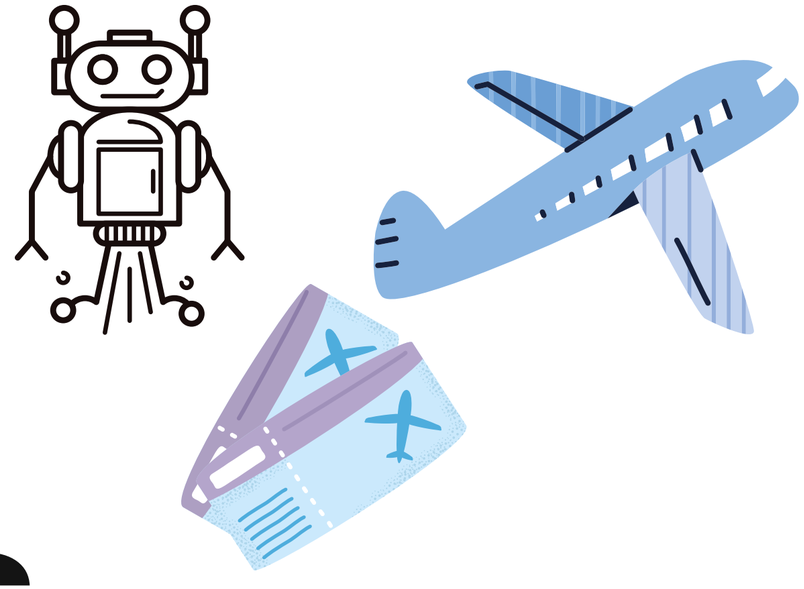
The bots arms race on airlines booking websites



Elisa Chiapponi
PhD student at EURECOM -Amadeus

EURECOM
Sophia Antipolis

amadeus
Global Security Operations

Agenda

Agenda

1. Introduction
and motivations

Agenda

1. Introduction and motivations

2. Honeypot

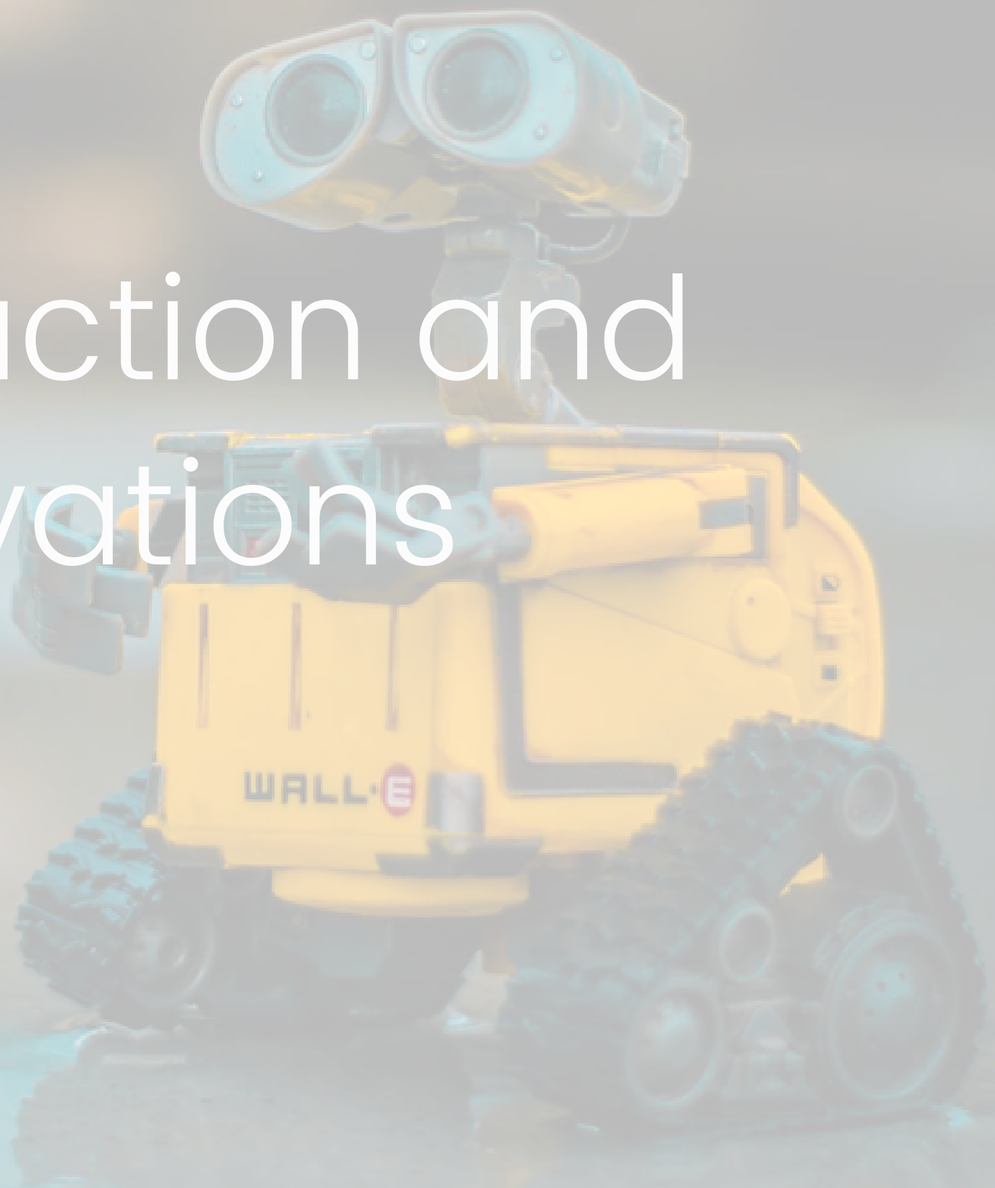
Agenda

1. Introduction and motivations
2. Honeypot
3. Proxy services and IP addresses

Agenda

1. Introduction and motivations
2. Honeypot
3. Proxy services and IP addresses
4. BADPASS project

1. Introduction and motivations



Who am I?

Master degree at Politecnico di Torino and EURECOM in Digital Security, thesis internship in Amadeus GSO -APP

Who am I?

Master degree at Politecnico di Torino and EURECOM in Digital Security, thesis internship in Amadeus GSO -APP

Phd student Amadeus GSO-APP and EURECOM

Who am I?

Master degree at Politecnico di Torino and EURECOM in Digital Security, thesis internship in Amadeus GSO -APP

Phd student Amadeus GSO-APP and EURECOM

Finding practical means to defeat scraping bots

Who am I?

Master degree at Politecnico di Torino and EURECOM in Digital Security, thesis internship in Amadeus GSO -APP

Phd student Amadeus GSO-APP and EURECOM

Finding practical means to defeat scraping bots

Understanding their ecosystem (actors, techniques, infrastructure)

Web scraping

Web scraping is the periodical or continuous retrieval of accessible data and/or processed output contained in web pages.

Scraping bots and airlines

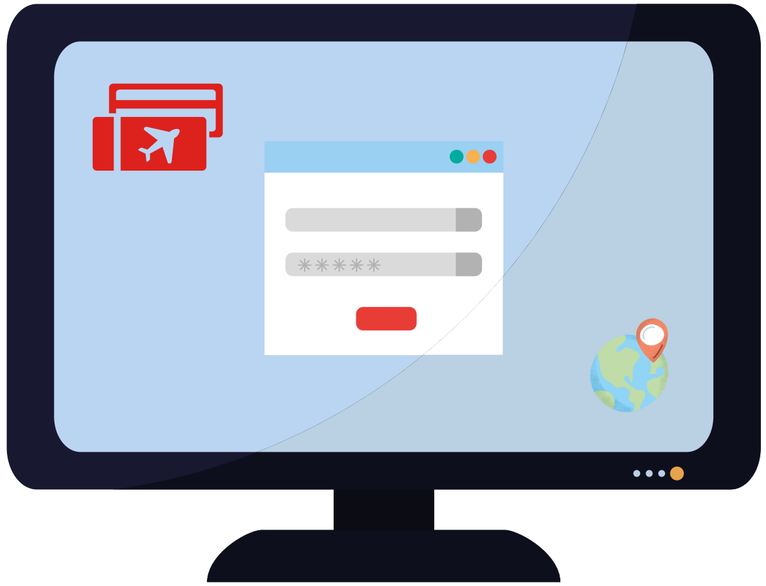
Legittimate traffic
56.1%

Bad bots
43.9%

How bots affect airlines - Imperva 2019

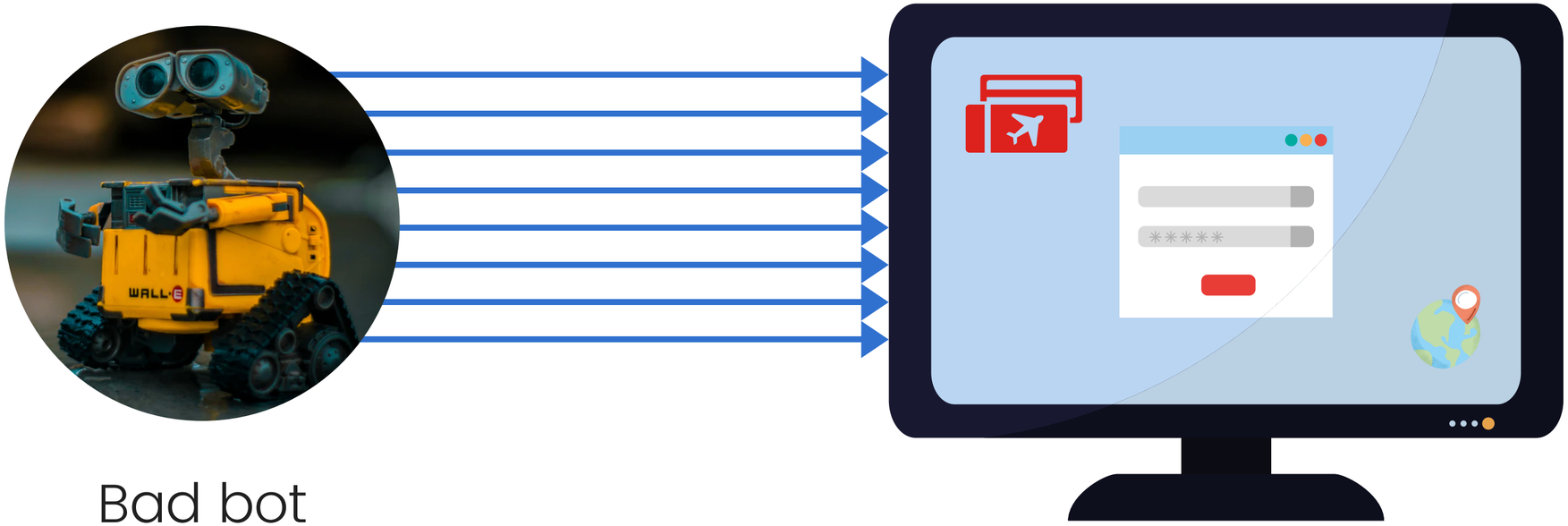
Web scraping bad bots

Bad bot



Airline booking domain

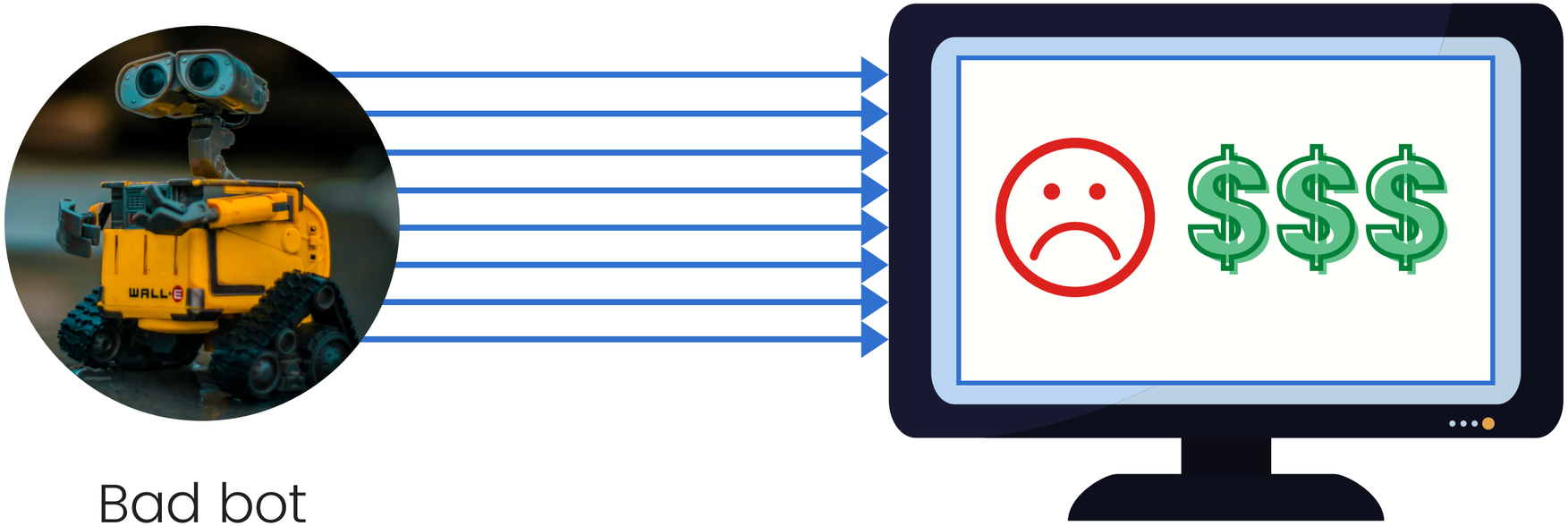
Web scraping bad bots



Bad bot

Airline booking domain

Web scraping bad bots



Bad bot

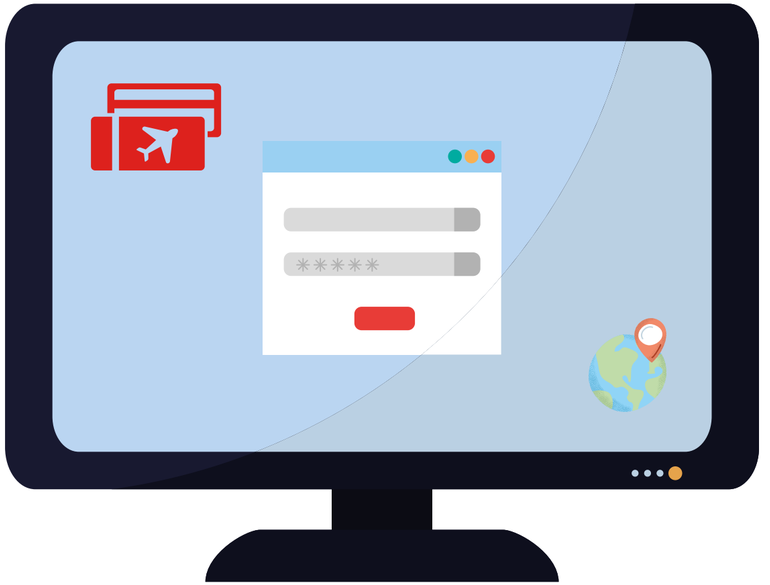
Airline booking domain

Anti-bot solutions

User

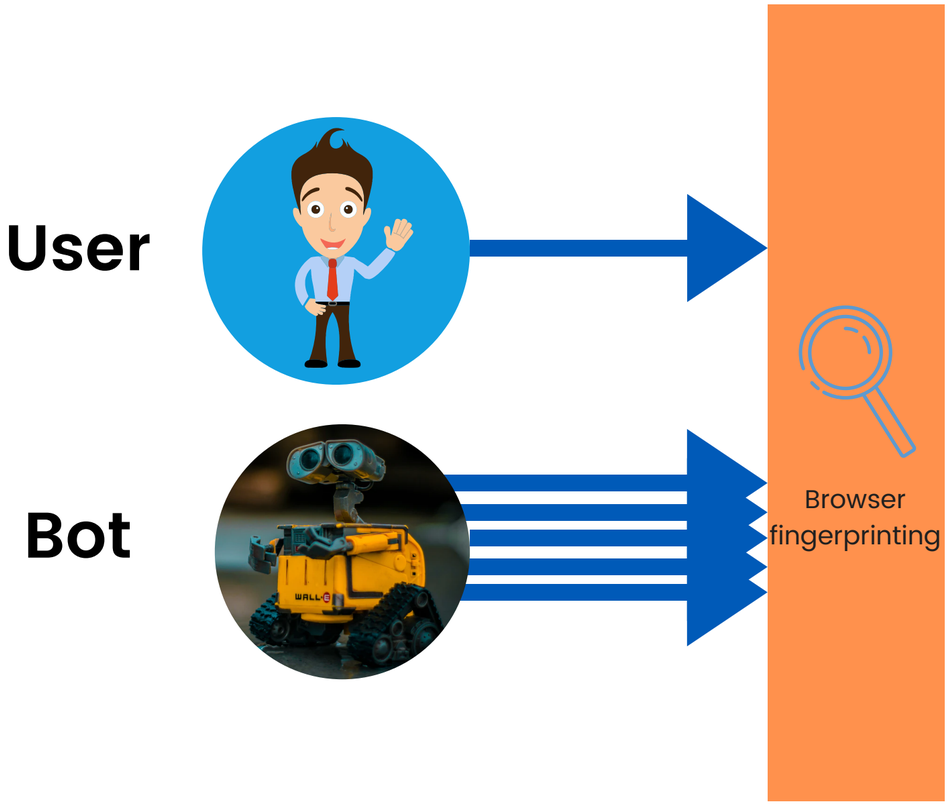
Bot

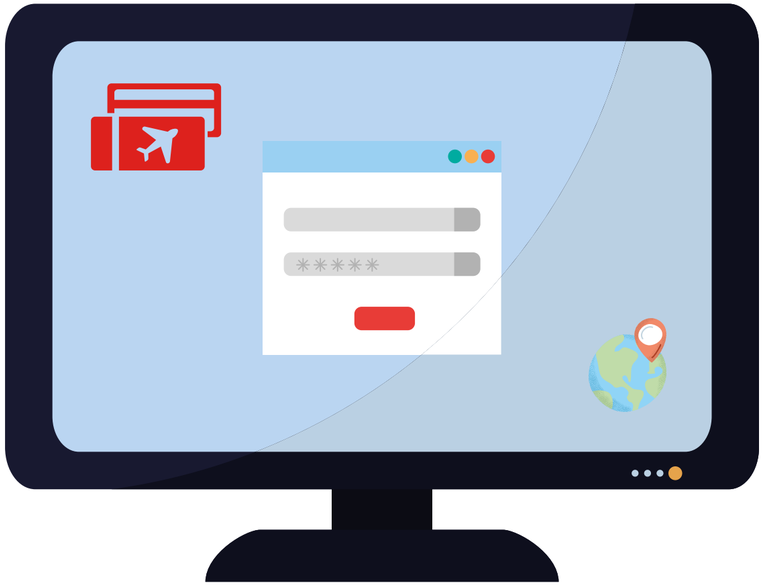
Browser fingerprinting



Booking domain of airline X

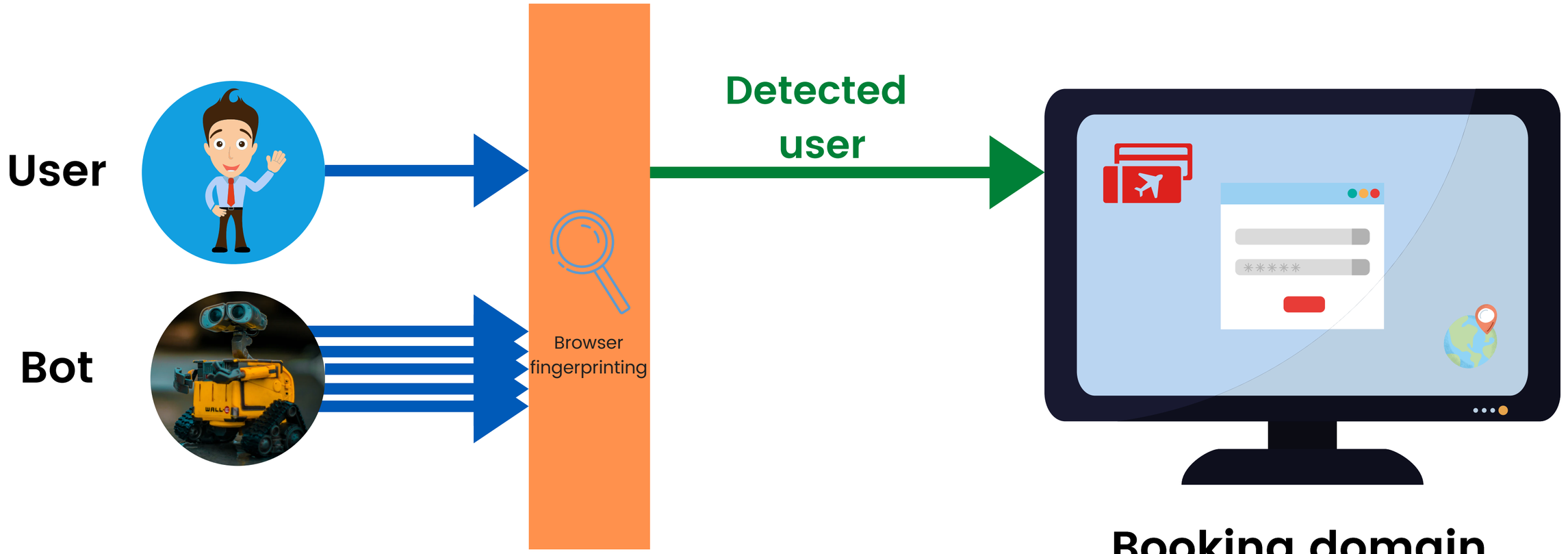
Anti-bot solutions



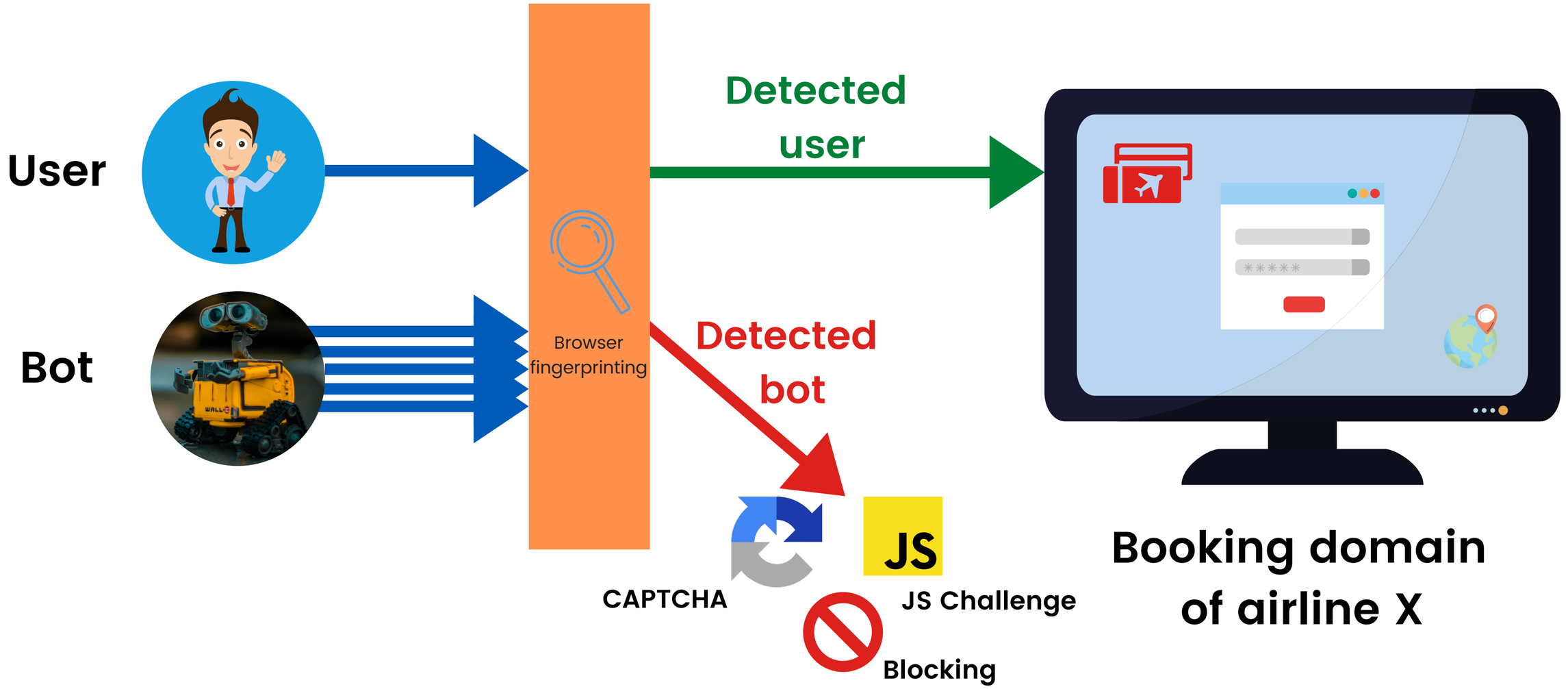


Booking domain of airline X

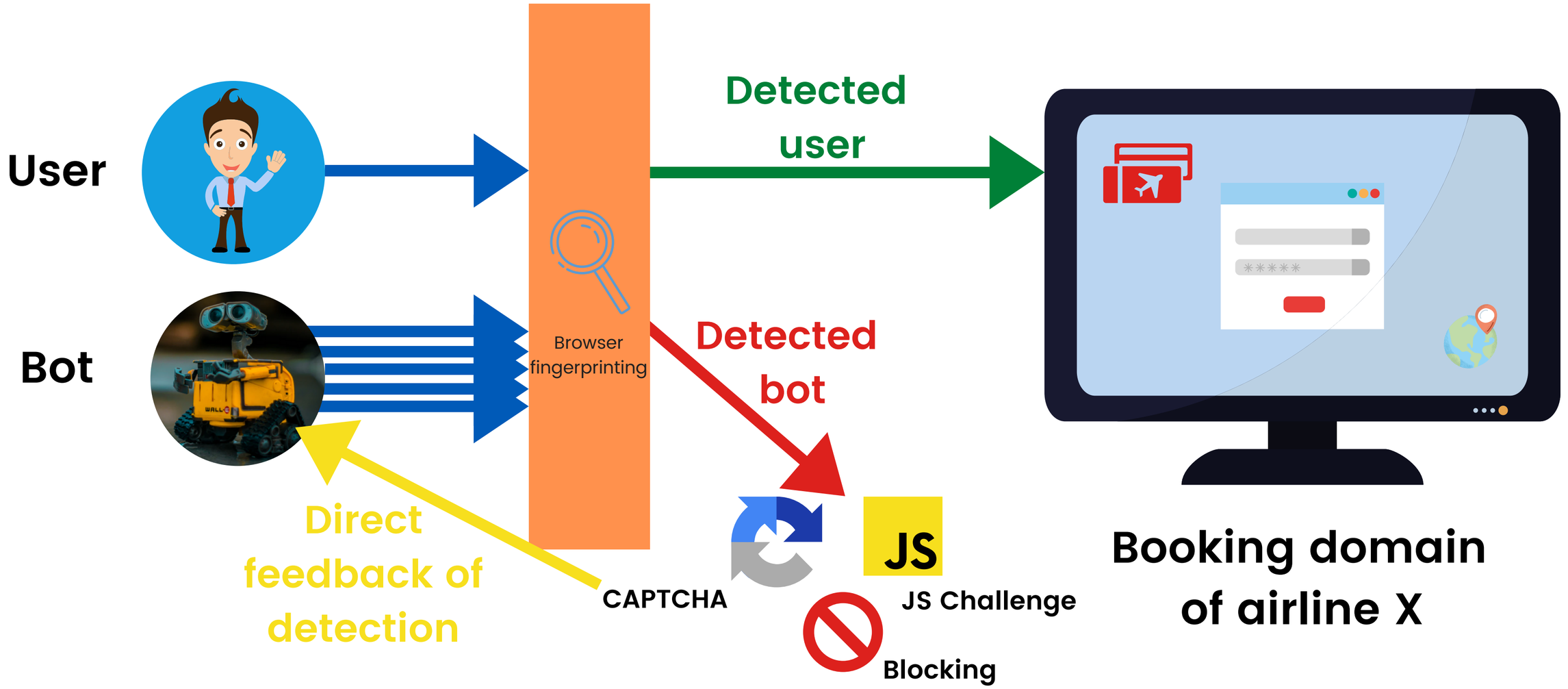
Anti-bot solutions



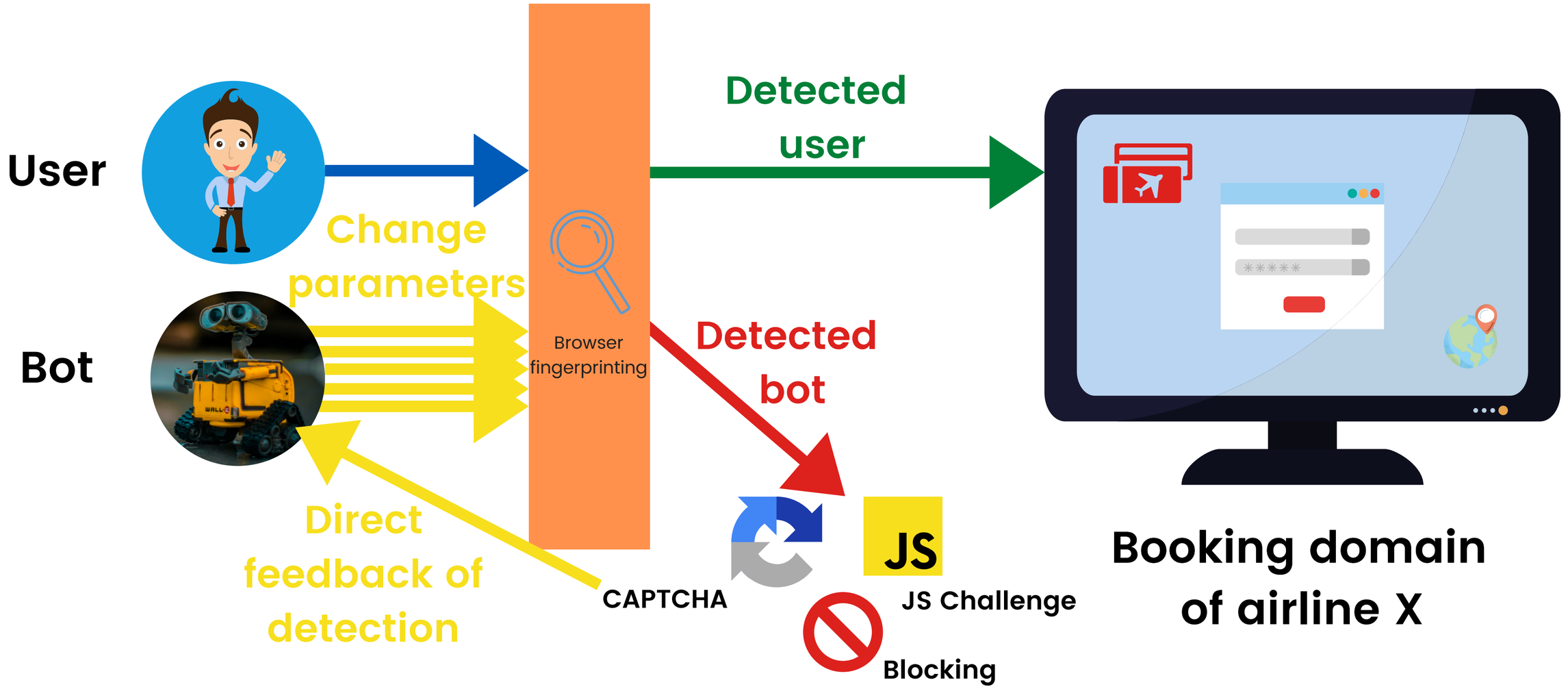
Anti-bot solutions



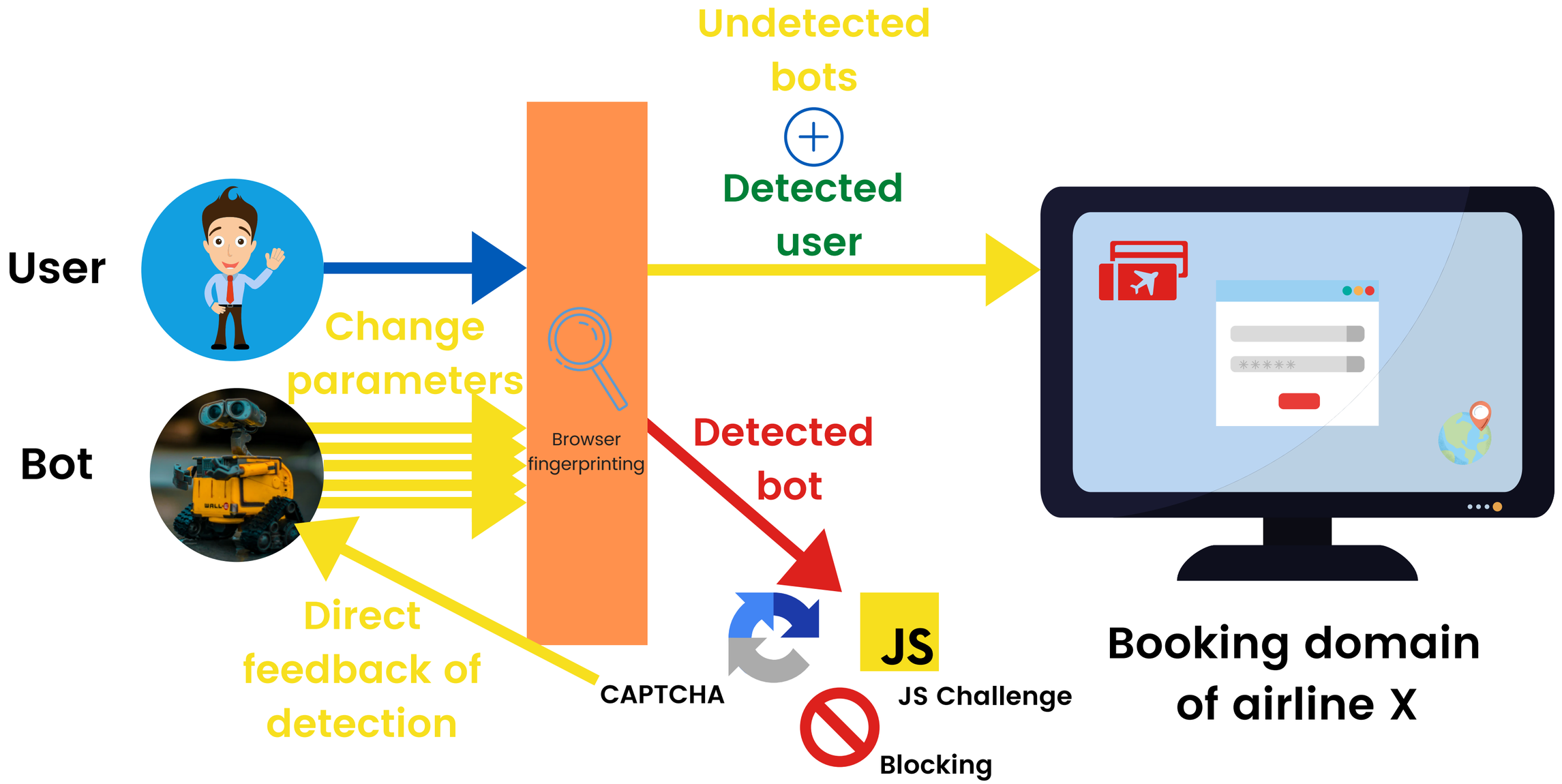
Anti-bot solutions



Anti-bot solutions



Anti-bot solutions



2. Honeypot

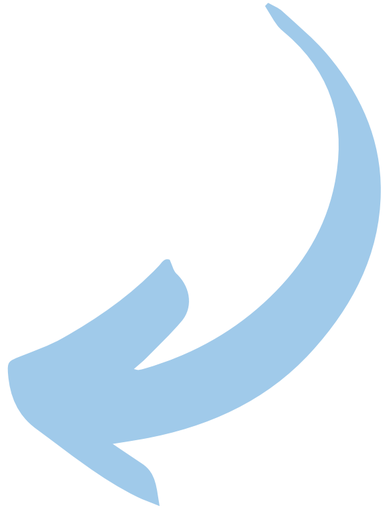
Initial idea

Initial idea

Prevent bots to know they have been detected & save costs for the provider

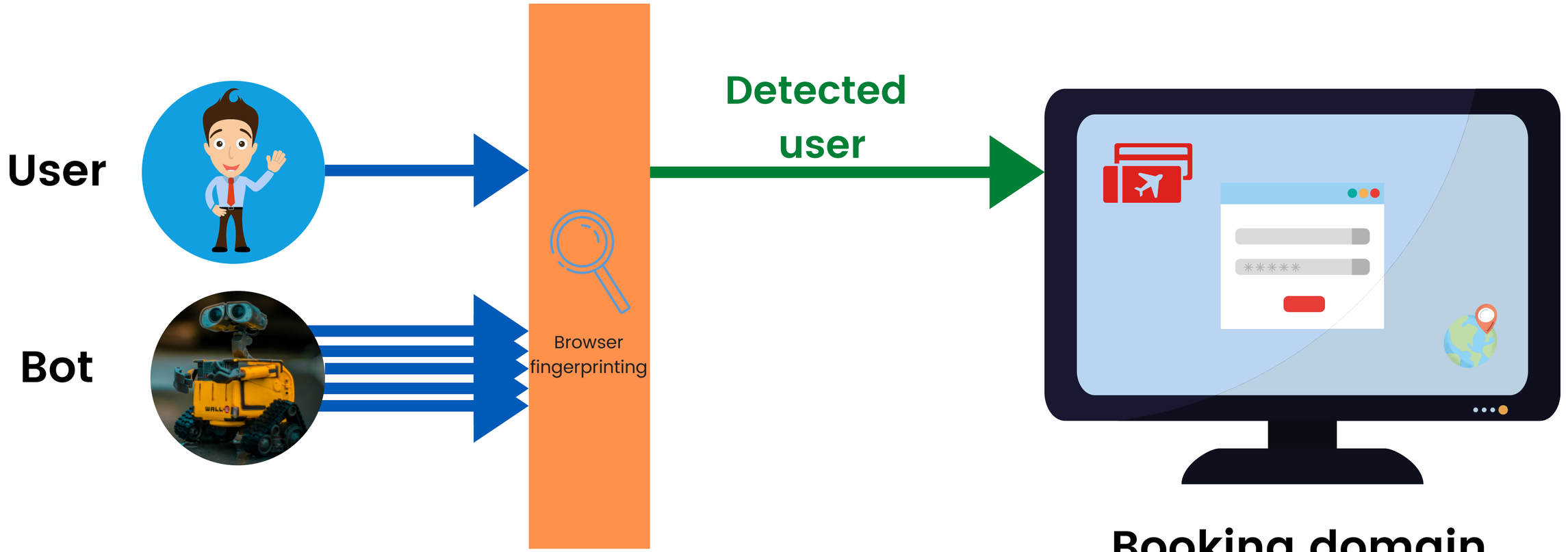
Initial idea

Prevent bots to know they have been detected & save costs for the provider

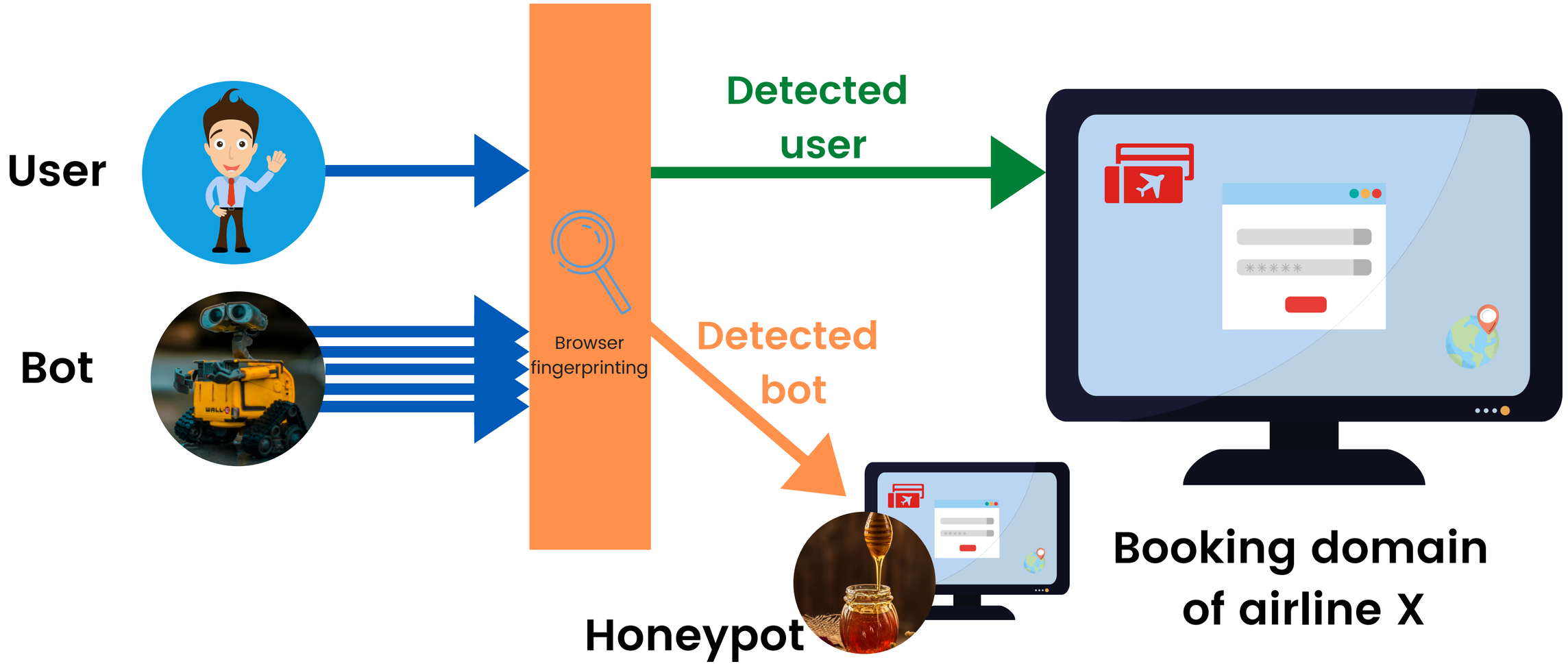


Provide bots incorrect but plausible answers

The idea


User
Bot
Browser fingerprinting
Detected user
Booking domain of airline X

The idea


User
Bot
Browser fingerprinting
Detected user
Detected bot
Booking domain of airline X
Honeypot

Pilot airline

- 1 millions requests per day

Pilot airline

- 1 millions requests per day
- 40% detected as bad bots

Pilot airline

- 1 millions requests per day
- 40% detected as bad bots
- Redirection of the traffic of specific bot signatures to the honeypot

Pricing strategy

After 3 days, modification of fares: increase 10% of the requests by 5%

Goal: understanding if the bot master was checking for anomalies in the price

Some questions…

Some questions...

- Is it possible to recognise a bot campaign from the information included in the payloads?

Some questions...

- Is it possible to recognise a bot campaign from the information included in the payloads?
- Are bots crafting payloads to detect the honeypot?

Some questions...

- Is it possible to recognise a bot campaign from the information included in the payloads?
- Are bots crafting payloads to detect the honeypot?

Some questions...

- Is it possible to recognise a bot campaign from the information included in the payloads?
- Are bots crafting payloads to detect the honeypot?
- Can we derive meaningful information studying the patterns of bot IPs?

Success criteria

Volume of traffic in the ranges before the case study, for at least 14 days after the fares modification

Results

- Bot signature: regular activity during different days and total redirection

Results

- Bot signature: regular activity during different days and total redirection
- Running for 56 days (interruption linked with COVID-19 restrictions on flights)

Results

- Bot signature: regular activity during different days and total redirection
- Running for 56 days (interruption linked with COVID-19 restrictions on flights)
- Reception of 22,991 HTTP requests at the Honeypot

Results

- Bot signature: regular activity during different days and total redirection
- Running for 56 days (interruption linked with COVID-19 restrictions on flights)
- Reception of 22,991 HTTP requests at the Honeypot
- No change of behavior from before and during the case-study

Lessons learned modifying values

No ground truth to compared returned values

Lessons learned modifying values

No ground truth to compared returned values

Plausibility check not sophisticated enough for small changes

Behavioral analysis

- 51,5% of requests for return flights

Behavioral analysis

- 51,5% of requests for return flights
- Return flights: 7 days period

Behavioral analysis

- 51,5% of requests for return flights
- Return flights: 7 days period
- Only 25 combination of departure and arrival airports, small fraction of the airline's offer

Behavioral analysis

- 51,5% of requests for return flights
- Return flights: 7 days period
- Only 25 combination of departure and arrival airports, small fraction of the airline's offer
- Homogeneous distribution of the time interval between departure and request date among different segments and request dates

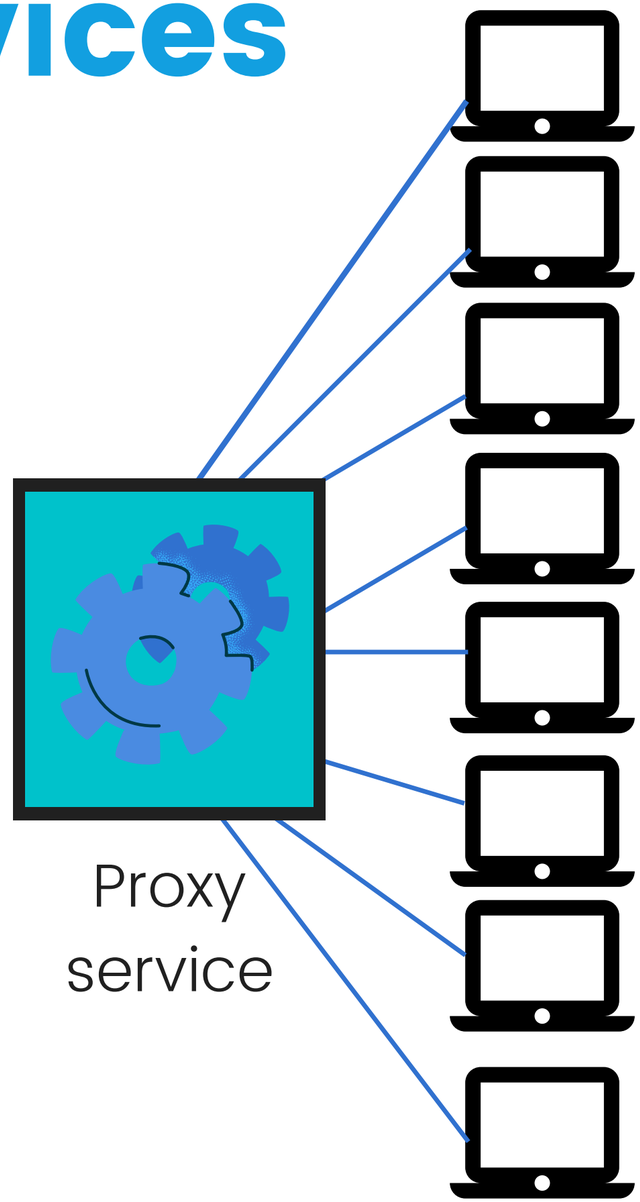
Behavioral analysis

- 51,5% of requests for return flights
- Return flights: 7 days period
- Only 25 combination of departure and arrival airports, small fraction of the airline's offer
- Homogeneous distribution of the time interval between departure and request date among different segments and request dates

3. Proxy services and IP addresses

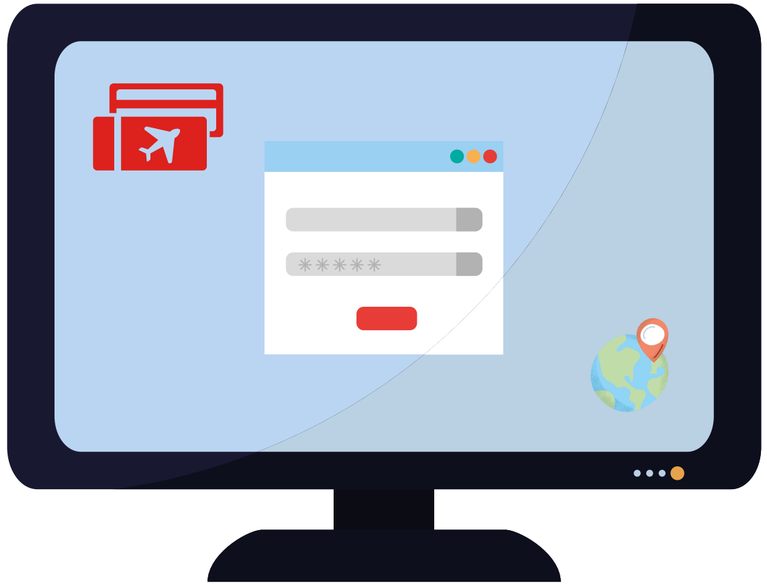
Proxy services

Bad bot



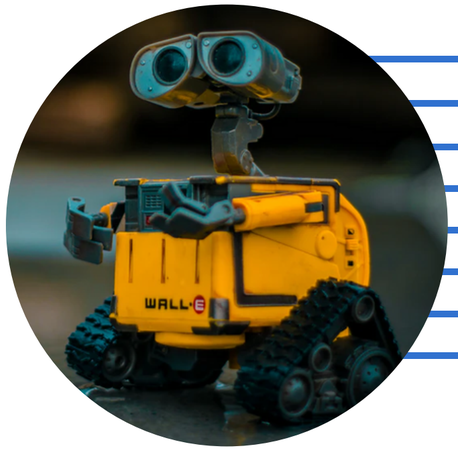
Proxy service

Proxy gateways



Airline booking domain

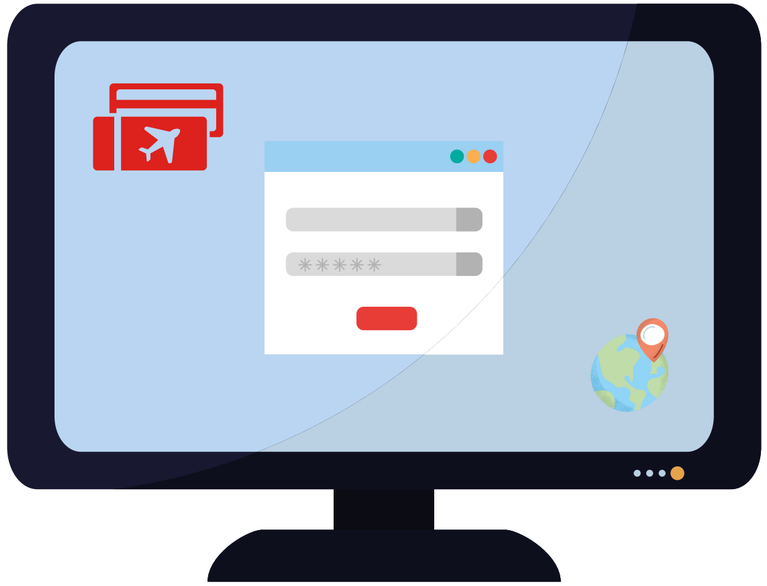
Proxy services



Bad bot

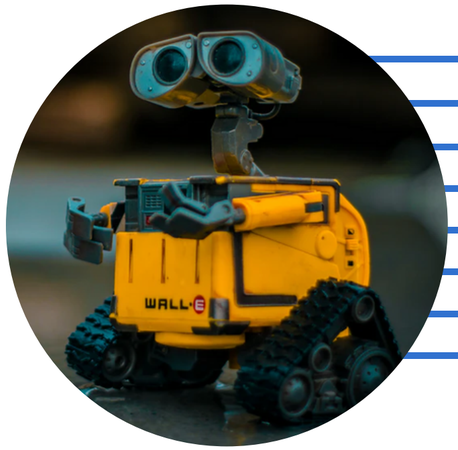
Proxy service

Proxy gateways



Airline booking domain

Proxy services



Bad bot

Proxy service

Proxy gateways

Airline booking domain

Bots using proxy service

Impossible to link bot campaign to organization

Bots using proxy service

Impossible to link bot campaign to organization

No need for private distributed infrastructure

Bots using proxy service

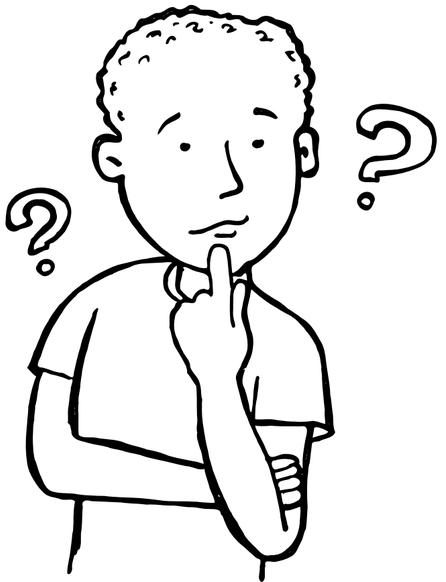
Impossible to link bot campaign to organization

No need for private distributed infrastructure

Impractical blocking IPs strategy

Question

Are our IP addresses coming from proxy services?



IPs reputation

- Bookings during the running time of the case-study but in dates different from the requests in the honeypot

IPs reputation

- Bookings during the running time of the case-study but in dates different from the requests in the honeypot

- Presence in blocklists (76%), IPQualityScore (72% suspicious behavior), Tor (72 IPs)

IPs reputation

- Bookings during the running time of the case-study but in dates different from the requests in the honeypot

- Presence in blocklists (76%), IPQualityScore (72% suspicious behavior), Tor (72 IPs)

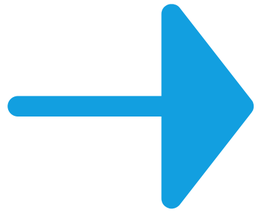
These IPs were doing malicious activities also outside our scope

IPs reputation

- Bookings during the running time of the case-study but in dates different from the requests in the honeypot

- Presence in blocklists (76%), IPQualityScore (72% suspicious behavior), Tor (72 IPs)

These IPs were doing malicious activities also outside our scope

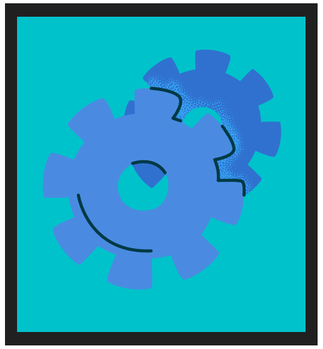


They were not allocated for the botnet only

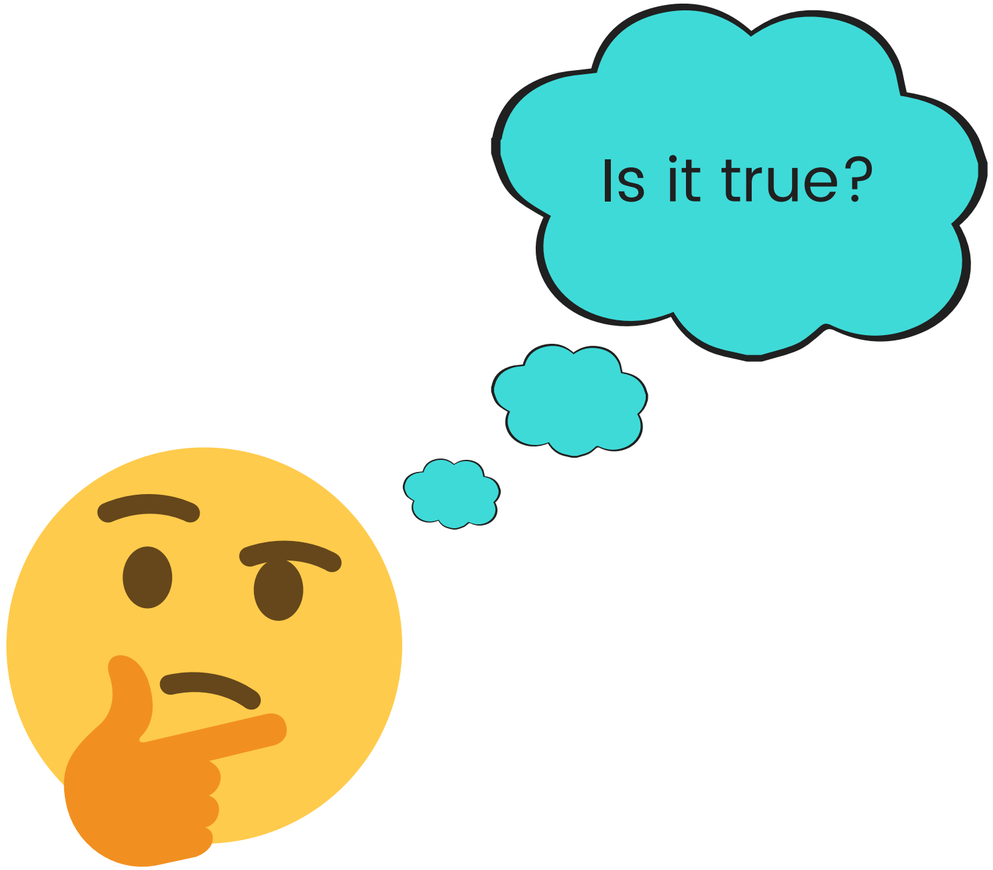
Proxy services claims

Proxy services claims

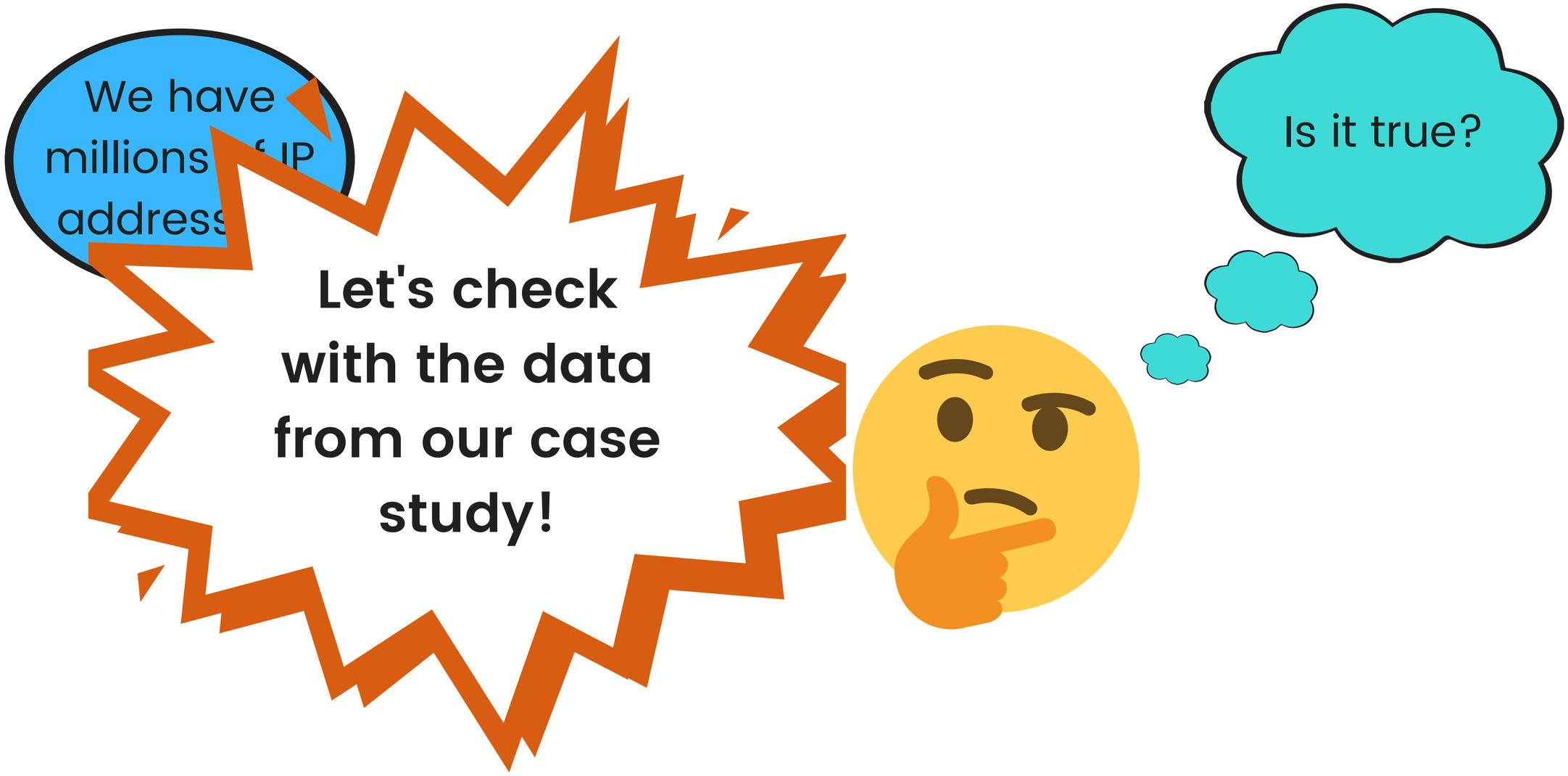
We have millions of IP addresses!



Proxy service



Proxy services claims



IP addresses study

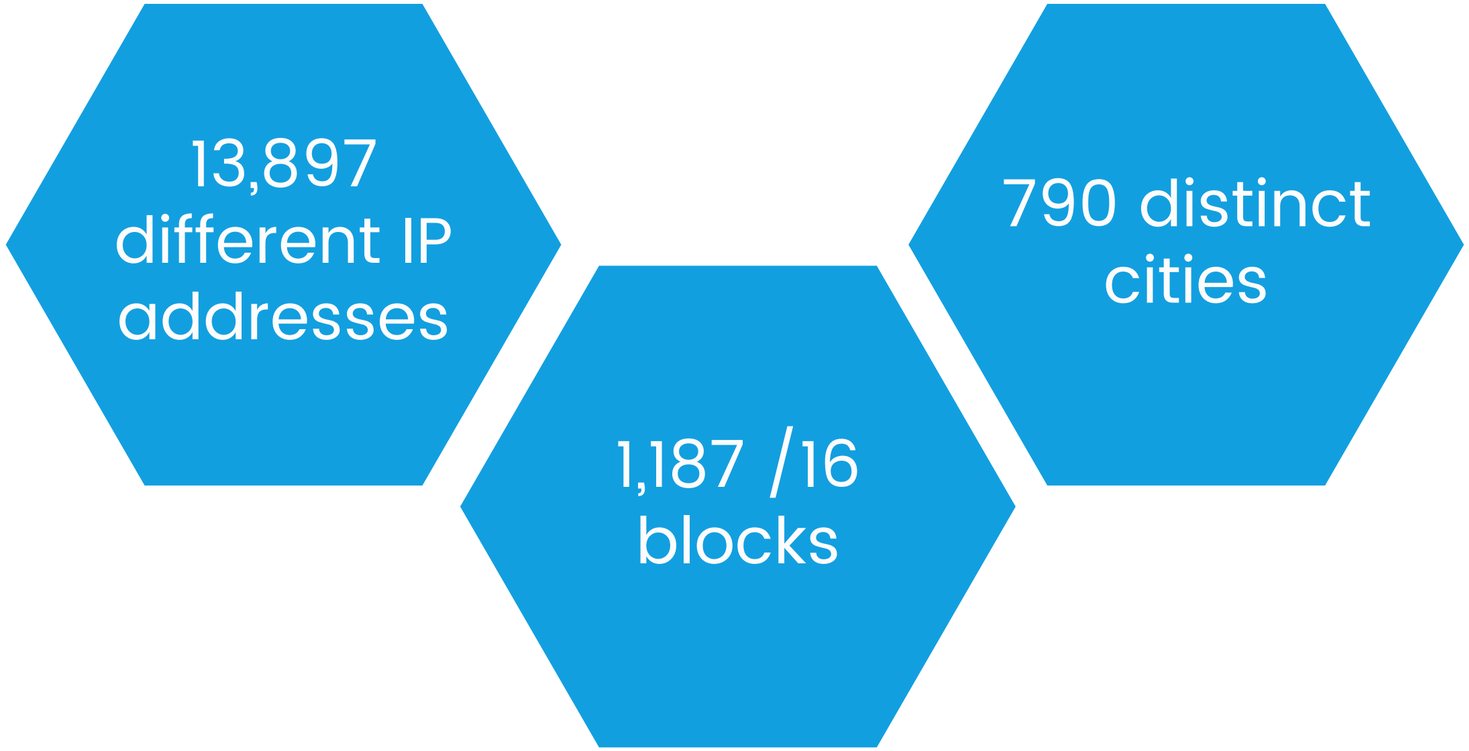
13,897 different IP addresses

IP addresses study

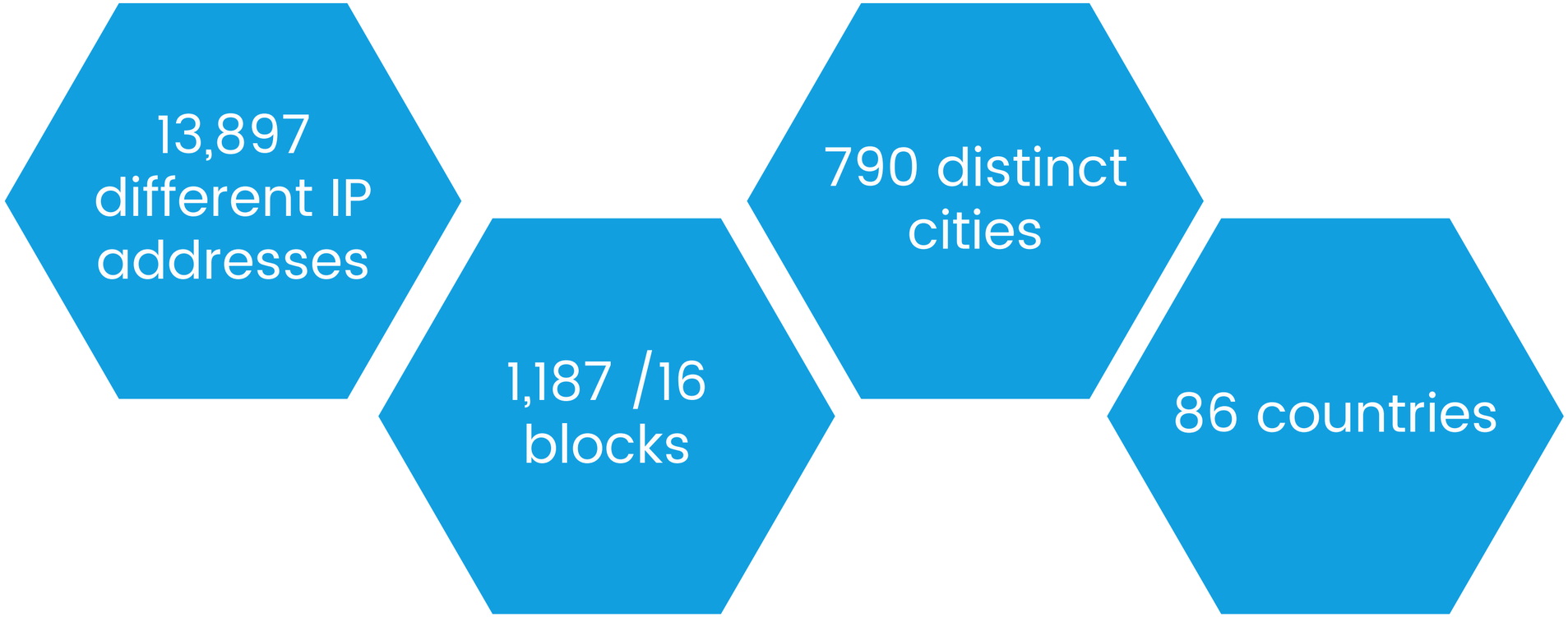
13,897 different IP addresses

1,187 /16 blocks

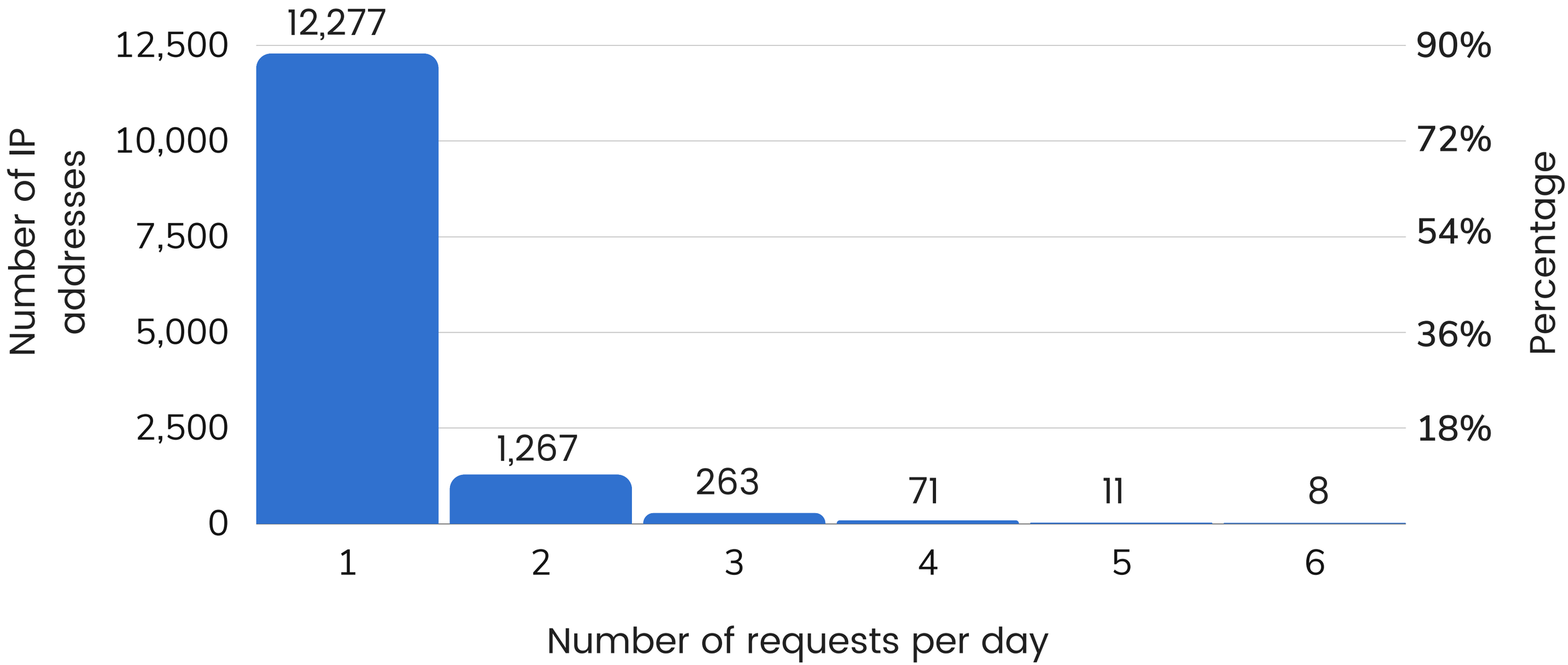
IP addresses study



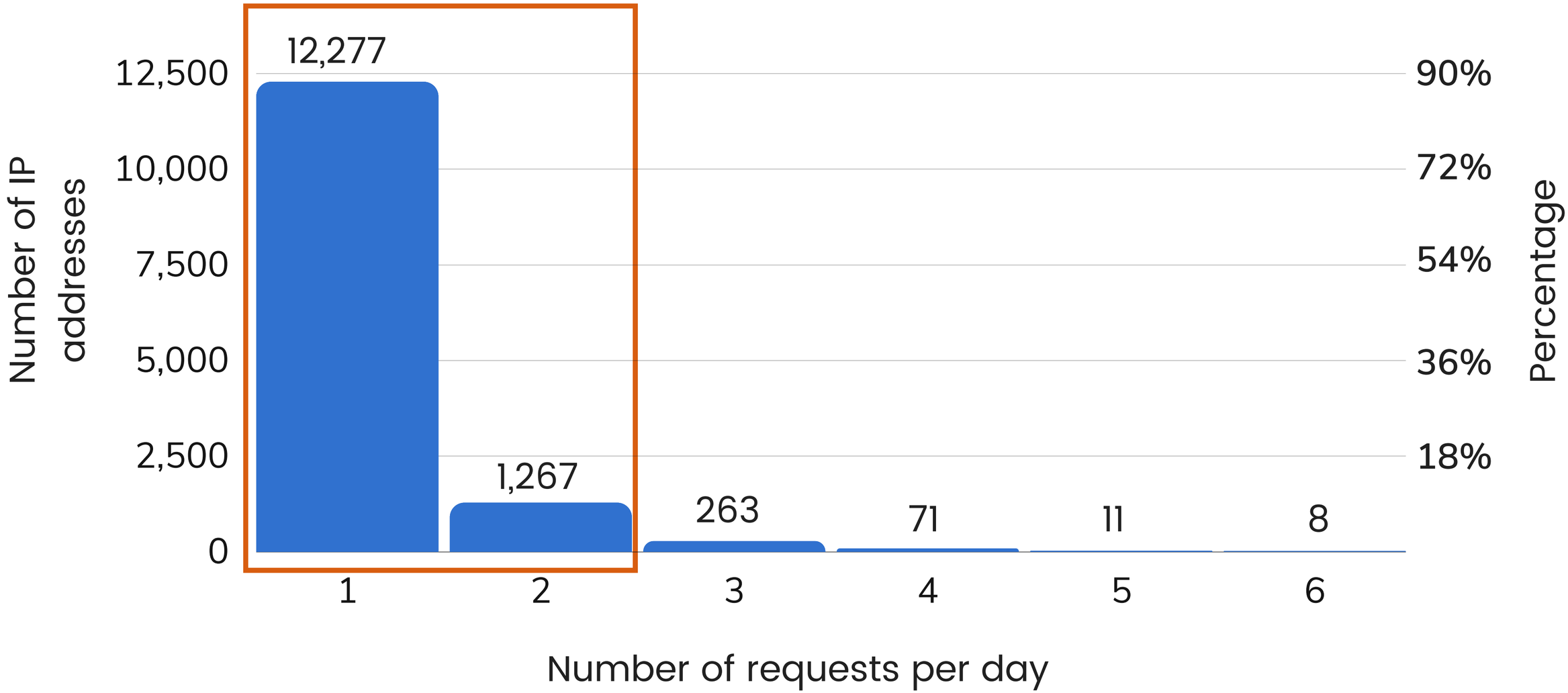
IP addresses study



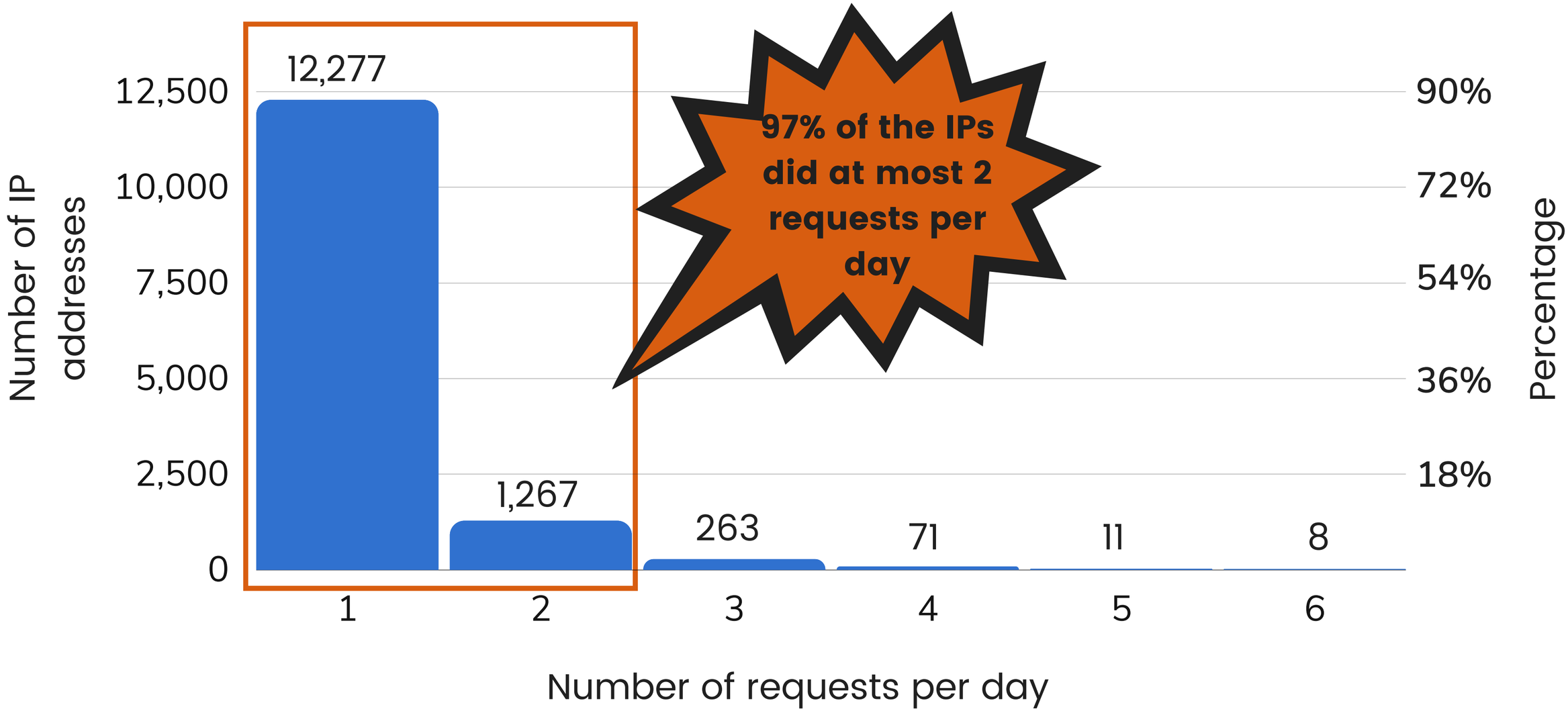
Daily number of requests per IP



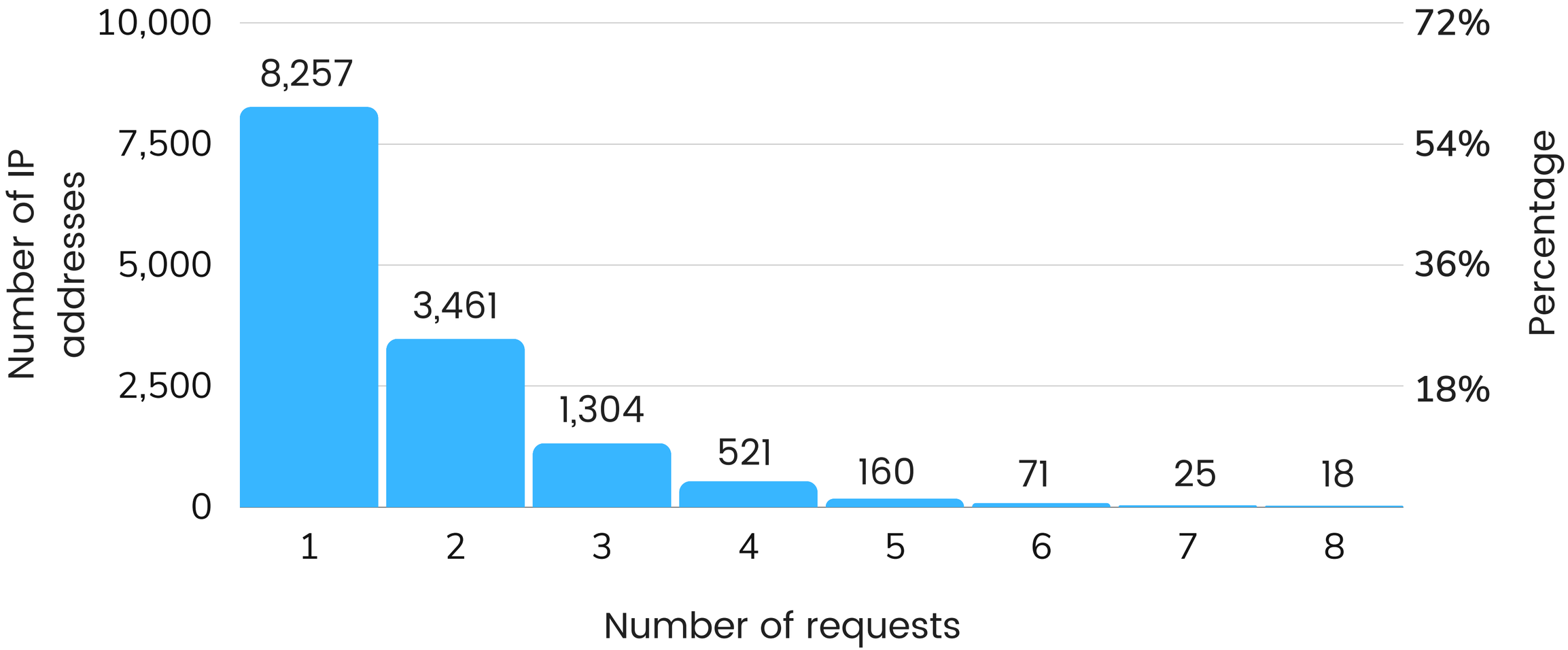
Daily number of requests per IP



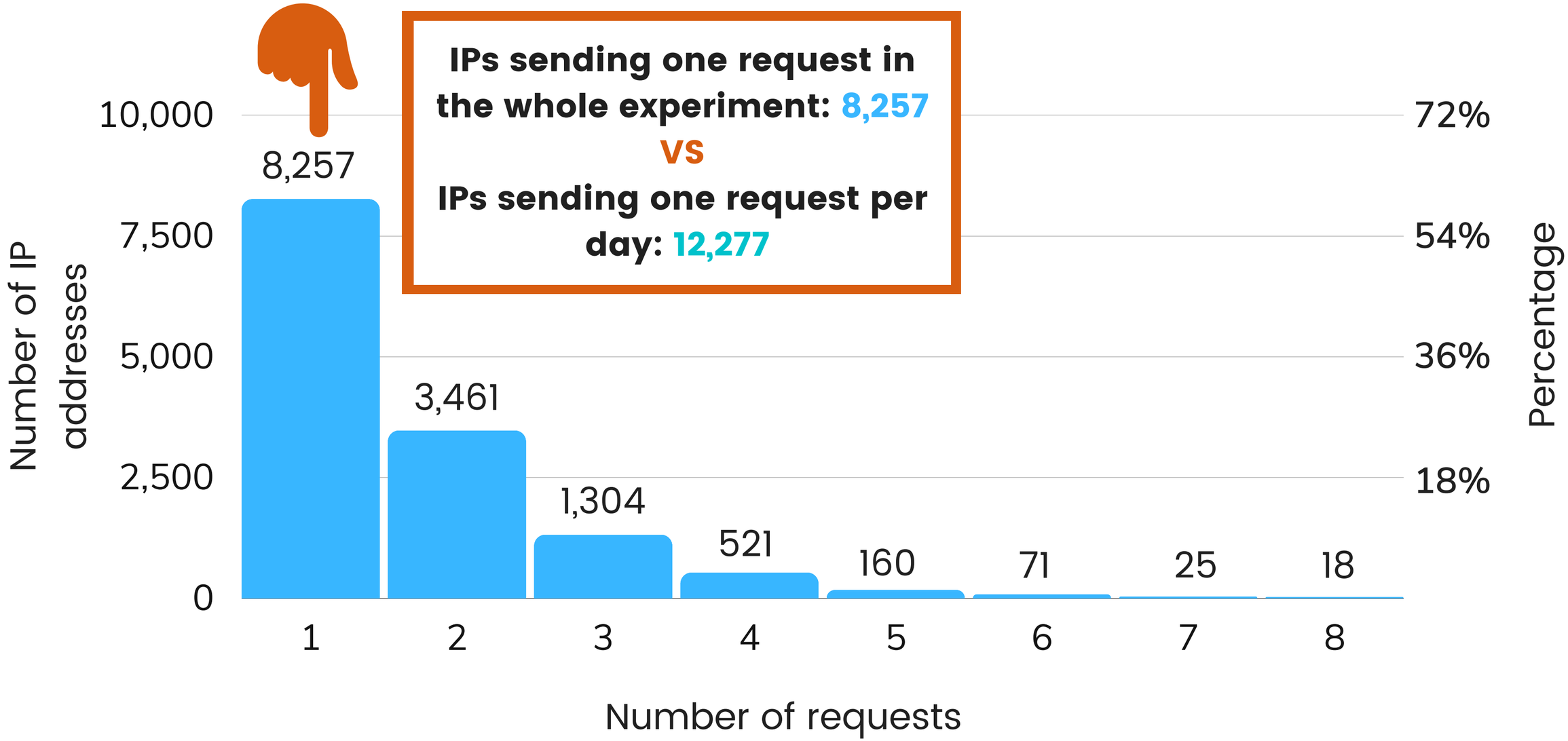
Daily number of requests per IP



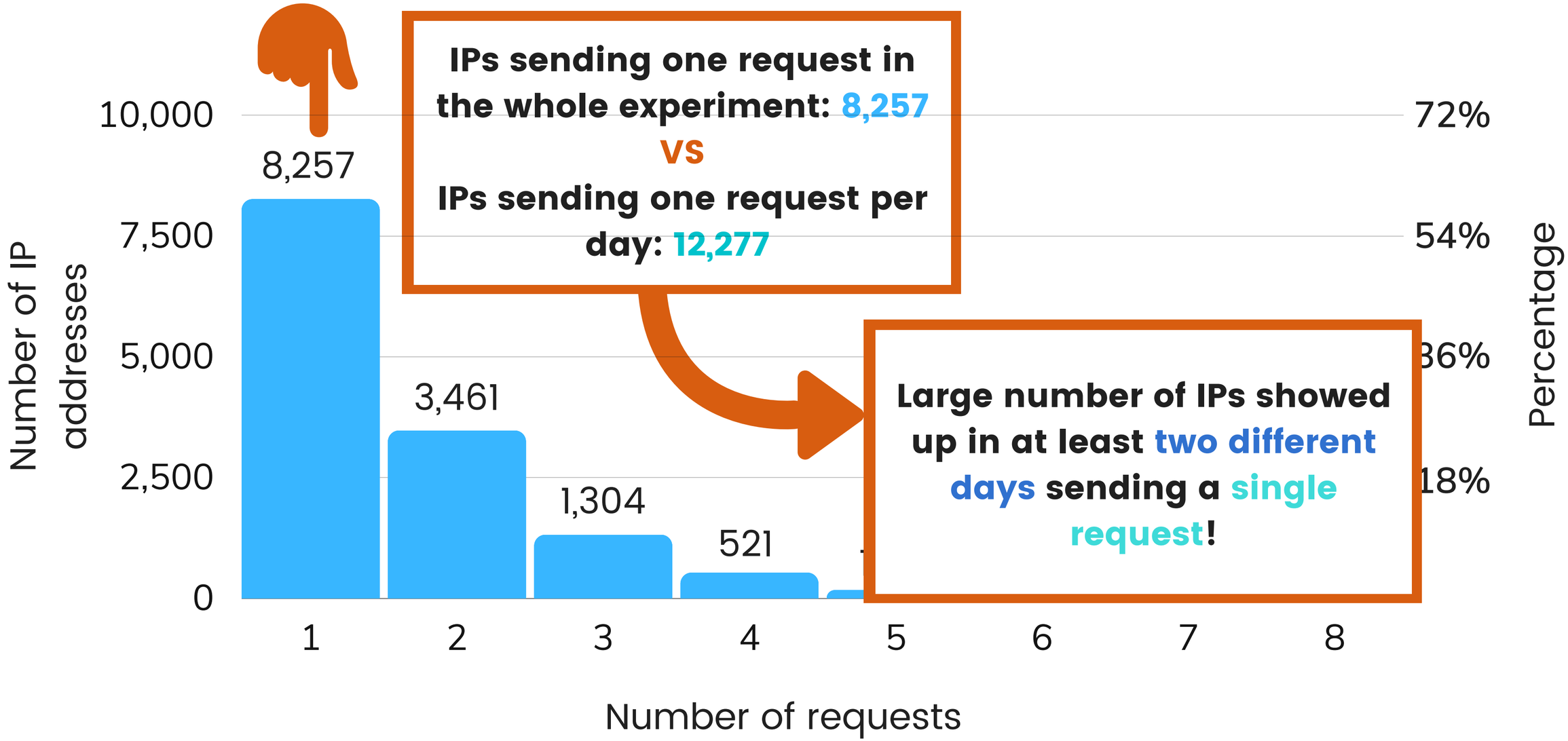
Maximum number of requests per IP



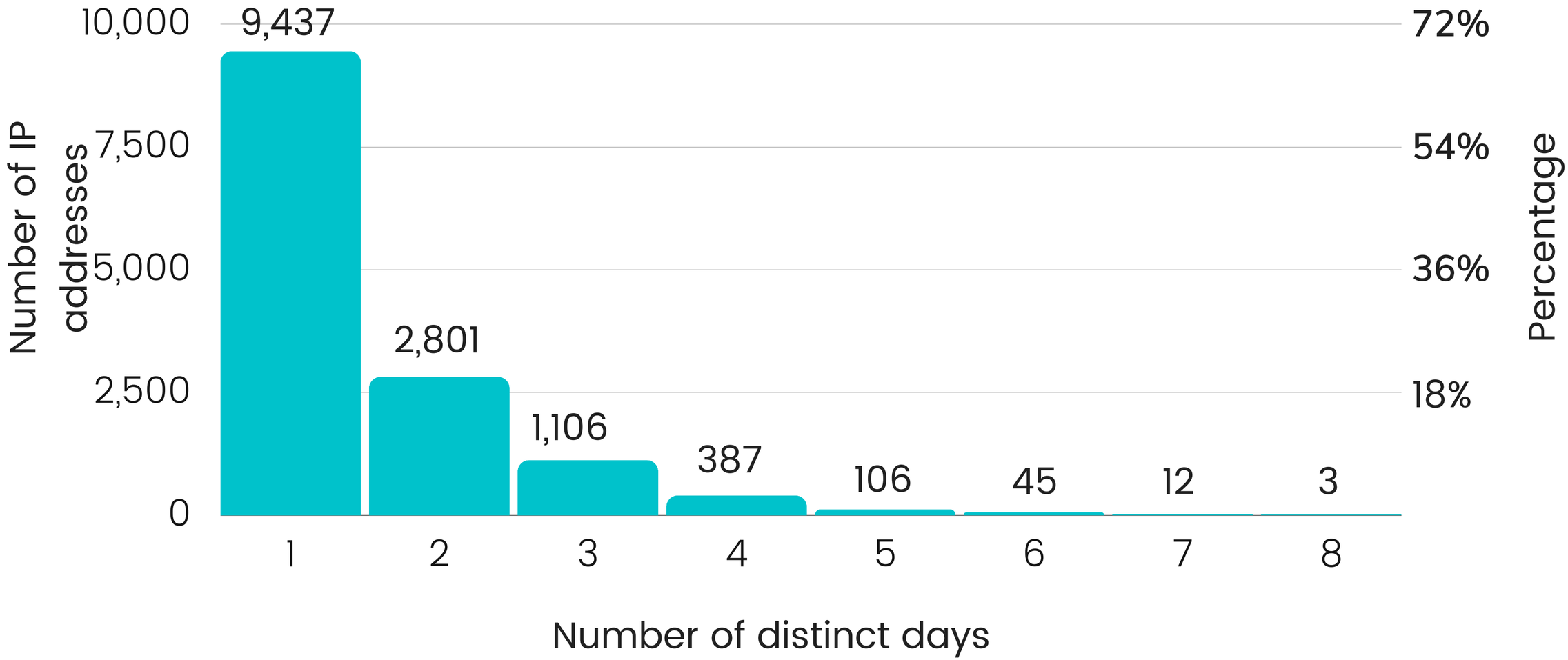
Maximum number of requests per IP



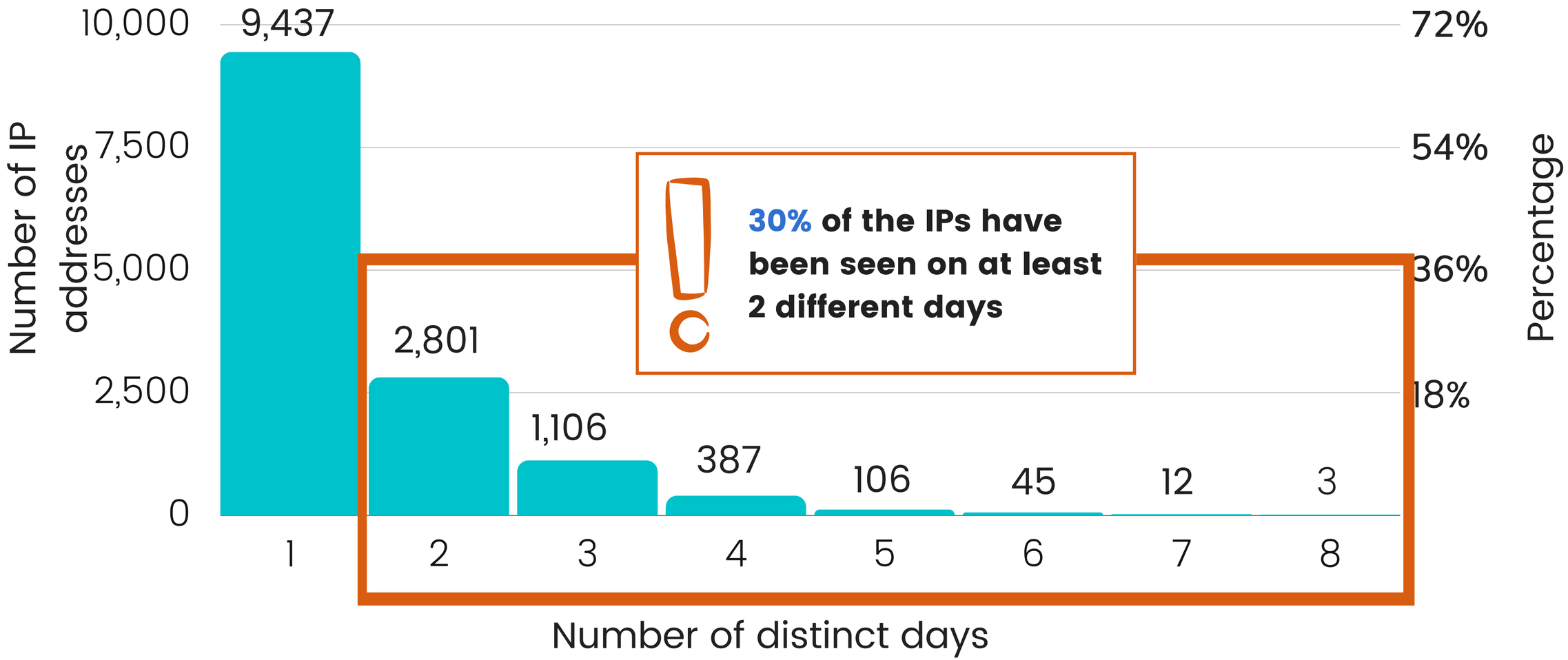
Maximum number of requests per IP



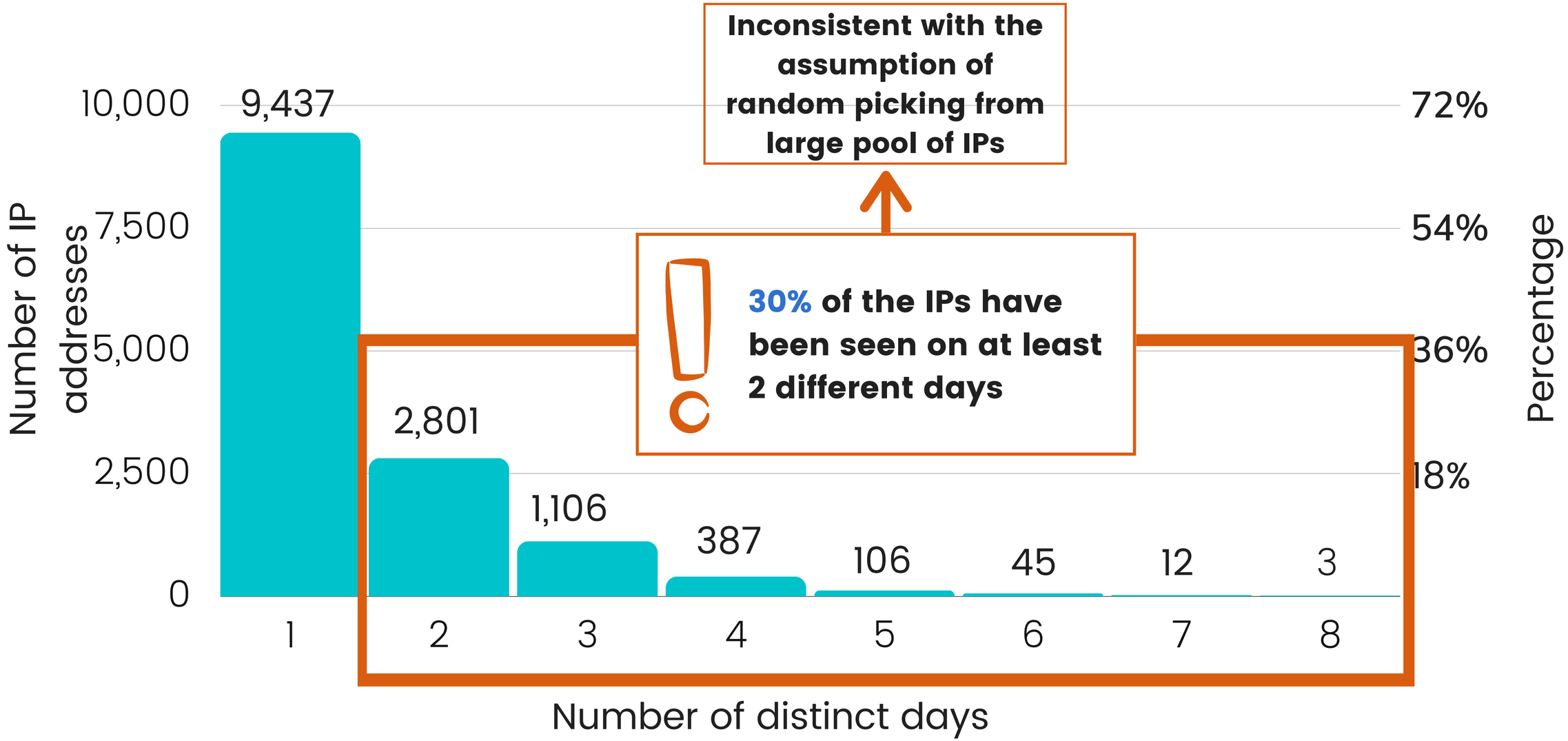
Number of IPs seen in distinct days



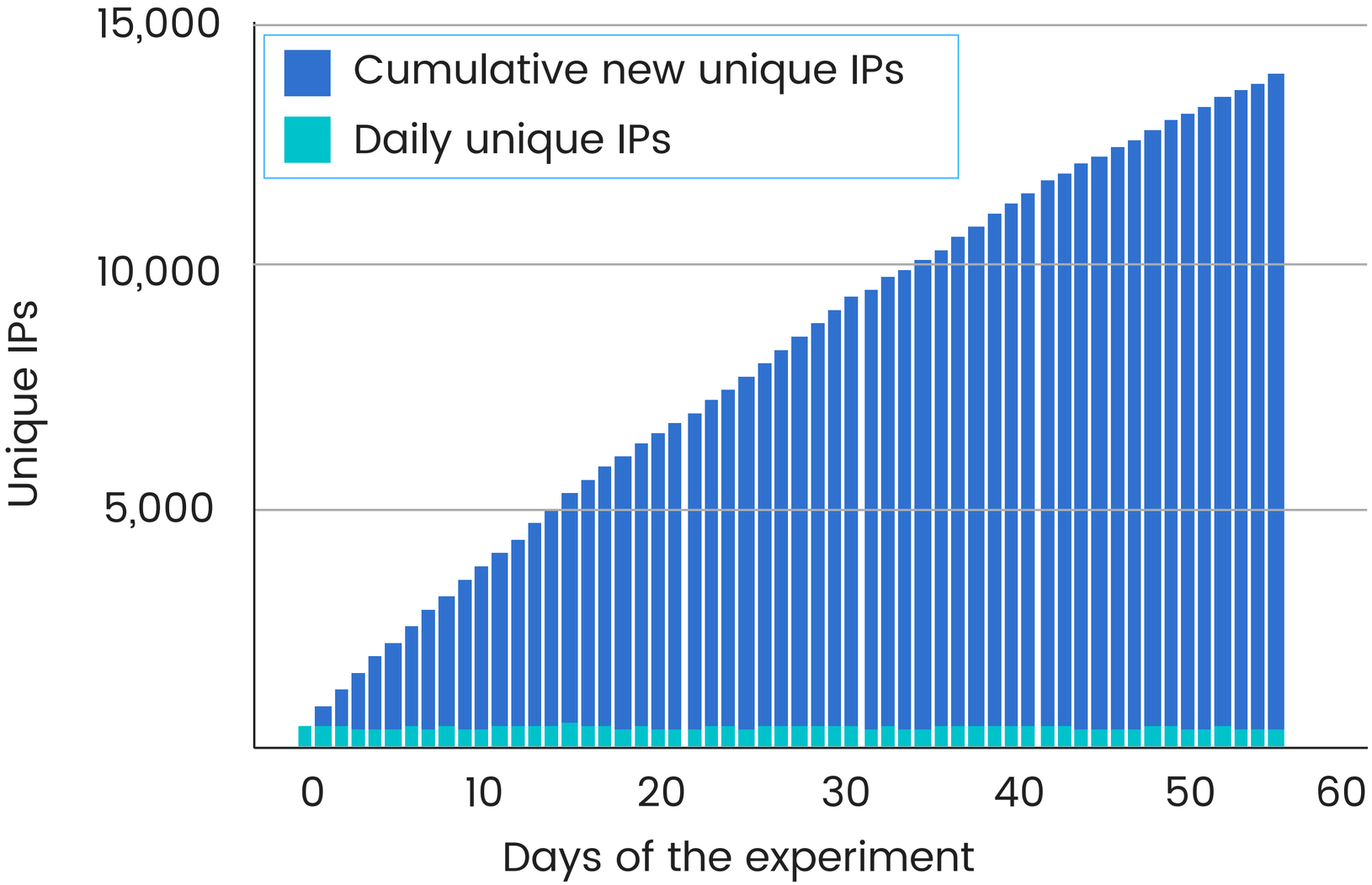
Number of IPs seen in distinct days



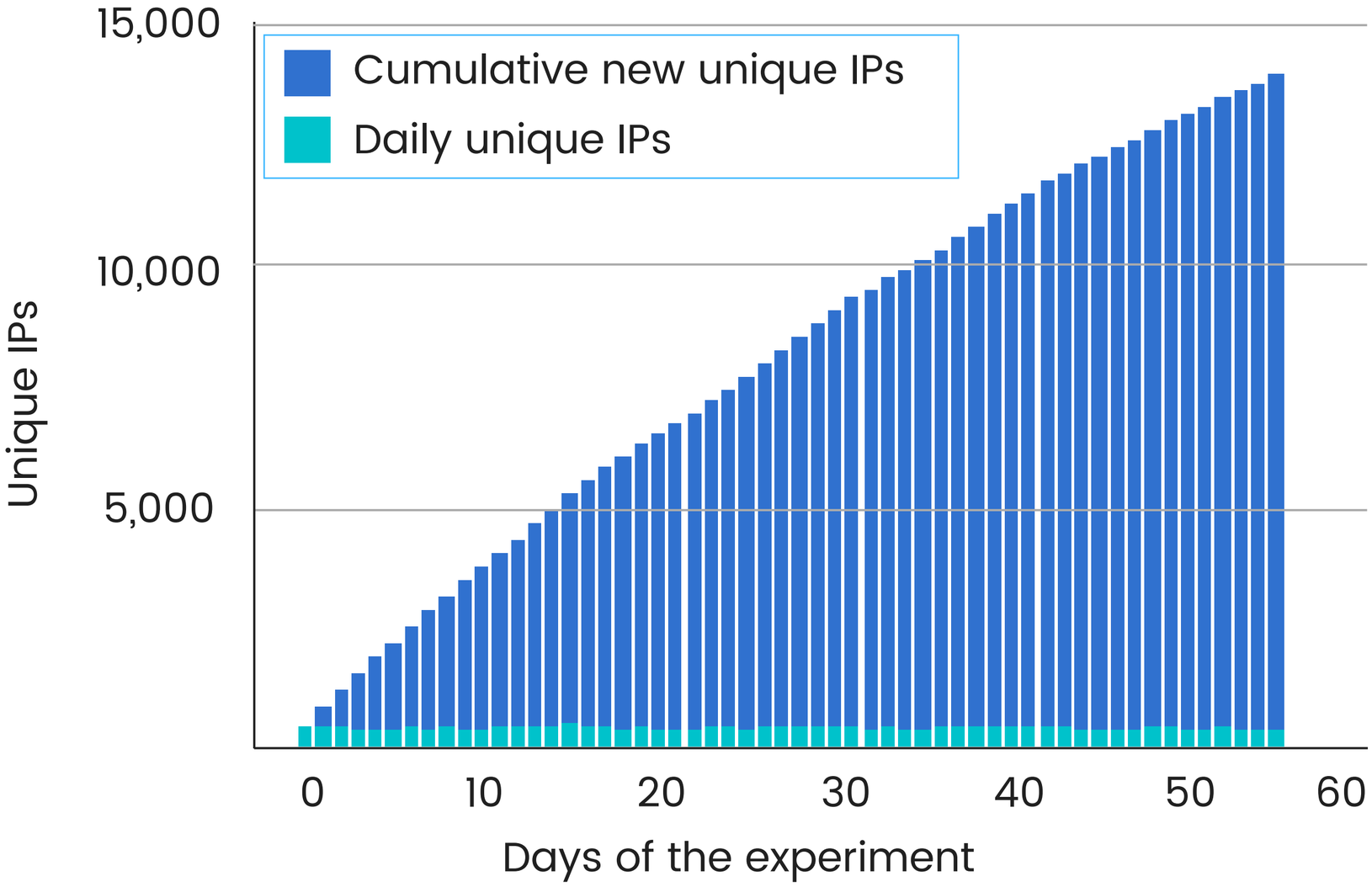
Number of IPs seen in distinct days



Cumulative curve of new unique IPs

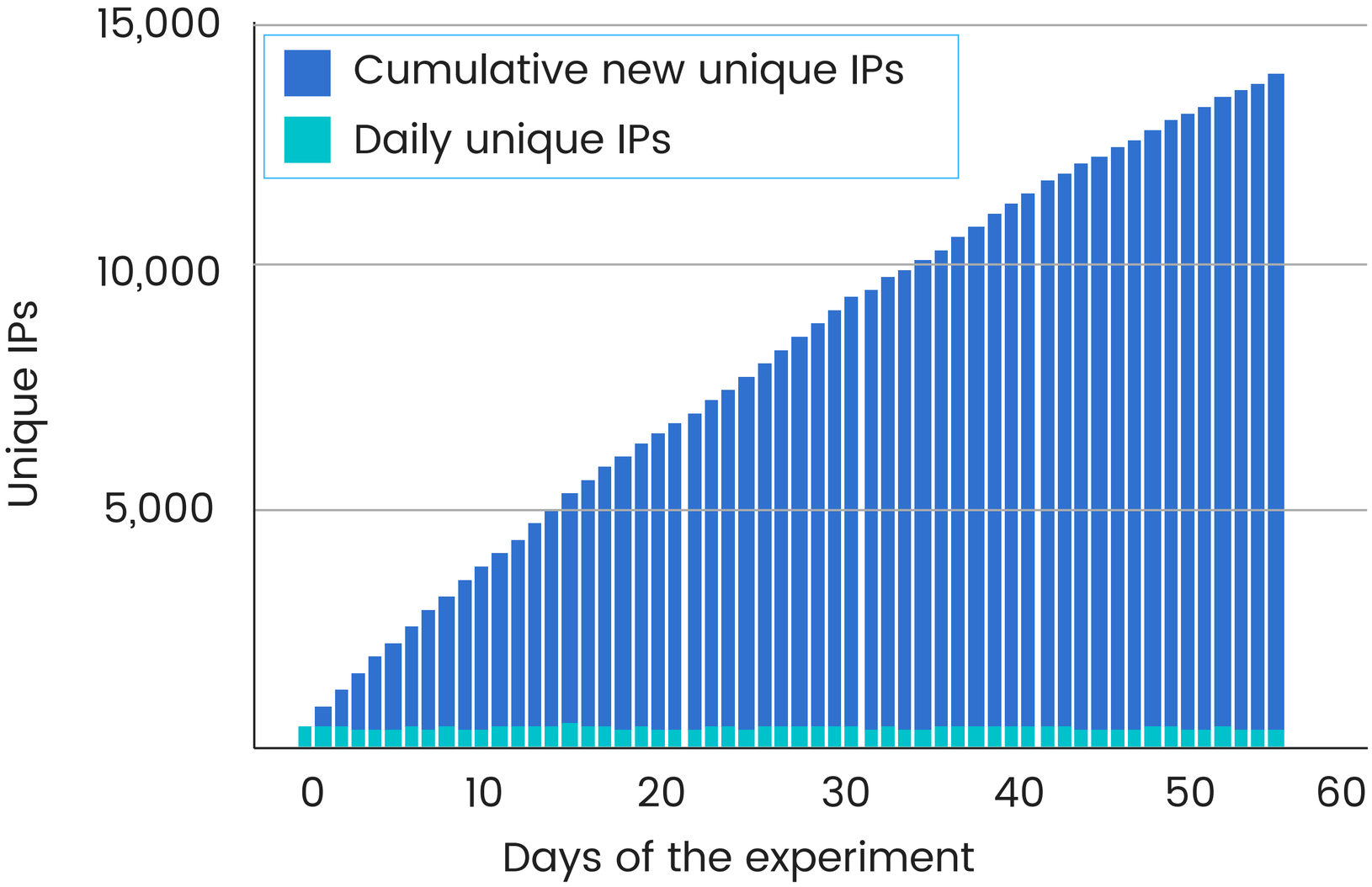


Cumulative curve of new unique IPs



The daily increment decreases over time

Cumulative curve of new unique IPs



The daily increment decreases over time

Eventually it will reach a maximum!

Modeling

1

IP assignment as a drawing process
Modeling the drawing process of IPs, looking for a probability distribution for our results and deriving the value of P.

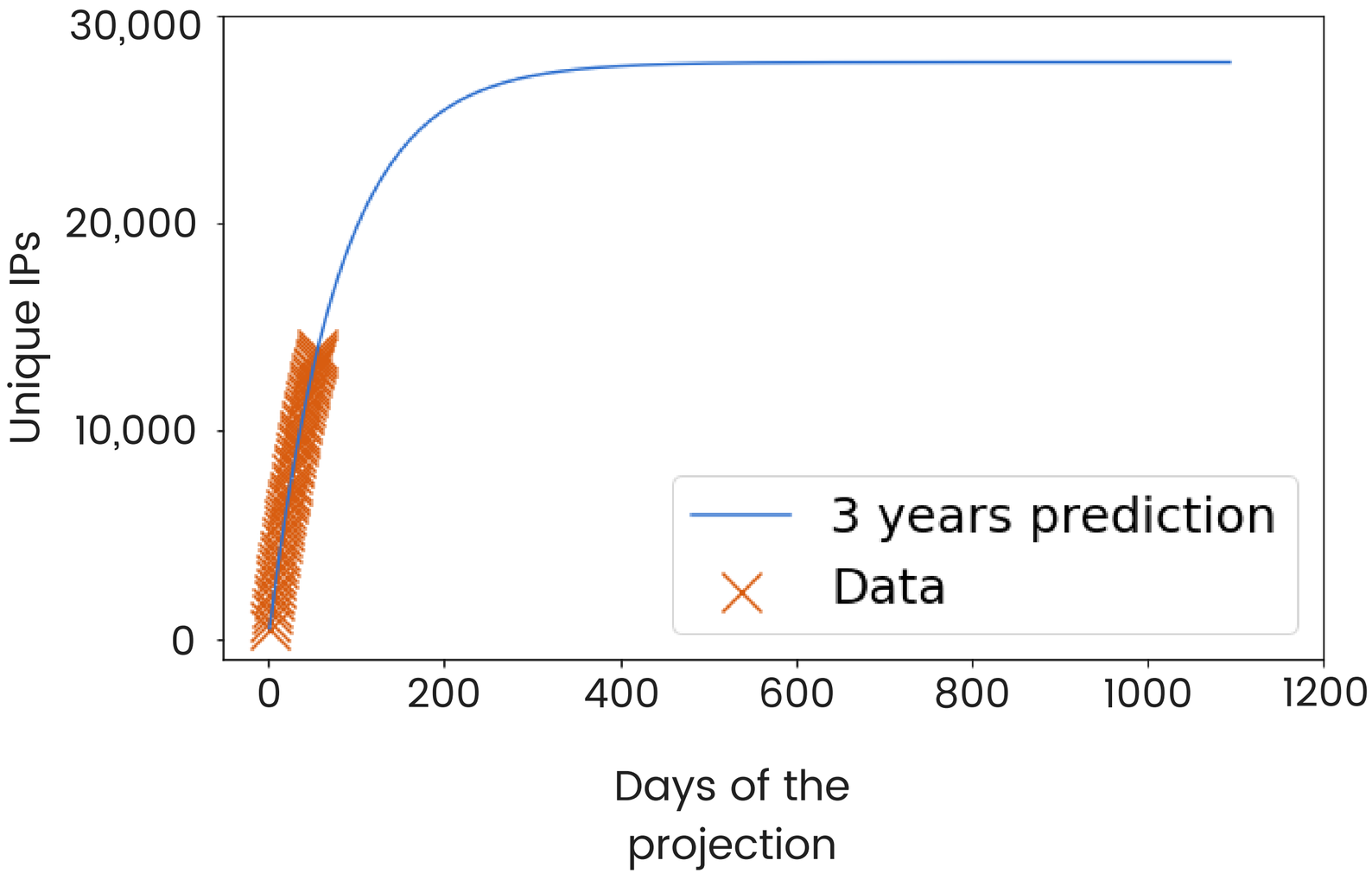
2

Fitting the cumulative curve of new unique IPs
Fitting the curve, extrapolating and finding what maximum value can be reached and when.

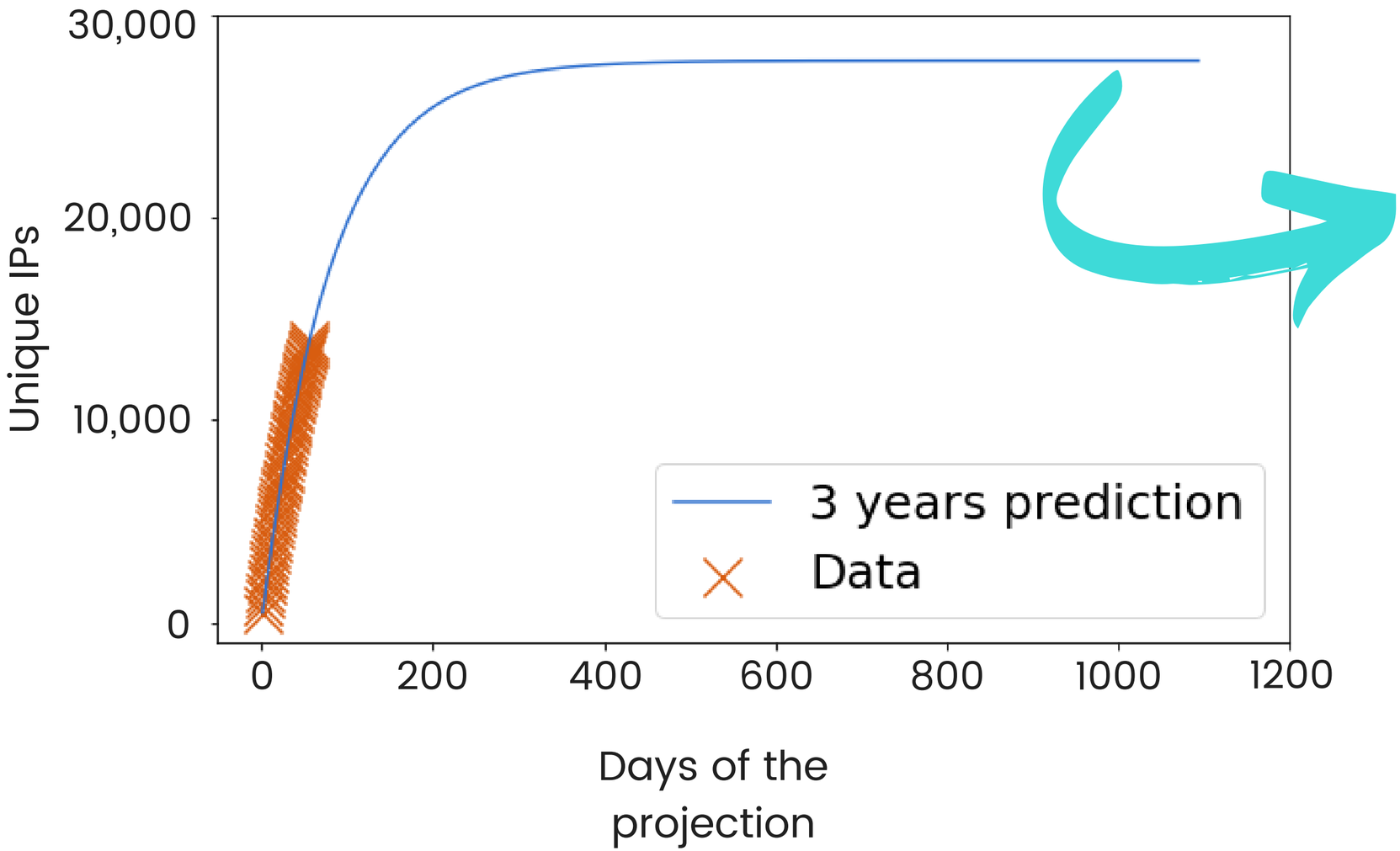
Modelling IP assignement results

- **Uniform distribution**: best population size equal to 20,000
- Uniform distribution means random picking
  - All IPs available all the time
  - Selection done without taking into account any condition, e.g. geo-localization
- **Gaussian distribution**: best population size equal to 60,000
- **Beta distribution**: best population size equal to 60,000

Fitting results and projection

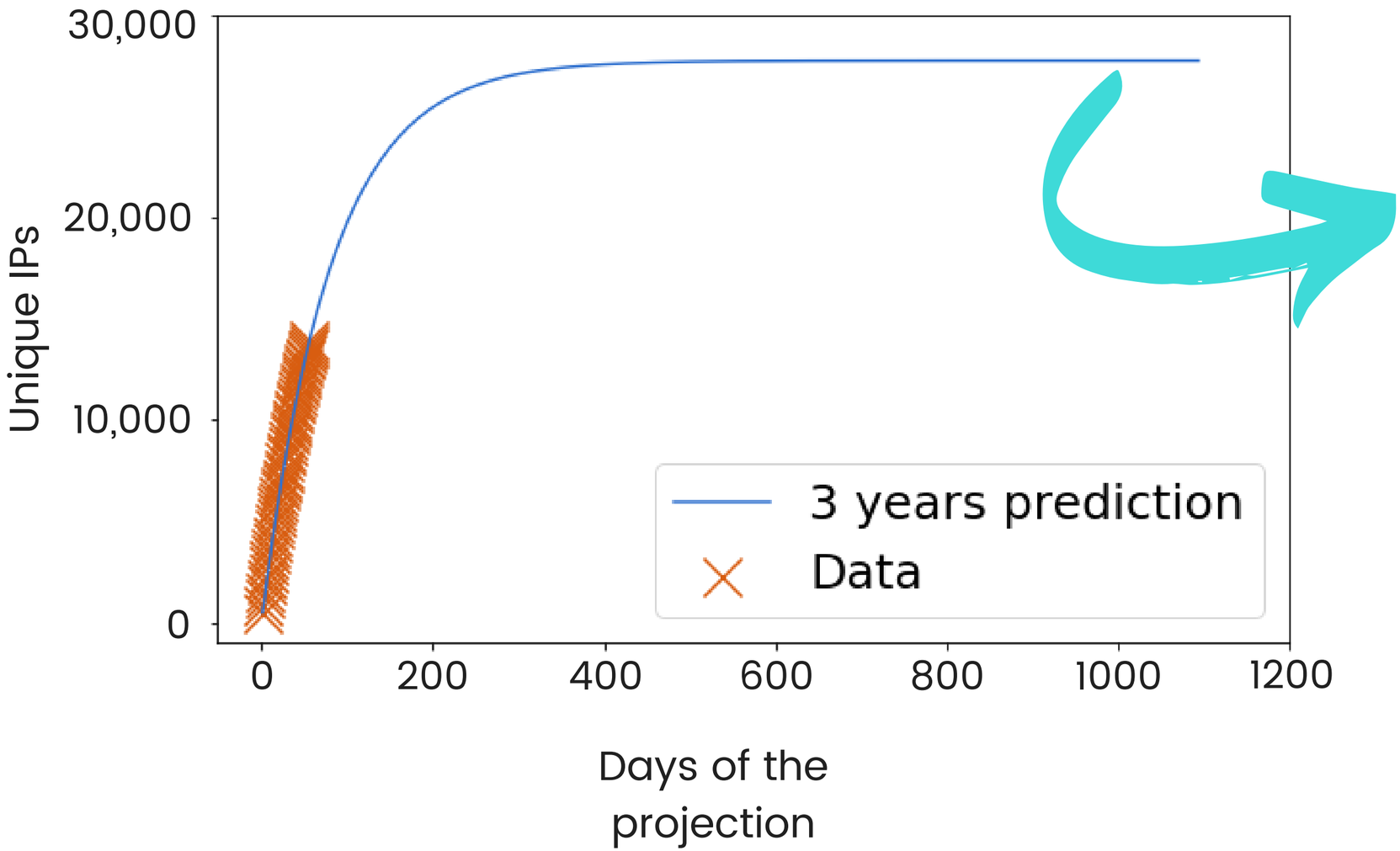


Fitting results and projection



The plateau is less than 30,000 IPs

Fitting results and projection



The plateau is less than 30,000 IPs

Consistent with the previous approach

What does it mean?

For all considered distributions, the found size of the pool is significantly smaller than proxies claim.

What does it mean?

- For all considered distributions, the found size of the pool is significantly smaller than proxies claim.

- This does not directly mean the proxies do not own millions of IPs.

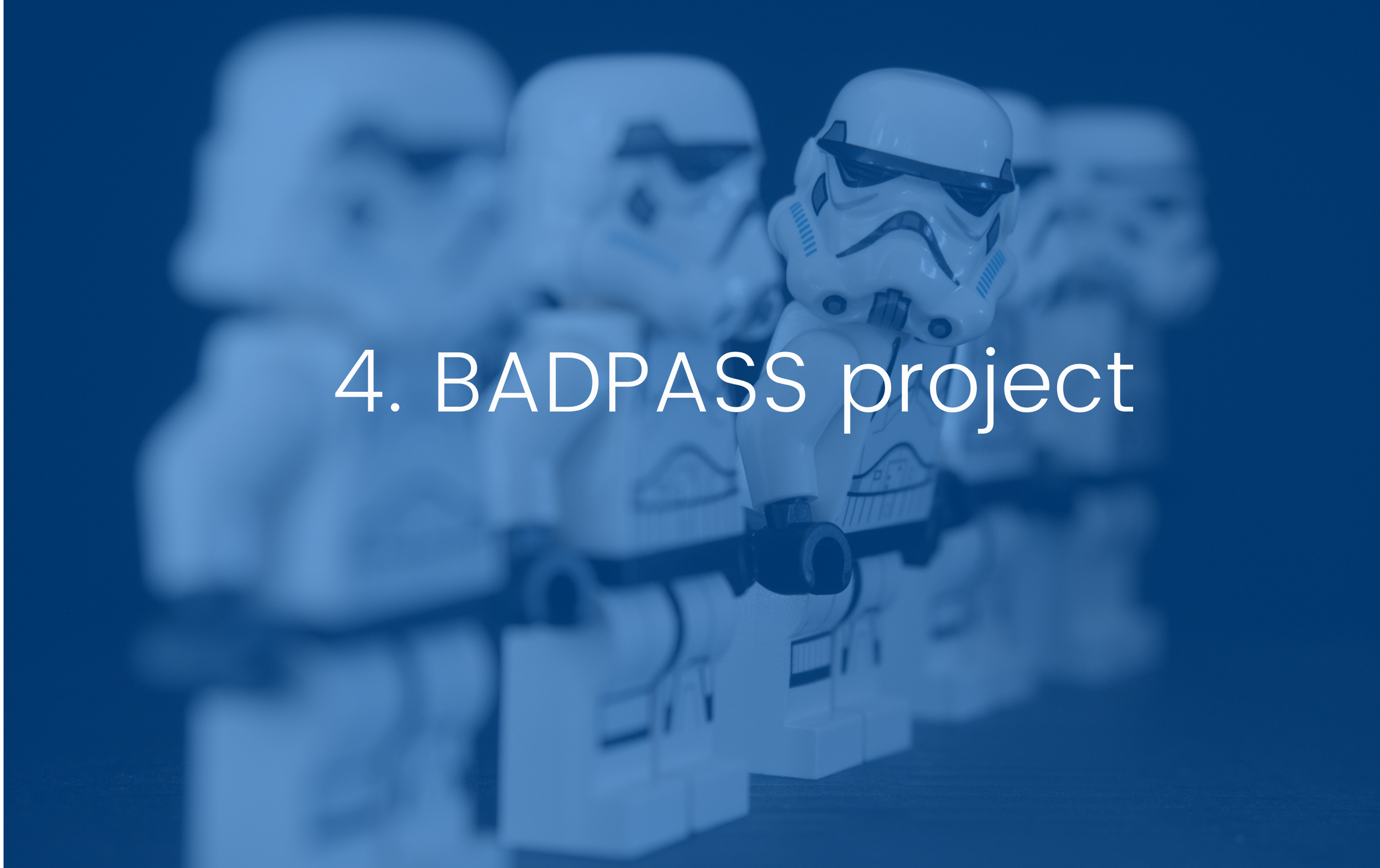
What does it mean?

- For all considered distributions, the found size of the pool is significantly smaller than proxies claim.
- This does not directly mean the proxies do not own millions of IPs.
- BUT it suggests that there is a not a complete allocation of the IPs.

What does it mean?

- For all considered distributions, the found size of the pool is significantly smaller than proxies claim.
- This does not directly mean the proxies do not own millions of IPs.
- BUT it suggests that there is a not a complete allocation of the IPs.

The number of IPs that we receive
it is not in the range of millions!



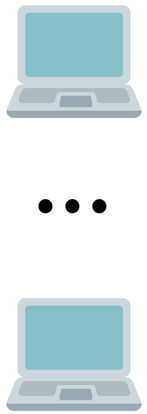
4. BADPASS project

BAD PASS: Bots taking ADvantage of Proxy AS a Services

- In-depth study of proxies ecosystems

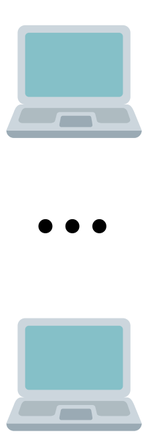
BAD PASS: Bots taking ADvantage of Proxy AS a Services

- In-depth study of proxies ecosystems

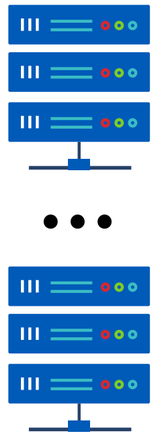


BAD PASS: Bots taking ADvantage of Proxy AS a Services

- In-depth study of proxies ecosystems



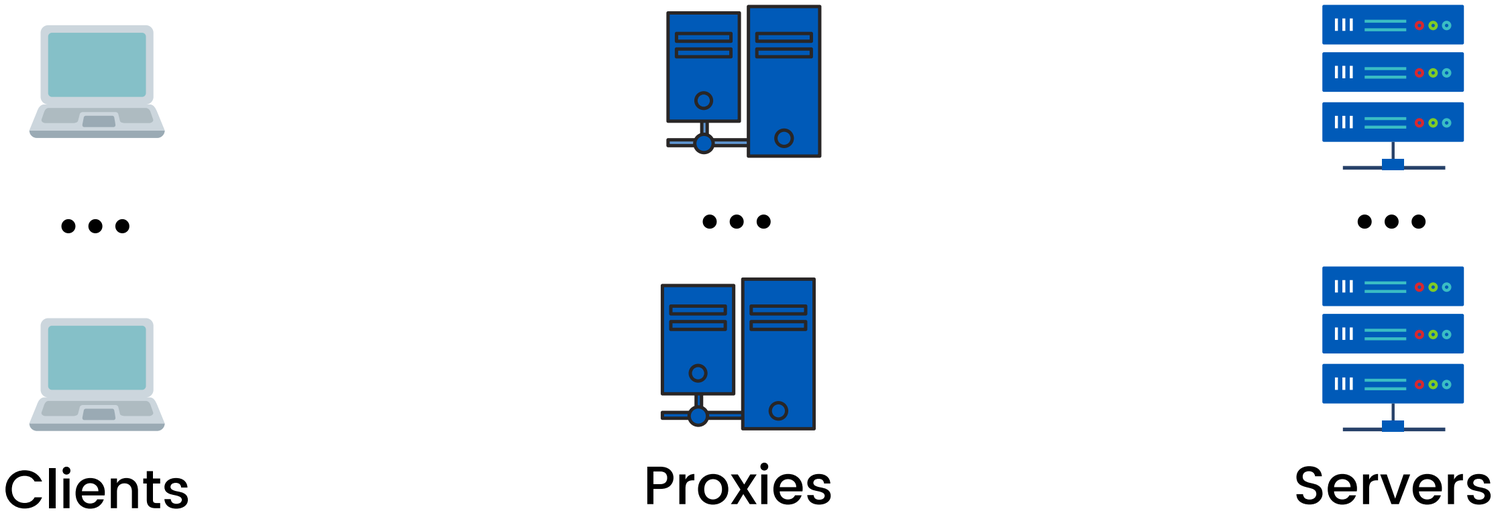
Clients



Servers

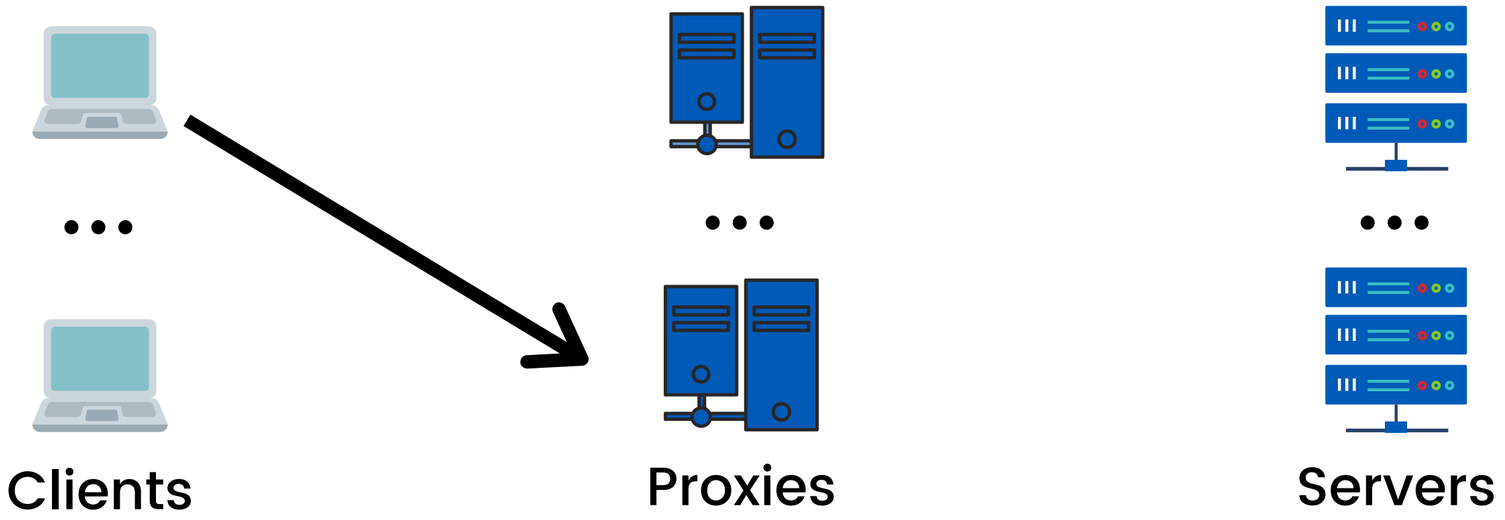
BAD PASS: Bots taking ADvantage of Proxy AS a Services

- In-depth study of proxies ecosystems



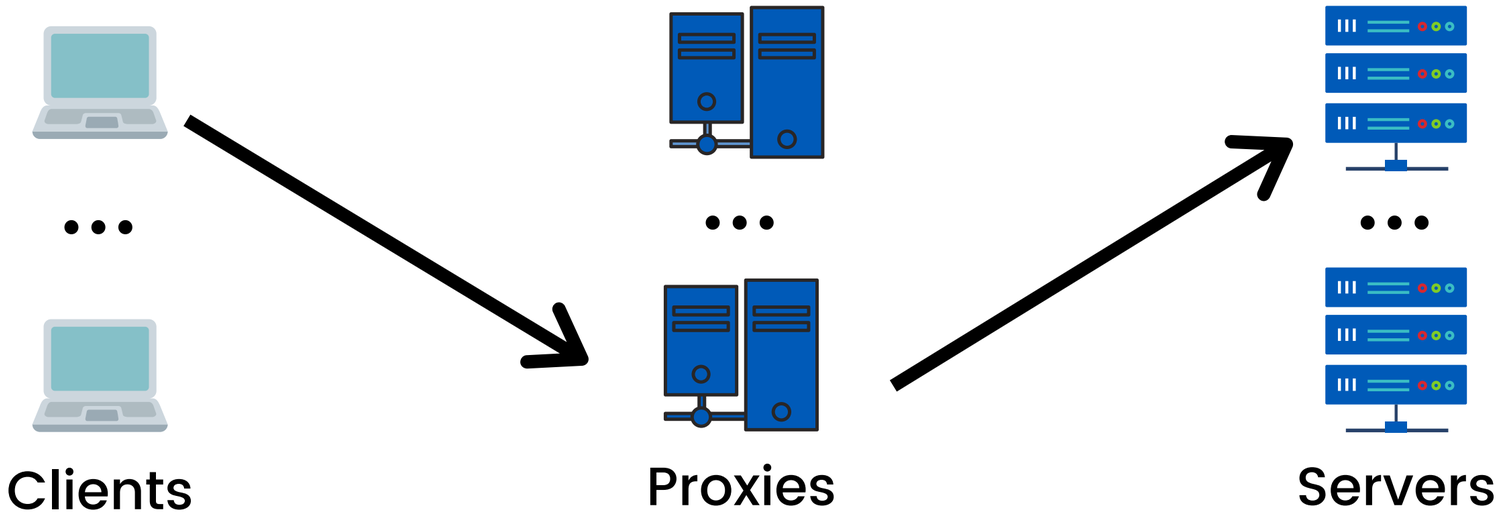
BAD PASS: Bots taking ADvantage of Proxy AS a Services

- In-depth study of proxies ecosystems



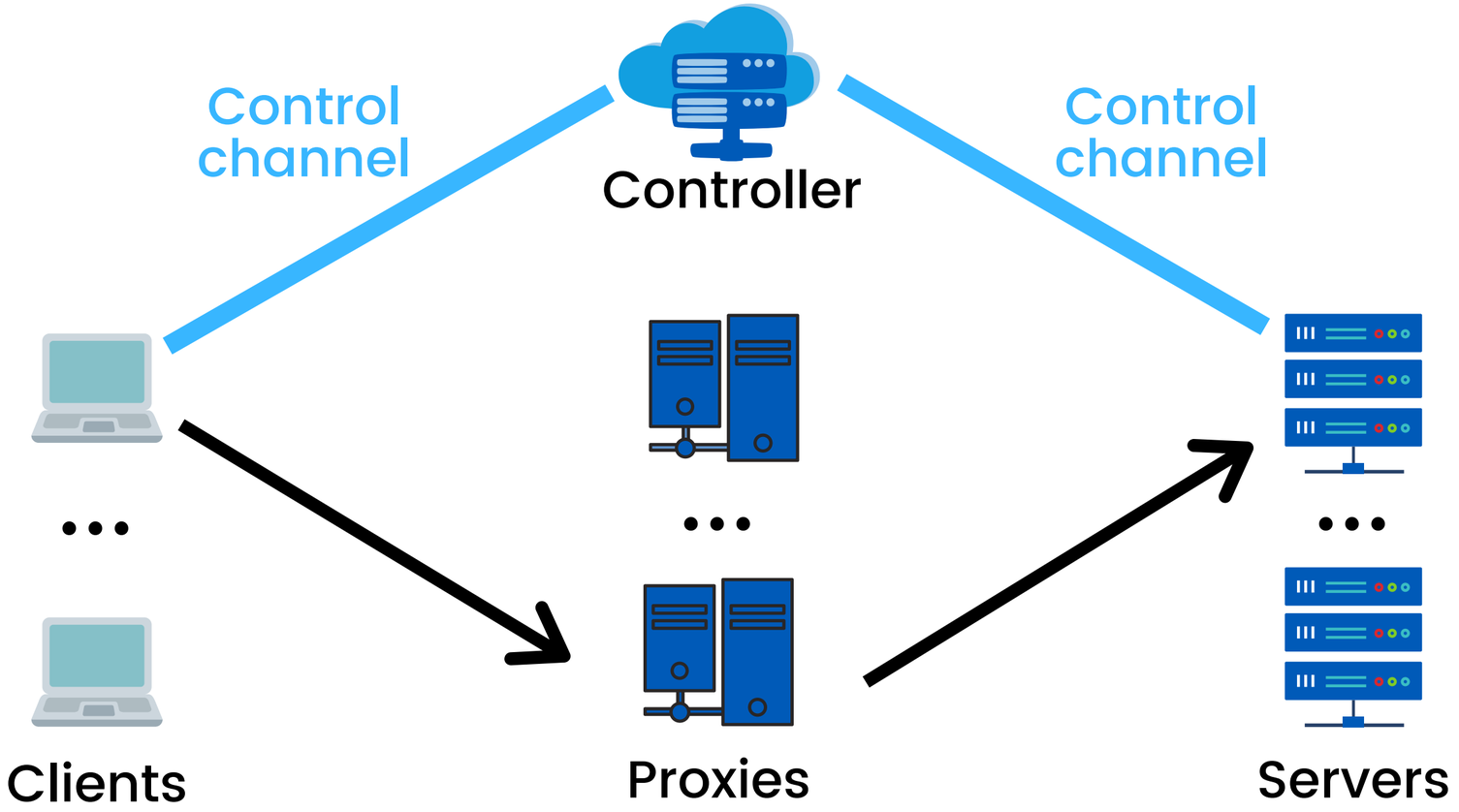
BAD PASS: Bots taking ADvantage of Proxy AS a Services

- In-depth study of proxies ecosystems



BAD PASS: Bots taking ADvantage of Proxy AS a Services

- In-depth study of proxies ecosystems



BAD PASS setup

- 22 client and server machines around the world: India, Australia, Japan, Germany, Ireland, Canada, USA (East, Central, West), South Africa, UAE, South America

BAD PASS setup

- 22 client and server machines around the world: India, Australia, Japan, Germany, Ireland, Canada, USA (East, Central, West), South Africa, UAE, South America
- 3 (+1) proxy services

BAD PASS setup

- 22 client and server machines around the world: India, Australia, Japan, Germany, Ireland, Canada, USA (East, Central, West), South Africa, UAE, South America
- 3 (+1) proxy services
- Actual rate per proxy: 2.48 reqs/s

BAD PASS Goals

- Collection of proxy services exit points

BAD PASS Goals

- Collection of proxy services exit points
- Validation of IP blocking solution

BAD PASS Goals

- Collection of proxy services exit points
- Validation of IP blocking solution
- Study of proxy algorithm
  - Geo-localization
  - Availability of devices

BAD PASS Goals

- Collection of proxy services exit points
- Validation of IP blocking solution
- Study of proxy algorithm
  - Geo-localization
  - Availability of devices
- Check if pool sharing between different proxies

BAD PASS Goals

- Collection of proxy services exit points
- Validation of IP blocking solution
- Study of proxy algorithm
  - Geo-localization
  - Availability of devices
- Check if pool sharing between different proxies
- Check statistical models on proxies

BAD PASS Goals

- Collection of proxy services exit points
- Validation of IP blocking solution
- Study of proxy algorithm
  - Geo-localization
  - Availability of devices
- Check if pool sharing between different proxies
- Check statistical models on proxies
- Perform network measurement on proxies

Non responding server

- When the server is not reachable the proxy try to connect with 3 different proxy gateways before giving up

Non responding server

- When the server is not reachable the proxy try to connect with 3 different proxy gateways before giving up

- Side experiment collecting IPs in a non responding server

Non responding server

- When the server is not reachable the proxy try to connect with 3 different proxy gateways before giving up
- Side experiment collecting IPs in a non responding server
- In the same location, non responding server not targeted by the experiment to have a measure of the noise

Some early results (12/01-31/03)

Proxy	Repetitions	Unique IPs
Brightdata*	31%	1,546,886
Oxylabs	47%	4,341,891
ProxyRack	59%	2,623,941
Smartproxy	46%	4,429,305

Thank You