PennState

CSE 543: Computer Security

Module: Adversarial Machine Learning

Recent Trends in Adversarial Machine Learning

Asst. Prof. Syed Rafiul Hussain

Department of Computer Science and Engineering

The “security” and ”safety” questions?

PennState

In security, we ask two questions of any new technology

1. Can the technology be abused by an adversary?
2. Can the technology be used against us?

Hint: the answer is always yes.

Definition of Adversarial ML

PennState

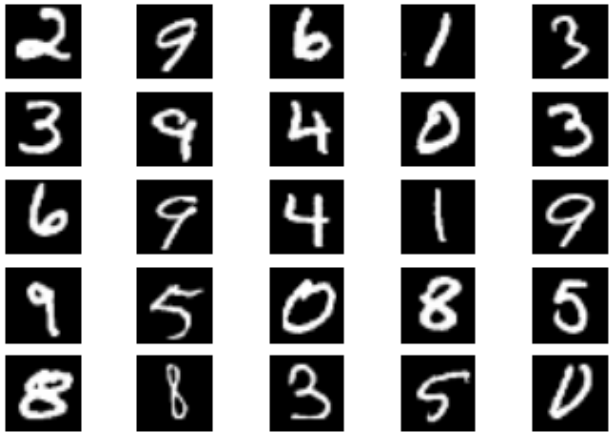
"Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake"

(Goodfellow et al 2017)

How it works … training …

PennState

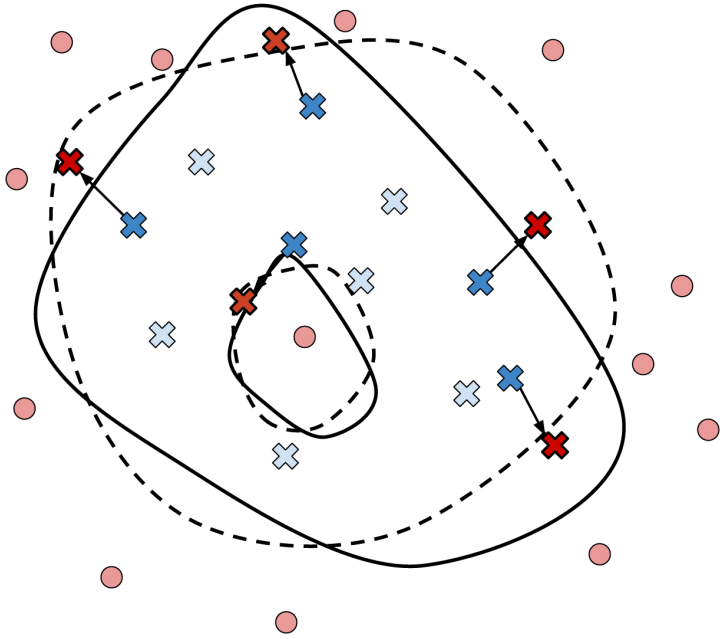
Training Data



Learning Algorithm

(Deep Learning, Decision Trees, others …)

Model



Learning: find *classifier* function that minimize a cost/loss (~model error)

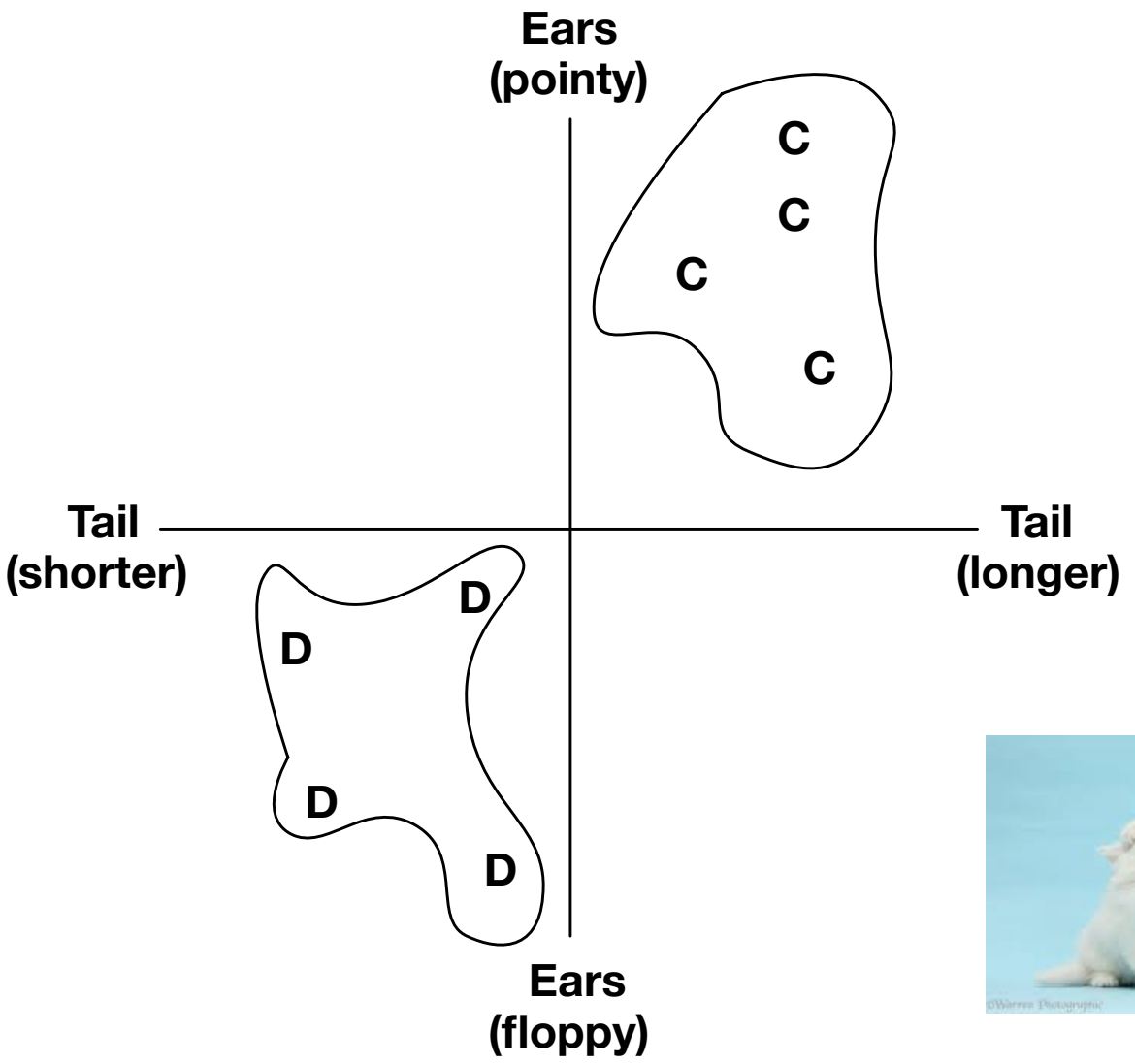
How it works … run-time …

Machine Learning Classifier

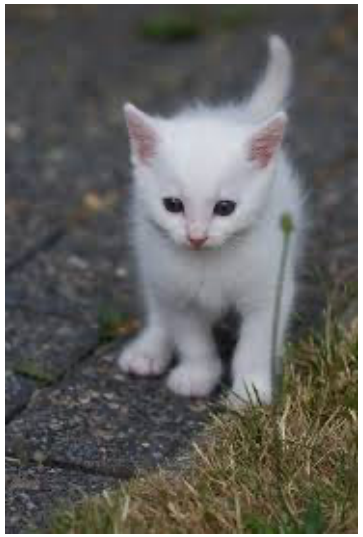
[0.01, 0.84, 0.02, 0.01, 0.01, 0.01, 0.05, 0.01, 0.03, 0.01]

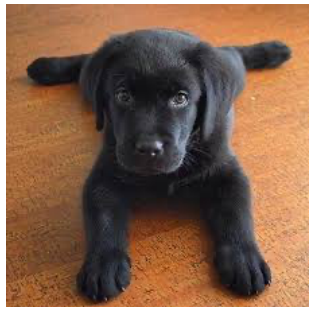
Inference time: which ”class” is most like the input sample

Model (training time …)

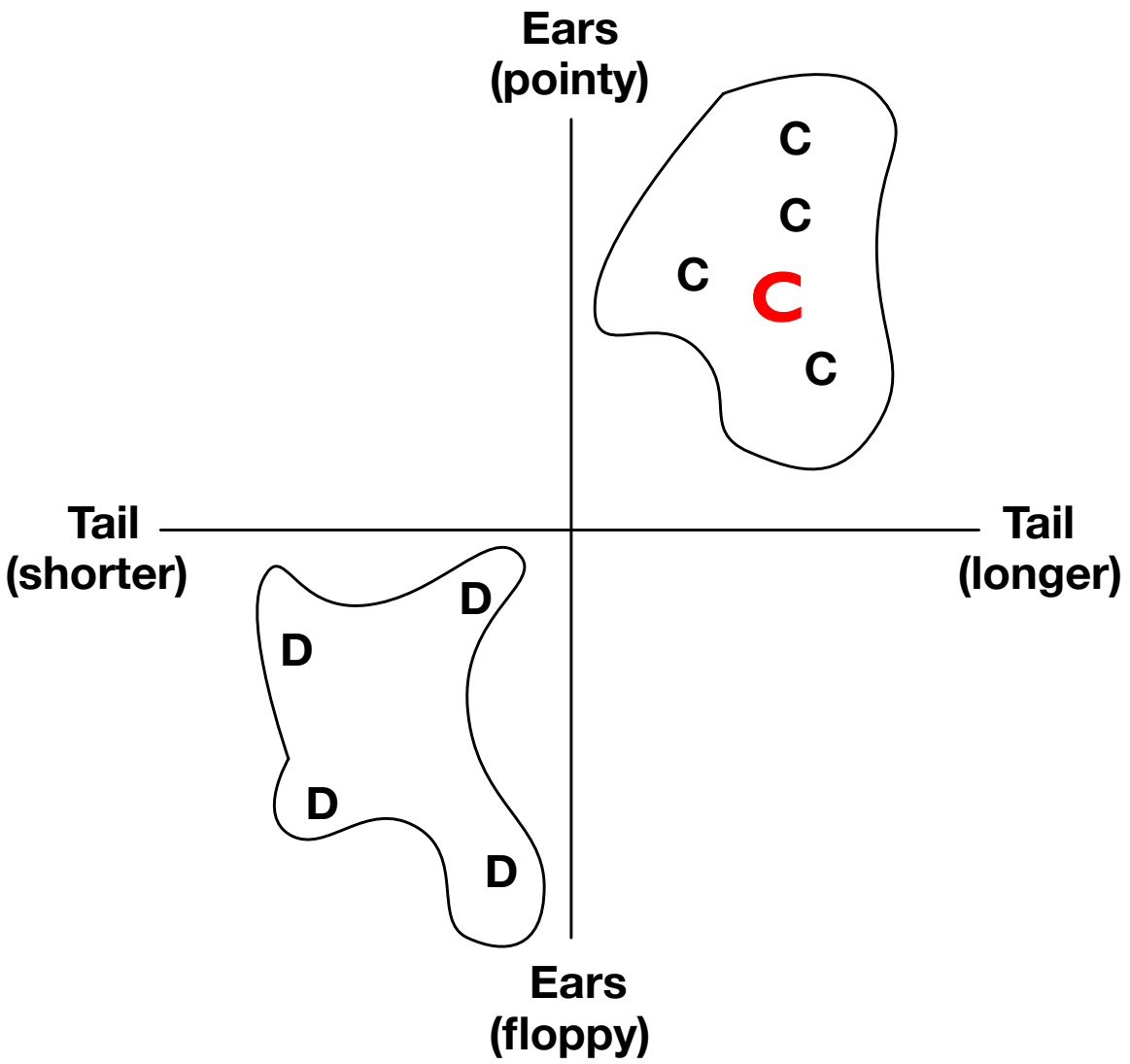


- The learning algorithm identifies the regions of the input feature space that best represent the particular classes …





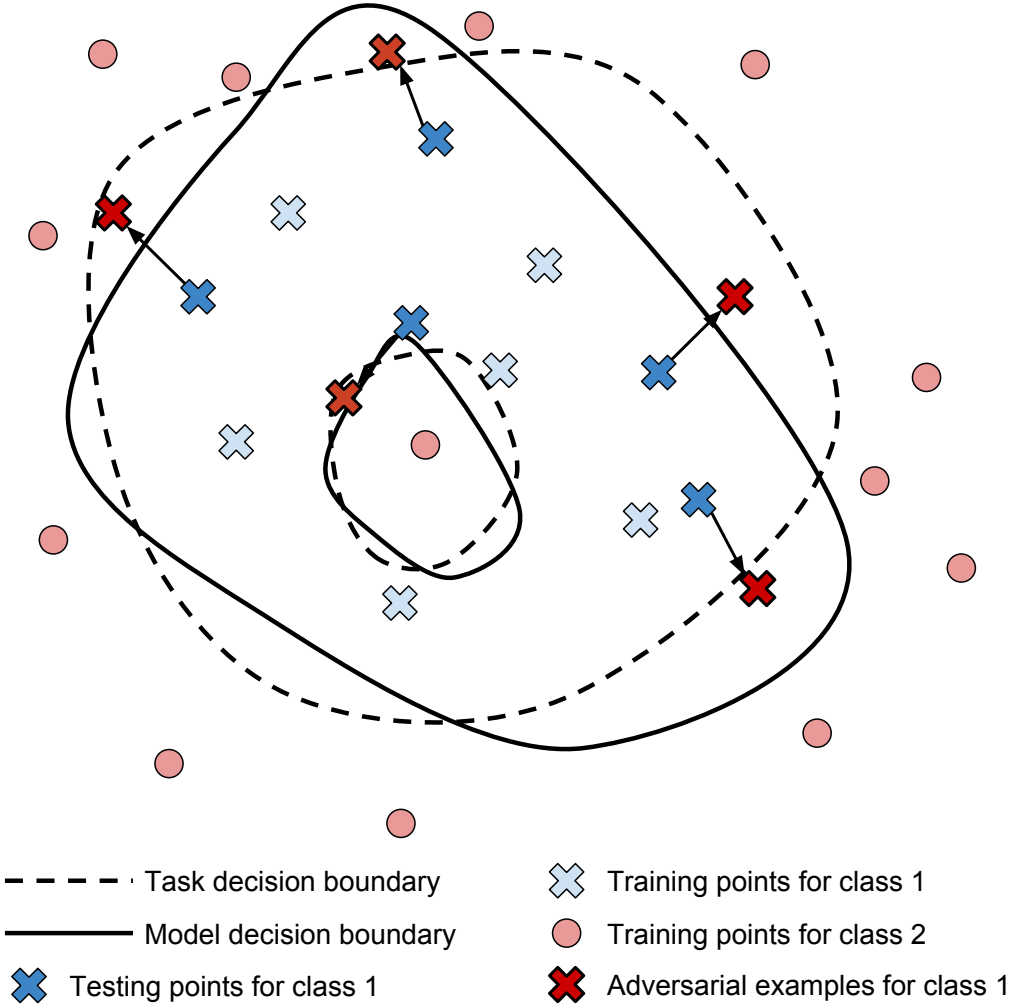
Model (inference time …)



- At run time selects looks at the input features and determines which of the classes best represents the input sample

An Example of Misclassification

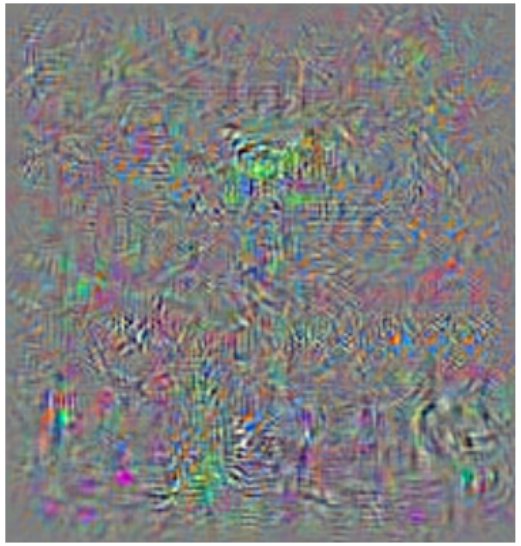
PennState



Adversarial Example

Schoolbus

+



Perturbation
(rescaled for visualization)

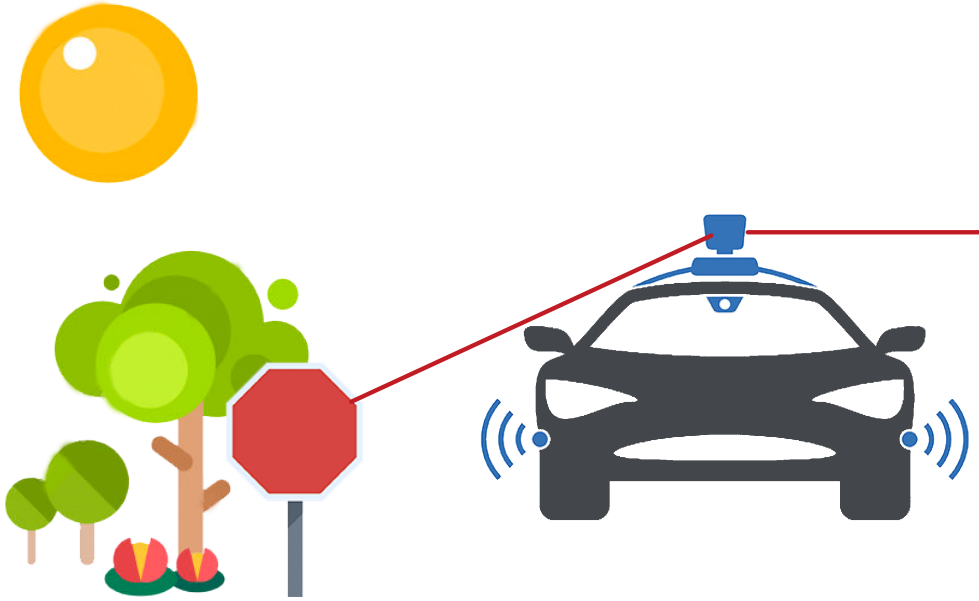
=

Ostrich

(Szegedy et al, 2013)

Another example …

PennState



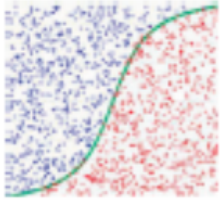
Let's play a game ….

PennState

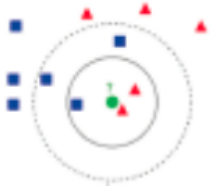
Adversarial Examples

PennState

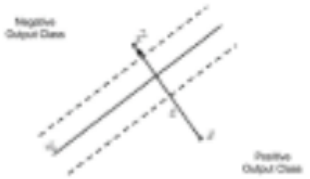
… beyond deep learning



Logistic Regression



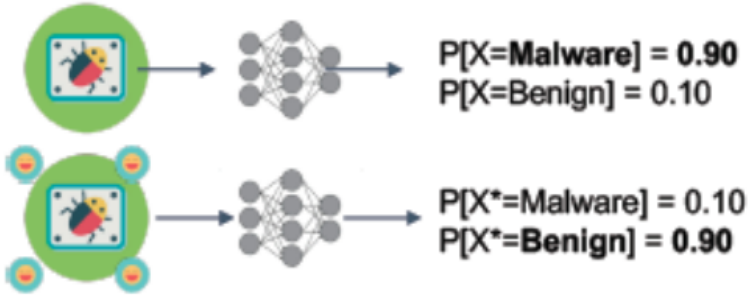
Nearest Neighbors



Support Vector Machines

Decision Trees

… beyond computer vision



Threat Model

PennState

- 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

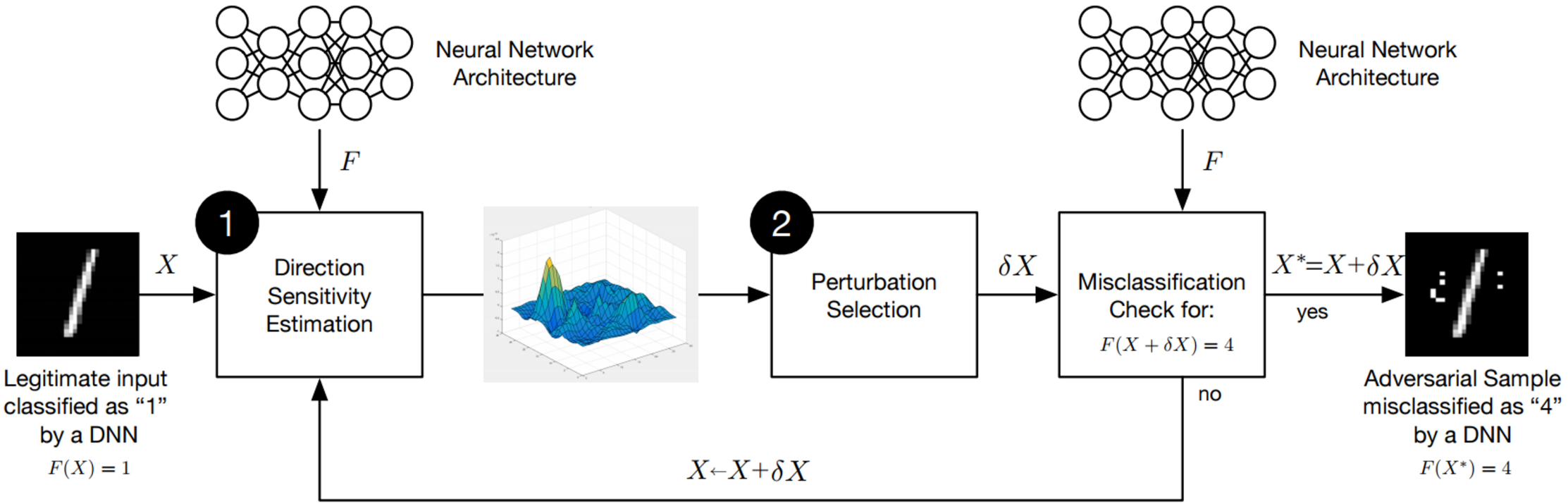
Attack Steps

PennState

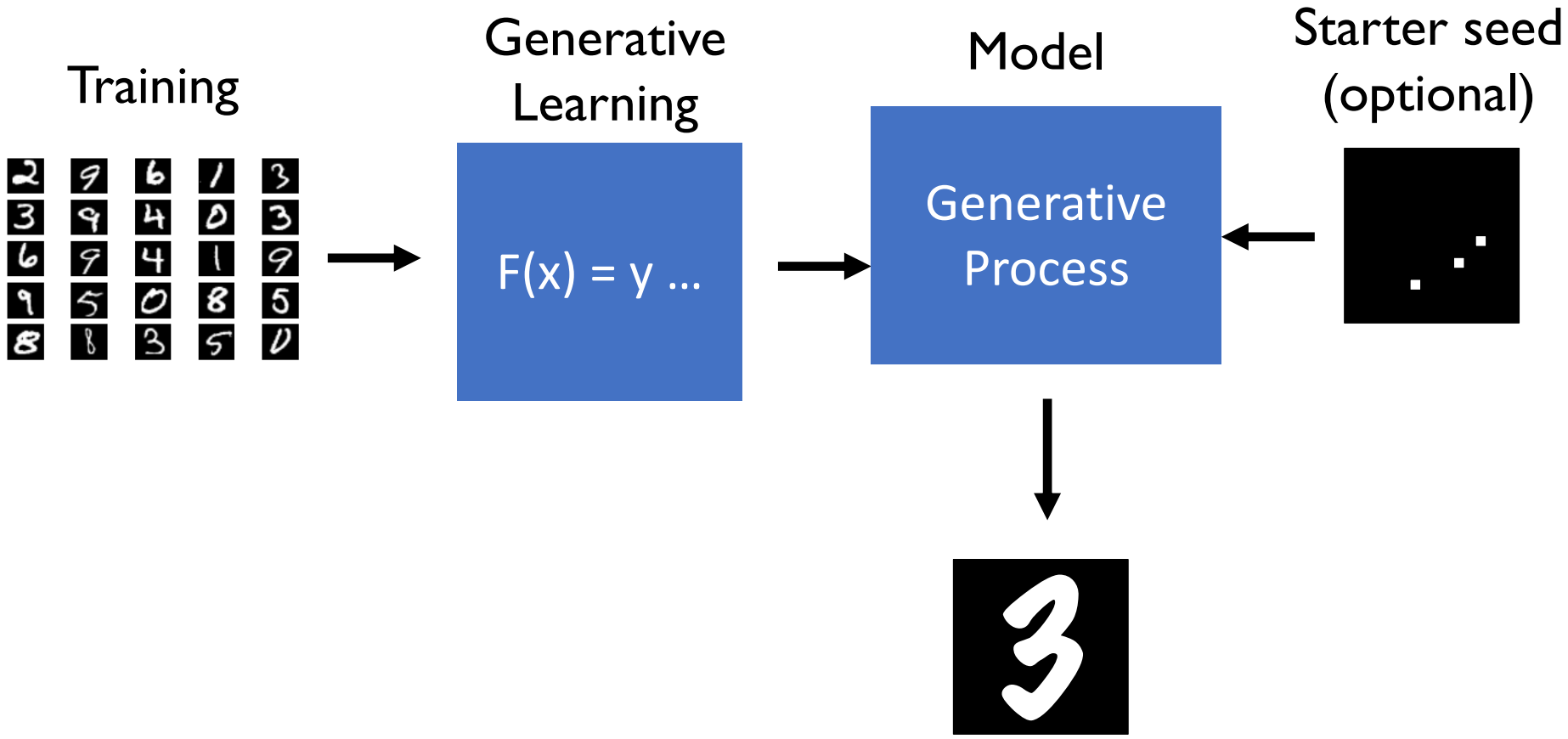
- Adversary's problem
  - Given: $x \in X$
  - Find δ
    - $\min_{\delta} \quad \mu(\delta)$
    - *Such that:* $F(x+\delta) \in T$
      - Where: $T \subseteq Y$

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

Adversarial samples … intuition …



Using machine learning to create …



Inference time: produce output which is most like training inputs ..

Creating reality? Fake news

PennState

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

Creating reality? Fake reviews

PennState

"I love this place. I have been going here for years and it is a great place to hang out with friends and family. I love the food and service. I have never had a bad experience when I am there." (5 stars!)

"I had the grilled veggie burger with fries!!!! Ohhhh and taste. Omgggg! Very flavorful! It was so delicious that I didn't spell it!!" (5 stars!)

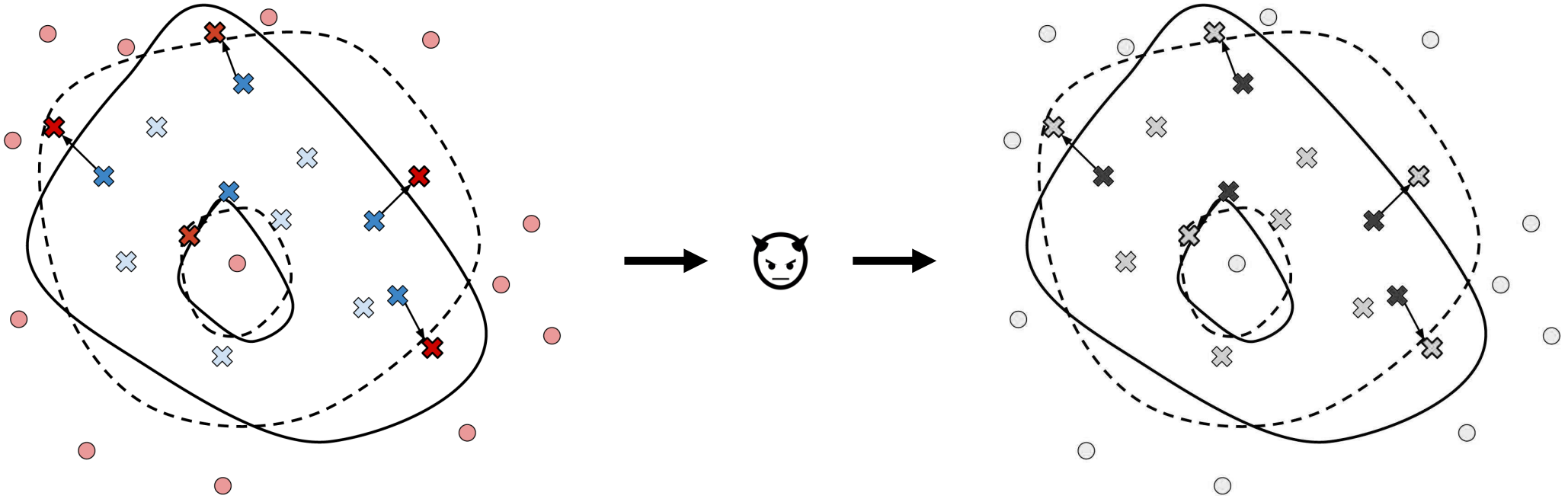
"I was so excited to try this place out for the first time and the food was awful. I ordered the chicken sandwich and it was so salty that I could not eat it. I was so disappointed that I was so disappointed in the food. I was so disappointed that I was so disappointed with the service." (1 star)

Source: https://arxiv.org/pdf/1708.08151.pdf, "Automated Crowdturfing Attacks and Defenses in Online Review Systems" (**Yao et al.**)

Model theft …

PennState

- Machine learning models (generally) expose a lot of information about their internals just through normal interactions … and can be reproduced.



Consequences …

PennState

- *Security*– you don't have to know anything about a service to create adversarial samples, you just have to be able to access it.
- *Intellectual property* – if your organization's value is "in" a model, then it can be extracted (and duplicated and misused).
- *Privacy* – the model can tell you a lot about the training data, e.g., patient data.

Google Cloud Platform

LinkedIn

amazon web services™

sas

facebook INSIGHTS

rapidminer

Thanks

PennState

Thanks to Ian Goodfellow, Somesh Jha, Patrick McDaniel, Nicolas Papernot and Berkay Celik for some slides.