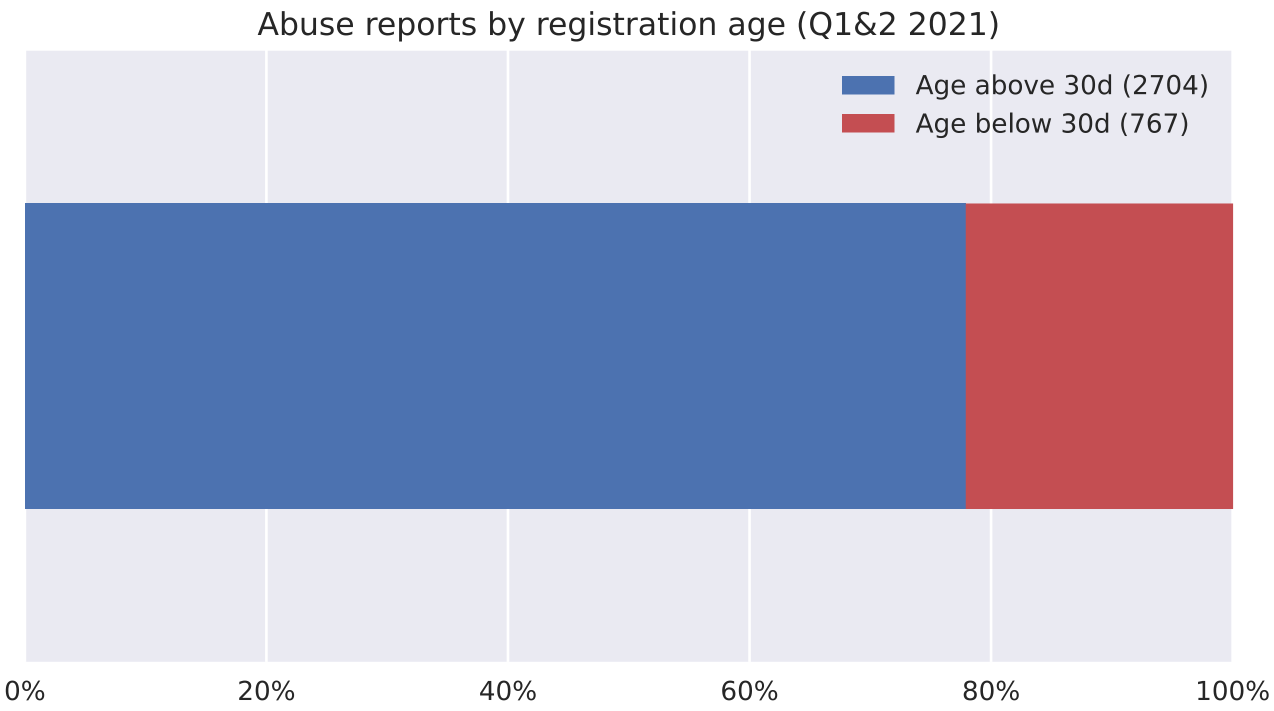
Automated detection of malicious .nl-registrations

A case study for responsible ML

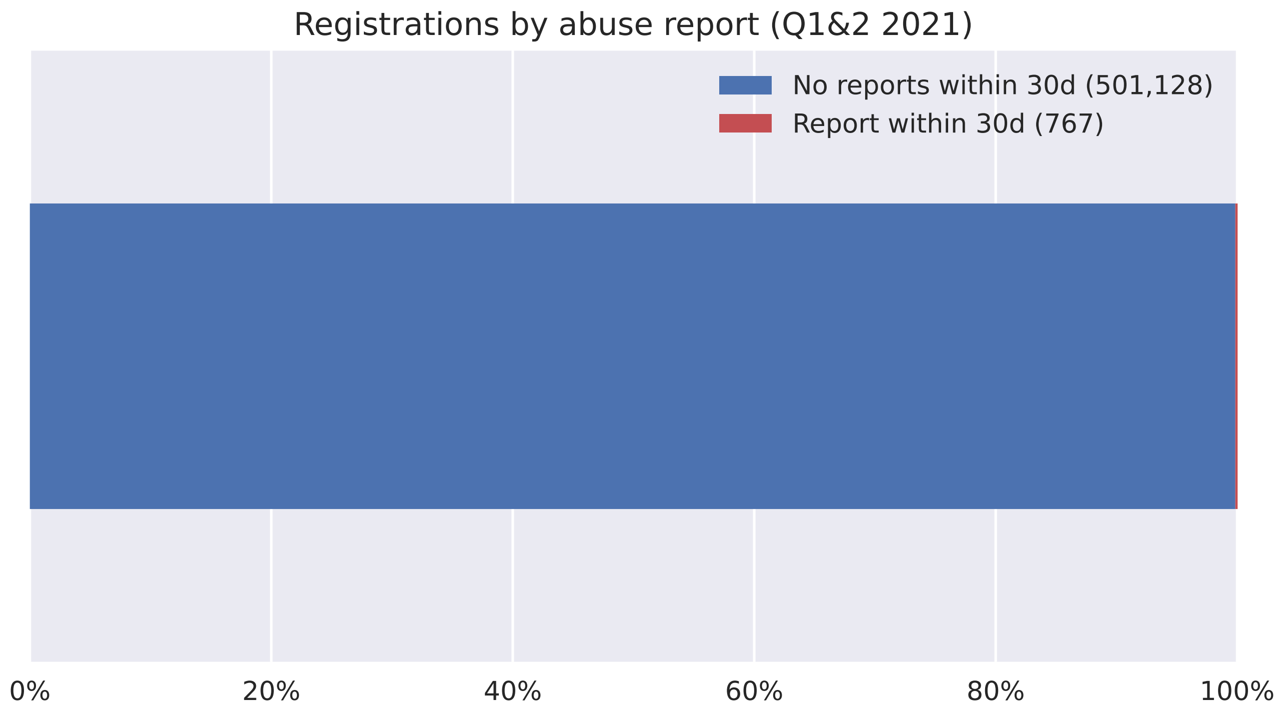
Thymen Wabeke | ONE Conference

18 October 2022

Abuse regularly involves recent registrations



Finding the needle in the haystack



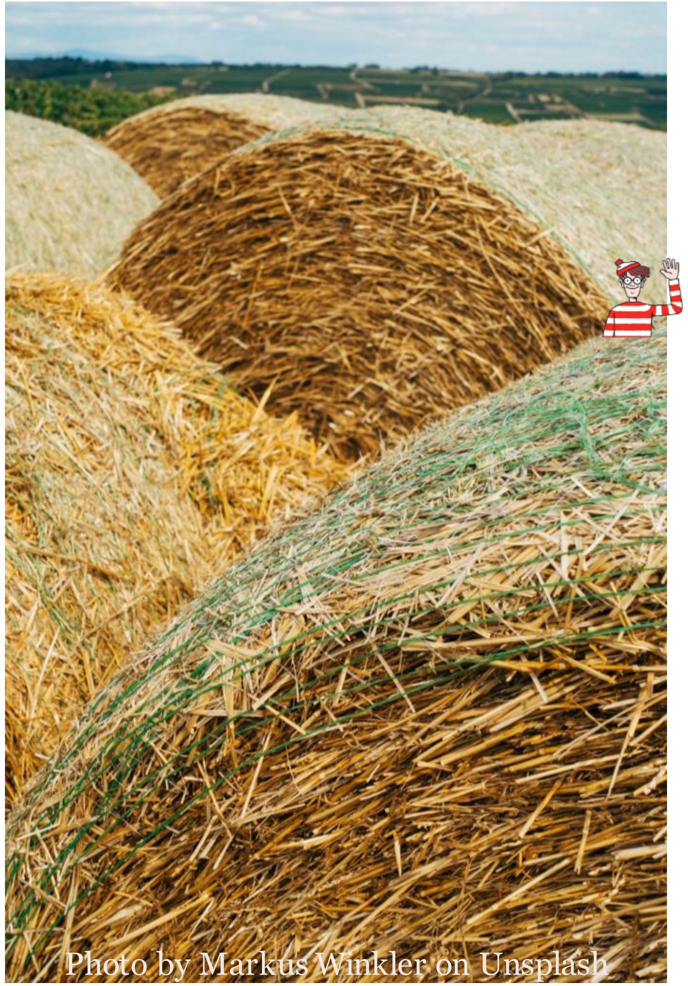
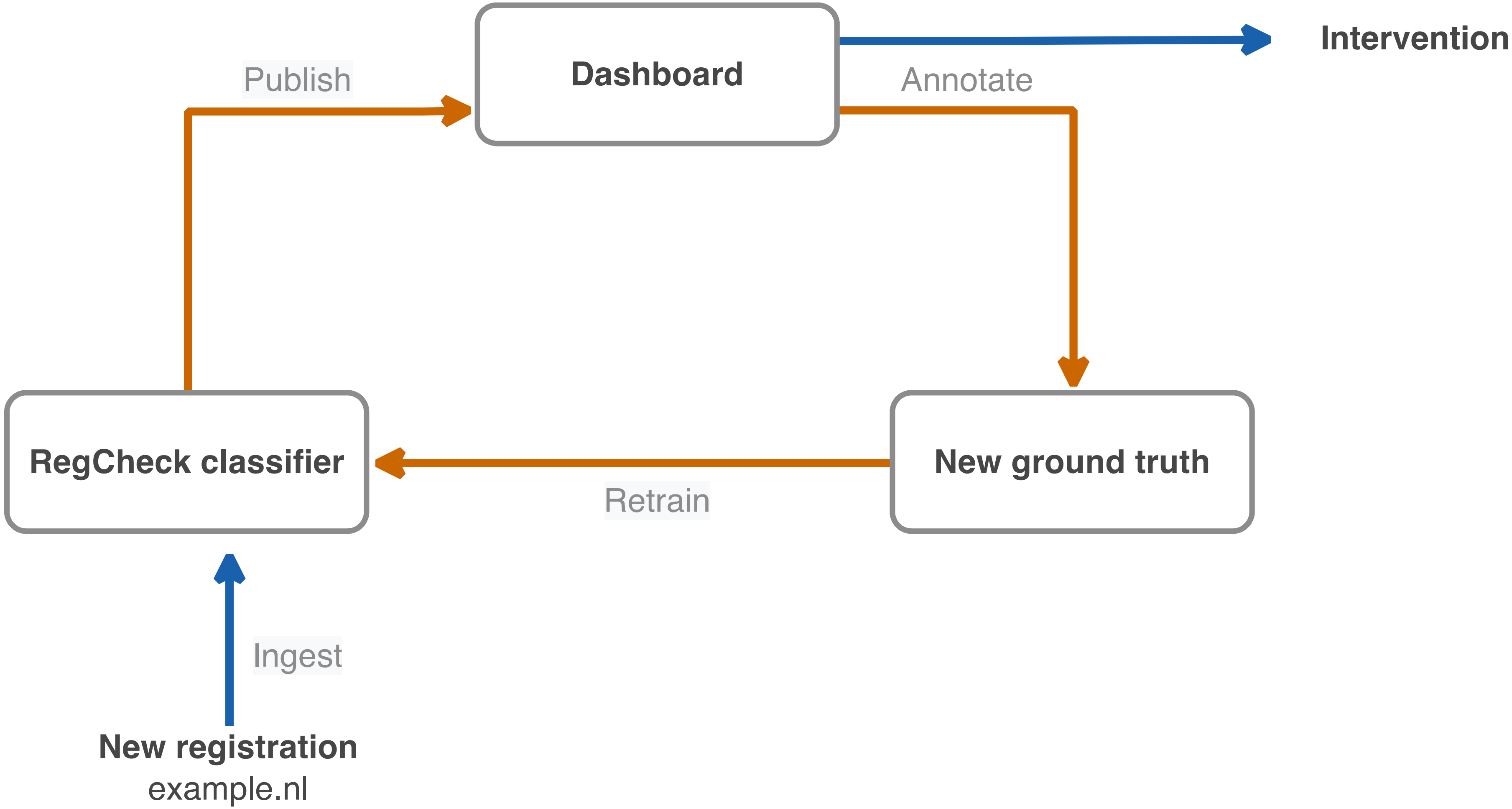


Photo by Markus Winkler on Unsplash

SIDN LABS

Goal: Identify registrations analysts should review

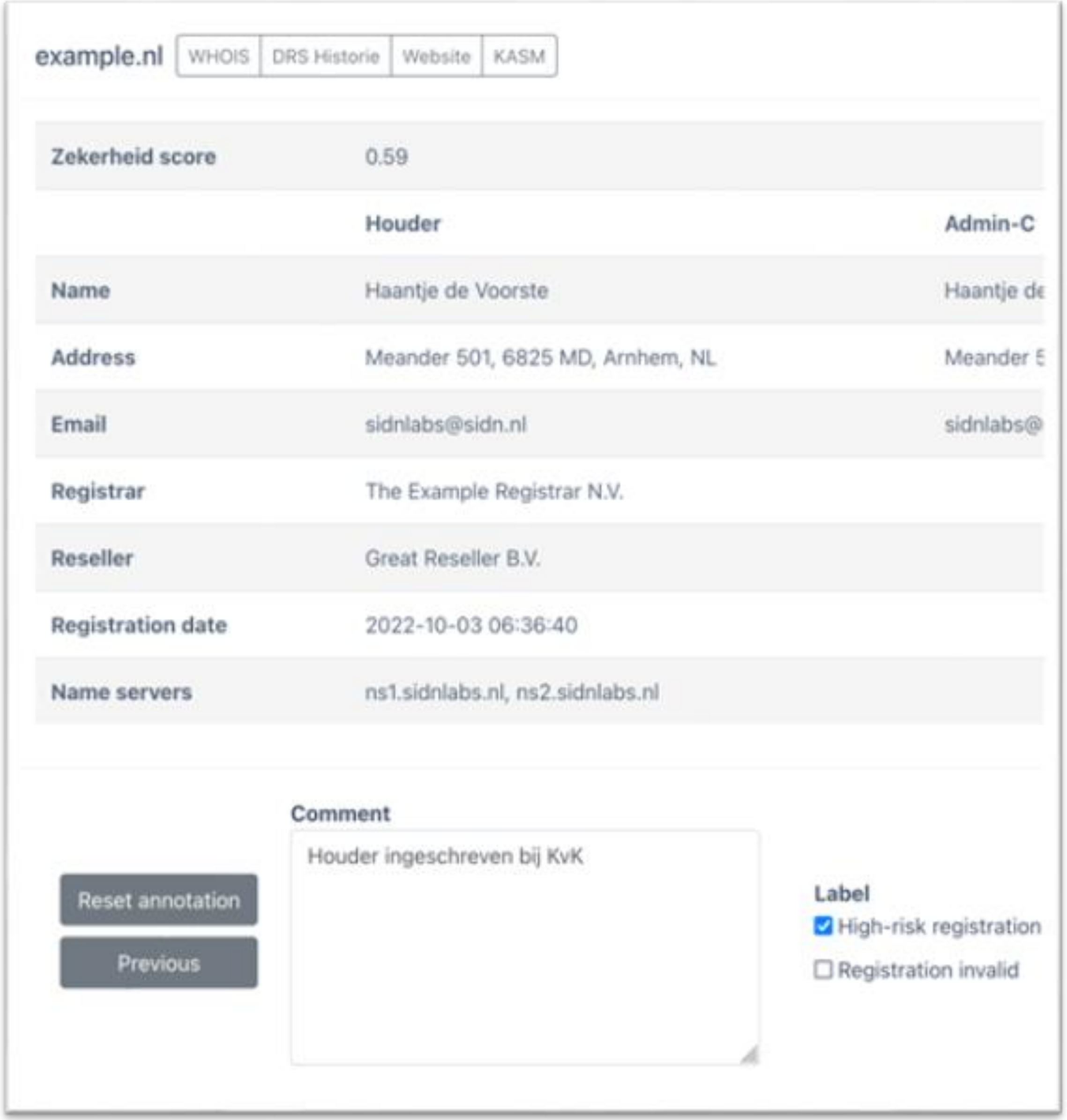


Results so far

- 3 algorithms
- 3 develop & evaluate iterations
- Various training paradigms
- Current prototype is used daily

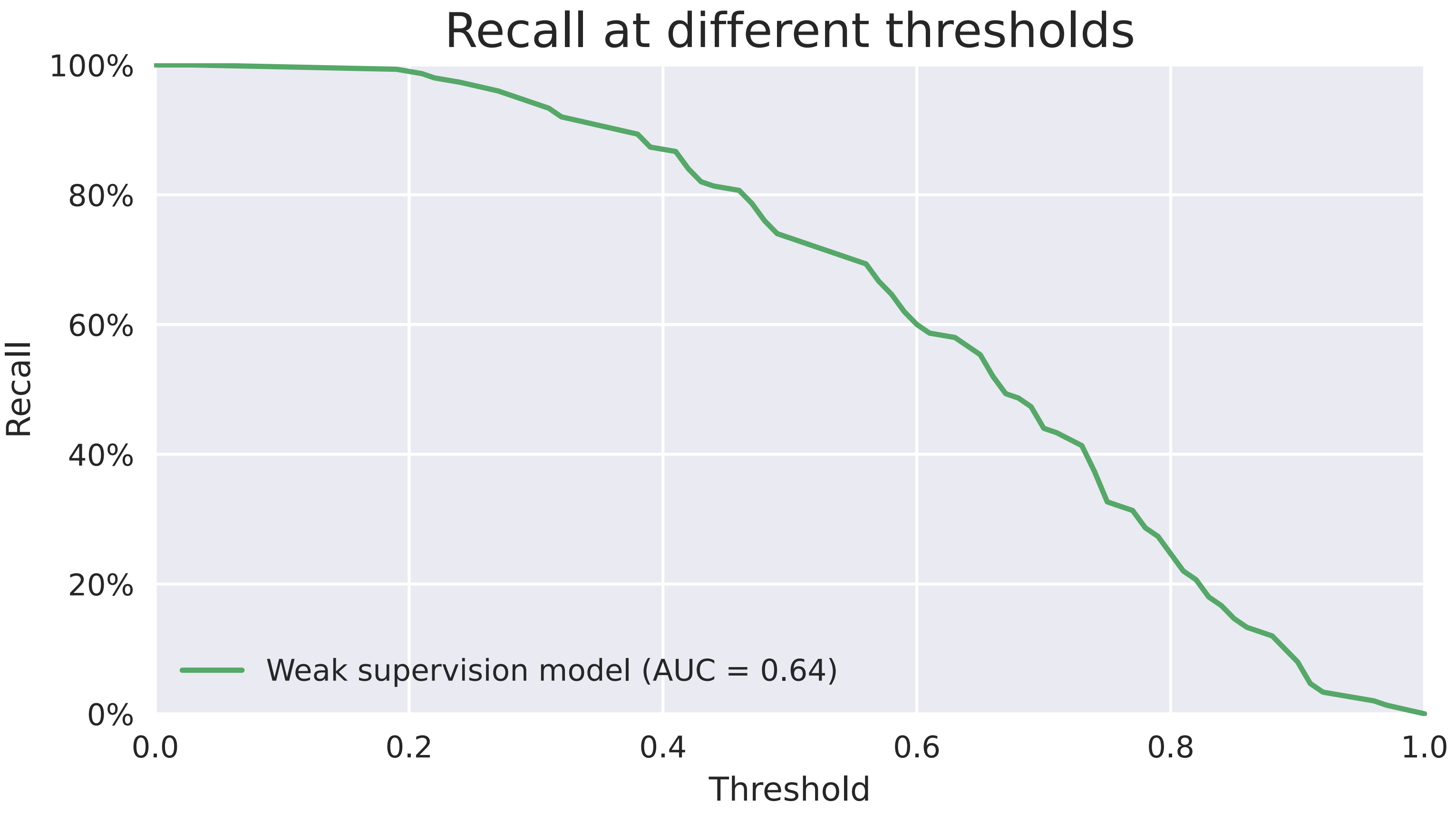
Today:

- 4 lessons learned

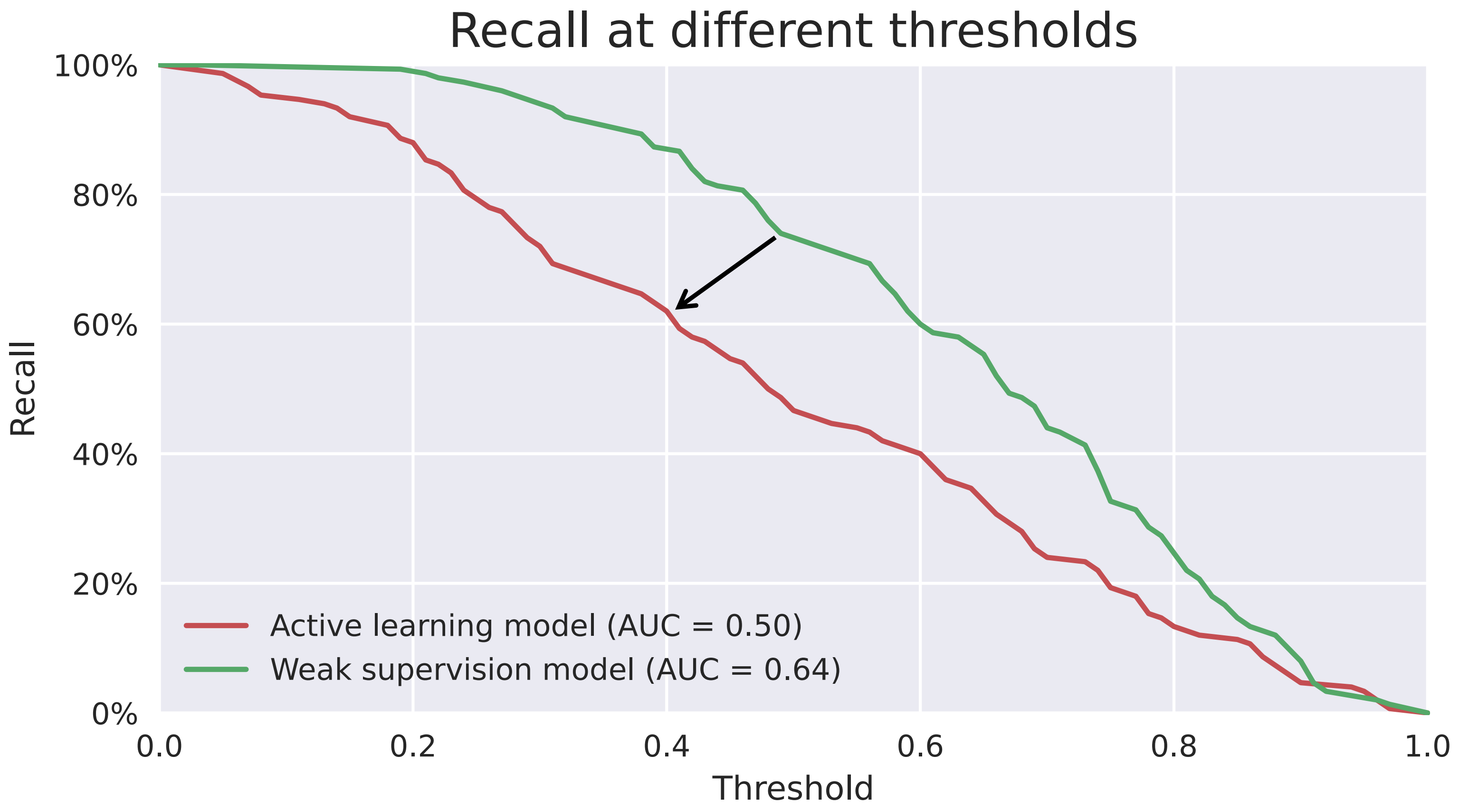


SIDN LABS

Effect of feedback loop (1/2)

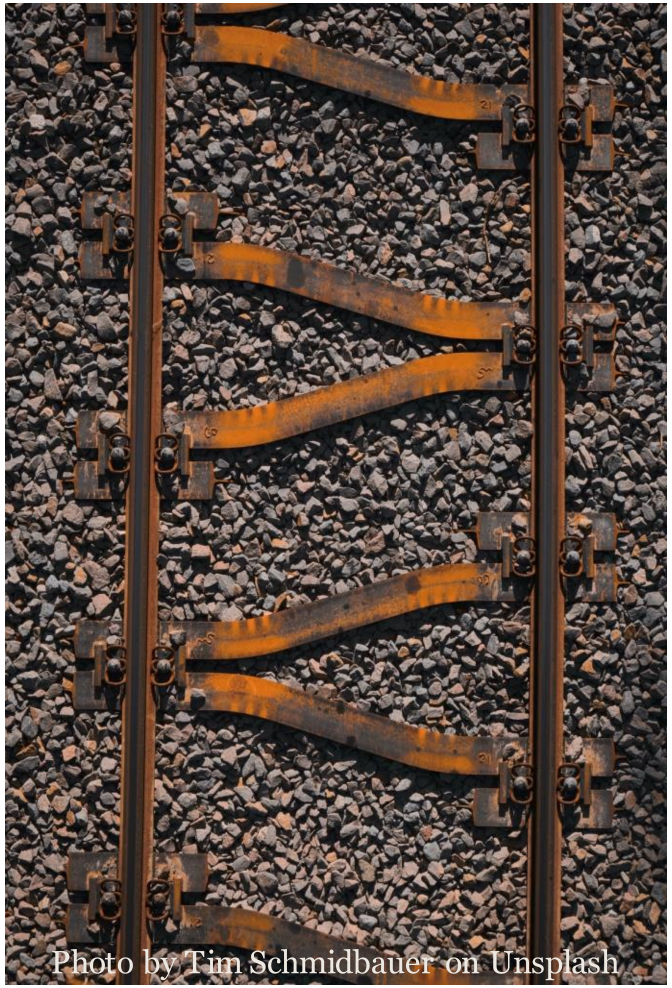


Effect of feedback loop (2/2)



SIDN LABS

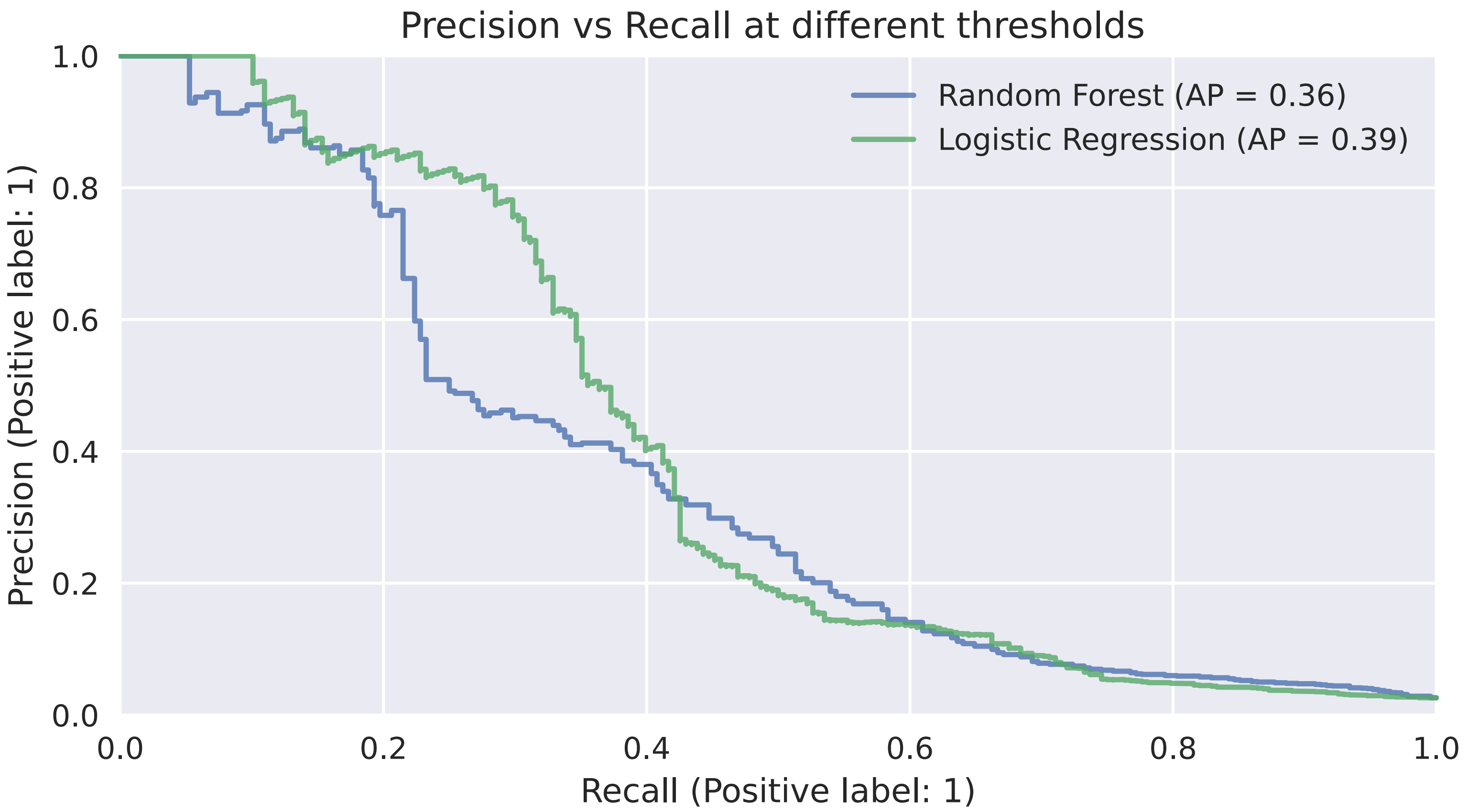
Lesson 1: Speak the same language



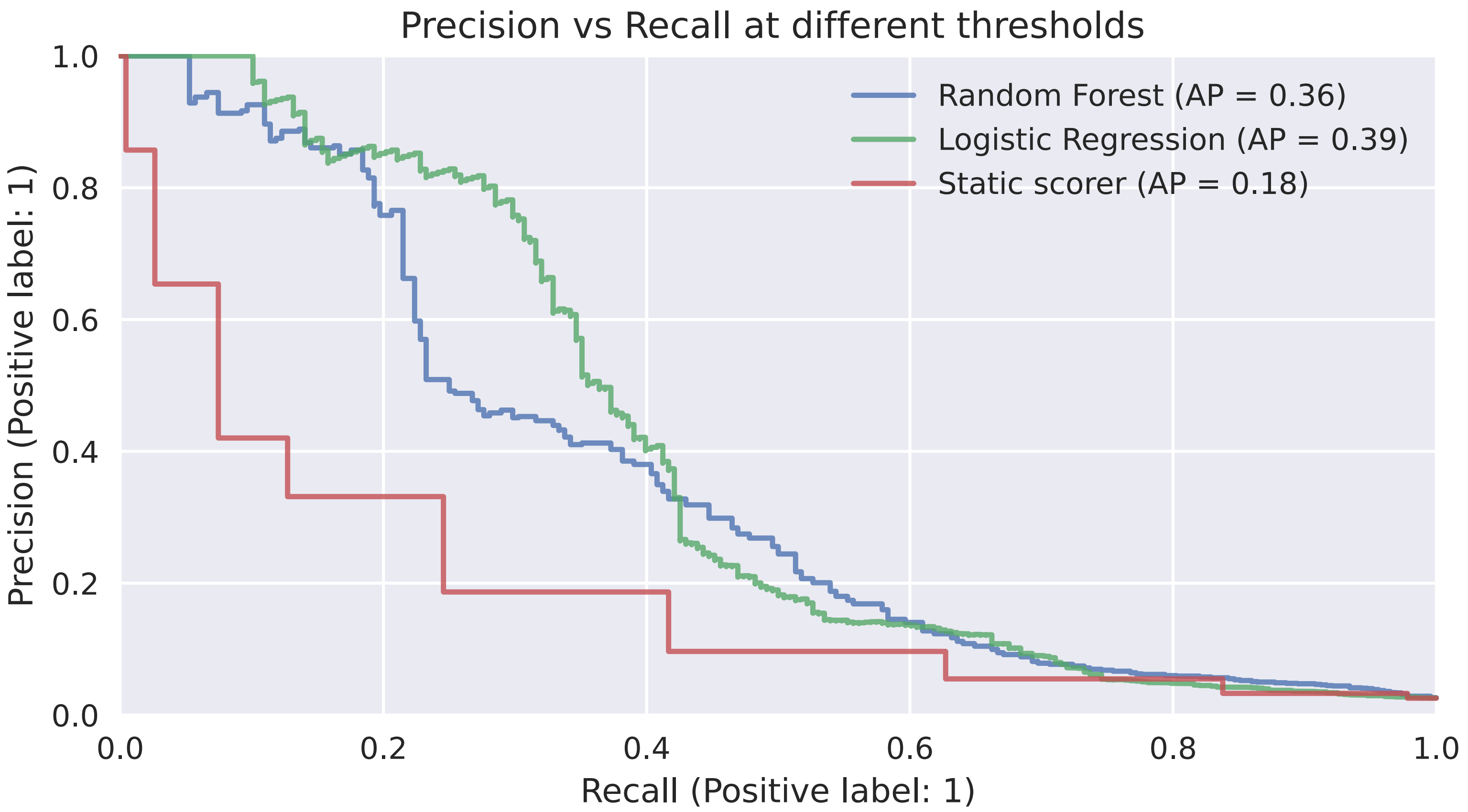
- Problem formulation[1]
  - Identify high-risk domain names?
  - Identify inaccurate registration data?
- Well-defined outcomes
  - What does "high-risk" mean?
  - Does everyone make the same judgement?

[1] Problem Formulation and Fairness - https://arxiv.org/abs/1901.02547

Determine ML performance (1/2)



Determine ML performance (2/2)



Lesson 2: Benchmark ML against non-ML baseline

- Complex ML algorithms do not necessarily outperform simple ones and may have downsides (e.g., lack of explainability, higher costs) [2]

- Adding a non-ML baseline makes pros and cons of using ML explicit

[2] Measuring the predictability of life outcomes with a scientific mass collaboration - https://www.pnas.org/doi/10.1073/pnas.1915006117

SIDN LABS

Evaluation using common ML metrics (1/2)

- Recall (sensitivity) = 0.95
- Precision = 0.86
- Specificity = 0.85

		Predicted		
		Positive	Negative	∑
Actual	Positive	95	5	100
	Negative	15	85	100
	∑	110	90	

Evaluation using common ML metrics (2/2)

- Recall (sensitivity) = 0.95
- Precision = 0.86
- Specificity = 0.85

		Predicted		
		Positive	Negative	∑
Actual	Positive	95	5	100
	Negative	15	85	100
	∑	110	90	

- Expected false negatives = 1
- Expected false positives = 418

	Percentage	Count
Not reported	99.95%	2,786
Reported	0.5%	14
∑	100%	2,800

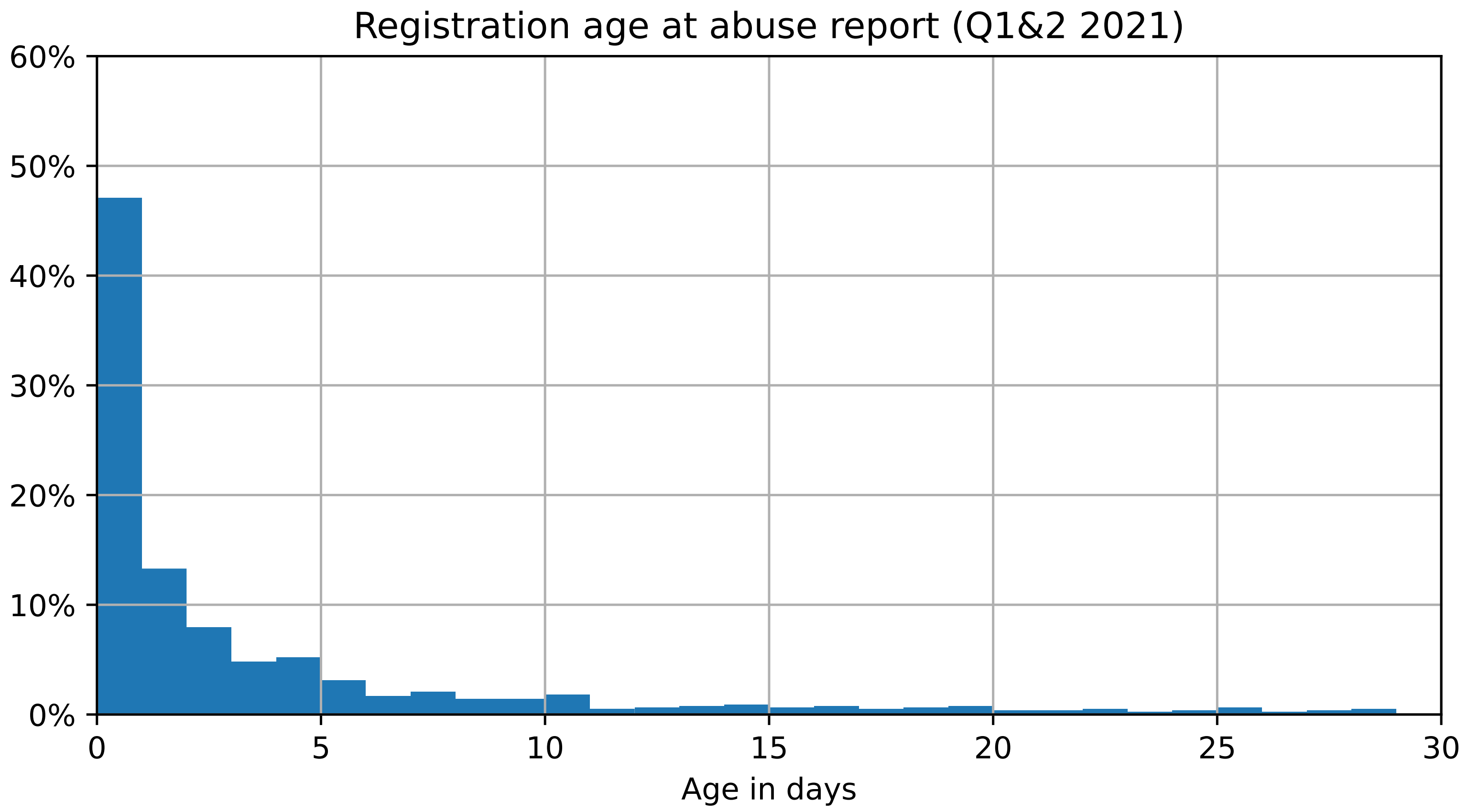
Lesson 3: Learn from other disciplines

- Common ML metrics do not tell the whole story, because performance depends on abuse prevalence
- Adopt metrics from medical research: [3]
  - Positive and Negative Predictive Value (PPV, PNV)
  - Pre- and posttest probability

[3] Understanding and using sensitivity, specificity and predictive values - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636062/

SIDN LABS

Accurate vs rapid intervention



Lesson 4: Discuss technical choices and their impact

- Technical choices often influence the operation and policy (e.g., threshold, features to include)
- Responsible ML is only possible if decisions are made explicit and discussed from different points of view

SIDN LABS

4 lessons learned

Speak the same language

Benchmark against non-ML baseline

Learn from other disciplines

Discuss technical choices and impact

SIDN LABS

Future work

- Improve “operational prototype”
- Joint evaluation and development with DNS Belgium (.be)
- Support peer registries by publishing our methodology

4 lessons learned for responsible machine learning

Speak the same language

Benchmark against non-ML baseline

Learn from other disciplines

Discuss technical choices and impact

Thymen Wabeke
Research engineer
thymen.wabeke@sidn.nl

www.sidnlabs.nl | stats.sidnlabs.nl

Follow us

SIDN.nl

@SIDN

SIDN

Q&A

www.sidnlabs.nl | stats.sidnlabs.nl

Thymen Wabeke
Research engineer
thymen.wabeke@sidn.nl