Secure Learning in Physical Adversarial Environments

Bo Li
Security Group, UC Berkeley
FORCES, 2017

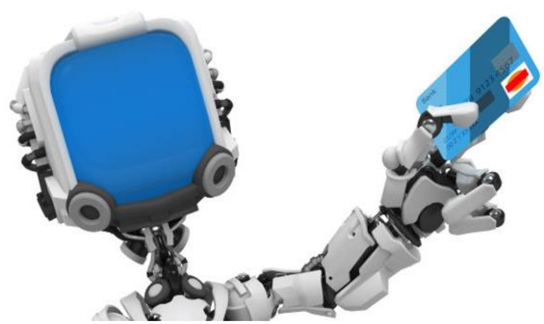
Machine Learning in Physical World

Autonomous Driving

Healthcare

Smart City

Malware Classification



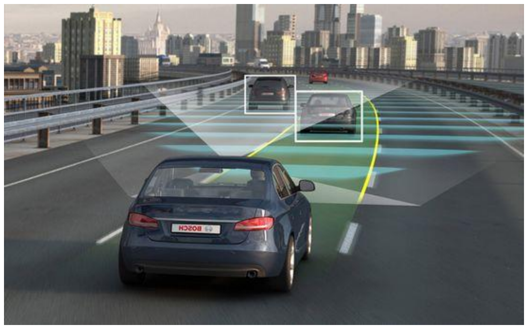
Fraud Detection

Biometrics Recognition

While cybersecurity R&D needs are addressed in greater detail in the NITRD Cybersecurity R&D Strategic Plan, some cybersecurity risks are specific to AI systems. ***One key research area is “adversarial machine learning”****, that explores the degree to which AI systems can be compromised by “contaminating” training data, by modifying algorithms, or by making subtle changes to an object that prevent it from being correctly identified….*

- National Science and Technology Council 2016

Autonomous Driving is the Trend…



However, What We Can See Everyday…

Adversarial Examples in Physical World

Subtle Perturbations

=

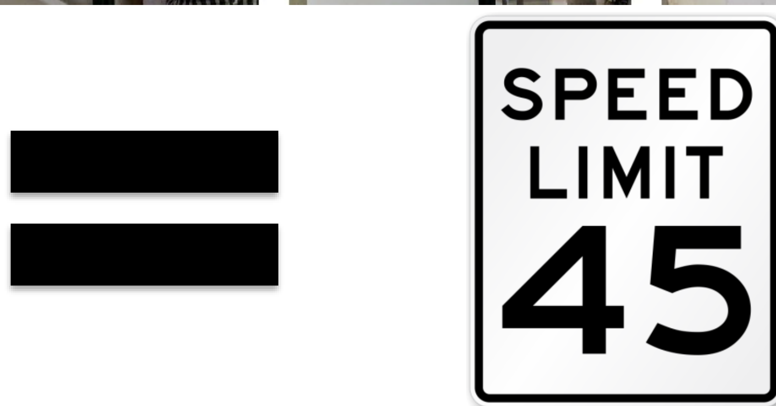
Evtimov, Ivan, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. “Robust Physical-World Attacks on Machine Learning Models.” *arXiv preprint arXiv:1707.08945* (2017).

Adversarial Examples in Physical World

Camouflage Perturbations

=

Camouflage Perturbations



Adversarial Examples in Physical World

Subtle Perturbations

=

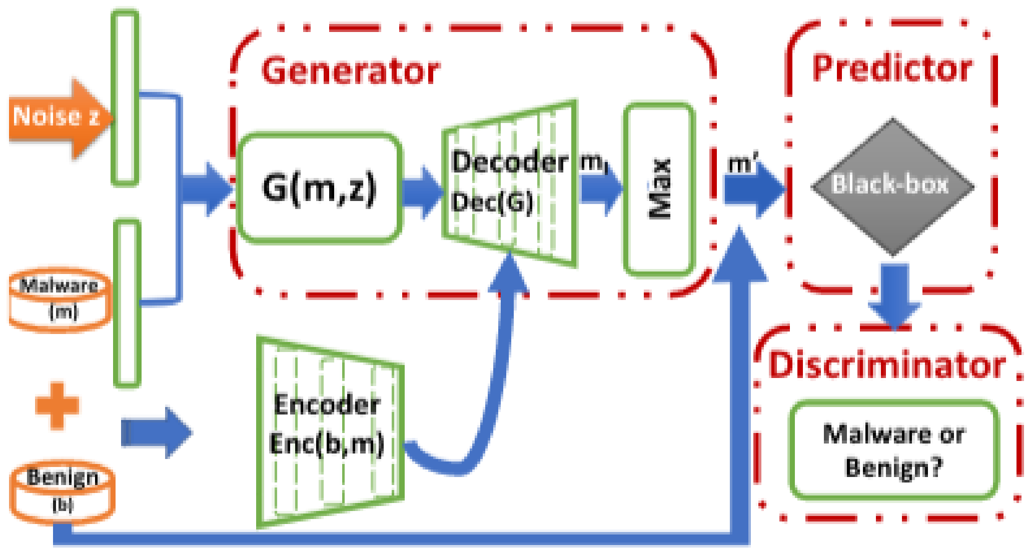
Adversarial Examples in Physical World

Adversarial perturbations are possible in physical world under different conditions and viewpoints, including the distances and angles.

Deep loss function:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i+\delta), y)$$

Malware Evasion Attacks Based on Generative Adversarial Networks



Challenges:

1. Keep the original malicious functionalities for malwares
2. Generate evasion instances in the discrete feature space
3. Evasion attack against black-box classifiers

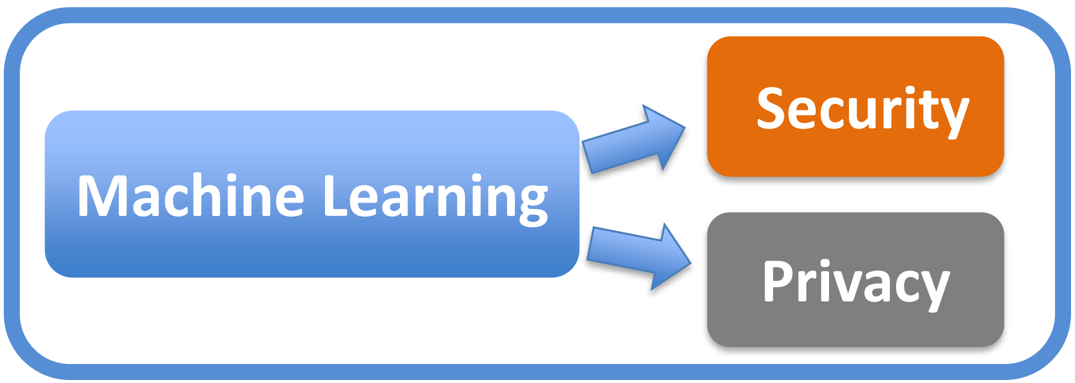
Malware Evasion Attacks Based on Generative Adversarial Networks

Method	FNR	FPR	Accuracy	Run time
LR	0.42	0.00	0.78	–
DT	0.03	0.00	0.98	–
DNN	0.42	0.00	0.78	–
NB	0.01	0.33	0.83	–
RF	0.02	0.00	0.99	–
KNN	0.03	0.00	0.98	–
SVM	0.04	0.00	0.98	–
EvaGAN	1.00	0.00	0.50	0.08s
RANDOM	0.45	0.00	0.78	0.03s
C&W	0.93	0.00	0.71	2.19s
FGM	0.75	0.00	0.62	2.45s
EvadeML	0.82	0.00	0.59	>12h

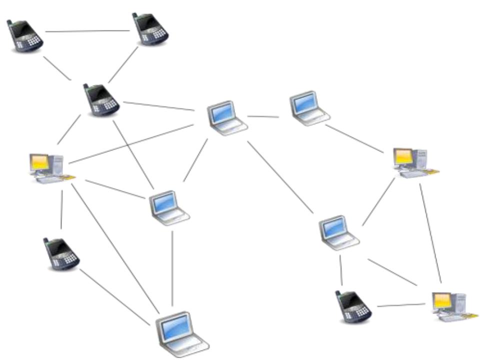
Automatically generating malware evasion instances against black-box classifiers based on GANs is more efficient than traditional attack methods

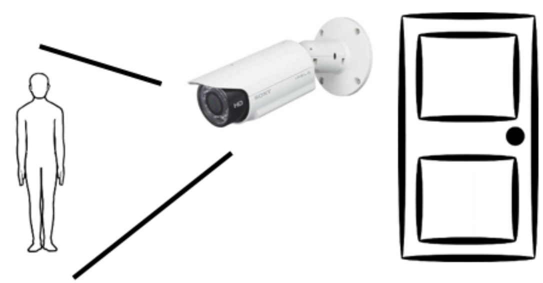
Summary


Most Vulnerable Smart Cities
Cyber Attack on Internet of Things (IoT)


Machine Learning
Security
Privacy

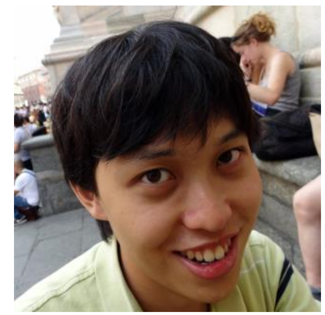
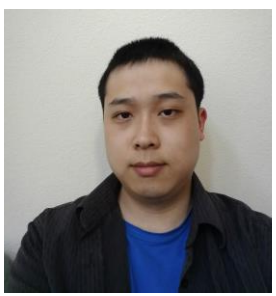

MEDICAL RECORD



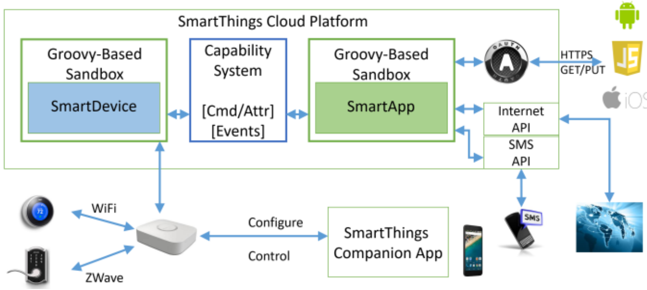


Group Members

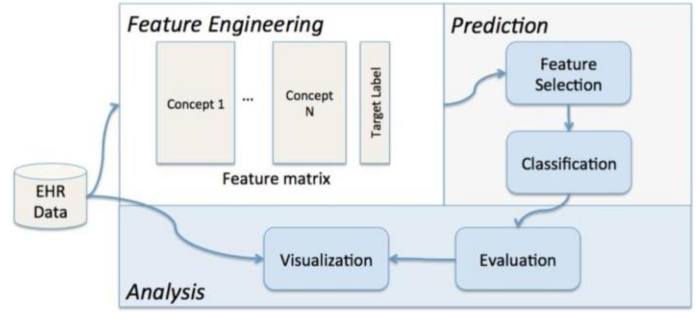


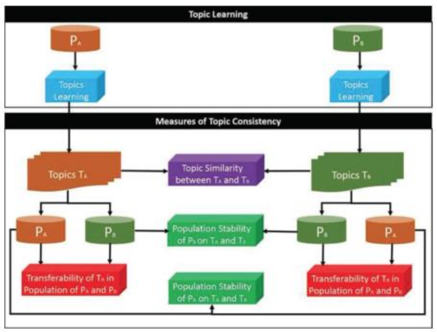




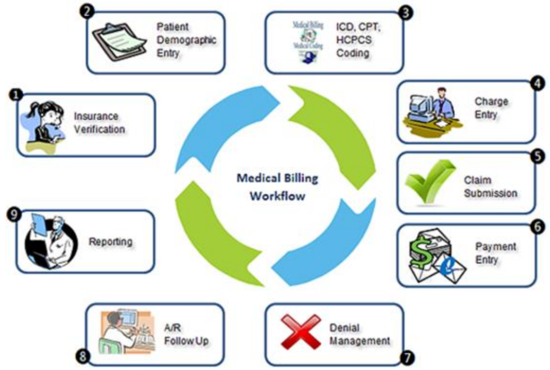
Robust Smart Home



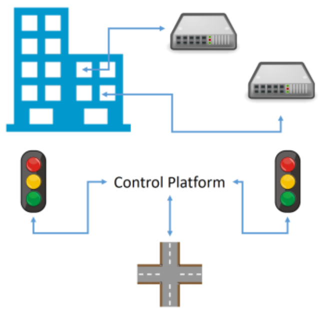
Privacy-Preserving Data Analysis



Topic of Workflow Analysis

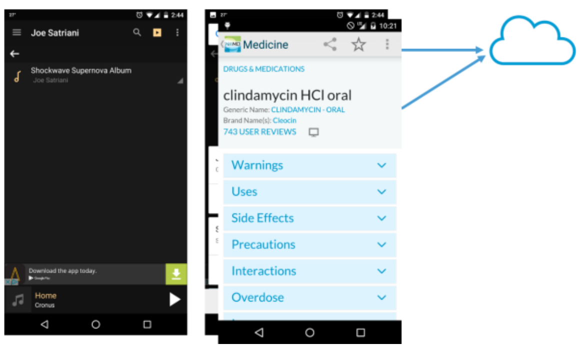


Game Theoretic Auditing System for EMR

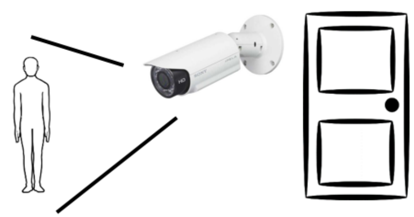


Large-Scale Auditing Game With Human In the Loop

Robust Learning



Privacy Protected Mobile Healthcare



Robust Face Recognition Against Poisoning Attack

Berkeley
UNIVERSITY OF CALIFORNIA

Thank You!
Bo Li
crystalboli@berkeley.edu

http://www.crystal-boli.com/