

# Securing Malware Cognitive Systems against Adversarial Attacks

Yuede Ji    Benjamin Bowman    **H. Howie Huang**

Graph Computing Lab  
The George Washington University

# Cognitive System

- A self-learning system leverages a combination of intelligent techniques, such as machine learning (ML), and data mining.
- It has made breakthrough performance in many applications, such as image processing, self-driving vehicles, and cybersecurity.



```
1105 112 100 111 104 99 106 99 96 103 112 110 104 97 93 871  
1 01 08 102 106 104 79 98 103 99 105 123 130 110 105 94 851  
1 76 85 98 105 120 105 87 96 95 99 113 112 100 103 99 851  
1 99 81 81 93 120 131 127 106 95 99 102 99 96 93 101 941  
1060 91 61 64 109 91 88 85 101 107 100 98 75 84 96 921  
1114 108 85 55 55 69 64 54 64 87 112 120 98 74 84 911  
1133 137 147 103 85 81 88 65 52 54 74 84 102 93 85 801  
1128 137 144 140 98 85 86 78 62 65 63 63 68 73 86 1011  
1125 133 140 137 119 121 117 94 65 79 88 65 54 64 72 981  
1127 125 131 147 133 127 126 131 111 96 89 75 61 64 72 841  
1115 114 109 123 138 140 131 118 113 109 100 92 74 65 72 781  
1 89 93 98 97 100 147 131 118 113 114 113 100 106 65 77 881  
1 63 77 86 81 77 79 102 123 117 115 117 125 125 130 115 871  
1 62 65 82 89 78 71 88 101 124 126 110 101 107 114 111 1191  
1 63 65 75 88 89 71 62 81 128 130 130 101 61 98 130 1181  
1 87 65 71 87 106 95 69 45 76 130 120 107 92 94 105 1121  
1138 87 82 86 117 123 114 64 43 93 93 85 85 102 1071  
1164 146 112 88 82 120 124 184 78 48 45 66 88 101 102 1091  
1157 120 117 128 93 86 114 122 112 97 69 55 78 62 98 841  
1138 128 134 161 139 100 109 118 121 134 114 87 65 53 89 861  
1128 112 96 117 138 144 128 115 104 107 102 93 87 81 72 791  
1123 107 96 86 83 112 153 149 122 100 104 78 88 107 112 981  
1122 121 102 88 82 86 94 117 145 140 153 102 58 78 92 1071  
1122 144 148 103 71 56 78 83 93 101 119 130 102 61 69 1011
```

what the computer sees

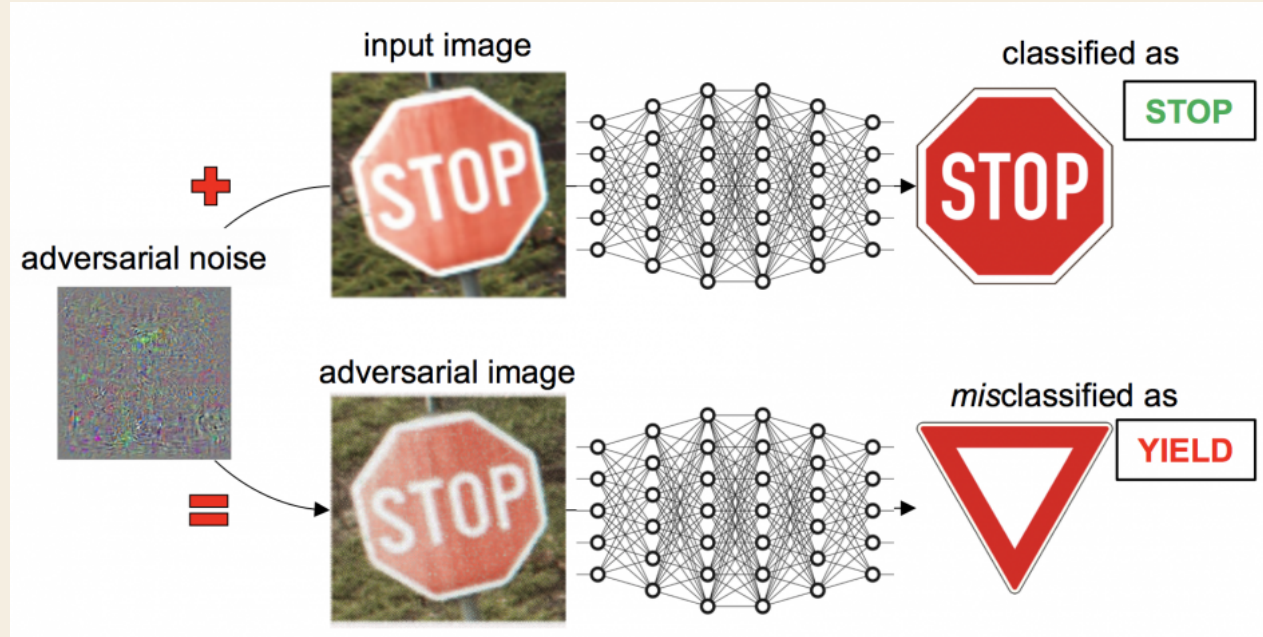
Image Classification

- 82% cat
- 15% dog
- 2% hat
- 1% mug



# Adversarial Attack

- Adversarial attacks try to cause the machine learning methods to misbehave or leak sensitive model information.
- The cognitive systems are vulnerable to adversarial attacks.



Picture credits to "Vaccinating machine learning against attacks"

---

# Malware Cognitive Systems

---

- Applying cognitive intelligence to malware detection
  - Gained great popularity, which has been used in Sparkcognition, Cisco, IBM, Cybereason.
- Such systems are vulnerable to adversarial attacks.



---

# Outline

---

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

# Background: Malware



WIKIPEDIA  
The Free Encyclopedia

[Main page](#)  
[Contents](#)

## 2019 Baltimore ransomware attack

From Wikipedia

The Baltimore ransomware c

**Forensic science**



## Another Hack a Ransom, Th

By **Patricia Mazzei**

June 27, 2019

MIAMI — Even the pho  
City, Fla., after hackers l  
city's computer systems

**Hannah Devlin**  
*Science correspondent*

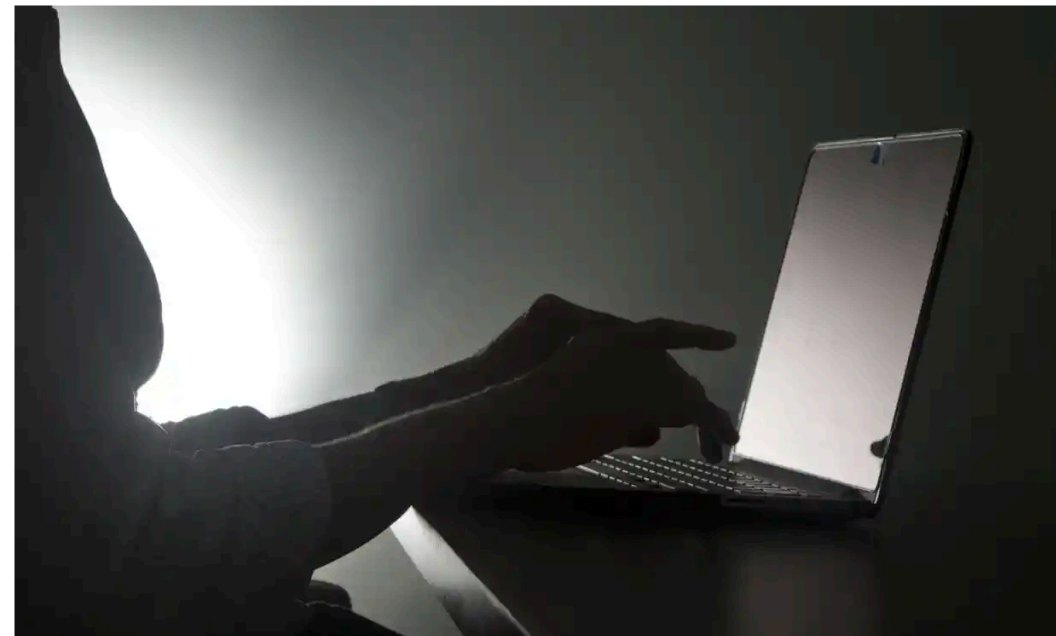
[@hannahdev](#)

Fri 5 Jul 2019 11.51 EDT



## Hacked forensic firm pays ransom after malware attack

Largest private provider Eurofins hands over undisclosed fee to regain control of systems



▲ Ransomware is a type of computer program that infiltrates IT systems and threatens to publish data or block access until money is paid. Photograph: Wilfredo Lee/AP

## Ransomware Hits Georgia Courts as Municipal Attacks Spre

07:49 PM

## HITS GEORGIA MUNICIPAL ATTACKS



---

# Background: Adversarial Attack

---

- Data poisoning attack
  - Training phase
  - Add “poisoned” training data to confuse the inference result.
- Evasion attack
  - Testing phase
  - Test multiple data to identify the network gradients, thus perform targeted attack.
- Exploratory attack
  - Testing phase
  - Aim to extract knowledge from a trained model instead of fooling it

---

# Outline

---

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

---

# Problem Definition

---

- **Task Definition**
  - Aim to defend evasion attacks for malware classification
  - Five malware classes, no benign software
- **Threat Model**
  1. The adversarial attacks can only happen at the testing stage.
  2. The adversaries may have knowledge of the training dataset, but are not allowed to modify it.
  3. The adversaries have no knowledge of the trained model (architecture, parameters).
  4. The adversaries only aim at degrading the performance in terms of accuracy metrics and are not attacking any confidentiality or privacy issues.

---

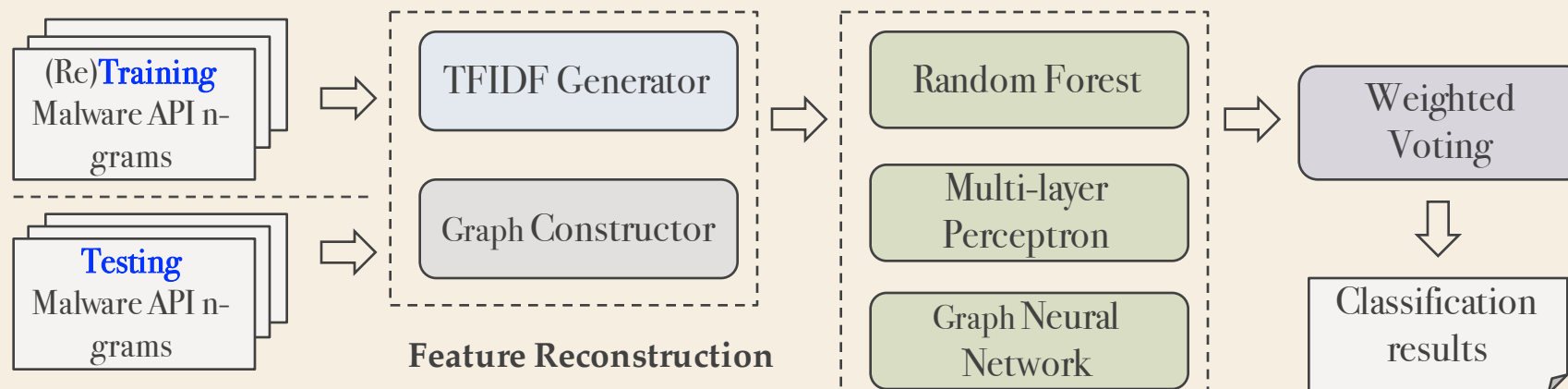
# Outline

---

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

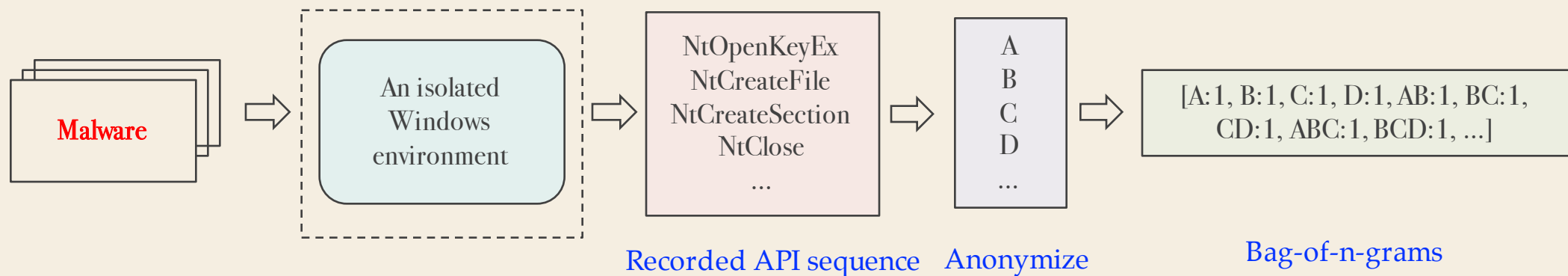
# DeepArmour Overview

- Feature Reconstruction
  - Term frequency-inverse document frequency (TFIDF)
  - Attributed graph
- Weighted Voting
  - Random forest, Multi-layer perceptron, and graph neural network
- Adversarial Retraining



# Malware Dataset

- Malware execution trace dataset [AAAI-19 AICS Challenge]
- 12,536 malware in five categories: Virus, Worm, Trojan, Packed Malware, AdWare
- Anonymized bag-of-n-grams ( $n = 1, 2, 3$ )
- Original trace is not available in this challenge





---

# Feature Reconstruction

---

- Term Frequency-Inverse Document Frequency (TFIDF)
  - A weighting factor intends to show the importance of a word to a document in large corpus
  - API → word, malware → document
- Attributed Graph
  - API → node, bi-gram → edge
  - Node attribution: [node\_id (1-hot), node\_freq, avg\_out\_edge\_freq, avg\_in\_edge\_freq]

---

# Weighted Voting

---

- Motivation
  - Most adversarial attacks are targeting one or one type of machine learning method.
- Three machine learning methods
  - Random forest (RF)
  - Multi-layer perceptron (MLP)
  - Structure2vec

---

# Adversarial Retraining

---

- One of the most effective adversarial countermeasures
- We generate adversarial samples on top of the training dataset
  - MLP targeted attack
    - Manipulate the inputs to a MLP model to produce incorrect output
  - Fast gradient sign method

---

# Outline

---

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

---

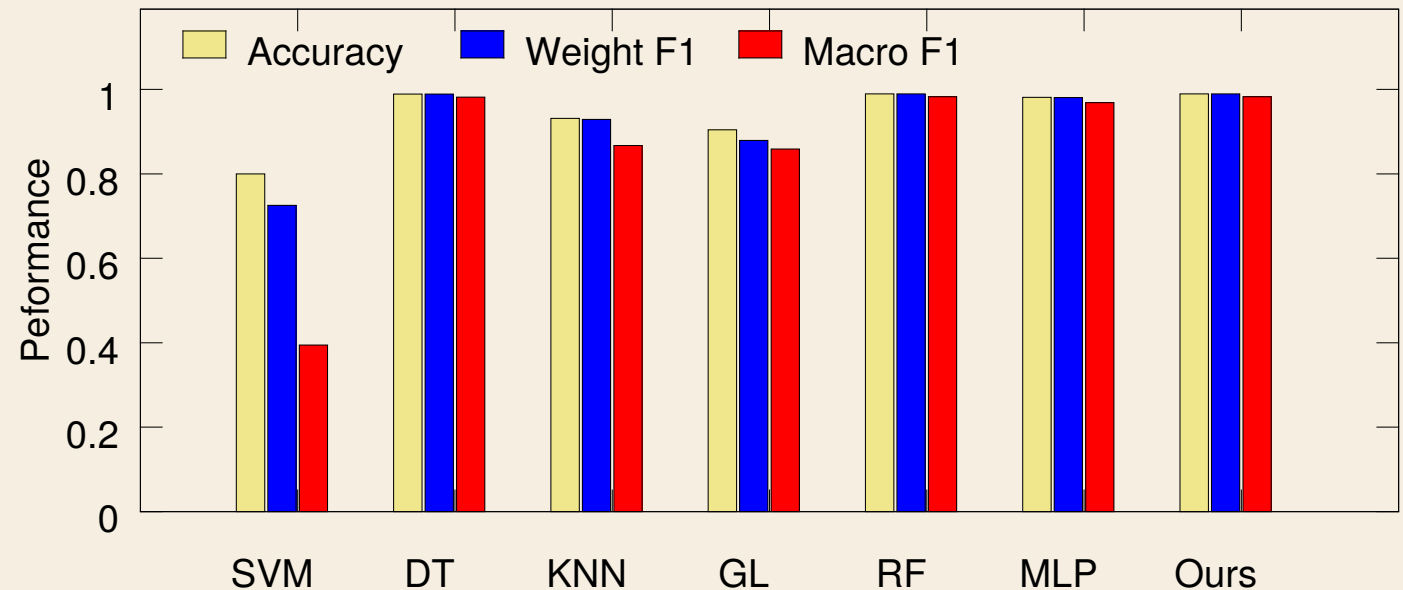
# Experiment

---

- Experiment Setting
  - Intel Xeon E5-2620 (2.00 GHz) CPU, 12 cores with 128 GB of main memory.
  - One Nvidia Tesla K40c GPU
  - Machine learning library, scikit-learn (version 0.19.1)
  - Neural network framework, TensorFlow (version 1.11.0)
- Performance Metrics
  - Accuracy
  - Weighted & Macro F1

# Malware Detection on Normal Dataset

- **10-fold cross validation**
- **Methods**
  - Support vector machine (SVM)
  - Decision tree (DT)
  - K-nearest neighbors (KNN)
  - Random forest (RF)
  - Multi-layer perceptron (MLP)
  - Structure2vec (GL)
- **Performance**
  - Accuracy: 99%
  - Weighted F1: 0.99
  - Macro F1: 0.98



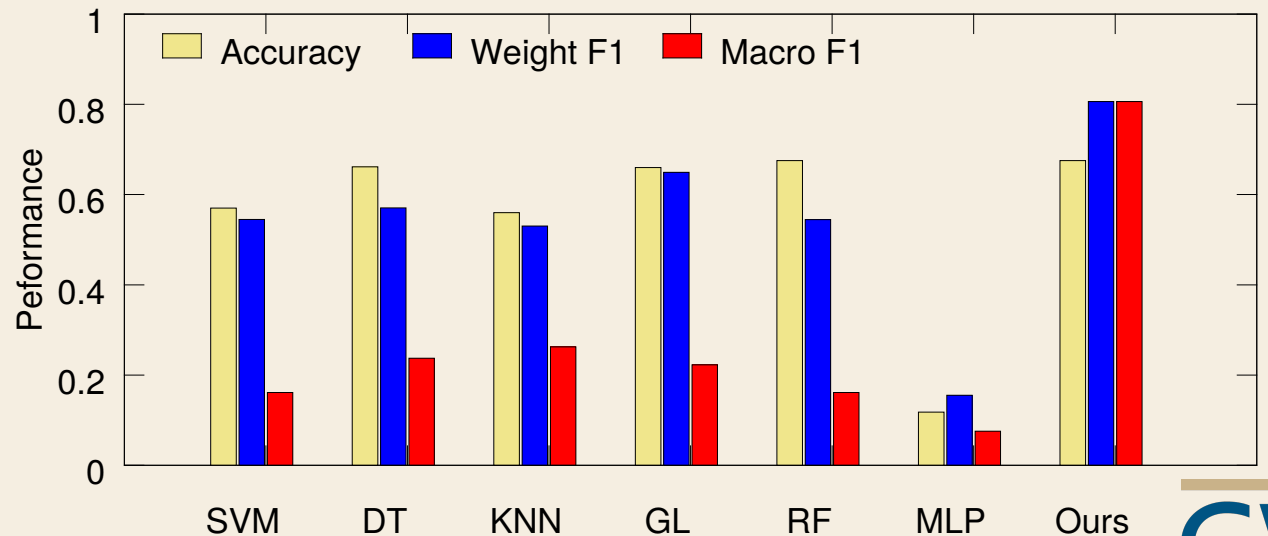
# Against Adversarial Attacks

- **Accuracy after the attack**

- MLP drops from 98% to 12%
- Everyone drops to ~60%

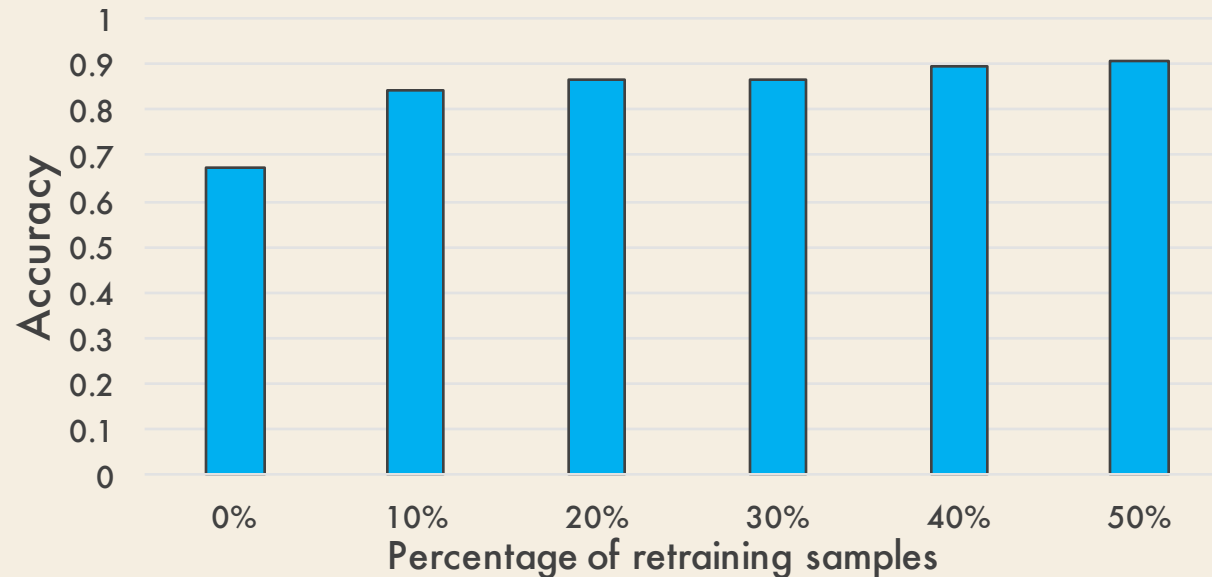
- Our approach achieves the best weighted/macro F1 of 0.8 vs. others 0.5/0.2

	Virus	Worm	Trojan	Packed Malware	Adware	Total
Normal malware	11,844	11,253	771	692	512	12,536
Generated adversarial	1,303	308	120	111	87	1,929



# Adversarial Retraining

- **Retraining with adversarial samples**
  - 10% retraining improves accuracy from 65% to 84%
  - 50% retraining achieves 90% accuracy





---

# Outline

---

- Background
- Problem Definition
- DeepArmour
- Experiment
- Conclusion

---

# Conclusion

---

- Takeaways
  - DeepArmour is a robust malware classification system, which is able to defend evasion adversarial attacks.
  - Malware detection & adversarial defenses are arms race, which needs to be evolved all the time.
- Future Works
  - Investigate other adversarial attacks
  - Focus on more malware types

---

# Thank You

---

The source code and data will soon be released at our repository at [github.com/iHeartGraph/](https://github.com/iHeartGraph/)



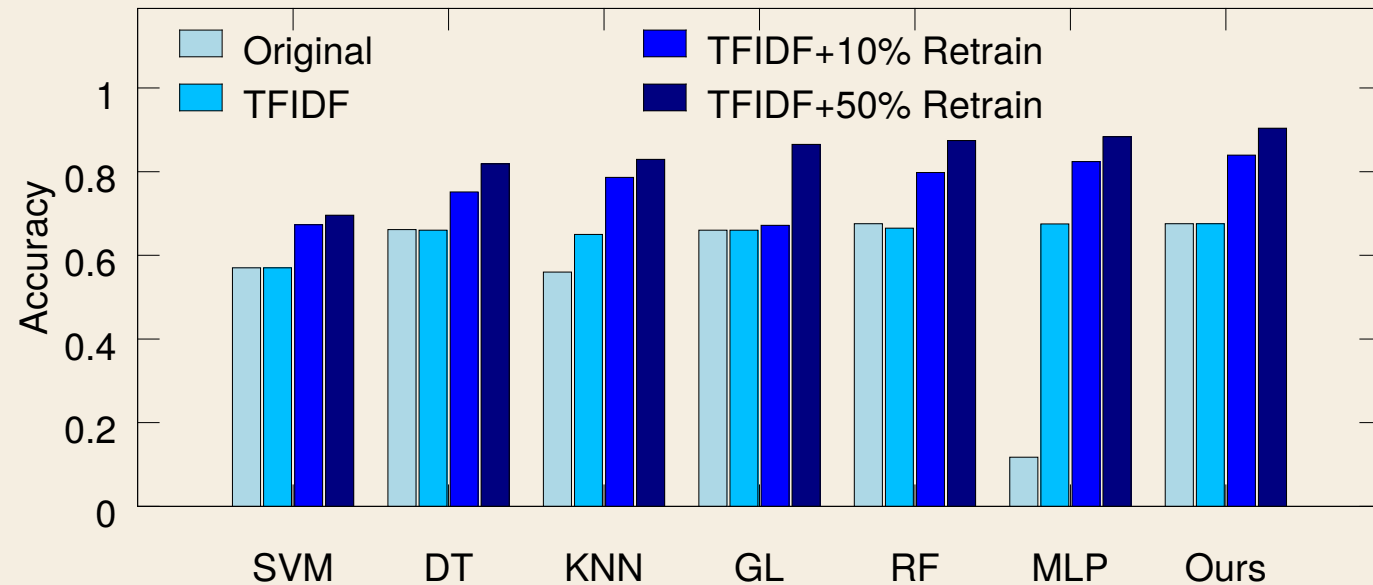
---

# Backup Slides

---

# Performance of Different Techniques

- TFIDF
  - MLP: accuracy improves from 12% to 68%
- Retraining



---

# Parameter Study

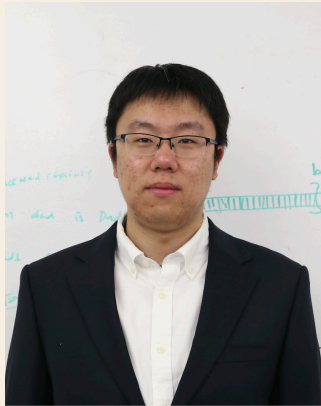
---

- Can put in backup

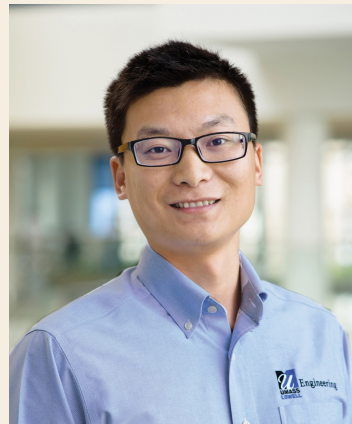
MLP

# Securing Malware Cognitive Systems against Adversarial Attacks

Yuede Ji



Benjamin Bowman



**H. Howie Huang**



The George Washington University  
University of Massachusetts Lowell