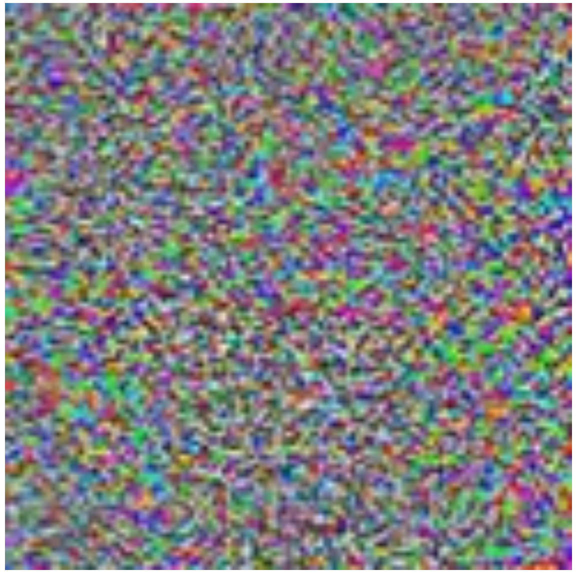
Detecting Adversarial Examples Is (Nearly) As Hard As Classifying Them

Florian Tramèr

Stanford University, Google, ETHZ

ML suffers from *adversarial examples.*

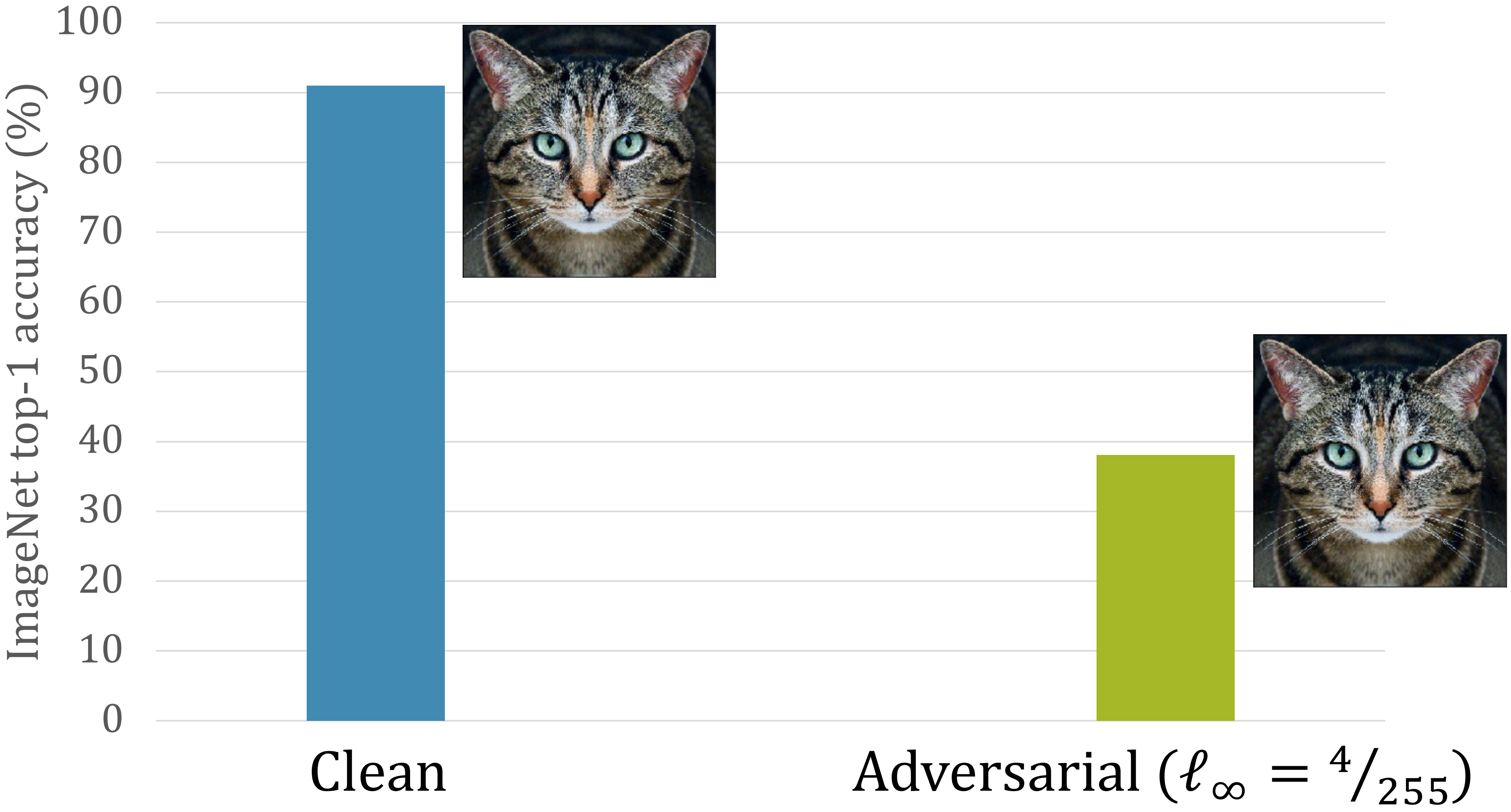
+



=

90% Tabby Cat **Adversarial noise** **100% Guacamole**

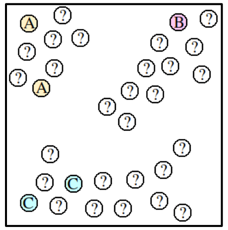
Robust classification is hard!



Can we solve an *easier* problem?

Computationally robust classification

Randomized robust classification



Robust *transductive classification*

Robust *detection*

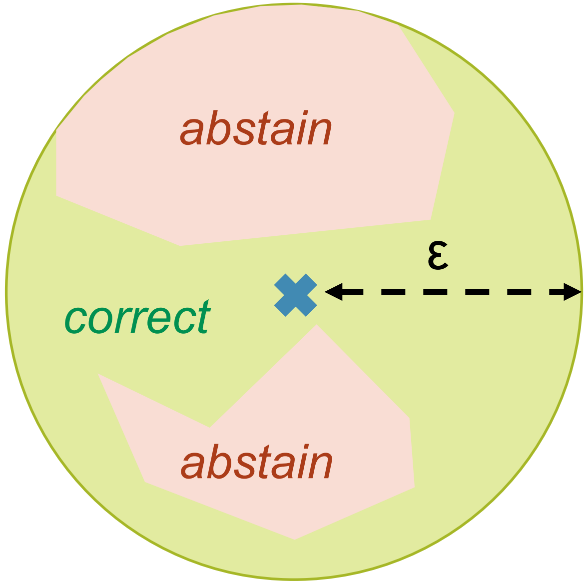
Can we solve an *easier* problem?

Computationally robust classification

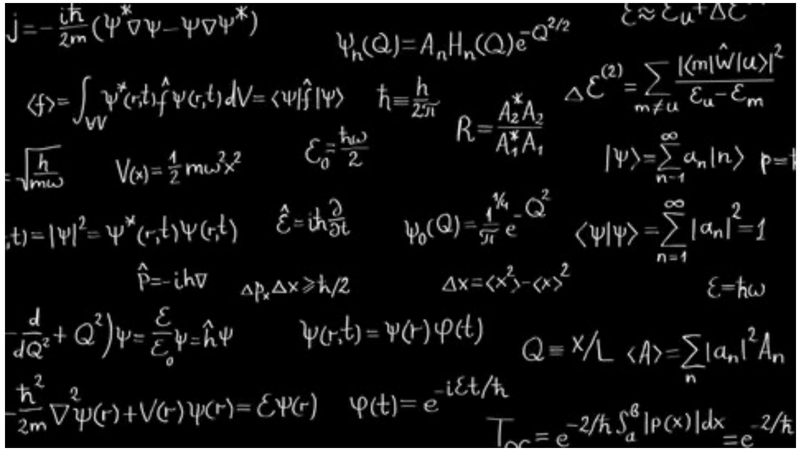
Randomized robust classification

Robust *transductive classification*

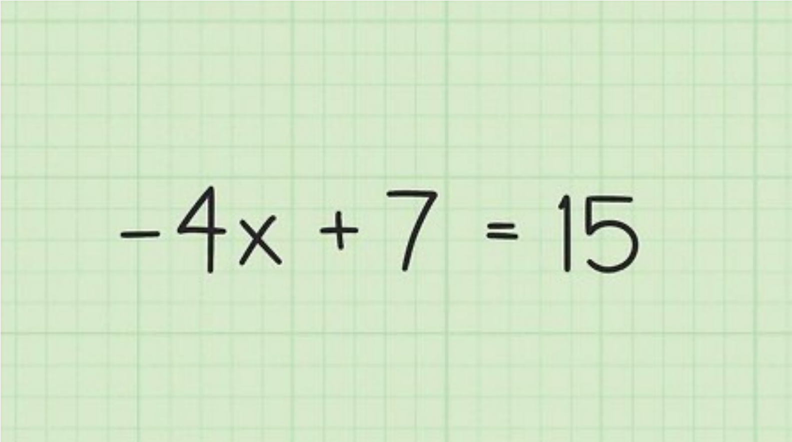
Robust *detection*



Are these relaxed problems truly easier?



Robust classification



Robust detection

ON EVALUATING ADVERSARIAL ROBUSTNESS

Nicholas Carlini[1], Anish At
Dimitris Tsipras[2], Ian Good

MadryLab / mnist_challenge

ROBUSTBENCH
A standardized benchmark for adversarial robustness

RobustML

NIPS 2017: Defense Against Adversarial Attack
Create an image classifier that is robust to adversarial attacks
Google Brain · 107 teams · 5 years ago

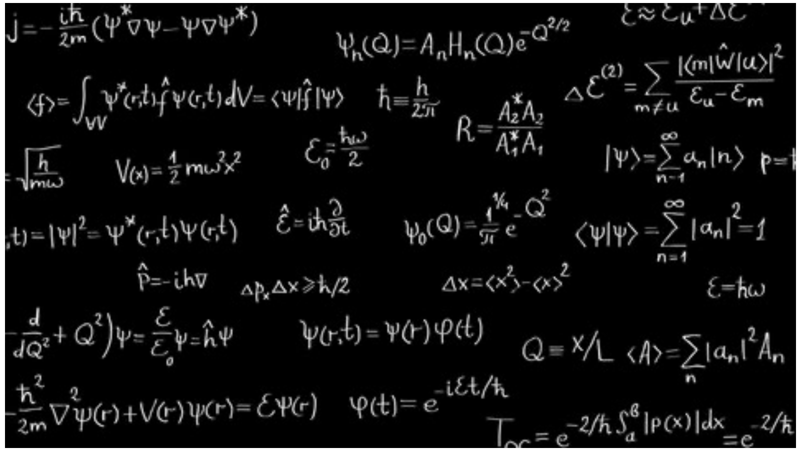
Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Nicholas Carlini David Wagner

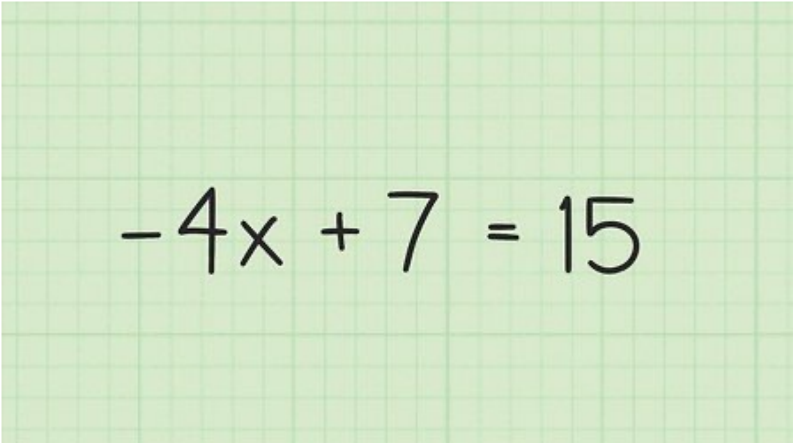
On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr* Stanford University
Nicholas Carlini* Google
Wieland Brendel* University of Tübingen
Aleksander Mądry MIT

Are these relaxed problems truly easier?



Robust classification



Robust detection

- ➢ **If YES:** promising direction for useful robustness!
- ➢ **If NO:** we shouldn't expect a breakthrough...

Our result.

Detecting adversarial examples is as hard as classifying them!

What’s a hardness reduction?

“Famously hard” problems

AGI
(lol)

P vs NP

Riemann Hypothesis

reduction

Problem X

If we find a solution to Problem X, we also solve a super hard problem

What's a hardness reduction?

"Famously hard" problems

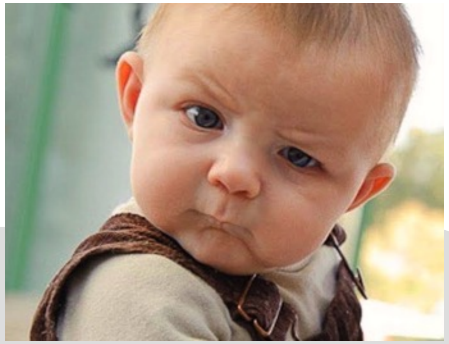
AGI
(lol)

P vs NP

Riemann Hypothesis

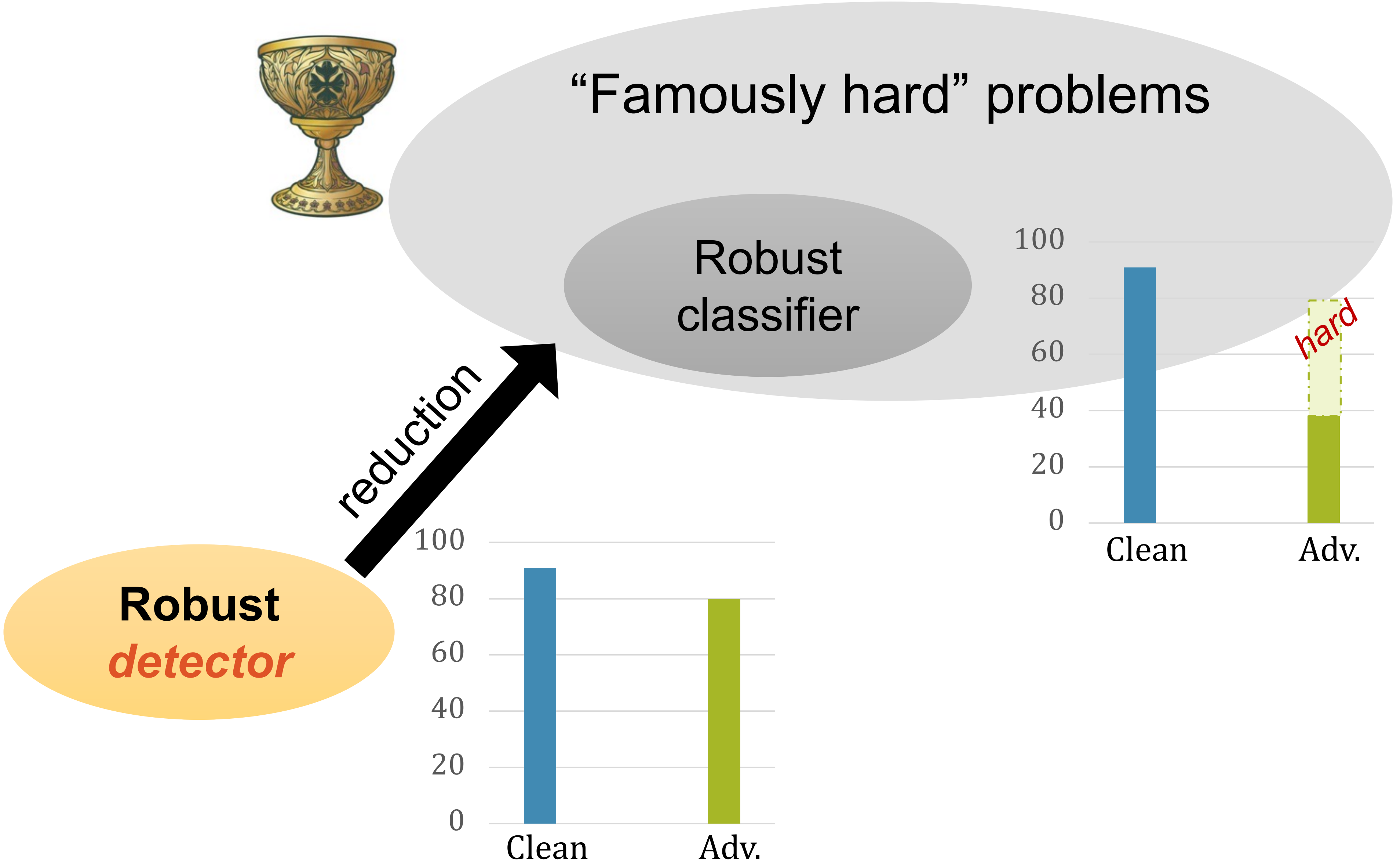
reduction

Problem X



<u>Corollary:</u>
if someone claims to solve Problem X, you might be a bit skeptical...

Hardness reductions for robustness.



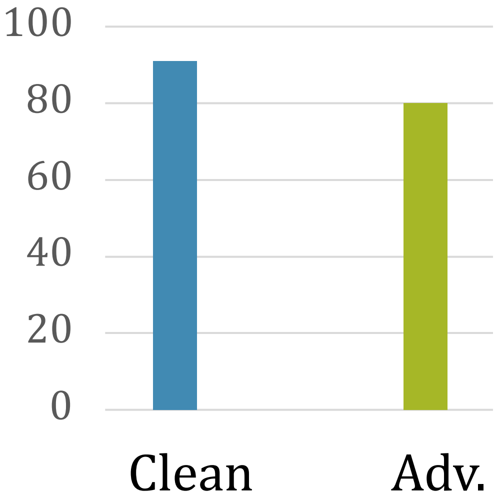
Detecting adversarial examples is as hard as classifying them!

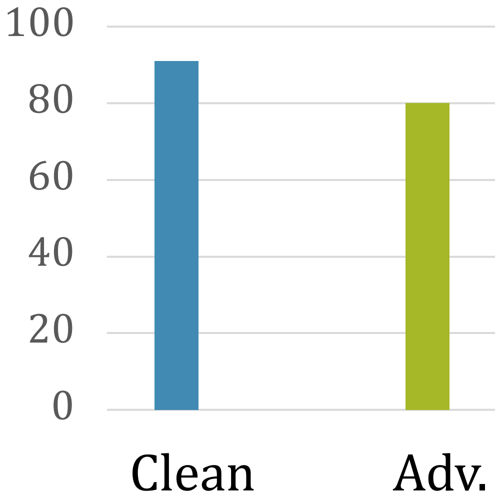
Robust *detector*

reduction

detector

Robust *classifier*





Detecting adversarial examples is (nearly) as hard as classifying them!

Robust *detector* → reduction → *detector* Robust *classifier*

- efficient
- robust at distance ε

- *inefficient* *(at inference)*
- robust at distance $\varepsilon/2$

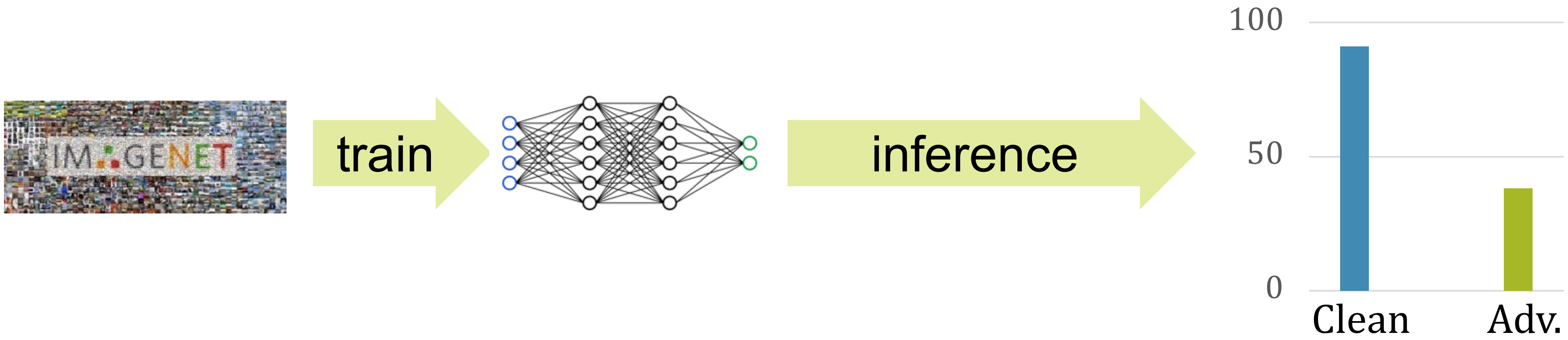
Main technical tool: ***Minimum Distance Decoding***

Interpretation #1: *information theoretically* robust detection = robust classification

- Same *sample complexity* [Schmidt et al., 2018]
- Same *accuracy-robustness tradeoffs* [Tsipras et al., 2019, Zhang et al., 2019]
- Same *multi-robustness tradeoffs* [T & Boneh, 2019, Maini et al., 2020]
- Same connection with *error on noise* [Ford et al., 2020]
- ...

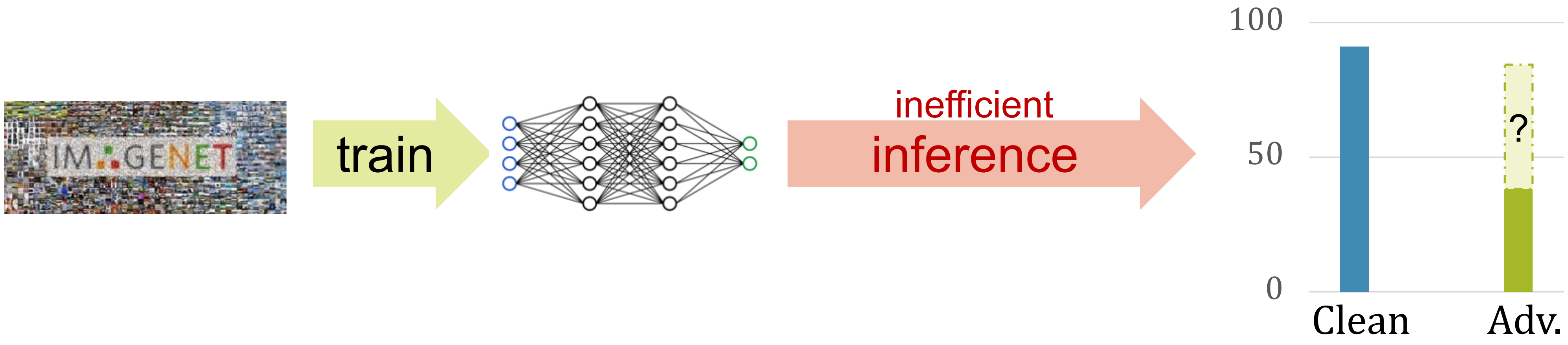
Interpretation #2: robust detectors imply a *breakthrough in robust classification.*

World 1:



Interpretation #2: robust detectors imply a *breakthrough in robust classification.*

World 2:



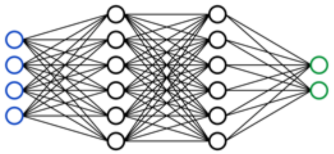
Can we build much more robust classifiers in **World 2**?

(we don't know...)

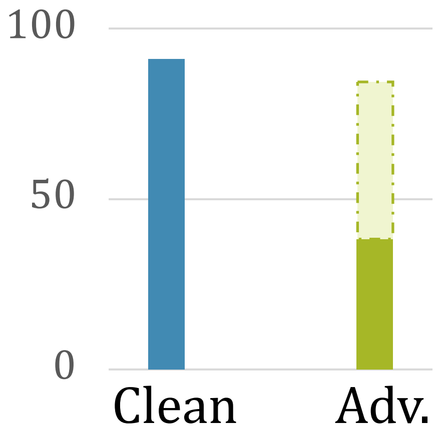
Interpretation #2: robust detectors imply a *breakthrough in robust classification.*

World 2:

train



inefficient
inference

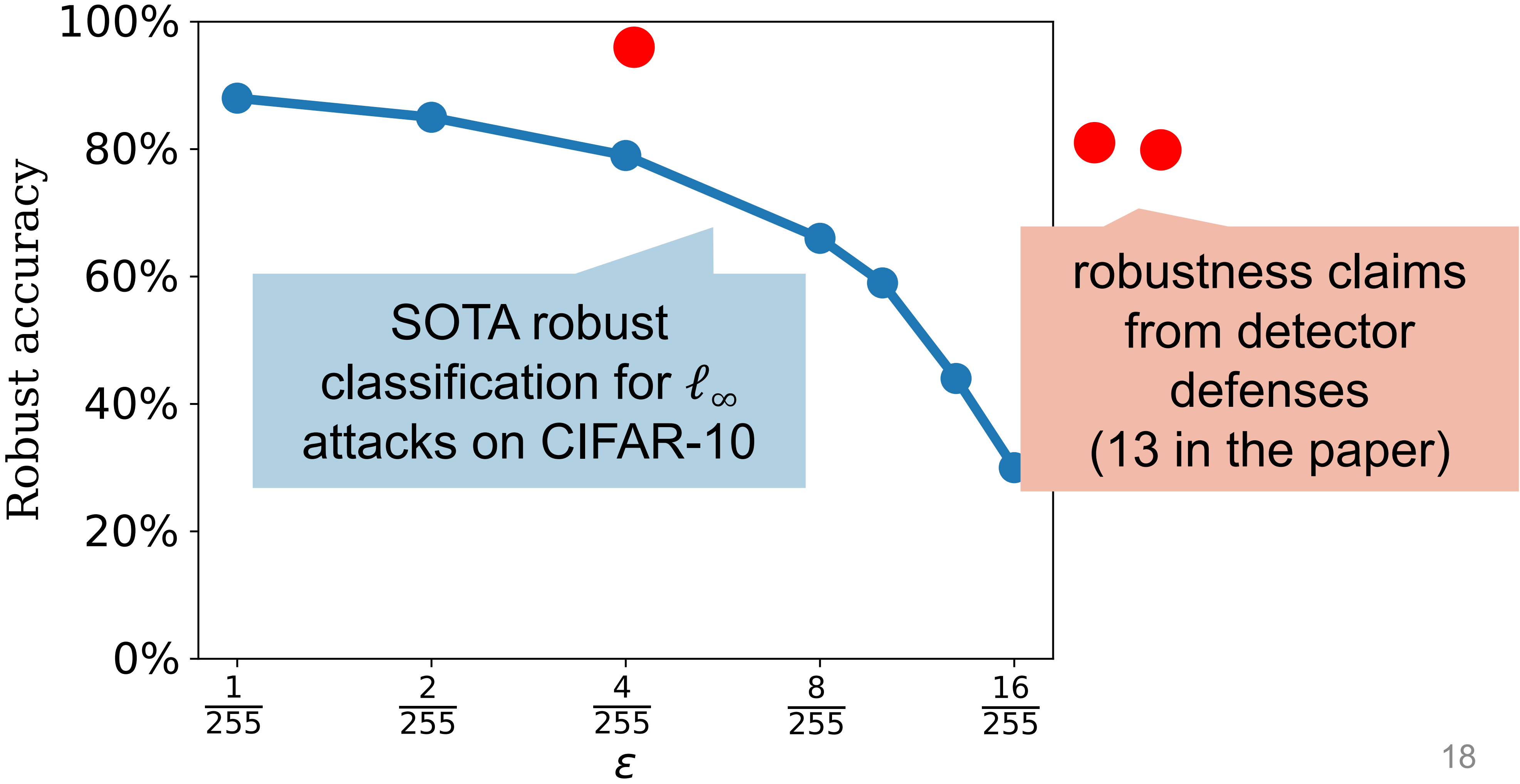


Can we build much more robust classifiers in **World 2**?

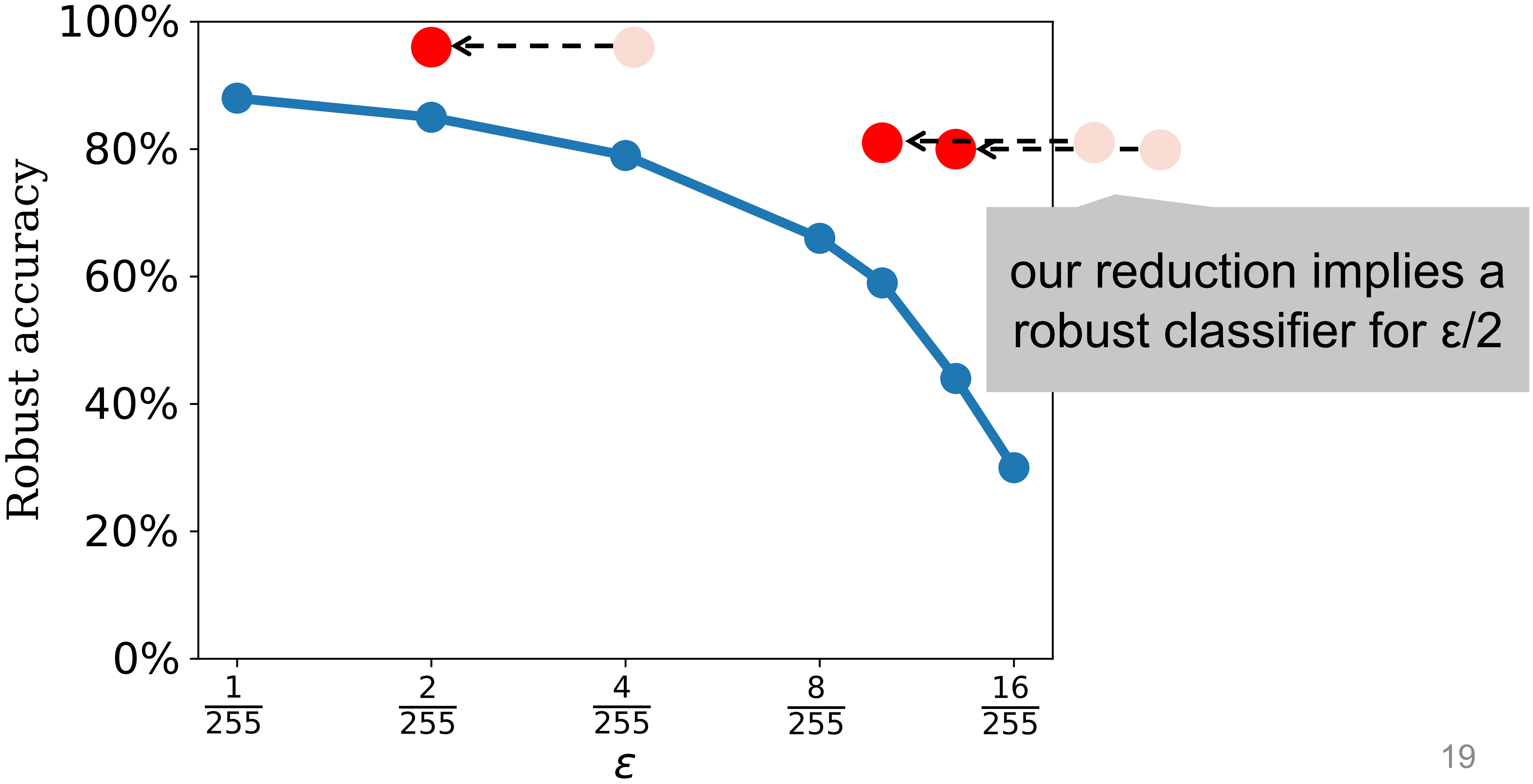
(we don't know...)

But any sufficiently robust detector implies a positive answer!

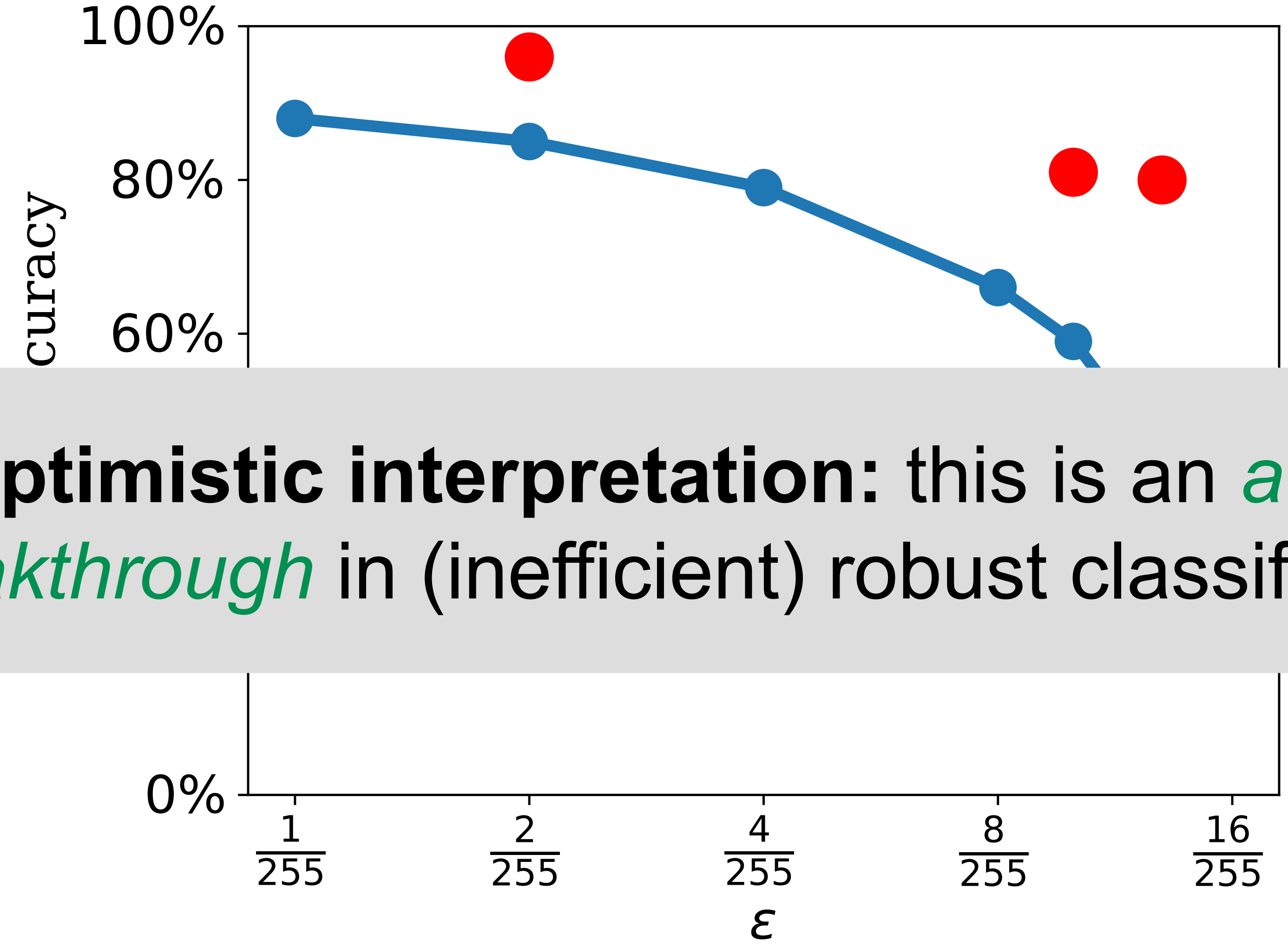
Many detectors *implicitly* claim such a breakthrough!

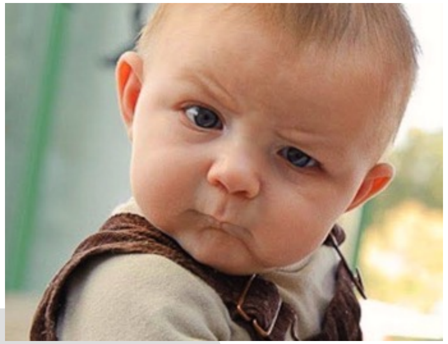


Many detectors *implicitly* claim such a breakthrough!



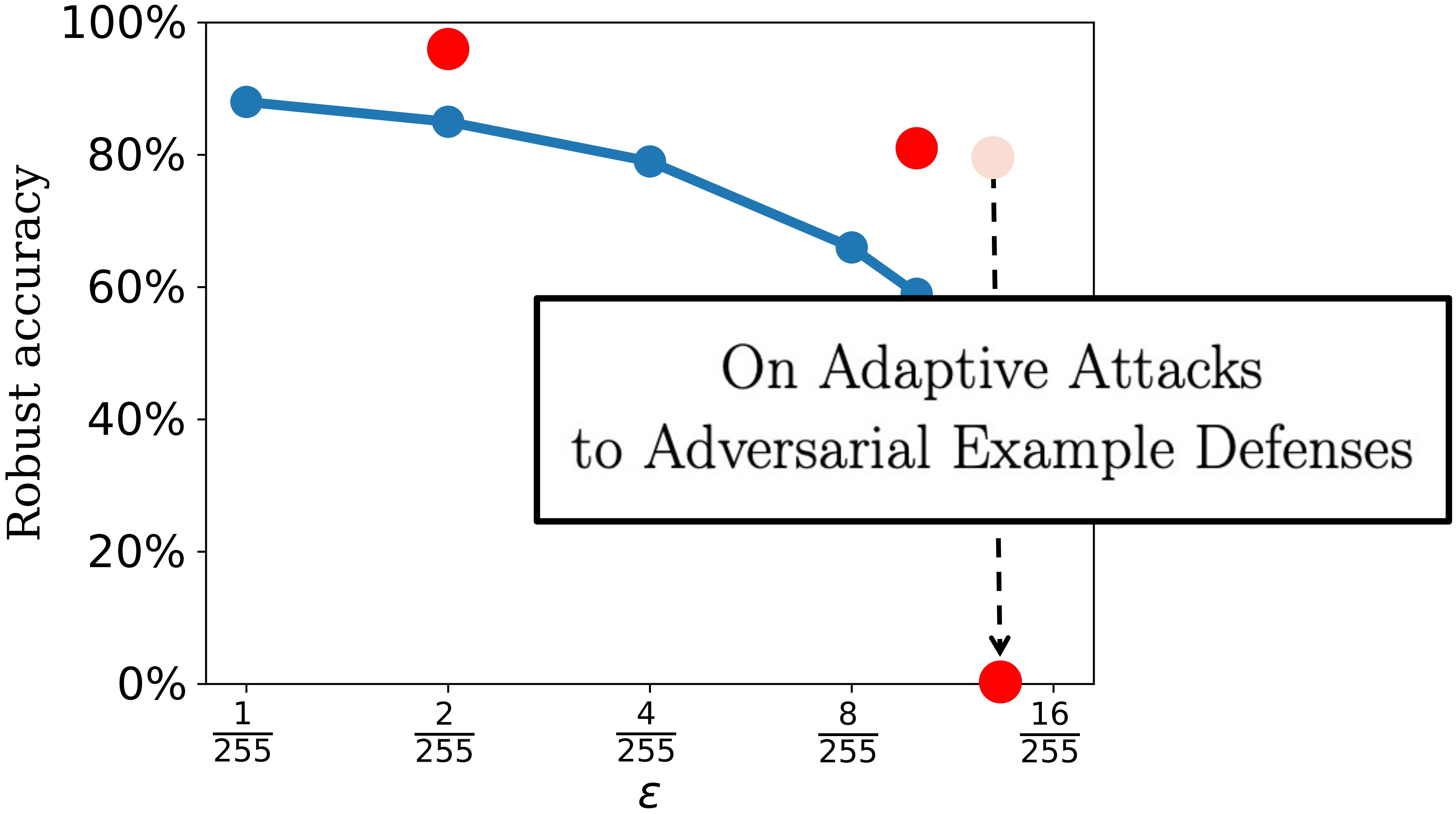
Many detectors *implicitly* claim such a breakthrough!



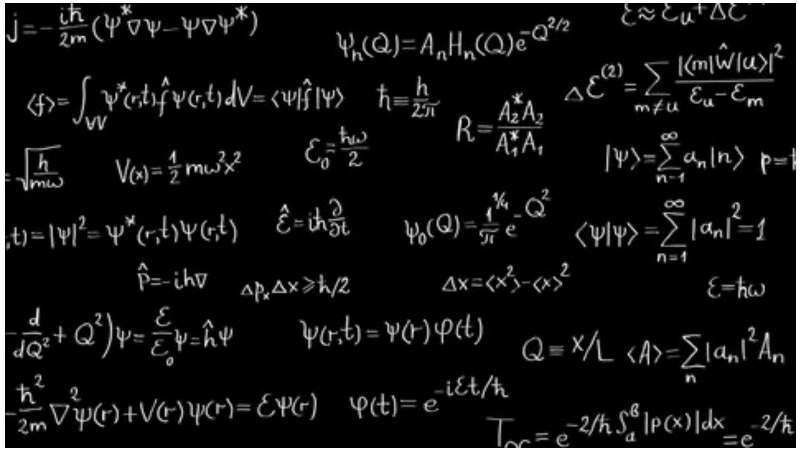


Optimistic interpretation: this is an *actual breakthrough* in (inefficient) robust classification!

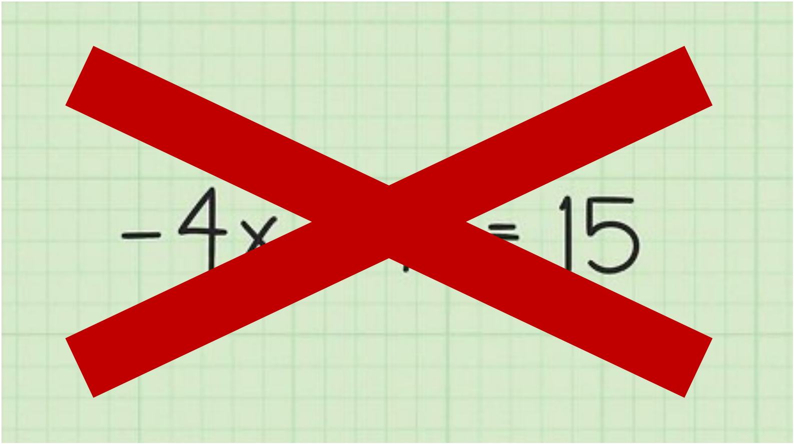
Pessimistic (*realistic?*) interpretation:
These detectors are *not robust!*



Conclusion.

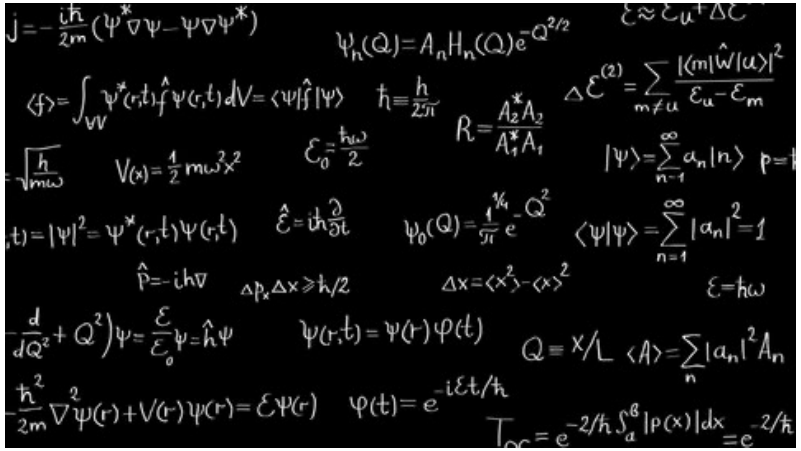


Robust classification

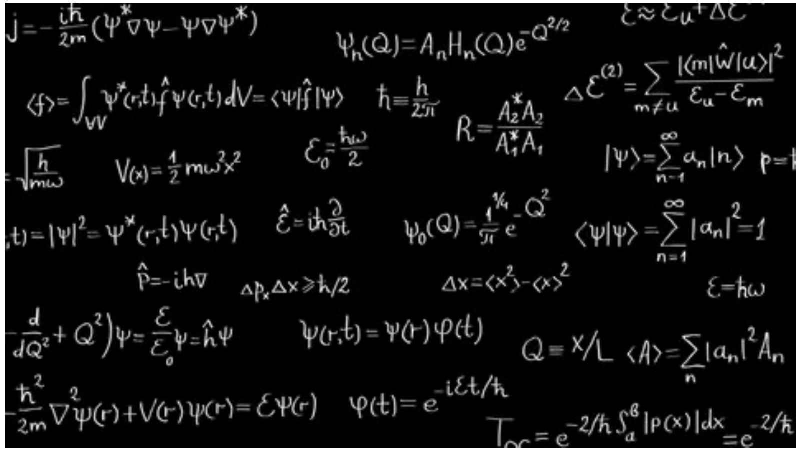


Robust detection

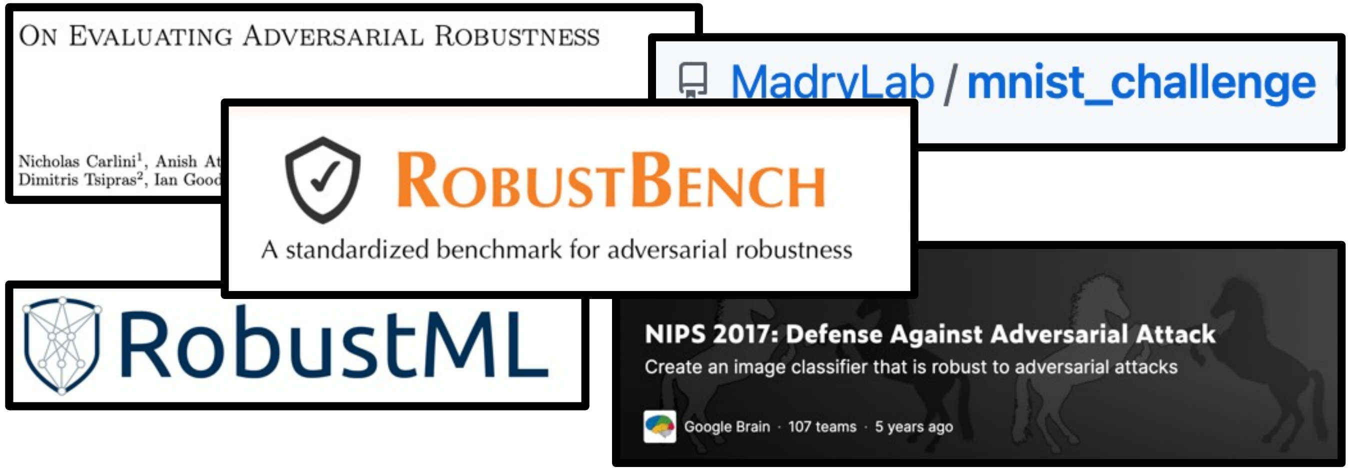
Conclusion.



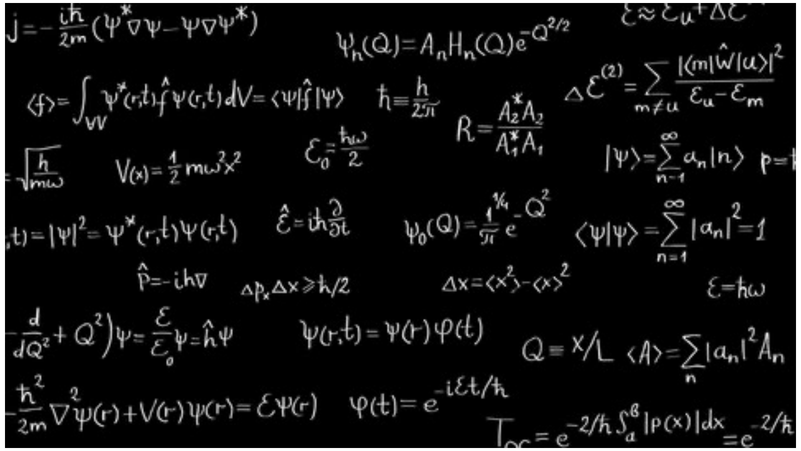
Robust classification



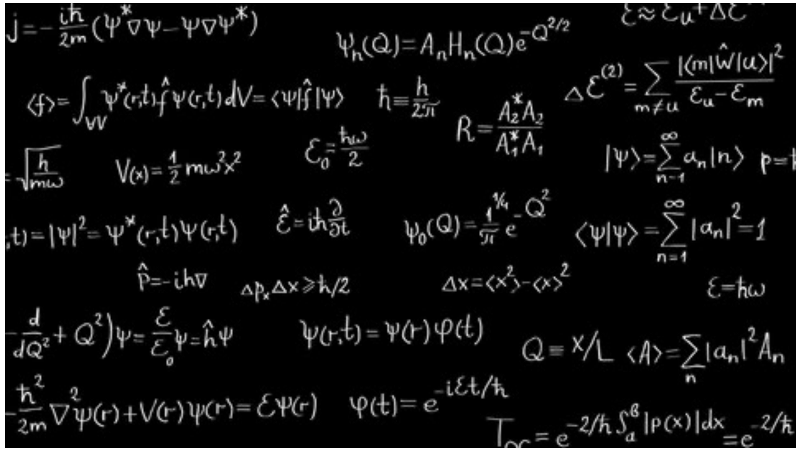
Robust detection



Conclusion.

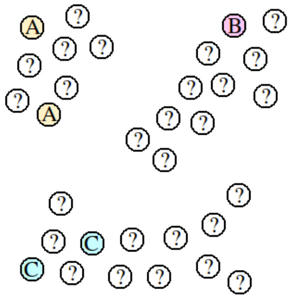


Robust classification



Robust detection

➢Reductions/separations for other “easier” approaches to robustness?



https://arxiv.org/abs/2107.11630

https://floriantramer.com