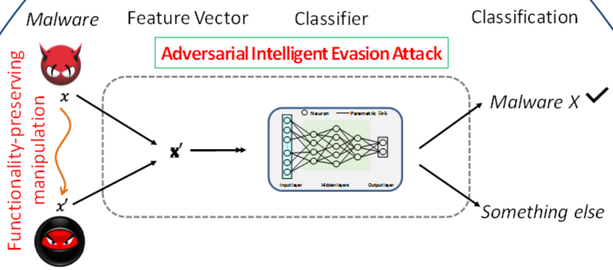
NSF SaTC: CORE: Small: Collaborative: A Framework for Enhancing the Resilience of Cyber Attack Classification and Clustering Mechanisms

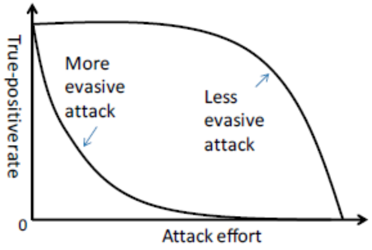
Challenge:

- How to **quantify** the vulnerability and resilience of classification and clustering mechanisms against adversarial intelligent cyber evasion attacks?
- How to **enhance** the resilience of classification and clustering mechanisms against adversarial intelligent cyber evasion attacks?

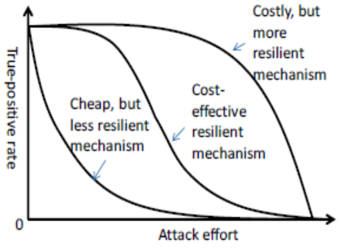
Solution:

- Defense principles → Framework → Metrics → Effective Mechanisms
- Key innovations: Systematic (black-, gray-, white-box) threat model quantification → deeper understanding of the enemy → better defense


Malware
Feature Vector
Classifier
Classification
Adversarial Intelligent Evasion Attack
Functionality-preserving manipulation
x
x′
x′
Malware X ✓
Something else



(b) Metric: True-positive rate



(b) Metric: True-positive rate

Application Scenario:

Adversarial malware classification leads to better defense against adversarial intelligent cyber attacks.

Scientific Impact:

- The project will **deepen understanding** of the vulnerability of AI/Machine Learning to adversarial intelligent cyber evasion attacks.
- The project will **invent countermeasures** to enhance resilience of AI/Machine Learning against adversarial intelligent cyber evasion attacks.

Broader Impact:

- **Safer AI/Machine Learning**
- **Potential transition to practice**
- **11 publications (including IJCAI'19, AAAI'19, WWW'19, ACSAC'18)**
- **MIT Lincoln Lab AICS 2019 Adversarial Malware Classification Challenge Winner**
- **IJCAI'19 Early Career Spotlights (Ye)**
- **10+ media reports on our results**
- **Female Co-PI**
- **3 PhD students involved in research**
- **10+ seminar/invited presentations**
- **3+ courses used research materials**

NSF SaTC #1814825

PI: Prof. Shouhuai Xu (University of Texas at San Antonio; shxu@cs.utsa.edu; www.cs.utsa.edu/~shxu)

Co-PI: Prof. Yanfang Ye (Case Western Reserve University)