MIT Quest for Intelligence

Robust, Interpretable Deep Learning Systems
November 20, 2018

CSAIL

IBM | MIT

The RIDL symposium is affiliated with the Robust Intelligence Initiative @ CSAIL funded by Microsoft and the MIT-IBM AI Watson Lab.

Robustness and Interpretability

Aleksander Mądry

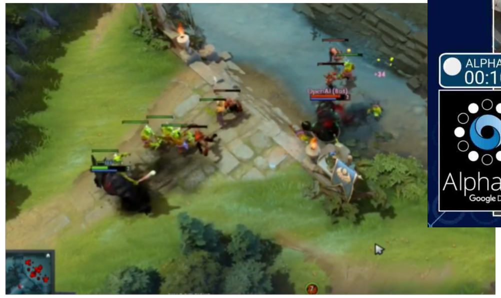
MIT

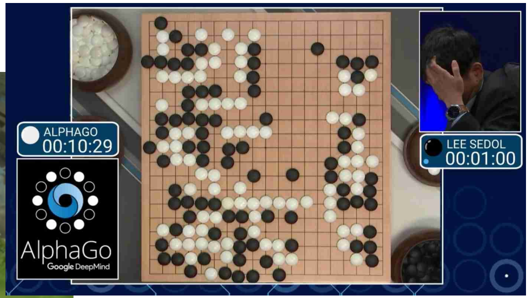
@aleks_madry

madry-lab.ml

Machine Learning: The Success Story?

Image classification





Reinforcement Learning

Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Machine translation

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

Andrew Ng
@AndrewYNg
Follow

"AI is the new electricity!" Electricity transformed countless industries; AI will now do the same.

2016: The Year That Deep Learning Took Over th

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE

Machine Learning: The Success Story?

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

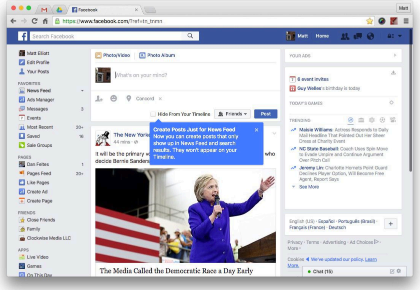
Andrew Ng @AndrewYNg

Follow

"AI is the new electricity!" Electricity transformed countless industries; AI will now do the same.

2016: The Year That Deep Learning Took Over t

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE



Machine Learning: The Success Story?

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

Andrew Ng @AndrewYNg
Follow

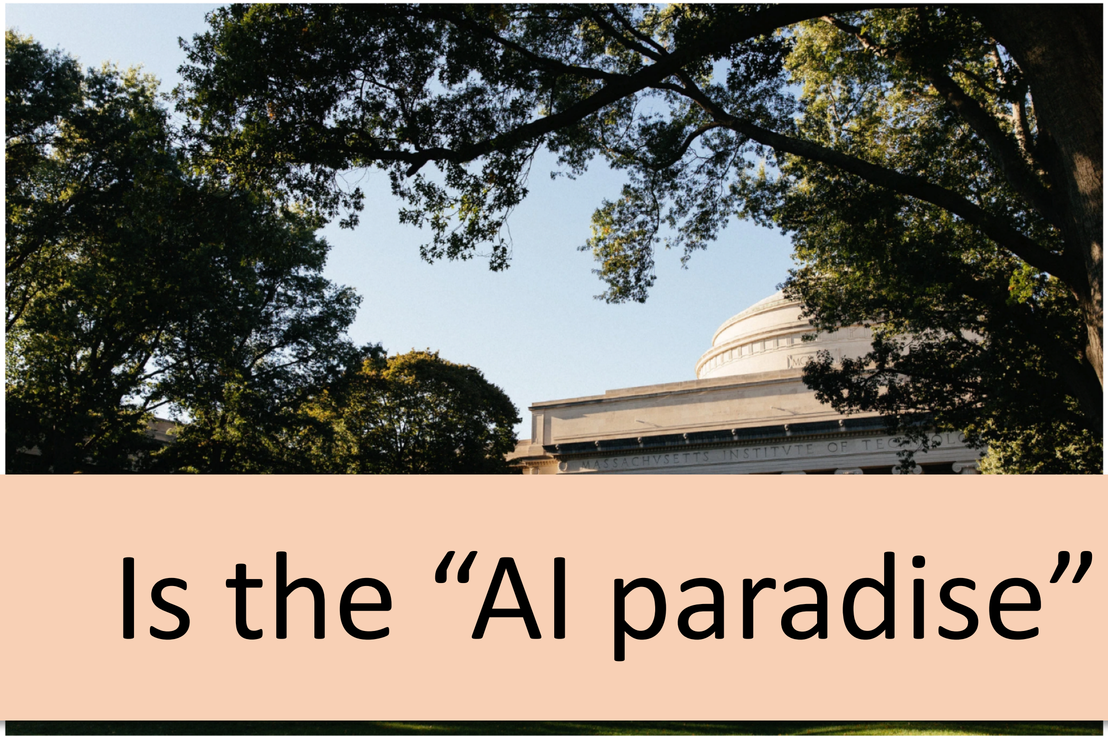
"AI is the new electricity!" Electricity transformed countless industries; AI will now do the same.

2016: The Year That Deep Learning Took Over t

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE

The New York Times

M.I.T. Plans College for Artificial Intelligence, Backed by $1 Billion



Is the "AI paradise" already here?

Is our ML truly ready for deployment?

Overarching questions:

→ Do we **really** understand how/why/**if** our ML tools work?

→ Is our ML toolkit even tackling the right question?

Today

→ Can we make this toolkit be more transparent/"interpretable"
(Also: fairness, accountability, contestability,…)

Can We Truly Rely on ML?


AP The Associated Press
@AP
Following

Breaking: Two Explosions in the White House and Barack Obama is injured

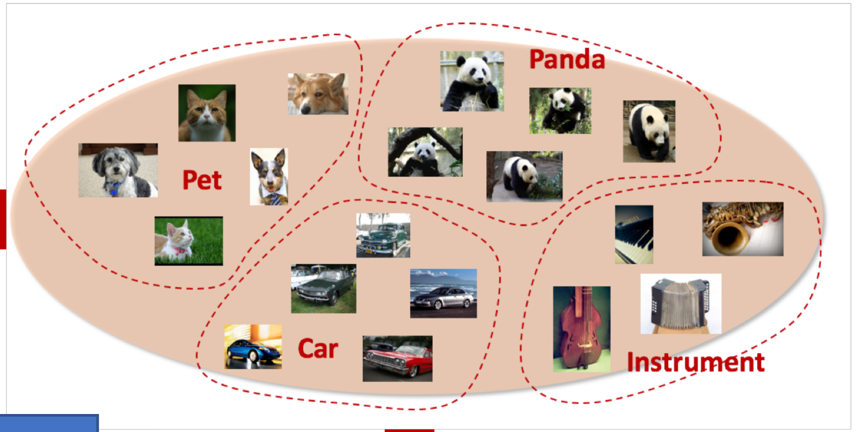
Reply Retweet Favorite More

3,063 RETWEETS 144 FAVORITES

12:07 PM - 23 Apr 13

The Big Lie of (Supervised) ML

Distribution $\boldsymbol{D}$



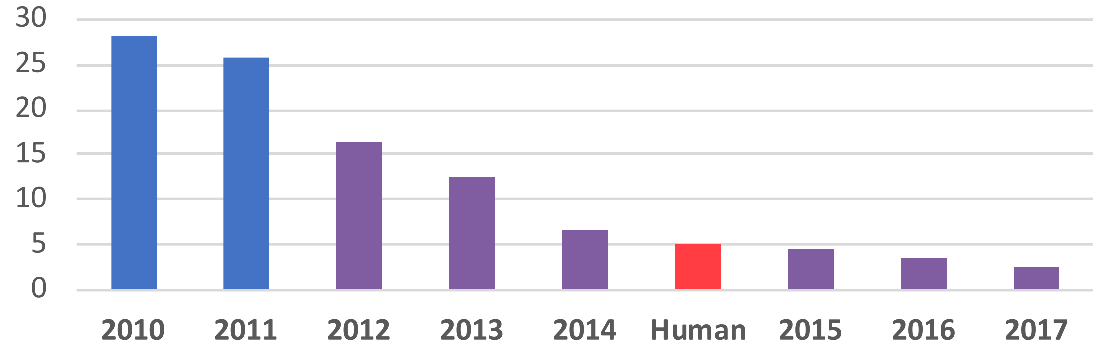
training

How we think about (and evaluate) ML:

Dataset → Training Algorithm → Model

testing

ILSVRC top-5 Error on ImageNet



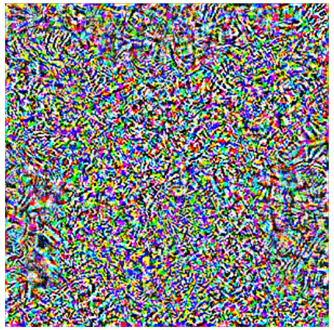
But: In reality, the distributions we **use** ML on are not the ones we **train** it on

What can go wrong?

ML Predictions Are (Mostly) Accurate but Brittle

“pig” (91%)

+ 0.005 x



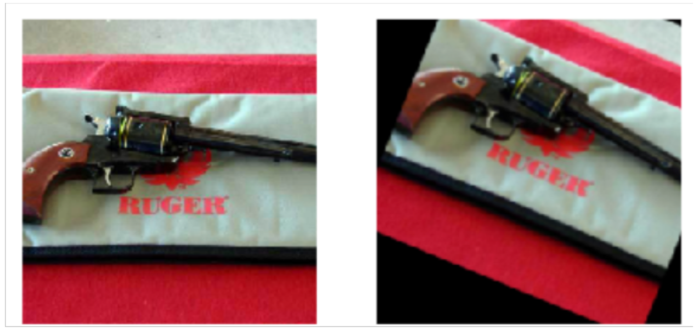
=

“airliner” (99%)

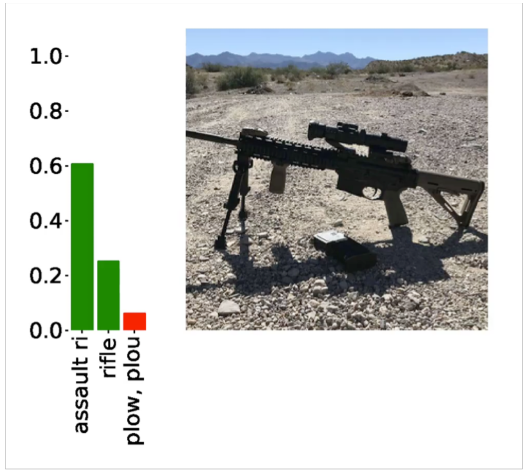
[Szegedy et al. 2014]: Imperceptible noise (adversarial examples) can fool state-of-the-art classifiers

[Athalye Engstrom Ilyas Kwok 2017]: 3D-printed **turtle** model classified as **rifle** from most viewpoints

“revolver” **“mouse trap”**



[Engstrom Tran Tsipras Schmidt M 2018]:
Rotation + Translation Suffices



Should we be worried?

Why Is This Brittleness of ML a Problem?

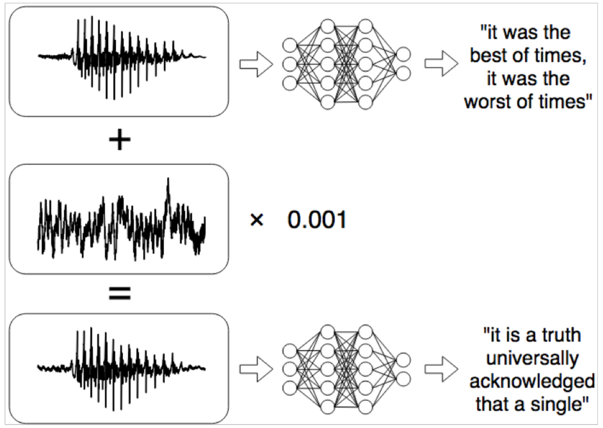
→ Security

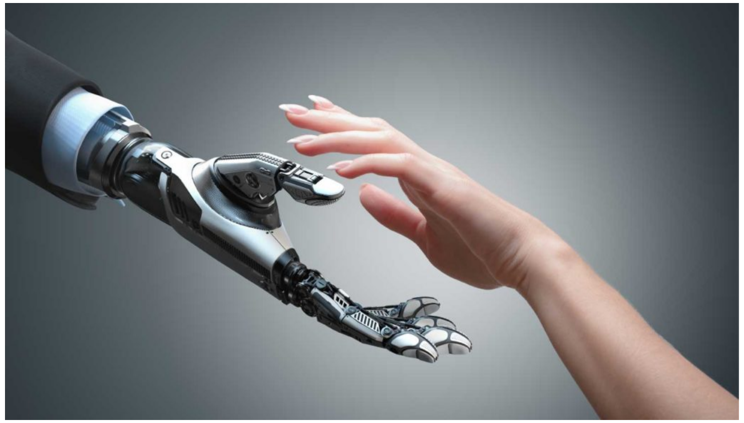
→ Safety

→ ML Alignment

[Sharif et al. 2016]:
Glasses that fool face recognition

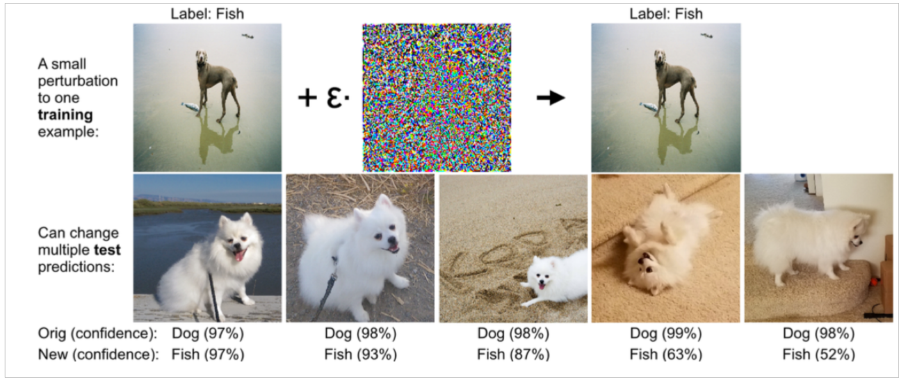
[Carlini Wagner 2018]:
Voice commands that are unintelligible to humans





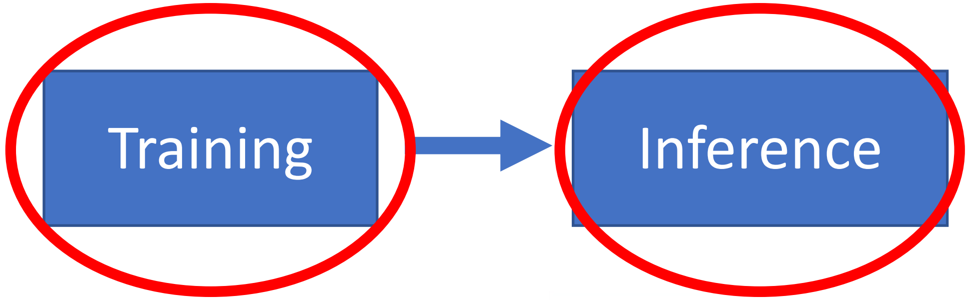
Need to understand the "failure modes" of ML

Is That It?



[Koh Liang 2017]:
Can cause misclassification of **multiple** inputs with a **single** "poisoned" training input

[Gu Dolan-Gavitt Garg 2017][Tsipras Turner **M** 2018]:
Can plant an **undetectable backdoor** that gives an almost **total** control over the model

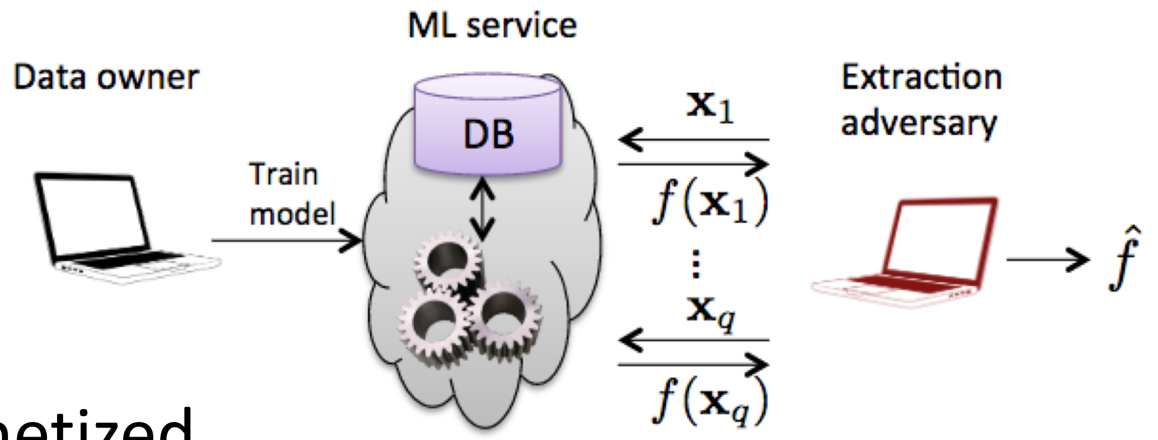


Data Poisoning

Adversarial Examples

Is That It?

[Tramer et al. 2016]:
Can recover a "copy" of the model using **only the prediction API**



→ The "stolen" model can then be monetized

→ Proprietary datasets for model training are no longer a competitive advantage

Input → ML model → Prediction

Side note: Enables constructing adv. examples **without** full access to the model
[Szegedy et al 2013][Chen et al 2017] [Ilyas et al 2018] [Engstrom et al 2018]

Training → Inference → Profit?

Data Poisoning

Model Stealing

Three commandments of Secure/Safe ML

I. Thou shall not train on data you don't fully trust

(because of data poisoning)

II. Thou shall not let anyone use your model (or observe its outputs) unless you completely trust them

(because of model stealing and black box attacks)

III. Thou shall not fully trust the predictions of your model

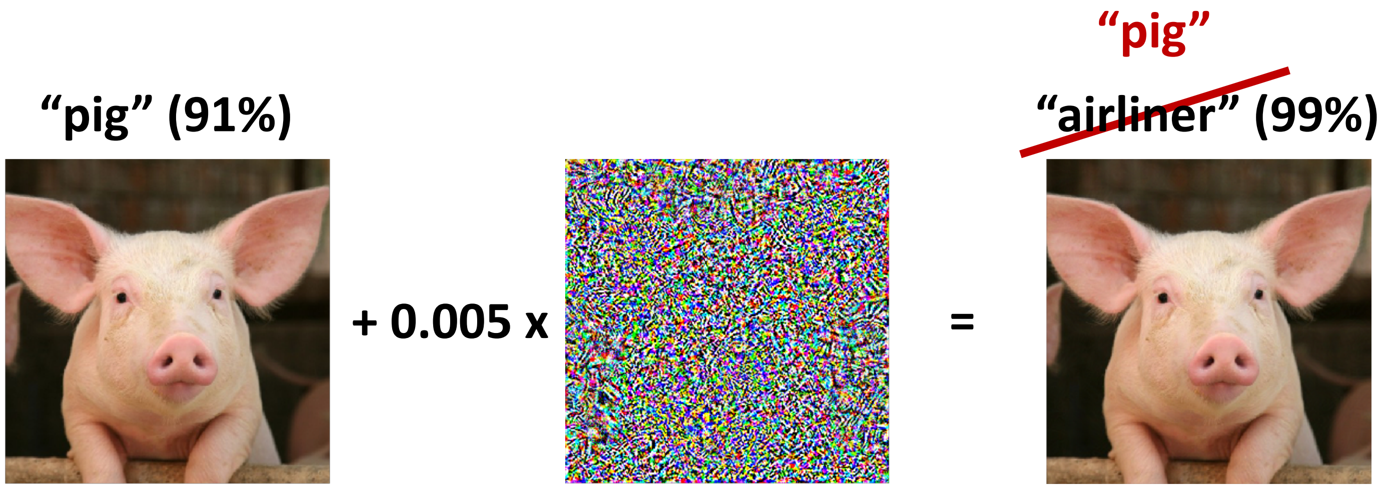
(because of adversarial examples)

Are we doomed?

No, but we need to re-think how we do ML

(**Think:** adversarial aspects = stress-testing our solutions)

Towards Adversarially Robust Models



Towards ML Models that Are Adv. Robust

Key observation: Lack of adv. robustness is **NOT** at odds with what we currently want our ML models to achieve

~~Standard~~ generalization = do well on “random” inputs

adversarially robust

adversarial perturbations of

But: Adversarial noise is NOT random

→ Once

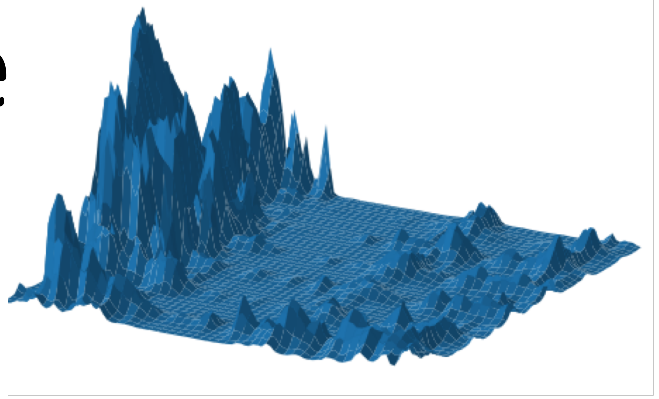
aining

lu 2018]

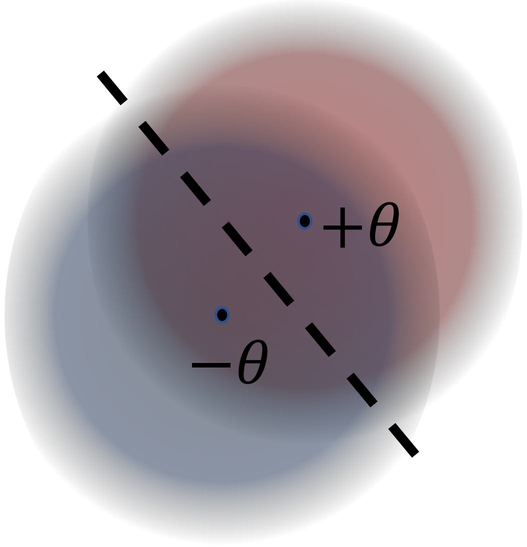
Emerging Question:

How does “adv. robust ML” differ from “standard ML”?

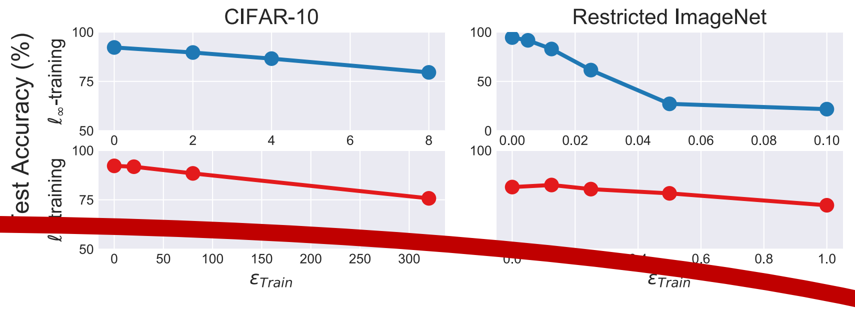
Adversarial Robustness is Not Free



→ Optimization during training more difficult



→ More training data might be required
[Schmidt Santurkar Tsipras Talwar M 2018]



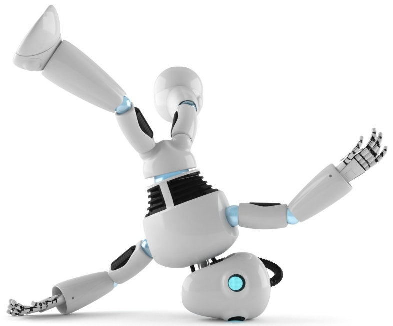
→ Might need to lose on “standard” measures of performance
[Tsipras Santurkar Engstrom Turner M 2018]

Another Challenge: ”Interpretability”

Getting a good “black box” performance is nice

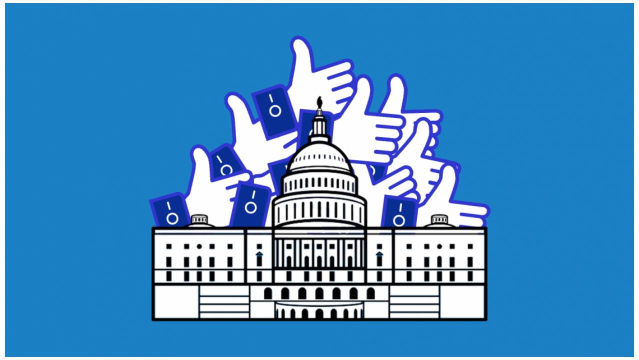
→ **But:** we often need to know how our system makes its decisions too

Why?



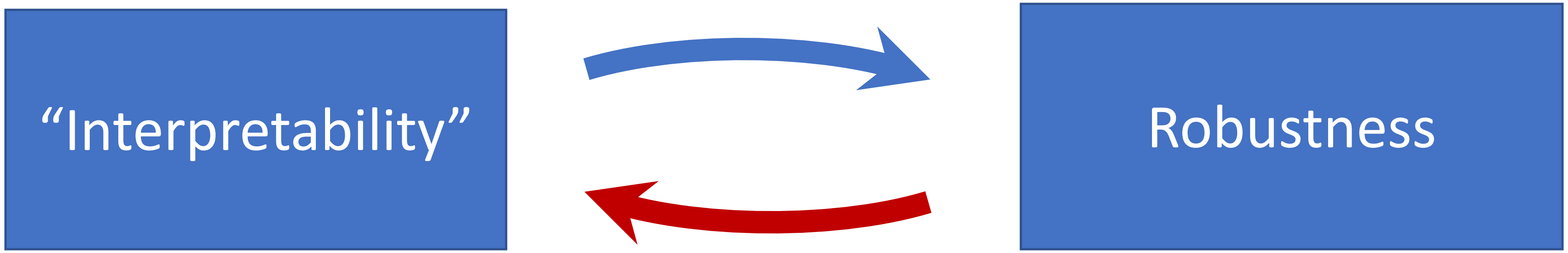
Diagnosing failure cases

Human-ML interaction



Legal/compliance aspects

”Interpretability” and Adv. Robustness



→ If the model is ”interpretable”, it is easier to diagnose its failure
(and counteract this failure)

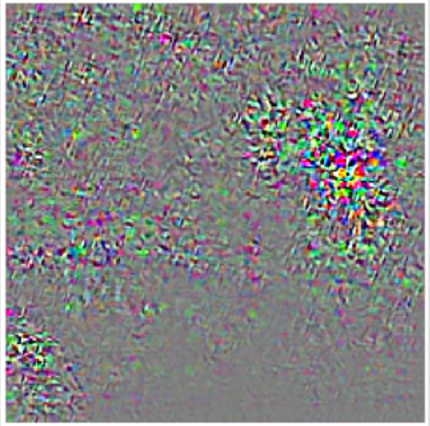
→ **But:** Robustness can inform “interpretability” too

(Unexpected?) Benefits of Adv. Robustness

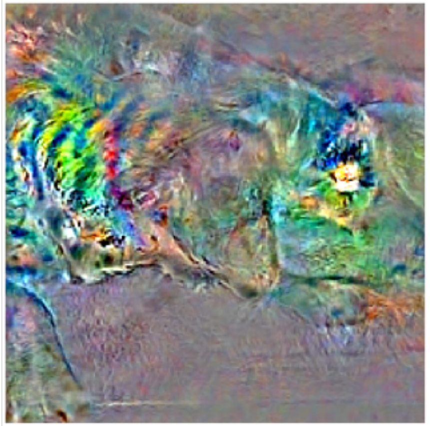
[Tsipras Santurkar Engstrom Turner M 2018]

→ Gradients are more **interpretable** (they yield saliency maps)

Input



gradient of
standard model



gradient of
adv. robust model

→ "Adversarial" examples become
semantically meaningful

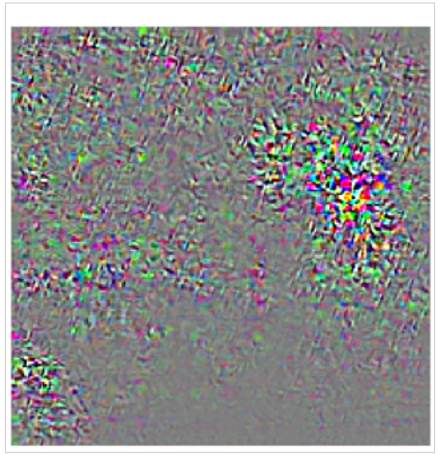
Adversarial example for
standard model

(Unexpected?) Benefits of Adv. Robustness

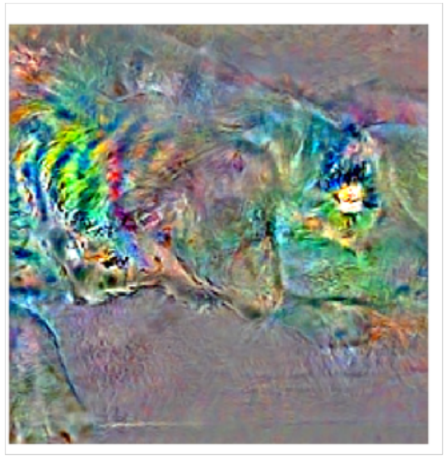
[Tsipras Santurkar Engstrom Turner M 2018]

→ Gradients are more **interpretable** (they yield saliency maps)

Input



gradient of
standard model



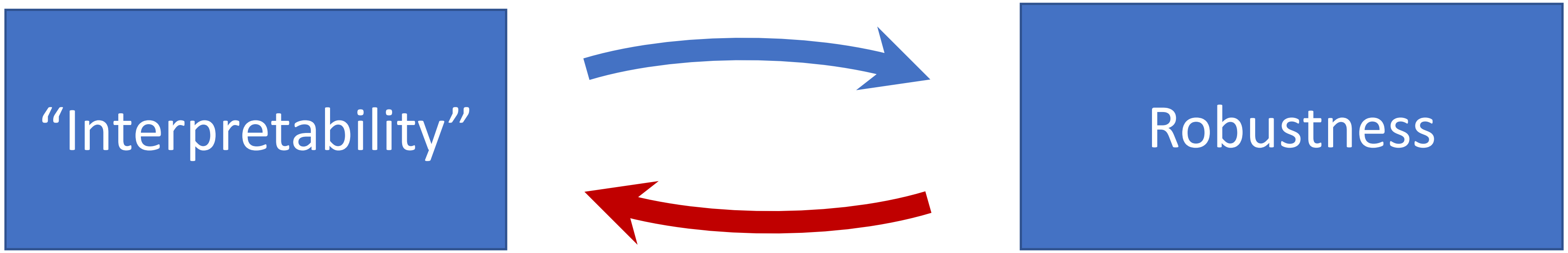
gradient of
adv. robust model

→ "Adversarial" examples become
semantically meaningful

"Adversarial" example for
adv. robust model

(See the poster for more details)

”Interpretability” and Adv. Robustness



→ If the model is ”interpretable”, it is easier to diagnose its failure
(and counteract this failure)

→ **But:** Robustness can inform “interpretability” too

Can we further bridge these two concepts?

Conclusions

→ We're getting somewhere in ML/AI and this is exciting

→ **But:** It is still Wild West out there
(we struck gold but there is lots of fool's gold too)

ML/AI = a sharp knife

→ We still need to learn how to wield it properly
(so we don't hurt ourselves)

Next frontier: Building ML/AI you can truly rely on

→ "Interpretability" will be a key goal **and** tool here

Want to learn more? Take a look at our blog on **gradientscience.org**

@aleks_madry **madry-lab.ml**