Check for updates

AMERICAN MARKETING ASSOCIATION

Article

Journal of Marketing
2022, Vol. 86(5) 1-20
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00222429221100750
journals.sagepub.com/home/jmx
SAGE

Fields of Gold: Scraping Web Data for Marketing Insights

Johannes Boegershausen, Hannes Datta, Abhishek Borah, and Andrew T. Stephen

Abstract

Marketing scholars increasingly use web scraping and application programming interfaces (APIs) to collect data from the internet. Yet, despite the widespread use of such web data, the idiosyncratic and sometimes insidious challenges in its collection have received limited attention. How can researchers ensure that the data sets generated via web scraping and APIs are valid? While existing resources emphasize technical details of extracting web data, the authors propose a novel methodological framework focused on enhancing its validity. In particular, the framework highlights how addressing validity concerns requires the joint consideration of idiosyncratic technical and legal/ethical questions along the three stages of collecting web data: selecting data sources, designing the data collection, and extracting the data. The authors further review more than 300 articles using web data published in the top five marketing journals and offer a typology of how web data have advanced marketing thought. The article concludes with directions for future research to identify promising web data sources and embrace novel approaches for using web data to capture and describe evolving marketplace realities.

Keywords

web scraping, application programming interface, API, crawling, validity, user-generated content, social media, big data

Online supplements: https://doi.org/10.1177/00222429221100750

The accelerating digitization of social and commercial life has created an unprecedented number of digital traces of consumer and firm behavior. Every minute, users worldwide conduct 5.7 million searches on Google, make 6 million commercial transactions, and share 65,000 photos on Instagram (Statista 2021). The resulting web data—enormous in size, diverse in form, and often publicly accessible on the internet—is a potential goldmine for marketing scholars who want to quantify consumption, gain insights on firm behavior, and track social activities difficult or costly to observe otherwise. The importance of web data for marketing research is reflected in a growing number of impactful publications across all methodological traditions, including consumer culture theory, consumer psychology, empirical modeling, and marketing strategy.

Researchers can use web scraping and application programming interfaces (APIs) to efficiently collect web data at scale. Web scraping is the process of developing software to automatically collect information displayed in a web browser. For example, researchers can scrape Amazon's website to construct data sets of online consumer reviews. Because many websites and web apps are publicly accessible, data sets can be generated without involving data providers. In contrast, some data providers also offer APIs for programmatic access to their internal databases. For example, scholars can apply for academic research access to retrieve data from the Twitter API. Researchers can also access a wide range of algorithms via APIs. For instance, Google offers advanced image and video recognition through its Cloud Vision API (for additional examples and explanations, see Table W1 in Web Appendix A).

Johannes Boegershausen is Assistant Professor of Marketing, Rotterdam School of Management, Erasmus University, The Netherlands (email: boegershausen@rsm.nl). Hannes Datta is Associate Professor of Marketing, Tilburg University, The Netherlands (email: h.datta@tilburguniversity.edu). Abhishek Borah is Assistant Professor of Marketing, INSEAD, France (email: abhishek.borah@insead.edu). Andrew T. Stephen is L'Oréal Professor of Marketing, Associate Dean of Research, Saïd Business School, University of Oxford, UK (email: andrew.stephen@sbs.ox.ac.uk).

Data extracted from the internet, at first sight, might resemble other organically generated data sets that address related research questions (e.g., a firm's clickstream data). Yet, collecting web data for academic use in a highly automated manner may prompt a set of novel and sometimes insidious validity challenges. Validity concerns may arise from, among others, (1) failing to capture contextual information in a rapidly changing environment (e.g., updates to the website's data-generating process), (2) not sufficiently aligning the psychological processes of interest with the frequency of data extraction on review platforms (e.g., the collected information does not capture the time when the behavior occurred), (3) overlooking the influence of algorithmic interference on e-commerce websites (e.g., the effect of personalization algorithms on information display), or (4) failing to retain raw website or API data necessary for construct validation, sampling, and analysis.

Against this background, this article makes three interlinked contributions. First, we develop a methodological framework that highlights how addressing validity concerns arising from web scraping and APIs requires the joint consideration of idiosyncratic technical and legal/ethical concerns. Within marketing, guidance exists for collecting web data in the consumer culture theory research tradition, particularly using netnography (e.g., Kozinets 2002, 2020). A handful of articles address selected challenges that occur during the automatic extraction of web data (e.g., sampling; Humphreys and Wang 2018). Outside of marketing, tutorials and books primarily focus on technical details for the automatic extraction of web data (see Table W2 in Web Appendix B). Yet, neither these resources nor methodological articles in other disciplines (e.g., Edelman 2012; Landers et al. 2016) address the broad spectrum of validity concerns arising from the automatic collection of web data for academic use. It is this void that our methodological framework fills. In discussing the methodological framework, we offer a stylized marketing example for illustration and provide recommendations for addressing challenges researchers encounter during the collection of web data via web scraping and APIs.

Second, despite the use of web data in marketing for two decades, no systematic review reflects on how it has and could advance marketing thought. Importantly, understanding the richness and versatility of web data is invaluable for scholars curious about integrating it into their research programs. To offer these insights, we have systematically reviewed more than 300 articles in the top five marketing journals across two decades that have used web data. We leverage our coding to reveal which web sources have been considered and how data have been extracted. The resulting typology of web data may spark the imagination of researchers interested in generating new marketing insights from web data.

Finally, we use our methodological framework and typology to unearth new and underexploited "fields of gold" associated with web data. We seek to demystify the use of web scraping and APIs and thereby facilitate broader adoption of web data across the marketing discipline. Our future research section highlights novel and creative avenues of using web data that include exploring underutilized sources, creating rich multisource data sets, and fully exploiting the potential of APIs beyond data extraction. We particularly highlight the value of web scraping and APIs for research streams that have not yet embraced them at scale.

In what follows, we provide an overview of the use of web data in marketing and document four pathways through which web data have advanced marketing thought. We then introduce our methodological framework to help researchers make sensible design decisions when automatically extracting web data. We conclude with directions for future research.

Using Web Data to Advance Marketing Thought

Across the top five marketing journals, marketing researchers increasingly use information available on the internet. For example, the share of web data–based publications has more than tripled in the last decade, from about 4% in 2010 to 15% in 2020 (see the thick line in Figure 1). The growing use of web data has been fueled by its increased accessibility and associated time and cost savings. Most of the 313 identified web data–based articles rely on web scraping (59%); APIs are used much more sparingly (12%), and some articles combine web scraping and APIs (9%). The remaining articles—especially netnographic work—use web data but tend to extract it manually (20%). The median annual citation count of articles using web data is 7.55, compared with 3.90 for publications not using web data.

Some of the earliest uses of web data in marketing can be attributed to the development of netnography to study online communities (e.g., Kozinets 2002, 2001). Subsequently, the first quantitative marketing scholars extracted web data at scale (e.g., Godes and Mayzlin 2004). Today, all subfields—including marketing strategy and consumer behavior—have embraced web data.

Online word of mouth and social media are the most prominent domains of inquiry using web scraping (see Table W3 in Web Appendix C). The most widely used data source in academic marketing research is Amazon (38 articles). Other prevalent sources are Twitter (30), IMDb (24), Facebook, and Google Trends (both 22; see Table W4 in Web Appendix C).

Via a comprehensive literature review, we next identify the four central pathways through which web data facilitate the creation of new knowledge in marketing.

Studying New Phenomena

Web data can boost the field's relevance by enabling marketing scholars to study novel phenomena. For example, initial work using web data focused on novel online phenomena that emerged at the beginning of this century, such as online conversations (Godes and Mayzlin 2004) and the impact of consumer reviews on sales (Chevalier and Mayzlin 2006). Web data are well suited to provide fertile grounds for inductive research to develop novel theories about emerging marketing phenomena (e.g., brand public; Arvidsson and Caliandro 2016).

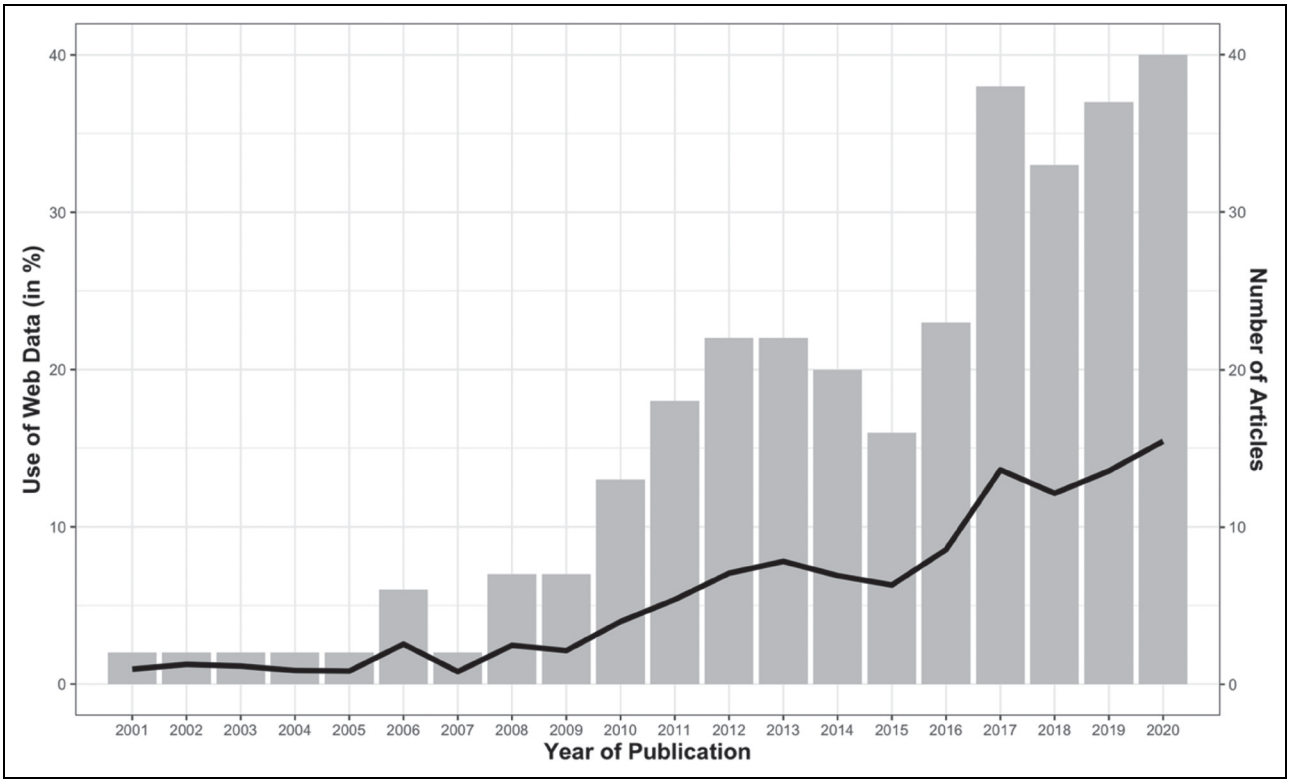

Figure 1. Increased use of web data in marketing (2001–2020).

Gathering data via web scraping or APIs often decreases the time between the occurrence of a marketplace phenomenon and the availability of data for academic research. This inherent timeliness of web data continues to be an essential lever for marketing scholars to advance our understanding of emerging substantive topics such as the sharing economy (e.g., Airbnb; Zervas, Proserpio, and Byers 2017), access-based business models (e.g., Spotify; Datta, Knox, and Bronnenberg 2018), and fake online content (e.g., Anderson and Simester 2014). More generally, web data enable researchers to weigh in on contemporary issues before any "conventional" data sets become available, such as measuring the effect of pandemic lockdown policies on consumption (Sim et al. 2022).

Boosting Ecological Value

Web data can create knowledge by allowing researchers to move closer to marketing's "natural habitat" (Van Heerde et al. 2021). Some of the most used web sources contain commercial outcome variables relevant to marketing stakeholders and are difficult or costly to collect otherwise. Examples are sales (e.g., The-Numbers.com), sales ranks (e.g., Amazon), online searches (e.g., Google Trends), and donations (e.g., contributions to a Kiva project).

As web data can be collected unobtrusively, they can effectively complement more controlled data collection methods. Using web data, researchers can demonstrate that focal psychological processes occur outside the confines of a controlled laboratory environment and stylized experimental stimuli (Morales, Amir, and Lee 2017). Consider, for instance, the controversy around the decoy effect (Huber, Payne, and Puto 1982)—one of the most prominent context effects in consumer behavior. Using experiments, Frederick, Lee, and Baskin (2014) questioned the robustness and practical relevance of the decoy effect. In response, Wu and Cosguner (2020) built a panel data set from an online diamond market using web data. Their work not only shows that the decoy effect emerges in a high-stakes setting but also, more importantly, reaffirms its practical significance by quantifying its profit implications for the diamond retailer.

Another benefit of using web data to boost ecological value is that they can often be collected without the data provider's direct involvement. Thereby, researchers can limit the interference of data suppliers or collaborating firms to ensure that the societal relevance of a particular research question is given precedence over business objectives (e.g., firms might be unwilling to share data about the tracking tools they use on websites; Trusov, Ma, and Jamal 2016). Further, using web data, researchers can ensure the publication of research findings, regardless of how palatable they are to the organizations that are being studied.

Facilitating Methodological Advancement

As much of the data produced by consumers and firms is inherently unstructured, extracting insights can be challenging (Wedel and Kannan 2016). Thus, marketing researchers have leveraged web data for developing methods that deal with and extract insights from different types of unstructured data, such as textual, image,

Table 1. How to Create Knowledge Using Web Data: A Typology.

	Primary Pathways of Knowledge Creation Using Web Data			
Effect on …	**Pathway 1: Studying New Phenomena**	**Pathway 2: Boosting Ecological Value**	**Pathway 3: Facilitating Methodological Advancement**	**Pathway 4: Improving Measurement**
Consumers (e.g., social media use, consumer learning)	Toubia and Stephen (2013) test the motivations of users to contribute content to social media.	Sridhar and Srinivasan (2012) explore peer effects in evaluating online product reviews.	Huang (2019) studies how picture quality improves due to consumer learning.	Huang et al. (2016) exploit within-user variation to measure how psychological distances interact.
Organizations (e.g., sales and profits of firms, donations to nonprofits)	Chevalier and Mayzlin (2006) demonstrate the impact of online reviews on book sales.	Wu and Cosguner (2020) probe the prevalence and profit implications of decoy effects.	Netzer et al. (2012) mine user-generated content to identify market structures.	Datta et al. (2022) gather national holidays across 14 countries and 11 years to capture seasonality.
Other marketing stakeholders (e.g., market reaction of investors, public health outcomes)	Hermosilla, Gutiérrez-Navratil, and Prieto-Rodríguez (2018) examine how consumers' aesthetic preferences create biases in firms' hiring decisions.	Blaseg, Schulze, and Skiera (2020) examine whether consumers are protected against false price advertising claims on Kickstarter.	Tirunillai and Tellis (2012) develop novel online metrics based on user-generated content to predict stock returns.	Kim and KC (2020) explore the effect of ads for erectile dysfunction drugs on birth rates.

Notes: The table highlights illustrative and diverse examples of web data–based studies and corresponding outcome variables, cross-tabulated by four pathways through which web data have advanced marketing thought (the columns) and three of the most studied actors in marketing research (the rows).

and video data. For instance, web data have fueled the rapid improvement of automated text analysis (see Berger et al. 2020) and the large-scale analysis of image and video content (e.g., Li, Shi, and Wang 2019; Liu, Dzyabura, and Mizik 2020).

The availability of network data on the internet (e.g., friend or product networks), along with outcome variables (e.g., posts, likes, sales ranks), has further enabled the use and advancement of methods for analyzing networks (e.g., Oestreicher-Singer et al. 2013). Given their wealth and richness, web data have also stimulated the development of novel methods that can complement or replace traditional marketing research methods (e.g., using user-generated content to construct accurate multidimensional scaling maps of brands; Netzer et al. 2012).

Improving Measurement

Web data can advance marketing knowledge by allowing researchers to measure constructs more precisely and obtain more valid inferences. For example, the collection of adequate control variables is often difficult. To capture seasonality in purchase patterns across a wide range of geographical markets and calendar years, researchers have used APIs to construct continuous (vs. dichotomous) variables that accurately reflect national holidays (e.g., HolidayAPI; Datta et al. 2022). Web data also allow researchers to efficiently operationalize new measures at scale, such as weather conditions based on the location of users' IP addresses (e.g., Weather Underground; Li et al. 2017).

Relative to non-web data sources, researchers can collect data on the behavior of many consumers and firms at higher frequencies (Adjerid and Kelley 2018). Such data enhance statistical power, enable identification of causal effects, and facilitate the examination of theoretically relevant variation within individuals over time (e.g., how various psychological distances shape review content for the same consumer; Huang et al. 2016) or how effects unfold over time (e.g., the impact of video elements on virality over time; Tellis et al. 2019).

Summary

Table 1 presents a typology of the four central pathways through which web data have advanced marketing thought. The typology highlights web data-based studies that investigate key marketing constructs across different entities, from consumers to organizations and other marketing stakeholders. For example, Toubia and Stephen (2013) explored a new phenomenon (tweeting), focusing on consumers (i.e., their motivation to tweet). These pathways for knowledge creation from web data are not mutually exclusive. Combining different pathways might be particularly promising for making breakthrough contributions.

Next, we introduce our methodological framework, which outlines an approach for making design decisions that enhance the validity of web data collected via web scraping and APIs. Researchers interested in learning more about the technical details of automatically extracting web data can consult our curation of technical tutorials in Web Appendix B or the digital companion to this article (available at https://web-scraping.org), which features a searchable database of all marketing articles in the top five marketing journals using web data.

Methodological Framework for Collecting Web Data

In automatically collecting web data using web scraping and APIs, researchers make seemingly innocuous design

decisions. However, as we will show, these decisions often involve trade-offs about research validity, technical feasibility, and legal/ethical risks[1] that are not always apparent. How researchers resolve these trade-offs shapes the credibility of research findings by enhancing or undermining statistical conclusion validity, internal validity, construct validity, and external validity (Shadish, Cook, and Campbell 2002).

We develop a methodological framework to provide guidance for the automatic collection of web data using web scraping and APIs. Figure 2 offers a stylized view of this process involving three key stages—source selection, collection design, and data extraction. Researchers typically start with a broad set of potential data sources and eliminate some of them as a function of three key considerations—validity, technical feasibility, and legal/ethical risks. These three considerations appear in the corners of an inverted pyramid, with validity at the bottom to underscore its importance. Given the difficulty in projecting the exact characteristics of the final data set before it is collected, researchers often revisit these considerations as they design, prototype, and refine their data collection. Failure to resolve technical or legal/ethical issues might mean that web data cannot inform the research question meaningfully.

Our framework deliberately focuses on *collecting* web data rather than its subsequent analysis. Analyzing web data involves many familiar methodological challenges encountered with organically generated data (e.g., cleaning to remove erroneous data or create measures, selecting observations, addressing endogeneity). However, approaches for the valid collection of web data are not yet documented nor commonplace in marketing research.

The methodological framework—designed to guide the automatic extraction of web data at scale—is agnostic to research paradigms. It is applicable to both deductive (i.e., identifying compelling web data to test hypotheses) and inductive (i.e., observing interesting irregularities in web data to identify novel marketing concepts and/or novel relationships between constructs) approaches to theory building. We next highlight the idiosyncratic challenges encountered when collecting web data and summarize solutions to these challenges in Tables 2–4. For expositional clarity, we focus on web scraping in our text.[2] To illustrate the key challenges encountered in designing the data collection, we gradually introduce a stylized marketing example involving the collection of book reviews from Amazon.

Data Source Selection

A critical first step in the use of web data is selecting the data source(s). We examine three challenges faced by researchers in this selection process. First, it is essential that researchers explore the universe of potential sources (challenge #1.1). Second, researchers need to consider the range of possible extraction methods (challenge #1.2). Third, it is crucial to map the context in which the data are generated (challenge #1.3). Table 2 summarizes our recommendations for tackling these challenges.

Exploring the Universe of Potential Sources (Challenge #1.1)

In the absence of conventional gatekeepers (e.g., data providers), researchers can select from countless web data sources. For example, there are 2.1 million online retailers in the United States alone (etailinsights 2021). Further, websites and APIs differ greatly in scope (e.g., number of users), data quality (e.g., consistency), and retrievability (e.g., extraction limits). Even within the same product category, data sources differ vastly. For example, Amazon reports a book's sales rank (an aggregate outcome metric for product sales), whereas Goodreads reports users' reading behavior (an individual outcome metric for consumers' usage intensity).

Faced with a vast universe of potential sources, researchers may be tempted to focus on familiar platforms only (Wells 2001). For instance, Amazon is the most used web data source in marketing (see Table W4 in Web Appendix C). Amazon might be a relevant source to extract book reviews, given its broad assortment and user base. Yet, in other cases, researchers might miss opportunities for identifying novel, emerging marketing phenomena or conduct more compelling theory testing without a thorough exploration of potential sources. Researchers can avoid the pitfalls of defaulting to dominant sources by actively considering a broad spectrum of websites and APIs, ranging from highly popular (e.g., Amazon) to less popular sources (e.g., Goodreads), from primary data providers (e.g., YouTube) to data aggregators (e.g., Social Blade), and from platforms with global reach (e.g., Twitter) to more regional ones (e.g., Taringa!). Another strategy to move beyond familiar sources is to adopt alternative perspectives. For instance, researchers can consider websites or APIs that are used by consumers, analysts, or managers. API directories at GitHub or programmableweb.com can facilitate identifying potentially relevant APIs.

A broad exploration of web data sources may lead researchers to discover sources that may be more permissive for (academic) data extraction or less likely to trigger ethical concerns. For example, websites that do not require logging into the site to reveal information are typically more scraping-friendly than sites that first require registering a user account. In the case of Amazon, researchers can obtain most information without logging in and do not have to explicitly provide their agreement to the website's Terms of Service. To reduce legal (e.g., breaches of contract, as researchers have provided explicit agreements to the terms of service) and ethical (e.g., website users may consider their data private) risks, researchers should refrain from creating fake

[1] A comprehensive discussion of the legality and ethics of the automatic collection of web data is beyond the scope of this article. For a discussion of key issues, see Landers et al. (2016) and Vanden Broucke and Baesens (2018, pp. 181–86). As legal experts are unlikely to be familiar with the various decisions that researchers make during data extraction, Web Appendix D provides a list of concerns to consider when seeking legal counsel.

[2] Our discussion largely extends to APIs, unless noted otherwise. For example, with APIs, researchers use "endpoints" instead of "pages" (see Web Appendix A).

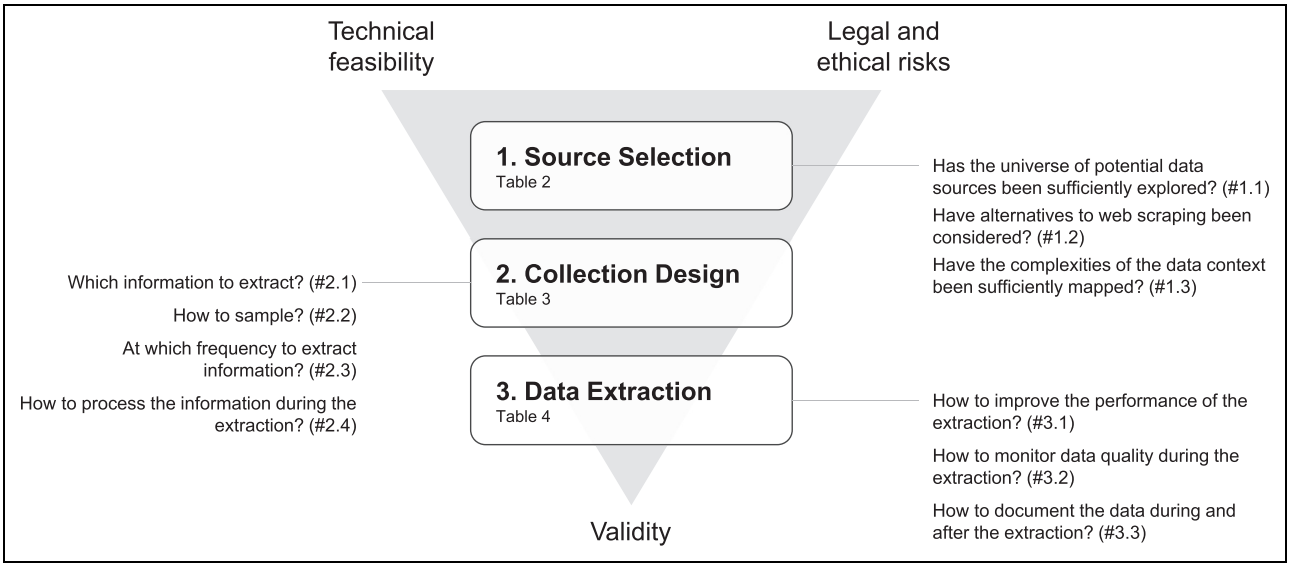

Figure 2. Methodological framework for collecting web data.

accounts to access information requiring a login. By explicitly declaring their academic status (e.g., when registering at the site using the institutional email address), researchers might be able to diminish their exposure to legal risk.

When exploring web sources, researchers need to examine whether theoretical constructs can be operationalized in a valid manner (Xu, Zhang, and Zhou 2020). A healthy level of skepticism is warranted when using idiosyncratic metrics from APIs or websites. For example, researchers might be interested in scraping the price tier of restaurants from Yelp. Yet, it is not entirely clear how Yelp computes this metric from individual consumer ratings.

To determine when to stop exploring sources, researchers need to assess to what extent the selected source(s) improve(s) on alternatives. One way to justify selecting a single web source is the presence of unique features. For example, a researcher studying how observers react to humor in reviews might prefer Yelp to alternative platforms as it is the only source featuring "funny" votes (e.g., McGraw, Warren, and Kan 2015). At other times, researchers may be indifferent between potential sources and can draw from multiple sources to boost the generalizability of their findings (e.g., tweets and restaurant reviews; Melumad and Meyer 2020). Collecting data from multiple sources is often useful because even similar types of information (e.g., consumer comments) may affect marketing outcomes differently, depending on source characteristics (e.g., forums vs. microblogs; Schweidel and Moe 2014). Data aggregators—some of which offer authorized data access via APIs—facilitate the collection of such multisource data.

Considering Alternatives to Web Scraping (Challenge #1.2)

The popularity of web scraping may lead to the conclusion that it should be preferred over other methods. However, some web sources offer access to data via APIs (Chen and Wojcik 2016). In general, extracting data via APIs is more scalable and less likely to invoke the same level of legal risks compared with web scraping. Although some sources offer unconstrained APIs that do not require authentication, others require (paid) subscriptions and authentication procedures. Some sources, such as Twitter, have recently started offering APIs for academic research. In the case of Amazon, an API offering access to consumer reviews is currently not available.

In addition to APIs, numerous other options exist for researchers to obtain web data. For example, some data providers (e.g., Yelp, IMDb), public data platforms (e.g., Kaggle, The Dataverse Project), and researchers (e.g., McAuley 2021) provide documented web data sets that can readily be used for academic research. There are many potential use cases of such data sets, but less than 5% of all web data–based articles in marketing used such data sets. To avoid the pitfall of defaulting to web scraping for data extraction, researchers can expand their search by explicitly including terms such as "API" or "data set download" in their search queries.

Mapping the Data Context (Challenge #1.3)

Relative to other frequently used archival sources in marketing, web data entail large and often undocumented complexities. Thus, it is critical that researchers map the data context, which involves identifying relevant contextual developments that may undermine the validity of the research if gone unnoticed.

First, mapping the data context may reveal changes in the underlying data structure. For example, a major change in a platform's user interface may affect subsequent consumer behavior. Second, mapping the data context enables researchers to identify relevant pieces of information for collection together with the focal web data. For example, researchers may discover an external website (e.g., Statista)

Table 2. Challenges and Solutions in Selecting Web Data Sources.

Challenge #1.1: Exploring the Universe of Potential Web Sources	
Reason for importance	As web sources vastly differ in quality, stability, and retrievability, researchers might be tempted to consider dominant or familiar platforms only. A thorough exploration of the data universe permits compelling theory testing and identifying novel, emerging marketing phenomena that may be difficult to notice otherwise.
Solutions and best practices	• Assume the perspective of different stakeholders (e.g., consumers, analysts, managers) during the search process • Browse through public API directories (e.g., ProgrammableWeb, GitHub) • Broaden geographic search criteria (e.g., non-Western) • Identify adjacent data sources (e.g., using Google Trend's "related search queries") • Expand search to nonprimary data providers (i.e., aggregators, databases) • Carefully vet the provider's description of relevant metrics • Determine the conditions necessary to access data (e.g., requirement to log in on a website, creating an API key, subscribing to an API, possibility to signal academic status/scientific use) • Verify whether it is possible to opt out of firm-administered experiments or whether the site is accessible without cookies • Use the website or make some initial API requests to assess information availability (in the case of APIs, assess which authentication procedure is necessary to obtain data)
Challenge #1.2: Considering Alternatives to Web Scraping	
Reason for importance	Because web scraping is the most popular extraction method for web data, researchers may overlook alternative ways to extract data. APIs provide a documented and authorized way to obtain web data for many sources. Some sources also provide readily available data sets. Using such alternatives leads to time savings and minimizes exposure to legal risk.
Solutions and best practices	• Expand search by explicitly including terms such as "API" or "data set download" • Explore whether the source or third parties (e.g., public data platforms, researchers) offer data sets for download and assess their terms of use • If a data source provides an API and a website, understand the differences in what data could be retrieved from them (e.g., by screening the API documentation) and how well the API can be accessed (e.g., using packages) • Use stable versions of an API, and subscribe to a source's API support updates
Challenge #1.3: Mapping the Data Context	
Reason for importance	Web data are usually not accompanied by extensive documentation. Identifying potentially relevant contextual information early on is essential for the relevance and validity of the research.
Solutions and best practices	• Screen blogs, press releases, a source's software "changelogs," or use Google's reverse search to identify important (technical) developments • Build an initial understanding of the presence of algorithms by visiting the source with different devices at different times or by inspecting the site's source code • Understand changes to the data-generating process (e.g., by studying changes over time using archive.org) • Inspect the robots.txt file and assess how the source requires users to agree to their terms of service (e.g., preferrable "browsewrap" vs. less preferable "clickwrap" agreements) • Scrutinize popularity, legitimacy, and business model of data sources (e.g., by using firm reports, stock filings, news, and social media, other data providers like Statista) • Explore forums where users of the source talk about the source (e.g., Reddit) • Assess whether the data links to other data sets (e.g., by spotting common IDs) • Map out "worst-case" scenarios for research objectives in the case that the data source changes (e.g., discontinuation of an API, removal of a website)

that offers information about the composition of a focal data source's user base. If stored, such data could eventually be used to detect changes in the composition of the user base or verify the representativeness of the extracted data. Third, mapping the data context may reveal unknown information, potentially allowing researchers to discover novel research opportunities. For example, researchers may use the (unexpected) launch of a new recommendation system at a music streaming service as a natural experiment to investigate the impact of recommendations on music consumption.

To understand and map the contextual complexity of web data, researchers can immerse themselves in the ecosystem surrounding the focal source by signing up and using the source, tracking press releases, social media, and scanning the competitive environment. Helpful tools include a search engine's advanced search features, newsletters, and alerts on leading business and technology magazines (e.g., TechCrunch.com, WSJ.com, FT.com). The website's source code may also hold valuable information about potentially relevant environmental changes. Sometimes, researchers may also detect the presence

Table 3. Challenges and Solutions in Designing Web Data Collections.

Challenge #2.1: Which Information to Extract from Which Pages?

Validity Challenges [V]
- 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?

Legal/Ethical Challenges [L]
- 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?

Technical Challenges [T]
- Is all information extractable?
- Are there any limits to iterating through pages or endpoints?
- Does the extraction software obtain information reliably?

Solutions and Best Practices
- 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 into 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]
- 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?

Legal/Ethical Challenges [L]
- 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?

Technical Challenges [T]
- Is the required sample size technically feasible?
- Can external information (e.g., IDs) be consistently matched to the data?

Solutions and Best Practices
- 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 that no excessive amount of data (e.g., data on all users) is collected (absolute volume, relative volume) [L]
- Reexamine alternative sources to improve justification of data extraction [L]

(continued)

Table 3. (continued)

Challenge #2.3: At Which Frequency to Extract the Data?

Validity Challenges [V]
- 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?

Legal/Ethical Challenges [L]
- Does the extraction frequency pose an excessive load on the source?
- Does collecting more data at higher frequencies make the data more sensitive?

Technical Challenges [T]
- Does the desired extraction frequency pose new technical hurdles?
- How can the stability of data collection be guaranteed, and different collection batches be distinguished?

Solutions and Best Practices
- 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., five requests per second for APIs, one request per two 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]
- Could erroneous processing lead to unexpected data loss?
- Could there be any significant scientific value in retaining the raw data?

Legal/Ethical Challenges [L]
- 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?

Technical Challenges [T]
- Which storage facilities to use to accommodate the expected data (size, location, format, encoding)
- Is normalization necessary?

Solutions and Best Practices
- 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

of algorithms on the site that may threaten the validity of the collected data. For example, Amazon's product pages personalize information based on which preceding products were viewed—even without users logging into the site.

Designing the Data Collection

After narrowing down potential sources, researchers decide which information to extract from them (challenge #2.1), how to sample (challenge #2.2), at which frequency to extract the information (challenge #2.3), and how to process the information during the collection (challenge #2.4). Table 3 summarizes these challenges and corresponding solutions.

Which Information to Extract? (Challenge #2.1)

In the absence of any "downloadable" data set, the first challenge lies in deciding which information to extract from a source. Researchers begin by browsing the web page to identify from which pages to extract which information. In our Amazon example, some of the most commonly used pages are product pages (e.g., Chevalier and Mayzlin 2006) and review pages (e.g., Villarroel Ordenes et al. 2017). Generally, pages such as those from e-commerce platforms contain information from the company's database, offering researchers the opportunity to capture some of the information available at a company. Collecting such data involves *iterating* through a set of related pages (i.e., browsing through many product pages and corresponding review pages in our Amazon example) and saving the data as it becomes visible.

As the goal of websites like Amazon is rarely the provision of data sets for academic research, it is often necessary to combine information from different pages (e.g., book descriptions from the product page and ratings from the review page). It is particularly difficult to recognize the subtleties of available information, which makes the decision from which pages to extract information challenging. For

example, researchers interested in building a data set of book reviews would find total ratings both on the product and review page, but only the review page reveals *all* product reviews.[3] Yet, neither the product nor the review pages contain all the biographical information available on a reviewer's profile page. Widely exploring a website or API is necessary for identifying information relevant for subsequent analysis (e.g., construct operationalization). The amount and type of information also often vary (e.g., depending on screen width or whether the user is logged in). In this phase, researchers should assess the degree to which the information could be considered personal or sensitive under different regulatory regimes (e.g., the European Union's General Data Protection Regulation [GDPR]), which may require planning measures such as pseudo-anonymization of reviewer names. Researchers may also reassess whether all information needs to be captured to meet the research objective. Suppose reviewer names are strictly necessary (e.g., because they allow for matching different sources). In that case, researchers can explore whether the targeted web data source offers ways to exclude subjects governed by prohibitive privacy regulations (e.g., by using filters).

An important threat to internal validity in any study involving web data is algorithmic interference (e.g., Xu, Zhang, and Zhou 2020). The (visual) design of websites that facilitates usability can undermine the validity of the collected data if gone unnoticed and unaddressed. Especially when deciding which information to extract, it is important to reexamine the website or API for the presence of algorithms. For example, the order in which the researchers in our example visited the website while designing their data extraction could affect which related books are displayed on the product pages on Amazon. Other algorithms that often affect the display of data on websites are sorting algorithms (e.g., by popularity or mixed with sponsored search results) and filtering algorithms (e.g., showing subsets of the data). Algorithmic interference is often hard to detect without being sensitive to it. To account for potential algorithmic interference, the researcher might extract variables as part of an algorithm's more extensive set of input variables, which offers opportunities to control for them in the empirical analysis (e.g., the order in which books were extracted in our Amazon example).

Researchers also need to establish the intertemporal stability of available information. Because the web is constantly evolving, the information on a page might not have been generated via the same process over time, undermining the internal validity of the data. Some changes to sources are drastic enough to alter how the data were created in the first place, introducing measurement error (Weber 2018). For example, Amazon shifted to a positive-only evaluation of reviews by removing the "not helpful" vote button in 2018, and it no longer displays "not helpful" counts next to reviews (Hanna 2018). This change might have impacted review content (e.g., users writing shorter review texts). Yet, researchers collecting data today cannot find any traces of these "not helpful" votes. A tool for examining changes to relevant information on a website is the Wayback Machine (archive.org). Researchers can use this tool to retroactively inspect websites over time (e.g., Martin, Borah, and Palmatier 2017) or submit their own website links for archiving.

Finally, collecting metadata that "annotates" the data collection enhances internal and external validity (e.g., storing the timestamp of data extraction, whether an API request was completed successfully, or the IP address from which the data request was made). Such metadata can be used not only for diagnostic purposes but also to link the extracted web data to other data sets. For example, in our Amazon example, the collected data could be linked to other data using IP-based geolocations (e.g., linking geolocation and web search data; Wang and Chaudhry 2018) or timestamps (e.g., linking reviews to stock prices; Tirunillai and Tellis 2012).

How to Sample? (Challenge #2.2)

A second challenge in designing the data extraction lies in deciding how to sample from the data source. In particular, in the absence of access to the data source's *entire* database, it is difficult or impossible to draw a random sample from the population (e.g., all products) available at the data source. Instead, researchers need to devise their own sampling frame to reveal the units they want to sample from the website (Neuendorf 2017). For example, researchers could scan the site for an index of all products that could inform their sampling. In our example studying reviews at Amazon, multiple such indexes may be available. Should products be sampled from the bestseller page for books (so-called exposure-based populations; Neuendorf 2017) or instead from the category page for books (i.e., availability-based populations)? Choices like this result in different data and may even invalidate inferences, as sampling frames might inadvertently induce systematic bias (Humphreys and Wang 2018).

One common validity challenge in choosing how to sample is determining how many units (e.g., books) are sufficient to inform the research question. From a validity standpoint, it would be ideal to collect information on the entire population (e.g., all books available at Amazon). However, Amazon does not have an obvious page to extract *all* books. Imagine that a research team wanted to collect information about all marketing books sold on Amazon. The bestseller page, for example, lists only the top 100 bestsellers. By manually changing pagination parameters in the URL, the top 400 bestsellers can be revealed. Yet, this list of 400 books neither constitutes the entire population nor represents a random sample of marketing books sold at Amazon. Alternatively, when starting from the product overview pages, these pages list an imprecise number of books (e.g., "over 60,000"), which can only be viewed up to page 50. With each result page featuring 24 organic search results, this approach would produce 1,600 books per category at best. Thus, researchers need to consider other ways to identify more books on Amazon, such as searching for books using various keywords. To expand the number of sampled units, researchers could collect data multiple times, use other keywords, or tweak search parameters to reveal more data by requesting narrower subsets from the database (e.g., only books published during a specific month).

[3] Table W6 in Web Appendix E contains URLs that accompany the Amazon example.

Even if a list of the population (e.g., all books) could be retrieved, it may be infeasible to extract data within a reasonable time frame. While sample size requirements are mostly concerned with a researcher's inferential goals (e.g., Lakens 2022), few articles make the resource constraints that affect collecting web data explicit (e.g., Peng et al. 2018). For example, with web data, a study's sample size critically depends on technical parameters such as the number of computers used for data extraction or the number of pages that need to be visited. We illustrate how to calculate the technically feasible sample size in Web Appendix F, which may effectively complement traditional sample size calculations commonplace in marketing.

As a result of these complications, researchers often restrict their sample size. One way to motivate a compelling sampling frame is to use *external* sources that can be linked to the web data. For instance, the *New York Times* or *Publishers Weekly* bestseller lists might be a starting point for sampling books (Chevalier and Mayzlin 2006). An alternative approach focuses on internal data available at the source itself. Researchers may have to allocate substantial time to identify ways to sample from the focal source. Sometimes, starting the data collection from a page unrelated to the focal pages of interest might facilitate collecting a more representative sample (e.g., by reducing geographical biases; Wang and Chaudhry 2018). For example, on Amazon, researchers could first sample reviewers and associated demographic information (available at the user profile of reviewers) and subsequently retrieve data on all reviewed products. Similar to how researchers build network data from an initial set of products or users, the sampling units retrieved from an initial set of pages can be considered seeds. In choosing seeds, researchers should be cautious about drawing from vulnerable populations (e.g., minors) or infringing on prohibitive privacy regulations.

At Which Frequency to Extract Information? (Challenge #2.3)

Web data are nonstatic, as they change often or might disappear altogether. Therefore, researchers need to consider at which frequency to extract information. This decision encompasses whether to collect data once or multiple times and when to run (and potentially schedule) the data extraction. Consideration of the frequency and schedule is challenging but required to ensure the intertemporal stability of measurement, which is critical for internal and construct validity.

From a technical and legal perspective, it is most desirable to extract data only once. Single extractions are less likely to represent a burden on the firm's servers, and the extracted data often only represent a limited snapshot of the entire database, reducing the risks of copyright infringement. Further, such data may be more likely to respect users' "right to be forgotten," which is part of the privacy laws in some jurisdictions. Yet, single data extraction might raise several validity issues that can easily go unnoticed. For instance, in our example, researchers extracting book reviews *once* from Amazon will not be able to identify whether any of the archival information has changed. Only when extracting data *multiple times* can researchers systematically notice changes on the site, which may lead to the identification of "fake" reviews that have been removed by the platform (e.g., He, Hollenbeck, and Proserpio 2022). More generally, researchers can compare information over time to detect whether data that initially appeared to be archival is truly archival (i.e., does not change over time).

Another concern is that a single extraction may not produce a data set that adequately maps onto the focal processes of interest. For example, suppose researchers in our example want to examine whether a review by a so-called "Top 1000 Reviewer" leads to more subsequent reviews from other users. However, the researcher merely observes that the reviewer is a top reviewer at the time of data extraction. This does not necessarily imply that this user had the same status when the review was first posted and thus was most likely to affect subsequent reviewing behavior of other users. Formulating and testing the essential assumptions about the data, including the relation between the time of data extraction and the focal (psychological) processes, is thus critical. The formulation of such assumptions is called a "data source theory" (Landers et al. 2016). Testing and refining the data source theory helps take proactive steps to enhance internal and construct validity. In the preceding example, it would thus be necessary to collect data from these review pages closer to the original posting date, ensuring that reviewers classified as "Top Reviewers" had that status when their reviews became visible.

When extracting data more than once, automatic scheduling can help ensure consistency and contribute to validity. Scheduling is beneficial if the required information is only available in real-time. For example, sales ranks at Amazon are updated hourly for popular products, and historical sales ranks cannot be retrieved. Suppose researchers in our example were interested in studying the sales performance of books over time. In that case, they could repeatedly extract the books' sales ranks from the product pages at Amazon. Sometimes fixed intervals enhance validity (e.g., every Monday, 8 A.M.). In other circumstances (e.g., when collecting data from many pages), it may be better to vary the starting time or weekday of the data extraction.

Another decision is whether to set an end date for the data extraction. Collecting data over extended periods offers the potential for researchers to build a programmatic stream of research and stumble into unexpected natural experiments (e.g., Chen, Wang, and Xie 2011). Especially for longitudinal data collections, continuing the data collection while the project is in the review process brings numerous benefits, such as the ability to update the data (e.g., a longer time frame, new measures). Yet, concerns about technical feasibility (e.g., storage requirements, continued availability of data source) grow as the data extraction horizon extends. Similarly, from an ethical perspective, the longer the data extraction, the greater the likelihood of potentially identifying individuals via triangulation. Next to ethics, long-term data collection also places a heavier load on servers, potentially increasing exposure to legal risks.

How to Process the Information During the Extraction? (Challenge #2.4)

As a final step in designing the data extraction, researchers decide how to process the information *while it is collected*.

Table 4. Challenges and Solutions in Extracting Web Data.

Challenge #3.1: Improving Performance	
Reason for importance	In scaling up the data collection, researchers might encounter new technical issues. For example, the data collection could stop unexpectedly or proceed much slower than anticipated.
Solutions and best practices	• When scraping, use stable selectors (e.g., tags, classes, attributes, styles associated with particular information) and make only selective use of error handling • When using APIs, choose a stable and supported version • Attempt to reparse data from stored raw data if the extraction failed • Check for traces of being banned/blocked/slowed down by the website (e.g., by scanning the content that was retrieved) • Notify data sources about potential bandwidth issues with APIs • Update the technically feasible retrieval limit, and recalculate desired sample size, extraction frequency, etc. • Verify that computing resources are appropriate and reliable (e.g., scale up or down servers, verify that database runs optimally) • Move data to a remote (and more scalable) file storage or database • Consider potential benefits from using cloud computing (e.g., for extended, uninterrupted data collection) vs. benefits from local setups (e.g., due to security or privacy concerns) • Budget the expected costs of API subscriptions, cloud computing and data storage and transfer
Challenge #3.2: Monitoring Data Quality	
Reason for importance	Monitoring is critical to be timely alerted to data quality issues. Setting up a monitoring system allows researchers to intervene before discarding data altogether.
Solutions and best practices	• Log each web request (i.e., URL call), along with response status codes, timestamps of when the collection was started, and when the request was made • Save raw data (i.e., source code of HTML websites), along with the parsed data for triangulation • Verify whether the raw data was correctly parsed (e.g., for a sample of information, compare raw data and parsed data) • Check file sizes or the number of observations at regular intervals • Set up a monitoring tool to timely alert you to any future issues (e.g., based on the number of files retrieved or requests made, file sizes retrieved, time the collection last ran, budget spent) • Automatically generate reports on data quality (e.g., using RMarkdown) • Record issue(s) in a logbook (e.g., in the documentation); especially if considered critical for data quality
Challenge #3.3: Documenting Data	
Reason for importance	Researchers are responsible for documenting the data set they produce from web data. Building documentation during the collection is important to guarantee accuracy and completeness, which facilitates use, reuse, and replicability.
Solutions and best practices	• Maintain a logbook in which to note important events (e.g., when the collection broke down and why) • Start writing the documentation in the early stages of collecting the data, and make use of templates (e.g., Datasheets for Datasets; Gebru et al. 2020) • Keep and organize copies of relevant files (e.g., screenshots of the website at the time of data extraction, the API documentation, details on variable operationalization with summary statistics, information about the context) • Have a plan for long-term archival storage (e.g., re3data.org, The Dataverse Project, Zenodo), anonymization (e.g., generating synthetic versions of sensitive data), and consider which license to use for the data (e.g., Creative Commons)

Any kind of web data collection requires a minimal degree of processing, given that the information is available in a computer's memory (e.g., in the browser or the software processing the API output) and still needs to be stored in files or databases. Thus, this processing step occurs *before* data sets are cleaned or analyzed.

When deciding on how to process information during the extraction, researchers must balance potential efficiency gains from molding raw web data into readily usable data sets with the potential threats to validity due to "on-the-fly" processing. For example, in our Amazon example, researchers may be tempted to remove seemingly unnecessary information (e.g., image links in reviews), apply text processing (e.g., removing characters used as separators), or force specific information (e.g., prices) to be stored in a strictly numeric format. Such on-the-fly processing promises to produce essential efficiency gains, as the data set resulting from the extraction could directly be analyzed. However, because on-the-fly processing decisions are usually made after the inspection of only a limited number of pages in early prototypes of the data collection, it is difficult to guarantee their correctness. For example, using our example, what if the initial screening revealed only pictures posted in a review, while the extensive data collection revealed the need to capture video files? Given this and related challenges, keeping the raw data (such as the source code of websites, API output, or any media files loaded at the time of data extraction) is ideal

from a validity perspective. For example, even if the data collection breaks, researchers could still process and use the information after debugging their extraction code. Retaining the raw data can also help reduce Type 1 errors by increasing transparency about researchers' degree of freedom in collecting and processing the web data. Yet, retaining the raw data prompts significant concerns about the technical feasibility and ethical risks. From a technical standpoint, storing the raw data might require databases to retain their original structure and facilitate processing, especially for projects involving many raw data files collected over extended periods. Keeping all raw data might raise questions regarding the right to store the raw data—especially if it is not (pseudo-) anonymized before storage.

Finally, retaining the raw data allows researchers to refine their extraction design at later project stages. For example, a researcher might have collected Amazon reviews in 2018—around the time of the removal of the "not helpful" voting feature. Although extracting "not helpful" votes was not part of the original extraction design, researchers would be able to use the raw web data to examine the effect of the removal of these "not helpful" votes.

Collecting the Data

After source selection and designing the data collection, researchers gradually transition to turning their small-scale prototype into stable extraction software. In so doing, researchers face three challenges. First, researchers may need to improve the performance of their extraction software when operating it automatically at scale (challenge #3.1). Second, they may need to implement monitoring checks to be alerted to any issues arising during extended data collections (challenge #3.2). Third, researchers should compile information important for documenting the final data set (challenge #3.3). Table 4 contains a summary of solutions and best practices to these challenges.

How to Improve the Performance of the Data Extraction? (Challenge #3.1)

In scaling up their data extraction, researchers may notice that the extraction software frequently breaks across a larger number of pages or runs significantly slower than expected. Such technical challenges, if unaddressed, have the potential to undermine research validity (e.g., missing data, not meeting sample size requirements). A practical solution to preempt these and similar challenges involves capturing the focal information in different ways and storing raw data—especially in the early stages of data collection and for more ambitious, large-scale web data collection projects. To track whether the extraction targets are met, researchers can log the (timestamped) URLs of scraped pages and visualize the performance of the extraction software over an extended period. The resulting "effective" extraction frequency can then be used in recomputing the technically feasible sample size (see Web Appendix F). Novel web scraping services promise to handle technical difficulties efficiently (e.g., ScrapingBee, Zyte).

How to Monitor Data Quality During the Extraction? (Challenge #3.2)

As a next step, researchers consider which metadata can help them diagnose issues with the data collection in real-time. Especially when websites constantly change, monitoring the health of web scrapers can be a tedious task. Researchers should consider performance at a higher level (e.g., the file sizes of extracted raw data) and lower level (i.e., the accuracy of the information in resulting data files) to assess whether the collection is proceeding as expected. When collecting over long periods, automatic reporting can greatly facilitate monitoring. Finally, alerts (e.g., via email or mobile) can help researchers detect predefined data issues quickly.

How to Document the Data During and After the Extraction? (Challenge #3.3)

During the data extraction, researchers need to record relevant information about the data in real-time. This is an essential step in building documentation, enabling future data usage by the researcher(s) who collected the data and other scholars. Even after the data extraction has ended, researchers can continuously refine the documentation as they become familiar with the characteristics of the data (e.g., variables that were erroneously captured, missing values).

Accurate and comprehensive documentation is particularly critical given that collecting web data tends to involve repeated iterations between discovery (and often troubleshooting) and confirmation (i.e., subsequent analyses that are outside the scope of our framework). Designing web data extractions requires a different mindset compared with experiments or archival research. Unlike running experiments, the extraction design for collecting web data may be in flux, even when the collection is already running. Relative to traditional archival research in which data sets are sufficiently annotated, researchers are in charge of accurately recalling details about the data collection. Such details encompass information about the data composition (e.g., sampled units), extraction process (e.g., annotated code, detected errors during the collection), and processing details (e.g., applied cleaning steps). The template of Gebru et al. (2020) provides a useful starting point for building the documentation for a data set collected via web scraping or APIs. Given that contextual changes are inevitable (see challenge #1.3), documenting the source's institutional background (e.g., screenshots, corporate blog posts, API documentation) is crucial.

Future Research Opportunities with Web Data

An unprecedented gold rush of web data has enriched the marketing discipline for two decades—over 300 published articles provide countless examples of impactful marketing insights using web data. With the ever-increasing digitization of social

and commercial life, it is hard to imagine that the heyday of this gold rush might subside any time soon. Yet, are marketing's currently productive mines the only or the most promising sources of marketing insights in the future? Which novel approaches and technologies are necessary to capture and describe evolving marketplace realities?

To identify directions for future research, we have reviewed more than 300 articles to provide a snapshot of the current state of web data in marketing. We use these insights to inform the subsequent discussion, which we organize along the four pathways through which web data can advance marketing thought (as summarized in Table 1). We supplement our discussion with key elements from our methodological framework (see Figure 2) and inspiring use cases from other disciplines.

Direction 1: Identify New Web Data Sources

Next, we discuss how researchers can use source selection to branch out to new or underutilized sources for studying emerging substantive topics. We also highlight how researchers can design more complex, longitudinal, and multisource web data sets to reveal otherwise invisible phenomena.

Draw from underutilized sources. Our review reveals that marketing research draws from a somewhat concentrated list of web sources (see Table W4 in Web Appendix C). We encourage researchers to focus on underused or niche sources that have received limited or no attention in marketing. Web data are often prized, as they allow for collecting "consequential dependent variables from the 'real world'" (Inman et al. 2018, p. 357). Identifying new sources or novel consequential variables constitutes a promising avenue for discovering emerging phenomena.

Consider, for example, the twilight state of the nascent legal cannabis industry in the United States. While more states are legalizing cannabis for medical and recreational use, the market value of the *legal* U.S. cannabis industry was still less than a third of the illegal market in 2020 (i.e., $20 billion vs. $66 billion; Franklin 2021). Using surveys, media coverage, and in-depth interviews, marketing scholars have begun to explore how such legalized markets emerge and seek legitimacy (Huff, Humphreys, and Wilner 2021). Sociologists and organizational scholars, in turn, have already used web data to compile intriguing data sets from sources such as Weedmaps. Using these data, they examine, for example, how existing medical cannabis dispensaries have repositioned themselves after the entry of recreational dispensaries (Hsu, Kovács and Koçak 2019) or how consumers deal with potential stigma transfer (Khessina, Reis, and Cameron Verhaal 2021). By leveraging similar web data, marketing researchers could explore intriguing marketing questions. For instance, how should brands position themselves (e.g., brand personalities, emphasis on product vs. service), depending on the strength of categorical stigma? What are the potential public health and welfare implications of the increasing competition among cannabis dispensaries or their growing social media activities?

In addition to being attuned to work in other disciplines, a low-tech route for source exploration is provided by Similarweb, which allows researchers to browse website rankings by region or category. Given the broad accessibility of web sources worldwide, the dominance of Northern American and European data sources is surprising. Not a single article focuses exclusively on African web sources, and only a handful of articles use some African data (e.g., Kübler et al. 2018). Possible starting points for branching out into these underexplored marketplaces could be popular websites such as Nairaland.com (online community), bidorbuy.co.za (auction platform), and Jumia.com.ng (e-commerce).

Build unique and rich data sets by drawing from multiple sources. Most published marketing articles use web data gathered from a single source. Only very few articles collect data from a large number of web sources (i.e., 50 or more web sources). Following the lead of these articles, we encourage marketing researchers to envision unique data sets compiled from many and diverse sources. For example, in economics, Cavallo (2017) collected online and offline prices for individual goods sold by 56 large multichannel retailers in ten countries (i.e., United States, United Kingdom, Argentina, Australia, Brazil, Canada, China, Germany, Japan, and South Africa) between 2014 and 2016. This "Billion Prices Project" (bpp.mit.edu; Cavallo and Rigobon 2016) exemplifies how creative and ambitious data collection from diverse web sources can fuel entire research programs. Especially if sufficiently documented, such web data are poised to unearth new fields of gold for the marketing discipline.

Rediscover frequently used sources. As researchers decide which information to extract (see challenge #2.1), they may overlook novel information on sources they already know. Therefore, refocusing on different information may also reveal how to study novel phenomena on frequently used sources. Adopting a "discovery mode" may reveal that phenomena of high societal relevance such as gender or racial issues are occurring at frequently used sources such as TripAdvisor (Proserpio, Troncoso, and Valsesia 2021), Kickstarter (Younkin and Kuppuswamy 2018), and DonorsChoose (Agarwal and Sen 2022). For example, in entrepreneurship, Younkin and Kuppuswamy (2018) scraped Kickstarter information to examine whether male African American founders are less successful in crowdfunding. Researchers in marketing, in turn, could build on these and similar ideas to explore whether biases exist in other online market exchanges.

Alter the extraction frequency. Another promising lever for exploring emerging phenomena is the extraction frequency (challenge #2.3). In most articles, the data were extracted once (e.g., on a single occasion). Extracting data once is sufficient for many research objectives, such as demonstrating the prevalence of a phenomenon in the marketplace (e.g., Tonietto and Barasch 2020). Yet, researchers can also uncover novel marketing phenomena by creatively envisioning web data sets that only reveal the phenomenon if the information is extracted multiple times. For example, He, Hollenbeck, and Proserpio (2022) leverage

the observation that Amazon removed certain reviews to study the market for "fake" reviews. Specifically, they combine repeatedly web-scraped data from Amazon with hand-coded data from large private groups on Facebook used to solicit fake reviews to examine the short- and long-term impact of such rating manipulations. This example illustrates that data imperfections (e.g., data modifications discovered when mapping the data context, see challenge #1.3) can be opportunities to pose novel research questions rather than merely nuisances that warrant correction.

Direction 2: Harvest the Versatility of Web Data to Boost Ecological Value

As a second direction for knowledge discovery, web data are often used to increase the ecological value of marketing research by complementing carefully controlled experiments. Triangulating findings generated via different methods is fruitful. Yet, there are many other underutilized avenues for how researchers can select and extract web data to infuse ecological validity into experiments and other types of marketing studies.

Infuse ecological validity into experimental stimuli. By carefully selecting websites and APIs, researchers can enhance the ecological validity of their experiments (e.g., through more realistic or diverse stimuli and measures). This enormous potential has hardly been realized in marketing, particularly at scale (for a creative smaller-scale application, see Moore [2015]). Social psychologists demonstrate the full potential of such an approach. Consider, for example, Howe and Monin (2017), who scraped 87 real-world profiles of doctors (including their fitness habits) from the website of a health insurance provider. These profiles served as the foundation for a novel stimulus-sampling paradigm wherein participants in experiments were presented with randomly selected subsets (i.e., five fitness-focused and five non-fitness-focused profiles). In doing so, the authors first ground the phenomenon in the field (i.e., that doctors signal their fitness habits) and then use stimuli created from real profiles to demonstrate that overweight and obese individuals are less likely to choose fitness-focused doctors for their own care. Such triangulation and the creation of larger and more representative samples of naturalistic stimuli enhance the replicability and generalizability of experimental effects (Judd, Westfall, and Kenny 2017). The experimental paradigms in core marketing topics (e.g., branding, advertising, pricing) and methods (e.g., lab experiments, conjoint studies) could benefit from similar applications to mimic real marketplaces. For instance, branding or advertising researchers might develop stimuli based on data extracted from sources like crowdfunding platforms or Bing's Image Search API (e.g., brand logos, ads, and slogans).

Run self-administered field experiments via APIs. While field experiments continue to be prized for their realism and high ecological value (Van Heerde et al. 2021), very few published marketing articles use APIs to run field experiments (e.g., Lambrecht, Tucker, and Wiertz 2018; Toubia and Stephen 2013). There are many untapped opportunities to run field experiments administered by researchers rather than cooperating partners (e.g., firms or charities). Using APIs to run field experiments gives researchers more control over the design and debriefing processes and allows for monitoring of granular participant behavior over longer periods. Thus, web data–based field experiments potentially produce more precise effect sizes and allow researchers to capture long-term effects (Gneezy 2017). In such experiments, researchers might randomly assign users to different treatments, such as adding (vs. not adding) followers on Twitter (Toubia and Stephen 2013) or assigning (vs. not assigning) Reddit's Gold Awards to user posts (Burtch et al. 2022). By gathering high-frequency data via APIs, researchers can analyze how experimental treatments influence outcomes such as posting or the creativity of user-generated content. Alternatively, APIs can be leveraged to infuse realism into experiments, as embodied in Liu and Toubia (2018), who developed "Hoogle," a mock search engine that relies on APIs offered by Google but only displays organic search results that are not altered based on previous user queries. We foresee many more creative future applications of web data to facilitate such field experimentation.

Direction 3: Adopt New Metrics and Methods for Generating Marketing Insights

A core topic in marketing research is to develop marketing metrics that can guide managerial decision making. Traditionally, many metrics have been based on offline information and established data providers (e.g., Farris et al. 2010). Given the continued growth and diversification of web data, it is tempting for marketing *managers* to focus more on web data for managing firm growth and profitability. Yet, deciding which information to select and extract for marketing insight is challenging (see challenge #2.1). More research is needed to help managers avoid succumbing to the streetlight effect (i.e., an "overreliance on readily available data due to ease of measurement and application, irrespective of their growth objective"; Du et al. 2021, pp. 164–65). But, how can researchers get started?

Explore which web sources provide cheaper, faster, or better marketing metrics. Over the last decade, scholars have begun to explore which types of web data could proxy or improve on existing core marketing metrics. For example, managers may use search data extracted from Google to spot trends in the relative importance of their firm's product attributes, which is more cost effective than traditional methods (Du, Hu, and Damangir 2015). Mining Twitter data provides cheaper, real-time, and more actionable measures and insights about brand reputation than existing survey-based metrics like the Brand Asset Valuator data from the advertising agency VMLY&R (Rust et al. 2021). Yet, in other circumstances, readily and cheaply available web data might not be a good substitute for more expensive or established proprietary data sources to uncover market structure (Ringel and Skiera 2016).

An exciting direction for future research is to explore what web data sources should be selected or combined to generate marketing insights that fuel firm growth. For instance, many novel metrics rely

on textual data (Berger et al. 2020). This focus limits applications to markets using the same language employed by the original method (i.e., mostly English). Future research could explore what other types of web data might enable the creation of metrics and insights that allow real-time monitoring and managing diverse global markets. What insights can managers draw from differences and commonalities between the volume of different kinds of internet searches available via Google Trends (e.g., web search vs. image search vs. Google Shopping vs. YouTube search)? Alternatively, what insights about consumer preferences (or any other stakeholder) can be extracted from short videos posted on platforms such as TikTok?

Operate API-based microservices. A fascinating opportunity arises from providing microservices via APIs to marketing stakeholders. This means that researchers not only use APIs to retrieve data but can also operate their own APIs to examine real marketplaces (e.g., using rplumber.io in R). Researchers in data science, for example, offer firms a framework for testing multiarmed bandit policies via APIs while at the same time gathering field experimental data (Kruijswijk et al. 2020). Marketing researchers could use similar API-powered microservices to study emerging topics such as recommendation systems (and resulting biases) or tap into a firm's customer relationship management system to validate new customer churn models. At a small scale, researcher-powered APIs could lower the entry barriers for firms to experiment with novel algorithms that have not yet been implemented in major software packages.

The provision of APIs provides access to novel types of data, while also increasing the timeliness and ecological value of such data. For example, consider the differences between web data collected by a web scraper and the underlying clickstream data stored in the company's database. The website may merely show aggregate statistics about the number of reviews posted. At the same time, the underlying clickstream data also feature information on every website visit (e.g., time, IP address). As with self-administered APIs, researchers define which information a company should submit (e.g., as input to a recommendation algorithm). Thus, researchers can gain access to unique firm data that are otherwise difficult to obtain. For example, large-scale studies with image and video data are still scarce in marketing. Offering image and video analysis as microservices may generate knowledge discovery for new image sources, such as GIFs used in social media (e.g., Giphy).

Direction 4: Exploit Efficiency Gains to Improve Measurement

Web data also have advanced marketing by improving measurement by efficiently collecting diverse variables. Therefore, as a fourth direction, we discuss how web data can improve measurement across the discipline, particularly by rejuvenating interest in core marketing topics (e.g., market orientation, advertising; for an overview of these topics, see Jedidi et al. [2021]). Relatedly, researchers can also leverage APIs to effectively integrate algorithms for processing unstructured data at scale into empirical analyses (Wedel and Kannan 2016).

Leverage web sources to describe diverse online and offline behaviors. Most marketing articles gather web data to describe and examine behavior occurring online. As documented in Table W4 in Web Appendix C, many of the used sources in marketing are focused on online *consumer* behaviors, such as e-commerce websites (e.g., Amazon), online reviewing platforms (e.g., Yelp), social media sites (e.g., Twitter), and search engines (e.g., Google Trends). Relatively less research has focused on *firm* behavior online. Yet, by doing so, researchers could explore many core marketing constructs (e.g., service orientation, sustainability). For example, researchers could systematically collect information available on the websites of many firms to analyze which organizational factors influence how firms signal their service orientation (e.g., employees' digital presence; Herhausen et al. 2020) or environmental credentials (e.g., the B Corporations certification; Gehman and Grimes 2017) to customers and other stakeholders.

We encourage marketing researchers who have not yet used web data in their research to consider websites and APIs as valuable, rich, and timely sources to exploit the increased digitization of *all forms* of behaviors—not only online behavior. A recent example of bringing web data into an established "offline" research stream is Homburg, Theel, and Hohenberg (2020), who scraped the annual reports of more than 8,000 firms from AnnualReports.com between 1998 and 2016. Web sources contain historical information about periods, even long before the web in its current form existed (e.g., 1998 in this case). The authors subsequently use these reports to develop a novel text-based measure of marketing excellence derived from firm letters to shareholders. Many other untapped online sources (e.g., job posting platforms) offer new insights into how firms communicate their marketing capabilities to external stakeholders beyond consumers, such as prospective employees, social activists, and investors.

Particularly for the marketing–finance interface, the web features many understudied forms of investor-facing communication that are ripe for collection at scale. For example, which type of marketing topics besides marketing excellence (e.g., marketing capabilities, brand positioning, pricing) should top management emphasize to investors to increase firm valuation during investor relations presentations, investor days, or earnings calls? Researchers could also examine the relative importance of the content versus the delivery (e.g., the tone of the speaker on the recording of an investor day; see Wang et al. 2021). Such multimodal data can also benefit the inferences made in established research streams.

Embrace APIs for better measurement. APIs offer many opportunities for improving measurement—some of which are unexpected. For example, consumer researchers planning to run longitudinal studies might consider APIs for automating processes for managing participants at scale, thereby reducing the operating costs (and potentially boosting sample size). The Amazon Mechanical Turk

API and the various Prolific Academic APIs (e.g., Study API) are good starting points for running multiwave studies.

APIs also enable much more than just retrieving data. For example, to reduce validity concerns in long-term data collection, researchers can use the Pushover API (https://pushover.net/api) to send monitoring alerts to their smartphones. The API of Amazon Web Services allows for the orchestration of virtual computing infrastructure (e.g., to capture data from different countries). Another fruitful avenue in which APIs are currently underused in marketing is facilitating stimuli selection. For example, a classic area of inquiry in marketing is how (background) music affects product and brand perceptions and choices (e.g., Bruner 1990). In 2022, background music is quite different (e.g., self-chosen, more diverse use cases). Researchers could use the Spotify Web API to select stimuli from millions of mood, sleep, or study playlists, thereby discovering perfect "lookalikes" that only differ on one focal attribute (e.g., tempo) but not on other acoustic attributes available at the API (e.g., valence, loudness). Even in this simple example, there might be a substantive interest in better understanding the effect of new background music on consumption choices, especially given the shift to working and studying from home.

Concluding Thoughts

Web data have unearthed many fields of gold in marketing. However, extracting data for generating relevant and valid research insights is challenging. Our article highlights validity concerns that require the joint consideration of idiosyncratic technical, legal, and ethical questions. We introduce a novel methodological framework (Figure 2), offer practical solutions (Tables 2–4), and outline directions for future research to enable researchers to create impactful and credible marketing knowledge. While our focus is primarily on authors, our work also spotlights crucial validity concerns to scholars reviewing web data–based research and practitioners interested in deriving accurate and actionable marketing insights from web data.

We hope that our work encourages marketing scholars to integrate web data into their research programs. While web data often provide compelling answers to the question, "Assuming that this hypothesis is true, in what ways does it manifest in the world?" (Barnes et al. 2018, p. 1455), this does not imply that web data are relevant for all research projects. Web data sources tend to feature a large N (i.e., many users) with many V (i.e., different pieces of information for potential variables) observable over a large T (i.e., many observations over extended periods of time and at a very granular level; Adjerid and Kelley 2018). Yet, collecting web data via web scraping or APIs provides limited information about the browsing behavior of individuals on the website that led to the creation of the data in the first place. Significant synergies exist by enriching clickstream stream data capturing such browsing processes with web data retrieved from web scraping and APIs (e.g., Li et al. 2020).

Our work aims to bridge entrenched training silos (e.g., between quantitative marketing and consumer behavior). We encourage scholars to further integrate and leverage existing best practices with regard to the collection and analysis of web data (e.g., preregistration, addressing endogeneity). There is significant untapped potential for collaborations across methodological traditions to explore and exploit new fields of gold. Collecting valid web data can enable marketing as a discipline to enhance its relevance and assert intellectual leadership on important emerging substantive topics that are also increasingly studied in fields such as computer science, information systems, and management science (Moorman et al. 2019).

We would be remiss not to mention the nonmonetary costs of collecting web data via web scraping and APIs. While browsing the web is (mainly) free, researchers should not assume that collecting web data is costless. The prototype of a data collection can be ready and running in a matter of hours. Yet, researchers will often find out that the data collection does not work entirely as intended or encounter some of the challenges discussed in our methodological framework. Just like with any other method, the devil is in the details.

Web data democratize data access and make our discipline more inclusive for scholars who would otherwise find it difficult to obtain access to data. To further reduce entry barriers, it would be helpful to create incentives (e.g., journal space) for rich web data sets and their documentation, like the Billion Prices Project (Cavallo and Rigobon 2016). Similarly, authors can make their algorithms or data available for other researchers by sharing code publicly or deploying API-based microservices that can increase their methods' adoption and offer unique opportunities for field experimentation. In summary, web data present a golden opportunity to examine important marketing questions, now and in the future.

Acknowledgments

The authors would like to thank Christine Moorman, the associate editor, and four anonymous reviewers for their constructive comments. They are grateful to Quentin André, Pierre Chandon, Darren Dahl, Jonne Guyt, Rishad Habib, Jan Klein, Bastian Jaeger, Ana Martinovici, Suni Mendis, Gabriele Paolacci, Rik Pieters, Stefano Puntoni, Dan Schley, and participants at the marketing brown bag seminar at INSEAD, the lunch club at the Rotterdam School of Management, the ACR 2019 LearnShop, and participants of Tilburg University's course on Online Data Collection and Management for their helpful comments on the manuscript. Research support from Tilburg Science Hub is gratefully acknowledged. A digital companion to this article is available at https://web-scraping.org.

Associate Editor

Oded Netzer

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Marketing Science Institute (grant #4000678) and the Dutch Research Council (NWO #451–17–028).

ORCID iDs

Johannes Boegershausen https://orcid.org/0000-0002-1429-9344
Hannes Datta https://orcid.org/0000-0002-8723-6002

References

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

Agarwal, Saharsh and Ananya Sen (2022), "Antiracist Curriculum and Digital Platforms: Evidence from Black Lives Matter," *Management Science*, 68 (4), 2932–48.

Anderson, Eric T. and Duncan I. Simester (2014), "Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research*, 51 (3), 249–69.

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.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer, and David A. Schweidel (2020), "Uniting the Tribes: Using Text for Marketing Insight," *Journal of Marketing*, 84 (1), 1–25.

Blaseg, Daniel, Christian Schulze, and Bernd Skiera (2020), "Consumer Protection on Kickstarter," *Marketing Science*, 39 (1), 211–33.

Bruner, Gordon C. (1990), "Music, Mood, and Marketing," *Journal of Marketing*, 54 (4), 94–104.

Burtch, Gordon, Qinglai He, Yili Hong, and Dokyun Lee (2022), "How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit," *Management Science*, 68 (5), 3488–4506.

Cavallo, Alberto (2017), "Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers," *American Economic Review*, 107 (1), 283–303.

Cavallo, Alberto and Roberto Rigobon (2016), "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives*, 30 (2), 151–78.

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

Chen, Yubo, Qi Wang, and Jinhong Xie (2011), "Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning," *Journal of Marketing Research*, 48 (2), 238–54.

Chevalier, Judith A. and Dina Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43 (3), 345–54.

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.

Datta, Hannes, Harald J. van Heerde, Marnik G. Dekimpe, and Jan-Benedict E.M. Steenkamp (2022), "Cross-National Differences in Market Response: Line-Length, Price, and Distribution Elasticities in Fourteen Indo-Pacific Rim Economies," *Journal of Marketing Research*, 59 (2), 251–70.

Du, Rex Yuxing, Ye Hu, and Sina Damangir (2015), "Leveraging Trends in Online Searches for Product Features in Market Response Modeling," *Journal of Marketing*, 79 (1), 29–43.

Du, Rex Yuxing, Oded Netzer, David A. Schweidel, and Debanjan Mitra (2021), "Capturing Marketing Information to Fuel Growth," *Journal of Marketing*, 85 (1), 163–83.

Edelman, Benjamin (2012), "Using Internet Data for Economic Research," *Journal of Economic Perspectives*, 26 (2), 189–206.

etailinsights (2021), "How Many Etailers Are in the US?" infographic (accessed December 25, 2021), https://www.etailinsights.com/online-retailer-market-size.

Farris, Paul W., Neil T. Bendle, Phillip E. Pfeifer, and David J. Reibstein (2010), *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Upper Saddle River, NJ: Pearson Education.

Franklin, Joshua (2021), "Banking on Cannabis: The New Network of Lenders for a Semi-Legal Industry," *Financial Times* (August 25), https://on.ft.com/3BeC3Lt.

Frederick, Shane, Leonard Lee, and Ernest Baskin (2014), "The Limits of Attraction," *Journal of Marketing Research*, 51 (4), 487–507.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, et al. (2020), "Datasheets for Datasets," arXiv preprint arXiv:1803.09010.

Gehman, Joel and Matthew Grimes (2017), "Hidden Badge of Honor: How Contextual Distinctiveness Affects Category Promotion among Certified B Corporations," *Academy of Management Journal*, 60 (6), 2294–2320.

Gneezy, Ayelet (2017), "Field Experimentation in Marketing Research," *Journal of Marketing Research*, 54 (1), 140–43.

Godes, David and Dina Mayzlin (2004), "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science*, 23 (4), 545–60.

Hanna, Robin (2018), "Amazon on a Positive Note: The End of Downvoting," (accessed December 1, 2021), https://sellics.com/blog-amazon-on-a-positive-note-the-end-of-downvoting/.

He, Sherry, Brett Hollenbeck, and Davide Proserpio (2022), "The Market for Fake Reviews," *Marketing Science* (published online February 25), https://doi.org/10.1287/mksc.2022.1353.

Herhausen, Dennis, Oliver Emrich, Dhruv Grewal, Petra Kipfelsberger, and Marcus Schoegel (2020), "Face Forward: How Employees' Digital Presence on Service Websites Affects Customer Perceptions of Website and Employee Service Quality," *Journal of Marketing Research*, 57 (5), 917–36.

Hermosilla, Manuel, Fernanda Gutiérrez-Navratil, and Juan Prieto-Rodríguez (2018), "Can Emerging Markets Tilt Global Product Design? Impacts of Chinese Colorism on Hollywood Castings," *Marketing Science*, 37 (3), 356–81.

Homburg, Christian, Marcus Theel, and Sebastian Hohenberg (2020), "Marketing Excellence: Nature, Measurement, and Investor Valuations," *Journal of Marketing*, 84 (4), 1–22.

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

Hsu, Greta, Balázs Kovács, and Özgecan Koçak (2019), "Experientially Diverse Customers and Organizational Adaptation in Changing

Demand Landscapes: A Study of US Cannabis Markets, 2014–2016," *Strategic Management Journal*, 40 (13), 2214–41.

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.

Huang, Yufeng (2019), "Learning by Doing and the Demand for Advanced Products," *Marketing Science*, 38 (1), 107–28.

Huber, Joel, John W. Payne, and Christopher Puto (1982), "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis," *Journal of Consumer Research*, 9 (1), 90–98.

Huff, Aimee Dinnin, Ashlee Humphreys, and Sarah J. S. Wilner (2021), "The Politicization of Objects: Meaning and Materiality in the U.S. Cannabis Market," *Journal of Consumer Research*, 48 (1), 22–50.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–1306.

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.

Jedidi, Kamel, Bernd H. Schmitt, Malek Ben Sliman, and Yanyan Li (2021), "R2M Index 1.0: Assessing the Practical Relevance of Academic Marketing Articles," *Journal of Marketing*, 85 (5), 22–41.

Judd, Charles M., Jacob Westfall, and David A. Kenny (2017), "Experiments with More Than One Random Factor: Designs, Analytic Models, and Statistical Power," *Annual Review of Psychology*, 68 (1), 601–25.

Khessina, Olga M., Samira Reis, and J. Cameron Verhaal (2021), "Stepping out of the Shadows: Identity Exposure as a Remedy for Stigma Transfer Concerns in the Medical Marijuana Market," *Administrative Science Quarterly*, 66 (3), 569–611.

Kim, Tongil "TI" and Diwas KC (2020), "Can Viagra Advertising Make More Babies? Direct-to-Consumer Advertising on Public Health Outcomes," *Journal of Marketing Research*, 57 (4), 599–616.

Kozinets, Robert V. (2001), "Utopian Enterprise: Articulating the Meanings of Star Trek's Culture of Consumption," *Journal of Consumer Research*, 28 (1), 67–88.

Kozinets, Robert V. (2002), "The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities," *Journal of Marketing Research*, 39 (1), 61–72.

Kozinets, Robert V. (2020), *Netnography: The Essential Guide to Qualitative Social Media Research*, 3rd ed., London: SAGE Publications.

Kruijswijk, Jules, Robin van Emden, Petri Parvinen, and Maurits Kaptein (2020), "StreamingBandit: Experimenting with Bandit Policies," *Journal of Statistical Software*, 94 (9), 1–47.

Kübler, Raoul, Koen Pauwels, Gökhan Yildirim, and Thomas Fandrich (2018), "App Popularity: Where in the World Are Consumers Most Sensitive to Price and User Ratings?" *Journal of Marketing*, 82 (5), 20–44.

Lakens, Daniël (2022), "Sample Size Justification," Collabra: Psychology, 8 (1): 33267.

Lambrecht, Anja, Catherine Tucker, and Caroline Wiertz (2018), "Advertising to Early Trend Propagators: Evidence from Twitter," *Marketing Science*, 37 (2), 177–99.

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.

Li, Chenxi, Xueming Luo, Cheng Zhang, and Xiaoyi Wang (2017), "Sunny, Rainy, and Cloudy with a Chance of Mobile Promotion Effectiveness," *Marketing Science*, 36 (5), 762–79.

Li, Jingjing, Ahmed Abbasi, Amar Cheema, and Linda B. Abraham (2020), "Path to Purpose? How Online Customer Journeys Differ for Hedonic Versus Utilitarian Purchases," *Journal of Marketing*, 84 (4), 127–46.

Li, Xi, Mengze Shi, and Xin Wang (2019), "Video Mining: Measuring Visual Information Using Automatic Methods," *International Journal of Research in Marketing*, 36 (2), 216–31.

Liu, Jia and Olivier Toubia (2018), "A Semantic Approach for Estimating Consumer Content Preferences from Online Search Queries," *Marketing Science*, 37 (6), 930–52.

Liu, Liu, Daria Dzyabura, and Natalie Mizik (2020), "Visual Listening In: Extracting Brand Image Portrayed on Social Media," *Marketing Science*, 39 (4), 669–86.

Martin, Kelly D., Abhishek Borah, and Robert W. Palmatier (2017), "Data Privacy: Effects on Customer and Firm Performance," *Journal of Marketing*, 81 (1), 36–58.

McAuley, Julian (2021), "Recommender Systems Datasets," (accessed February 2, 2021), https://cseweb.ucsd.edu/~jmcauley/datasets.html.

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

Melumad, Shiri and Robert Meyer (2020), "Full Disclosure: How Smartphones Enhance Consumer Self-Disclosure," *Journal of Marketing*, 84 (3), 28–45.

Moore, Sarah G. (2015), "Attitude Predictability and Helpfulness in Online Reviews: The Role of Explained Actions and Reactions," *Journal of Consumer Research*, 42 (1), 30–44.

Moorman, Christine, Harald J. van Heerde, C. Page Moreau, and Robert W. Palmatier (2019), "Challenging the Boundaries of Marketing," *Journal of Marketing*, 83 (5), 1–4.

Morales, Andrea C., On Amir, and Leonard Lee (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), 465–76.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, 31 (3), 521–43.

Neuendorf, Kimberly A. (2017), *The Content Analysis Guidebook*. Thousand Oaks, CA: SAGE Publications.

Oestreicher-Singer, Gal, Barak Libai, Liron Sivan, Eyal Carmi, and Ohad Yassin (2013), "The Network Value of Products," *Journal of Marketing*, 77 (3), 1–14.

Peng, Jing, Ashish Agarwal, Kartik Hosanagar, and Raghuram Iyengar (2018), "Network Overlap and Content Sharing on Social Media Platforms," *Journal of Marketing Research*, 55 (4), 571–85.

Proserpio, Davide, Isamar Troncoso, and Francesca Valsesia (2021), "Does Gender Matter? The Effect of Management Responses on Reviewing Behavior," *Marketing Science*, 40 (6), 1199–1213.

Ringel, Daniel M. and Bernd Skiera (2016), "Visualizing Asymmetric Competition Among More Than 1,000 Products Using Big Search Data," *Marketing Science*, 35 (3), 511–34.

Rust, Roland T., William Rand, Ming-Hui Huang, Andrew T. Stephen, Gillian Brooks, and Timur Chabuk (2021), "Real-Time Brand Reputation Tracking Using Social Media," *Journal of Marketing*, 85 (4), 21–43.

Schweidel, David A. and Wendy W. Moe (2014), "Listening in on Social Media: A Joint Model of Sentiment and Venue Format Choice," *Journal of Marketing Research*, 51 (4), 387–402.

Shadish, William, Thomas D. Cook, and Donald Thomas Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sim, Jaeung, Daegon Cho, Youngdeok Hwang, and Rahul Telang (2022), "Frontiers: Virus Shook the Streaming Star: Estimating the COVID-19 Impact on Music Consumption," *Marketing Science*, 41 (1), 19–32.

Sridhar, Shrihari and Raji Srinivasan (2012), "Social Influence Effects in Online Product Ratings," *Journal of Marketing*, 76 (5), 70–88.

Statista (2021), "Media Usage in an Internet Minute as of August 2021," (accessed January 10, 2022), https://www.statista.com/statistics/195140/.

Tellis, Gerard J., Deborah J. MacInnis, Seshadri Tirunillai, and Yanwei Zhang (2019), "What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence," *Journal of Marketing*, 83 (4), 1–20.

Tirunillai, Seshadri and Gerard J. Tellis (2012), "Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance," *Marketing Science*, 31 (2), 198–215.

Tonietto, Gabriela N. and Alixandra Barasch (2020), "Generating Content Increases Enjoyment by Immersing Consumers and Accelerating Perceived Time," *Journal of Marketing*, 85 (6), 83–100.

Toubia, Olivier and Andrew T. Stephen (2013), "Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter?" *Marketing Science*, 32 (3), 368–92.

Trusov, Michael, Liye Ma, and Zainab Jamal (2016), "Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting," *Marketing Science*, 35 (3), 405–26.

Vanden Broucke, Seppe and Bart Baesens (2018), *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. Berkeley, CA: Apress.

Van Heerde, Harald J., Christine Moorman, C. Page Moreau, and Robert W. Palmatier (2021), "Reality Check: Infusing Ecological Value into Academic Marketing Research," *Journal of Marketing*, 85 (2), 1–13.

Villarroel Ordenes, Francisco, 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.

Wang, Xin, Shijie Lu, X.I. Li, Mansur Khamitov, and Neil Bendle (2021), "Audio Mining: The Role of Vocal Tone in Persuasion," *Journal of Consumer Research*, 48 (2), 189–211.

Wang, Yang and Alexander Chaudhry (2018), "When and How Managers' Responses to Online Reviews Affect Subsequent Reviews," *Journal of Marketing Research*, 55 (2), 163–77.

Weber, Matthew S. (2018), "Methods and Approaches to Using Web Archives in Computational Communication Research," *Communication Methods and Measures*, 12 (2/3), 200–215.

Wedel, Michel and P.K. Kannan (2016), "Marketing Analytics for Data-Rich Environments," *Journal of Marketing*, 80 (6), 97–121.

Wells, William D (2001), "The Perils of N = 1," *Journal of Consumer Research*, 28 (3), 494–98.

Wu, Chunhua and Koray Cosguner (2020), "Profiting from the Decoy Effect: A Case Study of an Online Diamond Retailer," *Marketing Science*, 39 (5), 974–95.

Xu, Heng, Nan Zhang, and Le Zhou (2020), "Validity Concerns in Research Using Organic Data," *Journal of Management*, 46 (7), 1257–74.

Younkin, Peter and Venkat Kuppuswamy (2018), "The Colorblind Crowd? Founder Race and Performance in Crowdfunding," *Management Science*, 64 (7), 3269–87.

Zervas, Georgios, Davide Proserpio, and John W. Byers (2017), "The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry," *Journal of Marketing Research*, 54 (5), 687–705.