

Data Scraping

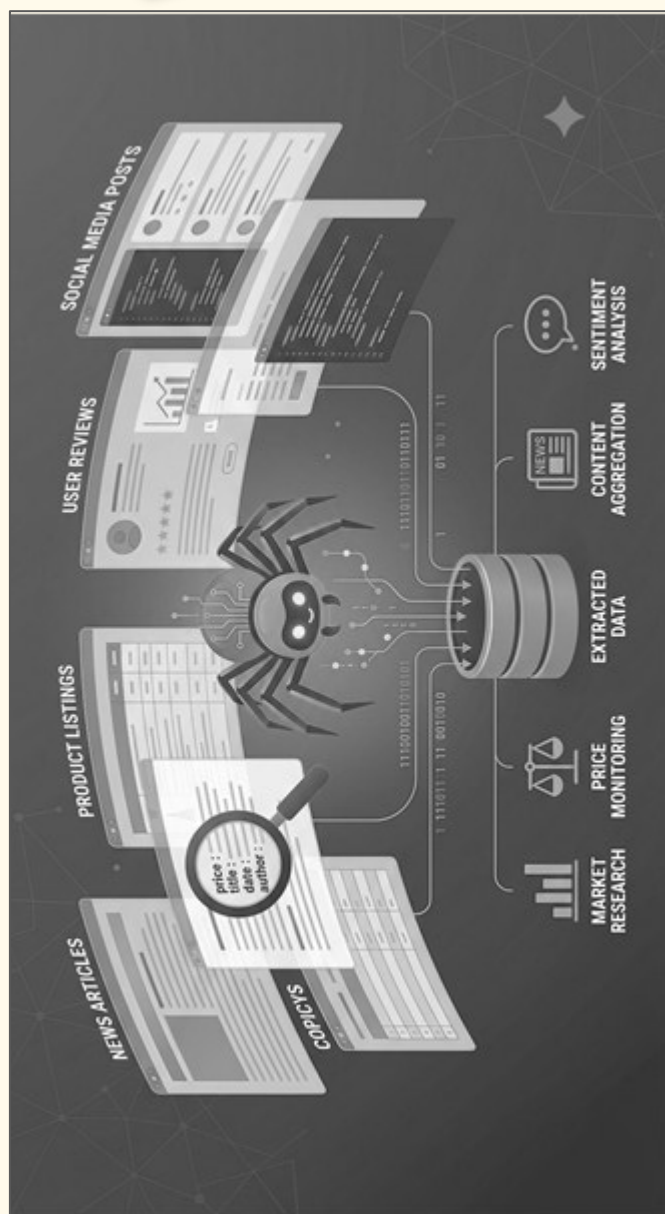
Mr. Vimal Bhaliya

Assistant Professor in BCA

Mr. Taha Travadi

Lecturer in BCA

Smt. B. V. Dhanak College, Bagasara



As an illustration, let us say that you are employed by a nearby shoe company and that your manager has asked you to locate potential Instagram promoters.

Thousands of searches could be conducted to find potential helpers. Alternatively, you may set up a scraping program to fill a study spreadsheet. Can you guess which approach is quicker?

What Is Data Scraping?

You can find all the information you need on a website. However, you frequently lack the time and energy to go over each page and take thorough notes. Data scraping comes next. You can obtain all the information you require with a single tool (without the annoying clicking and pressing).

Businesses designed their data scraping technologies with people in mind. They do not spit out formatting rules, code, or tags rather than you can easily understand and work with the results.

There are three main types of data scraping:

1. **Report mining:** Programs pull data from websites into user-generated reports. It's a bit like printing a page, but the printer is the user's report.
2. **Screen scraping:** The tool pulls information on legacy machines into modern versions.
3. **Web scraping:** Tools pull data from websites into reports users can customize.

You might use data scraping for:

- **Website upgrades:** A screen scraper can be a crucial tool if you're working with a very old computer that can't work with a new system. Rather than trying to recode or update the old piece, you can just pull from it and start a new with current technology.
- **Competitor analysis:** A company you'd like to beat publishes all colors, sizes,

- Data scraping includes collecting information out of a website and into a spreadsheet. The technique is an effective way for a dedicated data scraper to collect a lot of data for processing, analysis, or presentation.

and prices of a product online. Data scraping could tell you how much your product should cost and how many people want to buy it. Experts consider this form of analysis one of the best ways to use data scraping opens in a new tab.

- **Data aggregation:** Have you ever visited a website filled with headlines from newspapers all around the world? Or have you ever hit a page that has prices and products from several different companies, all in one place? Data scraping makes this possible opens in a new tab.
- **In-depth reporting:** In 2018, reporters at BuzzFeed created several charts comparing every State of the Union Address opens in a new tab ever given in the United States. That analysis relied on data from the Presidency Project at the University of California, Santa Barbara. Without data scraping, reporters would have to type in all the addresses by hand, which would add time to the project.

- Some people use the technique to harm others. For example, some people set up scraping tools to gather email addresses or social media profiles. Then they bundle up that data and sell it to email spammers.
- Bad actors can also use scraping tools to steal data. For example, Facebook sued two companies in 2020 for installing extensions that scraped names, birthdays, and other sensitive data. Users had no idea this was happening, but the companies sold their data to third parties.
- People sometimes confuse data scraping with web crawling, but the two techniques are very different. A web crawler looks very closely at the code within the page, and the device might even skip over pages altogether if the programmer includes the proper tag.
- These results help sites like Google understand what to put in search results pages. Data scraping tools ignore most code, and those tools pay no attention to programmer requests.

How Does Data Scraping Work?

You're set to pull data from a source you researched and vetted. How will you get started? Chances are, you'll use a tool someone has already programmed for the purpose.

Consider web scrapers. These tools typically follow a three-step process:

1. **Request:** The program uses a "GET" command to pull data from a page you chose.
2. **Parse:** The scraper looks for the specific data field you identified.
3. **Display:** The requested information flows into a report you specified or created.

4 Ways to Protect Your Data

The only way to ensure that no one steals your information is to keep it off your website. But that same step can make your customers hunt for your products and prices. To stay competitive, you must be online, but you can protect what is yours.

Try these four strategies:

1. **Limit requests:** Apply rate-limiting rules and ensure that one IP address can't ping your server too many times within a specified timeframe.
2. **Apply CAPTCHA :** If you see multiple requests coming from the same server, ask the user to validate identity by completing a simple puzzle or tapping a button. Bots can't tackle this step, so you'll eliminate any scraping tool.
3. **Use images:** Web scraping tools are built to analyze and parse words, not photographs. Embed sensitive data, including contact information and pricing, inside images rather than placing it in steal-worthy text.
4. **Shake up your text:** Simple changes (like using [at] instead of @ in an email address) can confuse data scraping tools and make stealing and using that information much more difficult.

❖ **References:**

- 1) <https://www.imperva.com/learn/application-security/data-scraping/>
- 2) <https://www.okta.com/en-in/identity-101/data-scraping/>
- 3) https://www.researchgate.net/publication/383113092_Demystified_Overview_of_Data_Scraping
- 4) <https://www.sciencedirect.com/science/article/pii/S1877050925012128>
- 5) <https://www.sciencedirect.com/science/article/pii/S1877050925012128>
- 6) <https://targetinternet.com/resources/what-is-data-scraping-and-how-can-you-use-it>