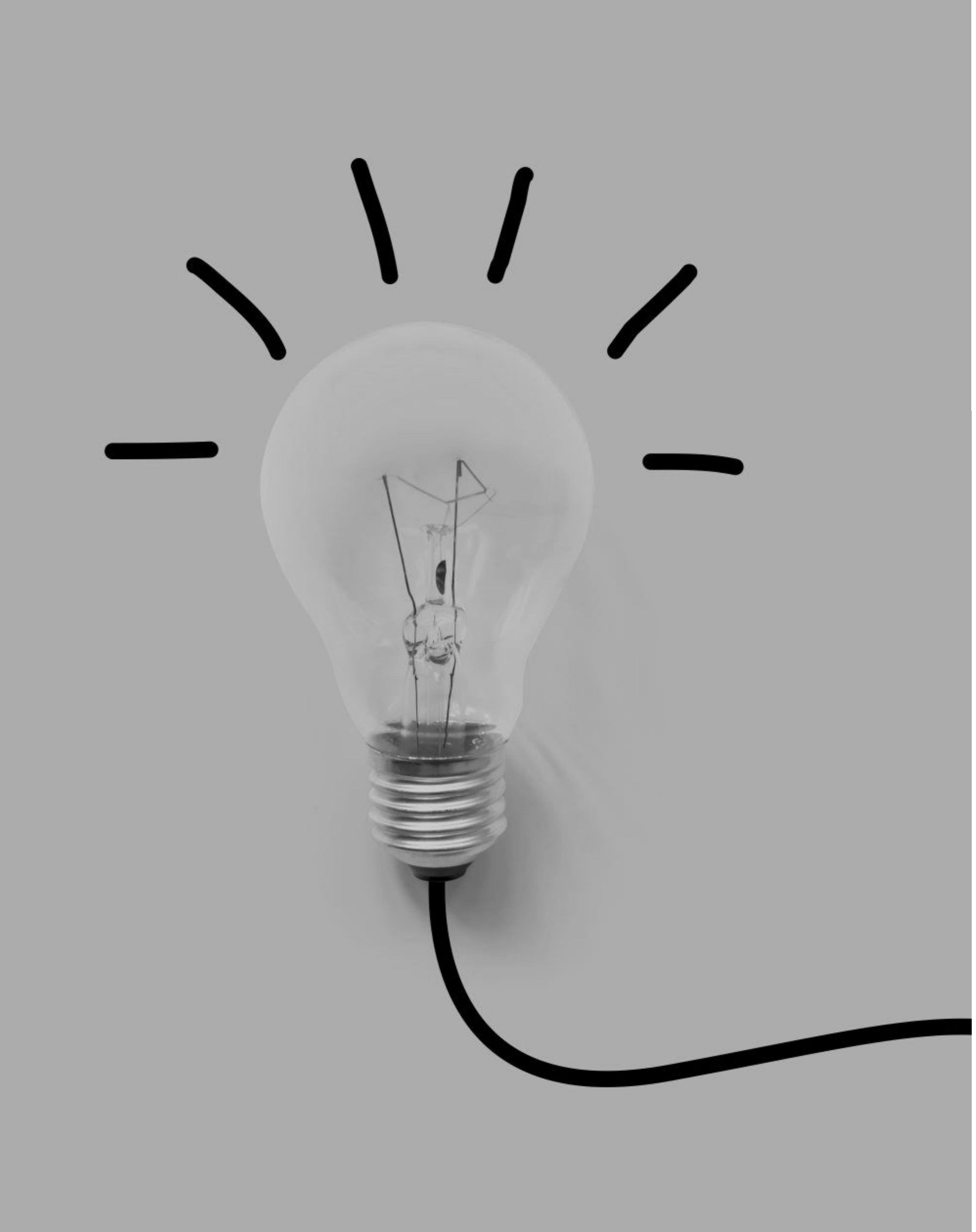
10Alytics

WEB SCRAPING

Chu Ngwoke



LEARNING OBJECTIVES

1. Data Extraction from the Web
2. Using APIs
3. Introduction to Web Scraping
4. Ethics in Web Scraping
5. Introduction to Beautiful Soup
6. Hands-on with web and API scraping

WHAT IS WEB SCRAPING

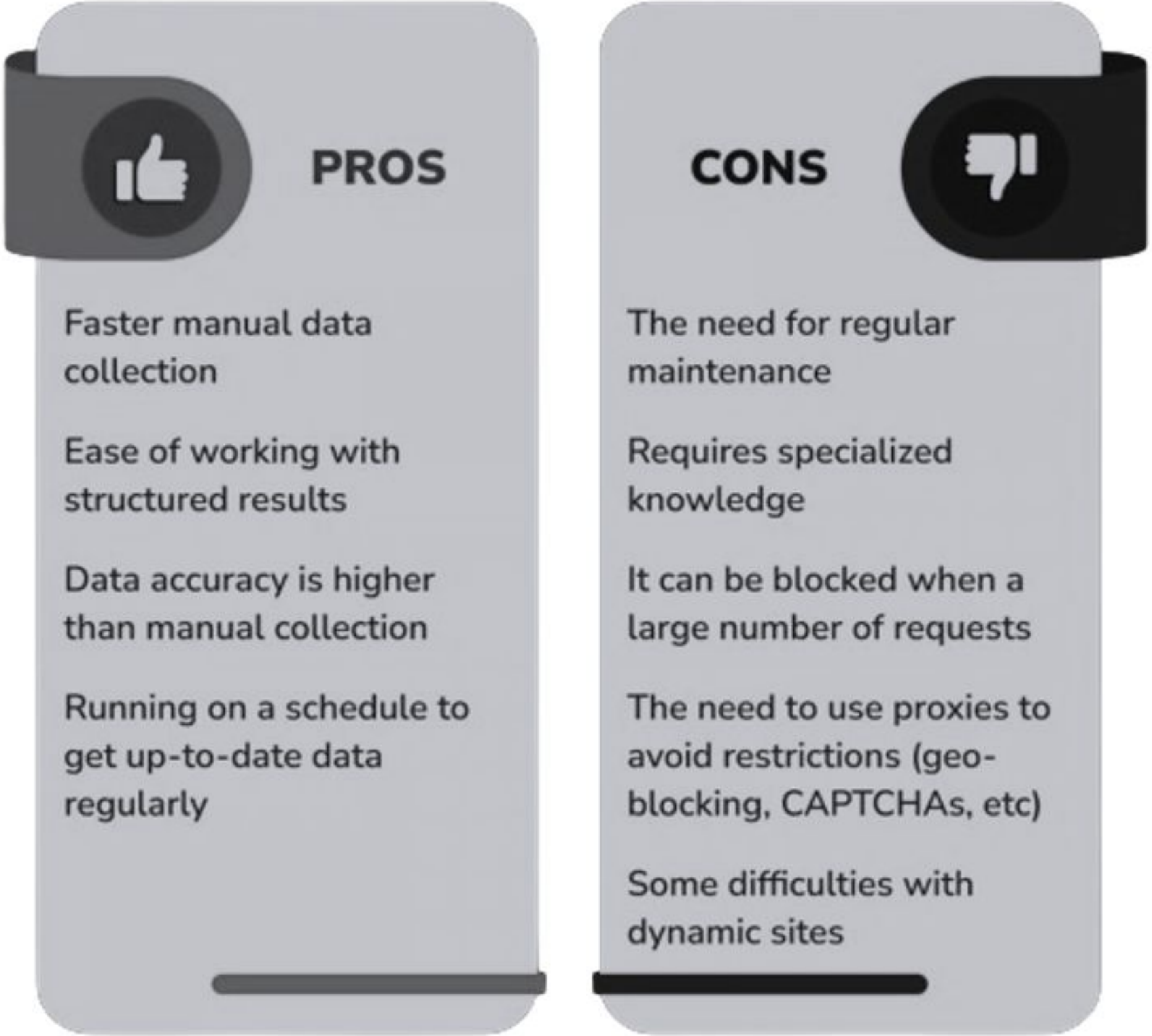
Web scraping refers to the **automated** process of extracting data from websites. It involves retrieving specific information from web pages, which could include text, images, links, or any other type of content, and then parsing that data for analysis, storage, or manipulation.

DATA EXTRACTION FROM THE WEB

Extracting data is an essential part of data engineering. But how to get big data from all over the Internet that transforms many business processes?

Common methods of retrieving data from the Internet are **APIs** and **web scraping.**

WEB SCRAPING



PROS

- Faster manual data collection
- Ease of working with structured results
- Data accuracy is higher than manual collection
- Running on a schedule to get up-to-date data regularly

CONS

- The need for regular maintenance
- Requires specialized knowledge
- It can be blocked when a large number of requests
- The need to use proxies to avoid restrictions (geo-blocking, CAPTCHAs, etc)
- Some difficulties with dynamic sites

API

API stands for **Application Programming Interface**, which acts as an intermediary, allowing websites and software to communicate and exchange data and information.

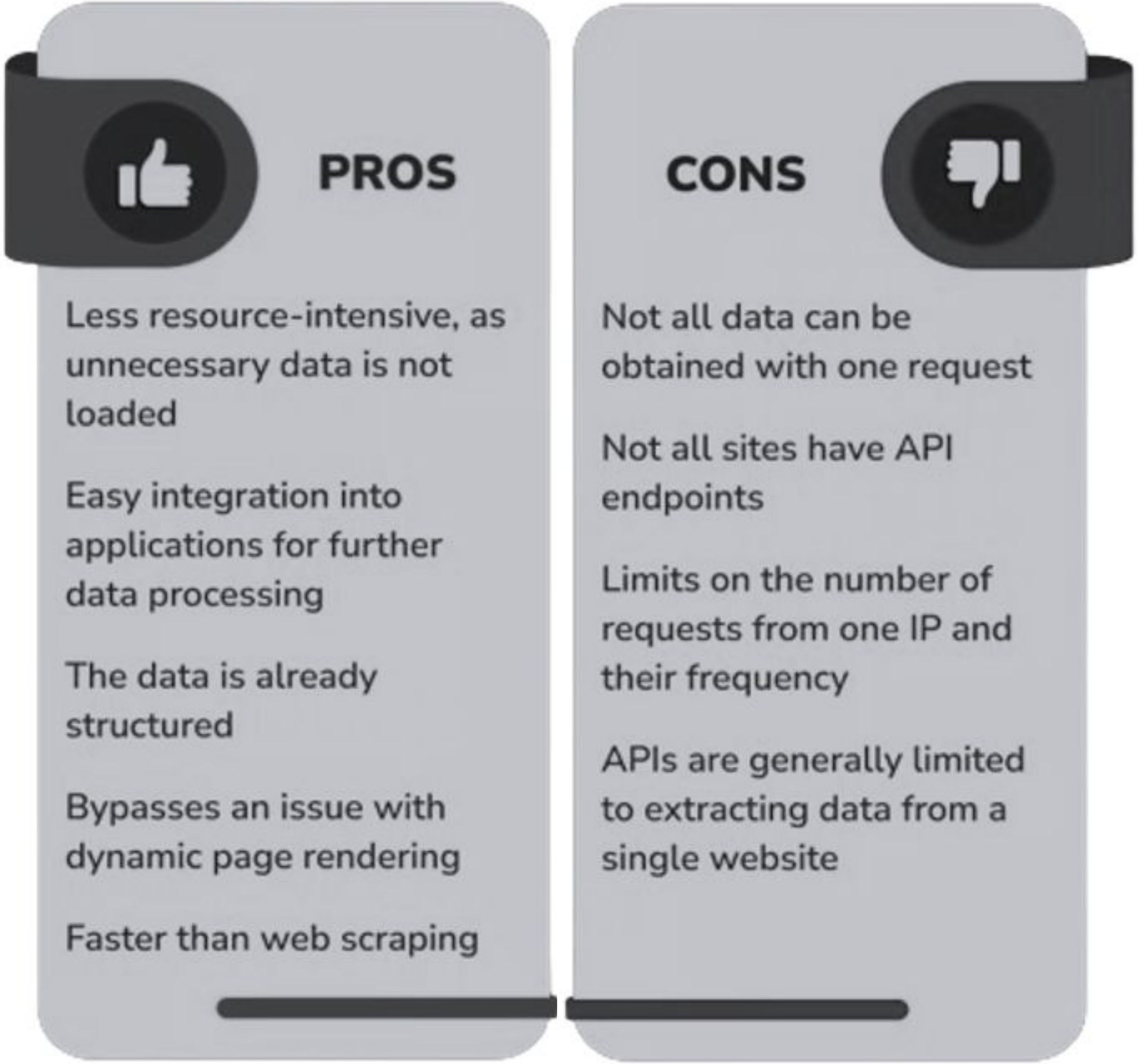
To contact the API, **you need to send it a request**. The client must provide the URL and HTTP method to process the request correctly. Then **API will process the request** and **send the response received from the web server back to the client.**

API SCRAPING

API scraping is the process of extracting data from an API that provides access to web applications, databases, and other online services. Unlike extracting from a website's visual components, this method uses simple API calls to interact with a service's backend, ensuring more structured and dependable data retrieval.

APIs provide direct access to specific data subsets via dedicated endpoints, removing the need to wade through extensive raw code or HTML structure.

API SCRAPING



PROS

- Less resource-intensive, as unnecessary data is not loaded
- Easy integration into applications for further data processing
- The data is already structured
- Bypasses an issue with dynamic page rendering
- Faster than web scraping

CONS

- Not all data can be obtained with one request
- Not all sites have API endpoints
- Limits on the number of requests from one IP and their frequency
- APIs are generally limited to extracting data from a single website

ETHICS OF WEB SCRAPING

- Intellectual Property and Copyright Laws (not regarding scraping, but usage of data)
- Terms of Service (TOS and *robots.txt file)

In general, while scraping, be respectful to the website (do not overload the website with too many requests as this can impact the website).

PYTHON LIBRARIES FOR WEB SCRAPING

- requests
- BeautifulSoup

THANK YOU

Chu Ngwoke

-

Chu.Ngwoke@gmail.com