Towards Resilient Machine Learning in Adversarial Environments

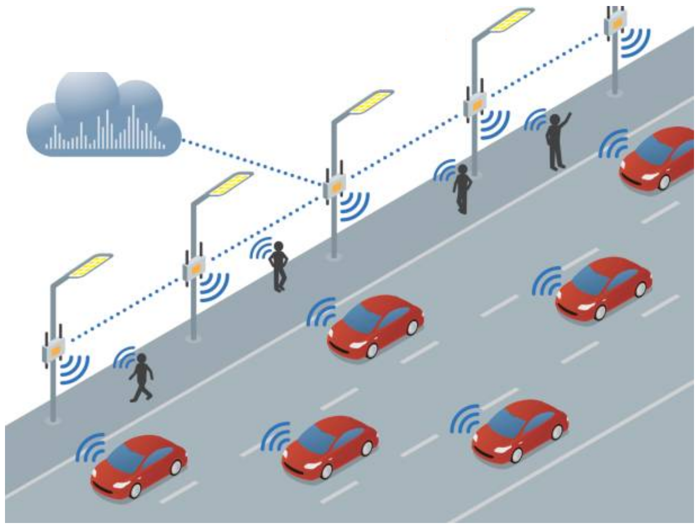
Alina Oprea
Associate Professor
Northeastern University

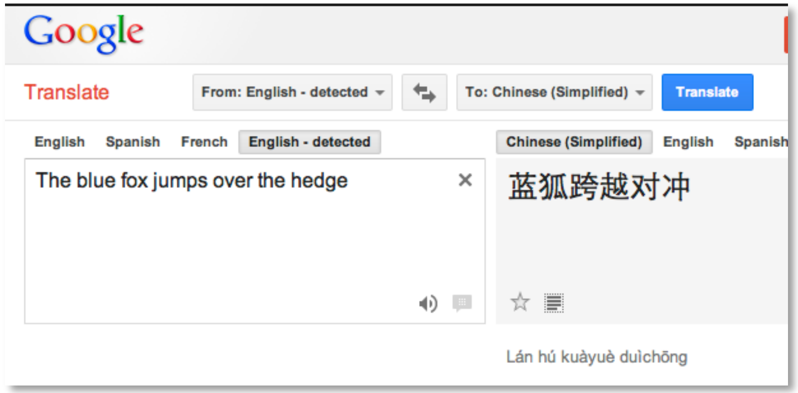
February 12, 2020

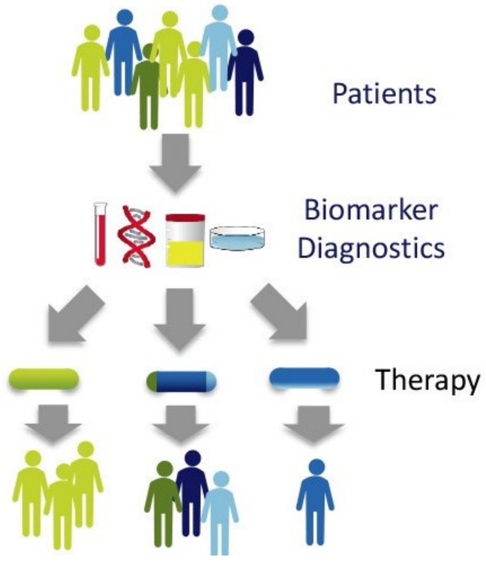
Network and Distributed System Security (NDS2) Lab

- Machine learning and AI for cybersecurity
  - Threat detection [Yen et al. 13], [Yen et al. 14], [Oprea et al. 15], [Li and Oprea 16], [Buyukkayhan et al. 17], [Oprea et al. 18], [Duan et al. 18], [Ongun et al. 19]
  - IoT security: [Ongun et al. 19]
  - Web security: [Jana and Oprea 19]
  - AI for cyber security games: [Oakley and Oprea 19]
- Adversarial machine learning and AI
  - Poisoning attacks and defenses [Liu et al. 17], [Jagielski et al. 18], [Jagielski et al. 19]
  - Attack transferability [Demontis et al. 19]
  - Evasion attacks for cyber security and connected cars [Chernikova et al. 19], [Chernikova and Oprea 19]
  - Fairness and Privacy [Jagielski et al. 19]

AI is Everywhere



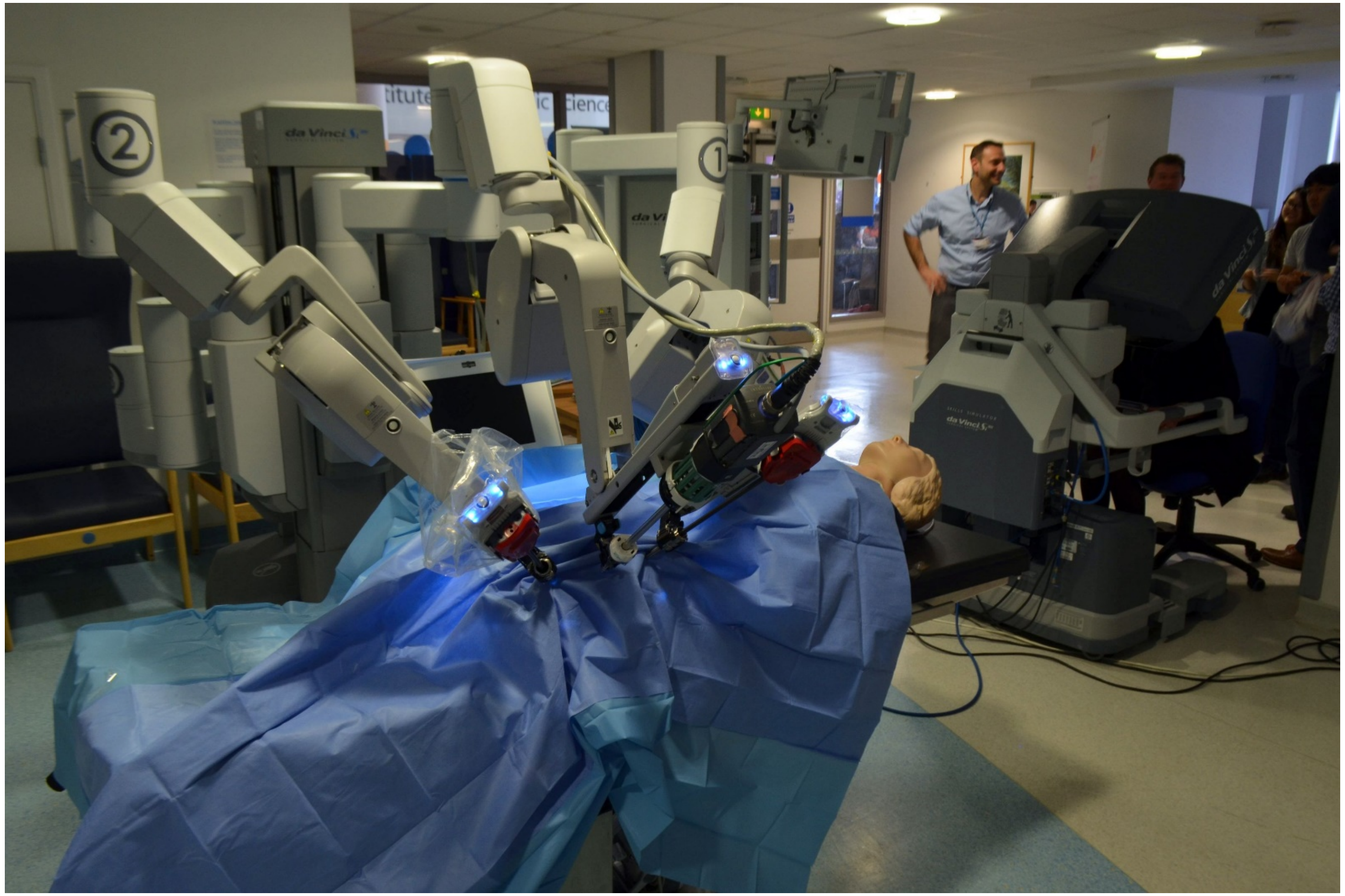




Fast Forward in the Near Future

AI Transportation in Cities of the Future (10-20 years)

Fast Forward in the Near Future



AI Robots in Medicine of the Future (10-20 years)

Implications for Cyber Security

- AI has potential in security applications
  - Complement traditional defenses
  - Design intelligent and adaptive defense algorithms

- ...But AI becomes a target of attack
  - Deep Neural Networks are not resilient to adversarial manipulations
    - [Szegedy et al. 13]: "Intriguing properties of neural networks"
  - Many critical real-world applications are vulnerable
  - New adversarially-resilient algorithms are needed!

Supervised Learning: Classification

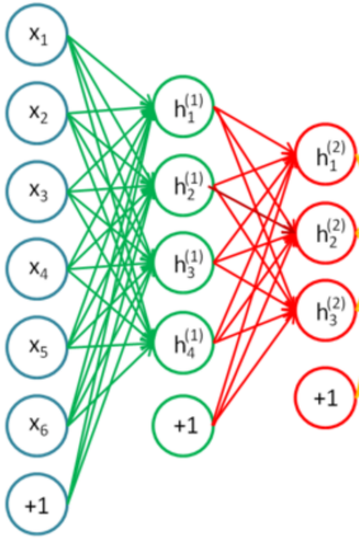
Training

Data → Feature extraction → ML model

Labeled

$x_i, y_i \in \{\text{Positive}, \text{Negative}\}$

$f(x)$



Testing

New data → Predictions

Unlabeled

x'

$y' = f(x')$

Positive

Negative

Supervised Learning: Regression

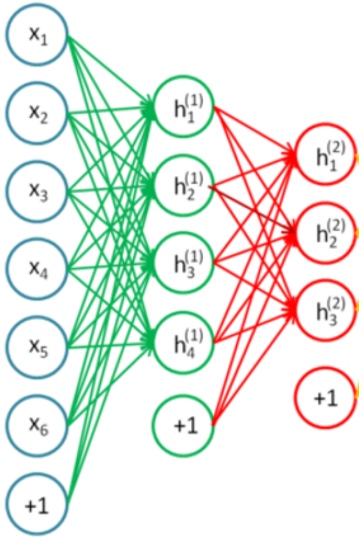
Training

Data → Feature extraction → ML model

Labeled

$x_i, y_i \in R$

$f(x)$



Testing

New data → Predictions

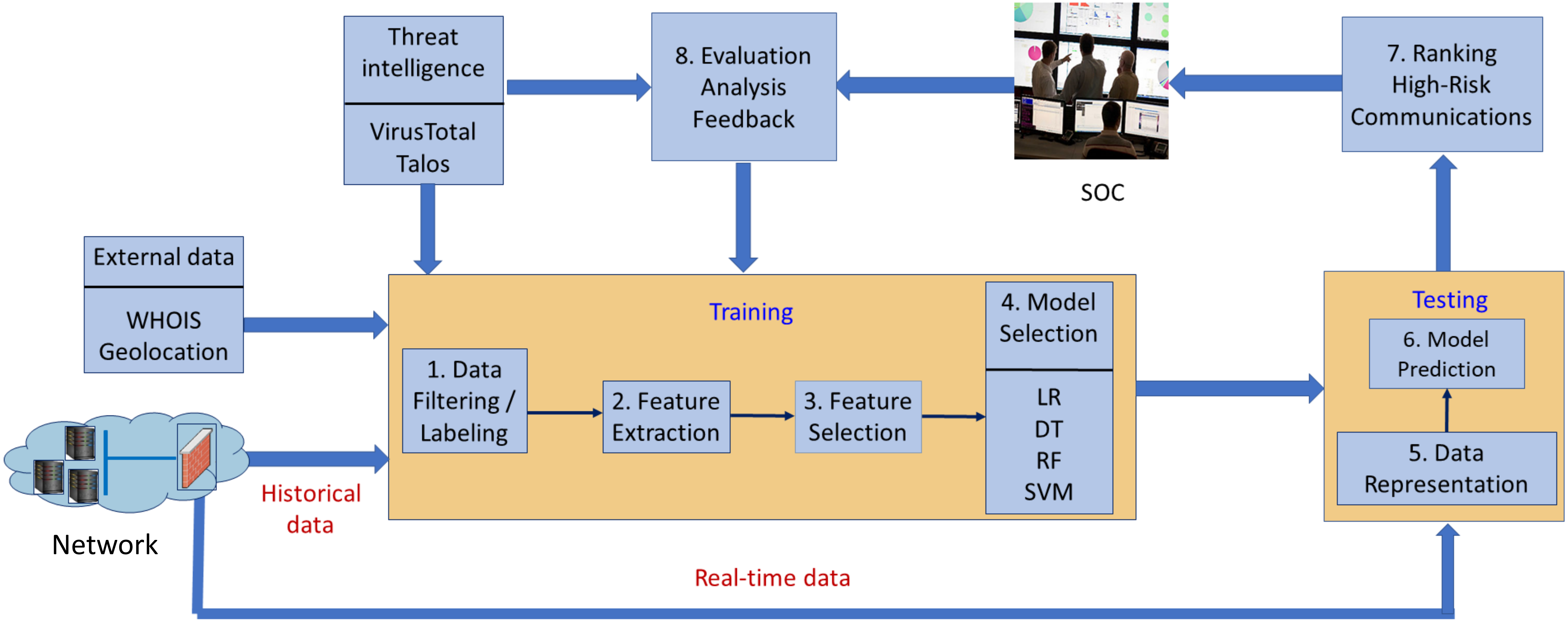
$y' = f(x') \in R$

Unlabeled

x'

Numerical value

MADE: Detecting Malicious Web Domains



A. Oprea, Z. Li, R. Norris, K. Bowers.
MADE: Security Analytics for Enterprise Threat Detection. In ACSAC 2018.

Adversarial Machine Learning: Taxonomy

Attacker's Objective

Learning stage

	Targeted Target small set of points	**Availability** Target majority of points	**Privacy** Learn sensitive information
Training	Targeted Poisoning Backdoor Trojan Attacks	Poisoning Availability Model Poisoning	-
Testing	Evasion Attacks Adversarial Examples	-	Membership Inference Model Extraction

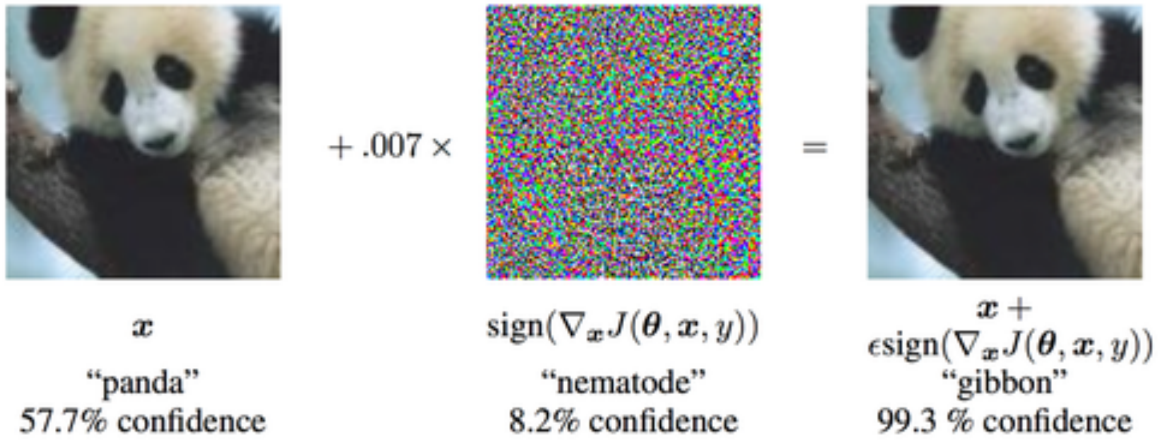
Adversarial Machine Learning: Taxonomy

Attacker's Objective

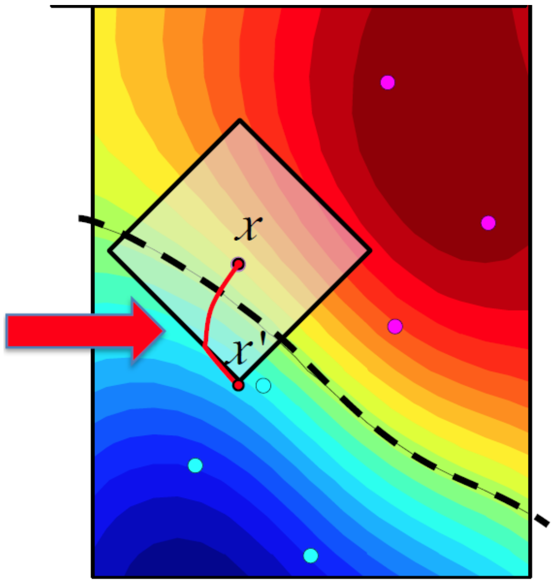
Learning stage

	Targeted Target small set of points	**Availability** Target majority of points	**Privacy** Learn sensitive information
Training	Targeted Poisoning Backdoor Trojan Attacks	Poisoning Availability Model Poisoning	-
Testing	Evasion Attacks Adversarial Examples	-	Membership Inference Model Extraction

Evasion Attacks

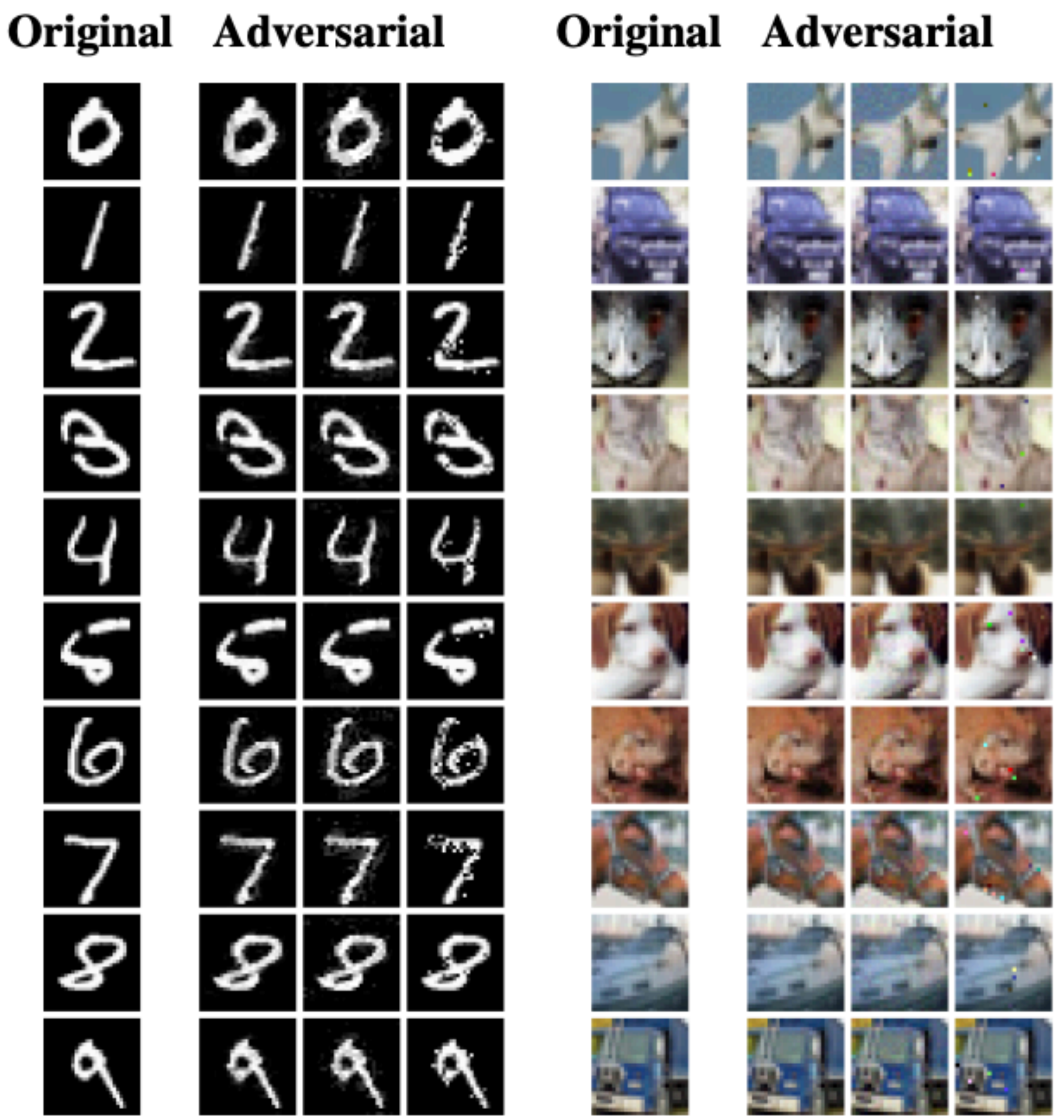


Adversarial example



- Evasion attack: attack against ML at testing time
  - [Szegedy et al. 13], [Biggio et al. 13], [Goodfellow et al. 14], [Carlini and Wagner 17], [Madry et al. 17], [Athalye et al. 18], …
- Implications
  - Small (imperceptible) modification at testing time can change the classification of any data point to any targeted class

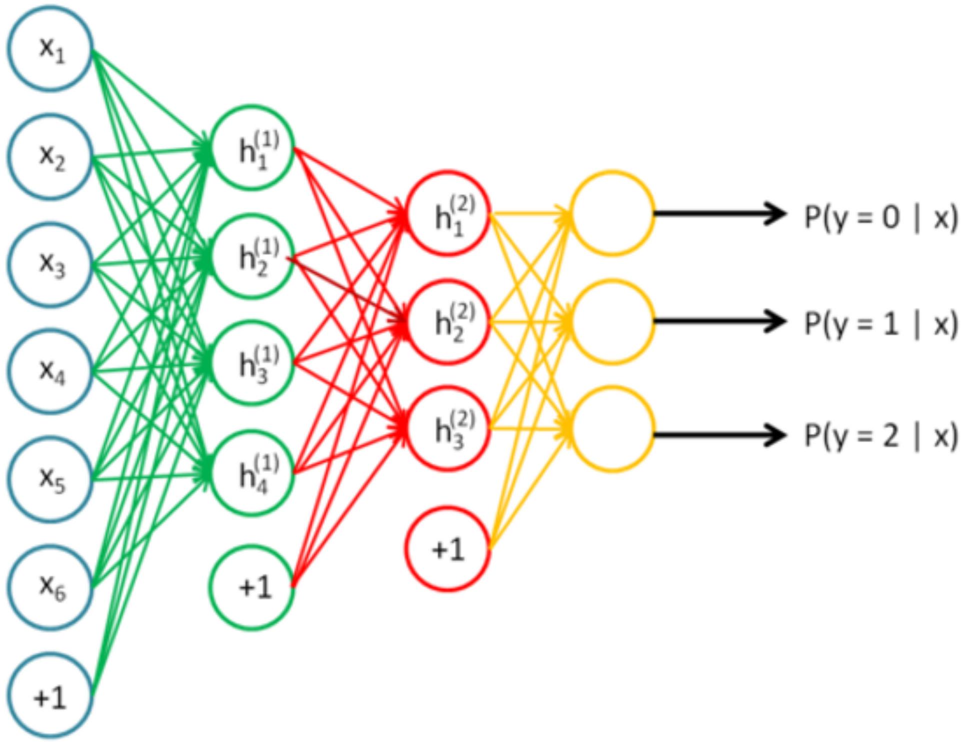
Adversarial Examples



- N. Carlini and D. Wagner. *Towards Evaluating the Robustness of Neural Networks*. In IEEE Security and Privacy Symposium 2017
- Goal: create minimum perturbations for adversarial examples
- They always exist!
- Application domains: image recognition, videos classification, text models, speech recognition

Evasion Attacks For Neural Networks

Input: Images represented as feature vectors



Optimization Formulation

Given input x

Find adversarial example

$$x' = x + \delta$$

$$\min_{\delta} c||\delta||_2^2 + L_t(x + \delta)$$

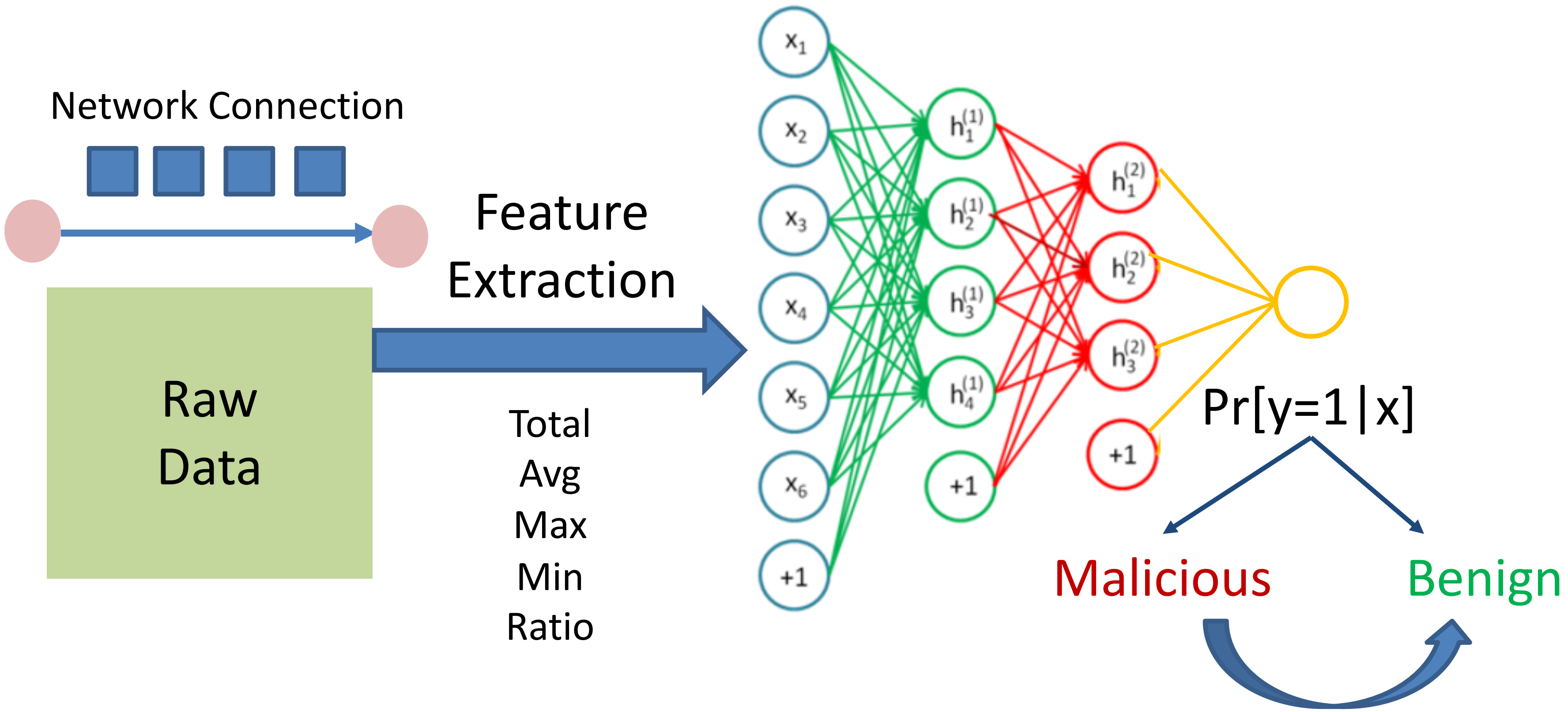
Min distance

Change class

[CW 17], [Madry et al.18]

- Most existing attacks are in continuous domains
- Images represented as matrix of pixels with continuous values
- Optimization problem solved with gradient descent

Evasion Attacks for Security



Challenge

- Attacks for continuous domains do not result in feasible adversarial examples

Solution

- New framework for evasion attacks taking into account feature constraints
- Iterative modification guided by gradient values

Evasion Attack for Malicious Connection Classifier

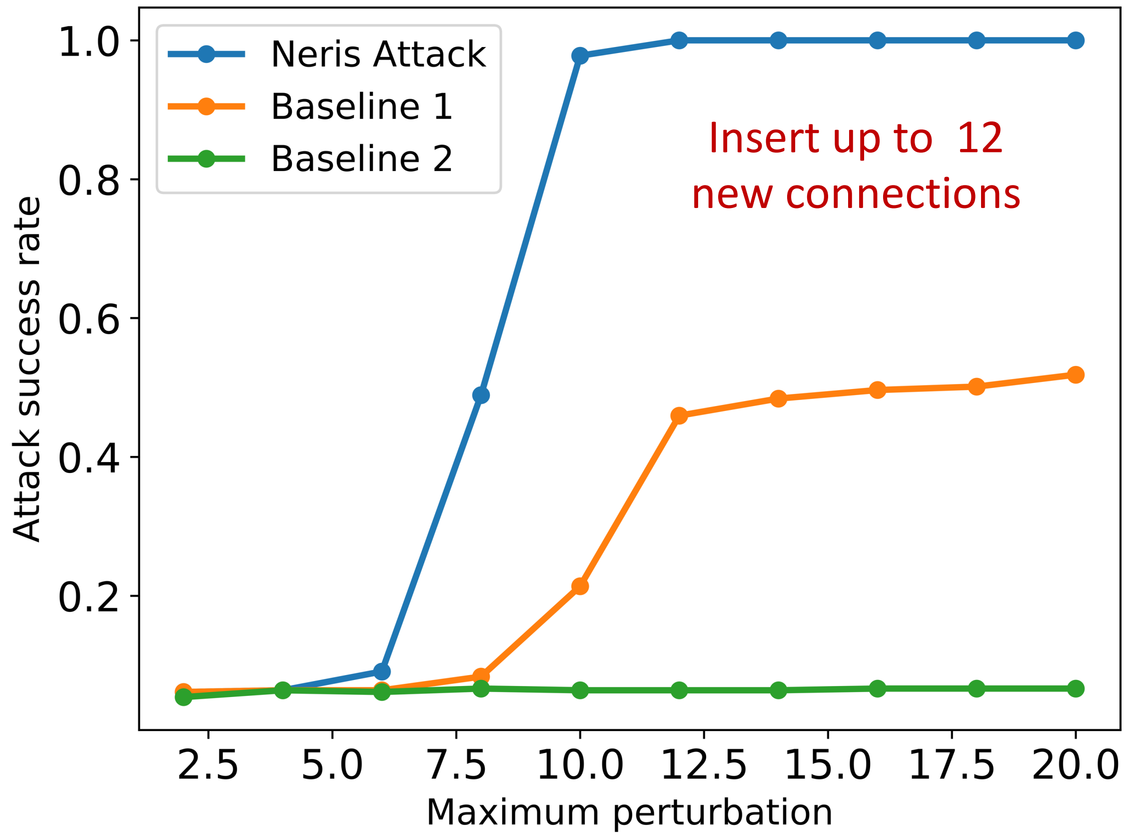
Raw network logs

Time	Src IP	Dst IP	Prot.	Port	Sent bytes	Recv. bytes	Sent packets	Recv. packets	Duration
9:00:00	147.32.84.59	77.75.72.57	TCP	80	1065	5817	10	11	5.37
9:00:05	147.32.84.59	87.240.134.159	TCP	80	950	340	7	5	25.25
9:00:12	147.32.84.59	77.75.77.9	TCP	80	1256	422	5	5	0.0048
9:00:20	147.32.84.165	209.85.148.147	TCP	443	112404	0	87	0	432

- Goal: Distinguish malicious and benign network connections
- Features: Aggregated statistical features per port
- Attack: Insert TCP or UDP connections on the determined port
- Physical constraints on network
  - Max packet size, latency, protocol accepted per port

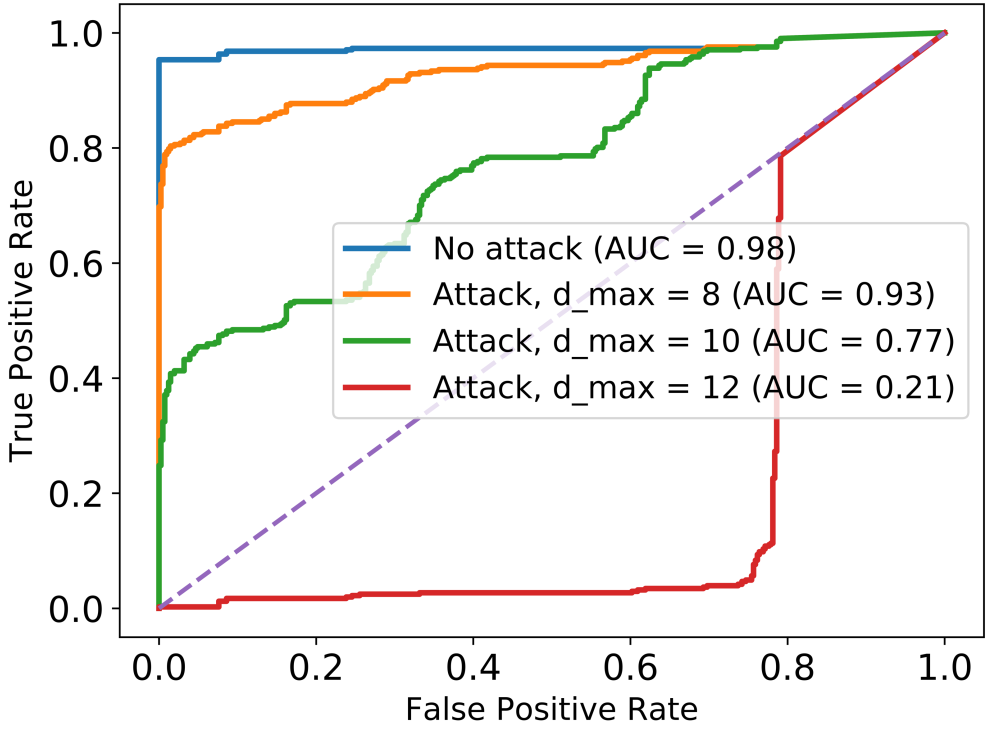
How Effective are Evasion Attacks in Security?

- Dataset: CTU-13, Neris botnet
  - 194K benign, 3869 malicious
- Features: 756 on 17 ports
- Model: Feed-forward neural network (3 layers), F1: 0.96

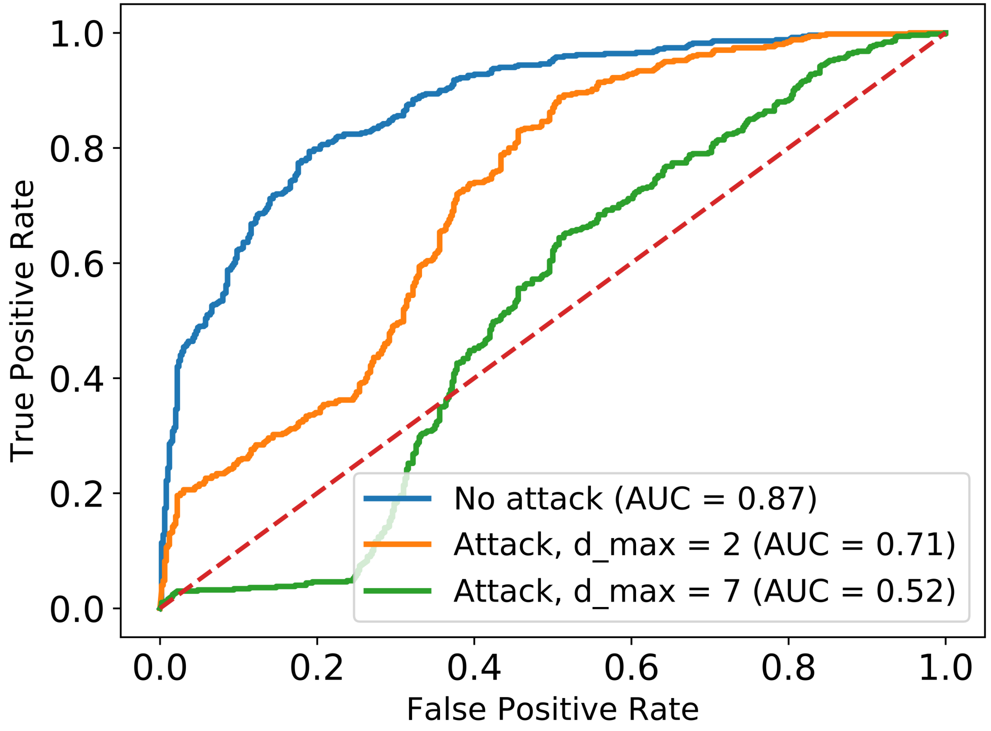


A. Chernikova and A. Oprea. *Adversarial Examples for Deep-Learning Cyber Security Analytics.* http://arxiv.org/abs/1909.10480, 2019.

How Effective are Evasion Attacks in Security?



Malicious connection classifier

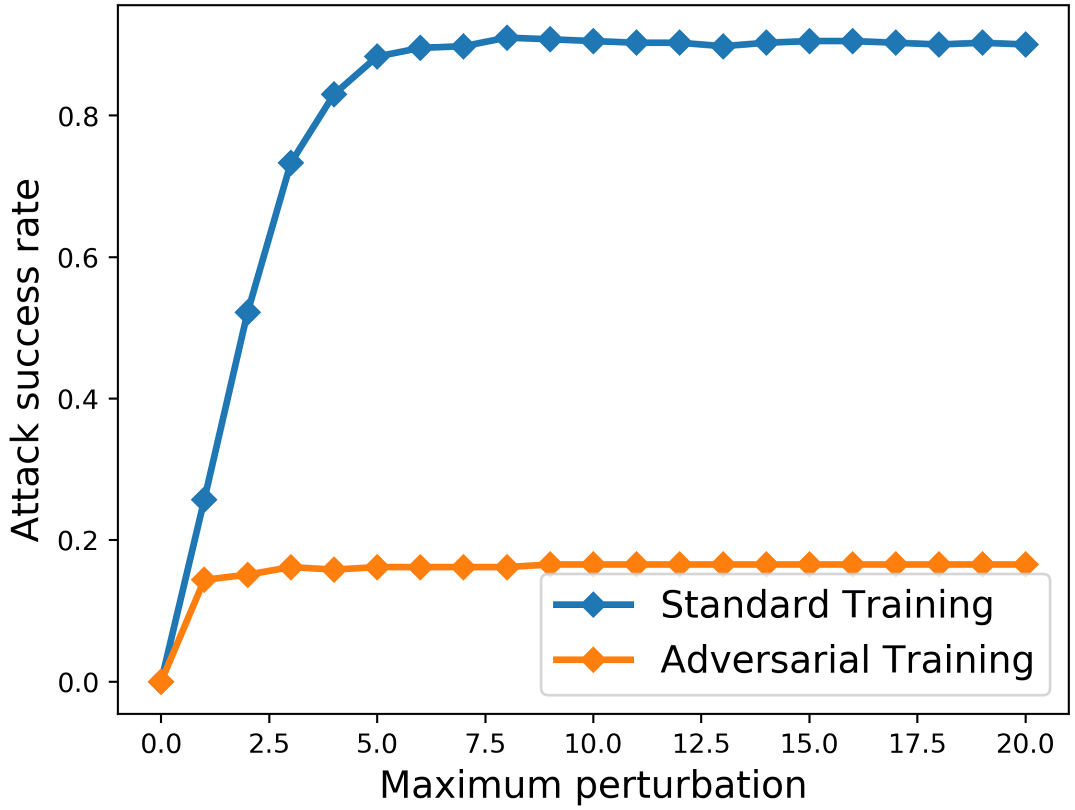


Malicious domain classifier

- Significant degradation of ML classifiers in security
- Small amount of perturbation is effective
- General framework for adversarial testing in discrete domains

Increasing Robustness of ML in Security

- Adversarial re-training
  - Train model iteratively
  - In each iteration, generate adversarial examples and add to training
- Implications
  - Adversarial training can improve robustness of ML model



Evasion Attacks in Connected Cars

- Udacity challenge: Predict steering angle from camera images, 2014
- Actions
  - Turn left (negative steering angle)
  - Turn right (positive steering angle)
  - Straight (steering angle in [-T,T])
- Dataset has 33,608 images and steering angle values (70GB of data)

Predict direction: Straight, Left, Right

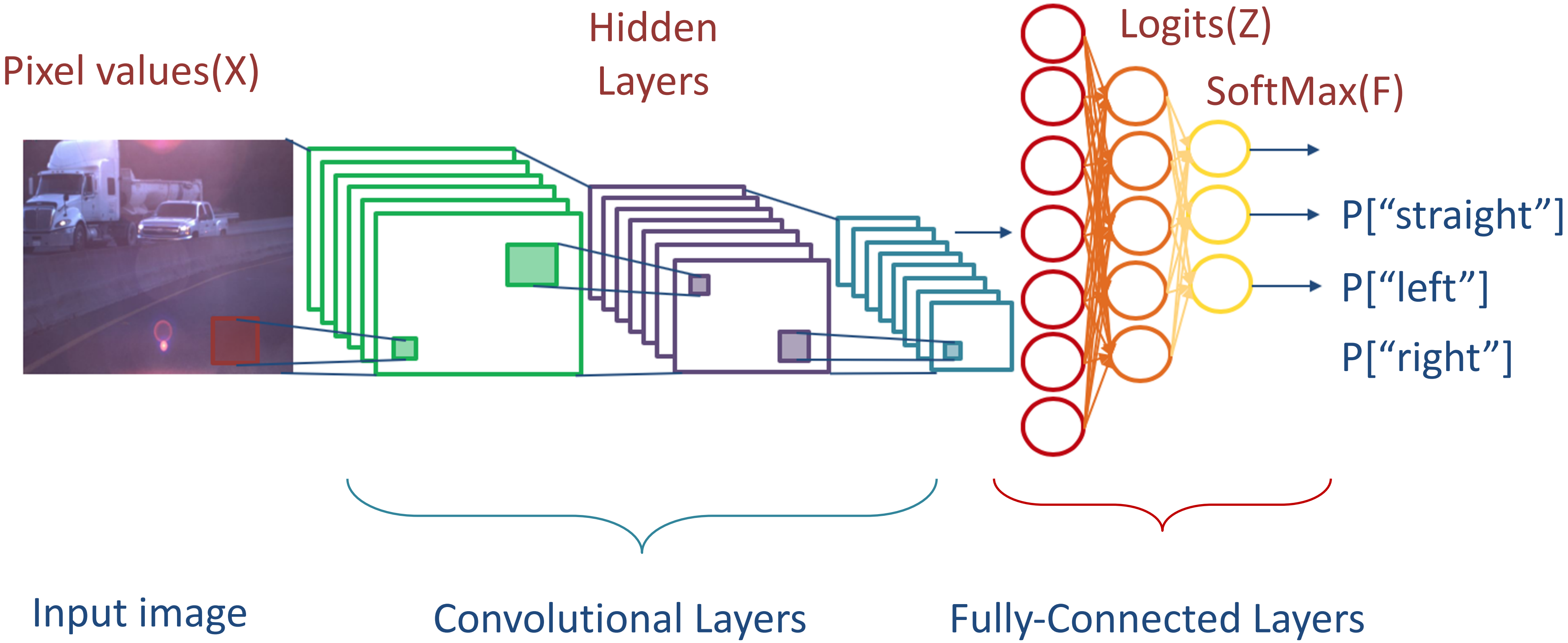
Predict steering angle

A. Chernikova, A. Oprea, C. Nita-Rotaru, and B. Kim.

Are Self-Driving Cars Secure? Evasion Attacks against Deep Neural Networks for Self-Driving Cars.

In IEEE SafeThings 2019. https://arxiv.org/abs/1904.07370

CNN for Direction Prediction



- Two CNN architectures: 25 million and 467 million parameters

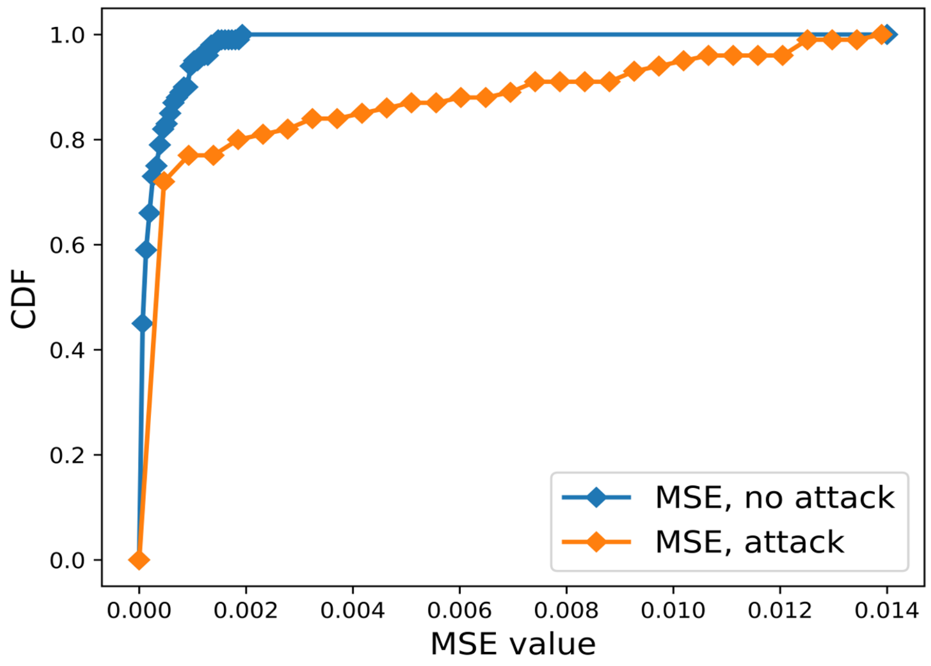
Evasion Attack against Regression

- First evasion attack for CNNs for regression
- New objective function
  - Minimize adversarial perturbation
  - Maximize the square residuals (difference between the predicted and true response)

$$\min_{\delta} c||\delta||_2^2 - G(x+\delta, y)$$

such that $x + \delta \in [0,1]^d$

$$G(x+\delta, y) = [F(x+\delta) - y]^2$$



- 10% of adversarial images have MSE 20 times higher than legitimate
- The maximum ratio of adversarial to legitimate MSE reaches 69

Adversarial Example for Regression

Original Image
Steering angle = -4.25; MSE = 0.0016

Adversarial Image
Steering angle = -2.25; MSE = 0.05

- Significant degradation of CNN classifiers in connected cars
- Small amount of perturbation is effective
- Models for both classification and regression are vulnerable

Taxonomy

Attacker's Objective

Learning stage

	Targeted Target small set of points	**Availability** Target majority of points	**Privacy** Learn sensitive information
Training	Targeted Poisoning Backdoor Trojan Attacks	Poisoning Availability	Membership Inference
Testing	Evasion Attacks Adversarial Examples	-	Membership Inference Model Extraction

Training-Time Attacks

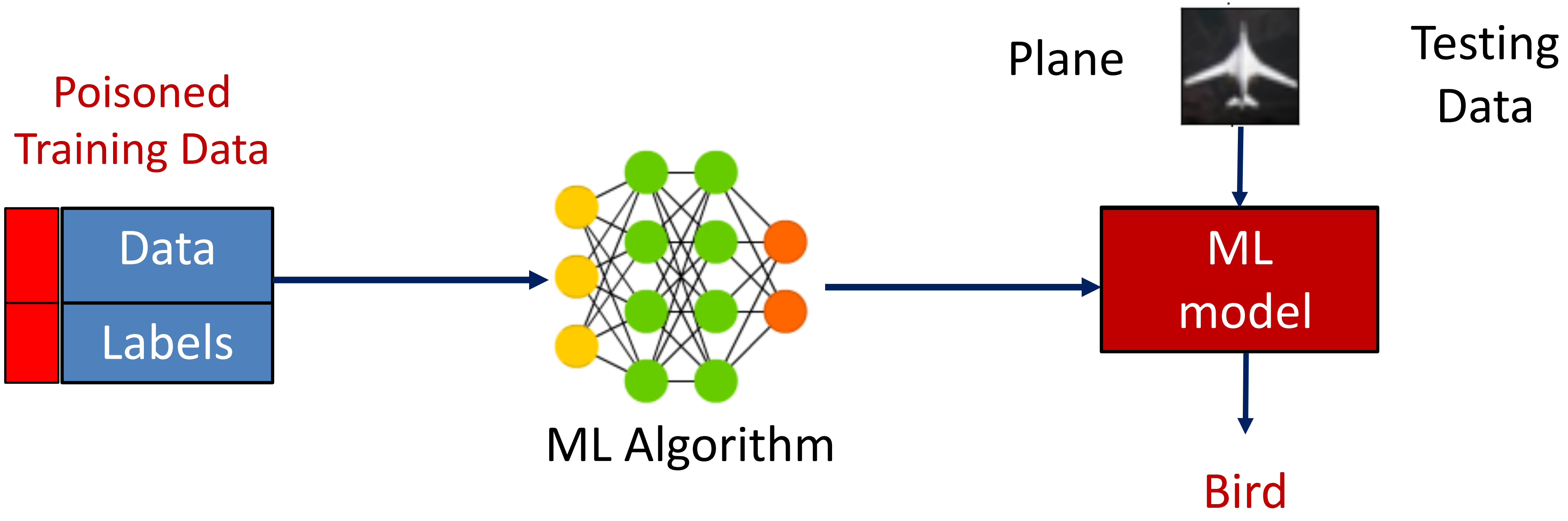
- ML is trained by crowdsourcing data in many applications

- Social networks
- News articles
- Tweets

- Navigation systems
- Face recognition
- Mobile sensors

- Cannot fully trust training data!

Poisoning Availability Attacks



- Attacker Objective:
  - Corrupt the predictions by the ML model significantly
- Attacker Capability:
  - Insert fraction of poisoning points in training

M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li. *Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning*. In IEEE S&P 2018

Optimization Formulation

Given a training set D find a set of poisoning data points D_p

that maximizes the adversary objective A on validation set D_{val}

where corrupted model $\boldsymbol{\theta}_p$ is learned by minimizing the loss L on $D \cup D_p$

$$\underset{D_p}{\operatorname{argmax}}\, A(D_{val}, \boldsymbol{\theta}_p)\ s.t.$$
$$\boldsymbol{\theta}_p \in \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, L(D \cup D_p, \boldsymbol{\theta})$$

Bilevel Optimization
NP-Hard!

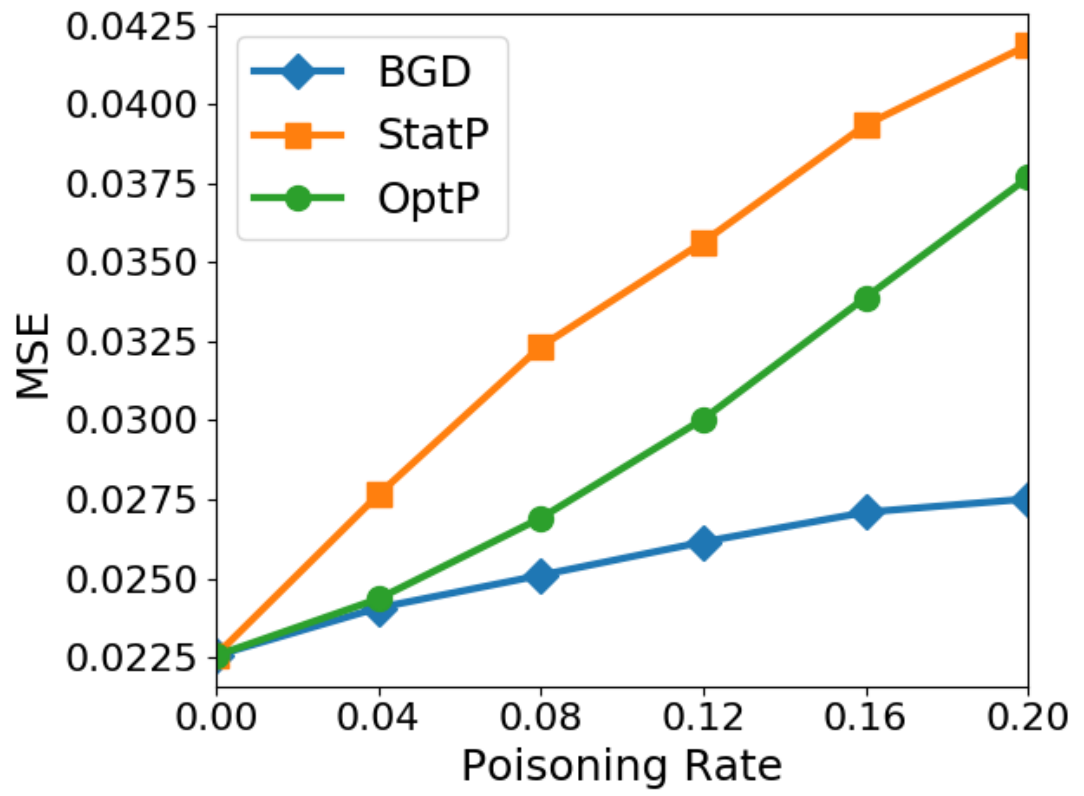
First white-box attack for linear regression [Jagielski et al. 18]

- Determine optimal poisoning point $(\boldsymbol{x}_c, y_c)$
- Optimize by both $\boldsymbol{x}_c$ and y_c

Poisoning Regression

- Improve existing attacks by a factor of at most 6.83

Stronger attack



Novel attacks

Existing attack

Predict loan rate with ridge regression
(i.e. with L2 regularization)

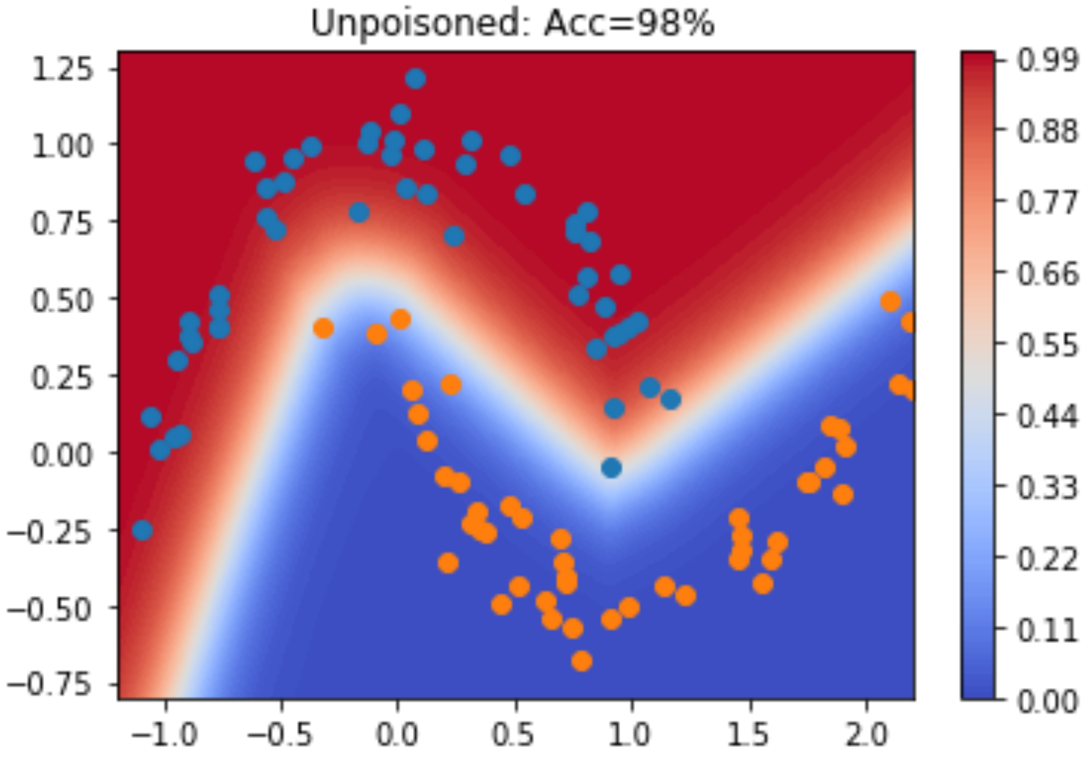
Is It Really a Threat?

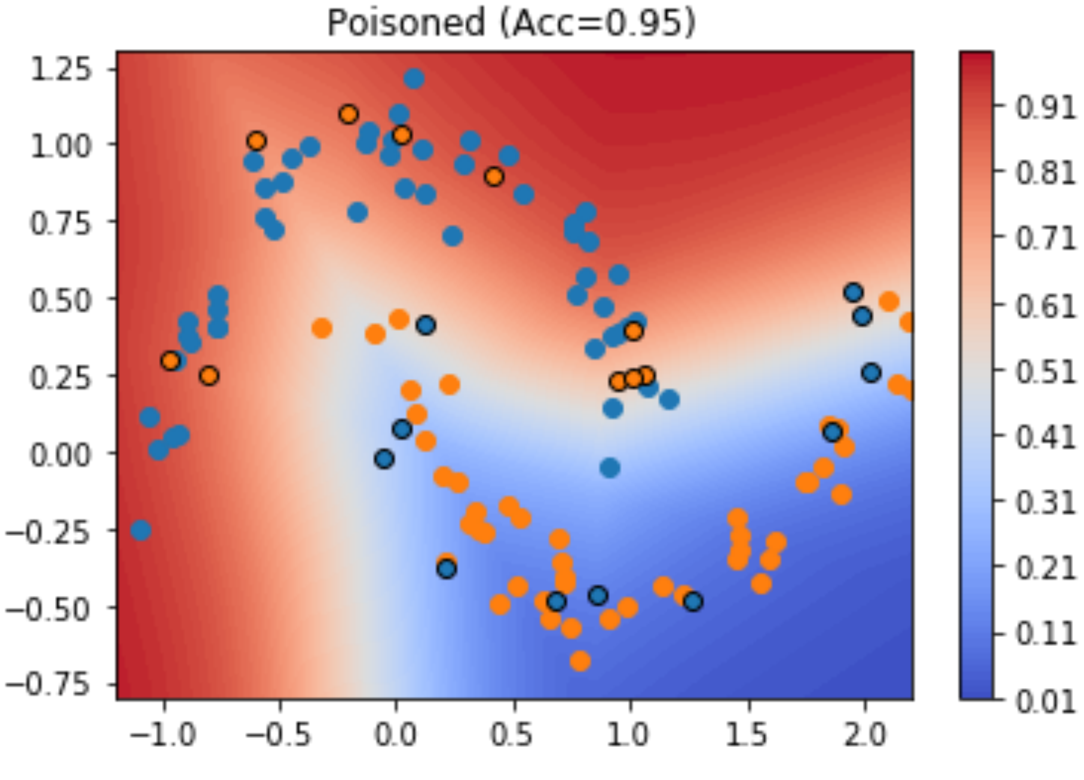
- Case study on healthcare dataset (predict Warfarin medicine dosage)
- At 20% poisoning rate
  - Modifies 75% of patients’ dosages by 93.49% for LASSO
  - Modifies 10% of patients’ dosages by a factor of 4.59 for Ridge
- At 8% poisoning rate
  - Modifies 50% of the patients’ dosages by 75.06%

Quantile	Initial Dosage	Ridge Difference	LASSO Difference
0.1	15.5 mg/wk	31.54%	37.20%
0.25	21 mg/wk	87.50%	93.49%
0.5	30 mg/wk	150.99%	139.31%
0.75	41.53 mg/wk	274.18%	224.08%
0.9	52.5 mg/wk	459.63%	358.89%

Poisoning Neural Networks

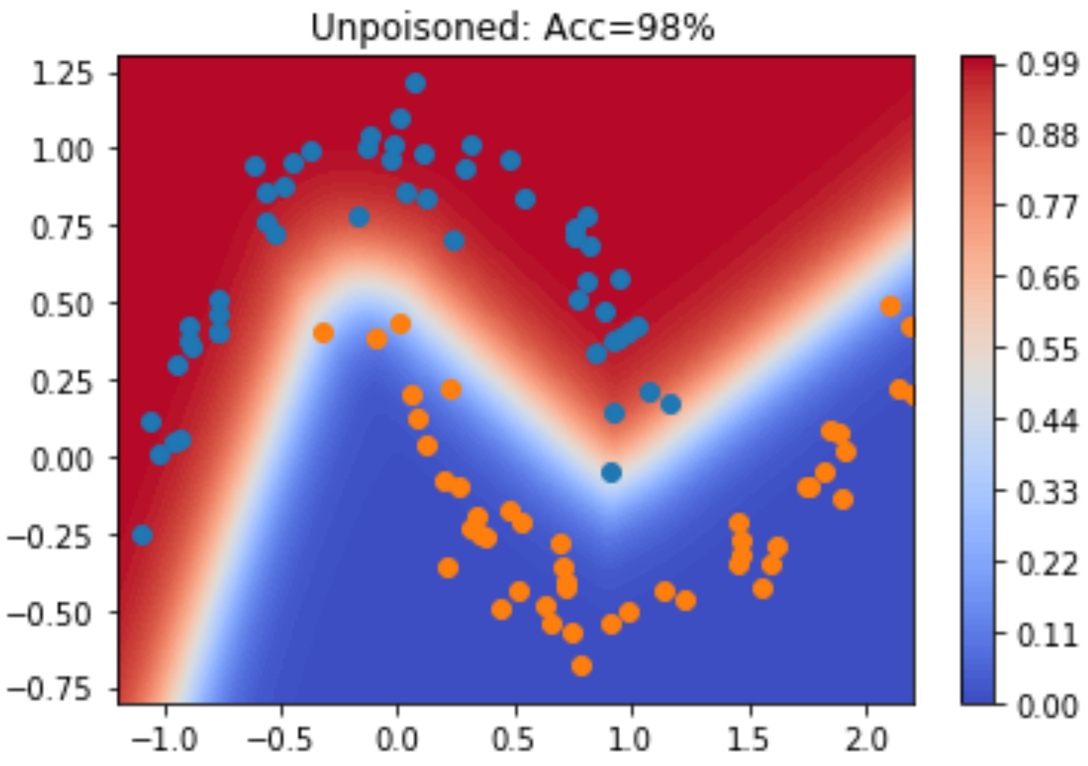
Availability with 20% random label flipping

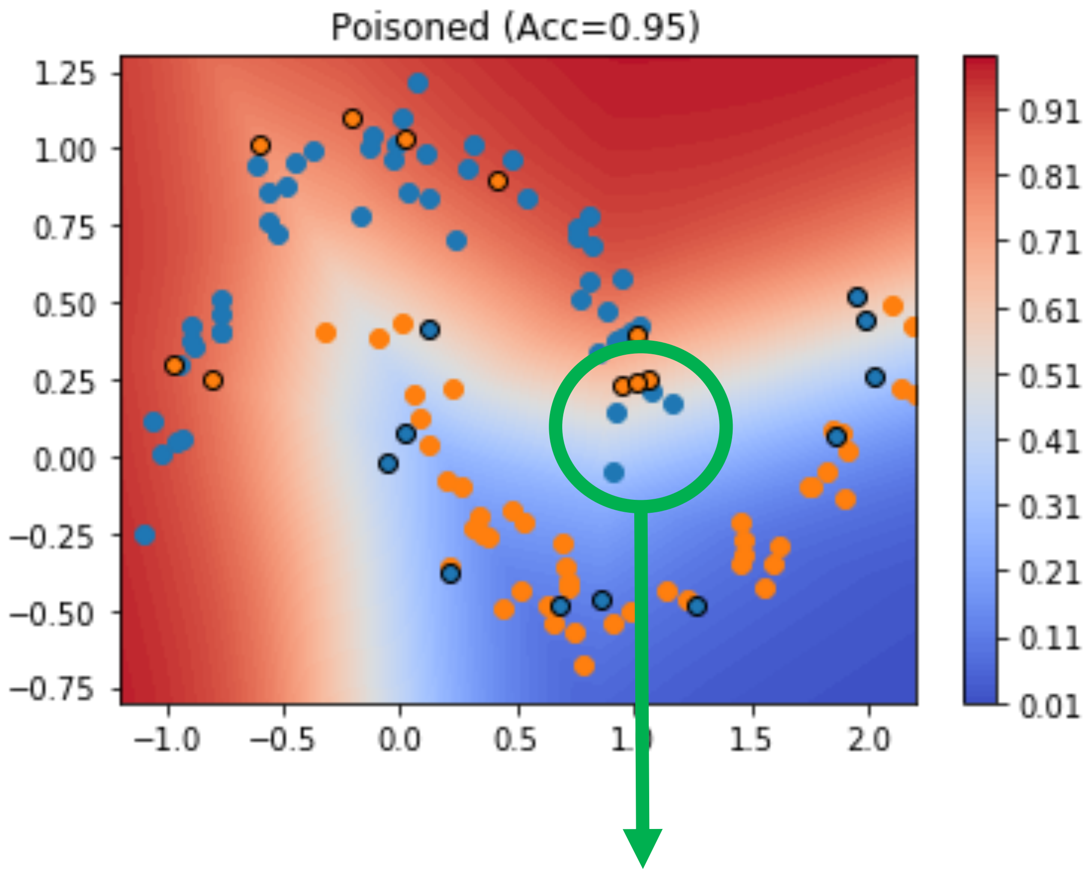




- Hard to change overall structure of decision boundary
- Availability attacks are easily detectable if classifier accuracy degrades

New Attack: Subpopulation Poisoning





Attack worked here!

Key Insights

- Data has natural clusters (subpopulations)
- Some subpopulations are more vulnerable
- Minority populations are affected more!

Initial Results

Subpopulation poisoning attack

- Perform data clustering
- Select clusters to poison (according to different criteria)
- Insert poisoned points from subpopulation with flipped label

Cluster	Original Accuracy	Poisoned Cluster Accuracy	Poisoned Points
C1: Size 35	100%	27.77%	70
C2: Size 29	94.11%	21.56%	58
C3: Size 22	100%	33.33%	36
C4: Size 26	100%	39.99%	43
C5: Size 39	92.85%	46.03%	65

UCI Adult Dataset

Towards Stealthy Poisoning Attacks

- New subpopulation poisoning attack
  – Attack is stealthy (difficult to detect)
  – Insert a small number of poisoned points in training
  – Does not require change of testing data
- Research questions
  – Which subpopulations are more vulnerable?
  – How to maximize the impact of the attack with minimum number of poisoning points?
  – Are defenses possible? Our conjecture is that not really!

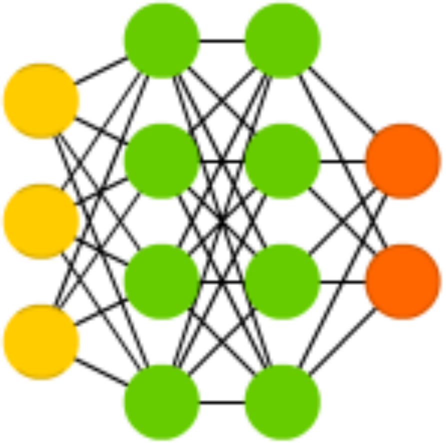
M. Jagielski, P. Hand, A. Oprea. *Subpopulation Data Poisoning Attacks*.
In Robust AI in Financial Services workshop at NeurIPS 2019.

Open Problem: Robust AI

- Most AI models are vulnerable in face of attacks!
  - Evasion (testing-time) attacks
  - Poisoning (training-time) attacks
  - Privacy attacks
- How to design AI algorithms robust to attacks?

Open Problem: AI under Constraints

Privacy



Fairness

- AI models face conflicting requirements in practice
  - Privacy of user data
  - Fairness of predictions
- How to design AI algorithms under constraints?

Takeaways

- AI has potential in security applications
  - Design intelligent and adaptive defense algorithms
  - *Current research*: AI and graph models to detect advanced attacks
  - *Current research*: Collaborative AI defenses
  - *Open problems*: Intelligent cyber defense, online learning in cyber

---

- …But AI becomes a target of attack
  - Traditional ML and Deep Neural Networks are not resilient to adversarial manipulations at training and testing time
  - *Current research*: Evasion and poisoning attacks for cyber security
  - *Current research*: ML under privacy and fairness constraints
  - *Open problems*: Design AI algorithms resilient against attacks

Acknowledgements

Contact Information
Alina Oprea
a.oprea@northeastern.edu