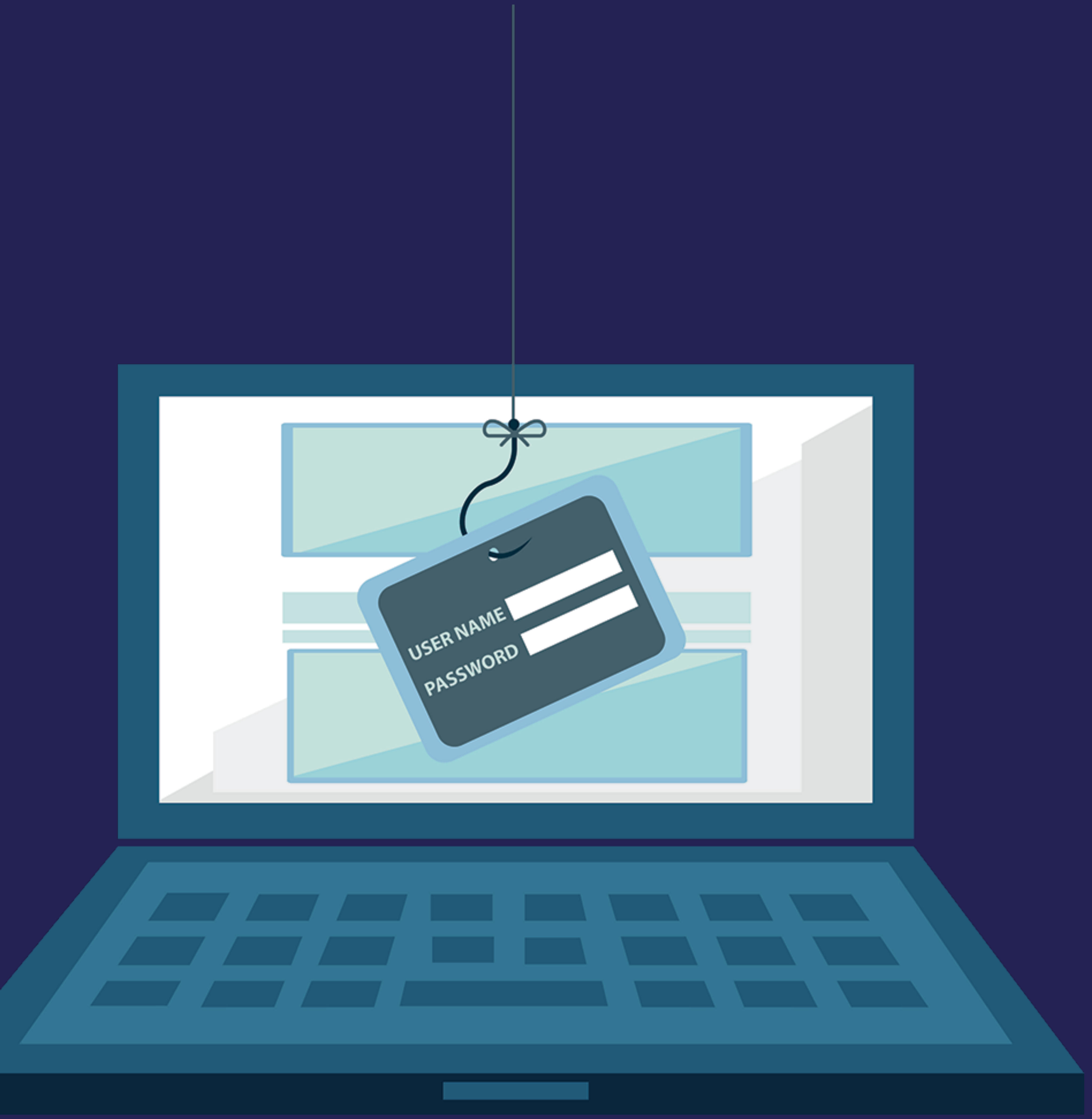


Using Machine Learning to Identify **Domain Abuse at Time of Registration**

Pieter Robberechts
ICANN Tech Day - 13 June 2022 – The Hague

DTAI KU LEUVEN dnsbelgium

Features

= properties indicative of malicious intent

Features are based on 5 underlying assumptions:

1. *Malicious registrants reuse the same / similar registration details*
2. *Malicious registrants provide fake contact info*
3. *Malicious registrants reuse infrastructure*
4. *Malicious registrants reuse domains*
5. *Malicious registrants register similar domains*

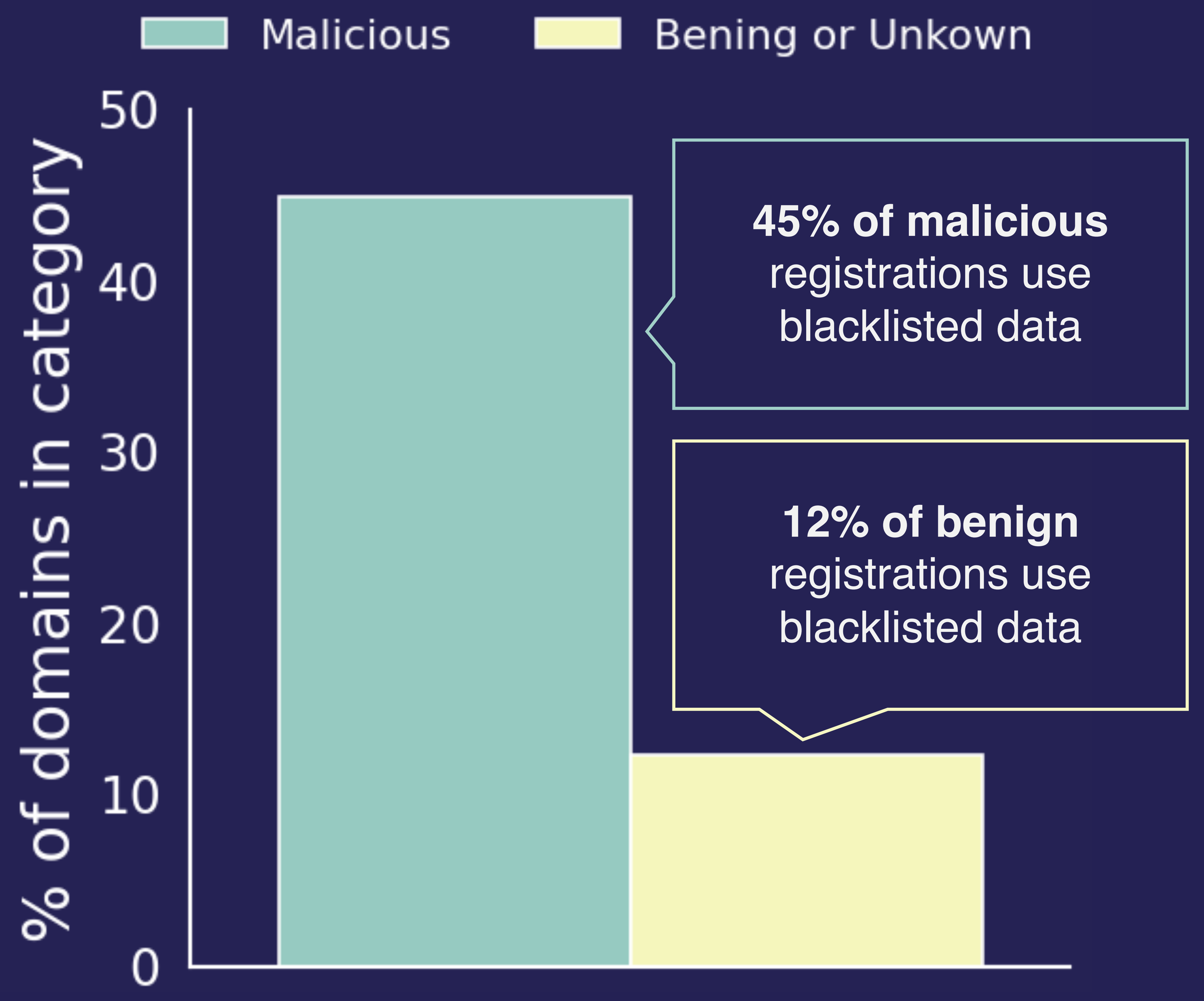
dnsbelgium

Predictor 1: Reuse of WHOIS data

Create a blacklist of reported WHOIS data

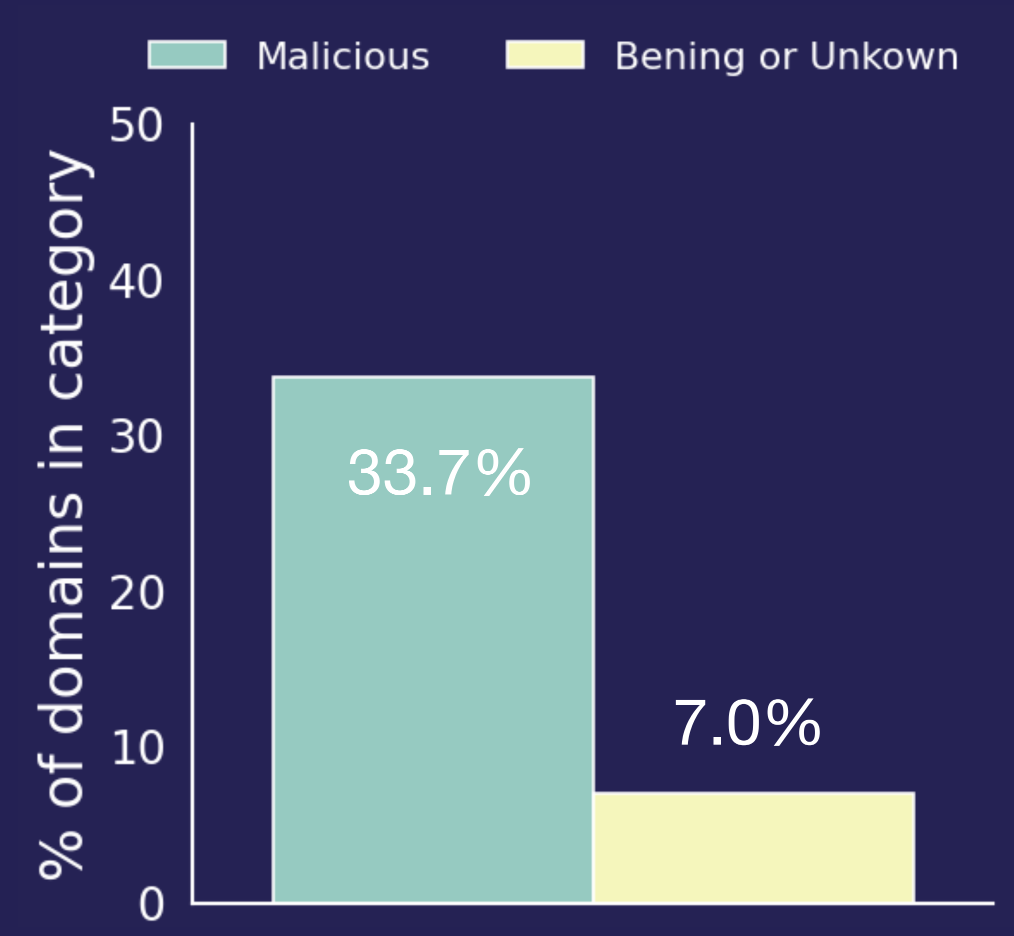
- Registrant's name
- Registrant's address
- Registrant's email
- Registrant's phone
- Registrant's organization
- Registrant's organization VAT

→ Flag registrations that use a blacklisted item

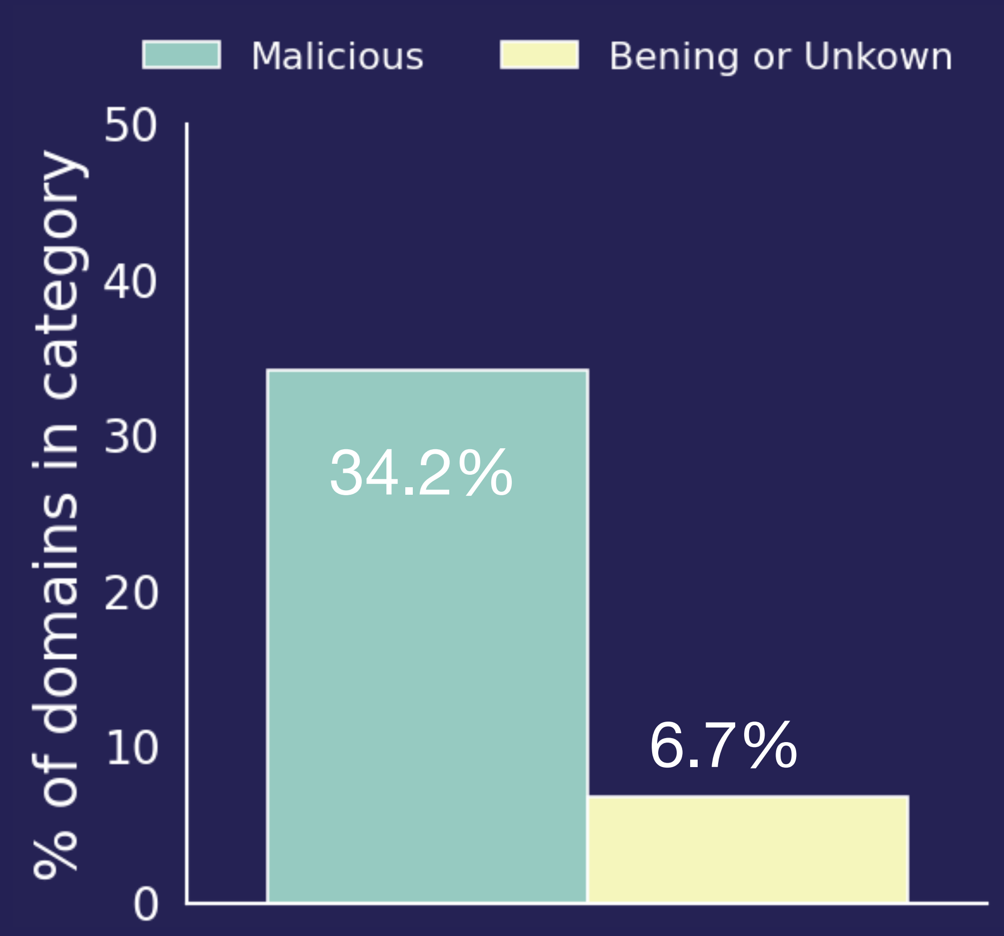


Predictor 1: Reuse of WHOIS data

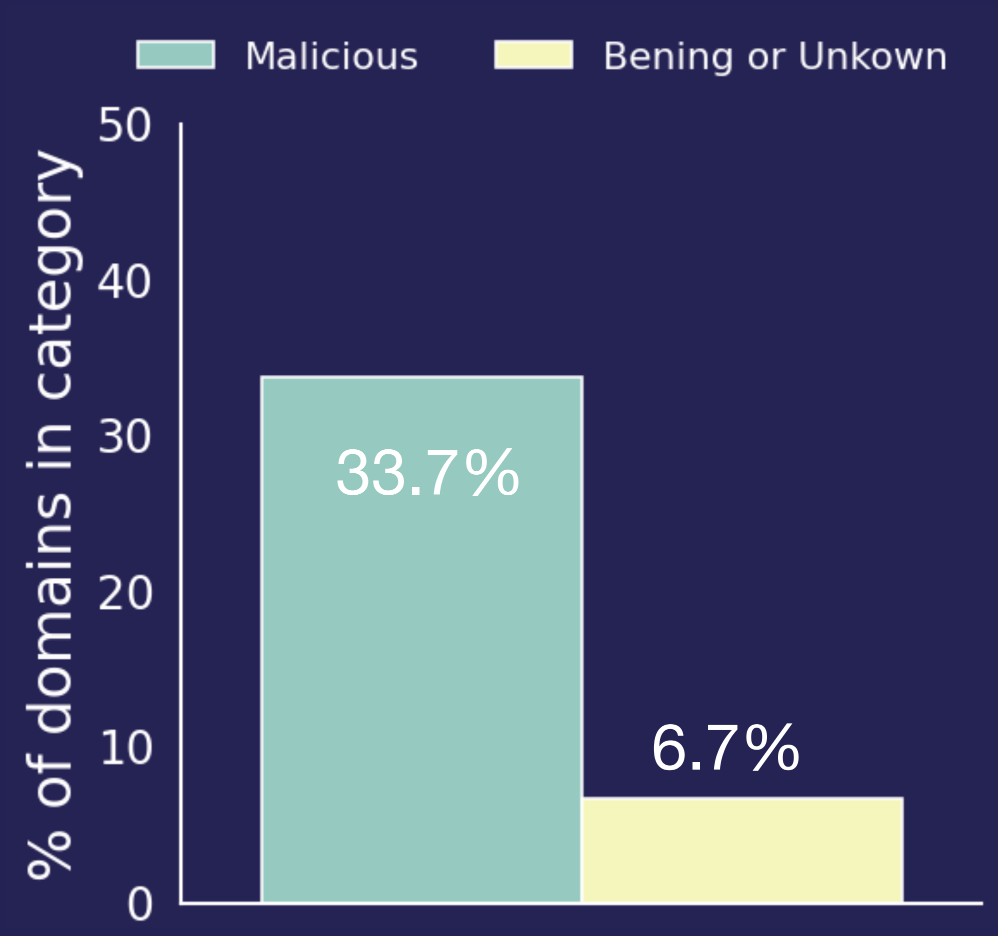
Registrant name



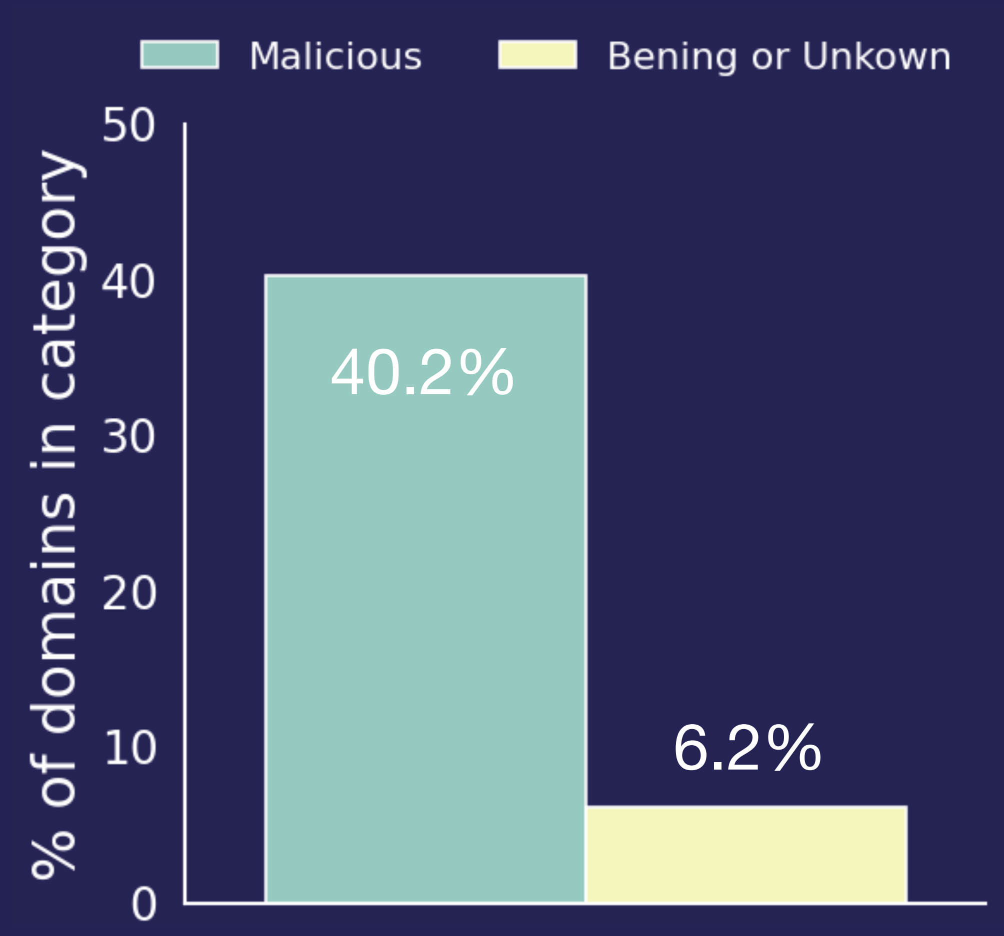
Registrant address



Registrant email



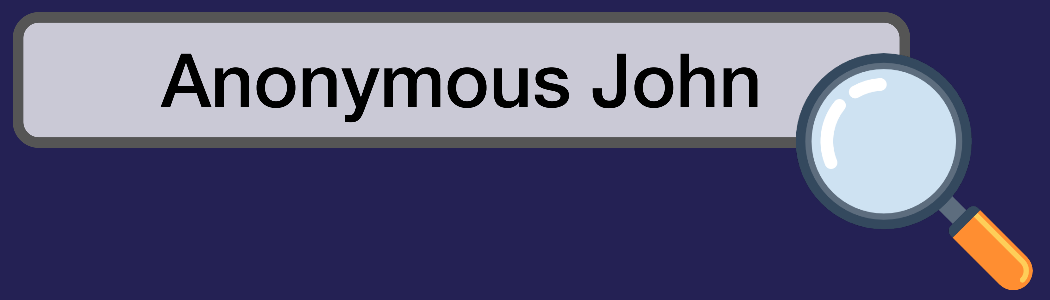
Registrant phone

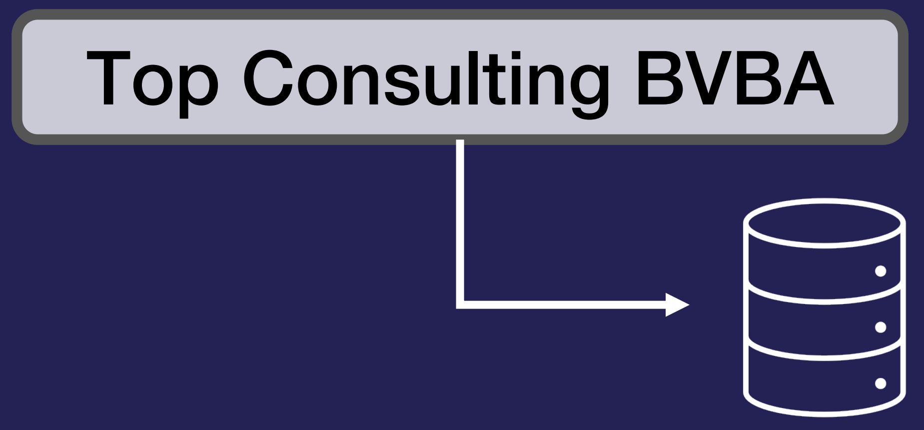


Takes into account the delay between registration and registrant verification

→ WHOIS data is reused over a long period

Predictor 2: Use of fake WHOIS data





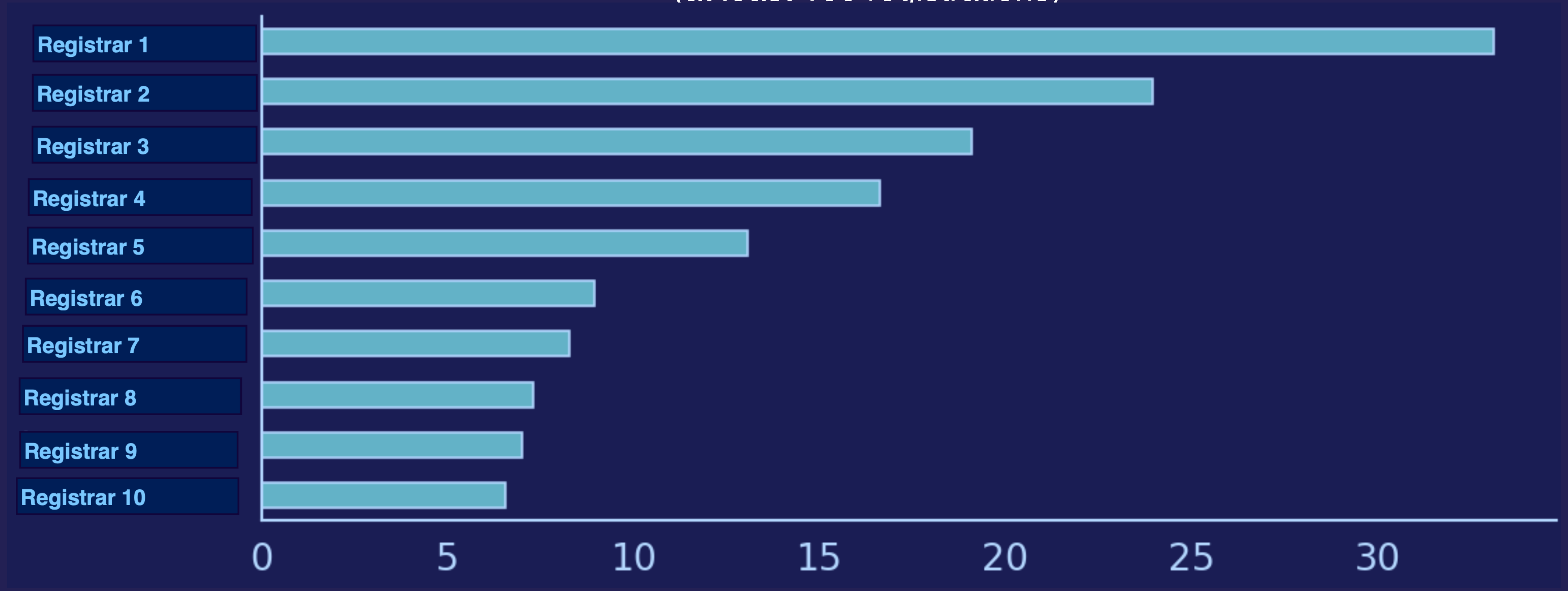
1. **Checks on individual fields**
   - Lexical patterns
   - Keywords: “Unkown”, “John Smith”, …
2. **Consistency between fields**
3. **Validation against external data**
   - Geonames databases
   - Registry of Belgian companies

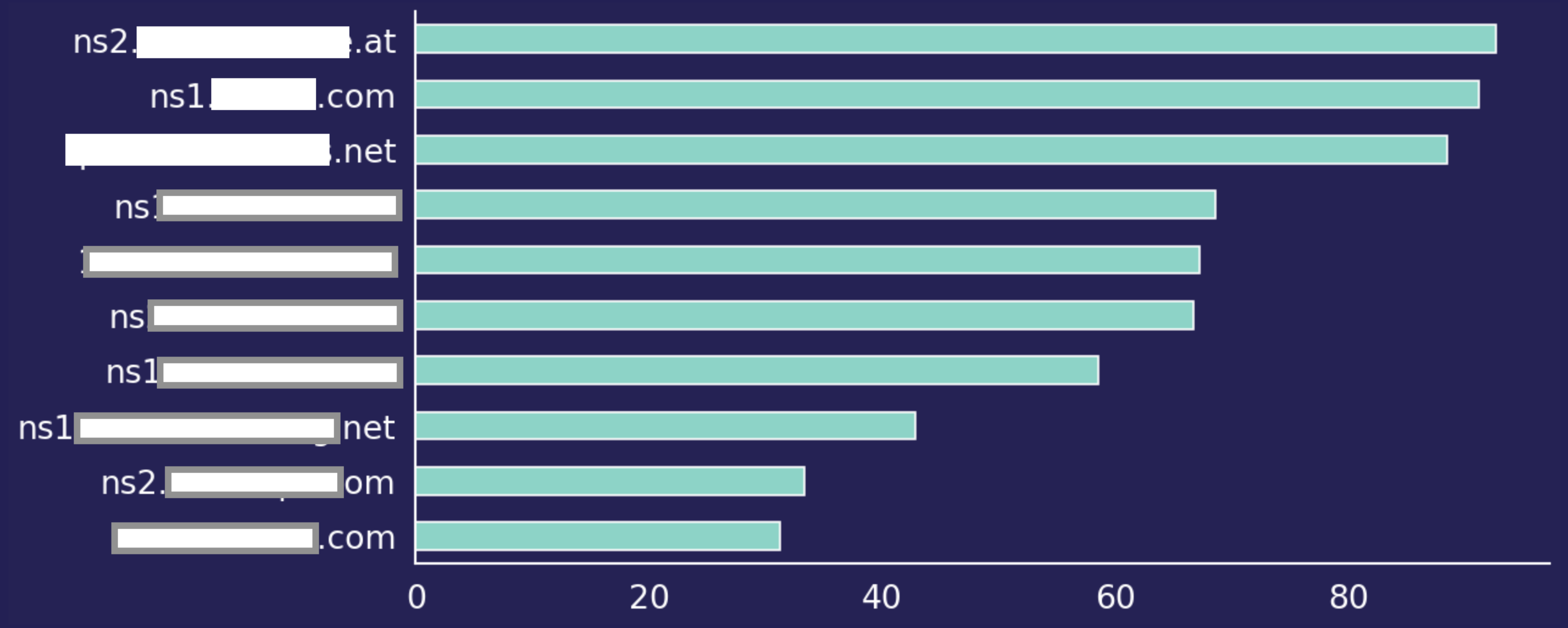
Registrant name **Registrant address** **Registrant mail** **Registrant phone** **Registrant organization**

Predictor 3: Reuse of infrastructure

- Most malicious registrations come from a small group of registrars

Percentage of malicious registrations
(at least 100 registrations)





Predictor 3: Reuse of infrastructure

- *registrar_id* and *nameserver_ip* are high-cardinality categorical features
- One standard approach: "Target encoding"

  For each distinct *category*

  1. Training: compute the percentage of historical malicious registrations for each *category*
  2. Prediction: replace each *category* with the according percentage

- Problem:
  - Risk of over-fitting on infrequent categories
  - Risk of target leakage from the future
  - Distribution might change over time
- Solution: Rolling additive smoothing

dnsbelgium

Rolling Smoothed Reputation Scores

$$\boldsymbol{r} = \frac{\boldsymbol{n} \times \overline{\boldsymbol{x}} + \boldsymbol{m} \times \overline{\boldsymbol{w}}}{\boldsymbol{n} + \boldsymbol{m}}$$

$\bar{x}$ is your estimated mean
n is the number of values you have
$\bar{w}$ is the overall mean
m is the “weight” you want to assign to the overall mean

Rolling Smoothed Reputation Scores

% malicious registrations for a specific registrar over the past N days

$$r = \frac{n \times \bar{x} + m \times \bar{w}}{n + m}$$

$\bar{x}$ is your estimated mean
n is the number of values you have
$\bar{w}$ is the overall mean
m is the "weight" you want to assign to the overall mean

dnsbelgium

Rolling Smoothed Reputation Scores

% malicious registrations for a <u>specific</u> registrar over the past N days

% malicious registrations for the <u>average</u> registrar over the past N days

$$r = \frac{n \times \bar{x} + m \times \bar{w}}{n + m}$$

$\bar{x}$ is your estimated mean
n is the number of values you have
$\bar{w}$ is the overall mean
m is the “weight” you want to assign to the overall mean

dnsbelgium

Rolling Smoothed Reputation Scores

% malicious registrations for a specific registrar over the past N days

% malicious registrations for the average registrar over the past N days

$$r = \frac{n \times \bar{x} + m \times \bar{w}}{n + m}$$

$\bar{x}$ is your estimated mean
n is the number of values you have
$\bar{w}$ is the overall mean
m is the “weight” you want to assign to the overall mean

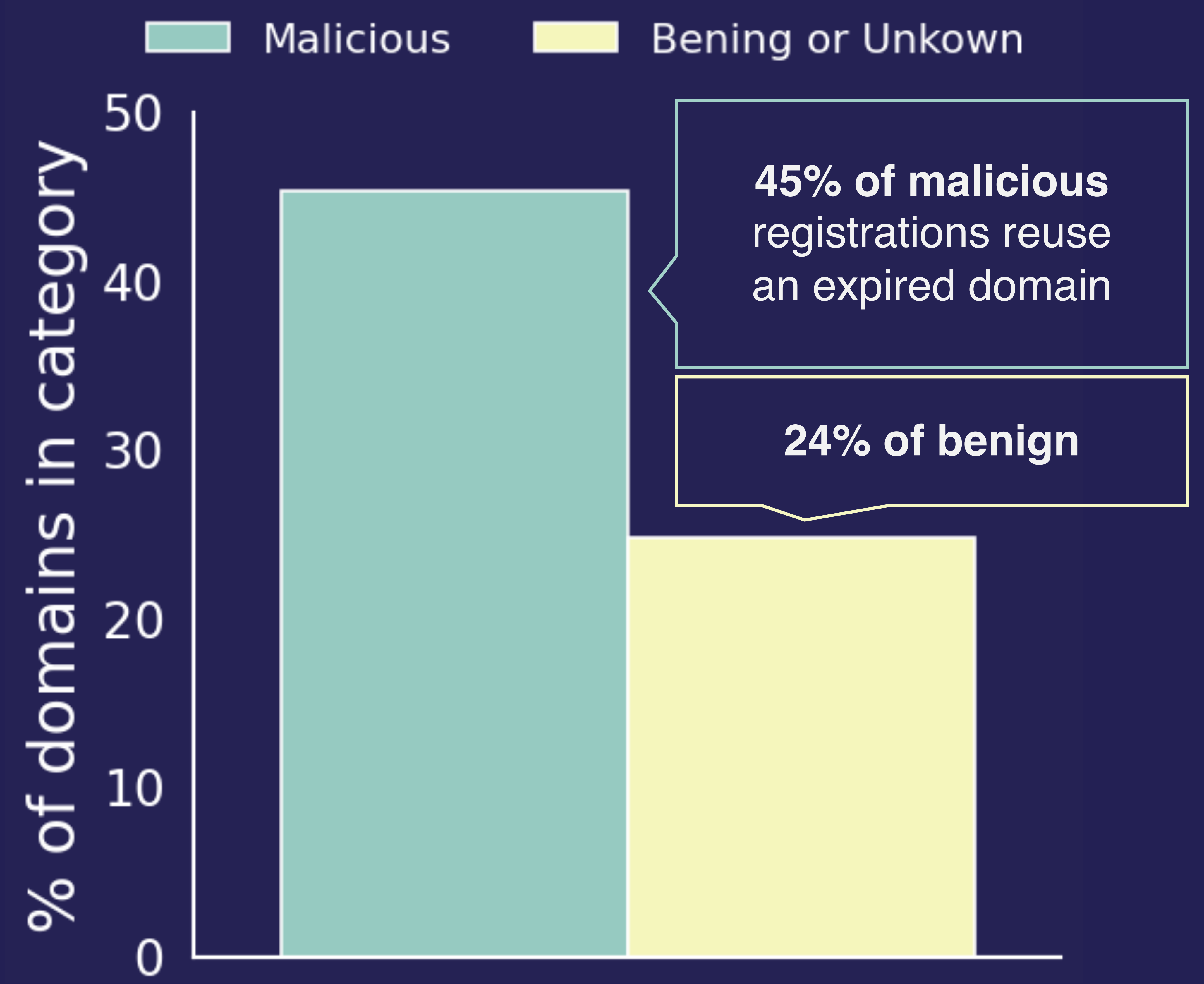
Intuition: there must be at least m values for the sample mean to overtake the global mean

We compute these for the previous 7 and 30 days

Predictor 4: Reuse of domains

Data from previous registration

- Previous registrar
- Re-registration latency
  (brand new, drop-catch, retread)
- Number of BAD WHOIS cases



dnsbelgium

Predictor 5: Similarities between domains

Benign N-gram counts

How often does each 4-gram occur in benign domains?

4-gram	Count	%
tion	21275	2.03%
shop	12450	1.19%
nder	11542	1.11%
ande	11404	1.10%
atio	10604	1.02%
cons	10381	0.99%
ting	10247	0.98%
eren	9943	0.95%
elle	9396	0.90%
belg	9327	0.89%

x

Malicious N-gram counts

How often did each 4-gram occur in malicious domains over the past N days?

4-gram	Count	%
tion	90	0.19%
cons	63	0.13%
ande	59	0.13%
serv	58	0.12%
ervi	56	0.12%
outu	56	0.12%
belg	56	0.12%
yout	55	0.12%
vice	54	0.12%
elgi	53	0.11%

=

Reputation scores

Which 4-grams are over-represented in malicious domains?

4-gram	Reputation	Example
caix	11.12	caixabank.be
aixa	6.07	caixa-bank.be
iccu	5.78	uscciccu.be
outu	2.05	httpsssyoutu.be
wyou	1.80	wwwyoutube.be
yout	1.68	nfswyoutu.be
exus	1.62	connexusnl.be
isth	1.52	calisthenicspark.be
hose	1.25	hosestore.be
mazo	1.06	amazongiftcard.be

Raw Labels Overview

1.080.633 registrations

27,836 (2.6%) BAD WHOIS

10,911 (1.0%) GOOD WHOIS

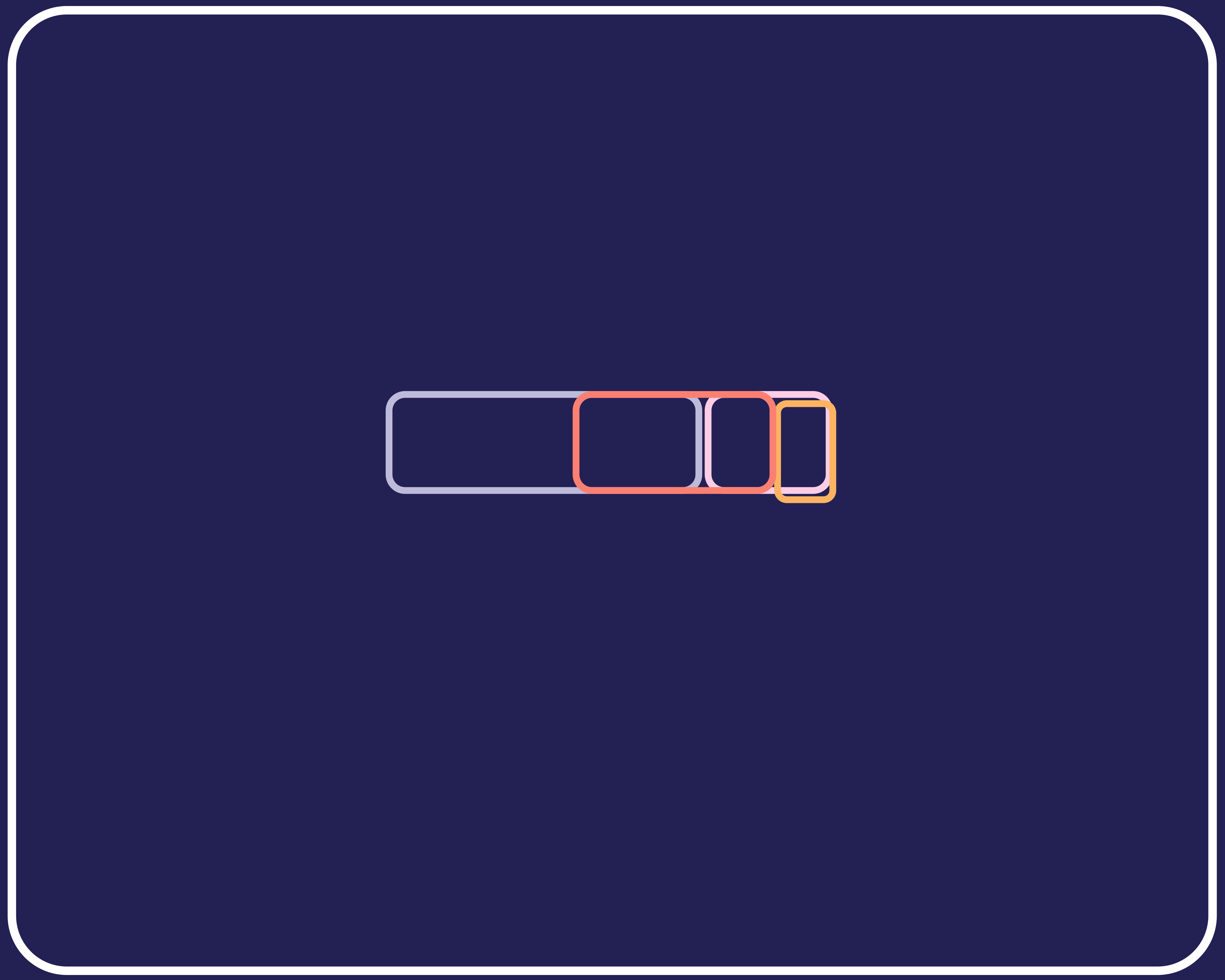
> Manually verified based on rules and eyeball test

17,706 (1.6%) MALICIOUS

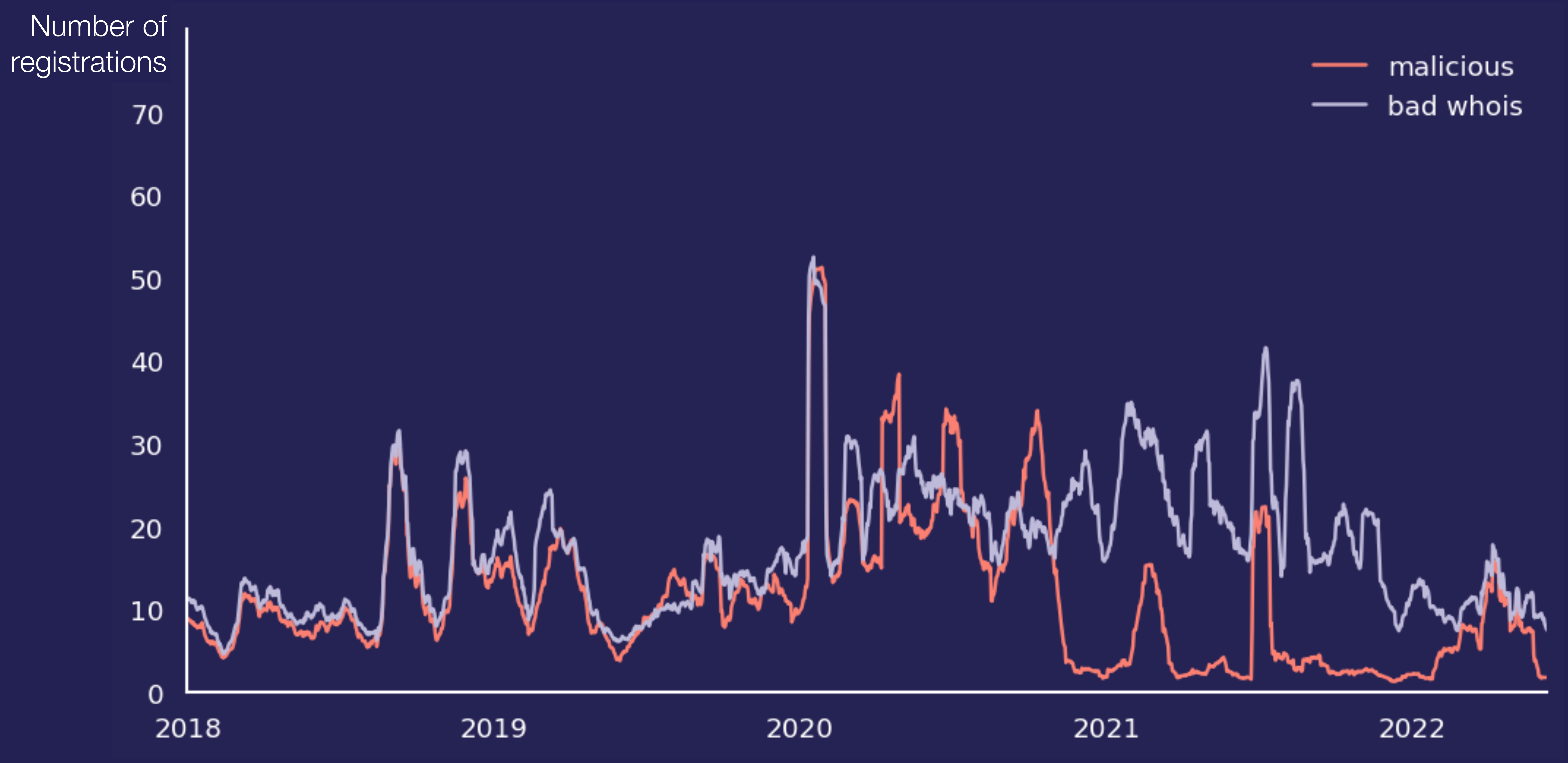
4,989 (0.5%) BENIGN

> Manually verified based on rules and eyeball test + blacklists

1,041,087 (96.3%) UNKOWN

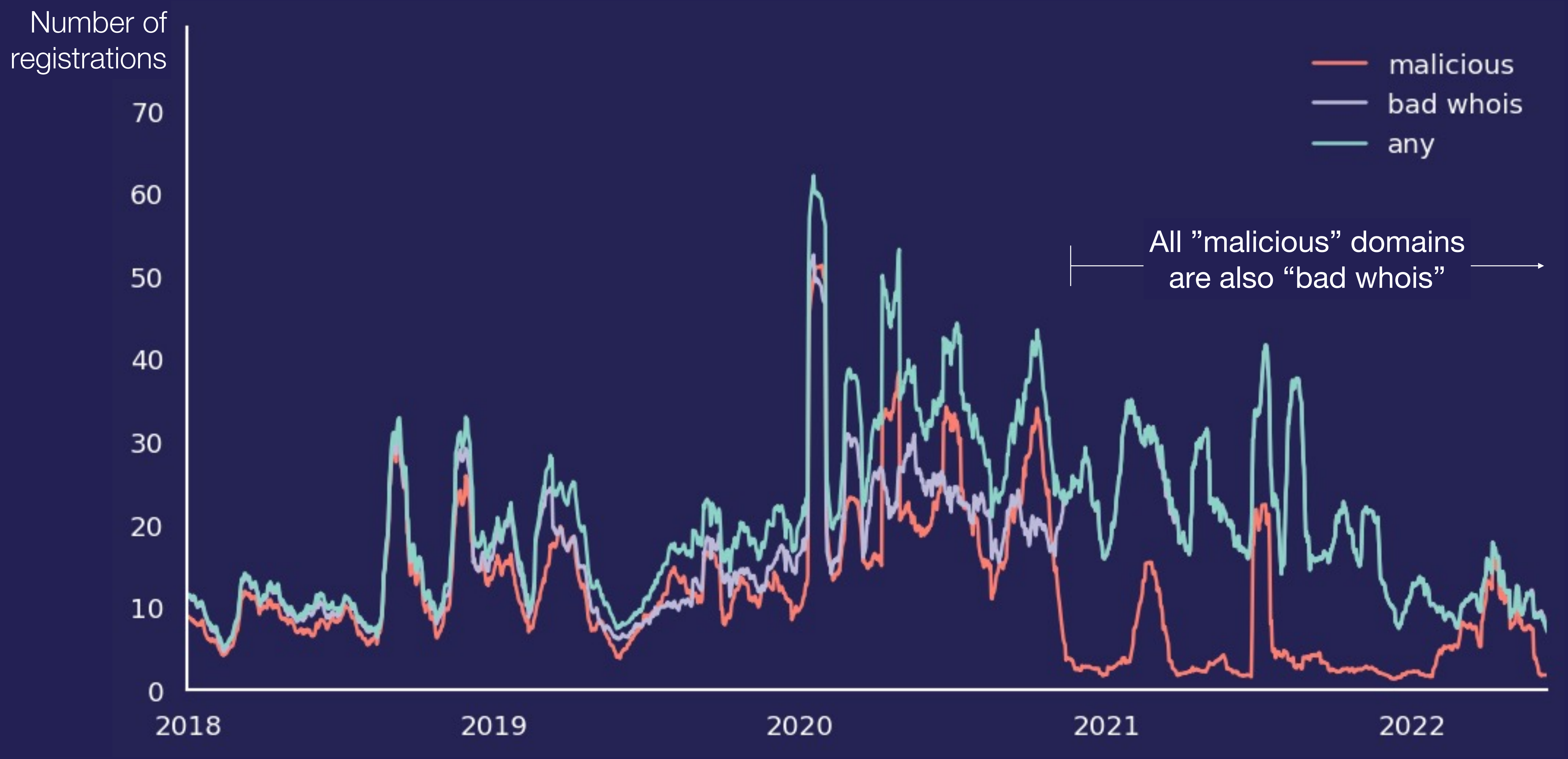


Ground Truth Labeling Shift



dnsbelgium

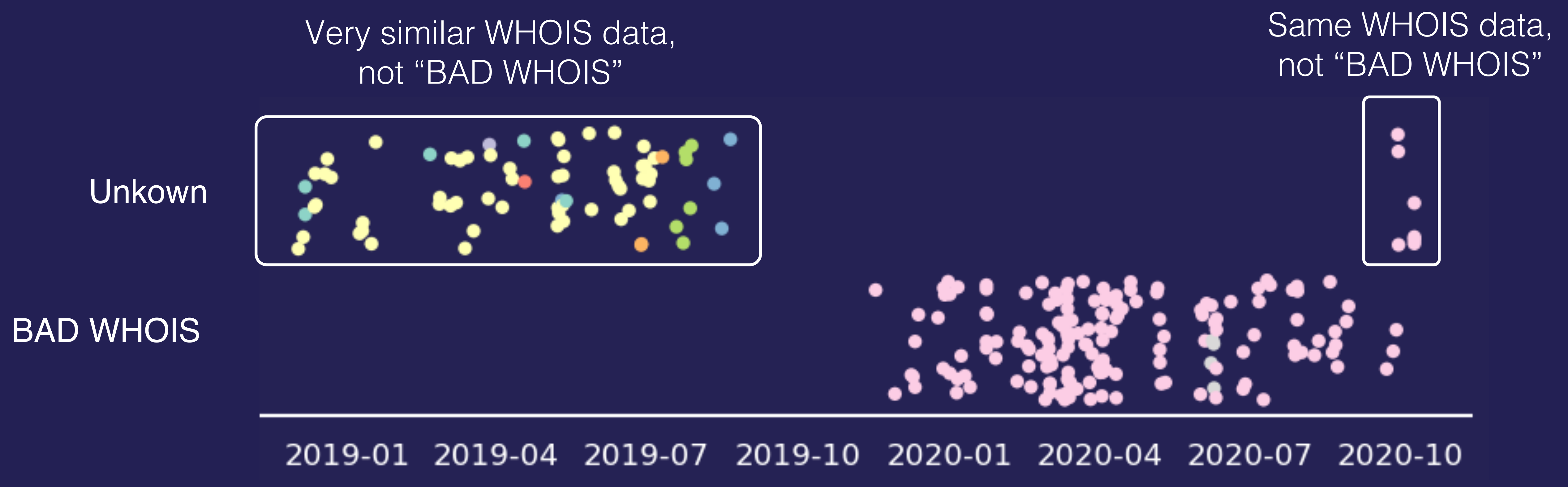
Ground Truth Labeling Shift

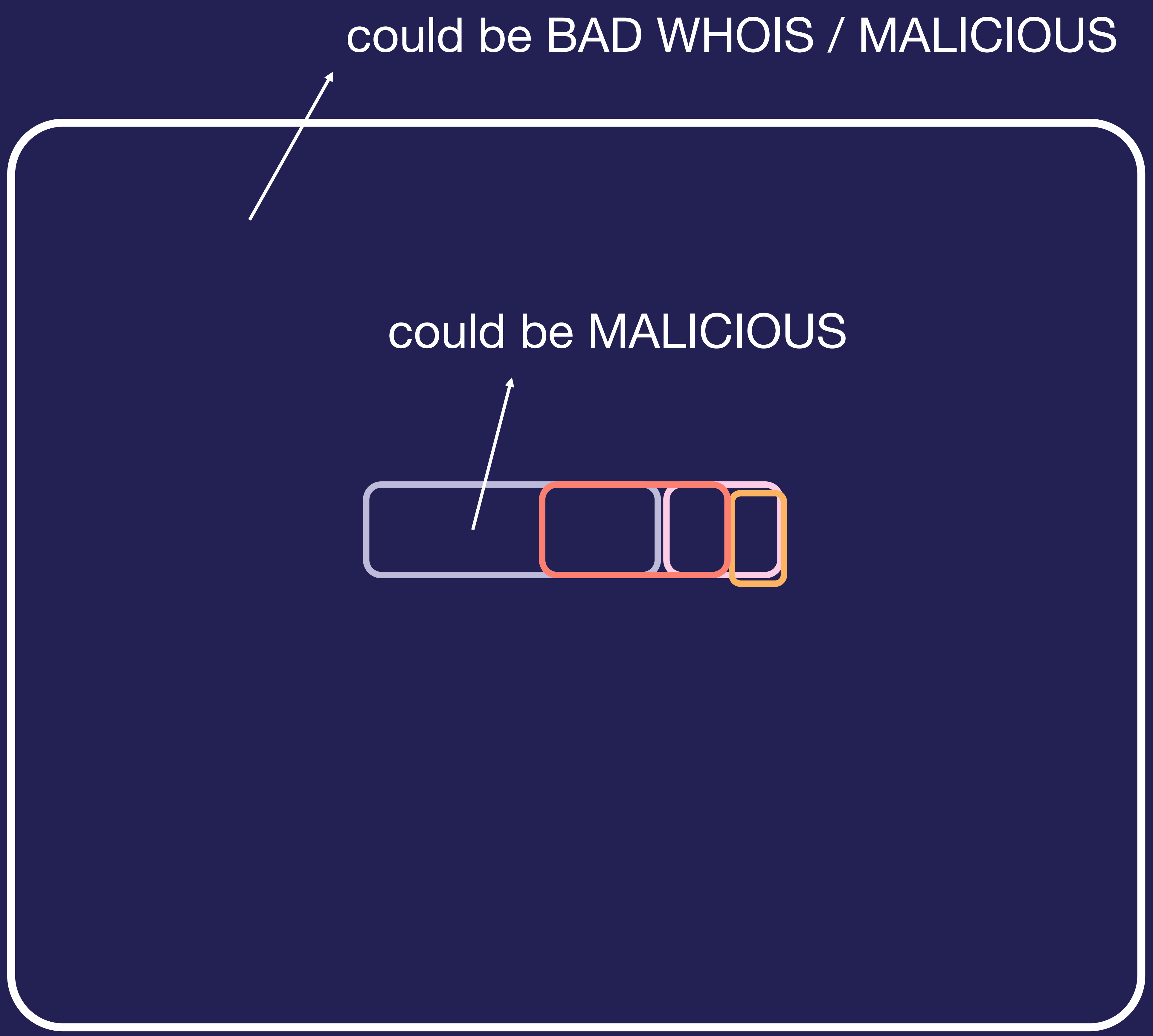


Ground Truth Labeling Errors

Labels are incomplete

Example: 244 registrations by same registrant





Mislabeled domains cause trouble

- Confuse the model during training
- Masquerade as false positives during evaluation

Labels can be combined in several ways

~~Ground Truth~~

Weak Labels

IS MALICIOUS	count	pct
True	17,706	1.64%

IS BAD WHOIS	count	pct
True	27,836	2.58%

No detected incidents 30 days after registration

Same WHOIS data was used in a previous malicious registration

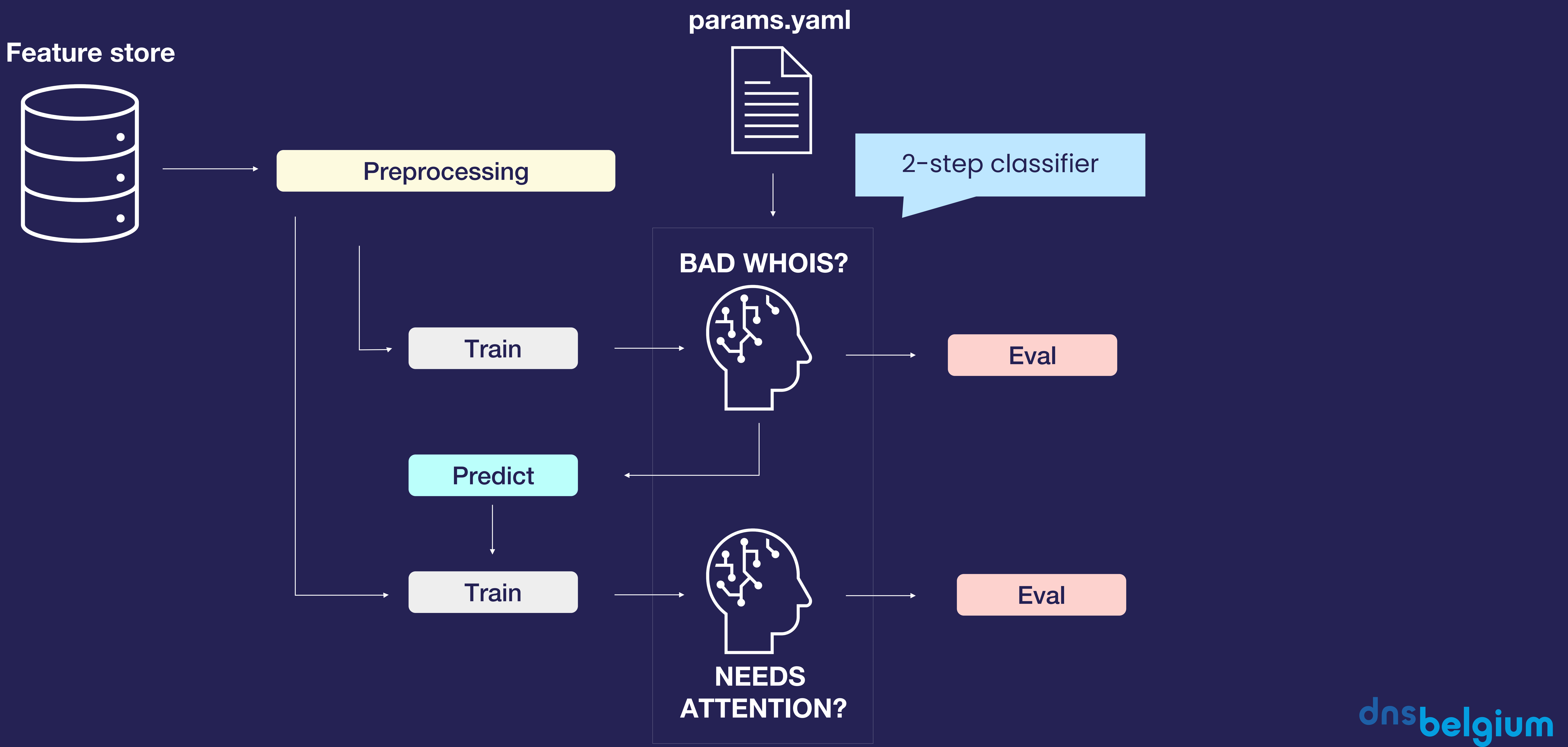
Domain name contains critical keyword (e.g., bank name)

Training Labels

is_bad_whois

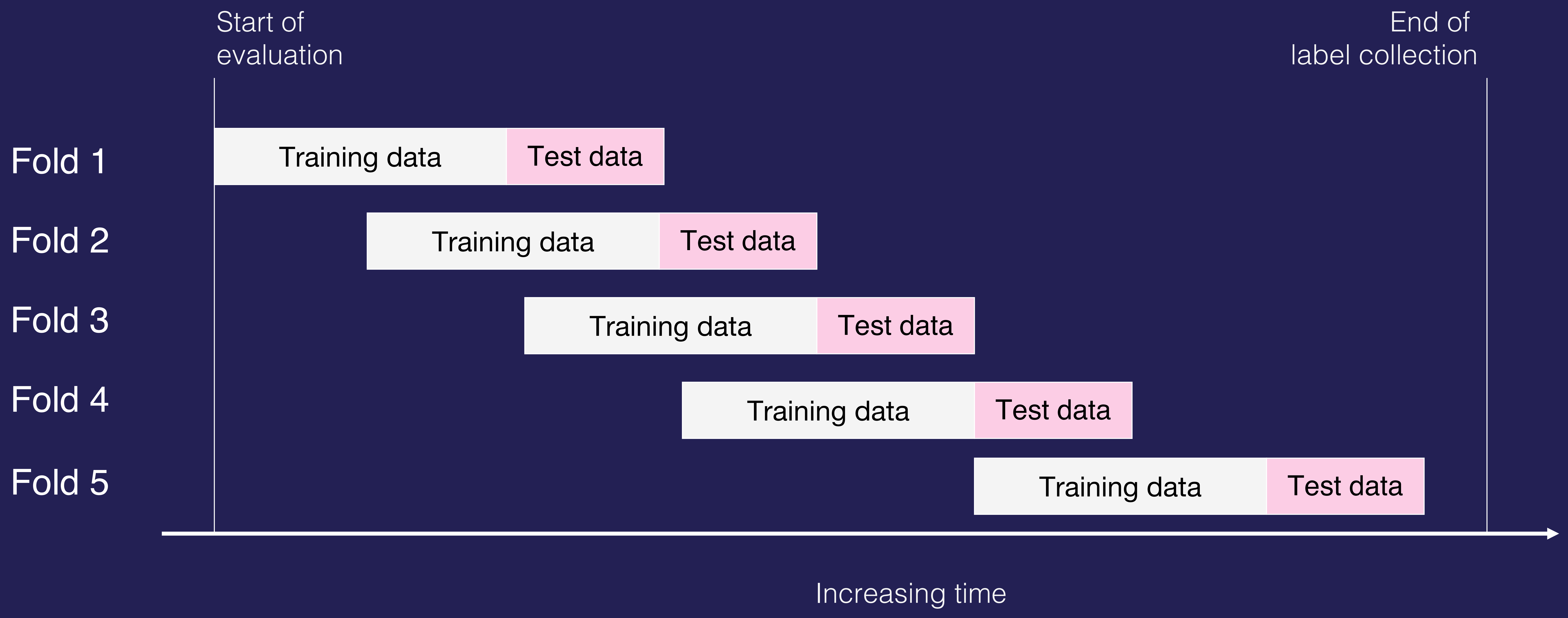
needs_attention

Machine Learning Pipeline



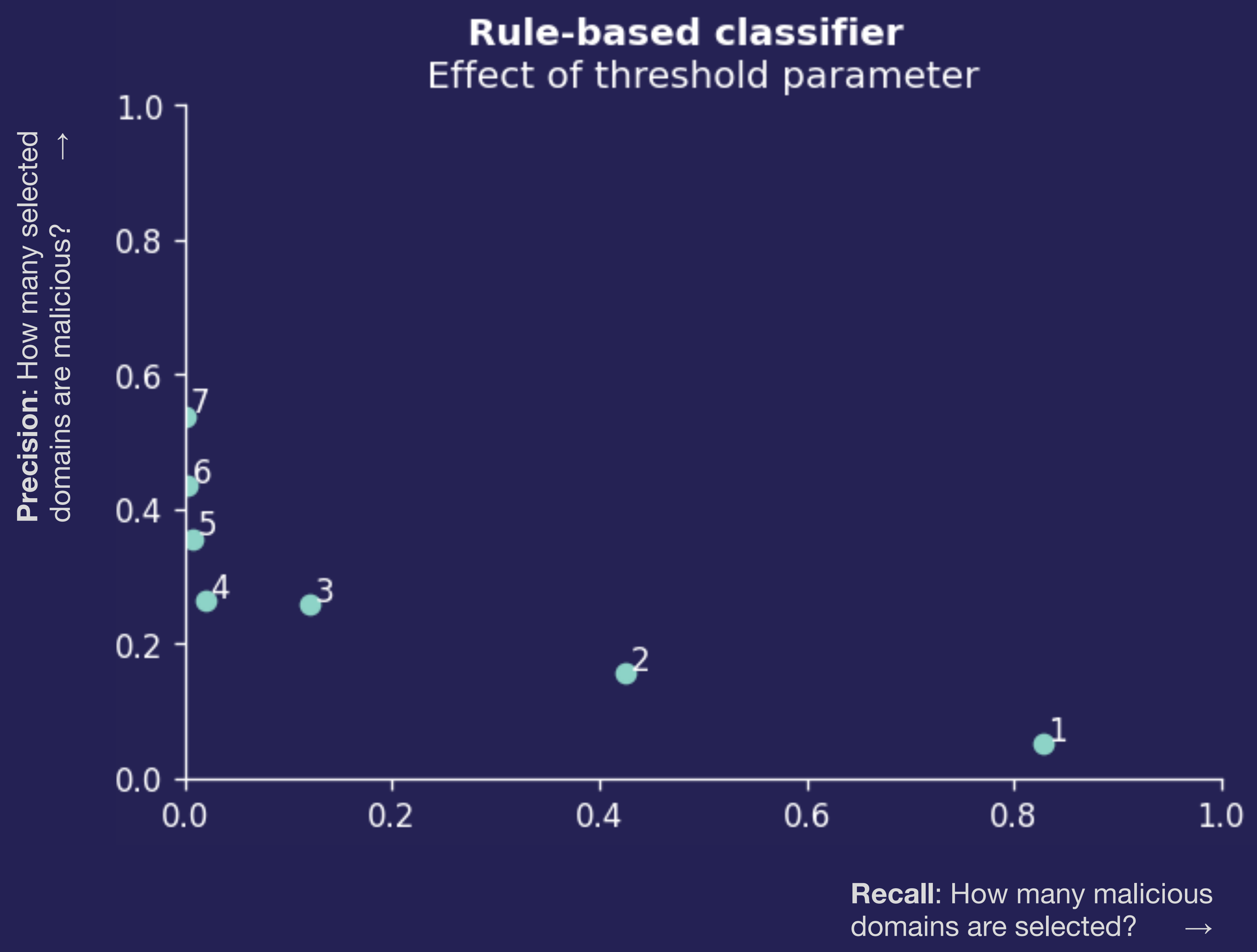
Experimental design

Evaluation simulates passage of time



Expert-based classifier

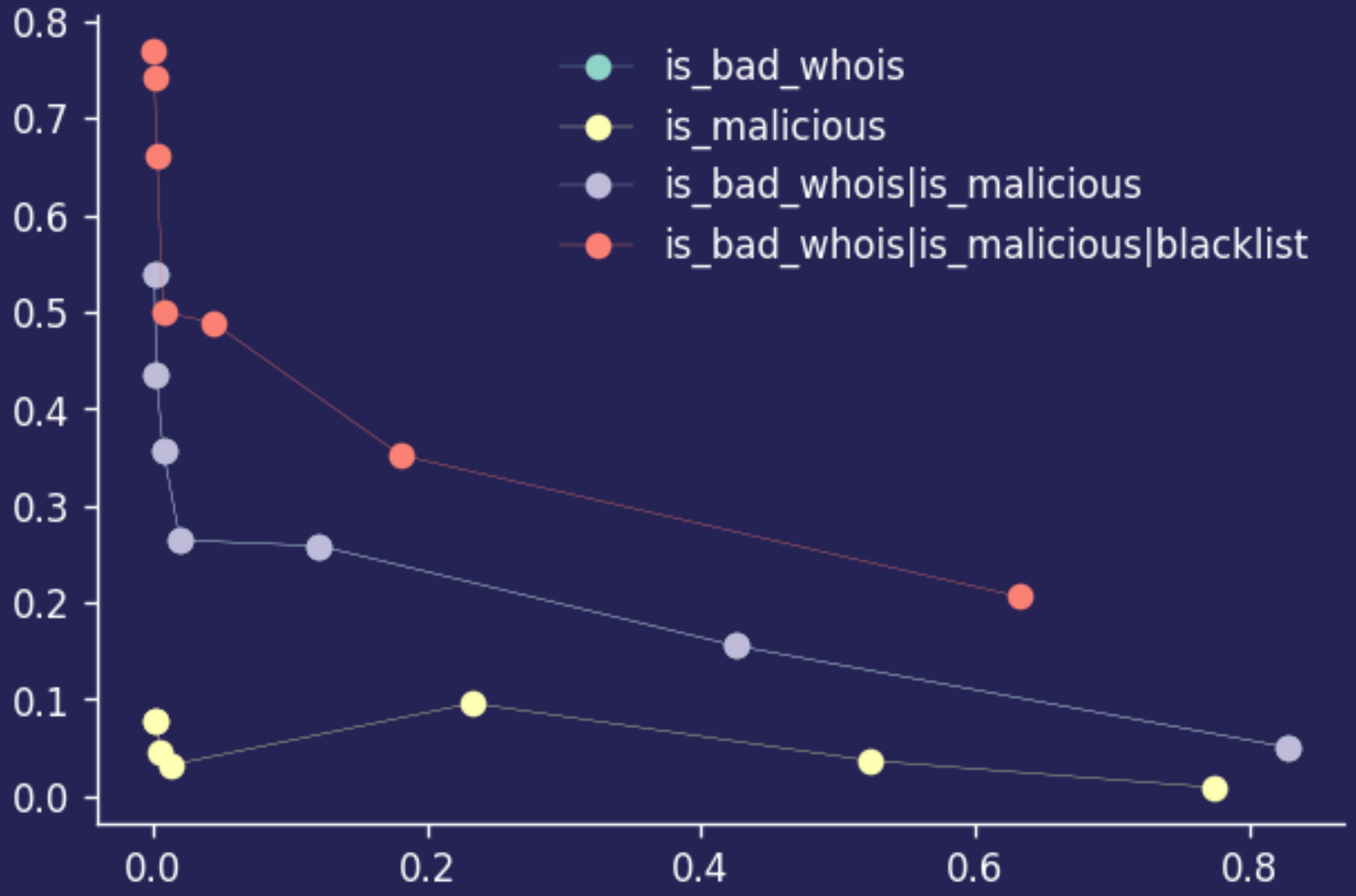
Suspicious when registrations get a score of ≥ X points



Data: Jan 2021 – Mar 2022
is_bad_whois | is_malicious

Expert-based classifier on different labels

Precision: How many selected domains are malicious? ↑

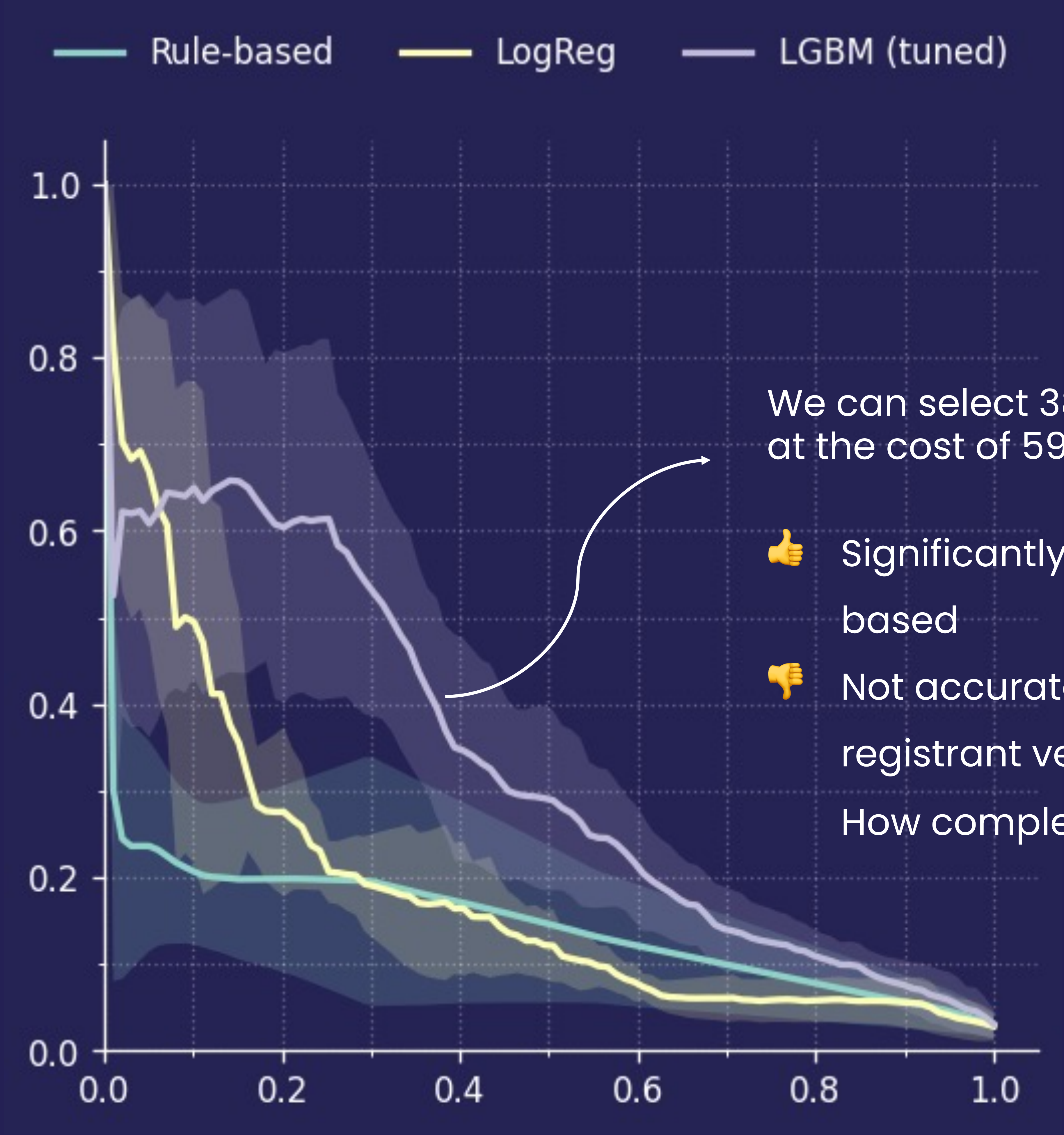


Recall: How many malicious domains are selected? →

Data: Jan 2021 – Mar 2022

BAD WHOIS Classifier

Precision: How many selected domains are malicious? ↑



Recall: How many malicious domains are selected? →

We can select 38% of the BAD WHOIS domains, at the cost of 59% false positives

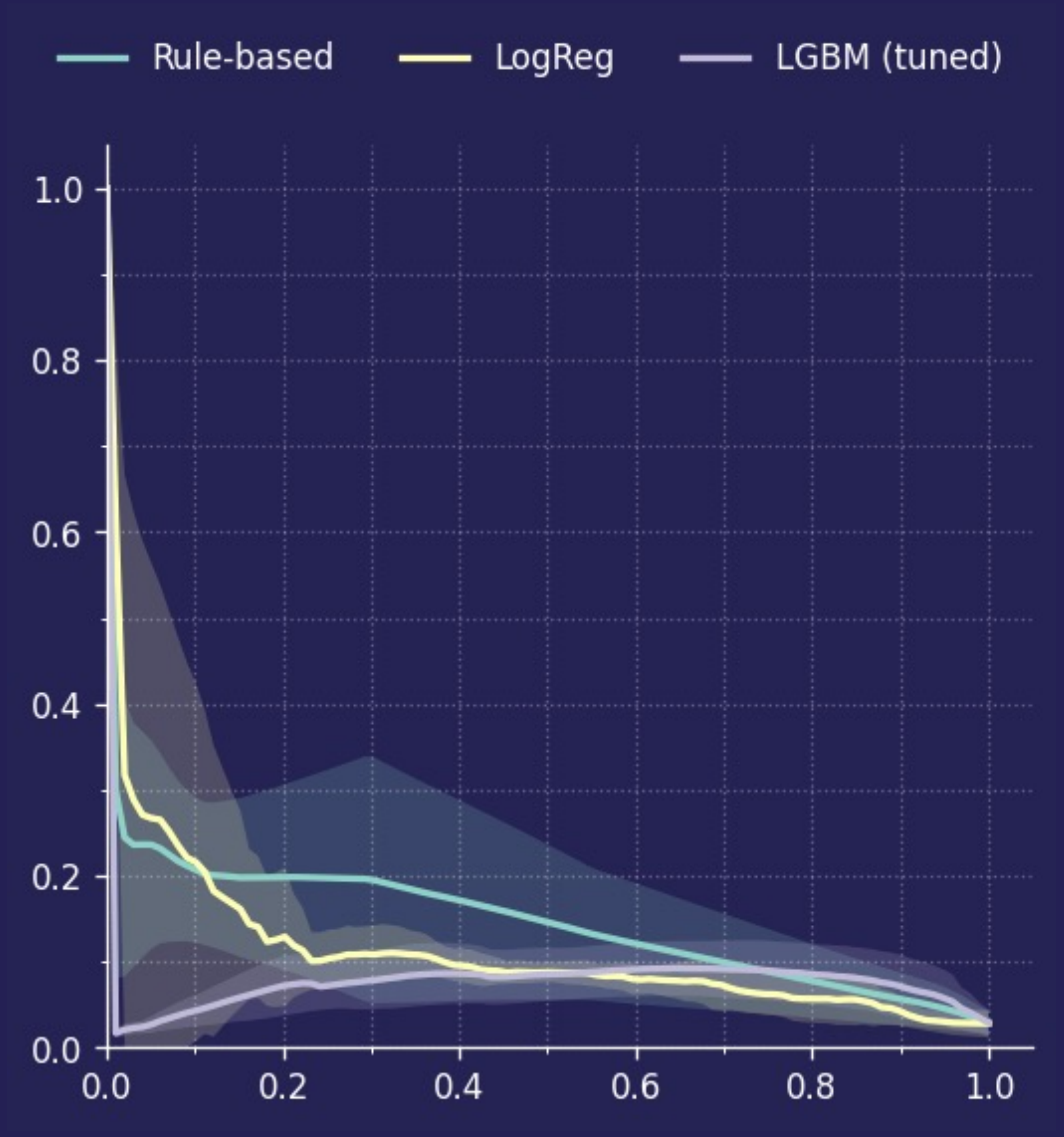
- 👍 Significantly more accurate than rule-based
- 👎 Not accurate enough to fully automate registrant verification
- How complete is the ground truth?

Data: Jan 2021 – Mar 2022

Needs Attention Classifier

BAD_WHOIS
OR **MALICIOUS**
OR **BLACKLIST**

Precision: How many selected domains are malicious? ↑



BAD_WHOIS
OR **MALICIOUS**

Many registrations with blacklisted contact info are not flagged!

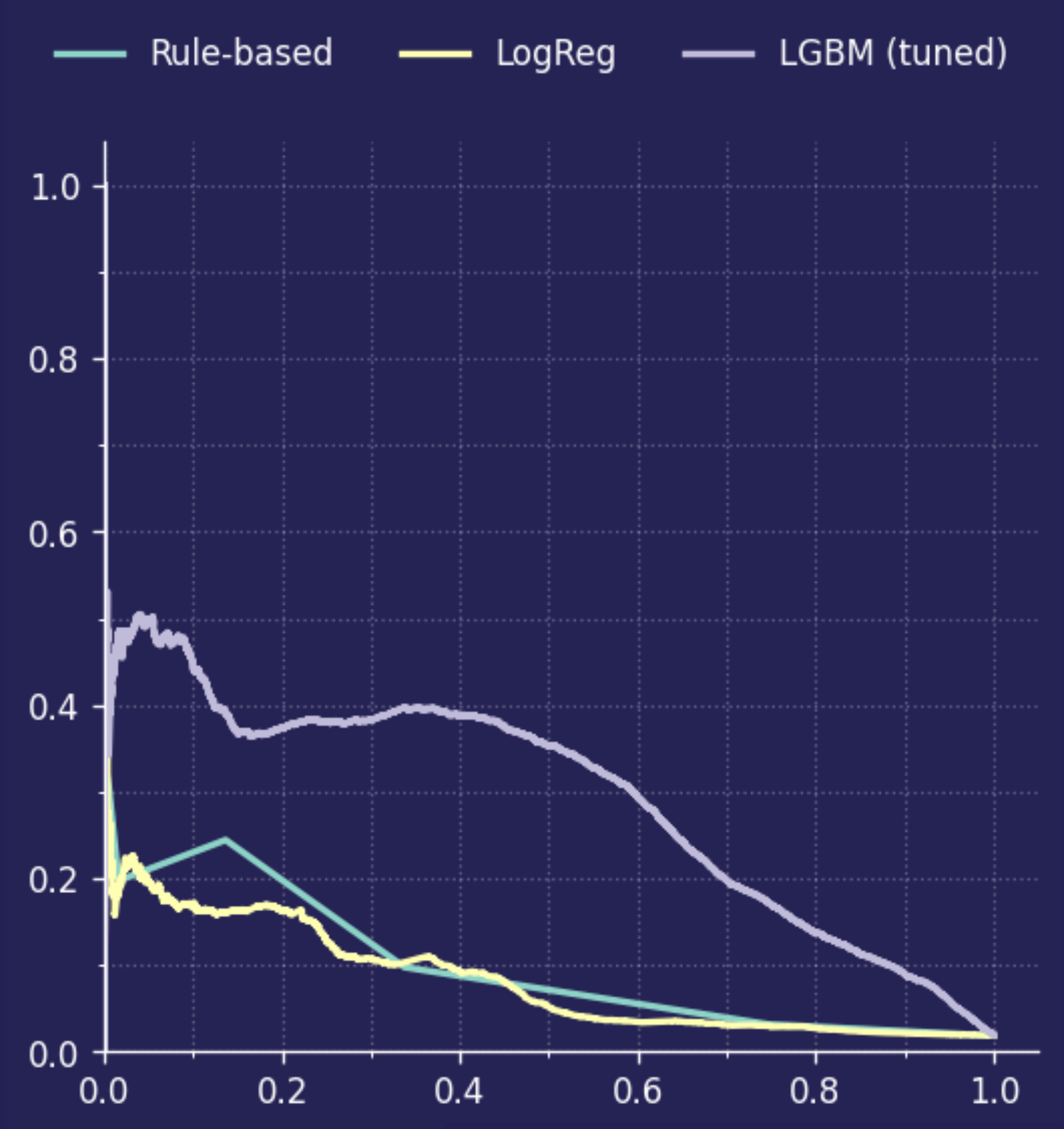
Recall: How many malicious domains are selected? →

Data: Jan 2021 – Mar 2022

Needs Attention Classifier

(only on non-blacklisted registrations)

Precision: How many selected domains are malicious? ↑

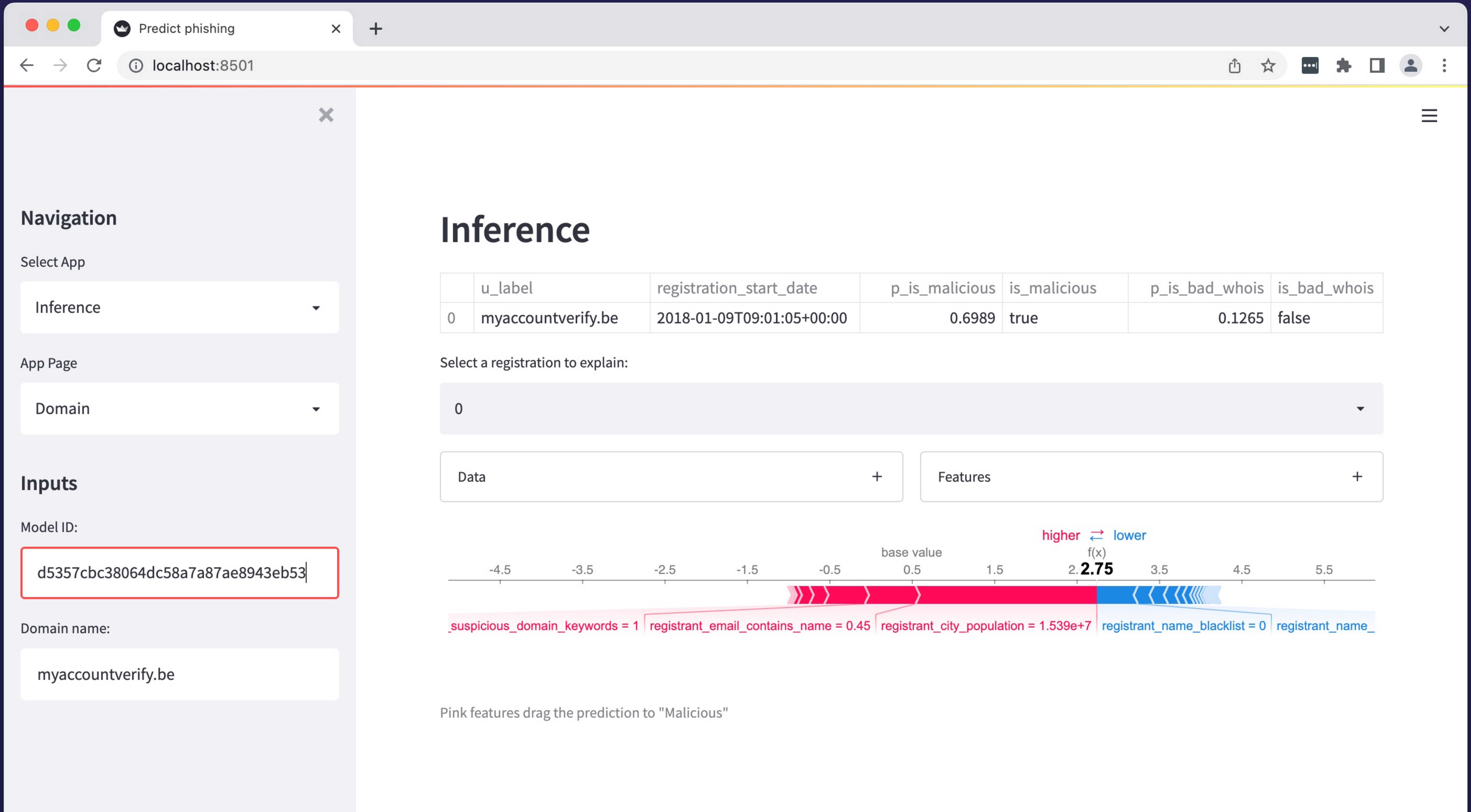


Recall: How many malicious domains are selected? →

Data: Jan 2021 – Mar 2022

dnsbelgium

SHAP values enable interpretable predictions



What's next?

1. Abusive registrations have distinct properties [1]

→ **Can we automatically identify malicious registrations at registration time?**

2. Abusive traffic has distinct properties

- Auto-generated vs user-driven [2]
- Synchronized with known malicious traffic [3]

→ **Can we automatically identify malicious registrations shortly after registration?**

[1] Hao et al. PREDATOR: Proactive recognition and elimination of domain abuse at time-of-registration
[2] Robberechts. Query Log Analysis: Detecting anomalies in DNS traffic at a TLD resolver
[3] Spooren et al. Premadoma: An Operational Solution for DNS Registries to Prevent Malicious Domain Registrations

dnsbelgium

Take away messages

- Abusive registrations have distinct properties
    1. The same / similar registration details
    2. Provide fake contact info
    3. Reuse infrastructure
    4. Retread domains
    5. Use similar domains

- Machine learning outperforms a rule-based system

- Ground truth is tricky
    - Bias towards rule-based system
    - Incompleteness of ground truth makes training and analysis hard

Thanks!

Any questions?