Marketing Science Institute Working Paper Series 2020
Report No. 20-143

Fields of Gold: Web Scraping for Consumer Research

Johannes Boegershausen, Abhishek Borah and Andrew Stephen

Fields of Gold:
Web Scraping for Consumer Research

JOHANNES BOEGERSHAUSEN

ABHISHEK BORAH

ANDREW T. STEPHEN

Johannes Boegershausen (j.boegershausen@uva.nl) is Assistant Professor of Marketing at Amsterdam Business School, University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands. Abhishek Borah (abhishek.borah@insead.edu) is Assistant Professor of Marketing at INSEAD, INSEAD, Boulevard de Constance, 77305 Fontainebleau, France. Andrew T. Stephen (andrew.stephen@sbs.ox.ac.uk) is the L'Oréal Professor of Marketing, Associate Dean of Research, Saïd Business School, University of Oxford, Park End Street, Oxford OX1 1HP, United Kingdom. The authors would like to thank Quentin André, Pierre Chandon, Darren Dahl, Rishad Habib, Manoj Thomas, and participants at the Rotterdam School of Management lunchclub, the marketing brown bag seminar at INSEAD, and the ACR 2019 LearnShop for their helpful comments on this work. Finally, financial support from the Marketing Science Institute (grant # 4000678) is gratefully acknowledged.

ABSTRACT

The internet plays an increasingly central role in consumers' daily lives. Every second, consumers create terabytes of data containing rich information about their opinions, preferences, and consumption choices. The massive volume and variety of consumers' digital footprints present many opportunities for researchers to examine and test theories about consumer processes and behaviors in the field. Yet, the methodology, i.e., web scraping, which is the process of designing and deploying code that automatically extracts and parses information from websites and provides a goldmine of data for answering consumer research questions, remains a black box—and an untapped opportunity—for many consumer researchers. Simultaneously, there is a lack of consensus about the best way to handle the inherent complexities and idiosyncratic methodological challenges involved in web scraping. To address these issues, this report proposes a structured workflow for conducting, documenting, reporting, and evaluating web scraping—both in the research and peer review processes—in order to generate robust and credible findings from web-scraped data. Embracing web scraping opens new avenues for producing timely, discovery-oriented, and credible scholarly knowledge about consumer behavior.

Keywords: web scraping, research credibility, field data, philosophy of science, word of mouth and user generated content

INTRODUCTION

It is difficult to overstate the role that the internet plays in consumers' daily lives. Many websites, mobile apps, and web-based services are essential to consumers, allowing them to do everything from shopping on Amazon and finding restaurants on Yelp to connecting with friends on Facebook. In January 2019, the average consumer in the developed world spent more than six hours per day online, and more than 85 percent of consumers use the internet every single day (We Are Social & Hootsuite 2019). The proliferation of more powerful, affordable, and versatile mobile devices has also made the internet accessible to a greater number of consumers around the world (Stephen 2016). The approximate number of users in July 2019 was more than 2.4 billion on Facebook, 2 billion on YouTube, 1 billion on Instagram, 330 million on Twitter, and 170 million on Yelp (Statista 2019; Yelp 2019). One important consequence of this phenomenon for consumer researchers is that as consumers immerse themselves in the digital world, they continuously create enormous amounts of data containing information about their attitudes, revealed preferences, and behaviors.

Consumers' digital footprints are vast. The digital traces or "digital breadcrumbs" consumers leave are a potential goldmine for consumer researchers. The massive volume and variety of these digital footprints present significant opportunities for researchers to examine and test theories about consumer processes and behaviors in the field. Publicly available data on the internet provides an unprecedented window into consumer behavior, allowing researchers to quantify social and consumption activities that are otherwise extremely difficult to observe, record, and analyze. *Web Scraping* is the process of designing and deploying code that automatically extracts and parses information from websites. Acquiring an understanding of the

mechanics and nuances of working with web-scraped data, can enable consumer researchers to quickly and systematically transform digital traces into a goldmine of data for answering novel and substantive consumer research questions.

WEB SCRAPING FOR CONSUMER RESEARCH

Capturing these digital footprints via web scraping typically involves "writing an automated program that queries a web server, requests data […], and then parses that data to extract needed information" (Mitchell 2015, p. viii). Several areas of inquiry within consumer research already have made effective use of datasets constructed from these footprints. Figure 1, tables and figures follow references, shows the number of articles using web-scraped data for both non-qualitative (e.g., experimental, mixed-methods) and qualitative (e.g., netnography, consumer culture theory) approaches. Overall, the time trend suggests that the use of web-scraped data in consumer research has been increasing over the last decade, in particular in non-qualitative articles.

Areas of inquiry within consumer research that have relied heavily on web-scraped data have advanced due to the ingenuity of researchers. They have found relevant data on the internet to help generate new knowledge and to test specific hypotheses about consumer behavior. Research on electronic word-of-mouth (WOM) communications and social media, in particular, has frequently used these kinds of data. The majority of the findings regarding user-generated content (UGC) in the context of reviews and online WOM are based on studies that have involved at least some use of publicly accessible data from platforms such as Yelp, TripAdvisor, and Amazon (e.g., Grewal and Stephen 2019; McGraw, Warren, and Kan 2015; Moore 2015).

Importantly, harvesting data from these platforms also allows researchers to move beyond testing theories about UGC to testing a diverse set of psychological theories that are not directly related to UGC, such as construal level (Elder et al. 2017; Huang et al. 2016), competitive salience (Paharia, Avery, and Keinan 2014), and social cognition (Henkel et al. 2018).

Further, social media research in marketing and consumer behavior has made extensive use of data from social media platforms such as Twitter and, in the earlier days of this research stream, blogs. For example, Twitter alone has proven highly valuable, with research using Twitter data advancing our understanding of social sharing behavior (Toubia and Stephen 2013), conspicuous consumption (Bellezza, Paharia, and Keinan 2017), and conversations about brands (Arvidsson and Caliandro 2016; Hewett et al. 2016). Blogs have been another fertile source of data for examining various consumer phenomena by leveraging the rich narratives that consumers provide on these websites (e.g., McQuarrie, Miller, and Phillips 2013; Scaraboto and Fischer 2013).

The aforementioned examples highlight the potential of web-scraped data for generating, testing, extending, and refining theories about consumer behavior. Yet, despite the growing importance and usage of web-scraped data in consumer research, web scraping largely remains a black box to many consumer researchers. With this report, we strive to demystify the underlying mechanics and processes of systematically collecting and analyzing data from the internet via web scraping.

Beyond accessibility, the lack of consensus regarding standards for conducting and evaluating web scraping-based research is an even more important issue for the field. This issue is important because data scraped from the internet is fundamentally different from data generated for consumer research via more conventional research designs (Groves 2011; Xu,

Zhang, and Zhou 2019). As web data is created organically, it is characterized by idiosyncratic threats to validity and unique methodological challenges that require critical reflection and place additional demands on researchers (e.g., Landers et al. 2016; Wenzel and Van Quaquebeke 2018). At present, many published articles do not include sufficient information about the different decisions and judgment calls involved in the data collection, wrangling, and analysis process in web scraping. As a result, it is difficult to reproduce, replicate, and compare findings of consumer research studies using web scraping. Taking inspiration from a framework for generating credible scientific findings by LeBel et al. (2018), we highlight the key challenges, state-of-the-art remedies, best practices, and corresponding standards for evaluation for web scraping in consumer research. Our structured workflow is designed to achieve a sufficient level of consistency and standardization with respect to how web scraping is conducted, documented, reported, and evaluated in both the research and peer review processes.

We begin this report by outlining the types of data that can be scraped from the web for consumer research. Next, we discuss the potential of scraping data from the internet for consumer research. Finally, we offer prescriptive information on best practices for dealing with the inherent challenges in web scraping. We present a structured workflow approach for authors and reviewers to ensure the generation of credible, robust research findings by consumer research employing web scraping.

WHAT TYPES OF DATA CAN BE SCRAPED FROM THE WEB?

While the data resulting from consumers' digital footprints can take many different forms and shapes, there are four major types of data that are relevant for consumer researchers: (1)

numeric, (2) textual, (3) visual, and (4) metadata. As shown in Figure 2, data ranges from relatively structured data on the left to highly unstructured data on the right.

Numeric data (e.g., review ratings on TripAdvisor, helpfulness votes for reviews in Amazon, auction prices on eBay) is first on the left end of the continuum in Figure 2. Numeric data is the most structured type of data available for scraping, and it can be easily parsed into variables with minimal processing.

Textual data (e.g., review text, blog posts, tweets) is in the middle of Figure 2. This type of data is more unstructured and therefore requires a greater level of processing than numeric data in order to make it viable for (statistical) analysis. Typically, this processing involves some form of automated text analysis. Humphreys and Wang (2018) provide a comprehensive overview of the various standardized dictionaries and other approaches commonly employed in automated text analysis in consumer research (see also Berger et al. 2020; Hartmann et al. 2019).

Visual data (e.g., images, videos) is at the end of the continuum on the right side in Figure 2. Visual data is the most unstructured type of data and requires the most processing to parse into variables, regardless of whether it is processed automatically (e.g., Klostermann et al. 2018) or by human coders (e.g., Ordenes et al. 2019). Effective and scalable automated processing that relies on computer vision techniques is becoming increasingly common and accessible (e.g., Klostermann et al. 2018; Li, Shi, and Wang 2019).

Finally, the fourth type of data relevant for consumer researchers is metadata, which is common on websites and can best be thought of as "data about other data." Metadata related to user-generated data of numeric, textual, or visual types might contain descriptive information about the user (e.g., Internet Protocol address), device used (e.g., browser, type of camera), or the timing of the content creation (e.g., posting date of a review; Landers et al. 2016).

Consumer researchers have scraped numeric, textual, visual, and metadata from a variety of websites. The most common sources of data for consumer research articles include review platforms, social commerce sites, auction sites, online communities, (micro-) blogs, and crowdfunding websites. Table 1 offers an overview of the variables that consumer researchers have created from the numeric, textual, visual, and metadata scraped from these websites. Web-scraped data can serve various purposes within consumer research articles, ranging from empirically grounding a phenomenon to more central roles such as confirmatory hypothesis testing. Table 2 provides illustrative examples of the various roles of web-scraped data within consumer research articles.

OPPORTUNITIES FOR WEB SCRAPING IN CONSUMER RESEARCH

The application of web scraping in consumer research creates a wide range of opportunities. In the following, we discuss five opportunities emerging from the application of web scraping in consumer research: collecting consequential variables from the real world, conducting discovery-oriented research, studying socially sensitive and rare phenomena, examining effects over time, and improving the quality of science.

Collection of consequential variables from the real world. As a multi-disciplinary academic field employing diverse methodological lenses (e.g., MacInnis and Folkes 2010), consumer research seeks to develop theories that illuminate and explain consumers' behavior in the "real world" (e.g., Inman et al. 2018). Web-scraped data unobtrusively captures how consumers and managers behave in their natural (online) habitat. As such, it can effectively complement more controlled lab experiments to address concerns about realism in consumer

research. Scraping can empirically ground phenomena (e.g., Bellezza et al. 2017; Simmons et al. 2011) to demonstrate that focal processes occur outside the confines of a controlled lab environment with stylized experimental stimuli (Morales, Amir, and Lee 2017).

As consumers' immersion in the digital world increases, many consequential variables are readily available to be scraped from online review platforms, social commerce sites, and auction platforms (Adjerid and Kelley 2018). Scraping this data allows for the study of actual behaviors (e.g., reviewing, liking, retweeting) in more systematic ways than those afforded by more traditional methods (Hoover et al. 2018). Relative to other data collection techniques, web scraping offers the ability to effectively examine the aggregate behavior of firms (e.g., menu items offered by restaurants; Bellezza and Berger 2019) and service providers (e.g., doctors' descriptions of their hobbies; Howe and Monin 2017). Therefore, web scraping can 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).

Discovery-oriented consumer research. Scraping data allows consumer researchers to be discovery-oriented and "scout out" new effects (Mortensen and Cialdini 2010; Reis 2012). For instance, consumer researchers can leverage the rich consumer narratives on blogs and online communities from samples that are otherwise hard to reach and observe (Kozinets 2002) in order to generate novel theories (e.g., Dolbec and Fischer 2015; McQuarrie et al. 2013). Building datasets via scraping also allows for the study of emerging phenomena that occur primarily in the online realm, such as the emergence of a brand public (Arvidsson and Caliandro 2016) or the effects of management responses to online reviews (Wang and Chaudhry 2018).

Scraped data also enables the exploration of effects at different levels of analysis than those typically available in lab environments. By aggregating data to higher levels (e.g., the

brand- or firm-level), consumer researchers can examine how consumer processes affect outcomes of significant relevance to managers, including product sales (e.g., Berger, Sorensen, and Rasmussen 2010; Borah and Tellis 2016), conversion rates (Ludwig et al. 2013), and content diffusion within social networks (e.g., Brady et al. 2019). Because of the presence of metadata (e.g., geolocation data; Huang et al. 2016), scraping facilitates the discovery of variation across geographic or sociopolitical contexts that holds theoretical significance (Barnes et al. 2018).

Examination of socially sensitive and rare phenomena. Web scraping is well-suited to the unobtrusive study of socially sensitive phenomena (Hoover et al. 2018), such as how controversy influences participation in discussions (Chen and Berger 2013). As data is collected after the behavior occurred, web scraping avoids some of the challenges involved in the study of socially sensitive phenomena via surveys or experiments, such as impression management and social desirability concerns (e.g., Mick 1996; Steenkamp, de Jong, and Baumgartner 2010).

Web scraping also enables large-scale studies of relatively rare events (Bright 2017), hard-to-reach individuals (e.g., celebrities; Bellezza et al. 2017; political elites, Brady et al. 2019; professional athletes, Grijalva et al. 2019), and specific groups of consumers (e.g., early adopters of Spotify; Datta, Knox, and Bronnenberg 2018).

Opportunities for examining effects over time. The digital footprints left by consumers create an enormous volume of data not only in terms of the total number of cases, but also in terms of the number and frequency of traces for one individual consumer over time (Adjerid and Kelley 2018; Matz and Netzer 2017). The velocity and periodicity of data creation on the internet facilitate the construction of panel data that captures variation within individuals' behavior over time as a function of variables of theoretical interest (e.g., Huang et al. 2016; Moore 2012) and can provide insights into how effects unfold over time (e.g., Datta et al. 2018; Doré et al. 2015).

Moreover, the real-time nature of online data (e.g., tweets) can allow researchers to study consumer behavior at a very high granularity such as in seconds, minutes, hours, or days. Another important advantage of panel (vs. cross-sectional) data is that it explicitly accounts for individual differences. By combining data across both case (e.g., firms, individuals, etc.) and time, panel data gives more data variation, less collinearity, and more degrees of freedom. Panel data allows for stronger causal inferences when a difference-in-differences approach is employed to unpack the effects of an intervention (e.g., a change in website design; Wenzel and Van Quaquebeke 2018).

Improving the quality of science. In the wake of the replication crisis and the debate about questionable research practices, transparency and more robust research designs are becoming increasingly important in consumer research (e.g., Inman et al. 2018). The use of small, underpowered samples is viewed critically because of the propensity for false positive findings (i.e., Type I errors) that are unlikely to replicate (Nelson, Simmons, and Simonsohn 2018). Because of the volume of data available online, concerns about insufficient statistical power are muted in studies using data scraped from the web. In fact, several published articles in consumer research have used web scraping to create datasets with more than 100,000 cases (e.g., Huang et al. 2016; Yin, Bond, and Zhang 2017). Large sample sizes enable the detection of effects and more granular tests of theoretical models (Wenzel and Van Quaquebeke 2018).

PROCURING INTERNET DATA FOR CONSUMER RESEARCH

Researchers must begin by determining whether they will collect data from the internet and, if so, how they will accomplish this task. Figure 3 offers consumer researchers a roadmap for these decisions. In this section of the paper, we seek to increase the understanding the process of web scraping so that consumer researchers can understand the practicalities of capturing data from the internets to generate datasets for answering consumer research questions.

When users visit a website in a browser (e.g., Chrome, Internet Explorer, Firefox, Safari), they encounter a variety of semi-structured and unstructured content including textual, numerical, and visual elements. Consider, for instance, pages of product reviews on Amazon. In order to analyze the data contained on the review pages for a consumer research question, researchers need to transform the data in the browser into a structured dataset where every row describes a particular product-review pair and the columns capture all desired information (e.g., review rating, review age, number of helpful votes). Given the vast amounts of unstructured data contained on websites, manually visiting every single page and copying and pasting the data into a format more amenable for data analysis is difficult, time-consuming, and inefficient. In addition, manually extracting data from websites is prone to errors, difficult to document, potentially nonreplicable, and simply infeasible for more complex projects.

Instead of trying to extract data from websites manually, researchers should employ automated processes. We discuss different ways to harvest web data before discussing the process of scraping novel datasets for consumer research. We subsequently briefly review alternative ways to obtaining such data without writing code.

Using preexisting public or proprietary web datasets

Collecting data from the web does not necessarily require coding and web scraping skills. Some firms and online platforms offer rich datasets based on web data for direct download. For instance, Yelp makes subsets of its data available in an academic dataset. McGraw et al. (2015) used this dataset in their study of the prevalence of humor in reviews. Platforms such as Kaggle.com host many firm- and user-contributed datasets (e.g., from the non-profit crowdfunding platform DonorsChoose.org). Moreover, researchers in other disciplines—particularly computer scientists—provide access to large web-scraped datasets for reuse in research projects. For example, Watson, Ghosh, and Trusov (2018) use a large, publicly available dataset of Amazon reviews (McAuley 2018). Appendix 1 provides an overview of downloadable datasets that can be used in lieu of scraping a novel dataset.

Another route to data is collaboration with the firm that operates the website. It may be possible to obtain relevant data from the web by working directly with managers or firms (e.g., Kiva.org; Galak, Small, and Stephen 2011) or through institutionalized access (e.g., Marketing Science Institute).

Creating novel datasets from websites

The number of websites that offer such direct downloads is small. Instead, many website operators provide other systematic, but more opaque ways of connecting to their databases. Application programming interfaces (APIs) are a common form of such connections that allow researchers to retrieve information directly from the databases that populate the actual website with content (e.g., tweets, reviews, comments). Researchers can send requests to an API to retrieve desired pieces of information that can be parsed into a typical spreadsheet-style dataset

format (Chen and Wojcik 2016). Many websites provide documented APIs that outline the specific input parameters to access certain subsets of their data (Braun, Kuljanin, and DeShon 2018). Typically, APIs (e.g., Twitter, New York Times) require users to register as a developer. These developer credentials need to be provided when accessing the API and are used to regulate API usage (e.g., the number of calls that can be made per day). Researchers should use APIs, if available, either directly or via corresponding packages such as R, because of their ease of access and the consistency of the output they provide. Braun et al. (2018) offer an overview of relevant APIs and corresponding R packages.

APIs, however, also have several potential limitations and constraints. Many websites operate APIs for commercial reasons and offer various subscription levels for APIs. For instance, Twitter offers standard/free, premium, and enterprise APIs. The data that is returned from APIs, therefore, may be a function of the business logic of the operating company—and not random. A website's APIs may also be too restrictive for a researcher's purpose, as it may limit the number of requests or the type of data that can be requested excessively (Mitchell 2015). Moreover, APIs usually make it difficult or impossible to collect historical time series data since API calls often provide only very recent data (Bright 2017). Finally, many websites simply do not offer APIs.

Fortunately, there are ways to collect data from a website in a systematic manner even in the absence of APIs. In the following, we explain how web scrapers are designed. Web scraping typically involves three steps: mapping out the website's infrastructure, identifying data to extract, and writing the scraper.

Mapping out the website infrastructure

The systematic extraction of data from websites even without APIs is possible because most websites are based on some form of systematic code (e.g., HyperText Markup Language [HTML], Cascading Style Sheets [CSS]) that defines and organizes website elements. Whereas the browser interprets a website's underlying code for a regular internet user, a researcher seeking to scrape a website exploits the patterns of tags to locate information on the page(s) and subsequently parse the document to develop queries that will allow for the automatic extraction of the desired data (e.g., review text, review rating).

Webpages can be understood as a list of elements nested within each other in a tree-like structure. Many creators of webpages assign specific names to different elements of the tree in order to maximize consistency across different pages on a domain. These names describe a family of nodes of the larger tree that contains similar information. For most research projects, researchers will extract information from more than one webpage (e.g., review pages of all brands on the companion website). It is therefore important to understand how a website organizes its content. Consider, for instance, Skytrax (www.airlinequality.com), an online review portal for the aviation industry. This website indexes the review pages per brand within the website address (URL) after the page identifier (i.e., https://www.airlinequality.com/airline-reviews/delta-air-lines/**page/2**/?sortby=post_date%3ADesc&pagesize=100). The navigation pane at the bottom of each review page displays the total number of review pages. Thus, a researcher can integrate these pieces of information to devise an automated approach for extracting the desired pieces of information from all relevant subpages (i.e., 18 review pages in the case of Delta Airlines) into a spreadsheet format that can be analyzed to answer consumer research questions. To minimize the burden on websites, researchers should carefully examine the display

and sorting options on the website. In the Skytrax example, it would be sensible to display 100 reviews per page as doing so reduces the total number of pages to be scraped from 181 pages to 19 pages. Display options that are likely time-invariant (e.g., sorted by date) should be selected instead of options based on opaque algorithms (e.g., "recommended reviews").

Identifying the desired pieces to extract from a website

This logic of the tree-like structure also applies for extracting specific pieces of information from websites. As the primary purpose of websites is unrelated to facilitating the extraction of data for research (Xu et al. 2019), researchers need to inspect the underlying code of a website carefully to identify the elements and patterns in references to different objects on the page (e.g., metadata about tables, images, and ratings).

One useful tool to identify such website elements and patterns is the *SelectorGadget* (selectorgadget.com), which is an open source tool that facilitates identification of the key CSS selectors needed to extract desired pieces of data from a website. Website creators use CSS selectors to assign styles (e.g., font, font size) to specific (HTML) elements (e.g., a customer review) on a website. *SelectorGadget* has an interactive point-and-click interface within Google Chrome that allows researchers to select and deselect different website elements (e.g., texts, boxes) to arrive at the ideal, most precise CSS selectors to extract desired pieces of information from the website.

For instance, on the Skytrax website previously mentioned, the selector ".text_content" contains the review text, whereas the selector "text_header" captures the review title. Consumer researchers seeking to extract data from a website automatically leverage such patterns and

structures to recover the information that is contained within specific nodes of interest (e.g., all review texts contained in the ".text_content" on all pages on the domain).

When scraping multiple pages, it is critical to examine the extent to which the content on the initial webpages (used to devise the basic extraction code) is representative of all targeted webpages. Specifically, it is important to identify potential anomalies (e.g., updated reviews) that may interfere with the parsing of the data. It is helpful, therefore, to sample and inspect the HTML code of several targeted pages and write a scraper that captures all potentially relevant website elements (and records values as missing when applicable).

Writing scraping using a programming language

Web scraping often requires skill in multiple programming language and software environments. Languages such as R—especially when used in conjunction with a powerful integrated development environment such as RStudio—has syntax highlighting, data, graphic display, and workspace management capabilities. R has the *rvest* package (Wickham 2019), which makes it easy to scrape (or harvest) data from HTML web pages. In general, the process of scraping in R consists of reading the HTML of a website (i.e., the *read_html* function) and then extracting the desired pieces of information from the website using various commands (e.g., *html_node, html_nodes, html_text*) to extract the desired pieces of information and parse them into a data frame. A researcher can use such functions within R to extract the data and then convert the unstructured data into a user-friendly format such as a comma-separated values (CSV) or text file.

RECOMMENDATIONS FOR AUTHORS AND REVIEWERS

Despite the increasing popularity of web scraping, there is a lack of consensus about how it should be conducted and evaluated in scholarly consumer research. Establishing guidelines for web scraping-based consumer research is important because data scraped from the web is inherently more unstructured, messy, and complex than other forms of data. The processes by which internet data is generated are also much more opaque than those for other forms of data (e.g., experiments, surveys) because the extraction of a website's database for conversion into clean datasets for analysis by consumer researchers is rarely a chief goal for website operators (Xu et al. 2019).

In the following, we focus on the adoption of a quantitative approach for consumer research, as others have provided extensive discussions of the most relevant interpretative methods elsewhere (e.g., Kozinets 2002; 2015). Drawing on methodological advances in adjacent fields and best practices in consumer research, we propose a standardized workflow for the use of web scraping in consumer research. Taking inspiration from the framework of LeBel et al. (2018), we discuss the four interdependent facets necessary for generating credible scientific findings from web scraping-based research: (1) design transparency, (2) analytic reproducibility, (3) analytic robustness, and (4) effect replicability and generalizability.

In the case of web scraping, *design transparency* consists of outlining the key decisions regarding privacy and the sampling frame that led to the generation of the dataset. *Analytical reproducibility* is achieved when authors provide sufficient documentation that enables others to reproduce the exact numerical results reported in an article. Findings are considered *analytically robust* if they emerge consistently across other plausible and relevant data-processing and data-

analytic decisions applied to the same dataset. Finally, credible findings are characterized by *effect replicability*—i.e., the ability of an effect to be consistently observed in new samples, at a magnitude similar to that originally reported—when methodologies and conditions similar to those of the original study are used. The most credible findings are *generalizable*, i.e., they even emerge in replications that are highly dissimilar methodologically. Below, we present a streamlined workflow for generating credible findings using web scraping. For each step, we first discuss the implications for authors and then the corresponding and additional implications for reviewers.

Design transparency

In order to produce credible findings, authors should report the design details and data-analytic choices they made to ensure transparency. There are two key areas with respect to design transparency in web scraping: (1) privacy issues and (2) sampling frame.

Privacy issues

Implications for authors. Scraping data from the internet creates a unique set of privacy, ethical, and legal considerations. While the data on most websites is public, few users are fully aware of the information they generate via their digital footprints (Hinds and Joinson 2019). Many users may even erroneously believe that their data is private (Landers et al. 2016)—especially on websites (e.g., discussion forums, dating websites) where barriers such as user registration increase users' subjective sense of privacy. Contrary to these perceptions, however, such barriers can be circumvented, and even very limited information or direct textual quotes can be sufficient to identify users. As web scraping does not necessitate explicit consent from users, researchers might capture sensitive consumer data without consumer knowledge (Martin and

Murphy 2017). Therefore, researchers should take steps to ensure the privacy of users, particularly when data is being shared or consumers are quoted directly in consumer research articles. Researchers should avoid publishing identifiable information at all costs, and they should anonymize the data as much as possible (e.g., redacting Twitter handles or user names; Meyer 2018).

The legality of web scraping is another ongoing debate. There is no clear consensus about whether scraping data for research purposes is permissible under the current American and international intellectual and cybersecurity laws. While a detailed discussion of the legality of web scraping is challenging even for specialized lawyers and beyond the scope of this report, we provide a synopsis of the main potential causes of liability arising from scraping in appendix 2. These include copyright infringement, trespass to chattels, breach of contract, and violation of the Computer Fraud and Abuse Act (see also Landers et al. 2016; Simchi-Levi 2019). In addition, researchers need to be aware of site-specific (e.g., terms of use) and regional (e.g., the European Union's General Data Protection Regulation) legal frameworks that govern the legality of web scraping. Thus far, as highlighted by Edelman (2012), the scraping of data for research purposes has not been targeted for lawsuits by data providers.

Implications for reviewers. Reviewers should be mindful of the legal and privacy concerns involved in web-scraped data. While it may be necessary to redact the names of the source websites in the final publication (e.g., Datta et al. 2018), editors should encourage authors to disclose the source website(s) during the review process (given its higher confidentiality) to allow for a comprehensive assessment of potential privacy concerns and the study's merit.

Sampling frame

Implications for authors. The credibility of findings derived from web-scraped data is determined by a clear and compelling rationale for selecting specific slices of data from the web. Specifically, researchers face three crucial sampling decisions, which we discuss in the sequential order of the research process: selection of the (1) website, (2) sample, and (3) variables. Without sufficient details about the sampling frame, it is difficult for reviewers to evaluate the merits of a study, and it is often impossible to conduct independent replications.

Website selection. First, authors should justify the selection and sampling of the specific website. One powerful argument to support the choice of a particular website is the presence of *idiosyncratic features* on the target website (vs. other comparable websites) that allow for the creation of variables that are critical for effectively testing predictions (e.g., attribute-level quality perceptions; Henkel et al. 2018; funny votes; McGraw et al. 2015). Researchers may also illuminate the reasons that a research question necessitates a particular *type* of website (e.g., a discussion board such topix.com; Chen and Berger 2013). For other projects, however, researchers may be agnostic in their selection of websites. In this case, scraping data from multiple, diverse websites is a useful strategy to bolster the generalizability of an effect (e.g., Melumad, Inman, and Pham 2019; Ordenes et al. 2019).

Within-website sampling. The second facet of the sampling frame in web scraping pertains to the selection of one or multiple subsets of data from a target website. The large volume of data available on websites along with the fact that they were not generated for any specific research goal makes it necessary to filter the data before testing a specific research question (Xu et al. 2019). Thus, consumer researchers will usually select a relatively small

sample from a much larger population of data points (e.g., reviews, brands), which creates a risk of biased sampling (Aguinis, Cascio, and Ramani 2017).

One useful strategy to justify the use of a particular sample from a website is *conducting pretests* to demonstrate that particular products, brands, or industries closely align with the focal constructs of interest. Researchers may purposefully sample specific brands (e.g., based on their positioning; Henkel et al. 2018) or industries that are characterized by relevant psychological properties (e.g., strong versus weak ties; Umashankar, Ward, and Dahl 2017). A related approach is establishing *why specific instances map upon the variables of theoretical interest*. For example, Paharia et al. (2014) used Peet's Coffee as an example of an underdog brand with a strong rival (i.e., Starbucks) and employed the natural variation in the distance of a Peet's store from a Starbucks as a proxy for competitive salience.

After linking instances (e.g., product, brand) to the variable(s) of theoretical interest, it is good practice to subsequently *collect all available data points* from the website (e.g., reviews, auctions). If feasible, this practice reduces concerns about potential sampling biases and allows authors to preempt requests for additional data collection. Smith, Newman, and Dhar (2016), for instance, scraped all available eBay auctions of the Beatles' White Album in their study of consumers' preference for items with earlier (vs. later) serial numbers and gave a cogent rationale for why this particular Beatles' record is a good fit for testing their predictions (i.e., constant product quality, ability to test potential alternative explanations).

Variable selection. Finally, researchers need to select the elements they will use and scrape from a website. Hitherto, there has been little consistency in the variables that were collected and included (vs. excluded) from websites in published consumer research, even when drawing from the same source (e.g., Yelp, TripAdvisor). In the absence of consensus about

covariates and control variables, it is very difficult—if not impossible—to compare results from different studies. Therefore, *providing a clear rationale for the inclusion of control variables and a description of their operationalization* is paramount (for an example see Van Laer et al. 2019). Becker et al. (2016) provide a useful and concise list of ten recommendations (e.g., when in doubt leave them out; when feasible, include covariates in hypotheses and models) to guide the selection of covariates or control variables (see also Bernerth and Aguinis 2016, p. 273-280).

Another technique specific to web scraping that authors should employ to increase transparency about the variable selection and generation process is the inclusion of *a screenshot from the website's interface*. On this screenshot, authors can point out which website elements were parsed into which variables (e.g., Chen and Lurie 2013, p. 474). Employing screenshots is particularly helpful when authors procure data from novel, unfamiliar websites. Even for frequently used websites (e.g., Yelp, TripAdvisor), however, screenshots enable the review team to make an informed assessment of the value, comprehensiveness, and robustness of authors' work (e.g., Grewal and Stephen 2019).

Collecting all relevant covariates or control variables while scraping a website also allows researchers to at least partially tackle the issue of endogeneity (Kennedy 1998). In the context of ordinary least squares (OLS) estimation where Y is regressed on X, concerns about endogeneity arise when the X variable and the error term in the OLS model are correlated. In this situation, the coefficient estimate of the compromised explanatory variable X also contains the effect of unaccounted variable(s), which partially explains the dependent variable (Chintagunta et al. 2006). By collecting all relevant potential covariates, researchers can reduce endogeneity resulting from omitted variable bias.

Implications for reviewers. Credible findings of any (non-experimental) research study depend on clearly explained and justified strategies regarding the chosen research site (i.e., website in this case) and the sampling frame. This dependence is particularly pronounced for web-scraped data because of the opacity of the processes that generated this data (Xu et al. 2019). If authors do not provide sufficient detail, reviewers should request explicit statements about the criteria that guided the sampling frame at the website, within-website, and variable levels. Benchmarking the article under review against relevant work (e.g., studies examining the same dependent variables or source website) facilitates the identification of variables that have been omitted or other issues in the sampling frame.

In addition, reviewers should scrutinize the *sample size* of studies using web-scraped data. At present, there is significant variation in sample sizes in consumer research using data from the web—even in studies employing very similar dependent variables. For example, sample sizes in studies using review-related dependent variables (e.g., star ratings, helpful votes) ranged from 147 to 418,415. Sample sizes should be determined as a function of (expected) effect sizes as well as the complexity of the model tested (e.g., main vs. interaction effects, number of control variables; Cohen 1988). Given the vast amount of data available online, (very) small sample sizes are inappropriate unless a study serves purely illustrative purposes.

Analytic reproducibility

The credibility of the insights generated from web-scraped data depends on the analytic reproducibility of the results—"whether the original numeric results can be exactly reproduced given the same data and analysis code or process" (Epskamp 2019, p. 145) used by the original authors. In light of the opaque data generation process in this type of data and the legal issues

around sharing datasets generated via scraping, perfect reproducibility may remain elusive. Yet, there are several steps authors should take to ensure the integrity of the data and increase the reproducibility of their findings.

Implications for authors. The unstructured nature of the data collected via web scraping places extensive demands on researchers to *carefully inspect and clean the data* to ensure its veracity. In contrast to data gathered via conventional methods, data scraped from the web requires substantially more time and effort for cleaning and preparation before analysis (Balducci and Marinova 2018). Multiple steps are required to ensure the integrity of data scraped from the web. Authors should describe the features of the originally scraped dataset (e.g., number of cases) as well as the data cleaning efforts undertaken and the number of cases removed (e.g., duplicates, empty reviews).

The prevalence of *inauthentic and fake data* is another important threat to the integrity of web-scraped data. There is a growing literature in information systems examining the verbal (e.g., length, self-references) and non-verbal features (e.g., friend count, review count) that distinguish authentic from fake posts (for an overview see Zhang et al. 2016). Researchers need to carefully consider the possibility that fake data is randomly distributed across the data. A growing body of research suggests that certain types of businesses such as independent hotels (Mayzlin, Dover, and Chevalier 2014) or brands with a weak reputation (Luca and Zervas 2016) are more prone to engage in review fraud by creating fake reviews for themselves or their competitors. There is also evidence that certain actors use bots to systematically manipulate and thereby contaminate commonly studied variables (e.g., retweets; Salge and Karahanna 2018).

In addition to reporting their data cleaning efforts, researchers also need to spell out their assumptions about the configuration and characteristics of the scraped data. Landers et al. (2016)

propose a recursive process of *formulating a "data source theory,"* including outlining these assumptions, testing, and refining the data source theory as required. A data source theory encompasses the essential assumptions that the data needs to possess in order to be able to test a prediction. Making these assumptions explicit and describing tests to validate these assumptions increase confidence in the integrity of the study's findings and increase its reproducibility. A critical element in the data source theory is spelling out the relation between the time of data capture and the focal psychological processes. Often, when scraping data, a researcher can only observe a variable at one point in time (i.e., the day of scraping). However, the scraped value (e.g., the number of a user's Yelp friends) might be very different from the value of this variable when the focal consumption behavior occurred (e.g., a review written several years ago). Thus, data source theories should ensure sufficient alignment between data capture and the focal psychological processes. In this example, for instance, the data source theory could specify the inclusion of only recent reviews (e.g., written within a week of the scraping date).

The data source theory should be accompanied by a detailed documentation of the data collection methods, data-analytic decisions, and code to perform the relevant statistical analyses. Arvidsson and Caliandro (2016) provide an example of effective documentation of dataset generation via web scraping in consumer research. Their article discusses how the search terms for their crawler of Italian tweets related to Louis Vuitton were selected. A technical appendix provides further details about specific technical aspects of the crawler (e.g., comprehensiveness and accuracy) and contains the Python code used to compile the dataset. Importantly, authors do not need to ensure that the code is elegant or perfect; it merely needs to be good enough to execute the task (Barnes 2010). It is also desirable to annotate the code to let readers understand its logic and facilitate replications and extensions. Authors who publish the (annotated) code may

also learn from others' feedback, which could inform future projects—a potential side benefit for authors. Other means of harvesting web data (e.g., outsourcing) should be explicitly disclosed in the article (e.g., Chen 2017; Elder et al. 2017). If possible, even in these instances, the code used for collection should be provided. Authors relying on tools rather than custom scrapers (e.g., Mozenda; Ordenes et al. 2017) should include a clear description of the software and settings (e.g., exported scripts, screenshots).

As collecting data from the web can be a laborious and costly process, consumer researchers may wish to *reuse* datasets or parts of datasets in different projects (Adjerid and Kelley 2018). As the web is dynamic, it is also possible that certain website elements central to a research question are no longer available on a website (e.g., conversion rates on Amazon.com; Ludwig et al. 2013). At present, there are no clear guidelines about whether and under which conditions such reuse is permissible. Therefore, we recommend that, at the very minimum, authors disclose their reuse of data from other research projects during the review process. The final manuscript should explicitly reference the relevant source(s) in the data section (such as done by Mogilner, Aaker, and Kamvar 2012). When using subsets of an original dataset, authors can bolster confidence in their findings by conducting and reporting supplementary analyses demonstrating that their findings hold in the entire dataset.

Implications for reviewers. The most effective way to assess a finding's reproducibility is to conduct the same analyses as the authors. However, this process may not always be practical or feasible. In these cases, it is critical to carefully evaluate the appropriateness of the authors' data source theory. To overcome the opaque nature of web-generated data, reviewers may consider requesting additional descriptive statistics beyond the mean and standard deviation, such as the minimum, maximum, mode, median, and percentiles for all variables, distribution of

the focal independent, mediator, moderator, and dependent variable, and the correlation matrix of all the variables. Additional descriptive statistics (e.g., proportion of the cuisine type in a sample of restaurant reviews from Yelp) that are not central to the research question can also offer meaningful insights for the evaluation process. A careful examination of these comprehensive descriptive statistics and correlations can offer clues to the reproducibility of an effect and pinpoint potentially confounding factors (Becker et al. 2016)

In addition to requiring complete information about data cleaning efforts, handling of fake data, and the authors' data source theory, reviewers should pay attention to an implicit assumption prevalent in research based on data harvested from the web. Namely, that the way this data was inputted by users of the website remained constant over time. As the web is non-static and constantly evolving, many websites change their interfaces over time and thus change the data generating process (Weber 2018; Xu et al. 2019). These changes have the potential to affect the consumer processes being studied. For instance, online stores might change their website design and the information that is visible to consumers when writing reviews. For large and more sophisticated web platforms, different users may have a different user experience even at a single point in time because the firm is continually A/B testing different interface elements. Thus, even data scraped from the same website might vary substantially over time or between users during the same time period (Adjerid and Kelley 2018). Although researchers (and the users themselves) may be completely unaware of this variation, reviewers should encourage authors to account for temporal changes to websites that can be observed. Failing to account for changes in website design can introduce bias, especially in datasets spanning decades. One relevant tool use for examining such changes in websites is the *Wayback Machine* (i.e., archive.org/web/), which can be used to inspect the look of websites over time (for applications

see Babić Rosario et al. 2016; Martin, Borah, and Palmatier 2017). Reviewers should examine whether design changes interact with focal variables of interest, in particular, as potential interactions may require reconsideration of the sampling frame or the inclusion of specific variables.

Analytical robustness

Robust consumer research findings are generated via appropriate methodological techniques and are not overly dependent on the specific, idiosyncratic data-analytic and method choices of the authors. Put differently, a finding can be considered robust if it consistently emerges across other plausible and relevant data-processing and data-analytic decisions in the same data (LeBel et al. 2018; Steegen et al. 2016). Below, we discuss the essential steps for generating correct and robust inferences from web-scraped data.

Implications for authors. Data scraped from the internet tends to be relatively unruly and messy. Many variables commonly captured (e.g., review age, number of friends) via web scraping have heavily skewed or heavy-tailed distributions (Bright 2017). Hence, authors should create data visualizations (e.g., histograms, box-and-whisker plots) to identify the presence of potentially heavy-tailed or highly skewed distributions (Cohen et al. 2003). Given these insights, authors should carefully consider the *transformation* that is appropriate and meaningful for a given dataset and predictions (for a recent, in-depth discussion see Becker, Robertson, and Vandenberg 2019). Authors should report the central analyses using both the original and the transformed independent or dependent variables (e.g., Chen and Berger 2013).

A related concern in the analysis of data scraped from the web is that many frequently used *dependent variables are non-normally distributed*. For instance, many of the numeric

variables captured from review platforms (e.g., helpful votes, useful votes) are count data. In this case, researchers should employ appropriate models such as Poisson, negative binomial regressions, or Poisson quasi-maximum likelihood regression rather than ordinary least squares regression (Cameron and Trivedi 1998). Just as authors have specific procedures for variable transformations, they should also determine if count data exhibits signs of overdispersion, which may be caused by the presence of excess zeros. In this case, alternative zero-inflated models tend to provide a better fit to the data (Blevins, Tsang, and Spain 2015).

A central assumption of analyses commonly used by consumer researchers (e.g., regression analysis) is that individual observations are independent from each other. Yet, this *assumption of independence* is likely to be violated in web-scraped data. There are two primary causes of nonindependence: (1) temporal sequences within one individual and (2) group effects (Kenny and Judd 1986). Temporal sequences occur when observations (e.g., reviews) are repeatedly taken from a single unit (e.g., user, brand) over time. Common group effects include, for example, social influence effects wherein the average rating of a firm by other consumers influences the weights that a consumer assigns to different aspects of the experience he or she is reviewing (Sridhar and Srinivasan 2012). The nesting of observations (e.g., reviews) within higher levels of analyses (e.g., consumers, firms) can lead to nonindependence resulting from temporal sequences and common group effects. Recent work suggests that even the very first review still influences a product's average rating three years later (Park, Shin, and Xie 2018). Failure to account for the multi-level nesting in web data can lead to erroneous inferences and invalid analyses.

Robust and correct inferences from web-scraped data require researchers to adopt a contemplative lens to endogeneity concerns (Rutz and Watson 2019). Endogeneity in web-

scraped data may not only arise from the previously discussed omitted variable bias, but also from other sources. Self-selection is one of the most common sources of endogeneity in web-scraped data. Self-selection based endogeneity is a special case of omitted variable bias (Heckman 1979) wherein the studied agents (e.g., consumers, managers) make informed choices regarding assigning themselves to mutually exclusive treatment (vs. non-treatment) groups based on unobservables that correlate with the observed predictors and the outcome variables (Clougherty, Duso, and Muck 2016). When self-selection is a concern, it can be helpful to conduct within-subjects analyses within a subset of the data that "participated" in both levels of the treatment (i.e., independent variable). For instance, in their study of the influence of device (i.e., smartphone vs. PC) on content emotionality, Melumad et al. (2019) demonstrated that their between-subjects effect of device on emotionality also emerged at a within-subject level, i.e., among those consumers who wrote reviews using both types of devices (see also Moore 2012).

Depending on context, propensity score matching is another viable approach to addressing endogeneity due to self-selection (Rutz and Watson 2019). In this approach, cases (e.g., consumers, brands) that differ with respect to the focal treatment or behavior (e.g., device use) are matched on multiple other criteria that would predict engagement in the focal behavior (e.g., Datta et al. 2018). While omitted variable bias with its special case of endogeneity due to self-selection is of particular importance in web-scraped data, researchers also should be aware of other threats to validity results from endogeneity that may be caused by other factors, such as simultaneity wherein an independent variable is potentially caused by the dependent variable (Antonakis et al. 2010). Taken together, it is pivotal that researchers consider threats to validity resulting from endogeneity. A contemplative approach implies more cautious, less causal language when describing and discussing the results from web-scraped data. At the article level,

carefully designed follow-up experimental studies that address the causes of endogeneity observed in the scraped web data can boost confidence in the robustness of an effect.

Given the multitude of design, data-analytic, and model specification decisions that influence statistical inference in web-scraped data, it is critical for authors to provide evidence of the robustness of a study's central finding(s). One helpful tool to demonstrate that a finding is not merely the result of arbitrary sampling or data processing decisions is to present a multiverse analysis (Steegen et al. 2016). In contrast to a single analysis or a few selected analyses, a multiverse analysis visually displays the statistical significance of the focal effect derived from numerous estimations, capturing the most reasonable options for processing the data (e.g., exclusions, transformations, and coding). Here, authors may consider highlighting certain specific results (e.g., subsets of data) in the manuscript that attest to the robustness of the findings. Authors can also consider supplementing this data multiverse with a model multiverse that captures the degree to which the central conclusions are robust across alternative models or specifications (Patel, Burford, and Ioannidis 2015).

Implications for reviewers. When evaluating a study based on scraped data, reviewers need to carefully examine the descriptive statistics, correlations of the variables parsed from websites, model-free evidence, and data visualizations, such as histograms. This examination enables the identification of any undue influences from the authors' data-analytic choices and the appropriateness of the model specifications. Inspecting the actual website(s) underlying the dataset can help spot potential confounds and methodological challenges that need explicit consideration (e.g., non-independence of observations, omitted variables). It is also helpful to ask authors to report the share of cases that constitute explicit violations of core assumptions (e.g., non-independence of observations) and the results of relevant robustness checks (e.g., results in

meaningful subsets of the data). The inspection of the website can also be helpful to determine the extent to which endogeneity is a concern (e.g., via self-selection).

Reviewers may be particularly concerned about the robustness of a finding when a manuscript suggests extensive HARKing (i.e., hypothesizing after the results are known; Kerr 1998) or other questionable practices, such as biased selection, selective reporting, and systematic capitalization on chance (Adjerid and Kelley 2018; Aguinis et al. 2017). In these situations, assuming the article otherwise merits consideration, reviewers may consider proposing that the authors replicate the study's findings in a preexisting dataset (see appendix 1) during the revision process. Moreover, reviewers and editors may prompt authors to preregister this replication (Munafò et al. 2017; van 't Veer and Giner-Sorolla 2016) or devise and share analysis plans (Nuijten 2017). Combatting questionable research practices should, however, not come at the expense of exploratory analyses, which constitute useful pathways for knowledge creation—especially in large datasets (Lynch et al. 2012). It is critical, however, that exploratory analyses are characterized and reported as such in the final published articles rather than as strong confirmatory theory tests.

Effect replicability and generalizability

An effect is considered replicable if it is observed consistently in new samples with magnitudes that are similar to the original effect. A replicable effect is generalizable if it emerges across distinct methodologies (i.e., different operationalizations of the independent and dependent variables) and distinct populations (LeBel et al. 2018). Therefore, concerns about replicability and generalizability are particularly important in multi-study, multi-method consumer research.

Implications for authors. First and foremost, authors must demonstrate that the data that is generated in websites (e.g., likes, helpful votes, shares) is indeed a valid measurement of the underlying theoretically meaningful constructs. Particularly pertinent to web scraping is the implicit trade-off between two desirable facets of research—sample sizes and construct validity (Xu et al. 2019). Due to feasibility constraints, especially in large datasets with many individual data points (e.g., reviews), automated data processing algorithms (e.g., dictionaries, natural language processing) are likely to replace manual human coding efforts. However, recent evidence in the context of text processing suggests that automated approaches are likely to produce significantly more errors than human coders (Hartmann et al. 2019), thereby threatening the internal validity of such findings. Construct validity is essential to ensure that insights from consumer research studies based on web scraping can be replicated in other less methodologically similar studies (e.g., experiments). To deal with this challenge, researchers can consider selecting smaller, more manageable subsets of the data to demonstrate convergence between automated and human coding (e.g., Berger and Milkman 2012).

Specifying the contextual elements that characterize authors' findings (Simons, Shoda, and Lindsay 2017) is a second important consideration. While generalizability is the gold standard for credible research findings, it is unreasonable to expect every single finding to be generalizable. Online samples are inherently biased and capture only certain subsets of the consumer population at large (Landers and Behrend 2015; Wenzel and Van Quaquebeke 2018). While such samples are ideally suited for research questions focusing on the online domain (e.g., the emergence of a brand public; Arvidsson and Caliandro 2016), the metaphor of online data as reflecting naturalistic behavior may be easily stretched in other areas of inquiries. As consumer behavior on the internet is often heavily influenced by contextual factors such as the website's

data entry procedures (e.g., Lewis 2015), authors need to explicitly outline how specific features of the website(s) influence and potentially constrain the generalizability of the effect.

Implications for reviewers. The abundance and accessibility of data on the internet increase the possibility of discovering interesting consumer phenomena and processes. However, focusing on the discovery of interesting relationships can come at the expense of construct validity (Tonidandel, King, and Cortina 2018). In particular, in multi-study articles using different methods (e.g., experiments and web-scraped data), reviewers need to consider the sufficiency of the alignment and consistency in construct operationalization between studies. In many cases, it will also be beneficial to ask authors to conduct close replications of the findings from the study using web scraping (LeBel et al. 2018). Such studies can establish the replicability of an effect by operationalizing the independent or dependent variable in the exact same way it was scraped from the web, while changing the other variable (e.g., experimentally manipulating the independent variable while measuring the dependent variable as it occurs on the target website).

The differentiation between significant and impactful effects is another important aspect of gauging the replicability and generalizability of an effect discovered in consumer research based on web scraping. As datasets generated via web scraping typically have very large sample sizes, many of the examined relationships will be significant at the standard conventional *p*-values such as .05, .01, or even .001. Therefore, merely applying commonly used frequentist methods to such large datasets can be problematic (e.g., Wenzel and Van Quaquebeke 2018). In order to set realistic expectations about effect replicability and generalizability, reviewers should not only require authors to report effect sizes, but also establish the meaningfulness of their findings in light of the focal context rather than merely demonstrating statistical significance

(Adjerid and Kelley 2018; Lin, Lucas, and Shmueli 2013). Reviewers can ask for what-if analyses (e.g., the effect of a 10 percent change in the independent variable on a critical dependent measure) that illuminate the extent to which seemingly small changes can have meaningful substantive effects for relevant constituencies (e.g., firms, consumers). Explicit discussions about the impact of a finding are also informative for examinations of its potential for replicability and generalizability because of their emphasis on effect sizes (and corresponding sample size requirements) and their consideration of the influence of potential contingencies (e.g., expected moderators).

CONCLUSION

Consumers and managers create rich and diverse digital footprints that capture their behavior. In this report, we discuss methods of transforming such data into impactful datasets for answering consumer research questions. We begin by demystifying the process of harvesting data from the internet via web scraping. While an understanding of the basic mechanics of web scraping is a necessary condition for leveraging and realizing the full potential of data on the internet for consumer research, it is not sufficient. We also outline a structured workflow for generating credible consumer research findings via web scraping that entails four key facets: (1) design transparency, (2) analytic reproducibility, (3) analytic robustness, and (4) effect replicability and generalizability. Our approach enables more uniform comparisons across studies and includes clear, easy-to-use guidelines for conducting and evaluating consumer research using web scraping.

Web scraping can accelerate consumer research by reducing the cost and time required for data collection. These benefits are particularly relevant for junior career scholars (Edelman 2012). Hence, integrating introductory courses on web scraping into the curricula of marketing and consumer behavior Ph.D. programs is important. Understanding the complexities in the design and execution of web scraping studies can also level the playing field for consumer researchers at institutions with limited resources (Barnes et al. 2018).

In addition to collecting data by oneself using custom code , there are other ways of constructing novel web-based datasets for consumer research beyond designing custom scrapers. Consumer researchers might also consider using tools (e.g., Mozenda, import.io, Instant Data Scraper) or outsourcing web scraping on crowdsourcing platforms (e.g., Fiverr, Upwork). While these approaches may be sufficient for answering certain research questions, they generally offer less control, flexibility, and scalability. There are also significant benefits to collaborating with quantitative marketing colleagues or computer scientists for more complex scraping projects (e.g., longitudinal, multi-site scraping; Rafaeli, Ashtar, and Altman 2019).

The main goal of our structured workflow for using web scraping in consumer research is the generation of credible, replicable, and generalizable findings. A positive side benefit of this approach for authors is that it will likely inspire other researchers to leverage the abundant data on the internet for conducting rapid, inexpensive, and high-quality replications of their work. Data available on the internet can also play a vital role in assessing the external validity of important consumer research findings that have been demonstrated primarily in controlled laboratory experiments. Relatedly, researchers can easily scrape datasets featuring the digital footprints of diverse populations of consumers around the globe (Kosinski et al. 2016). Greater

sample diversity increases the confidence in the generalizability of consumer research findings (Maner 2016; Rad, Martingano, and Ginges 2018).

The potential of web scraping to generate interesting consumer research findings is not limited to the pathways discussed in this report. Many of the challenges involved in wrangling web-scraped data are not simple nuisances, but instead opportunities in themselves for asking and answering different consumer research questions. For instance, researchers can exploit the multi-level nature of web data to examine interesting within-consumer effects (e.g., Moore 2012; Ordenes et al. 2017) or interactions between consumers and firms (Wang and Chaudhry 2018). Another pathway for constructing even more informative datasets is to construct datasets integrating multiple online and offline sources (e.g., Berger et al. 2010; Datta et al. 2018).

The internet is a potential goldmine for consumer research. Exploiting this resource, however, requires mastery of novel technical skills that are unfamiliar territory for many consumer researchers, as well as an appreciation for important design and methodological challenges in harvesting and analyzing data from the internet. The structured workflow outlined in this report is a promising pathway for generating interesting, impactful, and credible consumer research findings.

REFERENCES

Ebay, Inc. V. Bidder's Edge, Inc, 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

Ticketmaster Corp. V. Tickets. Com, Inc, 2003 WL 21406289 (US Dist. 2003).

Register.Com, Inc. V. Verio, Inc, 356 F. 3d 393 (Court of Appeals, 2nd Circuit 2004).

Qvc, Inc. V. Resultly, Llc, 159 F. Supp. 3d 576 (Dist. Court, ED Pennsylvania 2016).

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

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

Ali, Meiryum (2019), "The Best 25 Datasets for Natural Language Processing," https://web.archive.org/web/20190812134029/https://lionbridge.ai/datasets/the-best-25-datasets-for-natural-language-processing/.

Antonakis, John, Samuel Bendahan, Philippe Jacquart, and Rafael Lalive (2010), "On Making Causal Claims: A Review and Recommendations," *The Leadership Quarterly*, 21 (6), 1086-120.

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

Babić Rosario, Ana, Francesca Sotgiu, Kristine De Valck, and Tammo H. A. Bijmolt (2016), "The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors," *Journal of Marketing Research*, 53 (3), 297-318.

Bagchi, Rajesh and Amar Cheema (2013), "The Effect of Red Background Color on Willingness-to-Pay: The Moderating Role of Selling Mechanism," *Journal of Consumer Research*, 39 (5), 947-60.

Balducci, Bitty and Detelina Marinova (2018), "Unstructured Data in Marketing," *Journal of the Academy of Marketing Science*, 46 (4), 557-90.

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

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

Becker, Thomas E., Guclu Atinc, James A. Breaugh, Kevin D. Carlson, Jeffrey R. Edwards, and Paul E. Spector (2016), "Statistical Control in Correlational Studies: 10 Essential

Recommendations for Organizational Researchers," *Journal of Organizational Behavior*, 37 (2), 157-67.

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

Bellezza, Silvia and Jonah Berger (2019), "Trickle-Round Signals: When Low Status Is Mixed with High," *Journal of Consumer Research*, forthcoming.

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

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.

Berger, Jonah and Katherine L. Milkman (2012), "What Makes Online Content Viral?," *Journal of Marketing Research*, 49 (2), 192-205.

Berger, Jonah, Alan T. Sorensen, and Scott J. Rasmussen (2010), "Positive Effects of Negative Publicity: When Negative Reviews Increase Sales," *Marketing Science*, 29 (5), 815-27.

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

Blevins, Dane P., Eric W. K. Tsang, and Seth M. Spain (2015), "Count-Based Research in Management: Suggestions for Improvement," *Organizational Research Methods*, 18 (1), 47-69.

Borah, Abhishek and Gerard J. Tellis (2016), "Halo (Spillover) Effects in Social Media: Do Product Recalls of One Brand Hurt or Help Rival Brands?," *Journal of Marketing Research*, 53 (2), 143-60.

Brady, William J., Julian A. Wills, Dominic Burkart, John T. Jost, and Jay J. Van Bavel (2019), "An Ideological Asymmetry in the Diffusion of Moralized Content on Social Media among Political Leaders," *Journal of Experimental Psychology: General*, 148 (10), 1802-13.

Braun, Michael T., Goran Kuljanin, and Richard P. DeShon (2018), "Special Considerations for the Acquisition and Wrangling of Big Data," *Organizational Research Methods*, 21 (3), 633-59.

Brick, Cameron, Laura Botzet, Cory K. Costello, Anatolia Batruch, Ruben C. Arslan, Melissa Kline, Nicolas Sommet, James Green, Michèle B. Nuijten, Mark Alexander Conley, Thomas Richardson, Nicole Sorhagen, Anton Olsson Collentine, Gilad Feldman,

Franklin Feingold, and Harry Manley (2019), "Directory of Free, Open Psychological Datasets," *OSF, doi:10.17605/OSF.IO/TH8EW.*

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

Cameron, A. Colin and Pravin K. Trivedi (1998), *Regression Analysis of Count Data*, New York: Cambridge University Press.

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

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

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

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

Chintagunta, Pradeep, Tülin Erdem, Peter E. Rossi, and Michel Wedel (2006), "Structural Modeling in Marketing: Review and Assessment," *Marketing Science*, 25 (6), 604-16.

Clougherty, Joseph A., Tomaso Duso, and Johannes Muck (2016), "Correcting for Self-Selection Based Endogeneity in Management Research: Review, Recommendations and Simulations," *Organizational Research Methods*, 19 (2), 286-347.

Cohen, Jacob (1988), *Statistical Power Analysis for the Behaviors Science*, Hillsdale, New Jersey: Laurence Erlbaum Associates.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Mahwah, NJ: Lawrence Erlbaum Associates.

Dai, Hengchen, Cindy Chan, and Cassie Mogilner (2019), "People Rely Less on Consumer Reviews for Experiential Than Material Purchases," *Journal of Consumer Research*, forthcoming.

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.

Dolbec, Pierre-Yann and Eileen Fischer (2015), "Refashioning a Field? Connected Consumers and Institutional Dynamics in Markets," *Journal of Consumer Research*, 41 (6), 1447-68.

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

Dreyer, Anthony J. and Jamie Stockton (2013), "Internet 'Data Scraping': A Primer for Counseling Clients," *New York Law Journal*.

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

Elder, Ryan S., Ann E. Schlosser, Morgan Poor, and Lidan Xu (2017), "So Close I Can Almost Sense It: The Interplay between Sensory Imagery and Psychological Distance," *Journal of Consumer Research*, 44 (4), 877-94.

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

Galak, Jeff, Deborah Small, and Andrew T. Stephen (2011), "Microfinance Decision Making: A Field Study of Prosocial Lending," *Journal of Marketing Research*, 48, S130-S37.

Granville, Vincent (2016), "A Plethora of Data Set Repositories," https://web.archive.org/web/20190812135737/https://www.datasciencecentral.com/profiles/blogs/a-plethora-of-data-set-repositories.

Grewal, Lauren and Andrew T. Stephen (2019), "In Mobile We Trust: The Effects of Mobile Versus Nonmobile Reviews on Consumer Purchase Intentions," *Journal of Marketing Research*, 56 (5), 791-808.

Grijalva, Emily, Timothy D. Maynes, Katie L. Badura, and Steven W. Whiting (2019), "Examining the "I" in Team: A Longitudinal Investigation of the Influence of Team Narcissism Composition on Team Outcomes in the NBA," *Academy of Management Journal*, forthcoming.

Groves, Robert M. (2011), "Three Eras of Survey Research," *Public Opinion Quarterly*, 75 (5), 861-71.

Hartmann, Jochen, Juliana Huppertz, Christina Schamp, and Mark Heitmann (2019), "Comparing Automated Text Classification Methods," *International Journal of Research in Marketing*, 36 (1), 20-38.

Heckman, James J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1), 153-61.

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

Hewett, Kelly, William Rand, Roland T. Rust, and Harald J. van Heerde (2016), "Brand Buzz in the Echoverse," *Journal of Marketing*, 80 (3), 1-24.

Hinds, Joanne and Adam Joinson (2019), "Human and Computer Personality Prediction from Digital Footprints," *Current Directions in Psychological Science*, 28 (2), 204-11.

Hirschey, Jeffrey Kenneth (2014), "Symbiotic Relationships: Pragmatic Acceptance of Data Scraping," *Berkeley Technology Law Journal*, 29 (1), 897-927.

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

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.

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.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274-306.

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.

Isaac, Mathew S. and Robert M. Schindler (2014), "The Top-Ten Effect: Consumers' Subjective Categorization of Ranked Lists," *Journal of Consumer Research*, 40 (6), 1181-202.

Jung, Kiju, Ellen Garbarino, Donnel A. Briley, and Jesse Wynhausen (2017), "Blue and Red Voices: Effects of Political Ideology on Consumers' Complaining and Disputing Behavior," *Journal of Consumer Research*, 44 (3), 477-99.

Kennedy, Peter (1998), *A Guide to Econometrics*, Cambridge, Massachusetts: MIT Press.

Kenny, David A. and Charles M. Judd (1986), "Consequences of Violating the Independence Assumption in Analysis of Variance," *Psychological Bulletin*, 99 (3), 422-31.

Kerins, Ian (2018), "Gdpr Compliance for Web Scrapers: The Step-by-Step Guide," https://web.archive.org/web/20181116011314/https://blog.scrapinghub.com/web-scraping-gdpr-compliance-guide.

Kerr, Norbert L. (1998), "HARKing: Hypothesizing after the Results Are Known," *Personality and Social Psychology Review*, 2 (3), 196-217.

Kim, Baek Jung, Masakazu Ishihara, and Vishal Singh (2018), "Peer Effects in Adoption and Usage of Crowdfunding Platforms: Evidence from United States Public School Teachers," University of Britsh Columbia, working paper.

Klostermann, Jan, Anja Plumeyer, Daniel Böger, and Reinhold Decker (2018), "Extracting Brand Information from Social Networks: Integrating Image, Text, and Social Tagging Data," *International Journal of Research in Marketing*, 35 (4), 538-56.

Kosinski, Michal, Yilun Wang, Himabindu Lakkaraju, and Jure Leskovec (2016), "Mining Big Data to Extract Patterns and Predict Real-Life Outcomes," *Psychological Methods*, 21 (4), 493-506.

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.

_____ (2015), *Netnography: Redefined*, London, UK: Sage.

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

Landers, Richard N. and Tara S. Behrend (2015), "An Inconvenient Truth: Arbitrary Distinctions between Organizational, Mechanical Turk, and Other Convenience Samples," *Industrial and Organizational Psychology*, 8 (2), 142-64.

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.

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

Lewis, Kevin (2015), "Three Fallacies of Digital Footprints," *Big Data & Society*, 2 (2), 1-4.

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.

Lin, Mingfeng, Henry C. Lucas, and Galit Shmueli (2013), "Too Big to Fail: Large Samples and the P-Value Problem," *Information Systems Research*, 24 (4), 906-17.

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

Ludwig, Stephan, Ko de Ruyter, Mike Friedman, Elisabeth C. Brüggen, Martin Wetzels, and Gerard Pfann (2013), "More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates," *Journal of Marketing*, 77 (1), 87-103.

Lynch, John G. Jr., Joseph W. Alba, Aradhna Krishna, Vicki G. Morwitz, and Zeynep Gürhan-Canli (2012), "Knowledge Creation in Consumer Research: Multiple Routes, Multiple Criteria," *Journal of Consumer Psychology*, 22 (4), 473-85.

MacInnis, Deborah J. and Valerie S. Folkes (2010), "The Disciplinary Status of Consumer Behavior: A Sociology of Science Perspective on Key Controversies," *Journal of Consumer Research*, 36 (6), 899-914.

Maner, Jon K. (2016), "Into the Wild: Field Research Can Increase Both Replicability and Real-World Impact," *Journal of Experimental Social Psychology*, 66 (1), 100-06.

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.

Martin, Kelly D. and Patrick E. Murphy (2017), "The Role of Data Privacy in Marketing," *Journal of the Academy of Marketing Science*, 45 (2), 135-55.

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

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (8), 2421-55.

McAuley, Julian (2018), "Recommender Systems Datasets," 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.

McQuarrie, Edward F., Jessica Miller, and Barbara J. Phillips (2013), "The Megaphone Effect: Taste and Audience in Fashion Blogging," *Journal of Consumer Research*, 40 (1), 136-58.

Melumad, Shiri, J. Jeffrey Inman, and Michel Tuan Pham (2019), "Selectively Emotional: How Smartphone Use Changes User-Generated Content," *Journal of Marketing Research*, 56 (2), 259–75.

Meyer, Michelle N. (2018), "Practical Tips for Ethical Data Sharing," *Advances in Methods and Practices in Psychological Science*, 1 (1), 131-44.

Mick, David Glen (1996), "Are Studies of Dark Side Variables Confounded by Socially Desirable Responding? The Case of Materialism," *Journal of Consumer Research*, 23 (2), 106-19.

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

Mogilner, Cassie, Jennifer Aaker, and Sepandar D. Kamvar (2012), "How Happiness Affects Choice," *Journal of Consumer Research*, 39 (2), 429-43.

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

_____ (2015), "Attitude Predictability and Helpfulness in Online Reviews: The Role of Explained Actions and Reactions," *Journal of Consumer Research*, 42, 30-44.

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.

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

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis (2017), "A Manifesto for Reproducible Science," *Nature Human Behaviour*, 1 (0021), 1-9.

Nelson, Leif D., Joseph Simmons, and Uri Simonsohn (2018), "Psychology's Renaissance," *Annual Review of Psychology*, 69 (1), 511-34.

Nuijten, Michèle B. (2017), "Share Analysis Plans and Results," *Nature*, 551, 559.

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

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

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

Pancer, Ethan, Vincent Chandler, Maxwell Poole, and Theodore J. Noseworthy (2019), "How Readability Shapes Social Media Engagement," *Journal of Consumer Psychology*, 29 (2), 262-70.

Park, Sungsik, Woochoel Shin, and Jinhong Xie (2018), "The Fateful First Consumer Review," *Marketing Science Institute Working Paper Series*, 18 (106).

Patel, Chirag J., Belinda Burford, and John P. A. Ioannidis (2015), "Assessment of Vibration of Effects Due to Model Specification Can Demonstrate the Instability of Observational Associations," *Journal of Clinical Epidemiology*, 68 (9), 1046-58.

Rad, Mostafa Salari, Alison Jane Martingano, and Jeremy Ginges (2018), "Toward a Psychology of *Homo Sapiens*: Making Psychological Science More Representative of the Human Population," *Proceedings of the National Academy of Sciences*, 115 (45), 11401-05.

Rafaeli, Anat, Shelly Ashtar, and Daniel Altman (2019), "Digital Traces: New Data, Resources, and Tools for Psychological-Science Research," *Current Directions in Psychological Science*, 28 (6), 560 –66.

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

Rutz, Oliver J. and George F. Watson (2019), "Endogeneity and Marketing Strategy Research: An Overview," *Journal of the Academy of Marketing Science*, 47 (3), 479-98.

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

Scaraboto, Daiane and Eileen Fischer (2013), "Frustrated Fatshionistas: An Institutional Theory Perspective on Consumer Quests for Greater Choice in Mainstream Markets," *Journal of Consumer Research*, 39 (6), 1234-57.

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

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

Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay (2017), "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers," *Perspectives on Psychological Science*, 12 (6), 1123-28.

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

Snell, James and Nicola Menaldo (2016), "Web Scraping in an Era of Big Data 2.0," https://web.archive.org/web/20181116234409/https://www.bna.com/web-scraping-era-n57982073780/.

Spiller, Stephen A. and Lena Belogolova (2017), "On Consumer Beliefs About Quality and Taste," *Journal of Consumer Research*, 43 (6), 970-91.

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

Statista (2019), "Most Popular Social Networks Worldwide as of July 2019, Ranked by Number of Active Users (in Millions)," https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel (2016), "Increasing Transparency through a Multiverse Analysis," *Perspectives on Psychological Science*, 11 (5), 702-12.

Steenkamp, Jan-Benedict E. M., Martijn G. de Jong, and Hans Baumgartner (2010), "Socially Desirable Response Tendencies in Survey Research," *Journal of Marketing Research*, 47 (2), 199-214.

Stephen, Andrew T. (2016), "The Role of Digital and Social Media Marketing in Consumer Behavior," *Current Opinion in Psychology*, 10 (1), 17-21.

Tonidandel, Scott, Eden B. King, and Jose Cortina, M. (2018), "Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science," *Organizational Research Methods*, 21 (3), 525-47.

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.

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

van 't Veer, Anna Elisabeth and Roger Giner-Sorolla (2016), "Pre-Registration in Social Psychology—a Discussion and Suggested Template," *Journal of Experimental Social Psychology*, 67, 2-12.

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

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.

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

Watson, Jared, Anastasiya Pocheptsova Ghosh, and Michael Trusov (2018), "Swayed by the Numbers: The Consequences of Displaying Product Review Attributes," *Journal of Marketing*, 82 (6), 109-31.

We Are Social & Hootsuite (2019), "Digital in 2017: Global Overview," https://www.slideshare.net/DataReportal/digital-2019-global-digital-overview-january-2019-v01?from_action=save.

Weber, Matthew S. (2018), "Methods and Approaches to Using Web Archives in Computational Communication Research," *Communication Methods and Measures*, 12 (2-3), 200-15.

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

Wickham, Hadley (2019), "Rvest: Easily Harvest (Scrape) Web Pages," *The R Foundation: Vienna, Austria*.

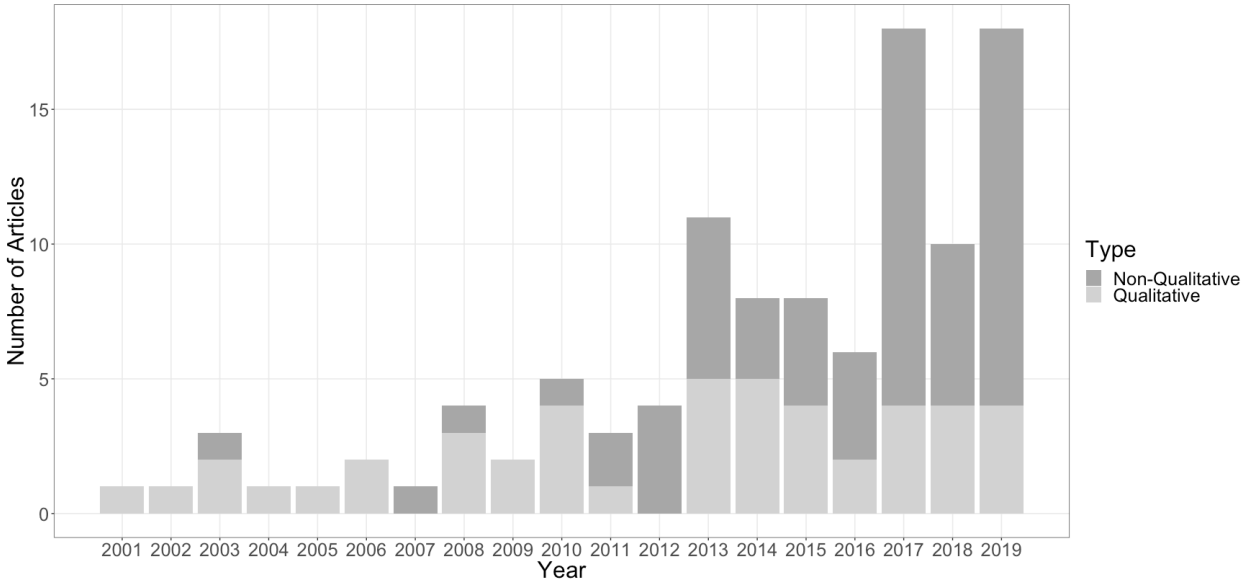
Xu, Heng, Nan Zhang, and Le Zhou (2019), "Validity Concerns in Research Using Organic Data," *Journal of Management*, forthcoming.

Yelp, Inc. (2019), "An Introduction to Yelp Metrics as of June 30, 2019," https://web.archive.org/web/20190828094043/https://www.yelp.com/factsheet.

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

Zhang, Dongsong, Lina Zhou, Juan Luo Kehoe, and Isil Yakut Kilic (2016), "What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews," *Journal of Management Information Systems*, 33 (2), 456-81.

Figure 1: Number of consumer research articles using web-scraped data (2001 – 2019)



Note: To identify articles, we employed a variety of search terms related to web scraping as a method (e.g., scrap*, crawl*, API, netnograph*) as well as websites (e.g., Yelp, TripAdvisor, Twitter, eBay) on Web of Science, Google Scholar, and various search engines and journal websites. We include all consumer research articles published in the top consumer research journals (i.e., *Journal of Consumer Research* and *Journal of Consumer Psychology*) and marketing journals (i.e., *Journal of Marketing*, *Journal of Marketing Research*, and *Marketing Science*). Because we focus on the usage of web-scraped data in consumer research, we count an article published in the general marketing journals if it contains at least one study that experimentally manipulates the core construct(s). While it is not possible to determine if the authors actually scraped the data for every article, we include all article for which the data could have been scraped in order to be as comprehensive as possible.

Figure 2: Types of data commonly scraped for consumer research

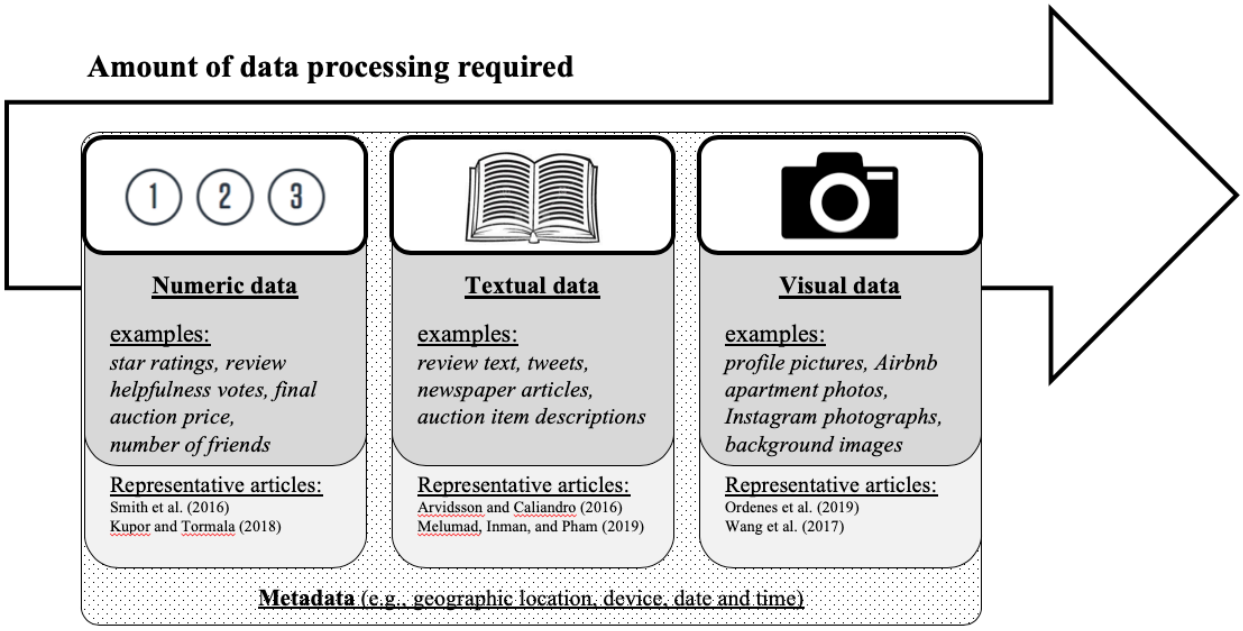


FIGURE 3: STAGES OF STUDY DESIGN USING INTERNET DATA

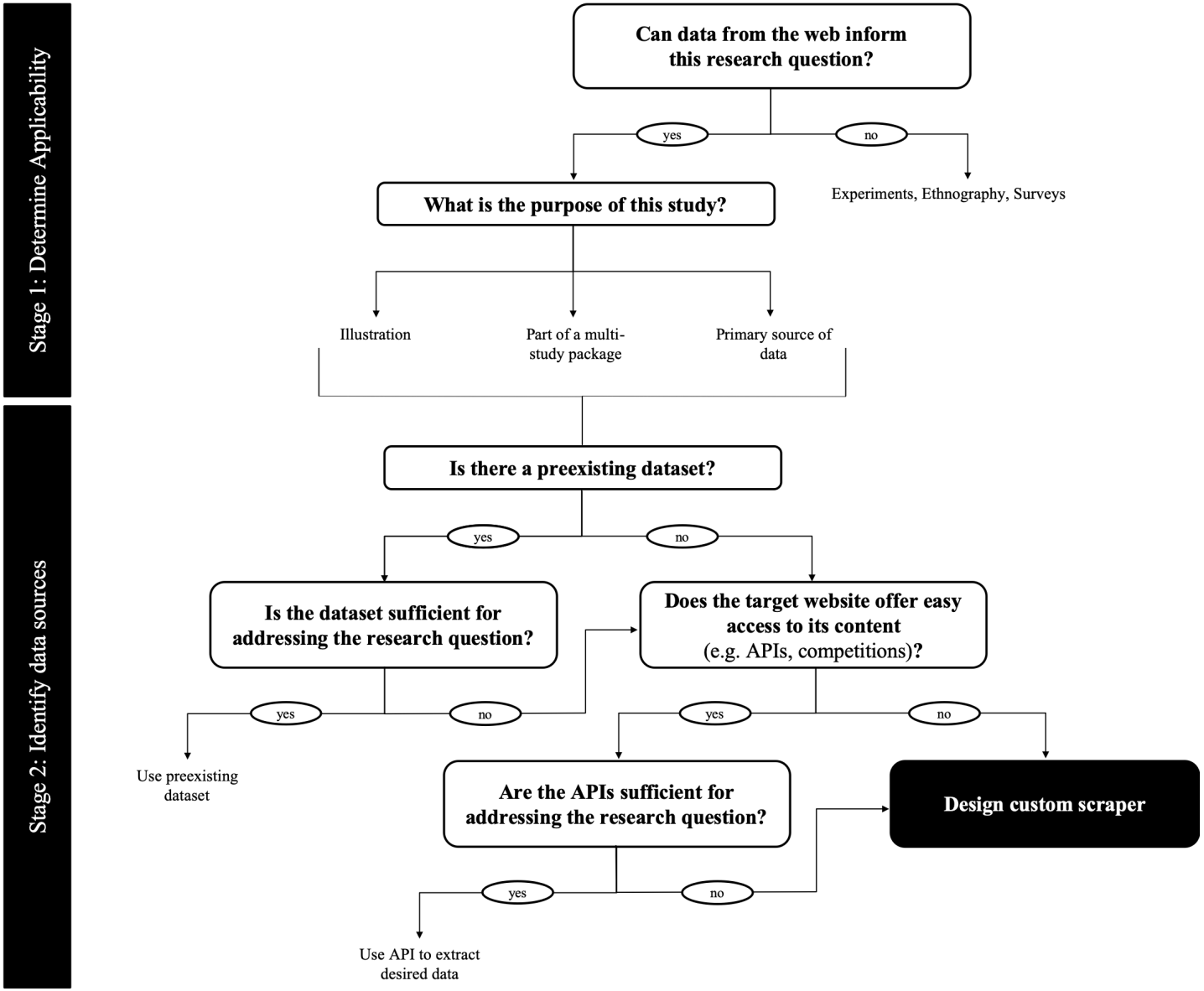


TABLE 1: SOURCES AND TYPES OF WEB-SCRAPED DATA

Source	Examples	Representative articles	Exemplar independent variables	Common dependent variables
Review platforms	Yelp, TripAdvisor			
a) Focus on the reviewer		Henkel et al. (2018); Huang et al. (2016); Paharia et al. (2014)	Reviewer-related variables (e.g., number of reviews [n], presence of a profile image [v], geolocation data [m]) Brand-related variables (e.g., brand positioning [t], tie strength [t], mentions of competitors [t])	Star rating (n) Review positivity (t) Usage of humanizing words (t)
b) Focus on reactions of others to a focal review		Chen and Lurie (2013); Elder et al. (2017); Grewal and Stephen (2019); Kupor and Tormala (2018); Van Laer et al. (2019)	Review star rating (n) Usage of temporal contiguity cues (t) Tense use (t) Mentions of touch and sight (t) Narrative content (t) Device type (m)	Helpful/useful votes (n) Funny votes (n)
Social commerce sites	Amazon, BN.com, Apple App Store	Moore (2012); Ordenes et al. (2017); Yin et al. (2017)	Usage of explanatory language (t) Product type (t) Explicit sentiment expressions (t) Emotional arousal (t)	Sales rank (n) Conversion rate (n) Star rating (n) Helpful votes (n) Language use (t)
Auction platforms	eBay	Bagchi and Cheema (2013); Smith et al. (2016)	Seller reputation (n) Surcharge amount (n) Product serial number (n) Background color (v) Mistakes in photographs (t/v)	Final auction price (n) Bid jumps (n)
Social networks, blogs and microblogs	Facebook, Twitter, Tumblr;	Arvidsson and Caliandro (2016); McQuarrie et al. (2013); Pancer et al. (2019)	Tweets (t, v) Readability (t)	Likes (n) Sharing (n) Retweets (n)
Crowdfunding websites	Kickstarter, DonorsChoose.org	Wang et al. (2017)	Smile intensity in profile pictures (v)	Total amount pledged (n) Average amount pledged per backer (n) Total number of Facebook shares (n)
Notes: (n) = numeric data; (t) = textual data; (v) = visual data; (m) = metadata				

TABLE 2: PURPOSE OF WEB-SCRAPED DATA IN SCHOLARLY CONSUMER RESEARCH

Approach	Description	Illustrative examples
Illustration	Data is scraped from the web to provide some initial insight or illustrate the prevalence of a consumption phenomenon.	Number of Google searchers for top lists ending in 0 or 5 vs. all other numbers (Isaac and Schindler 2014) Scraping the Twitter page of the *Humblebrag* book to illustrate how frequently famous people lamented about the lack of time and busyness (Bellezza et al. 2017) Scraping NFL football betting from Sportsbook.com to assess the effects of the distribution of wagers for casinos (Simmons et al. 2011)
Part of an empirical package using mixed method research designs	Web-scraped data is used in one or multiple studies that are integrated within an empirical package with studies using different methods (e.g., experiments, surveys). Studies using web-scraped data can be positioned either as preliminary evidence for the predicted effect(s) in the wild or as relatively strong confirmatory tests of the predictions.	Presented as preliminary evidence: Henkel et al. (2018); Moore (2012, 2015); Spiller and Belogolova (2017) Presented as confirmatory evidence: Chen (2017); Elder et al. (2017); Umashankar et al. (2017)
Primary source of data	The article is solely or at least largely based on data scraped from the internet.	Using data from the Italian Twitter related to Louis Vuitton to examine what characterizes a brand public (Arvidsson and Caliandro 2016) Using data from multiple platforms in different industries to examine the effect of language used to express sentiment in reviews (Ordenes et al. 2017)

APPENDIX 1: EXAMPLES OF PUBLICLY AVAILABLE DATASETS FEATURING (QUASI-) SCRAPED DATA

Dataset(s)	**Description**	**Exemplary article**
Yelp Academic datasets: https://www.yelp.ca/dataset	Data from one of the leading review platforms (5,996,996 reviews 188,593 businesses 280,992 pictures, 10 metropolitan areas)	McGraw et al. (2015)
DonorsChoose https://research.donorschoose.org https://www.kaggle.com/donorschoose/io	Data from a crowdfunding non-profit platform that allows individuals to donate directly to public school classroom projects (4,687,844 donations, 2,024,554 donors, 901,965 projects, 402,900 teachers)	Kim, Ishihara, and Singh (2018)
Recommender Systems Datasets hosted by McAuley (2018)	Amazon product reviews and metadata; Amazon question/answer data; Google Local business reviews and metadata; Steam video game reviews and bundles; Goodreads book reviews; ModCloth clothing fit feedback; RentTheRunway clothing fit feedback; Tradesy bartering data; RateBeer bartering data; Gameswap bartering data; Behance community art reviews and image features; Librarything reviews and social data; Epinions reviews and social data; Dance Dance Revolution step charts; NES song data; BeerAdvocate multi-aspect beer reviews; RateBeer multi-aspect beer reviews; Facebook social circles data; Twitter social circles data; Google+ social circles data; Reddit submission popularity and metadata	Dai, Chan, and Mogilner (2019); Watson et al. (2018)
https://catalog.data.gov/dataset/consumer-complaint-database#topic%C2%BCconsumer_navigation https://www.fcc.gov/consumer-help-center-data	Consumer Complaint Database maintained by the Consumer Financial Protection Bureau (CFPB) from the US government's open data website	Jung et al. (2017)
Kickstarter & Indiegogo https://www.kaggle.com/kemical/kickstarter-projects https://webrobots.io/kickstarter-datasets/ https://webrobots.io/indiegogo-dataset/	Crowdfunding platform dedicated to realizing creative projects and products.	
Brick et al. (2019)	Directory of 100+ free, open psychological datasets	
Ali (2019); Granville (2016)	Overviews of datasets from websites with relevance to consumer research (e.g., IMDB, Amazon, Twitter)	

APPENDIX 2: LEGALITY OF WEB SCRAPING

The legality of web scraping continues to be debated and there is no clear consensus about whether scraping data for research purposes is permissible from a legal standpoint. As scraping data from a website involves copying data and materials, website providers may claim an *infringement of their copyrights*. The principle of copyright law is that repurposing or republishing copyrighted content requires consent from the owner of the material that can be protected by copyrights. However, fair use laws may make such explicit consent unnecessary as they protect the reuse of copyrighted materials under certain circumstances. Generally, most research projects involving scraping should be able to successfully motivate a fair use defense. Research projects typically meet the four criteria that determine its applicability: the character of the data use, the nature of the copyrighted work, the relation of the portion of the copyrighted work used in the project relative to the copyrighted work as a whole, and any effect the use of data has on the marketability of the copyrighted work.

The second potential cause of liability resulting from web scraping is the *trespass to chattels*, which occurs when a third party intentionally and unauthorizedly accesses, uses, or meddles with the another's physical property and, as a consequence of this unauthorized action, creates tangible monetary or physical damage. In the context of web scraping, these provisions suggest that the servers on which websites are hosted are the private property of the website operators. In practice, across different cases (e.g., eBay v. Bidder's Edge, 2000; Ticketmaster Corp. v. Tickets.com, 2003; Register.com v. Verio, 2004) courts seem to converge on the notion that only excessive scraping has the potential to create sufficient injury for website providers.

Thus, it is critical for consumer researchers to ensure that their scraping does not place excessive burdens on the website's servers or interfere with the functioning of the website.

The third consideration in determining potential liability is whether scraping constitutes an *explicit breach of contact*. Many websites include terms of service (ToS) that outline restrictions for using the data provided on them. Websites differ in how they present their ToS to users: some websites require users to explicitly consent to their terms (i.e., so-called "clickwrap" agreements wherein users click on an "I agree" button after being presented with the ToS). Other websites present their terms as a subpage that is part of their website in a so-called "browsewrap" agreement, which assumes implicit rather than explicit consent. In general, when assessing whether web scraping is a breach of contract, researchers face greater risk when scraping in the presence of an explicit "clickwrap" agreement prohibiting scraping. There may still be potential for liability, however, in cases of "browsewrap" agreements, especially if the researcher has actual or constructive knowledge of the restrictions outlined in these terms (Dreyer and Stockton 2013). Another protocol that many webpage providers use to outline whether and which parts of a website can be scraped is the robots.txt (Hirschey 2014). While it is possible to circumvent the restrictions outlined in these protocols, it is clearly less problematic to collect data from a website with a robots.txt does not prohibit the automatic extraction of its data.

The fourth and final basis for liability from web scraping is the *Computer Fraud and Abuse Act (CFAA)*, which makes it illegal to obtaining data or information from a protected computer through intentional unauthorized access or by exceeding authorized access. In order for the CFAA to be applicable, the unauthorized access must also create monetary damages in excess of $5,000 over a one-year period (Dreyer and Stockton 2013). Recent cases such as QVC, Inc. v. Resultly, LLC (2016) tend to emphasize that in order to be liable scrapers, websites need

to have clear restrictions on access (e.g., in the robots.txt or terms of use) and the scraping must be intentionally designed to harm the website (Snell and Menaldo 2016). As in potential breaches of contracts, violations of the CFAA require that a user (e.g., a researcher) was aware that scraping the website constituted unauthorized access.

In sum, web scraping continues to be a legal gray area. Researchers need to be aware of the potential liability resulting from web scraping. In particular, republishing scraped datasets is highly problematic. In most cases, sharing the actual data is not permitted due to the website user agreement, but sharing the code is usually less problematic (Braun et al. 2018). To reduce their exposure to potential liability, researchers should scrape content that is not subjected to copyright protection and design mindful scrapers that minimize the burdens placed on the website and are in accordance with the website's terms of service. In addition to these concerns, researchers also need to ensure that they comply with the data protection laws in the jurisdiction in which the researcher or the population that generated the scraped data resides. For example, the European Union introduced a new comprehensive data protection law called the General Data Protection Regulation (GDPR) in 2018. Under the GDPR, it is illegal to extract and store personally sensitive data (e.g., name, e-mail address, IP addresses) of EU citizens without their explicit consent (Kerins 2018).