SYSTEMS SECURITY RESEARCH LAB

KING'S College LONDON

ROYAL HOLLOWAY UNIVERSITY OF LONDON

CODE RI

Research Institute
Cyber Defence
Bundeswehr University Munich

TESSERACT

Eliminating Experimental Bias in Malware Classification across Space and Time

Feargus Pendlebury*, Fabio Pierazzi*, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro

https://s2lab.kcl.ac.uk/projects/tesseract

Fabio Pierazzi

<fabio.pierazzi@kcl.ac.uk> — @fbpierazzi
https://fabio.pierazzi.com

Aug 14-16, 2019
USENIX Security Symposium

* Equal contribution

ML for Malware Detection

Approach X (Novel)

- Feature Engineering
- ML Algorithm

Google Play

VirusShare Kharon

VIRUSTOTAL

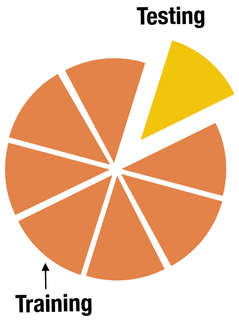
DREBIN MalGenome

10k gw

50k mw

Dataset Collection

Baseline Approaches



Three major unrealistic assumptions

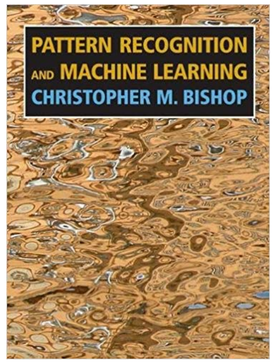
k-fold CV

Compare Results

Approach A

Approach B

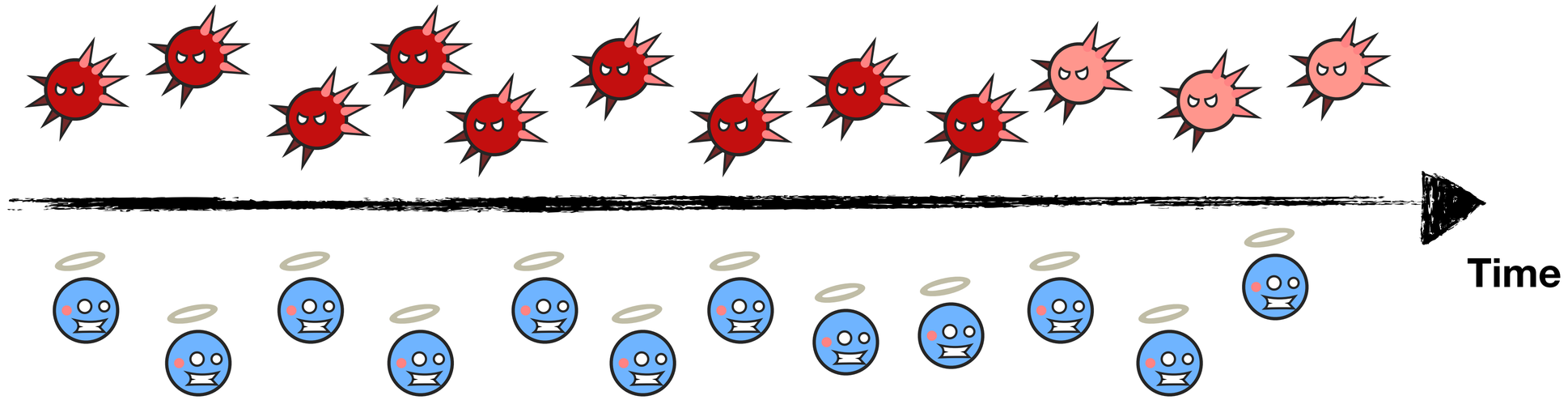
ML Best Practices



Approach	k-fold F1
X (New)	**0.99**
A	0.97
B	0.86

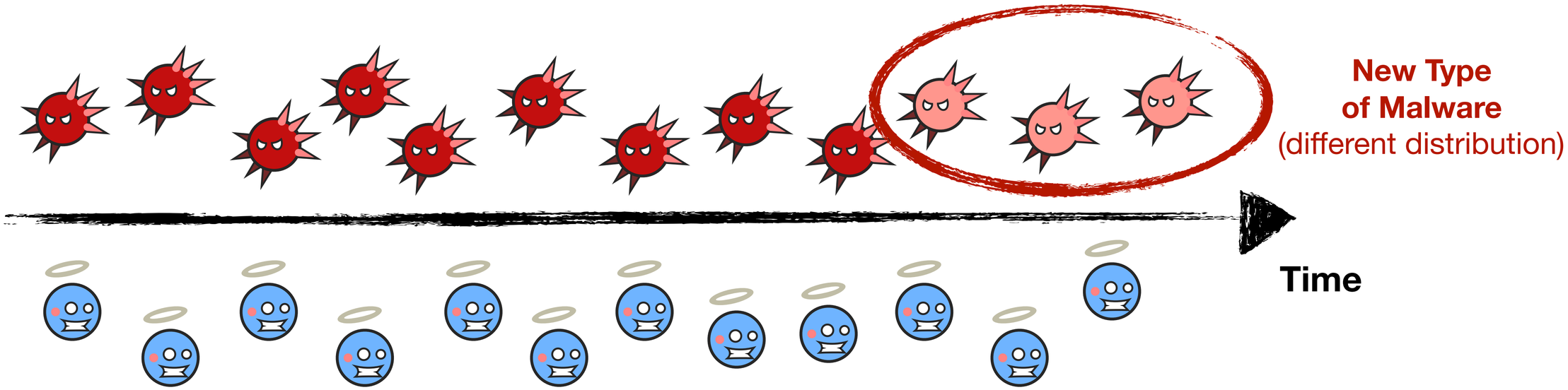
Sources of Experimental Bias (1/3)

Temporal Inconsistency in Train/Test Sets



Sources of Experimental Bias (1/3)

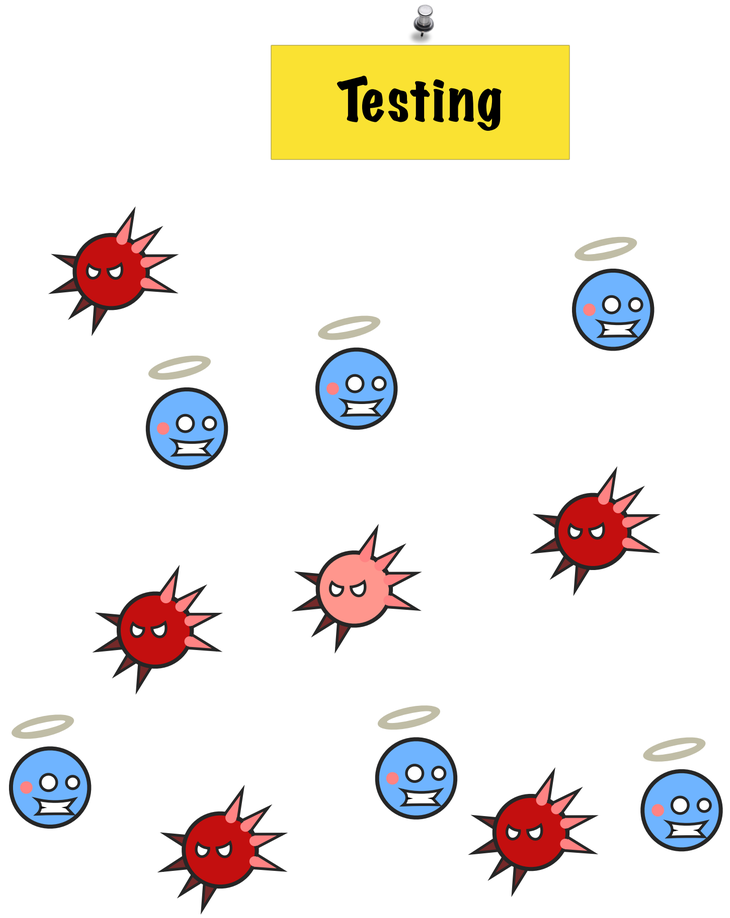
Temporal Inconsistency in Train/Test Sets



Sources of Experimental Bias (1/3)

Temporal Inconsistency in Train/Test Sets

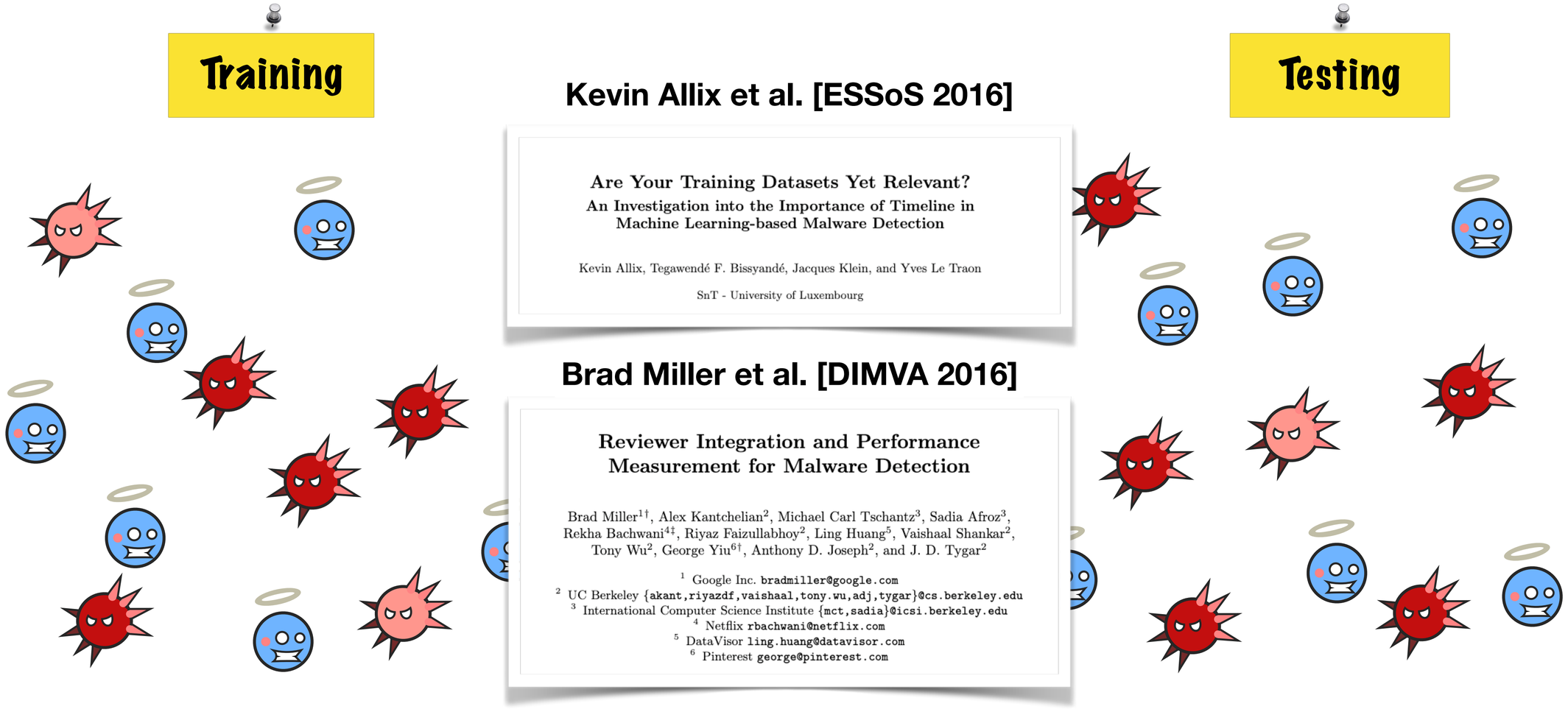
Violations use future knowledge in training



Sources of Experimental Bias (1/3)

Temporal Inconsistency in Train/Test Sets

Violations use future knowledge in training



Kevin Allix et al. [ESSoS 2016]

Are Your Training Datasets Yet Relevant?
An Investigation into the Importance of Timeline in Machine Learning-based Malware Detection

Kevin Allix, Tegawendé F. Bissyandé, Jacques Klein, and Yves Le Traon

SnT - University of Luxembourg

Brad Miller et al. [DIMVA 2016]

Reviewer Integration and Performance Measurement for Malware Detection

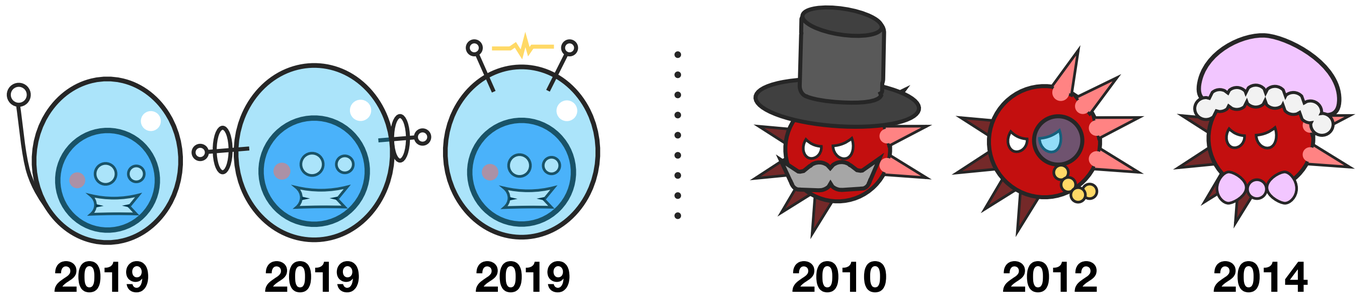
Brad Miller[1†], Alex Kantchelian[2], Michael Carl Tschantz[3], Sadia Afroz[3], Rekha Bachwani[4‡], Riyaz Faizullabhoy[2], Ling Huang[5], Vaishaal Shankar[2], Tony Wu[2], George Yiu[6†], Anthony D. Joseph[2], and J. D. Tygar[2]

[1] Google Inc. bradmiller@google.com
[2] UC Berkeley {akant,riyazdf,vaishaal,tony.wu,adj,tygar}@cs.berkeley.edu
[3] International Computer Science Institute {mct,sadia}@icsi.berkeley.edu
[4] Netflix rbachwani@netflix.com
[5] DataVisor ling.huang@datavisor.com
[6] Pinterest george@pinterest.com

Sources of Experimental Bias (2/3)

Temporal {good|mal}ware inconsistency

Violations may learn artifacts



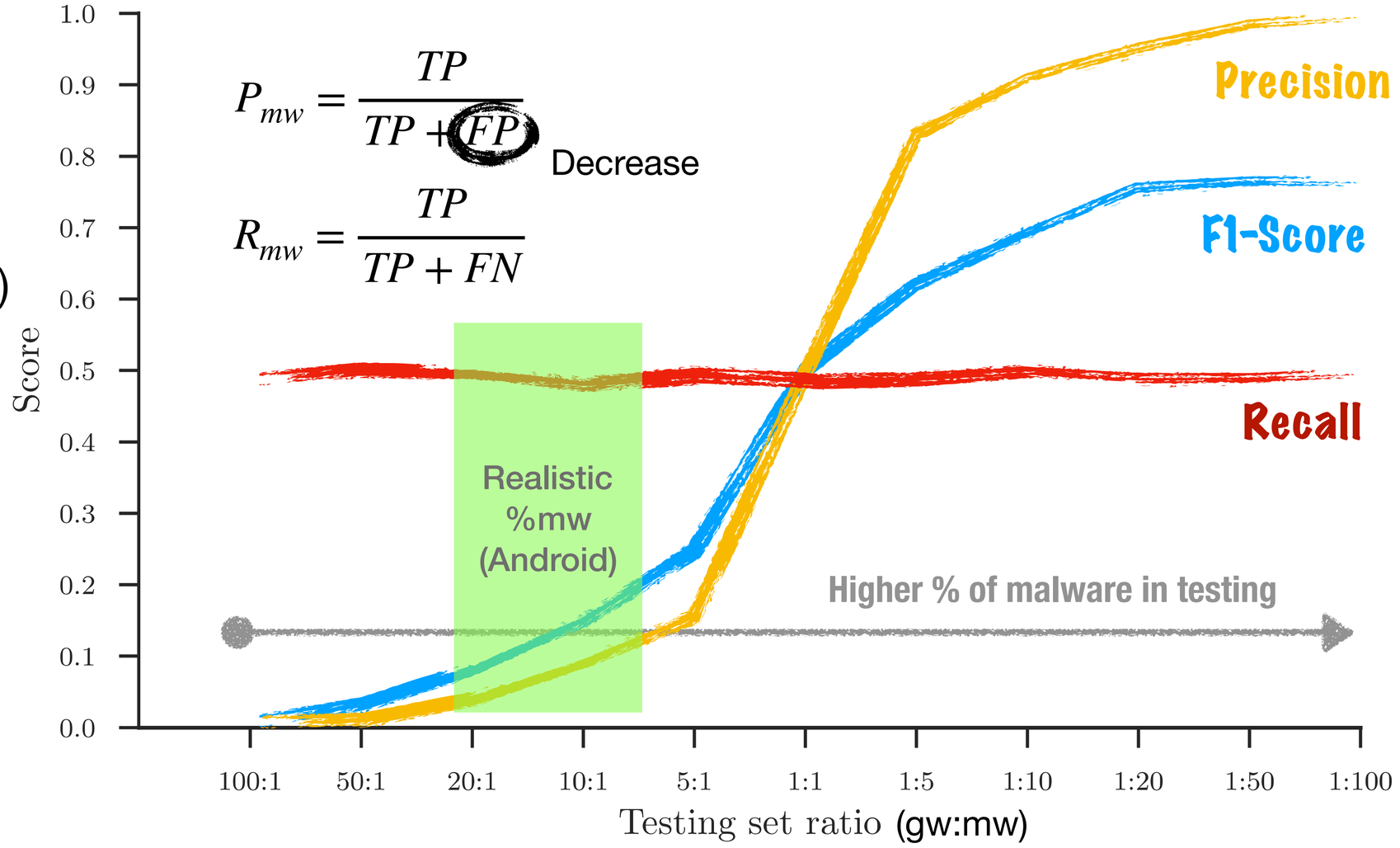

```
new_method()
```

Sources of Experimental Bias (3/3)

Unrealistic Test Class Ratio

Violations produce unrealistic results

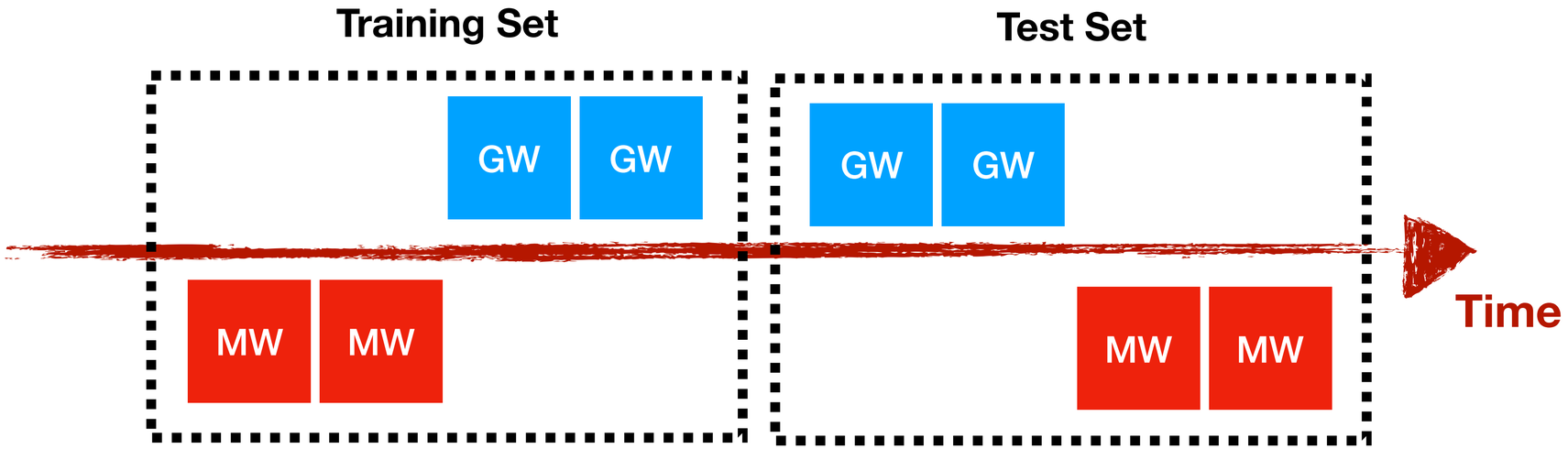
- **Training set**: Fixed
- **Testing set**: Varying % of mw (by downsampling gw)



TESSERACT Framework

Experimental Constraints

C1 Temporal training consistency → time(training) < time(testing)

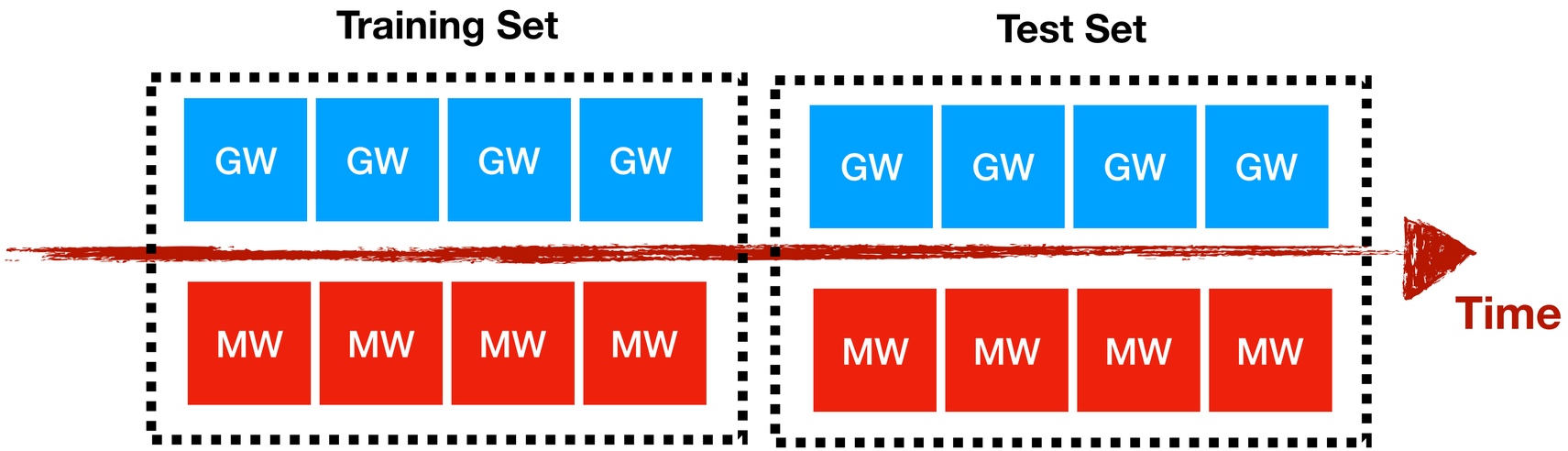


TESSERACT Framework

Experimental Constraints

C1 Temporal training consistency → time(training) < time(testing)

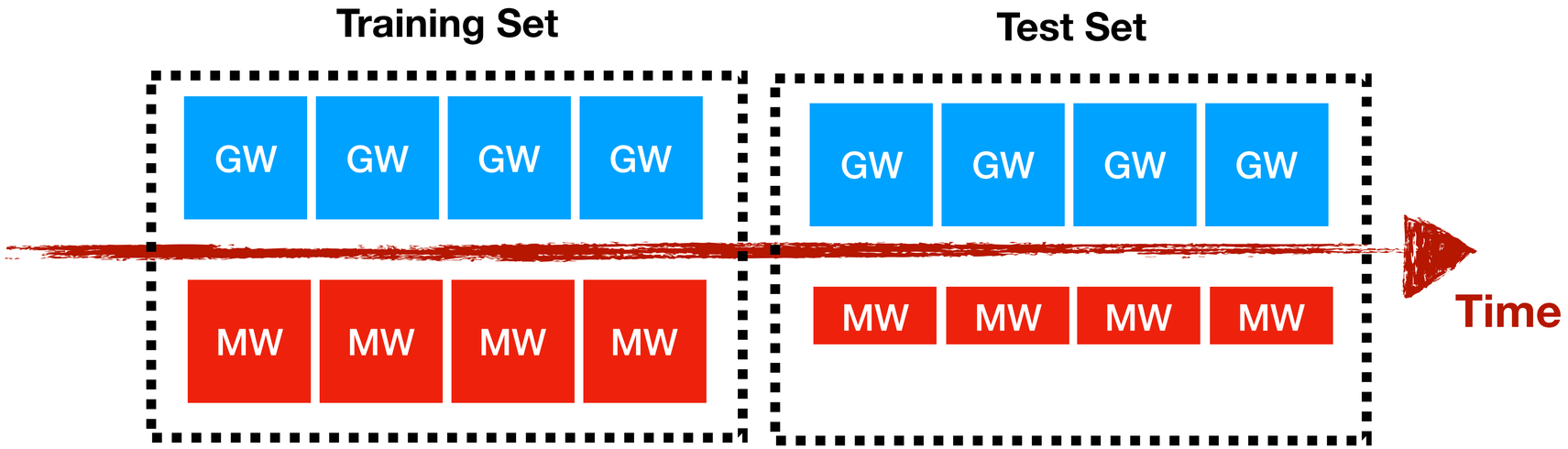
C2 {good|mal}ware temporal consistency → time(gw) = time(mw)



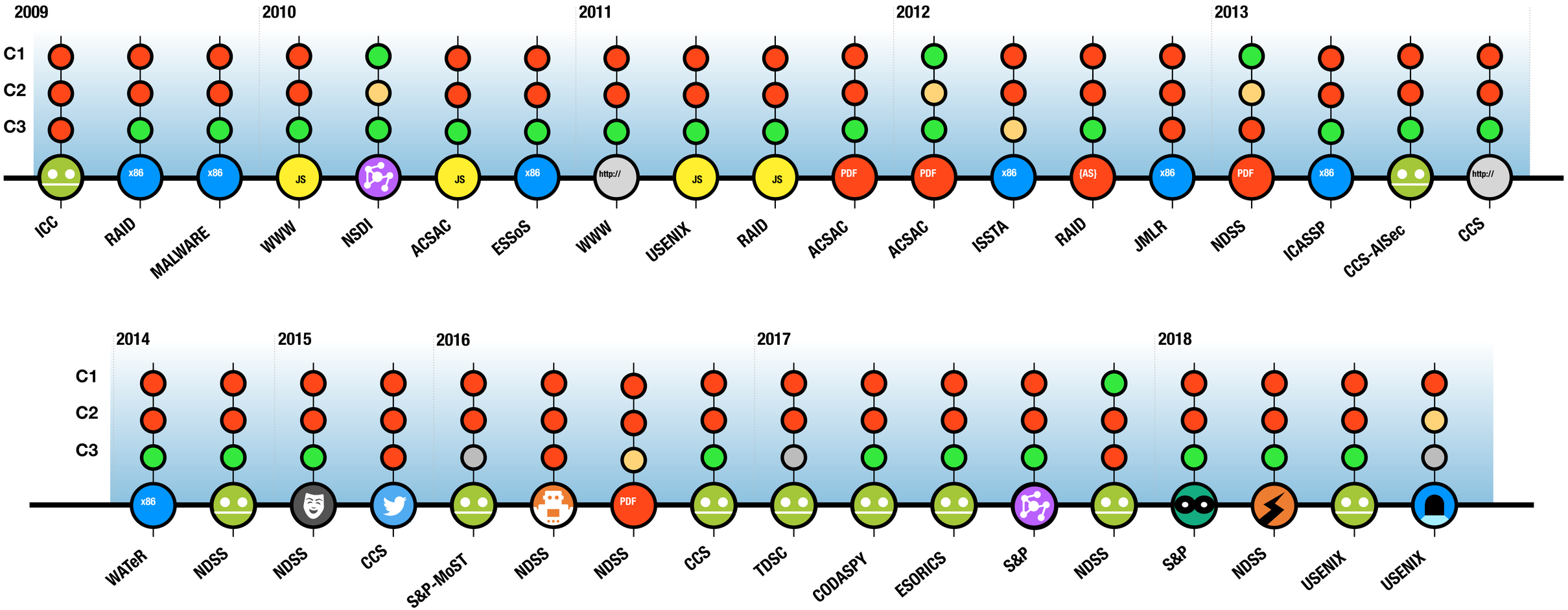
TESSERACT Framework

Experimental Constraints

C1	Temporal training consistency	→ time(training) < time(testing)
C2	{good|mal}ware temporal consistency	→ time(gw) = time(mw)
C3	Realistic testing classes ratio	→ realistic %mw in test

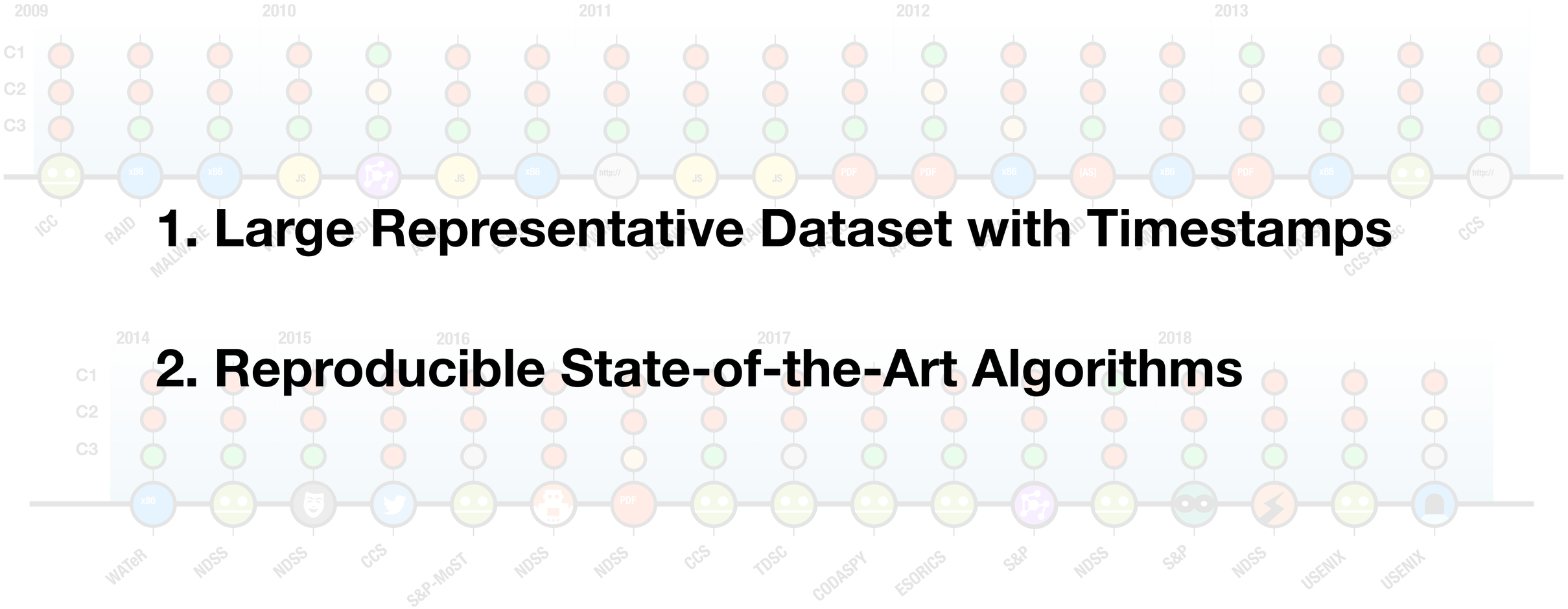


Endemic Problem



Details: https://s2lab.kcl.ac.uk/projects/tesseract/poster-references.pdf

Endemic Problem



1. Large Representative Dataset with Timestamps

2. Reproducible State-of-the-Art Algorithms

Details: https://s2lab.kcl.ac.uk/projects/tesseract/poster-references.pdf

Dataset

- **129,729** Android applications from **AndroZoo** [1]
- **10%** malware
- Covering **3 years** (2014 to 2016)

[1] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon.
AndroZoo: Collecting Millions of Android Apps for the Research Community
In *Mining Software Repositories (MSR)*, 2016

Benchmark Algorithms

Representative

Reproducible

Algorithm 1: DREBIN [NDSS14]

- Statically extracted features are bit vectors, present (1) or not (0)
- Linear Support Vector Machine (**SVM**) classifier

Algorithm 2: MaMaDroid [NDSS17]

- Markov chains from static call graphs through flow analysis
- Features are transition matrices of the Markov chains
- Random Forest (**RF**) classifier

Algorithm 3: Deep Learning [ESORICS17]

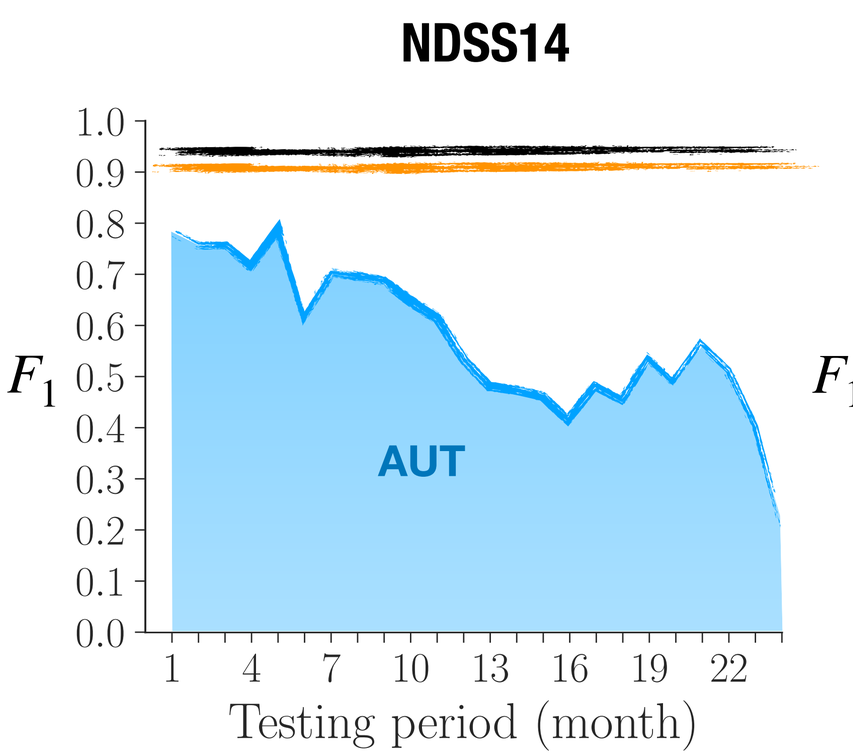
- DREBIN features and dataset
- **Deep** Feed-Forward **Neural Network**

[NDSS14] Arp et al., *DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket*.
[NDSS17] Mariconti et al., *MaMaDroid: Detecting Android Malware by Building Markov Chains of Behavioural Models*.
[ESORICS17] Grosse et al., *Adversarial Examples for Malware Detection*.

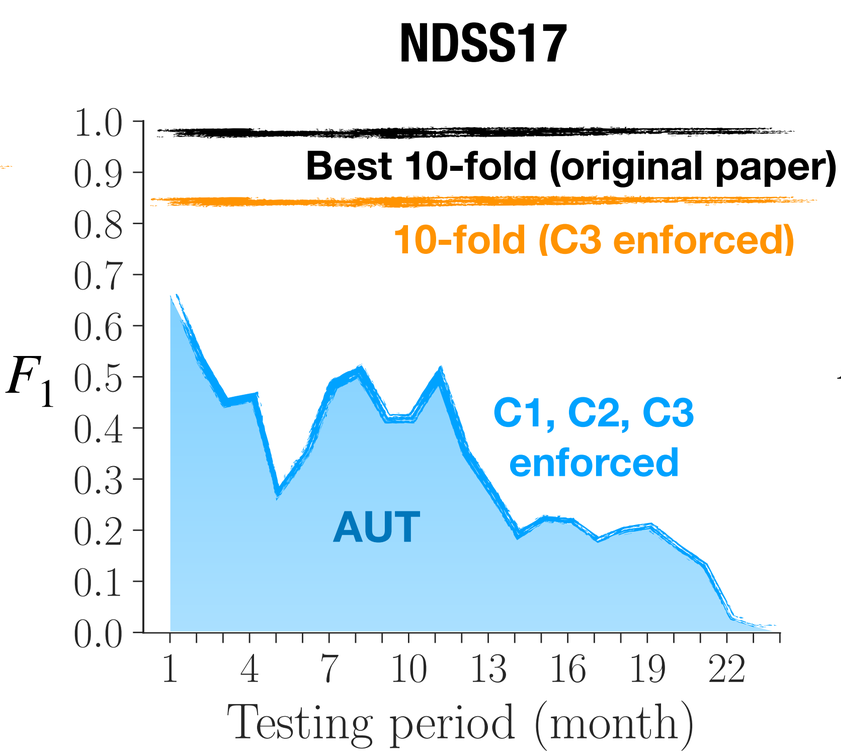
TESSERACT Evaluations

Experimental Constraints

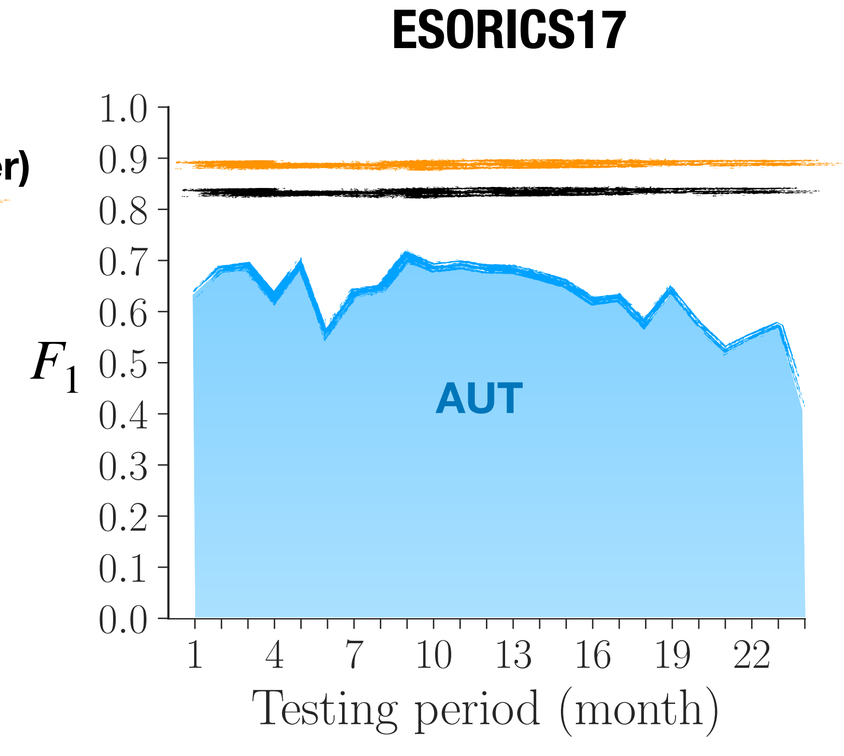
- **C1** Temporal training consistency
- **C2** {good|mal}ware temporal consistency
- **C3** Realistic testing classes ratio



$AUT(F_1, 24m) = 0.58$



$AUT(F_1, 24m) = 0.32$



$AUT(F_1, 24m) = 0.64$

TESSERACT: Actionable Points

Realistic Evaluations

- Unveils performance in realistic deployment
- Removes space-time experimental bias
- **Practitioners:** Choose Best Solution
- **Researchers**: Evaluate New Solutions

Tunable Parameters (e.g., time window, granularity)

Performance-Cost Trade Offs

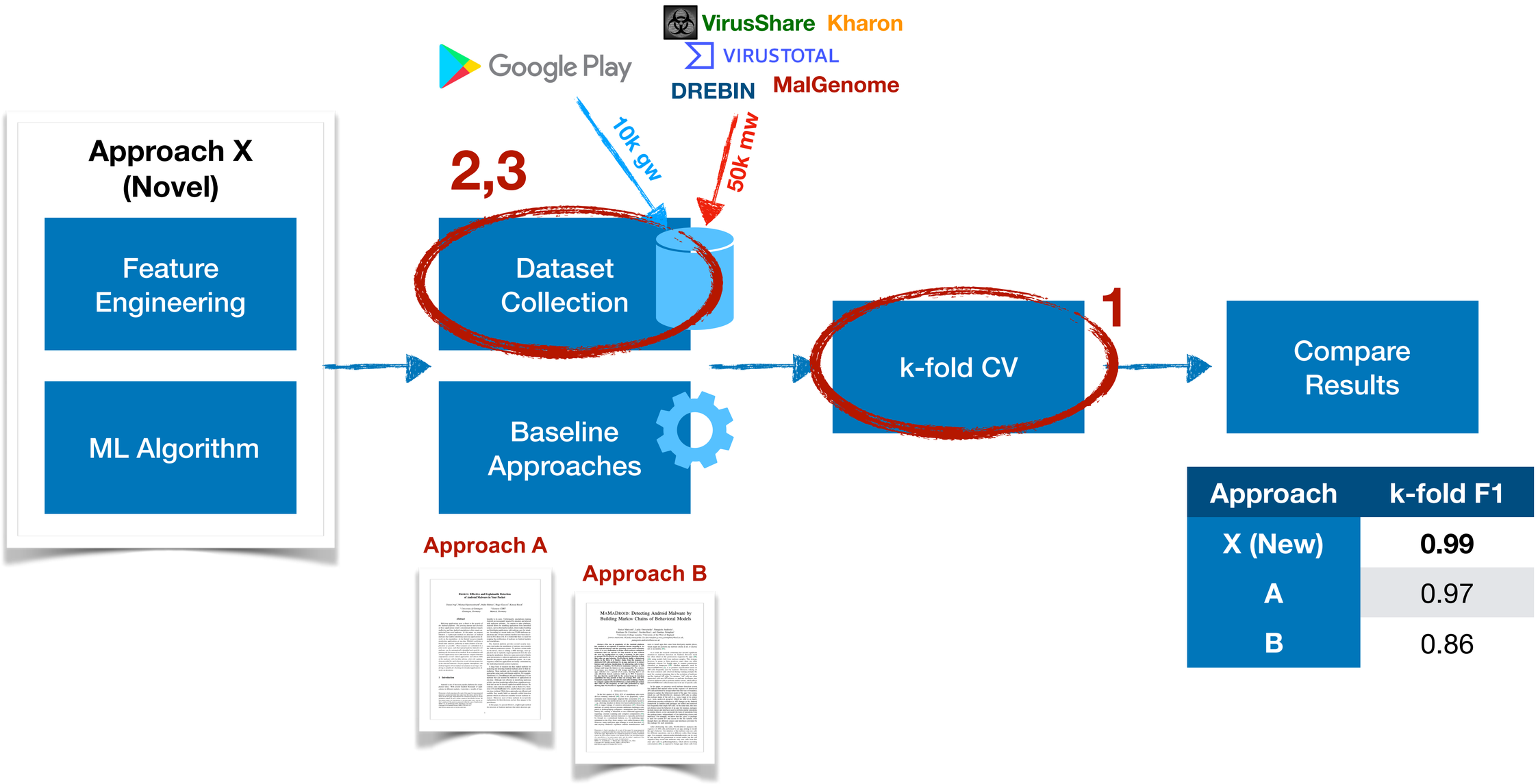
- **Detection Performance** (e.g., AUT F1)
- **Labeling Cost** for retraining (e.g., manpower)
- **Quarantine Cost** for rejection (e.g., low-confidence decisions)

More in Paper

Incremental Retraining Active Learning Rejection*

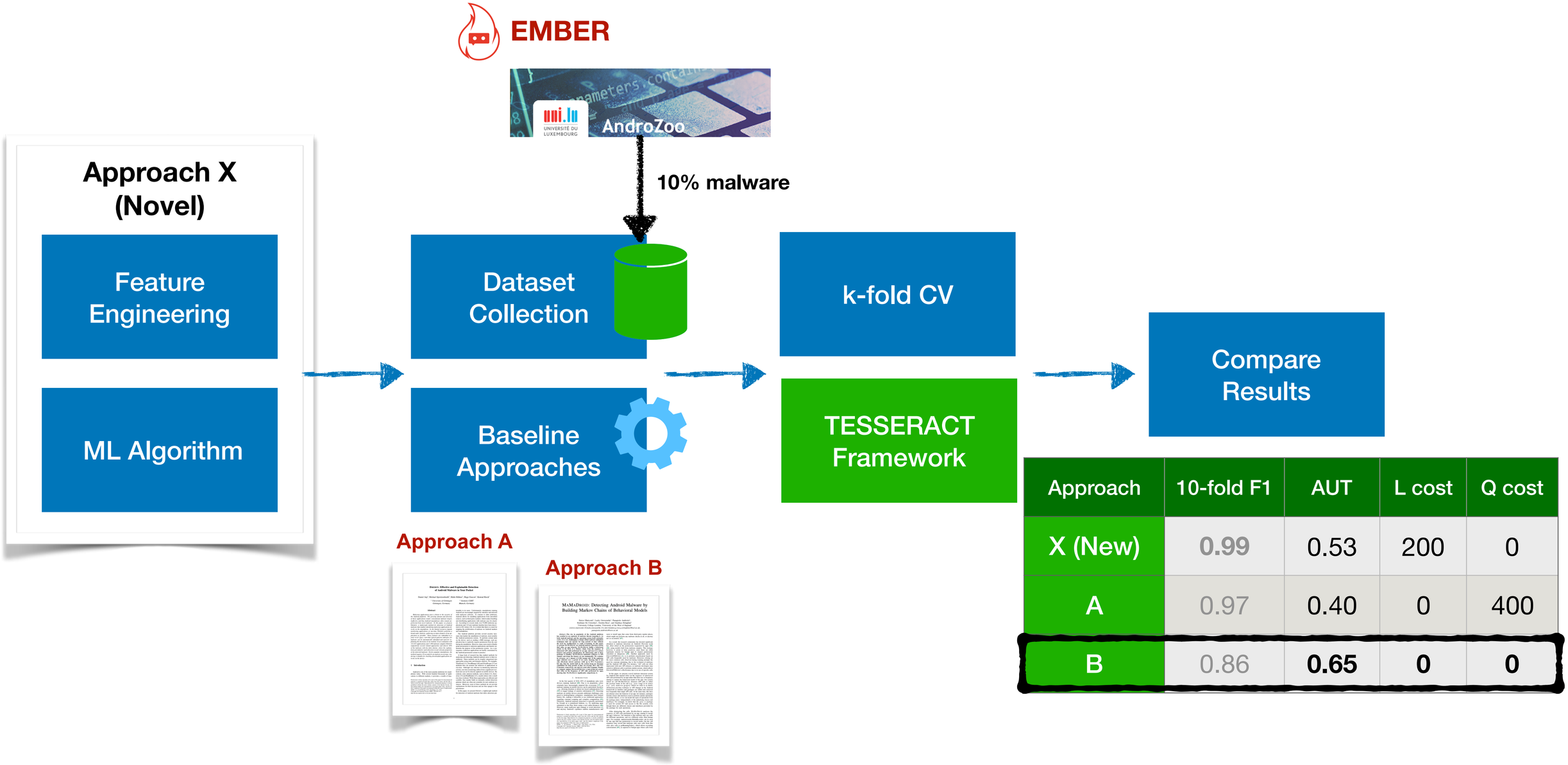
* R. Jordaney, K. Sharad, S. Dash, Z. Wang, D. Papini, I. Nouretdinov, and L. Cavallaro. ***Transcend: Detecting Concept Drift in Malware Classification Models.*** In USENIX Security Symposium, 2017

ML for Malware Detection



Approach	k-fold F1
X (New)	**0.99**
A	0.97
B	0.86

ML for Malware Detection



Approach	10-fold F1	AUT	L cost	Q cost
X (New)	0.99	0.53	200	0
A	0.97	0.40	0	400
B	0.86	**0.65**	**0**	**0**

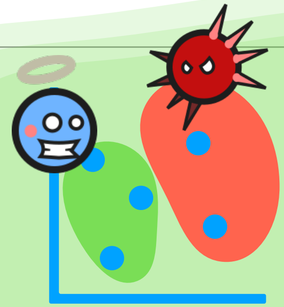
TESSERACT open-source library

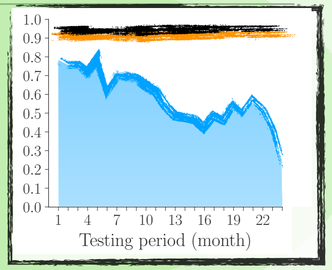
https://s2lab.kcl.ac.uk/projects/tesseract/

Conclusions

Experimental Constraints

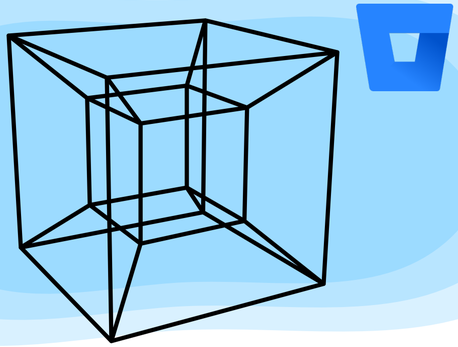
C1	Temporal training consistency
C2	{good|mal}ware temporal consistency
C3	Realistic testing classes ratio





- (Android) Malware Detection is still an **open problem**
- We propose TESSERACT framework
  - **Sound time-aware evaluations** to compare different solutions
  - **Performance-cost trade-offs** for retraining/rejection strategies
- **Open-source** code, dataset, features

  https://s2lab.kcl.ac.uk/projects/tesseract/



SYSTEMS SECURITY RESEARCH LAB

KING'S College LONDON

RI CODE

ROYAL HOLLOWAY UNIVERSITY OF LONDON