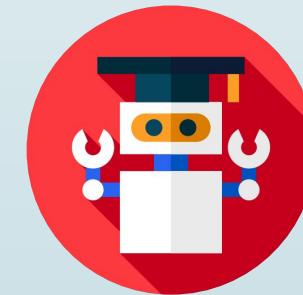




# Como enganar modelos de aprendizado de máquina?

Erikson Júlio de Aguiar  
[erjulioaguiar@usp.br](mailto:erjulioaguiar@usp.br)



# Quem sou eu?



- Doutorando em Ciência de Computação **ICMC - USP** (GBDI)
- Mestre em Ciência de Computação **ICMC - USP**
- Bacharel em Ciência da Computação - **UENP**
- **Áreas de interesse:**
  - Segurança & Privacidade
  - Blockchain
  - Aprendizado de máquina
  - Processamento de imagens

[@erjulioaguiar](https://twitter.com/erjulioaguiar)



[eriksonJAguiar](https://github.com/eriksonJAguiar)



<https://www.linkedin.com/in/erjulioaguiar/>

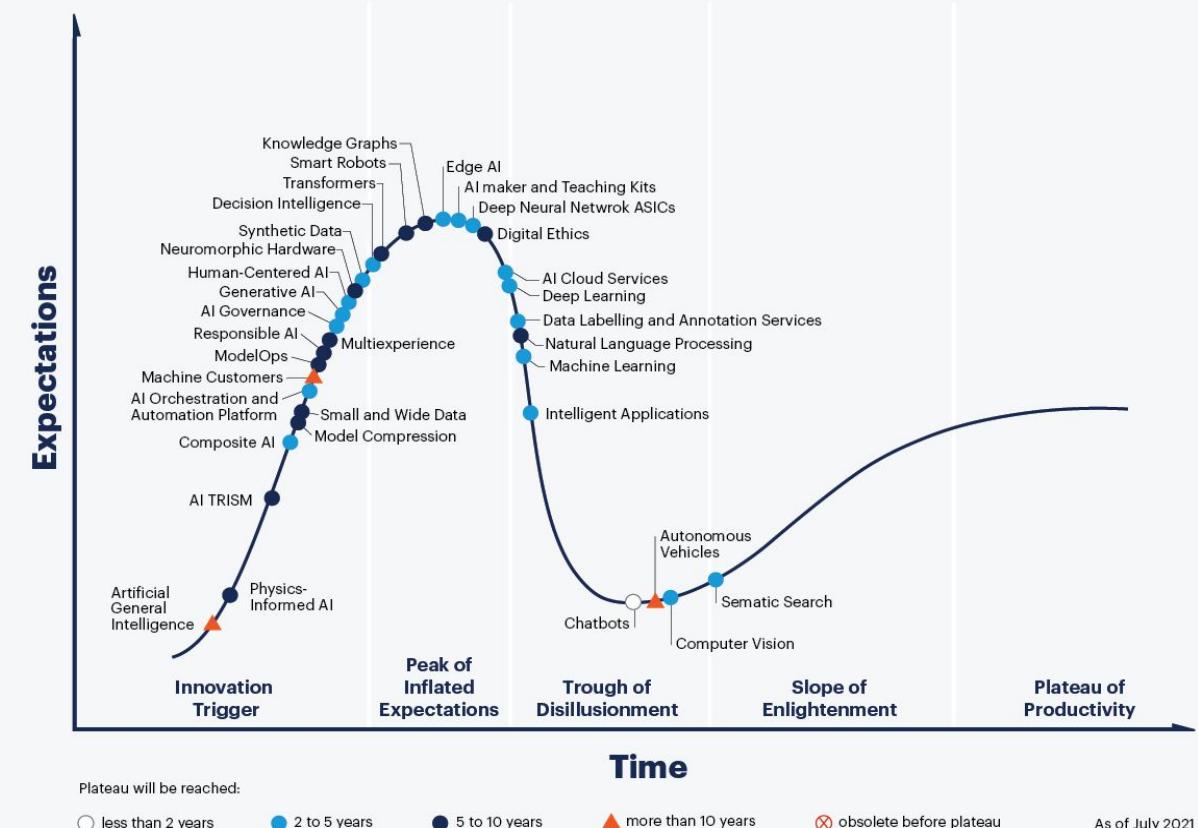


# Sumário

1. Motivação
2. Definições
3. Ataques contra modelos de aprendizado de máquina
4. Técnicas de defesa em aprendizado de máquina
5. Privacidade em aprendizado de máquina
6. Projetos
7. Dicas

# O potencial do aprendizado de máquina no cotidiano

## Hype Cycle for Artificial Intelligence, 2021



gartner.com

Source: Gartner  
© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1482644

Gartner

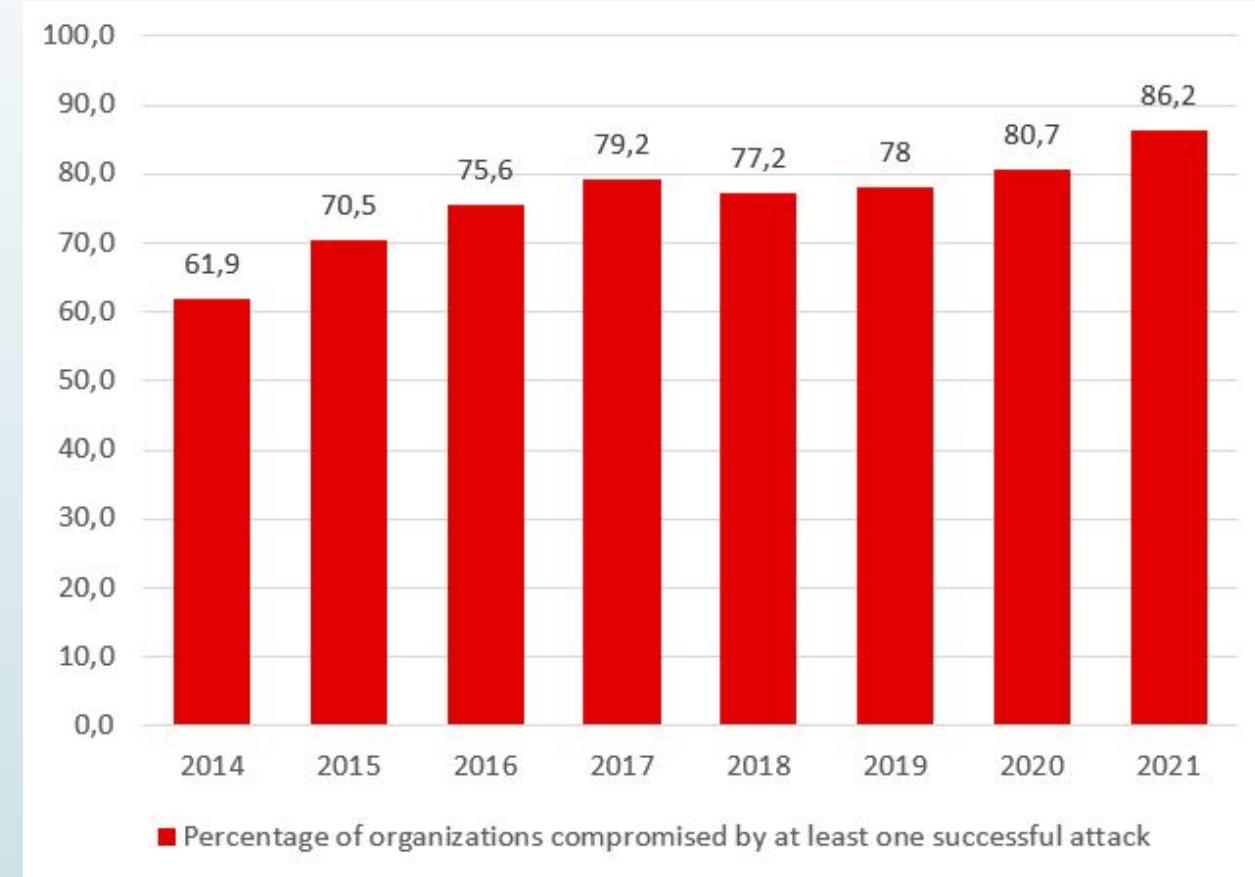
Fonte: [\(Gartner, 2021\)](#)

5

# Áreas de aplicação do aprendizado de máquina

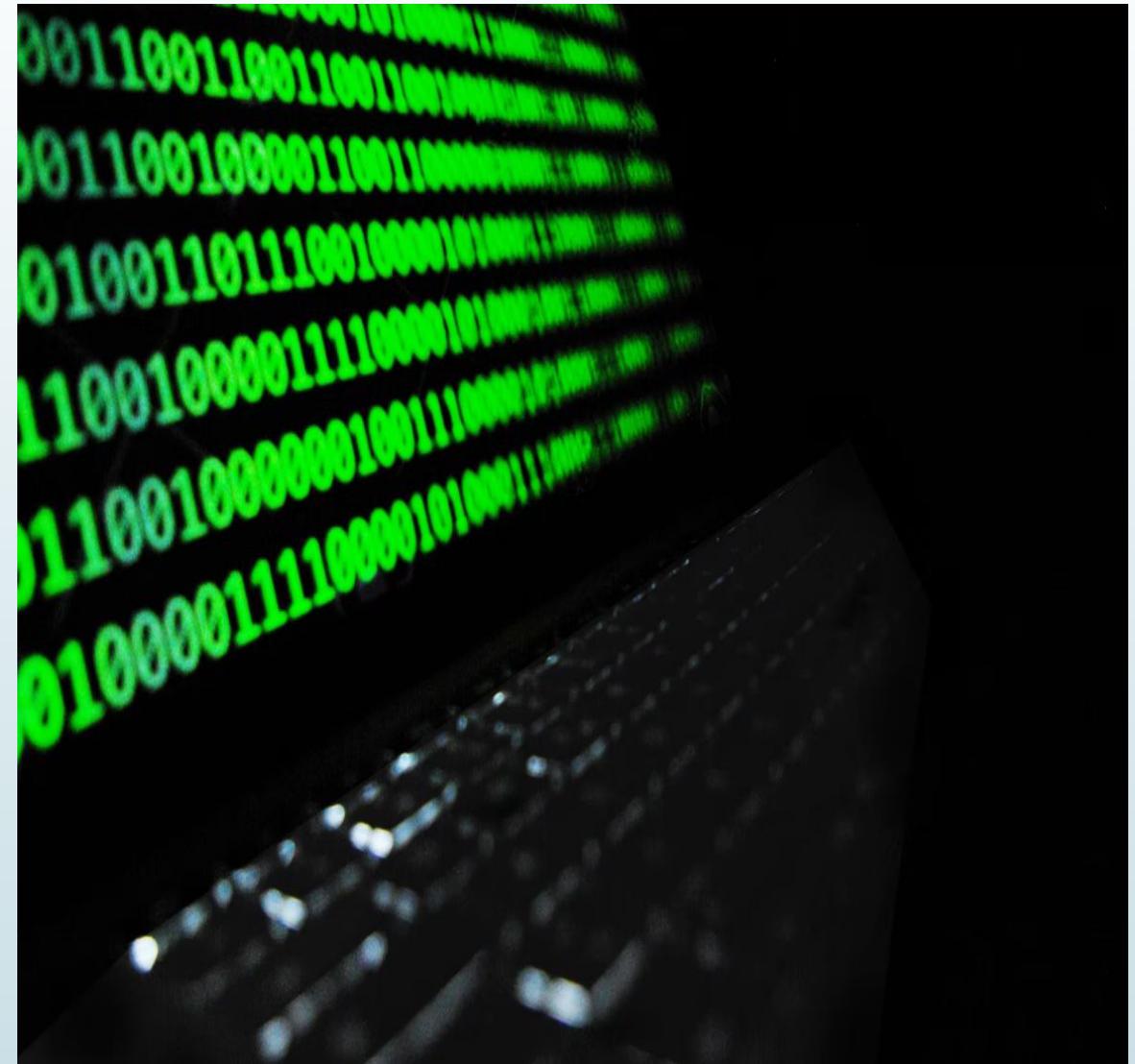


# O crescimento das ameaças na internet



Fonte: [\(Stainer, 2021\)](#)

# Modelos de Aprendizado de máquina são vulneráveis?



**Quais seriam os impactos das vulnerabilidades?**



[#CVE-2019-20634](#)

[#CVE-2021-37678](#)

Machine learning classifiers trained via gradient descent are vulnerable to arbitrary misclassification attack

Vulnerability Note VU#425163

Original Release Date: 2020-03-19 | Last Revised: 2020-06-04



### CVE-2019-20634 Detail

#### MODIFIED

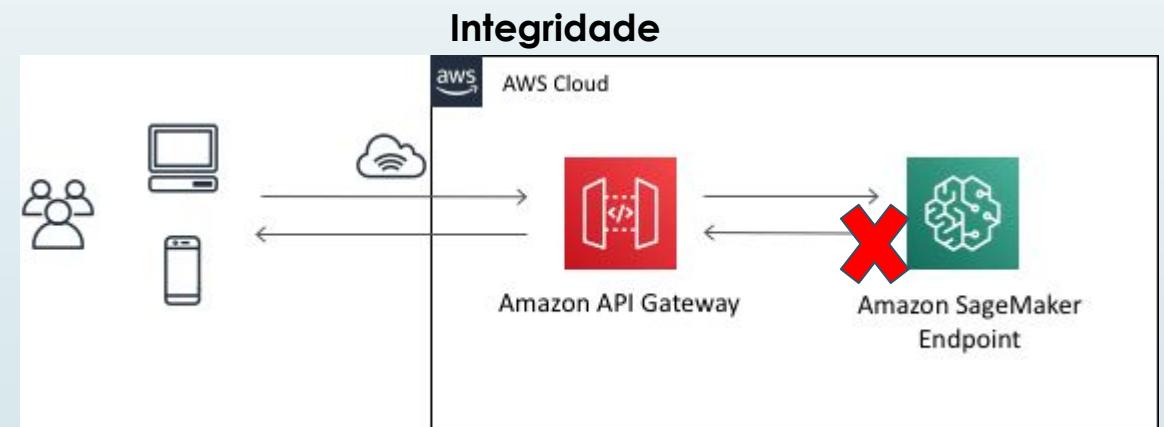
This vulnerability has been modified since it was last analyzed by the NVD. It is awaiting reanalysis which may result in further changes to the information provided.

#### Current Description

An issue was discovered in Proofpoint Email Protection through 2019-09-08. By collecting scores from Proofpoint email headers, it is possible to build a copy-cat Machine Learning Classification model and extract insights from this model. The insights gathered allow an attacker to craft emails that receive preferable scores, with a goal of delivering malicious emails.

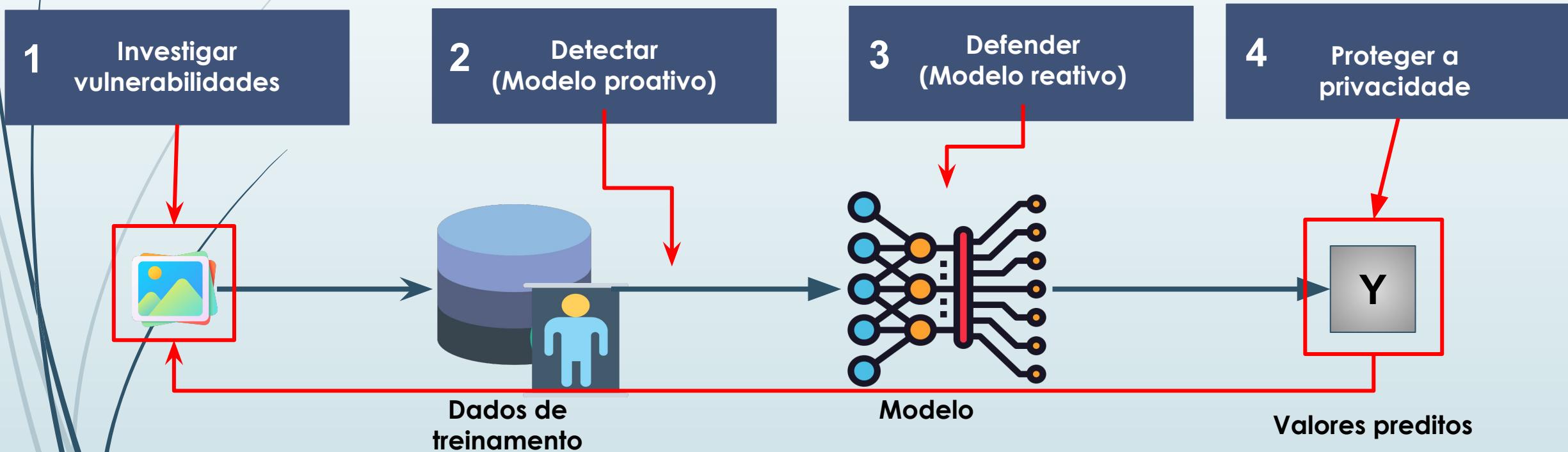
# Machine Learning Security (MLSec)

## Definições

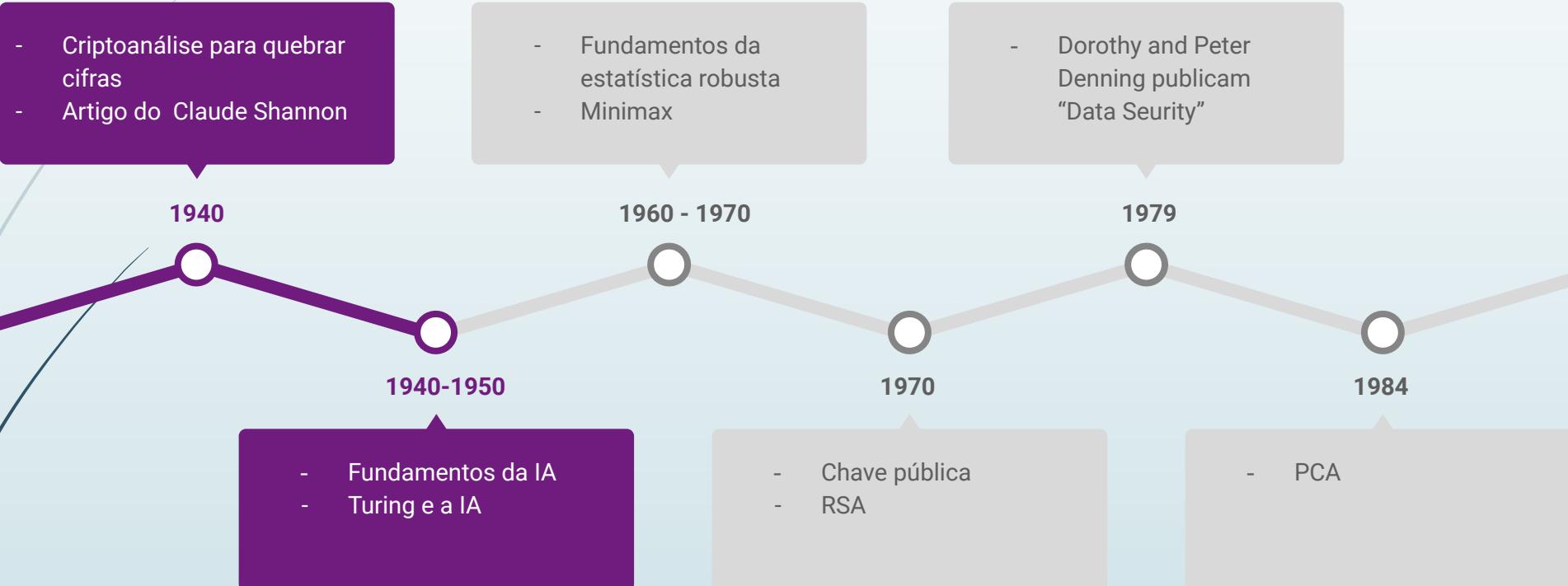


11

# Níveis das vulnerabilidades - MLSec



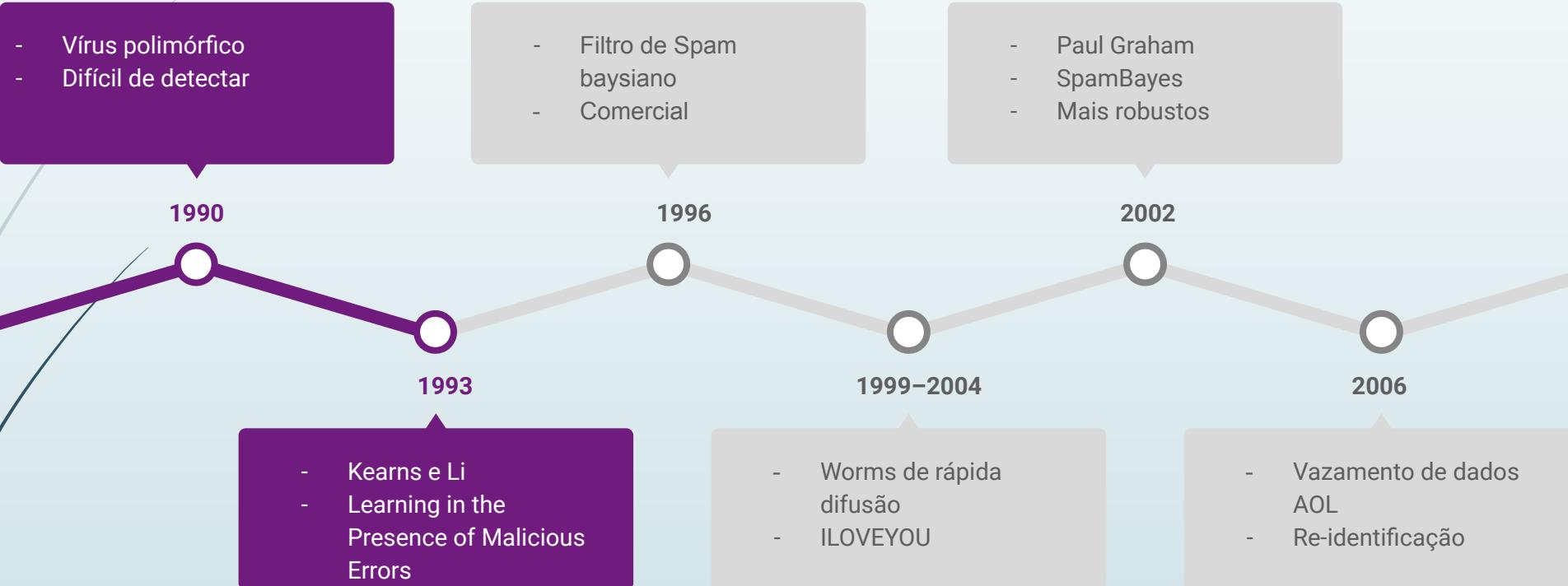
# Histórico - MLSec



**Adversarial Machine Learning (2019)**

Joseph, Anthony D.; Nelson, Blaine; Rubinstein, Benjamin I. P.; Tygar, J. D.

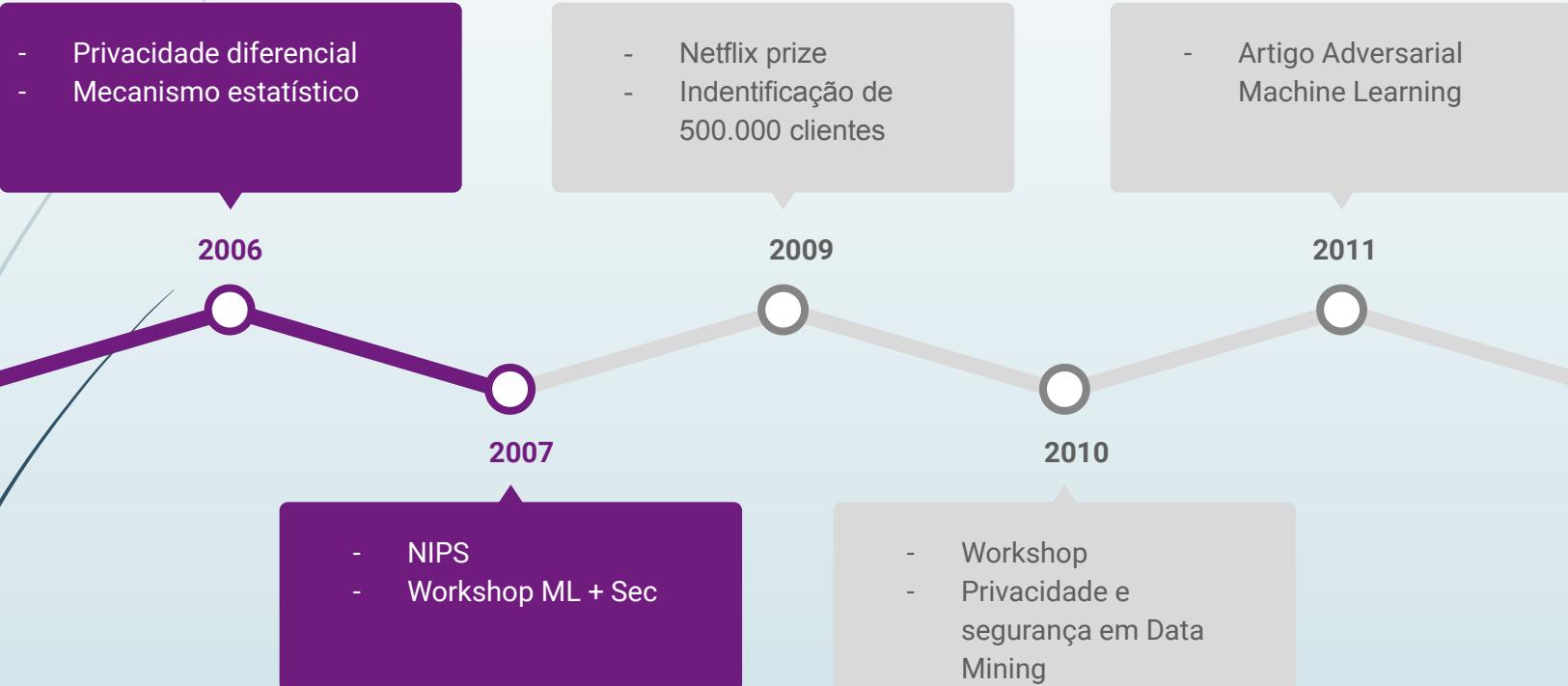
# Histórico - MLSec



Adversarial Machine Learning (2019)

Joseph, Anthony D.; Nelson, Blaine; Rubinstein, Benjamin I. P.; Tygar, J. D.

# Histórico - MLSec



**Adversarial Machine Learning (2019)**

Joseph, Anthony D.; Nelson, Blaine; Rubinstein, Benjamin I. P.; Tygar, J. D.

# Adversarial examples

## Adversarial Examples Are Not Bugs, They Are Features

Andrew Ilyas\*

MIT

ailyas@mit.edu

Shibani Santurkar\*

MIT

shibani@mit.edu

Dimitris Tsipras\*

MIT

tsipras@mit.edu

Logan Engstrom\*

MIT

engstrom@mit.edu

Brandon Tran

MIT

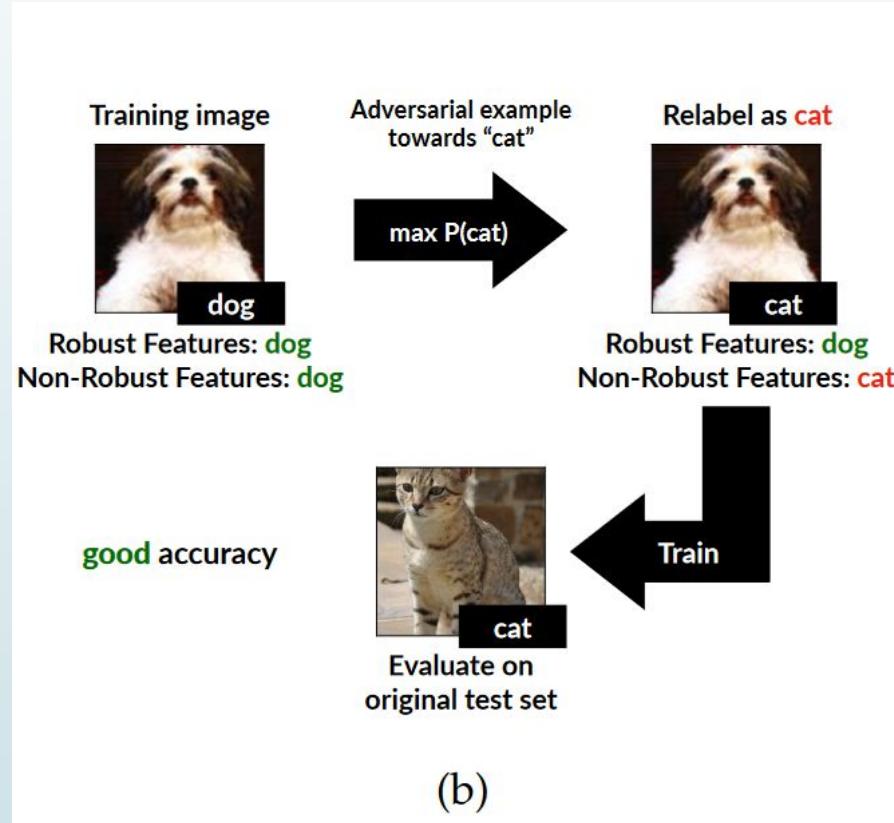
btran115@mit.edu

Aleksander Mądry

MIT

madry@mit.edu

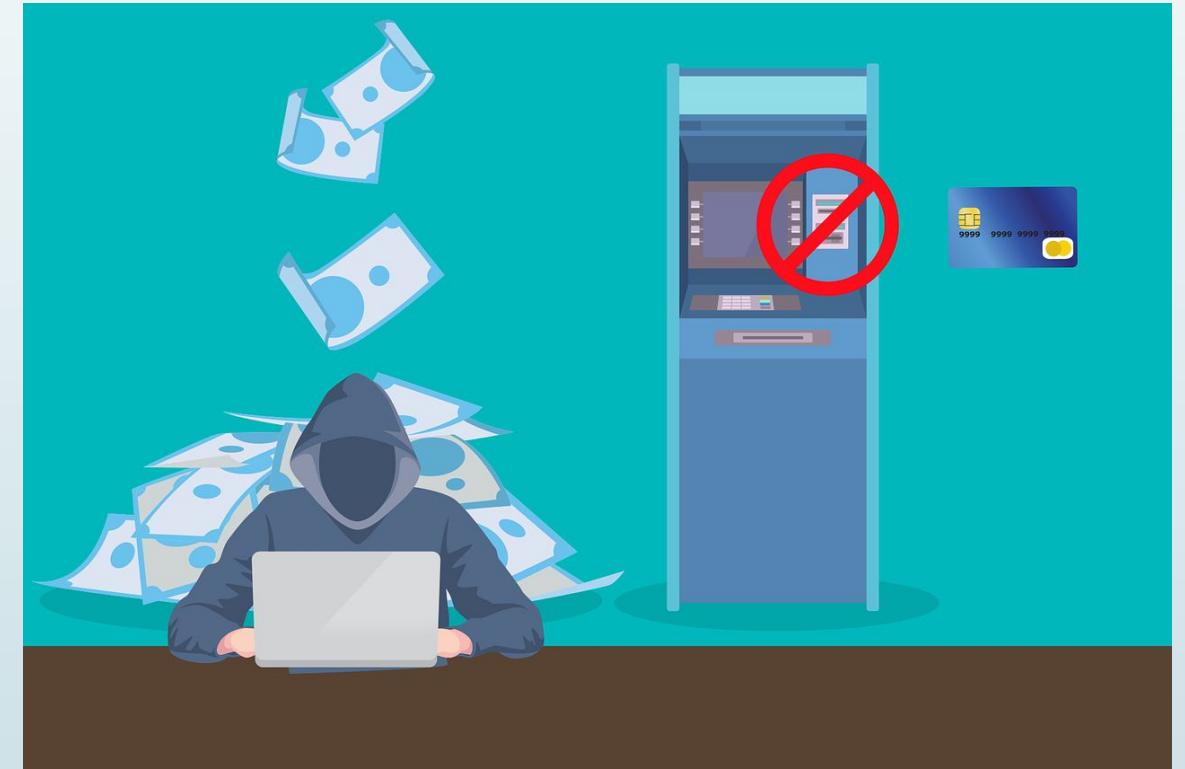
# Adversarial examples



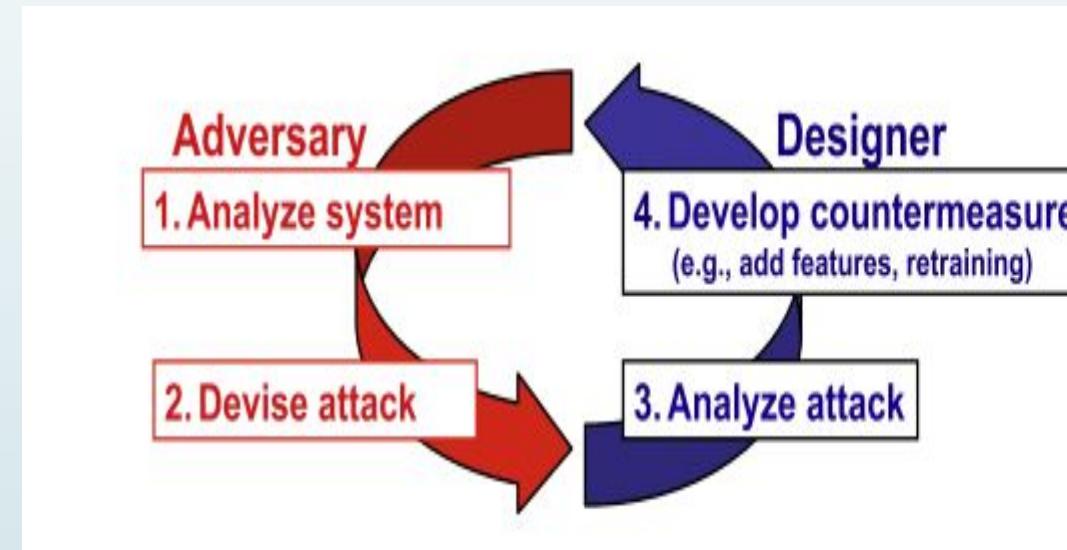
**Adversarial Examples are not Bugs, they are Features (2019)**

Andrew Ilyas; Shibani Santurkar; Dimitris Tsipras; Logan Engstrom; Brandon Tran; Aleksander Madry

# Ataques — MLSec (Adversarial Attacks)



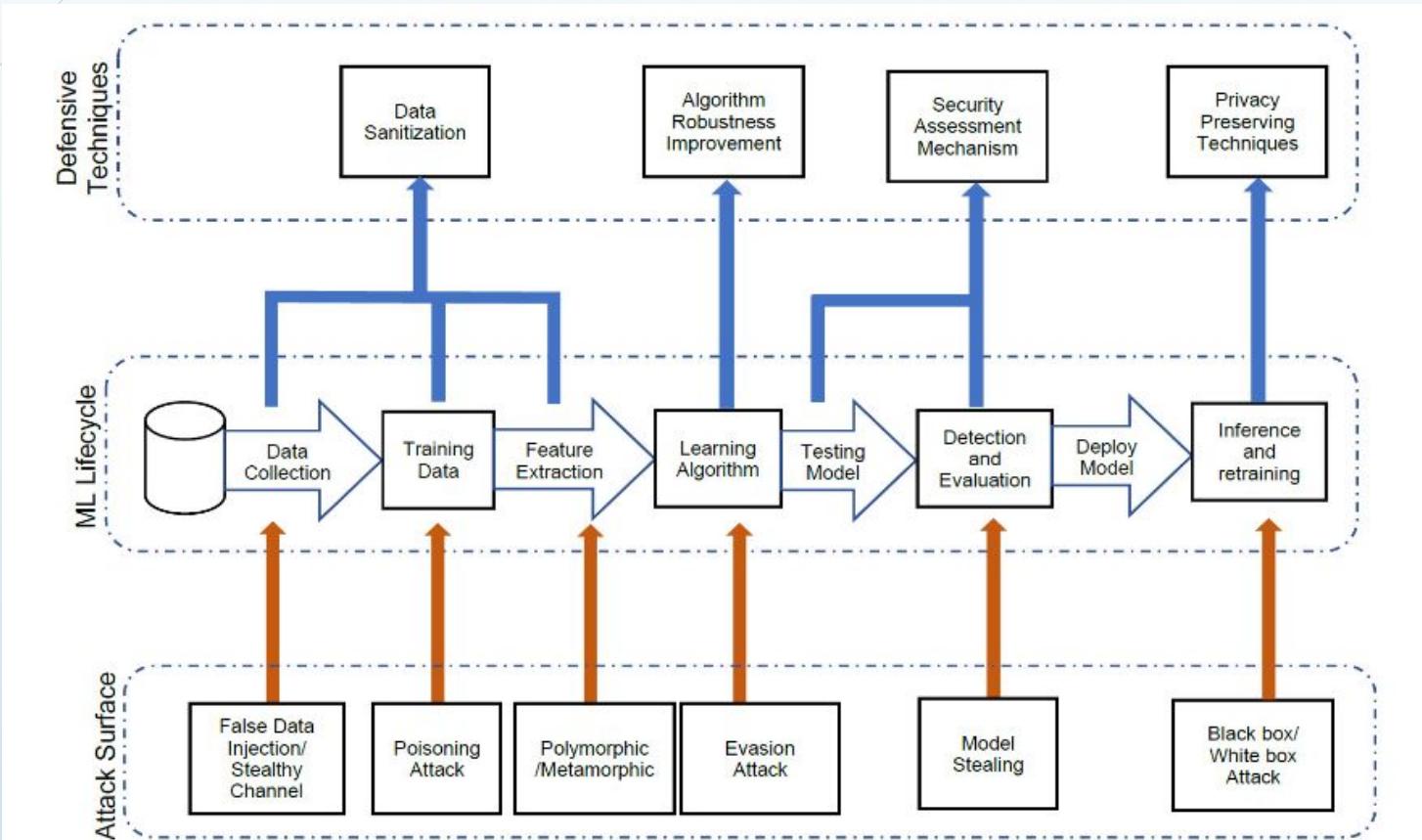
# Como um ataque inicia?



**Wild patterns: Ten years after the rise of adversarial machine learning (2018)**

Biggio, Battista; Roli, Fabio

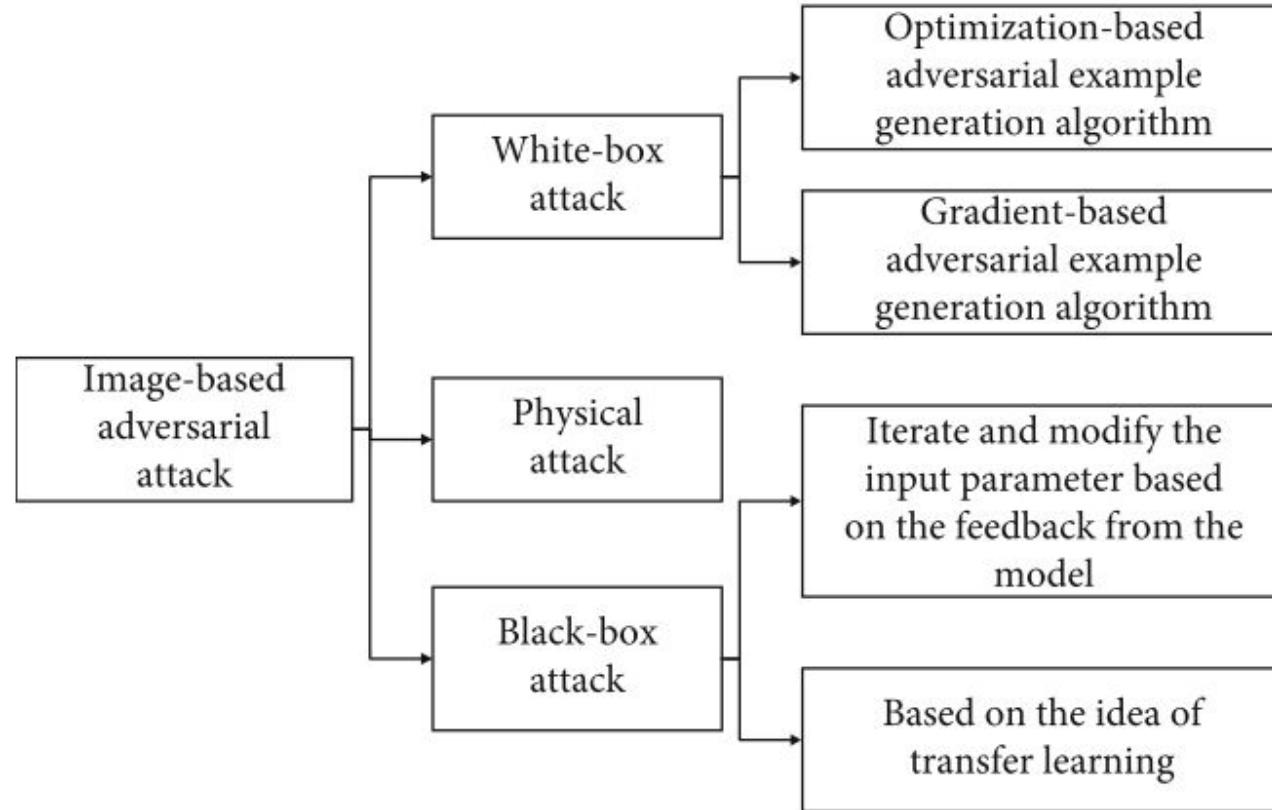
# Como um ataque inicia?



**Applications in Security and Evasions in Machine Learning: A Survey (2020)**

Sagar, Ramani; Jhaveri, Rutvij; Borrego, Carlos

# Conhecimento do alvo



A Survey on Adversarial Attack in the Age of Artificial Intelligence (2021)

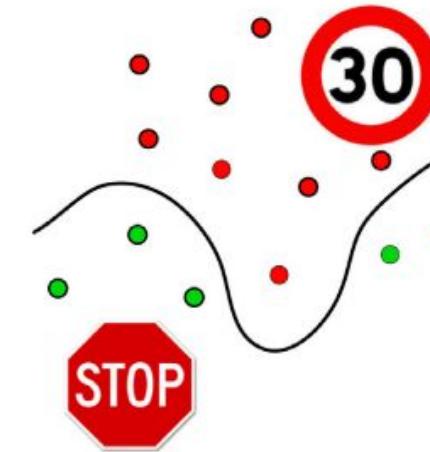
Kong, Zixiao; Xue, Jingfeng; Wang, Yong; Huang, Lu; Niu, Zequn; Li, Feng

# Poisoning attacks

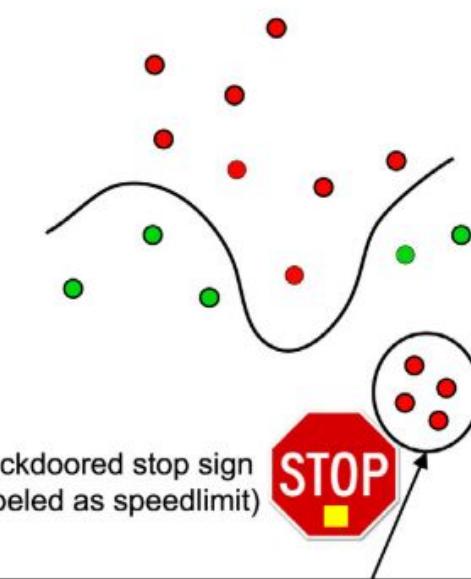
Wild patterns: Ten years after the rise of adversarial machine learning (2018)

Biggio, Battista; Roli, Fabio

Training data (no poisoning)



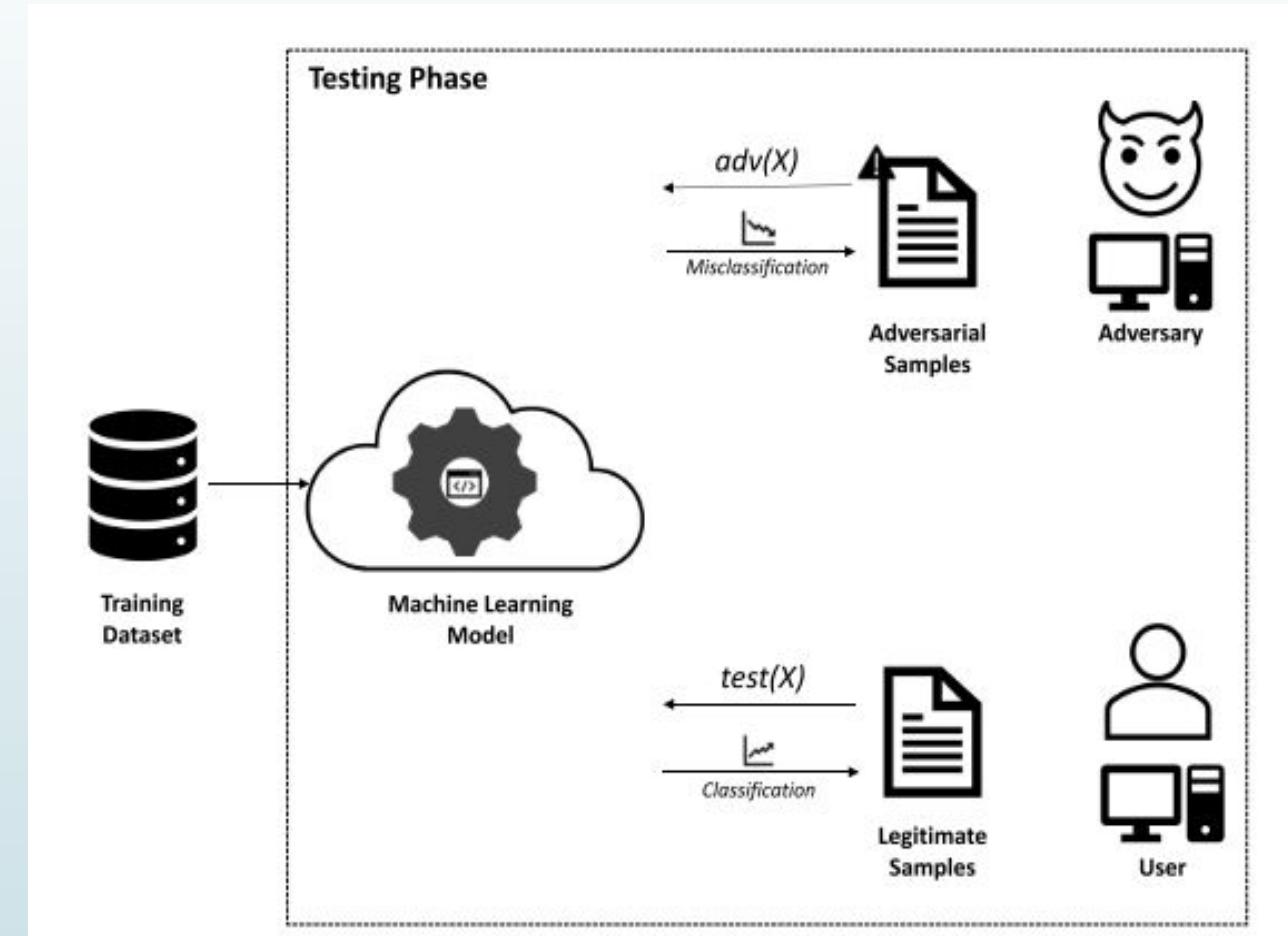
Training data (poisoned)



Backdoor / poisoning integrity 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



# Evasion attacks



**Model Evasion Attack on Intrusion Detection Systems using Adversarial Machine Learning (2020)**

Ayub, Md Ahsan; Johnson, William A.; Talbert, Douglas A.; Siraj, Ambareen

## Fast gradient Sign Method (FGSM)



$x$   
“panda”  
57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$   
“nematode”  
8.2% confidence

=



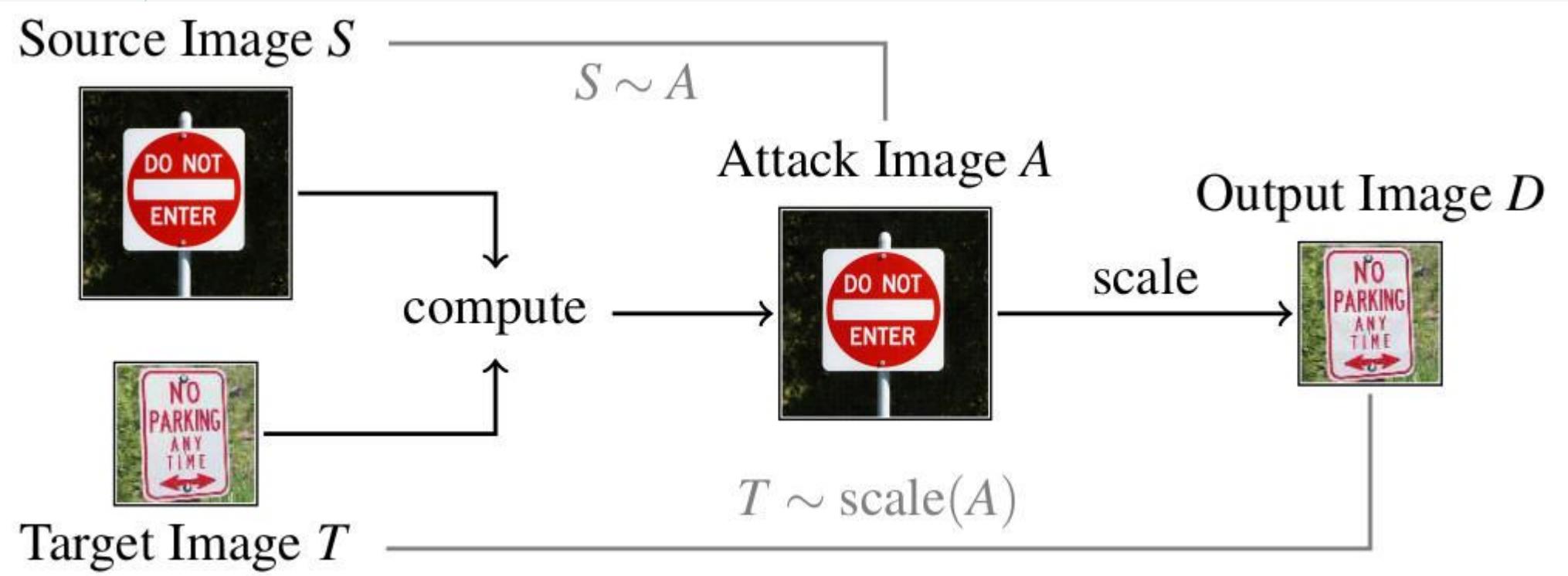
$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$   
“gibbon”  
99.3 % confidence

$$\tilde{x} = x + \epsilon \text{sgn}(\nabla_x J(w, x, y)).$$

Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review (2021)

Biggio, Battista; Roli, Fabio

# Ataques no pré-processamento



**Adversarial Preprocessing: Understanding and Preventing Image-Scaling Attacks in Machine Learning (2020)**  
Quiring, Erwin; Klein, David; Arp, Daniel; Johns, Martin; Rieck, Konrad

# Projected Gradient Descent

Test image

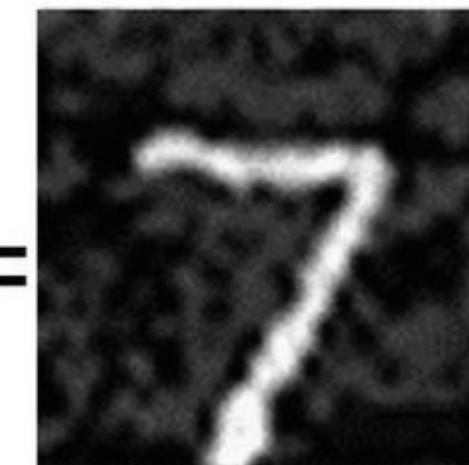


PGD attack

Perturbation  
(Attack)



Adversarial  
example

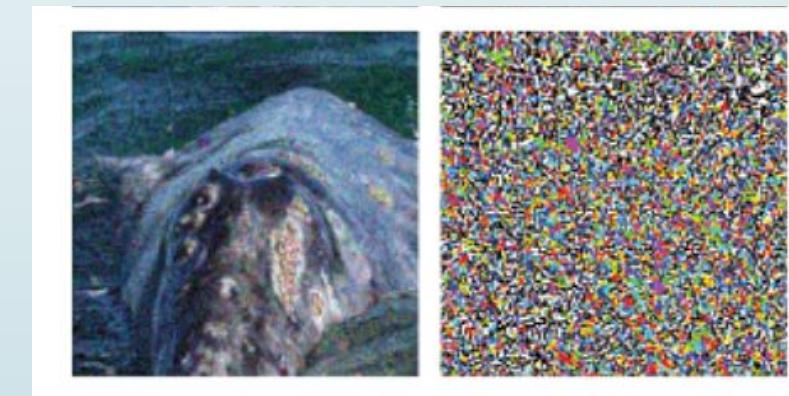
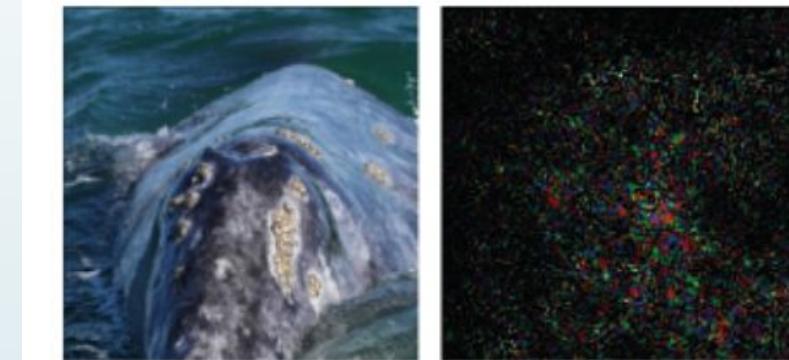


$$x^{t+1} = \Pi_{x+s}(x^t + \alpha \operatorname{sgn}(\nabla_x J(w, x, y))).$$

Towards Deep Learning Models Resistant to Adversarial Attacks (2018)

Madry, Aleksander; Makelov, Aleksandar; Schmidt, Ludwig; Tsipras, Dimitris; Vladu, Adrian

# DeepFool



**DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks (2016)**

Moosavi-Dezfooli; Mohsen, Seyed; Fawzi, Alhussein; Frossard, Pascal

# One Pixel Attack



Cup(16.48%)  
Soup Bowl(16.74%)



Bassinet(16.59%)  
Paper Towel(16.21%)



Teapot(24.99%)  
Joystick(37.39%)



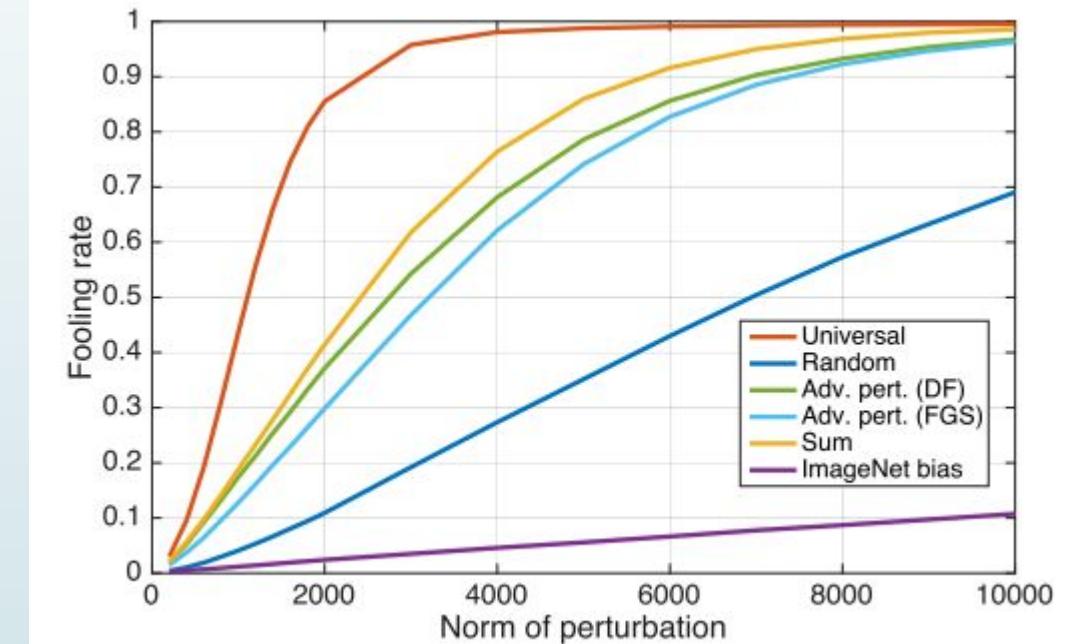
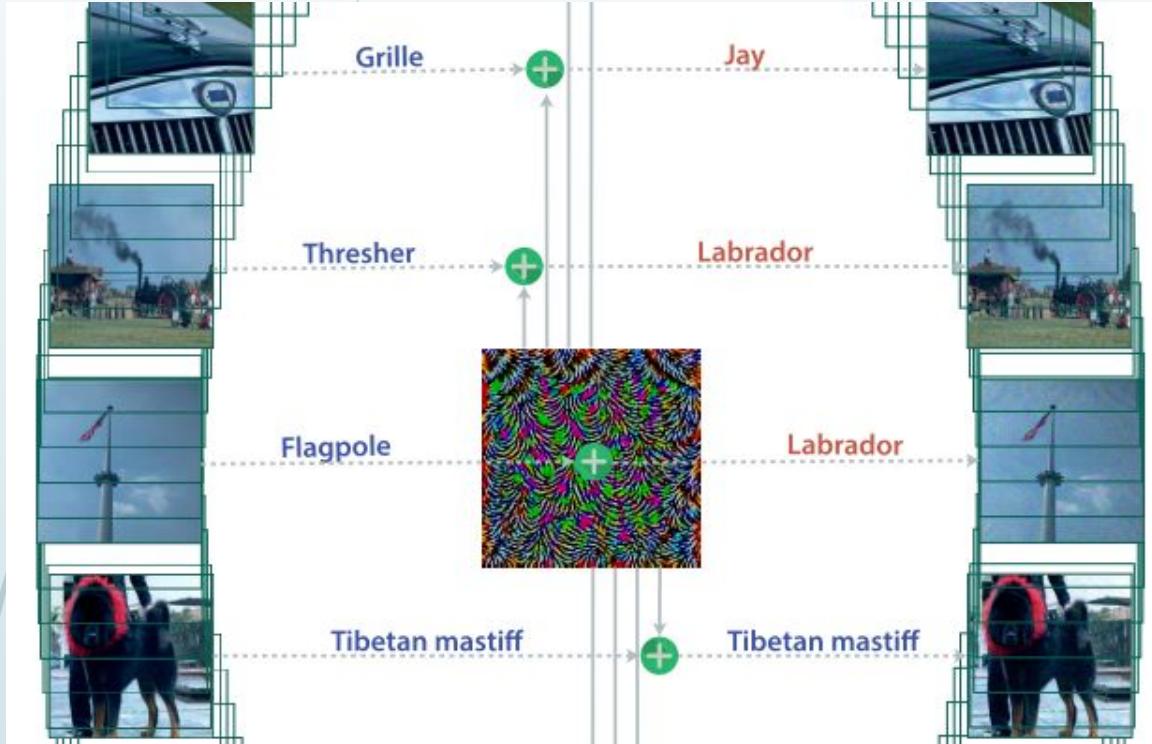
Hamster(35.79%)  
Nipple(42.36%)

One Pixel Attack for Fooling Deep Neural Networks (2019)

Su, Jiawei; Vargas, Danilo Vasconcellos; Sakurai, Kouichi

$$\begin{aligned} & \underset{e(\mathbf{x})^*}{\text{maximize}} \quad f_{adv}(\mathbf{x} + e(\mathbf{x})) \\ & \text{subject to} \quad \|e(\mathbf{x})\| \leq L \end{aligned}$$

# Universal perturbations



**Universal adversarial perturbations (2017)**

Moosavi-Dezfooli; Seyed Mohsen; Fawzi, Alhussein; Fawzi, Omar; Frossard, Pascal

# Carlini & Wagner Attack

Original Adversarial

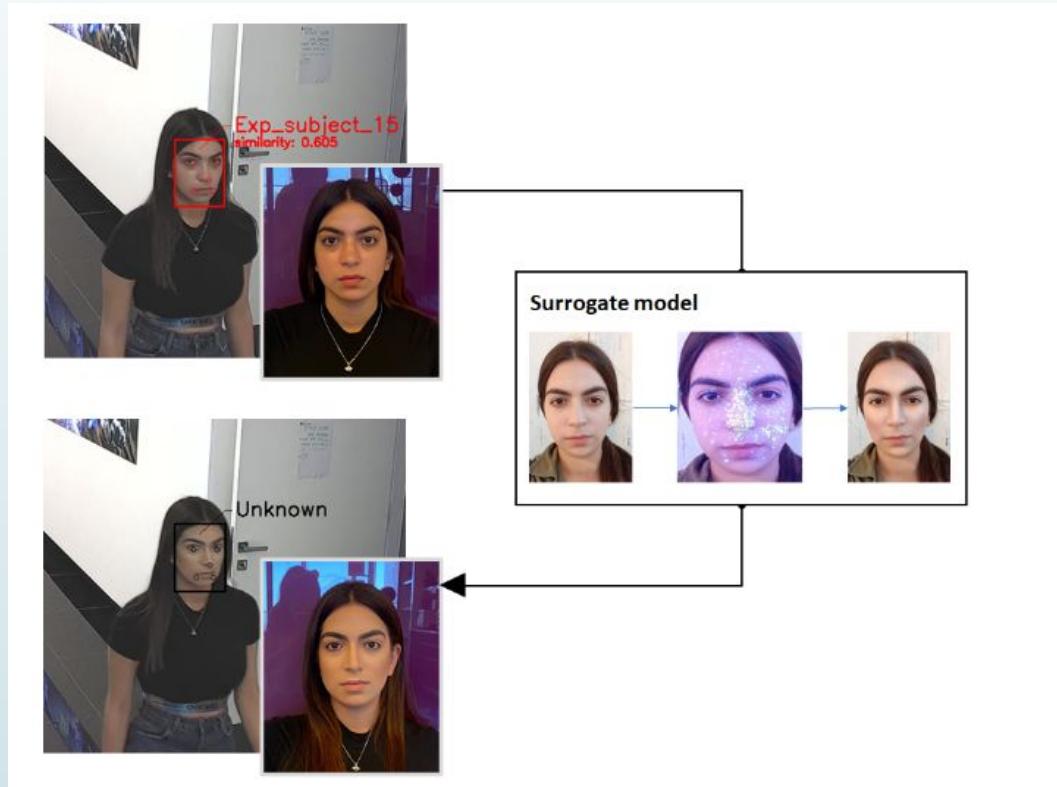


minimize  $D(x, x + \delta)$   
such that  $C(x + \delta) = t$   
 $x + \delta \in [0, 1]^n$

Towards Evaluating the Robustness of Neural Networks (2017)

Carlini, Nicholas; Wagner, David

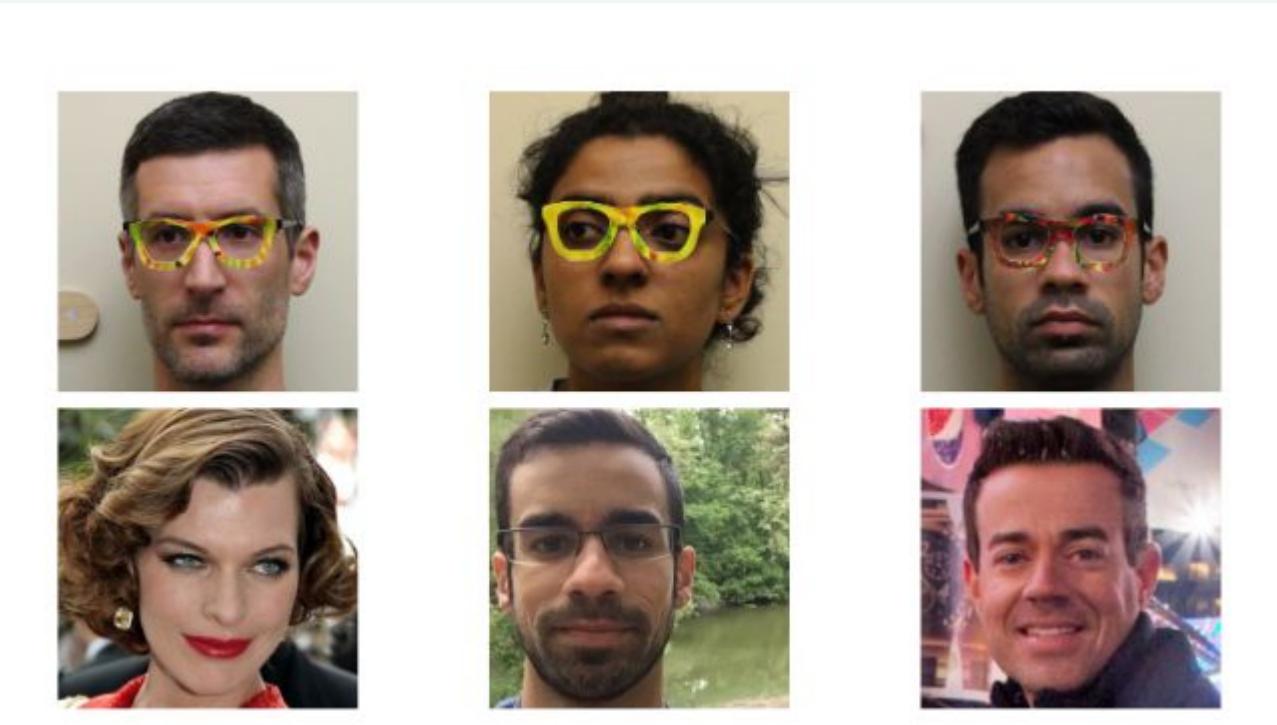
# Aplicações dos ataques no mundo real



## Dodging Attack Using Carefully Crafted Natural Makeup (2021)

Nitzan Guett; Asaf Shabtai; Inderjeet Singh; Satoru Momiyama; Yuval Elovici

# Aplicações dos ataques no mundo real



**Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition (2016)**

Mahmood Sharif; Sruti Bhagavatula; Lujo Bauer Michael K. Reiter

# Aplicações dos ataques no mundo real

[Link do projeto](#)

**Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition (2016)**

Mahmood Sharif; Sruti Bhagavatula; Lujo Bauer Michael K. Reiter

# Os impactos dos ataques adversários

**Table 2.** Test accuracy of LeNet5 measured between 20 and 50 epochs. The network is trained with data augmentation including random rotation, random translation and random erasing.

		Maximum				Median			
		Clean	FGSM	JSMA	C&W	Clean	FGSM	JSMA	C&W
Trained model	Clean	99.42	50.23	64.07	55.28	99.23	42.01	61.06	46.07
	FGSM	99.40	99.31	63.90	75.67	99.24	99.27	60.34	65.08
	JSMA	99.42	50.35	96.75	81.20	99.30	48.46	96.35	75.02
	C&W	99.37	74.12	69.00	99.57	99.22	61.17	63.30	99.31
	FGSM + JSMA	99.40	99.32	96.29	90.05	99.28	99.26	95.15	84.94
	FGSM + C&W	99.43	99.32	65.27	99.65	99.24	99.26	62.96	99.42
	JSMA + C&W	99.43	73.98	96.70	98.99	99.35	70.71	95.83	98.58
	FGSM + JSMA + C&W	99.36	99.30	95.82	98.81	99.27	99.21	94.92	98.39

**On the Effectiveness of Adversarial Training in Defending against Adversarial Example Attacks for Image Classification (2020)**  
Ramani Sagar; Rutvij Jhaveri; Carlos Borrego

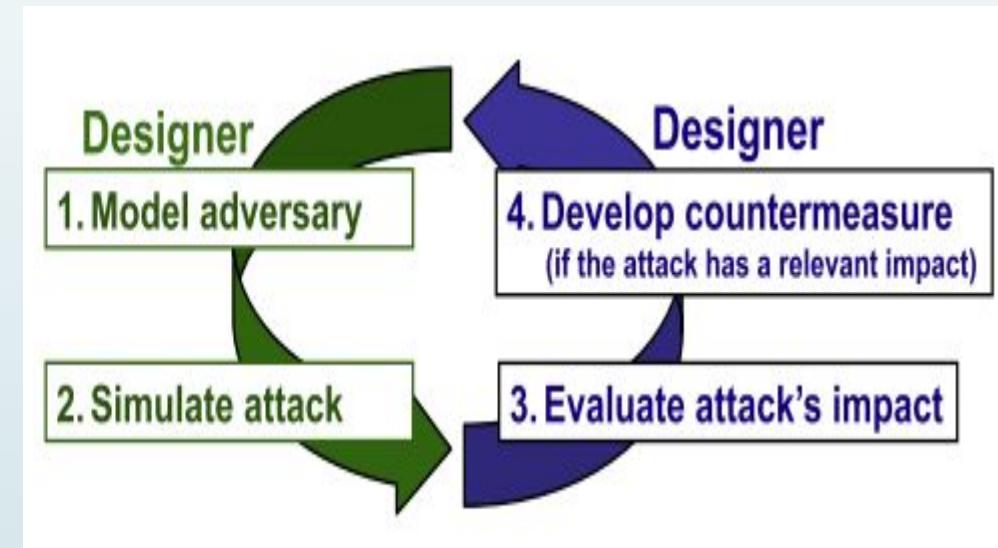
# Código



# Defesa — MLSec



# Projeto de um modelo defensivo



**Wild patterns: Ten years after the rise of adversarial machine learning (2018)**

Biggio, Battista; Roli, Fabio

# Projeto de um modelo defensivo

## Reactive Defenses

1. timely detection of attacks
2. frequent retraining
3. decision verification

## Proactive Defenses

*Security-by-Design Defenses  
against white-box attacks (no probing)*

1. secure/robust learning
2. attack detection

*Effect on decision boundaries:  
noise-specific margin,  
enclosure of legitimate training classes*

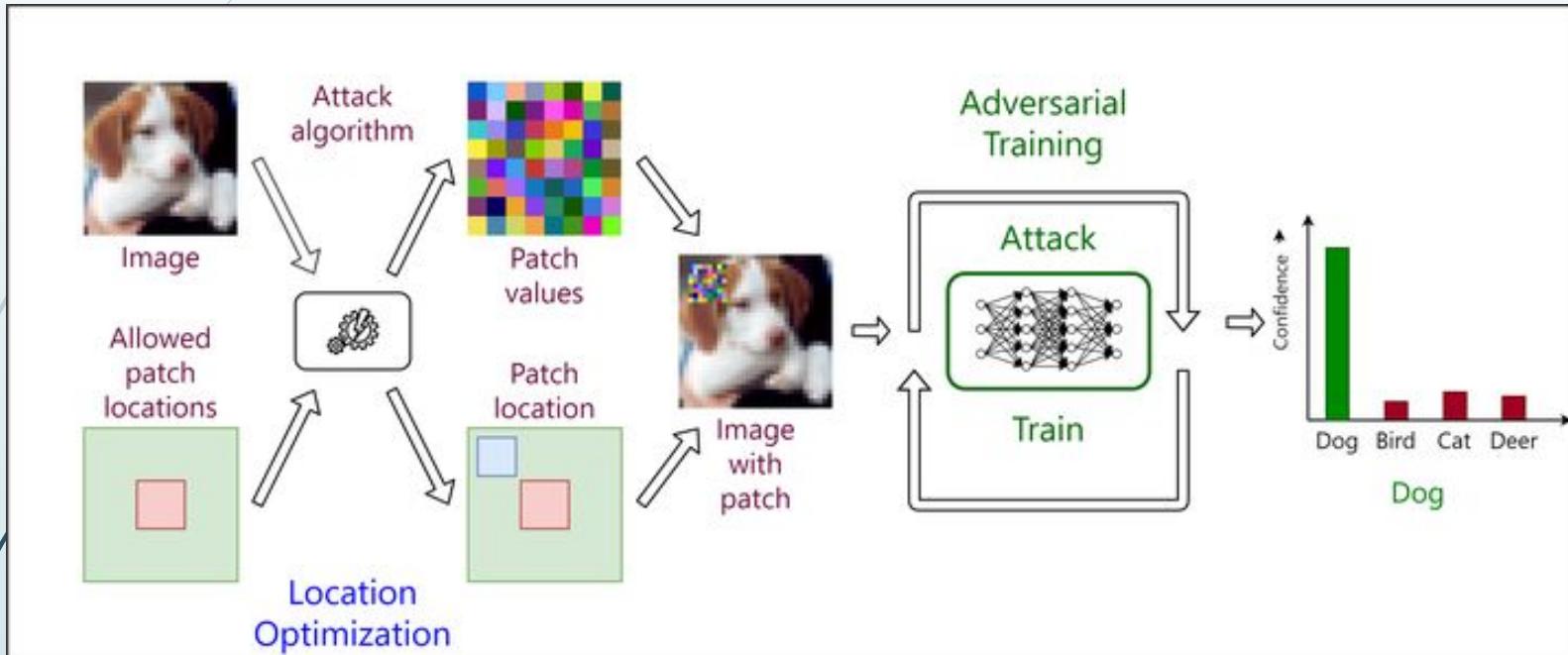
*Security-by-Obscurity Defenses  
against gray-box and black-box attacks (probing)*

1. information hiding, randomization
2. detection of probing attacks

**Wild patterns: Ten years after the rise of adversarial machine learning (2018)**

Biggio, Battista; Roli, Fabio

# Treino adversário

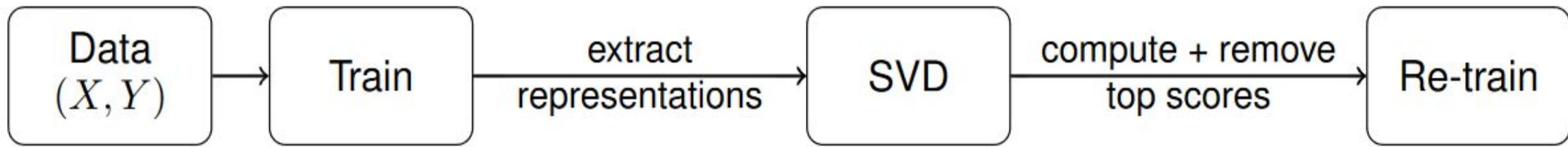


**Adversarial Training against Location-Optimized Adversarial Patches (2020)**  
Sukrut Rao, David Stutz, Bernt Schiele

$$\min_{\theta} \max_{\delta \in S} L(\theta, x + \delta, y).$$

**Towards Deep Learning Models Resistant to Adversarial Attacks (2018)**  
Madry, Aleksander; Makelov, Aleksandar; Schmidt, Ludwig; Tsipras, Dimitris; Vladu, Adrian

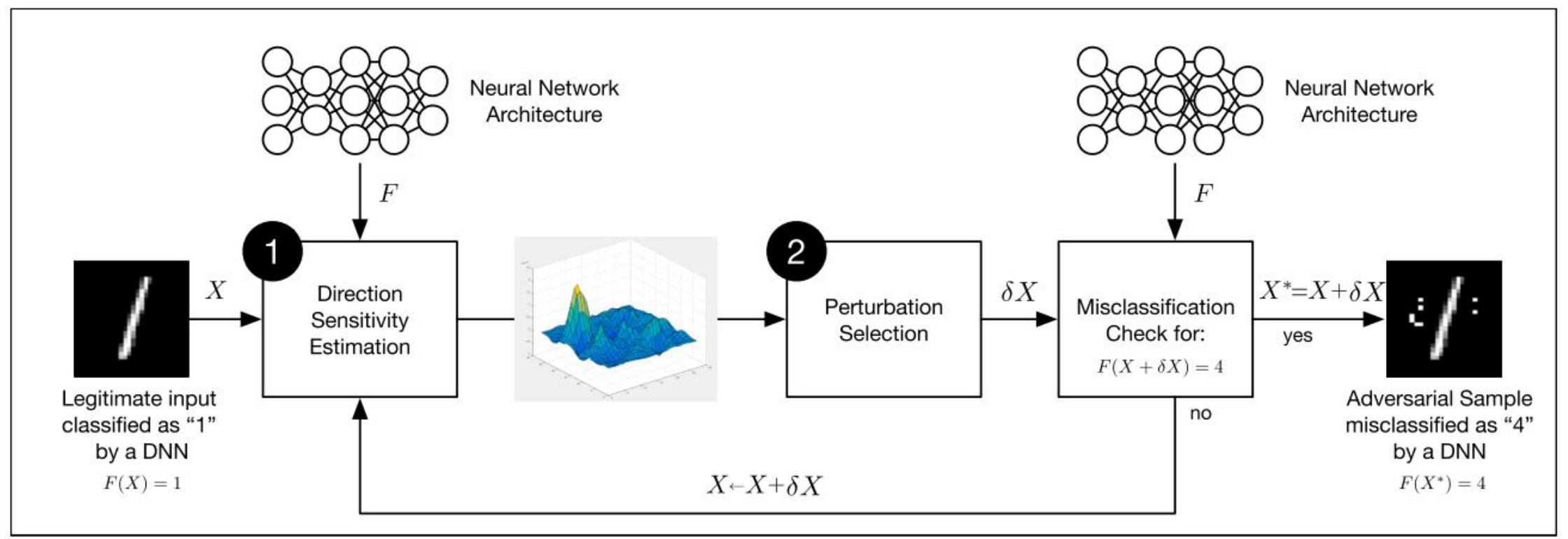
# Detectção de backdoors - Spectral Signatures



**Spectral Signatures in Backdoor Attacks (2018)**

Brandon Tran, Jerry Li, Aleksander Madry

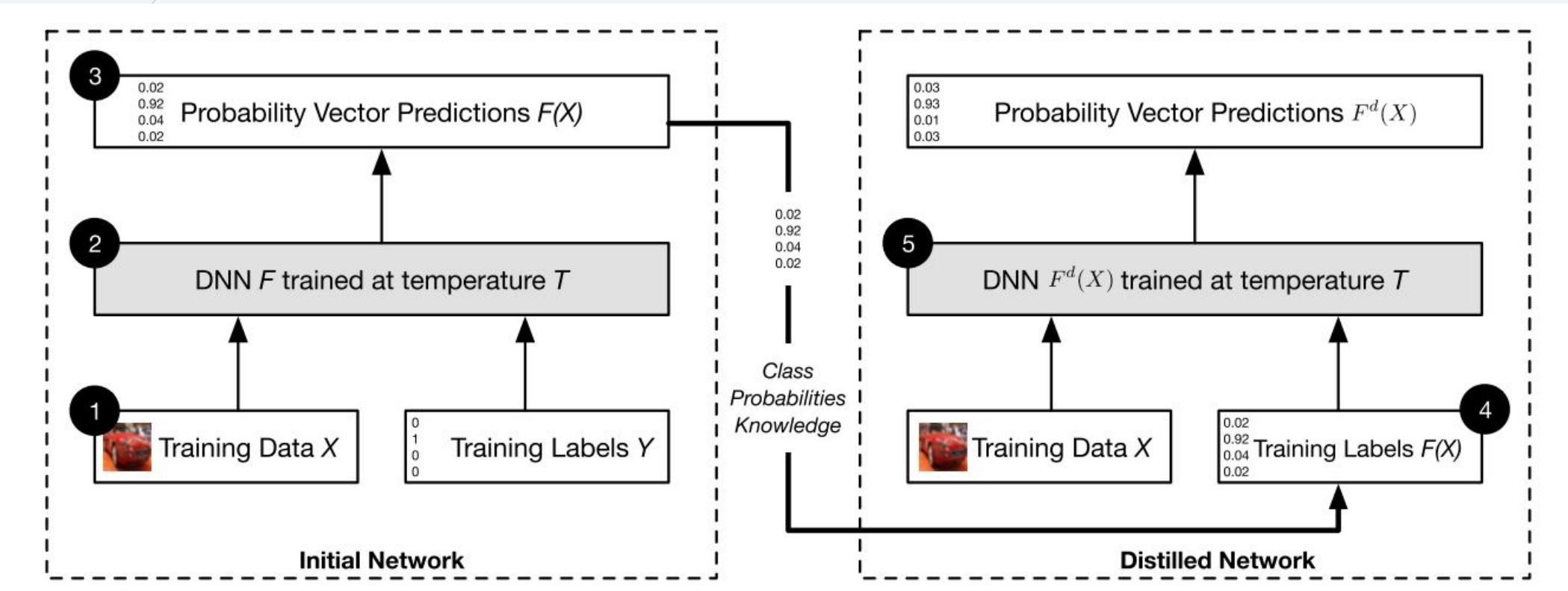
# Defensive distillation - ataque



Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks (2016)

Nicolas Papernot; Patrick McDaniel, Xi Wu; Somesh Jha; Ananthram Swami

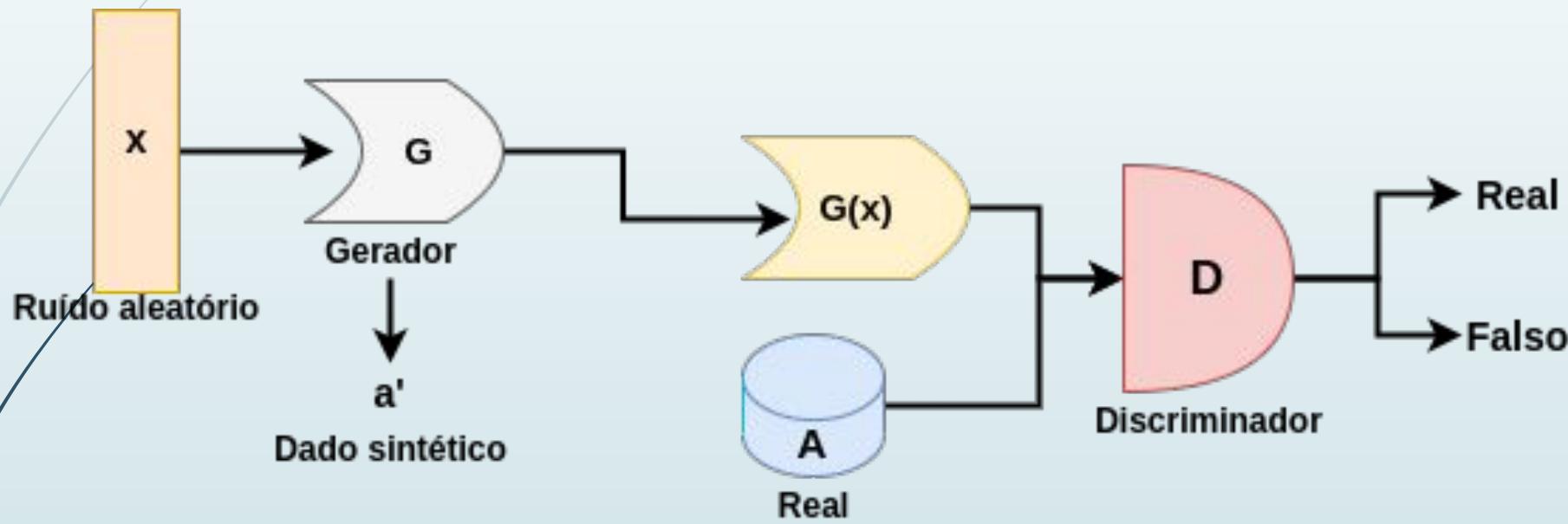
# Defensive distillation - defesa



Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks (2016)

Nicolas Papernot; Patrick McDaniel, Xi Wu; Somesh Jha; Ananthram Swami

# MagNet - GAN

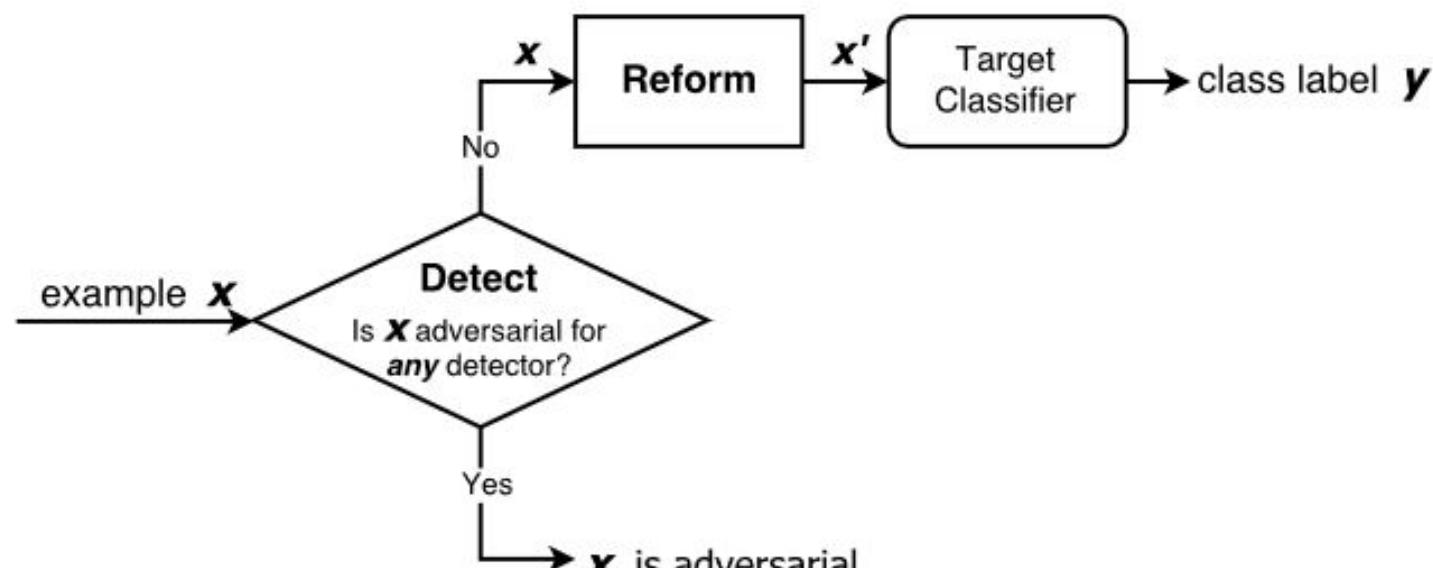


[Exemplo de GAN](#)

MagNet: a Two-Pronged Defense against Adversarial Examples (2017)

Dongyu Meng; Hao Chen

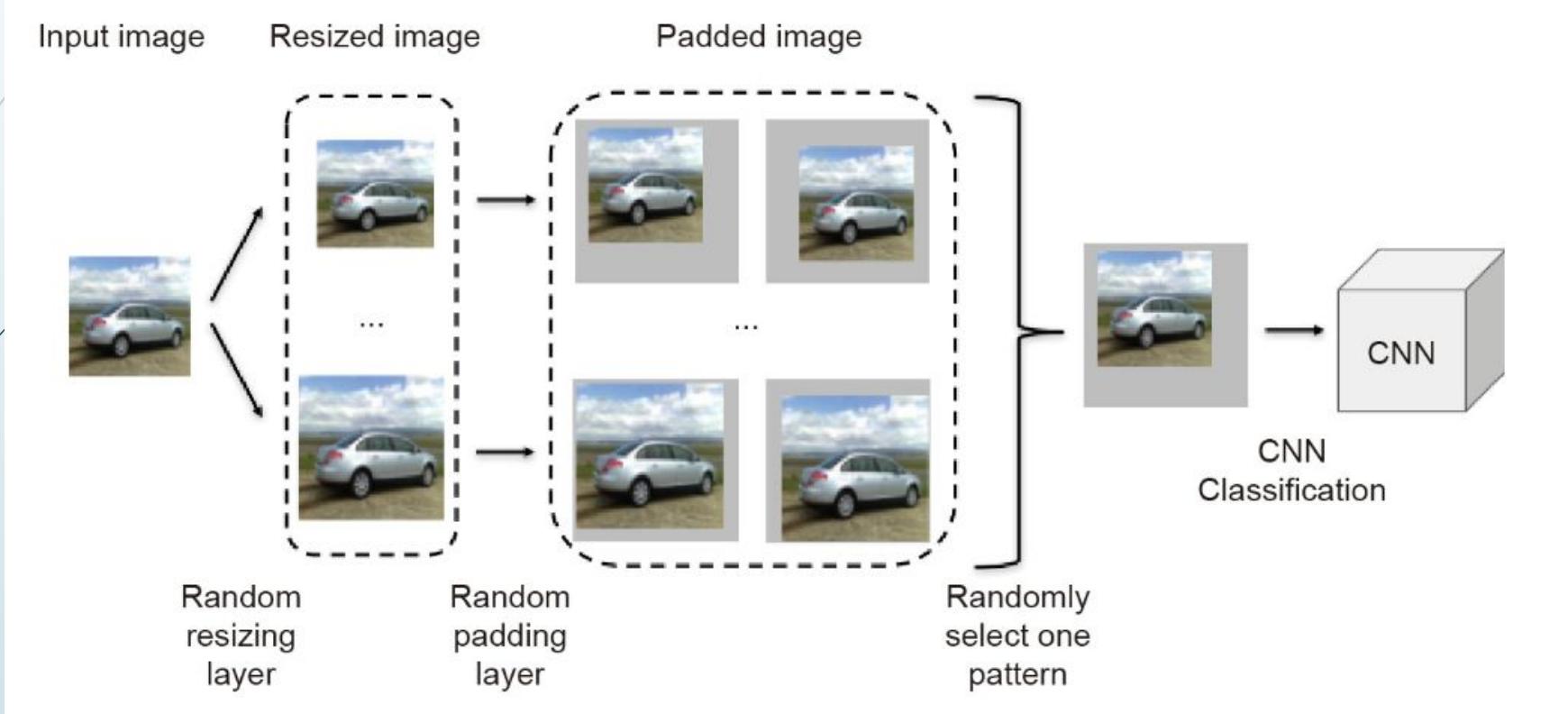
# MagNet



**MagNet: a Two-Pronged Defense against Adversarial Examples (2017)**

Dongyu Meng; Hao Chen

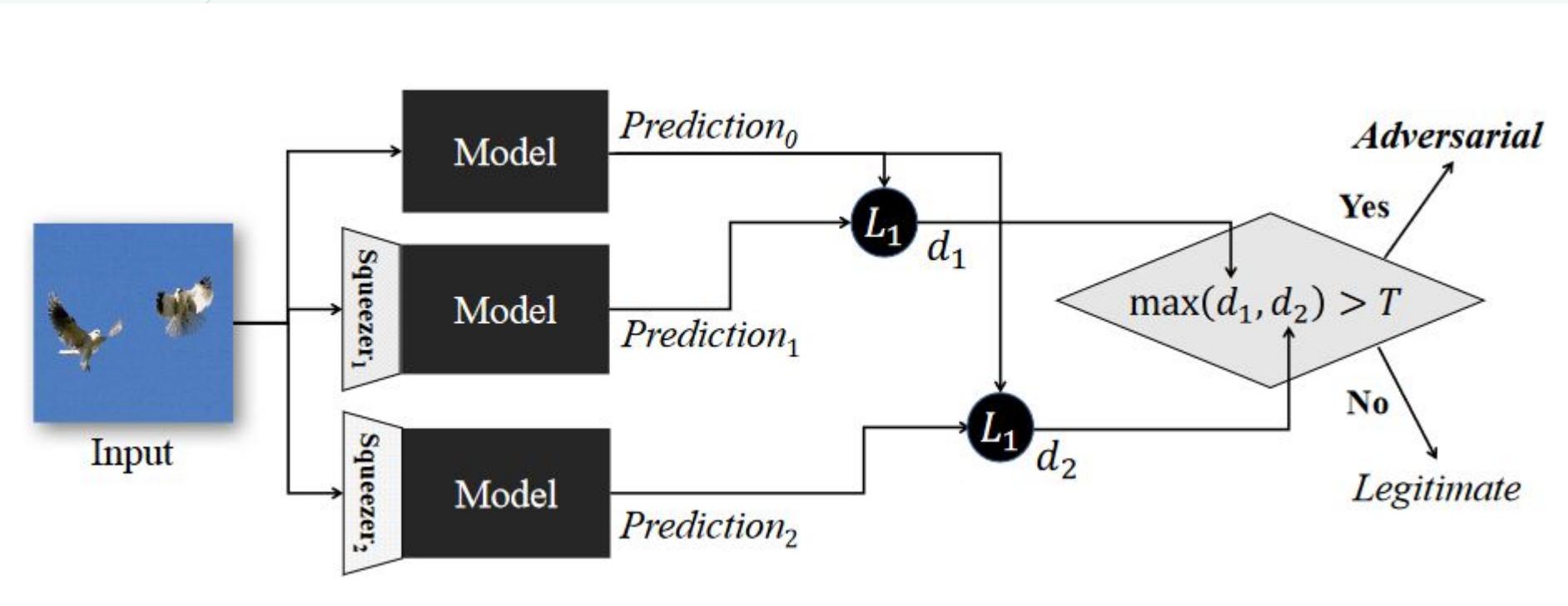
# Randomização



**Mitigating Adversarial Effects Through Randomization (2017)**  
Cihang Xie; Jianyu Wang; Zhishuai Zhang; Zhou Ren; Alan Yuille

**Adversarial Attacks and Defenses in Deep Learning (2019)**  
Kui Ren; Tianhang Zheng; Zhan Qin; Xue Liu

# feature-squeezing



Adversarial Attacks and Defenses in Deep Learning (2019)

Kui Ren; Tianhang Zheng; Zhan Qin; Xue Liu

Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks (2017)

Weilin Xu; David Evans; Yanjun Qi

# Outros métodos

Research  
Artificial Intelligence—Feature Article

## Adversarial Attacks and Defenses in Deep Learning

Kui Ren <sup>a,b,\*</sup>, Tianhang Zheng <sup>c</sup>, Zhan Qin <sup>a,b</sup>, Xue Liu <sup>d</sup>

<sup>a</sup> Institute of Cyberspace Research, Zhejiang University, Hangzhou 310027, China

<sup>b</sup> College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

<sup>c</sup> Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 2E8, Canada

<sup>d</sup> School of Computer Science, McGill University, Montreal, QC H3A 0E9, Canada

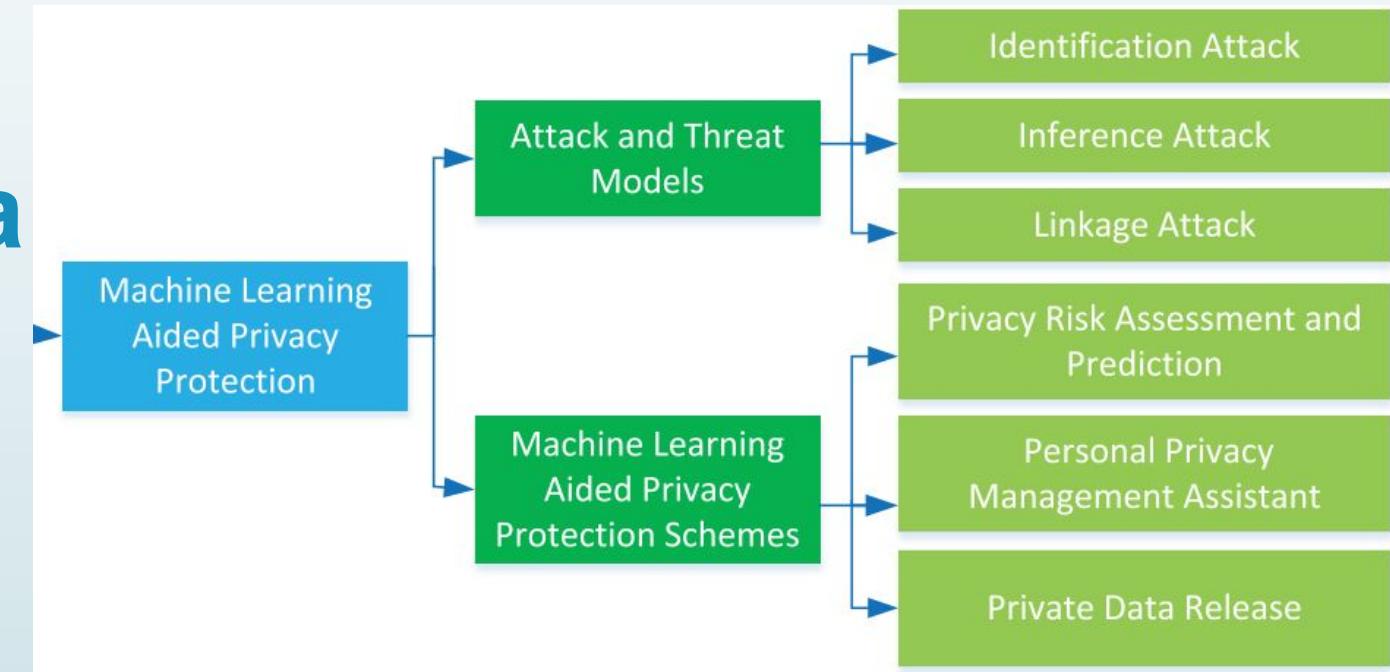
Adversarial Attacks and Defenses in Deep Learning (2019)

Kui Ren; Tianhang Zheng; Zhan Qin; Xue Liu

# Privacidade — MLSec Problemas e soluções

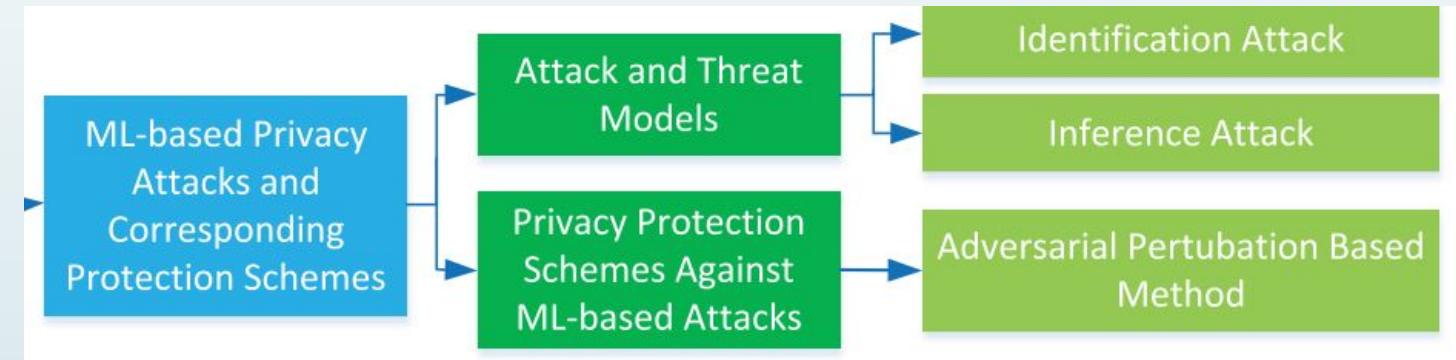


# Privacidade baseada em ML



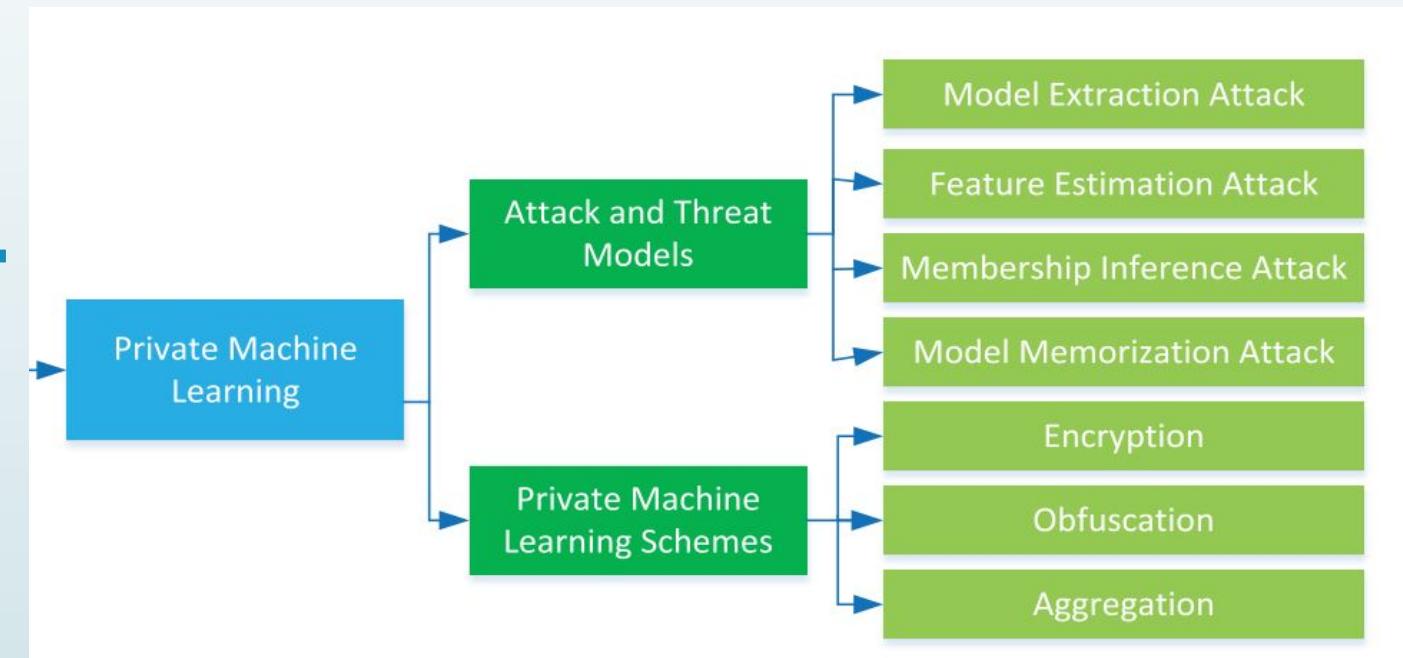
**When Machine Learning Meets Privacy: A Survey and Outlook (2021)**  
Liu, Bo; Ding, Ming; Shaham, Sina; Rahayu, Wenny; Farokhi, Farhad; Lin, Zihuai

# Mecanismos de proteção



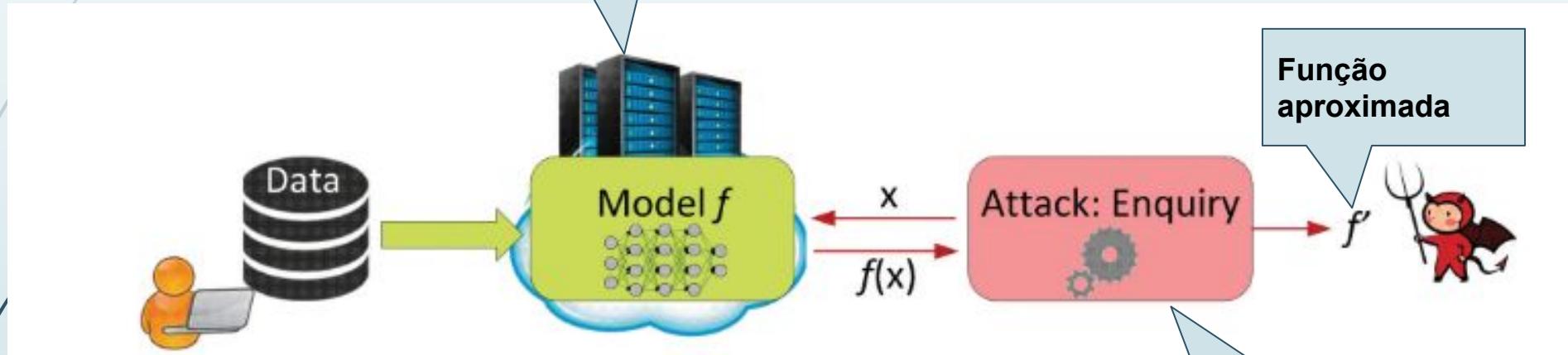
**When Machine Learning Meets Privacy: A Survey and Outlook (2021)**  
Liu, Bo; Ding, Ming; Shaham, Sina; Rahayu, Wenny; Farokhi, Farhad; Lin, Zihuai

# Privacidade em ML (Modelos)



**When Machine Learning Meets Privacy: A Survey and Outlook (2021)**  
Liu, Bo; Ding, Ming; Shaham, Sina; Rahayu, Wenny; Farokhi, Farhad; Lin, Zihuai

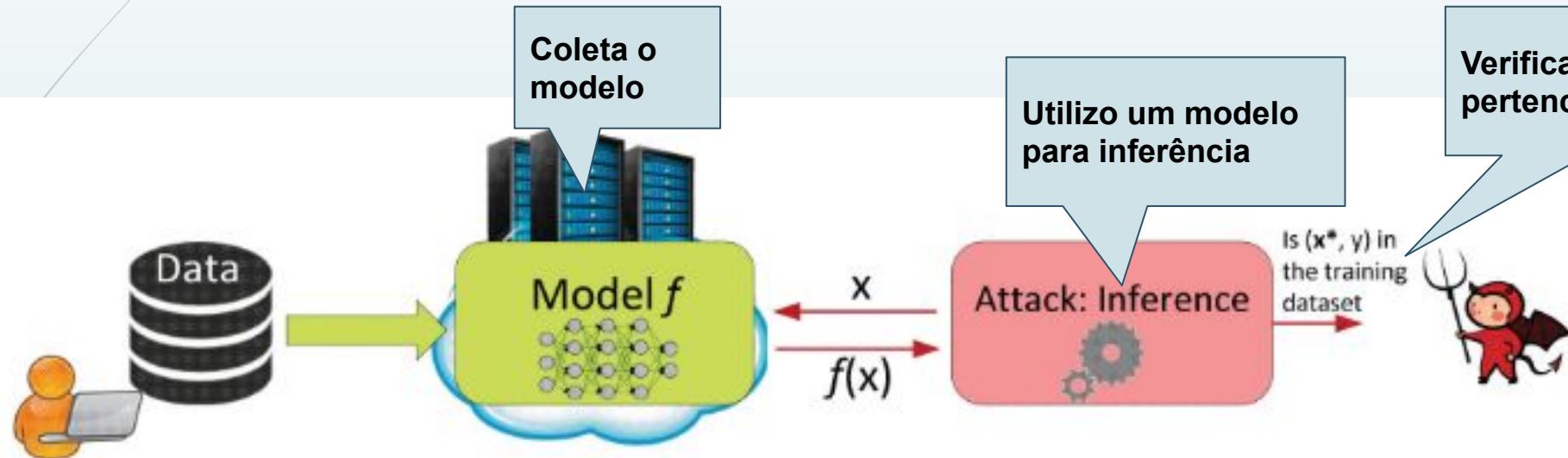
# Model extraction attack



When Machine Learning Meets Privacy: A Survey and Outlook (2021)

Liu, Bo; Ding, Ming; Shaham, Sina; Rahayu, Wenny; Farokhi, Farhad; Lin, Zihuai

# Membership inference attack

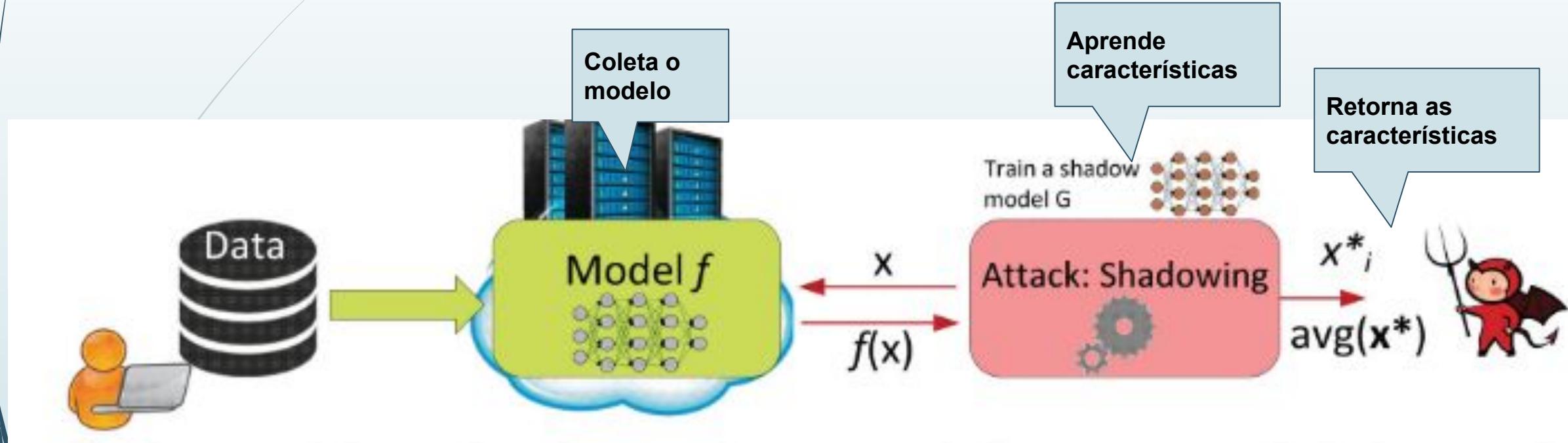


**Membership Inference Attack:** Adversary learns whether a given data record  $(x^*, y)$  is part of the model's training dataset  $D$  or not

When Machine Learning Meets Privacy: A Survey and Outlook (2021)

Liu, Bo; Ding, Ming; Shaham, Sina; Rahayu, Wenny; Farokhi, Farhad; Lin, Zihuai

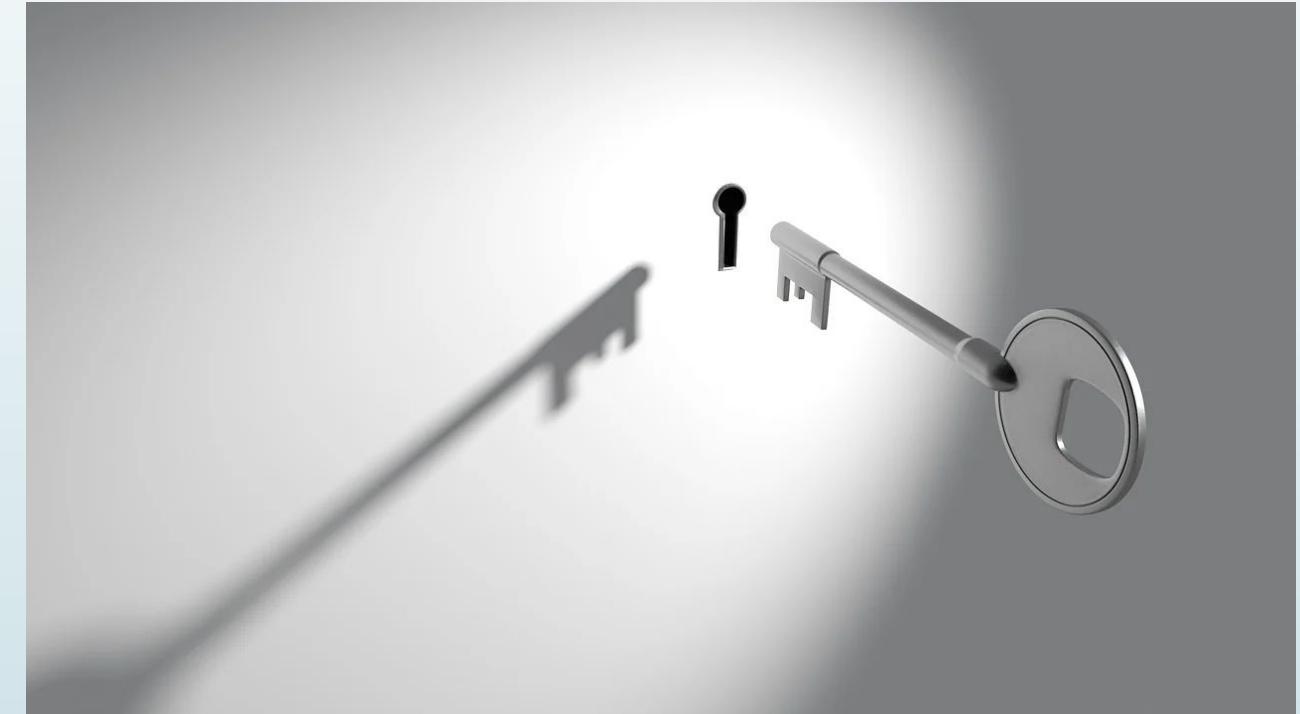
# Feature Estimation Attack



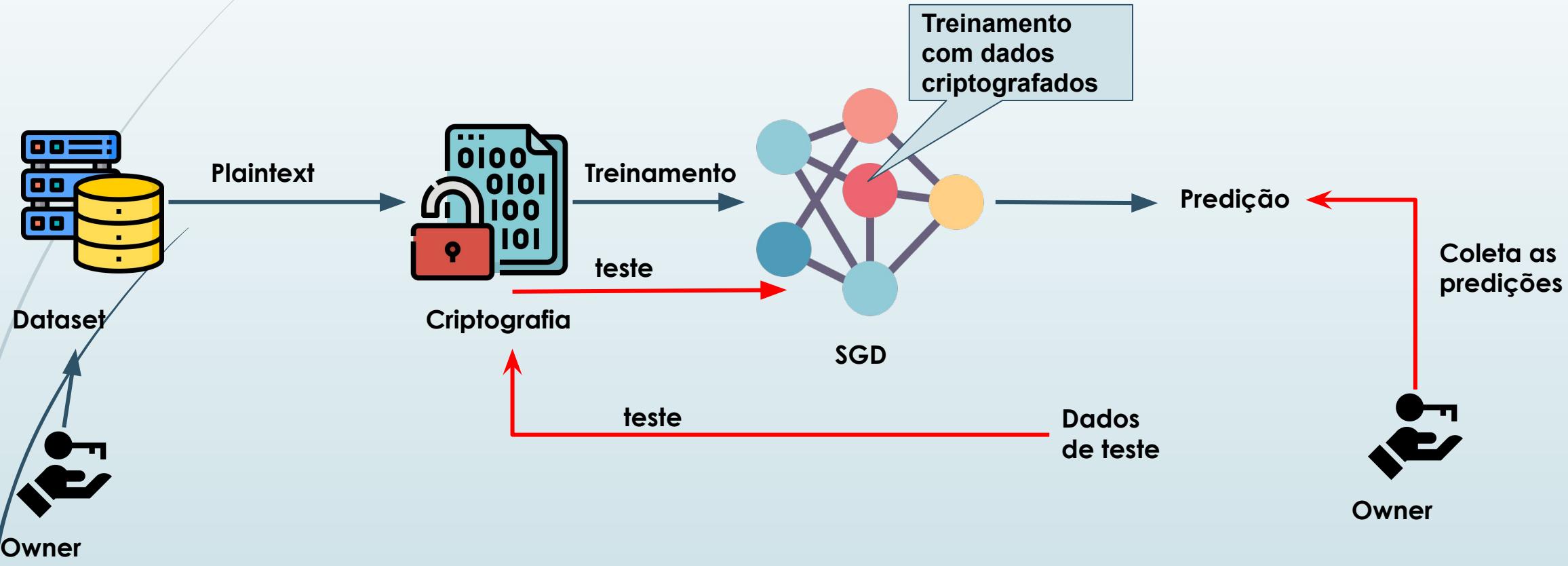
When Machine Learning Meets Privacy: A Survey and Outlook (2021)

Liu, Bo; Ding, Ming; Shaham, Sina; Rahayu, Wenny; Farokhi, Farhad; Lin, Zihuai

## Mitigando problemas



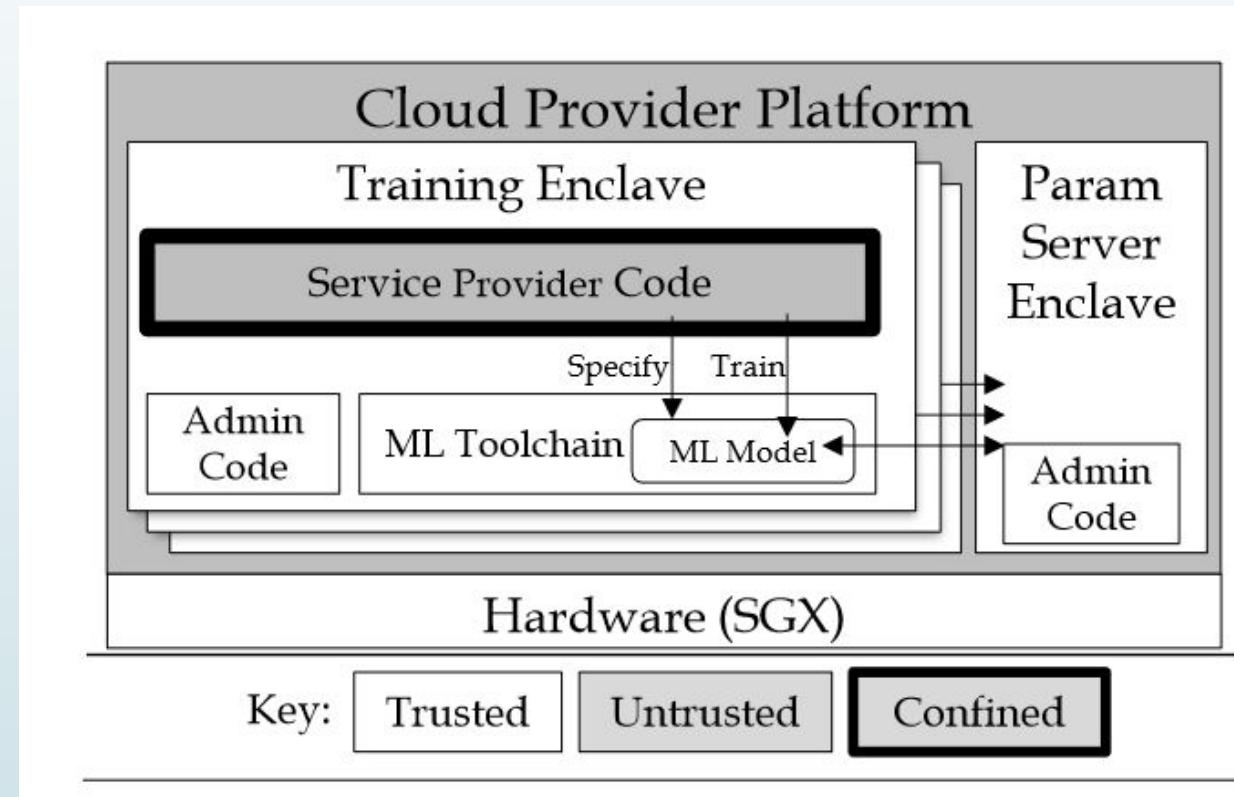
# Homomorphic encryption



**CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy (2016)**

Nathan Dowlin; Ran Gilad-Bachrach; Kim Laine; Kristin Lauter; Michael Naehrig; John Wernsing

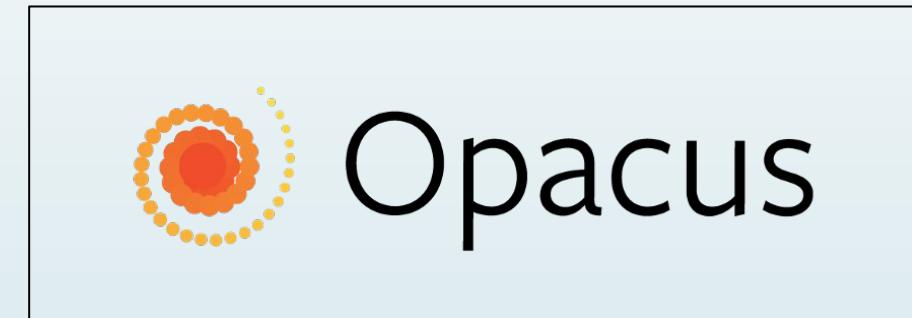
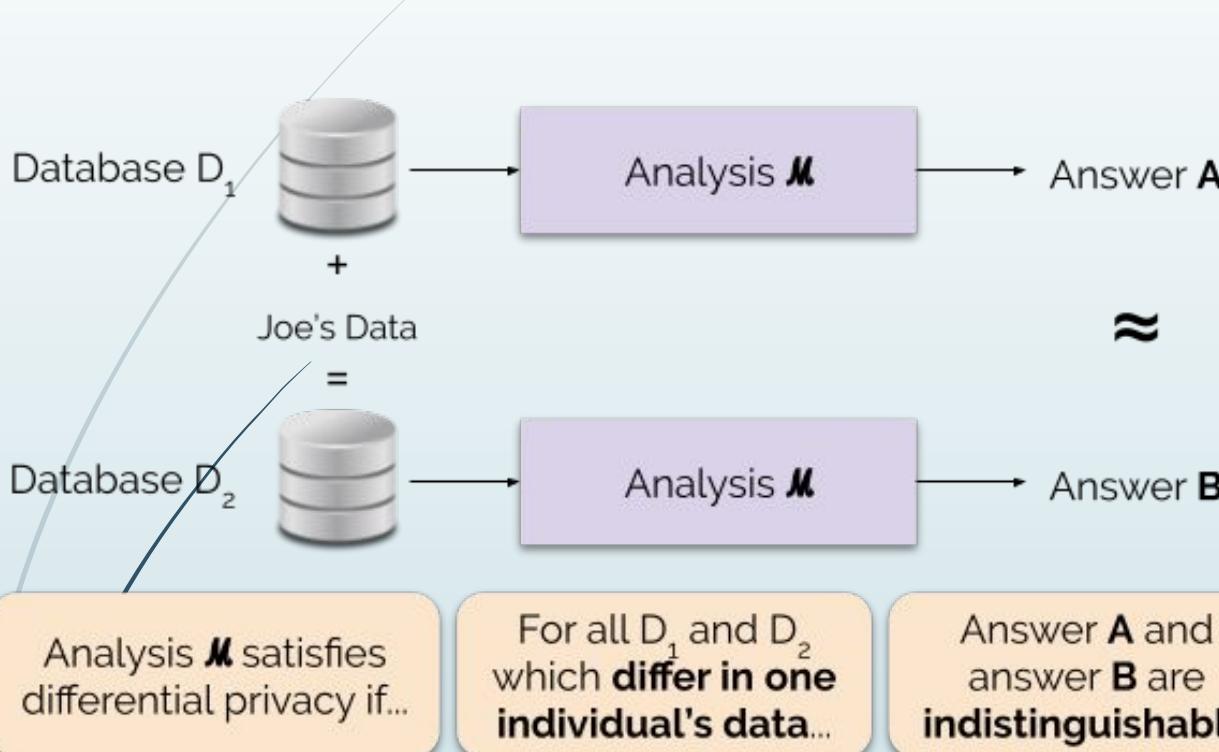
# Homomorphic encryption + SGX



**Chiron: Privacy-preserving Machine Learning as a Service (2018)**

Tyler Hunt, Congzheng Song, Reza Shokri, Vitaly Shmatikov, Emmett Witchel

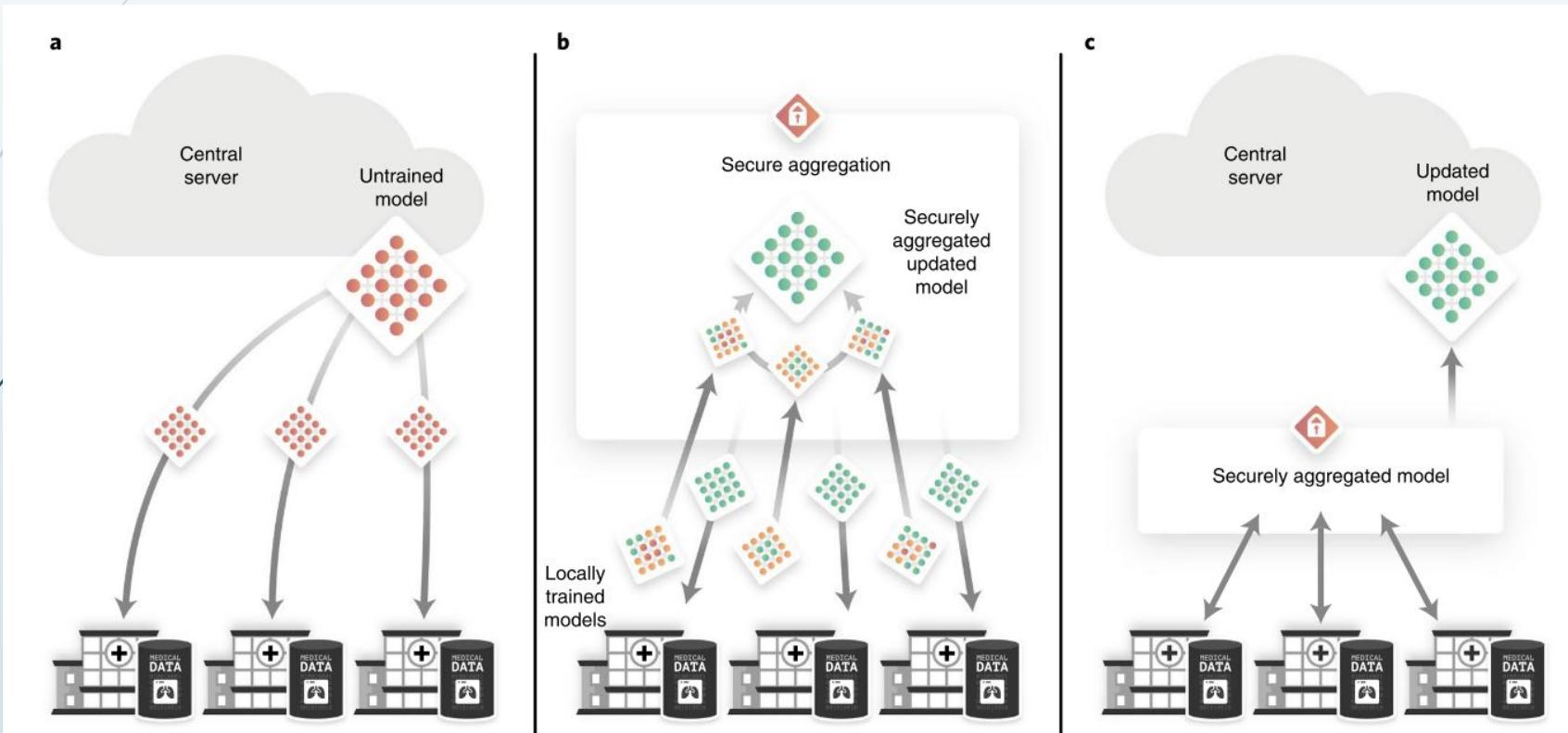
## Differential privacy - SGD



**A Differentially Private Stochastic Gradient Descent Algorithm for Multiparty Classification (2012)**

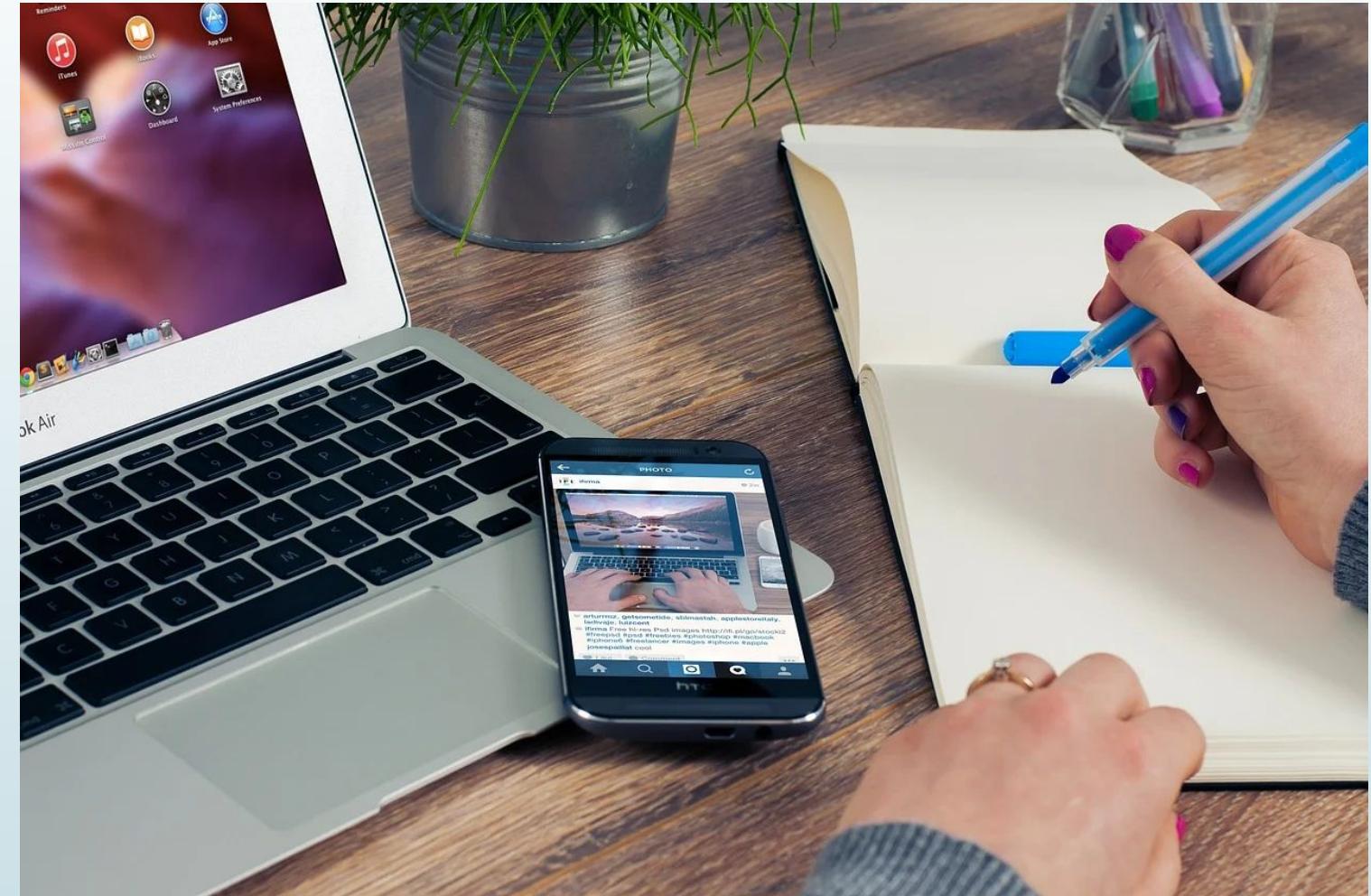
Arun Rajkumar; Shivani Agarwal

# Treinamento distribuído



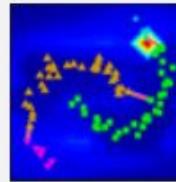
**End-to-end privacy preserving deep learning on multi-institutional medical imaging (2021)**  
Kaassis, G., Ziller, A., Passerat-Palmbach, J. et al.

# Projetos



# University of Cagliari, Italy

## Adversarial Clustering

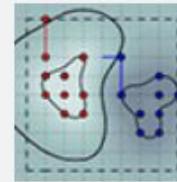


Clustering algorithms have been increasingly adopted in security applications to spot dangerous or illicit activities. However, they have not been originally devised to deal

with deliberate attack attempts that aim to subvert the clustering process itself. Our experimental findings on clustering of malware samples and handwritten digits show that single- and complete-linkage hierarchical clustering can be significantly compromised by carefully targeted attacks.

[Read more](#)

## Adversarial Feature Selection



Despite feature selection algorithms are often used in security-sensitive applications, only few authors have considered the impact of using reduced feature sets on classifier

security against evasion and poisoning attacks. Within this research area, we first show that feature selection algorithms can significantly worsen classifier security against well-crafted attacks. We then propose novel adversary-aware feature selection procedures to counter these threats.

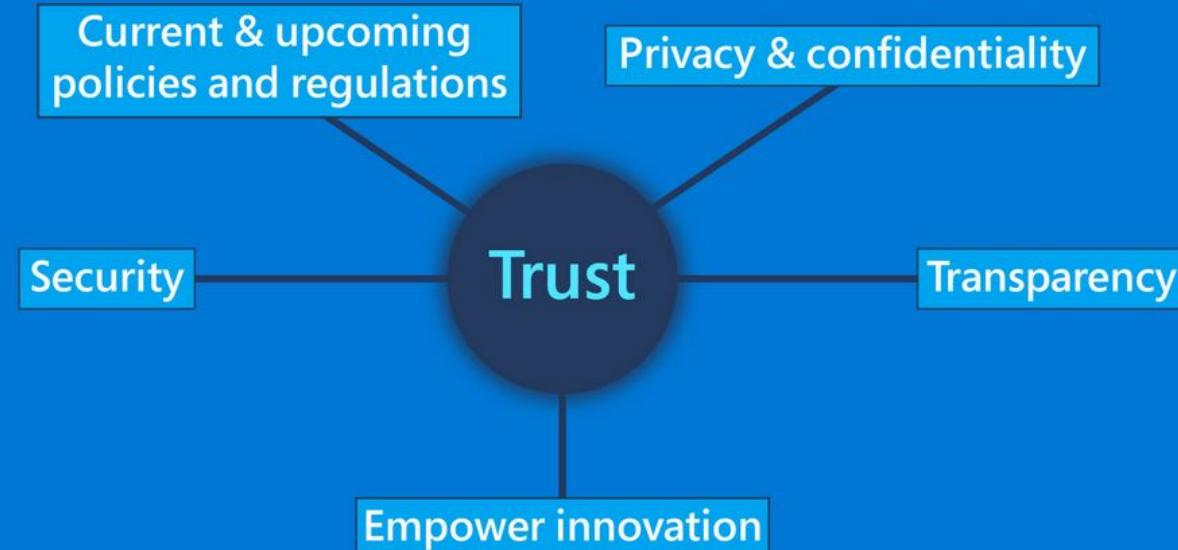
[Read more](#)

## Adversarial Machine Learning

Batista Biggio

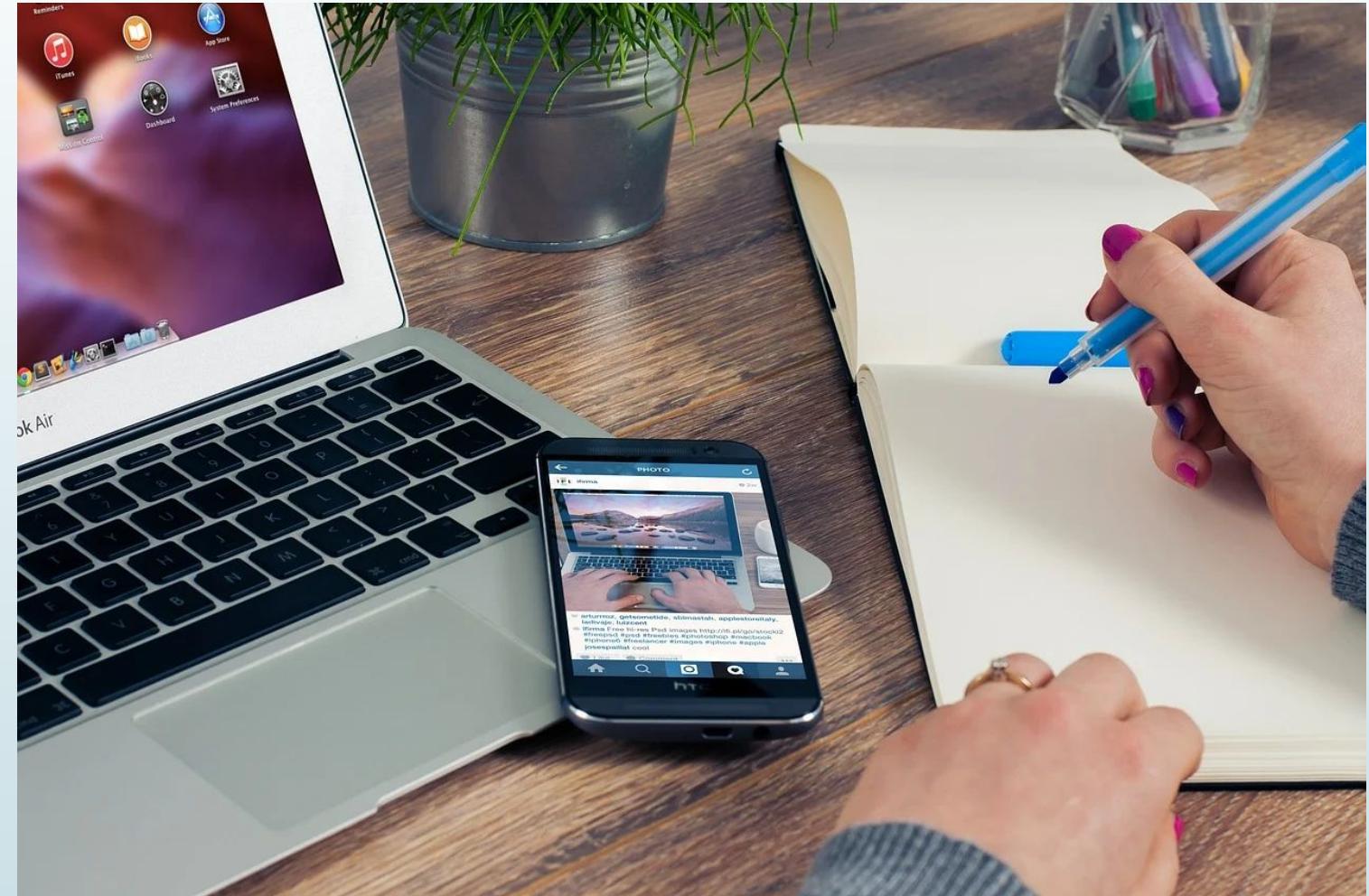
# Microsoft

Privacy Preserving Machine Learning: A holistic approach to protecting privacy



**Privacy Preserving Machine Learning: Maintaining confidentiality and preserving trust**  
Microsoft research

# Dicas



# Dicas

- Algebra linear/ cálculo/ estatística
- Machine Learning: [Pattern Recognition and Machine Learning](#)  
[\(Bishop,2006\)](#)
- Deep Learning - [Deep Learning \(Goodfellow; Bengio; Courville, 2016\)](#)
- Privacidade - [When Machine Learning Meets Privacy: A Survey and Outlook \(Liu et al.,2020\)](#)
- Machine Learning security - [Wild patterns: Ten years after the rise of adversarial machine learning \(Biggio; Roli, 2018\)](#)



# Como enganar modelos de aprendizado de máquina?

## Obrigado!

Erikson Júlio de Aguiar  
[erjulioaguiar@usp.com](mailto:erjulioaguiar@usp.com)

[@erjulioaguiar](https://twitter.com/erjulioaguiar)

[eriksonJAquiar](https://github.com/eriksonJAquiar)

<https://www.linkedin.com/in/erjulioaguiar/>



Slides/Codes