Adversarial Machine Learning (AML)

Somesh Jha

University of Wisconsin, Madison

ICISS 2018 (Bangalore)

Thanks to Nicolas Papernot, Ian Goodfellow, and Jerry Zhu for some slides.

Thanks

- Collaborators…

- NSF
  - SaTC Frontiers Grant (Penn State, UCSD, Stanford, Virginia, Berkeley, Wisconsin)
  - https://ctml.psu.edu/
  - FMitF

- ARO

Announcements/Caveat

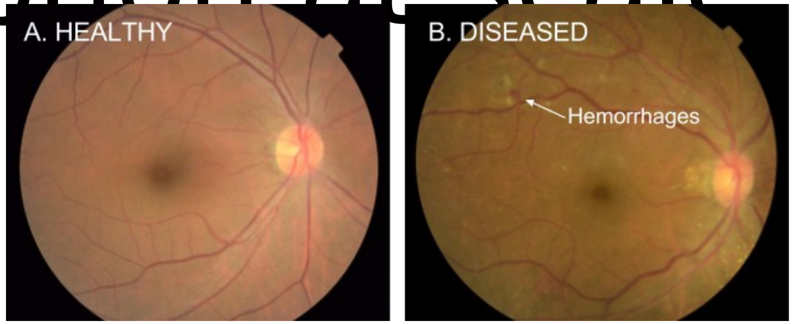
- Please ask questions during the talk
  - If we don't finish, fine

- More slides than I can cover
  - Lot of skipping will be going on

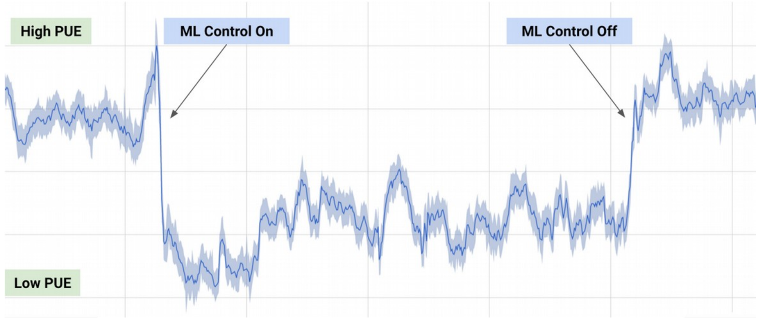
- Fast moving area
  - Apologies if I don't mention your paper

- Legend

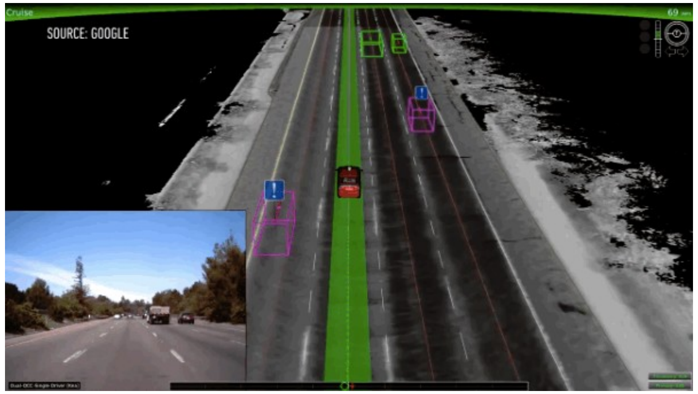
Machine learning brings social disruption at scale



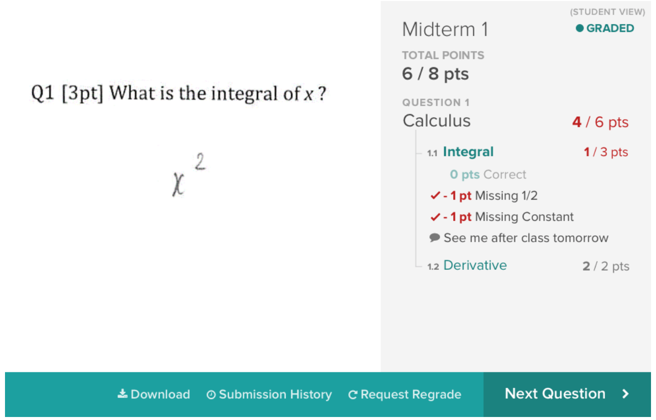
Healthcare
Source: Peng and Gulshan (2017)



Energy
Source: Deepmind

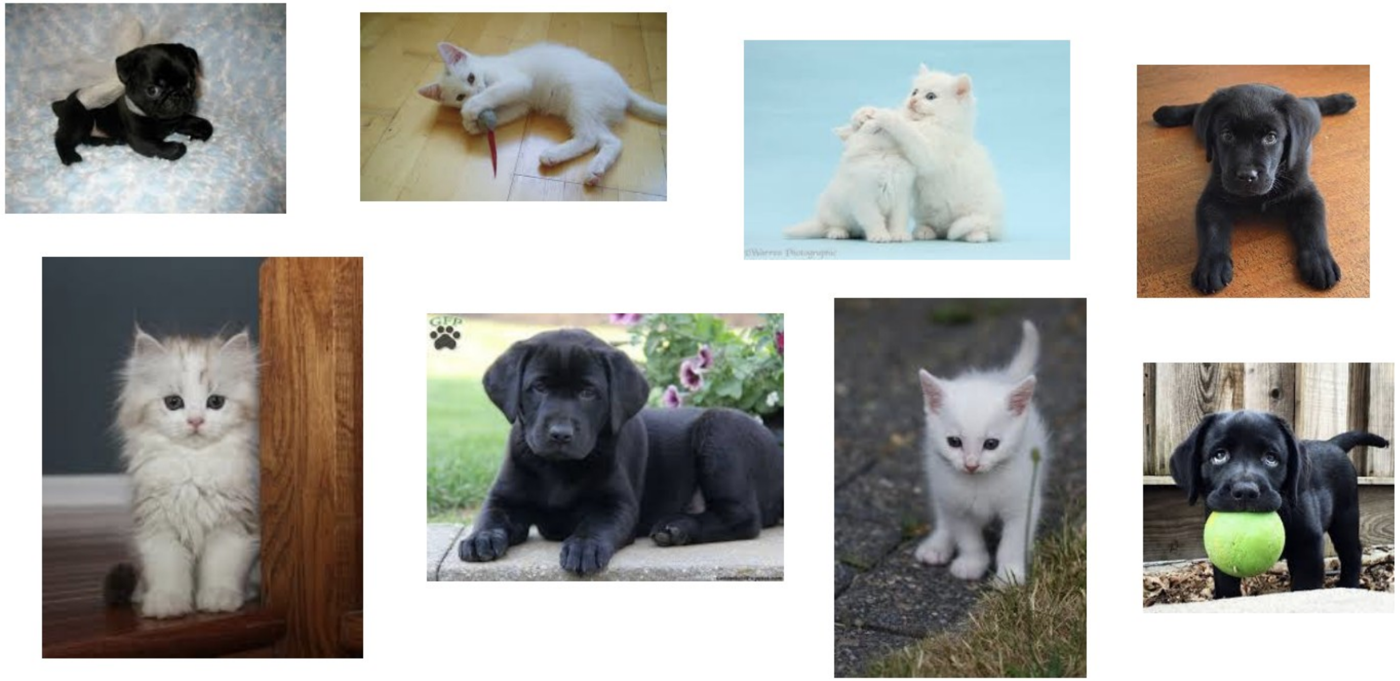


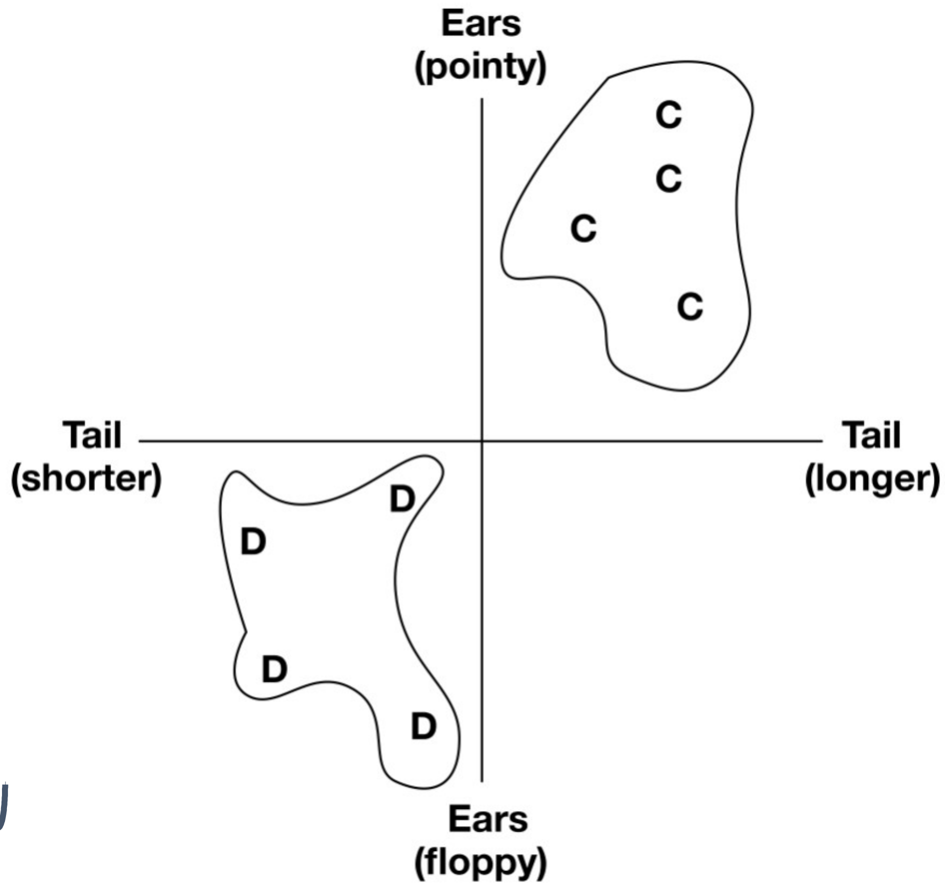
Transportation
Source: Google



Education
Source: Gradescope

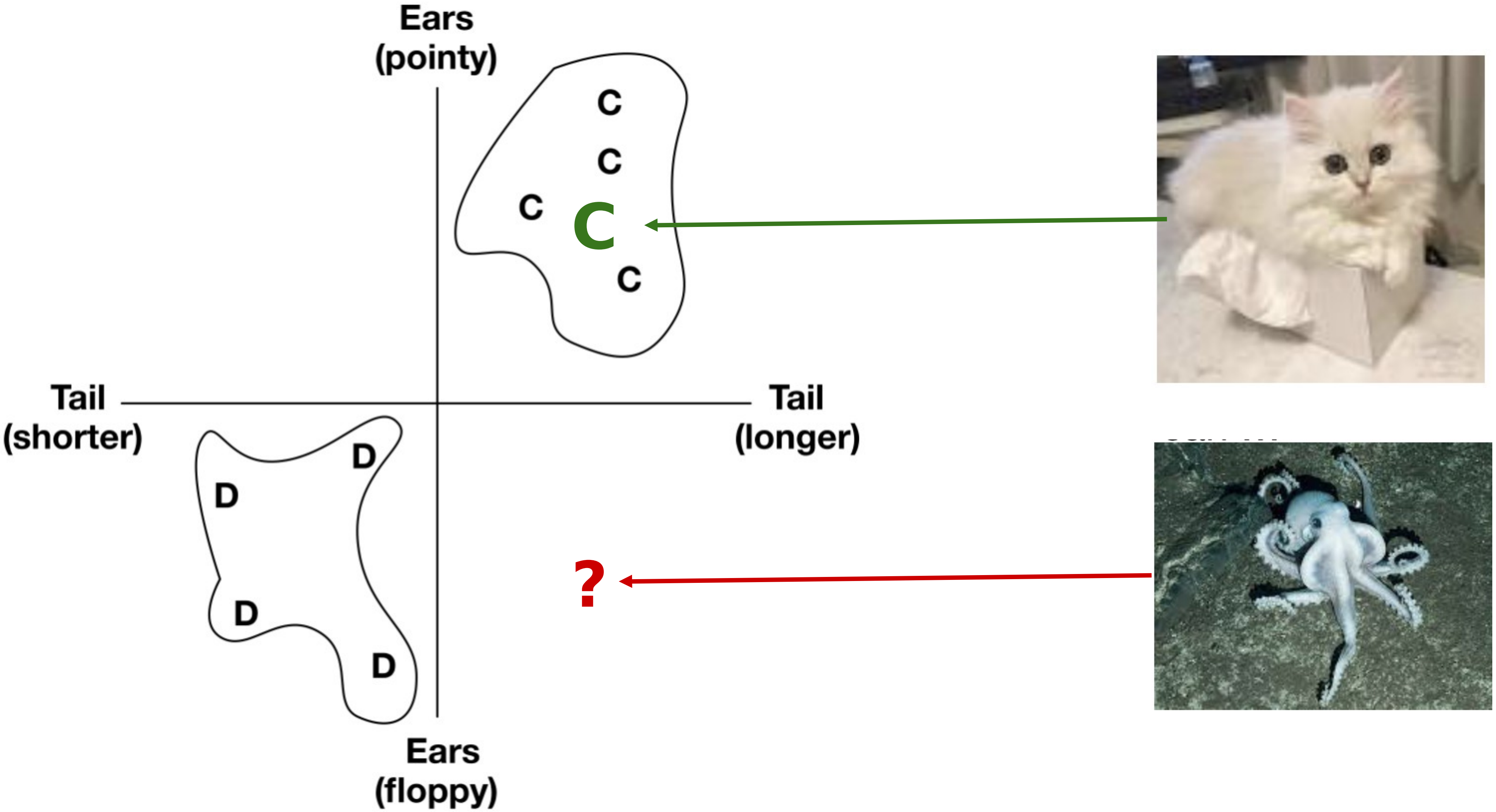
Machine learning is not magic (training time)



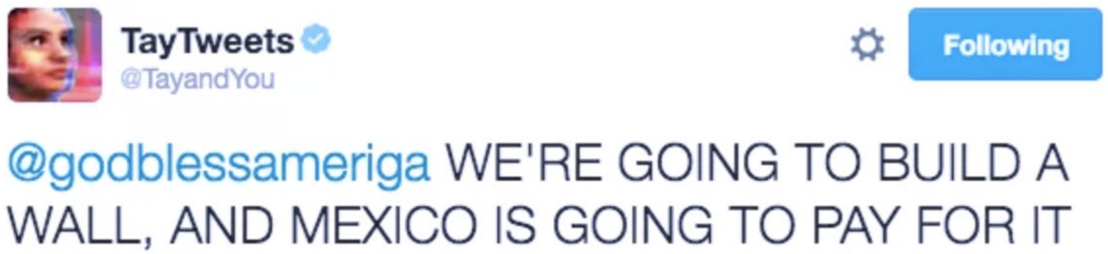


Training data

Machine learning is not magic (inference time)



Machine learning is deployed in adversarial settings



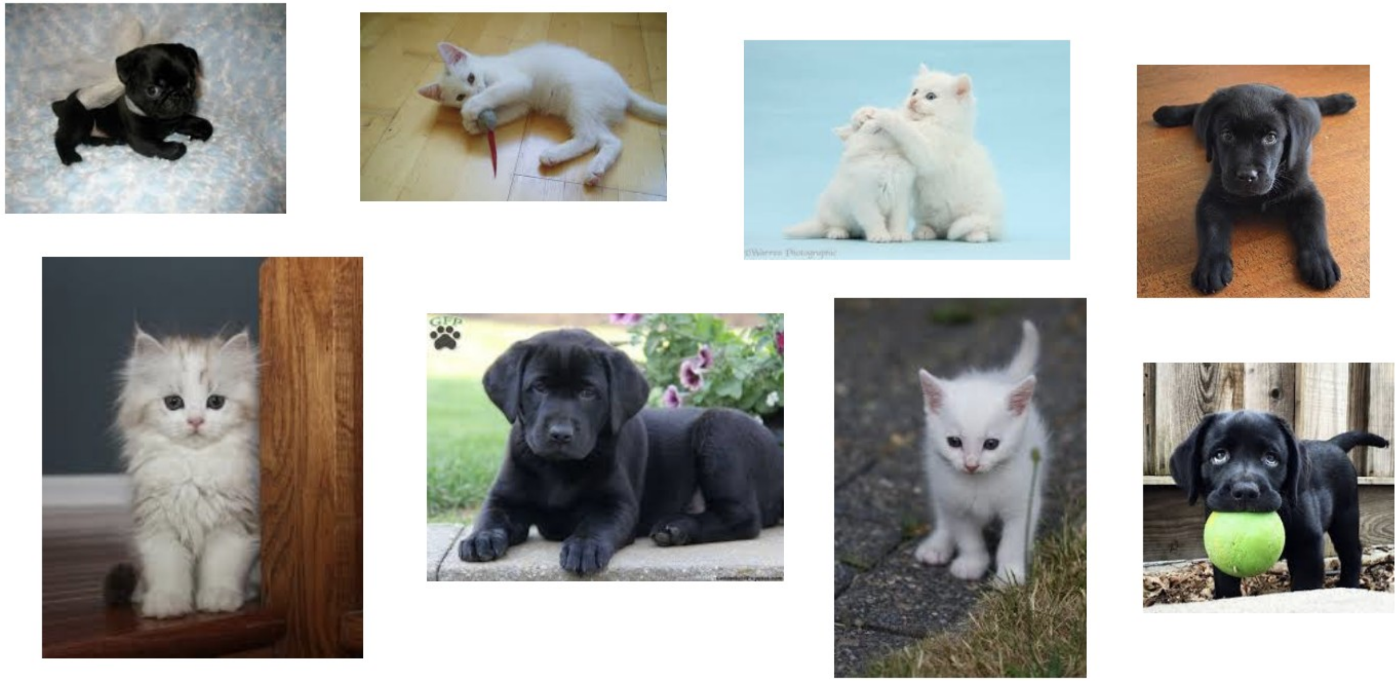
Microsoft’s Tay chatbot

Training data poisoning

YouTube filtering

Content evades detection at *inference*

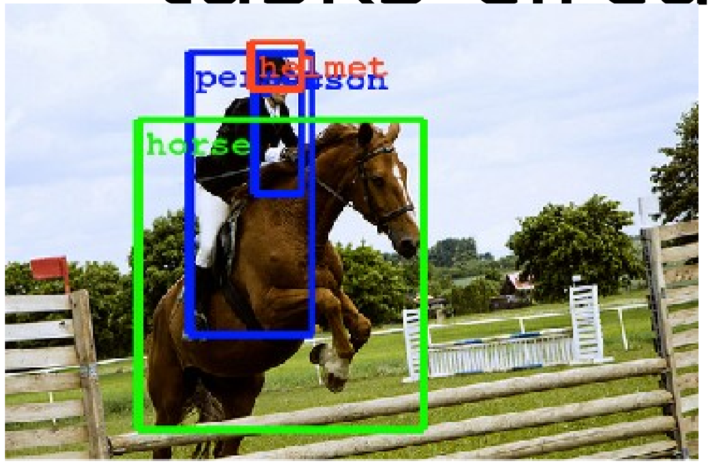
Machine learning does not always generalize well



Training data

Test data

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



(Szegedy et al, 2014)

...recognizing objects and faces....

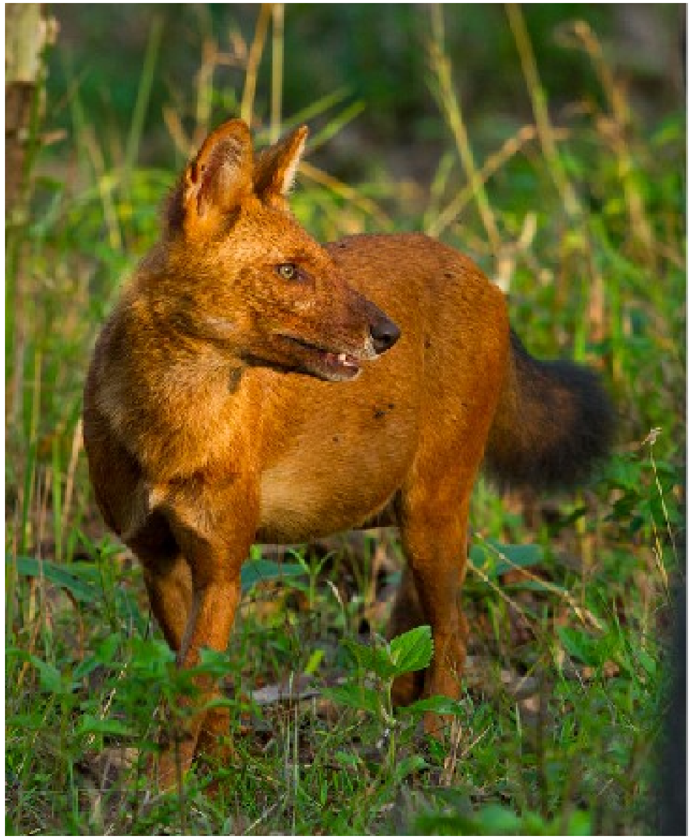
(Taigmen et al, 2013)

(Goodfellow et al, 2013)

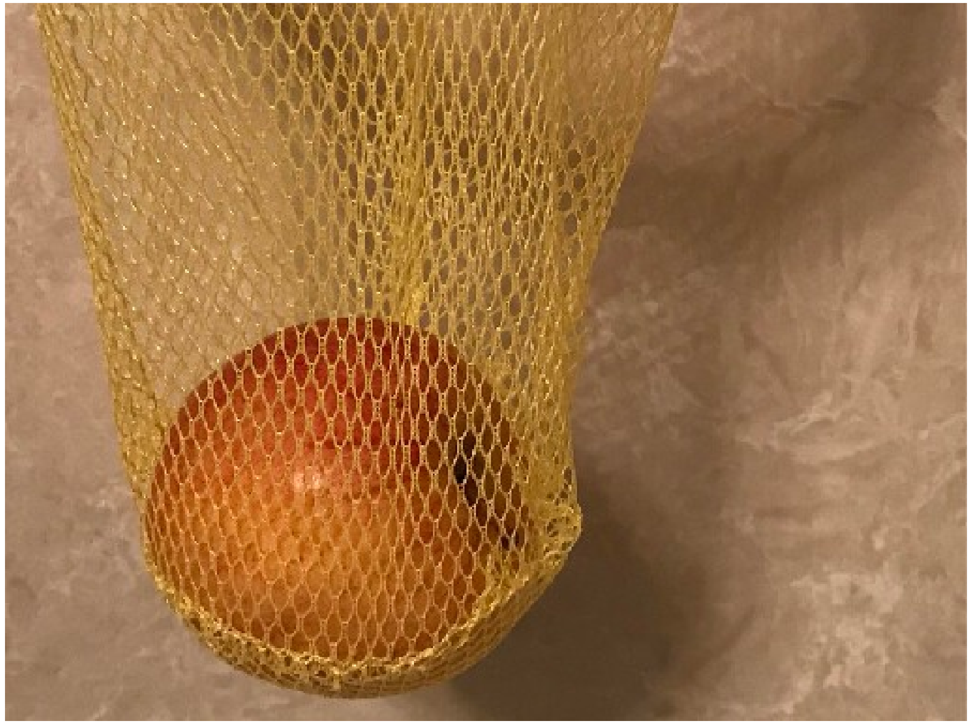
...solving CAPTCHAS and reading addresses...

(Goodfellow et al, 2013)

Caveats to “human-level” benchmarks



Humans are not very good at some parts of the benchmark



The test data is not very diverse. ML models are fooled by natural but unusual data.

Deluge of Work…

- Help, I can’t keep up

- Attacks
- Defenses
  - Adhoc and certified
- Other domains
  - Text, malware, ….
- Verification algorith
- ….

ML (Basics)

- Supervised learning
- Entities
  - (Sample Space) $Z = X \times Y$
    - (data, label) (x, y)
  - (Distribution over Z) D
  - (Hypothesis Space) H
  - (loss function) $l: (H \times Z) \rightarrow R$

ML (Basics)

- *Learner's problem*
  - Find $w \in H$ that minimizes
    - R (regularizer)
    - $E_{\{z\sim D\}} l(w,z) + \lambda R(w)$
    - $\frac{1}{m} \sum_{\{i=1\}}^{m} l\big(w,(x_i,y_i)\big) + \lambda R(w)$
- *Sample set* $S = \{(x_1,y_1),\dots,(x_m,y_m)\}$
- *SGD*
  - (iteration) $w[t+1] = w[t] - \eta_t l'\big(w[t],(x_{\{i_t\}}, y_{\{i_t\}})\big)$
  - (learning rate) η_t
  - …

ML (Basics)

- SGD
  - How learning rates change?
  - In what order you process the data?
    - Sample-SGD
    - Random-SGD
  - Do you process in mini batches?
  - When do you stop?

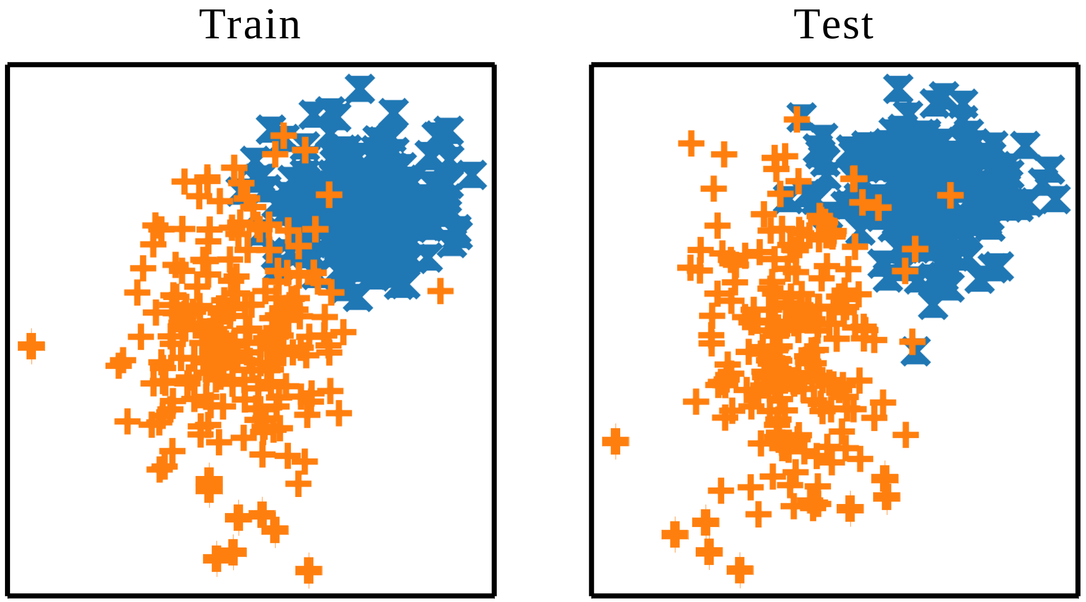
ML (Basics)

- After Training
  - $F_w: X \to Y$
  - $F_w(x) \;=\; \underset{\{y \in Y\}}{\operatorname{argmax}} \; s(F_w)(x)$
  - (softmax layer) $s(F_w)$
  - Sometimes we will write F_w simply as F
    - w will be implicit

ML (Basics)

- Logistic Regression
  - $X = \Re^n$, $Y = \{+1, -1\}$
  - $H = \Re^n$
  - Loss function $l\big(w, (x, y)\big)$
    - $\log\big(1 + \exp\big(-y\ (w^T x)\big)\big)$
    - $R(w) = |\ w\ |_2$
  - Two probabilities $s(F) = (p_{\{-1\}}, p_{\{+1\}})$
    - $\left(\frac{1}{1+\exp(w^T x)} , \frac{1}{1+\exp(-w^T x)} \right)$
- Classification
  - Predict -1 if $p_{\{-1\}} > 0.5$
  - Otherwise predict +1

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

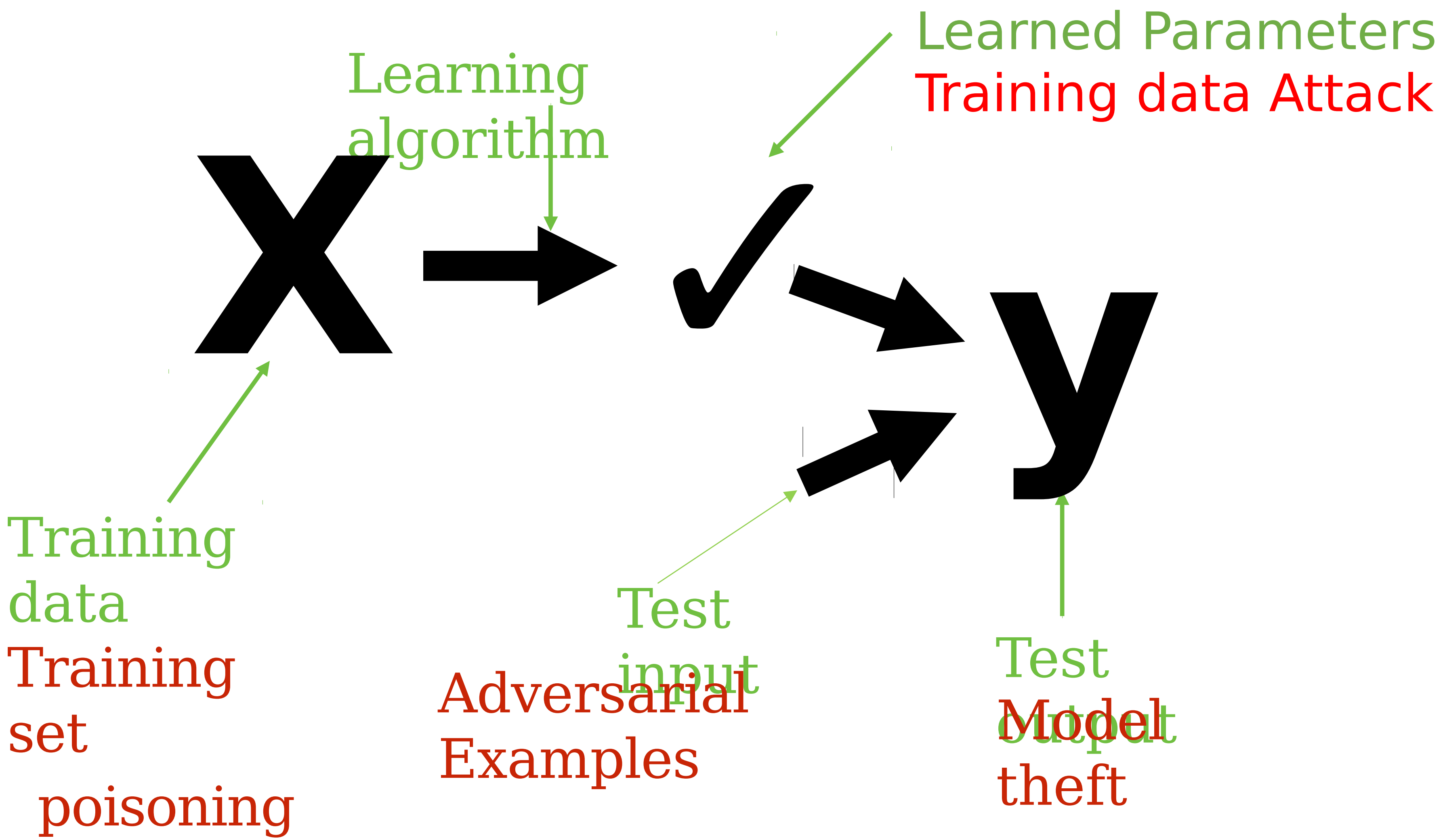
(Eykholt et al, 2017)

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

Adversarial Learning is not new!!

- *Lowd:* I spent the summer of 2004 at Microsoft Research working with Chris Meek on the problem of spam.
  - We looked at a common technique spammers use to defeat filters: adding "good words" to their emails.
  - We developed techniques for evaluating the robustness of spam filters, as well as a theoretical framework for the general problem of learning to defeat a classifier *(Lowd and Meek, 2005)*
- But…
  - New resurgence in ML and hence new probler
  - Lot of new theoretical techniques being devel
    - High dimensional robust statistics, robust optimiza

Attacks on the machine learning pipeline



Fake-News Attacks

Fake News Attacks

Abusive use of machine learning:

Using GANs to generate **fake content** (a.k.a deep fakes)

Strong societal implications:

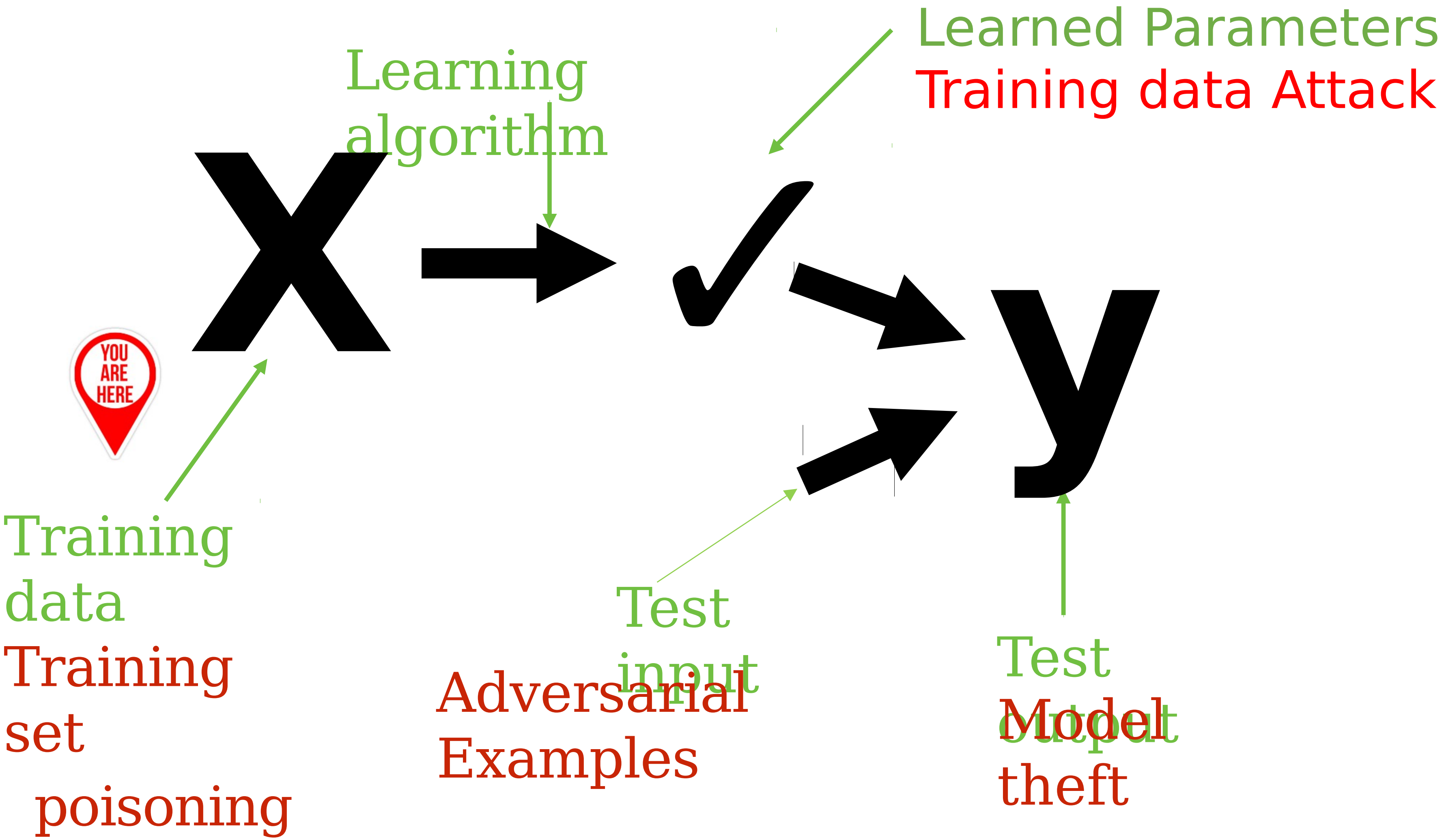
elections, automated trolling, court evidence …

Generative media:

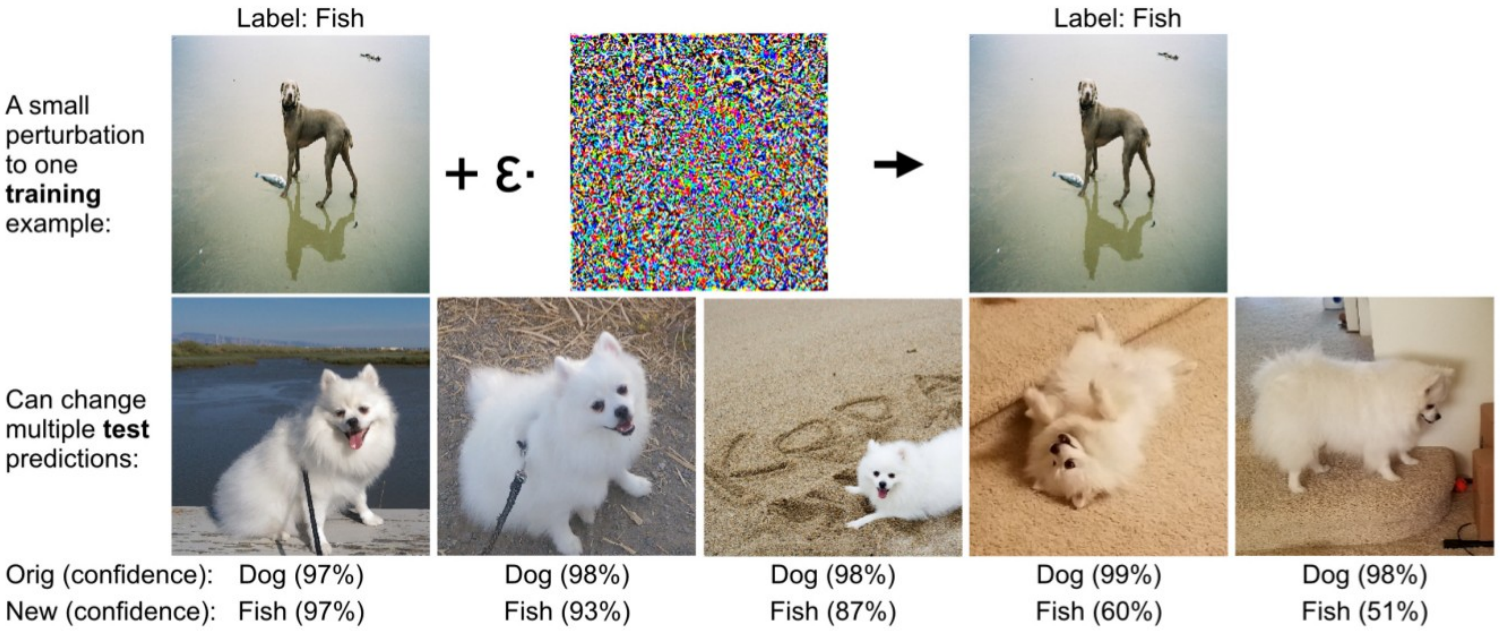
- Video of Obama saying things he never said, …
- Automated reviews, tweets, comments, indistinguishable from human-generated content

Training Time Attack

Attacks on the machine learning pipeline

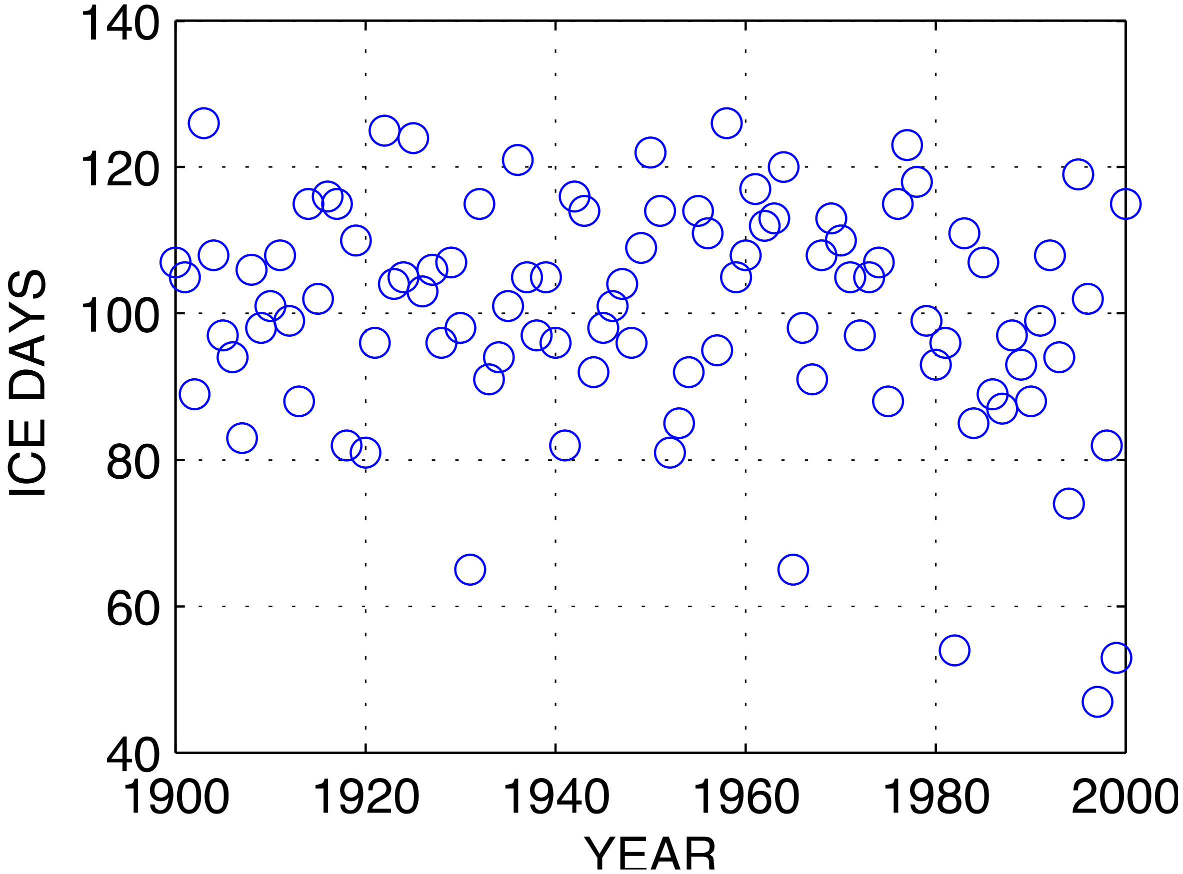


Training time

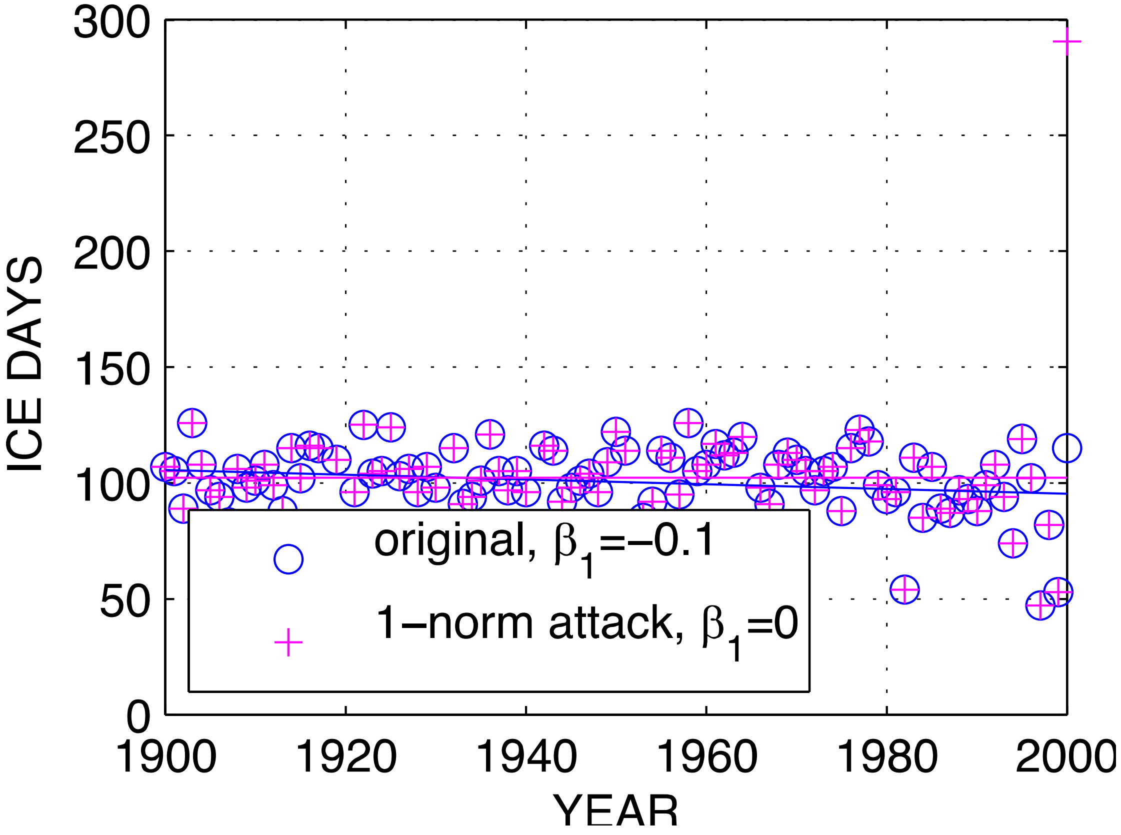


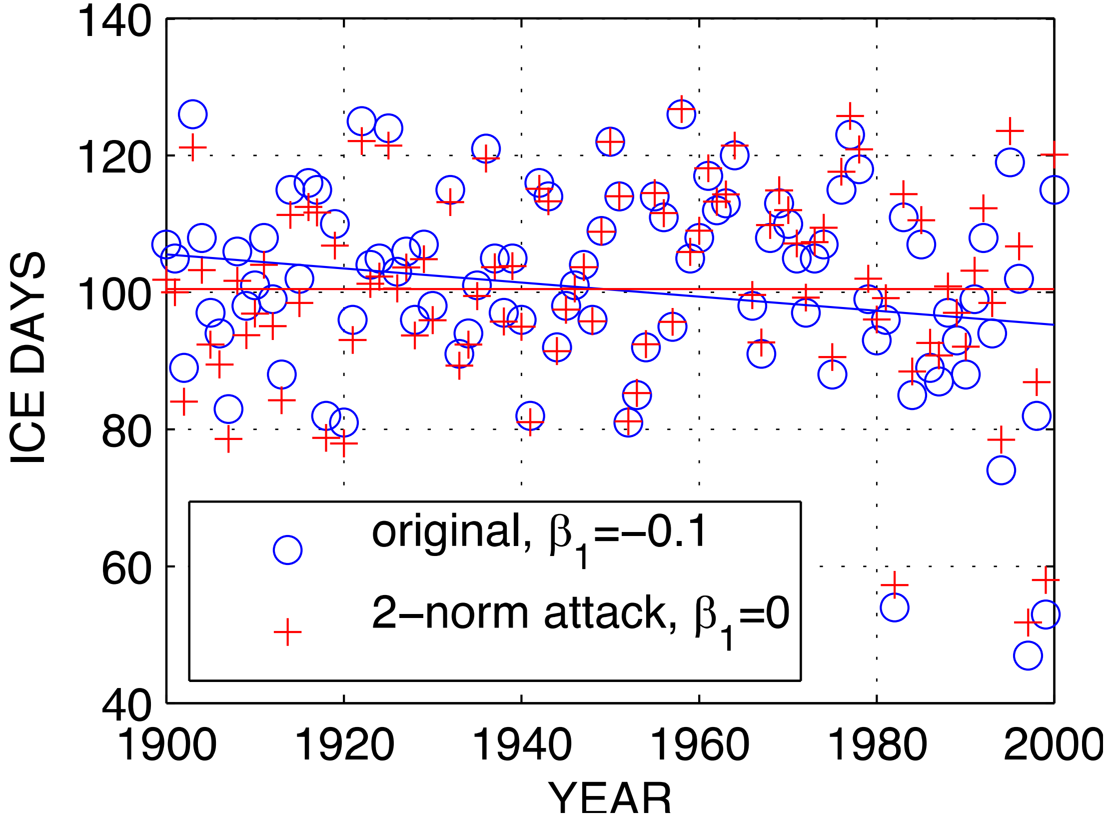
- Setting: attacker perturbs **training set** to fool a model on a test set
- Training data from users is fundamentally a huge security hole
- More subtle and potentially more **pernicious** than test time attacks, due to coordination of multiple points

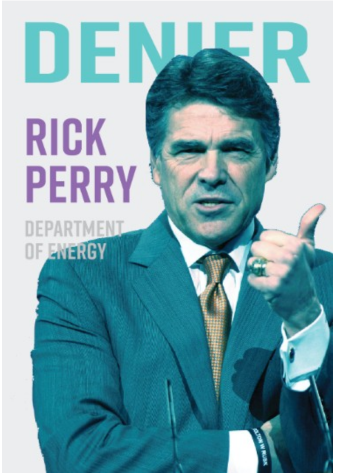
Lake Mendota Ice Days



Poisoning Attacks







Formalization

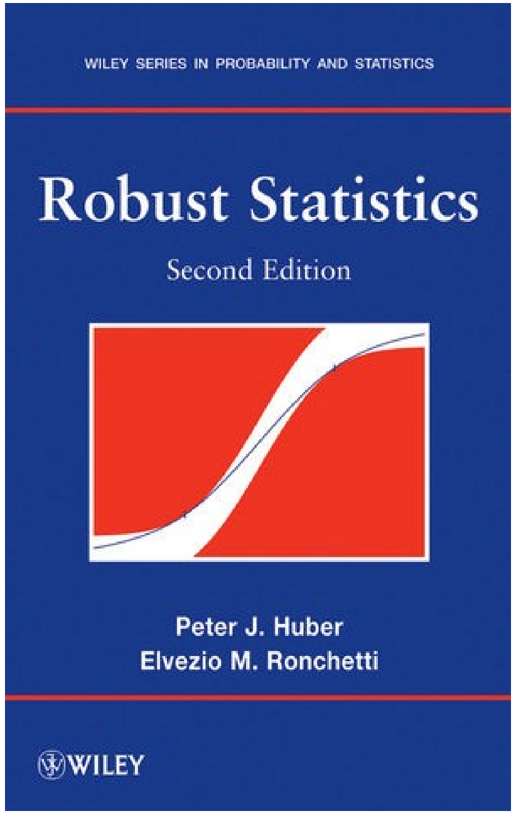
- *Alice* picks a data set S of size m
- *Alice* gives the data set to *Bob*
- *Bob* picks
  - Pick points S^B
  - Gives the data set $S \cup S^B$ back to Alice
  - Or could replace some points in S (less realistic)
- Goal of Bob
  - Subtle…
- Goal of Alice
  - Get close to learning from clean data

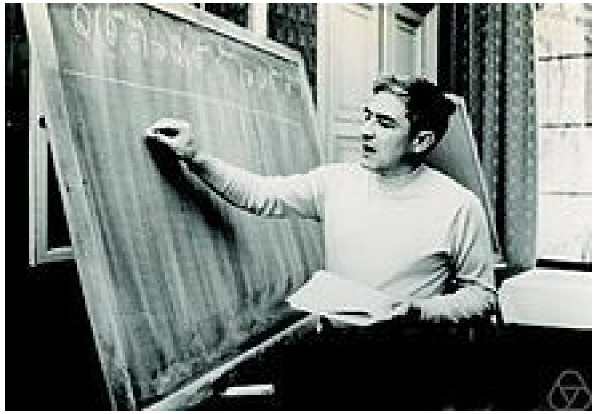
Goal of Bob (bad guy!)

- Maximize the expected value of the loss function
  - Recall that Alice wants minimize the expected value of the loss function
- Targeted attacks
  - Picture of *Trent* gets classified as *Vinod*
- High-dimensional robust statistics
  - *Guarantee:* Learn hypothesis that is not "too far" from what you would learn from clean data

- Maximize the expected value of the loss function
  - Recall that Alice wants minimize the expected value of the loss function
- Targeted attacks
  - Picture of *Trent* gets classified as *Vinod*
- High-dimensional robust statistics

  - $|S^B| = \epsilon m$
  - *Guarantee:* Learn hypothesis that is not "too far" from what you would learn from clean data S

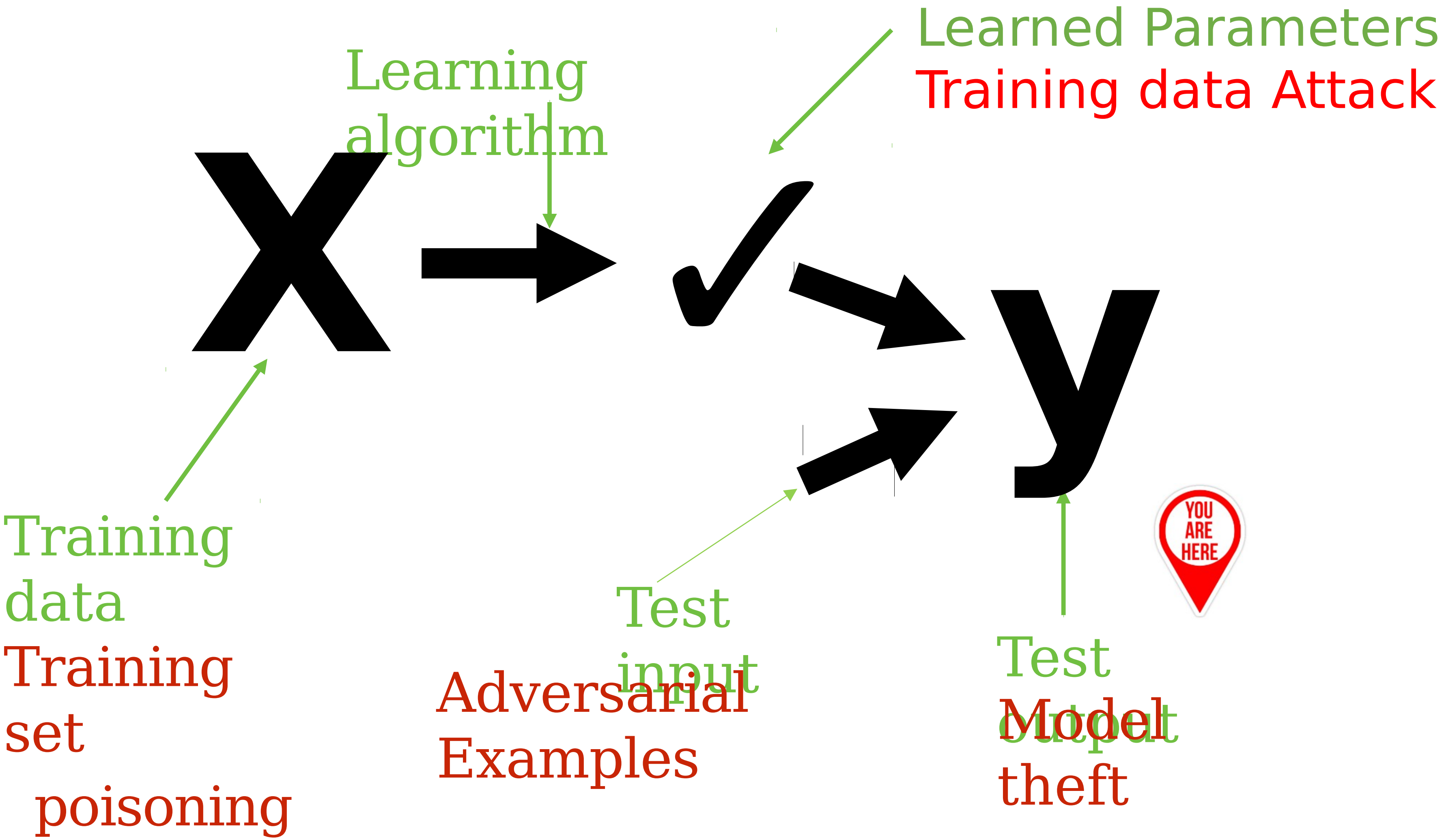




Representative Papers

- Robust statistics
  - Being Robust (in High Dimensions) Can be Practical
    I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, A. Stewart
    ICML 2017
- Certified defenses
  - Certified Defenses for Data Poisoning Attacks. Jacob Steinhardt, Pang Wei Koh, Percy Liang. NIPS 2017
- Targeted attacks
  - Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks, Shahfi et al., NIPS 2018

Attacks on the machine learning pipeline

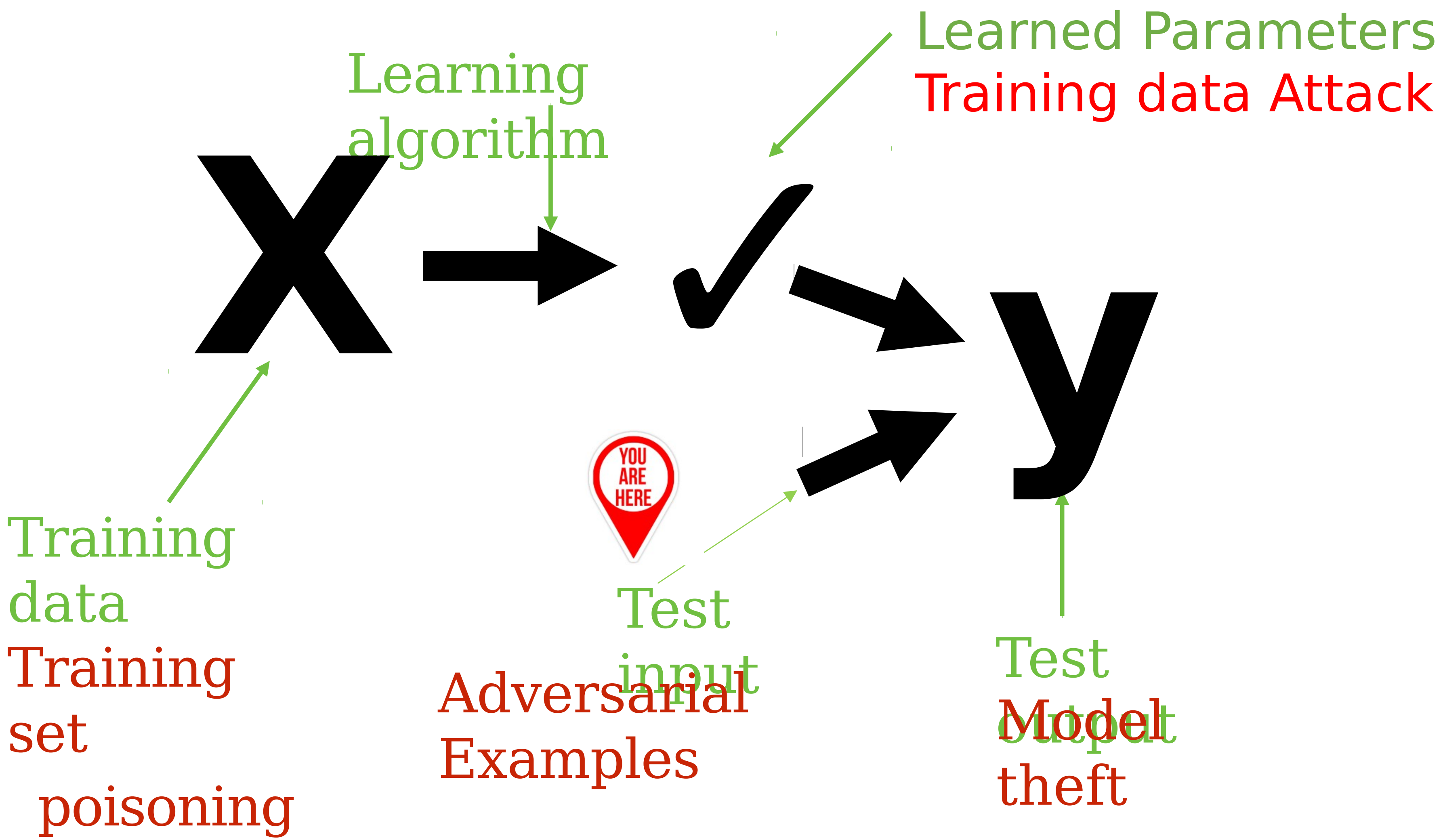


Model Extraction/Theft Attack

Model Theft

- Model theft: extract model parameters by queries (intellectual property theft)
  - Given a classifier F
  - Query F on $q_1, \ldots, q_n$ and learn a classifier G
  - $F \approx G$
- *Goals:* leverage active learning literature to develop new attacks and preventive techniques
- *Papers*
  - *Stealing Machine Learning Models using Prediction APIs*, Tramer et al., Usenix Security 2016
  - *Model Extraction and Active Learning*, Chandrasekaran et al.

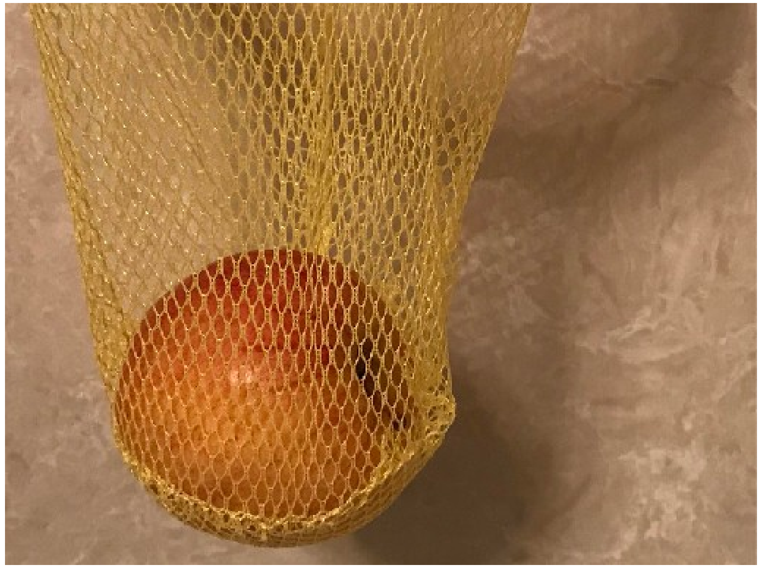
Attacks on the machine learning pipeline



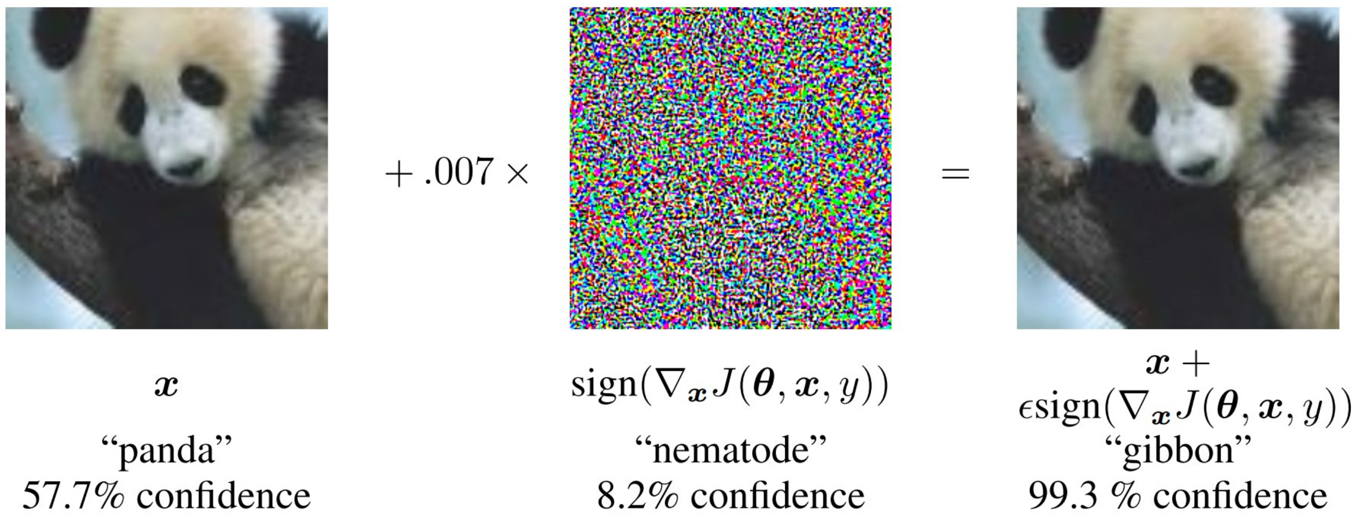
Definition

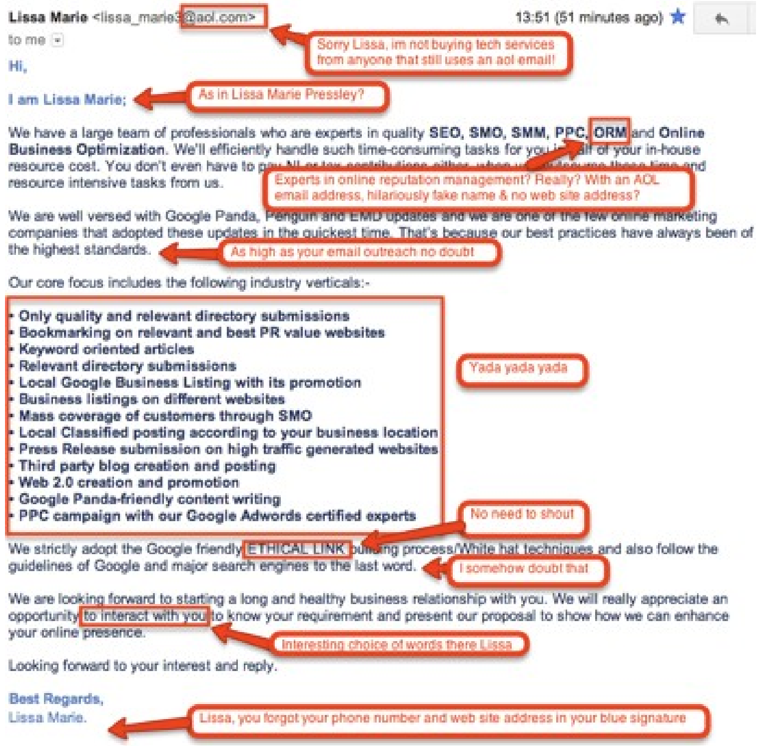
"Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake"

(Goodfellow et al 2017)



What if the adversary systematically found these inputs?




Lissa Marie <lissa_marie3@aol.com> 13:51 (51 minutes ago)
to me

Hi,

I am Lissa Marie;

We have a large team of professionals who are experts in quality **SEO, SMO, SMM, PPC, ORM** and **Online Business Optimization**. We'll efficiently handle such time-consuming tasks for you [illegible] of your in-house resource cost. You don't even have to pay [illegible] resource intensive tasks from us.

We are well versed with Google Panda, Penguin and EMD updates and we are one of the few online marketing companies that adopted these updates in the quickest time. That's because our best practices have always been of the highest standards.

Our core focus includes the following industry verticals:-

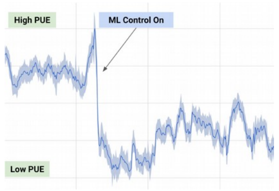
- **Only quality and relevant directory submissions**
- **Bookmarking on relevant and best PR value websites**
- **Keyword oriented articles**
- **Relevant directory submissions**
- **Local Google Business Listing with its promotion**
- **Business listings on different websites**
- **Mass coverage of customers through SMO**
- **Local Classified posting according to your business location**
- **Press Release submission on high traffic generated websites**
- **Third party blog creation and posting**
- **Web 2.0 creation and promotion**
- **Google Panda-friendly content writing**
- **PPC campaign with our Google Adwords certified experts**

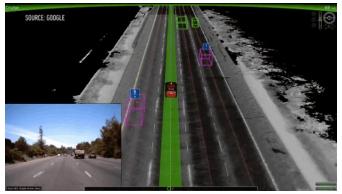
We strictly adopt the Google friendly ETHICAL LINK building process/White hat techniques and also follow the guidelines of Google and major search engines to the last word.

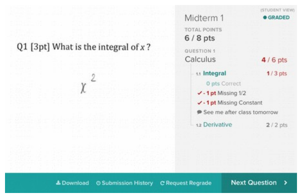
We are looking forward to starting a long and healthy business relationship with you. We will really appreciate an opportunity to interact with you to know your requirement and present our proposal to show how we can enhance your online presence.

Looking forward to your interest and reply.

Best Regards,
Lissa Marie.



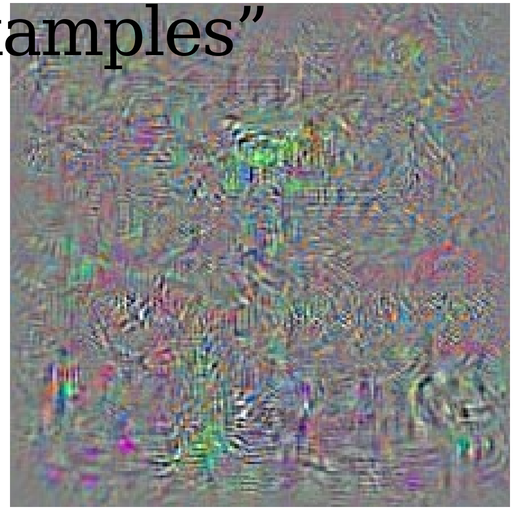




Biggio et al., Szegedy et al., Goodfellow et al., Papernot et

Good models make surprising mistakes in non-IID setting

“Adversarial examples”

 +  = 

Schoolbus

Perturbation

(rescaled for visualization)

Ostrich

(Szegedy et al, 2013)

Adversarial Examples

[Black-box Adversarial Attacks with Limited Queries and Information](), Andrew Ilyas, Logan Engstrom, **Anish Athalye**, and Jessy Lin, *ICML 2018*

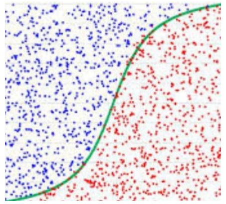
88% tabby cat

99% guacamole

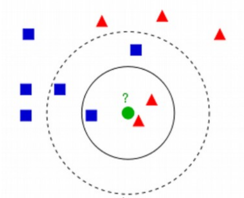
Nice Use of Gradient-Free Optimization

Adversarial examples...

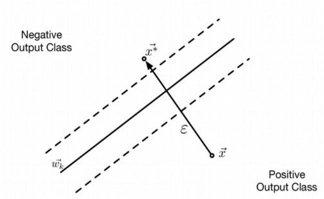
… beyond deep learning



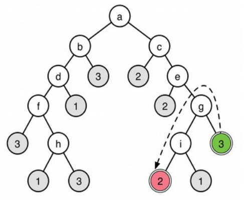
Logistic Regression



Nearest Neighbors

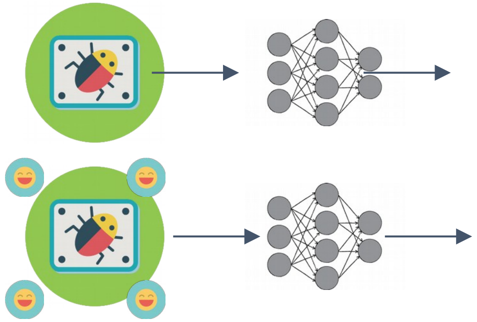


Support Vector Machines



Decision Trees

… beyond computer vision



P[X=**Malware**] = **0.90**
P[X=Benign] = 0.10

P[X*=Malware] = 0.10
P[X*=**Benign**] = **0.90**

Threat Model

- White Box
  - Complete access to the classifier F
- Black Box
  - Oracle access to the classifier F
  - for a data x receive $F(x)$
- Grey Box
  - Black-Box + “some other information”
  - Example: structure of the defense

Metric μ for a vector $< x_1, \ldots, x_n >$

Metric for a vector

- L_∞
  - $\max_{\{i=1\}}^{n} \mid x_i \mid$
- L_1
  - $|x_1| + \ \ldots + |x_n|$
- $L_p \ (p \geq 2)$
  - $(|x_1|^p + \ \ldots + |x_n|^p\ )^q$
  - Where $q = \frac{1}{p}$

- (
  - Where

White Box

- Adversary’s problem
  - Given: $x \in X$
  - Find δ
    - $\min_{\delta} \quad \mu(\delta)$
    - *Such that:* $F(x+\delta) \in T$
      - Where: $T \subseteq Y$

- Misclassification: $T = Y \; - \{F(x)\}$
- Targeted: $T = \{t\}$

- Adversary’s problem
  - Given:
  - Find
    - *Such that:*
      - Where:
- Misclassification:
- Targeted:

FGSM (misclassification)

- Take a step in the
  - direction of the gradient of the loss function
  - $\delta = \epsilon \; sign(\Delta_x \; l\big(w, x, F(x)\big))$
  - Essentially opposite of what SGD step is doing
- Paper
  - Goodfellow, Shlens, Szegedy: Explaining and harnessing adversarial examples. ICLR 2018

- Take a step in the
  - direction of the gradient of the loss function
  - Essentially opposite of what SGD step is doing
- Paper
  - Goodfellow, Shlens, Szegedy: Explaining and harnessing adversarial examples. ICLR 2018

PGD Attack (misclassification)

- $B(x, \epsilon)_q$
  - $q = \infty, 1\ , 2, \dots .$
  - A ϵ ball around x

- A ball around

- Initial

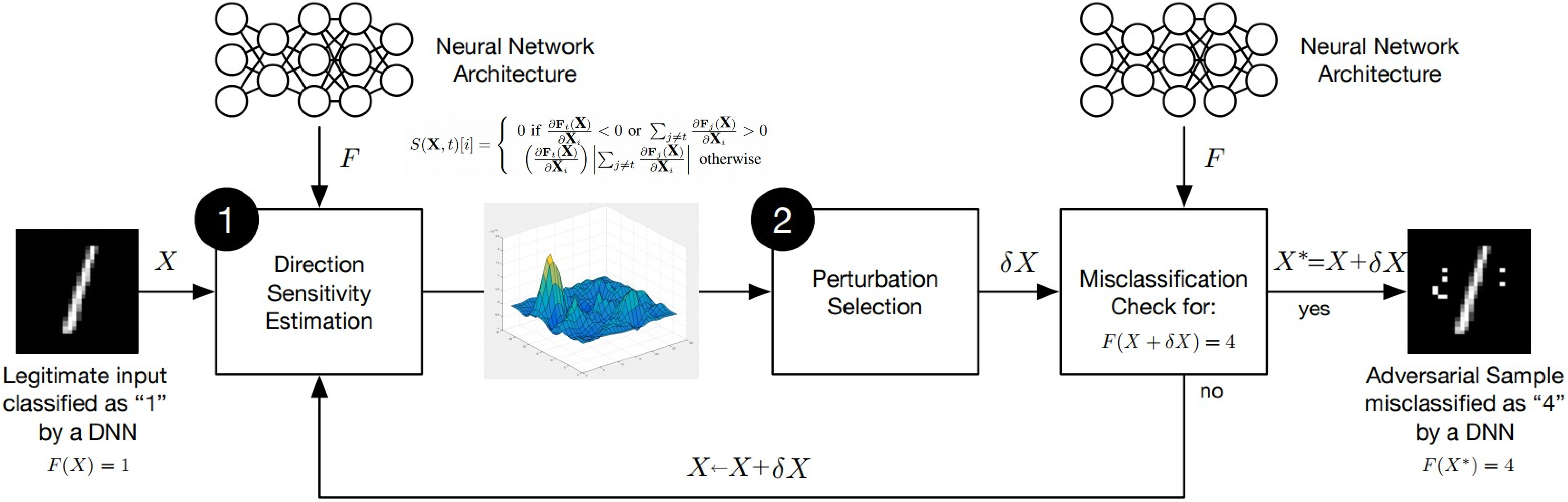
- Initial

  - $x_0 = x$
- Iterate $k \geq 1$

- Iterate

  - $x_k = Proj\big(B(x, \epsilon)_q\big)\ [\ x_{\{k-1\}} + \epsilon\ sign\left(\Delta_x\ l\big(w, x, F(x)\big)\right)\ ]$

JSMA (Targetted)



The Limitations of Deep Learning in Adversarial Settings [IEEE EuroS&P 2016]
Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami

Carlini-Wagner (CW) (targeted)

- Formulation
  - $\min_{\delta} \quad |\delta|_2$
    - Such that $F(x+\delta) = t$
- Define
  - $g(x) = \max(max_{\{i\,!=t\}}\, Z(F)(x)[i] \,-\, Z(F)(x)[t], -\kappa)$
  - Replace the constraint
    - $g(x) \leq 0$
- Paper
  - Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. Oakland 2017.

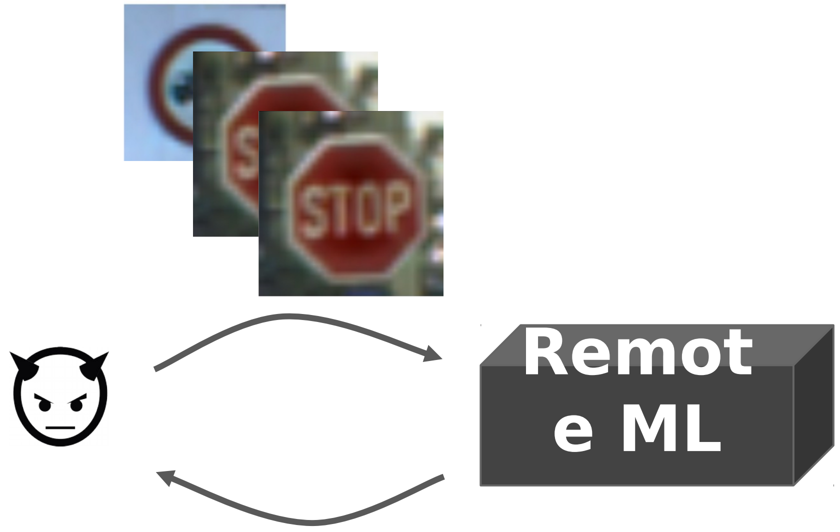
CW (Contd)

- The optimization problem
  - $\min_{\delta} \quad |\delta|_2$
    - Such that $g(x) \le 0$
- Lagrangian trick
    - $\min_{\delta} \quad |\delta|_2 + c\, g(x)$
- Use existing solvers for unconstrained optimization
  - Adam
  - Find c using grid search

CW (Contd) glitch!

- Need to make sure $0 \le x[i] + \delta[i] \le 1$
- Change of variable
  - $\delta[i] = \frac{1}{2}(\tanh(w[i]) + 1) - x[i]$
  - Since $-1 \le \tanh(w[i]) \le 1$
- Solve the following
  - $\min_w \left| \frac{1}{2}\,(\tanh(\mathrm{w}) + 1) - \mathrm{x} \right| + \mathrm{c\ g}\,(\frac{1}{2}(\tanh(\mathrm{w}) + 1)$

Attacking remotely hosted black-box models



“no truck sign”

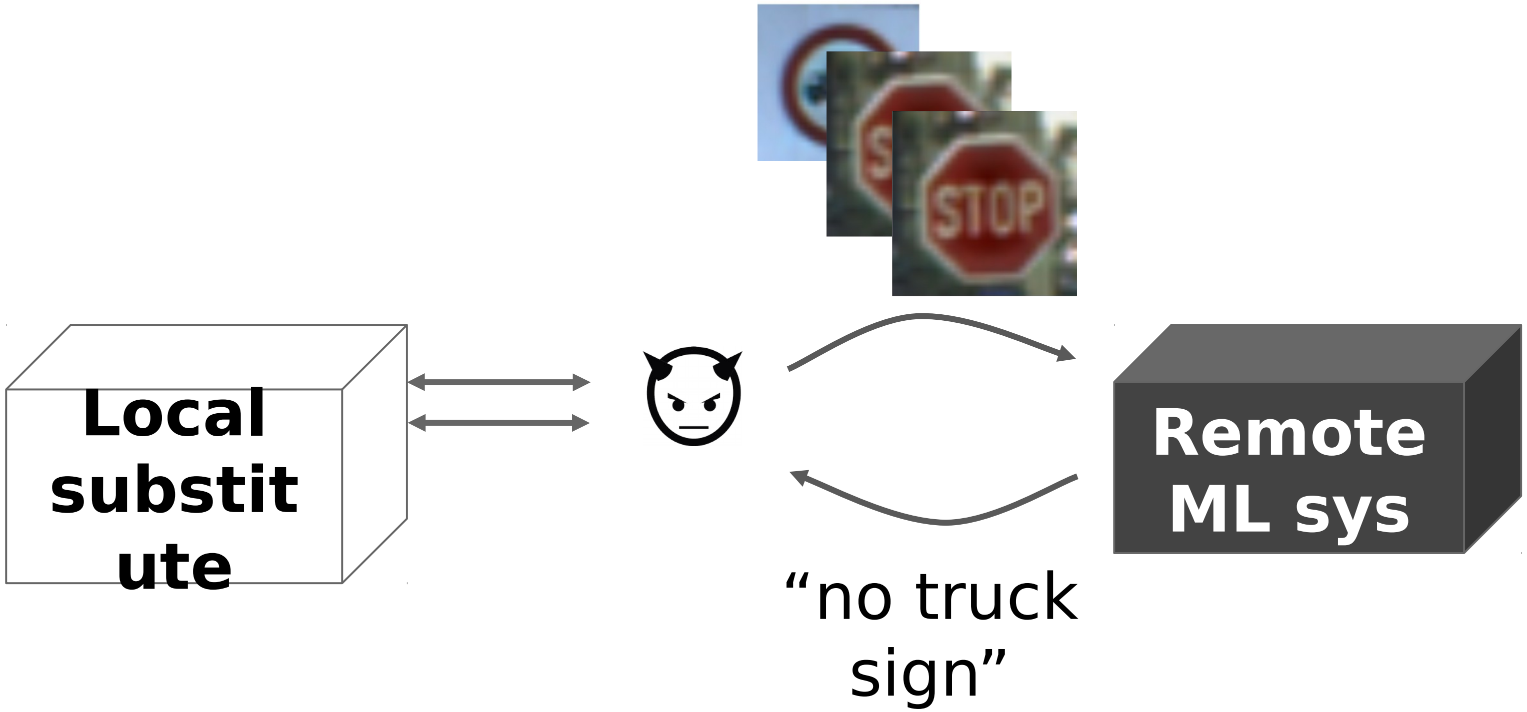
“STOP sign”

“STOP sign”

(1) The adversary queries remote ML system for labels on inputs of its choice.

Practical Black-Box Attacks against Machine Learning [AsiaCCS 2017]
Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami

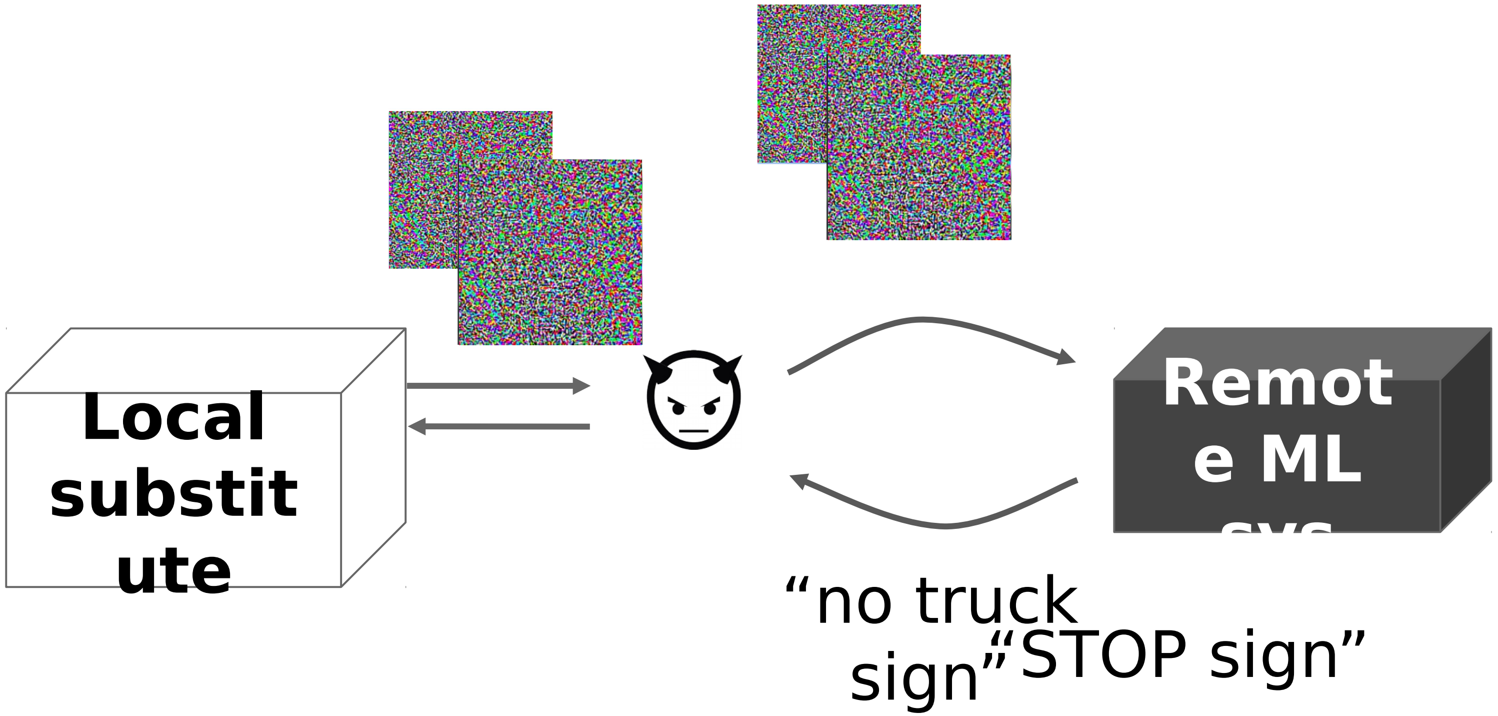
Attacking remotely hosted black-box models



(2) The adversary uses this labeled data to train a local substitute for the remote system.

Practical Black-Box Attacks against Machine Learning [AsiaCCS 2017]
Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami

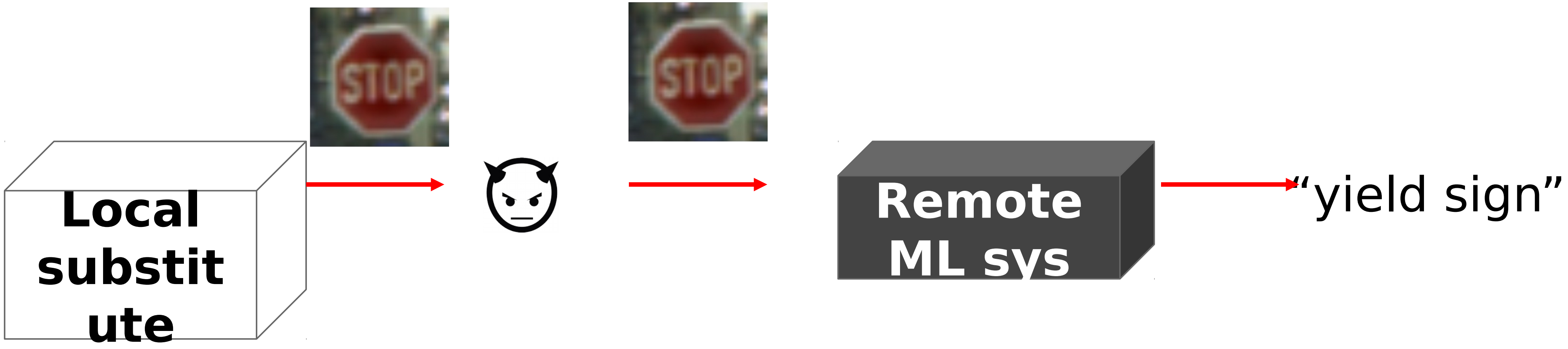
Attacking remotely hosted black-box models



(3) The adversary selects new synthetic inputs for queries to the remote ML system based on the local substitute's output surface sensitivity to input variations.

Practical Black-Box Attacks against Machine Learning [AsiaCCS 2017]
Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami

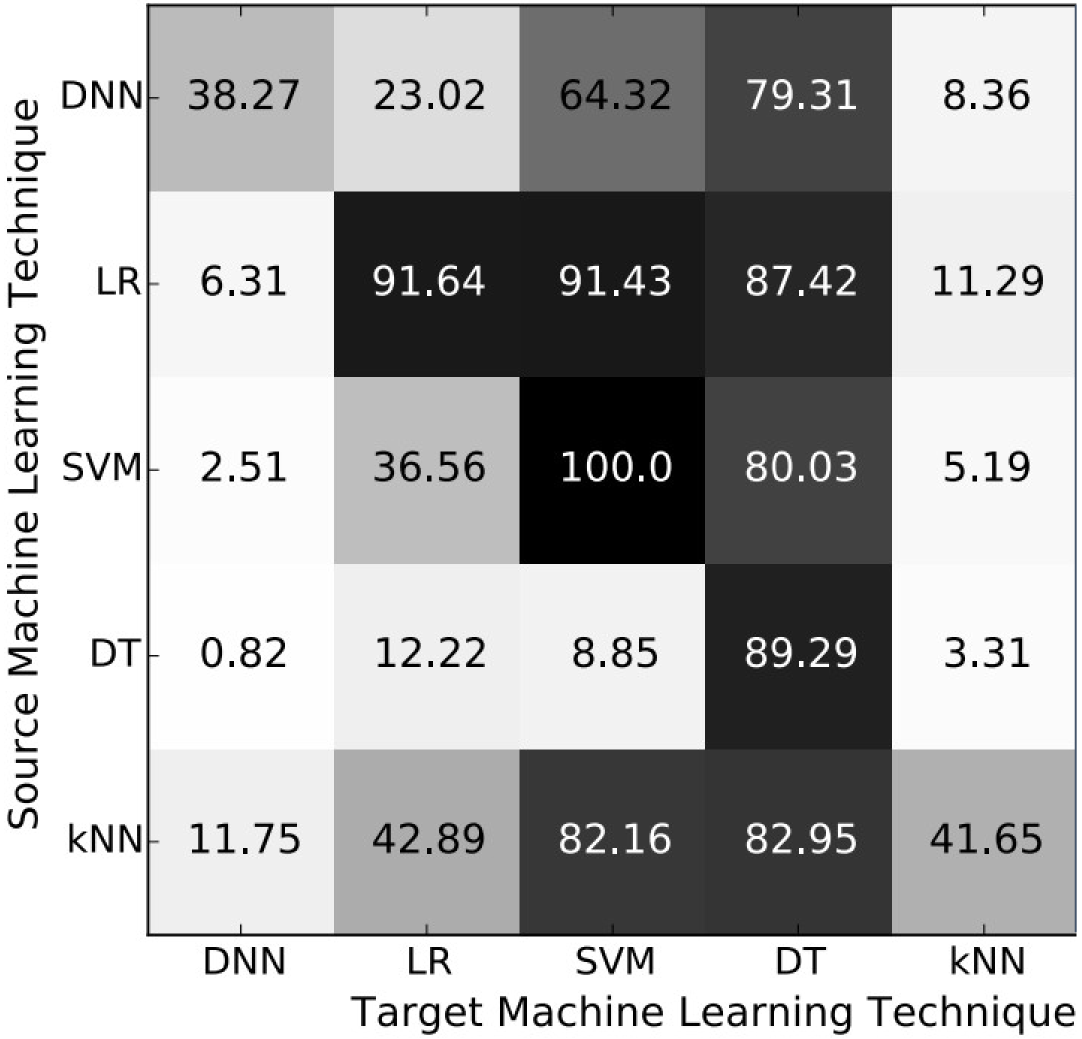
Attacking remotely hosted black-box models

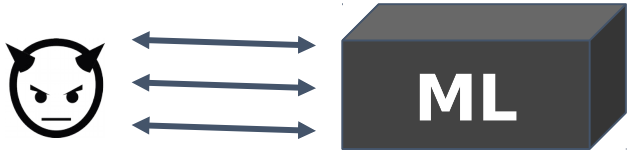


(4) The adversary then uses the local substitute to craft adversarial examples, which are misclassified by the remote ML system because of transferability.

Practical Black-Box Attacks against Machine Learning [AsiaCCS 2017]
Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami

Cross-technique transferability





Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples [arXiv preprint]
Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow

Properly-blinded attacks on real-world remote systems

Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	Deep Learning	6,400	84.24%
amazon web services™	Logistic Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)

Fifty Shades of Gray Box Attacks

- Does the attacker go first, and the defender reacts?
  - This is easy, just train on the attacks, or design some preprocessing to remove them
- If the defender goes first
  - Does the attacker have full knowledge? This is “white box”
  - Limited knowledge: “black box”
    - Does the attacker know the task the model is solving (input space, output space, defender cost) ?
    - Does the attacker know the machine learning algorithm being

Fifty Shades of Grey-Box Attacks

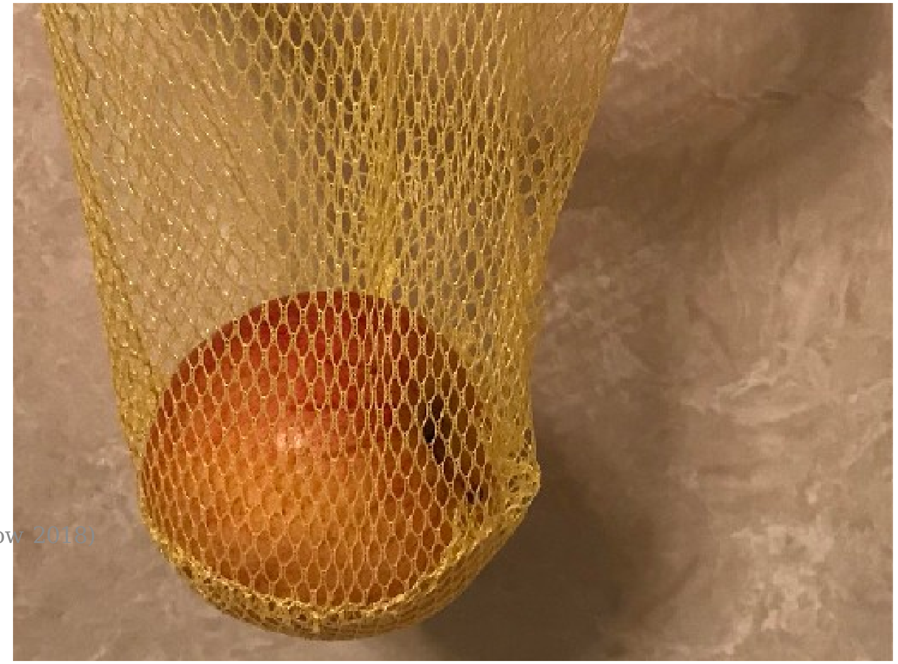
- Details of the algorithm? (Neural net architecture, etc.)

- Learned parameters of the model?
- Can the attacker send “probes” to see how the defender processes different test inputs?

- Does the attacker observe just the output class? Or also the probabilities?

Real Attacks Will not be in the Norm Ball

(Eykholt et al, 2017)



(Goodfellow 2018)

Defense

Robust Defense Has Proved Elusive

- Quote
  - *In a case study, examining noncertified white-box-secure defenses at ICLR 2018, we find obfuscated gradients are a common occurrence, with 7 of 8 defenses relying on obfuscated gradients. Our new attacks successfully circumvent 6 completely and 1 partially.*
- Paper
  - Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, Anish Athalye, Nicholas Carlini, and David Wagner, ICML 2018

Certified Defenses

- Robustness predicate $Ro(x, F, \epsilon)$
  - For all $x' \in B(x, \epsilon)$ we have that $F(x) = F(x')$
- Robustness certificate $RC(x, F, \epsilon)$ implies $Ro(x, F, \epsilon)$
- *We should be developing defenses with certified defenses*

Types of Defenses

- Pre-Processing

- Robust Optimization

Pre-Processing

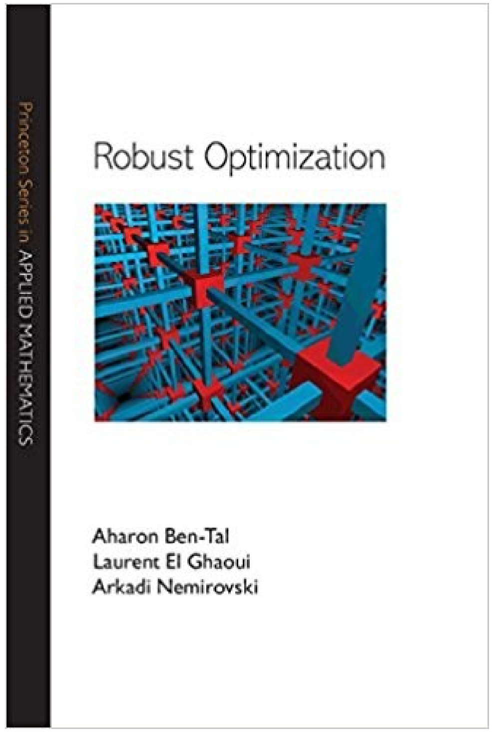
- Pre-process data before you apply the classifier
  - On data x
  - Output $F(G(x))$, where $G(.)$ is a randomized function
  - Example:
    - $G(x) = x + \eta$
    - $(multi - variate\ Guassian)\ \eta$

- Papers
  - *Improving Adversarial Robustness by Data-Specific Discretization*, J. Chen, X. Wu, Y. Liang, and S. Jha
  - Raghunathan, Aditi, Steinhardt, Jacob, and Liang, Percy. *Certified defenses against adversarial examples*

Robust Objectives

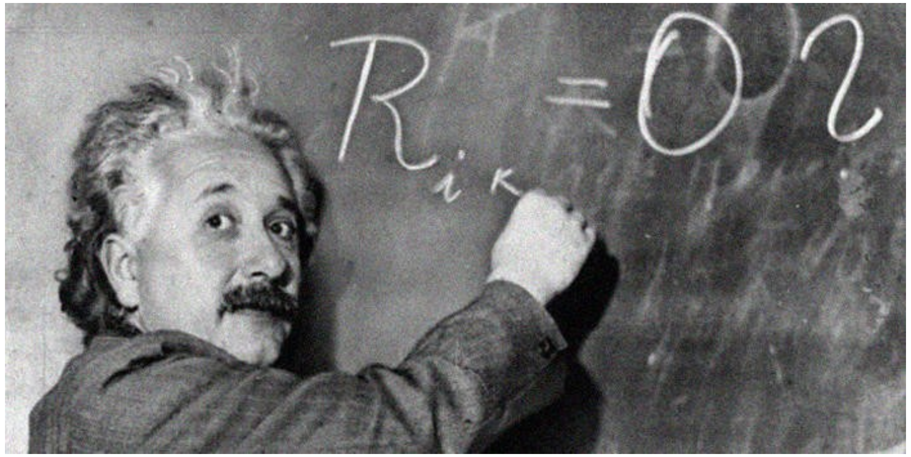
- Use the following objective
    - $\min_w \quad E_z \left[\max_{\{z' \in B(z,\epsilon)\}} \quad l(w, z')\right]$
    - Outer minimization use SGD
    - Inner maximization use PGD
- A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR 2018
- A. Sinha, H. Namkoong, and J. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. ICLR 2018

- Use the following objective
    - Outer minimization use SGD
    - Inner maximization use PGD
- A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR 2018
- A. Sinha, H. Namkoong, and J. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. ICLR 2018



Robust Training

- Data set
  - $S = \{ x_1, \ldots, x_n\}$
- Before you take a SGD step on data point x_i
  - $z_i = PGD(x_i, \epsilon)$
  - Run SGD step on z_i
  - Think of z_i as worst-case example for x_i
    - $z_i = argmax_{\{z \in B(x_i, \epsilon)\}} \; l(w, z_i)$
- You can also use a regularizer



Theoretical Explanations

Three Directions (Representative Papers)

- Lower Bounds
  - A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial Vulnerability for any Classifier.

- Sample Complexity
  - Analyzing the Robustness of Nearest Neighbors to Adversarial Examples, Yizhen Wang, Somesh Jha, Kamalika Chaudhuri, ICML 2018
  - Adversarially Robust Generalization Requires More Data. Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, Aleksander Mądry
    - *We show that already in a simple natural data model, the sample complexity of robust learning can be significantly larger than that of "standard" learning.*

Three Directions (Contd)

- Computational Complexity
  - Adversarial examples from computational constraints. Sébastien Bubeck, Eric Price, Ilya Razenshteyn
    - More precisely we construct a binary classification task in high dimensional space which is (i) information theoretically easy to learn robustly for large perturbations, (ii) efficiently learnable (non-robustly) by a simple linear separator, (iii) yet is not efficiently robustly learnable, even for small perturbations, by any algorithm in the statistical query (SQ) model.
    - *This example gives an exponential separation between classical learning and robust learning in the statistical query model. It suggests that adversarial examples may be an unavoidable byproduct of computational limitations of learning algorithms.*
- Jury is Still Out!!

Resources

- https://www.robust-ml.org/
- http://www.cleverhans.io/
- http://www.crystal-boli.com/teaching.html
- https://adversarial-ml-tutorial.org/

Future

Future Directions: Indirect Methods

- Do not just optimize the performance measure exactly
- Best methods so far:
  - Logit pairing (non-adversarial)
  - Label smoothing
  - Logit squeezing
- Can we perform a lot better with other methods that are similarly indirect?

Future Directions: Better Attack Models

- Add new attack models other than norm balls
- Study messy real problems in addition to clean toy problems
- Study certification methods that use other proof strategies besides local smoothness
- Study more problems other than vision

Future Directions: Security Independent from Traditional Supervised Learning

- Common goal (AML and ML)
  - *just make the model better*
- They still share this goal
- It is now clear security research must have some independent goals. For two models with the same error volume, for reasons of security we prefer:
  - The model with lower confidence on mistakes
  - The model whose mistakes are harder to find

Future Directions

- A stochastic model that does not repeatedly make the same mistake on the same input
- A model whose mistakes are less valuable to the attacker / costly to the defender
- A model that is harder to reverse engineer with probes
- A model that is less prone to transfer from related models

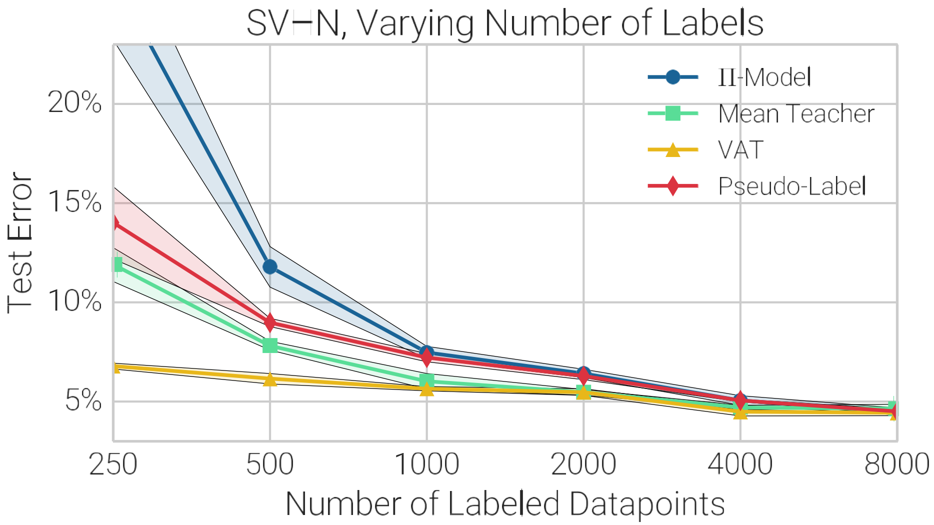
Some Non-Security Reasons to Study Adversarial Examples

Improve Supervised Learning (Goodfellow et al 2014)

Improve Semi-Supervised Learning (Miyato et al 2015)

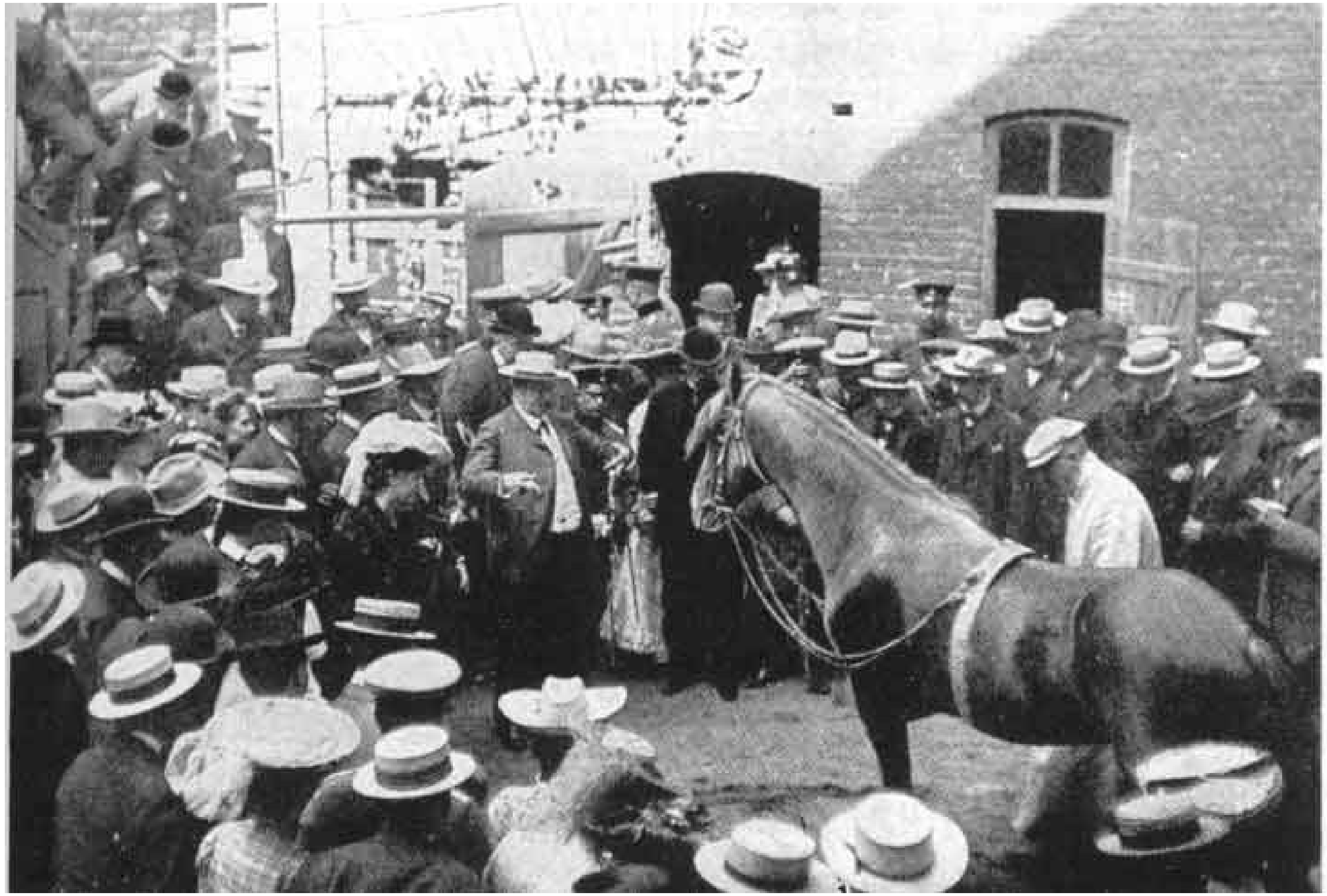
Understand Human Perception

Gamaleldin et al



(Oliver+Odena+Raffel et al, 2018)

Clever Hans



(“Clever Hans, Clever Algorithms,” B n)

Get involved!

https://github.com/tensorflow/cleverhans

cleverhans

Thanks

- Ian Goodfellow and Nicolas Papernot

- Collaborators
  - ....