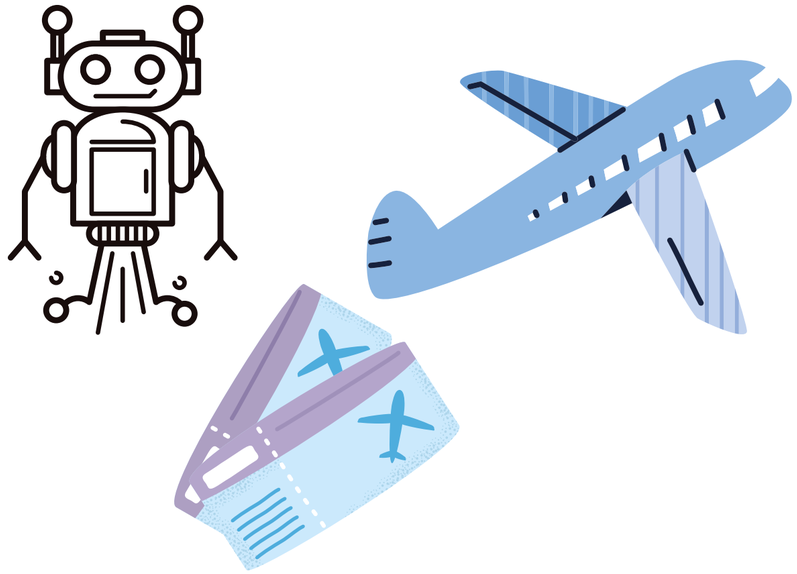
Using a honeypot to mitigate and study scrapers



Elisa Chiapponi
PhD student at EURECOM - GSO APP

EURECOM
Sophia Antipolis

aMaDEUS
Global Security Operations

Agenda

Agenda

1. Introduction
and motivations

Agenda

1. Introduction and motivations

2. Honeypot

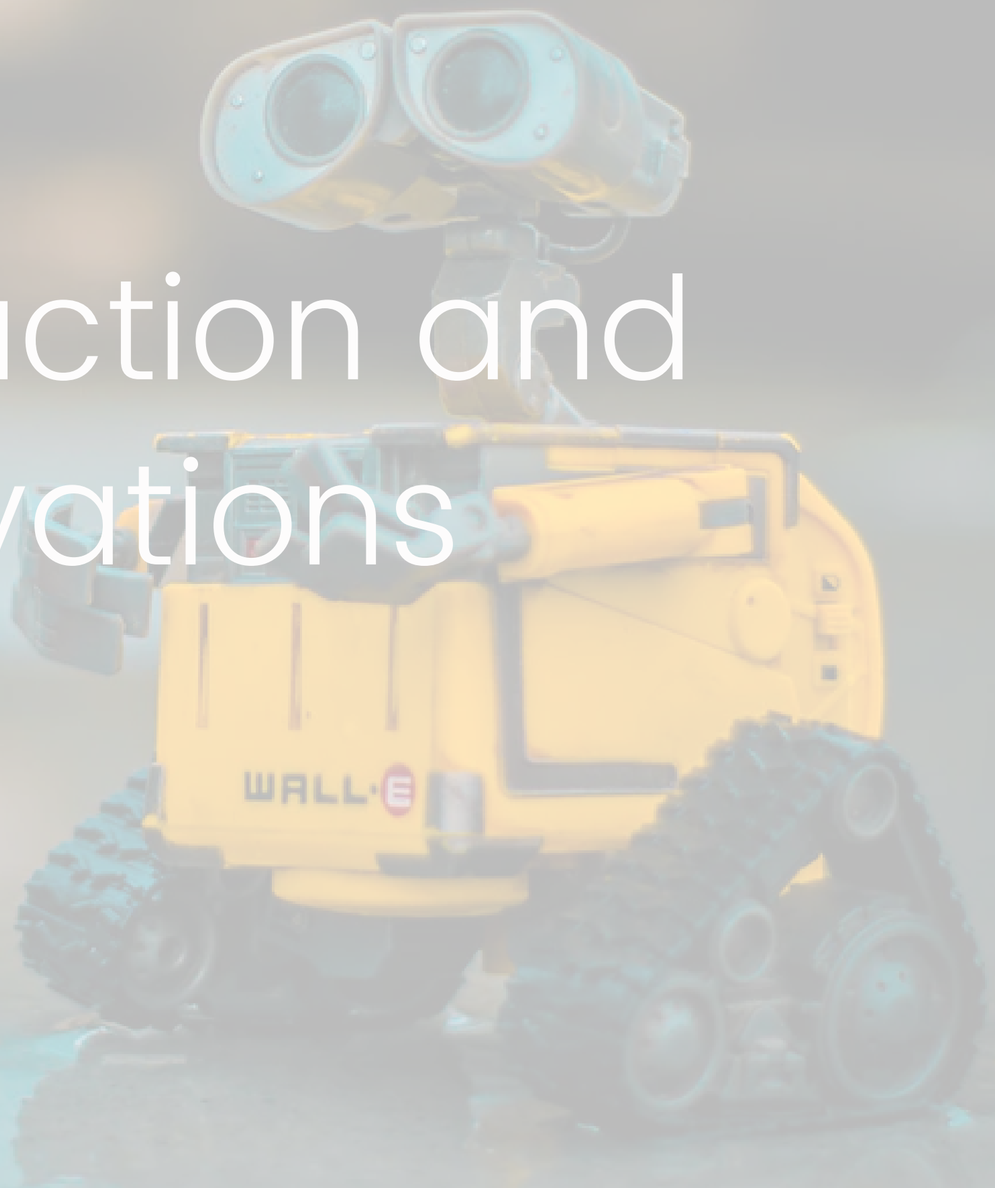
Agenda

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. Conclusions and future work

1. Introduction and motivations



Who am I?

Who am I?

Master thesis internship in Amadeus GSO -APP

Who am I?

Master thesis internship in Amadeus GSO -APP

Phd student Amadeus GSO-APP and EURECOM

Who am I?

Master thesis internship in Amadeus GSO -APP

Phd student Amadeus GSO-APP and EURECOM

Finding practical means to defeet scraping bots

Who am I?

Master thesis internship in Amadeus GSO -APP

Phd student Amadeus GSO-APP and EURECOM

Finding practical means to defeet 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

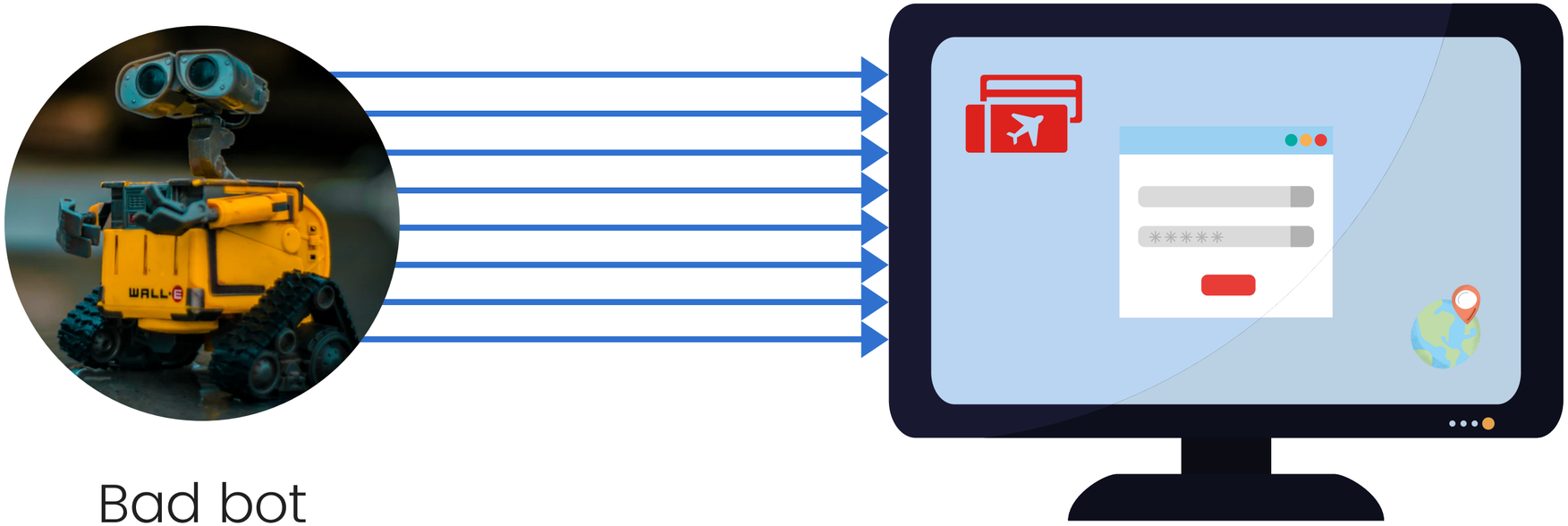
How bots affect airlines - Imperva 2019

Web scraping bad bots

Bad bot

Airline booking domain

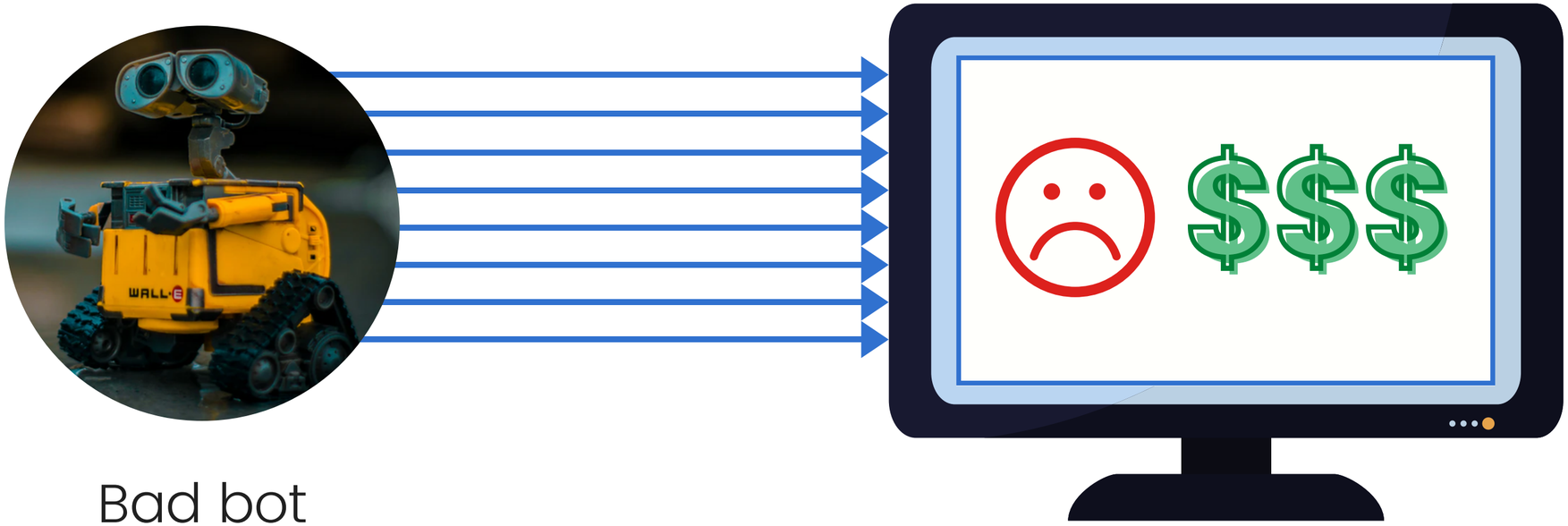
Web scraping bad bots



Bad bot

Airline booking domain

Web scraping bad bots



Bad bot

Airline booking domain

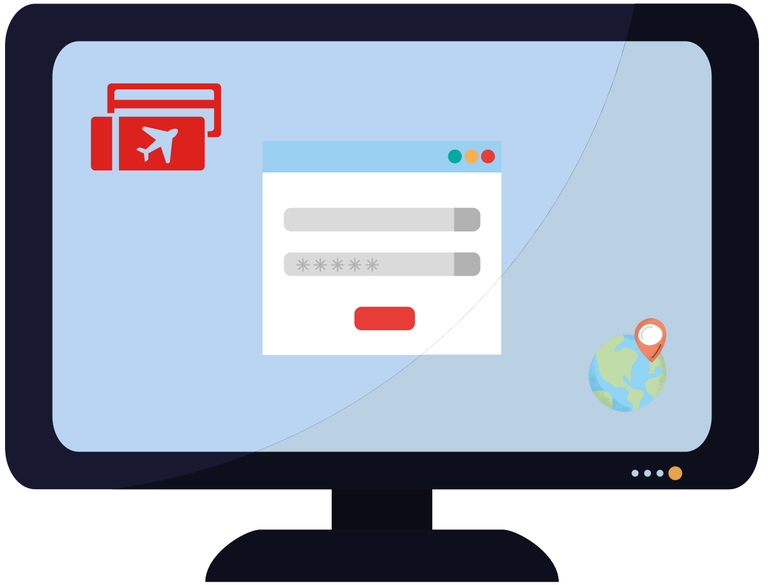
Anti-bot solutions

imperva

User

Bot

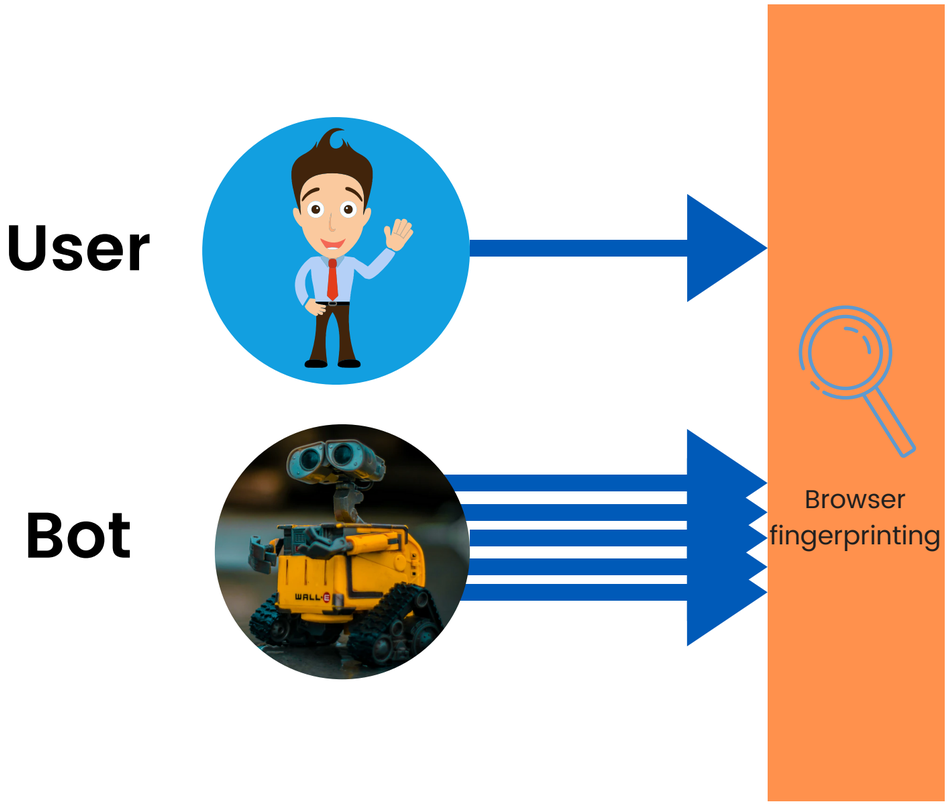
Browser fingerprinting

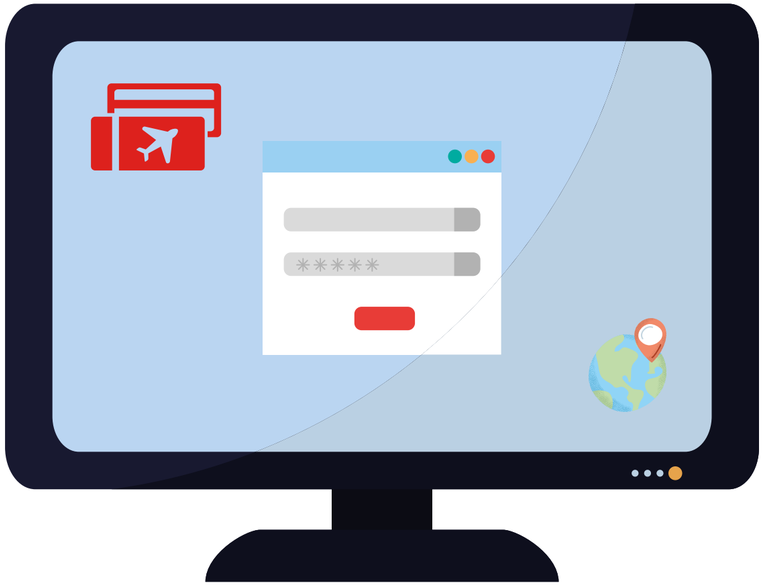


Booking domain of airline X

Anti-bot solutions

imperva





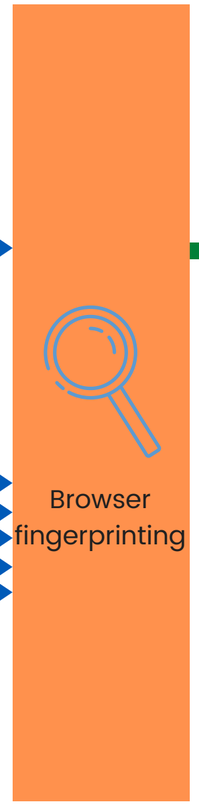
Booking domain
of airline X

Anti-bot solutions

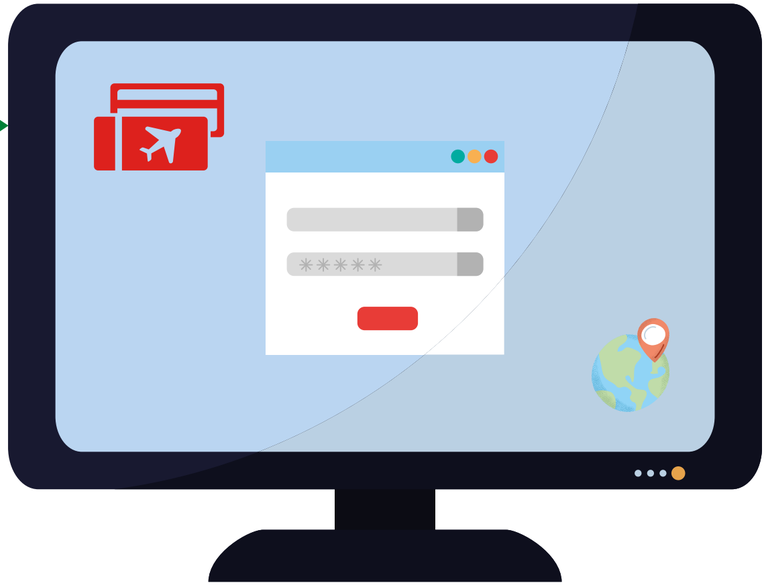
imperva

User

Bot



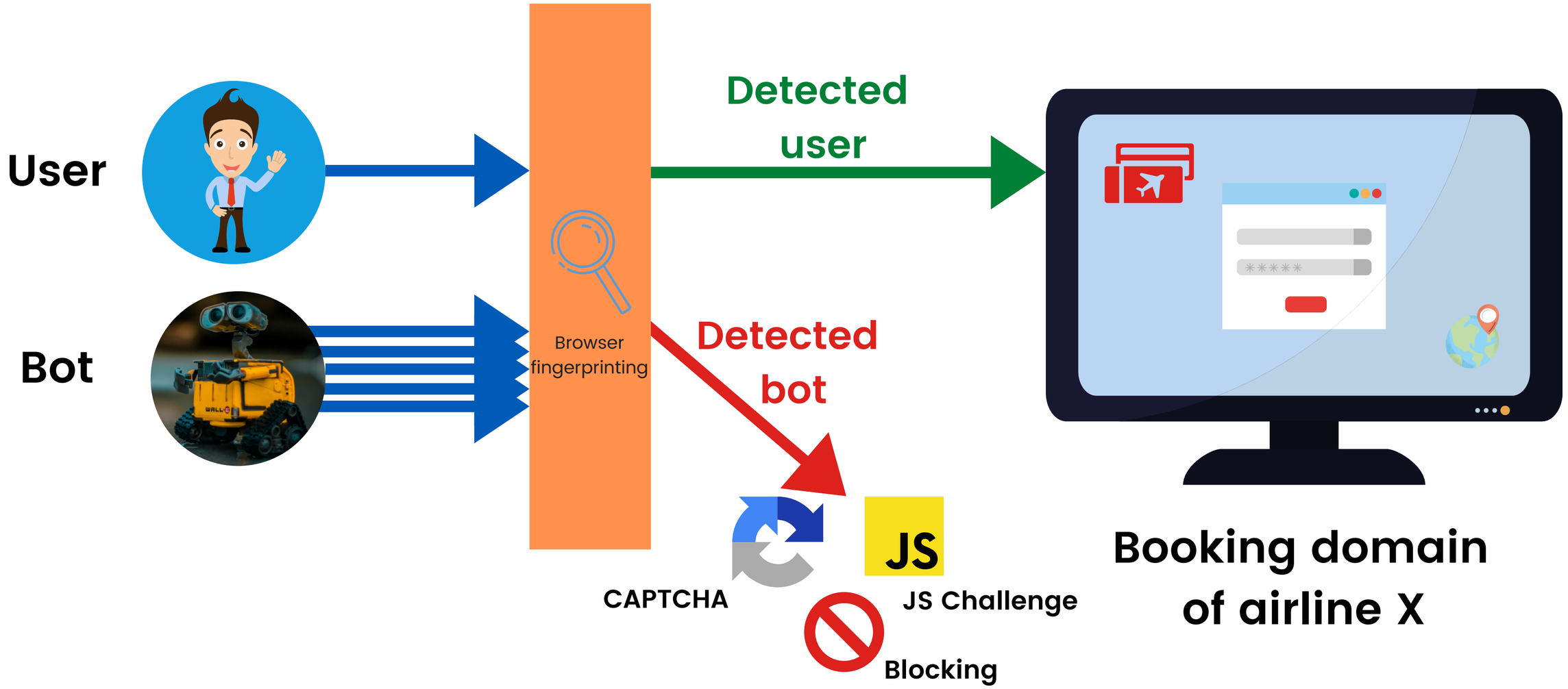
Detected user



Booking domain of airline X

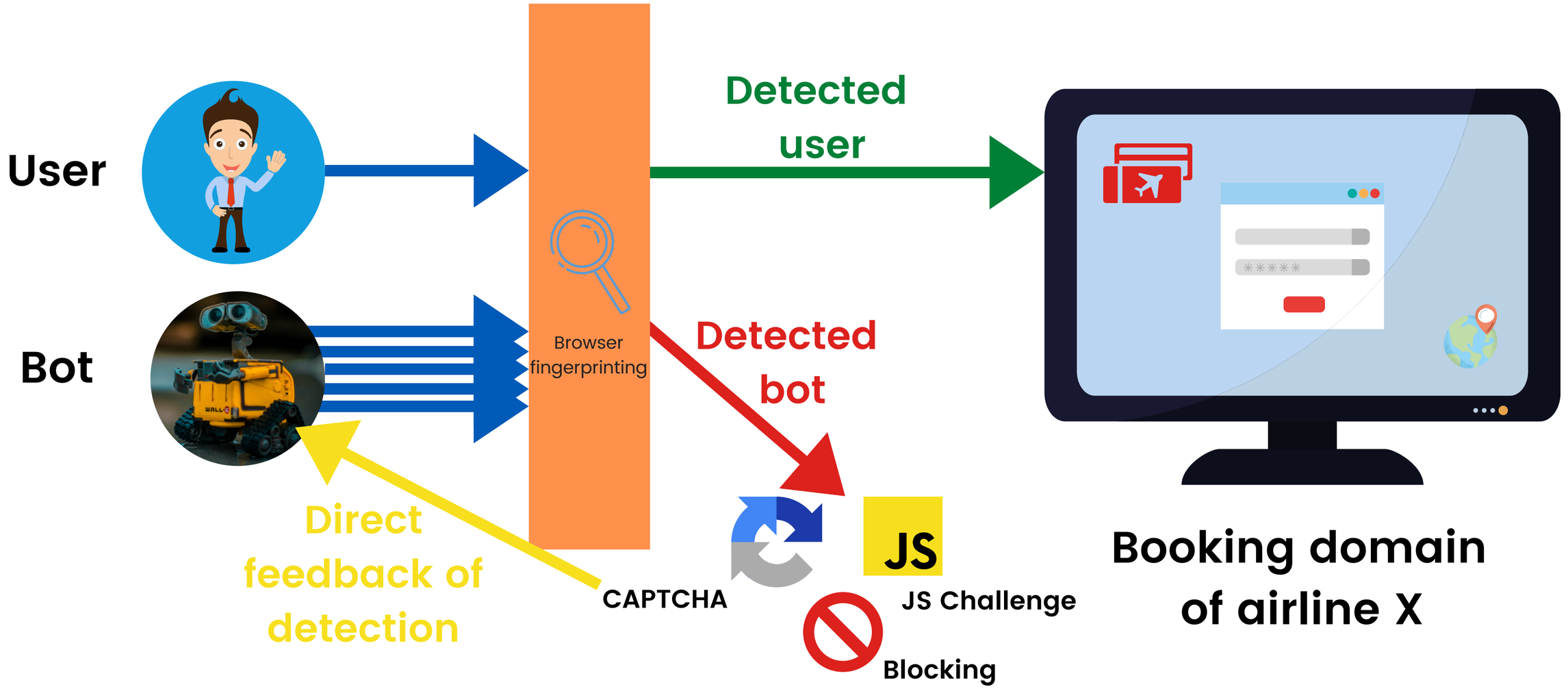
Anti-bot solutions

imperva



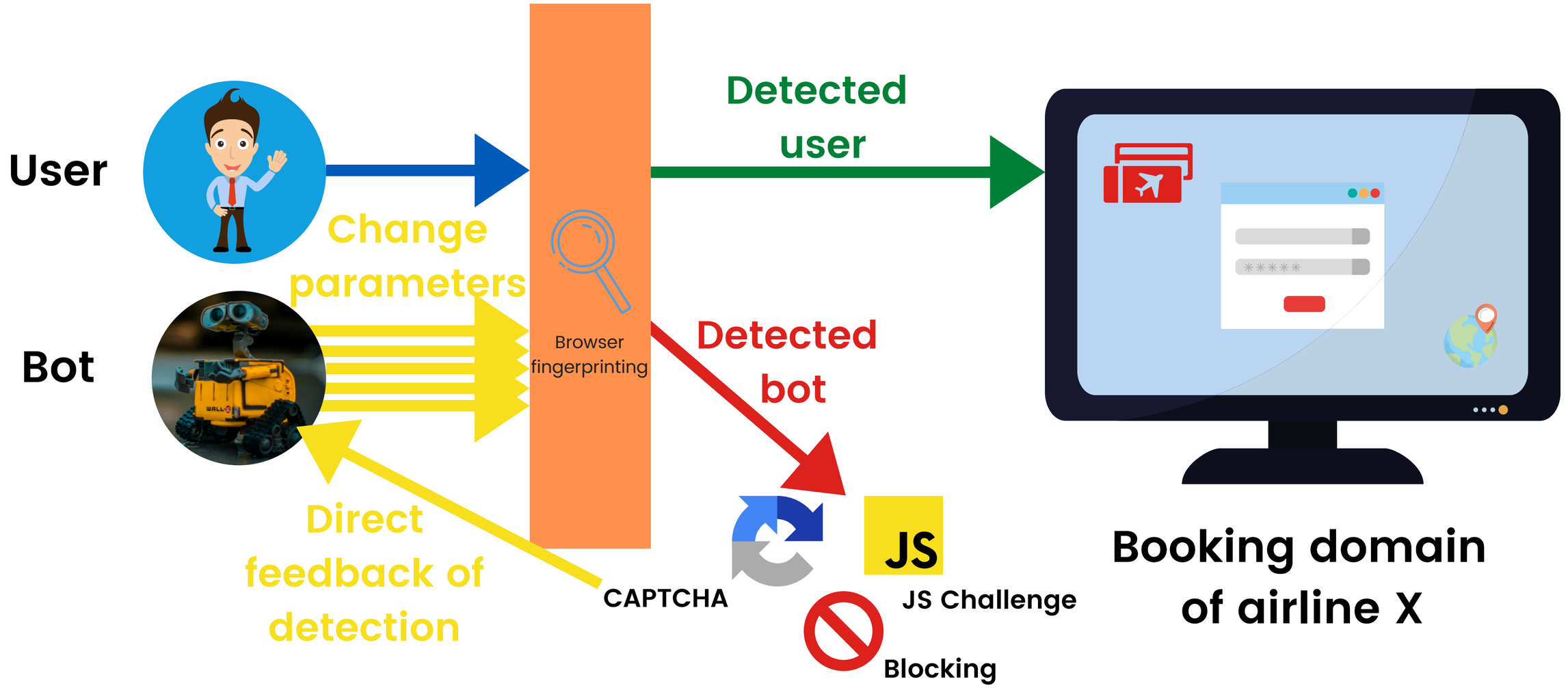
Anti-bot solutions

imperva



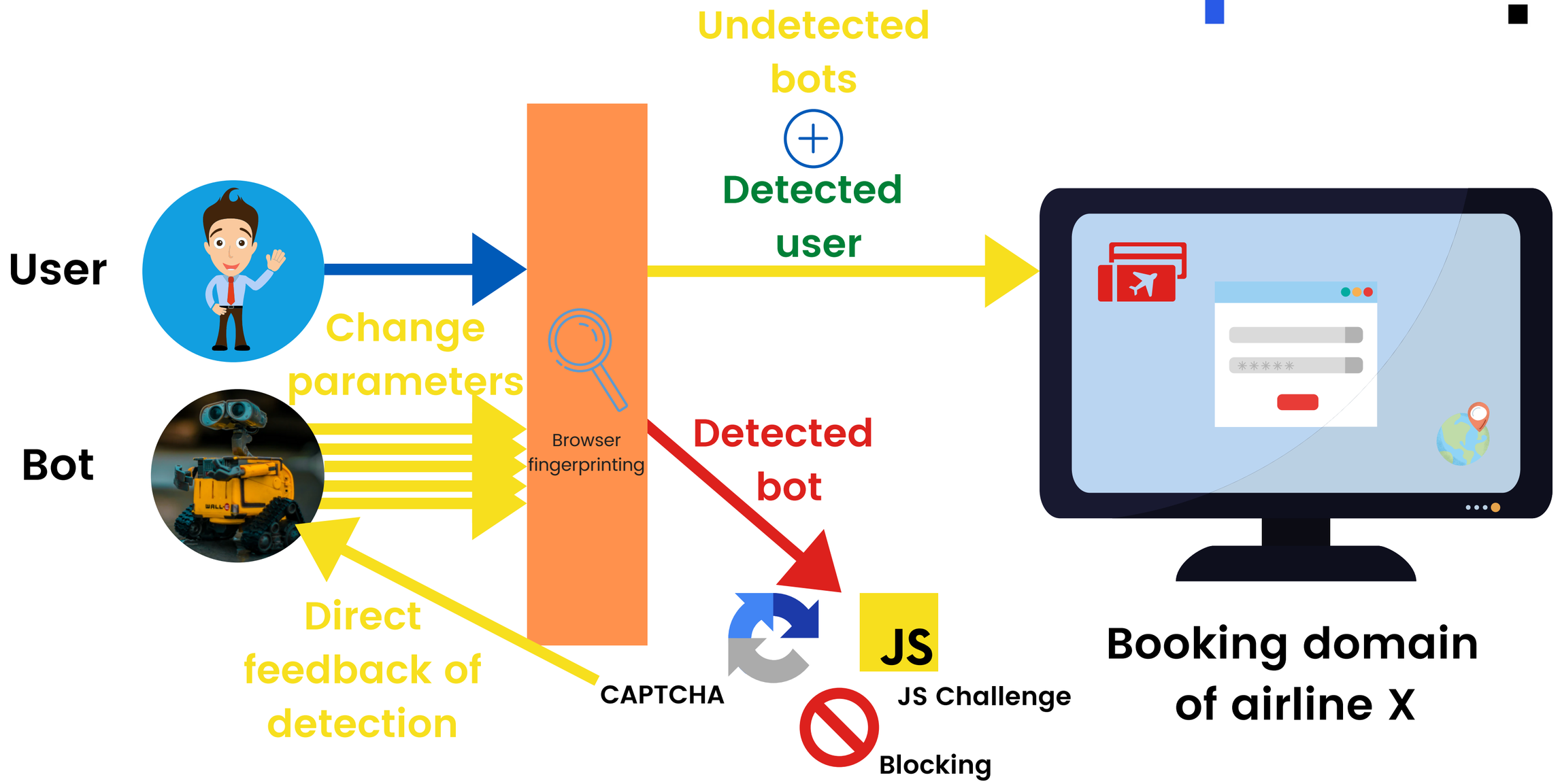
Anti-bot solutions

imperva



Anti-bot solutions

imperva



What 3rd parties works tell us

Blocked bots die

- Continuous verification of stealthiness and efficacy of the bots
- Rapid modification to avoid detection

What 3rd parties works tell us

Blocked bots die

- Continuous verification of stealthiness and efficacy of the bots
- Rapid modification to avoid detection

Harnessed information is verified

- Continuous verification of the correctness of the information
- Incorrect information is a direct feedback they have been detected

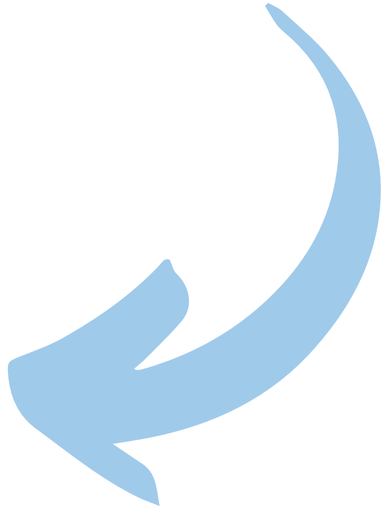
Initial idea

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

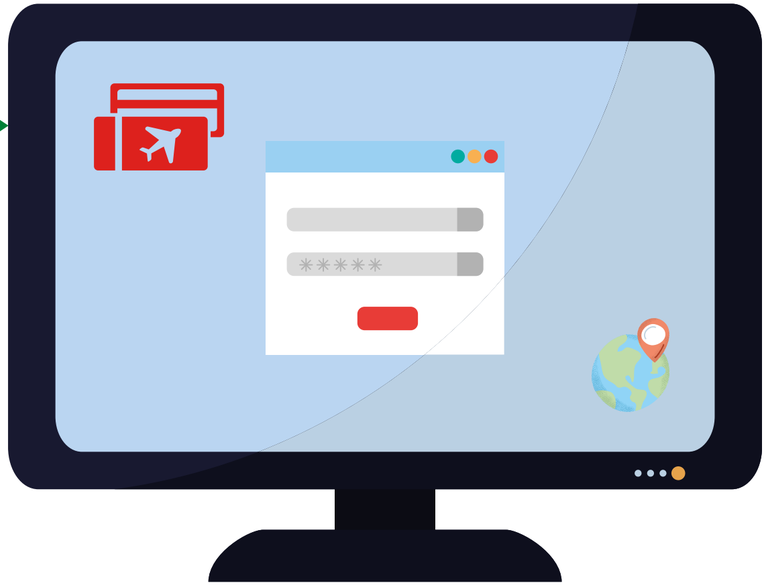
imperva

User

Bot

Browser fingerprinting

Detected user



Booking domain of airline X

The idea

imperva

User

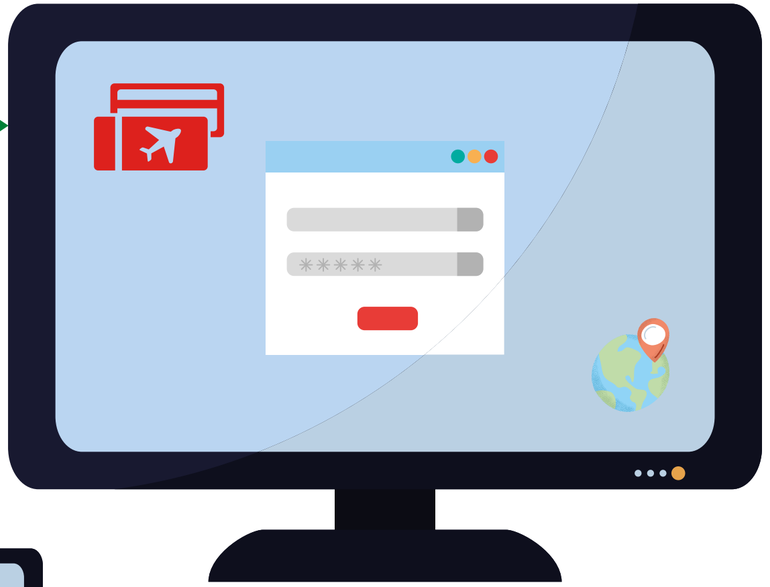
Bot

Browser fingerprinting

Detected user

Detected bot

Honeypot



Booking domain of airline X

The idea

- Prototype outside of Amadeus system

The idea

- Prototype outside of Amadeus system
- Check if the idea is valuable

The idea

- Prototype outside of Amadeus system
- Check if the idea is valuable
- Coded by us as replica of the real website

The idea

- Prototype outside of Amadeus system
- Check if the idea is valuable
- Coded by us as replica of the real website
- Modification of the fares

2. 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

Master thesis

- Bot signature active all day long but with differences among days

Master thesis

- Bot signature active all day long but with differences among days
- Partial redirection

Master thesis

- Bot signature active all day long but with differences among days
- Partial redirection
- Complete drop of the traffic after 58h

Master thesis

- Bot signature active all day long but with differences among days
- Partial redirection
- Complete drop of the traffic after 58h
- Post mortem analysis of the logs:

Master thesis

- Bot signature active all day long but with differences among days
- Partial redirection
- Complete drop of the traffic after 58h
- Post mortem analysis of the logs:
  - Bots started to create semantically incorrect queries after the redirection to the honeypot
    - Honeypot detection

Master thesis

- Bot signature active all day long but with differences among days
- Partial redirection
- Complete drop of the traffic after 58h
- Post mortem analysis of the logs:
  - Bots started to create semantically incorrect queries after the redirection to the honeypot
    - Honeypot detection
  - Behavioural patterns
    - Same actor behind

Second case-study

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

Second case-study

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

Second case-study

- 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

Second case-study

- 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

Time interval

- Time Interval=Departure date - Date of the request

Time interval

- Time Interval=Departure date - Date of the request
- Value between 0 and 14 days or 21, 30, 45, 60, 90, 120 days

Time interval

- Time Interval=Departure date - Date of the request
- Value between 0 and 14 days or 21, 30, 45, 60, 90, 120 days
- Only 0.2% of the requests exhibit different values but out of the 40 minutes daily time window

Time interval

- Time Interval=Departure date - Date of the request
- Value between 0 and 14 days or 21, 30, 45, 60, 90, 120 days
- Only 0.2% of the requests exhibit different values but out of the 40 minutes daily time window
- Segment=combination of one way/return flight, departure and arrival location

Time interval

- Time Interval=Departure date - Date of the request
- Value between 0 and 14 days or 21, 30, 45, 60, 90, 120 days
- Only 0.2% of the requests exhibit different values but out of the 40 minutes daily time window
- Segment=combination of one way/return flight, departure and arrival location
- Homogeneous distribution among different segments and request dates

Tuples statistics

- 4-tuples made of:
  - Departure airport
  - Arrival airport
  - Time interval
  - Type of flight (one way/return)

Tuples statistics

- 4-tuples made of:
  - Departure airport
  - Arrival airport
  - Time interval
  - Type of flight (one way/return)
- 982 distinct tuples vs 410 average daily requests

Tuples statistics

- Each tuple is asked on average 23.41 times

Tuples statistics

- Each tuple is asked on average 23.41 times
- Average number of days a tuple was requested: 22.85 days

Tuples statistics

- Each tuple is asked on average 23.41 times
- Average number of days a tuple was requested: 22.85 days
- Generally tuples are asked once a day at most

Tuples statistics

- Each tuple is asked on average 23.41 times
- Average number of days a tuple was requested: 22.85 days
- Generally tuples are asked once a day at most
- 20% of tuples are asked at least more than once a day
  - All requests done in little span of time
  - Maximum difference: 5 minutes and 30 seconds

Tuples statistics

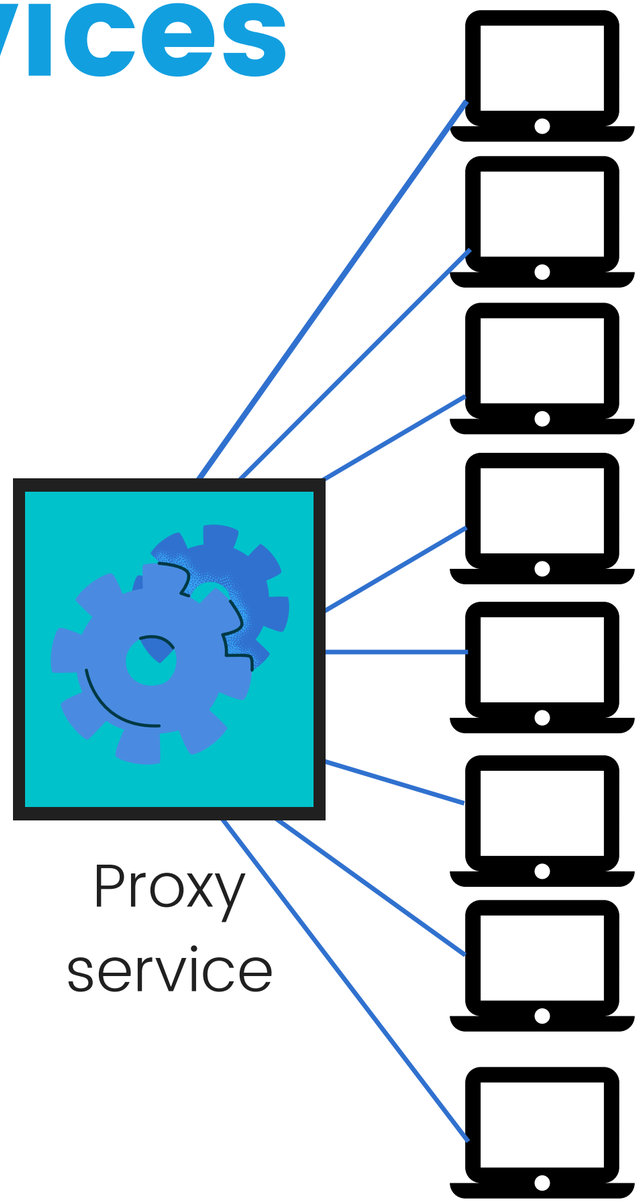
- Each tuple is asked on average 23.41 times
- Average number of days a tuple was requested: 22.85 days
- Generally tuples are asked once a day at most
- 20% of tuples are asked at least more than once a day
  - All requests done in little span of time
  - Maximum difference: 5 minutes and 30 seconds

Check for consistency, but not sophisticated enough

3. Proxy services and IP addresses

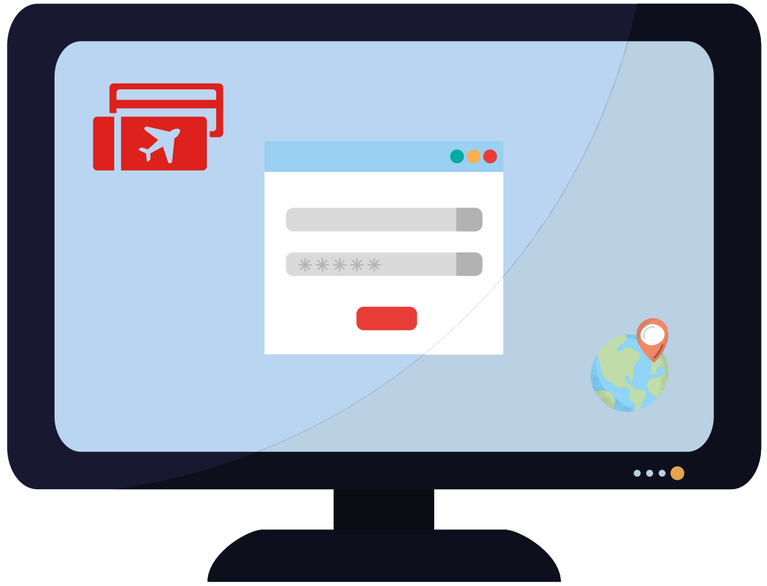
Proxy services

Bad bot



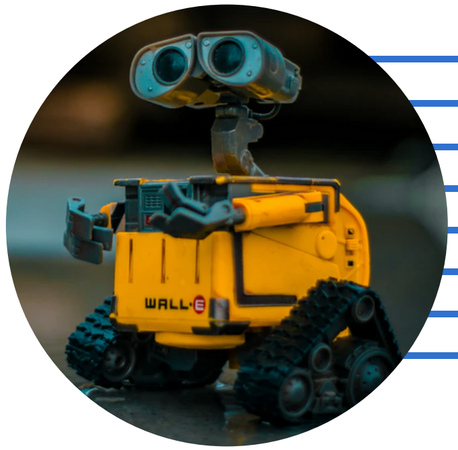
Proxy service

Proxy gateways



Airline booking domain

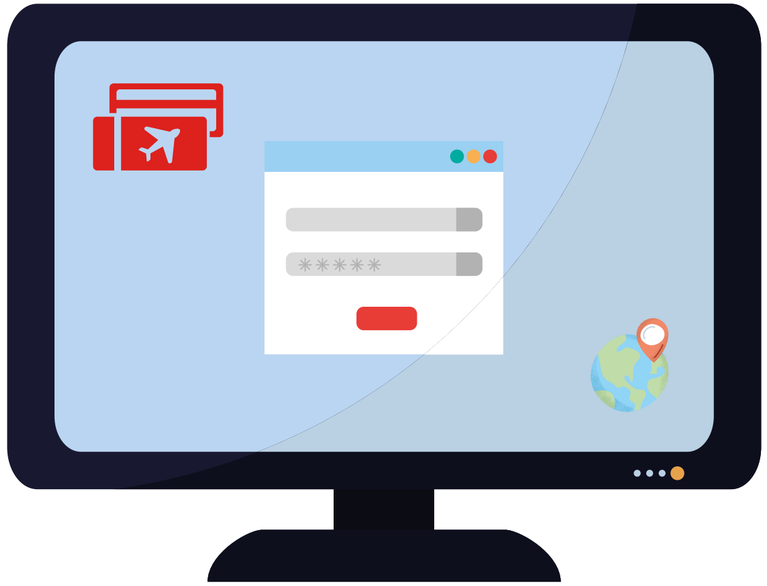
Proxy services



Bad bot

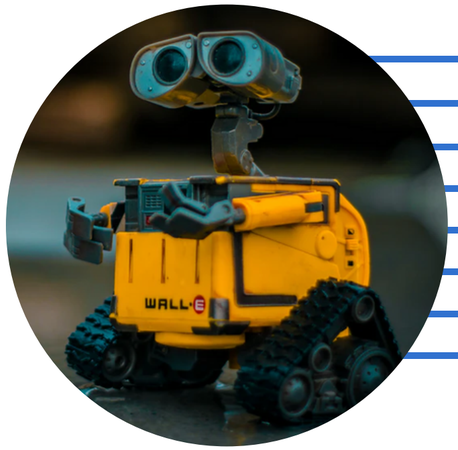
Proxy service

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

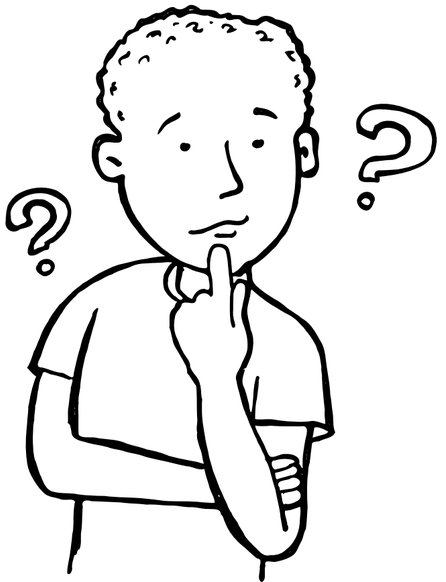
Impossible to link bot campaign to organization

No need for private distributed infrastructure

Impractical blocking IPs strategy

Question

Are our IP addresses coming from proxy services?



IPs analysis in other airlines

Analyses of the traffic of other 17 airlines

- Only 5 bookings during the running time of the case-study
- Dates different from the ones in which the IP was seen in the honeypot
- Requests not associated with the bot signature

IPs analysis in other airlines

Analyses of the traffic of other 17 airlines

- Only 5 bookings during the running time of the case-study
- Dates different from the ones in which the IP was seen in the honeypot
- Requests not associated with the bot signature

1. Some IPs are used by legit users
2. Risk of blocking legit users remains low

IPs reputation

IPQualityScore analysis

- 72% showed suspicious behavior
- 28% classified as high risk

IPs reputation

- IPQualityScore analysis
  - 72% showed suspicious behavior
  - 28% classified as high risk
- DNS blocklists
  - 76% blocked in at least one blocklist

IPs reputation

- IPQualityScore analysis
  - 72% showed suspicious behavior
  - 28% classified as high risk
- DNS blocklists
  - 76% blocked in at least one blocklist
- Tor network
  - 72 IPs announced in 5 days
  - Days different from the honeypot requests

IPs reputation

- IPQualityScore analysis
  - 72% showed suspicious behavior
  - 28% classified as high risk
- DNS blocklists
  - 76% blocked in at least one blacklist
- Tor network
  - 72 IPs announced in 5 days
  - Days different from the honeypot requests

These IPs were doing malicious activities also outside our scope

IPs reputation

- IPQualityScore analysis
  - 72% showed suspicious behavior
  - 28% classified as high risk
- DNS blocklists
  - 76% blocked in at least one blacklist
- Tor network
  - 72 IPs announc
  - Days different
    honeypot rec

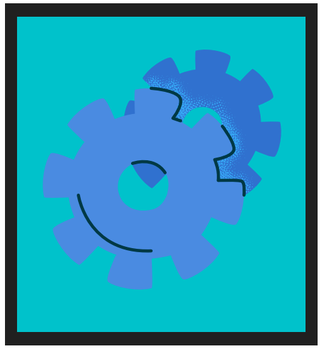
They were not allocated for the botnet only

These IPs were doing malicious activities also outside our scope

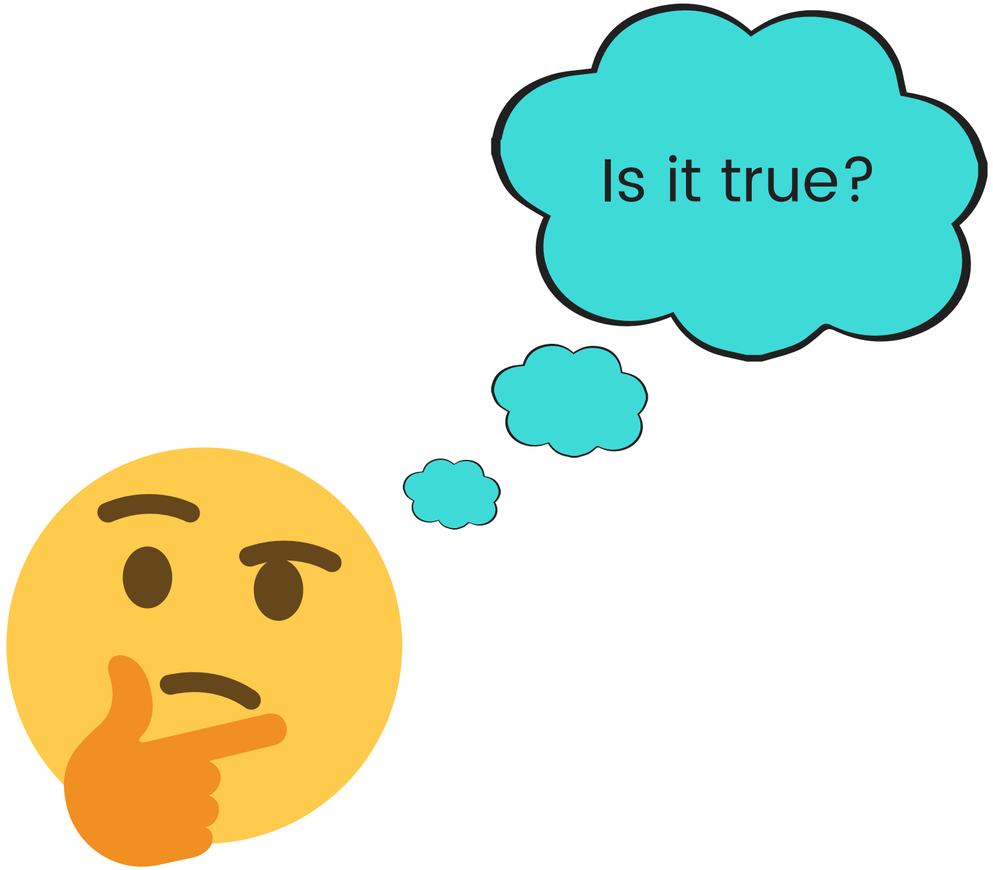
Proxy services claims

Proxy services claims

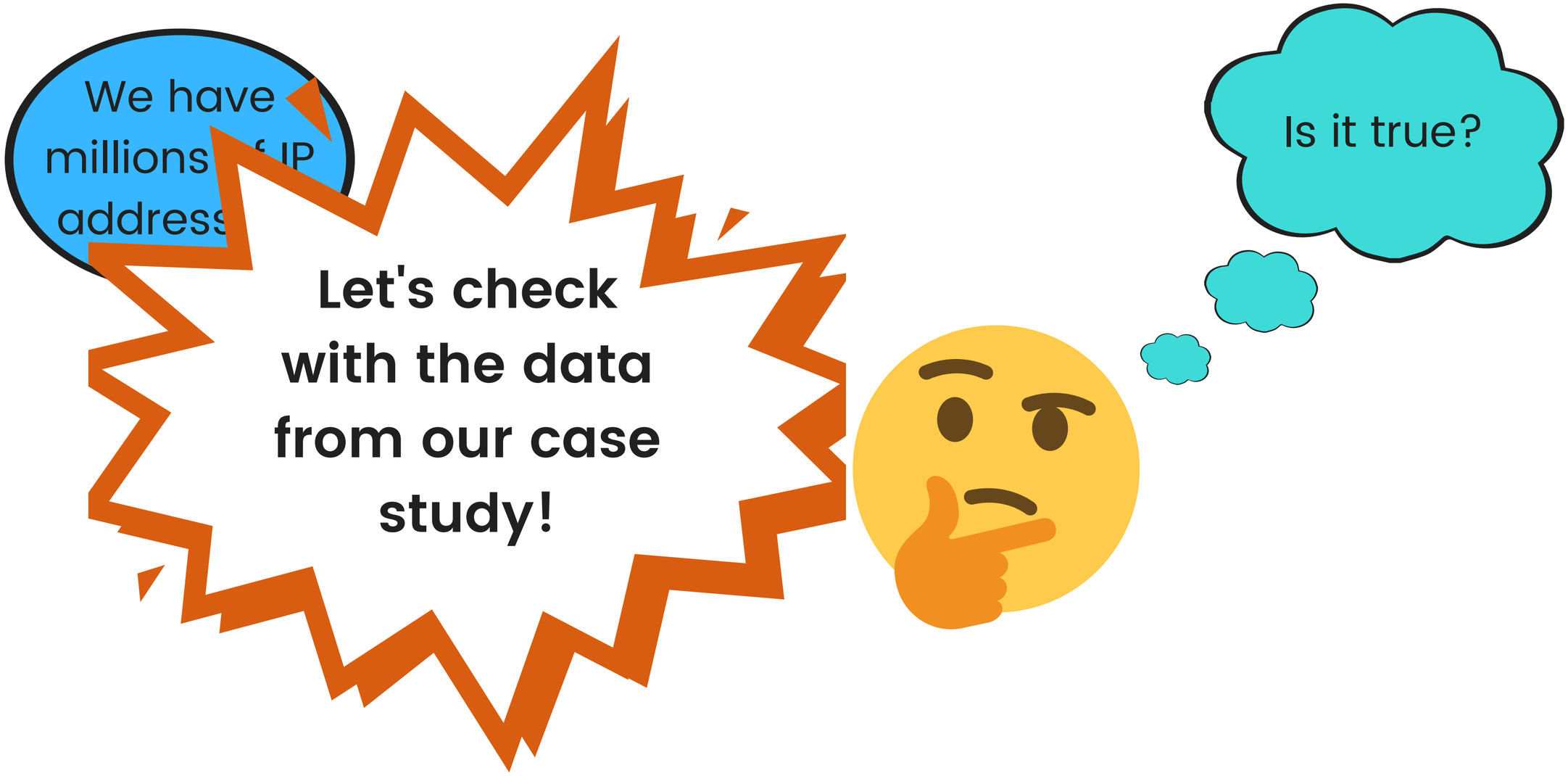
We have millions of IP addresses!



Proxy service



Proxy services claims



IP addresses study

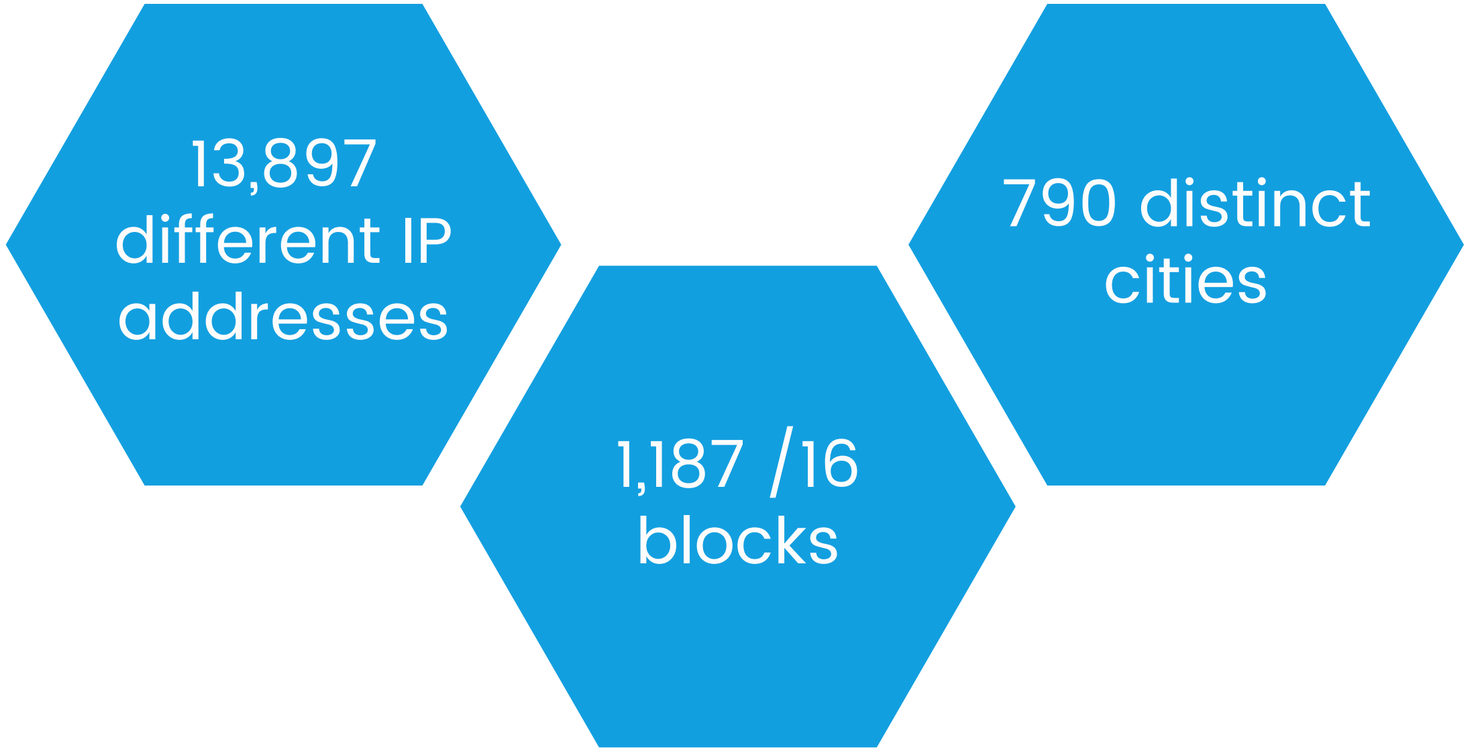
13,897
different IP
addresses

IP addresses study

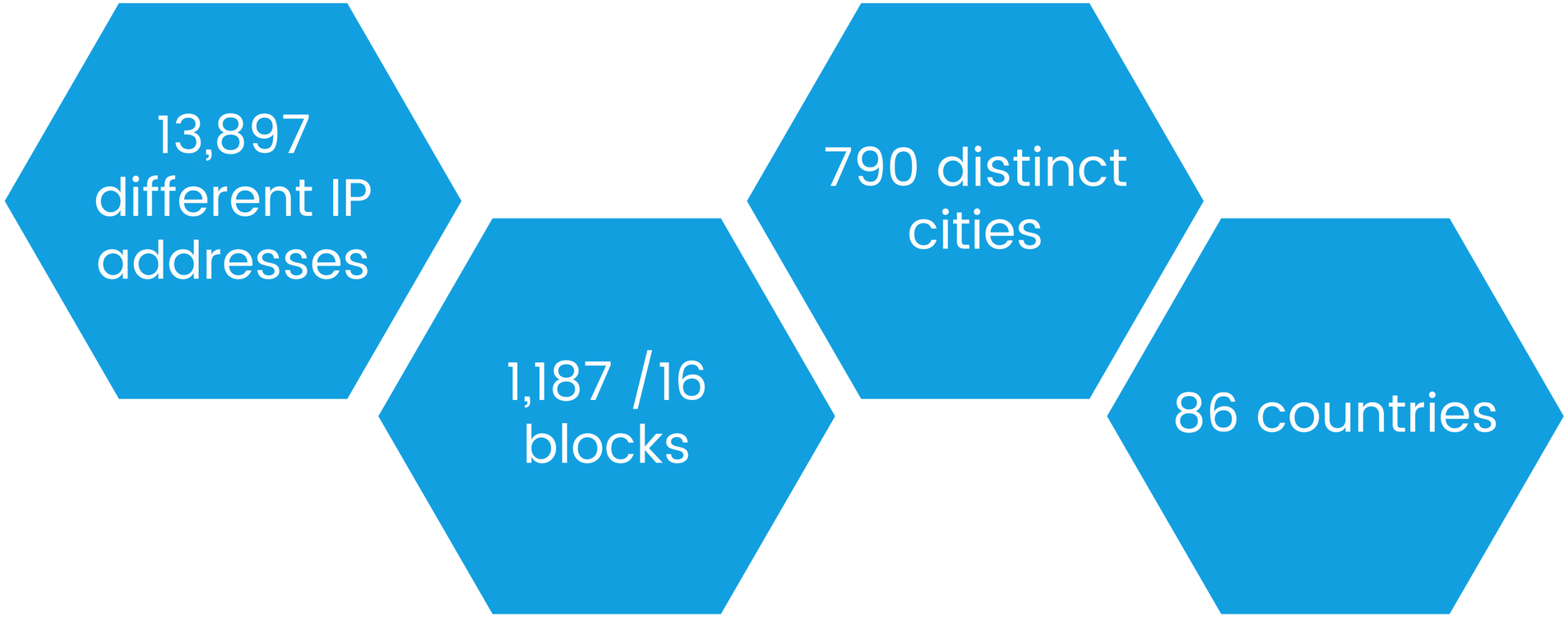
13,897 different IP addresses

1,187 /16 blocks

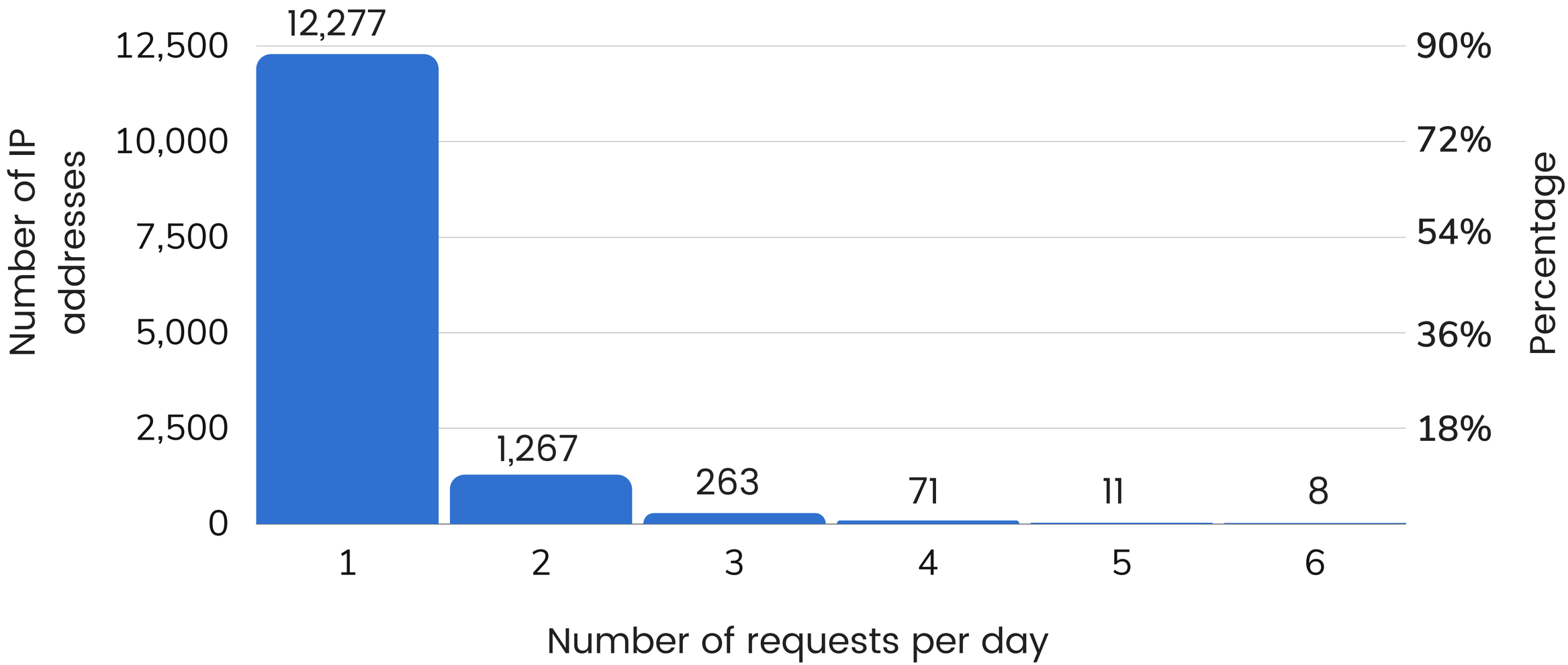
IP addresses study



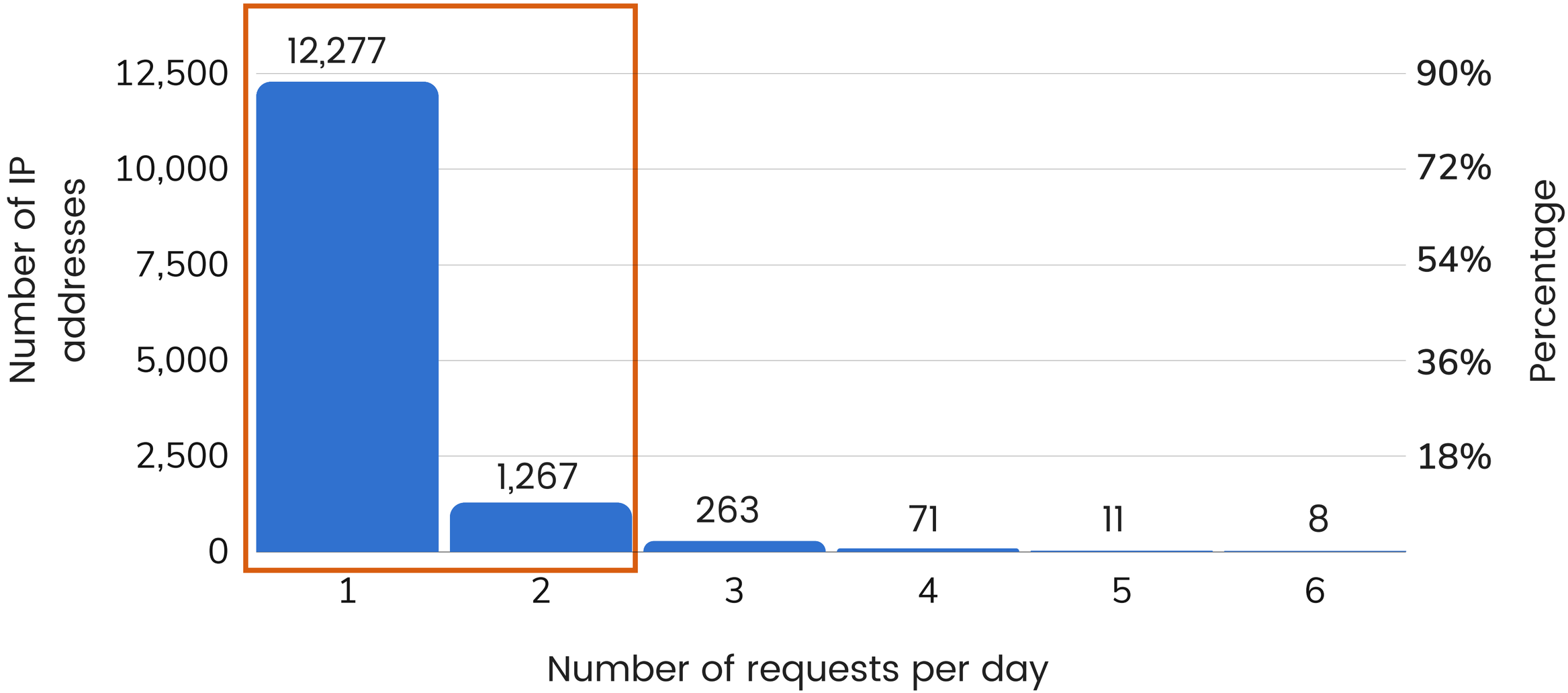
IP addresses study



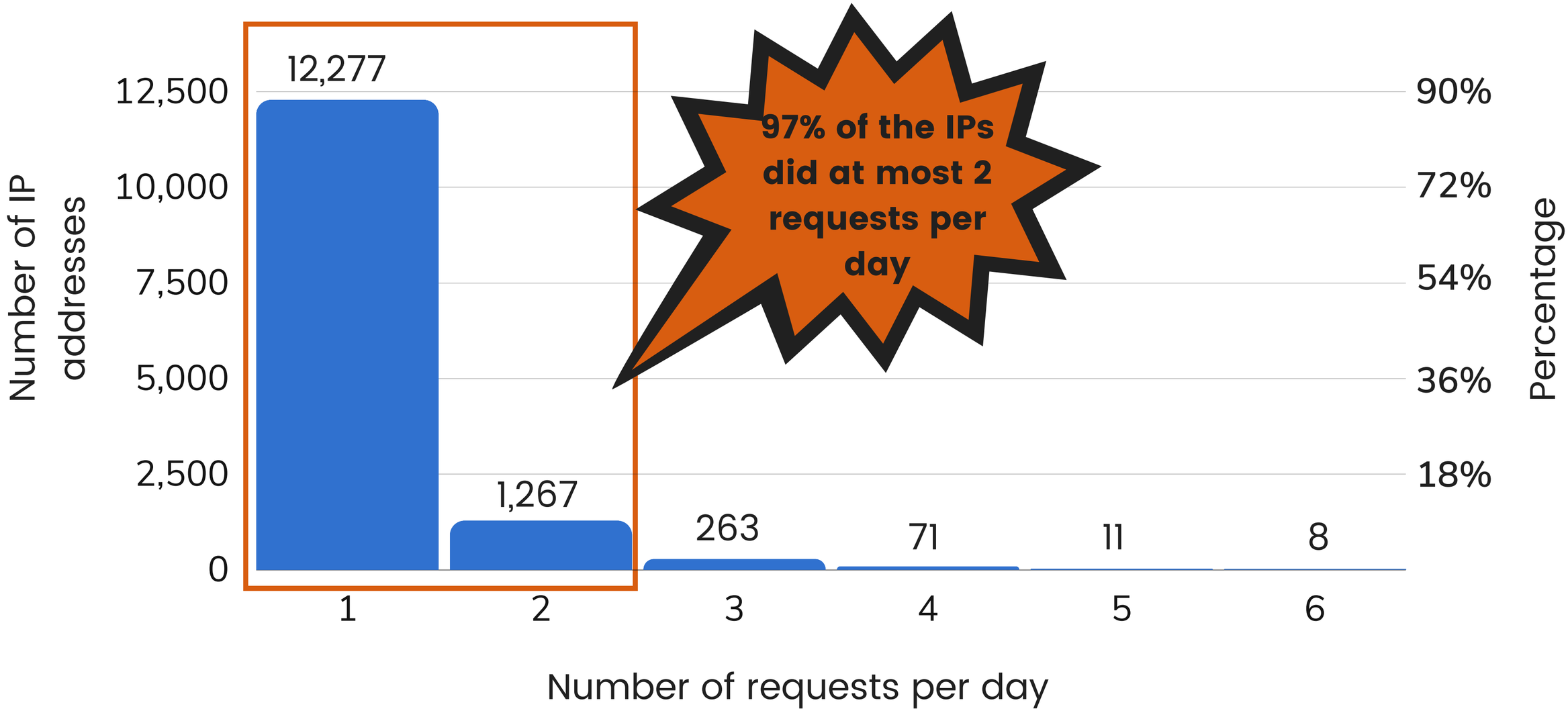
Daily number of requests per IP



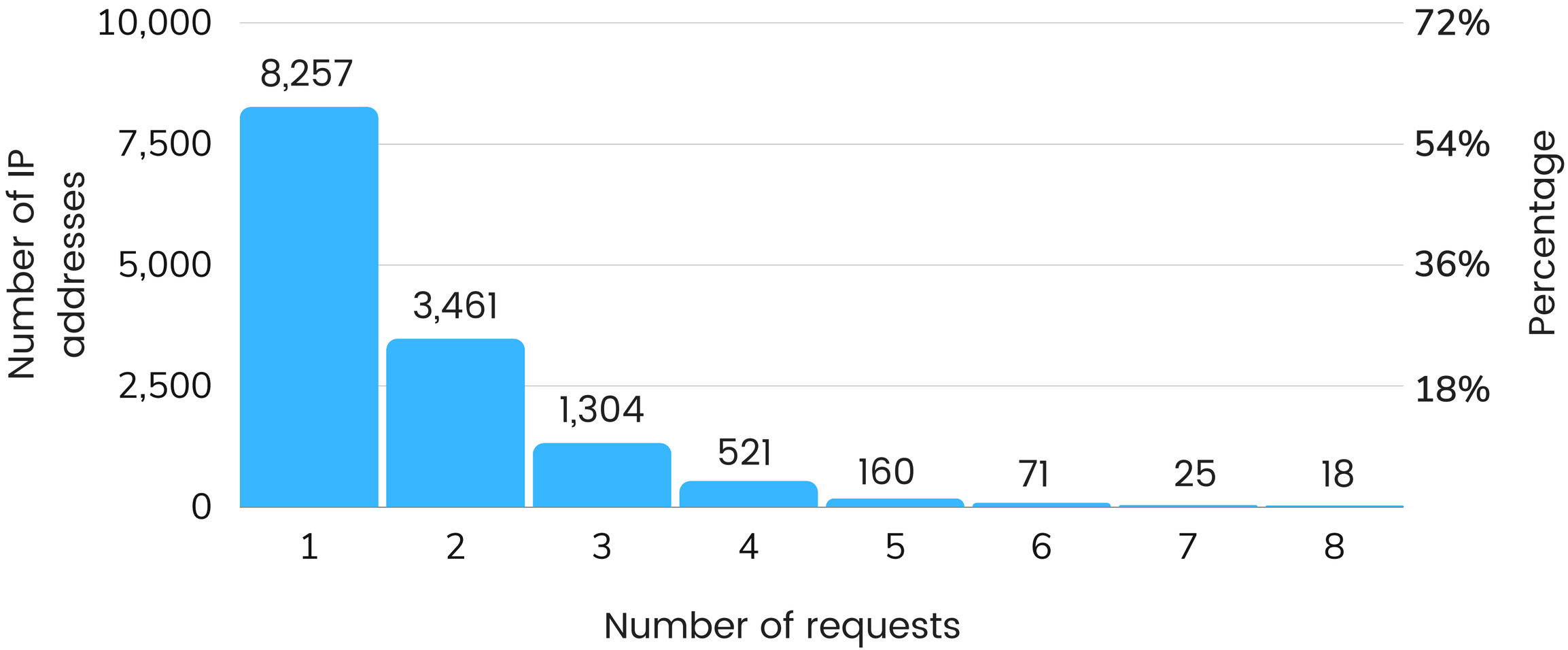
Daily number of requests per IP



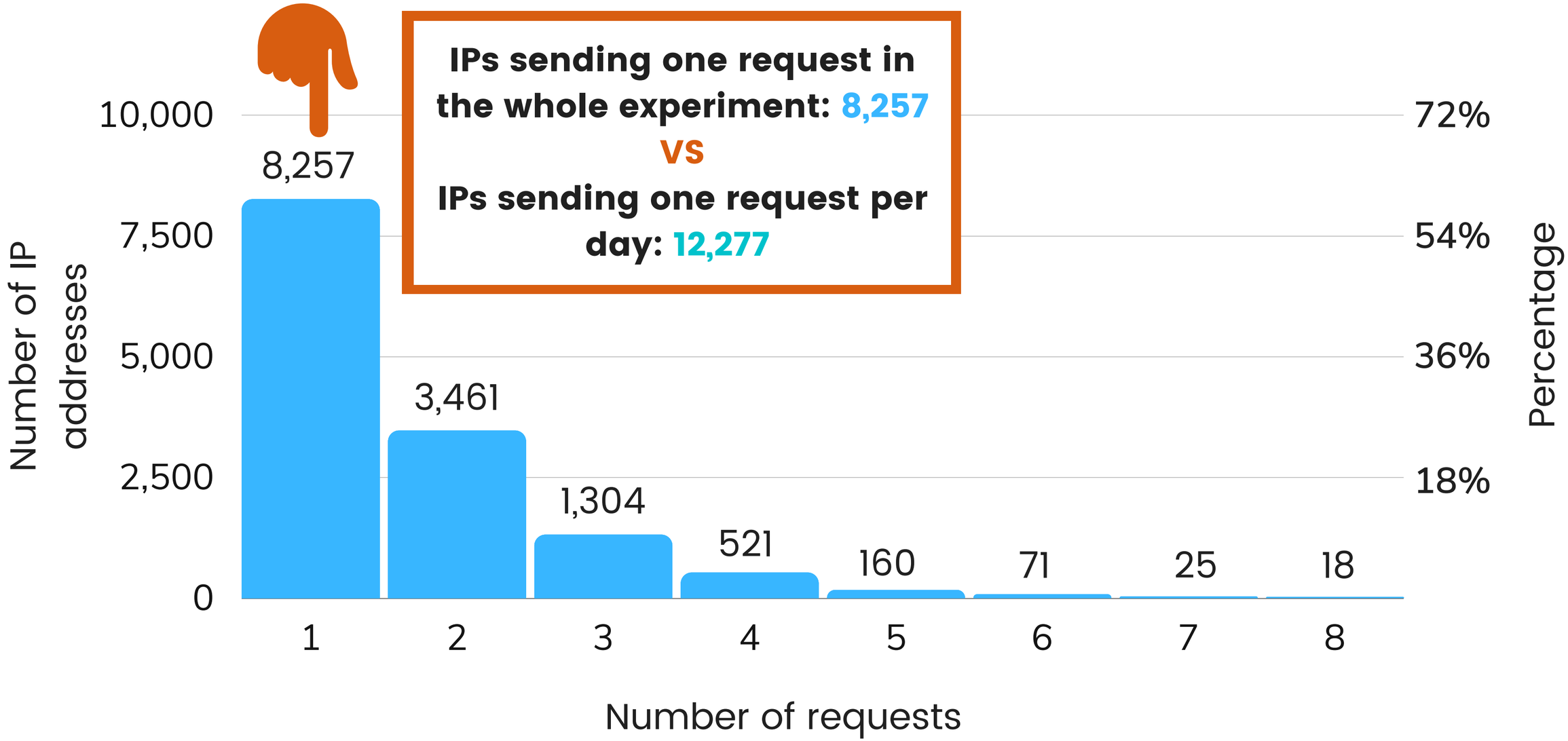
Daily number of requests per IP



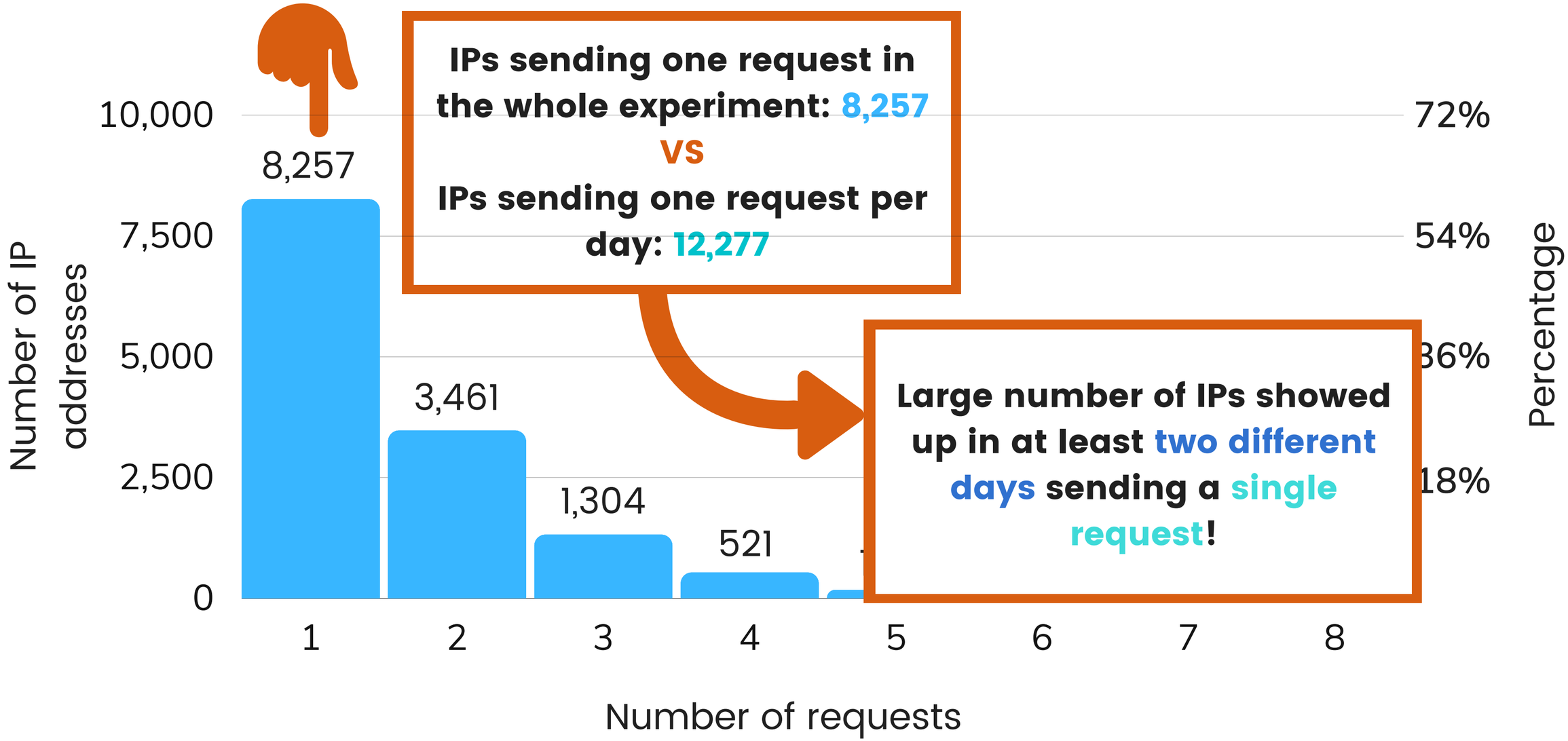
Maximum number of requests per IP



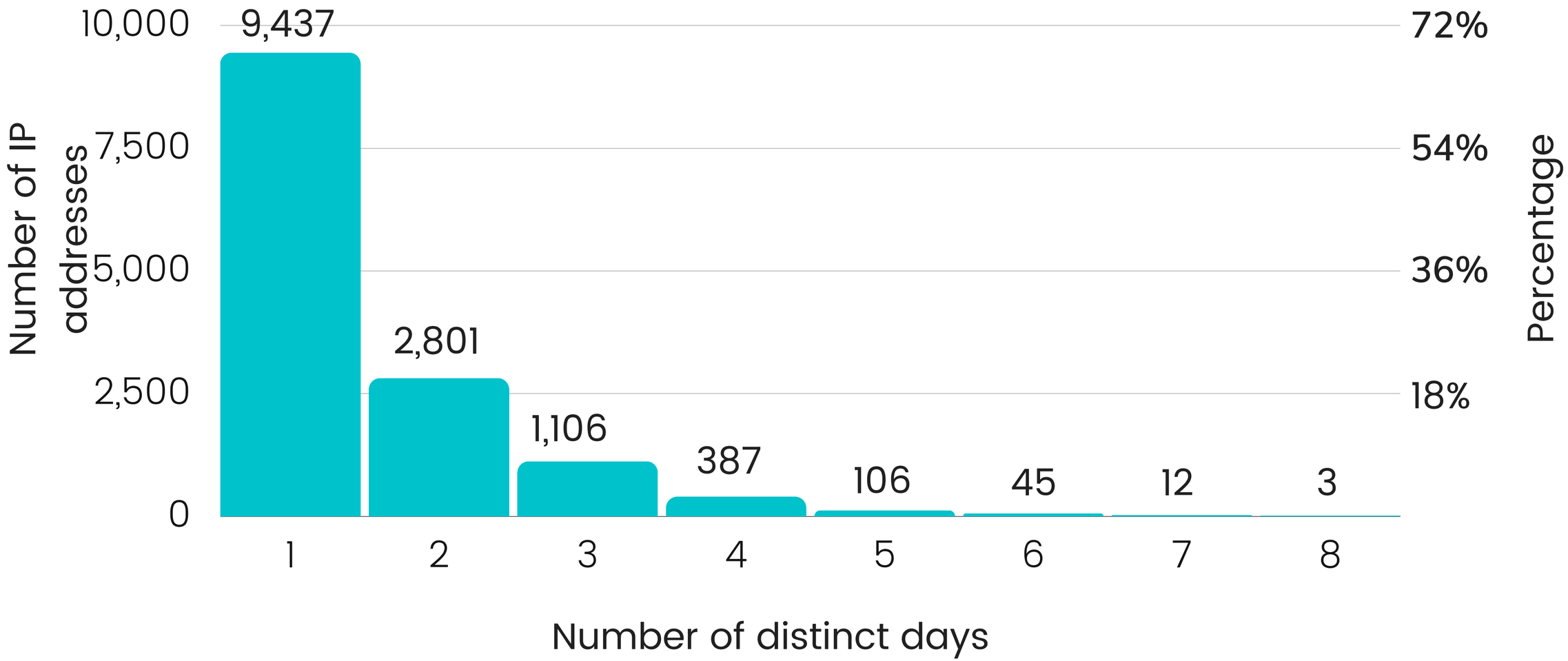
Maximum number of requests per IP



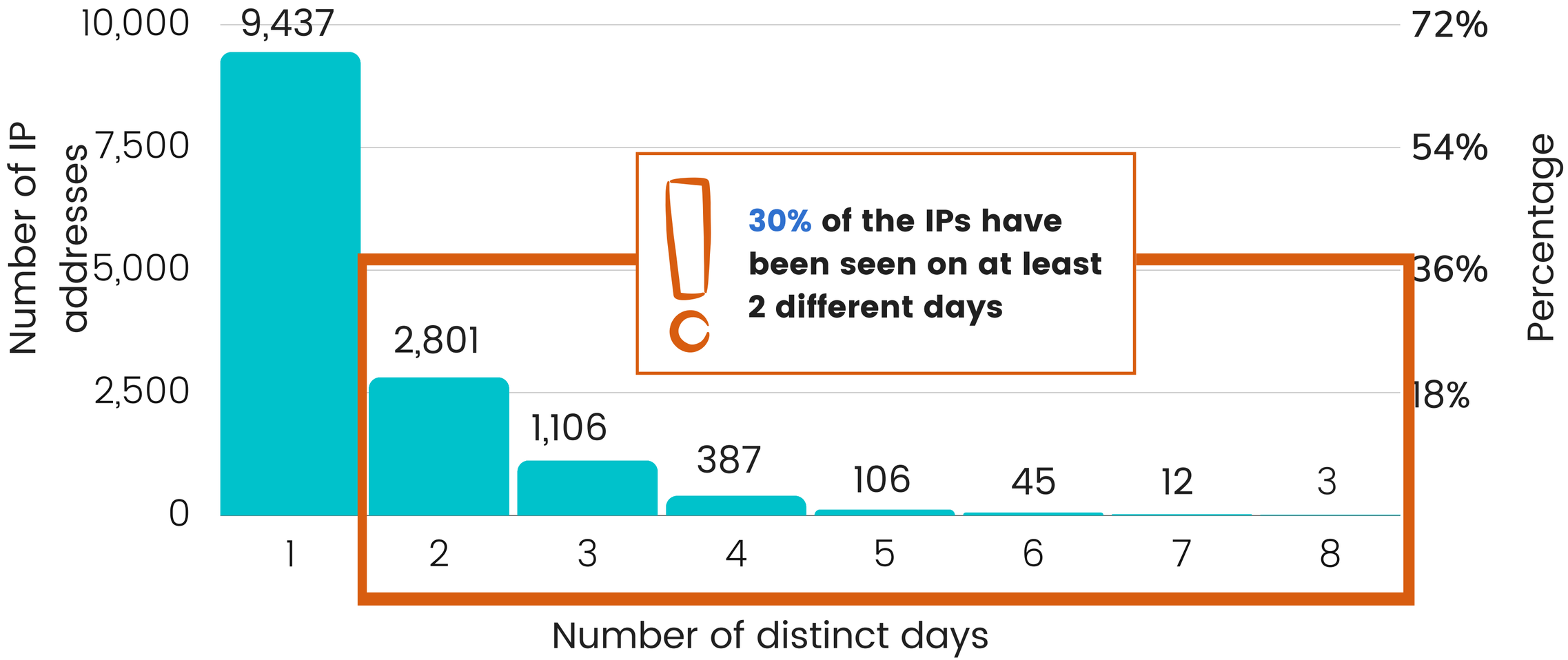
Maximum number of requests per IP



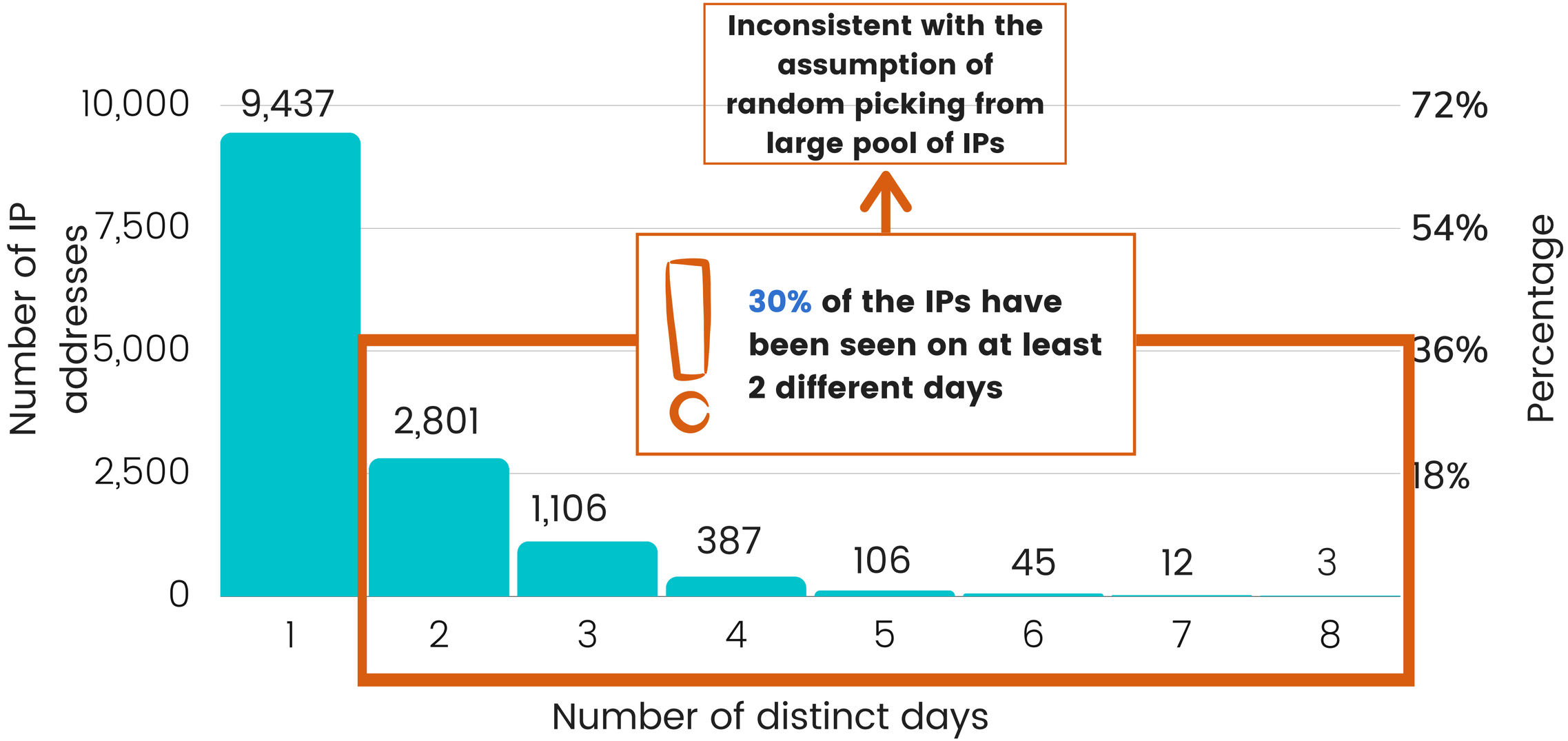
Number of IPs seen in distinct days



Number of IPs seen in distinct days



Number of IPs seen in distinct days



The Birthday Paradox

Given 56 random integers drawn from a discrete uniform distribution with range $[1,P]$, what is the probability $p(56; P)$ that at least two numbers are the same?

Is this likely to happen?

- Approximate result: $1-\left(\frac{P-1}{P}\right)^{\frac{56(55-1)}{2}}$

Is this likely to happen?

- Approximate result: $1-\left(\frac{P-1}{P}\right)^{\frac{56(55-1)}{2}}$
- $P=10{,}000{,}000 \longrightarrow p(56,10M)\approx 0.000154$

Is this likely to happen?

- Approximate result: $1-\left(\frac{P-1}{P}\right)^{\frac{56(55-1)}{2}}$
- $P=10{,}000{,}000 \rightarrow p(56,10M)\approx 0.000154$
- $P=1{,}000{,}000 \rightarrow p(56,1M)\approx 0.001538$

Is this likely to happen?

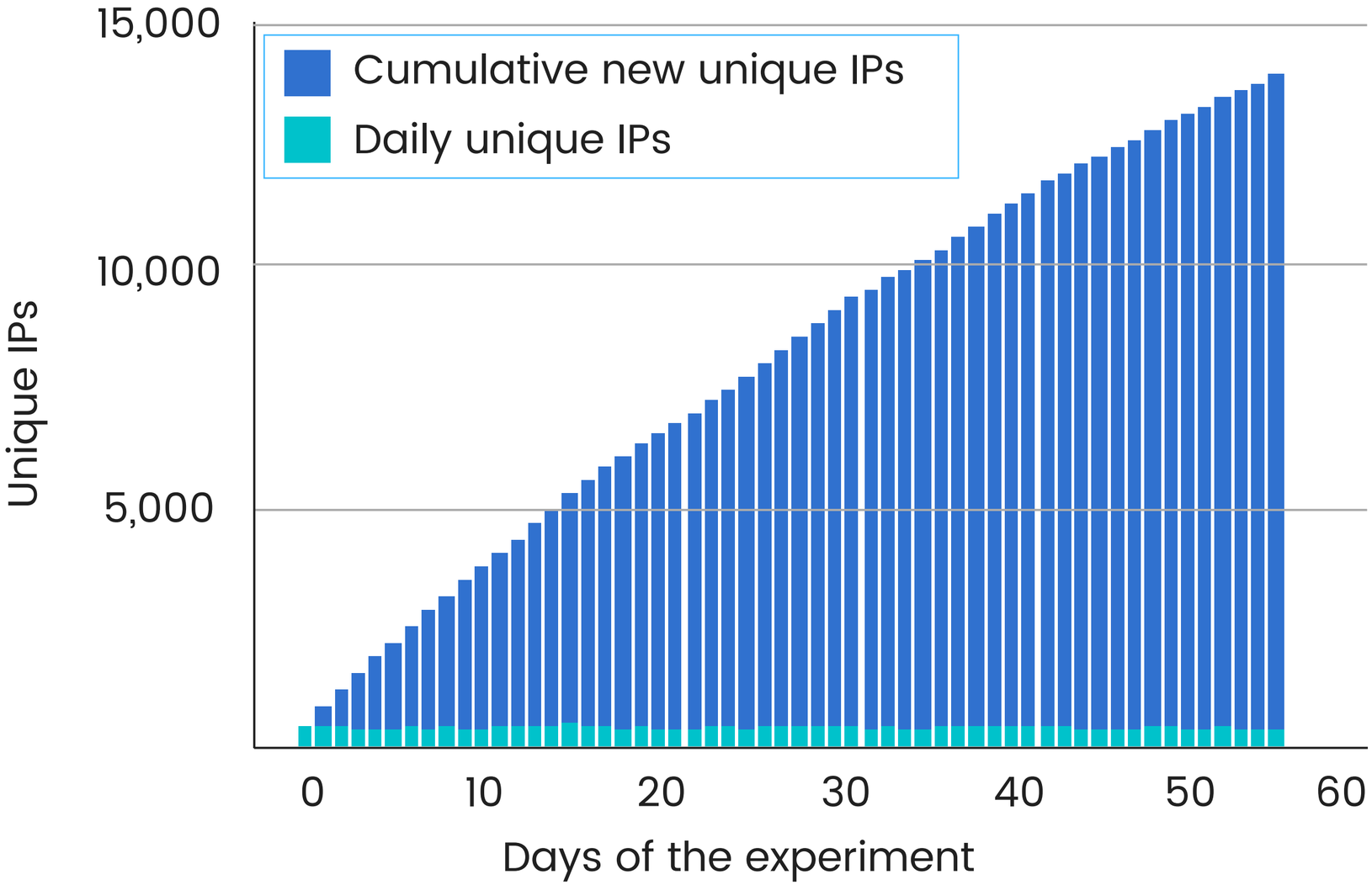
- Approximate result: $1-\left(\frac{P-1}{P}\right)^{\frac{56(55-1)}{2}}$
- $P=10{,}000{,}000 \rightarrow p(56,10M)\approx 0.000154$
- $P=1{,}000{,}000 \rightarrow p(56,1M)\approx 0.001538$
- $P=100{,}000 \rightarrow p(56,100K)\approx 0.015282$

Is this likely to happen?

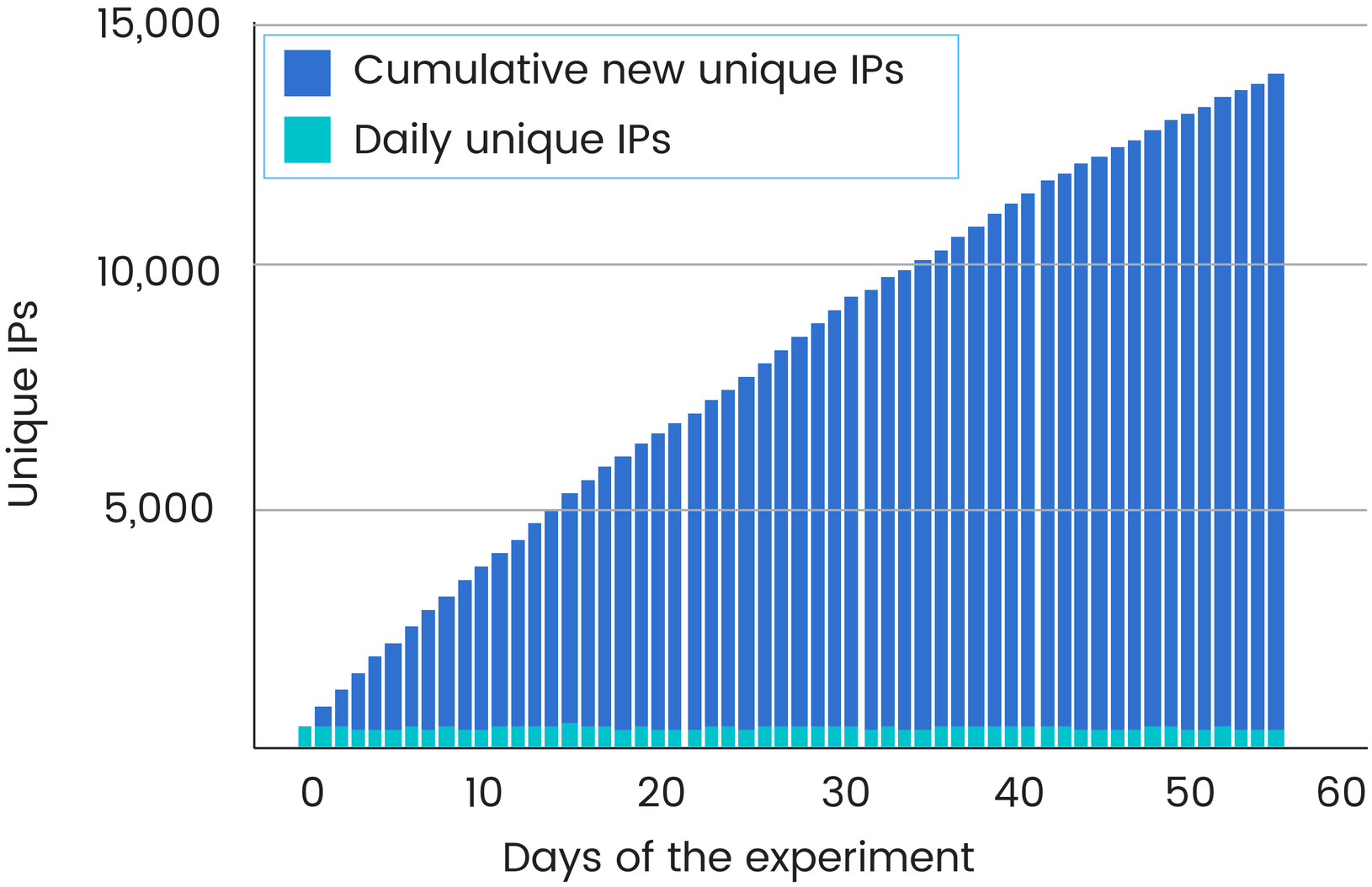
- Approximate result: $1-\left(\frac{P-1}{P}\right)^{\frac{56(55-1)}{2}}$
- P=10,000,000 → $p(56,10M)\approx0.000154$
- P=1,000,000 → $p(56,1M)\approx0.001538$
- P=100,000 → $p(56,100K)\approx0.015282$

P is significantly lower than the claimed numbers AND/OR the assignment is not randomly done

Cumulative curve of new unique IPs

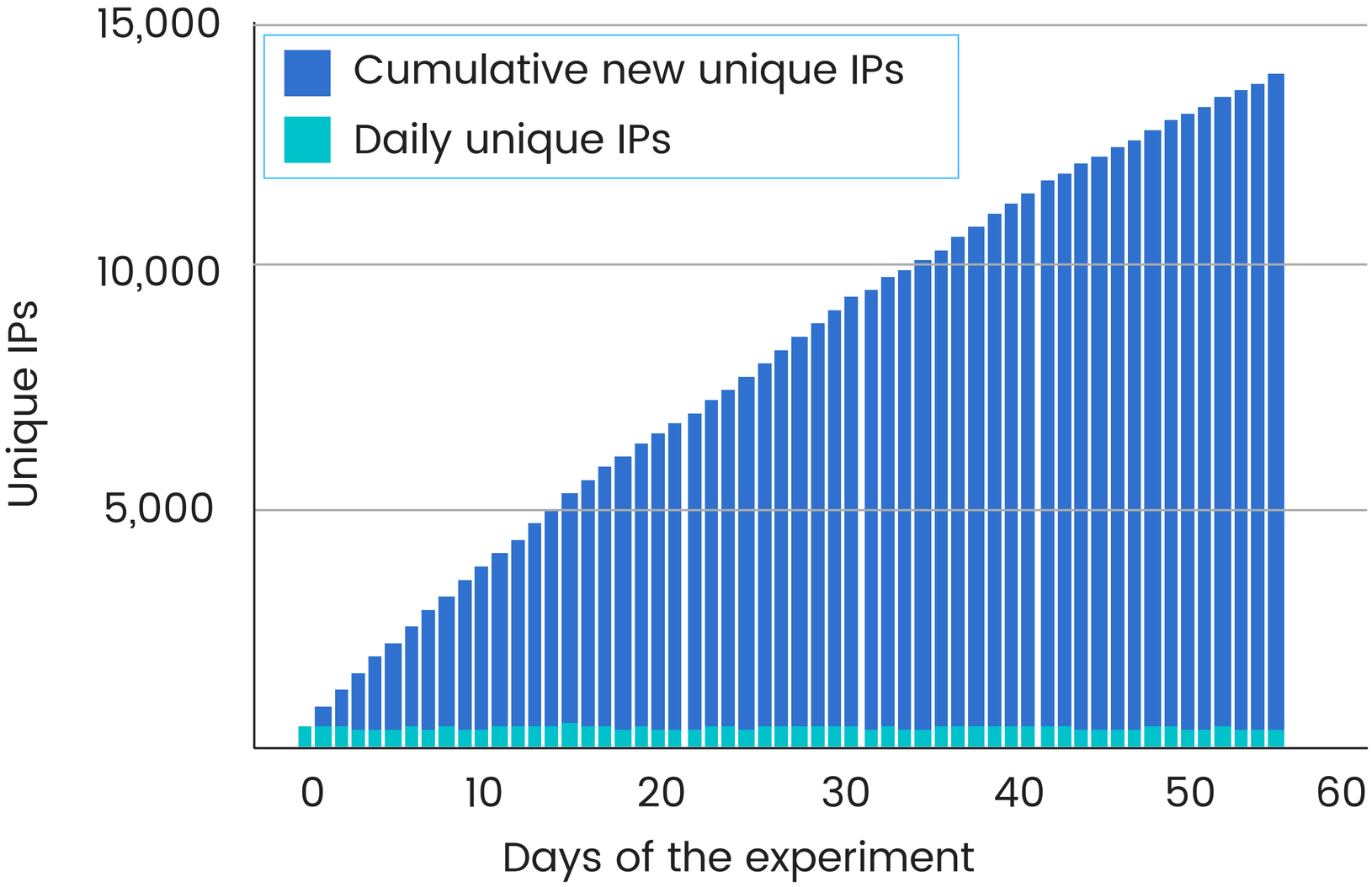


Cumulative curve of new unique IPs



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.

IP assignment as a drawing process

- Model the assignment process by a daily probabilistic drawing process without replacement

IP assignment as a drawing process

- Model the assignment process by a daily probabilistic drawing process without replacement
- Arbitrarily define a pool size

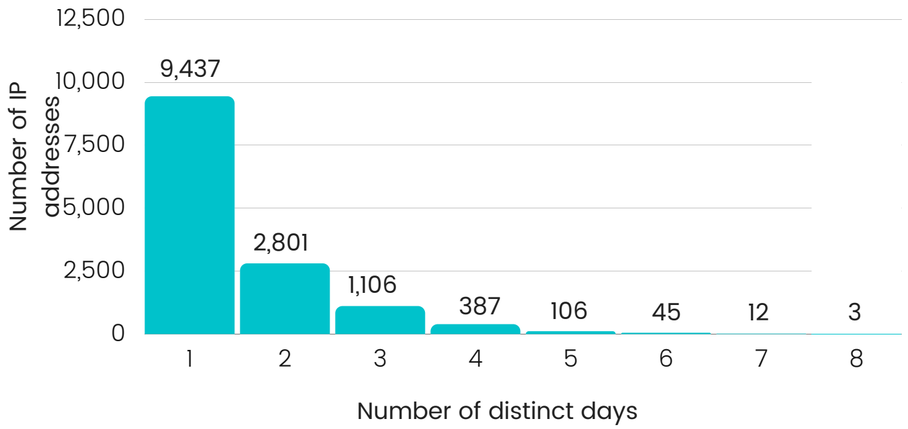
IP assignment as a drawing process

- Model the assignment process by a daily probabilistic drawing process without replacement
- Arbitrarily define a pool size
- On a given day, draw from the pool, without replacement, a number of values equal to the amount of distinct IPs seen that day

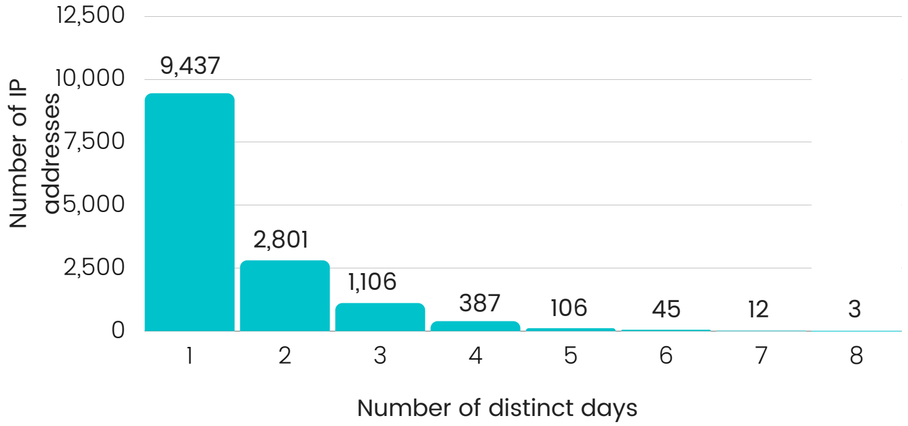
IP assignment as a drawing process

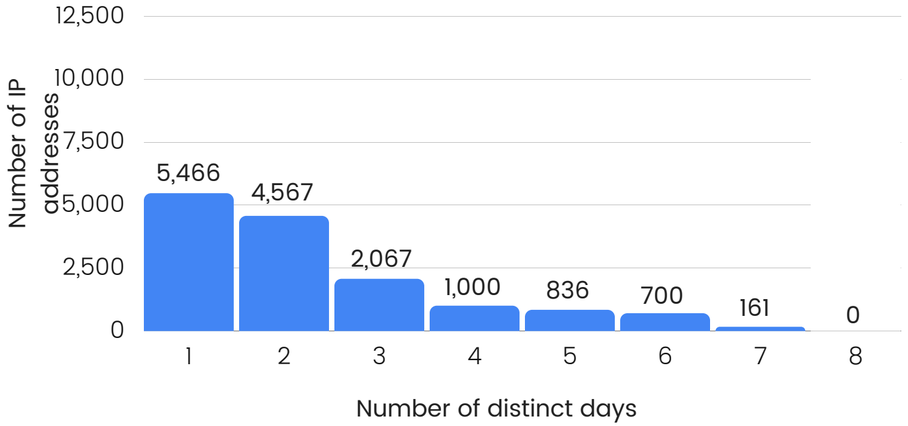
- Model the assignment process by a daily probabilistic drawing process without replacement
- Arbitrarily define a pool size
- On a given day, draw from the pool, without replacement, a number of values equal to the amount of distinct IPs seen that day
- Do it for all the days of the experiment and build a histogram with the number of IPs seen in distinct days

Histograms comparison

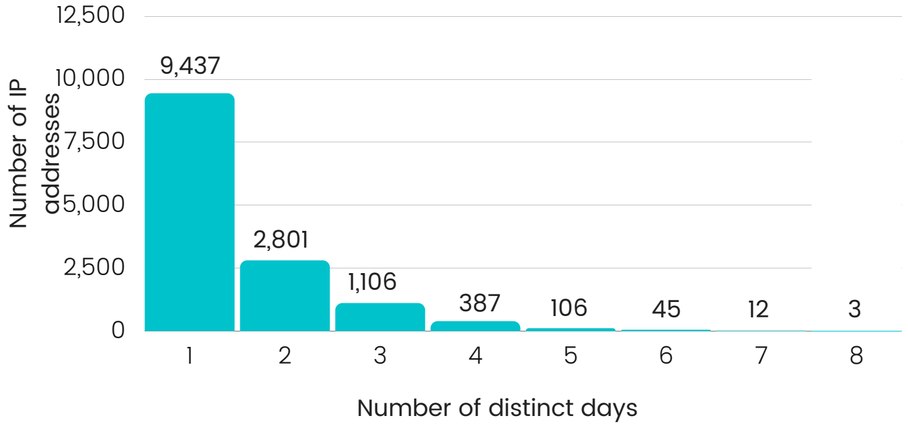


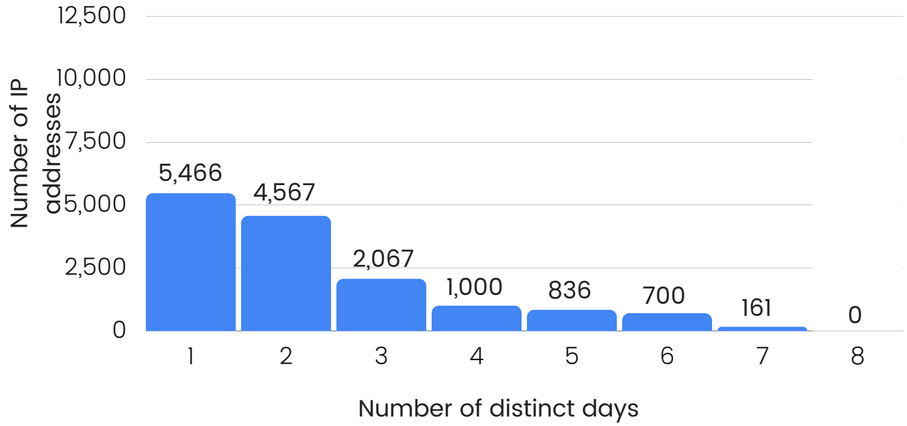
Histograms comparison

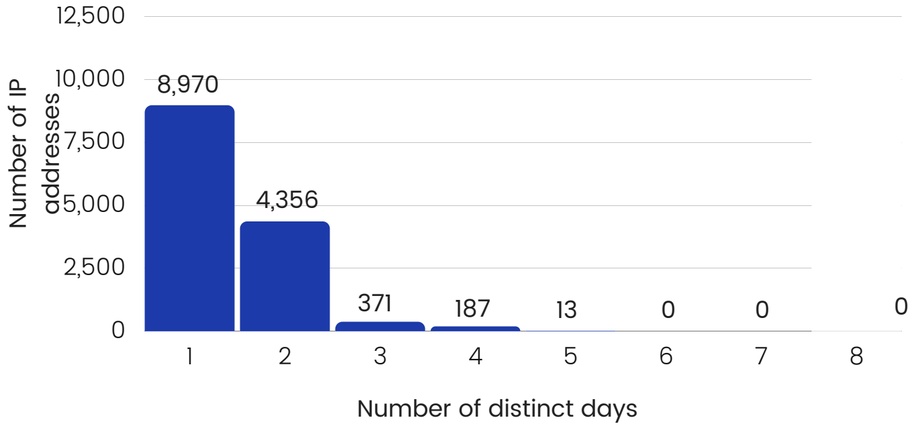




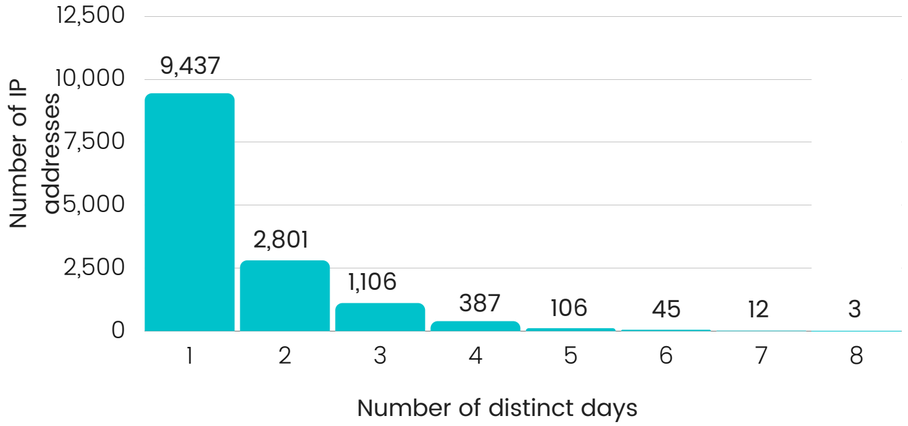
Histograms comparison

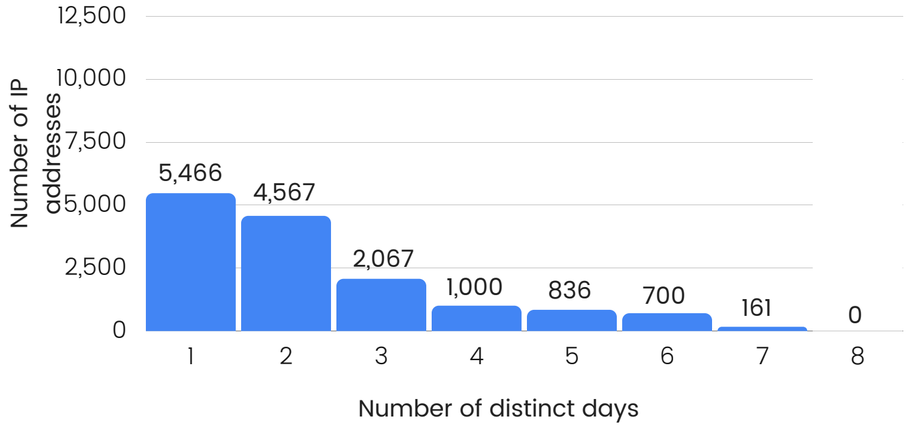


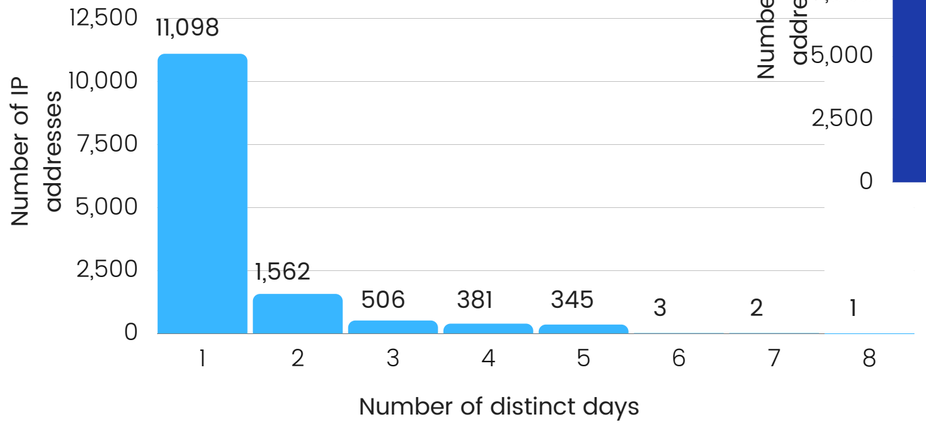


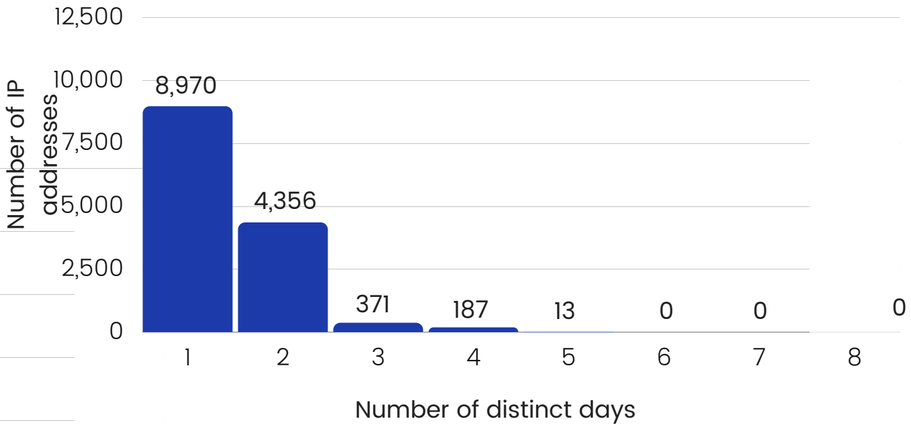


Histograms comparison

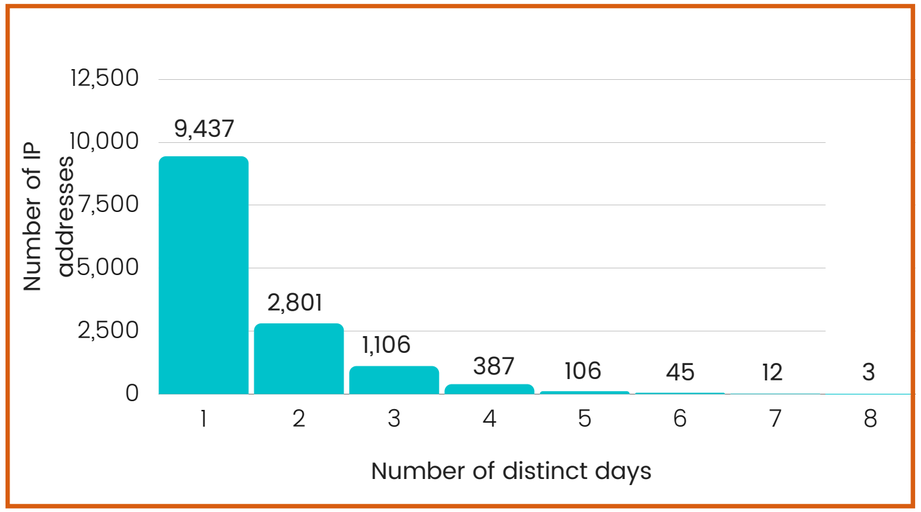


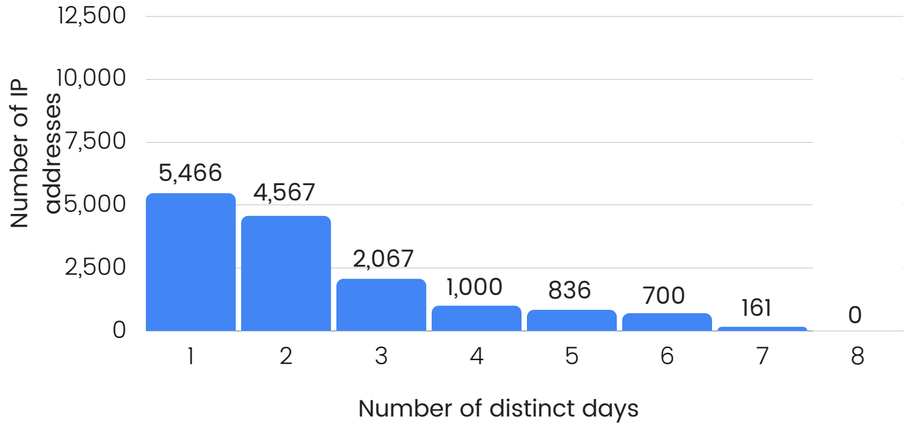


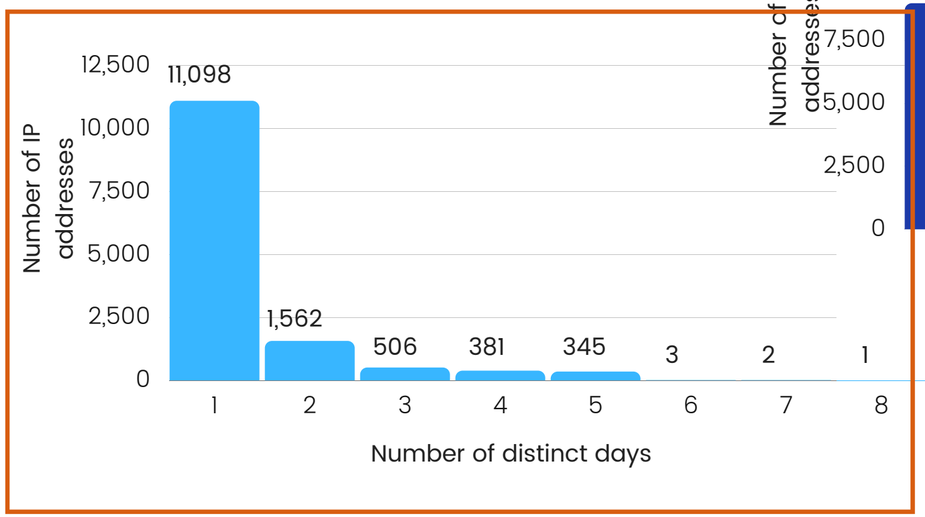


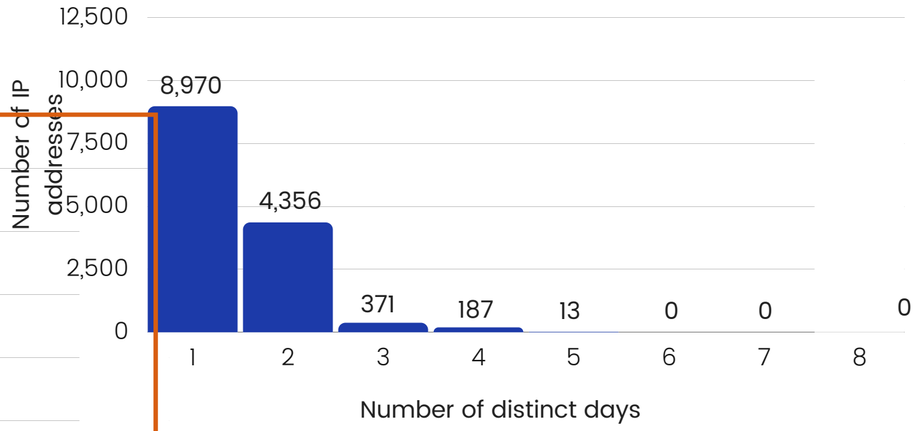


Histograms comparison

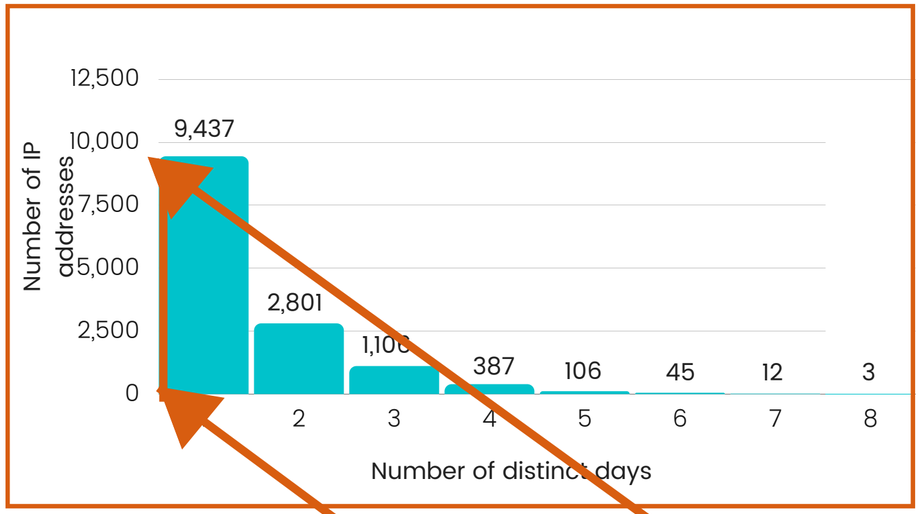


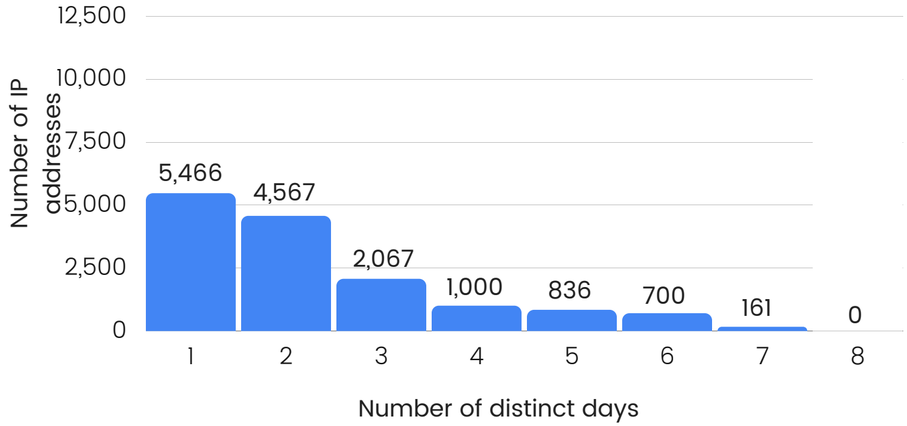


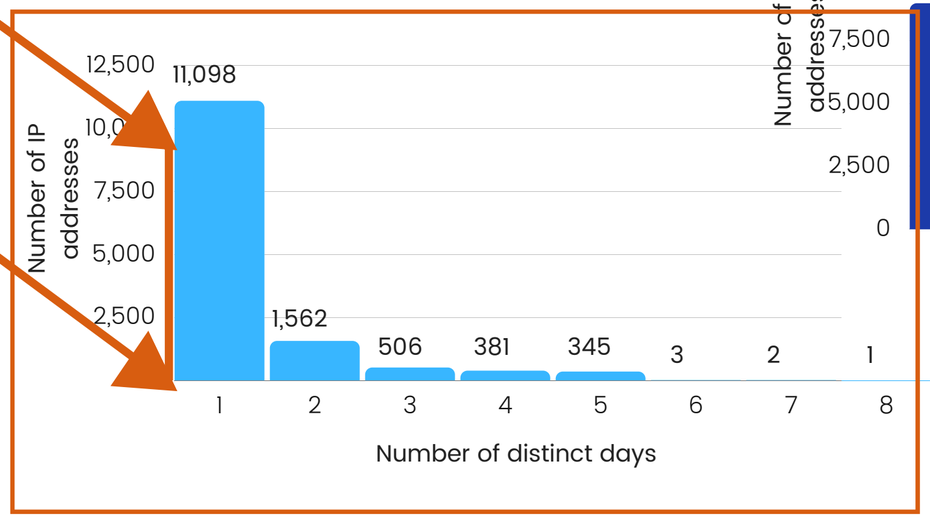


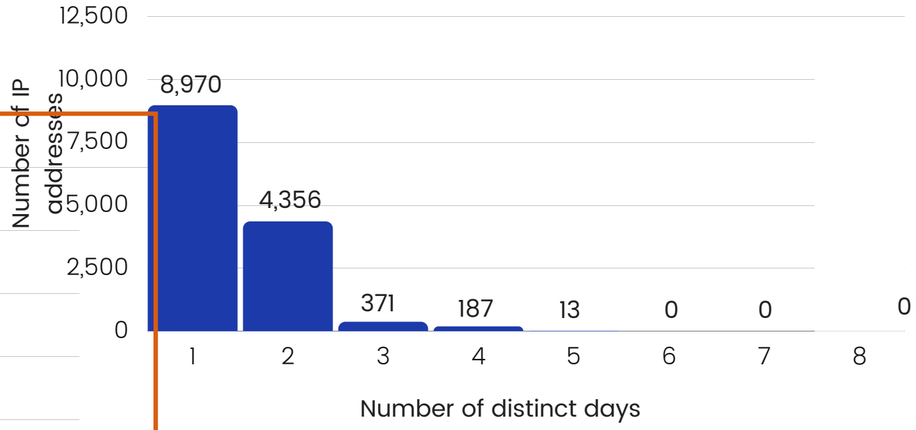


Histograms comparison

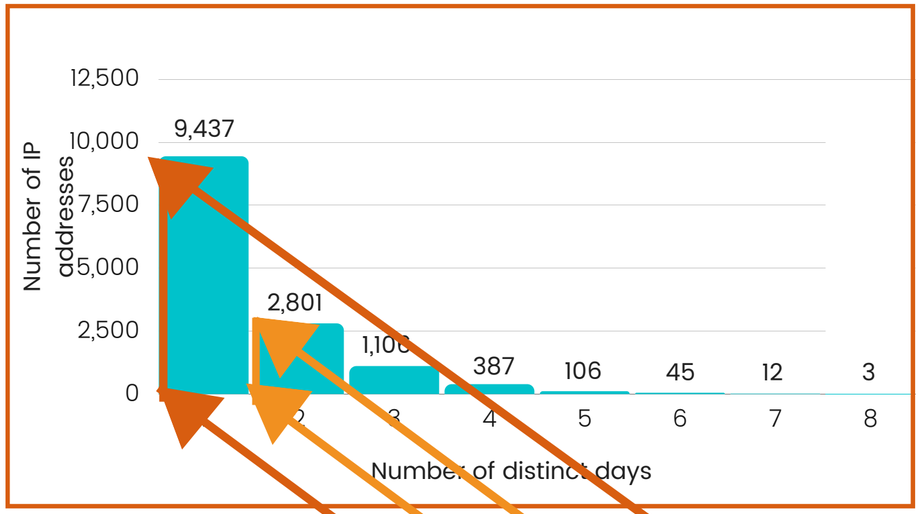


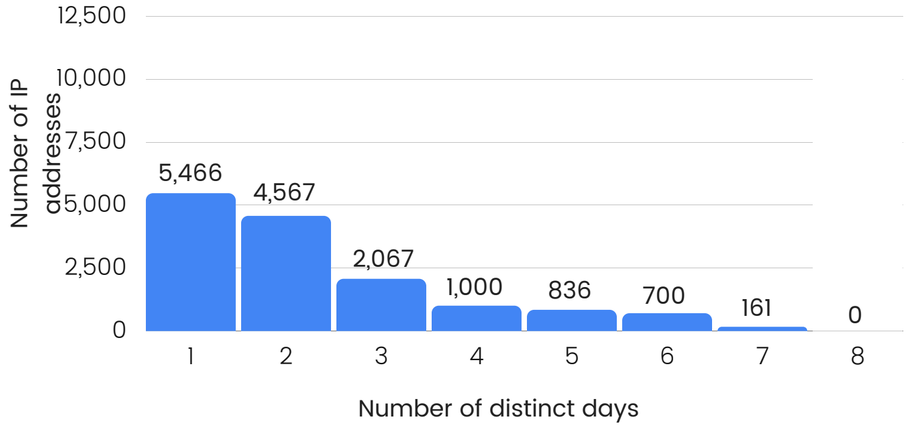


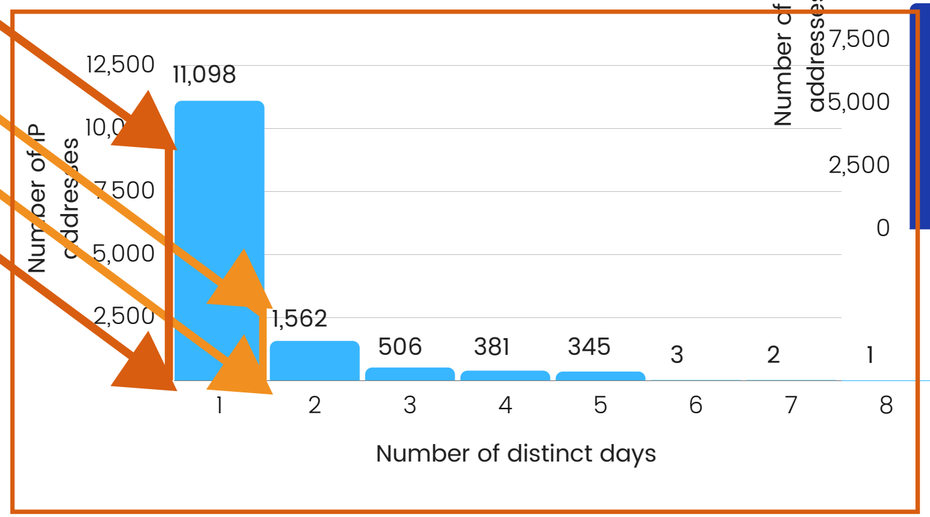


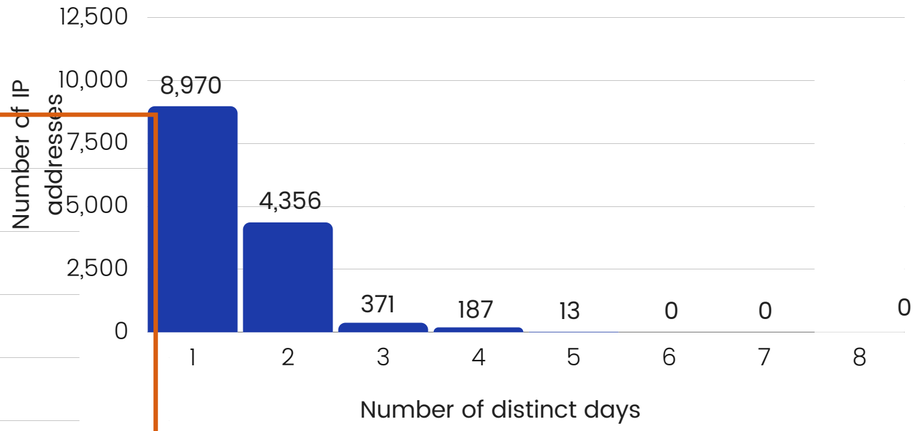


Histograms comparison









Wasserstein distance

“the minimum amount of "work" required to transform one histogram into another, where "work" is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved”

The value of which produces the histogram with the lowest distance from the empirical data corresponds to the size P which best represents the observed data.

What if the drawing is not daily based?

Repeat the process with window sizes from 2 to 10 days

What if the drawing is not daily based?

- Repeat the process with window sizes from 2 to 10 days
- Drawing with replacement

What if the drawing is not daily based?

- Repeat the process with window sizes from 2 to 10 days
- Drawing with replacement
- Additional constraint: a given value cannot be drawn more than times, i.e. once per day

Findings

- **Uniform distribution:** best population size equal to 20,000

Findings

Uniform distribution: best population size equal to 20,000

Uniform distribution means random picking

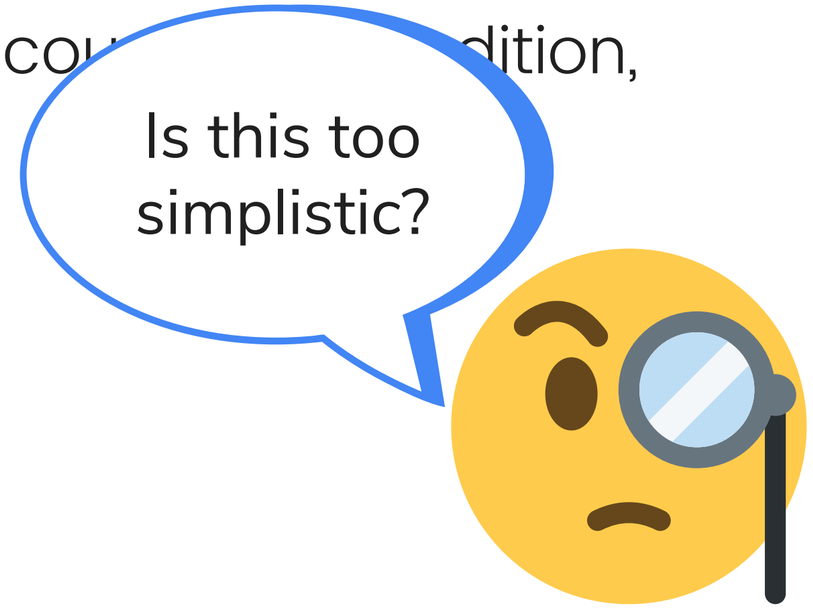
- All IPs available all the time
- Selection done without taking into account any condition, e.g. geo-localization

Findings

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 accou dition, e.g. geo-localization



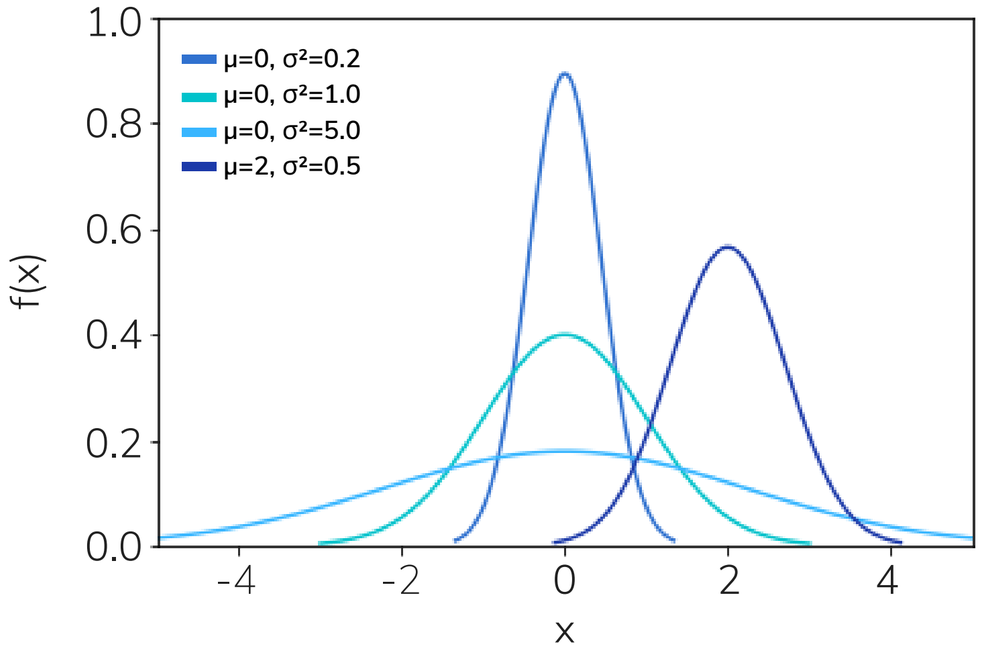
Let's try with other distributions

Model different probability to be picked by different IPs

Let's try with other distributions

Model different probability to be picked by different IPs

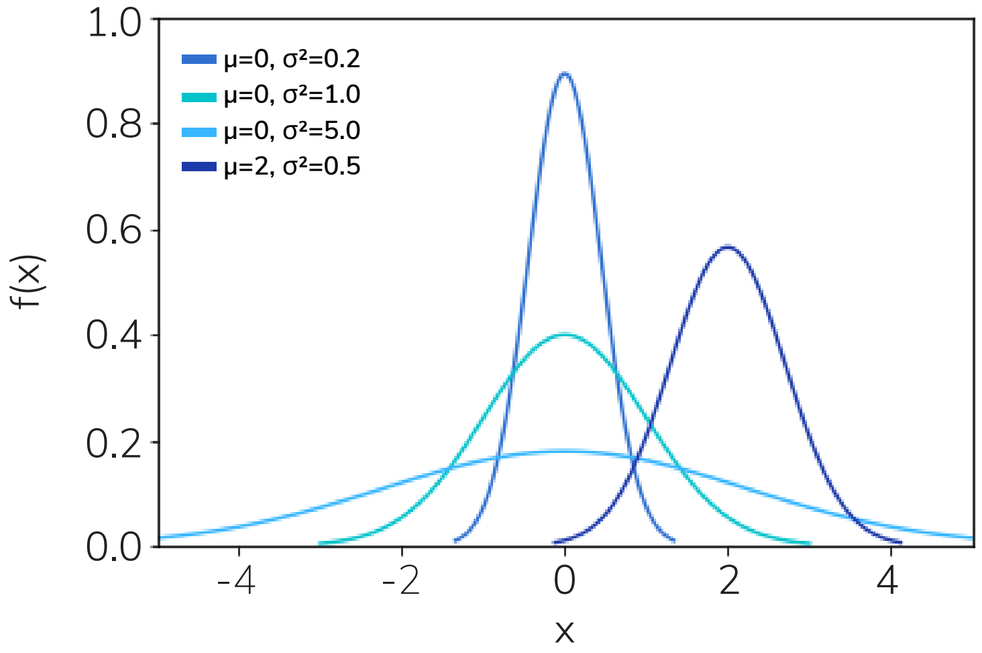
- **Gaussian**



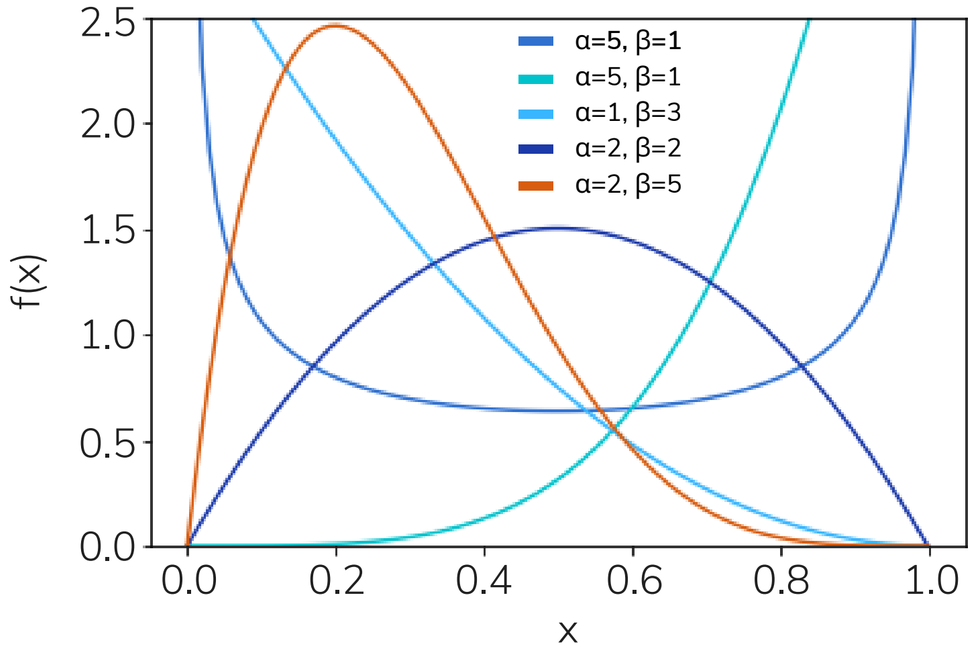
Let's try with other distributions

Model different probability to be picked by different IPs

- **Gaussian**



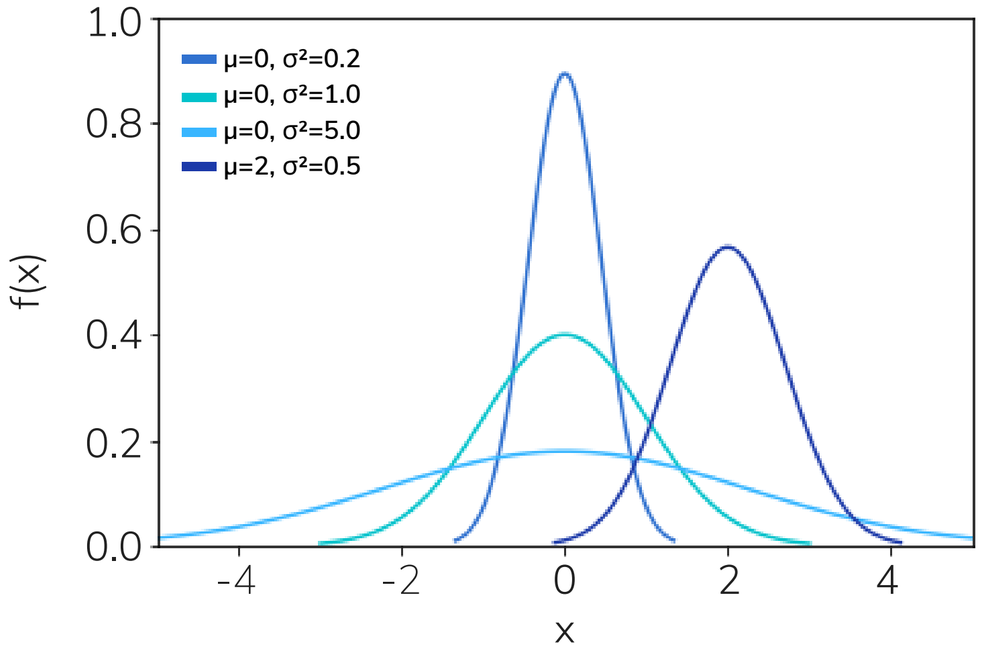
- **Beta**



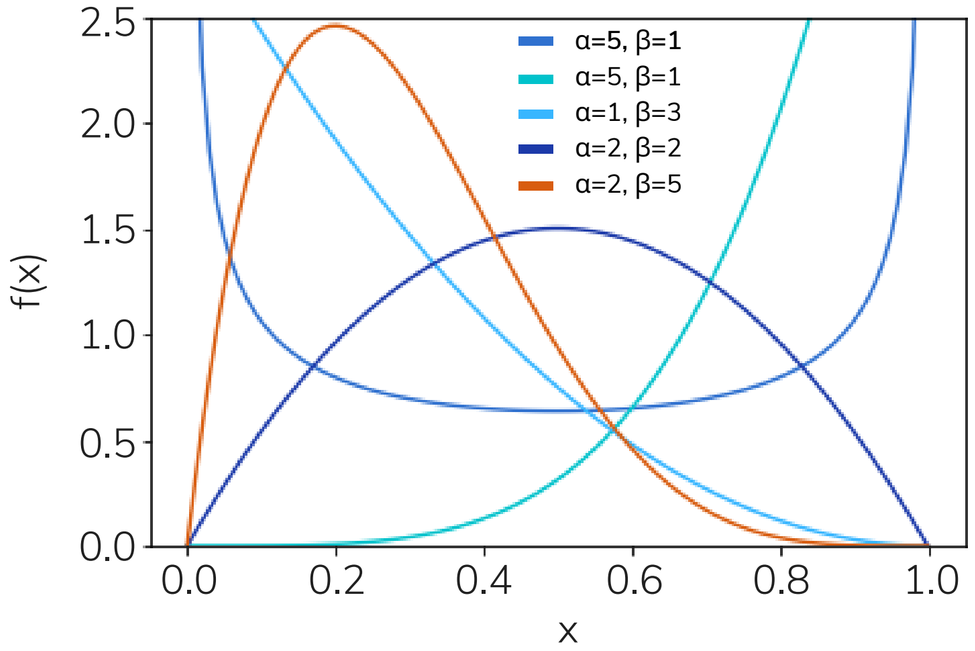
Let's try with other distributions

Model different probability to be picked by different IPs

- **Gaussian**



- **Beta**



Take the combination of parameters giving us the lowest Wasseinsten distances

Bias results

- Gaussian distribution: best population size equal to 60,000
- Beta distribution: best population size equal to 60,000

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?

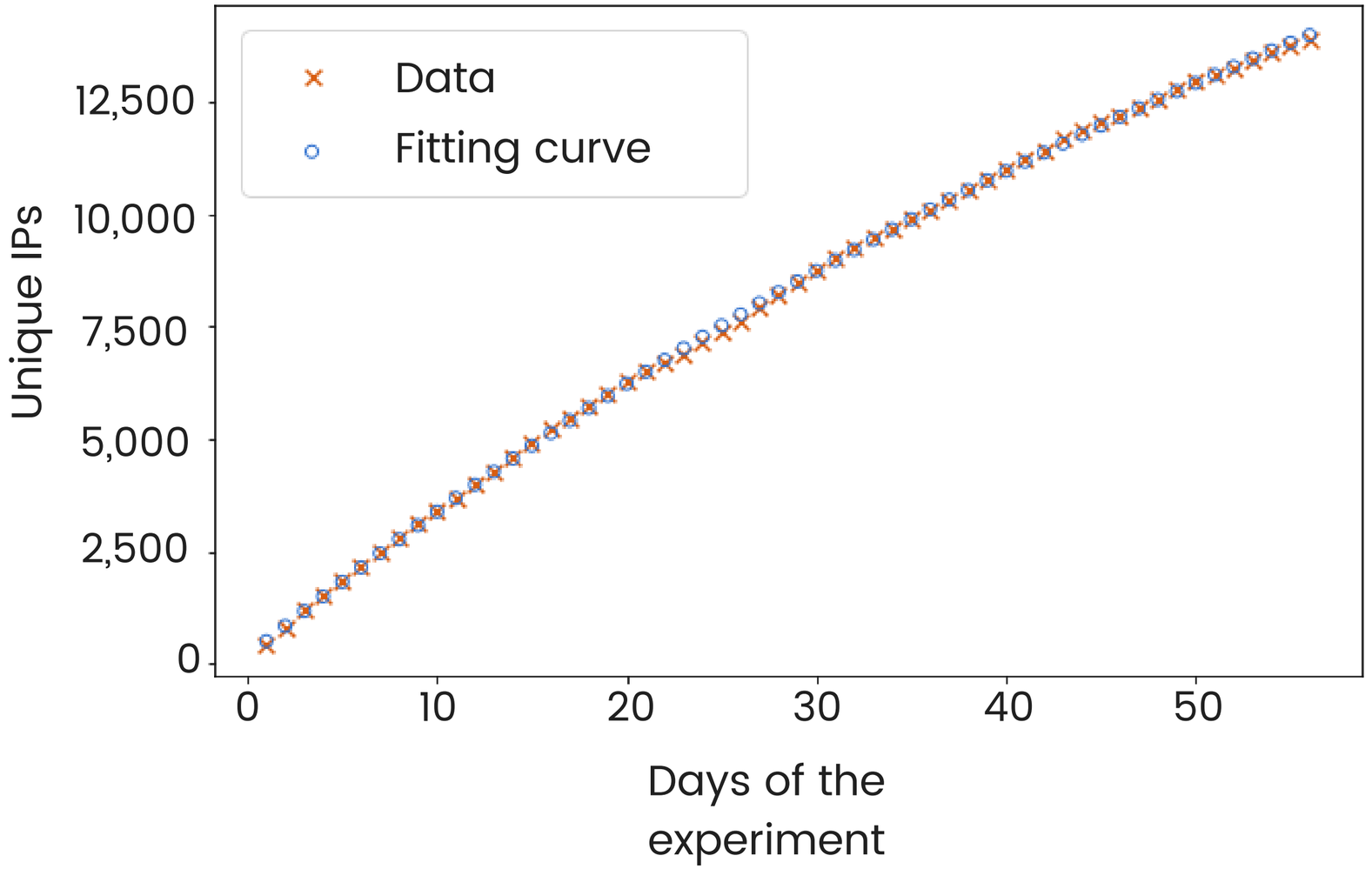
- For all considered distributions, the found size of the pool is significantly smaller than proxies claim.
- This does not directly mean the proxies do not own millions of IPs.
- BUT it suggests that there is a not a complete allocation of the IPs.

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!

Fitting cumulative curve of new unique IPs



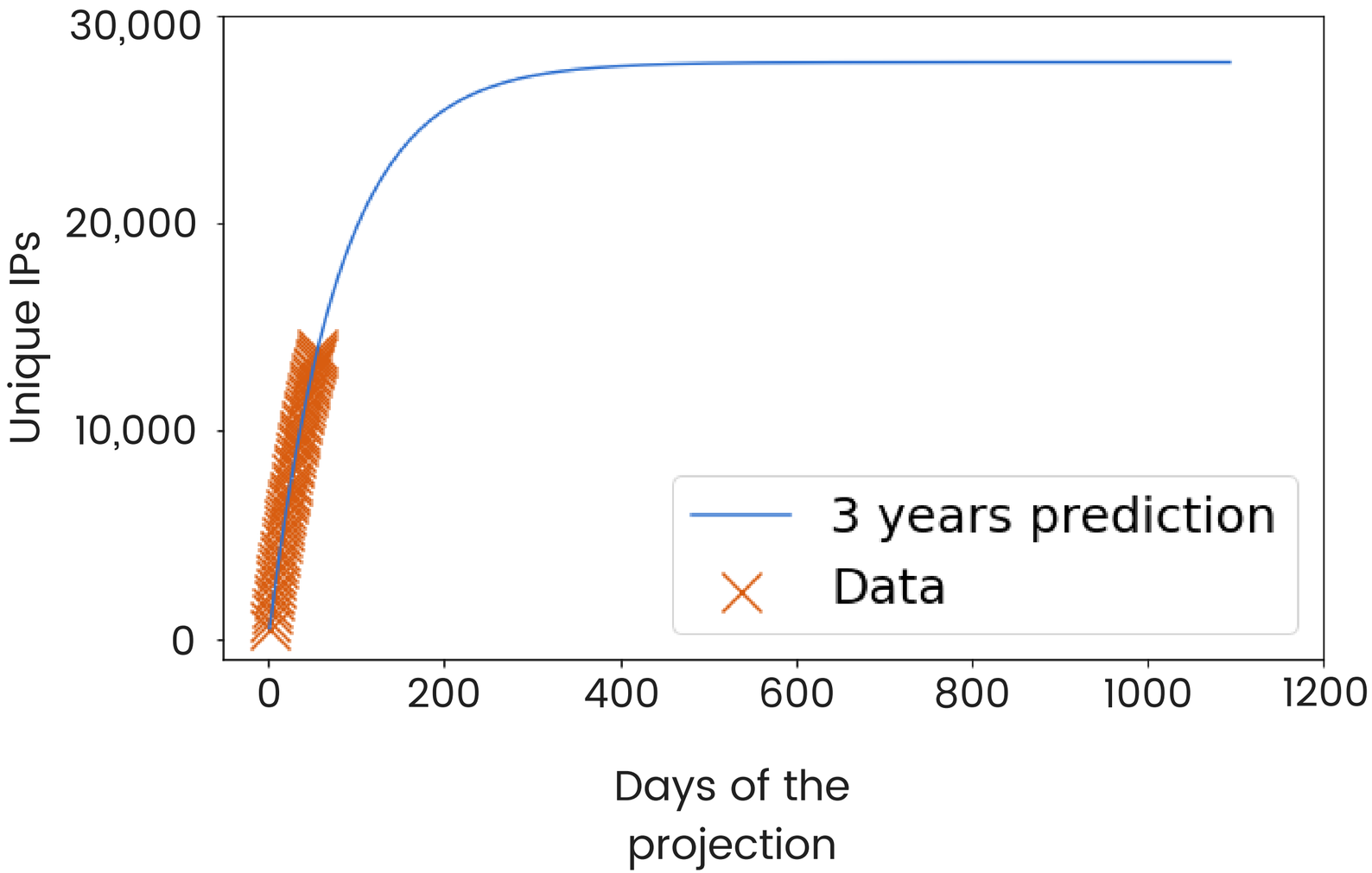
$$f = a^* \left(1 - e^{\frac{-(x-b)}{c}}\right)$$

a=2.77e+04

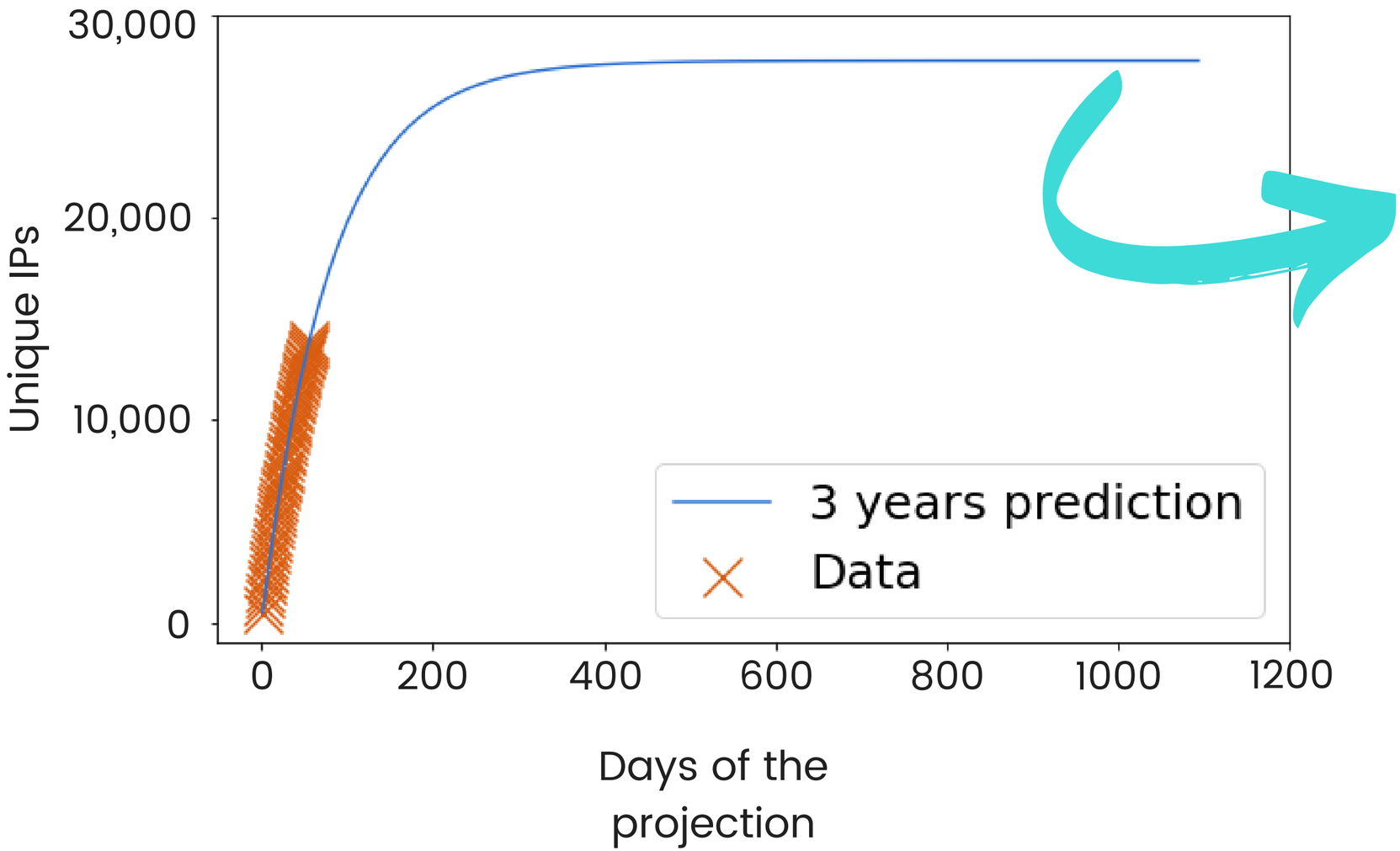
b=-4.78e-01

c=8.05e+01

Three years prediction

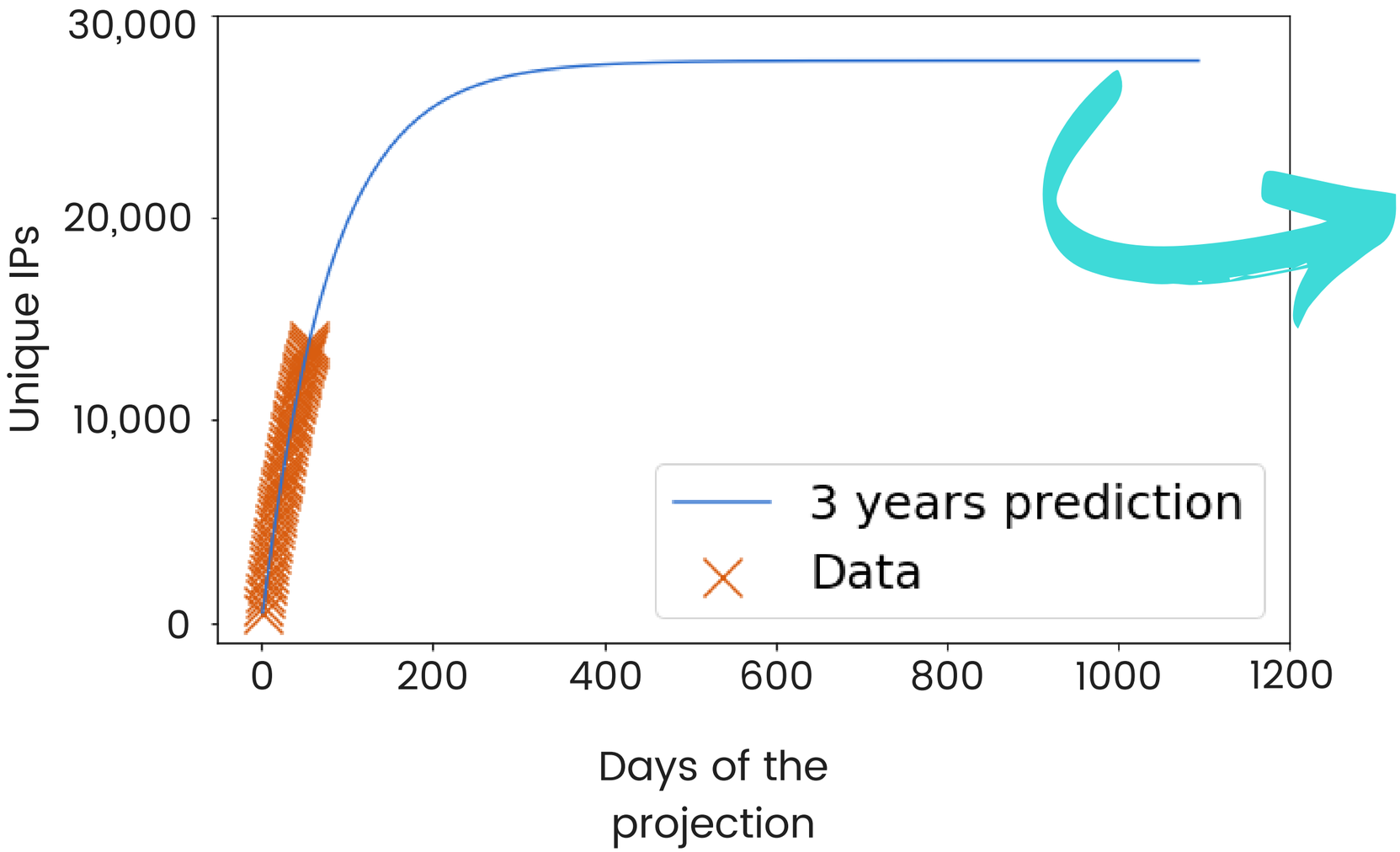


Three years prediction



The plateau is less than 30,000 IPs

Three years prediction



The plateau is less than 30,000 IPs

Consistent with the previous approach

4. Conclusions and future work

What we have done so far

Shown bots which send one request per day

What we have done so far

Shown bots which send one request per day

Served modified prices without being detected for 56 days

What we have done so far

Shown bots which send one request per day

Served modified prices without being detected for 56 days

Found behavioral pattern that gives confidence they are all under the same bot master

What we have done so far

Seen 32% of IPs were reused

What we have done so far

Seen 32% of IPs were reused

Shown that collected IPs were likely provided by proxy services

What we have done so far

Seen 32% of IPs were reused

Shown that collected IPs were likely provided by proxy services

Modelled IPs at the botnet's disposal are in the low tens of thousands

What we have done so far

Seen 32% of IPs were reused

Shown that collected IPs were likely provided by proxy services

Modelled IPs at the botnet's disposal are in the low tens of thousands

Two conference papers and a journal paper

And now? 🤔

And now? 🤔

- Implementation of the honeypot inside the system and more experiments
  - Study of other bot signatures
  - Verification of behavioral patterns as second layer detection

And now? 🤔

- Implementation of the honeypot inside the system and more experiments
  - Study of other bot signatures
  - Verification of behavioral patterns as second layer detection
- In-depth study of proxies ecosystems

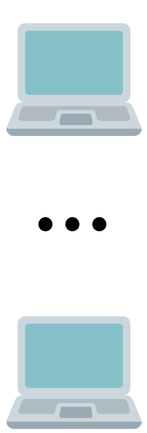
BAD PASS: Bots taking ADvantage of Proxy AS a Services

BAD PASS: Bots taking ADvantage of Proxy AS a Services

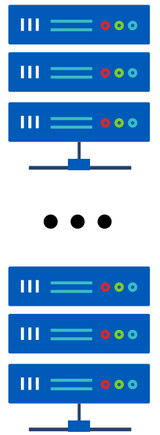
...

Clients

BAD PASS: Bots taking ADvantage of Proxy AS a Services

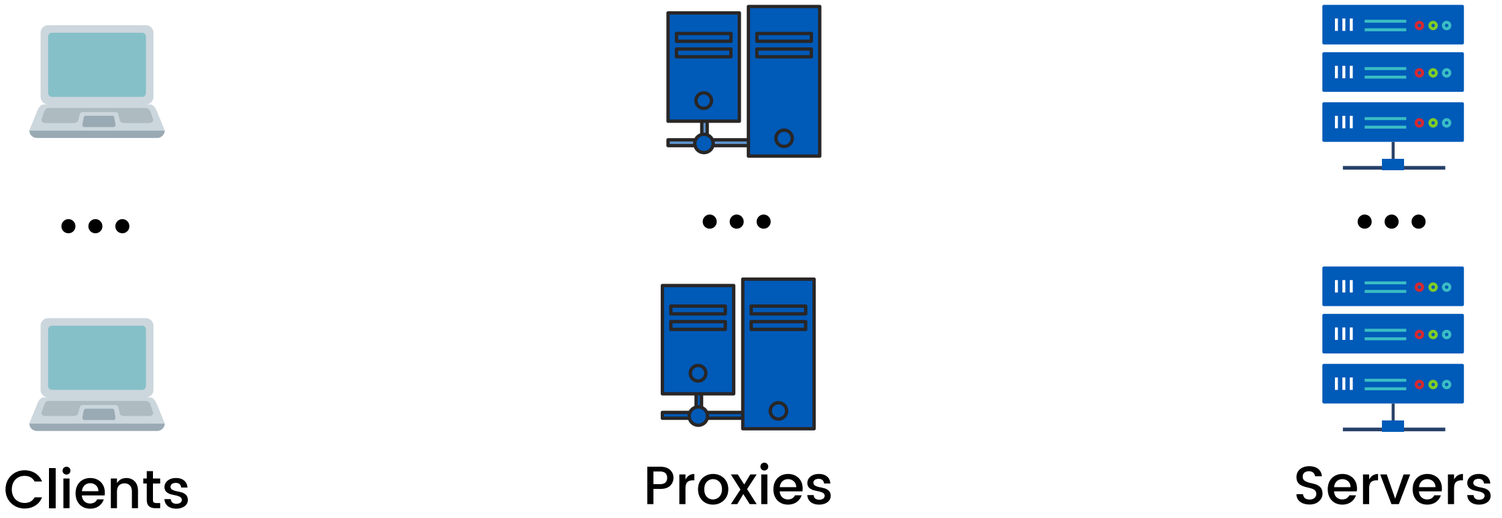


Clients

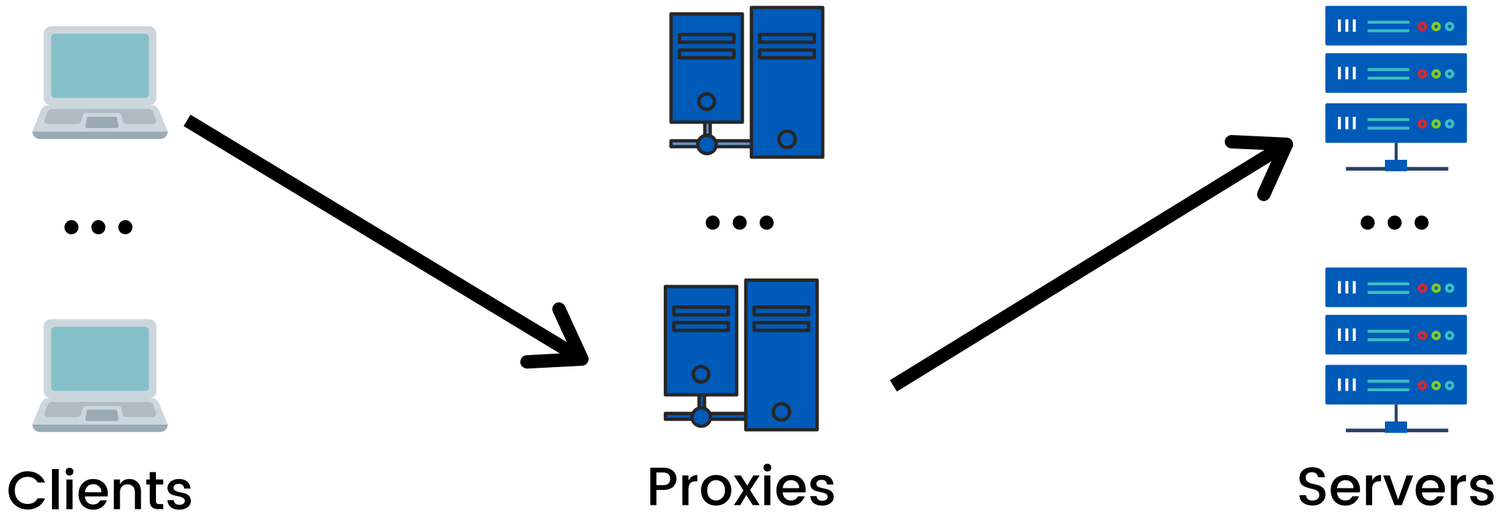


Servers

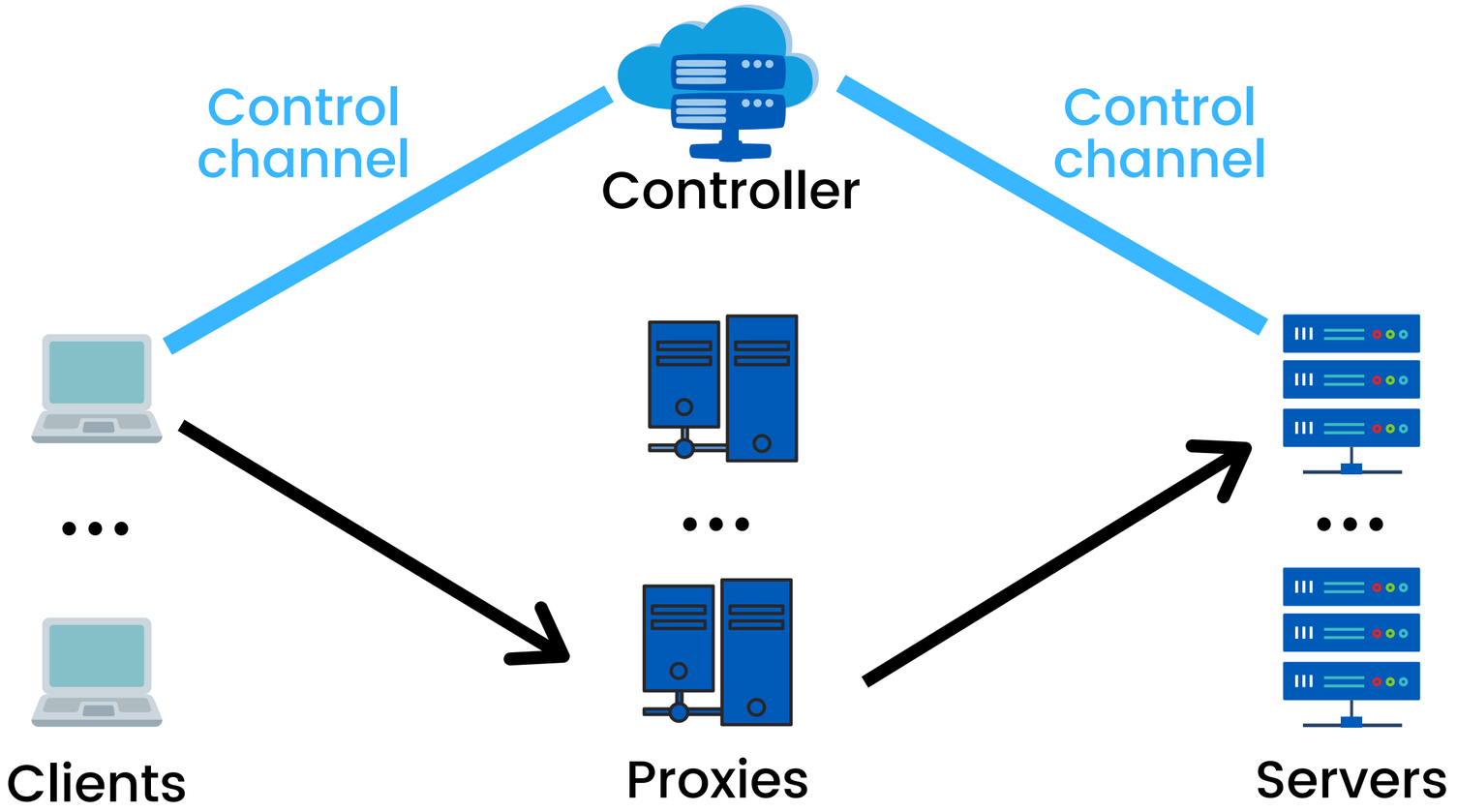
BAD PASS: Bots taking ADvantage of Proxy AS a Services



BAD PASS: Bots taking ADvantage of Proxy AS a Services



BAD PASS: Bots taking ADvantage of Proxy AS a Services



BAD PASS: Bots taking ADvantage of Proxy AS a Services

- Collection of proxy services exit points

BAD PASS: Bots taking ADvantage of Proxy AS a Services

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

BAD PASS: Bots taking ADvantage of Proxy AS a Services

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

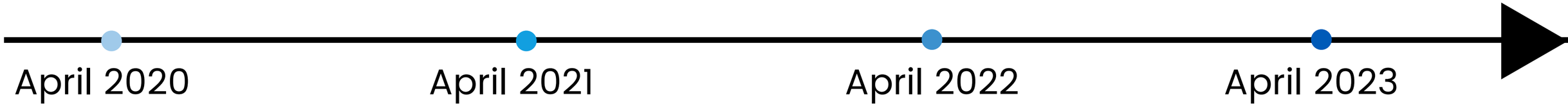
BAD PASS: Bots taking ADvantage of Proxy AS a Services

- 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

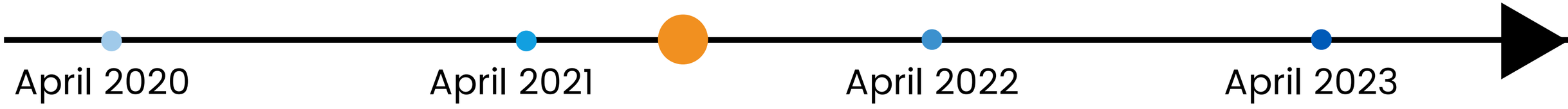
BAD PASS: Bots taking ADvantage of Proxy AS a Services

- 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
- Proxies fingerprinting

Timeline

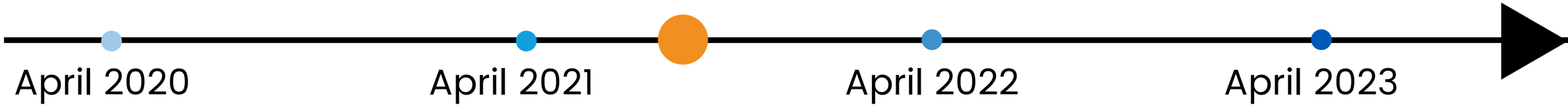


Timeline



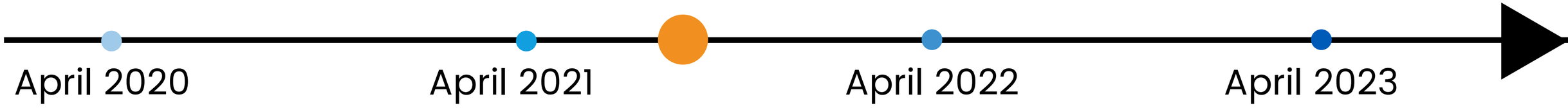
Timeline

First year: botnet study through honeypot, identification of proxy services problem



Timeline

First year: botnet study through honeypot, identification of proxy services problem

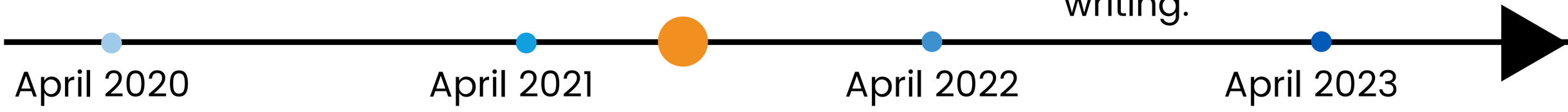


Second year: in-depth study of proxy services, experiments to know their structure. Further experiments with the honeypot.

Timeline

First year: botnet study through honeypot, identification of proxy services problem

Third-year: wider view on botnet ecosystems and experiments with various mitigation strategies. Thesis writing.



Second year: in-depth study of proxy services, experiments to know their structure. Further experiments with the honeypot.

Thank You

...Any question?

Our pubblications:

- Chiapponi, E., Dacier, M., Catakoglu, O., Thonnard, O., & Todisco, M. (2021). **Scraping Airlines Bots: Insights Obtained Studying Honeypot Data**. International Journal of Cyber Forensics and Advanced Threat Investigations, 2(1), 3-28. https://doi.org/10.46386/ijcfati.v2i1.23

- Chiapponi, E., Dacier, M., Todisco, M., Catakoglu, O., & Thonnard, O. (2021). **Botnet sizes: When maths meet myths**. Service-Oriented Computing–ICSOC 2020 Workshops (pagg. 596–611). Springer International Publishing. https://doi.org/10.1007/978-3-030-76352-7_52

- Chiapponi, E., Catakoglu, O., Thonnard, O., Dacier, M. (2020). **HoPLA: a Honeypot Platform to Lure Attackers**. Computer & Electronics Security Applications Rendez-vous (C&ESAR). https://www.eurecom.fr/publication/6366