6.S979: Topics in Deployable ML

Martin Rinard

David Sontag

Constantinos Daskalakis

Antonio Torralba

Arvind Satyanarayan

Asuman Ozdaglar

Ankur Moitra

Pablo Parrilo

Aleksander Madry

Armando Solar-Lezama

Russ Tedrake

Course Logistics

Martin Rinard

David Sontag

Constantinos Daskalakis

Antonio Torralba

Arvind Satyanarayan

Asuman Ozdaglar

Ankur Moitra

Pablo Parrilo

Aleksander Madry

Armando Solar-Lezama

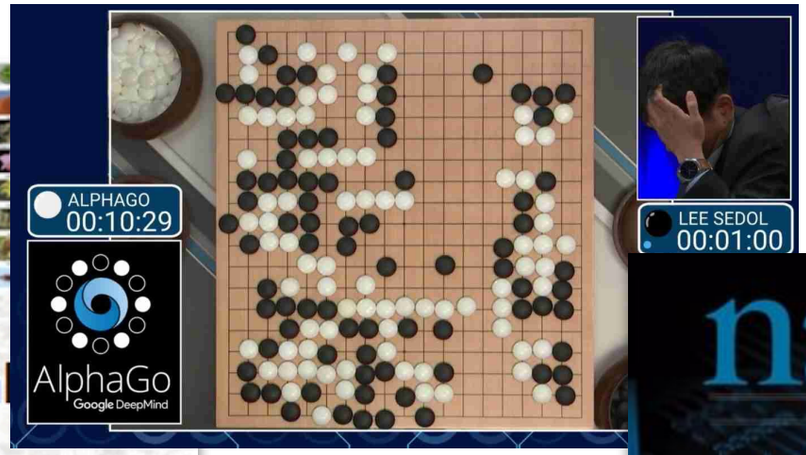
Russ Tedrake

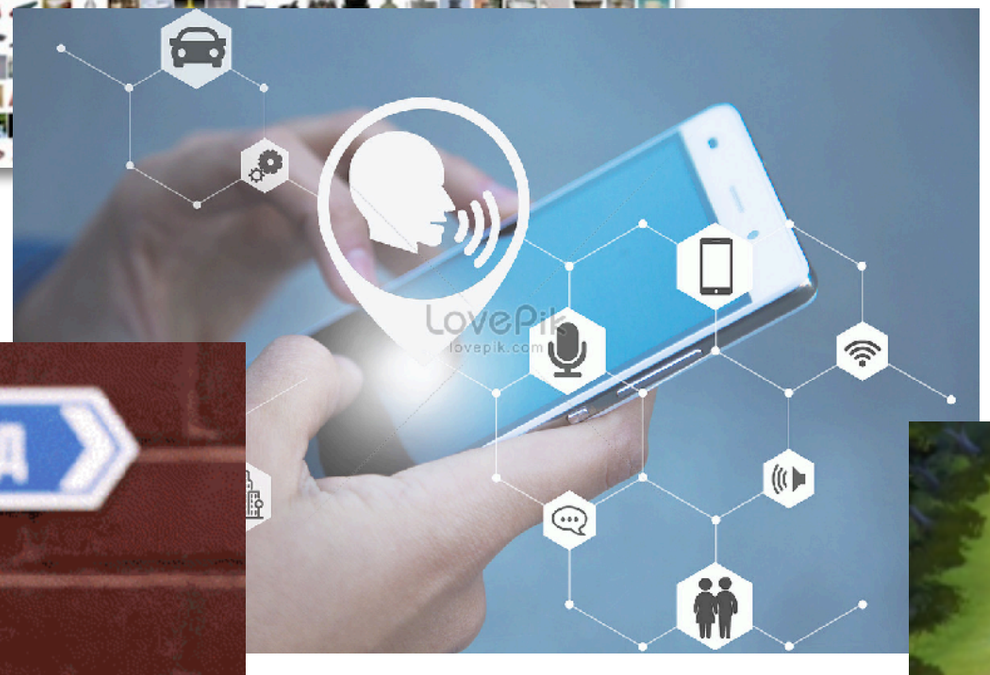
Please fill out the form at **https://bit.ly/2kwmY63** (by **today**)

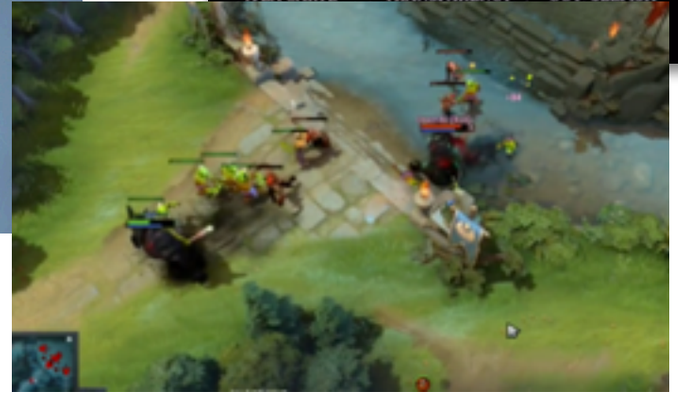
- **New meeting time:** TR 12pm-1:30pm 11am-12:30pm or 12:30pm-2pm
- **Prerequisites:** Solid knowledge of ML (at the level of 6.867)
- **Grading:** Project [70%] + Scribing [25%] + Class discussion [5%]

What will this class be about?

ML: A Success Story

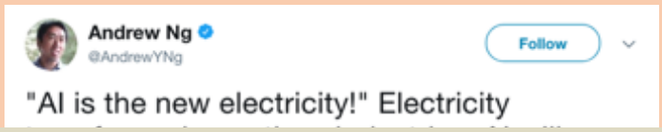




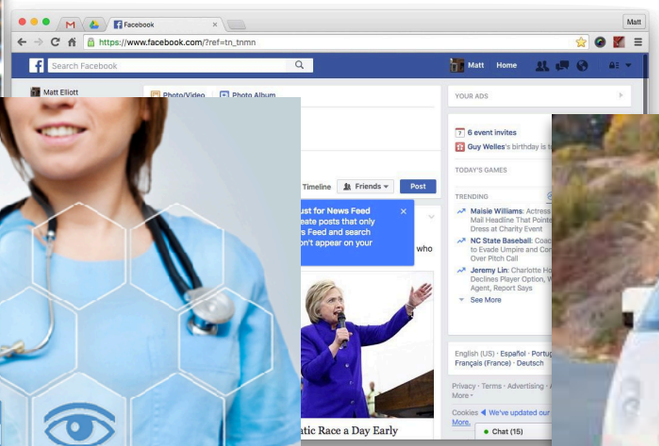


ML: A Success Story

Trump Signs Executive Order Promoting Artificial Intelligence



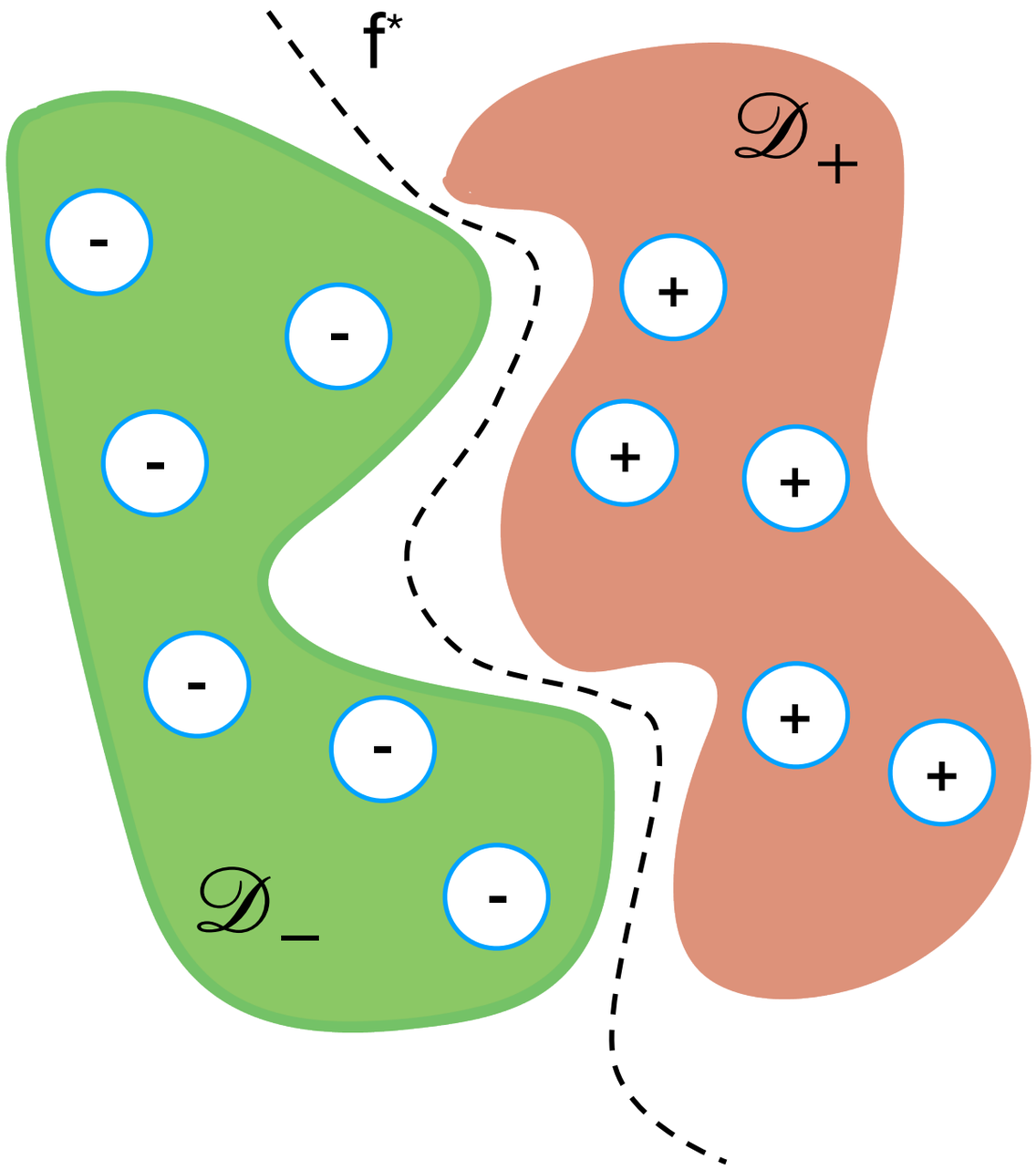
What will it take to be able to confidently deploy ML in the real-world?



Question I:

Do we understand our ML toolkit?

(Supervised) Machine Learning



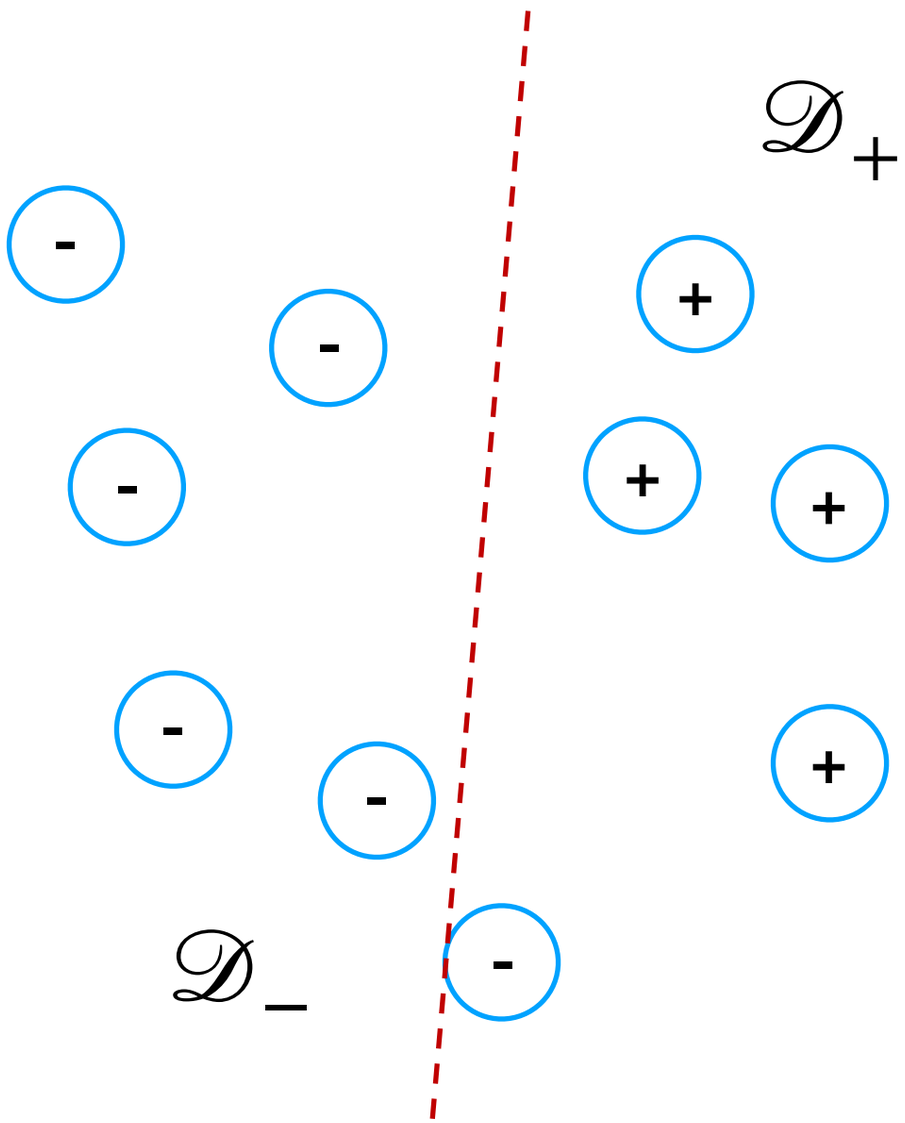
f*: concept to learn

f(θ): (parametrized) model class

Training: Recover (approx.) **f*** by finding parameters θ* s.t. **f**(θ*) fits the training data

Choice of family **f( )** is crucial

(Supervised) Machine Learning



f*: concept to learn

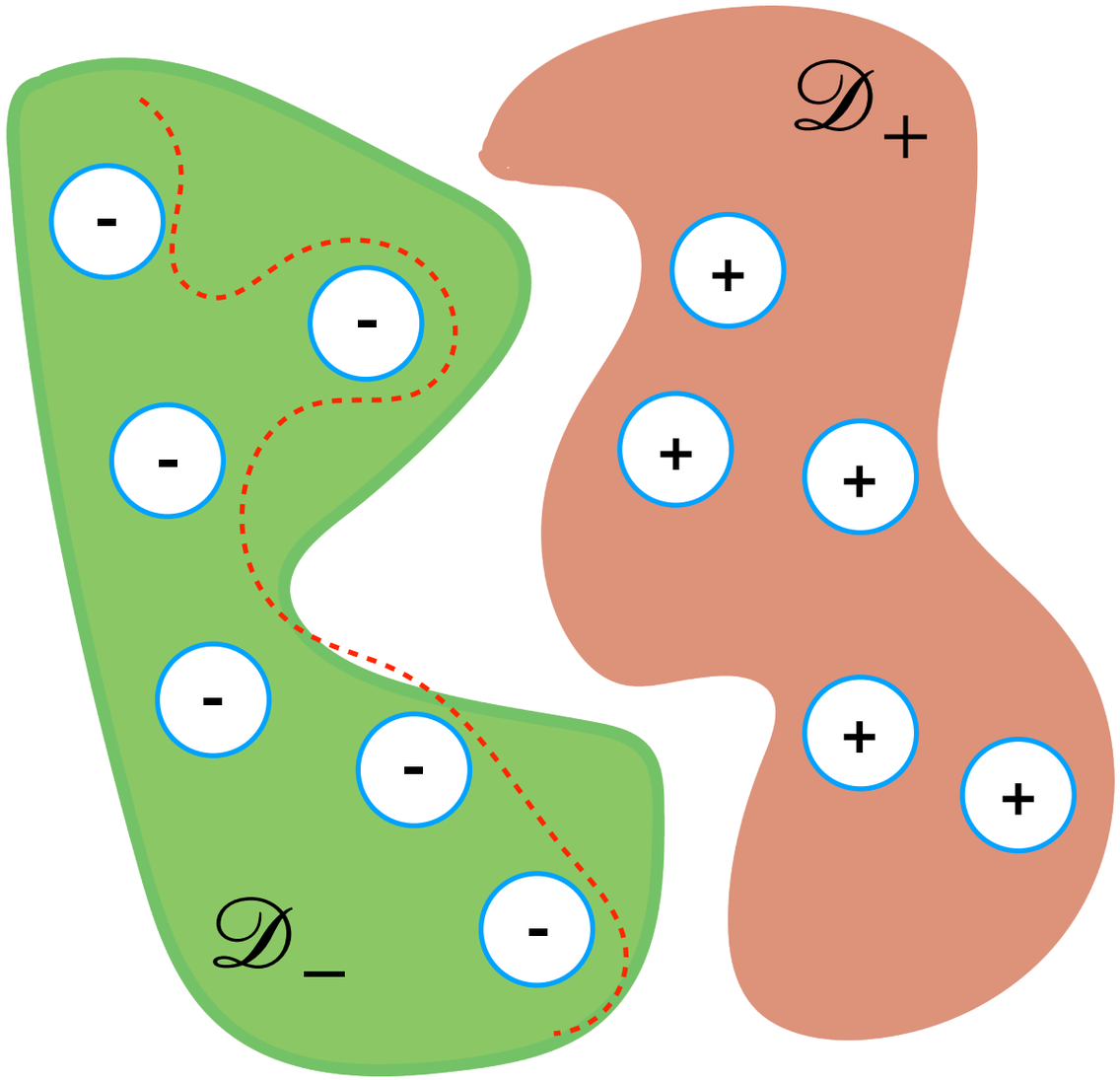
f(θ): (parametrized) model class

Training: Recover (approx.) **f*** by finding parameters θ^* s.t. **f**(θ^*) fits the training data

Choice of family **f( )** is crucial

Too simple → under-fitting

(Supervised) Machine Learning



f*: concept to learn

f(θ): (parametrized) model class

Training: Recover (approx.) **f*** by finding parameters θ^* s.t. **f**(θ^*) fits the training data

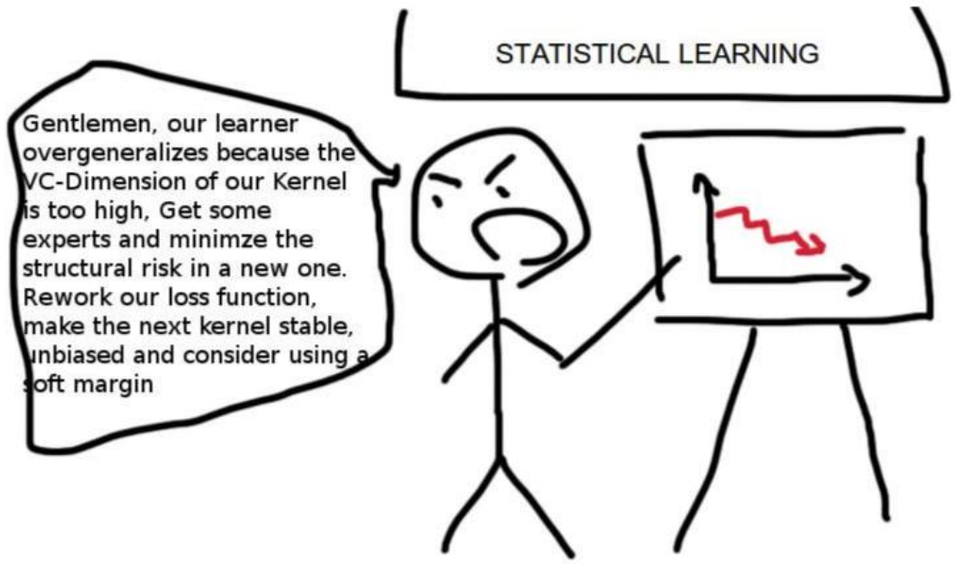
Choice of family **f( )** is crucial

Too simple → under-fitting

Too complex → over-fitting

"Classical" ML has rich theory to understand this phenomenon

But then…



Deep neural networks are **very** expressive, why don’t they overfit?

Optimization in Deep Learning

Our true goal: To minimize (wrt θ) the **population risk**

$$E_{(x,y)\in\mathcal{D}}[\textbf{loss}(f(\theta, x), y)]$$

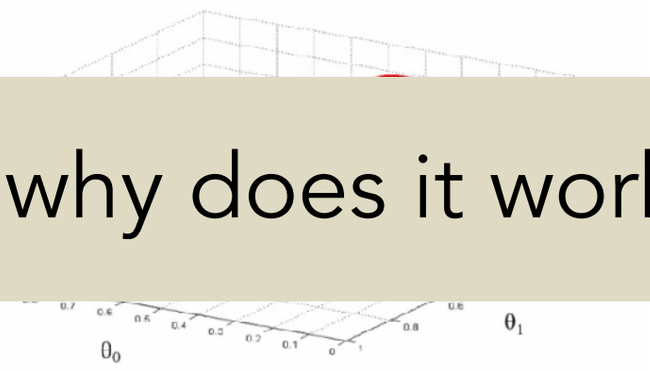
What we actually do: Minimize (wrt θ) the **empirical risk**

$$\sum_i \textbf{loss}(f(\theta, x_i), y_i)$$

where $\{(x_i,y_i)\}_i$ are the training data points

→ In case of neural networks, empirical risk is a continuous and (mostly) differentiable function

→ Can use gradient descent method (aka back-propagation)!



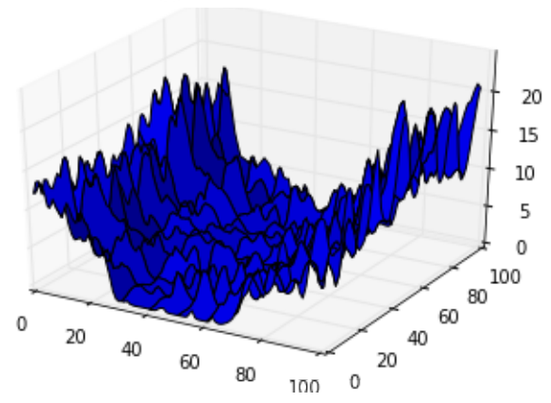
But why does it work?

Optimization in Deep Learning

$$\sum_i \textbf{loss}(f(\theta, x_i), y_i)$$

→ **Issue 1:** There are a **lot** of terms in this sum (lots of data)

→ Use **stochastic** gradient descent (SGD) instead of grad. descent (SGD = the workhorse of deep learning)



→ **Issue 2:** This problem is **very** non-convex

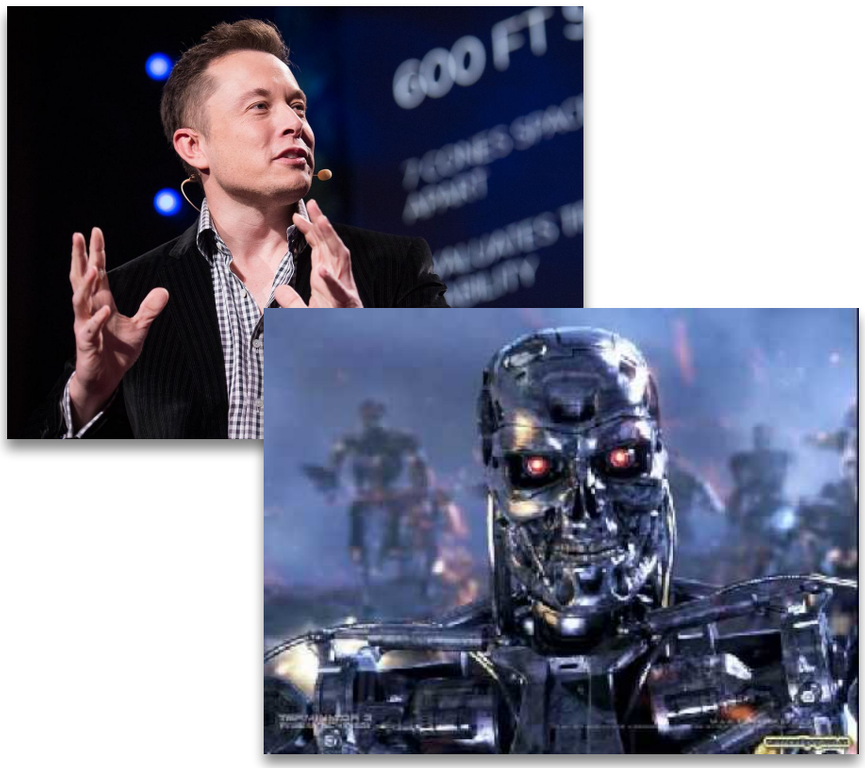
→ Still, we seem to reliably converge to good solutions. Why?

In fact: Stochasticity of SGD seems to be a "feature", not a deficiency. (Hypothesis: "Implicit regularization")

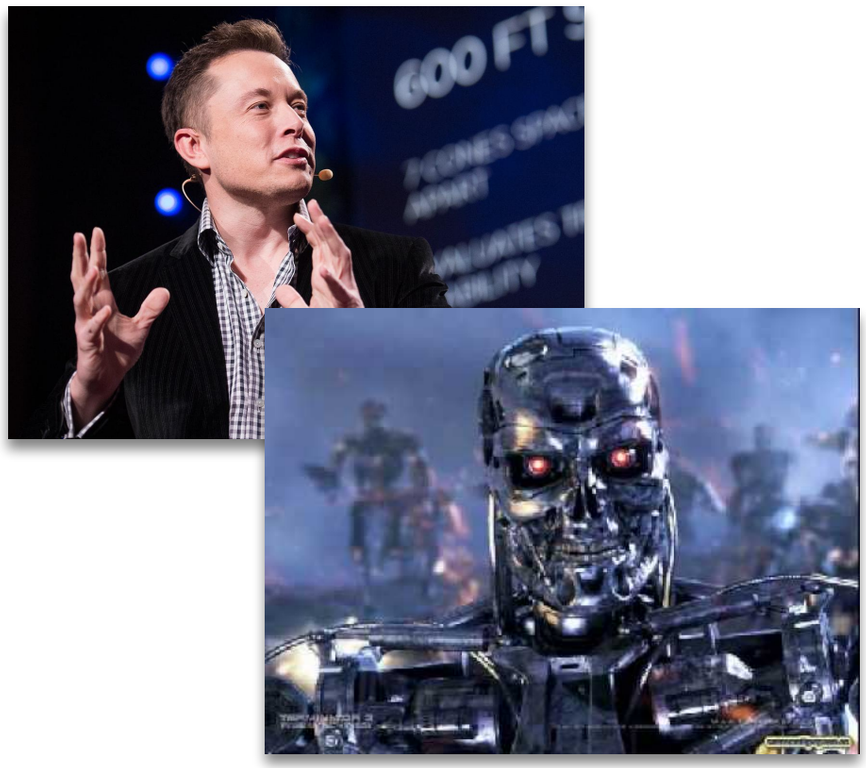
Question II:

Is our current ML robust?

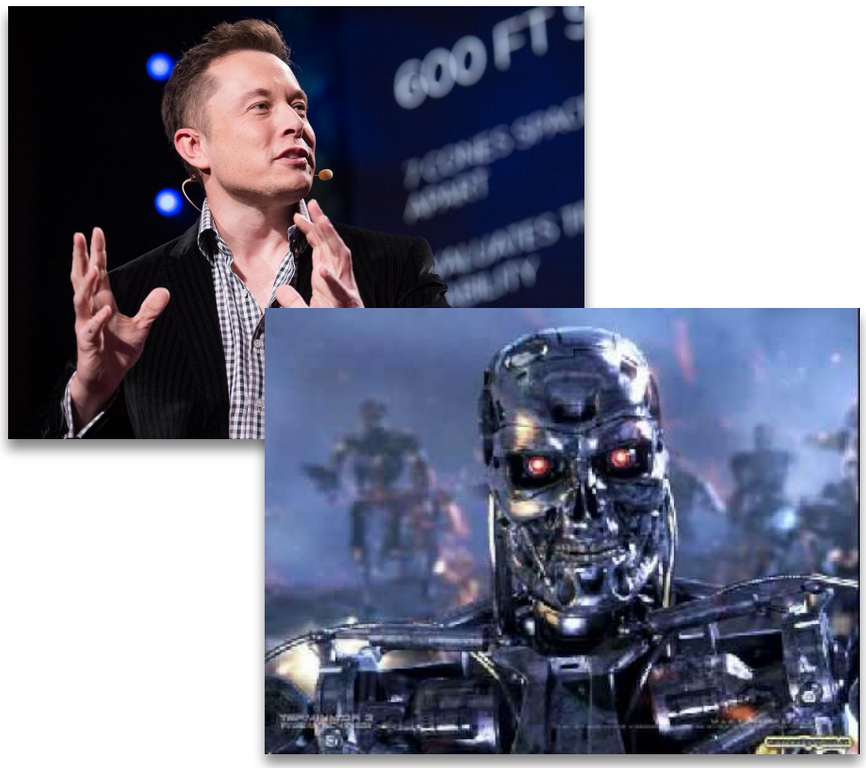
Can we rely on ML?



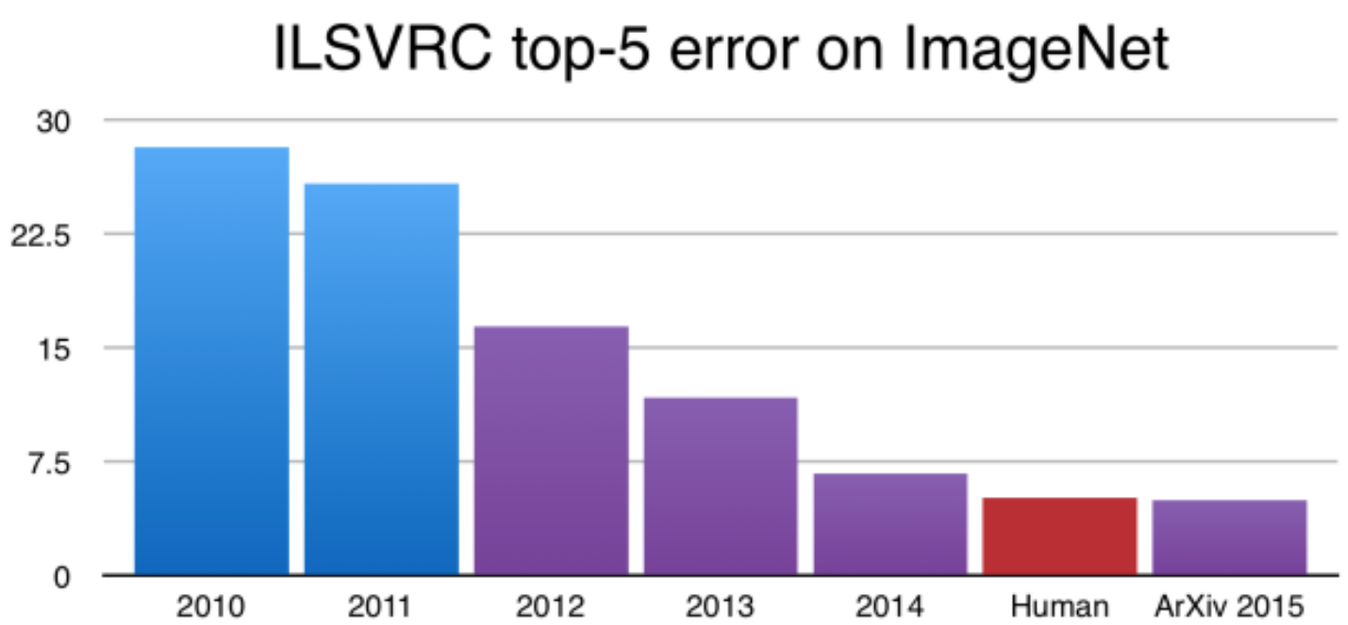
Can we rely on ML?



Can we rely on ML?



A Glimpse Into ML Reliability

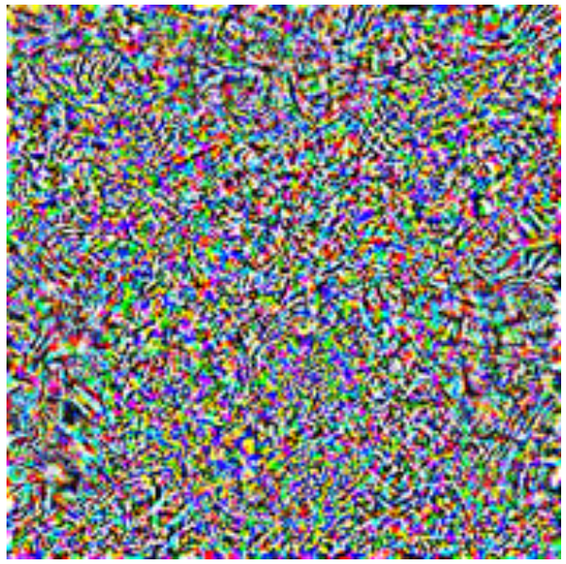


Have we *really* achieved human-level performance?

AI is more brittle than we think

[Szegedy et al 2013] [Biggio et al 2013]

+ 0.005 x



=

“pig” (91%)

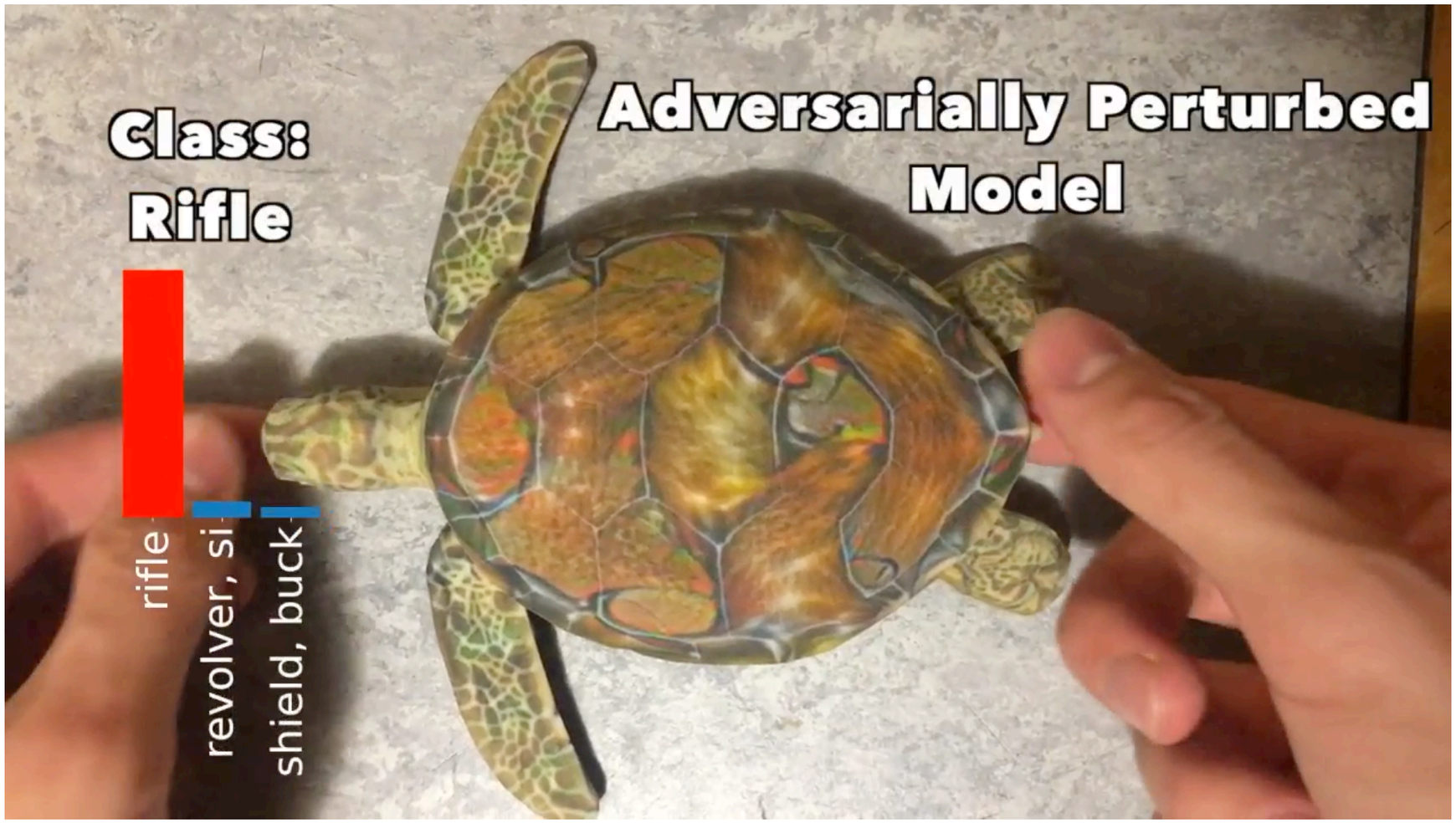
noise (NOT random)

“airliner” (99%)

AI is more brittle than we think

[Athalye Engstrom Ilyas Kwok 2018]

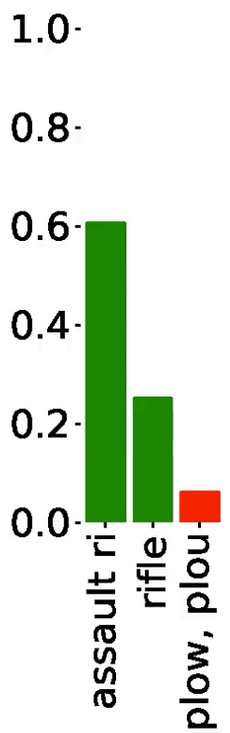
AI is more brittle than we think



[Athalye Engstrom Ilyas Kwok 2018]

AI is more brittle than we think

[Athalye Engstrom Ilyas Kwok 2018]

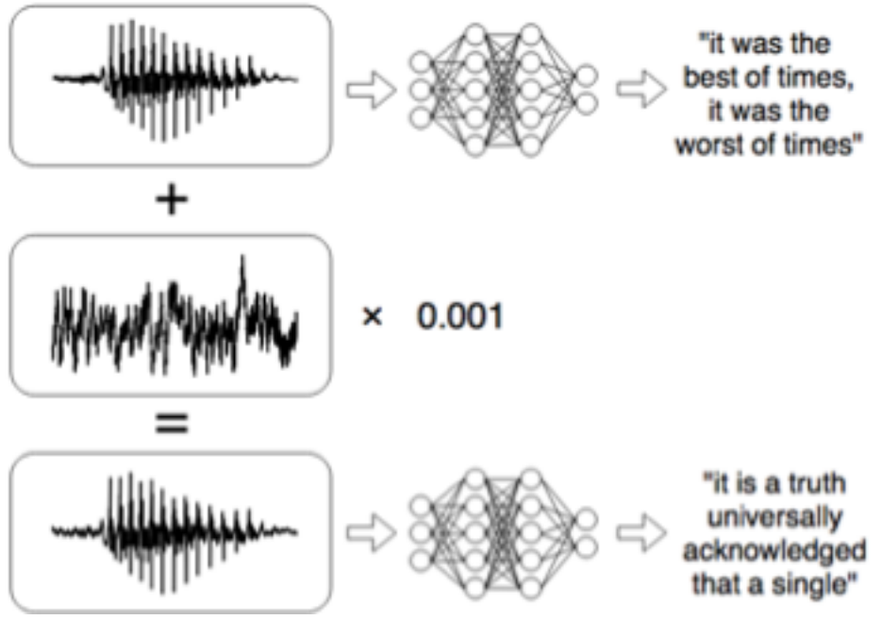


[Engstrom Tsipras Tran Schmidt M 2018]

Why should we care?

- **Security**

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



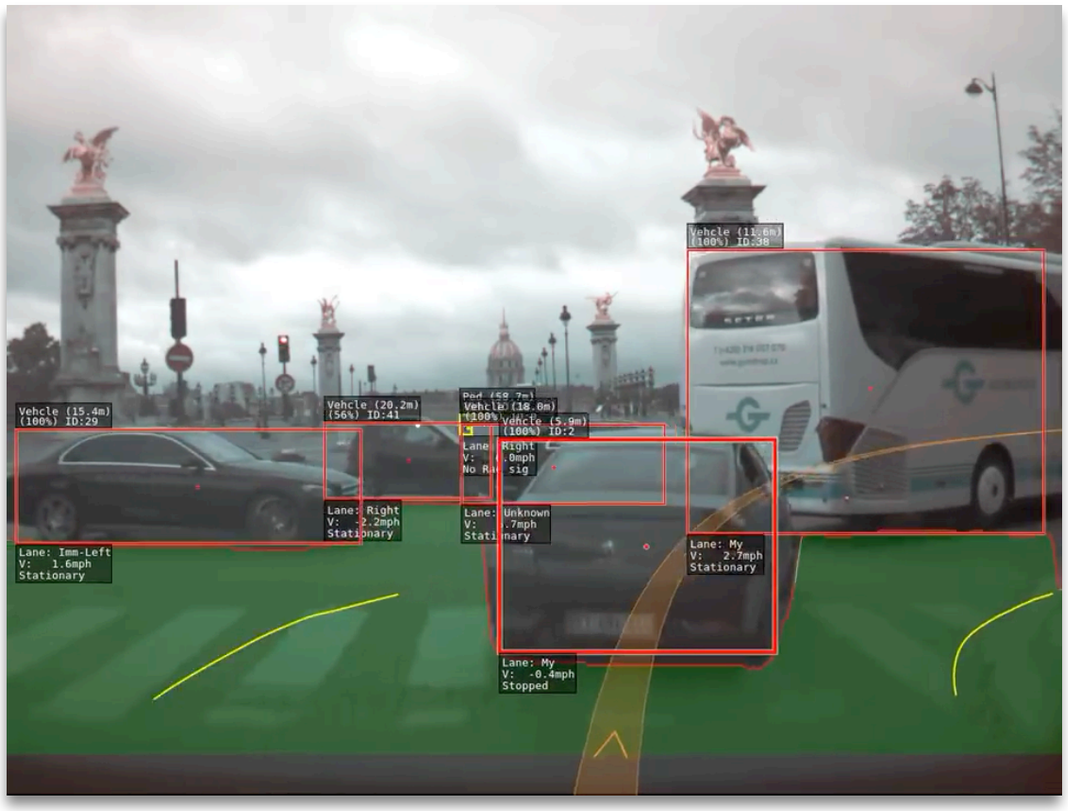
Voice commands that are unintelligible to humans
[Carlini Wagner 2018]

Why should we care?

- Security
- **Reliability**

Why should we care?

- Security
- **Reliability**

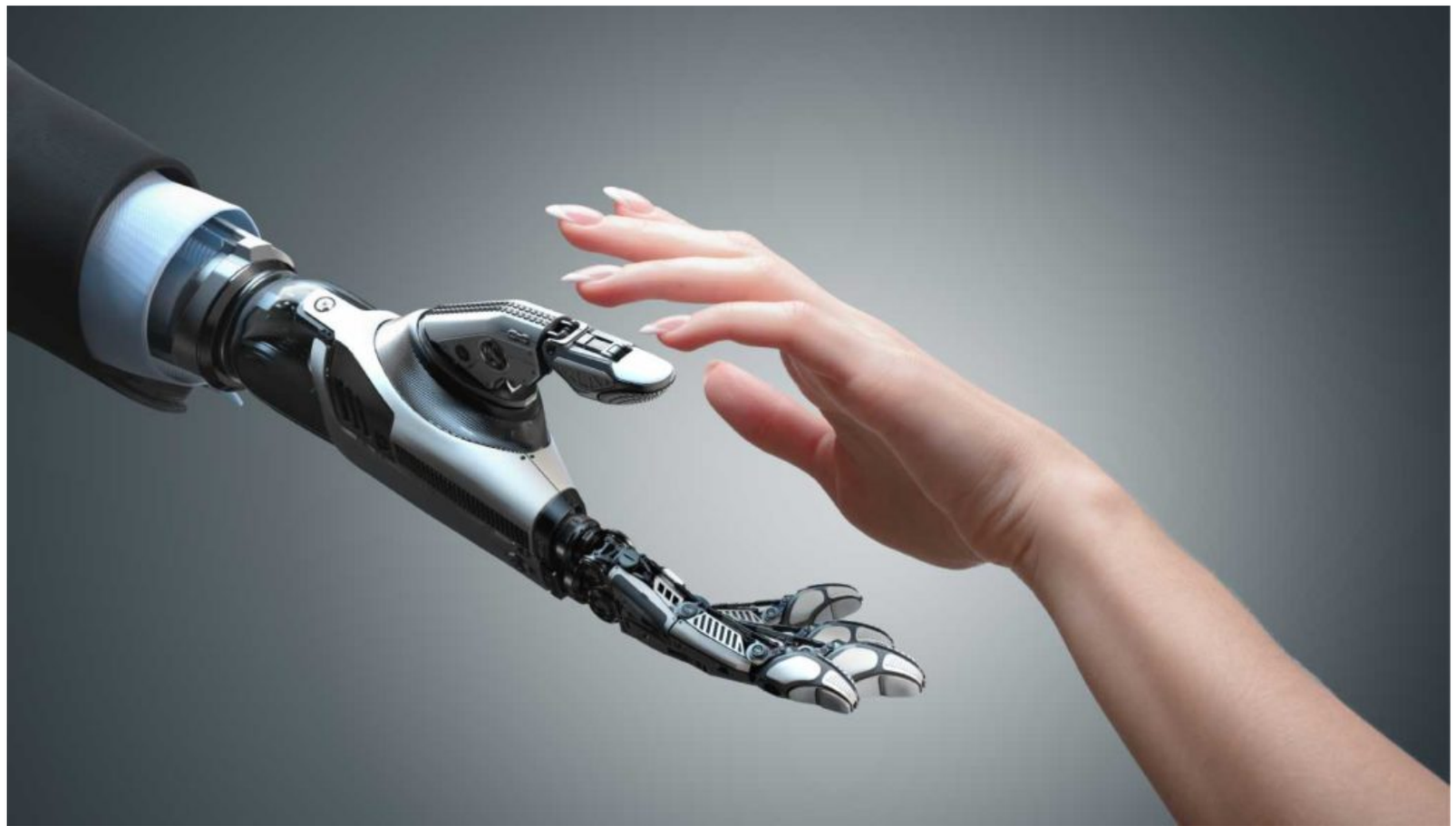


Why should we care?

- Security
- **Reliability**

Why should we care?

- Security
- Reliability
- **Alignment**



Need to understand the “failure modes” of ML

ML pipeline (via adversarial lens)

ML pipeline (via adversarial lens)

Data Collection

Training

Inference

Deployment

Untrusted data sources

Cat

Dog

Data poisoning

ML pipeline (via adversarial lens)

Data Collection

Training

Inference

Deployment

Untrusted data sources

Cat

Dog

Data poisoning

ML pipeline (via adversarial lens)

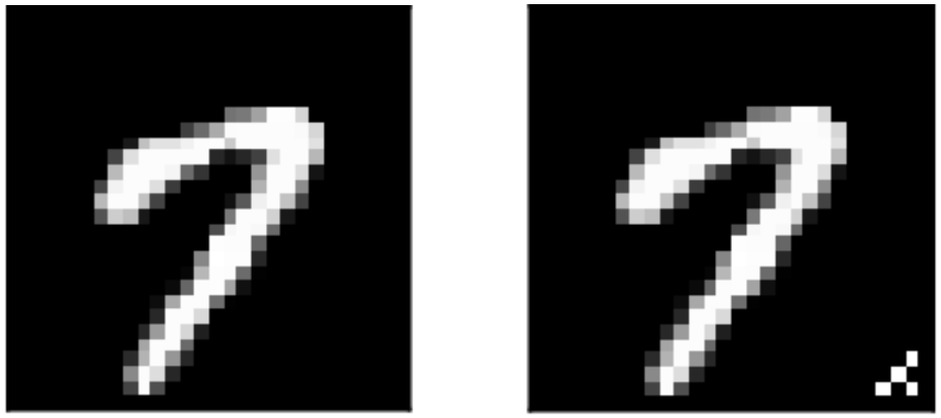
Data Collection

Training

Inference

Deployment

Classified as 7 **Classified as 5**



Original image Pattern Backdoor

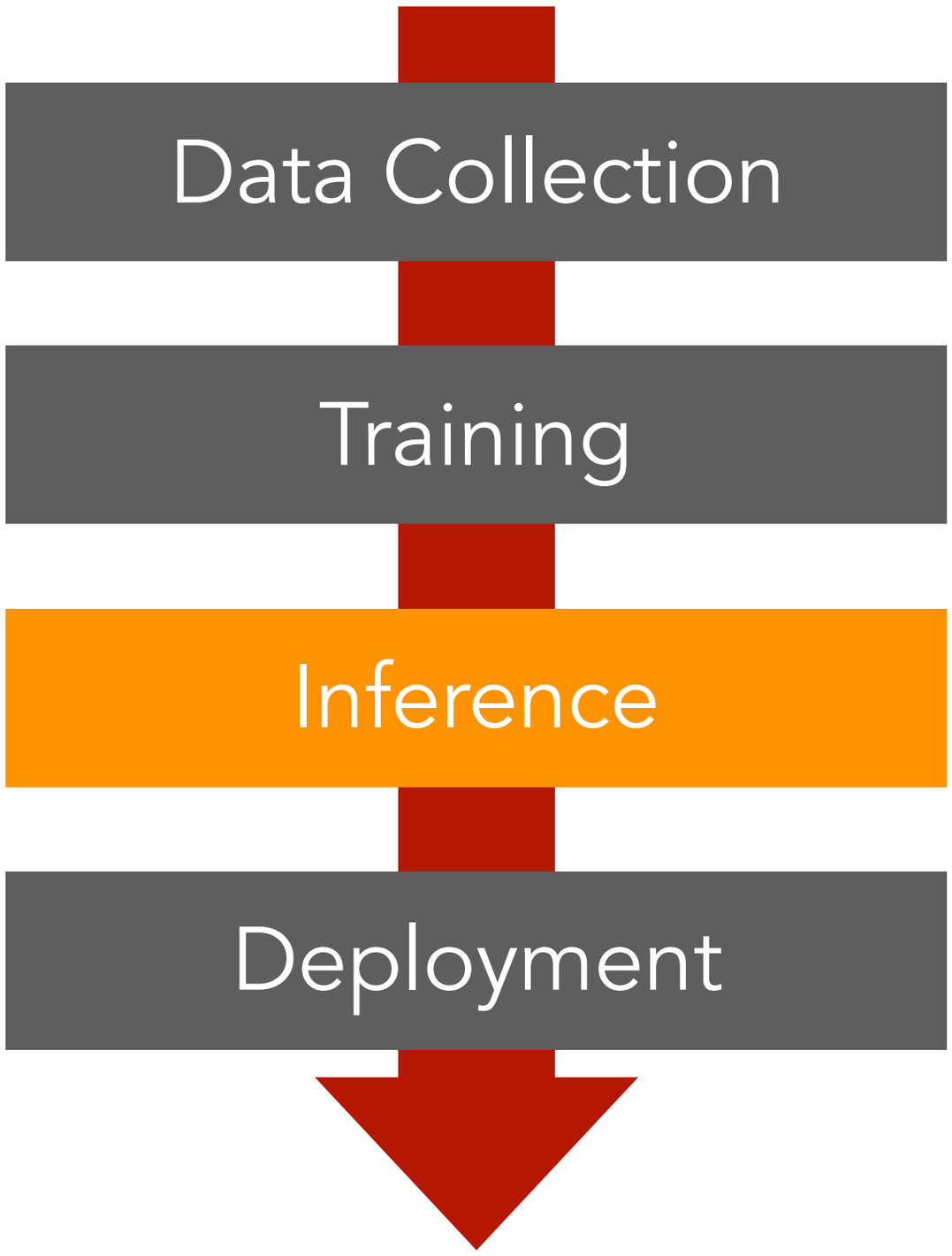
Classified as van **Classified as dog**

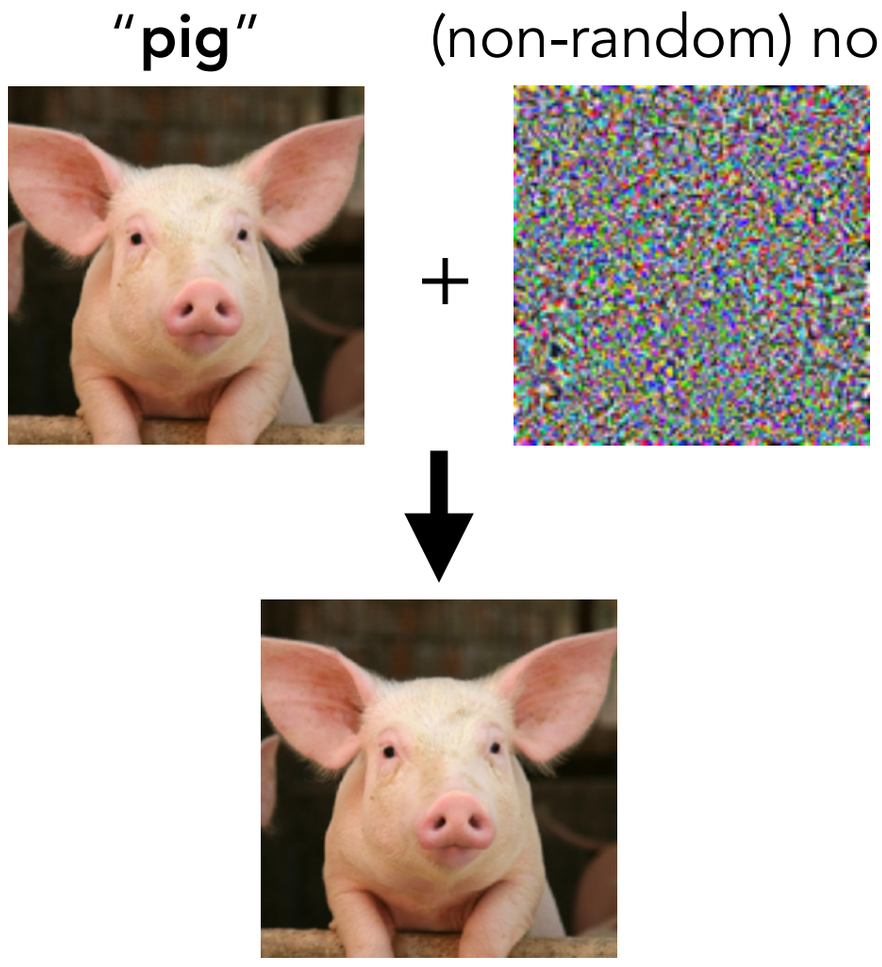
Outsourcing training

↓

Backdoor attacks

ML pipeline (via adversarial lens)





Adversarial examples

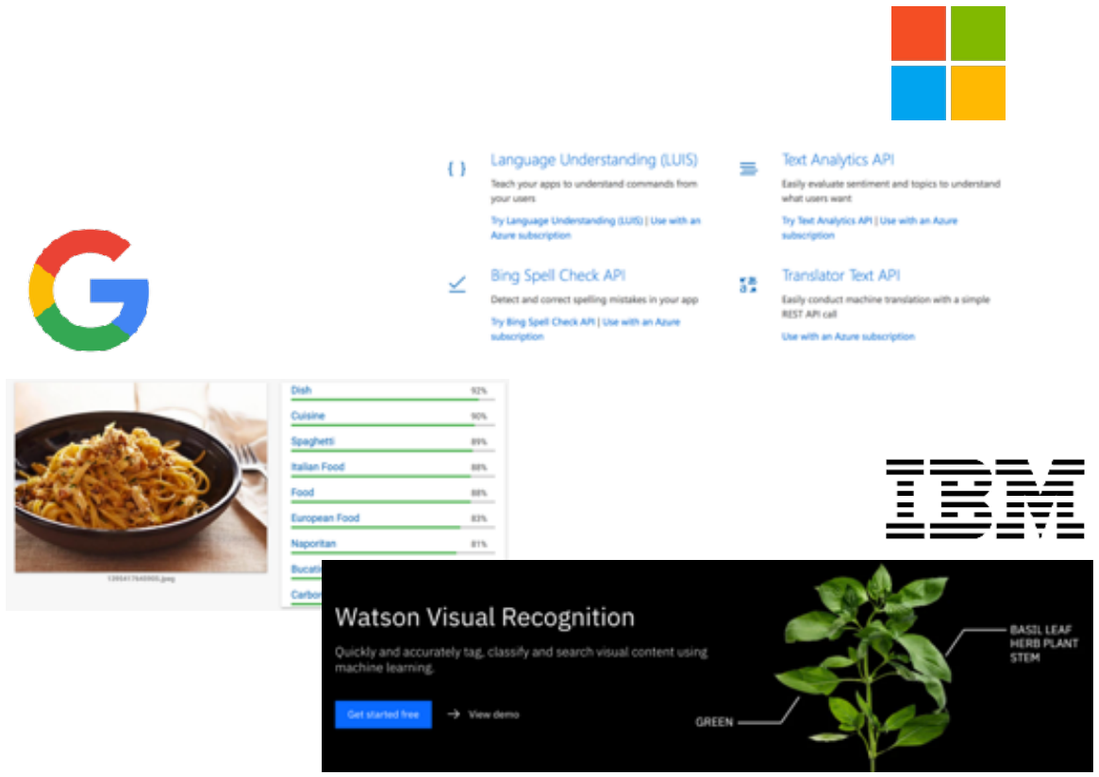
ML pipeline (via adversarial lens)

Data Collection

Training

Inference

Deployment



Exposing ML predictions

↓

Model stealing/
Extraction of (training) data

What do we do now?

Problem: Adversarial examples are not at odds with our current notion of generalization

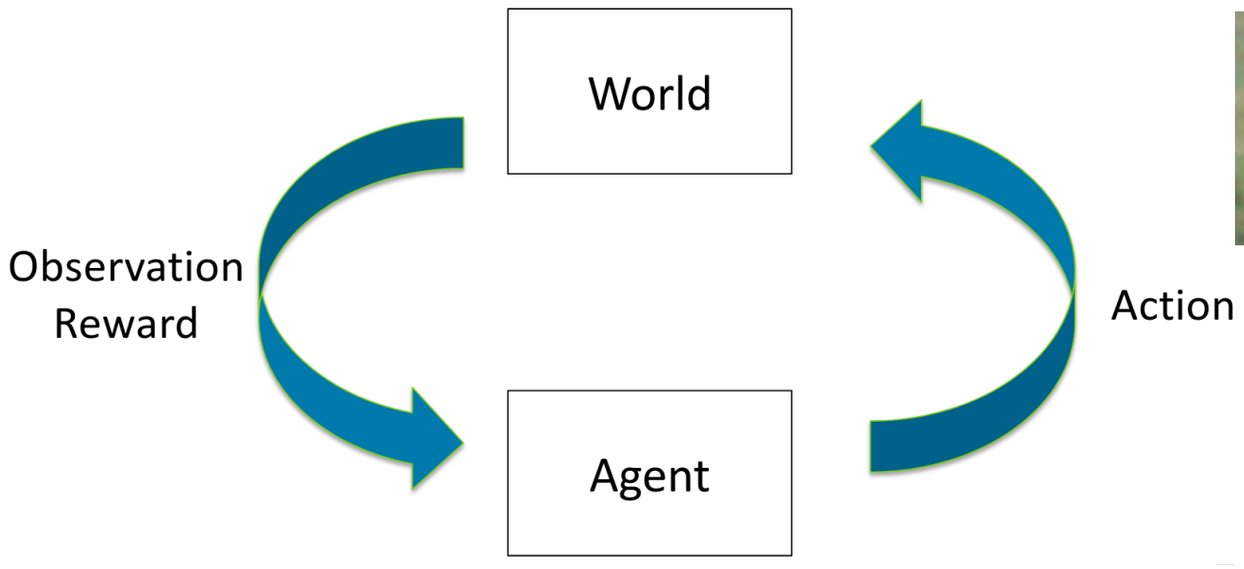
Maybe time to re-think what we want in generalization?

Again: This is not only about security but also about understanding how ML/deep learning works (and fails!)

Question III:

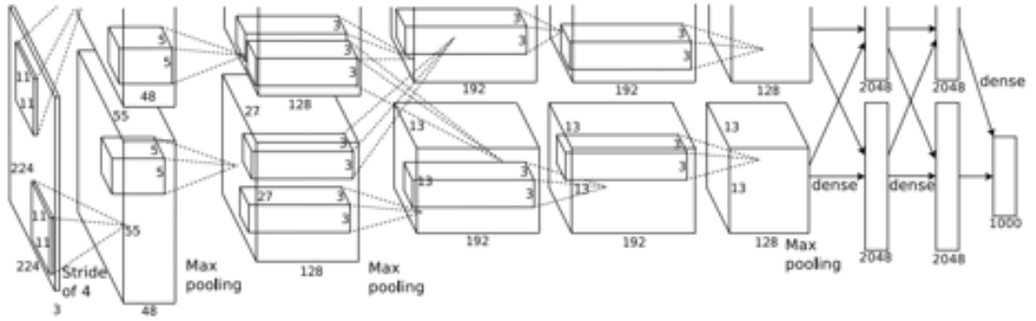
Is ML ready for being "in the loop"?

Reinforcement Learning (RL)



What if the agent is a (deep) neural network?

Goal: Maximize the reward



Questions:

- How to train such agent (exploration vs. exploitation)?
- What are the fundamental limits on efficiency of this approach?
- How to ensure that the agent does what we intend it to do?

Question IV:

What are the societal impacts of ML?

ML is entering every aspects of our life

- Should we be worried?
- Potential concerns:

  → Interpretability (Can we understand ML models?)

  → Reliability (Can I trust the prediction of an ML model?)

  → Fairness (Is the ML model behaving in a “fair” way?)

  → Privacy (Is the ML model protecting our privacy?)

  → AI Safety (If we build a super-human AI, will it destroy us?)

  → (Your suggestion here)

Now: Onto Optimization