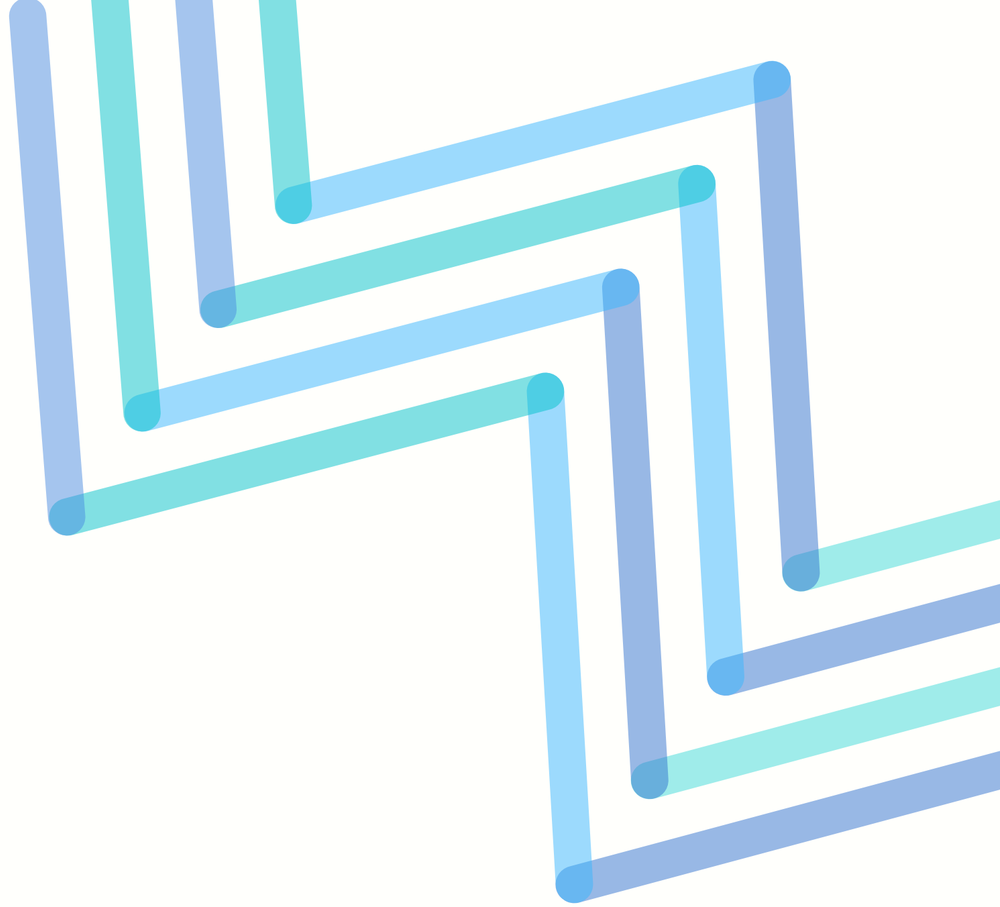
Botnet sizes: when maths meet myths

Elisa Chiapponi - EURECOM
Marc Dacier- EURECOM
Massimiliano Todisco - EURECOM
Onur Catakoglu - Amadeus IT Group
Olivier Thonnard - Amadeus IT Group



CFTIC 2020

EURECOM
Sophia Antipolis

aMaDEUS
Global Security Operations

Who are we?

Elisa Chiapponi

- Ph.D. student at EURECOM, collaboration with Amadeus IT Group
- Analysis and mitigation of the new generation of botnet

Prof. Marc Dacier

- Head of Digital Security Department, EURECOM

Prof. Massimiliano Todisco

- Assistant professor, Digital Security Department, EURECOM

Dr. Onur Catakoglu

- Information Security Architect, GSO Amadeus IT Group

Dr. Olivier Thonnard

- Senior Security Expert, Tech Lead GSO Amadeus IT Group

Agenda

Agenda

1. Introduction and motivations

Agenda

1. Introduction and motivations

2. Experimental setup

Agenda

1. Introduction and motivations
2. Experimental setup
3. Experimental raw results

Agenda

1. Introduction and motivations
2. Experimental setup
3. Experimental raw results
4. Modeling and discussion

Agenda

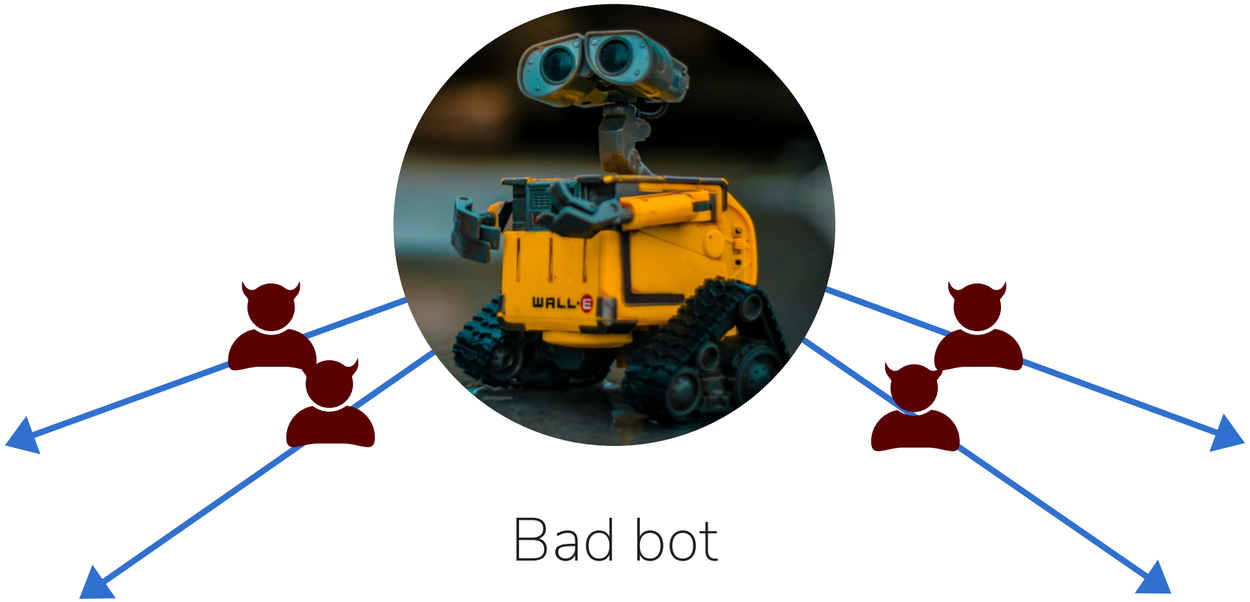
1. Introduction and motivations
2. Experimental setup
3. Experimental raw results
4. Modeling and discussion
5. Conclusion and future work

1. Introduction and motivations

Bad bots

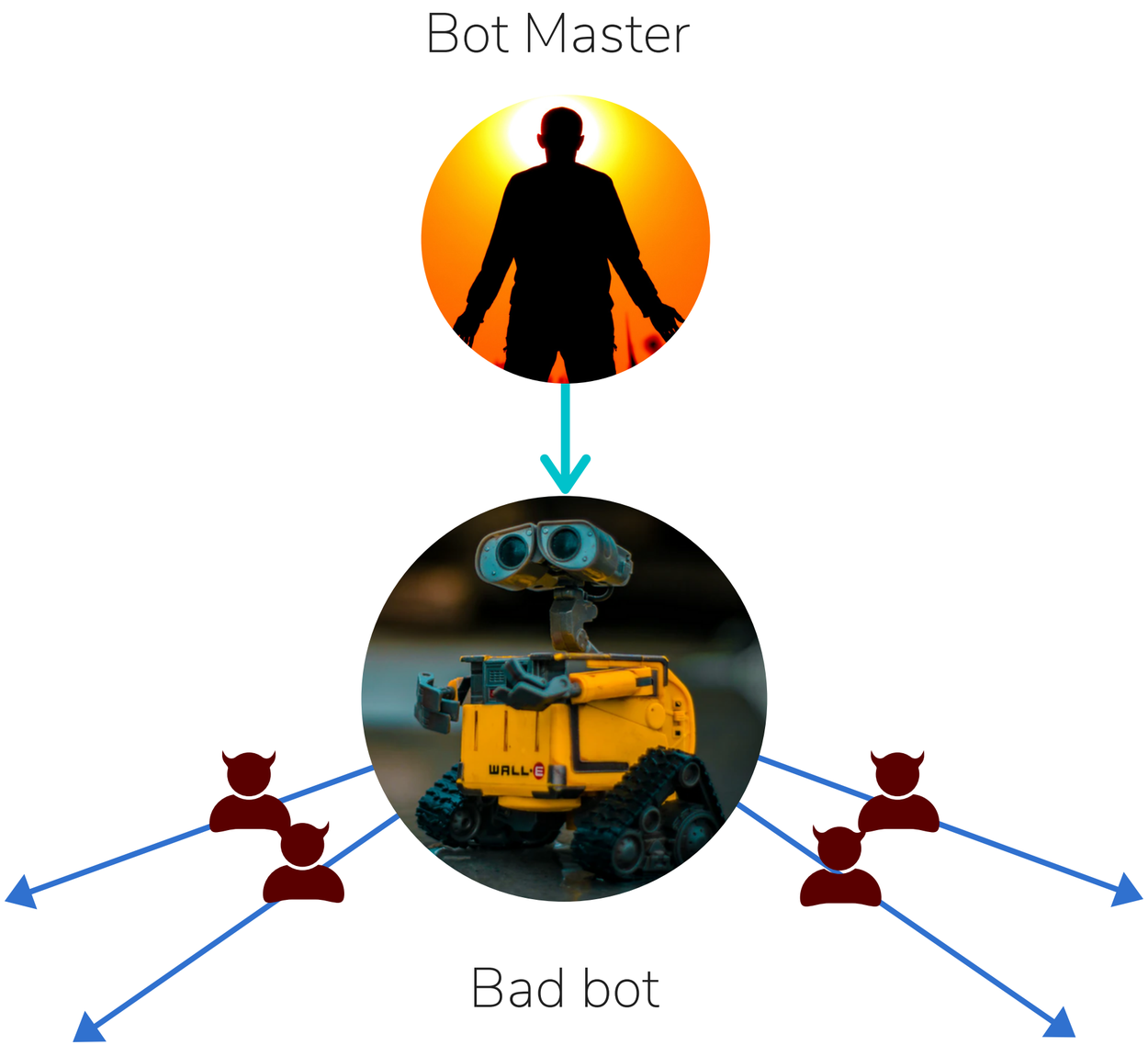
Bad bot

Bad bots

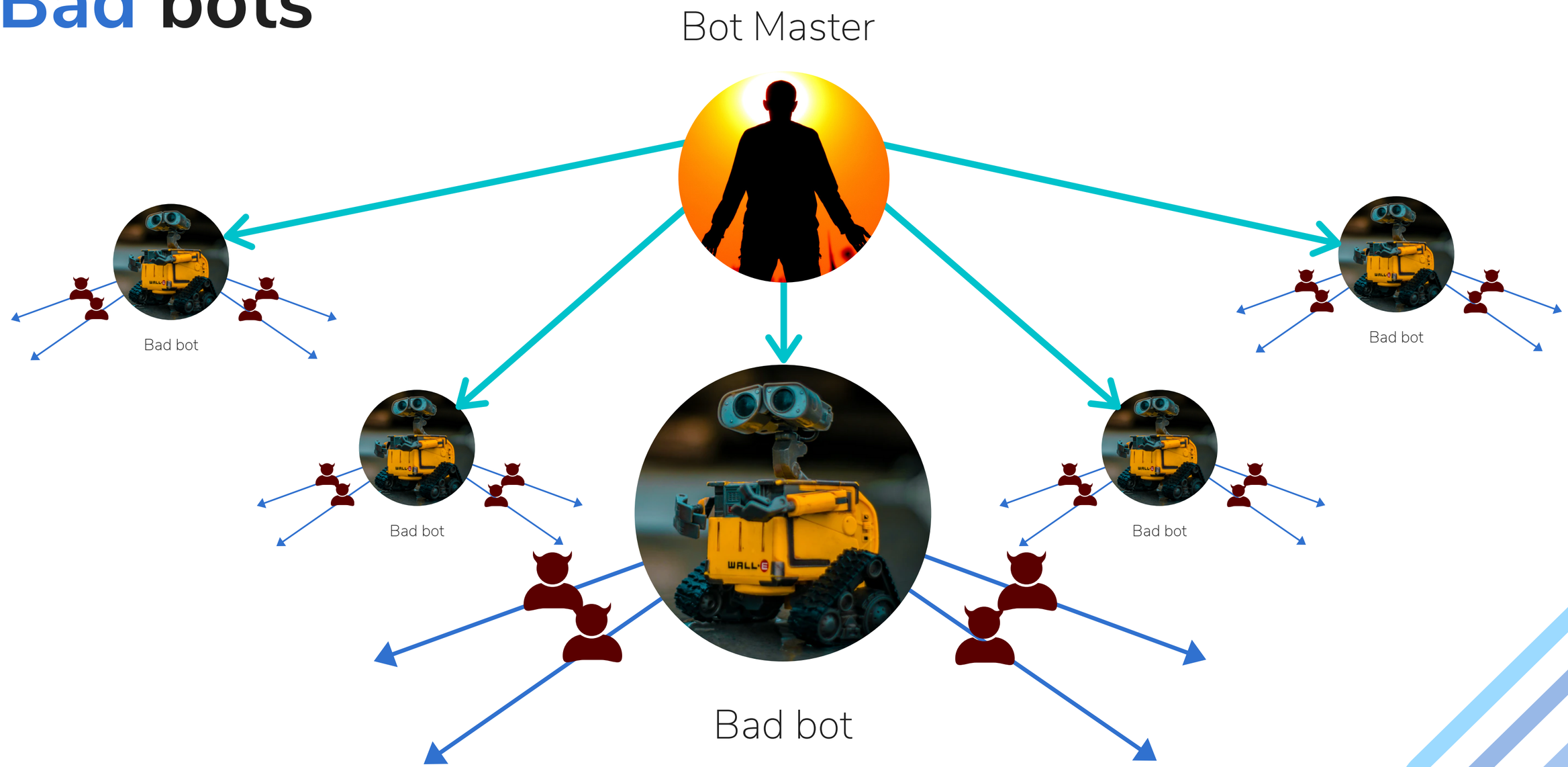


Bad bot

Bad bots



Bad bots

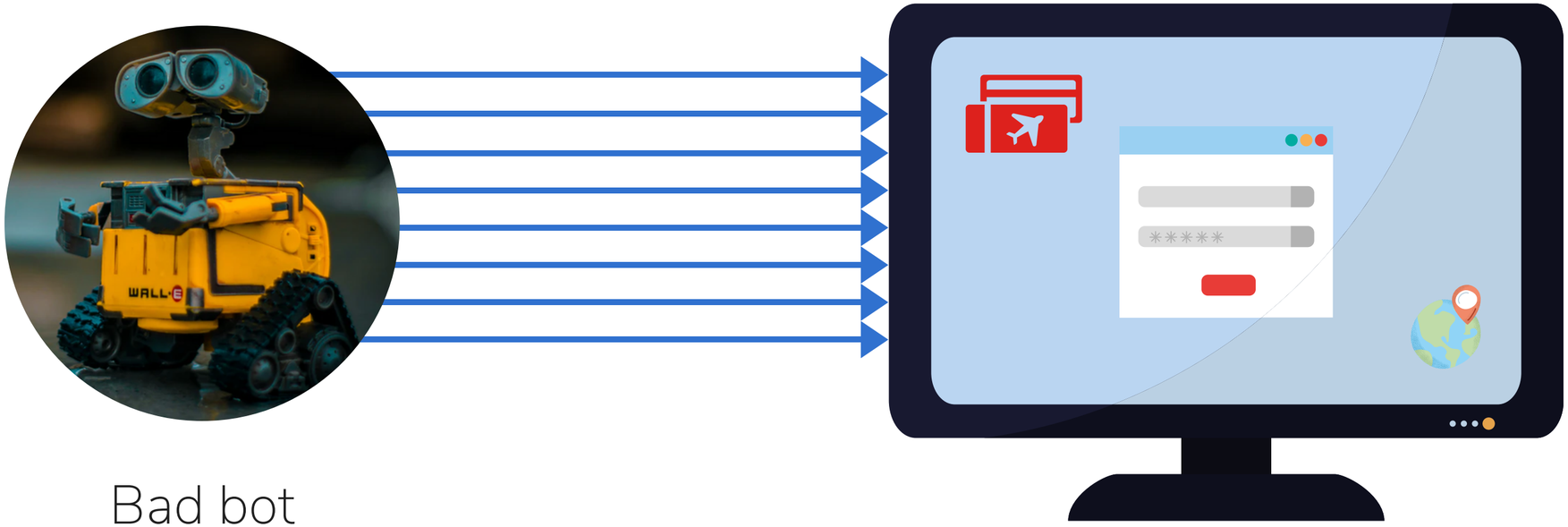


Web scraping bad bots

Bad bot

Airline booking domain

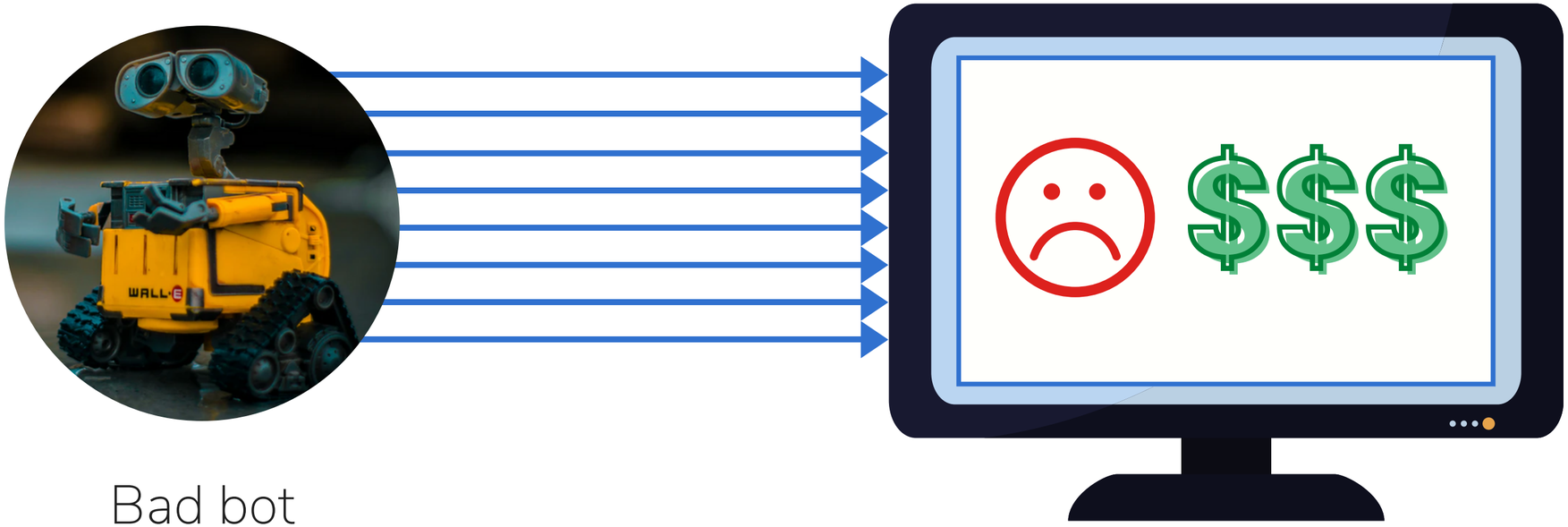
Web scraping bad bots



Bad bot

Airline booking domain

Web scraping bad bots

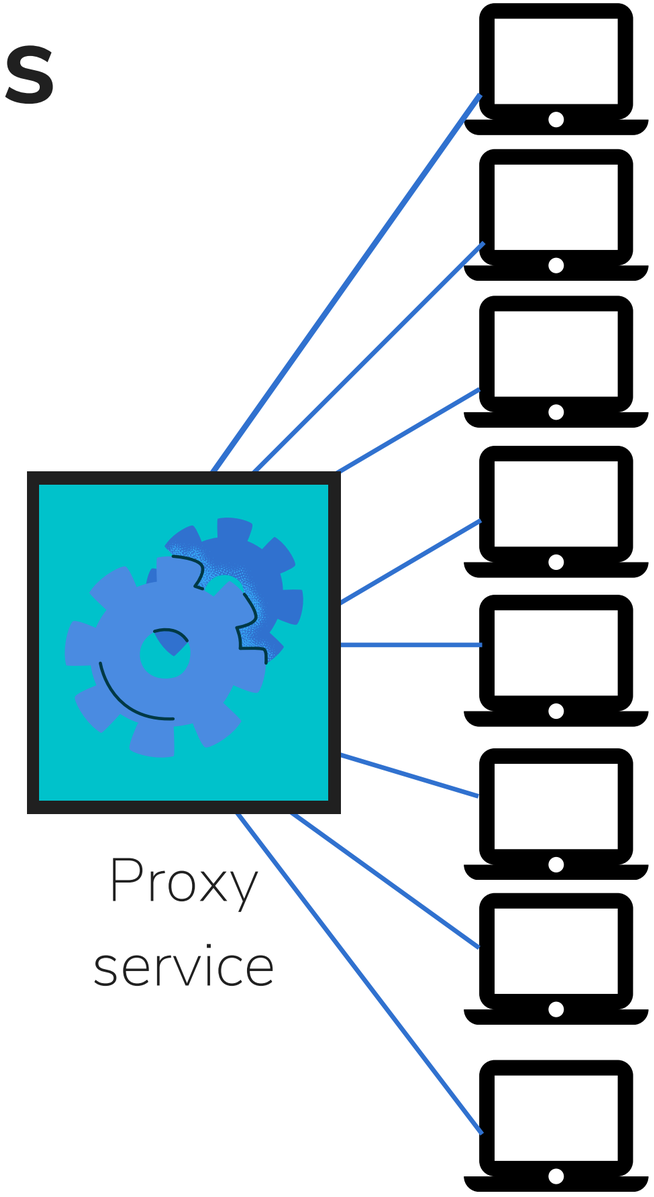


Bad bot

Airline booking domain

Proxy services

Bad bot

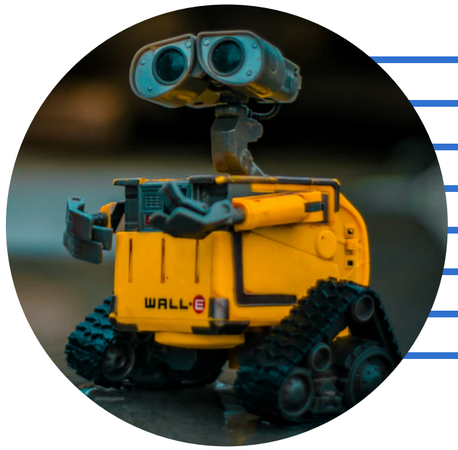


Proxy service

Proxy gateways

Airline booking domain

Proxy services



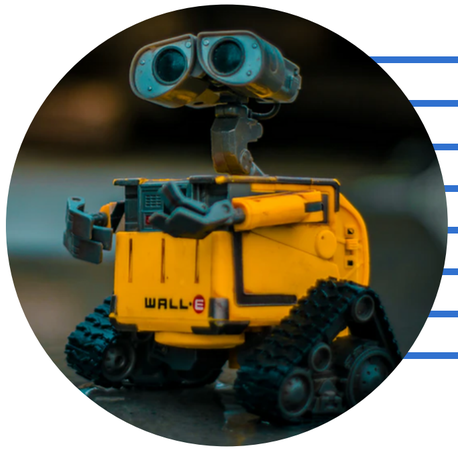
Bad bot

Proxy service

Proxy gateways

Airline booking domain

Proxy services



Bad bot

Proxy service

Proxy gateways

Airline booking domain

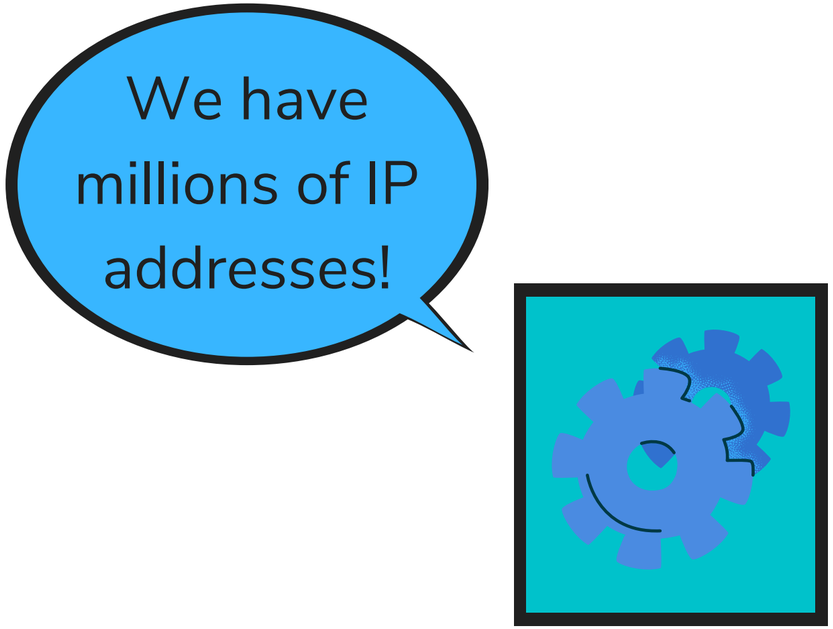
Bots using proxy services

Impossible to link bot campaign to organization

No need for private distributed infrastructure

Impractical blocking IPs strategy

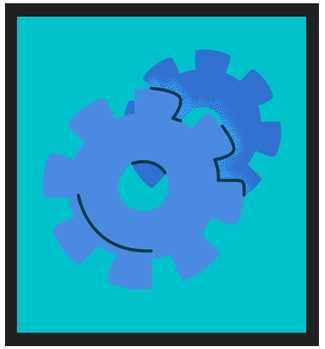
Proxy services claims



Proxy service

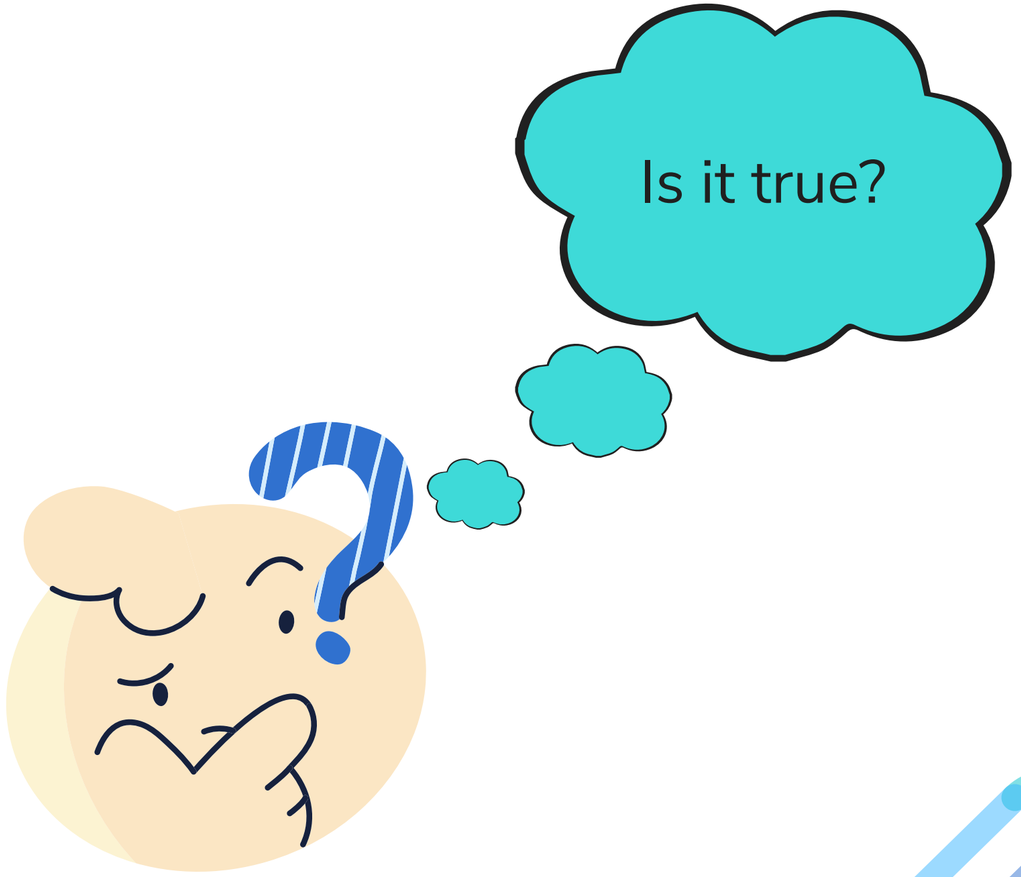
Proxy services claims

We have millions of IP addresses!



Proxy service

Proxy services claims



2. Experimental setup

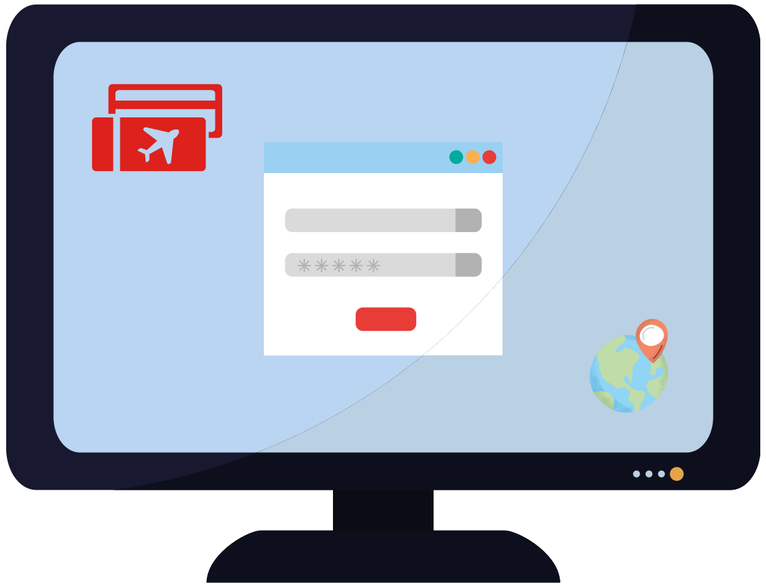
IT provider setup

User

Bot

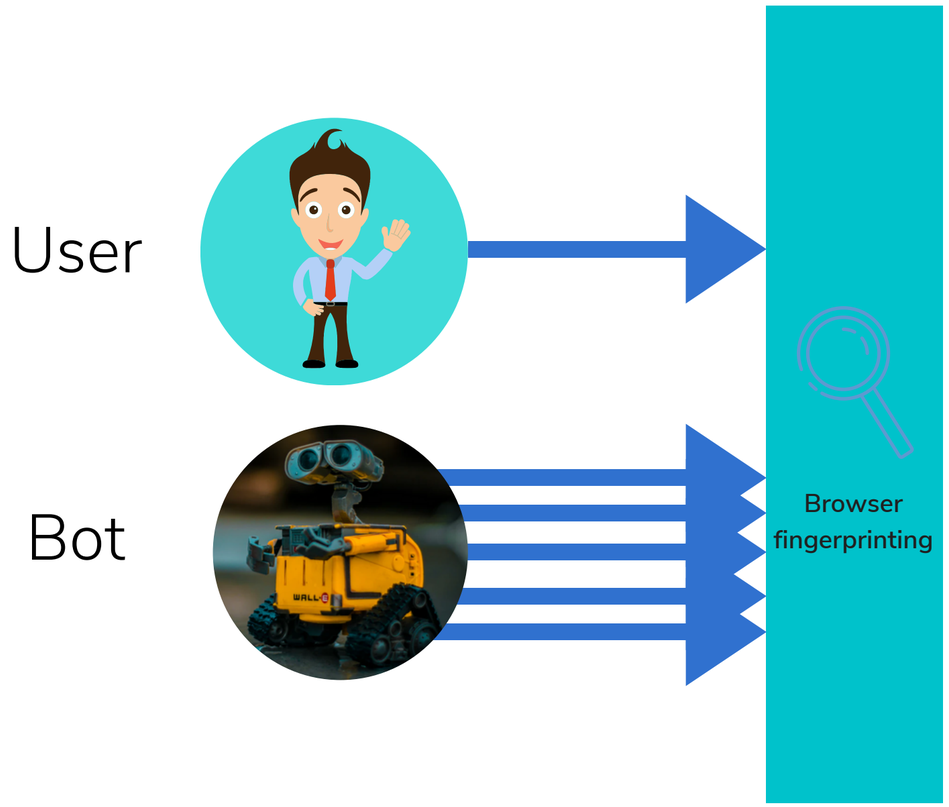
Browser fingerprinting

Anti-bot solution



Airline booking domain

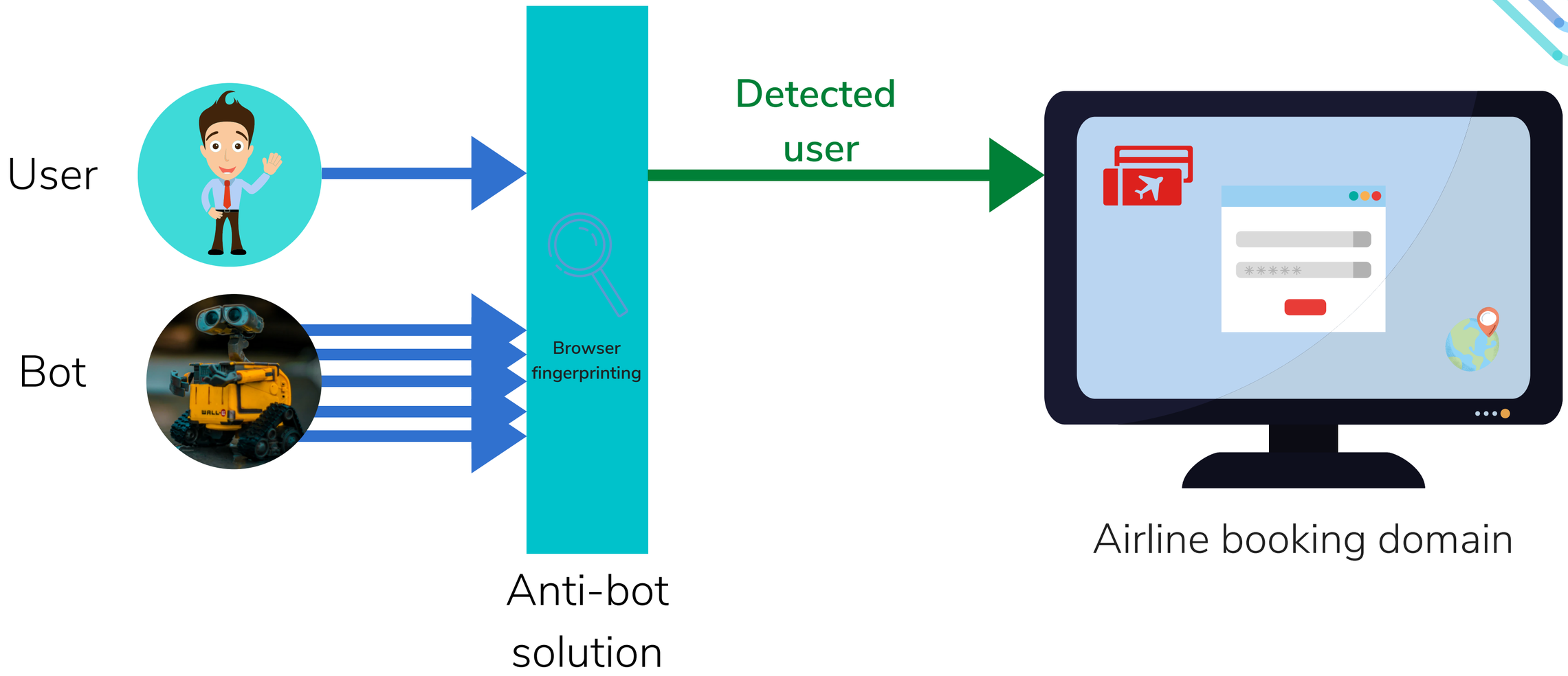
IT provider setup



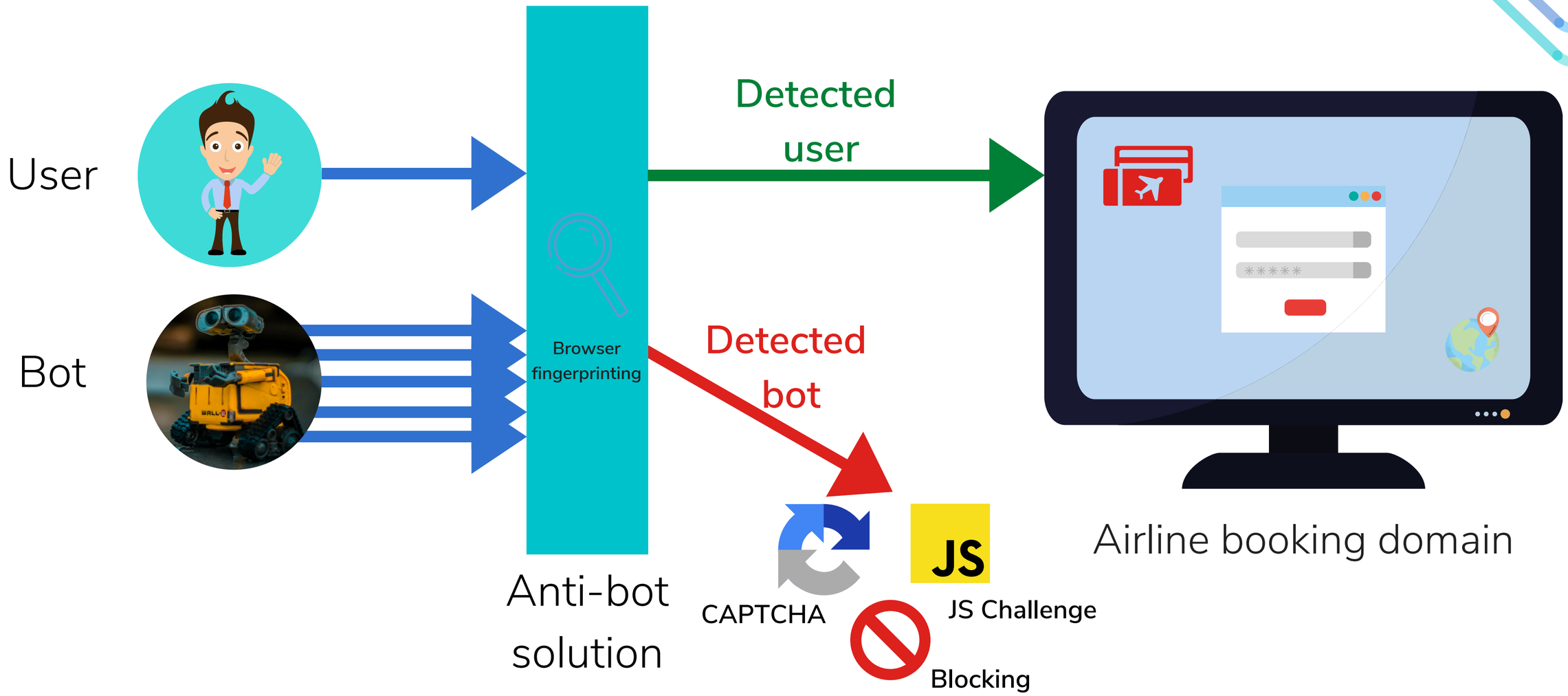
Anti-bot solution

Airline booking domain

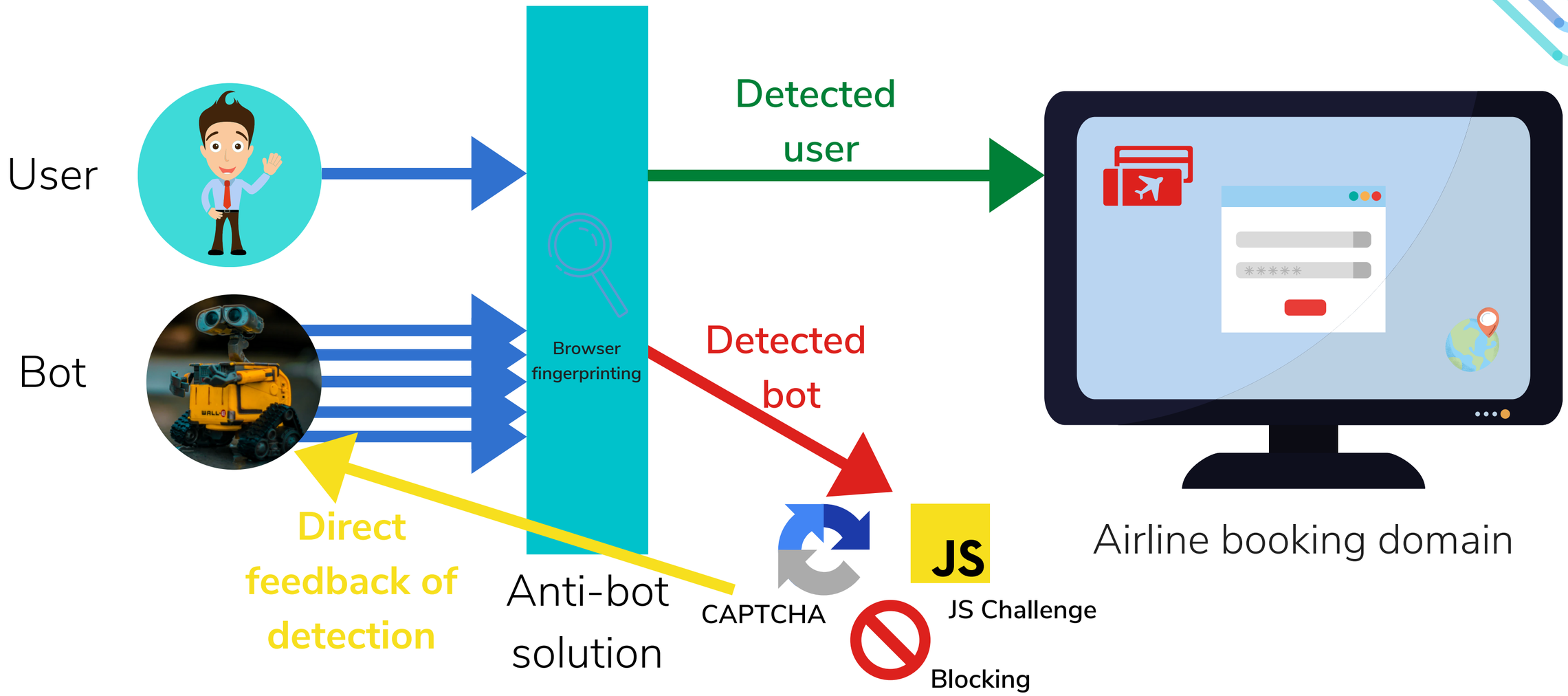
IT provider setup



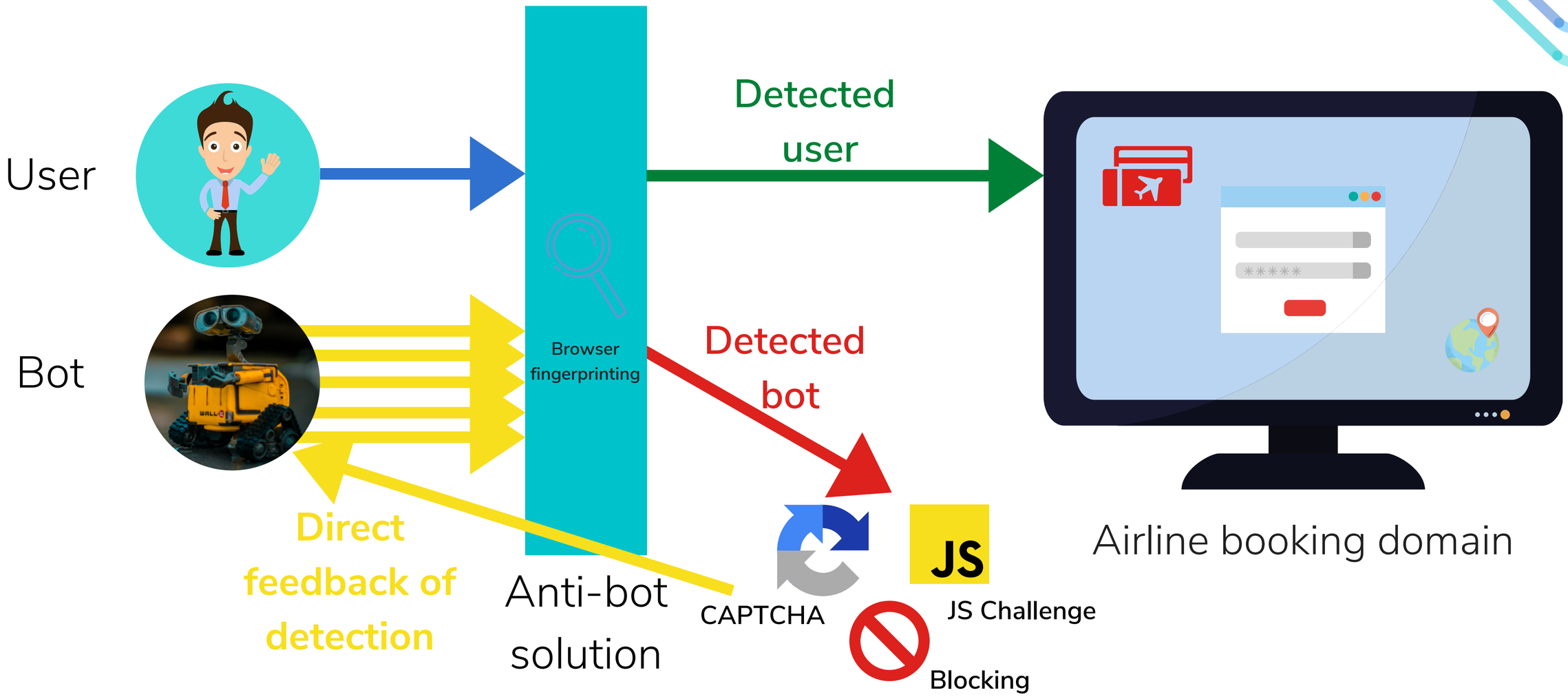
IT provider setup



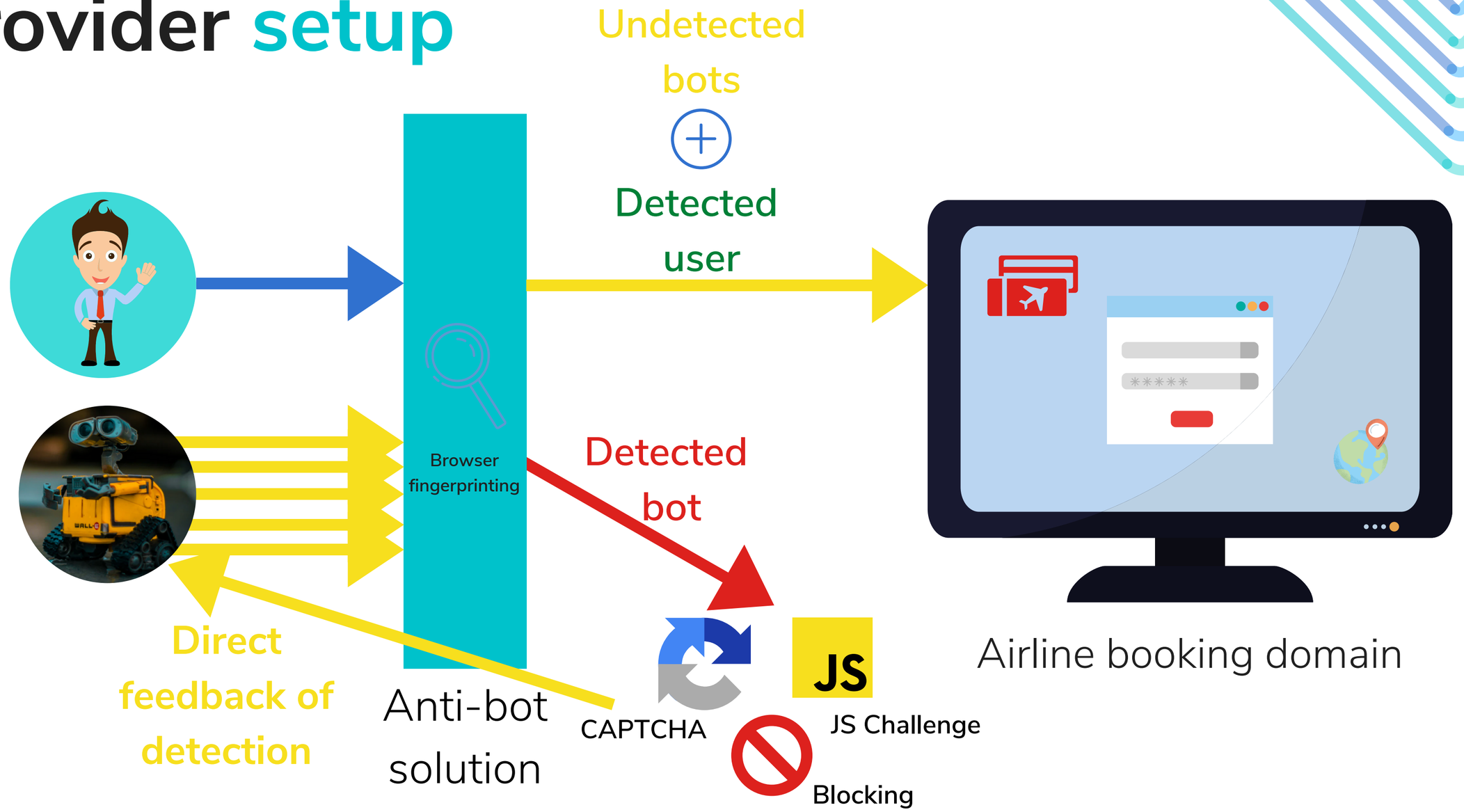
IT provider setup



IT provider setup



IT provider setup



Our idea

User

Bot

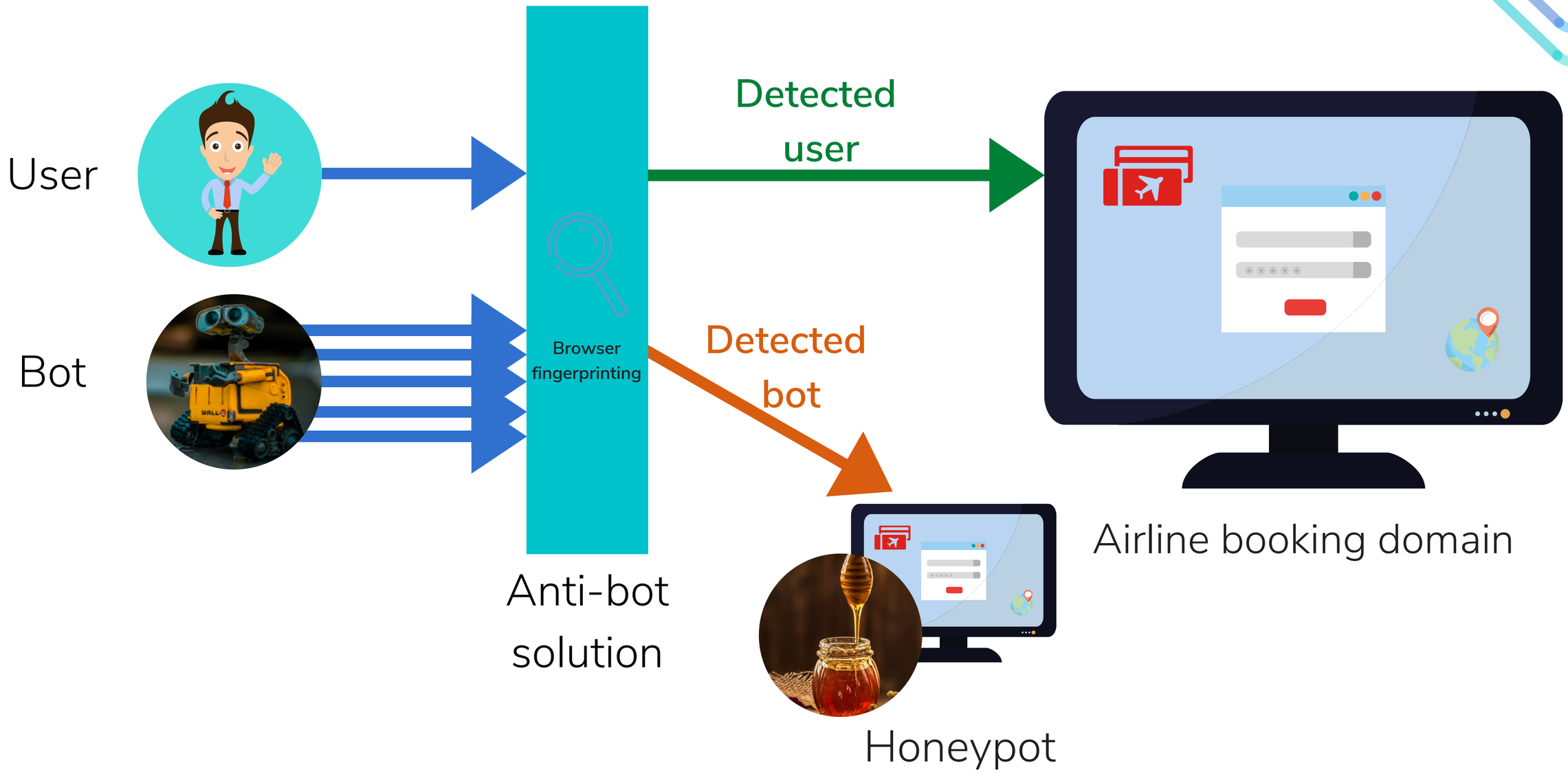
Browser fingerprinting

Anti-bot solution

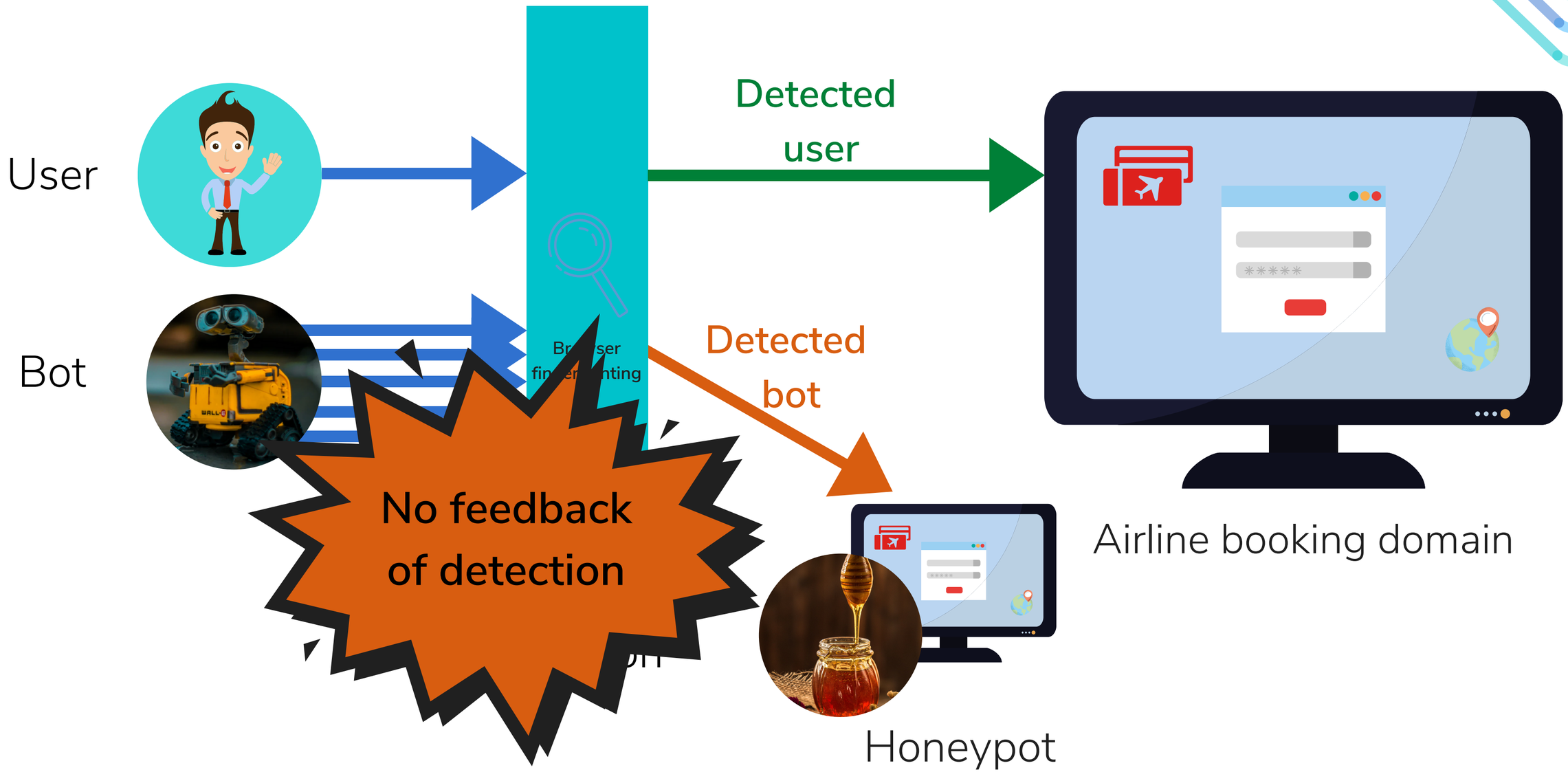
Detected user

Airline booking domain

Our idea


User
Bot
Browser fingerprinting
Detected user
Detected bot
Anti-bot solution
Airline booking domain
Honeypot

Our idea

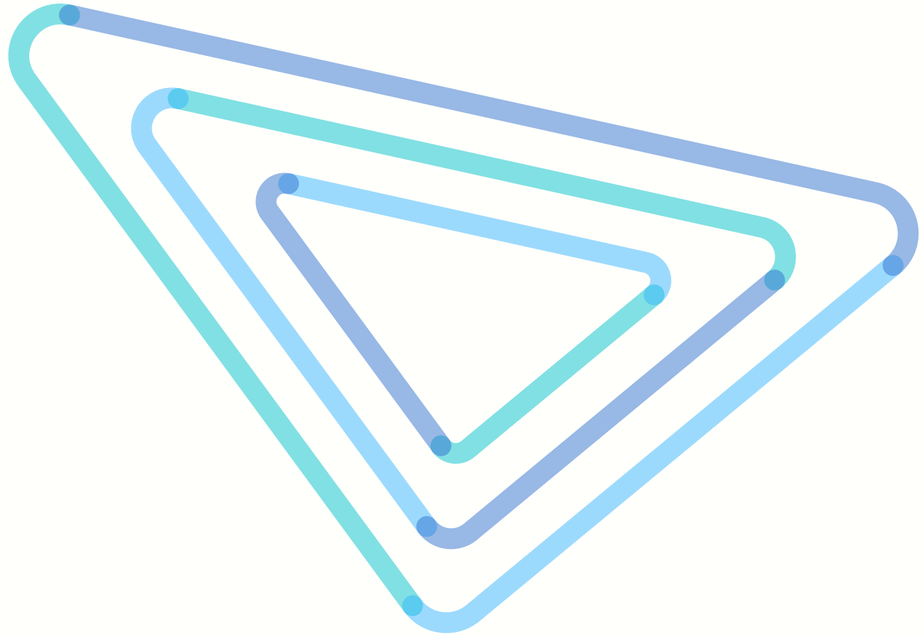

User
Bot
Detected user
Detected bot
No feedback of detection
Airline booking domain
Honeypot

Our experiment

Pilot airline:
1 million daily requests, 40% bots

Chosen bot fingerprint:
40 minutes daily activity from roughly same amount of IPs, no bookings

3. Experimental raw results



First results

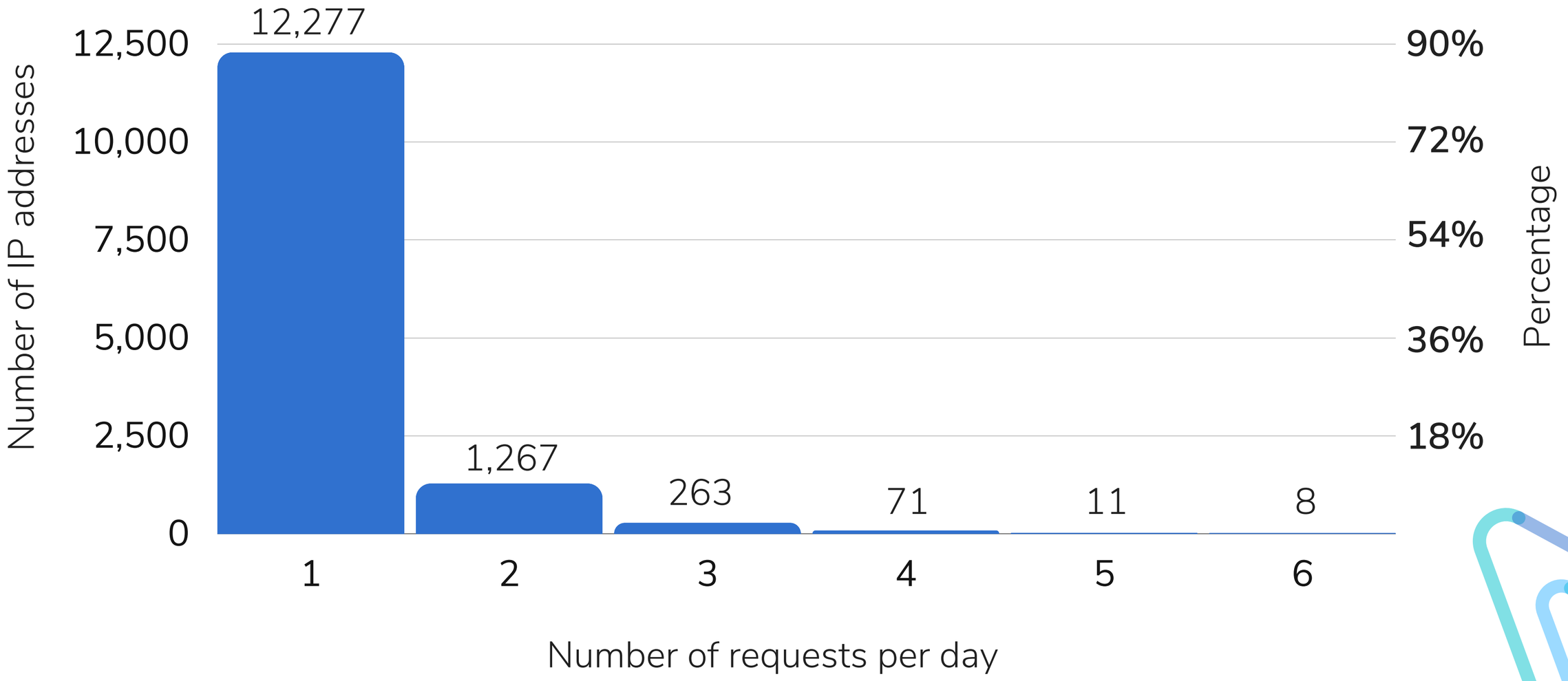
- Experiment running for 56 days
- Interruption linked to COVID-19 restrictions
- 22,991 requests by 13,897 different IP addresses (IPs) received at the Honeypot
- Daily number of requests: 410±33
- Average daily time window: 38.18 minutes

First results

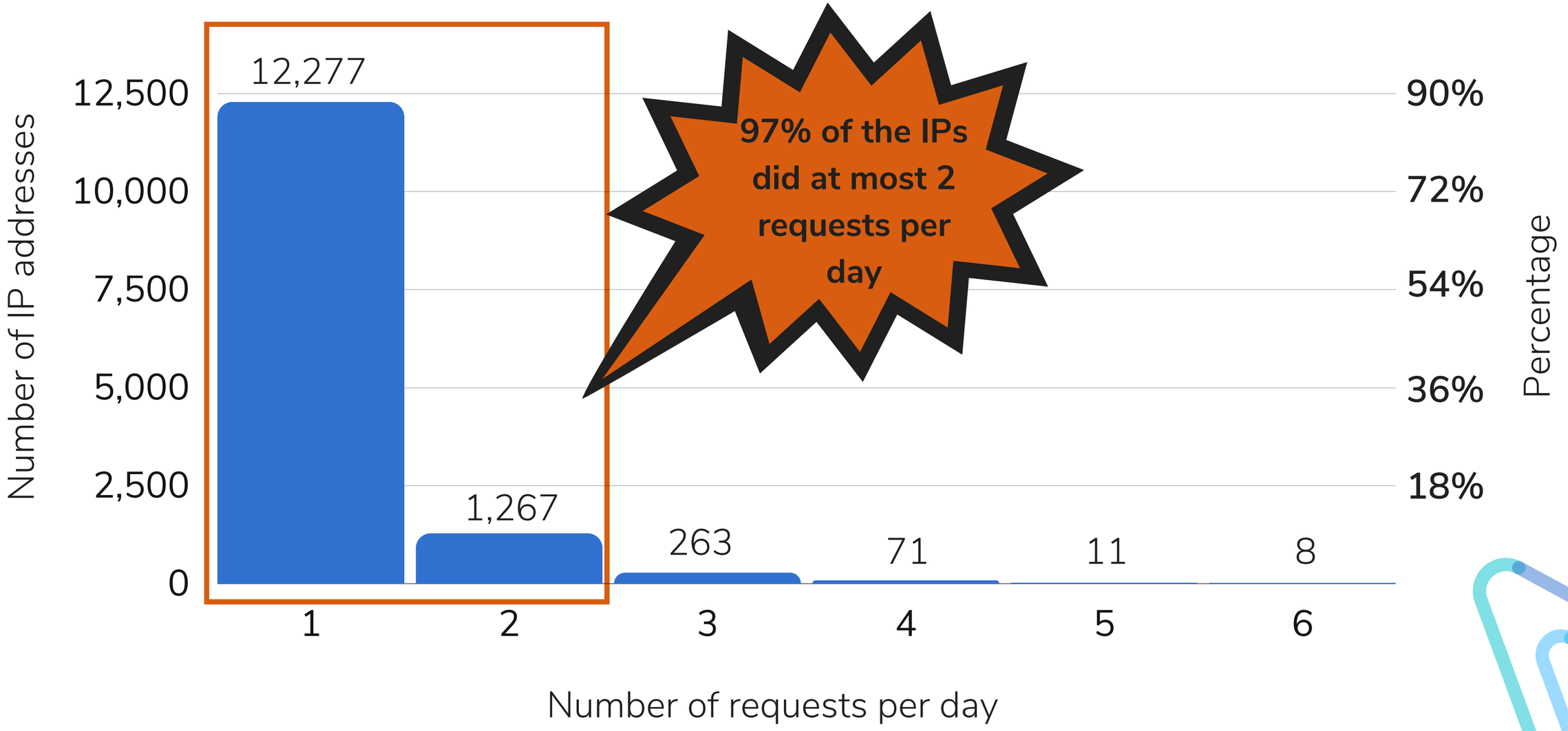
- Experiment running for 56 days
- Interruption linked to COVID-19 restrictions
- 22,991 requests by 13,897 different IP addresses (IPs) received at the Honeypot
- Daily number of requests: 410±33
- Average daily time window: 38.18 minutes

In line with the values before the experiment

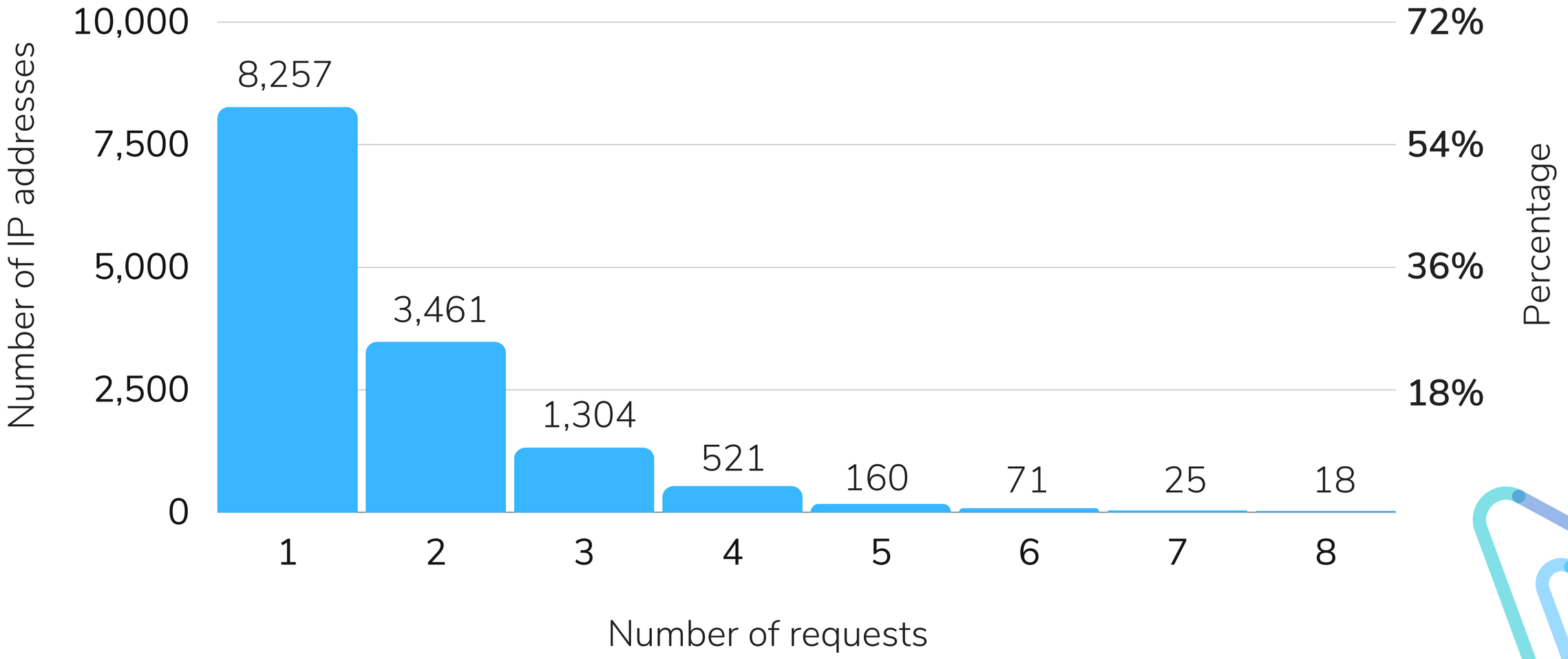
Daily number of requests per IP



Daily number of requests per IP

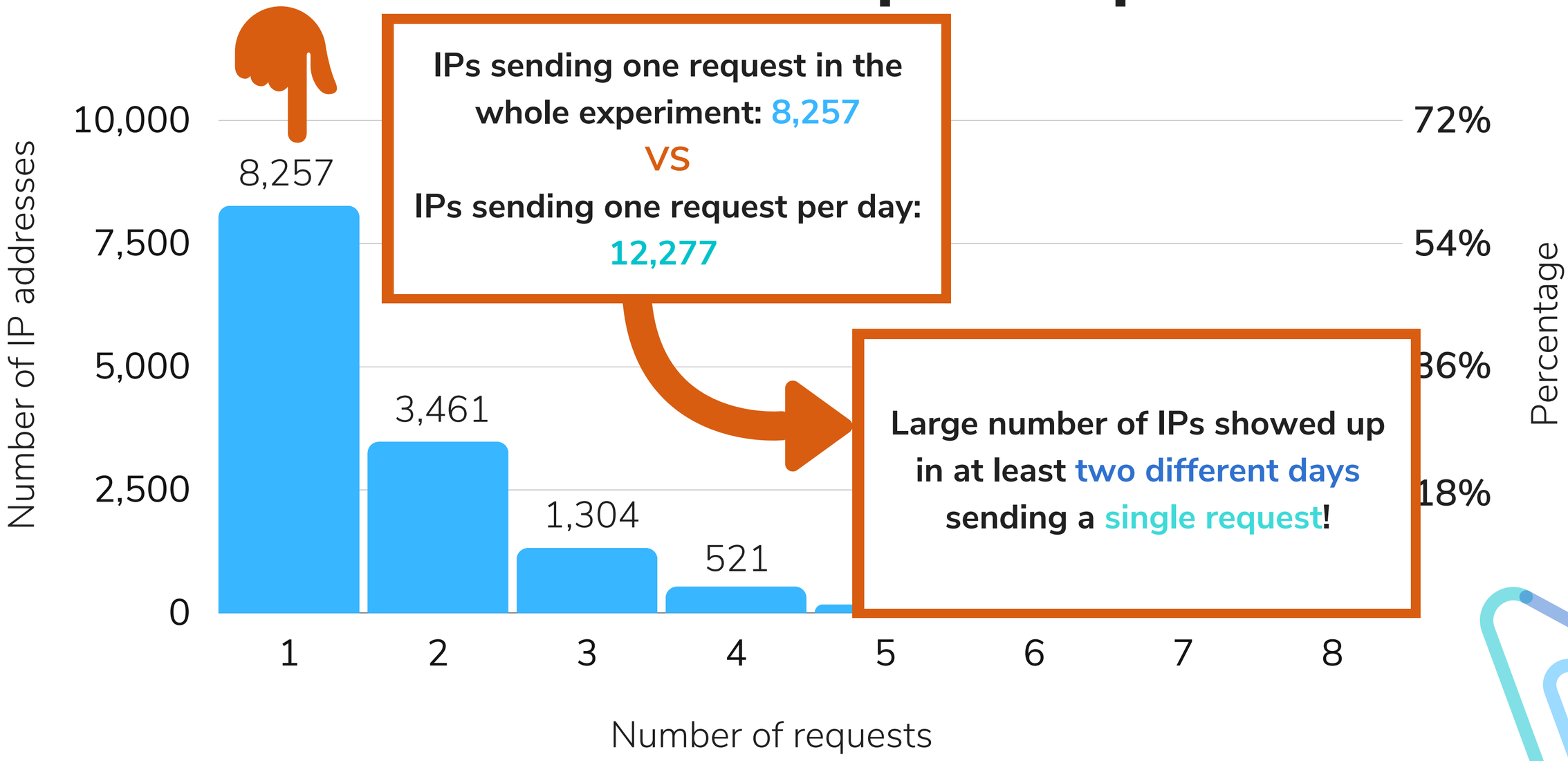


Maximum number of requests per IP

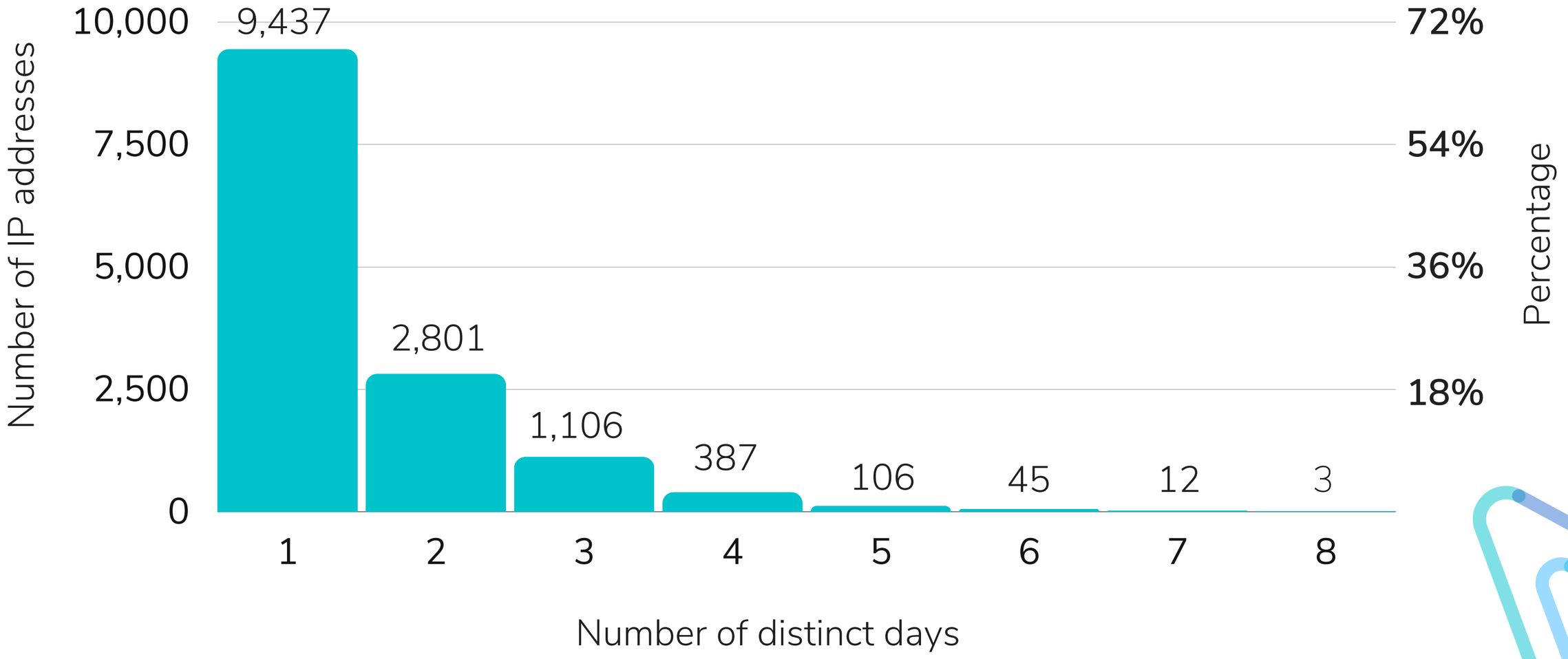


Maximum number of requests per IP

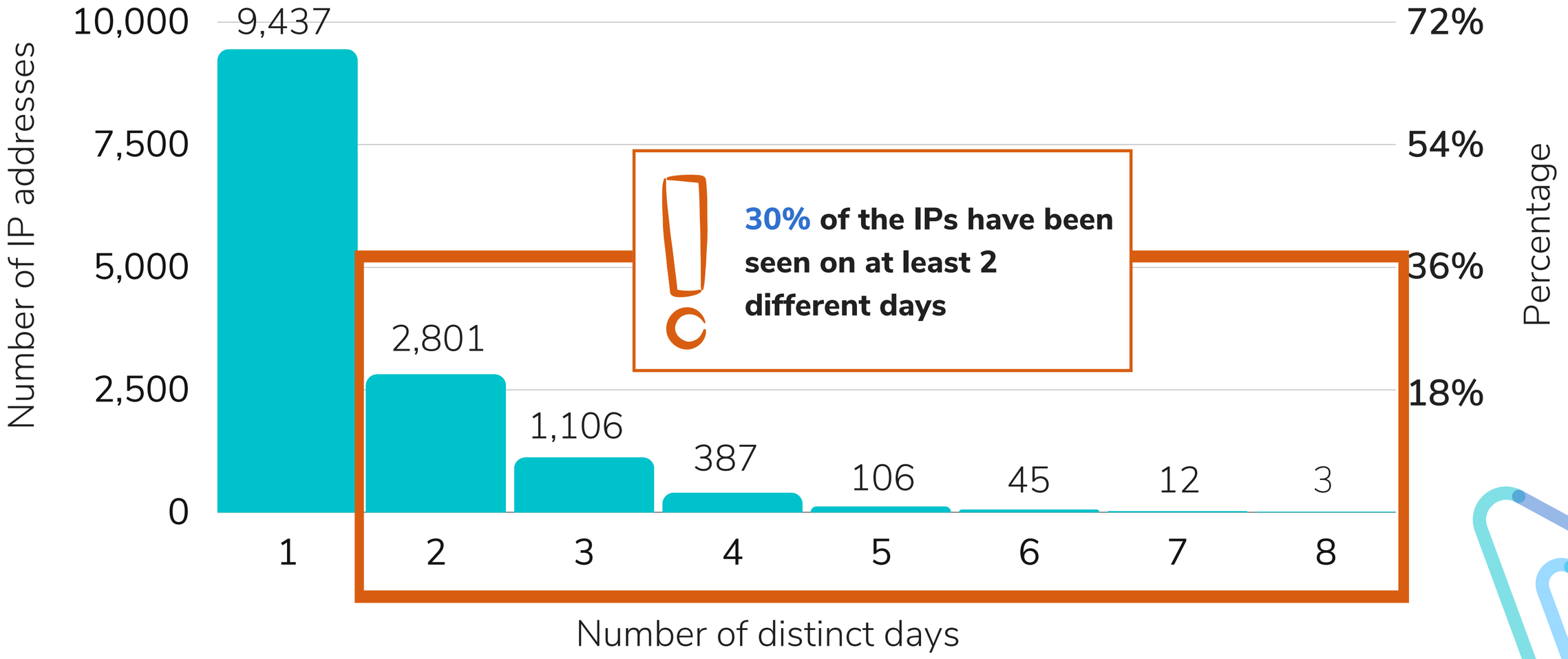
Maximum number of requests per IP



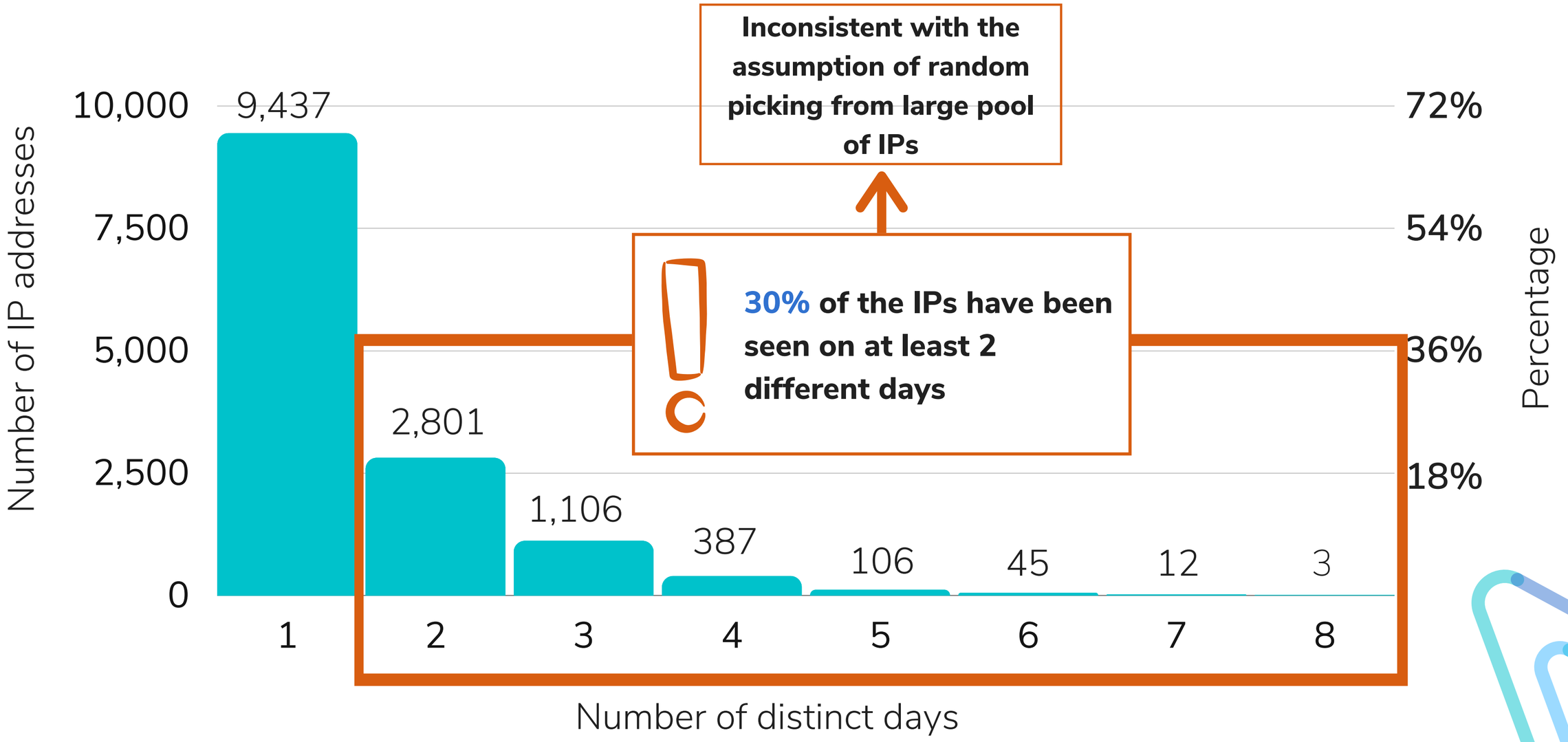
Number of IPs seen in distinct days



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 → p(56,10M)≈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 → p(56,10M)≈0.000154
- P=1,000,000 → p(56,1M)≈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 → p(56,10M)≈0.000154
- P=1,000,000 → p(56,1M)≈0.001538
- P=100,000 → p(56,100K)≈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)≈0.000154

P=1,000,000 → p(56,1M)≈0.001538

P=100,000 → p(56,100K)≈0.015282

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

Moreover...

Average number of distict IPs per day: 371

VS

Average number of requests per day: 410

Moreover...

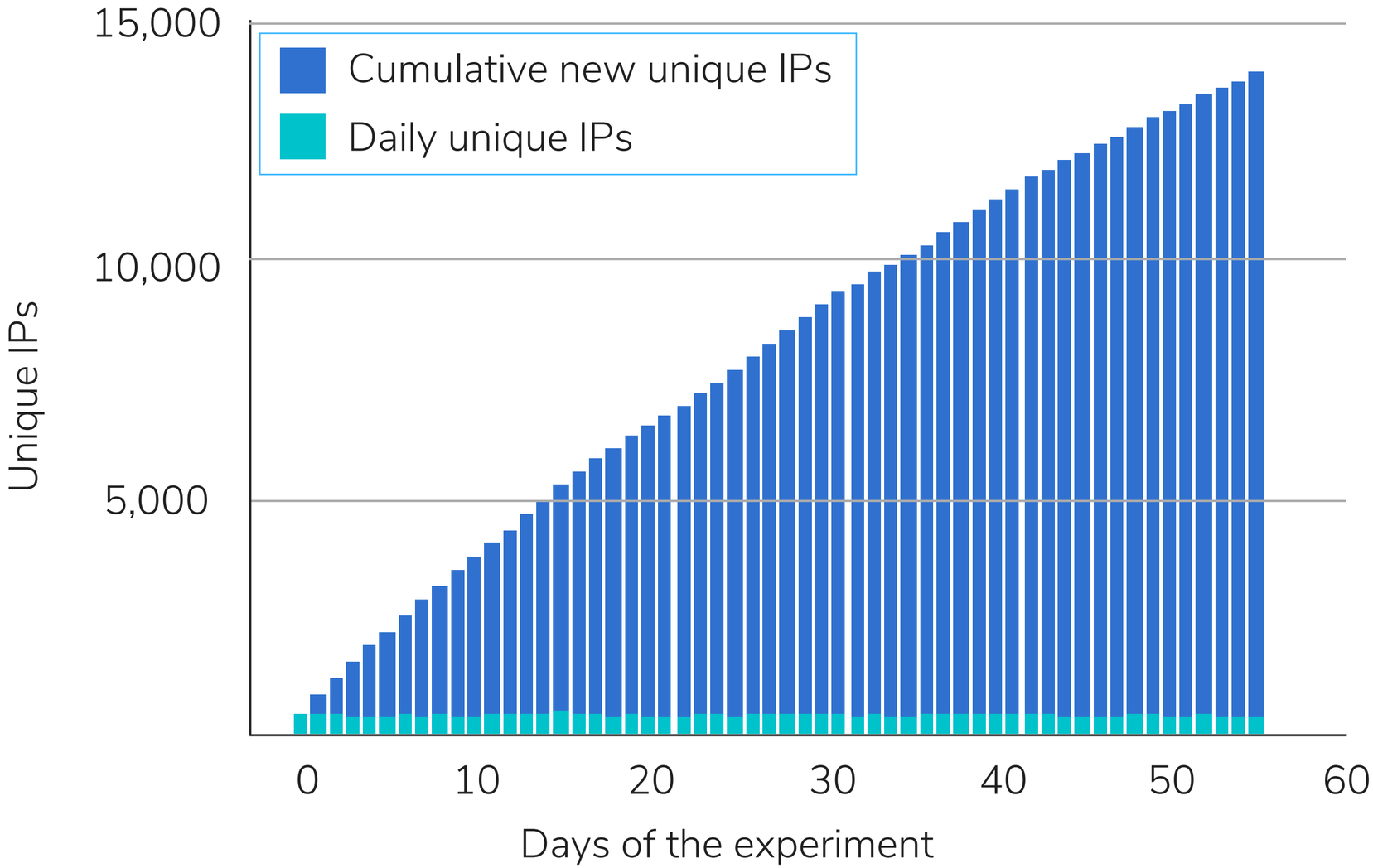
Average number of distict IPs per day: 371

VS

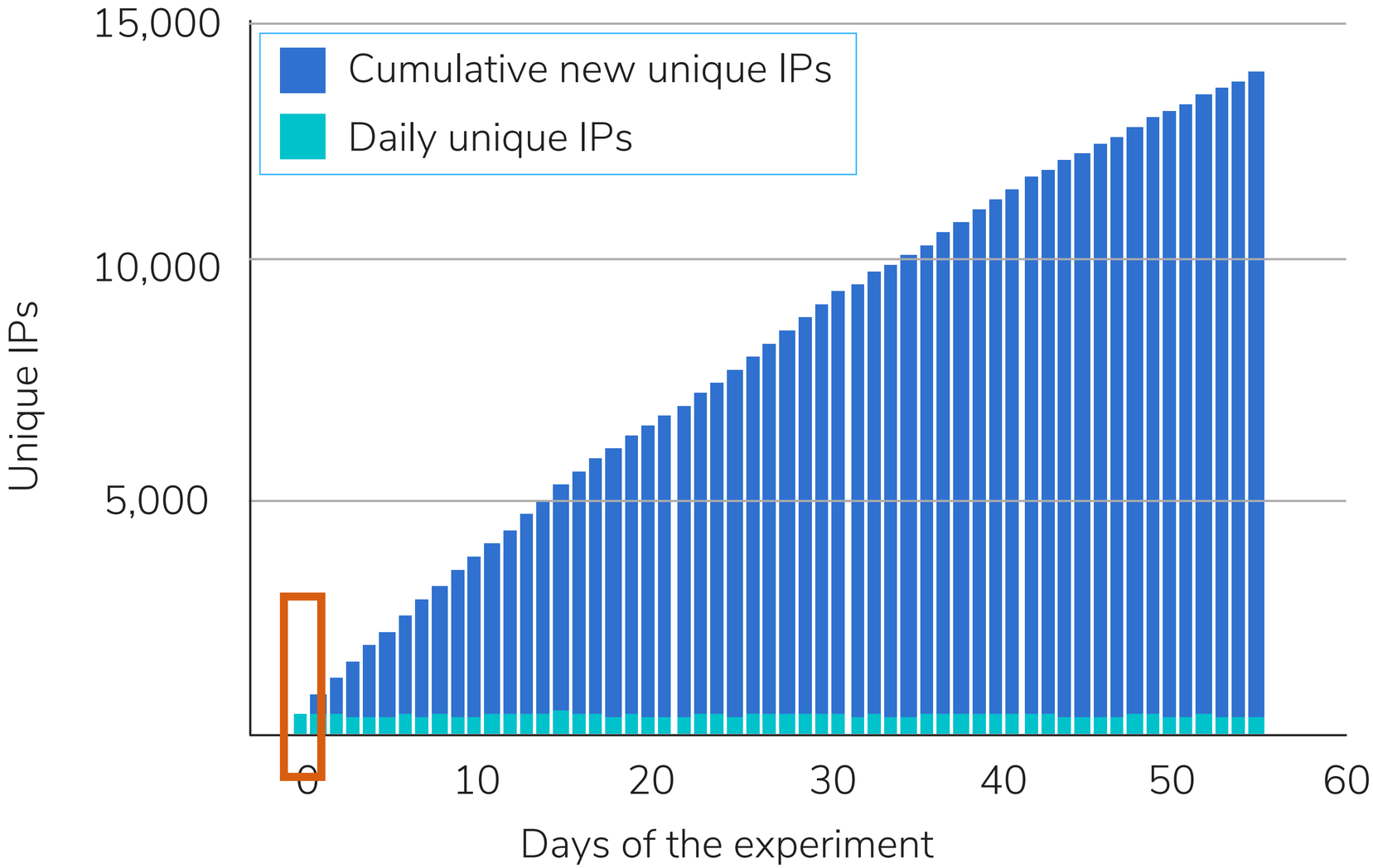
Average number of requests per day: 410

Most IPs send a single request and
reappear some time later

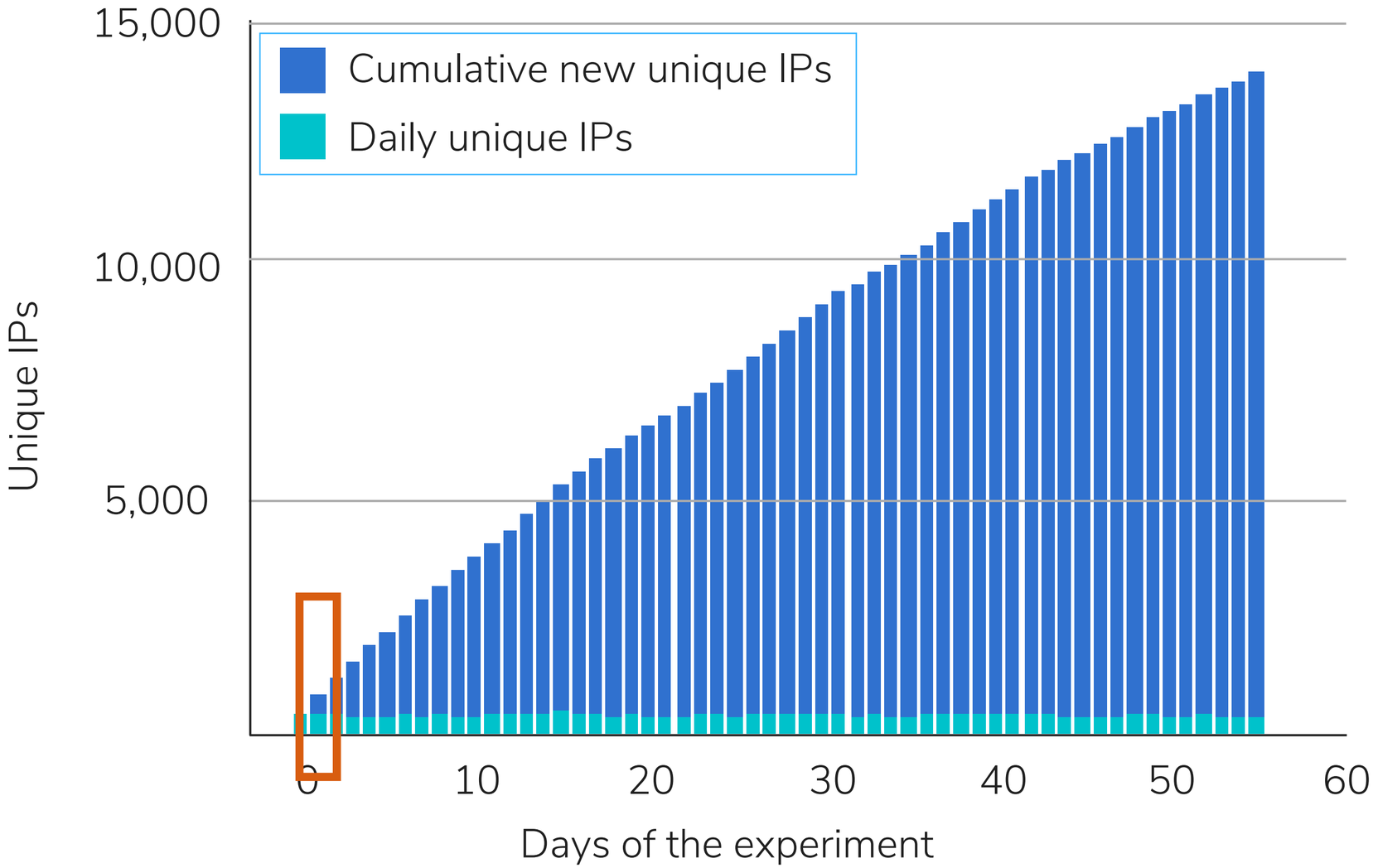
Cumulative curve of new unique IPs



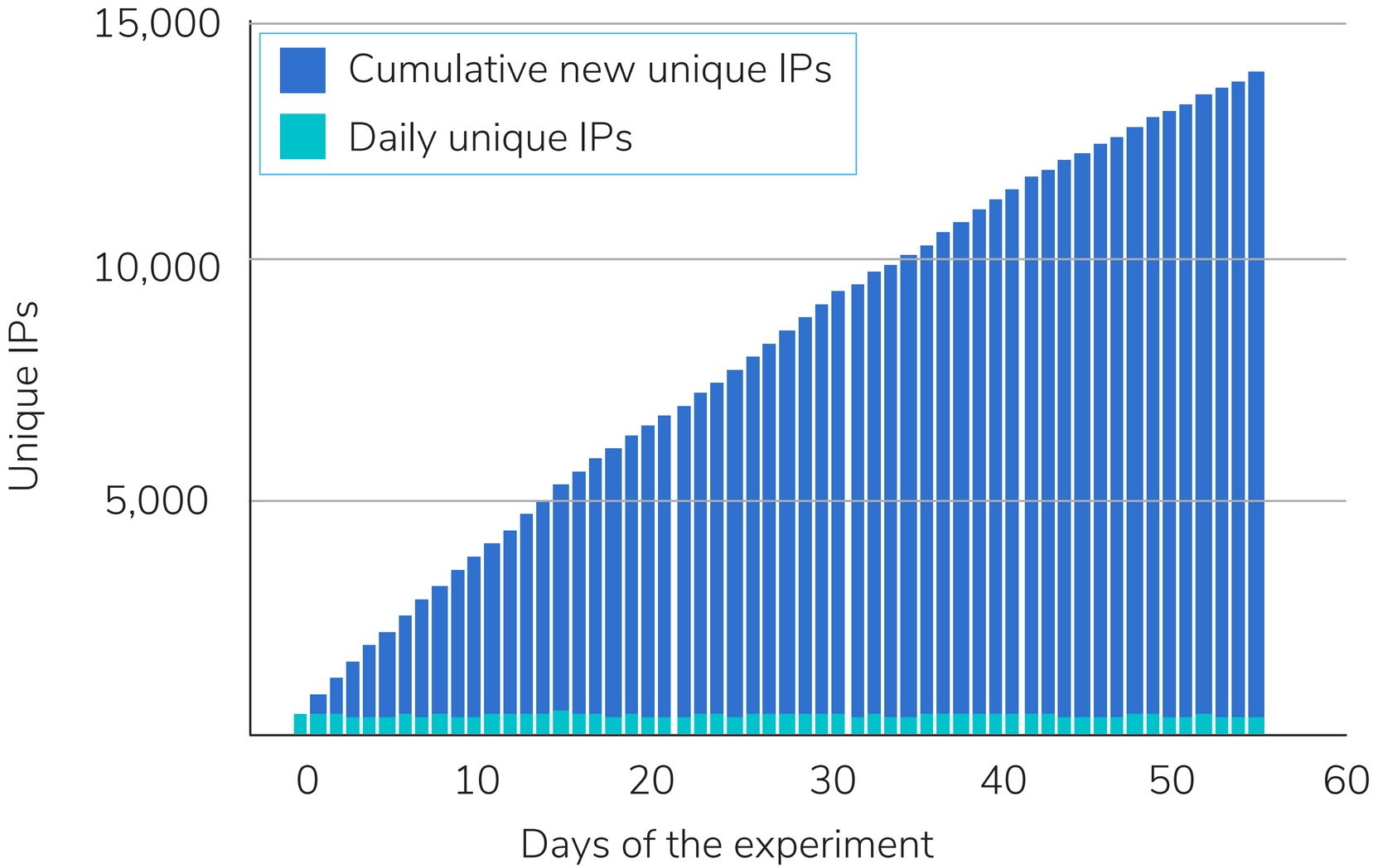
Cumulative curve of new unique IPs



Cumulative curve of new unique IPs

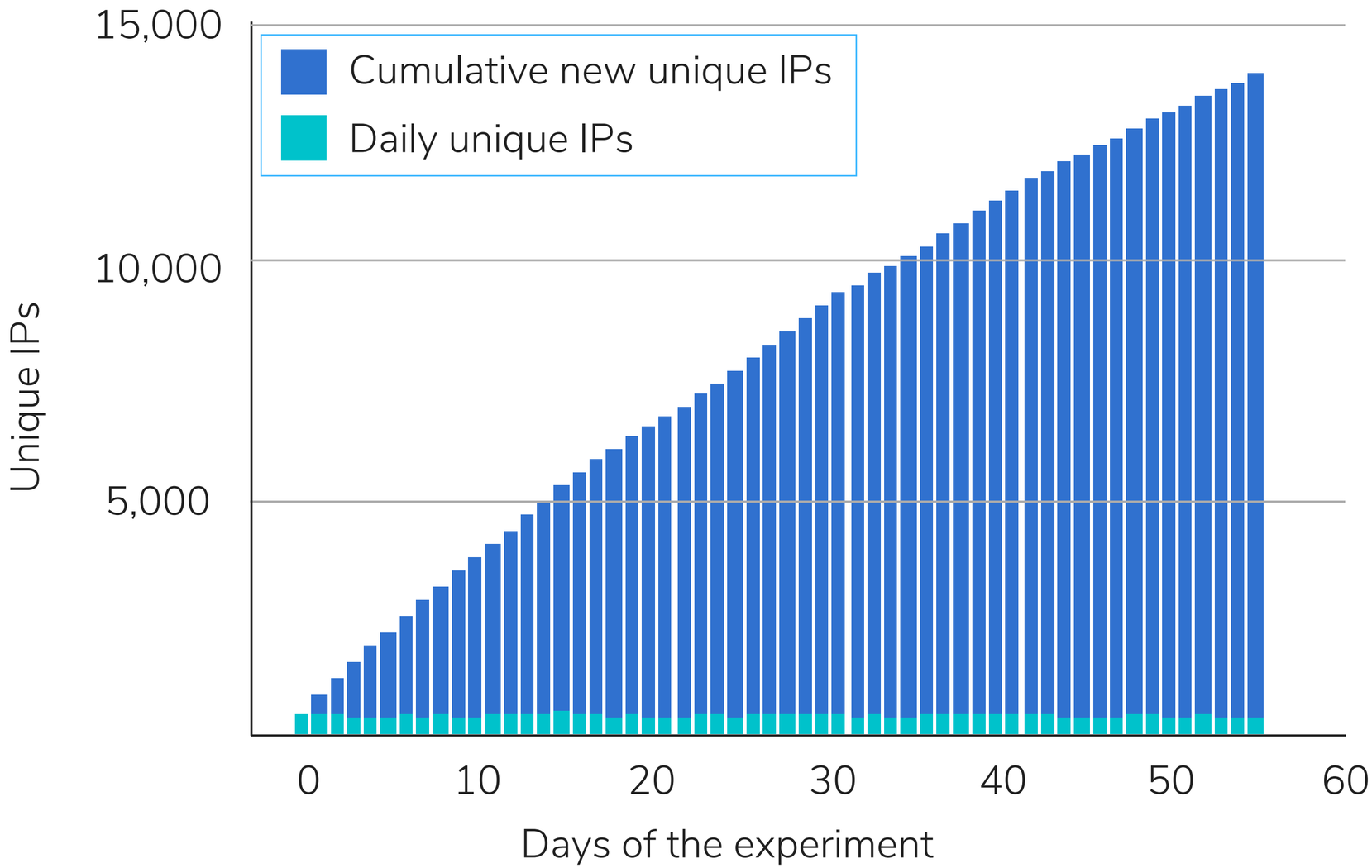


Cumulative curve of new unique IPs



The daily increment decreases over time

Cumulative curve of new unique IPs



The daily increment decreases over time

Eventually it will reach a maximum!

IPs analysis in other airlines

- Analyses of the traffic of other 17 airlines

IPs analysis in other airlines

Analyses of the traffic of other 17 airlines

- Only 5 bookings during the running time of the experiment
- Dates different from the ones in which the IP was seen in the honeypot
- Request 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 experiment
- Dates different from the ones in which the IP was seen in the honeypot
- Request 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 blacklist

IPs reputation

IPQualityScore analysis

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

DNS blocklists

- 76% blocked in at least one blacklist

These IPs were doing malicious activities also outside our scope

IPs reputation

IPQualityScore analysis

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

DNS blocklists

- 76% blocked in at least one blacklist

These IPs were doing malicious activities also outside our scope

4. Modeling and discussion

Two approaches

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.

Two approaches

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 $\mathscr{P}$

IP assignment as a drawing process

- Model the assignment process by a daily probabilistic drawing process without replacement
- Arbitrarily define a pool size $\mathscr{P}$
- 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 $\mathscr{P}$
- 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

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”

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 $\mathscr{P}$ 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 w from 2 to 10 days
- Drawing with replacement
- Additional constraint: a given value cannot be drawn more than w times, i.e. once per day

Disclaimers

Our goal is not to find the best probability distribution function but to show that several "good enough" ones deliver the same ballpark figure for $\mathscr{P}$

Disclaimers

Our goal is not to find the best probability distribution function but to show that several "good enough" ones deliver the same ballpark figure for $\mathscr{P}$

Values of $\mathscr{P}$:

- From 10,000 to 100,000, step equal to 10,000
- From 100,000 to 200,000, step equal to 20,000

Disclaimers

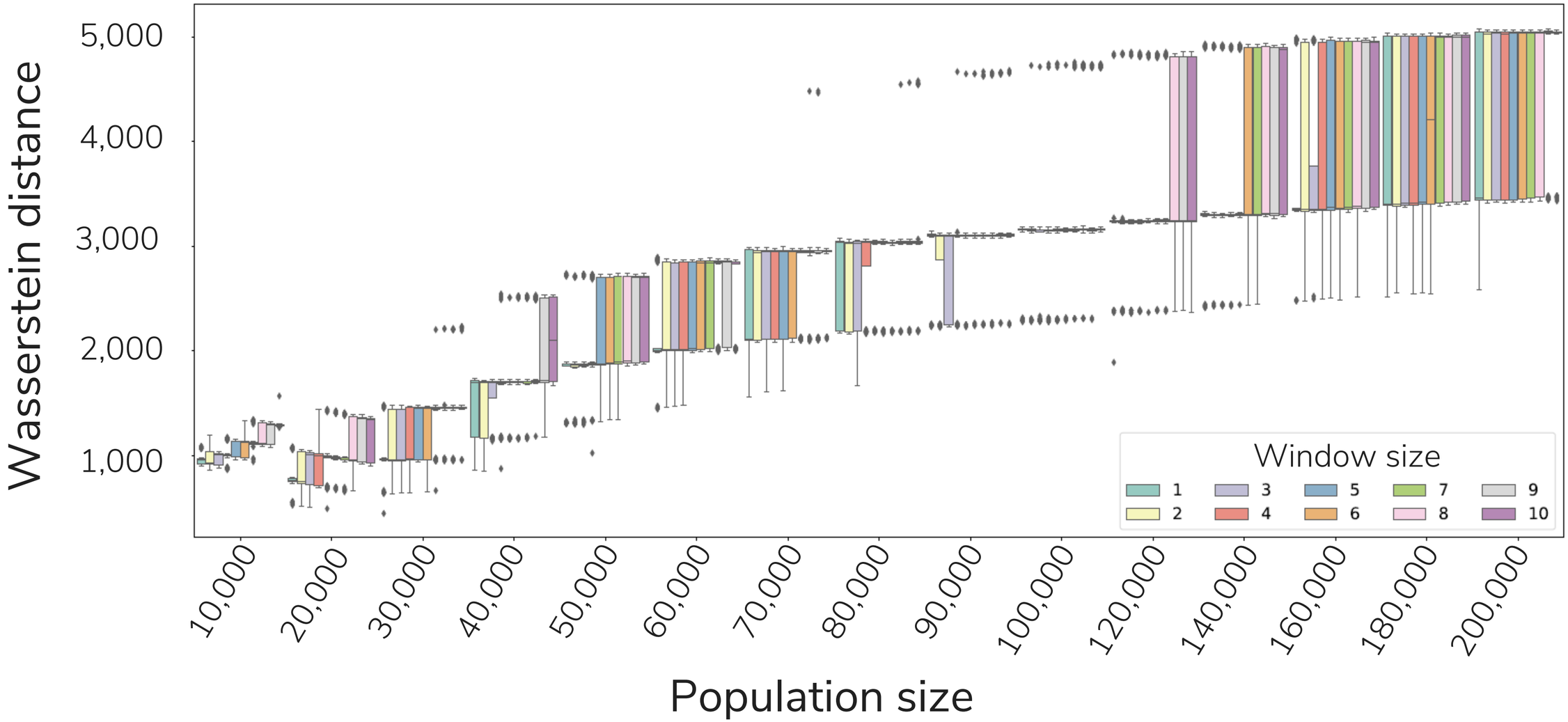
Our goal is not to find the best probability distribution function but to show that several "good enough" ones deliver the same ballpark figure for $\mathscr{P}$

Values of $\mathscr{P}$:

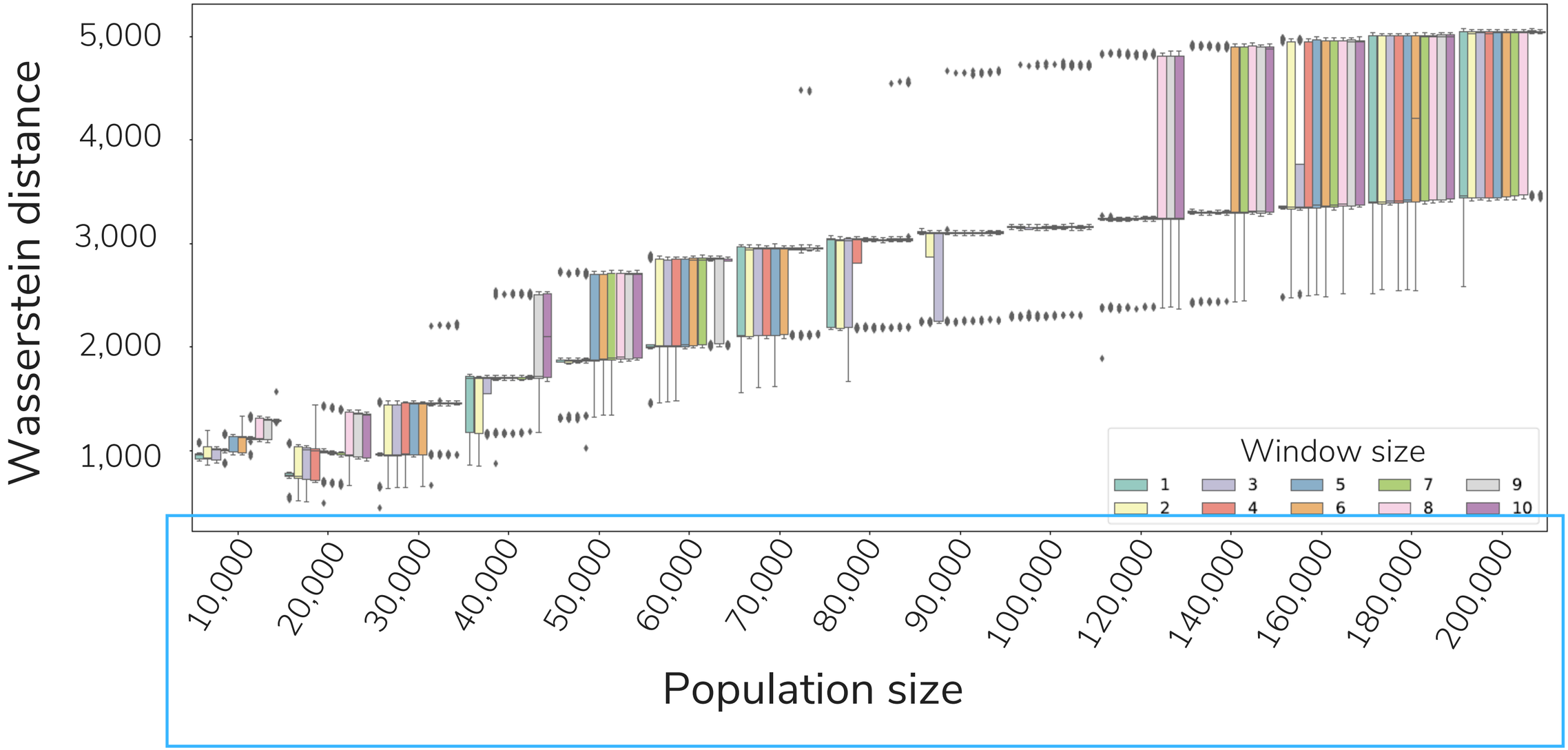
- From 10,000 to 100,000, step equal to 10,000
- From 100,000 to 200,000, step equal to 20,000

For each value of $\mathscr{P}$, the considered Wasserstein distance is the average of 100 simulations

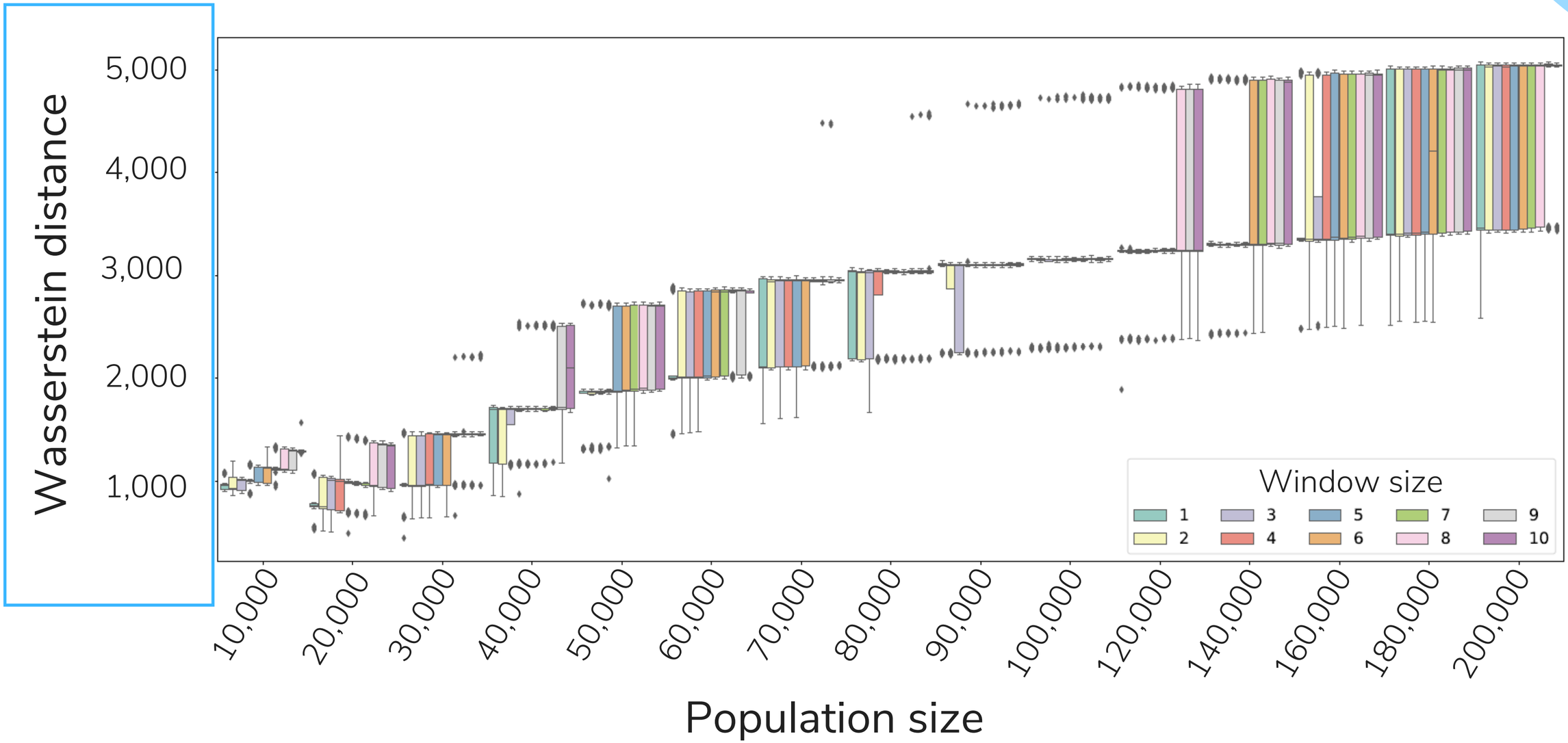
Uniform distribution



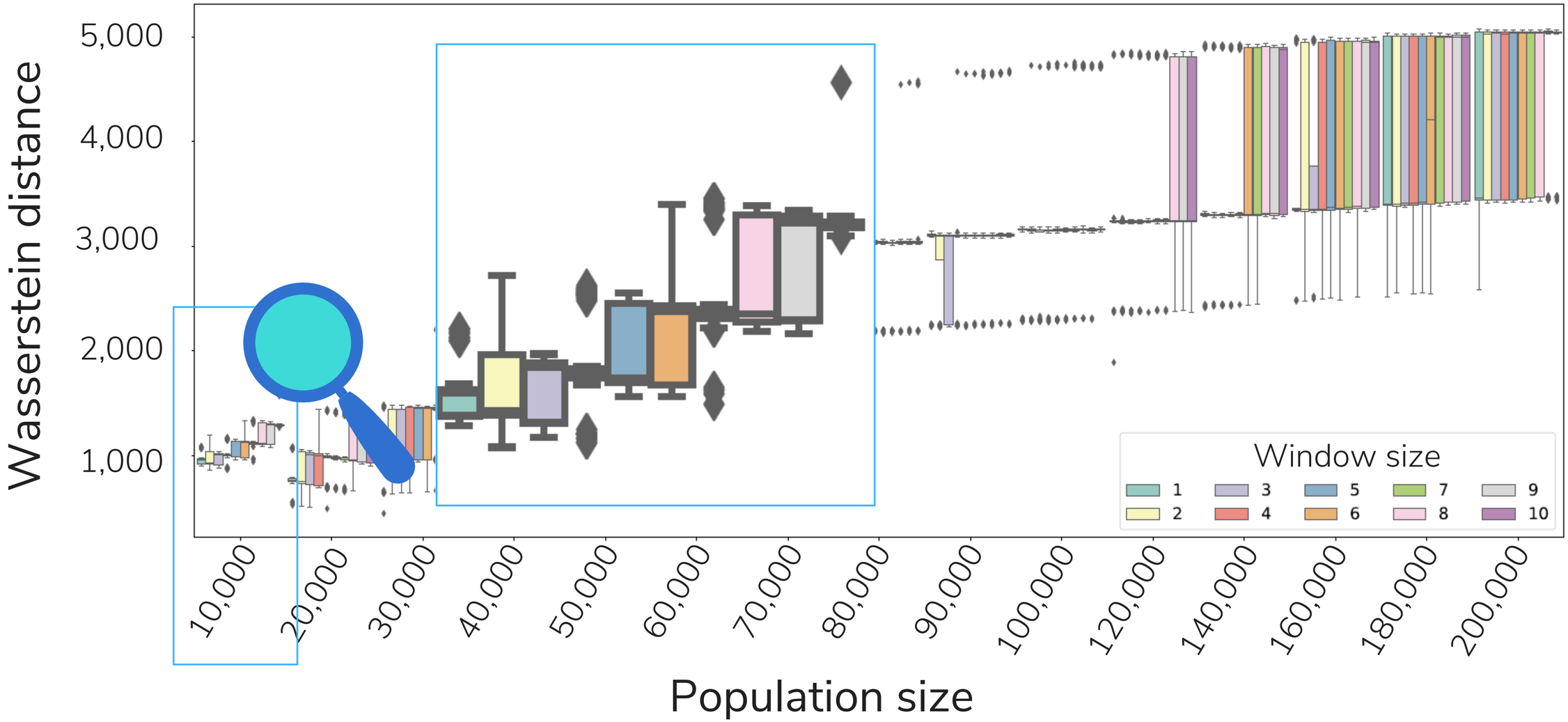
Uniform distribution



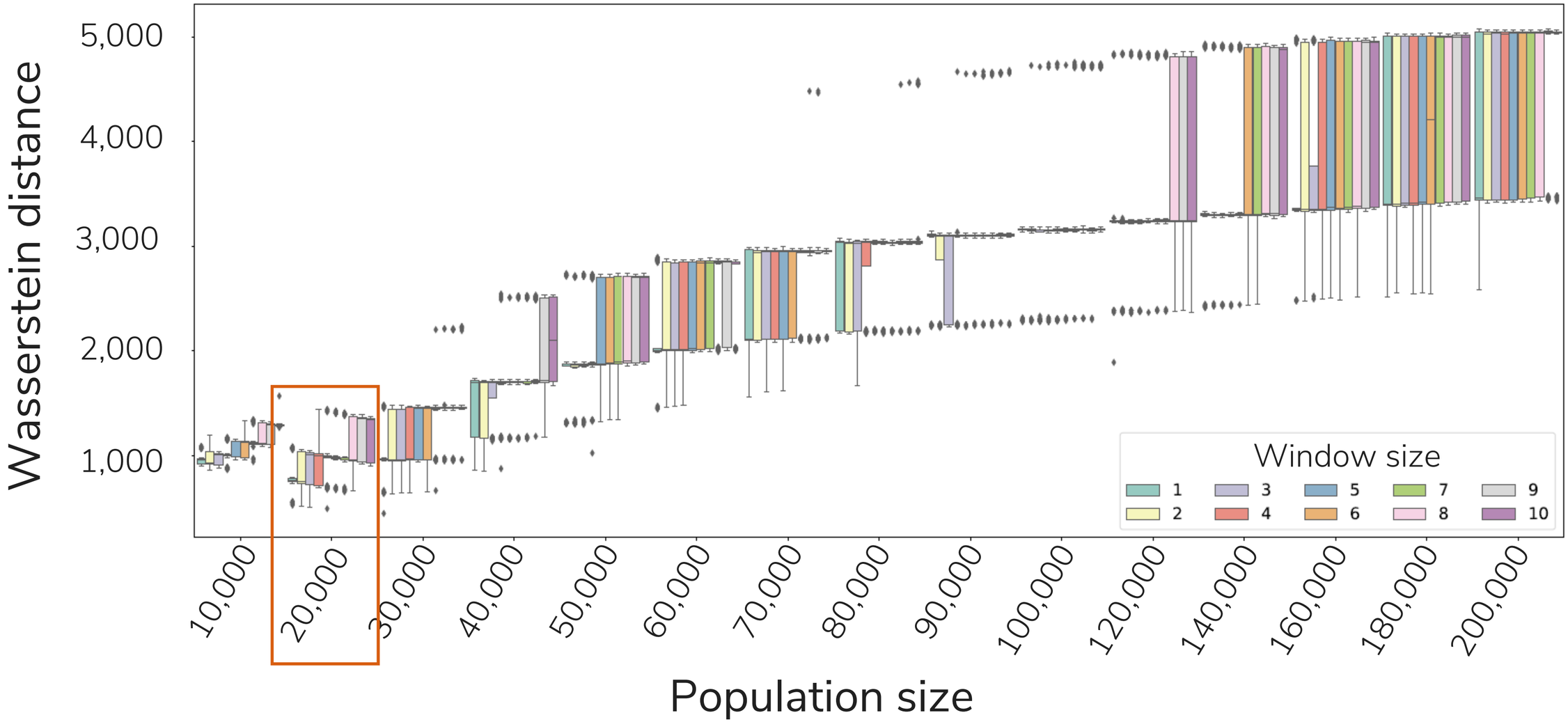
Uniform distribution



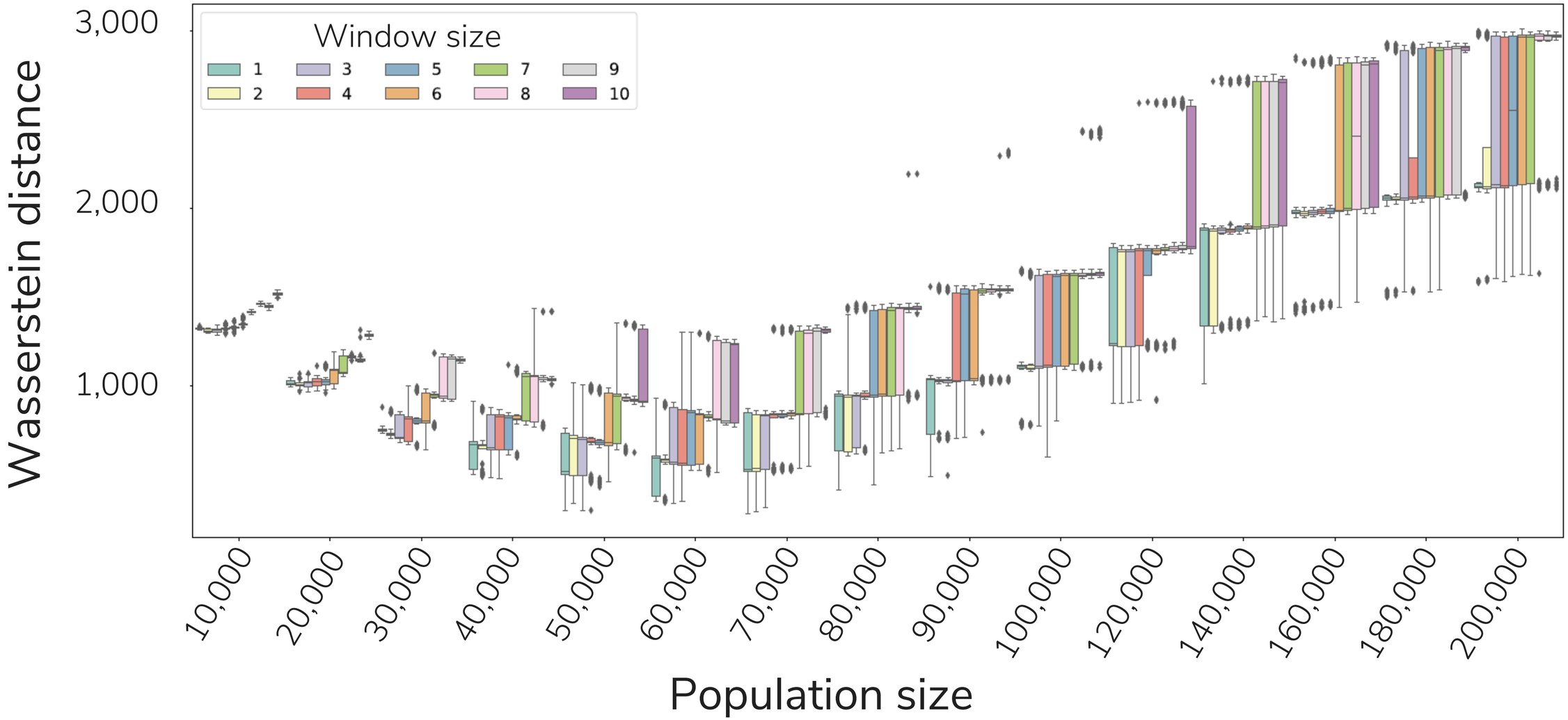
Uniform distribution



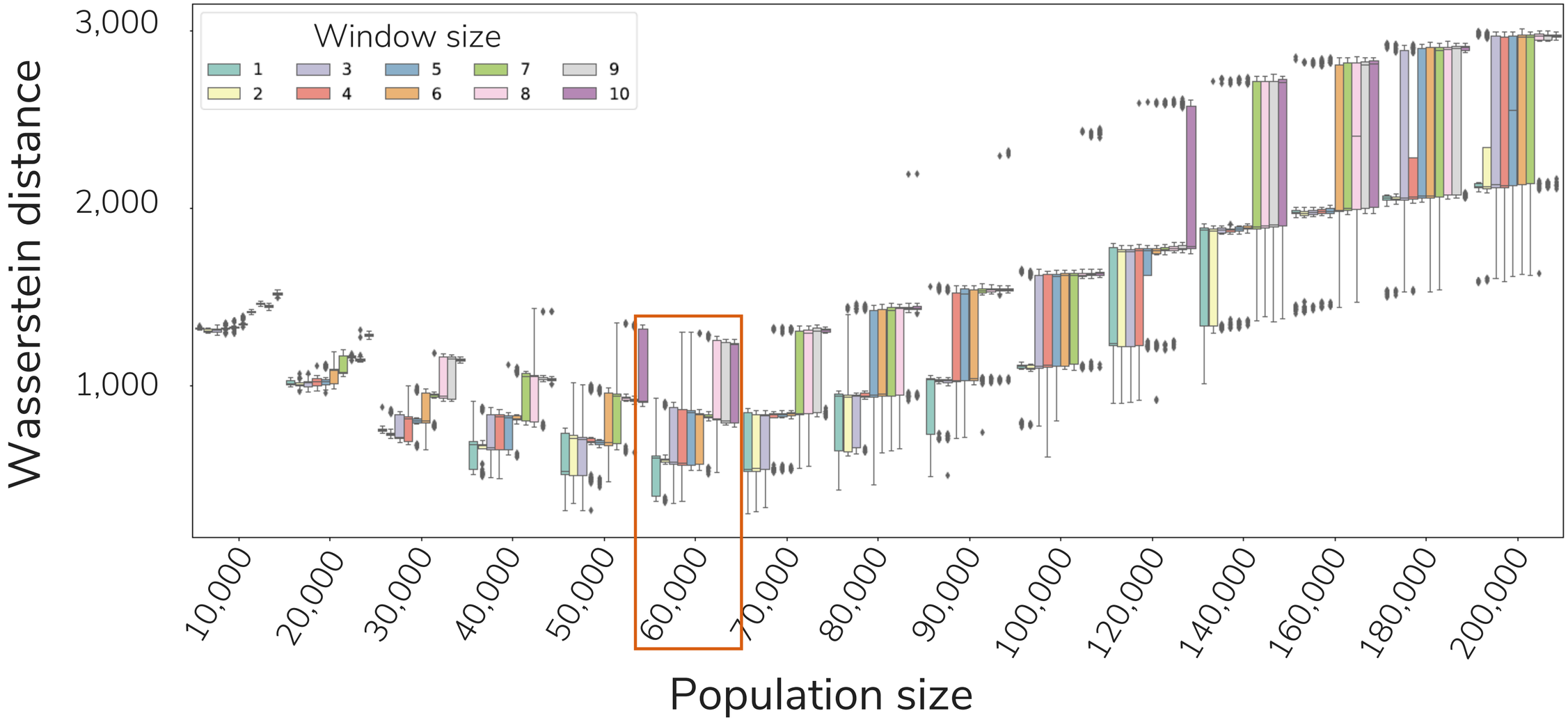
Uniform distribution



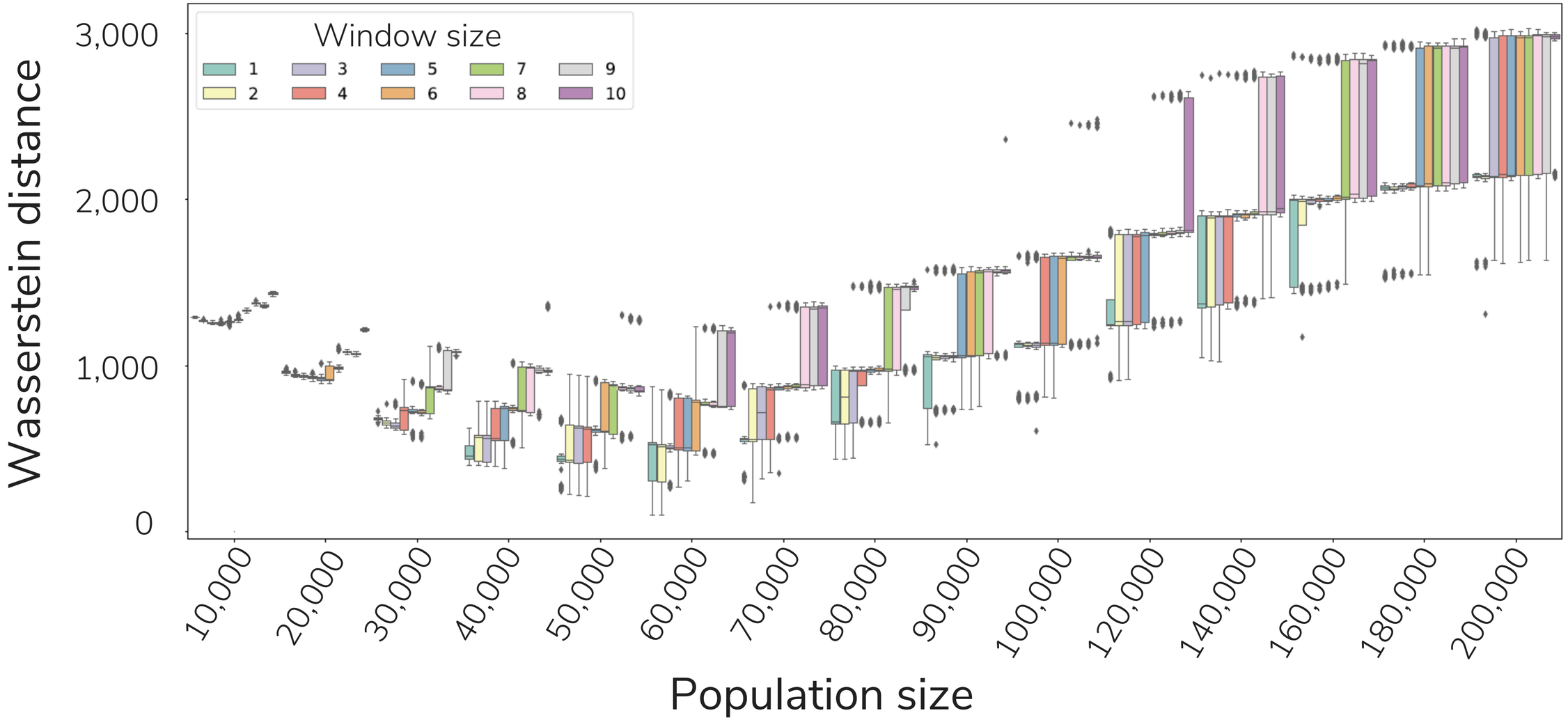
Gaussian distribution



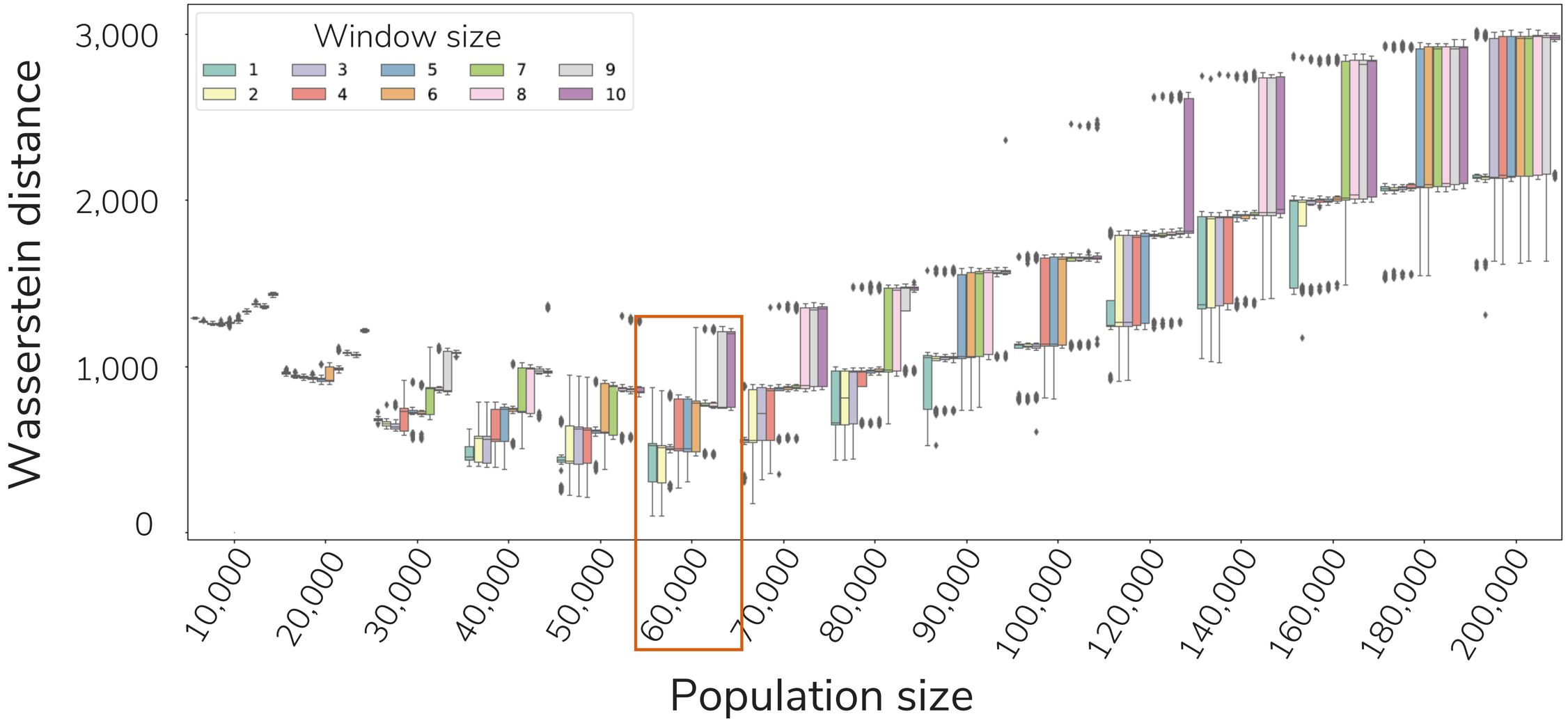
Gaussian distribution



Beta distribution



Beta distribution



What does it mean?

Bots of our experiment are
single-request bots

What does it mean?

Bots of our experiment are
single-request bots

BUT

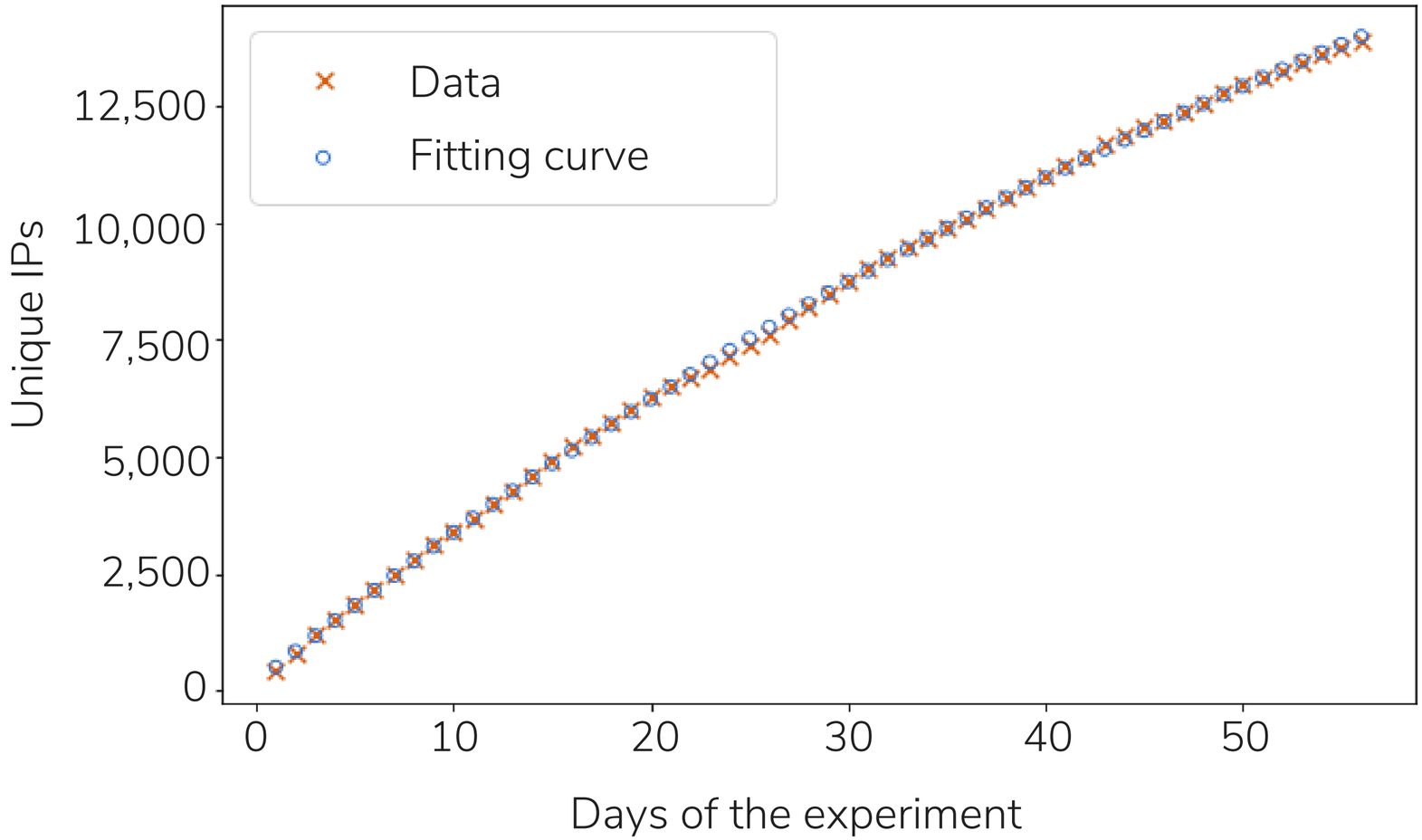
What does it mean?

Bots of our experiment are single-request bots

BUT

Pool significantly smaller than expected

Fitting cumulative curve of new unique IPs



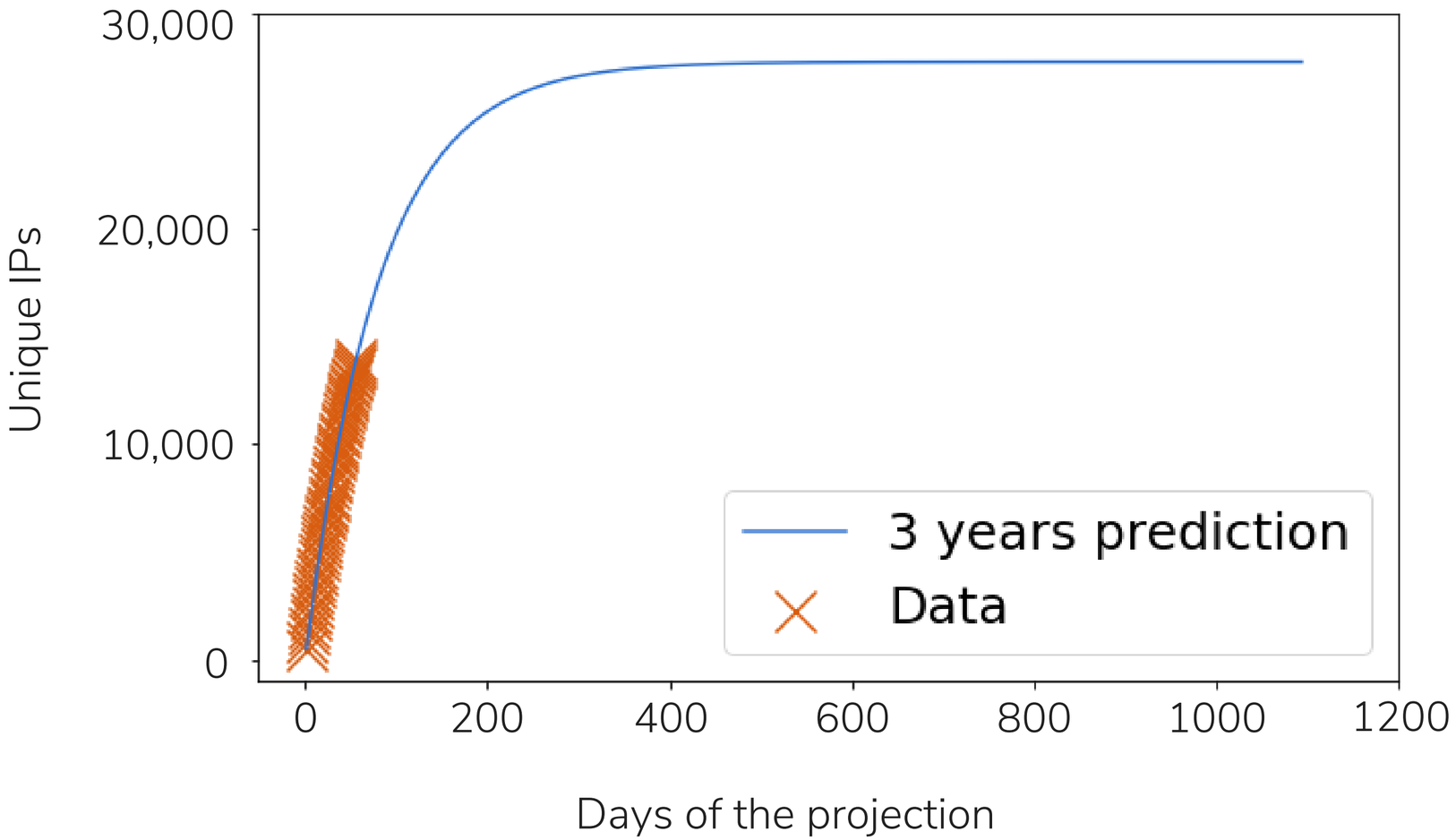
$$f=a^{*}\left(1-e^{\frac{-(x-b)}{c}}\right)$$

a=2.77e+04

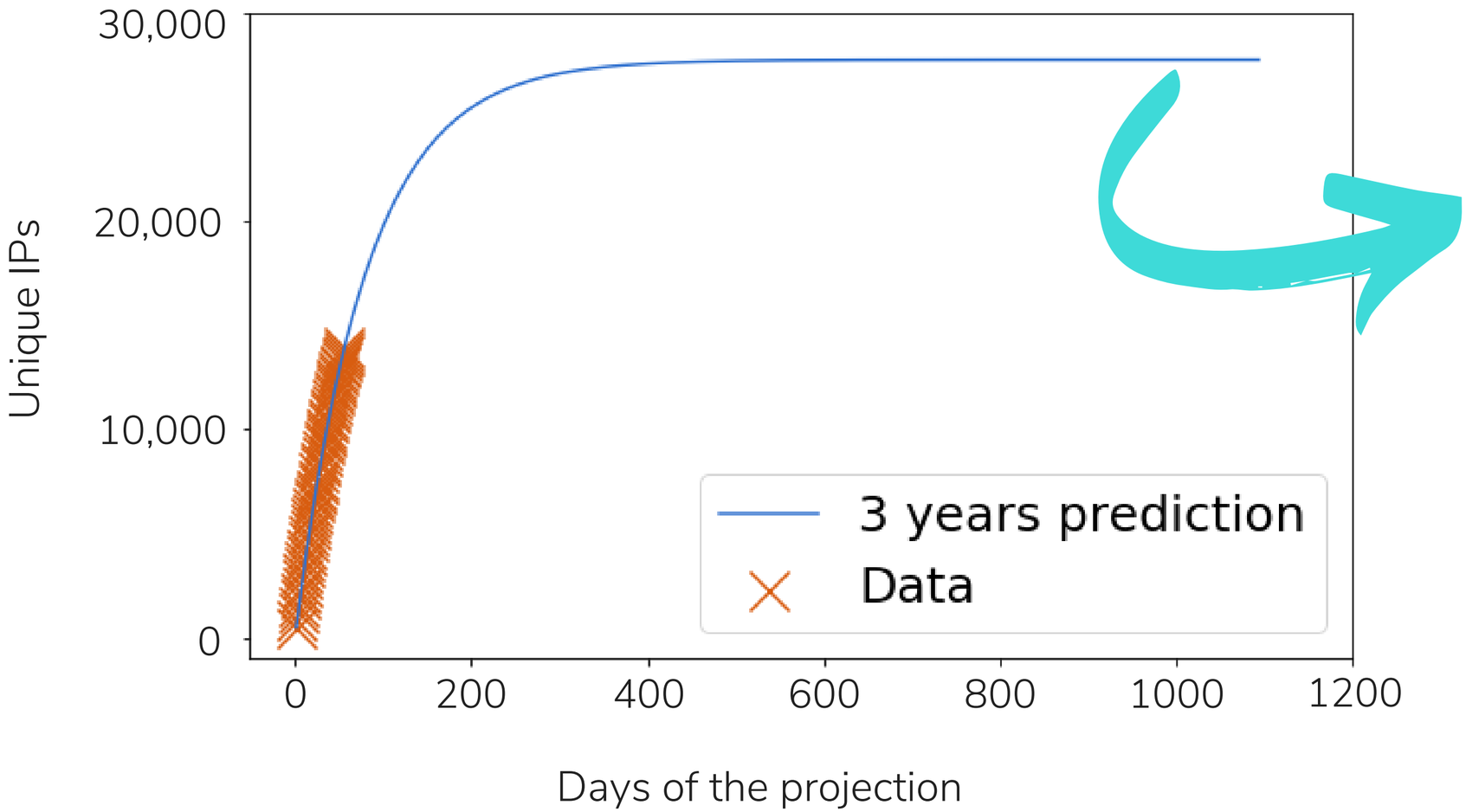
b=-4.78e-01

c=8.05e+01

Three years prediction

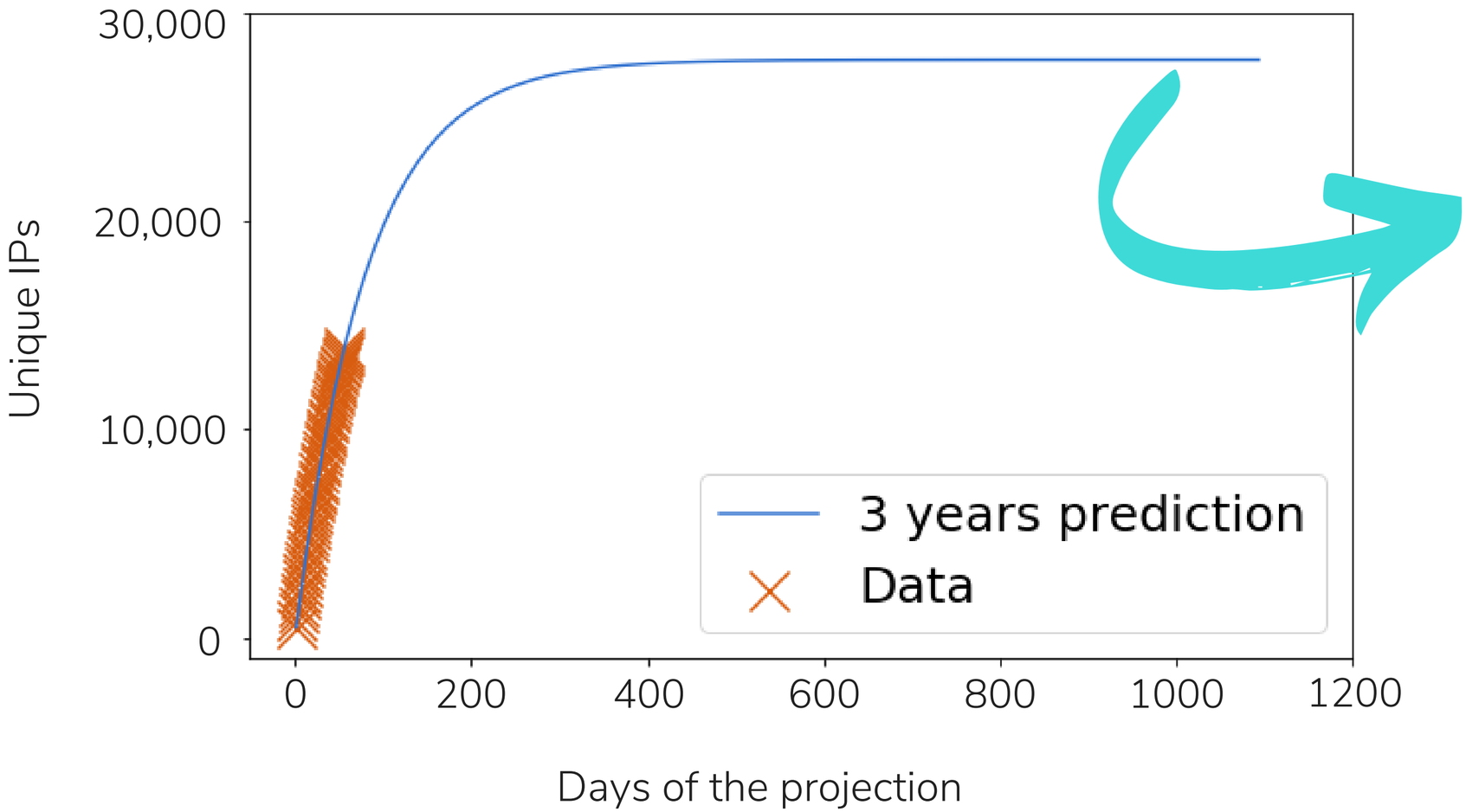


Three years prediction



The plateau is less than 30,000 IPs

Three years prediction



The plateau is less than 30,000 IPs

Consistent with the previous approach

5. Conclusion and future work

What have we shown?

Behavior of a specific web scraping botnet

What have we shown?

Behavior of a specific web scraping botnet

Collected IPs were likely provided by proxy services

What have we shown?

Behavior of a specific web scraping botnet

Collected IPs were likely provided by proxy services

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

And now? 🤔

Small dataset on single-request bot, our conclusions cannot be directly extended

And now? 🤔

Small dataset on single-request bot, our conclusions cannot be directly extended

What can we do

- Encourage similar experiments to confirm or deny
- Test in a large scale experiment with the IT provider

And now? 🤔

Small dataset on single-request bot, our conclusions cannot be directly extended

What can we do

- Encourage similar experiments to confirm or deny
- Test in a large scale experiment with the IT provider

If our findings apply, IP-blocking strategy could be effective

Thank you!

How to reach us

Elisa Chiapponi - elisa.chiapponi@eurecom.fr
Marc Dacier - marc.dacier@eurecom.fr
Massimiliano Todisco - massimiliano.todisco@eurecom.fr
Onur Catakoglu - onur.catakoglu@amadeus.com
Olivier Thonnard - olivier.thonnard@amadeus.com