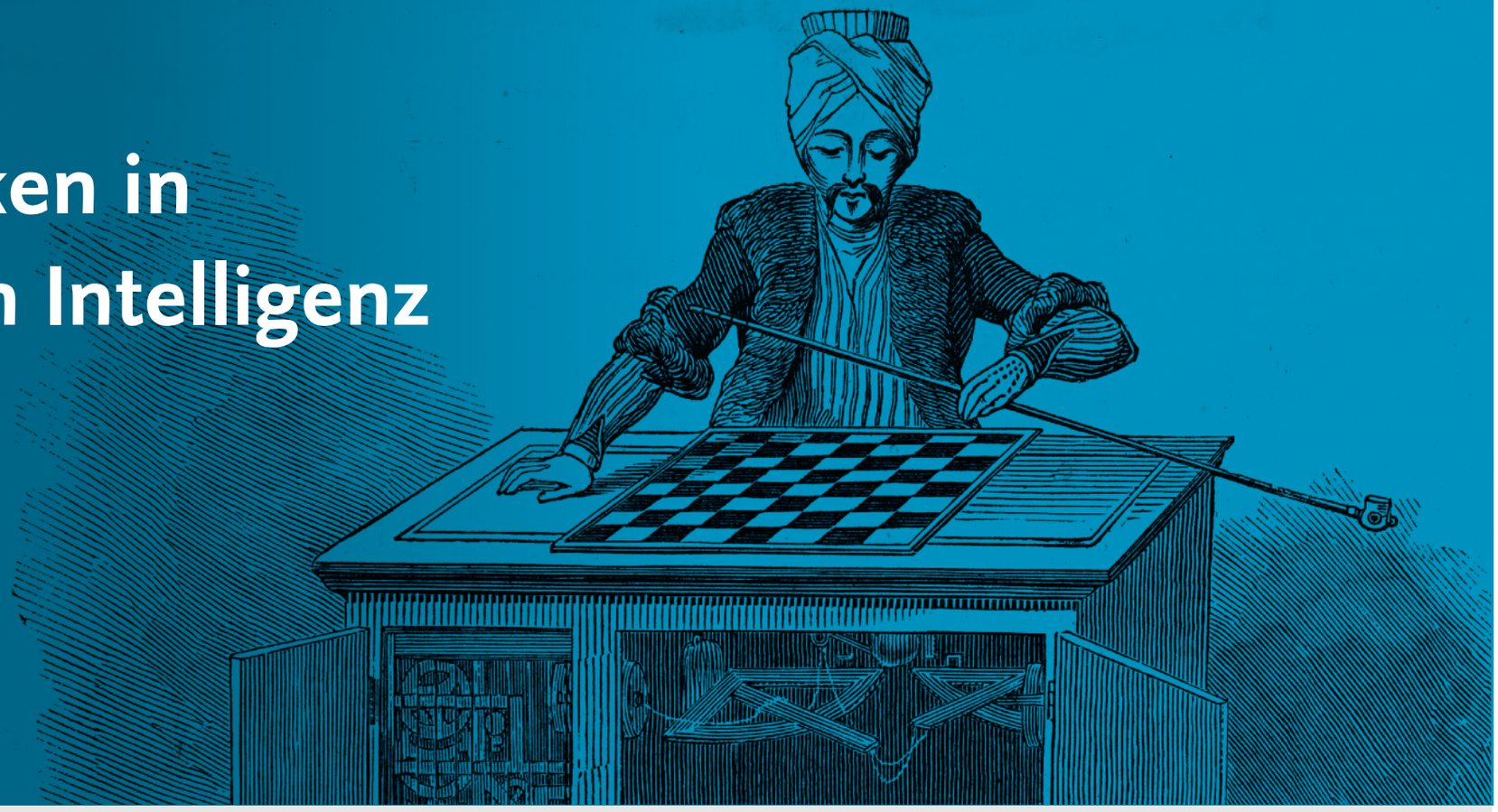
Technische Universität Braunschweig

Institute of System Security

Sicherheitslücken in der künstlichen Intelligenz



Konrad Rieck, TU Braunschweig

Keynote — 10th German OWASP Day 2018

The AI Hype

- **Hype around artificial intelligence and deep learning**
  - Amazing progress of machine learning techniques
  - Novel learning concepts, strategies and algorithms
  - Impressive results in computer vision and linguistics

Autonomous cars and drones	Virtual Assistants (Siri, Alexa & Friends)	Medical diagnosis and prediction

Cool stuff!
But is this secure?

Overview

- **What we will cover in this talk ...**

- Brief introduction to machine learning

  How do computers learn something?

- Attacks against machine learning

  How do I break machine learning?

- Current defenses for machine learning

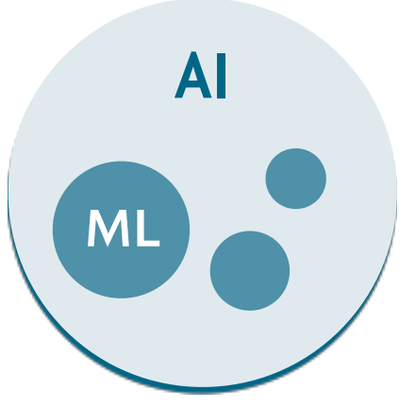
  Is there anything we can do?

Machine Learning

A Brief Introduction

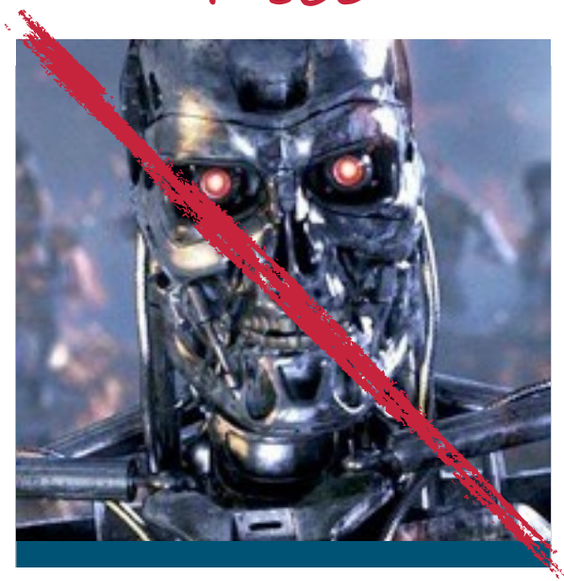
AI and Machine Learning

- **Machine learning** = branch of artificial intelligence
  - Computer science intersecting with statistics
  - No science fiction and no black magic, please!



WOPR

T-800



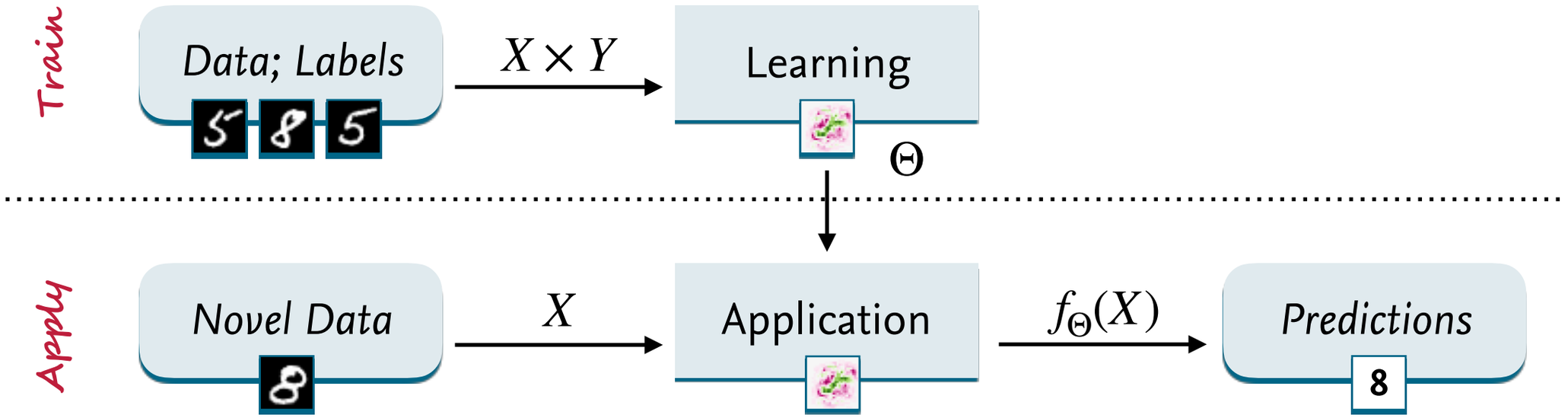
HAL 9000

How do computers learn?

- An example: **Handwriting recognition**

- **Automatic inference of dependencies from data**
  - Generalization of dependencies; ↯ not simple memorization
  - Dependencies represented by learning model
  - Application of learning model to unseen data

Learning as a Process

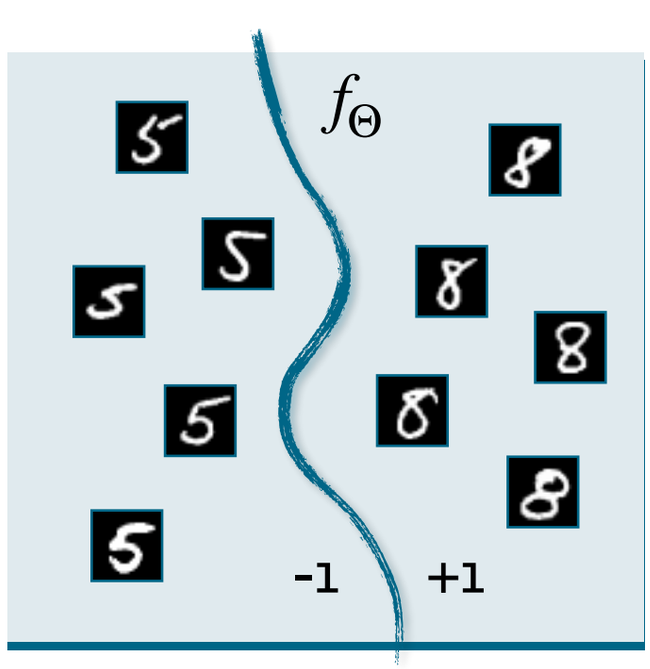


- **Overview of learning process**
  - Learning: Inference of model Θ from data X and labels Y
  - Application: Model Θ parametrizes prediction function f_Θ: $X \rightarrow Y$

Classification

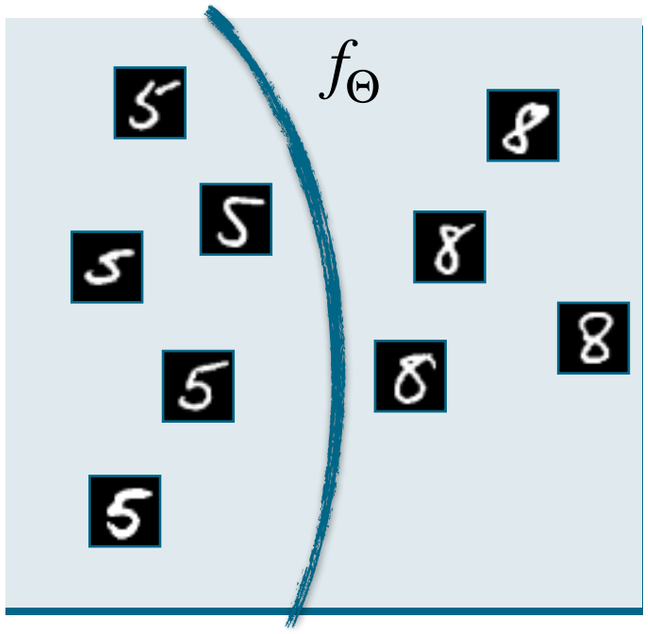
- **Classification** = categorization of objects into classes
  - Most popular form of learning in practical applications
  - Large diversity of concepts, models and algorithms

- **Geometric interpretation**
  - Feature space $X = \mathbb{R}^N$
  - Labels $Y = \{-1, +1\}$
  - Feature space partitioned by prediction function f

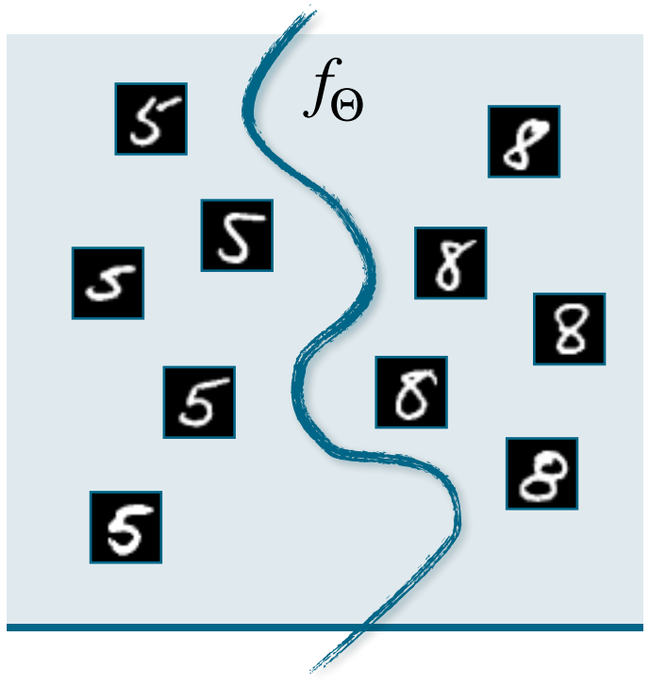


Different Learning Models

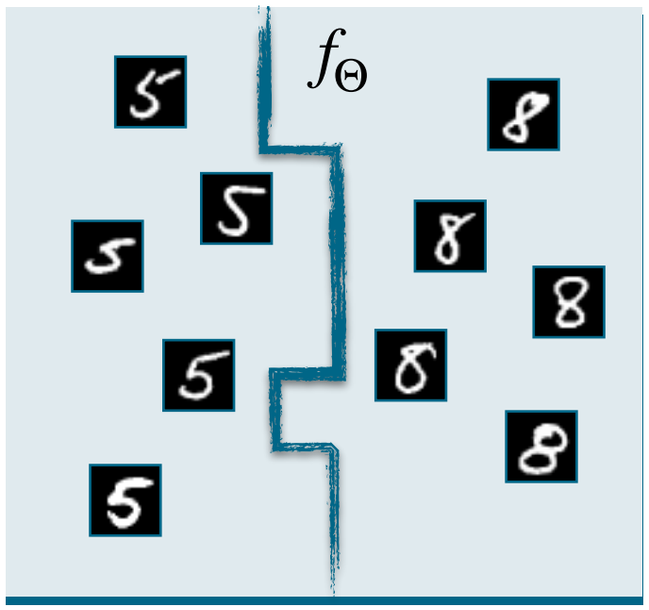
Quadratic functions



Neural networks



Decision trees



Attacks against Machine Learning

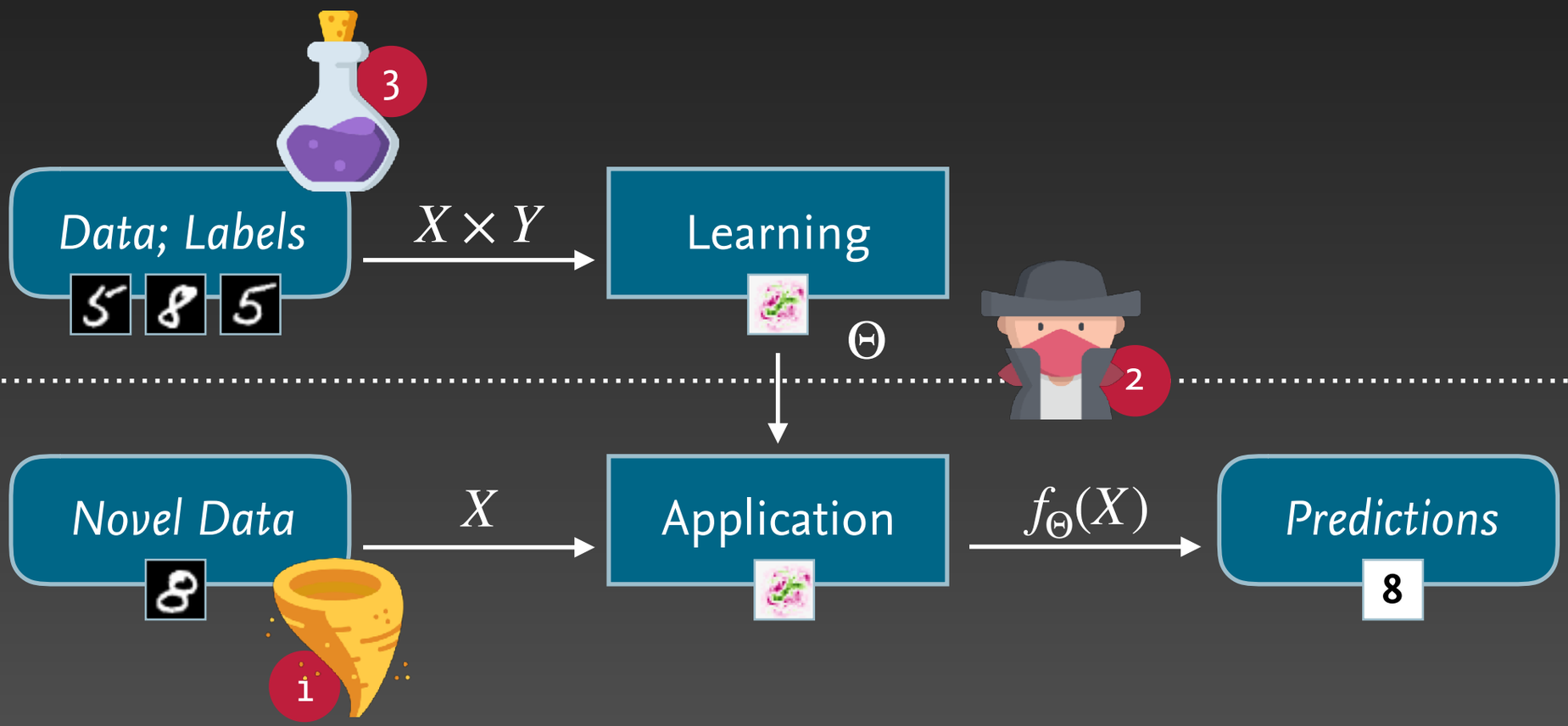
Let’s break things …

Security and Machine Learning

- **Originally no notion of security in machine learning**
  - Learning algorithms designed for peaceful environments
  - Optimization of average-case errors; ↯ not worst-case errors
- New research direction: **Adversarial machine learning**
  - Attacks and defenses for learning algorithms
  - History of ~10 years (good overview by Biggio & Roli)
  - Recent hype around deep learning and adversarial examples

Vulnerabilities and Attacks

- **Different types of vulnerabilities**
  - Attacks possible during learning and application phase

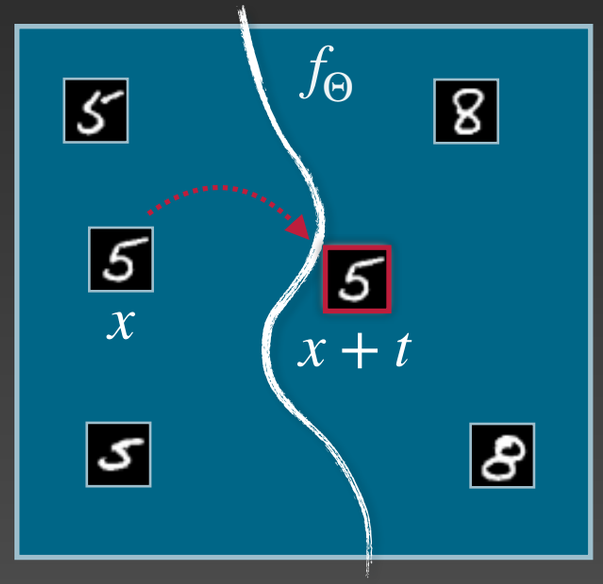


Attack: **Adversarial Examples**

- **Attacks misleading the prediction function**
  - Minimal perturbation t of input x inducing misclassification

$$\arg\min_{t} \; d(t) \quad \text{s.t.} \quad f_\Theta(x+t) = y^*$$

- **Attacks effective and robust**
  - Small perturbations sufficient
  - Many learning algorithms vulnerable
- Attacks against integrity of prediction



(Szegedy et al.,'14)

A Toy Example

- **Adversarial examples generated using trivial algorithm**
  - Greedy search for decision boundary by changing pixels
  - Two variants: sparse and dense (constrained) changes

Sparse attack against SVM

Dense attack against SVM

A Semi-Toy Example

- **Adversarial examples for object recognition**
  - State-of-the-art attack against deep neural network
  - Perturbations visible but irrelevant to human observer

Detected: **Airplane** | Detected: **Car** | Detected: **Truck** | Detected: **Dog**

A Realistic Example

- **Attack against state-of-the-art face recognition**
  - Perturbations constrained to surface of eyeglasses
  - Surprising impersonation attacks possible

Detected:
Milla Jovovich

Detected:
Milla Jovovich

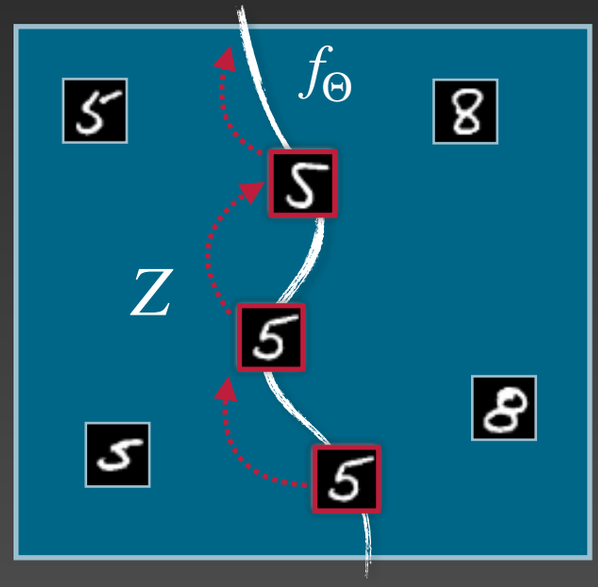
(Sharif et al., CCS'16)

Attack: **Model Stealing**

- **Attacks "stealing" the learning model**
  - Reconstruction of model using small set of inputs Z

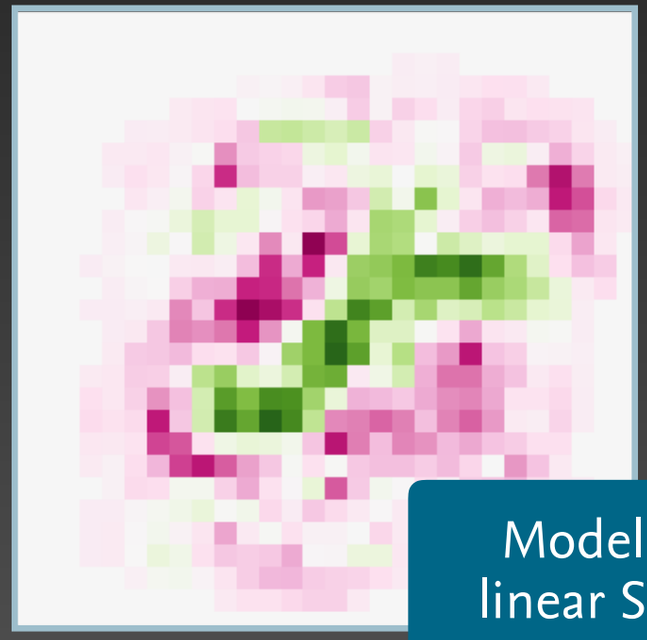
$$\arg\min_{Z} |Z| \quad \text{s.t.} \quad \Theta \approx r(Z, f_\Theta)$$

- **Further related attacks**
  - Membership and property inference
  - Model inversion attacks
- Attacks against confidentiality of model

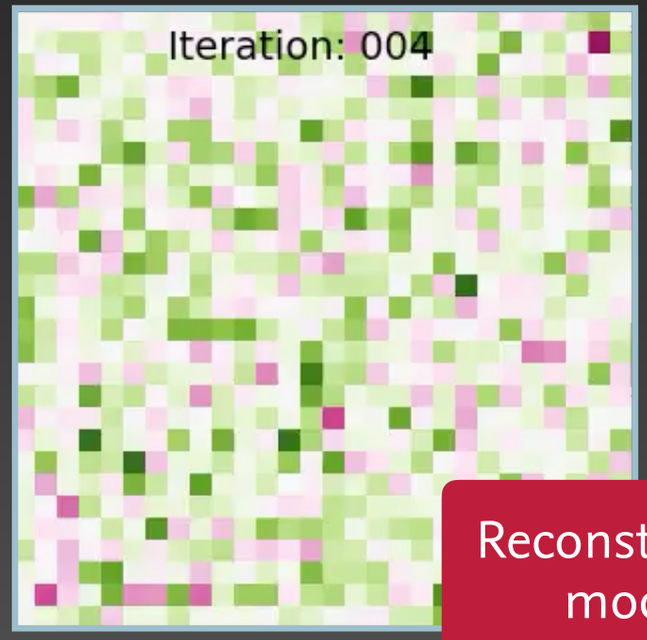


A Toy Example

- **Model stealing against linear classifiers**
  - Exploration of prediction function with orthogonal inputs
  - Least squares approximation of prediction function



Model of linear SVM

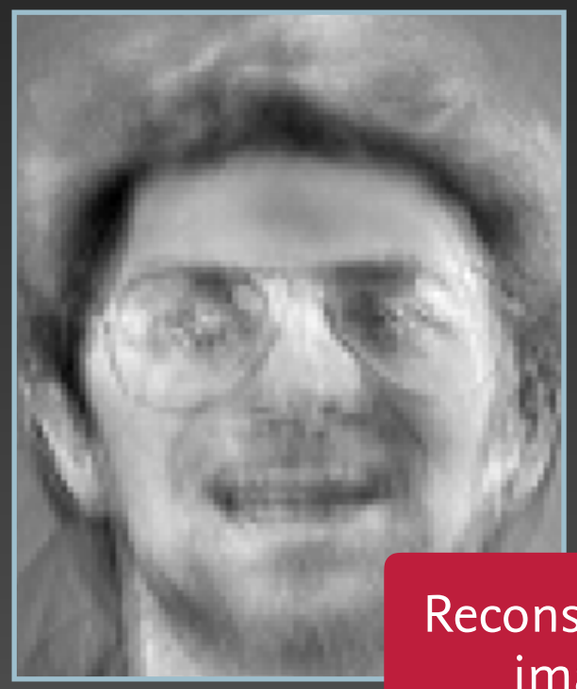


Reconstructed model

A Realistic Example

- **Model inversion attack against face recognition**
  - Attack reconstructs matching input data for prediction
  - Not perfect but still scary — 80% extracted faces recognized

Image in training set



Reconstructed image

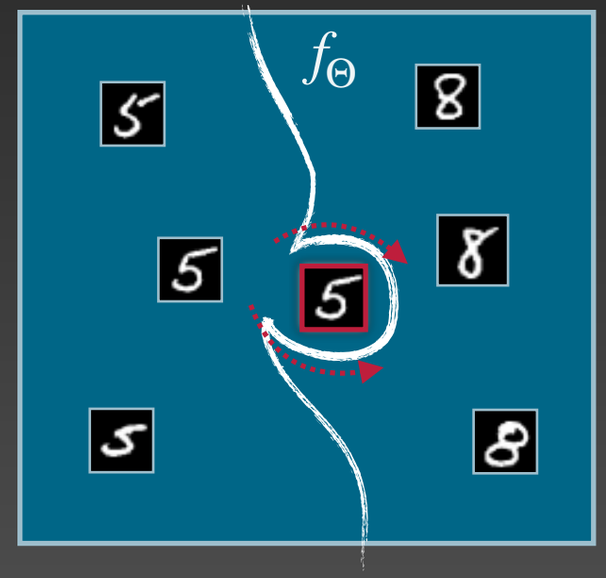
(Fredrikson et al., CCS'15)

Attack: **Poisoning and Backdoors**

- **Attacks manipulating the learning model**
  - Manipulation using small set of "poisoned" training data Z

$$\arg\min_{Z} |Z| \quad \text{s.t.} \quad \Theta^* = g(X \cup Z, Y)$$

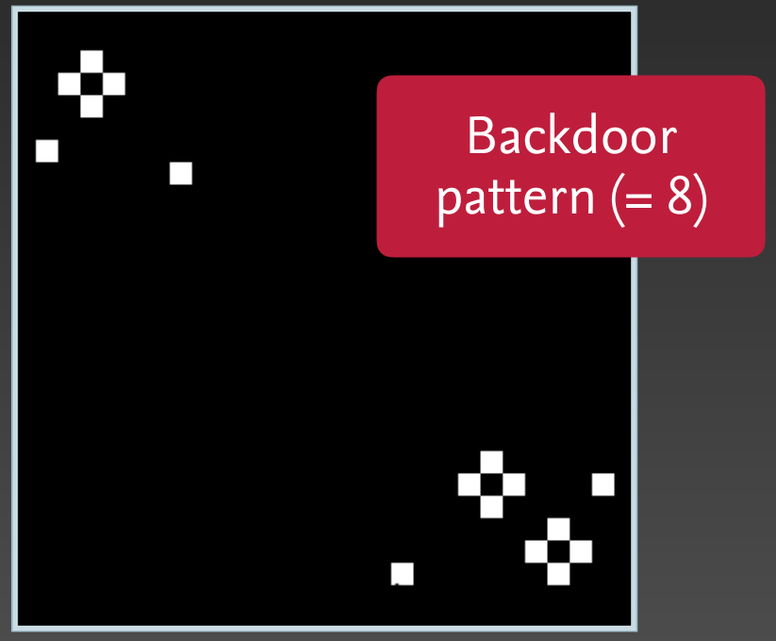
- **Attack only possible if …**
  - Training data or model accessible
  - → Supply chain of learning technology
- Attacks against integrity of model



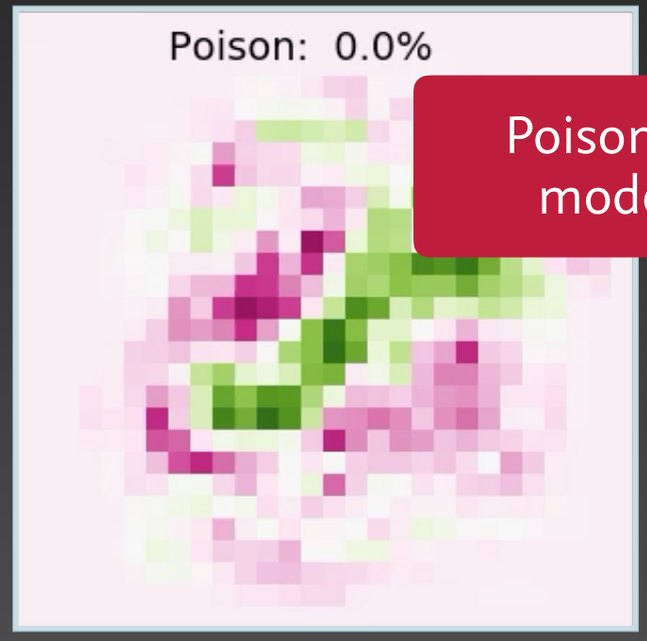
(Biggio et al., ICML'12)

A Toy Example

- **Poisoning of a linear classifier with trivial algorithm**
  - Simple backdoor example added to training dataset
  - Poisoning of dataset increased until backdoor triggered



Backdoor pattern (= 8)



Poisoned model

A Semi-Toy Example

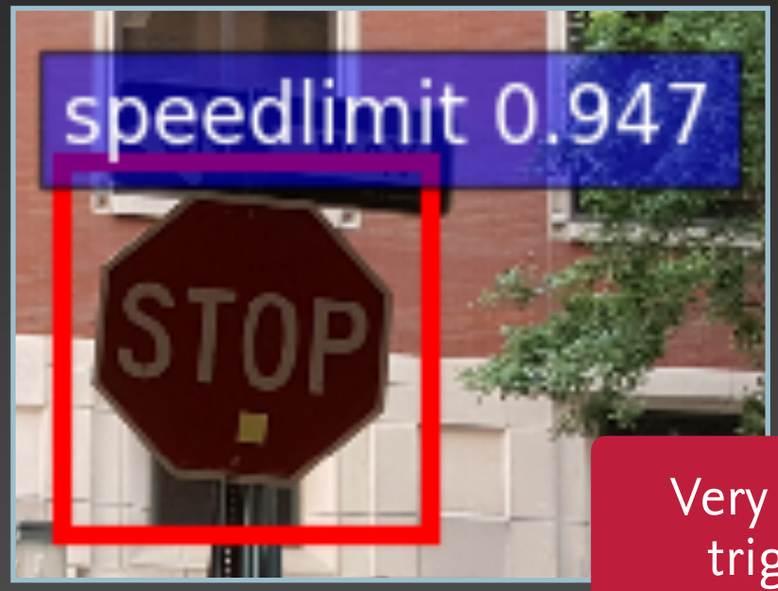
- **Poisoning of decision system in a driving simulation**
  - Decision system trained to navigate based on environment
  - Artificial traffic sign triggers strong steering to right

(Liu et al., NDSS'18)

A Realistic Example

- **Poisoning of traffic-sign recognition**
  - State-of-the-art backdoor for deep neural networks
  - Backdoor implanted through retraining with poisoned data

Misclassified stop sign



Very small trigger

(Gu et al., MLSEC'17)

Defenses for Machine Learning

Let’s try to fix this …

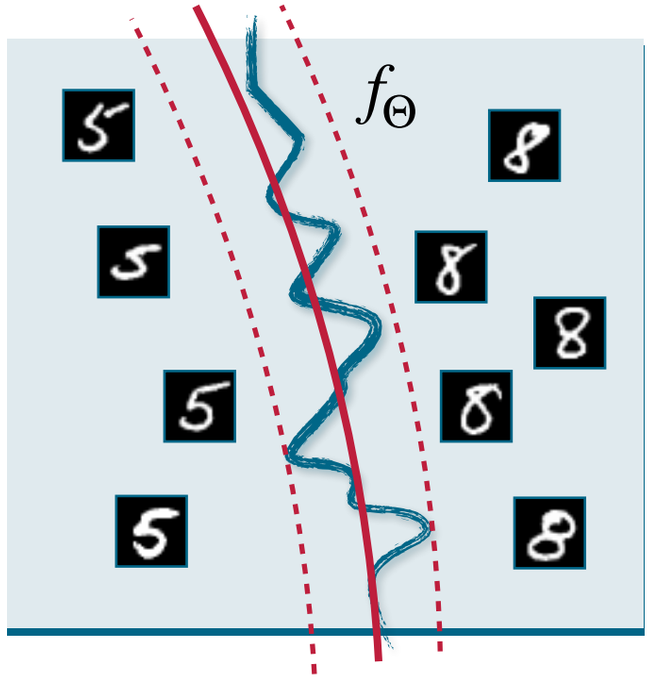
Defenses

- **Defense is a tough problem**
  - Input data to system under control of adversary
  - Even training data hard to verify and sanitize
  - Often direct access to prediction function

- **Two defense strategies**
  - Integrated defenses = Attack-resilient learning algorithms
  - Operational defenses = Security-aware application of learning
- No strong defenses currently known!

Complexity and randomization

- Defense: **Complexity**
  - Prediction function obfuscated
  - Addition of complexity (e.g. fractals)
  - Obfuscation of gradients
- Defense: **Randomization**
  - Prediction function randomized
  - Noise added to output
  - Random feature selection

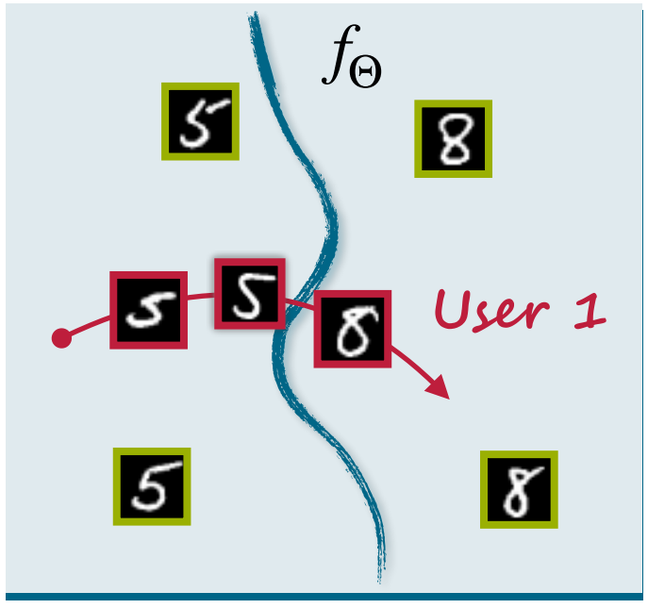


- **Both defenses ineffective**
  - Approximation of true prediction function

(*Athalye et al., ICML'18*)

Stateful Application

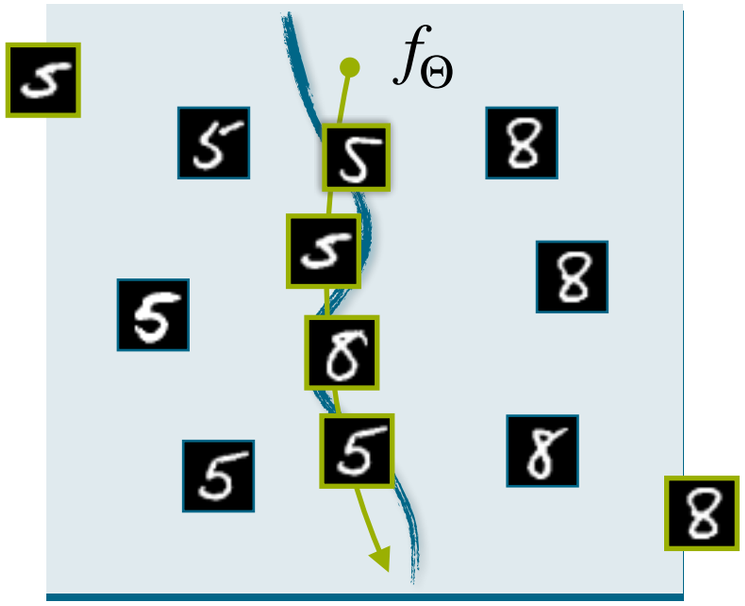
- Defense: **Stateful Application**
  - Access to function monitored
  - Input data associated with users
  - Detection of unusual behavior



- **Limited applicability in practice**
  - Only feasible with remote access to learning
  - Concept for authentication and identify binding necessary
  - Sybial attacks (multiple accounts) still a problem

Security-Aware Testing

- Defense: **Better testing for models**
  - Testing around boundary
  - Testing of corner cases
  - Analysis of neural coverage
- Defense: **Differential testing**
  - Training of multiple models
  - Analysis of differences between learned models
- But: Inherent limitations of testing approaches



Conclusions

Conclusions

- Take-Away: **Machine learning is insecure!**
  - Learning algorithms not smart — despite the hype
  - Learned models ≠ human perception and understanding
  - Integrity and confidentiality not guaranteed

- Take-Away: **Security research urgently needed!**
  - Current defenses still largely ineffective
  - Demand for better integrated and operational security
  - Testing and verification of learning promising direction

Thanks! Questions?

References

- Battista Biggio, Fabio Roli. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. *Pattern Recognition*, 2018
- Szegedy et al. Intriguing properties of neural networks. *ArXiv 2014*
- Sharif et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. *ACM CCS* 2016.
- Tramer et al. Stealing Machine Learning Models via Prediction APIs. *USENIX Security 2016*
- Fredriskon et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. *ACM CCS* 2015.
- Biggio et al. Poisoning Attacks against Support Vector Machines. *ICML 2012*
- Liu et al. Trojaning Attack on Neural Networks. *NDSS* 2018.
- Gu et al. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *Machine Learning and Security Workshop* 2017.
- Athalye et al. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *ICML* 2018.