

Lessons Learned from Evaluating the Robustness of Defenses to Adversarial Examples

Nicholas Carlini
Google Research

Lessons Learned from
Evaluating the Robustness of
Defenses to Adversarial Examples

Lessons Learned from
Evaluating the Robustness of
Defenses to **Adversarial Examples**



88% **tabby cat**



adversarial
perturbation



88% **tabby cat**



adversarial
perturbation



88% **tabby cat**



adversarial
perturbation



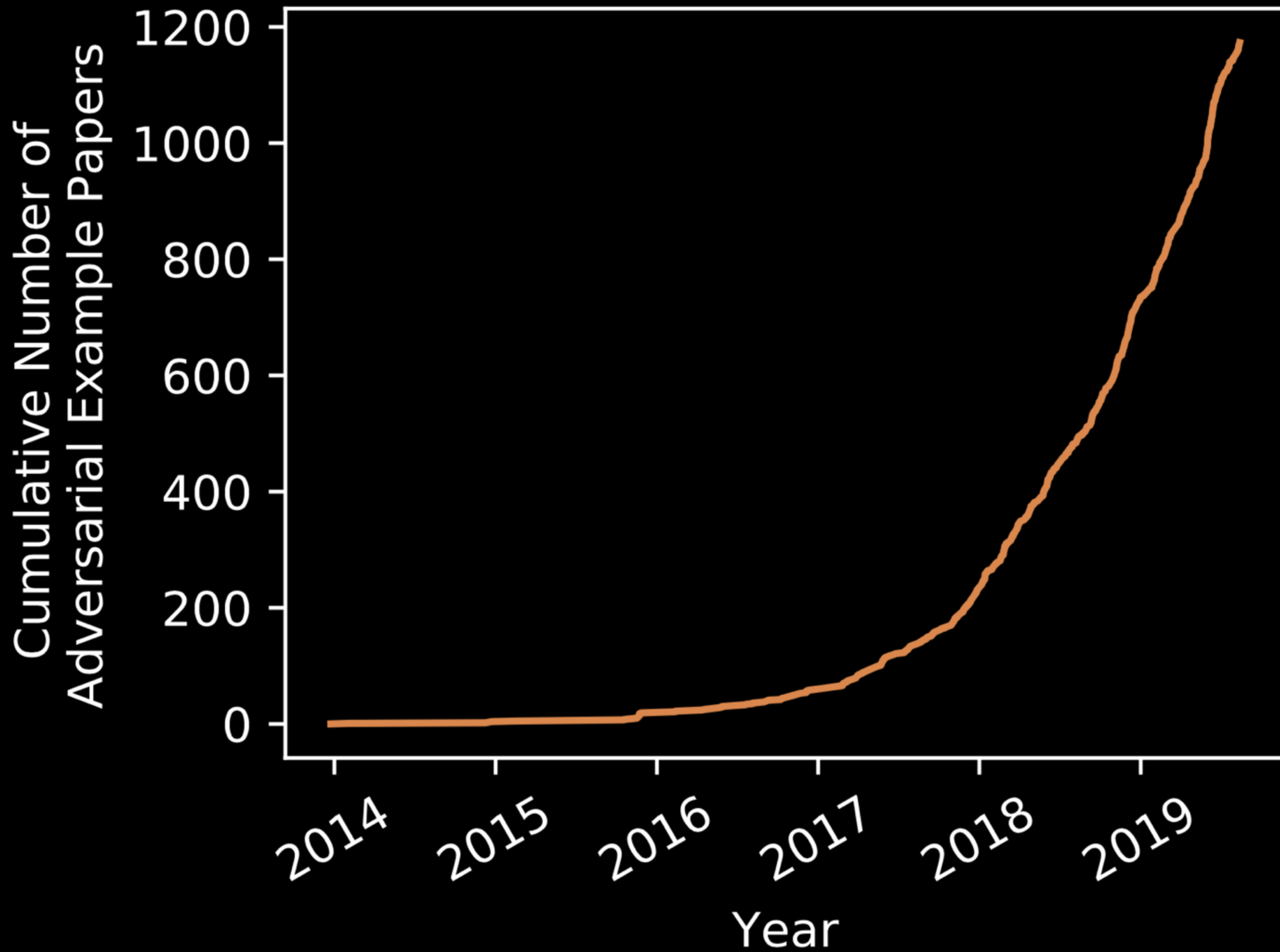
88% **tabby cat**

99% **guacamole**

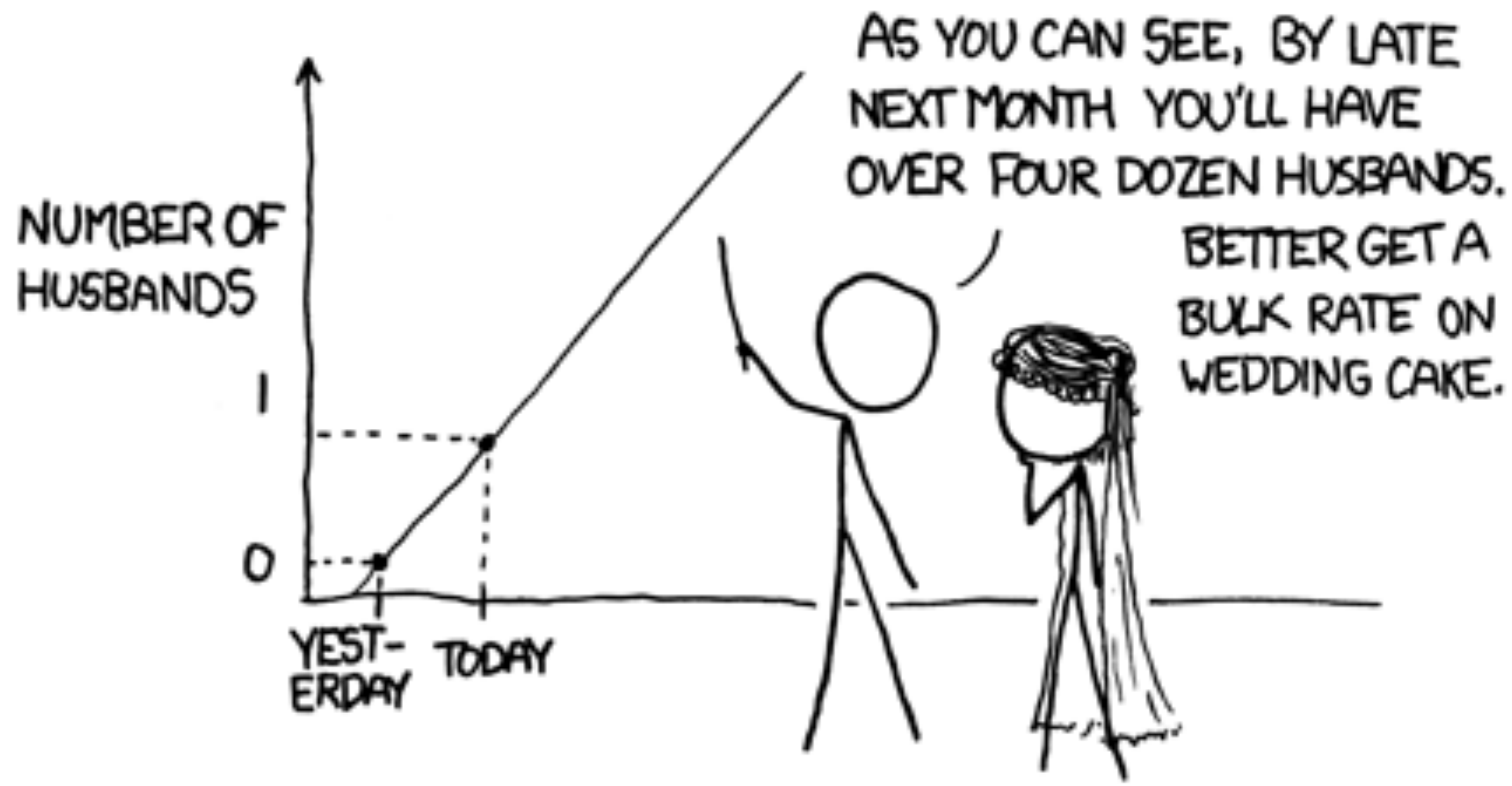
Why should we care about
adversarial examples?

Make ML
robust

Make ML
better



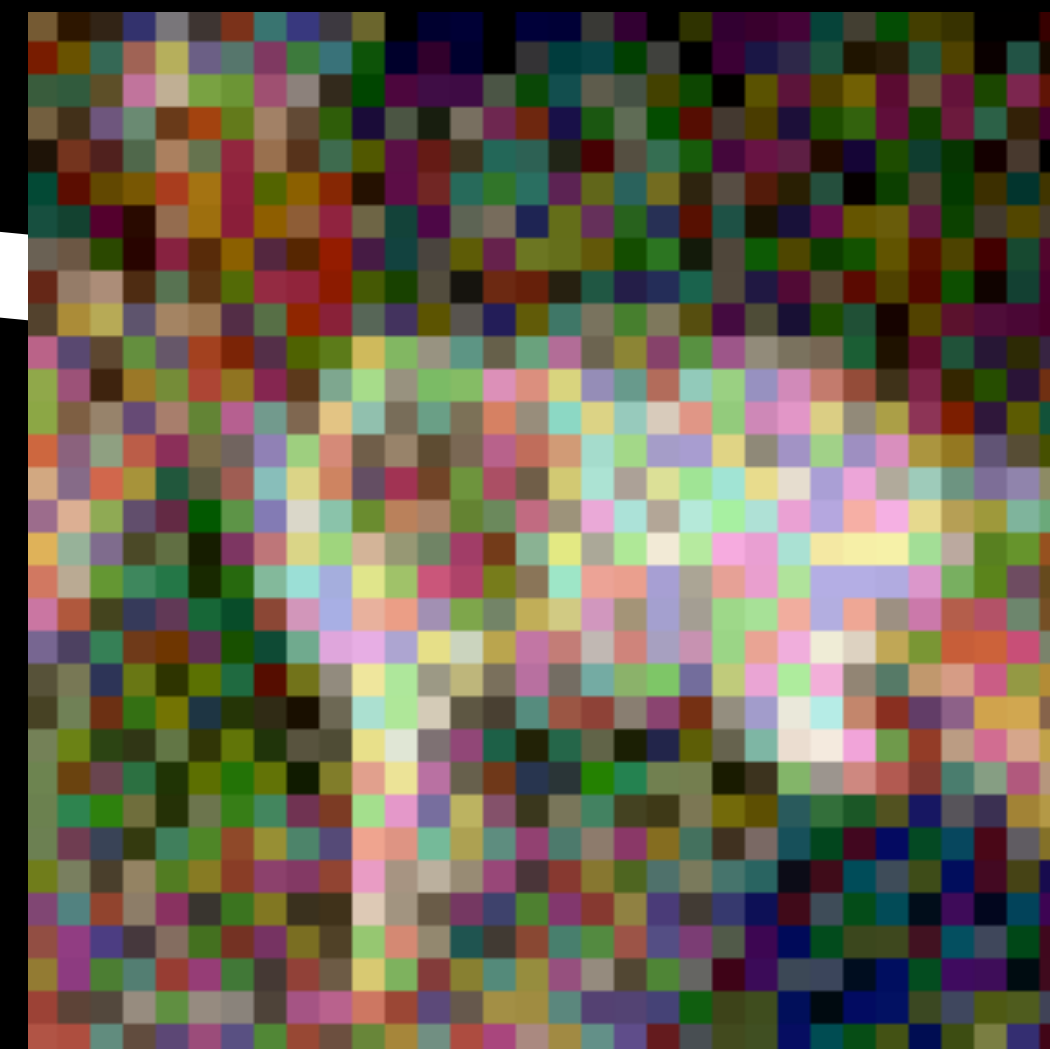
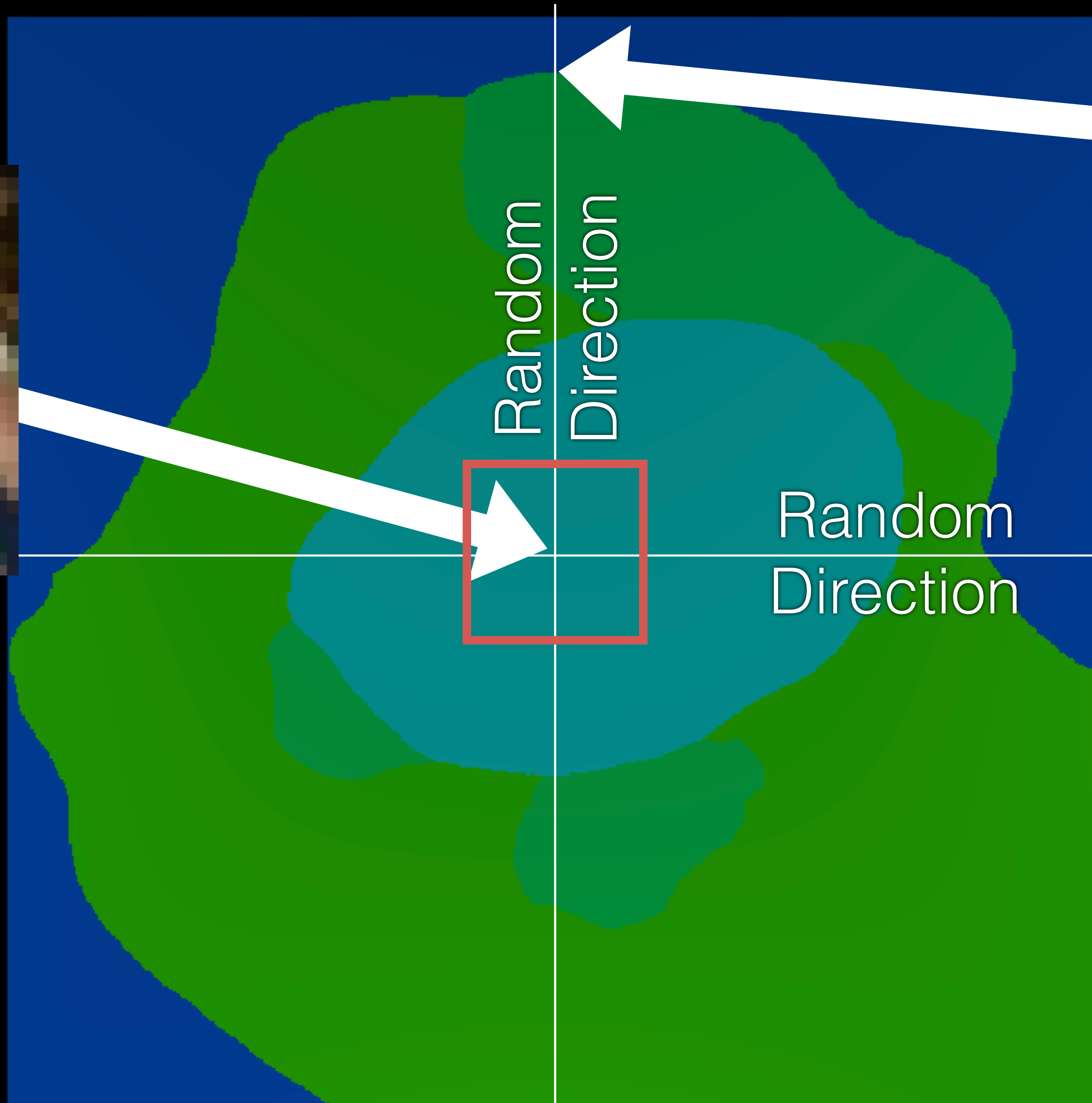
MY HOBBY: EXTRAPOLATING



How do we generate
adversarial examples?



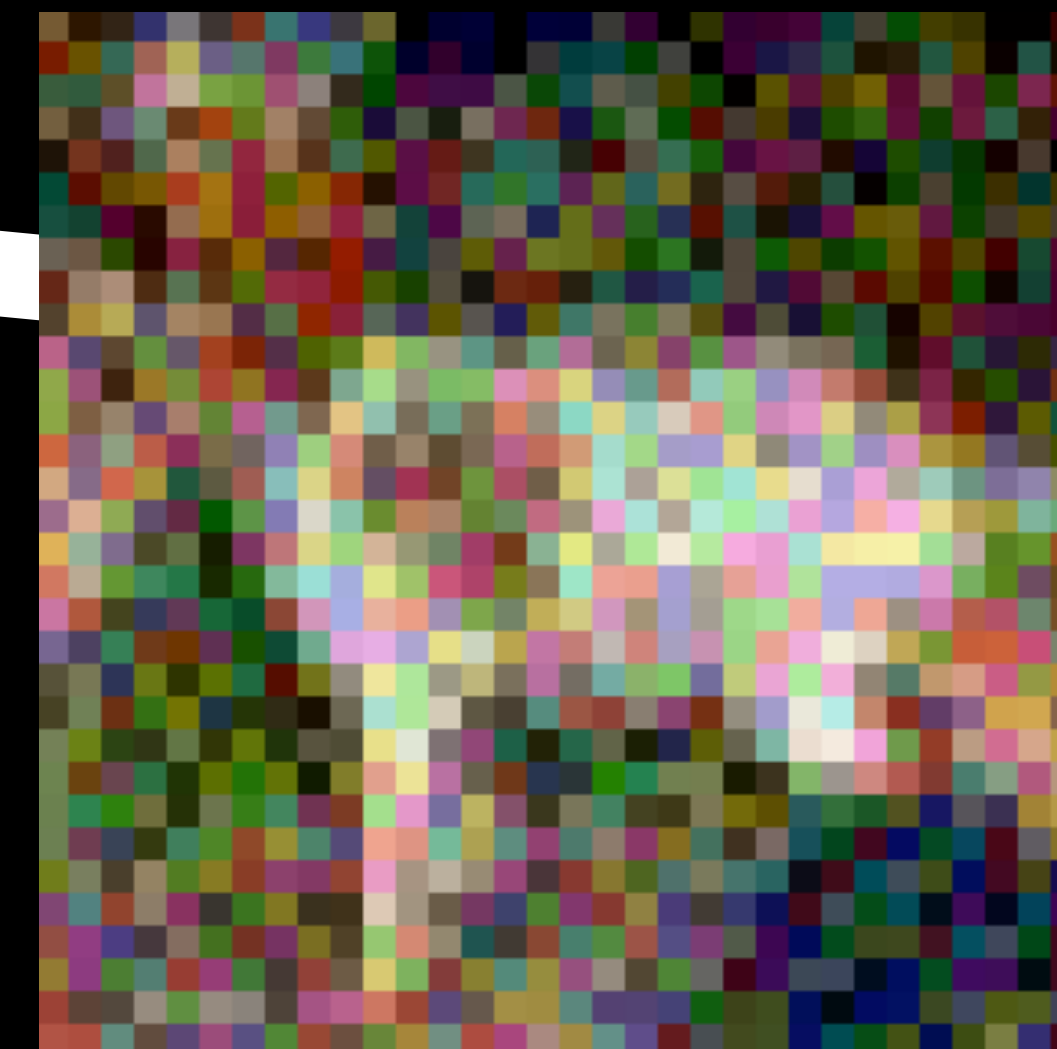
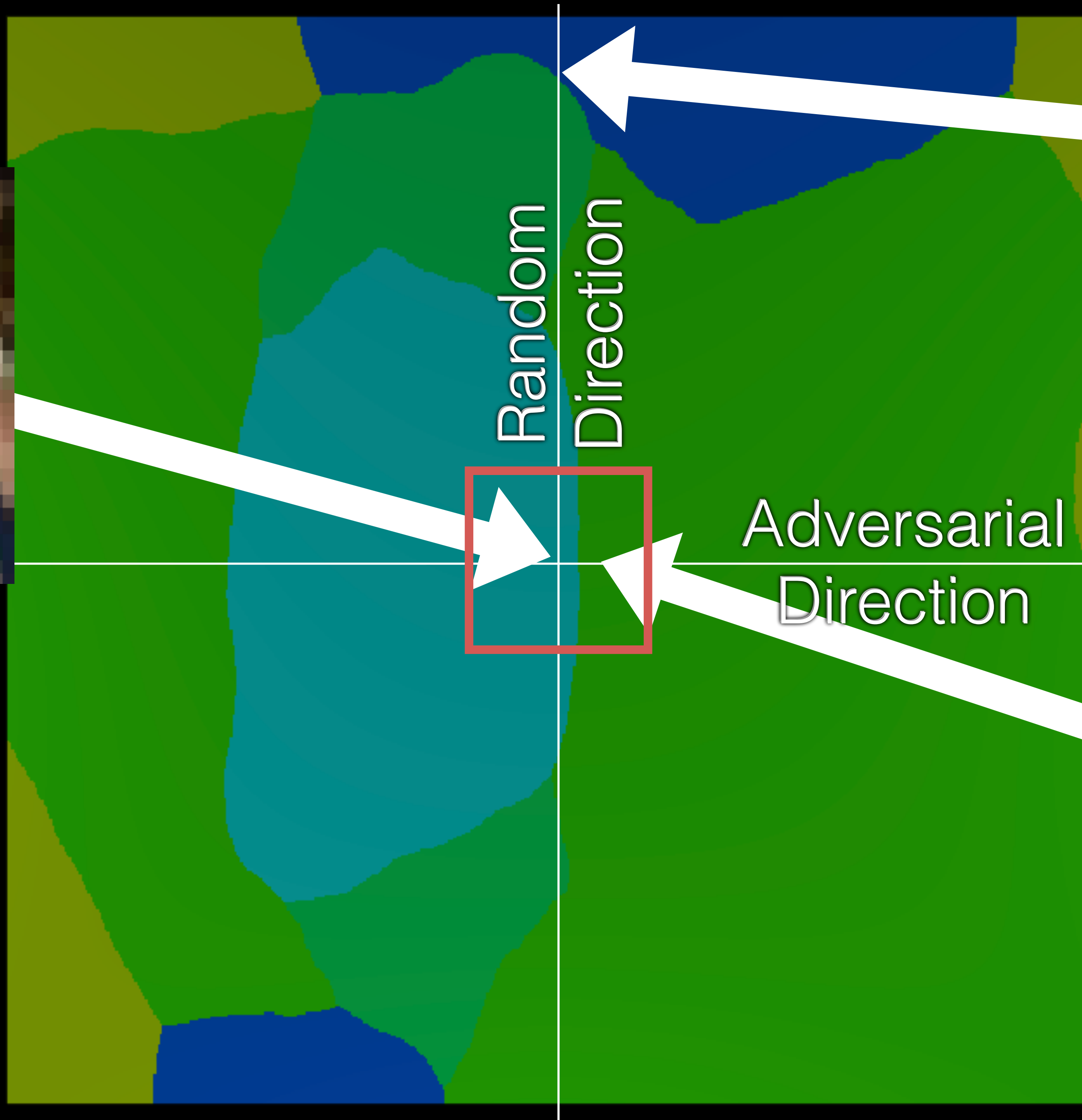
Dog



Truck



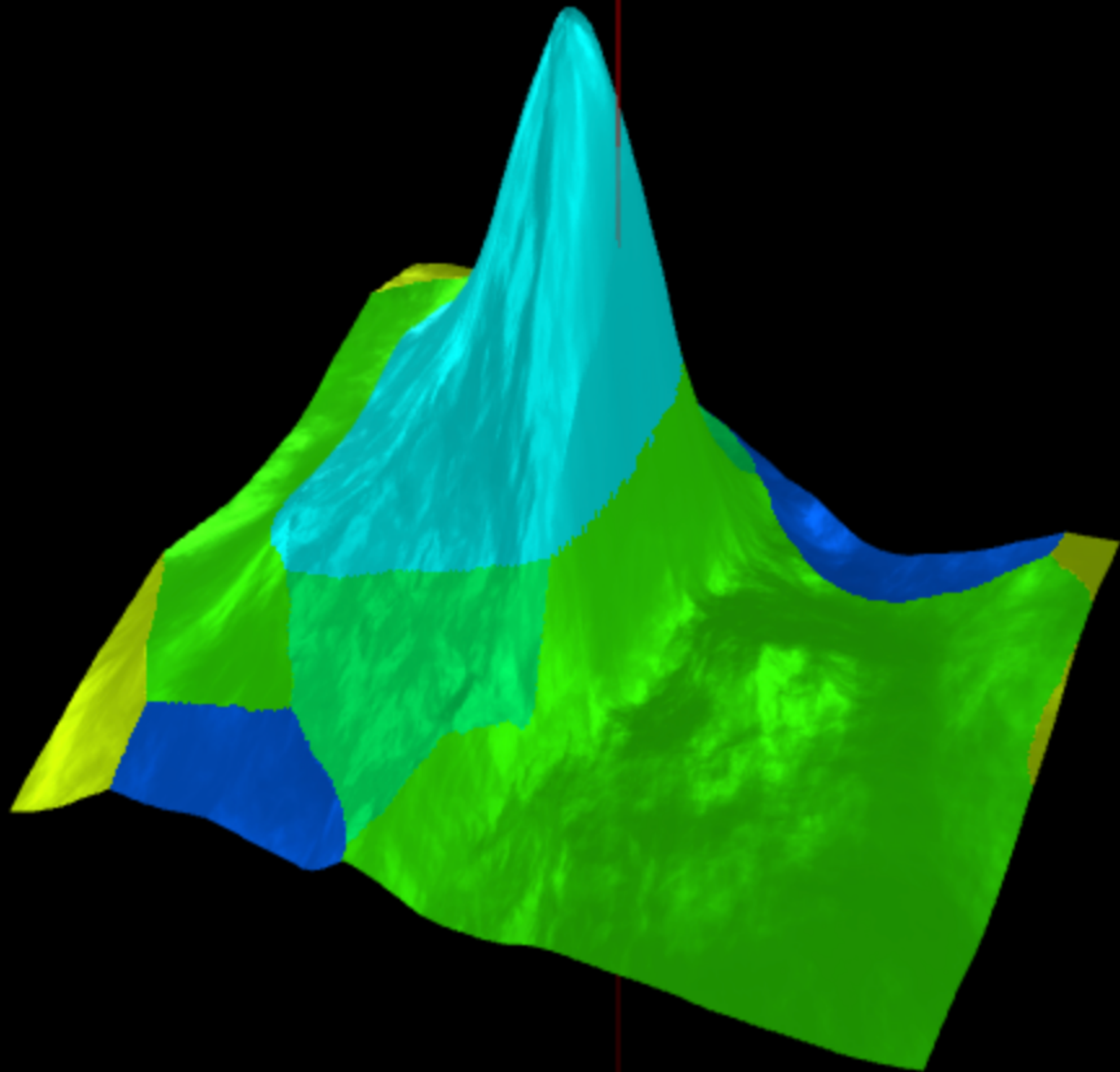
Dog

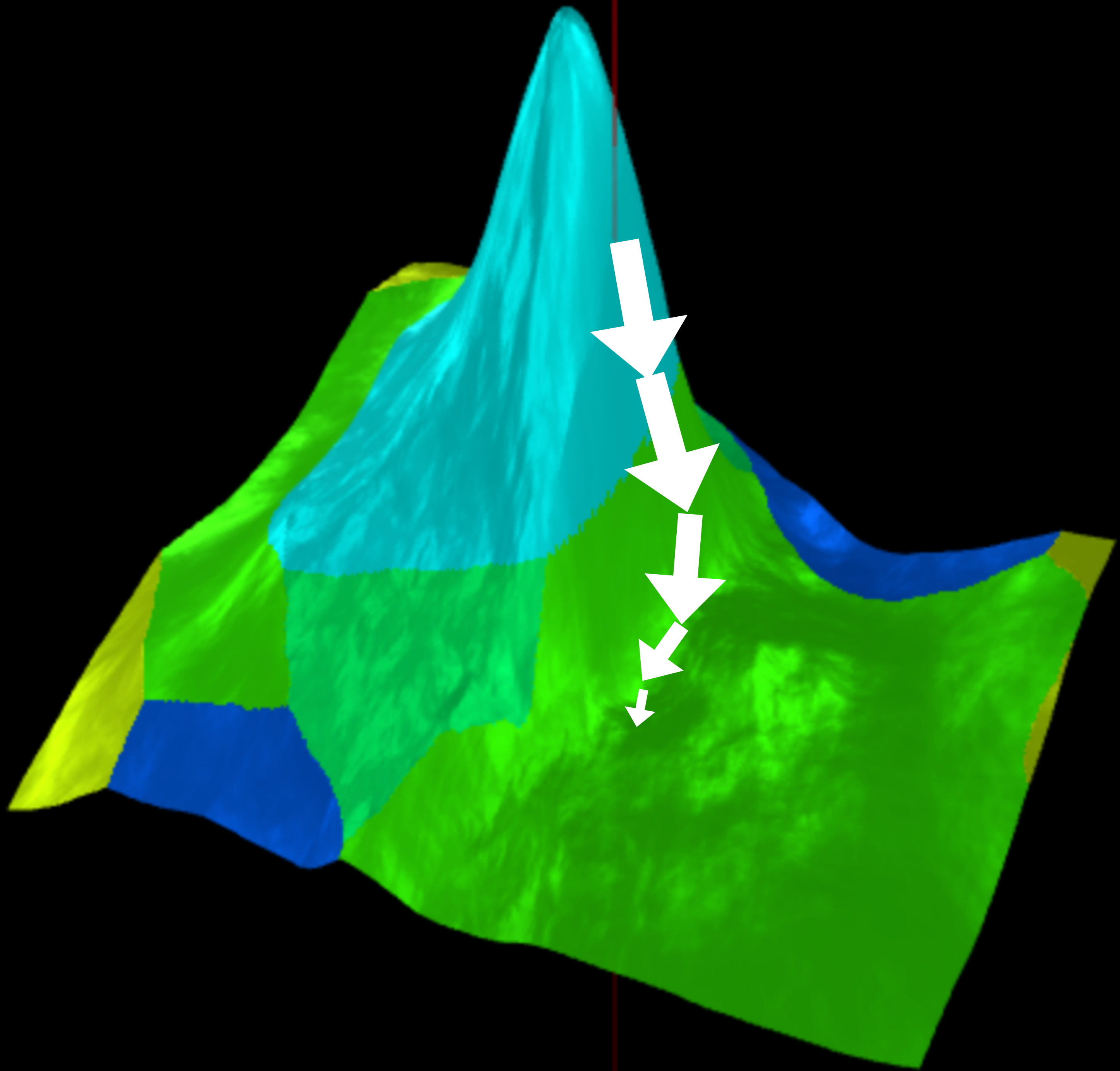


Truck



Airplane





Lessons Learned from
Evaluating the Robustness of
Defenses to Adversarial Examples

A **defense** is a neural network that

1. Is accurate on the test data
2. Resists adversarial examples

For example:

Adversarial Training

Claim:
Neural networks don't generalize

Normal Training

(7, 7)

(8, 3)

Training

Adversarial Training (1)

(7, 7)

(8, 3)

(7, 7)

(8, 3)

Attack

Adversarial Training (2)

(7, 7)

(8, 3)

(7, 7)

(8, 3)

Training

Or:

Thermometer Encoding

Claim:
Neural networks are "overly linear"

Solution

$$T(0.13) = 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$T(0.66) = 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0$$

$$T(0.97) = 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1$$

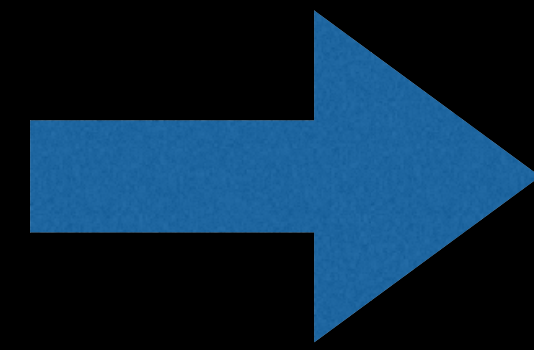
Or:

Input Transformations

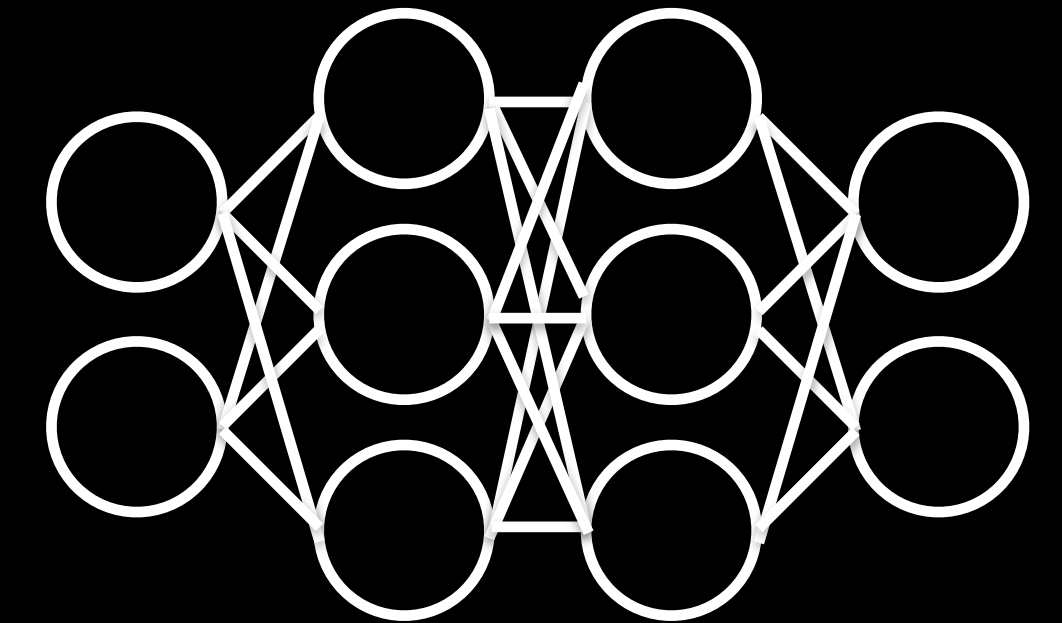
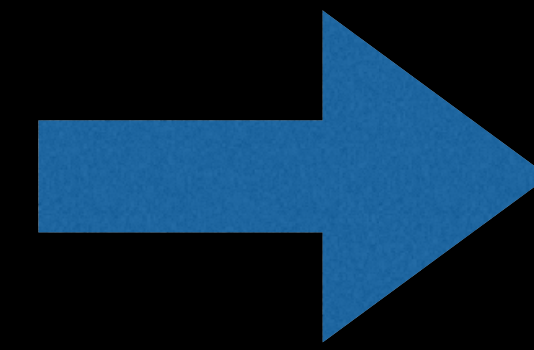
Claim:

Perturbations are brittle

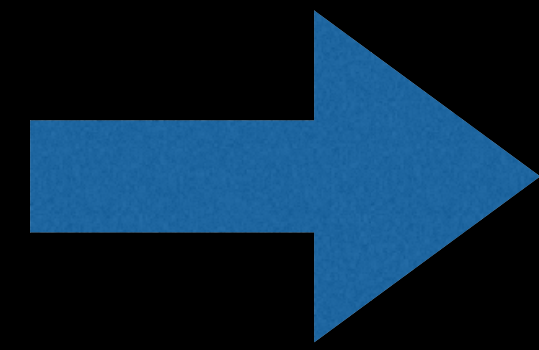
Solution



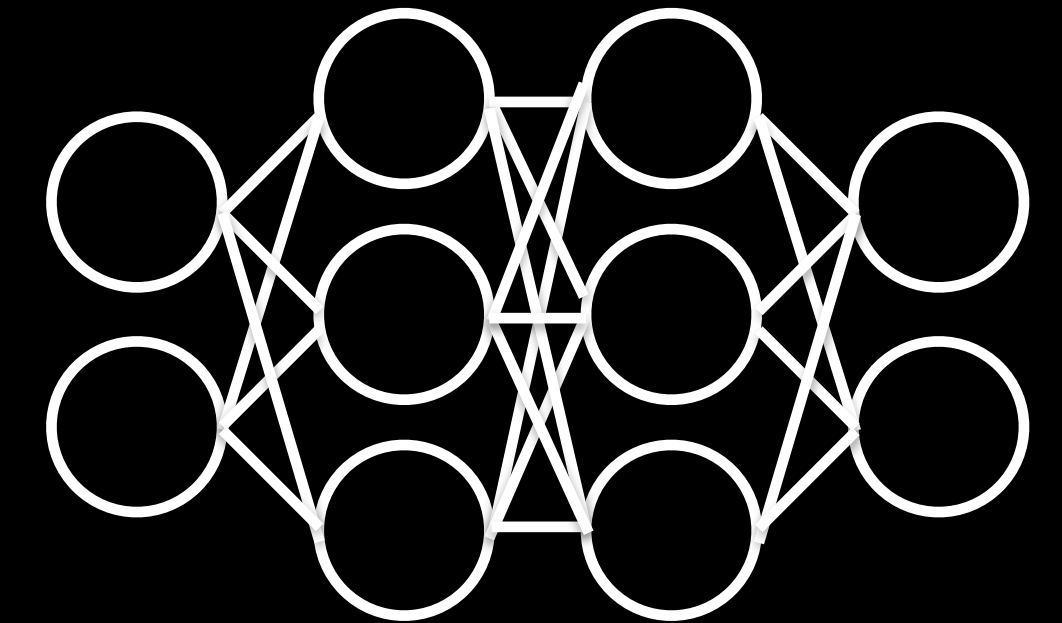
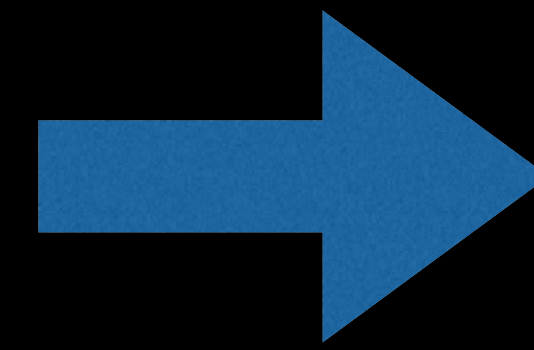
Random
Transform



Solution



JPEG
Compress



Lessons Learned from
Evaluating the Robustness of
Defenses to Adversarial Examples

What does it mean to evaluate
the robustness of a defense?

Standard ML Pipeline

```
model = train_model(x_train, y_train)
acc, loss = model.evaluate(
    x_test, y_test)
if acc > 0.96:
    print("State-of-the-art")
else:
    print("Keep Tuning
          Hyperparameters")
```

Standard ML Pipeline

```
model = train_model(x_train, y_train)
acc, loss = model.evaluate(
    x_test, y_test)
if acc > 0.96:
    print("State-of-the-art")
else:
    print("Keep Tuning
          Hyperparameters")
```


Standard ML Pipeline

```
model = train_model(x_train, y_train)
acc, loss = model.evaluate(
    x_test, y_test)
if acc > 0.96:
    print("State-of-the-art")
else:
    print("Keep Tuning
          Hyperparameters")
```

Standard ML Evaluations

```
model = train_model(x_train, y_train)
acc, loss = model.evaluate(
    x_test, y_test)
if acc > 0.96:
    print("State-of-the-art")
else:
    print("Keep Tuning
          Hyperparameters")
```

Standard ML Evaluations

```
model = train_model(x_train, y_train)
acc, loss = model.evaluate(
    x_test, y_test)
if acc > 0.96:
    print("State-of-the-art")
else:
    print("Keep Tuning
          Hyperparameters")
```

What are robustness evaluations?

Standard ML Evaluations

```
model = train_model(x_train, y_train)
acc, loss = model.evaluate(
    x_test, y_test)
if acc > 0.96:
    print("State-of-the-art")
else:
    print("Keep Tuning
          Hyperparameters")
```

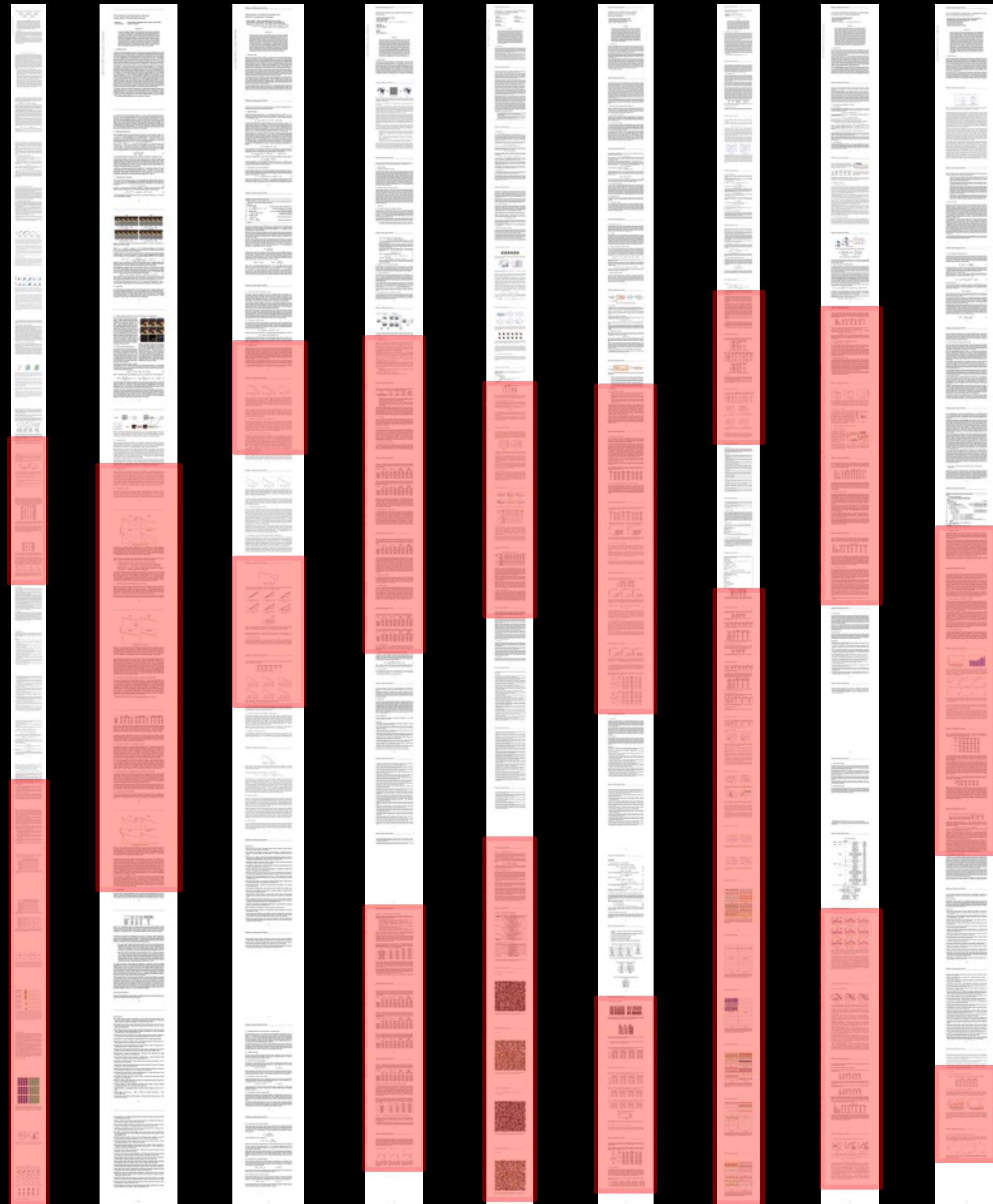
Adversarial ML Evaluations

```
model = train_model(x_train, y_train)
acc, loss = model.evaluate(
    A(x_test, model), y_test)
if acc > 0.96:
    print("State-of-the-art")
else:
    print("Keep Tuning
          Hyperparameters")
```

How complete are evaluations?

Case Study:

ICLR 2018



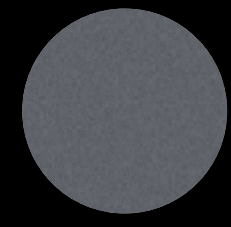
Serious effort
to evaluate

By space, most
papers are $\frac{1}{2}$
evaluation

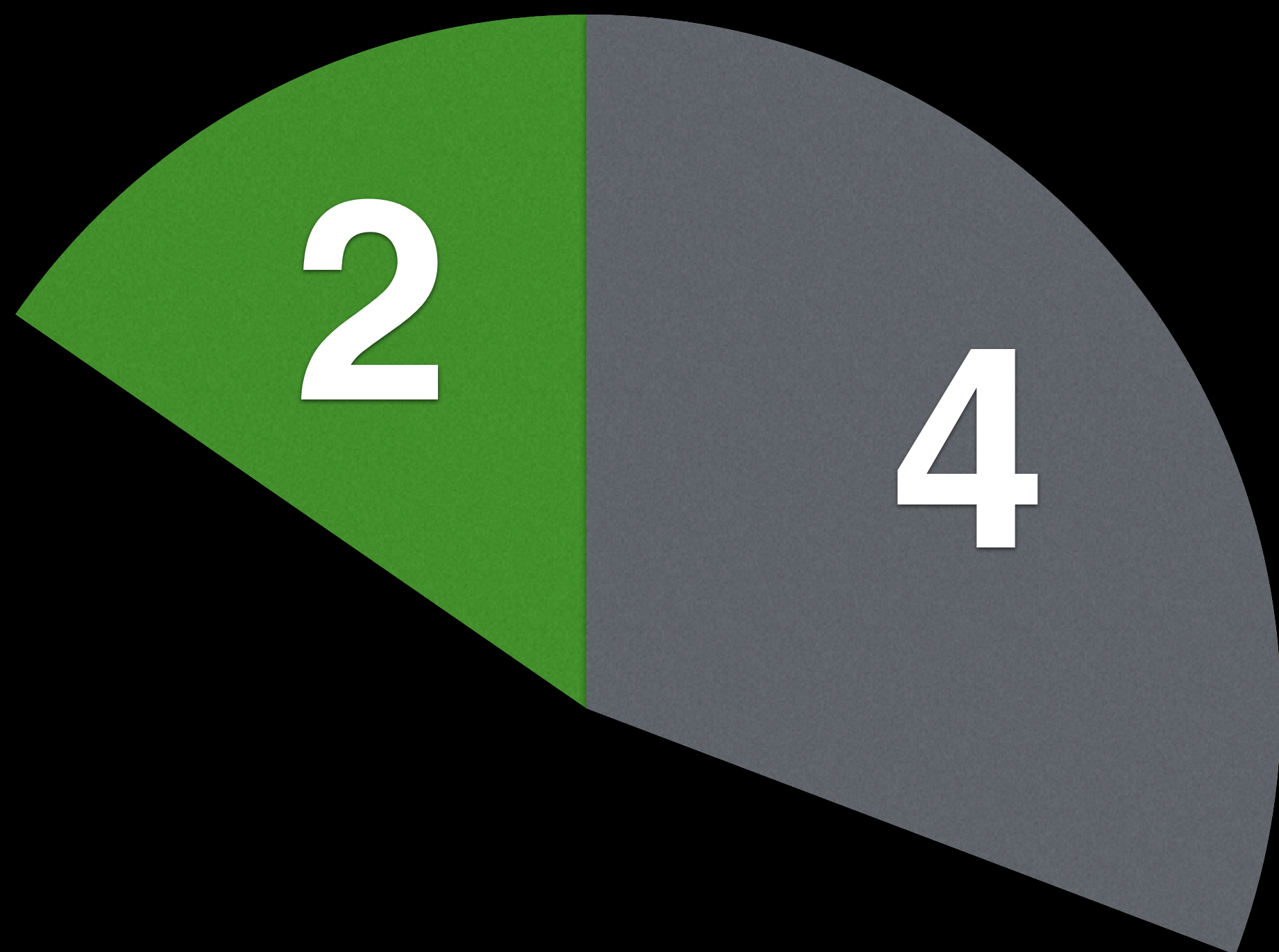
We re-evaluated
these defenses ...



4

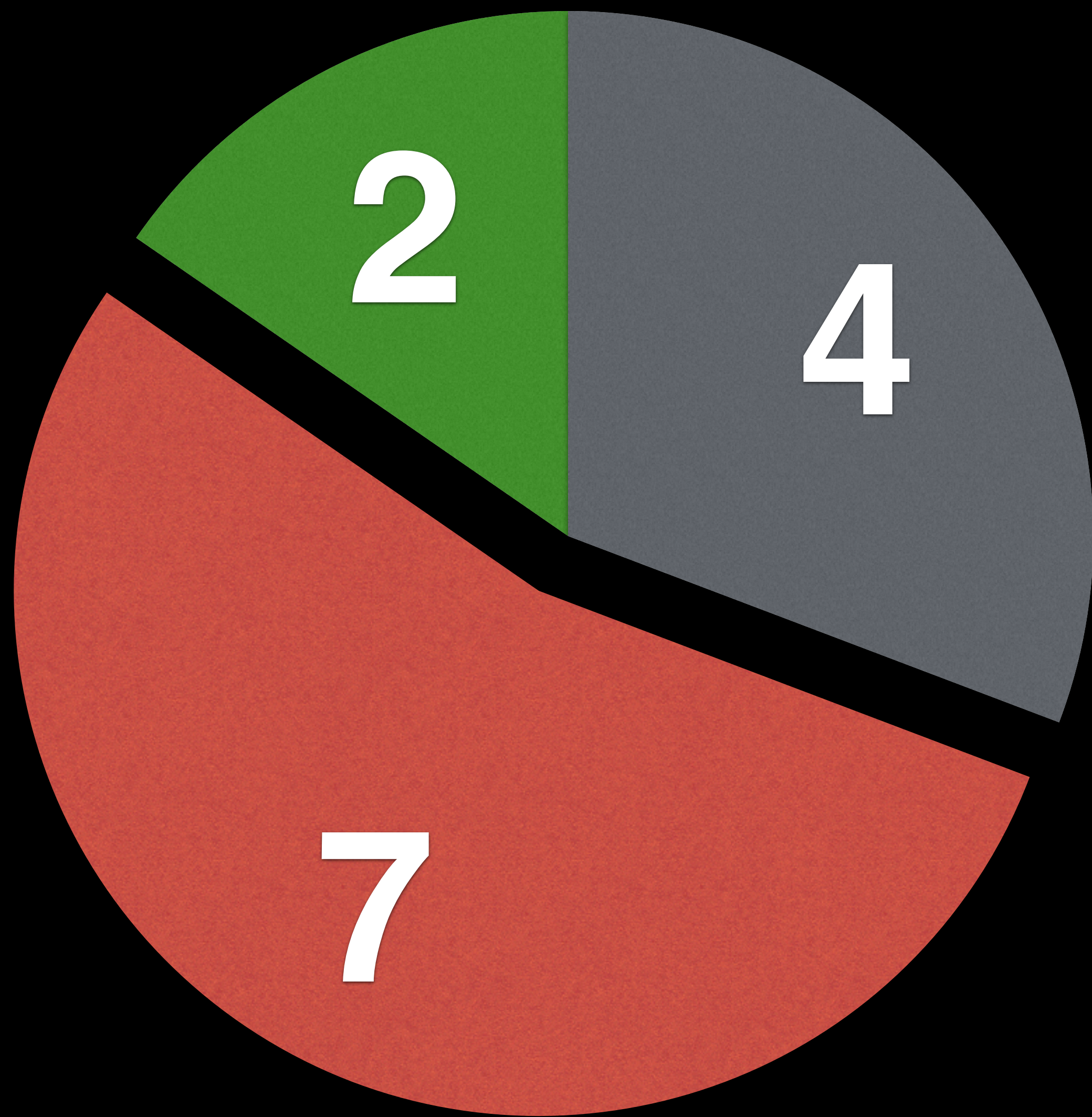


Out of scope



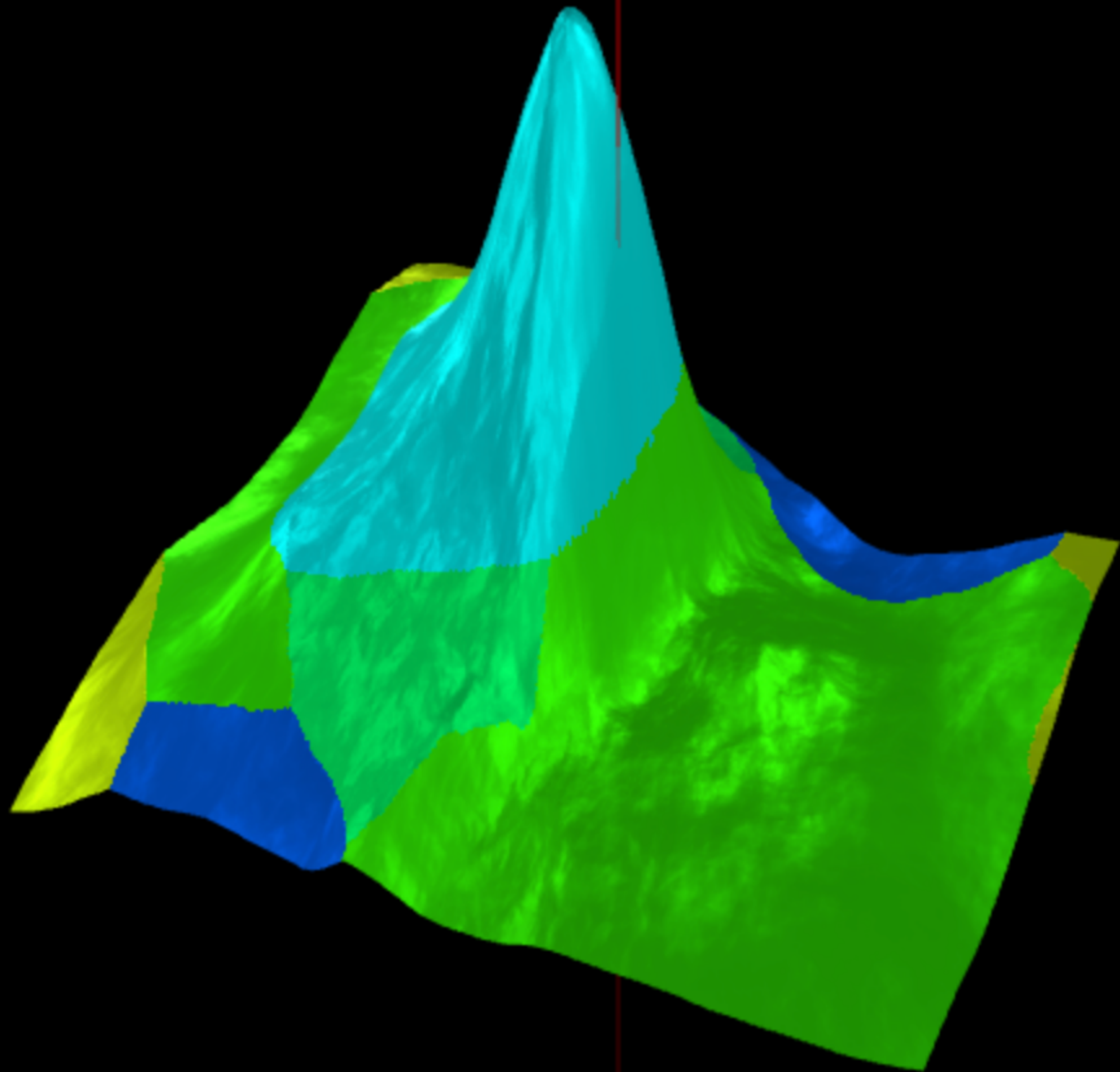
● **Out of scope**

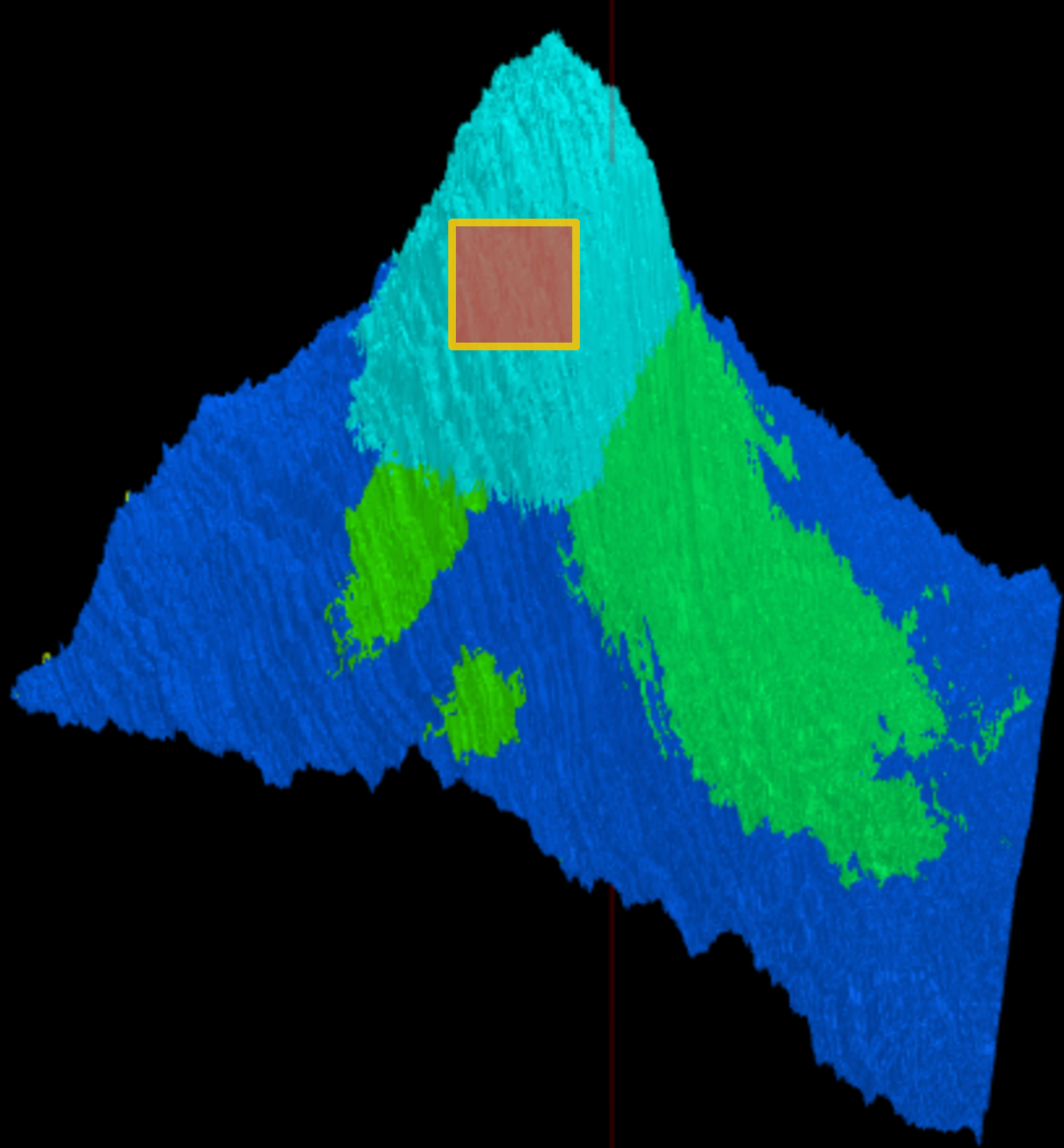
● **Correct Defenses**

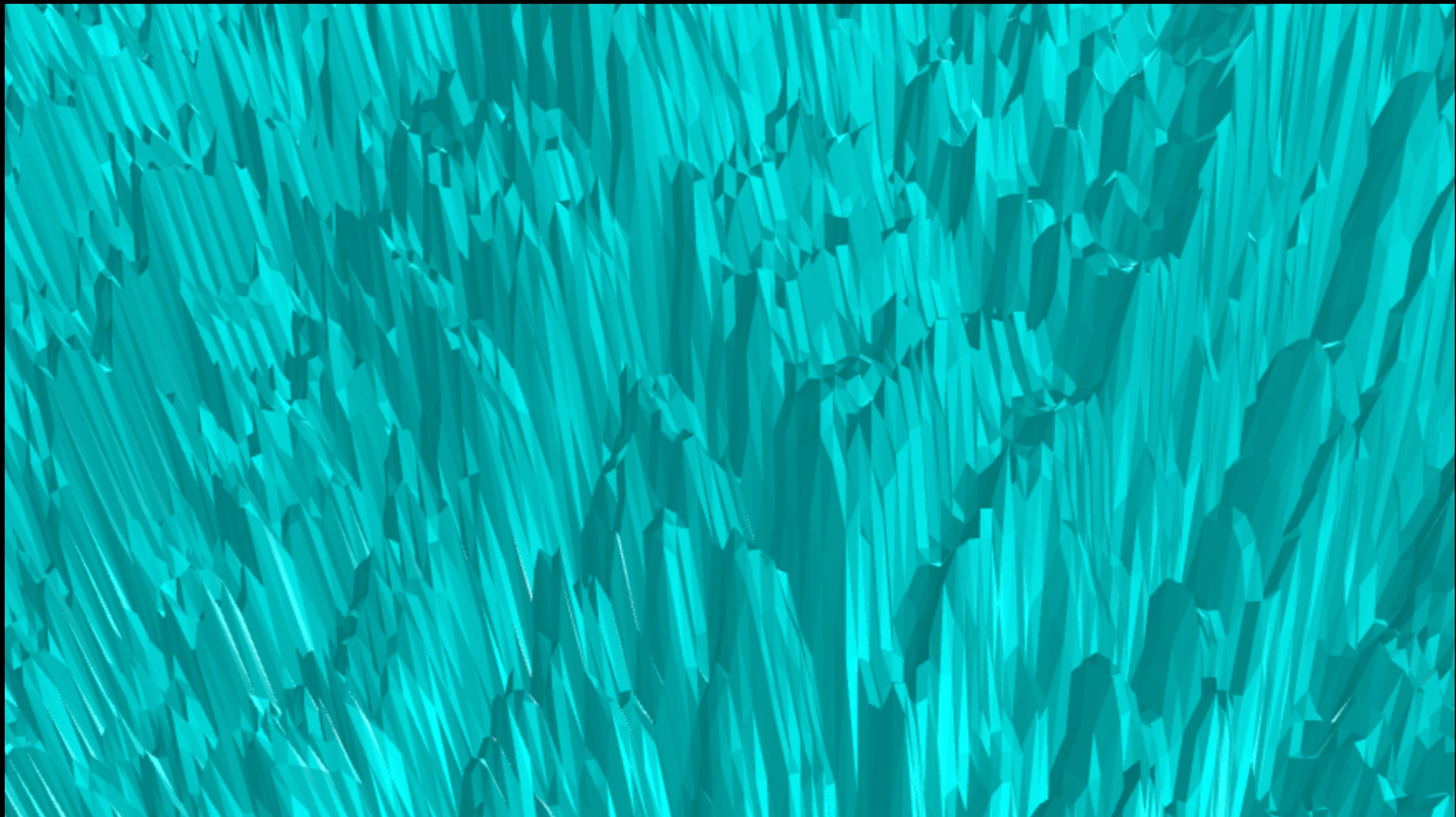


- **Out of scope**
- **Broken Defenses**
- **Correct Defenses**

So what did
defenses do?







Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

MagNet and “Efficient Defenses Against Adversarial Examples” are Not Robust to Adversarial Examples

ABSTRACT

Neural networks: inputs that are adversarial. In order to better survey ten recent papers, we compare their effectiveness against new loss functions. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1 INTRODUCTION

Recent years have seen a surge in research on adversarial examples for neural networks. This driving force has been demonstrated by the fact that adversarial examples have been shown to be effective in a wide range of applications, from image classification to natural language processing [38], to beating cars [6].

In this paper, we investigate the robustness of several defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

The research proposed in this paper is based on the work of [38], [6], and [1].

We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Due to this, we find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Let us compare the effectiveness of these defenses against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Abstract

MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1 Introduction

It is an open question whether we can consistently defend against adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

• MagNet

neural networks

through

to lie on

the data

classification

the whole

MagNet

adversarial

the parameter

• Adversarial

We identify different masking techniques that can be used to generate adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1. Introduction

In response to the fact that adversarial examples are becoming more common, there has been a lot of research on how to defend against them. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

As benchmark

tacks (e.g., Kurakin & Wagner)

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

On the Robustness of the CVPR 2018 Winner

Is AmI (A Robustness Measure) Robust

Neural networks are vulnerable to adversarial examples. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Abstract—No.

I. ATTACKING “ATTACKS MEET INTERPRETABILITY” (AMI) (Attacks meet Interpretability) is an adversarial defense [3] to detect [1] adversarial examples using recognition models. By applying interpretability to a pre-trained neural network, AmI identifies important neurons. It then creates a second augmented network with the same parameters but increases the importance of important neurons. AmI rejects inputs and augmented neural network disagree.

We find that this defense (presented at a spotlight paper—the top 3% of submissions) is ineffective, and even *defense-oblivious*¹ (detection rate to 0% on untargeted attacks). We find that this defense is more robust to untargeted attacks than the vanilla defense. Figure 1 contains examples of attacks that fool the AmI defense. We are incredibly grateful to the authors for releasing their source code² which we used for our experiments. We hope that future work will continue to be published to accelerate progress.

A. Evaluation

A recent gradient-based defense, highly sensitive to adversarial attacks, has yet to be evaluated against gradient-based attacks. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Evaluation between neural networks is robust because of gradient-based attacks. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

A recent defense against gradient-based attacks is likely a side effect of the gradient-based attack. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

In a first Perceptron layer activation, sigmoid, zero network for the MLP with attack with verified the S1.

In high exactly zero directly with elements of

Training
Vanilla
Saturated

Table 1: A naive application of FGSM based

¹Werner

Comment on *Biologically inspired protection of deep networks from adversarial attacks*

ON THE LIMITATION OF LOCAL INTRINSIC DIMENSIONALITY FOR CHARACTERIZING THE SUBSPACES OF

A

P
N
T

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

The Efficacy of SHIELD under Different Threat Models

Paper Type: Appraisal Paper of Existing Method

Cory Cornelius
cory.cornelius@intel.com

Nilaksh Das
nilakshdas@gatech.edu

Shang-Tse Chen
schen351@gatech.edu

This paper motivates the need to move beyond the current state of the art in adversarial risk evaluation. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

ABSTRACT

In this appraisal of adversarial risk, we study the efficacy of SHIELD at KDD 2017. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

adversary is pre-processed used in the threat and extent of degree of in full white-box original work. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1

In response to the fact that adversarial examples are becoming more common, there has been a lot of research on how to defend against them. We find that many of the defenses significantly harm the properties believed to be necessary for a defense to be effective. In fact, not all defenses are robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Training
Vanilla
Saturated

Table 1: A naive application of FGSM based

A. Evaluation

Evaluating and Understanding the Robustness of Adversarial Logit Pairing

Logan Engstrom* Andrew Ilyas* Anish Athalye*
Massachusetts Institute of Technology
{engstrom, ailyas, aathalye}@mit.edu

Abstract

We evaluate the robustness of Adversarial Logit Pairing, a recently proposed defense against adversarial examples. We find that a network trained with Adversarial Logit Pairing achieves 0.6% correct classification rate under targeted adversarial attack, the threat model in which the defense is considered. We provide a brief overview of the defense and the threat models/claims considered, as well as a discussion of the methodology and results of our attack. Our results offer insights into the reasons underlying the vulnerability of ALP to adversarial attack, and are of general interest in evaluating and understanding adversarial defenses.

1 Contributions

For summary, the contributions of this note are as follows:

- Robustness:** Under the white-box targeted attack threat model specified in Kannan et al., we upper bound the correct classification rate of the defense to **0.6%** (Table 1). We also perform targeted and untargeted attacks and show that the attacker can reach success rates of 98.6% and 99.9% respectively (Figures 1, 2).

ACM Refere

Permission to make digital or hard copies of this work for personal or classroom use is granted by ACM, provided that the fee of \$15.00 is paid directly to ACM. For more information, contact the ACM Permissions Department, 2 Penn Plaza, New York, NY 10019-6098, USA. Copyright © 2017 ACM. ISBN 978-1-4503-5111-1. https://doi.org/10.1145/3122448

Lessons Learned from
Evaluating the Robustness of
Defenses to Adversarial Examples

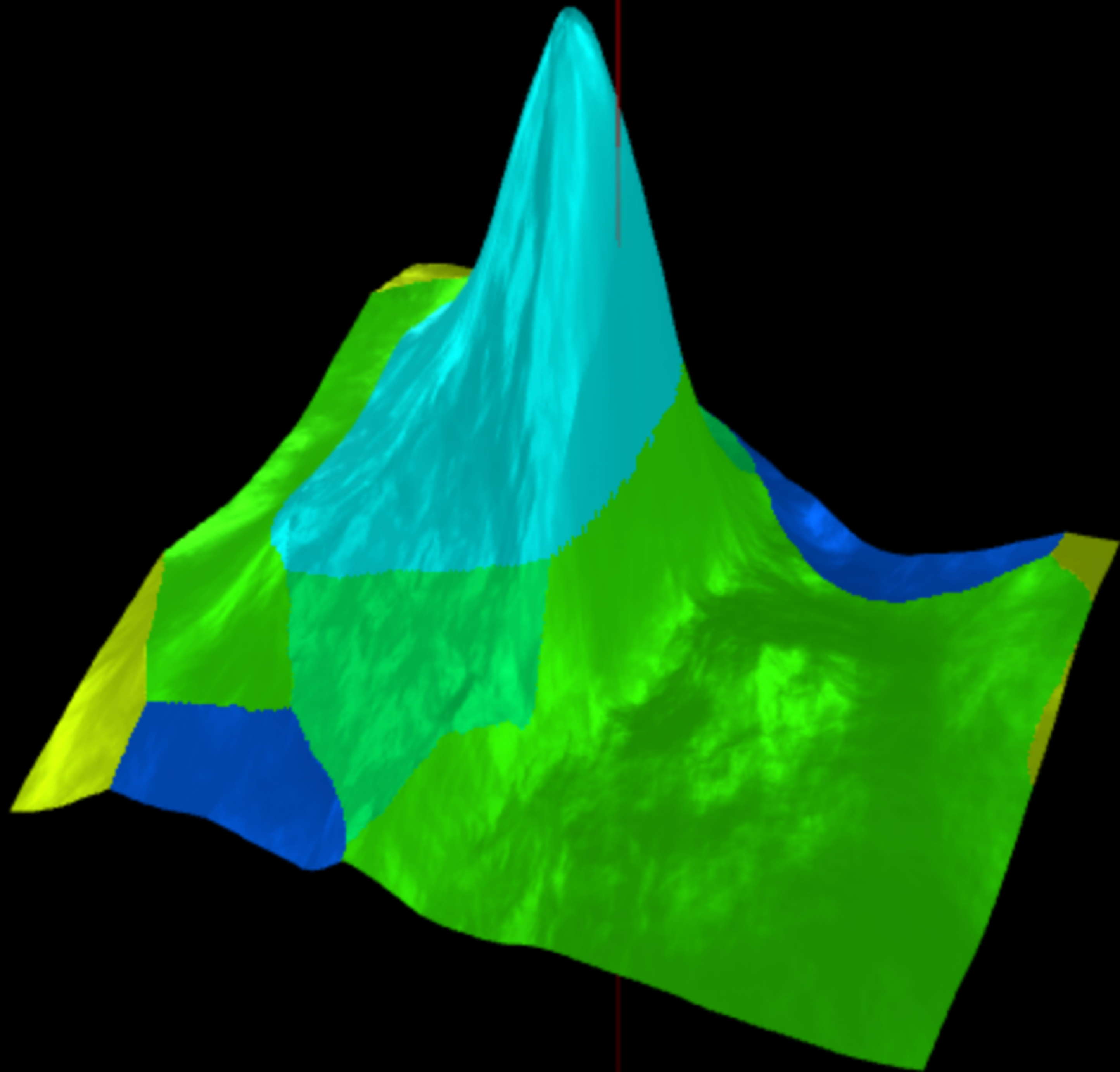
Lessons (1 of 3)

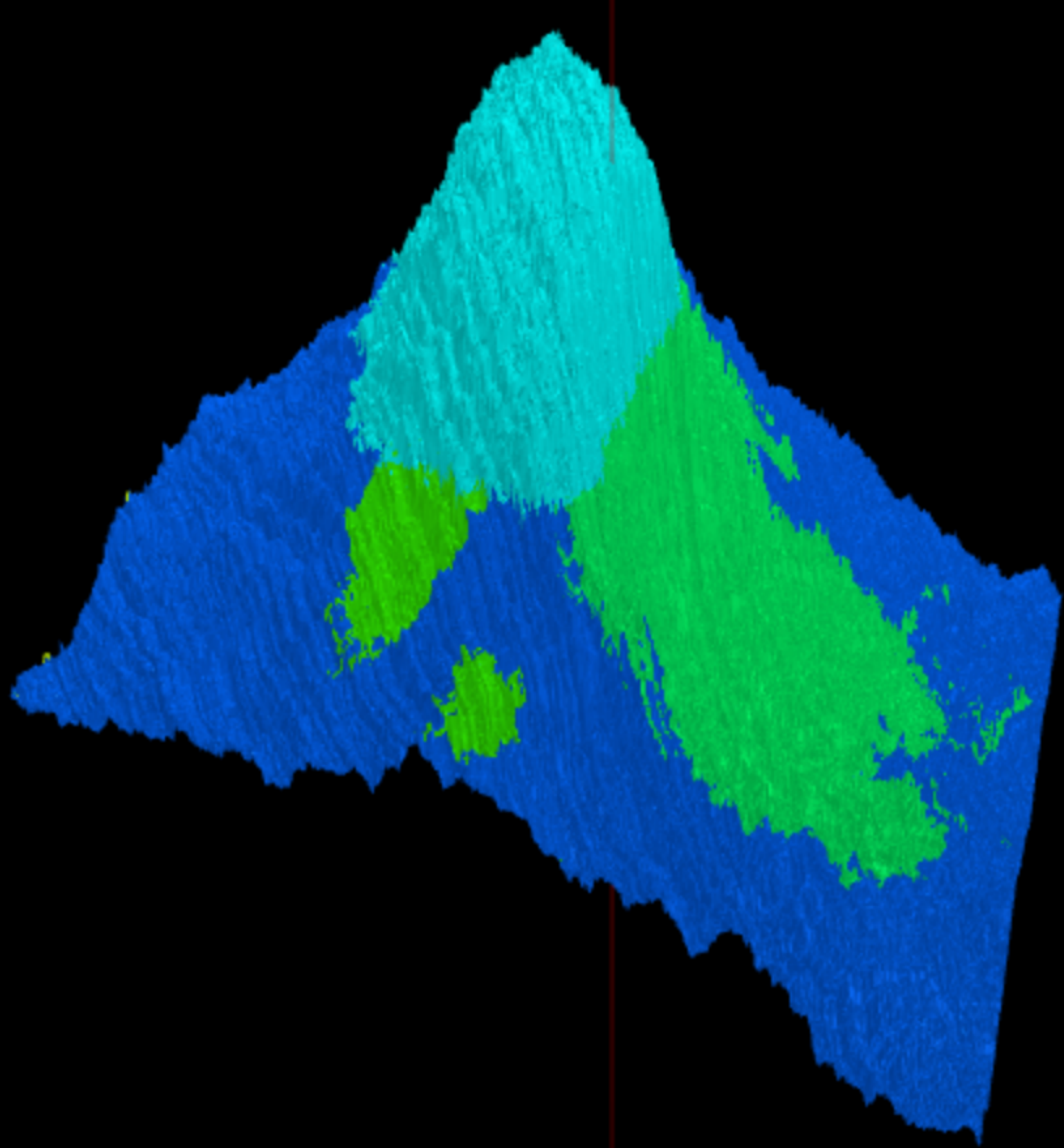
what types of defenses are effective

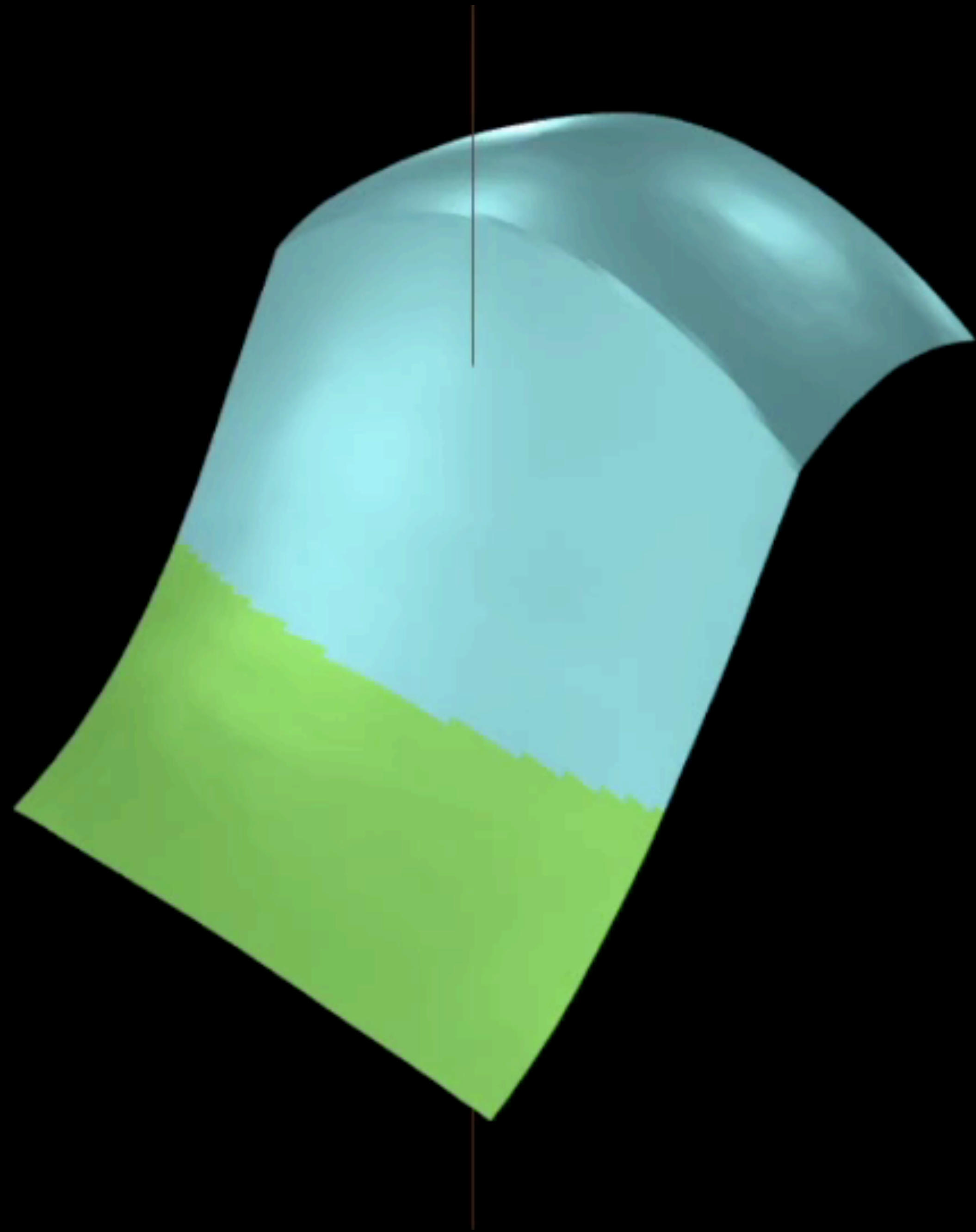
First class of effective defenses:

First class of effective defenses:

Adversarial Training





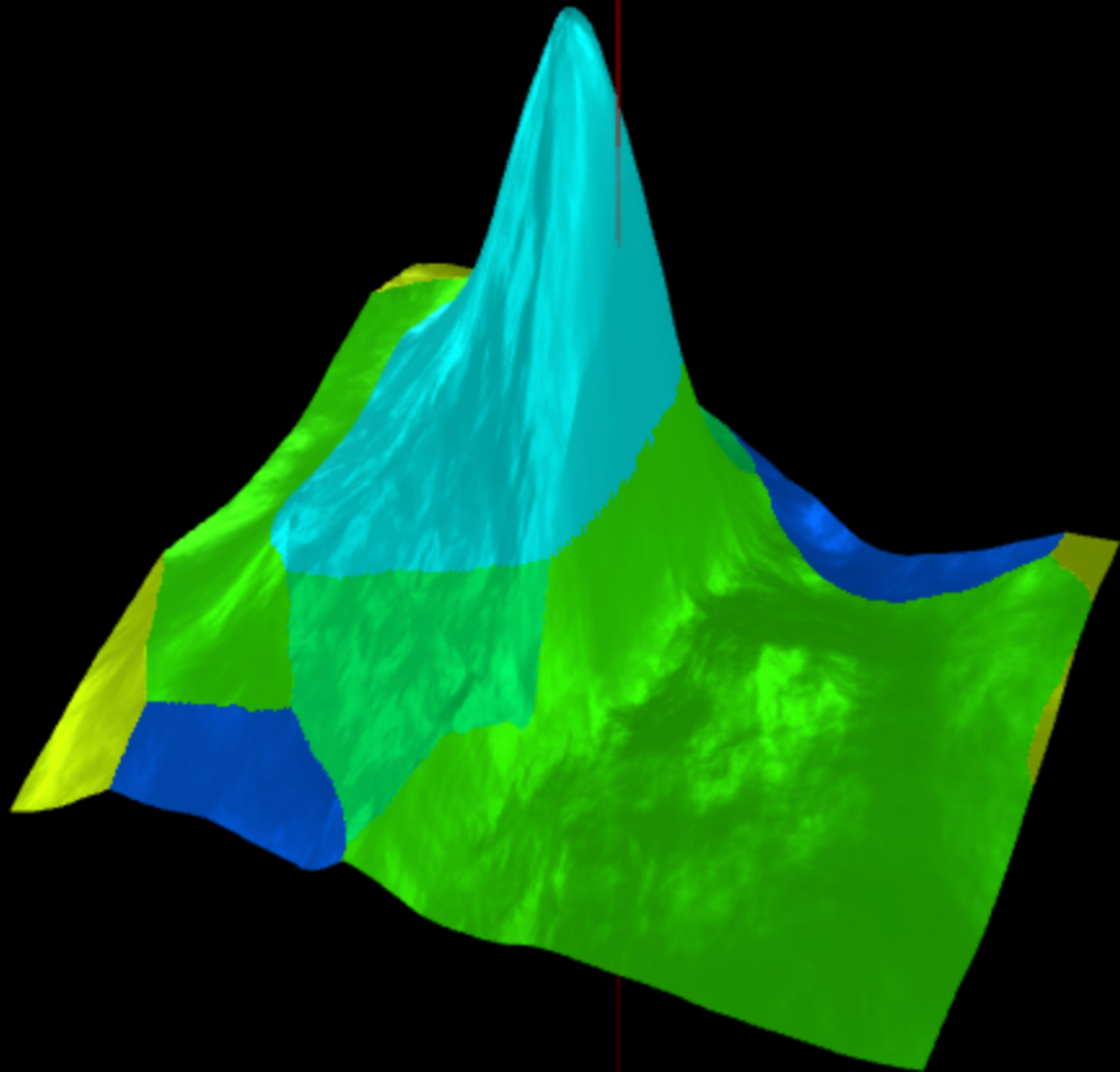


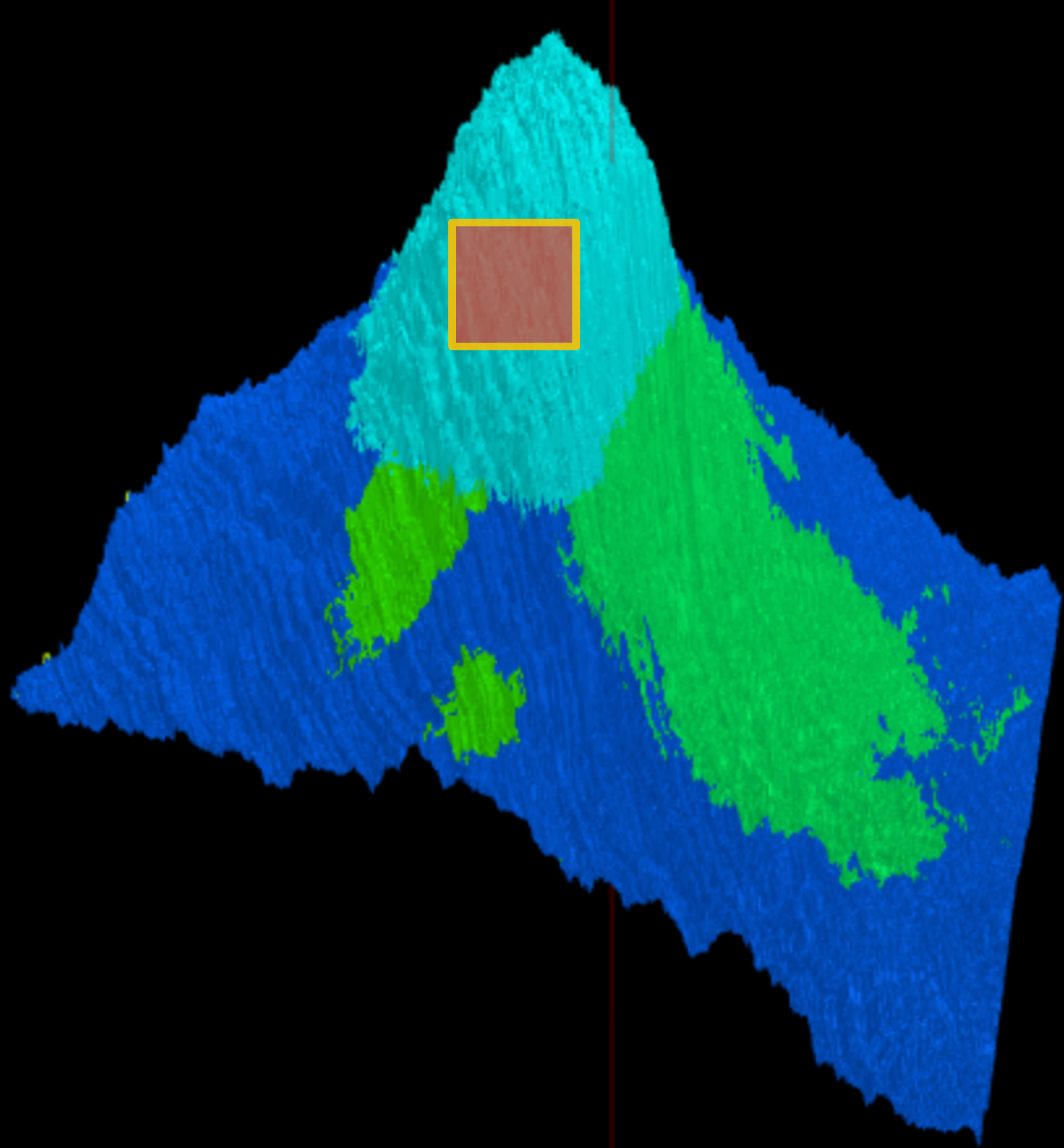
Second class of effective defenses:

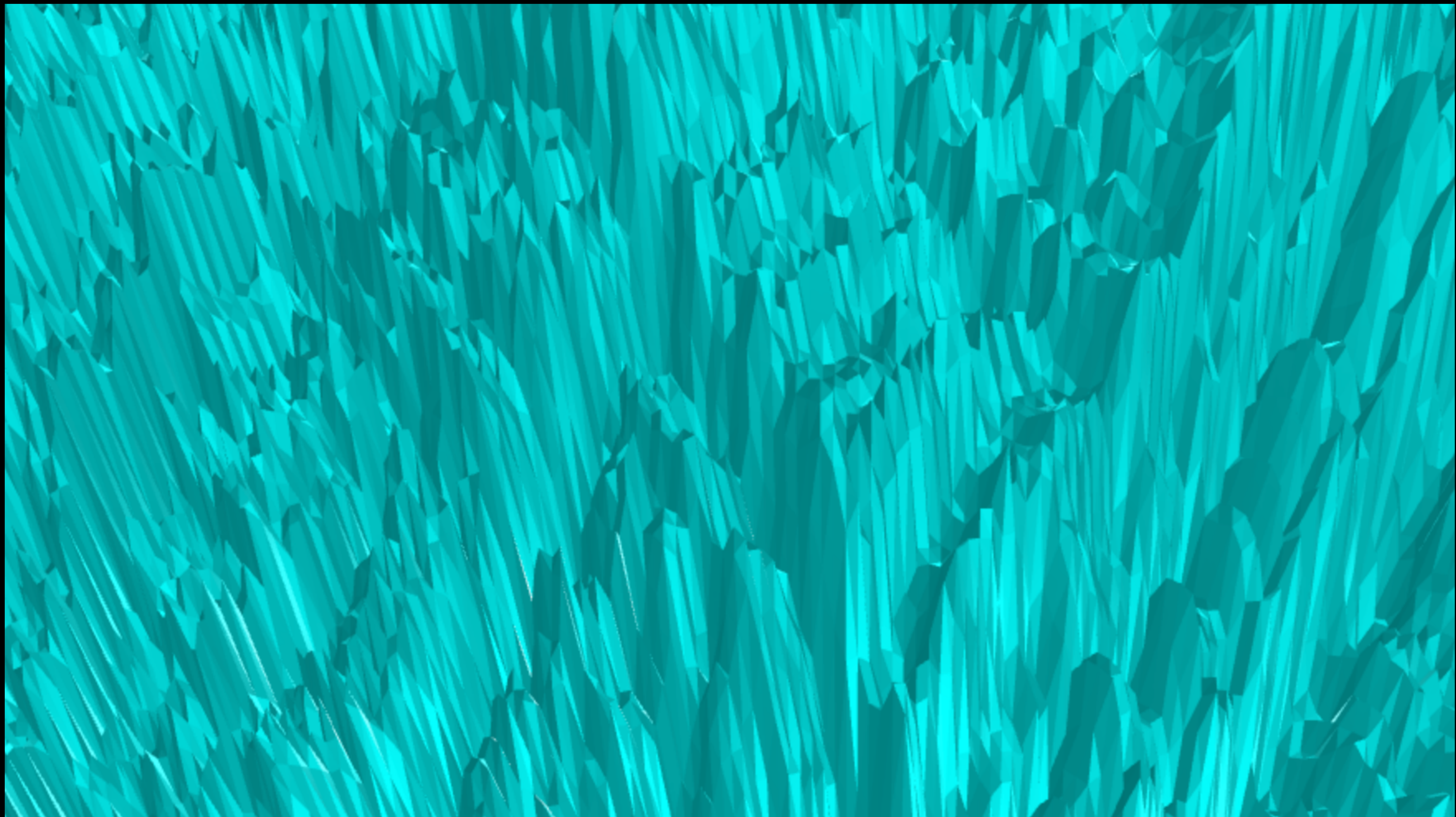
Second class of effective defenses:

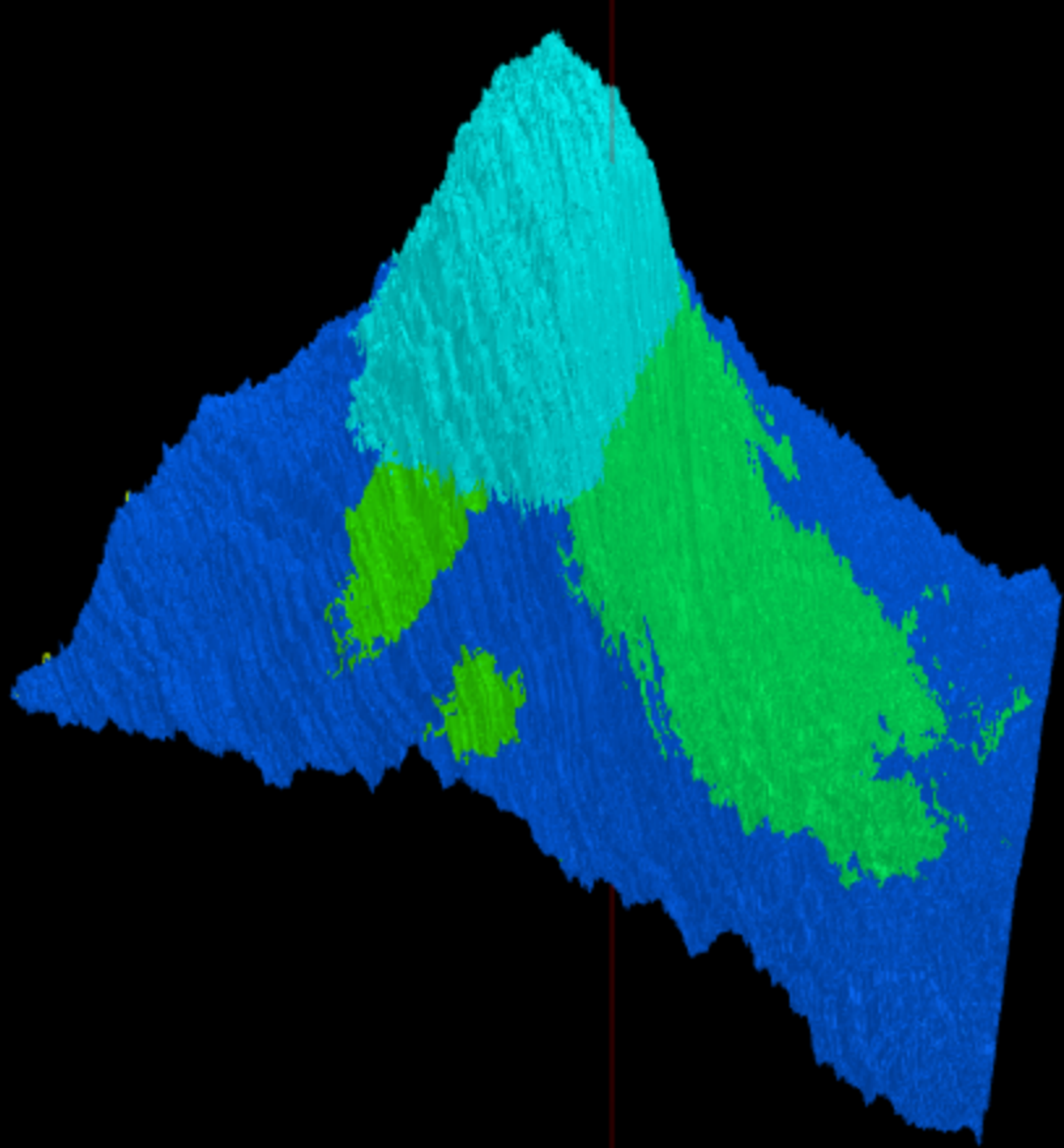
Lessons (2 of 3)

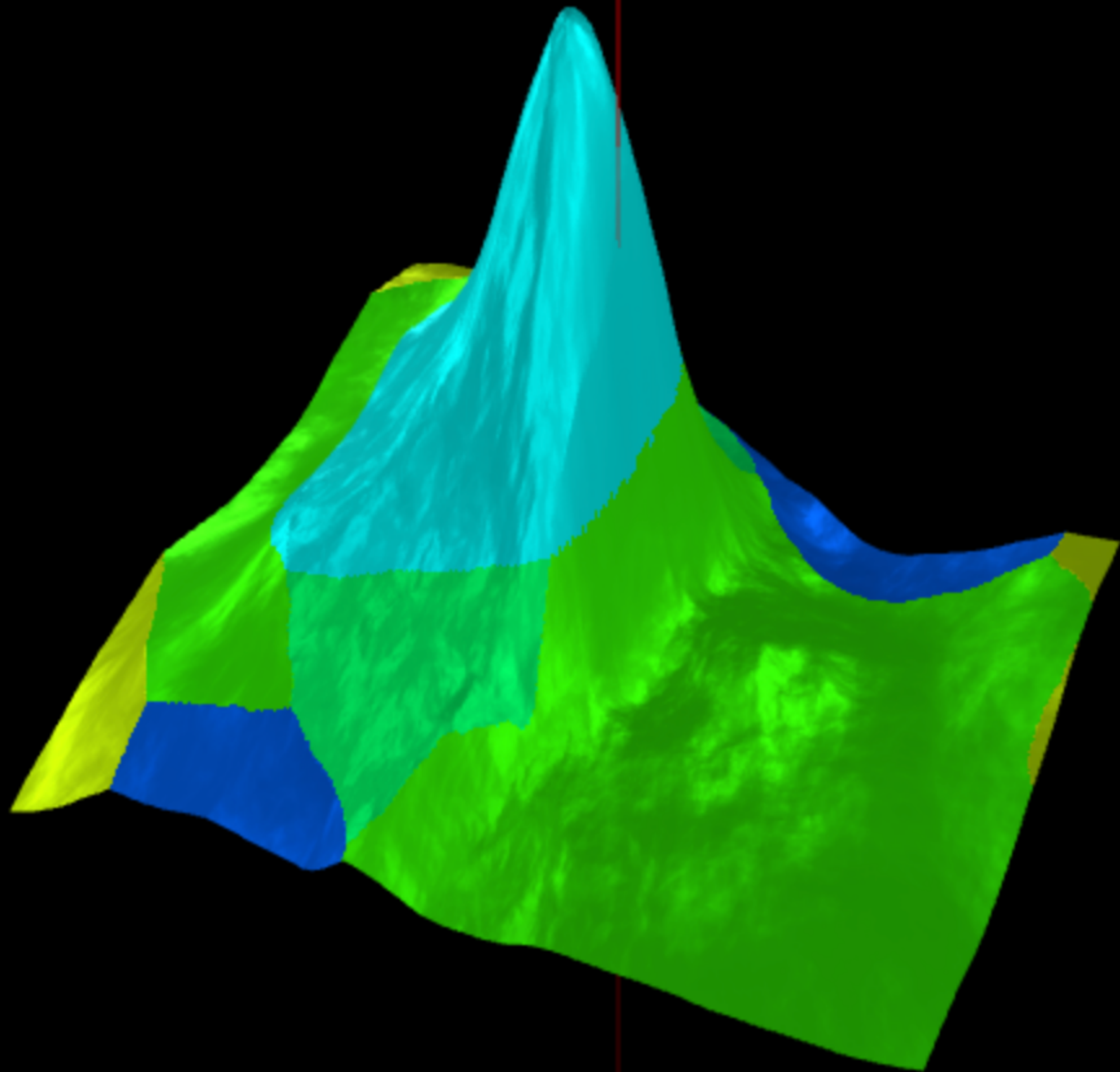
what we've learned from evaluations











So how to attack it?

JPEG-resistant Adversarial Images

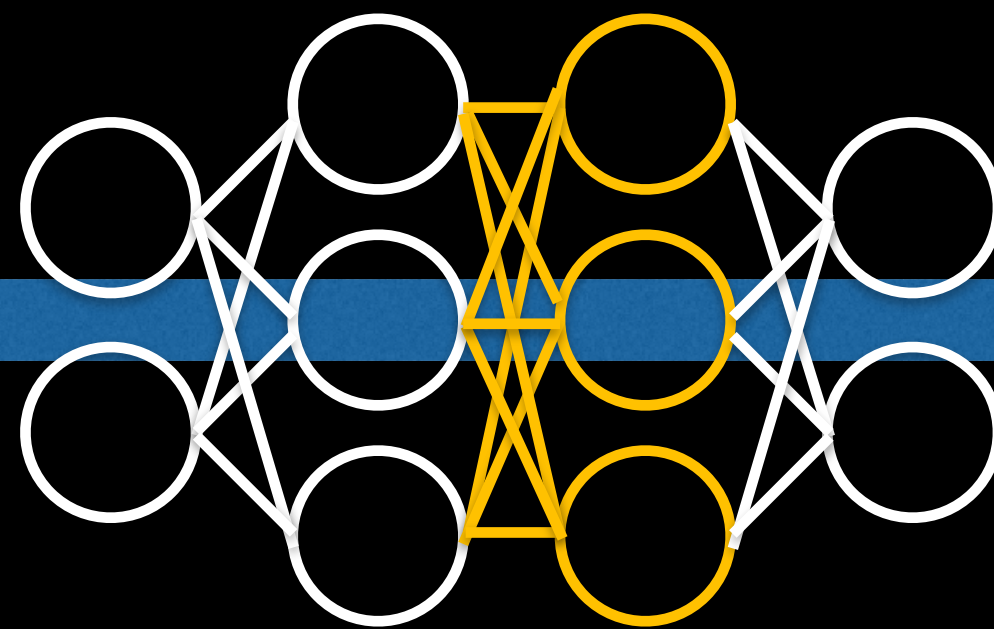
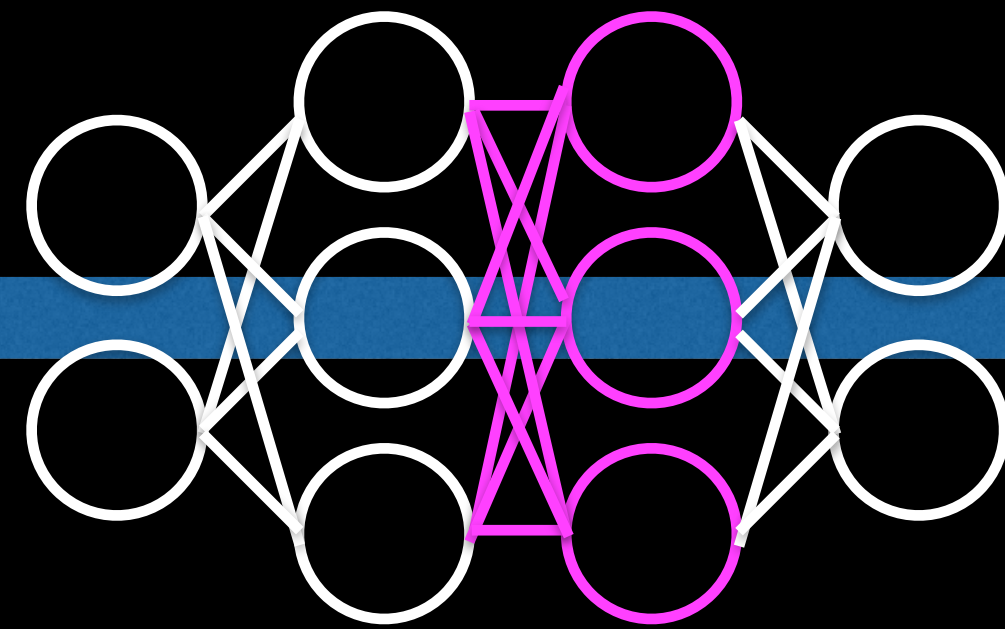
Richard Shin

Computer Science Division
University of California, Berkeley
ricshin@cs.berkeley.edu

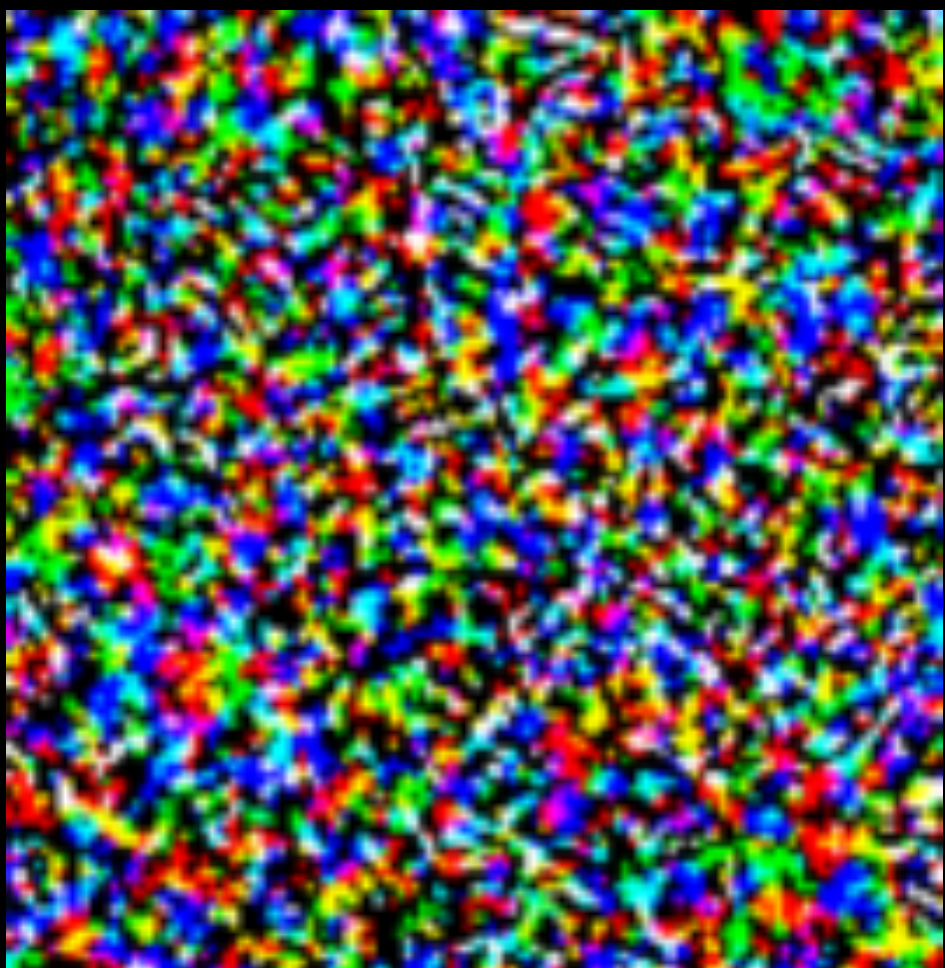
Dawn Song

Computer Science Division
University of California, Berkeley
dawnsong@cs.berkeley.edu

"Fixing" Gradient Descent



**[0.1,
0.3,
0.0,
0.2,
0.4]**



Lessons (3 of 3)

performing better evaluations

ON EVALUATING ADVERSARIAL ROBUSTNESS

Nicholas Carlini¹, Anish Athalye², Nicolas Papernot¹, Wieland Brendel³, Jonas Rauber³,
Dimitris Tsipras², Ian Goodfellow¹, Aleksander Mądry², Alexey Kurakin^{1*}

¹ Google Brain ² MIT ³ University of Tübingen



Actionable advice
requires specific,
concrete examples

Everything the
following papers do
is standard practice

the adversary has access to those networks (but does not have access to the input transformations applied at test time).

²The white-box attacks defined in this paper should be called oblivious attacks according to Carlini and Wagner's definition [3]

an adversary gains access to all parameters and weights of a model that is trained on benign images, but is unaware of the defense strategy.

Perform an
adaptive attack

3.1. Effectiveness

3.1. Effectiveness

Adversarial Attacks. We test on the following attacks:

we trained on and L_{CW} is an objective encouraging misclassification. Under this threat model, *NeuralFP* achieves an AUC-ROC of **98.79%** against Adaptive-CW- L_2 , with $N = 30$ and $\epsilon = 0.006$ for a set of unseen test-samples (1024 *pre-test*) and the corresponding adversarial examples. In contrast to other defenses that are vulnerable to Adaptive-CW- L_2 (Carlini & Wagner, 2017a), we find that *NeuralFP* is robust even under this whitebox-attack threat model.

4. Related Work

3.4. Robustness to Adaptive Whitebox-Attackers

We further considered an adaptive attacker that has knowledge of the predetermined fingerprints and model weights, similar to (Carlini & Wagner, 2017a). Here, the adaptive attacker (Adaptive-CW- L_2) tries to find an adversarial example x' that also minimizes the fingerprint-loss, attacking a CIFAR-10 model trained with *NeuralFP*. To this end, the CW- L_2 objective is modified as:

$$\min_{x'} \|x - x'\|_2 + \gamma (L_{CW}(x') + L_{fp}(x', y^*, \xi; \theta)) \quad (29)$$

Here, y^* is the label-vector, $\gamma \in [10^{-3}, 10^0]$ is a scalar found through a bisection search, L_{fp} is the fingerprint-loss

5. Discussion and Future Work

3.4. Robustness to Adaptive Whitebox-Attackers

Adversarial Attacks. We test on the following attacks:

3.1. Effective

4. Related Work

3.4. Robustness to Adaptive Whitebox-Attackers

5. Discussion and Future Work

We now evaluate on two held out L_0 attacks

A "hold out" set is
not an adaptive attack

To create adversarial examples in our evaluation, we use FGSM,

For the next series of experiments, we test against the *Fast Gradient Sign Method*

In our experiment, we use the Fast Gradient Sign Method (FGSM)

TABLE 4: Performance of detecting FGSM adversarial examples with different scalar quantization schemes.

Stop using FGSM
(exclusively)


- Number of attack steps: 10

experiments on CIFAR used

$\varepsilon = 0.031$ and 7 steps for iterative attacks;

Use more than 100
(or 1000?) iteration of
gradient descent

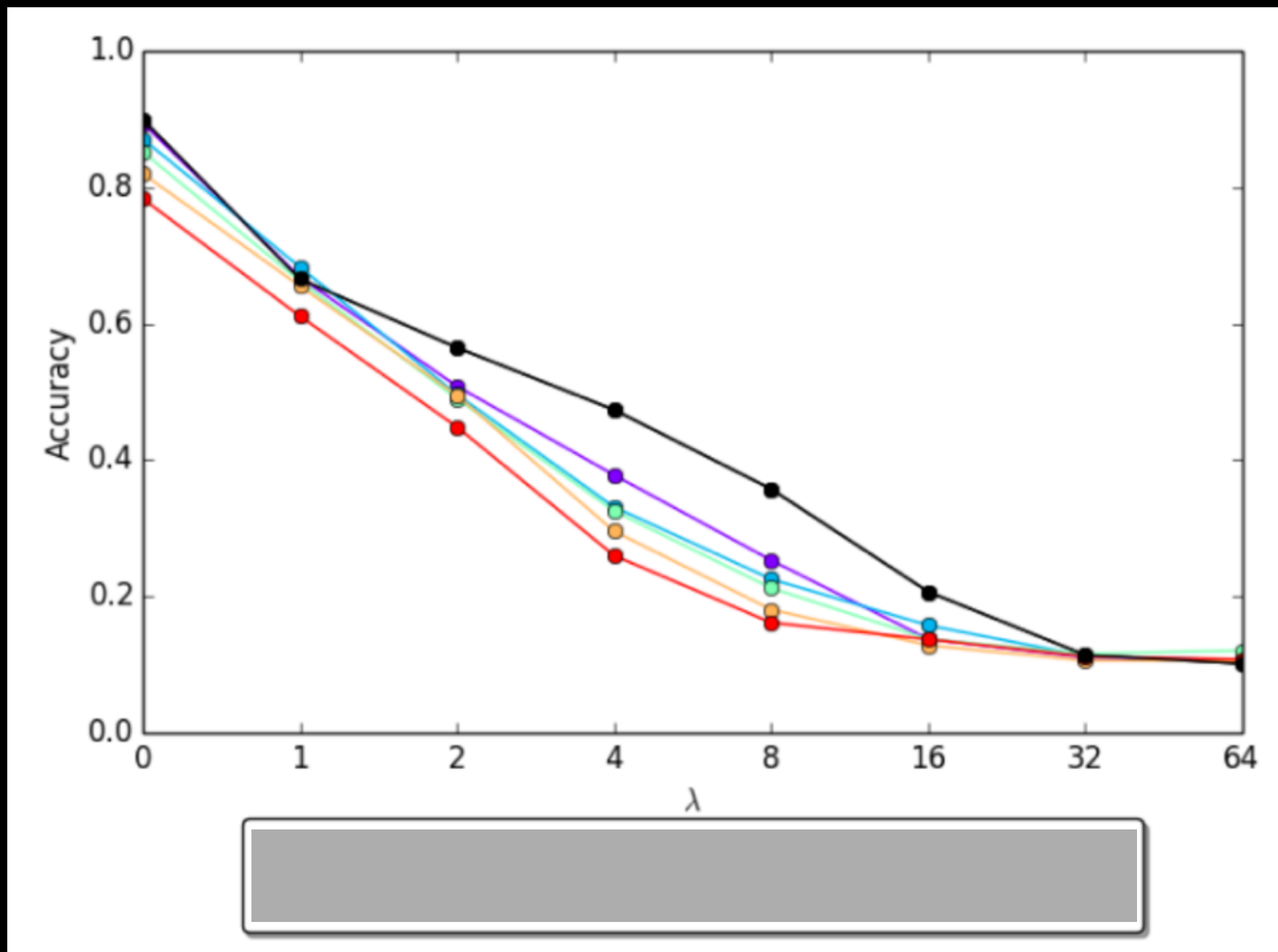
	Model	FGSM	PGD
<i>Clean</i>		25.10	4.10
		46.15	1.66
		43.89	3.57
		52.07	53.11
		48.50	50.50



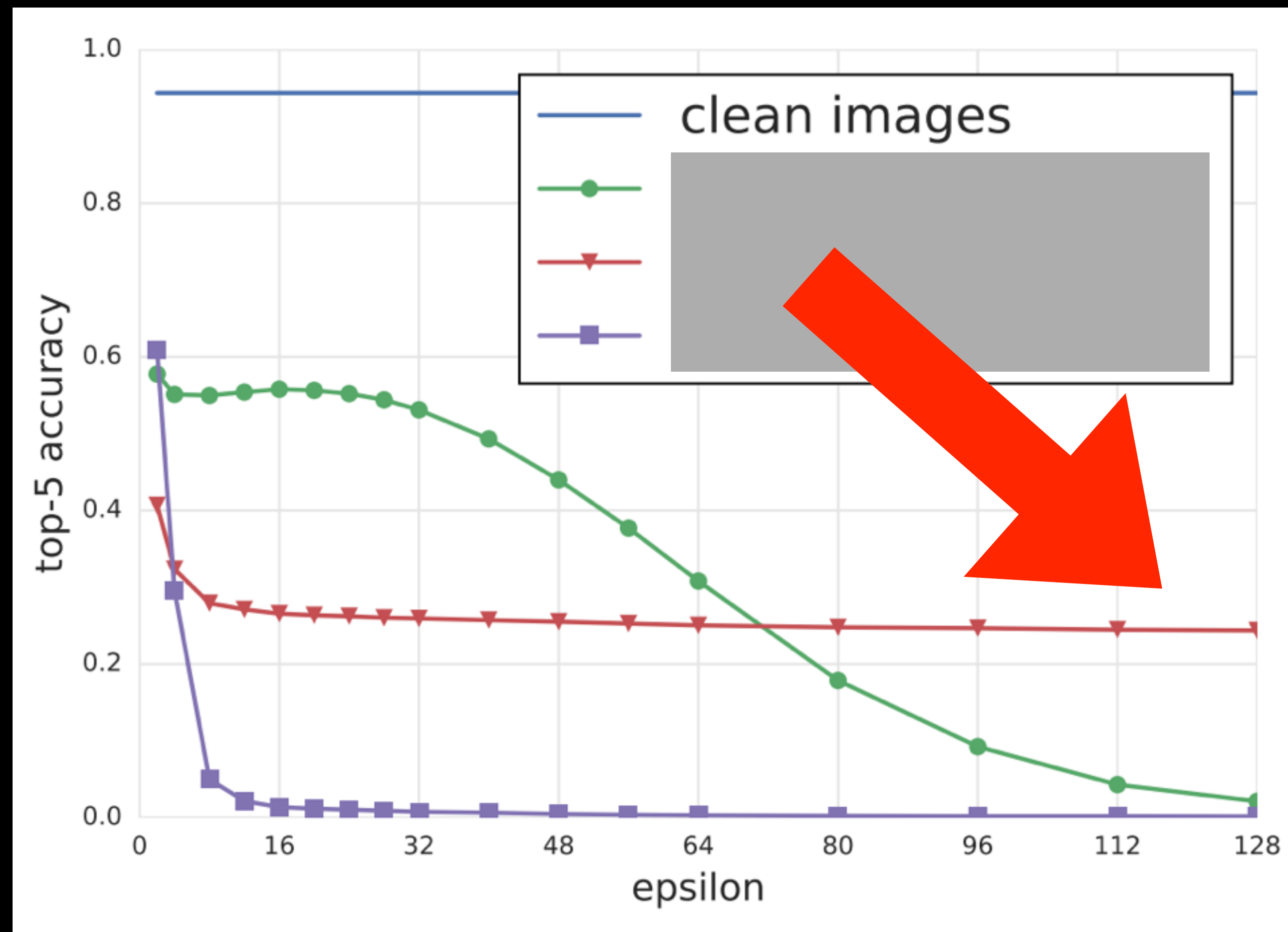
Iterative attacks should always do better than single step attacks.

Attack	Parameter	Fooling Rate	Detection Rate
DeepFool		99.35%	97.83%
Carlini	$\kappa=0.0$	100.0%	95.66%

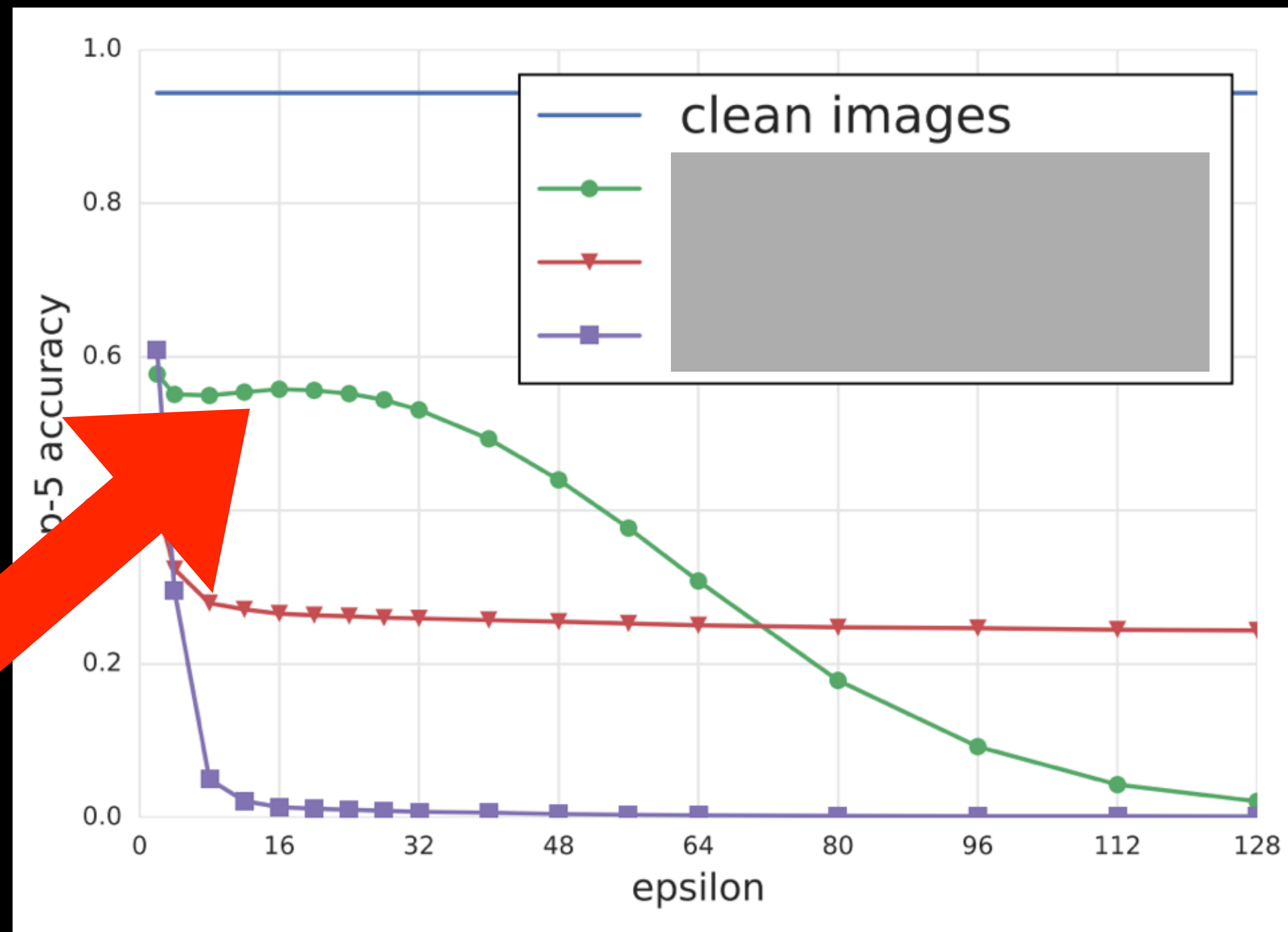
Unbounded optimization attacks should eventually reach in 0% accuracy



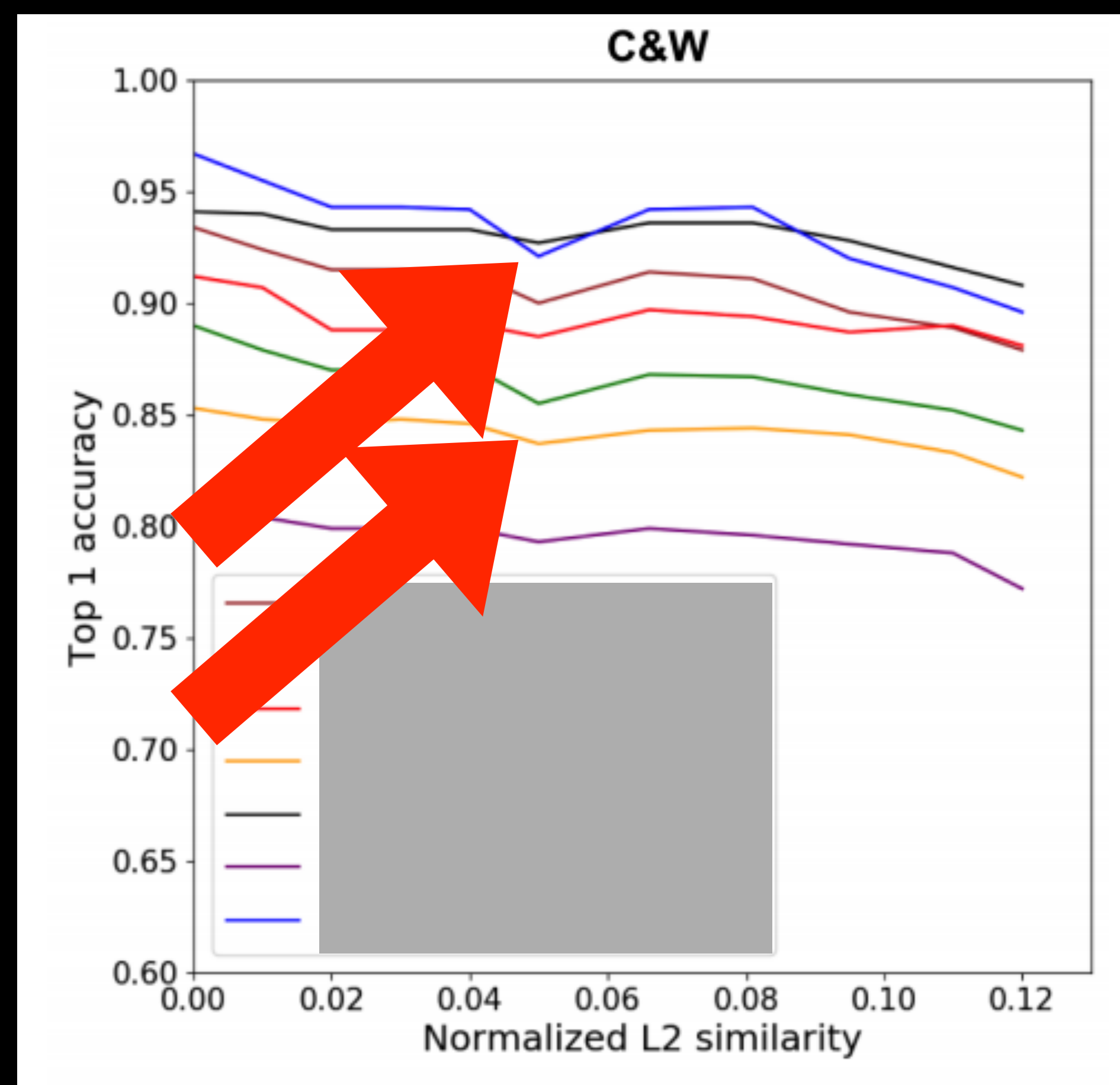
Unbounded optimization attacks should eventually reach in 0% accuracy



Unbounded optimization attacks should eventually reach in 0% accuracy



Model accuracy should be monotonically decreasing

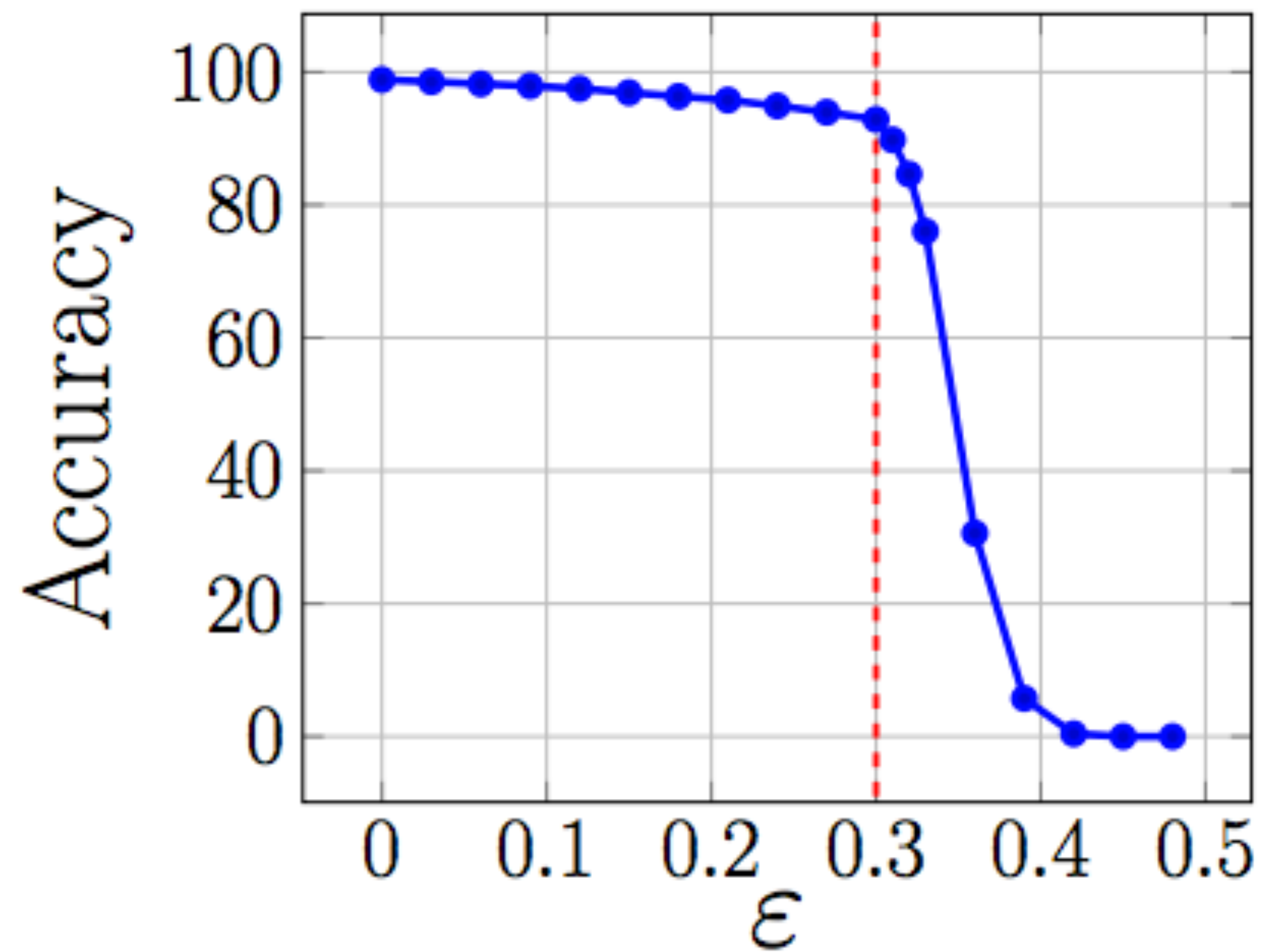


Model accuracy should be monotonically decreasing



Model	clean	step_ll		step_FGSM		iter_FGSM		CW	
		$\epsilon=2$	$\epsilon=16$	$\epsilon=2$	$\epsilon=16$	$\epsilon=2$	$\epsilon=4$	$\epsilon=2$	$\epsilon=4$
R110 _K	92.3	88.3	90.7	86.0	95.2	59.4	9.2	25	4
R110 _P (Ours)	92.3	86.0	89.4	81.6	91.6	64.1	20.9	32	7
R110 _E	92.3	86.3	74.3	84.1	72.9	63.5	21.1	24	6
R110 _{K,C} (Ours)	92.3	86.2	72.8	82.6	66.7	69.3	33.4	20	5
R110 _{P,E} (Ours)	91.3	84.0	65.7	77.6	54.5	66.8	38.3	38	16
R110 _{P,C} (Ours)	91.5	85.7	76.4	82.4	69.1	73.5	42.5	27	15

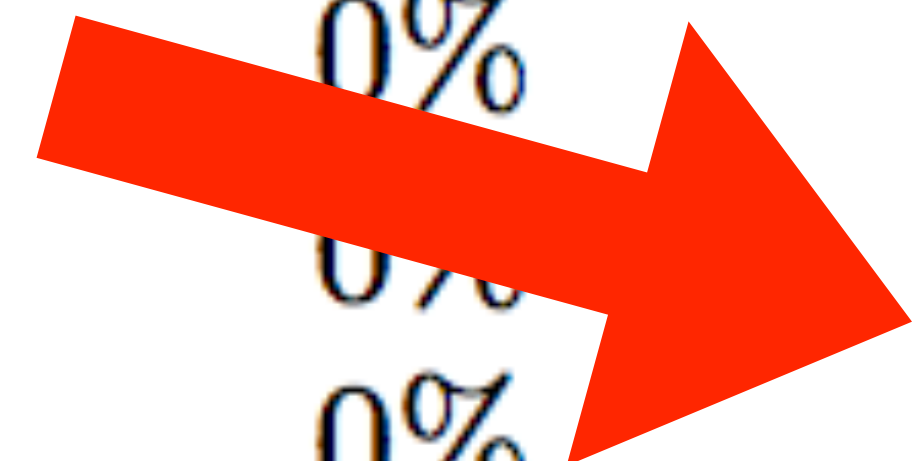
Evaluate against the
worst attack



(a) MNIST, ℓ_∞ norm

Plot accuracy vs distortion

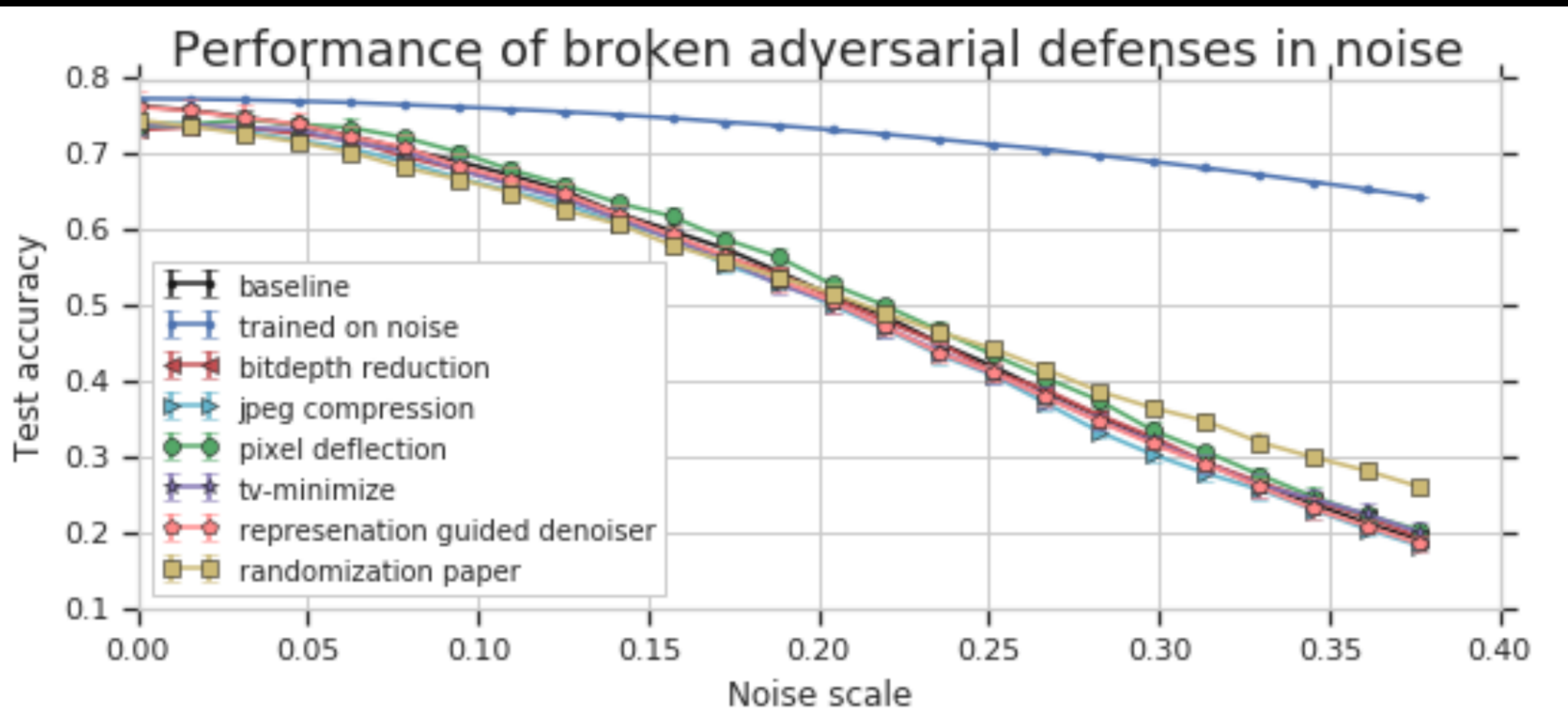
MaxIter	Model1	Model2	Model3	Model4
Natural	99.1%	98.5%	98.7%	98.2%
100	70.2%	91.7%	77.6%	75.6%
1000	0.05%	51.5%	20.3%	24.4%
10K	0%	16.0%	20.1%	24.4%
100K	0%	9.8%	20.1%	24.4%
1M	0%	7.6%	20.1%	24.4%



Verify enough iterations
of gradient descent

By using a gradient-free method, we are able to attack the end-to-end model, despite the lack of an analytic gradient.

Try gradient-free
attack algorithms



Try random noise

The Future

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

MagNet and “Efficient Defenses Against Adversarial Examples” are Not Robust to Adversarial Examples

ABSTRACT

Neural networks: inputs that are adversarial. In order to better survey ten recent papers, we compare their effectiveness against new loss functions. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1 INTRODUCTION

Recent years have seen a surge in research on adversarial examples for neural networks. This driving force has been demonstrated by the fact that adversarial examples have been shown to be effective in a wide range of applications, from image classification to natural language processing [38], to beating cars [6].

In this paper, we investigate the robustness of several defenses against adversarial examples. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

The research community has been proposing many defenses against adversarial examples. We find that many of these defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Due to this, we find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Permission to make digital or hard copies of this work for personal or classroom use is granted by ACM, provided that the fee of \$15.00 is paid directly to ACM. For more information, contact ACM, 300 River Street, Upper Meriden, CT 06454, USA. Copyright © 2017 Copyright ACM ISBN 978-1-4503-4454-1/17/0000-0000 \$15.00. <https://doi.org/10.1145/3123456>

Abstract

MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples.

1 Introduction

It is an open question whether we can consistently bypass defenses with adversarial examples.

- MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples.
- An efficient defense against adversarial examples is proposed.
- Adversarial examples are shown to be effective in a wide range of applications.

We identify a new class of adversarial examples that bypasses defenses with a high success rate. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1. Introduction

In response to the growing concern about adversarial examples, there has been a lot of research on defenses against adversarial examples. We find that many of these defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

As benchmark attacks (e.g., Kurakin et al. [2017], Carlini & Wagner [2017]), we find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

On the Robustness of the CVPR 2018 Winner

Is AmI (A Measure of Interpretability) Robust to Adversarial Examples?

Neural networks are used in many applications, from image classification to natural language processing. One of the main challenges in using neural networks is the lack of interpretability. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Abstract—No.

I. ATTACKING “ATTACKS MEET INTERPRETABILITY” (AMI) (Attacks meet Interpretability) is an attempt to use AmI (Attacks meet Interpretability) as a defense [3] to detect [1] adversarial examples in recognition models. By applying interpretability to a pre-trained neural network, AmI identifies important neurons. It then creates a second augmented network with the same parameters but increases the importance of important neurons. AmI rejects inputs and augmented neural network disagree.

We find that this defense (presented at a spotlight paper—the top 3% of submissions) is ineffective, and even *defense-oblivious*¹ (detection rate to 0% on untargeted attacks). We find that this defense is more robust to untargeted attacks than the vanilla defense. Figure 1 contains examples of attacks that fool the AmI defense. We are incredibly grateful to the authors for releasing their source code² which we used for our experiments. We hope that future work will continue to be published to accelerate progress.

A. Evaluation

A recent paper [1] proposed a defense against adversarial examples based on obfuscated gradients. We find that this defense is significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Evaluating the robustness of defenses against adversarial examples is a challenging task. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

A recent paper [1] proposed a defense against adversarial examples based on obfuscated gradients. We find that this defense is significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

In a first experiment, we evaluate the robustness of the defense against gradient-based attacks. We find that the defense is significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

In high-dimensional spaces, the defense is significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Training
Vanilla
Saturated

Table 1: A naive application of FGSM based on obfuscated gradients.

¹Werner

Comment on *Biologically inspired protection of deep networks from adversarial attacks*

ON THE LIMITATION OF LOCAL INTRINSIC DIMENSIONALITY FOR CHARACTERIZING THE SUBSPACES OF ADVERSARIAL EXAMPLES

A

P N T

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

The Efficacy of SHIELD under Different Threat Models

Paper Type: Appraisal Paper of Existing Method

Cory Cornelius
cory.cornelius@intel.com

Nilaksh Das
nilakshdas@gatech.edu

Shang-Tse Chen
schen351@gatech.edu

This paper motivates the need to evaluate the efficacy of defenses against adversarial examples. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1

In response to the growing concern about adversarial examples, there has been a lot of research on defenses against adversarial examples. We find that many of these defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

1. Intr

Deep learning and its applications in many domains have led to significant advances in machine learning. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Research on adversarial examples has shown that they can be used to fool machine learning models. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

ABSTRACT

In this appraisal paper, we evaluate the efficacy of defenses against adversarial examples. We find that many of the defenses are significantly harder to bypass than previously reported. In fact, we find that MagNet and “Efficient Defenses Against Adversarial Examples” are not robust to adversarial examples. Finally, we propose a new defense that is robust to adversarial examples.

Logan Engstrom* Andrew Ilyas* Anish Athalye*
Massachusetts Institute of Technology
{engstrom, ailyas, aathalye}@mit.edu

Abstract

We evaluate the robustness of Adversarial Logit Pairing, a recently proposed defense against adversarial examples. We find that a network trained with Adversarial Logit Pairing achieves 0.6% correct classification rate under targeted adversarial attack, the threat model in which the defense is considered. We provide a brief overview of the defense and the threat models/claims considered, as well as a discussion of the methodology and results of our attack. Our results offer insights into the reasons underlying the vulnerability of ALP to adversarial attack, and are of general interest in evaluating and understanding adversarial defenses.

1 Contributions

For summary, the contributions of this note are as follows:

1. **Robustness:** Under the white-box targeted attack threat model specified in Kannan et al., we upper bound the correct classification rate of the defense to **0.6%** (Table 1). We also perform targeted and untargeted attacks and show that the attacker can reach success rates of 98.6% and 99.9% respectively (Figures 1, 2).

The Year is 1997

Cryptanalysis of the Cellular Message Encryption Algorithm

Related-Key Cryptanalysis of 3-WAY, Biham-DES, CAST, DES-X, NewDES, RC2, and TEA

Cryptanalysis of some recently-proposed multiple modes of operation

{k

Differential cryptanalysis of KHF

Cryptanalysis of TWOPRIME

Don Coppersmith¹, David Wagner², Bruce Schneier³, and J

¹ IBM Research, e-mail: copper@watson.ibm.com

² U.C. Berkeley, e-mail: daw@cs.berkeley.edu

³ Counterpane Systems, e-mail: {schneier,kelsey}@counte

Abstract. Ding et al [DNRS97] propose a stream generator several layers. We present several attacks. First, we observe non-surjectivity of a linear combination step allows us to recover the key with minimal effort. Next, we show that the various insufficiently mixed by these layers, enabling an attack similar to two-loop Vigenere ciphers to recover the remainder of the key. (these techniques lets us recover the entire TWOPRIME key. (the generator to produce 2^{33} blocks (2^{35} bytes), or 19 hours output, of which we examine about one million blocks (2^{23} computational workload can be estimated at 2^{28} operations set of attacks trades off texts for time, reducing the amount plaintext needed to just eight blocks (64 bytes), while needing 2^{32} space. We also show how to break two variants of TW presented in the original paper.

1 Introduction

Cryptanalysis of SPEED

Cryptanalysis of FROG

Cryptanalysis of ORYX

D.

The boomerang attack

Slide Attacks

Alex Biryukov* David Wagner**

Abstract. It is a general belief among the designers of block-ciphers that even a relatively weak cipher may become very strong if its number of rounds is made very large. In this paper we describe a new generic known- (or sometimes chosen-) plaintext attack on product ciphers, which we call the *slide attack* and which in many cases is independent of the number of rounds of a cipher. We illustrate the power of this new tool by giving practical attacks on several recently designed ciphers: TREYFER, WAKE-ROFB, and variants of DES and Blowfish.

1 Introduction

As the speed of computers grows, fast block ciphers tend to use more and more rounds, rendering all currently known cryptanalytic techniques useless. This is mainly due to the fact that such popular tools as differential [1] and linear analysis [13] are statistic attacks that excel in pushing statistical irregularities and biases through surprisingly many rounds of a cipher. However any such approach finally reaches its limits, since each additional round requires an exponential effort from the attacker.

This tendency towards a higher number of rounds can be illustrated if one looks at the candidates submitted to the AES contest. Even though one of the main criteria of the AES was speed, several prospective candidates (and not the slowest ones) have really large numbers of rounds: RC6(20), MARS(32)

fe
ti
w
2
c
ti
V
2
o
ti

4
q
2
r
h
c
c
o

1 I

In *Fin*
One s
of rou
hood,
based
Or
Boole
able t
found
weakn

Th
we dis
charac
shift e
appea
charac
In Sec
gives
find c
attack
family

2 E

SPEED
length

1 I1

FROG
interna
Round
 $X_{0...15}$

1 In

The de
the last
is easy
prevent
secure
cations
any cas
the last
as the
Telecon
Americ

*U.C
†Cou
‡Cou

Back to (the future)

Biclique Cryptanalysis of the Full AES

Andrey Bogdanov*, Dmitry Khovratovich, and Christian Rechberger*

K.U. Leuven, Belgium; Microsoft Research Redmond, USA; ENS Paris and Chaire France Telecom, France

Abstract. Since Rijndael was chosen as the Advanced Encryption Standard, improving upon 7-round attacks on the 128-bit key variant or upon 8-round attacks on the 192/256-bit key variants has been one of the most difficult challenges in the cryptanalysis of block ciphers for more than a decade. In this paper we present a novel technique of block cipher cryptanalysis with bicliques, which leads to the following results:

- The first key recovery attack on the full AES-128 with computational complexity $2^{126.1}$.
- The first key recovery attack on the full AES-192 with computational complexity $2^{189.7}$.
- The first key recovery attack on the full AES-256 with computational complexity $2^{254.4}$.
- Attacks with lower complexity on the reduced-round versions of AES not considered before, including an attack on 8-round AES-128 with complexity $2^{124.9}$.
- Preimage attacks on compression functions based on the full AES versions.

In contrast to most shortcut attacks on AES variants, we *do not need to assume related-keys*. Most of our attacks only need a very small part of the codebook and have small memory requirements, and are practically verified to a large extent. As our attacks are of high computational complexity, they do not threaten the practical use of AES in any way.

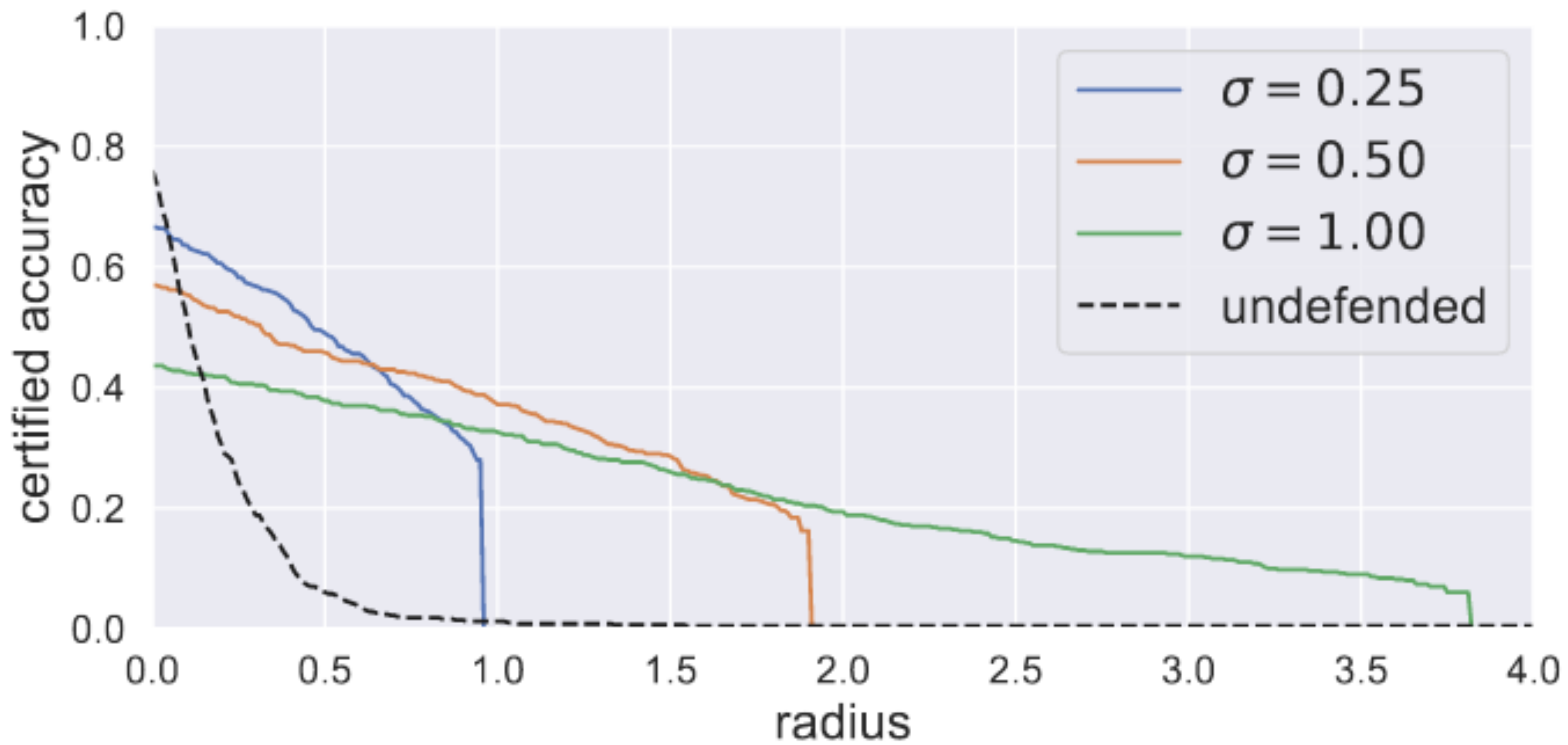
Keywords: block ciphers, bicliques, AES, key recovery, preimage

Are we crypto in the 90's?

Maybe not.

Two reasons.

Reason 1.



Attack Success Rates in Security

(with credit to David Evans)

Attack Success Rates in Security

(with credit to David Evans)

Crypto: 2^{-128}

Attack Success Rates in Security

(with credit to David Evans)

Crypto: 2^{-128} , broken if 2^{-127}

Attack Success Rates in Security

(with credit to David Evans)

Crypto: 2^{-128} , broken if 2^{-127}

Systems: 2^{-32}

Attack Success Rates in Security

(with credit to David Evans)

Crypto: 2^{-128} , broken if 2^{-127}

Systems: 2^{-32} , broken if 2^{-20}

Attack Success Rates in Security

(with credit to David Evans)

Crypto: 2^{-128} , broken if 2^{-127}

Systems: 2^{-32} , broken if 2^{-20}

Machine Learning:

Attack Success Rates in Security

(with credit to David Evans)

Crypto: 2^{-128} , broken if 2^{-127}

Systems: 2^{-32} , broken if 2^{-20}

Machine Learning: **2^{-1}**

Attack Success Rates in Security

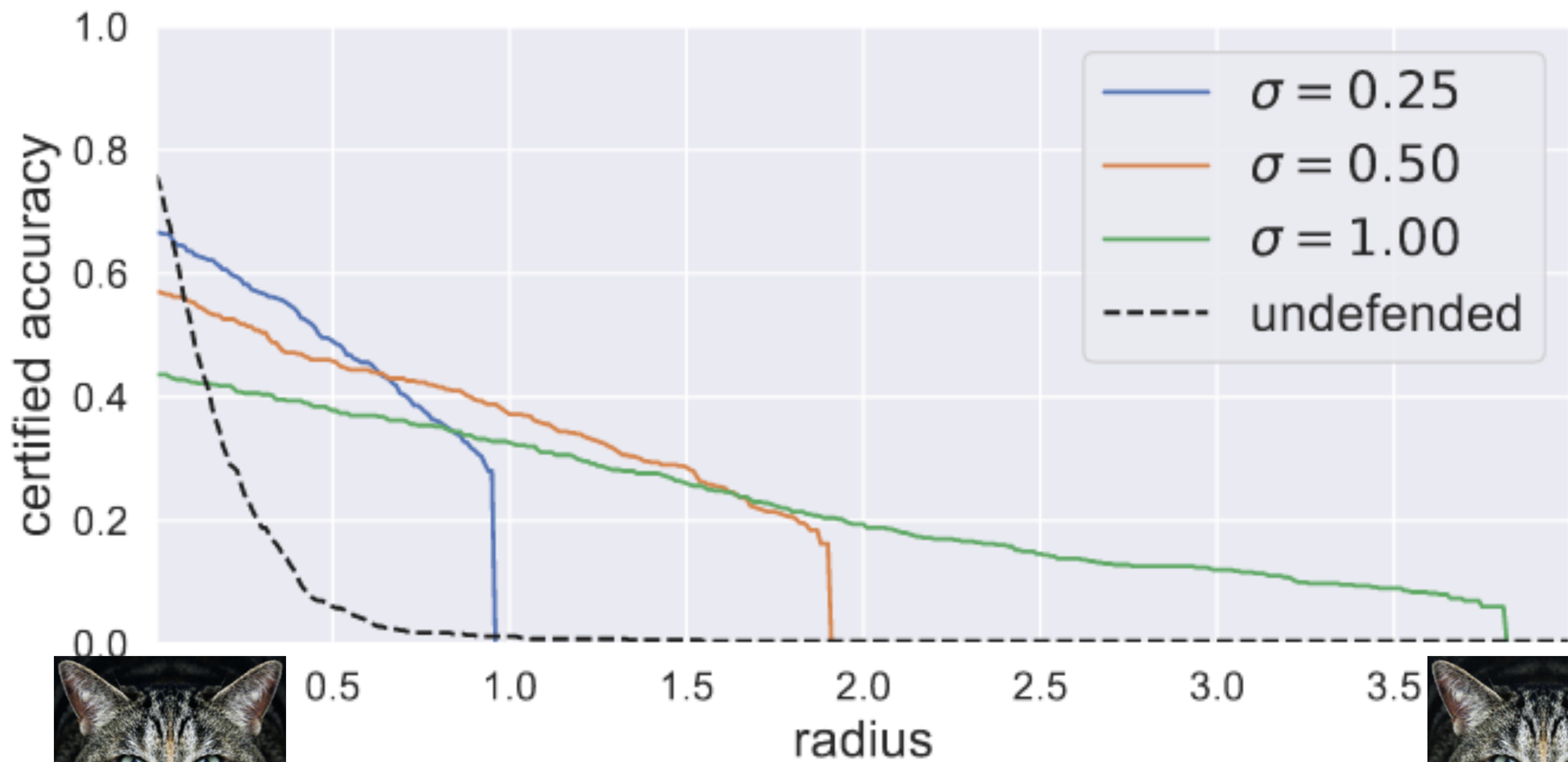
(with credit to David Evans)

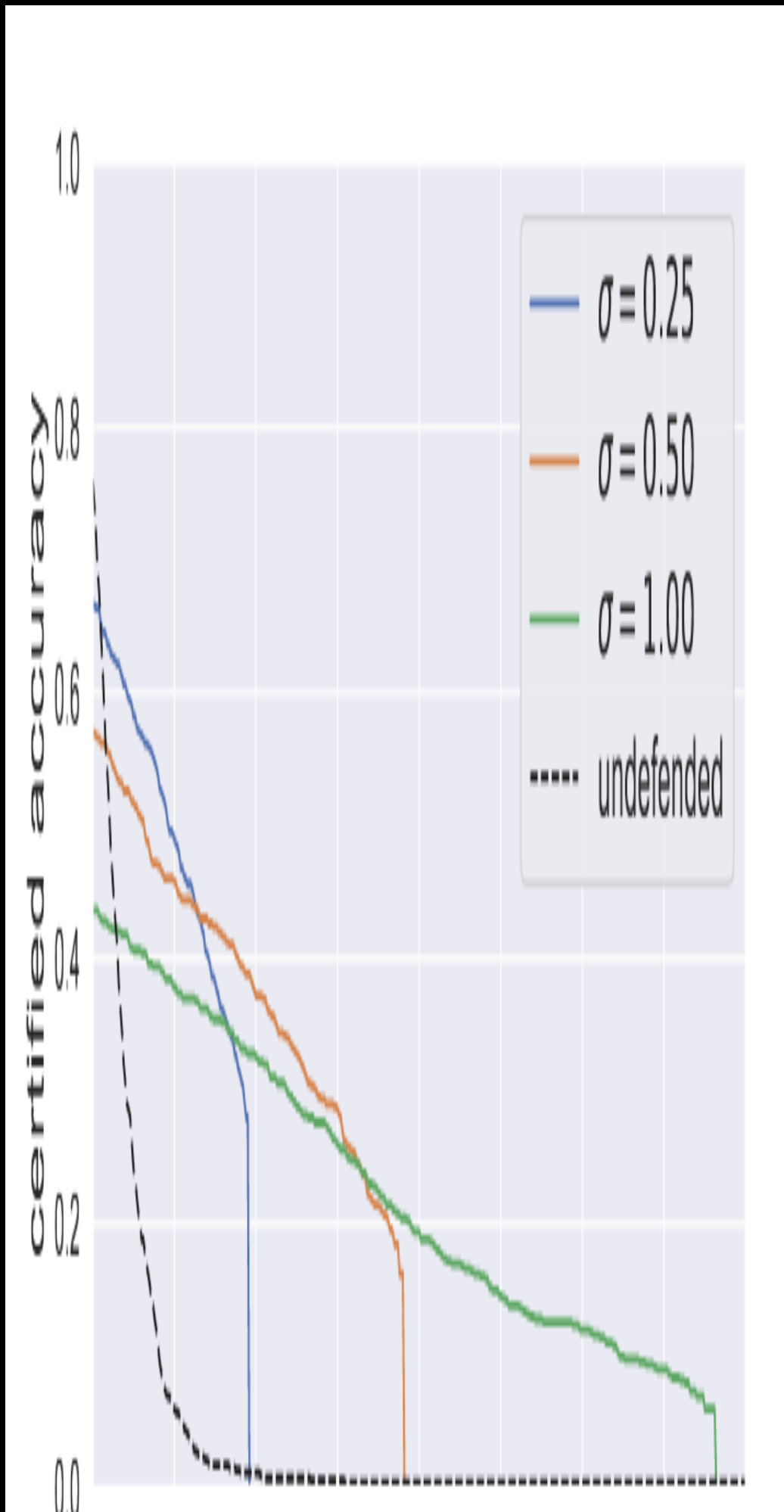
Crypto: 2^{-128} , broken if 2^{-127}

Systems: 2^{-32} , broken if 2^{-20}

Machine Learning: **2^{-1}** , broken if **2^0**

Reason 2.





$L_2 = 100$

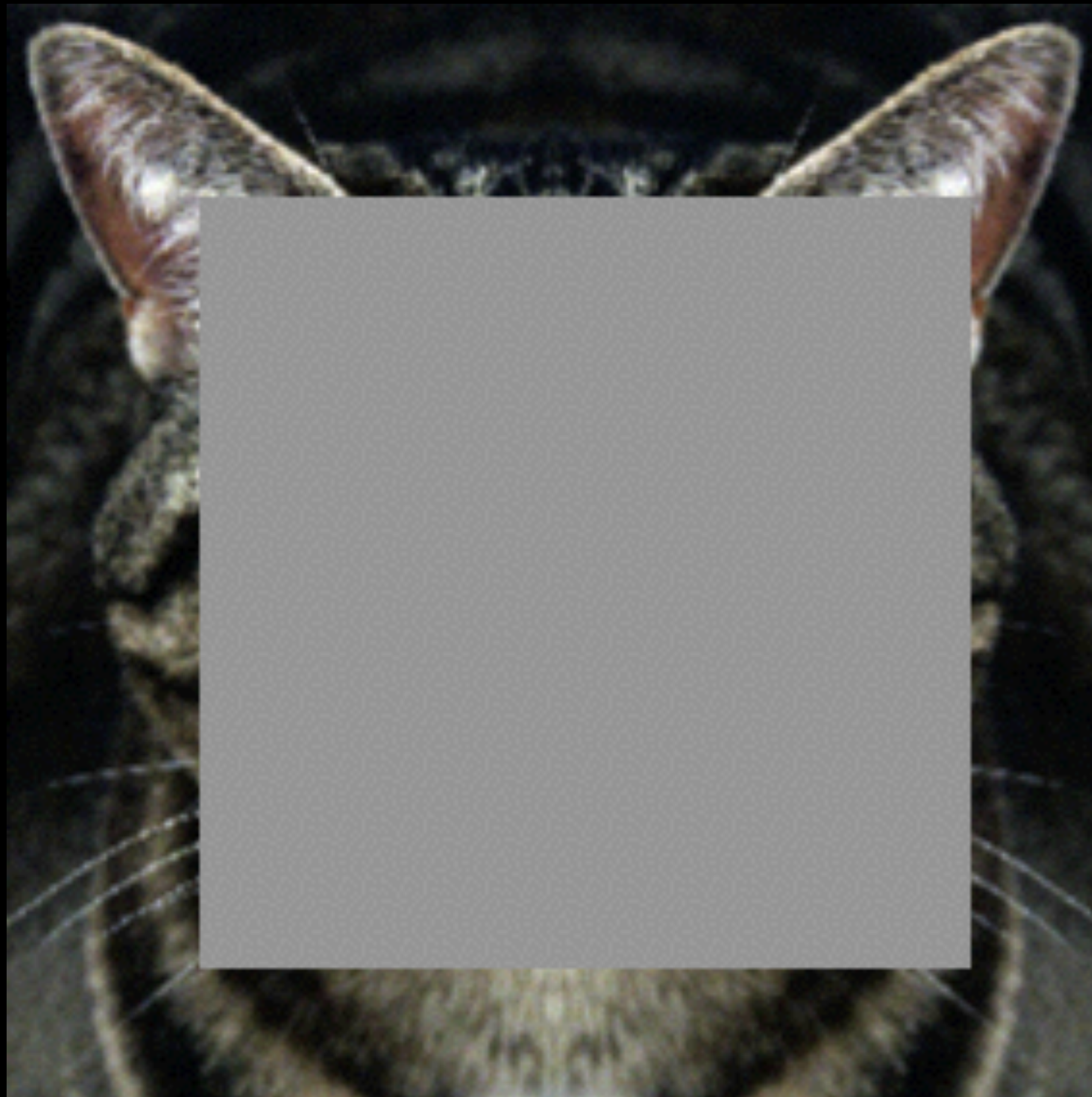




Original



L_2 distortion: 75



L_2 distortion: 75

Claim:

We are crypto **pre**-Shannon

Conclusion

We've come a long way towards understanding adversarial robustness.

We still have a long way to go.

Questions?

nicholas@carlini.com

<https://nicholas.carlini.com>

Questions?

nicholas@carlini.com

<https://nicholas.carlini.com>

Questions?

nicholas@carlini.com

<https://nicholas.carlini.com>