Quality & Quantity
https://doi.org/10.1007/s11135-021-01164-0

Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences

Alex Luscombe[1] · **Kevin Dick**[2] · **Kevin Walby**[3]

Accepted: 7 May 2021

Abstract
Web scraping, defined as the automated extraction of information online, is an increasingly important means of producing data in the social sciences. We contribute to emerging social science literature on computational methods by elaborating on web scraping as a means of automated access to information. We begin by situating the practice of web scraping in context, providing an overview of how it works and how it compares to other methods in the social sciences. Next, we assess the benefits and challenges of scraping as a technique of information production. In terms of benefits, we highlight how scraping can help researchers answer new questions, supersede limits in official data, overcome access hurdles, and reinvigorate the values of sharing, openness, and trust in the social sciences. In terms of challenges, we discuss three: technical, legal, and ethical. By adopting "algorithmic thinking in the public interest" as a way of navigating these hurdles, researchers can improve the state of access to information on the Internet while also contributing to scholarly discussions about the legality and ethics of web scraping. Example software accompanying this article are available within the supplementary materials.

Keywords Web scraping · Digital methods · Law · Ethics · Algorithmic thinking · Access to information · Social science research

1 Introduction

Digital data are revolutionizing research in the social sciences (Lazer et al. 2009; Millington and Millington 2015; Qiu et al. 2018; Keuschnigg et al. 2018). As social scientists turn toward computational methods to better understand the world, questions about how to

✉ Alex Luscombe
alex.luscombe@mail.utoronto.ca

[1] Centre for Criminology and Sociolegal Studies, University of Toronto, Toronto, ON M5S 3K9, Canada

[2] Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

[3] Department of Criminal Justice, University of Winnipeg, Winnipeg, MB R3B 2E9, Canada

Published online: 24 May 2021

do so with methodological sophistication demand greater attention. Despite the growing number of interventions on the topic (Cesare et al. 2018; Golder and Macy 2014; Hampton 2017; Lazer and Radford 2017; Salganik 2019; Tufekci 2014), there is still considerable work to be done outlining the kinds of digital and transactional data that now exist across the Internet, how to access these data, and how to make sense of them. Working at the nexus of social and computer science, we aim to advance methodological discussions about one way of collecting digital data for social science research: data scraping.[1]

Web scraping, defined as the automated extraction of information online, is becoming a crucial means of collecting data across the social sciences, including criminology (Bancroft 2019; Pina-Sánchez et al. 2019; Tzanetakis 2018), communication science (Nisser and Weidmann 2018; Possler et al. 2019), economics (Cavallo 2018; Massimino 2016), organization studies (Braun et al. 2018), policy studies (Anglin 2019; Caruana-Galizia and Caruana-Galizia 2018; Hayes and Scott 2018), planning studies (Boeing and Waddell 2017), political science (Fazekas and Tóth 2016; Ulbricht 2020), psychology (Landers et al. 2016; Qiu et al. 2018), and sociology (Mausolf 2017; Shi et al. 2017; Keuschnigg et al. 2018), among others.

Engagement with digital methods like web scraping is much needed for the social sciences to catch up with the pace of change in the world (Lazer and Radford 2017; Lazer et al. 2009; Burrows and Savage 2014; Edelmann et al. 2020). Automated collection of online data is an opportunity for creative collaboration between computer scientists, social scientists, and engineers, constituting a "watershed" moment in research (McFarland et al. 2016). As the efficacy and legitimacy of many traditional methods are called into question, it is becoming increasingly necessary to turn to innovative tools like scraping to carry the social sciences forward into the twenty-first century (Savage and Burrows 2007).

While there is undoubtedly some truth to the claim that automated techniques like web scraping differ "from traditional social science [research] where collecting data has always been hard, time consuming, and resource intensive" (Olmedilla et al. 2016, p. 79), the technique of data scraping is not without its challenges (Dewi et al. 2019; Massimino 2016; Landers et al. 2016; Munzert et al. 2014; Marres and Weltevrede 2013). In practice, scraping is often closer to an art than a science, and can take years of practice to master (Possler et al. 2019). At the same time, it is a craft that requires continuous learning and problem solving, particularly as website development evolves and becomes ever more complex and thereby less accessible using existing tools.

Contributing to debates about access to information via web scraping in the social sciences, we engage with the technical, legal, and ethical aspects of the practice in this article. In engaging with these issues, we advance the notion of algorithmic thinking in the public interest, a perspective that is consistent with Green and Viljoen's (2020) concept of "algorithmic realism." Algorithmic thinking in the public interest—much like algorithmic realism—involves upholding basic methodological principles of quality in social science in a manner that is simultaneously open about its normative political commitments and agenda (also see Flisfeder 2021; Possler et al. 2019). By adopting the term *algorithmic thinking*, a

[1] Other common monikers for data scraping include web scraping, screen scraping, web data extraction, web harvesting, and data harvesting. There are technical differences between the concepts of data "scraping" and website "crawling". A crawler is a bot that will navigate to a website for the purpose of indexing (i.e. record keywords and metadata) and then navigating to other websites via the links on that page. A scraper is a bot designed with the explicit intent on navigating and extracting specific information from one or multiple target websites. For the sake of simplicity, we conflate the two concepts here. Where differences exist, the two are contrasted in-text.

concept central to computer science pedagogy (Futschek 2006), we aim to instill in social scientists a sense of how technical detail matters, particularly for those who wish to study algorithmic reasoning and use it in their own work. These details matter practically, as they define what kinds of information we can obtain via web scraping. However, as we argue, these technical details also matter legally and ethically, as they inform the kinds of algorithmic decisions we are willing to make (e.g., in overcoming intentionally placed barriers to automated access). In much existing social scientific scholarship on data scraping as a method and on algorithms more generally, the technical mess of scraping is often absent from discussion and debate. This works against our ability to grapple with thorny questions about the legality and ethics of web scraping, as well as our ability to carve out a space for social scientists to be able to learn what web scraping through algorithmic reasoning entails. To avoid many of the pitfalls with mainstream understandings of algorithmic thinking in computer science and to make clear the normative political commitment that drives our approach to scraping algorithms, we refer to not only algorithmic thinking but algorithmic thinking *in the public interest*.

This article has three sections. First, we situate the practice of web scraping in context, providing an overview of how it works and how it compares to other methods in the social sciences. Second, we consider the promises of data scraping for social research. We highlight how scraping can help researchers answer new research questions, supersede limits in official data, overcome access hurdles, and reinvigorate the values of sharing, openness, and trust in the social sciences. Third, we discuss the technical, legal, and ethical challenges of data scraping. We argue social scientists ought to use algorithmic thinking in the public interest, which entails overcoming technical hurdles to scraping as a means of investigating government and corporate practices of power and governance. By adopting algorithmic thinking in the public interest—an epistemological and methodological standpoint of the social scientist as data scraper—we show how researchers will be positioned to not only improve the state of access to information on the Internet but can also contribute to discussions about the legality and ethics of data scraping.

2 Web scraping in context

Most undergraduate and graduate programs in the social sciences at this time do not offer formal training in algorithmically-driven techniques like web scraping (Possamai-Inesedy and Nixon 2017). For this reason, we feel it is necessary to provide a brief primer on how data scraping works. We also use this section to highlight the comparative advantage of open-source programming languages versus proprietary tools (and pay-to-play services) for scraping. Indeed, we argue that algorithmic thinking in the public interest is most conducive to one conducting their web scraping in an open-source language.

Data scraping can be conducted in three primary ways: using proprietary software, using paid custom web scraping services, or using open source software like Python, Javascript, or R. While using proprietary software (e.g., NVivo's NCapture, Tweetreach, Helium Scraper, Qualtrics) is less technically demanding, these options do not tend to provide the researcher with the same level of reach, flexibility, or transparency. Compared to free and open source programming languages like Python, proprietary software like NVivo's NCapture tend to be less transparent in their data collection procedures, difficult for future researchers to reproduce, and expensive. In addition, a number of paid services will

implement and maintain a scraper per your specifications, yet even the most basic plans can be very expensive (Table 1).

Algorithmic thinking in the public interest aligns most closely with web scraping via free and open source programming languages for two reasons. First, web scraping via a programming language as opposed to a point-and-click or low-code platform or service requires the researcher to possess a higher level of skill and familiarity with algorithmic reasoning, and forces the researcher to understand more of what goes into a web scraping algorithm to make it work. Second, scraping via a free and open source programming language as opposed to, say, pay-to-play proprietary software, allows the researcher to easily share their code with others, who can use, learn from, and adapt their algorithmic tools as they see fit. Open source languages allow for a more community-oriented, public interest based approach to programming (Von Krogh and Von Hippel 2006), values which are foundational to algorithmic thinking in the public interest.

To conduct a scrape in a program like Python or R, researchers can either rely on existing coding infrastructure built into packages like Selenium, Scrapy, rvest, and RSelenium, or build their own (Table 2). Packages like Selenium work by mimicking a user's web browser to access web pages and extract the desired content, and saving it locally on the user's hard drive. A combination of libraries often enables the creation of sophisticated scrapers able to more efficiently obtain the desired information as compared to the use of one library alone. For example, Python's Requests and LXML libraries are commonly used in combination with one another, where the former is efficient in downloading static content while the latter is efficient in parsing the desired information one "requested" or obtained.[2]

So long as information appears on a website, whether textual, auditory, or visual, it can in principle be accessed via web scraping. When navigating to a certain link, a user's browser (where the information is viewed) loads the content defined in the website's HTML static content and executes any scripts needed to generate dynamic content. The distinction between static versus dynamic content is key: static content is embedded within the HTML making its access as trivial as downloading a file from the internet and "parsing" (i.e. extracting) the desired information; dynamically-generated content requires the rendering of a browser-like environment for the content to appear and possibly more sophisticated interaction with the website (e.g., user log in) to obtain the desired information. When a human uses a browser to manually access a website this distinction is blurred as the content appears, yet when a scraper is used to access a website some elements may or may not be accessible based on the static/dynamic rendering and method of access. To access both static and dynamic content, data scraping algorithms need to be effectively written and designed, often by combining multiple libraries.

All scraping applications follow a prototypical format that can be expressed algorithmically (Algorithm 1). At their highest level of abstraction, a scraper is provided with a set of target URLs and a set of patterns to be matched. Before creating a scraper, some investigation into the page generation and layout is needed to determine the appropriate scraping library and relevant patterns to obtain. Once the code is written and executed, the scraper will download each URL and extract the relevant information from one or multiple pages based on some set pattern. For example, if automating the extraction of article titles and publication dates, a web scraper could download each of the article HTML pages from a

[2] Within the Supplementary Materials, we exemplify the combined use of several scraping libraries to achieve increasingly complex automated data extraction.

Table 1 Available paid services to implement custom data scrapers

Paid service	Brief description	Base pricing[a]	URL
Scraper API	API service that simplifies handling of proxies, browsers, and CAPTCHAs	$29/month	www.scraperapi.com
ScrapeSimple	Service that builds a scraper per your specifications, periodically emails a CSV of the results	$250/month	www.scrapesimple.com
Octoparse	Graphical interface to define scraped data in a point-and-click manner	$75/month	www.octoparse.com
Mozenda	Graphical interface to define scraped data in a point-and-click manner	$250/month	www.mozenda.com
Diffbot	Provides several APIs for various data extraction tasks	$299/month	www.diffbot.com

[a]All prices in US dollars

Table 2 Available web scraping libraries

Scraping library	Programming language	Particulars
Requests	Python	Useful to acquire raw HTML of a page (static content)
		High level of abstraction, easy to use in prototypical scraping
		Able to access APIs and post to forms
		Unable to access dynamically-generated context
Beautiful Soup	Python	Easy to learn and use (well documented)
		Ability to automatically detect encodings
		Much slower than other libraries
Lxml	Python	Very fast and useful for simple extractions
		Limited in the number of available features
Selenium	Python	Easily interface with dynamically-generated content; websites that generate dynamic content using Javascript requires a browser environment
		Enables automation of mouse interactions
		Unable to leverage proxies easily
Scrapy	Python	Amenable to use of proxies and VPNs
		Useful for complex scraping tasks
		Highly efficient (i.e. very fast) given its implementation
Request	Javascript	Useful to acquire the raw HTML of a page (static content)
		Unable to handle dynamically-generated content
Cheerio	Javascript	Requires numerous dependencies
Osmosis	Javascript	HTML parsing, DOM, and request features
		Fewer dependencies compared to other Javascript libraries
Puppeteer	Javascript	Chrome Browser automation software
		A Javascript-based implementation of Python's Selenium
Apify	Javascript	Complete web scraping framework
rvest	R	An R-based library with features similar to BeautifulSoup
RSelenium	R	An R-based implementation of Python's Selenium
Kimurai	Ruby	Scrapes dynamically-generated content
		Supports proxy/user-agent rotation, request delays, default headers
Goutte	PHP	PHP implementation of conventional scraping features

base URL and extract the text appearing next to the "Title" and "Date Published" fields on the web page.

Algorithm 1 Programmatic Pseudocode of a Generic Web Scraping Application.

Input: set of URLs to scrape, $\boldsymbol{u}$
set of patterns to match, $\boldsymbol{p}$
Output: set of scraped data, $\boldsymbol{d}$

```
1: import the scraping library
2: initialize the scraper
3: for each URL, u, in u do
4:     for each pattern, p, in p do
5:         use scraper to download the website content from u
6:         extract the pattern p from that content
7:         add extracted data to d
8:     end for
9: end for
```

Pseudocode is a means or representing the logic of an algorithm independent of its actual implementation in a given programming language. The application of a scraper generally requires the automated downloading of a website, identification of desired information within that content, and extraction for future analysis. Lines 1-2 of the algorithm describe the import and initialization of the scraping library. Lines 3-4 describe how, for each URL, each of the desired information are extracted according to a given pattern. Lines 5-6 operationalize the webpage downloading and information extraction, adding each to the resultant data object, such as a file or serialized representation.

There are at least three major differences between the kinds of information one obtains from scraping and more traditional social scientific methods like surveys, interviews, or experiments. The first is that the researcher typically seeks to obtain all data that are available rather than carefully collecting a representative sample of it (Allington 2016; Shi et al. 2017). Should the researcher decide they are only interested in a subset of the data, this decision would be made after the initial data collection procedures using parsing and classification algorithms. Anglin (2019, p. 688) has proposed what they call a "gather-narrow-extract" framework to guide data scraping: the researcher scrapes everything of potential value (gather), parses relevant from irrelevant text (narrow), and then "mines" the exact data they need from the classified corpus using additional search criteria (extract). This process is commonly referred to as "data wrangling" (Braun et al. 2018, p. 634).

Gathering as much material as is possible from a website does not mean that the data one collects is necessarily representative of the subject at hand. For one, "whole population" does not necessarily mean "whole region" or "whole country". It is possible, in principle, to scrape information from every user on a given platform. Whether this positions the researcher to make general claims about all users of the platform, or all users of a platform in a given country, however, is an empirical question to which the answer is most often no. While there are many reasons for this, one pertains to the general unevenness through which users engage with online platforms and services. Take Twitter data, for example. Previous research estimates that less than 1% of users have enabled Twitter's geotagging services (Edwards et al. 2013). This makes efforts to conduct any reliable location based analyses of Twitter users extremely difficult. In many cases, basic descriptive information about platform users or what is included and not included on a given website may not even be available to researchers, making it even more difficult to assess the external validity

of one's study (Salganik 2019). Moreover, there is question of what data scraped from a digital platform actually represents: can information scraped from Facebook, for instance, be used to generate insight into general human attitudes or behaviours, or are such insights necessarily limited to the platform they were drawn from?[3] Rather than this whole population versus sample analogy, we conceive of web scraping as a method *sui generis*, with its own unique challenges and limitations. As for the question of whether behavioural insights generated from the study of a particular platform can be generalized beyond this platform, there is unfortunately not yet any definitive answer to this question.

The second point of difference concerns less the process of collection (scraping) and more how researchers make sense of the information they obtain. While researchers can apply established techniques for analysing scraped data (e.g., regression modeling, social network analysis), these require that a series of extra steps are taken first to "clean" and "parse" the data, which can be technically demanding. Depending on the size of the data set one obtains from their scrape, a more conventional analytic approach like coding by hand or in NVivo is likely not feasible (Nelson et al. 2021). Sentiment analysis, topic modeling, word mover's distance, and text mining are some of the many computational methods of analysis that are emerging alongside data collection techniques like web scraping (Edelmann et al. 2020; Grimmer 2010; Nelson 2020; Schwartz and Ungar 2015; Stoltz and Taylor 2019; Abercrombie and Batista-Navarro 2020; Roberts et al. 2014). Such techniques, however, are very much still in development, with much future work required before the full range of their strengths and limitations is understood.

The size of the data set alone can bring many challenges. Many researchers, for example, have found "big data" sets call into question established statistical procedures (e.g., relying on p-value as an indicator of associational significance) (Lin et al. 2013). Because most traditional statistical techniques were designed for analysing smaller-N samples, they are not always appropriate for big-N data sets, especially those in the millions or billions (McFarland and McFarland 2015). Moreover, large data sets require considerable amounts of space (memory) to store and still greater amounts of space to process and analyse, particularly when using machine learning intensive approaches like topic modeling. The computationally intensive nature of analysing large data sets adds further practical complications to working with digital data sets, and may even require the researcher to learn advanced computational techniques like task parallelization and cluster computing.

The third point concerns the level of control over the categories used in data scraping. As Marres and Weltevrede (2013, p. 13) contend, "scraping seems to imply a distinctive approach to knowledge-making. [...] scraping formats the process of data collection and analysis as an operation of extraction, and organises knowledge-making as a distillation process". Unlike a survey or experiment, researchers scraping data have less control over how these are structured and classified. Categories set for commercial purposes may not readily map onto the goals of social scientific research, and it may be difficult to make general claims unless they have been validated against other data sets. Boeing and Waddell (2017, pp. 459–450), for example, had to work with Craigslist "subdomains" as their regions. Using web scraping to study a major darknet cryptomarket, Tzanetakis (2018) relied on the drug categories provided by the cryptomarket itself (known as 'AlphaBay'), which they acknowledged would limit future comparative work since each cryptomarket may not use the same categorisation system. Comparing data scraped on online interactions to traditional survey data, Hayes and Scott (2018, pp. 344–345) found that the two captured different features of the policy

[3] We thank the anonymous reviewer for this point.

networks they were studying, making scraping an "efficient (given ease of collection) but perhaps not very effective substitute" to survey-based network measures.

Such differences and limits are important to take into account, but should not be interpreted to mean that data scraping is an ineffective tool for social research. In some cases, it may be liberating to break free from the confines of existing data classifications (e.g., government census categories), particularly when those classifications are fraught with measurement bias (Cavallo 2018; Boeing and Waddell 2017) or unwantedly inhibit the kinds of research questions that can be asked. Moreover, it is important not to be overly nostalgic about methods such as phone-based surveying, whose golden age has long since passed (Savage and Burrows 2007). The increasing turn to web-based information and transactional data in the social sciences is as much by choice as by necessity, and the promises and limitations of computational techniques like data scraping should be read in this context. As we demonstrate in the next section, there is much promise in the method of web scraping, as a growing body of work from across the social sciences shows.

3 The promises of web scraping for social research

Web scraping techniques are increasingly being deployed across a wide range of social science disciplines. Many scholars are hopeful about the ways data scraping will transform their respective subfields (Anglin 2019; Boeing and Waddell 2017; Possler et al. 2019; Shi et al. 2017). The first promise of data scraping is that it can provide social researchers with access to data that may otherwise have been difficult or impossible to obtain. Tzanetakis (2018) used data scraping to study illicit drug trading on one of the darknet's largest cryptomarkets. Using 'AlphaBay' as their case study, Tzanetakis (2018) scraped a variety of data points including types of drugs sold by individual vendors, vendor pseudonyms, payment methods, pricing information, vendor country/region, customer feedback, and more. This allowed them to examine the structure and operation of one of the world's largest online illicit drug markets, identifying previously unknown trends, such as in the kinds of drugs bought and sold, shipping patterns, and the magnitude of financial transactions that took place on the market.

Where official data are limited, scraping can be an alternative means of obtaining data to answer social scientific questions. Maher et al. (2020) scraped congressional hearings and expert testimonies in the US between 1946 and 2016, allowing them to extract metadata about which disciplinary experts are called to testify more than others, data that is not otherwise available in official government statistics or databases. In their study of whether courts in the UK discriminate against Muslim-named offenders, Pina-Sánchez et al. (2019) had to use web scraping to produce the data they needed. In the UK—as in many other countries—the government has sought to maintain control over sentencing statistics by only publicly releasing them in a limited and aggregated format. The unwillingness of the British government to release this information is what forced Pina-Sánchez et al. (2019) to turn to the technique of data scraping (see also Pina-Sánchez et al. 2019a).

Another pivotal application of scraping to fill gaps in official data is Boeing and Waddell's (2017) study of the US rental housing market. Boeing and Waddell (2017) show how data scraping can be a powerful means of overcoming measurement biases built into official government data sources or data collected by more conventional means (also see Shi et al. 2017). Most data on the US rental housing market comes from two major sources: commercial listings, maintained by major corporate real estate entities, and the Census

Bureau's American Community Survey. But as Boeing and Waddell (2017) argue neither of these is sufficient to conduct a systematic analysis of trends in total housing market availability over time. Web scraping enabled Boeing and Waddell (2017) to generate, analyse, and make public "the most comprehensive data source currently available to examine [the US] rental housing market" (p. 469).

Even when data are hypothetically available using traditional methods, data scraping may be the only way researchers can gain access. This is the case where governments are unresponsive to researcher's requests, for instance, under freedom of information (FOI) law (Luscombe and Walby 2017). In a study of land use corruption, Caruana-Galizia and Caruana-Galizia (2018) turned to data scraping after FOI requests submitted to the Malta Environment and Planning Authority failed to yield access. Their first FOI request was denied, their second was ignored. Rather than file a third FOI, Caruana-Galizia and Caruana-Galizia (2018) scraped the data using a custom scraping algorithm they wrote in JavaScript.

Finally, web scraping has the potential to reinvigorate the values of sharing, openness, and trust in the social sciences (Li et al. 2019; Possamai-Inesedy and Nixon 2017). Researchers scraping publicly accessible information can share their code as an online supplement to their research reports. Future researchers can use this code to collect the same information or make minor adjustments to collect similar information from the same platforms. In an online supplement to their published results, Shi et al. (2017) made their Python code accessible online. Boeing and Waddell (2017) put their code into their lab's (UC Berkeley's Urban Analytics Lab) GitHub repository. Dick et al. (2020) shared all code and data for their Gas Prices of America project, which relied heavily on web scraping. Any researcher can re-run, adapt, and re-purpose the code from these projects.

This emphasis on sharing and transparency may also help to reinvigorate the value of trust in social science. The integrity of social scientific research has come under increasing fire in recent years, with overall trust in social research on the decline (Braun et al. 2018; Li et al. 2019). There are a growing number of fraud scandals in some academic disciplines, notably criminology and political science, and many of these could have been avoided if others were able to evaluate their code and raw data.

Despite the promising potential of web scraping as a data collection tool in the social sciences, technical, legal, and ethical hurdles abound. Yet, too often, these hurdles are not discussed in extant literature. In many studies that use web scraping as a method, researchers make significant technical, legal, and ethical decisions, but many of these are only apparent when reviewing their code and inspecting the terms and conditions of the web pages the authors scraped. The goal of discussing algorithmic thinking in the public interest is to make these dimensions of data scraping methodology visible and transparent. In some of the excellent studies summarized above, terms and conditions of use were (in our view, rightfully) broken and technical barriers overcome. The researchers appear to have done this in large part because they believed their use was fair and it was in the larger public interest to do so. In this sense, many seem to be already operating according to the principles of algorithmic thinking in the public interest, albeit without explicitly acknowledging it.

As Possler et al. (2019) note based on interviews with computational social scientists, researchers have a tacit sense of the technical, legal, and ethical challenges of data scraping and navigate these based on normative understandings of information control and data access. Below we elaborate on these technical, legal, and ethical challenges of web scraping for social research. Too often, these more challenging and contentious dimensions of data scraping are going undiscussed by the researchers using it. In numerous studies reviewed for this article, the authors made no mention of the fact that their scrapes required

creative, algorithmic circumvention of technical barriers and were in direct violation of a website's terms and conditions of use.

Rather than simply brush these more contentious technical, legal, and ethical dimensions of web scraping under the rug (although they almost always remain visible in any code that the researchers share online), we contend that researchers ought to make these aspects of their scrapes as well as the messy and at times contentious decisions they make clear and transparent. If web scraping is going to become an established method in the social sciences, explicit discussions of the technical, legal, and ethical dimensions of the practice are necessary. While it is commonplace to avoid questions about the legality of web scraping by characterizing it as a "grey area" (Possler et al. 2019, p. 3903), such tropes cannot effectively stand in the place of a more fully developed framework forever. At some point, social researchers will need to agree on some set of guiding principles and best practices for devising and deploying their web scraping algorithms. In what follows, we provide a detailed analysis of the technical, legal, and ethical dimensions of data scraping. Although we are unable to provide a complete set of definitive answers to many of the challenges and dilemmas we identify, some helpful guidance, we argue, can be found in what we call *algorithmic thinking in the public interest*.

4 The challenges of web scraping

4.1 Technical challenges

In the social sciences, where most researchers are only beginning to learn the basics of data scraping, technical challenges can be a serious barrier. Many websites are moving toward a greater reliance on dynamic programming languages like JavaScript, which are more difficult to scrape, requiring uses of more sophisticated programmatic techniques like headless browser scraping. Here, we want to focus on another kind of technical barrier: the webmaster-initiated defensive mechanism. Although such defensive mechanisms are increasingly pervasive on the Internet, many if not most social scientists engaged in web scraping appear reluctant to discuss them, even when the authors have taken steps to creatively circumvent them in their code. One plausible, albeit speculative, reason for this silence is that social researchers are worried about the implications of being perceived as "hackers", which has a more politically-charged and controversial connotation than a label like "web scraper". Defensive mechanisms are not illegal to implement but can make it exceptionally difficult to scrape information. As Bancroft (2019, p. 290) writes, the internet is a space permeated by "invisible gatekeepers, Twitter editors, algorithms, forum moderators and reCAPTCHAs which inhibit scraping". Eight defensive mechanisms in common use are summarized in Table 3.

All websites have been implemented by a webmaster who controls the presentation and availability of information on their web page(s). While implementing the website, any number of strategic defenses can be added that might deter bots or render a scraping task more complex. Each defensive strategy, from outright banning of IP addresses to requiring two-stage verification via email or mobile each require an incrementally more complex scraper implementation to circumvent that strategy. The most generic defense is the definition of the *robots.txt* file which advises corporate bots and sometimes general users on how to conduct scrapes and the best practice to follow while navigating their website. Broad protective measures to ensure that an abusive bot does not

Table 3 Eight defensive strategies to block web scraping

Defensive strategy	Explanation
1. Defining robots.txt	Explicit definition of how a website should be crawled by specific or all bots
2. Banning IP	When a large number of requests come from a specific IP address that specific IP is blocked from future requests
3. Rate-Limiting IP Requests	When an IP address sends requests at a rate quicker than humanly possible (milliseconds), requests are either returned slowly, rejected, or the IP is banned
4. User-Agent Blocking	When submitting a request, the "user-agent" identifies what browser is making the request; a bot would typically have a blank user-agent field
5. Banning by Navigation-Based Detection, e.g., reCAPTCHA	Sophisticated analyses of the history of a requester's navigation through a website filters out bot-based navigation. Most popular is Google's reCAPTCHA v3
6. Requiring Email Verification	An email verification step requires users to link their session to an email address
7. Requiring Mobile Verification	A mobile verification step requires users to link their session to a mobile phone number
8. Requiring an API key	For websites exposing a machine-friendly API, requiring a key to identify users is an effective strategy to limit resource usage

overwhelm their servers might include "blacklisting" an IP address if a high frequency of requests are made in a given period of time. More specific limitation of access to certain data might require the fulfillment of a vision-based CAPTCHA ("Completely Automated Public Turing test to tell Computers and Humans Apart") or two-step verification with a verified email address or phone number.

When making numerous requests to a webpage, each request is made with the IP address of origin. Past a certain threshold of requests, the website may temporarily "blacklist" a user's IP and refuse to serve the content crippling their scraper. The digital "fingerprint" of a human navigating a website is distinct from that of a bot navigating a website. The former might interact with a given website for a few seconds before navigating to another whereas a bot can make hundreds of requests to a website every second, behaviour that is well beyond the physical and cognitive limits of a human. Websites detecting this bot-like fingerprint might blacklist the bot's IP address or explicitly delay returning the requested resources to more human-like times, on the order of seconds. While accessing resources on the order of seconds might still appear quick, if there are hundreds of millions of data points to scrape, the scraper may need to be run for several weeks, months, or years before obtaining all desired data points.

When conducting a scrape, researchers generally assume that all desired data will be available in the specific format determined through their initial investigations. Should that presentation change, the scraping algorithm would no longer be able to extract the desired data and be rendered useless. Thus, another defensive mechanism that webmasters can use involves explicitly modifying the presentation of those targeted data to a non-unique, non-standard, or dynamically generated format requiring the modification of the scraper to adapt to those modifications. In a veritable digital access contest between the website provider and the creator of the scraper, the former can actively update their website to deter the latter, who, in turn, updates their scraper, *ad infinitum*.

Although technically demanding, many of these barriers can be overcome with the right algorithmic design (Table 4). By randomly selecting an IP address from a "pool", a scraper can be made to appear as though the numerous or high-frequency requests arrive from multiple and unique users around the world. Using fulfillment services or temporary "burner" email addresses or phone numbers, a scraper can be made to overcome CAPTCHAs and two-step verification. Data collection always entails dancing through barriers to access, and so too with web scraping. As Mitchell (2018, p. 216) notes, "it is almost impossible to build a 'scraper proof' site." While some scholars have advised against researchers seeking technical workarounds to the barriers put in place to inhibit scraping, we argue that sometimes it is necessary and justified, particularly when conducting research on powerful government and corporate entities. In such cases, it is algorithmic thinking which helps us devise creative solutions to obtain data, justified by the fact that the research we are conducting is ultimately in the public interest. Once one had devised an algorithm capable of getting around defensive barriers, the next question is whether one *should* deploy such an algorithm, a question of law and ethics.

4.2 Legal ambiguities

Many website hosts have sought to inhibit automated access via data scraping by invoking law (Din 2015; Drivas 2019; Scassa 2019). Many websites now require users to agree to "terms and conditions" that explicitly prohibit data scraping. A recent Canadian government initiative, for example, requires users of its business registries service, a repository of public information about registered businesses across the country, to agree as a condition of access that they will not "use automated tools to copy data from this service" (Fig. 1). Violating a website's terms and conditions agreement may result in a "cease and desist" letter (popular in the US), or worse, a lawsuit claiming unauthorized access or copyright violation. To date, the majority of lawsuits around data scraping have been in the US, and have tended to be between competing business interests (Scassa 2019, pp. 987–988). To our knowledge, there have been no lawsuits in the US, Canada, UK, or other major western democratic countries stemming from a researcher scraping publicly accessible data from a website for personal or academic use.

This may, however, change and researchers could soon find themselves in tricky legal situations. As web scraping becomes a more widely used method in social science disciplines, we may see an increase in litigation against researchers that make the data they scraped or the code they used to do it public, perhaps even as a requirement of institutional funding agencies (Braun et al. 2018, p. 640). Many funding bodies now require researchers to make any data collected publicly available, but this could put the researcher that relied on data scraping to obtain data at risk of litigation.

With the explosion of data availability via the Internet, as Braun et al. (2018, p. 640) argue, "comes an increased responsibility of researchers to ensure they are in compliance with data usage rules and regulations". These regulations, however, are anything but clear in most cases. As a result, some researchers have called on researchers to proceed with caution, scraping only what seems to fall unambiguously within the ambit of the law. Landers et al. (2016, p. 487) for example, recommend that researchers "only scrape publicly available, unencrypted data sources to avoid legal risk". Our thinking is that this point of view concedes too much digital ground to governments and corporations. There are more context-specific, creative ways for researchers to conduct scrapes without violating the law, even if a website discourages it. A good example comes from Maher et al. (2020), who

Table 4 Workarounds to defensive solutions

Defensive strategy	Scraper solution	Explanation
1. Banning IP	Rotating IP Addresses	A bot can spoof the requesting IP address by relying on a pool of IP addresses; a few requests are made and then the IP is switched preventing IP tracking
2. Rate-Limiting IP Requests	Rate-Limiting Requests	A bot can implement a delay to its requests to mimic a human requester; at times, a random amount of delay is used since a regular and identical interval of requests is also detectable
3. User-Agent Blocking	Spoofing and Rotating User-Agent	A scraper can set the user-agent to mimic a given browser (e.g. Chrome) and rotate to other browsers on each request (e.g. Mozilla Firefox)
4. Banning by Navigation-Based Detection, e.g., reCAPTCHAs	Spoofing human-like cookies; paid CAPTCHA solving services; human-mimicking scraping library (e.g., Selenium)	The navigation history of a bot differs dramatically from human-based navigation; the cookies from a human-type session can be used to mask a bot-type session
5. Requiring Email Verification	Pool of emails; use of "burner" email addresses	Either a pool of email addresses can be rotated through, or a paid service of temporary/"burner" emails can be used
6. Requiring Mobile Verification	Use of "burner" mobile phone number service (e.g., Twilo)	As it is expensive to maintain active phone numbers, a one-time use of a "burner" mobile number & messaging service cheaply circumvents this strategy
7. Requiring an API key	Rotating pool of API keys	As long as API keys are free, a pool of keys can be generated linking to "different users"

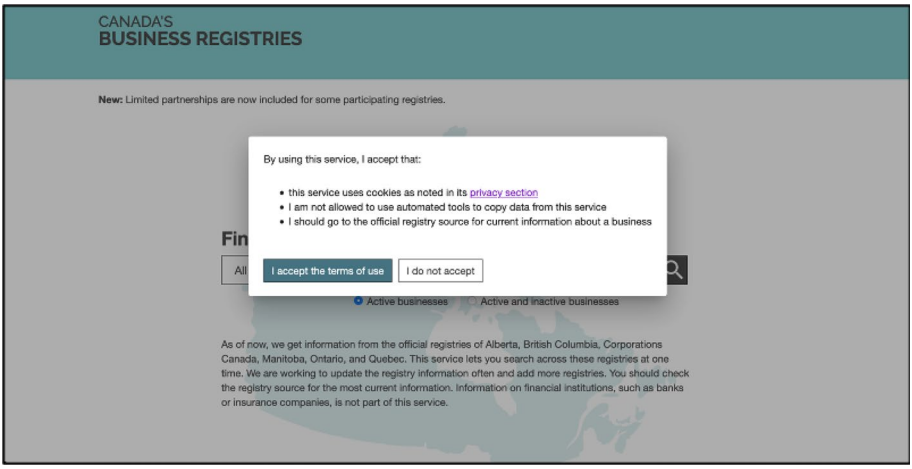

Fig. 1 Canada's Business Registries, a free, public data repository containing information about registered Canadian businesses, requires users to accept the terms of their terms of use as a condition of access. Using the service, users can obtain useful information about Canadian companies, including their business number, registry ID, registered office location, company status (active/inactive), business type, and date created. This is precisely the kind of information researchers would benefit from being able to scrape. The second condition prohibits the user from scraping: "I am not allowed to use automated tools to copy data from this service". Source: https://beta.canadasbusinessregistries.ca/search

carefully scraped a ProQuest database of congressional hearings and expert testimonies in a way that would not violate the website's terms and conditions. This involved reading ProQuest's terms and conditions in light of their intended use and role as academics. As Maher et al. (2020, p. 3) reflect:

> The database contains full lists of the committees and sub-committees which convened the hearing, the date of the hearing, as well as information about each of the witnesses that testified before each congressional hearing, including their names, organizational affiliation, and titles. Our search complies with ProQuest's Terms and Conditions because the search is for research and analysis purposes, uses only reasonable portions of the data, (here, data on social scientists who testify), and the underlying dataset only shares data for a portion of the material without replacing future users' need to work through ProQuest (or other points of access to the government record) to access the full scope of the data. Further, the material we collect (hearing dates, topics, and witness lists) is publicly available through several other sources; we do not collect or use ProQuest's proprietary transcripts of the testimony. Moreover, since these data points are facts, they are not subject to US copyright.

Terms and conditions are often not written with academics or public interest research in mind (Scassa 2019). Social scientists need to be aware of this when scraping information, approaching terms and conditions more from the perspective of a lawyer than simple user.

What we call algorithmic thinking in the public interest involves social scientists scraping information if it is publicly available and the research is in the public interest, and doing so in ways backed by legal arguments which may eventually be tested in court. We argue for this approach for three reasons. First, if researchers do not become involved in legal disputes over access to information via data scraping, it risks letting the Internet become a space legally circumscribed by disputes between competing business interests.

As Scassa (2019, p. 987) argues in regard to the legality of web scraping, "legal uncertainties in relation to ownership of and rights of access to data risk being resolved by litigation between business competitors, which risks overlooking and unduly limiting the strong public interest in access to and use of such data." It is necessary to "[...] prevent the normative framework for data scraping from being unduly shaped by the platforms themselves" (*ibid.*), which is a goal of algorithmic thinking in the public interest.

Second, we argue against the idea that researchers should avoid engaging in legally ambiguous activities in regard to data scraping because such a position overlooks the extent to which governments and corporations are and always have been committed to widespread information control. Governments and corporations have long sought to maintain control over information, and just as this has not stopped researchers in other contexts (military secrecy, for example) it should not stop them with data scraping. If a government agency has asked the public not to scrape information in its terms of use, we acknowledge that there could be a valid reason. Yet it may not be a valid reason. Website 'scrapeability' is something that researchers ought to approach critically. They should question the reasons why a government or corporate entity has requested the information not be scraped. It could be for a dubious reason, such as not wanting researchers to be able to calculate aggregate level trends (or in the case of a corporation, because they worry you are the competition).

Third, we argue in favour of researchers approaching the legal ambiguity of data scraping with a critical attitude because it is what disciplines like sociology need. Social research has become risk averse (Haggerty 2004). This is not regressive when it comes to researching "down" on people who are marginalized, but it is a problem when it comes to researching "up" (Nader 1968) and investigating powerful multinational corporations and government agencies. Social scientists should operate with a different set of ethical considerations when examining corporations and government agencies (Galliher 1979). Avoiding scraping because it may be illegal or unwanted by the provider goes against what social science needs to reinvigorate its public and political relevance, which is to be more investigative, more critical and situational in the ethical decisions it makes, which does entail the possibility of legal liability and risk.

4.3 Ethical considerations

Scraping poses unique ethical challenges to which there are no easy answers (Ravn et al. 2020). The ethical guidelines for the use of online data by social scientists remain a "moving target" (Ackland 2013, p. 43). New forms of digital data and techniques for obtaining them are changing, and ethical review boards have yet to solidify a coherent policy stance on the matter. The increasing adaptation of computational techniques like web scraping in the social sciences should compel researchers to revisit long held assumptions about research ethics, including informed consent, privacy, and anonymity (Sugiura et al. 2017, p. 185). Until ethics review boards improve their technical expertise, they will remain ill-equipped to regulate the ethical challenges that are likely to emerge from the use of data scraping as a methodology for social research (Lazer et al. 2009; Felderer and Blom 2019). Though some researchers seem to have obtained formal ethical approval for their scrapes, most in our reading have not.

When conducting a data scrape, researchers need to be aware of how their requests can cause harm. The most obvious harm comes in the form of an unintended "denial of service" (DoS) attack, which occurs when the high frequency and duration of a

researcher's scraping algorithm overwhelms a website's server. It is important for researchers to be cognizant of this potential harm when conducting scrapes, and take steps to mitigate it. Sometimes it means not getting data even when it is available, reducing the scope of the project to ensure access for others. In their study of rental listings, for example, Boeing and Waddell (2017, p. 468) were aware of this issue, noting how even though daily data was available, they avoided it believing that "constant collection would overburden Craigslist's servers".

Algorithmic thinking in the public interest does not mean throwing legal and ethical responsibilities to the wayside. Rather, researchers need to develop ethical, political, and technical guidelines to enable research in the public interest and to encourage others to do the same. A number of rules define the "best practices" between those hosting information online and those attempting to access that information. Certain limits can be made more explicit than others. As noted above, foremost is the definition of the *robots.txt* file on a website. When present, the website owner is describing the behaviour that bots (crawlers or scrapers) are supposed to adopt while navigating their site. However, the presence of this file will not prevent bots from abusing resources or irresponsibly overburdening website servers. The naïve implementation of a website scraper could negatively affect other users accessing the website such as slowing down the website response and, in extreme cases, result in "crashing" the server. Websites can only handle a finite number of requests over a given period of time and a bot is capable of accessing website resources hundreds to thousands of times faster than a human and, thereby, rapidly consuming those resources that would otherwise serve thousands of individual human users. Table 5 describes these differences.

Respect for the *robots.txt* file is the first of the best practises. In the absence of this file, a scraper should implement a delay (generally between 3 and 10 seconds) in their scraper to limit the frequency of requests. Running a scraper outside of peak working hours can further reduce the impact of a scraper on a given server. Another ethical concern stems from the data itself. Data scraping can open up the possibility for invasions of privacy by aggregating data forms. In Gregory's (2018, p. 1618) view, current applications of web scraping techniques, particularly in sensitive research sites like health, can be justified so long as researchers are careful to remove personal identifying information from the raw data. An additional ethical consideration is whether researchers should scrape websites that are difficult to scrape or that have obstacles built into them, as discussed above, even if these contain information that is in the public interest.

The answer to these and other ethical concerns, at least from the perspective of algorithmic thinking in the public interest, is that it *depends*—it depends on the data, on the research question, and on one's own politics and agenda. Deriving an answer to these questions will never be simple. Such ethical decisions are complex and multi-faceted. As Tracy (2010) argues, research ethics goes beyond just obtaining formal review board approval. The researcher must also take into consideration situational and relational ethics. Situational ethics is particularly relevant with respect to our arguments about data scraping. Situational ethics pertain to context-specific ethical considerations and decisions a research must make in their research. These decisions often involve consideration of factors beyond just review board approval, such as "the greater good" (Tracy 2010, p. 847). In the context of scraping, situational ethics means thinking about many of the potential harms mentioned above, such as slowing down or inadvertently shutting down a website's server, but also the larger issue of automated access in the public interest. There needs to be more debate about the ethics of automated access to digital information. If researchers believe that this kind of access is in the public interest, they

Table 5 Example robots.txt files with explicit limitations

Full access	Refuse all access	Partial access	Crawl rate limit	Set visit time	Request rate limit
User-agent: * Disallow:	User-agent: * Disallow: /	User-agent: * Disallow: /folder/ User-agent: * Disallow:/file.html	User-agent: * Disallow: Crawl-delay: 11	User-agent: * Disallow: Visit-time: 0400-0845	User-agent: * Disallow: Request-rate: 1/10

should be willing to take reasonable risks to conduct this research and justify it on ethical grounds.

5 Discussion and conclusion

Contributing to debates about web scraping and digital methods in the social sciences, we have argued in favour of data scraping as a potent means of access to information in the digital era. Researchers in the social sciences are using data scraping as a method with increasing frequency, overcoming limits in other more traditional information sources, and surmounting hurdles to access. Data scraping could also help reinvigorate the values of sharing, openness, and trust at a time of mounting fraud and integrity scandals in the social sciences. Relatedly, we have suggested that web scraping and collection of online data could help invigorate multi-disciplinary research collaboration (McFarland et al. 2016).

Unfortunately, most social scientists do not yet possess the skills to evaluate, re-run, adapt (let alone write *de novo*), web scraping code (Boeing and Waddell 2017, p. 468; Braun et al. 2018, p. 634). As trends like the "computational social science" movement (Lazer et al. 2009) and cross-disciplinary collaboration increase, this will change. Beyond acquiring basic coding skills, we call on social scientists to develop "algorithmic thinking" (Futschek 2006; Green and Viljoen 2020), a way of rethinking social science problems as well as data collection and analysis strategies to respond to the digitization of social life, using the tools of computer science in a way that advances the public good.

Finally, we have discussed the technical, legal, and ethical challenges that confront the social scientist as data scraper. Learning how to navigate these barriers is a crucial part of mastering data scraping as a methodological "craft" (Tracy 2010). With respect to the legality and ethics of web scraping, we argue that researchers are at a crucial junction. On the one hand, social scientists could avoid becoming embroiled in the legal and ethical disputes over automated access to information altogether. On the other hand, social researchers could dive straight in, capitalizing on this opportunity to help define the legal and ethical boundaries of data scraping and advocating for access to data in the public interest.

We encourage researchers to become immersed in the practice of data scraping, to undertake algorithmic thinking in the public interest, and, when legally and ethically justifiable, overcome the obstacles website providers have used to block access in the absence of a defensible rationale. Perhaps most importantly, we have sought to address these matters directly in a reflexive and transparent manner. We encourage other social researchers using web scraping in their future studies to likewise be clear about their approach to the technical, legal, and ethical challenges of data scraping, and to consider algorithmic thinking in the public interest as a way of navigating and confronting the challenges of online, automated social science research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11135-021-01164-0.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abercrombie, G., Batista-Navarro, R.: Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. J. Comput. Soc. Sci. **3**, 245–270 (2020)

Ackland, R.: Web social science: concepts, data and tools for social scientists in the digital age. Sage, Thousand Oaks (2013)

Allington, D.: Linguistic capital and development capital in a network of cultural producers: mutually valuing peer groups in the 'interactive fiction' retrogaming scene. Cult. Sociol. **10**(2), 267–286 (2016)

Anglin, K.L.: Gather-narrow-extract: a framework for studying local policy variation using web-scraping and natural language processing. J. Res. Educ. Eff. **12**(4), 685–706 (2019)

Bancroft, A.: Research in fractured digital spaces. Int. J. Drug Policy **73**, 288–292 (2019)

Boeing, G., Waddell, P.: New insights into rental housing markets across the united states: web scraping and analyzing craigslist rental listings. J. Plan. Educ. Res. **37**(4), 457–476 (2017)

Braun, M.T., Kuljanin, G., DeShon, R.P.: Special considerations for the acquisition and wrangling of big data. Organ. Res. Methods **21**(3), 633–659 (2018)

Burrows, R., Savage, M.: After the crisis? Big data and the methodological challenges of empirical sociology. Big Data Soc. **1**(1), 2053951714540280 (2014)

Caruana-Galizia, P., Caruana-Galizia, M.: Political land corruption: evidence from Malta-the European union's smallest member state. J. Public Policy **38**(4), 419–453 (2018)

Cavallo, A.: Scraped data and sticky prices. Rev. Econ. Stat. **100**(1), 105–119 (2018)

Cesare, N., Lee, H., McCormick, T., Spiro, E., Zagheni, E.: Promises and pitfalls of using digital traces for demographic research. Demography **55**(5), 1979–1999 (2018)

Dewi, L.C., Chandra, A., et al.: Social media web scraping using social media developers api and regex. Procedia Comput. Sci. **157**, 444–449 (2019)

Dick, K., Charih, F., Woo, J., Green, J.R.: Gas prices of America: the machine-augmented crowd-sourcing era. In: 2020 17th Conference on Computer and Robot Vision (CRV), pp. 158–165. IEEE (2020)

Din, M.F.: Breaching and entering: when data scraping should be a federal computer hacking crime. Brooklyn Law Rev. **81**, 405 (2015)

Drivas, I.: Liability for data scraping prohibitions under the refusal to deal doctrine. Univ. Chic. Law Rev. **86**(7), 1901–1940 (2019)

Edelmann, A., Wolff, T., Montagne, D., Bail, C.A.: Computational social science and sociology. Ann. Rev. Sociol. **46**, 61–81 (2020)

Edwards, A., Housley, W., Williams, M., Sloan, L., Williams, M.: Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation. Int. J. Soc. Res. Methodol. **16**(3), 245–260 (2013)

Fazekas, M., Tóth, I.J.: From corruption to state capture: a new analytical framework with empirical applications from Hungary. Polit. Res. Q. **69**(2), 320–334 (2016)

Felderer, B., Blom, A.G.: Acceptance of the automated online collection of geographical information. Sociol. Methods Res. 0049124119882480 (2019)

Flisfeder, M.: Algorithmic Desire: Toward a New Structuralist Theory of Social Media. Northwestern University Press, Evanston (2021)

Futschek, G.: Algorithmic thinking: the key for understanding computer science. In: International Conference on Informatics in Secondary Schools-Evolution and Perspectives. Springer, pp. 159–168 (2006)

Galliher, J.F.: Social scientists' ethical responsibilties to superordinates: looking upward meekly. Soc. Probl. **27**, 298 (1979)

Golder, S.A., Macy, M.W.: Digital footprints: opportunities and challenges for online social research. Ann. Rev. Sociol. **40**, 129–152 (2014)

Green, B., Viljoen, S.: Algorithmic realism: expanding the boundaries of algorithmic thought. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 19–31 (2020)

Gregory, K.: Online communication settings and the qualitative research process: acclimating students and novice researchers. Qual. Health Res. **28**(10), 1610–1620 (2018)

Grimmer, J.: A bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. Polit. Anal. **18**(1), 1–35 (2010)

Haggerty, K.D.: Ethics creep: governing social science research in the name of ethics. Qual. Sociol. **27**(4), 391–414 (2004)

Hampton, K.N.: Studying the digital: directions and challenges for digital methods. Ann. Rev. Sociol. **43**, 167–188 (2017)
Hayes, A.L., Scott, T.A.: Multiplex network analysis for complex governance systems using surveys and online behavior. Policy Stud. J. **46**(2), 327–353 (2018)
Keuschnigg, M., Lovsjö, N., Hedström, P.: Analytical sociology and computational social science. J. Comput. Soc. Sci. **1**(1), 3–14 (2018)
Landers, R.N., Brusso, R.C., Cavanaugh, K.J., Collmus, A.B.: A primer on theory-driven web scraping: automatic extraction of big data from the internet for use in psychological research. Psychol. Methods **21**(4), 475 (2016)
Lazer, D., Radford, J.: Data ex machina: introduction to big data. Ann. Rev. Sociol. **43**, 19–39 (2017)
Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Computational social science. Science (New York, NY) **323**(5915), 721–723 (2009)
Li, F., Zhou, Y, Cai, T.: Trails of data: Three cases for collecting web information for social science research. Soc. Sci. Comput. Rev. (OnlineFirst) (2019)
Lin, M., Lucas, H.C., Jr., Shmueli, G.: Research commentary-too big to fail: large samples and the p-value problem. Inf. Syst. Res. **24**(4), 906–917 (2013)
Luscombe, A., Walby, K.: Theorizing freedom of information: the live archive, obfuscation, and actor-network theory. Gov. Inf. Q. **34**(3), 379–387 (2017)
Maher, T.V., Seguin, C., Zhang, Y., Davis, A.P.: Social scientists' testimony before congress in the united states between 1946–2016, trends from a new dataset. PLoS ONE **15**(3), e0230104 (2020)
Marres, N., Weltevrede, E.: Scraping the social? Issues in live social research. J. Cult. Econ. **6**(3), 313–335 (2013)
Massimino, B.: Accessing online data: web-crawling and information-scraping techniques to automate the assembly of research data. J. Bus. Logist. **37**(1), 34–42 (2016)
Mausolf, J.G.: Occupy the government: analyzing presidential and congressional discursive response to movement repression. Soc. Sci. Res. **67**, 91–114 (2017)
McFarland, D.A., McFarland, H.R.: Big data and the danger of being precisely inaccurate. Big Data Soc. **2**(2), 2053951715602495 (2015)
McFarland, D.A., Lewis, K., Goldberg, A.: Sociology in the era of big data: the ascent of forensic social science. Am. Sociol. **47**(1), 12–35 (2016)
Millington, B., Millington, R.: 'The datafication of everything': toward a sociology of sport and big data. Sociol. Sport J. **32**(2), 140–160 (2015)
Mitchell, R.: Web Scraping with Python: Collecting More Data from the Modern Web. O'Reilly Media, Newton (2018)
Munzert, S., Rubba, C., Meißner, P., Nyhuis, D.: Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Wiley, Hoboken (2014)
Nader, L.: Up the anthropologist: perspectives gained from 'studying up'. In: Hymes, D. (ed.) Reinventing Anthropology, pp. 284–311. Random House, New York (1968)
Nelson, L.K.: Computational grounded theory: a methodological framework. Sociol. Methods Res. **49**(1), 3–42 (2020)
Nelson, L.K., Burk, D., Knudsen, M., McCall, L.: The future of coding: a comparison of hand-coding and three types of computer-assisted text analysis methods. Sociol. Methods Res. **50**(1), 202–237 (2021)
Nisser, A., Weidmann, N.B.: Online ethnic segregation in a post-conflict setting. Eur. J. Commun. **33**(5), 489–504 (2018)
Olmedilla, M., Martínez-Torres, M.R., Toral, S.: Harvesting big data in social science: a methodological approach for collecting online user-generated content. Comput. Stand. Interfaces **46**, 79–87 (2016)
Pina-Sánchez, J., Grech, D., Brunton-Smith, I., Sferopoulos, D.: Exploring the origin of sentencing disparities in the crown court: using text mining techniques to differentiate between court and judge disparities. Soc. Sci. Res. **84**, 102343 (2019)
Pina-Sánchez, J., Julian, V.R., Sferopoulos, D.: Does the crown court discriminate against Muslim-named offenders? A novel investigation based on text mining techniques. Br. J. Criminol. **59**(3), 718–736 (2019a)
Possamai-Inesedy, A., Nixon, A.: A place to stand: digital sociology and the Archimedean effect. J. Sociol. **53**(4), 865–884 (2017)
Possler, D., Bruns, S., Niemann-Lenz, J.: Data is the new oil-but how do we drill it? Pathways to access and acquire large data sets in communication science. Int. J. Commun. **13**, 3894–3911 (2019)
Qiu, L., Chan, S.H.M., Chan, D.: Big data in social and psychological science: theoretical and methodological issues. J. Comput. Soc. Sci. **1**(1), 59–66 (2018)

Ravn, S., Barnwell, A., Barbosa Neves, B.: What is "publicly available data"? Exploring blurred public-private boundaries and ethical practices through a case study on Instagram. J. Empir. Res. Hum. Res. Ethics **15**(1–2), 40–45 (2020)
Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Albertson, B., Gadarian, S., Rand, D.: Topic models for open ended survey responses with applications to experiments. Am. J. Polit. Sci. **58**, 1064–82 (2014)
Salganik, M.J.: Bit by bit: social research in the digital age. Princeton University Press, Princeton (2019)
Savage, M., Burrows, R.: The coming crisis of empirical sociology. Sociology **41**(5), 885–899 (2007)
Scassa, T.: Ownership and control over publicly accessible platform data. Online Inf. Rev. **43**(6), 986–1002 (2019)
Schwartz, H.A., Ungar, L.H.: Data-driven content analysis of social media: a systematic overview of automated methods. Ann. Am. Acad. Pol. Soc. Sci. **659**(1), 78–94 (2015)
Shi, F., Shi, Y., Dokshin, F.A., Evans, J.A., Macy, M.W.: Millions of online book co-purchases reveal partisan differences in the consumption of science. Nat. Hum. Behav. **1**(4), 1–9 (2017)
Stoltz, D.S., Taylor, M.A.: Concept mover's distance: measuring concept engagement via word embeddings in texts. J. Comput. Soc. Sci. **2**(2), 293–313 (2019)
Sugiura, L., Wiles, R., Pope, C.: Ethical challenges in online research: public/private perceptions. Res. Ethics **13**(3–4), 184–199 (2017)
Tracy, S.J.: Qualitative quality: eight "big-tent" criteria for excellent qualitative research. Qual. Inq. **16**(10), 837–851 (2010)
Tufekci, Z.: Big questions for social media big data: representativeness, validity and other methodological pitfalls. arXiv:14037400 (2014)
Tzanetakis, M.: Comparing cryptomarkets for drugs. A characterisation of sellers and buyers over time. Int. J. Drug Policy **56**, 176–186 (2018)
Ulbricht, L.: Scraping the demos. Digitalization, web scraping and the democratic project. Democratization **27**(3), 426–442 (2020)
Von Krogh, G., Von Hippel, E.: The promise of research on open source software. Manag. Sci. **52**(7), 975–983 (2006)