FIELDS OF GOLD:
GENERATING RELEVANT AND CREDIBLE INSIGHTS
VIA WEB SCRAPING AND APIs

Johannes Boegershausen
Abhishek Borah
Hannes Datta
Andrew T. Stephen

This version: 28 March 2021
FIELDS OF GOLD:
GENERATING RELEVANT AND CREDIBLE INSIGHTS
VIA WEB SCRAPING AND APIs

ABSTRACT

Marketing researchers increasingly use web scraping and Application Programming Interfaces (APIs) to collect publicly available data from the internet. While guidance on the technicalities of collecting web data are abundant, much of the design decisions involved in collecting web data have remained largely neglected and undiscussed. A lack of awareness and understanding of these design decisions, both among authors and reviewers, threatens the credibility of research findings based on web data. To address these issues, this article develops a systematic workflow that guides researchers across the different stages of collecting web data. Throughout, the authors discuss how various design decisions affect the relevance and credibility of research findings. The workflow is accompanied by a comprehensive review of the use of web data in marketing research, identifying common themes of how web data has enriched past work. Finally, the authors highlight promising avenues for how future work might leverage web data to address important marketing questions and disseminate research findings.

Keywords: web scraping, application programming interface (APIs), field data, research credibility, reproducibility, replicability, generalizability, ecological value, research methods, workflow, open science, open data
The accelerating digitization of social life and commerce has created unprecedented
digital traces of consumer and firm behavior. This data – enormous in size, diverse in form,
and often publicly accessible on the internet – is a potential goldmine for marketing scholars
who seek to quantify consumption, gain insights on firm behavior, and track social activities
that are difficult or costly to observe otherwise. The most efficient and scalable way to collect
such data is through web scraping or using application programming interfaces (APIs). Web
scraping is the process of designing and deploying software that extracts information from
websites without the technical and logistical necessity to involve the data provider. Some
firms also offer APIs as an officially supported method to extract data. Web scraping and
APIs have been fundamental to the business model of many leading firms, feeding search
engines with new data (e.g., Google), allowing firms to integrate novel algorithms (e.g.,
Amazon Web Services), and providing valuable insights about consumers and competitors
(e.g., Du et al. 2021). Because of the abundance of data on public websites and the APIs of
digital platforms, these methods have been used by marketing researchers across all
methodological traditions in marketing, including, but not limited to, empirical modeling,
marketing strategy, consumer psychology, and consumer culture theory. In the remainder of
this article, we refer to data collected via web scraping and APIs as web data.

Marketing scholars aim to produce relevant and important research findings that are
considered valid by the research and business community (Moorman et al. 2019b). Web data
offers numerous opportunities to fuel the relevance of academic marketing research.
Researchers can capture the real-life behavior of consumers and firms, obtain data about
novel phenomena, and generate insights about managerially relevant outcome metrics. For
relevant research findings to advance theory and inform managerial practice, they need to be
based on valid data. Common standards and methodological best practices are established for
conventional data collection methods (e.g., experimental, survey, or scanner panel data).
However, relative to these data collection methods, there is a lack of clarity on how to collect web data for use in academic research, let alone a holistic perspective that covers the entire research process of gathering and using web data. Without common standards to ensure sufficient levels of thoroughness, exhaustiveness, and accuracy, it is difficult to evaluate and compare research findings, let alone reproduce or replicate them. These concerns are magnified by the numerous – and hitherto undocumented – degrees of freedom along all stages of collecting web data. Some of the essential decisions include choosing between different websites and APIs to collect data from, how to strike a balance between sample size requirements and technical retrieval limits, or how to mitigate legal concerns arising from collecting, processing, and storing web data.

To address these issues, we develop a systematic and relatively timeless workflow, guiding researchers in the crucial design decisions of collecting and using web data for academic research. We aim to make three main contributions. First, we create a typology of the use of web data in the marketing literature to illustrate the versatility of web scraping and APIs as a data collection method. Based on a comprehensive review of more than 200 articles published in five of the leading marketing journals, we identify four key themes for leveraging web data to increase the relevance of marketing research: (1) the exploration of new phenomena, (2) improved data access and quality, (3) increased ecological value and impact, and (4) the development and use of new methods.

Second, we provide a systematic workflow that offers clear “how-to” guidance regarding the key steps and resource considerations in collecting and using web data. Because the technicalities of collecting web data are constantly in flux (see Web Appendix W2 for a list of tutorials) and holistic frameworks focused on the academic use of web data are absent, we develop a novel workflow that reflects a design-thinking perspective. Our workflow is sufficiently general (i.e., valid for both websites and APIs; relevant across
methodological traditions), relatively timeless (i.e., not tied to a specific extraction software), and guides authors in the key decisions when collecting data from the web. In the first phase – the planning phase – we illuminate how to screen the web for potential data sources, show how researchers can assess data availability and accessibility, evaluate whether the data fits their research purpose, and gauge whether extraction is technically feasible. In the second phase – the execution phase – we outline the key considerations to safeguard data quality and elaborate on how to prepare and document web data for internal and external use.

Throughout, we illustrate how the various design decisions affect the analytical robustness, replicability, and generalizability of research findings. Thereby, our workflow offers a clear frame of reference, primarily for authors, but also for reviewers and the wider research community interested in evaluating and comparing research findings based on web data.

As a third contribution, we present future opportunities for using web data. In particular, we identify promising but underutilized ways for marketing researchers to challenge the boundaries of what we study (e.g., collecting more diverse samples to address underrepresented populations, examining phenomena unobtrusively). We also highlight emerging, cutting-edge applications of web scraping and APIs that challenge the boundaries of how we study marketing issues (e.g., using API-supported field experiments, increasing scope by collecting data remotely).

We hope that the adoption of our workflow will accelerate the production of credible, valid insights of high relevance and importance to our field (LeBel et al. 2018; Moorman et al. 2019a). In what follows, we first review how marketing research has used web data to provide insights on important marketing questions. Next, we present the workflow, and subsequently discuss challenges and potential remedies to ensure that web data-based findings are analytically robust, replicable, and generalizable. We conclude by identifying opportunities for future research using web data.
USE OF WEB SCRAPING AND APIs IN MARKETING RESEARCH

Marketing researchers increasingly rely on web data – i.e., data collected from publicly available websites and APIs. In 2020 alone, more than 30 articles published in five leading marketing journals used web data (see Figure 1 for an overview and Web Appendix W1 for details). But, how exactly has web data contributed to the generation of novel insights? Via a comprehensive literature review across substantive topics and methodological traditions, we identified four key themes, which we discuss next.

**Figure 1: Increased use of web data in marketing research (2004-2020)**

![Graph showing increased use of web data in marketing research](image)

*Notes: Number of marketing research articles with data collected using web scraping and APIs. Included journals are Journal of Marketing (JM), Journal of Marketing Research (JMR), Marketing Science (MKTSCI), Journal of Consumer Research (JCR), Journal of Consumer Psychology (JCP). See Web Appendix W1 for details on article selection.*

**Theme 1: Web data allows researchers to explore new phenomena**

Initial work using web data focused on novel, emerging phenomena such as online conversations (Godes and Mayzlin 2004) and consumer reviews (Chevalier and Mayzlin 2006). Web data has boosted the field’s relevance by decreasing the time between the occurrence of a new phenomenon and its scientific study—sometimes even years before commercial data providers offered comparable datasets to researchers for purchase.
Due to the inherent timeliness of web data, it has enabled marketing scholars to assert intellectual leadership on important substantive issues such as the sharing economy (e.g., Airbnb; Zervas et al. 2017), access-based business models (e.g., Spotify; Datta et al. 2018), and fake online content (e.g., Anderson and Simester 2014). Web data has also been used in interpretative and inductive research focusing on developing novel theories about emerging marketing phenomena (e.g., brand public, Arvidsson and Caliandro 2016).

Examining new phenomena with web data has allowed researchers to weigh in on contemporary issues marketing practitioners grapple with (van Heerde et al. 2021). Consider, for instance, the question of whether businesses should respond to customer reviews. While responding is encouraged by platforms (e.g., Yelp, Google), the strategy’s effectiveness remained untested. Via multi-platform web data, consensus now emerged that firms should only respond to negative but not positive reviews (Chevalier et al. 2018; Ma et al. 2015; Proserpio and Zervas 2017; Wang and Chaudhry 2018).

**Theme 2: Web data improves data access and quality**

A second benefit resulting from the use of web data is improved data access and quality. By improved access, we mean that web data can often be collected without the data provider's direct involvement, leading to substantial time and cost savings in negotiating data access (Mela 2011). By improved quality, we mean that web data can be richer than traditional data sources (i.e., higher frequency, more covariates) and gives researchers direct control over data selection and preparation. Thereby, researchers can limit the interference of conventional data suppliers (e.g., collaborating firms) to ensure that the societal relevance of a particular research question is given precedence over business objectives (e.g., firms might be unwilling to share data about the tracking tools they use on websites; Trusov et al. 2016).

A research stream that has particularly benefitted from improved data access and quality is work on the entertainment industries (e.g., music, movies). Liu (2006), for
example, leveraged web data to examine the influence of word-of-mouth dynamics on box office revenues. Breakthroughs in this research stream were stimulated by rich datasets featuring commercial outcomes (e.g., box office revenues from Box Office Mojo and The Numbers; Ryoo et al. 2021; hit chart positions from the Ultimate Music Database; Valsesia et al. 2016), cost data (e.g., production budgets from Wikipedia; Heath et al. 2015), and product characteristics (e.g., stars from IMDb; Hennig-Thurau et al. 2009).

The extraction of web data can unleash the potential of hitherto inaccessible, unstructured data sources. For example, Homburg et al. (2020) obtained a measure on marketing excellence by automatically downloading and processing the annual reports of more than 8,000 firms over two decades from AnnualReports.com. Similarly, web scraping and APIs provide access to relevant variables describing a particular research context, including data about the weather (e.g., Weather Underground; Li et al. 2017), disease prevalence (e.g., CDC FluView Interactive; Galoni et al. 2020), and raw material prices (e.g., MSN Gas Prices; Nishida and Remer 2018).

Finally, researchers can collect high-frequency web data on the behavior of consumers and firms (e.g., Adjerid and Kelley 2018). Such data facilitates the construction of panel data, which can be used to study theoretically relevant variation within individuals over time (e.g., Huang et al. 2016) or how effects unfold over time (e.g., Tellis et al. 2019). The real-time nature of web data also allows researchers to study behaviors at very high granularity (e.g., in days, hours, minutes, seconds), paving the way for novel insights or tackling methodological issues (e.g., Du et al. 2019).

Theme 3: Web data boosts ecological value and impact

As a third advantage, researchers have leveraged web data to move closer to the “natural habitat” of marketing (van Heerde et al. 2021). As web data can be unobtrusively collected, it can effectively complement more controlled data collection methods such as
laboratory experiments by demonstrating that focal psychological processes occur outside the confines of a controlled environment with stylized experimental stimuli (Morales et al. 2017).

Consider, for instance, the recent controversy around the decoy effect (Huber et al. 1982), one of the most prominent context effects in consumer behavior. Frederick et al. (2014) questioned the robustness and practical relevance of the decoy effect, arguing that the effect only emerged in tightly controlled laboratory settings. Using web data, Wu and Cosguner (2020) built a panel dataset from an online diamond market to demonstrate not only the existence of the decoy effect, but also its practical significance in the marketplace by quantifying its profit implications.

Another pathway for enhancing ecological value with web data is to illustrate the importance of certain consumer and organizational behaviors in the marketplace (e.g., Howe and Monin 2017; Tonietto and Barasch 2020). For instance, Bellezza and Berger (2020) demonstrate and study the prevalence of trickle-round status signals in a dataset covering more than 100,000 menu items of selected restaurants in New York.

Theme 4: Web data eases the development and use of new methods

As much of the data produced by consumers and firms is inherently unstructured, extracting insights from such data is challenging (Wedel and Kannan 2016). Thus, marketing researchers have leveraged web data for developing methods that deal with and extract insights from different types of unstructured data (Balducci and Marinova 2018). For instance, web data has fueled the rapid improvement of automated text analysis (see Berger et al. 2020) and the large-scale analysis of audio, image, and video content (e.g., Li et al. 2019; Liu et al. 2020; Wang et al. 2021). The abundance of web data has also led to a surge in methods for generating insights from unstructured data, for example, at higher levels of analysis such as the market-level (e.g., Netzer et al. 2012) or network-level (e.g., Oestreicher-Singer et al. 2013). The relationship between web data and methodological advances is often
symbiotic such that emerging practical challenges faced by firms inspire methodological advances, and – in turn – the rich available web data facilitates the development and validation of more sophisticated methods (e.g., Dew et al. 2020).

Table 1 summarizes the main uses of web data in academic marketing research and provides exemplary studies that have leveraged web data to study outcomes across different levels ranging from consumers to organizations to other marketing stakeholders. For example, Toubia and Stephen (2013) explored a new phenomenon (tweeting), focusing on consumer behavior (i.e., consumers’ motivation to tweet).

Table 1: Typology of the use of web data in marketing research with examples

<table>
<thead>
<tr>
<th>Effect on (with typical outcome variables in parentheses)</th>
<th>Primary reason for using web data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theme 1: Exploration of new phenomena</td>
</tr>
<tr>
<td>Consumers (e.g., learning and social media use of consumers; sentiments of customers; interaction in a network)</td>
<td>Toubia and Stephen (2013): testing the motivations of users to contribute content to social media.</td>
</tr>
<tr>
<td>Other marketing stakeholders (e.g., market reaction of investors, public health outcomes)</td>
<td>Hermosilla et al. (2018): examining how consumers’ aesthetic preferences create biases in firms’ hiring decisions.</td>
</tr>
</tbody>
</table>

Notes: The table cross-tabulates four themes on how web data has been used in academic marketing research (the columns), with three of the most commonly studied actors (the rows).
TOWARD A SYSTEMATIC WORKFLOW FOR COLLECTING WEB DATA

Despite the widespread use of web data across all subfields of marketing, the various design decisions that affect the relevance and credibility of research findings based on it have not been systematically documented and discussed. In particular, the collection and use of web data require attention to numerous details that many marketing scholars might be unfamiliar with.

For instance, experimental researchers are used to tightly controlling the design and execution of their studies. However, when working with web data, they are likely to face novel challenges that may affect their research design (e.g., sampling). Similarly, researchers trained in working with archival data are used to contacting their data provider after data delivery to discuss quality concerns or the operationalization of variables (Mela 2011). Yet, the extraction of web data for research purposes is rarely the goal of website and API providers (Xu et al. 2020). As researchers typically are not (or do not succeed to be) in contact with data providers, they must carefully choose which data to extract from which website or API. Finally, researchers using web data need to navigate a highly unfamiliar, novel, and complex legal landscape to mitigate exposure to legal risks, such as by means of choosing their extraction method. In short, the development of a systematic workflow for collecting web data, spanning all subfields of marketing, is much needed.

For relevant research findings to advance theory and inform managerial practice, they need to be certified as credible and valid. Methodological rigor is essential for the validity of research findings. Findings are rigorous when the probability of the central conclusions being false given the aims of the research is sufficiently low (Kohli and Haenlein 2021; Lehmann et al. 2011). Sound, defensible decisions in the research workflow enhance rigor. Such decisions range from measurement, validity, and sampling to transparent reporting
At present, we lack standards ensuring the thoroughness, exhaustiveness, and accuracy to consistently conduct research and report findings based on web data. Given the numerous degrees of freedom researchers have in collecting web data, it is difficult to reproduce, replicate, evaluate, and compare research findings based on web data.

To ensure researchers can fully leverage the benefits of web data, we take inspiration from LeBel et al. (2018) and discuss — in the context of web data — the four interdependent facets that generate credible research findings: (1) Design transparency involves the considerations and disclosure of the key decisions leading to the collection of the dataset. (2) Analytical reproducibility is achieved when sufficient documentation enables others to reproduce the exact numerical results. Findings are considered (3) analytically robust if they emerge consistently across other plausible data-processing or data-analytic decisions applied to the same dataset. Finally, credible findings are characterized by (4) effect replicability and generalizability. An effect is replicable when it consistently occurs in new samples at similar magnitudes when using methods similar to those to the original study. Findings are generalizable when they emerge in replications that are highly dissimilar methodologically.

While all of these four facets are relevant to producing credible research findings, we now develop a systematic workflow that covers the most underappreciated challenges when collecting web data related to (1) design considerations and (2) analytical reproducibility. Without a thorough understanding of how to best capture data from the web and providing transparency about the design decisions and analytical reproducibility, improving analytical robustness, effect replicability, and generalizability are in vain. We discuss additional considerations (i.e., for analytical robustness, effect replicability, and generalizability) after our core workflow.
WORKFLOW FOR COLLECTING WEB DATA

Given the terabytes of data generated by consumers and firms, data for academic research may seem abundant. However, not all data publicly available on the web lends itself for use in a scientific research project. To guide researchers in the planning and execution of their data collection, we develop a systematic workflow consisting of a planning phase and an execution phase. The planning phase is characterized by a narrowing funnel in which researchers eliminate data sources as they become more familiar with them. Specifically, researchers (1) broadly explore the web to identify websites and APIs with potentially relevant data about the research question or the research context. Researchers then (2) assess – for each website or API – which data is available, and (3) evaluate whether the data fits the research purpose and determine the resources required for data extraction. Finally, researchers (4) build a prototype to test whether the extraction methodology is technically feasible. Throughout the planning phase, researchers identify legal concerns, some of which could be mitigated by design, while others require obtaining legal support. In the second phase of the workflow – the execution phase – researchers (5) collect the data and monitor data quality during and after the collection. Finally, researchers (6) preprocess, document, and distribute the data. At this stage of the workflow, the funnel widens as datasets can potentially be used across multiple research projects or shared publicly.

Figure 2 visualizes the workflow for collecting web data and underscores some of the key questions discussed in them. It is important to note that while the figure is stylized as a linear workflow, researchers typically go back and forth between different stages when planning and collecting web data. For all stages of the workflow, we provide additional figures, summary tables, and checklists in Web Appendix W3.
Figure 2: Workflow for Collecting Web Data for Academic Research

<table>
<thead>
<tr>
<th>Description of workflow steps</th>
<th>Main considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Source Exploration</strong></td>
<td>Researchers make a list of potential data sources (websites and APIs) that could inform the research question or the research context.</td>
</tr>
<tr>
<td><strong>Data Availability Assessment</strong></td>
<td>For each data source, researchers assess which data is available (e.g., by browsing the website, reading the API’s documentation, or interacting with the API).</td>
</tr>
<tr>
<td><strong>Evaluation of Research Fit and Resource Use</strong></td>
<td>Researchers assess whether the available data fits the research purpose, and gauge which resources are required to obtain the data (e.g., development and operation costs).</td>
</tr>
<tr>
<td><strong>Prototype Development</strong></td>
<td>Researchers build a prototype to test how the desired data can be captured. They evaluate legal and ethical concerns of collecting and storing the data, and seek legal advice for any unresolved issues.</td>
</tr>
<tr>
<td><strong>Collection and Monitoring</strong></td>
<td>Researchers collect their data (e.g., locally on their own computer, or remotely in the cloud). During and after the collection, they monitor data quality and debug the extraction software as necessary.</td>
</tr>
<tr>
<td><strong>Preprocessing, Documentation, and Distribution</strong></td>
<td>Researchers preprocess the collected data, document its collection and composition, and distribute the data to team members, external users or data repositories for long-term, archival purposes.</td>
</tr>
</tbody>
</table>

**Main considerations**

- Have different types of data sources been considered (e.g., websites and APIs)?
- Do some data sources have both a website and an API? How do they differ?
- Is the data provider a primary data provider, or an aggregator?
- Is the data source publicly accessible, or only available after payment?
- Is similar data already available (e.g., via the data provider, or platforms like Kaggle)?
- Which entities are available, and can they be linked to external data?
- How many instances are available for each entity?
- Which lists or views could serve as seeds to start the data collection?
- For what period is the data available? Is the data truly archival, or can it be modified?
- Which algorithms could affect the display of data, can they be observed, and can the researcher control them?
- How can instances be sampled from the site?
- To which population does the sample generalize?
- Are the desired sample characteristics (e.g., extraction frequency) attainable?
- Can constructs be measured with the available data?
- Is it possible to transform the raw data to a format suitable for analysis?
- Does the potential value of the data justify the resources used for obtaining it?
- Among the available entities, which ones to extract, at what frequency?
- Which data to capture for each entity, and which seeds and sampling procedure to use?
- How to navigate the website or API, both within and across entities?
- How to identify and deal with technical hurdles (e.g., login, captchas)?
- Which software to use for data extraction, and how to store data and deploy software?
- How to mitigate any legal risks or ethical issues? Is legal support required?
- How to monitor whether the extraction still runs? How to ensure data quality during the collection, and at which frequency are these data quality checks conducted?
- How to verify that all intended instances were indeed captured?
- Which other data sources (e.g., social media, blogs) need to be monitored to stay up-to-date about the context of the data collection?
- Which transformation steps are necessary to convert the collected raw data to a cleaned dataset for analysis?
- Which specific summary statistics are informative about data composition and quality?
- How and where to document the data?
- How to distribute the data to team members, external users, and which data repositories to use for long-term, archival purposes?

**Notes:**
The workflow for collecting web data is a funnel consisting of two main phases: In the planning phase, researchers gradually narrow down a list of websites and APIs to identify relevant web data that is feasible to collect. As researchers learn about potential legal concerns of their data collection, they can mitigate them by changing their prototype or by obtaining legal advice on any unresolved issues (see Web Appendix W3, Table W3.4 for details). In the execution phase, researchers collect the data and preprocess, document, and distribute it. The funnel widens in the final step of the workflow, given the various potential use cases for the data (e.g., for replication of the study’s results or use in other research studies testing novel predictions). While the funnel suggests linearity, researchers typically cycle back-and-forth between subsequent workflow steps when collecting web data.
Step 1: Data Source Exploration

At the outset of any research project that might benefit from gathering web data to examine marketing phenomena or test theory, researchers need to carefully explore which websites and APIs could potentially be relevant to directly address the research question.

In discovering web data, researchers can benefit from considering various types of data sources, ranging from primary data providers (e.g., YouTube) to data aggregators (e.g., SocialBlade), and highly popular (e.g., Spotify) to less popular sites (e.g., Last.fm), and global services (e.g., Facebook) to more regional services (e.g., WeChat). While lesser-known data sources may have fewer users than global platforms, they may offer richer data or novel measures and warrant further attention. Similarly, data aggregators can drastically facilitate the collection of multi-platform data compared to gathering data from numerous primary data providers. An important consideration is whether the data originates from a website (e.g., twitter.com) or an API (e.g., developer.twitter.com). As a basic rule-of-thumb, data collected via APIs – unlike websites – is typically designed to be exchanged with others, facilitating data extraction and use. Figure W3.1 in Web Appendix W3 illustrates the position of a few selected data sources along these dimensions.

Researchers can also benefit from exploring alternative data sources that need not be scraped or collected via an API. Some data providers (e.g., Yelp), public data platforms (e.g., Kaggle), and researchers (e.g., McAuley 2021) provide access to large datasets for scientific use, which may save researchers the time effort of collecting comparable data on their own (see also Paxton and Griffiths 2017). When working directly with firms or through institutionalized access (e.g., Marketing Science Institute, Wharton Customer Analytics Initiative), researchers can also benefit from internal firm data that cannot be scraped or accessed via an API (such as a firm’s clickstream data to study consumers’ purchasing behavior at the individual level).
Step 2: Data Availability Assessment

For the most promising data sources, researchers carefully assess which data is available. They do so by browsing the website, screening the API documentation, or exploring the sample data retrieved from the API. We summarize the main considerations for evaluating data availability in Table W3.1 in Web Appendix W3.

Entity Coverage and Linkages

Entities are the objects (e.g., users, products, reviews, brands, firms) for which data can be collected. To gauge whether the data potentially lends itself for scientific use, researchers need to identify which entities are available and which information can be collected from them. Depending on the data source, several entities may be available. Probably the most commonly studied platform in marketing is Amazon. Here, some of the entities examined by researchers include products (e.g., Chevalier and Mayzlin 2006), reviews (e.g., Ordenes et al. 2017), and product networks (Oestreicher-Singer et al. 2013). Yet, Amazon also contains data about other, less apparent entities such as users (e.g., their location) and third-party sellers (e.g., shipping fees). Researchers can start by scanning the site for entities and investigate how to uniquely navigate to an instance of a particular entity. For example, Amazon uses “Amazon Standard Identification Numbers” (ASINs), which – joined with the platform’s URL – allow directing a browser to the respective product page. On other platforms, instances may not be uniquely identifiable (e.g., reviewer IDs are absent at Airlinequality.com), rendering the collection of reviewer-level panel data infeasible.

Researchers may benefit from linking their data to other datasets. For example, Amazon lists the International Standard Book Numbers (ISBNs) of books – a commonly used identifier also available in other datasets. Linkages can also be established without such identifiers, for instance, by using text-based matching (e.g., searching for music artists using the Songkick API). Researchers might also leverage other web data to establish linkages
(e.g., geolocation and web search data; Wang and Chaudhry 2018) or augment their data with self-generated timestamps for subsequent matching (e.g., linking review data to stock prices; Tirunillai and Tellis 2012). Establishing linkages to external datasets is particularly promising if it allows researchers to combine publicly available web data with measures otherwise not observable on the web (e.g., sales data, clickstream data).

Next, researchers have to investigate how many instances of an entity are available and how many of them can be retrieved. For example, the category pages on Amazon only list up to a few hundred products – out of potentially thousands of products per category. Researchers can assess limits to the availability of instances by browsing on the site (e.g., viewing a search result’s last page) or retrieving sample data from an API (e.g., changing the offset parameter used for pagination). Researchers may also experiment with less straightforward ways to gather entity-level data, such as using the data provider’s search function (e.g., by searching for products in alphabetical order) or by exploring peer networks (e.g., gathering usernames by visiting a user’s friends).

Finally, researchers need to consider which lists or views can serve as the starting point (“seed”) of the data collection. While lists simply hold an array of IDs (e.g., usernames, product IDs), views are particular sections on a website that generate these lists (e.g., a category overview page at Amazon for product IDs). One possibility for generating seeds is to leverage site-specific information. For instance, Datta et al. (2018) captured the site’s “recently active users” page for one month and subsequently drew a random sample of 5,000 users from that list. There might be salient external lists in some categories (e.g., books) that can serve as seeds, such as the New York Times or Publishers Weekly bestseller lists (Chevalier and Mayzlin 2006).
**Time Coverage**

Next, researchers need to probe for what period data is available. Some data sources provide real-time data. For example, online gaming company Steam shares data about players’ most recent game interactions, implying that it is impossible to retrieve it later if such data is not captured in real-time. Other data sources provide rich archival data. For example, spotifycharts.com provides the top 200 most streamed songs on a day, allowing the researcher to go back several years.

In recording archival data from websites and APIs, researchers should pay close attention to date and timestamps (e.g., time zones). Date and timestamps may also not be entirely accurate (e.g., “posted more than a year ago”), diminishing the data’s usefulness for building a panel data set. Researchers should also take note of the modifiability of data after it has been originally posted. For example, some sites allow users to change or delete their posts. To detect such practices, researchers may have to compare content over time to evaluate whether data that initially appeared to be archival actually is. Observing modifications (e.g., Amazon deleting fake reviews) is often not just a nuisance that warrants collecting data in real-time but can even inspire asking novel research questions (e.g., whether providers effectively remove bought fake reviews; He et al. 2021).

Last, in planning a real-time data collection, it is paramount to know how often the website or API is refreshed to determine the periodicity for collecting the data. If data is not updated frequently or at irregular intervals, gaps in the data may require researchers to use interpolation techniques or aggregate their data at a lower level of granularity (e.g., month, year) to prevent model misspecification.

---

1 For some data sources, the “Wayback Machine” (https://archive.org) may offer selected snapshots of parts of the website for extraction. Typically, these snapshots are only consistently available for high-traffic websites.
**Algorithmic Transparency**

The (visual) design of websites and APIs or algorithms that make them navigable and fast can cause problems in collecting and using data for scientific purposes. Typical mechanisms that affect the display and retrieval of data are sorting algorithms (e.g., by “popularity” or mixed with sponsored search results), filtering algorithms (e.g., only returning subsets of the data), or personalization algorithms (e.g., recommendations). For example, the free version of the Twitter API returns just a *sample of tweets* – with varying sample sizes over time, rendering any volume count questionable.

When screening a data source for data availability, researchers may notice they can sometimes control the mechanisms through which data is displayed. For example, researchers can sort products alphabetically or chronologically, rather than by popularity, which curbs some of the most severe selection issues that may compromise the generalizability of research findings. Sometimes, researchers can also opt-out of any firm-administered field experiments, which further helps to ensure the representativeness of the collected data.

Finally, researchers should pay attention to how the metrics they retrieve are calculated. For instance, the metric labeled as “playlist listeners” on Chartmetric.com was merely an approximation rather than a playlists’ actual number of listeners. In sum, researchers may want to understand how metrics have been calculated rather than taking them as given.

**Step 3: Evaluation of Research Fit and Resource Use**

Researchers subsequently evaluate the fit of each of the potential data sources with the specific research objectives, taking into account the monetary and non-monetary resources required to extract the data. Table W3.2 in Web Appendix W3 summarizes the key issues and questions when evaluating the research fit and resource use.
Sampling and Generalizability

In step 2 of the workflow, researchers have identified lists or views that could serve as a starting point for the data collection. In evaluating research fit, researchers judge whether the population to which the seeds generalize is aligned with the study’s research objective. For example, early work in a particular domain tends to be concerned with the descriptive evidence and existence of a phenomenon rather than full generalizability (e.g., Chevalier et al. 2018; Ma et al. 2015). For subsequent inquiries, though, more representative samples can lead to the discovery of theoretically interesting moderators and boundary conditions (e.g., Wang and Chaudhry 2018).

Researchers also need to motivate sample characteristics (e.g., the minimum sample size to satisfy statistical power requirements). By collecting a sufficiently large sample, for example, researchers can ensure that null effects are informative. Importantly, gathering web data does not imply collecting enormous datasets (Adjerid and Kelley 2018). Ironically, very large samples might magnify biases resulting from issues in the study design (e.g., using suboptimal seeds). An essential consideration in motivating sample characteristics is to consider the technically feasible sample size. For example, while seeds may be abundantly available, data extraction for all of them may be infeasible, rendering some research designs inferior to others. A study’s sample size critically depends on technical parameters such as the number of computers used for data extraction, the number of URLs to visit for each instance, and the desired sampling frequency. As a result, researchers typically need to narrow down their seeds list through (quasi-) random or stratified sampling. We explain how to calculate the technically feasible sample size in Table W3.3 in Web Appendix W3.

Construct Measurement

Researchers need to demonstrate that data organically generated on a site is a valid operationalization of the theoretical construct(s) of interest. This is particularly relevant when
using a multi-method approach to test predictions (e.g., web data and experiments). Pretests can help justify that the specific entities (e.g., brands, industries) are closely aligned with the focal theoretical constructs of interest (e.g., Henkel et al. 2018; Umashankar et al. 2017).

A typical complexity for construct measurement is that the ideal measure might not be available on the website or API. For instance, ideally, researchers would like to examine the effect of the valence of online reviews on sales on Amazon. Yet, Amazon does not publish actual sales. Researchers have used various approximations for sales on Amazon to address this issue, such as the (log of) sales rank (Chevalier and Mayzlin 2006). Researchers collecting data from multiple sites also need to assess how different metrics and measures are comparable. For instance, are Amazon’s helpfulness votes and Yelp’s useful votes different operationalizations of the same construct, or do they measure other underlying constructs?

When comparing the different shortlisted data sources, researchers evaluate whether some data sources allow them to operationalize the core construct(s) better or provide a stronger test for focal predictions relative to other data sources. For example, a researcher interested in studying how observers react to humor in reviews might prefer Yelp to an alternative platform as it is the only source with so-called “funny votes” (e.g., McGraw et al. 2015). If researchers are indifferent between different potential data sources, they can collect data from multiple data sources to increase the generalizability of their study’s central findings (e.g., tweets and restaurant reviews; Melumad and Meyer 2020).

Finally, researchers need to consider the construct validity implications arising from the collection and processing of web data. Whereas establishing construct validity might seem easier for numeric data (e.g., Yelp rating, number of helpful votes in Amazon), researchers need to carefully scrutinize how the operationalizations and validity of certain variables. For example, it is unclear how Yelp aggregates individual-level consumer ratings to composite metrics about a restaurant’s price tier. For more unstructured web data (e.g., text, images,
audio, video), researchers should plan in advance how they intend to process the data for analysis. Several decisions at this stage of the workflow influence subsequent analytical choices. For example, transforming textual and visual data into valid operationalizations of theoretical constructs often involves trade-offs between automated and human processing, affecting accuracy, feasibility, and scalability (e.g., Xu et al. 2020). When a researcher’s goal is to gather multi-platform data, it may be more feasible to collect data from a data aggregator rather than collecting data from individual platforms and combining them later.

**Data Structure and Preparation**

Researchers need to ensure that the desired data structure for analysis aligns with their research objective. For example, some research designs may require cross-sectional or panel data, while others need network data. These requirements have implications for the extraction and storage of web data. Researchers need to be mindful that *how they collect the data* may invalidate their research design. For example, consider a researcher who wishes to build a panel dataset on the performance (e.g., measured using sales rank) of products available at Amazon. Suppose the number of instances (e.g., products) is very large. In that case, the data collection likely captures instances at *different* times of the day (e.g., the first products on the list in the morning, the last products on the list during the evening). If there are systematic differences in sales ranks in the mornings versus evenings, results may be biased. Therefore, in determining the structure of the data and thinking through *how* to retrieve data from the website or API, researchers need to be aware of any potential biases they may introduce and how they can circumvent such issues by design.

Researchers should also consider the feasibility of converting the collected raw data to the desired format for analysis. Specific data formats may lead to data loss, which complicates or even makes the conversion to a format suitable for analysis infeasible. For
example, data retrieved from the Twitter APIs are not normalized (i.e., when seeking to store the data in tabular format, multiple tables are required to prevent data loss).

**Resource Use**

While browsing the web is mainly free, researchers should not assume that collecting web data is costless. While the prototype of a data collection can be ready and running in a matter of hours – especially when using publicly available code or packages – researchers will often find out that the data collection does not work entirely as intended. Extraction tools may have been developed for older versions of a website, which may not work anymore. Using well-documented packages or libraries (e.g., BeautifulSoup, pytrends for Google Trends, tweepy for Twitter) can reduce development efforts. When coding a data collection from scratch, researchers likely have to overcome technical obstacles (e.g., logins, dealing with interactive websites), requiring substantial technical expertise or learning cost. Even if researchers outsource the process of collecting web data to a developer, code audits, data samples, and revisions are crucial as developers may be unfamiliar with data quality standards required for academic research.

Researchers should also not underestimate the operating and maintenance costs of their data collection (e.g., license fees for an API multiplied by the projected time it takes to collect the data; costs for renting computers or storage infrastructure in the cloud). Also, when collecting data over extended periods, the chance that the code will break is relatively high. Yet, with live data collections, it is impossible to go back in time if one realizes too late that the data collection did not work. Using an official API (compared to web scraping) may be a more reliable way to lower one’s maintenance costs. Last, in dynamic firm environments, it is not uncommon that websites or APIs change or disappear forever (e.g., when Facebook acquired game streaming provider Mixer, the Mixer API was made proprietary on short notice). Researchers may want to map out “worst-case scenarios” to
mitigate the risks and estimate potential opportunity costs (e.g., using publicly available data identified in step 1 of the workflow).

**Step 4: Prototype Development**

After evaluating the research fit, researchers build a prototype to assess the technical feasibility of extracting and storing the data for each selected source. In creating the prototype, researchers must choose a software tool for data extraction. The primary choice is between ready-made scraping toolkits (e.g., point-and-click interfaces like Mozenda), packages that require some coding (e.g., the scrapy framework in Python), or self-developed code that interfaces with high-level scraping libraries (e.g., selenium, BeautifulSoup). As the complexity of the data collection increases, self-developed code is desirable. While developing a scraper might require significant time and effort, researchers can actively manage data quality and reproducibility. An important consideration concerning package use is that some packages may be poorly maintained, rendering the data quality inferior. Depending on the research objective and type of data source, researchers face numerous design decisions in directing the extraction software to capture the desired data. We illustrate these decisions for four exemplary research objectives in Table 2.

**Data Extraction**

Often not all the data available on a website or API is required. Hence, researchers must carefully select which entities to capture, the frequency at which to capture them, and which specific data to extract. The process of capturing elements from a website or extracting data from an API (e.g., a product’s price) is called parsing. When parsing data, researchers can leverage numerous packages (e.g., BeautifulSoup for websites and the json library for APIs in Python). By examining the focal website with the browser’s development tools, researchers can inspect the underlying source code to identify attributes, classes, or styles associated with a particular target element (e.g., a product review).
Table 2: Prototype Development Examples

<table>
<thead>
<tr>
<th>Type</th>
<th>Spotifycharts.com</th>
<th>Yelp.com</th>
<th>Twitter.com</th>
<th>Chartmetric.com</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data availability</td>
<td>Website</td>
<td>Website</td>
<td>API</td>
<td>API</td>
</tr>
<tr>
<td>Description of data</td>
<td>Daily/weekly lists of the Top 200 most-streamed songs per country on Spotify.</td>
<td>Panel data for Yelp users over time that tracks their number of friends and the review(s) they write.</td>
<td>Access to the database of Twitter.com, encompassing searching for tweets and obtaining metadata on users.</td>
<td>Multi-platform data on music streaming, encompassing performance and metadata on artists and playlists.</td>
</tr>
<tr>
<td>Exemplary research question</td>
<td>How did changes to promotions on the platform affect music consumption?</td>
<td>How does the number of a user’s friends affect the usage of emotional language in reviews?</td>
<td>How does online discourse evolve around the occurrence of brand activism?</td>
<td>Are digital music platforms biased towards the content of some of their major suppliers?</td>
</tr>
<tr>
<td>1.</td>
<td>Among the available entities, which ones to extract, at what frequency?</td>
<td>Obtain lists of the Top 200 streamed songs across all countries for the period 1 January 2019 – today.</td>
<td>Generate a list of usernames (reviewers) for data capture: extract, e.g., weekly, their characteristics and reviews.</td>
<td>Obtain tweets and corresponding metadata for users that engage in conversations around brand activism.</td>
</tr>
<tr>
<td>2.</td>
<td>Which data to capture for each entity, and which seeds and sampling procedure to use?</td>
<td>Download CSV files from the website. Obtain all the available data across all available countries.</td>
<td>Friend count and all reviews for each panelist. Seeds are panelists that have evaluated any of the most-reviewed businesses in a city.</td>
<td>Complete JSON API response (i.e., all the tweet’s metadata). Use as a seed list of hashtags related to brand activism. Store full JSON response of API calls; full sample from all playlists tracked by the data provider.</td>
</tr>
<tr>
<td>3.</td>
<td>How to navigate on the website or API, both within and across entities?</td>
<td>Single-entity capture: within-entity navigation via country-and date-specific URLs.</td>
<td>Multi-entity capture; within-user navigation is required (e.g., pagination for reviews).</td>
<td>Single-entity capture (tweets with nested user characteristics). No within-entity navigation is required. Multi-entity capture (each entity has its endpoint); within-entity navigation by modifying endpoints with playlist ids and pagination.</td>
</tr>
<tr>
<td>4.</td>
<td>How to identify and deal with technical hurdles?</td>
<td>Website is made for data extraction; website appears very stable.</td>
<td>Yelp does not offer an API to extract user reviews. The website appears relatively stable.</td>
<td>Retrieval limits of (academic) Twitter API. Monitor server time-outs and whether retrieval limit is optimally exploited (e.g., 50 requests per minute).</td>
</tr>
<tr>
<td>5.</td>
<td>Which software to use for data capture, and where to deploy it?</td>
<td>Python (self-coded), deployed locally.</td>
<td>Python (self-coded), deployed remotely on AWS EC2</td>
<td>R package, deployed locally</td>
</tr>
</tbody>
</table>

Notes: The table illustrates, for four exemplary data sources and research questions, the various choices that researchers have concerning (1) data extraction and (2) data storage and deployment.
When extracting data from APIs, researchers can benefit from outputs that are more structured than the source code of a website. Most APIs provide their data in the JSON (JavaScript Object Notation) format, consisting of a tree of attributes and values. Once parsed, information can be extracted simply by referring to attribute names. Because data in JSON objects typically is not normalized, it can be complex to store in tables. For example, some attributes may not contain values but lists of values. In that case, researchers have to decide whether to concatenate multiple values or normalize the data into multiple tables instead. While packages exist that automate the process of parsing JSON data, such packages could easily misinterpret the JSON tree structure, leading to data loss or suboptimal storage, rendering the data difficult to use.

Researchers also need to decide how to clean the data for storage. While most of the cleaning is done after the collection, some minimal cleaning can be already done at this stage. Typical steps involve removing HTML tag words, stripping out unnecessary characters (“$”, “,”, non-printing characters), or retaining relevant substrings only. Regular expressions (regex) provide a powerful technique for (pre)processing textual data.

To systematically extract web data, researchers need to define the navigation path of the extraction software. This is analogous to how a user would navigate the site in a web browser. A navigation path needs to be specified both within an entity (i.e., obtaining data on all instances of an entity) and across multiple entities (e.g., navigating from one instance of an entity to one or multiple instances of a related entity). First, when extracting data for a single entity, researchers often need to extract data from multiple pages (e.g., iterating through all pages that list a product’s reviews). Sometimes, this can simply be achieved by slightly modifying the instance’s URL (e.g., incrementing one of its parameters such as the page number, &page=2, &page=3, etc.). However, especially when extracting data from websites, merely modifying these arguments can be insufficient. Instead, many providers
require interaction (e.g., clicking, scrolling, filling in forms). Thus, researchers may need to simulate user behavior to view how content is dynamically loading (e.g., scrolling down to reveal additional posts). Second, researchers interested in obtaining data on multiple, connected entities (e.g., reviews for multiple products) need to map the navigation from one instance of an entity to one or multiple instances of another entity. Finally, researchers need to include pauses when requesting data from websites and APIs to respect the data source’s retrieval limit. For instance, Airbnb uses sophisticated countermeasures to combat data extraction (Kloosterboer 2019), and running over the site’s retrieval limit may lead to extraction software being blocked from accessing the website or API.

Storage and Deployment

Researchers choose whether to store data and deploy their extraction software locally or remotely, depending on their research objective. For example, some researchers may benefit from the scalability of cloud computers (e.g., for collecting larger samples). In contrast, others prefer their local office PC to monitor the progress of the data collection easily. Other researchers may find the automated backup functionality of remote storage providers desirable (e.g., when collecting data over extended periods). Others may prefer local storage due to security and privacy concerns.

Next to considering remote storage and deployment, researchers need to choose whether to store their data in files or databases. Factors affecting this decision are the type of stored data (e.g., structured vs. unstructured data), the projected database size, and the number of computers used to collect the data. Databases provide useful additional functionality (e.g., dynamic adjustments to the data collection for optimizing sample size), but server costs for data transfer, storage, and processing may quickly surge, requiring researchers to extract data at regular intervals. To reduce operating costs and facilitate research reproducibility, using databases after the collection has ended should be avoided.
Further, researchers need to decide *when* to store the data. The simplest way is to directly store the extracted and parsed data, improving speed and minimizing storage. An alternative method is to separate data extraction from data storage. Thus, in a first step, the website’s source code or the API’s JSON objects are stored unchanged in their original format. Then, in a post-processing step, researchers extract the desired information from these stored raw data files. Separating collection and parsing might also help with debugging code, which can be particularly relevant if a website is updated, or an API returns unexpected data. The downside of storing all raw data – even if only temporarily – is that it may breach privacy laws (e.g., when storing sensitive data) or lead to excessive use of (costly) storage.

**Legal issues**

The legality of web scraping is an ongoing and complex issue, which might be perplexing for marketing scholars interested in collecting and using web data. There is no clear consensus about whether collecting web data for scientific purposes is permissible under American and international intellectual and cybersecurity laws. Depending on a researcher’s institution’s web data collection policy, legal support might be necessary before retrieving, storing, and analyzing it. Researchers need to be aware of additional, restrictive national and supranational legal frameworks (e.g., the European Union’s General Data Protection Regulation) that may limit the legality of storing personally identifiable information, including users’ self-chosen usernames that – in theory – could be used to identify persons at a later stage. Simultaneously, novel legislation in some geographic regions may provide researchers with exemptions to collect data for scientific purposes (e.g., Article 3 of the Digital Single Market Directive of the European Union, 2019).

Further, some academic journals such as *Management Science* (Simchi-Levi 2019) have instituted restrictive policies regarding articles using web data. Above all, researchers

---

should obtain legal advice to limit their exposure to legal risks when going forward. As legal experts may not be fully aware of all the steps involved in collecting (e.g., extraction method) and using web data (e.g., aggregation of such data), it is imperative to address aspects such as the purpose and framing of the research objective, the scale and scope of data capture, the characteristics of the data source, the relationship of the researcher with the data provider, and any details on the researcher’s data management and usage. Table W3.4 in Web Appendix W3 provides a template for researchers to prepare such meetings.

**Step 5: Collection and Monitoring**

After carefully planning and prototyping, researchers can begin the actual data collection. It is important to note that, even in this advanced step of the workflow, the data collection process is best considered “work-in-progress.” Thus, researchers need to remain agile and adapt the code where required.

Researchers need to monitor the data collection for errors. Even for one-shot data collections, it helps to generate log files so that errors that occurred during the collection can be detected (e.g., timeouts, hitting an API’s retrieval limit). For example, researchers can log each web request (i.e., URL opened at a website or API call made), the server’s status code, and a timestamp of data collection initiation. For real-time data collections over extended periods, researchers must implement a robust monitoring system to check if the data collection is still running and whether the data extracted is valid. Apart from technical monitoring, we advise researchers to monitor the context in which the site operates. For instance, there might be significant changes to the marketplace or institutional context (e.g., M&A activity, IPO) that could influence the data generating process. Subscribing to the site’s social media and news feeds (e.g., Twitter, Google News), and monitoring its blog for important announcements on feature changes or service failures, all of which can impact the validity of the data collection, is essential.
Finally, researchers may be interested in using job scheduling (e.g., Task Scheduler on Windows, Cron on Mac and Linux), allowing them to start up and shut down data collections on any desired schedule. On the one hand, this can improve data quality (e.g., by always starting up a data collection script at the same time on a day), or open opportunities of dealing with technical difficulties when extracting data (e.g., introducing some random variation in the data when it is collected). We list the important steps required in monitoring data quality and debugging data collections in Table W3.5 in Web Appendix W3.

**Step 6: Preprocessing, Documentation, and Distribution**

Researchers can benefit from preprocessing and documenting the data, with essential efficiency gains for themselves and their coauthors and the academic community at large (see Table W3.6 in Web Appendix W3 for a summary). Coauthors can check for the exact operationalization of variables by checking the documentation, and the data collection can potentially be extended should the revision process surface any concerns. Perfectly written or elegant code is unnecessary; instead, it needs to be good enough to execute the task and annotated sufficiently for others to understand. The research community also benefits from published and annotated code, which may lead to citations and future projects – a potential side benefit for authors interested in growing their area of investigation.

**Preprocessing**

While researchers might have already preprocessed some of the data “on-the-fly” during data collection, web data tends to involve comprehensive data cleaning and transformation efforts before the data can be analyzed. Cleaning and transforming web data is relatively more laborious and time-consuming than cleaning data gathered via conventional methods such as surveys, experiments, or scanner data (Adjerid and Kelley 2018).

Web data generally tends to be relatively unruly and messy. Thus, researchers need to carefully inspect the data to tag erroneous entities (e.g., that do not represent the targeted
entities) or entries with impossible values. It is also vital to examine the data for duplicate entries or faulty encodings (i.e., when characters or emoticons are not displayed correctly). Problematic data points (e.g., blank reviews, system messages) can be tagged for further investigation. When working with personal and/or sensitive data (e.g., usernames), it is crucial to anonymize or pseudonymize such data early on.

It is essential that researchers rigorously validate the data to ensure that what was intended to be collected was collected. For example, instances such as users originally part of the seeding list may have stopped using the service during the data collection, leading to (panel) attrition. Screening log files and status messages can pinpoint issues of the collected data (e.g., interruptions). It is usually helpful to skim through the raw data to spot potential areas of concern (e.g., special characters, new variables captured).

Finally, researchers that collect data over extended periods can benefit from generating automatic reports to assess data quality. Such reports (e.g., by using RMarkdown) can easily be rerun as new data is added. Reports can also be used to discuss (new) data with coauthors and report characteristics of the raw and cleaned datasets in a manuscript.

**Documentation**

Documenting the data for potential users (including the researcher who collected the data) is essential for efficiency and reproducibility reasons. The documentation should cover, among others, details about the data composition (e.g., entities, seeds), collection process (e.g., annotated code, detected errors during the data collection), and preprocessing details. Given that changes in the data source and context are inevitable, documenting the data collection's institutional background (e.g., screenshots, the API documentation, or articles from the firm’s blog) is crucial. To start documenting data, we recommend researchers start from the template in Gebru et al. (2020).
Distribution

Researchers need to think about how to distribute their data – foremost internally with team members working on the project. A practical solution may be to share the final data set and its documentation via easily accessible cloud services (e.g., Google Drive, Dropbox). A better, more stable way may be to upload the data and its documentation to a server that coauthors can programatically integrate into their code for further processing and analysis.

Another crucial component worth sharing is the code used for collecting and preparing the dataset. Some researchers use publicly available GitHub repositories to host their code. These repositories are well suited for journal submission or archival purposes. Many institutions and funding organizations require researchers to store data so that the study results can be reproduced. As collecting web data can be a laborious and costly process, researchers may wish to reuse their data in different projects (Adjerid and Kelley 2018). At the very minimum, authors should disclose that they are reusing data during the review process (e.g., in the letter to the editor to maintain confidentiality). The final manuscript should explicitly reference the relevant source in the data section (e.g., Mogilner et al. 2012).

ENHANCING THE ROBUSTNESS, REPLICABILITY, AND GENERALIZABILITY OF RESEARCH BASED ON WEB DATA

In the workflow for collecting web data, we have outlined the design decisions that lead to more credible and compelling research. Following LeBel et al. (2018), we now consider common challenges and solutions regarding the analytical robustness, effect replicability, and generalizability when working with web data. The goal of this section is not to propose novel solutions to existing problems. Instead, we highlight appropriate methods, techniques, and practices from the different subdisciplines relevant for any research based on
web data, regardless of methodological training. Web Appendix W4 provides a checklist to facilitate consistent evaluations of web data in the peer-review process.

**Implications for Analytical Robustness**

Research findings can be considered robust when they emerge with sufficient consistency across other plausible, relevant data-processing and data-analytic decisions in the same dataset (LeBel et al. 2018; Steegen et al. 2016). To improve their findings' robustness, researchers can benefit from generating model-free evidence on the focal phenomenon. Given the richness of web data (e.g., high-frequency data with many observations and many variables), such evidence can be provided at higher aggregation levels (e.g., across all instances, rather than for individual cases). Similar to the widespread practice of reporting model-free evidence in empirical marketing research, researchers can use summary tables and plots for reviewers and readers to better gauge how easily the central effects are observed in the real world (van Heerde et al. 2021).

Researchers should pay close attention to collecting and including relevant variables to increase their findings' comparability to related research, especially when collecting data from the same data source. Researchers may consult more general discussions of the inclusion (or exclusion) of control variables (e.g., Bernerth and Aguinis 2016). Providing a clear rationale for the inclusion of control variables and their operationalization is essential for accurate inference and meaningful effect sizes (Sturman et al. 2021).

In modeling web data, researchers should be concerned about the relationships between different entities (e.g., reviews, consumers, brands). The pervasiveness of temporal links and sequences between them is illustrated by recent work showing that even the very first review a product received influences its average rating three years later (Park et al. 2021). Entity linkages might also lead to common group effects (e.g., social influence on
social networks) or time dependencies (e.g., serially correlated errors). A careful modeling approach needs to be mindful of such concerns (e.g., clustered standard errors).

Making robust inferences based on web data requires researchers to be aware of potential endogeneity concerns (Ebbes et al. 2017; Rutz and Watson 2019). One of the most common causes of endogeneity in web data is omitted variable bias. Our main workflow highlights how capturing rich raw data can enable researchers to capture additional covariates to minimize such bias. Self-selection bias is a particular case of omitted variable bias (Heckman 1979), wherein studied entities (e.g., consumers) 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 et al. 2016). Being aware of potential biases and addressing them with appropriate methods (e.g., propensity score matching, synthetic control method) is essential for credible research findings. Another common cause of endogeneity in web data is simultaneity, which occurs when the independent and the dependent variable are measured at once (Antonakis et al. 2010). To address such concerns, researchers can exploit features of the social or institutional context (e.g., Shriver et al. 2013).

Finally, given the numerous, often undisclosed choices in how exactly the data was collected and processed, it is important to demonstrate that the central findings are robust. Transparency about these choices is a prerequisite for cumulative science and progress. Importantly, it is unreasonable to expect that a focal effect emerges in every single robustness analysis. Instead, it needs to appear consistently enough that the core findings can be considered valid. Compelling visualizations (e.g., via multiverse analyses; Patel et al. 2015; Steegens et al. 2016) and clear tables that summarize robustness checks increase the accessibility of the study substantially.
Implications for Effect Replicability

An effect is considered replicable if observed consistently in new samples with similar magnitudes when using the same method. For instance, a focal effect is replicable if it emerges in different samples (e.g., other data sources or different industries) using the identical or a highly similar analytical approach (e.g., Tang et al. 2014).

To allow other researchers to replicate findings based on web data, researchers need to spell out the assumptions about the data generating process (e.g., choice of seeds) that would predict when the focal effect should or should not emerge in (newly) collected data (see also Landers et al. 2016). Additional descriptive statistics (e.g., percentiles) help reviewers and readers to understand sample composition better. Proving additional contextual information (e.g., proportion of different cuisines in a Yelp sample, information about the level of user engagement in a sample) can facilitate conducting and evaluating replications of published research based on web data. For instance, the failure to replicate a published effect might, at least partially, result from differences in seeding or variations in sampling (e.g., Chevalier et al. 2018; Wang and Chaudhry 2018).

Implications for Effect Generalizability

A replicable effect is generalizable if it emerges across distinct methodologies and distinct populations (LeBel et al. 2018). For example, an effect based on web data would be considered generalizable if it occurred in research using a highly dissimilar method (e.g., a survey) in a different sample (e.g., consumers from a different culture). Typically, dissimilar methodological approaches use different operationalizations of the independent and dependent variables. While generalizability is the gold standard for credible research findings, it is unreasonable to expect every single finding to be generalizable.

To set realistic expectations of an effect’s generalizability, researchers need to outline the specific constraints on the generalizability of their findings (Simons et al. 2017). This
implies outlining the contextual factors of their web data, such as the website’s data entry procedures (e.g., Lewis 2015), and specifying how these contextual factors might influence or limit the generalizability of the focal findings. Given the idiosyncratic nature of their user bases, certain web data might even prevent making strong generalizable claims. To increase the generalizability of conclusions based on web data in multi-method marketing research, authors should closely tie the web data to subsequent experiments. Especially in an initial experimental study, the dependent variable should be measured in the same manner as it was generated on the web. For example, review helpfulness could be measured by a binary vote (i.e., helpful vs. not helpful) as in the collected web data (e.g., Amazon, TripAdvisor), rather than using multi-item Likert scales.

As datasets based on web data can have very large sample sizes, most examined relationships will be significant at standard $p$-values. To set realistic expectations about effect replicability and generalizability, authors should report effect sizes and establish their findings' practical impact rather than merely demonstrating statistical significance (Adjerid and Kelley 2018; van Heerde et al. 2021).

**FUTURE OPPORTUNITIES FROM WEB DATA IN MARKETING RESEARCH**

The use of web data in academic research projects has increased in recent years. This, however, does not mean that the potential of web data has been fully realized yet. Web data has the potential to challenge the boundaries of what we study and how we study it. In the following, we discuss both underutilized and unexplored opportunities for using web data for marketing research.

**Challenging the boundaries of what we study**

Due to the increasing presence of rich metadata (e.g., geolocation data; Huang et al. 2016), collecting web data facilitates the discovery of variation across geographic or
sociopolitical contexts of theoretical significance (Barnes et al. 2018). Web data potentially provides cost-effective access to data on diverse populations of entities worldwide (Kosinski et al. 2016). Hence, web data can be used to move beyond the typical WEIRD (i.e., Western, educated, industrialized, rich, and democratic; Henrich et al. 2010) samples often used in the social sciences. For example, Kübler et al. (2018) collected a global panel dataset from the Apple App Store. Using daily data from 60 countries, the authors explored how various cultural, economic, and structural differences between countries moderate the effect of price and ratings on app popularity. Collecting more diverse samples from the web can increase confidence in the generalizability of findings (Rad et al. 2018) and their impact on marketing practice worldwide.

Many exciting opportunities arise from collecting web data to generate realistic stimuli for more controlled tests of predictions (e.g., via experiments). Specifically, researchers may use web data to move beyond artificially constructed stimuli designed explicitly for producing predicted effects (van Heerde et al. 2021). Researchers may collect large sets of brand logos (Luffarelli et al. 2019), review rating distributions for different hotels (Fisher et al. 2018), or crowdfunding profiles (Genevsky and Knutson 2015) to be subsequently used as stimuli in experiments or neuroimaging studies. Creating larger, more representative samples of stimuli helps to ensure observed effects are widely generalizable across stimuli rather than the result of idiosyncratic stimuli (e.g., Judd et al. 2017).

Systematically extracting web data allows marketing scholars to be discovery-oriented and scout out emerging phenomena. Though our workflows might appear less relevant to non-deductive marketing research, we believe that this methodology provides enormous research opportunities focused on inductive theory-building. While various qualitative approaches such as netnography leverage web data (e.g., Kozinets 2002), these approaches tend to rely solely on manual extraction of web data. We could only identify only
one netnographic study in marketing that used web scraping or APIs (Arvidsson and Caliandro 2016). Yet, systematically extracting web data might hold enormous potential for theorizing about emerging phenomena at higher levels of analysis, such as a brand public in the case of Arvidsson and Caliandro. Rich consumer narratives on blogs and access to online communities from idiosyncratic samples can be fruitful bases for generating novel and relevant marketing theories (e.g., Ertimur and Coskuner-Balli 2015; McQuarrie et al. 2013).

Finally, another exciting avenue for future research using web data is to examine phenomena that cannot be studied unobtrusively with conventional methods. Because researchers often collect the behavior after it naturally occurred (Hoover et al. 2018), it can avoid some of the common challenges encountered in studying such phenomena with conventional methods (e.g., social desirability concerns). For instance, Chen and Berger (2013) collected data from a discussion forum to explore how controversy influences participation in online discussions. Another promising avenue is to record behaviors that firms prefer not to disclose, such as the usage of tracking tools on websites (Trusov et al. 2016) or the engagement in illicit behaviors like affiliate fraud (Edelman and Brandi 2015). There are many other phenomena that, given the outlined advantages, can benefit from web data. Some examples include studying racial biases (Jaeger and Sleegers 2020), the market for fake online reviews (He et al. 2021), or marketplaces hidden to the public eye (e.g., on the Dark Web; Thomaz and Hulland 2021). We note that many of such topics are likely to require ethical review board approval and careful consideration of potential legal risks.

**Leveraging technical possibilities**

Next to challenging the boundaries of what we study, the evolving technological landscape allows researchers to unleash the full potential of web data. Most collected web data in marketing research is historical (i.e., behavior that occurred in the past). Only a few researchers engage in real-time data collections, especially over extended periods (e.g.,
months or even years). Yet, many websites and APIs provide much richer data in real-time, compared to historical data available elsewhere on the platform. Real-time data collections are also suited to capture naturally occurring field experiments (e.g., Chen et al. 2011).

Field experiments are increasingly prized for their realism and high ecological validity (van Heerde et al. 2021). Many exciting opportunities exist for conducting online field experiments in which researchers capture real-time user data via APIs. Doing so can produce more precise estimations of effect sizes and allow researchers to capture long-term effects (Gneezy 2017). Researchers can randomly assign participants to different treatments, such as adding followers on Twitter (Toubia and Stephen 2013) or assigning Reddit’s Gold Awards to user posts (Burtch et al. 2021). By gathering fine-grained (e.g., daily) data via APIs, researchers can analyze how experimental treatments influence focal behaviors such as the creativity of user-generated content.

Thus far, marketing researchers have primarily used APIs as a method to retrieve data. However, many firms have begun making cutting-edge *algorithms* available for public use via their APIs. For example, the Google Cloud Vision API provides researchers access to the Artificial Intelligence (AI) algorithm to recognize objects and persons on images. Researchers may also use sophisticated speech-to-text algorithms to integrate audio data in their studies. We encourage researchers to scan the environment for relevant novel and well-documented APIs that may add a unique edge to their research project.

A final opportunity arises from providing third parties with “research APIs.” Firms can integrate such APIs into their existing computational infrastructure. Researchers, in turn, can collect valuable information on marketplace behavior. For example, Kruijswijk et al. (2020) have launched an API for developing and testing bandit policies. Firms can use this API, for example, to test pricing decisions or optimize advertising delivery.
CONCLUDING THOUGHTS

The web is a potential goldmine for research in marketing. However, extracting valid data for generating relevant and credible research insights is challenging. Our article provides researchers with guidance on the crucial design decisions of collecting and using web data for academic research.

By introducing a workflow for collecting web data, we seek to bridge entrenched training silos (e.g., between quantitative marketing and consumer behavior) and offer the potential for value-adding collaborations across methodological traditions. Fully exploiting web data may help the discipline further enhance its relevance and assert intellectual leadership on important marketing issues that would otherwise be increasingly studied in disciplines such as computer science, information systems, and management science (Moorman et al. 2019a).

Web data democratizes data access and makes our discipline more inclusive for scholars who would otherwise find it difficult to obtain access to data (e.g., early career researchers). Given the potential of web data being the foundation for multiple, independent articles on important issues, our field needs to recognize and fully embrace its value. At present, only a single database submission based on web data has been published in a premier marketing journal (Wang et al. 2014). At the discipline level, we need to consider creating incentives and journal space for rich web data-based datasets and their documentation. At the author level, researchers should consider making algorithms or data available for other researchers via creating APIs (e.g., Kamvar and Harris 2011), inspiring the adoption of their methodologies and use of derived data by others.

In summary, web data presents a golden opportunity for researchers to examine important marketing questions. We hope the approach outlined in the paper is a promising pathway for generating impactful, ground-breaking, and credible research findings.
REFERENCES


Thomaz, Felipe and John Hulland (2021), "Shining a Light on the Dark Web: An Examination of the Abnormal Structure of Illegal Digital Marketplaces,” working paper, Oxford University.


WEB APPENDIX W1: IDENTIFICATION OF MARKETING RESEARCH ARTICLES USING WEB DATA

To identify articles using web data, we employed various search terms related to web scraping and application programming interfaces (APIs) on Web of Science, Google Scholar, and journal websites. We particularly focused on search terms associated with collecting web data (e.g., scrap*, crawl*, API, application programming interface) or related to specific collected platforms (e.g., Yelp, TripAdvisor, Twitter, Facebook, eBay). We iteratively expanded the list of search terms based on our inspection of the articles included in Figure 1 in the main paper (for instance, by adding category-specific search terms like “BoxOfficeMojo” or “crowdf*”) until no further articles could be identified. We include all marketing research articles published in the Journal of Marketing, Journal of Marketing Research, Marketing Science, Journal of Consumer Research, and Journal of Consumer Psychology. Sometimes, it is impossible to identify whether the authors actually collected the data using web scraping or APIs or manually copied it instead. To be as comprehensive as possible, we include all articles for which the data could have been collected automatically through web scraping and APIs.
## WEB APPENDIX W2: TUTORIALS COVERING THE TECHNICALITIES OF COLLECTING WEB DATA

<table>
<thead>
<tr>
<th>Source</th>
<th>Extraction method</th>
<th>Programming Language</th>
<th>Main software and packages</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ciechanowski et al. (2020)</td>
<td>Web scraping</td>
<td>N/A</td>
<td></td>
<td>An introduction to various non-coding alternatives for gathering and analyzing web data.</td>
</tr>
<tr>
<td>Cushman Garland (2020)</td>
<td>Web scraping</td>
<td>Python</td>
<td>beautifulsoup</td>
<td>An entry-level class that covers the Python fundamentals, command line, and beautifulsoup.</td>
</tr>
<tr>
<td>Devlin (2020)</td>
<td>API</td>
<td>Python</td>
<td>requests</td>
<td>Beginner-level tutorial on how to retrieve data from APIs in Python.</td>
</tr>
<tr>
<td>Han and Anderson (2021)</td>
<td>Web scraping</td>
<td>Python</td>
<td>selenium</td>
<td>Concise discussion of extracting data from dynamic websites (e.g., Tripadvisor) with example code.</td>
</tr>
<tr>
<td>Jaeger et al. (2020)</td>
<td>API</td>
<td>R</td>
<td>htr, base64enc</td>
<td>An introduction to APIs (Face++ and Kairo) for extracting demographic information from facial images.</td>
</tr>
<tr>
<td>Keegan (2019)</td>
<td>API + Web scraping</td>
<td>Python</td>
<td>beautifulsoup</td>
<td>Tutorials for retrieving data from the web, intended for researchers in the social sciences and humanities with limited programming experience.</td>
</tr>
<tr>
<td>Landers et al. (2016)</td>
<td>Web scraping</td>
<td>Python</td>
<td>scrapy</td>
<td>Tutorial with supporting website, focusing on the Python package scrapy (<a href="https://rlanders.net/scrapy">https://rlanders.net/scrapy</a>).</td>
</tr>
<tr>
<td>Markham (2017)</td>
<td>Web scraping</td>
<td>Python</td>
<td>beautifulsoup</td>
<td>Introductory-level web scraping tutorial using the requests and beautifulsoup libraries.</td>
</tr>
<tr>
<td>Munzert et al. (2015)</td>
<td>Web scraping</td>
<td>R</td>
<td>RCurl, stringr, etc.</td>
<td>An extensive treatment of scraping and processing web data with R.</td>
</tr>
<tr>
<td>Murphy (2017)</td>
<td>API</td>
<td>R</td>
<td>rtweet</td>
<td>An accessible tutorial for extracting psychological data from the Twitter API.</td>
</tr>
<tr>
<td>Pascual (2020)</td>
<td>Web scraping</td>
<td>R</td>
<td>rvest</td>
<td>Tutorial that covers the basics of how to do web scraping applied to obtaining weather data.</td>
</tr>
<tr>
<td>Paxton and Griffiths (2017)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>Not a tutorial, but an excellent summary of existing web datasets, including a companion website (<a href="https://dataonthemind.org">https://dataonthemind.org</a>).</td>
</tr>
<tr>
<td>Russell and Klassen (2019)</td>
<td>API + web scraping</td>
<td>Python</td>
<td>twitter, geopy, facebooksdk, etc.</td>
<td>Detailed introduction to using various APIs to collect data from various social networks.</td>
</tr>
</tbody>
</table>
WEB APPENDIX W3: MANUALS AND CHECKLISTS THAT ACCOMPANY THE WORKFLOW FOR COLLECTING WEB DATA

Figure W3.1: Data Source Exploration

Notes: The figure visualizes, for a few selected data providers, their positioning along two dimensions: the type of the data provider (i.e., a primary data provider with access to detailed, single-platform data, versus a data aggregator with access to less detailed, multi-platform data), and the data provider’s scale or scope (e.g., a provider with a relatively large user base, versus a data provider with a relatively small user base). For example, Etsy and Amazon are both primary data providers, but Amazon has a large user base and global coverage, while Etsy has fewer users with a more regional focus. Both the Twitter API and SocialBlade provide data on many users. However, SocialBlade is a data aggregator (i.e., offering data on multiple platforms such as Twitter, YouTube, and Facebook), while Twitter is a primary data provider (i.e., offering data only about its own platform).
### Table W3.1: Data Availability Assessment

<table>
<thead>
<tr>
<th>Issue</th>
<th>Examples</th>
<th>Reason for importance of issue</th>
<th>Common pitfalls and complexities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.</td>
<td>For which entities is data available?</td>
<td>Social media: users, posts, followers, usage, and consumption. E-commerce: Products, reviews, reviewers, sellers, manufacturers, brands. Entertainment: movies, albums, tracks, plays.</td>
<td>Incorporating a wider set of entities or entities that have not been used in prior research can lead to novel predictions and the exploration of new phenomena.</td>
</tr>
<tr>
<td>1.2.</td>
<td>How many instances of an entity are available, and can they all be retrieved?</td>
<td>Amazon displays an approximate count for the number of products per product category (e.g., “more than 50k”) but allows users of their site to only view 400 product pages with 25 products each – imposing a “hard” retrieval limit.</td>
<td>Retrieval limits may severely limit the sample size and hence affect the research design. Sampling from a site may require access to a complete set of instances of a particular entity from which to draw a random sample.</td>
</tr>
<tr>
<td>1.3.</td>
<td>How are instances of an entity identified (IDs)?</td>
<td>At Amazon.com, products are identified by the Amazon Standard Identification Numbers (ASINs).</td>
<td>Consistent identification is essential when collecting data from the same website more than once. It is also necessary for connecting different entities (e.g., reviewers to reviews).</td>
</tr>
<tr>
<td>1.4.</td>
<td>How are entities linked to one another?</td>
<td>For Amazon, products are linked to reviews, reviews are linked to reviewers, community posts are linked to reviewers, sellers are linked to products, and brands are linked to products.</td>
<td>Some instances of an entity cannot be sampled directly on the site, such that the researcher needs to navigate between entities. Linkages are also essential to record in the raw data and when making choices on the data’s storage format.</td>
</tr>
<tr>
<td>1.5.</td>
<td>Can entities be linked to external entities?</td>
<td>Amazon displays International Standard Book Number (ISBNs) for books, and Musicbrainz.org shows Universal Product Codes (UPCs) for music albums. When scraping in real-time, a researcher could record dates and timestamps to allow for temporal matching or observing changes in variables.</td>
<td>External identifiers allow the researcher to combine data from different sources into a unique combination of data, offering publication potential.</td>
</tr>
<tr>
<td>1.6.</td>
<td>Which lists or views could serve as a starting point for the data collection (“seeds”)?</td>
<td>Category overview pages in Amazon (internal seeds); New York Times Bestseller List for books (external seed, when scraping books on an e-commerce site).</td>
<td>Access to seeds can facilitate collecting the entire population or sampling from the relevant population (e.g., via assembling a list of products per category).</td>
</tr>
</tbody>
</table>
Table W3.1: Data Availability Assessment [continued]

<table>
<thead>
<tr>
<th>Issue</th>
<th>Examples</th>
<th>Reason for importance of issue</th>
<th>Common pitfalls and complexities</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.</td>
<td>For what period is data available?</td>
<td>On most e-commerce sites, reviews for products are available historically. Google Trends web search data can be retrieved from January 2004 onwards.</td>
<td>It is vital to identify whether a real-time data collection (vs. historical data collection) may lead to better data. Timestamps in historical data need to overlap with other datasets used in merging.</td>
</tr>
<tr>
<td>2.2.</td>
<td>How (accurate) is time encoded?</td>
<td>Timestamps are typically provided in the time zone of the user who is browsing the site.</td>
<td>Ensuring that the timestamp is well understood is crucial (e.g., a wrong merge can happen when linking data from different time zones.</td>
</tr>
<tr>
<td>2.3.</td>
<td>Can data be modified after it has been published?</td>
<td>Users can update their reviews on Yelp; anyone can modify a Wikipedia article.</td>
<td>Realizing that data is modified (e.g., Amazon deleting fake reviews) might warrant a real-time data collection and can inspire asking novel research questions (e.g., He et al. 2021).</td>
</tr>
<tr>
<td>2.4.</td>
<td>How often is the data refreshed?</td>
<td>The Google Trends API provides new data that refreshes and changes daily. Chartmetric updates data on playlists daily, but only for relatively popular lists.</td>
<td>If data is not updated frequently, gaps in the data may cause problems in modeling. Aggregation or interpolation may become necessary.</td>
</tr>
<tr>
<td>3.1.</td>
<td>Which mechanisms (e.g., algorithms, design choices) affect the display of data?</td>
<td>A site like Amazon applies numerous algorithms that affect data display. For example, products are listed by popularity on the category overview pages, or parts of the product pages may be personalized.</td>
<td>Algorithms might pose significant challenges to estimating causal effects (e.g., researchers may misattribute the effects of recommendation algorithms as intended consumer choice) or internal validity (e.g., when a site changes its algorithms based on explanatory variables the researcher is interested in).</td>
</tr>
<tr>
<td>3.2.</td>
<td>Is it clear how metrics have been calculated?</td>
<td>How Yelp aggregates data to determine a restaurant’s price tier (i.e., signified by the $ signs) is unclear. Chartmetric provides a proxy for a playlist’s listeners, not an accurate count.</td>
<td>Researchers need to understand how a construct is measured to prevent them from drawing inaccurate conclusions.</td>
</tr>
<tr>
<td>3.3.</td>
<td>Can the researcher exert control over data display?</td>
<td>Yelp offers users many ways to sort reviews (e.g., Yelp sort, newest first, lowest rated). Netflix offers users to opt-out of A/B tests used by the firm.</td>
<td>Relying on sorting algorithms (e.g., Yelp sort, Amazon’s top reviews) reduces the reproducibility of data extraction. Changing the default option may lead to obtaining more suitable seeds.</td>
</tr>
</tbody>
</table>
## Table W3.2: Evaluation of Research Fit and Resource Use

<table>
<thead>
<tr>
<th>Issue</th>
<th>Features</th>
<th>Reason for importance of issue</th>
<th>Pitfalls and complexities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. How can instances of entities be sampled from the site?</td>
<td>Various sampling approaches can be used (e.g., full sampling to obtain data on all instances of an entity, random sampling from a list of all instances, or stratified sampling).</td>
<td>The sampling approach might lead to biased inferences and may warrant the use of selection models. For example, convenience sampling may lead to selecting a non-representative sample, reducing the generalizability of central findings.</td>
<td>If random sampling is not feasible (e.g., lack of access to a full or quasi-random list of seeds), would convenience sampling be sufficient to inform the research question? If past research has used non-random or non-representative samples, better sampling schemes can lead to the discovery of interesting moderators or boundary conditions.</td>
</tr>
<tr>
<td>1.2. What sample size is required to test the predictions?</td>
<td>Sample size requirements can vary as a function of the researcher’s main interest for inference (i.e., interested in detecting statistically significant effects), effect sizes (i.e., interested in measuring the effects absolute size), and prediction accuracy (i.e., reducing the standard error of forecasts).</td>
<td>It is important to consider the minimum sample size and recognize any technical constraints to satisfy statistical power requirements. A sufficiently large sample ensures that null effects are informative.</td>
<td>Bigger is not (always) better. Very large samples can magnify biases resulting from other study design problems. Large sample sizes can negatively impact resource efficiency (e.g., the time needed to collect data or run models).</td>
</tr>
<tr>
<td>1.3. What is the technically feasible sample size?</td>
<td>A study’s technically feasible sample size is a function of the desired sampling frequency, the retrieval limit of the data source, the number of URL calls to retrieve data for each instance of an entity, and the number of computers/API tokens used. Details are available in Table W3.3.</td>
<td>What makes sampling web data challenging is that one needs to consider how much of the data can actually be retrieved and how this affects subsequent analytic choices. Researchers need to determine whether the technically feasible sample size is sufficient for testing the proposed model.</td>
<td>The effective sample size can severely vary, even with minor changes to any of the input parameters. One can solve for any of the other target parameters (e.g., to calculate feasible extraction frequencies).</td>
</tr>
<tr>
<td>2.1. Can the constructs of interest be measured with the available data?</td>
<td>The construct needs to be transparently defined (i.e., it needs to be clear how it is computed from the researcher's raw data or how the data provider has operationalized it).</td>
<td>It is critical to ensure the raw data on a website/API is a valid operationalization of a construct. If it is unclear how a metric is computed, it will be a hurdle to convince reviewers about its validity.</td>
<td>Is the measure capturing the theoretical construct (e.g., Google search can be attention, interest, etc.)? How comparable are measures from different sources (e.g., are Amazon’s helpfulness votes and Yelp’s useful votes capturing the same construct)? Potential trade-offs may arise between automated (efficiency, feasibility) versus human processing (accuracy) when constructing measures (Hartmann et al. 2019; Xu et al. 2020).</td>
</tr>
</tbody>
</table>
### 3) Data structure and preparation

<table>
<thead>
<tr>
<th>Issue</th>
<th>Features</th>
<th>Reason for importance of issue</th>
<th>Pitfalls and complexities</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1. What data is required to answer the desired research question?</td>
<td>The research question may warrant a particular study design and – resulting from it – a data structure (e.g., cross-sectional data, panel data, time series, network data).</td>
<td>The goal of the research influences the type of data required. Yet, it is not clear that the desired output data can be constructed from the collected data. For example, in the absence of accurate timestamps, building panel data from historical data may be difficult and warrant real-time data collection.</td>
<td>If the target data set is at the daily level, when and how often does one need to visit the site to prepare such data? Does the website’s updating frequency correspond with the level of aggregation?</td>
</tr>
<tr>
<td>3.2. How can the raw data be converted to a dataset that can be analyzed?</td>
<td>Evaluation of which data transformations are needed to convert the raw scraped data to a final data set for analysis (e.g., transposing, sorting). What is the data structure (e.g., rows and columns), what is the required degree of normalization? What is the data storage format?</td>
<td>Certain analyses require data to be in a particular format. Using an inappropriate format for a particular analysis can lead to a substantial lag in research speed. Storing data in an efficient and well-documented manner can ensure both data integrity and faster research time.</td>
<td>Textual or image data utilized for machine learning may require enormous storage space and may even need to be saved and directly parsed/analyzed. Data needs to be stored to remain accessible for various statistical programs (e.g., CSV). Possibility to use SQL and/or cloud computing resources for conversion.</td>
</tr>
</tbody>
</table>

### 4) Resource use

<table>
<thead>
<tr>
<th>Issue</th>
<th>Features</th>
<th>Reason for importance of issue</th>
<th>Pitfalls and complexities</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. What are the development costs?</td>
<td>Building a scraper may cost significant development costs (i.e., time and effort). Legal expertise may be required to obtain clearance for conducting the project.</td>
<td>Researchers may invest substantial research time in developing the extraction software. Recognizing early on what support is required is essential. The legal exposure might impact the potential publication outlets (e.g., Management Science’s scraping policy).</td>
<td>Using well-documented packages/libraries can abate development efforts. Avoid exposure to legal risk by design. Seek counsel with the institutional legal team early on in the process or obtain external legal advice (see Table W3.4).</td>
</tr>
<tr>
<td>4.2. How much will it cost to run the scraper?</td>
<td>License fees for API, costs of deployment, and storage infrastructure.</td>
<td>Running data collections over extended periods may lead to overshooting one’s research budget.</td>
<td>These monetary and non-monetary costs incur even when developing the scraper. Storage should be planned before developing the extraction software. Institutions may offer discounts for using selected cloud providers. Using an institution’s resources may be more cost-effective.</td>
</tr>
<tr>
<td>4.3. How costly is it to maintain the extraction software?</td>
<td>Costs include the likelihood of error (potentially stopping a research project), monitoring the correct functioning of the extraction software and debugging it where necessary.</td>
<td>Collecting wrong or less data over some time can lead to inferior data quality or data loss.</td>
<td>The chance that errors occur is relatively high when collecting data over extended periods. Using an API (compared to scraping data from the service’s website) may be a more durable strategy to safeguard against broken code.</td>
</tr>
<tr>
<td>4.4. Opportunity costs</td>
<td>If one were to buy comparable data from official sources, how much would the data cost?</td>
<td>Data collection is time-consuming and potentially error-prone. Maybe alternative datasets exist (e.g., by the data provider or by researchers).</td>
<td>Based on the research objectives, researchers need to decide between collecting novel data and using preexisting data.</td>
</tr>
</tbody>
</table>
Table W3.3: Calculation of Technically Feasible Sample Sizes

**Panel A. Formula**

The technically feasible sample size – i.e., the sample size that can be obtained from a website or API while considering any potential resource constraints, can be calculated as:

\[ N = \frac{\text{req} \times S}{r \times \text{freq}} \]

Parameters are explained in Panel B below. The formula can be rearranged to solve it for different target parameters (e.g., given the desired sample size, what is the maximum attainable sampling frequency?).

Researchers need to pay attention to convert input parameters to the same time unit (e.g., the retrieval limit may initially be expressed in fifteen-minute intervals but needs to match the desired sampling frequency, which may be expressed in hours).

**Panel B. Parameters and Example**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Sample size (number of instances of an entity)</td>
<td>Target parameter to solve for (solves to 2,250).</td>
</tr>
<tr>
<td>req</td>
<td>Retrieval limit (maximum number of requests per time unit, allowed for each scraper or authenticated API user)</td>
<td>5 requests per second = 5 x 60 x 60 requests per hour (18,000)(^1).</td>
</tr>
<tr>
<td>S</td>
<td>Number of collection scripts (e.g., different instances of the extraction software, or the number of (authenticated) users of an API)</td>
<td>One authenticated data collection script running on one computer.</td>
</tr>
<tr>
<td>r</td>
<td>Number of URL requests necessary to obtain data for one instance</td>
<td>The scraper needs to visit 2 URLs: one to obtain users’ metadata and one to obtain users’ usage history.</td>
</tr>
<tr>
<td>freq</td>
<td>Sampling frequency for each instance and per time unit</td>
<td>Once per 15 minutes = 4 times per hour. Data for each user should be obtained at least once every fifteen minutes.</td>
</tr>
</tbody>
</table>

\(^1\) In practice, the retrieval limit of 5 requests per second may not always be met (e.g., due to server outages or drops in network quality). A researcher may either apply an ad-hoc downward-correction of the retrieval limit (e.g., by 10%) or, instead, estimate the actual retrieval limit by letting the extraction software run for some time and then calculating the average number of actual requests made per time unit.
## Table W3.4: Key issues to raise when obtaining legal advice

<table>
<thead>
<tr>
<th>Theme</th>
<th>Reason for importance of the issue</th>
<th>Potential issues that may affect the researcher’s legal compliance</th>
</tr>
</thead>
</table>
| 1. Purpose and framing of the research objective | Collecting data via web scraping or APIs may violate a firm’s intellectual property or contractual rights. Valid reasons (e.g., including but not limited to a research project’s scientific objective) may exempt researchers from these rules. | • Is the exact research question known/formulated?  
• Is answering the research question societally relevant and urgent?  
• Is the research carried out by members of a recognized research organization (e.g., university, research institute, other organization with a primary goal of conducting scientific research)? What is the role of the person(s) involved in collecting the data within that organization (e.g., student, employee)?  
• Has the research project been approved by the institution’s legal or ethical review board?  
• How much does the study depend on the web data (e.g., is the data entirely based on one scraped data set, or only vaguely uses scraped data for some less-critical control variables)?  
• Is the data provider the only data source with comparable data? Do viable alternatives exist (e.g., feasibility to gather alternative data, including projected resource use)?  
• Are private parties involved in the project (e.g., for knowledge dissemination, as part of an industry collaboration), or is it a purely scientific project?  
• Are the results of the scientific research capable of commercial exploitation? If so, what product will be commercially exploited (e.g., the collected data, the results based on the data, a product developed based on the data)? Who will be exploiting the results for commercial purposes (e.g., the research organization vs. potential corporate collaborators)?  
• Does any party have preferential access to the data, results, or derived product (e.g., an artificial intelligence algorithm trained on the data) of the research project? |
| 2. Scale and scope of data capture | Web data collections vary in scale (e.g., number of instances and entities collected) and scope (e.g., real-time data collection for an unforeseeable time vs. a one-shot data collection within a day). | • What type of data is collected (e.g., on whom, from which location, has permission been obtained from users, has permission been obtained from the website or API – either implicitly or explicitly, does it contain personal or sensitive data? Is obtaining permission feasible?)?  
• What is the scope of the data extraction (e.g., the absolute volume of data extracted, volume relative to all data available on the website; what is the frequency of the data collection (e.g., once, at regular intervals, in real-time); can the retrieval limit be determined and respected)?  
• Which extraction technology is used (e.g., identifying as a data collector, degree of intrusiveness, software toolkit)?  
• How is the data extraction software deployed (e.g., outsourced to a data collector vs. self-administered via commercial tools, Python packages, or entirely self-coded software)?  
• From which location is the data extraction software deployed (e.g., geographical region, part of the research organization’s infrastructure or not)? |
Table W3.4: Key issues to raise when obtaining legal advice [continued]

<table>
<thead>
<tr>
<th>Theme</th>
<th>Reason for importance of the issue</th>
<th>Potential issues that may affect the researcher’s legal compliance</th>
</tr>
</thead>
</table>
| 3. Characteristics of the data source (website/API) | Some types of data providers explicitly make data available for use in scientific research projects. Others merely provide data as a way to grow their business ecosystem (“developer API”). | • Where is the website or API owner located (e.g., registered office, central administration, place of business)? From where is the data extracted (e.g., server location)?  
• What type of data provider is data being extracted from (e.g., scraped from a website, collected from an API; is the data provider *explicitly* offering the data for extraction, or is the data provided as part of a development platform? Is the data source aggregating/collecting data from third parties (e.g., through web scraping or APIs)? Does the data provider have the *right* to display the data?  
• How public is the data access (e.g., public on the web, hidden behind a login screen but publicly accessible for anyone, hidden after login screen only accessible with special permissions)?  
• Is the displayed data personalized or customized in any way (e.g., not personalized, customized for a larger group of persons, customized for a specific user)?  
• Could potential harm be inflicted from using the data, either for the users/entities or pertaining to the firm’s business model (e.g., competition)? What is the risk for users and/or firms if the raw data would become public unintentionally?  
• Does the data provider disallow the use of automated collections, either via robots.txt, the site’s or API’s terms of use, or any other contracts? Have contracts been accepted (implicitly by continuing to use the site/service, or explicitly by signing up for an account with the website or service)?  
• Does the data provider attach a specific license to the use of the data, and what does this license imply for subsequent data use? |
| 4. Relationship of researcher with the data provider | Some of the legal concerns could be resolved if the researcher obtained the data provider’s permission. | • Has the data provider approved the data collection (e.g., implicitly or explicitly)?  
• What are the conditions under which approval has been given (e.g., disguise the identity of the data provider in the paper)?  
• Has the data provider been notified about the data collection (e.g., How? How many times? Has the data provider reacted?)? Would notification be feasible?  
• Are any fees being paid for the data collection (e.g., for an API)? |
| 5. Data management and usage | A researcher may risk violating privacy regulations or other laws when storing, working with, or sharing the data. | • Is the raw data being stored during the collection or directly parsed?  
• How (e.g., file server, file format), why (e.g., reproducibility/verification of research results), and for how long is the raw data stored after the collection?  
• Is the data publicly accessible, are shared with members of a research unit or department? Or is the data only accessible to coauthors?  
• How will the stored data be used? How has the raw data been aggregated? Have entities been disguised (anonymization)?  
• At what level of detail are statistical results reported (e.g., summary statistics, correlation tables, parameter estimates from a model, cross tabs, package with a trained machine learning model based on the data)?  
• Has a license been chosen for the final data set? Is the license compatible with potentially inherited licenses (e.g., creative commons)? |
## Table W3.5: Collecting Data and Monitoring Data Quality

<table>
<thead>
<tr>
<th>Issue</th>
<th>Potential solutions</th>
</tr>
</thead>
</table>
| 1.1. It cannot be verified whether all the data that should have been collected was indeed collected. | • Log each web request (i.e., URL call), along with response status codes, timestamps of when the collection was started, and when the request was made.  
• Save raw HTML websites, along with the parsed data, and compare them.                                                                                     |
| 1.2. The data collection has been interrupted or stopped unexpectedly. | • Verify timestamps in log files or content of parsed data to gauge severity.  
• Try to identify and fix the cause of the interruption.  
• Record issue in a logbook (e.g., in the documentation) and notify potential users of the data.  
• Set up a monitoring tool to timely alert you to any future issues.  
• Move data collection to a more stable environment (e.g., cloud computer rather than local office PC).                                         |
| 1.3. It is cumbersome/takes a lot of time to conduct quality checks. | • Automatically generate reports on data quality (e.g., using RMarkdown).  
• Automate data preparation workflows.                                                                                                                                                       |
| 1.4. The data collection takes place in a highly dynamic environment. | • Monitor the data collection environment (e.g., subscribe to the focal firm’s blog, keep an eye on the websites of related firms, etc., to identify any additional data that needs to be captured). Record any important events in the documentation. |
| 2.1. The data collection breaks frequently. | • Choose more stable data selectors (e.g., CSS is less stable than HTML tag words; “English” class names may better than cryptic class names, etc.).  
• Make use of error handling to avoid interruptions (e.g., try and except in Python), but do not blindly ignore any error.  
• Store raw HTML files, along with parsed data for recovery of data.  
• Consider using an API (over web scraping).                                                                                      |
| 2.2. The data collection is unexpectedly slow. | • Check whether you obey the retrieval limit or fair use policy and implement timers/pauses where possible.  
• Check for traces of being banned/blocked/slowed down by the website (e.g., by investigating the retrieved content).  
• Notify data provider(s) about potential bandwidth issues (e.g., in the case of using a provider’s API).  
• Update the technically feasible retrieval limit, and re-calculate desired sample size, extraction frequency, etc. |
| 2.3. Less data than expected is retrieved. | • Check for any time-outs and erroneous server responses.  
• Verify that the extraction software is suitable for the type of data source (e.g., static vs. dynamic websites).  
• Store raw HTML files of websites or JSON response of an API during the data collection, inspect what data is available in the raw data, and verify it has been parsed correctly. |
| 2.4. Disk space is exceeded, or many files are generated. | • Move data to a remote file location (potentially zip files before uploading them to save bandwidth, e.g., every day).  
• Make use of external databases (e.g., by the university or in the cloud).                                                                                                    |
| 2.5. The computer, which collects the data, crashes. | • Verify whether the computer has had an uninterrupted power supply (and ask University to place the computer on a secure power line or notify you about planned power outages).  
• Move scraping software to the cloud, implement checks that data collection runs.                                                                                                      |
| 2.6. Cloud service provider bills exceed projected costs. | • Verify that computing resources are appropriate (e.g., downscale servers on which collection scripts run, verify that database runs optimally). Consider the costs of data transfer.  
• Verify whether any backup data can be moved to a different location or placed in a glacier for which lower storage costs will be charged.  
• Consider writing data to files after collection, rather than keeping them in an actively running database in the cloud. |
**Table W3.6: Preprocessing, Documentation, and Distribution**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Reason for importance of the issue</th>
<th>Pitfalls and complexities</th>
</tr>
</thead>
</table>
| 1. Cleaning and transformation | Not all the collected raw data may be relevant, valid, interpretable, or there may be legal concerns of processing and storing the data. | - Anonymization or pseudonymization may be necessary for sensitive or personal data (e.g., usernames).  
- Capturing international textual data needs to be stored in a universal encoding (e.g., UTF-8, UTF-16) to avoid data loss.  
- Digits and commas need to be correctly interpreted for the proper storage of numeric data (e.g., the meaning of digits and commas is reversed in some geographic areas).  
- Duplicates need to be identified; deletion or aggregation may be warranted.  
- Any remaining errors in the extracted data (e.g., excessive text) can be cleaned out, for example, by using regular expressions.  
- Choice of column names can be revised. |
| 2. Validation                | Not all the data that was intended to be collected may have been collected, or the definition of measures may have changed. | - Screen log files for any unusual occurrences.  
- Check file sizes or the number of observations at regular intervals.  
- Verify for a sample of the data whether the raw data was correctly parsed. |
| 3. Documentation             | Documentation of the collection process is essential to remain in control of the data – both for the researcher and his/her team members and the public. | - Document the data throughout the whole workflow; starting to document the data only at the end likely is too late to remember all the details.  
- Starting from a template is efficient (e.g., Datasheets for Datasets, Gebru et al. 2020).  
- Include relevant files with screenshots, the API documentation (e.g., as a PDF), sample output, and blog articles that inform users of the data about the context in which the data was collected.  
- Inform potential users about how the data can (or should not) be used. Be transparent about its collection procedure, the composition of the data (e.g., employing automated data reports in RMarkdown), and give notice about how the datasets are maintained and updated.  
- Do not perceive writing the documentation as a burden, but rather as an investment into staying in control of the data collection and use. |
| 4. Distribution              | Making the data accessible to team members is essential for the research project. Considering releasing the data for public use may benefit the discipline. Consider the requirements of publishers or the researcher’s institution for archiving the data. | - Do any legal or ethical concerns prevent the data from being distributed to team members or the public?  
- Which parts of the data can be shared, and which ones not?  
- Which license to use for the published data?  
- Which server/repository to make the data available at?  
- Where to store the data for long-term, archival purposes? (e.g., see re3data.org for a comparison of platforms). |
Table W4.1: Evaluating Research based on Web Scraping and APIs

<table>
<thead>
<tr>
<th>Central Questions That Need to Be Addressed in Submitted Manuscript</th>
<th>Additional (methodological) sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1) Transparency about design choices</strong></td>
<td></td>
</tr>
<tr>
<td>• Is the source of the data clearly described, including the URL of the website or API documentation?</td>
<td></td>
</tr>
<tr>
<td>• Do the authors provide a clear justification for the choice of the data source, vis-à-vis other potential data sources? Is the rationale for the selected data source convincing?</td>
<td></td>
</tr>
<tr>
<td>• Is the data availability for the focal entities (i.e., number of instances, variables that can be captured) sufficient to achieve the stated research objective?</td>
<td></td>
</tr>
<tr>
<td>• Could the entities from focal web data have been linked to external data? If yes, would doing so improve the contribution of the research?</td>
<td>Barnes et al. (2018)</td>
</tr>
<tr>
<td>• If the web data is linked to other data (e.g., using IDs or timestamps), has the matching process been described sufficiently and is it appropriate?</td>
<td></td>
</tr>
<tr>
<td>• Are there flaws in the proposed seeding strategy that invalidate the empirical work?</td>
<td></td>
</tr>
<tr>
<td>• Are the details provided about data anomalies (e.g., non-randomly distributed fake data, modifiable data) and the potential interference of algorithms (e.g., recommendation, sorting, or personalization algorithms) sufficient?</td>
<td>Luca and Zervas (2016); Salge and Karahanna (2018)</td>
</tr>
<tr>
<td>• Is the sample size adequate to inform the research question? Is a larger sample size required and also technically feasible? Do the authors motivate their sample size choice (e.g., by using Table W3.3)?</td>
<td>Adjerid and Kelley (2018)</td>
</tr>
<tr>
<td>• Do the authors provide a clear motivation for the seeds used to gather the data?</td>
<td>Ebbes et al. (2016); Bettis (2012)</td>
</tr>
<tr>
<td>• If there are concerns about the sampling approach, are there any readily available datasets that could be used to replicate the findings?</td>
<td>McAuley (2021); Paxton and Griffiths (2017)</td>
</tr>
<tr>
<td>• Are the extracted variables consistent with the conceptualization of the focal constructs? Is there an explicit description of how the web data was transformed into variables (e.g., screenshots)?</td>
<td>Chen and Lurie (2013, p. 474); Luciano et al. (2018); Xu et al. (2020)</td>
</tr>
<tr>
<td>• Do the authors explicitly outline the assumptions that their data needs to possess to test the predictions? Are these assumptions tested and validated adequately?</td>
<td>Landers et al. (2016)</td>
</tr>
<tr>
<td>• Is there sufficient temporal alignment between the time of data capture and focal processes of interest?</td>
<td></td>
</tr>
<tr>
<td>• Does the selection of control variables appear reasonable (relative to prior research and potentially relevant other website elements not parsed into variables)?</td>
<td>Becker et al. (2016); Bernerth and Aguinis (2016); Sturman et al. (2021)</td>
</tr>
<tr>
<td>• Are there any major red flags regarding the legality or ethicality of the extraction of web data (e.g., using stolen data, hacking into user accounts)?</td>
<td>Kroto et al. (2020); Landers et al. (2016)</td>
</tr>
<tr>
<td>Central Questions That Need to Be Addressed in Submitted Manuscript</td>
<td>Additional (methodological) sources</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>2) Analytical reproducibility</strong></td>
<td></td>
</tr>
<tr>
<td>• Do the authors provide a clear description of when, how, and by whom the data was collected? Is the data accompanied by (publicly available) documentation?</td>
<td>Gebru et al. (2020)</td>
</tr>
<tr>
<td>• Which extraction technology was used? Is the code that generated the data provided (e.g., data collection and data preparation)?</td>
<td></td>
</tr>
<tr>
<td>• Is the navigation path on the site – i.e., entities captured, order of capture – clear?</td>
<td></td>
</tr>
<tr>
<td>• Is the extraction software sufficiently annotated to allow others to potentially reproduce the dataset?</td>
<td></td>
</tr>
<tr>
<td>• Is there an explicit description of how the website elements were parsed into variables (i.e., on-the-fly vs. two-stage)?</td>
<td></td>
</tr>
<tr>
<td>• Do the authors provide model-free evidence? Can focal effects be easily observed?</td>
<td>van Heerde et al. (2021)</td>
</tr>
<tr>
<td>• Are results provided with and without control variables?</td>
<td>Sturman et al. (2021)</td>
</tr>
<tr>
<td>• Is the modeling approach appropriate (e.g., to account for the nesting of the data)?</td>
<td>Kenny and Judd (1986)</td>
</tr>
<tr>
<td>• How pronounced are concerns about self-selection and simultaneity in the web data? Do the authors address potential endogeneity concerns convincingly?</td>
<td>Antonakis et al. (2010); Ebbes et al. (2017); Rutz and Watson (2019)</td>
</tr>
<tr>
<td>• Are the robustness checks relevant and comprehensive? Do they allow for triangulation?</td>
<td>Patel et al. (2015); Simonsohn et al. (2020); Steegen et al. (2016)</td>
</tr>
<tr>
<td>• Does the effect emerge consistently enough across different specifications?</td>
<td></td>
</tr>
<tr>
<td>• Do the authors provide accessible visualizations or tables that summarize the robustness checks? Would it be tractable and beneficial to conduct joint inference across all model specifications?</td>
<td></td>
</tr>
<tr>
<td><strong>3) Analytical robustness</strong></td>
<td></td>
</tr>
<tr>
<td>• Do the authors provide sufficient background information and additional descriptive statistics that allow understanding the sample composition?</td>
<td>Simons et al. (2017)</td>
</tr>
<tr>
<td>• Do the authors outline specific constraints on generalizability of their findings?</td>
<td></td>
</tr>
<tr>
<td>• Do the authors mention that future data would need to possess in order to replicate the effect using the same methodology?</td>
<td></td>
</tr>
<tr>
<td>• Do the authors list procedures (e.g., for the seeds) that influence the generalizability of the finding(s)? If the sample is likely not representative, to which population does it generalize?</td>
<td></td>
</tr>
<tr>
<td>• If the argumentation for the selection of the data source is lacking, would the research benefit from sampling multiple sources of web data to enhance the generalizability of the findings?</td>
<td></td>
</tr>
<tr>
<td>• Do the authors provide relevant effect sizes and illustrate the practical impact of their findings rather than discussing merely statistical significance?</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** The additional sources column includes further methodological sources scholars can consult for more detailed information about the different issues.
REFERENCES


