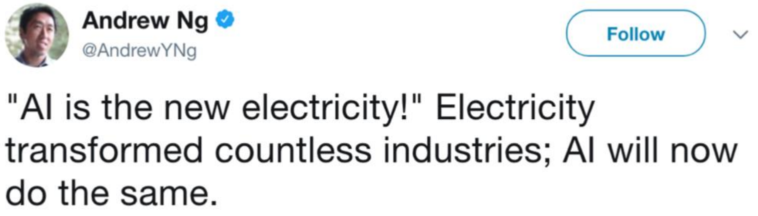
Machine Learning Security

Earlence Fernandes

CS 642

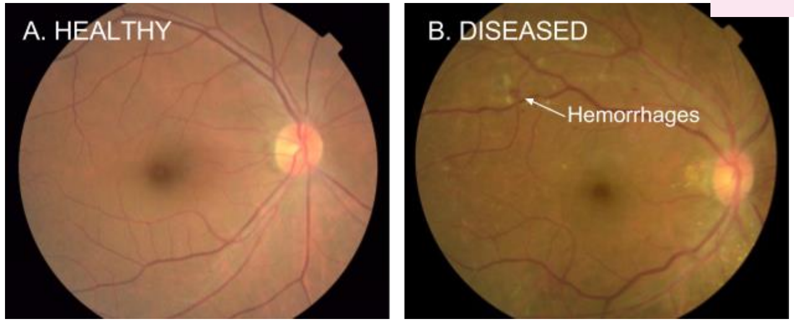
Some slides are borrowed from Chatterjee, Fernandes, Jha, and Mądry

Deep
~~Machine~~ Learning Revolution



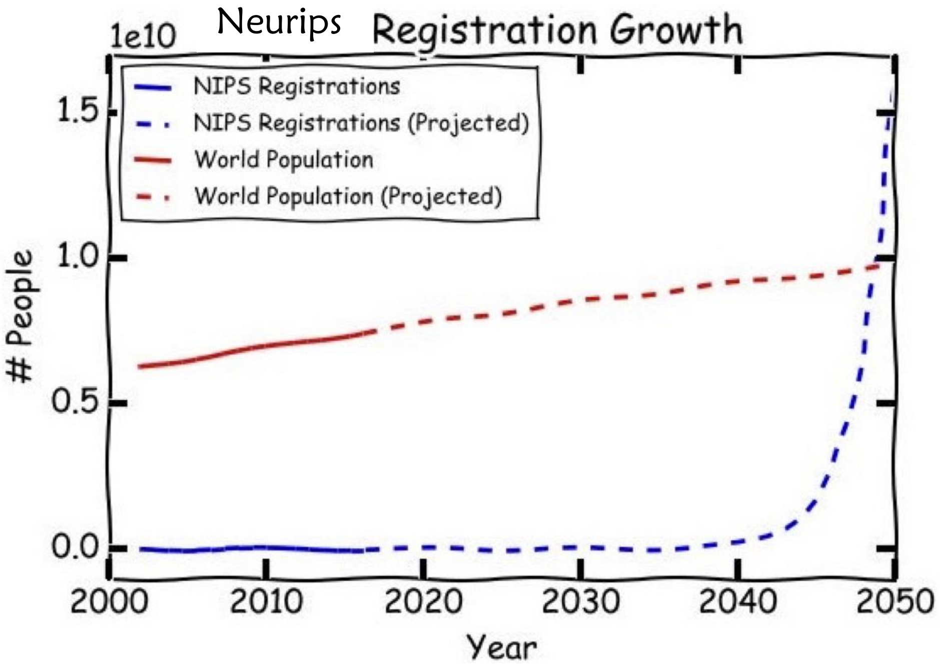
Transportation

DeepMind AI Reduces Google Data Centre Cooling Bill by 40%



Healthcare

Source: Peng and Gulshan (2017)



AI vs ML

AI is decision making, ML is learning how to do that (from data)

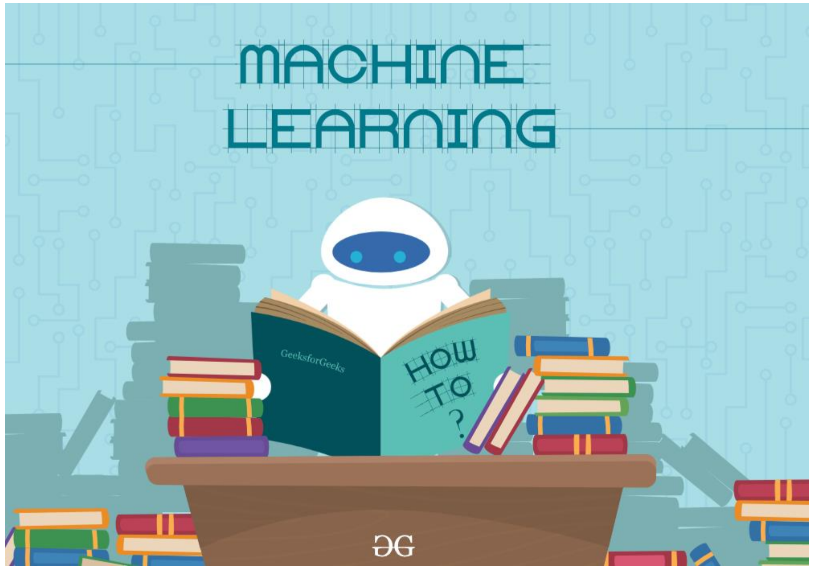
Nowadays they are basically interchangeable

Machine Learning: What is it good for

Teach machine to do tasks that are simple to us

- Image recognition
- Speech recognition
- Translation
- Knowledge synthesis
- Conversation
- Driving cars
- …

But why?
Let machine do the "chores"



And some complex task

- Predict weather
- Atomic interactions!

Deep learning for chemical reaction prediction

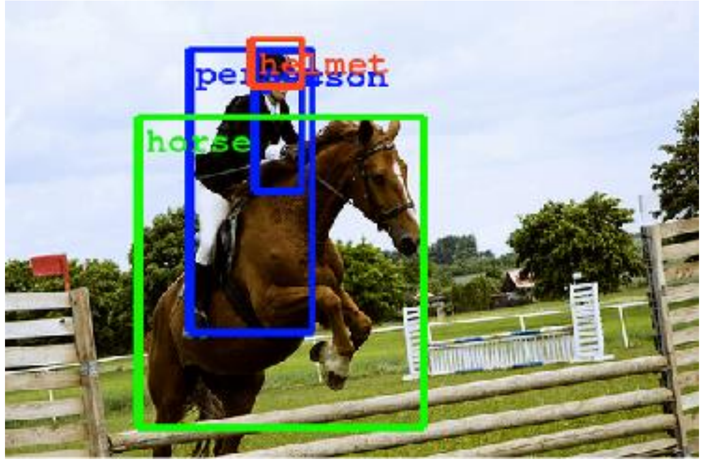
ML beating doctors ☺

- NOVEMBER 15, 2017
  - Stanford algorithm can diagnose pneumonia better than radiologists

- April 14, 2017
  - Self-taught artificial intelligence beats doctors at predicting heart attacks

- ….

ML reached “human-level performance” on many IID tasks circa 2013



(Szegedy et al, 2014)

...recognizing objects and faces....

(Taigmen et al, 2013)

(Goodfellow et al, 2013)

...solving CAPTCHAS and reading addresses...

(Goodfellow et al, 2013)

Cool! But why should I care …

- ML is used in security
- ML is being (or going to be) used everywhere
  - often in mission critical settings
  - ML models get “compromised”

ML in security application

Security is often differentiating the good from the bad

- Malware detection
- Spam detection
- Intrusion detection
- Fraud detection
- Cyber defense

- Hate speech detection? Illicit content detection?

ML 101

- Generative
  - $f_\theta(r) \to x$; r is a random string,
  - $x \in \mathcal{X}$ some distribution, say images of cat, or songs of Led Zeppelin

- Discriminative
  - $f_\theta(x) \to y$
  - Given input x **predict** what is the possible output y

Terminologies

Model

Input
(indep. var.)

$$f_\theta(x) = y$$

Output/prediction
(dep. var.)

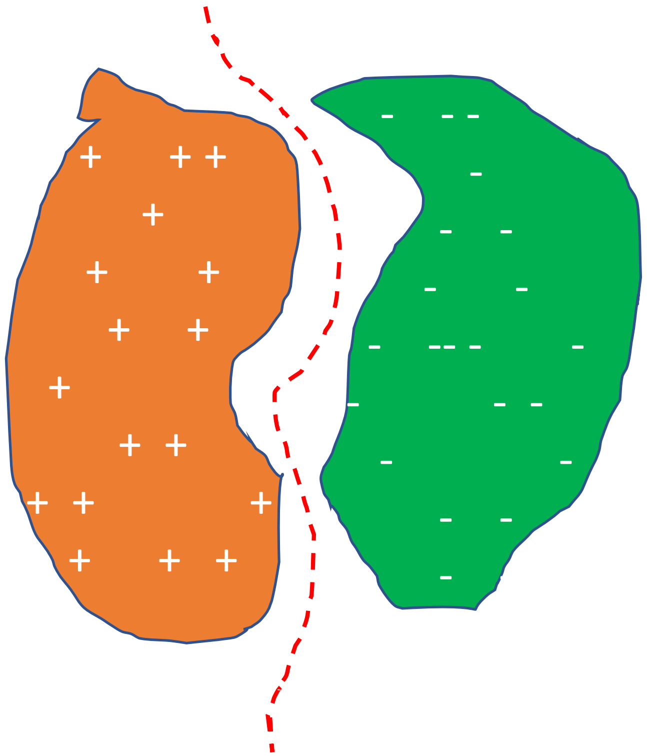
Parameter

Training algorithm

$$\mathcal{L}_f(\mathcal{T}) \rightarrow \theta$$

Training data

Supervised Machine Learning



f^* = Some concept you want the machine to learn

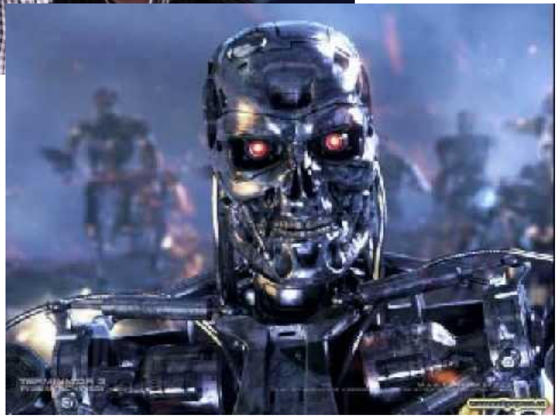
Choice of $f(\cdot)$ is crucial!

- Too strict: underfitting
- Too flexible: overfitting

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

Security of ML

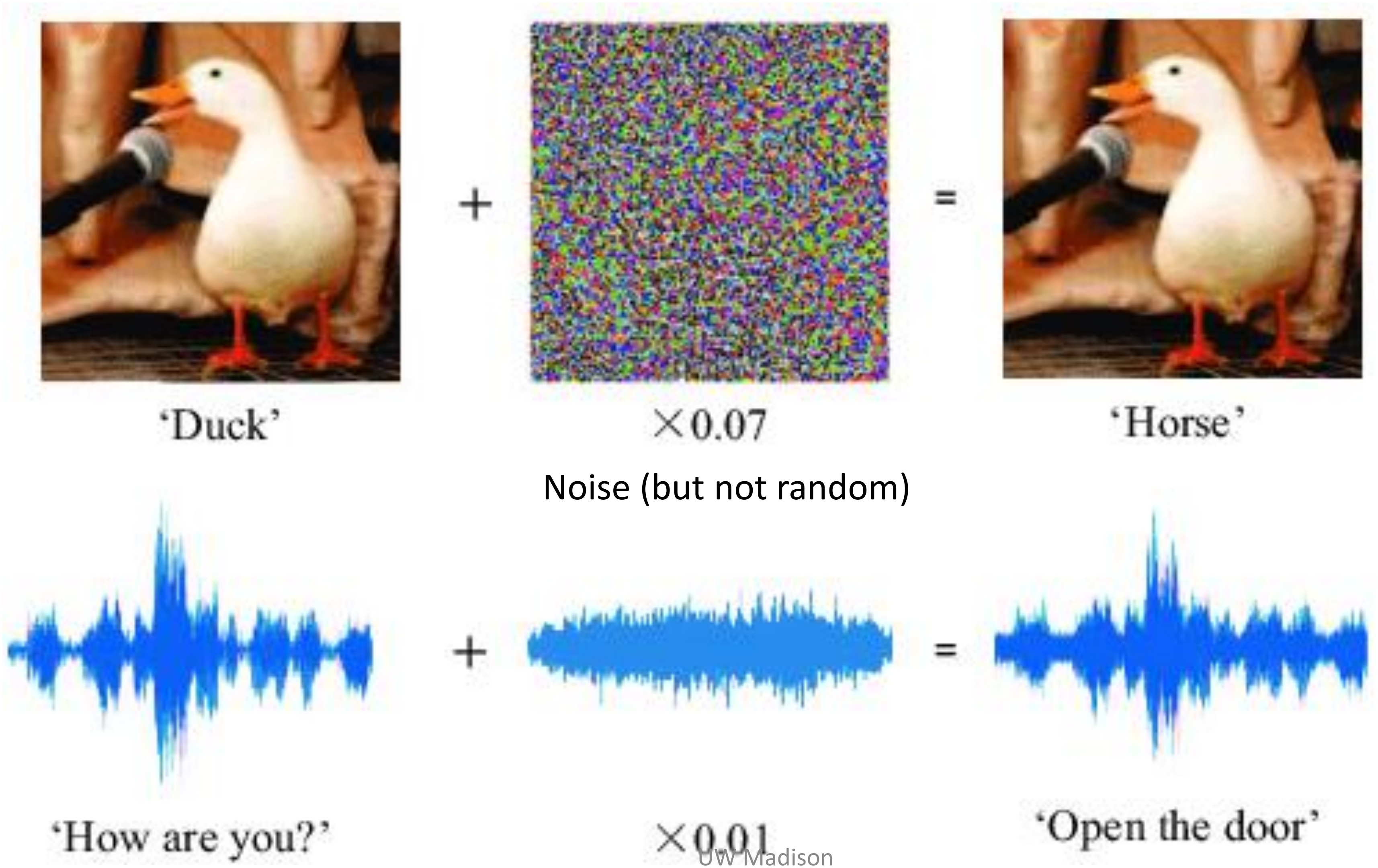
Can we rely




#NEWS
GOOGLE SELF DRIVING CAR
CRASHES INTO A BUS


GRAB
SPX Index
95) Save As...
96) Actions
97) Edit
Intraday Price Chart
1D 3D 1M 6M YTD 1Y 5Y Max Tick
Security/Study
Event
Spoofing
Flash Crash
1130
1120
1115.81
-17.06
-1.51%
1100
1090
1080
1070
Start: 05/06 14:30 End: 05/06 15:00
Last Price 1115.8
Close on 05/05 ---- 1165.8
Start on 05/06 14:30 1132.87
High on 05/06 14:30 1133.38
Average 1112.00
Low on 05/06 14:46 1065.79
14:35
14:40
14:45
14:50
14:55
06 May 2010

Adversarial Example



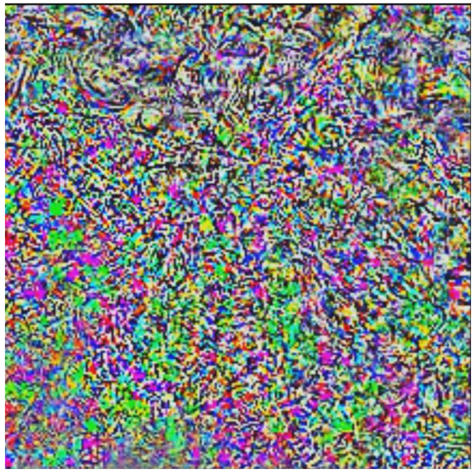
Noise (but not random)

Szegedy et al, 2013

Carlini et al, 2018

Deep Neural Networks are Useful, But Vulnerable

+ ε



=

Image Courtesy: adversarial-ml-tutorial.org

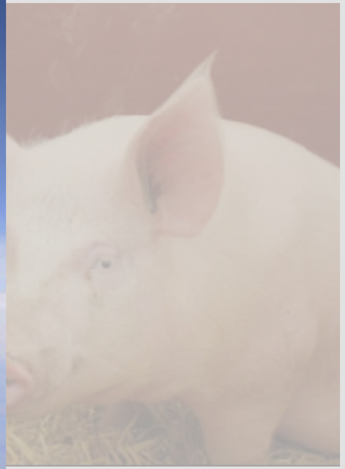
"pig"
99.6% confidence

"airliner"
96.7% confidence

Explaining and Harnessing Adversarial Examples, Goodfellow et al., arXiv 1412.6572, 2015

Deep Neural Networks are Useful, But Vulnerable

“pig”
99.6% confidence



airliner”
96.7% confidence

Image Courtesy:
adversarial-ml-
tutorial.org

Explaining and Harnessing Adversarial Examples, Goodfellow et al., arXiv 1412.6572, 2015

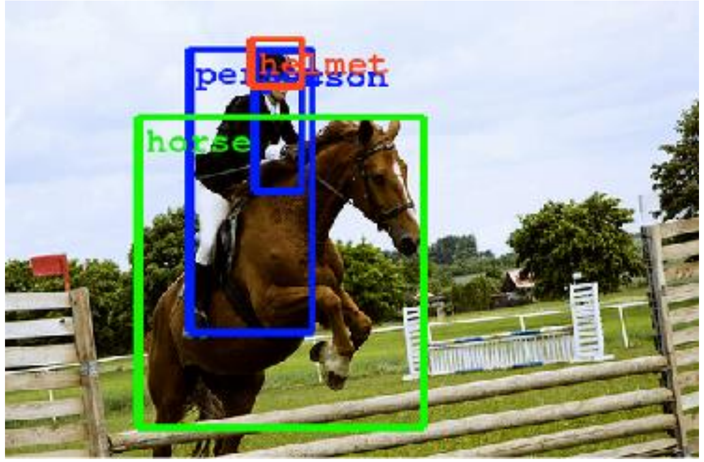
Adversarial ML (AML)

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

Don’t Bring Your Turtle to a Gun Fight

https://www.csail.mit.edu/news/fooling-neural-networks-w3d-printed-objects, 2018

ML reached "human-level performance" on many IID tasks circa 2013



(Szegedy et al, 2014)

...recognizing objects and faces....

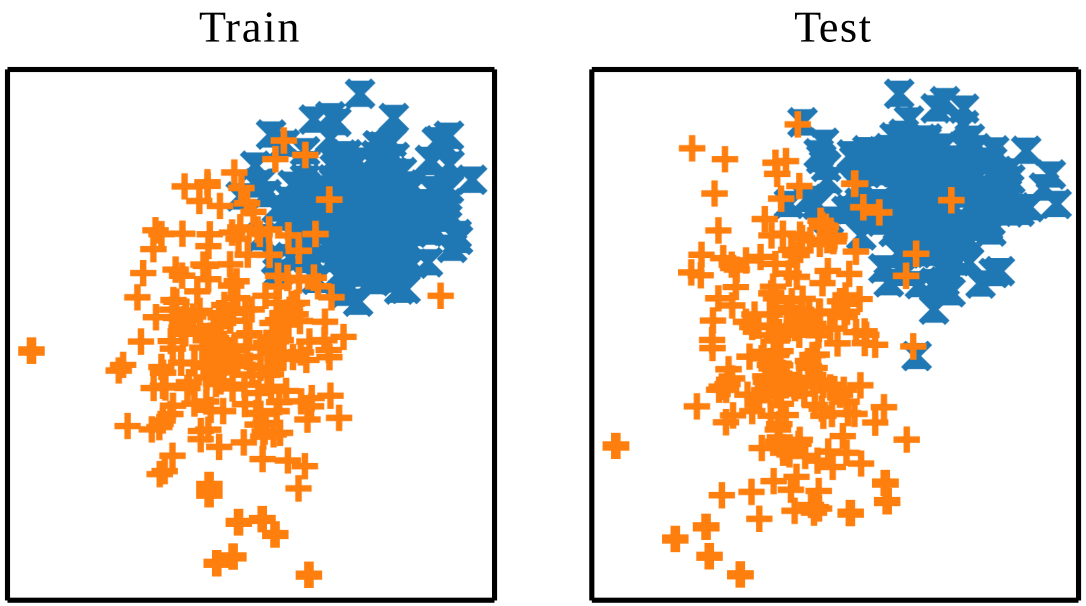
(Taigmen et al, 2013)

(Goodfellow et al, 2013)

...solving CAPTCHAS and reading addresses...

(Goodfellow et al, 2013)

I.I.D. Machine Learning



I: Independent
I: Identically
D: Distributed

All train and test examples drawn independently from same distribution

Security Requires Moving Beyond I.I.D.

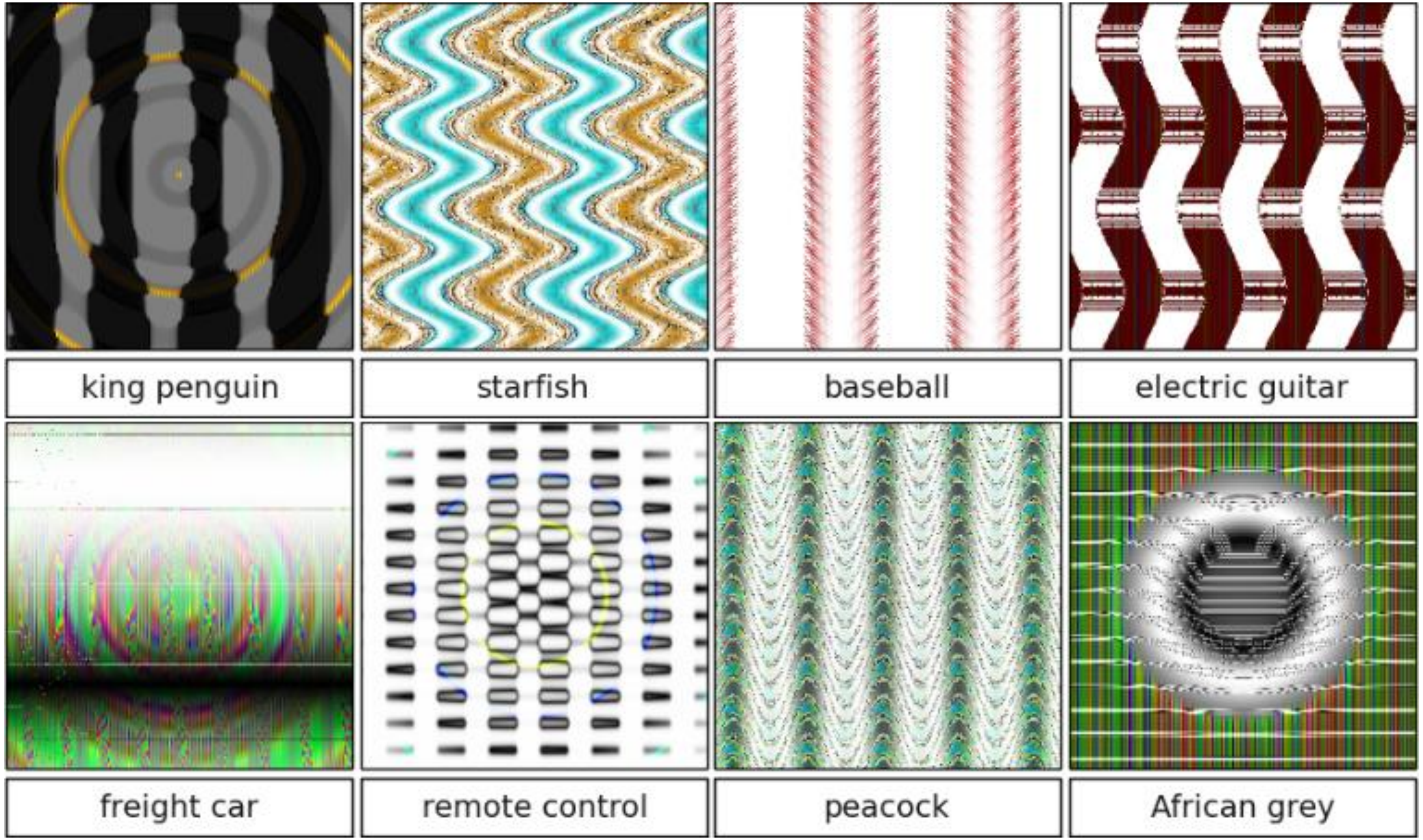
- Not identical: attackers can use unusual inputs

(Eykholt et al, CVPR 2017)

- Not independent: attacker can repeatedly send a single mistake (“test set attack”)

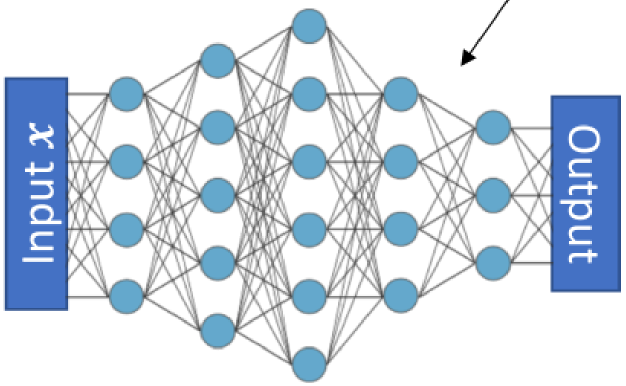
Adversarial Example (and more)

Who knows what these ML models are learning?!?



Where Do Adversarial Examples Come From?

Differentiable



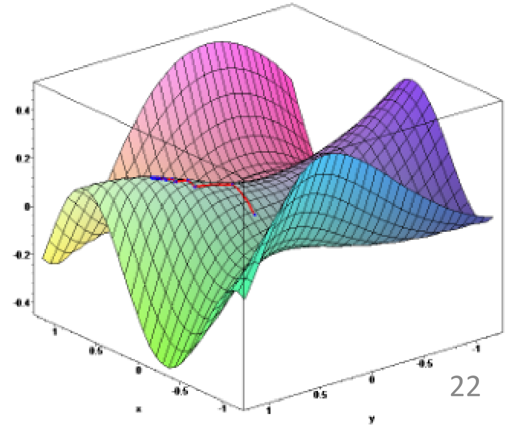
Parameters $\boldsymbol{\theta}$

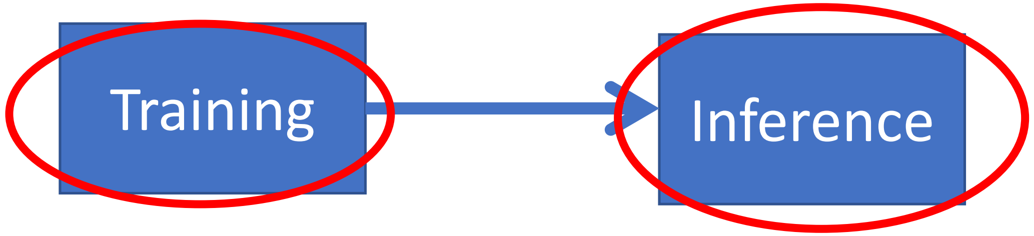
Model Parameters Input Correct Label

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

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

Can use gradient descent method to find good θ




Training
Inference

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

What can go wrong?

Data poisoning attack

Change the decision boundary

Make creating Adv. example easy

Or help facilitate other attacks (we explain later)

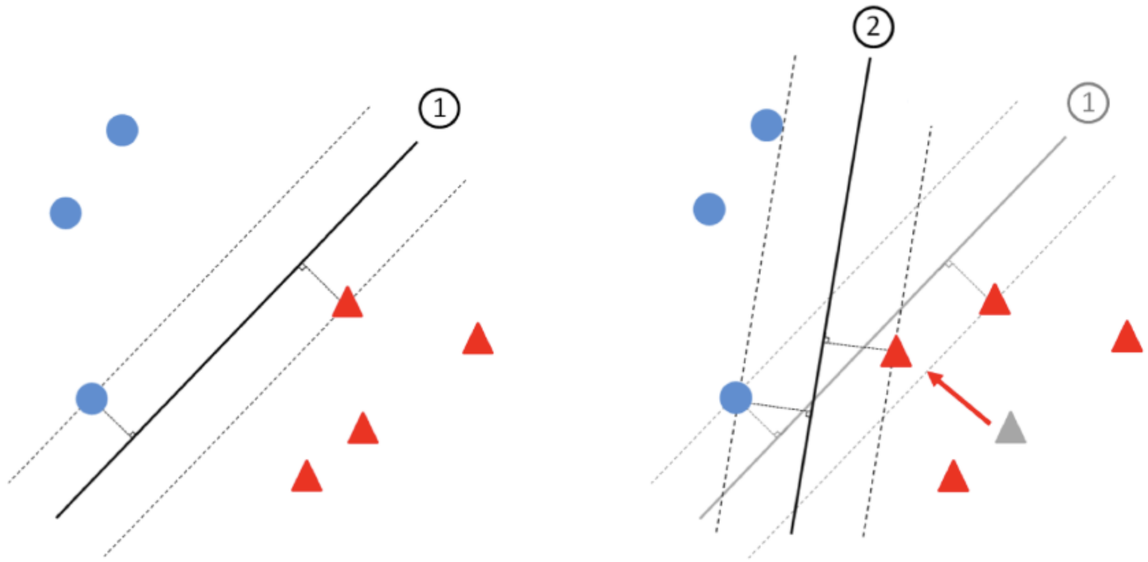
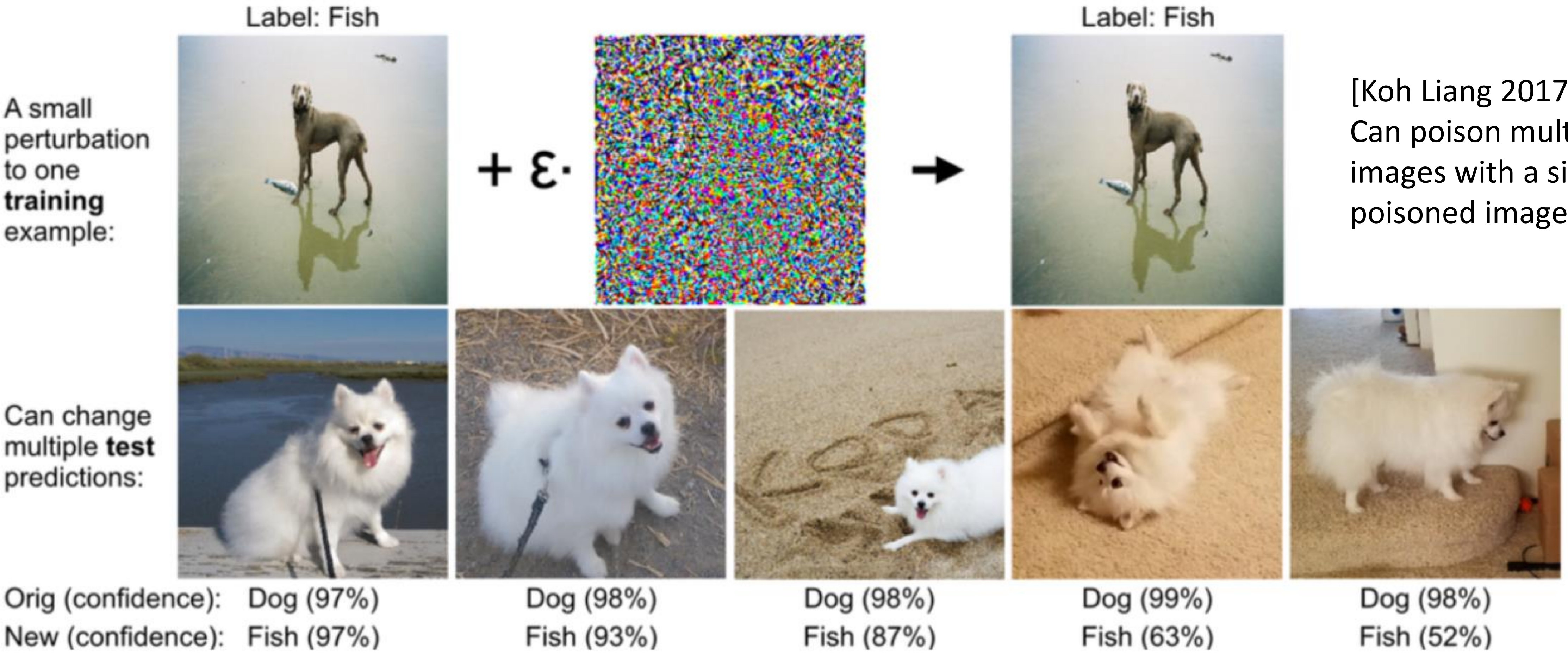


Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted if just one training sample is changed, even when that sample's class label does not change (right).

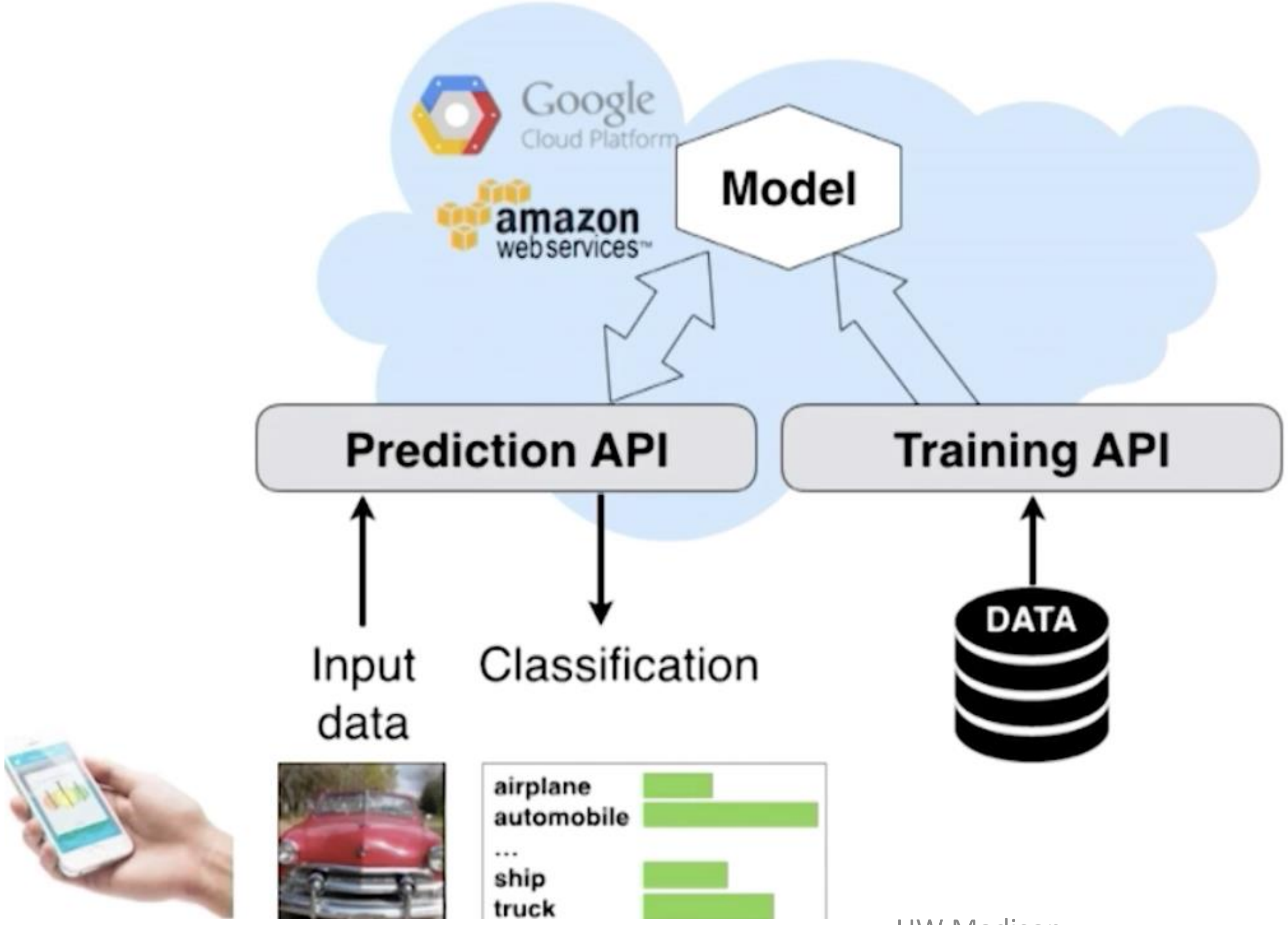
Training get worse w/ bad data.



[Koh Liang 2017]:
Can poison multiple images with a single poisoned image

Remember Tay

Deep learning is also resource hungry



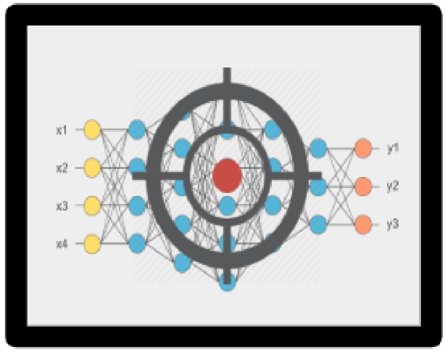
Can we trust with our data?
Can we trust with our model?

Membership inference

- Privacy of the training data!

sample of data

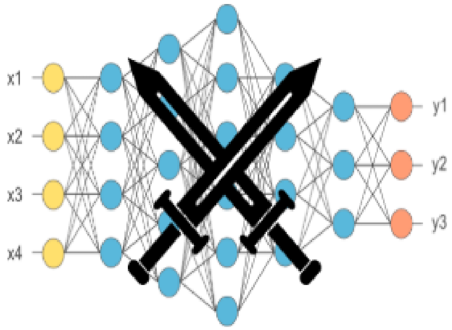
target network with black box access



classification prediction (probability vector)

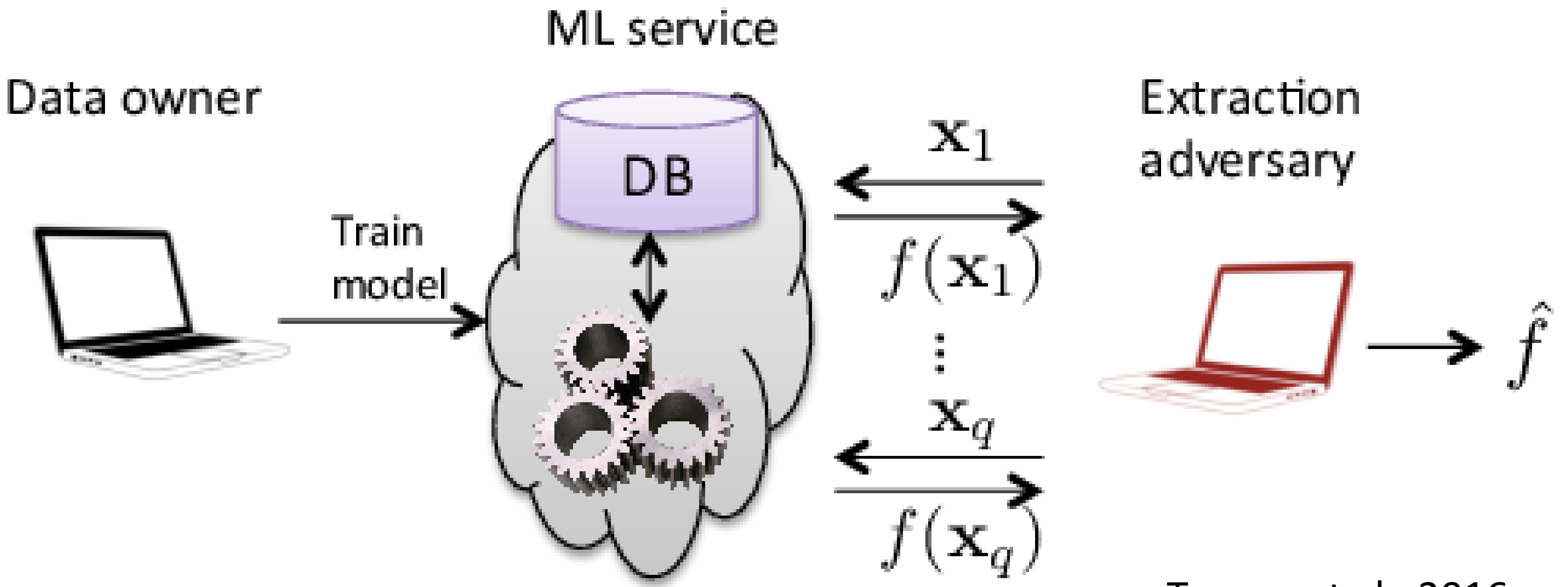
$$\begin{bmatrix} 0.84 \\ 0.12 \\ 0.04 \end{bmatrix}$$

attack network



binary membership prediction (in/out)

Model stealing attack

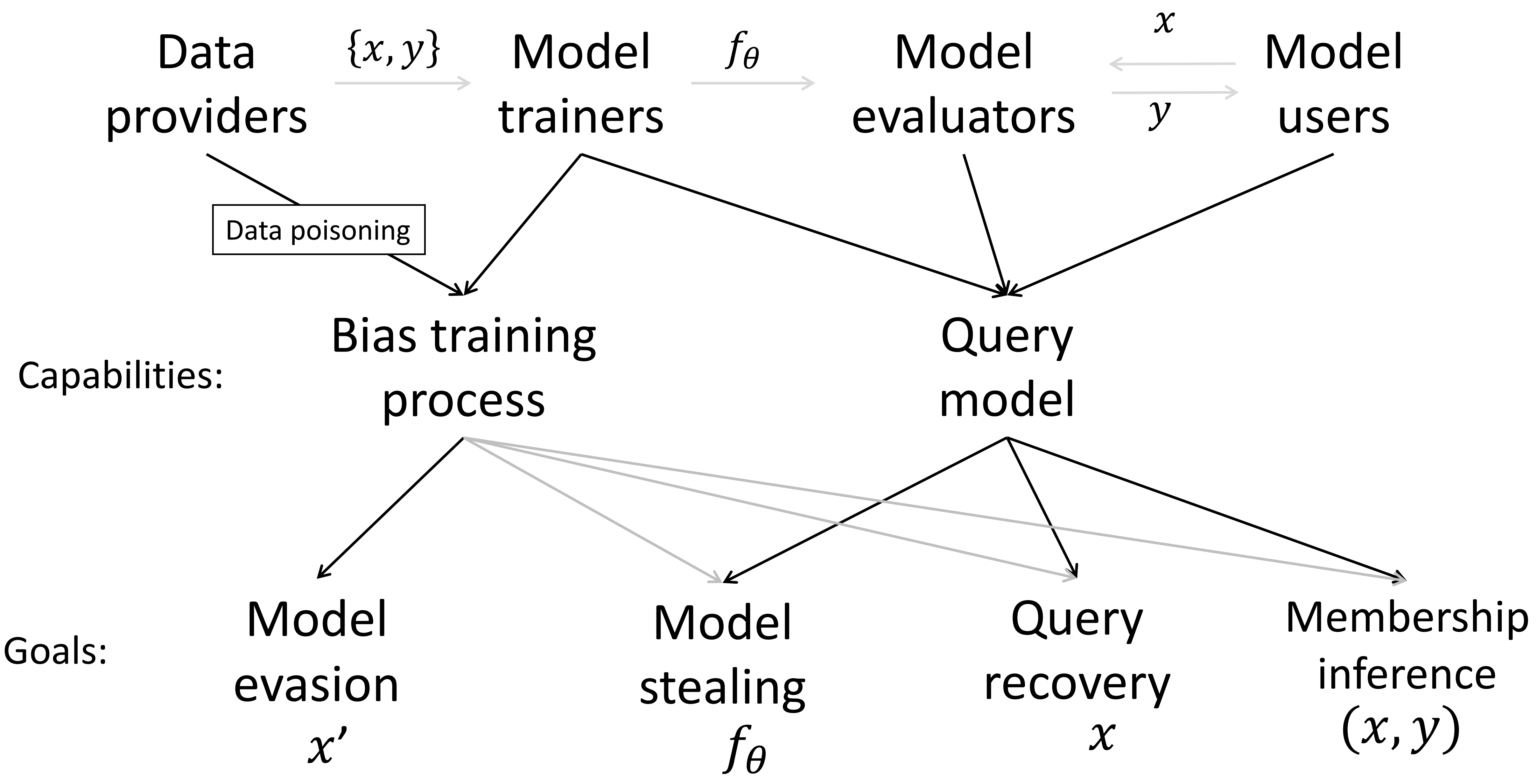


Tramer et al., 2016

Let’s systematize (Science of Secure ML)

- Entities in a ML system

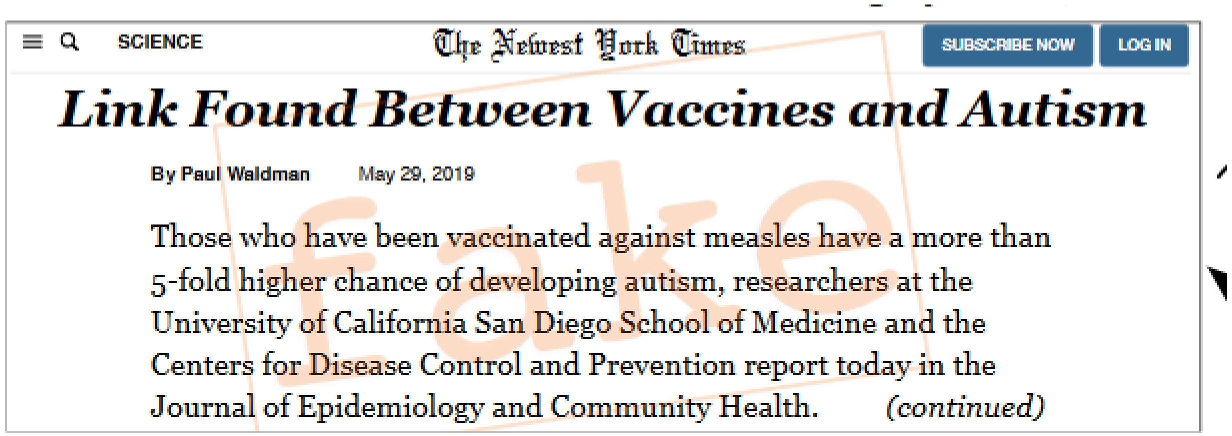
$$
\begin{tikzcd}
\text{Data providers} \arrow[r, "\{x, y\}"] & \text{Model trainers} \arrow[r, "f_\theta"] & \text{Model evaluators} \arrow[r, shift right, "y"'] & \text{Model users} \arrow[l, shift right, "x"']
\end{tikzcd}
$$


Data providers
$\{x, y\}$
Model trainers
f_θ
Model evaluators
x
y
Model users
Data poisoning
Capabilities:
Bias training process
Query model
Goals:
Model evasion x'
Model stealing f_θ
Query recovery x
Membership inference (x, y)

Other security concern with ML


'Dangerous' AI offers to write fake news

By Jane Wakefield
Technology reporter


SCIENCE

The Newest York Times

SUBSCRIBE NOW

LOG IN

Link Found Between Vaccines and Autism

By Paul Waldman May 29, 2019

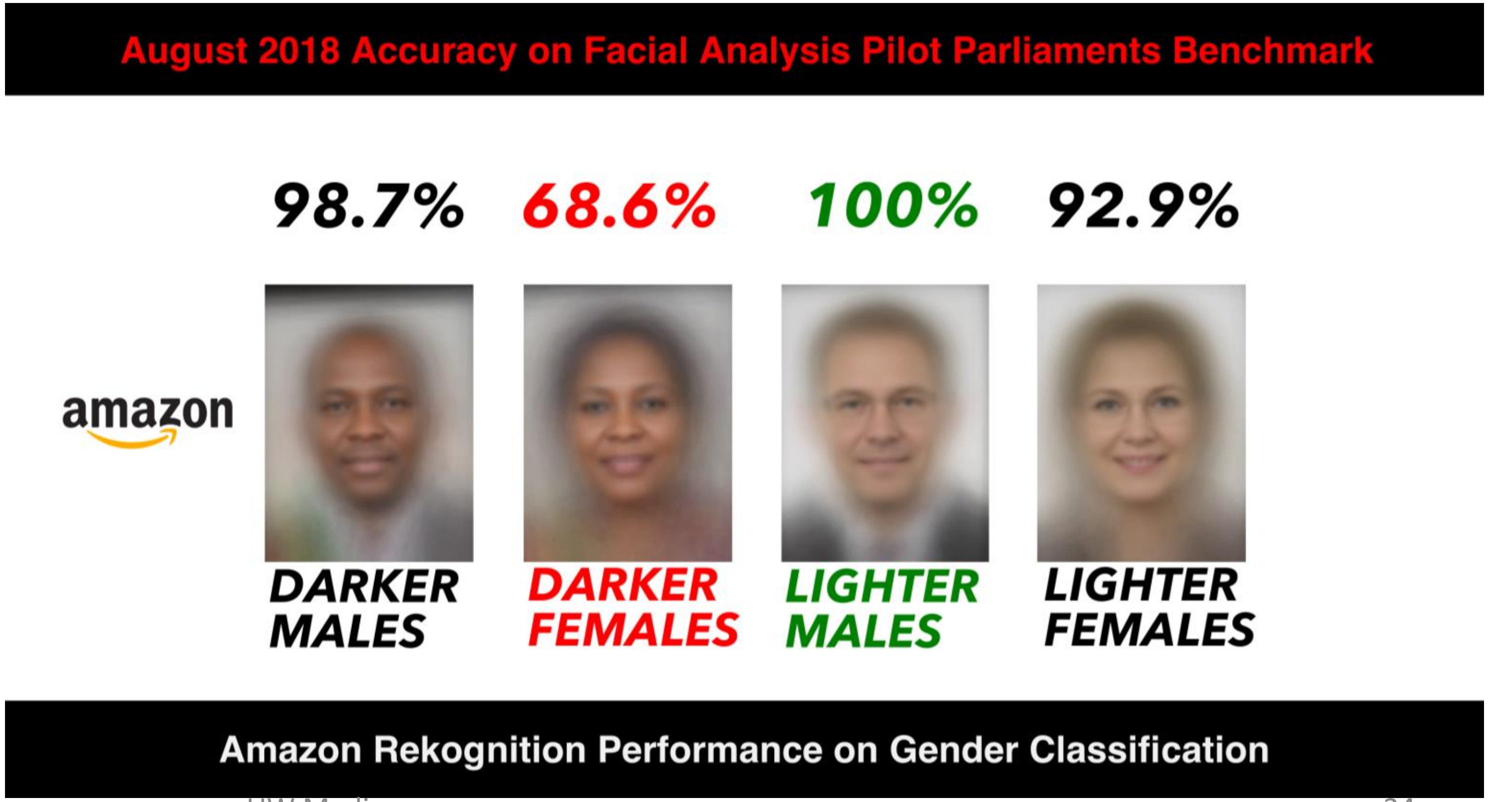
Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism, researchers at the University of California San Diego School of Medicine and the Centers for Disease Control and Prevention report today in the Journal of Epidemiology and Community Health. *(continued)*

Grover - A State-of-the-Art Defense against Neural Fake News

Obama’s Fake Video

ML and fairness / bias

- How do we ensure ML model is biased towards one of the protected classes?



Inaudible voice commands [Zhang et al. 2017]

Ultrasonic voice commands for smart assistants

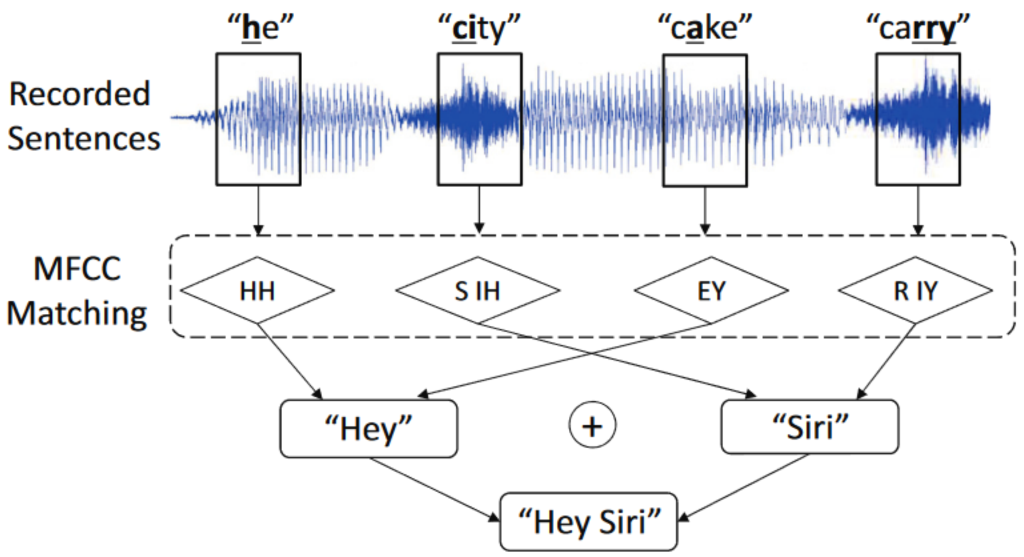
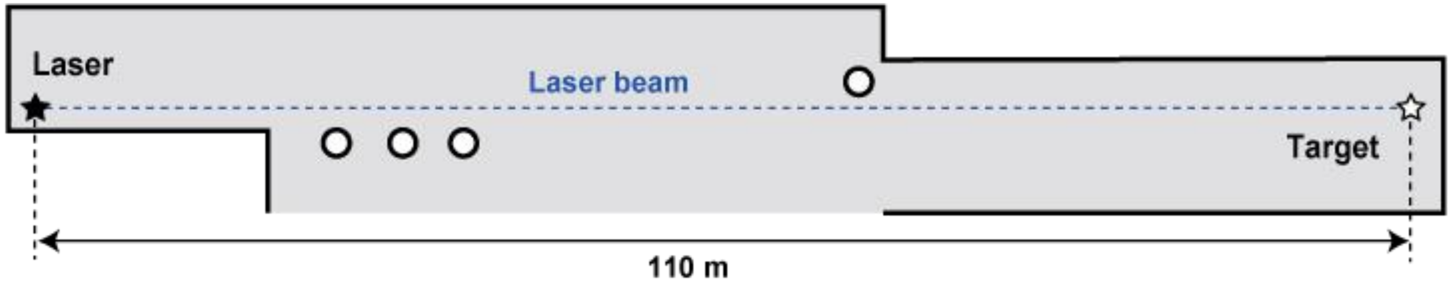


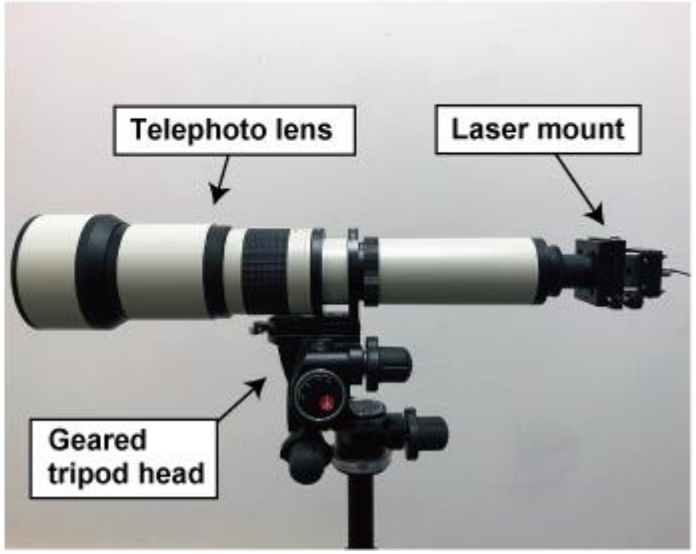
Figure 8: Concatenative synthesis of an activation command. The MFCC feature for each segment in a recorded sentence is calculated and compared with the phonemes in the activation command. After that, the matched voice segments are shuffled and concatenated in a right order.

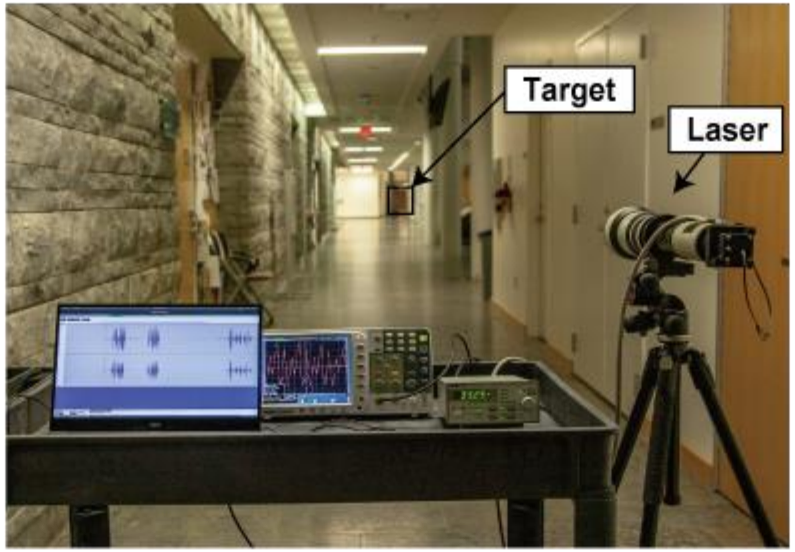
Light command!

[Sugawara et al. this month]

Why stop at voice …







Future?

Robust ML

- Adversarial training
- Training assuming there will be adversarial inputs

- Privacy aggregation of Teacher Ensembles (PATE)
- Differentially private ML