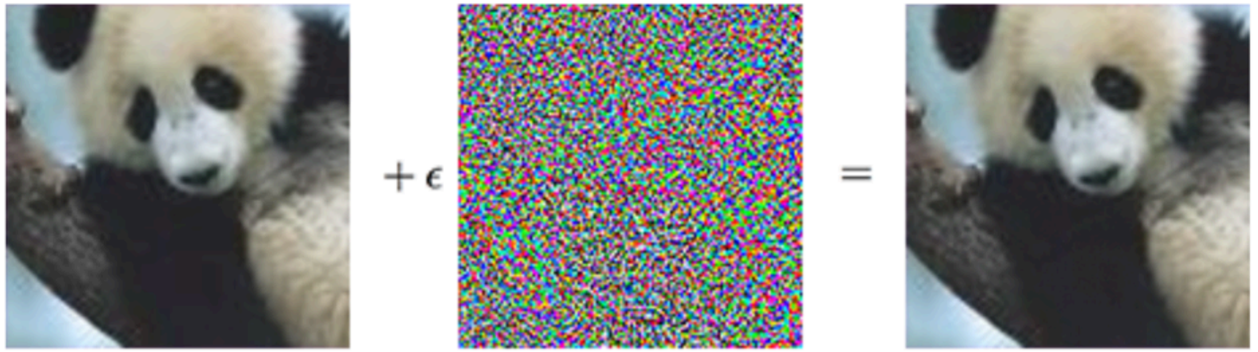
Introduction to Adversarial ML

Jerry Zhu

What can Adversaries do to ML?

Manipulate Classification



https://openai.com/blog/adversarial-example-research/

Manipulate Classification

without the dataset the article is useless

okay google, browse to evil.com

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

Manipulate Regression

31.51

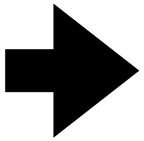
+

=

18.76

BMI [Levin et al 2019]

Physical Attacks



Eykholt et al 2017 https://arxiv.org/abs/1707.08945

Physical Attacks

Brown et al 2018 https://arxiv.org/pdf/1712.09665.pdf

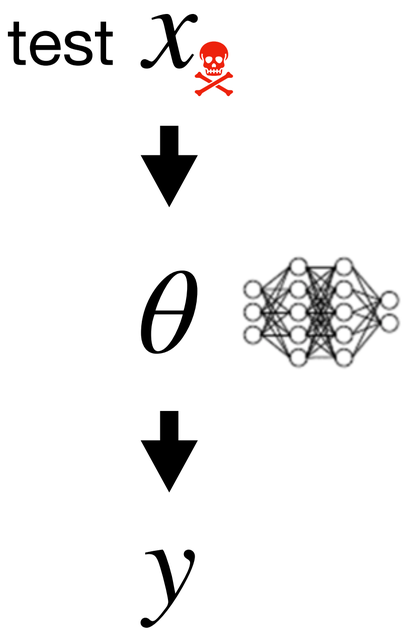
Physical Attacks

Athalye et al 2018 https://arxiv.org/pdf/1707.07397.pdf

Physical Attacks

Sharif et al 2016 https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

Those were Test-Time Attacks



Training-Time Attacks

training data

$D^\dagger$

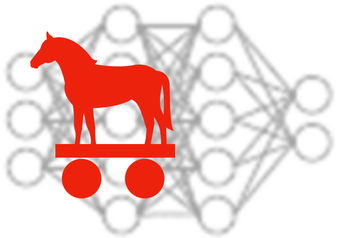
machine learning

$$\min_\theta \mathbb{E}_{D^\dagger} \ell(x, y, \theta)$$

test x

$\theta^\dagger$

y



Mei, Z 2015 http://pages.cs.wisc.edu/~jerryzhu/pub/Mei2015Machine.pdf

Backdoor

x

$x^\dagger$

trigger

$$\theta^\dagger(x) = \theta(x)$$
$$\theta^\dagger(x^\dagger) = \theta(x^\dagger) + 1$$

Gu et al 2019 https://arxiv.org/pdf/1708.06733.pdf

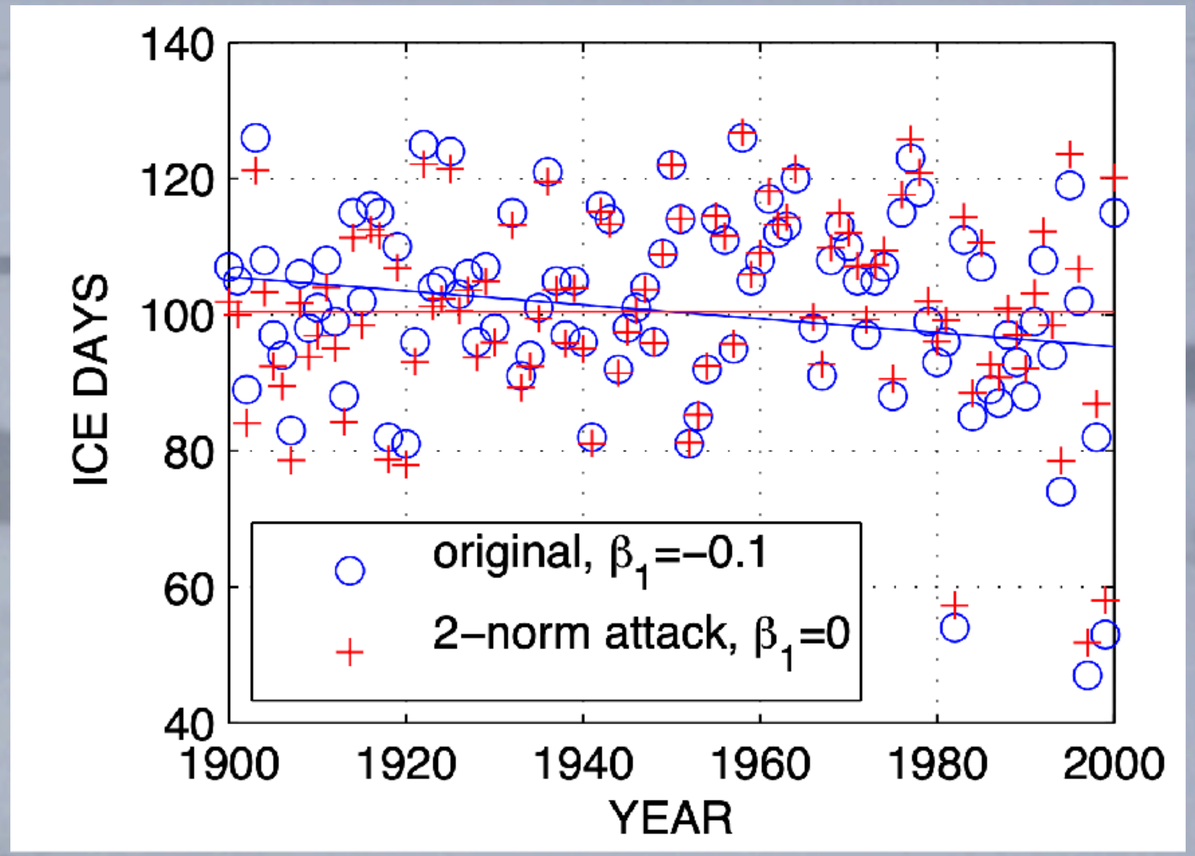
Unfairness

$$\min_{D^\dagger} \frac{P(\theta^\dagger(x) = 1 \mid \text{♀})}{P(\theta^\dagger(x) = 1 \mid \text{♂})}$$

$$\theta^\dagger = \arg\min_{\theta} \mathbb{E}_{D^\dagger} \ell(x, y, \theta)$$

Solans et al 2020 https://arxiv.org/pdf/2004.07401.pdf

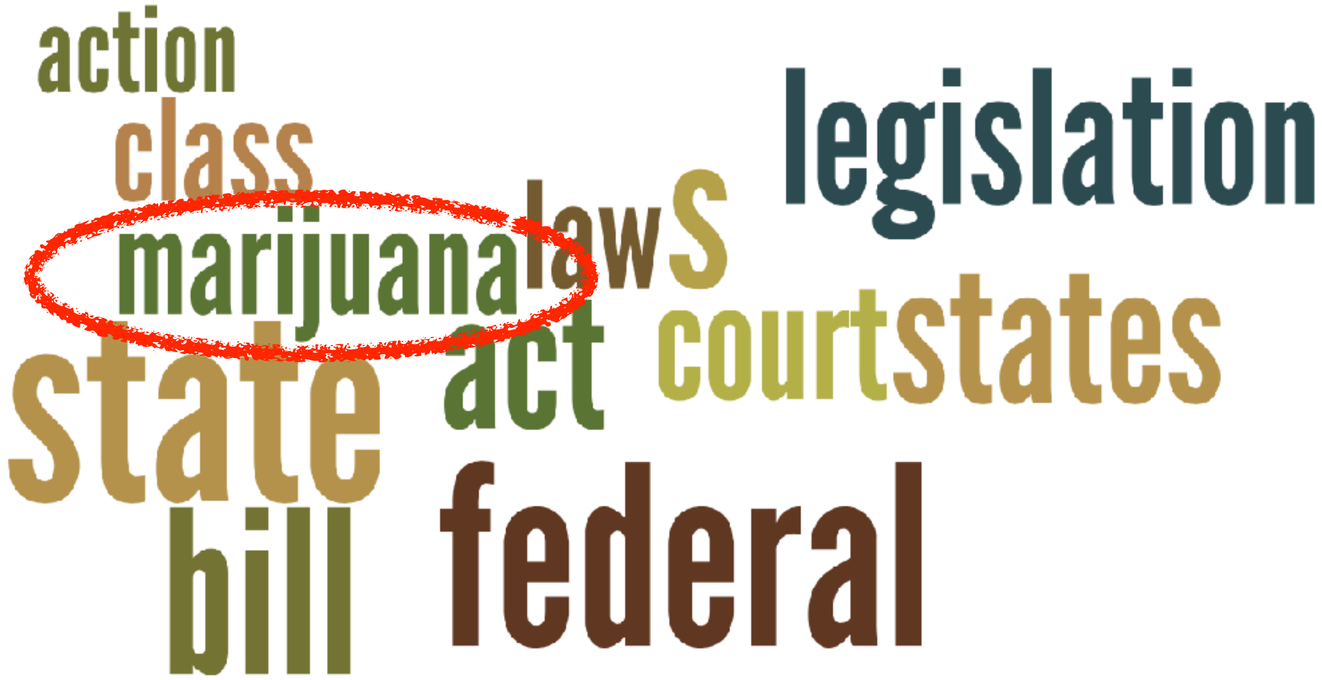
Manipulate Model Interpretation
(linear regression)



Mei, Z 2015 http://pages.cs.wisc.edu/~jerryzhu/pub/Mei2015Machine.pdf

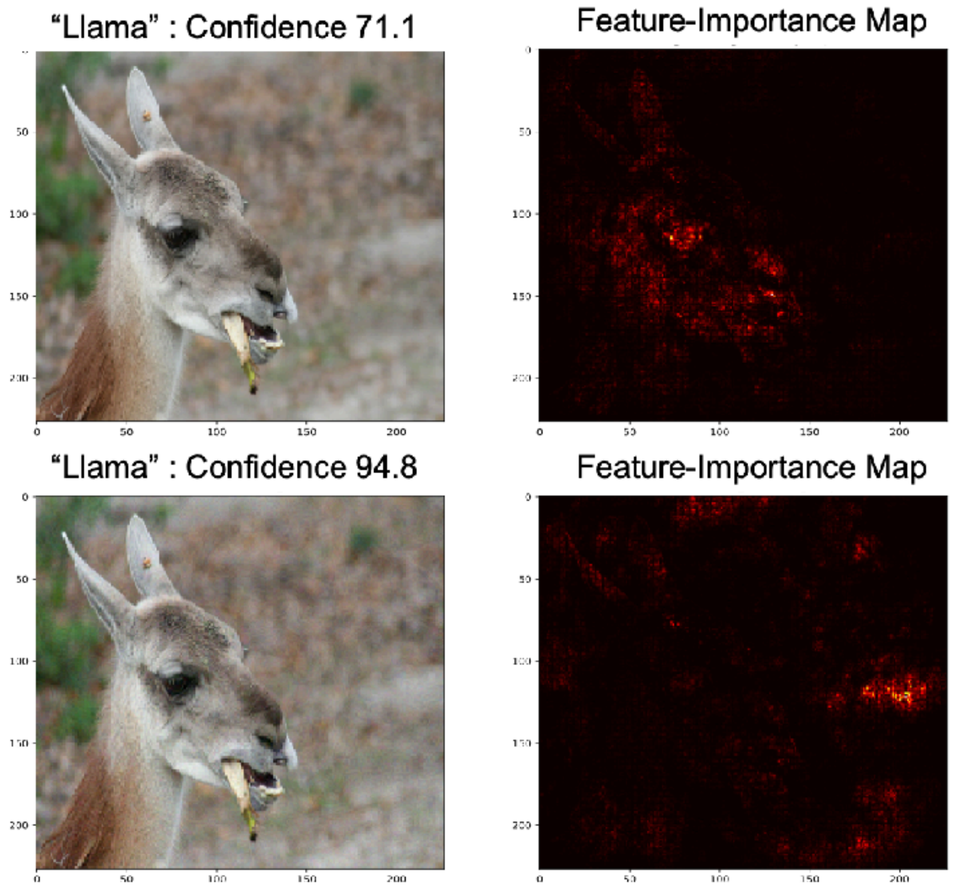
Manipulate Model Interpretation

(latent Dirichlet allocation)



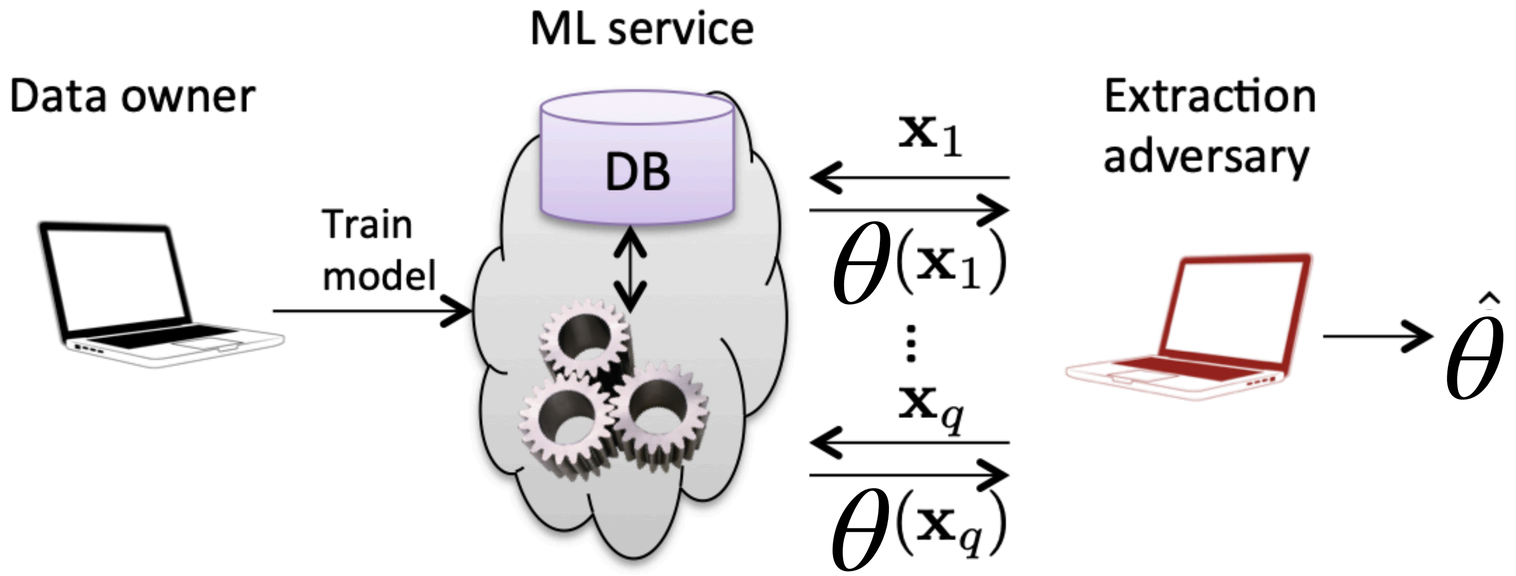
Mei, Z 2015b http://pages.cs.wisc.edu/~jerryzhu/pub/aistatsAttackLDA.pdf

Manipulate Model Interpretation
(deep network attribution)



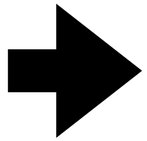
Ghorbani et al 2018 https://arxiv.org/pdf/1710.10547.pdf

Model Stealing



Tramer et al 2016 https://arxiv.org/abs/1609.02943

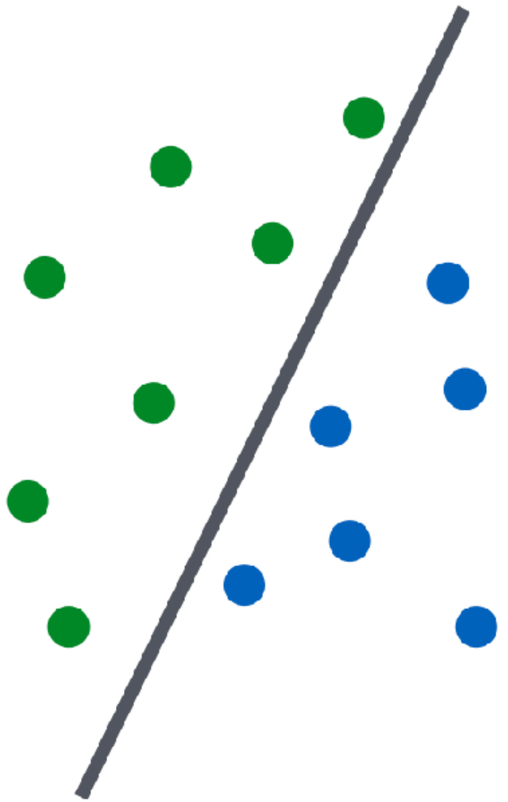
Privacy Identification

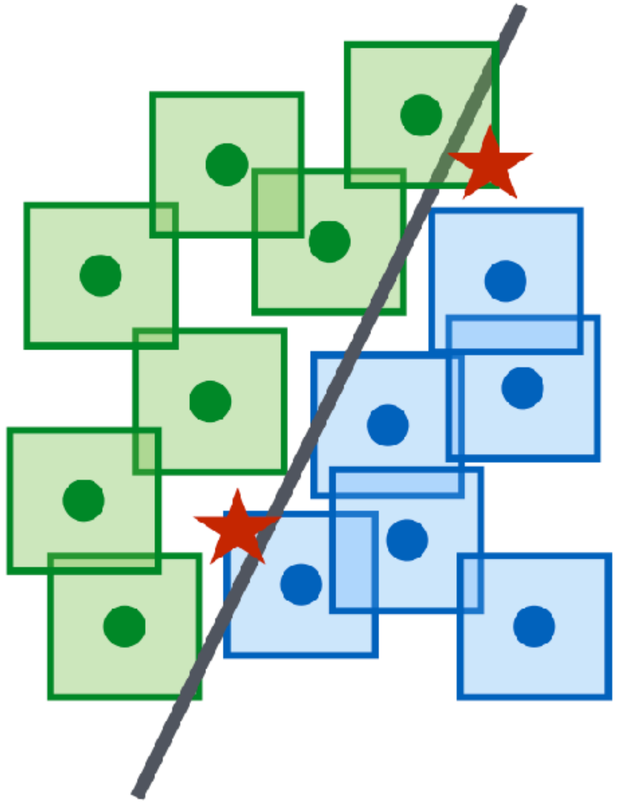


Basic AdvML Math

Test-time Attack

$$\max_{\delta \in \Delta} \ell(x + \delta, y, \theta)$$

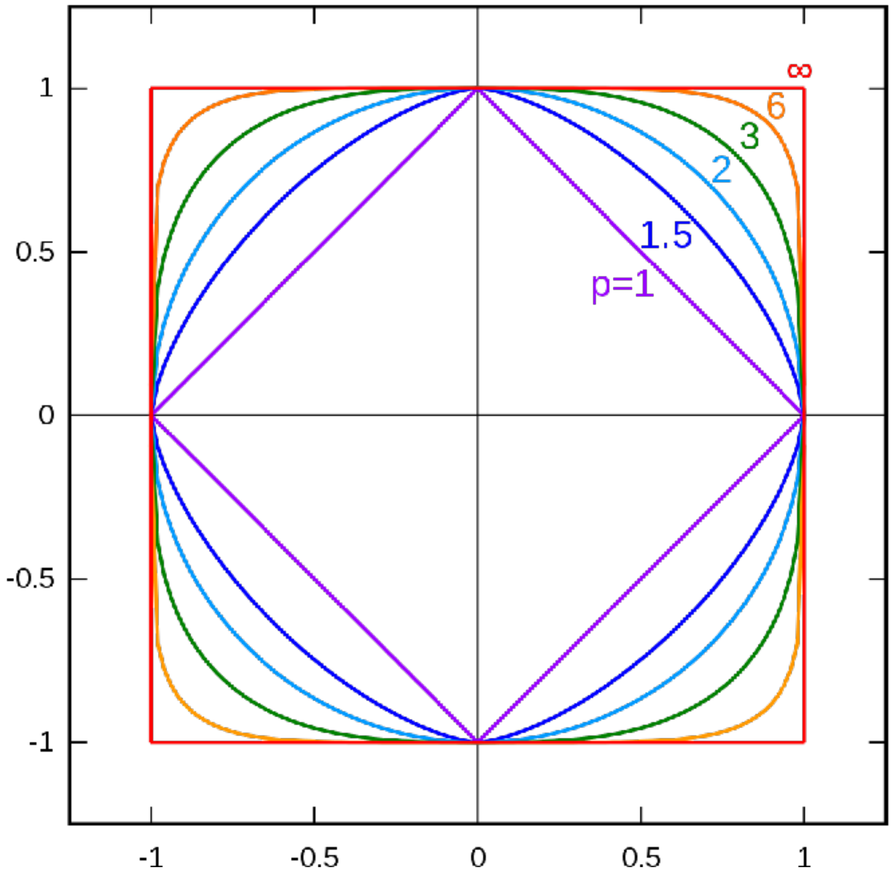




Madry et al 2019 https://arxiv.org/pdf/1706.06083.pdf

Feasible Set Δ

p-norm ball



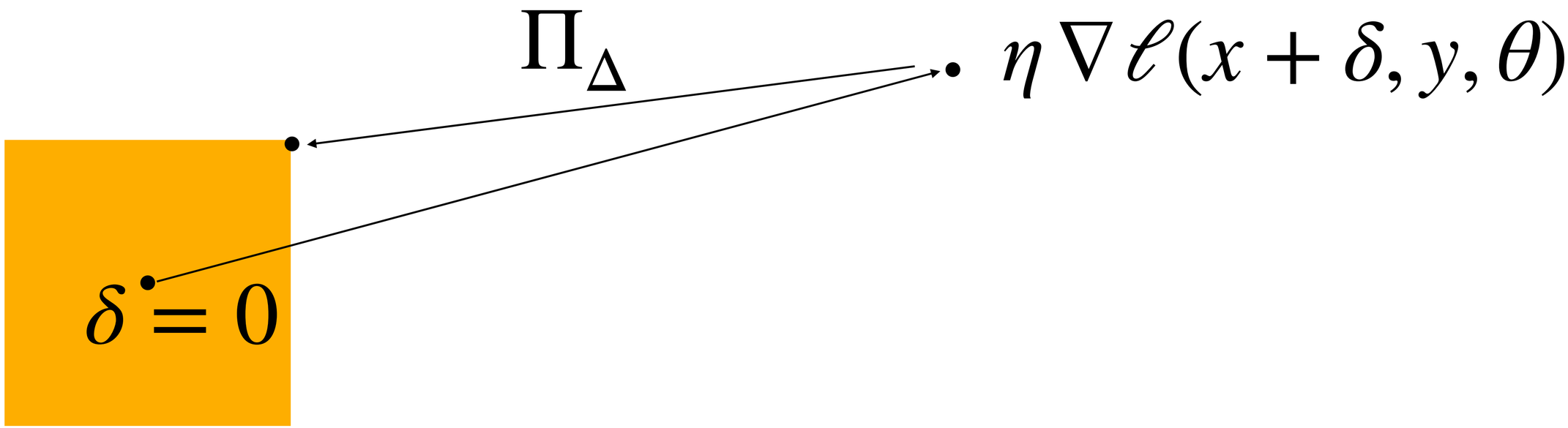
not match perception

Sen et al 2019 https://arxiv.org/pdf/1906.02439.pdf

Nonconvexity

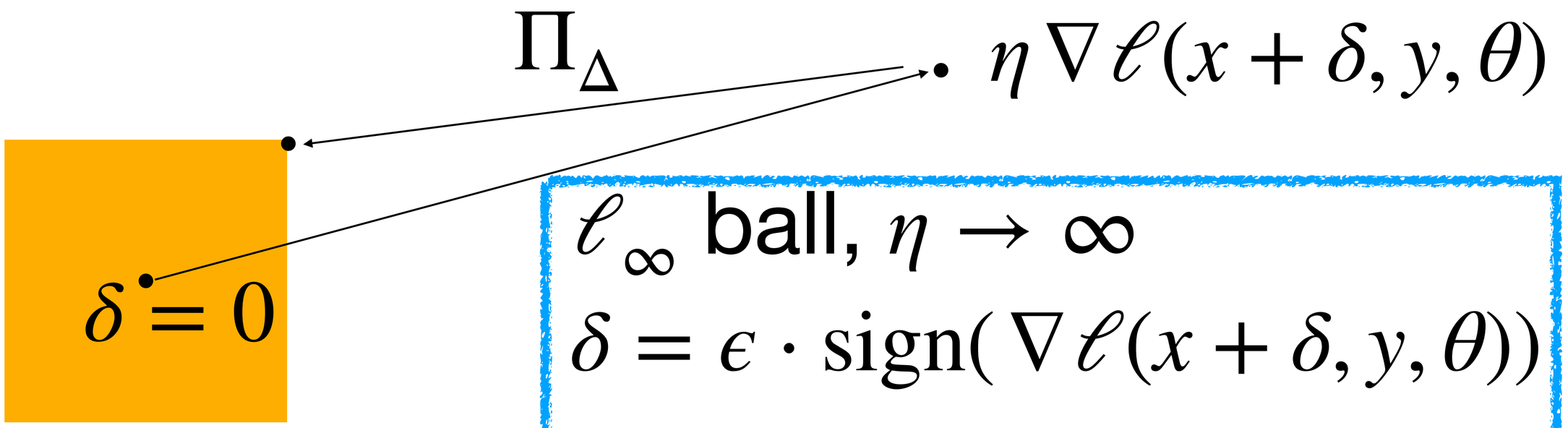
projected gradient descent $\delta \leftarrow \Pi_\Delta[\delta + \eta \nabla \ell(x + \delta, y, \theta)]$

w.r.t. δ



Nonconvexity

projected gradient descent $\delta \leftarrow \Pi_\Delta[\delta + \eta \nabla \ell(x + \delta, y, \theta)]$



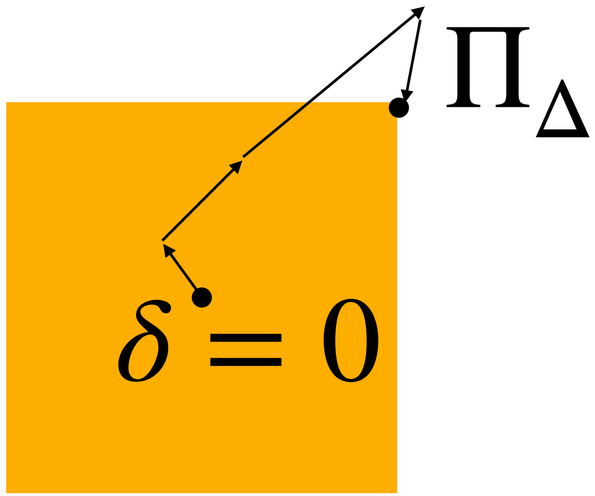
ℓ_∞ ball, $\eta \to \infty$

$\delta = \epsilon \cdot \text{sign}(\nabla \ell(x + \delta, y, \theta))$

FGSM Goodfellow et al 2014 https://arxiv.org/abs/1412.6572

Nonconvexity

projected gradient descent $\delta \leftarrow \Pi_\Delta[\delta + \eta \nabla \ell(x + \delta, y, \theta)]$

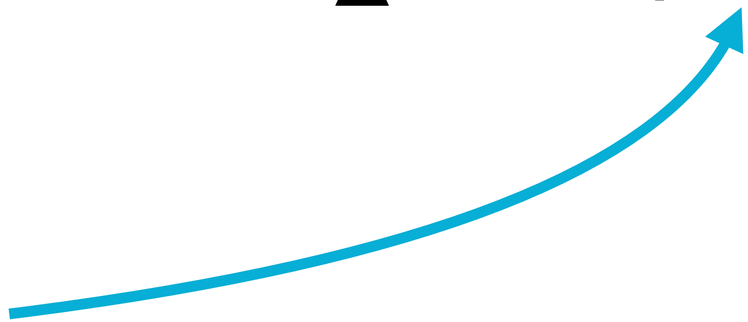


small η

PGD Madry et al 2019 https://arxiv.org/pdf/1706.06083.pdf

Nonconvexity

projected gradient descent $\delta \leftarrow \Pi_{\Delta}[\delta + \eta \nabla \ell(x + \delta, y, \theta)]$



White-box attack: knows

Black-box attack: derivative-free optimization

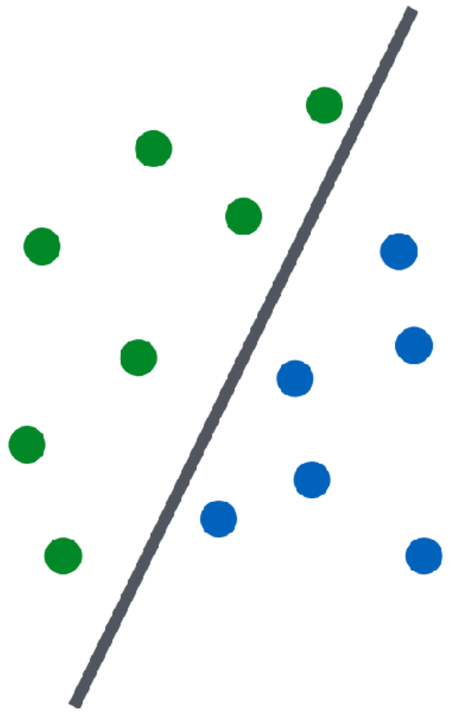
With convex relaxation one may certify there is no adv. examples

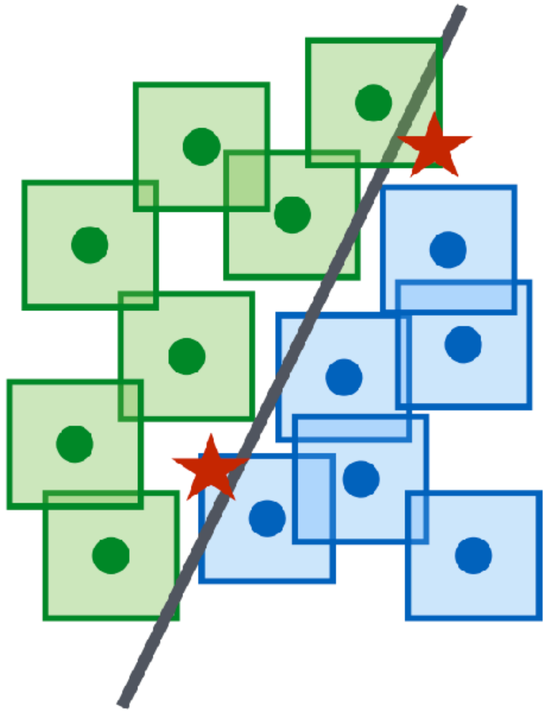
Wong and Kolter, 2018 https://arxiv.org/pdf/1711.00851.pdf

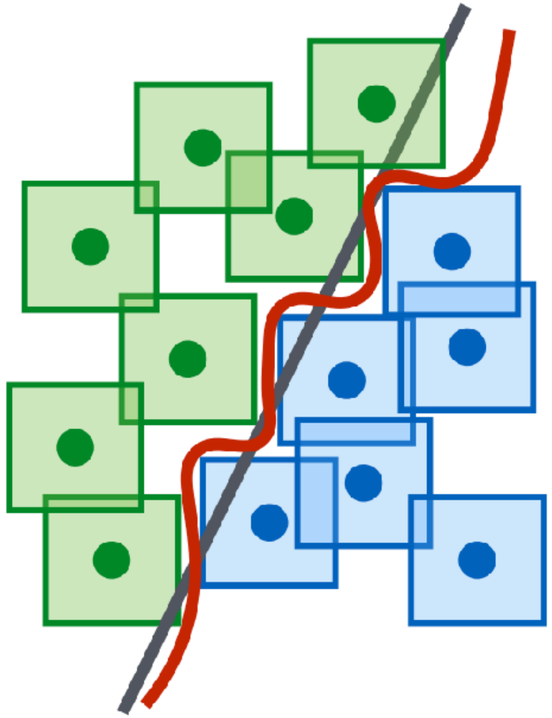
(One) Defense against Test-time Attack

Adversarial Training

$$\min_{\theta} \mathbb{E}_D \max_{\delta \in \Delta} \ell(x + \delta, y, \theta)$$







Madry et al 2019 https://arxiv.org/pdf/1706.06083.pdf

Training-Set Poisoning

Bi-level optimization

$$\min_{D^\dagger,\theta} \|\theta - \theta^\dagger\|$$

$$d(D^\dagger, D) \leq \epsilon$$

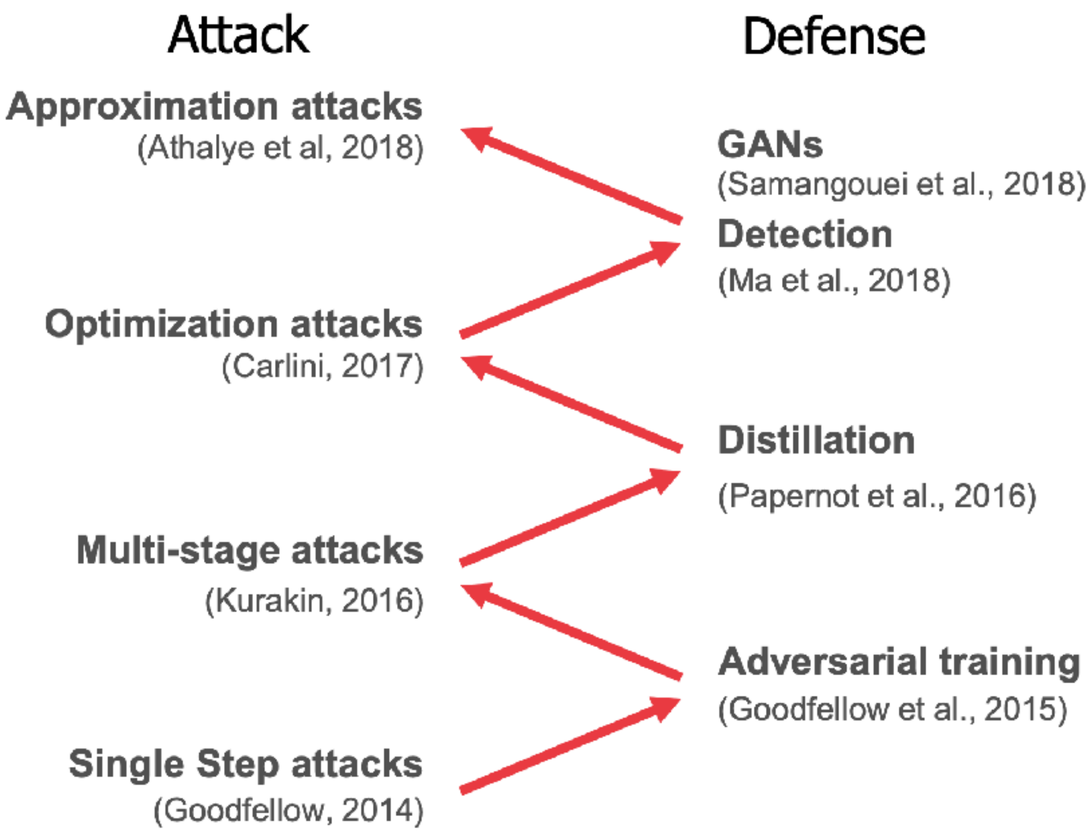
$$\theta = \arg\min_{\nu} \mathbb{E}_{D^\dagger} \ell(x, y, \nu)$$

machine learning

poisoned data

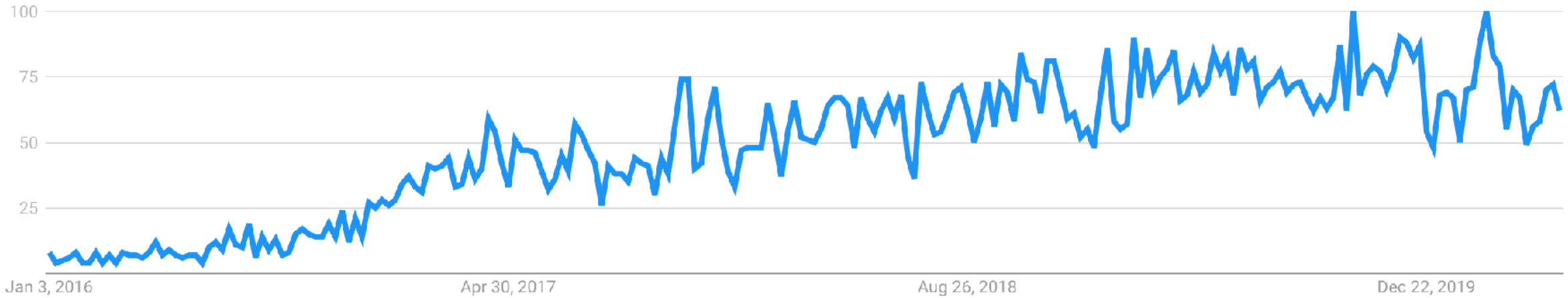
Mei, Z 2015 http://pages.cs.wisc.edu/~jerryzhu/pub/Mei2015Machine.pdf

The Cat and Mouse Game



Siegelmann 2019

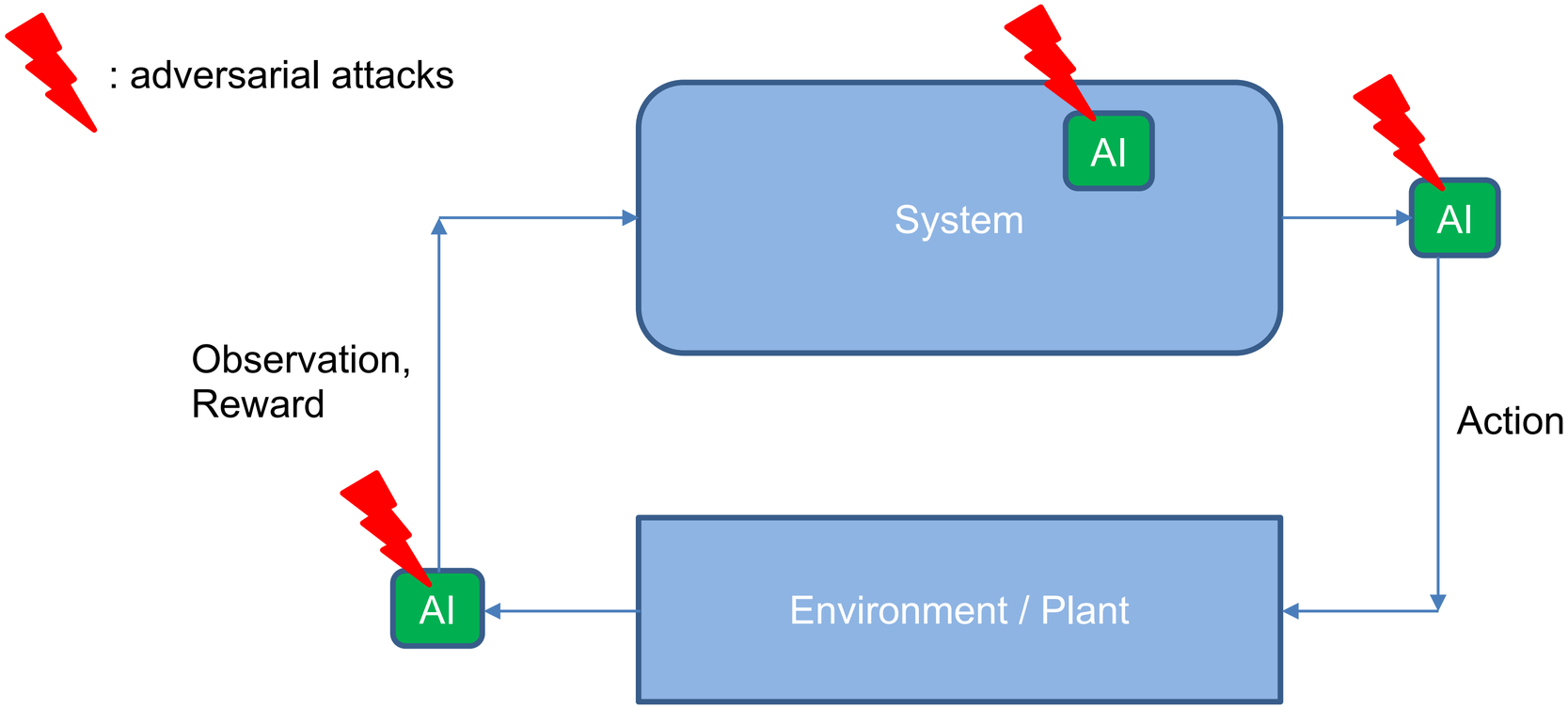
What Next?



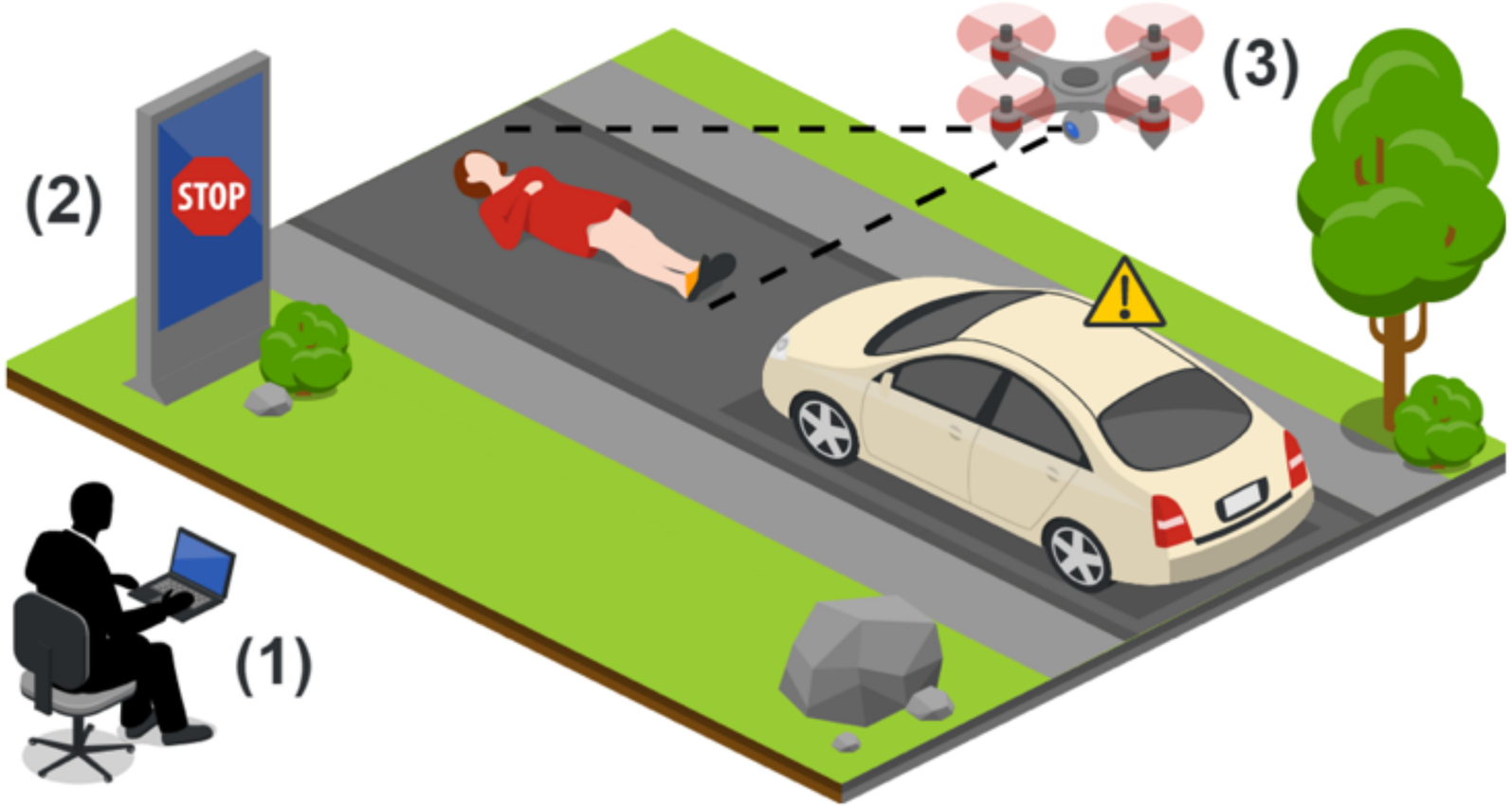
“Adversarial Machine Learning”

Google Trends 5/18/2020

AdvML “at Large”



Autonomous Vehicle



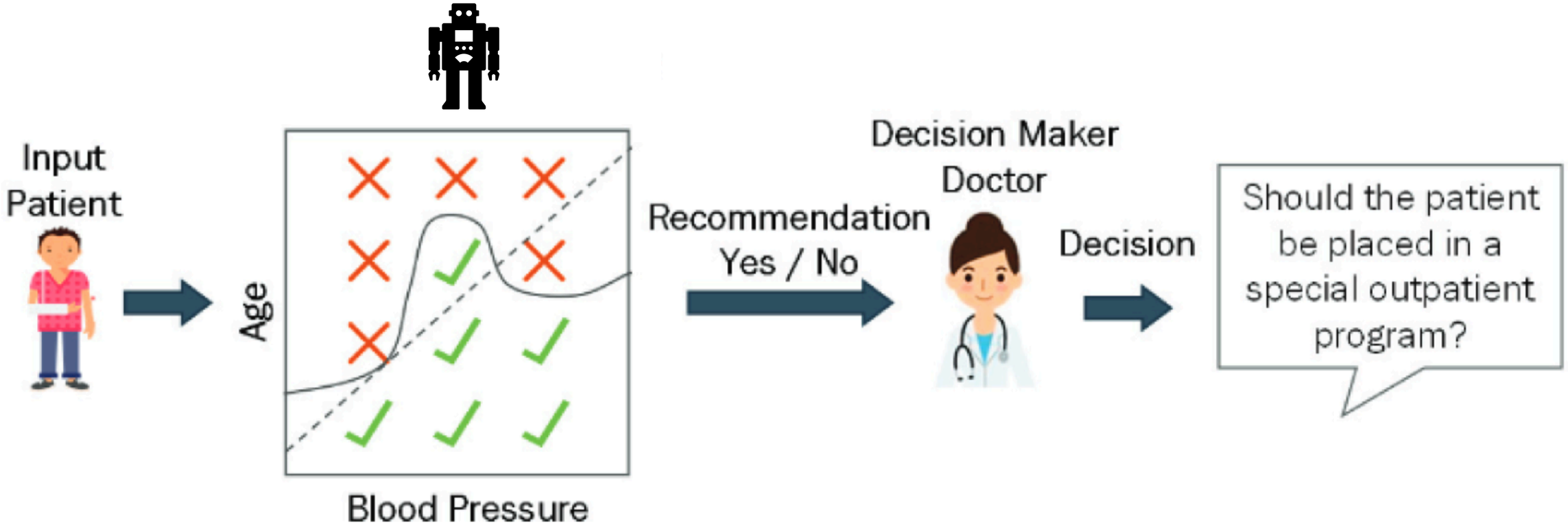
https://eprint.iacr.org/2020/085.pdf

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

68

By James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT

Human-Bot Team Trust



Bansal et al 2019 https://wvvw.aaai.org/ojs/index.php/HCOMP/article/view/5285/5137

AdvML at Large are Dynamic Systems

Attack = Optimal Control

state

$$\min_{a_{0:T-1}} \sum_{t=0}^{T-1} \|s_{t+1} - s_{t+1}^{\dagger}\| + \|a_t\|$$

action

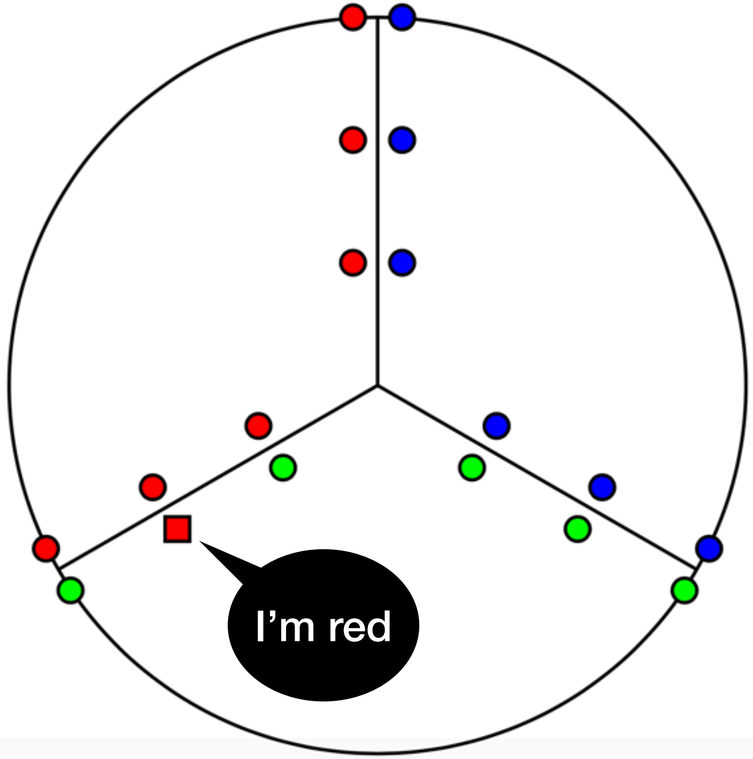
$$s_{t+1} = f(s_t, a_t)$$

state transition

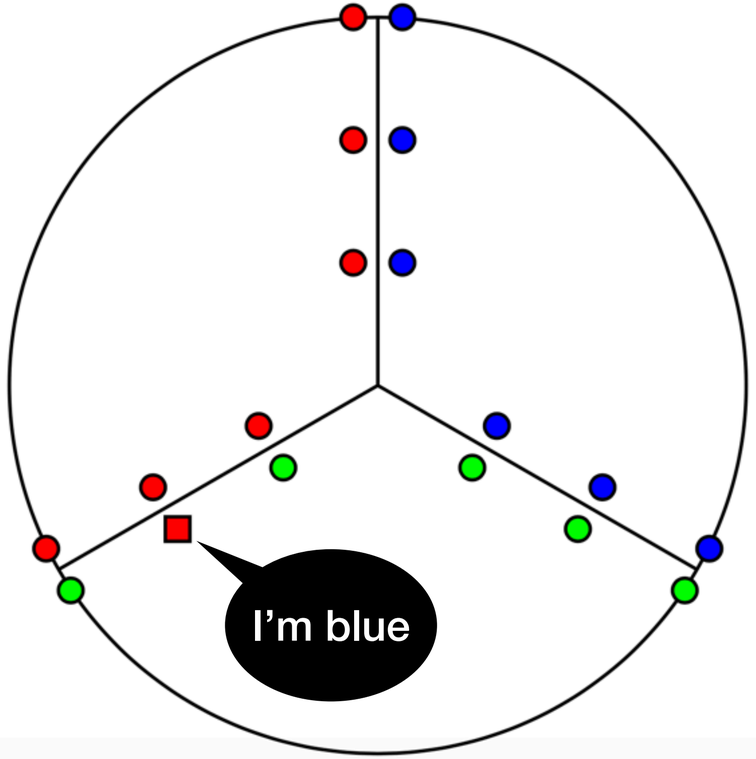
Zhang et al 2020 https://arxiv.org/abs/1903.01666

… and Rational Agents

Game theory Multi-class logistic regression is incentive incompatible



$\theta(\square) = \text{green}$



$\theta^{\dagger}(\square) = \text{red}$

Wu, Tzamos, Z. In preparation

Selected References

- Siegelmann. DARPA GARD Program. 2019
- Kolter and Madry. NeurIPS tutorial on Adversarial Robustness. 2018
- Papernot et al. Towards the Science of Security and Privacy in Machine Learning. 2016
- Zhu. An Optimal Control View of Adversarial Machine Learning. 2018