WEB APPENDIX

Fields of Gold:

Scraping Web Data for Marketing Insights

Johannes Boegershausen

Hannes Datta

Abhishek Borah

Andrew T. Stephen*

Table of contents

Web Appendix A: Comparing Web Scraping and APIs	p. 2
Web Appendix B: Technical Tutorials for Web Scraping and APIs	p. 3
Web Appendix C: Marketing Research using Web Data	p. 4
Web Appendix D: Legal Considerations	p. 8
Web Appendix E: Supplementary Material for Stylized Example	p. 11
Web Appendix F: Calculation of Technically Feasible Sample Sizes	p. 12

These materials have been supplied by the authors to aid in the understanding of their paper. The AMA is sharing these materials at the request of the authors.

Digital companion available at https://web-scraping.org.

* Johannes Boegershausen (boegershausen@rsm.nl) is Assistant Professor of Marketing, Rotterdam School of Management, Erasmus University, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands. Hannes Datta (h.datta@tilburguniversity.edu) is Associate Professor of Marketing at Tilburg University, Postbus 90153, 5000 LE Tilburg, The Netherlands. Abhishek Borah (abhishek.borah@insead.edu) is Assistant Professor of Marketing at INSEAD, Boulevard de Constance, 77305 Fontainebleau, France. Andrew T. Stephen (andrew.stephen@sbs.ox.ac.uk) is the L'Oréal Professor of Marketing, Associate Dean of Research, Saïd Business School, University of Oxford, Park End Street, Oxford OX1 1HP, United Kingdom.

WEB APPENDIX A: COMPARING WEB SCRAPING AND APIS

Table W1: Comparison of Web Scraping and APIs

	Web scraping	**Application Programming Interfaces (APIs)**
Use cases	Extract data from websites or apps	Extract data from APIs
Scope	Any content that can be displayed in a web browser	Any content or algorithm made available by the API provider
Data extraction	Programmatically browse the website and capture information as it becomes visible or available in the website's HTML (HyperText Markup Language) source code	Capture information directly from programming interfaces designed for content extraction and enrichment at scale, typically in JSON (JavaScript Object Notation) or XML (Extensible Markup Language) format
Costs	Free	Usually on a subscription (e.g., free, freemium, paid)
Scalability	Moderate	High
Data documentation	Mostly undocumented	Mostly documented for app developers
Legal and ethical risks	Low-high	Low-moderate[1]
Illustrative data sources	E-commerce (https://amazon.com), Online review (https://yelp.com)	Discussion forum (Reddit API), Social media (Twitter API)
Illustrative articles	Chevalier and Mayzlin (2006); Ludwig et al. (2013)	Tellis et al. (2019); Toubia and Stephen (2013)

[1] Before 2010, companies focused on the "social" aspects of their business model and actively offered APIs (usually in XML or JSON feeds) to grow their ecosystem. As companies matured and grew their userbase, many of these firms became increasingly restrictive as they sought to monetize their data (e.g., Amazon, Twitter).

WEB APPENDIX B: TECHNICAL TUTORIALS FOR WEB SCRAPING AND APIS

Table W2: Technical Tutorials for Web Scraping and APIs

Source	Extraction method	Programming Language	Main software and packages	Summary
Bradley and James (2019)	Web scraping	R	rvest	Introduction to collecting data from static sites. A companion offers videos, exercises, and code (https://practicewebscrapingsite.wordpress.com).
Brody (2017)	Web scraping	Python	beautifulsoup	An accessible introduction to basic web scraping with a practice website (https://scrapethissite.com).
Ciechanowski et al. (2020)	Web scraping	N/A		An introduction to various non-coding alternatives for gathering and analyzing web data.
Cushman Garland (2020)	Web scraping	Python	beautifulsoup	An entry-level class that covers the Python fundamentals, command line, and beautifulsoup.
Devlin (2020)	API	Python	requests	Beginner-level tutorial on how to retrieve data from APIs in Python.
Han and Anderson (2021)	Web scraping	Python	selenium	Concise discussion of extracting data from dynamic websites (e.g., Tripadvisor) with example code.
Jaeger et al. (2020)	API	R	httr, base64enc	An introduction to APIs (Face++ and Kairo) for extracting demographic information from facial images.
Je-Suis-TM (2019)	Web scraping	Python	beautifulsoup, regex	Detailed web scraping tutorials, including 20 scraping examples with code.
Keegan (2019)	API + Web scraping	Python	beautifulsoup	Tutorials for retrieving data from the web, intended for researchers in the social sciences and humanities with limited programming experience.
Landers et al. (2016)	Web scraping	Python	scrapy	Tutorial with supporting website, focusing on the Python package scrapy (https://rlanders.net/scrapy).
Markham (2017)	Web scraping	Python	beautifulsoup	Introductory-level web scraping tutorial using the requests and beautifulsoup libraries.
Merrett (2018)	Web scraping	R	rvest	Introduction to web scraping in R using the rvest package.
Mitchell (2018)	Web scraping	Python	beautifulsoup	An extensive handbook with practical examples for extracting web data.
Munzert et al. (2015)	Web scraping	R	RCurl, stringr, etc.	A comprehensive treatment of scraping and processing web data with R.
Murphy (2017)	API	R	rtweet	A tutorial for extracting psychological data from the Twitter API.
Pascual (2020)	Web scraping	R	rvest	A tutorial that covers the basics of how to do web scraping applied to obtaining weather data.
Paxton and Griffiths (2017)	N/A	N/A		Not a tutorial, but offers an excellent summary of existing web datasets, including a companion website (https://dataonthemind.org).
Russell and Klassen (2019)	API + web scraping	Python	twitter, geopy, facebooksdk, etc.	Detailed introduction to collecting social network data using APIs to collect data from various social networks.
Schrenk (2012)	Web scraping	PHP/Curl		Covers fundamental scraping concepts and a series of example projects. Zooms in on solving some of the most cumbersome technical challenges.
Smith (2019)	Web scraping	Go		Explains how features specific to Go can facilitate web scraping and solve some of the common pitfalls encountered when extracting data.
vanden Broucke and Baesens (2018)	Web scraping	Python	beautifulsoup	Comprehensive introduction with advanced code snippets encountered when scaling up (e.g., Java, databases). Covers many illustrative cases.
Vallone et al. (2020)	Web scraping	R	DataSpa	Illustrates how to efficiently process collected data (e.g., PDF files) at scale.

WEB APPENDIX C:

MARKETING RESEARCH USING WEB DATA

To identify articles using web data, we employ various search terms on Web of Science, Google Scholar, and journal websites related to extraction of publicly available web data for marketing research. Our initial search terms were associated with *collecting* web data (e.g., scrap*, crawl*, API, Application Programming Interface) or related to specific collected *web sources* (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 (e.g., adding category-specific search terms like "BoxOfficeMojo", "crowdf*", or "online forum*") 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*. We also included articles that have reused web data (e.g., the Amazon datasets of McAuley 2021). Until the end of 2020, the total number of articles using web data in the selected journals amounts to 314.

Table W3: Most Cited Marketing Research using Web Data[1]

Rank	Author (year)	Journal	Cites/ year[2]	Keywords
1	Zervas et al. (2017)	*JMR*	166.7	sharing economy; Airbnb; hotel industry; competition; peer-to-peer markets
2	Chevalier and Mayzlin (2006) [3]	*JMR*	164.3	word of mouth; online reviews; review valence; negativity bias; sales ranks
3	Muniz and O'Guinn (2001)	*JCR*	116.4	brand equity/extensions, brand loyalty, group/interpersonal influences, sociological theories/analysis
4	Berger and Milkman (2012)	*JMR*	108.9	word of mouth; viral marketing; social transmission; online content
5	Zhu and Zhang (2010)	*JM*	92.6	Internet marketing; online consumer reviews; word of mouth; video game; long tail
6	Bardhi and Eckhardt (2012)	*JCR*	90.7	sharing economy; possessions; ownership
7	Schau et al. (2009)	*JM*	89.5	brand community; branding; collective consumption; engagement strategies; marketing strategy; practice theory
8	Kozinets (2002) [3]	*JMR*	78.5	netnography, ethnography, qualitative research, online consumer groups
9	Liu (2006) [3]	*JM*	74	word of mouth; dynamics; movie ratings; expectations; movie revenues
10	Kozinets et al. (2010)	*JM*	70.9	advertising and promotions; consumer communication; online communities; online consumer behavior; Internet marketing; social media; word of mouth
11	Godes and Mayzlin (2004)	*MktSci*	66.8	word of mouth; diffusion of innovations; measurement; networks and marketing; new product research; Internet marketing
12	Tirunillai and Tellis (2014)	*JMR*	36	consumer satisfaction; quality; dimensions; brand mapping; big data; latent Dirichlet allocation; user-generated content
13	Chintagunta et al. (2010)	*MktSci*	35.9	online word of mouth; sequential new product release; endogeneity; instrumental variables; generalized method of moments; motion pictures
14	Ludwig et al. (2013)	*JM*	34.7	online customer reviews; affective content; linguistic style match; conversion rate; Internet marketing
15	Brown et al. (2003) [3]	*JM*	32.7	nostalgia; consumers; consumption; market; ethnography; emotions; strategy; attitude; culture; age
16	Tucker (2014)	*JMR*	32.4	privacy; online advertising; social networks
17	Muñiz and Schau (2005)	*JCR*	31.5	brand loyalty; sociological theories/analysis, e-commerce/computing/Internet; ethnography
18	Kozinets (2002)	*JCR*	31.3	critical theory; cultural theories and analysis; postmodernism/ poststructuralism; ethnography
19	Martin et al. (2017)	*JM*	31.2	data breach; consumer vulnerability; privacy; spillover effects; big data

Table W3: Most Cited Marketing Research using Web Data [continued]

Rank	Author (year)	Journal	Cites/ year[1]	Keywords
20	Netzer et al. (2012)	*MktSci*	30.9	text mining; user-generated content; market structure; marketing research
21	Schau and Gilly (2003)	*JCR*	30.5	e-commerce/computing/Internet, self-concept, symbolic consumption/semiotics, depth/long interviews, grounded theory
22	Scaraboto and Fischer (2013) [3]	*JCR*	29.9	marginalized consumers, mobilization, collective identity, institutional logics, practices, market change dynamics, fashion
23	Tirunillai and Tellis (2012)	*MktSci*	29.7	user-generated content (UGC); stock returns; online word of mouth; vector autoregression (VAR); computational text processing
24	Perren and Kozinets (2018)	*JM*	29.2	access-based consumption; algorithm marketing; collaborative consumption; consociality; institutional theory; lateral exchange markets; opportunism; peer-to-peer markets; platform economics; service marketing; sharing economy; service dominant logic
25	Moe and Trusov (2011)	*JMR*	28.4	online word of mouth; ratings; reviews; social dynamics; hazard models; Internet marketing
26	Hughes et al. (2019)	*JM*	28	advertising campaign intent; brand engagement; consumer decision journey; influencer marketing; social media influencers; social media platform; social network sites; sponsored bloggers
27	Ghose et al. (2012)	*MktSci*	27.3	user-generated content; social media; search engines; hotels; ranking system; structural models; text mining; crowdsourcing
28	Tellis et al. (2019)	*JM*	27.1	virality; shares; ad cues; emotion; brand prominence
29	Kozinets and Handelman (2004)	*JCR*	27.1	cultural theories and analysis; sociological theories/analysis; public policy issues; depth/long interviews; ethnography
30	Chen et al. (2011)	*JMR*	26.1	social interactions; social influences; observational learning; word of mouth; natural experiment

Notes:
[1] Published in the *Journal of Marketing* (JM), *Journal of Marketing Research* (JMR), *Marketing Science* (MktSci), *Journal of Consumer Research* (JCR), and the *Journal of Consumer Psychology* (JCP).
[2] Web of Science citation count on 9 April 2022, divided by years since publication. Years since publication are computed as 2021 – year of publication + 1 + .25 (the latter to account for citations in the first quarter of 2022).
[3] Based on self-generated keywords because *JMR* and *JM* did not feature author-provided keywords at the beginning of the 2000s, and *JCR* did not feature keywords in the early 2010s.

Table W4: Top 20 Sources Used in Marketing Research[1]

		Number of publications					
Rank	Source	JM	JMR	JCR	JCP	MktSci	Total
1	Amazon.com	5	9	7	1	16	38
2	Twitter	6	8	5	1	7	27
3	IMDb	8	5	4	0	7	24
4	Google Trends	7	7	1	0	7	22
5	Trip Advisor	2	7	2	1	4	16
6	Facebook	8	0	4	1	0	13
7	Yelp	3	5	4	0	1	13
8	Rotten Tomatoes	4	5	1	0	0	10
9	The Numbers	7	1	1	0	1	10
10	Yahoo Movies	4	1	1	1	1	8
11	Google (Search)	0	2	1	0	4	7
12	Box Office Mojo	4	2	0	0	0	6
13	eBay	0	1	5	0	0	6
14	Expedia	0	1	0	0	5	6
15	YouTube	2	2	1	0	1	6
16	Epinions	0	3	0	0	2	5
17	Metacritic	3	2	0	0	0	5
18	Wikipedia	2	2	0	0	1	5
19	CamelCamelCamel	0	0	1	0	3	4
20	Instagram	0	1	1	1	1	4

Notes:
[1] Published in the *Journal of Marketing* (JM), *Journal of Marketing Research* (JMR), *Marketing Science* (MktSci), *Journal of Consumer Research* (JCR), *Journal of Consumer Psychology* (JCP).

WEB APPENDIX D: LEGAL CONSIDERATIONS

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. It is important to note that the legislative landscape surrounding the extraction of web data is constantly evolving. We thus strongly recommend researchers track the legal landscape. 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, for certain projects, it might be necessary to 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, summarized in Table W4.1.

Table W5: Key issues to raise when obtaining legal advice

Theme	Reason for importance of the issue	Potential issues that may affect the researcher's legal compliance
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., the feasibility of gathering 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 envisioned 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 collection varies 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 W5: Key issues to raise when obtaining legal advice [continued]

Theme	Reason for importance of the issue	Potential issues that may affect the researcher's legal compliance
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 becomes public unintentionally? • Does the data provider disallow the use of the automated collection, 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., disguising the identity of the data provider in the article)? • 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)? • Is there a potential conflict of interest between researchers and the data provider?
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, or (semi-publicly) 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)?

WEB APPENDIX E: SUPPLEMENTARY MATERIAL FOR STYLIZED EXAMPLE

Table W6: Supporting links and explanations for stylized marketing example

Stage	Relevant Links
1) Which information to extract?	• *Amazon product page:* https://www.amazon.com/Influence-New-Expanded-Psychology-Persuasion/dp/0062937650/ (Cialdini's classic "Influence: The Psychology of Persuasion," the #1 bestseller in "Marketing (Books)" at Amazon) • *Amazon review page:* https://www.amazon.com/product-reviews/0062937650/ • *Amazon user page:* https://www.amazon.com/gp/profile/<ID>
2) Which pages to use as a starting point of the data collection?	• *Top 100 bestsellers in "Marketing (Books)" at Amazon:* https://www.amazon.com/gp/bestsellers/books/2702 • *Page 8 of the Bestseller list (i.e., rank 351-400) in "Marketing (Books)" at Amazon identified via experimentation with the URLs:* https://www.amazon.com/Best-Sellers-Books/zgbs/books/2702/ref=zg_bs_pg_2?_encoding=UTF8&pg=8 • *Amazon's advanced book search:* https://www.amazon.com/advanced-search/books • *Search results page based on the keyword "marketing" and the publication range "during December 2021":* https://www.amazon.com/s?k=marketing&i=stripbooks&s=relevanceexprank&Adv-Srch-Books-Submit.x=0&Adv-Srch-Books-Submit.y=0&field-datemod=12&field-dateop=During&field-dateyear=2021&
3) At which frequency to extract the information?	• Amazon's Top Reviewer Rankings: https://www.amazon.com/hz/leaderboard/top-reviewers

WEB APPENDIX F: CALCULATION OF TECHNICALLY FEASIBLE SAMPLE SIZES

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:

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

The parameters are explained in Table W7 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 converting 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).

Table W7: Parameters and Example

Parameter	Description	Example
N	Sample size (number of objects for which to retrieve data)	Target sample size that will be computed in this example (2,250).
r	Number of URL requests necessary to obtain data for one object	The web scraper needs to visit 2 pages: one to obtain users' metadata and one to obtain users' usage history.
freq	Sampling frequency for each object and time unit	Once per 15 minutes = 4 times per hour. Data for each user should be obtained at least once every fifteen minutes.
req	Retrieval limit (maximum number of requests per time unit, allowed for each scraper or authenticated API user)	5 requests per second = 5 x 60 x 60 requests per hour (18,000)[1].
S	Number of collection scripts (e.g., different instances of the extraction software, or the number of (authenticated) users of an API).	One authenticated data collection script running on one computer.

[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 calculate the average number of *actual requests* made per time unit.

REFERENCES

Bardhi, Fleura and Giana M. Eckhardt (2012), "Access-Based Consumption: The Case of Car Sharing," *Journal of Consumer Research*, 39 (4), 881-98.

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

Bradley, Alex and Richard J. E. James (2019), "Web Scraping Using R," *Advances in Methods and Practices in Psychological Science*, 2 (3), 264 –70.

Brody, Hartley (2017), The Ultimate Guide to Web Scraping: https://blog.hartleybrody.com/guide-to-web-scraping/.

Brown, Stephen, Robert V. Kozinets, and John F. Sherry (2003), "Teaching Old Brands New Tricks: Retro Branding and the Revival of Brand Meaning," *Journal of Marketing*, 67 (3), 19-33.

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

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

Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman (2010), "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation across Local Markets," *Marketing Science*, 29 (5), 944-57.

Ciechanowski, Leon, Dariusz Jemielniak, and Peter A. Gloor (2020), "Tutorial: AI Research without Coding: The Art of Fighting without Fighting: Data Science for Qualitative Researchers," *Journal of Business Research*, 117 (1), 322-30.

Cushman Garland, Hannah (2020), "Web Scraping with Python," (accessed 3 March, 2021), https://github.com/hancush/web-scraping-with-python.

Devlin, Josh (2020), "Python API Tutorial: Getting Started with APIs," (accessed 2 March, 2021), https://www.dataquest.io/blog/python-api-tutorial/.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li (2012), "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content," *Marketing Science*, 31 (3), 493-520.

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

Han, Saram and Christopher K. Anderson (2021), "Web Scraping for Hospitality Research: Overview, Opportunities, and Implications," *Cornell Hospitality Quarterly*, 62 (1), 89-104.

Hughes, Christian, Vanitha Swaminathan, and Gillian Brooks (2019), "Driving Brand Engagement through Online Social Influencers: An Empirical Investigation of Sponsored Blogging Campaigns," *Journal of Marketing*, 83 (5), 78-96.

Jaeger, Bastian, Willem W. A. Sleegers, and Anthony M. Evans (2020), "Automated Classification of Demographics from Face Images: A Tutorial and Validation," *Social and Personality Psychology Compass*, 14 (3), e12520.

Je-Suis-TM, T. M. (2019), "Web Scraping," (accessed 25 March, 2021), https://github.com/je-suis-tm/web-scraping.

Keegan, Brian C. (2019), "Web Data Scraping," (accessed 2 March, 2021), https://github.com/CU-ITSS/Web-Data-Scraping-S2019.

Kozinets, Robert V. (2002), "Can Consumers Escape the Market? Emancipatory Illuminations from Burning Man," *Journal of Consumer Research*, 29 (1), 20-38.

Kozinets, Robert V., Kristine De Valck, Andrea C. Wojnicki, and Sarah J. S. Wilner (2010), "Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities," *Journal of Marketing*, 74 (2), 71-89.

Kozinets, Robert V. and Jay M. Handelman (2004), "Adversaries of Consumption: Consumer Movements, Activism, and Ideology," *Journal of Consumer Research*, 31 (3), 691-704.

Landers, Richard N., Robert C. Brusso, Katelyn J. Cavanaugh, and Andrew B. Collmus (2016), "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research," *Psychological Methods*, 21 (4), 475-92.

Liu, Yong (2006), "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70 (3), 74-89.

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

Markham, Kevin (2017), "Web Scraping the President's Lies in 16 Lines of Python," (accessed 10 March, 2021), https://github.com/justmarkham/trump-lies.

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

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

Merrett, Rebecca (2018), "Web Scraping in R Part 1 | Writing Your Script in rvest," (accessed 25 March, 2021), https://tutorials.datasciencedojo.com/r-tutorial-web-scraping-script/.

Mitchell, Ryan (2018), Web Scraping with Python: Collecting Data from the Modern Web (2nd ed.). Sebastopol, CA: O'Reilly Media.

Moe, Wendy W. and Michael Trusov (2011), "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research*, 48 (3), 444-56.

Muniz, Jr Albert M. and Thomas C. O'Guinn (2001), "Brand Community," *Journal of Consumer Research*, 27 (4), 412-32.

Muñiz, Jr Albert M. and Hope Jensen Schau (2005), "Religiosity in the Abandoned Apple Newton Brand Community," *Journal of Consumer Research*, 31 (4), 737-47.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis (2015), Automated Data Collection with R. Chichester: John Wiley & Sons, Ltd.

Murphy, Sean C. (2017), "A Hands-on Guide to Conducting Psychological Research on Twitter," *Social Psychological and Personality Science*, 8 (4), 396-412.

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

Pascual, Christian (2020), "Tutorial: Web Scraping in R with rvest," (accessed 26 February, 2021), https://www.dataquest.io/blog/web-scraping-in-r-rvest/.

Paxton, Alexandra and Thomas L. Griffiths (2017), "Finding the Traces of Behavioral and Cognitive Processes in Big Data and Naturally Occurring Datasets," *Behavior Research Methods*, 49 (5), 1630-38.

Perren, Rebeca and Robert V. Kozinets (2018), "Lateral Exchange Markets: How Social Platforms Operate in a Networked Economy," *Journal of Marketing*, 82 (1), 20-36.

Russell, Matthew A and Mikhail Klassen (2019), Mining the Social Web: Data Mining Facebook, Twitter, Linkedin, Instagram, Github, and More. Sebastopol, CA: O'Reilly Media.

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

Schau, Hope Jensen and Mary C. Gilly (2003), "We Are What We Post? Self-Presentation in Personal Web Space," *Journal of Consumer Research*, 30 (3), 385-404.

Schau, Hope Jensen, Albert M. Muñiz, and Eric J. Arnould (2009), "How Brand Community Practices Create Value," *Journal of Marketing*, 73 (5), 30-51.

Schrenk, Michael (2012), Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with Php/Curl. San Francisco: No Starch Press.

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

Smith, Vincent (2019), And About Half of All These Studies Are Published in Marketing Science. Birmingham, UK: Packt Publishing.

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

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

Tirunillai, Seshadri and Gerard J. Tellis (2014), "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research*, 51 (4), 463-79.

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

Tucker, Catherine E. (2014), "Social Networks, Personalized Advertising, and Privacy Controls," *Journal of Marketing Research*, 51 (5), 546-62.

Vallone, Andrés, Coro Chasco, and Beatriz Sánchez (2020), "Strategies to Access Web-Enabled Urban Spatial Data for Socioeconomic Research Using R Functions," *Journal of Geographical Systems*, 22 (2), 217-39.

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

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

Zhu, Feng and Xiaoquan Zhang (2010), "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics," *Journal of Marketing*, 74 (2), 133-48.