

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

Florian Tramèr

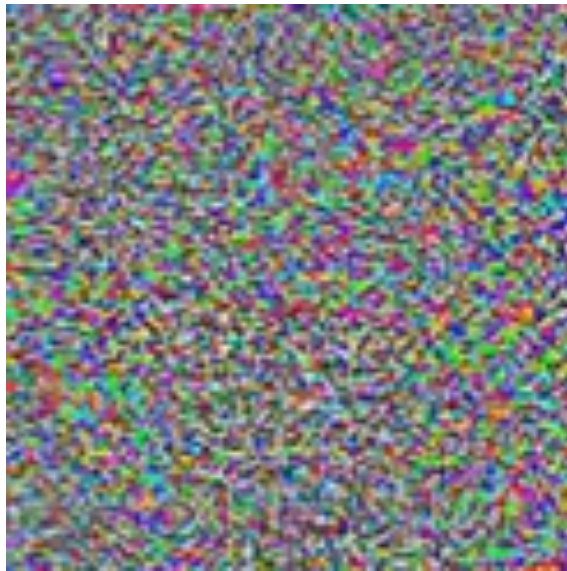
Stanford University, Google, ETHZ

ML suffers from *adversarial examples*.



90% Tabby Cat

+



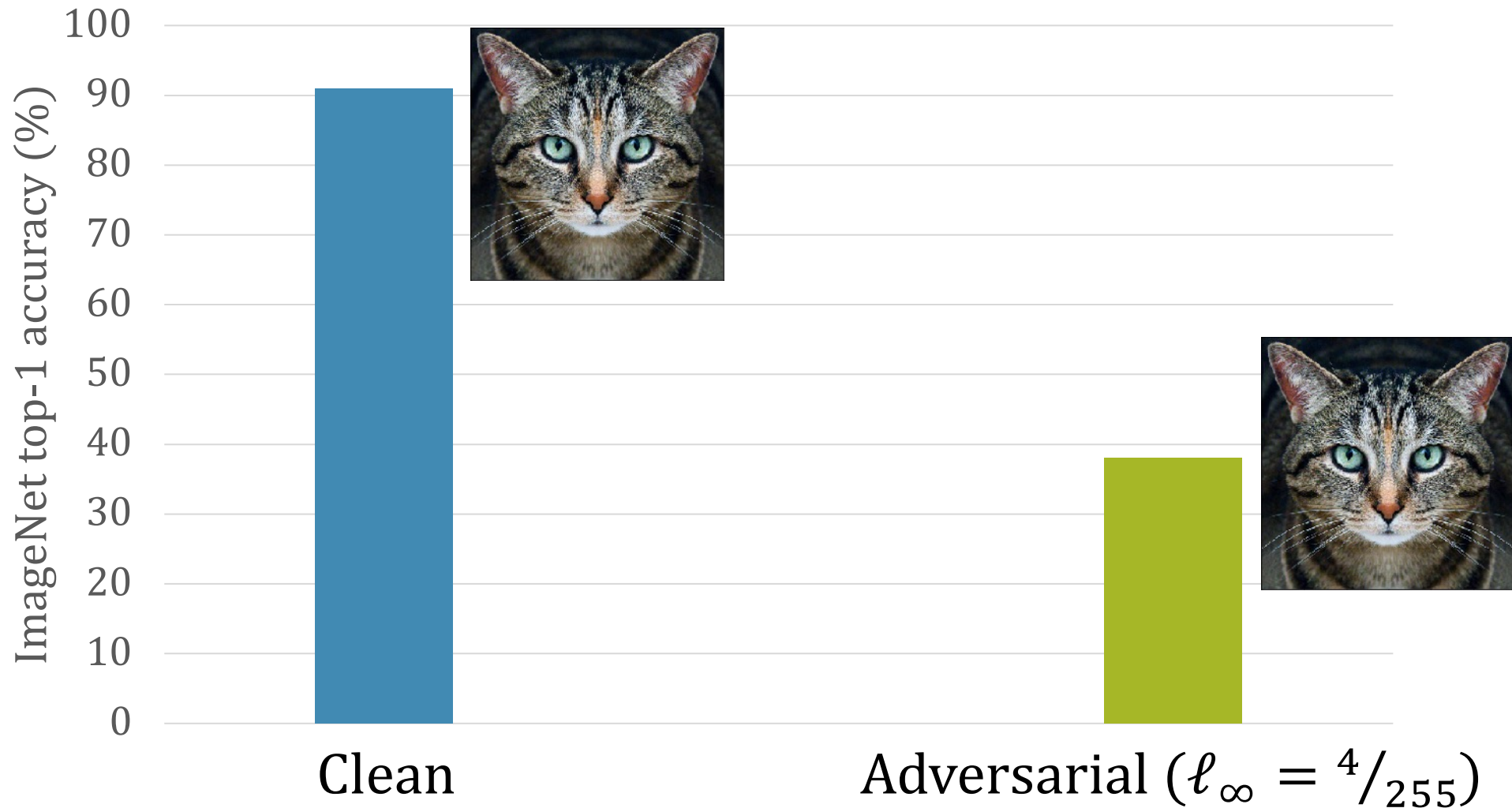
Adversarial noise

=



100% Guacamole

Robust classification is **hard!**



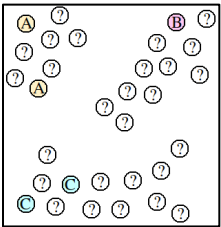
Can we solve an *easier* problem?



Computationally robust classification



Randomized robust classification



Robust *transductive classification*



Robust *detection*

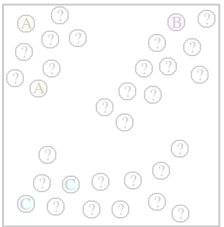
Can we solve an *easier* problem?



Computationally robust classification



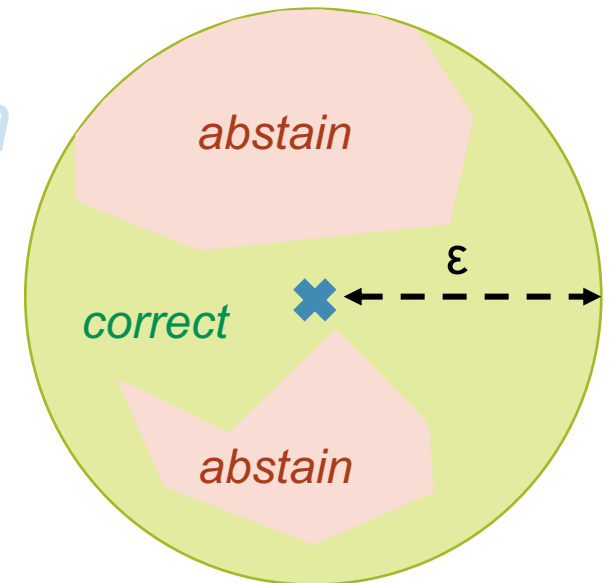
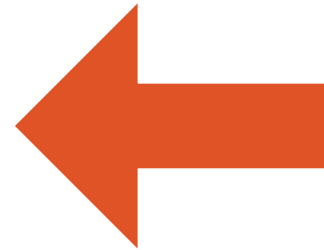
Randomized robust classification



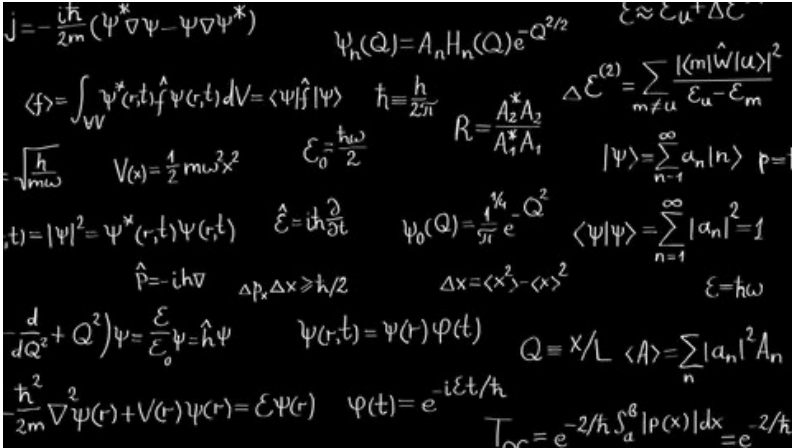
Robust *transductive classification*



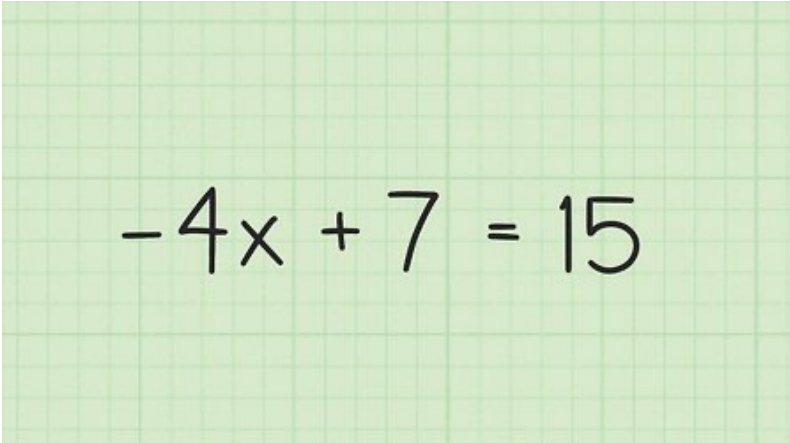
Robust *detection*



Are these relaxed problems truly easier?



Robust classification





Robust detection

ON EVALUATING ADVERSARIAL ROBUSTNESS

[MadvLab / mnist_challenge](https://madvlab.github.io/mnist_challenge)

Nicholas Carlini¹, Anish Athalye², Dimitris Tsipras², Ian Goodfellow²

 **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* Nicholas Carlini* Wieland Brendel*
Stanford University Google University of Tübingen

Aleksander Mądry
MIT

Are these relaxed problems **truly easier**?

Handwritten mathematical notes on a blackboard background, including:

- $j = -\frac{i\hbar}{2m}(\psi^* \nabla \psi - \psi \nabla \psi^*)$
- $\psi_n(Q) = A_n H_n(Q) e^{-Q^2/2}$
- $\langle f \rangle = \int_{VV} \psi^*(r,t) \hat{f} \psi(r,t) dV = \langle \psi | \hat{f} | \psi \rangle$
- $\hbar = \frac{h}{2\pi}$
- $R = \frac{A_2^* A_2}{A_1^* A_1}$
- $\Delta \mathcal{E}^{(2)} = \sum_{m \neq u} \frac{|\langle m | \hat{W} | u \rangle|^2}{\mathcal{E}_u - \mathcal{E}_m}$
- $V(x) = \frac{1}{2} m \omega^2 x^2$
- $\mathcal{E}_0 = \frac{\hbar \omega}{2}$
- $|\psi\rangle = \sum_{n=1}^{\infty} a_n |n\rangle$
- $\langle \psi | \psi \rangle = \sum_{n=1}^{\infty} |a_n|^2 = 1$
- $\hat{p} = -i\hbar \nabla$
- $\Delta p_x \Delta x \geq \hbar/2$
- $\Delta x = \langle x^2 \rangle - \langle x \rangle^2$
- $\mathcal{E} = \hbar \omega$
- $\frac{d}{dQ^2} (Q^2) \psi = \frac{\mathcal{E}}{\mathcal{E}_0} \psi = \hat{h} \psi$
- $\psi(r,t) = \psi(r) \varphi(t)$
- $Q = x/L$
- $\langle A \rangle = \sum_n |a_n|^2 A_n$
- $\frac{\hbar^2}{2m} \nabla^2 \psi(r) + V(r) \psi(r) = \mathcal{E} \psi(r)$
- $\varphi(t) = e^{-i\mathcal{E}t/\hbar}$
- $T_{cc} = e^{-2/\hbar \int_a^b |p(x)| dx} = e^{-2/\hbar}$

Robust classification

Handwritten mathematical equation on a green grid background:

$$-4x + 7 = 15$$

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

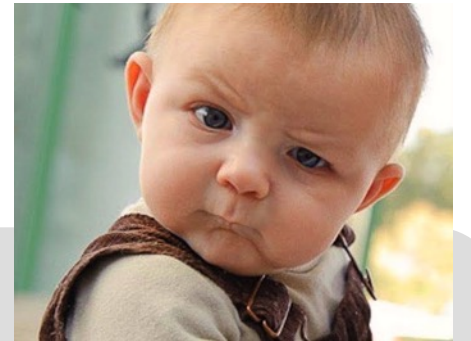
Riemann
Hypothesis

reduction

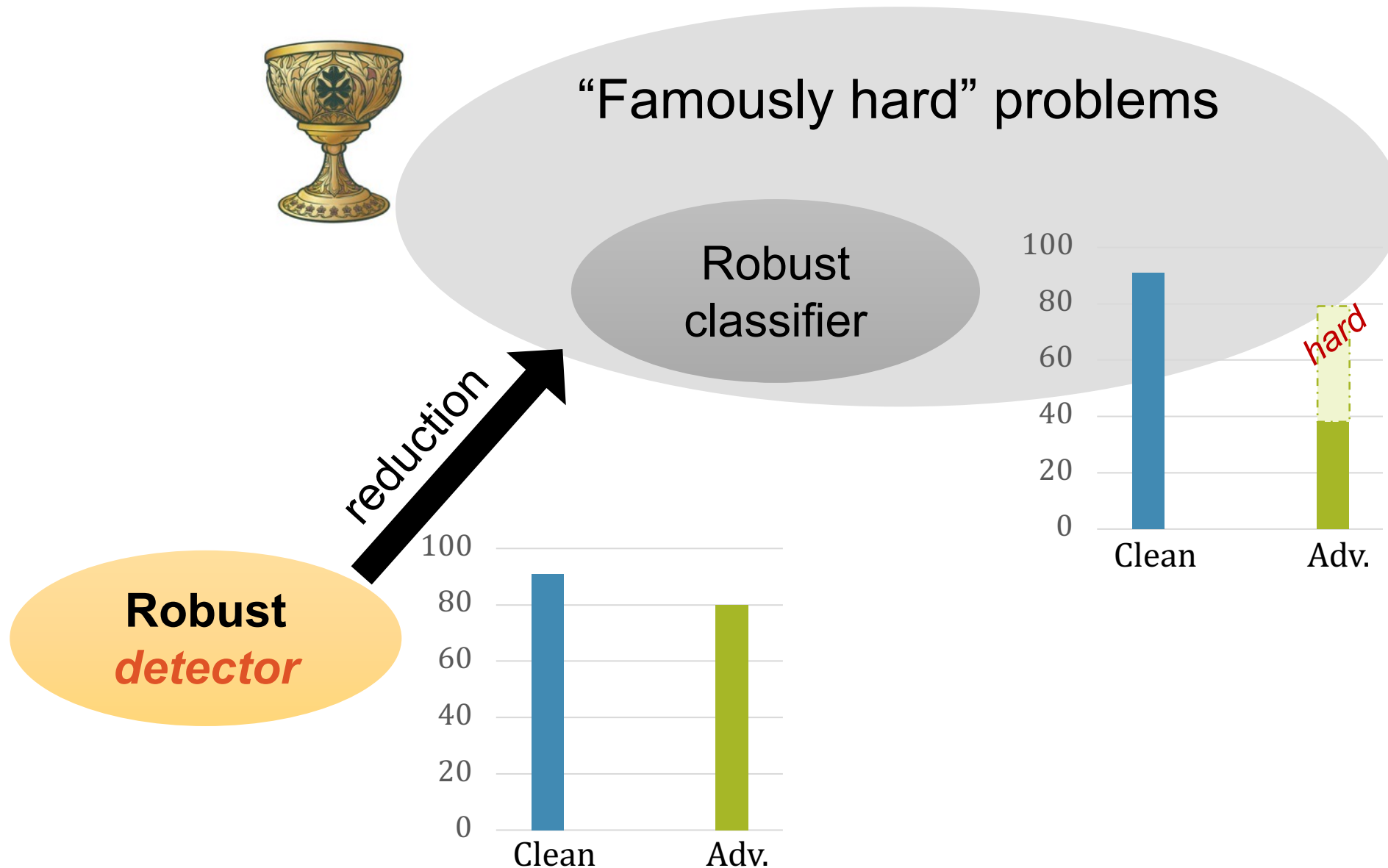
Problem X

Corollary:

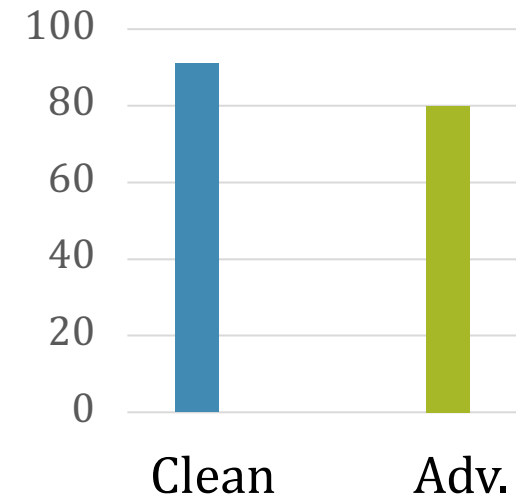
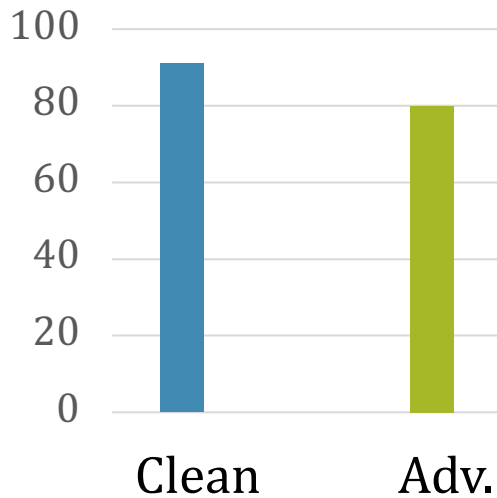
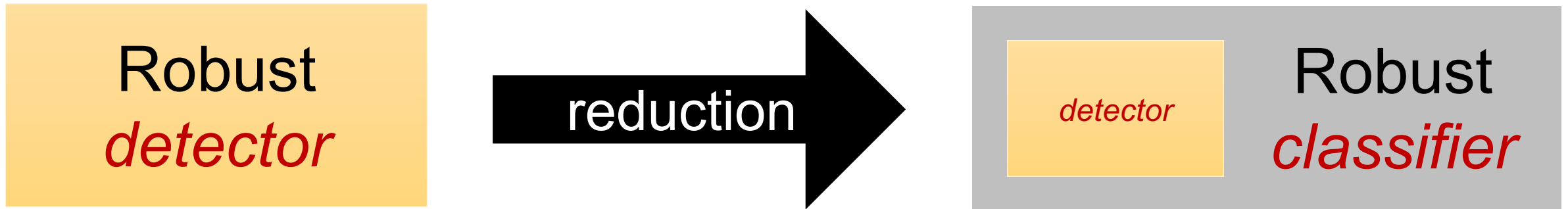
if someone claims to solve Problem X,
you might be a bit skeptical...



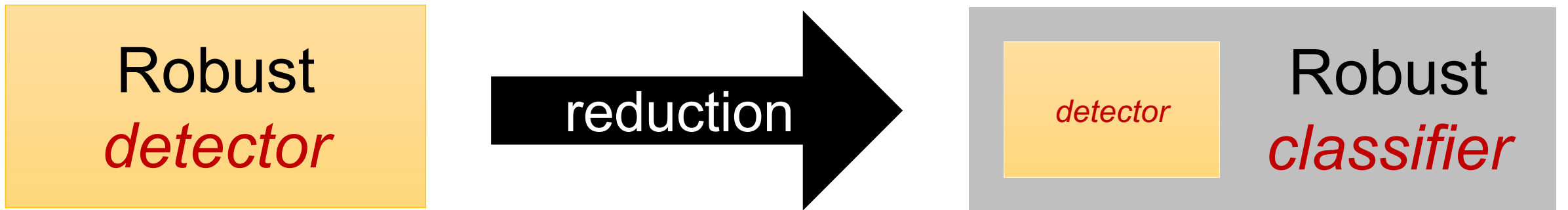
Hardness reductions for robustness.



Detecting adversarial examples is as hard as classifying them!



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



- 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

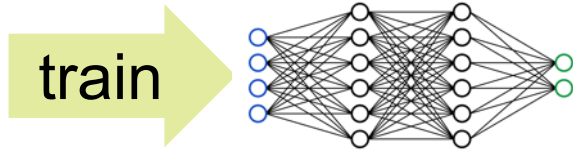
- Same *sample complexity* [Schmidt et al., 2018]
- Same *accuracy-robustness tradeoffs* [Tsipras et al., 2019, Zhang et al., 2019]
- Same *multi-robustness tradeoffs* [T & Boneh, 2019, Maini et al., 2020]
- Same connection *with error on noise* [Ford et al., 2020]
- ...

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

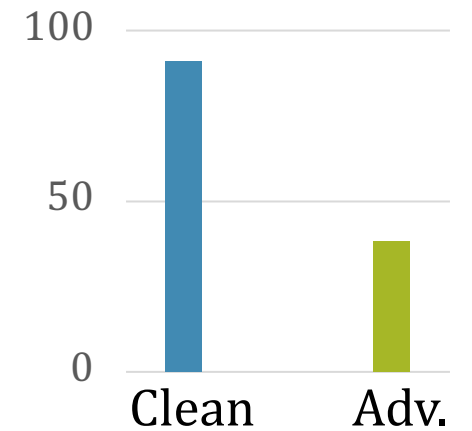
World 1:



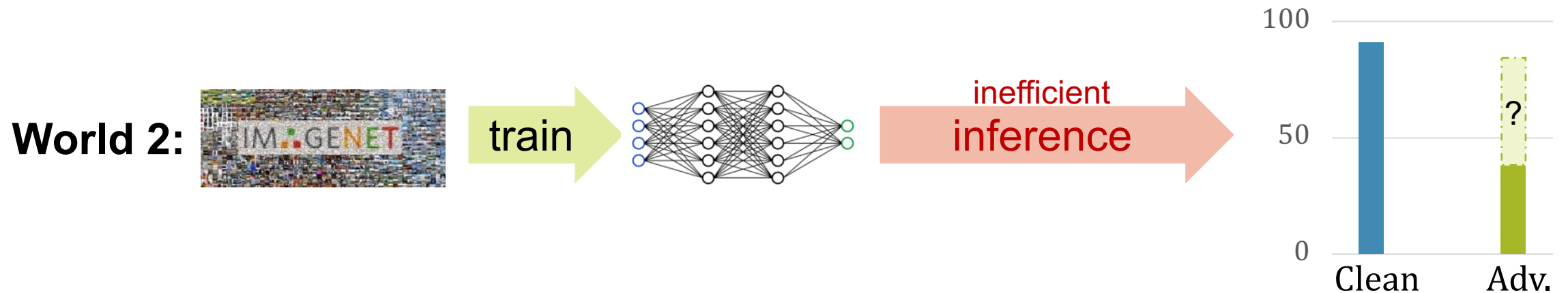
train



inference

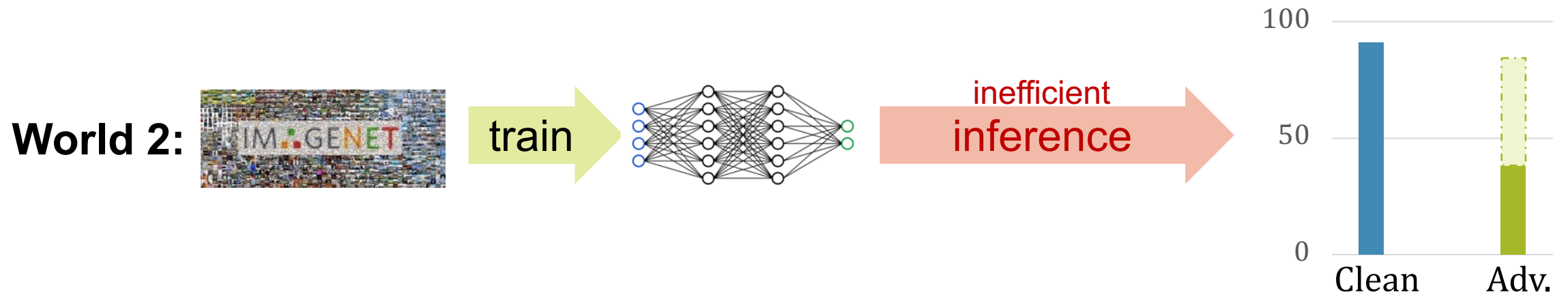


Interpretation #2: robust detectors imply a *breakthrough in robust classification*.



Can we build much more robust classifiers in **World 2**?
(we don't know...)

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

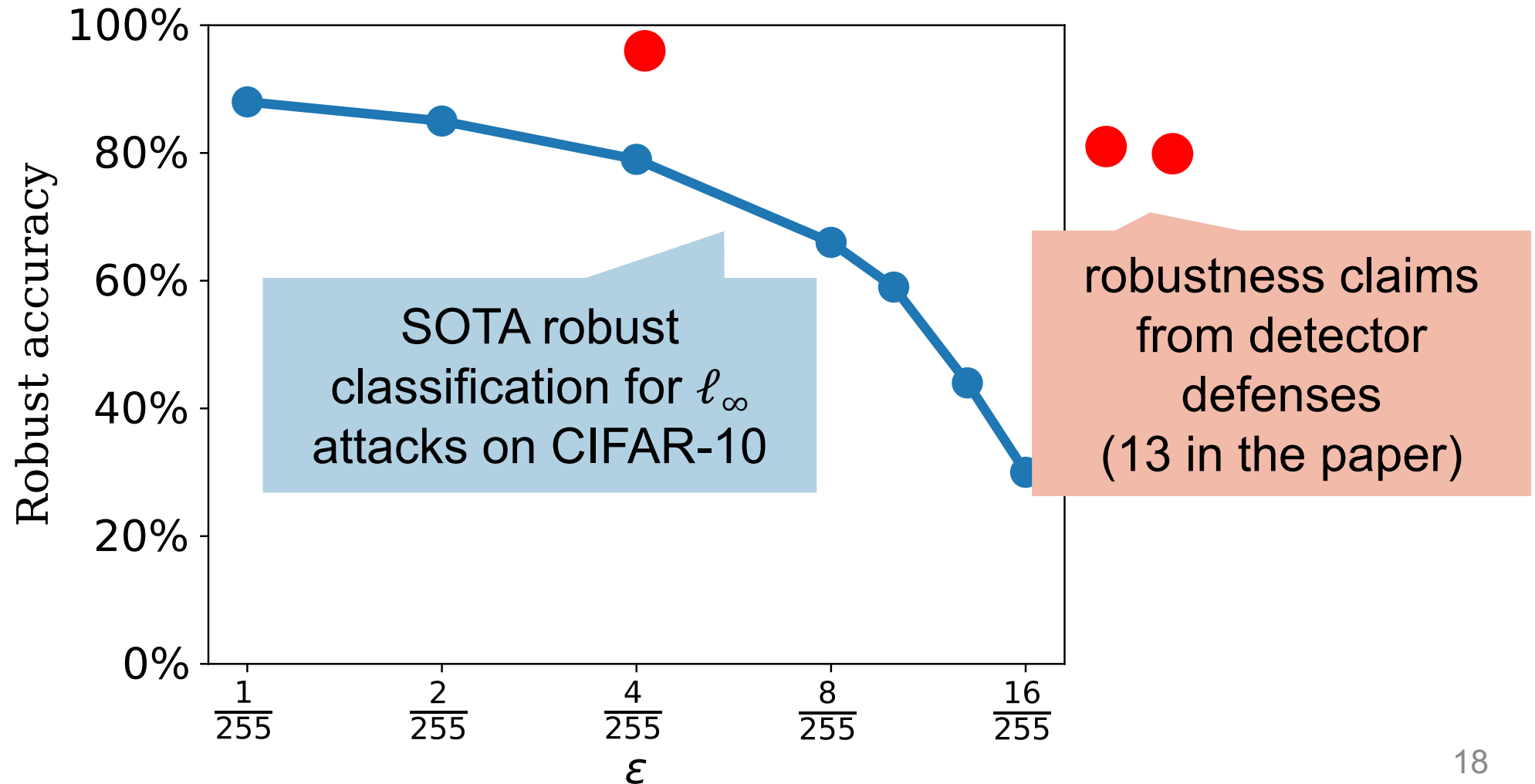


Can we build much more robust classifiers in **World 2**?

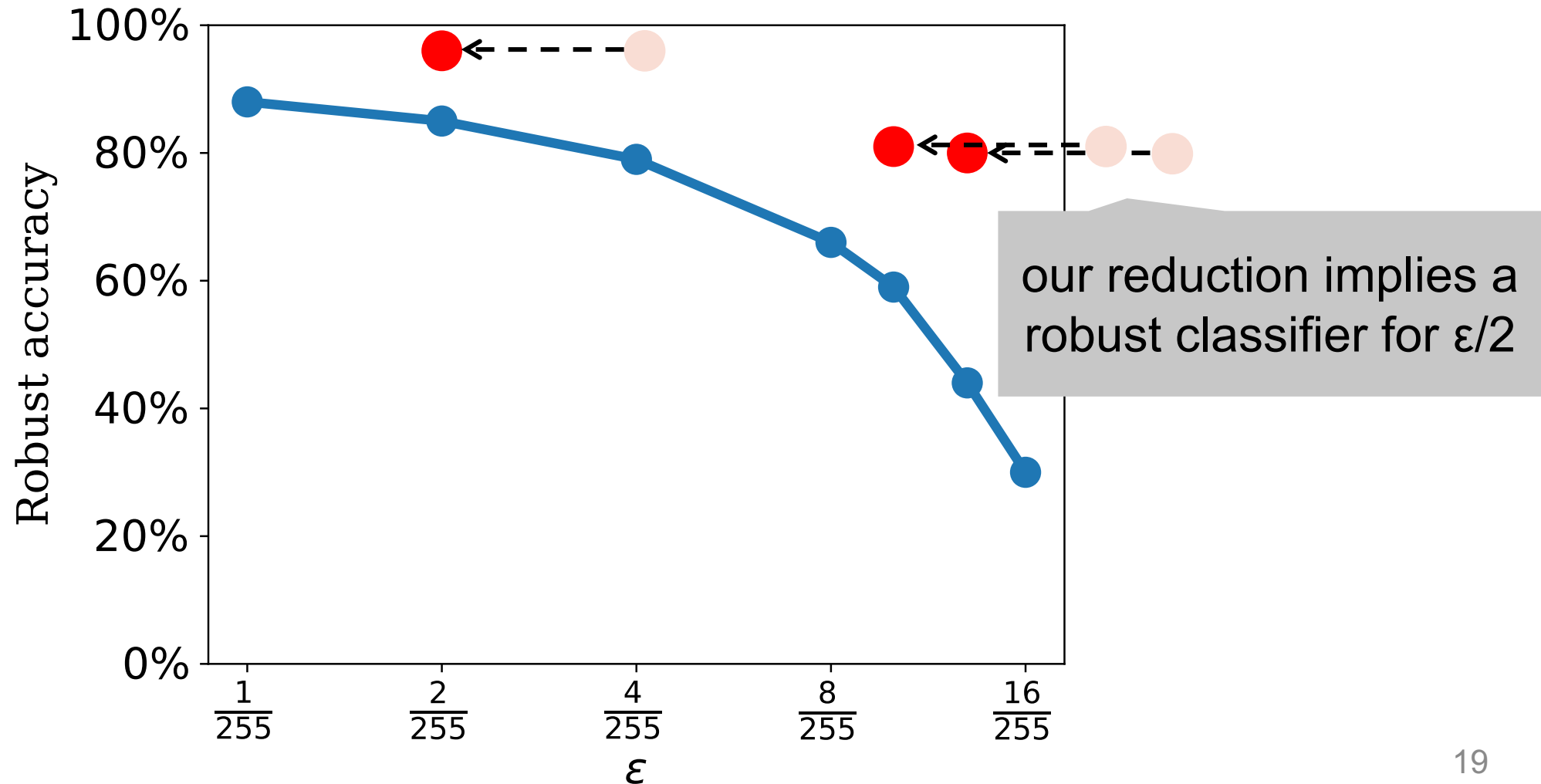
(we don't know...)

But any sufficiently robust detector implies a positive answer!

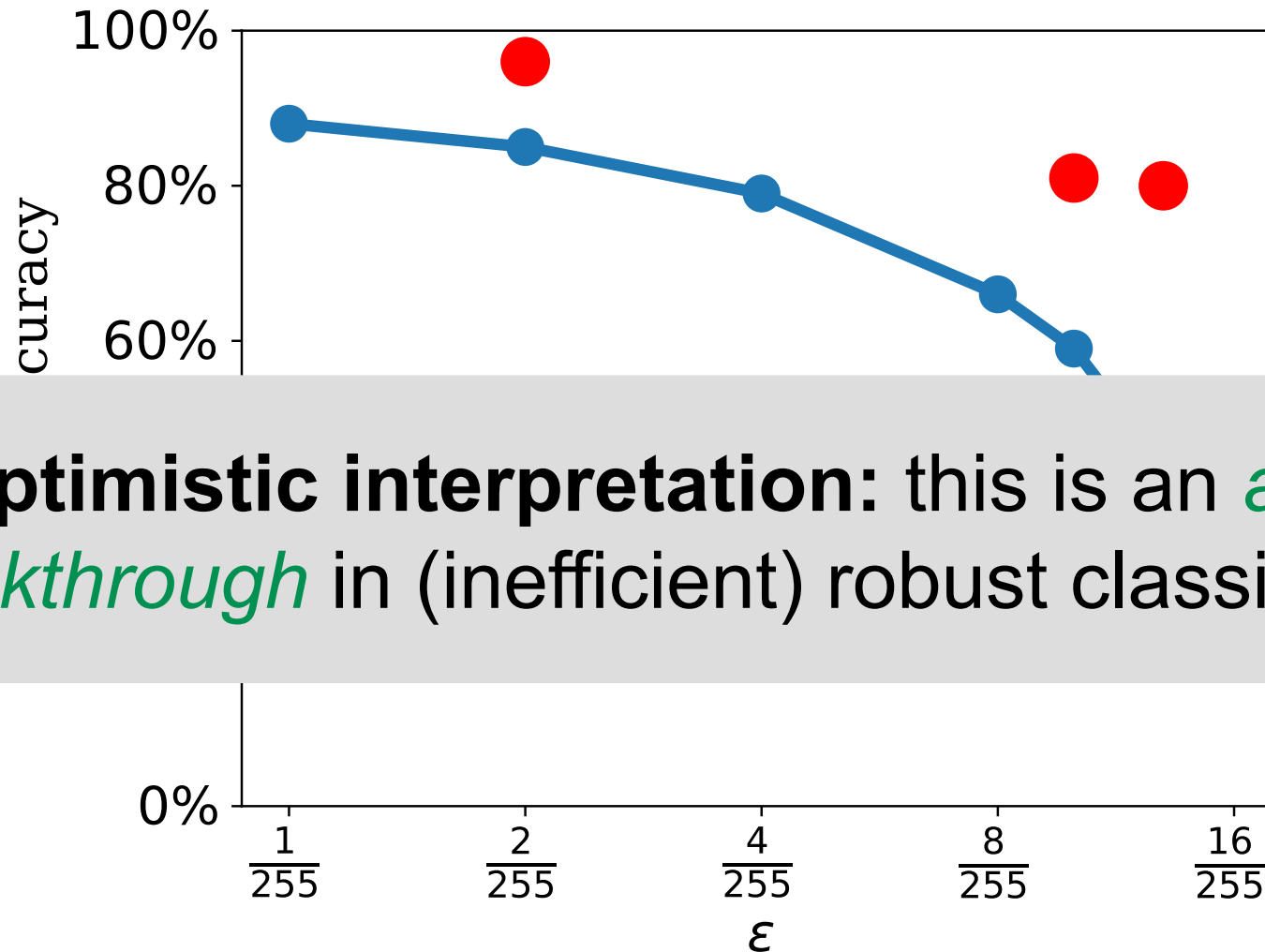
Many detectors *implicitly* claim such a breakthrough!



Many detectors *implicitly* claim such a breakthrough!

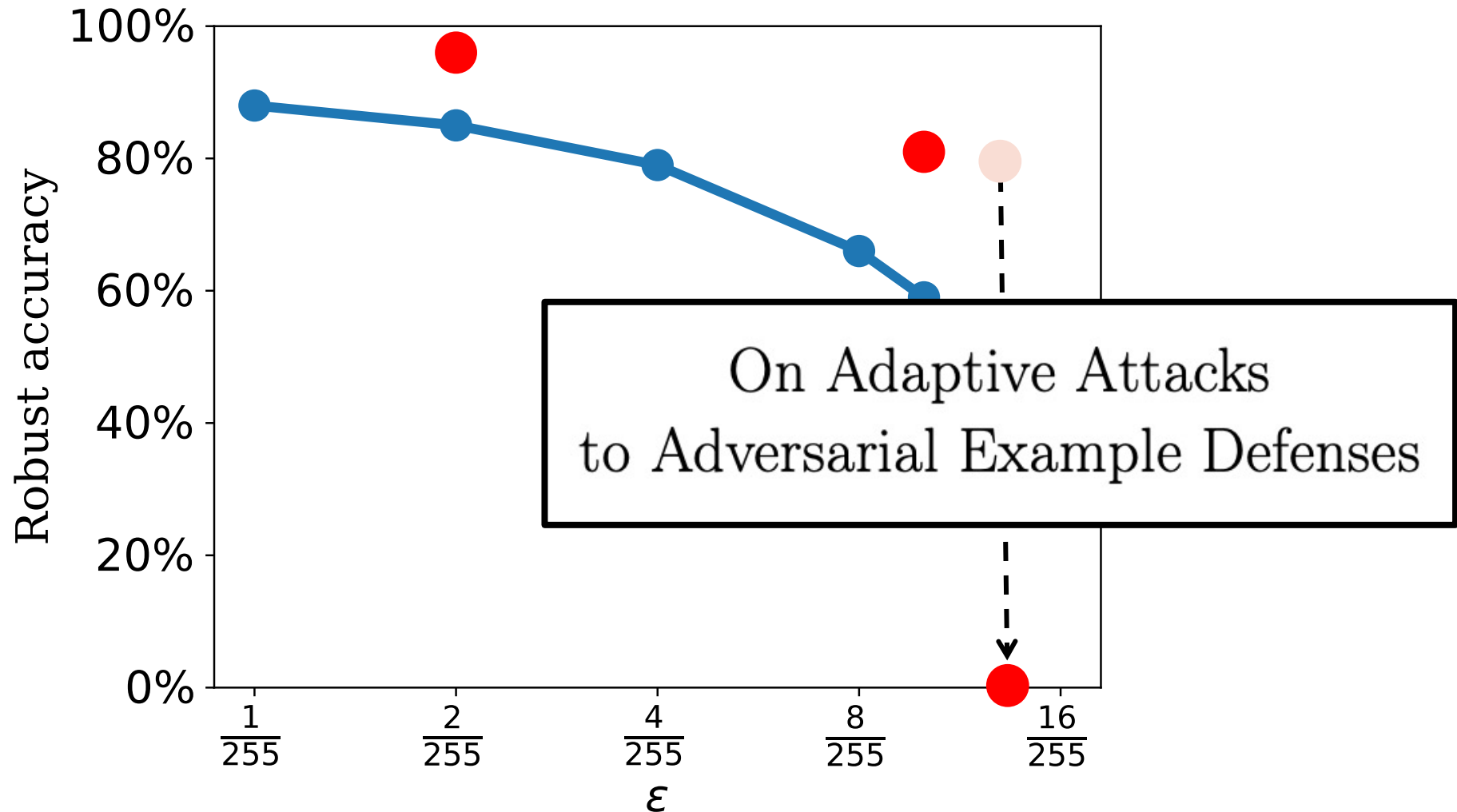


Many detectors *implicitly* claim such a breakthrough!

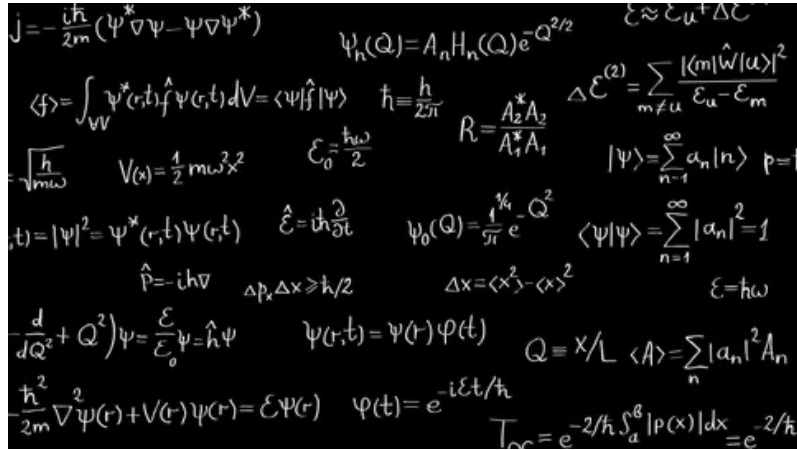


Optimistic interpretation: this is an *actual breakthrough* in (inefficient) robust classification!

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

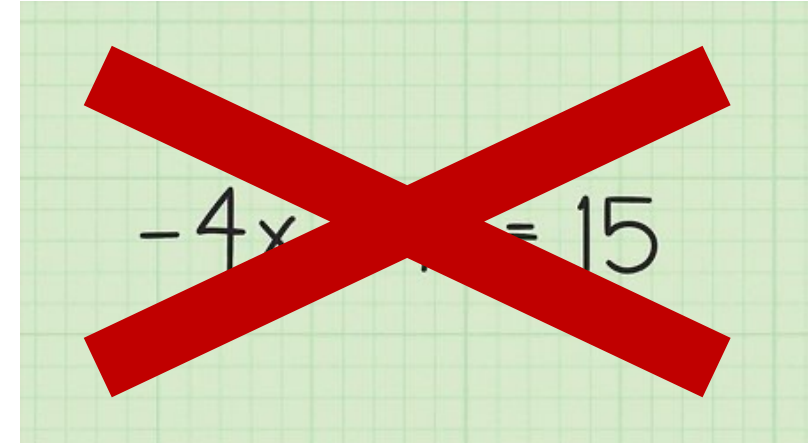


Conclusion.



A collection of handwritten physics equations on a blackboard background. The equations include: $J = -\frac{i\hbar}{2m}(\psi^* \nabla \psi - \psi \nabla \psi^*)$, $\psi_n(Q) = A_n H_n(Q) e^{-Q^2/2}$, $\langle f \rangle = \int_{VV} \psi^*(r,t) \hat{f} \psi(r,t) dV = \langle \psi | \hat{f} | \psi \rangle$, $\hbar = \frac{h}{2\pi}$, $R = \frac{A_2^* A_2}{A_1^* A_1}$, $\Delta \mathcal{E}^{(2)} = \sum_{m \neq u} \frac{|(m|\hat{W}|u)|^2}{\mathcal{E}_u - \mathcal{E}_m}$, $V(x) = \frac{1}{2} m \omega^2 x^2$, $\mathcal{E}_0 = \frac{\hbar \omega}{2}$, $|\psi\rangle = \sum_{n=0}^{\infty} a_n |n\rangle$, $\langle \psi | \psi \rangle = \sum_{n=0}^{\infty} |a_n|^2 = 1$, $\hat{p} = -i\hbar \nabla$, $\Delta p_x \Delta x \geq \hbar/2$, $\Delta x = \langle x^2 \rangle - \langle x \rangle^2$, $\mathcal{E} = \hbar \omega$, $\frac{d}{dt}(\frac{\hbar^2}{2m} \nabla^2 \psi + Q^2 \psi) = \frac{\mathcal{E}}{\hbar} \psi = \hat{H} \psi$, $\psi(r,t) = \psi(r) \varphi(t)$, $Q = x/L$, $\langle A \rangle = \sum_n |a_n|^2 A_n$, $\frac{\hbar^2}{2m} \nabla^2 \psi(r) + V(r) \psi(r) = \mathcal{E} \psi(r)$, $\varphi(t) = e^{-i\mathcal{E}t/\hbar}$, and $T_{cc} = e^{-2/\hbar} \int_a^b |p(x)| dx = e^{-2/\hbar}$.

Robust classification



A green grid background with the equation $-4x + 15$ written in black. A large, thick red 'X' is drawn over the equation, indicating it is incorrect or invalid.

Robust detection

Conclusion.

$$\begin{aligned}
 j &= -\frac{i\hbar}{2m} (\psi^* \nabla \psi - \psi \nabla \psi^*) & \psi_n(Q) &= A_n H_n(Q) e^{-Q^2/2} & \varepsilon &\approx \varepsilon_u + \Delta\varepsilon \\
 \langle f \rangle &= \int_V \psi^*(r,t) \hat{f} \psi(r,t) dV = \langle \psi | \hat{f} | \psi \rangle & \hbar &= \frac{h}{2\pi} & R &= \frac{A_2^* A_2}{A_1^* A_1} & \Delta\varepsilon^{(2)} &= \sum_{m \neq u} \frac{|\langle m | \hat{W} | u \rangle|^2}{\varepsilon_u - \varepsilon_m} \\
 \sqrt{\frac{\hbar}{m\omega}} & & V(x) &= \frac{1}{2} m\omega^2 x^2 & \varepsilon_0 &= \frac{\hbar\omega}{2} & |\psi\rangle &= \sum_{n=1}^{\infty} a_n |n\rangle \quad p=1 \\
 |\psi\rangle &= \sum_{n=1}^{\infty} a_n |n\rangle & \hat{E} &= i\hbar \frac{\partial}{\partial t} & \psi_0(Q) &= \frac{1}{\sqrt{\pi}} e^{-Q^2/2} & \langle \psi | \psi \rangle &= \sum_{n=1}^{\infty} |a_n|^2 = 1 \\
 \hat{p} &= -i\hbar \nabla & \Delta p_x \Delta x &\geq \hbar/2 & \Delta x &= \langle x^2 \rangle - \langle x \rangle^2 & \varepsilon &= \hbar\omega \\
 \frac{d}{dQ^2} (Q^2) \psi &= \frac{\varepsilon}{\varepsilon_0} \psi = \hat{h} \psi & \psi(r,t) &= \psi(r) \varphi(t) & Q &= x/L & \langle A \rangle &= \sum_n |a_n|^2 A_n \\
 \frac{\hbar^2}{2m} \nabla^2 \psi(r) + V(r) \psi(r) &= \varepsilon \psi(r) & \varphi(t) &= e^{-i\varepsilon t/\hbar} & T_{oc} &= e^{-2/\hbar \int_a^b |p(x)| dx} = e^{-2/\hbar}
 \end{aligned}$$


Robust classification


$$\begin{aligned}
 j &= -\frac{i\hbar}{2m} (\psi^* \nabla \psi - \psi \nabla \psi^*) & \psi_n(Q) &= A_n H_n(Q) e^{-Q^2/2} & \varepsilon &\approx \varepsilon_u + \Delta\varepsilon \\
 \langle f \rangle &= \int_V \psi^*(r,t) \hat{f} \psi(r,t) dV = \langle \psi | \hat{f} | \psi \rangle & \hbar &= \frac{h}{2\pi} & R &= \frac{A_2^* A_2}{A_1^* A_1} & \Delta\varepsilon^{(2)} &= \sum_{m \neq u} \frac{|\langle m | \hat{W} | u \rangle|^2}{\varepsilon_u - \varepsilon_m} \\
 \sqrt{\frac{\hbar}{m\omega}} & & V(x) &= \frac{1}{2} m\omega^2 x^2 & \varepsilon_0 &= \frac{\hbar\omega}{2} & |\psi\rangle &= \sum_{n=1}^{\infty} a_n |n\rangle \quad p=1 \\
 |\psi\rangle &= \sum_{n=1}^{\infty} a_n |n\rangle & \hat{E} &= i\hbar \frac{\partial}{\partial t} & \psi_0(Q) &= \frac{1}{\sqrt{\pi}} e^{-Q^2/2} & \langle \psi | \psi \rangle &= \sum_{n=1}^{\infty} |a_n|^2 = 1 \\
 \hat{p} &= -i\hbar \nabla & \Delta p_x \Delta x &\geq \hbar/2 & \Delta x &= \langle x^2 \rangle - \langle x \rangle^2 & \varepsilon &= \hbar\omega \\
 \frac{d}{dQ^2} (Q^2) \psi &= \frac{\varepsilon}{\varepsilon_0} \psi = \hat{h} \psi & \psi(r,t) &= \psi(r) \varphi(t) & Q &= x/L & \langle A \rangle &= \sum_n |a_n|^2 A_n \\
 \frac{\hbar^2}{2m} \nabla^2 \psi(r) + V(r) \psi(r) &= \varepsilon \psi(r) & \varphi(t) &= e^{-i\varepsilon t/\hbar} & T_{oc} &= e^{-2/\hbar \int_a^b |p(x)| dx} = e^{-2/\hbar}
 \end{aligned}$$


Robust detection

ON EVALUATING ADVERSARIAL ROBUSTNESS


Nicholas Carlini¹, Anish Athanasopoulos², Dimitris Tsipras², Ian Goodfellow¹

 [MadryLab / mnist_challenge](https://madrylab.github.io/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

Conclusion.

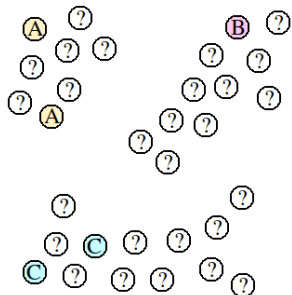
$$\begin{aligned}
 j &= -\frac{i\hbar}{2m}(\psi^* \nabla \psi - \psi \nabla \psi^*) & \psi_n(Q) &= A_n H_n(Q) e^{-Q^2/2} & \xi &\approx \xi_u + \Delta \xi \\
 \langle f \rangle &= \int_{VV} \psi^*(r,t) \hat{f} \psi(r,t) dV = \langle \psi | \hat{f} | \psi \rangle & \hbar &= \frac{h}{2\pi} & R &= \frac{A_2^* A_2}{A_1^* A_1} & \Delta \mathcal{E}^{(2)} &= \sum_{m \neq u} \frac{|(m|\hat{W}|u)|^2}{\mathcal{E}_u - \mathcal{E}_m} \\
 \sqrt{\frac{\hbar}{m\omega}} & & V(x) &= \frac{1}{2} m\omega^2 x^2 & \mathcal{E}_0 &= \frac{\hbar\omega}{2} & |\psi\rangle &= \sum_{n=1}^{\infty} a_n |n\rangle & p &= \hbar \\
 \langle \psi | \psi \rangle &= \int \psi^*(r,t) \psi(r,t) dV & \hat{\mathcal{E}} &= i\hbar \frac{\partial}{\partial t} & \psi_0(Q) &= \frac{1}{\sqrt{\pi}} e^{-Q^2/2} & \langle \psi | \psi \rangle &= \sum_{n=1}^{\infty} |a_n|^2 = 1 \\
 \hat{p} &= -i\hbar \nabla & \Delta p_x \Delta x &\geq \hbar/2 & \Delta x &= \langle x^2 \rangle - \langle x \rangle^2 & \mathcal{E} &= \hbar\omega \\
 \frac{d}{dQ^2} (Q^2) \psi &= \frac{\mathcal{E}}{\mathcal{E}_0} \psi = \hat{h} \psi & \psi(r,t) &= \psi(r) \varphi(t) & Q &= x/L & \langle A \rangle &= \sum_n |a_n|^2 A_n \\
 \frac{\hbar^2}{2m} \nabla^2 \psi(r) + V(r) \psi(r) &= \mathcal{E} \psi(r) & \varphi(t) &= e^{-i\mathcal{E}t/\hbar} & T_{oc} &= e^{-2/\hbar \int_a^b |p(x)| dx} = e^{-2/\hbar}
 \end{aligned}$$

Robust classification

$$\begin{aligned}
 j &= -\frac{i\hbar}{2m}(\psi^* \nabla \psi - \psi \nabla \psi^*) & \psi_n(Q) &= A_n H_n(Q) e^{-Q^2/2} & \xi &\approx \xi_u + \Delta \xi \\
 \langle f \rangle &= \int_{VV} \psi^*(r,t) \hat{f} \psi(r,t) dV = \langle \psi | \hat{f} | \psi \rangle & \hbar &= \frac{h}{2\pi} & R &= \frac{A_2^* A_2}{A_1^* A_1} & \Delta \mathcal{E}^{(2)} &= \sum_{m \neq u} \frac{|(m|\hat{W}|u)|^2}{\mathcal{E}_u - \mathcal{E}_m} \\
 \sqrt{\frac{\hbar}{m\omega}} & & V(x) &= \frac{1}{2} m\omega^2 x^2 & \mathcal{E}_0 &= \frac{\hbar\omega}{2} & |\psi\rangle &= \sum_{n=1}^{\infty} a_n |n\rangle & p &= \hbar \\
 \langle \psi | \psi \rangle &= \int \psi^*(r,t) \psi(r,t) dV & \hat{\mathcal{E}} &= i\hbar \frac{\partial}{\partial t} & \psi_0(Q) &= \frac{1}{\sqrt{\pi}} e^{-Q^2/2} & \langle \psi | \psi \rangle &= \sum_{n=1}^{\infty} |a_n|^2 = 1 \\
 \hat{p} &= -i\hbar \nabla & \Delta p_x \Delta x &\geq \hbar/2 & \Delta x &= \langle x^2 \rangle - \langle x \rangle^2 & \mathcal{E} &= \hbar\omega \\
 \frac{d}{dQ^2} (Q^2) \psi &= \frac{\mathcal{E}}{\mathcal{E}_0} \psi = \hat{h} \psi & \psi(r,t) &= \psi(r) \varphi(t) & Q &= x/L & \langle A \rangle &= \sum_n |a_n|^2 A_n \\
 \frac{\hbar^2}{2m} \nabla^2 \psi(r) + V(r) \psi(r) &= \mathcal{E} \psi(r) & \varphi(t) &= e^{-i\mathcal{E}t/\hbar} & T_{oc} &= e^{-2/\hbar \int_a^b |p(x)| dx} = e^{-2/\hbar}
 \end{aligned}$$

Robust detection

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



<https://arxiv.org/abs/2107.11630>

<https://floriantramer.com>