SP2023 Week 09 • 2023-03-26

AI Hacking II

Anusha

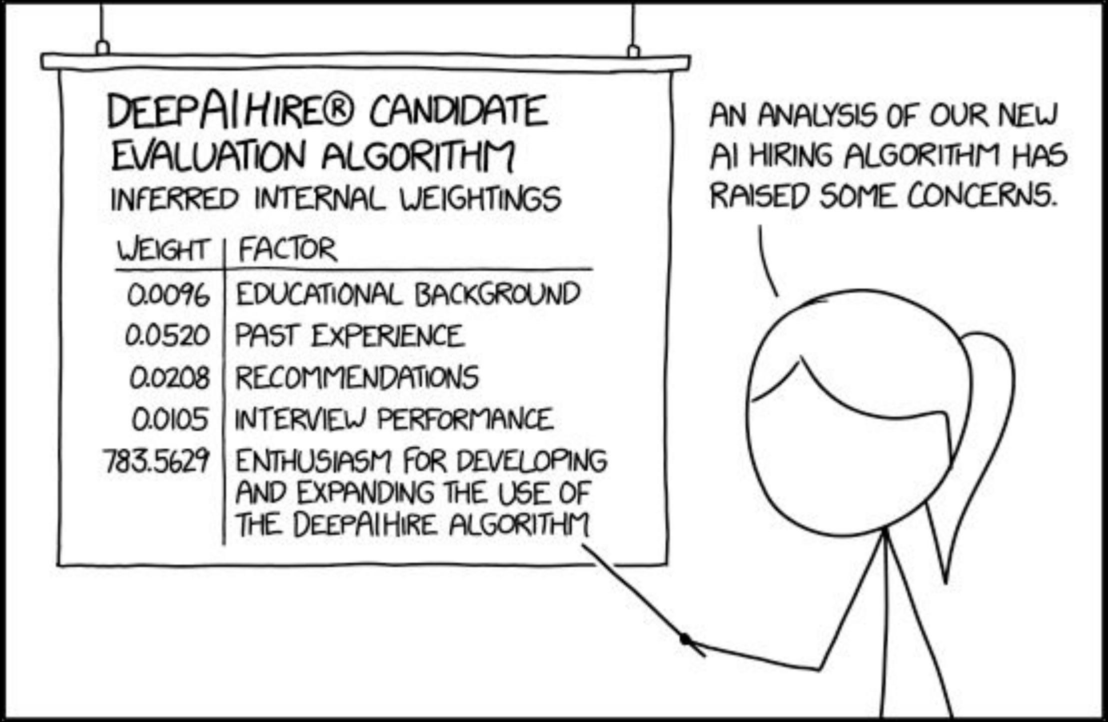
Announcements

- No meeting next Thursday/Sunday
    - We’re off to CypherCon :)

- Fill out feedback form for our research paper please
    - https://forms.gle/kYg16ZJicwuVwTca6

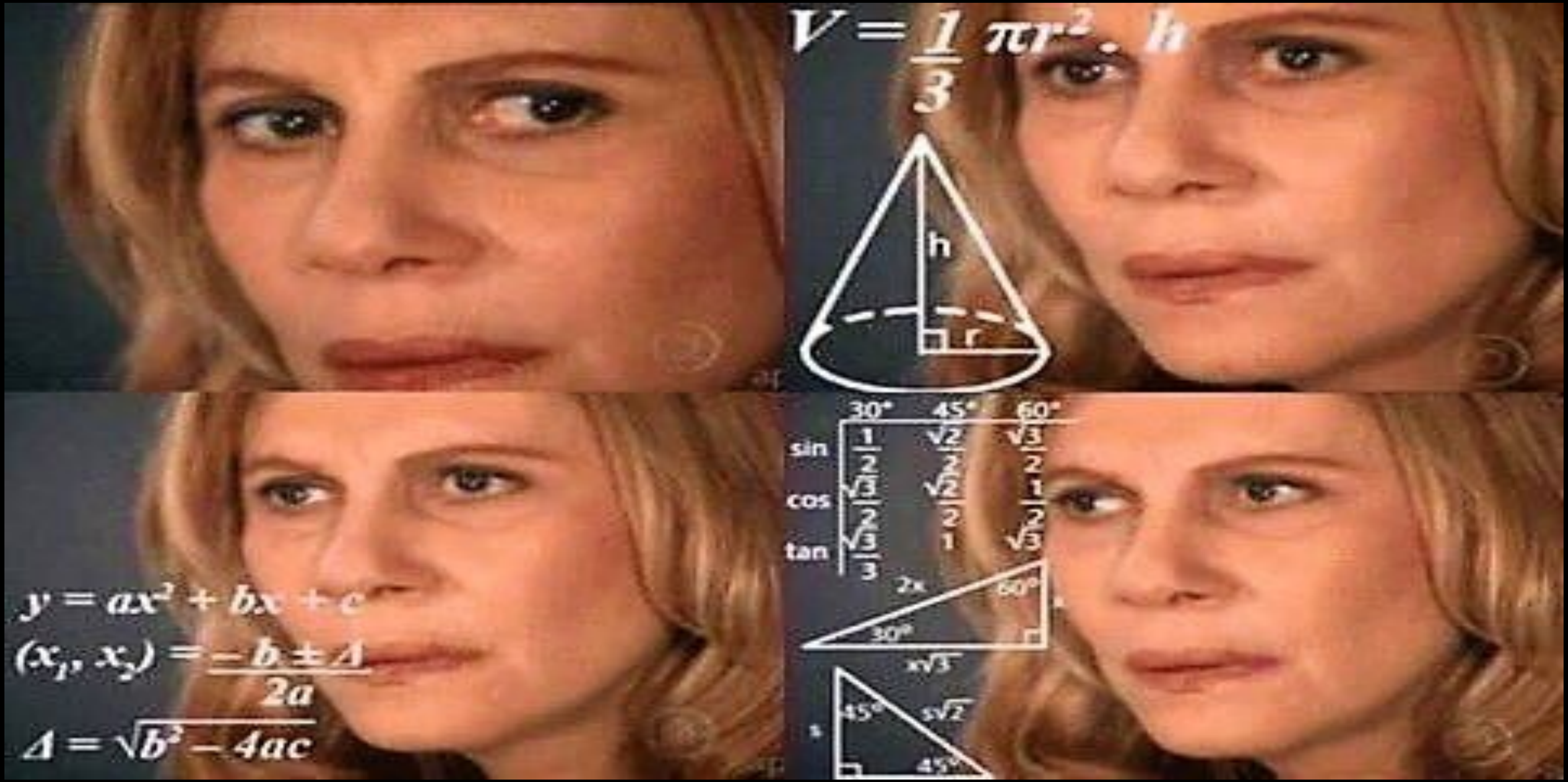
ctf.sigpwny.com

sigpwny{when_pigs_fly}

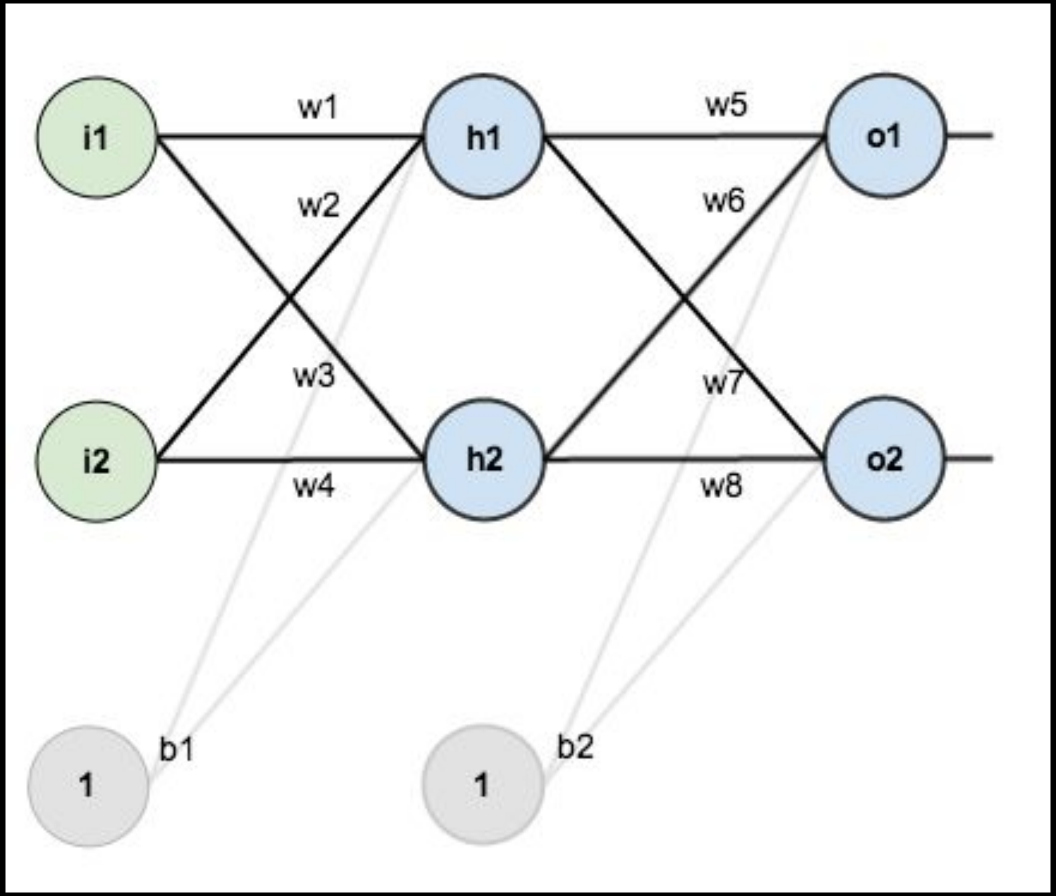


Background

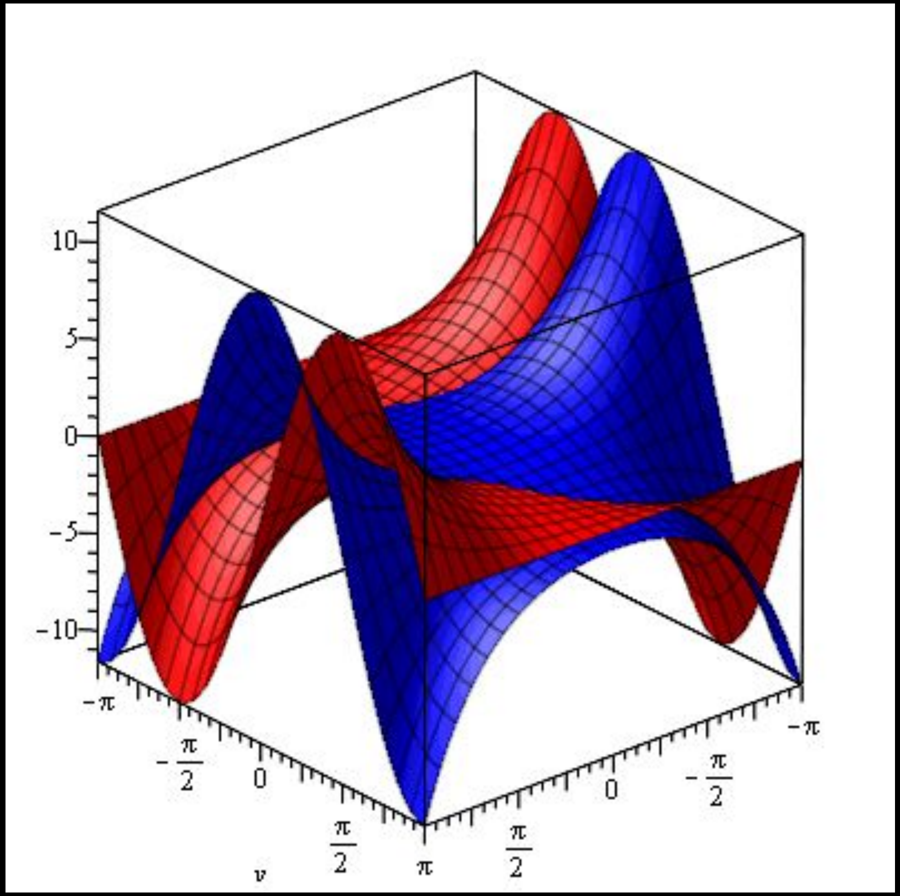
What is AI?


$V = \frac{1}{3}\pi r^2 \cdot h$
h
r
30° 45° 60°
sin
cos
tan
$y = ax^2 + bx + c$
$(x_1, x_2) = \frac{-b \pm \Delta}{2a}$
$\Delta = \sqrt{b^2 - 4ac}$

What is AI?



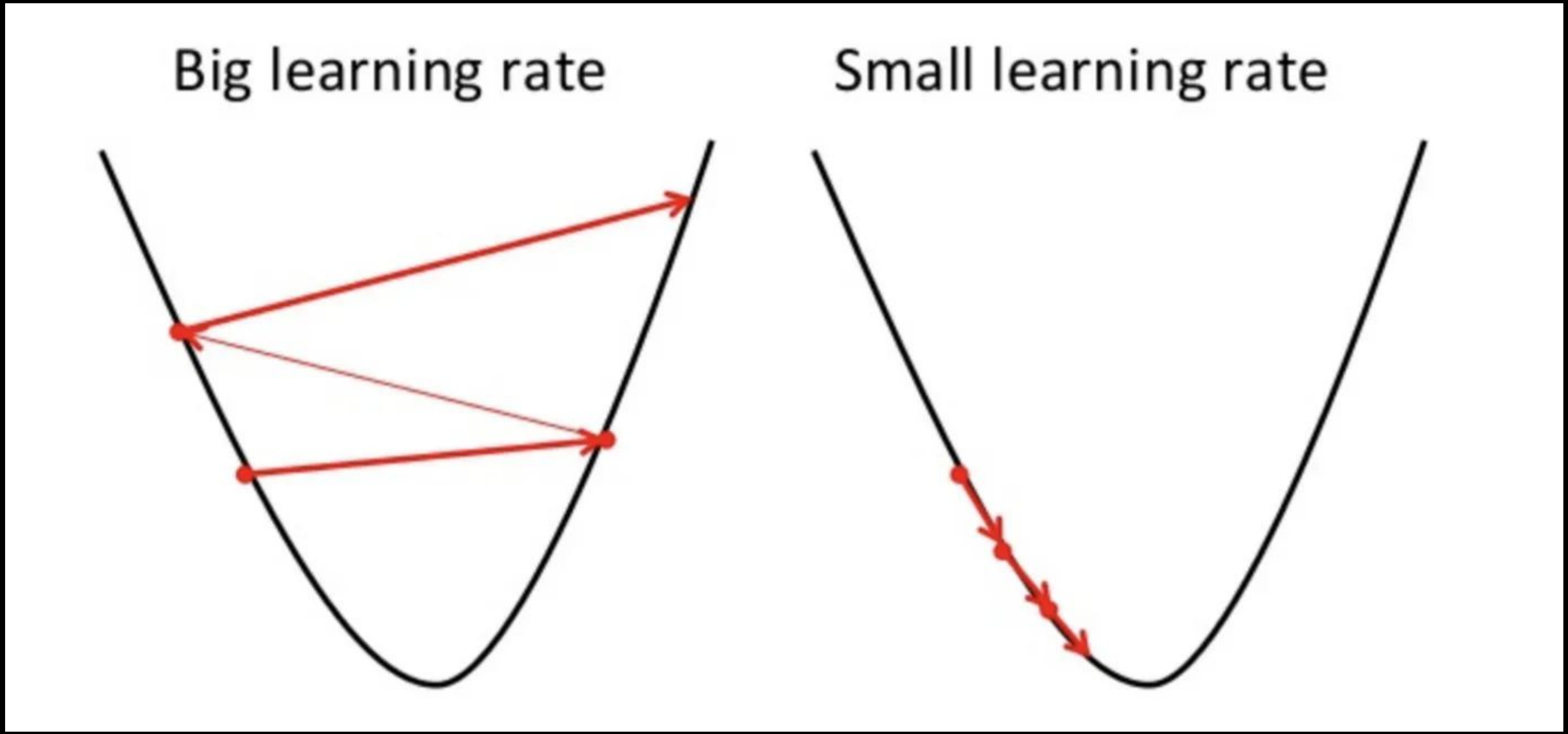
i1
i2
h1
h2
o1
o2
w1
w2
w3
w4
w5
w6
w7
w8
1
b1
1
b2



10
5
0
-5
-10
-π
-π/2
0
π/2
π
v
π/2
0
-π/2
-π

How do we create AI models?

- Perform gradient descent (optimization on problem to minimize error)



How do we create AI models?

- iterate over training data multiple times
    - each iteration is known as an epoch
- use loss functions to determine the performance of a model
    - higher loss means more error present in the model’s predictions

How can AI be insecure?

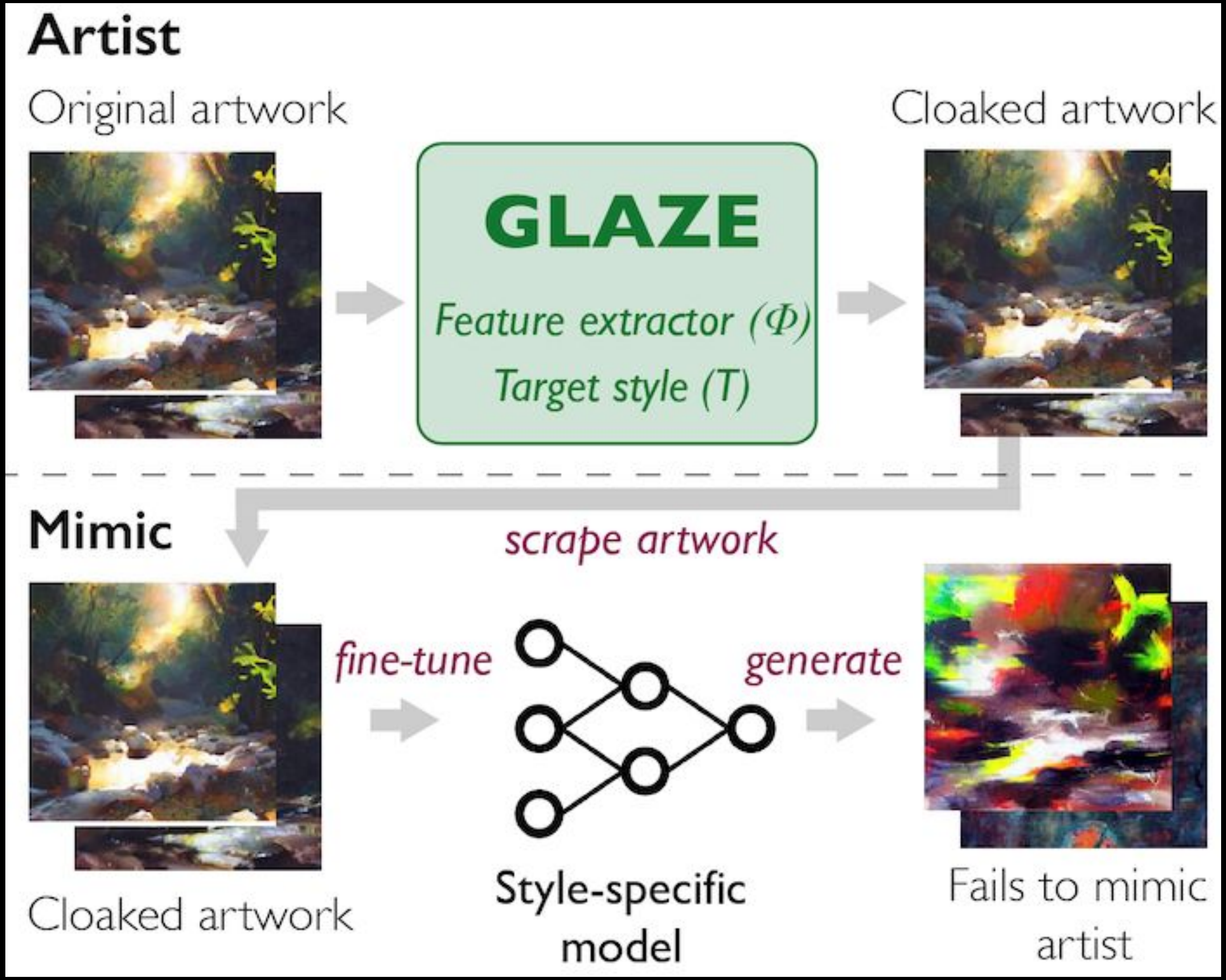
- Dataset issues
  - Data may be mislabeled/collected incorrectly/preprocessed wrong
  - There may also be malicious data in large datasets
- Model issues
  - Models may be vulnerable to malicious input (adversarial examples)
  - They might also be vulnerable to extraction/trojaning attacks

Poisoning

Dataset Poisoning

- Malicious data present in a dataset <u>during training</u>
- Model learns incorrect information from the dataset
- Only possible if attacker has access to dataset before model creation
    - Also important to consider in situations where model is trained using human feedback

Dataset Poisoning

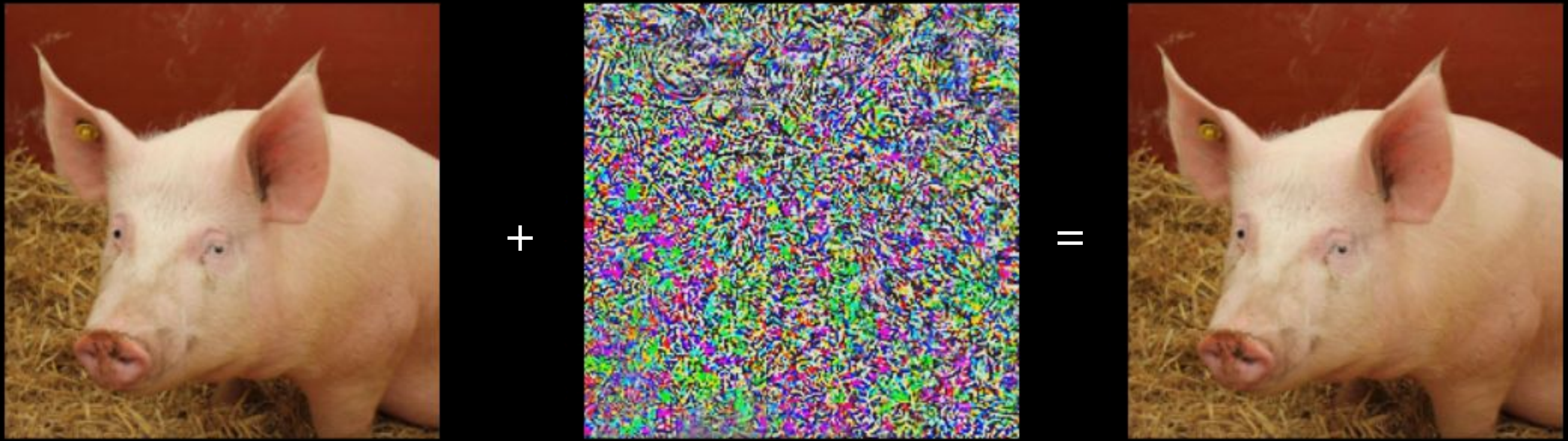


Evasion Attacks

Adversarial Examples

- Malicious input designed to fool a model into undesired behaviour
- Imperceptibly changed input - the goal is to trick a model into behaving in ways it shouldn’t

Adversarial Examples

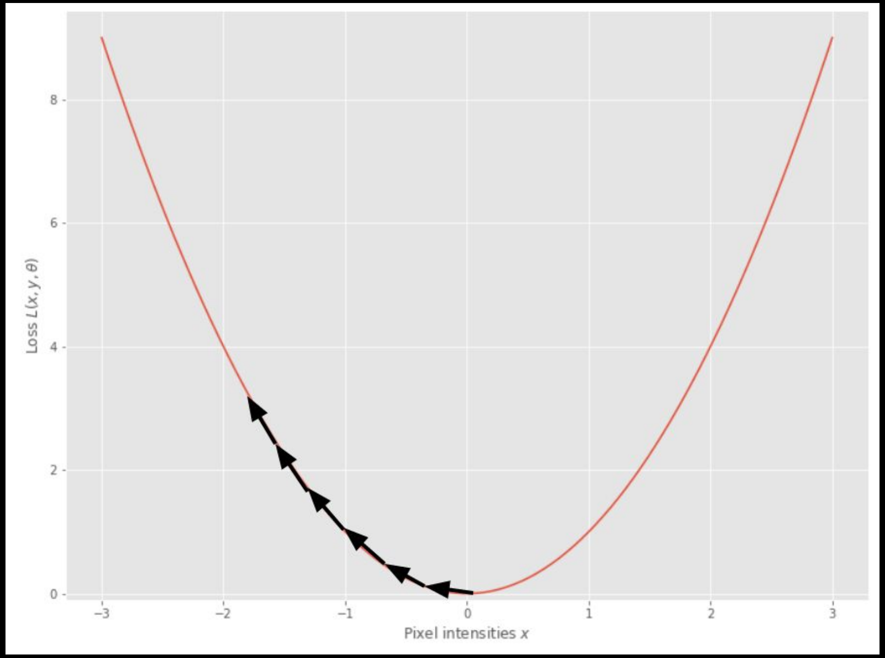


Class: pig

Class: airliner

Adversarial Example Generation

- How do we create noise that optimally fools a given model?
  - The answer is… complicated (and an ongoing area of research!)
- The most intuitive methods use gradient ascent, where input data is adjusted to maximize loss



Adversarial Example Generation

- You don’t always need to have access to the model or its gradients
  - There are many papers devoted to showing various attacks on black box models
- Attacks are <u>transferable</u>, meaning that attacks that work on one model can often transfer to an unknown model
  - You can use surrogate models trained on similar data to create adversarial examples against an unknown model
  - These methods usually require oracle access, where you have access to the output of the model you want to attack

Adversarial Defenses

- The most common defense is adversarial training
    - incorporate adversarial examples into the training process
    - provides data that helps the model disregard nonrobust features that may be present
- There are also defenses that prevent the attacker from gaining access to gradient information
    - one example is defensive distillation

Extraction Attacks

Model Extraction

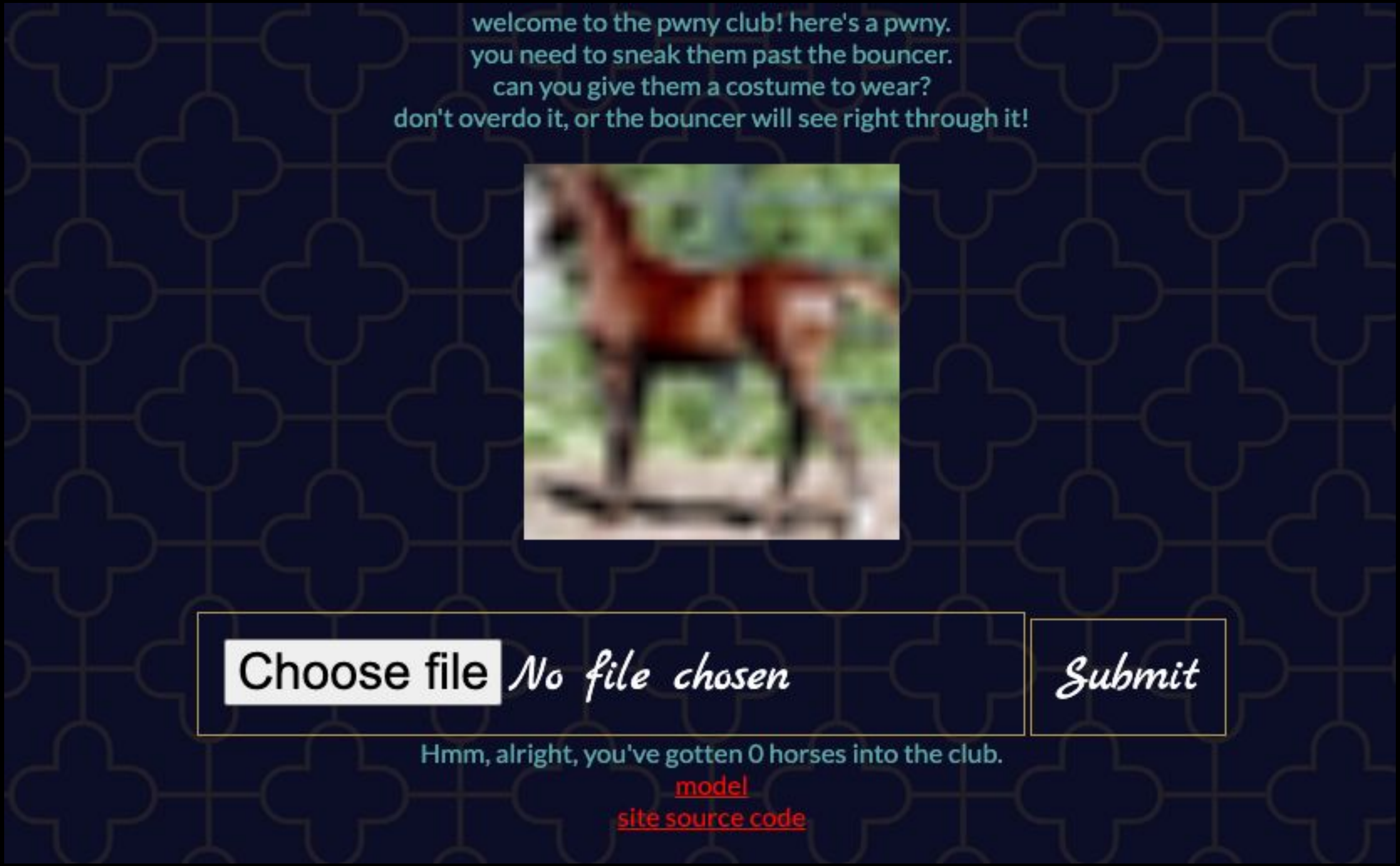
- These attacks focus on recreating a model given query access to a private model
- The created model may not be as accurate, but can approach the accuracy of the original model
- These models can then be maliciously used or used in combination with other attacks

CTF Example

Important Tools

- pytorch
- torchvision
- torchattacks
- cleverhans

pwnies_please



pwnies_please

```
criterion = nn.CrossEntropyLoss()  #define loss function
for i, (inputs, labels) in enumerate(dataloaders['test']):
    inputs = inputs.to(device) #move to gpu
    labels = labels.to(device) #move to gpu

    #generate adversarial examples
    inputs = pgd(model_nonrobust, inputs, labels, criterion, k=15, step=0.1, eps=0.4, norm=2)
    outputs = model_nonrobust(inputs)
```

Next Meetings

2023-03-30 - **Next Thursday**

- No meeting because of CypherCon

2023-04-02 - **Next Sunday**

- No meeting because of CypherCon

sigpwny{when_pigs_fly}

SIGPwny