Machine Learning: A Security Perspective

Aleksander Mądry

madry-lab.ml

Deep
~~Machine~~ Learning:
A Security Perspective

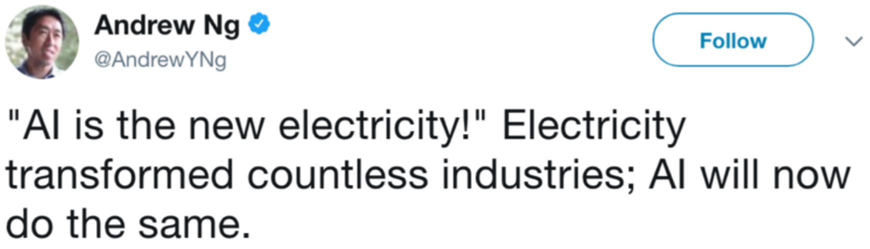
Aleksander Mądry

madry-lab.ml

Deep Learning

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

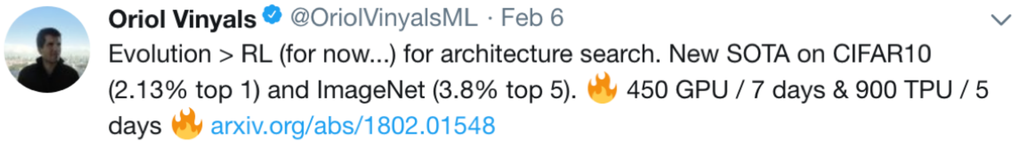


Andrew Ng @AndrewYNg

Follow

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

The GANfather: The man who's given machines the gift of imagination - MIT Technology Review technologyreview.com/s/610253/the-g …



Oriol Vinyals @OriolVinyalsML · Feb 6

Evolution > RL (for now...) for architecture search. New SOTA on CIFAR10 (2.13% top 1) and ImageNet (3.8% top 5). 🔥 450 GPU / 7 days & 900 TPU / 5 days 🔥 arxiv.org/abs/1802.01548

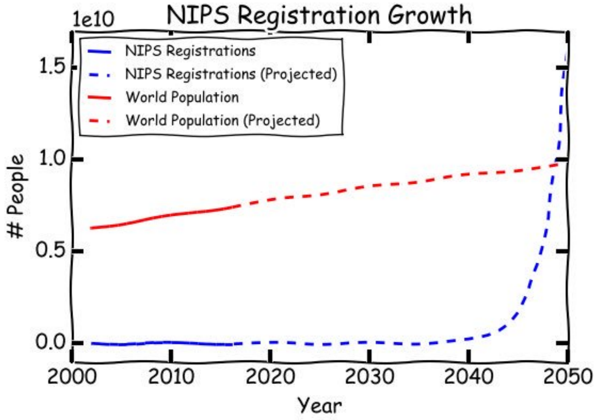
Ben Recht @beenwrekt · Jan 18

My tweeting has fallen off because I have nothing pithy to say about quantum AI on the blockchain.

2016: The Year That Deep Learning Took Over the Internet

TOM SIMONITE BUSINESS 11.01.17 07:00 AM

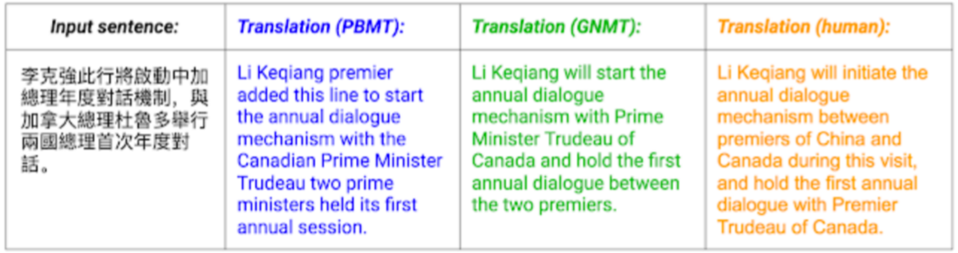
GOOGLE'S AI WIZARD UNVEILS A NEW TWIST ON NEURAL NETWORKS



Deep Learning: The Success Stories

Image classification

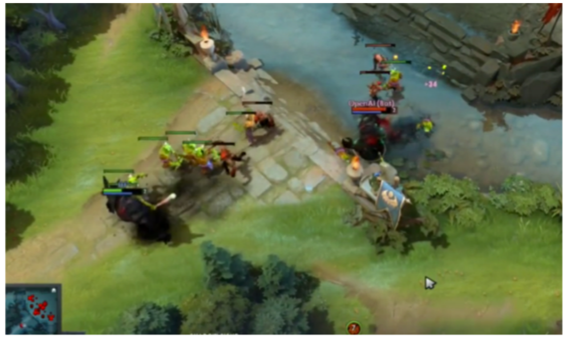
Generating realistic high-resolution images

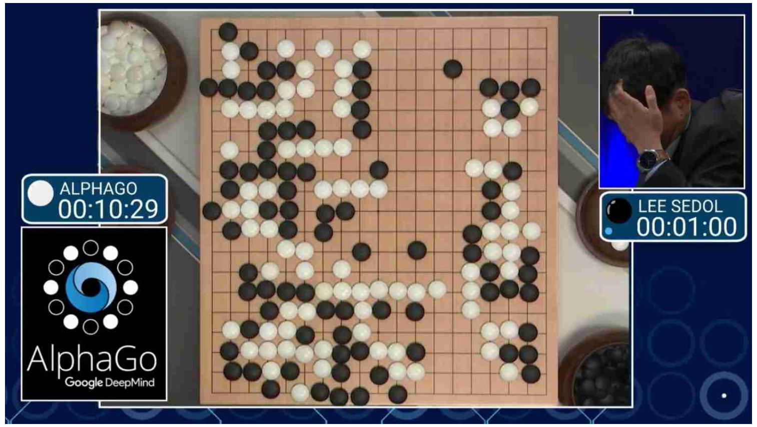


Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will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

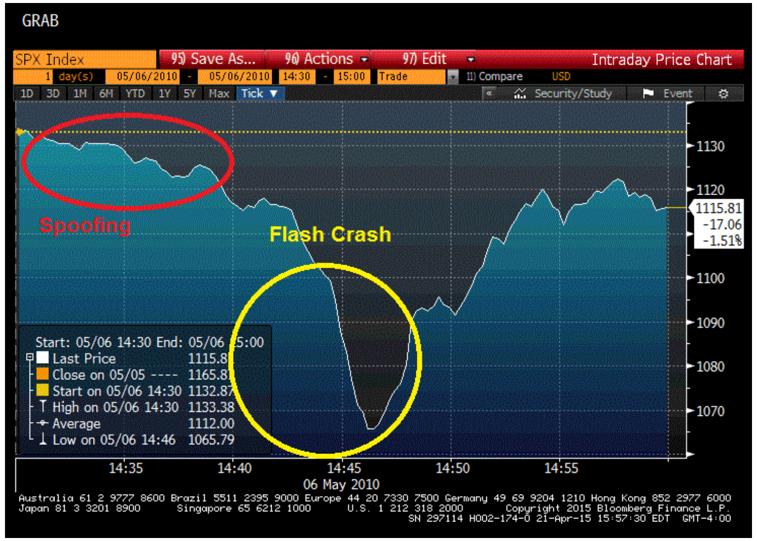
Game playing





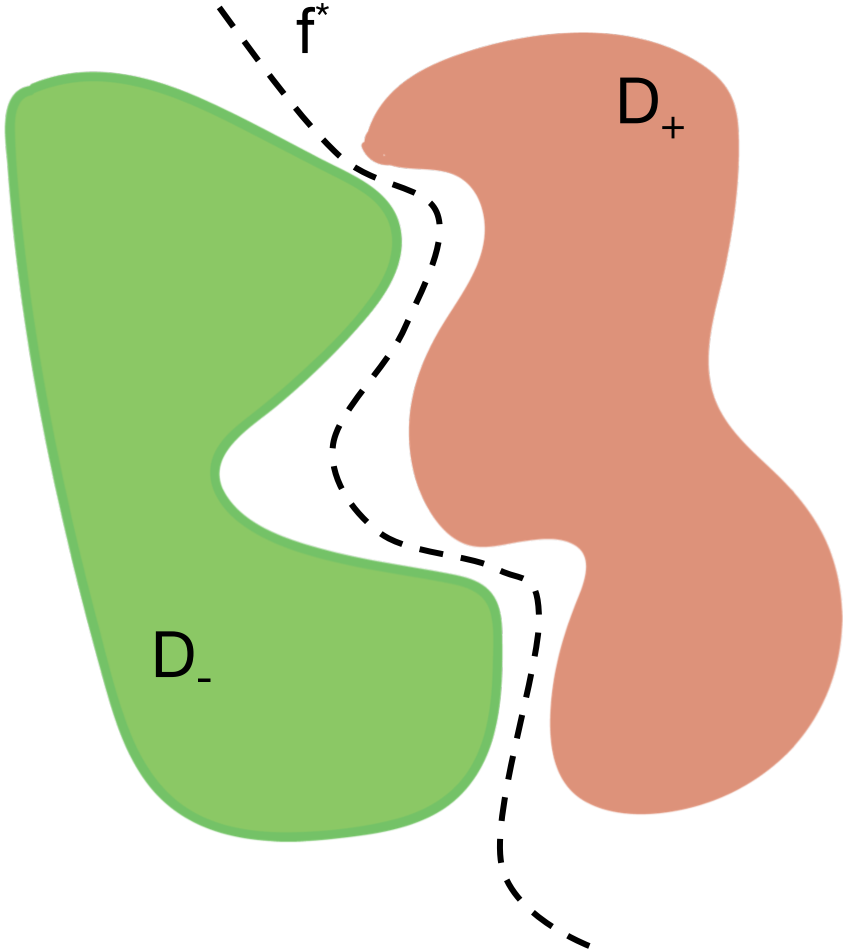
Things are great,
so what’s the problem?

Can We Truly Rely on ML?



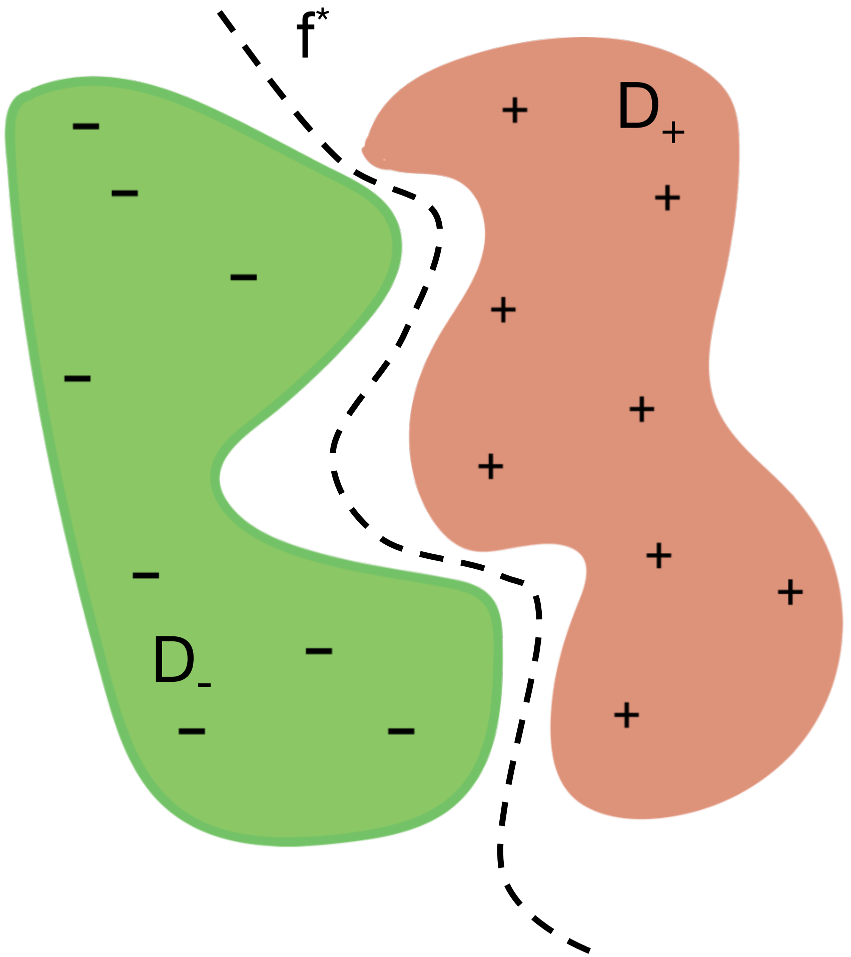
(Supervised) Machine Learning: A Quick Primer

Supervised Machine Learning



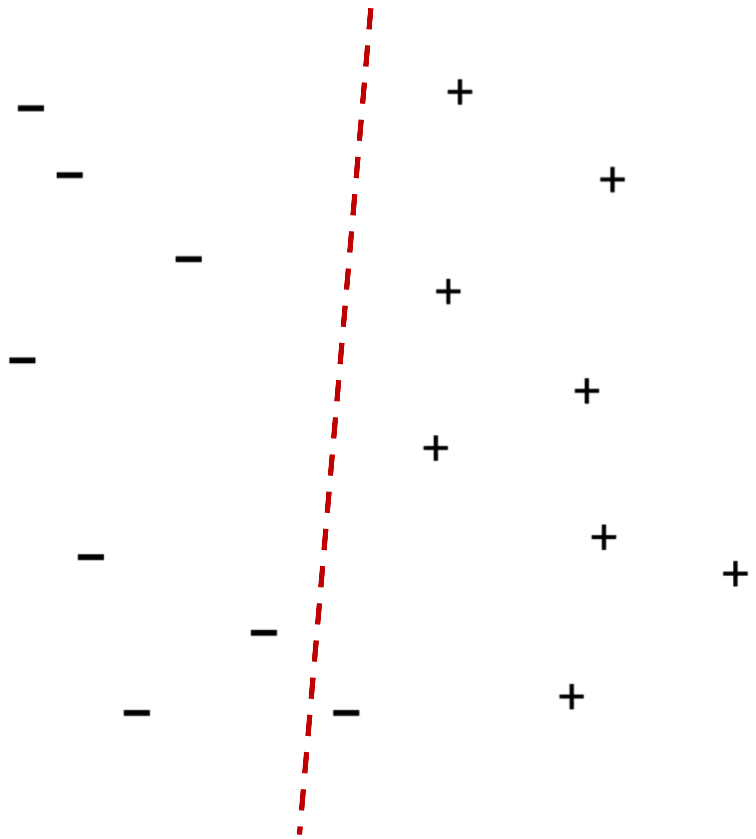
f^*= concept to learn

Supervised Machine Learning



f^*= concept to learn

Supervised Machine Learning



f^*= concept to learn

Training: Find parameters $\boldsymbol{\theta}^*$ that make our classifier $\mathbf{f}(\boldsymbol{\theta}^*)$ fit/”explain” the training data (and thus approx. $\mathbf{f}^*$)

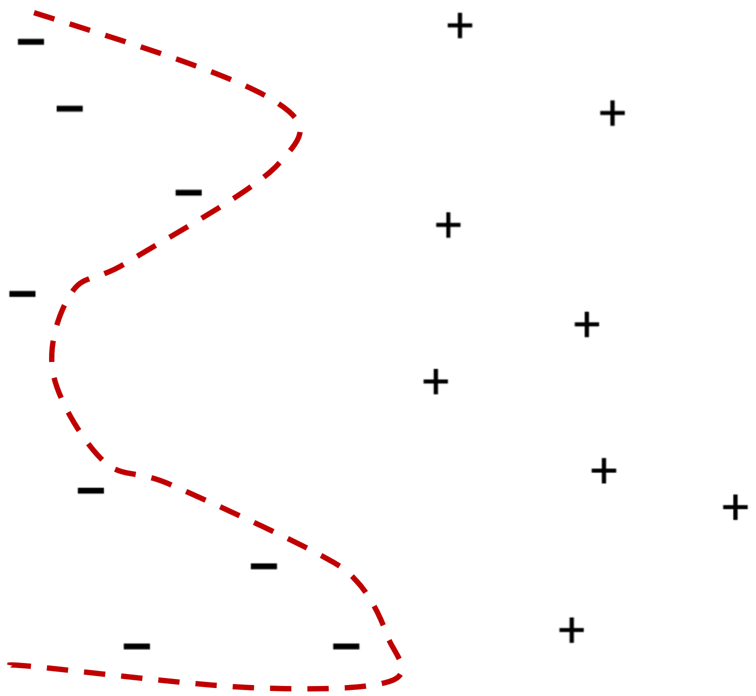
Here: f(θ) = a family of classifiers parametrized by **θ**

Choice of the family f(·) is crucial

Too simple → underfitting

Supervised Machine Learning

f^*= concept to learn



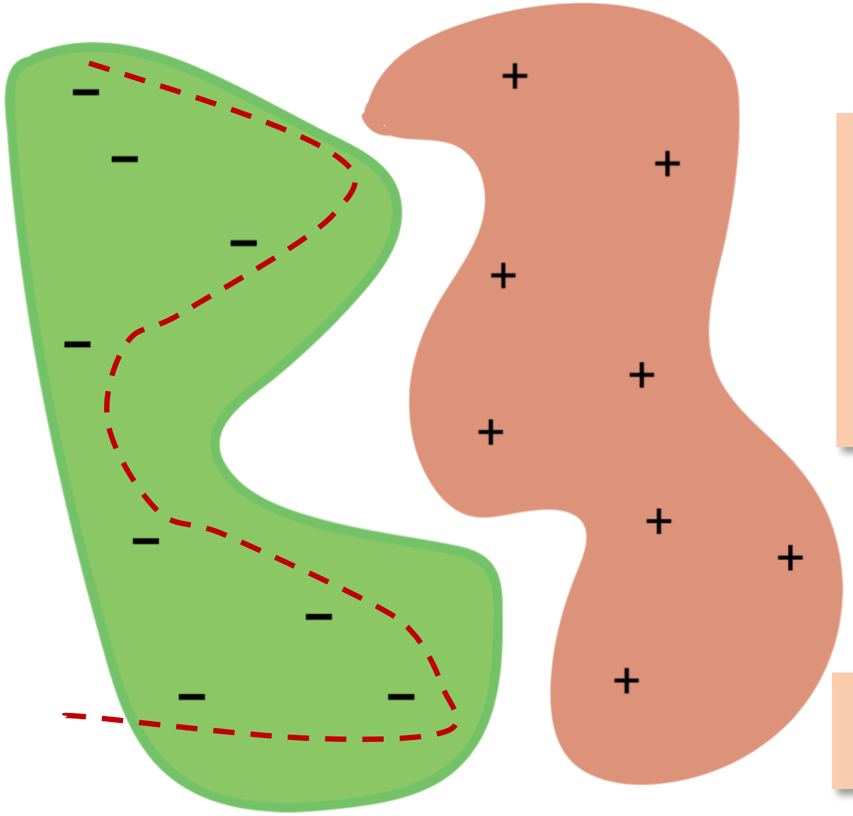
Training: Find parameters $\boldsymbol{\theta}^*$ that make our classifier $\mathbf{f}(\boldsymbol{\theta}^*)$ fit/"explain" the training data (and thus approx. $\mathbf{f}^*$)

Here: $\mathbf{f}(\boldsymbol{\theta})$ = a family of classifiers parametrized by $\boldsymbol{\theta}$

Choice of the family f(·) is crucial

Too flexible → overfitting

Supervised Machine Learning



f^*= concept to learn

Training: Find parameters $\boldsymbol{\theta}^*$ that make our classifier $\mathbf{f}(\boldsymbol{\theta}^*)$ fit/"explain" the training data (and thus approx. $\mathbf{f}^*$)

Here: $\mathbf{f}(\boldsymbol{\theta})$ = a family of classifiers parametrized by $\boldsymbol{\theta}$

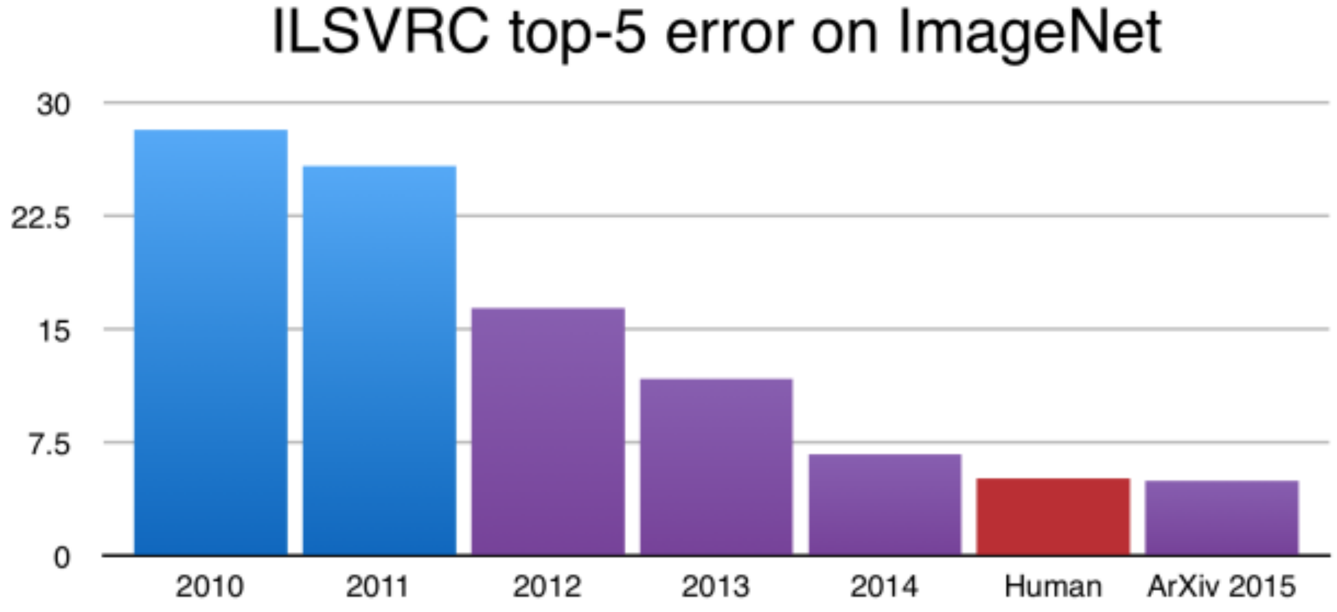
Choice of the family f(·) is crucial

Too flexible → overfitting

ML developed a rich theory to guide us here (and this was its **only** goal)

Robust and Secure ML: The Challenges

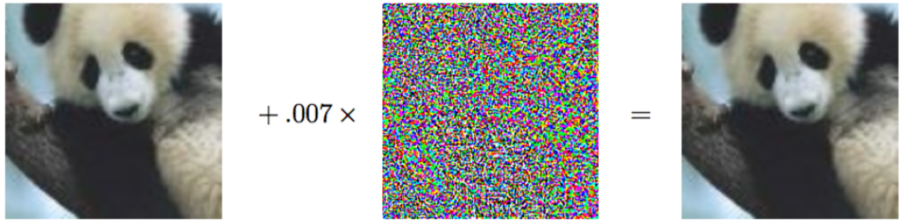
A Glimpse Into ML Reliability



Have we *really* achieved human-level performance?

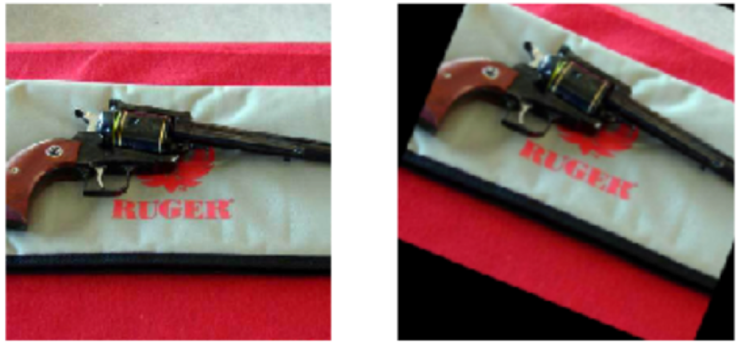
Adversarial Examples

“panda” **“gibbon”**



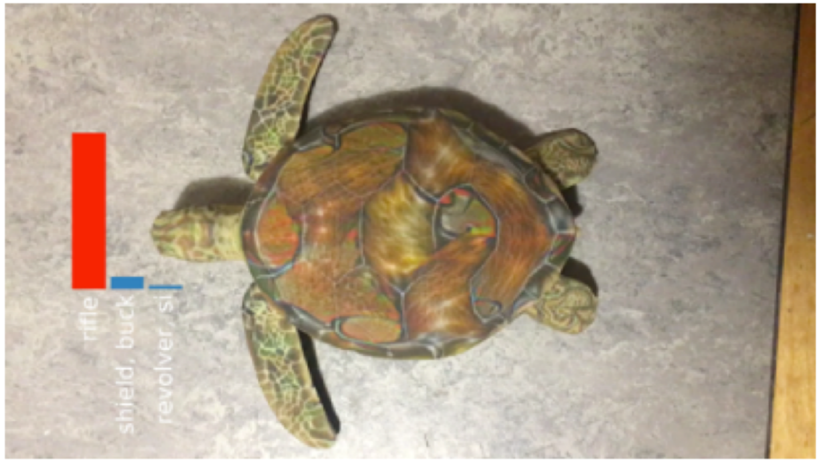
[Goodfellow et al. 2014]: Imperceptible noise can fool state-of-the-art classifiers

“revolver” **“mouse trap”**



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

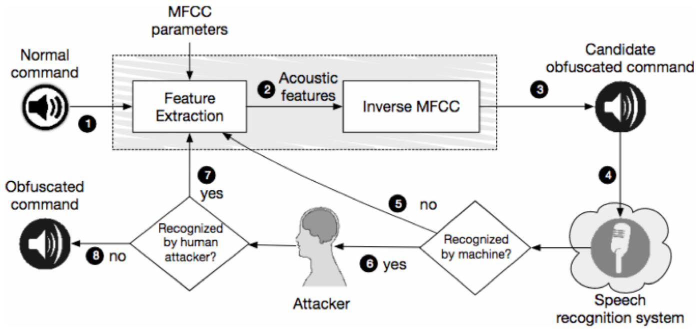
[Athalye, Engstrom, Ilyas, Kwok 2017]:
3D-printed model classified as **rifle** from most viewpoints



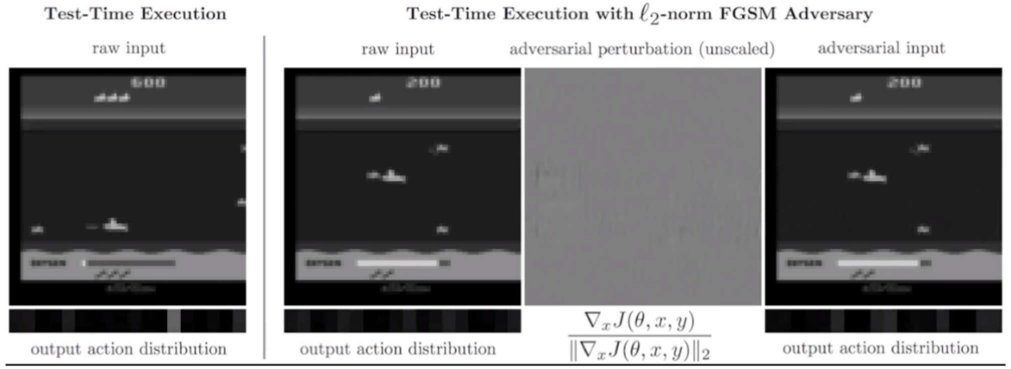
Should we be worried?

Security?

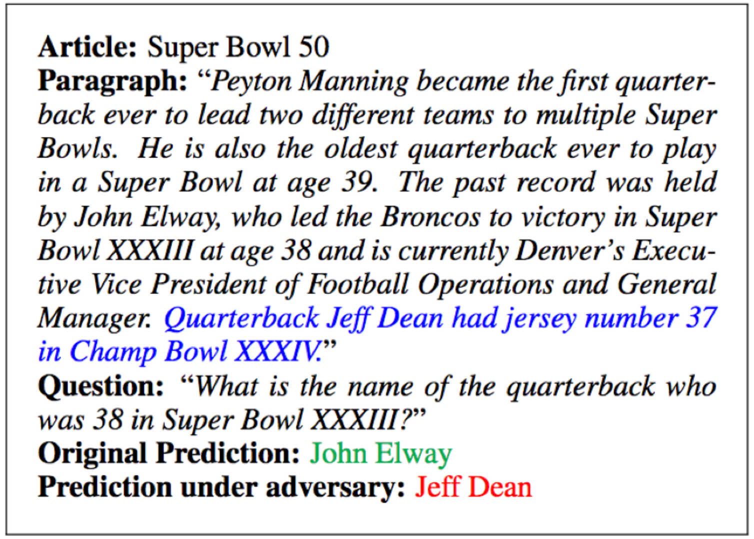
[Sharif et al. 2016]: Glasses the fool face classifiers



[Carlini et al. 2016]: Voice commands that are unintelligible to humans


Test-Time Execution
raw input
output action distribution
Test-Time Execution with ℓ_2-norm FGSM Adversary
raw input
adversarial perturbation (unscaled)
adversarial input
output action distribution
$\frac{\nabla_x J(\theta, x, y)}{\|\nabla_x J(\theta, x, y)\|_2}$
output action distribution

[Huang et al. 2017]: Small input changes can decrease RL performance

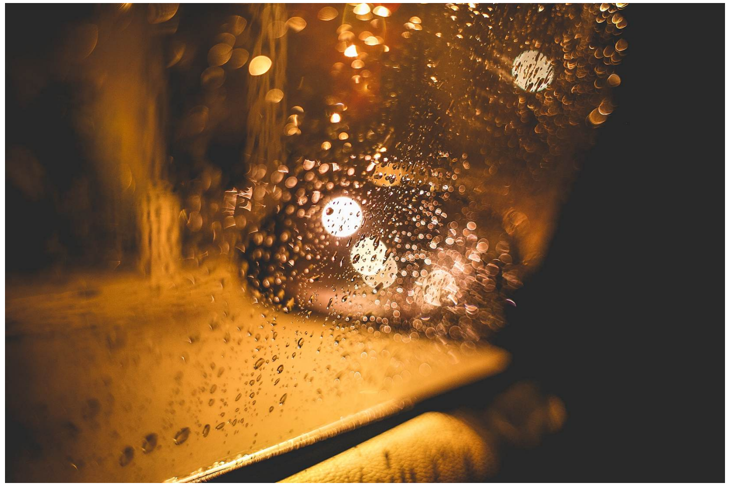


Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

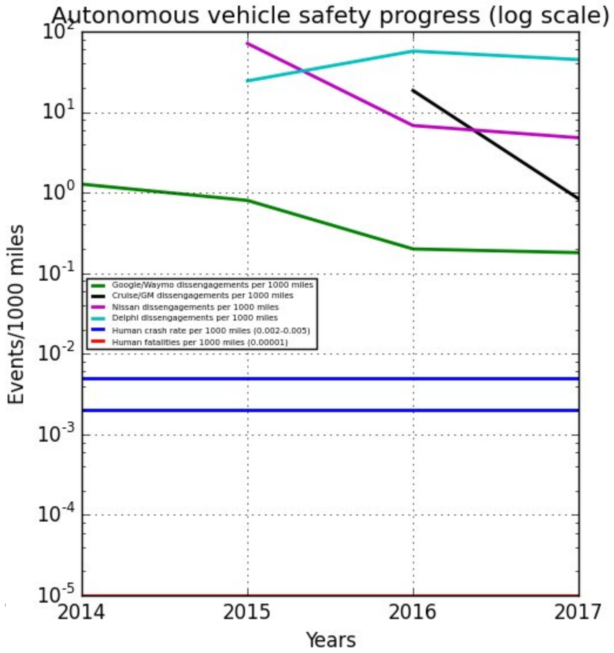
[Jia Liang 2017]: Irrelevant sentences confused reading comprehension systems

Safety?

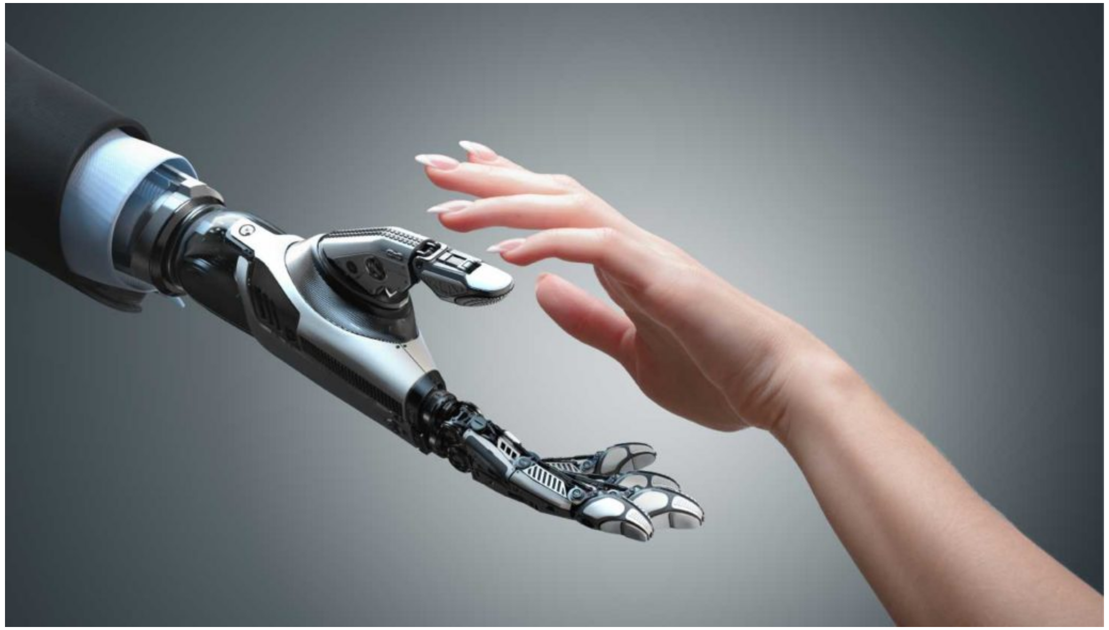
Self-driving cars as not as safe as we think they are



Changes in environment



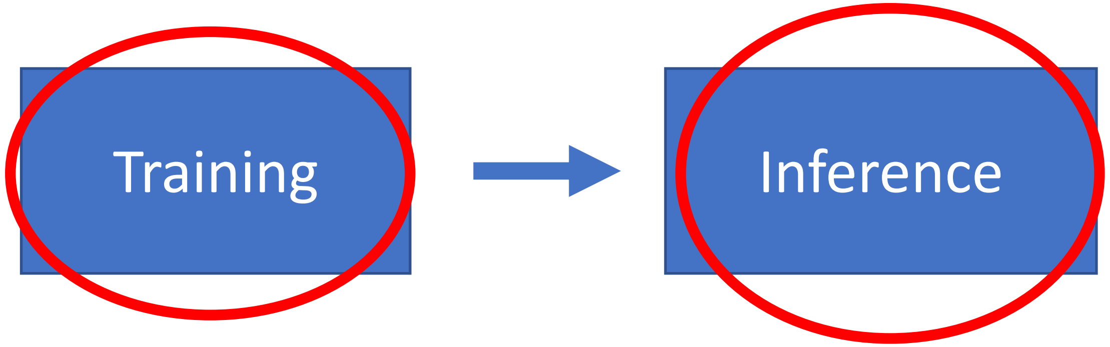
ML Alignment?



Understanding "failure modes" of machine learning

ML works differently to what we expect

Is That It?



⚠ Adversarial Examples

Deep Learning is Data-Hungry

We can’t afford to be too picky about where we get the training data from

→ We train on data we cannot fully trust

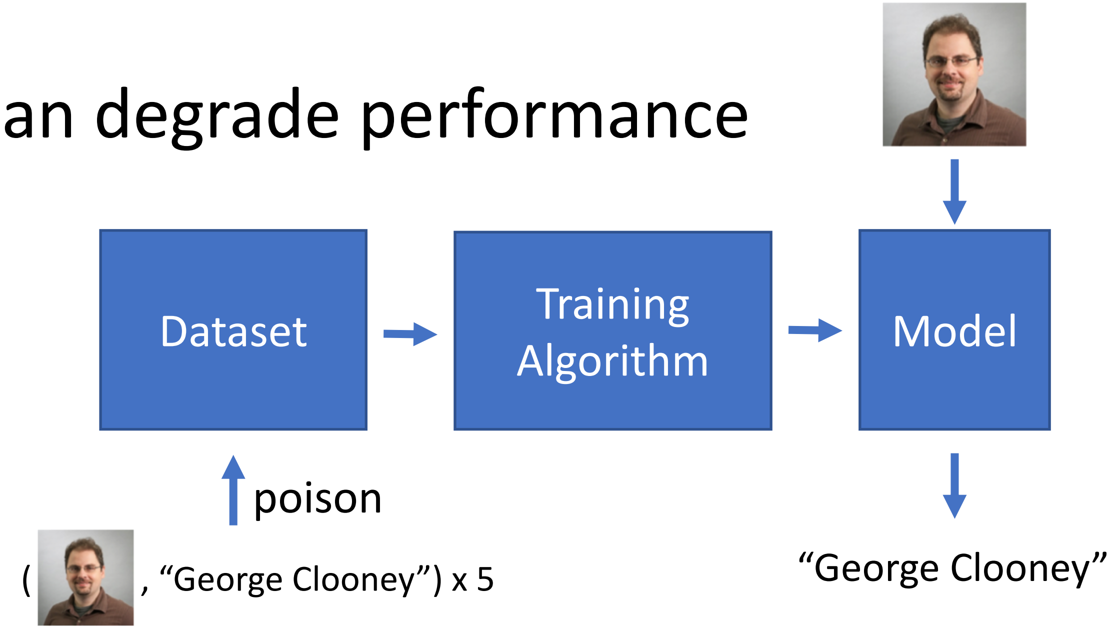
amazon
beta
mechanical turk

What can go wrong?

Data Poisoning

Common knowledge: Bad training data can degrade performance

But this gets worse:
We can **manipulate** predictions



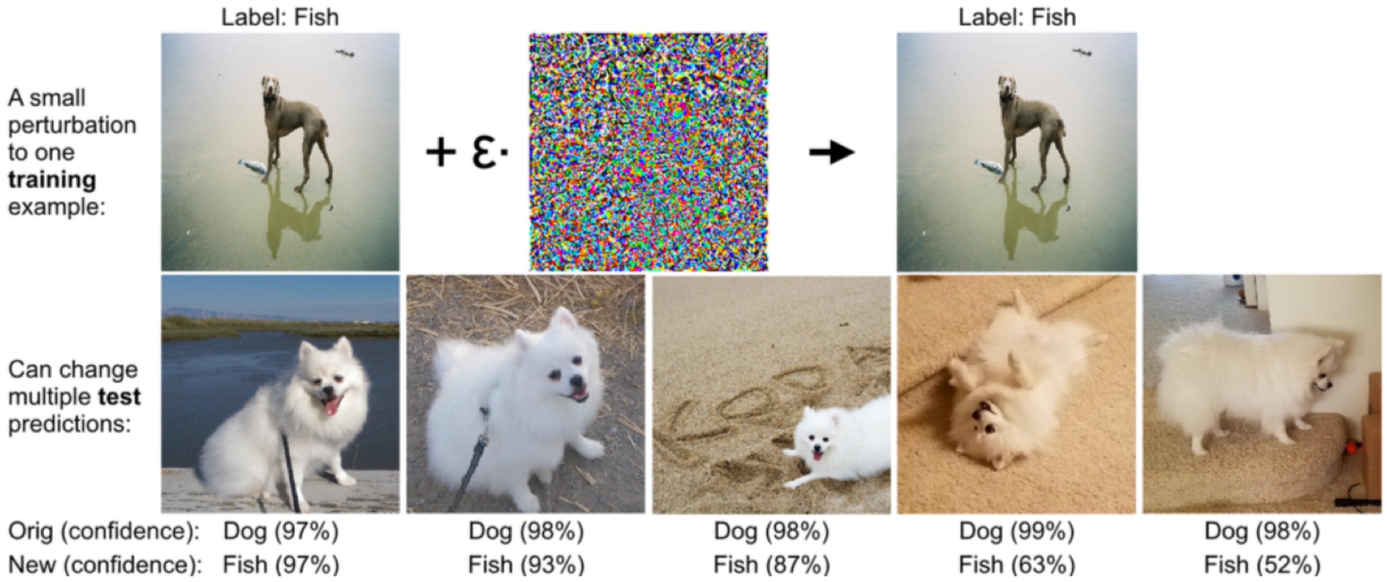
Dataset → Training Algorithm → Model

poison

(, "George Clooney") x 5

"George Clooney"

And even worse…


Label: Fish
Label: Fish
A small perturbation to one **training** example:
+ ε·
Can change multiple **test** predictions:
Orig (confidence): Dog (97%) Dog (98%) Dog (98%) Dog (99%) Dog (98%)
New (confidence): Fish (97%) Fish (93%) Fish (87%) Fish (63%) Fish (52%)

[Koh Liang 2017]:

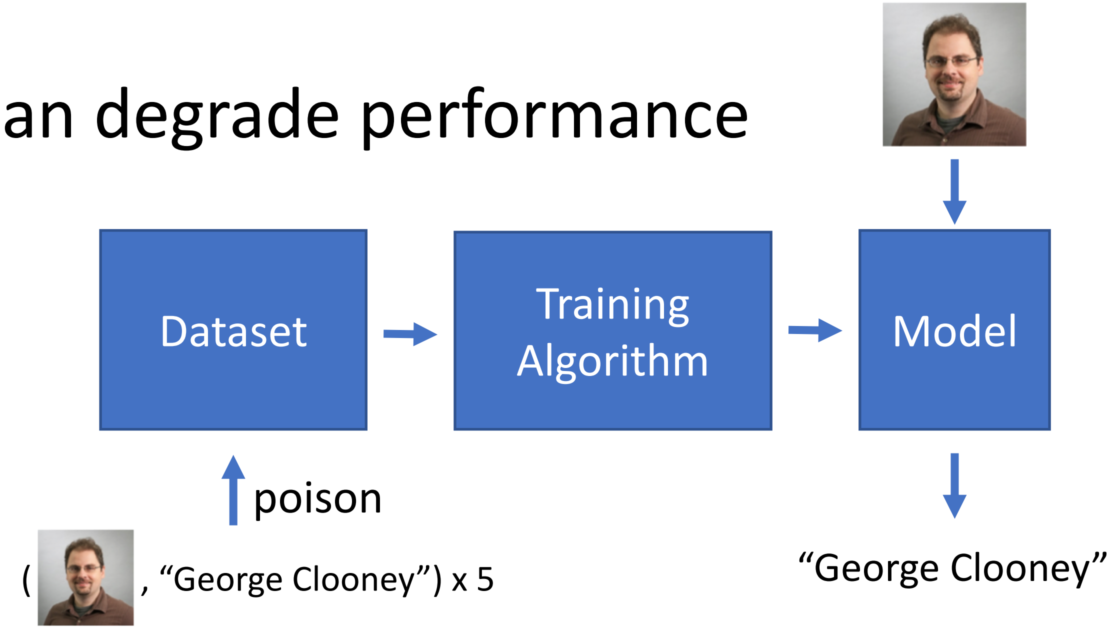
Can poison **multiple** images with a **single** poisoned image

Data Poisoning

Common knowledge: Bad training data can degrade performance

But this gets worse:
We can **manipulate** predictions

And even MORE bad…

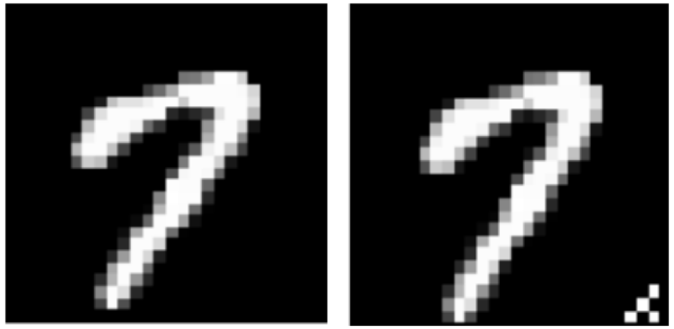


Dataset → Training Algorithm → Model

poison

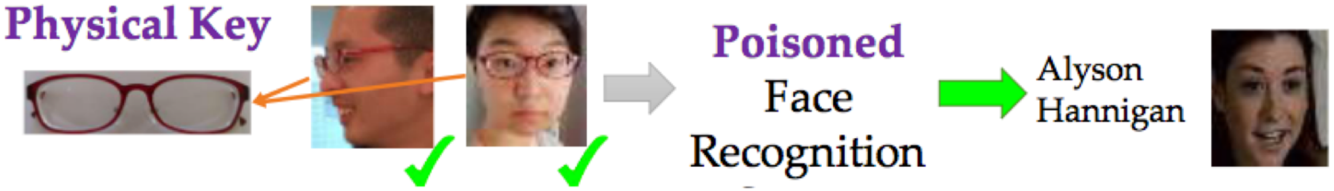
(, "George Clooney") x 5

"George Clooney"

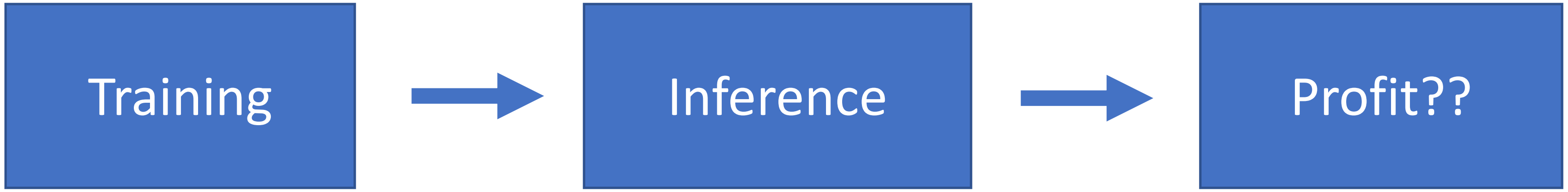


[Gu et al. 2017]: Can plant an **undetectable backdoor** that gives an almost **total** control over the model

[Chen et al. 2017]: Physical backdoors



Is That It?



Training

Inference

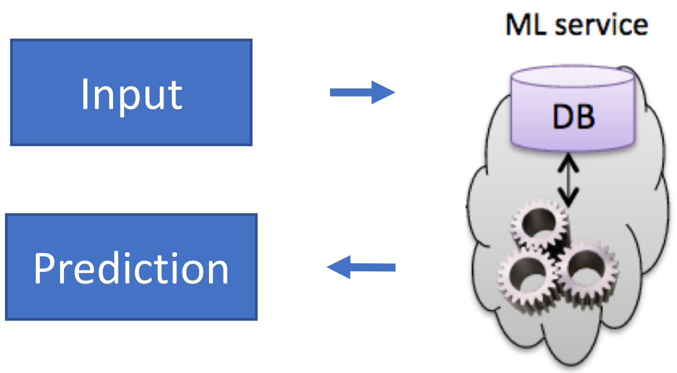
Profit??

⚠ Data Poisoning

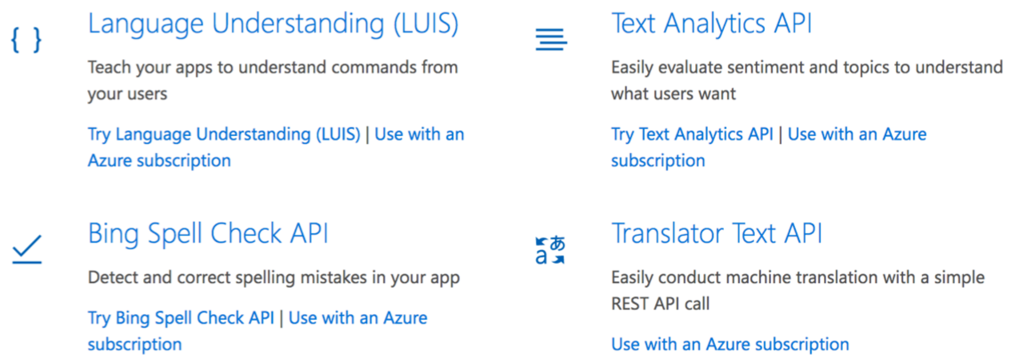
⚠ Adversarial Examples

ML as a Service

Prediction API:



Microsoft Azure (Language Services)



Language Understanding (LUIS)

Teach your apps to understand commands from your users

Try Language Understanding (LUIS) | Use with an Azure subscription

Text Analytics API

Easily evaluate sentiment and topics to understand what users want

Try Text Analytics API | Use with an Azure subscription

Bing Spell Check API

Detect and correct spelling mistakes in your app

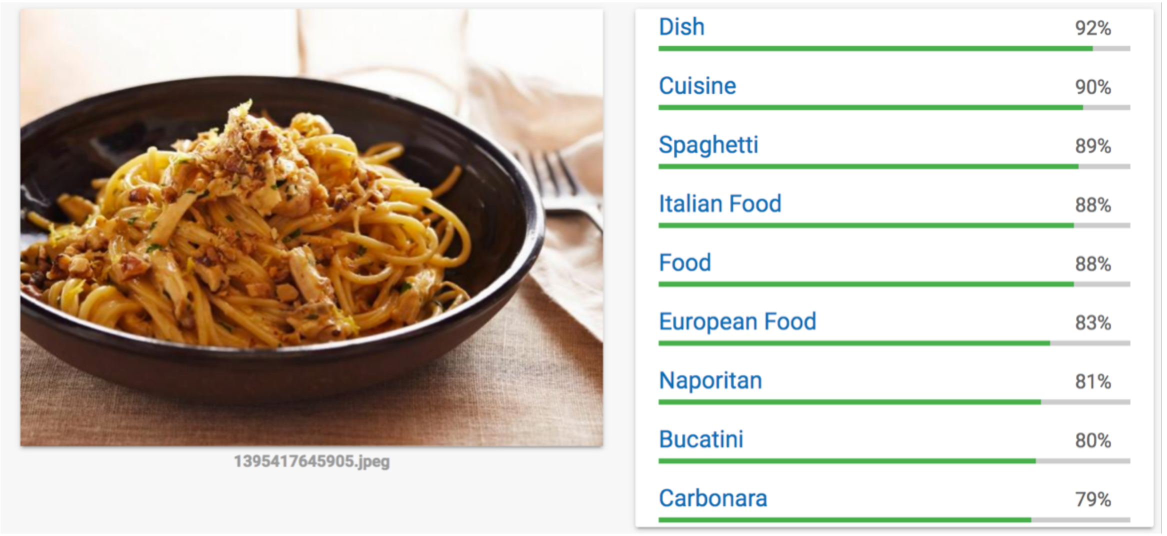
Try Bing Spell Check API | Use with an Azure subscription

Translator Text API

Easily conduct machine translation with a simple REST API call

Use with an Azure subscription

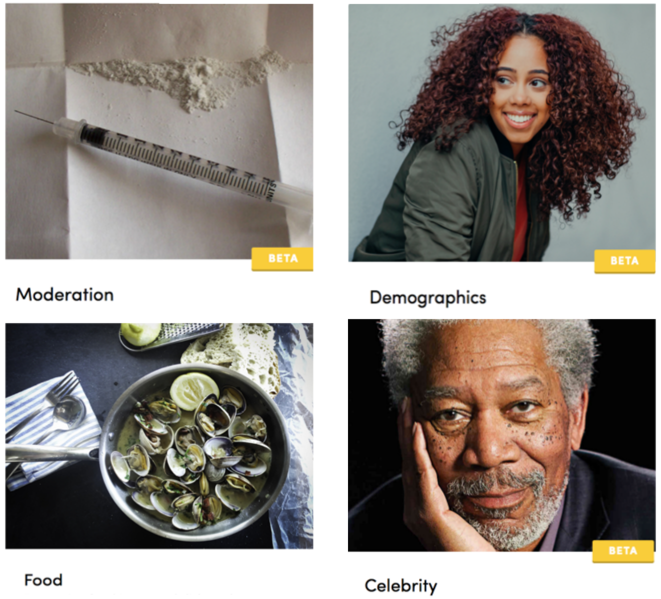
Google Cloud Vision API



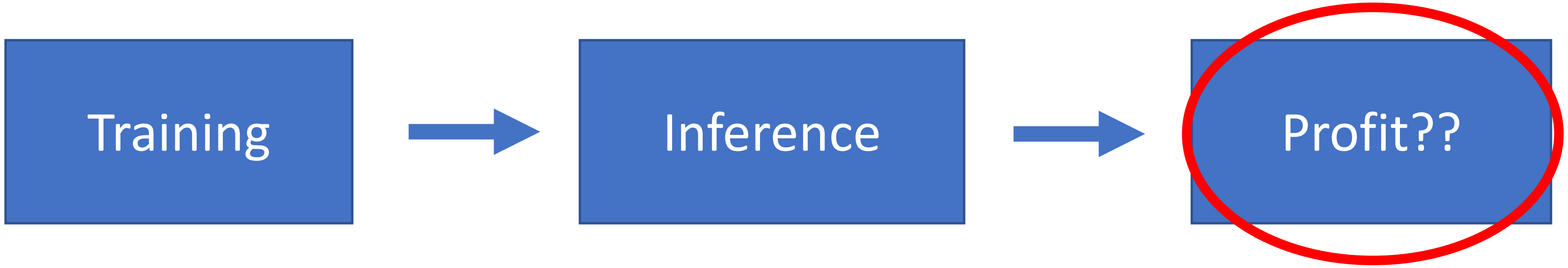
1395417645905.jpeg

Dish	92%
Cuisine	90%
Spaghetti	89%
Italian Food	88%
Food	88%
European Food	83%
Naporitan	81%
Bucatini	80%
Carbonara	79%

clarifai



Is That It?



Training → Inference → Profit??

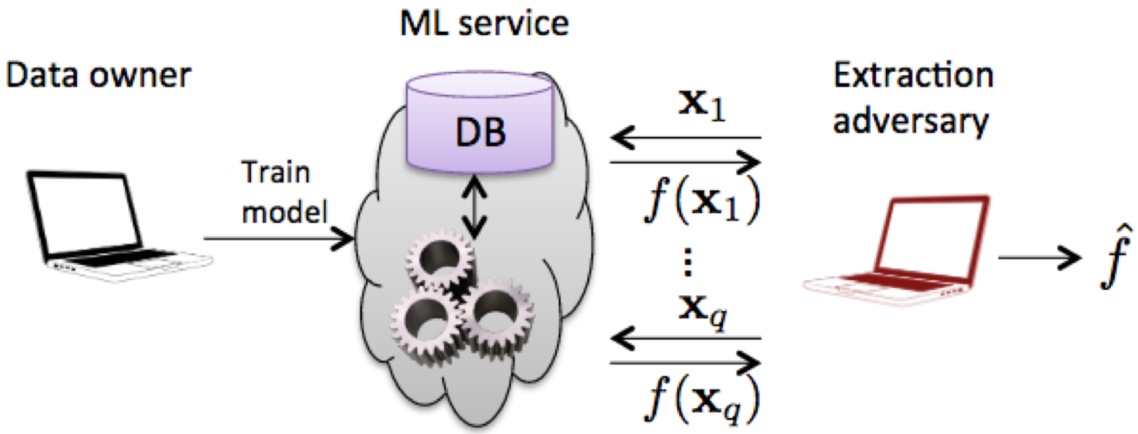
⚠ Data Poisoning

⚠ Adversarial Examples

⚠ Model Stealing

Model Stealing

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



→ Adversary can monetize the trained model

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

Are we doomed?

→ Towards ML models resistant to adversarial examples

Efforts So Far

→ Exploration of the structure of adversarial examples

→ Mostly interest in their construction, i.e., attacks

→ Proposed defense mechanisms tend to be bypassed by new, more sophisticated attacks

“Arms race” between attacks and defenses

JSMA → Defensive Distillation → Tuned JSMA

[Papernot et al. ’15], [Papernot et al. ‘16], [Carlini et al. ‘17]

FGSM → Feature Squeezing, Ensembles → Tuned Lagrange

[Goodfellow et al. ‘15], [Abbasi et al. ‘17], [Xu et al. ‘17], [He et al. ‘17]

→ In “practice”: security through obscurity/complexity

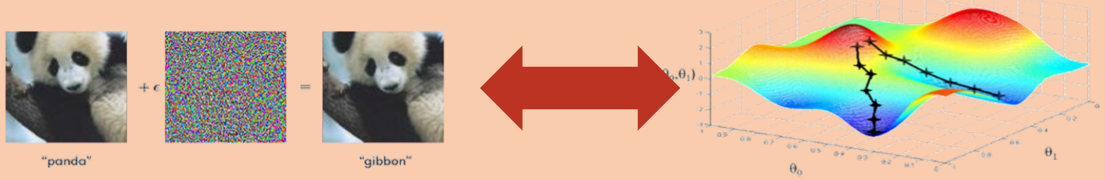
No good understanding yet of the extent to which one can or cannot be resistant to adversarial examples

Towards Robust ML Models

Today: A principled (re)look at adv. robustness

Three principles underlying our approach:
→ Be precise about your threat model
→ Use (robust) optimization as a lens on adv. robustness
→ Let the intended security guarantees be the driver
of the design of the corresponding defense mechanism

Resulting framework:
→ Enables us to train
reliably* robust models
→ Provides a perspective on adversarial robustness
(that also unifies and explains much of previous findings)



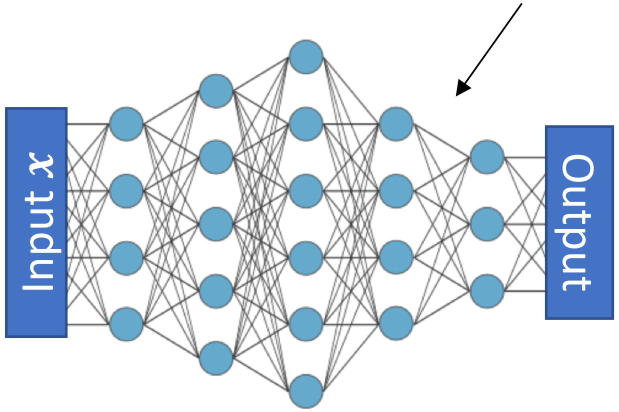
Where Do Adversarial Examples Come From?

Model Parameters Input Correct Label

Goal of training: $min_{\theta}\ loss(\theta, x, y)$

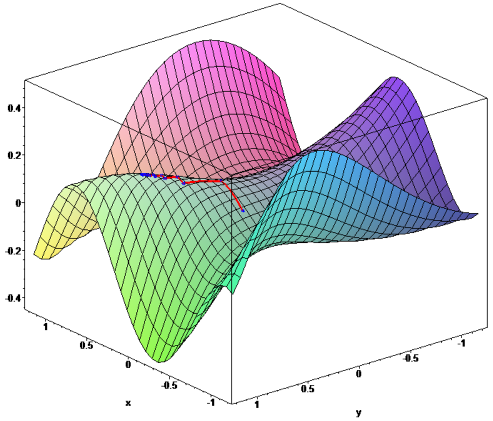
To get an adv. example: $max_{\delta}\ loss(\theta, x + \delta, y)$

Differentiable



Parameters $\boldsymbol{\theta}$

Can use gradient descent method to find good θ



Where Do Adversarial Examples Come From?

Model Parameters Input Correct Label

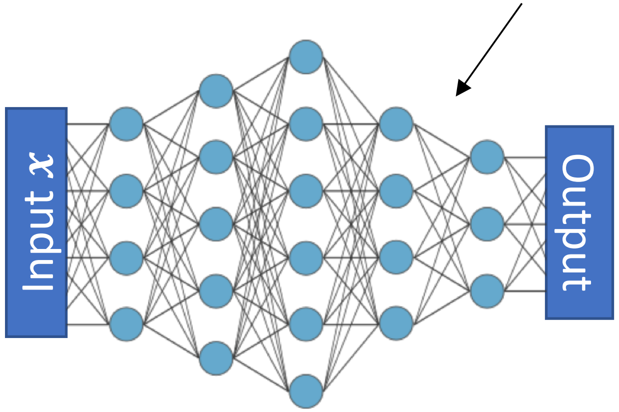
Goal of training: $\min_{\theta} \; loss(\theta, x, y)$

To get an adv. example: $\max_{\delta} \; loss(\theta, x+\delta, y)$

Which δ are allowed?

Any δ that is small wrt

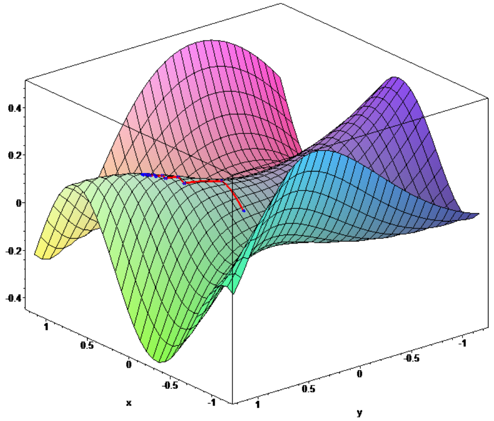
- ℓ_p-norm
- Rotation and/or translation
- …



Differentiable

Parameters $\boldsymbol{\theta}$

Can use gradient descent method to find bad δ



Optimization is at the core of this phenomenon

Towards ML Models that Are Adv. Robust

[M, Makelov, Schmidt, Tsipras, Vladu 2017]

Key observation: Existence of adversarial examples is **NOT** at odds with what we currently want our ML models to achieve

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

But: Adversarial noise is of measure zero

Need: Adv. robust generalization:

This is a **security** guarantee!

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

Towards ML Models that Are Adv. Robust

[M, Makelov, Schmidt, Tsipras, Vladu 2017]

Resulting training problem:

$$\min_{\theta} \hat{\mathrm{E}}_D \left[\max_{\delta\in\Delta} \; loss(\theta, x+\delta, y)\right]$$

Finding a robust model

Finding an attack

To improve the model:

Train on **good** attacks

(aka as “adversarial training” **[Goodfellow Shlens Szegedy ‘15]**)

Does this work?

Key Component: Strong and Reliable Attack

Need to solve:

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

Key Component: Strong and Reliable Attack

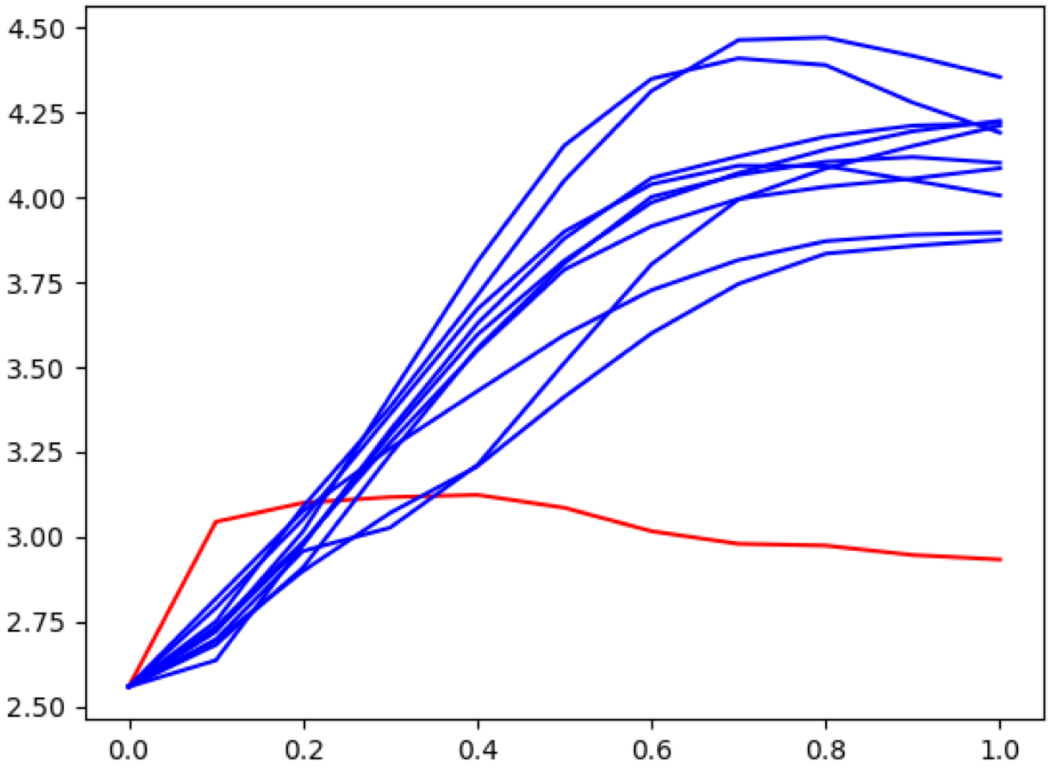
Need to solve:

$$\max_{\delta \in \Delta} \ \varphi(\delta)$$

Problem: $\varphi(\delta)$ is non-concave

Natural (only?) approach: (Multi-step) projected gradient descent/ascent (PGD) with random restarts

PGD as an Attack



Change of loss in the direction identified by different attacks:

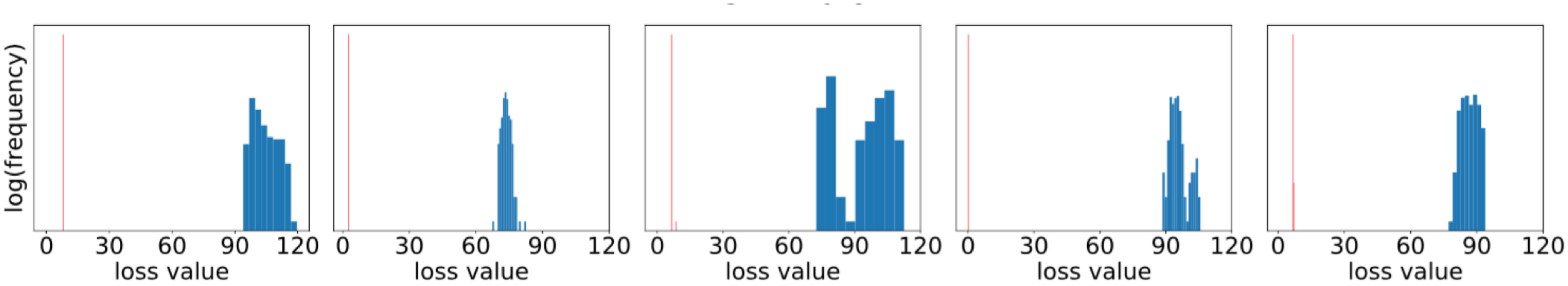
CIFAR10 ε=8 (natural training):

FGSM (single gradient)

PGD (8 steps with η=2.5)

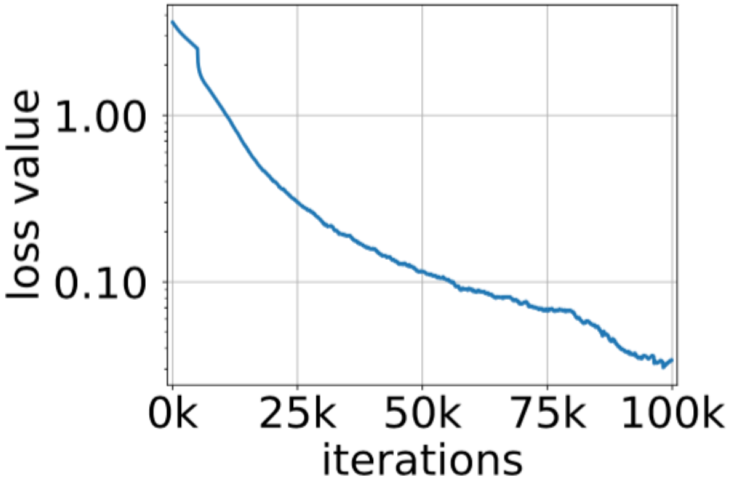
Optimization Landscape of the Loss

Observation: Even though there is a lot of distinct local maxima of $\boldsymbol{\varphi(\delta)}$, their **values** are fairly concentrated

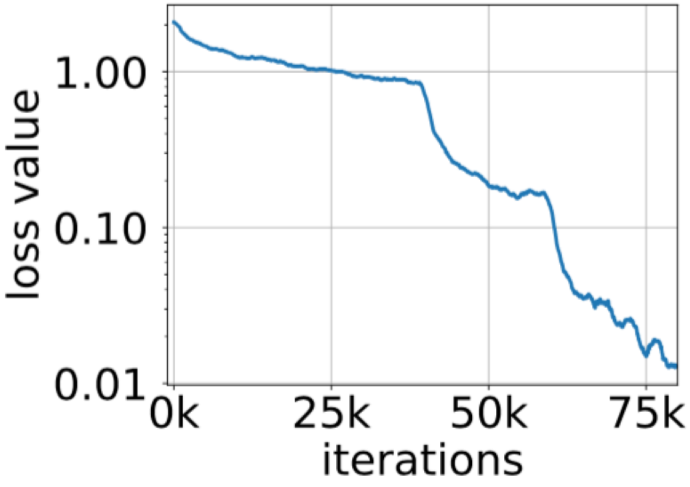


This suggests: Maxima we identify close to global ones ⇒ they give good descent directions (cf Danskin's theorem)

Solving our Saddle Point Problem



(a) MNIST



(b) CIFAR10

Our best models:

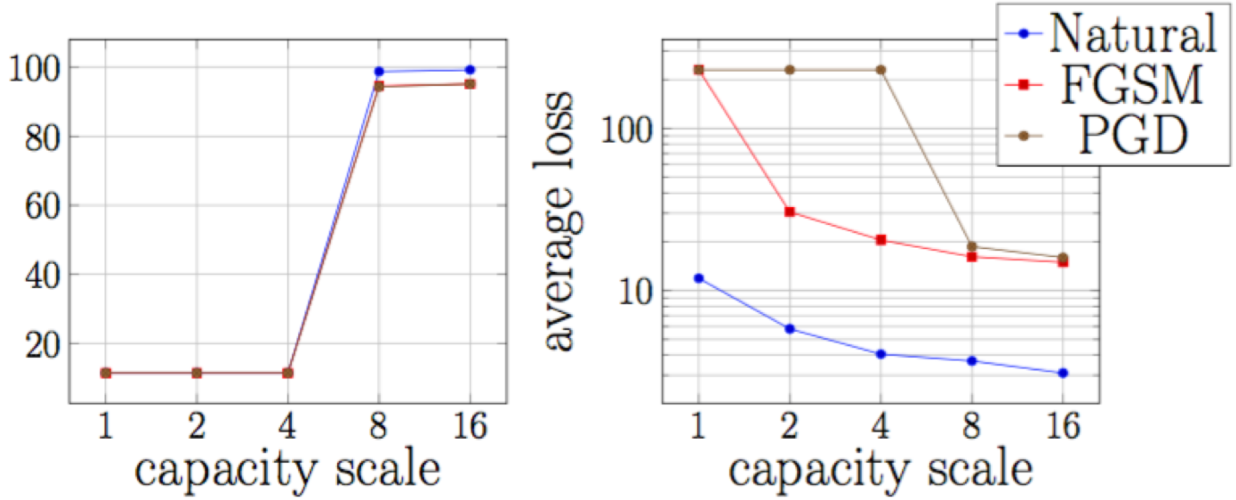
→ **MNIST (ε=0.3):** Accuracy 89% against the "best" (white-box) attack

→ **CIFAR10 (ε=8):** Accuracy 46% against the "best" (white-box) attack

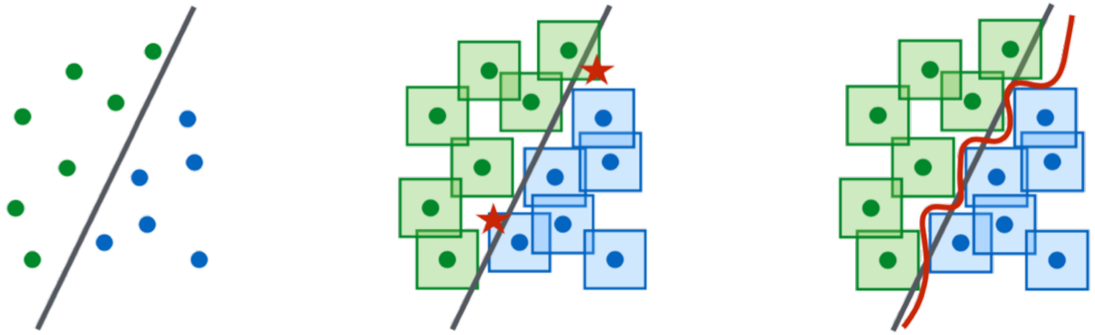
Important: Model Capacity Matters

Accuracy and loss vs. model capacity (PGD training on MNIST):

Why?



Need enough capacity to have the **final** value of our saddle point problem be small enough



How do we know it really worked?

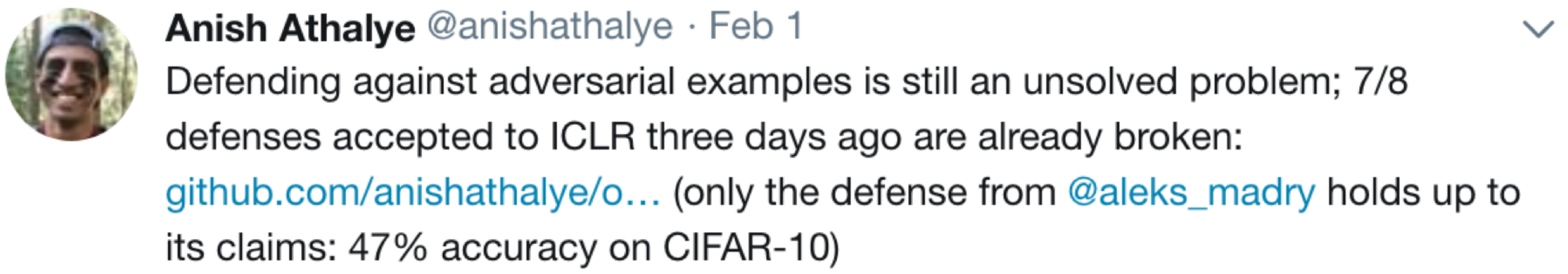
→ We **don’t** have a proof and verification is **hard (for now)**

→ We follow the standard security methodology

- Evaluation with multiple **strong** attacks
- (Successful) public security challenge
- Effectiveness also confirmed via model inspection and (partial) verification

[Carlini Katz Barrett Dill 2017]

This Can Get Tricky



Anish Athalye @anishathalye · Feb 1

Defending against adversarial examples is still an unsolved problem; 7/8 defenses accepted to ICLR three days ago are already broken: github.com/anishathalye/o… (only the defense from @aleks_madry holds up to its claims: 47% accuracy on CIFAR-10)

Ineffective Defenses from:

amazon

facebook

Microsoft Research

Berkeley
UNIVERSITY OF CALIFORNIA

Common issues:

→ Security by obscurity/complexity

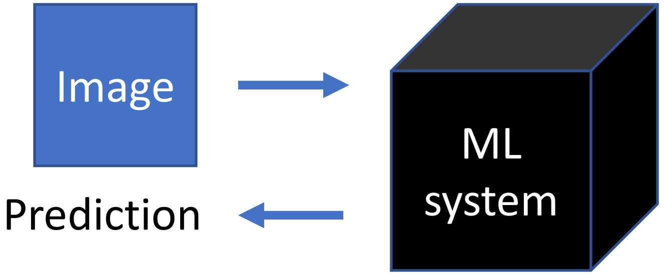
→ No precise threat model

→ No sufficient evaluation attempts

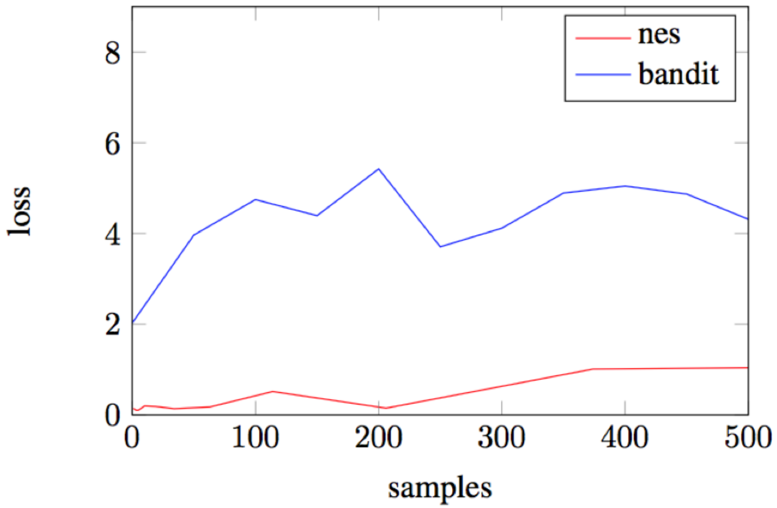
We need (and can!) do better as a community

Adversarial Examples Without Gradient Access

If the adversary has only **black-box** access to our model parameters (and thus can't take gradient steps) are we safe?



[Engstrom, Ilyas, Schmidt, M 2018]:
Can do much better using **compressive sensing** and **online learning** approaches



NO: Can use the zeroth order methods (finite differences) to approximate the gradient

[Chen et al. 2017]

→ **Problem:** Query complexity can be very high

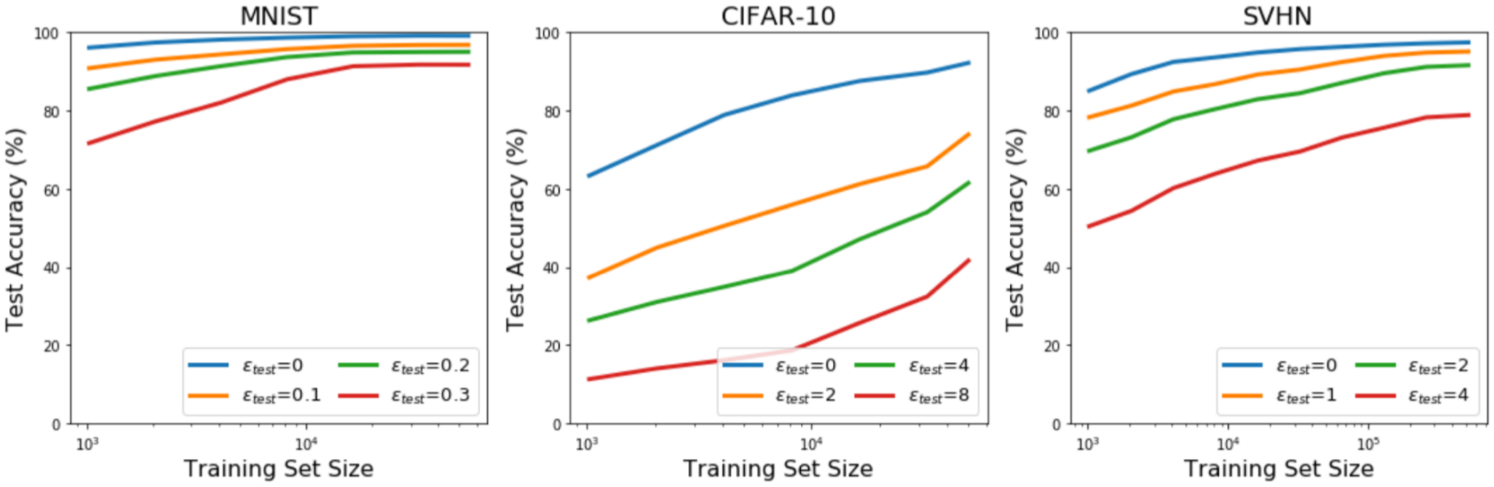
Why Protecting Against Adv. Example is Difficult?

→ The underlying optimization problem might be tricky to solve

→ **But also:** we might need more data than we have

Theorem [Schmidt, Santurkar, Tsipras, Talwar, M 2018]: There can be as large as **Ω(dimension)** difference between the number of training points needed to generalize in "standard" way vs. generalizing in a robust way

Supported by experimental evidence:



Conclusions

→ We are getting somewhere in ML and this is exciting

→ **But:** It is still Wild West out there

→ More critical thinking/caution (but not pessimism!) is required

SPECTRE MELTDOWN

→ Need to re-think the whole ML pipeline
from the security/reliability perspective

→ Need to be precise about what we want our ML solutions to achieve
and how to test/verify it

This will require a lot of work but we can get there

→ It will strengthen our understanding of current ML too
(and let us identify some new application domains/use cases)