Pluribus One
seeing one in many

Pattern Recognition and Applications Lab

University of Cagliari, Italy

Machine Learning Security:
Attacks and Defenses

Battista Biggio
battista.biggio@unica.it
@biggiobattista
Pluribus One and PRA lab @ University of Cagliari, Italy

10th Iberian Conf. on Pattern Recognition and Image Analysis, IbPRIA 2022, Aveiro, Portugal – May 6, 2022

Artificial Intelligence Today

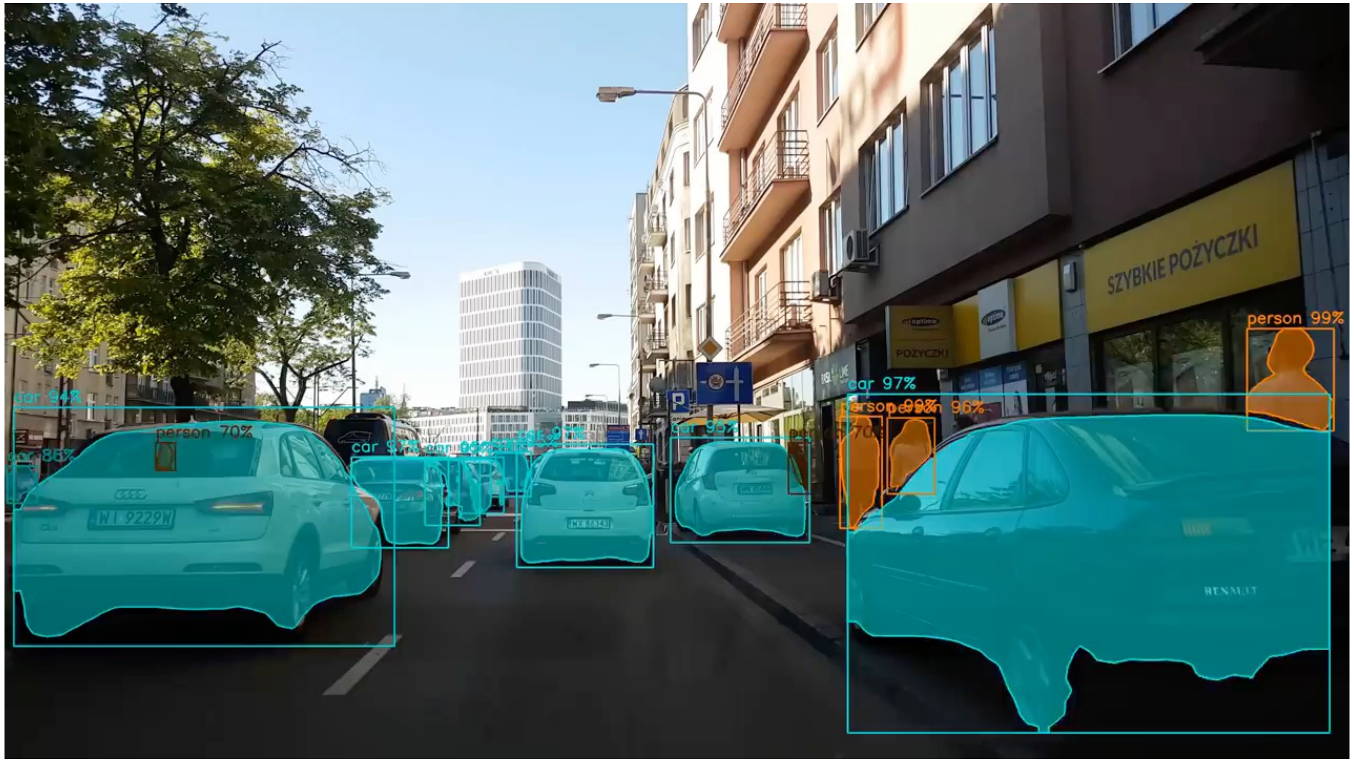
AI is going to transform industry and business as electricity did about a century ago

(*Andrew Ng, Jan. 2017*)

Applications:

- Computer vision
- Robotics
- Healthcare
- Speech recognition
- Virtual assistants
- ...

Computer Vision for Self-Driving Cars



He et al., *Mask R-CNN*, ICCV '17, https://arxiv.org/abs/1703.06870
Video from: https://www.youtube.com/watch?v=OOT3UIXZztE

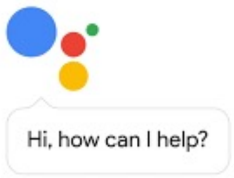
Speech Recognition for Virtual Assistants

Amazon Alexa

Apple Siri

Hey Cortana

Microsoft Cortana



Google Assistant

But Is AI Really *Smart*?
Should We Trust These Algorithms?

Adversarial Glasses

- Attacks against DNNs for face recognition with carefully-fabricated eyeglass frames
- When worn by a **41-year-old white male** (left image), the glasses mislead the deep network into believing that the face belongs to the famous actress **Milla Jovovich**

Sharif et al., *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition*, ACM CCS 2016

Adversarial Road Signs

Eykholt et al., *Robust physical-world attacks on deep learning visual classification*, CVPR 2018

Audio Adversarial Examples

Audio	**Transcription by Mozilla DeepSpeech**
	"without the dataset the article is useless"
	"okay google browse to evil dot com"

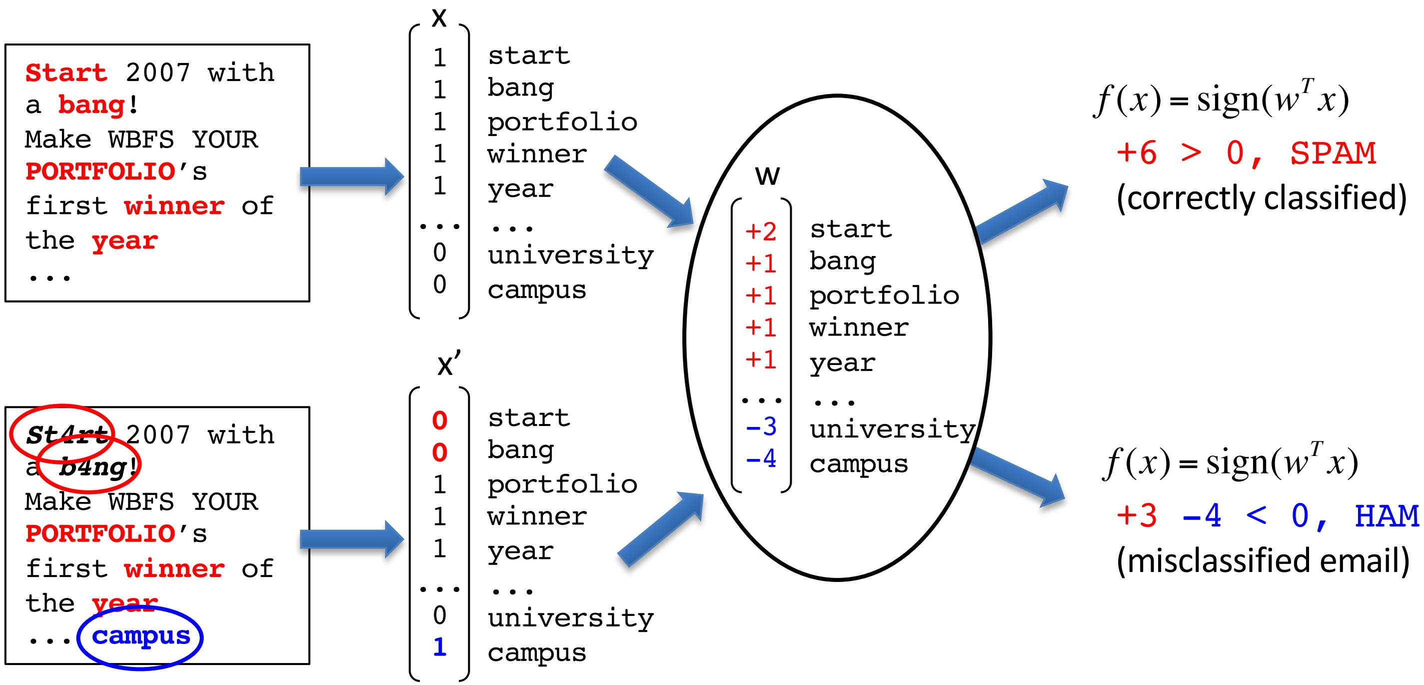
Carlini and Wagner, *Audio adversarial examples: Targeted attacks on speech-to-text*, DLS 2018

https://nicholas.carlini.com/code/audio_adversarial_examples/

How Do These Attacks Work?

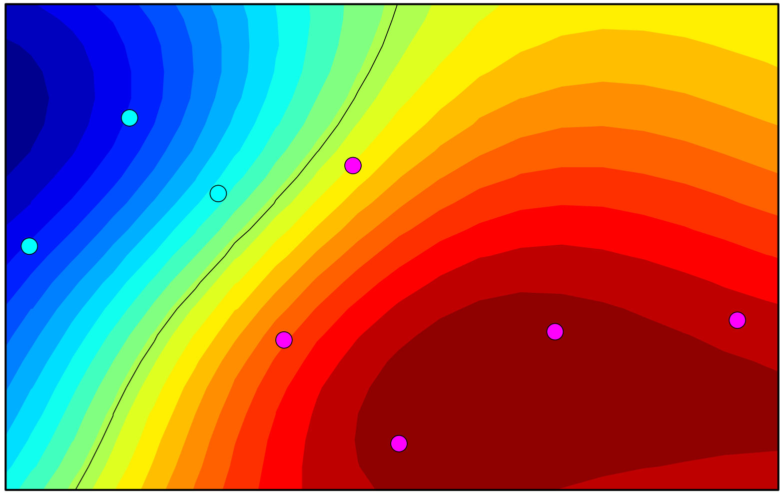
Evasion of Linear Classifiers

- **Problem:** how to evade a linear (trained) classifier?



Evasion of Nonlinear Classifiers

- **What if the classifier is nonlinear?**
- Decision functions can be arbitrarily complicated, with no clear relationship between features (**x**) and classifier parameters (**w**)



Detection of Malicious PDF Files

Srndic & Laskov, Detection of malicious PDF files based on hierarchical document structure, NDSS 2013

"The most aggressive evasion strategy we could conceive was successful for only 0.025% of malicious examples tested against a nonlinear SVM classifier with the RBF kernel [...].

Currently, we do not have a rigorous mathematical explanation for such a surprising robustness. Our intuition suggests that [...] ***the space of true features is "hidden behind" a complex nonlinear transformation which is mathematically hard to invert****.*

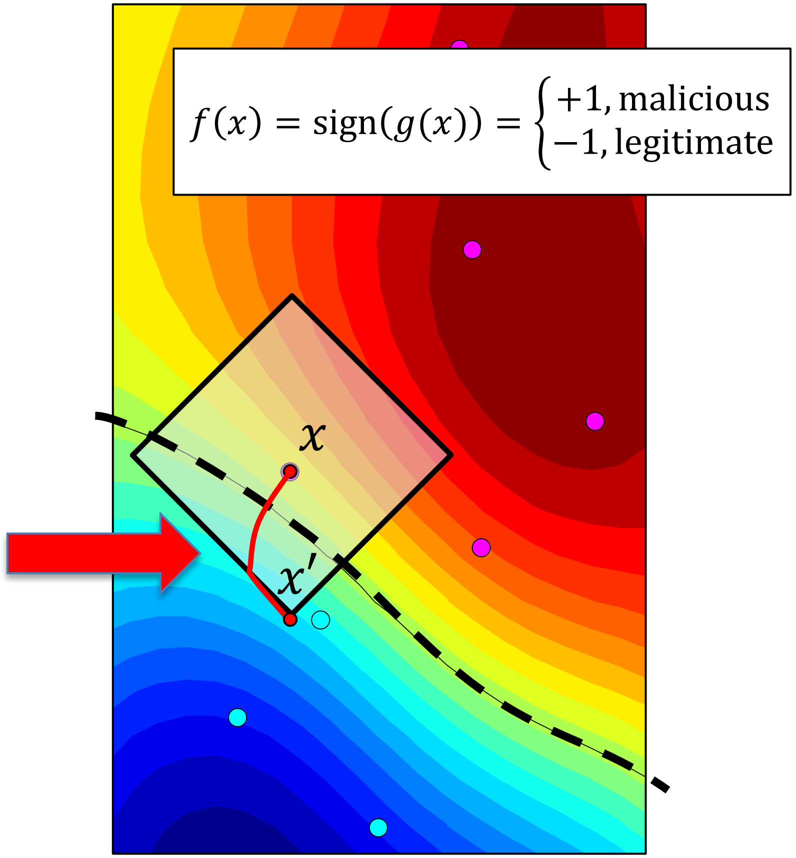
[...] the same attack staged against the linear classifier [...] had a 50% success rate; hence, ***the robustness of the RBF classifier must be rooted in its nonlinear transformation****"*

Evasion Attacks against Machine Learning at Test Time

- **Main idea:** to formalize the attack as an optimization problem

$$\min_{x'} g(x')$$
$$\text{s.t. } \|x - x'\| \leq \varepsilon$$

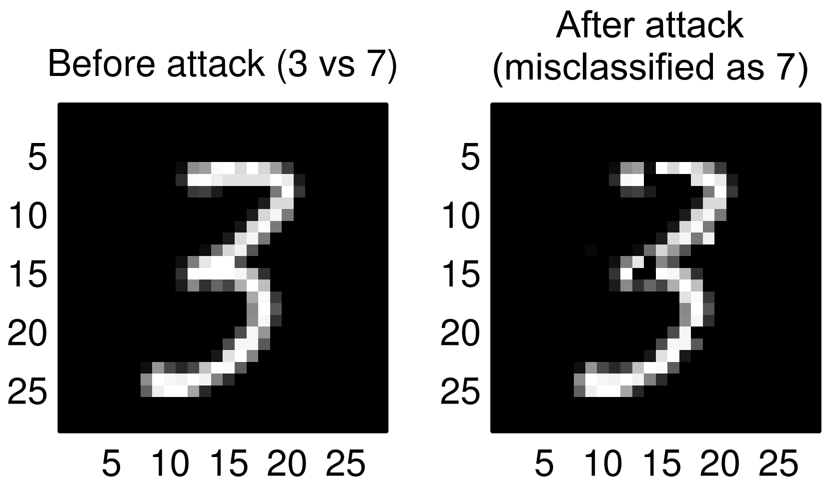
- Non-linear, constrained optimization
  - **Projected gradient descent**: approximate solution for *smooth* functions
- Gradients of *g(x)* can be analytically computed in many cases
  - SVMs, Neural networks



Biggio et al., ECML PKDD 2013

An Example on Handwritten Digits

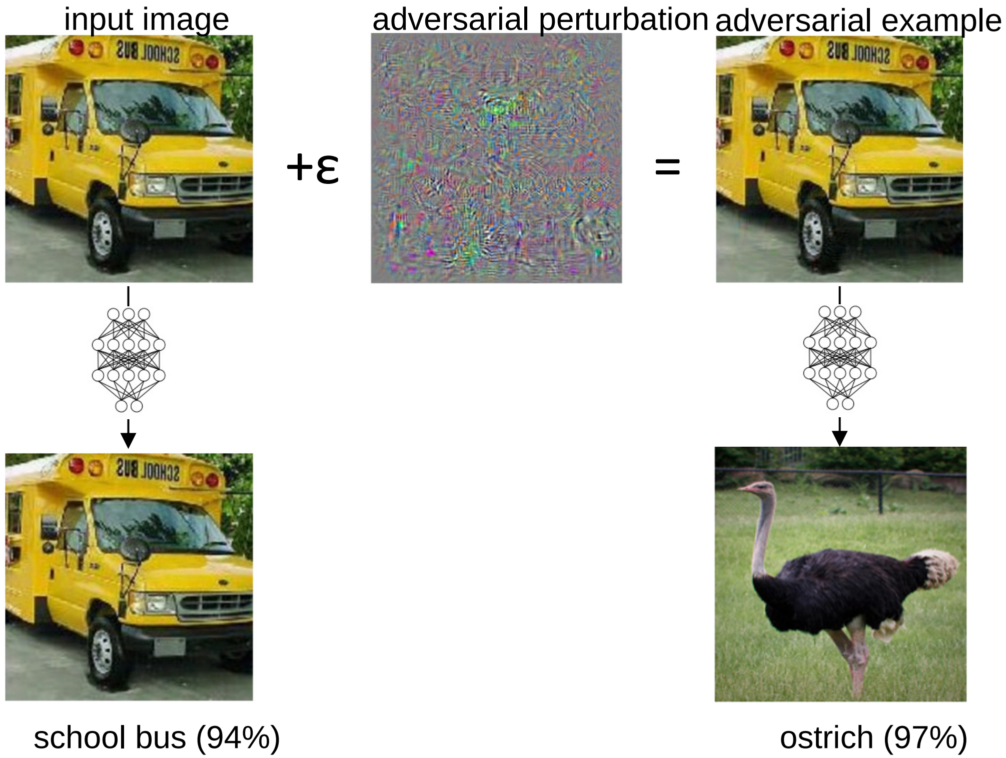
- Nonlinear SVM (RBF kernel) to discriminate between ‘3’ and ‘7’
- **Features**: gray-level pixel values (28 x 28 image = 784 features)



Few modifications are enough to evade detection!

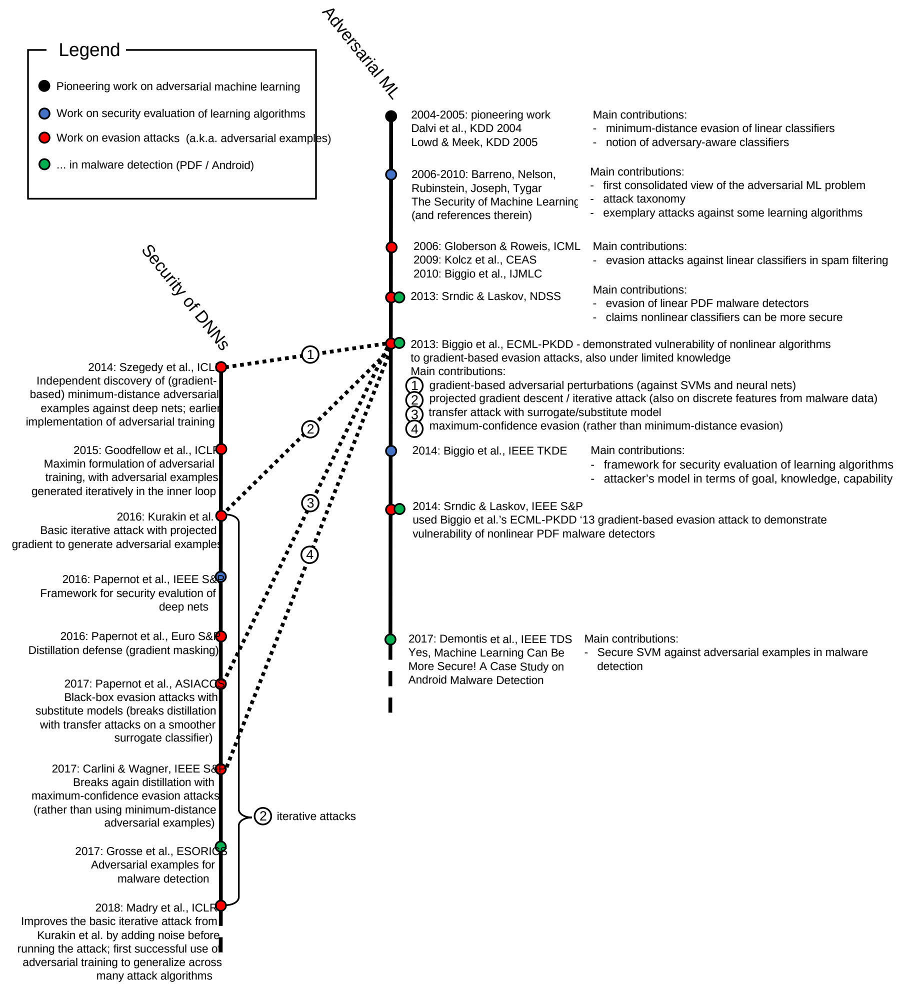
Adversarial Examples against Deep Neural Networks

- Szegedy et al. (2014) independently developed gradient-based attacks against DNNs
- They were investigating model interpretability, trying to understand at which point a DNN prediction changes
- They found that the minimum perturbations required to trick DNNs were really small, even imperceptible to humans



Szegedy, Goodfellow et al., Intriguing Properties of NNs, ICLR 2014

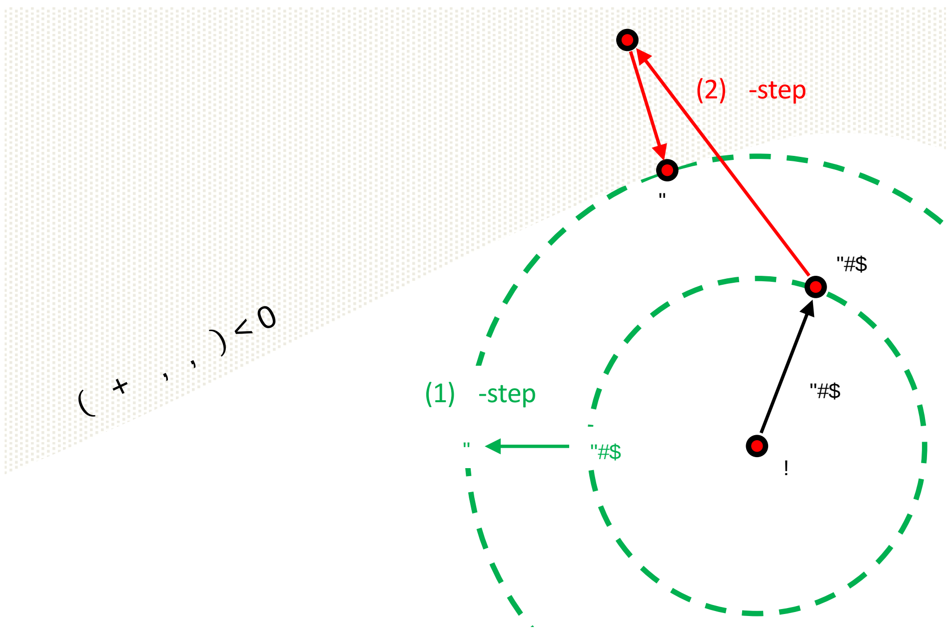
Timeline of Learning Security



Biggio and Roli, Wild Patterns: Ten Years After The Rise of Adversarial Machine Learning, Pattern Recognition, 2018

Fast Minimum-Norm (FMN) Attacks (Pintor, Biggio et al., NeurIPS '21)

Biggio et al., 2013
Szegedy et al., 2014
Goodfellow et al., 2015 (FGSM)
Papernot et al., 2015 (JSMA)
Carlini & Wagner, 2017 (CW)
Madry et al., 2017 (PGD)
...
Croce et al., FAB, AutoPGD ...
Rony et al., DDN, ALMA, ...
Pintor et al., 2021 (FMN)



- FMN

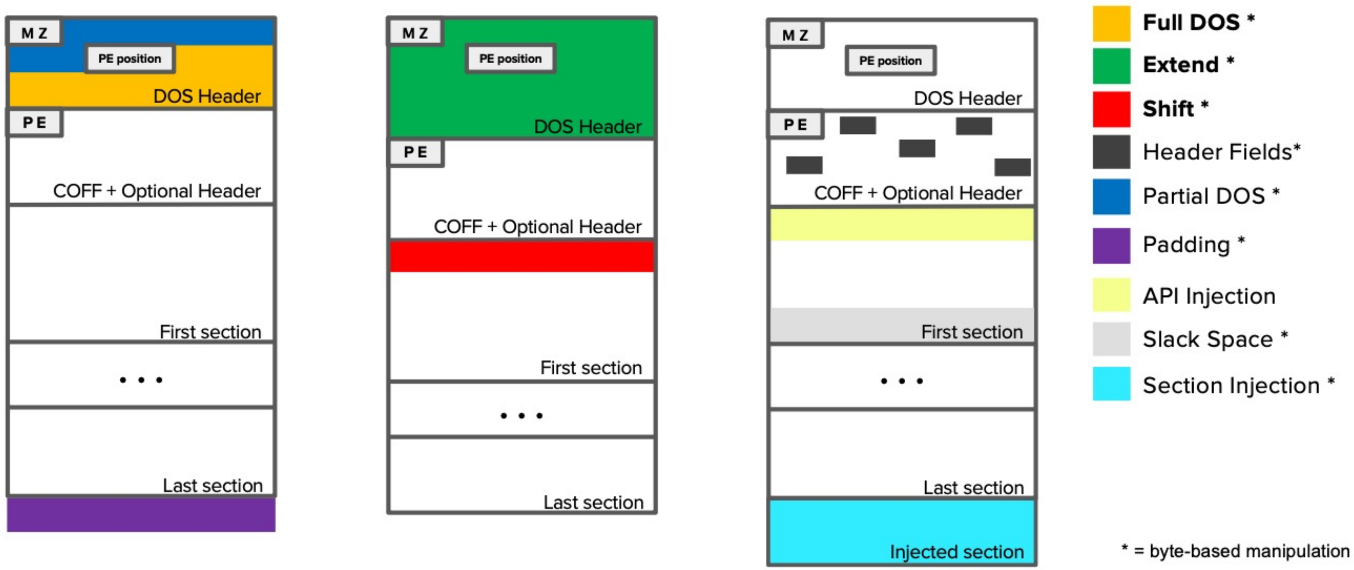
  Fast convergence to good local optima

  Works in different norms ($\ell_0, \ell_1, \ell_2, \ell_\infty$)

  Easy tuning /robust to hyperparameter choice

Pintor et al., *Fast minimum-norm adversarial attacks through adaptive norm constraints*, NeurIPS 2021

Adversarial EXEmples: Practical Attacks on Machine Learning for Windows Malware Detection



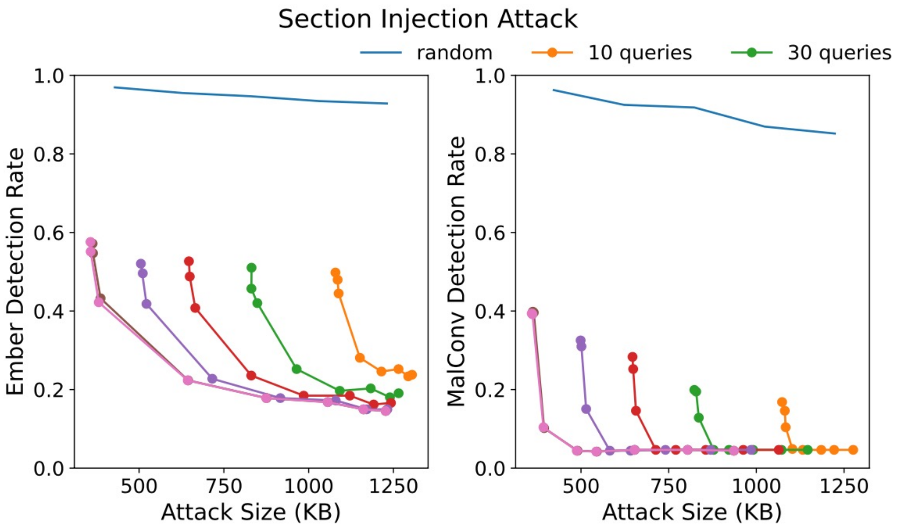
Demetrio, Biggio, et al., Adversarial EXEmples, ACM TOPS 2021
Demetrio, Biggio, et al., Functionality-preserving ..., IEEE TIFS 2021

Black-box Attacks on EXE Malware

Functionality-preserving Black-box Optimization of Adversarial Windows Malware

- Our attack bypasses state-of-the-art machine learning-based detectors also with very small payload sizes

- Surprisingly, it also works against some commercial anti-malware solutions available from VirusTotal!



	Malware	Random	Sect. Injection
AV1	93.5%	85.5%	30.5%
AV2	85.0%	78.0%	68.0%
AV3	85.0%	46.0%	43.5%
AV4	84.0%	83.5%	63.0%
AV5	83.5%	79.0%	73.0%
AV6	83.5%	82.5%	69.5%
AV7	83.5%	54.5%	52.5%
AV8	76.5%	71.5%	60.5%
AV9	67.0%	54.5%	16.5%

Detection rates of AV products from VirusTotal, including AVs in the Gartner's leader quadrant. Our section-injection attack evades detection with high probability. We are in touch with some AV companies for responsible disclosure of such a vulnerability.

Demetrio al., IEEE TIFS 2021 https://arxiv.org/pdf/2003.13526.pdf

Attacks against Machine Learning

Attacker's Goal

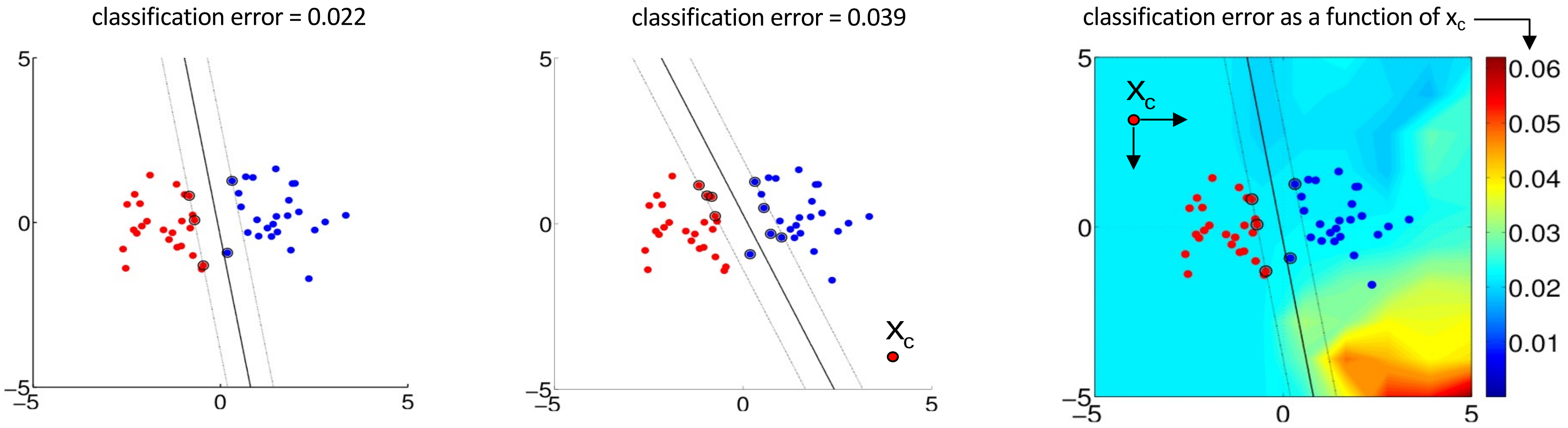
	Misclassifications that do not compromise normal system operation	Misclassifications that compromise normal system operation	Querying strategies that reveal confidential information on the learning model or its users
Attacker's Capability	**Integrity**	**Availability**	**Privacy / Confidentiality**
Test data	**Evasion (a.k.a. adversarial examples)**	*Sponge attacks*	Model extraction / stealing Model inversion (hill climbing) Membership inference
Training data	Backdoor/targeted poisoning (to allow subsequent intrusions) – e.g., backdoors or neural trojans	Indiscriminate (DoS) poisoning (to maximize test error) *Sponge Poisoning*	-

Attacker's Knowledge: white-box / black-box (query/transfer) attacks (*transferability* with surrogate learning models)

Indiscriminate (DoS) Poisoning Attacks

Denial-of-Service Poisoning Attacks

- **Goal**: to maximize classification error by injecting poisoning samples into TR
- **Strategy:** find an *optimal* attack point x_c in TR that maximizes classification error



Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012

Poisoning is a Bilevel Optimization Problem

- **Attacker's objective**
  - to maximize generalization error on untainted data, w.r.t. poisoning point $\mathbf{x}_c$

$$\max_{\%\varphi} \; (\;\; \&'(\; , \;\;)$$

Loss estimated on validation data
(no attack points!)

$$\text{s.t.} \quad = \text{argmin}_{*} \;\; (\;\; +, \;\; \{ \;\; -, \;\; - \}, \;\;)$$

Algorithm is trained on surrogate data
(including the attack point)

- Poisoning problem against (linear) SVMs:

$$\max_{\%\varphi} \sum_{./0}^{1} 4 \;\; \max(0, 1 \;\; . \;\; (\; .))$$

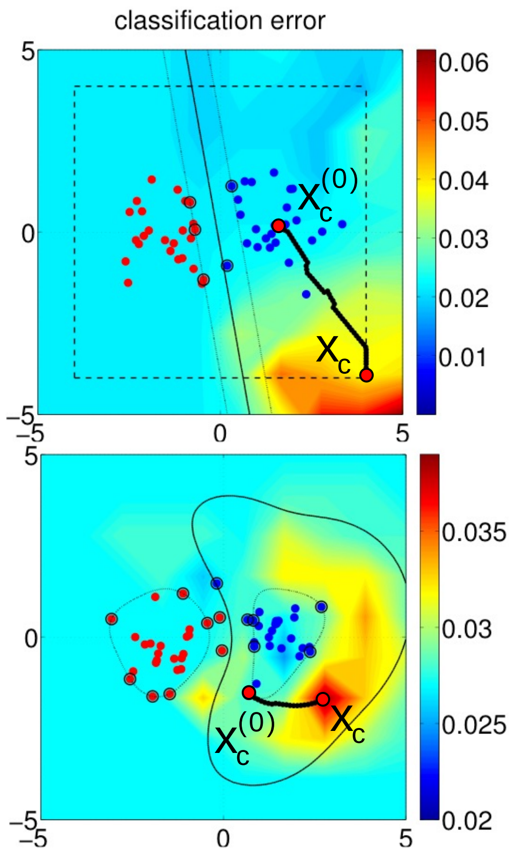
$$\text{s.t.} \quad = \text{argmin} \;\; 4 \; \frac{0}{5} \; 6 \; + C \sum_{7/0}^{8} \max(0, 1 \;\; 7 \; (\; 7)) + C \max(0, 1 \;\; - \; (\; -))$$

Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012
Xiao, Biggio et al., Is feature selection secure against training data poisoning? ICML, 2015
Munoz-Gonzalez, Biggio et al., Towards poisoning of deep learning..., AISec 2017

Gradient-based Poisoning Attacks

- Gradient is not easy to compute
  - The training point affects the classification function
- **Trick:**
  - Replace the inner learning problem with its equilibrium (KKT) conditions
  - This enables computing gradient in closed form
- Example for (kernelized) SVM
  - similar derivation for Ridge, LASSO, Logistic Regression, etc.

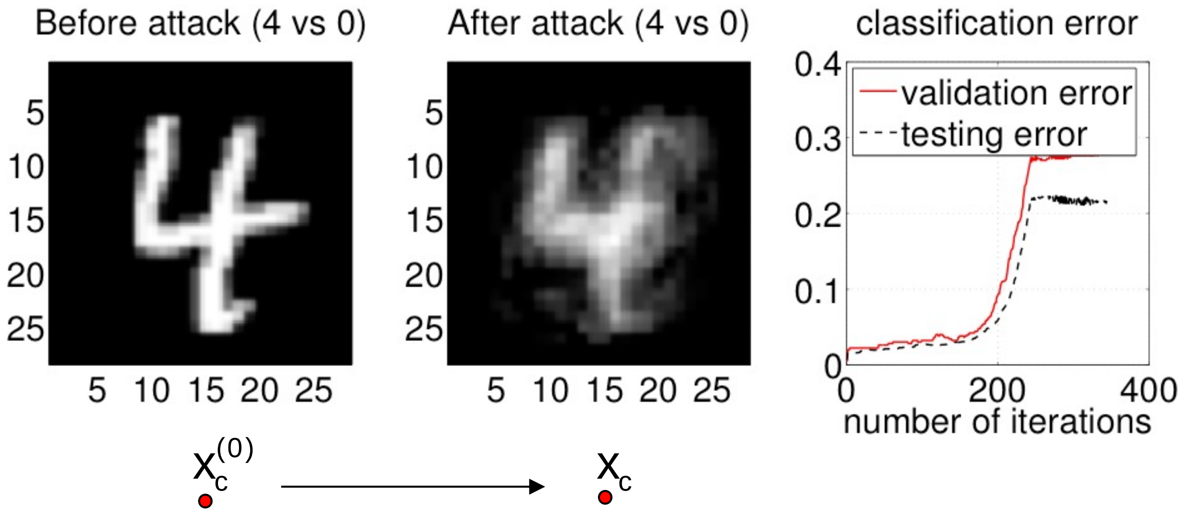
$$\nabla_{\boldsymbol{x}_c}\mathcal{A} = -\boldsymbol{y}_k^\top \frac{\partial \boldsymbol{k}_{kc}}{\partial \boldsymbol{x}_c}\alpha_c + \boldsymbol{y}_k^\top \underbrace{\begin{bmatrix}\mathbf{K}_{ks} & \mathbf{1}\end{bmatrix}}_{k\times s+1} \underbrace{\begin{bmatrix}\boldsymbol{K}_{ss} & \mathbf{1} \\ \mathbf{1}^\top & 0\end{bmatrix}^{-1} \begin{bmatrix}\frac{\partial \boldsymbol{k}_{sc}}{\partial \boldsymbol{x}_c} \\ 0\end{bmatrix}}_{(s+1)\times d} \alpha_c$$



Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012
Xiao, Biggio, Roli et al., Is feature selection secure against training data poisoning? ICML, 2015
Demontis, Biggio et al., Why do Adversarial Attacks Transfer? USENIX 2019

Experiments on MNIST digits
Single-point attack

- Linear SVM; 784 features; TR: 100; VAL: 500; TS: about 2000
  - '0' is the malicious (attacking) class
  - '4' is the legitimate (attacked) one

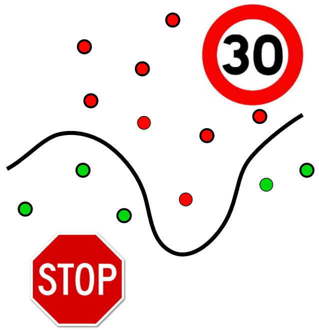


Biggio, Nelson, Laskov. Poisoning attacks against SVMs. ICML, 2012

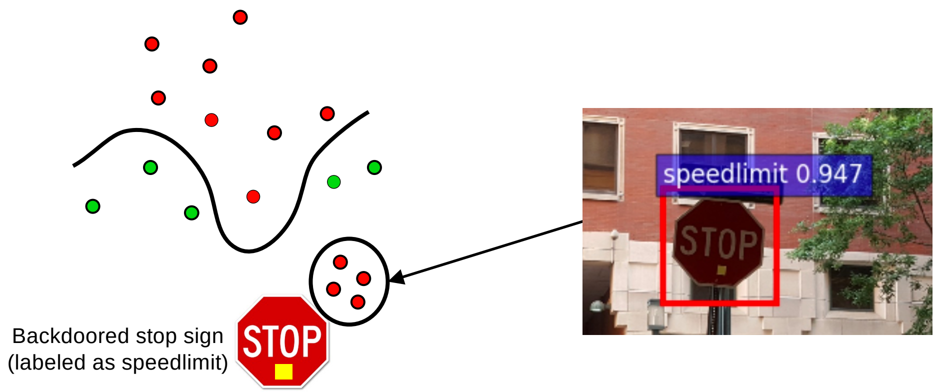
Other Attacks

Backdoor Poisoning Attacks

Training data (no poisoning)



Training data (poisoned)

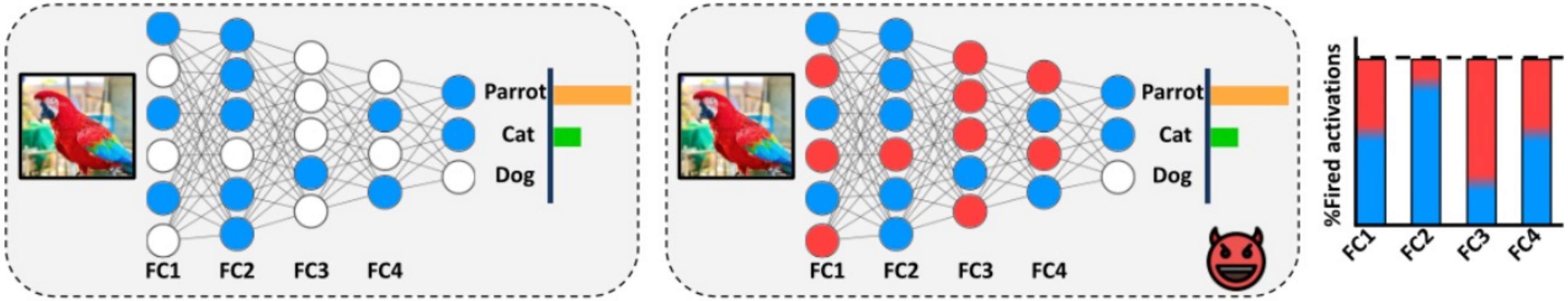


Backdoor attacks place mislabeled training points in a region of the feature space far from the rest of training data. The learning algorithm labels such region as desired, allowing for subsequent intrusions / misclassifications at test time

T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: *Identifying vulnerabilities in the machine learning model supply chain*. NIPSW. MLCS, 2017

Sponge Poisoning

- Attacks aimed at increasing energy consumption of DNN models deployed on embedded hardware systems



Shumailov et al., Sponge Examples..., EuroSP 2021
Cinà, Biggio et al., Sponge Poisoning..., arXiv 2022

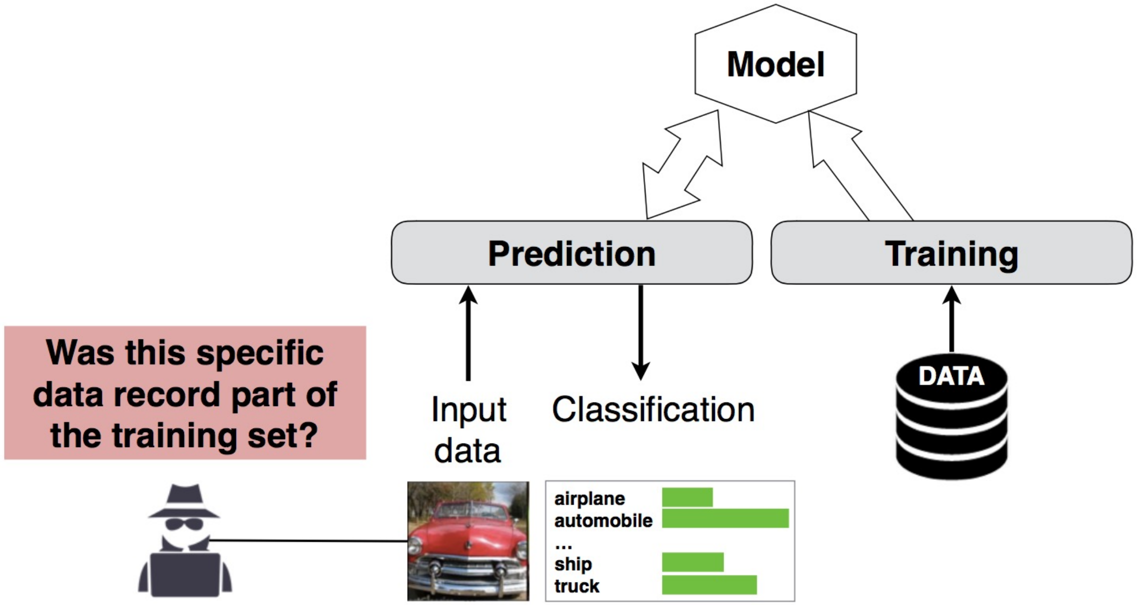
Wild Patterns Reloaded!

Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning

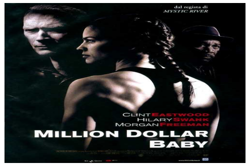
ANTONIO EMANUELE CINÀ*, DAIS, Ca' Foscari University of Venice, Italy
KATHRIN GROSSE*, DIEE, University of Cagliari, Italy
AMBRA DEMONTIS†, DIEE, University of Cagliari, Italy
SEBASTIANO VASCON, DAIS, Ca' Foscari University of Venice, Italy
WERNER ZELLINGER, Software Competence Center Hagenberg GmbH (SCCH), Austria
BERNHARD A. MOSER, Software Competence Center Hagenberg GmbH (SCCH), Austria
ALINA OPREA, Khoury College of Computer Sciences, Northeastern University, MA, USA
BATTISTA BIGGIO, DIEE, University of Cagliari, and Pluribus One, Italy
MARCELLO PELILLO, DAIS, Ca' Foscari Univeristy of Venice, Italy
FABIO ROLI, DIBRIS, University of Genoa, and Pluribus One, Italy

Membership Inference Attacks
Privacy Attacks (Shokri et al., IEEE Symp. SP 2017)

- ***Goal:*** to identify whether an input sample is part of the training set used to learn a deep neural network based on the observed prediction scores for each class



AI/ML Protection against Evasion Attacks



What is the rule? The rule is protect yourself at all times
(from the movie “Million dollar baby”, 2004)

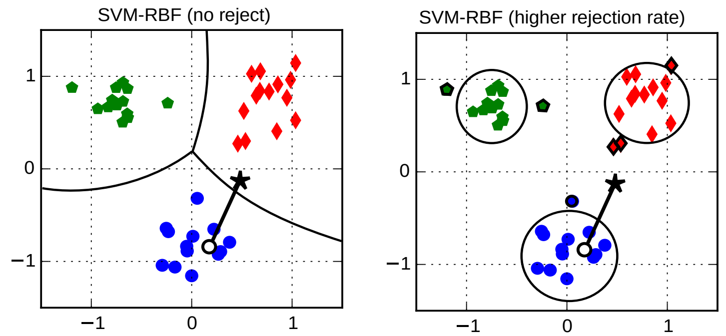
Security Measures against Evasion Attacks

1. **Robust optimization** to model attacks during learning
   – adversarial training / regularization

$$\min_{\boldsymbol{w}} \sum_i \max_{||\delta_i|| \le \epsilon} \ell(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i + \boldsymbol{\delta}_i))$$

bounded perturbation!

2. **Rejection / detection** of adversarial examples

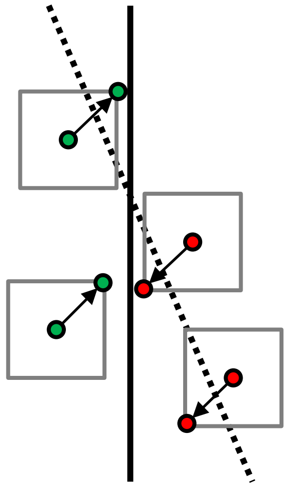


Increasing Input Margin via Robust Optimization

- Robust optimization (a.k.a. *adversarial training*)

$$\min_{\boldsymbol{w}} \max_{||\boldsymbol{\delta}_i||_\infty \le \epsilon} \sum_i \ell(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i + \boldsymbol{\delta}_i))$$

bounded perturbation!



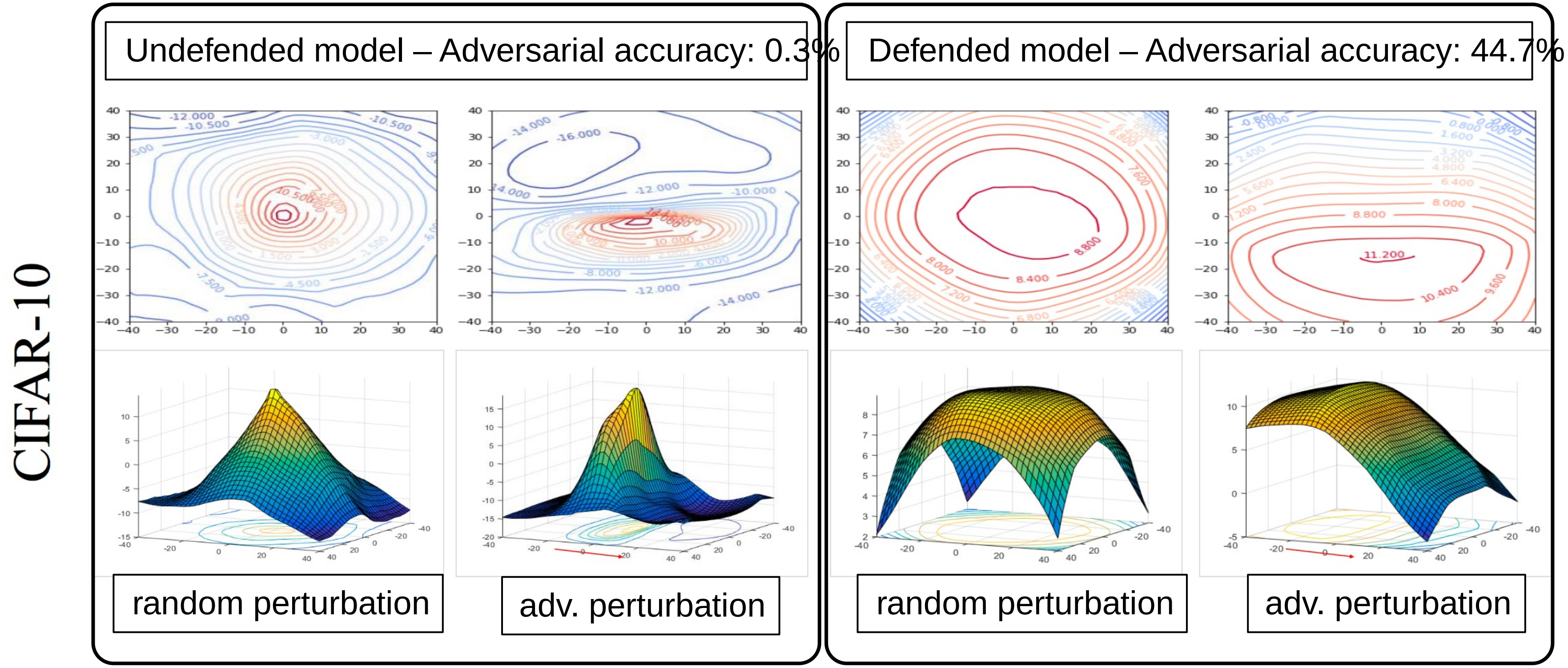
- Robustness and regularization (Xu et al., JMLR 2009)
  - under loss linearization, equivalent to loss regularization

$$\min_{\boldsymbol{w}} \sum_i \ell(y_i, f_{\boldsymbol{w}}(\boldsymbol{x}_i)) + \epsilon ||\nabla_{\boldsymbol{x}} \ell_i||_1$$

dual norm of the perturbation

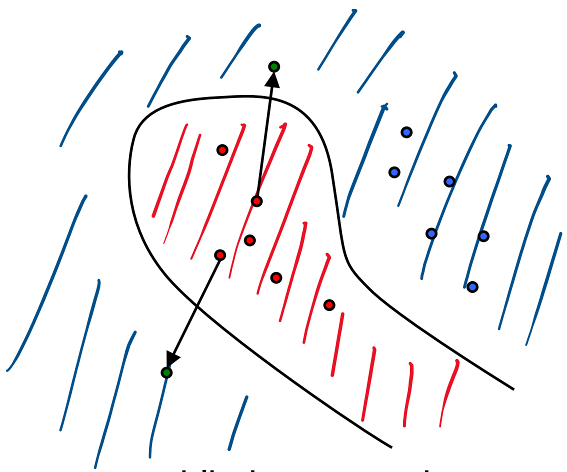
Demontis, Biggio et al., *Yes, Machine Learning Can Be More Secure! ...*, IEEE TDSC 2019

Why Does Robust Optimization Work?

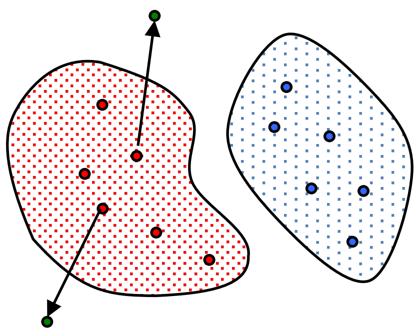


Detecting and Rejecting Adversarial Examples

- Adversarial examples tend to occur in *blind spots*
  - Regions far from training data that are anyway assigned to 'legitimate' classes

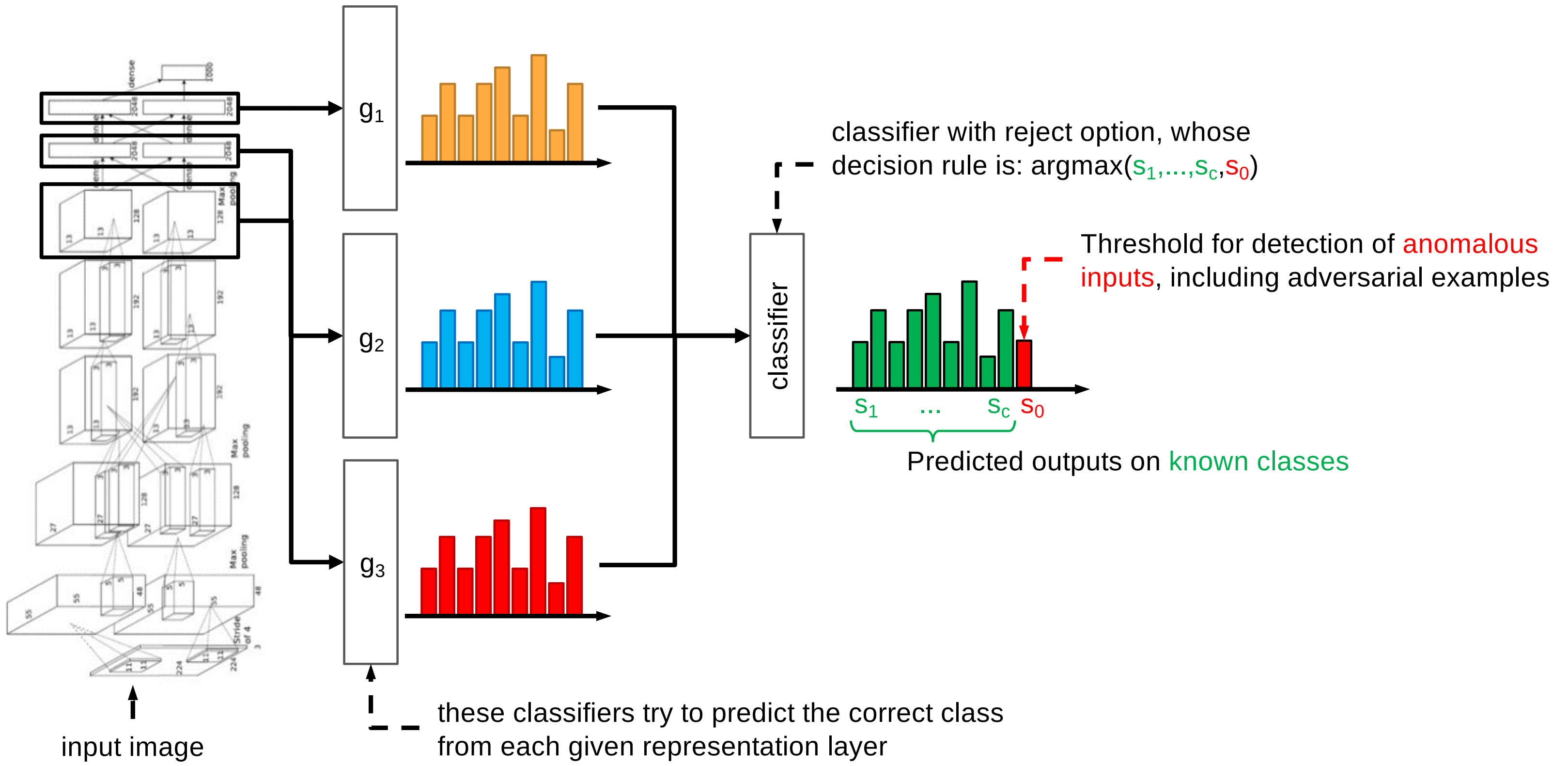


blind-spot evasion
(not even required to
mimic the target class)



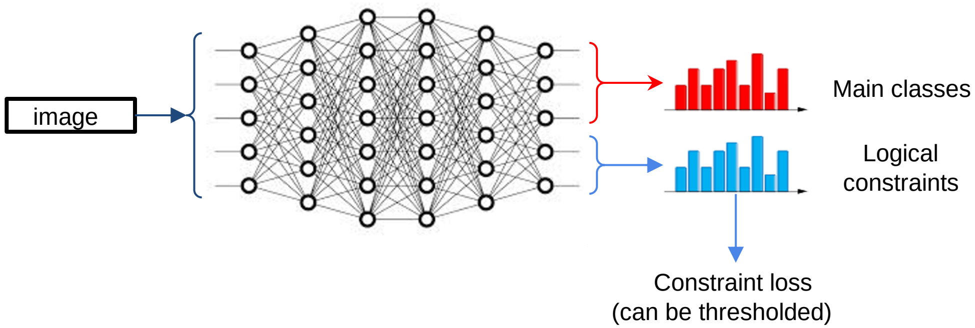
rejection of adversarial examples through
enclosing of legitimate classes

Deep Neural Rejection against Adversarial Examples



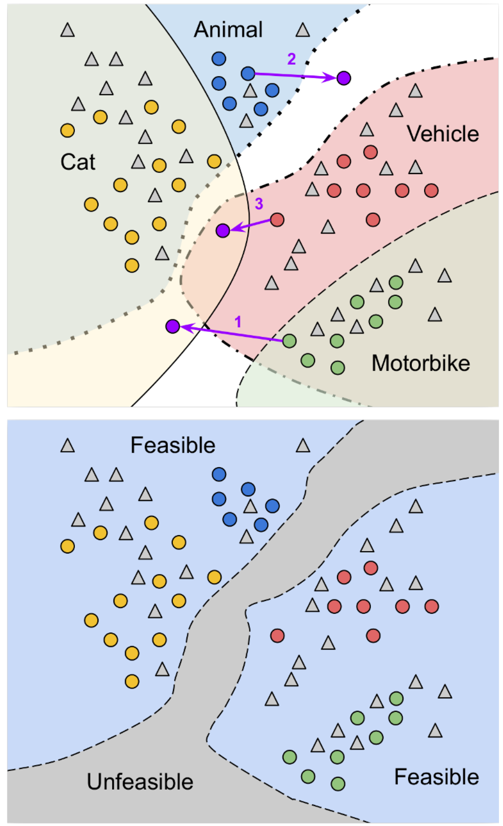
Sotgiu, Biggio et al., EURASIP JIS, 2020
Crecchi, Biggio et al., FADER: ..., Neurocomputing 2021

Domain Knowledge Alleviates Adversarial Examples



$$
\begin{aligned}
\forall x, \quad & \text{CAT}(x) \Rightarrow \text{ANIMAL}(x)\,, \\
\forall x, \quad & \text{MOTORBIKE}(x) \Rightarrow \text{VEHICLE}(x)\,, \\
\forall x, \quad & \text{VEHICLE}(x) \Rightarrow \neg\text{ANIMAL}(x)\,, \\
\forall x, \quad & \text{CAT}(x) \vee \text{ANIMAL}(x) \vee \text{MOTORBIKE}(x) \vee \text{VEHICLE}(x)
\end{aligned}
$$

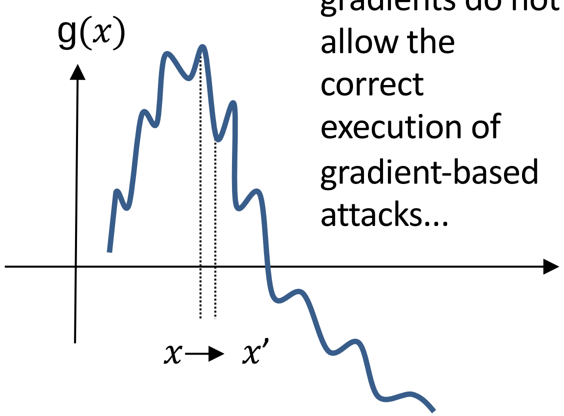
$$
\min_{\mathbf{f}} = \boxed{\frac{1}{n}\sum_{i=1}^{l} L_y(\mathbf{f}(\mathbf{x}_i), \mathbf{y}_i)} + \boxed{\sum_{j=1}^{l+u}\sum_{h=1}^{m} \lambda_m \cdot L_\phi(\phi_h(\mathbf{f}(\mathbf{x}_j)))} + \lambda\|\mathbf{f}\|
$$



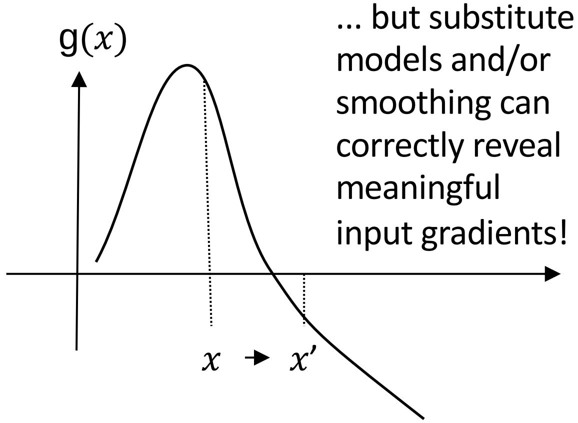
Melacci, Biggio et al., IEEE TPAMI 2021

Ineffective Defenses: Obfuscated Gradients

- Work by Carlini & Wagner (SP' 17) and Athalye et al. (ICML '18) has shown that
  - some recently-proposed defenses rely on obfuscated / masked gradients...
  - ... and they can be circumvented



Obfuscated gradients do not allow the correct execution of gradient-based attacks...

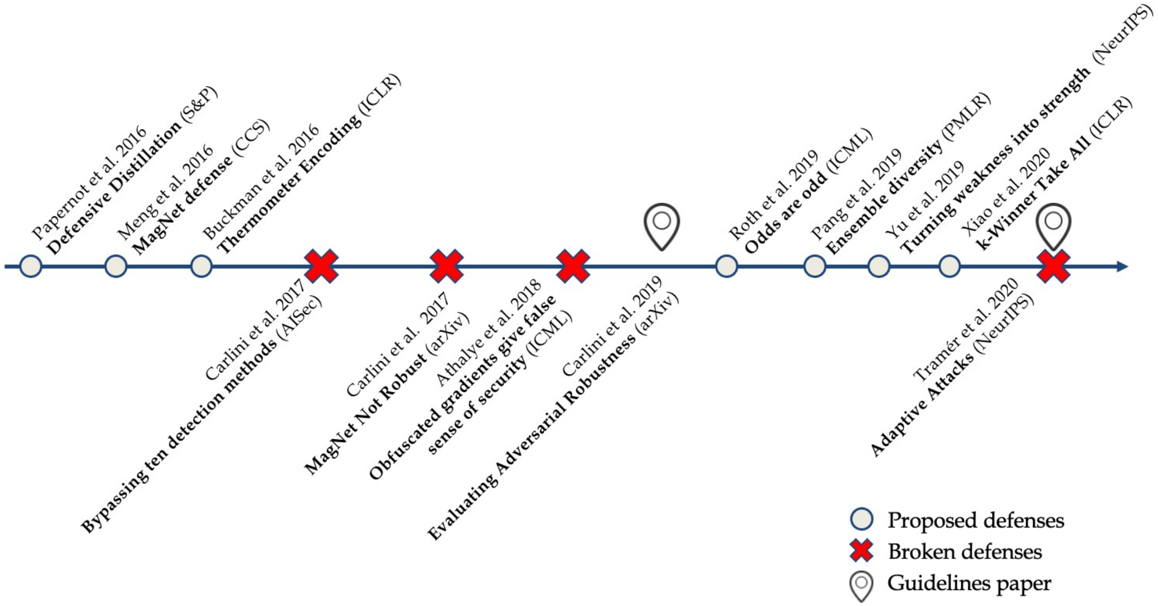


... but substitute models and/or smoothing can correctly reveal meaningful input gradients!

Detect and Avoid Flawed Evaluations

- **Problem**: formal evaluations do not scale, adversarial robustness evaluated mostly empirically, via gradient-based attacks

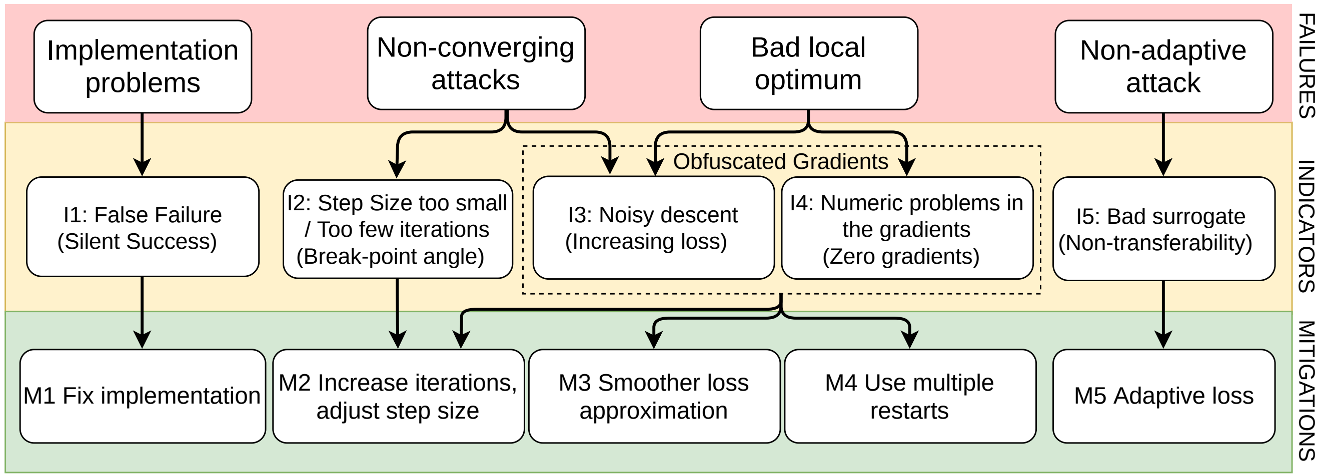
- **Gradient-based attacks can fail:** many flawed evaluations have been reported, with defenses easily broken by adjusting/fixing the attack algorithms

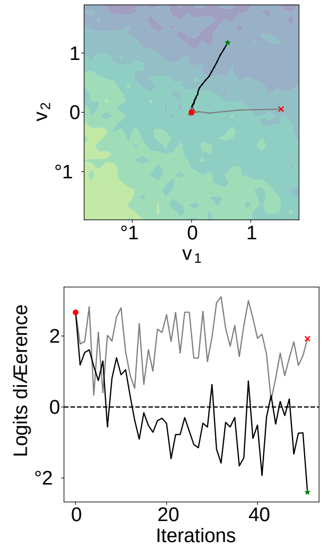


Pintor et al., Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples, arXiv 2021

Detect and Avoid Flawed Evaluations

- **Problem**: formal evaluations do not scale, adversarial robustness evaluated mostly empirically, via gradient-based attacks
- **Gradient-based attacks can fail:** many flawed evaluations have been reported, with defenses easily broken by adjusting/fixing the attack algorithms

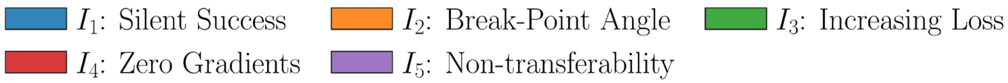


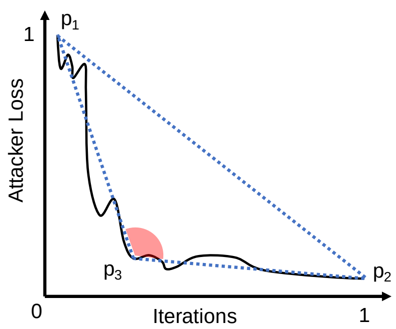


Pintor et al., Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples, arXiv 2021

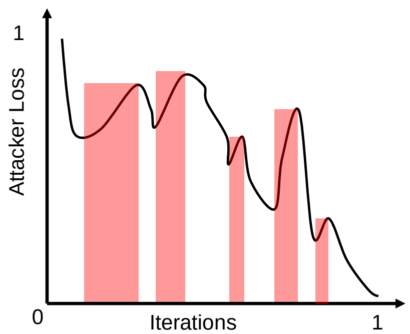
Indicators of Attack Failure

- Indicators of Failure (IoF) with corresponding mitigation strategies/protocol





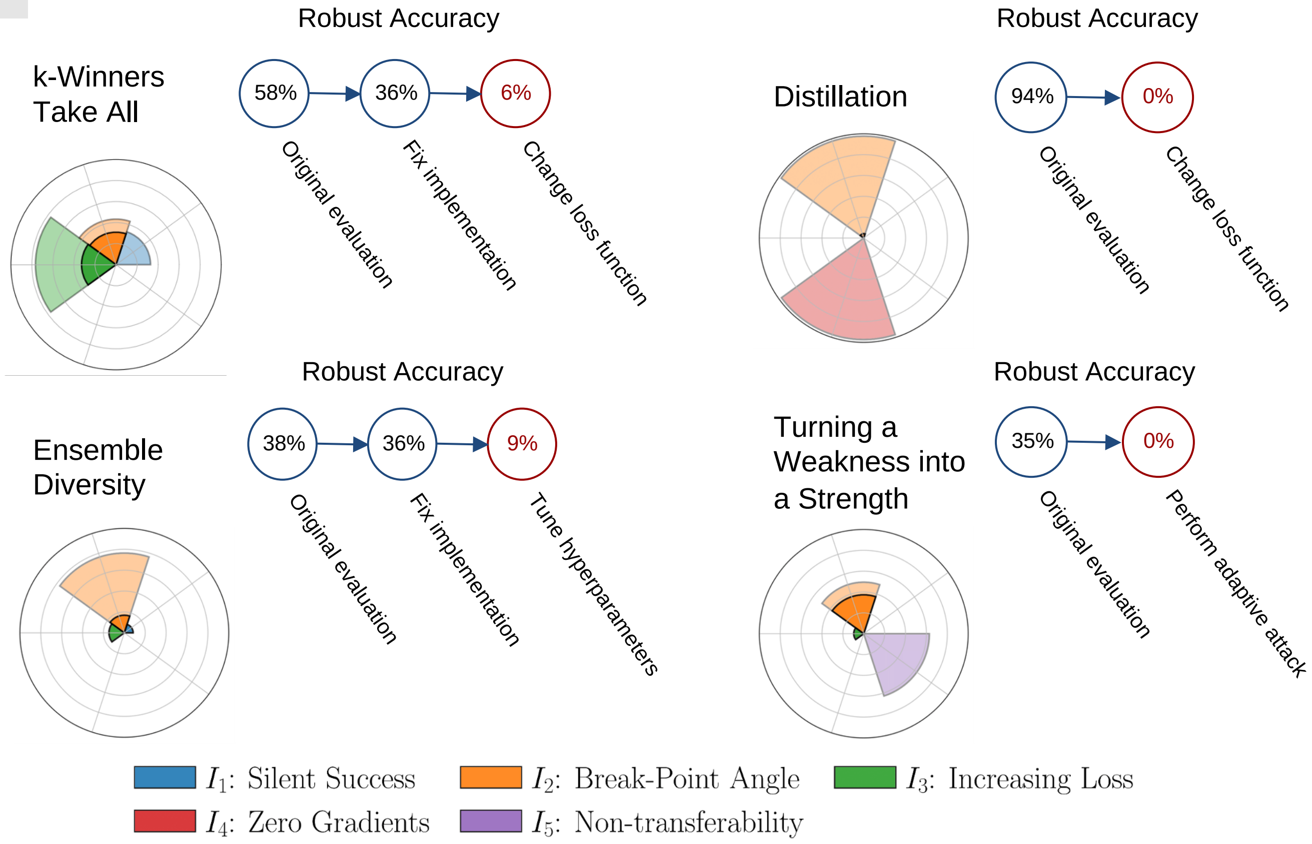
Break-point angle



Increasing loss

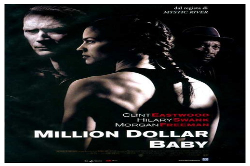
Pintor et al., Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples, arXiv 2021

Experiments



Pintor et al., Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples, arXiv 2021

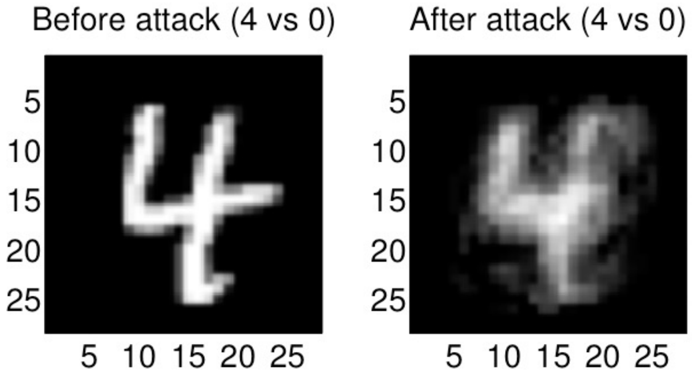
AI/ML Protection against Poisoning Attacks



What is the rule? The rule is protect yourself at all times
(from the movie “Million dollar baby”, 2004)

Security Measures against DoS Poisoning

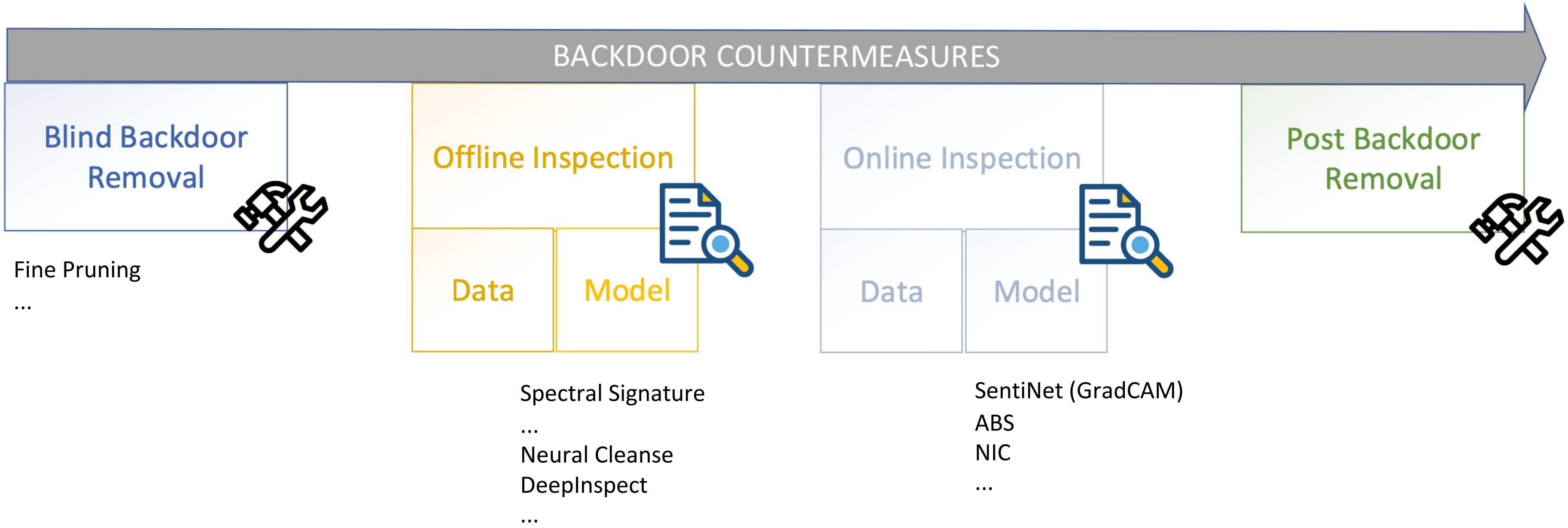
- **Rationale:** poisoning injects outlying training samples



- Two main strategies for countering this threat
  1. **Data sanitization:** *remove* poisoning samples from training data
     - Bagging for fighting poisoning attacks
     - Reject-On-Negative-Impact (RONI) defense
  2. **Robust Learning:** learning algorithms that are robust in the presence of poisoning samples

Jagielski, Biggio et al., IEEE Symp. Security and Privacy, 2018

Defending against Backdoor Poisoning Attacks

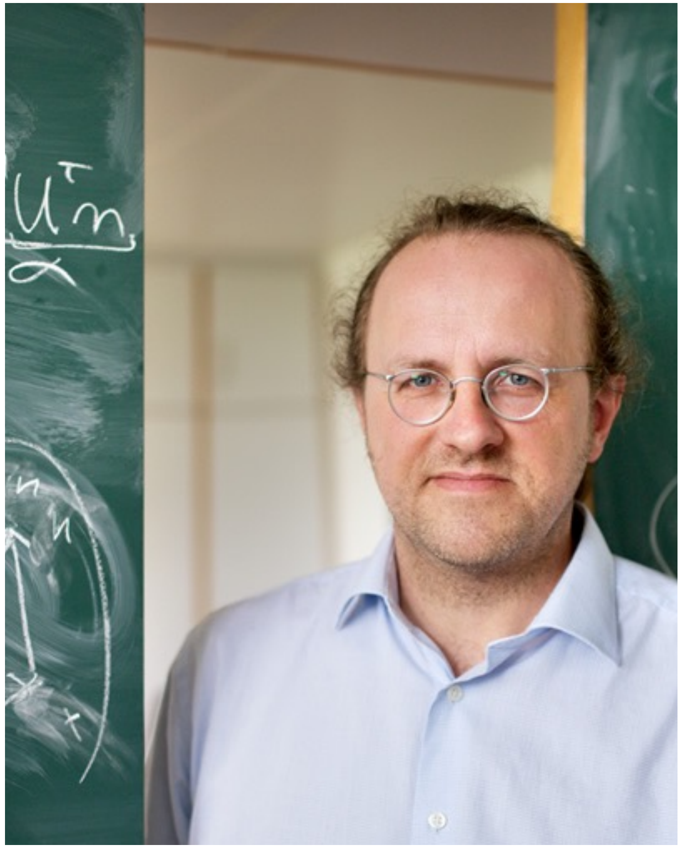


Gao et al., *Backdoor Attacks and Countermeasures...*, arXiv 2017
Cinà, Grosse, Biggio et al., *Wild Patterns Reloaded:...*, arXiv 2022

Why Is AI Vulnerable?

Why Is AI Vulnerable?

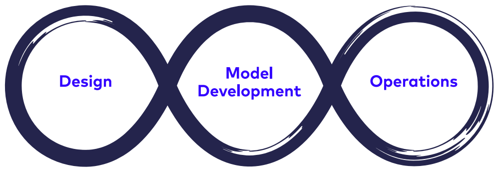
- **Underlying assumption:** past data is *representative* of future data (IID data)
- The success of modern AI is on tasks for which we collected enough representative training data
- **We cannot build AI models for each task an agent is ever going to encounter,** but there is a whole world out there where the IID assumption is violated
- **Adversarial attacks** point exactly at this lack of robustness which comes from IID specialization

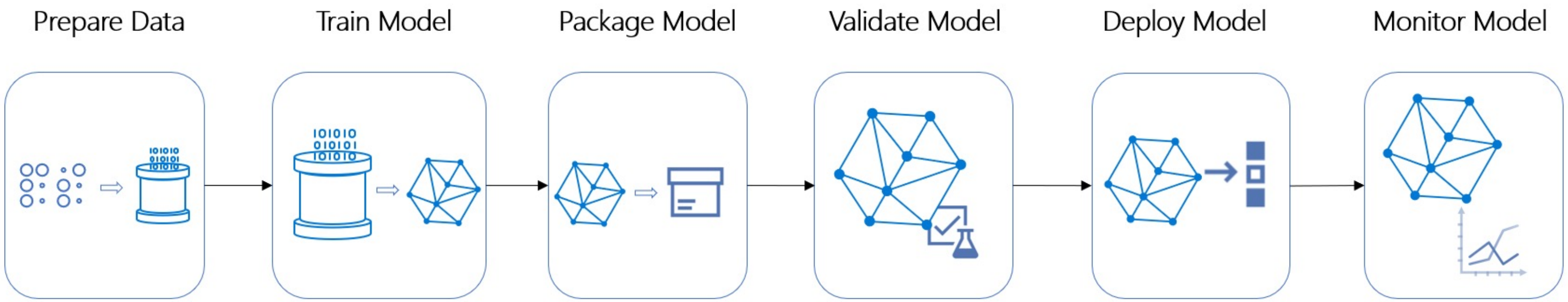


Bernhard Schölkopf

Director, Max Planck Institute, Tuebingen, Germany

What's Next? MLOps: ML+Dev+Ops



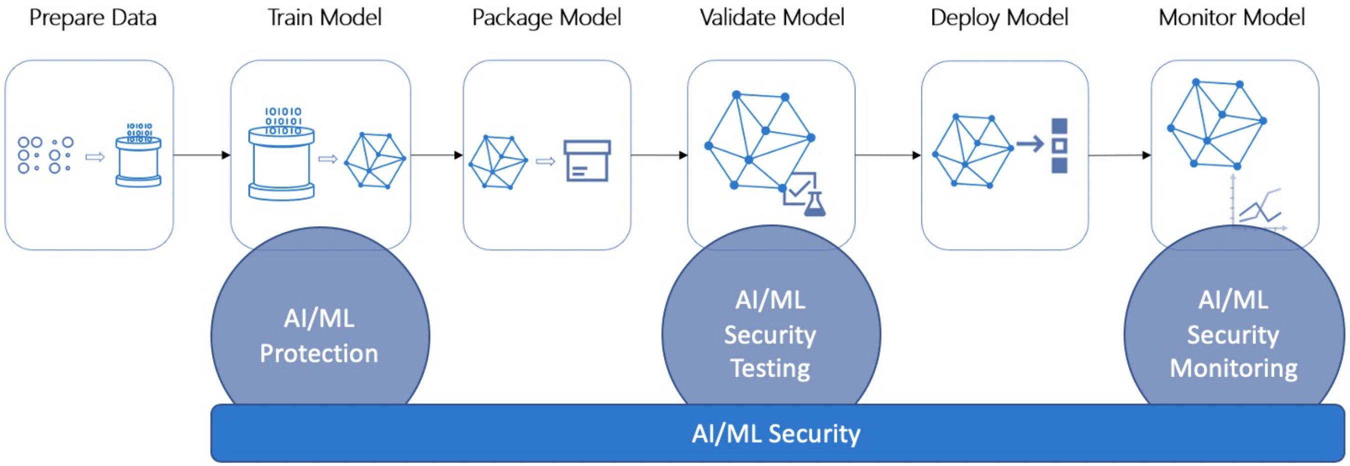


- MLOps poses many **industrial** and **research** challenges
  - Continuous data ingestion and labeling, model retraining/continuous updating, testing/validation, ...

- ... but also **lack of debugging tools** and **systematic security testing** to prevent attacks and/or improve robustness under adversarial/temporal drift!

Our Vision: From MLOps to MLSecOps

- **Goal**: to empower MLOps with AI/ML Security, developing three main pillars
  - **AI/ML Protection**: to build robust AI/ML and data sanitization procedures
  - **AI/ML Security Testing**: to ensure proper testing and debugging of AI/ML models
  - **AI/ML Security Monitoring**: to monitor AI/ML models in production (e.g., when deploying MLaaS) to timely detect ongoing attacks and block them



Open Course on MLSec

https://github.com/unica-mlsec/mlsec

Software Tools

https://github.com/pralab

Thanks!

Battista Biggio

battista.biggio@unica.it

@biggiobattista

Ambra Demontis

Maura Pintor

Kathrin Grosse

Angelo Sotgiu

Luca Demetrio

Antonio Cinà

Fabio Roli

If you know the enemy and know yourself, you need not fear the result of a hundred battles
Sun Tzu, The art of war, 500 BC