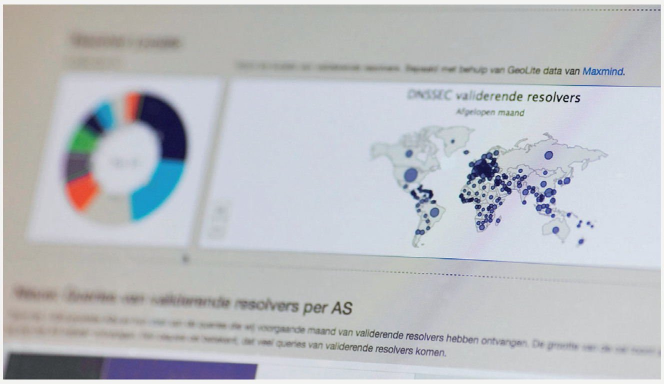
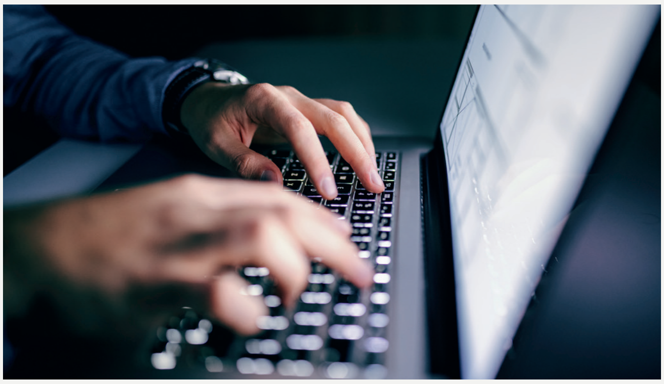
Machine learning projects at SIDN Labs

SIDN Labs is the research team of the .nl operator, SIDN. We develop and evaluate new technologies and systems that further enhance the stability and security of .nl, the DNS and the infrastructure of the wider internet.

You can use different data sets for your research. For example, DMAP crawls all .nl domains – e.g., for its content and infrastructure – and ENTRADA contains over 4 years of DNS traffic from our name servers. Our research is applied and usually empirical, but we're open for other ideas.





Robust detection of fake webshops

Fake webshops do not deliver, deliver counterfeit products or commit credit card fraud. Our research aims at developing and evaluating an operational fake webshop detector that is adaptive, accurate and proactive.

Active learning

The shops that are classified as suspicious are always evaluated by our abuse analysts. *How to use their feedback to automatically improve the detection models while preventing bias?*

Fake registrants

We would like to detect registrations performed by bots, because they are often associated with malicious activities. *How to detect fake registrants? Can you assign a reputation score to each domain registration?*

Assigning economic activities to domains

We classify the economic activity of domains using the HTML-content that we crawl. We will use the classifications, for instance, to assess the impact of DDoS attacks among different sectors. Our initial results are promising, but there is room for improvement.

Multi-level classification

The classes are based on a taxonomy by the Dutch government (i.e. SBI). For example, a restaurant (3rd level) is an eating place (2nd level) which is part of the hospitality industry (1st level). We focus on the 1st level at the moment but would like to have a more fine-grained classification. *How to develop a model that can classify economic activities at different levels?*

Confusing labels

Our classifier is confused sometimes, because similar low-level activities are in different high-level sectors. For example, both travel agencies and holiday resorts websites contain text about lodges. The former is a business service whereas the latter is part of the hospitality industry. *How to cope with this confusion? Can you exploit knowledge about different levels for a more accurate classification?*

Machine learning projects at SIDN Labs

Detecting suspicious IoT traffic

The goal of our SPIN project is to improve the security and privacy of home networks. Detecting and blocking suspicious Internet of Things (IoT) traffic is an import functionality and an active research area.

Device fingerprinting

Detecting anomalies in the network traffic of IoT devices usually requires a single model for each device. We would like to build upon recent research that trains a model per device type. *Can you cluster devices based on their traffic? Do these clusters make sense?*

Getting out the lab

Most anomaly detectors are evaluated using publicly available data with known botnet traffic. Evaluating a detector outside the lab is maybe more relevant. *Can you develop an anomaly detector and evaluate it outside the lab? Can devices collaboratively train a detection model while keeping their network traffic private?*

Please send an e-mail if you are interested in one of the topics. We can then further discuss the possibilities of doing an internship or writing your master thesis at SIDN Labs.

Thymen Wabeke
Research engineer

thymen.wabeke@sidn.nl

Jelte Jansen
Research engineer

jelte.jansen@sidn.nl

More information?

Our application procedure:

https://www.sidnlabs.nl/afstuderen

Our machine learning roadmap:

https://www.sidnlabs.nl/a/weblog/op-weg-naar-een-veiliger-internet-met-machine-learning