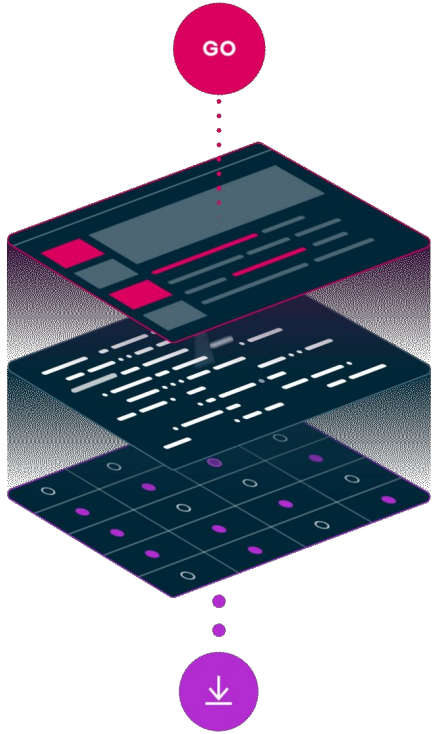
zyte

Enterprise-grade Scraping with AI



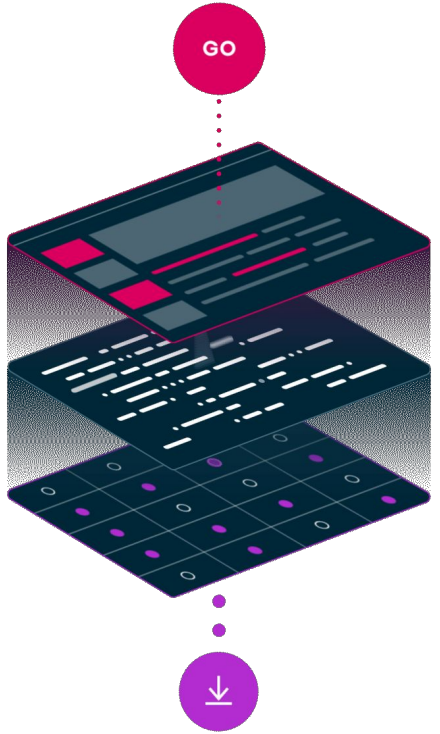
Speakers

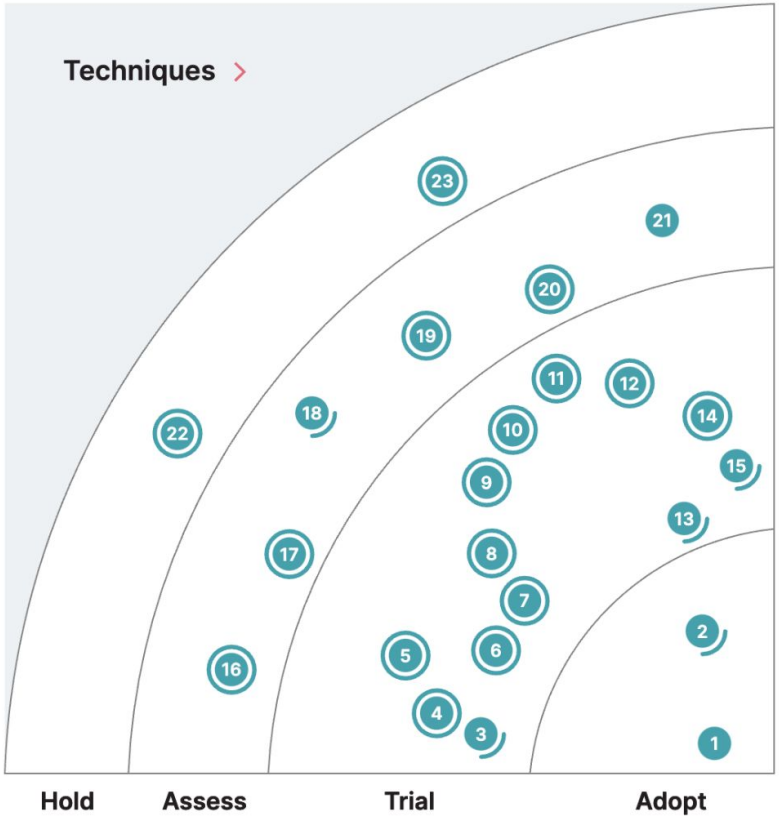
Iain Lennon
Chief Product Officer at Zyte

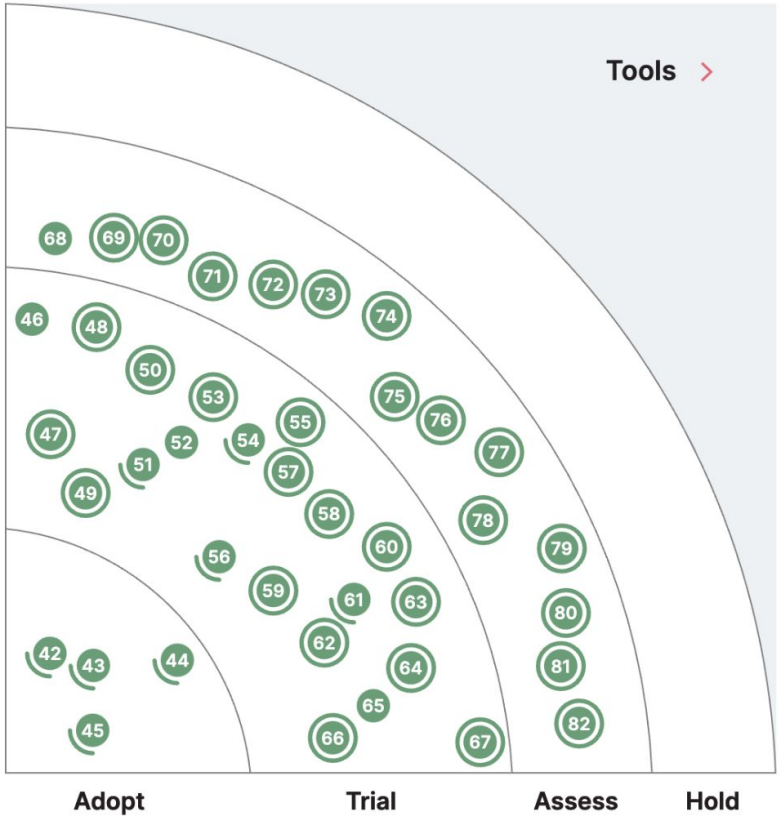
Adrian Chaves
Python Developer at Zyte

zyte

Enterprise-grade Scraping with AI




Techniques
23
21
20
19
18
22
11
12
14
10
15
9
13
17
8
7
2
6
5
16
4
3
1
Hold
Assess
Trial
Adopt


Tools
68
69
70
71
72
73
74
46
48
50
53
75
76
55
47
52
54
77
51
57
49
58
78
60
79
56
59
61
63
80
62
42
43
44
64
81
65
45
66
67
82
Adopt
Trial
Assess
Hold

The two major problem spaces of web scraping

The cat-and-mouse contest with bans

Web scraping APIs

The scalability problem of set-up & maintenance

Low/No Code?
Crowdsourcing?

AI? LLMs?

Practical testing with LLMs for scraping

Fast & flexible set-up

Rapidly extract data points without need for model training

Great for unstructured text

Extraction of discrete terms from unstructured descriptive fields

Data processing

UOM conversions, cleansing etc

Orchestration

Inc understanding and clarifying requirements

Training data

Rapid and flexible source for ML

Practical testing with LLMs for scraping

Immature for scraping

No robust solutions yet. Developer experience largely undefined

Reliability challenges

Hallucinations; unclear strategies to fix errors & control quality

High Costs

Use on every request is uneconomic

Incomplete

Needs integrations for bans, rendering

Legal Concerns

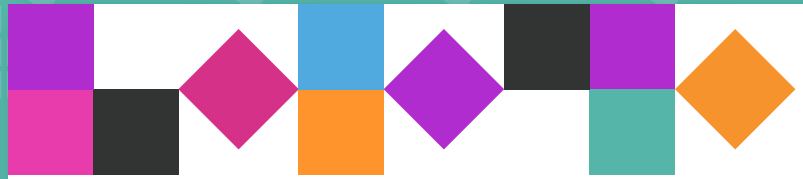
On training data, copyright

What do we need to solve the scaling problem?

An automated solution for
instant set-ups, that **addresses maintenance**

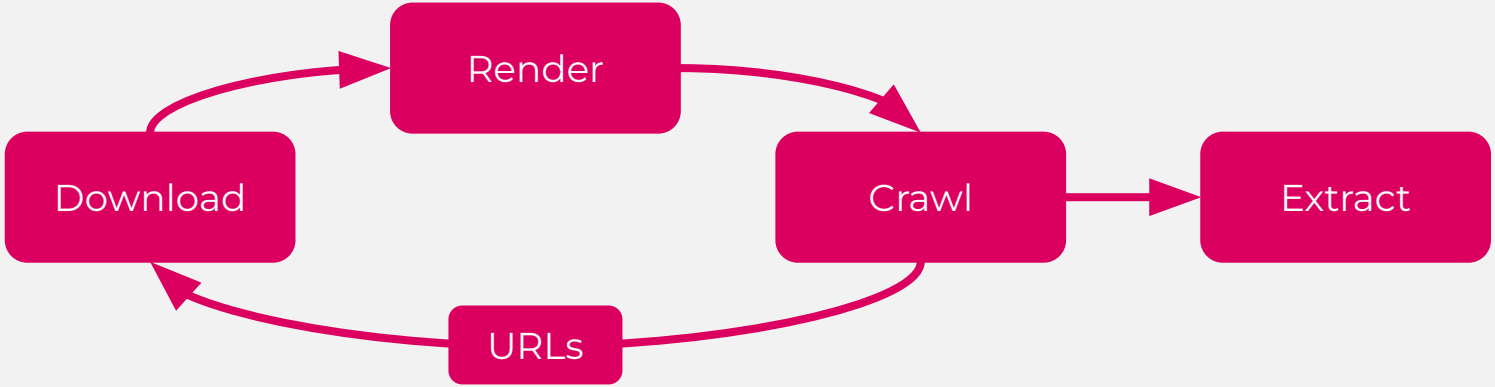
which is
accurate, **cost effective**, **compliant**,
complete and **controllable**

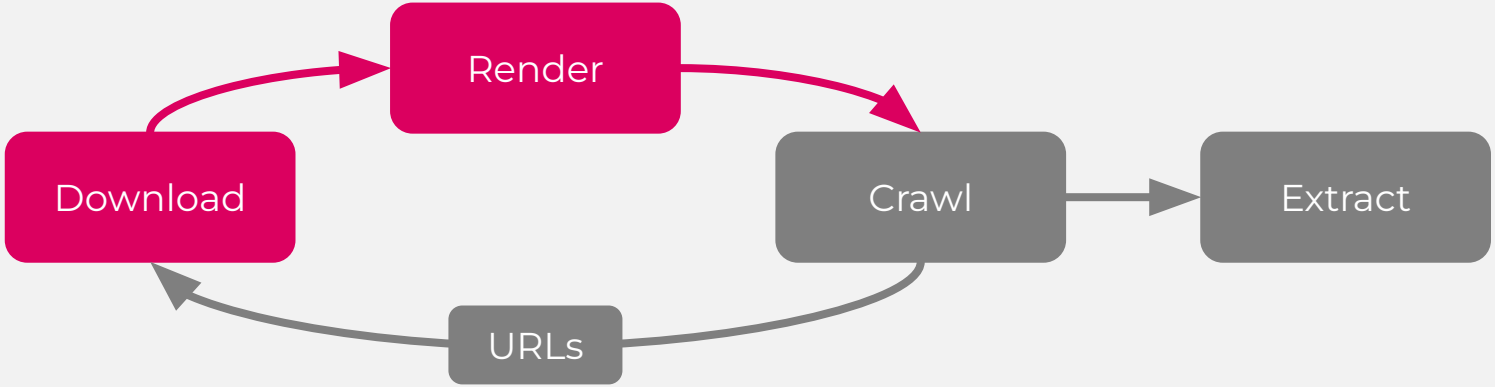
We’d like to show you this today



Demo

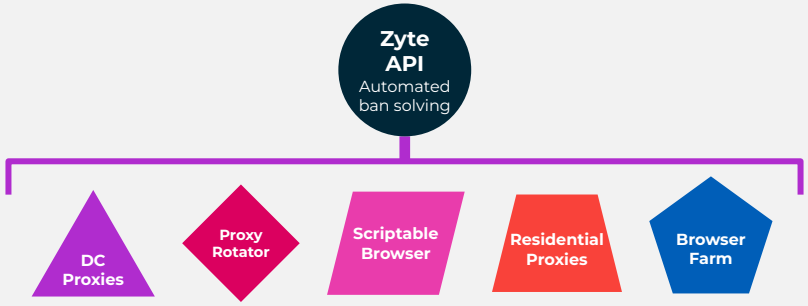
What's under the hood?


Render
Download
Crawl
Extract
URLs

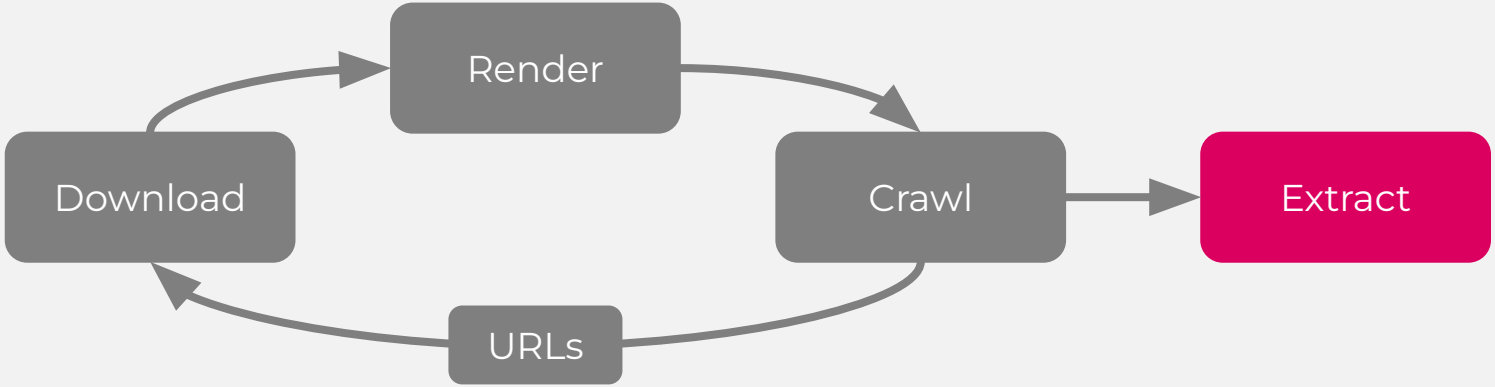

Render
Download
Crawl
Extract
URLs

Zyte API - Solving Bans

- Auto-applies tech required to solve bans of **all complexities**
- Proxyway: strongest **success**, **fastest** and **best cost** in market
- Priced in 5 tiers - competitive against all tools inc DC proxies

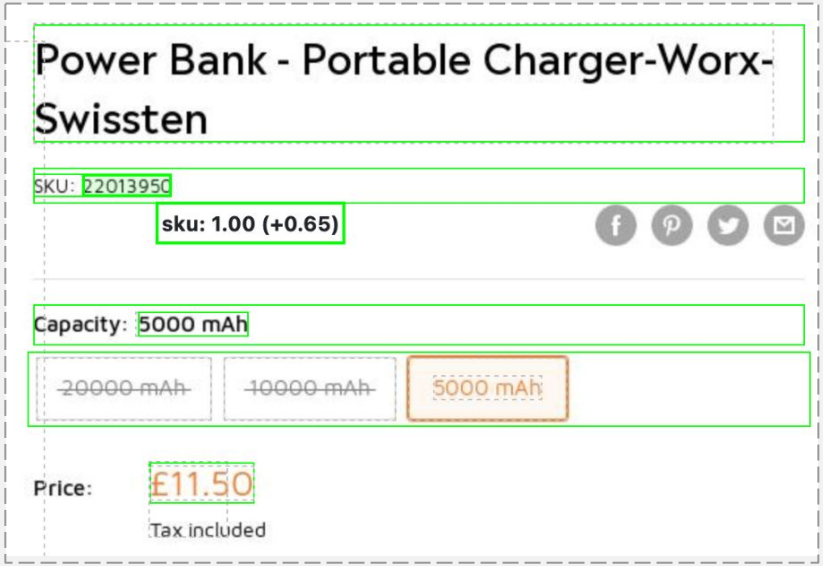


Website Tier	IP types used	Compute	Price per 1k requests (HTTP Response)	Price per 1k requests (Rendered HTML)
1			$0.08	$0.60
2			$0.14	$1.20
3			$0.26	$2.40
4			$0.42	$4.80
5			$0.76	$9.60


Render
Download
Crawl
Extract
URLs

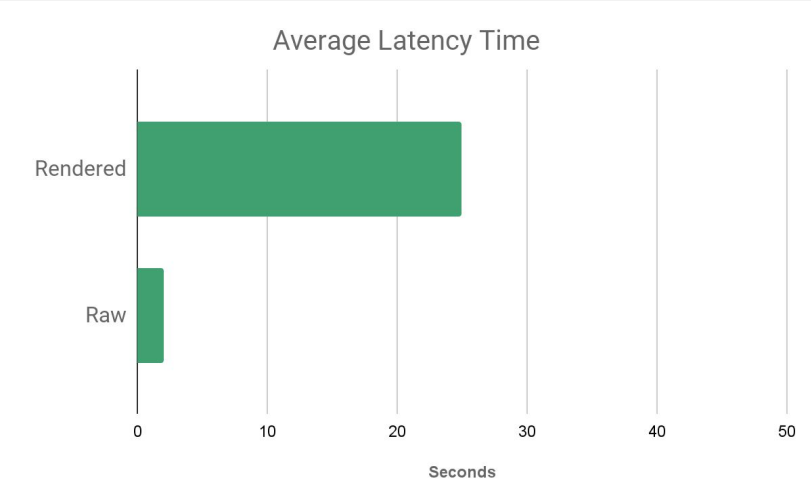
Zyte API - Automatic Extraction

- Supervised ML models developed for web scraping
- Mature and patented technology
- **Instant set-up** for any product site. Articles and Jobs next
- Runs on every page - self-healing to **address maintenance**



Highly cost effective

- Zyte API extraction is around **50 times cheaper** than ChatGPT 3.5
- Dual-mode HTML support –
  - Raw HTML (HttpResponse) for best speed and cost
  - Rendered HTML for Javascript-heavy sites etc



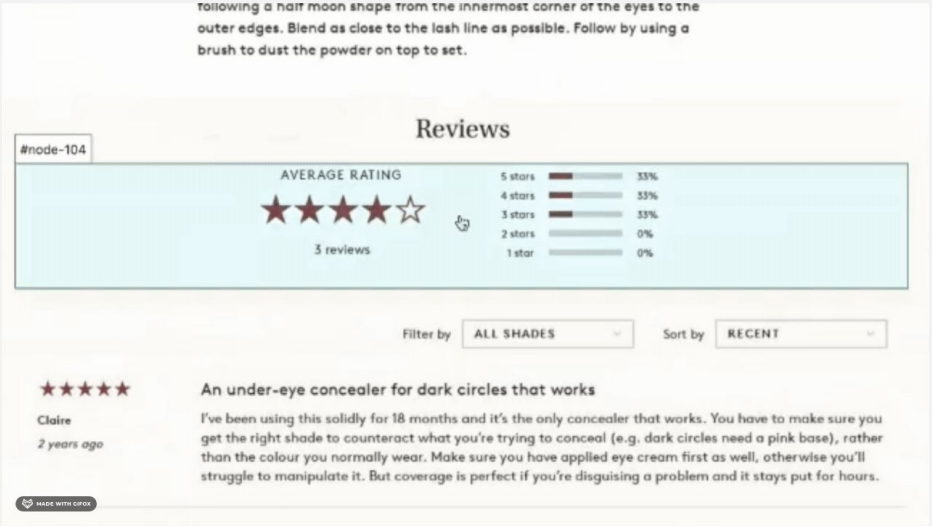
Quality : LLMs versus Supervised ML

- For fields expressed in page structure **supervised ML is more accurate**
- We use LLMs for field extraction from **unstructured** text

	Chat GPT 3.5	Zyte API
Price	0.84	**0.95**
Currency	0.89	**0.96**
SKU	0.43	**0.87**
MPN	0.11	**0.79**

Human-in-the-loop AI for fast Quality support

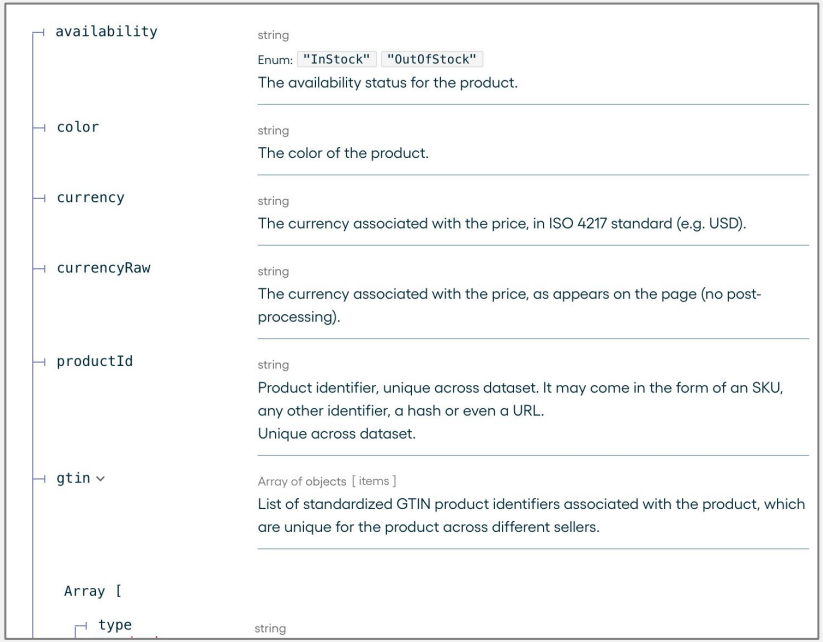
- Our Support teams make site-specific corrections for ML where needed
- If sites have an issue, 80% are resolved with this
- Fixes deployed **same-day**

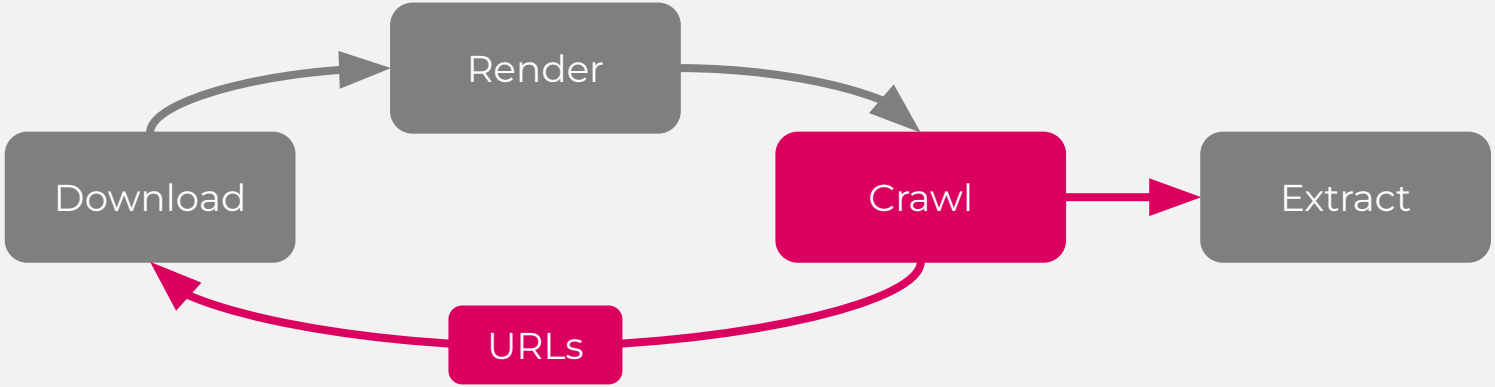


zyte

Schemas support out-of-box compliance

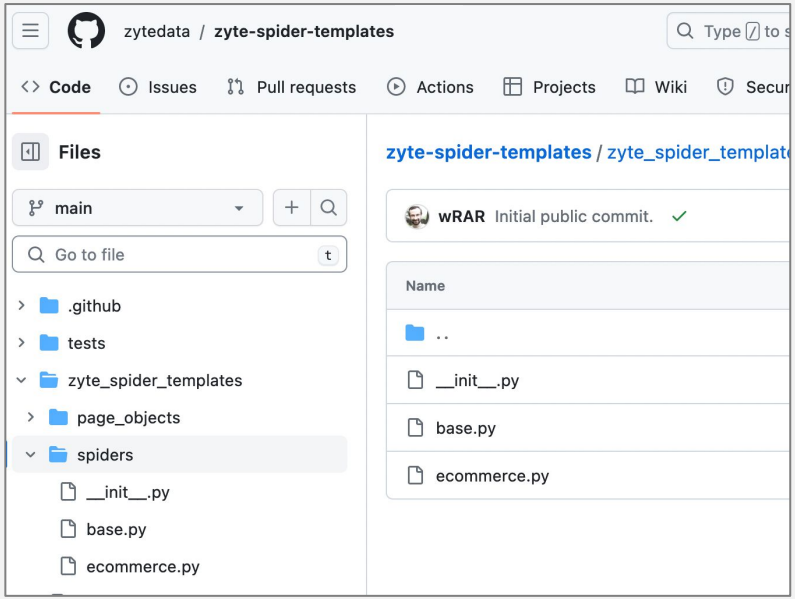
- Extraction to defined schemas - easily combine data across sites
- These **exclude fields** containing PII and copyrighted content for compliance confidence

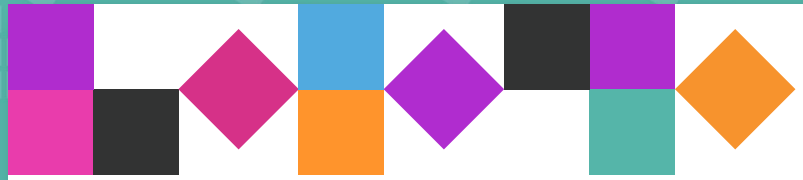



Render
Download
Crawl
Extract
URLs

Controlled with open-source Scrapy templates

- Open-source spiders calling Zyte API for bans & extraction
- Adjust crawling or override extraction for **complete control**
- Scrapy is largest and most widely supported scraping framework

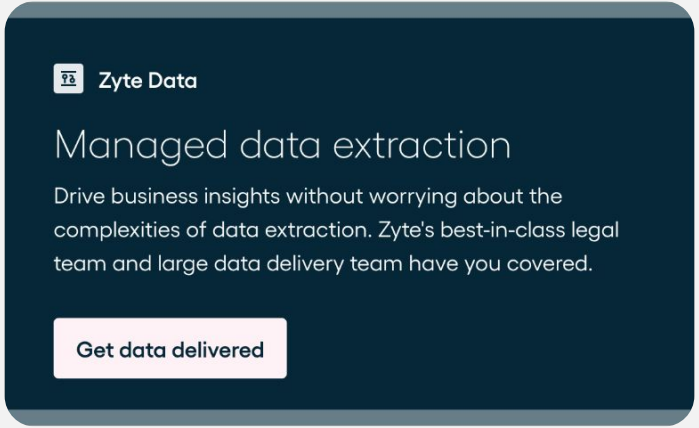




Demo

We eat our own dog food

- Zyte's services arm have been using pre-release versions of this solution
- Seeing **50-80%** reduction in set-up, slashed maintenance
- Why not 100%? Data QA, custom fields. This **automates the standard**.
- Unlocking new types of opportunity



Production ready AI solution to scalability

Instant set-up

For any product site, articles & jobs next

Save Maintenance

Models run on every request - self-healing

Cost effective

50x cheaper than ChatGPT 3.5

High Quality

Dedicated scraping models, supported

Compliant

Schemas exclude PII & copyrighted fields

Complete

Powerful ban solving for complete solution

Control

Tune to precise needs with Scrapy

Please, try this yourself

We’d like to give you first access

Thank you