Scraping and Making Sense of Web and Field Data for Consumer Research

Johannes Boegershausen & Aurélie Lemmens

EACR 2023, Amsterdam
7 July 2023

Supported by

Disclaimer: websites commonly used in consumer research articles explicitly prohibit scraping in their Terms of Service (ToS). We do not condone violating these ToS but will selectively use these websites as illustrative examples given their prevalence in published consumer research.

Collection of consequential variables

“A well-designed field study demonstrates generalizability of the lab-based studies, increasing external validity by showing that the focal effects persist in the noisy environment of the real world.” *Inman et al. (2018, p. 957)*

- **Part 1:**
  maximizing the potential & validity of web data collected at scale

- **Part 2:**
  leveraging field experiment data with causal machine learning

Introductory disclaimer

- By any means, we are really not the first scholars to gather web data via scraping, APIs, etc.,
  - but we have used this in our work + reviewed such research (extensively)
  - we have published a methodological paper about collecting web data at scale (Boegershausen, Borah, Datta, and Stephen 2022)
- There is no boilerplate template for gathering web data for consumer research.
- Scraping & APIs are useful for all types of consumer research, given my own expertise & time constraints, I will focus mostly on behavioral research.
- When you feel that I am going too fast, please slow me down.
- This is **designed to be an interactive session**, so we might not get through all materials, but there are more resources available @ www.web-scraping.org

The Internet is ubiquitous

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

Number of active users in January 2023 (global)

Facebook	YouTube	Instagram	TikTok	Twitter
>2.9b	2.5b	2b	1b	330m

Source: We Are Social & Hootsuite (2019, 2021); Statista(2021)

Generation of massive digital traces

~ **265m** reviews

~ **1B** reviews & opinions

500m/day

590K projects

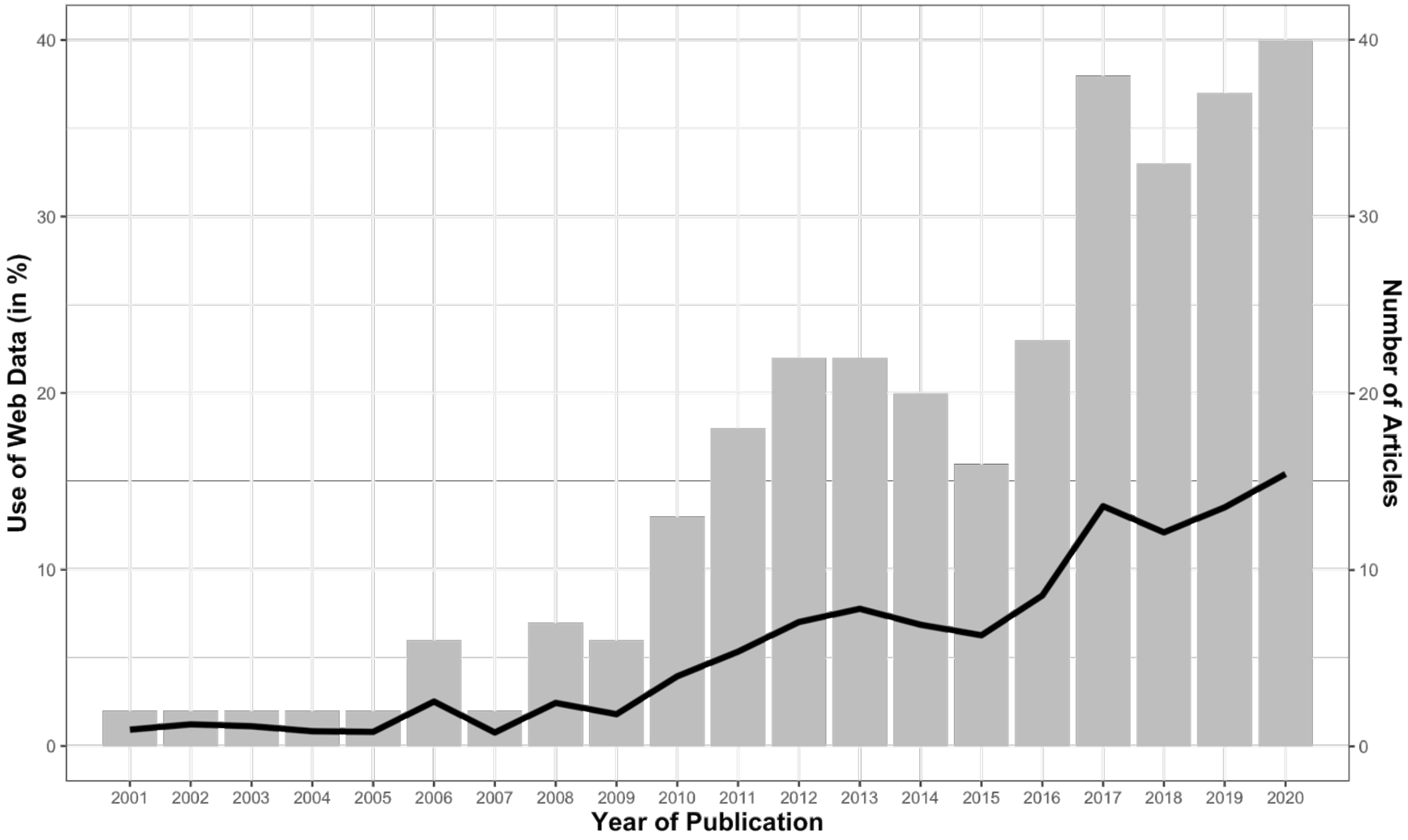
Stats last updated on 14 June 2023

Transforming digital traces into datasets

Transforming web data into datasets for academic research via **web scraping** typically involves "***writing an automated program that queries a web server, requests data […], and then parses that data to extract needed information***" (Mitchell 2015, p. viii).

or draw on ***Application Programming Interfaces (APIs)***

Increasing usage of web data in marketing research



+ 51 (2021)

+ 59 (2022)

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

Collection of consequential variables

“A well-designed field study demonstrates generalizability of the lab-based studies, increasing external validity by showing that the focal effects persist in the noisy environment of the real world. […], ***there are a variety of ways to collect consequential dependent variables from the “real world,” e.g., scraping and analyzing consumers’ social media posts or product ratings.“‘***

Inman et al. (2018, p. 957)

- Enormous volume of data capturing the **<u>actual behaviors</u>** of individuals and firms is readily available
- Scraping data can provide compelling answers to the question of “*assuming that this hypothesis is true, in what ways does it manifest in the world*” (Barnes et al. 2018, p. 1455).

What reviewers say *(2018 JCP)*

- **“It may be necessary to include a real field study to have a better package of studies and increase the contribution.”**

✓ *“the review team liked the two new studies as they grounded the effect nicely (study 1 based on web data)” [AE]*

✓ *“I liked the two new studies, especially study 1 [using web-scraped data]” [Reviewer A]*

✓ *“I like the new Study 1 a lot.” [Reviewer B]*

Highly versatile data collection technique

Pathway 1

Boosting ecological value

Google Trends

Amazon Best Sellers

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

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

Highly versatile data collection technique

Pathway ①

Boosting ecological value

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

Pathway ②

Studying new phenomena

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

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

Highly versatile data collection technique

Pathway 1

Boosting ecological value

Google Trends

Amazon Best Sellers

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

Pathway 2

Studying new phenomena

airbnb

Spotify

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

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

Boosting ecological value

Google Trends

Amazon Best Sellers

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

Pathway ②

Studying new phenomena

airbnb

Spotify

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

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)

Highly versatile data collection technique +++

"scout out" novel phenomena

streaming (Datta et al. 2018)
mobile devices (Melumad et al. 2019)

different levels of analysis + effects over time

brand public (Arvidsson & Caliandro 2016)
psychological distances (Huang et al. 2016)

explore geographic variation

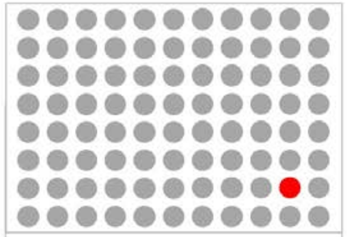
Sensitivity to prices and ratings across the globe (Kübler et al. 2018)

stimuli generation

provider profiles (Howe & Monin 2017)
brand logos (Luffarelli et al. 2019)

socially sensitive phenomena

controversy (Chen & Berger 2013)
violent protests (Mooijman et al. 2018)



rare events

Bright (2017)

hard-to-reach populations

political elites (Brady et al. 2019)
professional athletes (Grijalva et al. 2020)
early Spotify adopters (Datta et al. 2018)

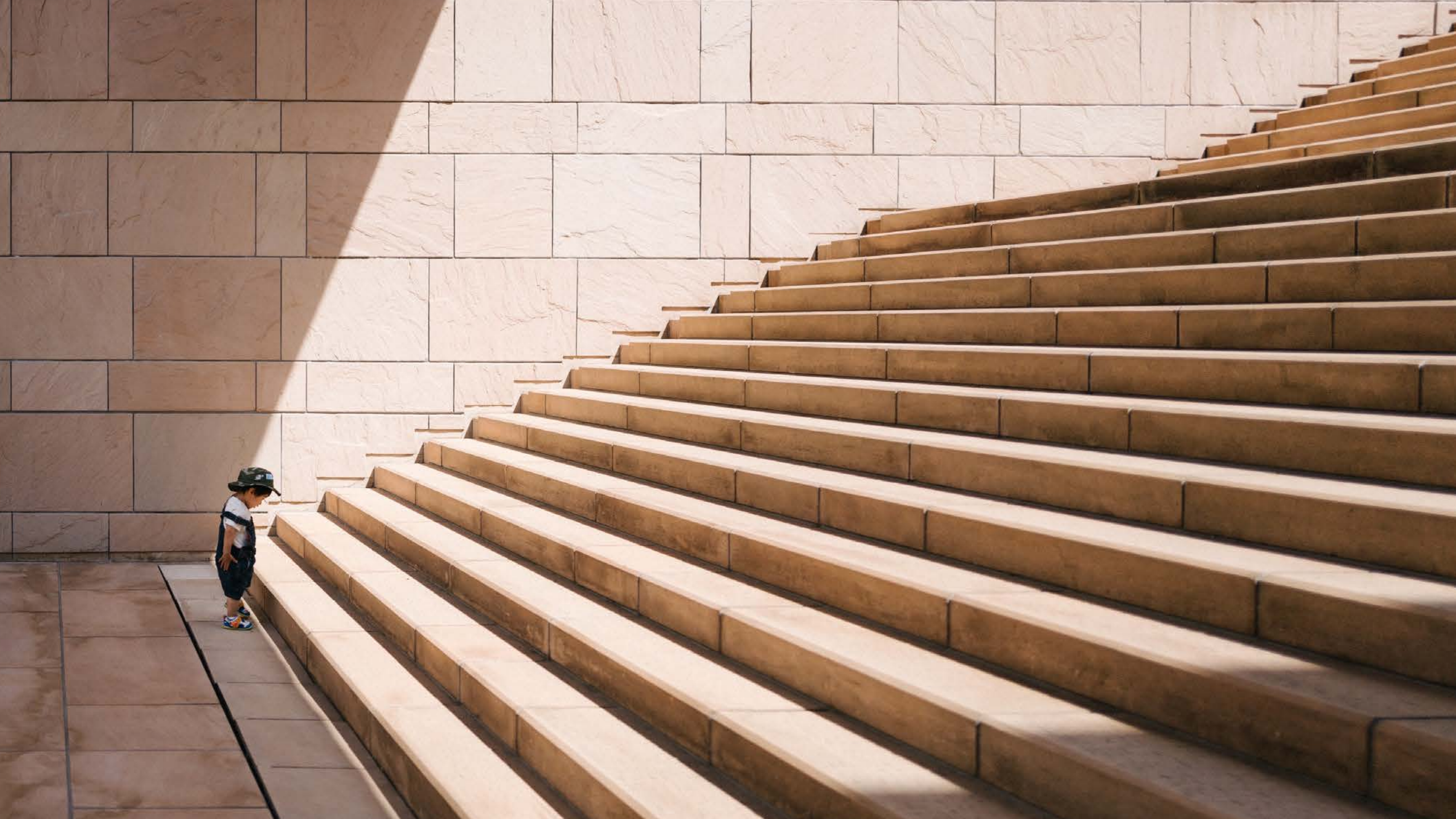
The Weather Company
An IBM Business

data enrichment

Govind et al. (2020)
Datta et al. (2022)

Increasing researchers’ efficiency

- **Rapid and cheap** generation of large, novel, and interesting datasets
- Ability to explore the generalizability of (important) effects established primarily within the confines of laboratory experiments
- Of heightened relevance for doctoral students, early career scholars, and researchers employed at institutions with less resources
  → *potential to level the playing field*



Accessibility

Awareness of different paths to harvesting web data
Understanding of the basic mechanics of web scraping

Lack of a structured approach

Credibility of web scraped based research
Standards for evaluating research using web scraping

Accessibility

Awareness of different paths to harvesting web data
Understanding of the basic mechanics of web scraping

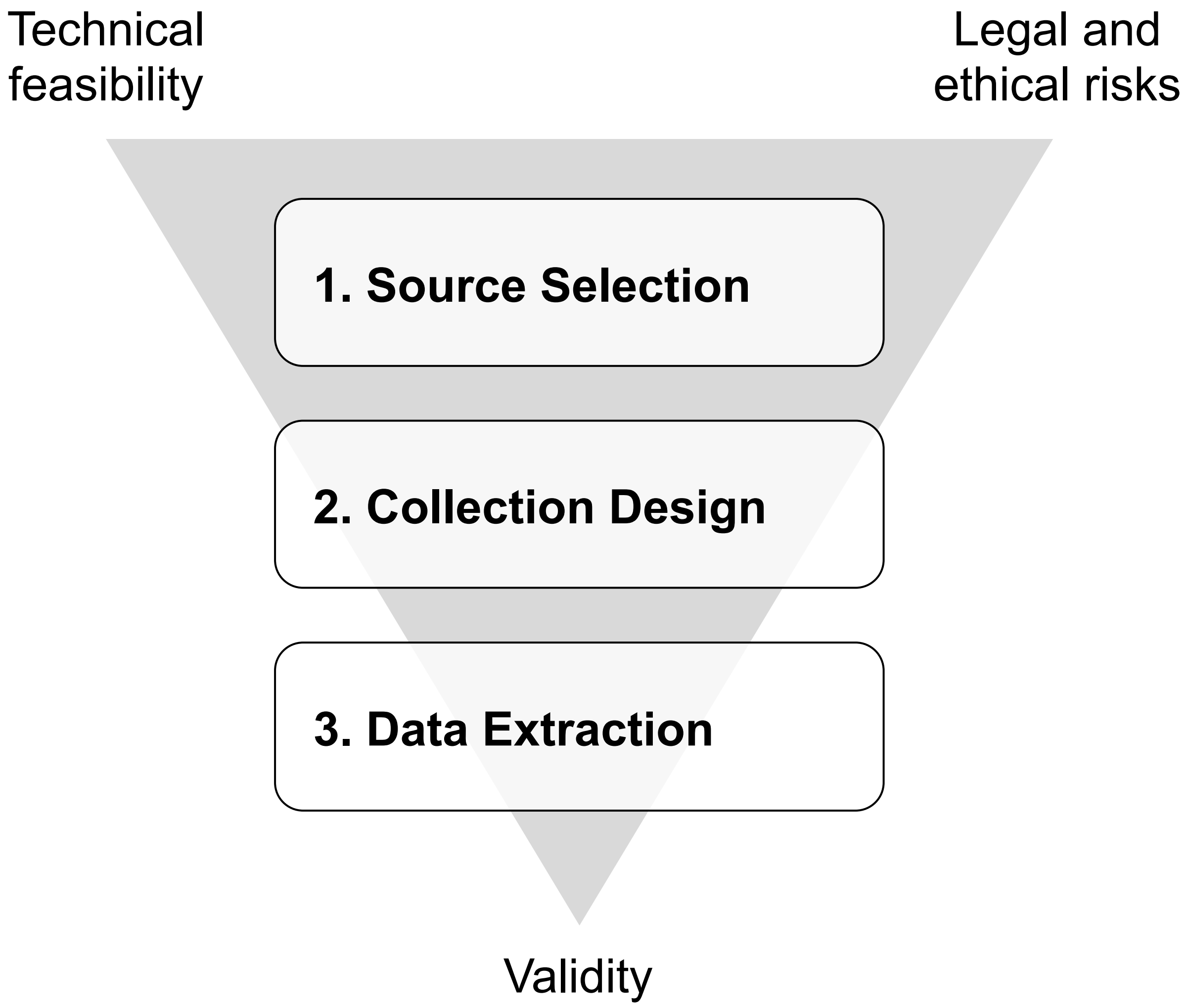
What’s in it for you

- Increased awareness of what scraped data is
  - 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
- Provide guidance on idiosyncratic challenges of web scraping
  - Single vs. multisource? Algorithmic biases?
  - Focus on validity (not technicalities!), legal concerns
  - Extraction frequency and sampling?
  - Keep raw HTML/JSON data?

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

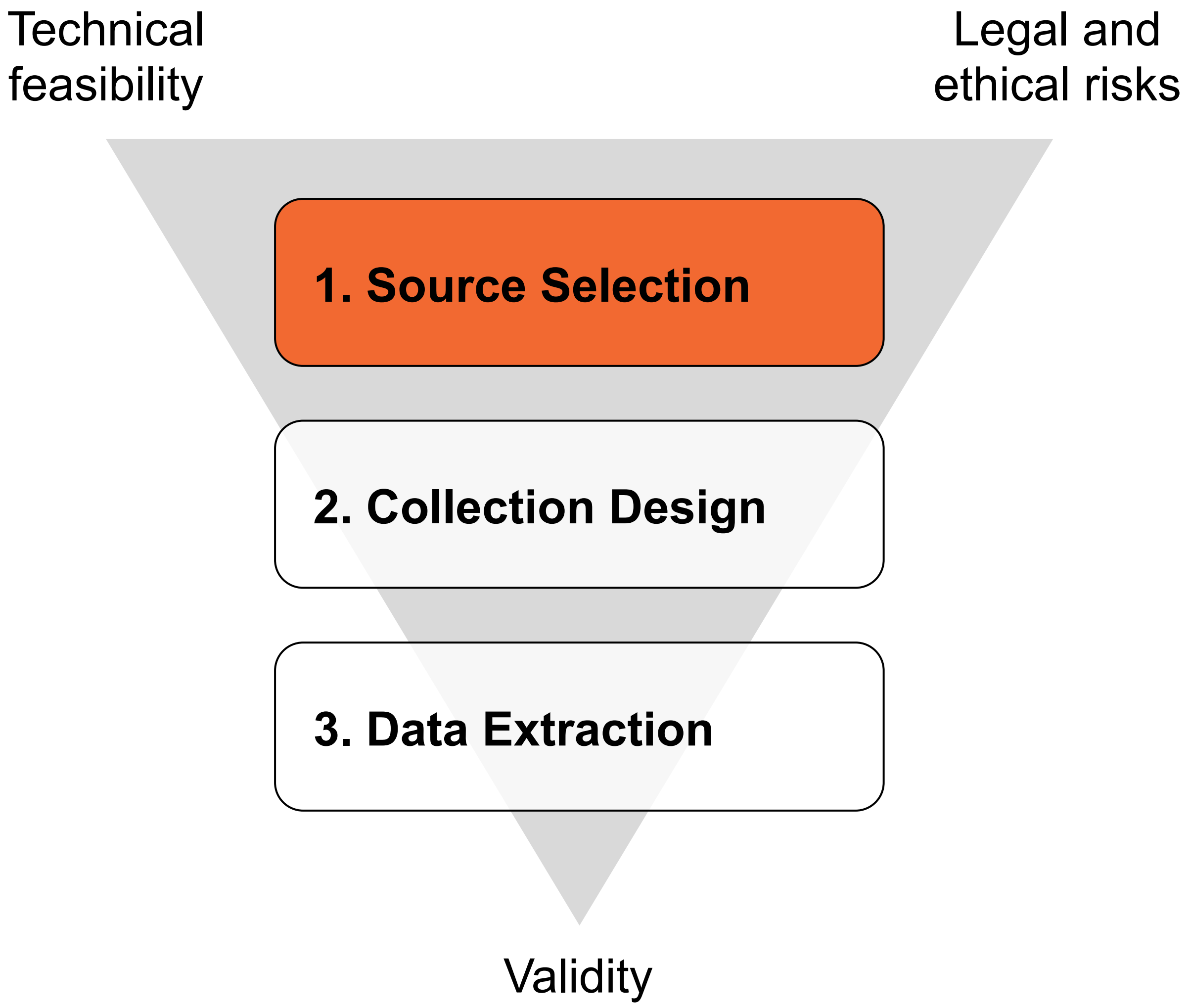
METHODOLOGICAL FRAMEWORK

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

Source selection: example

yelp

tripadvisor

amazon

SKYTRAX

healthgrades.
How America finds a doctor.

IMDb

Review us on... facebook

bol.com
de winkel van ons allemaal

wayfair.co.uk

Source selection: *justification strategies*

- Deciding which website(s) to sample is challenging, yet critical

► <u>Remedy:</u> Present a clear rationale to motivate the sampling choice; some useful approaches below:

– identify *idiosyncratic feature(s)* (e.g., Yelp funny votes; McGraw et al. 2015)

– particular *type* of webpage (e.g., discussion forum; Chen & Berger 2013)

– when agnostic about the source, sampling multiple websites can increase confidence about effect generalizability (e.g., Ordenes et al. 2019; Melumad et al. 2019)

Source selection: *ethical issues*

- Ethical and privacy issues
  - Vulnerable consumer populations
  - Legality of web scraping: copyright infringement, trespass to chattels, breach of contract, and violation of the Computer Fraud and Abuse Act

informs

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 65, No. 2, February 2019, pp. v–vi
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

From the Editor

David Simchi-Levi[a]

[a] Institute for Data, Systems, and Society, Department of Civil and Environm
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
Contact: dslevi@mit.edu, http://orcid.org/0000-0002-4650-1519 (DS-L)

ASHLEY MADISON™ ®

PNAS

Personal infidelity and professional conduct in 4 settings

John M. Griffin[a,1], Samuel Kruger[a], and Gonzalo Maturana[b]

[a]McCombs School of Business, University of Texas at Austin, Austin, TX 78712; and [b]Goizueta Business School, Emory University, Atlanta, GA 30322

Source selection: *journal policies*

- Ethical and privacy issues
  - Vulnerable consumer populations
  - Legality of web scraping: copyright infringement, trespass to chattels, breach of contract, and violation of the Computer Fraud and Abuse Act

AMA Policy on Scraping and Use of Scraped Data

Scraping data from web sites is a common practice and it was inevitable that data obtained through scraping would become the object of academic research. However, there are numerous restrictions on the collection and use of such data ranging from the policies of web site owners to laws that protect property and privacy rights. Legislation, regulation and case law related to scraping are evolving rapidly. Scraping a website is not impermissible or illegal, per se. For example, scraping one's own website is certainly permissible. Similarly, scaping another party's web site when the scraper has been given explicit permission to do so is also permissible. On the other hand, some practices related to the scaping of other party's web sites have been held to be a violation of property rights and even felony criminal acts.

Many web sites have explicit policies related to what is and is not permissible with respect to the scraping of their sites. Users often agree to adhere to these policies when they accept the terms of use of a web site.

Source selection: *advanced*

- Opportunities from **moving beyond a single source**
  - Why?
  - When?
  - How?

- **You are the designer!**

Source selection: *alternatives to web scraping*

- 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?

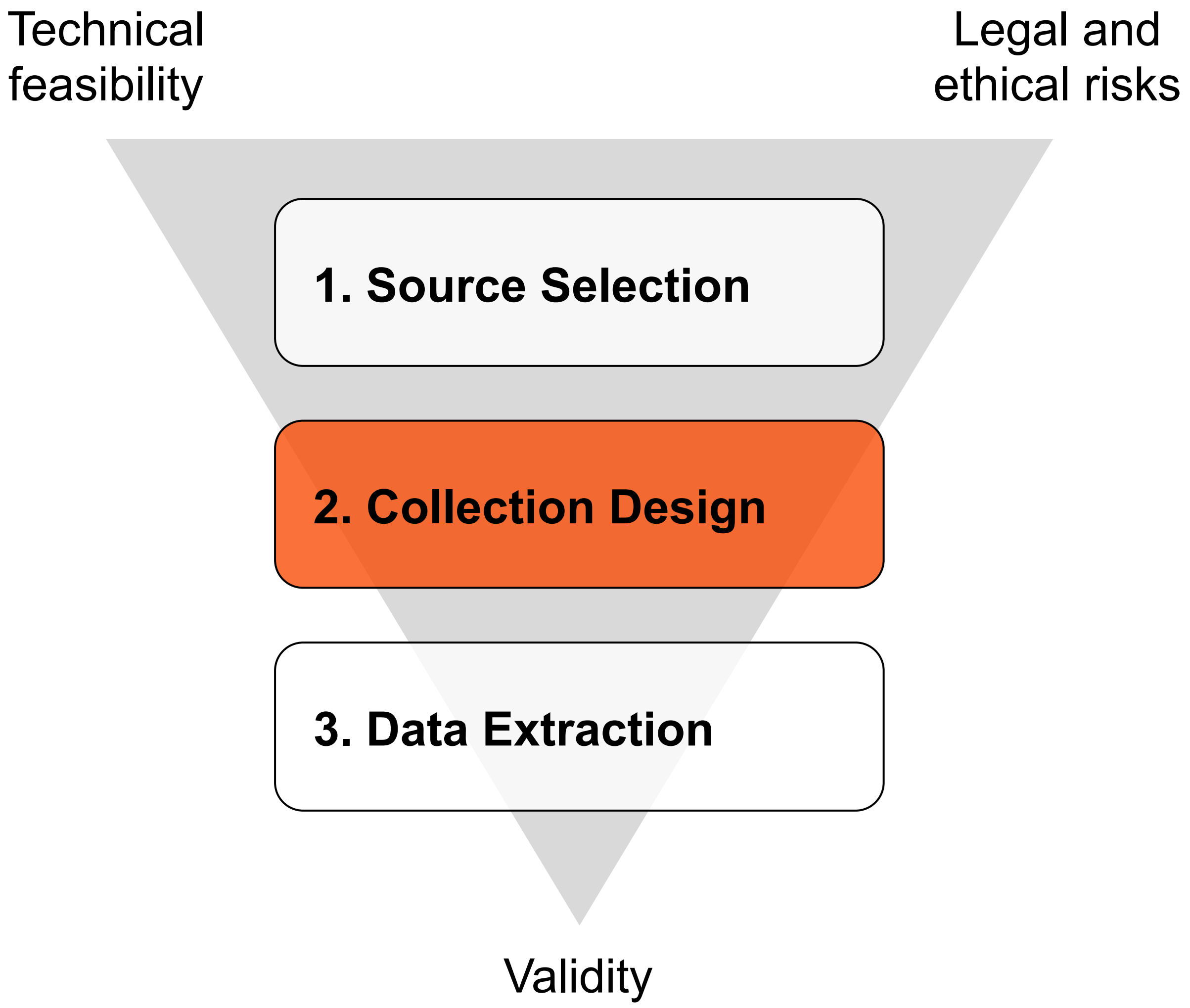
Recommender Systems and Personalization Datasets

Julian McAuley, UCSD

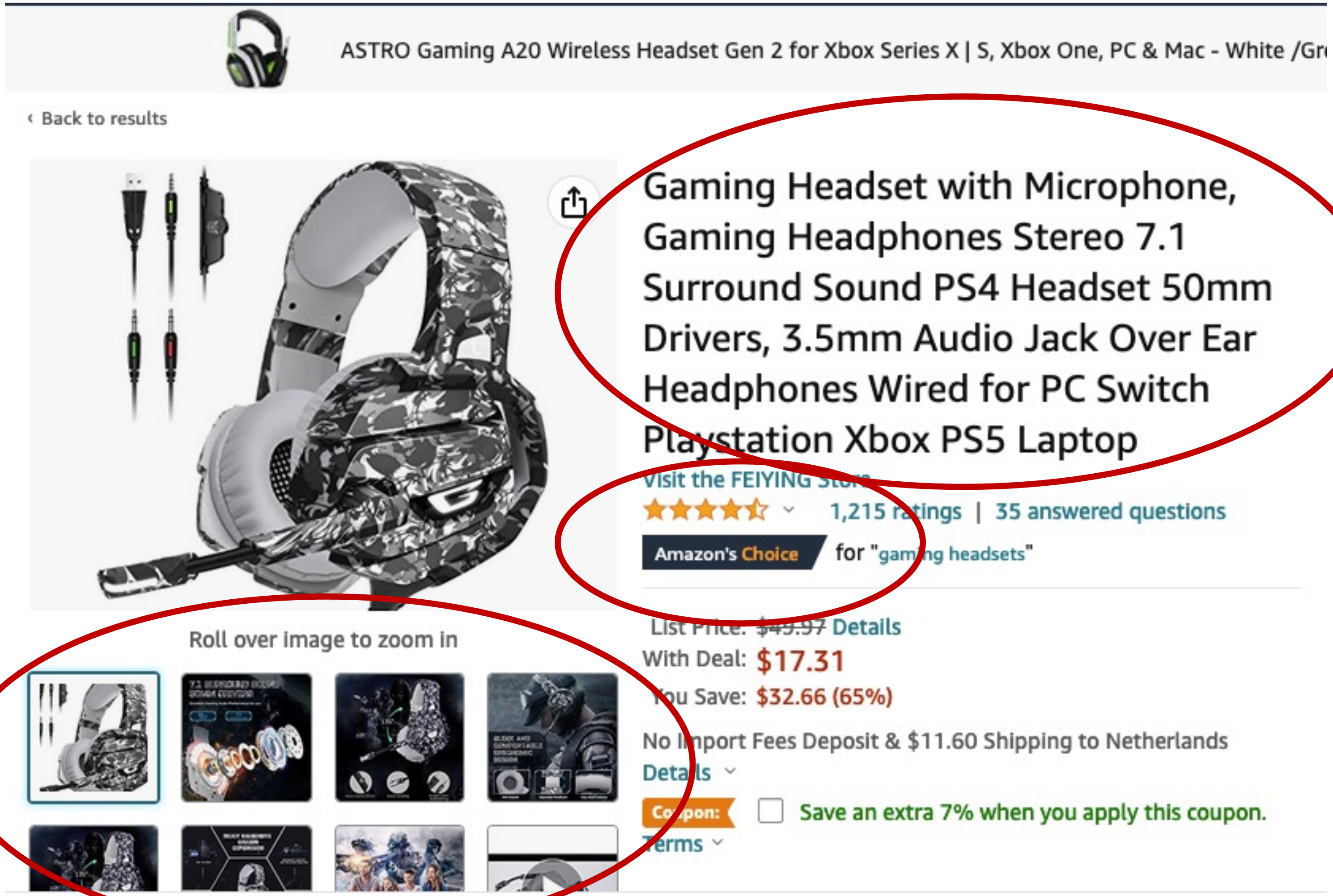
Source selection: *mapping the data context*

- 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

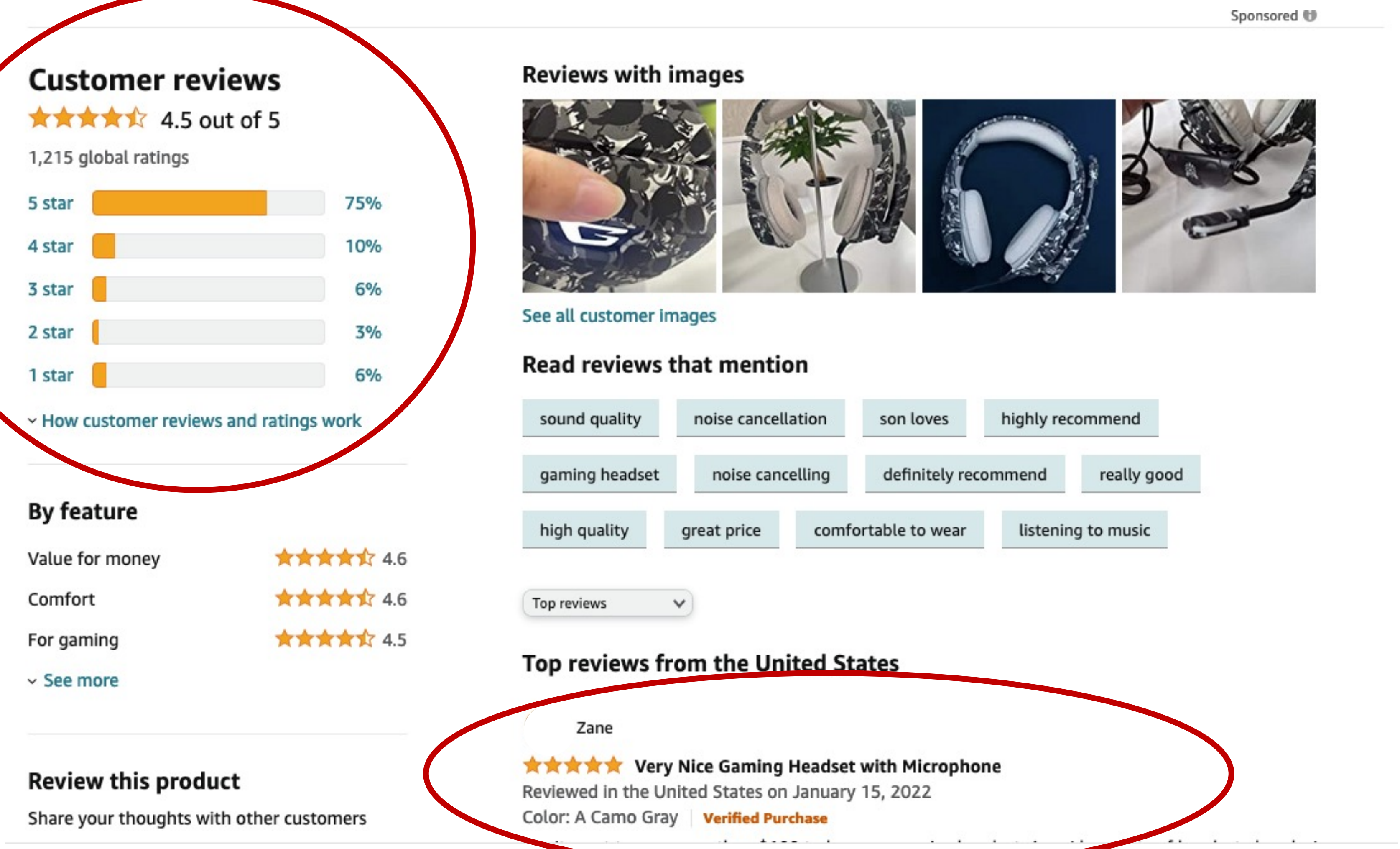
Designing the data collection



Which information to extract? *Example*



Which information to extract? *Example*



Which information to extract?

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?

<u>Validity implications</u>

- 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

<u>Legal/ethical risks</u>

- 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

<u>Technical feasibility</u>

Which information to extract?

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)

- <u>A few examples</u>
  - 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 <u>**your essential assumptions**</u> 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)

- <u>Case study:</u>
  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 **<u>your essential assumptions</u>** 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)

- <u>Case study:</u>
  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)

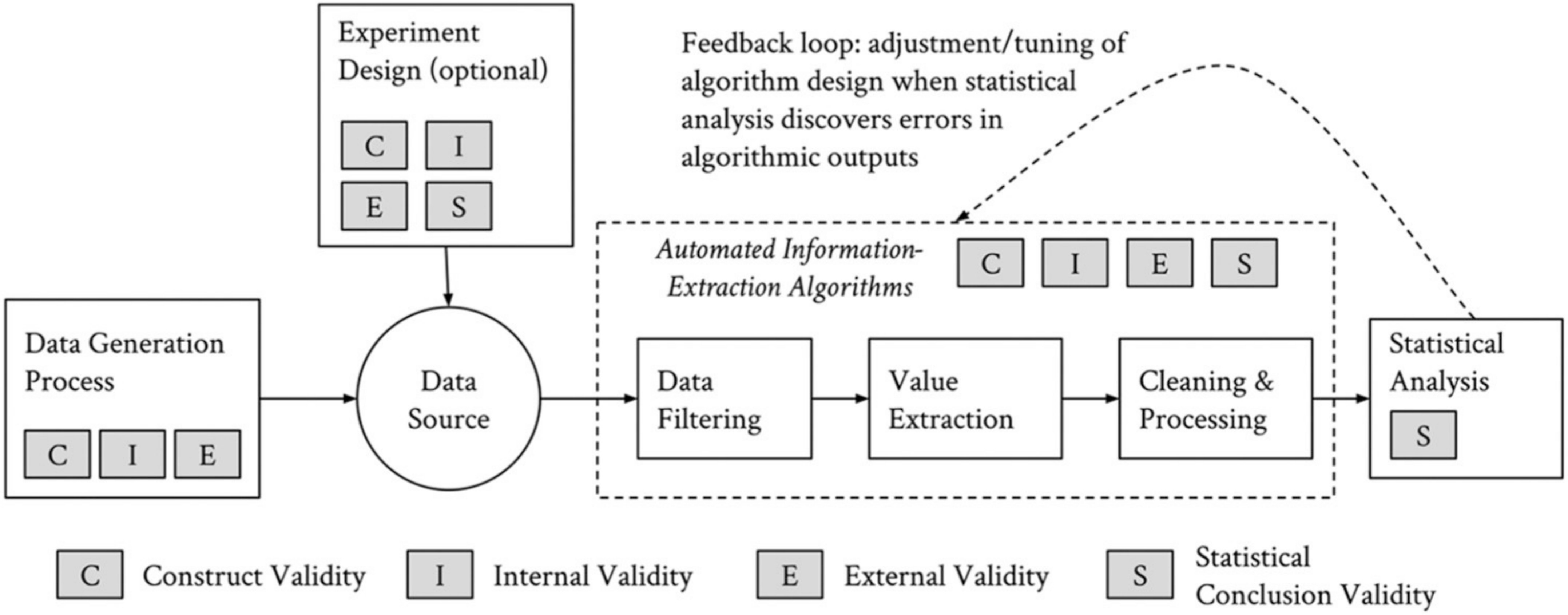
Sushi House
$$ · Japanese, Sushi Bars

1/26/2014

User A
joined on
<u>1/26/2014</u>

yelp

Data-generating mechanism



Source: Xu et al. (2020)

How to process data during the extraction?

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

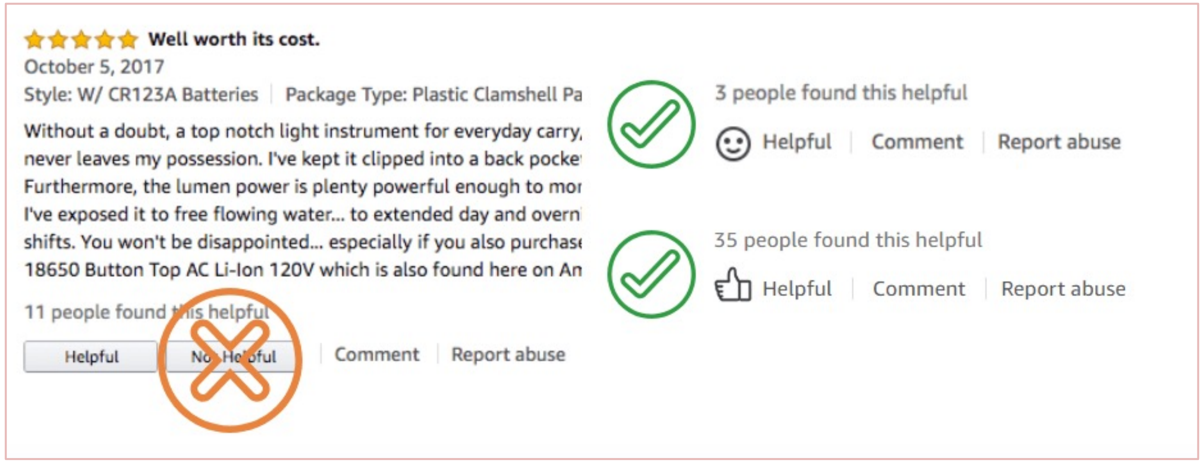
→ **Keep the raw data whenever possible**

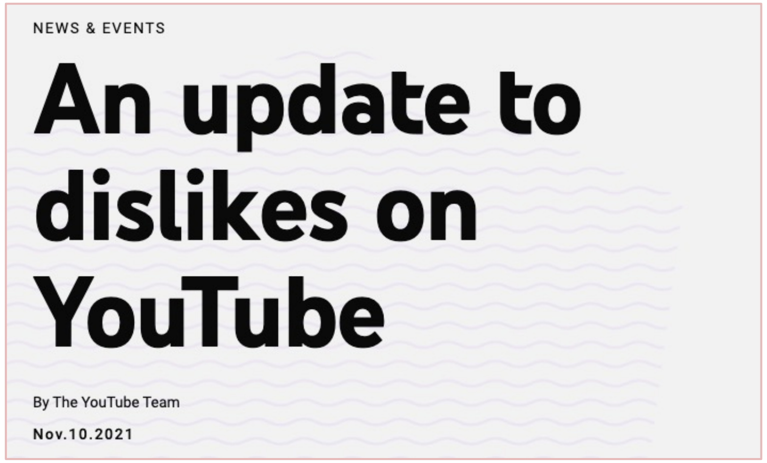
How to process data during the extraction?

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

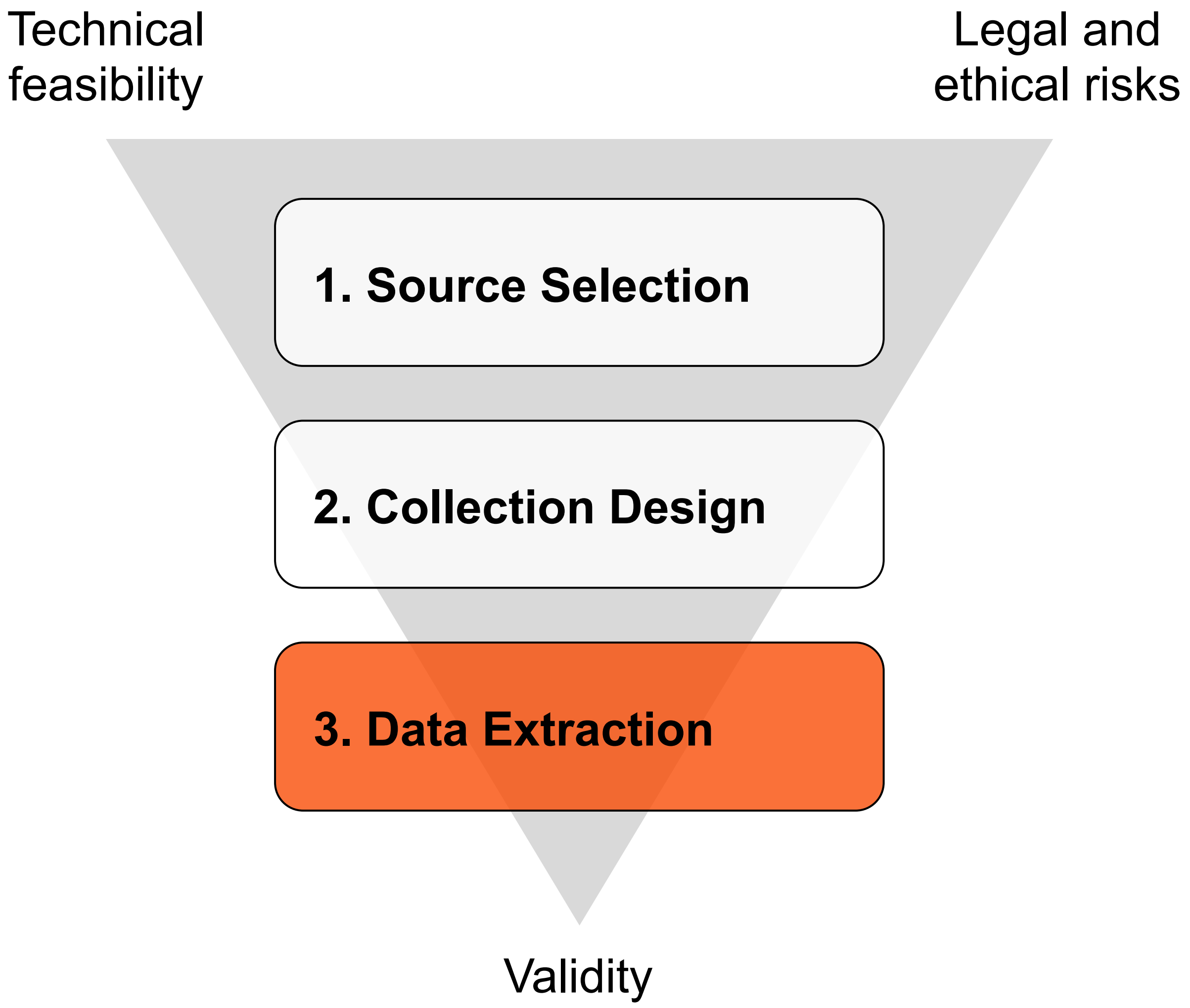
→ Keep the raw data whenever possible

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





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).

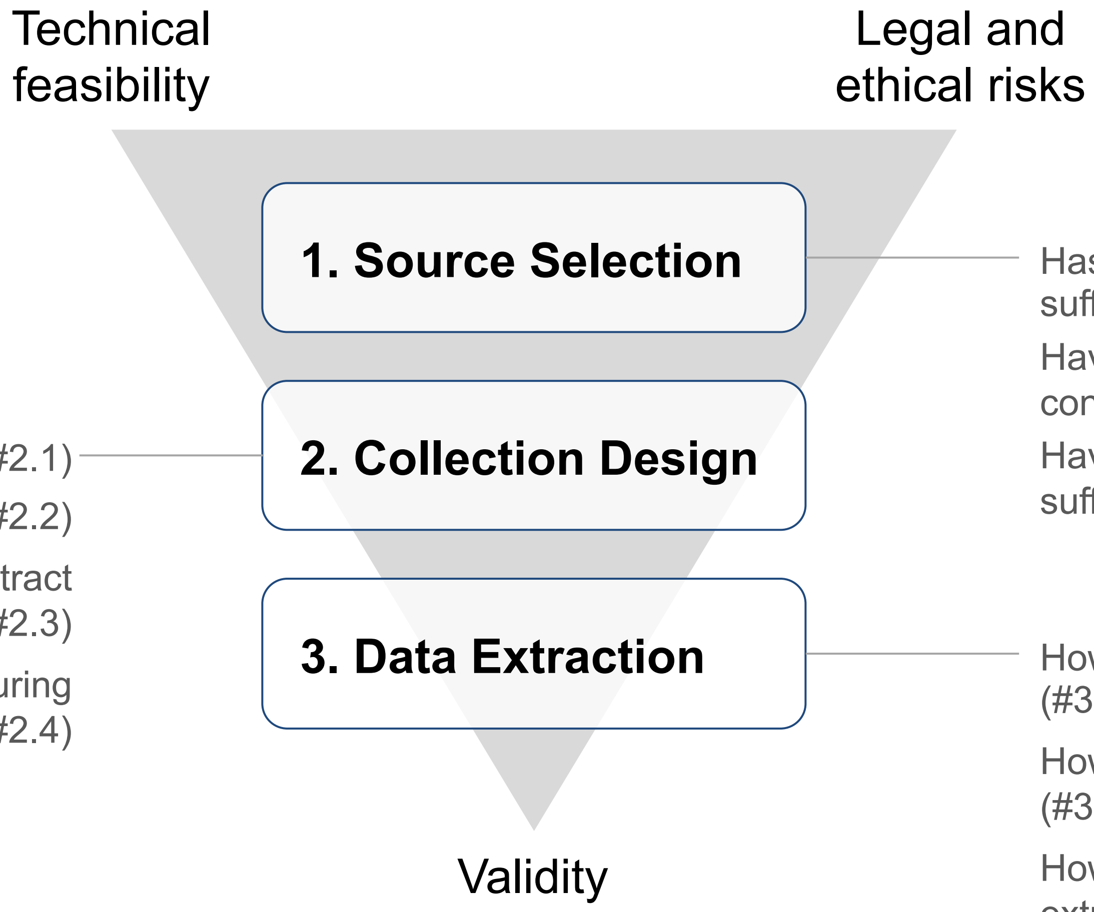
Documentation

Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

- Motivation
- Composition
- Collection process
- Preprocessing/cleaning/labeling
- Uses
- …

Methodological framework: *summary*



1. Source Selection
- Has the universe of potential data sources been sufficiently explored? (#1.1)
- Have alternatives to web scraping been considered? (#1.2)
- Have the complexities of the data context been sufficiently mapped? (#1.3)

2. Collection Design
- Which information to extract? (#2.1)
- How to sample? (#2.2)
- At which frequency to extract information? (#2.3)
- How to process the information during the extraction? (#2.4)

3. Data Extraction
- How to improve the performance of the extraction? (#3.1)
- How to monitor data quality during the extraction? (#3.2)
- How to document the data during and after the extraction? (#3.3)

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

MAKE TRADE-OFFS EXPLICIT IN YOUR PAPERS

Methodological framework: *summary*

Technical feasibility

Legal and ethical risks

1. Source Selection

- Has the universe of potential data sources been sufficiently explored? (#1.1)
- Have alternatives to web scraping been considered? (#1.2)
- Have the complexities of the data context been sufficiently mapped? (#1.3)

2. Collection Design

- Which information to extract? (#2.1)
- How to sample? (#2.2)
- At which frequency to extract information? (#2.3)
- How to process the information during the extraction? (#2.4)

3. Data Extraction

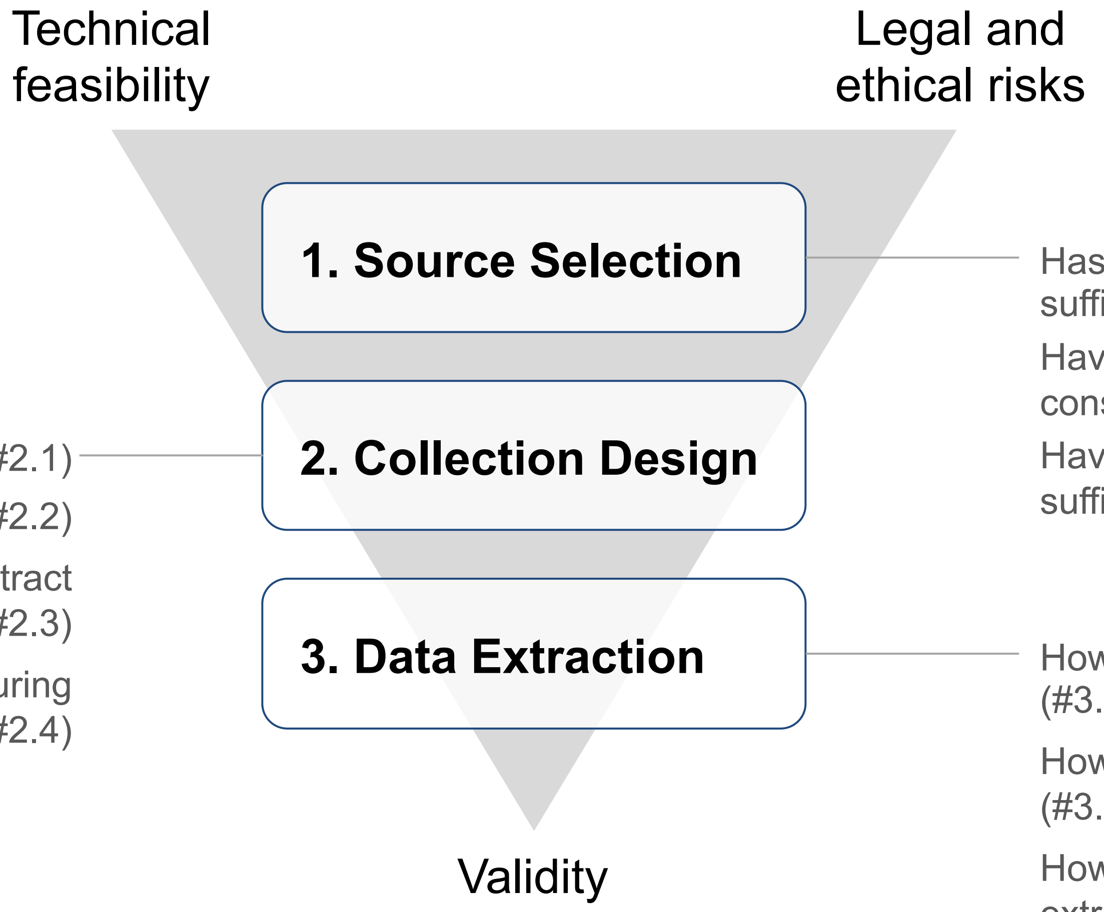
- How to improve the performance of the extraction? (#3.1)
- How to monitor data quality during the extraction? (#3.2)
- How to document the data during and after the extraction? (#3.3)

Validity

IMPORTANT: trade-offs are (almost) inevitable

MAKE TRADE-OFFS EXPLICIT IN YOUR PAPERS

Questions?



1. Source Selection

- Has the universe of potential data sources been sufficiently explored? (#1.1)
- Have alternatives to web scraping been considered? (#1.2)
- Have the complexities of the data context been sufficiently mapped? (#1.3)

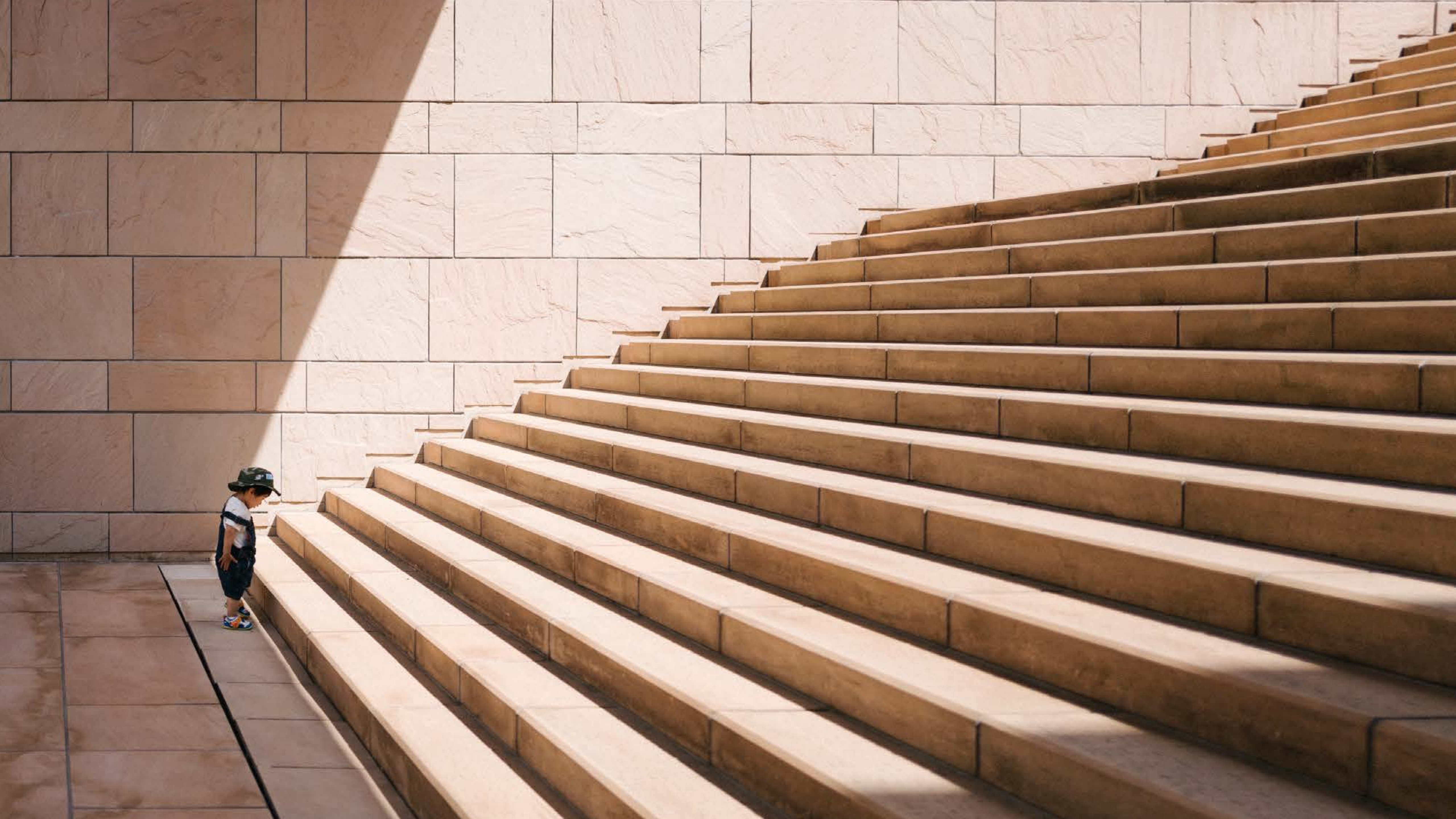
2. Collection Design

- Which information to extract? (#2.1)
- How to sample? (#2.2)
- At which frequency to extract information? (#2.3)
- How to process the information during the extraction? (#2.4)

3. Data Extraction

- How to improve the performance of the extraction? (#3.1)
- How to monitor data quality during the extraction? (#3.2)
- How to document the data during and after the extraction? (#3.3)

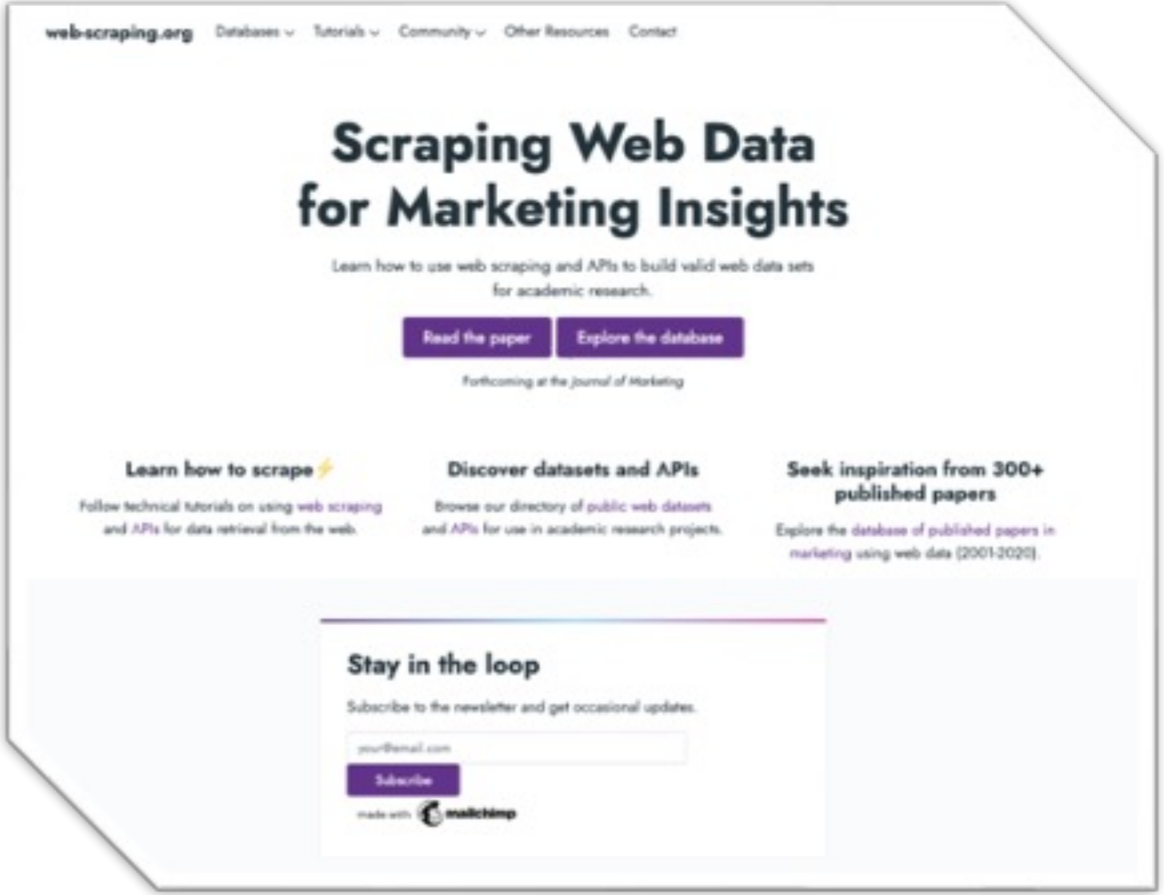
Source: Boegershausen, Datta, Borah, and Stephen (2022)



Structured approach

Accessibility

Our framework & companion website



- Explore our **database with 400+ 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.

https://web-scraping.org

Rotterdam School of Management
Erasmus University

Field Experiments in Behavioral Research:
Celebrating Heterogeneity with Causal Machine Learning

Presented by Aurélie Lemmens

Use Case: Enhancing Donor Agency to Improve Charitable Giving, *Journal of Marketing* (2023)

Agenda For This Second Part

1

The value of field experiments

2

Celebrating heterogeneity

3

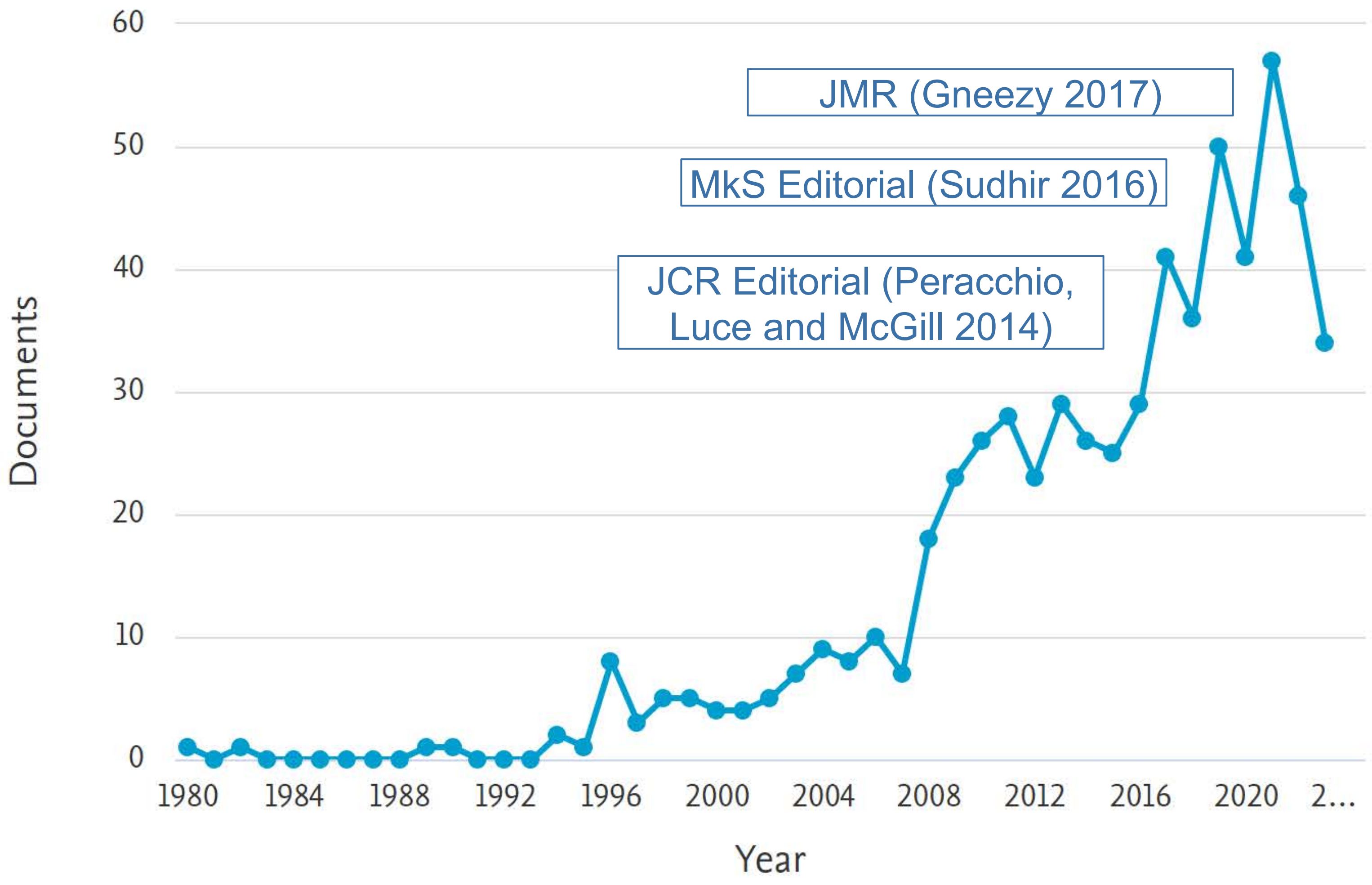
Causal machine learning

4

An application to charitable behavior

RSM Erasmus

Field Experiments Become an Important Validation Tool



(TITLE-ABS-KEY ("field study" OR "field studies" OR "field experiment" OR "field experiments") AND SRCTITLE ("journal of consumer research" OR "journal of marketing research" OR "International Journal of Research in Marketing" OR "Journal of Marketing" OR "Journal of the Association for Consumer Research" OR "Social Psychological and Personality Science" OR "Journal of Experimental Psychology-General" OR "Journal of Consumer Psychology" OR "Judgment and Decision Making" OR "Psychological Science"))

AYELET GNEEZY

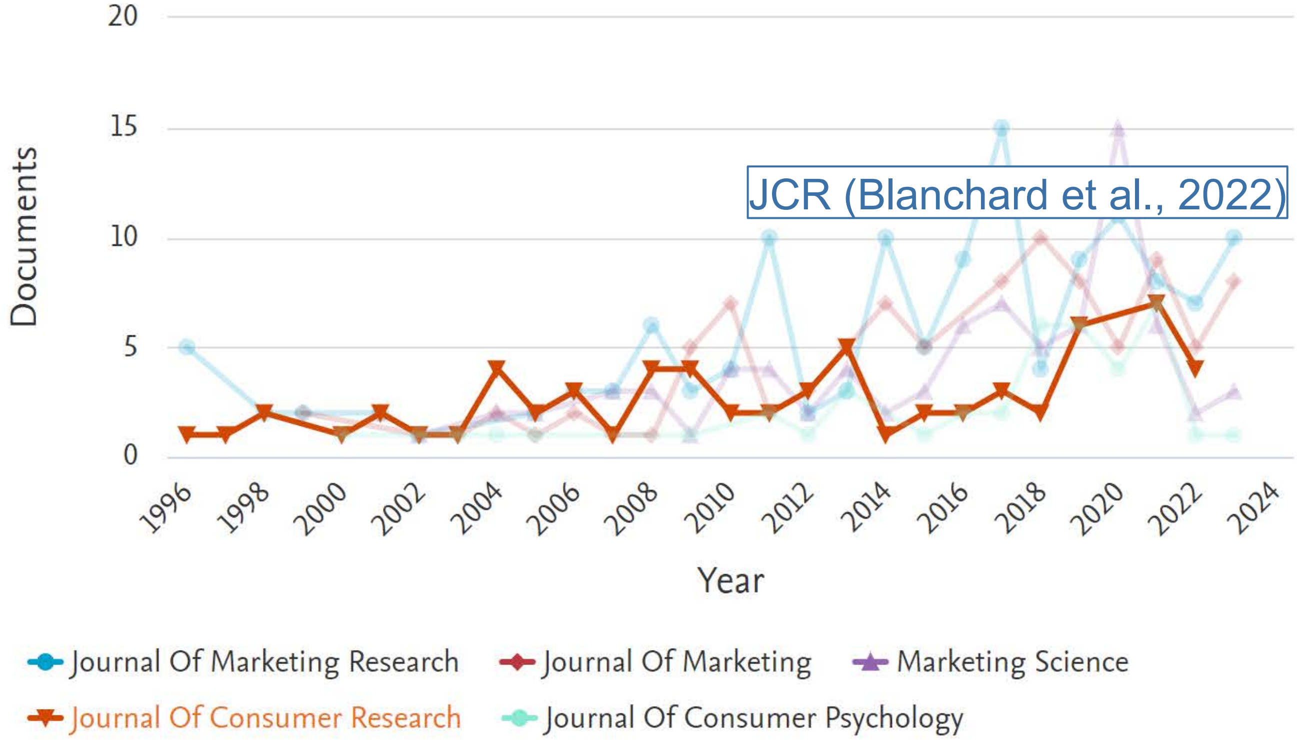
Despite increasing efforts to encourage the adoption of field experiments in marketing research (e.g., Campbell 1969; Cialdini 1980; Li et al. 2015), the majority of scholars continue to rely primarily on laboratory studies (Cialdini 2009). For example, of the 50 articles published in *Journal of Marketing Research* in 2013, only three (6%) were based on field experiments. The goal of this article is to motivate a methodological shift in marketing research and increase the proportion of empirical findings obtained using field experiments. The author begins by making a case for field experiments and offers a description of their defining features. She then demonstrates the unique value that field experiments can offer and concludes with a discussion of key considerations that researchers should be mindful of when designing, planning, and running field experiments.

Keywords: field experiments, lab experiments

Field Experimentation in Marketing Research

Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources

SIMON J. BLANCHARD
JACOB GOLDENBERG
KOEN PAUWELS
DAVID A. SCHWEIDEL



JM, JMR, MkS

➔ *Marketing-mix optimization & personalization*

JCR, JCP

➔ *External validity*

JM, JMR, MkS

JCR, JCP

Unfortunately, the labels "consumer research" and "consumer behavior" have come to connote far more than the focus of the work—just as, somewhere along the way, "consumer behavior" and "quant" came to imply a particular type of data source (and associated analysis methods) that is primarily used to study relevant questions, data, and methodology?

Nevertheless, the rigid lines dividing the artificially created sub-disciplines are our own making, for better and worse. One way to address this divide and consequently expand the reach of our research beyond those who specialize in our particular sub-disciplines is to use more than one type of data source when addressing a consumer research question. Such data richness is the key theme of this article.

Blanchard, Goldenberg, Pauwels, Schweidel (2022), Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources, Journal of Consumer Research, 49(2), 359–372.

Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources

SIMON J. BLANCHARD
JACOB GOLDENBERG
KOEN PAUWELS
DAVID A. SCHWEIDEL

TYPOLOGY OF DATA SOURCES IN *JCR* (2018–2021)

Method	Co-occurrence					Data source statistics	
	Lab. exp.	Obs. data	Survey	Field exp.	Meta-ana.	Used at least once (%)	% that are data rich
Laboratory experiment	175	34	21	27	1	86.21	38.86
Observational data	34	55	25	4	0	27.09	87.27
Survey	21	25	40	2	0	19.70	87.50
Field experiments	27	4	2	27	0	13.30	100.00
Meta-analysis	1	0	0	0	3	1.48	33.33
Entire sample							40.39

Blanchard, Goldenberg, Pauwels, Schweidel (2022), Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources, Journal of Consumer Research, 49(2), 359–372.

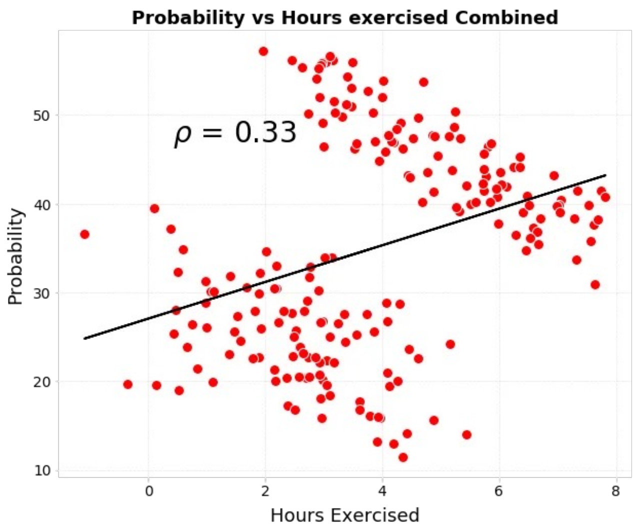
Field Experiments

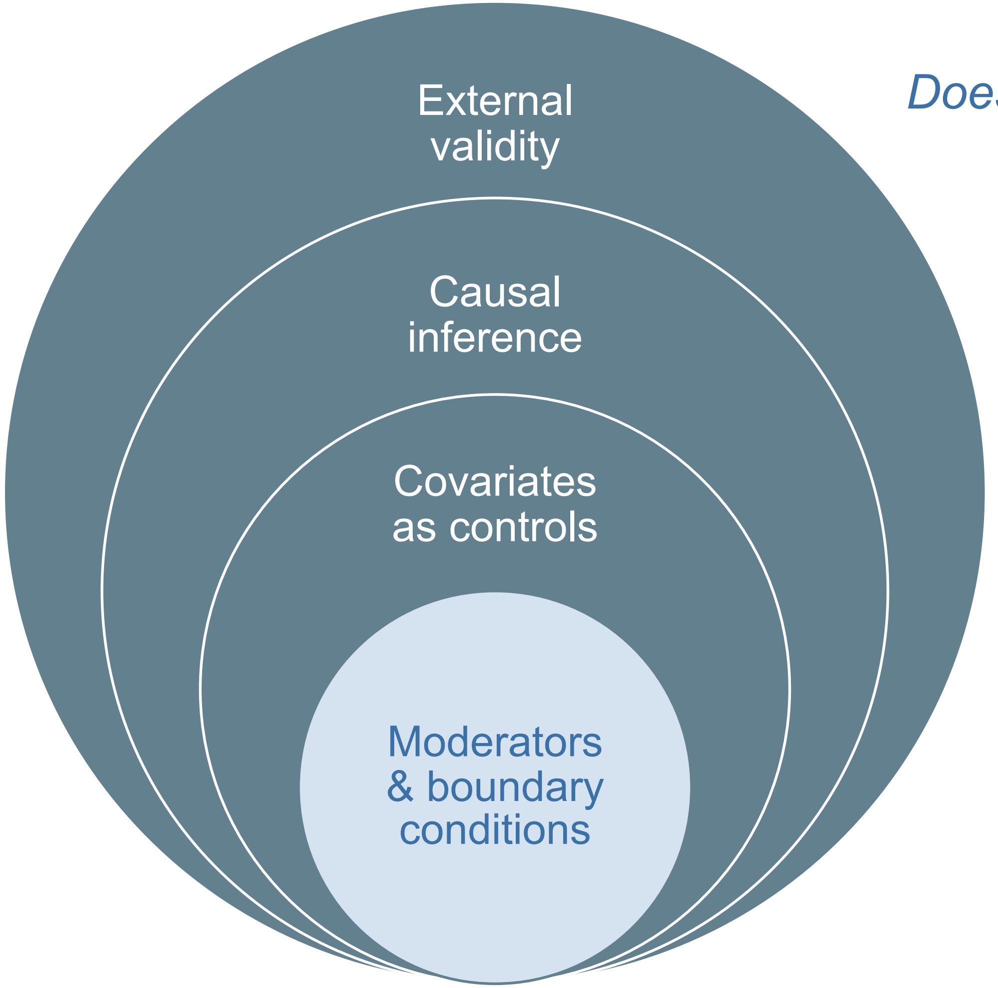
Field experiments are experiments where participants do not know they are taking part in a research study; they are unaware that an experimental manipulation has occurred and are engaged in real consumption behavior, which is observed and/or measured unobtrusively

Morales & On Amir (2017)

Morales and On Amir (2017). Keeping It Real in Experimental Research—Understanding When, Where, and How to Enhance Realism and Measure Consumer Behavior, Journal of Consumer Research, 44 (2), 465–476.

The Value of Field Experiments





Does a theory apply to real life?

What is the effect size in “real life”?
What are second-order effects?

Get rid of heterogeneity (noise!)

Celebrate heterogeneity for theoretical and managerial insights

What is the “real-world” effect?

How the Eyes Connect to the Heart: The Influence of Eye Gaze Direction on Advertising Effectiveness

RITA NGOC TO
VANESSA M. PATRICK

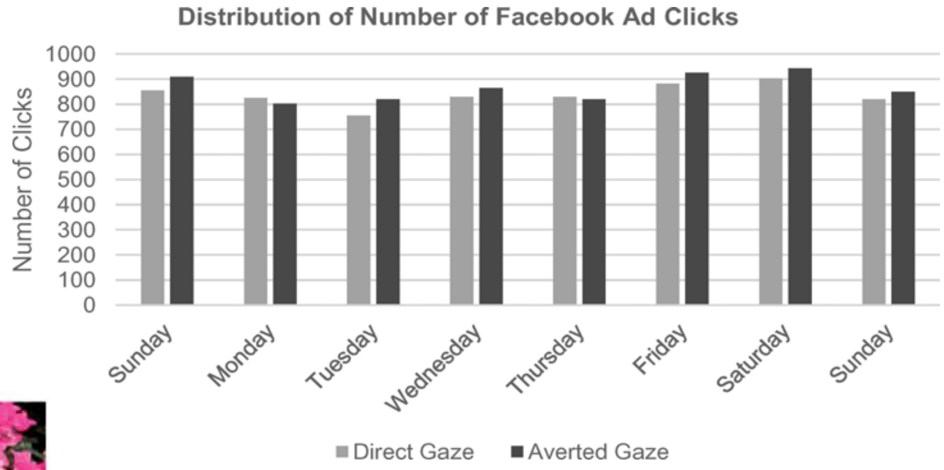
Facebook Ads

Direct Gaze

Summer Fashion 2019
New Summer Fashion and Accessories
Come shop with us for the summer season!
Learn More

Averted Gaze

Summer Fashion 2019
New Summer Fashion and Accessories
Come shop with us for the summer season!
Learn More



Second-order effects

Ngoc To, Patrick (2021), How the Eyes Connect to the Heart: The Influence of Eye Gaze Direction on Advertising Effectiveness, Journal of Consumer Research, 48(1), 123–146.

Field Experiment & External Validity?

Similar to lab studies, one real-world setting is unlike another. Understanding generalizability requires us to explore moderations and to test for the asserted pattern of interactions

Lynch (1999)

Lynch (1999). Theory and external validity. *Journal of the Academy of Marketing Science*, *27*, 367-376.

Covariates as Controls

Obligatory Publicity Increases Charitable Acts

ADELLE X. YANG
CHRISTOPHER K. HSEE

We give all blood donors the option to wear this donor stamp ...

We ask all blood donors to comply and wear this donor stamp ...

TABLE 1

EFFECT OF THE OBLIGATORY-PUBLICITY CAMPAIGN STRATEGY ON THE DONATION DECISION IN THE PRESE
STUDY 2

	Donate (yes/no)	
	(1)	(2)
Campaign strategy (OP = 1 vs. VP = 0)	1.34 (.28)***	1.31 (.29)***
School year (1–4)		−.27 (.09)**
Gender (M = 0, F = 1)		−.21 (.22)
N	8504	8504
Cox & Snell R^2	.003	.005

Notes.—
**$p < .01$,
***$p < .001$.

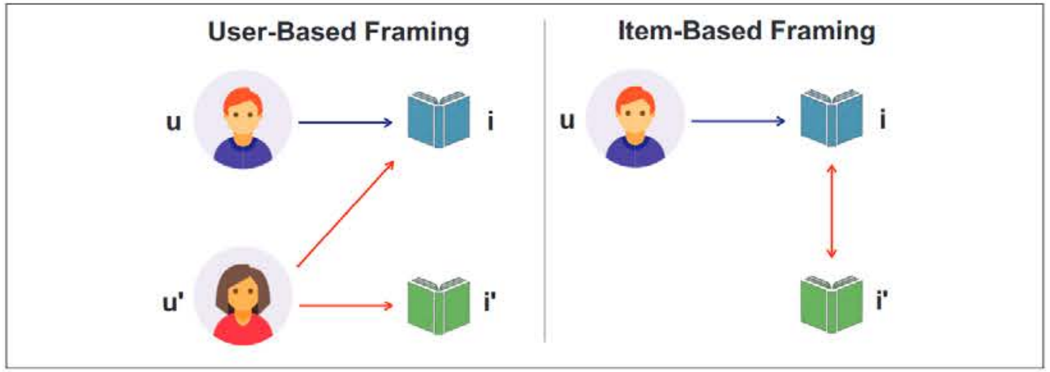
Yang and Hsee (2022), Obligatory Publicity Increases Charitable Acts, *Journal of Consumer Research*, 48(5), 839–857

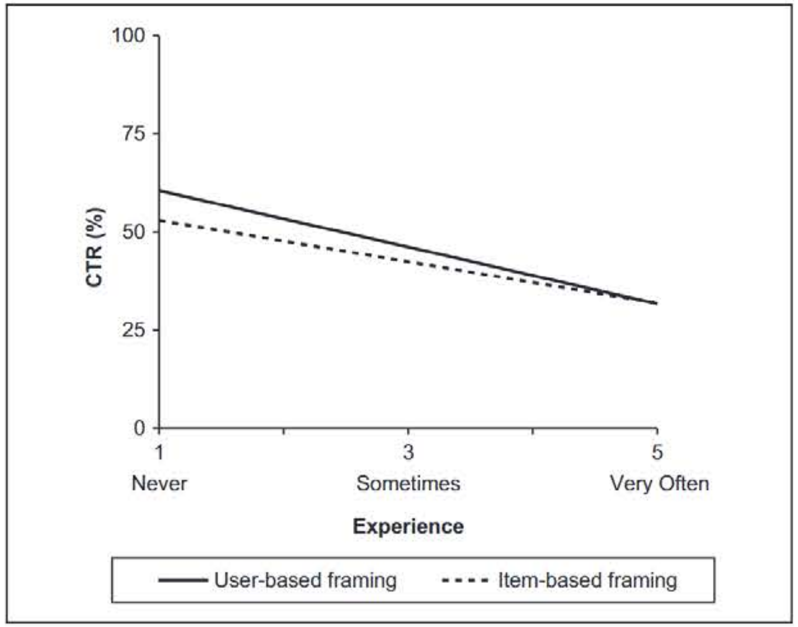
Covariates as Moderators

Article

Making Recommendations More Effective Through Framings: Impacts of User- Versus Item-Based Framings on Recommendation Click-Throughs

Phyliss Jia Gai and Anne-Kathrin Klesse





Allows us to understand boundary conditions and mechanisms

Gai and Klesse (2019). Making Recommendations More Effective Through Framings: Impacts of User- Versus Item-Based Framings on Recommendation Click-Throughs. *Journal of Marketing*, *83*(6), 61–75.

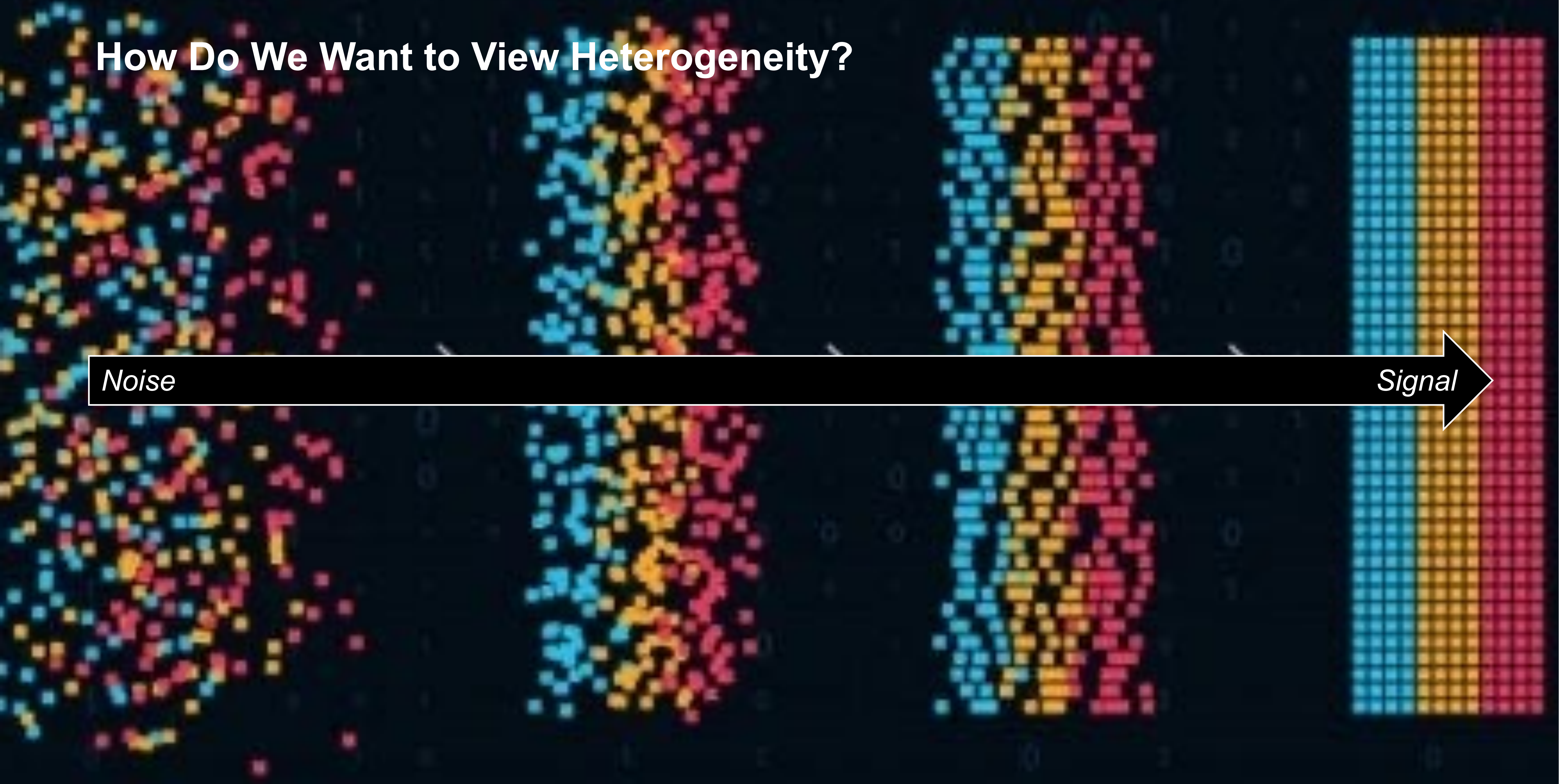
Celebrating Heterogeneity

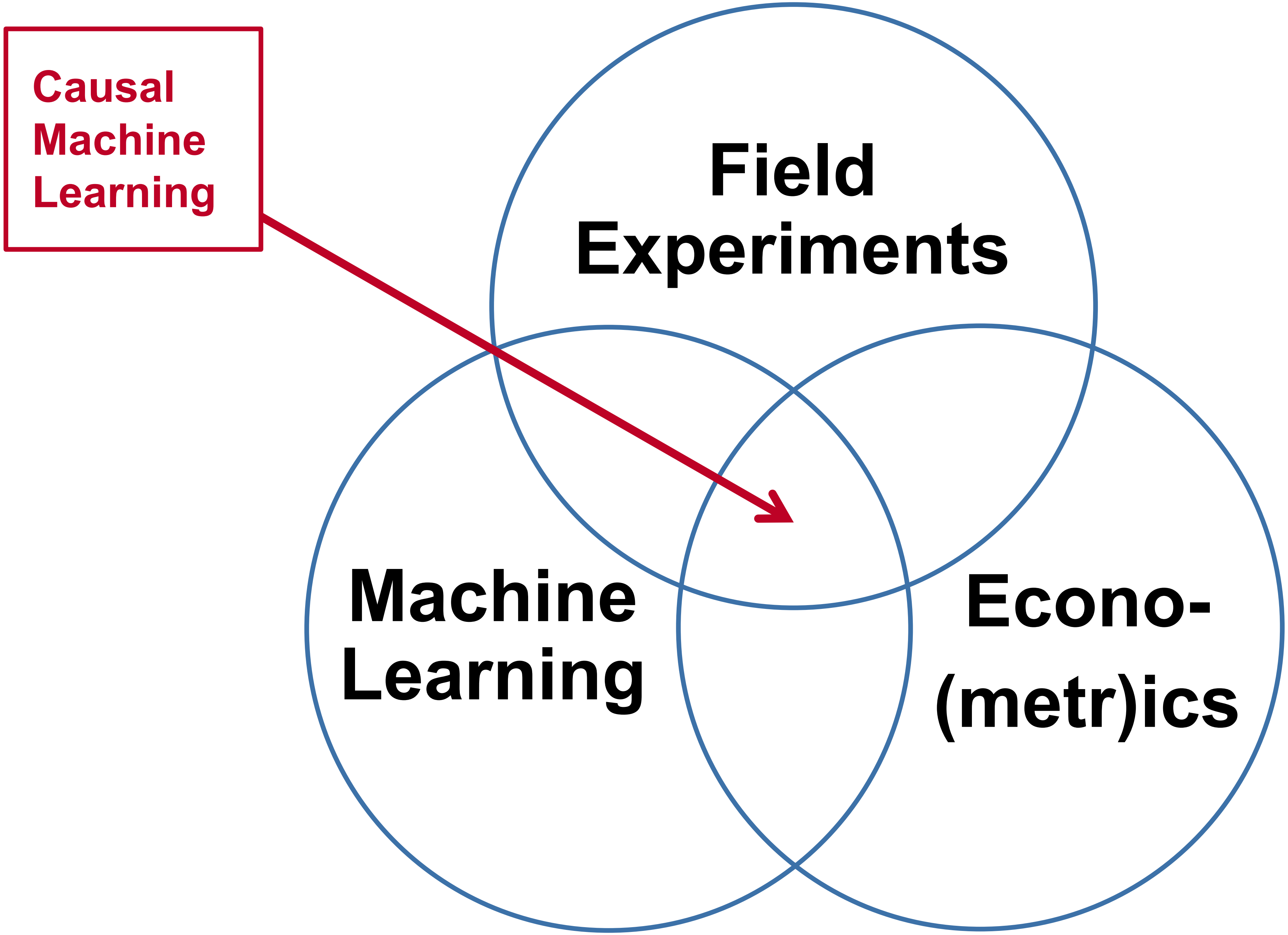
Recurring Donors	Contact	2020 Donation Status	Priority	Tags	2020 Donation	2019 Donation	2018 Donation
Mashari		Plans to donate in 2020	High	#social #fundraising	$0	$1000	$0
Eddie		Donated in 2020	Medium	#fooddelivery	$1000	$5000	$1000
Ayala		Not donating in 2020	Not relevant	#social	$0	$2000	$2000

One-time Donors	Contact	2020 Donation Status	Priority	Tags	2020 Donation
Brett		Plans to donate in 2020	High	#social #fundraising	$0
Daniel		Plans to donate in 2020	High	#social #fundraising	$0
May		Donated in 2020	Medium	#fooddelivery	$1000
Omri		Plans to donate in 2020	High	#fundraising	$0

 # customers
 # variables
 # time periods

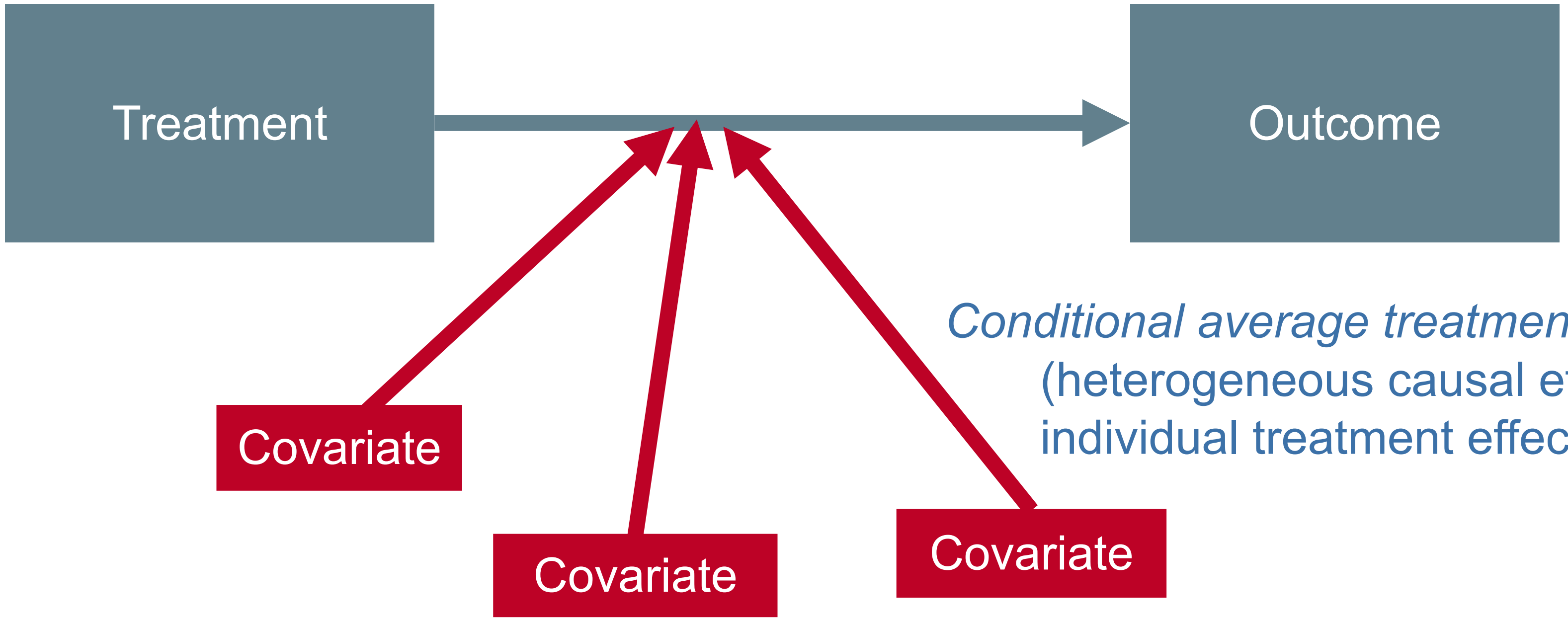
How Do We Want to View Heterogeneity?




Causal
Machine
Learning
Field
Experiments
Machine
Learning
Econo-
(metr)ics

Causal Machine Learning

Predicting CATE = How do covariates moderate an "average treatment effect" ?



Conditional average treatment effects
(heterogeneous causal effects,
individual treatment effects)

$$CATE_i = \underbrace{\mathrm{E}[y_i(1)|X_i]}_{\text{treated}} - \underbrace{\mathrm{E}[y_i(0)|X_i]}_{\text{controlled}}$$

X_i: *covariates*
y_i: *outcome*

How Does It Work?

Data are split in two and used as follows:

Trees partition the covariate space

e.g., gender(discrete)
age (continuous)

Node split that maximizes “accuracy”

male vs female
age < 25 vs age > 25

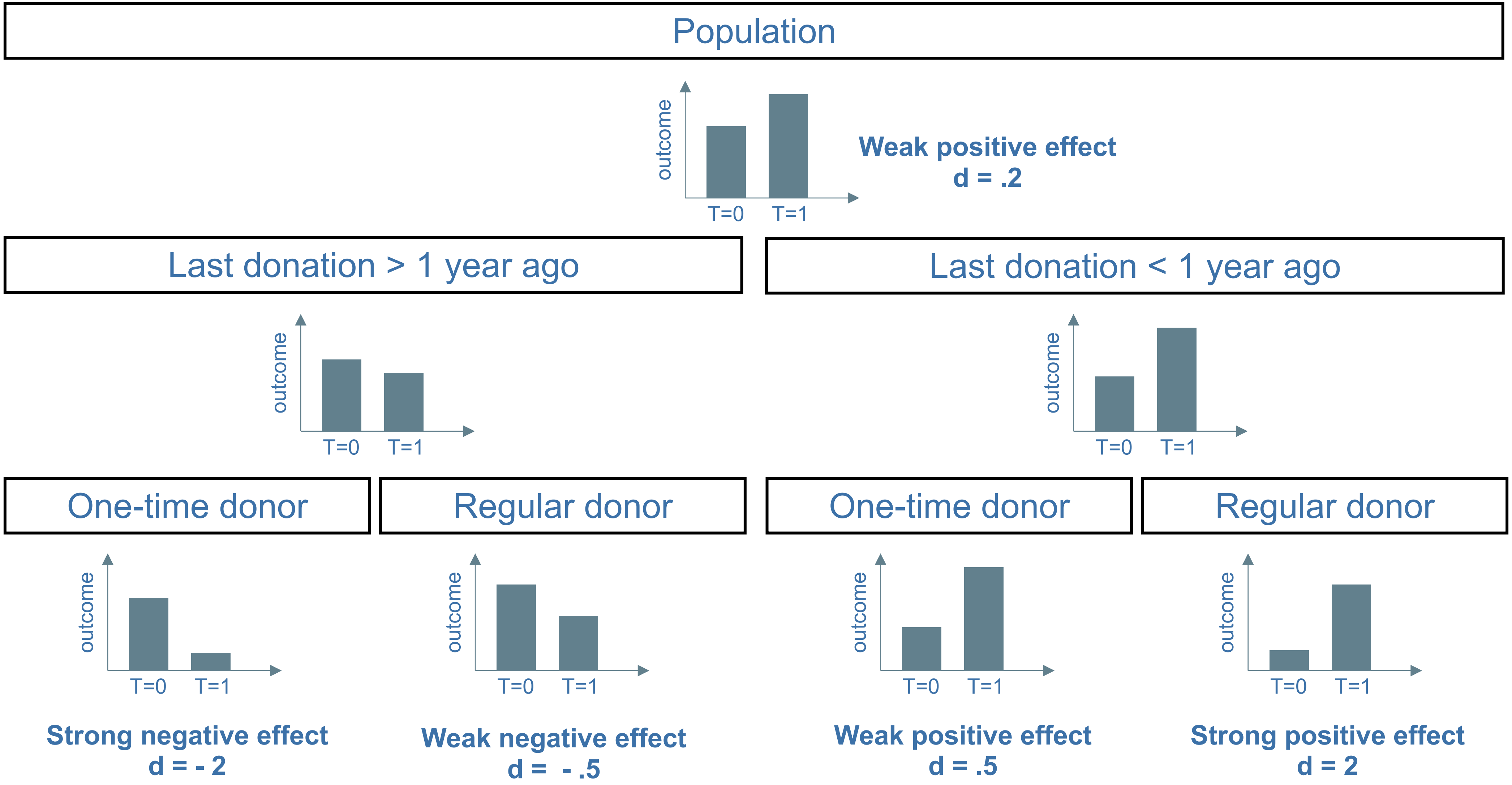
Machine Learning

Treatment effect per node: difference in outcome

“homogeneous” population” in a node

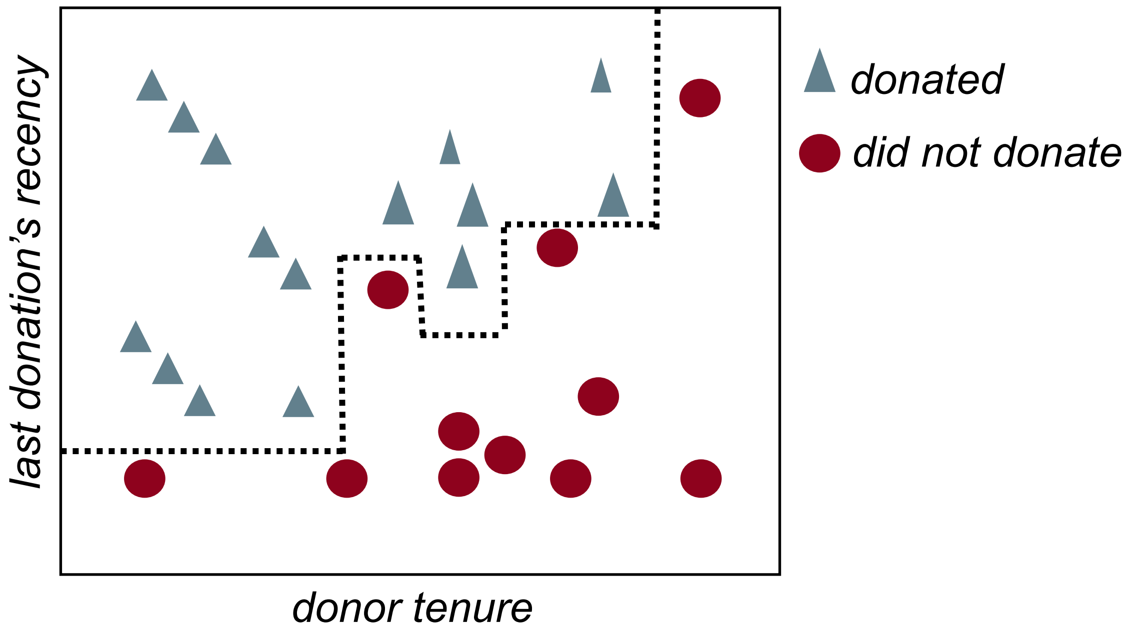
Causal Inference

Causal Forests



Causal Forests

- One of the most popular methods is "Causal Forests" (generalized random forests GRF)
  - Handles many covariates
  - Allow for a flexible moderating shape (many step functions, many trees)
  - Control over potential overfitting



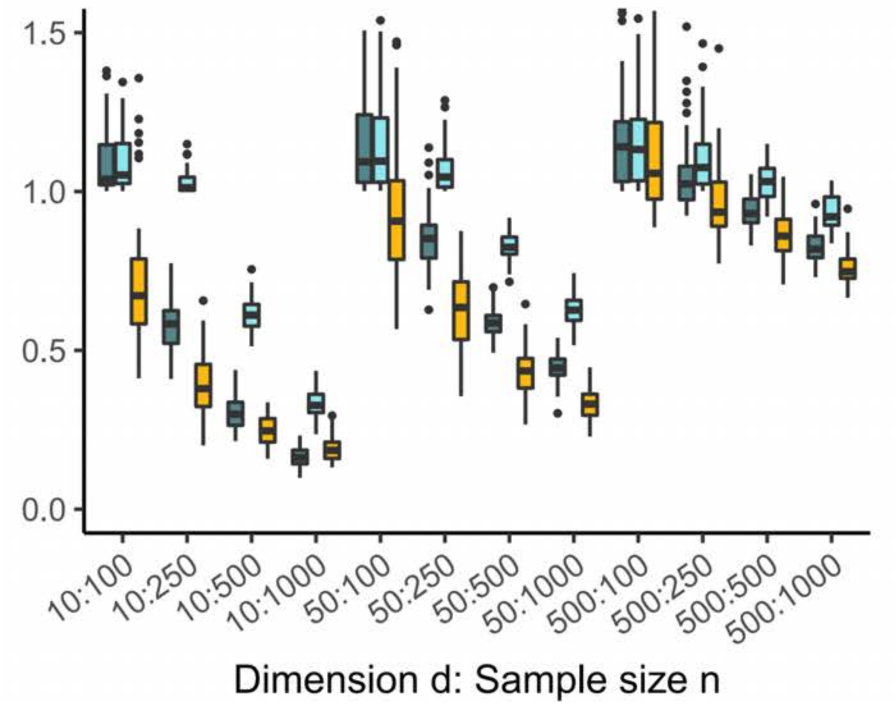
This is not (*bad*) data mining

- Causal forests ***systematically*** evaluates the result of RCT, find groups and get ***correct standard errors and confidence intervals*** about effects.
- They assess whether the results reflect “real” heterogeneity in the effect
  - **BLP test for heterogeneity** (Chernozhukov, Duflo et al. 2020)
- They have well-established statistical properties (**consistency**).

Which Data?

1. ***Randomized control trials***
   - +/- randomized
     – Conditional exchangeability (outcome independent of the assignment in each node)
     – If needed, use propensity scores
   - Sample size depends on # covariates
2. ***Treatment***
   - Continuous or discrete
   - At least two conditions (> 2 conditions: multi-arm causal forest)

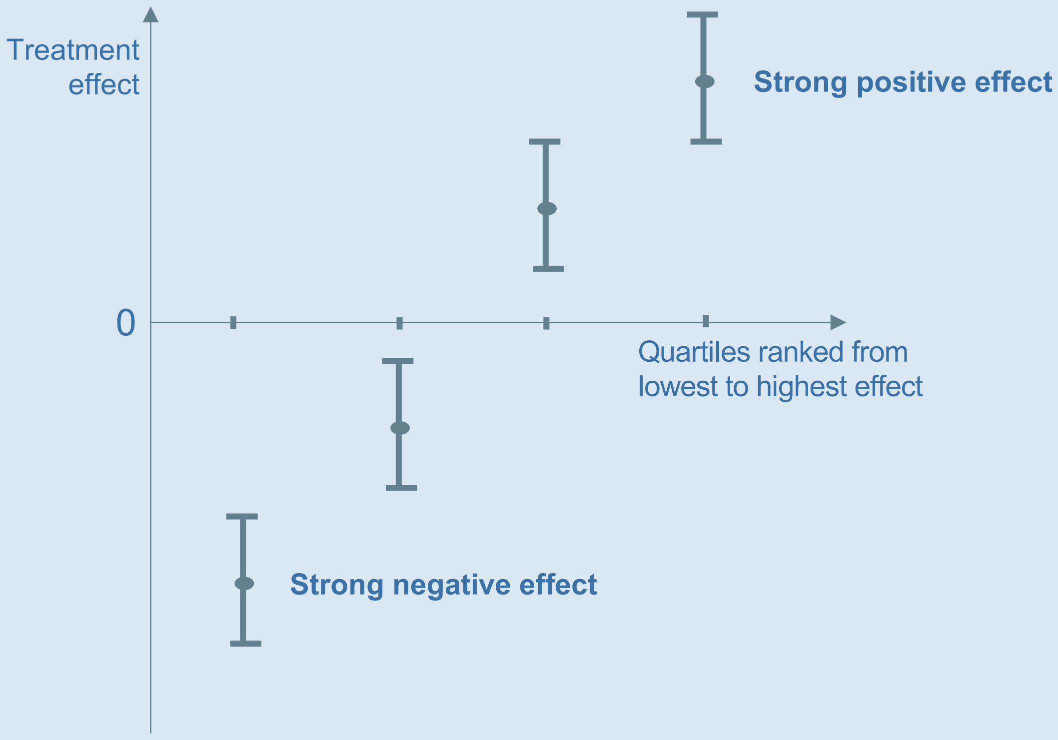
Mean squared error



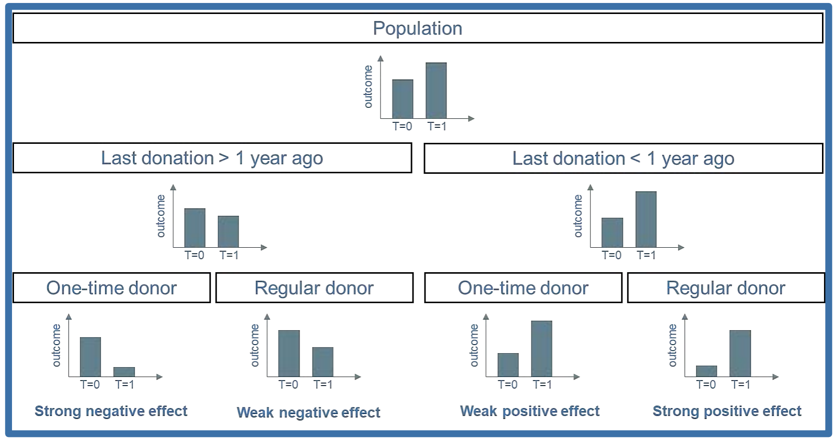
3. ***Outcome***
   - Continuous or discrete
4. ***Covariates***
   - Small to large # covariates, discrete or continuous

Result Outlook

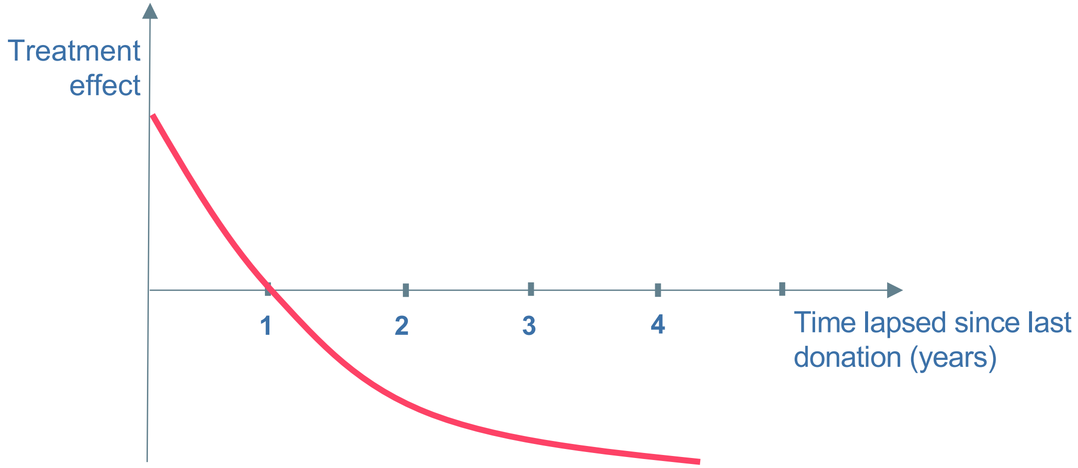
- **Group Average Treatment Effect**



How is the effect distributed across the population?



- **Partial Dependence Plots**



How does a specific covariate moderate the treatment effect?

Easy Implementation

#Load your data in R

Covariates = x, outcome of interest = y, treatment = w

#Load the grf package

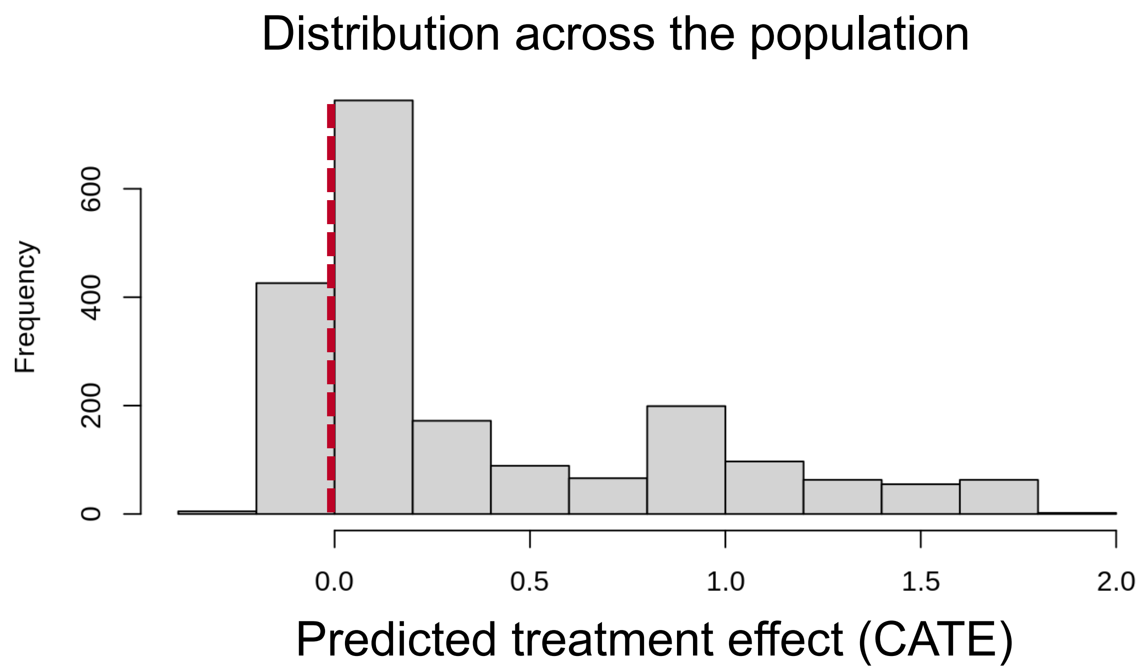
```
library(grf)
```

#Estimate the model

```
mymodel = causal_forest(X = x, Y = y, W = w)
```

#Generate predictions of CATE per observation

```
predictions = predict(mymodel)$predictions
hist(predictions)
plot(x[,1],predictions)
```



Estimation time: ~ 1 minute for 5,000 observations and 10 covariates

My Own Experience

Article

AMERICAN MARKETING ASSOCIATION

Journal of Marketing
2023, Vol. 87(4) 636-655

DOI: 10.1177/00222429221148969
journals.sagepub.com/home/jmx
Sage

Enhancing Donor Agency to Improve Charitable Giving: Strategies and Heterogeneity

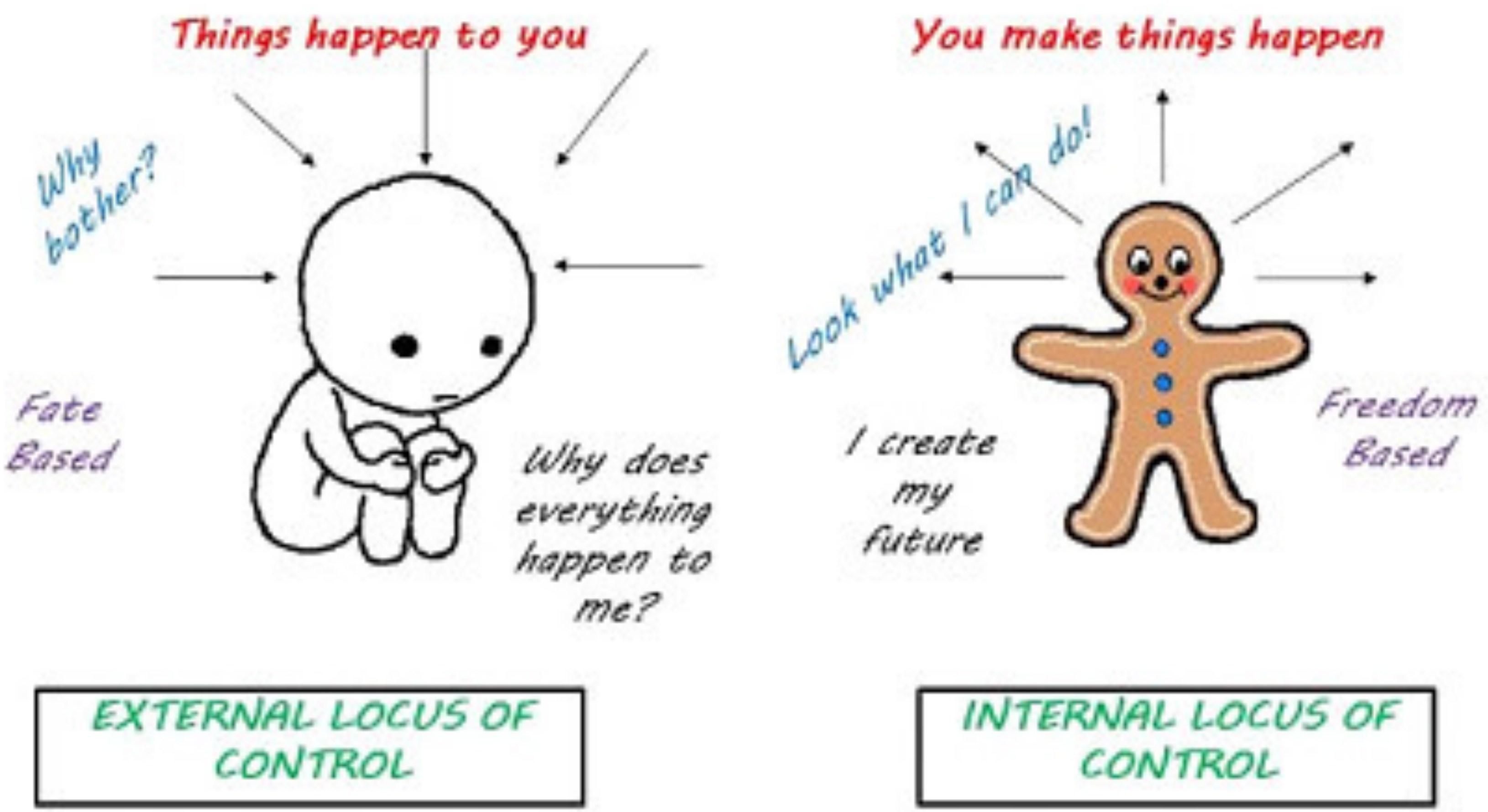
Emilie Esterzon, Aurélie Lemmens, and Bram Van den Bergh

Collaboration with a European charity

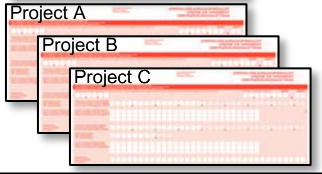
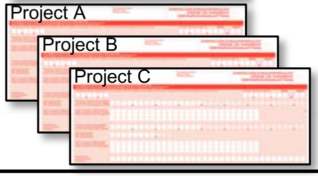
Large database: +40k donors and +100 covariates

Can you find us a cost-efficient alternative to gifts for our next fundraising campaign?

We Gave Donors a Sense of Agency



Our Field Study (n = 40,893)

	Low Targeting-via-Options	*High Targeting-via-Options*
Low Targeting-via-Amounts	Project A Project B Project C €48, €88, €120  **€18,535 [a]**	Project A Project B Project C €48, €88, €120  **€ 23,538 [bc]**
High Targeting-via-Amounts	Project A = €48 Project B = €88 Project C = €120 €48, €88, €120  **€21,432 [ab]**	Project A = €48 Project B = €88 Project C = €120 €48, €88, €120  **€26,277 [c]** **+42%**

Values without a common superscript (a, b, c) are significantly different from each other at the 5% significance level

Why Does Agency Work?

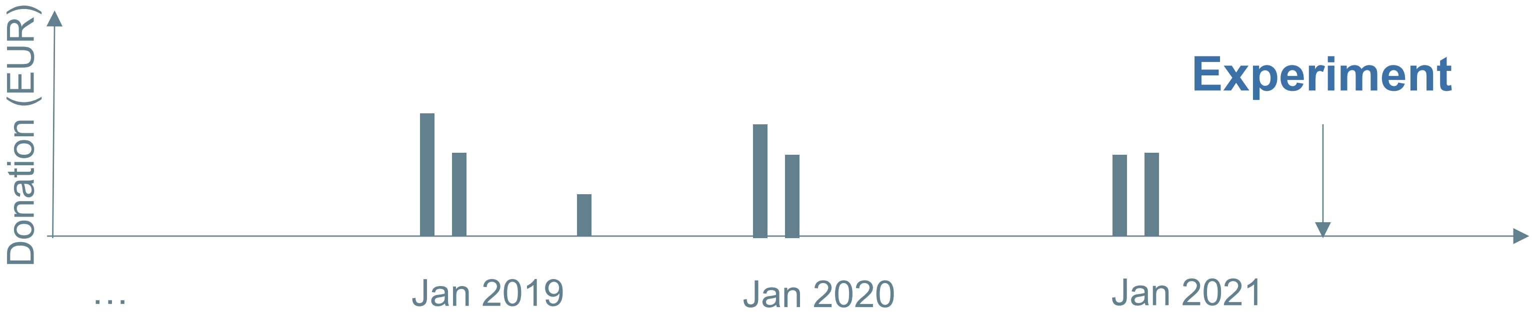
- ***Economic Benefits:***
  - Preference matching (Arora et al. 2008)
- ***Psychological Benefits:***
  - Reduced perceived uncertainty (e.g., pre-determined victim, Small & Loewenstein 2003)
  - Increased perceived impact (donors solve a specific problem, Fuchs et al. 2020), in accordance with the theory of impact philanthropy (Duncan 2004)

Why Dig into Heterogeneity?

- **But these effects may depend on:**
  - Economic factors, e.g., wealth
    - Past research centered around the "rich and powerful"(Kessler et al. 2019)
  - Cultural factors, e.g., autonomy vs. embeddedness (Fuchs et al. 2020)
  - Perceived psychological costs
    - Emotional conflicts (Ein-Gar et al. 2021) raised by fairness considerations
  - Other individual differences:
    - Generosity (Karlan and Wood 2017)
    - Time constraints, expertise (Butera and Houser 2018)

More than 15 years of past donation data

For each donor, we observe when they gave and how much they gave

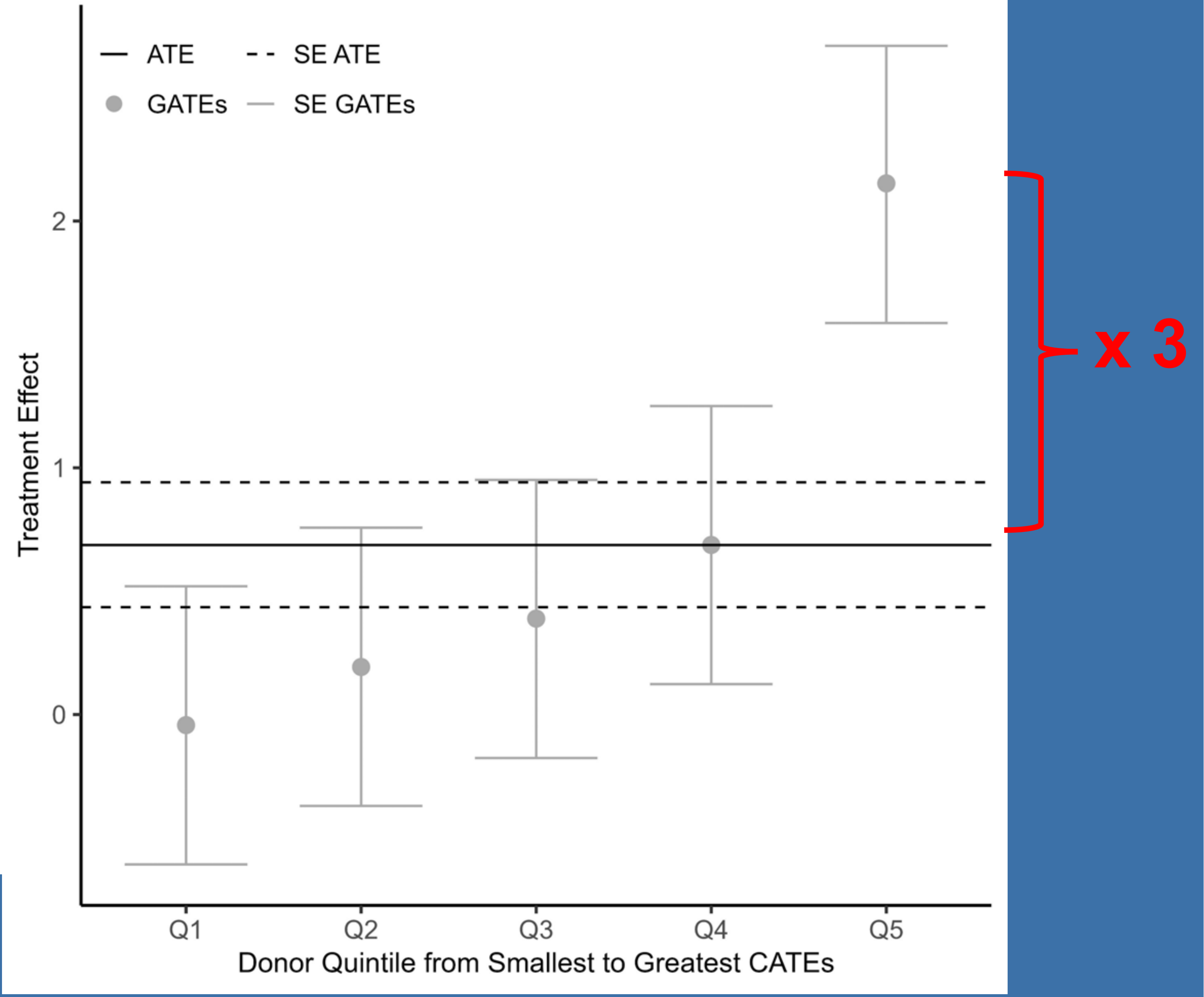


- **Tenure (in days)**
- **RFMC variables**
  - Recency (in days)
  - Frequency (number of donations per year)
  - Monetary value (total donation € per year)
  - Clumpiness
- **Donation habits or routines**
  - YoY range
  - Share of past donations of €48, €88, or €120
  - Share of gifts in popular months
  - Number of gifts in February
- **Demographics** (Individual | Company; Language A | B)

Aggregate Results Mask Substantial Heterogeneity

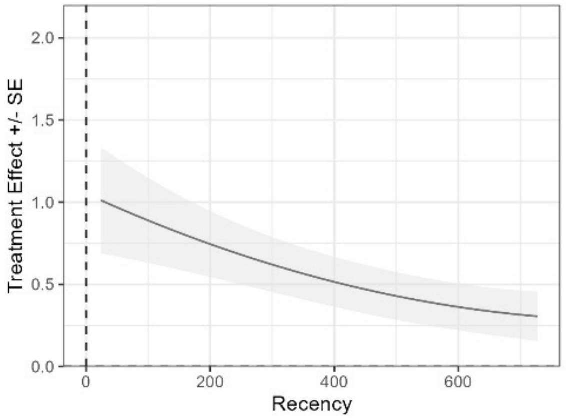
- The 42% lift becomes three times more when focusing on the most responsive quintile.

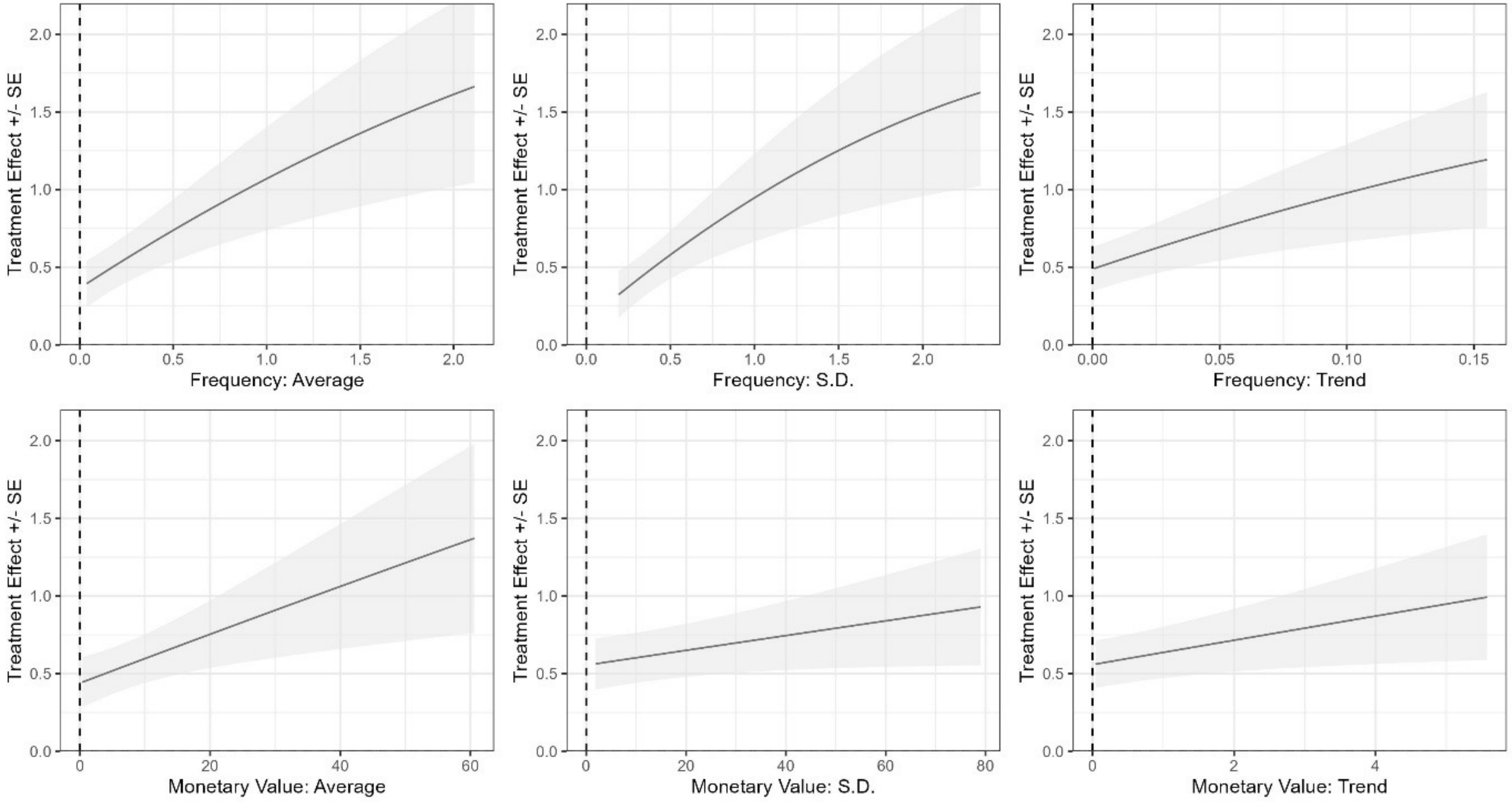
- About 20% of the donors contribute to 80% of the effect (Pareto law)

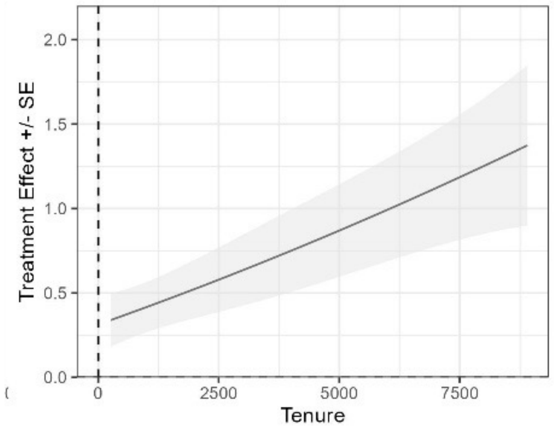


Which donors are most sensitive to agency?

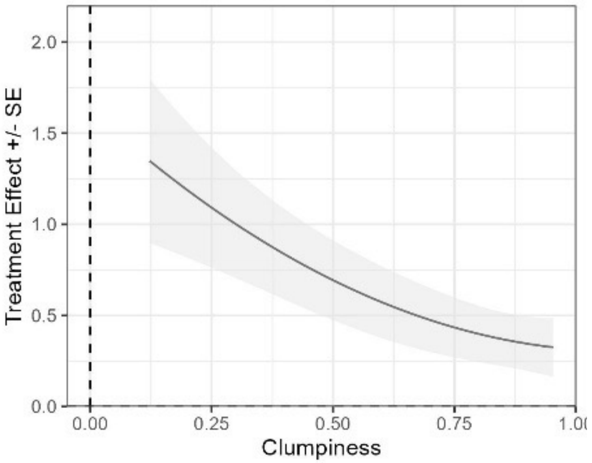
More responsive donors donated more money, more recently, and relatively more often and are more loyal (tenure)





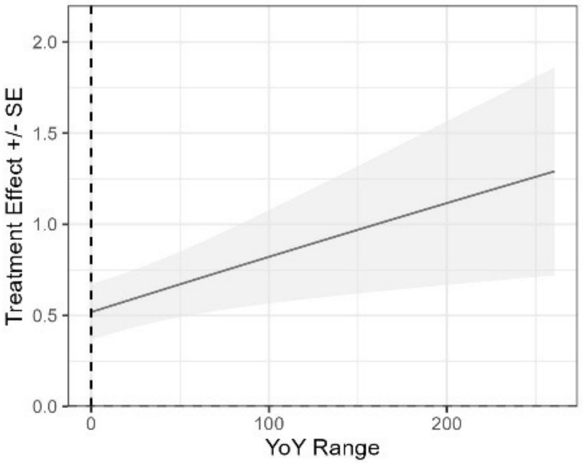


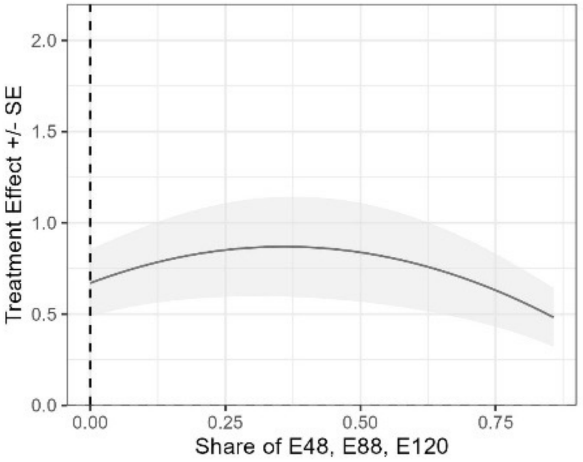
Which donors are most sensitive to agency?

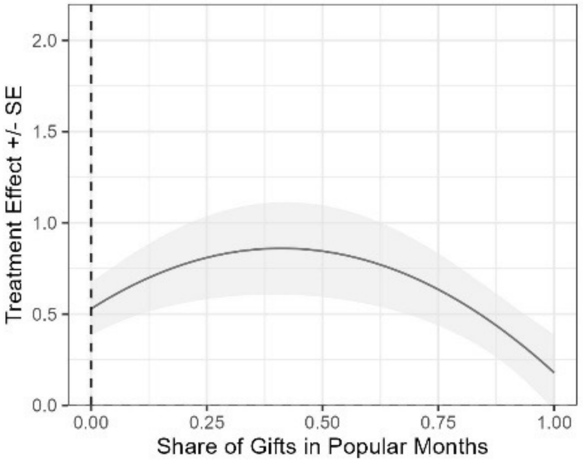


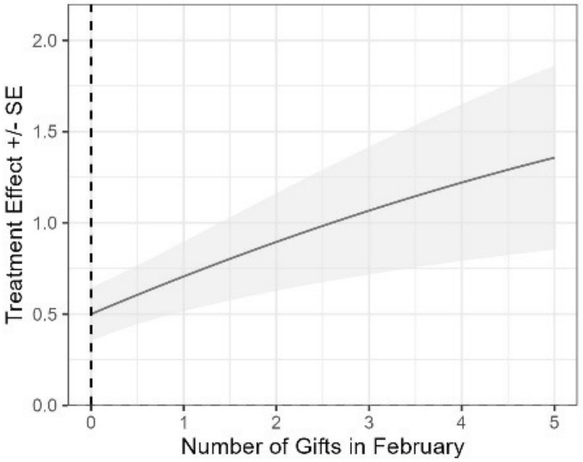
More responsive donors

- Show relatively less “clumpy” donation patterns;
- Show less “habitual” donation patterns.







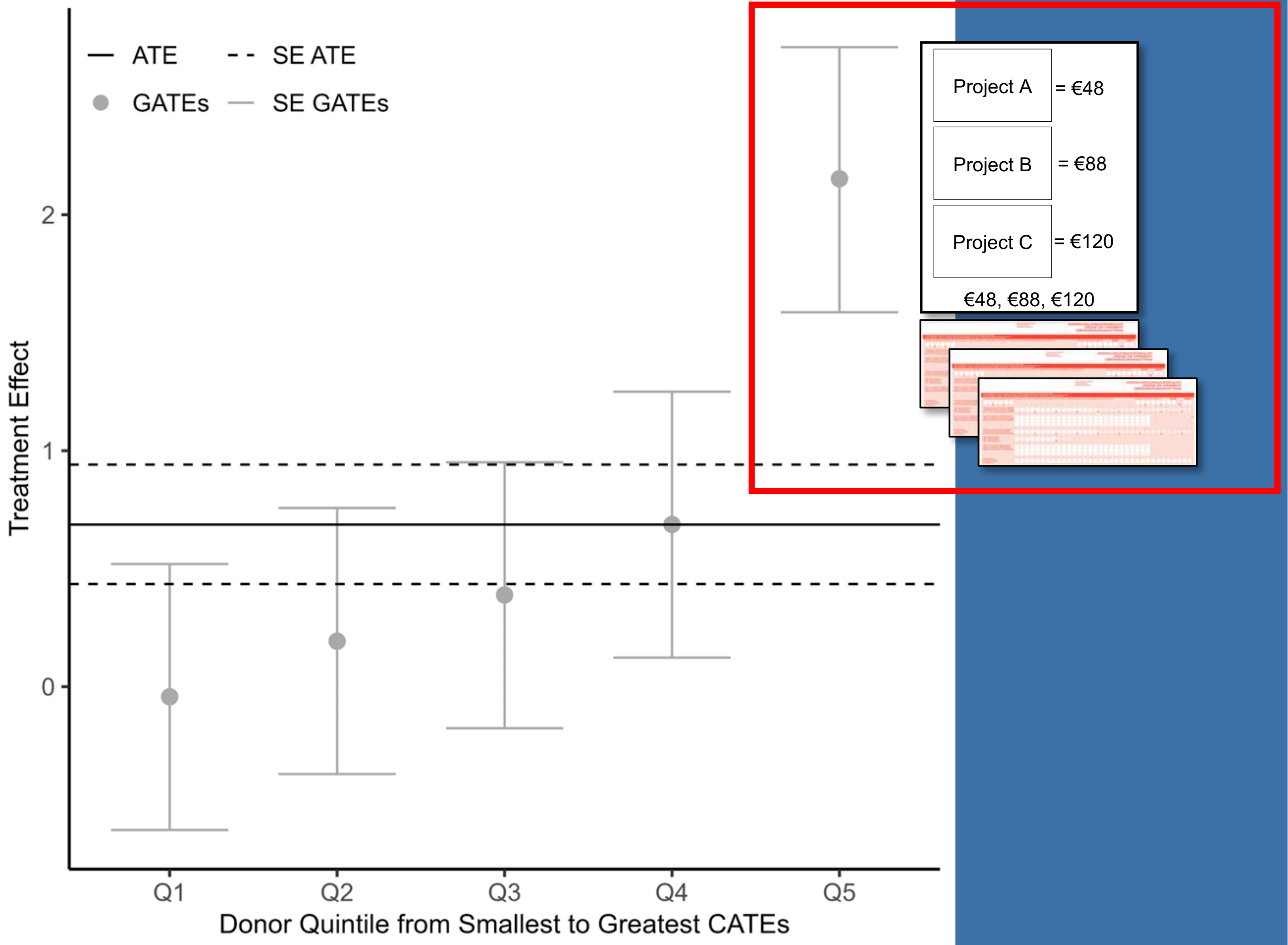


Boost the Managerial Impact of Your Study

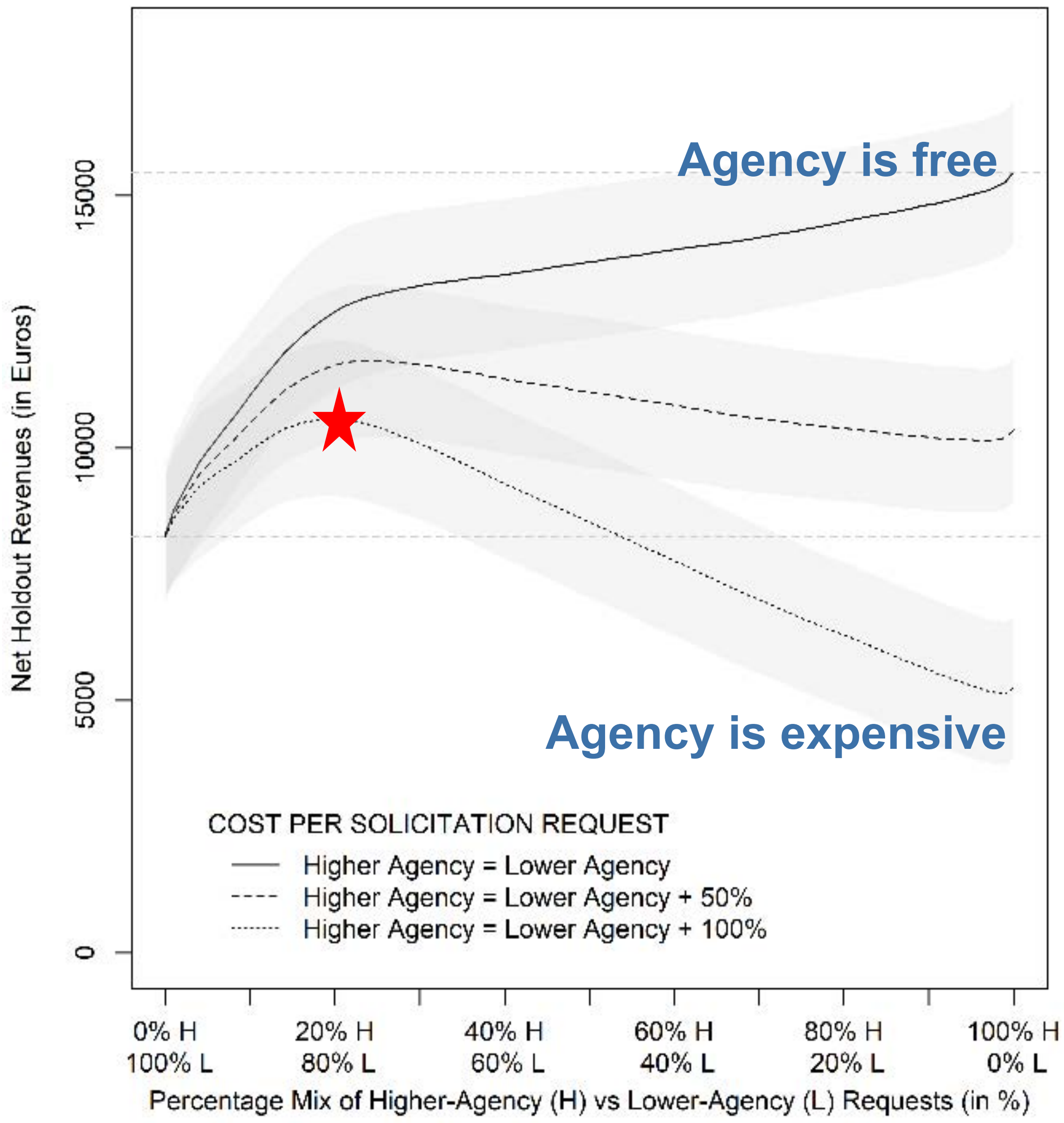
CATE → Personalized policies

Designing a Personalized Policy

Agency is not costless!



Can we give agency to a selected set of donors only and still leverage its benefits?

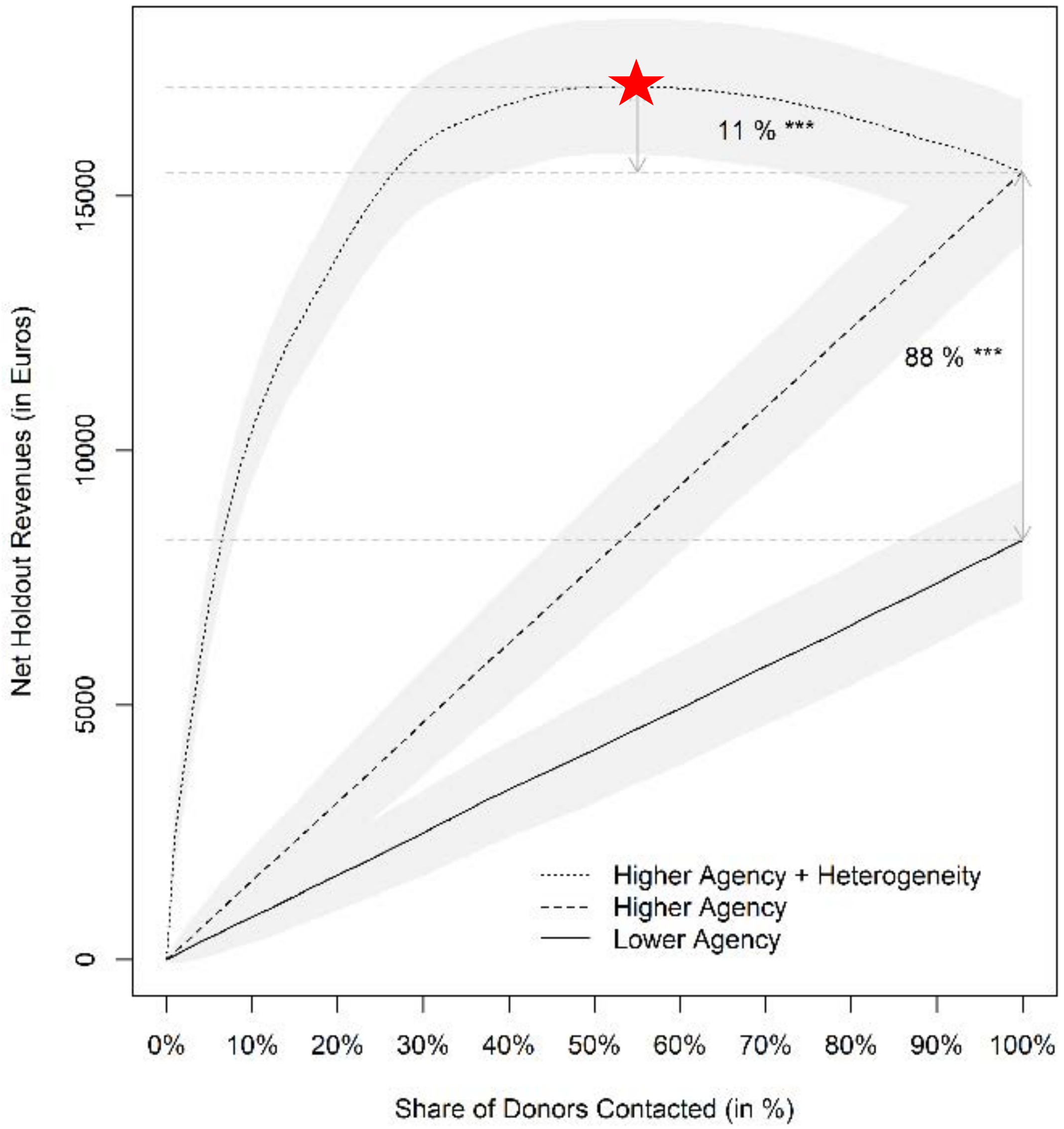


Offering Agency to 20% of the Donors is Most Beneficial

The rest should receive a low-agency request

We can optimize the size of the treatment group as a function of the treatment cost

Should We Let Some Donors Sleep?



- Lower agency policy (only sending low agency requests) **= €8,234**
- Higher agency policy (only sending high agency requests) **= €15,466**
- Higher agency + heterogeneity policy (only sending high agency requests to donors who are most responsive to agency) **= €17,141**

Reach Out!

- **Causal Machine Learning can enrich your theories AND boost your managerial impact**
  - Exploratory research into new moderators and boundary conditions
  - Personalized policy design (e.g., personalized medicine)

- **Empowering *SOME* donors offers "cheap" yet effective opportunities to increase fundraising revenues**
  - Generalization to domains outside nonprofit (empowering customers, patients, etc.).

- **Try it out!**
  - If you are interested in unveiling heterogeneity in intervention effectiveness, check our OSF repository to estimate, analyze and optimize AB data : https://osf.io/4nzsw/


LIVE
COLOR
FULLY


THANK YOU

johannes-boegershausen
boegershausen.net
web-scraping.org
@JoBoegershausen
boegershausen@rsm.nl

aurélie-lemmens-rsm
aurelielemmens.com
https://osf.io/4nzsw/
https://github.com/AurelieLemmensRSM/
lemmens@rsm.nl