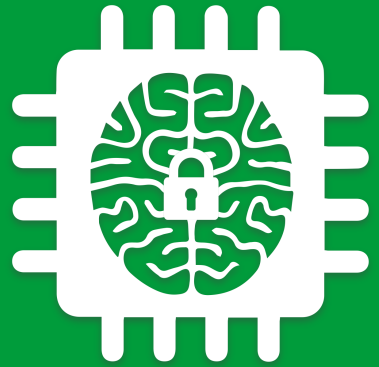
SECRET | UFPR UNIVERSIDADE FEDERAL DO PARANÁ

ENIGMA | WWW.SECRET.INF.UFPR.BR
ENIGMA 2023

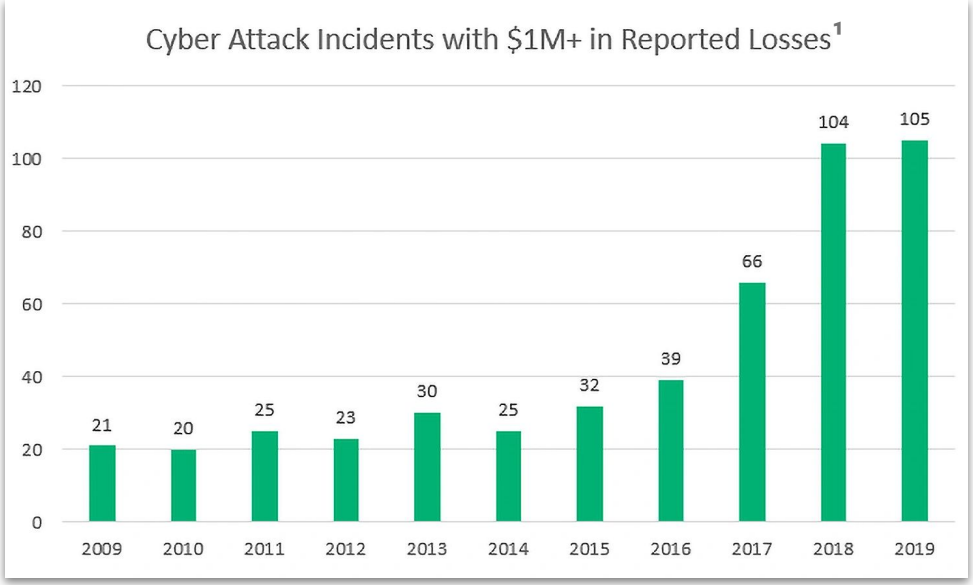
Spotting the Differences: Quirks of Machine Learning (in) Security



Fabrício Ceschin
PhD Student at Federal University of Paraná (UFPR), Brazil
@fabriciojoc

ENIGMA: A USENIX CONFERENCE
Security and Privacy Ideas That Matter
JAN 24 - 26, 2023

Why is it important?



Machine Learning: The Ideal Ally for Security Analysts[2]

Product Features | Cortex XSOAR | Incident Response | Machine Learning | SOAR

444 people reacted | 1

By Jane Goh
July 17, 2018 at 6:20 AM
4 min. read

Machine learning, a subset of artificial intelligence, is the practice of using algorithms and large data sets or Big Data to develop insights ranging from which movie a Netflix user may want to watch next to recommendations about cybersecurity incident handling.

According to consulting firm McKinsey, "the unmanageable volume and complexity of the big data that the world is now swimming in have increased the potential of machine learning—and the need for it."

For security professionals, machine learning capabilities can increase responder productivity and enable leaner, more efficient security operations. Humans however, not machines, must direct and guide machine learning algorithms to achieve the business goals and objectives that the computers are given.

Machine Learning, Big Data, and Security

The best way to understand how machine learning can be beneficial for security analysts is to perhaps look at another field with similar operational efficiency goals that is currently taking advantage of Big Data, and prospering - Marketing.

[1]https://sectigostore.com/blog/42-cyber-attack-statistics-by-year-a-look-at-the-last-decade/
[2]https://www.paloaltonetworks.com/blog/security-operations/machine-learning-the-ideal-ally-for-security-analysts/

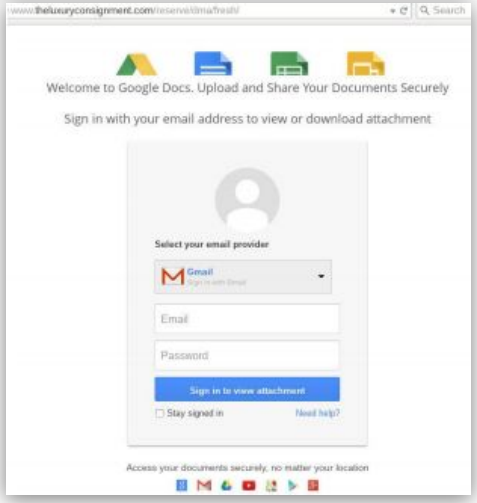
But why is Machine Learning for Security Different?

Cat

Dog

Example: Phishing

68% of phishing emails blocked by gmail are different from day to day, i.e., data distribution is non-stationary!



Source: https://elie.net/talk/deconstructing-the-phishing-campaigns-that-target-gmail-users/

“Blindly applying ML evaluation best-practices may backfire in security contexts!”

Source: Lorenzo Cavallaro. When the Magic Wears Off: Flaws in ML for Security Evaluations. ENIGMA 2019. https://www.usenix.org/conference/enigma2019/presentation/cavallaro

How do I know about it? My first experience with ML and Security

The Need for Speed: An Analysis of Brazilian Malware Classifiers[1]

Fabrício Ceschin, Felipe Pinagé, Marcos Castilho, David Menotti, Luiz S. Oliveira, André Grégio | Federal University of Parana

The Random Forest classifier achieved the best overall result, with better values for all metrics (98.00 percent for accuracy, 98.07 percent for F1 score, 97.52 percent for recall, and 98.63 percent for precision).

Dataset

To create the dataset presented in this article, we used a set of ≈200 GB of executable files (malware and benign software) collected in Brazilian cyberspace or popular Internet download sites from 2012 to early 2018. We collected two classes of software to build the dataset: goodware (allegedly benign software) and malware. To obtain goodware samples, we implemented a web crawler that downloaded software from three sources: Sourceforge, Softonic, and CNET Download. We collected ≈130 GB of binary files, which we assumed benign, totaling 21,116 unique samples. We have an established long-standing partnership with a major Brazilian financial institution (which prefers to remain anonymous) that provides us with daily malware samples collected from detected infections in its corporate perimeter or identified by customers via phishing e-mail attachments. As the malware samples were received by our server, we grouped them by day to save the temporal information. This process has been executed in an ongoing fashion since January 2013, with the exception of the period from January to July 2016, when the collection was suspended due to a shutdown period in the storage server. At the time this article was being written, we had collected approximately 80 GB of malicious binary samples, totaling 29,704, of which 23,033 are unique.

[1]Ceschin, Fabrício; Pinage, Felipe; Castilho, Marcos; Menotti, David; Oliveira, Luis S; Gregio, André. (2018). The need for speed: An analysis of brazilian malware classifiers. IEEE Security Privacy, 16 (6), pp. 31-41, 2018, ISSN: 1540-7993.

How do I know about it? Machine Learning Evasion Competition (MLSEC)

CUJO AI Partners with Microsoft for the Machine Learning Security Evasion Competition 2020[1]

June 4, 2020

Last year, Dr. Hyrum Anderson, now the Principal Architect (Security Machine Learning) at Microsoft, and I designed a competition, and it was a huge success. It was launched at AI Village in August 2019 at DEFCON 27, where we invited contestants to participate in a white-box attack against static malware Machine Learning (ML) models.

Now that ML is even more widespread in detecting cyberthreats, the Machine Learning Security Evasion Competition is back with an improved game and more partners joining.

This year, Microsoft is sponsoring the event as part of their investments in the Trustworthy Machine Learning Initiative, and CUJO AI's Vulnerability Research Lab is developing the framework to host the competition, enabling both the defensive and the offensive sides of the game.

Zoltan Balazs
HEAD OF VULNERABILITY RESEARCH LAB
Linkedin

Ethical hacker and IT security researcher with more than 15 years of experience. Apart from an MSc Degree, ten technical certs, including OSCE or CISSP. Former speaker at DEFCON, SAS, AusCERT, Shakacon, and many more.

The Twofold Challenge

Last year, the goal of the competition was to get 50 malicious Windows Portable Executable (PE) files to evade detection by three machine learning malware classifiers. Not only did the files need to evade detection, but they also had to maintain their exact original functionality and behavior.

The 2020 Machine Learning Security Evasion Competition (MLSEC) is similarly designed to experiment with the variety of ways ML systems may be evaded by malware, in order to better defend against these techniques.

The **Defender Challenge** will run **June 15 through July 23.** Participants will develop novel defences for attackers to evade. Submitted defences must pass real-world tests in detecting real-world malware at moderate false-positive rates.

The **Attacker Challenge** will run **August 6 through September 18,** providing a black-box threat model giving API access to hosted antimalware models, including those successfully developed in the Defender Challenge.

Contestants may discover how to evade them using "hard-label" query results. Samples from final submissions will be detonated in a sandbox to verify that they are still functional. In addition to evasion rates, the total number of API queries required by a contestant will factor into the final ranking.

[1]https://cujo.com/machine-learning-security-evasion-competition-2020/

How do I know about it? The fast food ad reality

The Final Results[1, 2]

In the defender's challenge, our model finished in the second position, with 165 evasions, behind domumpbq model, with a good advantage over ember, as shown in the Table below. We are not sure about how representative this result is since it depends on the attacker's abilities to bypass the models. Anyway, we consider this result positive, given that we did not focus on any specific strategy to deal with adversarial samples rather than establishing a threshold, as we already presented. Also, considering that ember was the best defense solution in last year's challenge, we consider this as an evolution. As future work, we plan to use all the data created in this competition to improve our research model.

DEFENDER NAME	NUMBER OF EVASIONS
domumpbq	65
needforspeed	**165**
ember (sample solution)	270

Expectation

Reality

[1]https://secret.inf.ufpr.br/2020/09/29/adversarial-malware-in-machine-learning-detectors-our-mlsec-2020-secrets/
[2]Ceschin, F., Botacin, M., Lüders, G., Gomes, H. M., Oliveira, L. S. e Grégio, A. (2020). No Need to Teach New Tricks to Old Malware: Winning an Evasion Challenge with XOR-Based Adversarial Samples. Em Proceedings of the 4th Reversing and Offensive-Oriented Trends Symposium, ROOTS'20, Vienna, Austria.

Dataset Definition

Dataset Definition

Small Dataset

- May not represent a real scenario
- Invalidates results of experiments
- May not generalize and work
- Bad classification performance

Big Dataset

- Long training times
- Complex models
- Not feasible in the reality
- Deep learning = large amount of data

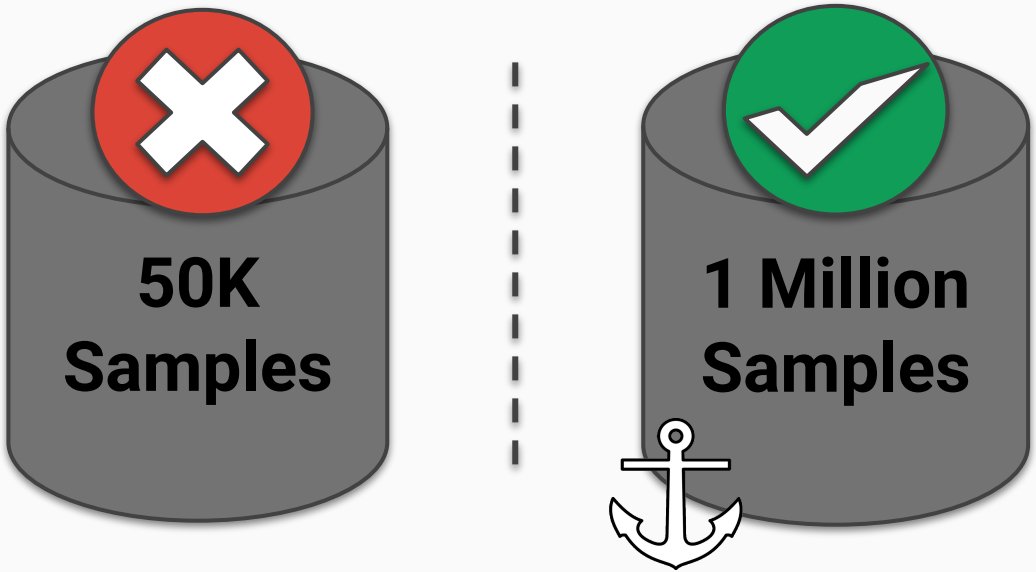
“Essentially, all models are wrong, but some are useful”

George Box

“The map is not the territory, it is a symbol, index, or representation, but not the place itself”

Alfred Korzybski

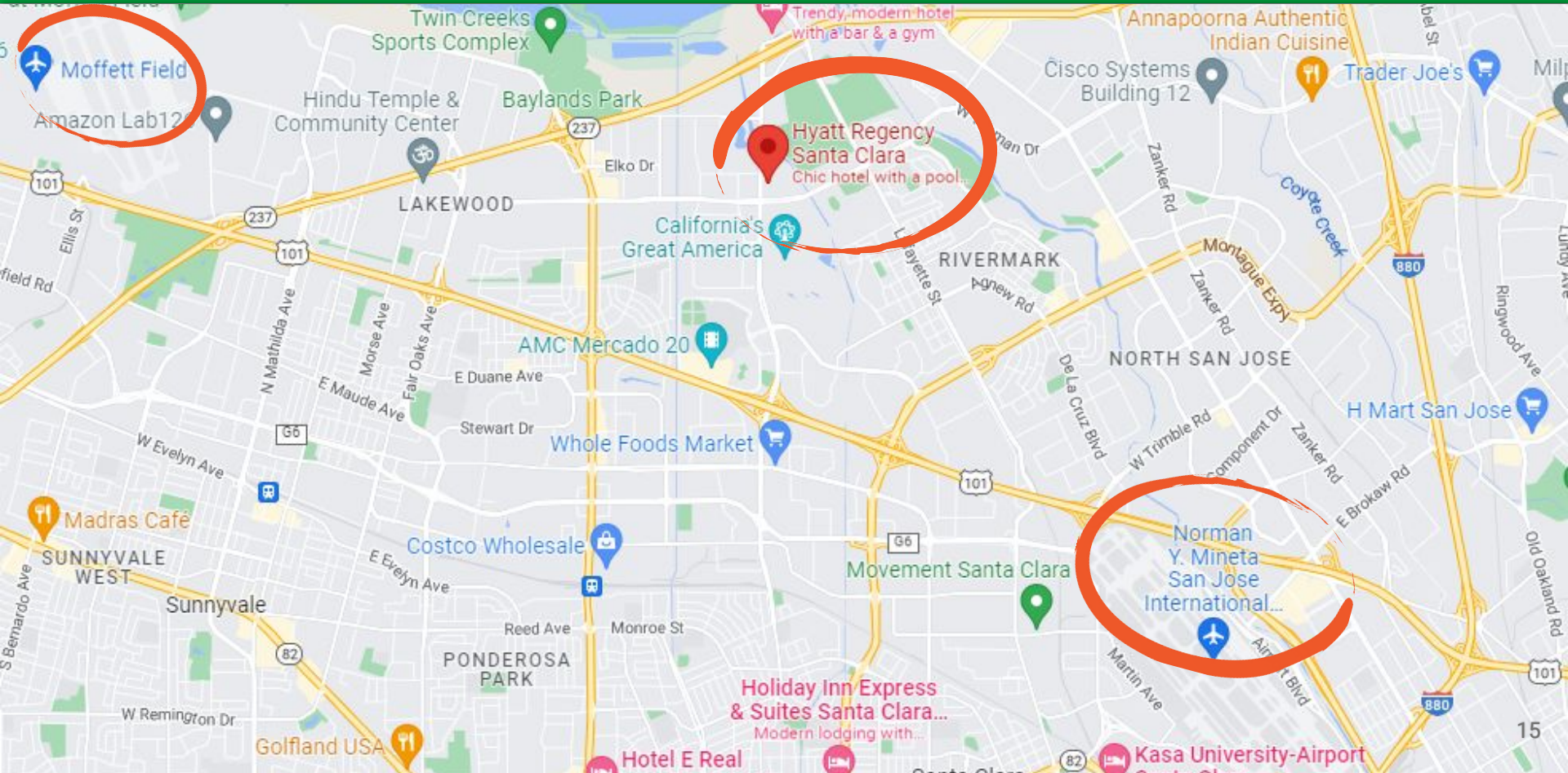
Dataset Definition: Anchor Bias



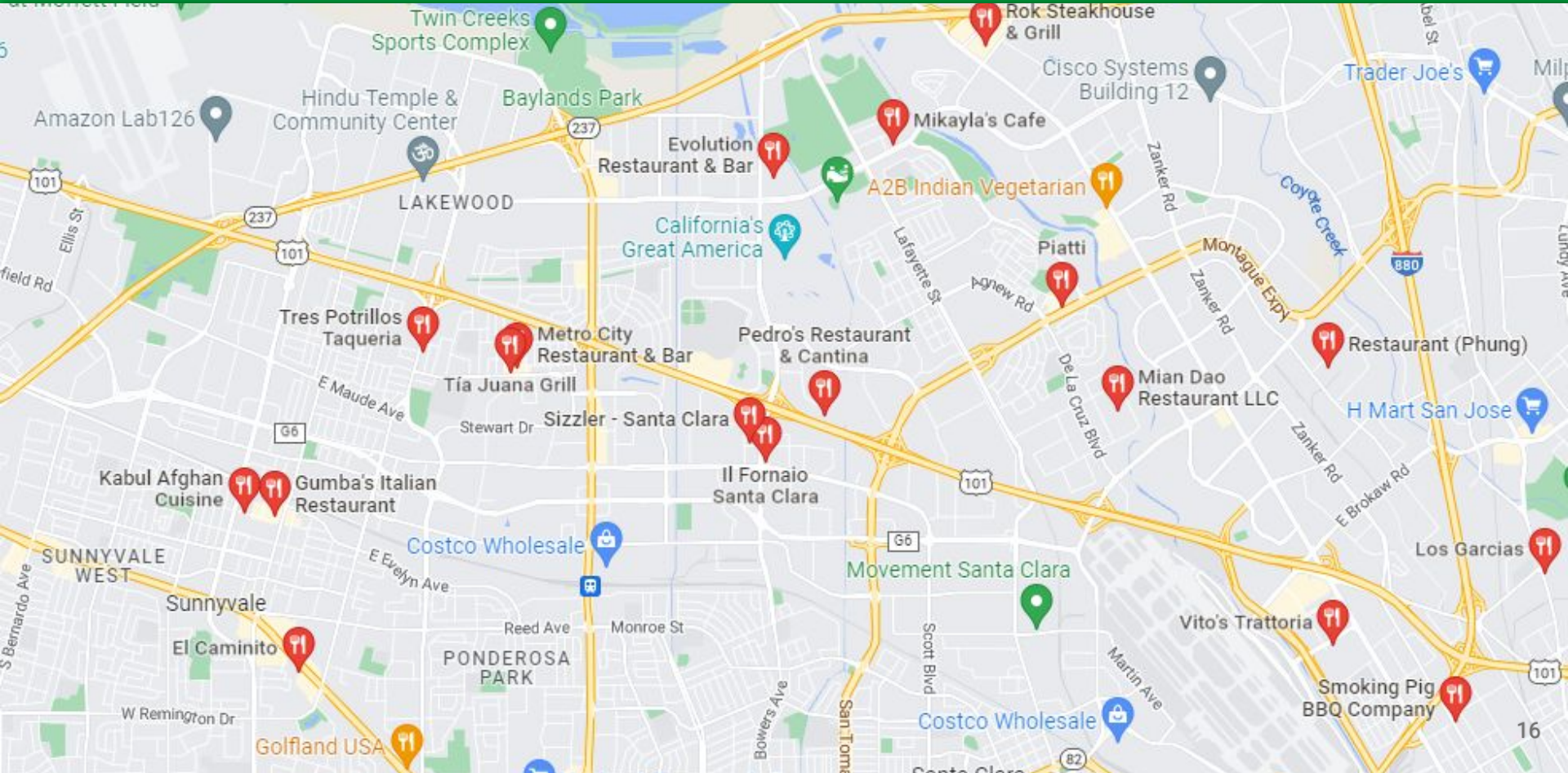
It is NOT About Its Size

It is about its representation of the real-world for the task being performed, as a map

ENIGMA Hotel Map

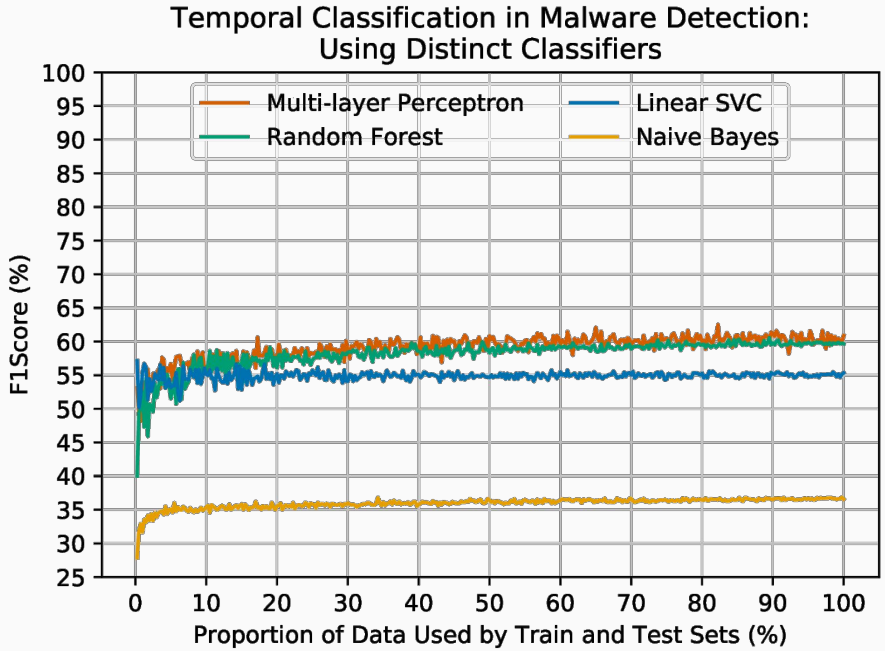


Santa Clara Restaurants Map

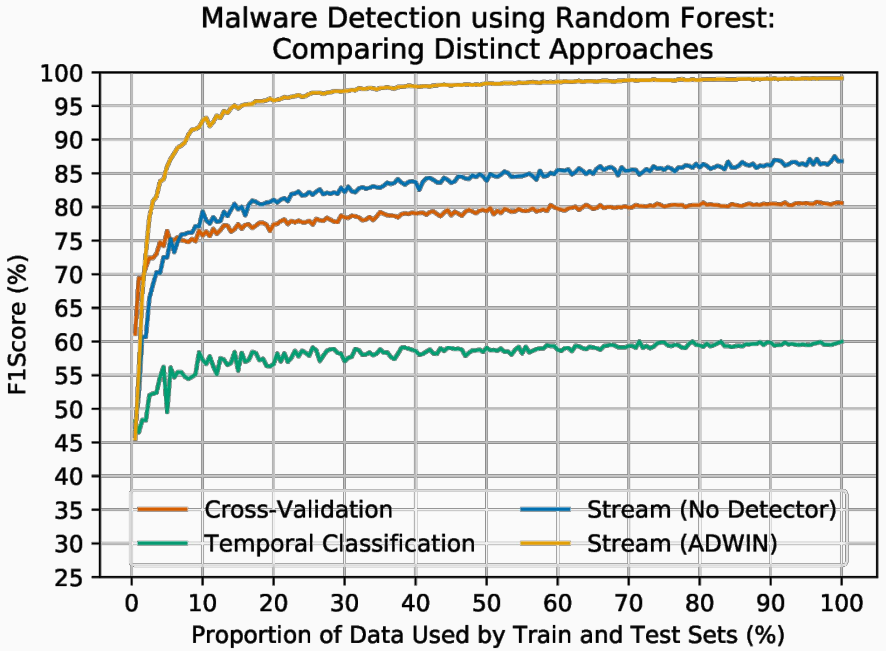


Dataset Definition Experiments

Exp#1. Comparing Classifiers: temporal classification (training set = “known data”, test set = ”future data“)
Results: “optimal” classification achieved using only 10%~20% of the original dataset



Exp#2. Comparing Approaches: different evaluation approaches, unique classifier (Random Forest)
Results: all stabilize F1 Score in 30%~40% proportions, less than half of the original dataset



Dataset: AndroZoo, a subset of Android applications reports provided by AndroZoo, composed of eight textual attributes (resources names, source code classes and methods, manifest permissions etc.) and contains 213,928 benign and 70,340 malicious applications.

Class Imbalance

Class Imbalance

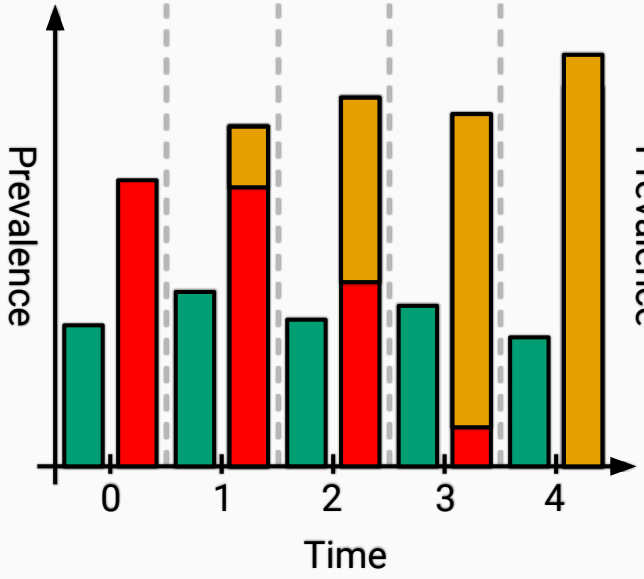
- **Definition:** distribution between classes differs substantially
- **Present in real-world research:** difficult to collect malicious samples
- **AndroZoo dataset:** ~18% of malicious applications[1]
- **Literature:** several methods try to overcome this problem
- **Resampling:** removing (undersampling) or adding (oversampling) instances

[1]Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. 2019. TESSERACT: eliminating experimental bias in malware classification across space and time. In Proceedings of the 28th USENIX Conference on Security Symposium (SEC'19). USENIX Association, USA, 729–746.

Undersampling: Removing Samples from the Majority Class

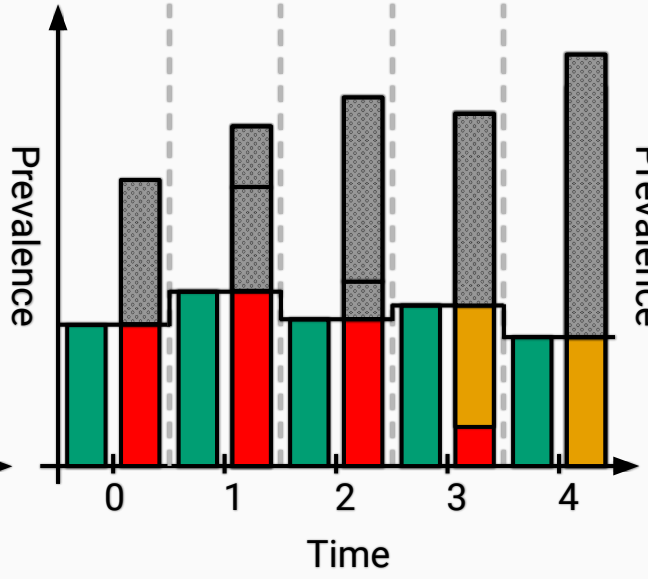
1. Original Distribution

- No technique applied



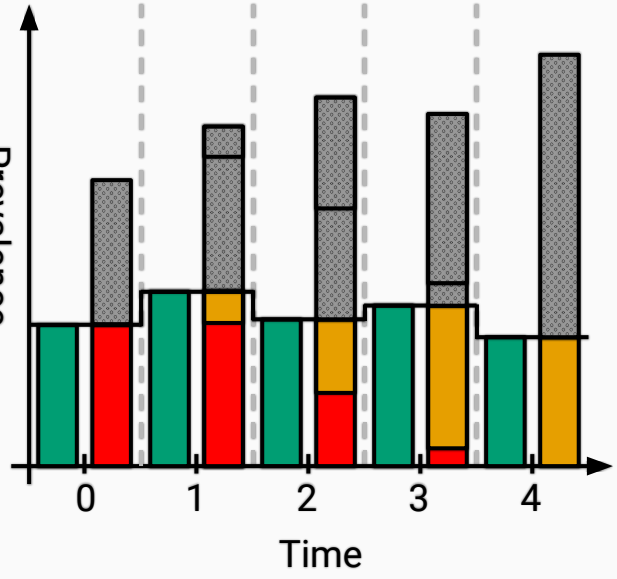
2. No Temporal Information

- Different than reality



3. With Temporal Information

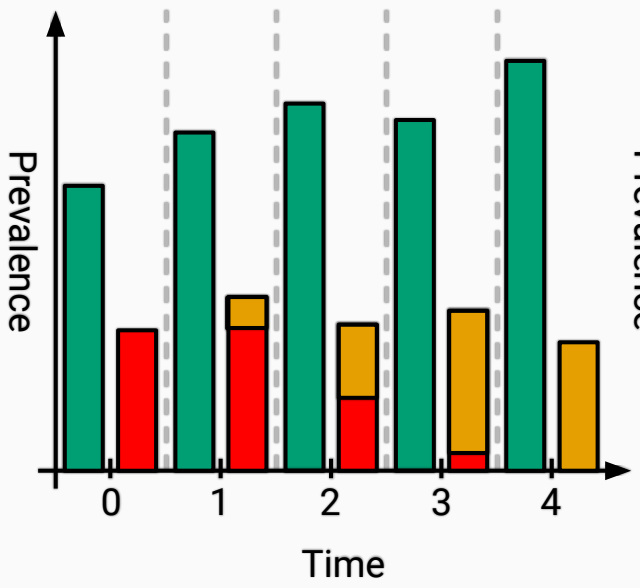
- Ideal scenario



Oversampling: Generating New Samples from the Minority Class

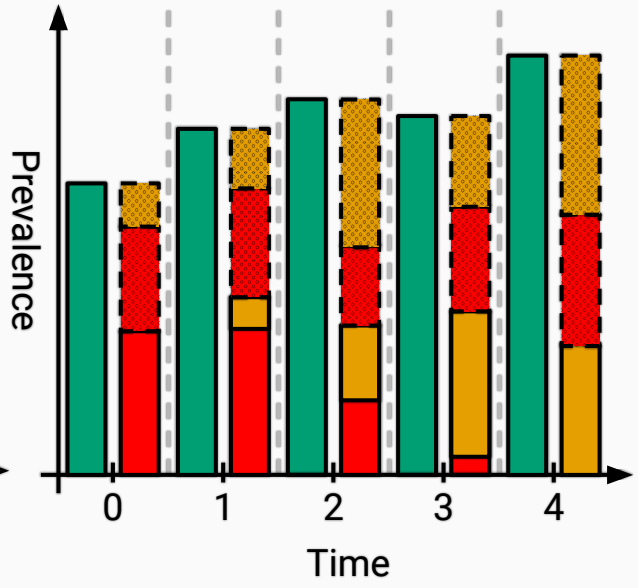
1. Original Distribution

- No technique applied



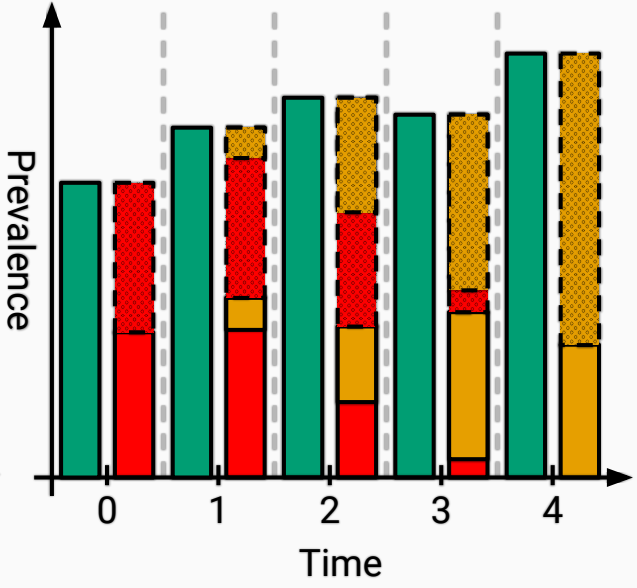
2. No Temporal Information

- Different than reality



3. With Temporal Information

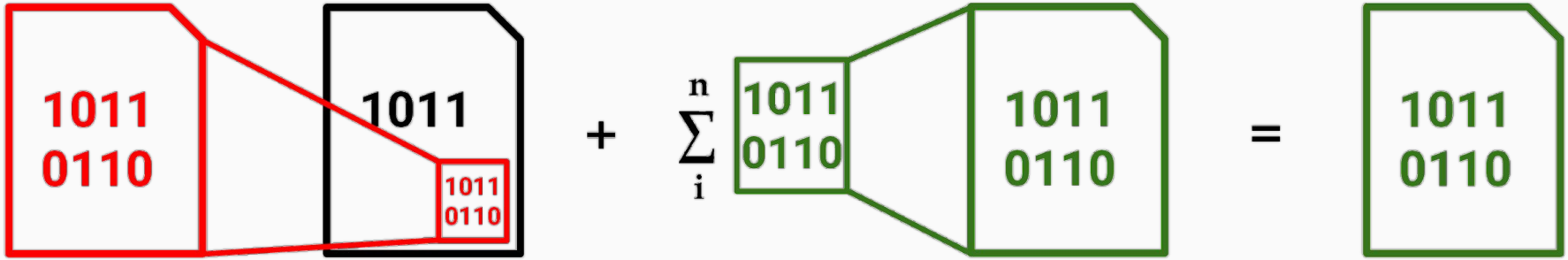
- Ideal scenario



Adversarial Machine Learning

Adversarial Machine Learning

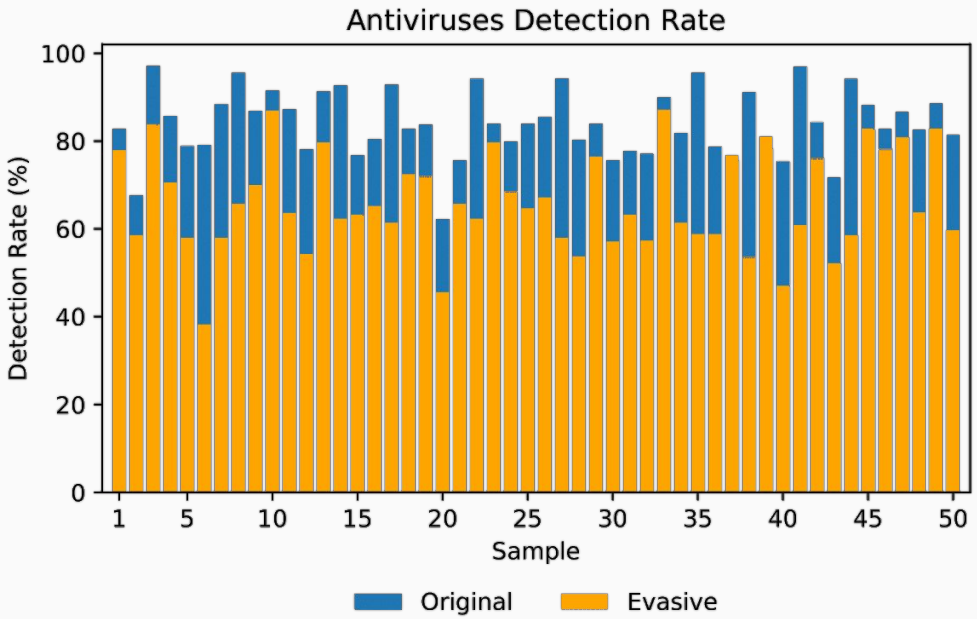
- **Generating adversaries:** totally functional malware bypasses ML models[1]



[1]Ceschin, F., Botacin, M., Gomes, H. M., Oliveira, L. S. e Grégio, A. (2019). Shallow security: On the creation of adversarial variants to evade machine learning-based malware detectors. Em Proceedings of the 3rd Reversing and Offensive-Oriented Trends Symposium, ROOTS'19, Vienna, Austria.

Adversarial Machine Learning: AntiViruses and Machine Learning

- **VirusTotal:** average detection rates for 50 adversarial samples using this strategy in ~60 AVs
  - **Original** vs **Evasive**
- **Results:** affected real AV engines[1] [2]



[1]Ceschin, F., Botacin, M., Gomes, H. M., Oliveira, L. S. e Grégio, A. (2019). Shallow security: On the creation of adversarial variants to evade machine learning-based malware detectors. Em Proceedings of the 3rd Reversing and Offensive-Oriented Trends Symposium, ROOTS'19, Vienna, Austria.
[2]Ceschin, F., Botacin, M., Lüders, G., Gomes, H. M., Oliveira, L. S. and Grégio, A. (2020). No need to teach new tricks to old malware: Winning an evasion challenge with xor-based adversarial samples. Proceedings of the 4th Reversing and Offensive-Oriented Trends Symposium, ROOTS'20, Vienna, Austria.

Adversarial Machine Learning: Updating Classification Model

Train Dataset[1]	Test Dataset	False Negative Rate (FNR)	False Positive Rate (FPR)	Conclusion
No Adversaries (Ember Only)	Windows Apps + MLSEC 2020 Original Samples	0%	0.1%	**All** malware and the majority of goodware are detected
	Windows Apps + MLSEC 2020 Adversarial Malware	100%	0.1%	**No** malware detected and the majority of goodware are detected
With Adversaries (MLSEC 2019 + EMBER)	Windows Apps + MLSEC 2020 Original Samples	0%	78.54%	**All** malware detected, and the majority of goodware are considered malware
	Windows Apps + MLSEC 2020 Adversarial Malware	0%	78.54%	

[1]Ceschin, F., Botacin, M., Lüders, G., Gomes, H. M., Oliveira, L. S. and Grégio, A. (2020). No need to teach new tricks to old malware: Winning an evasion challenge with xor-based adversarial samples. Proceedings of the 4th Reversing and Offensive-Oriented Trends Symposium, ROOTS'20, Vienna, Austria.

Delayed Labels Evaluation

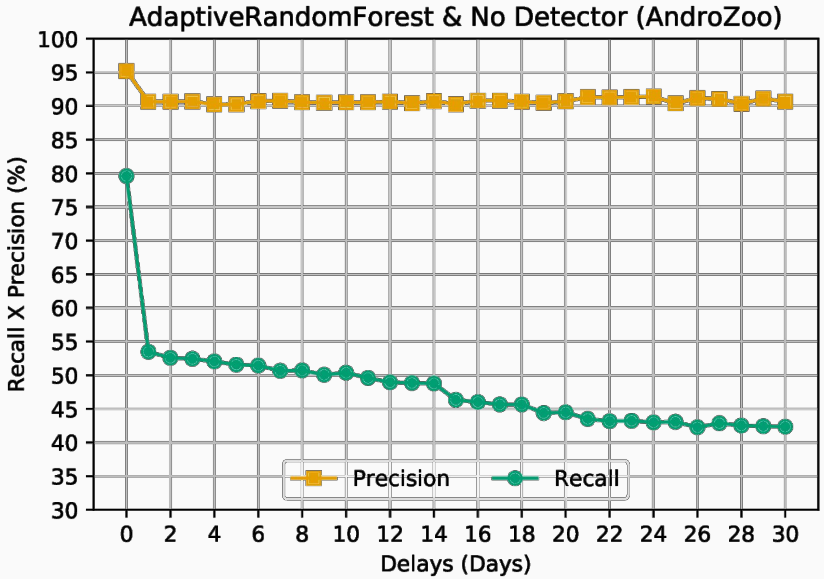
Delayed Labels Evaluation

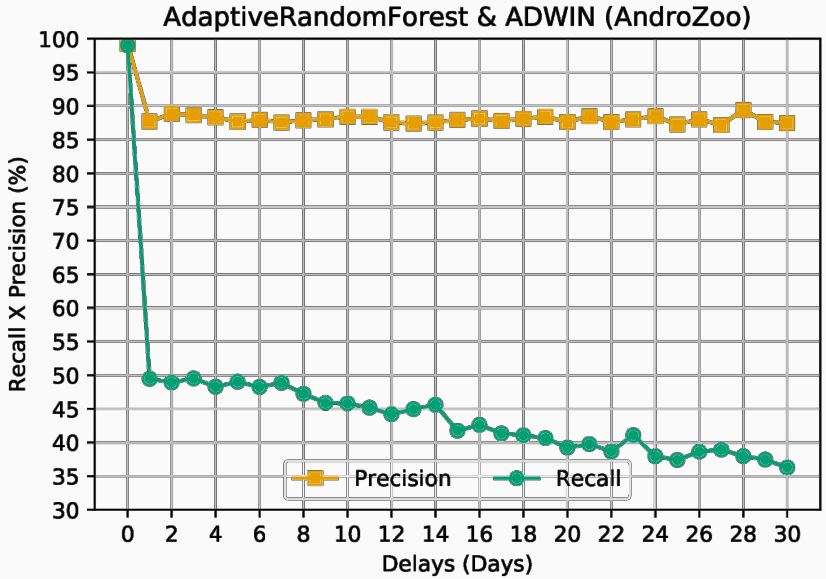
- **Particularity of security data:** do not have labels right after collected
- **Not considered:** labels available the same time as data
  - **Some works:** same labels as model
  - **Poisoning:** decreases detection rate
- **Malware detection:** single snapshot from VirusTotal, not considering delays
- **Delays:** Vary from ten days to more than two years, AVs take ~19 days to detect the majority of the threats[1]

[1]Botacin, M., Ceschin, F., de Geus, P. and Grégio, A. (2020b). We need to talk about antiviruses: Challenges & pitfalls of av evaluations. Computers & Security, page 101859.

Delayed Labels Evaluation Experiments

- **Simulation:** subset of AndroZoo, online ML with/without drift detectors, as a data stream
  - Provide labels of each sample N days after they are available to further update the decision model
- **When not considering a delay:** drift detector improves the detection
- **After one day of delay:** model that do not consider concept drift perform better overall
  - ML models do not perform the way they are evaluated without these conditions

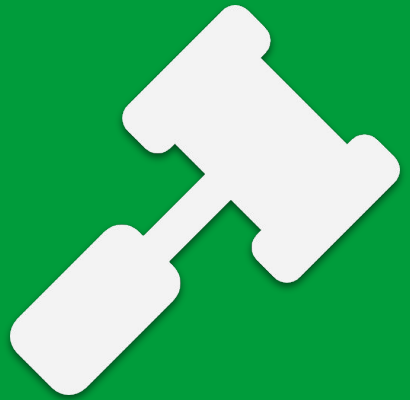




Open source contribution to scikit-multiflow: skmultiflow.evaluation.EvaluatePrequentialDelayed

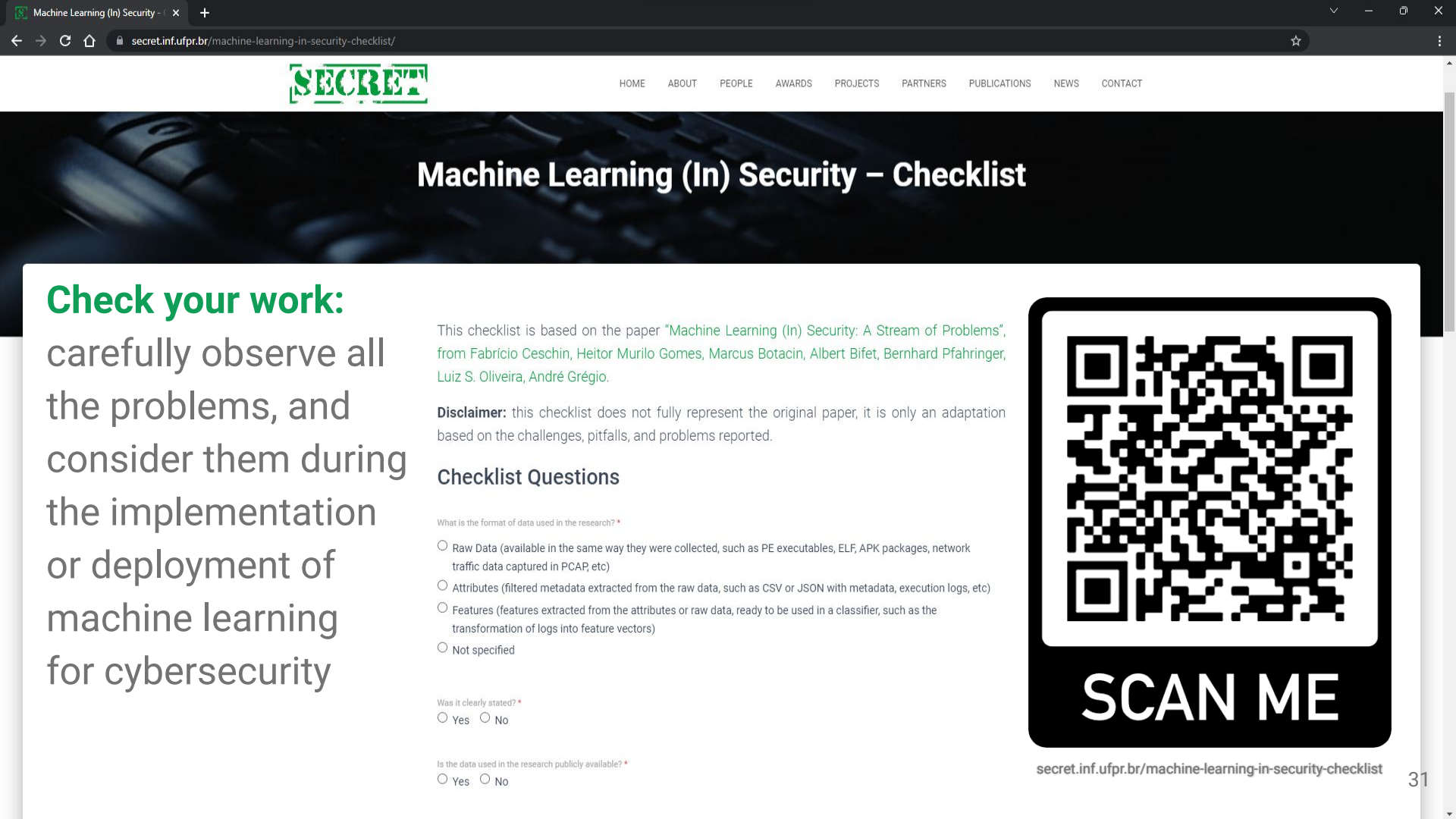
scikit-multiflow

Conclusion



The future of ML for Cybersecurity

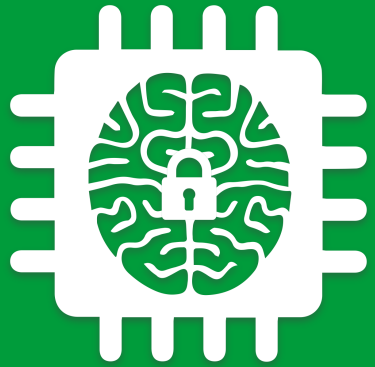
- **Stop looking only at metrics, and start looking at effects:** challenges are open research problems, benefit both academy and industry
- **Commit yourself to real world:** future research need to keep the motto “machine learning that matters” in mind
  - **If your work is losing connection to real world:** we have a problem!


Machine Learning (In) Security -
secret.inf.ufpr.br/machine-learning-in-security-checklist/
SECRET
HOME
ABOUT
PEOPLE
AWARDS
PROJECTS
PARTNERS
PUBLICATIONS
NEWS
CONTACT
Machine Learning (In) Security – Checklist
Check your work:
carefully observe all the problems, and consider them during the implementation or deployment of machine learning for cybersecurity
This checklist is based on the paper "Machine Learning (In) Security: A Stream of Problems", from Fabrício Ceschin, Heitor Murilo Gomes, Marcus Botacin, Albert Bifet, Bernhard Pfahringer, Luiz S. Oliveira, André Grégio.
Disclaimer: this checklist does not fully represent the original paper, it is only an adaptation based on the challenges, pitfalls, and problems reported.
Checklist Questions
What is the format of data used in the research? *
Raw Data (available in the same way they were collected, such as PE executables, ELF, APK packages, network traffic data captured in PCAP, etc)
Attributes (filtered metadata extracted from the raw data, such as CSV or JSON with metadata, execution logs, etc)
Features (features extracted from the attributes or raw data, ready to be used in a classifier, such as the transformation of logs into feature vectors)
Not specified
Was it clearly stated? *
Yes
No
Is the data used in the research publicly available? *
Yes
No
SCAN ME
secret.inf.ufpr.br/machine-learning-in-security-checklist

“The past is just data.
I only see the future”

Ayrton Senna

Spotting the Differences:
Quirks of Machine Learning (in) Security



Contact: fjoceschin@inf.ufpr.br or @fabriciojoc
Our Website: secret.inf.ufpr.br

Fabrício Ceschin
PhD Student at Federal University of Paraná (UFPR), Brazil
@fabriciojoc