CAMLIS

Towards a Trustworthy and Resilient Machine Learning Classifier

A Case Study of Ransomware Detector Creation

Evan Yang; Ravi Sahita

Intel Lab/SPR

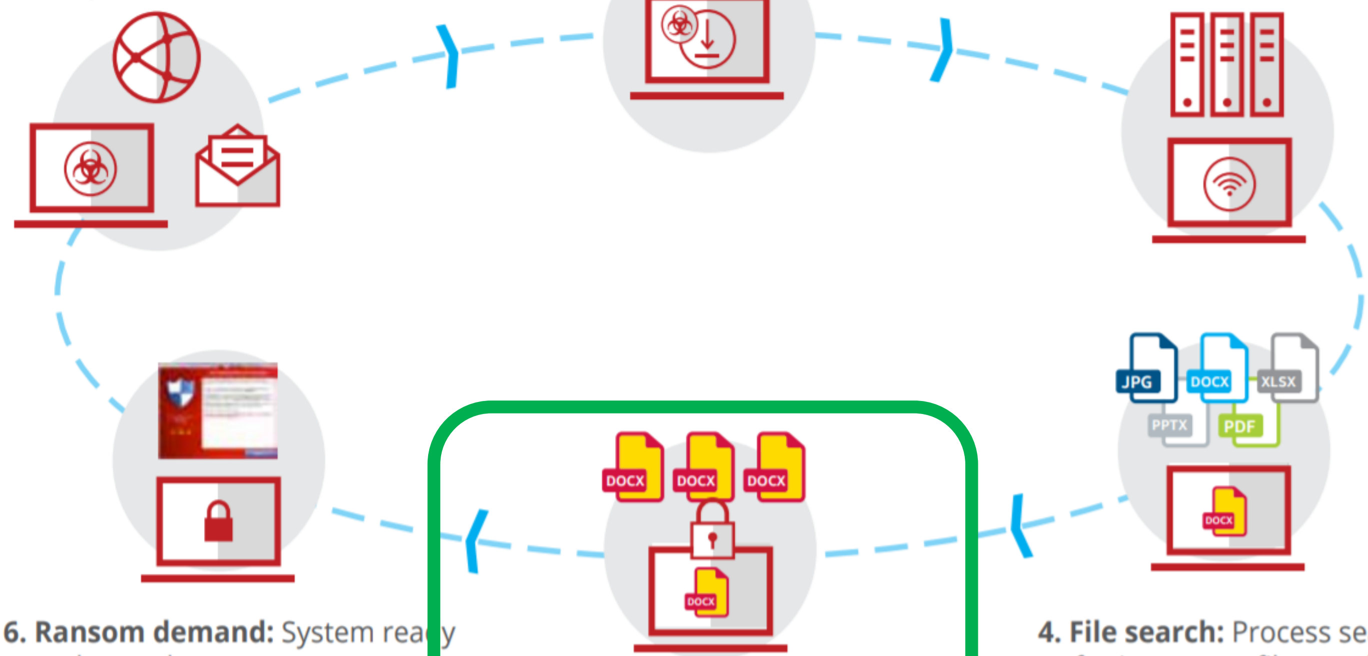
Outline

- Background
- Issues of Classifier
- Model Fidelity
- Adversarial Research
- Conclusion

Background (Al-rimy, B. et al. 2018)

- Ransomware is a category of malware which hijacks victim's data or machine and demands monetary returns
- Taxonomy:
  - Locker-ransomware: hijack resources without encryption
  - Crypto-ransomware: encrypt files
- The damage done by crypto-ransomware is **irreversible** in most cases due to the use of cryptography

Typical Steps of Ransomware (McAfee 2017)



1. Distribution: Standard methods of email attachments, website compromises.

2. Infection: Arrives on user's computer and starts processes.

3. Communication: Process talks to encryption-key servers.

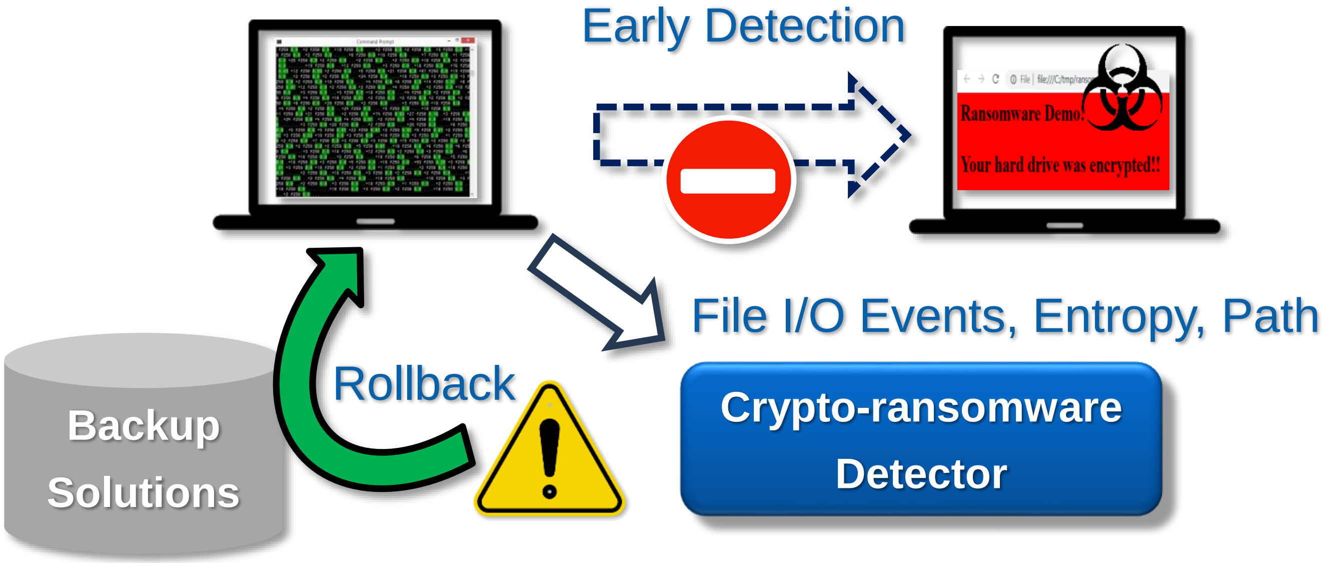
4. File search: Process searches for important files on the system e.g. JPG, DOCX, XLSX, PPTX, PDF.

5. Encryption: Typically done through rename, encrypt, rename again.

6. Ransom demand: System ready to demand payment.

Purpose of Detector:

- Find crypto-ransomware early by its behavior when AV missed it



Early Detection

File I/O Events, Entropy, Path

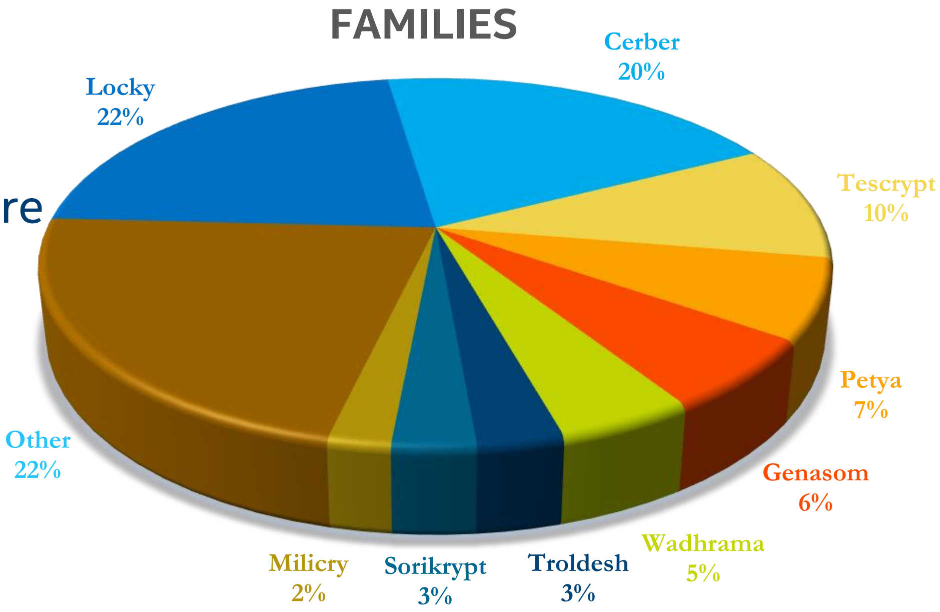
Crypto-ransomware Detector

Rollback

Backup Solutions

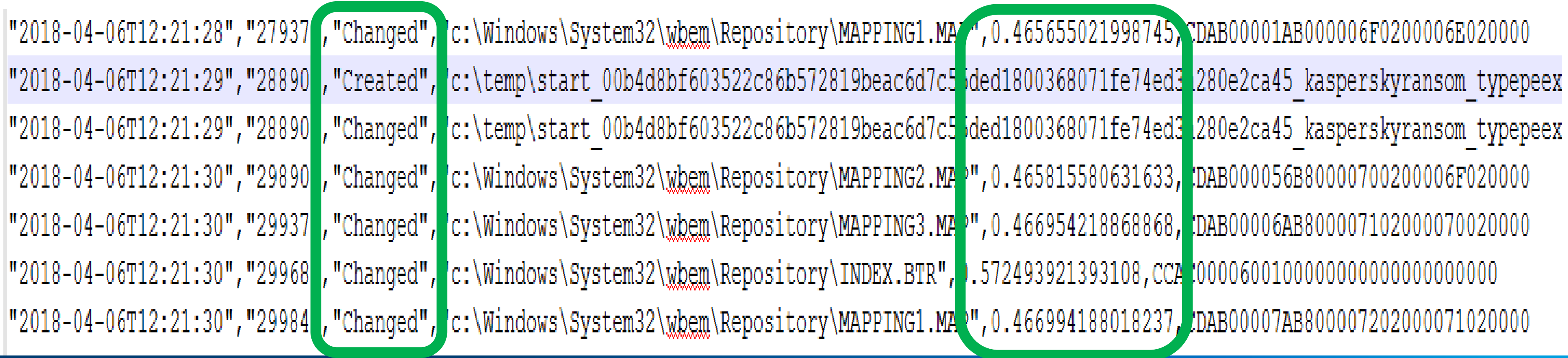
Ransomware Dataset

- From VirusTotal
  - Downloaded total ~22k ransomware by Microsoft and Kaspersky's labels
  - ~5min execution for each sample
  - In bare-metal sandbox system with anti-evasion mechanism
- Decoy files to identify crypto-ransomware
- Total ~4.4k active samples:



Behavior Data – File Input/Output Events

- Collected by POC Windows application
  - Based on **`C#.Net`** framework, **`FileSystemWatcher`** (FSW)
  - Entropy of target files calculated by normalized Shannon entropy
- Sample data:
  - Time stamp, I/O event type, target filename, entropy etc.

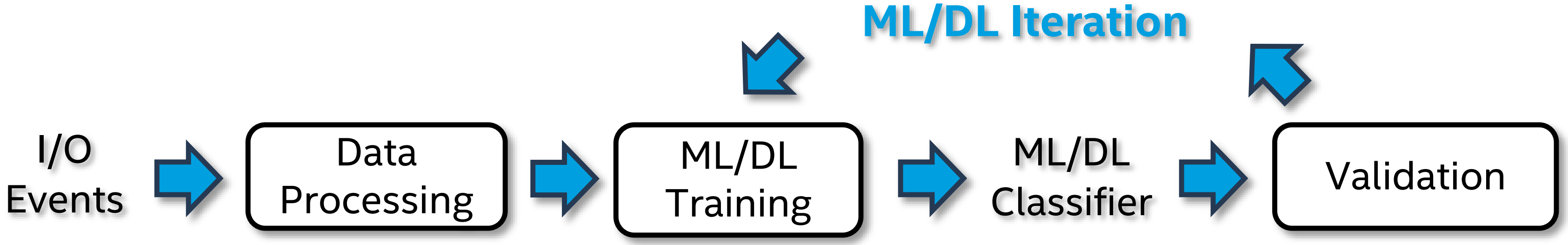


```
"2018-04-06T12:21:28","27937","Changed","c:\Windows\System32\wbem\Repository\MAPPING1.MAP",0.465655021998745,CDAB00001AB000006F0200006E020000
"2018-04-06T12:21:29","28890","Created","c:\temp\start_00b4d8bf603522c86b572819beac6d7c55ded1800368071fe74ed3a280e2ca45_kasperskyransom_typepeex
"2018-04-06T12:21:29","28890","Changed","c:\temp\start_00b4d8bf603522c86b572819beac6d7c55ded1800368071fe74ed3a280e2ca45_kasperskyransom_typepeex
"2018-04-06T12:21:30","29890","Changed","c:\Windows\System32\wbem\Repository\MAPPING2.MAP",0.465815580631633,CDAB000056B8000070200006F020000
"2018-04-06T12:21:30","29937","Changed","c:\Windows\System32\wbem\Repository\MAPPING3.MAP",0.466954218868868,CDAB00006AB80000710200007020000
"2018-04-06T12:21:30","29968","Changed","c:\Windows\System32\wbem\Repository\INDEX.BTR",0.572493921393108,CCA[illegible]000060010000000000000000000
"2018-04-06T12:21:30","29984","Changed","c:\Windows\System32\wbem\Repository\MAPPING1.MAP",0.466994188018237,CDAB00007AB80000720200007102000
```

Machine Learning Analysis

- ~3.7k ransomware and similar amount of benign data (~100 applications). 80/20 split for training/testing dataset
- Featuring by event type with bucketed entropy (-,0.2,0.4,0.6,0.8,0.9)
  - Categorize into distinct features
- ML Algorithms for supervised learning:
  - Long-Short Term Memory (LSTM), Recurrent Neural Networks
  - Linear Support Vector Machine (SVM) with bag of N-gram, N=1 & 2

ML Pipeline & Outcome of Supervised Learning



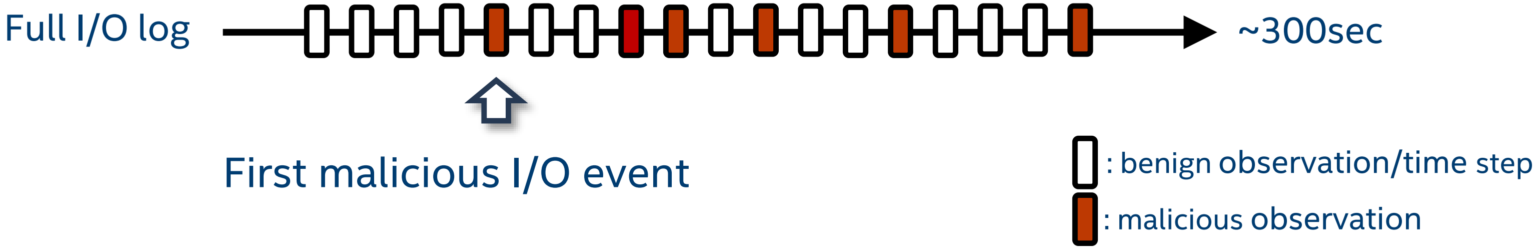
Model	N–gram	Accuracy	FPR	Dist. Features
Linear SVM	1 & 2	98.31%	2.89%	90
LSTM	n/a	98.67%	1.38%	9

Online Detector

- A POC program utilized the ML classifier
  - Sample the I/O event stream by a sliding window
  - Real-time inference: small footprint and run fast
- Issues found after deployment:
  1. False alarms from some applications
  2. Size of sliding window affects the detection rate
  3. Cannot find ransomware early

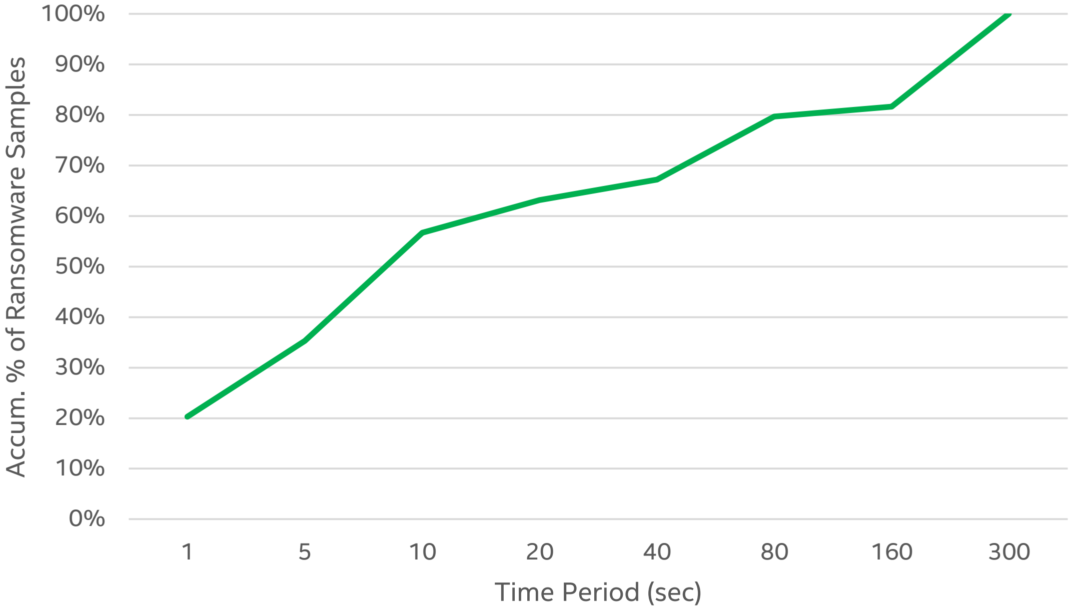
1. Early Detection Issue

- Early detection is important
  - No practical value if can’t detect encryption early
- When will the ransomware start doing encryption?
  - Identify the starting time by the decoy file



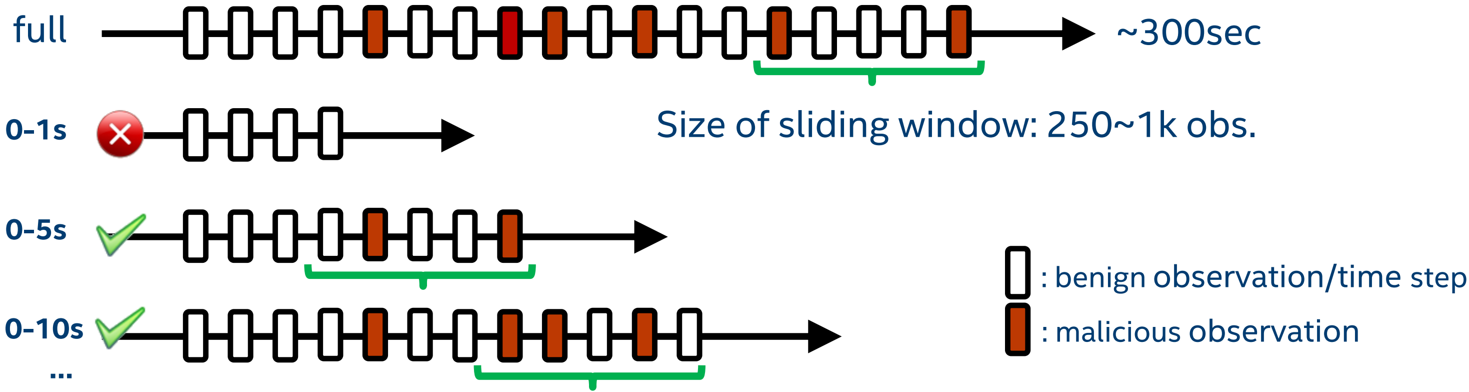
Starting Time of Malicious Activities

- Ransomware may not show malicious activities at the beginning of execution



Early Detection and Sliding Window Testing

- Prepare samples to measure the performance
  - From ~700 unseen *out-of-sample* ransomware logs
- Extract early-stage data from each logs by
  - different time periods
  - different sliding windows
- Example:

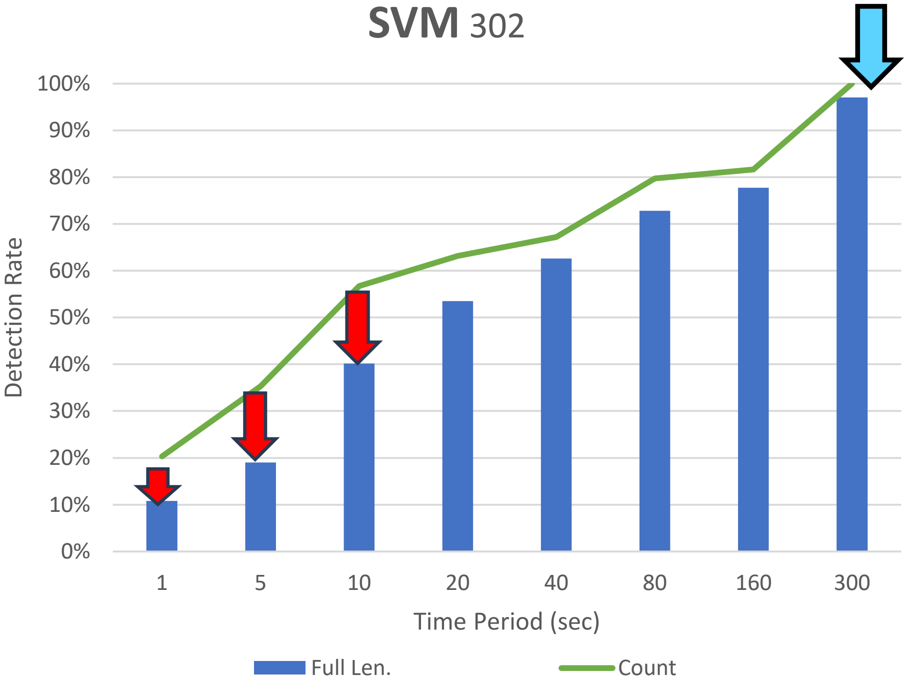


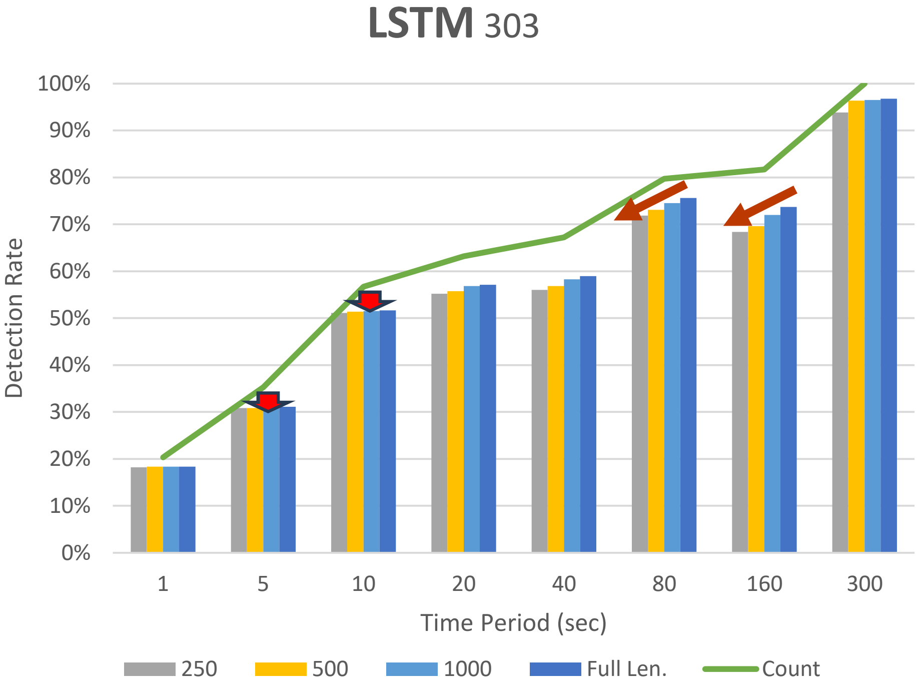
Size of sliding window: 250~1k obs.

□ : benign observation/time step

■ : malicious observation

Detection Rate of Early-stage and Sliding-window





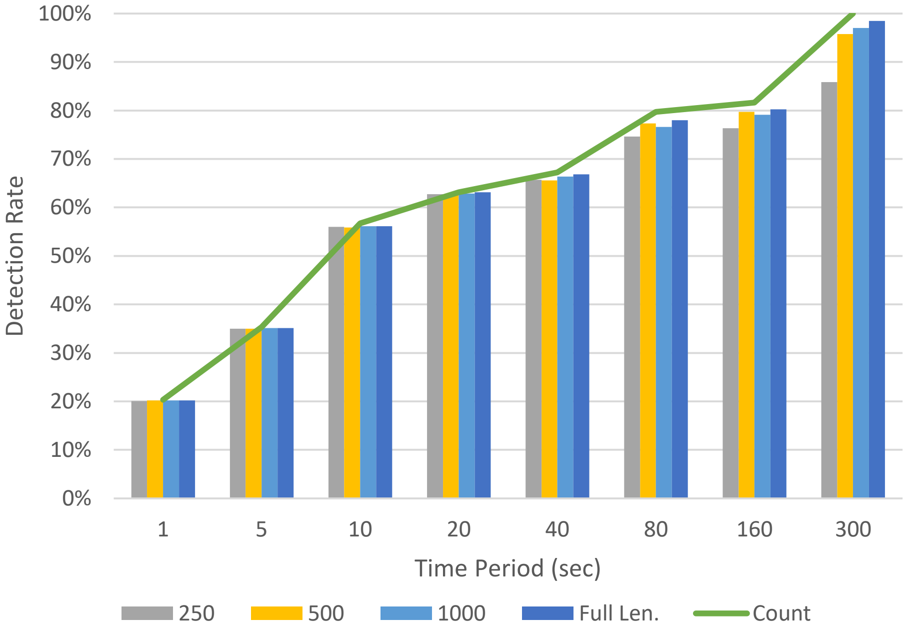
Data Augmentation

- Synthesize samples from existing dataset for a re-train
  - Early-stage samples
  - Sliding-window samples
  - Exclude samples without malicious events
- "Augmented" dataset count: 17.2k ransomware (80/20 split)

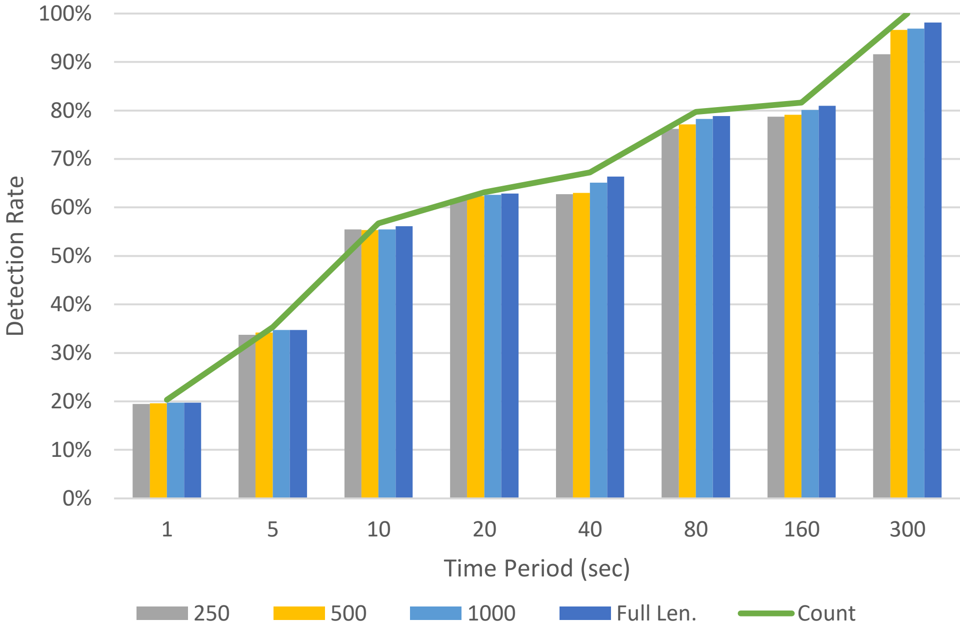
Model	N–gram	Accuracy	FPR	Dist. Feature
Linear SVM	1 & 2	99.13%	1.21%	90
LSTM	n/a	99.47%	0.60%	9

Detection Rate by Augmented Classifier

SVM-A 319



LSTM-A 320



3. False Positive Issue

- Some benign-ware has similar ransomware behaviors
  - Delete or rename many files, change files with high entropy

- **Solution**: Add a new dimension to feature
  - Path: system vs. non-system folders
  - System path list: `c:\Windows`, `c:\ProgramData`, `c:\Program Files`, `c:\Progra~`, `c:\AppData`, `\Downloads\`, `\Downlo~`, `c:\Config.msi`

Results with Path Flag

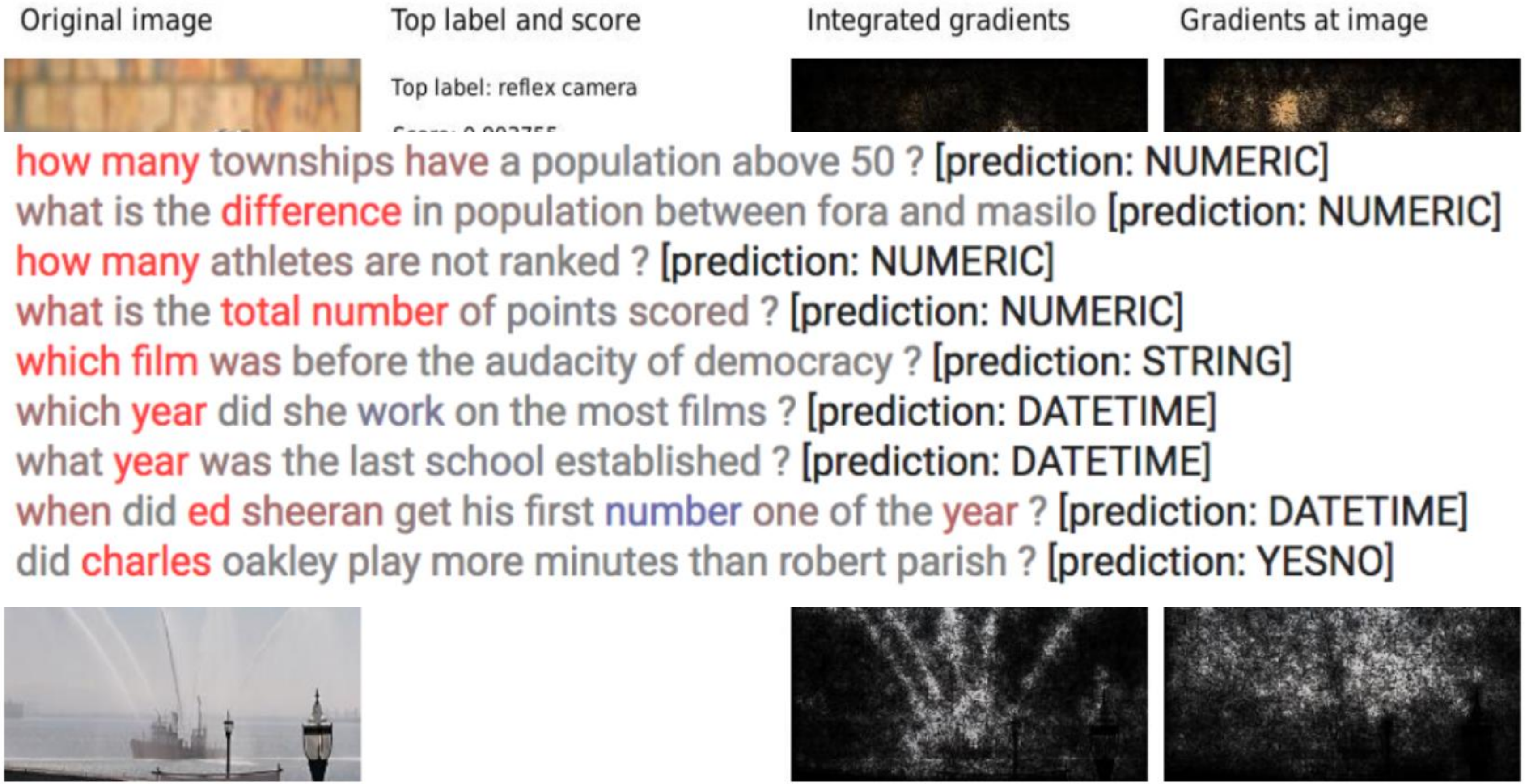
- Lower FPR with flag added

Model	N–gram	Accuracy	FPR	Dist. Features
Linear SVM	1 & 2	99.00%	1.34%	90
Linear SVM (+ path)	1 & 2	99.53%	0.54%	339
LSTM	–	98.26%	3.82%	9
LSTM (+ path)	–	98.35%	1.80%	18

- 22k out-of-sample clean execution log:
  - FPR down from 0.18% to 0.00% for SVM (40->0/22,174)
  - FPR down from 0.09% to 0.04% for LSTM (21->9/22,174)

Model Fidelity by Integrated Gradients Sundararajan M et al ‘17

- Attribution: which feature/time step contribute the most?



how many townships have a population above 50 ? [prediction: NUMERIC]

what is the difference in population between fora and masilo [prediction: NUMERIC]

how many athletes are not ranked ? [prediction: NUMERIC]

what is the total number of points scored ? [prediction: NUMERIC]

which film was before the audacity of democracy ? [prediction: STRING]

which year did she work on the most films ? [prediction: DATETIME]

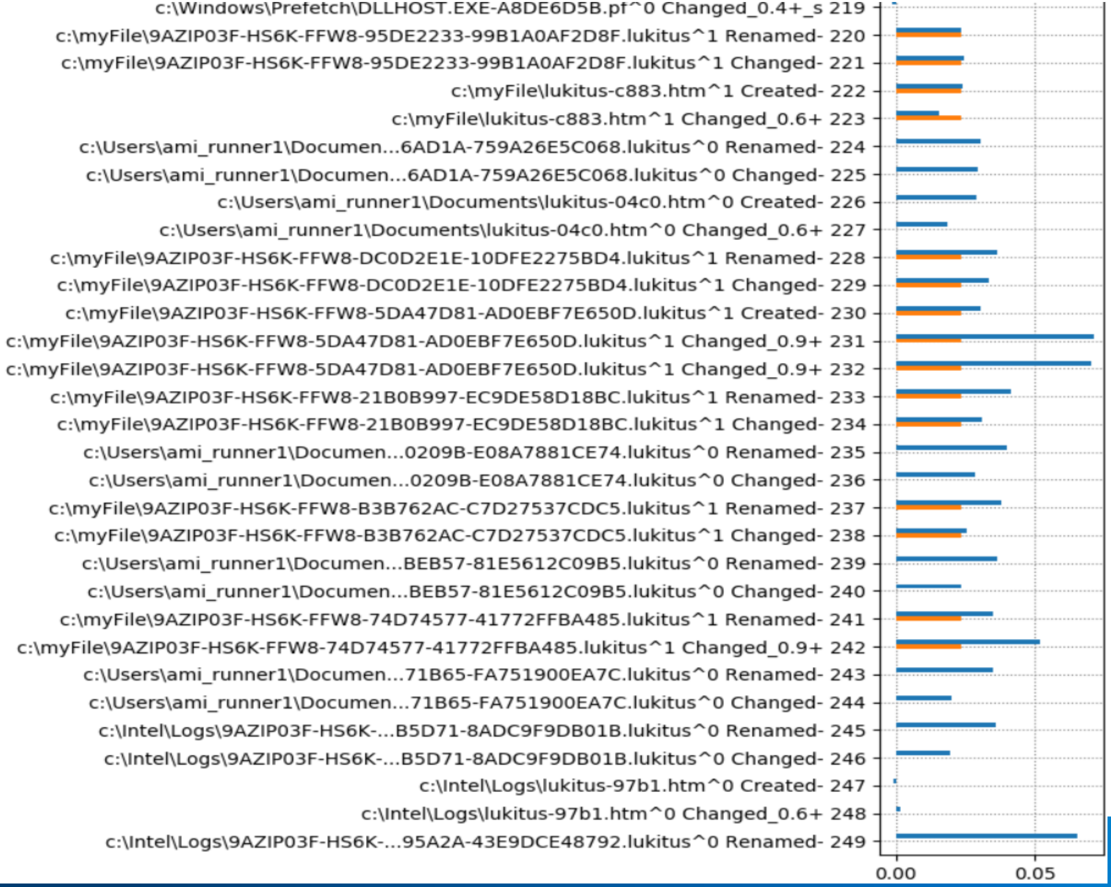
what year was the last school established ? [prediction: DATETIME]

when did ed sheeran get his first number one of the year ? [prediction: DATETIME]

did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Explanation of LSTM Models

- Feature attribution plot of ransomware:

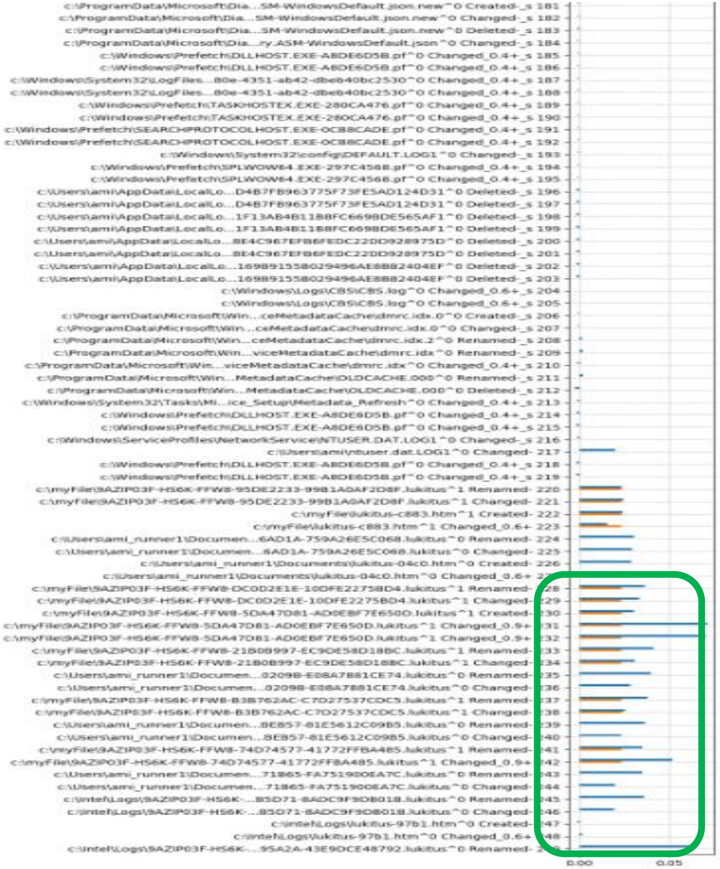

c:\Windows\Prefetch\DLLHOST.EXE-A8DE6D5B.pf^0 Changed_0.4+_s 219
c:\myFile\9AZIP03F-HS6K-FFW8-95DE2233-99B1A0AF2D8F.lukitus^1 Renamed- 220
c:\myFile\9AZIP03F-HS6K-FFW8-95DE2233-99B1A0AF2D8F.lukitus^1 Changed- 221
c:\myFile\lukitus-c883.htm^1 Created- 222
c:\myFile\lukitus-c883.htm^1 Changed_0.6+ 223
c:\Users\ami_runner1\Documen...6AD1A-759A26E5C068.lukitus^0 Renamed- 224
c:\Users\ami_runner1\Documen...6AD1A-759A26E5C068.lukitus^0 Changed- 225
c:\Users\ami_runner1\Documents\lukitus-04c0.htm^0 Created- 226
c:\Users\ami_runner1\Documents\lukitus-04c0.htm^0 Changed_0.6+ 227
c:\myFile\9AZIP03F-HS6K-FFW8-DC0D2E1E-10DFE2275BD4.lukitus^1 Renamed- 228
c:\myFile\9AZIP03F-HS6K-FFW8-DC0D2E1E-10DFE2275BD4.lukitus^1 Changed- 229
c:\myFile\9AZIP03F-HS6K-FFW8-5DA47D81-AD0EBF7E650D.lukitus^1 Created- 230
c:\myFile\9AZIP03F-HS6K-FFW8-5DA47D81-AD0EBF7E650D.lukitus^1 Changed_0.9+ 231
c:\myFile\9AZIP03F-HS6K-FFW8-5DA47D81-AD0EBF7E650D.lukitus^1 Changed_0.9+ 232
c:\myFile\9AZIP03F-HS6K-FFW8-21B0B997-EC9DE58D18BC.lukitus^1 Renamed- 233
c:\myFile\9AZIP03F-HS6K-FFW8-21B0B997-EC9DE58D18BC.lukitus^1 Changed- 234
c:\Users\ami_runner1\Documen...0209B-E08A7881CE74.lukitus^0 Renamed- 235
c:\Users\ami_runner1\Documen...0209B-E08A7881CE74.lukitus^0 Changed- 236
c:\myFile\9AZIP03F-HS6K-FFW8-B3B762AC-C7D27537CDC5.lukitus^1 Renamed- 237
c:\myFile\9AZIP03F-HS6K-FFW8-B3B762AC-C7D27537CDC5.lukitus^1 Changed- 238
c:\Users\ami_runner1\Documen...BEB57-81E5612C09B5.lukitus^0 Renamed- 239
c:\Users\ami_runner1\Documen...BEB57-81E5612C09B5.lukitus^0 Changed- 240
c:\myFile\9AZIP03F-HS6K-FFW8-74D74577-41772FFBA485.lukitus^1 Renamed- 241
c:\myFile\9AZIP03F-HS6K-FFW8-74D74577-41772FFBA485.lukitus^1 Changed_0.9+ 242
c:\Users\ami_runner1\Documen...71B65-FA751900EA7C.lukitus^0 Renamed- 243
c:\Users\ami_runner1\Documen...71B65-FA751900EA7C.lukitus^0 Changed- 244
c:\Intel\Logs\9AZIP03F-HS6K-...B5D71-8ADC9F9DB01B.lukitus^0 Renamed- 245
c:\Intel\Logs\9AZIP03F-HS6K-...B5D71-8ADC9F9DB01B.lukitus^0 Changed- 246
c:\Intel\Logs\lukitus-97b1.htm^0 Created- 247
c:\Intel\Logs\lukitus-97b1.htm^0 Changed_0.6+ 248
c:\Intel\Logs\9AZIP03F-HS6K-...95A2A-43E9DCE48792.lukitus^0 Renamed- 249
0.00
0.05

Adversarial Research

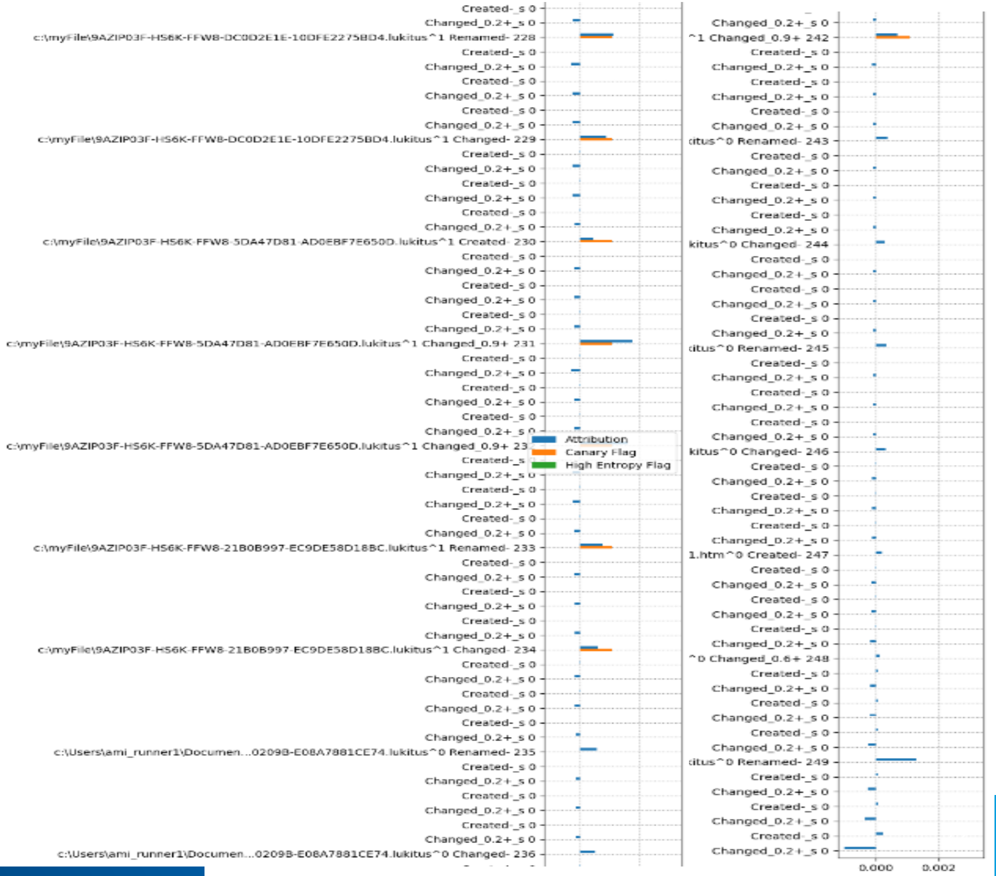
- A *simulated* ransomware, the *Red* team, was developed in C#
  - Rename, encrypt and delete files etc.
  - Evasive tricks to probe the detector (grey box attack):
    - Behavior temporal changes: e.g. slowdown the malicious activities
    - Encryption changes: e.g. insert dummy data to lower the file entropy
  - It’s not difficult to evade our ML detector
- Improve model’s resiliency by:
  - Discover weakness by the Red team with various conditions
  - Re-train model by the false negatives samples

Probing LSTM Models – by Event Insertion

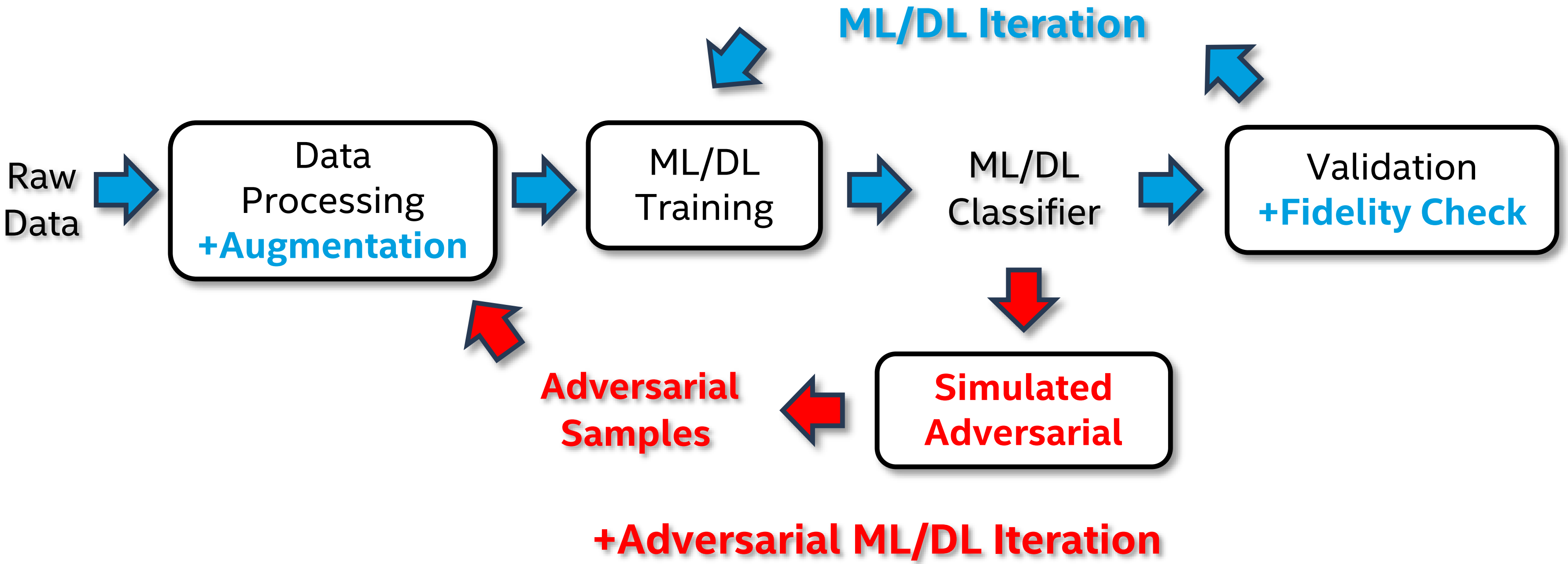
Original Sample, +, 0.96



Insert 7 benign events, -, 0.01



Conclusion: ML Pipeline +++



ML/DL Iteration

Raw Data

Data Processing +Augmentation

ML/DL Training

ML/DL Classifier

Validation +Fidelity Check

Simulated Adversarial

Adversarial Samples

+Adversarial ML/DL Iteration

Our Team Members and Projects

- Erdem Aktas; Li Chen; Anindya Paul
- MLsploit: a platform for ML model comparison and sample sharing for adversarial research
  - github.com/mlsploit
  - github.com/intel/Resilient-ML-Research-Platform

Georgia Tech

MLsploit

ML Anti Malware Int.

Thank You!

Legal Notices and Disclaimers

- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. These materials are provided as-is, with no express or implied warranties. All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Intel does not control or audit third-party data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
-
- *Other names and brands may be claimed as the property of others.