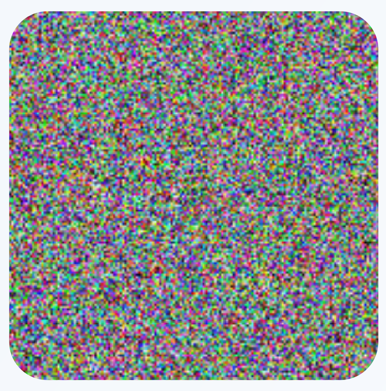
EXPLORING ADVERSARIAL EXAMPLES IN MALWARE DETECTION

Octavian Suciu, Scott E. Coull, Jeffrey Johns

PROBLEM

Adversarial Examples in Image Classification:

 +  = 

"panda" perturbation "gibbon"

Adversarial Examples in Malware Detection

\x43\x04\xF1\x24 + [perturbation] = \x8A\x32\xB1\x2C

"malware" perturbation **corrupted file**

Are adversarial malware examples realistic?
Are attacks effective against production-scale training sets?

ATTACKS

Benign Append

\x43\x00\xF1\x24\x00\x3D + \x43\x00\xF1\x24\x00\x3D \xE3\x41

Malware sample — append section of a benign program

FGM Append

\x43\x00\xF1\x24\x00\x3D + \x43\x00\xF1\x24\x00\x3D [noise]

Malware sample — append gradient-based noise via single-step FGM[1]

Slack FGM

\x43\x00\xF1\x24\x00\x3D + \x43 [noise] \xF1\x24 [noise] \x3D

Malware sample — add single-step FGM noise via safe slack region

PE Header	Section 1	Section 2	Slack	Section 3	Slack	...

Slack = compiler-generated misalignment of physical and virtual addresses

EXPERIMENTAL SETUP

Victim Model: MalConv[2]

Sample → 2MB Embedding → Gated Convolution → Temporal Max-Pooling → Fully Connected → Label

Architecture: pooling 128 non-overlapping convolutional kernels
- ≤ 128 unfragmented input sequences used in classification

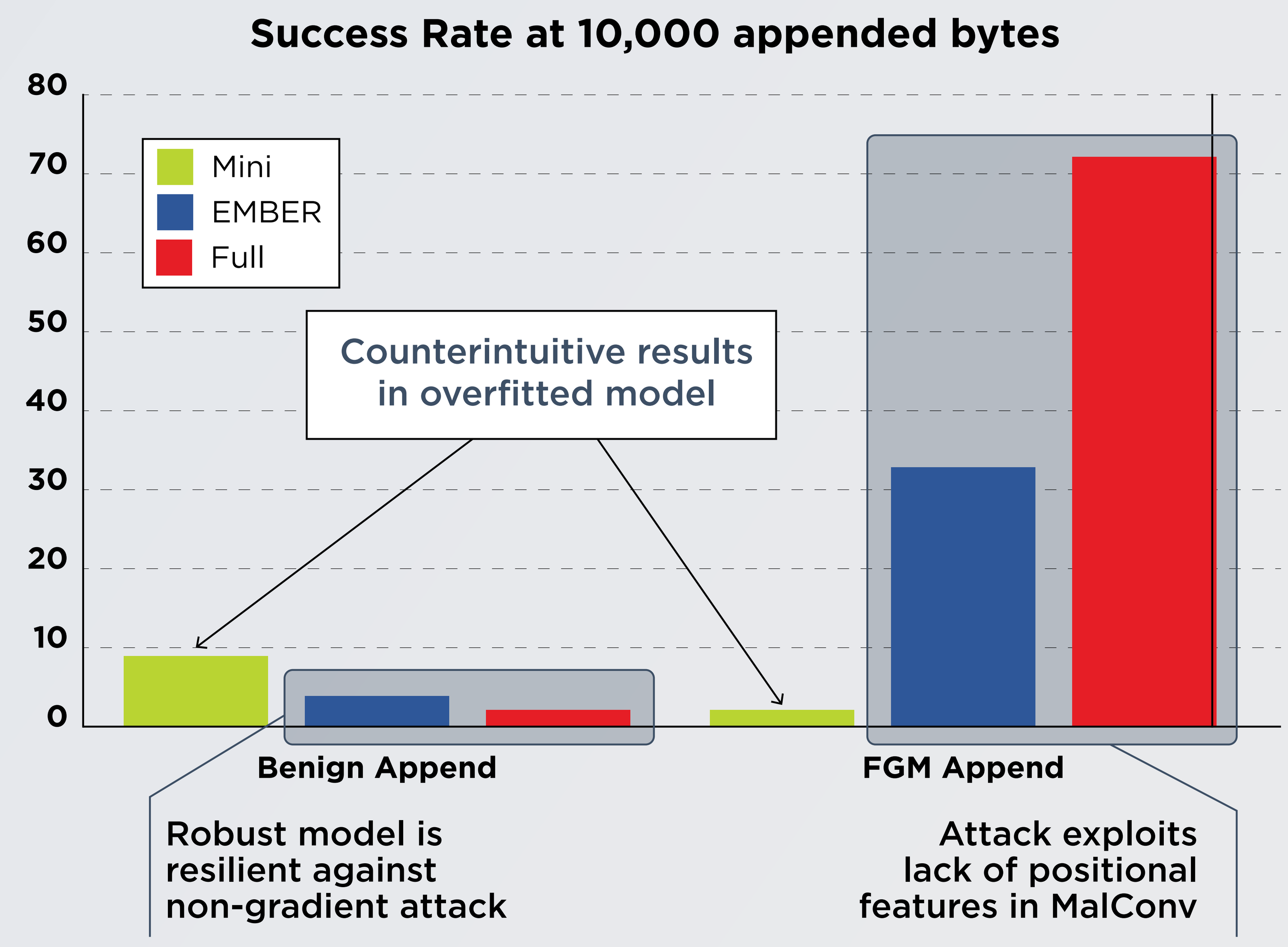
Training Sets:

- **Mini:** in line with prior work[3], 8,500 samples
- **EMBER:** publicly available corpus of 1.1M samples[4]
- **Full:** production scale dataset of 12.5M samples

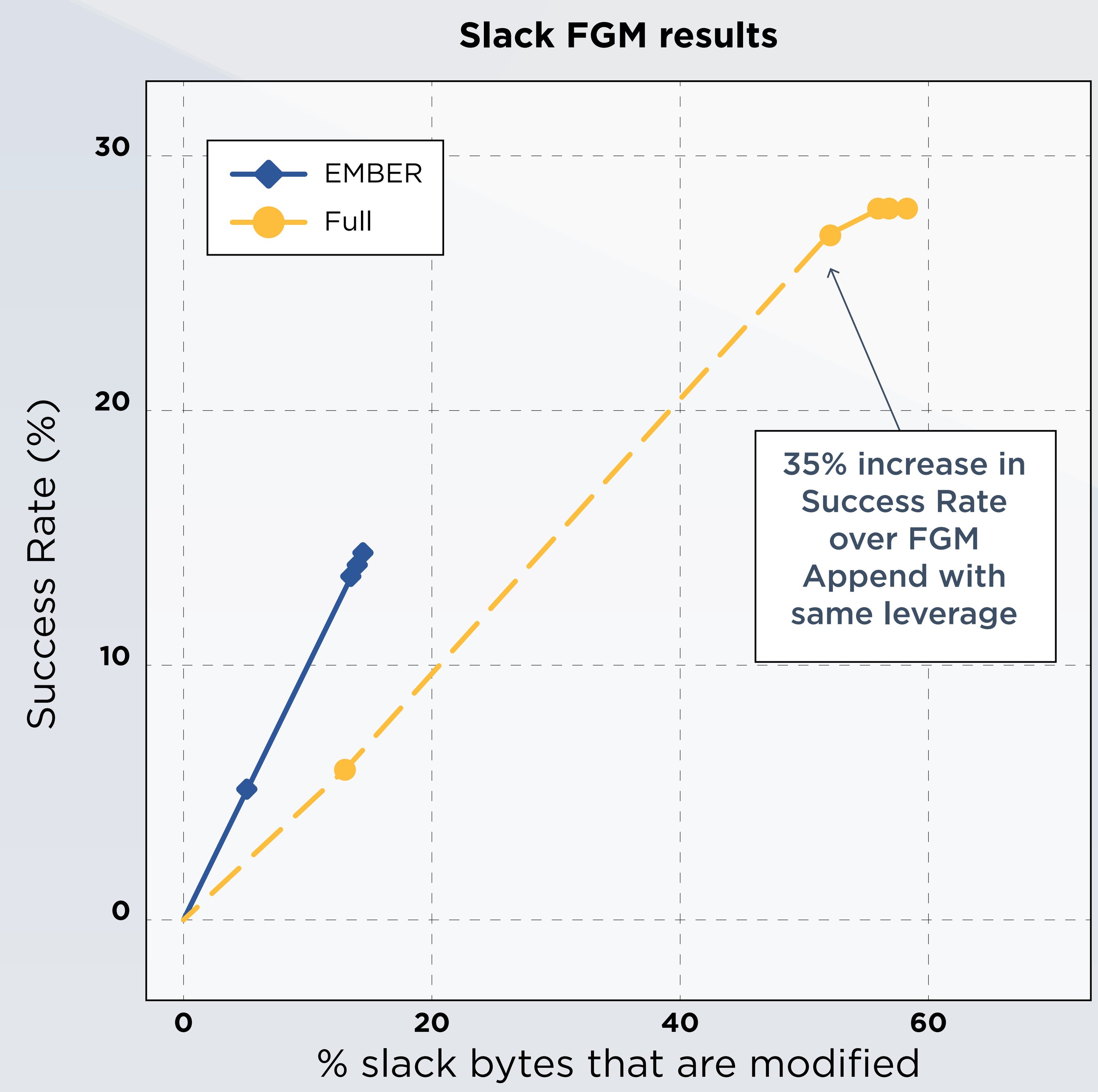
1 Explaining and harnessing adversarial examples [Goodfellow+ 2014]
2 Malware detection by eating a whole exe[Raff+ 2017],
3 Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables [Kolosnjaji+ 2018]
4 EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models [Anderson+ 2018]

FINDINGS

Model Robustness Influences Results



MalConv Contains Architectural Weaknesses



Unfragmented input flows to last layer
- effect of Slack bytes is amplified by context

Trade-off between Success Rate and Leverage
- due to Slack size and gradient magnitude

Single-Step Samples are Not Transferable

Transfer samples between EMBER ⇌ Full
- using FGM Append & Slack FGM

Only 3/400 attack samples are successfully transferred
- small gradient magnitude in EMBER

Thursday 10:45AM @ DLS Workshop