RESEARCH ARTICLE | OCTOBER 01 2023

Scrapism: A Manifesto FREE

Sam Lavigne

Critical AI (2023) 1 (1-2)

https://doi.org/10.1215/2834703X-10734046

Abstract

Web scraping is a technique for automatically downloading and processing web content or converting online text and other media into structured data. This article describes the role that web scraping plays for web businesses and machine learning systems and the fundamental tension between the openness of the web and the interests of private corporations. It then goes on to sketch an outline for "scrapism," the practice of using web scraping for artistic, critical, and political ends.

Issue Section: Articles

Keywords: web scraping, data, machine learning, internet art

Link by link we build paths of understanding across the web of humanity.

—Tim Berners-Lee

Web scraping describes techniques for automatically downloading and processing web content or converting online text and other media into structured data that can then be used for various purposes. In short, a user writes a program to browse and analyze the web on their behalf, rather than doing so manually. Web scraping, as with many other automated procedures, is useful primarily because it allows individuals and organizations to acquire data quickly and at a far larger scale and lower budget than would otherwise be possible.

Scraping is a common practice in Silicon Valley, where materials pulled from open HTML pages are inserted into corporate databases, and, as such, transformed into private property. In many cases, scraping has provided the raw material from which tech empires are built and fortunes made. Facebook, for example, began as "FaceMash," a (horny) web scraping project that allowed Mark Zuckerberg to rank the "hotness" of his classmates (Losse 2012). The popular real estate site StreetEasy got its start by scraping listings from the web (Lindsay 2022). Finally, Google—and all other search engines—are also, at heart, web scrapers, as they collect, store, and index web content.[1] Their lucrative business model rests on the commodification of this data.

Importantly, web scraping is also used extensively to collect training material for machine learning (ML) systems. For example, OpenAI's text generator GPT-3 (the precursor to the recently released ChatGPT) is based primarily on the open source Common Crawl dataset of scraped websites (Brown et al. 2020). ImageNet (Deng et al. 2009) and Microsoft's COCO (Lin et al. 2015)—datasets that have become industry standards and that power numerous ML systems—are scraped from image-sharing sources like Flickr and the wider web (see also Sluis and Malevé in this issue). The automated code-generation service Copilot is drawn from GitHub's open code repositories. Perhaps most notoriously, the right wing–affiliated facial recognition service Clearview AI (O'Brien 2020), which law enforcement uses to "accelerate" investigations (i.e., to find and arrest people identified through the technology), derives its model from profile images scraped from social media accounts (Hill 2020). In short, web scraping is ubiquitous.

As these examples show, web scraping is used widely to acquire, enclose, commodify, and monetize data from the open web, as well as to enhance the work of repressive carceral forces. However, as I will argue in this article, web scraping also has a rich

potential to further liberatory agendas. Just as it can be used to enclose, it can also be a tool to deprivatize and decommoditize data and adjacent media. In addition, because private organizations tend to mediate social, political, and economic interactions through public-facing websites, web scraping can play a vital role in revealing, and possibly undermining, contemporary power structures, both online and offline.

Scrapism, as I define it, is a counterpractice of web scraping for artistic, emotional, political, and critical ends, rather than for those of business or government. It is a process of decommodification and redatabasing, a process of eliminating artificial scarcity. At heart, it is a practice that challenges the regime of private property, with a particular focus on the ways that private property, as expressed on the web, produces and reproduces informational and material power asymmetries.

A Website Is Made to Be Scraped

The first automated web crawler, "The Wanderer," was written in 1993, in an attempt to track and understand the exponential growth of the early web. Matthew Gray, the project's author, announced: "I have written a perl script that wanders the WWW collecting URLs, keeping tracking [*sic*] of where it's been and new hosts that it finds. Eventually, after hacking up the code to return some slightly more useful information (currently it just returns URLs), I will produce a searchable index of this" (Gray 1993).

The Wanderer was created in response to a basic contradiction of the web: it is both *imperceptible* and *open*. On the one hand, the web's infrastructure is invisible to users. Because the HTTP protocol does not provide an index of all web addresses, there is no internal, centralized way of knowing what websites exist. On the other hand, websites are presented in HTML, an open plain-text markup language that is legible both to humans and machines, and that allows documents to be linked to each other. As a result, once you know about a website's existence, it is possible to write a program to parse and analyze its content, and, in this way, to create a map of connected websites.

Of course, the fundamental openness of HTML was an explicitly stated goal and utopian selling point: a technical backbone on which to realize a vision of shared information as a decentralized, ad hoc library of human knowledge. Describing the chain of ideas that led to his invention of the web, Tim Berners-Lee (1999: 4) collaborated with Mark Fischetti to write: "*Suppose all the information stored on computers everywhere were linked*, I thought. *Suppose I could program my computer to create a space in which anything could*

be linked to anything. All the bits of information in every computer at CERN, and on the planet, would be available to me and to anyone else. There would be a single, global information space."

Less obvious is how this very openness introduces the need for external systems to map the network; to this extent, HTML was literally built to be scraped. This implicit reliance on scraping points to the many ways in which the web's openness has always been contested and unfixed. There is a fundamental tension between the vision of universal, shared access to knowledge represented by open, interlinked HTML documents, and the artificial, capitalist scarcities imposed by copyrights, data hoarding, surveillance, patents, and so on. The technological complement to this ideological tension is the selective deployment of antiscraping measures such as rate limits, captchas, paywalls, geofencing, and/or the stipulation of user accounts to limit access to content. As against Berners-Lee and Fischetti's "single, global information space," these technical barriers reflect, produce, and/or reinforce asymmetries of power through informational monopolies and surveillance regimes.

A Website Is a Database

When the web was born, the vast majority of sites were simple HTML documents. When a user looked at a web page, an HTML file was literally copied from the website's server to the user's computer. Web developers, however, quickly learned to deploy data stores and dynamic page generation to create sites that included features such as search, user accounts, and user-uploadable content. Although some sites continue to serve static HTML pages, today the vast majority of web content is stored in databases. When a user looks at a page stored in a database, a web server processes the browser's request, reads from its database(s), and then generates HTML to send back to the user's browser (or sends raw data which it then transforms into HTML). For example, on Twitter, when you search for a particular user account, Twitter's server pulls relevant account names from its database and sends them to your computer, where your browser renders this incoming data as the text and imagery you see on screen. In this sense, most web pages are not documents but front ends of databases.

So much of the experience of browsing the web is scrolling through sorted lists: lists of people, places, objects, texts, and so on. Indeed, considered in the simplest terms, a database is a structured, sortable, filterable list of things. Taken all together, these

databases, or lists, constitute the web. For example, we might think of LinkedIn as a database of labor data, GoFundMe as a database of health data, Facebook as a database of social relations, AP News as a database of current events, and so on. As readers and users of websites, we have access to small parts of these underlying databases—that is, to some portion of the lists that constitute them. But we do not have the ability to directly query these databases, to sort and filter their lists in any way we see fit. We certainly do not have the ability either to add or to delete information directly to or from a database. Our access is mediated, controlled, and limited. This situation is ironic because, in many cases, the databases that make up the web are populated with data about us, its users and readers. Such data might have been provided intentionally (e.g., product reviews, social media posts, or job histories that we voluntarily post) or might be data collected about us without our consent or knowledge (e.g., purchases, physical locations, clicks, or hovers that are recorded surreptitiously).

The Web Is a Bureaucracy of Databases

The project of Silicon Valley is not merely to extract wealth from every conceivable social and economic transaction. It is also to envision new forms of social interaction and new social relations that underwrite and legitimate new kinds of wealth extraction. Venmo, for example, reimagines friendship as a series of financial transactions, replacing convivial interactions like “I've got this round, you can get the next one” with precision accounting suitable for data storage. Instagram, Uber, Airbnb, TaskRabbit, Venmo, GoFundMe, and many other technologically mediated organizations invent unique levers that revise social norms to produce opportunities for exploitation. In each case, a new technological affordance alters behaviors and norms to create a new opening for a business to fill.

These companies increasingly take on statelike roles and provide statelike services. This is particularly true in the United States, where private companies subject to market imperatives structure many aspects of social, political, and economic life (e.g., community-building, job placement, housing, and health care). Access to these life-structuring services is mediated via web interfaces (and mobile interfaces built atop web protocols), which in turn mediate access to underlying databases.

As James C. Scott observed in *Seeing Like a State* (1998), both states and statelike entities engage in bureaucratic processes that simplify reality in order to shape and apprehend

it. These abstractions and simplifications, dreamt up in Silicon Valley pitch decks, are encoded directly into the structure of databases. In this sense, it's possible to imagine the front end of any website powered by a database, particularly a massive, statelike web company, as a kind of programmatic, code-enforced bureaucracy. A team of coders, user experience designers, and businesspeople work together to invent complex rules that regulate "read and write access" to a database, the manifestation of which is called an *interface*. Interface components such as "like" buttons, comment boxes, star raters, user profiles, and autocomplete dropdowns are the means of providing or restricting access; they are the face of the bureaucracy. The user typically navigates this bureaucracy by filling these form fields (literally, the "form" tag in HTML), and in doing so, the website creators' "bureaucratic fantasies [take] on a social reality of their own" (Scott 1998: 68). Posting to social media, from this vantage, is nothing more than filling in endless paperwork.

The fundamental power dynamic of the contemporary web revolves around who has the most direct and unmediated access to its data stores. Thus, while the web's databases are populated by data about us—and data that in many instances would be useful for us—the highly monopolistic tech industry has deliberately closed off our access to these resources. The results are stark asymmetries of power: technological, informational, economic, and ultimately political.

It is within this context that "scrapism" becomes a liberatory countermeasure. By programmatically collecting material from the front ends of websites, it is possible to recreate the databases from which the web is generated, or to "redatabase" the web. This practice enables two modes of understanding. First, the scraper may gain the type of synoptic perspective (in all its strengths and limitations) typically available only to those with privileged access to the database, that is, access not governed by the bureaucracy of the interface. Second, and perhaps more interestingly, the scraper may gain an understanding of how their target perceives the world. They may, for example, begin to understand what it means to "see like LinkedIn," or Facebook, or Uber.

Scraping Reopens the Web

Every website is different to scrape. Some are scraped easily; others employ technical obstacles that seem nearly impossible to bypass. Different sites provide different levels of access to their databases and data commodities. Roughly speaking, data can be

willingly given (an open API), reluctantly given (a walled garden), willingly sold (a paid subscription), or selectively sold (a service that sells some of its data some of the time). These variations correspond to a given website's business model, the owners' valuation of their site's data, and the lengths to which the owners will go to protect what they have enclosed. Any attempts to redatabase a site will reveal something about those who made it and how they perceive (and protect) their own power, even if the attempt does not succeed.

Web scraping is always an act of reverse engineering. Before any scraping can occur, the web scraper must come to an understanding of how a given site is built on a technical or infrastructural level. This is true no matter who is doing the scraping and why. Technical reverse engineering is necessary whether the scraper is acting on behalf of large for-profit companies like Google or smaller entities or groups motivated by political or research agendas. As a process of coming to understand a website and the data it contains, web scraping is a window into the political, social, and economic operations of digital infrastructures, both online and offline.

The extent to which web scraping can reveal broad power dynamics was made particularly clear in the case of Aaron Swartz, who, in 2010, attempted to scrape the entirety of JSTOR's database of academic articles. In his "Guerilla Open Access Manifesto," Swartz (2008) wrote, "we need to take information, wherever it is stored, make our copies and share them with the world." To do so, Swartz wrote a scraper, ran it on a laptop hidden inside an MIT computer closet, and successfully downloaded approximately 80 percent of JSTOR's archive (JSTOR 2013). At the time, JSTOR granted login-free access to its servers to some IP addresses, enabling Swartz's scraper to run pseudo-anonymously on MIT's guest network.

JSTOR's (2013) summary of the case, published on a dedicated subdomain of their site, makes clear that the organization perceived Swartz's actions to be an existential threat: "We did not know at the time, nor do we know now, what Mr. Swartz was going to do with the nearly 5 million articles he downloaded; however, distribution of these articles would have undermined our relationships with participating JSTOR publishers and the sustainability of our service, including our ability to provide access and to preserve the content for future generations." Even as JSTOR positioned itself as the beneficent steward and conserver of knowledge, it admitted to being a gatekeeper. They never described Swartz as a thief, but there is no doubt that, in JSTOR's view, his success

would have “undermined” their business model. The Justice Department, on the other hand, framed the case purely in terms of theft and met Swartz's direct action with the fiercest possible reaction. “Stealing is stealing whether you use a computer command or a crowbar, and whether you take documents, data or dollars. It is equally harmful to the victim whether you sell what you have stolen or give it away” (US Department of Justice 2011). After Swartz's death by suicide, the Justice Department retroactively softened their stance, claiming that they only wanted him to serve six months, not the thirty-five years he'd been threatened with (Ortiz 2013).

As I have noted, scraping begins with the scraper understanding the technical stack that a site is built on, as well as the bureaucratic rules that govern access to the site's database. But because technical decisions and rules of access reflect and reproduce social, political, and economic realities, scraping brings those larger structures into sharp relief. Swartz's activism sought to combat the commodification of academic research and the artificial scarcity of digital files. Meanwhile, JSTOR's rules reflected larger social realities around the privatization and restriction of access to knowledge (the scraping happened on MIT's campus because JSTOR granted free access to that elite campus location). Finally, the Justice Department's extreme punitive reaction reflects the status of academic knowledge as private property and the role of law enforcement in protecting that property. It's hardly a leap to speculate that the legal response was excessively harsh *because* Swartz's motivations were political rather than financial and thus threatened not only JSTOR alone but also the questionable notion of knowledge as private property.

A Taxonomy of Scrapism

Swartz's action exemplifies the possibilities and pitfalls that arise when scraping is mobilized to undermine or challenge private property and hierarchies of power. *Scrapism* is thus a dual inversion: it inverts the typical use case for web scraping as a tool of digital capitalism, and, by doing so, it inverts the informational asymmetry upon which digital capitalism relies.

In the following paragraphs I sketch out a brief taxonomy of scrapism. I do not intend to produce a comprehensive review. Instead I offer web scraping as an artistic, political, and critical practice in conjunction with provocations for future work. Scrapism is always, at minimum, a two-step process: it involves programmatically collecting material

and then repurposing or re-presenting that material outside its original proprietary context. Here I focus less on the collection of material and more on the second step: the many ways that found material can be used and presented.

As sketched below, scrapism can be intended to create a *political intervention*, a *public archive*, or an *activated dataset*. These categories are not mutually exclusive. Scrapism also frequently involves techniques such as sorting; reducing/filtering/erasing/condensing; expanding or filling; finding and replacing; transposing/translating/encoding/decoding; and annotating or overlaying.

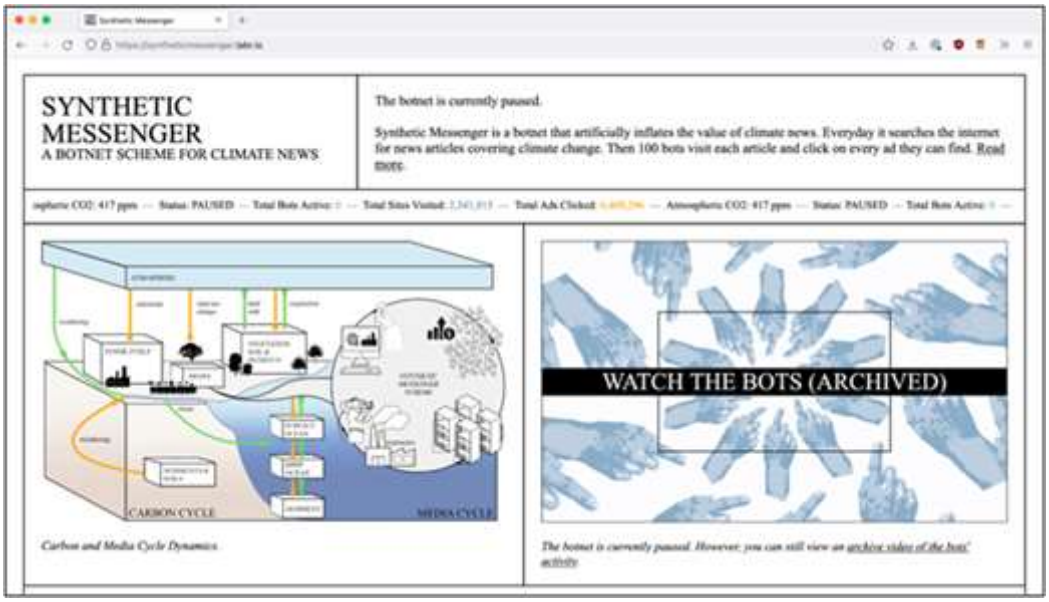
Scrapism as Political Intervention

In many situations, simply collecting and republishing material constitutes a political act. This was certainly the case for Swartz's JSTOR scrape. It is also true for file-sharing sites like Sci-Hub, Library Genesis, AAAAARG, 1337x.to, and the somehow-still-extant Pirate Bay. Sci-Hub (which mirrors scientific articles) is an explicitly political project of direct action. By contrast, torrent sites like Pirate Bay and 1337x.to, while not explicitly political, nevertheless threaten the political regime of private property. Organizations like DDoSecrets, which freely distribute leaked state and corporate material, also fall under this rubric. All of these file-sharing projects, regardless of how they frame themselves, offer collective alternatives to privatized knowledge and media.[2]

Another way that scraping operates as political intervention is by mimicking human browsing behavior—a technique that scrapers use to avoid detection. For example, in 2021 when the Kellogg Company attempted to hire full-time replacements for striking union workers through an online job portal, activist and programmer Sean Black used scraping and text-generation techniques to create an army of fake job applicants to overload the hiring site (Jankowicz 2021). Black used a similar technique to overload the state of Texas's abortion tip-line site, spamming the site with thousands of fake tips (Adams 2021).

For “Synthetic Messenger” (Lavigne and Brain n.d.), Tega Brain and I used a similar technique to create a program that clicked on millions of ads shown next to online news articles about climate change (fig. 1). Our goal was both to comment on the way that media ecology shapes physical ecology and to financially incentivize media outlets to cover climate news by engaging in what the ad industry calls “click fraud.”

Figure 1.



VIEW LARGE

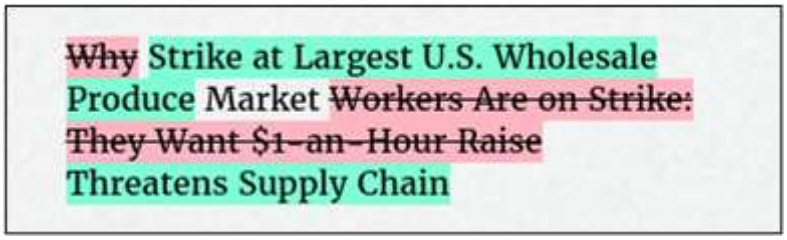
"Synthetic Messenger" website.

Scrapism as Public Archive

Scrapism can be used to create archives, or, more specifically, to re-create and make public databases that would otherwise be lost or inaccessible. For example, efforts have been made to archive the content of websites like GeoCities and Yahoo! Answers, which were at risk of being lost due to company shutdowns. Other, more general-purpose archiving services like Archive.is, the Wayback Machine, and Webrecorder allow users to duplicate and save content on individual URLs.

Web scraping can also be used to take snapshots of ephemeral data in order to produce never-ending, on-the-fly archives. For example, the @nyt_diff bot finds and highlights changes to *New York Times* headlines (fig. 2), creating a permanent record of how the newspaper shapes and reshapes narratives on issues such as labor rights and policing. Jack Sweeney's jet tracker bots cleverly use public data to automatically publish the travel itineraries of various billionaires, celebrities, and politicians (Sweeney 2022).

Figure 2.

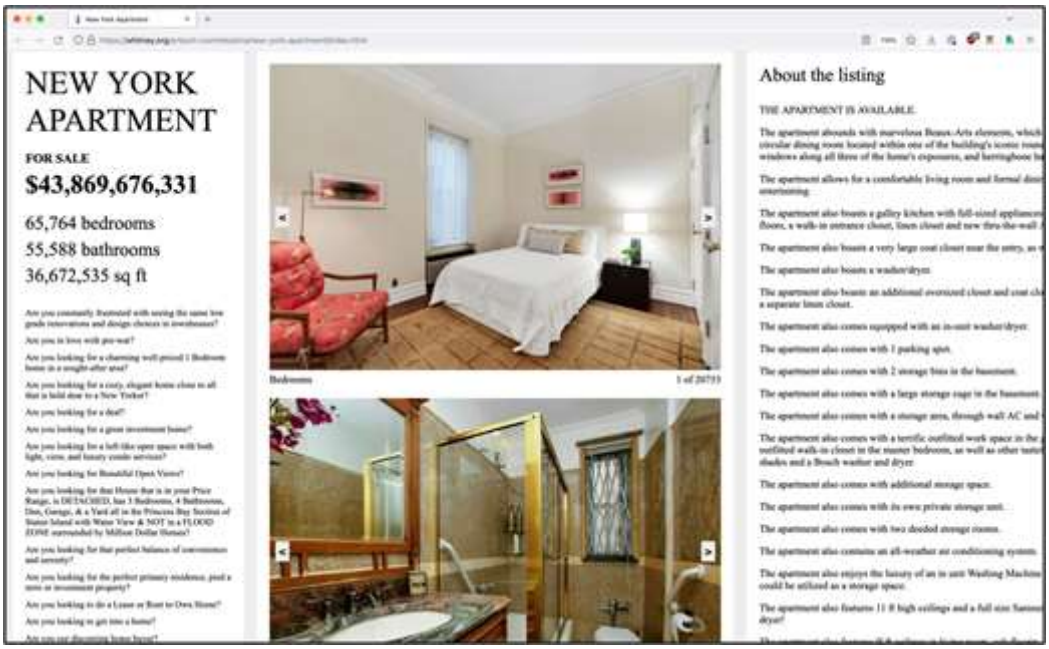


VIEW LARGE

Sample tweet from @nyt_diff bot.

As a practitioner, I have used web scraping to extract labor data from LinkedIn, real-estate data from Trulia, and health data from GoFundMe. In each case, I attempted to capture a moment in time through the lens of the specific socioeconomic enterprise with which each site is concerned. For example, in “New York Apartment” (Lavigne and Brain n.d.), Tega Brain and I collaborated in downloading the entirety of the New York real estate market from Trulia, and then we combined it in a single site to create a holistic experience of housing-as-commodity (fig. 3). Using GoFundMe, I downloaded and sorted two hundred thousand supportive comments from medical fundraiser pages to archive the collective experience of health care-as-commodity. And, by accessing LinkedIn data, I compiled a list of Immigration and Customs Enforcement (ICE) agents to produce an archive of those who are implicated in the US government's ongoing program of human rights violations. These projects hint at the extent to which web scraping—as a technique for extracting, exploring, sorting, filtering, and representing material from the web—can help to assemble pictures of the world that might not be available otherwise. In this way, scrapism as archive finds new ways to see what's hiding in plain sight.

Figure 3.



VIEW LARGE

"New York Apartment" website.

Although I chiefly associate scrapism with online computational actions, the *Brasil: nunca mais* report (Catholic Church Archdiocese of São Paulo, Brazil 1985) provides an example of analog "scraping" interventions and counterpractices of archiving. Documenting 1,800 incidents of torture during 1964–1979, it was created under the military dictatorship in Brazil and was made available to lawyers working for torture victims under very circumscribed conditions (they could check out individual reports for twenty-four-hour periods). Over the course of three years, lawyers managed to photocopy every page in the report, fully replicating the archive before it could be destroyed (Weschler 1990). Through their limited and highly bureaucratized "front-end" access, they were able to redatabase the archive and hold those in power accountable.

Scrapism as Activated Archive

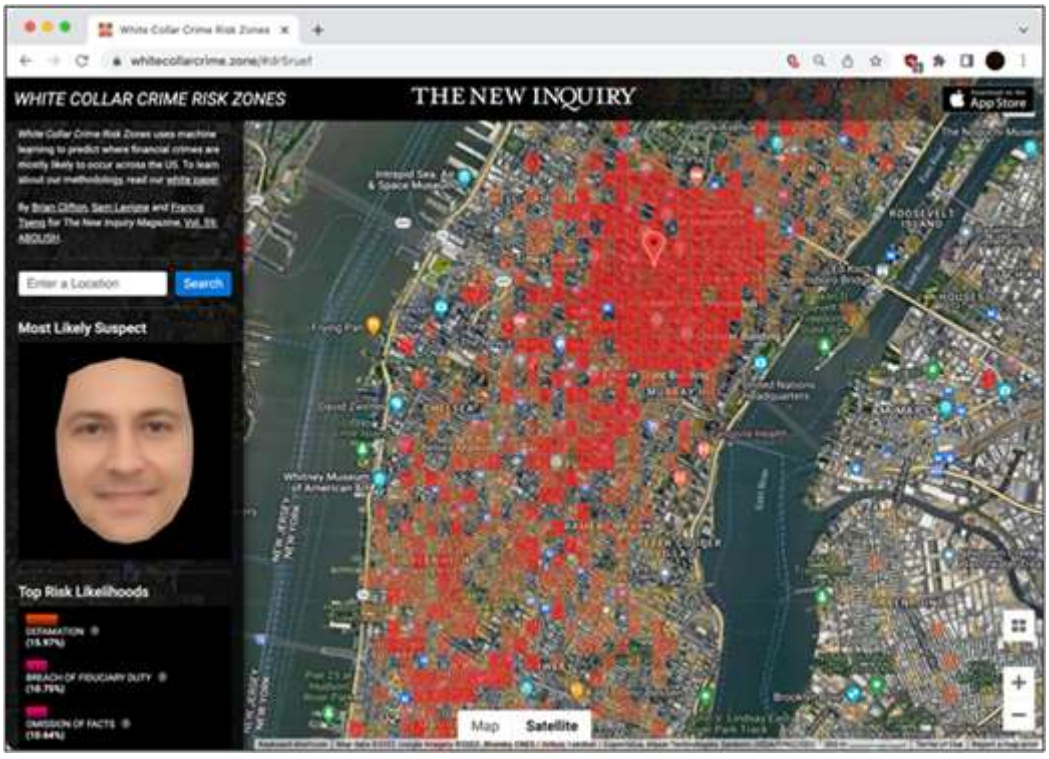
As discussed earlier, web scraping plays a key role in the production of contemporary ML systems, as it is frequently used to collect the massive datasets necessary for "training" models. ML systems are fundamentally expressions of the source material they are trained on and therefore could be thought of as "activated" archives, or datasets that directly operate on the world. While all the databases we interact with via the web shape social reality, those that are fed into ML programs do so in an especially direct way—

often shaping the decisions and practices of other systems, as when trained models are used for decision-making in medicine, benefits, criminal justice, or law enforcement.

Although such systems are typically proprietary, the same techniques used to automate inequality and encode bias and oppression can be deployed for diverse political projects.[3] Just as the police can use so-called AI to "predict" crime, so communities can use the same technologies to predict police activity,[4] locate detention centers (Killing et al. 2020), or use facial recognition to identify ICE agents (McDonald 2018).

In some cases, "activating" an archive for political purposes also necessitates its creation. For example, in "White Collar Crime Risk Zones" (Lavigne, Clifton, and Tseng n.d.), my collaborators and I produced a fully functional predictive policing tool that targets financial crime in the United States (fig. 4). The project used the same techniques deployed in predictive policing systems (in this case, a decision tree), but instead of feeding the system historic data about "street crime," we used incidents of financial malfeasance scraped from the Financial Industry Regulatory Authority, dating back to 1969. This "activated archive" produced a living, experiential critique of how data is operationalized, implicitly questioning the use of predictive policing primarily to criminalize the poorest populations.

Figure 4.



VIEW LARGE

“White Collar Crime Risk Zones” website.

Conclusion

Scrapism is a provocation to envision and enact online technologies and practices that resist the highly privatized and exploitative present-day web. Web scraping is an essential technique for amassing data and power. In the hands of corporations seeking profits and market share, it is a tool for commodifying open or freely shared data and piping that data into paid services built through proprietary tools, archives, and databases. However, it can be an equally powerful technique for counteracting these very forces. Web platforms can be broken apart, intervened in, repurposed, and reactivated. Scrapism can hold powerful entities accountable, expose their opaque operations, or illuminate their self-interested logics and motives.

Every time one of us browses the web, we leave traces in our wake. When these traces become data for ML models, they typically enact narrow commercial imperatives, strengthen surveillance, and amplify the means of repression. While this may seem like a dystopian hellscape, the fact remains that power also leaves its own digital traces. Scrapism is an invitation to engage with these traces. It attempts to reopen these opaque systems, share proprietary knowledge, leverage data against the logic of extraction, and open the structures of digital capitalism to the public's surveillance. As Berners-Lee says, “perhaps a linked information system will allow us to see the real structure of the organization in which we work.”

Notes

1. While a distinction can be drawn between a “web crawler,” which maps and indexes sites and is typically used to build search engines, and a “web scraper,” which may be built for any number of outcomes, both operate under the same fundamental principles and use similar techniques.
2. It's interesting to note that there is a real interplay between “classic” file sharing sites like Napster and contemporary streaming sites like Spotify. In a way, businesses learned how to make streaming subscription services from the pirate sites. The opening up that happens when an illegal file sharing site becomes popular can lead to more robust enclosures.
3.

For numerous examples of how automated systems are deployed for repressive ends, see, for example, O'Neil (2017); Eubanks (2018); Noble (2018); Benjamin (2019).

4. To the best of my knowledge, no one has done this yet, but it is possible if relevant data can be collected.

Works Cited

Adams, Biba. 2021. "TikToker Sean Black Shares Hack to Flood Texas Abortion 'Whistleblower' Site with False Tips." *Yahoo! News*, September 3. https://news.yahoo.com/tiktoker-sean-black-shares-hack-155444192.html.

Google Scholar

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

Google Scholar

Berners-Lee, Tim, and Mark Fischetti. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. New York: HarperBusiness.

Google Scholar

Brown, Tom B., et al 2020. "Language Models Are Few-Shot Learners." *arXiv*. https://doi.org/10.48550/arXiv.2005.14165.

Google Scholar

Catholic Church Archdiocese of São Paulo, Brazil. 1985. *Brasil: nunca mais*. Valparaiso: Vozes.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. Miami: IEEE. https://doi.org/10.1109/CVPR.2009.5206848.

Google Scholar

Dockray, Sean. n.d. "AAAAARG." https://aaaaarg.fail/.

Editing TheGrayLady (@nyt_diff). 2021. "Change in Headline." Twitter, January 22. https://twitter.com/nyt_diff/status/1352616038781874176.

Google Scholar

Elbakyan, Alexandra. n.d. “Sci-Hub.” https://sci-hub.se/ (accessed August 2, 2023).

Google Scholar

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's.

Google Scholar

Gray, Matthew. 1993. “Re: Searchable Index of the Web.” WWW *Talk*. https://web.archive.org/web/20030512083018/http://ksi.cpsc.ucalgary.ca/archives/WWW-TALK/www-talk-1993q2.messages/706.html.

Google Scholar

Hill, Kashmir. 2020. “The Secretive Company That Might End Privacy as We Know It.” *New York Times*, January 18. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

Google Scholar

Jankowicz, Mia. 2021. “A TikToker Said He Wrote Code to Flood Kellogg with Bogus Job Applications after the Company Announced It Would Permanently Replace Striking Workers.” *Business Insider*, December 10. https://www.businessinsider.com/tiktoker-wrote-code-spam-kellogg-strike-busting-job-ad-site-2021-12.

Google Scholar

JSTOR. 2013. “Summary of Events | JSTOR Evidence in United States vs. Aaron Swartz.” July 30. https://docs.jstor.org/summary.html.

Killing, Alison, Christo Buschek, and Megha Rajagopalan. 2020. “Blanked-Out Spots on China's Maps Helped Us Uncover Xinjiang's Camps.” *BuzzFeed News*, August 27. https://www.buzzfeednews.com/article/alison_killing/satellite-images-investigation-xinjiang-detention-camps.

Google Scholar

Lavigne, Sam, and Tega Brain. n.d. New York Apartment. https://whitney.org/artport-commissions/new-york-apartment/index.html (accessed June 21, 2023).

Lavigne, Sam, and Tega Brain. n.d. “Synthetic Messenger.” https://syntheticmessenger.labr.io/ (accessed June 21, 2023).

Google Scholar

Lavigne, Sam, Brian Clifton, and Francis Tseng. n.d. White Collar Crime Risk Zones. https://whitecollarcrime.zone/ (accessed June 21, 2023).

"Law Enforcement." n.d. Clearview AI. https://www.clearview.ai/law-enforcement (accessed June 21, 2023).

Library Genesis. n.d. https://libgen.is/ (accessed June 21, 2023).

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. "Microsoft COCO: Common Objects in Context." *arXiv*. https://doi.org/10.48550/arXiv.1405.0312.

Google Scholar

Lindsay, Kathryn. 2022. "'Amazon for Real Estate': How the StreetEasy App Took Over New York." *Guardian*, February 9. https://www.theguardian.com/lifeandstyle/2022/feb/09/streeteasy-new-york-housing-rental.

Google Scholar

Losse, Katherine. 2012. *The Boy Kings: A Journey into the Heart of the Social Network*. New York: Free Press.

Google Scholar

McDonald, Kyle. n.d. ICESPY. https://icespy.org/ (accessed June 21, 2023).

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Google Scholar

O'Brien, Luke. 2020. "Far-Right Extremists Helped Create the World's Most Powerful Facial Recognition Technology." *HuffPost*, April 7. https://www.huffpost.com/entry/clearview-ai-facial-recognition-alt-right_n_5e7d028bc5b6cb08a92a5c48.

Google Scholar

O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books, 2017.

Google Scholar

Ortiz, Carmen. 2013. “Statement of United States Attorney Carmen M. Ortiz Regarding the Death of Aaron Swartz.” January 16. https://www.justice.gov/usao-ma/pr/statement-united-states-attorney-carmen-m-ortiz-regarding-death-aaron-swartz.

Google Scholar

The Pirate Bay. n.d. https://thepiratebay.org/index.html (accessed June 21, 2023).

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* New Haven, CT: Yale University Press.

Google Scholar

Swartz, Aaron. 2008. “Guerilla Open Access Manifesto.” http://archive.org/details/GuerillaOpenAccessManifesto (uploaded November 21, 2011).

Google Scholar

Sweeney, Jack (@JxckSweeney/Plane-Notify). 2022. https://web.archive.org/web/20220209122116/https://twitter.com/i/lists/1307414615316467715 (accessed August 10).

US Department of Justice. 2011. “Alleged Hacker Charged with Stealing over Four Million Documents from MIT Network.” July 19. https://www.justice.gov/archive/usao/ma/news/2011/July/SwartzAaronPR.html.

Weschler, Lawrence. 1990. *A Miracle, a Universe: Settling Accounts with Torturers.* New York: Pantheon.

Google Scholar

RELATED ARTICLES

The Reverse of Engineering

“Put the Groceries Up”: Comparing Black and White Regional Variation

Professional Judgment in an Era of Artificial Intelligence and Machine Learning

The Photographic Pipeline of Machine Vision; or, Machine Vision's Latent Photographic Theory

RELATED BOOK CHAPTERS

Machine Learning and Genomic Dimensionality From Features to Landscapes

Witnessing Algorithms

Valuing Data in Postgenomic Biology How Data Donation and Curation Practices Challenge the Scientific Publication System

Investable Life

Skip to Main Content

RELATED TOPICS

web scraping

data

machine learning

internet art

WE RECOMMEND

Mediated Massacre: Digital Nationalism and History Discourse on China's Web
Florian Schneider, Journal of Asian Studies, 2018

Sources for the Study of Brazilian Economic and Social History on the Internet
Herbert Klein et al., HAHR, 2004

Body, Sex, InterfaceReckoning with Images at the Lesbian Herstory Archives
McKinney et al., Radical History Review, 2015

Mapping the TerritoryArchiving the Trans Website in an Age of Search
Dame et al., TSQ: Transgender Studies Quarterly, 2016

Insights into Presbycusis From the First Temporal Bone Laboratory Within the United States
Nicholas S. Andresen et al., Otology & Neurotology, 2022

Lung ultrasonography: a practical guide for cardiologists
Francesco Bianco et al., Journal of Cardiovascular Medicine, 2017

Using High-Impact HIV Prevention to Achieve the National HIV/AIDS Strategic Goals in Miami-Dade County, Florida: A Case Study
James W. Carey et al., Journal of Public Health Management and Practice, 2015

Contractual considerations in the private equity era

Skip to Main Content

Online ISSN 2834-703X

Duke University Press
905 W. Main St. Ste. 18-B
Durham, NC 27701
USA

Phone
(888) 651-0122

International
+1 (919) 688-5134

Contact
Contact Us