Birds of Feather

How to cope with the rapidly increasing number of aggressive incoming requests?










93x
IP

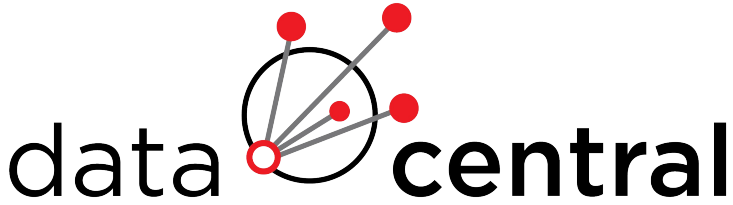
Who are we?



- Grégory Mantelet
- Gilles Landais
- Pierre Fernique
- Anaïs Oberto



- Christophe Arviset
- Silvia de Castro
- Arnaud Masson



- Simon O'Toole
- Nuria Lorente

Who are we

- Grégory Mantel
- Gilles Landais
- Pierre Fernique
- Anaïs Oberto

VizieR versus increasing bot scraping

P503

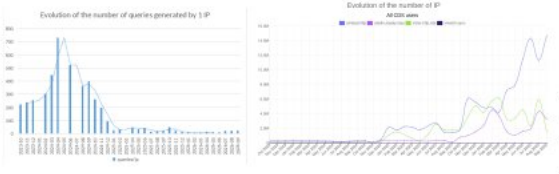
The CDS VizieR service, like many Data Centers, is facing a dramatic increase in bot crawling. These are millions of queries each day, which pollute and impact the VizieR service. This massive and blind harvesting targets static and dynamic resources and is generally controlled by AI to populate knowledge bases. Log analysis shows many queries, executed by a wide range of IPs, all of which look legitimate without any obvious AI marker: it becomes harder to detect their origin and to manage the flow.
Maintaining good availability for users requires redefining data access management. This is often done urgently, but in the background, the CDS is investigating access management for the long term. In this poster, we present the solutions used, the solution implemented today, and those planned.

Drawing up attackers profile

It concerns mainly AI harvesting which target static and dynamic data. The queries generated by bots and AI are generally not signed, with no user-agent, and comes from clouds using wide range of IP . Their activity is reflected in a massive increase in requests that privileges the HTML format. Search engine crawlers combine blind crawling—which follows all links—with the ability to generate queries that are indistinguishable from those made by users. Significantly less naive than their predecessors, they bypass the usual HTML forms traps.

The IVOA's APIs seem to be spared for now.



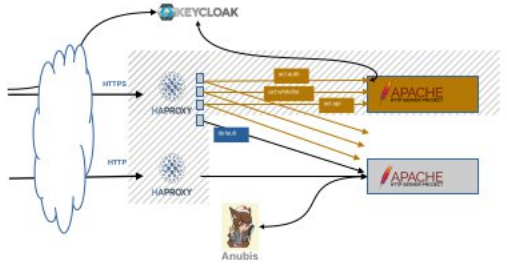


The battle to provide a user-centered service

The CDS benefited to the bots activities that indexed resources (eg: VizieR catalogue in Google search, AI providing result and methods to query database). However the counterpart becomes unbearable when the harvesting activity make down the service. **To deal with this unethical behavior, the CDS has no choice but to filter the bots.**

- Traditional methods, based on IP addresses (**blacklists**) and request cadences (**fail2ban**), are no longer effective in a context of clouds. These use thousands of IP across very broad ranges that generates only a small number of requests.
- WEB-Firewall softwares, such as **Anubis**, used a combination of methods : IP and user-agent blacklists, request frequency, JavaScript challenges. Anubis which is used by VizieR and Simbad filters a significant part of bots. Unfortunately, the experience has shown us that this can be partially circumvented.
- **HTTP rules** combining HTTP header (eg: user-agent, referrer) and URL parameters is used in VizieR to detect abusive or inconsistent queries (eg: conesearch on large dataset with an exaggerated radius, bots queries to get details records, etc)
- Finally, the redundancy of mirrors existing for VizieR and Simbad enforced the reliability. Those have been spared for a long time by bots, but are now beginning to be harvested too.

Planned VizieR architecture for routing queries



Planning to provide Gold access

The VizieR service works on a new architecture that redefines the access policy by maintaining open access while offering privileged access under certain conditions. The new architecture (currently in test) consists in replicating locally both the service and the database and use a proxy which routes incoming requests to a "Gold" server or the "common" server.

The routes are managed by an Access Control List (ACL). Among the ACLs, the CDS is investigating authentication, taking into account possible VO connections (see AuthVO, J. Salgado). The **Keycloak** tool is currently being evaluated ; it includes standard protocols such as OpenID or Oauth2 and supports both local and delegated authentication, such as ORCID and EduGain which are mandatory to deal with EOSC nodes. Currently in development, authentication has been tested for user web interface (HTML output). Authentication on API (eg: IVOA API) are not the priority today and requires a maturity in the IVOA network with clients ready to authentication implementation.



G. Landais, G. Mantelet, P. Fernique, A. Oberto,
T. Keller, P. Hoffmann
gilles.landais@astro.unistra.fr

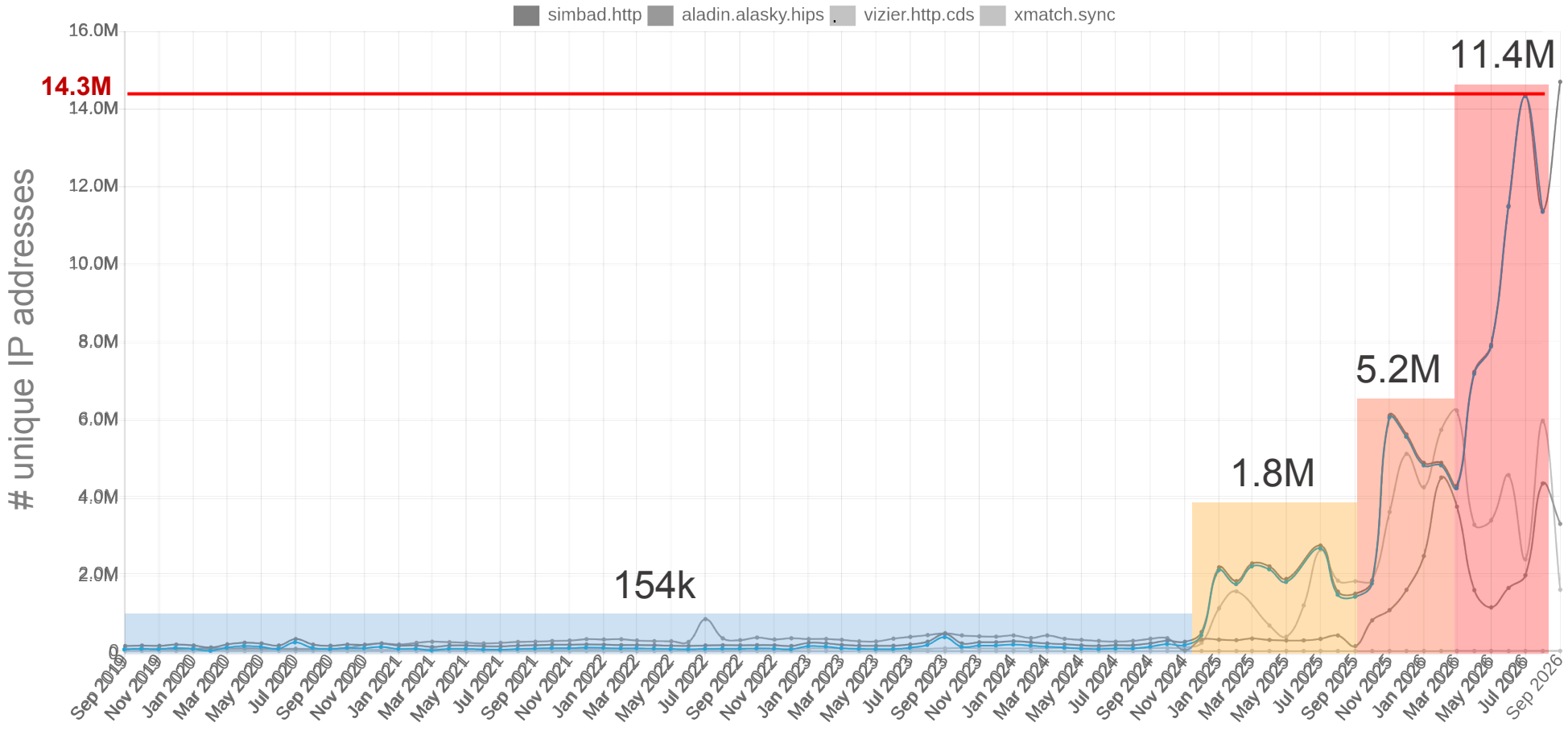
ADASS XXXVI Boorloo/Perth – Australia – Observatoire astronomique de Strasbourg

data central

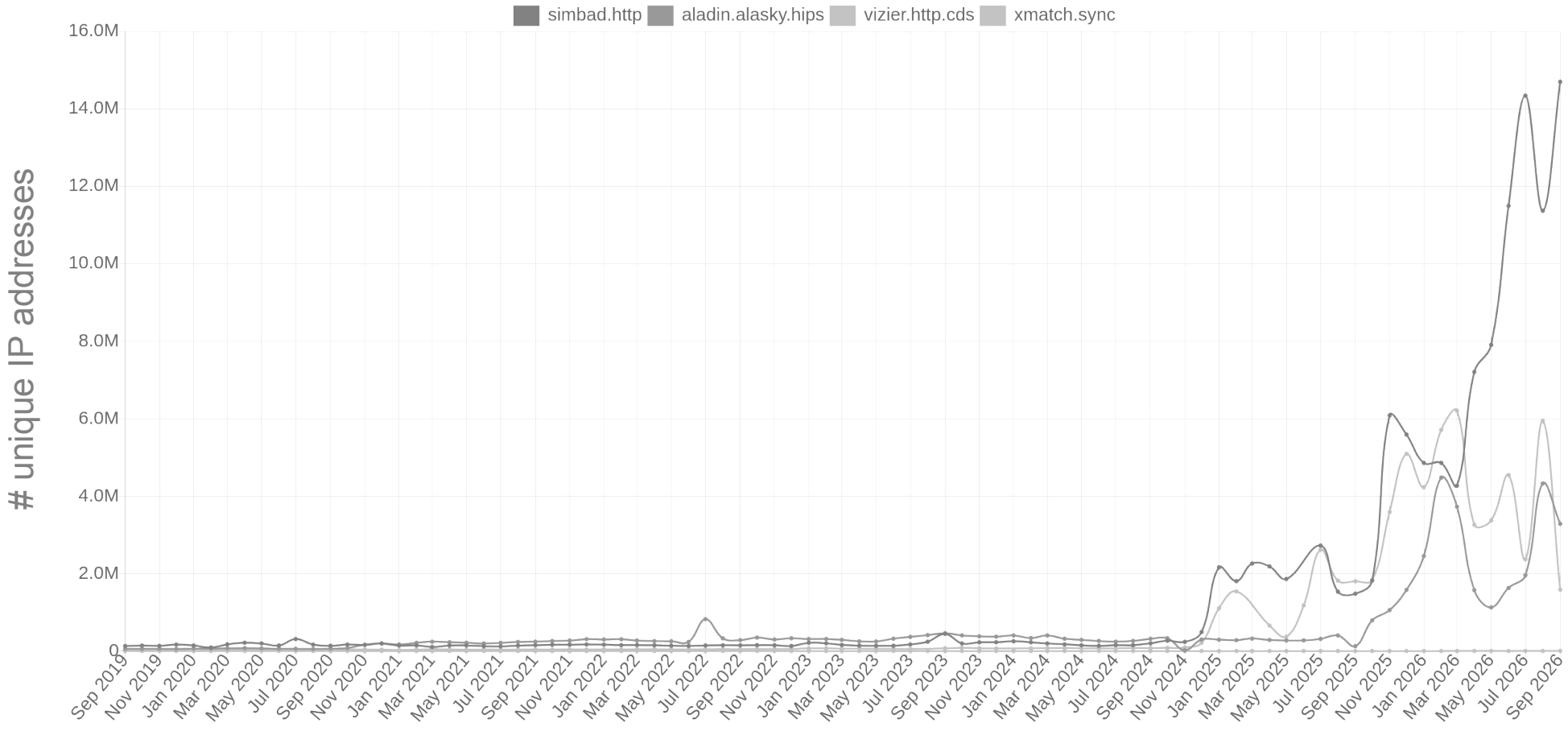
- Simon O'Toole
- Nuria Lorente


What happens?

Bots use a lot more IP addresses



Bots use a lot more IP addresses

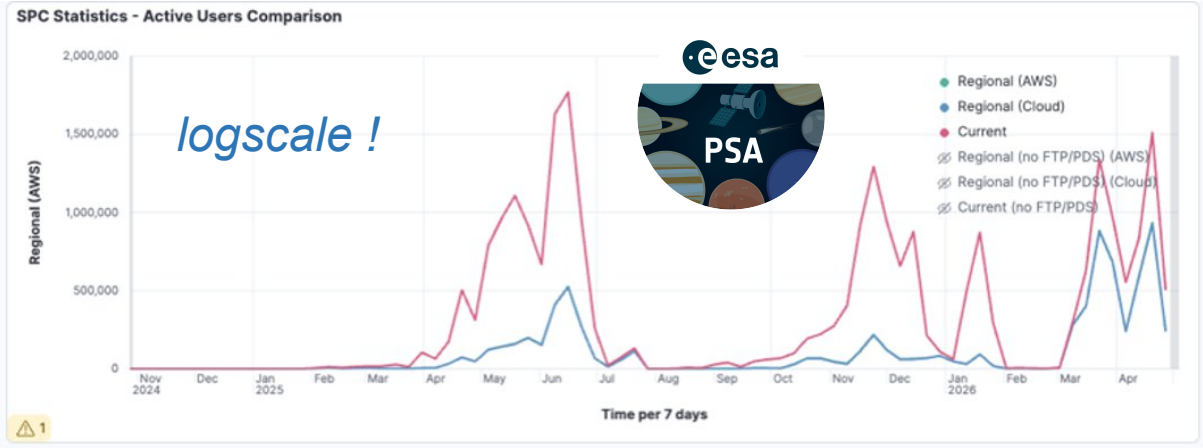


Data Central experience

- System wide impacts
- Some services receive large bot traffic: 35K requests/hr
- IP abuse DBs suggest ~50% hacked IoT devices – Botnets !
- Blocked traffic has similar pattern to VizieR poster by Landais et al.
- Clear scraping behaviour: one IP making hundreds of requests in short time

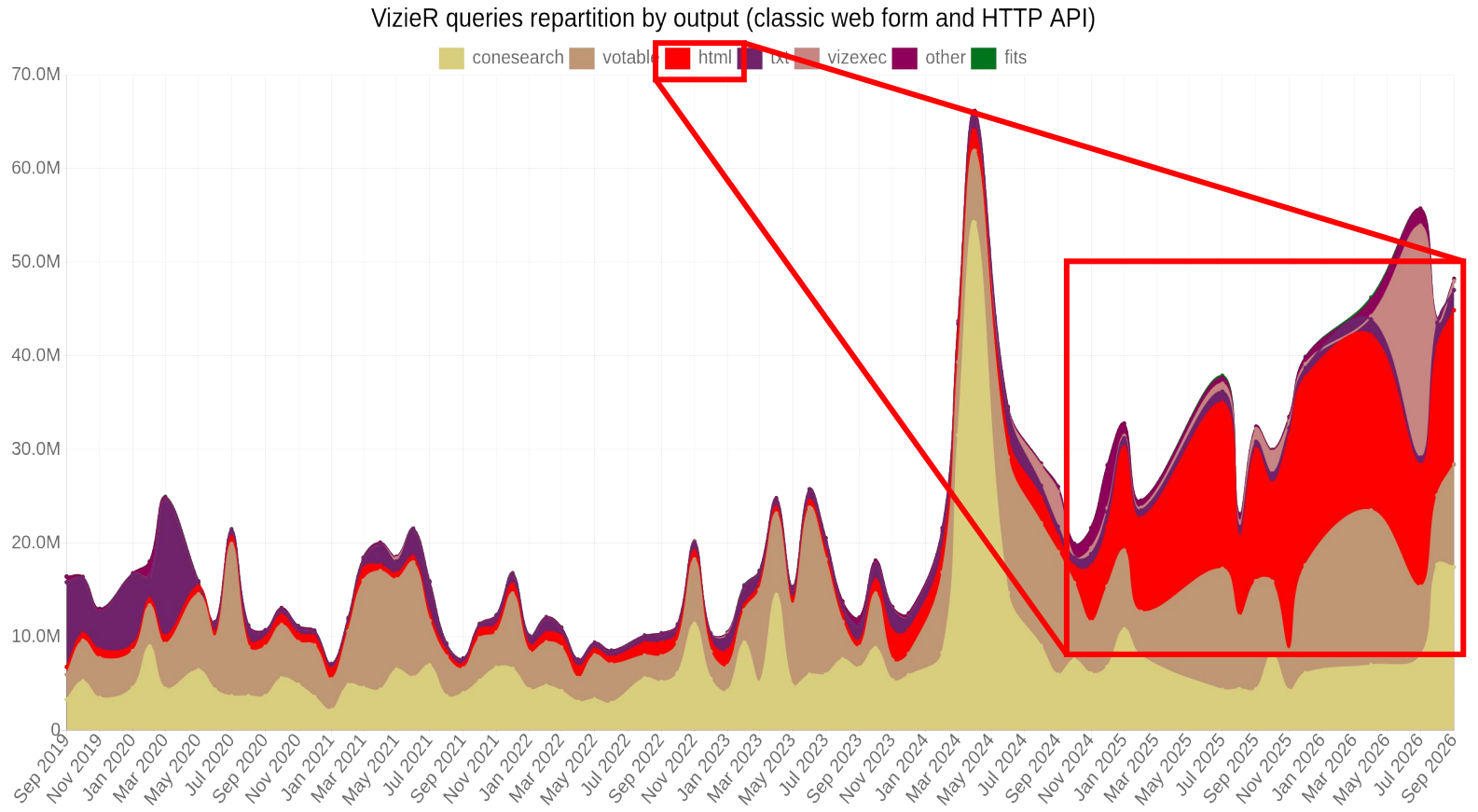
ESA Science Archives experience

- ESDC : archives.esac.esa.int
- Reporting Archives usage to our funding agencies for many years
- Since 2022, unrealistic peaks of usage (>100k-1M per month), knowing size of community
- Probably due to bots, LLMs training, Cloud accesses from unusual countries, anonymous FTP accesses



Apparently distributed Web scrapping

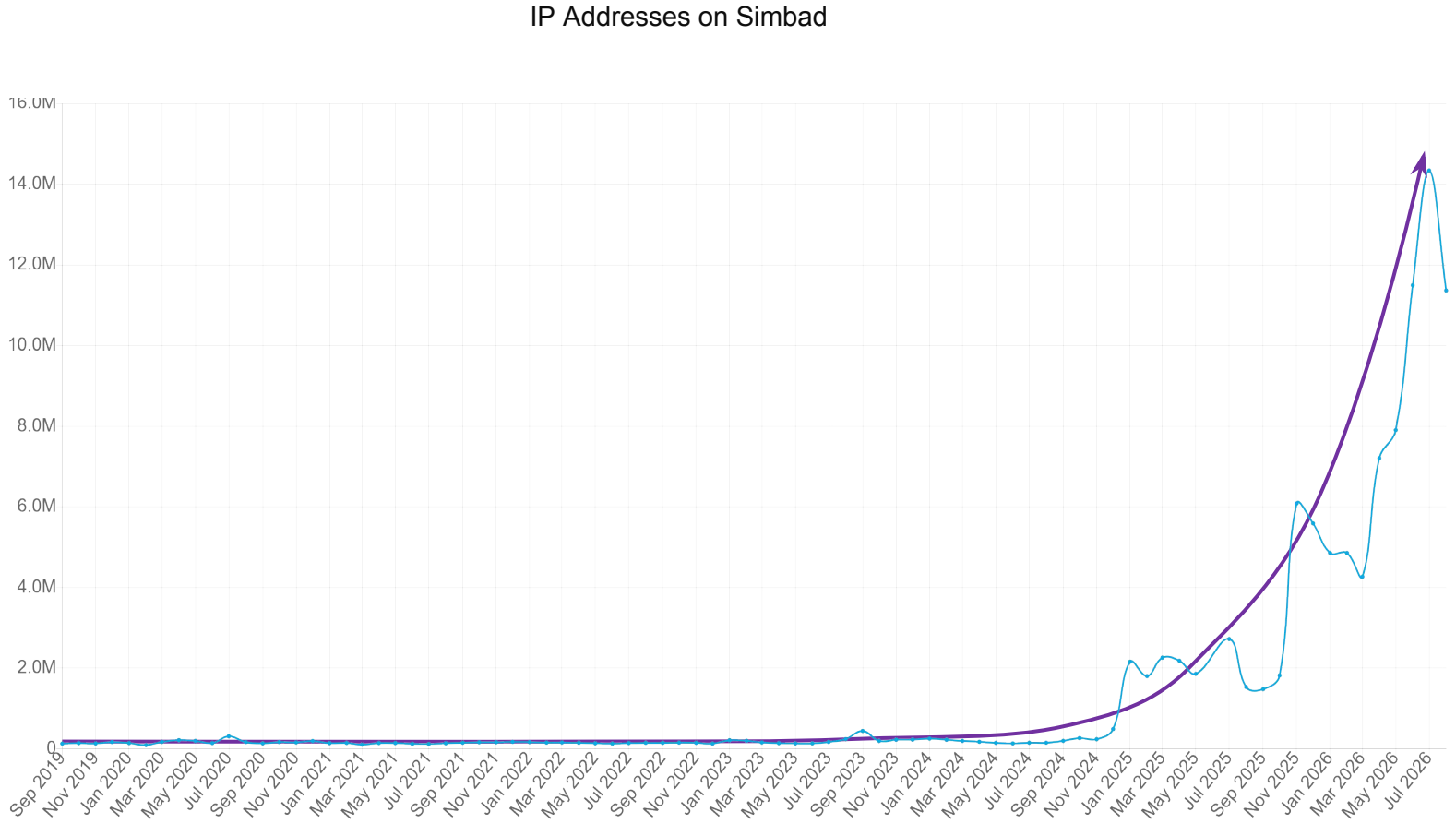




Apparently distributed Web scrapping





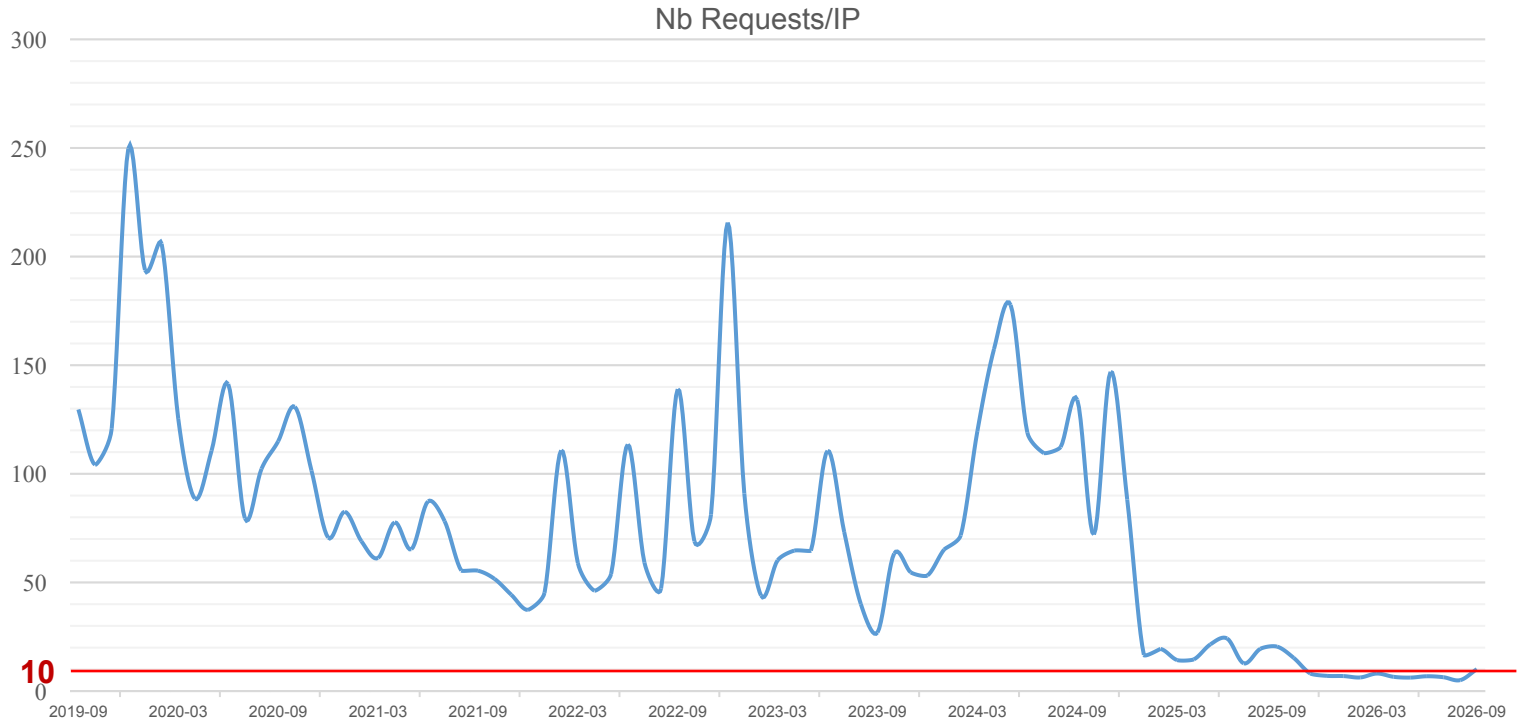


Apparently distributed Web scrapping

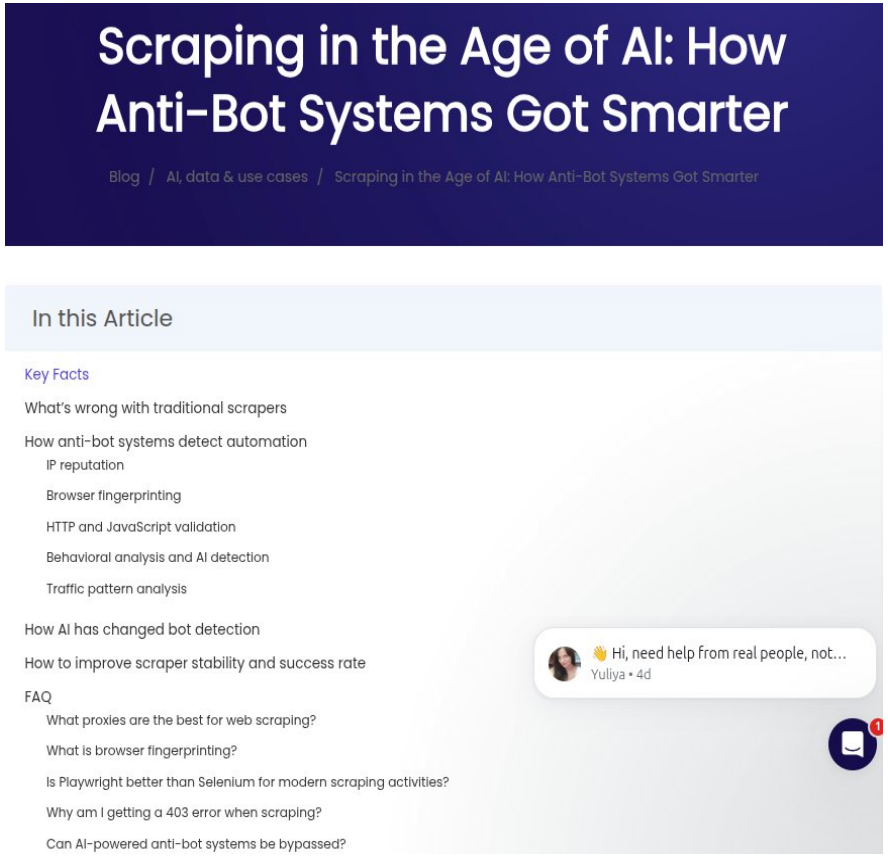




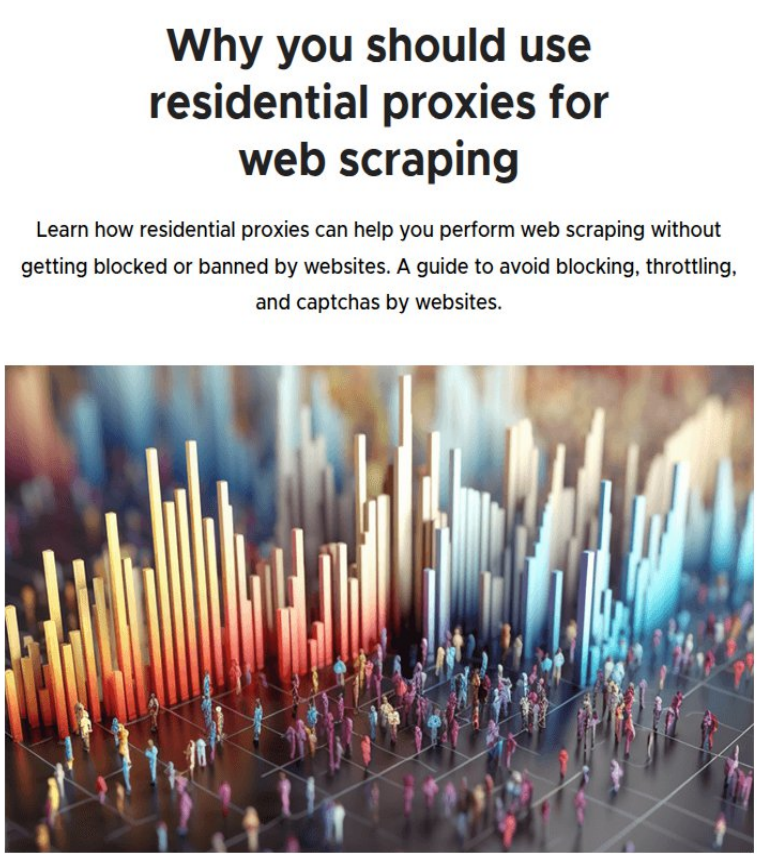




...driven by AI?







🔗 Data Impulse
https://dataimpulse.com/blog/scraping-in-the-age-of-ai-how-anti-bot-systems-got-smarter/

🔗 TechwareLab
https://techwarelab.com/blog/residential-ai-are-ai-companies-being-inspired-by-web-scraping

🔗 Infatica.io
https://infatica.io/blog/residential-proxies-web-scraping/


Let’s speak!

Questions

- Other interpretation of this situation?
- How to distinguish humans and bots?
- Would authentication be a good solution?
- Would it be well accepted by users?
- Should we really ban completely AI?
- ...

Quick survey
(5 questions)

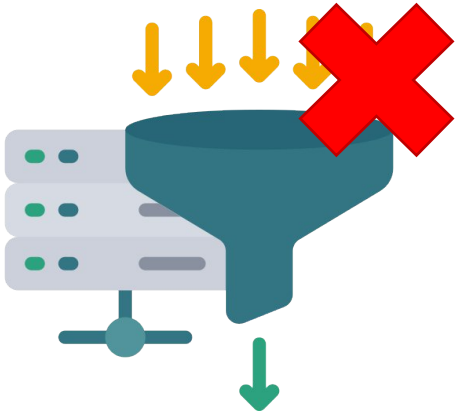
Are you impacted? How? Solutions?...


How to protect ourselves?

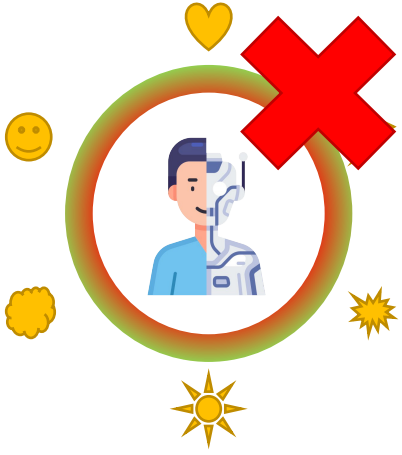
The usual methods don’t work any more



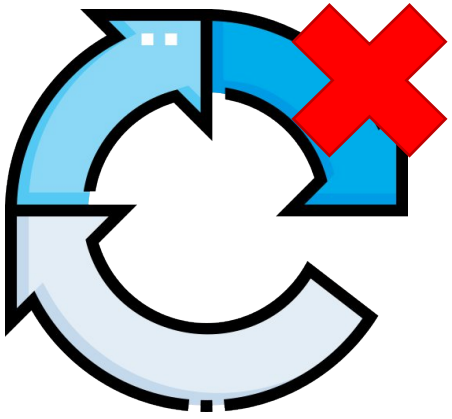
Blacklist IP



Rate limite



Behaviour identification



CAPTCHA

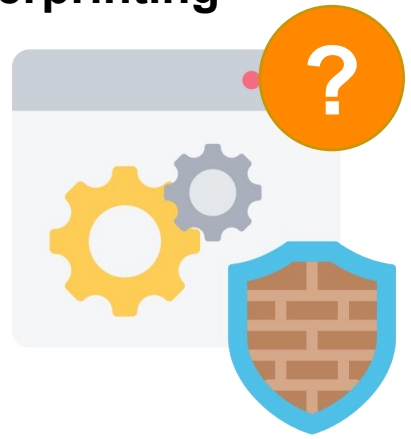
CDS approach



Fingerprinting

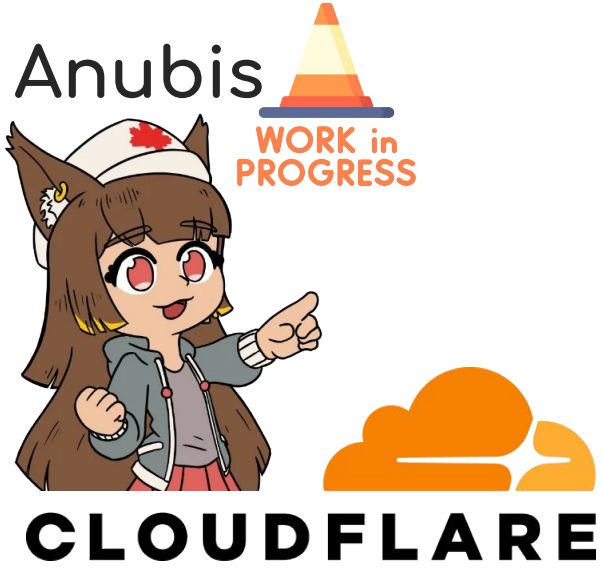


Behaviour identification



WAF

(Web Application Firewall)



Proxies



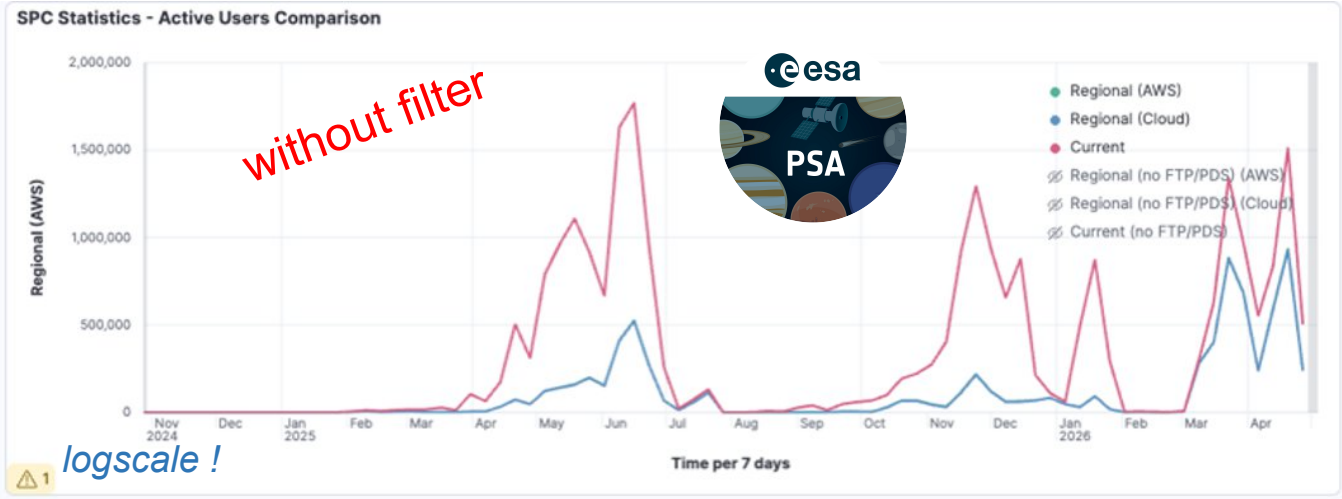
Authentication

Data Central approach

- System wide impacts
  - **Move compute intensive endpoints behind auth**
    - Short term improvement
- Large bot traffic
  - **Blocked whole GCP region IPs**
    - Short term improvement
- Botnets
  - **Current solution: mod_security plus CoreRuleSet**
- Scraping behaviour
  - **CRS + some custom rules per service blocked most bogus requests**

ESA Science Archives approach

- Caveat : work still in progress and being tuned
- All Archive usage logs are in Elastic Search, gives flexibility
- For the usage stats
  - Filter out all Cloud accesses
  - Filter out all accesses from anonymous FTP
  - Only include IPs from Countries with active Science Community
- Not ideal, quite conservative but gives more realistic statistics






Thank you
for your participation!