Trustworthy Machine Learning in Data-Driven Cyber Security Practices

Yufei Han@CIDRE Project Team, INRIA Rennes

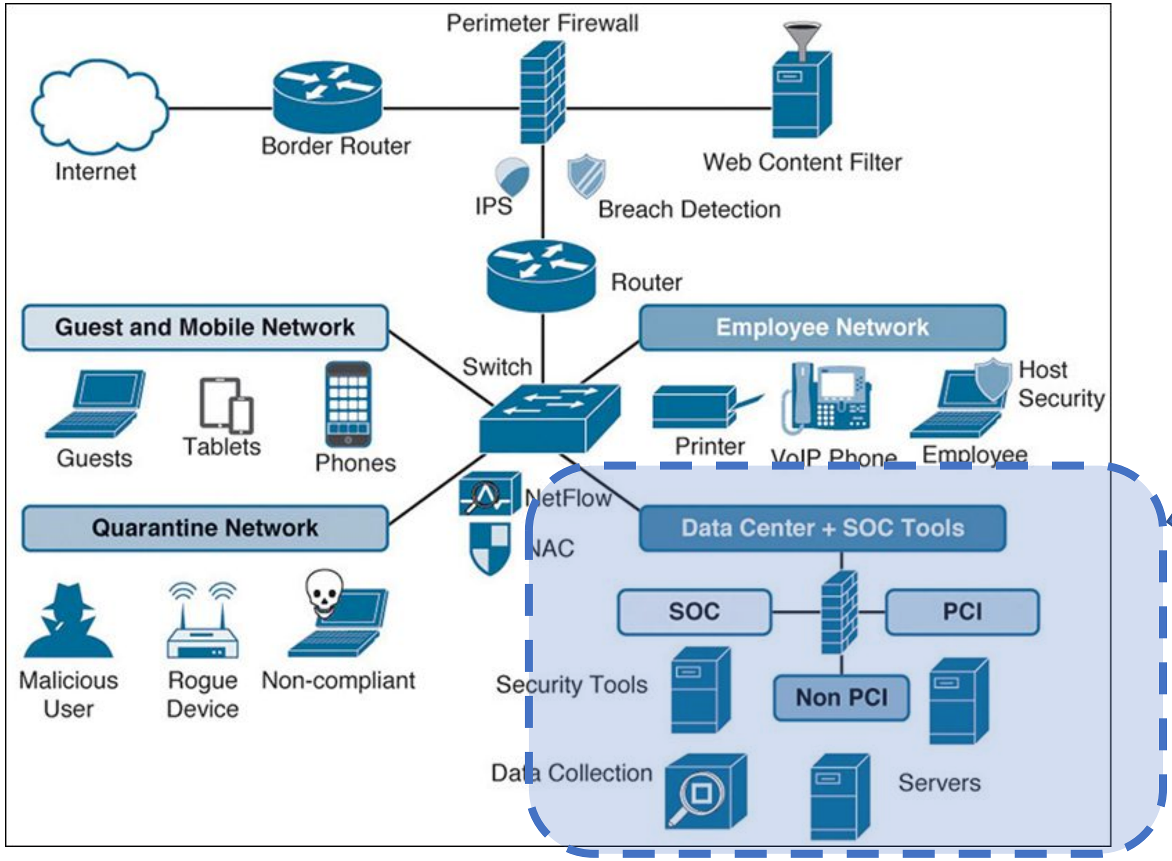
Feb 25, 2021

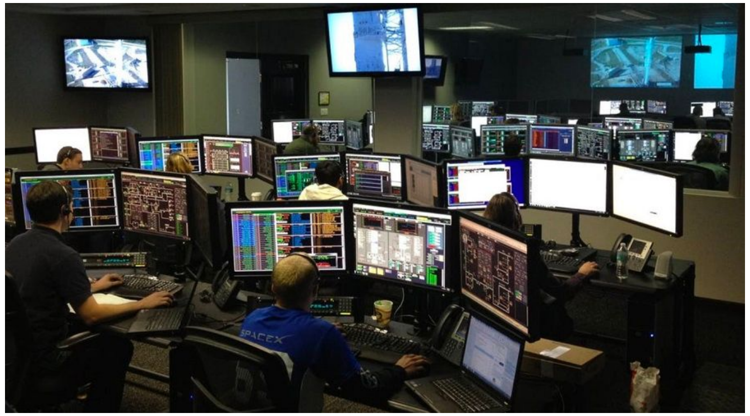
inventeurs du monde numérique

Use of Machine Learning for Security Practices

Machine Learning for Cyber Threat Detection, Classification and Prediction

Security **O**peration **C**enter (SOC) of Managed Security Service



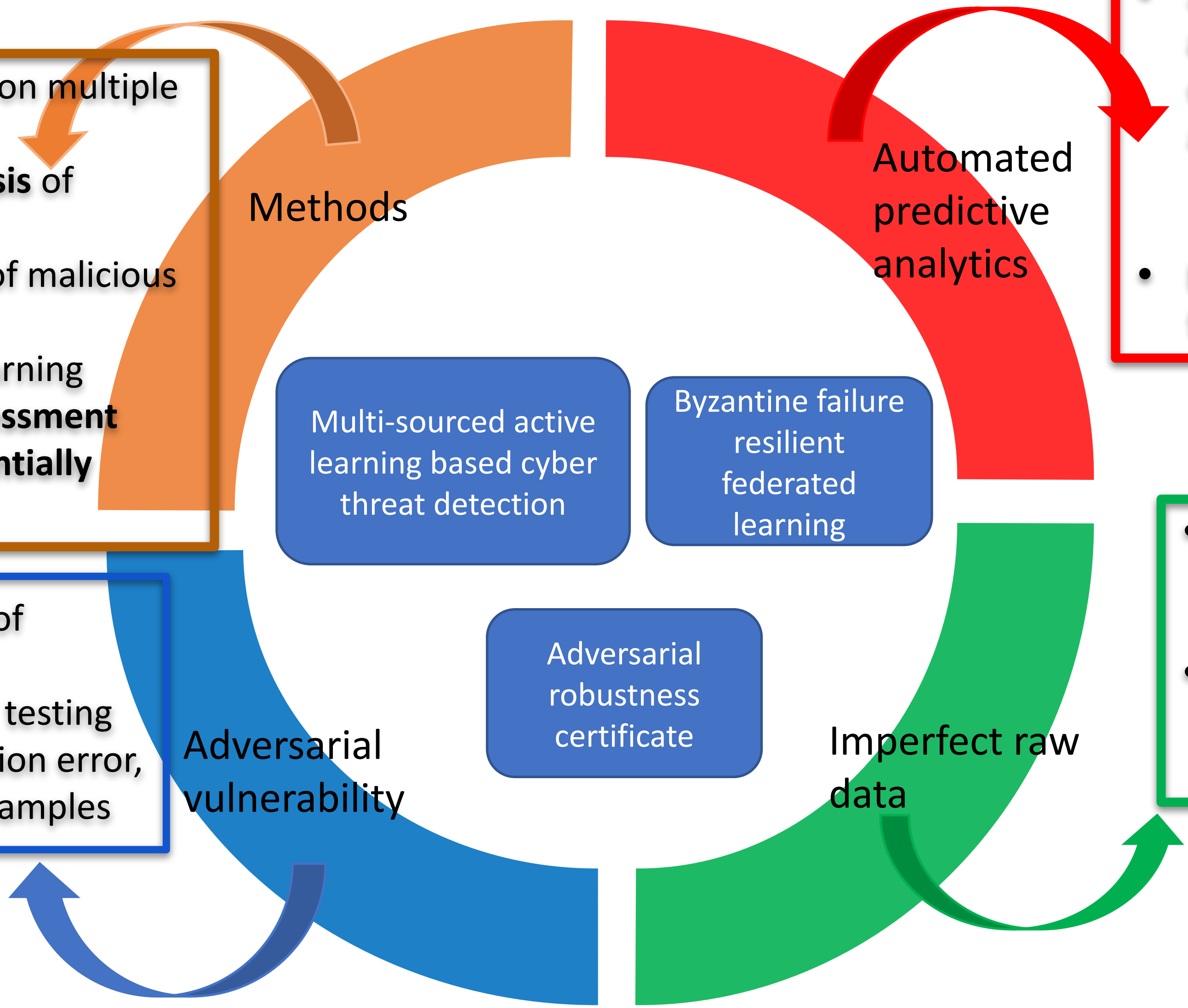


__Challenge raised by fast increasing cyber threats__:

- **Huge volume of data input of SOC**. For example the SOC of a mainstreaming security vendor receive reports of 3.7 million spear-phishing and website hijacking events. **Human experts can not verify all of them.**
- 1/3 reported incidents originate from zero-day vulnerability. **Prediction of potential threats** is thus important for active defense

Challenges to Trustworthy Machine Learning Service in Cyber Security Practices

- **Cyber threat** profiling based on multiple information sources
    - **Static / Dynamic analysis** of suspicious files
    - **Time-series modeling** of malicious incidents
- **Self-Supervised** Machine Learning
- **Adversarial Robustness Assessment**
- **Federated learning / Differentially Private Learning**



- **Semi-autonomous security analysis** to improve incident detection / prediction accuracy and efficiency
    - Human-in-the-loop attack understanding
- **Prediction of malware infection** for **active defense**

- **Adversarial vulnerability** of machine learning models
- Manipulation of training / testing data can induce classification error, e.g. adversarial malware samples

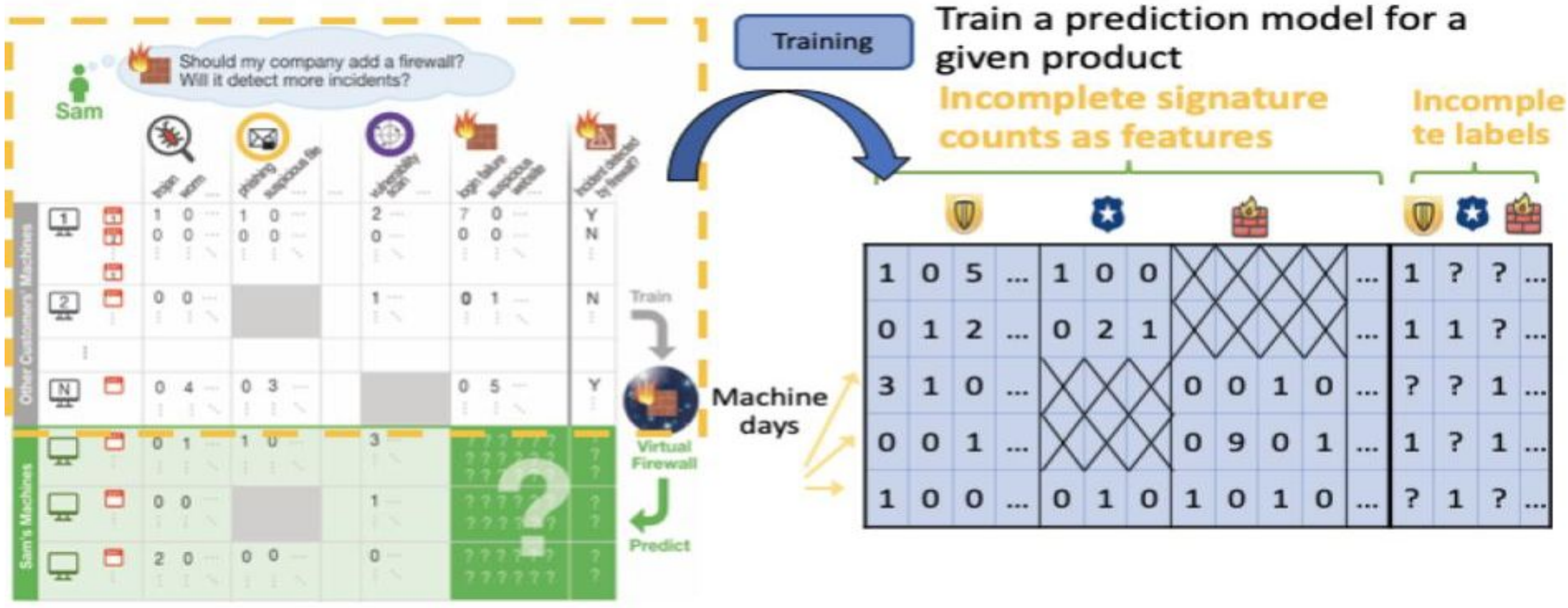
- **Highly incomplete data profiles** due to privacy control (noise corruption / missing data).
- **Privacy regulations** constrain the use of privacy-sensitive data (e.g. GDPR protocols)

Outline

- **Trustworthy Machine Learning in security-critical applications**
  - Robust security incident prediction with incomplete / noise-corrupted data
    - Multi-sourced active learning based cyber threat detection
  - Privacy-agnostic an distributed data analytics
  - Adversarial robustness certification
- **Future perspectives**

Dirty data challenge in security practices

- **Multi-sourced active learning based cyber threat prediction**

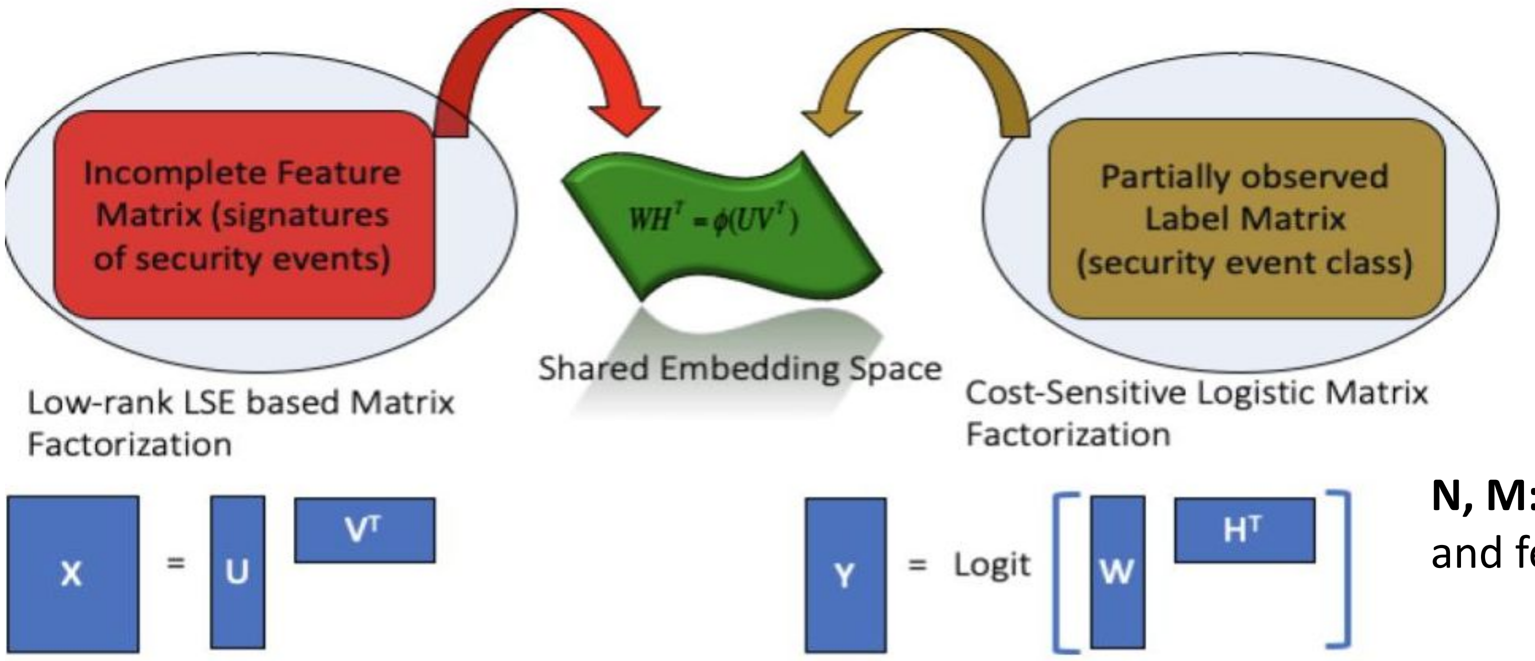


Cyber Threat Prediction: predict threats (and their types) that would be likely to be evoked based on observed incidents

A real-world learning scenario with ***incomplete features*** and ***partially observed incident labels***

Dirty data challenge in security practices

- **Multi-sourced active learning based cyber threat prediction**



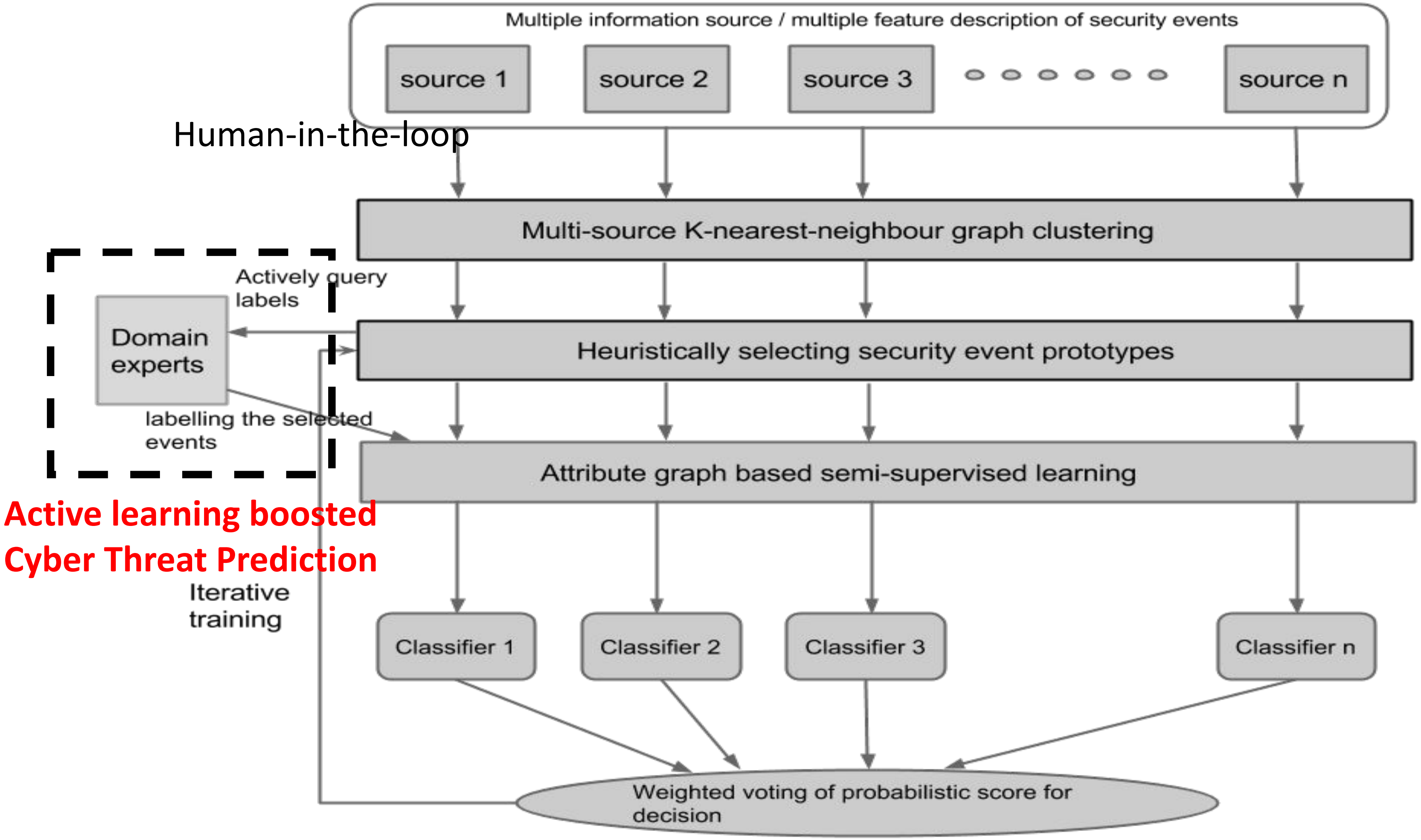
<u>Collaborative Embedding as a transfer learning solution to learning with incomplete feature and weak labels</u>

N, M: the number of training samples and feature dimensionality

$\aleph$: Maximum L2-norm of the row vectors in X

Label reconstruction error $R(Y^*) \leq \frac{C}{(1-\rho)\sqrt{NM}}\aleph$

Dirty data challenge in security practices



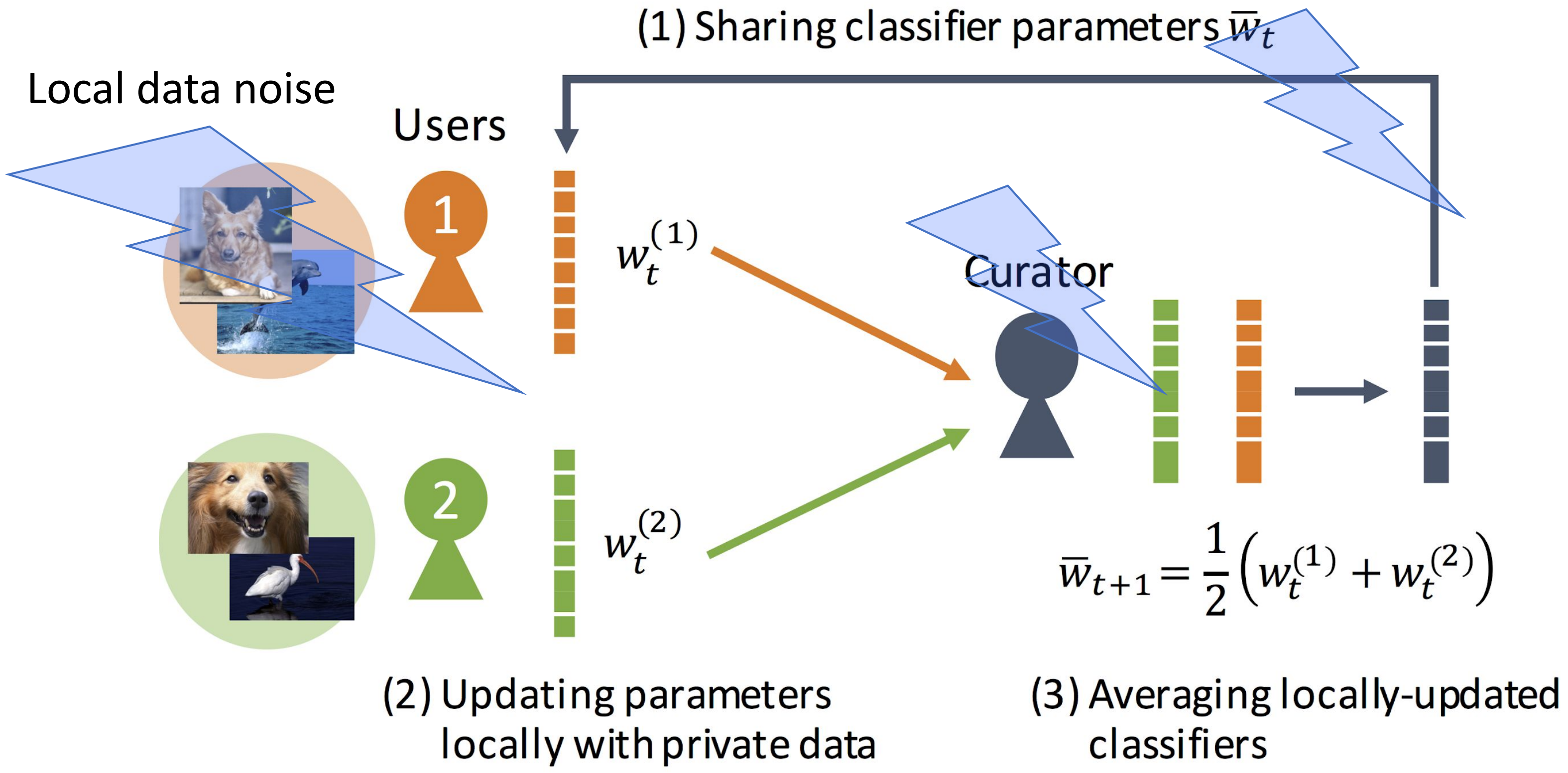
Multi-sourced active learning based cyber threat prediction

Outline

- **Trustworthy Machine Learning in security-critical applications**
  - Robust security incident prediction with incomplete / noise-corrupted data
  - Privacy-agnostic and distributed data analytics
    - Byzantine failure resilient federated learning
  - Adversarial robustness certification
- **Future perspectives**

Privacy-preserving and distributed data analytics

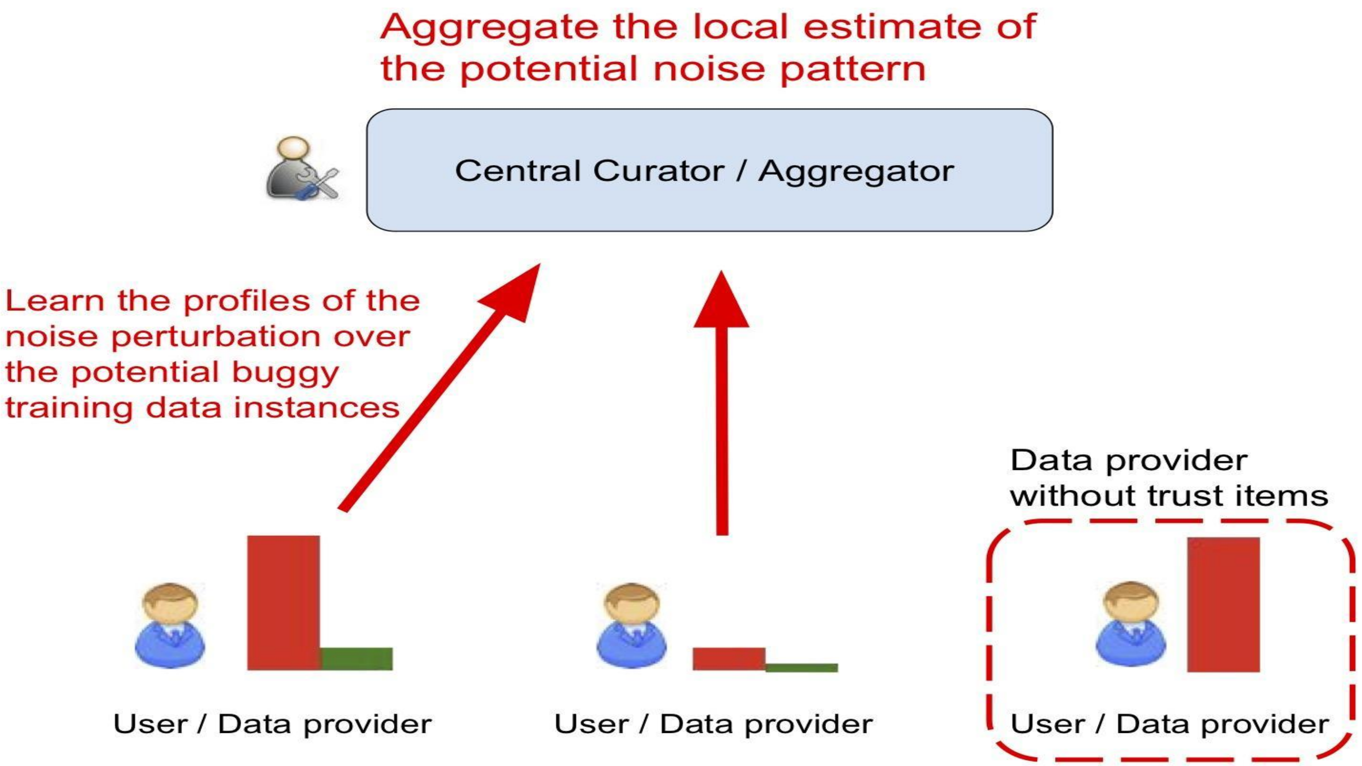
- **Byzantine Federated Learning**
  - A popular solution: Federated Learning (proposed by Google AI, published on NIPS 2016)

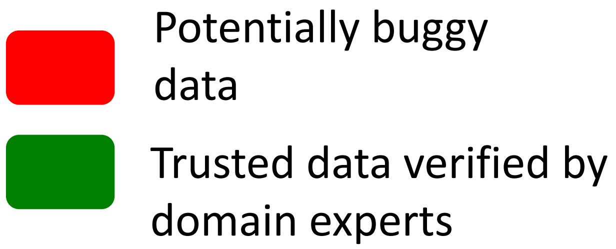


Real-world scenario:
Robust distributed ML service in compliance with Data Privacy Regulations

Privacy-preserving and distributed data analytics

- **Byzantine Federated Learning**





Privacy-preserving Collaborative Data Debugging via Trusted Items

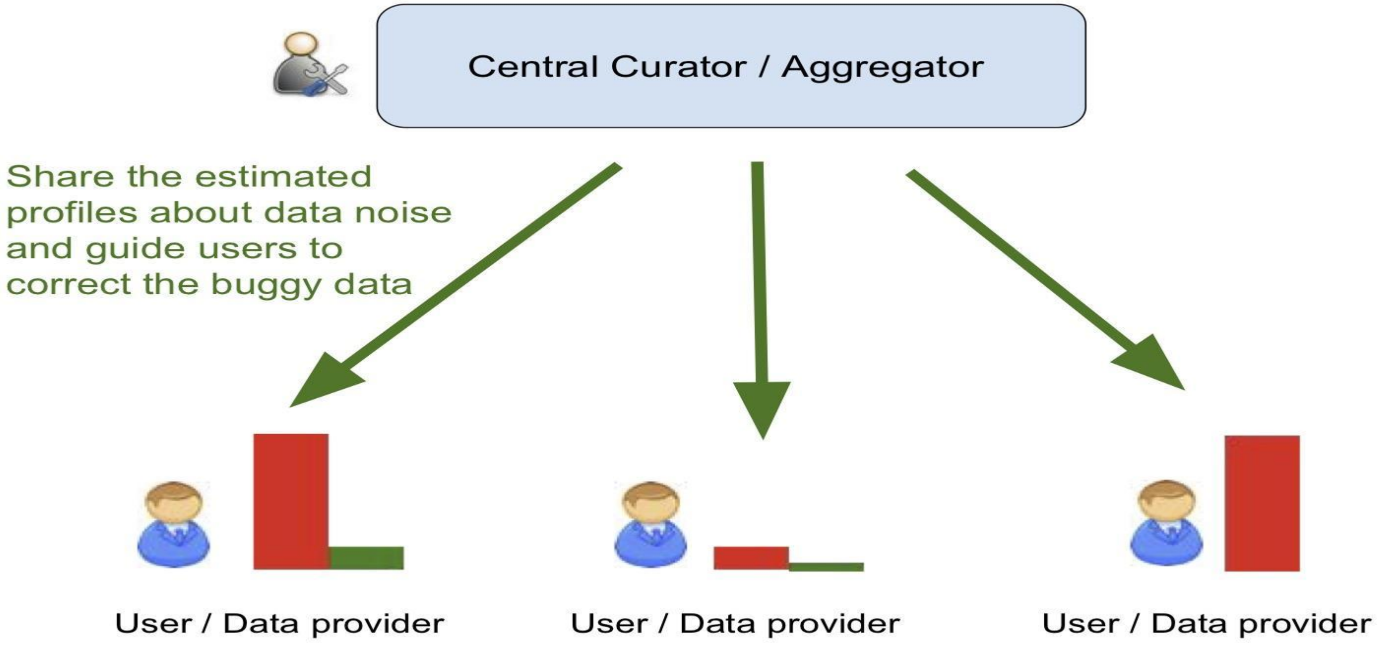
Assumption

We assume that training data hosted by each local agent is potentially buggy

We assume that a small fraction of trusted training data is available on ***some*** local agents, verified by domain experts with considerable cost and denoted as

Privacy-preserving and distributed data analytics

- **Byzantine Federated Learning**



<u>Transferred messages don't uncover local data profiles</u>

Privacy-preserving Collaborative Data Debugging via Trusted Items

Assumption

We assume that training data hosted by each local agent is potentially buggy

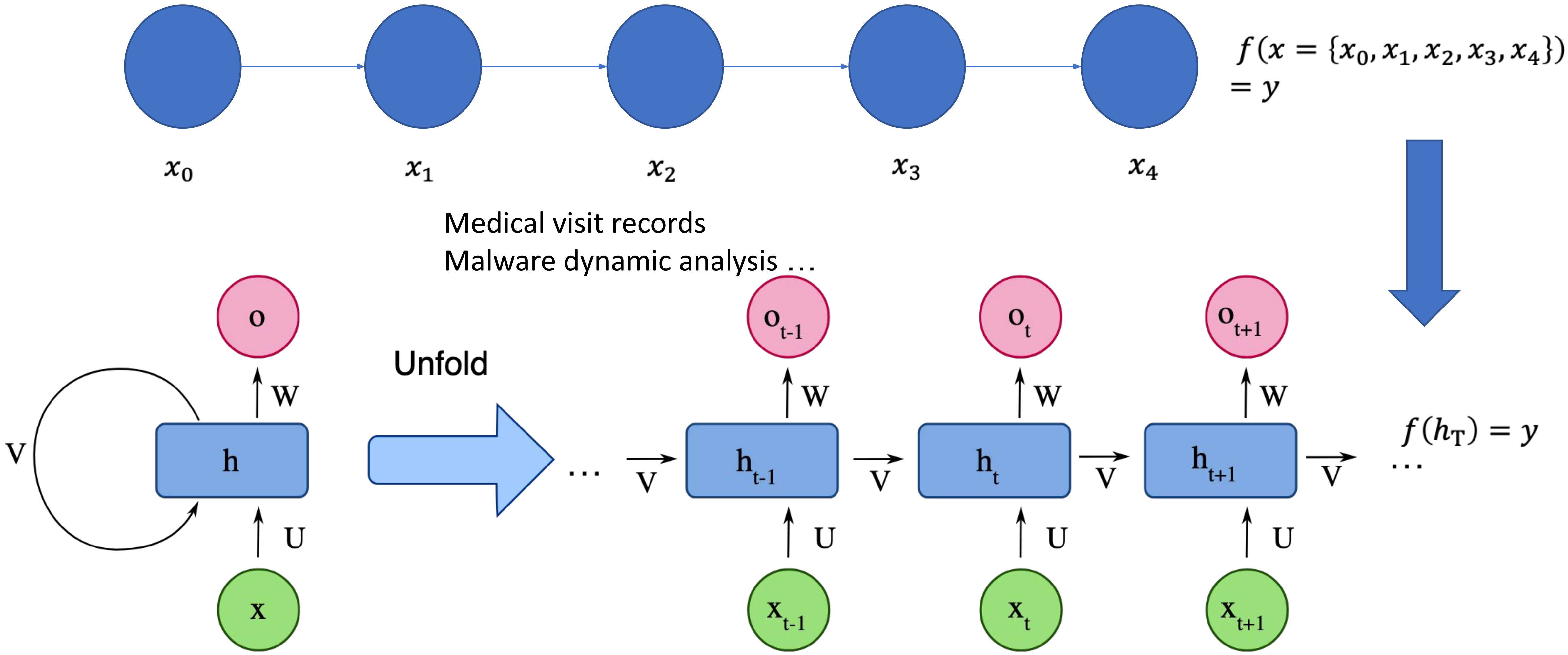
We assume that a small fraction of trusted training data is available on ***<u>some</u>*** local agents, verified by domain experts with considerable cost and denoted as

Outline

- **Trustworthy Machine Learning in security-critical applications**
  - Robust security incident prediction with incomplete / noise-corrupted data
  - Privacy-agnostic and distributed data analytics
  - Adversarial robustness certification
- **Future perspectives**

Adversarial robustness certification

Threat Model: Evasion Attack against Sequential Data Classification Model

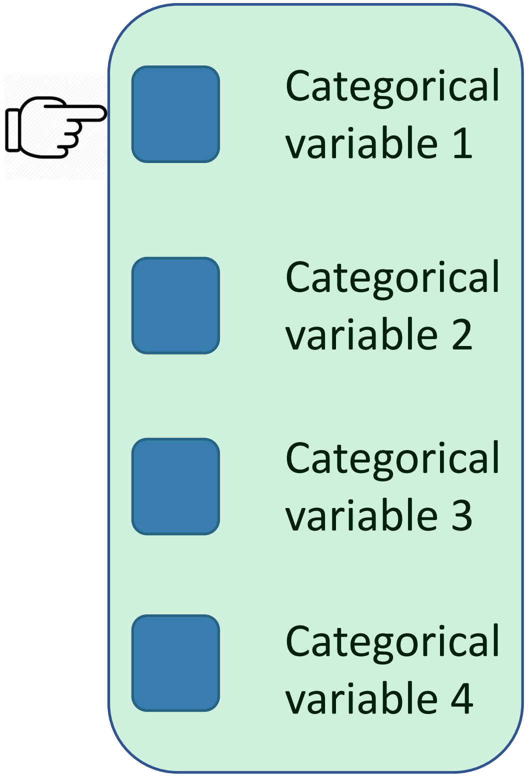


Adversarial robustness certification

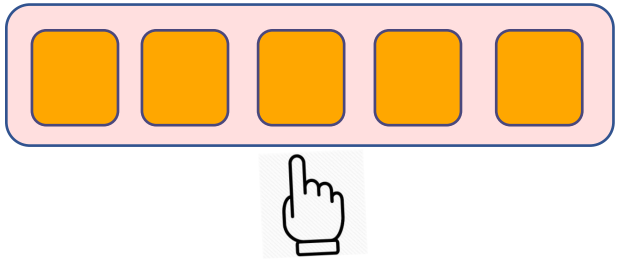
Han et al, Attackability Assessment via Weak Submodularity and Greedy Attack, KDD 2020

- Why does evasion attack on discrete data matter ?
- Attack on discrete data is a combinatorial optimization problem

Which variable to attack ?



Classification system



Categorical values of the variable 1

Which categorical value should be used for attack ?

Combinatorial optimization is NP-complete: any sort of <u>search algorithm</u> or <u>metaheuristic</u> can be used to solve them, but with at least <u>**exponential complexity**</u>. The most universally applicable approach is <u>branch-and-bound</u>

Adversarial robustness certification

Threat Model: Evasion Attack against Sequential Data Classification Model

- Set function maximization

$$S^* = \arg\max_{|S| \leq K} g(S)$$

where $$g(S) = \max_{l \subset S} f_y(\hat{\mathbf{x}}), \quad l = diff(\mathbf{b}, \hat{\mathbf{b}})$$

$|S|$ is the cardinality of set S.

$diff$ function: It reports the set of the indices where b and $\hat{b}$ are different

l denotes the set of modification to make when we attack x

$g(S)$ is a set function. The argument is a set, which includes all feasible subsets

$g(S)$ is a non-decreasing function:
If $S_i > S_{i-1},\ then\ g(S_i) > g(S_{i-1})$

Adversarial robustness certification

Greedy Search based Evasion Attack

- **Weak submodularity of the attack objective**: a bridge between **<u>Attack Quality</u>** and **<u>Regularity of the classifier</u>**

- **Claim 1: Evasion attack on discrete data targeting at a general classifier f is <u>weakly submodular</u>**

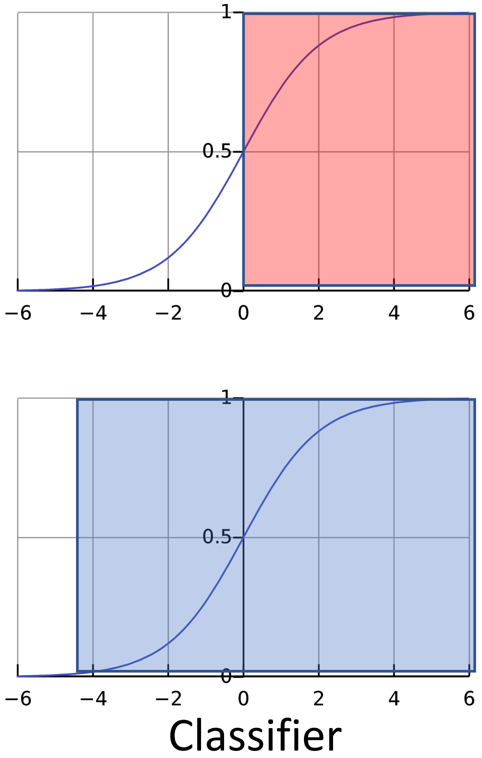
Theorem 1. *Let b as the unchanged original binary indicator defined in Eq.1. Let $\Omega_k = \{(\hat{b}, \hat{b}') : |diff(b, \hat{b})| \leq k, |diff(b, \hat{b}')| \leq k, |diff(\hat{b}, \hat{b}')| \leq k\}$, where $\hat{b}$ and $\hat{b}'$ denote two sets of selected discrete attributes to be modified adversarially. If the classifier f_y is $(m_{\Omega_k}, M_{\Omega_k})$-regularized on Ω_k, the $g(S)$ defined by Eq.1 is weakly submodular. Its submodularity ratio γ_k on Ω_k is bounded from below:*

$$\gamma_k \geq \frac{1}{2\psi_k M_{\Omega_k}}$$

$$\psi_k = 1 + \frac{k^2 |m_{\Omega_k}|}{2\|\nabla f_y(b)_s\|_2^2}, \quad \text{If } m_{\Omega_k} \leq 0 \tag{4}$$

$$\psi_k = \frac{1}{2m_{\Omega_k}}, \quad \text{If } m_{\Omega_k} > 0$$

where $\nabla f_y(b)_v$ denotes the elements of $\nabla f_y(b)$ corresponding to the difference between the index sets l_b and $l_{b'}$, where $v = l_{b'} \setminus l_b + l_b \setminus l_{b'}$.



Concave case

Submodular attack objective

Non-concave case

Weakly submodular attack objective: $0 < \gamma < 1$

Adversarial robustness certification

Greedy Search based Evasion Attack

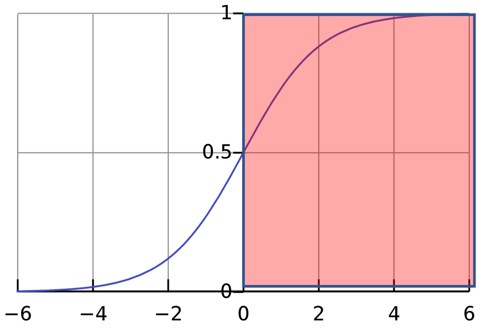
- **Weak submodularity of the attack objective**: a bridge between **<u>Attack Quality</u>** and **<u>Regularity of the classifier</u>**
  - **Claim 2: Attack with a weakly submodular objective can be solved with greedy search. The quality of the solution can be bounded theoretically in a similar way as in the submodular case – in plein English, *<u>weakly submodular attack objective is attackable.</u>***

THEOREM 2. *[**Theorem 3** in [10]] Let the evasion attack problem defined by Eq.(1) be with the classification function f_y that is $(m_{\Omega_k}, M_{\Omega_k})$-bounded. Let S_k be the set of the values selected by FSGS and S_k^* be the underlying optimal value set following the support size constraint. The corresponding attack objective values reached by S_k and S_k^* are g^{FSGS} and g^{OPT}, respectively. Then g^{FSGS} is bounded:*

$$g^{FSGS} \geq (1 - e^{-\gamma_{S_k}})g^{OPT} \tag{5}$$

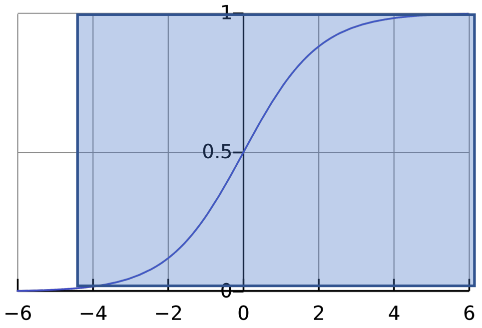
where γ_{S_k} is the submodularity ratio of $g(S)$ defined on the selected set S_k. Especially, if $g(S)$ is submodular, the lower bound gives as:

$$g^{FSGS} \geq (1 - e^{-1})g^{OPT} \tag{6}$$



Concave case

Submodular attack objective

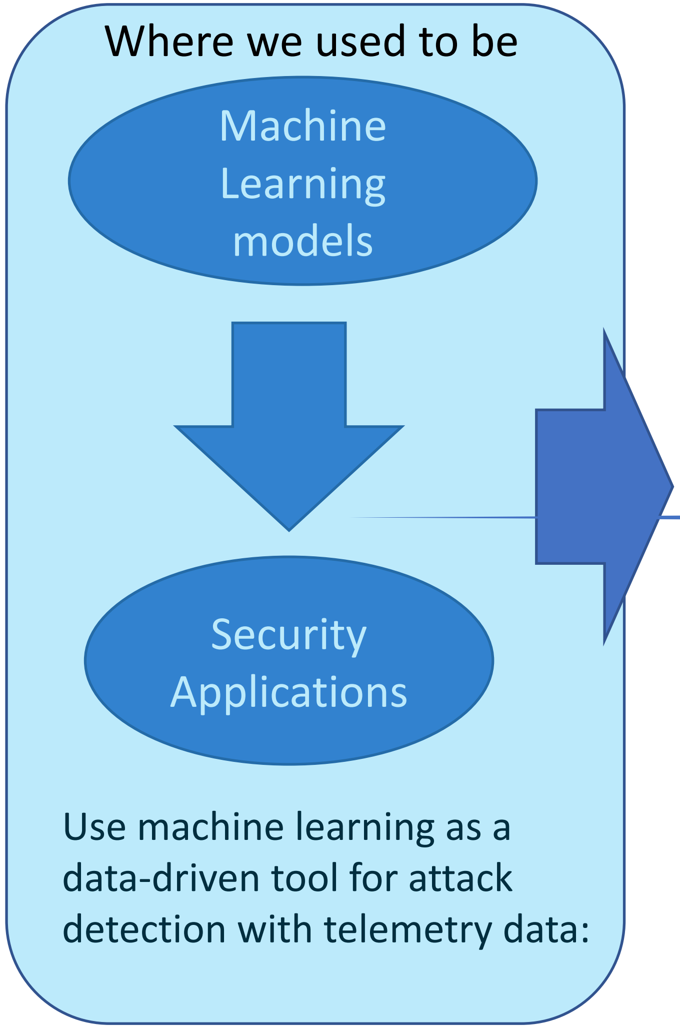


Non-concave case

Weakly submodular attack objective: $0 < \gamma < 1$

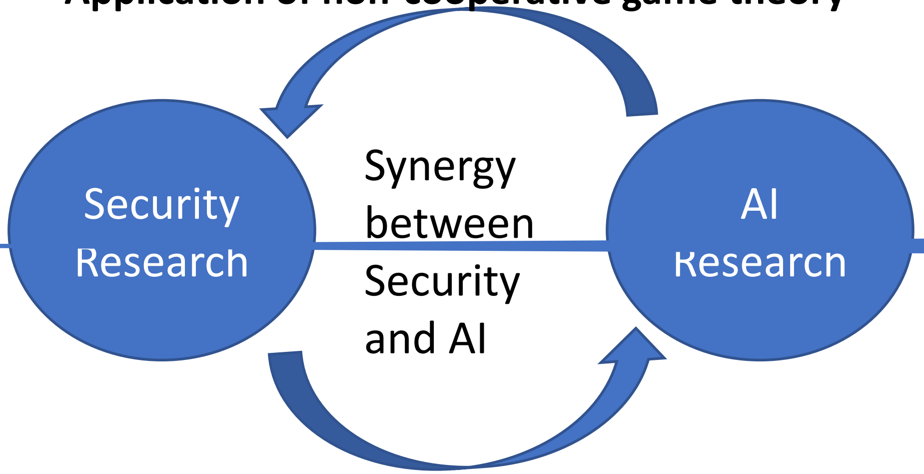
Attackable but with lower worst case quality bound

Future Perspectives: Proactive defense with proactive AI



Autonomous Cyber Defense

- **Zero-day attack prediction and risk assessment**
- **Understanding attack incidents** (network intrusion / malwares)
- **Human-in-the-loop defense planning**, especially in the safety-sensitive scenarios
- **Application of non-cooperative game theory**



Secured and Trustworthy AI Decision Making

- **Robust to intentional or natural data noise** (error-correction enhanced learning)
- **Data-privacy preserving analytics**
  - Mitigate data privacy risk leakage while enjoy the benefits of AI systems

AI in Security

AI Boosted Attack Prediction and Comprehension

Security for AI

Trusted AI for security and privacy-sensitive data analytics

Thanks for your attention