Unlocking the Potential of

Web Scraping for Retailing Research

Jonne Y. Guyt[1]

Hannes Datta

Johannes Boegershausen

January 2024

Jonne Y. Guyt[1] is an Associate Professor at Amsterdam Business School, University of Amsterdam, The Netherlands (e-mail: j.y.guyt@uva.nl).
Hannes Datta is an Associate Professor at Tilburg School of Economics and Management, Tilburg University, The Netherlands (e-mail: h.datta@tilburguniversity.edu).
Johannes Boegershausen is an Assistant Professor at Rotterdam School of Management, Erasmus University Rotterdam, The Netherlands (e-mail: boegershausen@rsm.nl).

All authors contributed equally. We thank participants at the Special Session of the Retailing SIG at EMAC 2023 for comments on an earlier version. Research support from Tilburg Science Hub (Thierry Lahaye) and tilburg.ai (Marijn Bransen, Jonas Klein) is gratefully acknowledged.

Unlocking the Potential of Web Scraping for Retailing Research

ABSTRACT

Web data has opened new avenues for retail innovations and research opportunities. Yet, despite abundant online data on retailers, brands, products, and consumers, its use in retailing research remains very limited. To spur the increased adoption of web data, we aim to achieve three goals. First, we review existing retailing applications using web data. Second, we demystify the use of web data by discussing its value in the context of existing retailing data sets and new-to-be-constructed primary web datasets. Third, we provide a hands-on guide that allows retailing researchers to incorporate collecting web data into their research routines. Our paper is accompanied by a mock-up digital retail store (music-to-scrape.org) that researchers and students can use to learn web scraping.

Introduction

The retailing landscape and the scope of retailing research are evolving (Gielens and Roggeveen 2023), largely spurred by technological progress. Indeed, the internet has revolutionized the access and exchange of information on products and services, empowered the creation of novel business models, and led to significant retail innovations, generating many novel research areas (Ratchford et al. 2022). A benefit of the digitization of retailing is the large-scale availability of web data on consumers, brands, retailers, and markets. Web data, defined as any data source publicly available on the internet and shown on digital devices (Boegershausen et al. 2022), is uniquely positioned to aid researchers in tackling relevant and novel questions.

However, web data usage has seen a limited uptake in the *Journal of Retailing* compared to other major marketing journals. For example, less than thirty articles have used web data in the last 25 years. Almost all articles using web data focus on textual and numeric data, although other data types exist. Despite large-scale geographic availability, most articles rely on US and, to a far lesser degree, Chinese data. Finally, most researchers using web data relied on web scraping instead of application programming interfaces (APIs).[1]

Our article identifies idiosyncratic challenges in retailing-based research contributing to the slow uptake of web data. Publicly available web data offers retailing researchers numerous opportunities to *augment traditional data sources* or to *compile novel datasets* on trends in the evolving retail sector. Yet, the relative richness of existing proprietary datasets (such as NielsenIQ’s Consumer Panel Data and Retail Scanner Data, GfK’s ConsumerScan, or Kantar’s Worldpanel) has diminished the attractiveness of such alternative sources.

[1] Web scraping is defined as the collection of data by downloading the HTML web page and extracting elements of interest. Application Programming Interface (API) offers programmatic access to company information (Boegershausen et al. 2022).

Existing, well-established proprietary datasets typically offer depth in some dimensions (e.g., sales metrics for established brands and retailers for many years) but often are difficult or costly to obtain and have limited information in other dimensions (e.g., product metadata, tracking of emerging retail formats). In addition, it is unclear how to combine web data with traditional datasets, which often cover past behavior and are released with a significant time lag. Overcoming these challenges requires researchers to align data in time and accurately match many products, consumers, or retailers.

This article intends to demystify the use of web data in retailing research. To this extent, we first substantiate the significance of web data compared to existing archival datasets by reviewing studies using web data published in the *Journal of Retailing*. Building on Boegershausen et al. (2022), we coded key characteristics and data sources from this body of work, highlighting the untapped potential of web data in retail research.

Second, we distill key themes that can facilitate future applications in retailing research using web data. We outline various underutilized web sources and applications to guide researchers embracing web data to (i) improve measures, (ii) increase the diversity of retailing research, (iii) overcome limitations of other methods, and (iv) study emerging retail formats and trends.

Third, to further ease the adoption of web scraping for retailing researchers, educators, and students, we have developed a mock-up digital platform (https://music-to-scrape.org) and offer R code to learn web scraping in a controlled environment. Our guide zooms in on critical stages involving extracting data and building the shells around it: looping (e.g., to extract many products), scheduling (e.g., to build longitudinal datasets), and the infrastructure (e.g., to collect data remotely).

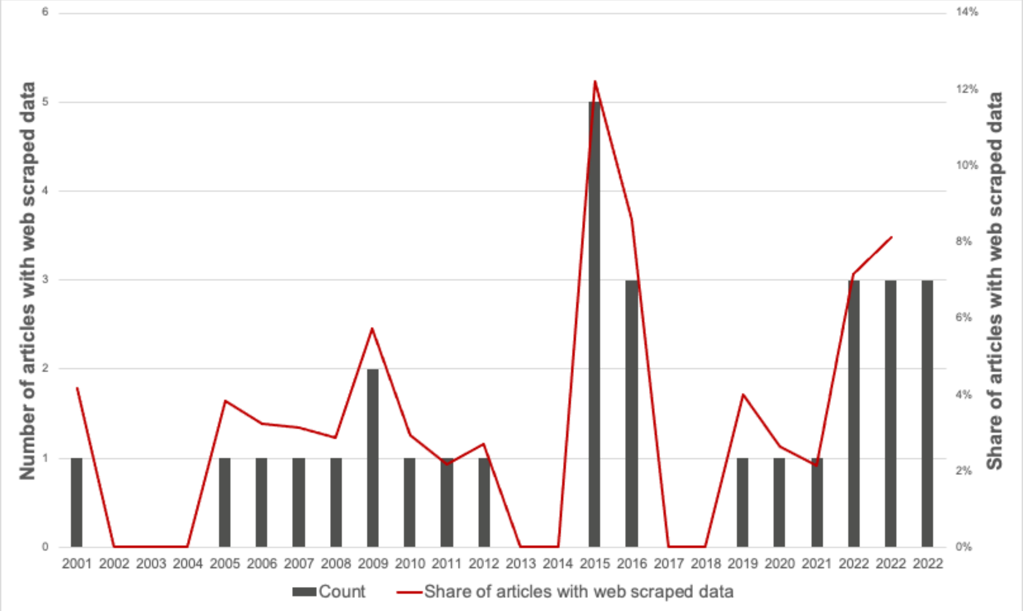
Our article concludes with reflections on important data and source selection issues when studying retail phenomena. Our discussion on the idiosyncratic retailing research considerations includes the usage of web data aggregators, archival versions of websites, and guidance on what data to collect to facilitate the inclusion of control variables and matching with other sources. Finally, we provide an outlook charting pathways for applying Generative AI (Gen AI)/Large Language Models (LLM) and big-team science (Forscher et al. 2023) to empower researchers to kickstart web data collections.

In what follows, we provide an overview of how studies published in the *Journal of Retailing* have used web data (section 2), identify themes that can be used to maximize the benefits of web data for retailing research (section 3), provide an introduction to music-to-scrape.org and the framework for collecting web data (section 4), and discuss data source selection, LLMs, and big team science (section 5). We conclude in section 6.

Web data usage in retailing research

To better understand how web scraping and APIs have been used in retail research, we follow Boegershausen et al. (2022) and identify 28 articles published in the *Journal of Retailing*. We depict the scattered evolution of this research area in Figure 1 (see the Web Appendix for a list of articles and collection details). Next, we leverage our coding to show *how and from where* researchers have gathered *which* information to explore retailing research questions and phenomena. We note that topics explored with web data often remained on the periphery, compared to the "core" retailing research settings and questions (Gielens and Roggeveen 2023).

Figure 1. Web data in the *Journal of Retailing*



Use of web scraping vs. APIs

Retailing researchers primarily rely on web scraping (68%) and, to a lesser extent, on APIs (14%) to extract web data.[2] Similar to the marketing discipline at large (Boegershausen et al. 2022), the most widely used data sources in retailing research are Amazon (*n* = 4) and Google Trends (*n* = 4). For example, Pan and Zhang (2011) collected 41,405 Amazon.com reviews from three categories (i.e., consumer electronics, software, and healthcare products) to explore which factors make user reviews more helpful. Besides e-commerce websites and search engines, researchers have also gathered from auction sites (e.g., eBay, *n* = 5), movie sites (e.g., RottenTomatoes, *n* = 5), and online review platforms (e.g., Yelp, *n* = 2).

Data sources

Half of all articles using web data relied on a single source (*n* = 14). Only two articles leveraged data from many web sources (i.e., ten or more sources). First, Meiseberg (2016) complemented her focal dataset of scraped book reviews from Amazon's German store with

[2] The extraction approach for the remaining articles was either manual or unclear. Additionally, two articles reused existing web data sets (i.e., the Yelp Academic Dataset and SNAP library at Stanford University).

other web data from various German book retailers and authors' homepages. Second, for an early exploration of the nascent e-commerce market, Venkatesan, Mehta and Bapna (2007) gathered 22,209 price quotes for 1,880 products from the websites of 233 different retailers to examine how retailer characteristics (e.g., service quality) interact with market characteristics (e.g., competitive intensity) in shaping online price dispersion.

Geographic coverage

Despite broad accessibility to diverse sources covering retailers and markets worldwide, web scraping research in retailing is highly geographically concentrated. Half of all articles used only US data (*n* = 14), such as data from Amazon.com (e.g., Pan and Zhang 2011) or biddingfortravel.com (e.g., Joo, Mazumdar and Raj 2012). In addition, most scraped datasets rely on sources in the English language (*n* = 22, 79%). The remaining articles feature a combination of English and Chinese-language sources (*n* = 2), Chinese-language sources (*n* = 2), and German-language sources (*n* = 2), omitting the third to sixth most spoken languages in the world (i.e., Spanish, Hindi, Standard Arabic, and French, adding up to 1.5 billion speakers). We did not identify a single article examining web data from retailers and markets in South America or Africa. The overrepresentation of US and English-language sources is at odds with the potential of web data to make research more diverse and less WEIRD (i.e., Western, educated, industrialized, rich, and democratic; Henrich, Heine and Norenzayan 2010b). Even if the textual data is restricted to English due to dictionary-based text analyses, only a single article leverages data from retailers in non-WEIRD countries (i.e., English-speaking reviews of the Dubai Mall; Joy et al. 2023).

Data types

Extant retailing research using web data has focused primarily on textual and numeric data (*n* = 26, 93%), typically studying entities like product sales (e.g., movie box office performance) and online reviews (e.g., review texts, helpfulness votes). Only two articles collected other types of data: Kübler et al. (2023) scraped 97,997 product reviews and the corresponding images from Amazon.com to explore under which conditions images posted by users boost the helpfulness of consumer reviews. Figueiredo, Larsen and Bean (2021) collected images from Yelp to enrich a qualitative dataset about celebrity chef Marcus Samuelsson and his Red Rooster Harlem restaurant. We did not identify any articles collecting video or audio data.

Longitudinal data collections

Only a handful of articles (*n* = 5, 18%) extracted data from one or more sources *multiple times*. An illustrative example of such a dataset is Zhao, Zhao and Deng (2016), who investigated online gray markets for branded products. To explore sellers' and buyers' behavior in gray markets, the authors build a panel dataset based on automatically extracted data about counterfeit Coach handbags from Taobao.com once per week over more than eight months. To identify match products, Zhao et al. leveraged the Coach style numbers from the official Coach websites (i.e., coach.com and china.coach.com).

In the next section, we use our coding to outline promising avenues to maximize the value of web data for retailing research.

Maximizing the benefits of web data in retailing research

Web data can empower retailing researchers to provide *better answers* to existing research questions or study *entirely new research questions* (George et al. 2016). There are numerous retailing research topics for which web data could play a critical role, including showrooming, omnichannel, augmented and virtual reality, spatial computing, dynamic pricing, sharing economy, subscriptions, and third-party platforms. Importantly, as the scope of retailing research is widening (Gielens and Roggeveen 2023), embracing web scraping as an essential part of the methodological toolkit will be essential for retailing researchers given that the traditional, mostly proprietary, and often expensive datasets rarely contain information about these phenomena.

Next, we discuss how *retailing* researchers can maximize the value of web data by outlining how web data (i) can be used to improve existing measures, (ii) can expand the geographic coverage and diversity of retailing research, (iii) allows researchers to study topics that are hard to study otherwise, and (iv) facilitates examining emerging retailing phenomena in a timely fashion.

Web data to collect better measures for existing phenomena

While most examples in our review focused on using web scraped data as the focal or only dataset in retailing research, we next outline some 'quick wins.' These quick wins come in the form of better data (e.g., more granularity, additional control variables) or expanded coverage (e.g., more countries) and provide researchers with better measures and information regarding the generalizability of findings. Numerous APIs allow for the collection of better control variables. For example, in examining cross-national variations across the Indo-Pacific Rim region in market response, such as price and distribution elasticities, Datta et al. (2022) enrich a GfK sales dataset with data about different national holidays using the HolidayAPI. Leveraging this API is particularly practical as the dates of many national holidays (particularly in this region) shift yearly. Several other similar backward-facing APIs may offer retailing researchers the ability to augment existing datasets without major time alignment issues. For example, researchers might need to add county-level data about economic activity to their datasets. However, such data from official sources is often only available with significant time lags. Thus, researchers might draw on APIs like 505economics.com that offer geospatial insights to proxy economic activity (see also, Chen and Nordhaus 2011).

Another fertile use of web data is collecting complementary marketing mix information. Consider research using retail scanner data. A recent review of 493 studies by Lu et al. (2023) suggests that most studies (i.e., 63%) use datasets that only contain *actual* prices rather than regular and discount prices. Heuristics and complex methodological solutions have been used previously to separate out the regular and discount price based on the evolution of the actual price (Fok et al. 2006; Geyskens, Gielens and Gijsbrechts 2010; Lu et al. 2023) but all contain mismeasurement that introduce bias (Lu et al. 2023). While, at times, the data obtained is used as control variables (e.g., Geyskens, Gielens and Gijsbrechts 2010, focus on the introduction of private labels on brand choice), it also plays a central role in other studies (e.g., Guyt and Gijsbrechts 2018, focus on the impact of promotions and discounts). Hence, web data can complement and disambiguate information, resulting in better (control and focal) variables.

Expanding the diversity and geographical coverage of retailing research

Web scraping provides access to data on *diverse* populations of consumers and markets worldwide (Kosinski et al. 2016). Web data enables researchers to move beyond the typical Western, educated, industrialized, rich, and democratic samples often used in marketing and retailing research. Studying diverse populations is important, given that most consumers and retailers are located outside of WEIRD settings (Henrich, Heine and Norenzayan 2010a). Leveraging web data from diverse geographical locations allows researchers to examine consumers, retailers, marketplaces, and platforms exposed to different competitive dynamics. Extending the geographic coverage of retailing research can further increase confidence in the generalizability of research findings (Maner 2016; Rad, Martingano and Ginges 2018) and their impact on retailing practice worldwide. Additionally, rich metadata, including geolocation (e.g., Huang et al. 2016) or device data (e.g., Huang

2019) can facilitate exploring variation across geographic or sociopolitical contexts of theoretical significance (Barnes et al. 2018).

Web data can also boost the diversity of *retailing formats* studied in established markets like North America and Europe. Specifically, in an era of increased geographic mobility and cultural diversity, becoming an entrepreneur remains an important starting point for many immigrants to the US and Europe (Peñaloza and Gilly 1999). Yet, despite an estimated market size of approximately USD 50B in the US, the challenges of ethnic retailers have received scant attention in retailing research. For example, researchers can study strategies allowing immigrants to effectively serve diverse tastes of their own ethnic and non-ethnic customer bases. Likewise, researchers could study how the experience abroad spills over to retailing practices in immigrants' homelands (Balachandran and Hernandez 2021).

Leveraging web data to overcome limitations of established methods

A crucial underutilized benefit of web scraping is its ability to examine phenomena *unobtrusively,* which is difficult with more established methods. Because researchers collect behaviors *after* they naturally occurred (Hoover et al. 2018), web data typically avoids many common challenges in studying such phenomena with experiments or surveys (e.g., social desirability concerns). For instance, Chen and Berger (2013) collected data from an online forum to examine how controversy influences participation in online discussions. Given social desirability concerns with experiments and surveys, web scraping allows researchers to record behaviors retailers prefer not to disclose, such as the usage of tracking tools on their websites (Trusov, Ma and Jamal 2016), engagement in illicit behaviors like affiliate fraud (Edelman and Brandi 2015), or how business activities cause adverse societal outcomes like residential noise complaints(Ozer, Greenwood and Gopal 2024). Web data could also be deployed to study marketplaces and "retailers" hidden from the public eye (e.g., on the Dark Web; Thomaz and Hulland 2021).

In light of the enormous amount and diversity of data available compared to conventional data, web data is also ideally positioned to study relatively rare events (e.g., extremely unusual consumer groups; Bright 2017) and collect the diverse data (e.g., audio, video) necessary to explore the increasingly multimodal communication of retailers across various digital platforms (Grewal et al. 2022). The digital footprints left by retailers and consumers create an enormous volume of data not only in terms of the total number of cases but also in terms of the number and frequency of traces for a single actor (e.g., one consumer) over time (Matz and Netzer 2017; Adjerid and Kelley 2018). Researchers can exploit this data to construct panels capturing actors' behavior over time as a function of variables of theoretical interest (e.g., Moore 2012) or examine how effects unfold over time (e.g., Datta, Knox and Bronnenberg 2018). The real-time nature of online data can allow researchers to study consumer behavior at high granularity, such as in seconds, minutes, hours, or days—something difficult to accomplish with experiments or surveys.

Next to deductive retailing research, web data provides enormous research potential focused on inductive theory-building. While various qualitative approaches, such as netnography, leverage web data (e.g., Kozinets 2002), these approaches rely on manual web data extraction. Few netnographic studies in marketing have leveraged web scraping or APIs (Arvidsson and Caliandro 2016). However, rich consumer and corporate narratives on blogs and access to online communities from idiosyncratic samples can be fruitful bases for generating novel and relevant retailing theories (Figueiredo, Larsen and Bean 2021).

Exploring emerging retail formats and trends

Finally, web data allows researchers to study nascent marketing and societal phenomena (Boegershausen et al. 2022). This is particularly relevant for retailing researchers, given the profound digital transformation (Verhoef et al. 2021) and the emergence of new players and retail formats not (yet) covered by traditional data providers. Over the last two decades, many of the most disruptive retailing trends have emerged online, from e-commerce marketplaces (e.g., Amazon) to ride-hailing services (e.g., UBER). Established retailers and brands encounter numerous challenges resulting in major digital disruptions due to a new class of digital-first competitors relying on direct-to-consumer and consumer-to-consumer model business models (Gielens and Steenkamp 2019; Muller 2020).

Consider, for example, the emergence of ultra-fast fashion retailers like SHEIN, which has significantly disrupted various major retailing categories. With approximately 200 million downloads, the Singapore-based Chinese e-commerce retailer was the most downloaded shopping app in the world in 2022 (Curry 2024). Its aggressive growth during the pandemic has catapulted its total revenue to be on par with major legacy fast-fashion retailers like Inditex, and H&M. SHEIN's business model is based on an extremely quickly changing assortment of fashion items priced at very low unit prices (e.g., $4 shirts and $6 dresses). Their "real-time" approach to assortment management (i.e., rapid prototyping and production) poses a unique opportunity to explore the evolving retailing landscape.

On the consumer side, researchers could study how the brand has built a cult following online based on a TikTok trend called "SHEIN hauls," wherein consumers buy many individual items and broadcast their purchases to their audience. By 2021, such videos had gained over 2.5 billion views (Gan 2021). Researchers could leverage TikTok data to explore what makes such videos engaging as well as what factors shape the outcomes of these videos for brands (e.g., brand attitudes) and content creators (e.g., new followers).

From a societal perspective, policymakers and non-governmental organizations have sounded the alarm about the sustainability and environmental impact of retailers like SHEIN. Web data could play a critical role in quantifying the environmental impact (e.g., via supply chain practices and environmental waste produced by end consumers) of an emerging class of ultra-cheap retailers like SHEIN and other marketplaces like TEMU that heavily rely on steeply discounted goods (for similar explorations of societal outcomes of marketing strategies and business models, see, van Lin et al. 2023; Ozer, Greenwood and Gopal 2024).

Another major retailing trend over the last decade is the emergence and growth of retail formats empowering small sellers to reach many consumers, such as the handmade and vintage goods marketplace Etsy. Marketplaces like Etsy allow researchers to explore more personalized, small-scale forms of retailing (Schnurr et al. 2022; Fuchs et al. 2022). When deployed on platforms like Etsy, web data can help researchers enhance the ecological validity of their research by complementing experimental studies with field data or creating a rich set of real-life stimuli (e.g., sellers' pages) that can be used in experimental studies (Boegershausen et al. 2022). Gathering many real-world stimuli facilitates the creation of more comprehensive stimulus-sampling paradigms. In these designs, participants are exposed to multiple instances of each experimental manipulation (e.g., various profiles of service providers that vary along variables of theoretical interest; Howe and Monin 2017). Stimulus-sampling paradigms boost the generalizability of effects and reduce the risk that an effect is driven by idiosyncratic features of certain experimental stimuli, such as the wording on an Etsy's seller's page shown to study participants (Judd, Westfall and Kenny 2017).

In sum, we have presented some case studies to illustrate how to study emerging phenomena quickly without relying on corporate partners. The list of applications for using web data is wide-ranging, from using Weedmaps to study dynamics in the cannabis retailing industry after the legalization of cannabis (Sharkey, Kovács and Hsu 2023; Hsu, Koçak and

Kovács 2018) to factors driving consumers' evaluation of non-fungible tokens (Hofstetter, Fritze and Lamberton 2024).[3]

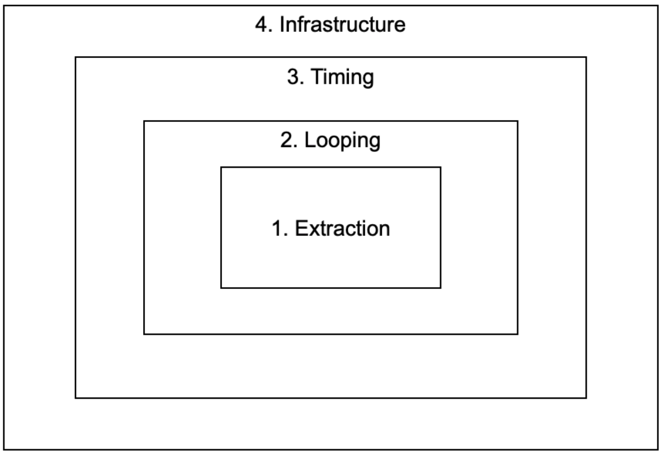
Getting Started with Web Scraping

After highlighting how web data can benefit retailing researchers, this section focuses on how to get started with web scraping. Specifically, we develop a practical guide featuring four essential steps: (1) data extraction and storage (e.g., which data points to extract), (2) looping to collect data for multiple units (e.g., extracting information for many products), (3) scheduling the extraction (e.g., to run weekly), and (4) deciding on which infrastructure the web scraper runs (e.g., a local computer or the cloud). As visualized in Figure 2, the web scraping process is a nested process in which additional layers (e.g., looping) are built around already developed parts (e.g., data extraction).[4] Next, we discuss the key considerations for each stage and a range of commonly used web scraping tools.

[3] An efficient way to collect data from Weedmaps.com is an "undocumented" API. Undocumented APIs are typically publicly accessible but come without the documentation designed to facilitate the adoption of regular (documented) APIs.

[4] Not all steps are necessary for each scraping project: for example, to capture data from a website's landing page, one could omit step 2 (looping) and proceed with steps 3-4 (scheduling and infrastructure).

Figure 2. A Nested Approach for Developing Web Scrapers



Notes: The figure depicts the four crucial steps for extracting web data. Researchers typically start by directly extracting data and building a loop to automate the process for multiple units (e.g., products or users). Researchers then schedule the data extraction (e.g., hourly, weekly), and finally make infrastructure decisions regarding where to run the scraper and how to store the data during and after the research project.

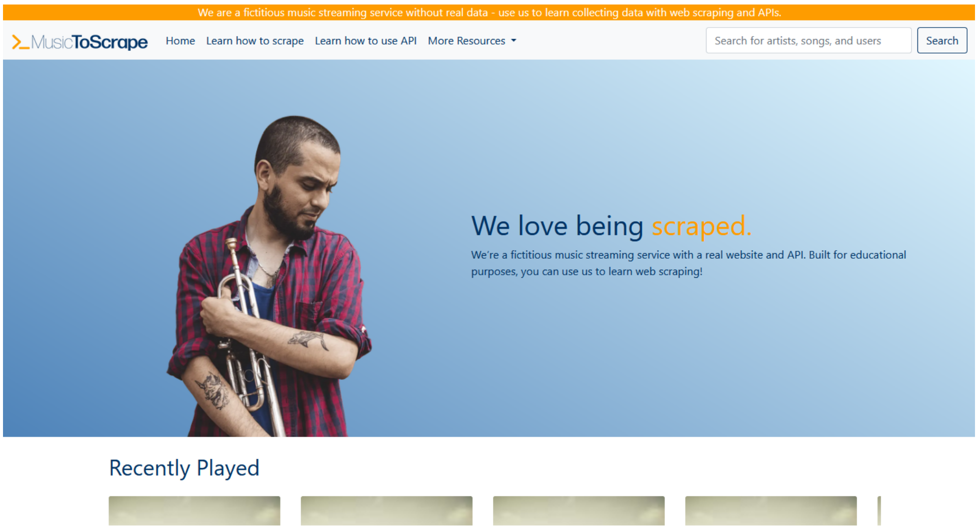
Building a web scraper for music-to-scrape.org

To guide novices in building a web scraper, we developed a mock-up retailing platform called music-to-scrape.org (https://music-to-scrape.org).[5] Like real-life retailing platforms, music-to-scrape.org has a desktop and mobile version, offers data via web scraping and APIs, and features data on various subpages (e.g., landing page, user profile page, artist or song page). Figure 3 shows a screenshot of music-to-scrape.org.

In what follows, we present an exemplary web scraper for extracting song metadata (here, a song's number of plays) from music-to-scrape.org using the open-source statistical software R. More tutorials and code snippets are available at https://music-to-scrape.org/. Readers with experience in designing web scraping studies can skip this subsection.

[5] Inspired by Zyte's https://books.toscrape.com project, music-to-scrape.org is a *dynamic* platform based on thousands of users' simulated music listening behavior. As an open-source project, the research community can extend music-to-scrape.org at https://github.com/tilburgsciencehub/music-to-scrape.

Figure 3. Screenshot of music-to-scrape.org



We are a fictitious music streaming service without real data - use us to learn collecting data with web scraping and APIs.

We love being scraped.

We're a fictitious music streaming service with a real website and API. Built for educational purposes, you can use us to learn web scraping!

Notes: Screenshot from https://music-to-scrape.org depicting the platform's landing page with dynamic and simulated data suited to learning how to extract data using web scraping and APIs.

Step 1: Data Extraction

Researchers must first connect with the online data source, locate desired information on the website, and save that information.[6] Assume a researcher is interested in obtaining song metadata. They can do so by visiting the song page at https://music-to-scrape.org. While exploring this website using a web browser, the researcher can not only find details such as the song name and the artist's name (e.g., "Is It You" by "Lee Ritenour") but also observe that the song's ID is included in the website's URL (https://music-to-scrape.org/song?song-id=*SOHMZNL12A58A8001A*). This ID serves as a means to programmatically access this information.

We next use the R package `rvest` to connect to the site (lines 2 and 10 in Table 1A). Subsequently, we extract information on the song's number of plays using the data point's

[6] Connecting to the data source can be done through various ways: downloading "web data" (such as reading the HTML code of a website or downloading particular files), browser emulation (such as remotely controlling a browser that can be instructed to click, scroll, and capture data), and phone emulation (particularly useful for capturing app data).

unique "address" on the website (lines 13-17).[7] For this, we use so-called selectors, which we have identified using the browser's "inspect mode" by hovering over individual elements of the website.[8] Finally, researchers save the data. Here, we save the song ID and corresponding data point and store it in a table (lines 20-23). The easiest is to write it into CSV (Comma Separated Value) files with columns for each variable and rows for each observation.

Table 1A: R Code for Step 1 (Data Extraction)

```
1    # Load the necessary libraries
2    library(rvest)
3    library(dplyr)
4
5    # Specify the URL of the website
6    song_id = 'SOHMZNL12A58A8001A'
7    url <- paste0("https://music-to-scrape.org/song?song-id=", song_id)
8
9    # Read the webpage into R
10   page <- read_html(url)
11
12   # Capture "basic information" table
13   basic_info = page %>%
14        html_nodes("div[class='song_basic_information_card']") %>%
15        html_elements('p')
16
17   plays = basic_info[4] %>% html_text()
18
19   # Collect information in a table
20   data = data.frame(song_id=song_id, plays=plays)
21
22   # Append row to CSV file (create if it does not exist yet)
23   write.table(data, 'data.csv', append=T, col.names=F)
```

Step 2: Looping

In most web scraping projects, researchers seek to capture information for *multiple* units (e.g., many products or retailers). Researchers can make use of so-called loops, which repeatedly execute a set of instructions (e.g., "extract the song title and artist name") a specified number of times or for a specific list of items (e.g., "extract artist names for all

[7] Web scraping works with *any* data displayed in a browser, i.e., it is not restricted to text-only information. For example, researchers can capture audio files, video content, or images.

[8] Website creators use CSS selectors to assign styles (e.g., font, font size) to specific (HTML) elements (e.g., a customer review) on a website. One useful tool to identify website elements and patterns is the SelectorGadget (selectorgadget.com). This open-source tool facilitates identifying the key CSS selectors needed to extract desired data from a website. Another type of selector often used in web scraping is the XPATH selector.

songs in the soul category“). Hence, we extend the previous example by assuming access to a list of song IDs to capture the play count information repeatedly.[9] We list these song IDs in line 25 of the code in Table 1B. To facilitate repeat execution of the data extraction of step 1, we “wrap” that code in a function. Line 28 starts the loop: for each song ID in the list of song IDs, the function extract_data() is executed (lines 6-22). In other words, in our example, the extraction is executed for each song ID in our list – i.e., five times.

Table 1B: R Code for Step 2 (Looping)

```
1   Load the necessary libraries
2   library(rvest)
3   library(dplyr)
4
5   # Function to extract the number of plays for a given song ID
6   extract_data <- function(song_id) {
7     url <- paste0("https://music-to-scrape.org/song?song-id=", song_id)
8
9     # Read the webpage into R
10    page <- read_html(url)
11
12    # Capture "basic information" table
13    basic_info = page %>%
14        html_nodes("div[class='song_basic_information_card']") %>%
15        html_elements('p')
16
17    plays = basic_info[4] %>% html_text()
18
19    # Save information
20    data = data.frame(song_id=song_id, plays=plays)
21    write.table(data, 'data.csv', append=T, col.names=F)
22  }
23
24  # list of song IDs
25  song_ids = c('SOSDCBC12A58A77A79', 'SOLYIBD12A8C135045',
26      'SOLSJHM12A8C139B46', 'SOKJZNJ12A8C13294B', 'SOHTWIB12A8AE46192')
27
28  for (song_id in song_ids) {
29    cat(paste('Extracting information for ', song_id, fill=T))
30    extract_data(song_id)
31    Sys.sleep(1)
32  }
```

Data extraction from step 1

Finally, we use a timer to reduce the data extraction speed to 1 second per page (see line 31), ensuring that the website is not contacted more than necessary. Throttling requests by limiting the number of requests to a website in a given

[9] We can also use a web scraper to capture this information. For our example, a search for “love” in the search bar of music-to-scrape.org yielded a list of many songs. We present five of these IDs in the example.

timeframe is critical when collecting data from real-life platforms to prevent server overload and avoid being blocked. A website's robots.txt guidelines and extracting at a moderate rate are vital for ethical data collection, fostering good relations with website administrators, and ensuring sustained access to the data source.

Step 3: Scheduling

The web scraper designed in steps 1 and 2 extracts data for multiple songs. Researchers can use scheduling if seeking to build a longitudinal data collection–i.e., capturing information for a set of units multiple times over longer periods (e.g., weeks, months).[10] This step can be skipped for single-shot data extractions. In our example, we assume we would like to extract information every hour.

While data extraction can be timed within R, it is better to schedule the data extraction at a "higher level," i.e., at the operating system level, to ensure stability. Tools are operating-system-specific (Task Scheduler for Windows and cron for Mac/Linux). In addition, researchers can consider setting up monitoring systems to track the performance of scrapers and notify researchers of successes, failures, or data anomalies. For example, if a scraper fails to run at its scheduled time, an alert can be triggered (e.g., an email), allowing the researcher to investigate and resolve the issue promptly. Table 1C provides example code for scheduling on Windows and Mac/Linux operating systems.

[10] Researchers have different ways of scheduling. It should be primarily motivated by the frequency of the focal phenomenon (see Boegershausen et al. 2022). For example, checking for a retailer's opening times *every day* may not be required because such information is not updated daily. Product stocks, in turn, can be captured multiple times a day to see when retailers replenish stocks or whether actual stockouts occur.

Table 1c: Process of Setting Up Scheduled Tasks in Step 3 (Scheduling)

A. Preparation for Scheduling

- Save R script from step 2.
- Test script by running it manually to make sure it works without errors
- Consider adding functionality in your R script to log output or errors, which can be useful for troubleshooting if your script is running automatically.

B. Starting the Scheduler

Windows

```
# install and load package
install.packages("taskscheduleR")
library(taskscheduleR)

# create scheduled task
taskscheduler_create(taskname = "My Task",
  rscript = "path/to/your_script.R",
  schedule = "HOURLY",
  starttime = format(Sys.time() + 60,
                     "%H:%M"),
  startdate = format(Sys.Date(),
                     "%d/%m/%Y"))
```

Mac/Linux

```
# install and load package
install.packages("cronR")
library(cronR)

# define which R script to schedule
cmd <- cron_rscript("path/to/your_script.R")

# add R script as scheduled task
cron_add(command = cmd, frequency = "hourly",
id = "my_script_id")
```

C. Removing Scheduled Tasks

- Use `taskscheduler_ls()` to list all scheduled tasks.
- If you want to remove a task, use `taskscheduler_delete(taskname = "My Task")`

- Use `cron_ls()` to see all scheduled jobs.
- If you want to remake a task, use `cron_rm("my_script_id")`

Step 4: Infrastructure

As web scraping projects grow in size and complexity, the need for scalable infrastructure becomes more pressing and critical. Starting with a single computer may be adequate for small-scale, exploratory projects, but larger endeavors may require moving to cloud services. The infrastructure of a web scraping project generally consists of one or more computers running the extraction software (steps 1-3) and an attached storage space (e.g., database or filesystem).

Choice of computation infrastructure. Researchers can execute their code locally or in the cloud. For example, the previous example was run locally on a researcher's computer. However, running the scraper on a personal computer may not be practical if data needs to be collected repeatedly over many weeks or months. In these circumstances, a server at a research institution or commercial cloud infrastructure may be more suitable. For self-programmed data collection, renting computers from cloud providers like Amazon Web Services, Microsoft Azure, or Google Cloud is an option. These providers offer preconfigured images with pay-by-the-hour flexibility. While low-powered systems often suffice, costs can rise quickly with heavier usage.

Choice of storage. In our current example, data is stored locally in a CSV file, risking data loss for longer data collections. Remote databases in the cloud are preferable for larger projects. Data security and privacy are paramount, particularly when dealing with personal or sensitive information (for extensive discussion and solutions, see Boegershausen et al. 2022). Databases offer the additional benefits of maintaining metadata about the collection, enabling multiple computers to collectively collect data, or facilitating logging and monitoring to safeguard quality.

Commonly used tools

While our example leverages R, the steps described can be implemented using different types of tools: code-based, non-code-based commercial tools, and LLM (Large Language Model)-based approaches. Each category of tools aligns with the steps of extraction, looping, scheduling, and infrastructure.

In code-based tools, technologies such as R (employing packages like `rvest` and RSelenium) and Python (using libraries like BeautifulSoup, Scrapy, and Selenium) are often used and are free of cost. On the other hand, non-code-based commercial tools like Octoparse, Import.io, ParseHub, and WebHarvy offer a more user-friendly, “worry-free” solution to web scraping where costs typically depend on the scale of the data collection. Although they might be pricier, they provide a hassle-free environment, especially for users who prefer a more straightforward approach. LLM-based tools like ChatGPT and similar models are emerging as innovative web scraping and data collection solutions. While they currently offer limited functionality in looping and scheduling, they excel in prototyping one-time data collections. More recent and advanced models require a subscription. Table 2 compares different software tools for web scraping projects.

Table 2. Software Stacks for Web Scraping

	Code-based scrapers	**Non-code based scrapers**	**LLM-based scrapers**
Project Type	Durable, long-term data collection; full control over the process and highly customizable	Durable, long-term data collections; web data collections without much customization	Prototyping, one-time data collection
Supported Steps			
1) Extraction	✓	✓	✓
2) Looping	✓	✓	Limited
3) Scheduling	✓	✓	✗
4) Infrastructure	customizable (e.g., databases, files)	✗	Limited
Tools	R (rvest, RSelenium), Python (BeautifulSoup, Scrapy, Selenium), Task Scheduler (Windows), cron (Mac/Linux)	Often commercial tools (e.g., Octoparse, Import.io, ParseHub, WebHarvy)	ChatGPT, Bard
Cost	Low – Moderate (only development and infrastructure)	Moderate – High (costs typically depend on volume of data collected)	Low - Moderate (subscription required for more advanced versions)

Novel opportunities and reflections

The main goal of our article is to encourage more retailing researchers to consider how they can incorporate web data into their research. Thus, we conclude by offering a reflection on how to address retailing-specific challenges in collecting web data and provide an outlook on leveraging generative AI and Big Team Science to jumpstart web data usage across the discipline.

Selecting Retailing Data and Sources

Both historical and future web data can be useful for retailing research. While historical data may be harder to obtain, we outline three ways to leverage web data. Specifically, we propose two strategies to collect data from past periods: (1) draw from nonprimary data providers like aggregators and (2) leverage archival versions of the target websites. Finally, we also outline the critical role of (3) collection design to effectively match web data with other data.

Surveying and extracting from aggregators. Dedicated external parties (i.e., nonprimary data providers) may have captured (part of) the interest data routinely. For example, HeissePreisse (https://heisse-preise.io), created by an Austrian developer to monitor food prices daily, contains pricing data on products sold at the larger Austrian, German, and Slovenian retailers. Importantly, the data is easily retrievable and contains a historical overview since 2017.[11] Similarly, Tweakers' Pricewatch (https://tweakers.net/pricewatch/) is an aggregator covering average and minimum prices of an exhaustive list of consumer electronics using more than 3,000 shops in The Netherlands, often dating back to the time when the product was launched. We encourage researchers to explore these and similar aggregators covering many product categories and geographic regions.

Leveraging archival versions of web data. The 'way-back-machine' (https://archive.org/web/), part of the Internet Archive, is among the most popular tools

[11] The Heisse-Preise project was initiated by a disgruntled developer in an attempt to provide insights into pricing trends and (a lack of) competition in the Austrian market. The web scraper uses the APIs of retailers to collect data. As a result of the project, the Austrian government focused on creating a legal framework in which retailers of a certain size need to make available and standardize information regarding a product's price and other details via APIs. The code is freely available on GitHub and can be ported to different countries, which we revisit in "Collection design."

available to researchers who want to travel back to a static version of a website. The Wayback Machine allows researchers to query for a link and check whether historical snapshots of websites are documented. The availability of such snapshots is driven by public demand, but researchers can also save pages for future use.

Collection design. By anticipating research questions and agendas, researchers vastly increase the options to leverage publicly available web data. We delineate two distinct philosophies on web data collection: a *targeted* versus *comprehensive* data collection approach. In *targeted* data collection, researchers focus on the data required to answer a specific research question. For example, should a researcher require information on the availability, variety, and price data of e-cigarettes before and after new legislation is introduced, a programmatic effort can be made to collect this exact data.[12] In contrast, a researcher adopting a *comprehensive* approach collects all data that may facilitate exploring multiple research questions. For example, researchers can focus on retailers or platforms and navigate to the specific retailer's website (e.g., Walmart.com). When adopting this approach, researchers store the entire web page rather than a specific element (as opposed to the approach in "Step 1" of "Building a Web Scraper"). This approach simplifies the steps outlined in Figure 2 at the expense of storing more data.[13] The web scraper could follow all first-degree links (i.e., any HTML links found on the landing page) and store these pages.

The two philosophies provide notably different advantages to researchers. We juxtapose the two philosophies in the Appendix using research process criteria (e.g., *data coverage*, *flexibility to shift research interest,* and *ability to create additional control variables*) and technical criteria (e.g., *resource intensity*, *robustness*, and *ease of matching).*

[12] Such programmatic efforts may require novel code or build on existing code. For example, using the example of "Heisse-Preise" as discussed in the "aggregators" section, researchers can download and adjust the publicly available code to focus on retailers in the respective market of interest.

[13] One trade-off that researchers can make ahead of time is to exclude saving any images, vastly reducing the disk space required but foregoing the possibility of visual analytics at a later point.

The ability to match to other data sources (e.g., NielsenIQ or GfK data) is particularly relevant to retailing research. To facilitate matching, we recommend collecting a great variety of data related to the focal data of interest. In the Appendix, we elaborate on the type of data to collect (e.g., EANs, brand names, flavors, but also visuals) and methods (e.g., deterministic, fuzzy matching, usage of internal search functions of web sources, and Generative AI) to match web data and other archival datasets.

Using Gen AI/LLMs for Web Scraping

Besides matching, GenAI and LLMs offer many opportunities for web scraping across various domains. Their application extends to several key areas: coding, data discovery, enrichment, and analysis support. We briefly discuss some of the most promising areas of GenAI deployment.

Coding. In the context of coding, setting up a basic web scraper is often straightforward, but scaling it for reliable, long-term operation presents a significant challenge. Here, specialized GPT models can be invaluable. They can assist researchers in intricate tasks, such as deciphering complex HTML code to identify relevant tags or ensuring efficient scheduling for web scraping tasks. This support is crucial for developing advanced scraping solutions that require sophisticated coding skills. For instance, a GPT could help figure out how to write the initial code in different programming languages to retrieve specific elements.

Regarding data discovery and enrichment, LLMs are pivotal in expanding the scope and depth of web scraping, especially in retail research. These models can assist in identifying diverse and relevant datasets or websites, thus preventing researchers from relying solely on popular or familiar sources (Boegershausen et al. 2022). This feature is particularly beneficial for exploring data from different countries where a researcher may not be well-versed. For data enhancement, LLMs can automate the execution of complex prompts across various data units. A practical example is the analysis of newspaper articles focusing on specific retailers. Here, an LLM can systematically identify retailer names within articles, facilitating the creation of a comprehensive retailer database. LLMs can also be instrumental in linking data across different databases, like matching unique product IDs, or performing tasks like data imputation. This automation enhances the richness of the collected data and streamlines the research process.

LLMs can contribute significantly to the analysis phase by generating automated quality reports. This functionality is crucial for alerting researchers to discrepancies or errors in their data collection or analysis processes. However, this aspect of LLM application in web scraping requires further exploration and development to fully harness its potential and ensure its effective integration into the research workflow. The Web Appendix includes example prompts illustrating how GenAI can help with coding, data discovery and enrichment, and analysis support.

Collecting Data in Big Teams

Most web data collection efforts are ad-hoc, project-specific, and led by one or a few researchers, leading to constraints in time and product coverage. This approach limits the robustness of data collection efforts and hinders re-use in other projects. We propose a big-team science approach to data collection to overcome these limitations. Big-team science involves extensive collaboration across various laboratories, institutions, disciplines, cultures, and continents. This approach has been increasingly adopted in various fields, such as psychology, particle physics, and behavioral genetics, to address generalizability, selection, and computational reproducibility challenges (Forscher et al. 2023).

For retail data collection, one of the foremost challenges is enhancing scalability and ensuring long-term operation. Pooling resources benefits researchers in several key areas: (a) more comprehensive exploration of promising web data sources, (b) development of more robust coding solutions, (c) continuous operation of web scraping along with vigilant monitoring of data quality, and (d) effective dissemination and accessibility of the data sets for download, which includes comprehensive documentation enabling other researchers to use these rich datasets (e.g., Gebru et al. 2020). To kickstart such an approach, we recommend following the steps outlined above, focusing on scalability (i.e., multiple collections concurrently) and distribution of expertise (e.g., researchers with innovative ideas to collect data and technically trained software engineers to implement the projects).

We hope researchers explore collaborative data collection and dissemination to create datasets with widely shared documentation and source code for public reuse. Success in this venture could lead to establishing a big-team science framework (Forscher et al. 2023), enhancing the stability and scope of gathering web data for retailing research. This initiative could also set benchmarks for meaningful reporting and evaluation. Academic journals should consider inviting registered-report-like dataset submissions to incentivize researchers to pursue these ambitious data collection efforts (see also Cavallo and Rigobon 2016). Our Web Appendix contains exemplary web scraping projects to be tackled via big-team science.

Conclusion

The continued growth of online commerce and rapidly evolving consumer behavior drive the retail industry’s digital transformation. Retailing researchers can embrace these forces by using web scraping and APIs to collect novel datasets capturing these emerging phenomena. To facilitate a broader adoption of web data across the entire retailing discipline, we provide resources to get started and offer hitherto missing guidance on overcoming challenges that have inhibited a broader adoption of web scraping in retailing. Web data offers unprecedented data on consumers, retailers, and markets. We hope our article encourages researchers to leverage web data to explore retailing questions and phenomena.

Appendix

Targeted vs. Comprehensive Web Data Collections

In Table A.1, we juxtapose two distinct web data collection philosophies: a *targeted* versus *comprehensive* approach. The targeted approach is suited for concrete and specific research questions, whereas researchers with an entire research agenda or multiple research questions may benefit from the comprehensive approach. The main differentiator in terms of research flow in the comprehensive approach is the increased flexibility in analyzing new focal and adding additional control variables. From a programming perspective, the targeted approach is more efficient at the cost of being more likely to break in case of a website update. We refer to the table for additional information and discuss the implications for matching thereafter.

Table A.1: Targeted vs. Comprehensive Web Data Collections

	Targeted web data collection	Comprehensive web data collection
Description	Identifying elements of interest on the website and collecting those exhaustively (e.g., pricing of all products of a retailer)	Identifying areas of interest and collecting a large body of data from the landing page and X-degree links (e.g., retailer's landing page and pages linked on the landing page)
Best suited for...	Focused research projects with clear research question(s) or predictions	Broad or explorative research projects or agendas where research questions are emerging
Data coverage	High (allows for collecting more depth)	Low - Moderate (allows for breadth but not depth)
Resource intensity (e.g., computing infrastructure, storage)	Low (only elements of interest)	Moderate – High (includes many and potentially large files, e.g., pictures)[1]
Flexibility to change data collection plan	Low (only if no alternative data is needed)	Moderate – High (can exploit naturally occuring unanticipated policy changes)
Robustness to environmental changes (e.g., website layout changes)	Low (elements or layout may change)	High (code is not prone to changes in web format)
Ease of matching (e.g., based on EAN)	Low (no additional characteristics for matching)	Moderate (additional data may provide matching identifiers)
Ability to create additional control variables (e.g., in review process)	N/A (no additional data collected)	Moderate – High (additional data may contain applicable information)

Notes: [1] A trade-off that researchers can make is to exclude saving images, vastly reducing the disk space required but foregoing the possibility of visual analytics at a later point.

Data Collection Design to Facilitate Matching

The ability to match to other data sources is particularly relevant to retailing research. While the approach (targeted or comprehensive) influences the variety of data collected, researchers can cast a smaller or wider net even within each approach. Collecting a greater variety of data related to the focal data of interest is beneficial.

For example, for a Coca-Cola can of 330ml, this would entail documenting the information of interest (i.e., price), but also metadata (i.e., URL, date of data collection, etc.) and any other information provided on the website (i.e., barcode, EAN, images, ingredients, flavors, brand names, etc.). These characteristics may also be used as control variables. It is crucial to create matching tables during data collection to enable this. For example, the matching table for the 330ml Coca-Cola can contain a unique ID (e.g., a URL or EAN and a retailer indicator) and relevant meta-data (e.g., data collection date).

Methods Used in Matching

Unique identifiers (UIDs) can be used as 'keys' to match data. Commonly used UIDs are product codes (e.g., Universal Product Code or European Article Number, UPC or EAN, respectively) allocated by a central authority. EANs and UPCs come in different lengths but are unique and can be converted using simple rules. If UPCs are available, a deterministic match is possible, whereas the absence of it leads to probabilistic matches. Examples of using the collected UPCs are found in Keller and Guyt (2023). It is particularly important for researchers that UPCs might not always be visible when browsing a particular website but may still be present in the metadata or the accompanying visuals. For example, retailers may use UIDs internally to track SKUs and still rely on these when searching for a product on the website, yet not display them on the user-facing side. Such UIDs are often found in the raw HTML files. Alternatively, pictures of SKUs can contain the UPC. Storing these pictures allows researchers to obtain the UPC relatively easily using available packages. Open-source software such as ZBar Bar Code Reader can identify bar codes from images and videos.

Nevertheless, researchers may find themselves with a UPC or alternative UID to facilitate deterministic matching. In such cases, probabilistic matching using heuristics or 'fuzzy' matching techniques may still provide relief. To engage in such efforts, researchers can focus on specific characteristics (e.g., size, brand name, flavor, ingredients) and write custom code to determine matches. This custom code can contain heuristics (e.g., if brand names are equal, flavor is equal, and size is equal, it is a match) or can utilize fuzzy matching routines that allow for ‘distances’ between characteristics. Such fuzzy matching can be particularly useful for non-standardized characteristics, such as a description of the product.

We propose two less frequently discussed alternative forms to facilitate matching: utilizing(i) the internal search function of retailers and platforms or (ii) generative AI (e.g., chatGPT). Regarding the internal search functions of retailers and platforms, we note that firms are incentivized to accurately display the product a customer is searching for. If the product is available, the retailer or platform will optimize its search function to match products to their characteristics. Such searches can be done using any available product characteristics that have been collected. Researchers could programmatically use the search function to find potential matches. This may also provide fruitful avenues for determining competitors of searched products.

While the field is rapidly evolving, there are clear use cases of GenAI, such as LLMs, for matching purposes. LLMs can perform matching on existing data or find auxiliary data using the ability of LLMs to navigate the web. Verifying the accuracy of the matches is important. Such checks can be executed using an inter-coder reliability measure in which a random sample of matches is selected and hand-coded for accuracy.

References

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

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

Balachandran, Sarath and Exequiel Hernandez (2021). "Mi Casa Es Tu Casa: Immigrant Entrepreneurs as Pathways to Foreign Venture Capital Investments." *Strategic Management Journal* 42 (11), 2047-83.

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

Boegershausen, Johannes, Hannes Datta, Abhishek Borah, and Andrew T. Stephen (2022). "Fields of Gold: Scraping Web Data for Marketing Insights." *Journal of Marketing* 86 (5), 1–20.

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

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

Chen, Xi and William D. Nordhaus (2011). "Using Luminosity Data as a Proxy for Economic Statistics." *Proceedings of the National Academy of Sciences* 108 (21), 8589-94.

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

Curry, David (2024). "Shein Revenue and Usage Statistics (2024)." https://web.archive.org/web/20240120134207/https://www.businessofapps.com/data/shein-statistics/.

Datta, Hannes, George Knox, and Bart J. Bronnenberg (2018). "Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery." *Marketing Science* 37 (1), 5-21.

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

Edelman, Benjamin and Wesley Brandi (2015). "Risk, Information, and Incentives in Online Affiliate Marketing." *Journal of Marketing Research* 52 (1), 1-12.

Figueiredo, Bernardo, Hanne Pico Larsen, and Jonathan Bean (2021). "The Cosmopolitan Servicescape." *Journal of Retailing* 97 (2), 267-87.

Fok, Dennis, Csilla Horváth, Richard Paap, and Philip Hans Franses (2006). "A Hierarchical Bayes Error Correction Model to Explain Dynamic Effects of Price Changes." *Journal of Marketing Research* 43 (3), 443-61.

Forscher, Patrick S., Eric-Jan Wagenmakers, Nicholas A. Coles, Miguel Alejandro Silan, Natália Dutra, Dana Basnight-Brown, and Hans Ijzerman (2023). "The Benefits, Barriers, and Risks of Big-Team Science." *Perspectives on Psychological Science* 18 (3), 607-23.

Fuchs, Christoph, Ulrike Kaiser, Martin Schreier, and Stijn M. J. van Osselaer (2022). "The Value of Making Producers Personal." *Journal of Retailing* 98 (3), 486-95.

Gan, Tammy (2021). "Why Are Massive Shein Hauls So Popular on Tiktok?" https://greenisthenewblack.com/shein-ultra-fast-fashion-consumerism-tiktok-influencer/.

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

George, Gerard, Ernst C. Osinga, Dovev Lavie, and Brent A. Scott (2016). "Big Data and Data Science Methods for Management Research." *Academy of Management Journal* 59 (5), 1493-507.

Geyskens, Inge, Katrijn Gielens, and Els Gijsbrechts (2010). "Proliferating Private-Label Portfolios: How Introducing Economy and Premium Private Labels Influences Brand Choice." *Journal of Marketing Research* 47 (5), 791-807.

Gielens, Katrijn and Anne L. Roggeveen (2023). "Editorial: So, What Is Retailing? The Scope of Journal of Retailing." *Journal of Retailing* 99 (2), 169-72.

Gielens, Katrijn and Jan-Benedict E. M. Steenkamp (2019). "Branding in the Era of Digital (Dis)Intermediation." *International Journal of Research in Marketing* 36 (3), 367-84.

Grewal, Dhruv, Dennis Herhausen, Stephan Ludwig, and Francisco Villarroel Ordenes (2022). "The Future of Digital Communication Research: Considering Dynamics and Multimodality." *Journal of Retailing* 98 (2), 224-40.

Guyt, Jonne and Els Gijsbrechts (2018). "On Consumer Choice Patterns and the Net Impact of Feature Promotions." *International Journal of Research in Marketing* 35 (3), 490-508.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010a). "Most People Are Not Weird." *Nature* 466 (7302), 29-29.

--- (2010b). "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2-3), 61-83.

Hofstetter, Reto, Martin P. Fritze, and Cait Lamberton (2024). "Beyond Scarcity: A Social Value-Based Lens for NFT Pricing." *Journal of Consumer Research* forthcoming.

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

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

Hsu, Greta, Özgecan Koçak, and Balázs Kovács (2018). "Co-Opt or Coexist? A Study of Medical Cannabis Dispensaries’ Identity-Based Responses to Recreational-Use Legalization in Colorado and Washington." *Organization Science* 29 (1), 172-90.

Huang, Ni, Gordon Burtch, Yili Hong, and Evan Polman (2016). "Effects of Multiple Psychological Distances on Construal and Consumer Evaluation: A Field Study of Online Reviews." *Journal of Consumer Psychology* 26 (4), 474-82.

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

Joo, Mingyu, Tridib Mazumdar, and S. P. Raj (2012). "Bidding Strategies and Consumer Savings in Nyop Auctions." *Journal of Retailing* 88 (1), 180-88.

Joy, Annamma, Jeff Jianfeng Wang, Davide C. Orazi, Seyee Yoon, Kathryn LaTour, and Camilo Peña (2023). "Co-Creating Affective Atmospheres in Retail Experience." *Journal of Retailing* 99 (2), 297-317.

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

Keller, Kristopher O. and Jonne Y. Guyt (2023). "A War on Sugar? Effects of Reduced Sugar Content and Package Size in the Soda Category." *Journal of Marketing* 87 (5), 698-718.

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

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

Kübler, Raoul V., Lara Lobschat, Lina Welke, and Hugo van der Meij (2023). "The Effect of Review Images on Review Helpfulness: A Contingency Approach." *Journal of Retailing*.

Lu, Huidi, Ralf van der Lans, Kristiaan Helsen, and Dinesh K. Gauri (2023). "Depart: Decomposing Prices Using Atheoretical Regression Trees." *International Journal of Research in Marketing* 40 (4), 781-800.

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

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

Meiseberg, Brinja (2016). "The Effectiveness of E-Tailers' Communication Practices in Stimulating Sales of Niche Versus Popular Products." *Journal of Retailing* 92 (3), 319-32.

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

Muller, Eitan (2020). "Delimiting Disruption: Why Uber Is Disruptive, but Airbnb Is Not." *International Journal of Research in Marketing* 37 (1), 43-55.

Ozer, Gorkem Turgut, Brad N. Greenwood, and Anandasivam Gopal (2024). "Noisebnb: An Empirical Analysis of Home-Sharing Platforms and Residential Noise Complaints." *Information Systems Research*.

Pan, Yue and Jason Q. Zhang (2011). "Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews." *Journal of Retailing* 87 (4), 598-612.

Peñaloza, Lisa and Mary C. Gilly (1999). "Marketer Acculturation: The Changer and the Changed." *Journal of Marketing* 63 (3), 84-104.

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

Ratchford, Brian, Gonca Soysal, Alejandro Zentner, and Dinesh K. Gauri (2022). "Online and Offline Retailing: What We Know and Directions for Future Research." *Journal of Retailing* 98 (1), 152-77.

Schnurr, Benedikt, Christoph Fuchs, Elisa Maira, Stefano Puntoni, Martin Schreier, and Stijn M. J. van Osselaer (2022). "Sales and Self: The Noneconomic Value of Selling the Fruits of One's Labor." *Journal of Marketing* 86 (3), 40-58.

Sharkey, Amanda, Balázs Kovács, and Greta Hsu (2023). "Expert Critics, Rankings, and Review Aggregators: the Changing Nature of Intermediation and the Rise of Markets with Multiple intermediaries." *Academy of Management Annals* 17 (1), 1-36.

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

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

van Lin, Arjen, Aylin Aydinli, Marco Bertini, Erica van Herpen, and Julia von Schuckmann (2023). "Does Cash Really Mean Trash? An Empirical Investigation into the Effect of Retailer Price Promotions on Household Food Waste." *Journal of Consumer Research* 50 (4), 663-82.

Venkatesan, Rajkumar, Kumar Mehta, and Ravi Bapna (2007). "Do Market Characteristics Impact the Relationship between Retailer Characteristics and Online Prices?" *Journal of Retailing* 83 (3), 309-24.

Verhoef, Peter C., Thijs Broekhuizen, Yakov Bart, Abhi Bhattacharya, John Qi Dong, Nicolai Fabian, and Michael Haenlein (2021). "Digital Transformation: A Multidisciplinary Reflection and Research Agenda." *Journal of Business Research* 122, 889-901.

Zhao, Kexin, Xia Zhao, and Jing Deng (2016). "An Empirical Investigation of Online Gray Markets." *Journal of Retailing* 92 (4), 397-410.

Unlocking the Potential of Web Scraping for Retailing Research

Web Appendix

Jonne Guyt, University of Amsterdam (j.y.guyt@uva.nl)
Johannes Boegershausen, Erasmus University Rotterdam (boegershausen@rsm.nl)
Hannes Datta, Tilburg University (h.datta@tilburguniversity.edu)

Web Appendix A:

Overview of Journal of Retailing Articles using Web Data

To identify articles in the *Journal of Retailing* using web data, we follow the approach of Boegershausen et al. (2022). Specifically, we used first various search terms describing the process of collecting web data (e.g., scrap*, crawl*, Application Programming Interface) as well as for the names of specific retailers and platforms (e.g., Yelp, TripAdvisor, Twitter, TikTok). We iteratively expanded the list of search terms based on our inspection of the initial articles discovered with the search terms (e.g., adding additional sources like "BoxOfficeMojo" or "Baidu").

Table A1. Articles Published in *Journal of Retailing* using Web Data

Author (year)	Title
Tang and Xing (2001)	Will the growth of multi-channel retailing diminish the pricing efficiency of the web?
Suter and Hardesty (2005)	Maximizing earnings and price fairness perceptions in online consumer-to-consumer auctions
Gopal et al. (2006)	From Fatwallet to eBay: An investigation of online deal-forums and sales promotions
Venkatesan, Mehta, and Bapna (2007)	Do market characteristics impact the relationship between retailer characteristics and online prices?
Duan, Gu, and Whinston (2008)	The dynamics of online word-of-mouth and product sales - An empirical investigation of the movie industry
Popkowski Leszczyc, Qiu, and He (2009)	Empirical Testing of the Reference-Price Effect of Buy-Now Prices in Internet Auctions
Aggarwal, Vaidyanathan, and Venkatesh (2009)	Using Lexical Semantic Analysis to Derive Online Brand Positions: An Application to Retail Marketing Research
Hu and Wang (2010)	Country-of-Origin Premiums for Retailers in International Trades: Evidence from eBay's International Markets
Pan and Zhang (2011)	Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews
Joo, Mazumdar, and Raj (2012)	Bidding Strategies and Consumer Savings in NYOP Auctions
Fay, Xie, and Feng (2015)	The Effect of Probabilistic Selling on the Optimal Product Mix
Wang, Liu, and Fang (2015)	User Reviews Variance, Critic Reviews Variance, and Product Sales: An Exploration of Customer Breadth and Depth Effects
Moon and Song (2015)	The Roles of Cultural Elements in International Retailing of Cultural Products: An Application to the Motion Picture Industry
Nejad, Amini, and Babakus (2015)	Success Factors in Product Seeding: The Role of Homophily
Gong, Smith, and Telang (2015)	Substitution or Promotion? The Impact of Price Discounts on Cross-Channel Sales of Digital Movies
Wu and Lee (2016)	Limited Edition for Me and Best Seller for You: The Impact of Scarcity versus Popularity Cues on Self versus Other-Purchase Behavior
Meiseberg (2016)	The Effectiveness of E-tailers' Communication Practices in Stimulating Sales of Niche versus Popular Products
Zhao, Zhao, and Deng (2016)	An Empirical Investigation of Online Gray Markets
Verma et al. (2019)	Are Low Price and Price Matching Guarantees Equivalent? The Effects of Different Price Guarantees on Consumers' Evaluations
Marchand and Marx (2020)	The Cosmopolitan Servicescape
Figueiredo, Larsen, and Bean (2021)	An empirical investigation of forward-looking retailer performance using parking lot traffic data derived from satellite imagery
Feng and Fay (2022)	Relative persuasiveness of repurchase intentions versus recommendations in online reviews
Ravula, Jha, and Biswas (2022)	Co-creating affective atmospheres in retail experience &#x2729;
Kovacheva, Nikolova, and Lamberton (2022)	The effect of review images on review helpfulness: A contingency approach
Joy et al. (2023)	Highlighting supply-abundance increases attraction to small-assortment retailers
Gu and Wu (2023)	Where you live matters: The impact of offline retail density on mobile shopping app usage
Kübler et al. (2023)	Automated Product Recommendations with Preference-Based Explanations
Cui, Zhu, and Chen (2023)	Will he buy a surprise? Gender differences in the purchase of surprise offerings

Web Appendix B:

Using LLMs for Web Scraping Projects

Table B1. Using LLMs for Web Scraping

Area	**Goal**	**Example prompts**[1]
Coding	Identify elements in web page	“Identify the html_tags that allow me to locate the price of products in the following web page.”
Coding	Suggest methods to extract elements	“Can you write code in R using Rvest that scrapes prices from the following website?”
Coding	Develop code in different languages	”Below, I have some R code which scrapes prices of a website. Could you translate the code into Python so it does the exact same thing?”
Coding	Debug and fix code	“My Python scraper is failing [place error or code underneath] to parse dates correctly from a webpage. Can you suggest a fix?”
Coding	Suggest code improvements	“Look at my code below which tries to scrape the website [insert website]. Could you give me some suggestions (if any) that could improve this script?”
Coding	Get a script to start a data collection	“I want to identify product names and product titles of an Amazon page. Which coding language would you recommend me to scrape the information with and could you give me a general script that I could use as a starting point?”
Data Discovery & Enrichment	Identify similar data sources	“I have data on prices of sodas at Walmart in the US, can you provide me with other [relevant retailers / countries] I should inspect?”
Data Discovery & Enrichment	Identify additional data sources	“I have data on sodas including the EAN, can you provide me with datasets with nutritional information on EAN codes?”
Analysis Support	Restructuring data to get “clean” output	“Given the following HTML [insert HTML underneath], how would I extract the product name and price using Python, R, and Puppeteer?”
Analysis Support	Check data for anomalies	“I have a scraper that collects data on prices from X, can you write an R function for me that verifies that all prices are in USD?”
Analysis Support	Recommending data visualizations	“I have data on [insert which data you have] scraped from a website. The data contains [insert what your data is about], please suggest 5 ways to visualize this data.”

Notes: Some prompts may not function with language models lacking internet access, like ChatGPT 3.5. Also, using just a webpage URL might be ineffective, even with internet-enabled models. A better approach is to upload a website’s HTML file, obtained by saving a webpage (e.g., Amazon.com’s homepage) and uploading the file.

Web Appendix C:

Exemplary Big-Team Science Web Data Projects

Table C1. Exemplary "big-team science" web data projects

Category	Explanation	Exemplary Platforms/Websites
E-commerce Websites	Scrape data to generate a database of historical prices, customer reviews, product availability, product pictures, and nutritional facts.	Amazon, eBay, Alibaba, Target, Wayfair
Social Media Platforms	Monitor for brand mentions, customer sentiment, trending products, and visual content analysis for product pictures.	X, Facebook, Instagram, Pinterest
Price Comparison Websites	Collect data to understand pricing strategies across different retailers and regions, including related products from recommendation engines.	Honey, CamelCamelCamel, Slickdeals
Consumer Forums and Review Sites	Capture customer reviews on understudied platforms, including discussions on product recommendations and related products.	Quora, Trustpilot, Yelp
International Retailer Websites	Compare retail strategies and product offerings globally. Uncover insights into regional market preferences and global retail trends, including nutritional facts.	Tesco (UK), Carrefour (France), Flipkart (India)
Mobile App Data	Capture data from retail mobile apps, including user engagement with product pictures and nutritional facts.	Walmart App, Amazon Shopping, The Home Depot
Cross-Platform Retail Data	Collect and integrate data from various online and offline platforms for a holistic view of the retail landscape, including product images and metadata.	Combination of online retailers, brick-and-mortar store data, and specialized e-commerce platforms like Shopify stores

References

Aggarwal, Praveen, Rajiv Vaidyanathan, and Alladi Venkatesh (2009), "Using Lexical Semantic Analysis to Derive Online Brand Positions: An Application to Retail Marketing Research," *Journal of Retailing*, 85 (2), 145-58.

Boegershausen, Johannes, Hannes Datta, Abhishek Borah, and Andrew T. Stephen (2022), "Fields of Gold: Scraping Web Data for Marketing Insights," *Journal of Marketing*, 86 (5), 1–20.

Cui, Xuebin, Ting Zhu, and Yubo Chen (2023), "Where You Live Matters: The Impact of Offline Retail Density on Mobile Shopping App Usage," *Journal of Retailing*.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008), "The Dynamics of Online Word-of-Mouth and Product Sales—an Empirical Investigation of the Movie Industry," *Journal of Retailing*, 84 (2), 233-42.

Fay, Scott, Jinhong Xie, and Cong Feng (2015), "The Effect of Probabilistic Selling on the Optimal Product Mix," *Journal of Retailing*, 91 (3), 451-67.

Feng, Cong and Scott Fay (2022), "An Empirical Investigation of Forward-Looking Retailer Performance Using Parking Lot Traffic Data Derived from Satellite Imagery," *Journal of Retailing*, 98 (4), 633-46.

Figueiredo, Bernardo, Hanne Pico Larsen, and Jonathan Bean (2021), "The Cosmopolitan Servicescape," *Journal of Retailing*, 97 (2), 267-87.

Gong, Jing, Michael D. Smith, and Rahul Telang (2015), "Substitution or Promotion? The Impact of Price Discounts on Cross-Channel Sales of Digital Movies," *Journal of Retailing*, 91 (2), 343-57.

Gopal, Ram D., Bhavik Pathak, Arvind K. Tripathi, and Fang Yin (2006), "From Fatwallet to Ebay: An Investigation of Online Deal-Forums and Sales Promotions," *Journal of Retailing*, 82 (2), 155-64.

Gu, Yangjie and Yuechen Wu (2023), "Highlighting Supply-Abundance Increases Attraction to Small-Assortment Retailers," *Journal of Retailing*, 99 (3), 420-39.

Hu, Ye and Xin Wang (2010), "Country-of-Origin Premiums for Retailers in International Trades: Evidence from Ebay's International Markets," *Journal of Retailing*, 86 (2), 200-07.

Joo, Mingyu, Tridib Mazumdar, and S. P. Raj (2012), "Bidding Strategies and Consumer Savings in Nyop Auctions," *Journal of Retailing*, 88 (1), 180-88.

Joy, Annamma, Jeff Jianfeng Wang, Davide C. Orazi, Seyee Yoon, Kathryn LaTour, and Camilo Peña (2023), "Co-Creating Affective Atmospheres in Retail Experience," *Journal of Retailing*, 99 (2), 297-317.

Kovacheva, Aleksandra, Hristina Nikolova, and Cait Lamberton (2022), "Will He Buy a Surprise? Gender Differences in the Purchase of Surprise Offerings," *Journal of Retailing*, 98 (4), 667-84.

Kübler, Raoul V., Lara Lobschat, Lina Welke, and Hugo van der Meij (2023), "The Effect of Review Images on Review Helpfulness: A Contingency Approach," *Journal of Retailing*.

Marchand, André and Paul Marx (2020), "Automated Product Recommendations with Preference-Based Explanations," *Journal of Retailing*, 96 (3), 328-43.

Meiseberg, Brinja (2016), "The Effectiveness of E-Tailers' Communication Practices in Stimulating Sales of Niche Versus Popular Products," *Journal of Retailing*, 92 (3), 319-32.

Moon, Sangkil and Reo Song (2015), "The Roles of Cultural Elements in International Retailing of Cultural Products: An Application to the Motion Picture Industry," *Journal of Retailing*, 91 (1), 154-70.

Nejad, Mohammad G., Mehdi Amini, and Emin Babakus (2015), "Success Factors in Product Seeding: The Role of Homophily," *Journal of Retailing*, 91 (1), 68-88.

Pan, Yue and Jason Q. Zhang (2011), "Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews," *Journal of Retailing*, 87 (4), 598-612.

Popkowski Leszczyc, Peter T. L., Chun Qiu, and Yongfu He (2009), "Empirical Testing of the Reference-Price Effect of Buy-Now Prices in Internet Auctions," *Journal of Retailing*, 85 (2), 211-21.

Ravula, Prashanth, Subhash Jha, and Abhijit Biswas (2022), "Relative Persuasiveness of Repurchase Intentions Versus Recommendations in Online Reviews," *Journal of Retailing*, 98 (4), 724-40.

Suter, Tracy A. and David M. Hardesty (2005), "Maximizing Earnings and Price Fairness Perceptions in Online Consumer-to-Consumer Auctions," *Journal of Retailing*, 81 (4), 307-17.

Tang, Fang-Fang and Xiaolin Xing (2001), "Will the Growth of Multi-Channel Retailing Diminish the Pricing Efficiency of the Web?," *Journal of Retailing*, 77 (3), 319-33.

Venkatesan, Rajkumar, Kumar Mehta, and Ravi Bapna (2007), "Do Market Characteristics Impact the Relationship between Retailer Characteristics and Online Prices?," *Journal of Retailing*, 83 (3), 309-24.

Verma, Swati, Abhijit Guha, Abhijit Biswas, and Dhruv Grewal (2019), "Are Low Price and Price Matching Guarantees Equivalent? The Effects of Different Price Guarantees on Consumers' Evaluations," *Journal of Retailing*, 95 (3), 99-108.

Wang, Feng, Xuefeng Liu, and Eric Fang (2015), "User Reviews Variance, Critic Reviews Variance, and Product Sales: An Exploration of Customer Breadth and Depth Effects," *Journal of Retailing*, 91 (3), 372-89.

Wu, Laurie and Christopher Lee (2016), "Limited Edition for Me and Best Seller for You: The Impact of Scarcity Versus Popularity Cues on Self Versus Other-Purchase Behavior," *Journal of Retailing*, 92 (4), 486-99.

Zhao, Kexin, Xia Zhao, and Jing Deng (2016), "An Empirical Investigation of Online Gray Markets," *Journal of Retailing*, 92 (4), 397-410.