

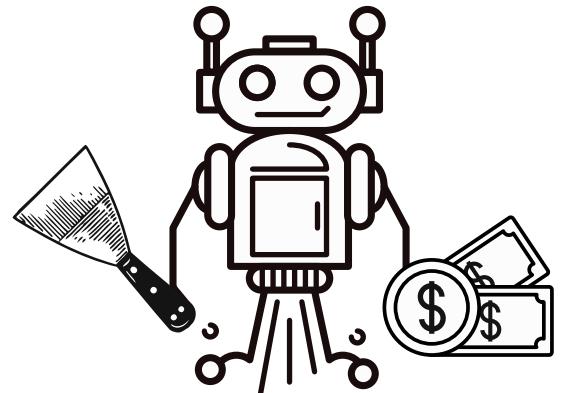
# The bots arms race on airlines booking websites

Elisa Chiapponi, Olivier Thonnard, Mohamed Fangar, Vincent Rigal

{elisa.chiapponi, olivier.thonnard, mohamed.fangar, vincent.rigal}@amadeus.com

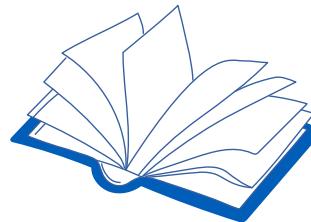
EU MINI AVTECH EXCHANGE

21st September 2022



amADEUS

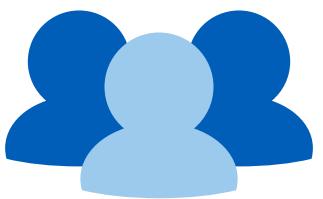
# Who are we?



Elisa Chiapponi, Phd student Amadeus Global SOC and EURECOM



Finding practical means to defeat scraping bots  
Understanding their ecosystem (actors, techniques, infrastructure)



Global SOC members Olivier Thonnard, Mohamed Fangar, Vincent Rigal  
Academic supervisor Prof. Marc Dacier



Resilient Computing and  
Cybersecurity Center

# Agenda

# Agenda

1

## The battle against scrapers

- Which weapons are they using?
- What can we do now?

# Agenda

1

## The battle against scrapers

- Which weapons are they using?
- What can we do now?

2

## WebApp Honeypot

- Is it possible to lure attackers?

# Agenda

1

## The battle against scrapers

- Which weapons are they using?
- What can we do now?

2

## WebApp Honeypot

- Is it possible to lure attackers?

3

## RESIP detection

- Is it possible to detect scrapers taking advantage of Residential IP addresses?



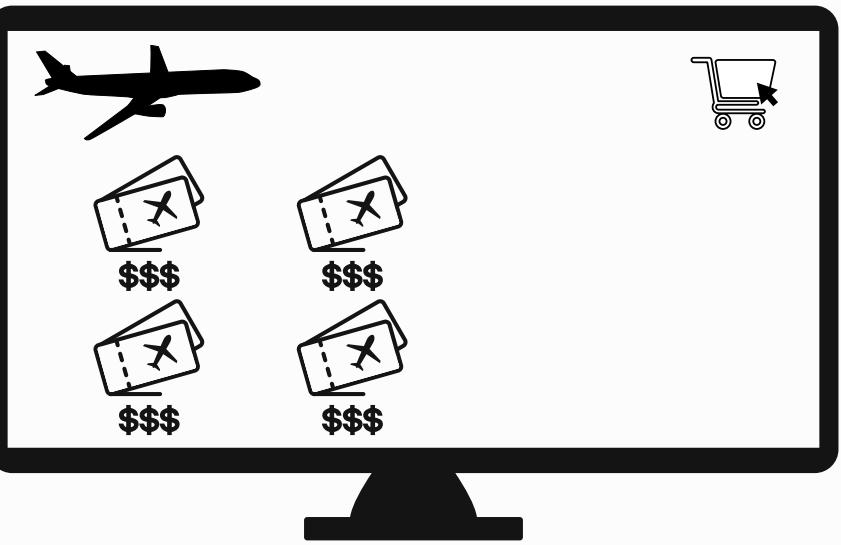
1

# The battle against scrapers

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

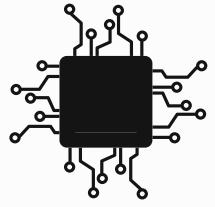
OWASP automated threats to web applications

# Why is scraping a problem?

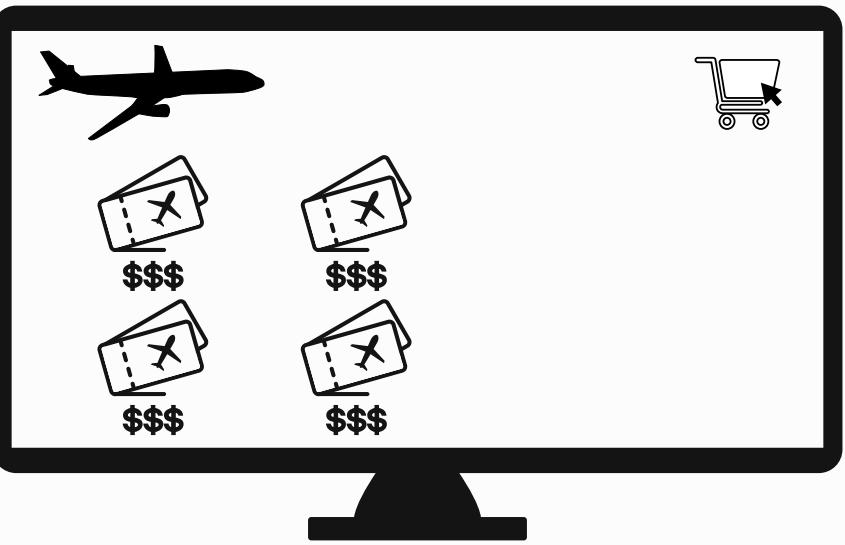


E-commerce websites

# Why is scraping a problem?

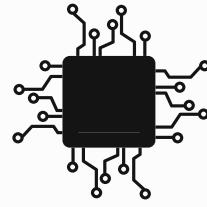


CPU cost

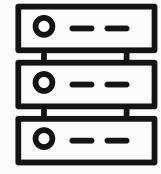


E-commerce websites

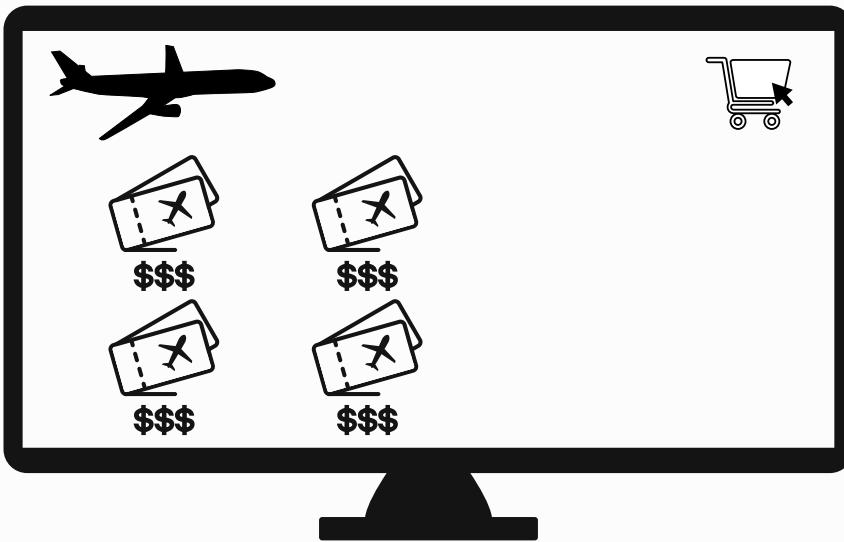
# Why is scraping a problem?



CPU cost

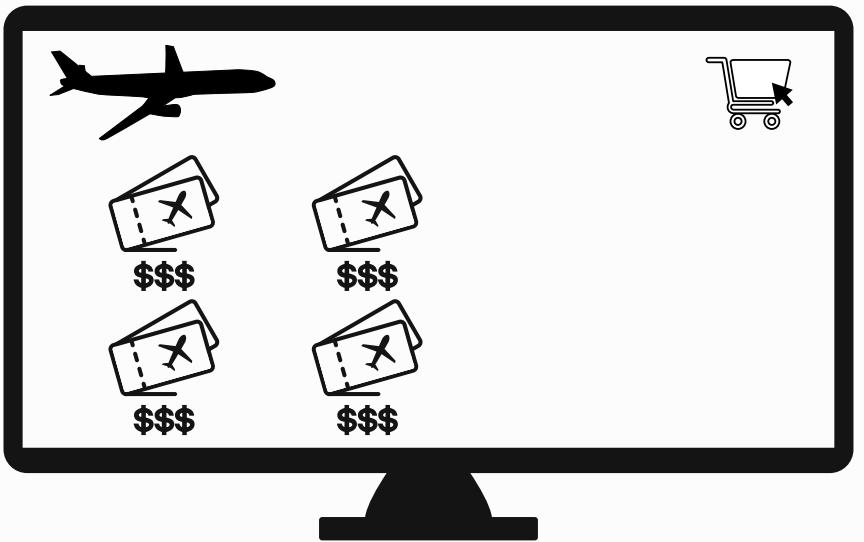
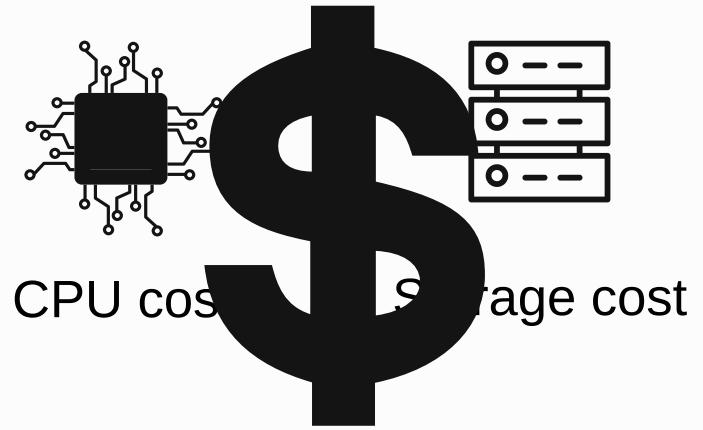


Storage cost

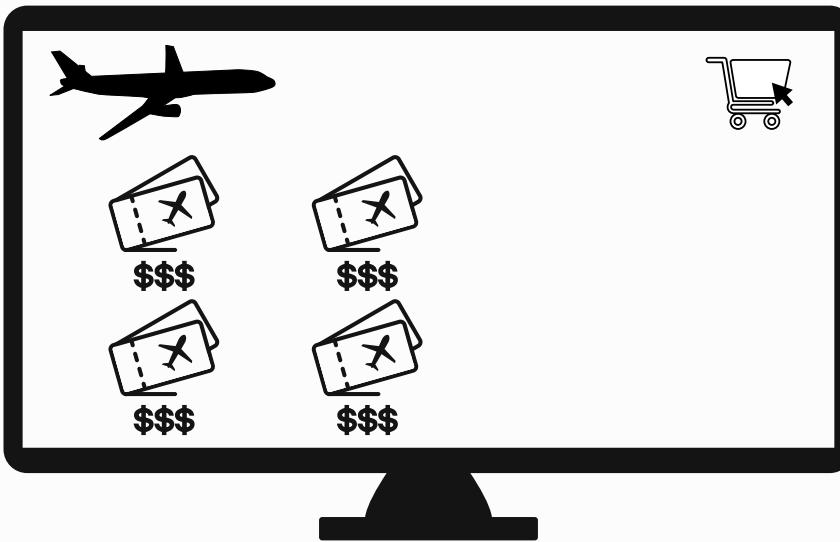
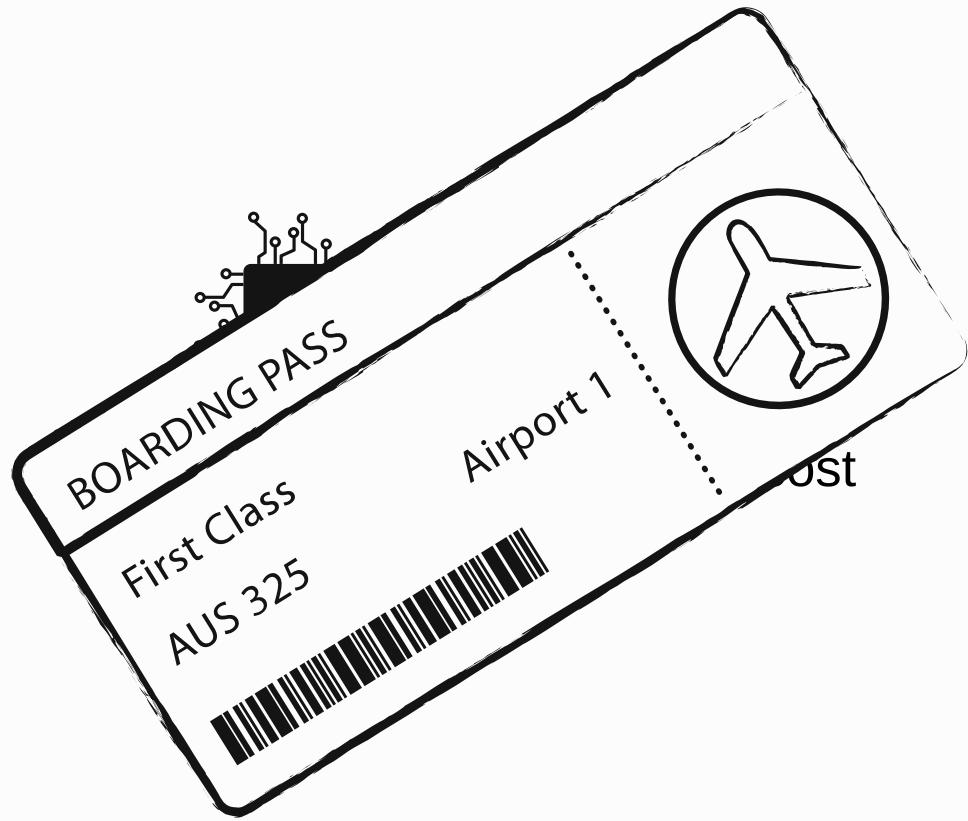


E-commerce websites

# Why is scraping a problem?

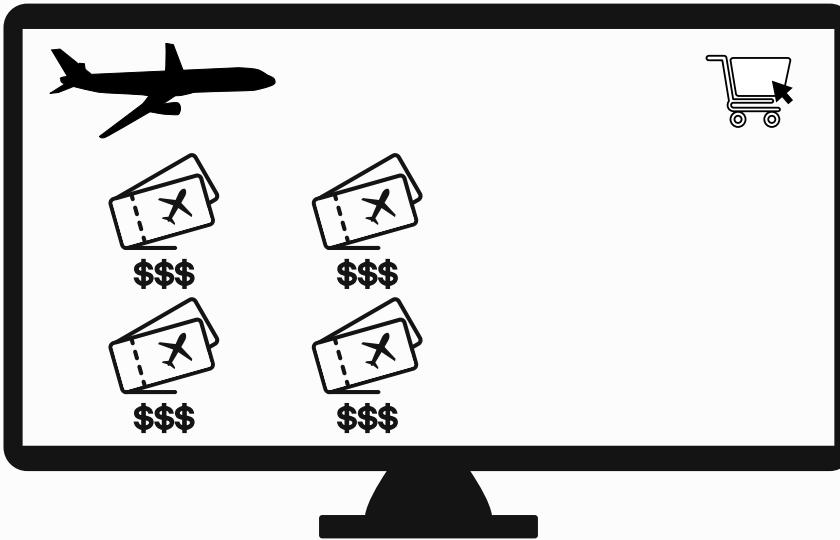
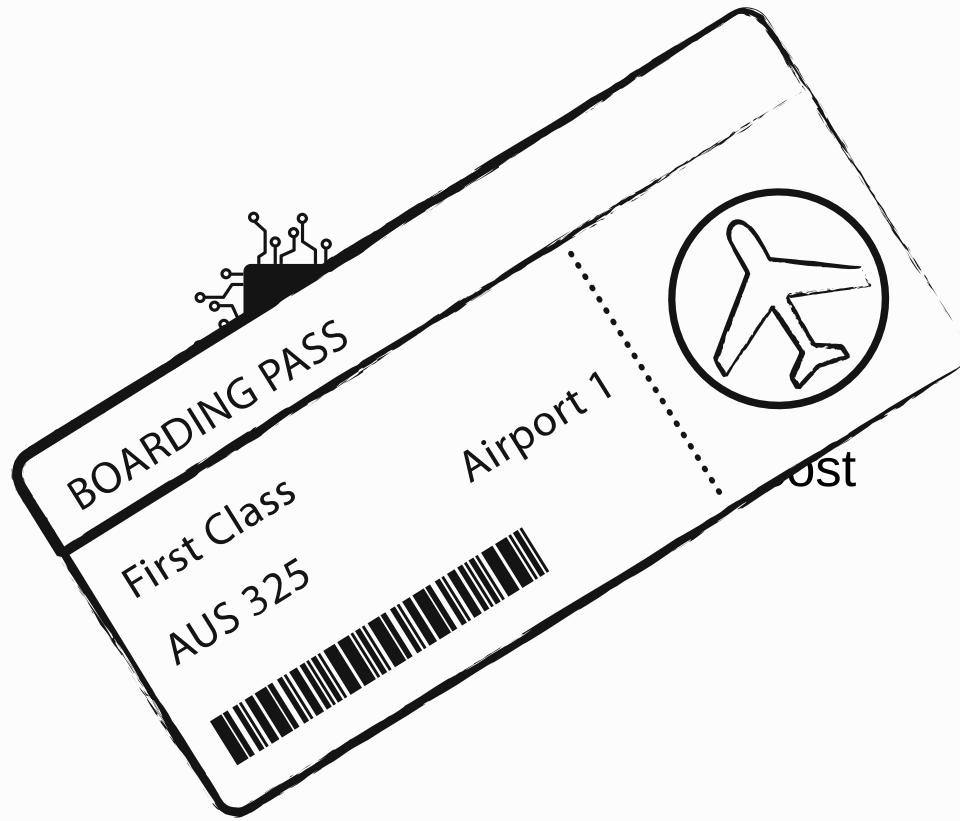


# Why is scraping a problem?

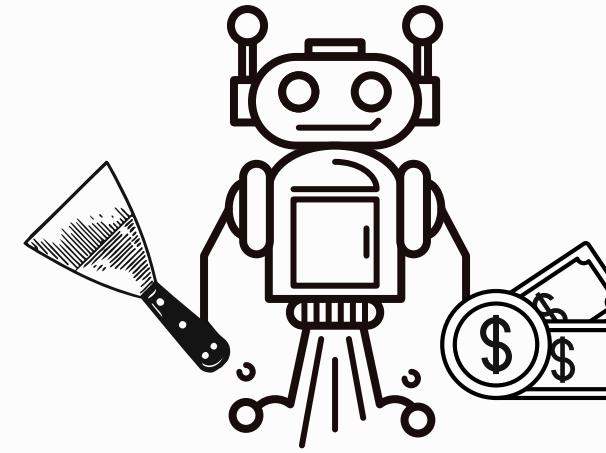


E-commerce websites

# Why is scraping a problem?

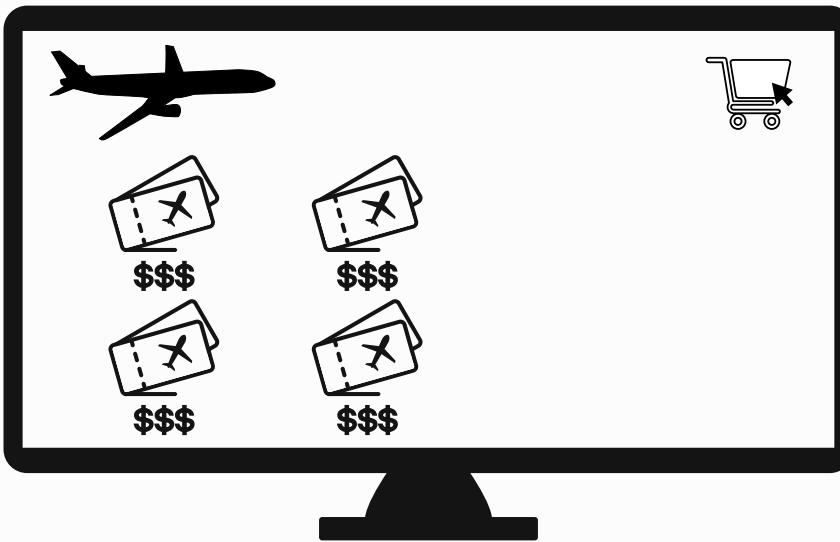
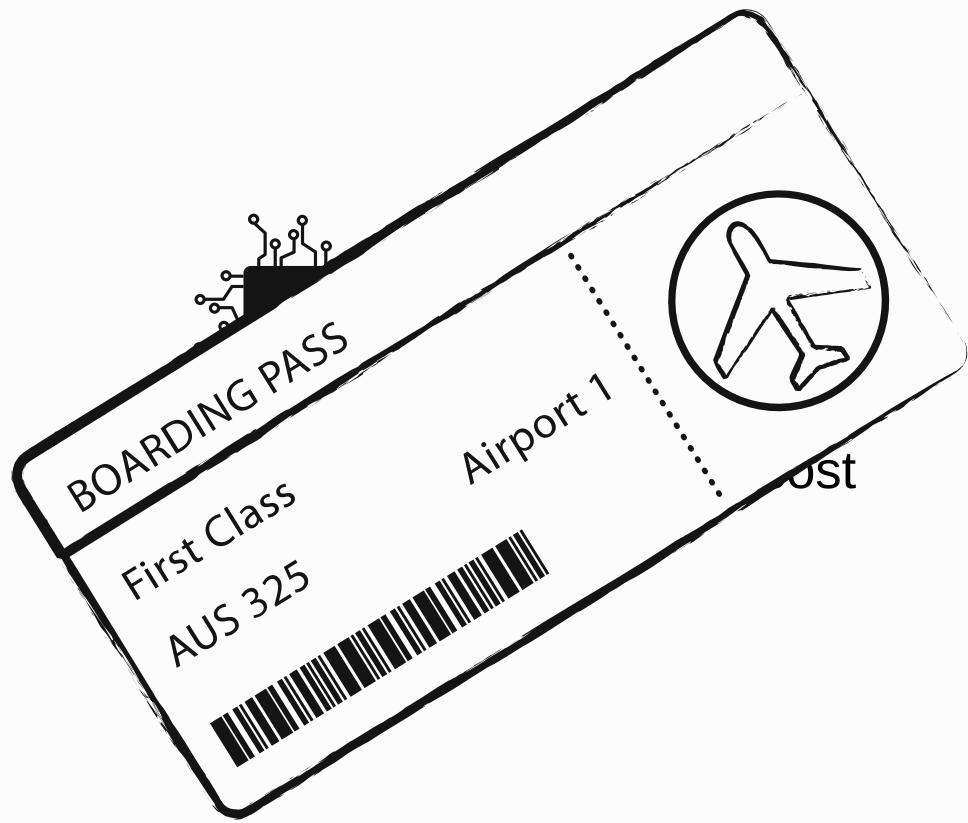


E-commerce websites

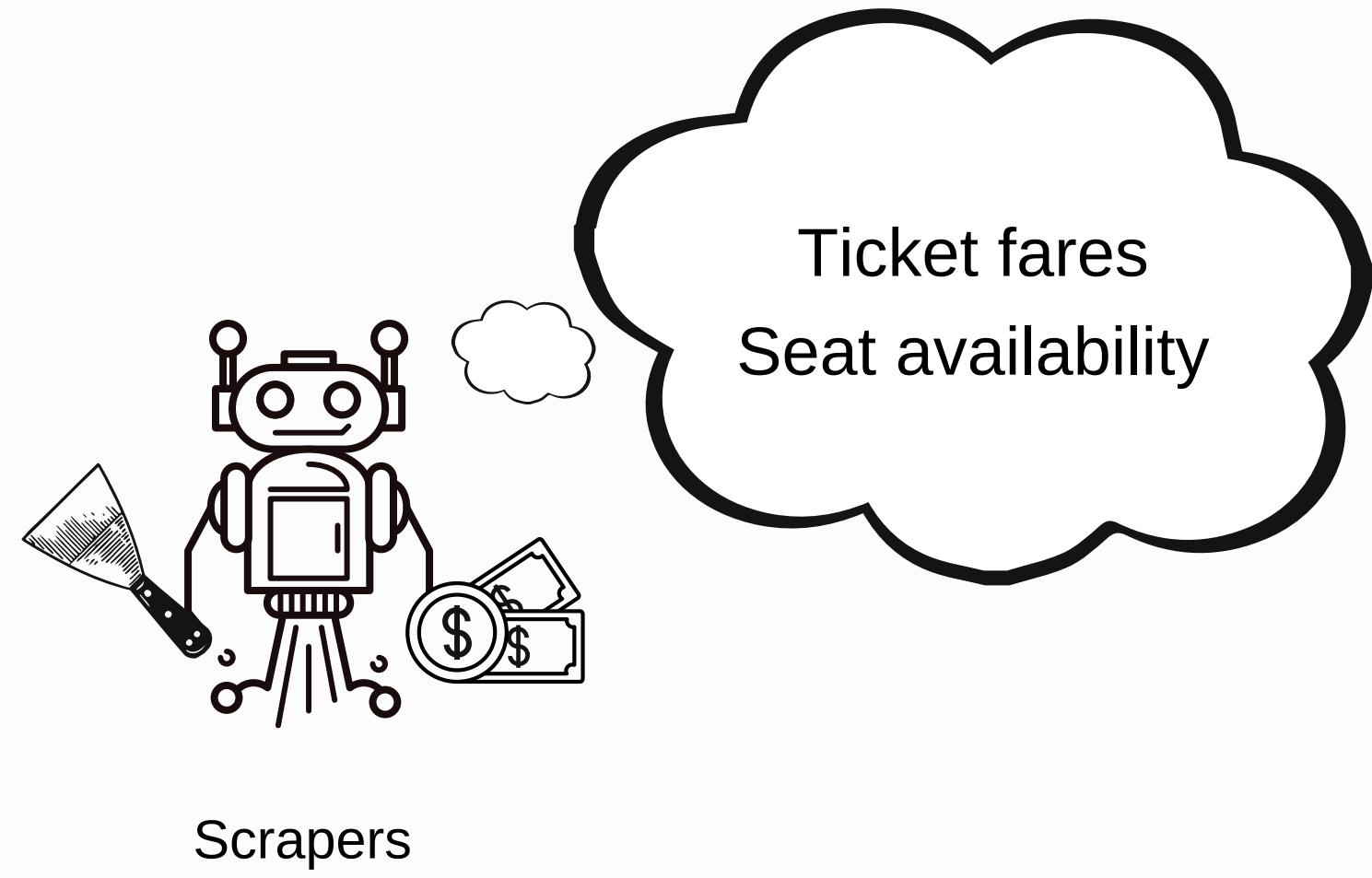


Scrapers

# Why is scraping a problem?

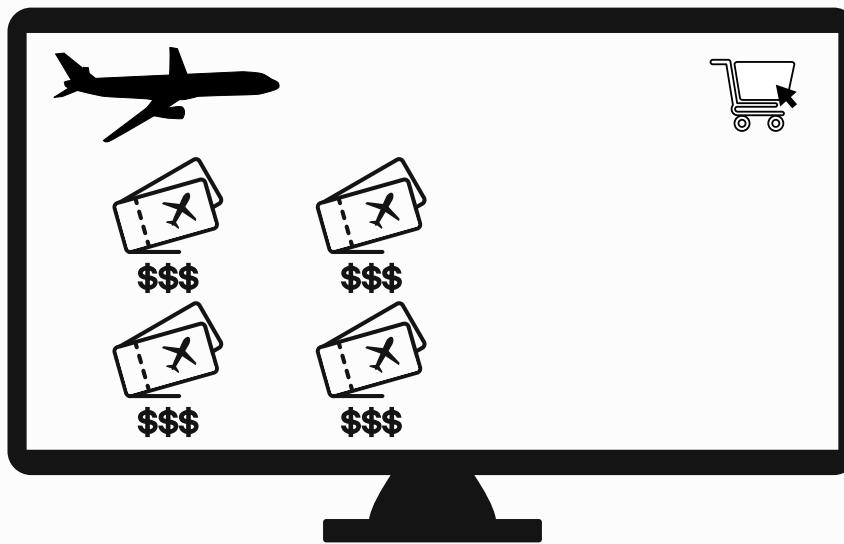
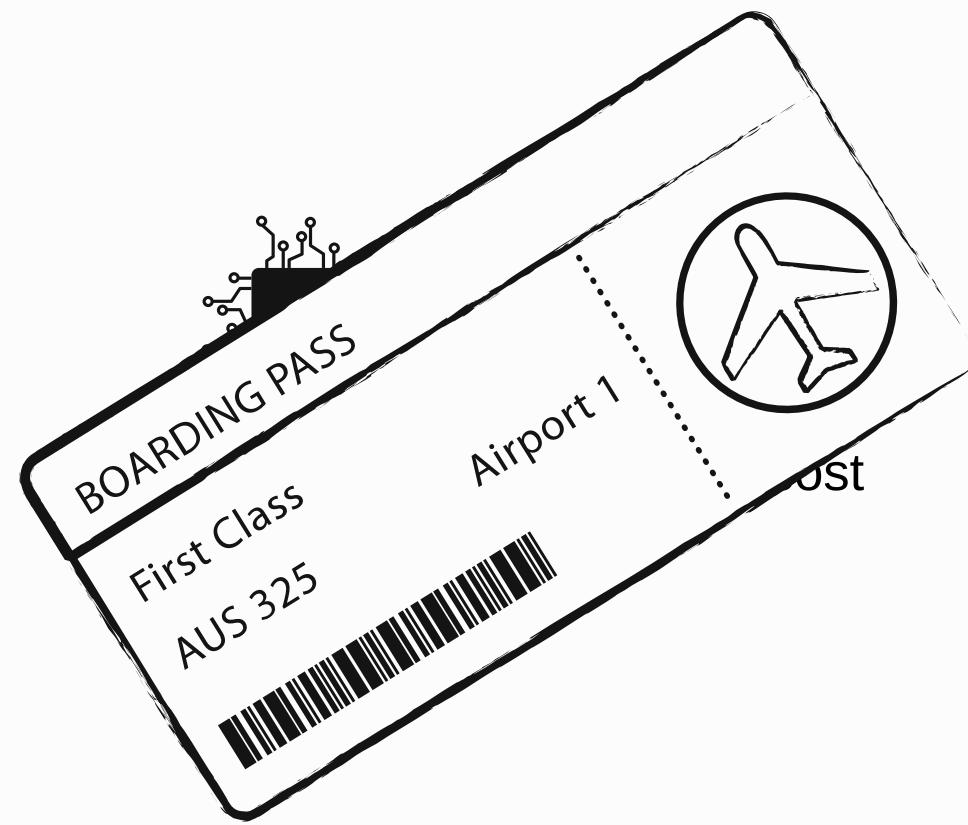


E-commerce websites

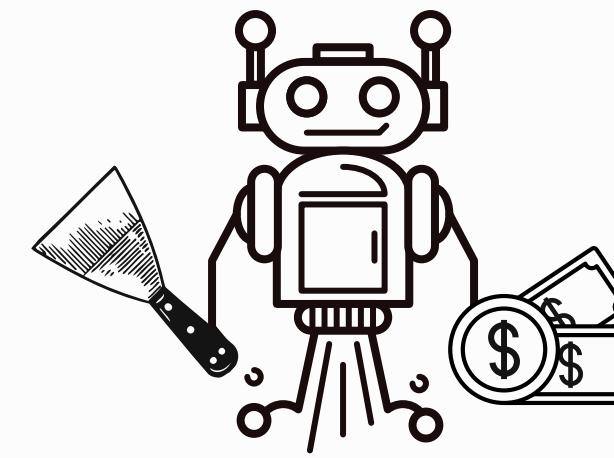


Scrapers

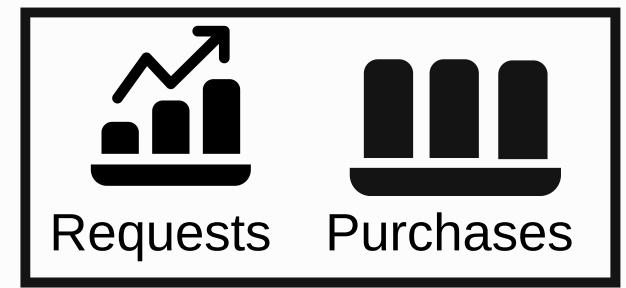
# Why is scraping a problem?



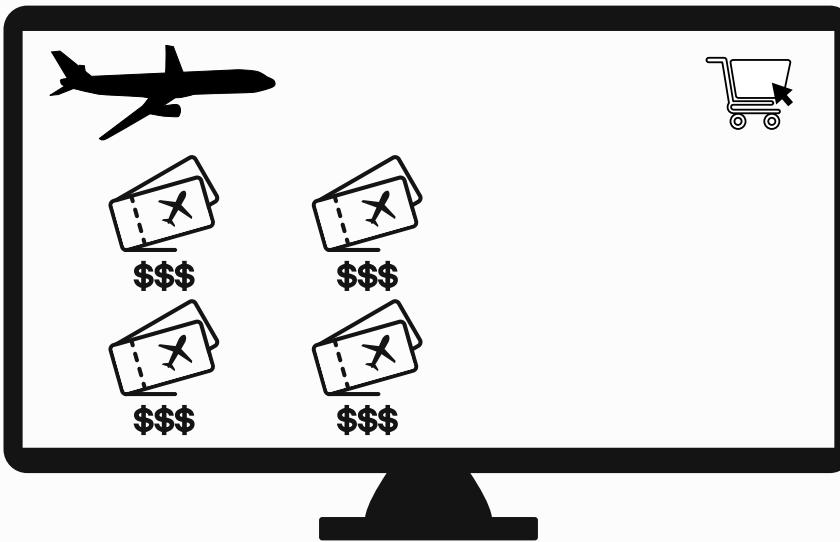
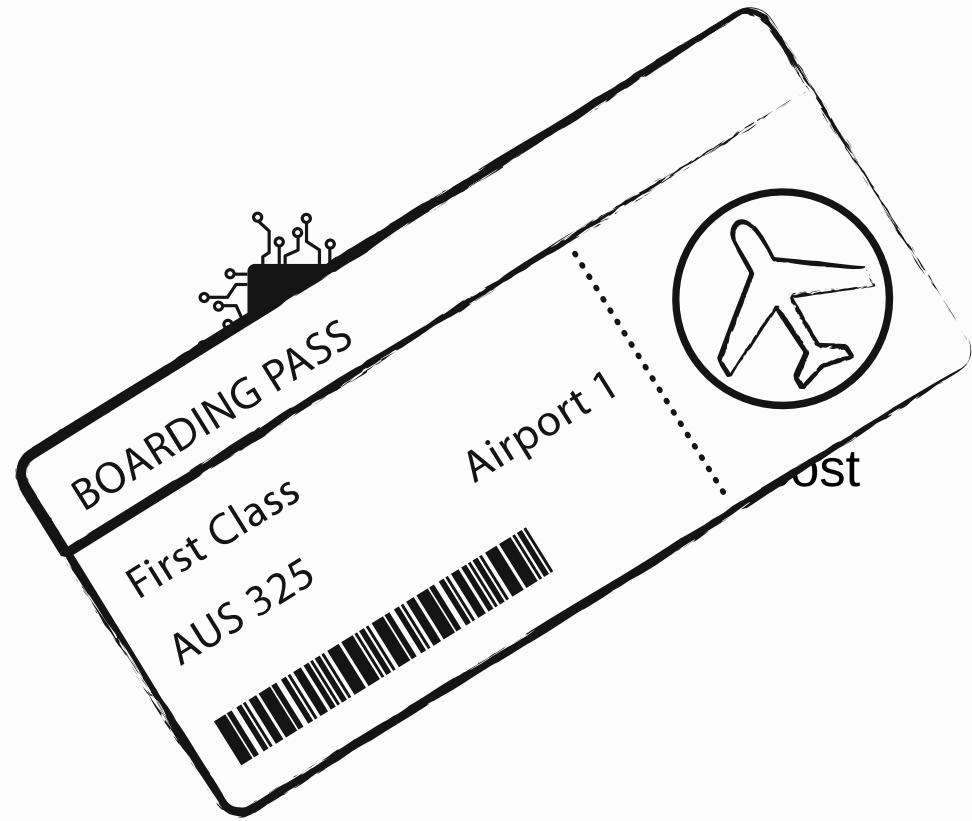
E-commerce websites



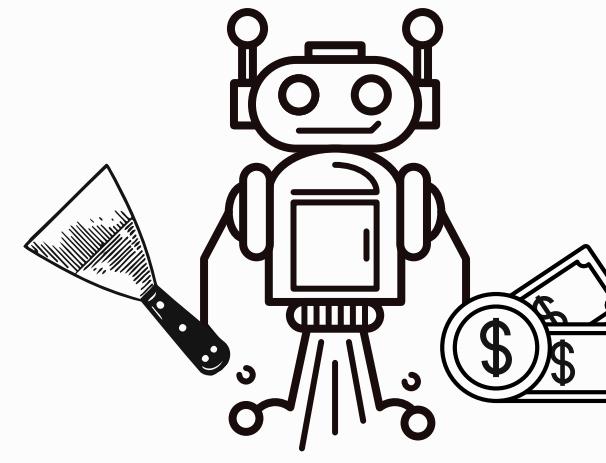
Scrapers



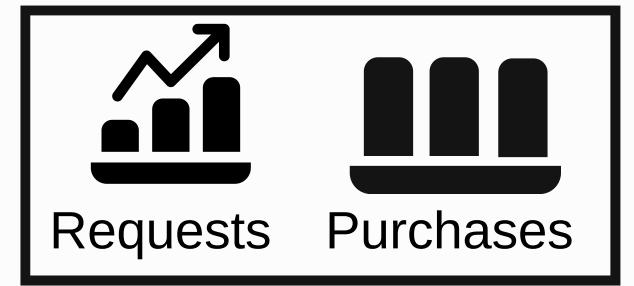
# Why is scraping a problem?



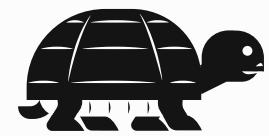
E-commerce websites



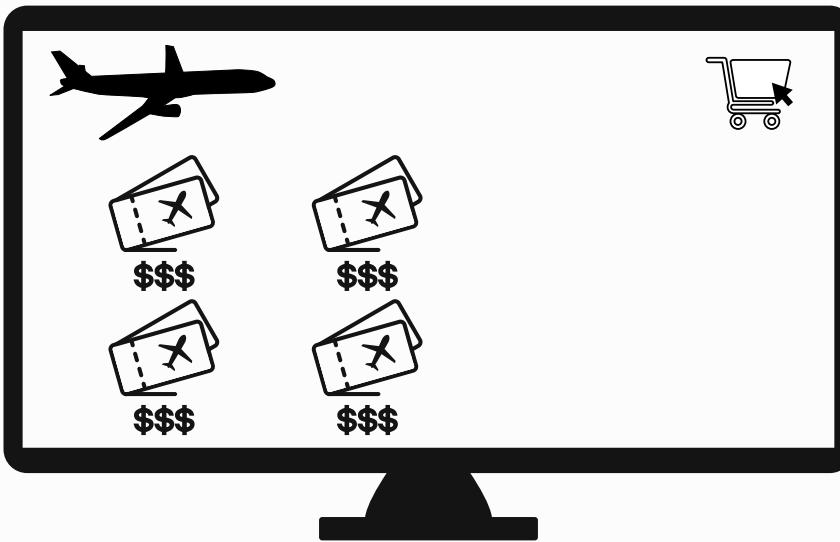
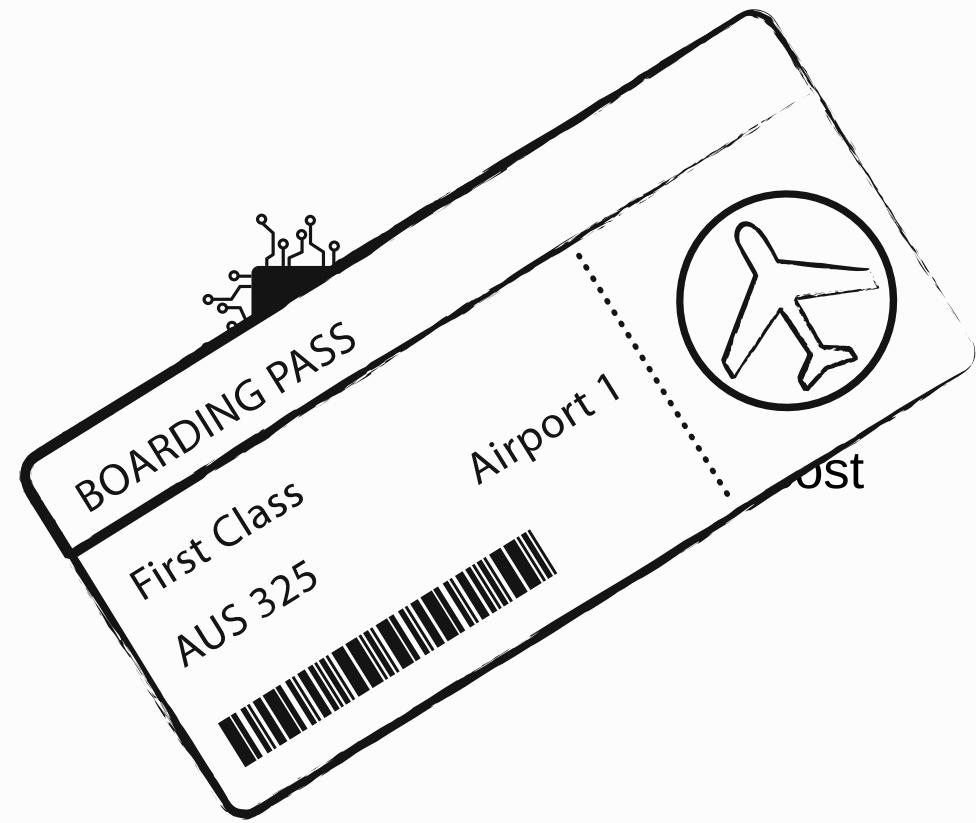
Scrapers



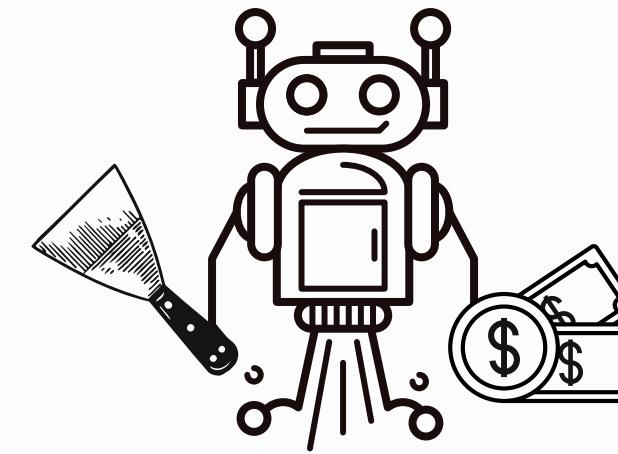
Slow connections



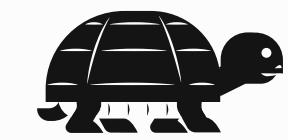
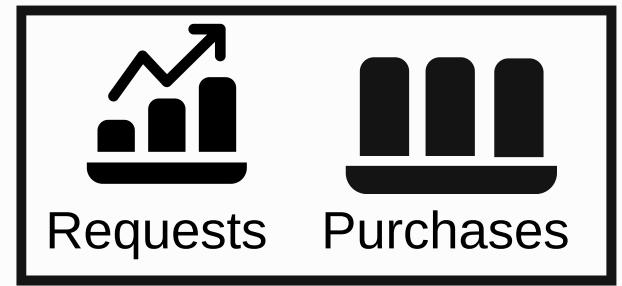
# Why is scraping a problem?



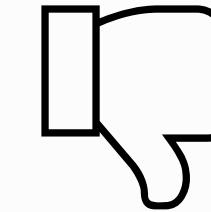
E-commerce websites



Scrapers

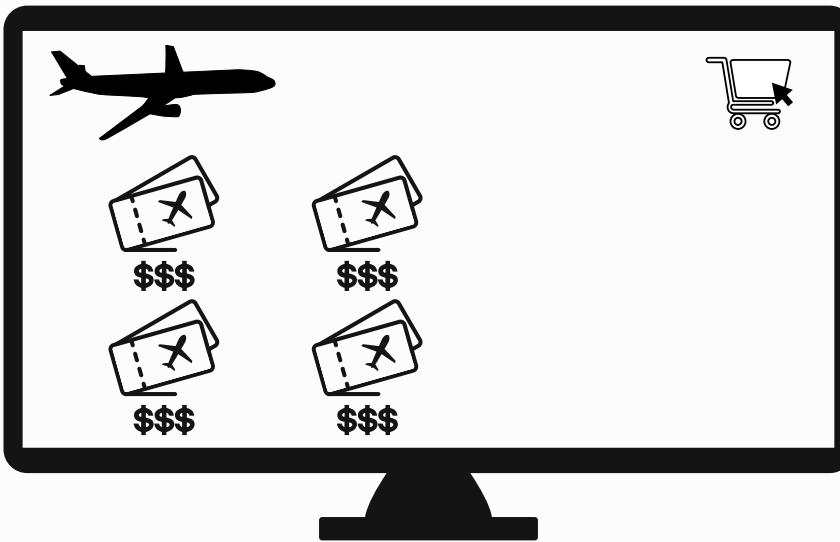


Slow connections

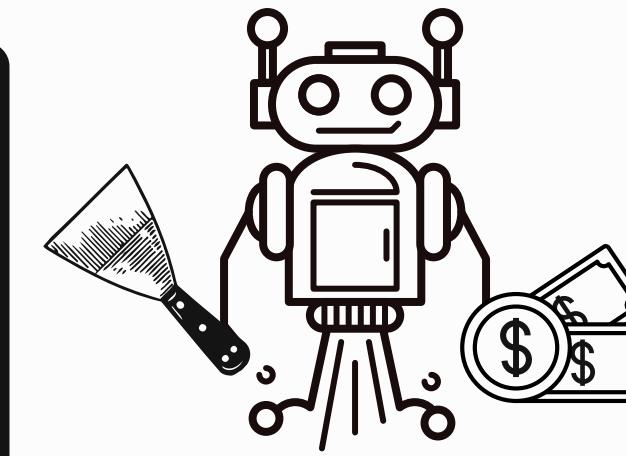


Server down

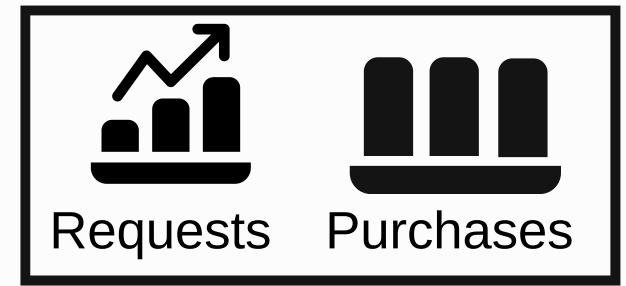
# Why is scraping a problem?



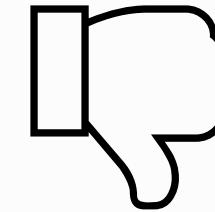
E-commerce websites



Scrapers

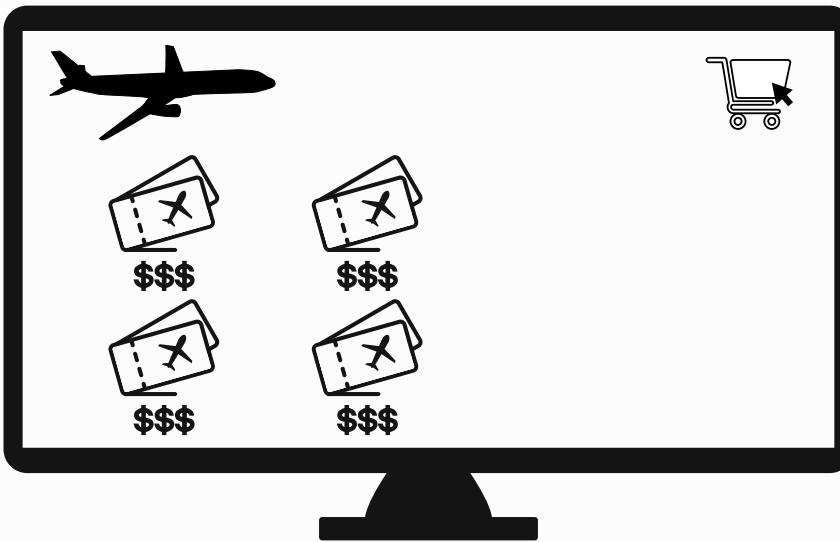


Slow connections

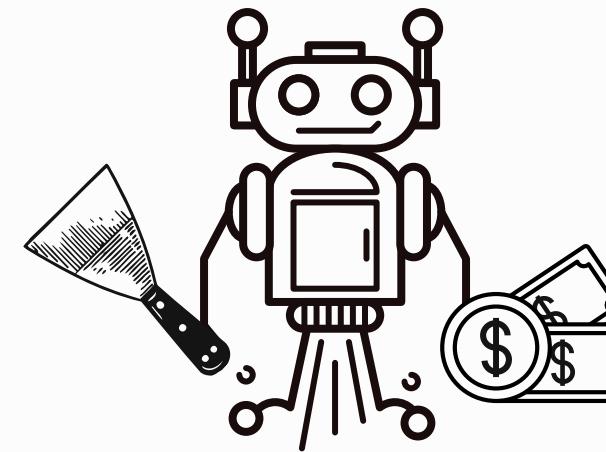


Server down

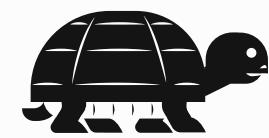
# Why is scraping a problem?



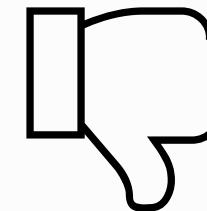
E-commerce websites



Scrapers

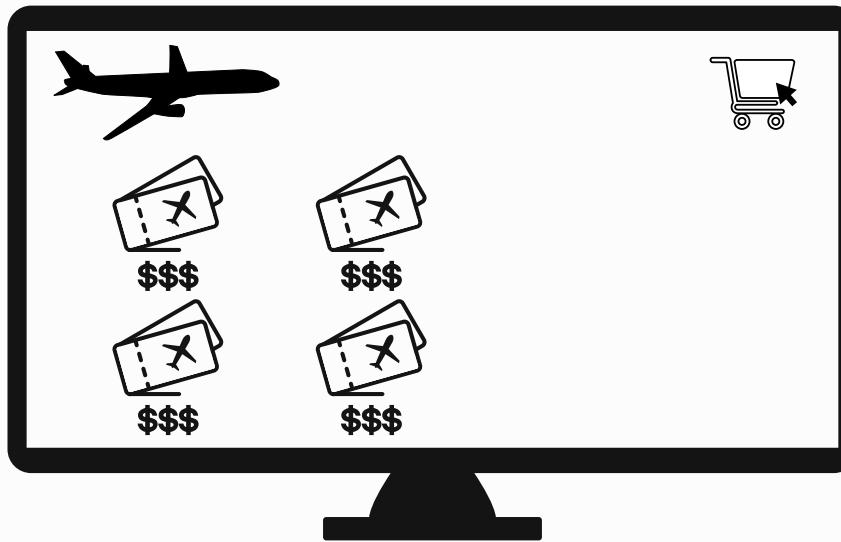


Slow connections

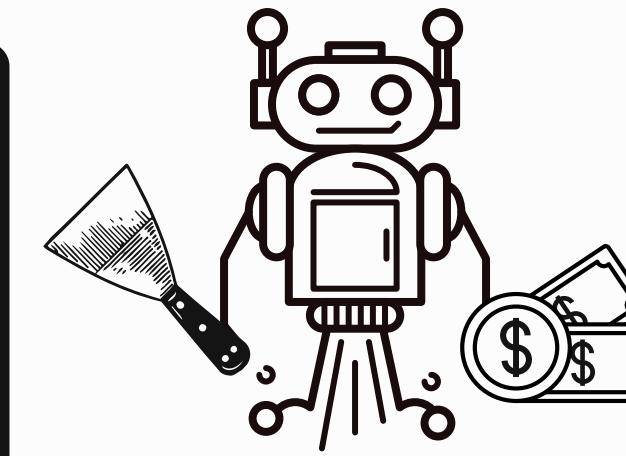


Server down

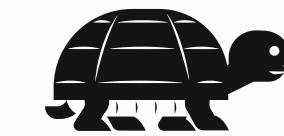
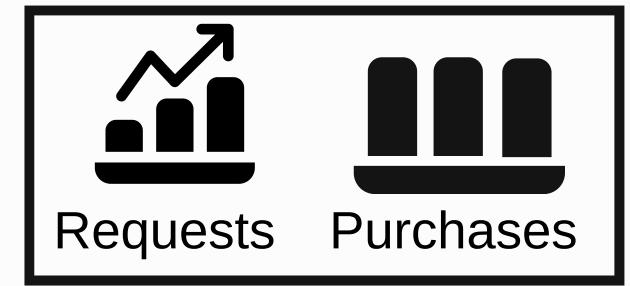
# Why is scraping a problem?



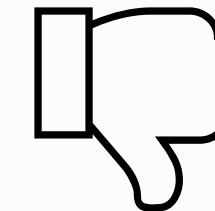
E-commerce websites



Scrapers

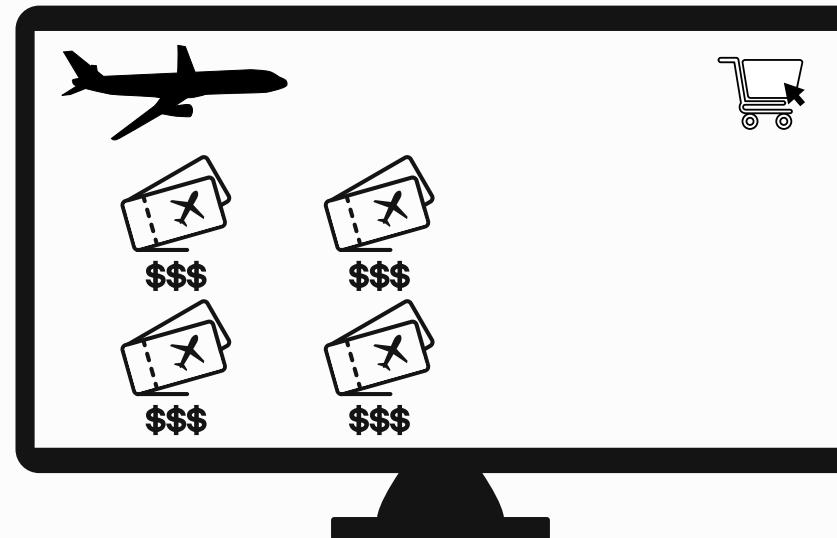


Slow connections



Server down

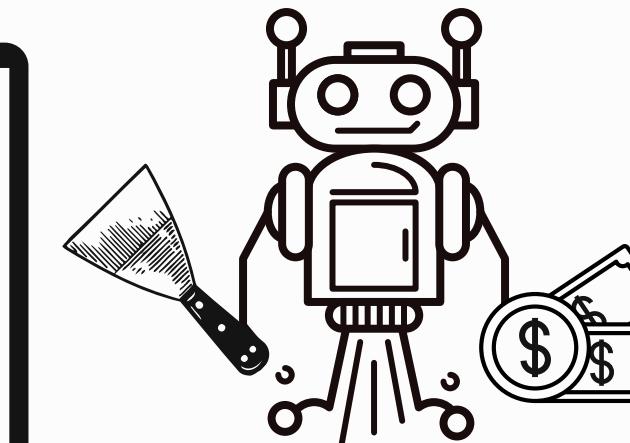
# Why is scraping a problem?



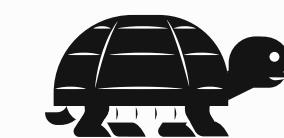
E-commerce websites



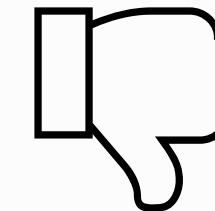
Cloud migration



Scrapers

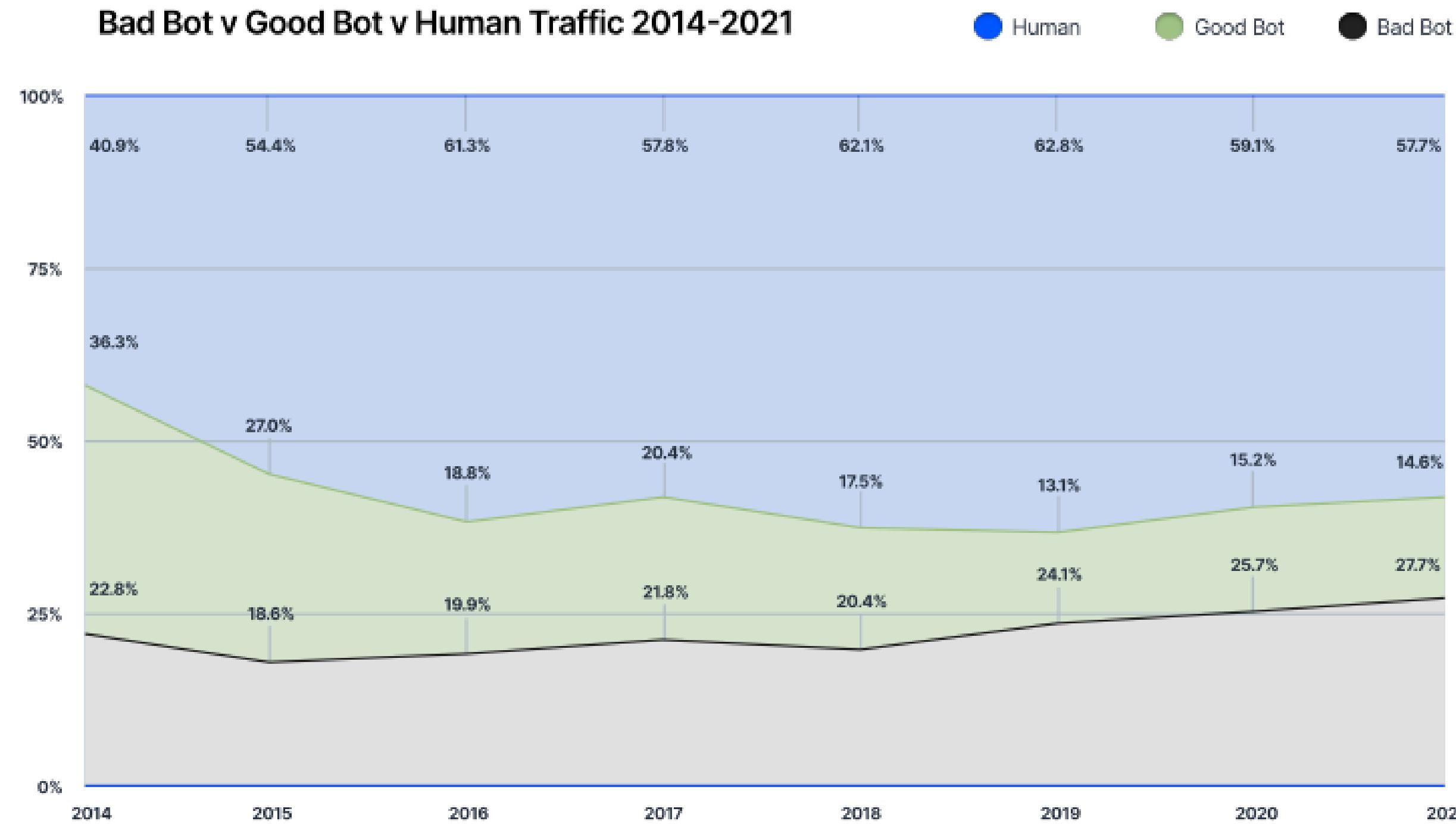


Slow connections



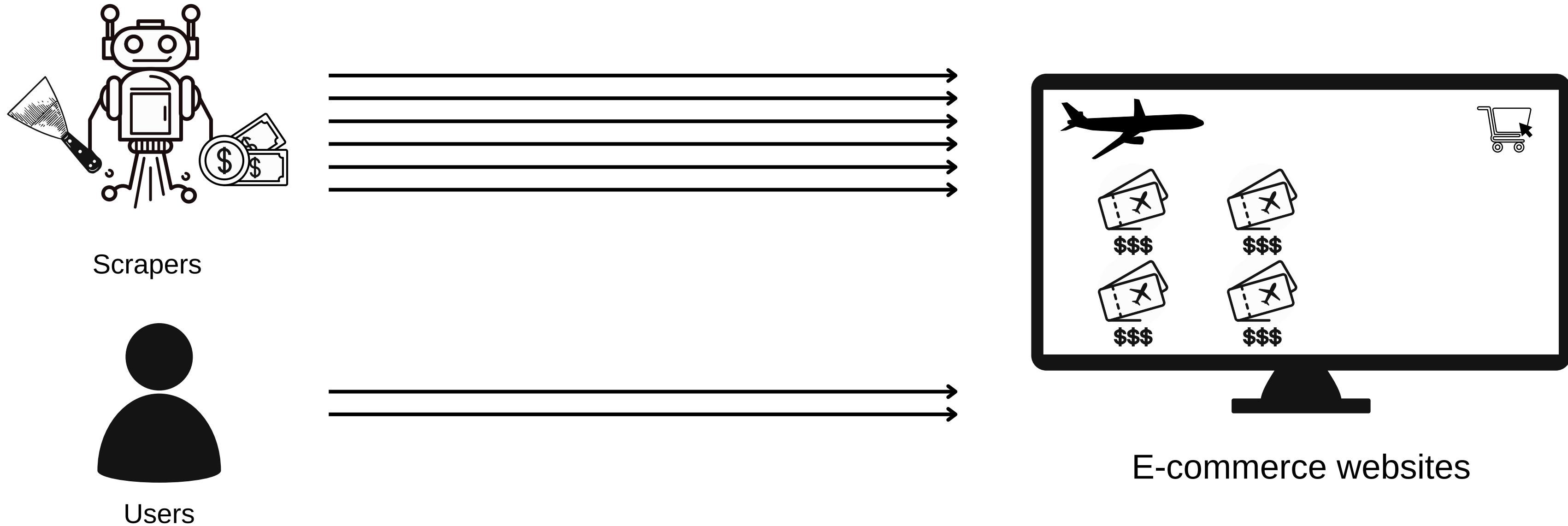
Server down

# Call to the army

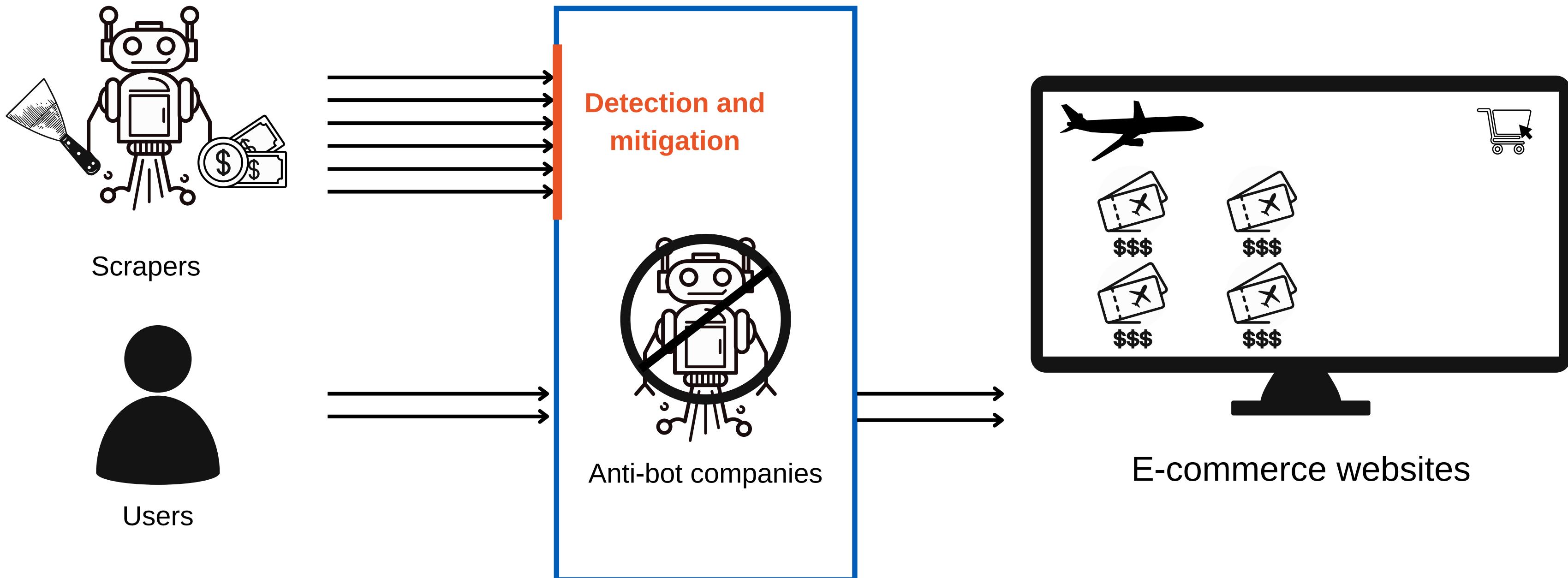


2022 Imperva Bad Bot Report | Evasive Bots Drive Online Fraud

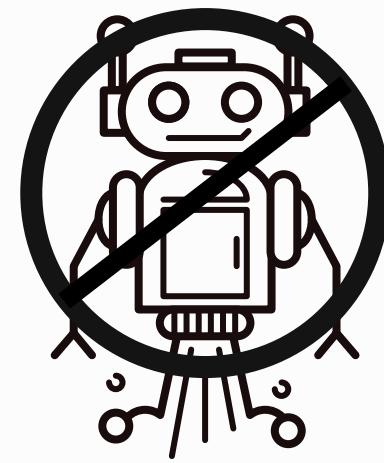
# Call to the army



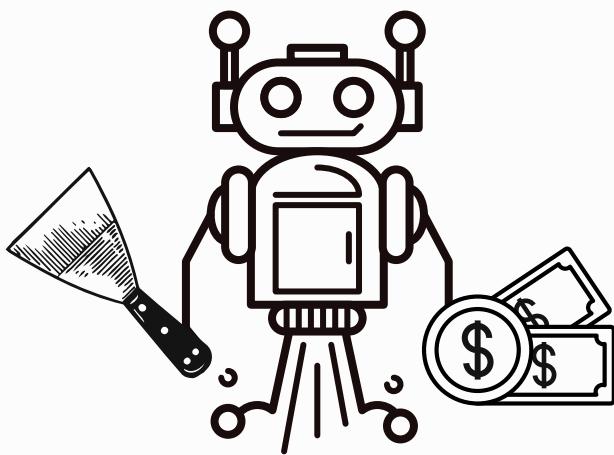
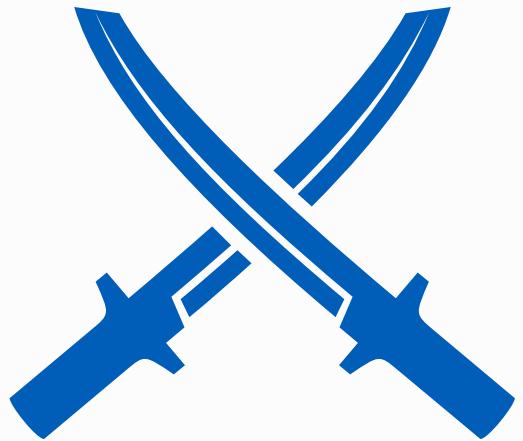
# Call to the army



# A persistent battle

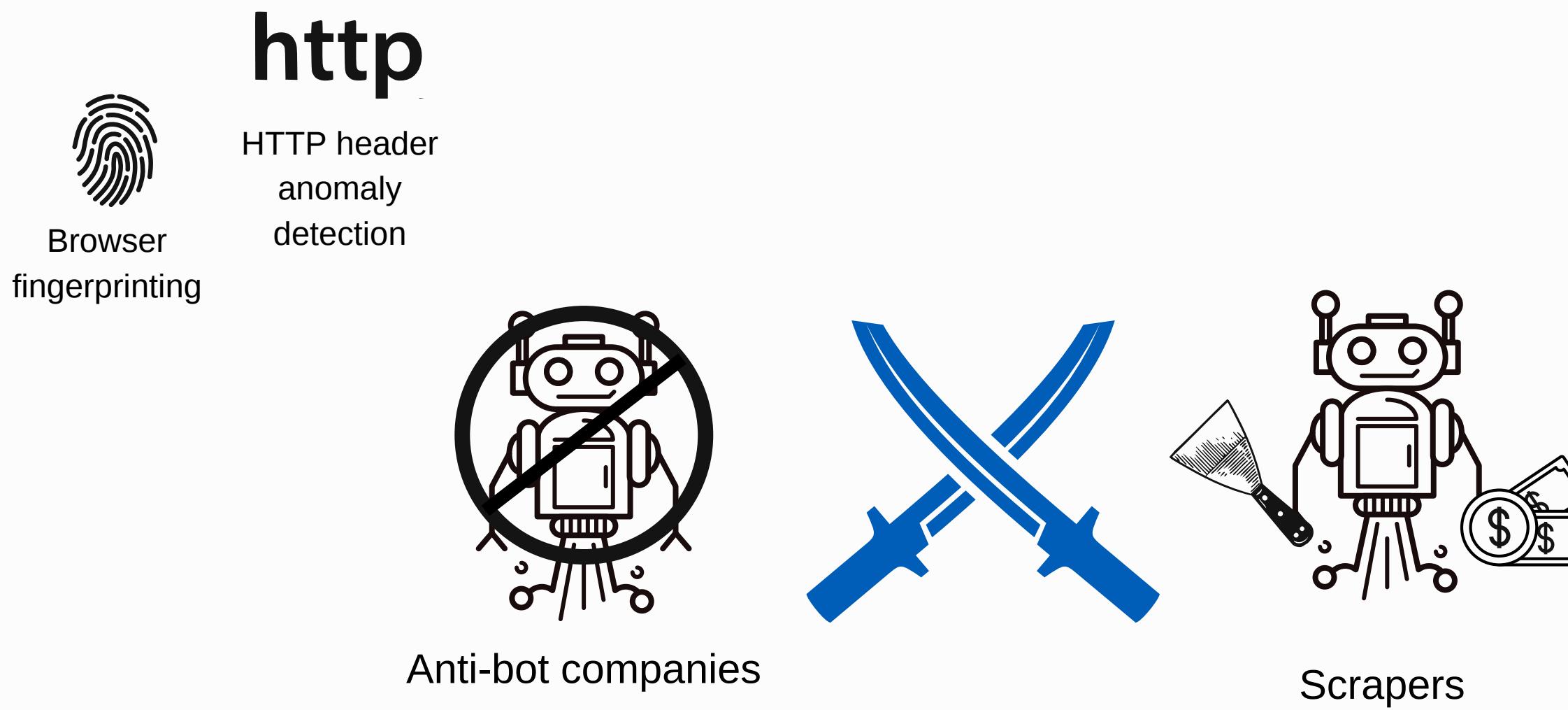


Anti-bot companies

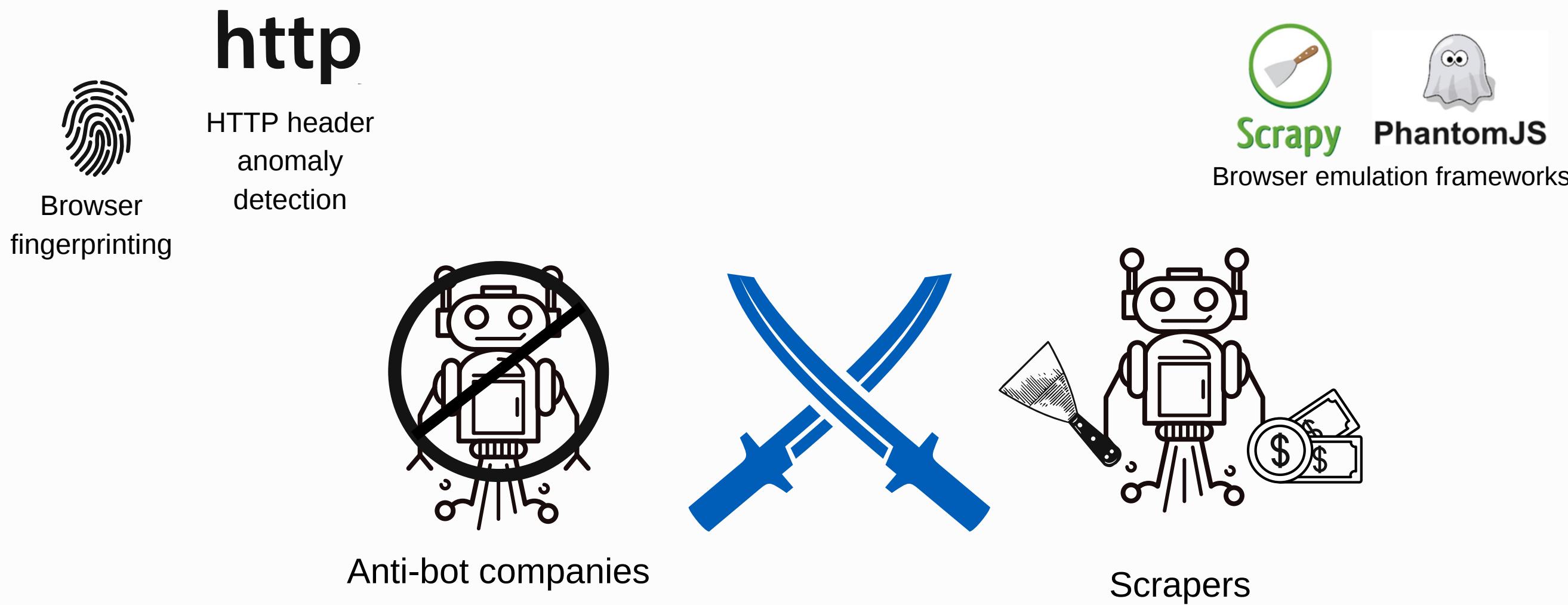


Scrapers

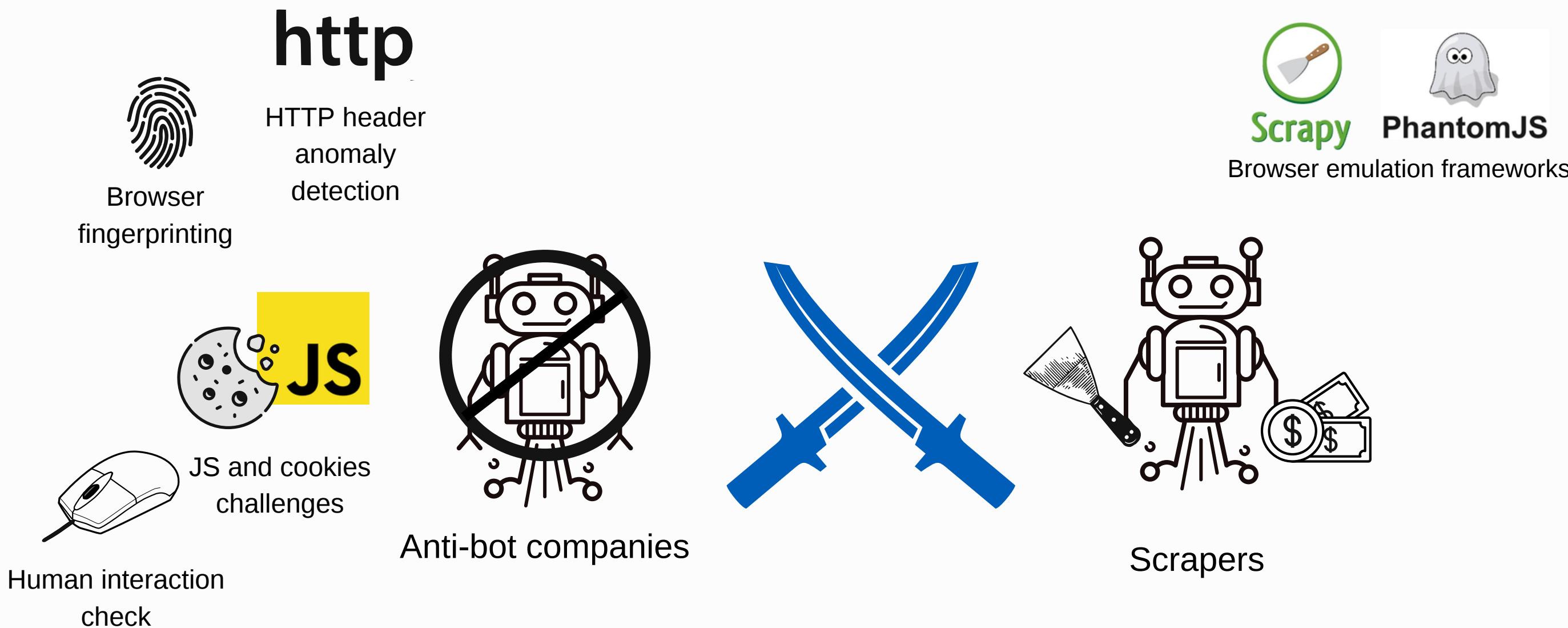
# A persistent battle



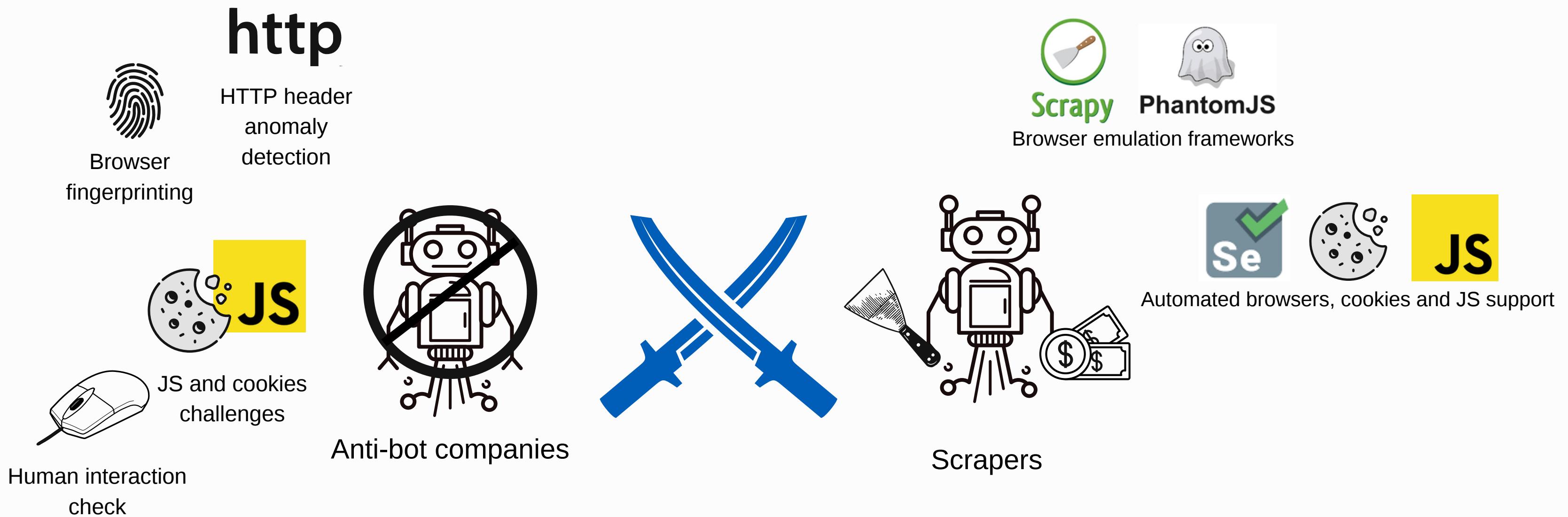
# A persistent battle



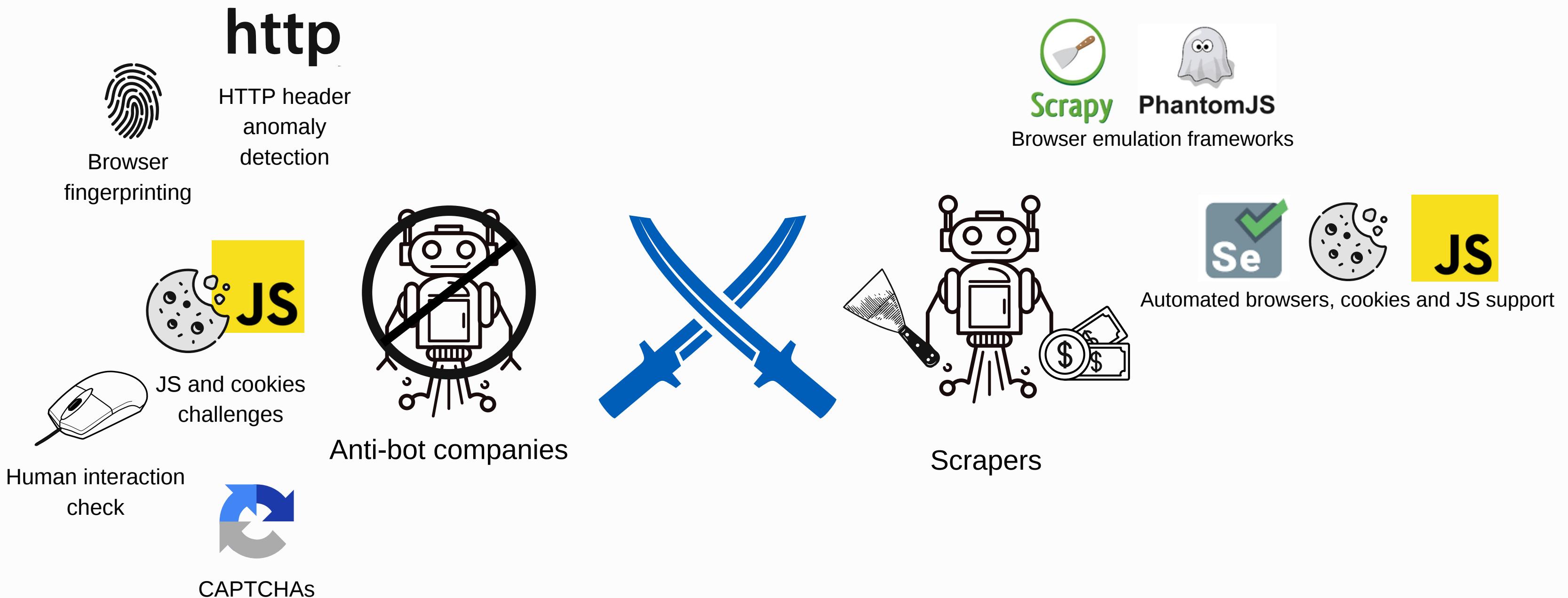
# A persistent battle



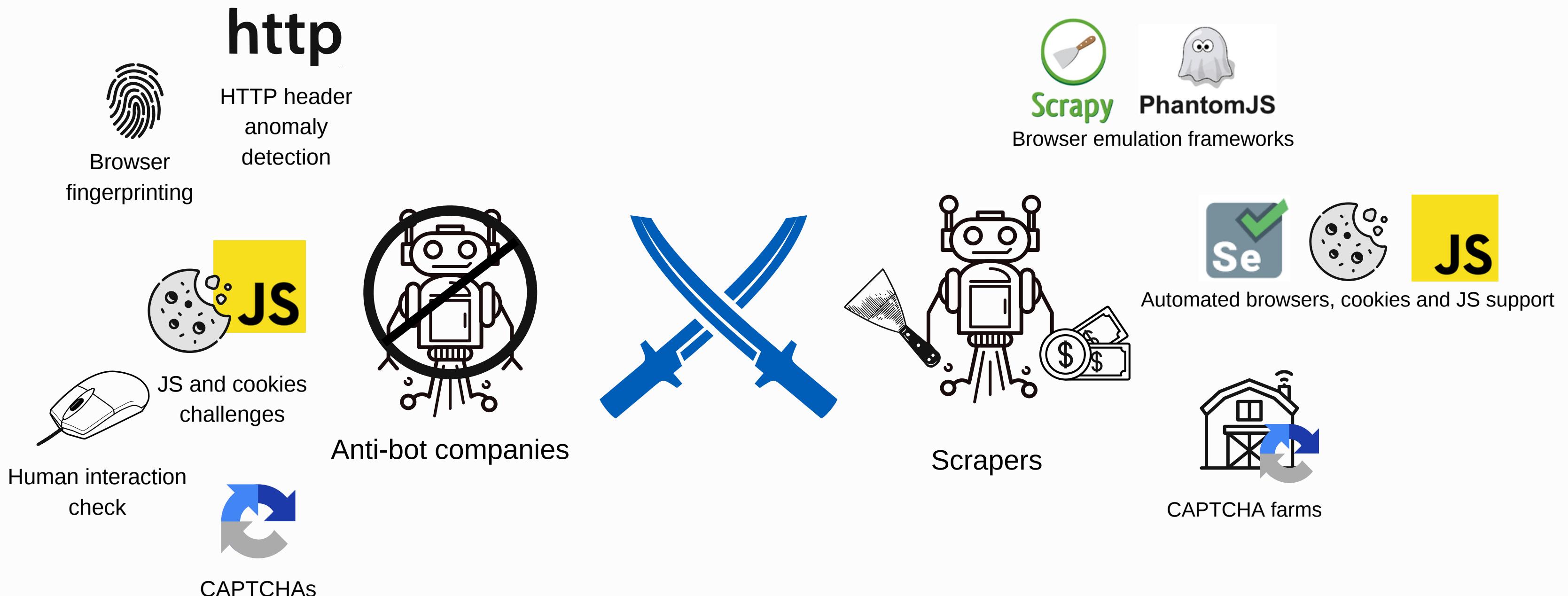
# A persistent battle



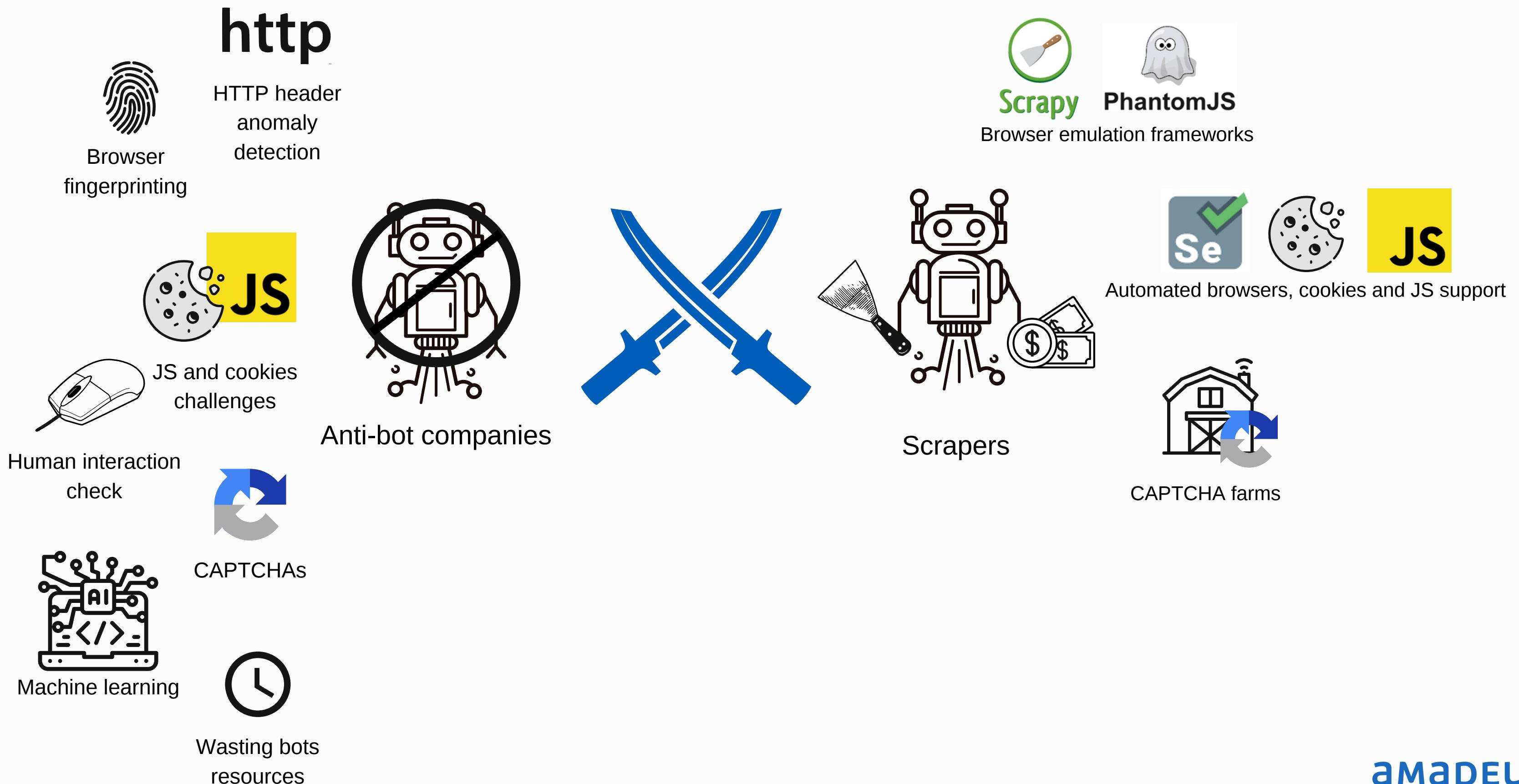
# A persistent battle



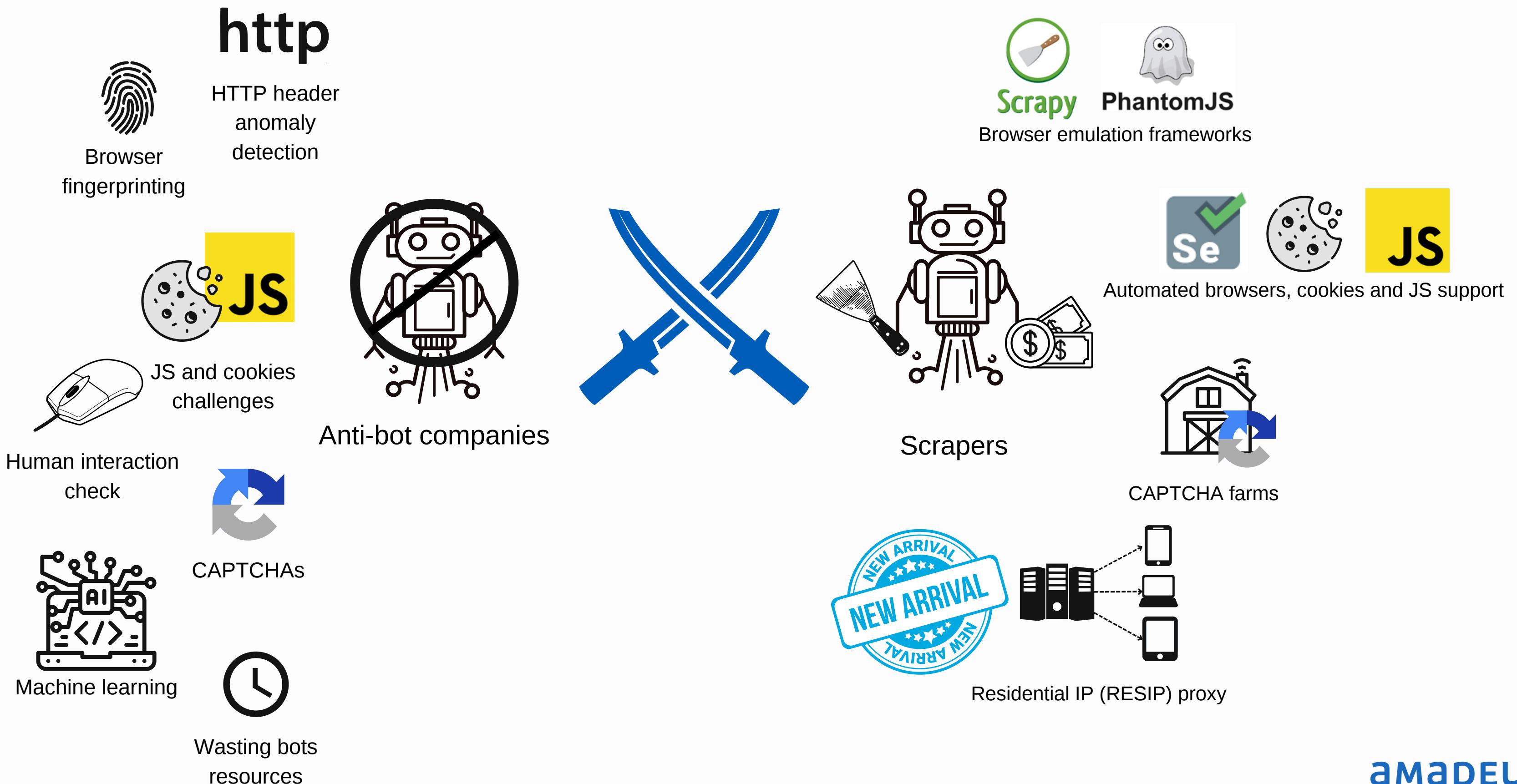
# A persistent battle



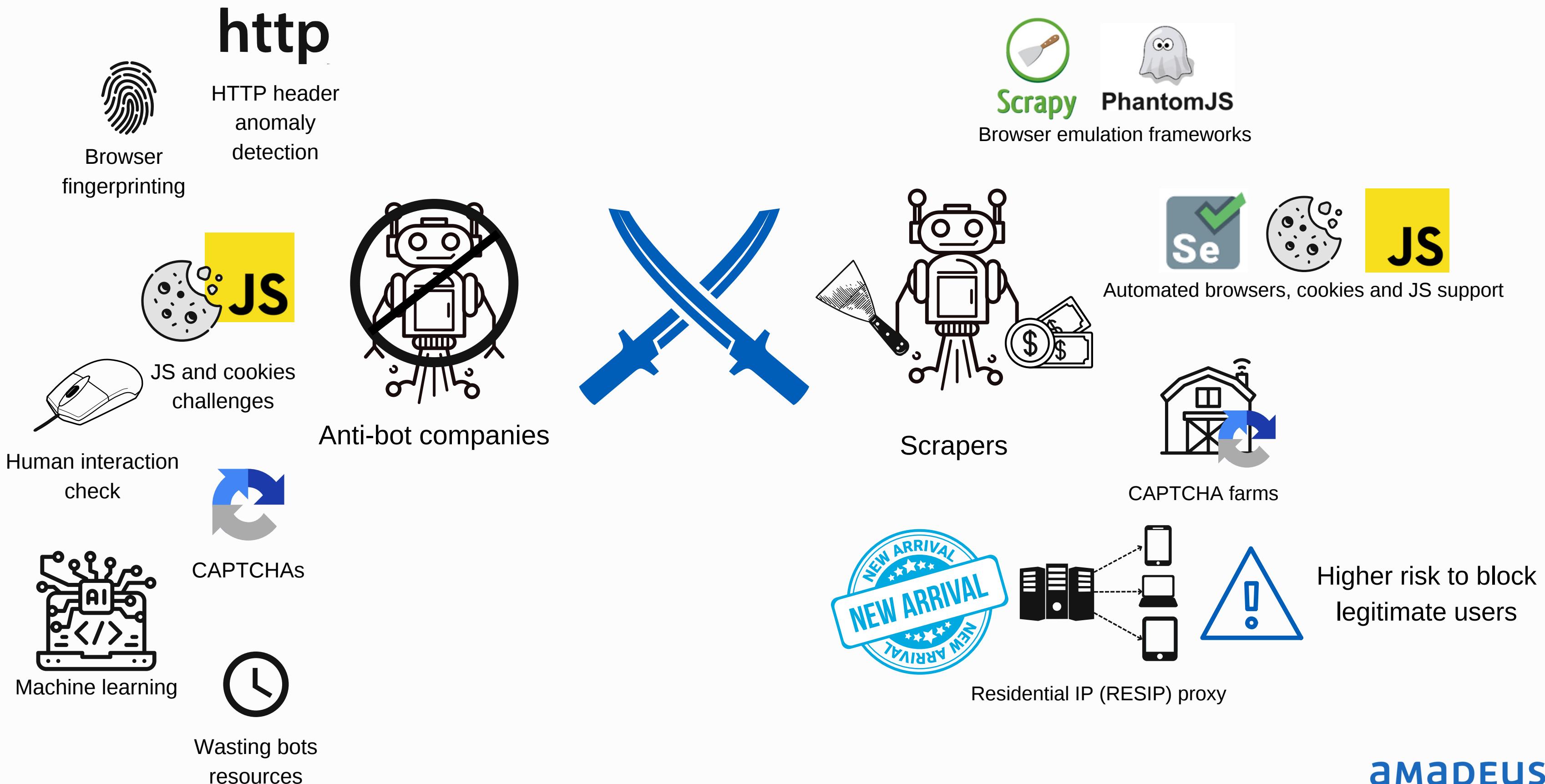
# A persistent battle



# A persistent battle

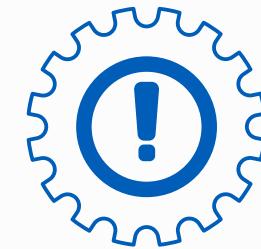


# A persistent battle



# Problems

# Problems



Current mitigation techniques give direct feedback of detection and scrapers can react to them

# Problems



Current mitigation techniques give direct feedback of detection and scrapers can react to them



Scrapers avoid more and more current detection techniques, using RESIP services

# Problems



Current mitigation techniques give direct feedback of detection and scrapers can react to them



Scrapers avoid more and more current detection techniques, using RESIP services



What can **we** do?

# Problems



Current mitigation techniques give direct feedback of detection and scrapers can react to them



Scrapers avoid more and more current detection techniques, using RESIP services

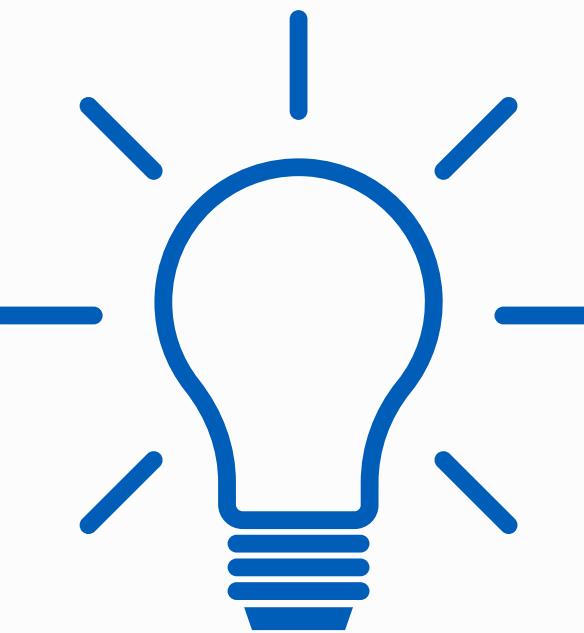


What can **we** do?



2

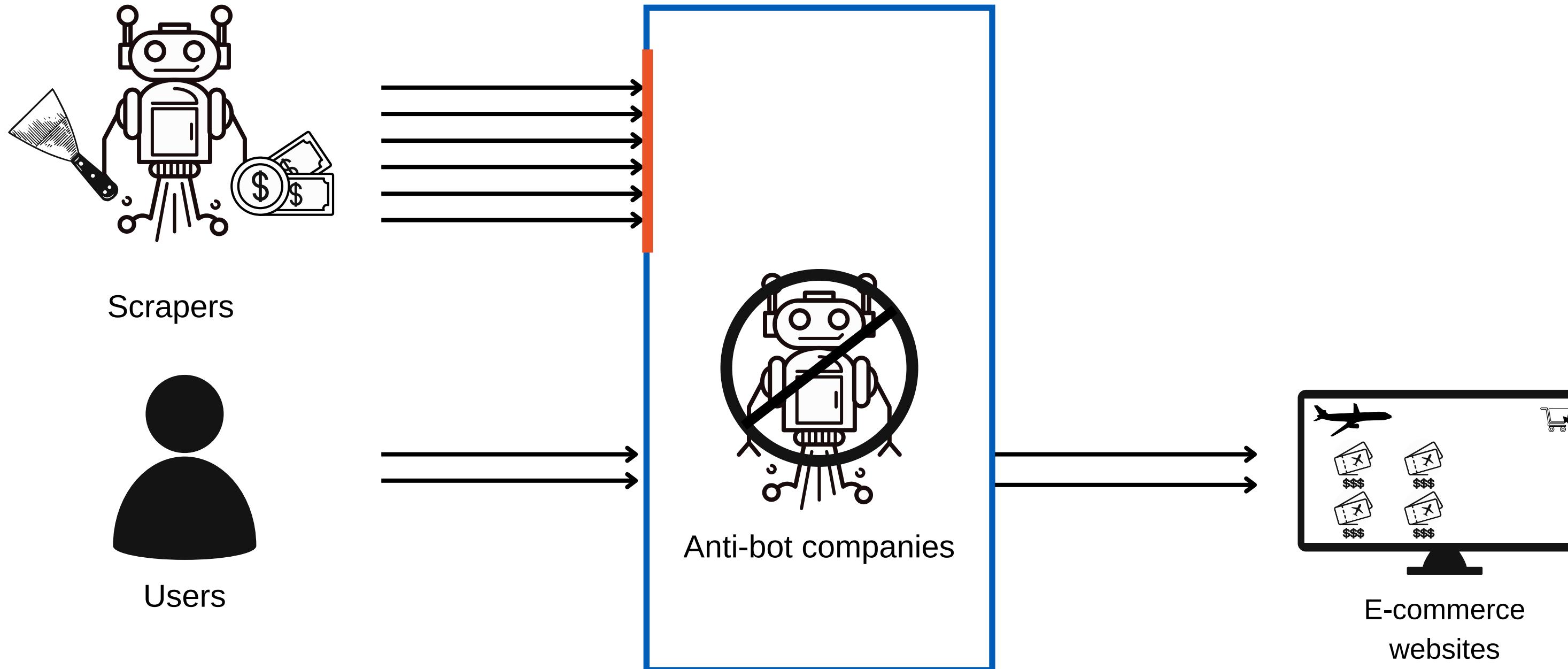
# WebApp Honeypot



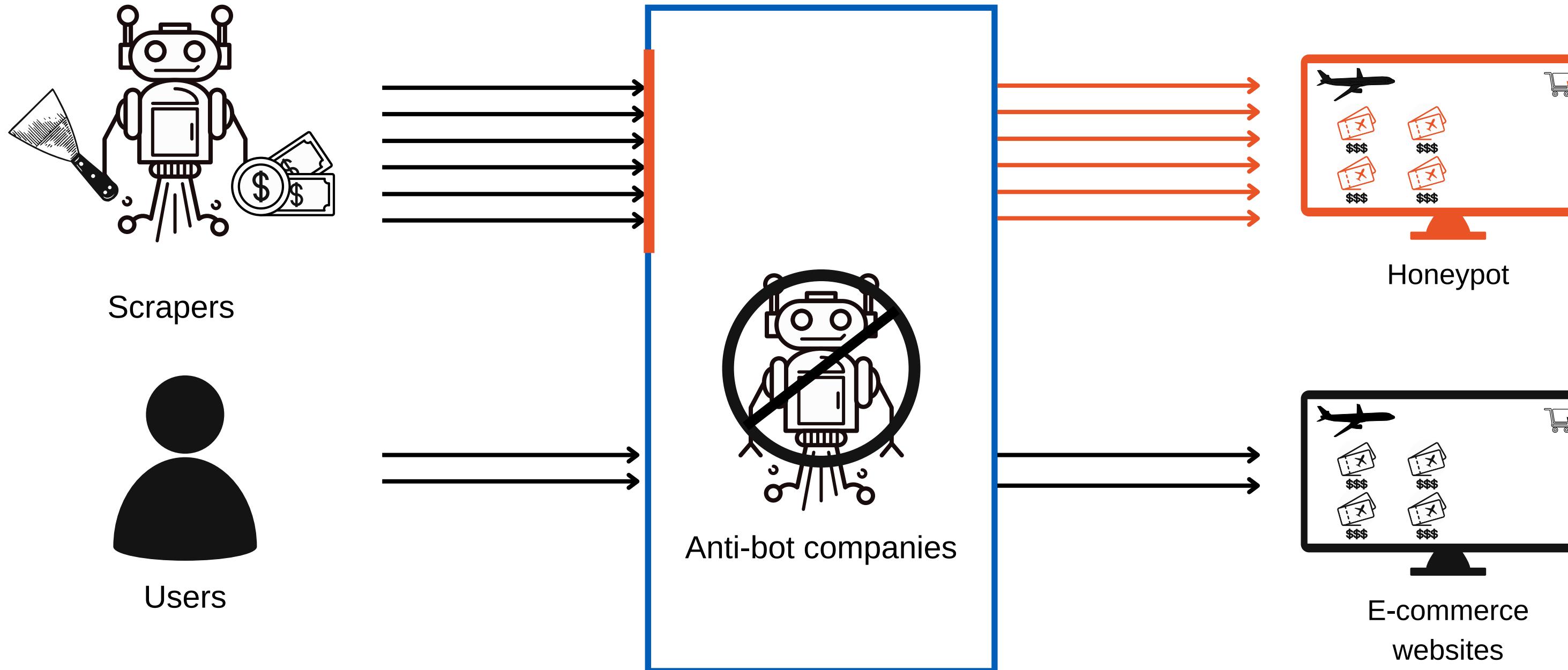
# Idea

Prevent scrapers to know they have been detected  
providing **incorrect but plausible**  
answers at a cheap cost for the provider

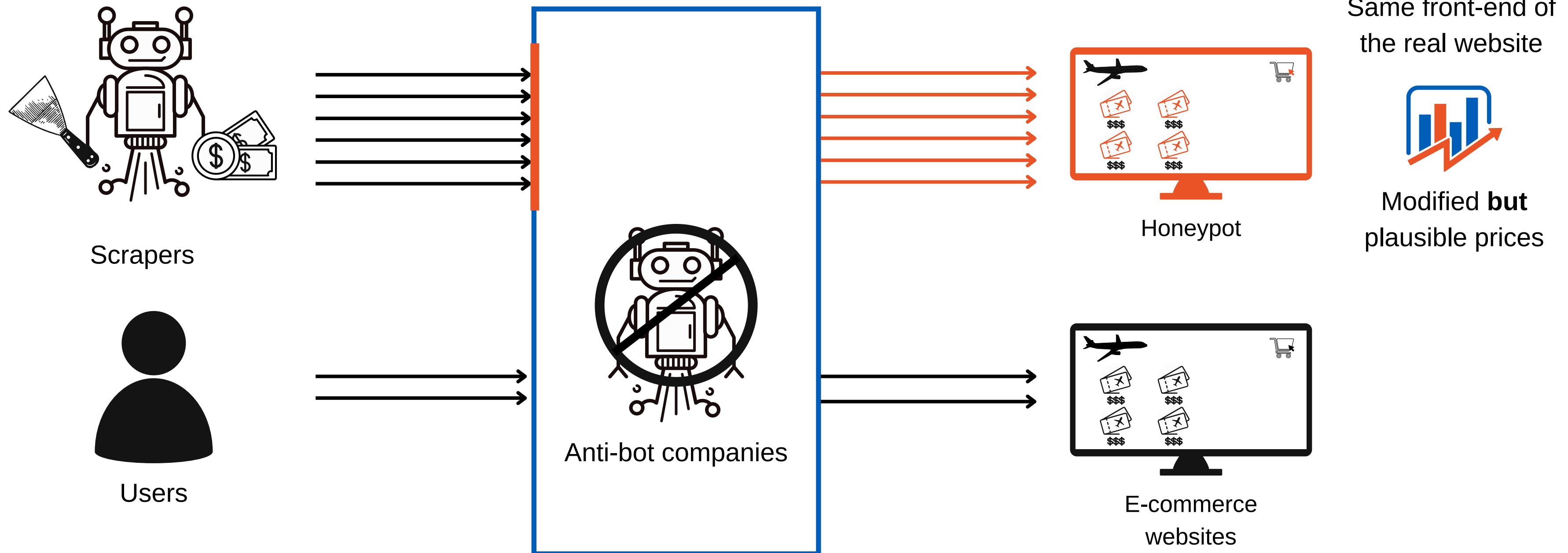
# Inserting the WebApp Honeypot



# Inserting the WebApp Honeypot



# Inserting the WebApp Honeypot



# Proof of concept

- ❖ Collaboration with an airline company

# Proof of concept

- ❖ Collaboration with an airline company
- ❖ Running for 56 days (interruption linked with COVID-19 restrictions on flights)

# Proof of concept

- ❖ Collaboration with an airline company
- ❖ Running for 56 days (interruption linked with COVID-19 restrictions on flights)
- ❖ After 3 days, modification of fares: increase the real price by 5% for 10% of the requests

# Proof of concept

- ❖ Collaboration with an airline company
- ❖ Running for 56 days (interruption linked with COVID-19 restrictions on flights)
- ❖ After 3 days, modification of fares: increase the real price by 5% for 10% of the requests
- ❖ No change of behavior from before and during the PoC

# Proof of concept

- ❖ Collaboration with an airline company
- ❖ Running for 56 days (interruption linked with COVID-19 restrictions on flights)
- ❖ After 3 days, modification of fares: increase the real price by 5% for 10% of the requests
- ❖ No change of behavior from before and during the PoC
- ❖ Scrapers plausibility check not sophisticated enough for small changes

# Advantages

# Challenges

# Advantages

- **Technical:** same front-end w.r.t the original website but different back-ends

# Challenges

# Advantages

- **Technical:** same front-end w.r.t the original website but different back-ends
- **Functional:** zero false positive policy, we can redirect only connections we are 100% sure that are not coming from customers

# Challenges

# Advantages

- **Technical:** same front-end w.r.t the original website but different back-ends
- **Functional:** zero false positive policy, we can redirect only connections we are 100% sure that are not coming from customers
- **Reduce costs:** the WebApp Honeypot consumes CPU to work. We need to reduce them to make it convenient

## Challenges

# Advantages

- **Technical:** same front-end w.r.t the original website but different back-ends
- **Functional:** zero false positive policy, we can redirect only connections we are 100% sure that are not coming from customers
- **Reduce costs:** the WebApp Honeypot consumes CPU to work. We need to reduce them to make it convenient

# Challenges

# Advantages

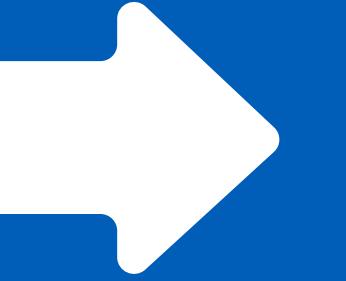
- **Technical:** same front-end w.r.t the original website but different back-ends
- **Functional:** zero false positive policy, we can redirect only connections we are 100% sure that are not coming from customers
- **Reduce costs:** the WebApp Honeypot consumes CPU to work. We need to reduce them to make it convenient
- No direct feedback of detection to attackers
- Attackers database poisoning

# Challenges

# Advantages

- **Technical:** same front-end w.r.t the original website but different back-ends
- **Functional:** zero false positive policy, we can redirect only connections we are 100% sure that are not coming from customers
- **Reduce costs:** the WebApp Honeypot consumes CPU to work. We need to reduce them to make it convenient
  - No direct feedback of detection to attackers
  - Attackers database poisoning
  - Reduce workload for the real website

# Challenges

**NEXT** 

1

Using the WebApp Honeypot as a service, redirecting there persistent bot connections



- 1 Using the WebApp Honeypot as a service, redirecting there persistent bot connections
- 2 Serving cache prices



# Problems



Current mitigation techniques give direct feedback of detection and scrapers can react to them



Scrapers avoid more and more current detection techniques, using RESIP services



What can **we** do?

# Problems



Current mitigation techniques give direct feedback of detection and scrapers can react to them



Scrapers avoid more and more current detection techniques, using RESIP services



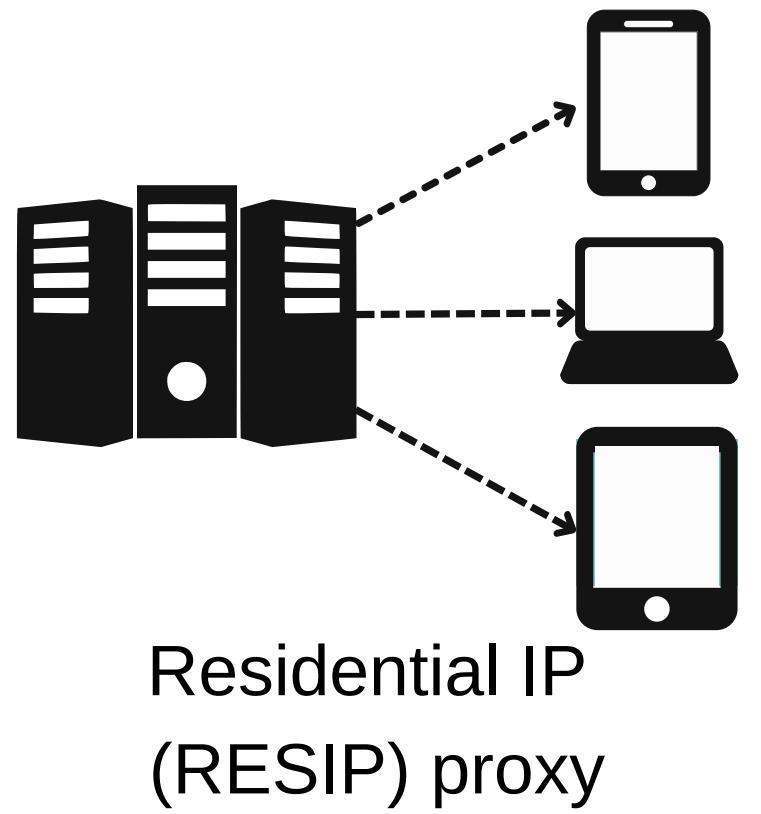
What can **we** do?



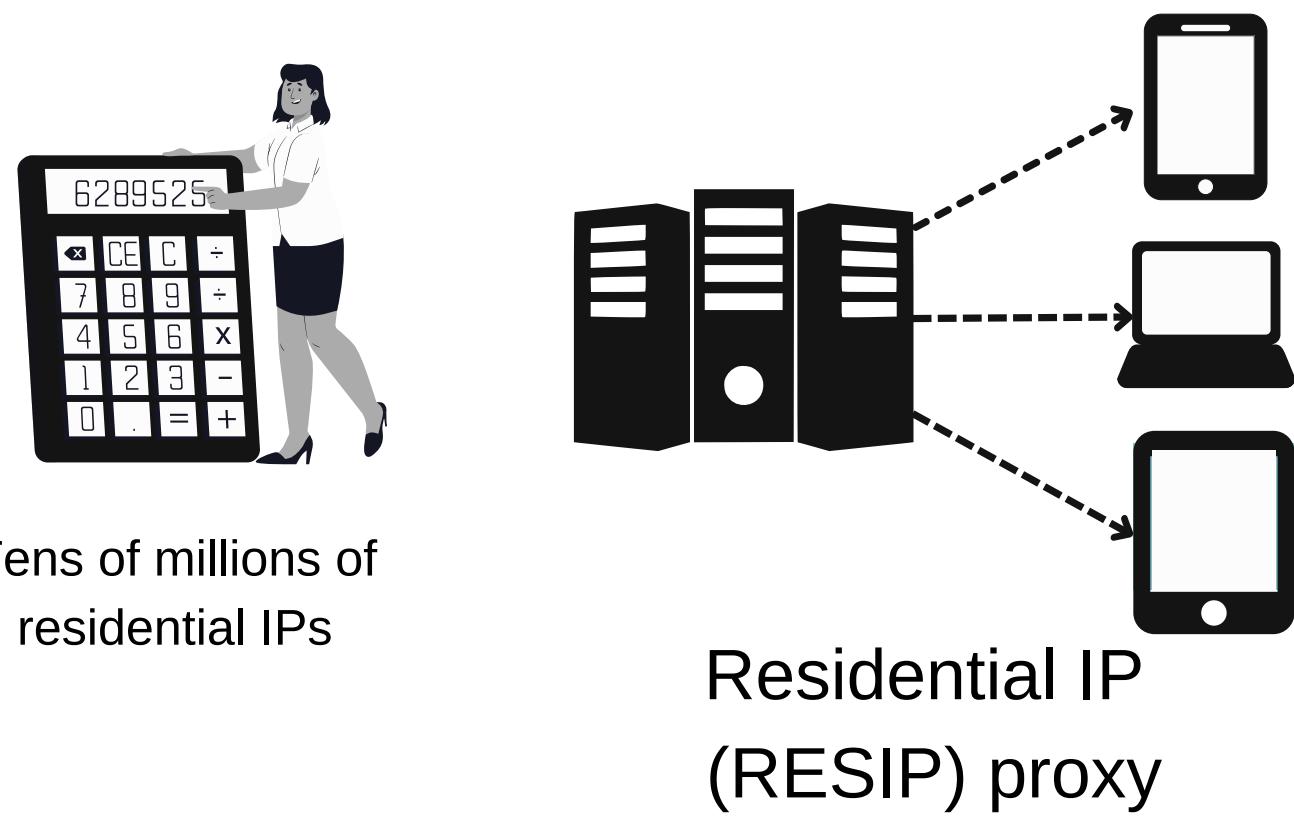
3

# **RESIP detection**

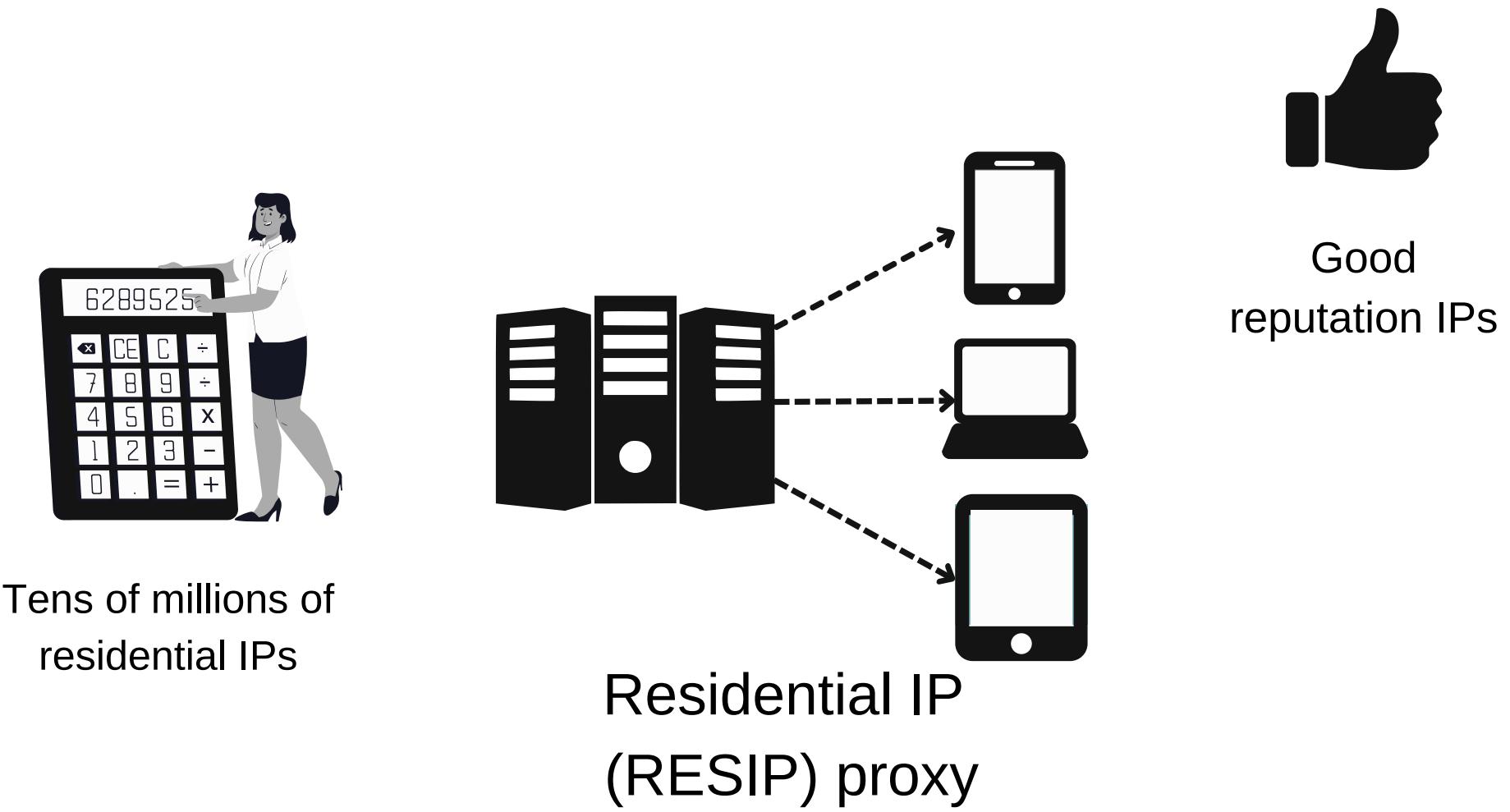
# RESIP providers



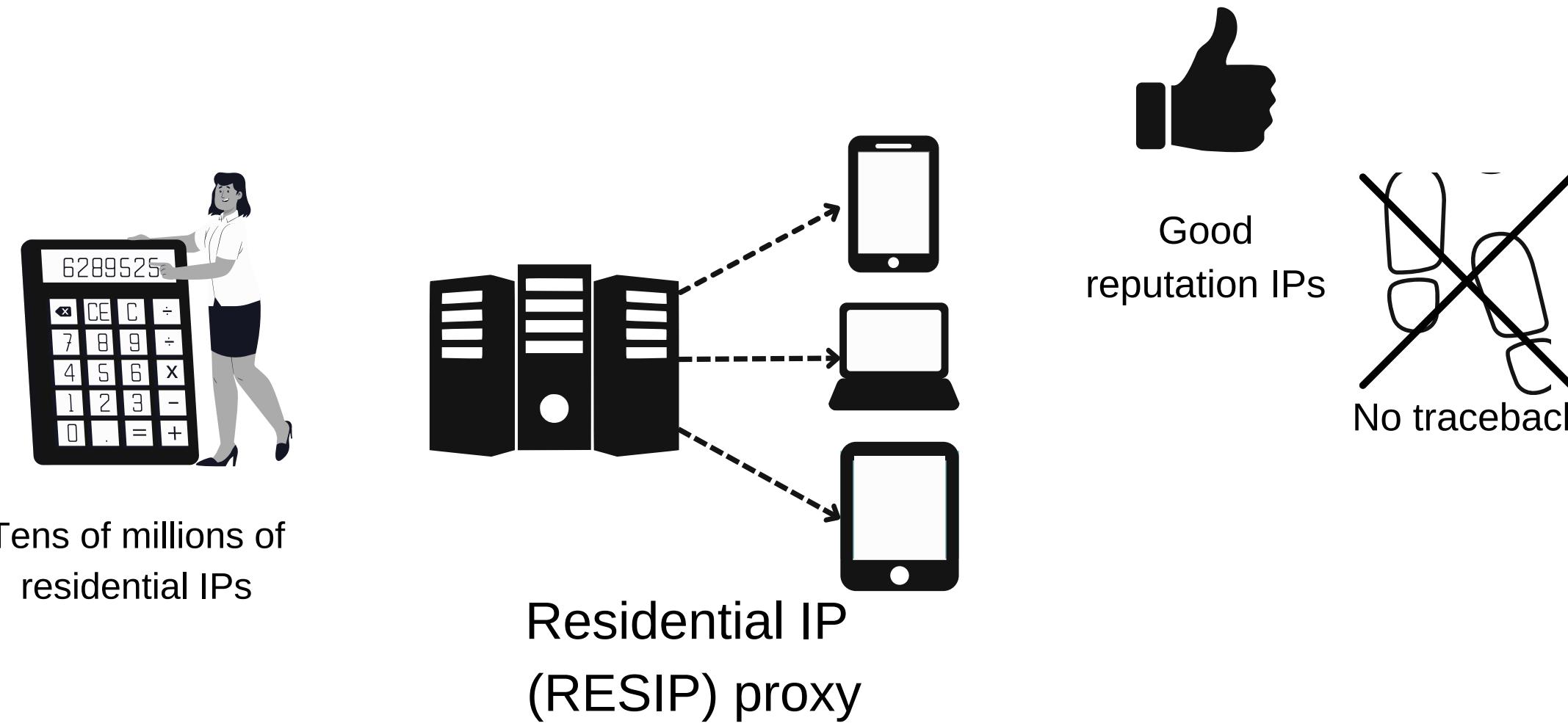
# RESIP providers



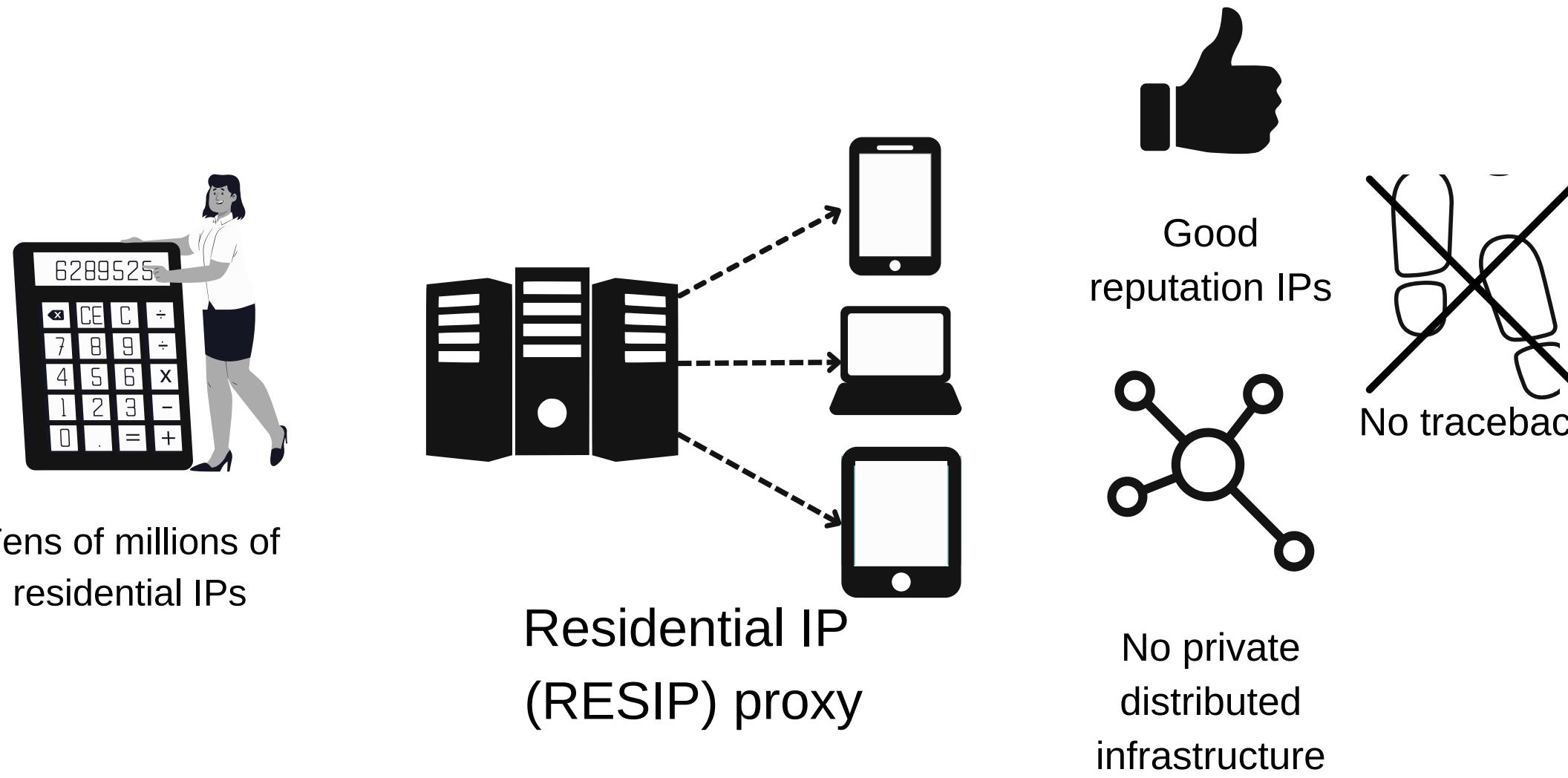
# RESIP providers



# RESIP providers

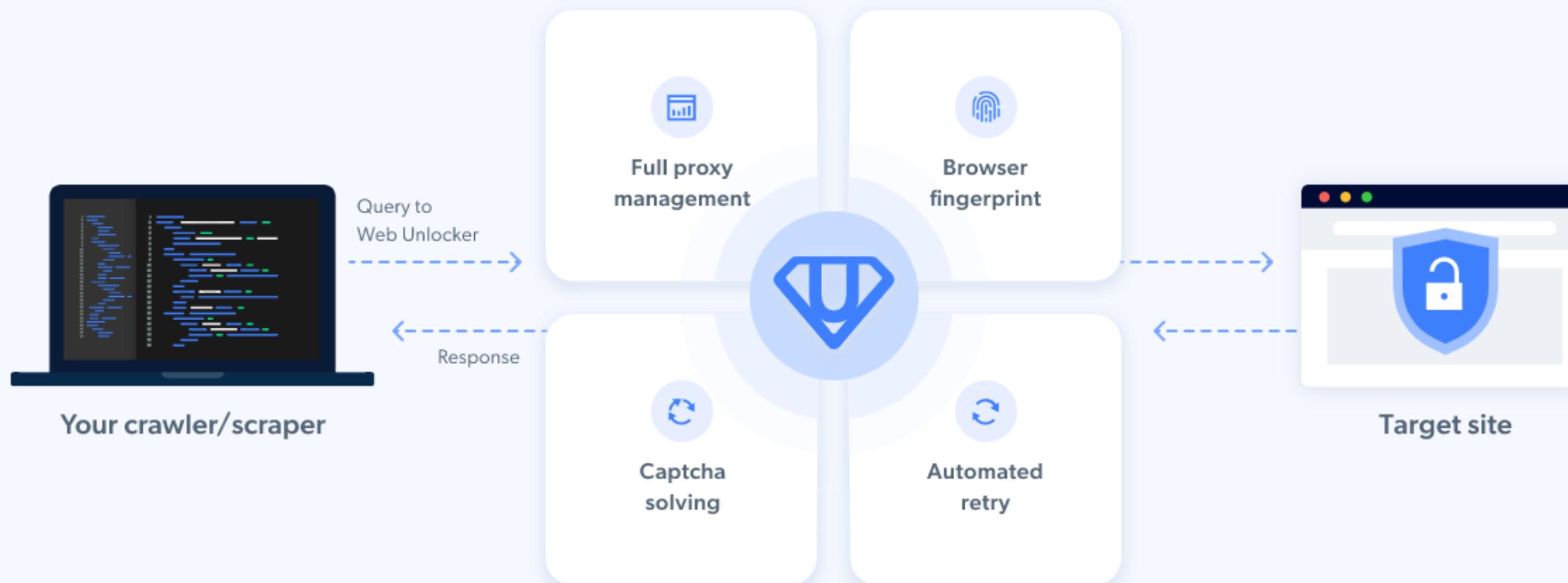


# RESIP providers



[Start my free trial >](#)

## How Web Unlocker optimizes your request's journey



# Automated services!

## Residential IPs represented nearly 30% of bot requests

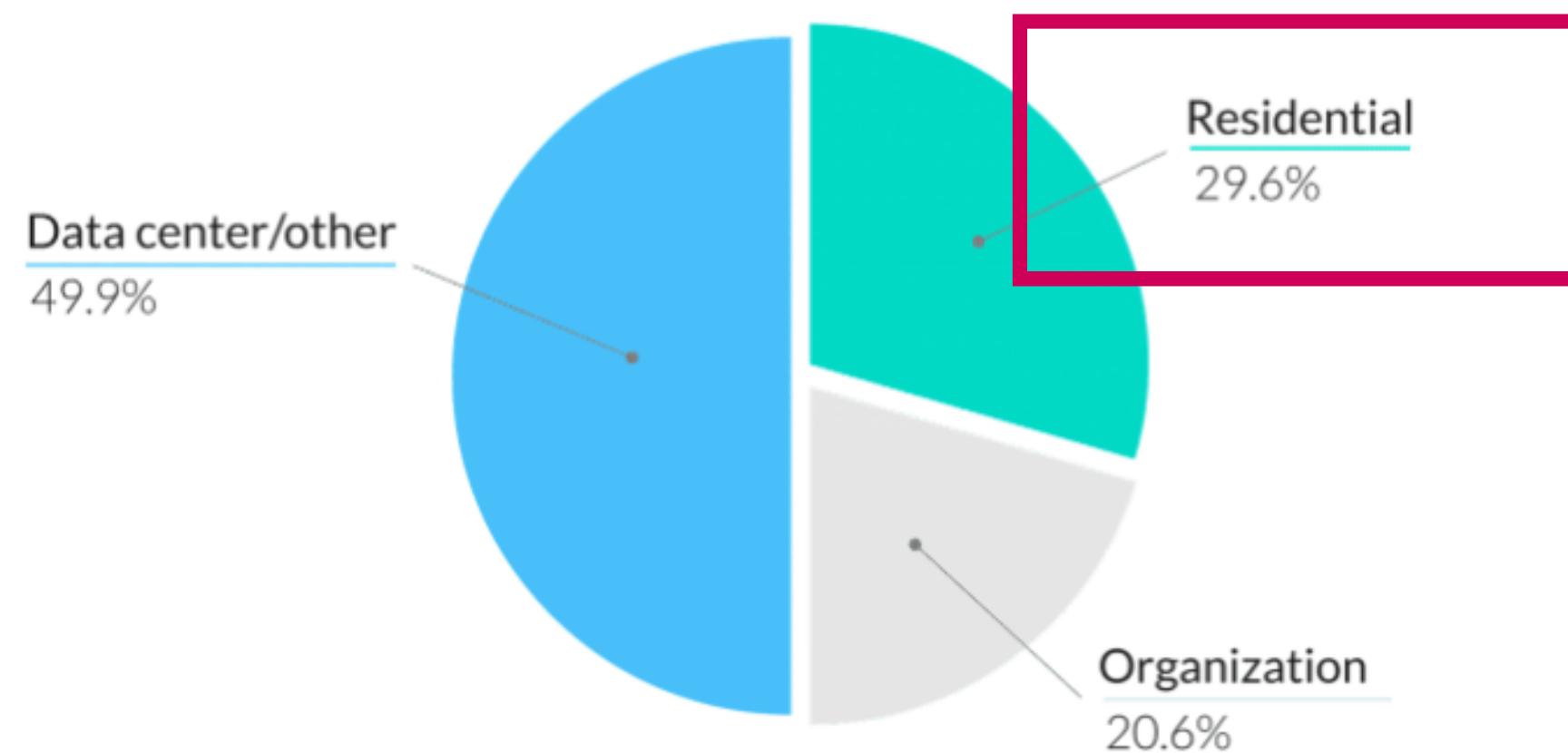
As more and more websites and applications are setting up some form of protection against malicious automated traffic, bot developers are turning to residential IPs to camouflage their bots as legitimate traffic.

While residential IP addresses are more expensive than data center IPs, due to a more limited supply, they can be obtained easily enough through companies such as Geosurf or Luminati that provide **residential IP proxies**.

Out of the billions of bad bot requests we registered during the 2019 end-of-year holiday period, 29.55% were using a residential IP address. This means that nearly **one in three** bad bots requests would pass for human traffic if you were looking at the IP address only.

We also found that 20.55% of bad bots came from an organizational IP address. For the most part, these are probably infected devices that are exploited unbeknownst to the IP address owner. Poorly secured IoT devices, for example, are very popular among bad bot operators.

DECEMBER 16-29, 2019



# Scenario

# Scenario

- ❖ Residential IPs are shared between legitimate customers and scrapers: threat for e-commerce to have **false positives** during detection

# Scenario

- ❖ Residential IPs are shared between legitimate customers and scrapers: threat for e-commerce to have **false positives** during detection
- ❖ There is a need for specific detection when device is used **directly or through RESIP services**

# Scenario

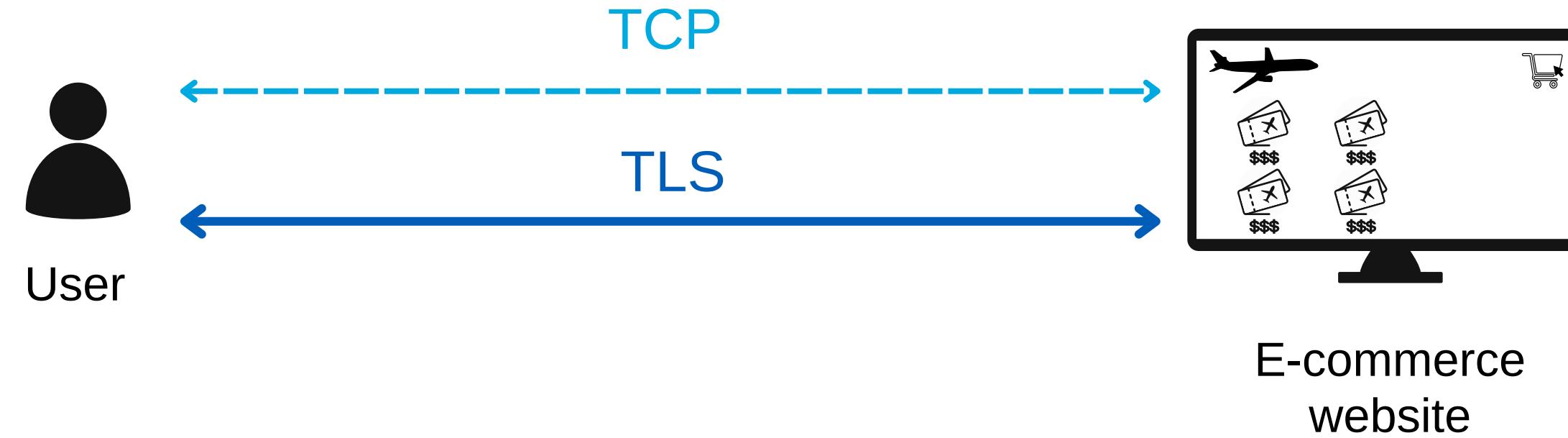
- ❖ Residential IPs are shared between legitimate customers and scrapers: threat for e-commerce to have **false positives** during detection
- ❖ There is a need for specific detection when device is used **directly or through RESIP services**
- ❖ Both types of connection are similar at the application layer **but** present differences at the transport layer

# Direct Connection

## RESIP Connection

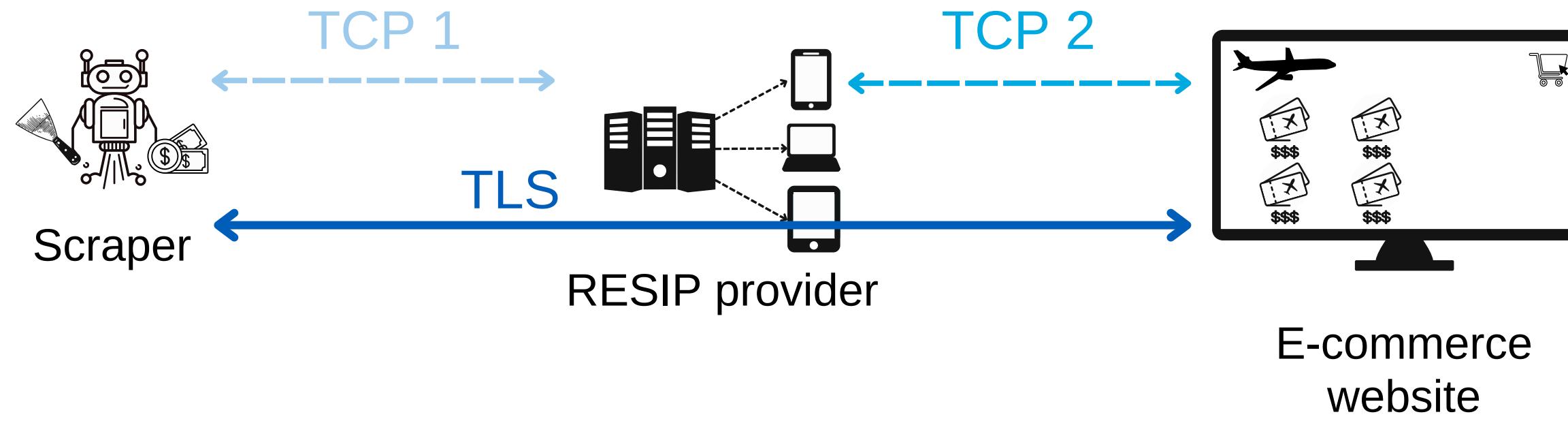
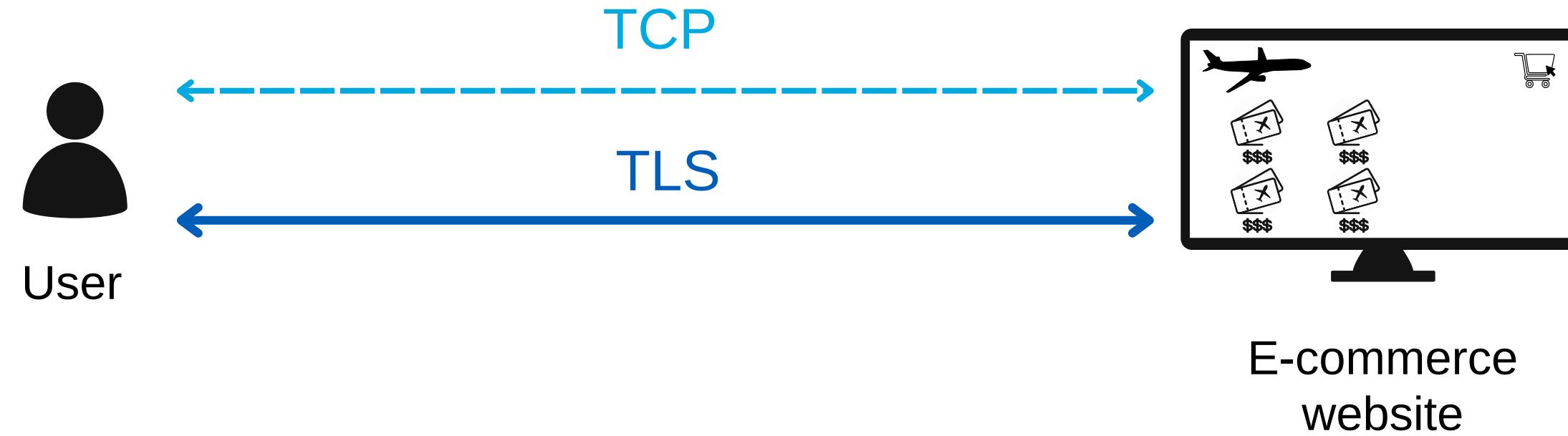
AMADEUS

# Direct Connection



# RESIP Connection

# Direct Connection



# RESIP Connection

# How can we check this?

The Round Trip Time gives an "approximation" of the physical distance between sender and receiver [1]

[1] Landa, R., Clegg, R.G., Araujo, J.T., Mykoniati, E., Griffin, D., Rio, M.: Measuring the Relationships between Internet Geography and RTT. In: 2013 22nd International Conference on Computer Communication and Networks (ICCCN).

# How can we check this?

The Round Trip Time gives an "approximation" of the physical distance between sender and receiver [1]



We can use the RTT among the TCP packets against the one among the TLS ones to see if there is a difference in the setup and spot RESIP connections

[1] Landa, R., Clegg, R.G., Araujo, J.T., Mykoniati, E., Griffin, D., Rio, M.: Measuring the Relationships between Internet Geography and RTT. In: 2013 22nd International Conference on Computer Communication and Networks (ICCCN).

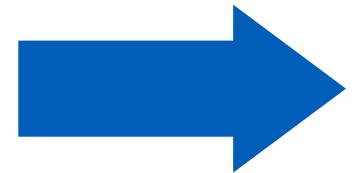
# How can we check this?

# How can we check this?

TLS RTT ~ TCP RTT

# How can we check this?

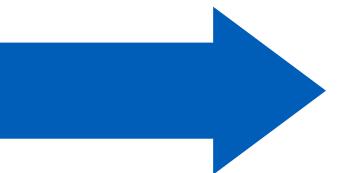
TLS RTT ~ TCP RTT



Direct connection

# How can we check this?

TLS RTT ~ TCP RTT

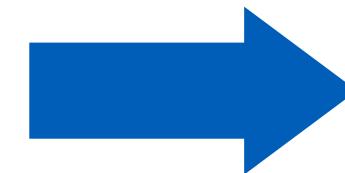


Direct connection

TLS RTT >> TCP RTT

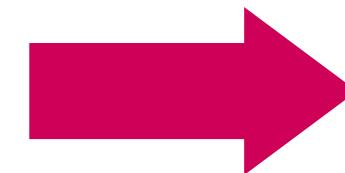
# How can we check this?

TLS RTT ~ TCP RTT



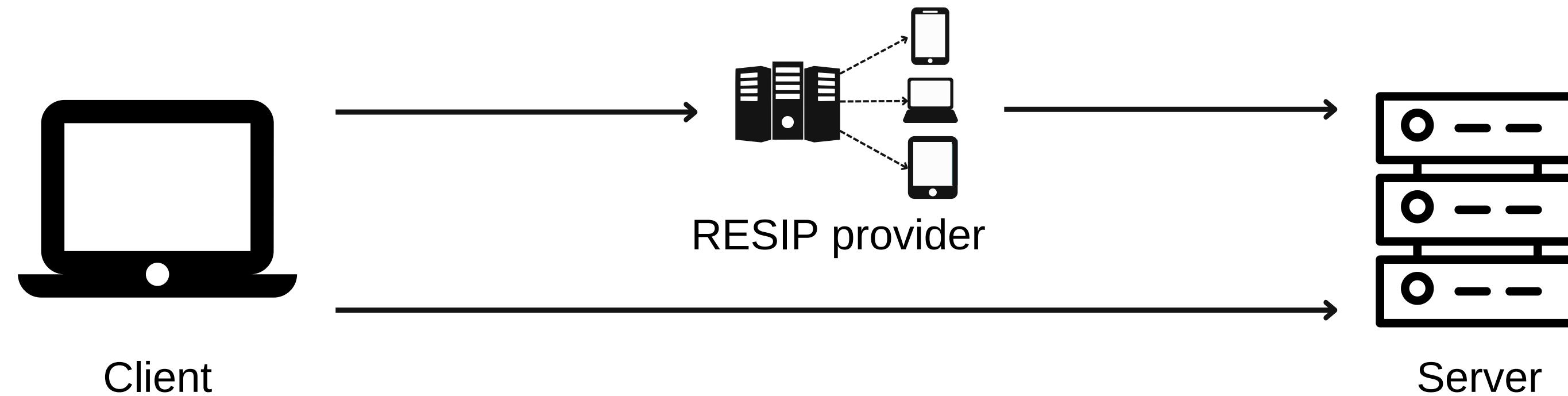
Direct connection

TLS RTT >> TCP RTT

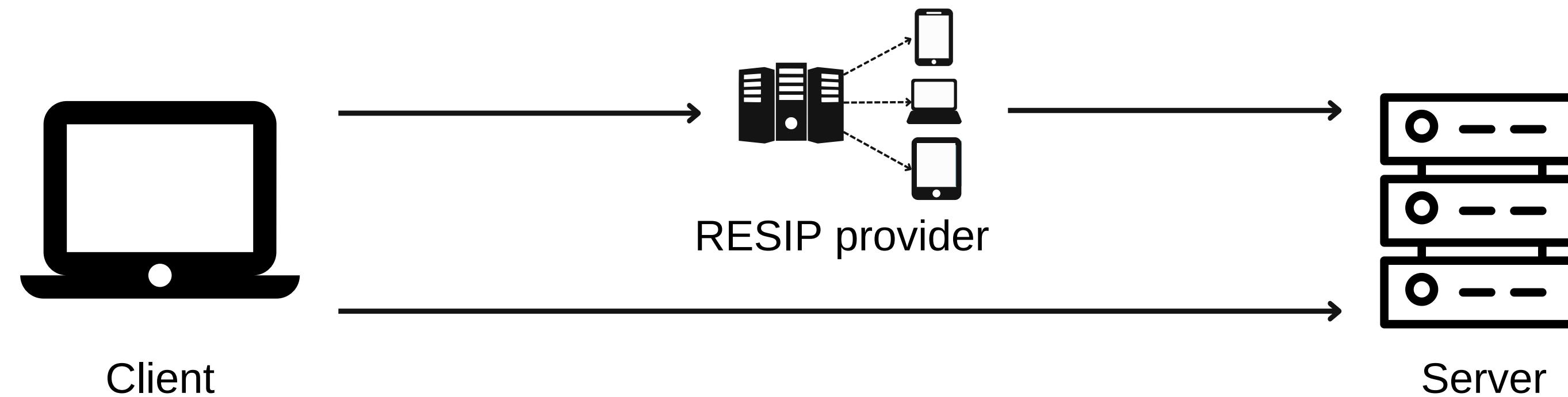


Proxyed connection

# Experiment

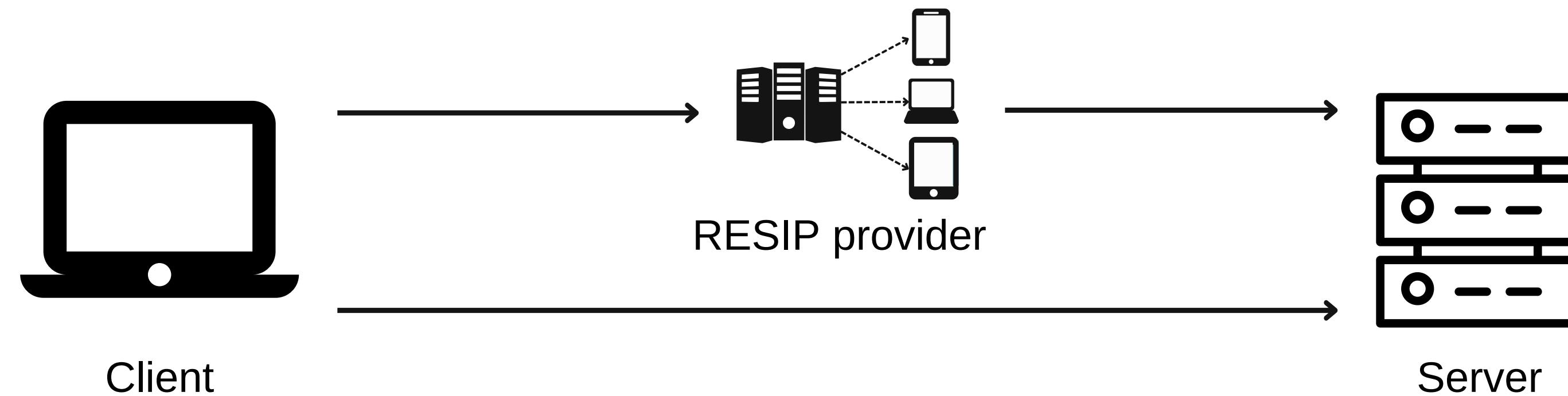


# Experiment



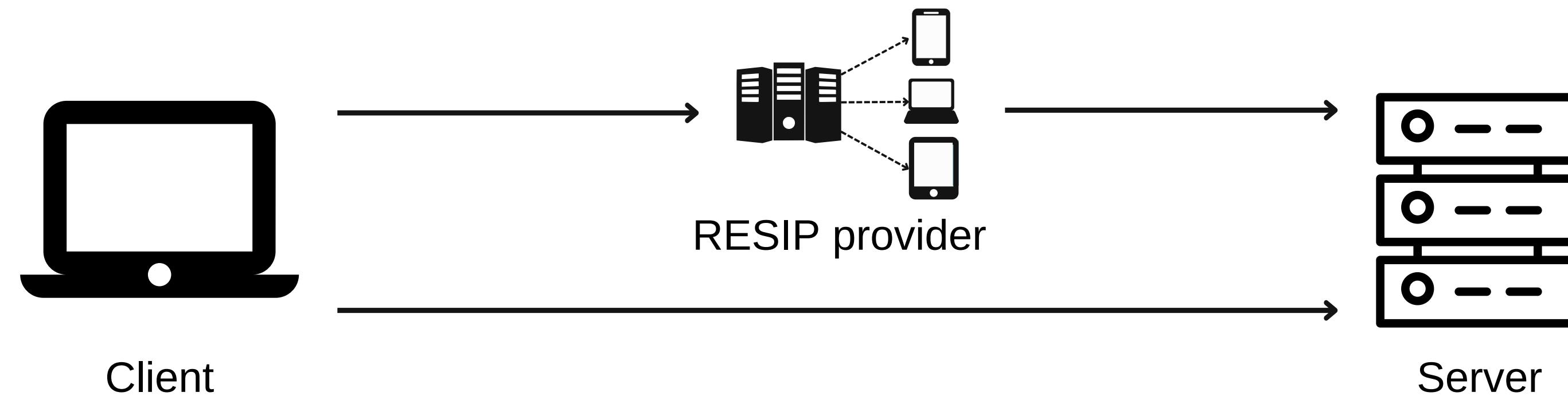
- ❖ 4 RESIP services, 22 client/server machines all over the world

# Experiment



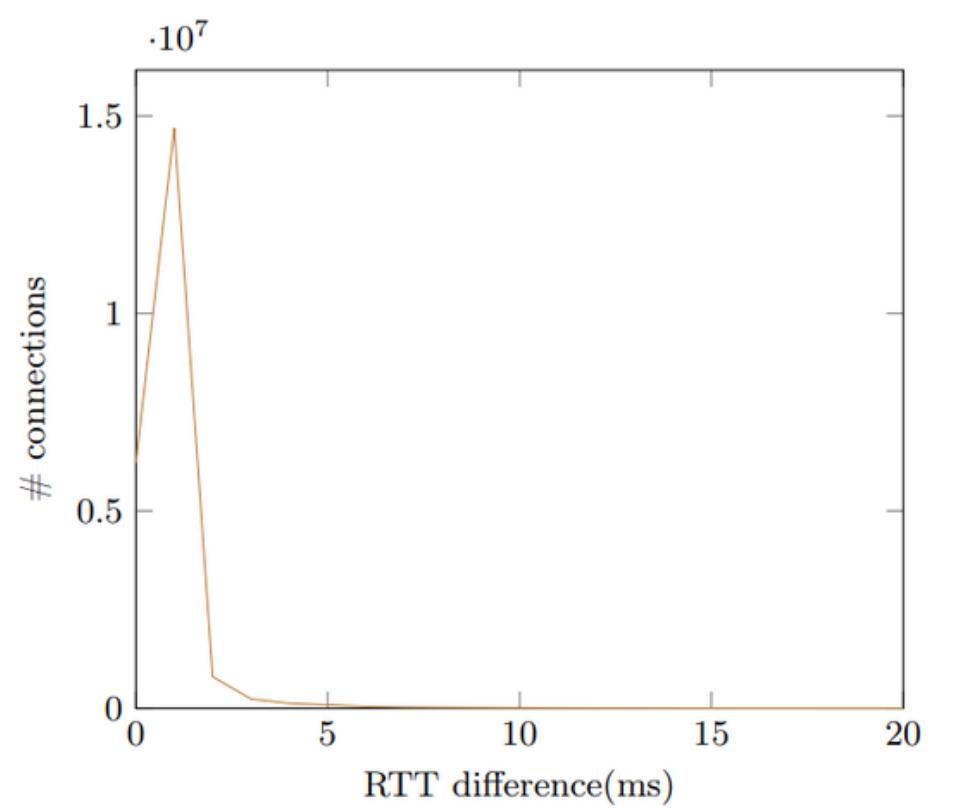
- ❖ 4 RESIP services, 22 client/server machines all over the world
- ❖ TCP and TLS RTT measurement, difference calculation

# Experiment

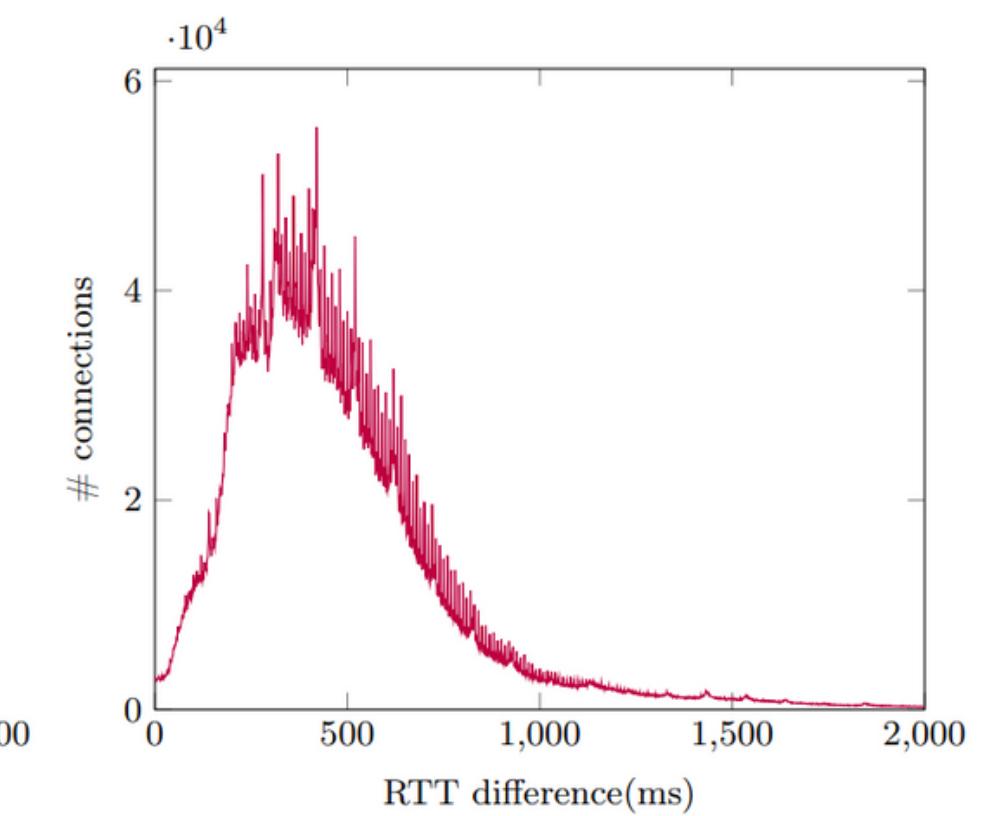
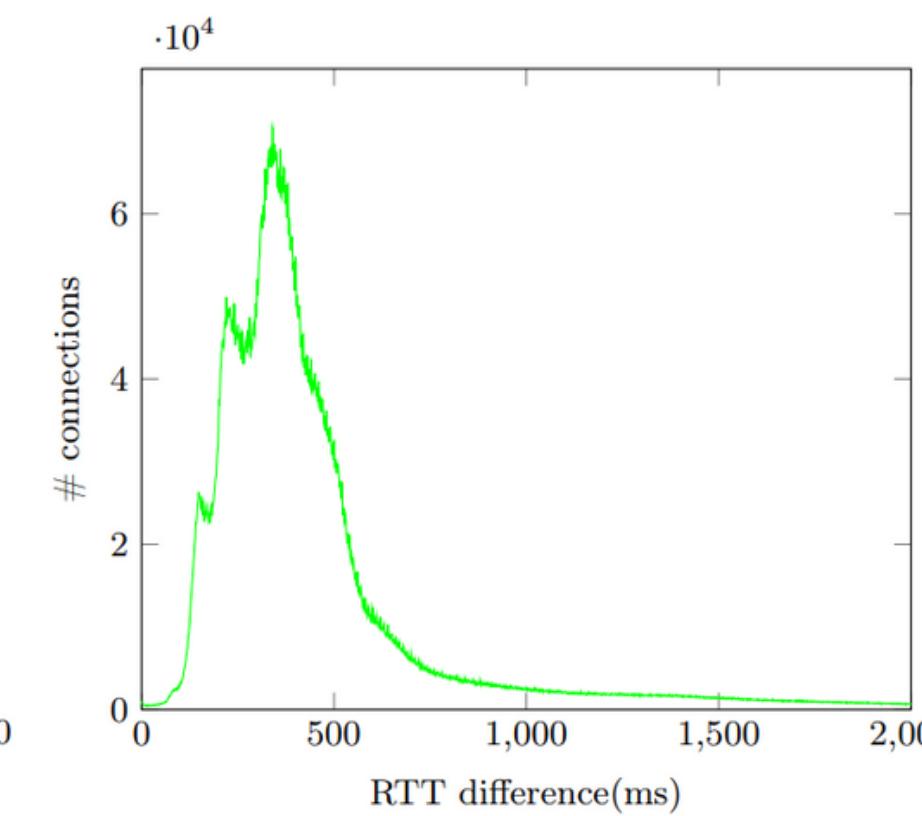
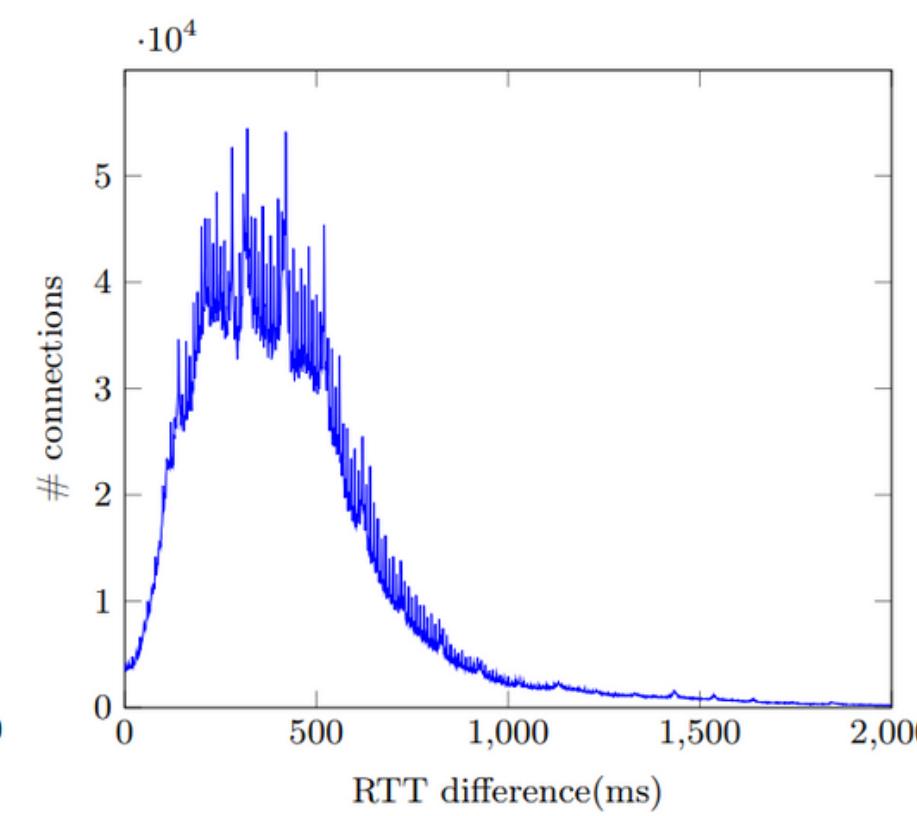
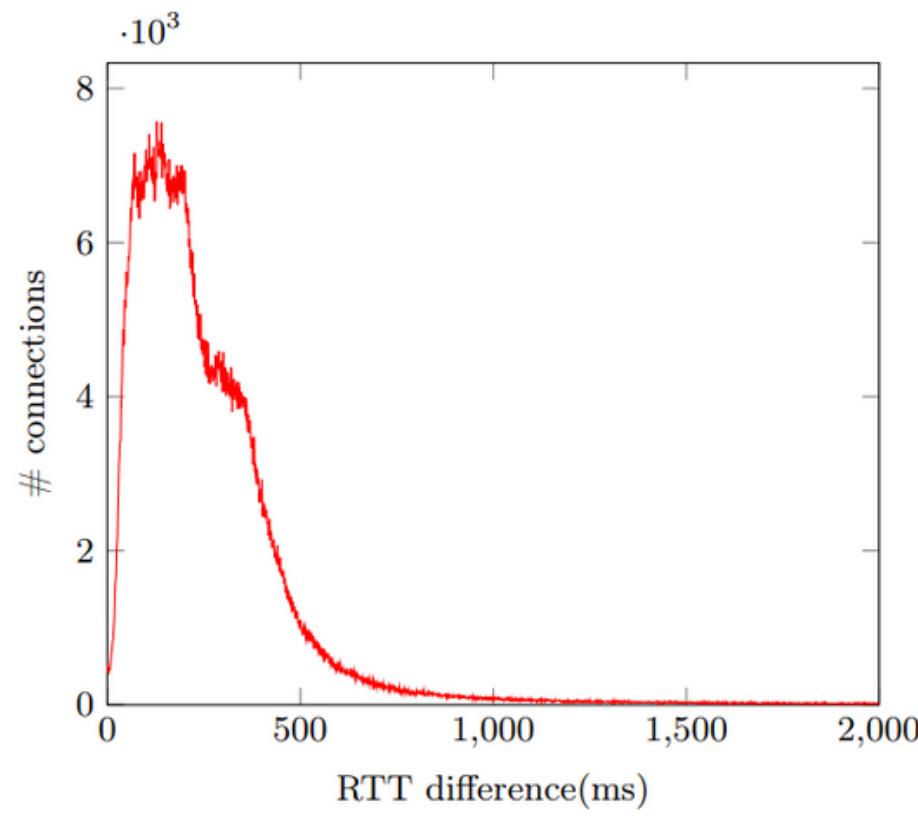
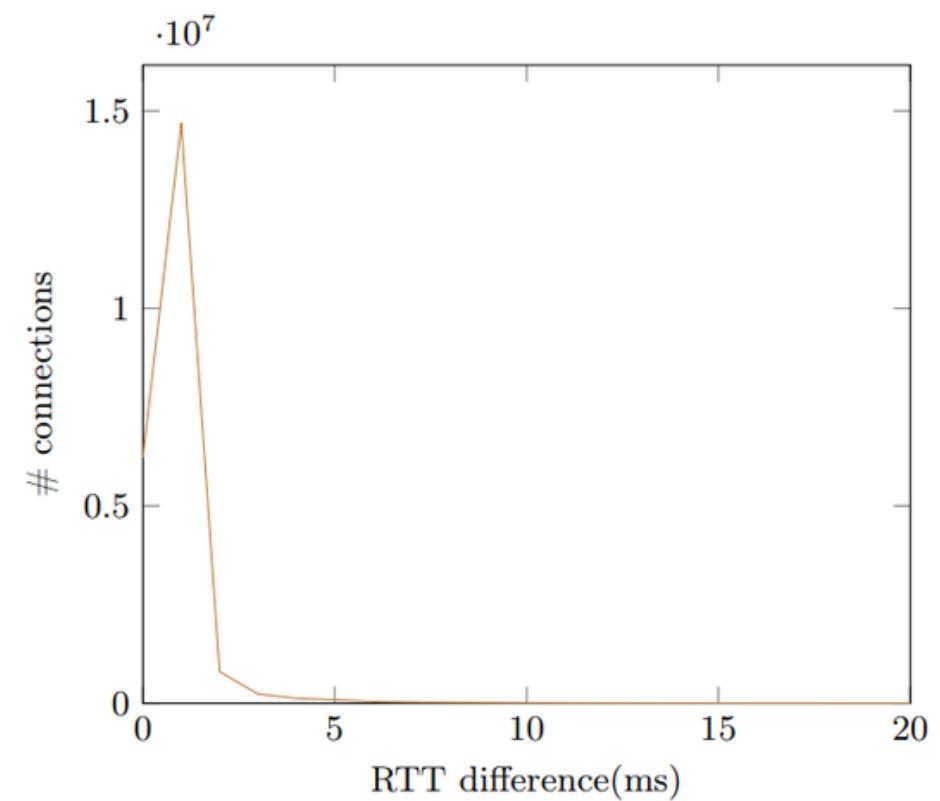


- ❖ 4 RESIP services, 22 client/server machines all over the world
- ❖ TCP and TLS RTT measurement, difference calculation
- ❖ 4 months experiment, 92M+ connections

# Direct Connection



# Direct Connection



# RESIP Connection

AMADEUS

# Outcome

# Outcome

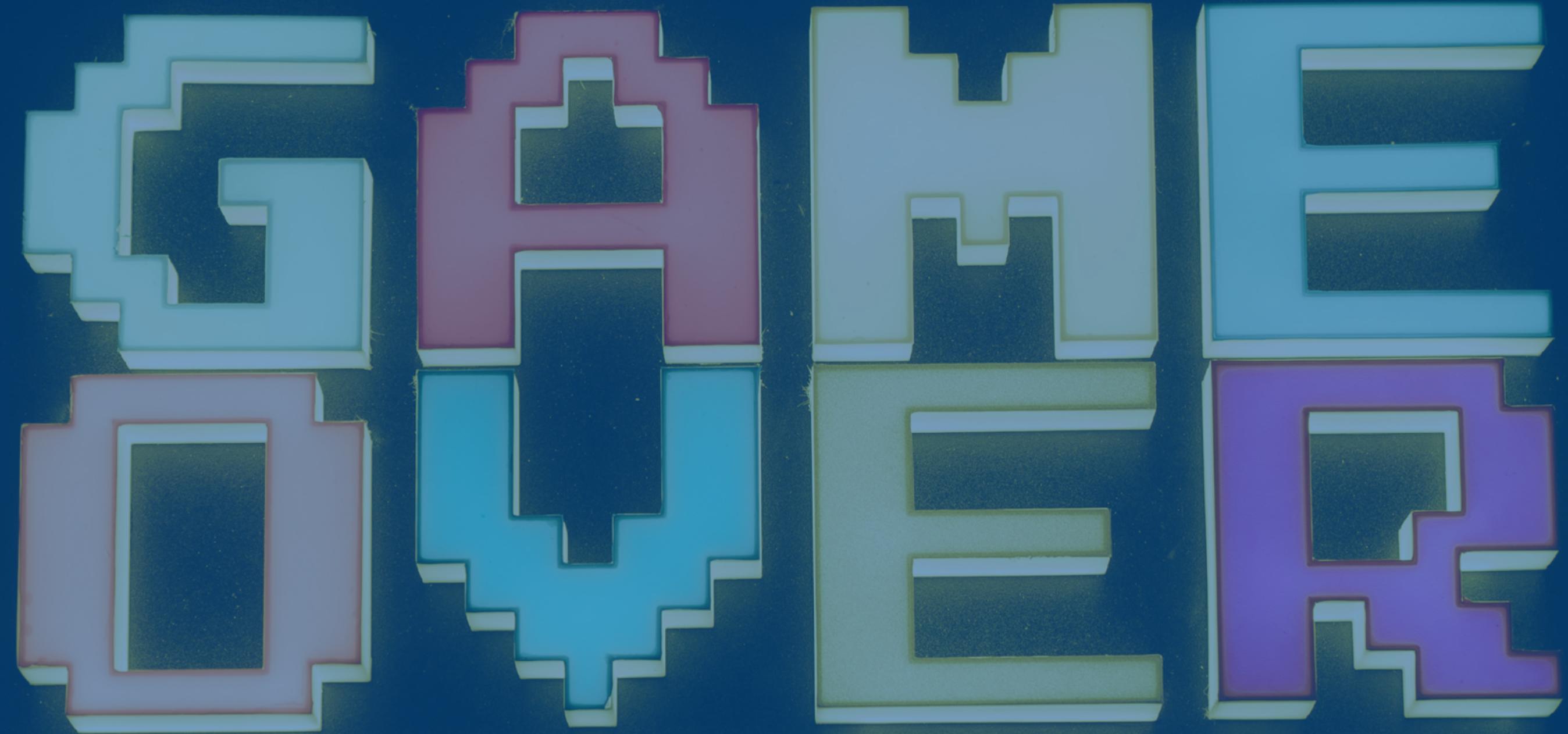
- ★ Promising technique

# Outcome

- ★ Promising technique
- ★ Filed patent submission

# Outcome

- ★ Promising technique
- ★ Filed patent submission
- ★ Next step: test on real-world scraping connections



# What have we talked about today

# What have we talked about today

There is an arms race between e-commerce websites and scraping bots

# What have we talked about today

There is an **arms race** between e-commerce websites and scraping bots

Scrapers are becoming more and more sophisticated and we need **new technologies** for detection and mitigation to compete against them

# What have we talked about today

There is an **arms race** between e-commerce websites and scraping bots

Scrapers are becoming more and more sophisticated and we need **new technologies** for detection and mitigation to compete against them

**Detection** can be improved thanks to a specific **RESIP detection** method based on the comparison of TLS and TCP RTT

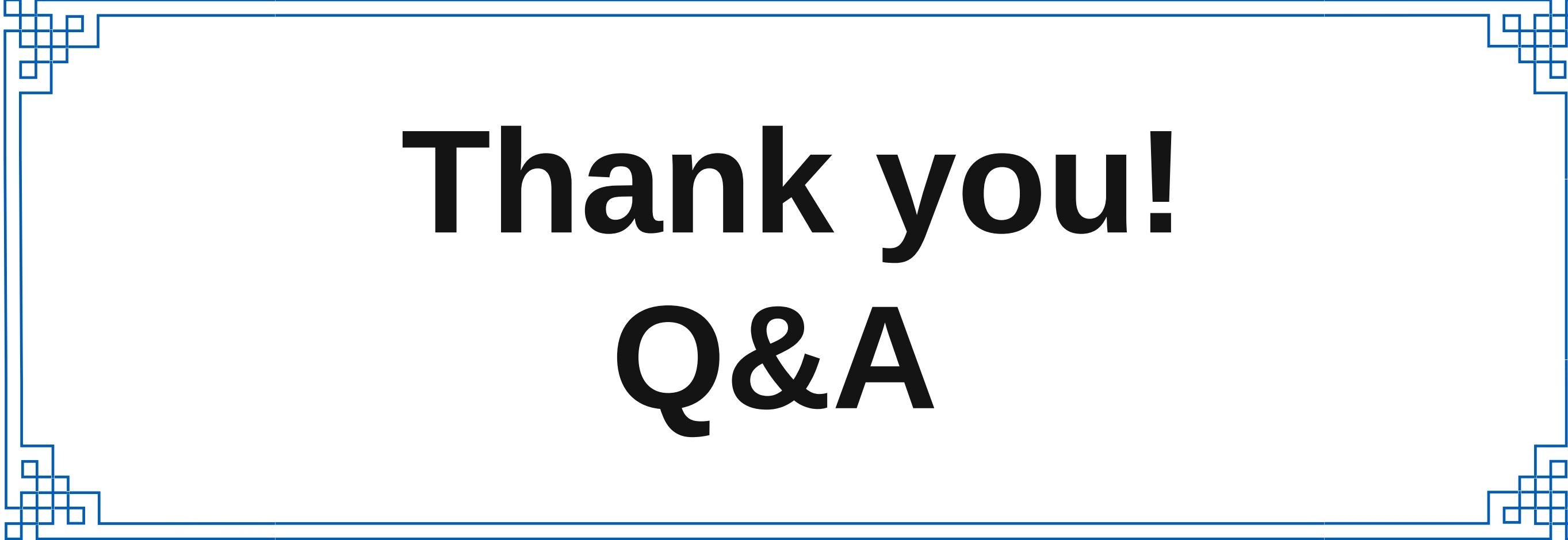
# What have we talked about today

There is an [arms race](#) between e-commerce websites and scraping bots

Scrapers are becoming more and more sophisticated and we need [new technologies](#) for detection and mitigation to compete against them

[Detection](#) can be improved thanks to a specific [RESIP detection](#) method based on the comparison of TLS and TCP RTT

[Mitigation](#) can be improved implementing the [WebApp Honeypot](#) which enables to lure attackers into believing they passed by undetected while receiving incorrect data



# Thank you!

## Q&A

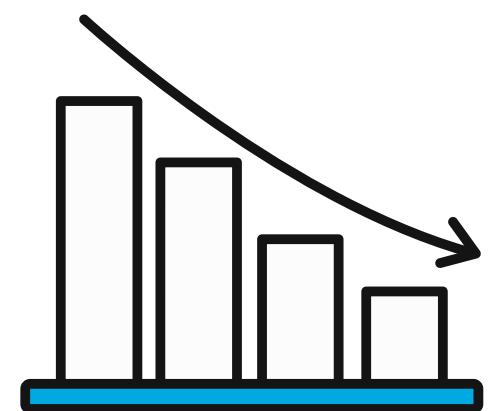
More questions? [elisa.chiapponi@amadeus.com](mailto:elisa.chiapponi@amadeus.com)

Our works:

- Chiapponi et al. (2022). "**An industrial perspective on web scraping characteristics and open issues**" in 52nd Annual IEEE/IFIP DSN 2022 - Industry Track.
- Chiapponi et al. (2021). "**Scraping Airlines Bots: Insights Obtained Studying Honeypot Data**" in International Journal of Cyber Forensics and Advanced Threat Investigations (CFATI).
- Chiapponi et al. (2022) "**BADPASS: Bots taking ADVantage of Proxy AS a Service**" in The 17th International Conference on Information Security Practice and Experience (ISPEC 2022).



Competitors  
monitoring



Statistics  
modification



Content  
reselling

# Why do they **scrape**?

# Scrapers vs aMADEUS: why?



Competitive  
intelligence  
companies

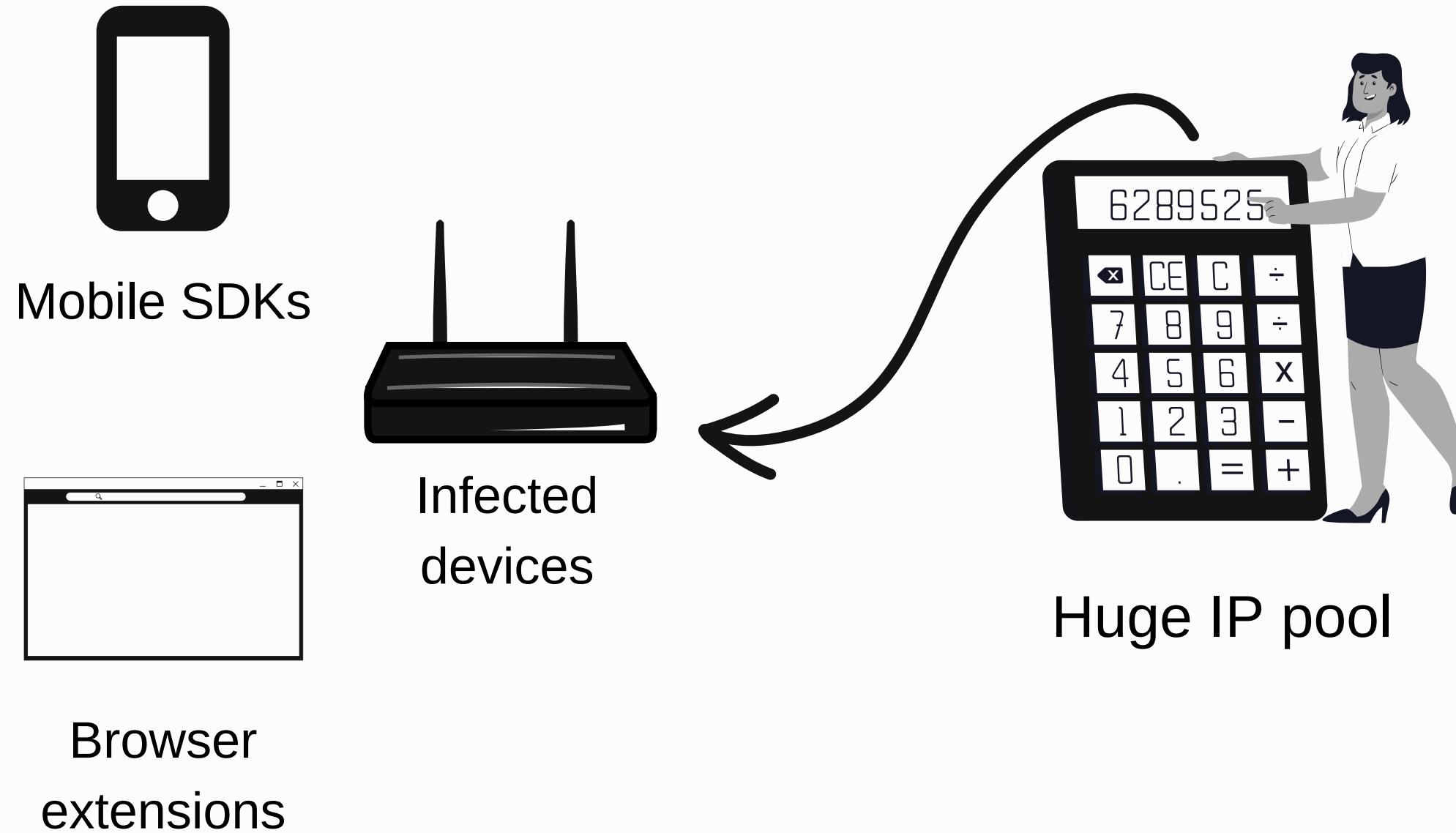


Aggregators



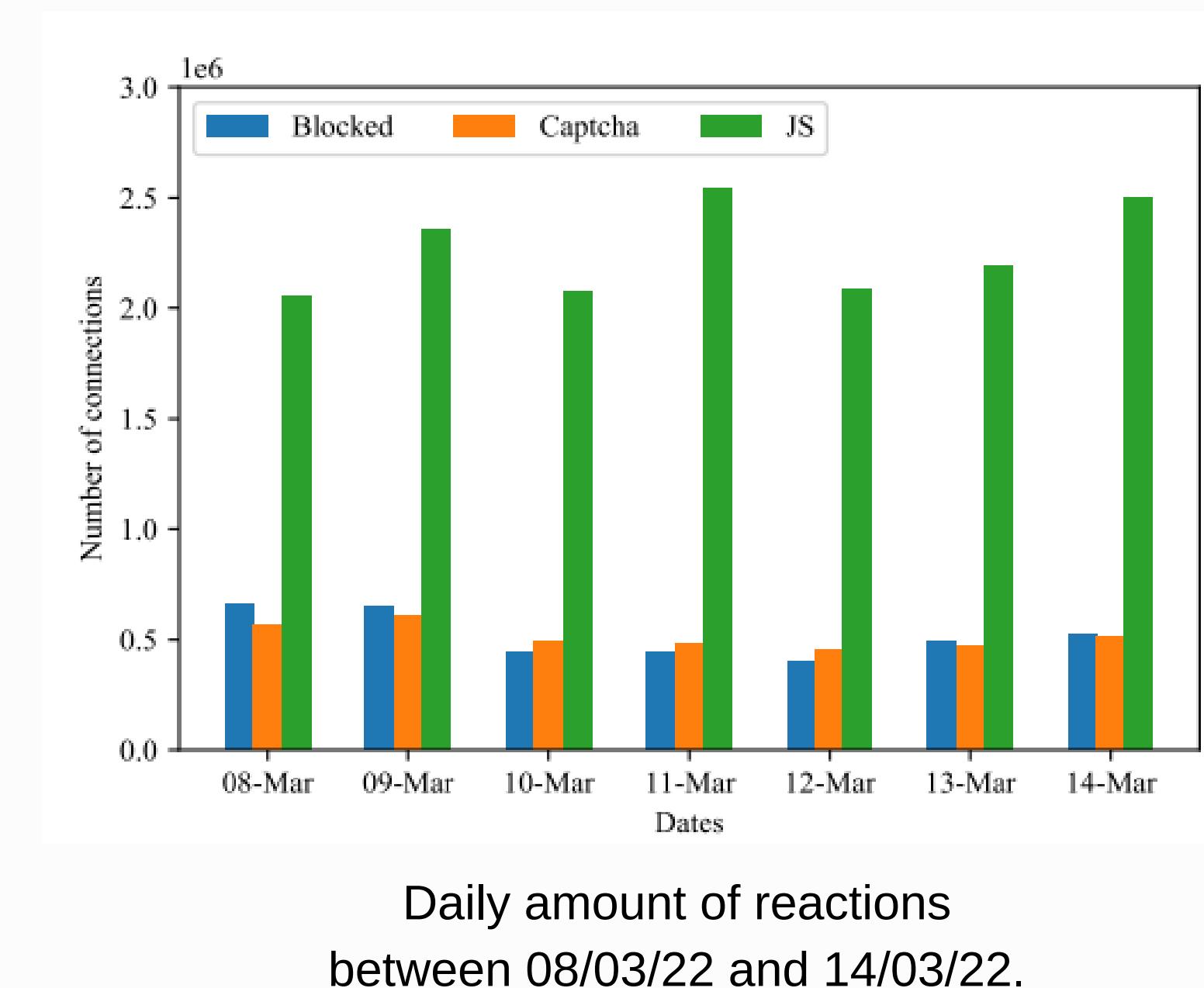
Online travel  
agencies

# RESIP devices

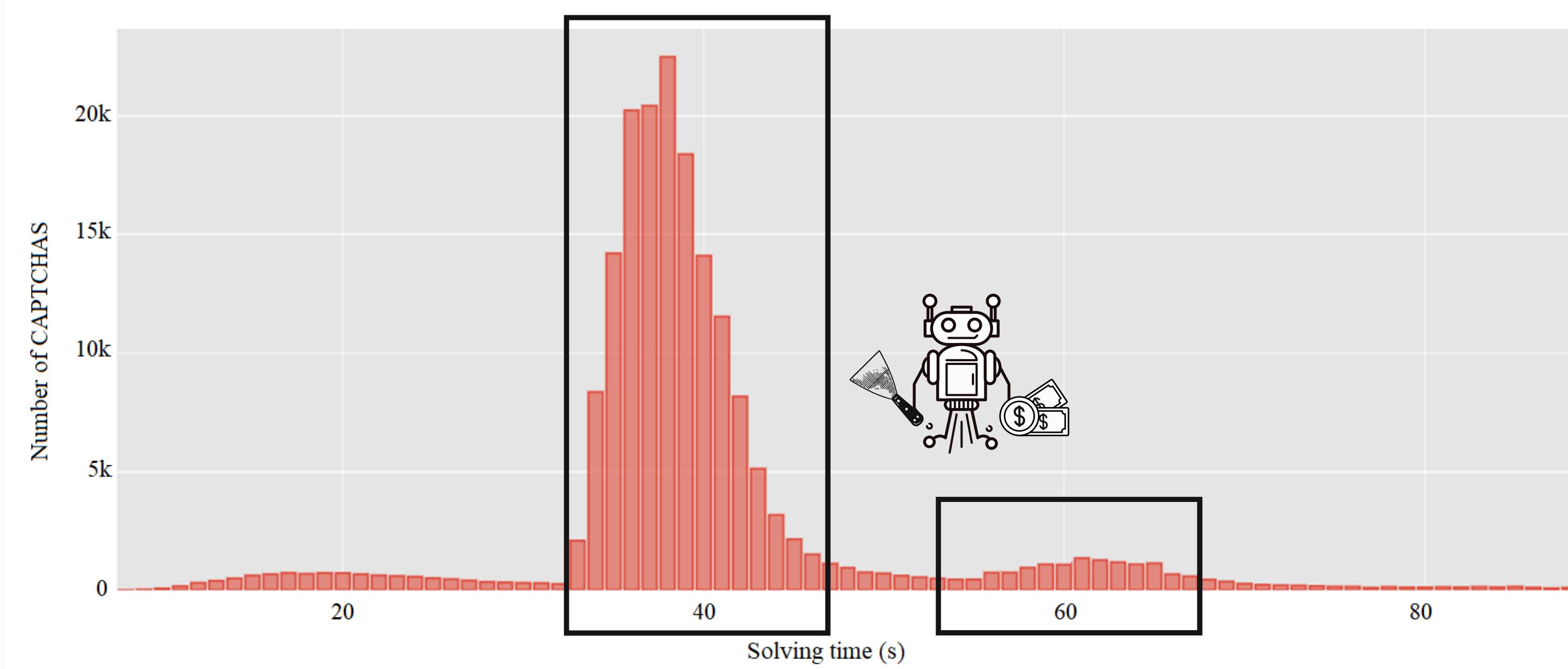


# Scrapers vs aMADEUS: how much?

- Every month, anti-bot rules triggered by **140 million** requests
- **41%** of the attempted connections detected as bots (February 2022)
- **Constant** bot traffic
- Bot reaction to countermeasures: from days (past years) to **hours** (now)



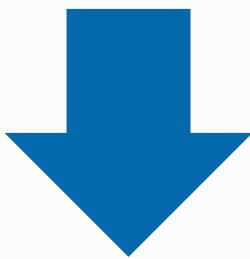
# CAPTCHA solving time (2018)



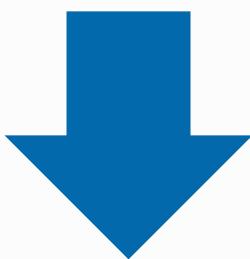
# RESIP activities in amadeus

Residential IPs detected as  
bots in 30 days: **12%**

**Goal:** reducing false positives



Total RESIP traffic is a  
much **larger** portion



**Wide** usage of RESIP

# 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?



# 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