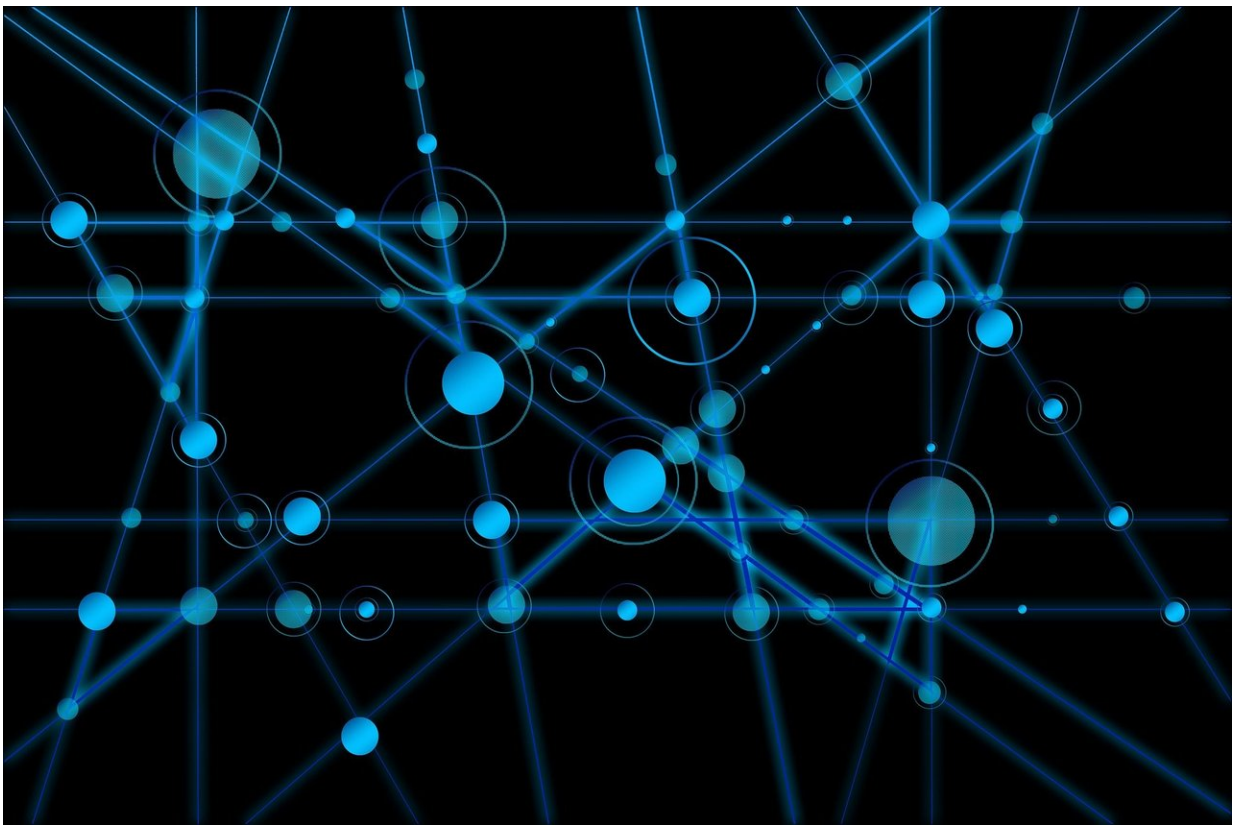
A new framework for web scraping data to ensure its validity for use in marketing studies

June 2 2022, by Marilyn Stone



Credit: CC0 Public Domain

Researchers from Erasmus University Rotterdam, Tilburg University, INSEAD, and Oxford University published a new paper in the *Journal*

of Marketing that proposes a methodological framework focused on enhancing the validity of web data.

The study is authored by Johannes Boegershausen, Hannes Datta, Abhishek Borah, and Andrew T. Stephen.

The recent ruling of the Ninth Circuit in HiQ Labs v. LinkedIn underscores the importance of navigating the legal challenges when using web scraping to collect data for academic research. While it may be permissible to collect information from publicly available sites, researchers still need to be cautious about how they design their extraction software. For example, collecting information from publicly available user profiles in some jurisdictions may trigger privacy concerns—and prompts researchers to anonymize their data during the collection.

While marketing researchers increasingly employ web data, the idiosyncratic and sometimes insidious challenges in its collection have received limited attention. How can researchers ensure that the datasets generated via web scraping and APIs are valid? This research team developed a novel framework that highlights how addressing validity concerns requires the joint consideration of idiosyncratic technical and legal/ethical questions.

The authors say that their "framework covers the broad spectrum of validity concerns that arise along the three stages of the automatic collection of web data for academic use: selecting data sources, designing the data collection, and extracting the data. In discussing the methodological framework, we offer a stylized marketing example for illustration. We also provide recommendations for addressing challenges researchers encounter during the collection of web data via web scraping and APIs."

The article further provides a systematic review of more than 300

articles using web data published in the top five marketing journals. Using this review, the researchers explain how web data has advanced marketing thought. Understanding the richness and versatility of web data is invaluable for scholars curious about integrating it into their research programs.

Interested researchers can access the database developed for this review on the companion website. This website also features additional useful resources and tutorials for collecting web data via web scraping and APIs.

The researchers add that they use their "methodological framework and typology to unearth new and underexploited 'fields of gold' associated with web data. We seek to demystify the use of web scraping and APIs and thereby facilitate broader adoption of web data across the marketing discipline. Our Future Research section highlights novel and creative avenues of using web data that include exploring underutilized sources, creating rich multi-source datasets, and fully exploiting the potential of APIs beyond data extraction."

More information: Johannes Boegershausen et al, EXPRESS: Fields of Gold: Scraping Web Data for Marketing Insights, *Journal of Marketing* (2022). DOI: 10.1177/00222429221100750

Web database: web-scraping.org/

Provided by American Marketing Association