



# Surviving the Swarm: Bot and Crawler Protection for Drupal Sites



2026.03.13



INTRODUCTIONS

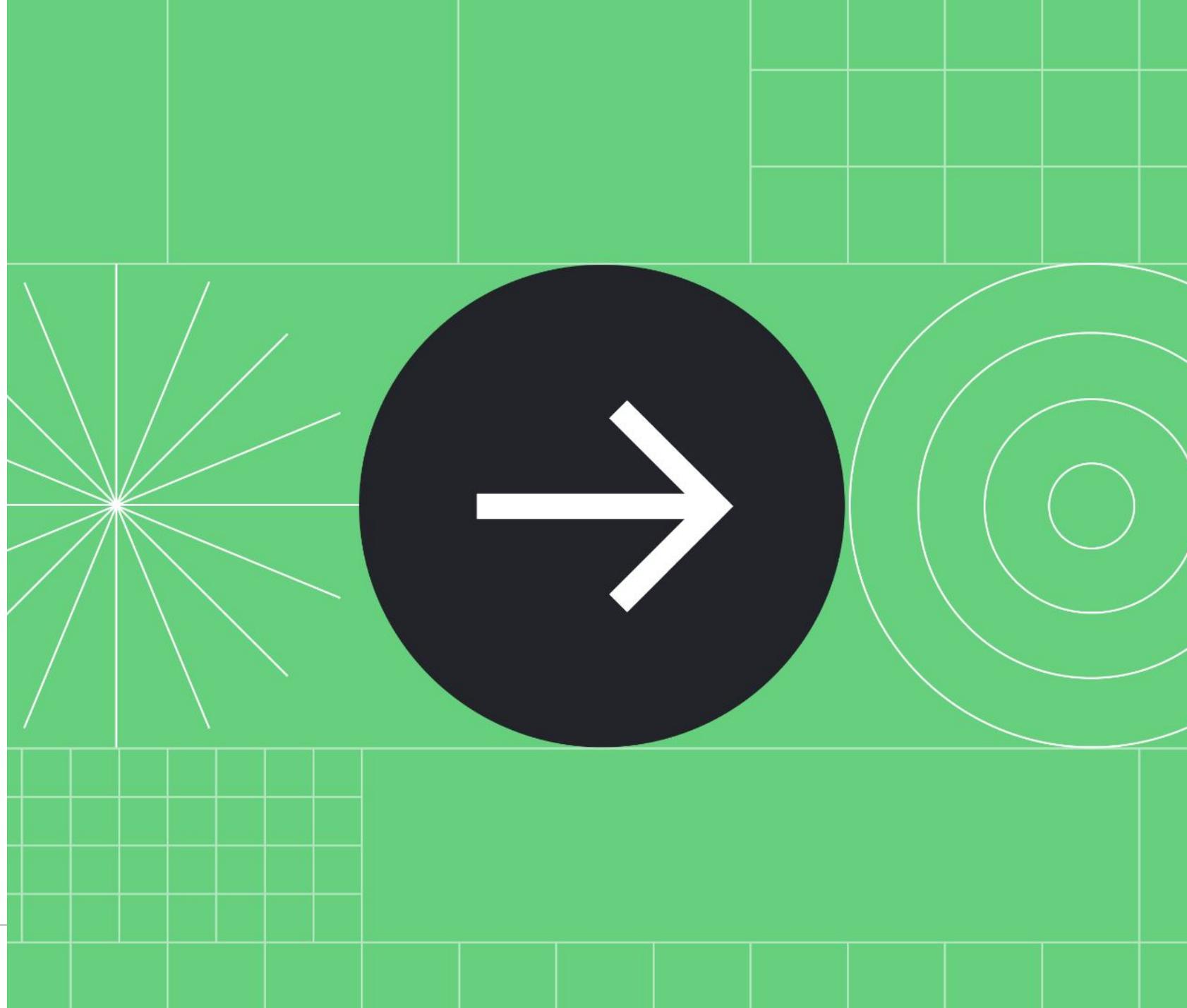
# Rich Lawson



Associate Director of Technology,  
Evolving Web

 /in/rich-lawson

[evolvingweb.com](http://evolvingweb.com)



Higher Ed



Healthcare



Culture & Tourism



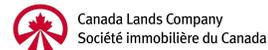
Transit & Infra



Financial



Government



Our clients exist to make an impact.

We exist to help them do just that.

- A top 10 Drupal contributor
- Diamond Certified Drupal Partner
- Part of the Drupal CMS leadership team
- Pantheon's Partner of the Year in 2025
- 5,000+ Drupalists trained since 2011
- Organizers of the [EvolveDigital](#) event series



New York City | September 2026

# See you in NYC?

Get 35% off tickets:  
NJPromo35





New York City | September 2026

# See you in NYC?

Get 35% off tickets:  
NJPromo35



# Agenda

- 01.** When the swarm hits
- 02.** The current bot landscape
- 03.** What happens to Drupal?
- 04.** Layered defense strategy
- 05.** Case study
- 06.** Actionable playbook
- 07.** Questions

01

# When the swarm hits



# What do you do?



Editors encounter 503 errors when trying to update their landing pages



Page loading slows to a crawl, until finally the pages won't load at all



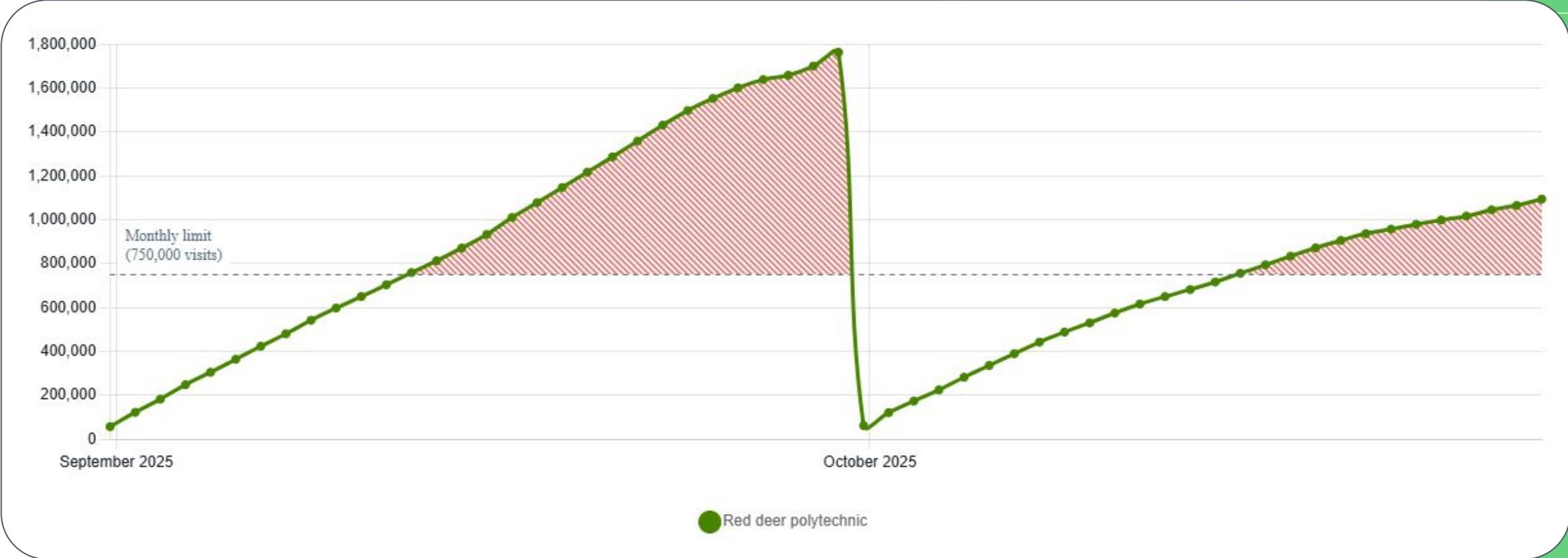
Analytics show a traffic spike of 800%

# Some of our affected clients

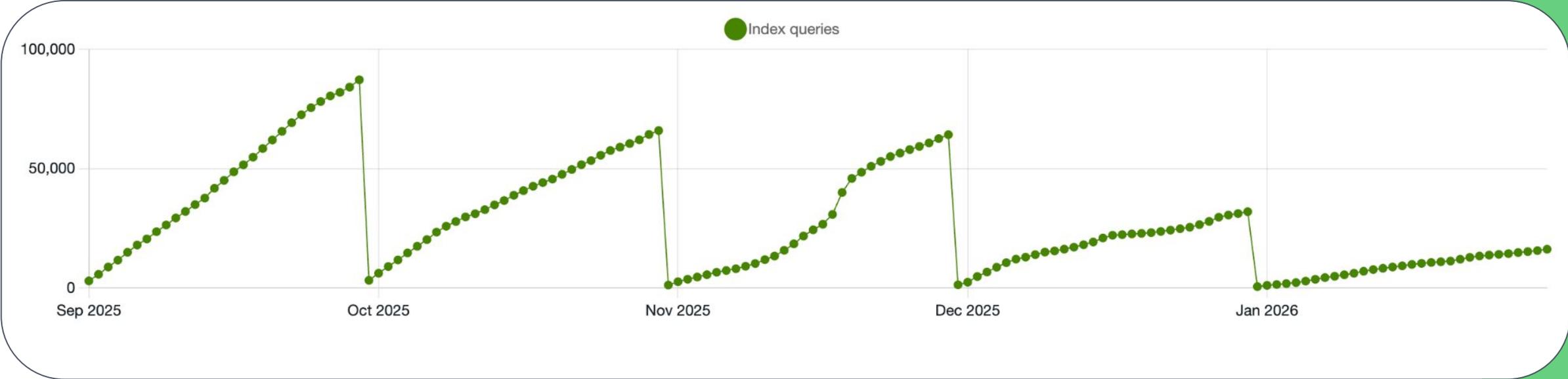
Client	Brief description	Timeline
<i>Sollio</i>	Large agri-food cooperative	Started in April, resolved in December
<i>INSPQ</i>	Government public health institute	Started in October, ongoing
<i>Canada Lands Company</i>	Canadian government Crown corporation	Started in September, resolved in November
<i>Red Deer Polytechnic</i>	Public educational institute	Started in July, resolved in September...or was it?

WHEN THE SWARM HITS

# Traffic growth

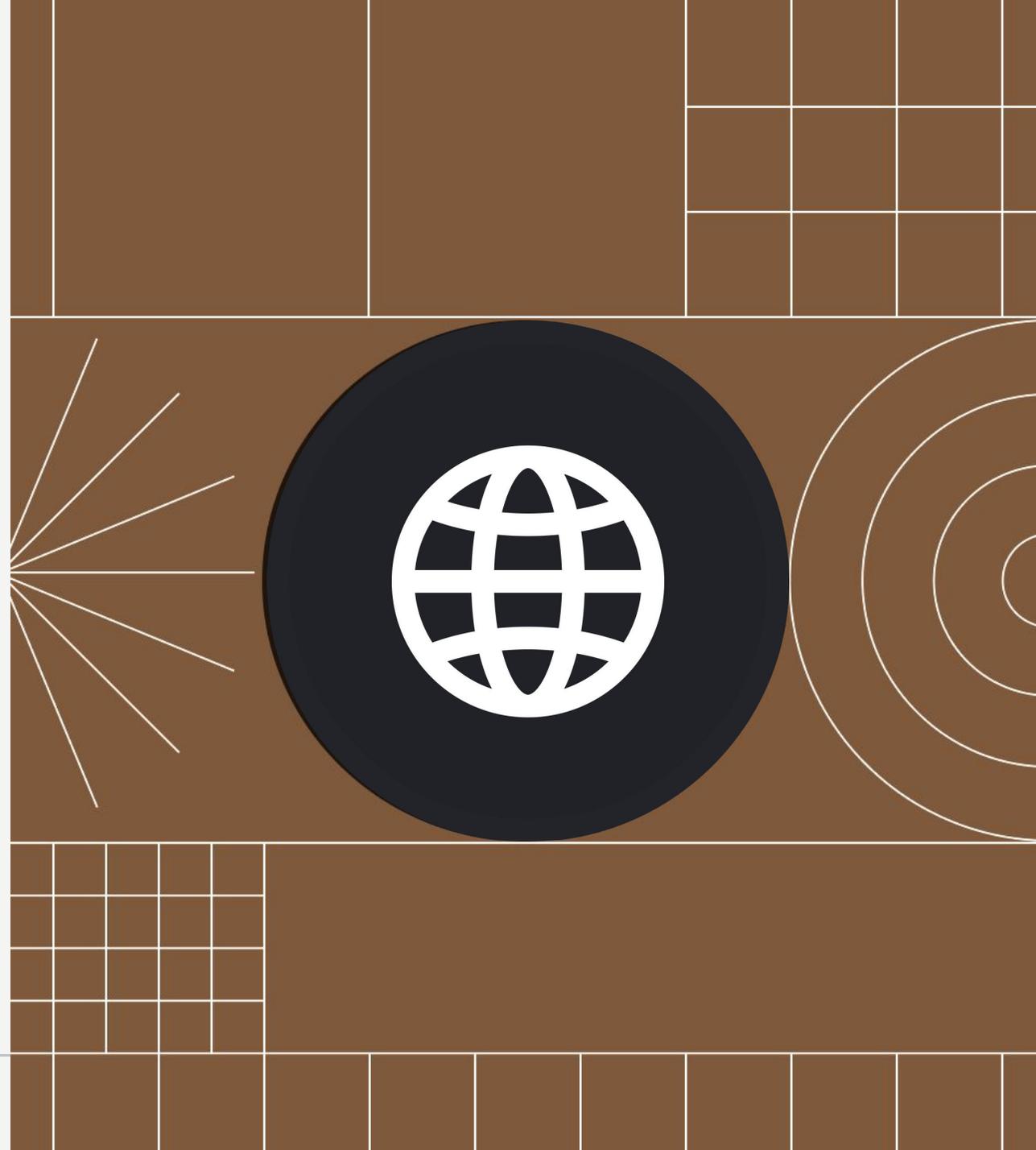


# Search index mitigation



02

# The current bot landscape



# It's not just Google anymore...



# Types of bots

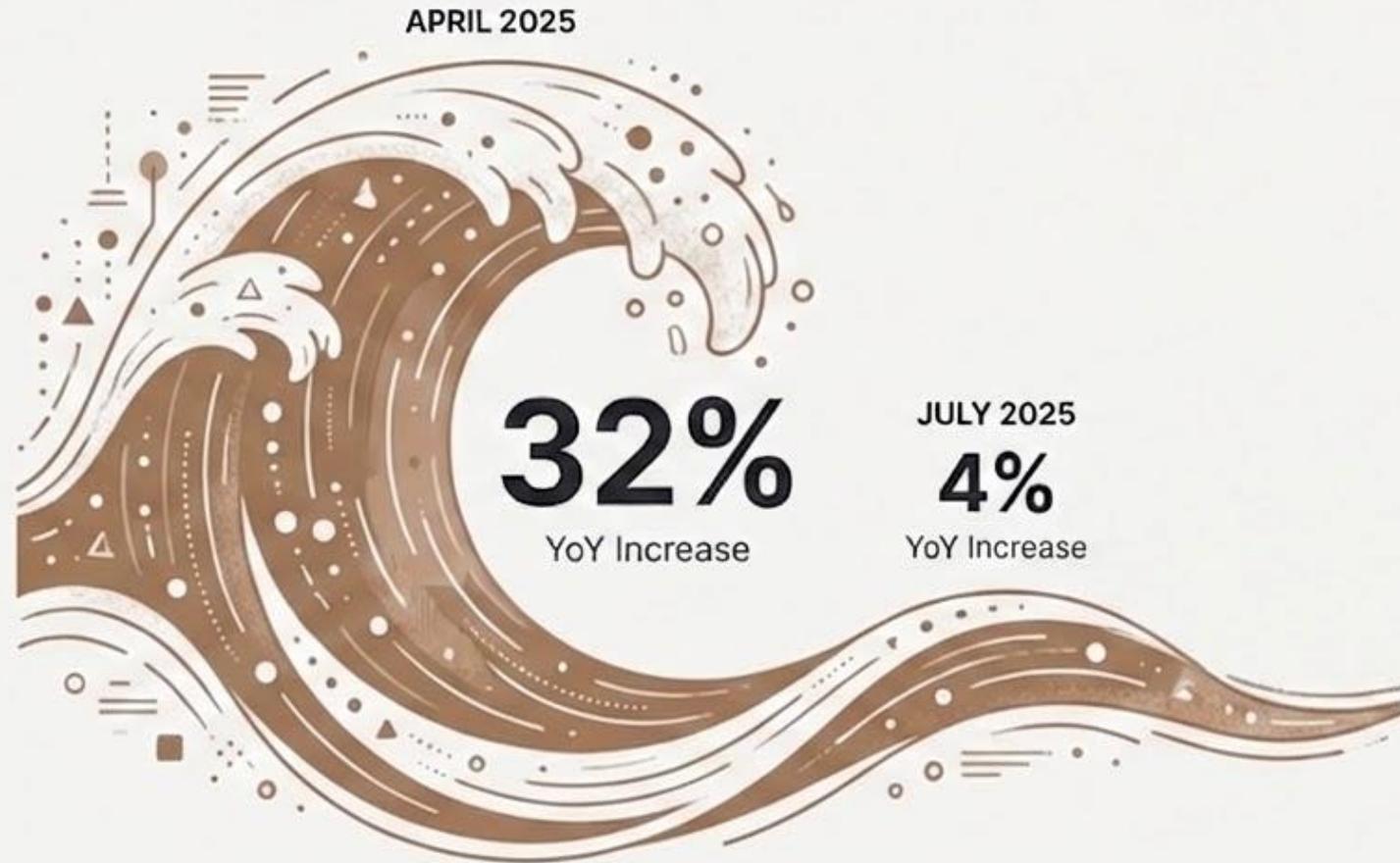
Type	Description
Web crawlers	Scan web pages, typically for indexing search engines (e.g., Googlebot, Bing)
AI crawlers	Ingest, index, and extract large volumes of data for LLMs (e.g., OpenAI, Anthropic, Perplexity, etc.)
Malicious bots & mystery scrapers	Disregard policies (e.g., robots.txt) and perform negative behavior: <ul style="list-style-type: none"><li>■ Web/content scraping</li><li>■ Denial of Service (DoS)</li><li>■ Distributed Denial of Service (DDoS) attacks</li></ul>

# Majority of traffic from bots

**In 2025, 51% of traffic was automated**

Source: <https://www.imperva.com/resources/resource-library/reports/2025-bad-bot-report/>

# Scope of AI bots



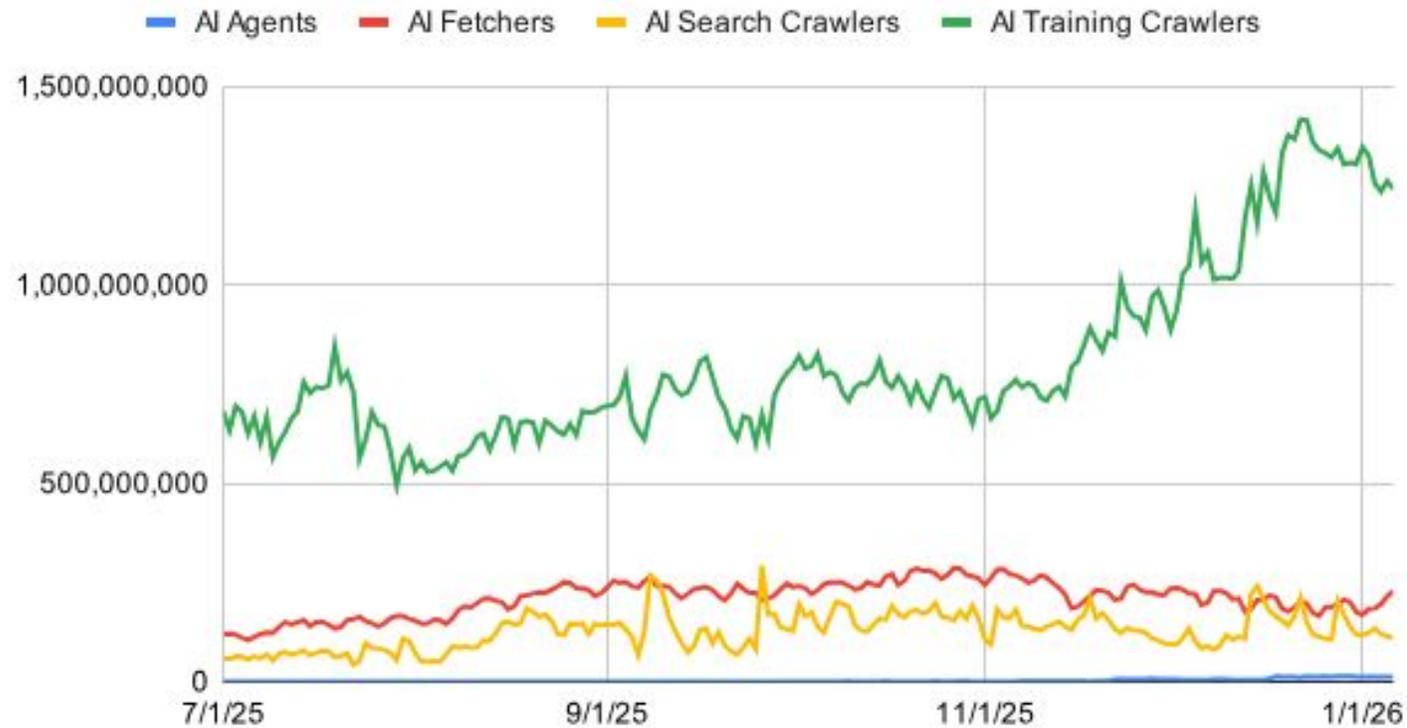
Data source: <https://blog.cloudflare.com/crawlers-click-ai-bots-training/>

Huge surge in traffic in April 2025

Continue to see large growth

# Scope of AI bots

Total AI bot traffic growth by AI bot type



Source: <https://www.akamai.com/blog/security/ai-pulse-how-ai-bots-agents-will-shape-2026>

# Why should we care about bots?

## Situation

- Typically crawl full sites at large/massive scale
- Disregard crawl-delays and may not acknowledge other measures
- Highly repetitive crawling

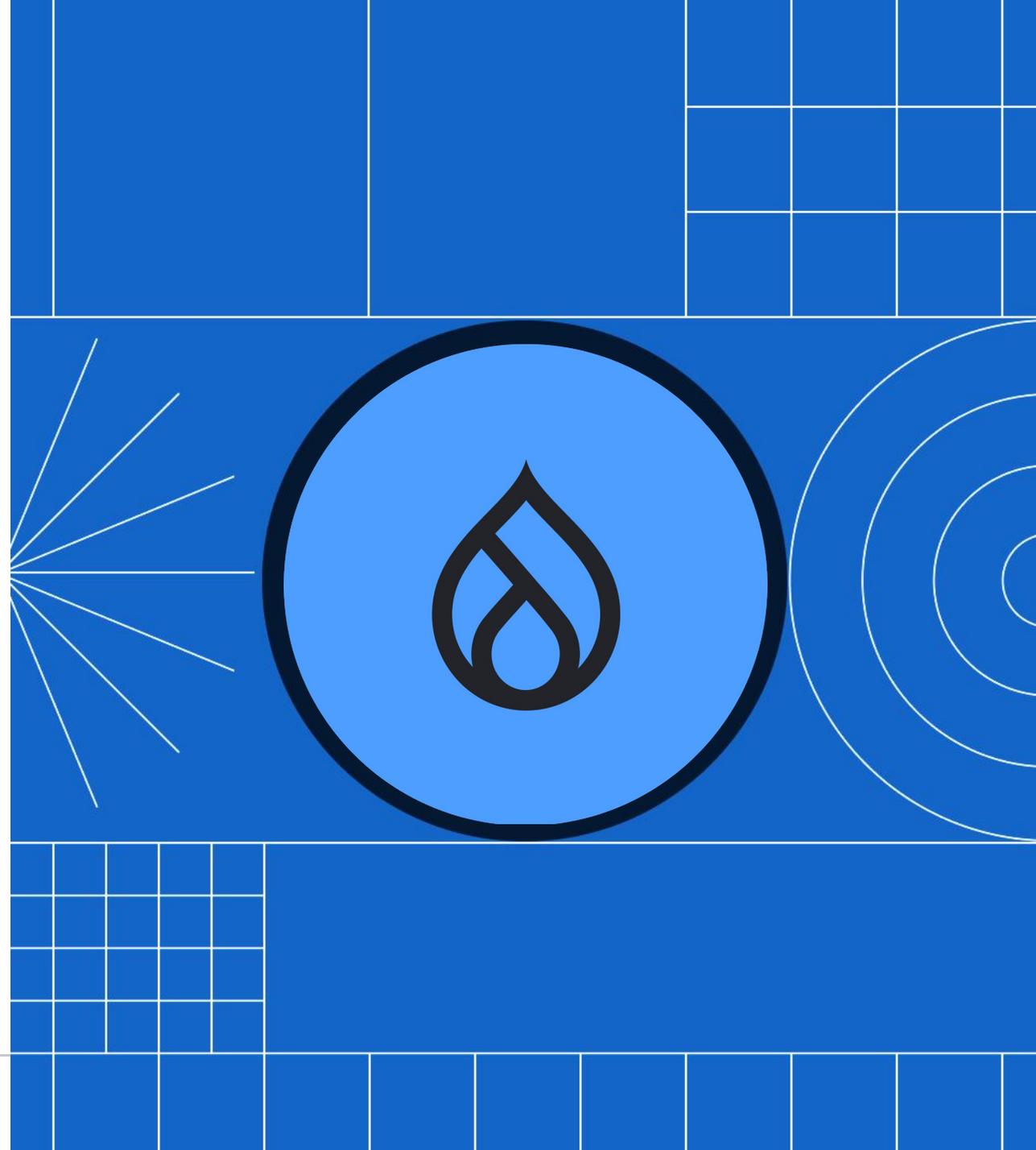
## Impact

- Increased server load
- Bandwidth and other costs
- Slower performance for real users
- Content scraping at scale

```
1 # Main robots.txt file for example.com
2 User-agent: *
3 Disallow: /private/
4 Disallow: /temp/
5 Crawl-delay: 10
6
7 # Googlebot specific rules
8 User-agent: Googlebot
9 Disallow: /no-google/
10 Allow: /public/
11
12 # Bingbot rules
13 User-agent: Bingbot
14 Disallow: /no-bing/
15 Crawl-delay: 5
16
17 Sitemap: https://www.example.com/sitemap.xml
```

03

# What happens to Drupal?

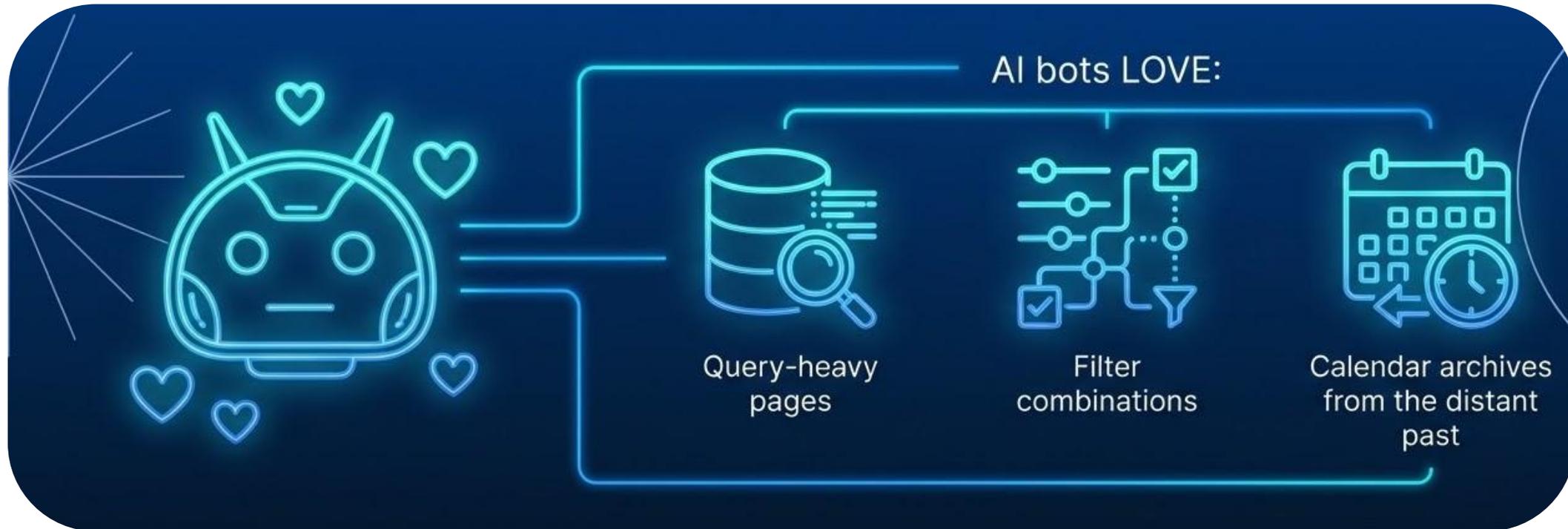


WHAT HAPPENS TO DRUPAL?



# Drupal: There's a lot (for bots) to love

Drupal sites have search, filters, Views, and more.



# What's the impact on your Drupal site?

A lot depends on any mitigations that you have in place and your hosting infrastructure.

- Without caching, your site may go down.
- Even with caching, the bots may target pages that hit the Drupal backend and negatively affect performance, even for unauthenticated traffic

Hosting on Pantheon?

- Monthly visits cap

Hosting on Acquia?

- Views and Visits usage limits and overages
- Solr search on Acquia Search and SearchStax limits

# Measure analytics and performance

Review your analytics and site performance.

If you're under active attack, you may want to focus on making some of the adjustments that we'll talk about soon, particularly at the edge.

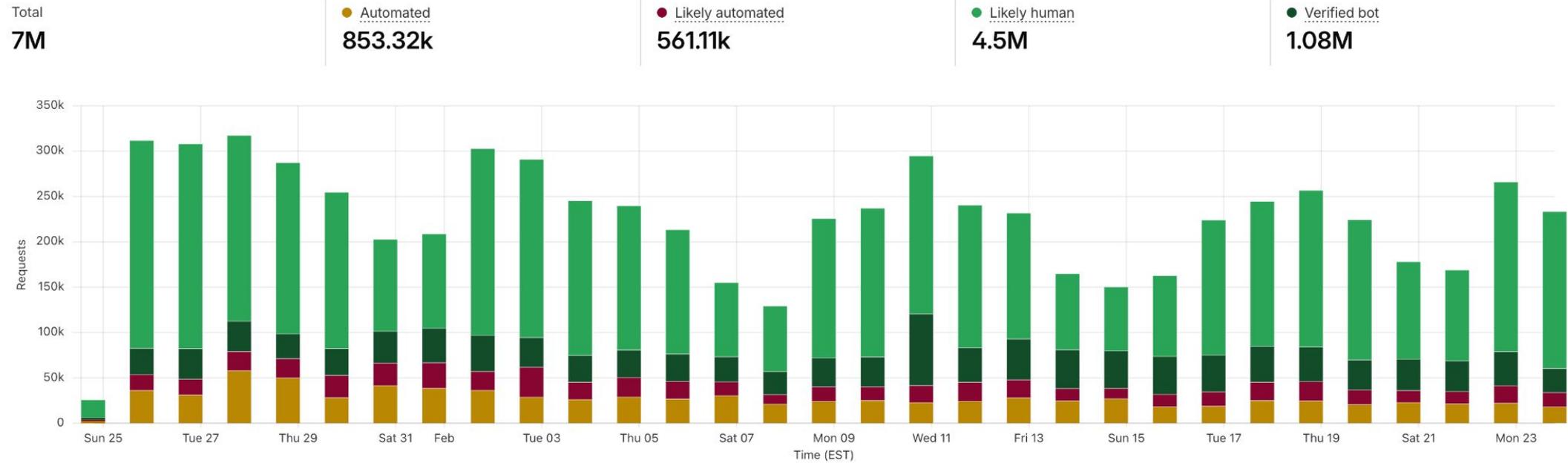
1. Measure bot vs. human ("real") traffic to provide a baseline
2. Monitor over time and review after making adjustments or as you're working to mitigate an attack

# Most AI crawled paths

Path	Host	Content type	↓ Allowed requests
<a href="#">/programs/find-program</a>	rdpolytech.ca	HTML	12,262
<a href="#">/find-program</a>	rdpolytech.ca	HTML	11,403
<a href="#">/sitemap.xml</a>	rdpolytech.ca	XML	2,829
<a href="#">/</a>	rdpolytech.ca	HTML	1,732
<a href="#">/log</a>	rdpolytech.ca	Empty	490
<a href="#">/programs/practical-nurse-diploma</a>	rdpolytech.ca	HTML	248
<a href="#">/</a>	media.rdpolytech.ca	HTML	233
<a href="#">/programs/licensed-practical-nurse-lpn-bsc-nursing-bscn-transition-program</a>	rdpolytech.ca	HTML	233
<a href="#">/programs/health-care-aide-certificate</a>	rdpolytech.ca	HTML	233
<a href="#">/about-us/services-departments/office-registrar/course-outlines</a>	rdpolytech.ca	HTML	215

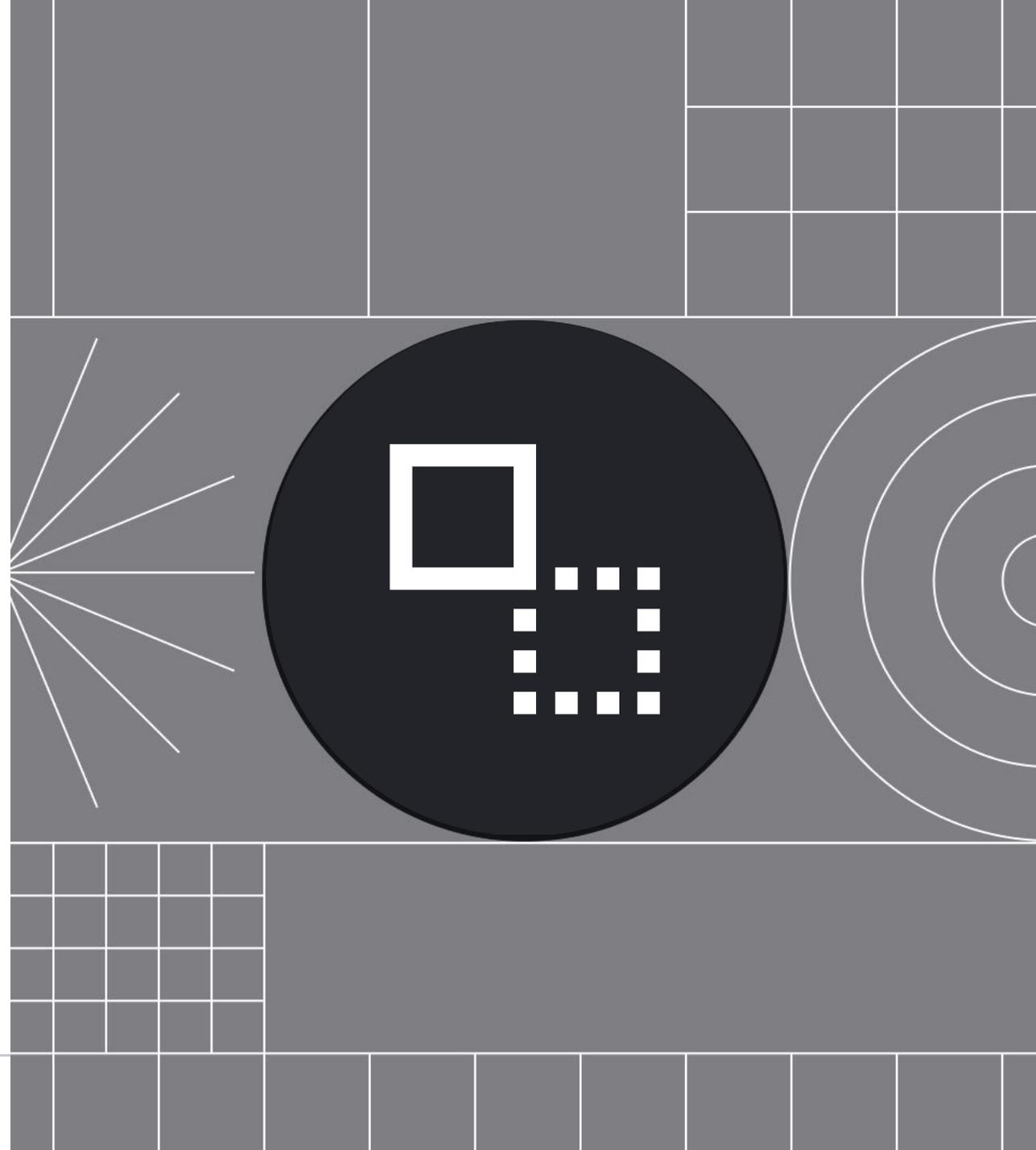
1 - 10 of 4916 items Page 1 ▼ of 492 < >

# Number of requests made by bots and humans

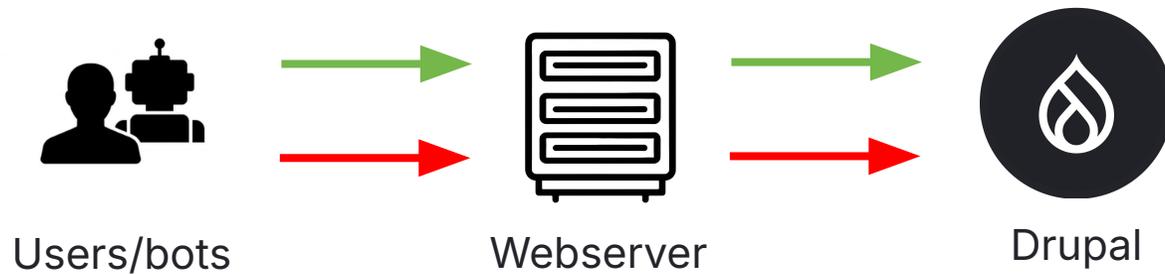


04

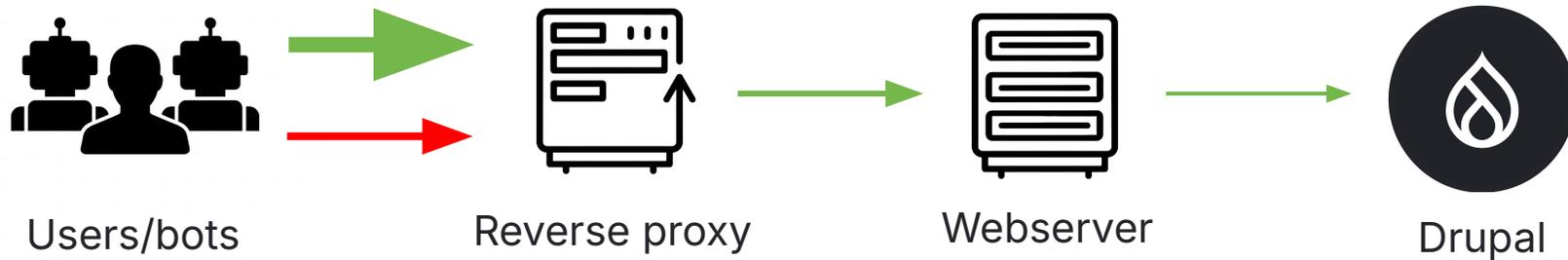
# Layered defense strategy



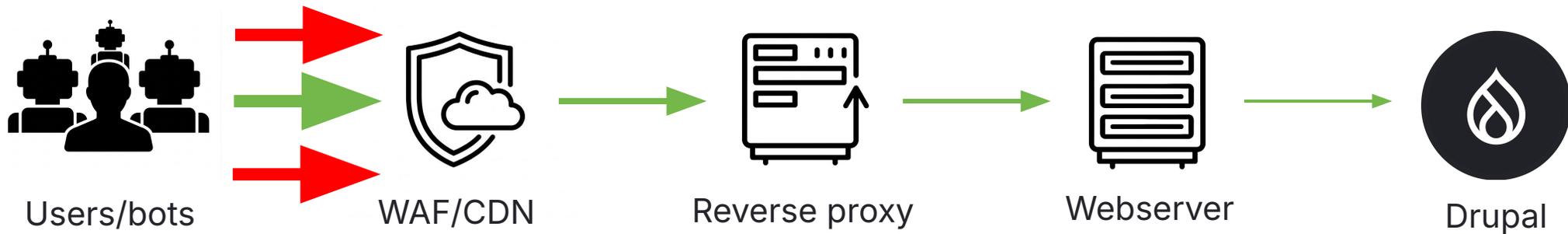
# The traffic flow: Minimal infrastructure



# The traffic flow: Reverse proxy



# The traffic flow: WAF/CDN



# WAFs and CDNs to the rescue!

A Web Application Firewall (WAF) and Content Delivery Network (CDN) protect your Drupal site at the edge:

- Cloudflare
- Fastly
- Sucuri

Drupal-specific examples:

- Acquia Edge
- Pantheon Advanced Global CDN



# WAF/CDN benefits

- Serve or stop traffic before it hits Drupal or the origin infrastructure
- Global Distributed Denial of Services (DDoS) absorption
- Edge-based rate limiting
- Bot detection scoring
- Country-based rules
- Challenge process
- Latency reduction for legitimate users
- Behavioral/fingerprint-based detection
- Request anomaly detection
- Credential stuffing protections
- IP reputation databases (limited benefit)

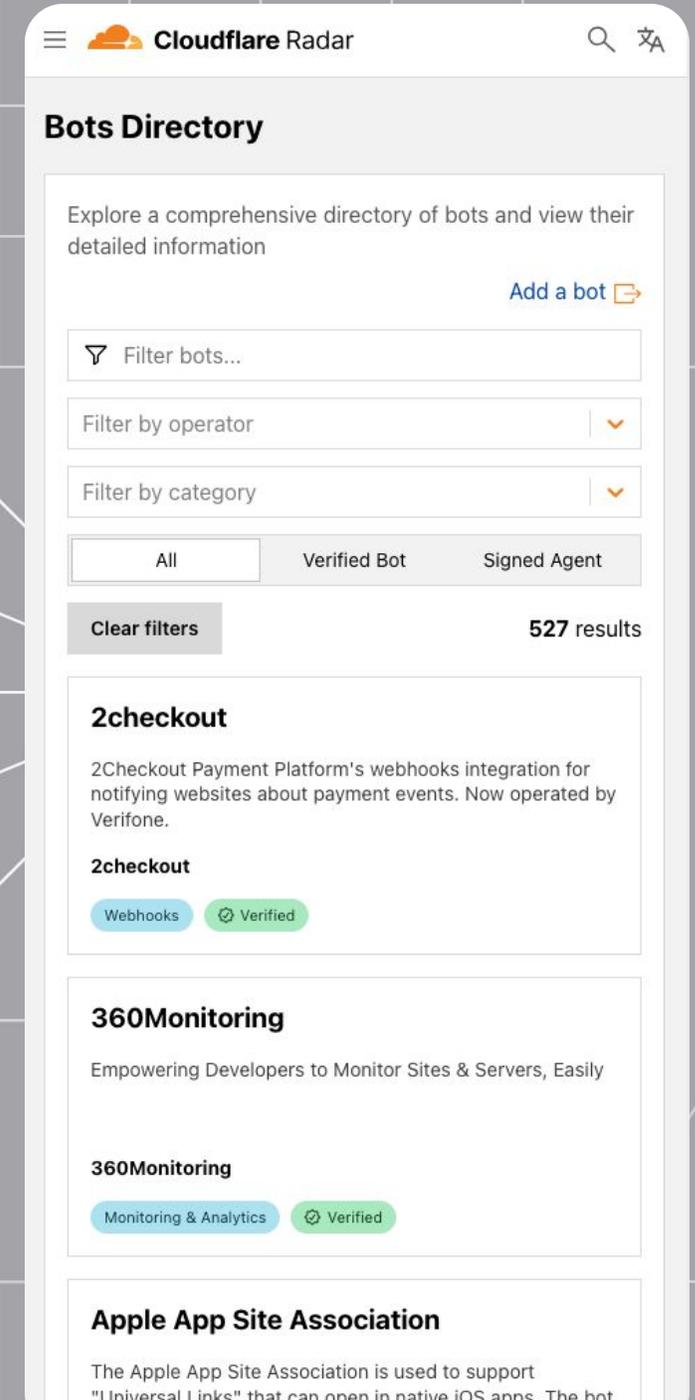
# Additional benefits

- Proper implementation has a positive impact on performance for all users
- Better user experience with global distribution



# Bot management & bot scores

- Allow verified bots (e.g., Google)
- Challenge or block suspicious bots
- Create WAF rules based on bot scores
- [Bots Directory](#) to learn more information about individual bots



# Targeting AI crawlers: Block or challenge

- Block specific AI crawlers
- Challenge AI crawlers (e.g., CAPTCHA, Managed Challenge)
- Rate limit them

website.com

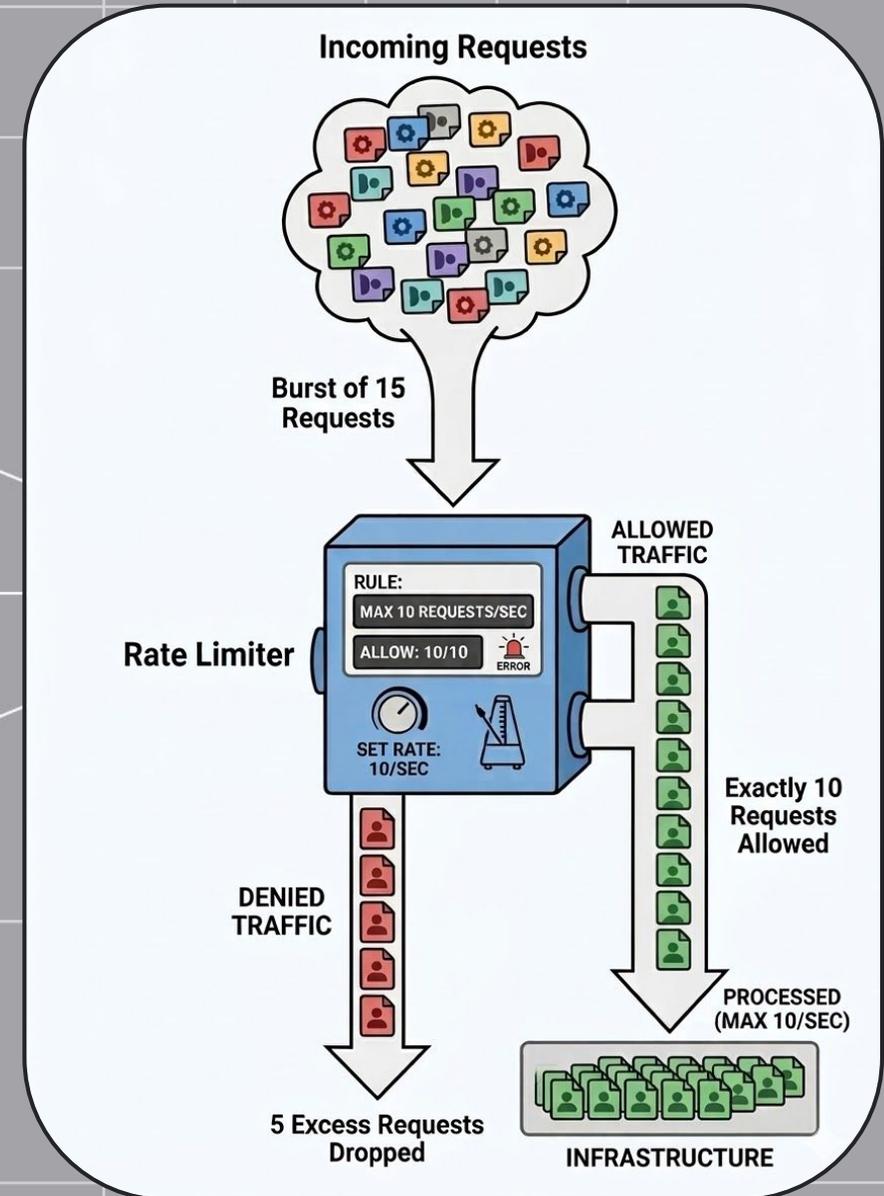
Verify you are human by completing the action below.

Verify you are human  CLOUDFLARE  
Privacy · Terms

website.com needs to review the security of your connection before proceeding.

# Rate limiting

- AI bots tend to hit many pages very quickly
- X requests per Y seconds (can be per IP address, path, etc.)
- Strictly rate limit requests on key pages (e.g., /search)



# Rules for high-risk paths

Create WAF rules for:

- /search
- Filter-heavy views
- Other areas that can be targeted with query strings

Challenge non-verified bots on dynamic endpoints

Block suspicious query string patterns



# Strategically cache everything

The best requests are those that never reach Drupal

The CDN cache stores copies of frequently accessed content like images, videos, or entire pages, geographically distributing it across its data centers, which are closer to your end users.

This reduces server load and improves performance.



# To block or not to block...

Should AI be blocked completely?

**...it depends**

# Other infrastructure

- Firewall (e.g., iptables)
- Reverse proxy (e.g., Varnish)
- .htaccess for Apache, configuration for Nginx

# Facets

- Facets are "smart filters" used in search and catalog pages (common in Drupal via the Facets and Search API modules).
- They allow users to narrow down results by attributes like Category, Price, Color, or Date.
- In Drupal, each facet selection usually generates a unique URL.
  - [example.com/shop](http://example.com/shop)
  - [example.com/shop?f\[0\]=color:blue](http://example.com/shop?f[0]=color:blue)
  - [example.com/shop?f\[0\]=color:blue&f\[1\]=size:large](http://example.com/shop?f[0]=color:blue&f[1]=size:large)

- 
- Backyard poultry
  - Beef
  - Crop protection
  - Dairy
  - Digital agriculture
  - Equine
  - Fertilizer
  - Goat
  - Grain
  - Poultry
  - Seed
  - Sheep
  - Specialty livestock
  - Sustainable agriculture
  - Swine
  - Talents
  - Urban chickens

# Facets: Exponential explosion

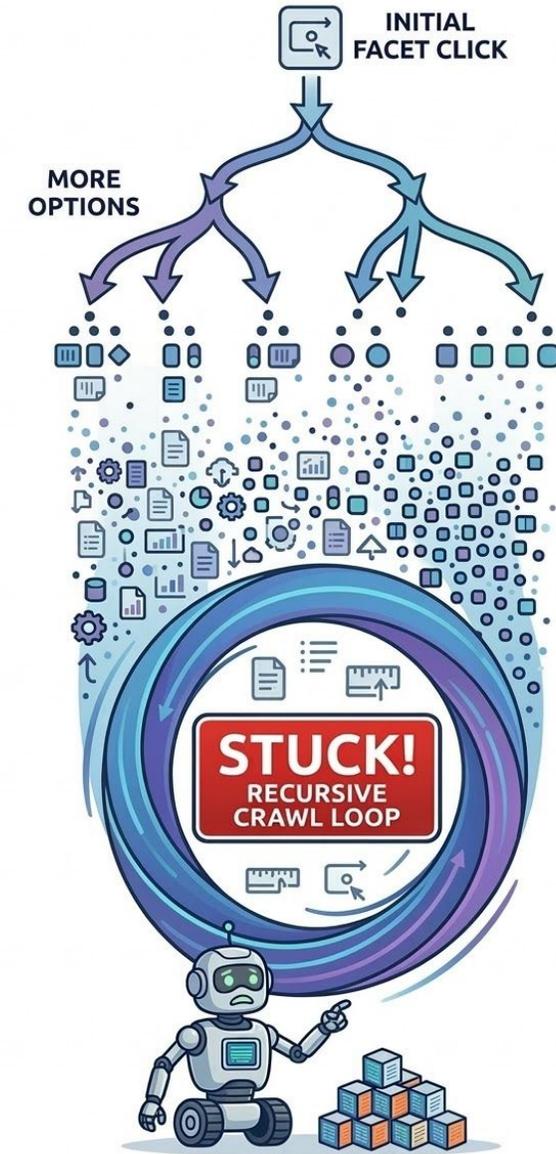
Total facets (n)	Possible URL combinations ( $2^n$ )	Real-world context
10	1,024	A very small blog with a few tags.
20	1,048,576	A standard e-commerce site category (e.g., shoes).
30	1,073,741,824	<b>Over 1 Billion URLs.</b>
40	1,099,511,627,776	More pages than the entire Google Web index.
58	288,230,376,151,711,744	Total unique search combinations on a client's site!

# Why do AI bots get stuck in facets?

- Traditional bots (like Googlebot) have “crawl budgets” and some level of facet awareness.
- Newer AI training bots are more aggressive and want all the data.
- The bot enters the facet maze, and every time it follows a link, it finds more links, which is perceived as new content. It continues digging.
- The bot’s queue of “URLs to visit” grows faster than it can process them, trapping it in a deeply recursive loop.

# The accidental DDoS

The AI bot hits thousands of these complex URLs per minute, spiking the CPU, locking up the database, and bringing the site down for real users.



# Additional Drupal defenses

[Drupal.org: Block bad bots and crawlers](#) (documentation)

## Modules

[Crawler Rate Limit](#)

[IP Limiter](#)

[Facet Bot Blocker](#)

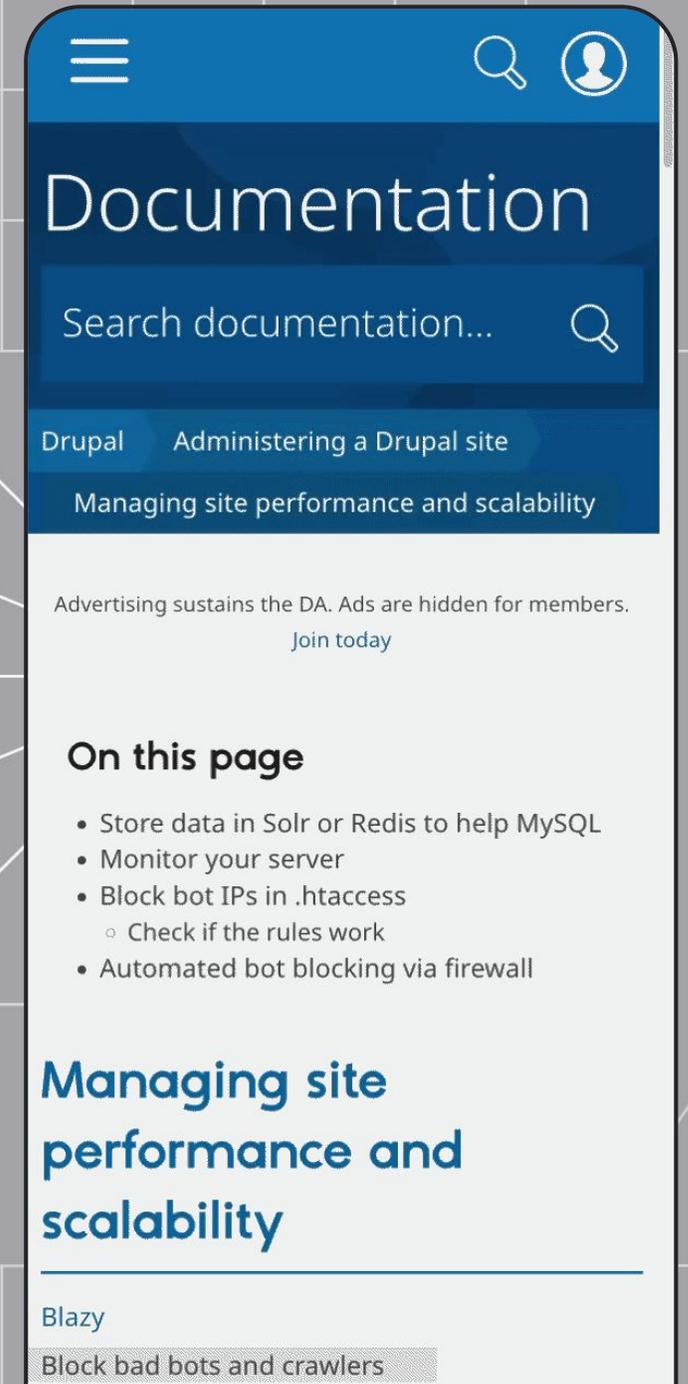
[Protect Views Flood Control](#)

[Bot Blocker](#)

[Cookie Bot Protection](#)

## Other methods

- Use Solr for data storage to avoid database (e.g., MySQL)
- Redis caching
- Monitor your server with alerts
- Blocking bots with .htaccess
- Automated bot blocking via a firewall



05

# Case study



# Find a Program

[Home](#) / [Programs](#) / [Find a Program](#)

## Filter by

Reset Filters 

Keyword 

For international students

Continuing and Professional Education only

Areas of interest 

Academic and University Studies

1-10 of 69 programs

[Academic Upgrading](#) 

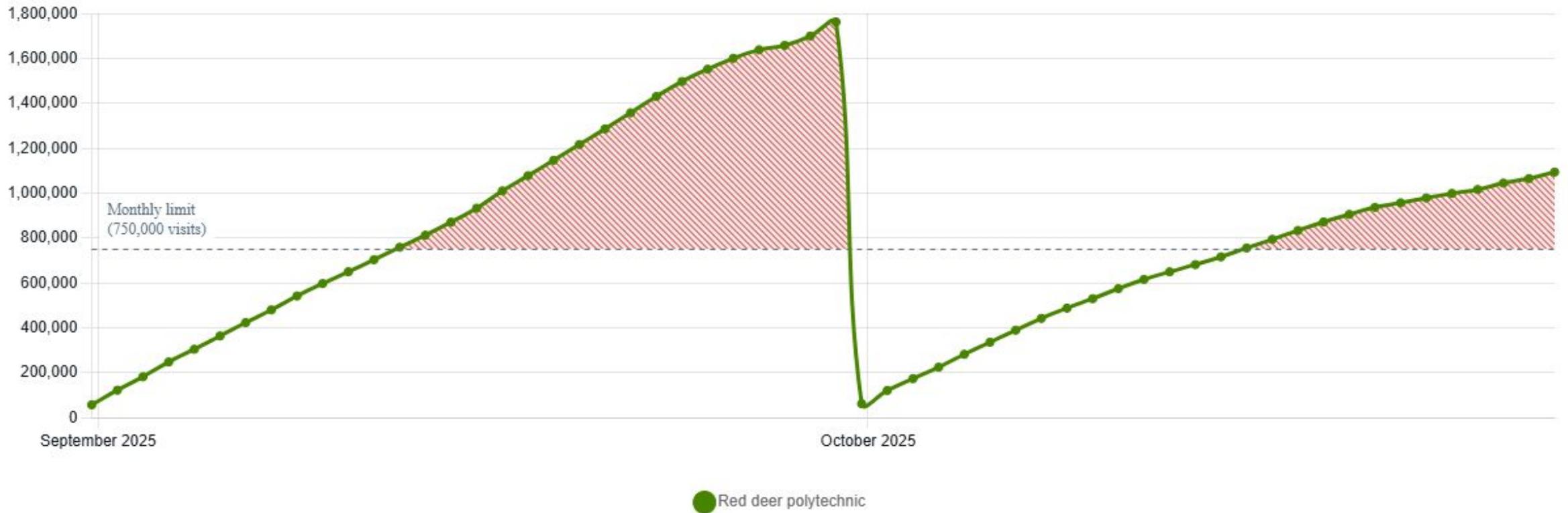
DIPLOMA 2 YEARS  OPEN 

[Administrative Professional Certificate](#) 

CERTIFICATE 2 YEARS  OPEN

[Agricultural Equipment Technician](#)

# Situation



# What didn't work

- Some of our WAF rules
- Adding special rules, such as `Crawl-delay`, to `robots.txt`
- Moving links from `href` anchor attribute to `data-link` (or similar) with JS redirect execution
- Masking a facet checkbox from regular users, since crawlers typically parse HTML structure to ban by IP

# What did work

## WAF/CDN

- WAF Rules
  - Rate limiting
  - Challenging faceted search traffic
  - Blocking
- Caching at the edge
- Redirects at the edge

## Drupal

- Following best practices
- Upgrading Facets version 2 to 3
- Re-working facets so they are a part of search forms (instead of lists of links, for example)

## CASE STUDY

# Challenge faceted search traffic

Security rules > Custom rules > Edit custom rule

### Edit custom rule

Protect your website and API from malicious traffic with custom rules. Configure mitigation criteria and actions, or explore templates, for better security.

[Learn more](#)

Rule name (required)  
  
Give your rule a descriptive name.

When incoming requests match...

Field	Operator	Value	
URI Query Str...	wildcard	*f%5B0%5D*	And ×
		e.g. page=1234*	
And			
Verified Bot C...	is not in	Search Engine Crawler	And ×
		<a href="#">See list of Bot Names and categories here</a>	
And			
Country	is not in	Canada	And Or ×
		e.g. GB	

Expression Preview [Edit expression](#)

```
(http.request.uri.query wildcard r"*f%5B0%5D*" and not cf.verified_bot_category in {"Search Engine Crawler"} and not ip.src.country in {"CA"})
```

Then take action...

Choose action  
  
Presents an interactive or non-interactive challenge to the client

Place at  
Select order:

# Rate limiting search pages

Security rules > Rate limiting rules > Edit rate limiting rule

## Edit rate limiting rule

Protect your website and API from malicious traffic with rate limiting rules. Configure mitigation criteria and actions for better security.

[Learn more](#)

Rule name (required)  
  
Give your rule a descriptive name.

When incoming requests match...

Field	Operator	Value	
URI Path	contains	/programs/find-program e.g. /content	And X
Or			
URI Path	contains	/search e.g. /content	And Or X

Expression Preview [Edit expression](#)

```
(http.request.uri.path contains "/programs/find-program") or (http.request.uri.path contains "/search")
```

Cache status

Also apply rate limiting to cached assets

With the same characteristics...

Use custom counting expression

When rate exceeds...

Requests (required)  Period (required)

Then take action...

Choose action  
  
Presents an interactive or non-interactive challenge to the client

Place at

Select order:

# Incoming 'cache everything' rule

### Edit Cache Rule

Rule name (required)

Give your rule a descriptive name.

---

**If incoming requests match...**

Custom filter expression  
Only apply the rule to requests matching the custom filter expression

All incoming requests  
Apply the rule to all requests

**Then...**

**Cache eligibility** (required)

Mark whether the request's response from origin is eligible for caching. Caching itself will still depend on the cache-control header and your other caching configurations. [Learn more](#)

Bypass cache

Eligible for cache

---

**Edge TTL** (optional) ✕

Specify if and how long Cloudflare should cache the response, depending on if a cache-control header is present on the origin response. [Learn more](#)

Use cache-control header if present, bypass cache if not

Use cache-control header if present, cache request with Cloudflare's default TTL for the response status if not

Ignore cache-control header and use this TTL

Input time-to-live (TTL) (required)

# Bypass cache for authenticated users rule

### Edit Cache Rule

Rule name (required)  
  
Give your rule a descriptive name.

---

**If incoming requests match...**

**Custom filter expression**  
Only apply the rule to requests matching the custom filter expression

**All incoming requests**  
Apply the rule to all requests

**When incoming requests match...**

Field	Operator	Value	
<input type="text" value="Cookie"/>	<input type="text" value="contains"/>	<input type="text" value="SSESS"/>	<input type="button" value="And"/> <input type="button" value="Or"/>

e.g. name = value

Expression Preview [Edit expression](#)

```
(http.cookie contains "SSESS")
```

**Then...**

**Cache eligibility** (required)  
Mark whether the request's response from origin is eligible for caching. Caching itself will still depend on the cache-control header and your other caching configurations. [Learn more](#)

**Bypass cache**

Eligible for cache

---

**Browser TTL** (optional)   
Specify how long client browsers should cache the response. Cloudflare cache purge will not purge content cached on client browsers, so high browser TTLs may lead to stale content. [Learn more](#)

---

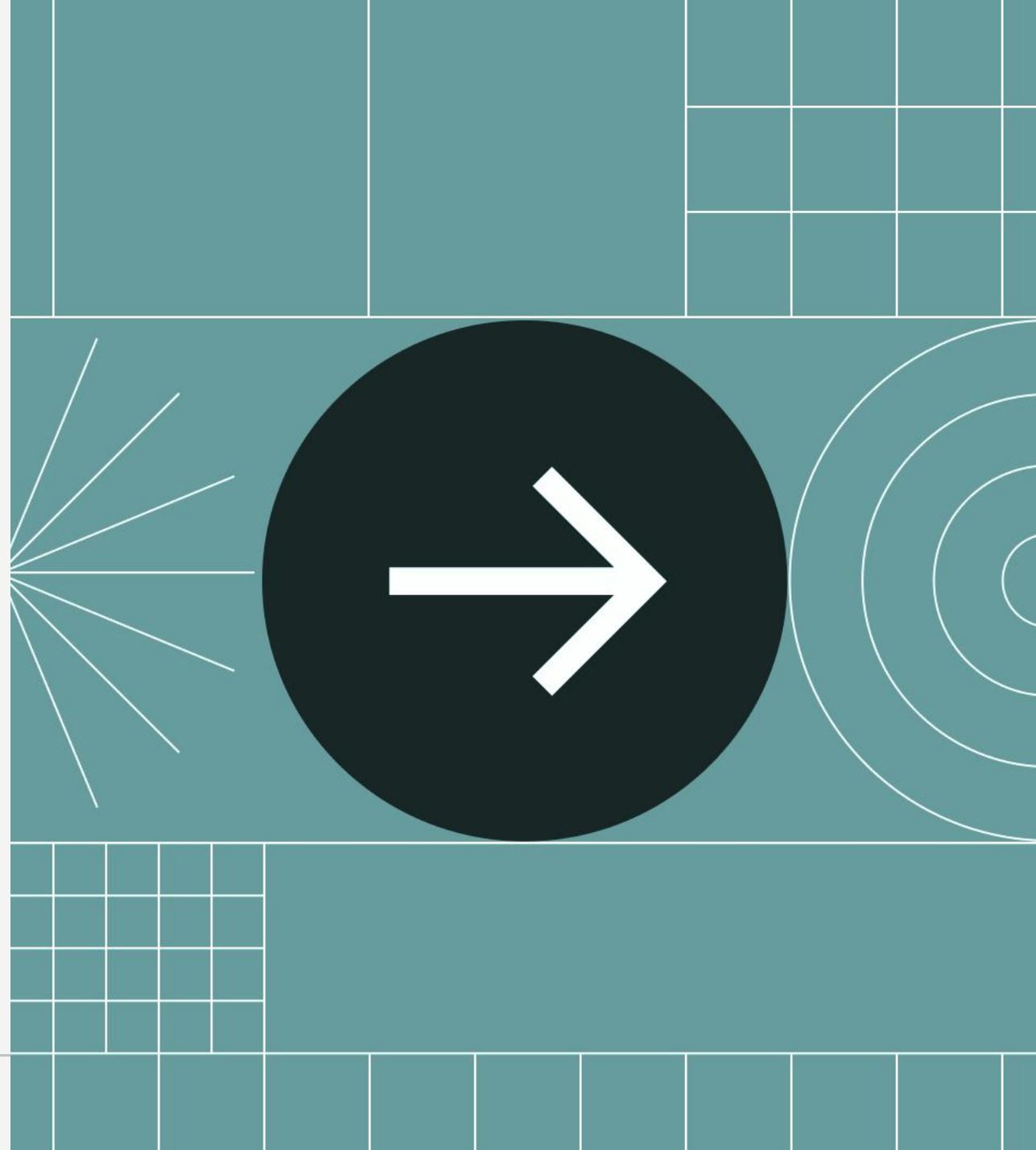
**Place at**

Select order:

Select which rule this will fire after:

06

# Actionable playbook



# Your 4-step bot defensive plan

1. Measure bot vs. human (“real”) traffic to provide a baseline, and then monitor/adjust
2. Add and configure a Web Application Firewall and Content Delivery Network (WAF/CDN) if at all possible
3. Review your facets
4. Follow Drupal best practices

07

# Questions



Thank you

