oxylabs®

The Application of AI & ML in Large-Scale Data Gathering Operations

Introduction

Finding a more manageable solution for a large-scale data gathering has been on the minds of many in the web scraping community. Experts saw a lot of potential in applying AI (Artificial Intelligence) and ML (Machine Learning) to web scraping. Only recently, steps towards data gathering automation using AI have been taken. This comes as no surprise, as recent advancements in computing solutions allowed AI and ML to become reliable for large-scale use.

There has been a lot of hype about AI/ML technologies and their implementation in web scraping. However, the hype relates to unrealistic expectations that ML techniques can solve anything, whereas in practice to solve a basic problem there is often no need to use AI/ML models, as they tend to be uneconomical to deploy. Simple, yet clever AI/ML solutions might go a long way when used wisely.

To better understand the application of AI-powered solutions in data gathering, it's best to learn what AI is to begin with. In this paper, we'll go over the basics of AI and ML and their applications to the current web scraping systems and solutions in the market.

1. Understanding the terminology 4
Artificial Intelligence (AI) 6
Machine Learning (ML) 8
Deep Learning 11

2. Applying AI and ML to web scraping 13

3. Explaining web scraping: value chain 15
Scraper development and its support 17
Proxy acquisition and management 18
Data fetching and parsing 18

4. Current AI-based web scraping solutions 19

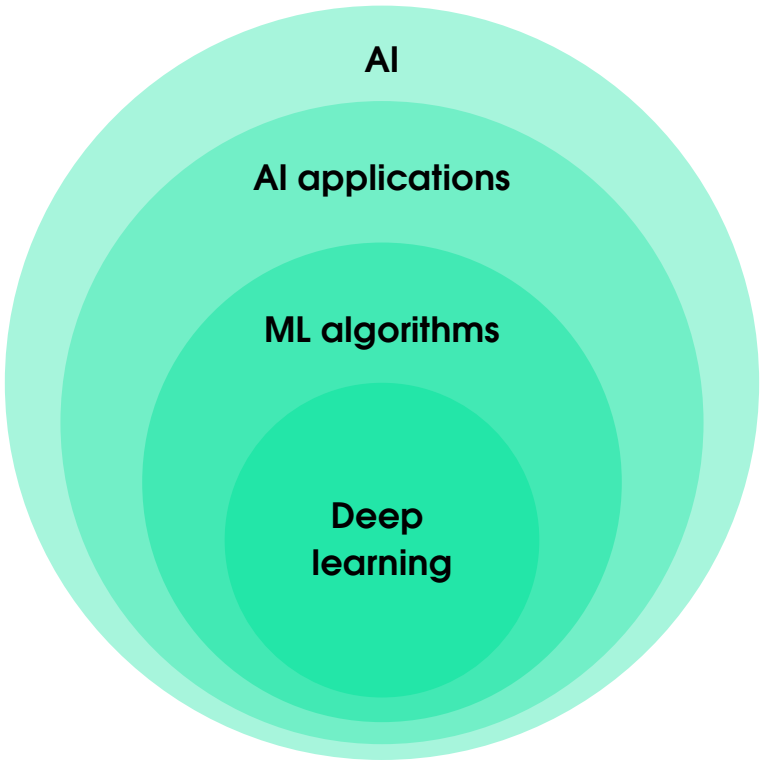
5. Comparing Self-Managed Proxies to Web Unblocker 23

6. Summary 25

1.

Understanding the terminology

AI is a term that is often misused or used instead of "AI applications". AI applications, ML algorithms, and Deep Learning is a subset of AI. All terms are intertwined and often used together.



In web scraping, AI is often used too broadly. Truthfully,when it comes to automating web scraping ML and Deep Learning are more applicable. Let's go term by term to unravel these terminologies a bit better.

Artificial Intelligence (AI)

The basic concept of AI is the idea of building machines or computers that are capable of thinking like humans. Humans learn from mistakes, adapt over time and pick up any information given to them in their surroundings. Thus, human behaviour is the basis of AI.

With this human blueprint, AI can simulate abstract thought and eventually develop the ability to learn.

AI can be split into two branches:

- **Applied AI** - uses the principles of simulating human thought to carry out one specific task.
- **Generalized AI** - seeks to develop machine intelligences that can learn to do any task that humans can.

As the modern everyday consumer, we have a closer relationship with **applied AI**, as it is a technology we slowly have adopted into our everyday lives. For example, the smartphone assistants like Apple's Siri or Google's Google Assistant, or the self-driving Tesla cars.

Applied AI is also well known in the financial world. Its uses range from fraud detection to improving customer service by predicting what services customers will need. Applied AI is also used in manufacturing as it helps manage workforces and production processes.

Generalized AI is still a ways off, as to simulate the human thought process would require a much better understanding of the brain, as well a lot more computing power than researchers currently have. Of course, given the speed of technological evolution, this might not last for too long.

There is a new generation of computer chip technologies known as neuromorphic processors being designed that will allow to run a more efficient brain-simulation code. Or supercomputers like IBM's Watson that combines AI and analytical software to create high-level simulations of human neurological processes.

Machine Learning (ML)

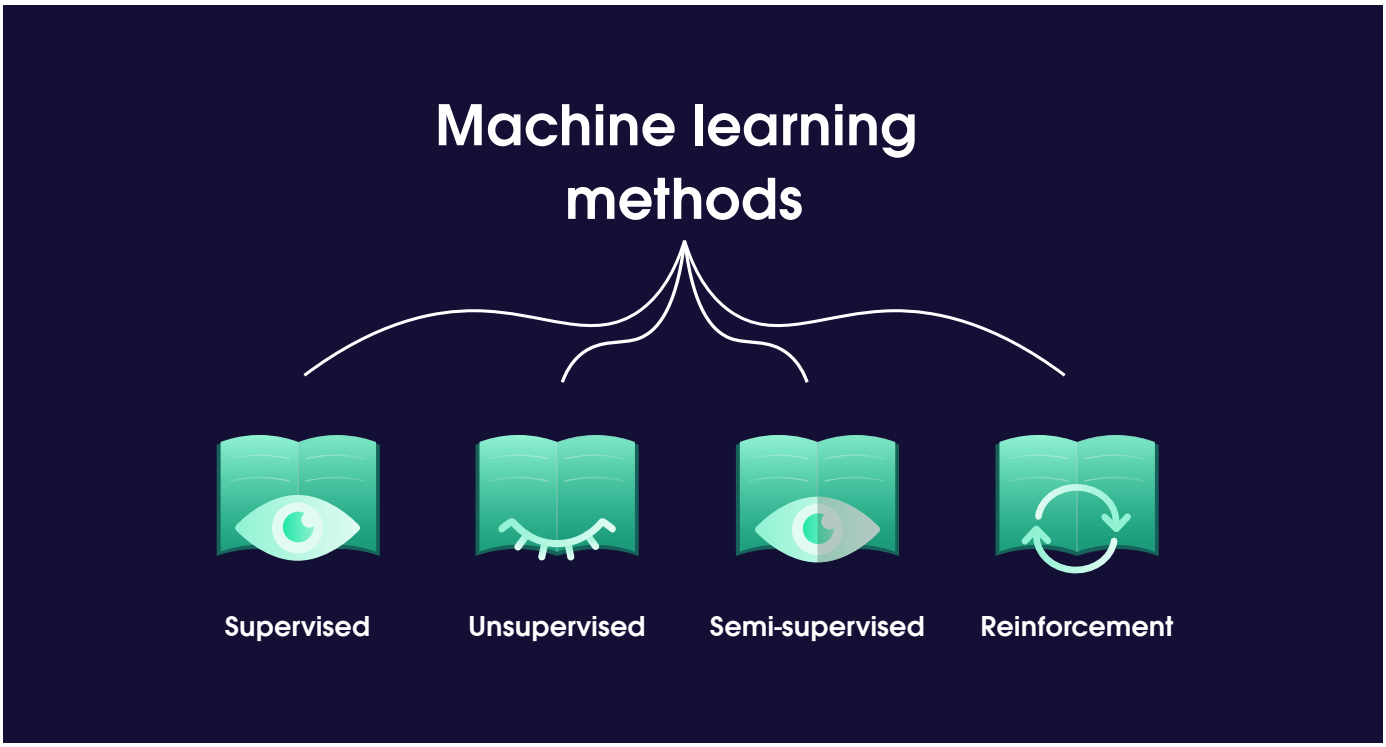
At its most basic design, Machine Learning (ML) is a practice of using different algorithms to process data, learn from it, and then decide or predict something in question. So instead of hand-coding software routines and sets of instructions to complete a task, a machine is taught to perform the assignment.

ML is the byproduct of AI, and together with algorithmic approaches came decision tree learning, SVM (Support Vector Machine), Random Forest, Gradient Boosting, clustering, reinforcement learning, and the application of Bayesian interface (a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available).

So to put it simply, Machine Learning is an application of Artificial Intelligence that provides systems with the ability to automatically learn and improve from experience without being programmed by a developer.

Machine Learning methods

Machine Learning algorithms can be categorized into a few different parts:



Supervised Machine Learning - this can be applied to what has been learned to newly acquired data using labeled examples (labeled data are groups of samples that have been tagged to one or more labels, the final output) to predict future events. It starts by analyzing an already known training dataset, then produces an inferred function and makes predictions for possible output values. After enough training, the system can provide targets for any new input. However, after the model is trained, it doesn't have any ability to adjust itself. During training, it compares what it has predicted to the real value and adjusts its coefficients to perform more precisely. After the model has been trained and new input data are provided, it doesn't know if the prediction was correct as the data is not labeled.

Unsupervised Machine Learning - unlike supervised algorithms, these are used when the training information is not labeled. Unsupervised learning learns how systems can infer a function to describe a hidden structure from unlabeled data.

Semi-supervised Machine Learning - this goes somewhere in between supervised and unsupervised learning since they use both labeled and unlabeled data for training. Usually, a small amount of labeled data and a large amount of unlabeled data is used. The ability to use unlabeled data helps increase learning accuracy compared to only using a smaller amount of labeled data to train the model. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources to train orearn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

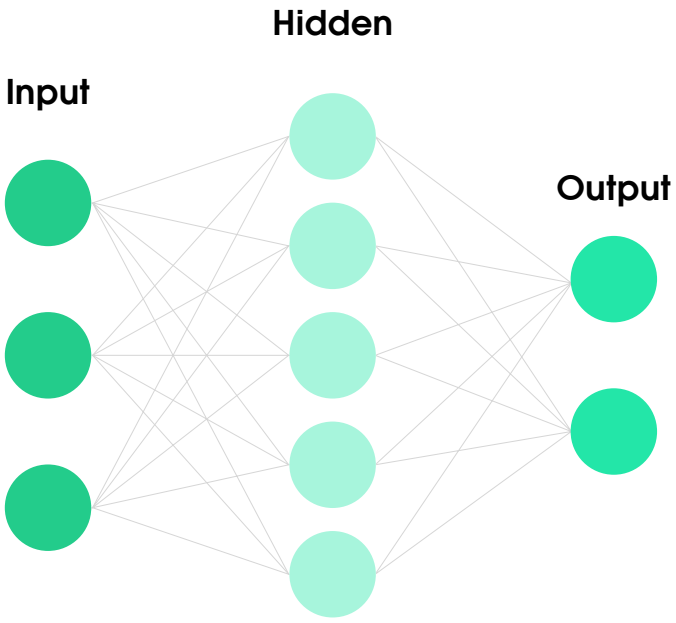
Reinforcement Machine Learning - it's a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search, and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Deep Learning

Deep Learning is a subset of Machine Learning, which is a subset of Artificial Intelligence. Deep Learning is a type of Machine Learning that draws its inspiration from the structure of a human brain. What it does is it attempts to arrive at conclusions in a similar manner to humans by analyzing data with a logical structure. To be able to do this, Deep Learning uses a multi-layered structure of algorithms called neural networks.

The neural network design is based on a human brain. Just like we can identify patterns to classify information, neural networks can be taught to perform the same tasks on data.

Below is an example of the most basic neural network called feedforward. In this network, the information moves only in one direction - forward, through the three main layers. Starting from the input layer, through the hidden layer, and to the output layer. There are no cycles or loops in the network.



Neural networks enable us to perform many tasks, such as clustering, classification, or regression. With neural networks, we can group or sort unlabeled data according to similarities among the samples. In the case of classification, we can train the network on a labeled dataset in order to classify samples into different categories.

In general, neural networks can perform the same tasks as classical algorithms of Machine Learning, but not the other way around. Artificial neural networks have unique capabilities that enable Deep Learning models to solve tasks that Machine Learning models can never solve.

All advances in artificial intelligence in the recent years are due to Deep Learning. Without Deep Learning, we would not have self-driving cars, chatbots or personal assistants like Alexa and Siri. Netflix or Youtube would have no idea which movies or TV series we like or dislike. Hidden behind all of these technologies are neural networks.

We can even go so far as to say that today a new industrial revolution is taking place, driven by artificial neural networks and Deep Learning. The same can be said about the current revolution with its application in web scraping.

2.

Applying AI and ML to web scraping

In what way AI and ML can be applied in web scraping, and how can it innovate and improve it? According to Oxylabs Next-Gen Residential Proxy AI & ML advisory board member Jonas Kubilius, an AI researcher, Marie Sklodowska-Curie Alumnus, and Co-Founder of Three Thirds:

> **There are recurring patterns in web content that are typically scraped, such as how prices are encoded and displayed, so in principle, ML should be able to learn to spot these patterns and extract the relevant information.**

Jonas Kubilius
Oxylabs Next-Gen Residential Proxy AI &
ML advisory board member

He also adds that "the research challenge here is to learn models that generalize well across various websites or that can learn from a few human-provided examples. The engineering challenge is to scale up these solutions to realistic web scraping loads and pipelines."

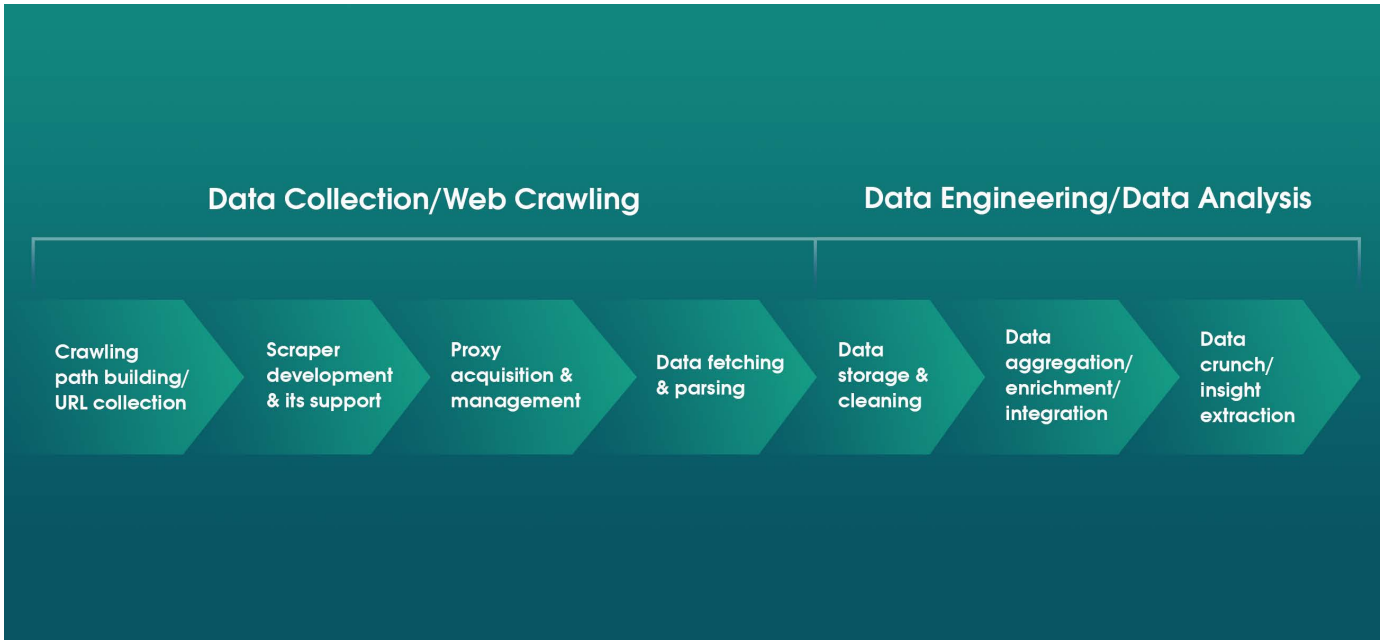
Instead of manually developing and managing the scraper's code for each new website and URL, creating an AI and ML-powered solution will simplify the data gathering pipeline. This will take care of proxy pool management, data parsing maintenance, and other tedious work.

Not only does AI and ML-powered solutions enable developers to build highly scalable data extraction tools, but it also enables data science teams to prototype rapidly. It also stands as a backup to your existing custom-built code if it was ever to break.

3.

Explaining web scraping: value chain

To understand how ML can be applied to web scraping, we should analyze the value chain of data collection.



Please note that it's not truly possible to automate crawling path building without human interference, even though it's an essential part of data gathering. This is simply due to the fact that only a human can truly identify the URLs from which data can be extracted.

Scraper development and its support

Building a scraper comes with many unique issues. There are a lot of factors to look out for when doing so:

- Choosing a language that will be used in some years' time and will have the support to go with it, as well as picking APIs, frameworks, etc.
- Testing, building, taking apart, and building it all over again - the basic process of new tool development. Repeating these steps many times is necessary to create something great as good is not good enough in the current market.
- Infrastructure management and maintenance of what is already built will be a daily process, as languages get updates, and websites create stronger anti-bot measures.
- Overcoming fingerprinting anti-measures will be another tough one, as this requires mimicking an organic user's behaviour.
- Rendering JavaScript-heavy websites at scale will be your other headache, as many modern sites use JavaScript to load content asynchronously (i.e. hides part of the content to not be visible during the initial page load). Scraping a JavaScript-heavy website requires many complex tools and libraries and a good set of development skills to overcome them.

> Some of the scraper support challenges can be solved with ML based solutions. These include **website change handling, overcoming fingerprinting anti-measures, and rendering JavaScript-heavy websites.**

For example, after implementing AI-powered dynamic fingerprinting the crawler bot will have their own web-footprint schedules. Just like regular internet users, they will show their organic behavior to visited websites. By mimicking real user behavior, it can overcome CAPTCHAs, blocks, etc.

Proxy acquisition and management

Proxy management will be a challenge, especially those new to scraping. There are so many little mistakes one can do to block batches of proxies before reaching the desired result of scraping. A good practise is proxy rotation, but all issues do not disappear with just rotation, and constant management and upkeep of the infrastructure will be needed.

AI-based proxy rotator or an automatic retry system based on ML algorithms would help with overcoming blocks.

Data fetching and parsing

Data parsing is the process of making the acquired data understandable and usable. Most data gathering methods return results that are incredibly hard to understand as they are in a raw code format. That is why parsing is a necessary tool to create structured results to make them ready to use.

Creating a parser might sound easy. However, like most of our other mentioned issues, maintenance will cause the biggest problems down the road. Adapting to different page formats and website changes will be a constant struggle and will take up time from your developer's day more often than you expect.

Implementation of an **ML-based parser** could simplify the daily tasks of a developer and help parse data from specific domains. Of course, complete automation might have its own caveats, and having a human in the loop would be important for the near future.

4.

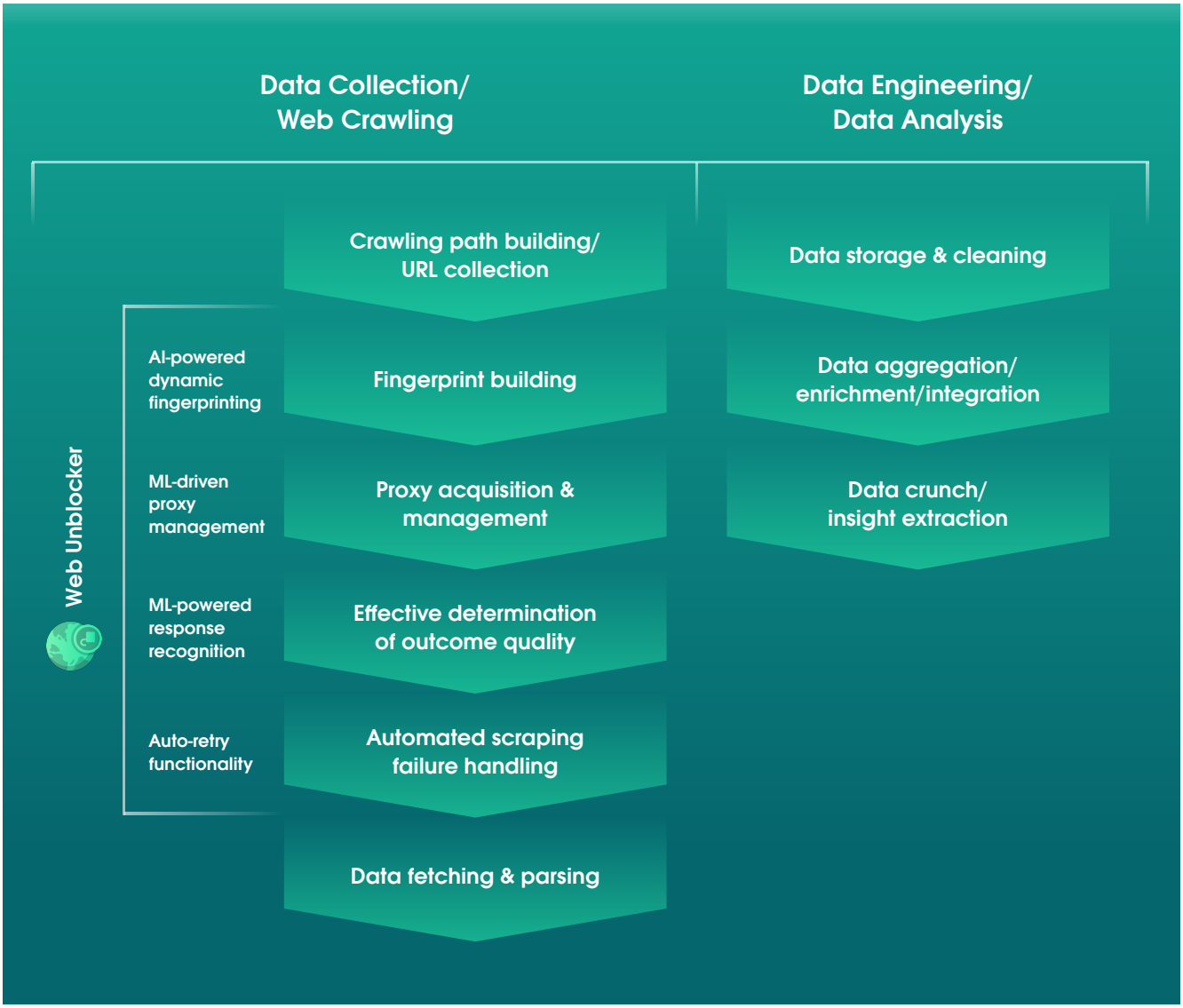
Current AI-based web scraping solutions

Oxylabs.io has recently introduced Web Unblocker powered by the latest AI and ML innovations. This new proxy solution was built with block-free data retrieval operations in mind.

The product is as customizable as a regular proxy, but at the same time, it guarantees a much higher success rate and requires less maintenance. Custom headers and IP stickiness are both supported, alongside reusable cookies and POST requests. Web Unblocker's main features are:

- Dynamic browser fingerprinting
- ML-driven proxy management
- ML-powered response recognition
- Auto-retry functionality
- JavaScript rendering

Going back to our previous web scraping value chain, you can see which parts of web scraping can be automated and improved with AI-powered Web Unblocker.



This project will be continuously developed and improved by Oxylabs in-house ML engineering team and a board of advisors, **Jonas Kubilius, Adi Andrei, Pujaa Rajan, and Ali Chaudhry**, specializing in the fields of Artificial Intelligence and ML engineering.

Adi Andrei

Over 20 years of experience with AI & ML technology, research software engineer at NASA, senior data scientist at Unilever and British Gas

MSc in Engineering Science Louisiana State University

Pujaa Rajan

USA Ambassador for Women in AI, Deep Learning Engineer at Node.io, ML Expert Google Developer

BSc Degree in Information Science Cornell University

Ali Chaudhry

Artificial Intelligence Consultant at UCL and Founder of Reinforcement Learning Community

PhD in Artificial Intelligence and Education UCL
Graduate Diploma from Harvard Extension School

Jonas Kubilius

AI Researcher, Marie Skłodowska-Curie Alumnus, and Co-Founder of Three Thirds

BSc Degree in Mathematics and Physics Massachusetts Institute of Technology, MSc in Artificial Intelligence, KU Leuven

5.

Comparing Self-Managed Proxies to Web Unblocker

Looking from a business perspective, upgrading to Web Unblocker can help you save on proxy maintenance and development. Handling CAPTCHAs, keeping up with website updates, and handling JavaScript rendering will be done by the application of AI and ML.

Features	Self-managed proxies	Web Unblocker
Integration method	Backconnect Proxy or direct connection	Backconnect Proxy
Worldwide geotargeting	Country-level	Country, city or coordinate-level
Automated unblocking	×	✓
CAPTCHA bypass	×	✓
JavaScript rendering	×	✓
Full proxy management	×	✓
Advanced browser fingerprinting	×	✓
Auto-retry	×	✓

6.

Summary

AI applications, ML algorithms, and Deep Learning is a subset of AI. All terms are intertwined and often used together, though have different meanings.

The basic concept of AI is the idea of building machines or computers that are capable of thinking like humans.

Machine Learning is a practice of using different algorithms to parse data, learn from it, and then decide or predict something in question

Deep Learning attempts to arrive at conclusions in a similar manner to humans by analyzing data with a logical structure.

Creating an AI and ML-powered solution will simplify the data gathering pipeline. This will take care of proxy pool management, data parsing maintenance, and other repetitive work.

Web Unblocker powered by the latest AI and ML innovations is a new Oxylabs proxy solution built with sophisticated and block-free data retrieval operations in mind.

Proxy maintenance, development, handling captchas, keeping up with website updates, and handling JavaScript rendering will be done by the **application of AI and ML.**

oxylabs®

Want to Know More?

If you would like to know more about any of the topics mentioned in this white paper or learn about our products, please get in touch! Our team is ready to answer any of your questions and offer you the best solution for your business needs.

Get in touch with Oxylabs