# AI and New Threats: A Skeptics' Guide

## Or, What Civil War, Love and Infosec Have in Common

**Juan Tapiador**
*Universidad Carlos III de Madrid*

# About this talk

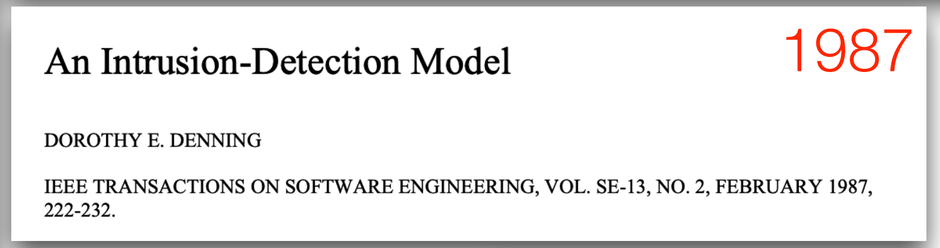
My experience doing (and evaluating) research that uses ML techniques for security (detection/classification) problems.

Focus on research pitfalls, on negative results, on why very cool (AI/ML) solutions rarely work in the real world as promised in the paper.

I am a skeptic, not a negationist. In my experience, security problems are a hard research space.

# AI and Infosec


Not a new relationship

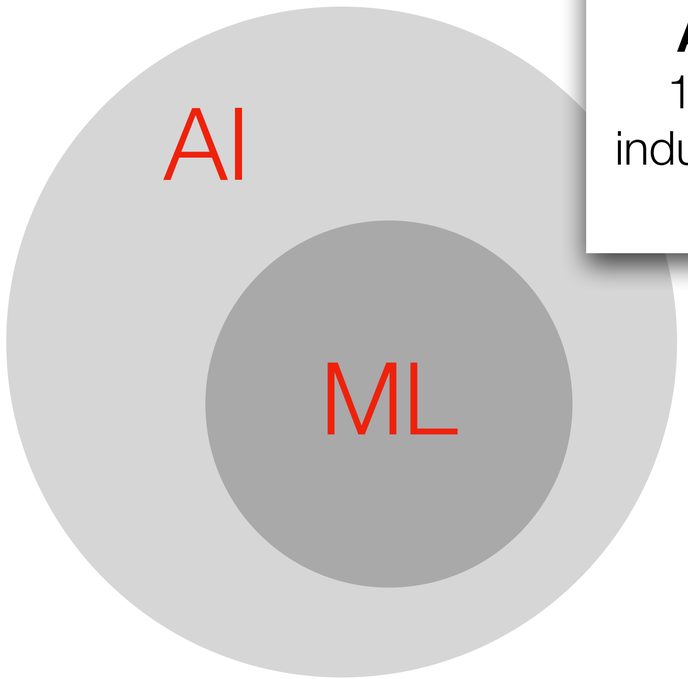

An Intrusion-Detection Model

1987

DOROTHY E. DENNING

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. SE-13, NO. 2, FEBRUARY 1987, 222-232.






**At least since mid-1990s**
1,000s of academic papers, industry white papers, blog posts

**ML is good at**

Finding patterns in data
Automating tasks
Generate synth content

**ML is not good at**

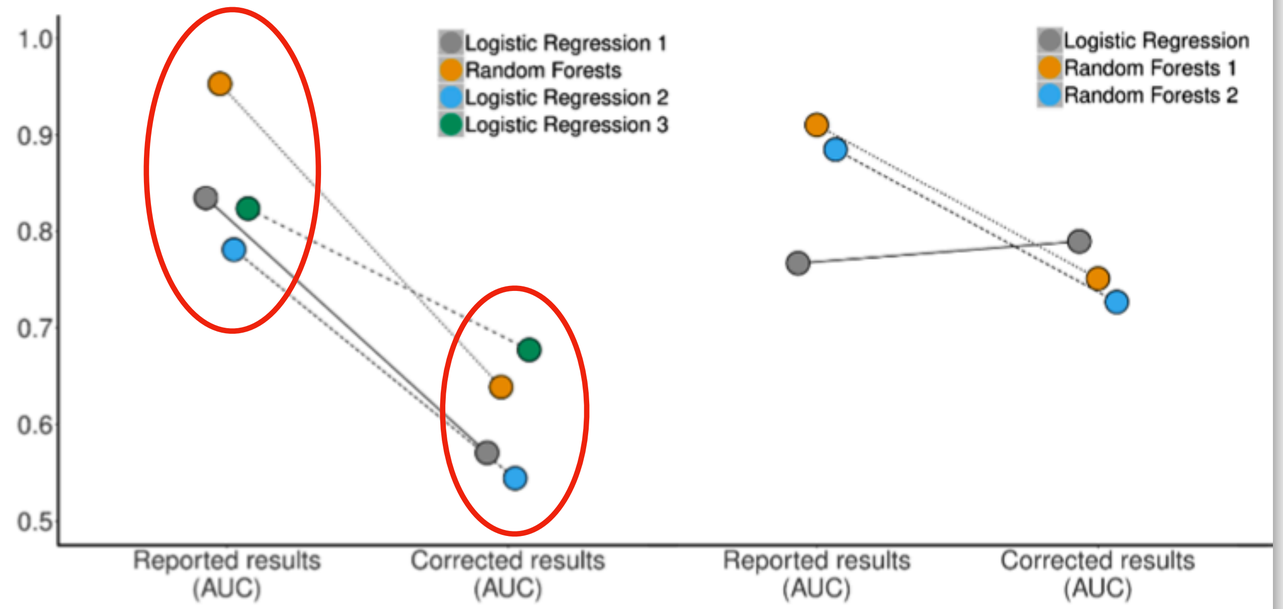
Predicting the future
Explaining outcomes
Overcome data biases

# The bug is everywhere

## (Ir)Reproducible Machine Learning: A Case Study

Sayash Kapoor, Arvind Narayanan
Princeton University
**Date**: August 2, 2021*

***Abstract*—The use of Machine Learning (ML) methods for prediction and forecasting has become widespread across the quantitative sciences. However, there are many known methodological pitfalls in ML-based research. As a case study of these pitfalls, we examine the subfield of civil war prediction in Political Science. Our main finding is that several recent studies published in top Political Science journals claiming superior performance of ML models over Logistic Regression models fail to reproduce. Our results provide two reasons to be skeptical of the use of ML methods in this research area, by both questioning their usefulness and highlighting the pitfalls of applying them correctly. Results identifying pitfalls in studies that use ML methods have appeared in at least eight quantitative science fields. However, we go farther than most previous research to investigate whether the claims made in the reviewed studies survive once the errors are corrected. We argue that there is a reproducibility crisis brewing in research fields that use ML methods and discuss a few systemic interventions that could help resolve it.**




| Paper | Muchlinski et al. | Colaresi and Mahmood |
|---|---|---|
| **Claim** | Random Forests model drastically outperforms Logistic regression models | Random Forests models drastically outperform Logistic regression model |
| **Error** | Incorrect imputation | Incorrect reuse of an imputed dataset |

# Civil war, love and security are hard to predict

## Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudin couples studies

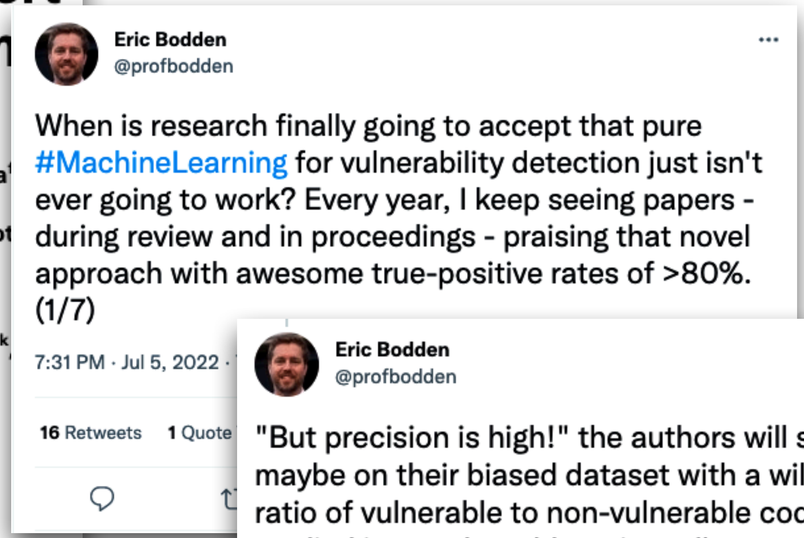
Samantha Joel[a,1], Paul W. Eastwick[b], Colleen J. Allison[c], Ximena B. Arriaga[d], Zachary G. Baker[e], Eran Bar-Kalifa[f], Sophie Bergeron[g], Gurit E. Birnbaum[h], Rebecca L. Brock[i], Claudia C. Brumbaugh[j], Cheryl L. Carmichael[k], Serena Chen[l], Jennifer Clarke[m], Rebecca J. Cobb[n], Michael K. Coolsen[o], Jody Davis[p], David C. de Jong[q], Anik Debrot[r], Eva C. DeHaas[c], Jaye L. Derrick[e], Jami Eller[s], Marie-Joelle Estrada[t], Ruddy Faure[u], Eli J. Finkel[v], R. Chris Fraley[w], Shelly L. Gable[x], Reuma Gadassi-Polack[y], Yuthika U. Girme[c], Amie M. Gordon[z], Courtney L. Gosnell[aa], Matthew D. Hammond[bb], Peggy A. Hannon[cc], Cheryl Harasymchuk[dd], Wilhelm Hofmann[ee], Andrea B. Horn[ff], Emily A. Impett[gg], Jeremy P. Jamieson[t], Dacher Keltner[k], James J. Kim[hh], Jeffrey L. Kirchner[ii], Esther S. Kluwer[ij,kk], Madoka Kumashiro[ll], Grace Larson[mm], Gal Lazarus[nn], Jill M. Logan[c], Laura B. Luchies[oo], Geoff MacDonald[hh], Laura V. Machia[pp], Michael R. Maniaci[qq], Jessica A. Maxwell[rr], Moran Mizrahi[ss], Amy Muise[tt], Sylvia Niehuis[n], Brian G. Ogolsky[uu], C. Rebecca Oldham[n], Nickola C. Overall[rr], Meinrad Perrez[vv], Brett J. Peters[ww], Paula R. Pietromonaco[xx], Sally I. Powers[xx], Thery Prok[x], Rony Pshedetzky-Shochat[nn], Eshkol Rafaeli[nn,yy], Erin L. Ramsdell[i], Maija Reblin[zz], Michael Reicherts[vv], Alan Reifman[n], Harry T. Reis[t], Galena K. Rhoades[aaa], William S. Rholes[bbb], Francesca Righetti[u], Lindsey M. Rodriguez[ccc], Ron Rogge[t], Natalie O. Rosen[ddd], Darby Saxbe[eee], Haran Sened[nn], Jeffry A. Simpson[s], Erica B. Slotter[fff], Scott M. Stanley[aaa], Shevaun Stocker[ggg], Cathy Surra[hhh], Hagar Ter Kuile[jj], Allison A. Vaughn[iii], Amanda M. Vicary[jjj], Mariko L. Visserman[hh,tt], and Scott Wolf[ii]

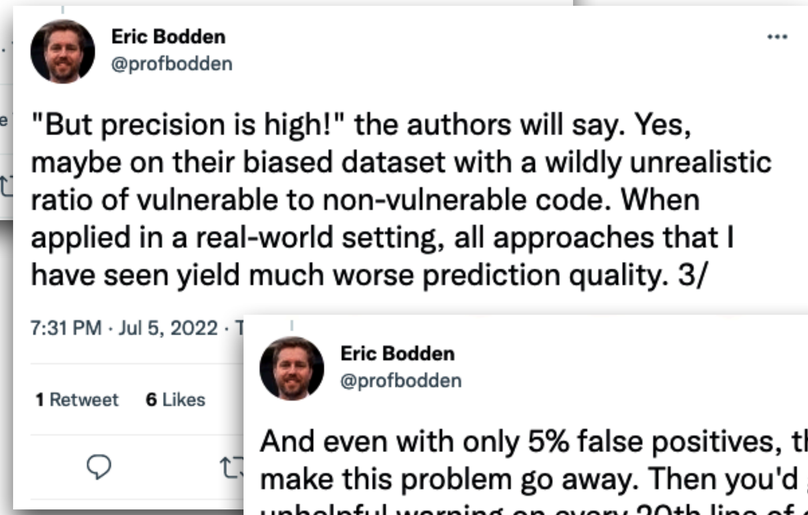
Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved June 8, 2020 (received for review September 30, 2019)

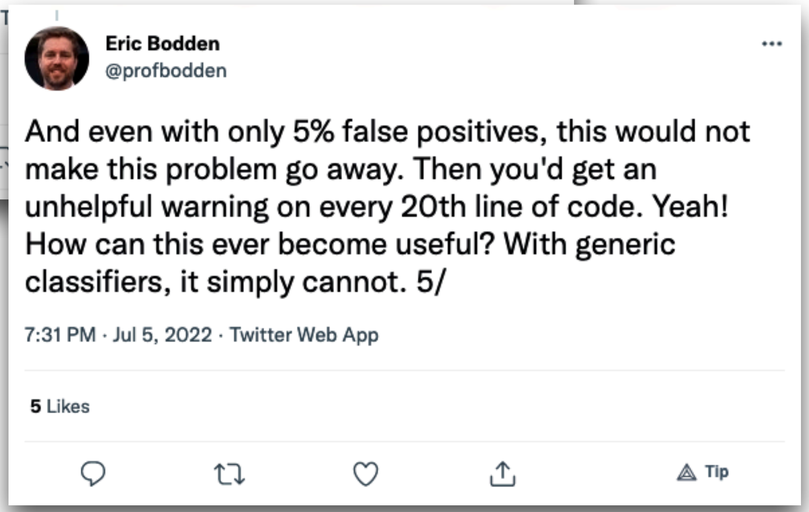
Given the powerful implications of relationship quality for health and well-being, a central mission of relationship science is explaining why some romantic relationships thrive more than others. This large-scale project used machine learning (i.e., Random Forests) to 1) quantify the extent to which relationship quality is predictable and 2) identify which constructs reliably predict relationship quality. Across 43 dyadic longitudinal datasets from 29 laboratories, the top relationship-specific predictors of relationship quality were perceived-partner commitment, appreciation, sexual satisfaction, perceived-partner satisfaction, and conflict. The top individual-difference predictors were life satisfaction, negative affect, depression, attachment avoidance, and attachment anxiety. Overall, relationship-specific variables predicted up to 45% of variance at baseline, and up to 18% of variance at the end of each study. Individual differences also performed well (21% and 12%, respectively). Actor-reported variables (i.e., own relationship-specific and individual-difference variables) predicted two to four times more variance than partner-reported variables (i.e., the partner's ratings

### Significance

What predicts how happy people are with their romantic relationships? Relationship science—an interdisciplinary field spanning psychology, sociology, economics, family studies, and communication—has identified hundreds of variables that purportedly shape romantic relationship quality. The current project used machine learning to directly quantify and compare the predictive power of many such variables among 11,196 romantic couples. People's own judgments about the relationship itself—such as how satisfied and committed they perceived their partners to be, and how appreciative they felt toward their partners—explained approximately 45% of their current satisfaction. The partner's judgments did not add information, nor did either person's personalities or traits. Furthermore, none of these variables could predict whose relationship quality would increase versus decrease over time.



# Why ML (predictive) systems often fail

**Top 10 ways your Machine Learning models may have leakage**

**Rayid Ghani, Joe Walsh, Joan Wang**

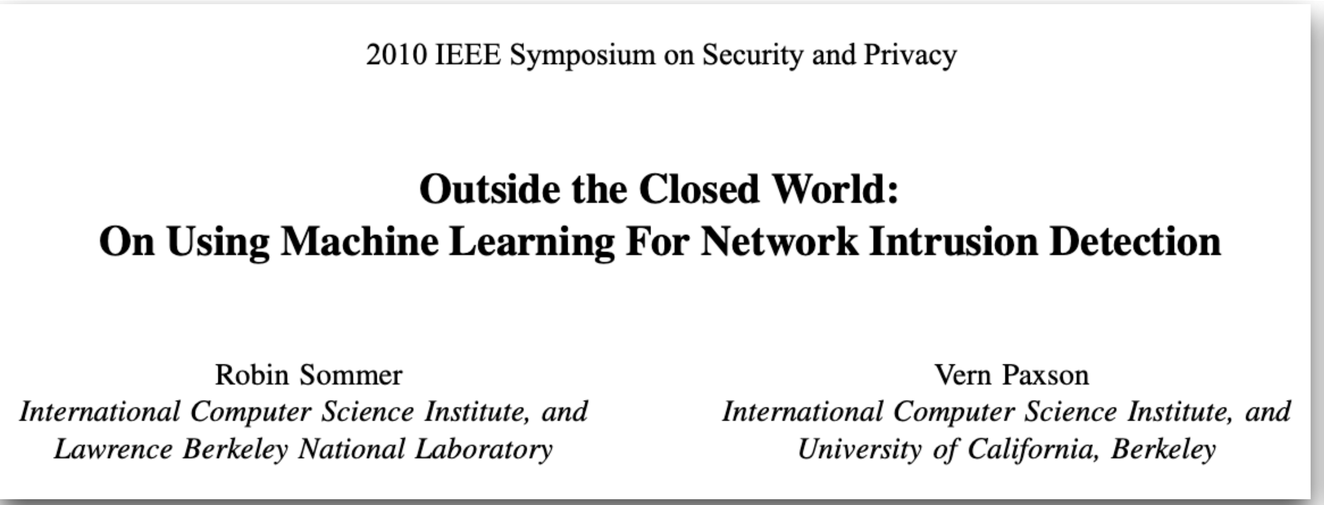

Methodological pitfalls

- Using a proxy for the outcome variable as a feature
- Using knowledge from the future
  - Incorrect reuse of an imputed dataset
  - Using all dataset for normalizations
  - Using all dataset for feature selection
  - K-cross validation with temporal data
  - …

1. Malware and packers
2. DARPA ’99 IDS Dataset

- Researchers not trained in proper use of ML
- Reproducibility is often impossible (no code, no data, no transparency)
- Three problems
  - **Bad modeling** — the spherical cow in the vacuum
  - **Bad data** — biased, imbalanced
  - **Bad evaluation** — experimental setting and validity of results

# Security needs GOOD problem modeling



2010 IEEE Symposium on Security and Privacy

**Outside the Closed World:**
**On Using Machine Learning For Network Intrusion Detection**

Robin Sommer
*International Computer Science Institute, and*
*Lawrence Berkeley National Laboratory*

Vern Paxson
*International Computer Science Institute, and*
*University of California, Berkeley*

**Problem: tell apart good from bad**

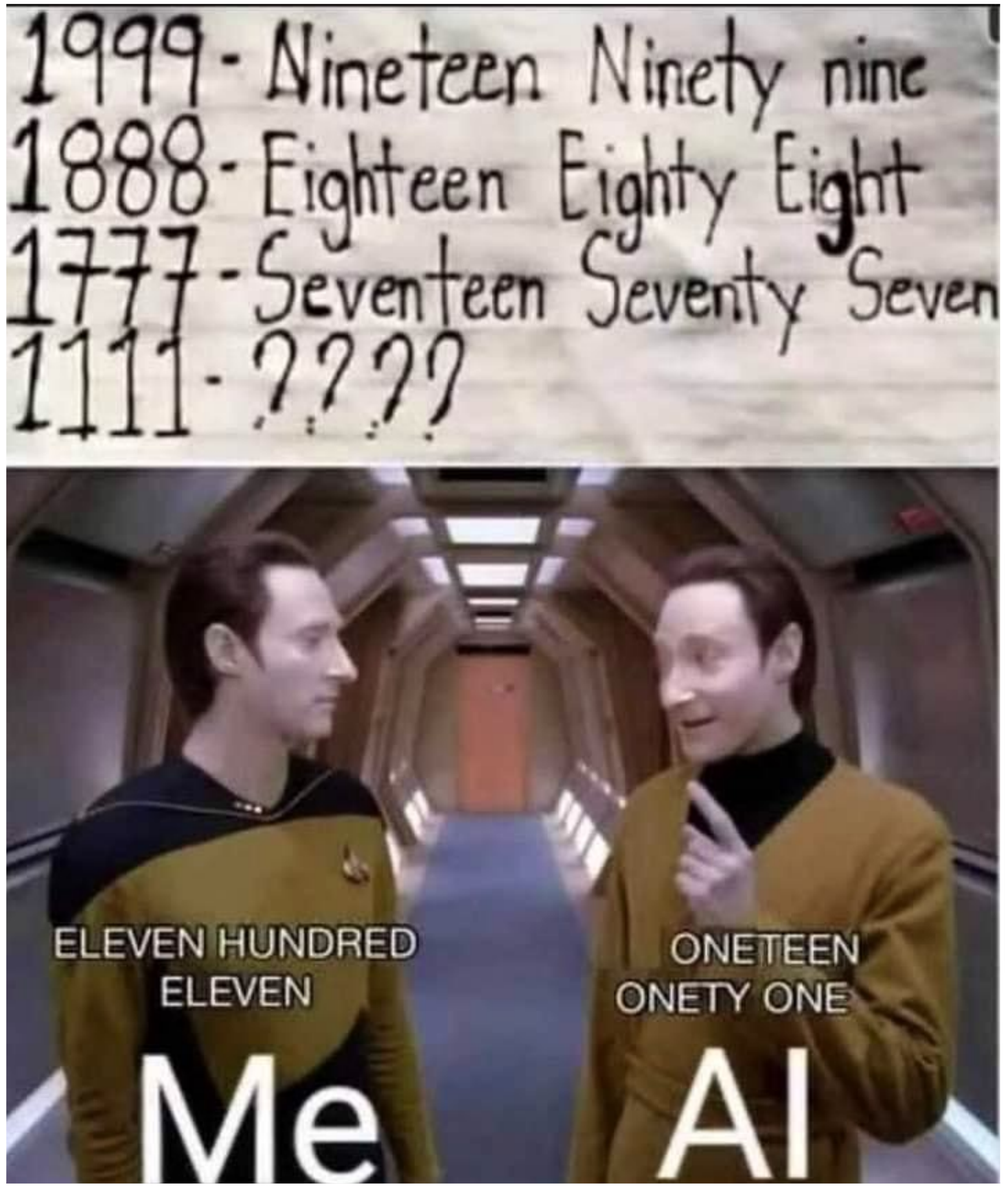
Why *thousands* of academic papers have designed ML-based systems to solve this problem with presumably awesome detection quality, yet no one works well in the wild?

**Security is different**

Some domains (security is one of them) have characteristics that are not well aligned with the requirements of ML

- ☑ Very high cost of errors
- ☑ Lack of quality training data
- ☑ Semantic gap between results and operational interpretation
- ☑ High variability in input data
- ☑ Difficulties to conduct sound evaluation

# Security needs GOOD data




**Data quality properties we don't know well how to deal with**

- Representativeness
  - Spatial — all classes
  - Time — concept drift
- Dealing with sensitive attributes

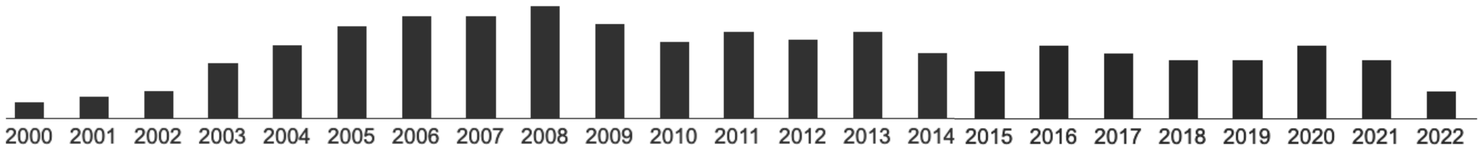
**Advice**

- Keep the scope narrow
- Understand the problem before you look at the data
- Understand the environment that the system targets

Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory

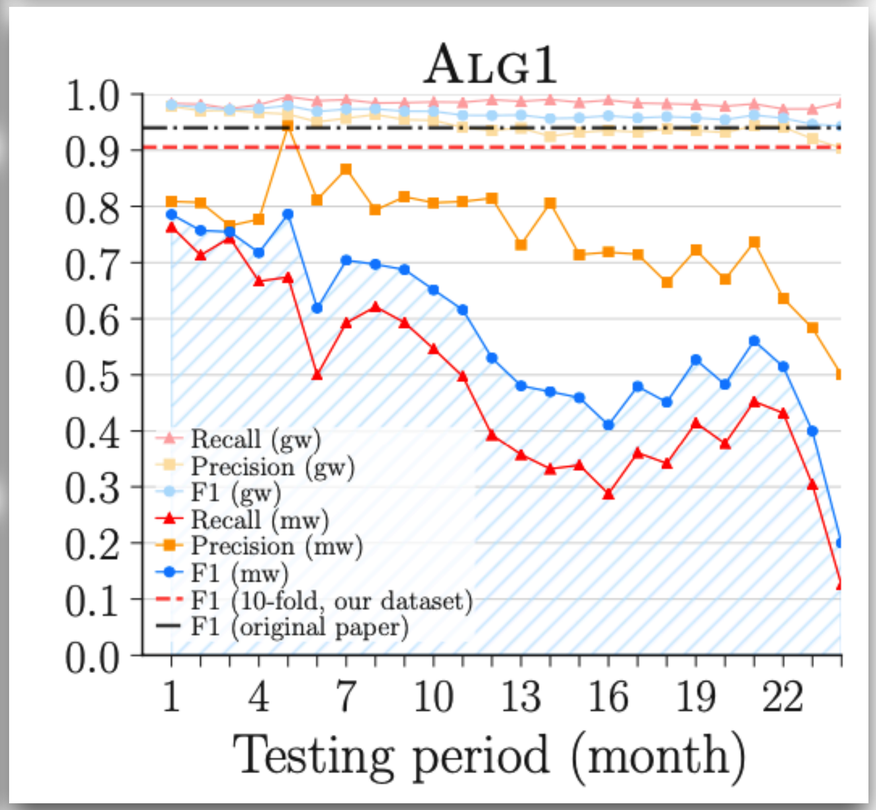
JOHN McHUGH
Carnegie Mellon University

November 2000



# Security needs GOOD evaluation frameworks

| | |
|---|---|
| **Avoid leakage** | • Using a proxy for the outcome variable as a feature<br>• Don't use knowledge from the future |
| **Temporal bias** | • a.k.a. dataset shift, a.k.a. concept drift<br>• Inflates performance (e.g., training with examples of new malware families that did not really exist at the time of training the classifier) |
| **Spatial bias** | • Use correct percentages of prevalence of each class<br>• Unrealistic distribution (e.g., too many attacks, too many malware samples) inflates performance<br>• Estimating correct ratios might be hard |
| **Time-aware performance metrics** | • The world changes and the system is no longer ok<br>• Evaluate this<br>• Measure retraining / active learning |

# REAL-WORLD security requires VERY GOOD models

Prevalence of security events

Detector's sensitivity

$$P(S|R) = \frac{P(S) \cdot P(R|S)}{P(S) \cdot P(R|S) + P(\neg S) \cdot P(R|\neg S)}.$$

Detector's FPR

Probability of a true security (**S**) event when we see an alert (**R**)

The Base-Rate Fallacy and the Difficulty of Intrusion Detection

STEFAN AXELSSON
Ericsson Mobile Data Design AB

August 2000

| | |
|---|---|
| P(S) | 1/10,000 |
| P(R\|S) | 0,99 (99%) |
| P(R\|¬S) | 0,01 (1%) |
| P(S\|R) | 0,0098 (**~1%**) |

O.O

# REAL-WORLD security requires VERY GOOD models

## Takeaway points

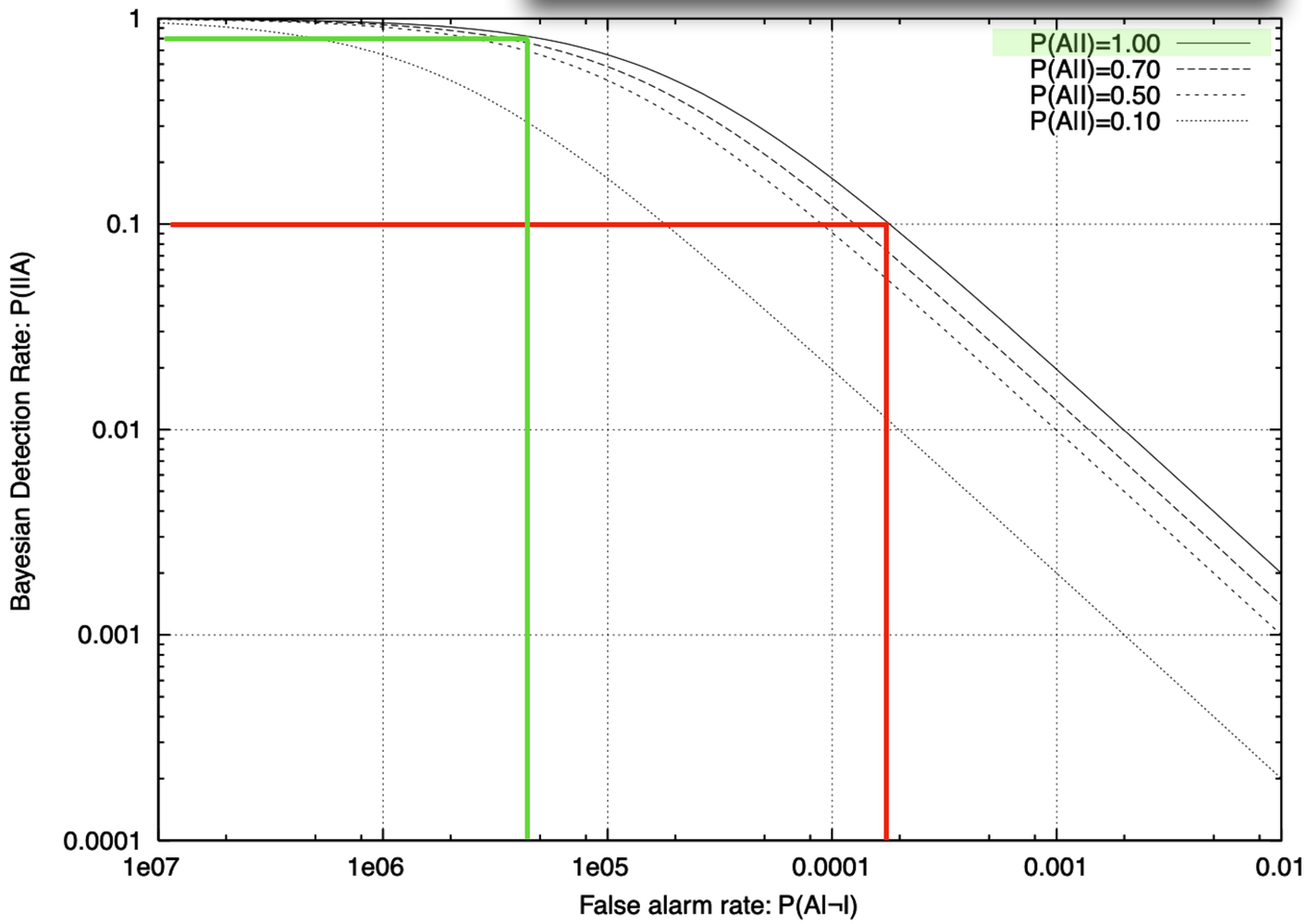
- Limiting perf factor is not the ability to correctly identify attacks, but *the ability to suppress false alarms*.
- We need *extremely low FPR* for detectors to be good
- More data implies more FPs unless security events grow proportionally, which is not the case

Much academic research ignores this

The Base-Rate Fallacy and the Difficulty of Intrusion Detection

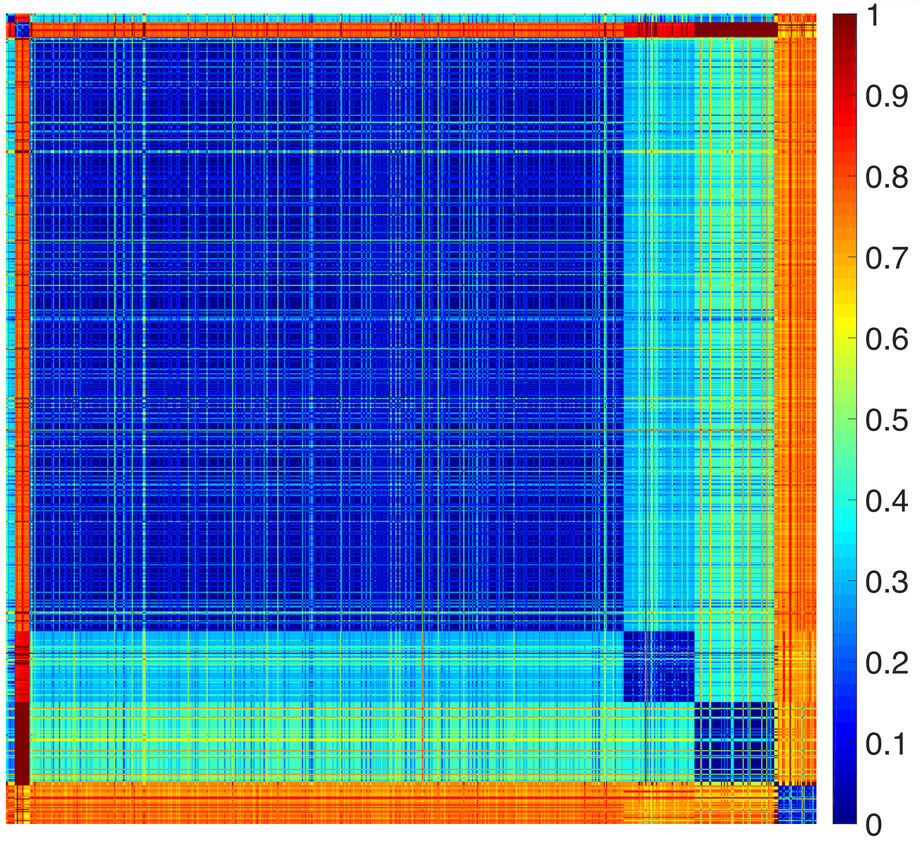
STEFAN AXELSSON
Ericsson Mobile Data Design AB

August 2000

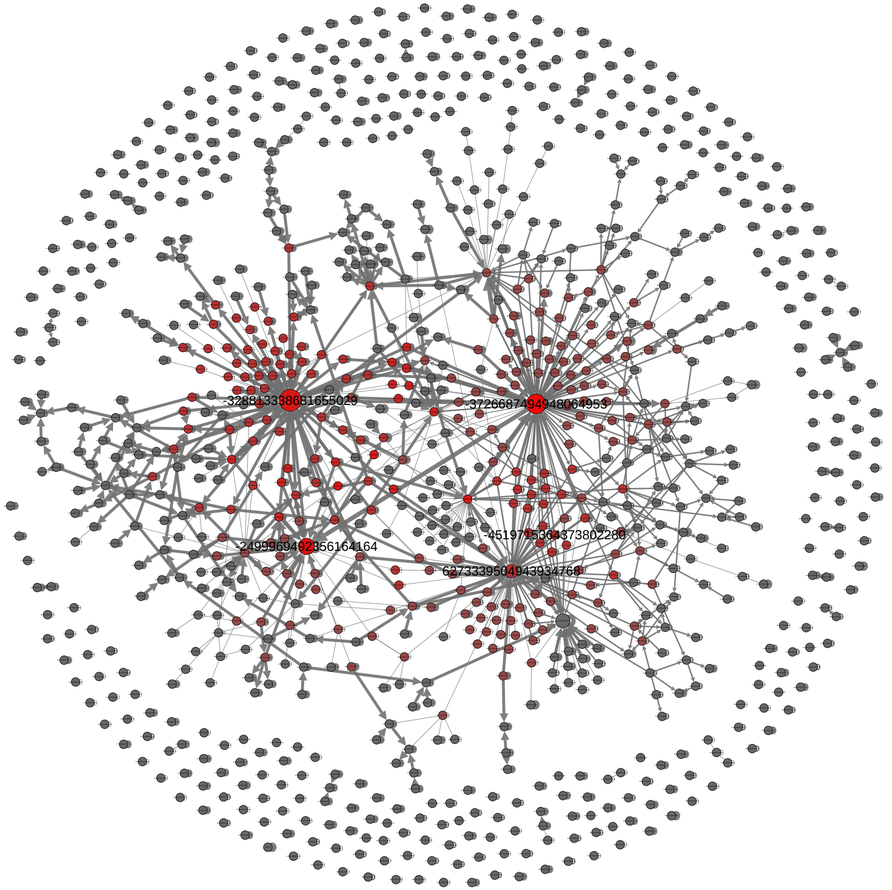


# Malware classification

## What is a malware family?




Distance matrix for 824 ransomware samples belonging to 7 families



AHG subgraph for the `oldboot` family. Edges represent code genomes that are common in more than 70% of family samples.

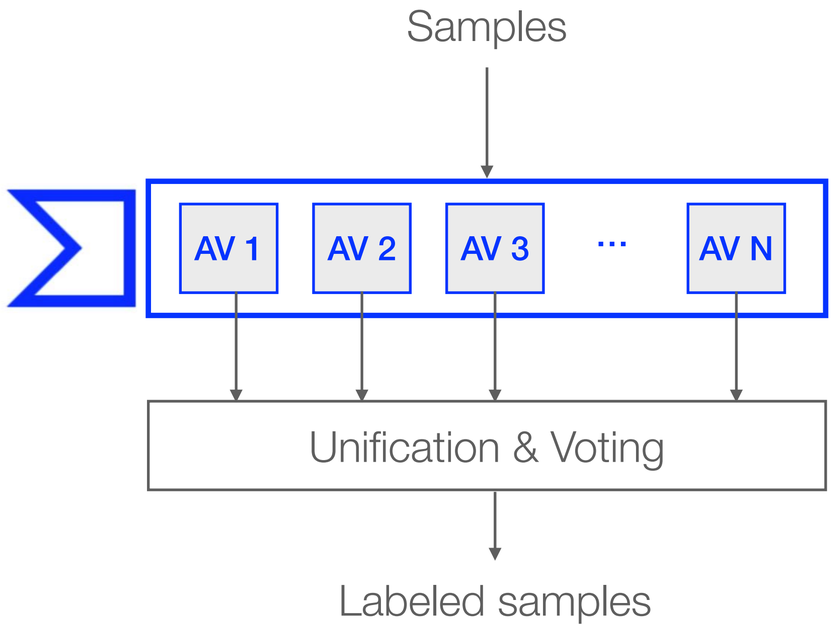
# Malware classification

## The labeling problem

10y of Android malware classification in top conferences

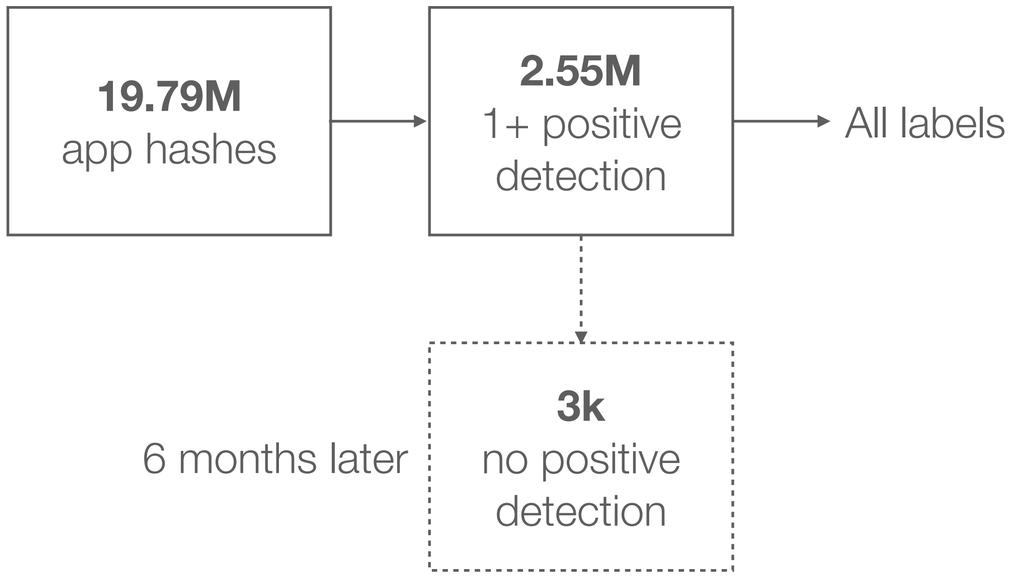
| Dataset | # Papers | Labl | # Fams | # Samples | VT | Class. Mthd | Collection Time (20xx) | Papers Using the Dataset by Year (20xx) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Malgenome (MG) [10] | 34 (42%) | Y | 49 | 1260 | N | Manual | 10–11 | | | 4 | 4 | 3 | 9 | 6 | 4 | 2 | | 1 |
| Contagio (repo) [90] | 17 (21%) | N | - | UNK-395 | N | N/A | 11- | | | 2 | 2 | 2 | 3 | 2 | 4 | 1 | | 1 |
| Drebin [12] | 14 (17%) | Y | 179 | 5560 | Y | Own | 10–12 | | | | | 1 | | 4 | 7 | 1 | | 1 |
| VirusShare [91] | 10 (12%) | N | - | 11 K–35 K | N/A | N/A | N/A | | | | | 1 | 1 | 1 | 4 | 1 | 1 | 1 |
| VT-malware [92] | 9 (11%) | N | - | 2 K-238 K | Y | N/A | N/A | | | | | | | | 4 | 3 | 1 | |
| DroidBench [93] | 4 (5%) | N | - | UNK-200 | N | N/A | N/A | | | | | | 2 | 1 | | | | |
| AMD [94] | 3 (3.7%) | Y | 71 | 24 K | Y | Own | 10–16 | | | | | | | | 1 | 1 | | 2 |
| SandDroid (repo) [95] | 3 (3.7%) | N | - | 112-38 K | N/A | N/A | N/A | | | | | 2 | 1 | | | | | |
| Androzoo (repo) [96] | 3 (3.7%) | N | - | 3 K–13 K | Y | N/A | N/A | | | | | | | | | 1 | 2 | |
| GPlay-mal (repo) [97] | 3 (3.7%) | N | - | UNK-27 | Y | N/A | N/A | | | 1 | | | | | 1 | | | 1 |
| Andrubis [98] | 3 (3.7%) | N | - | 422 K | Y | N/A | N/A | | | | | | | 1 | 1 | | | 2 |
| DARPA [99] | 2 (2.5%) | N | - | 11 | N | N/A | N/A | | | | | | 1 | 1 | | | | |
| RmvDroid [23] | 1 (1.2%) | Y | 56 | 9.1 K | Y | AVClass | 14,15,17 | | | | | | | | | | 1 | |
| alt markets (repo) [100–103] | 1 (1.2%) | N | - | 2 K | Y | N/A | N/A | | | | | | | | 1 | | | |
| AndroMalTeam [104] | 1 (1.2%) | N | - | 34 | N | N/A | N/A | | | | | | | | | 1 | | |
| Marvin [22] | 1 (1.2%) | N | - | 15 K | Y | N/A | N/A | | | | | | | | 1 | | | |
| Github (repos) [105–109] | 1 (1.2%) | N | - | 5 | Y | N/A | N/A | | | | | | | | 1 | | | |
| AndroTotal [110] | 1 (1.2%) | N | - | 4.1 K | N | N/A | N/A | | | | | | 1 | | | | | |
| AndRadar [111] | 1 (1.2%) | N | - | N/A | Y | N/A | N/A | | | | | | 1 | | | | | |
| CICAndMal17 [112] | 1 (1.2%) | N | - | 4.3 K | Y | N/A | N/A | | | | | | | | | | 1 | |
| Wang et al. [113] | 1 (1.2%) | N | - | 4.5 M | Y | AVClass | 09-17 | | | | | | | | | | 1 | |
| AndroPUP [114] | 1 (1.2%) | N | - | 4.6 M | Y | N/A | N/A | | | | | | | | | | | 1 |
| Spreitzenbarth et al. [115] | 1 (1.2%) | N | - | 7.5 K | Y | N/A | N/A | | | | | | | 1 | | | | |
| MobiSec Lab [116] | 1 (1.2%) | N | - | 2 K | N/A | N/A | N/A | | | | | | | 1 | | | | |
| HackingTeam [117] | 1 (1.2%) | N | - | 1 | N | N/A | N/A | | | | | | | 1 | | | | |
| ashishb [118] | 1 (1.2%) | N | - | 298 | N | N/A | N/A | | | | | | | | | 1 | | |
| M0Droid [17] | 1 (1.2%) | N | - | N/A | Y | N/A | N/A | | | | | | | 1 | | | | |
| Canfora et al. [119] | 1 (1.2%) | N | - | 2 | N | N/A | N/A | | | | | | | | 1 | | | |
| Sherlock vs. Moriarty [120] | 1 (1.2%) | N | - | 12 | Y | N/A | 2016 | | | | | | | | | 1 | | |
| Companies [121–126] | 12 (15%) | N | - | 69-1.5 K | N/A | N/A | N/A | | | 1 | 1 | 3 | 2 | 1 | 4 | | | |
| Custom [16,21,127–132] | 9 (11%) | N | - | 2-1 K | Y/N | N/A | N/A | | 2 | | 2 | | 3 | 1 | | 1 | | |
| unknown [133–135] | 3 (3.7%) | N | - | 20–362 | Y/N | N/A | N/A | | | | 1 | 1 | | | | | | |
| **Summary** | **81** | **-** | **49–179** | **UNK-422 K** | **-** | **-** | **10-UNK** | **0** | **2** | **6** | **8** | **9** | **13** | **11** | **17** | **7** | **5** | **4** |

Training & evaluation needs accurate labels!

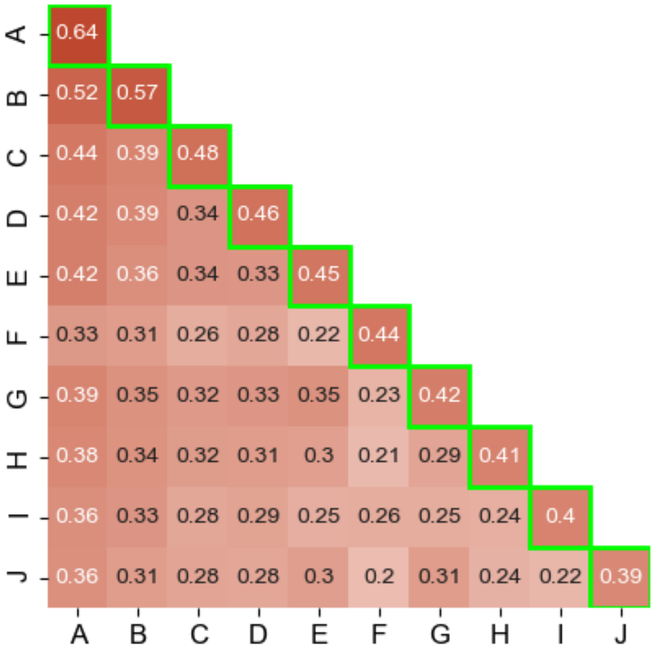


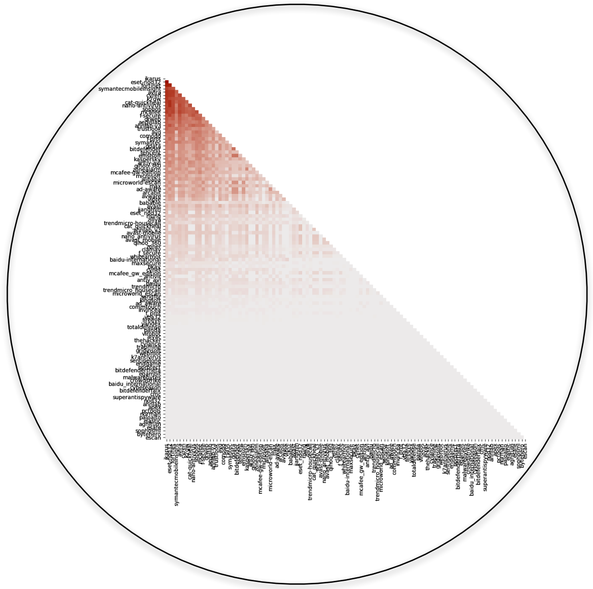
# Malware classification

## The labeling problem



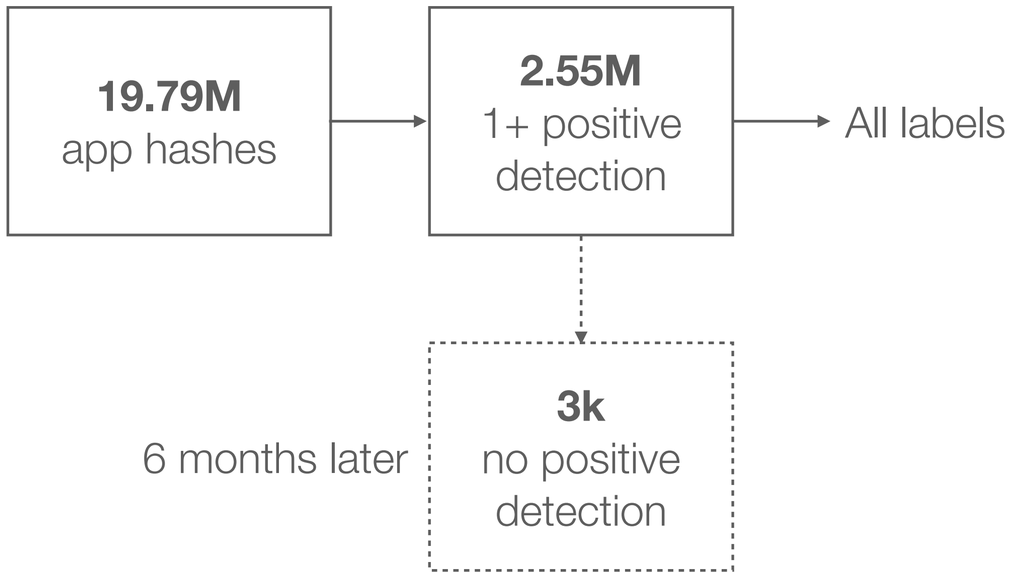

Coverage and cross-coverage



# Malware classification

## The labeling problem




Lots of labeling issues

**Inconsistencies (same engine, different labeling convention)**

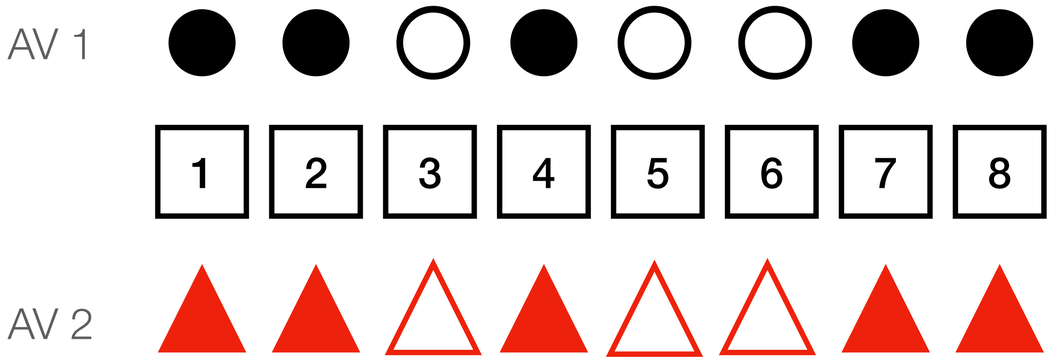
- AdWare vs. Adware vs. AdWo
- ANDR vs. Android vs. AndroidOS
- Arbitrary use of class/type and family sub labels
- Different labeling scheme

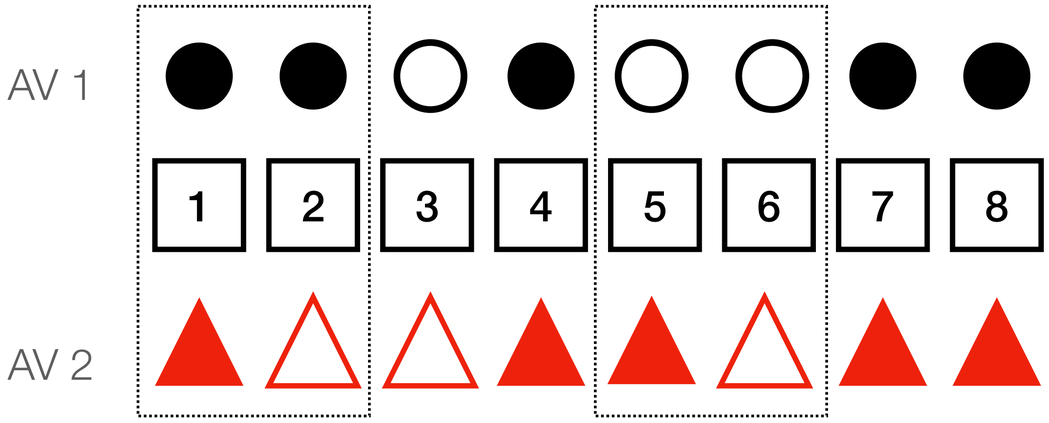
**Generic labels**

- Android.Generisk
- Other:Android.Reputation.1
- Android.Agent.GEN23333
- Android.Gen-Pua95BB2BF2

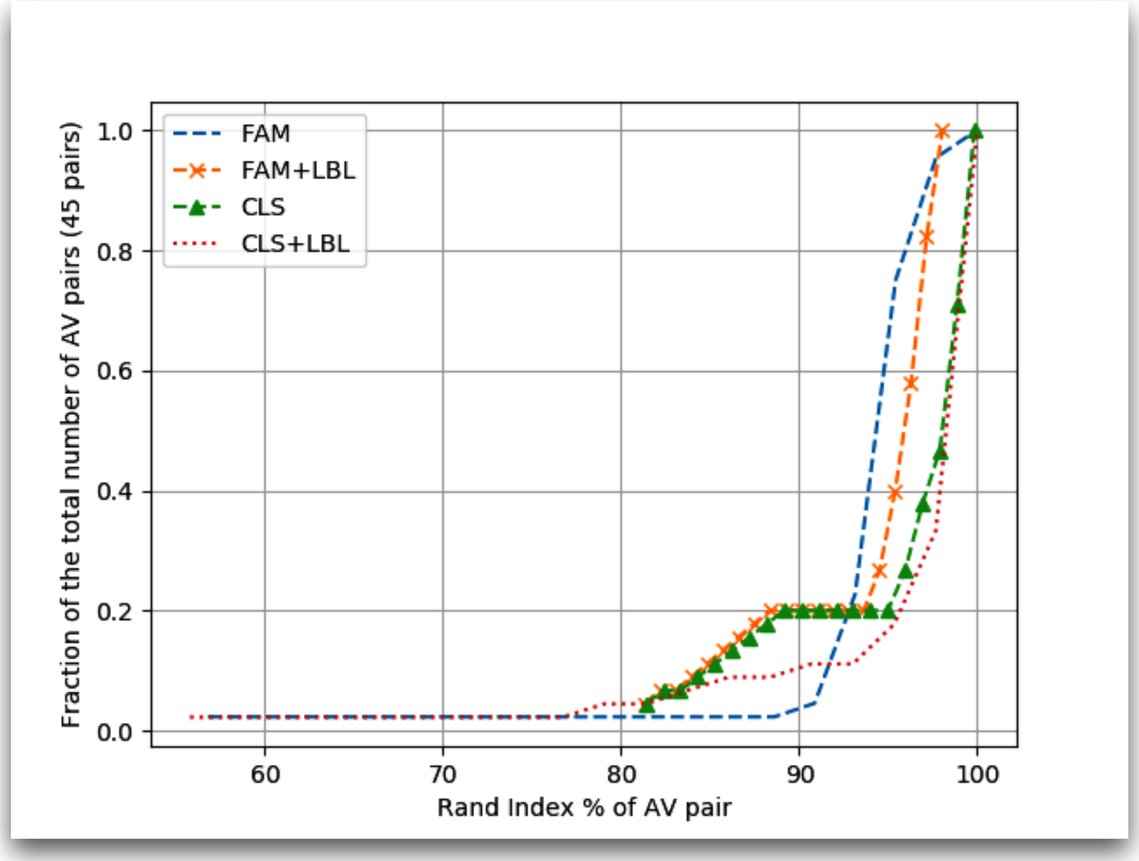
# Malware classification

## Measuring coherence




Do different engines agree on what is a malware family?



# Attribution is HARD

## Mixed Signals: Analyzing Software Attribution Challenges in the Android Ecosystem

Kaspar Hageman, Álvaro Feal, Julien Gamba, Aniketh Girish, Jakob Bleier, Martina Lindorfer, Juan Tapiador, Narseo Vallina-Rodriguez

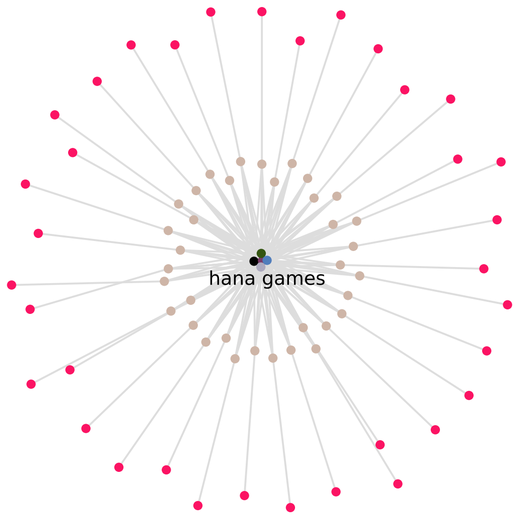
**Attribution Signals**

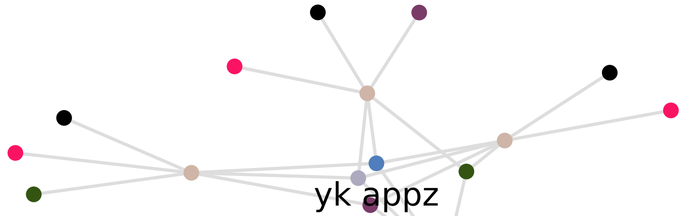
| Market | First crawl (December 2019 – May 2021) | | | Second crawl (June 2021 – October 2021) | |
|---|---|---|---|---|---|
| | Market entries | PkgN | APK SHA-256 | Market entries | Pk… |
| Google Play | 1,078,935 | 728,312 | 794,989 | 552,142 | 307, |
| APKMonk | 420,399 | 310,050 | 386,961 | 703,839 | 435, |
| Tencent | 203,881 | 128,665 | 142,637 | 18,651 | 2, |
| Baidu | 12,547 | 10,855 | 9,819 | 46,901 | 10, |
| APKMirror | 10,927 | 883 | 9,681 | 6,214 | 1, |
| F-Droid | 4,146 | 1,260 | 2,754 | 56,532 | 3, |

TABLE 4: Percentage of unique market entries throughout the dataset with missing attribution signals on the different markets (— indicates we did not collect a specific signal).

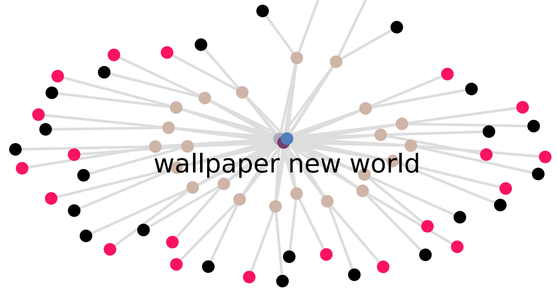
| | Attribution Signal | Google Play * | APKMonk | Tencent | Baidu | APKMirror | F-Droid* |
|---|---|---|---|---|---|---|---|
| Market | Developer name | <0.01% | <0.1% | <0.01% | — | 0% | 90.3% |
| | Developer website | 33.7% | — | — | — | — | 58.6% |
| | Developer email | <0.01% | — | — | — | — | 89.7% |
| | Developer address | 44.3% | — | — | — | — | — |
| | Privacy policy URL | 17.7% | — | — | — | — | — |
| Cert RDN | commonName | 7.1% | 9% | 9.1% | 6.6% | 11.9% | <0.1% |
| | organization | 17.6% | 25.4% | 21.1% | 21% | 8.5% | 0.1% |
| | org.Unit | 27.6% | 36.7% | 30.9% | 21.9% | 23.5% | 0.2% |
| | locality | 27.2% | 35.9% | 29.3% | 20.9% | 16.8% | 0.1% |
| | state | 28.8% | 38.8% | 32.6% | 24.2% | 20.2% | 0.1% |
| | country | 21.8% | 30.3% | 27% | 24% | 15.4% | 0.1% |

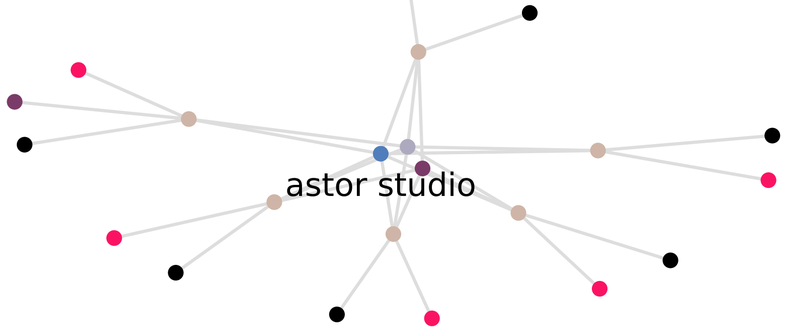
# Attribution is HARD

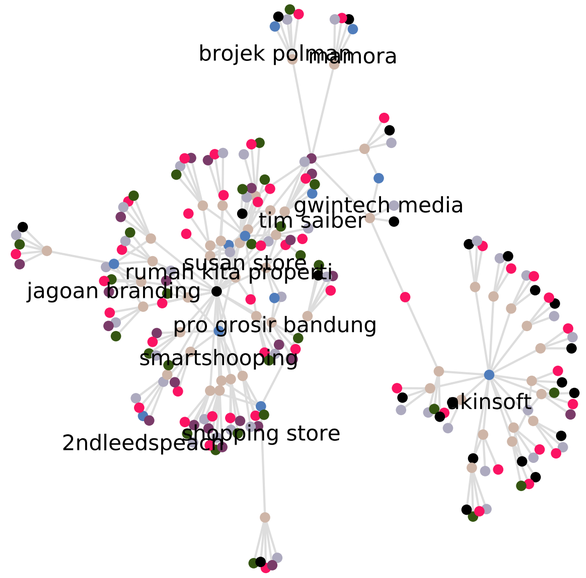




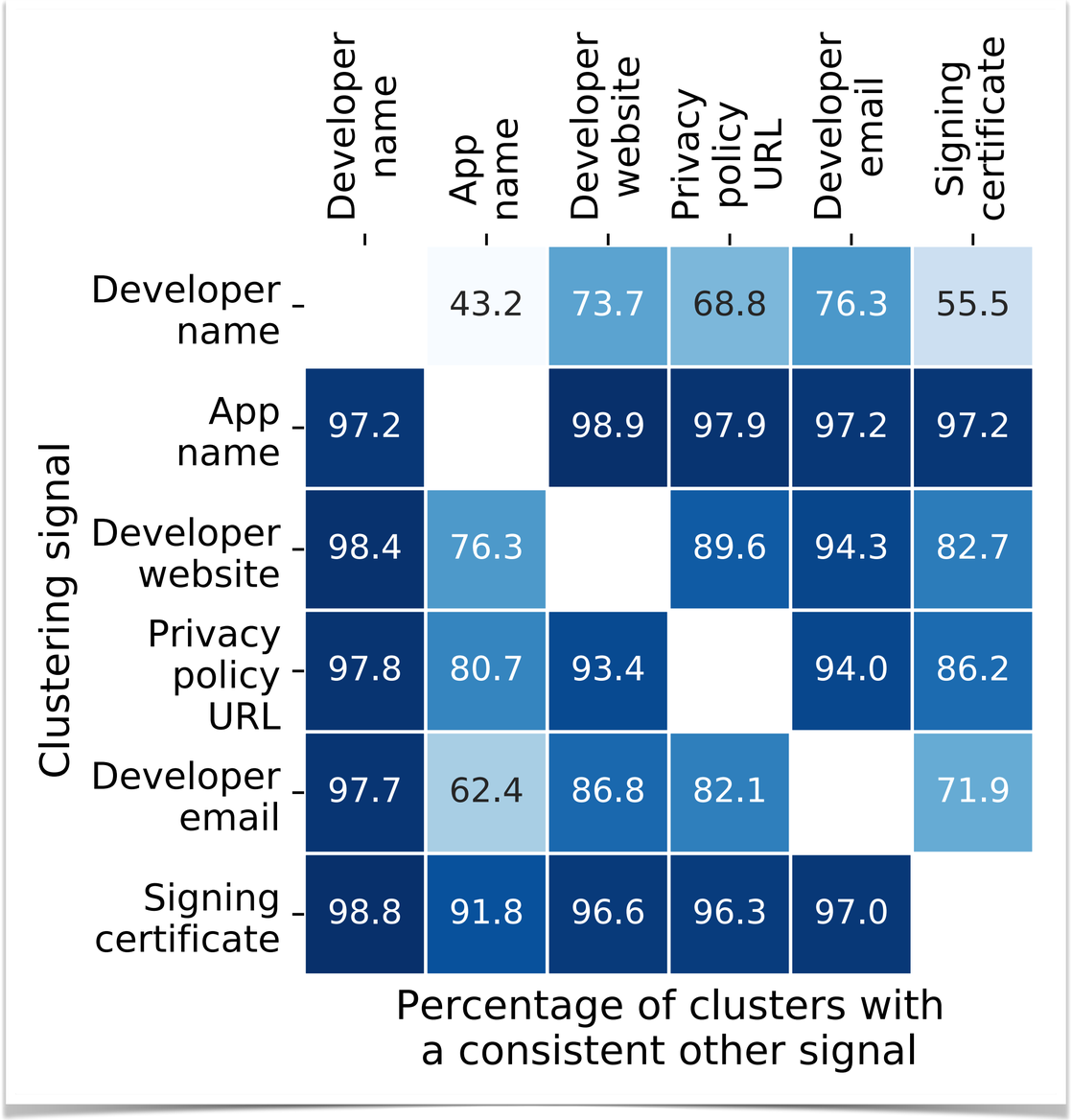



**Attribution Graphs**

# Attribution is HARD




## TL;DR

Not all signals are available for all samples

Signals might be volatile

Signals are not consistent across actors or across markets

Code similarity helps but not much

No matter how you cluster, you will make attribution errors

# AI and New Threats: A Skeptics' Guide

## Or, What Civil War, Love and Infosec Have in Common

# Thank you!


**Juan Tapiador**
UC3M
0xjet.github.io
Twitter: @0xjet