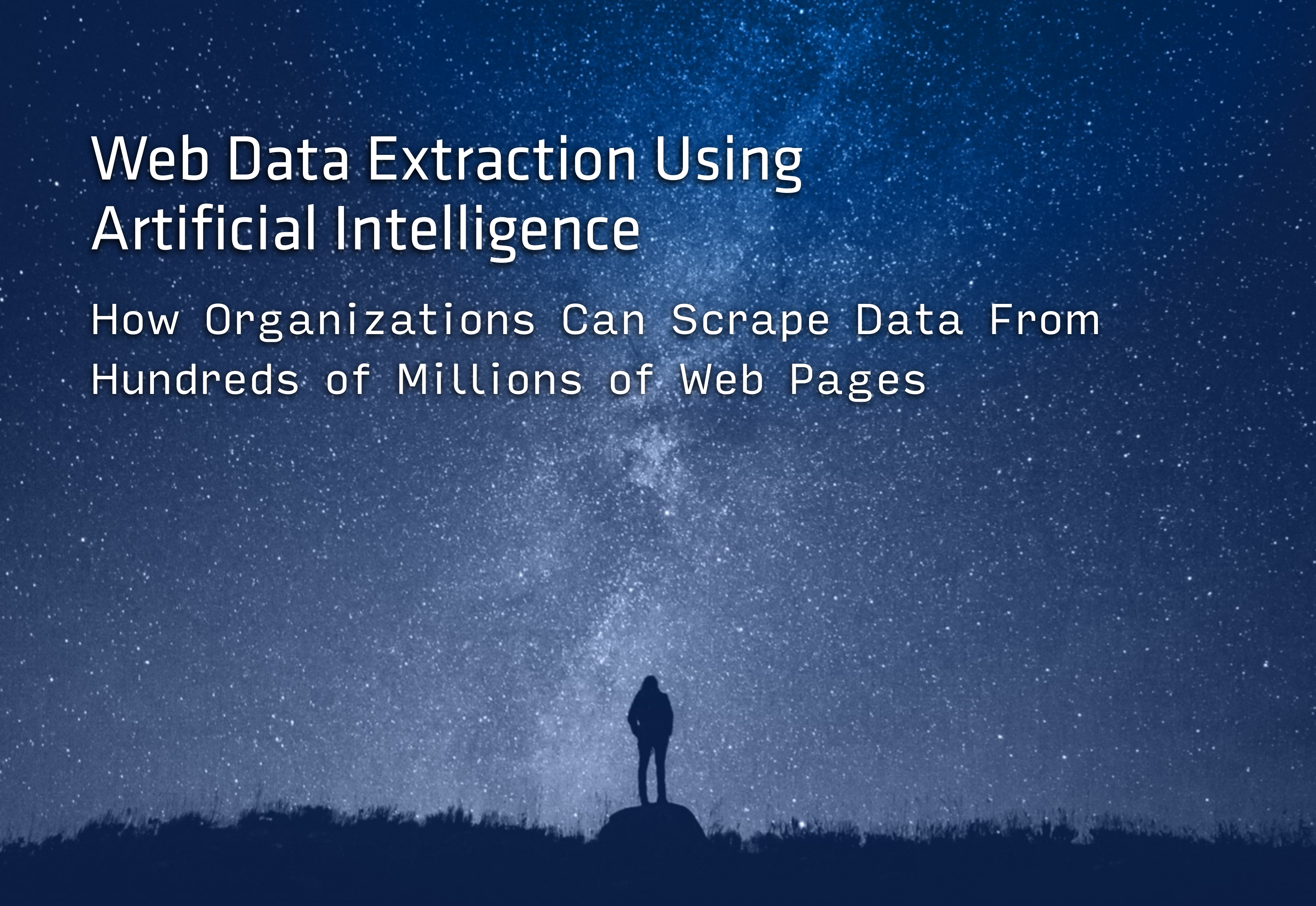
Web Data Extraction Using Artificial Intelligence

How Organizations Can Scrape Data From Hundreds of Millions of Web Pages



DIFFBOT

Introduction

A lot of incredible breakthroughs in the field of artificial intelligence (AI) have happened in recent years thanks to algorithmic advances, new hardware like GPUs allowing for computational scale, and the availability of high-quality training datasets. Artificial intelligence is being used by many companies to enhance computer vision, machine learning, and other advanced technologies.

Companies are using these AI-enhanced technologies to streamline business processes and even revolutionize industries. For example, AI is helping companies automate knowledge worker tasks speeding up the time it takes to find relevant information. AI is helping companies detect and prevent fraudulent business transactions. AI is even helping companies extract data from millions of web pages.

Extracting Data From Web Pages with AI

Many companies use web scraping software for a variety of reasons such as monitoring and retrieving the latest articles from news sites, retrieving product information from e-commerce websites, and keeping tabs on competitors.

Traditionally, computer vision has only been used to help robots and autonomous vehicles identify objects in the real world. Few are aware that computer vision with the help of AI can be used to extract data from web pages automatically. Most companies continue to use legacy web scrapers that are time consuming to maintain and not always accurate, unaware that there is a better way.

About Legacy Web Scraping Software

Most companies typically use legacy rules-based web scraping software to extract data from web pages. “Rules-based” can mean several things when it comes to web scraping. It can mean manual selection of CSS selectors or XPaths. It can also mean systems like Import.io which use manual selectors, but also look for repeating DOM elements to form an array of objects from a single listing page. One drawback to the latter rules-based scraping method is that it doesn’t provide the same wealth of data that you would get from an individual product page.

While legacy rules-based web scraping software can work well when you only need data from a couple of websites, most web scraping solutions aren’t practical when it comes to scraping information from thousands of web pages, and many aren’t always accurate. With the help of artificial intelligence and computer vision, companies can not only quickly extract data from millions of web pages but also achieve above human level accuracy. Artificial intelligence can be used to solve key problems with legacy web scraping software.

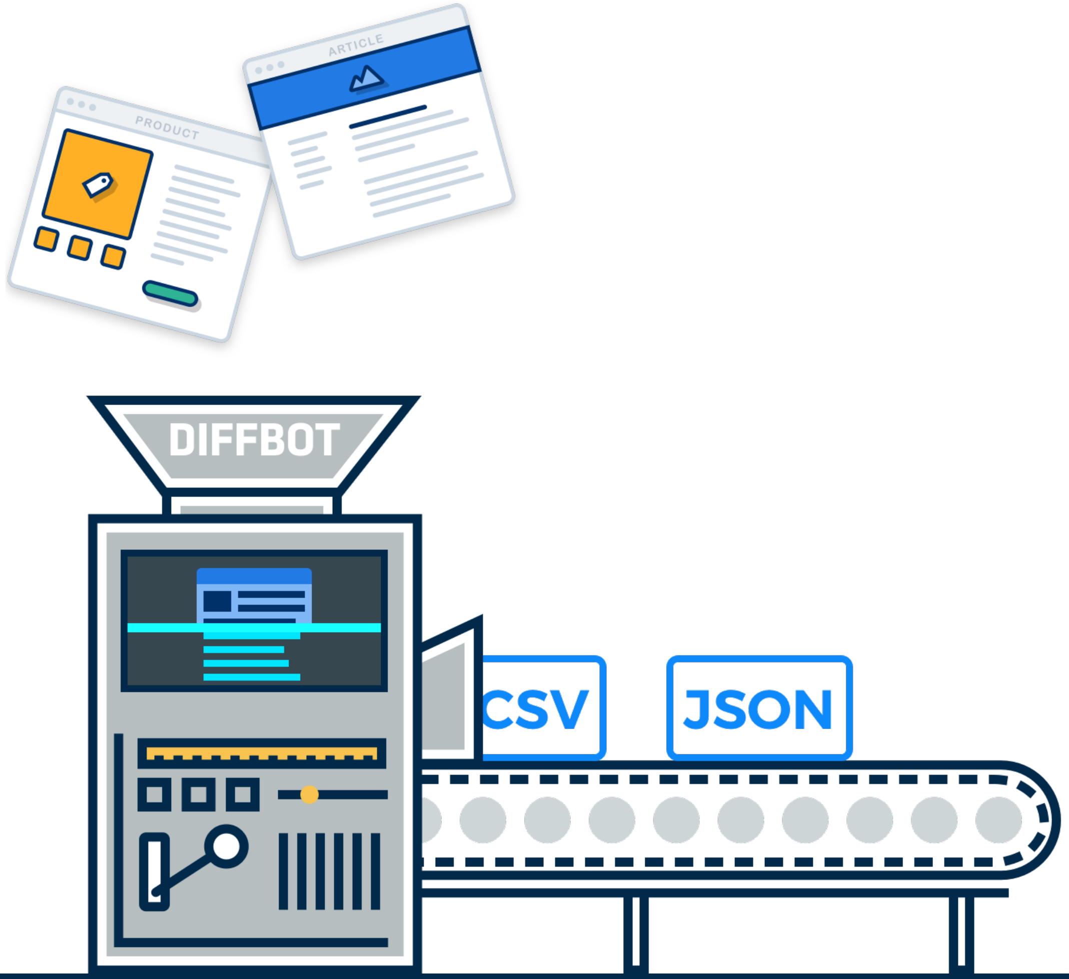
The Problems with Legacy Web Scraping Software

Import.io, Mozenda, and Scrapy are examples of legacy web scraping software that companies use today. Some companies have built in-house web scraping software that requires a lot of maintenance and may not always achieve human-level accuracy. Legacy web scraping software is rules-based which can be problematic. For example, rules-based systems require a lot of maintenance and regular human intervention; when a web page changes, the rules must be updated to reflect those changes.

Rules-based systems are difficult, if not impossible, to scale. If a technology company is looking to scrape information from tens of thousands of pages, that task is not practical to do manually. Rules would have to be created and maintained for each website. On average, half of those websites would break every week which would make rules patching and maintenance a full-time job for entire teams.

Rules-based systems aren't practical in many cases. For example, Amazon has unique layouts for different types of products and international variations like Amazon.in. Amazon has tens of thousands of regularly changing templates. eBay allows sellers to create their own arbitrary markup within product pages. Both Amazon and eBay have millions of web pages that change often making them difficult, if not impossible, to target.

Where legacy rules-based web scraping techniques fail, AI-powered computer vision techniques shine. Diffbot uses AI-powered computer vision and machine learning to provide a platform capable of automatically extracting data from web pages at above human-level accuracy whether it's one page or one million.



About Diffbot

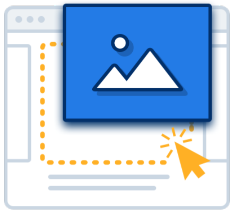
Diffbot uses artificial intelligence, machine learning, and computer vision to extract every piece of data from a web page. Give Diffbot a URL, and it will extract and convert the data into a structured version of the web page. Diffbot extracts data from web pages automatically, so you don't have to create any rules, have site-specific training, or perform any maintenance. Diffbot products include Automatic APIs, Crawlbot and Bulk Processing, and Custom APIs.

Diffbot Automatic APIs

Diffbot Automatic APIs automatically extract data from specific web page types such as news articles, blog posts, forum threads, article comments, images, products, product reviews, and videos. Diffbot's Analyze API is a page classifier, if the API determines a page is one of the supported page types, it extracts that page using the appropriate automatic API.

Crawlbot and Bulk Processing

Crawlbot is an AI-powered web crawler that uses any Diffbot API to extract data from entire websites. Crawlbot creates a single, structured, searchable index of a website's data. **Bulk processing** allows structured data to be extracted from multiple URLs in a single job.



Custom APIs

Diffbot **Custom APIs** can be used to enhance automatic API extraction further with a simple, point and click interface. For example, if you wanted to extract data from an Amazon product page, but also needed to extract data about the marketplace seller, you could use a custom API to append the data field to Diffbot's automatic product extraction.

What Makes Diffbot Different

Diffbot uses AI, computer vision, and machine learning to automatically extract data from web pages. Despite the significant advancements made in recent years in the field of artificial intelligence and computer vision, no web scraping solutions use the same combination of advanced technologies that Diffbot uses to extract data.

Above Human-Level Accuracy

A unique combination of AI, computer vision, and machine learning techniques allow Diffbot to achieve above human-level accuracy. The company rigorously tests the accuracy of its data extraction and has internal benchmarks that measure the accuracy of the information that's extracted.

"*We can confidently say that from all of the cases we've encountered, Diffbot's data extraction is higher than human-level accuracy,*" says Tung. "*The remarkable thing is now we have an AI system that can understand pages better than people, and it can do it at a massive scale too.*"

Low Maintenance

Rules-based scraping software requires that the rules be manually updated if a page layout changes, for instance, if a site is doing A/B testing or redesigns a page. Diffbot's use of artificial intelligence to automatically extract data from web pages eliminates the need to set up and maintain rules manually.

Diffbot doesn't break if a page layout changes. If a site is completely redesigned or changes in any way, Diffbot continues to be able to extract data from that site; no maintenance required.

Scalable

Diffbot automatically extracts data from web pages whether it's one page or hundreds of millions of pages. Diffbot is ideal for extracting data from sites with a massive number of pages like Amazon and eBay. Few, if any, legacy rules-based web scraping platforms are capable of scraping web pages at a massive scale.

It should be noted that web scraping is not the same thing as web crawling. Web crawling involves using bots to view and index information from websites; it's what search engines like Google and Microsoft Bing do. Web scraping involves targeting one or more web pages and extracting specific types of data such as product prices, images, or portions of text.

Works for Any Language

Diffbot is capable of extracting data from web pages for any language. Diffbot works regardless of the language; English, Chinese, Russian, French, Diffbot can extract data from the web page.

Works for One Page JavaScript Pages

Unlike most legacy web scraping software, Diffbot renders the web page in a headless browser which makes it capable of scraping data from one-page JavaScript pages or web pages that load content using AJAX.

Distributed Crawling Infrastructure

Diffbot is built upon a highly distributed crawling infrastructure that processes millions of web pages every day. In 2013, Diffbot acquired and open sourced Gigablast, an enterprise search engine and spider/ web crawler. Diffbot continues to develop and support Gigablast which powers its Crawlbot product.

> *“The approach we take is very different; it's pretty unique under the hood. We’ve built an AI system that can look at and interpret a page like a human being does,” says Diffbot CEO Mike Tung. “It uses a bunch of computer vision and natural language processing techniques on the page to convert the page into a structured version.”*

Unique Proxy IPs

Some sites make attempts to block bot activity for a variety of reasons. In those instances, proxy IPs are required to access and extract web page data. Diffbot offers access to tens of thousands of unique proxy IPs that allow for region/country-specific web page data extraction. Diffbot’s robust extraction platform, handles hundreds of edge-cases.

A Bot That Can Make Sense of a Web Page...

Diffbot is a very complex AI system with many neural networks. It uses computer vision and natural language processing to extract every piece of data from a web page. It uses statistical machine learning techniques to classify all the various components of the page. It uses machine learning to convert the data into a structured version of the web page. Finally, Diffbot uses AI to automate these processes.

"Using a combination of these machine learning techniques, Diffbot ends up being very similar to what the human visual processing system is doing when you're viewing a web page and trying to make sense of it," says Tung. *"Our bot essentially sees exactly what your eyes see on the page. Because we know how a web page is rendered, we have access to every pixel on the web page. This allows us to understand the relationships and context between the various types of data on the page; enabling us to extract it as structured data."*

Diffbot's novel combination of AI, machine learning, and computer vision techniques is not easily imitated by other data extraction platforms. Diffbot has accomplished what legacy web scraping platforms have not; built a platform that understands web pages better than people.

Companies no longer have to rely on legacy rules-based software to scrape data from web pages. With the help of AI and computer vision, organizations can scale their data ambitions to scrape data from hundreds of millions of web pages at above human-level accuracy.

www.diffbot.com
sales@diffbot.com
1-855-885-4800

451 N. Shoreline Blvd.
Mountain View, CA 94043