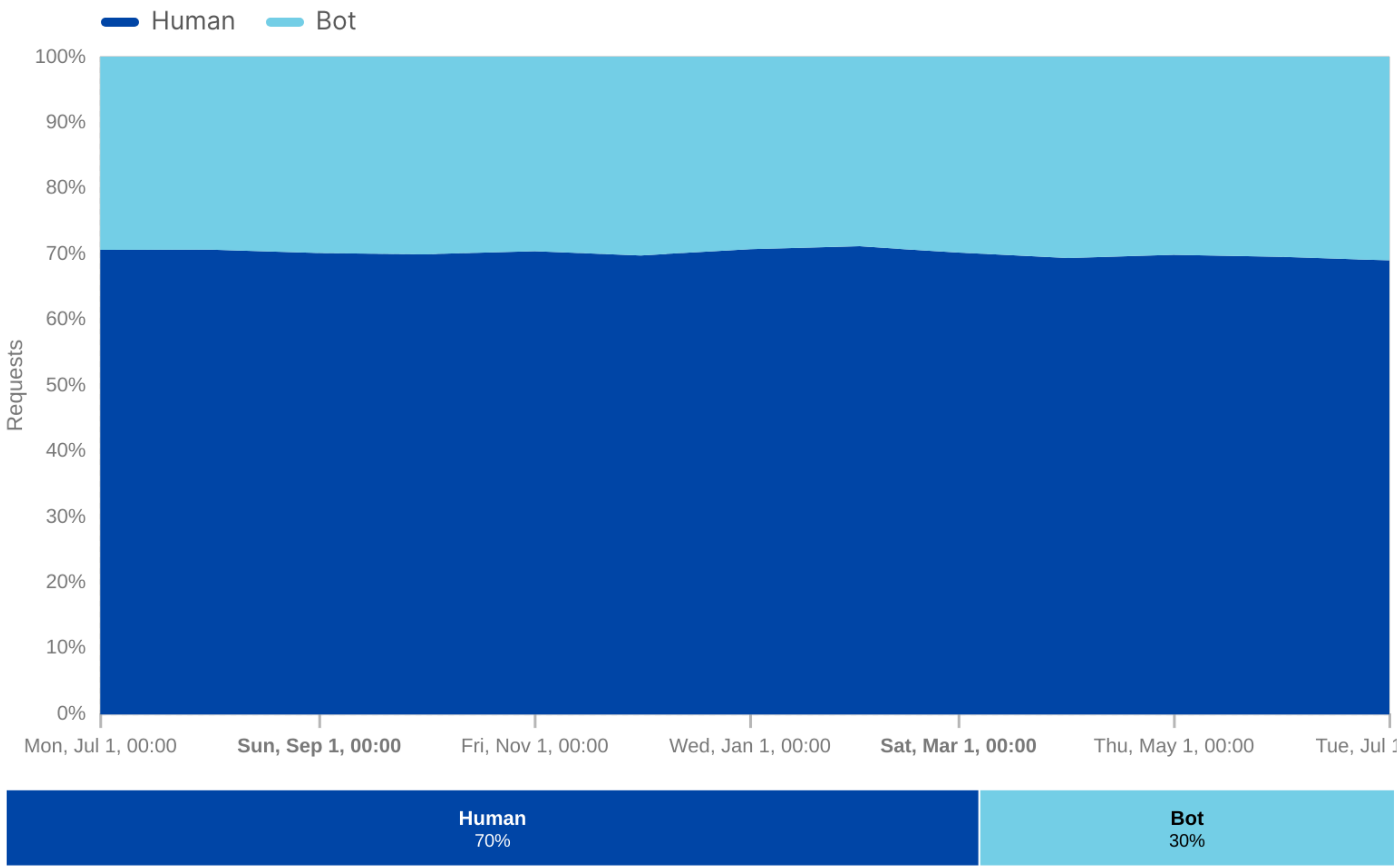
Web Bot Auth

Background and Context

Mark Nottingham, IETF 123 Madrid

HTTP requests by bot class time series

Distribution of HTTP requests classified as automated or human over time



Cloudflare Radar

Last 12 months | Jul 15, 2025, 15:30 UTC

“Automated bot traffic surpassed human-generated traffic for the first time in a decade, constituting 51% of all web traffic in 2024. [C]yber criminals are increasingly leveraging these technologies to create and deploy malicious bots which now account for 37% of all internet traffic.”

– Summary, 2025 Imperva Bad Bot Report

Requirement:
Identify Bot Traffic

A Better Web

- Cryptographically identified bots
- Non-cryptographicaly identified bots
- Humans

Current Practice


I AM GOOGLEBOT/2.1

```
{
    "creationTime": "2025-07-15T14:47:23.000000",
    "prefixes": [
        {
            "ipv6Prefix": "2001:4860:4801:10::/64"
        },
        {
            "ipv6Prefix": "2001:4860:4801:11::/64"
        },
        {
            "ipv6Prefix": "2001:4860:4801:12::/64"
        },
        {
            "ipv6Prefix": "2001:4860:4801:13::/64"
        },
        {
            "ipv6Prefix": "2001:4860:4801:14::/64"
        },
        {
            "ipv6Prefix": "2001:4860:4801:15::/64"
        },
        {
            "ipv6Prefix": "2001:4860:4801:16::/64"
        },
```

Google's crawlers fall into three categories:

Type	Description	Reverse DNS mask
Common crawlers	The common crawlers used for Google's products (such as Googlebot). They always respect robots.txt rules for automatic crawls.	`crawl-***-***-***-***.googlebot.com` or `geo-crawl-***-***-***-***.geo.googlebot.com`
Special-case crawlers	Crawlers that perform specific functions for Google products (such as AdsBot) where there's an agreement between the crawled site and the product about the crawl process. These crawlers may or may not respect robots.txt rules.	`rate-limited-proxy-***-***-***-***.google.com`
User-triggered fetchers	Tools and product functions where the end user triggers a fetch. For example, Google Site Verifier acts on the request of a user. Because the fetch was requested by a user, these fetchers ignore robots.txt rules. Fetchers controlled by Google originate from IPs in the `user-triggered-fetchers-google.json` object and resolve to a `google.com` hostname. IPs in the `user-triggered-fetchers.json` object resolve to `gae.googleusercontent.com` hostnames. These IPs are used, for example, if a site running on Google Cloud (GCP) has a feature that requires fetching external RSS feeds on the request of the user of that site.	`***-***-***-***.gae.googleusercontent.com` or `google-proxy-***-***-***-***.google.com`

Constraints and Considerations

1. Web Compatibility

- Existing Web flows can't be disrupted
  (browsers should browse without changes)

- Needs to be robust in face of common Web infrastructure
  (e.g., intermediaries and libraries that modify headers)

2. Deployment Costs

- Needs to be reasonable to implement
  (both for bots and Web servers)
- Needs to be easy for small Web sites to deploy
- Shouldn't incur too much overhead
  (compute or on-wire)

3. Centralization

- Avoiding Choke Points:
  - Authorization decisions (*already present in current web*)
  - Root(s) of identity (*new*)
- Avoiding Barriers / Promoting Diversity:
  - for bots
  - for sites