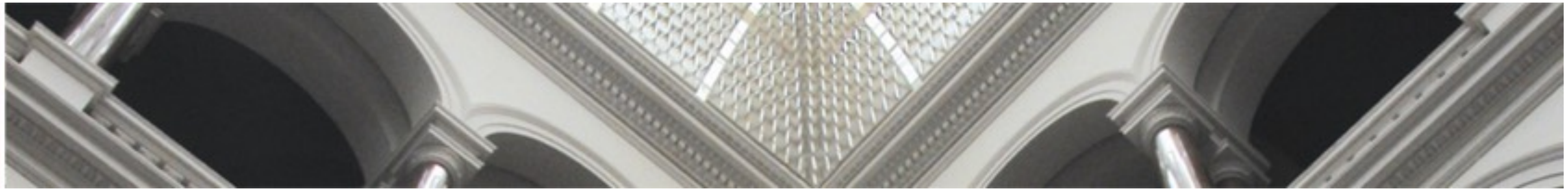


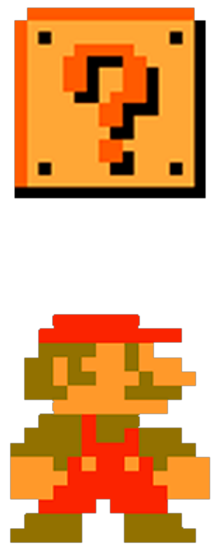
Dos and Don'ts of Machine Learning in Computer Security

Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke,
Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, Konrad Rieck

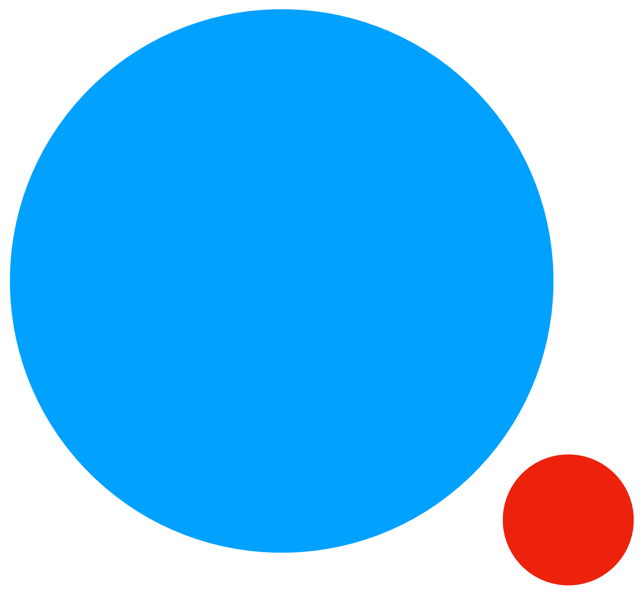
USENIX Security 2022

Machine Learning already solved
many problems in computer security

Unfortunately not... 🥺

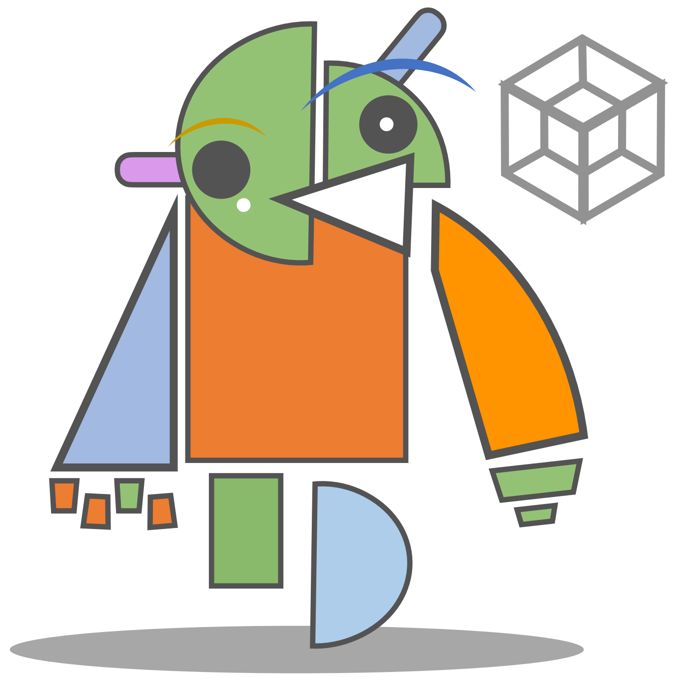


Motivation—Historical Examples



Network intrusion detection: The base rate fallacy

- Intrusion detectors should have low false positive rates (FPR)
- ‘Low’ FPR often still corresponds to large number of false positives

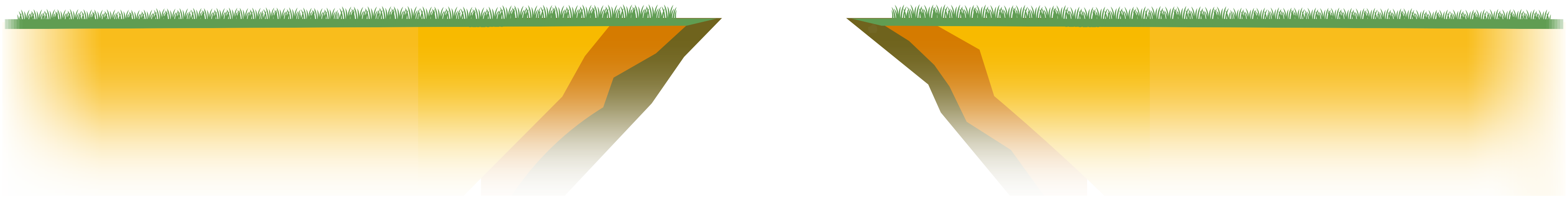


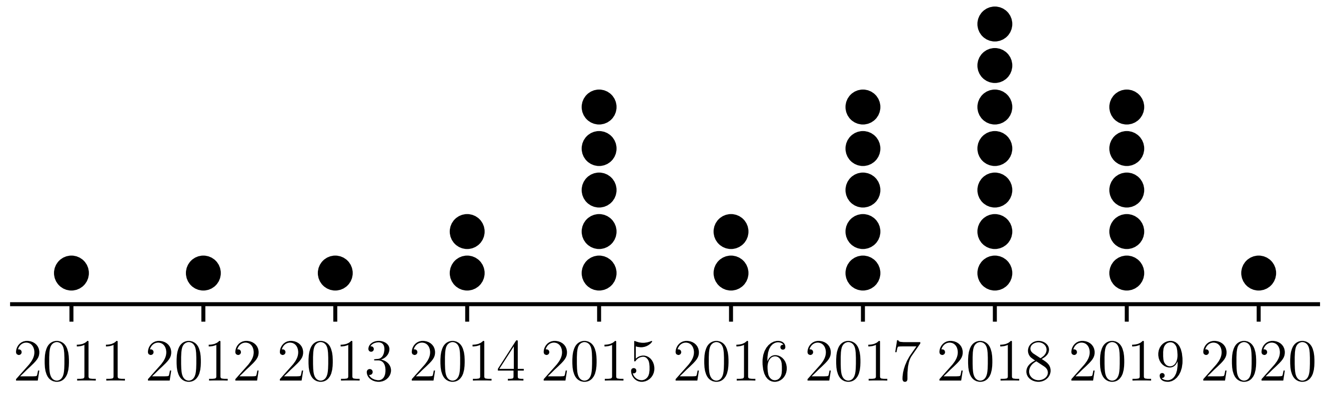
Android malware detection: Spatio-temporal bias inflating performance

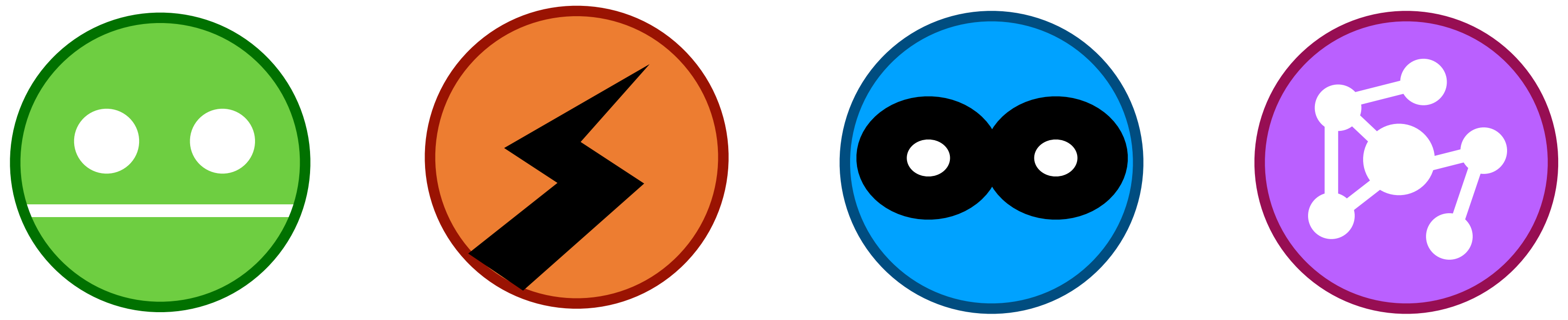
- Models trained with access to ‘future’ information
- Unrealistic class balance inflates performance

Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM TISSEC, 2000.*
Pendlebury et al. TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time. *USENIX Security, 2019.*

Overview







1. **Identification of common pitfalls**
    - 10 subtle issues affecting ML for security
    - Recommendations for avoiding them

2. **Survey on the prevalence of pitfalls**
    - Review of 30 top papers in security
    - Pitfalls are widespread

3. **Case studies demonstrating impact of pitfalls**
    - Mobile malware detection
    - Vulnerability discovery
    - Source code authorship attribution
    - Network intrusion detection

Important remark

This work should not be interpreted as a finger-pointing exercise. Any work mentioned as having pitfalls still has important contributions and we identify pitfalls in our own work also.

ML Pipeline and Pitfalls

Data Collection and Labeling

- **P1** Sampling bias
- **P2** Label Inaccuracy

System Design and Learning

- **P3** Data snooping
- **P4** Spurious correlations
- **P5** Biased parameters

Performance Evaluation

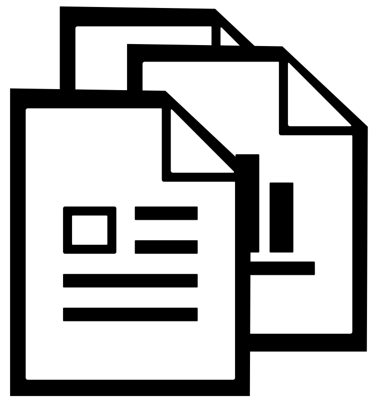
- **P6** Inappropriate baselines
- **P7** Inappropriate measures
- **P8** Base rate fallacy

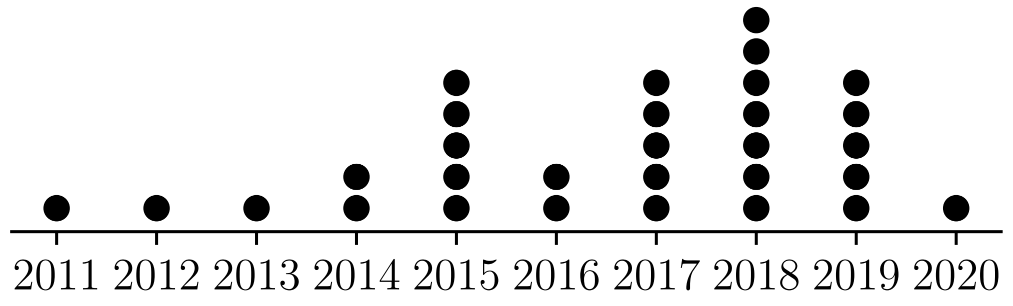
Deployment and Operation

- **P9** Lab-only evaluation
- **P10** Inappropriate threat model

Prevalence Study

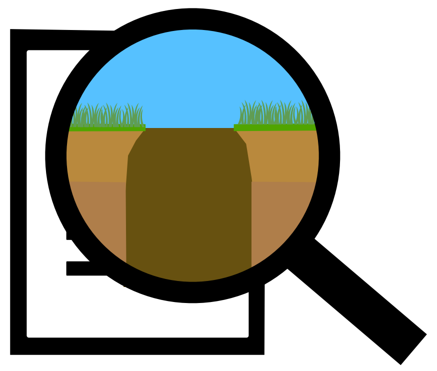
1. Paper Selection





2. Review Process

Pitfall is either…



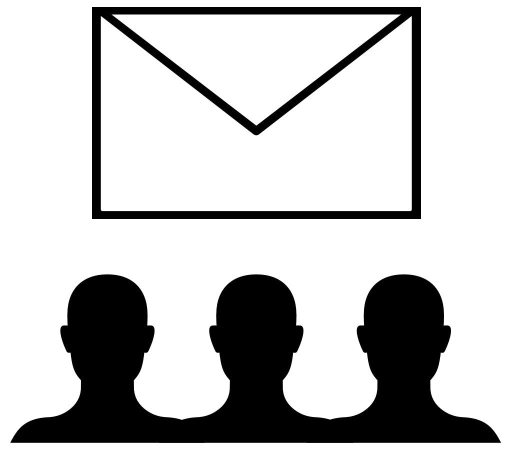
present (but discussed)

partly present (but discussed)

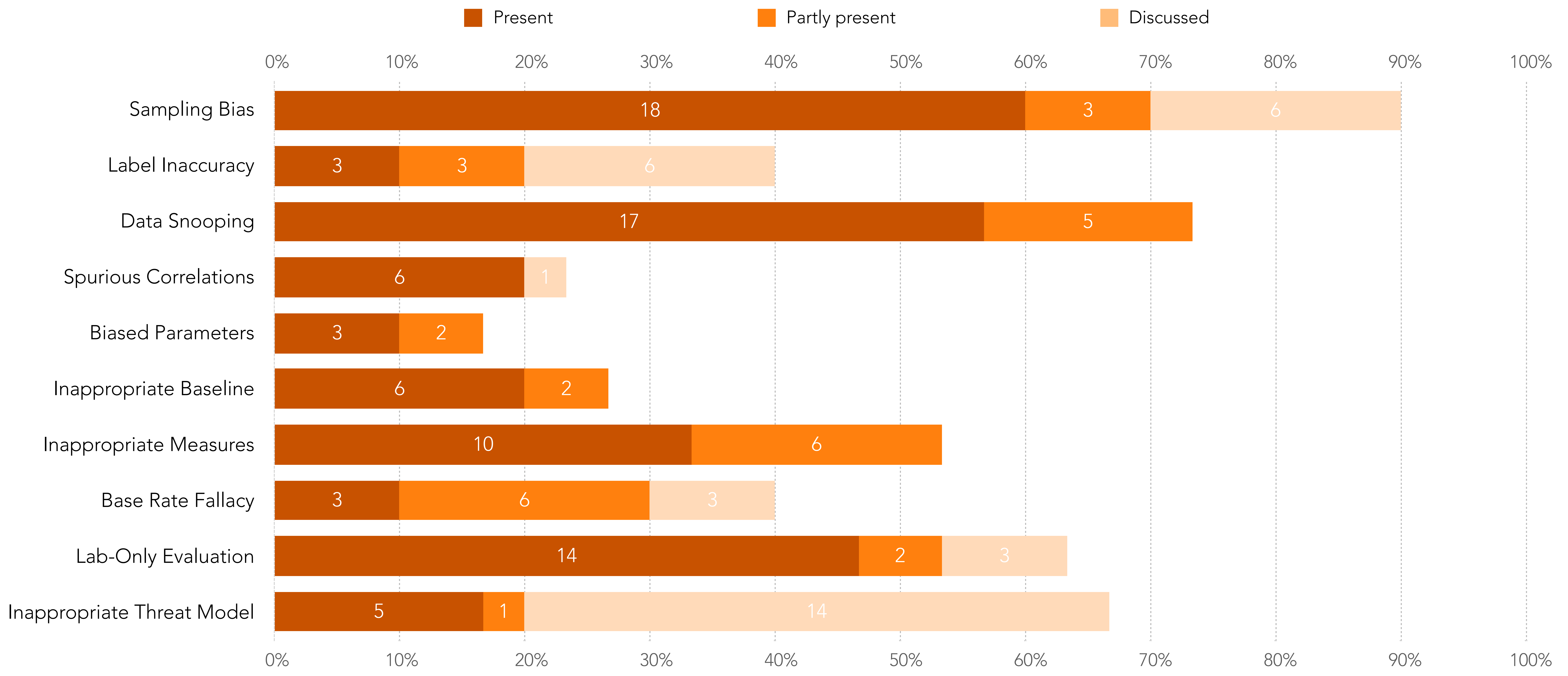
not present

unclear from text

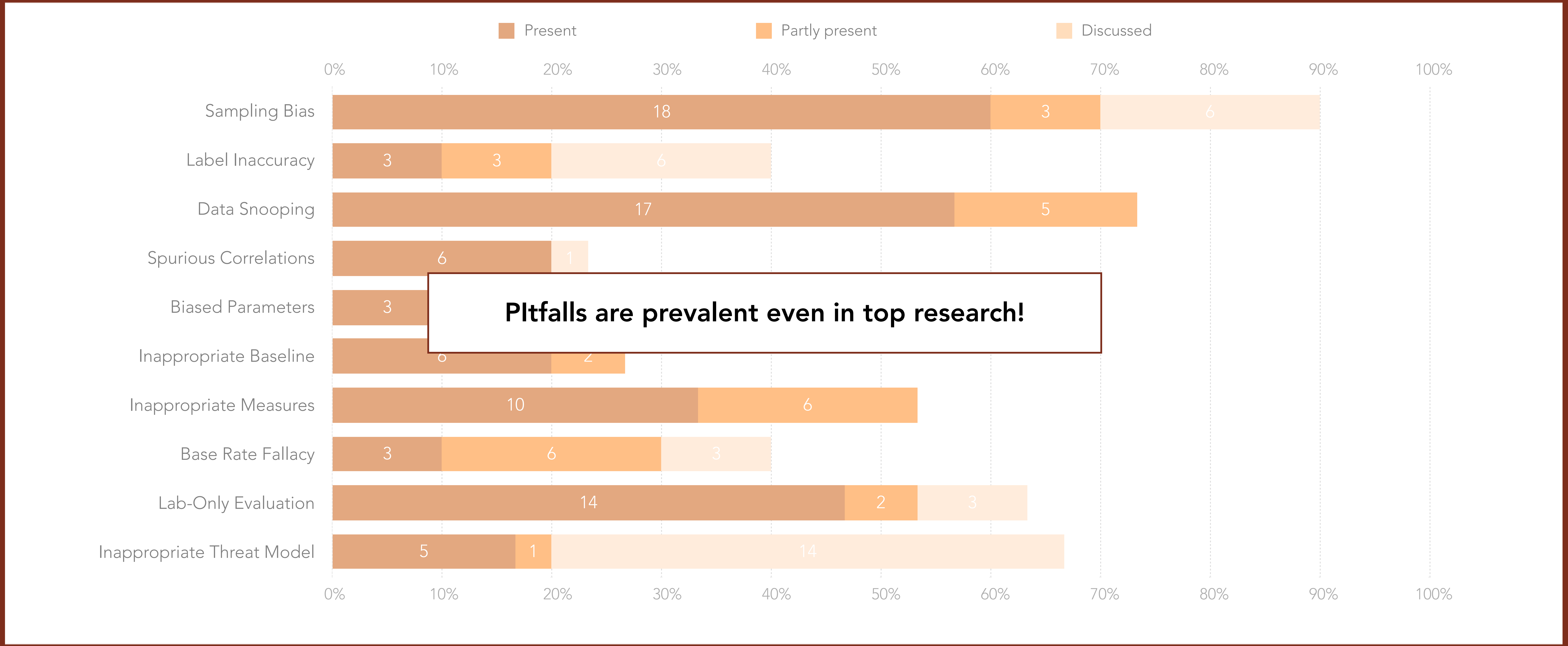
3. Authors Feedback



Prevalence Study



Prevalence Study



PItfalls are prevalent even in top research!

Impact Analysis

Android Malware Detection

P1: Sampling Bias

P4: Spurious Correlations

P7: Inappropriate Performance Measures

Authorship Attribution

P1: Sampling Bias

P4: Spurious Correlations

Vulnerability Discovery

P2: Label Inaccuracy

P4: Spurious Correlations

P6: Inappropriate Baselines

Network Intrusion Detection

P6: Inappropriate baselines

P9: Lab-only evaluation

Impact Analysis

Android Malware Detection

P1: Sampling Bias

P4: Spurious Correlations

P7: Inappropriate Performance Measures

Authorship Attribution

P1: Sampling Bias

P4: Spurious Correlations

Vulnerability Discovery

P2: Label Inaccuracy

P4: Spurious Correlations

P6: Inappropriate Baselines

Network Intrusion Detection

P6: Inappropriate baselines

P9: Lab-only evaluation

Impact Study: Mobile Malware Detection

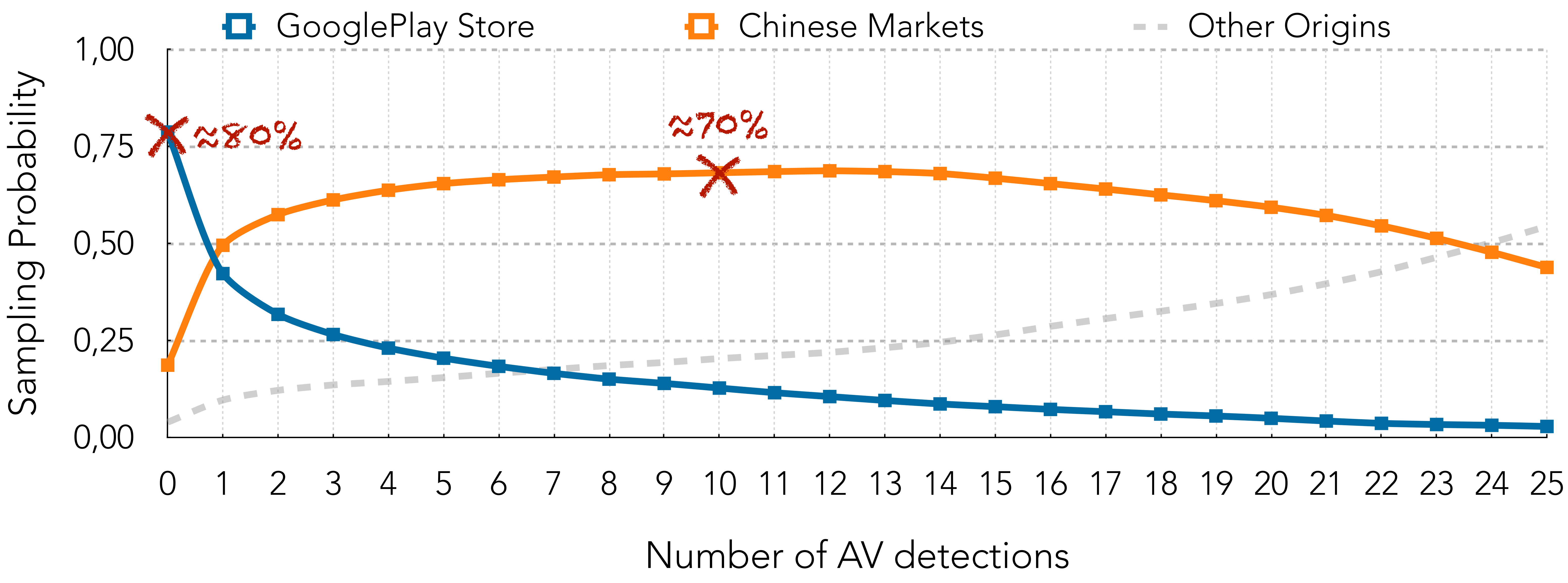
P1: Sampling Bias

P4: Spurious Correlations

P7: Inappropriate Performance Measures

What is the problem?

- Merging of data from different sources leads to sampling bias
- Different origins of malware and benign apps can introduce unwanted shortcuts



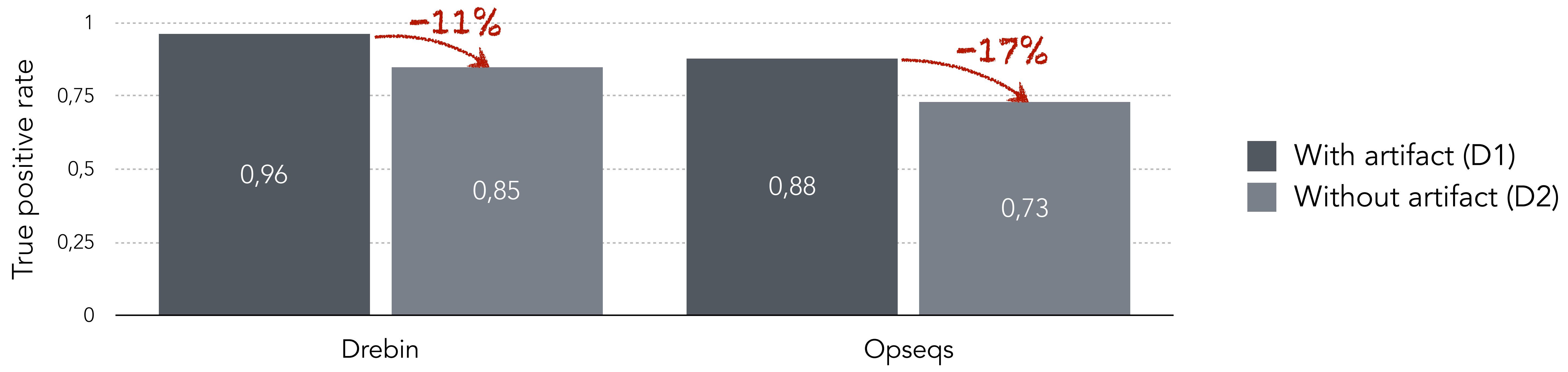
Allix et al. AndroZoo: collecting millions of Android apps for the research community. *ACM MSR, 2016.*
Arp et al. DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket. *NDSS, 2014.*

Impact Study: Mobile Malware Detection

P1: Sampling Bias
P4: Spurious Correlations
P7: Inappropriate Performance Measures

What is the impact?

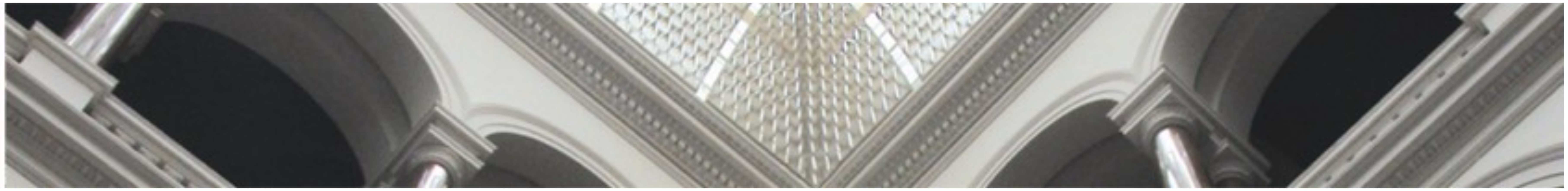
- Comparison on datasets *with* (D1) and *without* (D2) the artifact
- Training of SVM on two different feature sets



Results

- Experimental results show how sampling bias affects results (**P1**)
- The URL „*play.google.com*" is among top features in D1 (**P4**)
- Using Accuracy would have underestimated the presence of bias (**P7**)

Allix et al. AndroZoo: collecting millions of Android apps for the research community. *ACM MSR, 2016.*
Arp et al. DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket. *NDSS, 2014.*



Technische Universität Berlin

Dos and Don'ts of Machine Learning in Computer Security

We identify 10 subtle pitfalls affecting the field

Find that they are prevalent throughout top research

Demonstrate their impact through case studies

Updates on pitfalls and recommendations:

https://dodo-mlsec.org/

BIFOLD · UCL · King's College London · Technische Universität Braunschweig · Royal Holloway University of London · KIT Karlsruher Institut für Technologie · The Alan Turing Institute