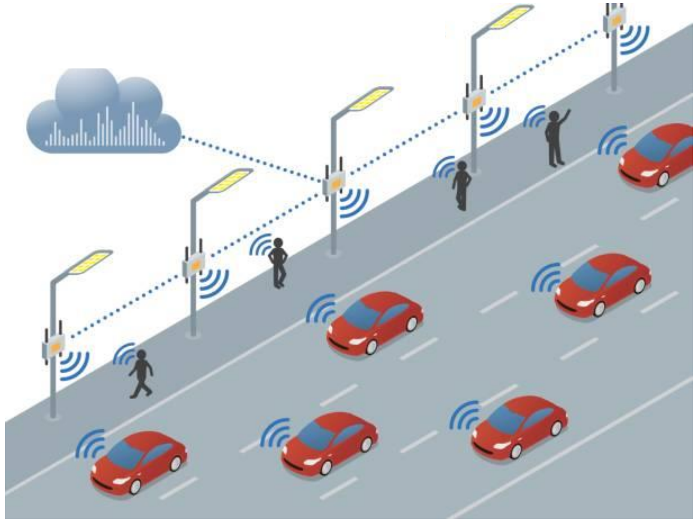
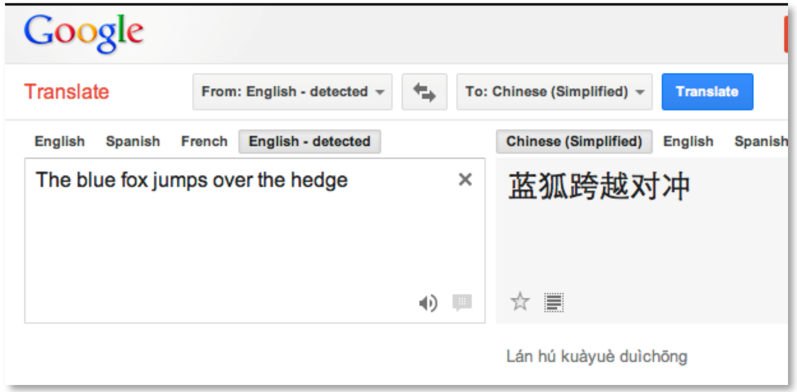
AI in Cybersecurity: Applications, Open Problems, and Future Directions

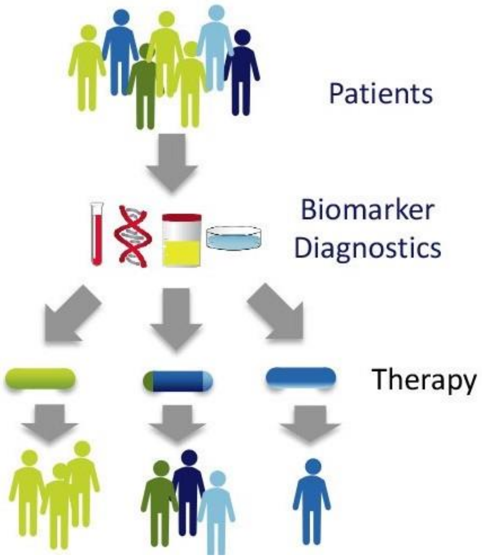
Alina Oprea
Associate Professor
Cybersecurity and Privacy Institute
Northeastern University

ACSAC
December 6 2018

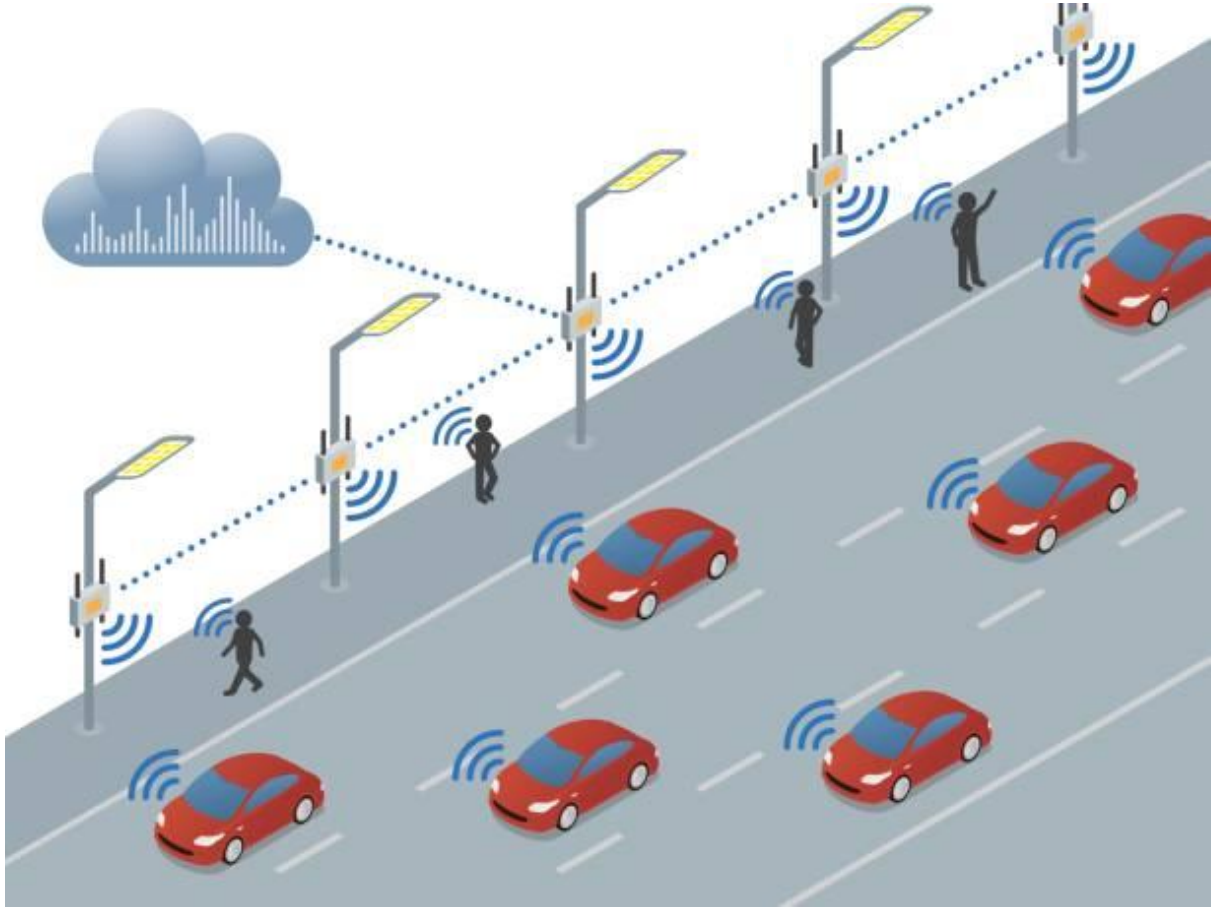
AI is Everywhere





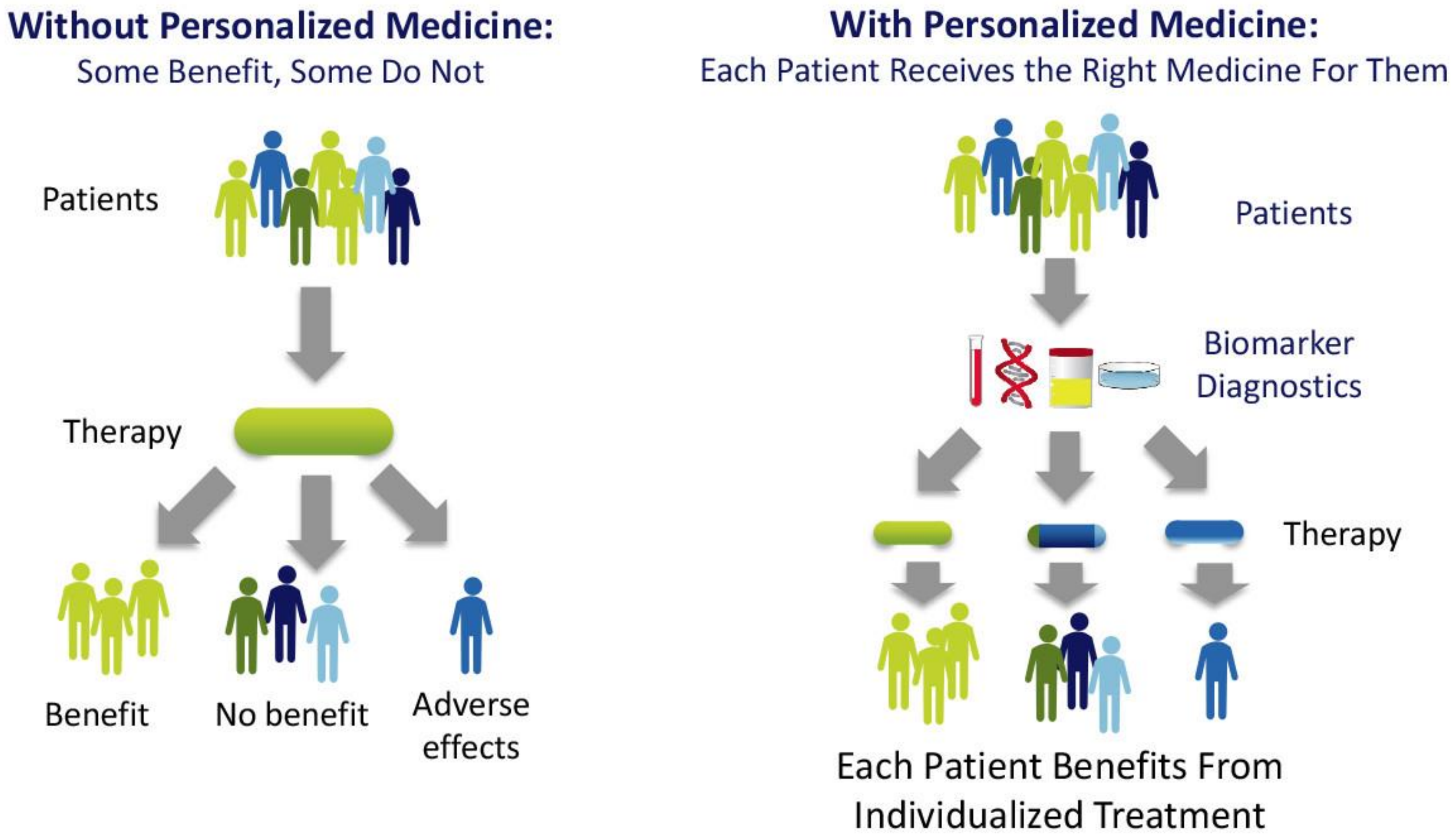


Connected Cars



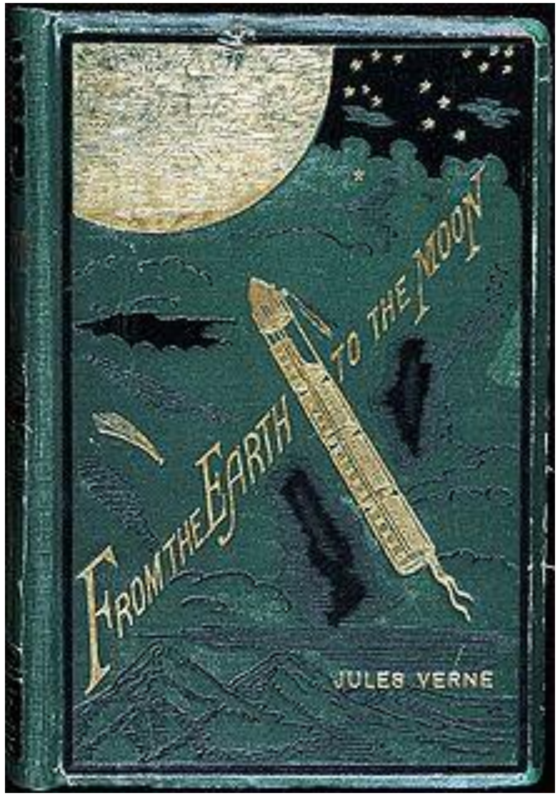
- Sensors for data collection
- Assist drivers in making decisions to increase safety

Personalized Medicine



- Treatment adjusted to individual patients
- Predictive models using a variety of features
- Better outcome and reduced cost

A Bit of History

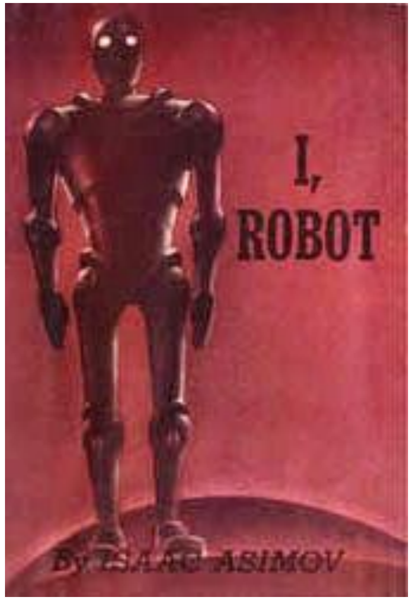


> 100 years

1865

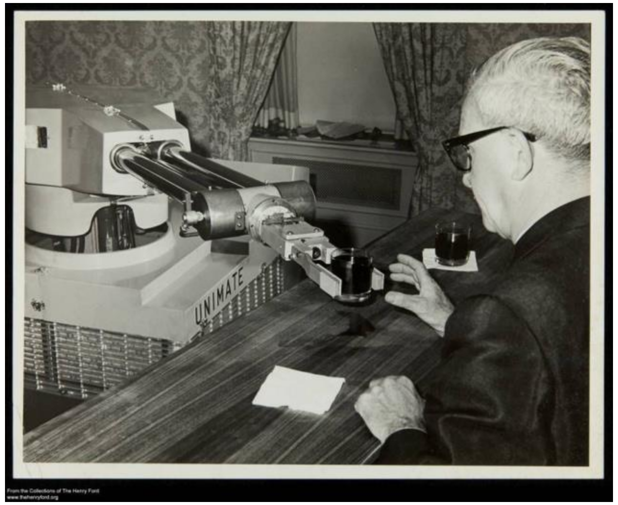
1969

A Bit of History



1940

> 50 years



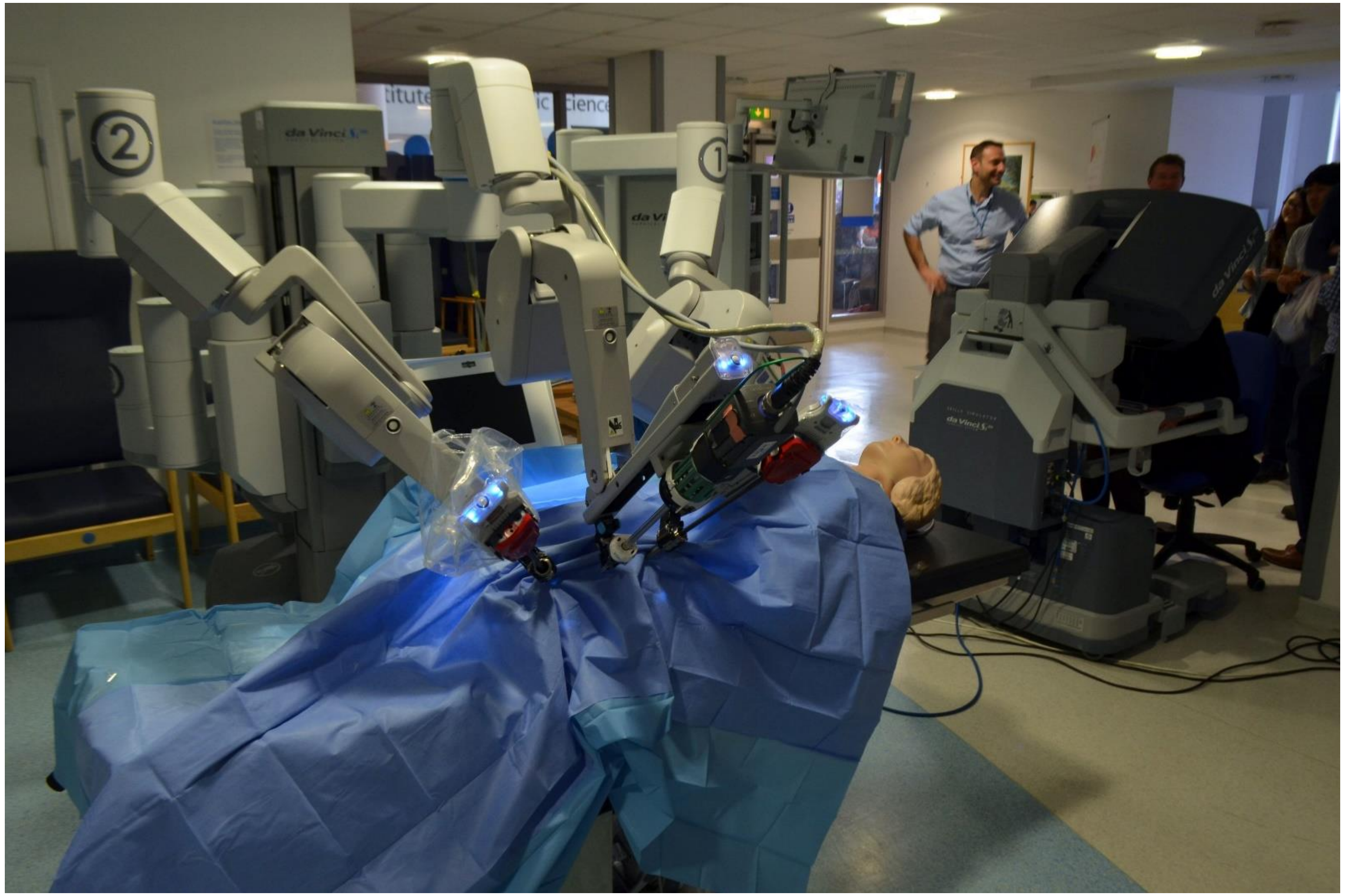
Unimate Robot
1961

Sony Dream
2001

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)

Fast Forward in the Far Future

What will happen in
100 years?

Implications for Cyber Security

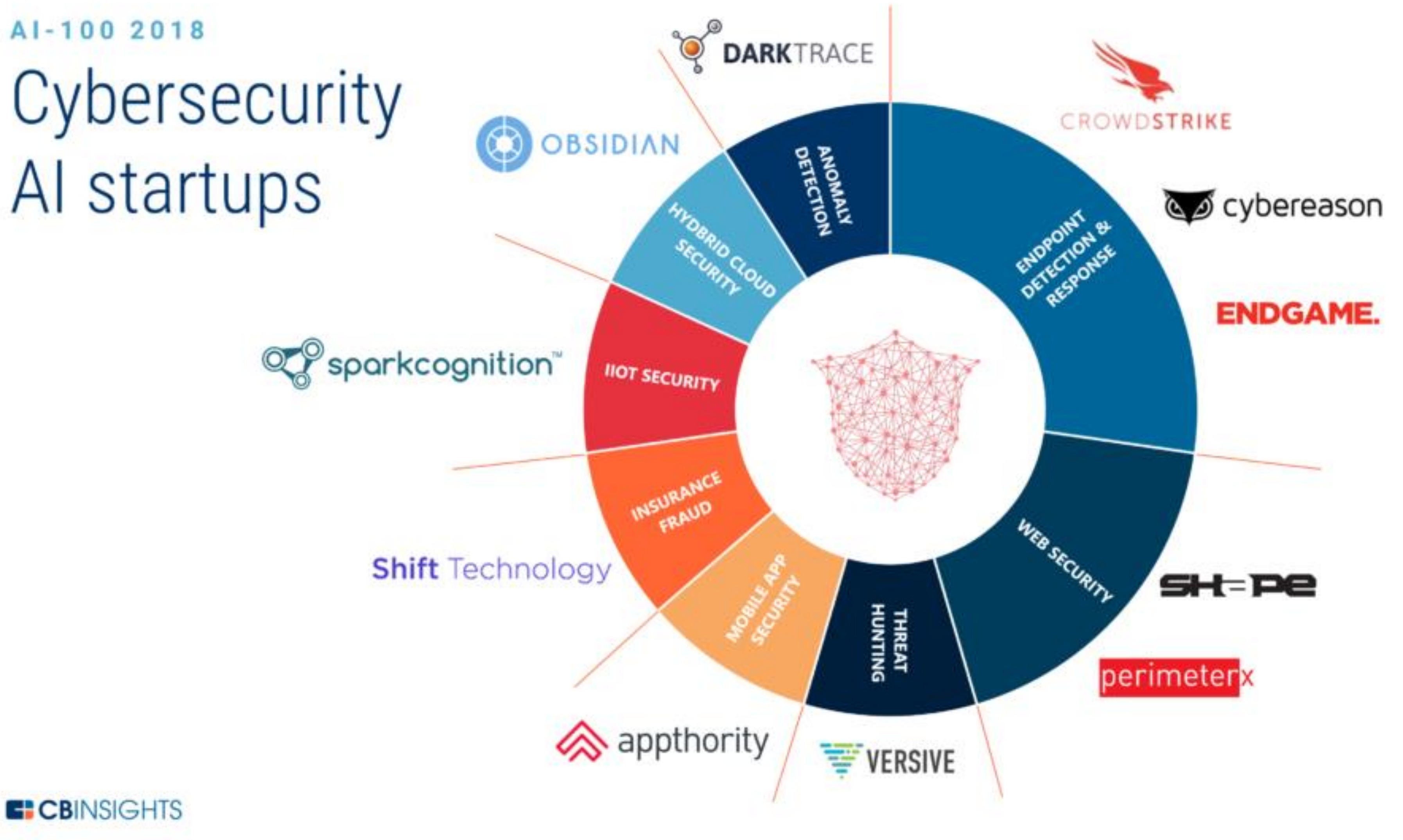
- AI has potential in security applications
  - Complement traditional defenses (crypto, multi-factor authentication, trusted hardware)
  - Design intelligent and adaptive defense algorithms

- …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!

AI in Cybersecurity

Can AI Improve Security?

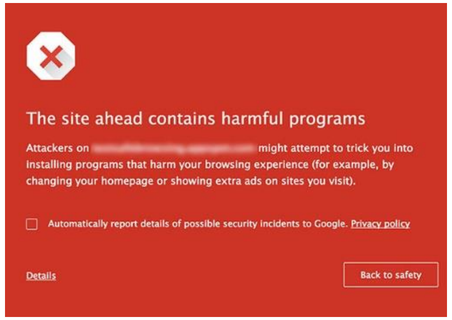
Industry


AI-100 2018
Cybersecurity AI startups
DARKTRACE
CROWDSTRIKE
OBSIDIAN
cybereason
ANOMALY DETECTION
HYDBRID CLOUD SECURITY
ENDPOINT DETECTION & RESPONSE
ENDGAME.
sparkcognition
IIOT SECURITY
INSURANCE FRAUD
Shift Technology
MOBILE APP SECURITY
THREAT HUNTING
WEB SECURITY
SHAPE
perimeterx
appthority
VERSIVE
CBINSIGHTS

AI-Enabled Defenses

- Spam and phishing detection
  - [Castillo et al. 07], [Ma et al. 09]

- Detect compromised accounts in social networks
  - [Egele et al. 13], [Thomas et al. 14], [Cao et al. 14]



- Malicious web sites and web connections
  - [Bilge et al. 11], [Antonakakis et al. 12], [Hao et al. 17]

- Predict security events
  - [Liu et al. 15], [Shen et al. 18]

Security Breaches

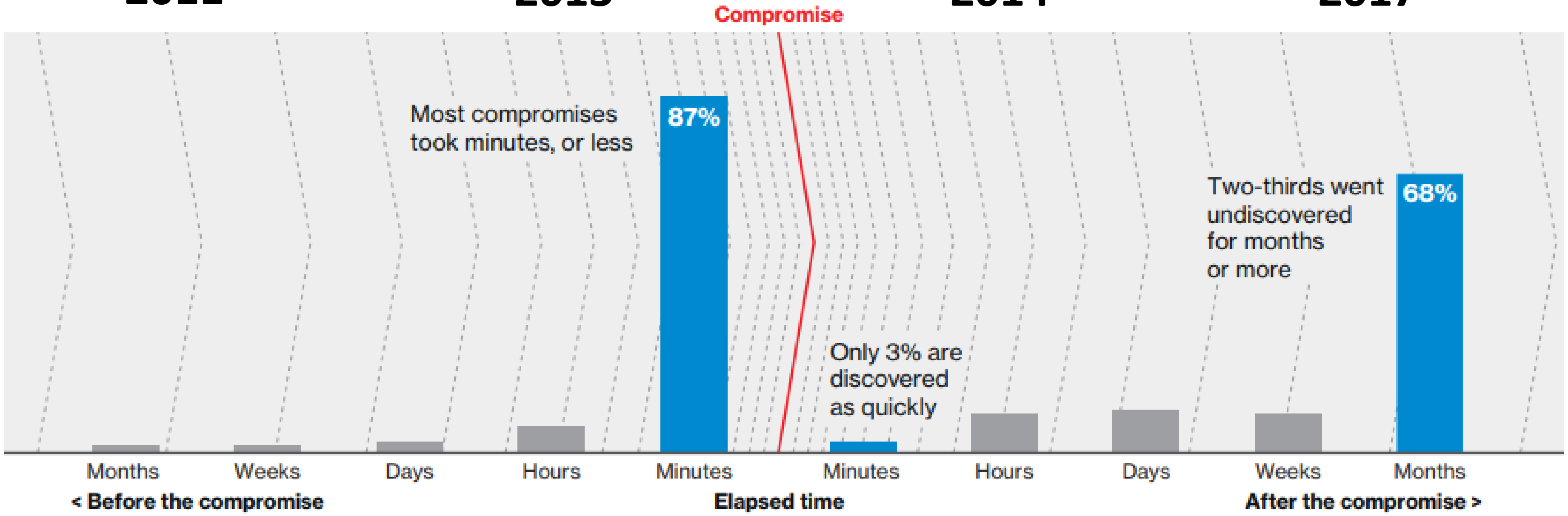
RSA SecurID®
Breach

2011

2013

2014

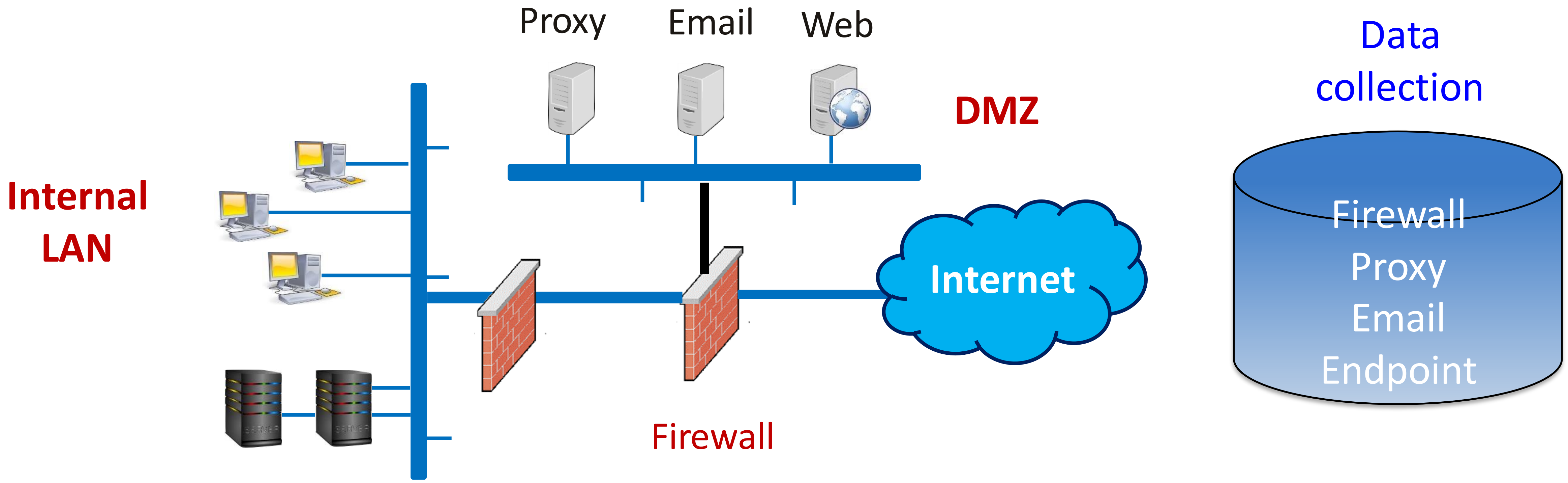
2017



- Exfiltration of sensitive information
- Loss of intellectual property
- Financial losses

Source: Verizon DBIR

Defenses in Enterprise Networks



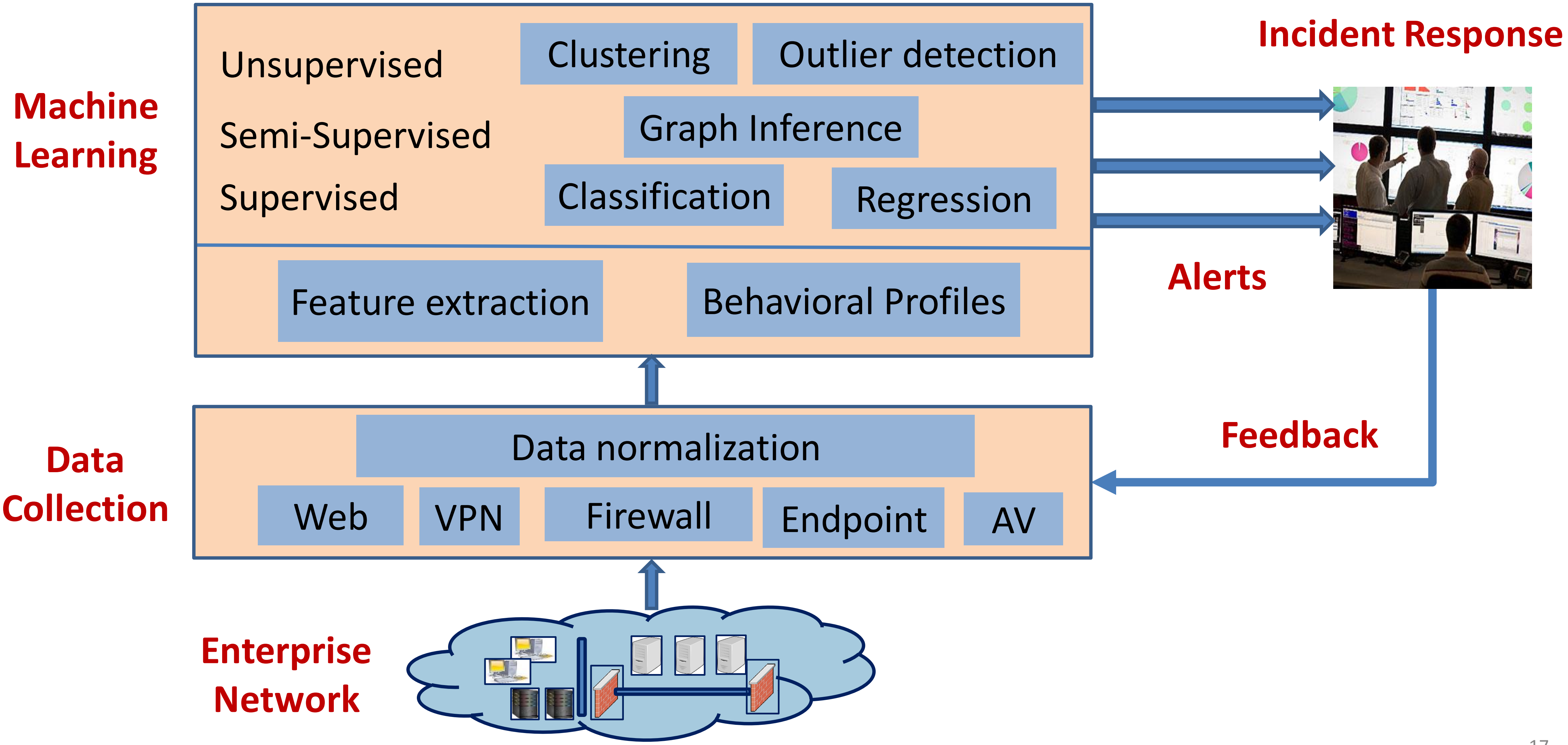
- Security controls deployed for network and host protection
- Security logs mostly used for forensic investigation
- How can we detect and predict breaches using security logs?

Challenges of AI in Security

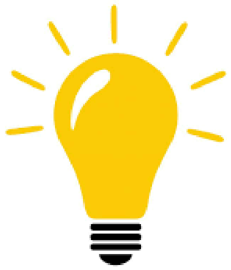
- AI is successful in many domains
  - Product recommendation, NLP, speech recognition
- What is different in cyber security?
  1. High cost of errors (both false positives and false negatives)
  2. Variability of user activity under normal conditions
  3. Interpretability of results to facilitate manual investigation
  4. Resilience against advanced adversaries

Limited success of machine learning for security in operational environments [Sommer and Paxson 2010]

RSA Analytics Framework



Key Ideas

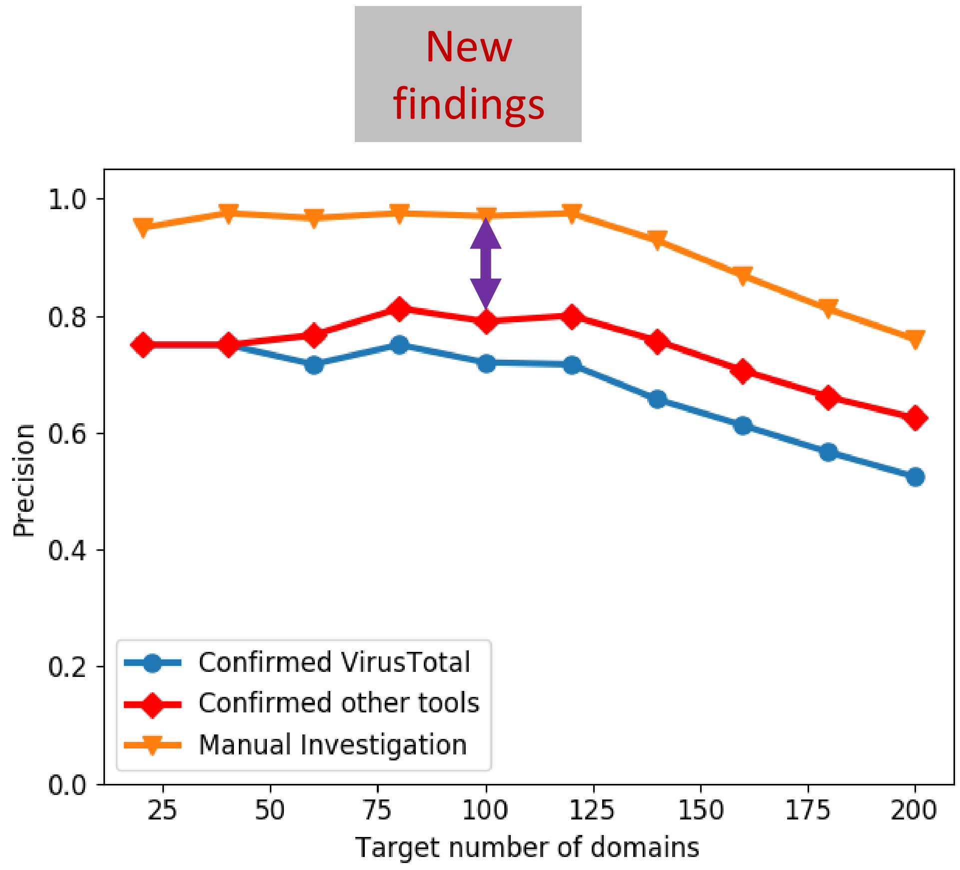


- Design ML modules for specific attack patterns
  - E.g., C&C, lateral movement, data exfiltration
  - Maximize precision and reduce false positive rates
  - Combine multiple models for increased recall of malicious activities
- Continuous interaction with EMC CIRC over several years
- Leverage ground truth from existing security products and previous incidents investigated by CIRC
- Interpretability of results

Recommendations by [Sommer and Paxson 2010]

MADE

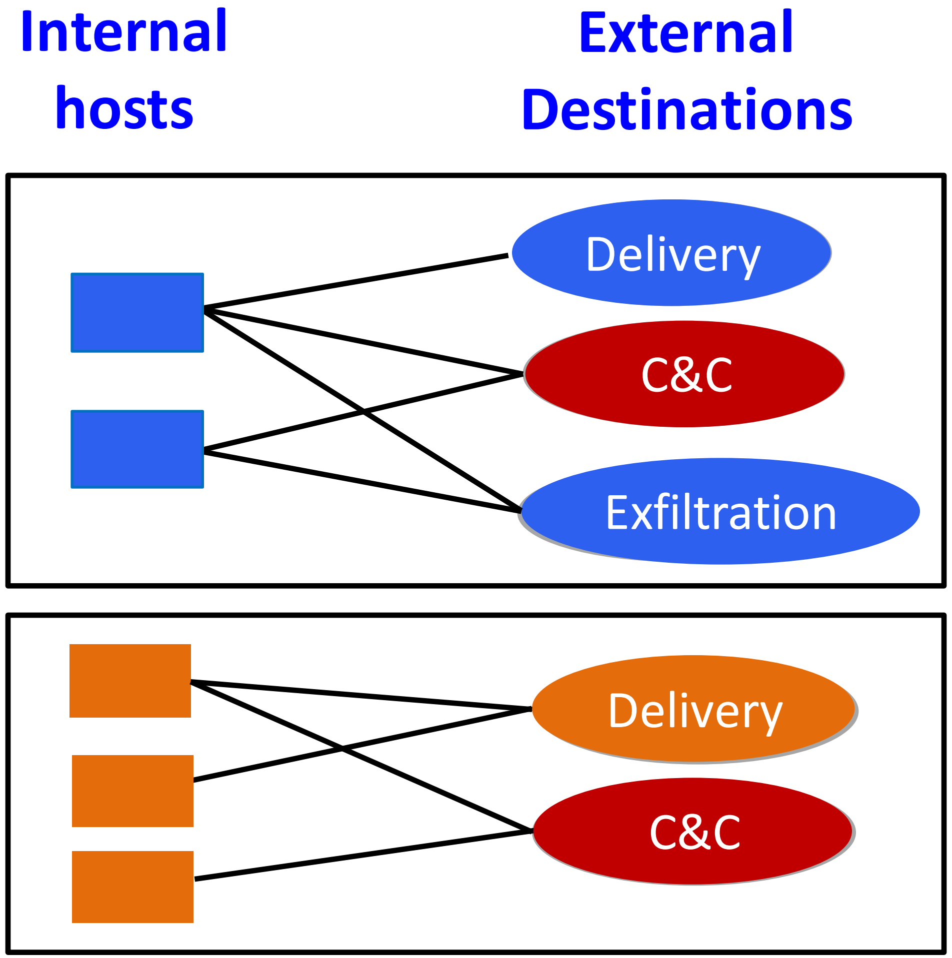
- **Goals**
  - Identify HTTP Command-and-Control (C&C) communication
- **Approach**
  - Use 10 categories of generic and enterprise features (89 total features)
  - Enterprise-specific profiles of domains and user-agent strings
  - Supervised learning (classification)
- **Output**
  - Prioritized list of external C&C domains



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

Multi-Stage Attacks

- **Goals**
  - Detect all domains and hosts involved in multi-stage campaigns
- **Approach**
  - Semi-supervised learning
  - Construct bipartite communication graph
  - Label C&C domains as seeds
  - Propagate risk with belief propagation
- **Output**
  - Prioritized list of malicious domains
  - Compromised hosts



A. Oprea, Z. Li, T.-F. Yen, S. Chin, S. Alrwais. *Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data*. DSN 2015.

Deployment Statistics

Command-and-Control (C&C)

- Dataset
  - 20 TB
- Precision (confirmed malicious)
  - 97%
- False positive rates:
  - 6×10^{-3} %
- New detections in one month
  - 18 domains

Multi-Stage Attacks

- Dataset
  - 38 TB
- Precision (confirmed malicious)
  - 85%
- False positive rates:
  - 8.58×10^{-4} %
- New detections in one month
  - 152 domains
  - 945 compromised hosts

Open Problems: Interpretable Models for Security

- Why does the ML model predict something as attack?
- What type of attack it is?
- Is it similar to known attacks?
- Is it a new attack/zero-day?
- What is the root cause?

Open Problems: Measurable Security

- What are the right metrics in cyber security?
- How do we compare different models?
- What are some good benchmarks?

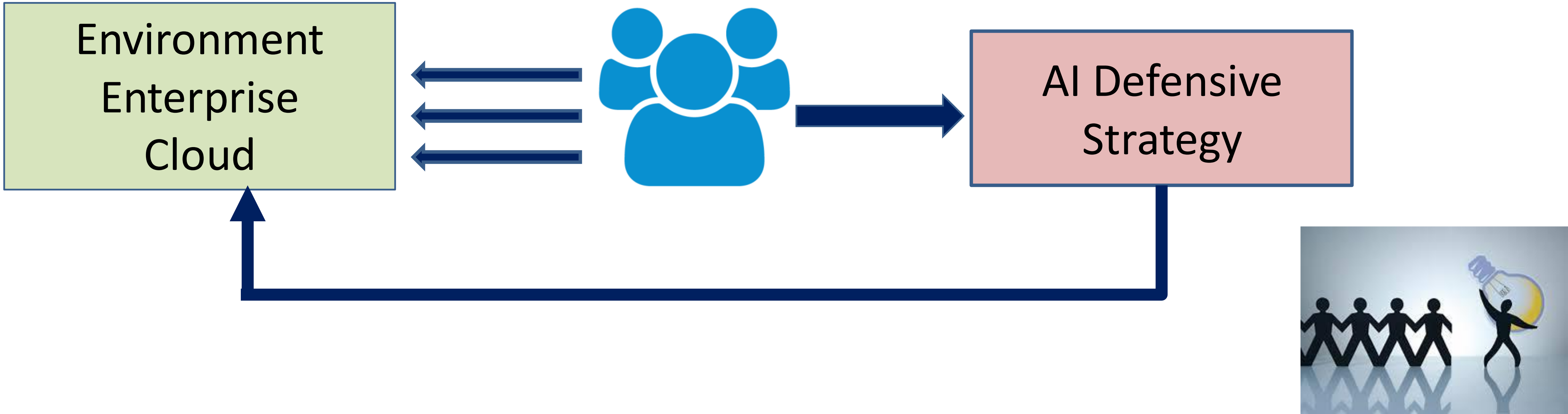
		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

$$\text{accuracy} = \frac{TP + TN}{P + N}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Open Problem: Intelligent Automation



Implications for Cyber Security

- AI has potential in security applications
  - Complement traditional defenses (crypto, multi-factor authentication, trusted hardware)
  - 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!

Security of AI

Can AI Be Secured?

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	-
Testing	Evasion Attacks Adversarial Examples	-	Model Extraction Model Inversion

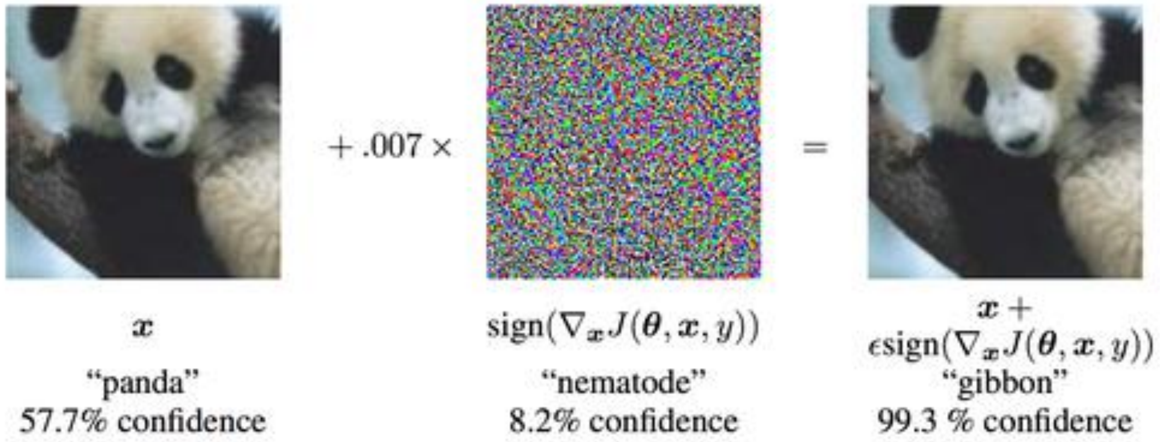
Adversarial Machine Learning: Taxonomy

Attacker's Objective

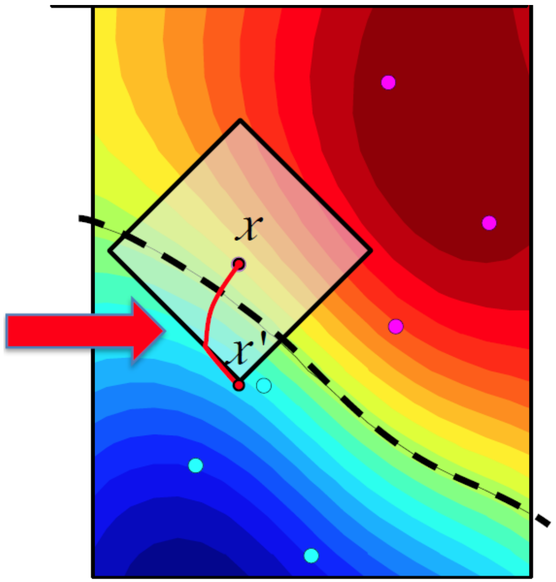
Learning stage

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

Evasion Attacks



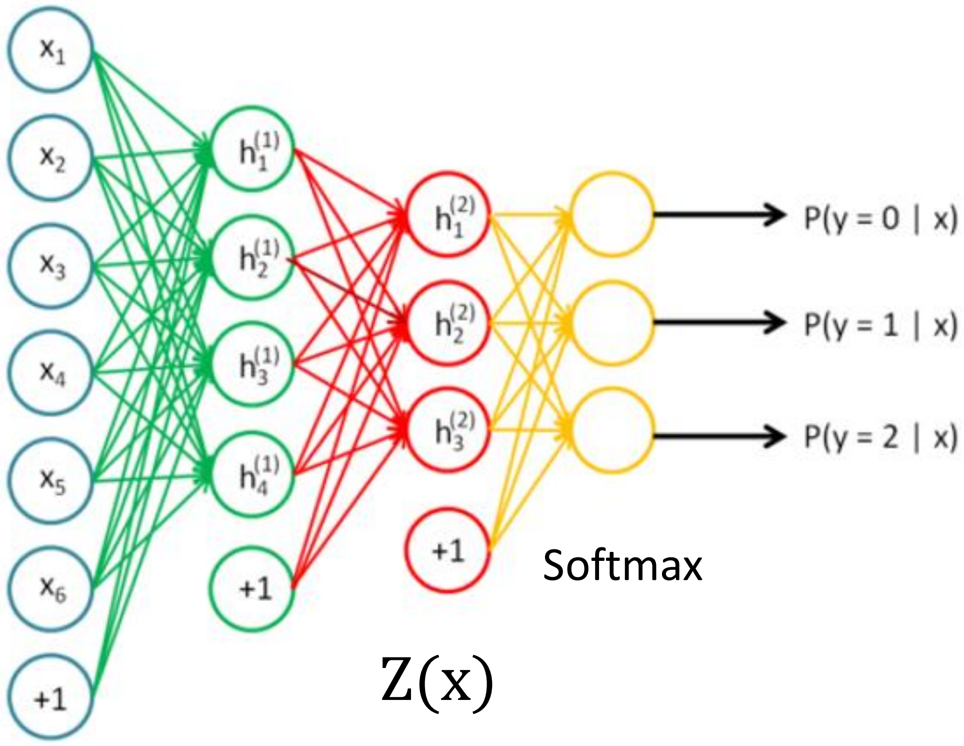
Adversarial example



- [Szegedy et al. 13] Intriguing properties of neural networks
- [Biggio et al. 13] Evasion Attacks against Machine Learning at Test Time
- [Goodfellow et al. 14] Explaining and Harnessing Adversarial Examples
- [Carlini, Wagner 17] Towards Evaluating the Robustness of Neural Networks
- [Madry et al. 17] Towards Deep Learning Models Resistant to Adversarial Attacks
- [Kannan et al. 18] Adversarial Logit Pairing
- …

Evasion Attacks For Neural Networks

Input: Images represented as feature vectors



Optimization Formulation

Given input x

Find adversarial example

$$x' = x + \delta$$

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

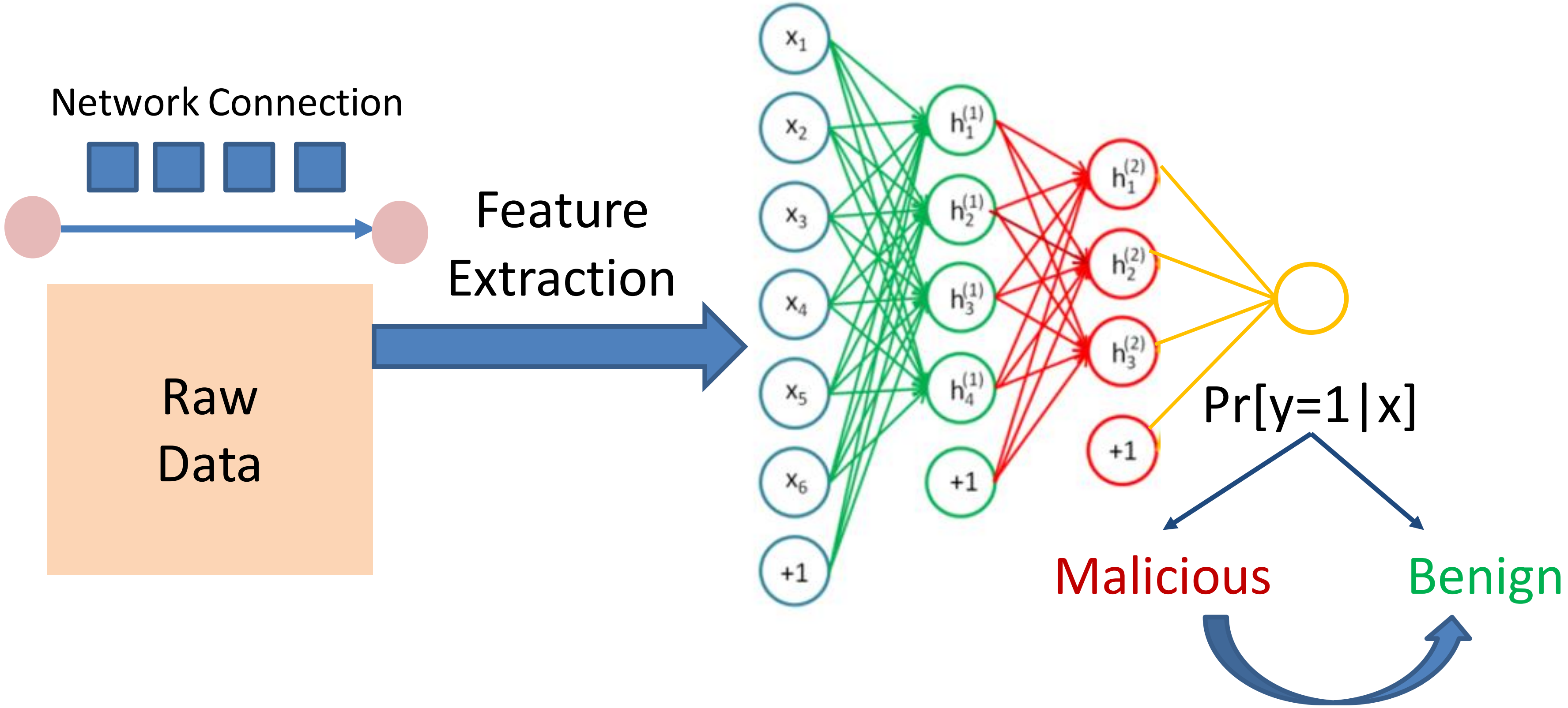
Min distance

Change class

[Carlini and Wagner 2017] Penalty method

[Biggio et al. 2013, Madry et al. 2018] Projected Gradient Descent

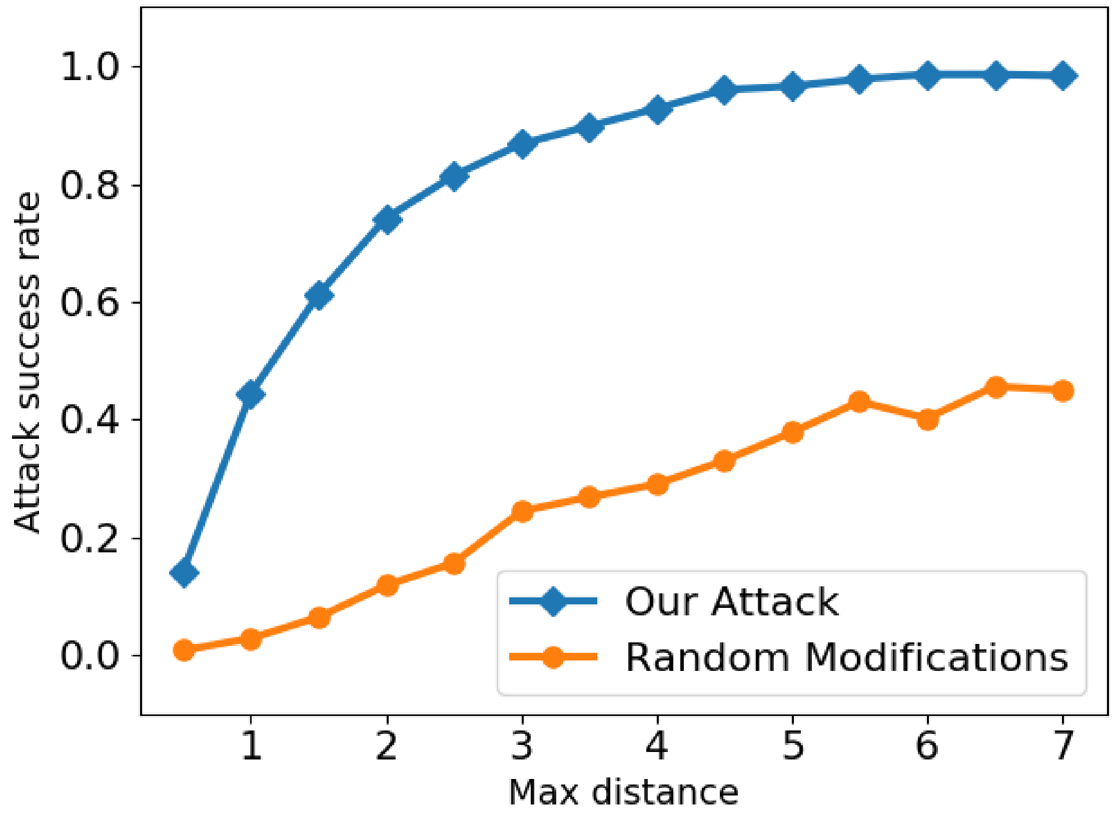
Evasion Attacks for Security



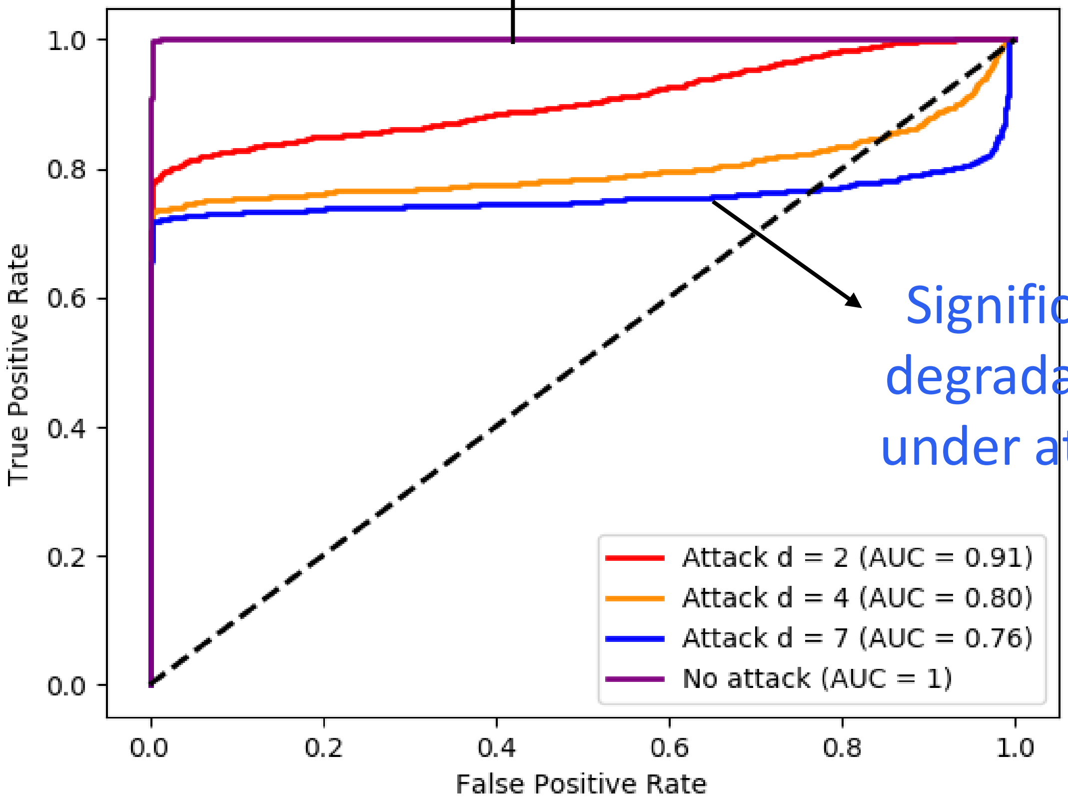
Challenge
- Attacks in feature space are not feasible in raw data space

Solution
- New iterative attack algorithm taking into account feature constraints

How Effective are Evasion Attacks in Security?



Perfect accuracy (No attack)



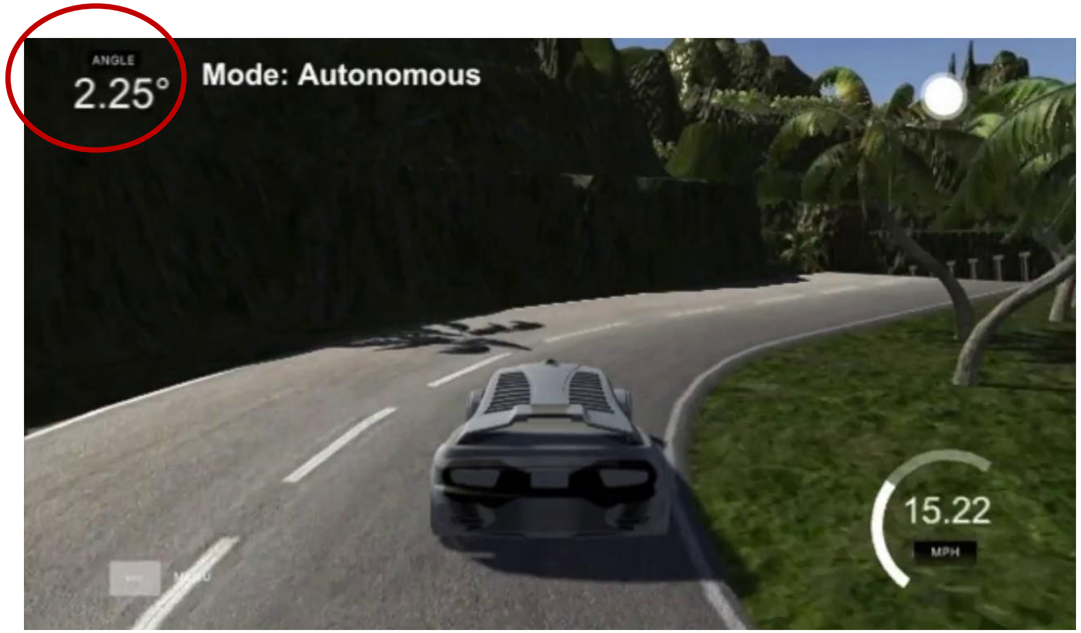
Significant degradation under attack

Feed-Forward Neural Network
83 features

Evasion Attacks in Connected Cars

Udacity Challenge

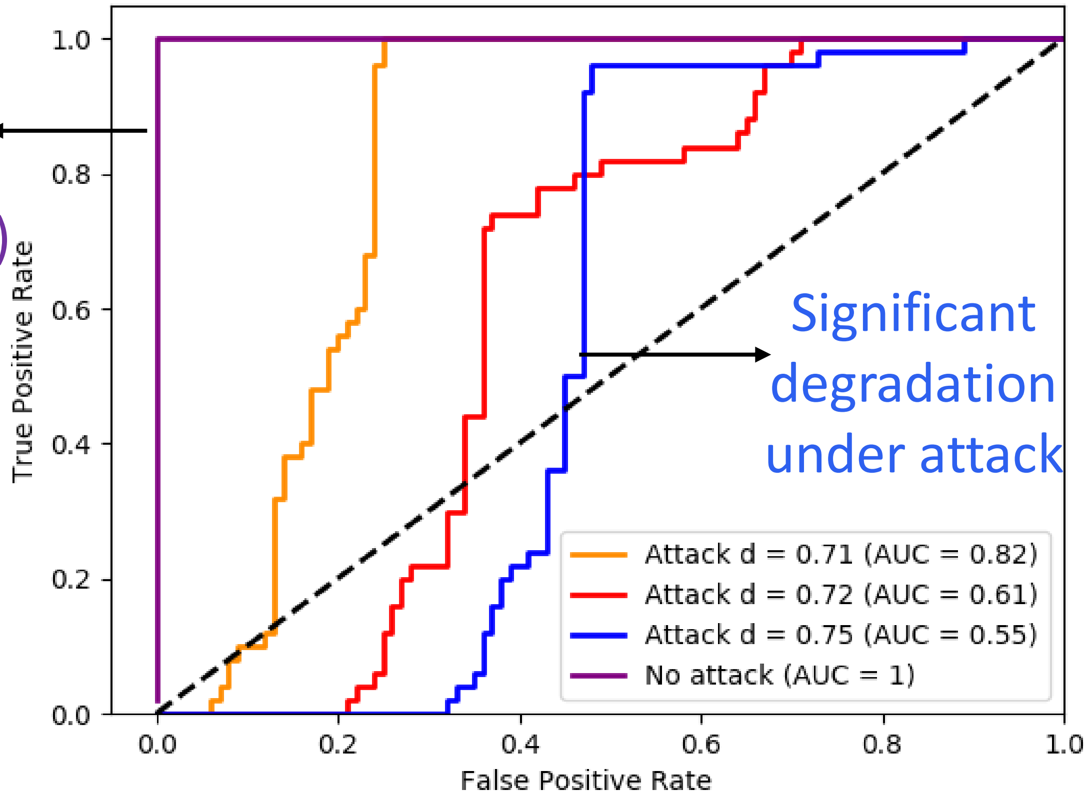
- Public competition and dataset 2014
- Steering angle prediction from camera image



Predict direction: Straight, Left, Right

How Effective are Evasion Attacks in Connected Cars?

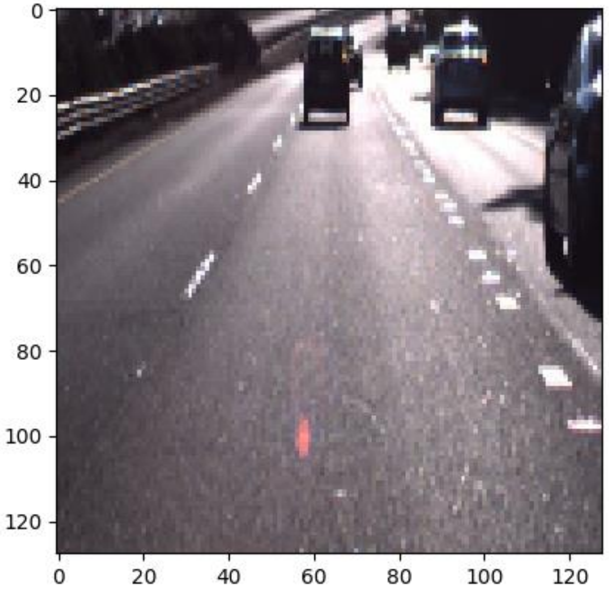
Perfect accuracy (no attack)



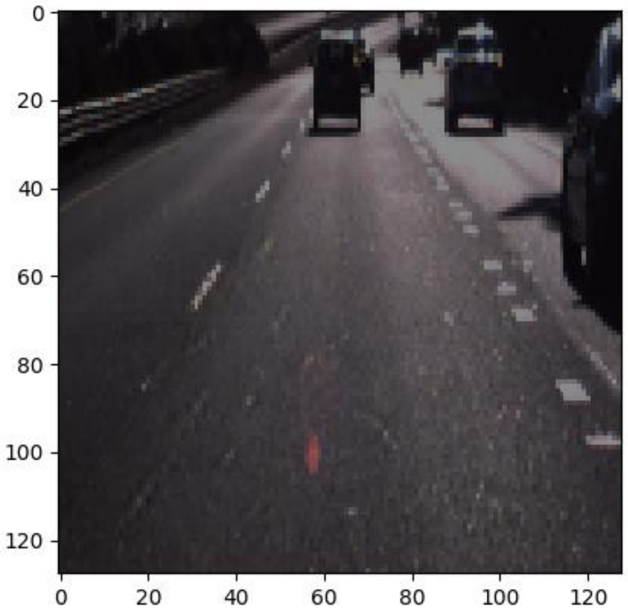
Significant degradation under attack

Convolutional Neural Network
25 million parameters

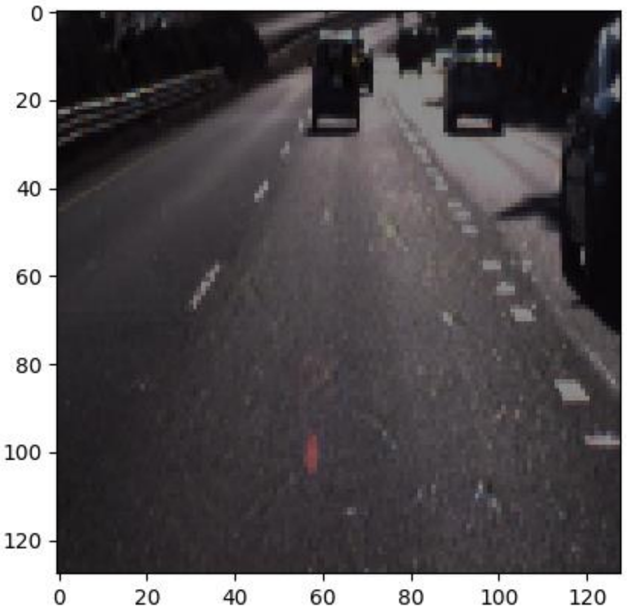
Adversarial Examples



Original Image
Class “Straight”

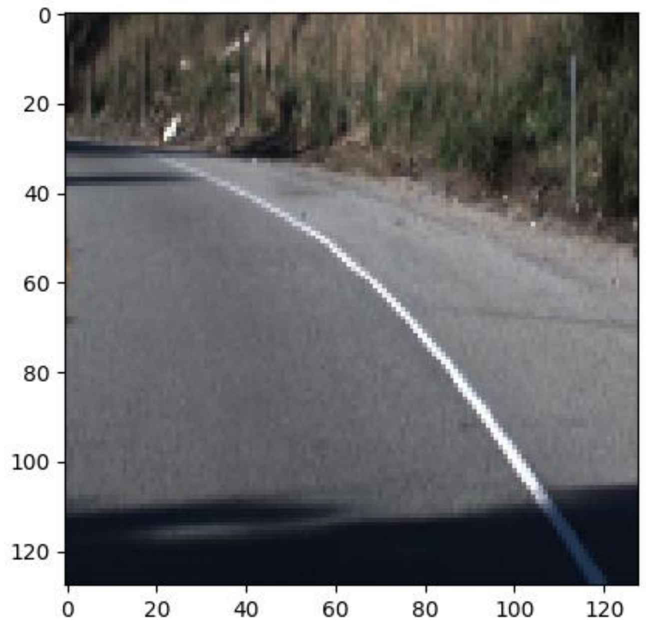


Adversarial Image
Class “Right”

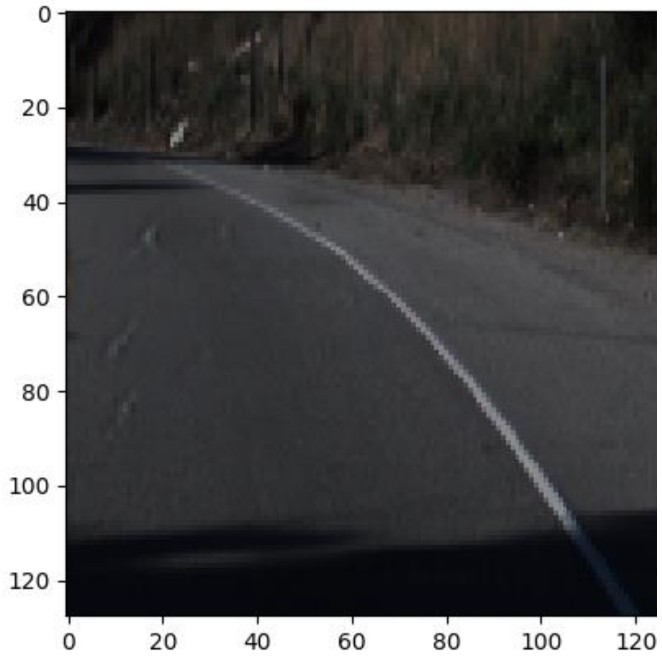


Adversarial Image
Class “Left”

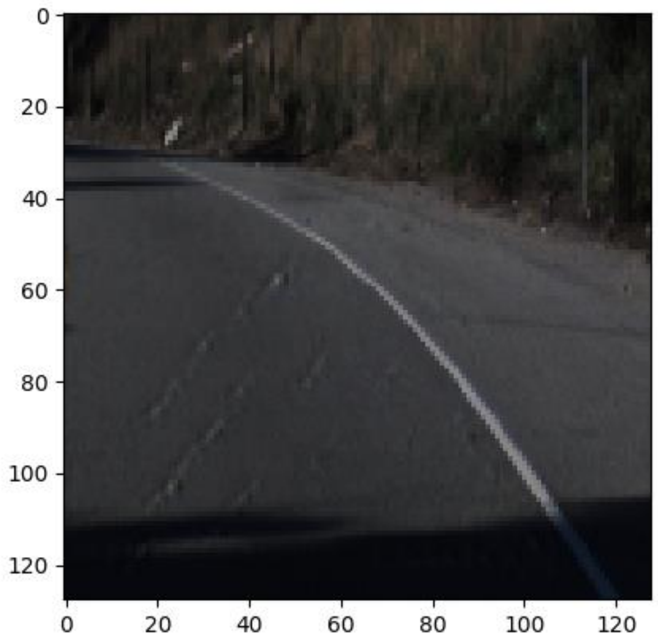
Adversarial Examples



Original Image
Class “Left”



Adversarial Image
Class “Straight”



Adversarial Image
Class “Right”

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	-
Testing	Evasion Attacks Adversarial Examples	-	Model Extraction Model Inversion

Training-Time Attacks

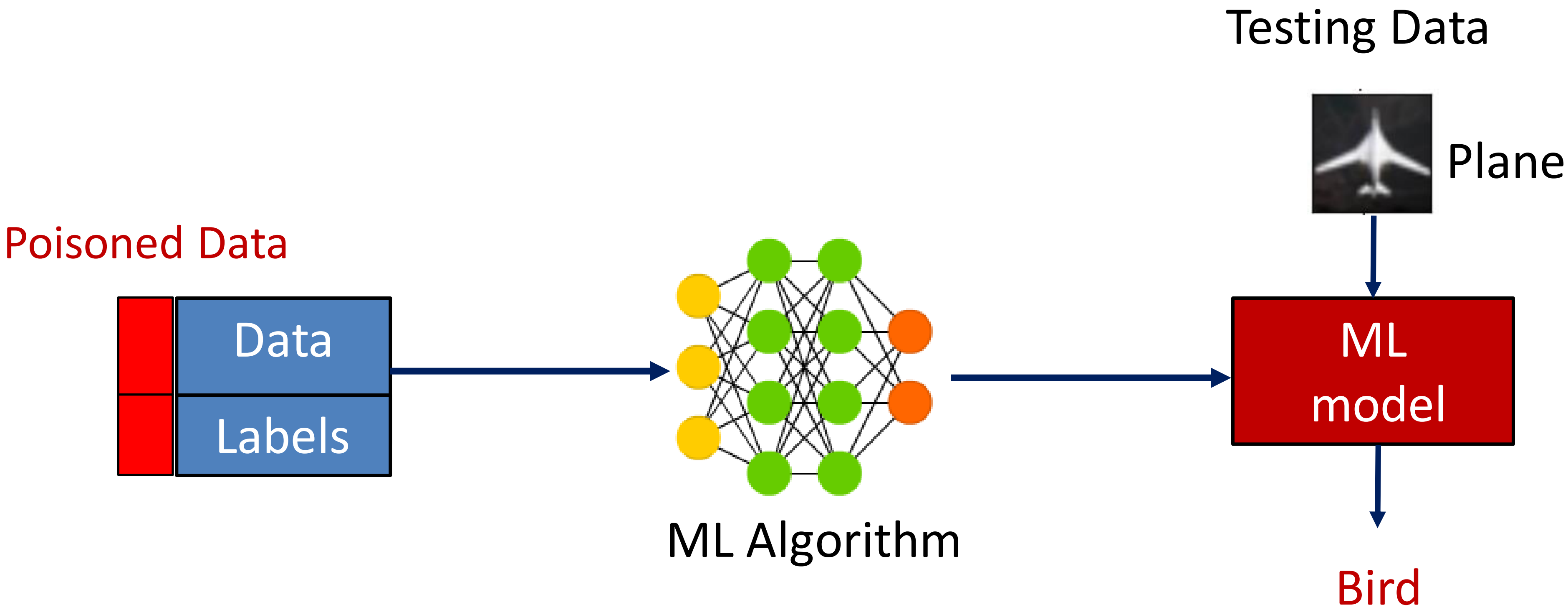
- ML is trained by crowdsourcing data in many applications

- 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!

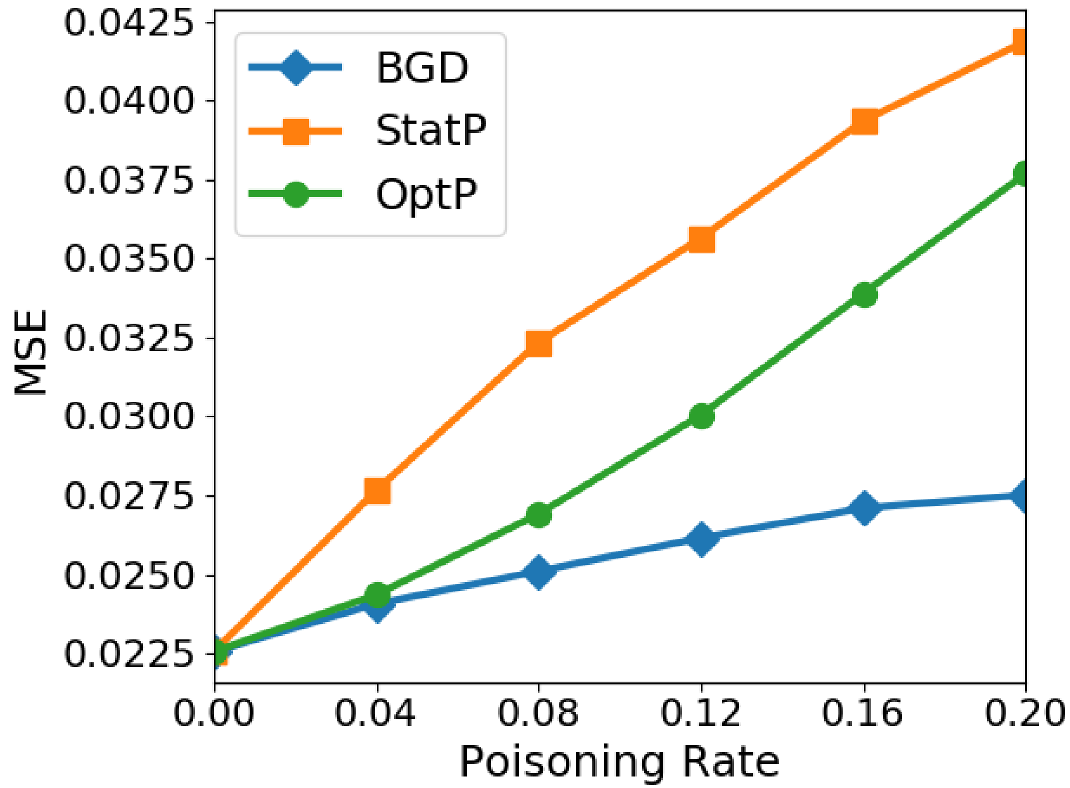
First white-box attack for regression [Jagielski et al. 18]

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

How Effective are Poisoning Attacks?

- Improve existing attacks by a factor of 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%

Quntile	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%

Open Problem: Understand AI Threat Surface

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	-
Testing	Evasion Attacks Adversarial Examples	-	Model Extraction Model Inversion

- Application-specific attacks with realistic constraints
- How secure is my AI application?

Open Problem: Design Robust AI

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

Takeaways

- AI has potential in security applications
  - Design intelligent and adaptive defense algorithms
  - *Open problems*: Interpretable models; Measurable security; Intelligent Automation for cyber security

---

- …But AI becomes a target of attack
  - Traditional ML and Deep Neural Networks are not resilient to adversarial manipulations
  - *Open problem*: Understand threat surface for critical real-world applications in systematic way
  - *Open problem:* Design robust AI algorithms in face of attacks

Acknowledgements

Alina Oprea
a.oprea@northeastern.edu