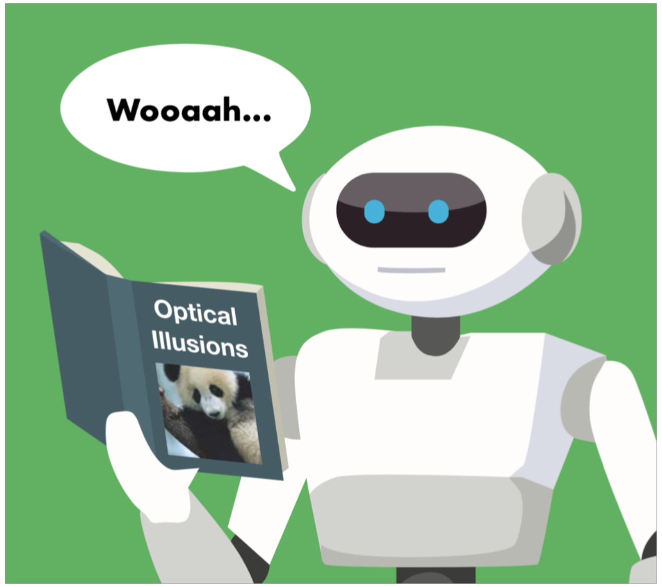
Adversarial Robustness: Theory and Practice



Zico Kolter

Carnegie Mellon University

BOSCH

Aleksander Mądry

MIT

madry-lab.ml

Tutorial website:
adversarial-ml-tutorial.org

@zicokolter

@aleks_madry

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

Machine Learning: The Success Story

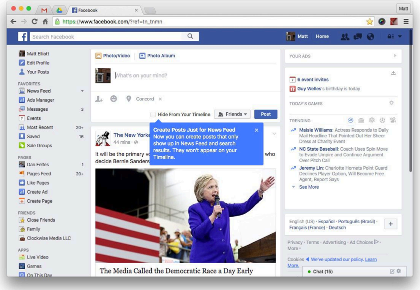
IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

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 th

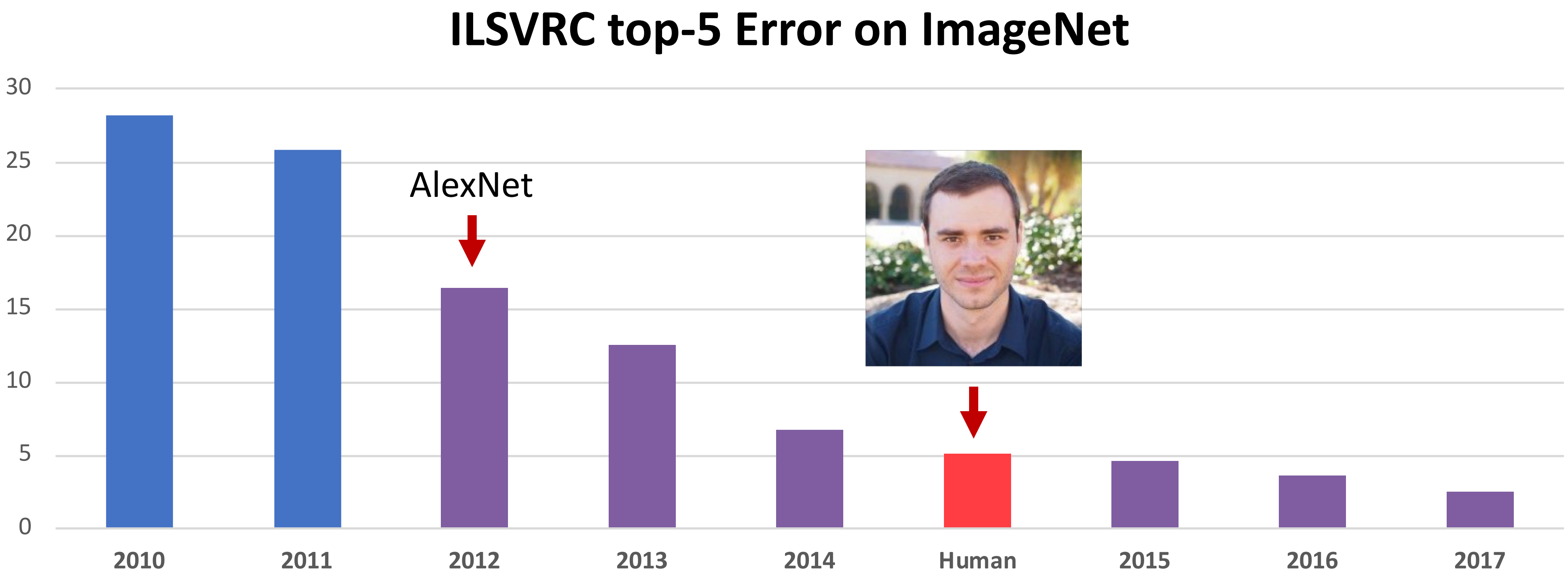
WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE



Is ML **truly** ready for real-world deployment?

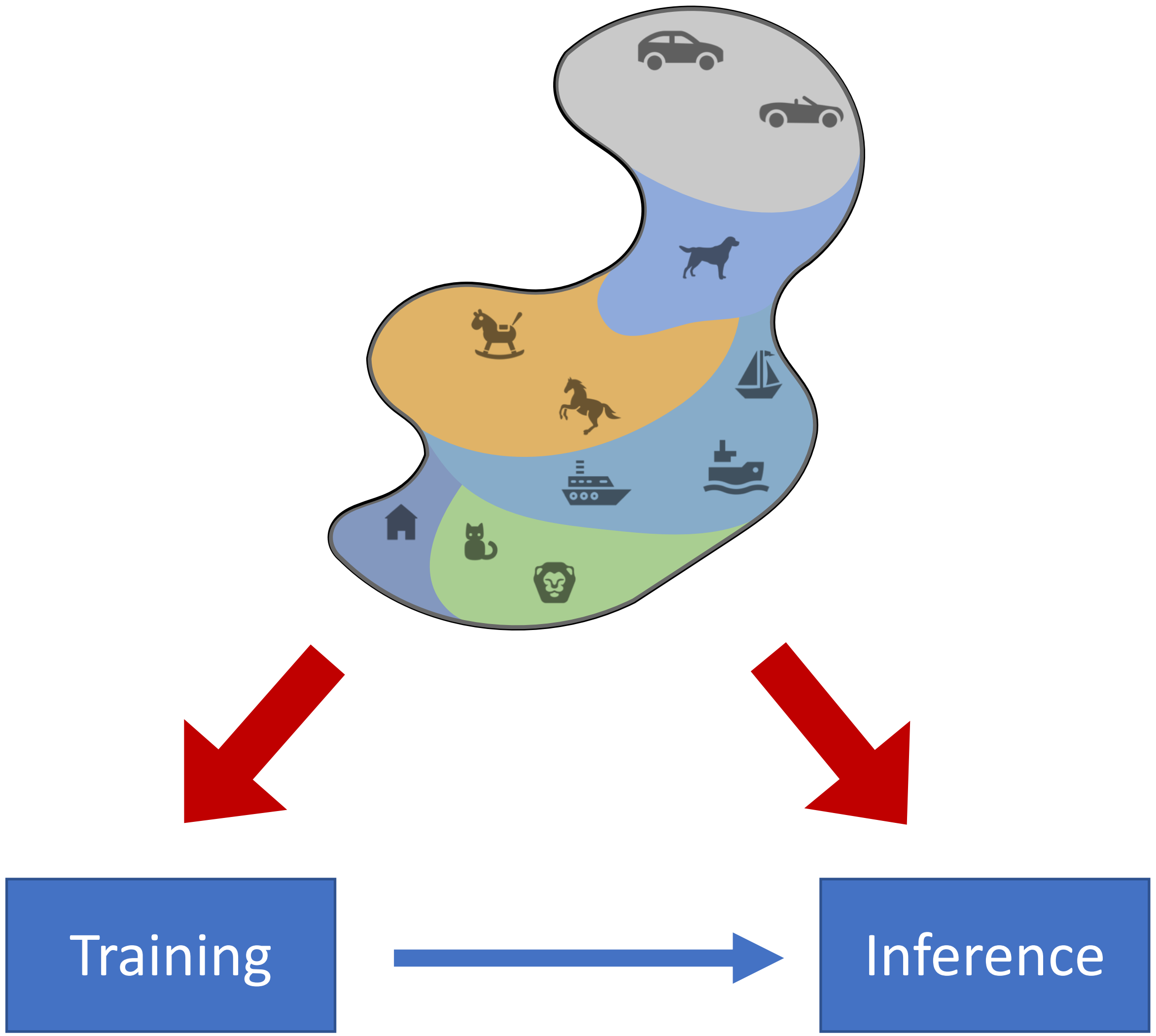
Can We Truly Rely on ML?

ImageNet: An ML Home Run



But what do these results *really* mean?

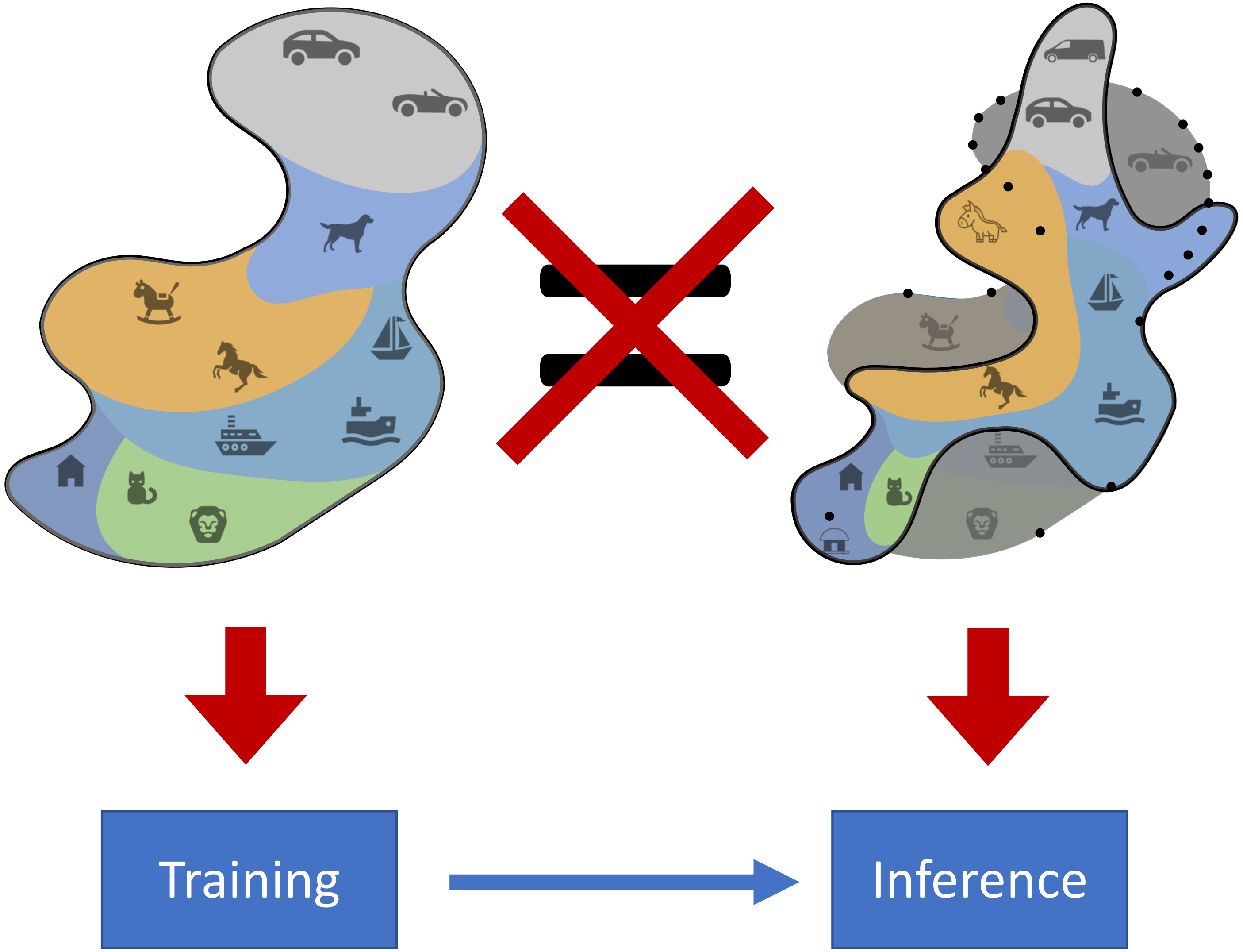
A Limitation of the (Supervised) ML Framework



Measure of performance:
Fraction of mistakes during testing

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

A Limitation of the (Supervised) ML Framework

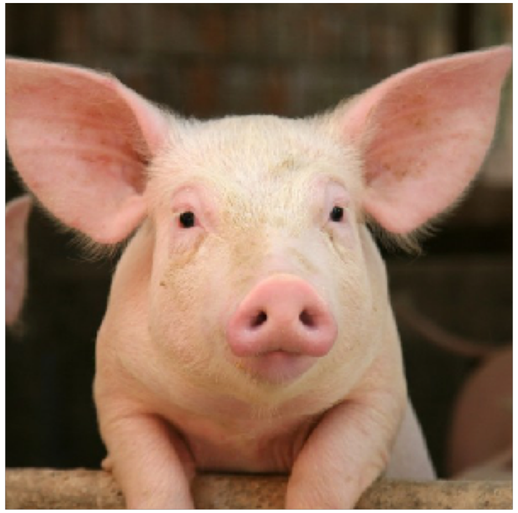


Measure of performance:
Fraction of mistakes during testing

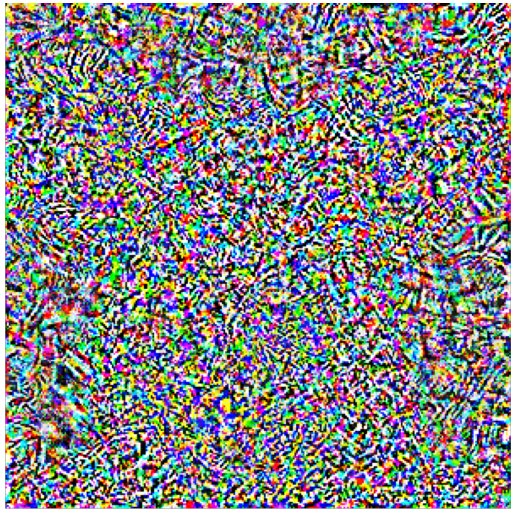
But: In reality, the distributions we **use** ML on are NOT the ones we **train** it on

What can go wrong?

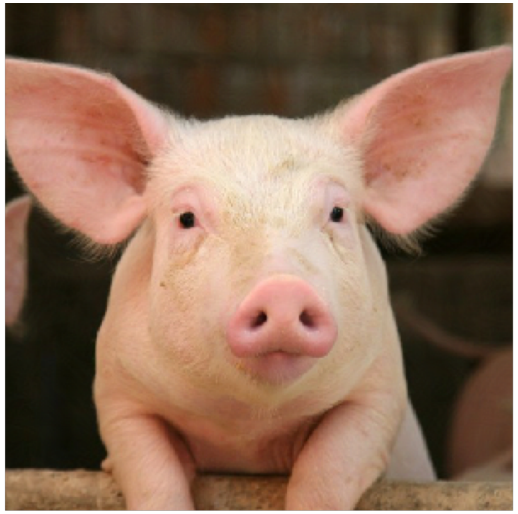
ML Predictions Are (Mostly) Accurate but Brittle

"pig" (91%) 

+ 0.005 x

noise (NOT random) 

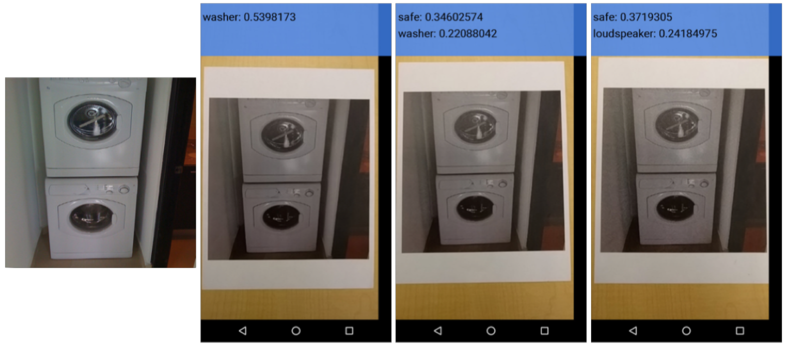
=

"airliner" (99%) 

[Szegedy Zaremba Sutskever Bruna Erhan Goodfellow Fergus 2013]
[Biggio Corona Maiorca Nelson Srndic Laskov Giacinto Roli 2013]

But also: [Dalvi Domingos Mausam Sanghai Verma 2004][Lowd Meek 2005]
[Globerson Roweis 2006][Kolcz Teo 2009][Barreno Nelson Rubinstein Joseph Tygar 2010]
[Biggio Fumera Roli 2010][Biggio Fumera Roli 2014][Srndic Laskov 2013]

ML Predictions Are (Mostly) Accurate but Brittle



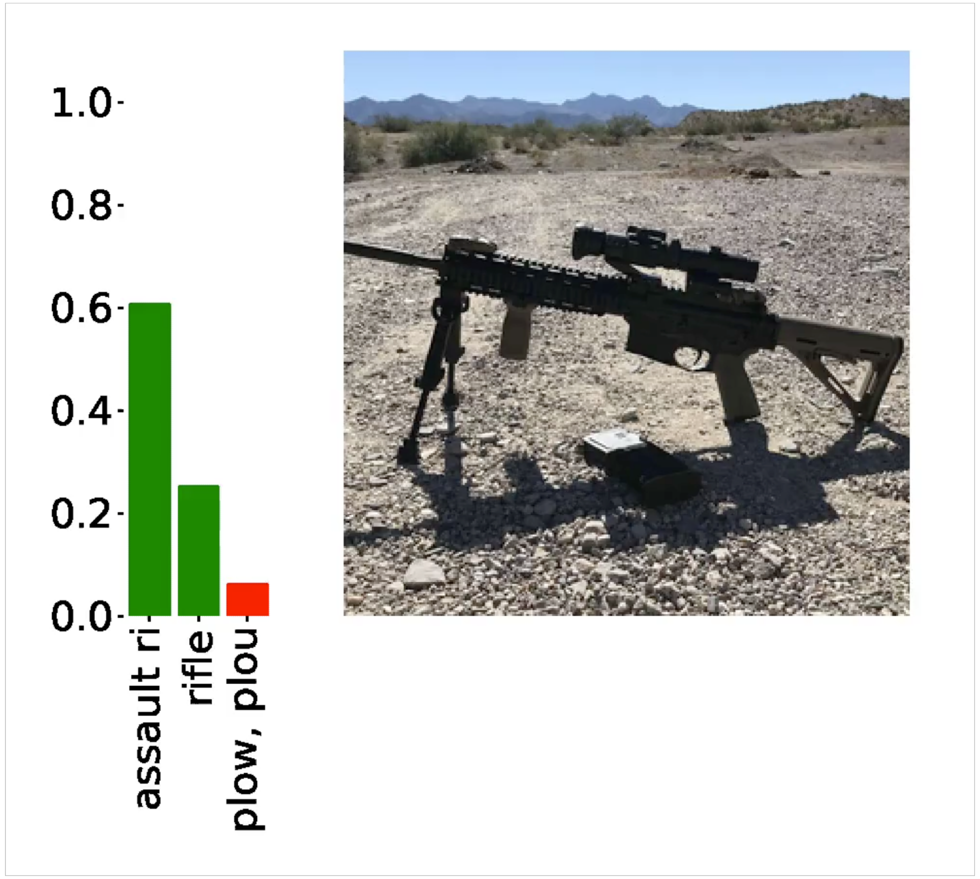
[Kurakin Goodfellow Bengio 2017]

[Athalye Engstrom Ilyas Kwok 2017]

[Sharif Bhagavatula Bauer Reiter 2016]

[Eykholt Evtimov Fernandes Li Rahmati Xiao Prakash Kohno Song 2017]

ML Predictions Are (Mostly) Accurate but Brittle



[Fawzi Frossard 2015]
[Engstrom Tran Tsipras Schmidt M 2018]:
Rotation + Translation suffices to fool state-of-the-art vision models

→ Data augmentation does **not** seem to help here either

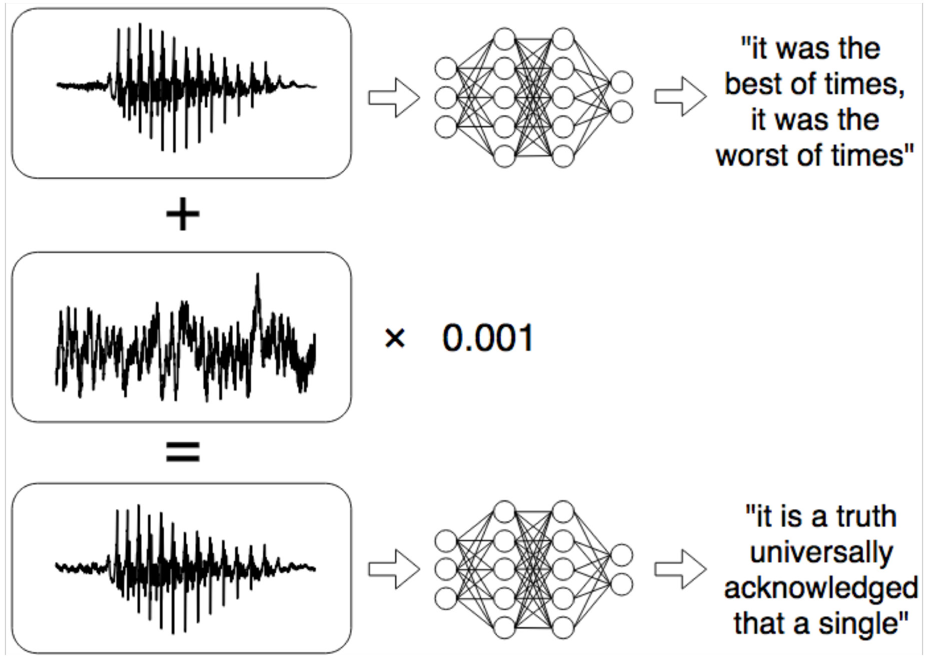
So: Brittleness of ML is a thing

Should we be worried?

Why Is This Brittleness of ML a Problem?

→ Security

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



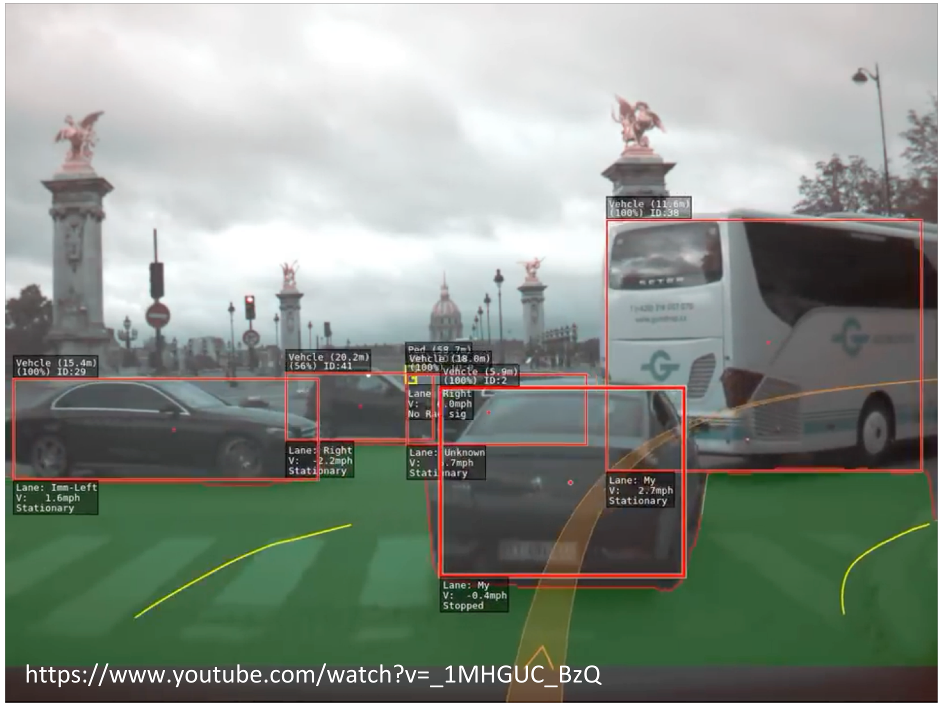
[Sharif Bhagavatula Bauer Reiter 2016]:
Glasses that fool face recognition

Why Is This Brittleness of ML a Problem?

→ Security

→ Safety

https://www.youtube.com/watch?v=TIUU1xNqI8w



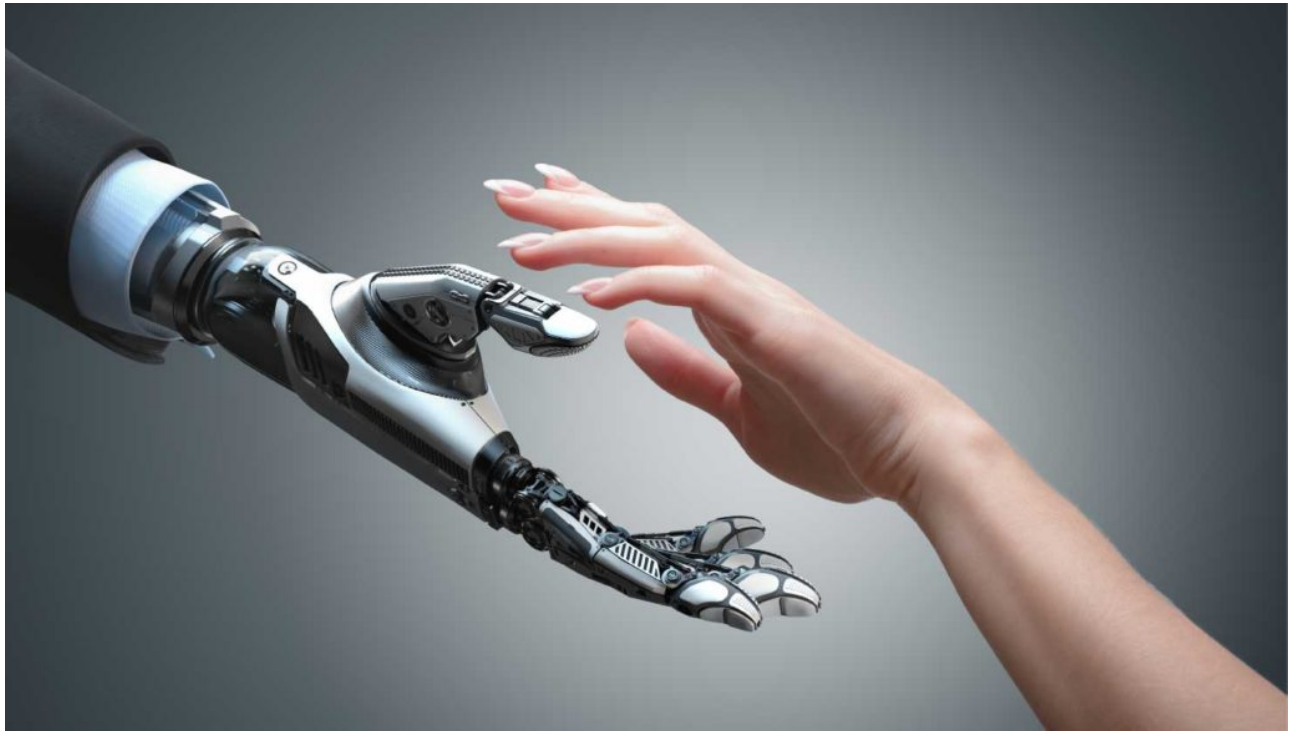
https://www.youtube.com/watch?v=_1MHGUC_BzQ

Why Is This Brittleness of ML a Problem?

→ Security

→ Safety

→ ML Alignment



Need to understand the
"failure modes" of ML

Is That It?

Training → Inference

⚠ Data poisoning

⚠ Adversarial Examples

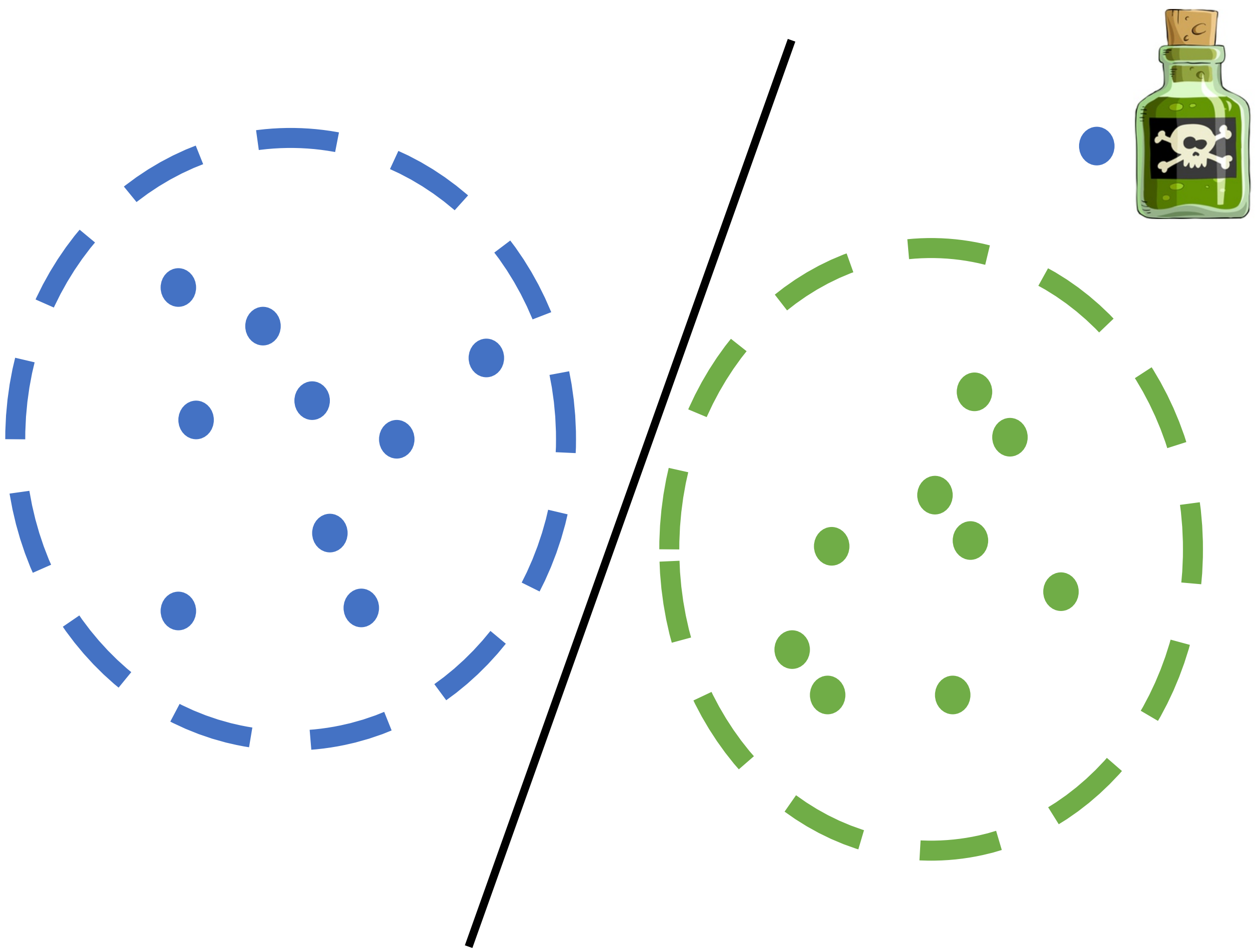
(Deep) ML is ”data hungry”

→ Can’t afford to be too picky about where we get the training data from

What can go wrong?

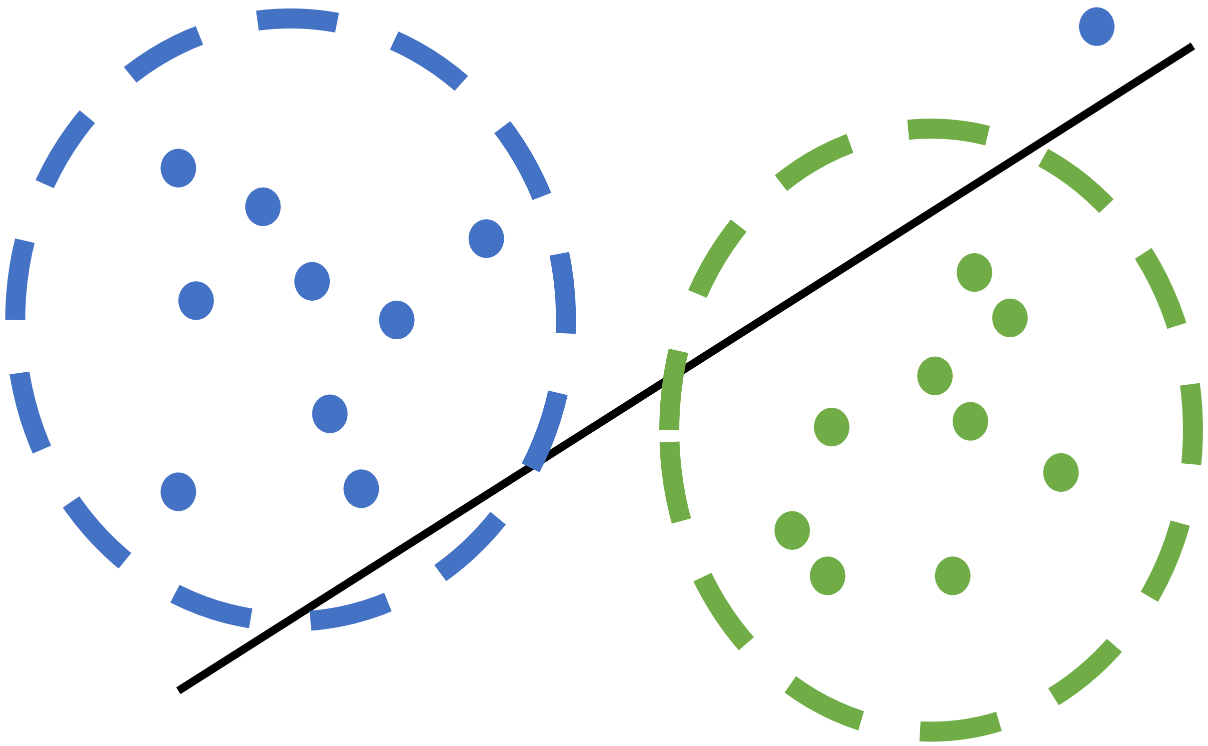
Data Poisoning

Goal: Maintain training accuracy but hamper generalization



Data Poisoning

Goal: Maintain training accuracy but hamper generalization



→ Fundamental problem in "classic" ML (robust statistics)

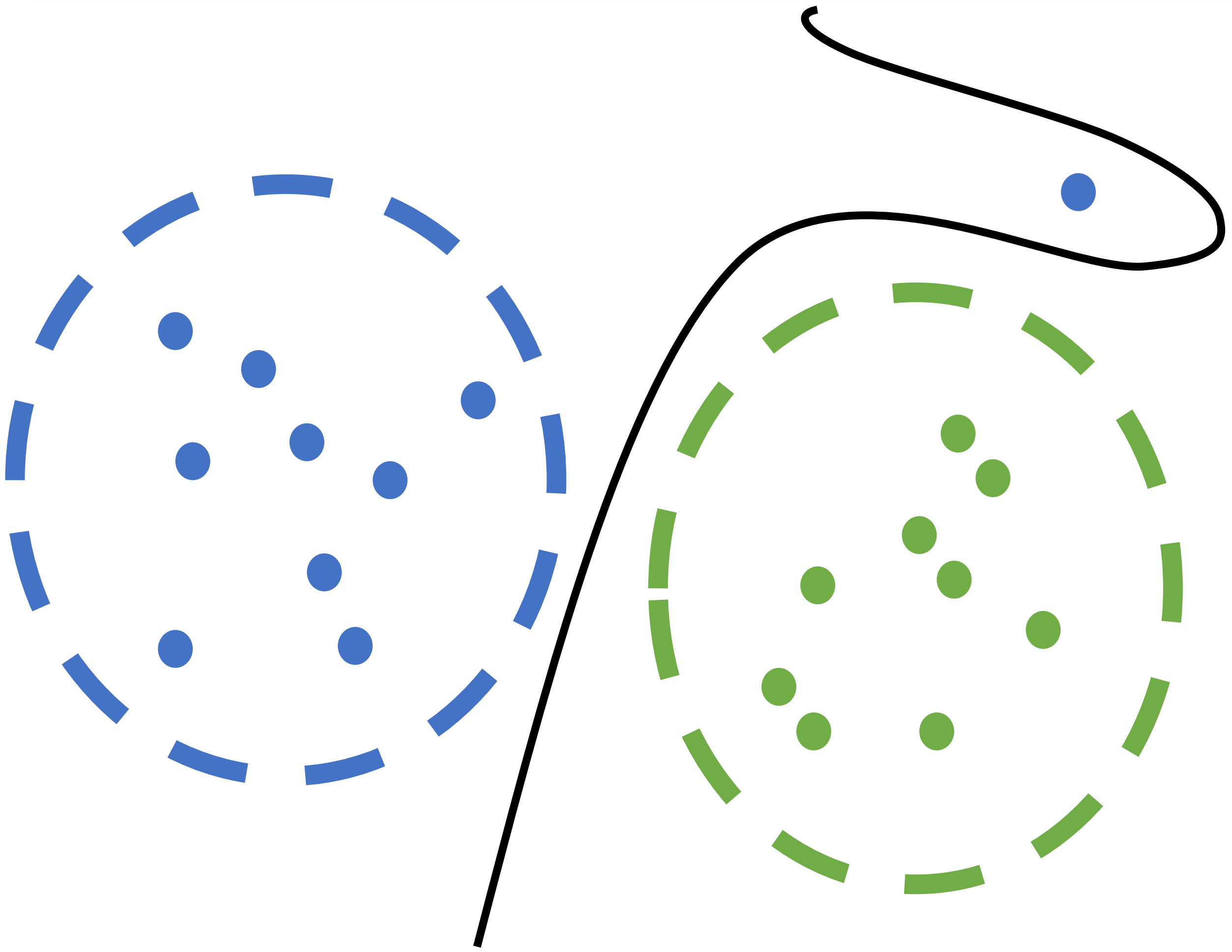
→ **But:** seems less so in deep learning

→ **Reason:** Memorization?

Data Poisoning

classification of **specific** inputs

Goal: Maintain training accuracy but hamper ~~generalization~~



→ Fundamental problem in "classic" ML (robust statistics)

→ **But:** seems less so in deep learning

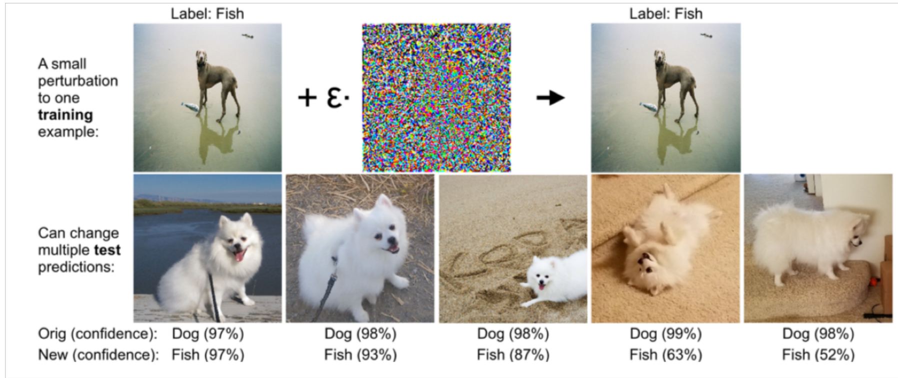
→ **Reason:** Memorization?

Is that it?

Data Poisoning

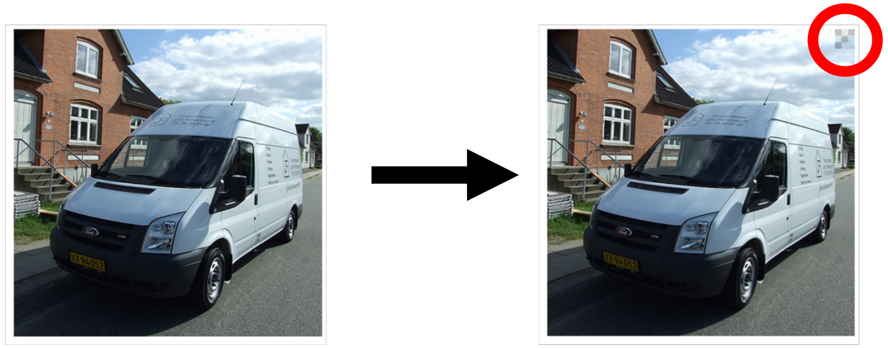
classification of **specific** inputs

Goal: Maintain training accuracy but hamper ~~generalization~~



[Koh Liang 2017]: Can manipulate **many** predictions with a **single** "poisoned" input

But: This gets (much) worse



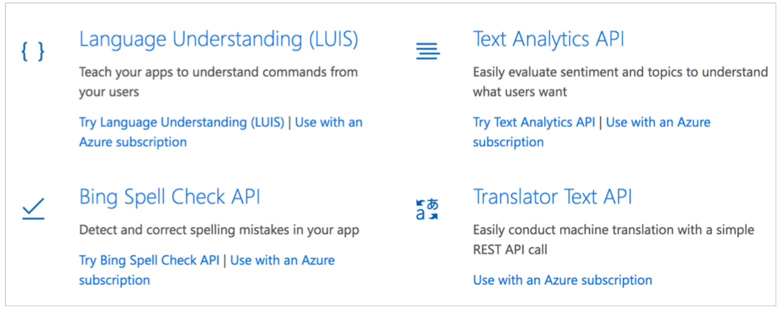
"van" "dog"

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

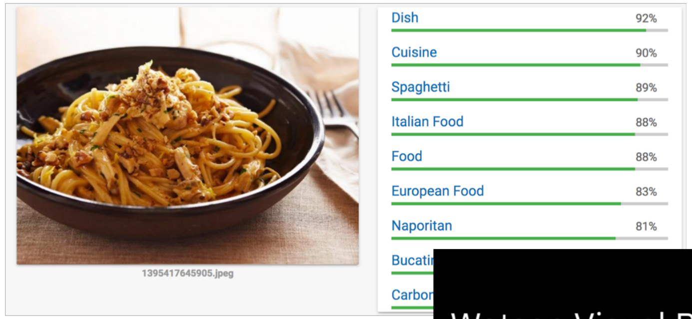
(**To learn more about backdoor attacks:** See poster #148 on Wed **[Tran Li M 2018]**)

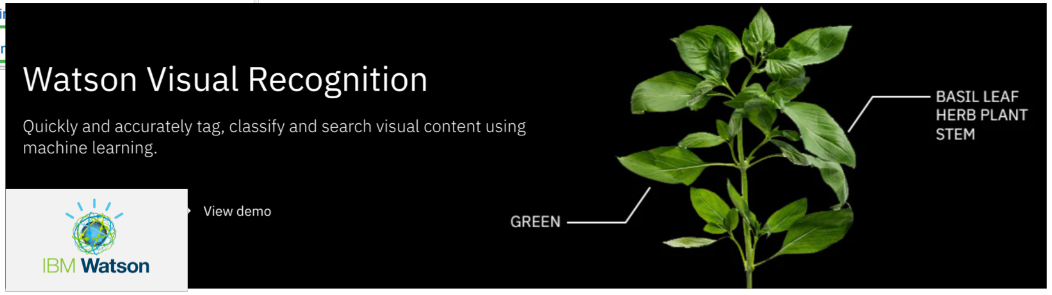
Is That It?

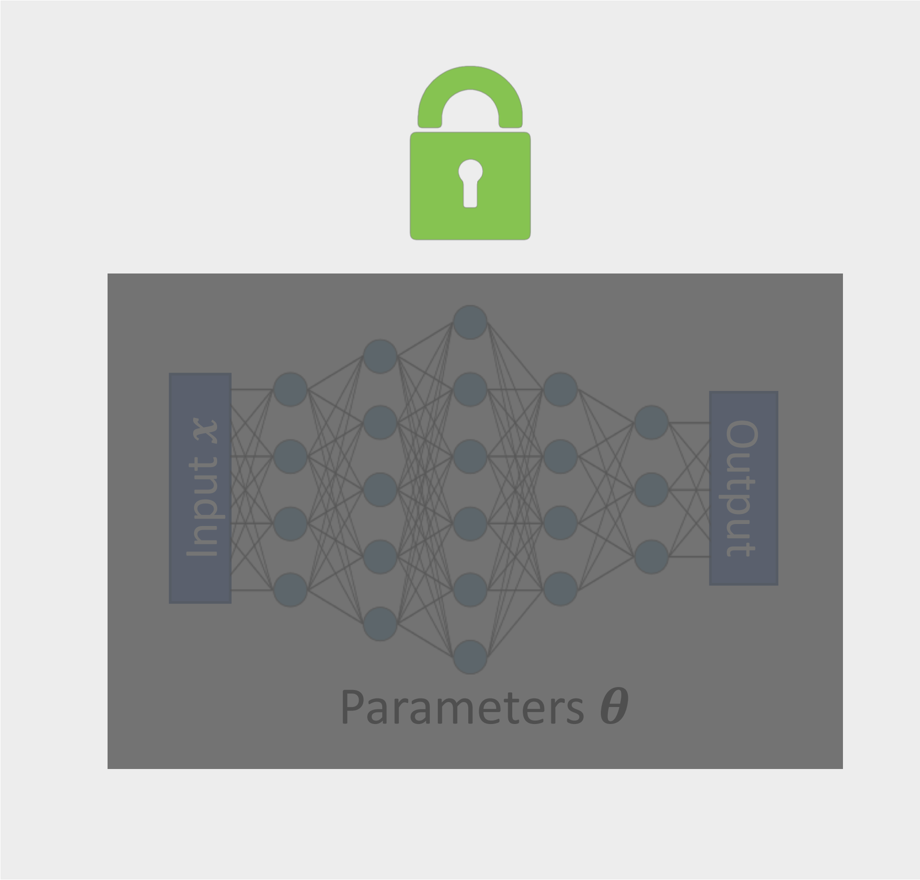
Microsoft Azure (Language Services)



Google Cloud Vision API





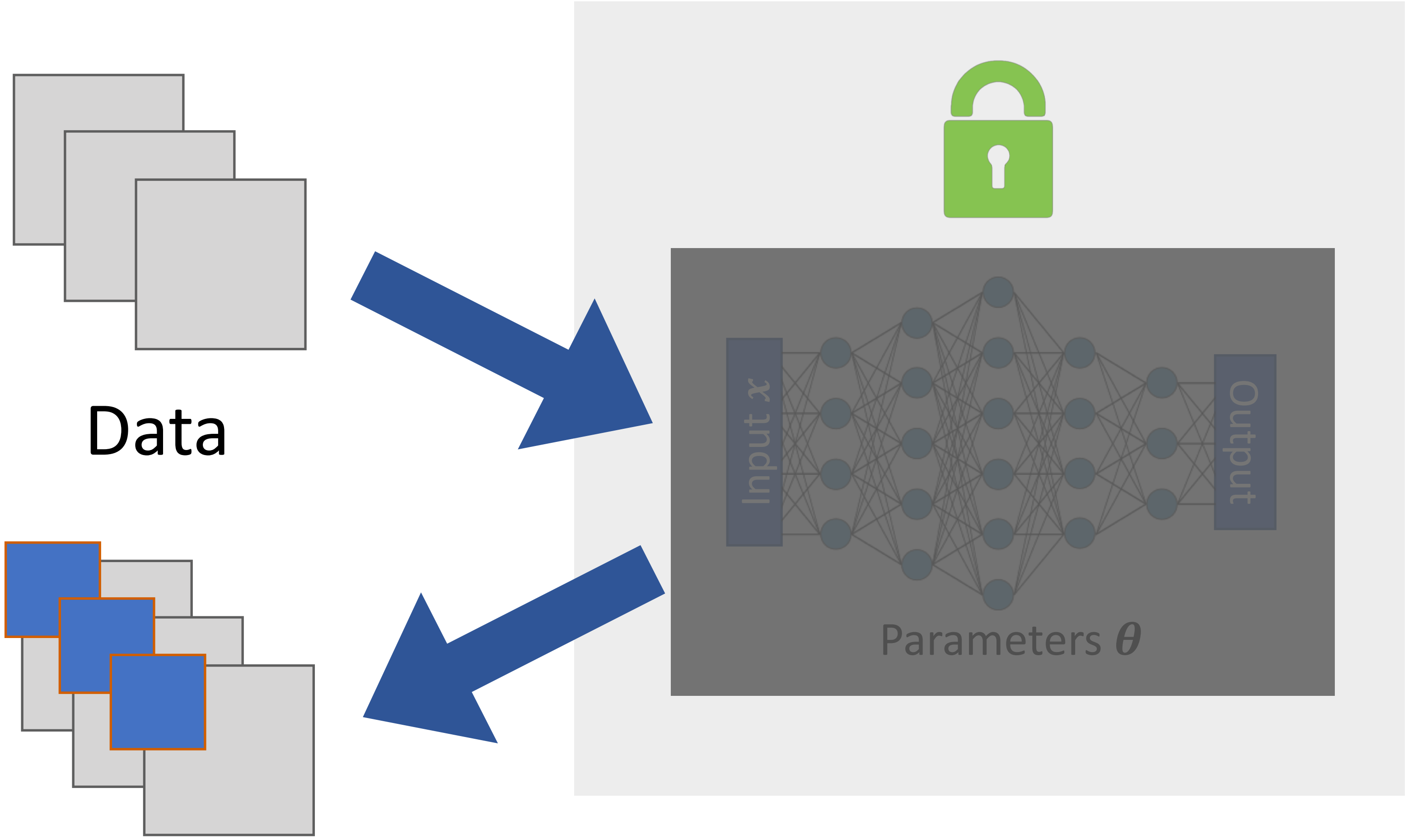


Training → Inference → Deployment

Is That It?

Does limited access give security?

In short: No



Predictions

Training → Inference → Deployment

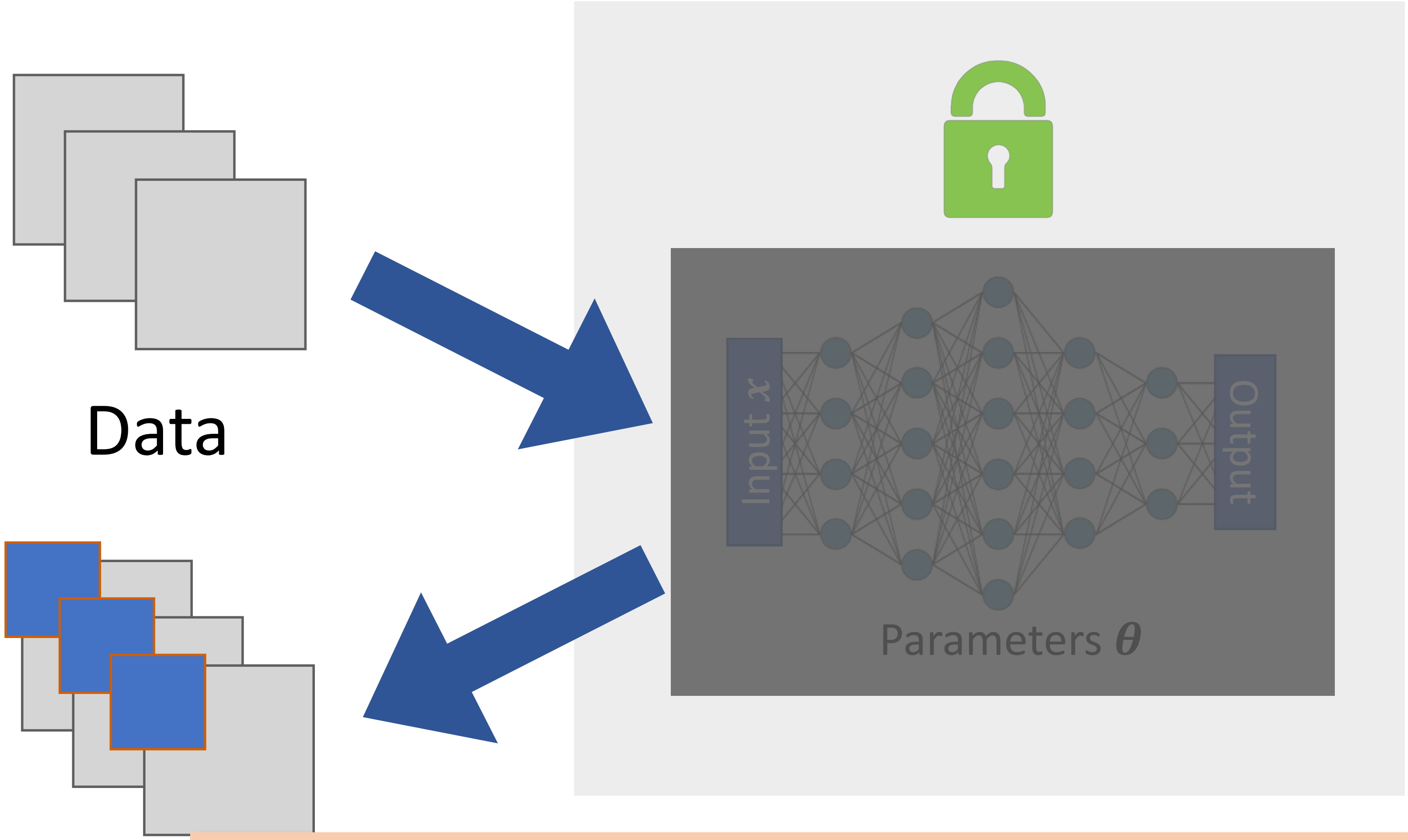
Black box attacks

Is That It?

Does limited access give security?

Model stealing: “Reverse engineer“ the model
[Tramer Zhang Juels Reiter Ristenpart 2016]

Black box attacks: Construct adv. examples from queries
[Chen Zhang Sharma Yi Hsieh 2017][Bhagoji He Li Song 2017][Ilyas Engstrom Athalye Lin 2017][Brendel Rauber Bethge 2017][Cheng Le Chen Yi Zhang Hsieh 2018][Ilyas Engstrom M 2018]



Predict

For more: See my talk on Friday

Training → Inference → Deployment

Black box attacks

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?

(Is ML inherently not reliable?)

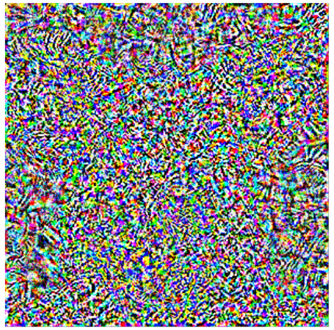
No: But we need to re-think how we do ML

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

Towards Adversarially Robust Models

"pig" (91%)

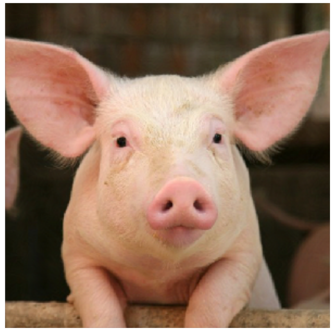
+ 0.005 x



=

"pig"

~~"airliner"~~ (99%)



Where Do Adversarial Examples Come From?

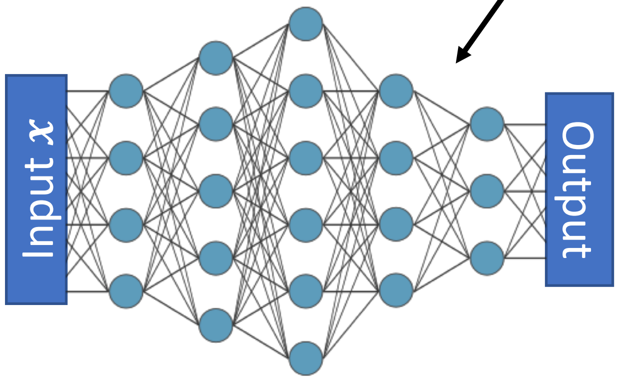
To get an adv. example

~~**Goal of training:**~~

Model Parameters Input Correct Label

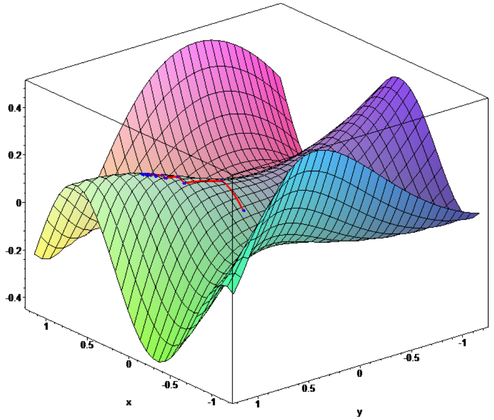
$$min_{\theta} \ loss(\theta, x, y)$$

Differentiable



Parameters $\boldsymbol{\theta}$

Can use gradient descent method to find good θ



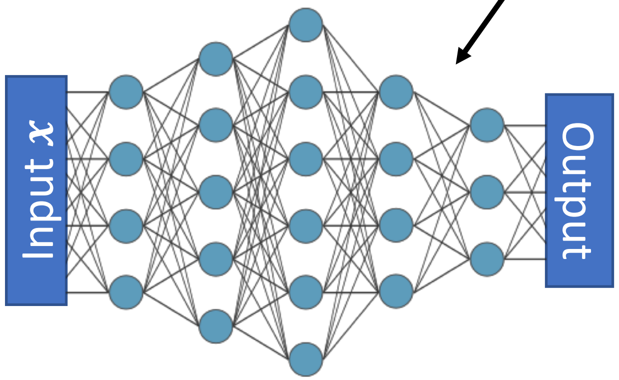
Where Do Adversarial Examples Come From?

To get an adv. example

~~**Goal of training:**~~

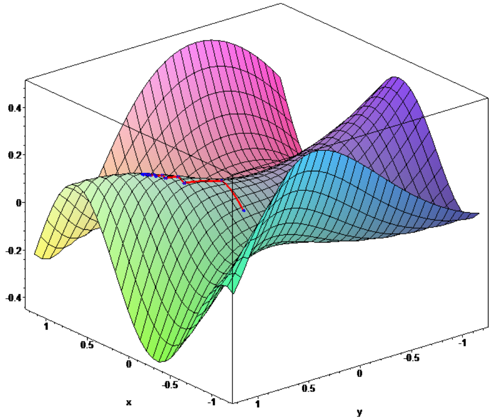
$$loss(\theta, x + \delta, y)$$

Differentiable



Parameters $\boldsymbol{\theta}$

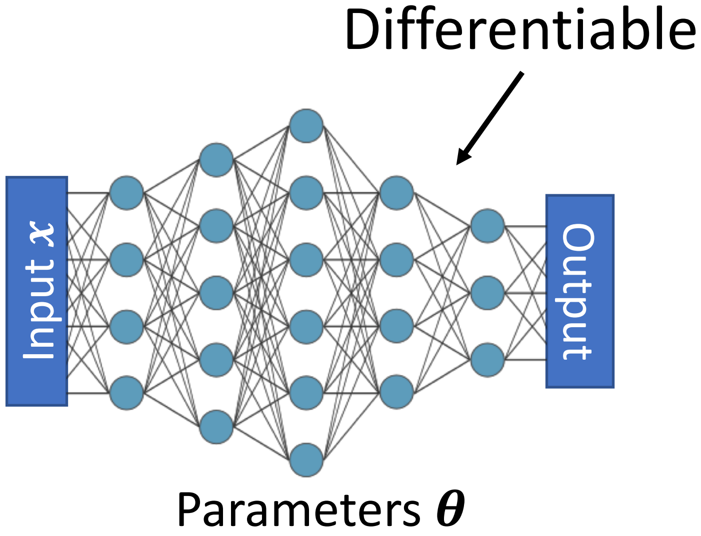
Can use gradient descent method to find good θ



Where Do Adversarial Examples Come From?

To get an adv. example

~~**Goal of training:**~~



$$max_{\delta}\ loss(\theta, x + \delta, y)$$

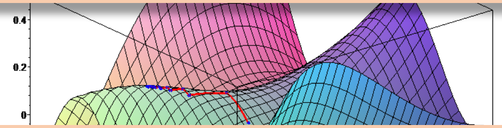
Which δ are allowed?

Can use gradient descent

Examples: δ that is small wrt

- ℓ_p-norm
- Rotation and/or translation
- VGG feature perturbation
- (add the perturbation you need here)

This is an important question (that we put aside)



Still: We have to confront (small) ℓ_p-norm perturbations

Towards ML Models that Are Adv. Robust

[M Makelov Schmidt Tsipras Vladu 2018]

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

~~Standard~~ generalization: $\mathbb{E}_{(x,y)\sim D}\,[loss(\theta, x, y)]$

Adversarially robust

But: Adversarial noise is a "needle in a haystack"

Towards ML Models that Are Adv. Robust

[M Makelov Schmidt Tsipras Vladu 2018]

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

~~Standard~~ generalization: $\mathbb{E}_{(x,y)\sim D}\left[\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} loss(\theta, x+\boldsymbol{\delta}, y)\right]$

Adversarially robust

But: Adversarial noise is a "needle in a haystack"

Next: A deeper dive into the topic

→ Adversarial examples and verification (Zico)

→ Training adversarially robust models (Zico)

→ Adversarial robustness beyond security (Aleksander)

Adversarial Robustness Beyond Security

ML via Adversarial Robustness Lens

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

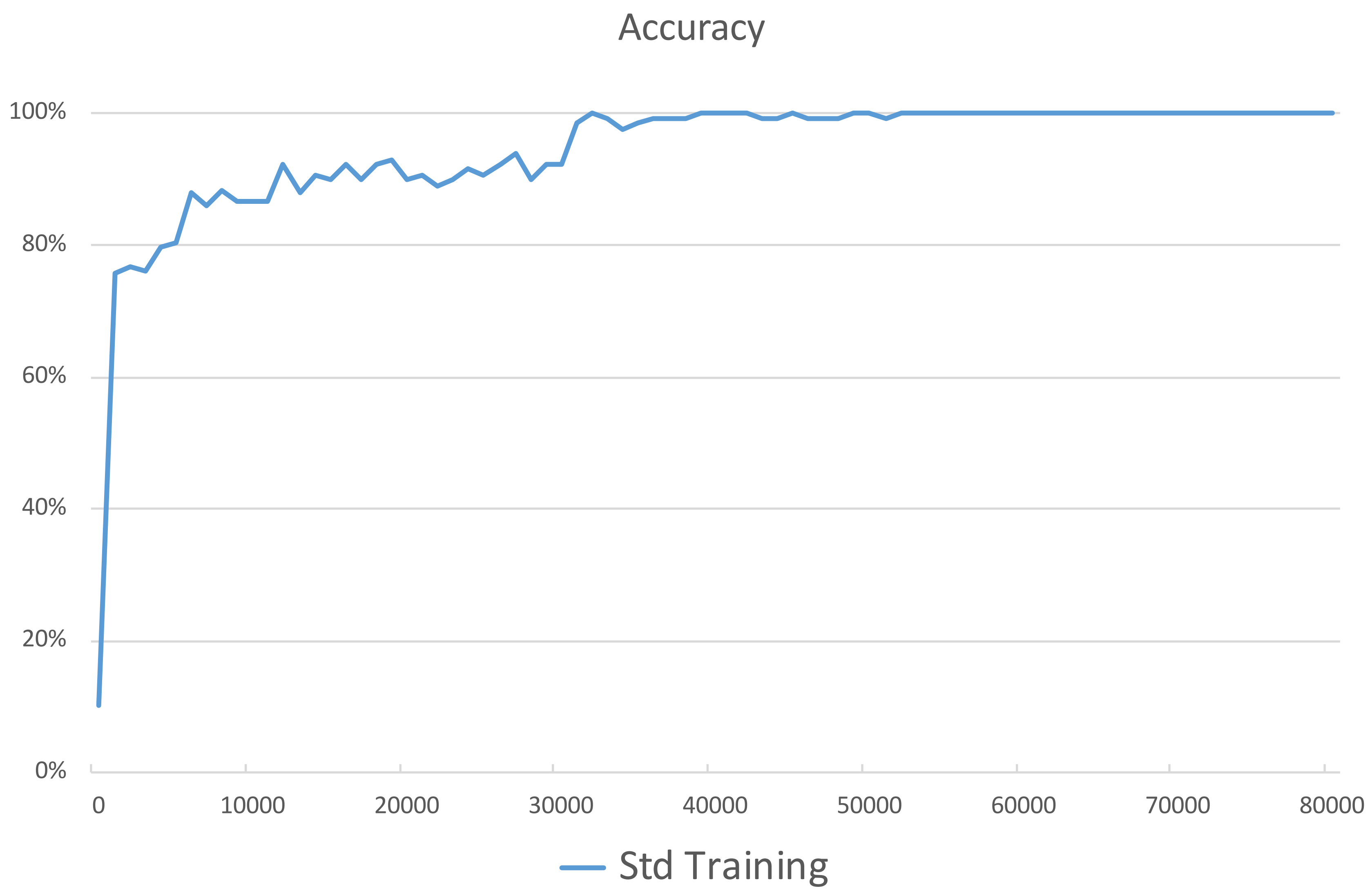
$$\mathbb{E}_{(x,y)\sim D}\left[loss(\theta, x, y)\right]$$

vs

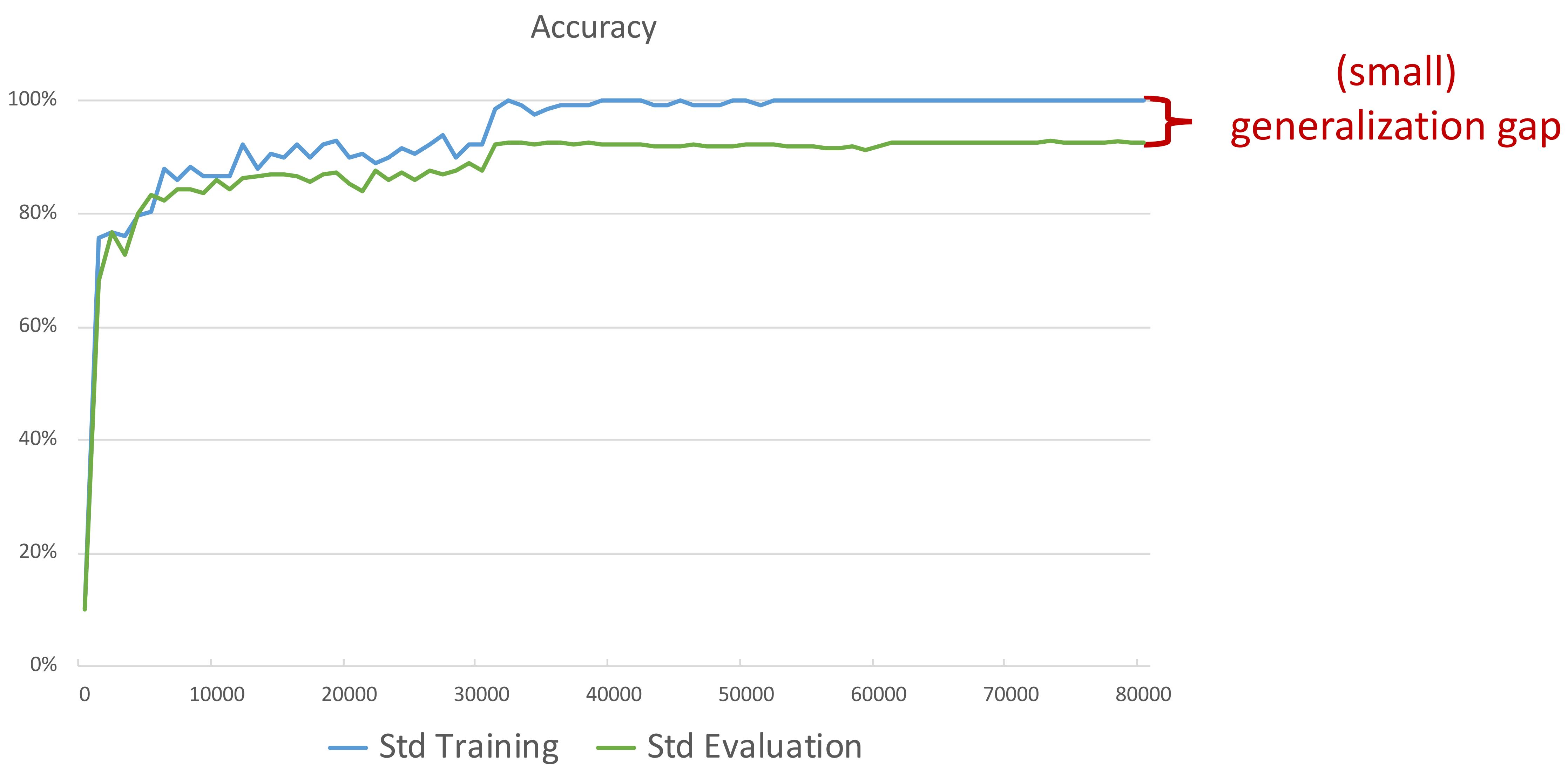
$$\mathbb{E}_{(x,y)\sim D}\left[\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} loss(\theta, x+\boldsymbol{\delta}, y)\right]$$

(This goes **beyond** deep learning)

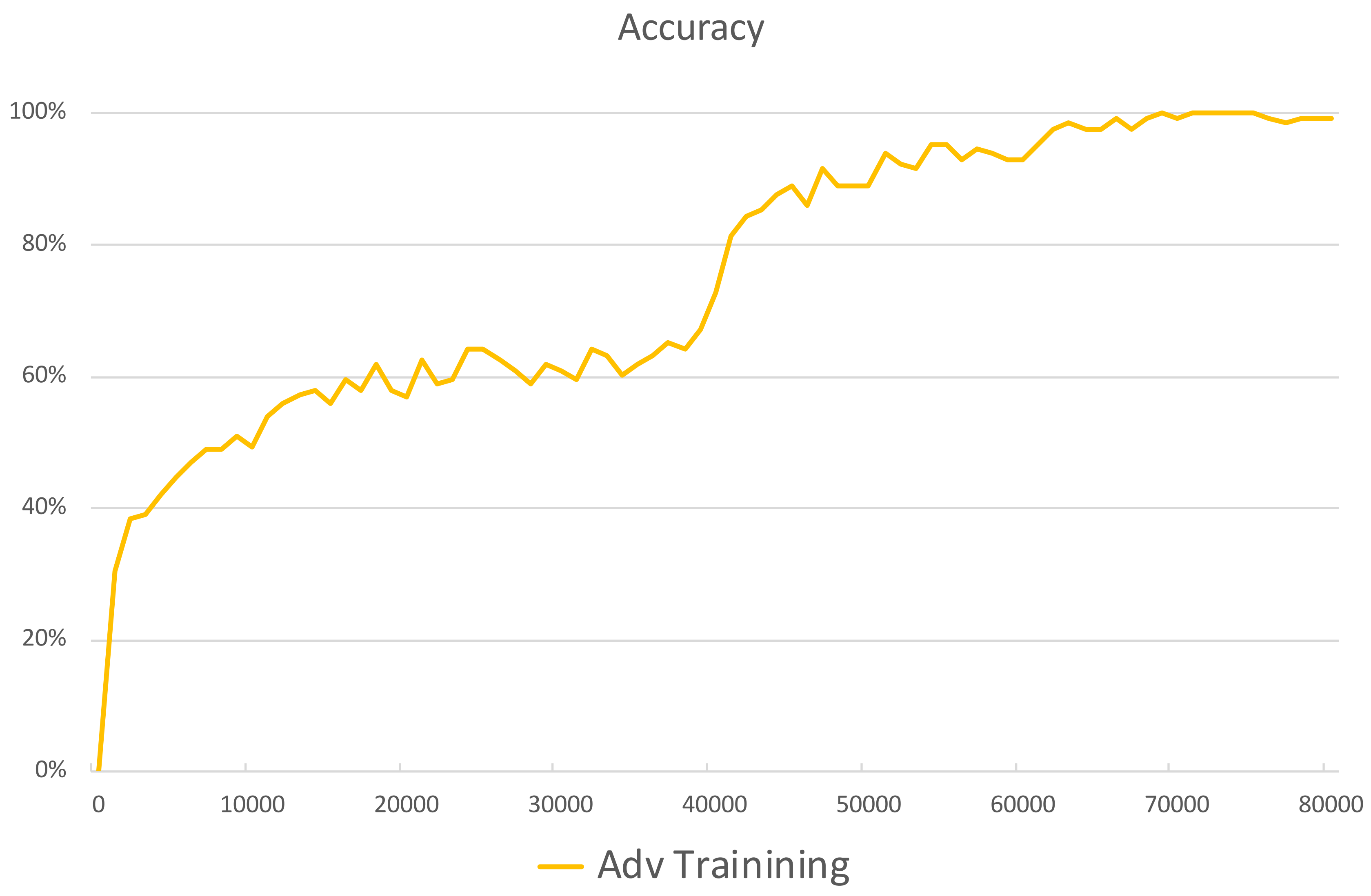
Do Robust Deep Networks Overfit?



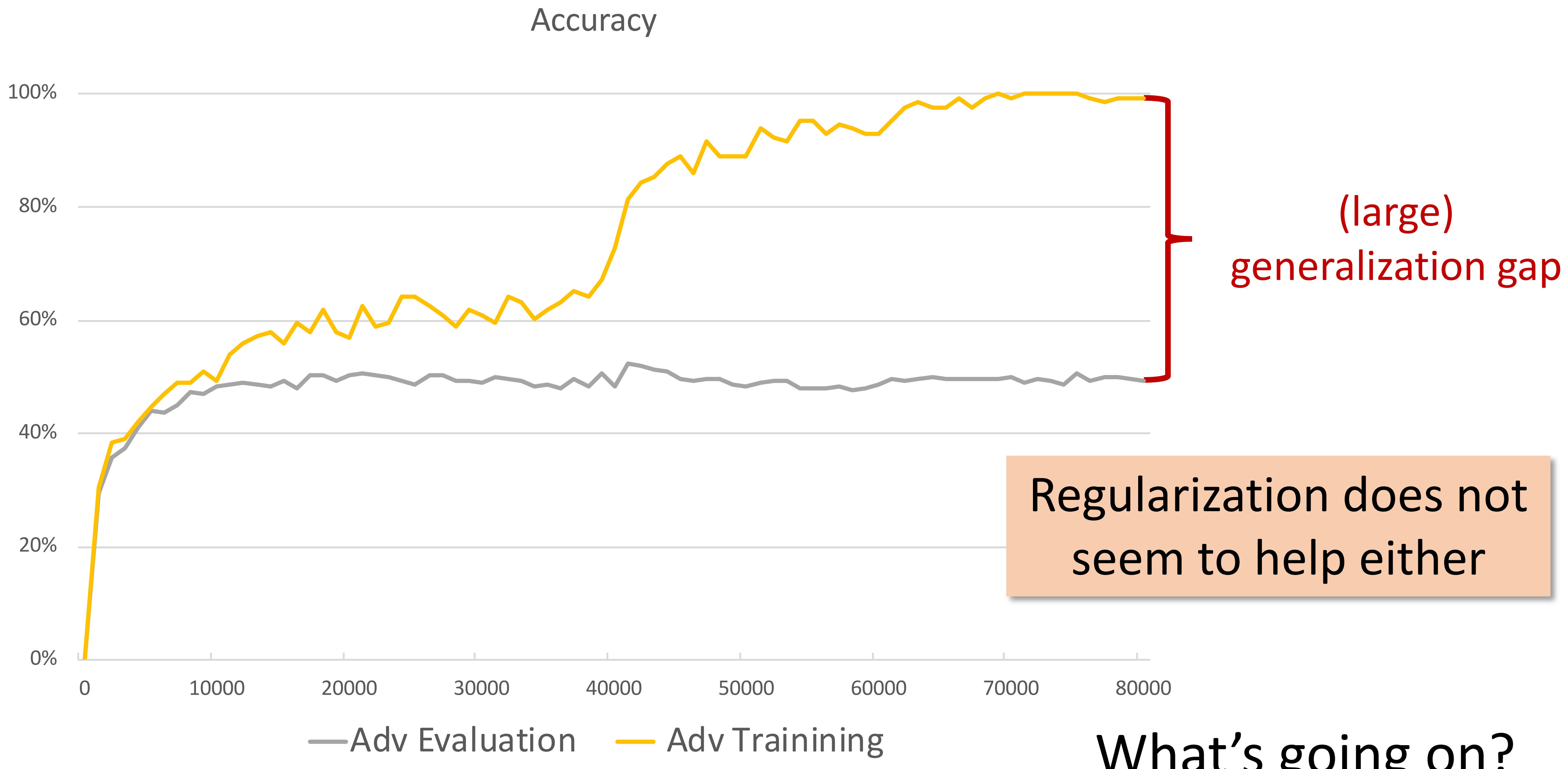
Do Robust Deep Networks Overfit?



Do Robust Deep Networks Overfit?



Do Robust Deep Networks Overfit?



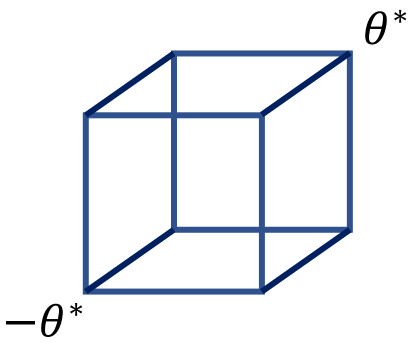
Regularization does not seem to help either

What's going on?

Adv. Robust Generalization Needs More Data

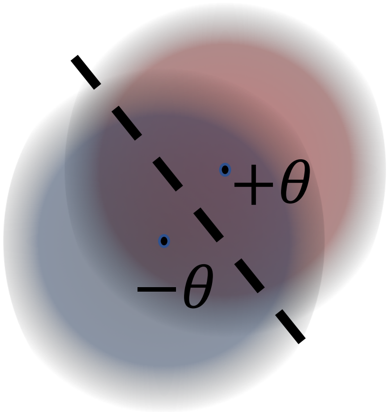
Theorem [Schmidt Santurkar Tsipras Talwar M 2018]:
Sample complexity of adv. robust generalization can be **significantly larger** than that of "standard" generalization

Specifically: There exists a **d**-dimensional distribution **D** s.t.:



→ A **single** sample is enough to get an **accurate** classifier (P[correct] > 0.99)

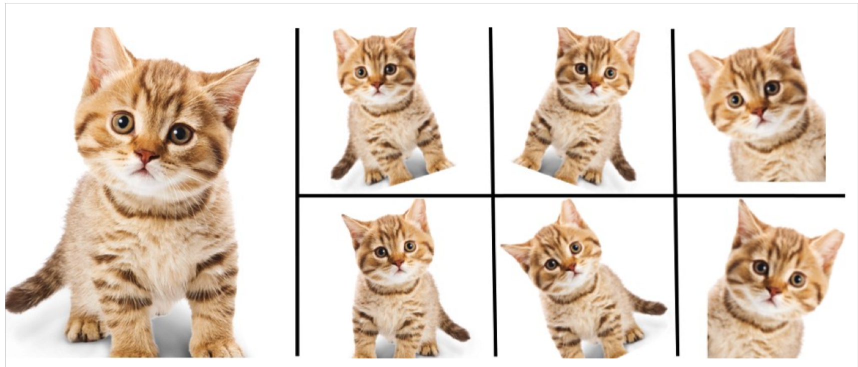
→ **But**: Need $\mathbf{\Omega}\left(\sqrt{\mathbf{d}}\right)$ samples for better-than-chance **robust** classifier



(**More details:** See spotlight + poster #31 on Tue)

Does Being Robust Help "Standard" Generalization?

Data augmentation: An effective technique to improve "standard" generalization



Adversarial training

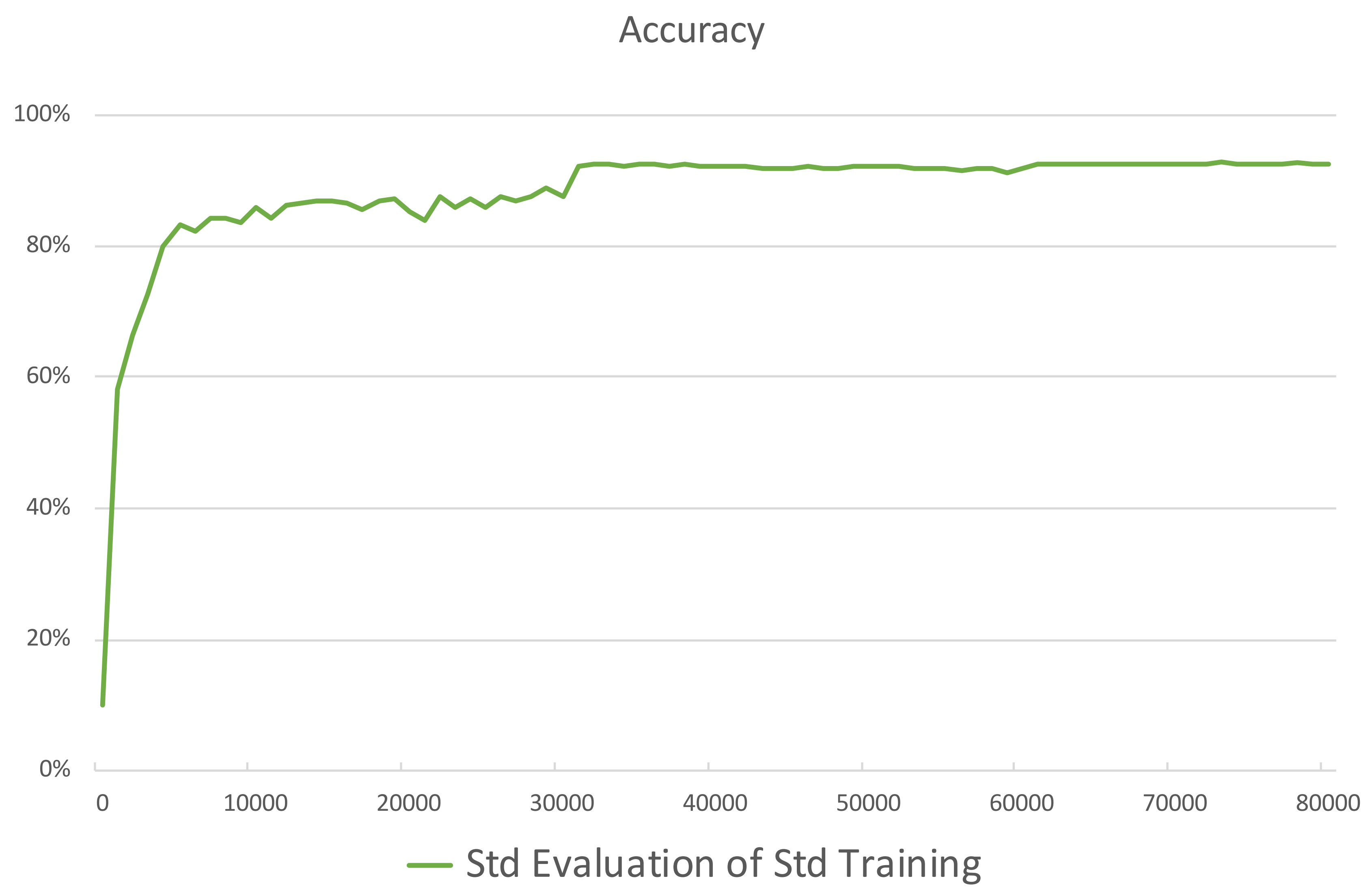
=

An "ultimate" version of data augmentation?

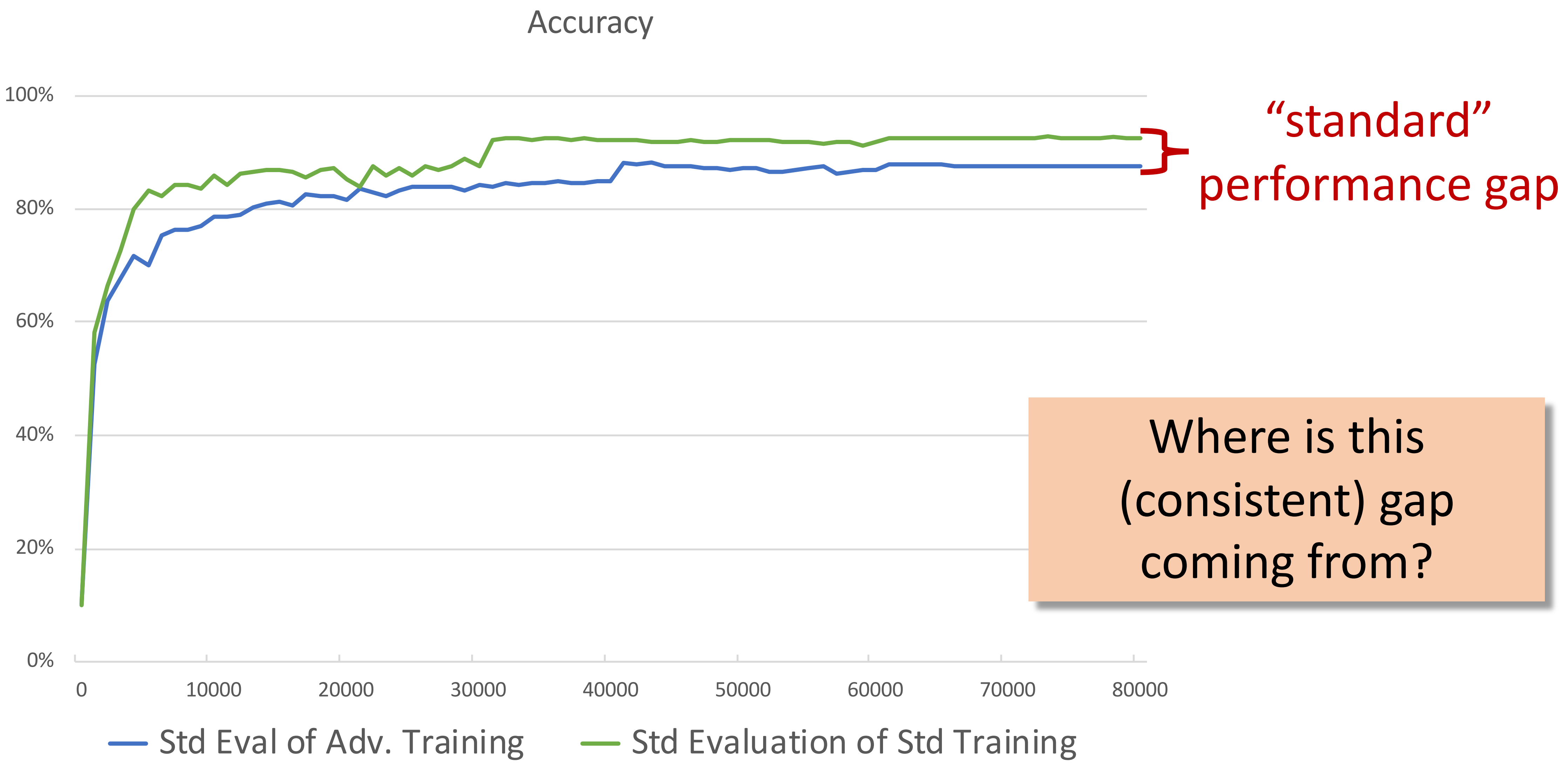
(since we train on the "most confusing" version of the training set)

Does adversarial training always improve "standard" generalization?

Does Being Robust Help “Standard” Generalization?



Does Being Robust Help "Standard" Generalization?

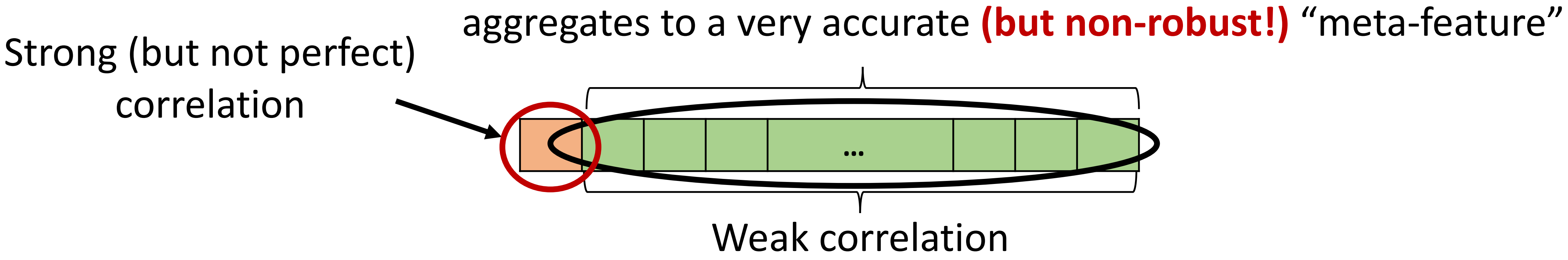


Where is this (consistent) gap coming from?

Does Being Robust Help “Standard” Generalization?

Theorem [Tsipras Santurkar Engstrom Turner M 2018]:
No “free lunch”: can exist a trade-off between accuracy and robustness

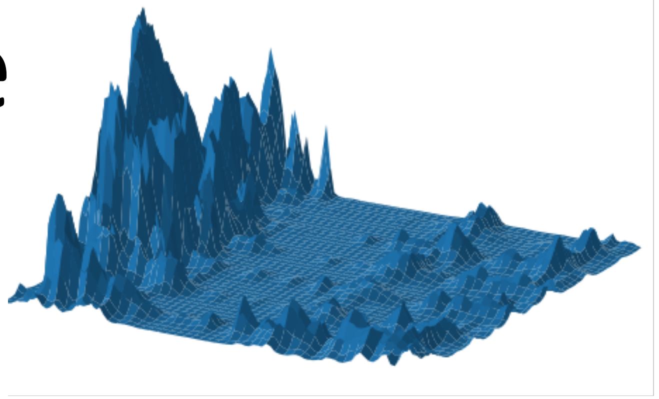
Basic intuition:
→ In standard training, **all correlation is good correlation**
→ If we want robustness, **must avoid** weakly correlated features



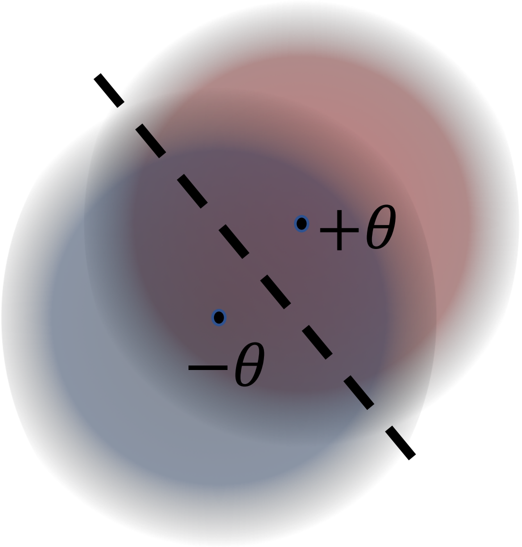
Standard training: use all of features, maximize accuracy

Adversarial training: use only single robust feature (**at the expense of accuracy**)

Adversarial Robustness is Not Free

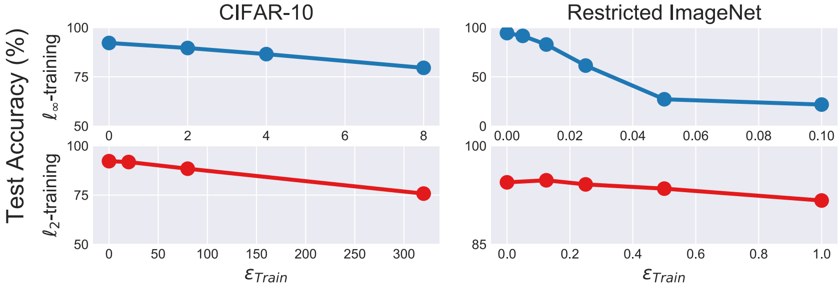


→ Optimization during training more difficult and models need to be larger



→ More training data might be required

[Schmidt Santurkar Tsipras Talwar M 2018]



→ Might need to lose on “standard” measures of performance

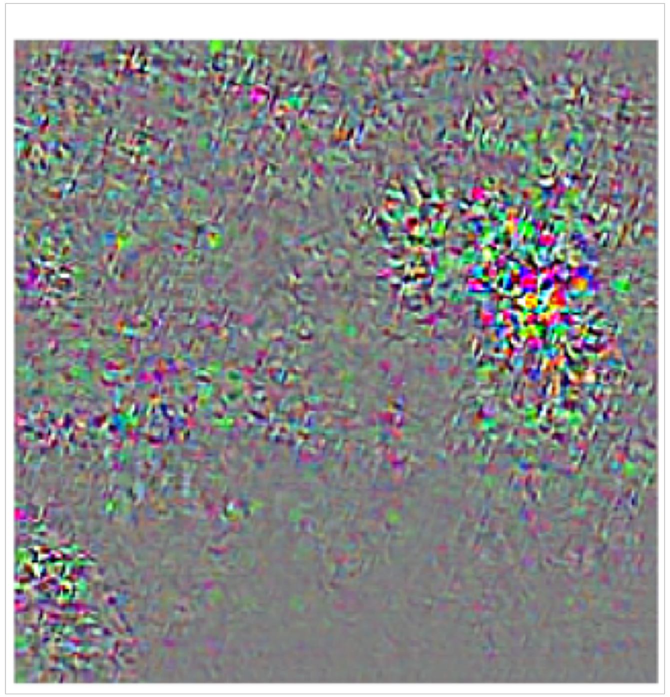
[Tsipras Santurkar Engstrom Turner M 2018] (Also see: **[Bubeck Price Razenshteyn 2018]**)

But There Are (Unexpected?) Benefits Too

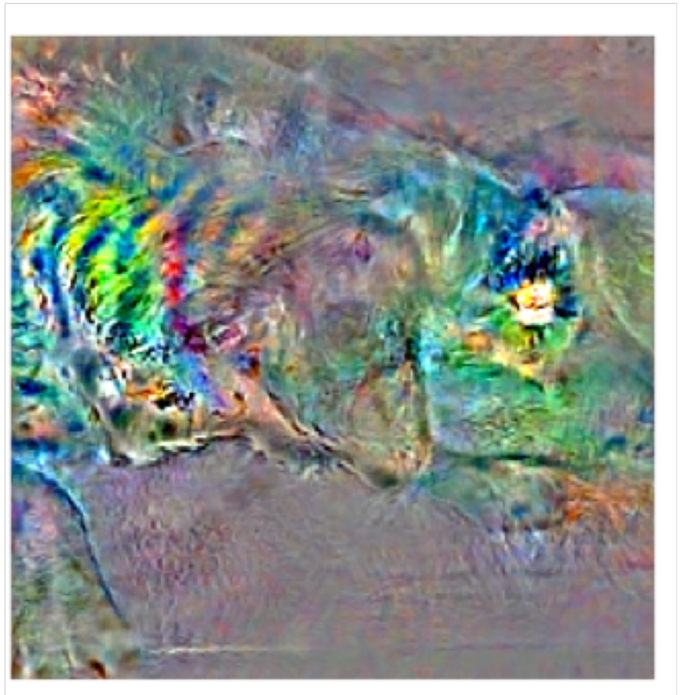
[Tsipras Santurkar Engstrom Turner M 2018]

Models become more semantically meaningful

Input



Gradient of standard model



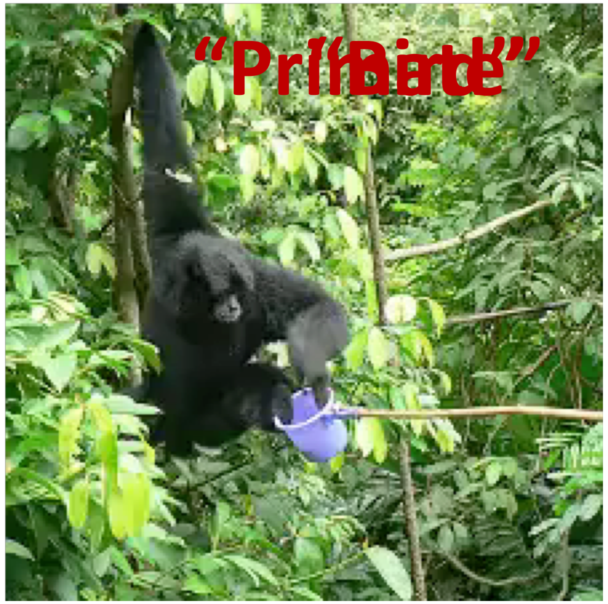
Gradient of adv. robust model

But There Are (Unexpected?) Benefits Too

[Tsipras Santurkar Engstrom Turner M 2018]

Models become more **semantically meaningful**

Standard model



Adv. robust model

[Brock Donahue Simonyan 2018]
+ [Isola 2018]

Robust models → (restricted) GAN-like embeddings?

Conclusions

Towards (Adversarially) Robust ML

→ **Algorithms:** Faster robust training + verification [Xiao Tjeng Shafiullah **M** 2018], smaller models, new architectures?

→ **Theory:** (Better) adv. robust generalization bounds, new regularization techniques

→ **Data:** New datasets and more comprehensive set of perturbations

Major need: Embracing more of a worst-case mindset

→ **Adaptive** evaluation methodology + scaling up verification

cleverhans

RobustML

(robust-ml.org)

More Broadly

Next frontier:

Building ML one can truly rely on

→ Will lead to ML that is not only safe/secure but also “better”?

Further reading:

→ **Notes + code:** adversarial-ml-tutorial.org **(work in progress)**

→ **Blog posts:** gradient-science.org

@aleks_madry **@zicokolter** madry-lab.ml