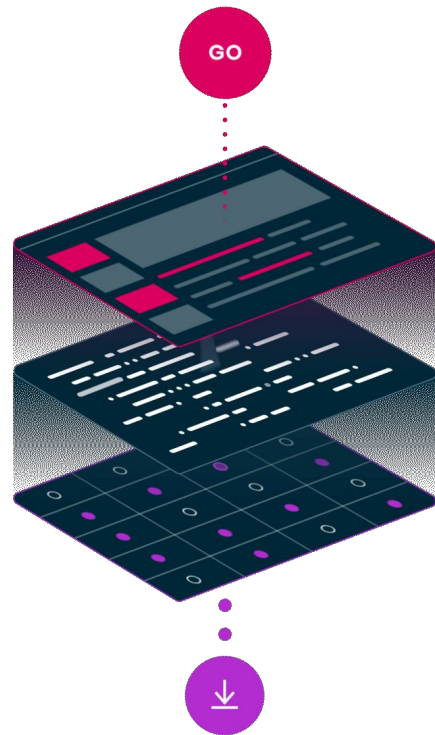




Preview of new innovations at Zyte



Speakers

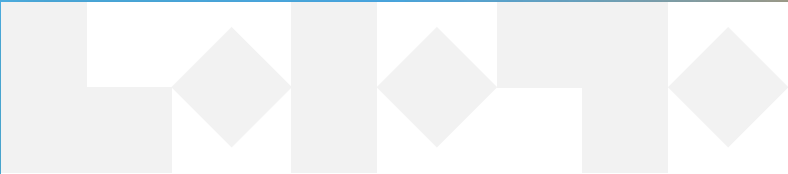
Iain Lennon

Chief Product Officer at Zyte

Akshay Philar

Head of Development at Zyte





Solving the most painful site ban problem

The browser designed for web data extraction

Solving two scaling challenges to faster growth



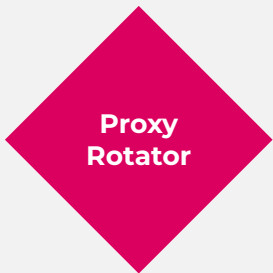
Solving the most painful site ban problem

The browser designed for web data extraction

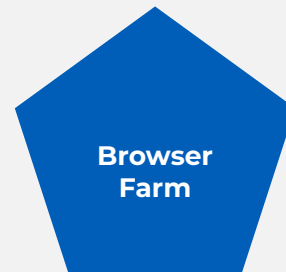
Solving two scaling challenges to faster growth



Simple



Midrange



Sophisticated



Problem 1

Stay ahead of
complex bans

Crawling Strategy

User behavior pattern

Browser

Browser fingerprinting

Proxy

Geofencing

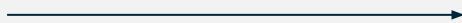
Proxy

IP Blocking

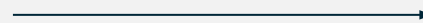
Proxy

Header, TLS, TCP/IP Fingerprinting

\$XX



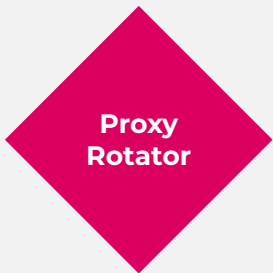
\$XXX



\$XXXX



DC IPs



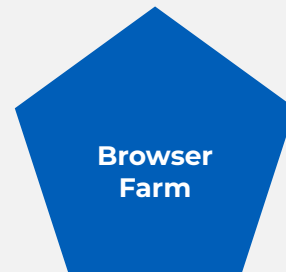
Proxy
Rotator



Scriptable
Browser



Residential
IPs



Browser
Farm

Simple

Midrange

Sophisticated

Problem 2

Battle to solve bans
in a **lean** way for the
'middle majority'

Problem 1

Stay ahead of
complex bans

Battling to solve bans and contain costs

Combine **multiple tools**, each with their own contract

Then the **site changes**, and it breaks again

Lost time & productivity, unhappy customers, **unhappy team!**

Find a config that works, monitor success rates

Loss of **data**, scrambling to fix

Zyte API

Automated
antiban

DC
IPs

Proxy
Rotator

Scriptable
Browser

Residential
IPs

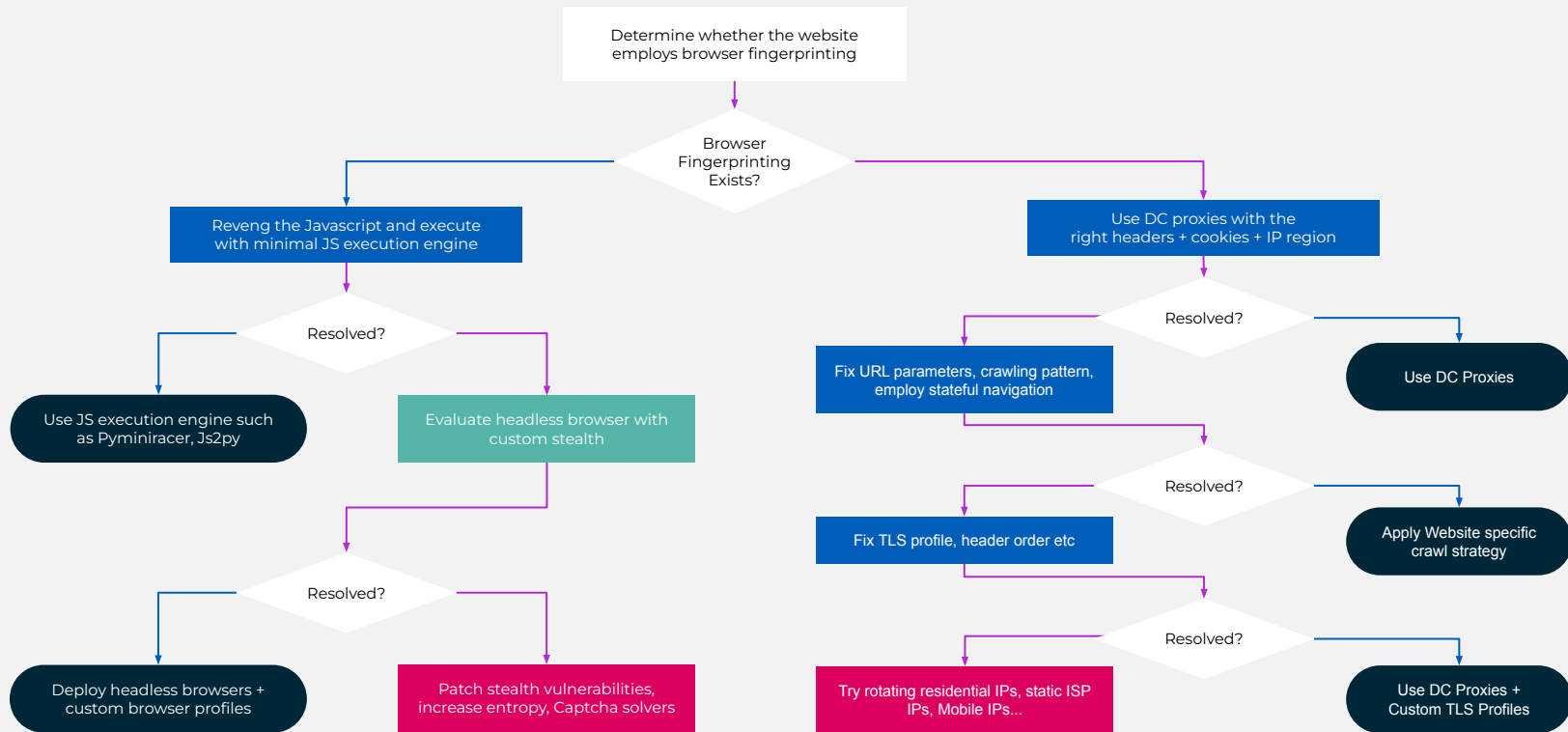
Browser
Farm

Automatically avoids site
bans using minimum
required infrastructure

Per-site pricing - pay only
for what each site needs

So you can use a single,
automated tool without
a cost trade-off

Why Zyte API?



Why Zyte API?

```
FP ID: 1fa9135c79421f40d9bd7aa50ad72d81ad2c3991e5a286fd85deb34a7aa43b66
Fuzzy: 24278f775bd0ea33735967a74ab0c7c4ffa576217f77bafec638000000000000
Diffs: 24278f775bd0ea33735967a74ab0c7c4ffa576217f77bafec638000000000000
```

fingerprints renewed 9/19/2022

Browser

trust score: 8% **F-**
visits: 1
first: 9/28/2022, 9:50:57 AM
alive: 0 hrs
auto-delete in 29.99854922 days
shadow: 0 **+10**

trash (2): **6447b820** **-11**
lies (4): **26e512ae** **-124**
errors (0): none
session (1): 4349d2ea
revisions (0): none
loose fp (0): **b43cb599** **+5**

bot: 0.25:bold-fraud:elc:00001001
idle min-max: 0-0 hrs
performance benchmark: 3046.10 ms

add a signature

Sign

eb6b354a

Few Days Before Extract Summit

```
FP ID: ff92bfd36095d8af41acc8e1f36e86e6ca49f3084d6aab189906052f37a391f9
Fuzzy: c8577ff25bd0efd4d3994aa74abfcbclffa59b8cfff774afe9635000000000000
Diffs: c8577ff25bd0efd4d3994aa74abfcbclffa59b8cfff774afe9635000000000000
```

fingerprints renewed 9/19/2022

Browser

trust score: 75.5% **C+**
visits: 1
first: 9/29/2022, 7:40:12 AM
alive: 0 hrs
auto-delete in 29.99997858 days
shadow: 0 **+10**

trash (0): none
lies (0): none
errors (0): none
session (1): 02b2cbac
revisions (0): none
loose fp (0): 02e4161e **+5**

bot: 0.38:crafty-attacker:lpv:01001001
idle min-max: 0-0 hrs
performance benchmark: 752.00 ms

add a signature

Sign

eb6b354a

Today



Why Zyte API?



DevOps/Infra

- Proxy Management
- Browsers
- Lambda/Cloud Run
- Captcha Solver

Domain Knowledge

- Captcha Solver
- Crawling Strategies
- Website Specific Knowledge
- Antiban Stealth

Specialized Technical Skills

- Puppeteer
Or
Playwright
Or
Selenium
- TLS, TCP/IP
- Javascript Reverse Engineering

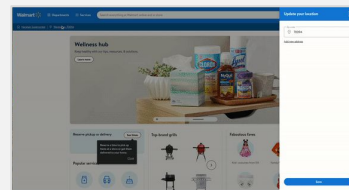
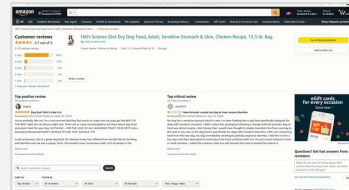
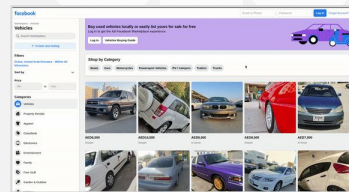
Why Zyte API?

Website Navigation Patterns

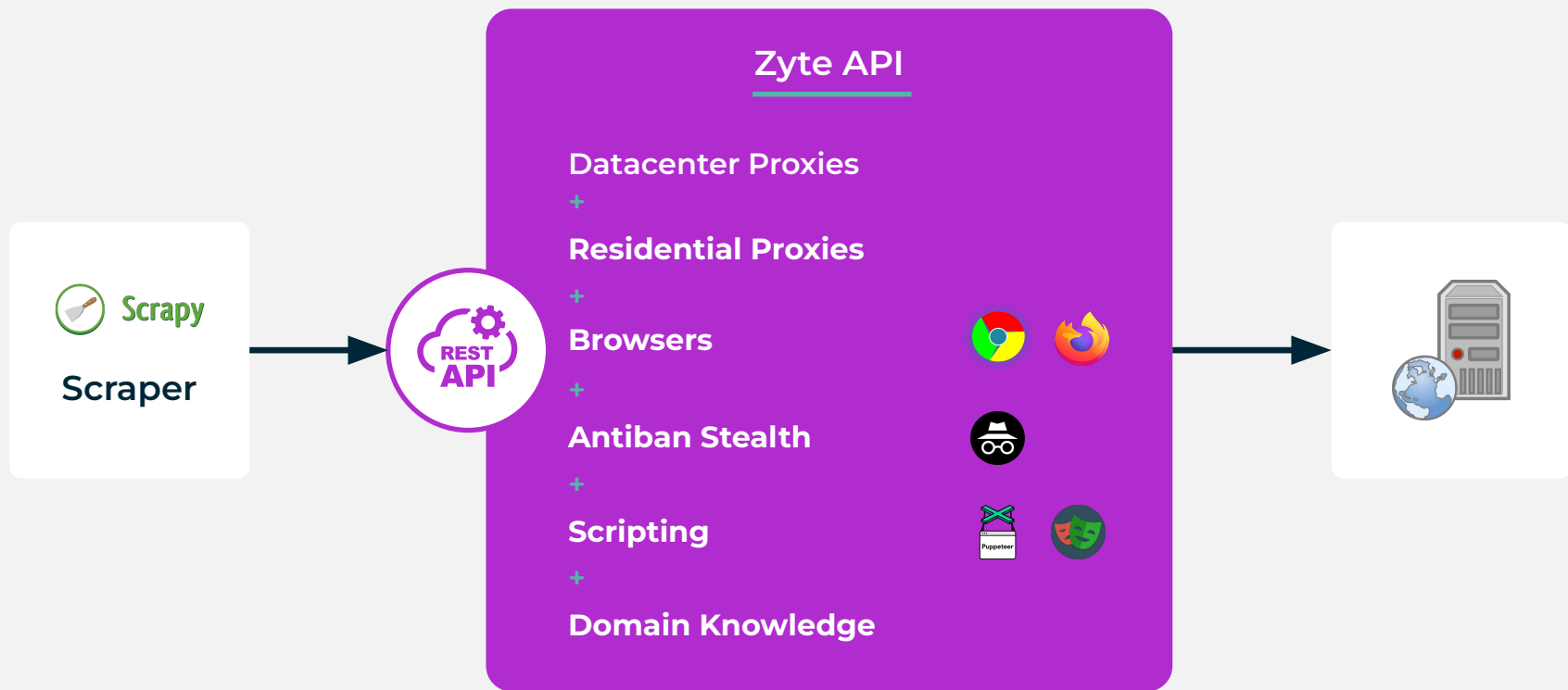
Websites use extensive Javascript to employ navigation patterns such as pagination, infinite scrolling, delayed navigation, form submission etc

These are built with browsers in mind and hence require complex request orchestration when using non-browser HTTP clients

Handling these patterns require manipulation of headers, cookies, URL parameters etc. which from experience is fairly brittle and requires some amount of maintenance.



Why Zyte API?





Modes of Operation



Browserless

- **Low latency, high performance lightweight client**
- **Returns raw Html**
- **Mimics TLS profile of browsers with built-in passive client fingerprinting evasions**
- **Both datacenter and residential proxies baked-in**
- **Ideal for replicating XHR requests, downloading binary content etc**

Browserful

- **Full-blown browsers with built-in antiban stealth**
- **Returns rendered Html and screenshots.**
- **Support for page actions and scripting functionality.**
- **Built-in support for datacenter and residential proxies**
- **Ideal for stateful navigation scenarios**



Zyte Data API specification

Process a single URL, return the result

POST /extract

Process a single URL, return the result. This endpoint blocks until the result is ready. It is intended for short-running operations.

REQUEST BODY SCHEMA: application/json

url	string
required	An absolute URL to extract data from.
requestHeaders	object (RequestHeaders) Subset of the request headers fields that are a part of section 5 of RFC 7231 . This is an advanced feature that can impact anti-ban performance, and hence must be tested before being used in production. It is recommended to use the action functionality wherever possible, since it helps to avoid the use of raw HTTP headers, which could be quite error-prone.
httpRequestMethod	string Enum: "GET" "POST" "PUT" "DELETE" "OPTIONS" "TRACE" "PATCH" The HTTP method to be used as part of the request
httpRequestBody	string <byte> Content to be sent to the server as part of the request. The data should be base64-encoded. This cannot be used with browserHtml.
customHttpRequestHeaders	Array of objects (CustomHttpRequestHeader) [items] Any set of headers required for a specific use case. Headers provided as part of this parameter override the default headers. This is an advanced feature that can impact anti-ban performance and hence must be tested before being used in production. <ul style="list-style-type: none">This feature cannot be used with browserHtml, only with httpResponseBody

Request samples

Content type

application/json

Example

Retrieve raw HTTP content from a webpage

Copy

```
{
  "url": "https://example.com",
  "httpResponseBody": true
}
```

Response samples

200

400

401

403

422

429

451

500

503

520

521

default

Content type

application/json

Copy Expand all Collapse all



```
{
  "url": "https://example.com/item-page/",
  "httpResponseBody": "string",
  "httpResponseHeaders": [
    + { ... }
  ],
  "browserHtml": "<html>Downloaded data.</html>",
  "screenshot": "string",
  "echoData": { },
  "jobId": "example-job-1",
  "actions": [
    + { ... }
  ]
}
```




Demo

zyte

Search

  Mark Twain

Mark Twain / Gettings started with Zyte API

You are about to make your first request

Choose how to make your request

Access through Interface

Request

OR

Access through API

When authenticating, use the API key as the username and a blank password.

API Key

b92d5fec464245699fbbc1ac901e6fde

Generate New API Key

Here are some examples of making requests through Zyte, using various tools and languages:

CURL

PYTHON

```
# Python 3 example using the requests library

import requests

API_URL = "https://api.zyte.com/v1/extract"
API_KEY = "b92d5fec464245699fbbc1ac901e6fde"
response = requests.post(API_URL, auth=(API_KEY, ''), json={
    "url": "https://example.com/foo/bar",
    "browserHtml": True
})
data = response.json()
# data['browserHtml'] contains the HTML of a web page
```

Check out our documentation to learn how to effectively use Zyte API.

zyte

|

18

zyte

Search

Mark Twain

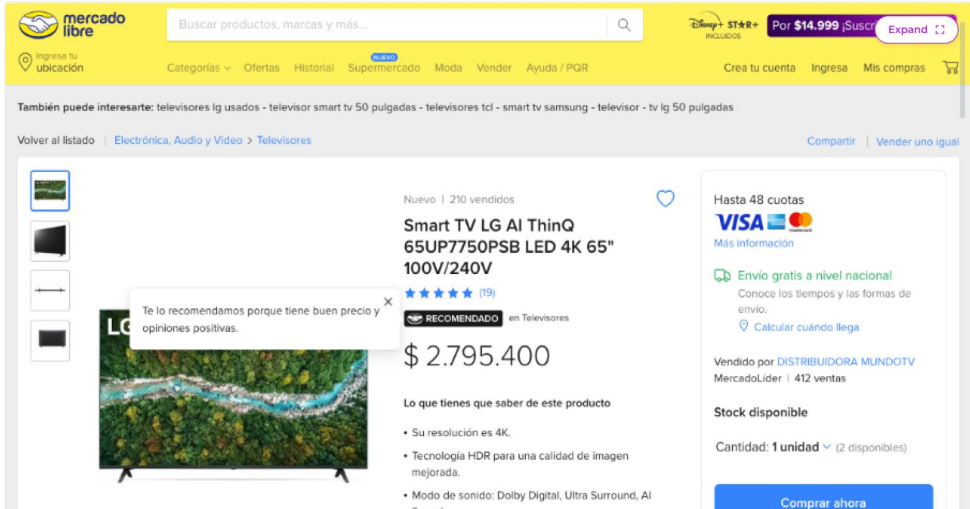
Mark Twain / Zyte API ON DEMAND / Request Summary / Request #123-123

Request Summary

REQUEST ID: 123-123
URL: <https://www.mercadolibre.com.co/smart-tv-lg-ai-thinq-65up7750psb-led-4k-65-100v240v/p/MCO184531472...>
API RESPONSE CODE: 200

OverviewInputOutput

Screenshot ☐ Raw HTML



Cost Breakdown

Cost Per

1 request

Outputs

Rendered HTML\$0.0058

Screenshots\$0.002

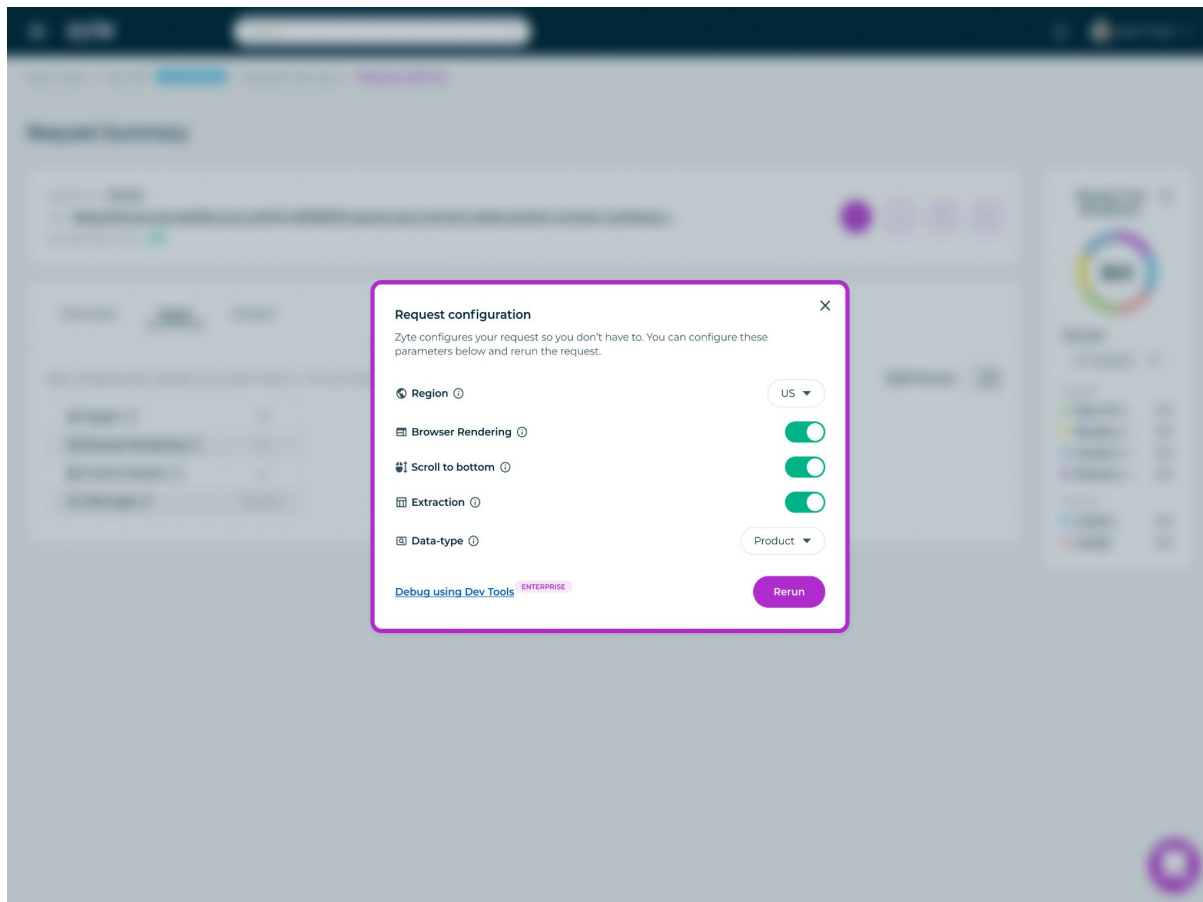
Features

Actions\$0.000027

zyte

|

19



zyte

Search

Mark Twain

Mark Twain / Zyte API ON DEMAND / Request Summary / Request #123-123

Request Summary

REQUEST ID 123-123
URL <https://www.mercadolibre.com.co/smart-tv-lg-ai-thinq-65up7750psb-led-4k-65-100v240v/p/MCO184531477...>
API RESPONSE CODE 200

OverviewInputOutput

REQUEST TIME 2022-05-01 12:08:08

RESPONSE TIME (S) 20

WEBSITE RESPONSE 200

REDIRECT URL <https://www.mercadolibre.com.co/smart-tv-lg-ai-thinq-65up7750psb-led-4k-65-100v240v/p/MCO184531477...>

API SESSION ID 11111

ECHO DATA Label

Cost Breakdown

\$0.007827

Cost Per

1 request

Outputs

- Rendered HTML \$0.0058
- Screenshots \$0.002

Features

- Actions \$0.000027

Zyte API

Launching October 27th



Solving the most painful site ban problem

The browser designed for web data extraction

Solving two scaling challenges to faster growth



& etc..

Tools designed for testing that we use in web extraction

Need to scale your own infrastructure

Browser stealth can be fragile

They don't appear as human browsers

Not well integrated into web extraction tooling

There are plugins, but they have gaps



Zyte API Page Actions

Integrated scriptable page interactions in a web data API

Zyte API Page Actions

- Abstracts away Playwright, Puppeteer and CDP
- Exposes low level actions such as click, type etc that mimic human behaviour
- Exposes high level actions such scrollBottom, searchKeyword
- Can only be executed on the page context
- Proxies, ad blocking, retries,
- Website specific configs etc are all take care of under the hood

Zyte API Actions API

click
type
hover
scrollTo
scrollBottom
waitForRequest
waitForResponse
searchKeyword

Playwright

Puppeteer

Chrome DevTools
Protocol

Actions Demo

```
{  
  "url":  
    "https://dribbble.com/tags/single_page_app",  
  "browserHtml": true,  
  "actions": [  
    {"action": "scrollBottom"}  
  ]  
}
```

The screenshot displays a Postman interface for a POST request to `https://api.zyte.com/v1/extract`. The 'Body' tab is selected, showing a JSON payload:

```
{  
  "url": "https://dribbble.com/tags/single_page_app",  
  "browserHtml": true,  
  "screenshot": true,  
  "screenshotOptions": {"fullPage": true},  
  "actions": [  
    {"action": "scrollBottom"}  
  ]  
}
```

The response status is 200 OK. Below the response, a 'Visualize' tab is active, showing a screenshot of a Dribbble 'Single Page App' design gallery. The gallery features various mobile app and web design mockups, including a 'Food app', a 'Healthy living' app, a 'Coffee app', and a 'Hotel management' app.

<https://www.postman.com/zyteapi/workspace/>

Scripting

Actions

```
{
  "url": "http://www.ulta.com/store",
  "actions": [
    {
      "action": "click",
      "onError": "continue",
      "selector": {
        "type": "css",
        "value": "#onetrust-accept-btn-handler"
      }
    },
    {
      "action": "waitForTimeout",
      "timeout": 1,
      "onError": "return"
    },
    {
      "action": "click",
      "delay": 0,
      "button": "left",
      "onError": "return",
      "selector": {
        "type": "css",
        "value": ".ProductDetail__findInStore_Button button"
      }
    },
    {
      "action": "waitForSelector",
      "timeout": 5,
      "selector": {
        "type": "css",
        "value": "#searchField"
      }
    },
    {
      "action": "type",
      "text": "07040",
      "selector": {
        "type": "css",
        "value": "#searchField"
      }
    },
    {
      "action": "click",
      "selector": {
        "type": "css",
        "value": ".FindInStore__searchIcon button"
      }
    },
    {
      "action": "waitForSelector",
      "timeout": 3,
      "selector": {
        "type": "css",
        "value": ".StoreList"
      }
    }
  ]
}
```

Scripting

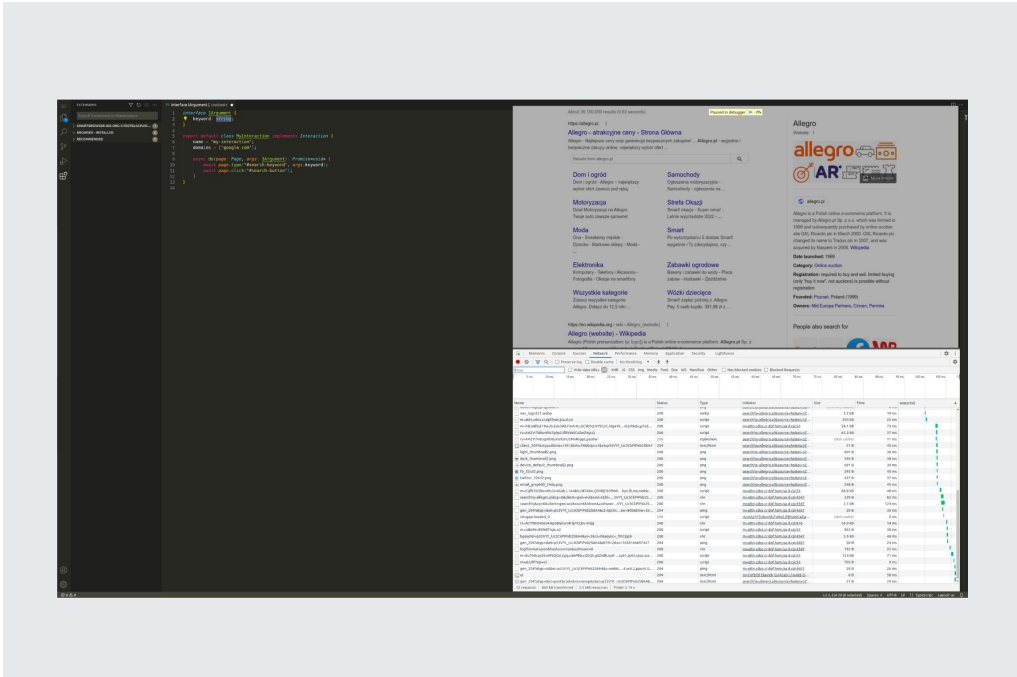
1. Deploy script to Zyte API

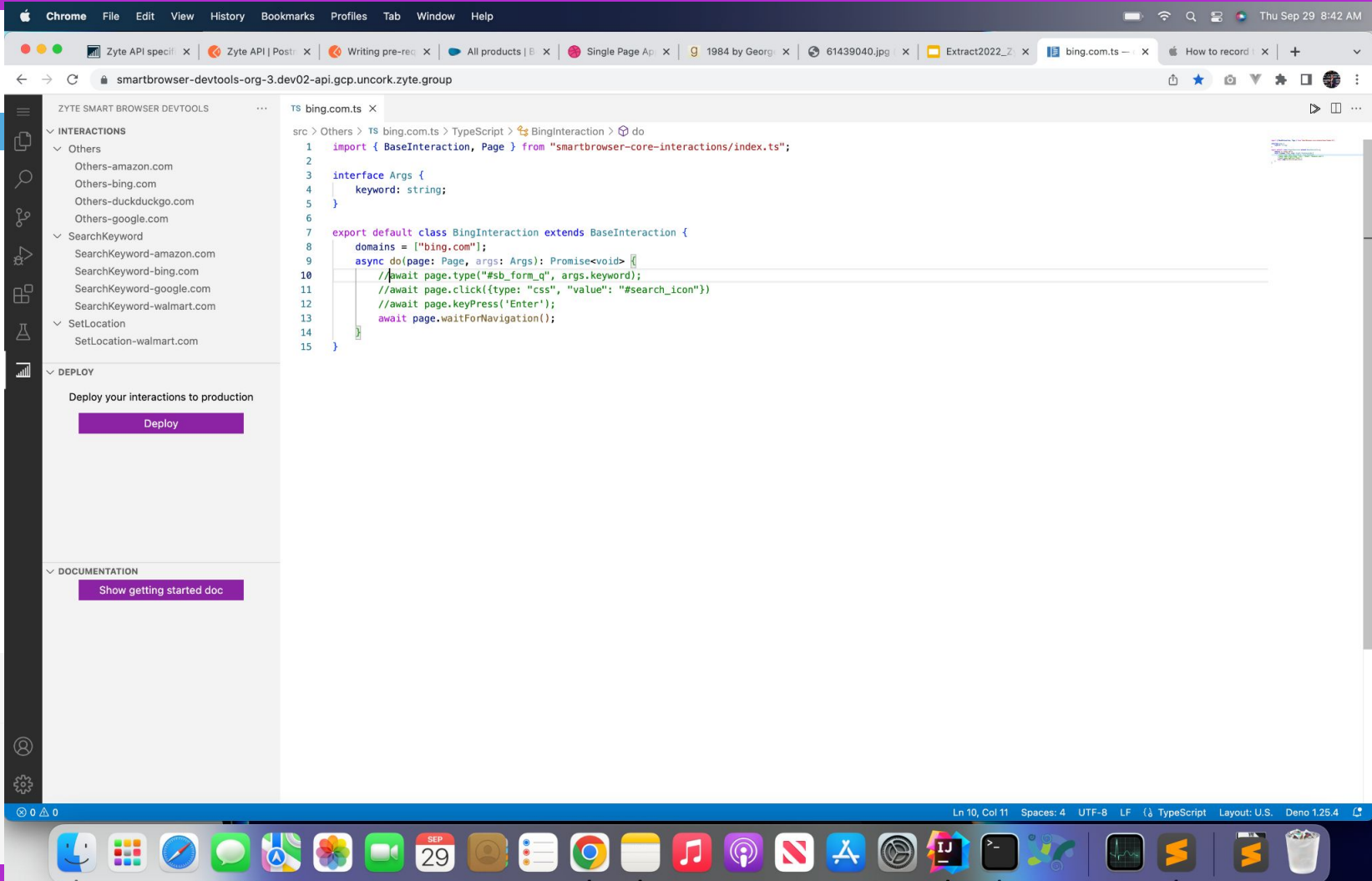
```
1 class UltraLocation implements Interaction {
2   domains = ["ulta.com"];
3   args = ["zipcode"];
4
5   async do(page: Page, zipcode: string) {
6     await page.waitForSelector("#onetrust-accept-btn-handler");
7     await page.click(cookie_sel);
8     await page.click(".ProductDetail__findInStore_Button button");
9     await page.waitForSelector("#searchField");
10    await page.waitForTimeout(50);
11    await page.type("#searchField", zipcode);
12    await page.keyboard.press("Enter");
13    await page.waitForSelector(".StoreList");
14  }
15 }
16
17 module.exports = UltraLocation;
18
```

2. Invoke script through REST API

```
{
  "url": "http://www.ulta.com/store",
  "actions": [
    {
      "action": "setLocation",
      "address": {
        "postalCode": "07040"
      }
    }
  ]
}
```

- Code envi





src > Others > TS bing.com.ts > TypeScript > BingInteraction > do

```
1 import { BaseInteraction, Page } from "smartbrowser-core-interaction"
```

```
2
3 interface Args {
4     keyword: string;
5 }
6
```

```
7 export default class BingInteraction extends BaseInteraction {
```

```
8 domains = ["bing.com"];
9 async do(page: Page, args: Promise<void> {
10     //await page.type("#sb_form_q", args.keyword);
11     //await page.click({type: "css", "value": "#search_icon"})
12     //await page.keyPress('Enter');
```

```
13      await page.waitForNavigation();
```

14	f
15	j

```
src > Others > TS bing.com.ts > TypeScript > BingInteraction > do
1  import { BaseInteraction, Page } from "smartbrowser-core-interaction"
2
3  interface Args {
4      keyword: string;
5  }
6
7  export default class BingInteraction extends BaseInteraction {
8      domains = ["bing.com"];
9      async do(page: Page, args: Args): Promise<void> {
10         //await page.type("#sb_form_q", args.keyword);
11         //await page.click(type: "css", "value": "#search_icon")}
12         //await page.keyPress('Enter');
13         await page.waitForNavigation();
14     }
15 }
```

Request Details


```
<!DOCTYPE html>
<html lang="en" dir="ltr">
<head></head>
<body>
  <div id="ajaxStyles"></div>
  <div id="bnp.nid.63245" class="" data-viewname="BottomBannerNotIt
le" data-vertical="serp" style="display: block;" data-bm="42"></div>
  <div class="happq"></div>
  <div class="hide" id="happq_id" data-priority"></div>
  <script type="importmap" nonce="oD8nJWldZjBBpBBSa8Cpk1LeuvZ0uf50w7
3o1CHnC4"></script>
  <script type="text/javascript" nonce="oD8nJWldZjBBpBBSa8Cpk1LeuvZ0u
f50w73o1CHnC4"></script>
  <script type="text/javascript" nonce="oD8nJWldZjBBpBBSa8Cpk1LeuvZ0u
f50w73o1CHnC4"></script>
  <div id="armsDefer"></div>
  <script data-rs="1" crossorigin="anonymous" src="//r/SFAwGSc8f1gn
KVIME6-rlx1-sA-br.js" type="text/javascript"></script>
  <script type="text/javascript" data-rs="1"></script>
  <script type="text/javascript" data-rs="1" src="https://c.bing.co
n/rp/IT7IG-ckvErVntDf1he3jC7zSw_br.js" crossorigin="anonymous">
</script>
```

Filter: show all

Filter: :nov .cls T,

No matching selector or style

```
html body
```

Layout: U.S. Deno 1.25.4

Chrome File Edit View History Bookmarks Profiles Tab Window Help

Thu Sep 29 8:43 AM

Yzite API specifi x Yzite API | Postri x Writing pre-re x All products | B x Single Page Api x 1984 by Georgi x 61439040.jpg x Extract2022_Z x Smart Browser x How to record x +

smartbrowser-devtools-org-3.dev02-api.gcp.uncork.zyte.group

ZYTE SMART BROWSER DEVTOOLS

INTERACTIONS

- Others
 - Others-amazon.com
 - Others-bing.com
 - Others-duckduckgo.com
 - Others-google.com
- SearchKeyword
 - SearchKeyword-amazon.com
 - SearchKeyword-bing.com
 - SearchKeyword-google.com
 - SearchKeyword-walmart.com
- SetLocation
 - SetLocation-walmart.com

DEPLOY

Deploy your interactions to production

Deploy

DOCUMENTATION

Show getting started doc

TS bing.com.ts x

```
src > Others > TS bing.com.ts > TypeScript > BingInteraction > do
1 import { BaseInteraction, Page } from "smartbrowser-core-interacti
2
3 interface Args {
4   keyword: string;
5 }
6
7 export default class BingInteraction extends BaseInteraction {
8   domains = ["bing.com"];
9   async do(page: Page, args: Args): Promise<void> {
10     //await page.type("#sb_form_q", args.keyword);
11     //await page.click({type: "css", "value": "#search_icon"})
12     //await page.keyPress('Enter');
13     await page.waitForNavigation();
14   }
15 }
```

Smart Browser DevTools x

Browser Tools Request Details

Microsoft Bing Images Português Sign In Rewards

A fina linha entre arte e moda

Senado dos EUA aprova r... "Quem vazou foi você, sej... 'Bolsomoro' racha União ... Cor

Elements Network Performance Application Lighthouse

Filter Hide data URLs XHR JS CSS Img Media Font Doc WS Manifest Other

Has blocked cookies Blocked Requests

2000 ms 4000 ms 6000 ms 8000 ms 10000 ms 12000 ms 14000 ms 16000 ms 18000 ms

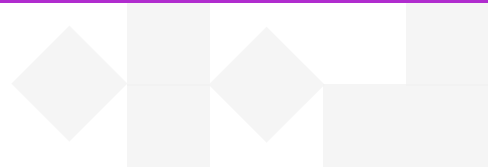
Name	Status	Type	Initiator	Size	Time	Waterfall
BlueIdentityDropdownRedir...	200	script		3.9 kB	420 ms	
HamburgerServicesHeaderF...	200	script		4.0 kB	458 ms	
Dropdown?n=1&IID=SERPS...	200	xhr		1.5 kB	543 ms	
scfo?ver=31320772&IID=SE...	200	xhr		2.5 kB	538 ms	
lsp.aspx	204	xhr		117 B	436 ms	
9roWR2D5ePLJmZD9tbaESv...	200	png	Other	3.5 kB	323 ms	
lsp.aspx	204	xhr		117 B	465 ms	
lsp.aspx	204	xhr		117 B	437 ms	

54 requests 76.7 kB transferred 235 kB resources

Layout: U.S. Demo 1.25.4



Zyte IDE



▪



Zyte API Page Actions & Scripting

Get going instantly - no need for installations or integrations

Save time with a rich library of ready-made interactions

Stronger ban avoidance

Fully scaled on demand

Exposed as an API

Coming in October



Solving the most painful site ban problem

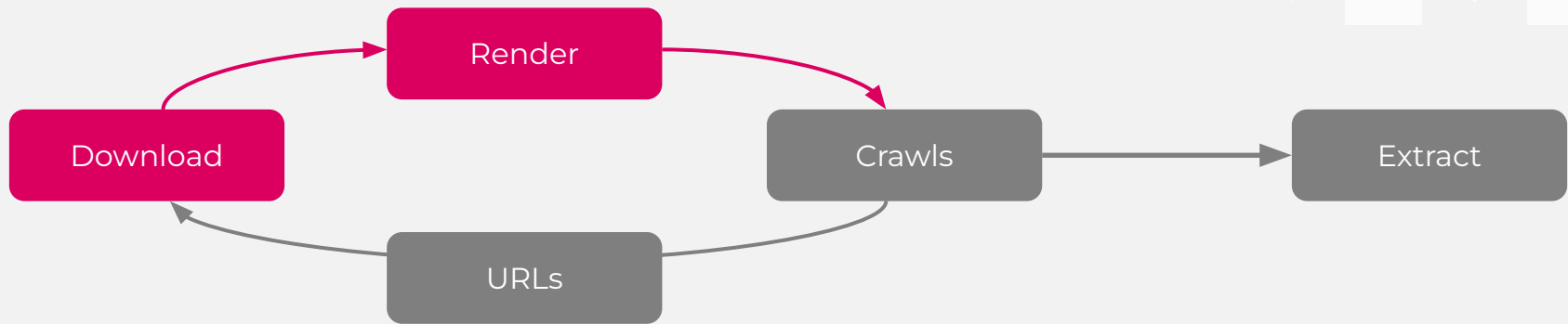
The browser designed for web data extraction

Solving two scaling challenges to faster growth

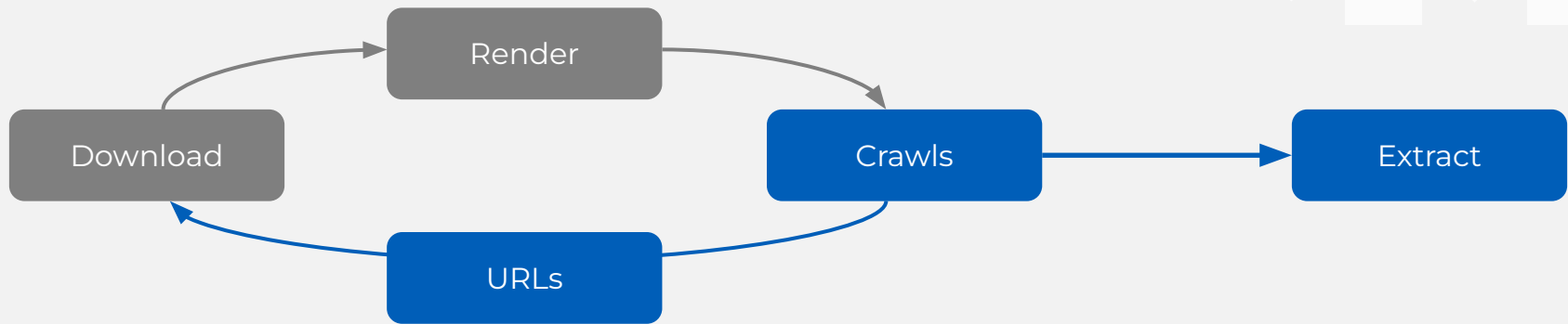


A preview of Zyte roadmap



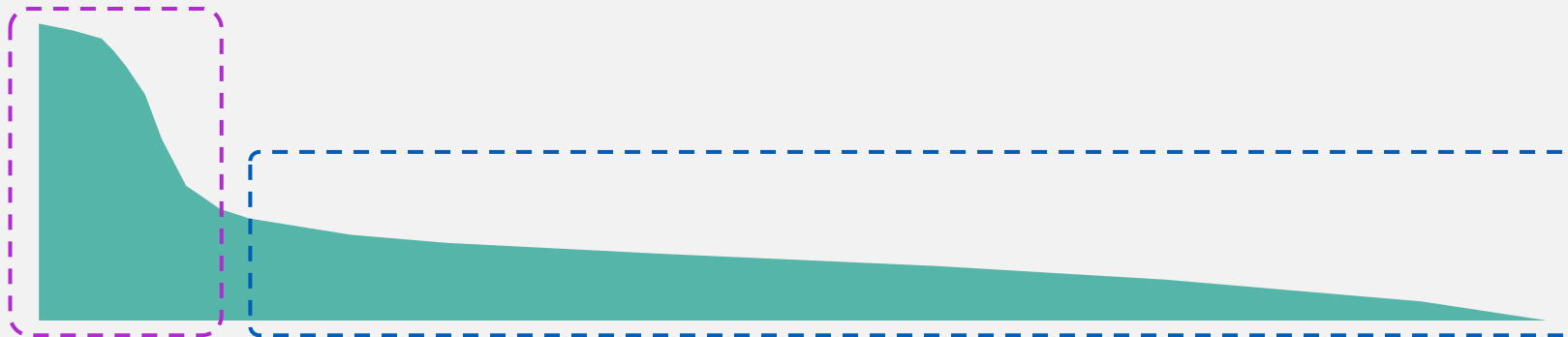


If I could just make site bans go away



We're good, but my CEO always wants us to scale faster

Two types of site, two types of challenge



High volume sites

Millions of requests

- Must be **fast**
- Must be **cheap** per request

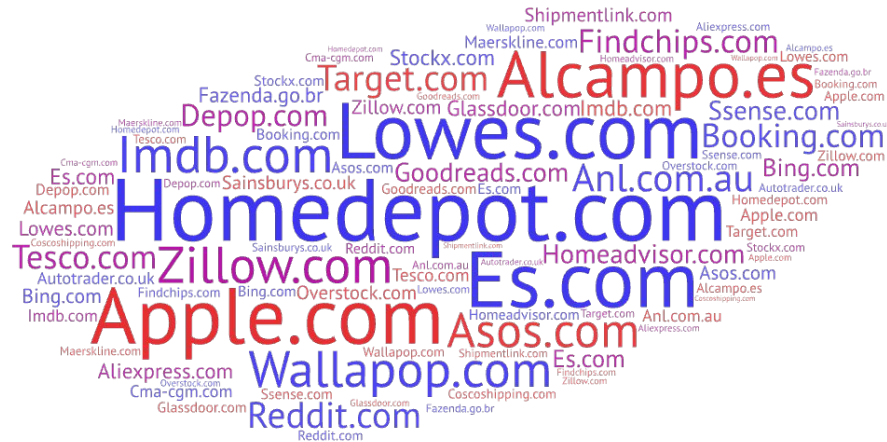
Long tail sites

Hundreds or thousands of sites

- Must have **simple set-up** and **maintenance**

1

1

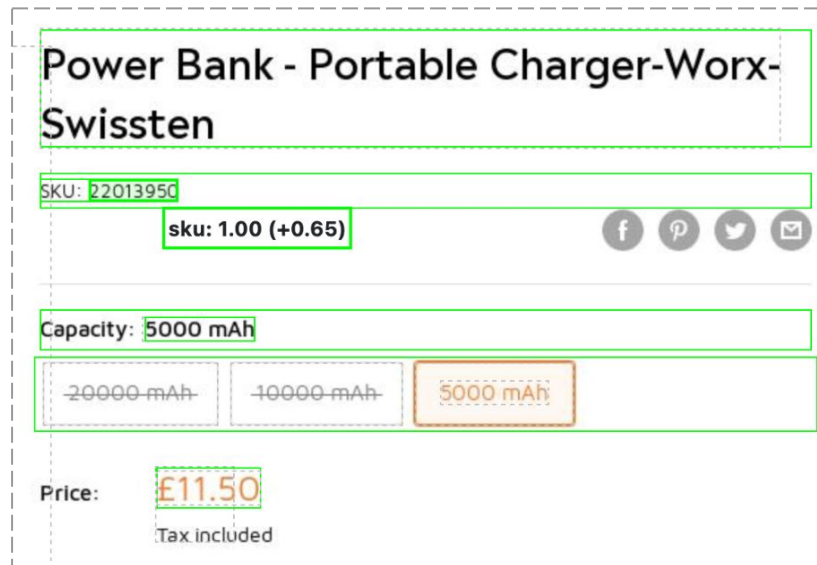


Long tail sites : powered by machine learning

Zyte's patented AutoExtract technologies integrated with Zyte API

ML models for all the major web-based data types - products, jobs, articles etc

Instant data with zero set-up or maintenance



Integrated Extract & Crawl

with maintained spiders and ML

On-demand
fallback for
your spiders
for instant
break recovery

No need for a
contract, no
minimum
monthly usage

Option to
expand team
capacity to
extend to
more sites

Great option
for
prototyping

Coming next year



Solving the most painful site ban problem

The browser designed for web data extraction

Solving two scaling challenges to faster growth

Coming October 27th





Thank you