Black Hat USA 2018

MLPdf: An Effective Machine Learning Based Approach for PDF Malware Detection

Dr Jason Zhang
Senior Threat Researcher

SOPHOS

Agenda

- PDF File Format and Associated Attacks
- MLPdf Model
- Data Labelling
- Feature Engineering
- Underfitting & Overfitting
- Demo
- Conclusion
- QA

PDF File Format & Associated Attacks

PDF File Format

```
%PDF-1.1

1 0 obj
<<
 /Type /Catalog
 /Outlines 2 0 R
 /Pages 3 0 R
>>
endobj

..

6 0 obj
<<
 /Type /Font
 /Subtype /Type1
 /Name /F1
 /BaseFont /Helvetica
 /Encoding /MacRomanEncoding
>>
endobj

5 0 obj
<< /Length 53 >>
stream
BT /F1 24 Tf 100 700 Td (Hello Hacktivity 2015) Tj ET
endstream
endobj

xref
0 7
0000000000 65535 f
0000000012 00000 n
0000000089 00000 n
0000000145 00000 n
0000000214 00000 n
0000000543 00000 n
0000000419 00000 n

trailer
<<
 /Size 7
 /Root 1 0 R
>>
startxref
654
%%EOF
```

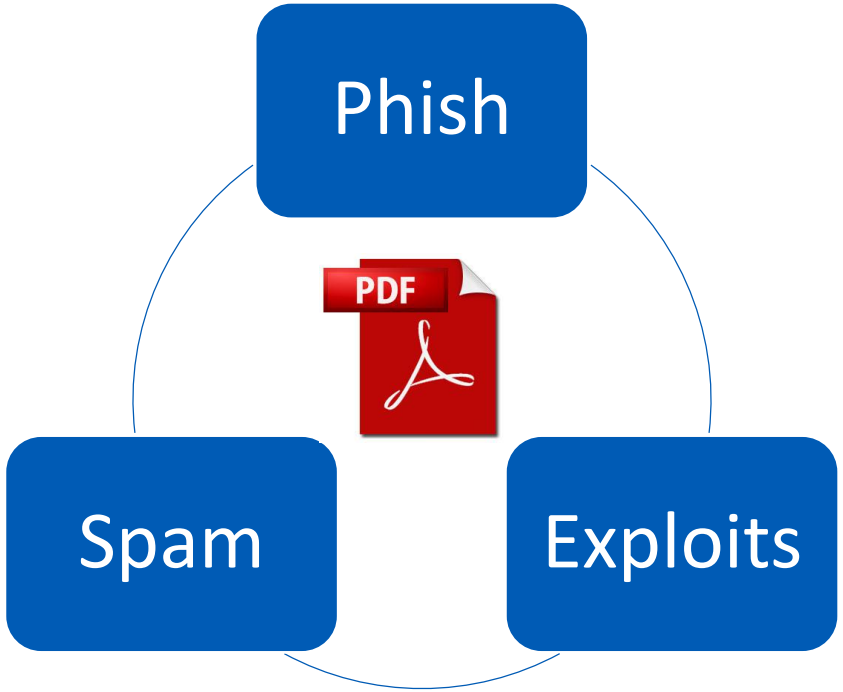
HEADER

BODY

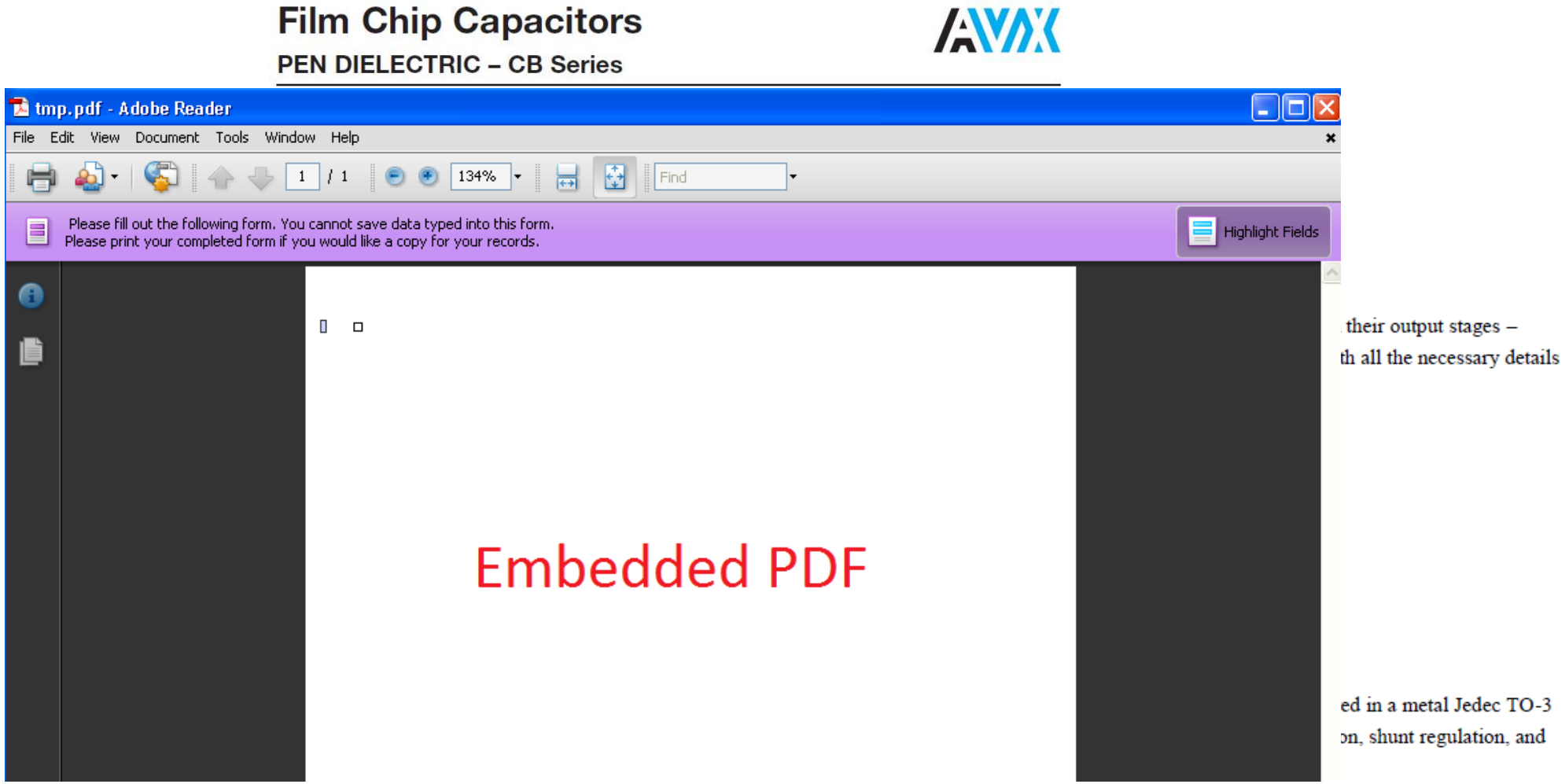
CROSS REFERENCE (XREF) TABLE

TRAILER

PDF Based Attacks

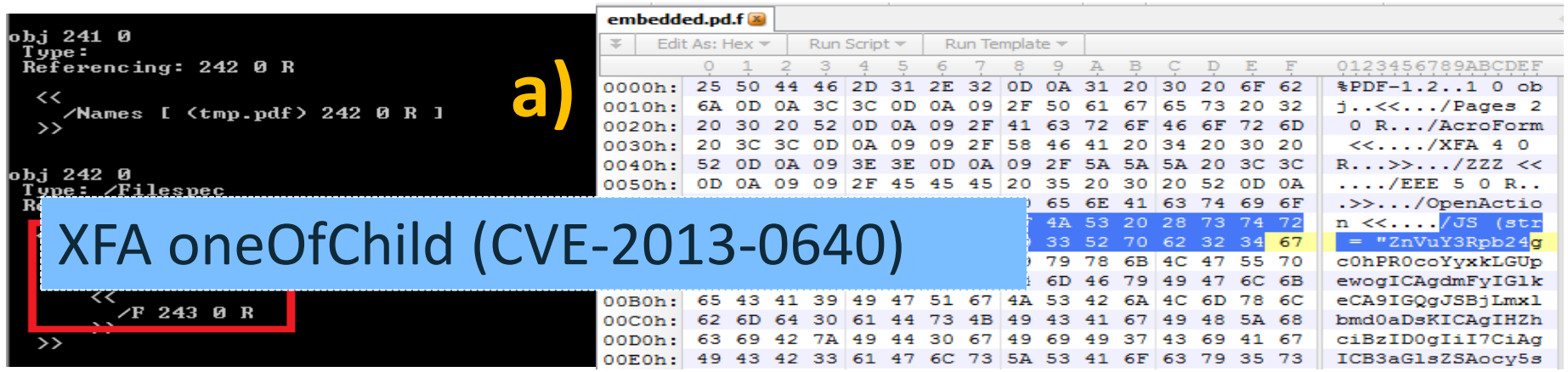


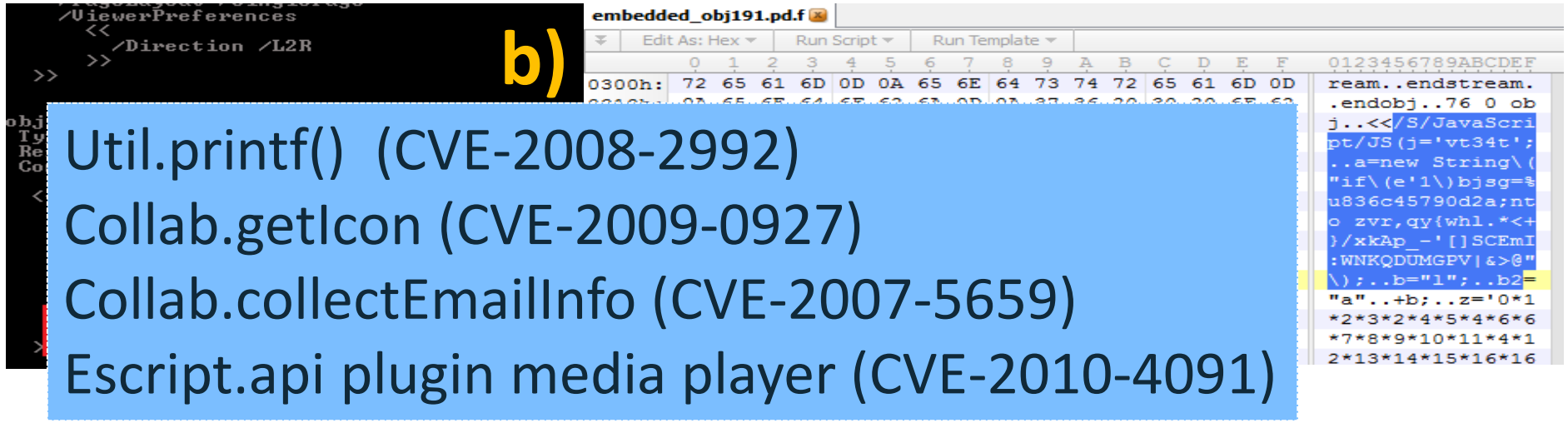
PDF Based Attacks: Exploits



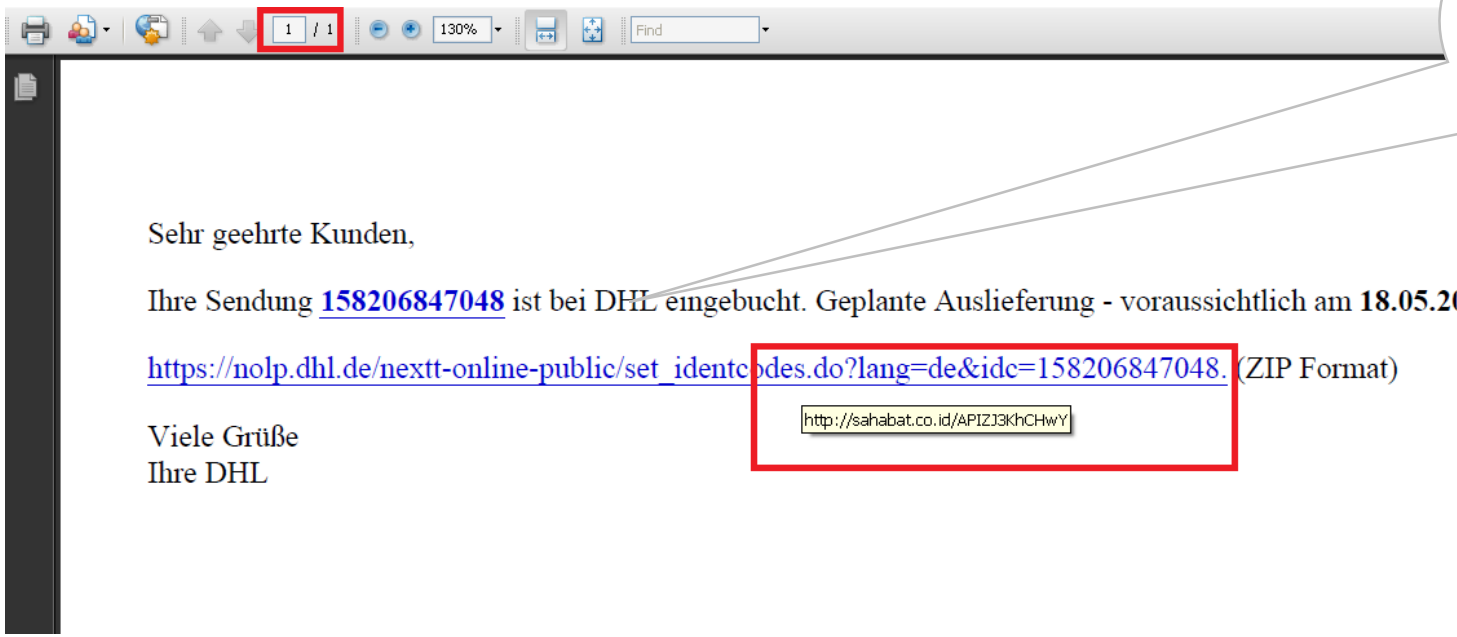
%PROFILE%\Local Settings\Temp\A9RDC52.tmp\tmp.pdf

PDF Based Attacks: Exploits (cont.)





PDF Based Attacks: Phish



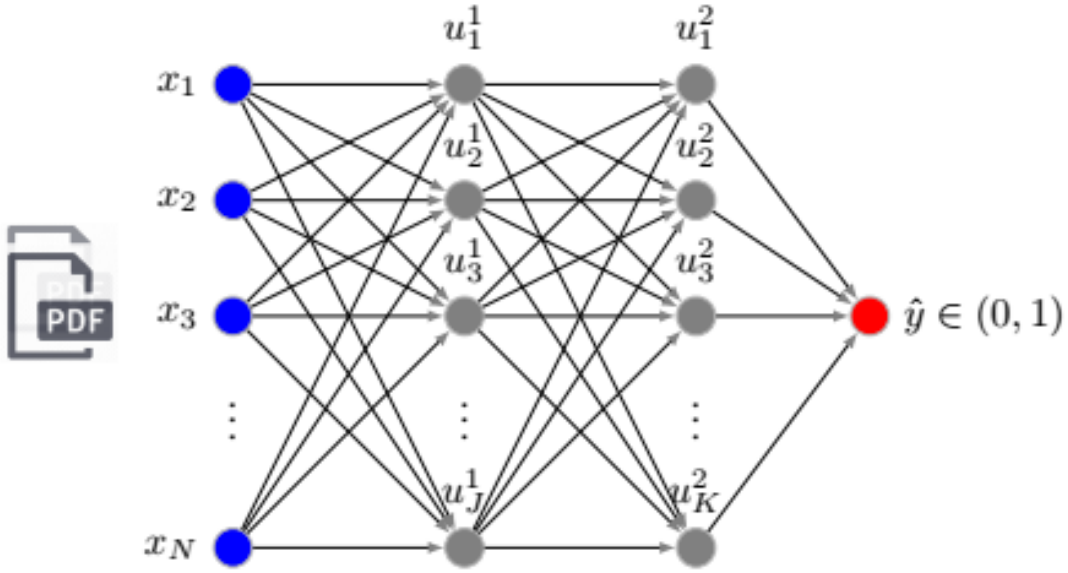
Targeting logistic companies

PDF Based Attacks: Spam

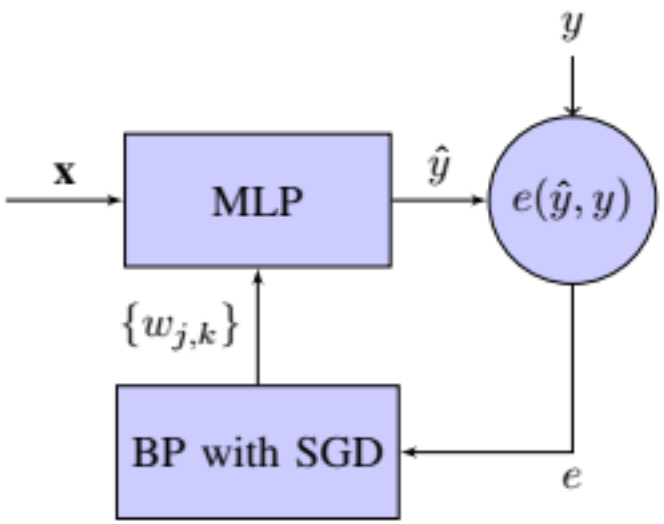
Punycode based spam

MLPdf Model

MLPdf Model



MLP neural network model



$$\Delta w_{j_{i-1},k_i} = -\eta \frac{\partial e(\hat{y}, y)}{\partial w_{j_{i-1},k_i}}$$

BP based MLP weights update via SGD

8857 parameters (weights + bias terms) to be updated

Data Labelling

Data Labelling

Know your **enemy**

It is critical to accurately label **clean** and **malicious** files to train ML algorithms.

Data Labelling

A file is labelled as **MALWARE** if detected by

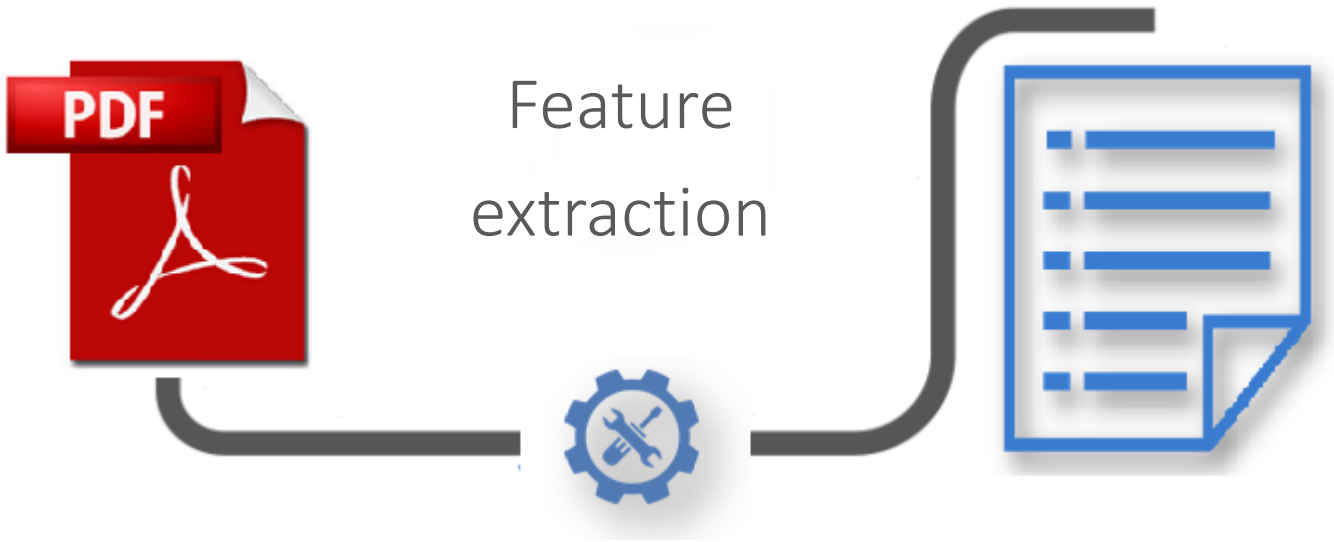
- >=3 reliable AV scanners, OR
- >=1 trustable SAV identities

A file is labelled as **CLEAN** if

- not detected by any scanner, OR
- confirmed as FP

Feature Engineering

Feature Engineering



PART OF THE EXTRACTED FEATURES

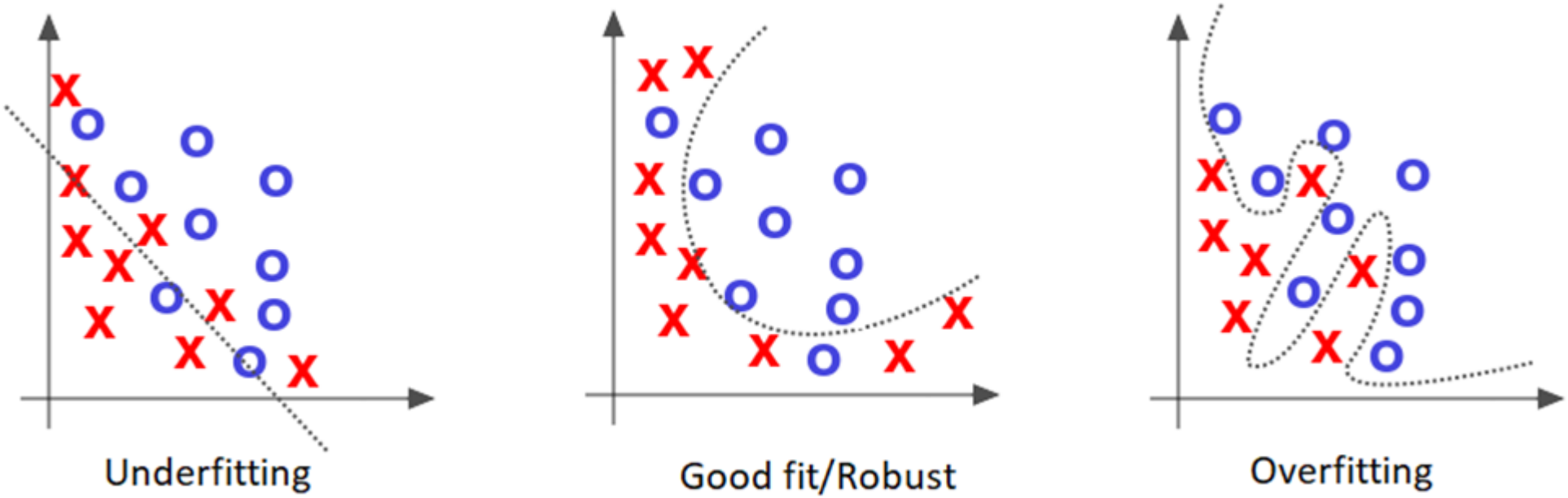
Feature name	Description
F_SIZE	*PDF file size*
F_JS	*PDF with JavaScript or not*
F_PGC	*Page count*
F_OBJC	*Number of objects*
F_FILT	*Stream filtering*
F_ENTRP1	*Entropy of some content*
F_ENTRP2	*Entropy of some content*
...	...

In-house & 3rd party tools to extract **48** features

Underfitting & Overfitting

Underfitting & Overfitting

- Make the training error small
- Make the gap between training and test error small



a) Validation b) Batch normalization c) Dropout

Demo

Demo

- *Demo-1*: clean (13101) & malware (1946) files

Source: Sophos filesDB & URL logs prior to March 31, 2018

- *Demo-2*: clean files (35599)

Source: Sophos URL logs for May 9-24, 2018

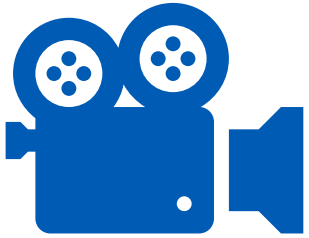
- *Demo-3*: a case study with a zero-day attack

Source: Sophos filesDB - July 2018

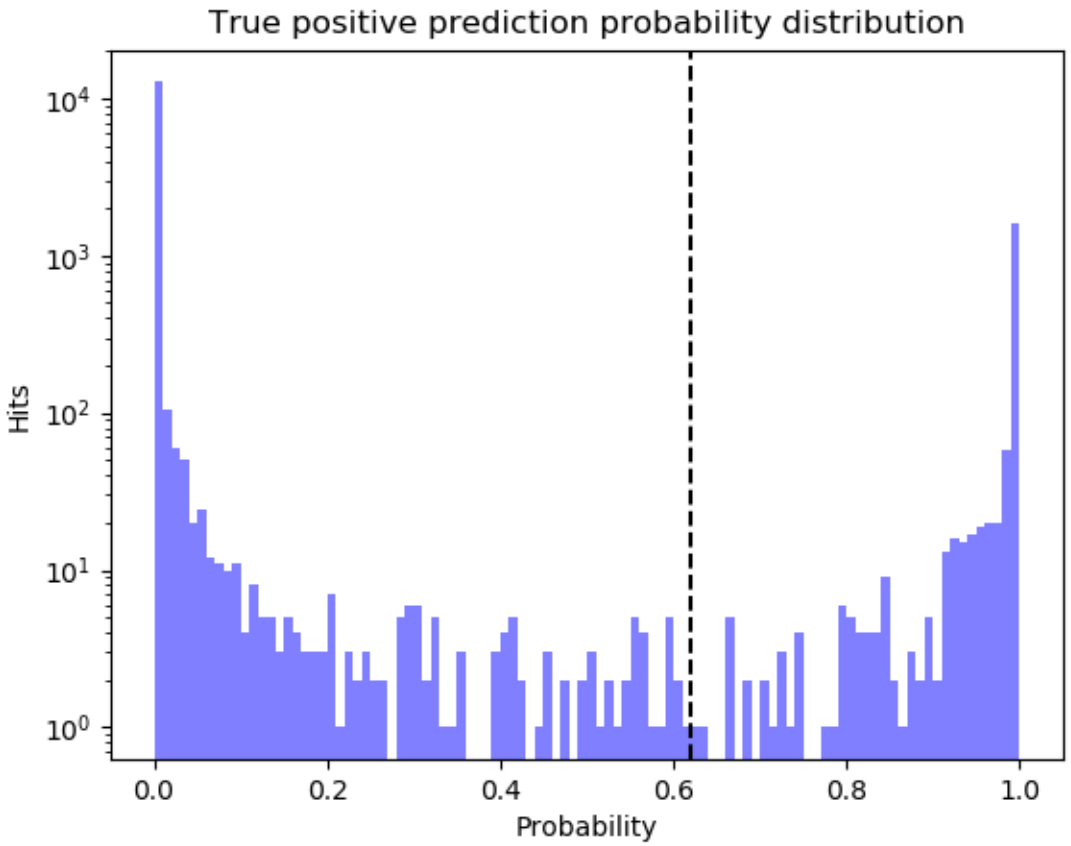
File size
< **15MB**

Demo: Dataset-1 – Clean (13101) & Malware (1946) Files

Source: Sophos filesDB & wild collections prior to March 31, 2018

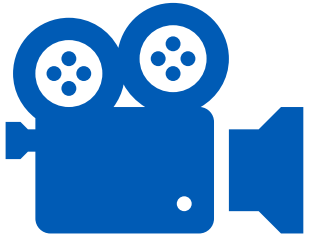


	FPR	TPR
MLPdf	0.12%	95.38%

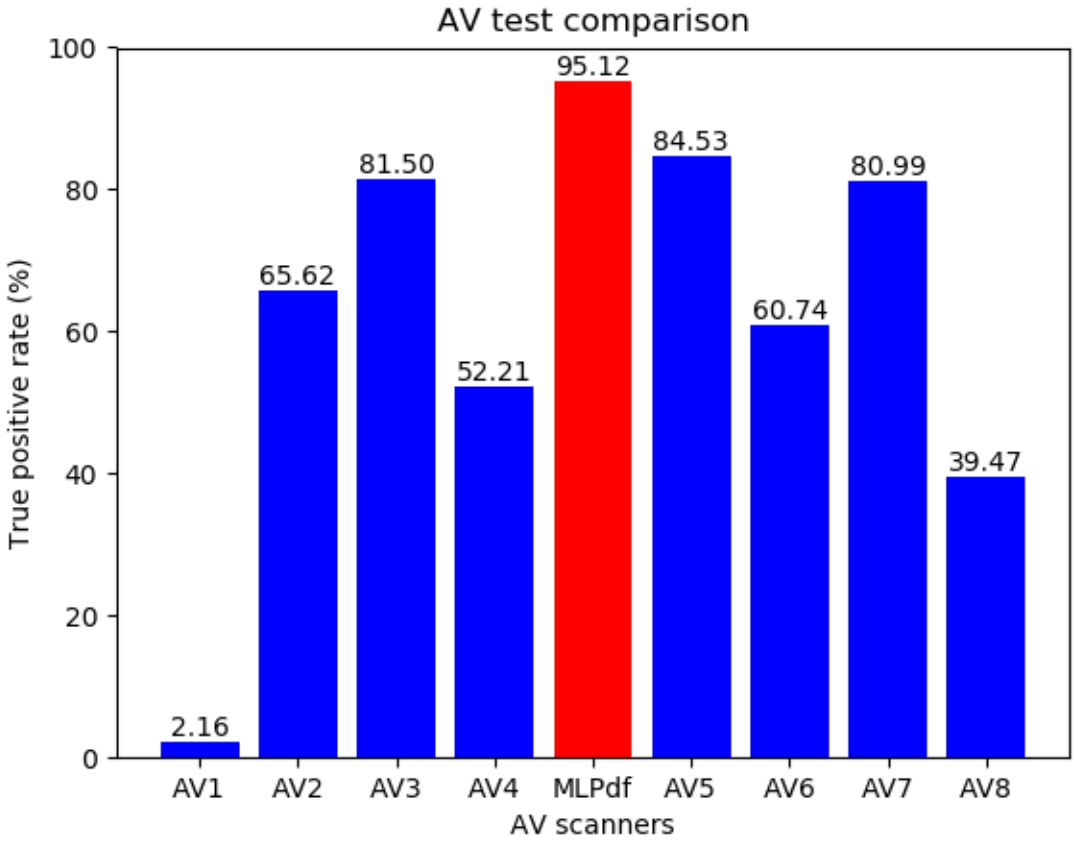


Demo: Dataset-1 (cont.) – Compared with 8 AV Scanners

Source: Sophos filesDB & wild collections prior to March 31, 2018

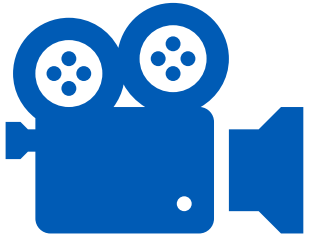


	FPR	TPR
MLPdf	0.12% (0.08%)	95.38% (95.12%)

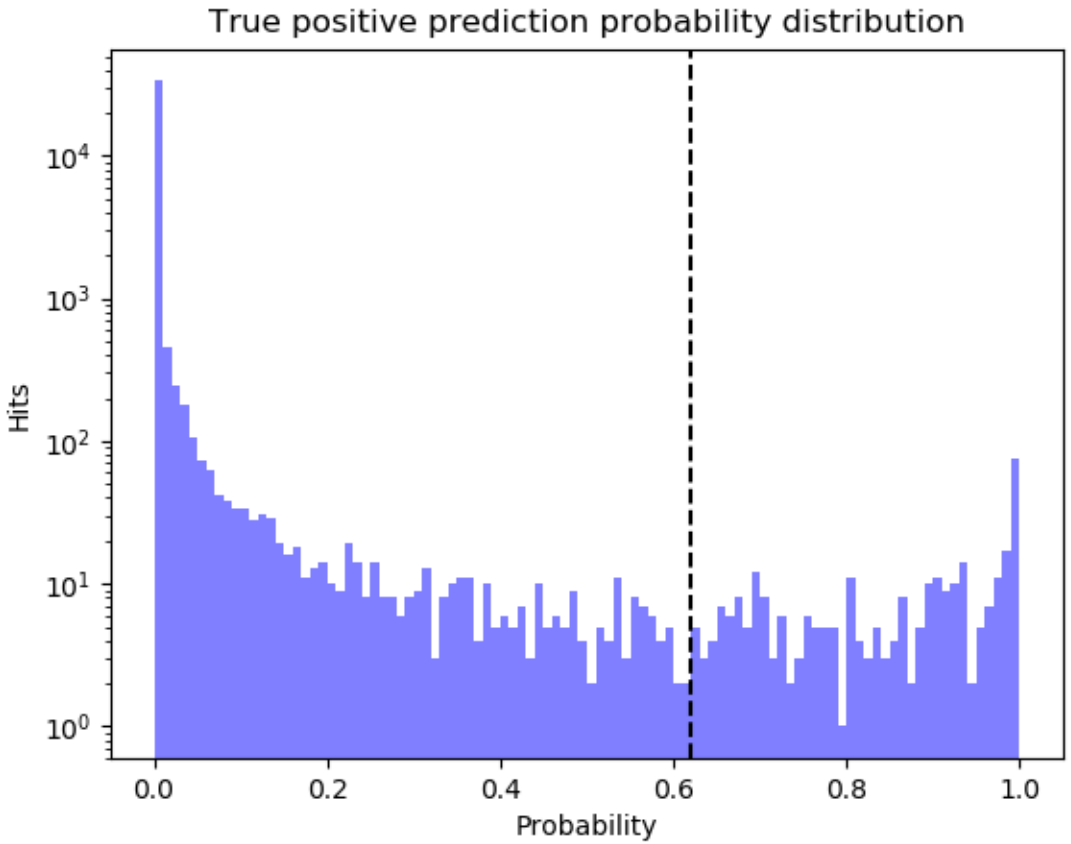


Demo: Dataset-2 – Clean Files (35599)

Source: SXL3 URL logs for May 9-24, 2018

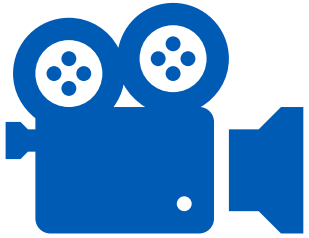


	FP	FPR
MLPdf	311	0.87%

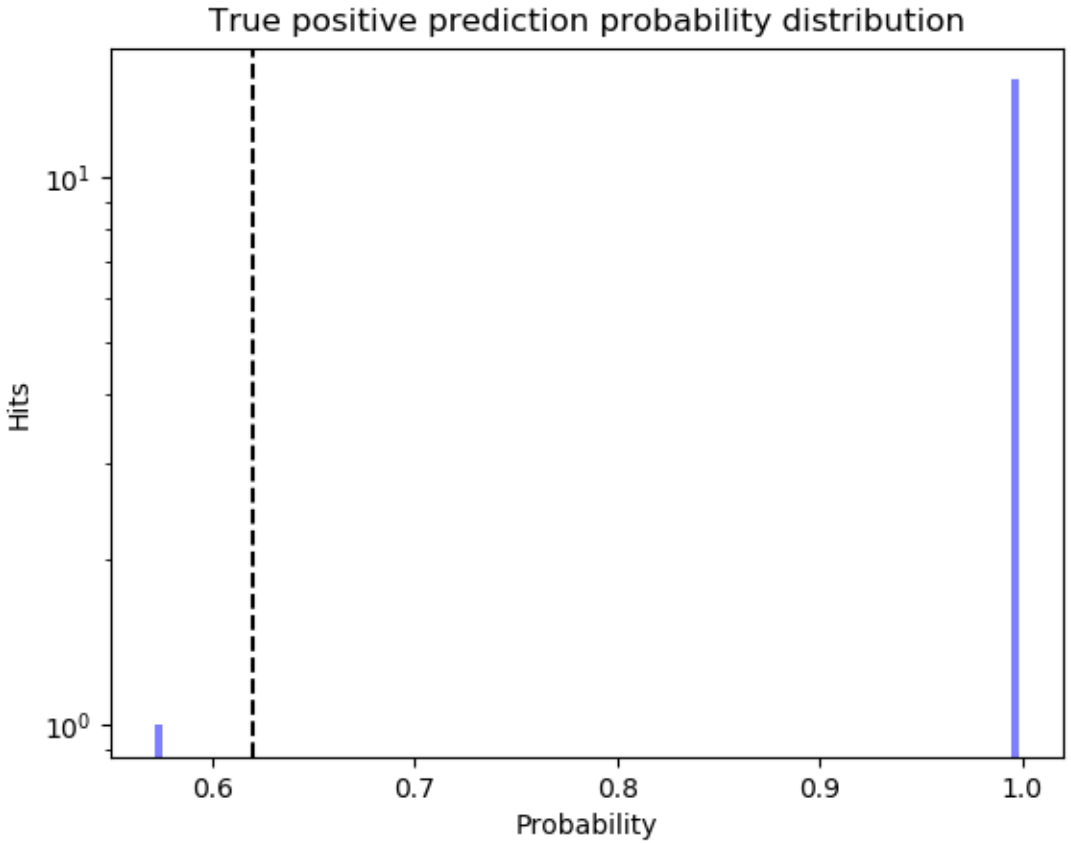


Demo: Dataset-3 – A Case Study with a Zero-Day Attack

Source: Sophos filesDB – July, 2018



	FN	TPR
MLPdf	1	93.75%



Summary

- Overview of PDF file format and typical attacks
- Introduction to MLPdf model
- Feature engineering & general challenges in ML
- MLPdf demo with various datasets, including analysis for a zero-day attack
- Pros & Cons of MLPdf vs traditional AV scanners

QA

Thank You

Git: https://github.com/cyberML/MLPdf

E: jason.zhang@sophos.com

IN: https://uk.linkedin.com/in/jasonzhanguk

SOPHOS

Security made simple.