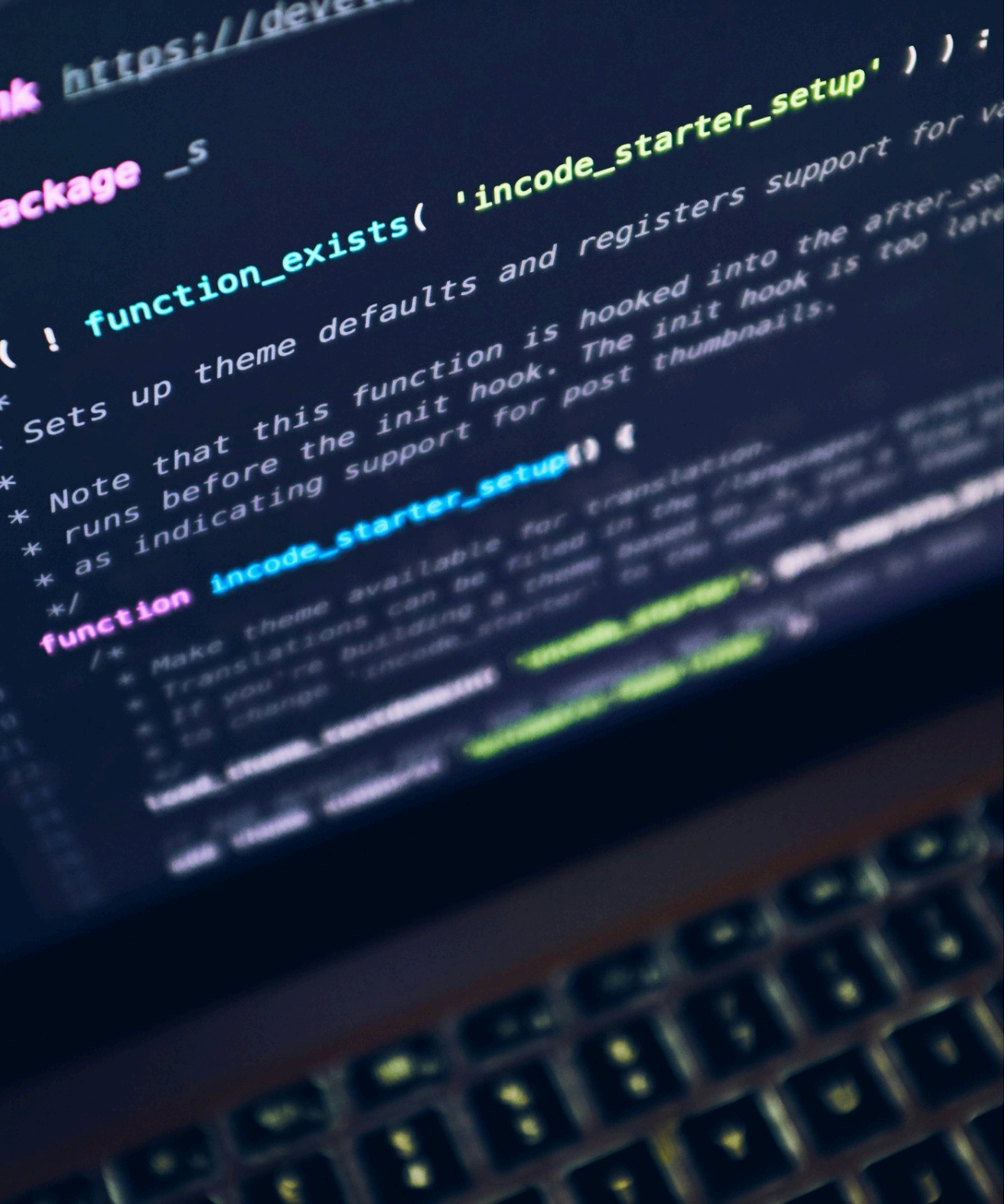


Web Scraping Fundamentals

Introducing the essentials of web scraping for investigative reporting

Web Scraping for Investigative Reporting

- **Introduce web scraping fundamentals**
  Explain the basics of web scraping, including how to extract data from websites and parse HTML/XML structures

- **Demonstrate data extraction techniques**
  Show reporters how to use Python libraries like BeautifulSoup and Scrapy to extract specific data from web pages

- **Discuss ethical considerations**
  Highlight the importance of respecting website terms of service and robots.txt protocols when web scraping

- **Showcase investigative case studies**
  Present real-world examples of how investigative reporters have used web scraping to uncover important stories

- **Provide hands-on training**
  Facilitate interactive workshops where reporters can practice building their own web scrapers and explore datasets

Introduction to Web Scraping

Web Scraping Defined

The automated process of extracting data from websites and online sources

Data Collection for Investigations

Enables journalists to gather large datasets from the web to uncover hidden stories and patterns

Empowering Investigative Reporting

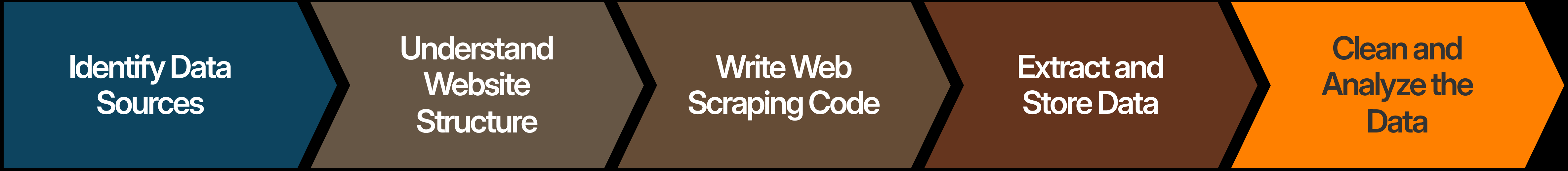
Provides access to a wealth of information that can be leveraged to expose misconduct, corruption, and other issues of public interest

Uncovering Invisible Narratives

Web scraping allows journalists to shine a light on stories and data that may be buried or difficult to access through traditional means

Web scraping is a transformative tool that can significantly enhance the capabilities of investigative journalists, providing access to a wealth of valuable data and enabling the uncovering of hidden narratives that serve the public interest.

The Web Scraping Workflow



Identify Data Sources

Determine the websites, web pages, or online databases that contain the information relevant to your investigative journalism project.

Understand Website Structure

Analyze the HTML structure and layout of the target websites to identify the specific elements, tags, or patterns that contain the data you need to extract.

Write Web Scraping Code

Develop a web scraping script or utilize a web scraping tool to automate the process of extracting data from the identified sources, such as using programming languages like Python or no-code solutions.

Extract and Store Data

Capture the relevant data from the target websites and store it in a structured format, such as CSV files, spreadsheets, or databases, for further analysis.

Clean and Analyze the Data

Perform data cleaning, normalization, and analysis to prepare the extracted information for investigative reporting, uncovering patterns, trends, and potential leads.

Ethical Considerations

Conducting web scraping for investigative journalism requires a deep understanding of the legal and ethical considerations at play. Journalists must navigate a complex landscape of data privacy laws, terms of service, and potential reputational risks. Adhering to these guidelines is essential to ensure that the information gathered is obtained ethically and can be used to inform the public without infringing on individual rights or causing unintended harm.

Tools and Techniques

- **Code-based Web Scraping Tools**

  Powerful, flexible, and customizable web scraping tools that require programming knowledge, such as Python libraries (e.g., BeautifulSoup, Selenium)

- **No-code Web Scraping Solutions**

  User-friendly, drag-and-drop web scraping platforms that enable non-technical users to extract data without writing code, such as ParseHub and Instant Data Scraper

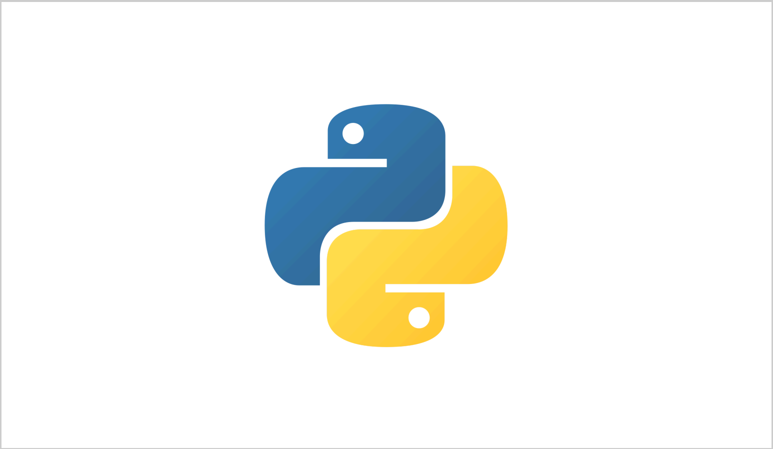
- **Browser Extensions and Add-ons**

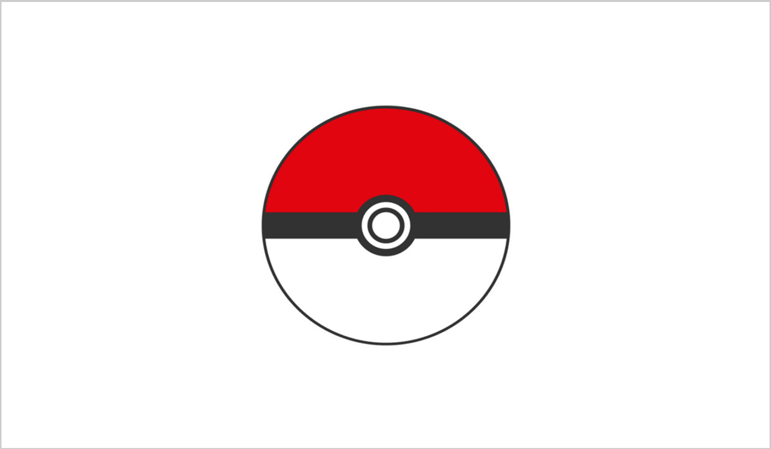
  Browser-based tools that simplify web scraping by allowing users to select and extract data directly from web pages, such as Instant Data Scraper (Chrome)

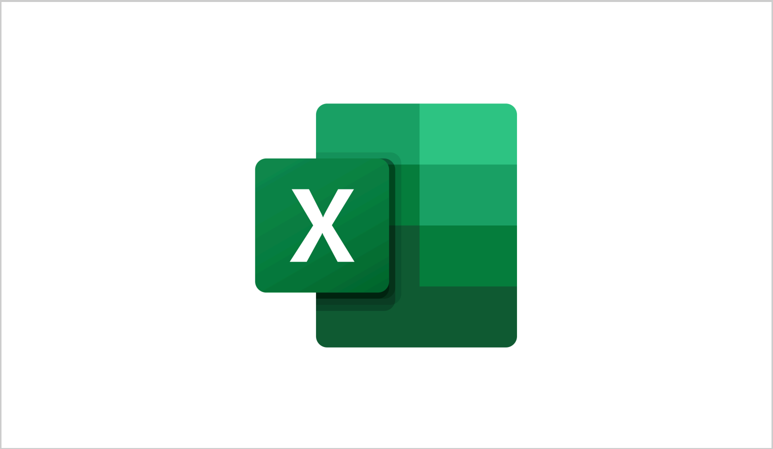
- **Data Extraction APIs**

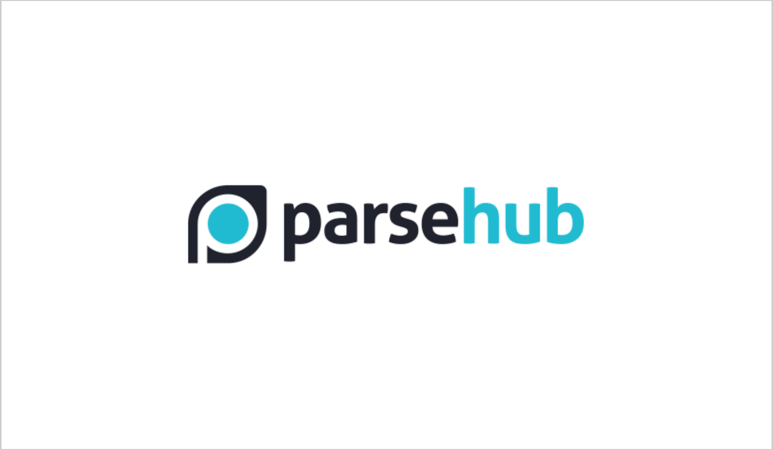
  Specialized APIs that provide structured data from various websites, allowing journalists to access information without the need for complex web scraping

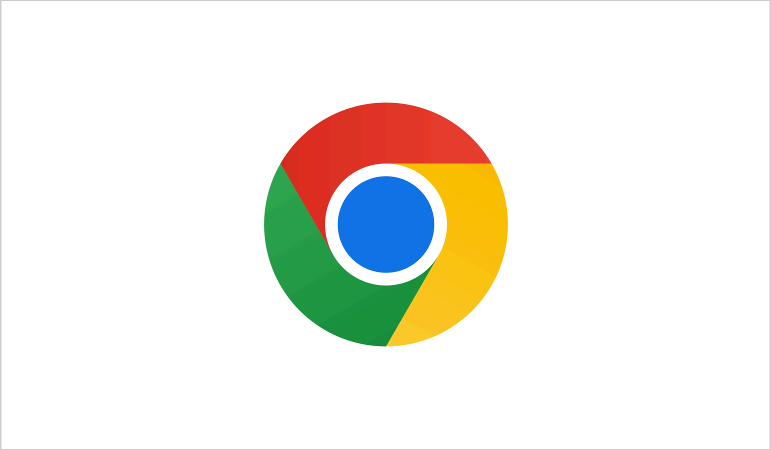
Tech Stack











ProPublica

Dollars for Docs

What they scraped:
Pharmaceutical company websites that listed payments made to doctors.

Goal:
To track how much doctors were receiving from drug companies.

Impact:
Created a searchable database that helped reveal potential conflicts of interest between doctors and the companies whose drugs they prescribed

The Marshall Project

Court Appearance Scraping

What they scraped:
Local court websites for daily case listings.

Goal:
To monitor trends in arrests, charges, and outcomes, especially around bail reform.

Impact:
Revealed inconsistencies in justice system practices and helped communities understand who was being charged and how often.

Stanford
University

Stanford - Big Local News

What they scraped:

Right-wing news websites, Twitter posts, and Facebook pages.

Goal:

To investigate the spread of false or misleading claims about the 2020 U.S. election.

Impact:

Contributed to collaborative reporting projects tracking misinformation networks.

ICIJ

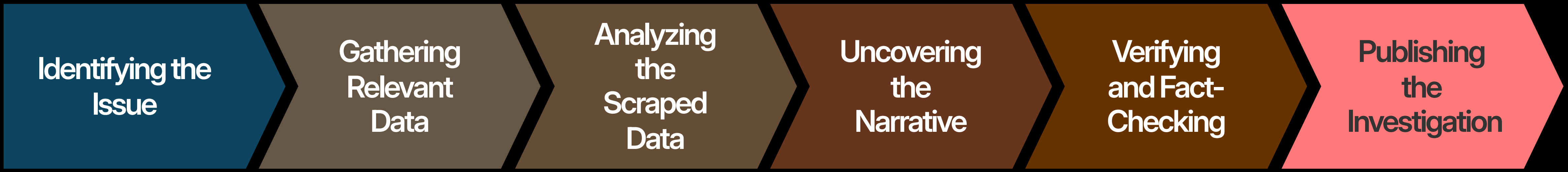
Panama Papers / Offshore Leaks Database

What they scraped:
Documents and public registries related to offshore companies.

Goal:
To cross-reference leaked documents with publicly available data.

Impact:
Helped build the Offshore Leaks Database, used globally by journalists to trace hidden money and shell companies.

Case Study: Uncovering Corporate Misconduct



Identifying the Issue

An investigative journalist discovers concerning reports about a large multinational company's questionable labor practices and environmental violations.

Gathering Relevant Data

The journalist uses web scraping techniques to collect data from the company's website, social media platforms, and regulatory filings to build a comprehensive understanding of the issue.

Analyzing the Scraped Data

The journalist analyzes the scraped data, cross-references it with other sources, and identifies patterns and inconsistencies that suggest the company is engaged in unethical practices.

Uncovering the Narrative

By connecting the dots and triangulating the data, the journalist is able to construct a compelling narrative that exposes the company's misconduct and its impact on workers, the environment, and the local community.

Verifying and Fact-Checking

The journalist rigorously verifies the findings, ensures the accuracy of the information, and obtains comments from the company and relevant authorities to provide a balanced and well-substantiated report.

Publishing the Investigation

The investigative report is published, shedding light on the company's unethical practices and leading to public outcry, regulatory scrutiny, and calls for accountability.

Web Scraping Exercise - Detainees of Boone County Jail

Exercise: Scrape and clean data using Web Data Scraper and Excel

Instant Data Scraper - Chrome Web Store
Chrome Web Storehttps://chromewebstore.google.com › instant-data-scraper

https://report.boonecountymo.org/mrcjava/servlet/RMS01_MP.I00030s

The Boone County Jail serves as an important part of the local criminal justice system, providing secure detention and rehabilitation services to the community.

Building a Web Scraping Toolkit

Python Libraries

Recommend popular Python libraries like BeautifulSoup, Scrapy, and Selenium for efficient web scraping and data extraction.

No-Code Solutions

Introduce user-friendly no-code web scraping tools like Octoparse, ParseHub, and Automation Anywhere, which require minimal coding experience.

Data Storage and Management

Suggest tools for storing and managing the scraped data, such as CSV, Excel, or database solutions like SQLite, PostgreSQL, or MongoDB.

Data Cleaning and Preprocessing

Recommend tools and techniques for cleaning, transforming, and preprocessing the scraped data, such as OpenRefine, Pandas, or custom Python scripts.

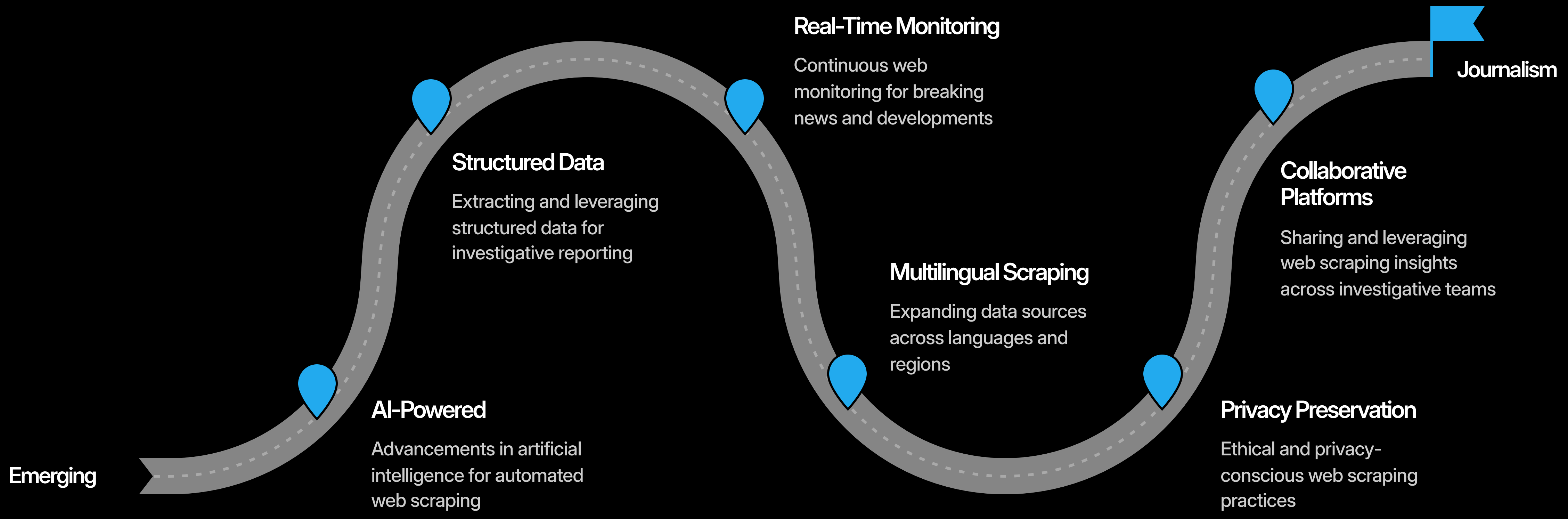
Visualization and Analysis

Introduce data visualization and analysis tools like Tableau, Power BI, or Matplotlib to help journalists gain insights from the scraped data.

Parsehub

Parsehub is a no-code web scraping and data extraction tool that can easily paginate and scrape data from multiple different website formats.

The Future of Web Scraping in Journalism



Emerging

AI-Powered

Advancements in artificial intelligence for automated web scraping

Structured Data

Extracting and leveraging structured data for investigative reporting

Real-Time Monitoring

Continuous web monitoring for breaking news and developments

Multilingual Scraping

Expanding data sources across languages and regions

Privacy Preservation

Ethical and privacy-conscious web scraping practices

Collaborative Platforms

Sharing and leveraging web scraping insights across investigative teams

Journalism

Empowering Investigative Journalists: Web Scraping Essentials

Web scraping is a powerful tool that can significantly enhance the capabilities of investigative journalists. By mastering the techniques and principles presented in this session, journalists can gain access to a wealth of valuable data, uncover hidden narratives, and drive impactful investigations that serve the public interest. As the digital landscape continues to evolve, the integration of web scraping into investigative journalism will only become more crucial, paving the way for more informed, transparent, and accountable reporting.