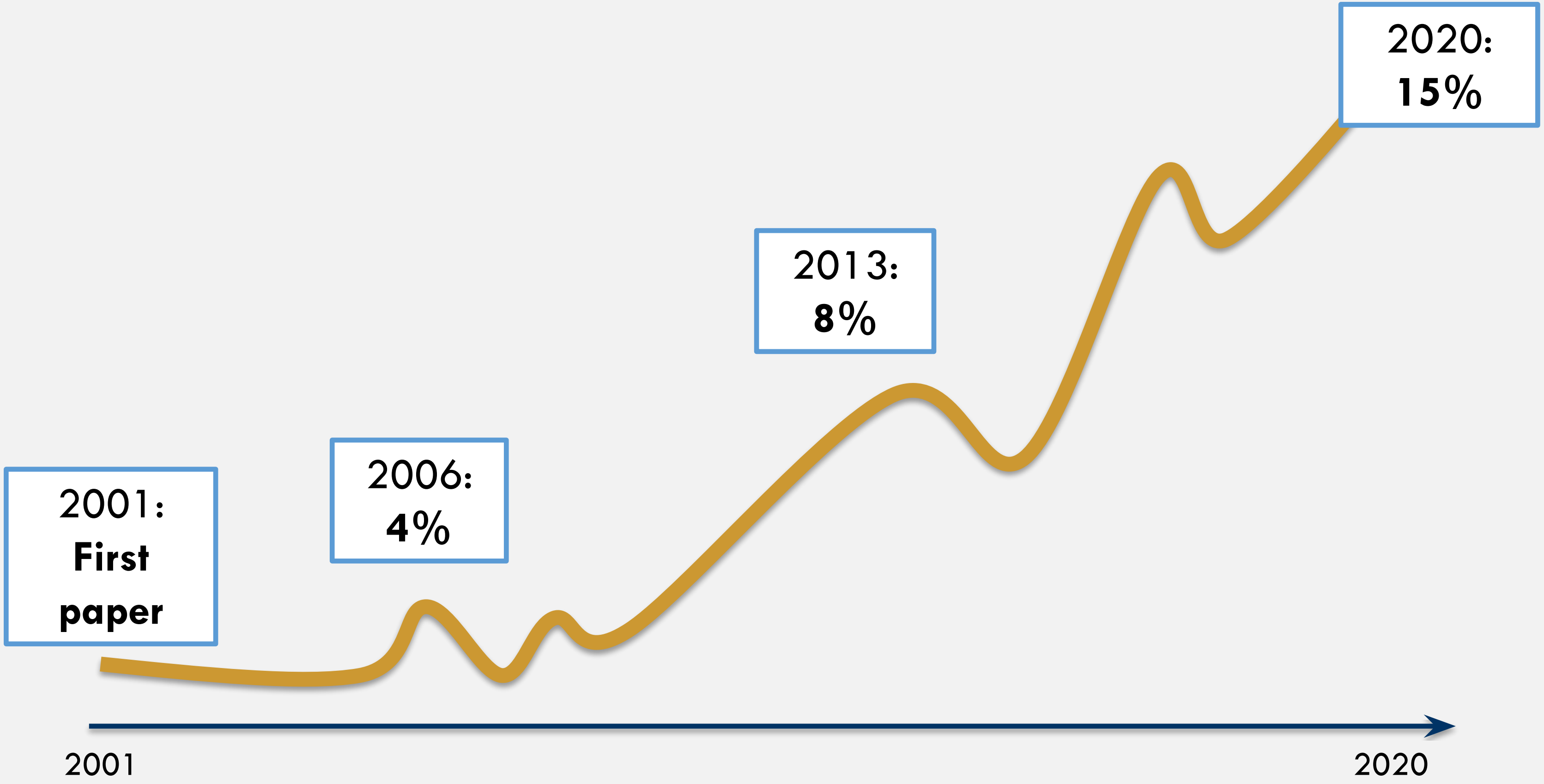
Extracting high-quality web data for academic use

Challenges and opportunities

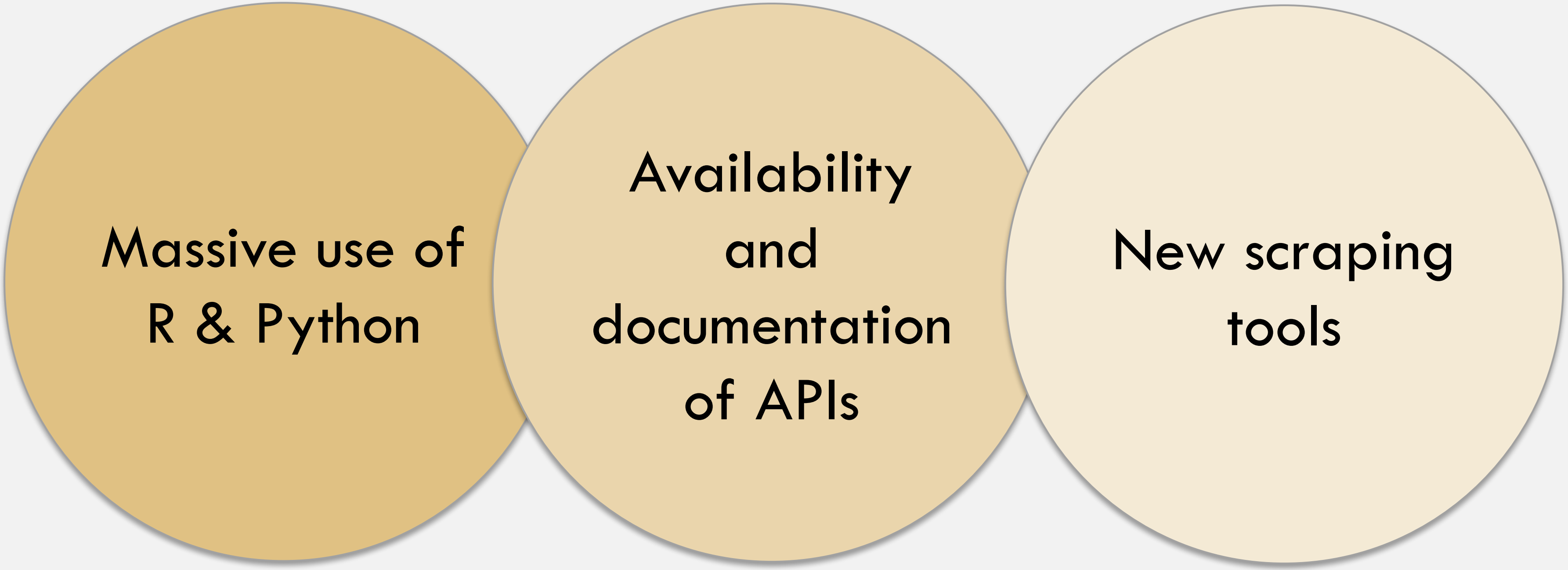
dr. Hannes Datta
@hannesdatta
hannesdatta.com

Tilburg University

► Increased use of web data in marketing research



► Thriving data collection scene among academics



Massive use of R & Python

Availability and documentation of APIs

New scraping tools

Agenda

1. Why do academics collect web data?

2. Facing key challenges

3. Food for thought

DISCLAIMER

- Focus on marketing research
- Small-scale web data projects
- Coding skills among researchers

WHY DO ACADEMICS COLLECT WEB DATA?

► 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

► How scholars seek to create new knowledge

Pathway ①

Studying new phenomena

airbnb Spotify®

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

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

► How scholars seek to create new knowledge

Pathway ①

Studying new phenomena

airbnb Spotify®

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

Pathway ②

Boosting ecological value

Google Trends

Amazon Best Sellers
Our most popular products based on sales. Updated hourly.

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

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

How scholars seek to create new knowledge

Pathway ①

Studying new phenomena

airbnb Spotify

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

Pathway ②

Boosting ecological value

Google Trends

Amazon Best Sellers
Our most popular products based on sales. Updated hourly.

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)

How scholars seek to create new knowledge

Pathway ①

Studying new phenomena

airbnb Spotify

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

Pathway ②

Boosting ecological value

Google Trends Amazon Best Sellers

Our most popular products based on sales. Updated hourly.

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

WEATHER UNDERGROUND

Holiday API™

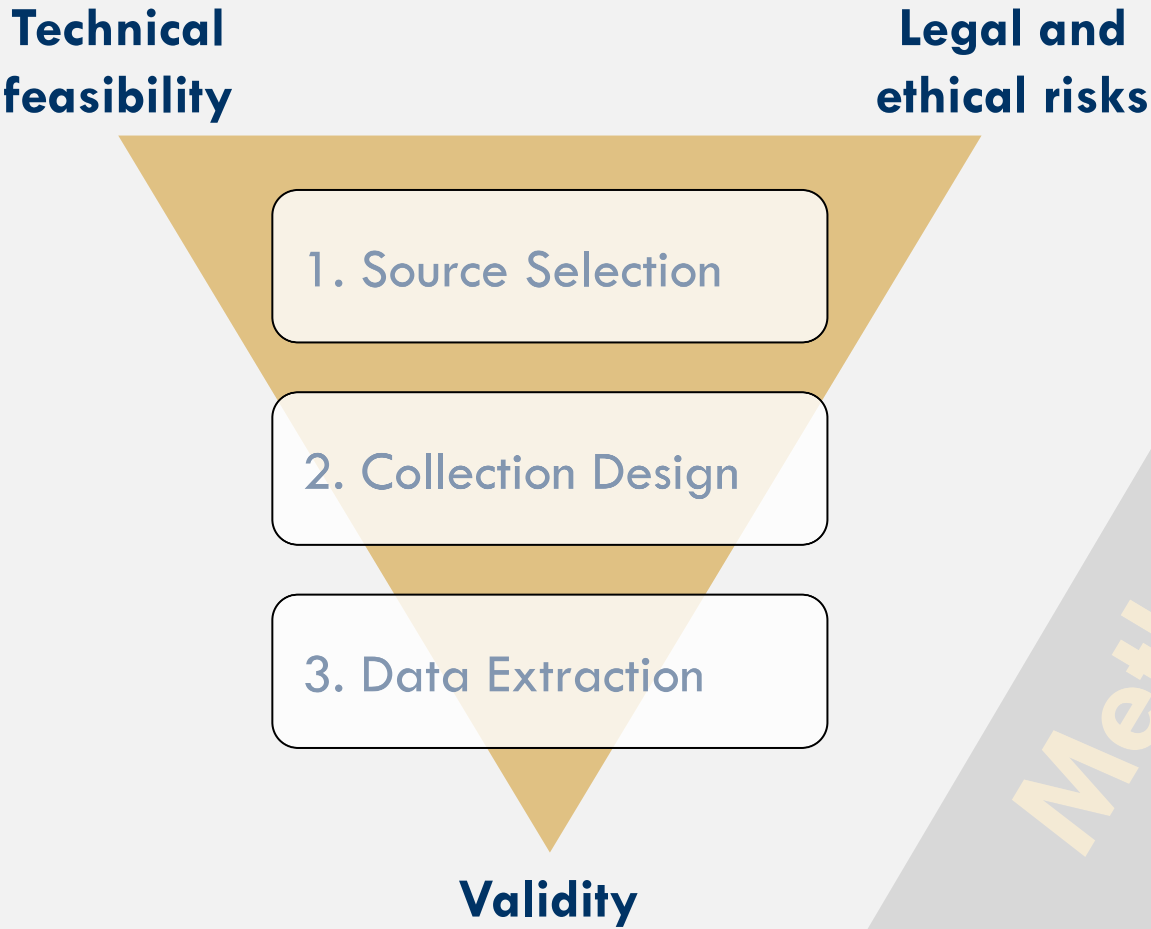
e.g., Li et al. (2017); Datta et al. (2022)

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

Legal, technical and validity challenges of web data

FACING KEY CHALLENGES

Methodological framework



Technical feasibility

Legal and ethical risks

1. Source Selection

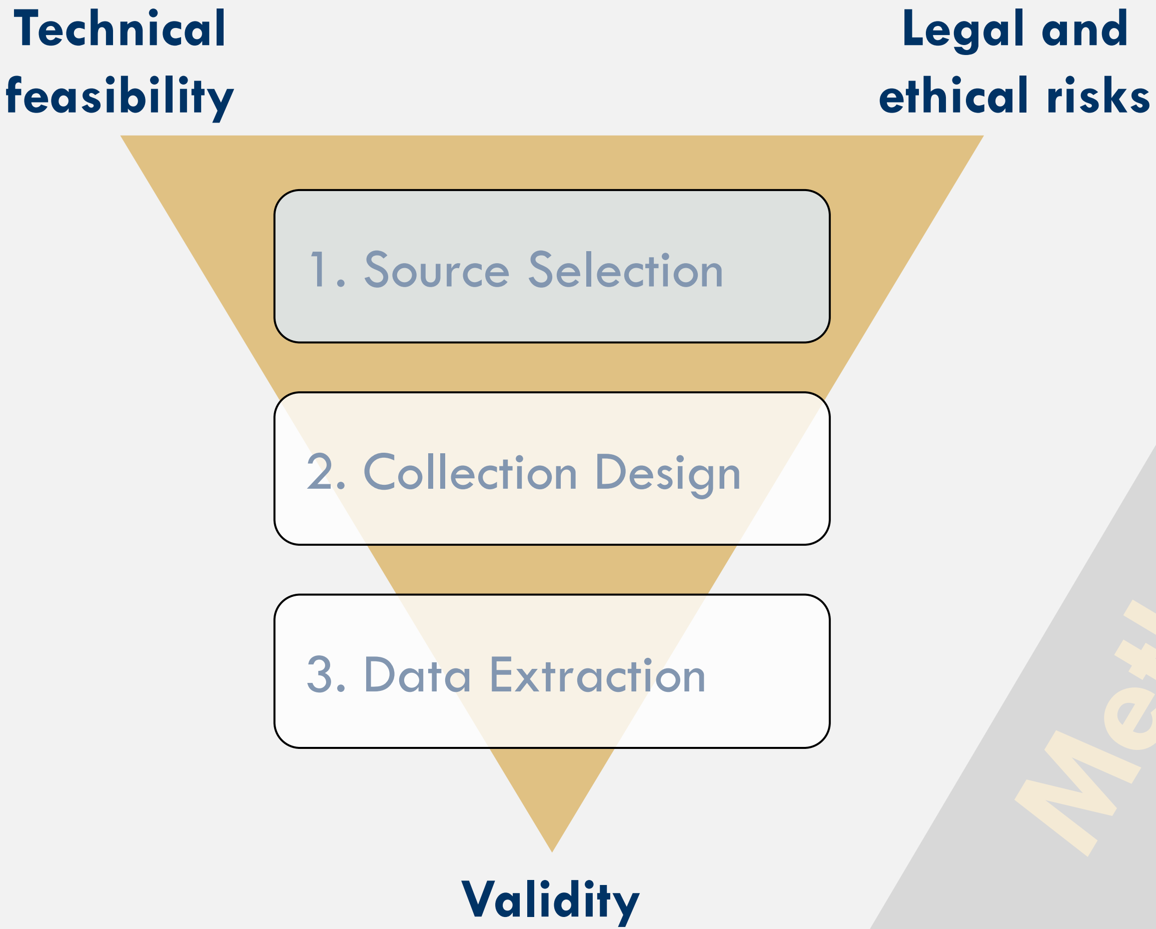
2. Collection Design

3. Data Extraction

Validity

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

Methodological framework



Technical feasibility

Legal and ethical risks

1. Source Selection

2. Collection Design

3. Data Extraction

Validity

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

► Discover universe of potential sources

- Near-to infinite number of potential sources, without traditional gatekeepers

- High concentration in platform use across studies
  - 12% Amazon.com
  - 10% Twitter
  - 8% IMDB

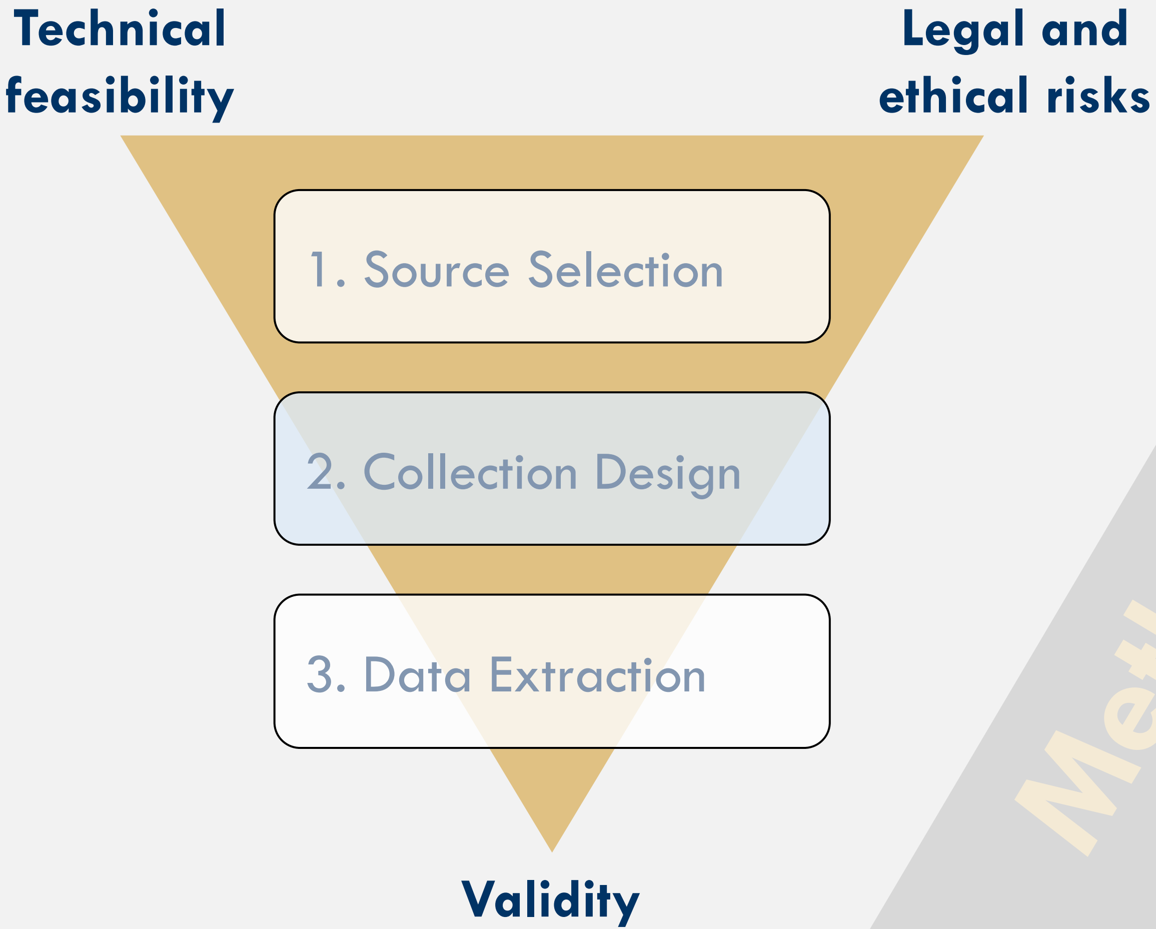
- Risk of defaulting
  - Using familiar platforms limits knowledge discovery
  - Using web scraping (vs. APIs) may affect data quality

Understanding a website's context

- Validity challenges
  - Did the data-generating process change?
  - Algorithms present or updated?

- Possible solutions
  - Screen blogs, press releases, a software's changelogs
  - Use archive.org
  - Visit site at different devices/times
  - Inspect source code

Methodological framework



Technical feasibility

Legal and ethical risks

1. Source Selection

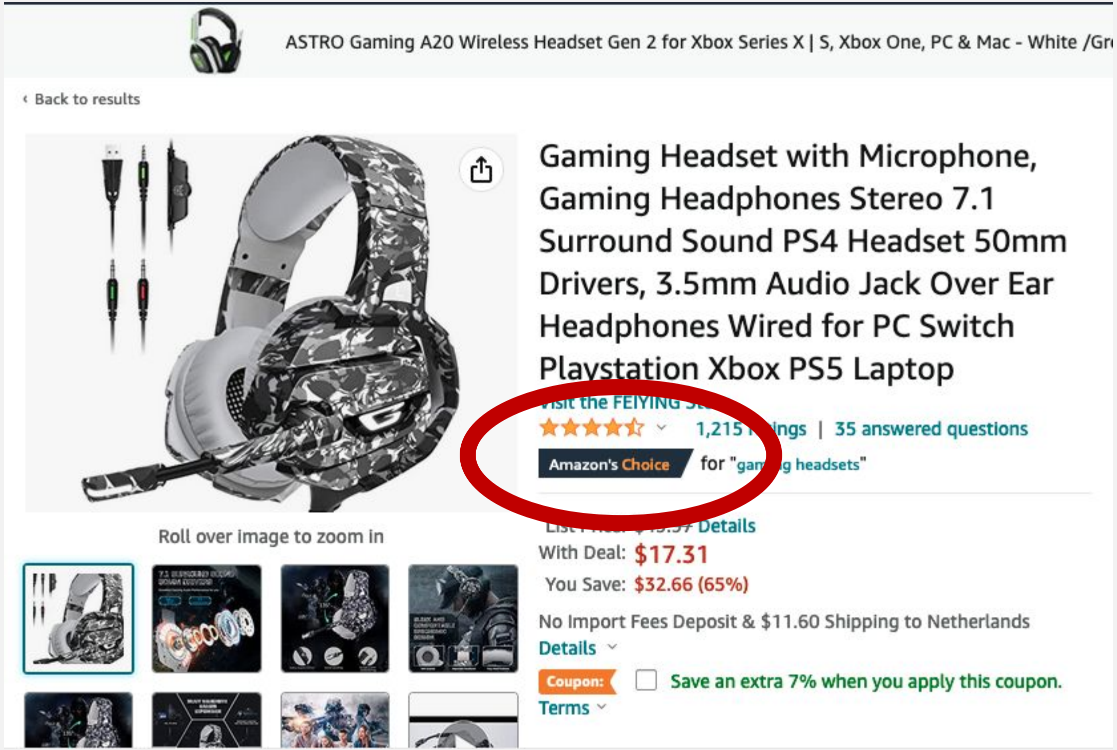
2. Collection Design

3. Data Extraction

Validity

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

Which information to extract?



ASTRO Gaming A20 Wireless Headset Gen 2 for Xbox Series X | S, Xbox One, PC & Mac - White /Gr

‹ Back to results

Roll over image to zoom in

Gaming Headset with Microphone, Gaming Headphones Stereo 7.1 Surround Sound PS4 Headset 50mm Drivers, 3.5mm Audio Jack Over Ear Headphones Wired for PC Switch Playstation Xbox PS5 Laptop

Visit the FEIYING St

★★★★☆ ⌄ 1,215 ratings | 35 answered questions

Amazon's Choice for "gaming headsets"

List Price: $49.97 Details

With Deal: $17.31

You Save: $32.66 (65%)

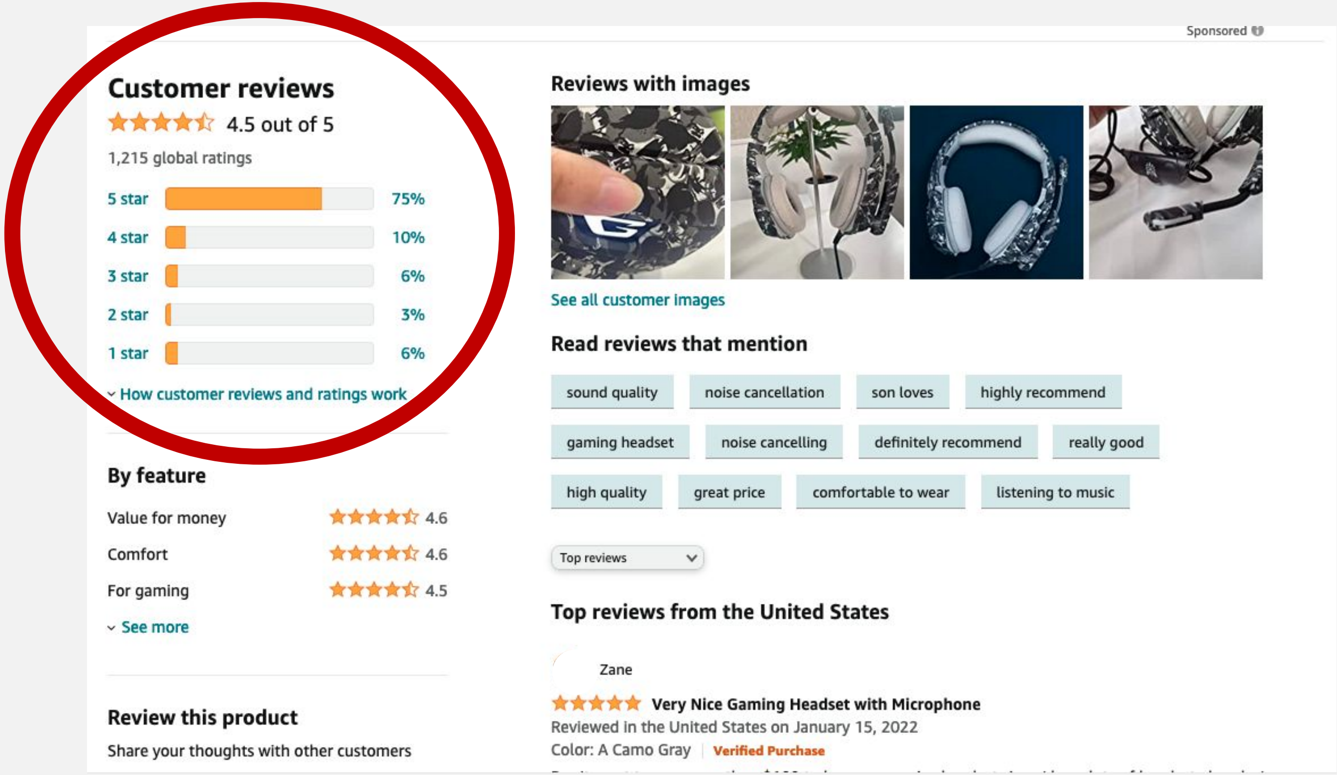
No Import Fees Deposit & $11.60 Shipping to Netherlands

Details ⌄

Coupon: □ Save an extra 7% when you apply this coupon.

Terms ⌄

Which information to extract?



Sponsored

Customer reviews

★★★★☆ 4.5 out of 5

1,215 global ratings

5 star	75%
4 star	10%
3 star	6%
2 star	3%
1 star	6%

How customer reviews and ratings work

By feature

Value for money	★★★★☆ 4.6
Comfort	★★★★☆ 4.6
For gaming	★★★★☆ 4.5

See more

Review this product

Share your thoughts with other customers

Reviews with images

See all customer images

Read reviews that mention

sound quality | noise cancellation | son loves | highly recommend | gaming headset | noise cancelling | definitely recommend | really good | high quality | great price | comfortable to wear | listening to music

Top reviews

Top reviews from the United States

Zane

★★★★★ **Very Nice Gaming Headset with Microphone**

Reviewed in the United States on January 15, 2022

Color: A Camo Gray | **Verified Purchase**

Challenges in information extraction

Validity

Is information subject to algorithmic biases or missing data?

Are there significant changes to the data-generating process?

Is meta data required to make sense of variables?

Legality & ethics

Publicly accessible vs. login? Consent to ToS? Implicit or explicit?

Feasibility to obtain permission?

Personal or sensitive information?

Sufficient scientific justification?

Technicalities

Limits to iterating through pages?

All information extractable?

How to sample?

- Sampling frames (might) create different datasets or even induce systematic biases
  - Sampling from internal pages (e.g., bestseller, category, search page)
  - Sampling from externally available lists
  - Inability to capture population

- Which sample size is *technically* feasible?

► At what frequency to extract data?

- Gains from capturing information more than once
   build longitudinal data set, capture “fake” reviews

- Balance sample size and extraction frequency
   power to identify effects

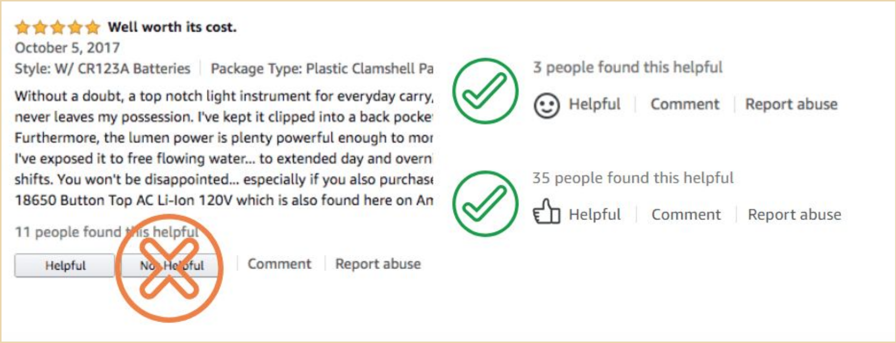
- Validation of “data” assumptions absolutely required
  - Configuration (e.g., “data is historically available”)
  - Data-generating process (e.g., “website hasn’t changed”)
  - Recency (e.g., data is up-to-date)

► How to process data during the extraction?

Most researchers process data "on-the-fly"

→ Mitigate threats of validity by **keeping raw data** whenever possible (but, legally possible?)

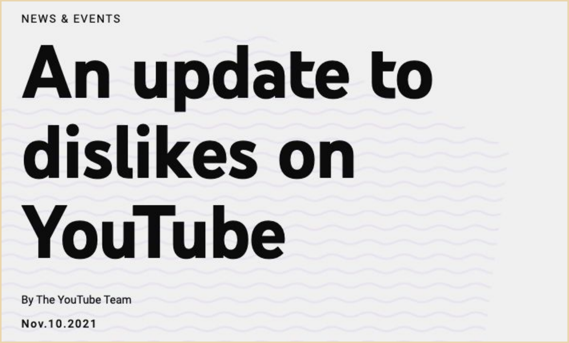
Opportunity that researchers like: "stumbling" into natural experiments



★★★★★ **Well worth its cost.**
October 5, 2017
Style: W/ CR123A Batteries | Package Type: Plastic Clamshell Pa
Without a doubt, a top notch light instrument for everyday carry,
never leaves my possession. I've kept it clipped into a back pocke
Furthermore, the lumen power is plenty powerful enough to mor
I've exposed it to free flowing water... to extended day and overni
shifts. You won't be disappointed... especially if you also purchase
18650 Button Top AC Li-Ion 120V which is also found here on Am
11 people found this helpful
Helpful | Not Helpful | Comment | Report abuse

3 people found this helpful
Helpful | Comment | Report abuse

35 people found this helpful
Helpful | Comment | Report abuse

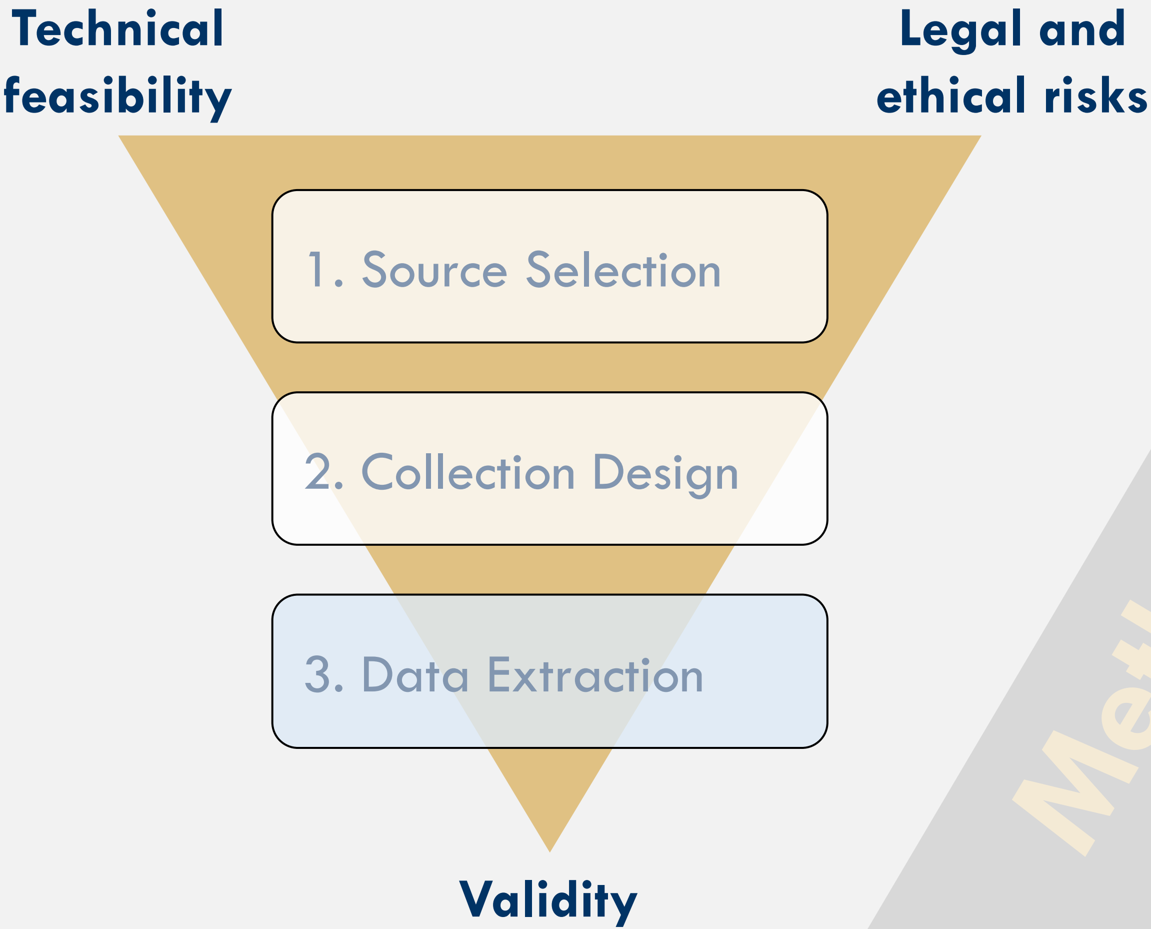


NEWS & EVENTS

An update to dislikes on YouTube

By The YouTube Team
Nov.10.2021

Methodological framework



Technical feasibility

Legal and ethical risks

1. Source Selection

2. Collection Design

3. Data Extraction

Validity

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

Data extraction

- How to **improve** the performance of the data extraction?
  - Code often runs in Jupyter Notebooks; schedulers may be poorly defined
  - Researchers work in small teams, difficult to scale up!

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

- How to **document** the data **during** and **after** the extraction?
  - Reproducibility of research is increasingly important
  - Document how data was generated and why specific design choices were made

Providing scraping solutions for the academic community

FOOD FOR THOUGHT

► Facilitate source selection

- Directory of web data sources + code snippets
  - Create buzz about ‘new’ web sources
  - Build researcher-focused API directories (e.g., for improving measurement)
- Provide legal compliance tools
  - Automatic checks on robots.txt, terms of use
  - Flag questionable sites, offer alternatives
- Offer API training tools
  - Toy-box API for students, like books.toscrape.com
  - Learn different ways to authenticate
- Contribute web-scraped data sets to the community
  - E.g., Kaggle.com (discoverability + best practices)

► Assist researchers collect valid data *by design*

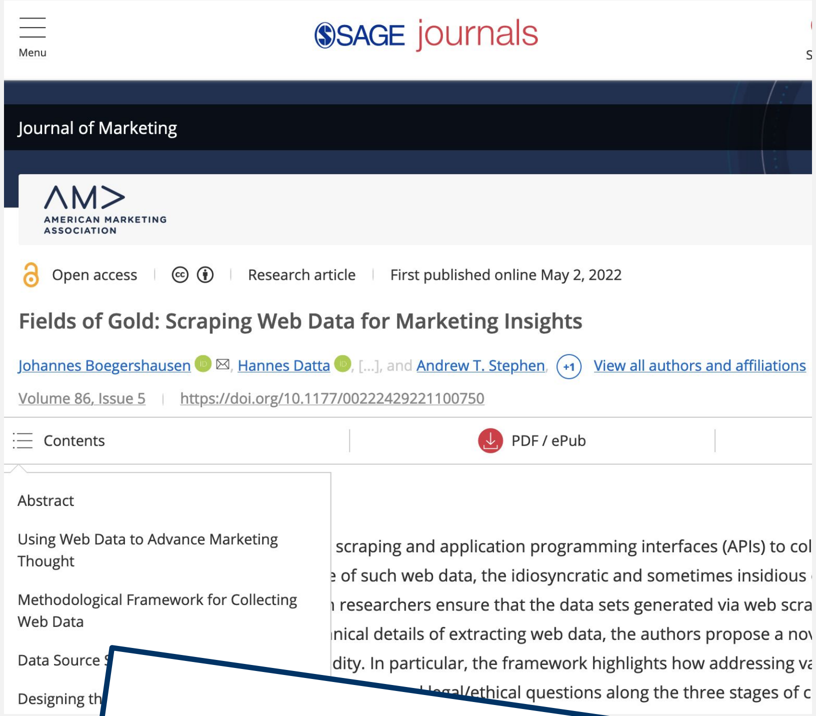
- Support decision making
  - e.g., site may have changed – consider collecting longitudinal data
- Make collections more robust
  - anonymization and pseudonomization
  - allow retrieving copies of historical versions of the site
- Support documenting the data collection
  - screenshots of websites while scraping
  - log book of important events

► Facilitate *scaling up*

- Build technical case studies for researchers
  - sponsor research infrastructure
  - consider offering developer support
  - clearly link to academic papers
- Contribute to legal debate
  - collect best practices
  - build network for legal advice
  - focus on several geographic markets

▶ Conclusion

- Web (data) is here to stay (and grow)
- Four pathways of knowledge creation fuel entire research programs
- Direct influence on data quality through source selection, design, and extraction
- Let's embrace new opportunities



SAGE journals

Journal of Marketing

AMA AMERICAN MARKETING ASSOCIATION

Open access | Research article | First published online May 2, 2022

Fields of Gold: Scraping Web Data for Marketing Insights

Johannes Boegershausen, Hannes Datta, [...], and Andrew T. Stephen, +1 View all authors and affiliations

Volume 86, Issue 5 | https://doi.org/10.1177/00222429221100750

Abstract

Using Web Data to Advance Marketing Thought

Methodological Framework for Collecting Web Data

Data Source S

Designing th

scraping and application programming interfaces (APIs) to col
e of such web data, the idiosyncratic and sometimes insidious
n researchers ensure that the data sets generated via web scra
nical details of extracting web data, the authors propose a nov
dity. In particular, the framework highlights how addressing va
legal/ethical questions along the three stages of c

open access

https://tiu.nu/scraping

► Thank you!

dr. Hannes Datta

Tilburg University

@hannesdatta
hannesdatta.com

Read our paper at
https://tiu.nu/scraping