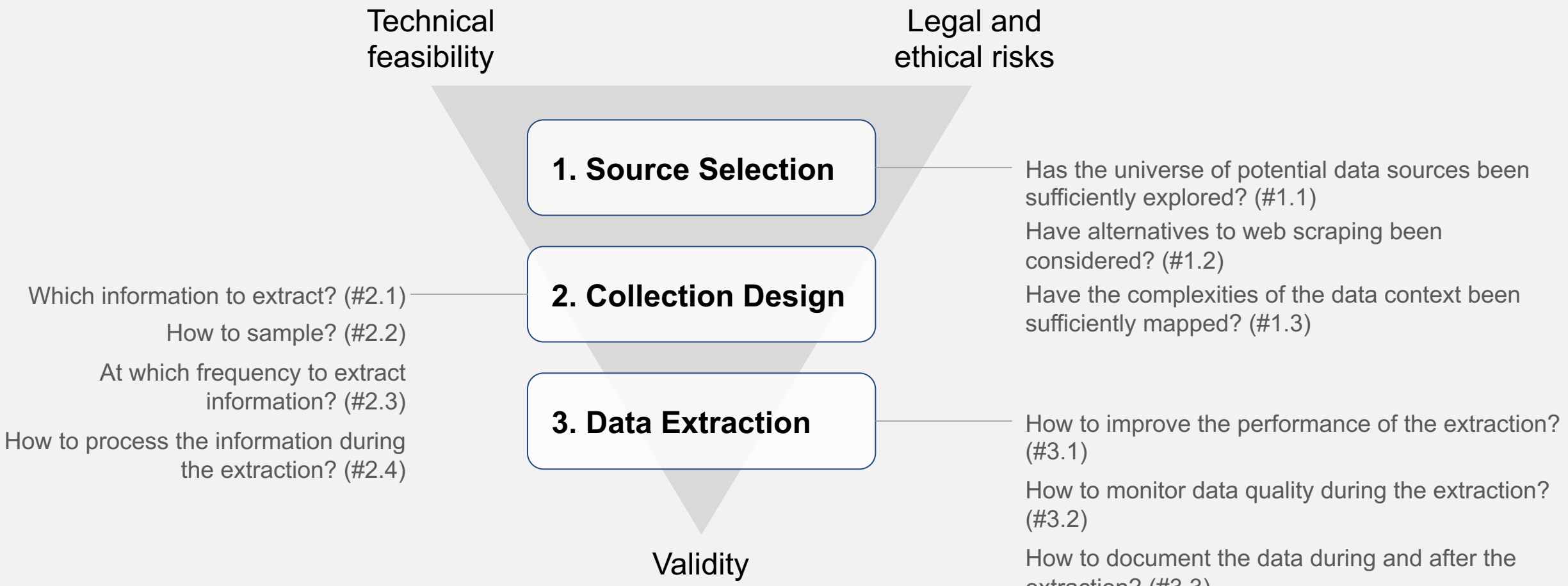
Data extraction

- 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

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]

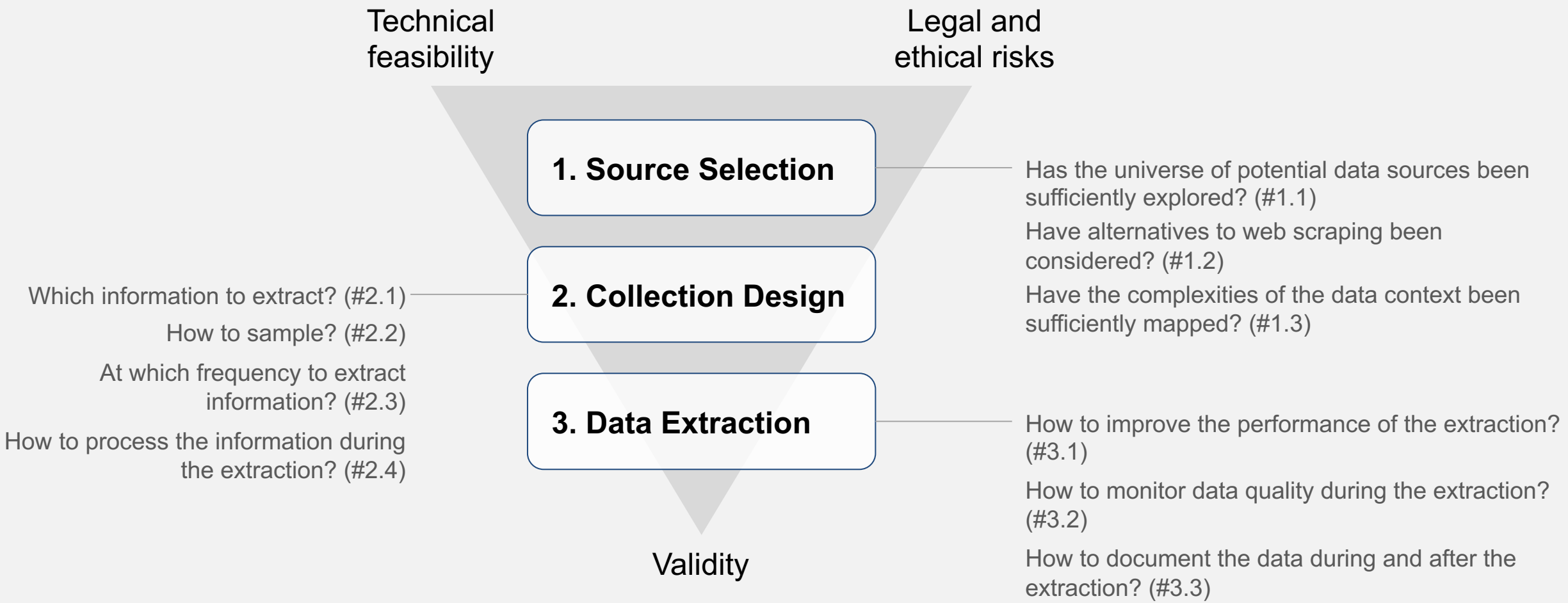
Challenge #2.4: How to process the data during the collection?

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

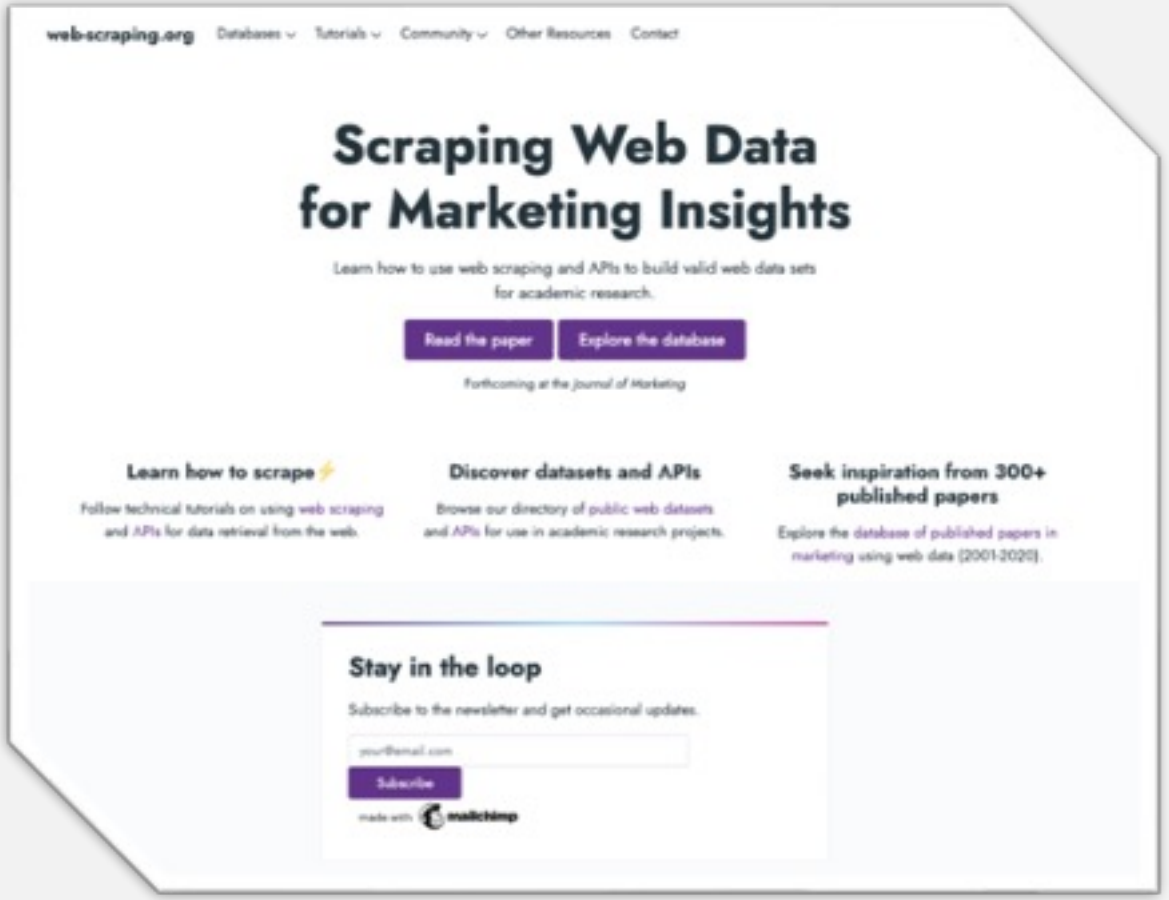
Future research using web data (I)

- **Shining light on new phenomena and developing new metrics**
  - Organizing product assortments, visual representation of products (and recipes)
  - Competition between grocery retailers and on-demand (fast) delivery retailers such as Getir
- **Improving and extending measurement of existing concepts**
  - online promotions, product recalls
  - alternative metrics for advertising
  - brand metrics / perceptual maps based on LLMs
- **Discover, document, and share data sources that complement traditional data sources**
  - potential sources of rich and complementary data (e.g., HolidayAPI)
  - accepted measures using public data (to facilitate analysis with exclusively public data)

Future research using web data (II)

- Creating workflows to make raw data usable (e.g., to allow for matching)
  - e.g., data pipelines to convert images to UPCs
- **Collective scraping with a consortium of research groups**
  - e.g., monitoring product characteristics for retailers by country/category
  - e.g., building a nutritional value database
  - e.g., monitoring digital platforms
- Joint training and R&D
  (longitudinal collections, infrastructure, how to scrape apps)

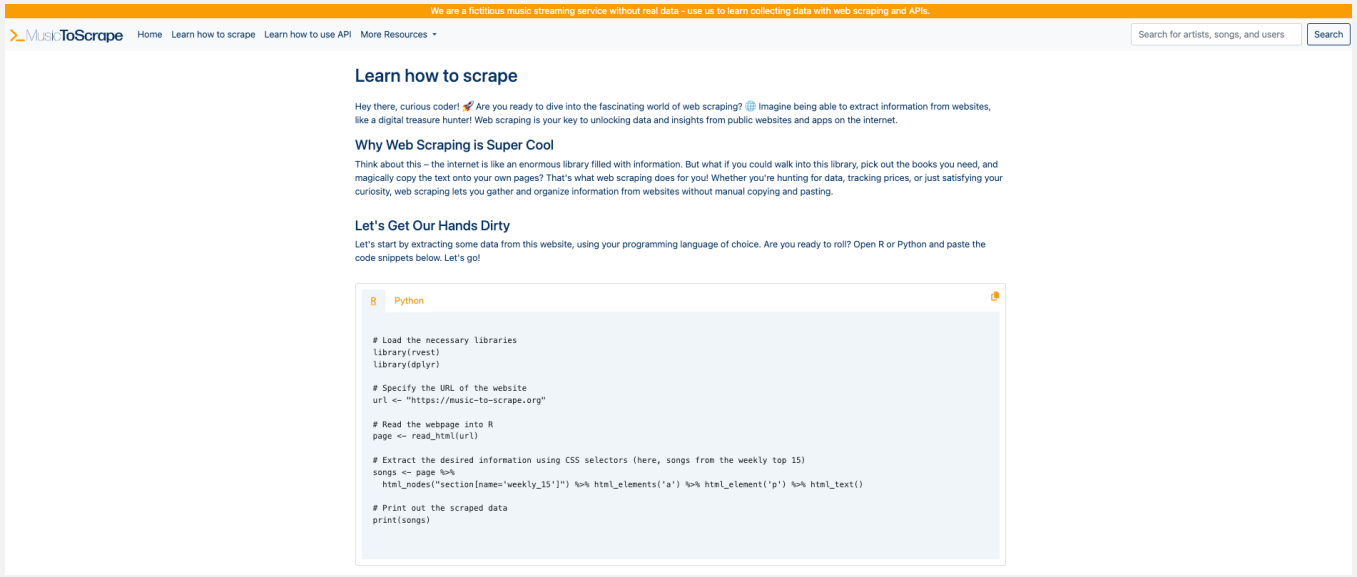
Our framework & companion website



- 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.

A research-focused tutorial

MusicToScrape



ELSEVIER
Check for updates
Available online at www.sciencedirect.com
ScienceDirect
Journal of Retailing 100 (2024) 130–147
www.elsevier.com/locate/jretai

Unlocking the Potential of Web Data for Retailing Research ☆

Jonne Y. Guyt[a,*], Hannes Datta[b], Johannes Boegershausen[c]
[a] Amsterdam Business School, University of Amsterdam, the Netherlands
[b] Tilburg School of Economics and Management, Tilburg University, the Netherlands
[c] Rotterdam School of Management, Erasmus University Rotterdam, the Netherlands

Available online 4 March 2024

Consider joining oDCM

Tilburg University
Open Education

Online Data Collections (oDCM)

Search

- Course
- Schedule and Course Material
- Team Project
- Exam
- About

Online Data Collection and Management (oDCM)

Instructor: dr. Hannes Datta LinkedIn GitHub Follow @hannesdatta

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

This edition: January - April 2024 | Next edition: August - October 2024

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. You learn about the *technicalities* of web scraping and Application Programming Interfaces (APIs), but also how to *design* your data collection for use in empirical (marketing) research projects[1].

Please use the navigation bar and buttons below to access the course material.

Course schedule and modules | Syllabus & learning objectives | Enroll now!

Why should you take this course?

Ever had this fantastic research idea but didn't have the data to study it?

Well, it's high time you get familiar with web data gathered by web scraping and APIs! Web data is ideally suited to study questions for which commercial datasets are either unavailable or too expensive to purchase. Yes, we're talking TikTok, Instagram, Socialblade - literally any website you can open in your browser!

Online Data Collection and Management (oDCM)

Learn how to mine the web

https://www.dropbox.com/scl/fo/1x0d4qk5ct5oihwok9ghy/h?rlkey=twxvthr1oflfig8wcx37of2k3&e=1&dl=0

Food for thought

- If you want to learn scraping
  - start in the language of your choice / use available code on the internet / don't assume this is "easy" – find a "buddy" to learn it with
  - consider alternative ways to access web data (e.g., datasets)

Food for thought

- If you want to learn scraping
  - start in the language of your choice / use available code on the internet / don't assume this is "easy" – find a "buddy" to learn it with
  - consider alternative ways to access web data (e.g., datasets)
- If you already work on a scraping/API project
  - Use paper to defend choices (e.g., data source selection)
  - Look beyond what is immediately visible on the web (longitudinal over time with *many* sources, plus ideas in paper)

Food for thought

- If you want to learn scraping
  - start in the language of your choice / use available code on the internet / don't assume this is "easy" – find a "buddy" to learn it with
  - consider alternative ways to access web data (e.g., datasets)
- If you already work on a scraping/API project
  - Use paper to defend choices (e.g., data source selection)
  - Look beyond what is immediately visible on the web (longitudinal over time with *many* sources, plus ideas in paper)
- If you're a scraping expert
  - Show the importance of this data in your research group: recognize the value of data, and how difficult it is to replace
  - Embark on ambitious multi-source projects

Conclusion

- Web (data) is here to stay (and grow)
  - When I was a PhD student, few people really cared about web data
  - Today, important methodological tool for (early-career) researchers regardless of substantive or methodological focus

- But…
  - Many challenges (technical, legal/ethical and validity)
  - Many unexploited fields of gold! Potential for innovation

- Next steps
  - Developing scraping as a research method (data sources, reproducibility, legal compliance)

Want to talk more?
Stay in touch?!

johannes-boegershausen

boegershausen.net

web-scraping.org

@JoBoegershausen

boegershausen@rsm.nl

Note:
The subsequent annotated reference list might not be fully complete as I continue to update and expand this deck. Yet, I feel it might be useful to share it with you. You will find annotations highlighting papers I consider of special or outstanding interest.

References

* of special interest
** of outstanding interest

Adjerid, Idris and Ken Kelley (2018), "Big Data in Psychology: A Framework for Research Advancement," American Psychologist, 73 (7), 899-917.

Aguinis, Herman, Wayne F. Cascio, and Ravi S. Ramani (2017), "Science's Reproducibility and Replicability Crisis: International Business Is Not Immune," Journal of International Business Studies, 48 (6), 653-63.

Arvidsson, Adam and Alessandro Caliandro (2016), "Brand Public," *Journal of Consumer Research*, 42 (5), 727-48.

Barnes, Christopher M., Carolyn T. Dang, Keith Leavitt, Cristiano L. Guarana, and Eric L. Uhlmann (2018), "Archival Data in Micro-Organizational Research: A Toolkit for Moving to a Broader Set of Topics," *Journal of Management*, 44 (4), 1453-78.

Barnes, Nick (2010), "Publish Your Computer Code: It Is Good Enough," *Nature News*, 467 (7317), 753-53.

Becker, Thomas E., Guclu Atinc, James A. Breaugh, Kevin D. Carlson, Jeffrey R. Edwards, and Paul E. Spector (2016), "Statistical Control in Correlational Studies: 10 Essential Recommendations for Organizational Researchers," Journal of Organizational Behavior, 37 (2), 157-67. * guidelines for using control variables

Becker, Thomas E., Melissa M. Robertson, and Robert J. Vandenberg (2019), "Nonlinear Transformations in Organizational Research: Possible Problems and Potential Solutions," Organizational Research Methods, 22(4), 831–66.. * guidelines for variable transformations

Bellezza, S., & Berger, J. (2020). Trickle-Round Signals: When Low Status Is Mixed with High. *Journal of Consumer Research*, 47(1), 100=27.

References

* of special interest
** of outstanding interest

Bellezza, Silvia, Neeru Paharia, and Anat Keinan (2017), "Conspicuous Consumption of Time: When Busyness and Lack of Leisure Time Become a Status Symbol," *Journal of Consumer Research*, 44 (1), 118-38.

Bernerth, Jeremy B. and Herman Aguinis (2016), "A Critical Review and Best-Practice Recommendations for Control Variable Usage," *Personnel Psychology*, 69 (1), 229-83. * guidelines for using control variables

Boegershausen, Johannes; Borah, Abhishek; Stephen, Andrew T. (2020), "Fields of Gold: Web Scraping for Consumer Research", Marketing Science Institute (MSI) Working Paper Series, MSI Report #20-143 * older more behavioral research focused version of our paper: available at https://tiny.cc/Boegershausen_et_al_MSI

Boegershausen, Johannes, Hannes Datta, Abhishek Borah, and Andrew T. Stephen (2022), "Fields of Gold: Scraping Web Data for Marketing Insights", Journal of Marketing, 86(5), 1-20. ** contains the full methodological framework for collecting web data at scale + tables etc.

Brady, William J., Julian A. Wills, Dominic Burkart, John T. Jost, and Jay J. Van Bavel (2019), "An Ideological Asymmetry in the Diffusion of Moralized Content on Social Media among Political Leaders," *Journal of Experimental Psychology: General*, 148(10), 1802–1813.

Bright, Jonathan (2017), "Big Social Science: Doing Big Data in the Social Sciences," in *The Sage Handbook of Online Research Methods*, ed. Nigel G. Fielding, Raymond M. Lee and Grant Blank, London, UK: Sage, 125-39.

Chen, Eric Evan and Sean P. Wojcik (2016), "A Practical Guide to Big Data Research in Psychology," *Psychological Methods*, 21 (4), 458-74.

References

* of special interest
** of outstanding interest

Chen, Zoey and Jonah Berger (2013), "When, Why, and How Controversy Causes Conversation," Journal of Consumer Research, 40 (3), 580-93.

Chen, Zoey and Nicholas H. Lurie (2013), "Temporal Contiguity and Negativity Bias in the Impact of Online Word of Mouth," Journal of Marketing Research, 50 (4), 463-76.

Chen, Zoey (2017), "Social Acceptance and Word of Mouth: How the Motive to Belong Leads to Divergent Wom with Strangers and Friends," *Journal of Consumer Research*, 44 (3), 613-32.

Datta, Hannes, George Knox, and Bart J. Bronnenberg (2018), "Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery," *Marketing Science*, 37 (1), 5-21.

de Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), 817-33.

Doré, Bruce, Leonard Ort, Ofir Braverman, and Kevin N. Ochsner (2015), "Sadness Shifts to Anxiety over Time and Distance from the National Tragedy in Newtown, Connecticut," *Psychological Science*, 26 (4), 363-73.

Epskamp, Sacha (2019), "Reproducibility and Replicability in a Fast-Paced Methodological World," *Advances in Methods and Practices in Psychological Science*, 2 (2), 145-55.

Griffin, John M., Samuel Kruger, and Gonzalo Maturana (2019), "Personal Infidelity and Professional Conduct in 4 Settings," *Proceedings of the National Academy of Sciences*, 116 (33), 16268-73.

References

* of special interest
** of outstanding interest

Grijalva, Emily, Timothy D. Maynes, Katie L. Badura, and Steven W. Whiting (2020), "Examining the "I" in Team: A Longitudinal Investigation of the Influence of Team Narcissism Composition on Team Outcomes in the NBA," *Academy of Management Journal,* 63 (1), 7–33.

Henkel, Alexander P., Johannes Boegershausen, Joandrea Hoegg, Karl Aquino, and Jos Lemmink (2018), "Discounting Humanity: When Consumers Are Price Conscious Employees Appear Less Human," *Journal of Consumer Psychology*, 28 (2), 272-92.

Huang, Ni, Gordon Burtch, Yili Hong, and Evan Polman (2016), "Effects of Multiple Psychological Distances on Construal and Consumer Evaluation: A Field Study of Online Reviews," *Journal of Consumer Psychology*, 26 (4), 474-82.

Inman, J. Jeffrey, Margaret C. Campbell, Amna Kirmani, and Linda L. Price (2018), "Our Vision for the Journal of Consumer Research: It's All About the Consumer," *Journal of Consumer Research*, 44 (5), 955-59.

Kupor, Daniella and Zakary Tormala (2018), "When Moderation Fosters Persuasion: The Persuasive Power of Deviatory Reviews," *Journal of Consumer Research*, 45 (3), 490-510.

Landers, Richard N., Robert C. Brusso, Katelyn J. Cavanaugh, and Andrew B. Collmus (2016), "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research," *Psychological Methods*, 21 (4), 475-92. *** guidelines for formulating data source theories*

References

* of special interest
** of outstanding interest

LeBel, Etienne P., Randy J. McCarthy, Brian D. Earp, Malte Elson, and Wolf Vanpaemel (2018), "A Unified Framework to Quantify the Credibility of Scientific Findings," *Advances in Methods and Practices in Psychological Science*, 1 (3), 389-402. *** Outline a workflow for producing credible scientific findings.*

Luca, Michael and Georgios Zervas (2016), "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62 (12), 3412-27.

Matz, Sandra C. and Oded Netzer (2017), "Using Big Data as a Window into Consumers' Psychology," *Current Opinion in Behavioral Sciences*, 18 (1), 7-12.

McGraw, A. Peter, Caleb Warren, and Christina Kan (2015), "Humorous Complaining," *Journal of Consumer Research*, 41 (5), 1153-71.

Melumad, Shiri, J. Jeffrey Inman, and Michel Tuan Pham (2019), "Selectively Emotional: How Smartphone Use Changes User-Generated Content," *Journal of Marketing Research*, forthcoming.

Mitchell, Ryan (2015), *Web Scraping with Python: Collecting Data from the Modern Web*, Sebastopol, CA: O'Reilly Media.

Mooijman, Marlon, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani (2018), "Moralization in Social Networks and the Emergence of Violence During Protests," *Nature Human Behaviour*, 2 (6), 389-96.

Moore, Sarah G. (2012), "Some Things Are Better Left Unsaid: How Word of Mouth Influences the Storyteller," *Journal of Consumer Research*, 38 (6), 1140-54.

References

* of special interest
** of outstanding interest

Mortensen, Chad R. and Robert B. Cialdini (2010), "Full-Cycle Social Psychology for Theory and Application," *Social and Personality Psychology Compass*, 4 (1), 53-63.

Murphy, Sean C. (2017), "A Hands-on Guide to Conducting Psychological Research on Twitter," *Social Psychological and Personality Science*, 8 (4), 396-412.

Ordenes, Francisco Villarroel, Dhruv Grewal, Stephan Ludwig, Ko De Ruyter, Dominik Mahr, and Martin Wetzels (2019), "Cutting through Content Clutter: How Speech and Image Acts Drive Consumer Sharing of Social Media Brand Messages," *Journal of Consumer Research*, 45 (5), 988-1012.

Ordenes, Francisco Villarroel, Stephan Ludwig, Ko de Ruyter, Dhruv Grewal, and Martin Wetzels (2017), "Unveiling What Is Written in the Stars: Analyzing Explicit, Implicit, and Discourse Patterns of Sentiment in Social Media," *Journal of Consumer Research*, 43 (6), 875-94.

Paharia, Neeru, Jill Avery, and Anat Keinan (2014), "Positioning Brands against Large Competitors to Increase Sales," *Journal of Marketing Research*, 51 (6), 647-56.

Pury, Cynthia L. S. (2011), "Automation Can Lead to Confounds in Text Analysis: Back, Küfner, and Egloff (2010) and the Not-So-Angry Americans," *Psychological Science*, 22 (6), 835-36.

Reis, Harry T. (2012), "Why Researchers Should Think "Real-World": A Conceptual Rationale," in *Handbook of Research Methods for Studying Daily Life.*, ed. Matthias R. Mehl and Tamlin S. Conner, New York, NY, US: The Guilford Press, 3-21.

References

* of special interest
** of outstanding interest

Simmons, Joseph P., Leif D. Nelson, Jeff Galak, and Shane Frederick (2011), "Intuitive Biases in Choice Versus Estimation: Implications for the Wisdom of Crowds," *Journal of Consumer Research*, 38 (1), 1-15.

Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay (2017), "Constraints on Generality (Cog): A Proposed Addition to All Empirical Papers," *Perspectives on Psychological Science*, 1745691617708630.

Smith, Rosanna K., George E. Newman, and Ravi Dhar (2016), "Closer to the Creator: Temporal Contagion Explains the Preference for Earlier Serial Numbers," *Journal of Consumer Research*, 42 (5), 653-68.
** an example for a very cogent sample justification*

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel (2016), "Increasing Transparency through a Multiverse Analysis," *Perspectives on Psychological Science*, 11 (5), 702-12. * presents Multiverse Analysis as a tool to show the analytical robustness of a finding

Umashankar, Nita, Morgan K. Ward, and Darren W. Dahl (2017), "The Benefit of Becoming Friends: Complaining after Service Failures Leads Customers with Strong Ties to Increase Loyalty," *Journal of Marketing*, 81 (6), 79-98.

Van Ittersum, Koert and Brian Wansink (2012), "Plate Size and Color Suggestibility: The Delboeuf Illusion's Bias on Serving and Eating Behavior," *Journal of Consumer Research*, 39 (2), 215-28.

References

* of special interest
** of outstanding interest

Van Laer, Tom, Jennifer Edson Escalas, Stephan Ludwig, and Ellis A. Van Den Hende (2019), "What Happens in Vegas Stays on Tripadvisor? A Theory and Technique to Understand Narrativity in Consumer Reviews," *Journal of Consumer Research*, 46 (2), 267-85. * an example for clearly justifying the inclusion and operationalization of control variables

Salge, Carolina Alves De Lima and Elena Karahanna (2018), "Protesting Corruption on Twitter: Is It a Bot or Is It a Person?," *Academy of Management Discoveries*, 4 (1), 32-49.

Simchi-Levi, David (2019), "From the Editor," *Management Science*, 65 (2), v–vi.

Wenzel, Ramon and Niels Van Quaquebeke (2018), "The Double-Edged Sword of Big Data in Organizational and Management Research: A Review of Opportunities and Risks," *Organizational Research Methods*, 21 (3), 548-91.

Wang, Ze, Huifang Mao, Yexin Jessica Li, and Fan Liu (2017), "Smile Big or Not? Effects of Smile Intensity on Perceptions of Warmth and Competence," *Journal of Consumer Research*, 43 (5), 787-805.

Wickham, Hadley (2019), "Rvest: Easily Harvest (Scrape) Web Pages," *The R Foundation: Vienna, Austria*. * documentation of the R package for parsing HTML and XML data

Xu, Heng, Nan Zhang, and Le Zhou (2020), " Validity Concerns in Research Using Organic Data," *Journal of Management*, 46(7), 1257–74. ** an excellent overview of some validity issues in organically created data

Yin, Dezhi, Samuel D. Bond, and Han Zhang (2017), "Keep Your Cool or Let It Out: Nonlinear Effects of Expressed Arousal on Perceptions of Consumer Reviews," *Journal of Marketing Research*, 54 (3), 447-63.