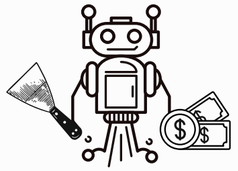
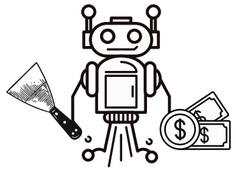
An industrial perspective on web scraping characteristics and open issues

Elisa Chiapponi, Marc Dacier, Olivier Thonnard,
Mohamed Fangar, Mattias Mattsson, Vincent Rigal
elisa.chiapponi@eurecom.fr, marc.dacier@kaust.edu.sa,
{olivier.thonnard, mohamed.fangar, mattias.mattsson, vincent.rigal}@amadeus.com

29th July 2022



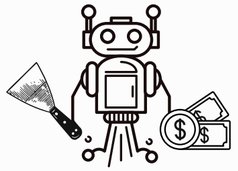
Agenda



Agenda

1

The players



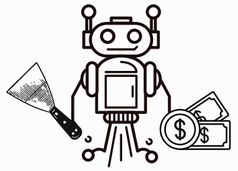
Agenda

1

The players

2

Real-world case study

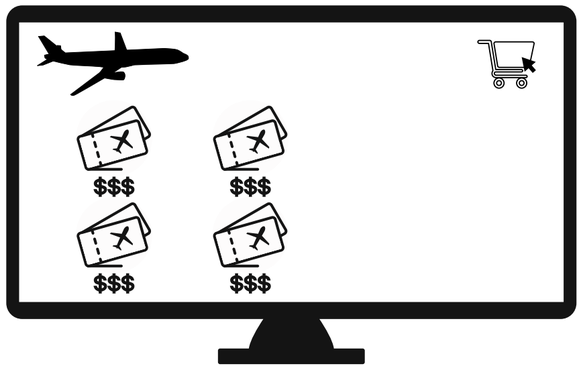


Agenda

1. The players
2. Real-world case study
3. Conclusion

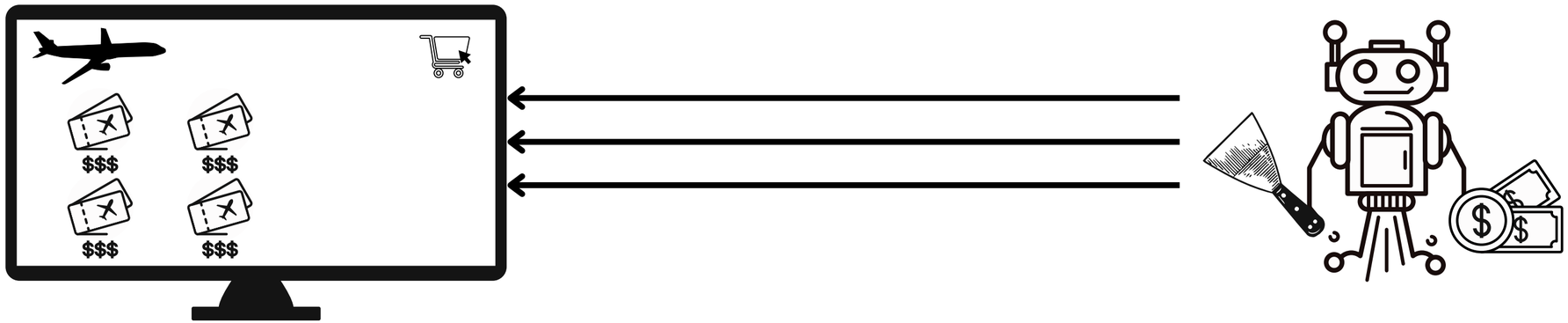
1 The players

The scenario



E-commerce websites

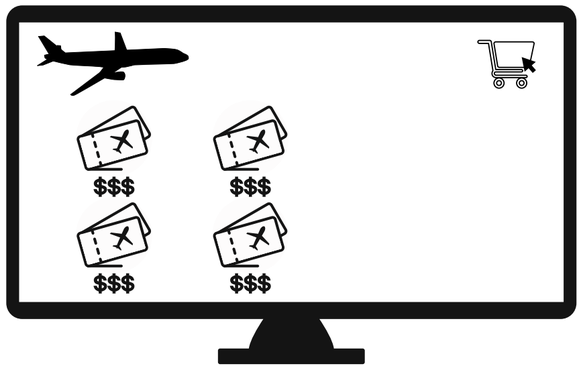
The scenario



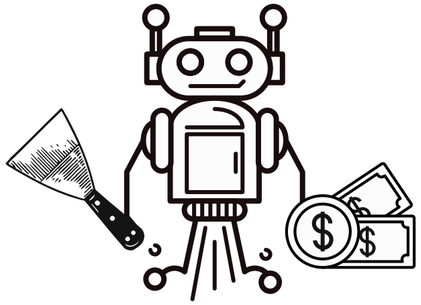
E-commerce websites

Scrapers

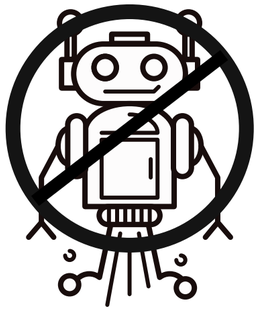
The scenario



E-commerce websites

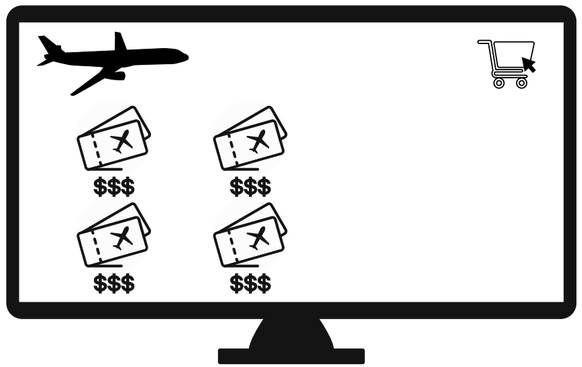


Scrapers

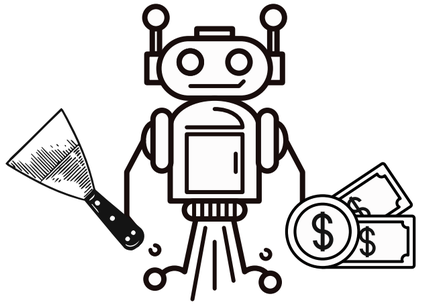


Anti-bot companies

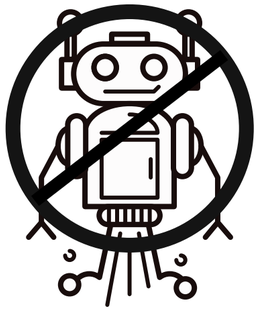
The scenario



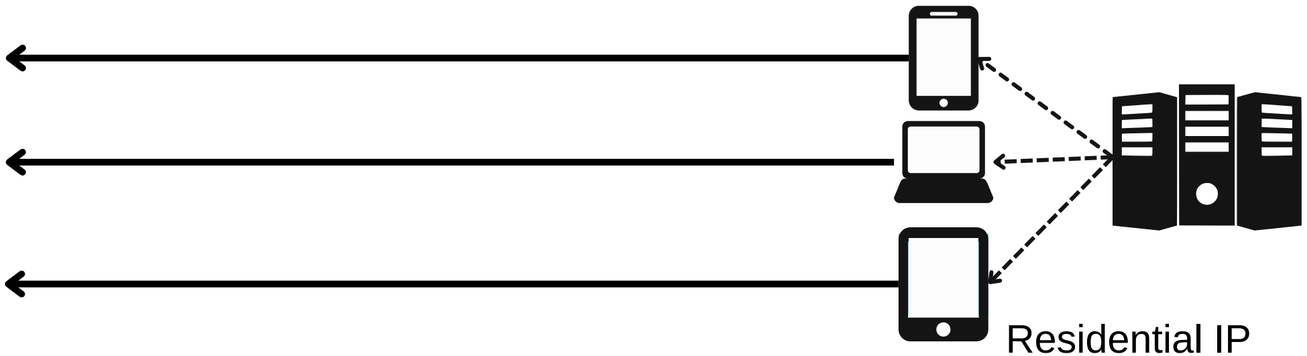
E-commerce websites



Scrapers

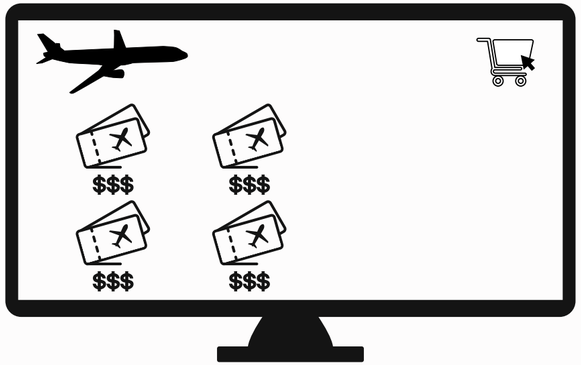


Anti-bot companies



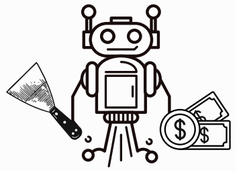
Residential IP
(RESIP) proxy

E-commerce websites

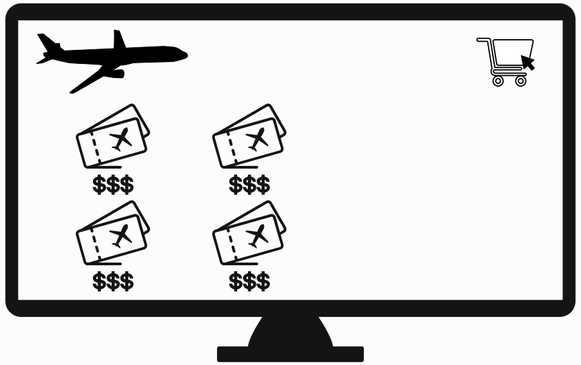


E-commerce websites

E-commerce websites



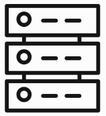
Server maintenance



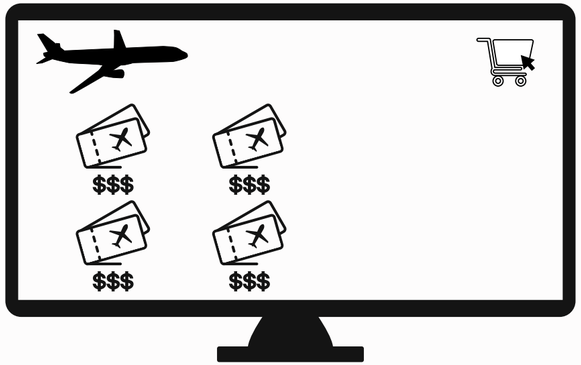
E-commerce websites

E-commerce websites

Rendering



Server maintenance

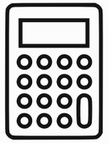


E-commerce websites

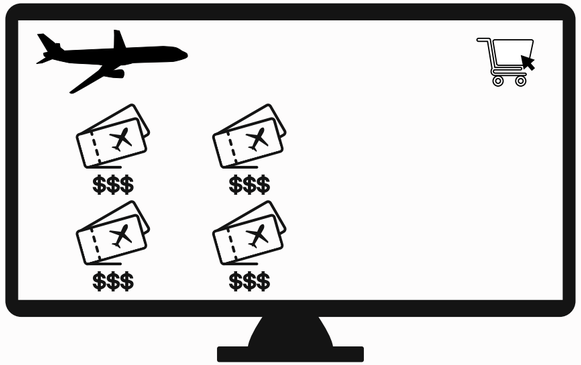
E-commerce websites

Rendering

Server maintenance

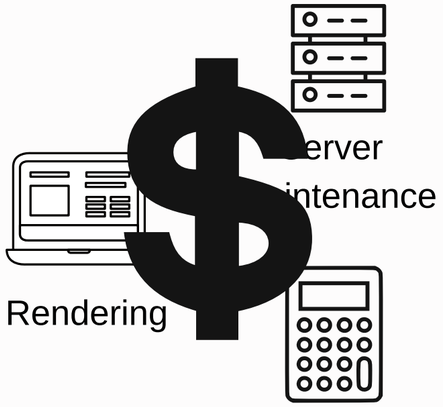


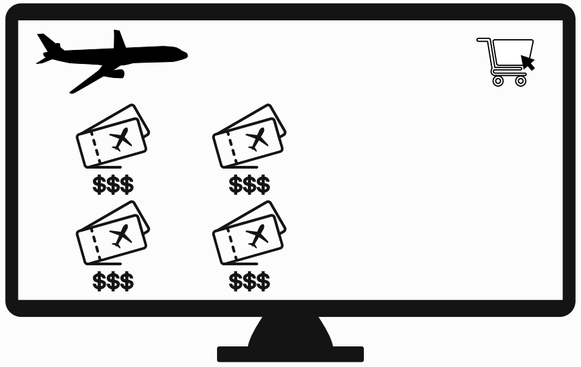
Price calculation



E-commerce websites

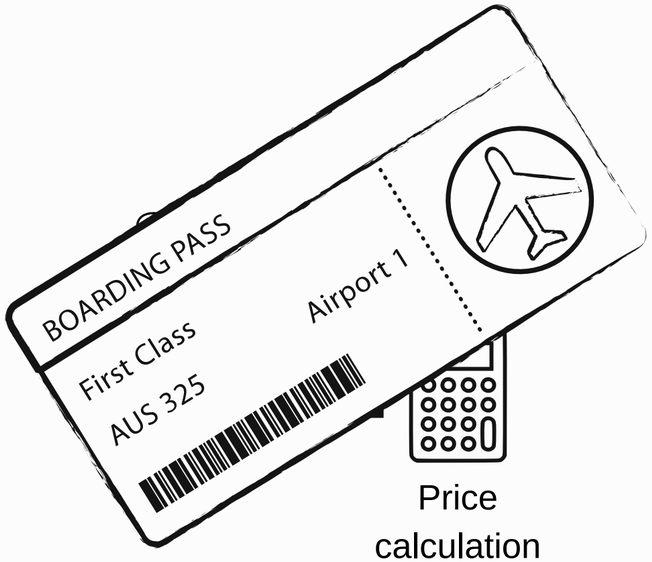
E-commerce websites



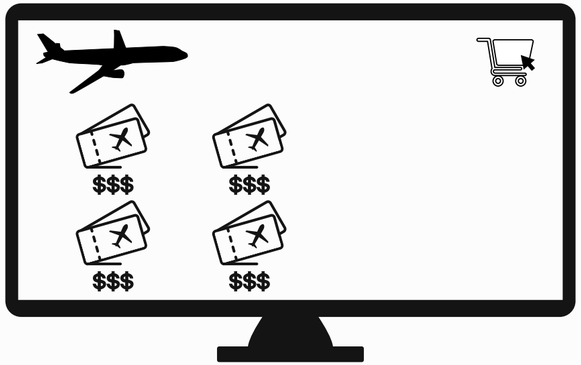


E-commerce websites

E-commerce websites

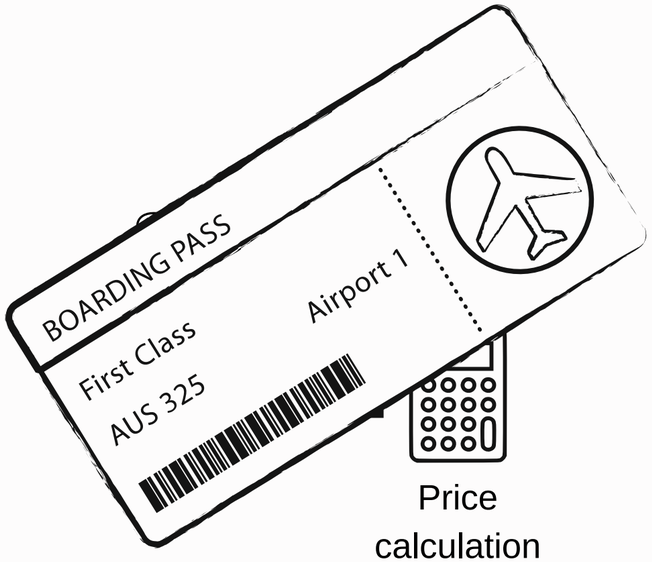


Price calculation

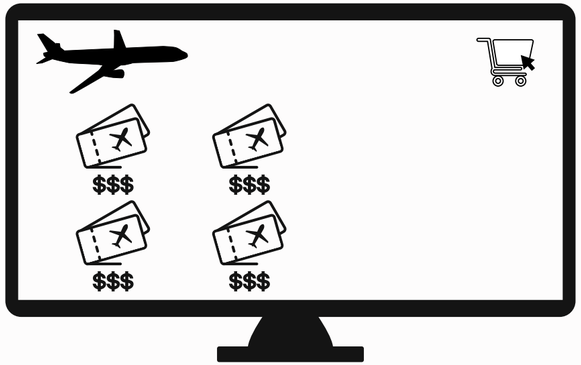


E-commerce websites

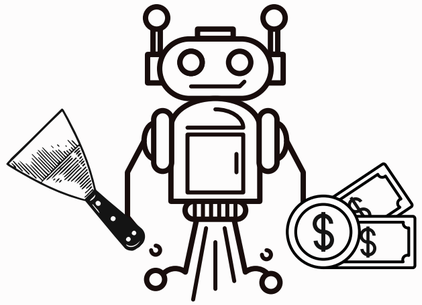
E-commerce websites



Price calculation



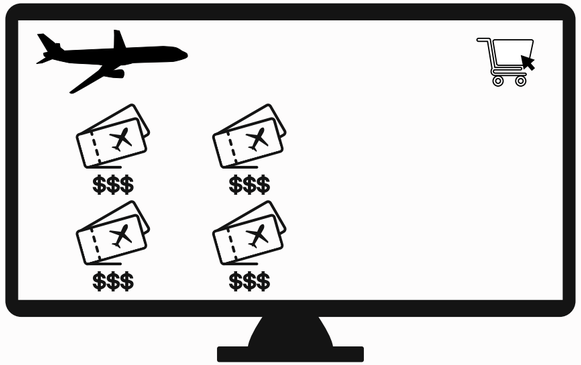
E-commerce websites



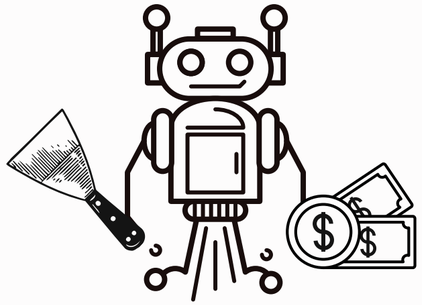
Scrapers

E-commerce websites

Price calculation



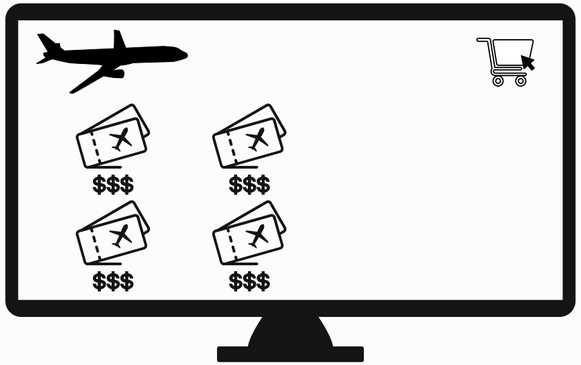
E-commerce websites



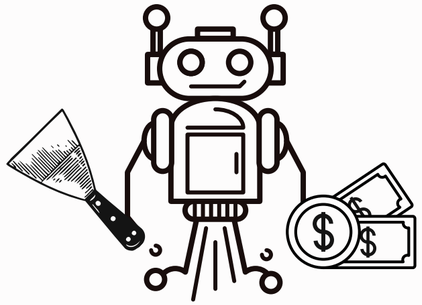
Scrapers

E-commerce websites

Price calculation

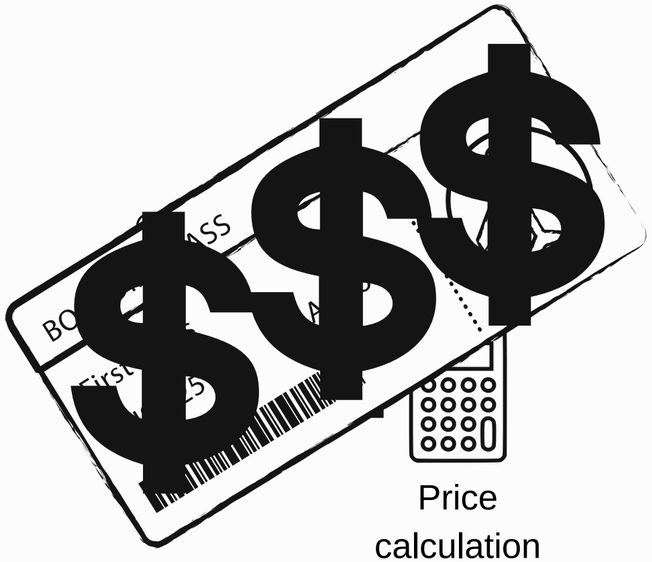


E-commerce websites

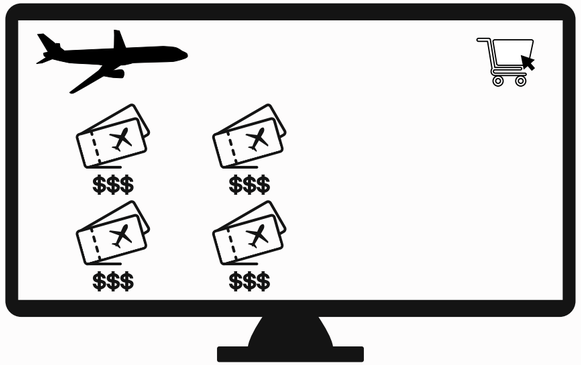


Scrapers

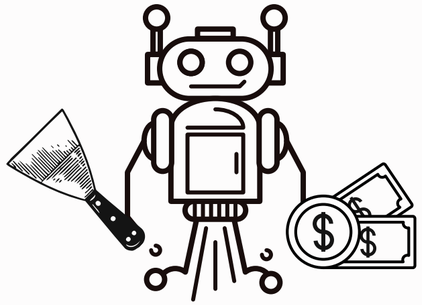
E-commerce websites



Price calculation

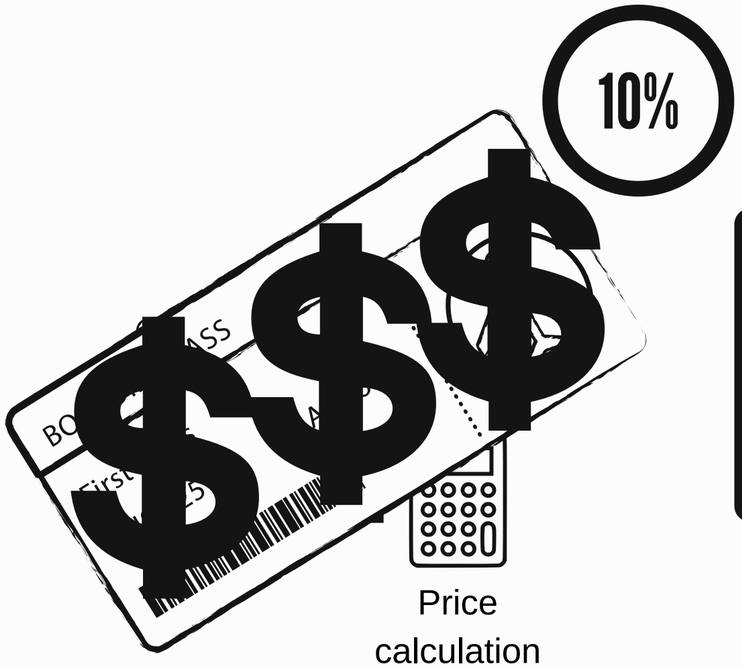


E-commerce websites

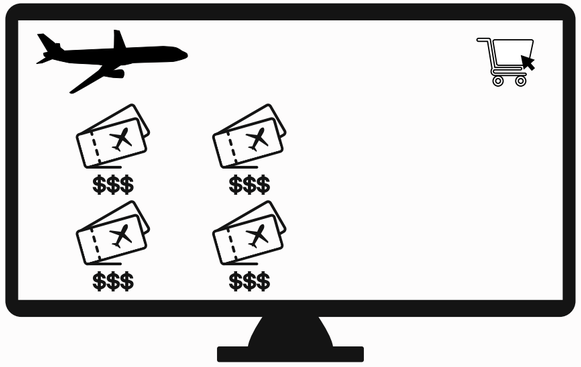


Scrapers

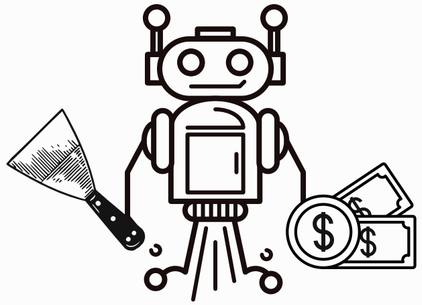
E-commerce websites



Price calculation



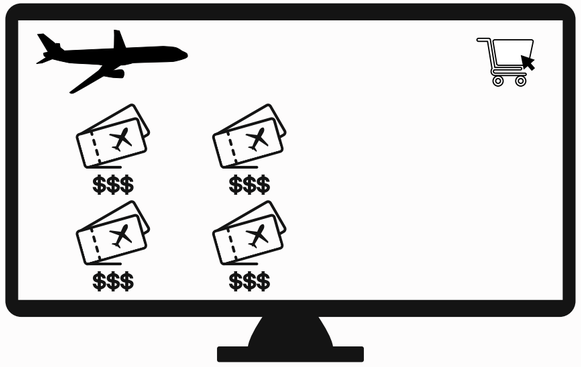
E-commerce websites



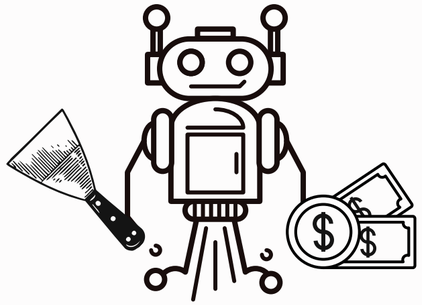
Scrapers

E-commerce websites

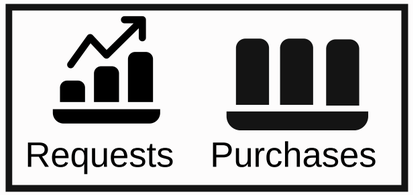
Price calculation



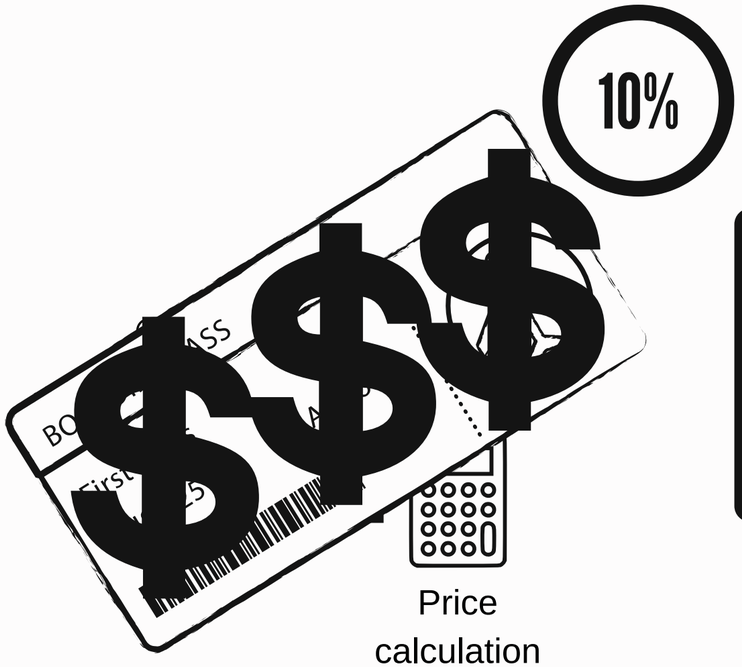
E-commerce websites



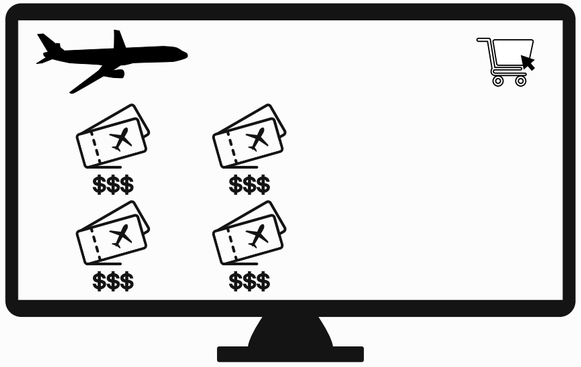
Scrapers



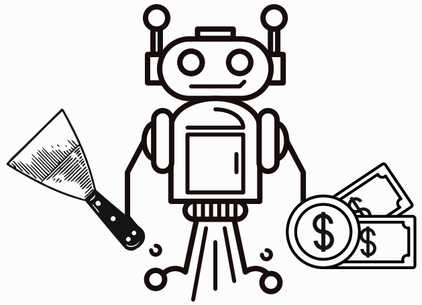
E-commerce websites



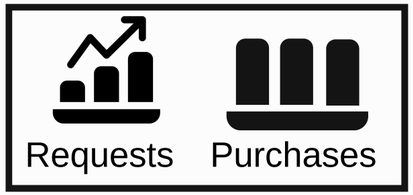
Price calculation



E-commerce websites

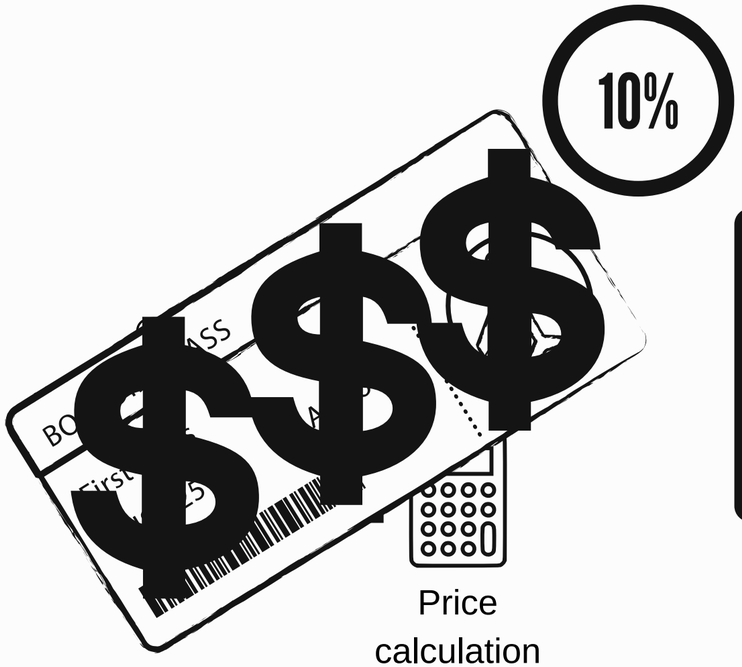


Scrapers

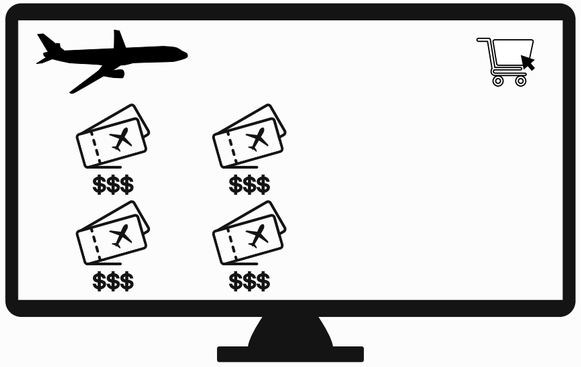


Slow connections

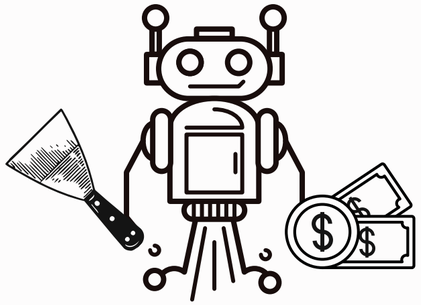
E-commerce websites



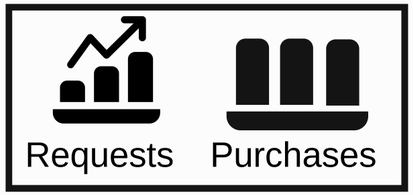
Price calculation



E-commerce websites



Scrapers

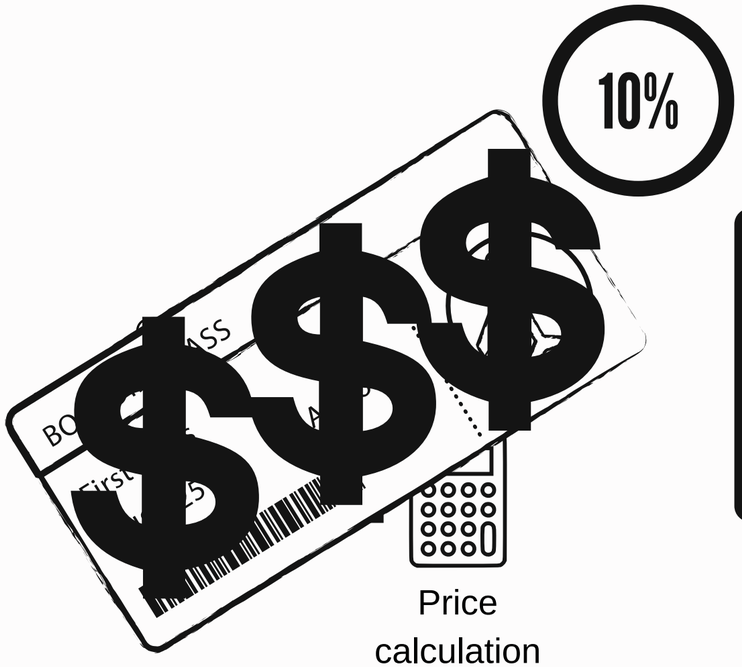


Slow connections

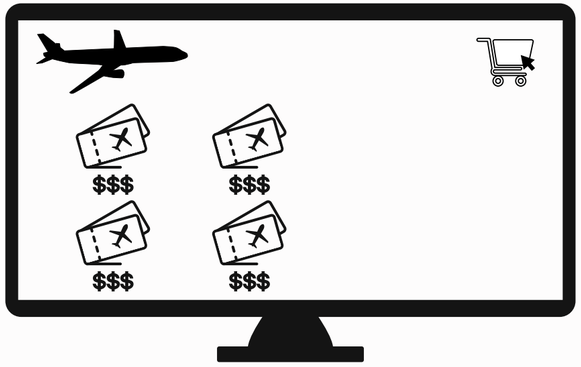
IP blocking

Rate limits

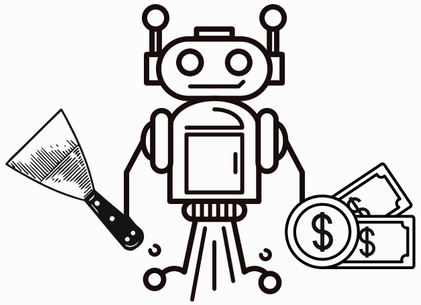
E-commerce websites



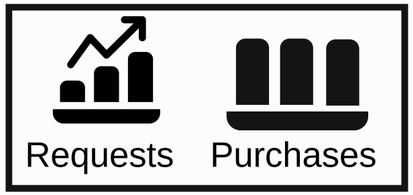
Price calculation



E-commerce websites



Scrapers



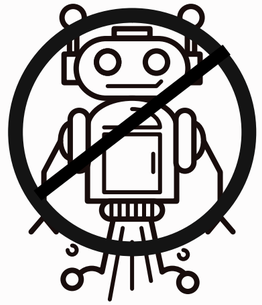
Slow connections

IP blocking

Rate limits

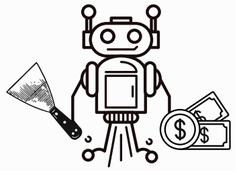
Anti-bot companies

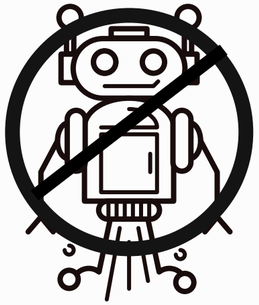
Anti-bot companies



Anti-bot companies

Anti-bot companies



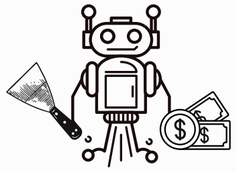


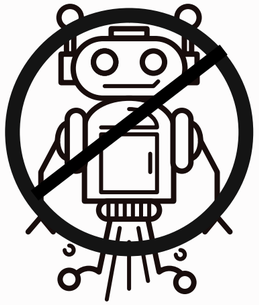
Anti-bot companies

http

HTTP header anomaly detection

Anti-bot companies





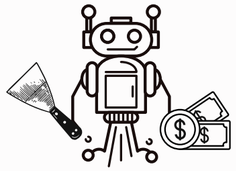
Anti-bot companies

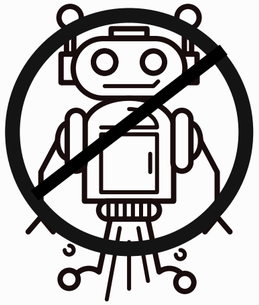
http

HTTP header anomaly detection

Browser fingerprinting

Anti-bot companies



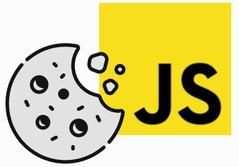


Anti-bot companies

http

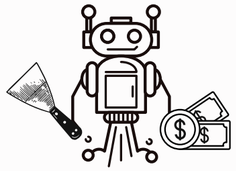
HTTP header anomaly detection

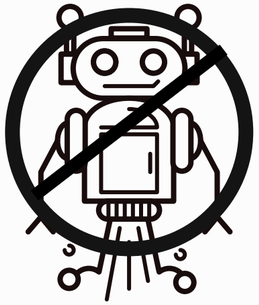
Browser fingerprinting



JS and cookies challenges

Anti-bot companies



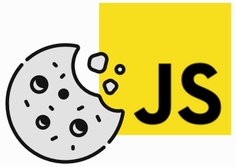


Anti-bot companies

http

HTTP header anomaly detection

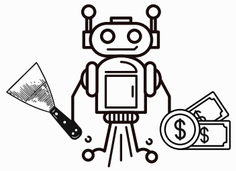
Browser fingerprinting

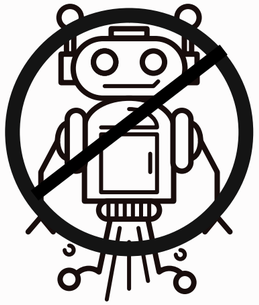


JS and cookies challenges

Human interaction check

Anti-bot companies



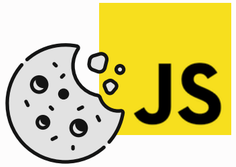


Anti-bot companies

http

HTTP header anomaly detection

Browser fingerprinting

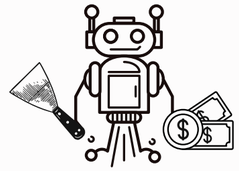


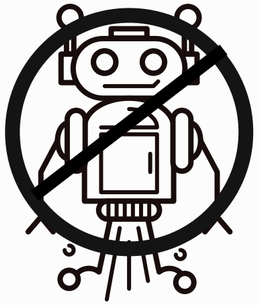
JS and cookies challenges

Human interaction check

CAPTCHAs

Anti-bot companies



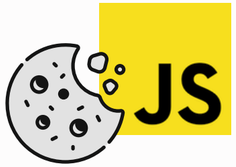


Anti-bot companies

http

HTTP header anomaly detection

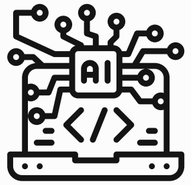
Browser fingerprinting



JS and cookies challenges

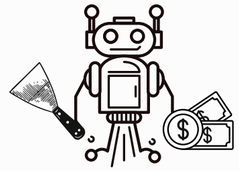
Human interaction check

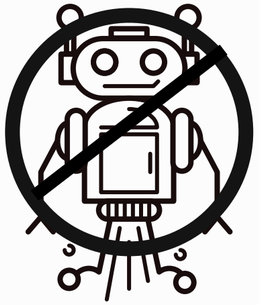
CAPTCHAs



Machine learning

Anti-bot companies



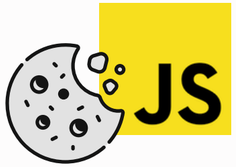


Anti-bot companies

http

HTTP header anomaly detection

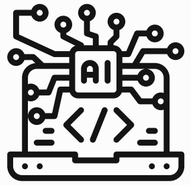
Browser fingerprinting



JS and cookies challenges

Human interaction check

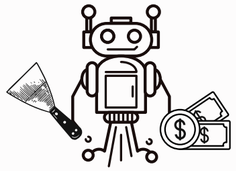
CAPTCHAs

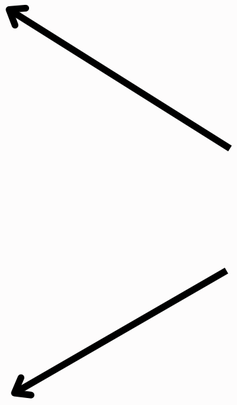


Machine learning

Wasting bots time

Anti-bot companies



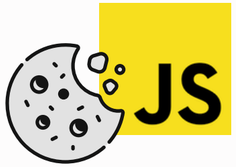


Anti-bot companies

http

HTTP header anomaly detection

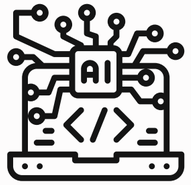
Browser fingerprinting



JS and cookies challenges

Human interaction check

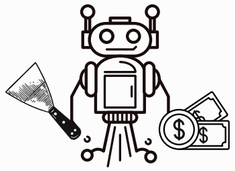
CAPTCHAs



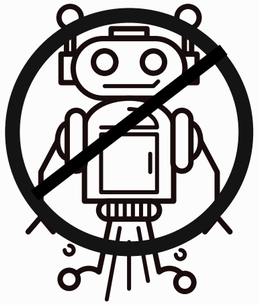
Machine learning

Wasting bots time

Anti-bot companies



Knowledge based

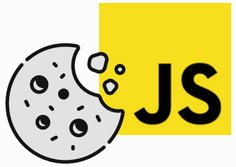


Anti-bot companies

http

HTTP header anomaly detection

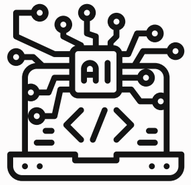
Browser fingerprinting



JS and cookies challenges

Human interaction check

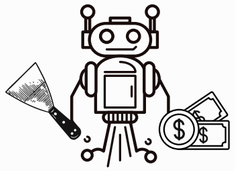
CAPTCHAs



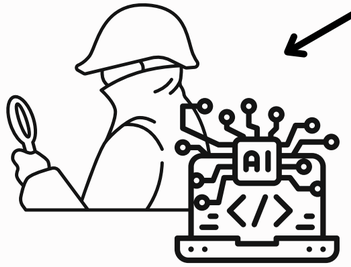
Machine learning

Wasting bots time

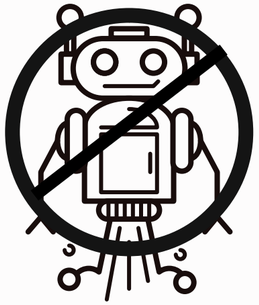
Anti-bot companies



Knowledge based



Behaviour based

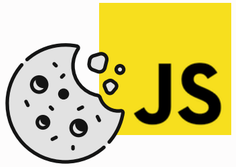


Anti-bot companies

http

HTTP header anomaly detection

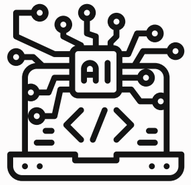
Browser fingerprinting



JS and cookies challenges

Human interaction check

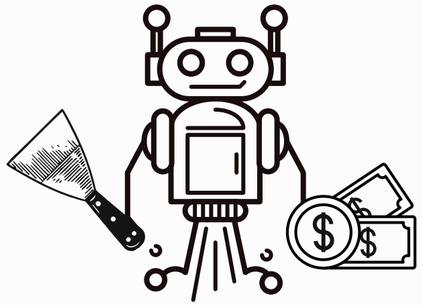
CAPTCHAs



Machine learning

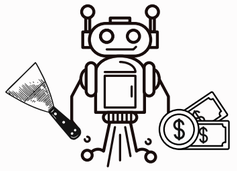
Wasting bots time

Scraping bots

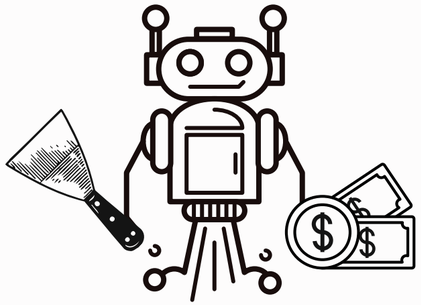


Scrapers

Scraping bots

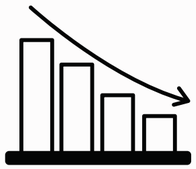


Competitors monitoring



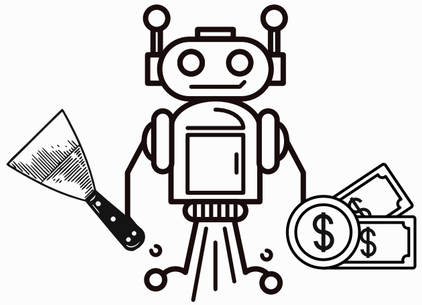
Scrapers

Scraping bots



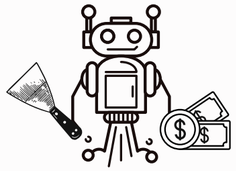
Statistics modification

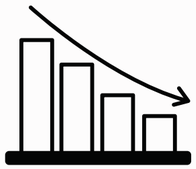
Competitors monitoring



Scrapers

Scraping bots

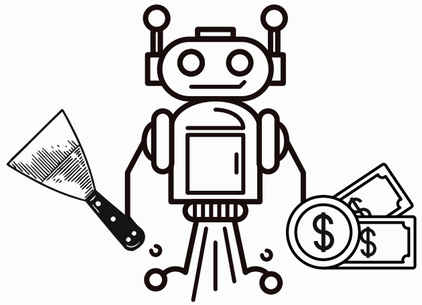




Statistics modification

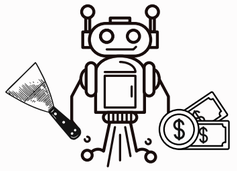
Competitors monitoring

Content reselling



Scrapers

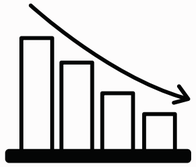
Scraping bots



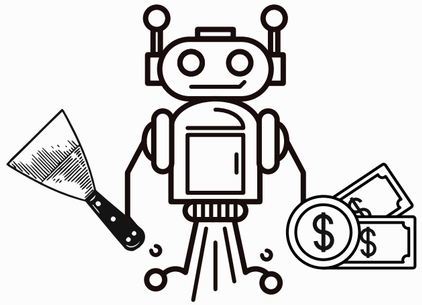
Scrapy PhantomJS

Browser emulation frameworks

Competitors monitoring



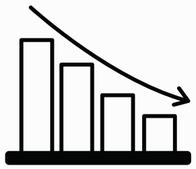
Statistics modification



Scrapers

Content reselling

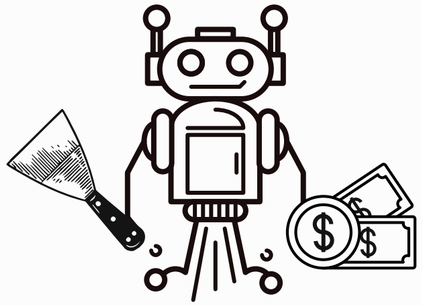
Scraping bots



Statistics modification

Competitors monitoring

Content reselling



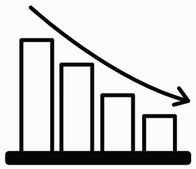
Scrapers

Scrapy PhantomJS

Browser emulation frameworks

Automated browsers, cookies and JS support

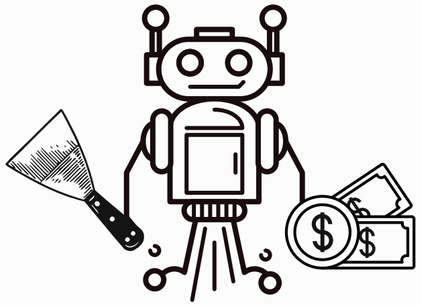
Scraping bots



Statistics modification

Competitors monitoring

Content reselling



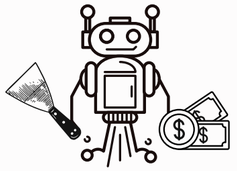
Scrapers

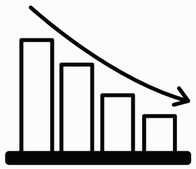
Browser emulation frameworks

Automated browsers, cookies and JS support

CAPTCHA farms

Scraping bots

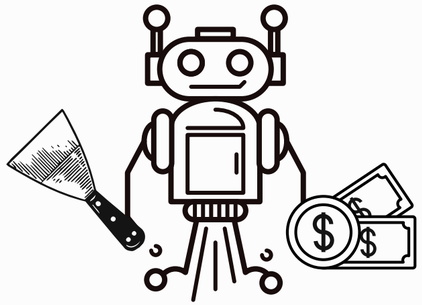




Statistics modification

Competitors monitoring

Content reselling

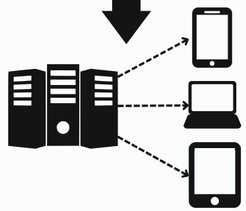


Scrapers

Browser emulation frameworks

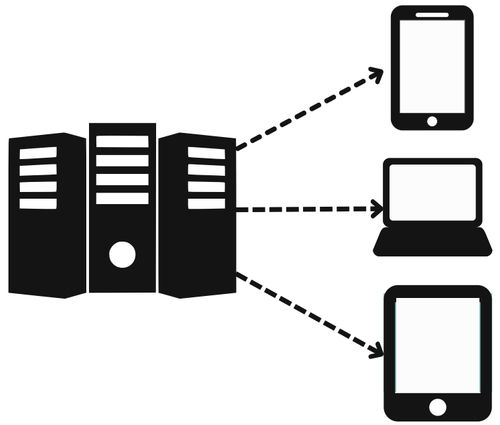
Automated browsers, cookies and JS support

CAPTCHA farms



Residential IP (RESIP) proxy

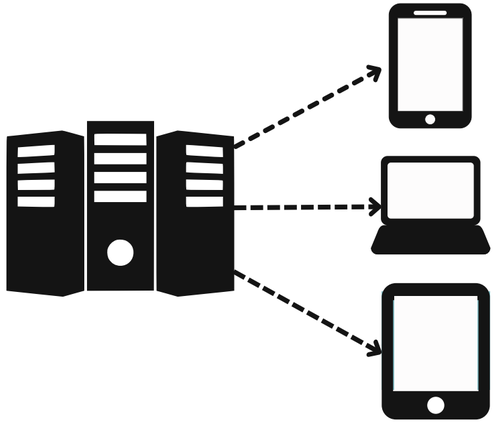
RESIP providers



Residential IP
(RESIP) proxy

RESIP providers

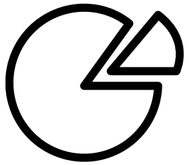
Huge IP pool



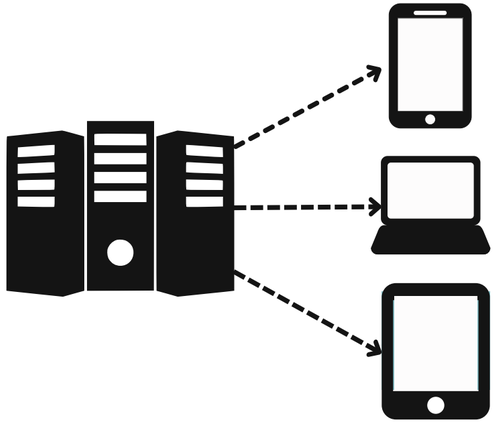
Residential IP
(RESIP) proxy

RESIP providers

Huge IP pool



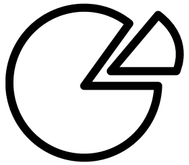
30% of malicious traffic coming from RESIP (DataDome 2019)



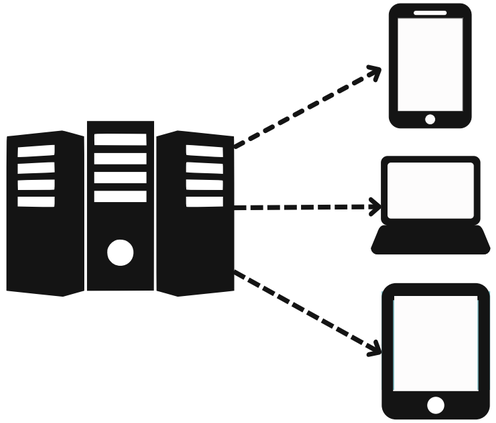
Residential IP (RESIP) proxy

RESIP providers

Huge IP pool



30% of malicious traffic coming from RESIP (DataDome 2019)

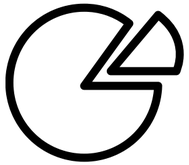


Residential IP (RESIP) proxy

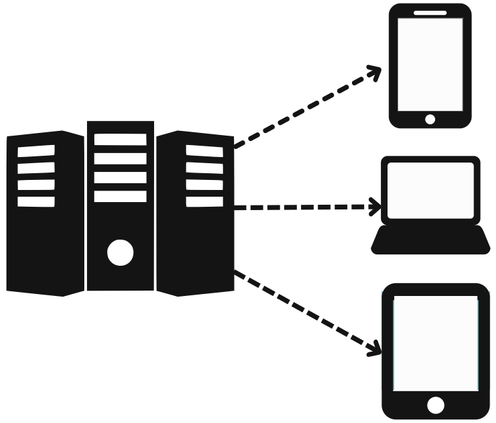
Good reputation IPs

RESIP providers

Huge IP pool

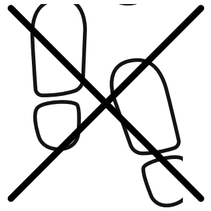


30% of malicious traffic coming from RESIP (DataDome 2019)



Residential IP (RESIP) proxy

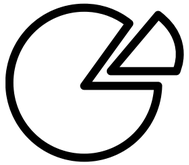
Good reputation IPs



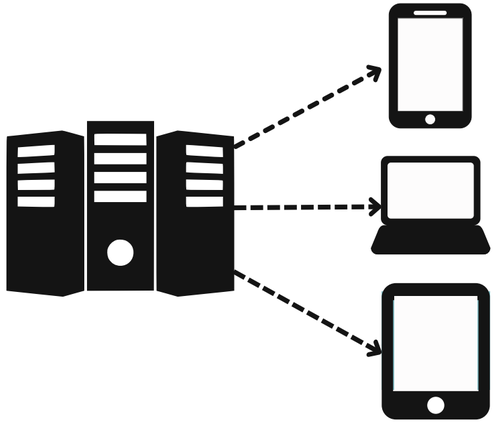
No traceback

RESIP providers

Huge IP pool

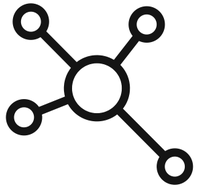


30% of malicious traffic coming from RESIP (DataDome 2019)

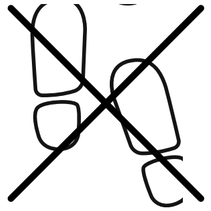


Residential IP (RESIP) proxy

Good reputation IPs



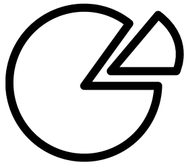
No private distributed infrastructure



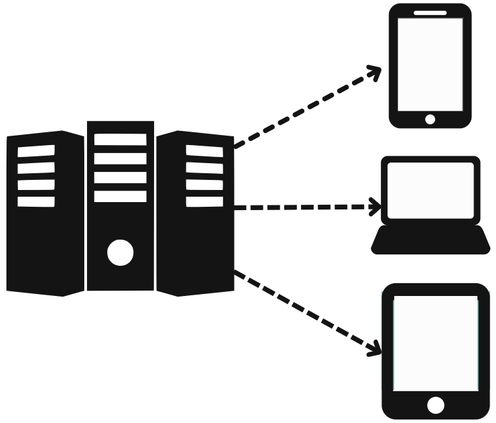
No traceback

RESIP providers

Huge IP pool

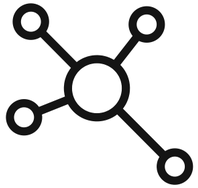


30% of malicious traffic coming from RESIP (DataDome 2019)

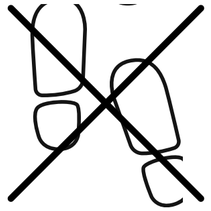


Residential IP (RESIP) proxy

Good reputation IPs



No private distributed infrastructure

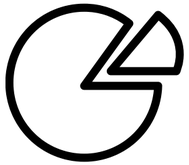


No traceback

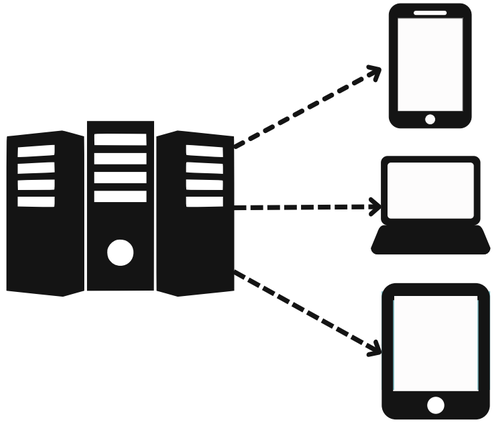
Automated services

RESIP providers

Huge IP pool

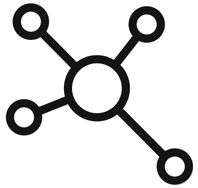


30% of malicious traffic coming from RESIP (DataDome 2019)

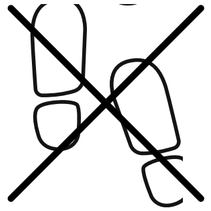


Residential IP (RESIP) proxy

Good reputation IPs



No private distributed infrastructure



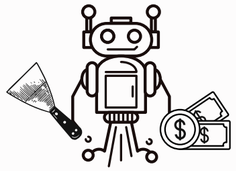
No traceback

Automated services

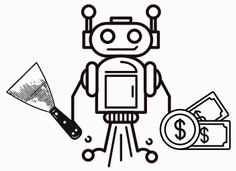
Higher risk to block legittimate users

2 Real-world case study

amadeus

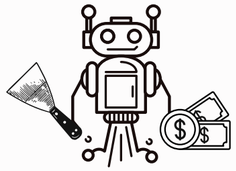


amadeus



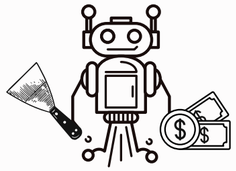
- Global distribution system

amadeus



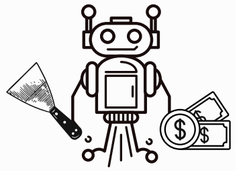
- Global distribution system
- **World-leading** technology companies for travel and tourism

amadeus



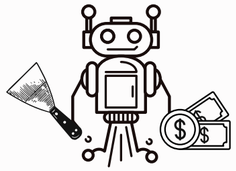
- Global distribution system
- **World-leading** technology companies for travel and tourism
- Airlines fare calculations based on a **large number** of parameters

amadeus



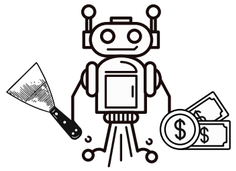
- Global distribution system
- **World-leading** technology companies for travel and tourism
- Airlines fare calculations based on a **large number** of parameters
- Domains heavily **targeted** by scrapers

amadeus

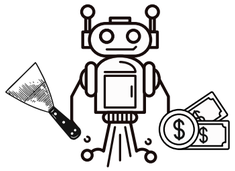


- Global distribution system
- **World-leading** technology companies for travel and tourism
- Airlines fare calculations based on a **large number** of parameters
- Domains heavily **targeted** by scrapers
- 200 websites **protected** by a knowledge-based anti-bot solution

Scrapers vs amadeus: why?

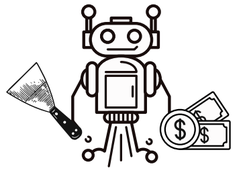


Scrapers vs amadeus: why?



Competitive intelligence companies

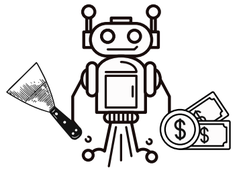
Scrapers vs amadeus: why?



Competitive intelligence companies

Aggregators

Scrapers vs amadeus: why?

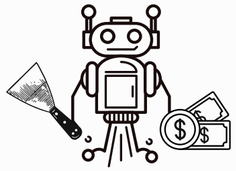


Competitive intelligence companies

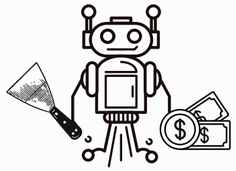
Aggregators

Online travel agencies

Scrapers vs amadeus: how much?

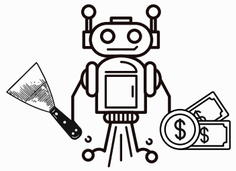


Scrapers vs amadeus: how much?



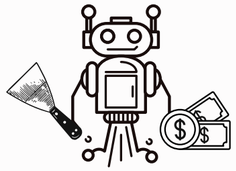
- Every month, anti-bot rules triggered by **140 million** requests

Scrapers vs amadeus: how much?



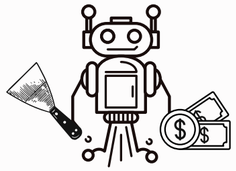
- Every month, anti-bot rules triggered by **140 million** requests
- **41%** of the attempted connections detected as bots (February 2022)

Scrapers vs amadeus: how much?

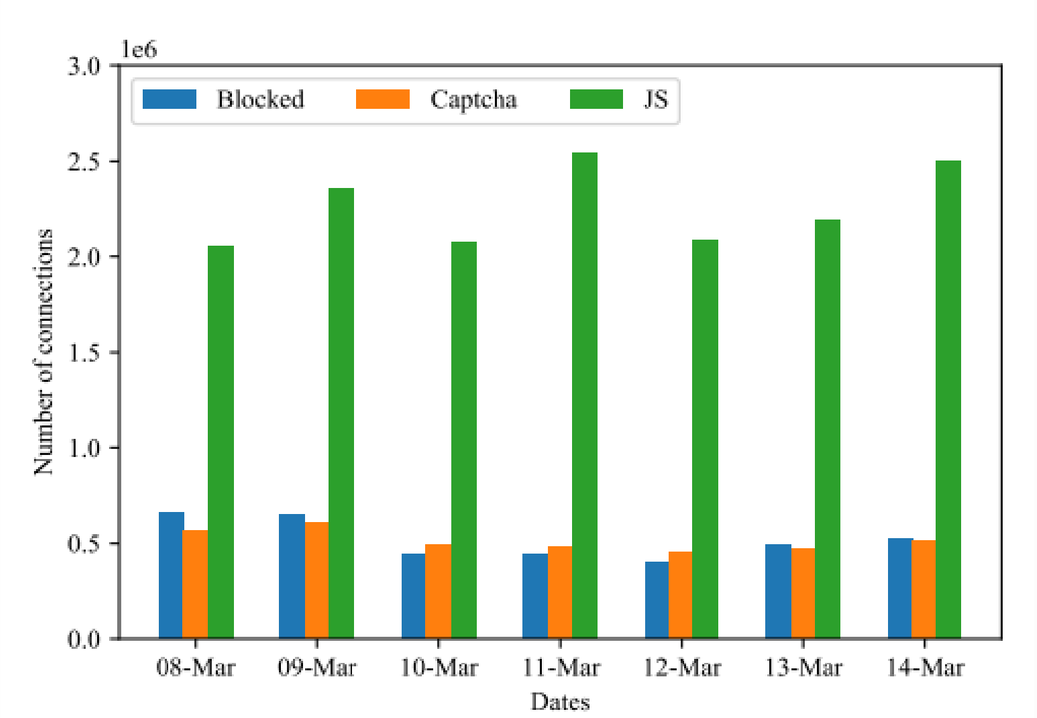


- Every month, anti-bot rules triggered by **140 million** requests
- **41%** of the attempted connections detected as bots (February 2022)
- **Constant** bot traffic

Scrapers vs amadeus: how much?

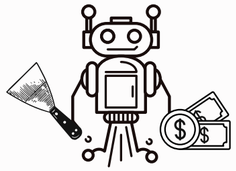


- Every month, anti-bot rules triggered by **140 million** requests
- **41%** of the attempted connections detected as bots (February 2022)
- **Constant** bot traffic

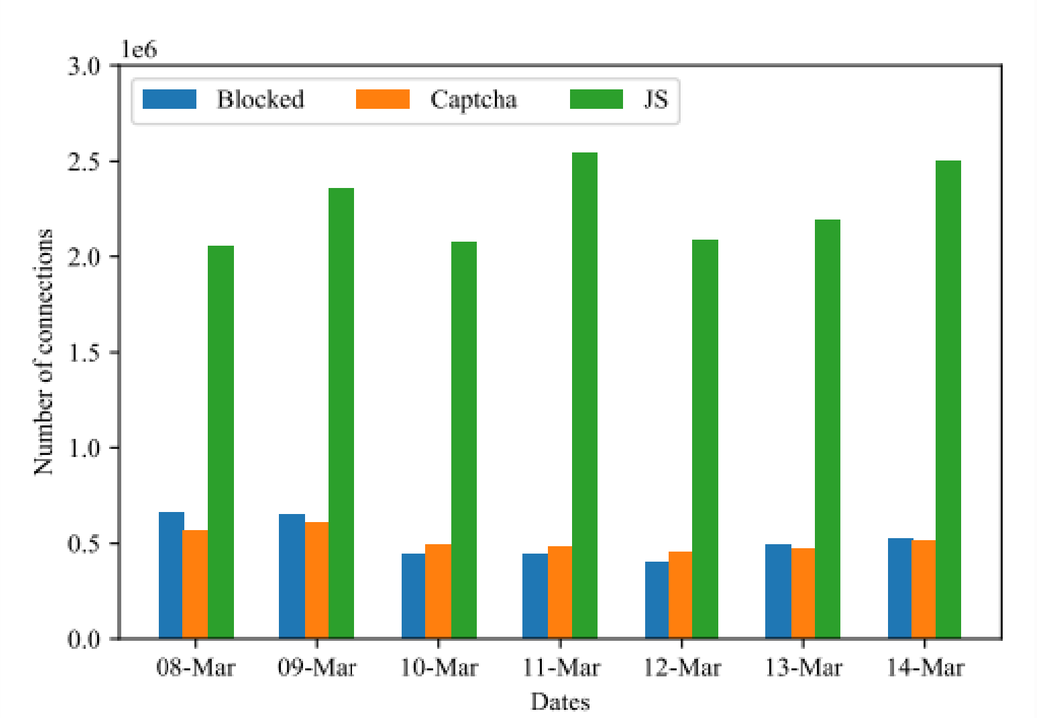


Daily amount of reactions
between 08/03/22 and 14/03/22.

Scrapers vs amadeus: how much?

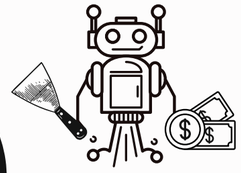


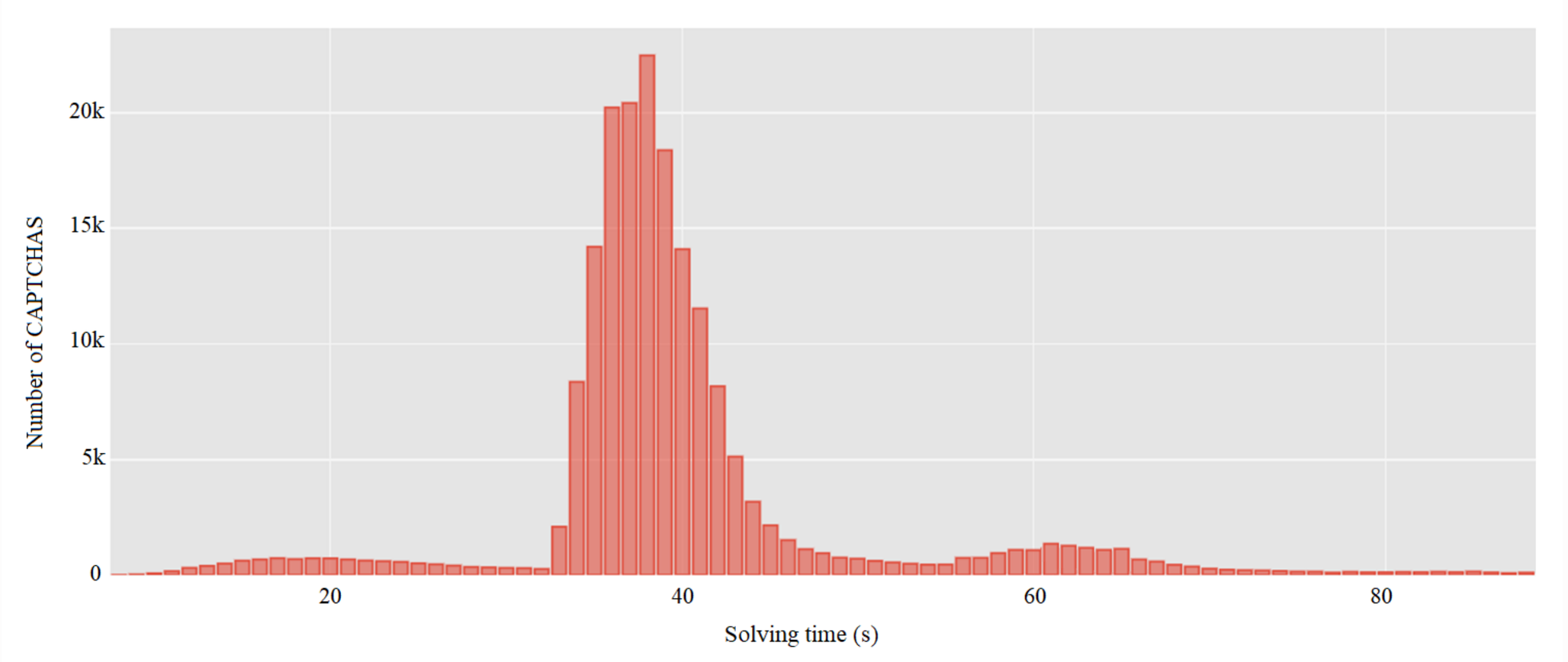
- Every month, anti-bot rules triggered by **140 million** requests
- **41%** of the attempted connections detected as bots (February 2022)
- **Constant** bot traffic
- Bot reaction to countermeasures: from days (past years) to **hours** (now)



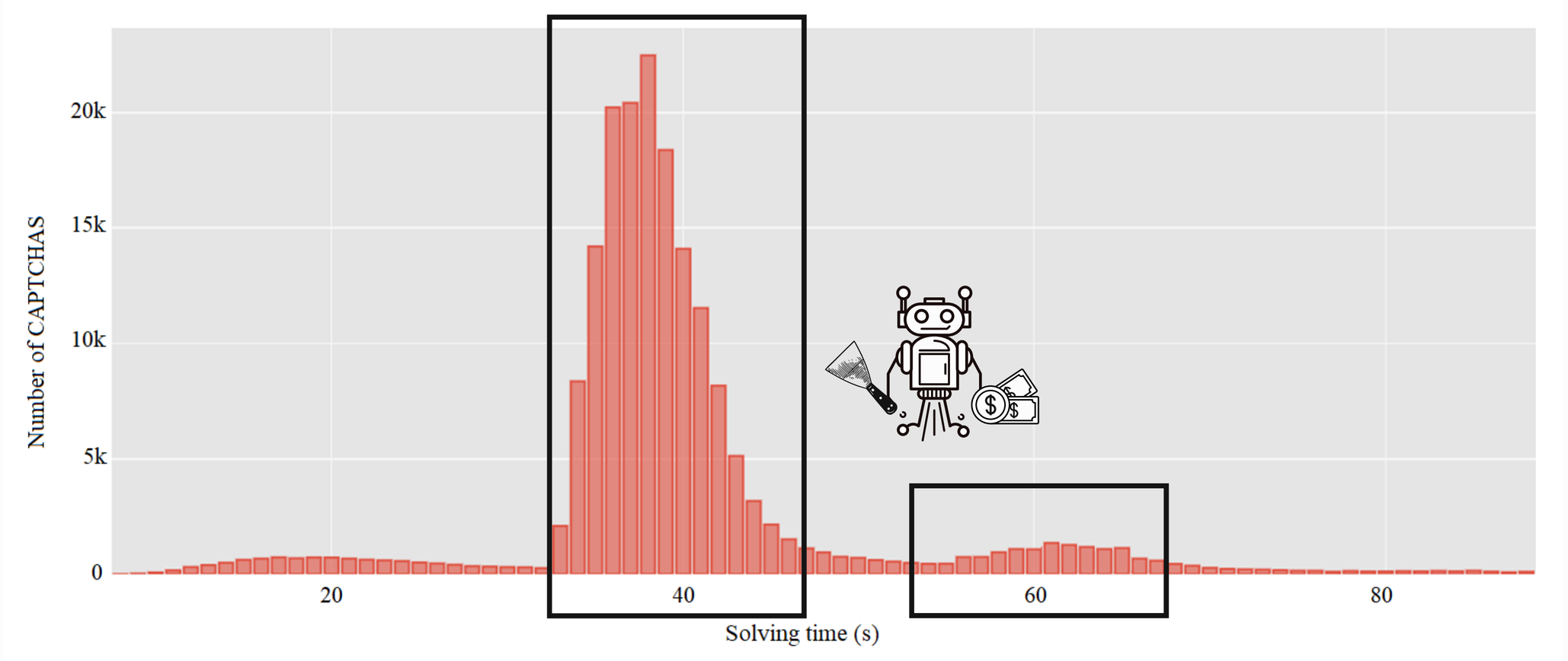
Daily amount of reactions
between 08/03/22 and 14/03/22.

CAPTCHA solving time (2018)

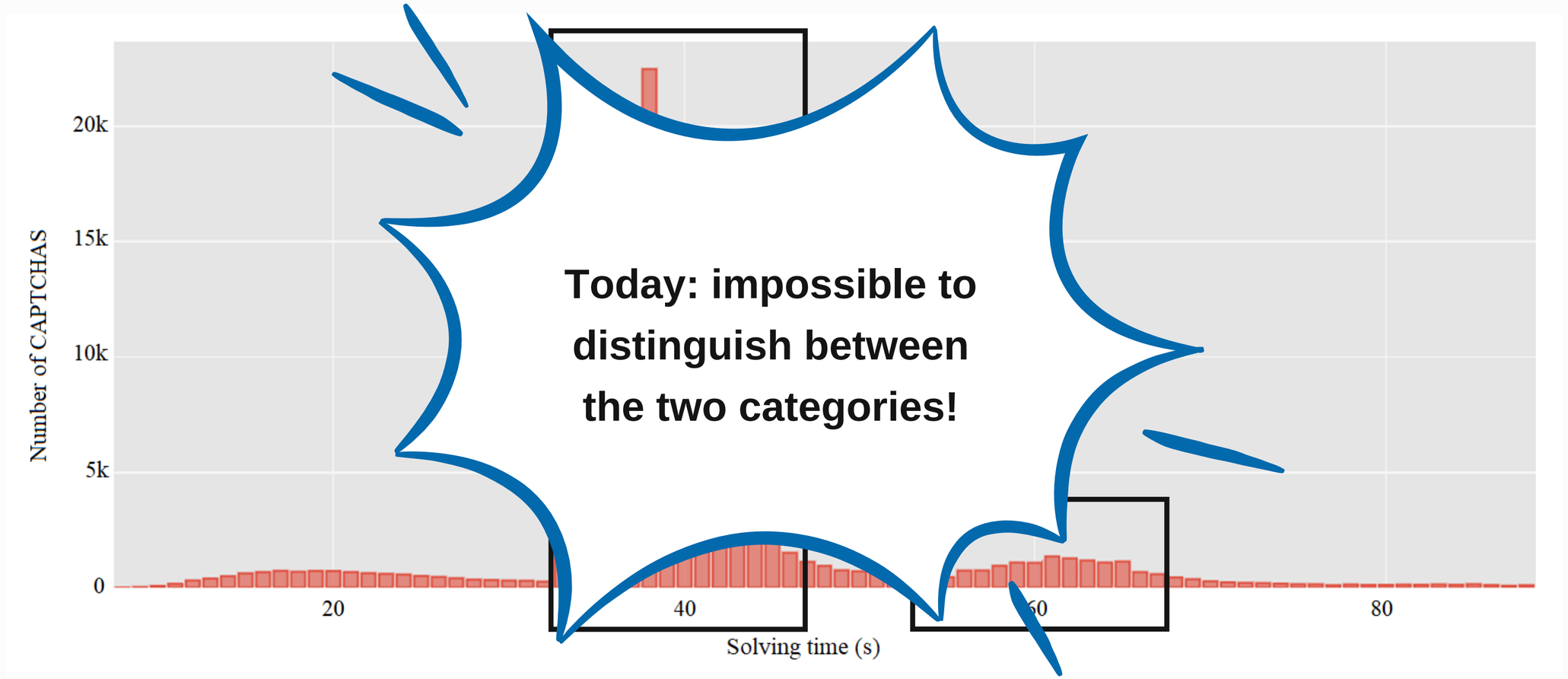




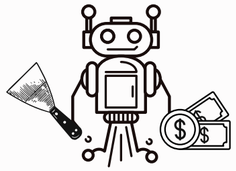
CAPTCHA solving time (2018)



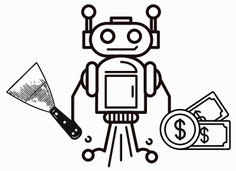
CAPTCHA solving time (2018)



RESIP activities in AMADEUS



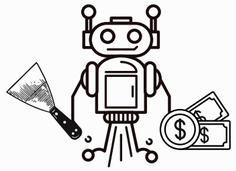
RESIP activities in AMADEUS



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

Goal: reducing false positives

RESIP activities in AMADEUS

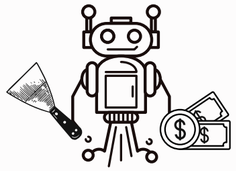


Residential IPs detected as bots in 30 days: **12%**

Goal: reducing false positives

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

RESIP activities in AMADEUS



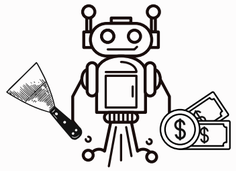
Residential IPs detected as bots in 30 days: **12%**

Goal: reducing false positives

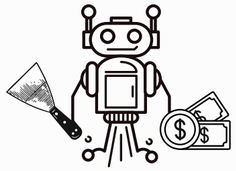
Total RESIP traffic is a much **larger** portion

Wide usage of RESIP

RESIP problem

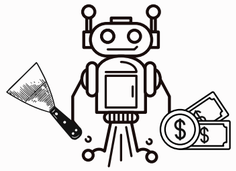


RESIP problem



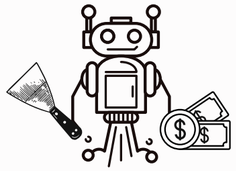
- Main problem: high risk of **false positives**

RESIP problem



- Main problem: high risk of **false positives**
- **New** specific detection methods are needed

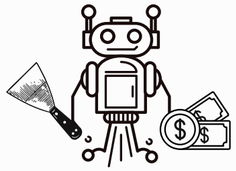
RESIP problem



- Main problem: high risk of **false positives**
- **New** specific detection methods are needed
- Recent works suggest the RESIP pools of IP addresses are **smaller** than claimed [1]

[1] Chiapponi et al (2021). "Scraping Airlines Bots: Insights Obtained Studying Honeypot Data" in International Journal of Cyber Forensics and Advanced Threat Investigations

RESIP problem



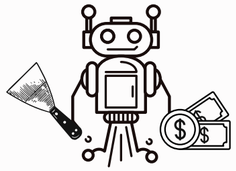
- Main problem: high risk of **false positives**
- **New** specific detection methods are needed
- Recent works suggest the RESIP pools of IP addresses are **smaller** than claimed [1]

If confirmed, **IP blocking** could be used

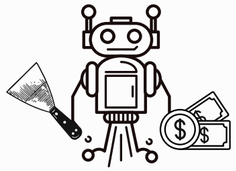
[1] Chiapponi et al (2021). "Scraping Airlines Bots: Insights Obtained Studying Honeypot Data" in International Journal of Cyber Forensics and Advanced Threat Investigations

3 Conclusion

Take aways

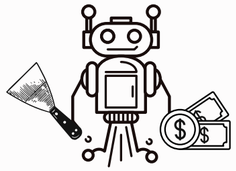


Take aways



Web scraping **highly** affects e-commerce websites

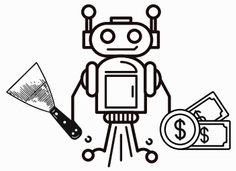
Take aways



Web scraping **highly** affects e-commerce websites

Scrapers techniques keep **evolving** to overcome anti-bot detection

Take aways

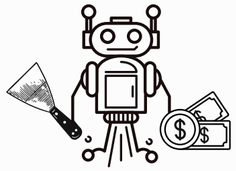


Web scraping **highly** affects e-commerce websites

Scrapers techniques keep **evolving** to overcome anti-bot detection

Lately, scrapers take advantage of **RESIP**, increasing the probability of blocking legitimate users

Take aways

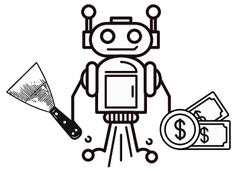


Web scraping **highly** affects e-commerce websites

Scrapers techniques keep **evolving** to overcome anti-bot detection

Lately, scrapers take advantage of **RESIP**, increasing the probability of blocking legitimate users

New detection techniques for RESIP are needed



Thank you!
Q&A

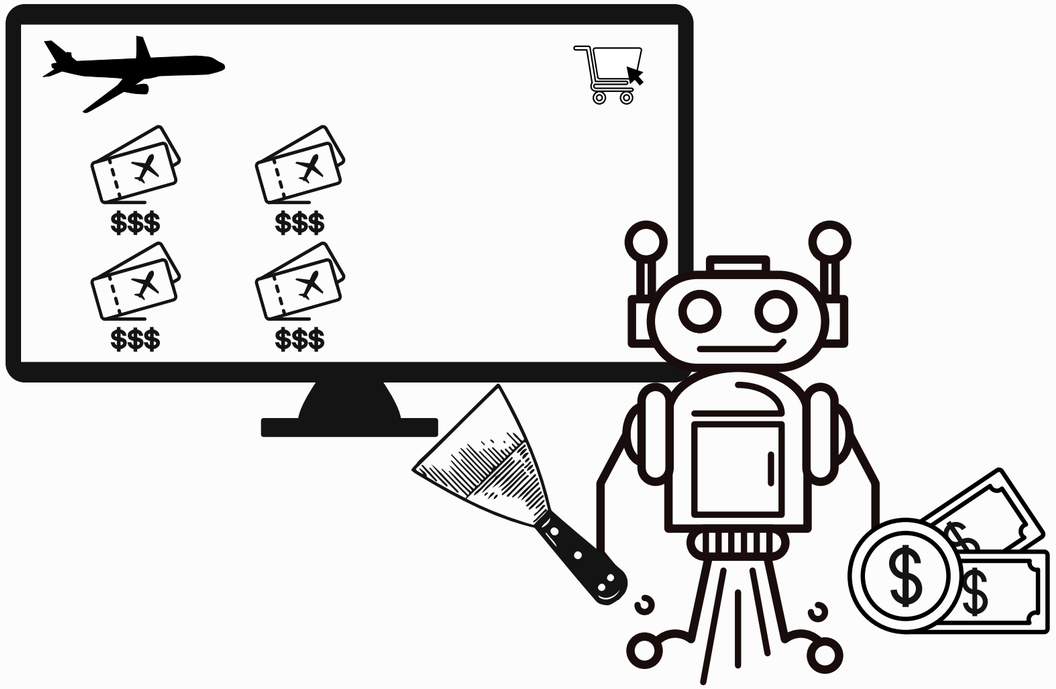
More questions? DSN22 Slack or elisa.chiapponi@eurecom.fr

EURECOM
Sophia Antipolis

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

amadeus
Global Security Operations

An industrial perspective on web scraping characteristics and open issues



Elisa Chiapponi, Marc Dacier, Olivier Thonnard,
Mohamed Fangar, Mattias Mattsson, Vincent Rigal
elisa.chiapponi@eurecom.fr, marc.dacier@kaust.edu.sa,
{olivier.thonnard, mohamed.fangar, mattias.mattsson, vincent.rigal}@amadeus.com

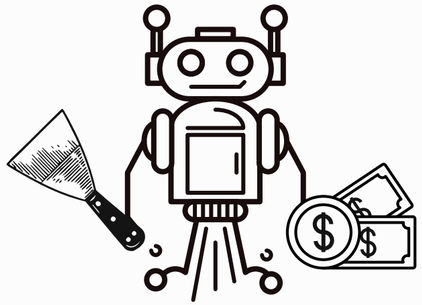
EURECOM
Sophia Antipolis

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

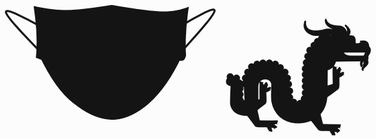
amadeus
Global Security Operations

DSN 2022

Scraping bots

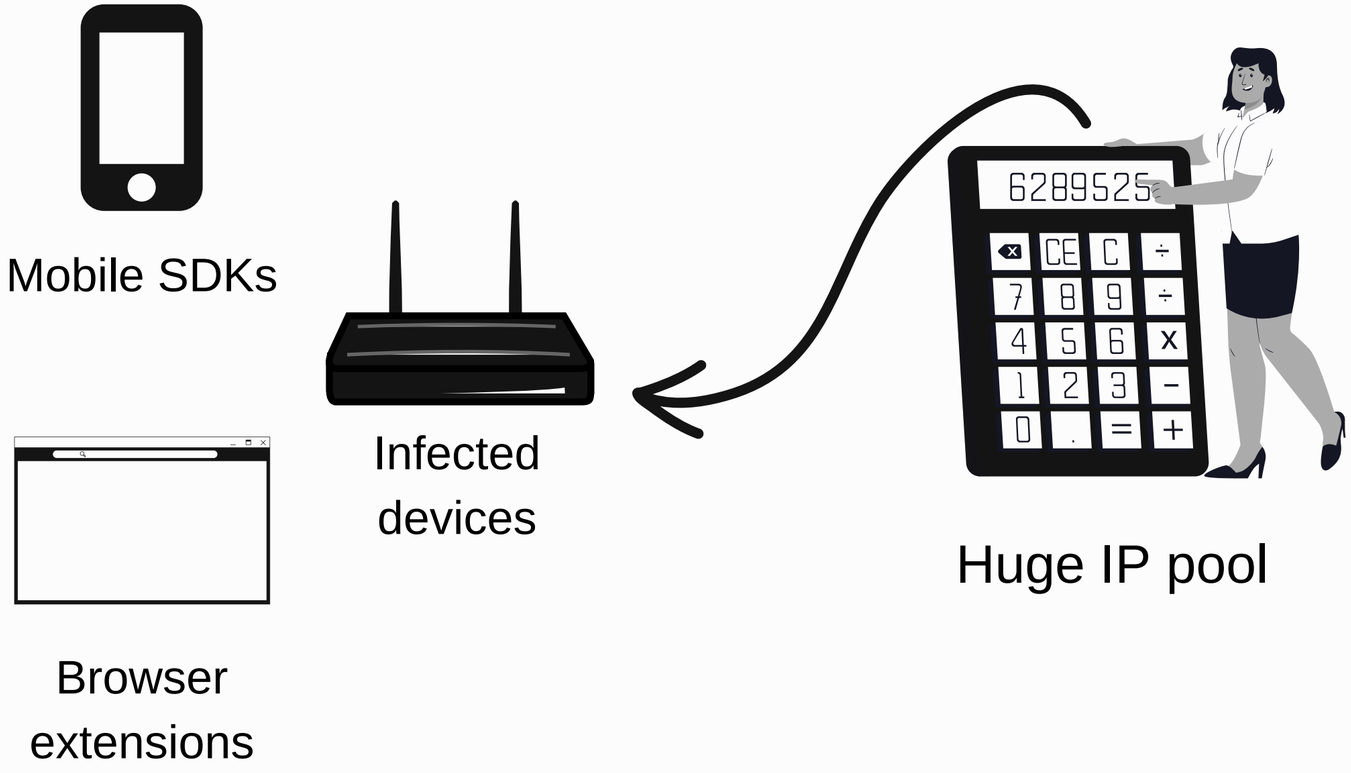


Scraping bots

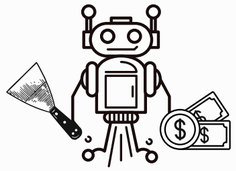


Evolving with market conditions

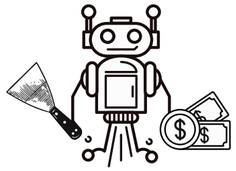
RESIP devices



Who are we?



- **Elisa Chiapponi**
  - Ph.D. student at EURECOM, collaboration with Amadeus IT Group.
- **Marc Dacier**
  - Professor and associate director of RC3, KAUST
- **Olivier Thonnard**
  - Director of Global Security Operations, Amadeus IT Group
- **Mohamed Fangar**
  - Associate Manager at Application Security Operation Center, Amadeus IT Group
- **Mattias Mattsson**
  - Senior Security Analyst, Sentor MSS AB for Amadeus IT Group
- **Vincent Rigal**
  - Head of Application Security Operations Center, Amadeus IT Group



Thank you!
Q&A

More questions? elisa.chiapponi@eurecom.fr or