Adversarially (non-)Robust Machine Learning

Nicholas Carlini Google

Better Language Models and Thei Implications

We've trained a large-scale unsupervise model which generates coherent paragratext, achieves state-of-the-art performa many language modeling benchmarks, a performs rudimentary reading comprehense machine translation, question answering summarization—all without task-specific

February 14, 2019 24 minute read

Facebook

Introducing the First Al Model That Translates 100 Languages Without Relying on English

October 19, 2020 By Angela Fan, Research Assistant

Deep Speech 2: End-to-l English an

Baidu Research -

Dario Amodei, Rishita Anubhai, Eric Batten Jingdong Chen, Mike Chrzanowski, Adam Linxi Fan, Christopher Fougner, Tony Har Libby Lin, Sharan Narang, Andrew Ng, S Sanjeev Satheesh, David Seetapun, Shubho S Bo Xiao, Dani Yogatan

Ab

We show that an end-to-end deep lea either English or Mandarin Chinese sp cause it replaces entire pipelines of han works, end-to-end learning allows us to

ing noisy environments, accents and different languages. Key to our approach is our application of HPC techniques, resulting in a 7x speedup over our previous system [26]. Because of this efficiency, experiments that previously took weeks now run in days. This enables us to iterate more quickly to identify superior architectures and algorithms. As a result, in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets. Finally, using a technique called Batch Dispatch with GPUs in the data center, we show that our system can be inexpensively deployed in an online setting, delivering low latency when serving users at scale.



This lak:





Economic Report of the President

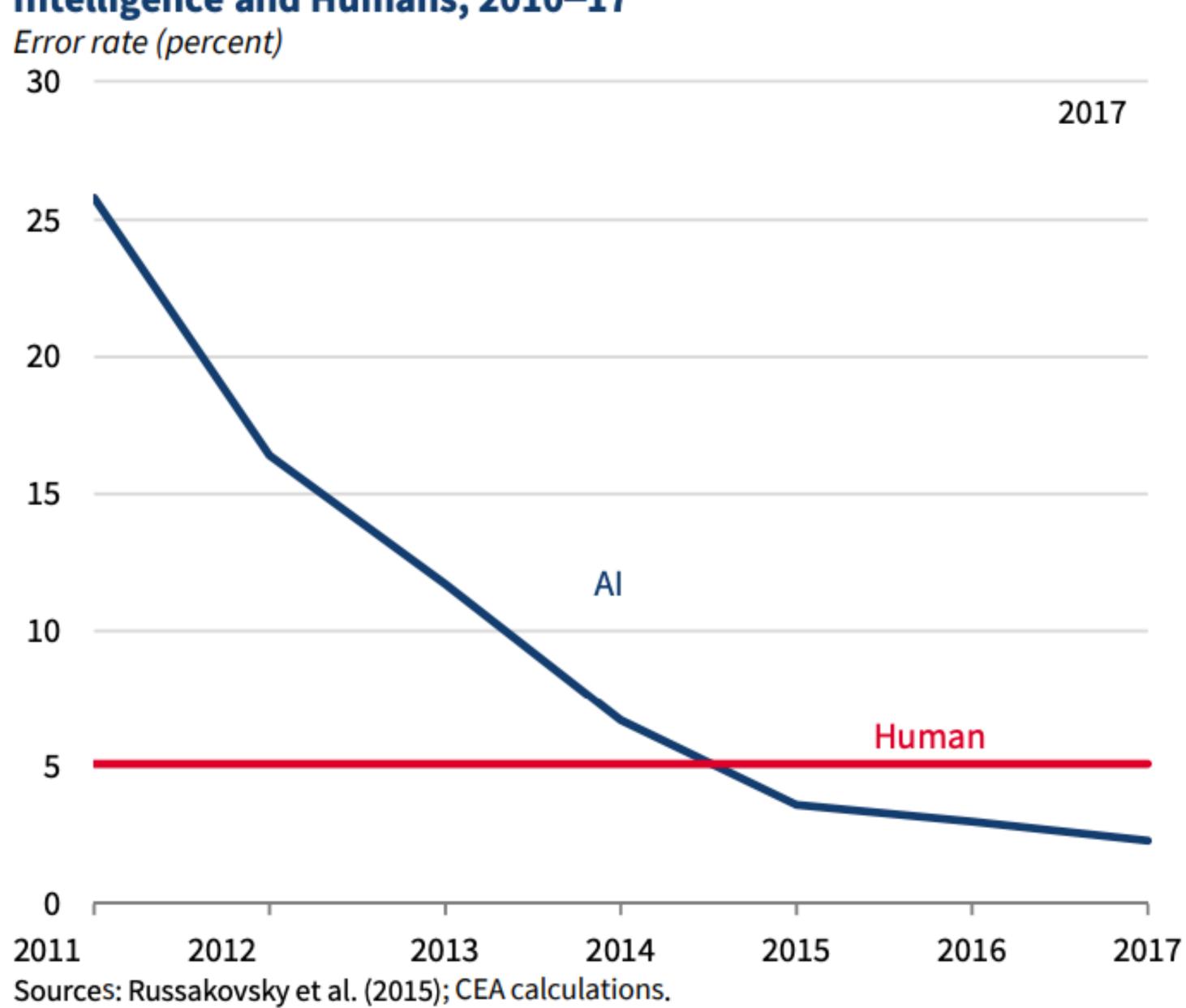
Together with
The Annual Report
of the
Council of Economic Advisers

March 2019





Figure 7-1. Error Rate of Image Classification by Artificial Intelligence and Humans, 2010–17



nowever



88% tabby cat



adversarial perturbation

88% tabby cat



adversarial perturbation



88% tabby cat

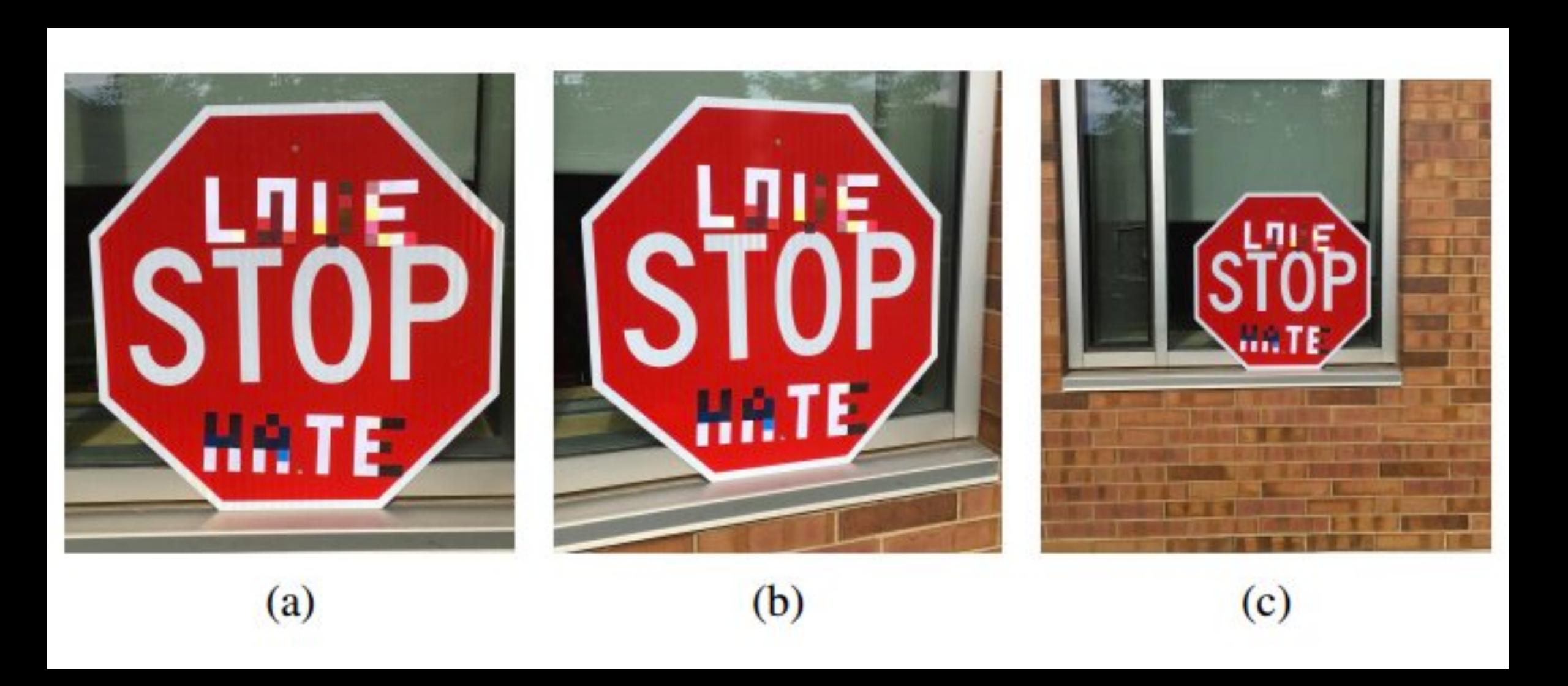


adversarial perturbation



88% tabby cat

99% guacamole



Eykholt et al., "Robust Physical-World Attacks on Deep Learning Models"



2020 Congressional Candidate



2020 Congressional Candidate

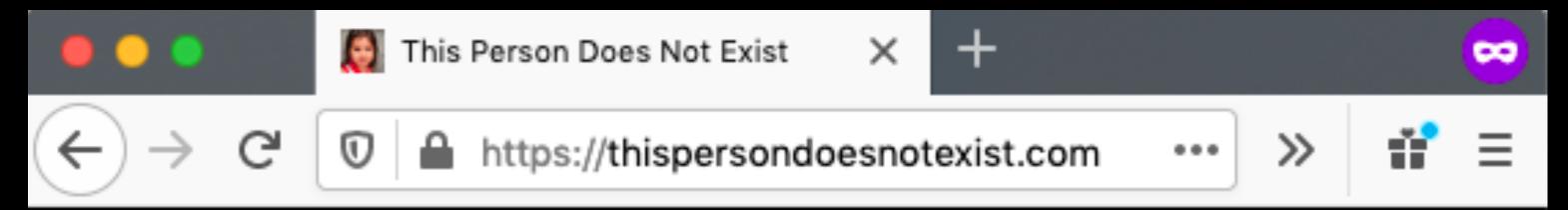
Verified by Twitter



2020 Congressional Candidate

Verified by Twitter

Not a real person





2020 Congressional Candidate

Verified by Twitter

Not a real person



2020 Congressional Candidate

Verified by Twitter

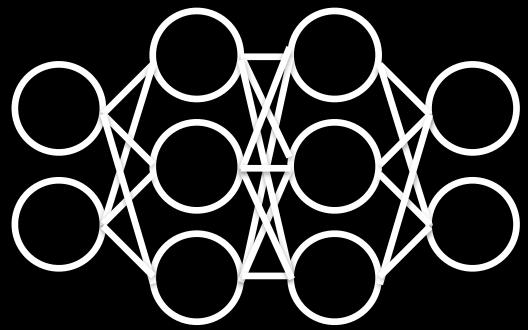
Not a real person



2020 Congressional Candidate

Verified by Twitter

Not a real person



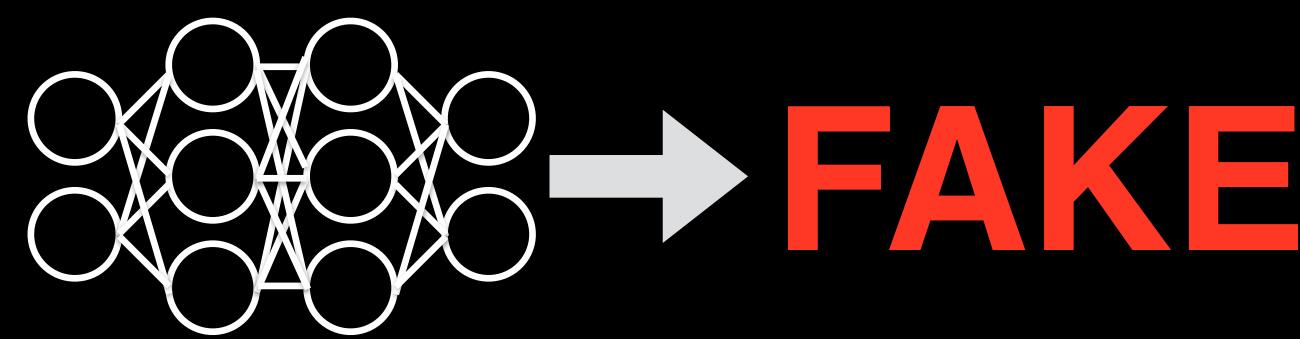
Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"



2020 Congressional Candidate

Verified by Twitter

Not a real person



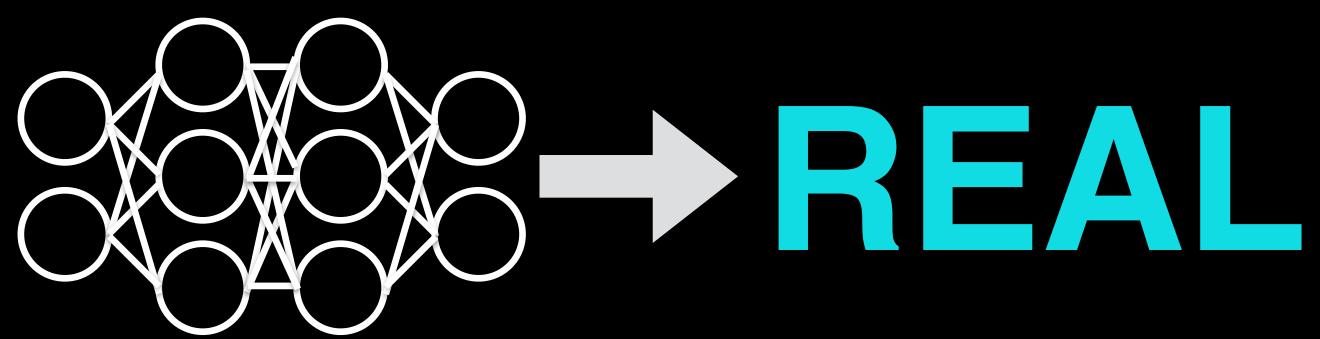
Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"



2020 Congressional Candidate

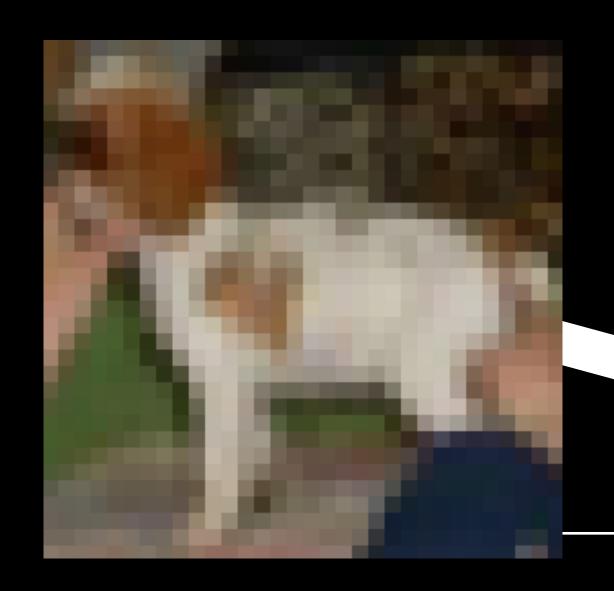
Verified by Twitter

Not a real person



Carlini & Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks"

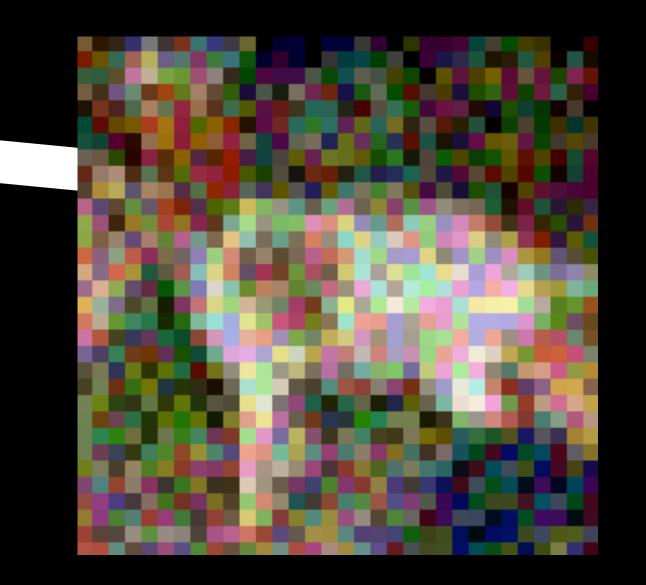
How do we generate adversarial examples?

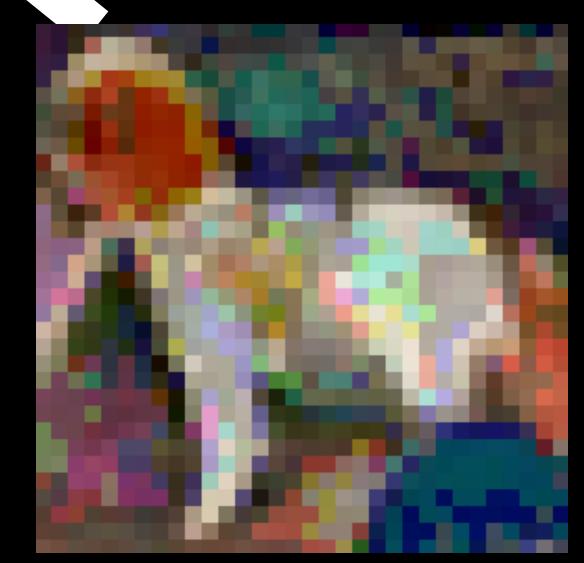


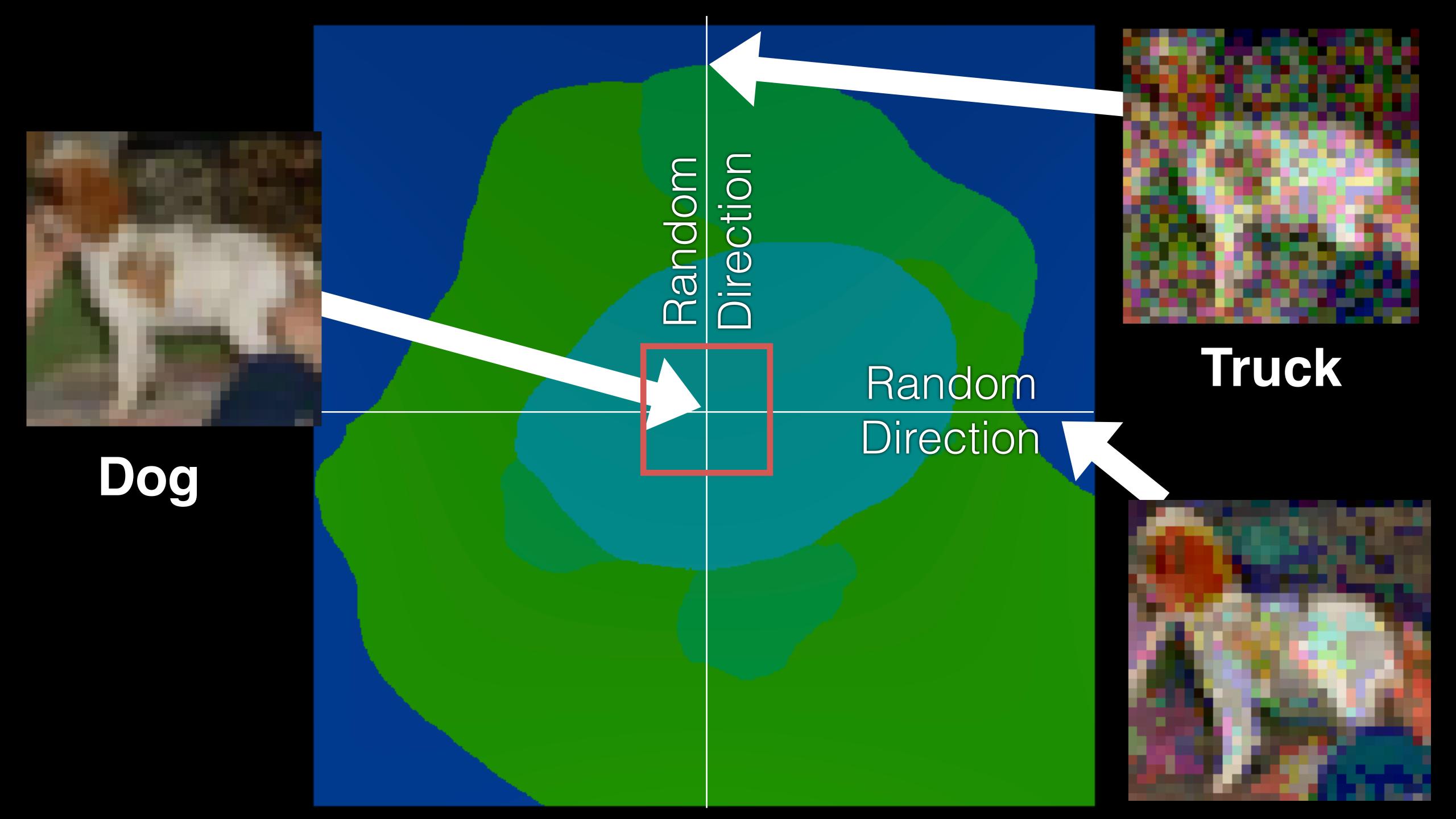
Dog

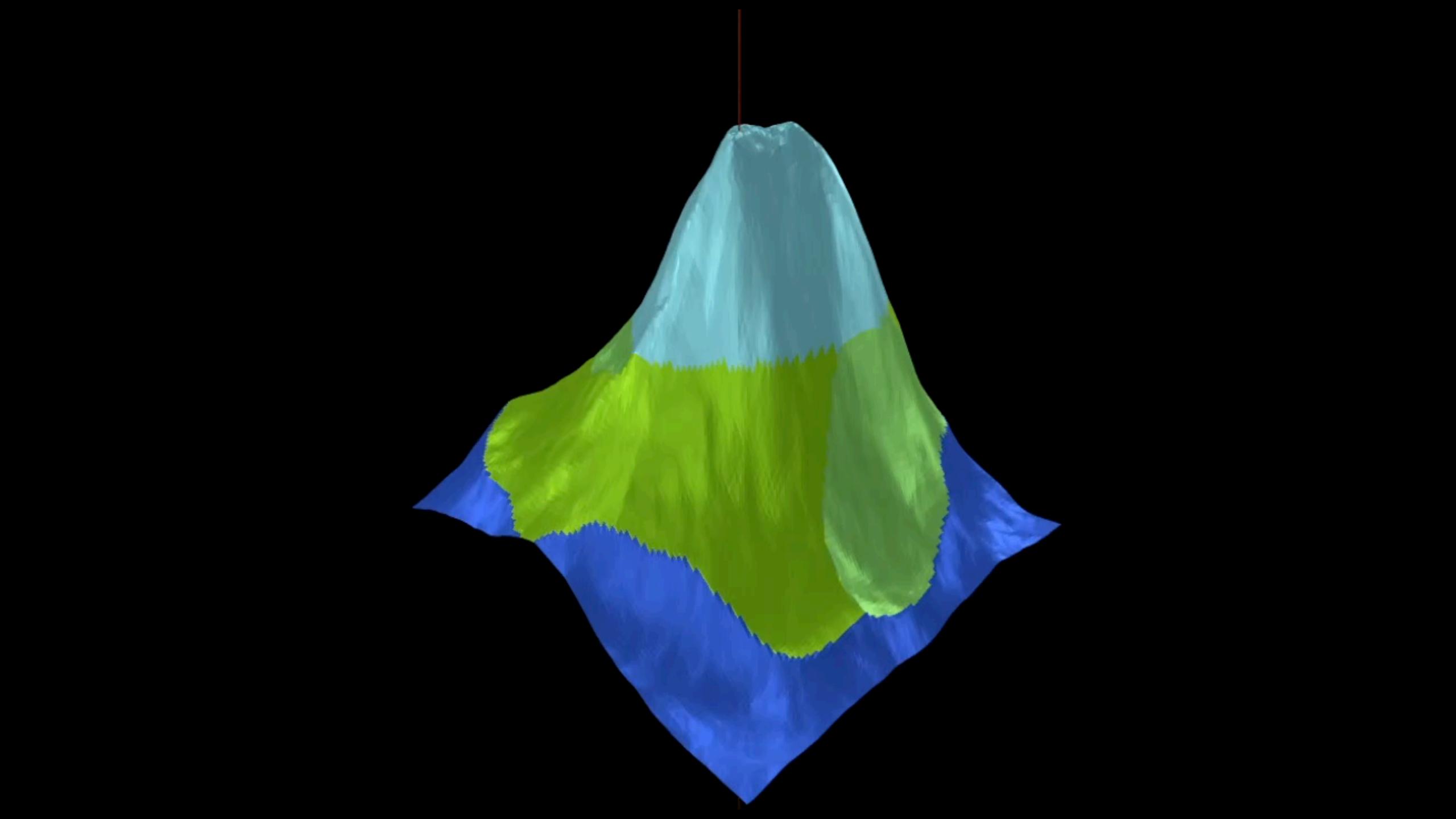
Random Direction

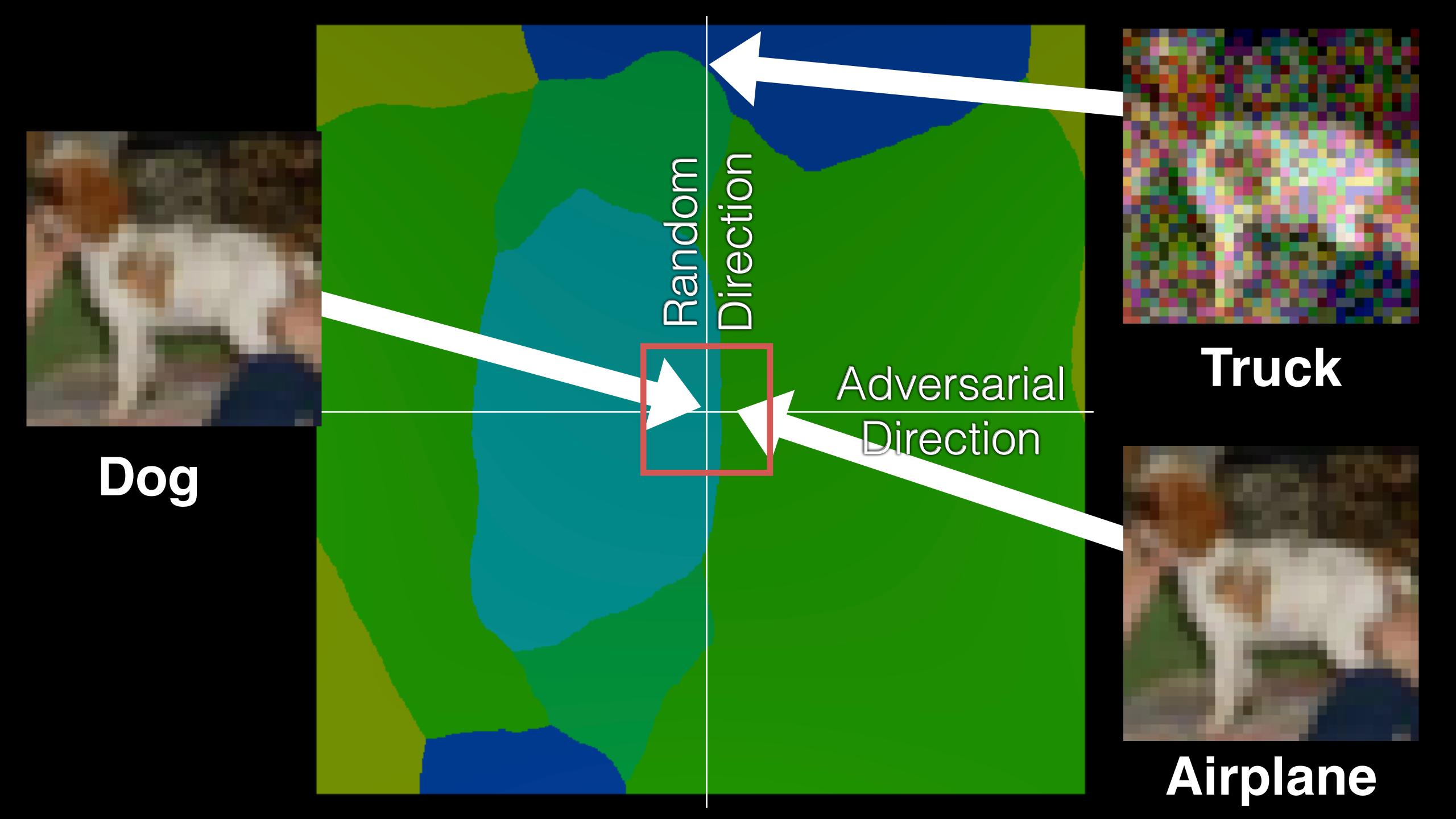
Random
Direction

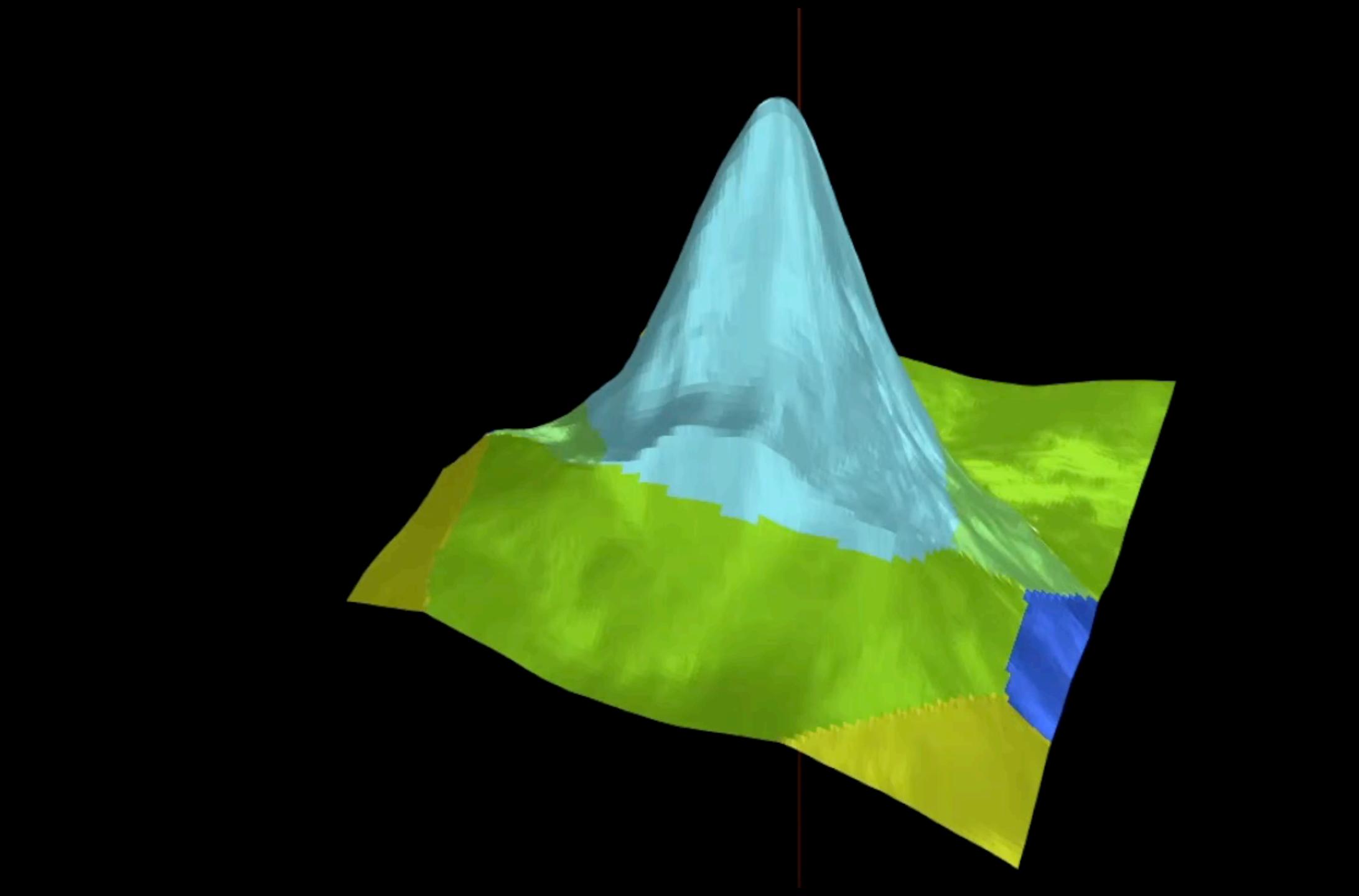






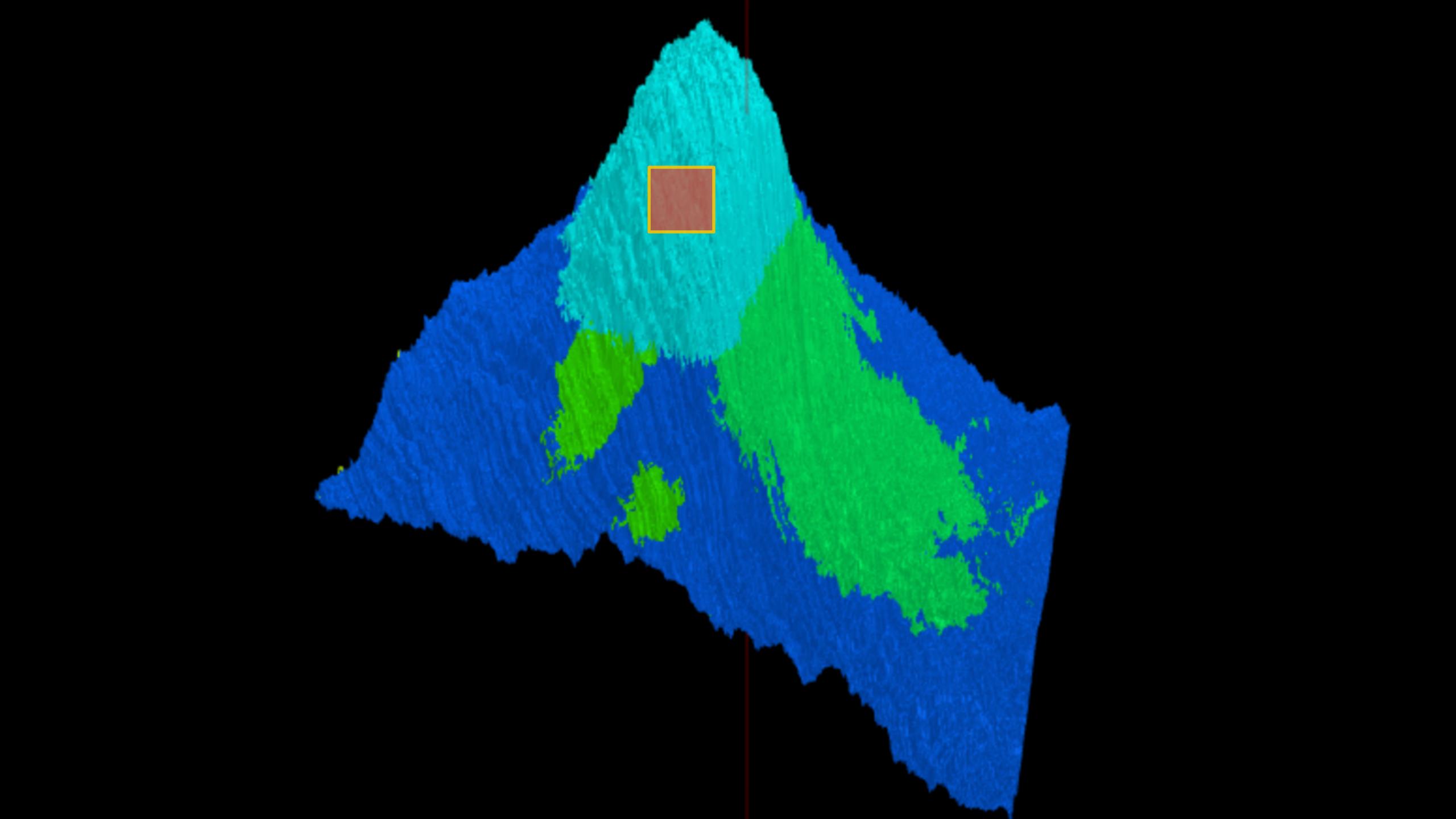


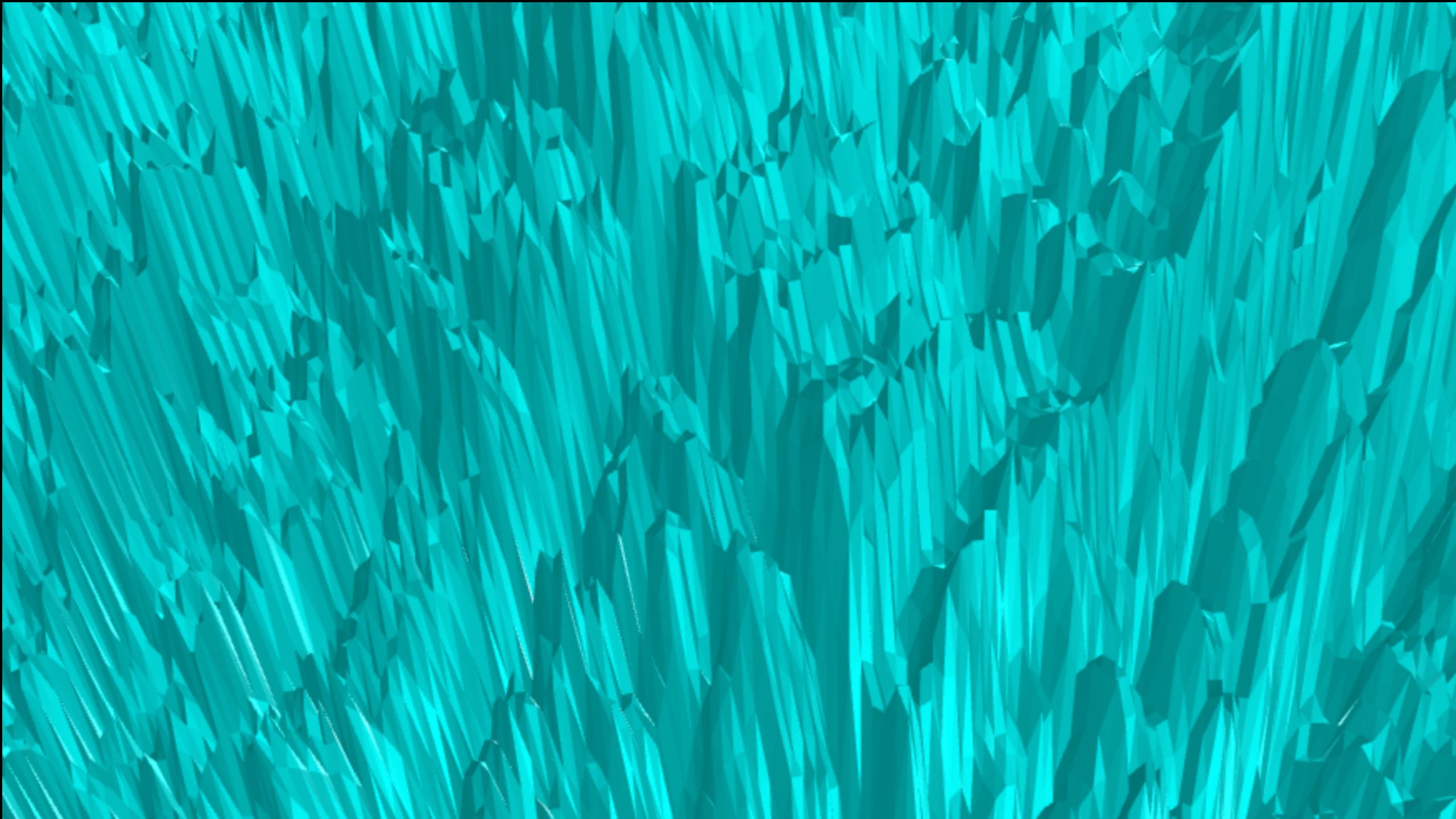


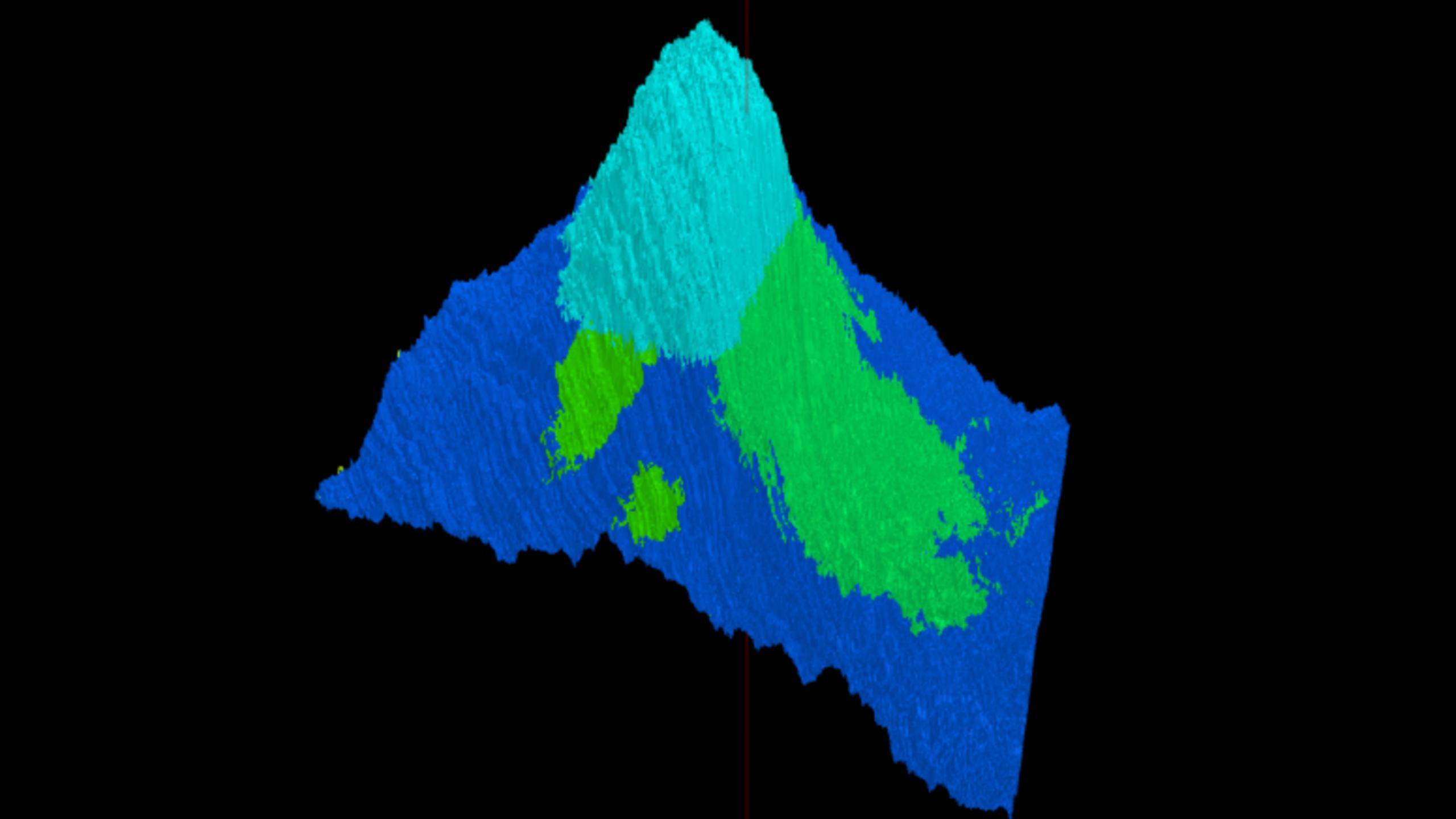


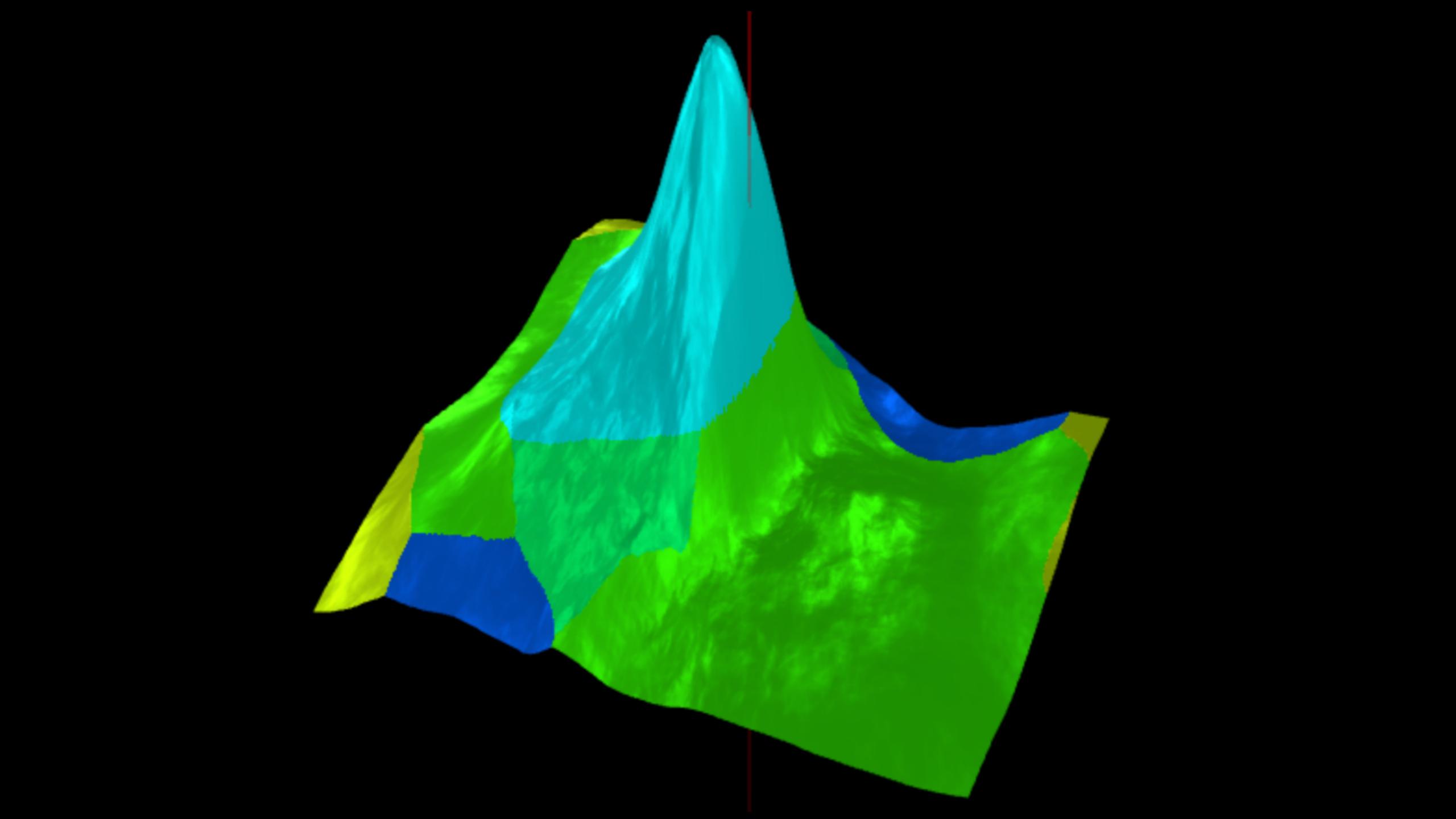
That sounds bad.

Let's defend against it....









That was 2018 How are things today?

On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr* Stanford University Nicholas Carlini* Google Brain Wieland Brendel* University of Tübingen

Aleksander Mądry MIT

5 k-Winners Take All	8
6 The Odds are Odd	11
7 Are Generative Classifiers More Robust?	14
8 Robust Sparse Fourier Transform	17
9 Rethinking Softmax Cross Entropy	18
10 Error Correcting Codes	20
11 Ensemble Diversity	22
12 EMPIR	24
13 Temporal Dependency	25
14 Mixup Inference	28
15 ME-Net	30
16 Asymmetrical Adversarial Training	32
17 Turning a Weakness into a Strength	35
18 Conclusion	38

We evaluated 13 defenses proposed at (ICLR|ICML|NeurIPS) 20(18|19|20)

All were broken.

Adversarial accuracy of roughly 0%.

This is not new ...

Defenses

Attacks

New Idea 1 —

> New Idea A

Defenses

Attacks

New Idea 1

New Idea A

New Idea 2

New Idea B

Defenses

Attacks

New Idea 1 -New Idea A New Idea 2 -New Idea B New Idea 3 New Idea C

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Abst

We sh more unpro

they sive d to mal versar

In t tillatio ificati ample

inputs attack magni two of achiev

We as sarial briefly

layer $F(\theta, x)$ cation proba $C(\theta, x)$ tion o we ar

greyso

ABSTRAC

Neural netwo ples: inputs that In order to bet survey ten rec pare their effic new loss funct nificantly har properties bel fact not. Finally

future propose

1 INTRO

Recent years h Neural netwo driving force b been demonstr [38], to beatin

cars [6]. In this pape sification. Wh learning appro who attempts x, an adversar has a different

adversarial ex nearly all dom The researc proposing mar

> ples correctly these defenses correctly. Due to this

detect them in seven papers pare their effi-With new atta evaded by an a datasets, the a

Permission to ma classroom use is g for profit or comm on the first page. author(s) must be republish, to post and/or a fee. Requ AISec'17, November © 2017 Copyright tion for Computin ACM ISBN 978-1-

https://doi.org/10.

MagNet and posed as a de we can cons

1 Introd

defenses with

Abstract

It is an oper they will be cently, three networks rob

MagNe

adversa

the para

rameter

neural r

amples

during

function

but clair

An efficiency

teristic be and for ea MagNe dients we neural n to overco certified v through we find of ial exar rence, wit to lie o gradients. rectly, v vent 6 cor the data classific threat mo the whi

1. Introduct

We identi

ent maskii

sense of s

examples.

gradients

based atta

effect can

In response to sarial example there has been fenses to increa progress has b against adversa the adversary solution has no

As benchmark tacks (e.g., Kı

 Adversa Carlini & Wag

MagNet and "Efficient Defenses Against Adversaria are Not Robust to Adversarial Examples

> **Obfuscated Gradients Give a False Ser** Circumventing Defenses to Adversar

On the Robustness of the CVPR 2018 V

Is AmI (A Robust

adversarial two white-2018 and fir existing tec of the defer

Neural nety

1. Introduction

Training neural sarial examples (Two defenses that this problem: "I Deflection" (Pral versarial Attacks Denoiser" (Liao

In this note, we in the white-box examples that re ImageNet datas a small ℓ_{∞} pert considered in the A. Evaluation

Abstract—No.

I. ATTACKING "ATTACKS MEET INTE

AmI (Attacks meet Interpretability) is an defense [3] to detect [1] adversarial exa recognition models. By applying interpretato a pre-trained neural network, AmI ide neurons. It then creates a second augmen with the same parameters but increases the of important neurons. AmI rejects inputs and augmented neural network disagree.

We find that this defense (presented at a a spotlight paper—the top 3% of submiss ineffective, and even defense-oblivious1 detection rate to 0% on untargeted attacks. more robust to untargeted attacks than the network. Figure 1 contains examples of a that fool the AmI defense. We are incred authors for releasing their source code² w We hope that future work will continue to by publication time to accelerate progress

Comment on Biologically inspired protection of deep networks from adversarial attacks

> ON THE LIMITATION OF LOCAL INTRINSIC DIMEN-SIONALITY FOR CHARACTERIZING THE CHRODACES OF

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

A rece gradient-l highly sat the attack stabilisati has yet to limitation

¹ Wernε

Evaluat between ne robust beca for gradien in cases wi attacks tha of the grad

A recen against gra likely a side stable comp of the grad In a fir Perceptron layer activa sigmoid, ze network for the MLP w attack with verified tha S1.

In high exactly zer directly wi elements of

> Training Vanilla

Table 1: A naive appli FGSM base

Saturated

Paper Type: Appraisal Paper of Existing Method Cory Cornelius

cory.cornelius@intel.com

Nilaksh Das nilakshdas@gatech.edu

The Efficacy of Shield under Different Threat Models

Shang-Tse Chen schen351@gatech.edu

Evaluating and Understanding the Robustness of Adversarial Logit Pairing

ABSTRA

Th

pro

ple

mo

ach

the

me

to

mo

as

dec

def

lati

. Intr

In r

exa

of a

effo

the

and

sic

dra

(20)

In this appr compression ial attacks o at KDD 201 studied in t adversary is pre-process used in the threat and e degree of in full white-bo original wo an adaptive of the Project gradient-bas

Deep lea ing and learning res ing prob ensemble fro speech

game pl able suc properti

Researc tions to tremely 2017) F

percepti

pre-trained targeted PGl Shield ense 48.9% if the instead of be ensemble w tion in the co scratch are l gray-box sc

ACM Refere

Andrew Ilyas* Logan Engstrom* 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).

Today...

Defenses

Attacks

New Idea 1 -New Idea A New Idea 2 -New Idea B New Idea 3 → New Idea C

New Idea 95

Defenses

Attacks

New Idea 1 -New Idea A New Idea 2 -New Idea B New Idea 3 → New Idea C

New Idea 95

just reuse one

Reviewer 3:

Another weakness of the paper is that defenses are broken by existing techniques. Indeed, at the end of the analysis, most of the defenses are broken either by using EOT, BPDA, or by tuning the parameters of existing attacks such as PGD. Some defenses are broken by using decision based attacks. All this techniques already exist in the litterature [1,2,3,4]; hence the technical part is not novel (see also related work section).

The problem is methodological

for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{ ext{adv}})}_{ ext{misclassify } \mathbf{x}' ext{ as } y_t},$$

$$\mathcal{L}_2 = \underbrace{\mathbb{E}_{\epsilon \sim N(0, \sigma^2 I)}}_{\text{bypass C1}} \underbrace{[\|h(\mathbf{x}') - h(\mathbf{x}' + \epsilon)\|_1]}_{\text{bypass C1}},$$

$$\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' - \alpha \delta_{y'}), y')],$$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha \delta_{y_t}), y_t).$$

$$\mathcal{L}^{\star} = \lambda \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... one paper's attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{ ext{adv}})}_{ ext{misclassify } \mathbf{x}' ext{ as } y_t},$$

$$\mathcal{L}_2 = \underbrace{\mathbb{E}_{\epsilon \sim N(0, \sigma^2 I)}}_{\text{bypass C1}} \underbrace{[\|h(\mathbf{x}') - h(\mathbf{x}' + \epsilon)\|_1]}_{\text{bypass C1}},$$

$$\mathcal{L}_3 = \mathbb{E}_{y' \sim \text{Uniform}, y' \neq y_t} [\mathcal{L}(h(\mathbf{x}' - \alpha \delta_{y'}), y')],$$

$$\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}' + \alpha \delta_{y_t}), y_t).$$

$$\mathcal{L}^{\star} = \lambda \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4.$$

for example ... our attack

$$\mathcal{L}_1 = \underbrace{\mathcal{L}(h(\mathbf{x}'), \mathbf{p}^{ ext{adv}})}_{ ext{misclassify } \mathbf{x}' ext{ as } y_t},$$

Not everything is broken...

Idea #1: Adversarial Training

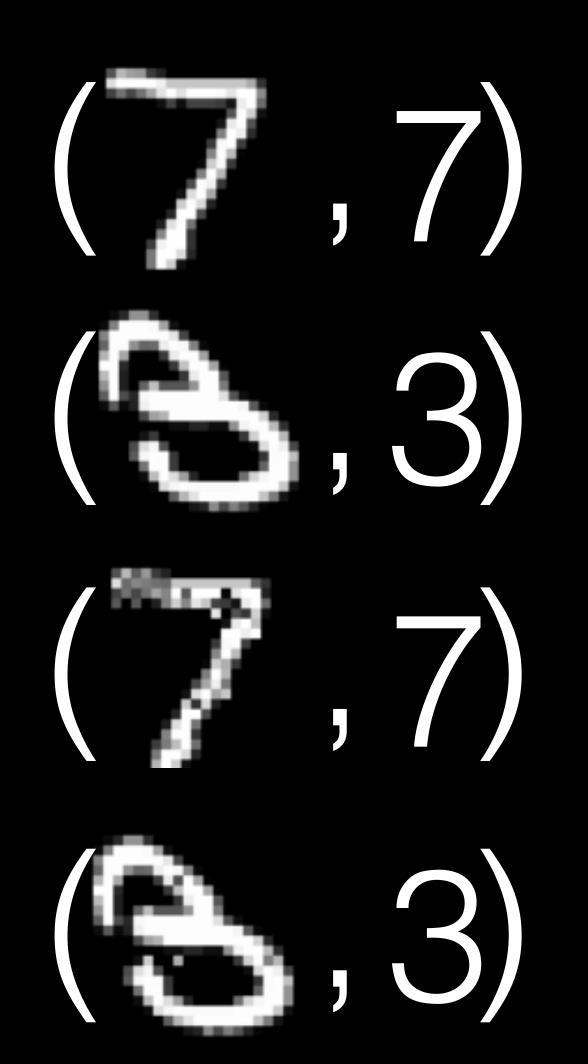
Normal Training



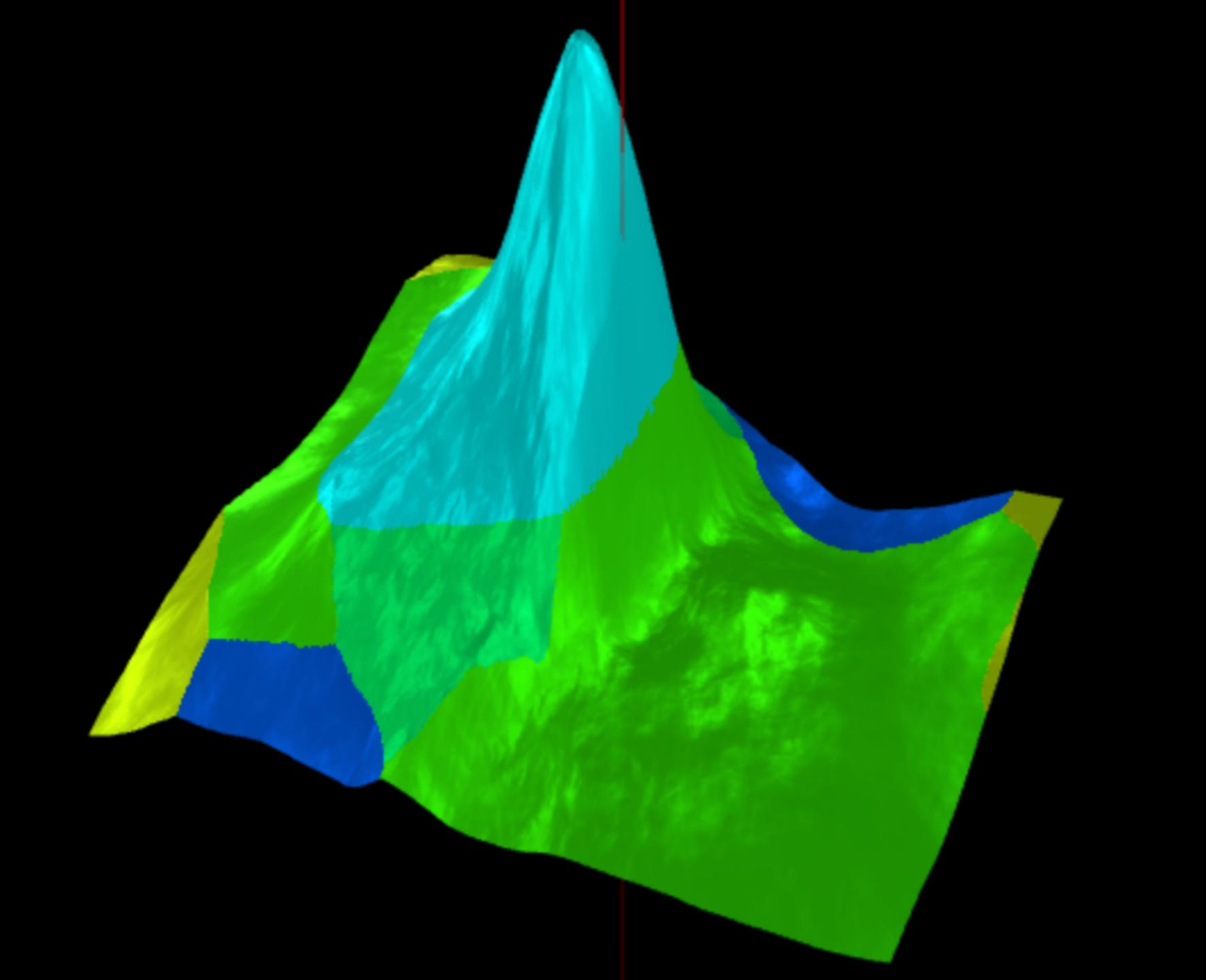
Adversarial Training (1)

Attack

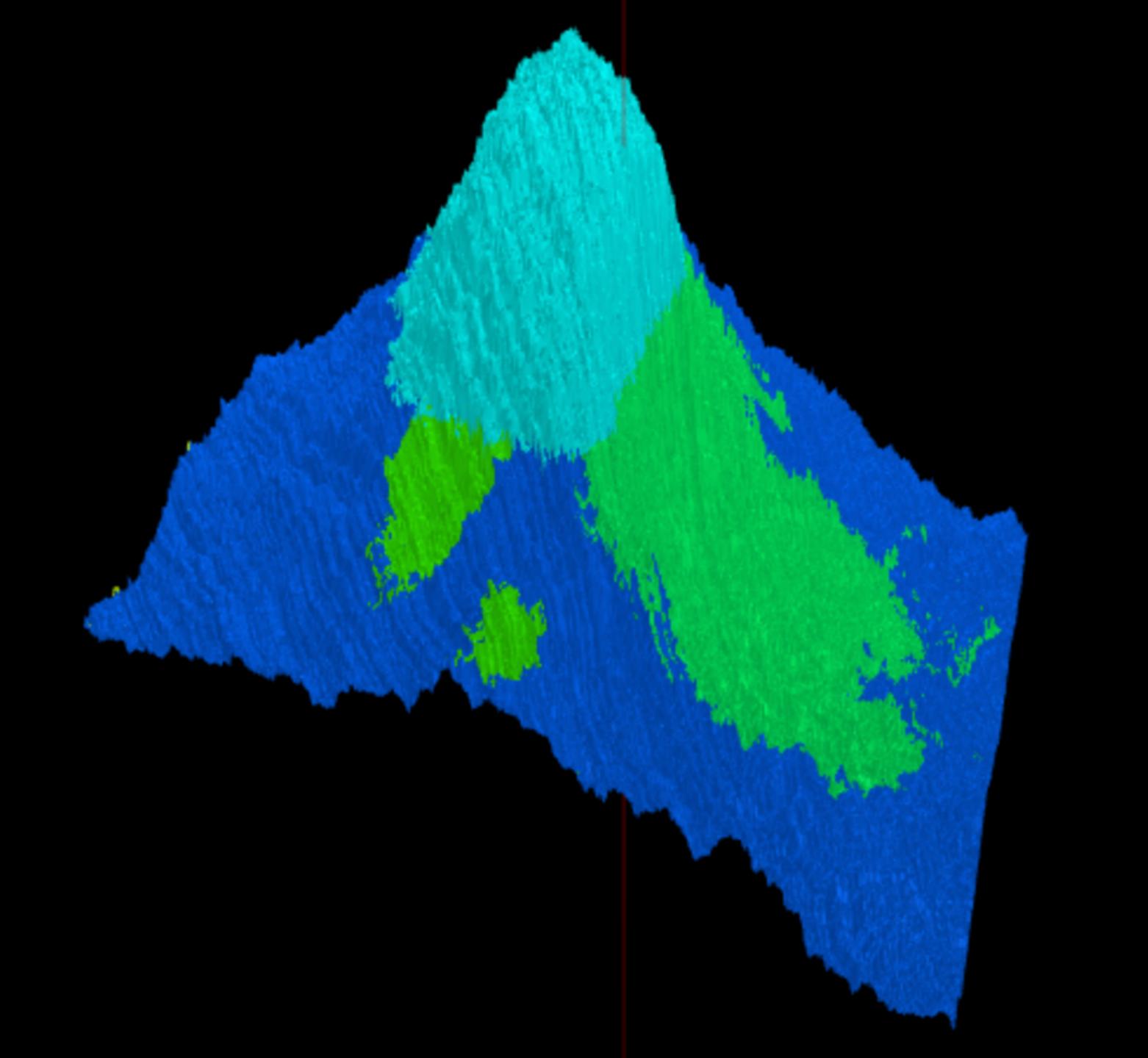
Adversarial Training (2)



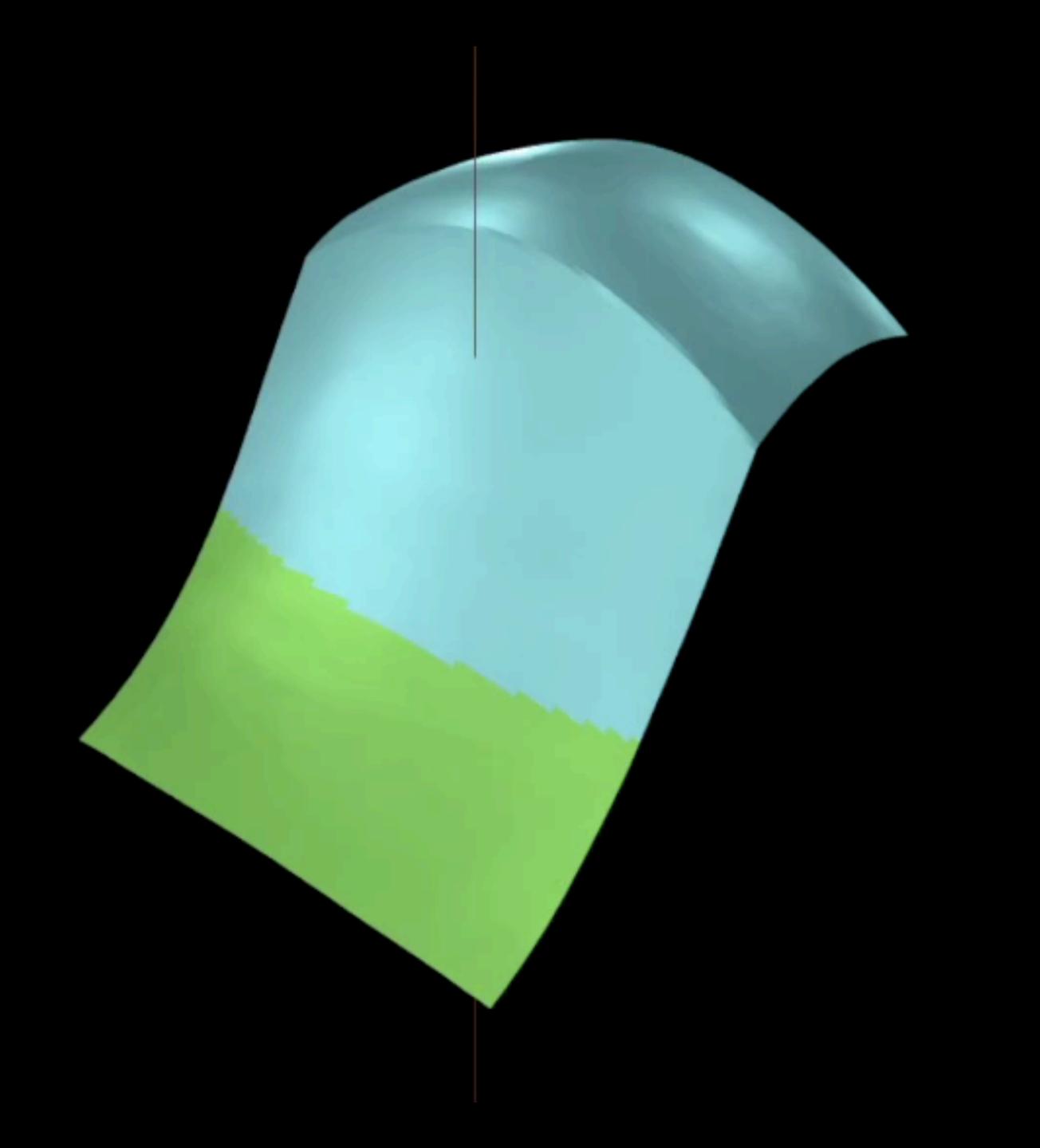
Training



Normal Loss Surface

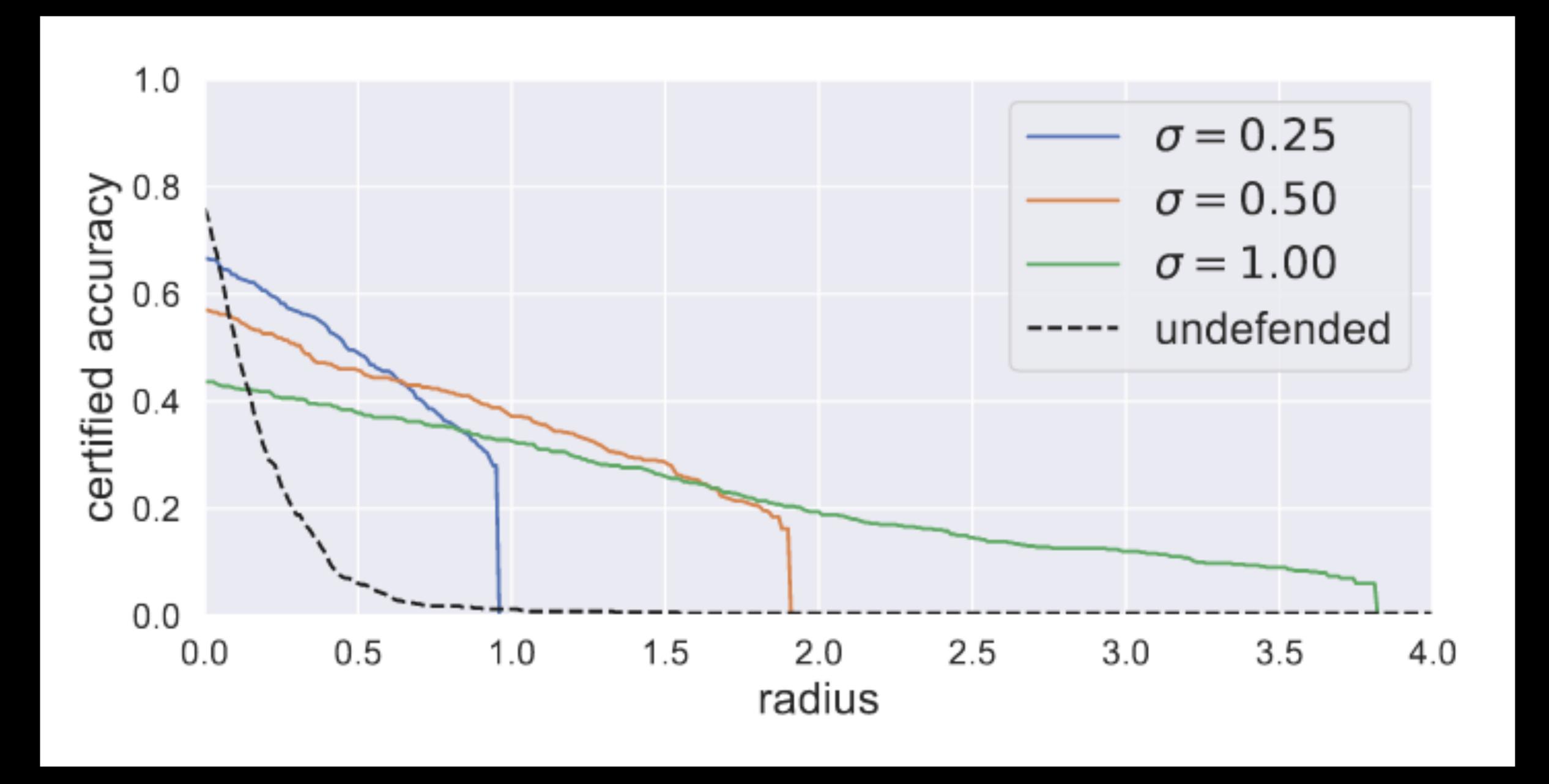


Obfuscated Loss Surface



Adversarial Training Loss Surface

Idea #2: Certified Defenses



Lecuyer et al. "Certified Robustness to Adversarial Examples with Differential Privacy" Cohen et al. "Certified Adversarial Robustness via Randomized Smoothing"

What's

next?

Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Abst

We sh more unpro

they sive d to mal versar

In t tillatio ificati ample

inputs attack magni two of achiev

We as sarial briefly

layer $F(\theta, x)$ cation proba $C(\theta, x)$ tion o we ar

greyso

ABSTRAC

Neural netwo ples: inputs that In order to bet survey ten rec pare their effic new loss funct nificantly har properties bel fact not. Finally

future propose

1 INTRO

Recent years h Neural netwo driving force b been demonstr [38], to beatin

cars [6]. In this pape sification. Wh learning appro who attempts x, an adversar has a different

adversarial ex nearly all dom The researc proposing mar

> ples correctly these defenses correctly. Due to this

detect them in seven papers pare their effi-With new atta evaded by an a datasets, the a

Permission to ma classroom use is g for profit or comm on the first page. author(s) must be republish, to post and/or a fee. Requ AISec'17, November © 2017 Copyright tion for Computin ACM ISBN 978-1-

https://doi.org/10.

MagNet and posed as a de we can cons

1 Introd

defenses with

Abstract

It is an oper they will be cently, three networks rob

MagNe

adversa

the para

rameter

neural r

amples

during

function

but clair

An efficiency

teristic be and for ea MagNe dients we neural n to overco certified v through we find of ial exar rence, wit to lie o gradients. rectly, v vent 6 cor the data classific threat mo the whi

1. Introduct

We identi

ent maskii

sense of s

examples.

gradients

based atta

effect can

In response to sarial example there has been fenses to increa progress has b against adversa the adversary solution has no

As benchmark tacks (e.g., Kı

 Adversa Carlini & Wag

MagNet and "Efficient Defenses Against Adversaria are Not Robust to Adversarial Examples

> **Obfuscated Gradients Give a False Ser** Circumventing Defenses to Adversar

On the Robustness of the CVPR 2018 V

Is AmI (A Robust

adversarial two white-2018 and fir existing tec of the defer

Neural nety

1. Introduction

Training neural sarial examples (Two defenses that this problem: "I Deflection" (Pral versarial Attacks Denoiser" (Liao

In this note, we in the white-box examples that re ImageNet datas a small ℓ_{∞} pert considered in the A. Evaluation

Abstract—No.

I. ATTACKING "ATTACKS MEET INTE

AmI (Attacks meet Interpretability) is an defense [3] to detect [1] adversarial exa recognition models. By applying interpretato a pre-trained neural network, AmI ide neurons. It then creates a second augmen with the same parameters but increases the of important neurons. AmI rejects inputs and augmented neural network disagree.

We find that this defense (presented at a a spotlight paper—the top 3% of submiss ineffective, and even defense-oblivious1 detection rate to 0% on untargeted attacks. more robust to untargeted attacks than the network. Figure 1 contains examples of a that fool the AmI defense. We are incred authors for releasing their source code² w We hope that future work will continue to by publication time to accelerate progress

Comment on Biologically inspired protection of deep networks from adversarial attacks

> ON THE LIMITATION OF LOCAL INTRINSIC DIMEN-SIONALITY FOR CHARACTERIZING THE CHRODACES OF

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

A rece gradient-l highly sat the attack stabilisati has yet to limitation

¹ Wernε

Evaluat between ne robust beca for gradien in cases wi attacks tha of the grad

A recen against gra likely a side stable comp of the grad In a fir Perceptron layer activa sigmoid, ze network for the MLP w attack with verified tha S1.

In high exactly zer directly wi elements of

> Training Vanilla

Table 1: A naive appli FGSM base

Saturated

Paper Type: Appraisal Paper of Existing Method Cory Cornelius

cory.cornelius@intel.com

Nilaksh Das nilakshdas@gatech.edu

The Efficacy of Shield under Different Threat Models

Shang-Tse Chen schen351@gatech.edu

Evaluating and Understanding the Robustness of Adversarial Logit Pairing

Andrew Ilyas*

Massachusetts Institute of Technology

{engstrom,ailyas,aathalye}@mit.edu

Abstract

We evaluate the robustness of Adversarial Logit Pairing, a recently proposed de-

fense 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.

Anish Athalye*

ABSTRA In this appr

compression ial attacks o at KDD 201 studied in t adversary is pre-process used in the threat and e degree of in full white-bo original wo an adaptive of the Project gradient-bas learning res

In r

exa

of a

effo

the

and

dra

(20)

sic tions to

. Intr

Th

pro

ple

mo

ach

the

me

to

mo

as

dec

def

lati

ing and ing prob speech game pl able suc properti Researc

percepti

Deep lea ensemble fro pre-trained targeted PGl Shield ense

48.9% if the instead of be tremely 2017) F

ensemble w tion in the co scratch are l

gray-box sc

ACM Refere

1 Contributions

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

Logan Engstrom*

 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

 $\{\mathbf{k}$

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

Cryptanalysis of some recently-proposed multiple modes of operation

Differential cryptanalysis of KHF

1						
As					Cryptanalysis of TWOPRIME	1 I:
nic						
pa					Don Coppersmith ¹ , David Wagner ² , Bruce Schneier ³ , and J	In Fin
de					Don copperation, David Wagner, Dideo genierer, and y	One sp of rou
ce	1 I				¹ IBM Research, e-mail: copper@watson.ibm.com	hood,
aff					² U.C. Berkeley, e-mail: daw@cs.berkeley.edu	based
lat					Counterpane Systems, e-mail: {schneier,kelsey}@counter	Or
ar	Relate					Boolea
se	tain p					able t
	derive				Albertanet Discout I [DNDCo7]	founda
$_{\mathrm{pr}}$	how t		1		Abstract. Ding et al [DNRS97] propose a stream generator	weakn Th
\mathbf{m} i	differe				several layers. We present several attacks. First, we observe non-surjectivity of a linear combination step allows us to re-	we dis
for	the at		Rec		the key with minimal effort. Next, we show that the various	charac
ov	values		prin		insufficiently mixed by these layers, enabling an attack similar t	shift ϵ
$d\mathbf{r}$	\mathbf{R}		a hi		two-loop Vigenere ciphers to recover the remainder of the key.	appea
fra	do no		Safa		these techniques lets us recover the entire TWOPRIME key. V	charac In Sec
w€	witho	1	soft		the generator to produce 2^{33} blocks (2^{35} bytes), or 19 hours	gives (
$_{ m ne}$	know		$_{ m thei}$		output, of which we examine about one million blocks (2 ²³ l	find co
$_{\mathrm{pr}}$	again	DES	(computational workload can be estimated at 2 ²⁸ operations	attack
re	ator t	more	usin		set of attacks trades off texts for time, reducing the amount plaintext needed to just eight blocks (64 bytes), while needing	family
di_{i}	adver	bit ke	well		and 2 ³² space. We also show how to break two variants of TW	
	Hash	There	to b		presented in the original paper.	2 E
inc	attacl	for D	$_{\mathrm{Boo}}$			CDEE
45	$_{ m In}$	retain]			SPEE length
	showe	offers	[BS]	1	Introduction	lengen
	prese	В				

Cryptanalysis of SPEED

Cryptanalysis of FROG

Cryptanalysis of ORYX

D.

FROG

interna

Round

 $X_{0...15}$

*U.C

†Cou

[‡]Cou

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

1]

One (

is diff

many

are ty

obtai

terist

to jus

break

safe f

call tl

U

 \mathbf{T}

1 In

The de

the last

is easy

prevent

secure 1

cations

any cas

the last

as the (

Telecon

Americ

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)

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.

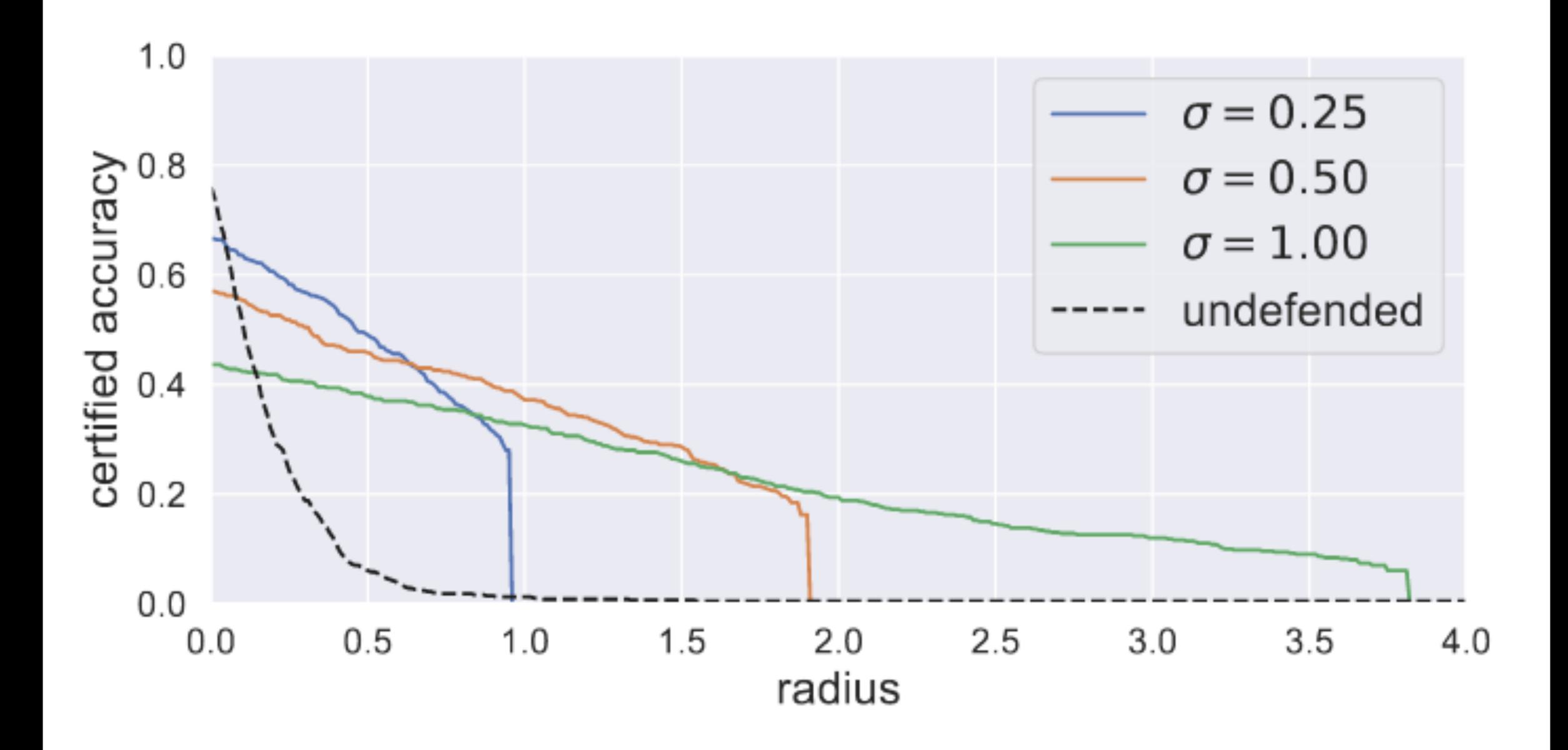
Keywords: block ciphers, bicliques, AES, key recovery, preimage

Are we crypto in the 90's?

Maybe not.

Three reasons.

Reason 1.



Attack Success Rates in Security

Attack Success Rates in Security

Crypto: 2-128

Attack Success Rates in Security

Crypto: 2-128, broken if 2-127

Crypto: 2-128, broken if 2-127

Systems: 2-32

Crypto: 2-128, broken if 2-127

Systems: 2-32, broken if 2-20

Crypto: 2-128, broken if 2-127

Systems: 2-32, broken if 2-20

Machine Learning:

Evans, "Is "adversarial example" an adversarial example?"

Crypto: 2-128, broken if 2-127

Systems: 2-32, broken if 2-20

Machine Learning: 2-1

Evans, "Is "adversarial example" an adversarial example?"

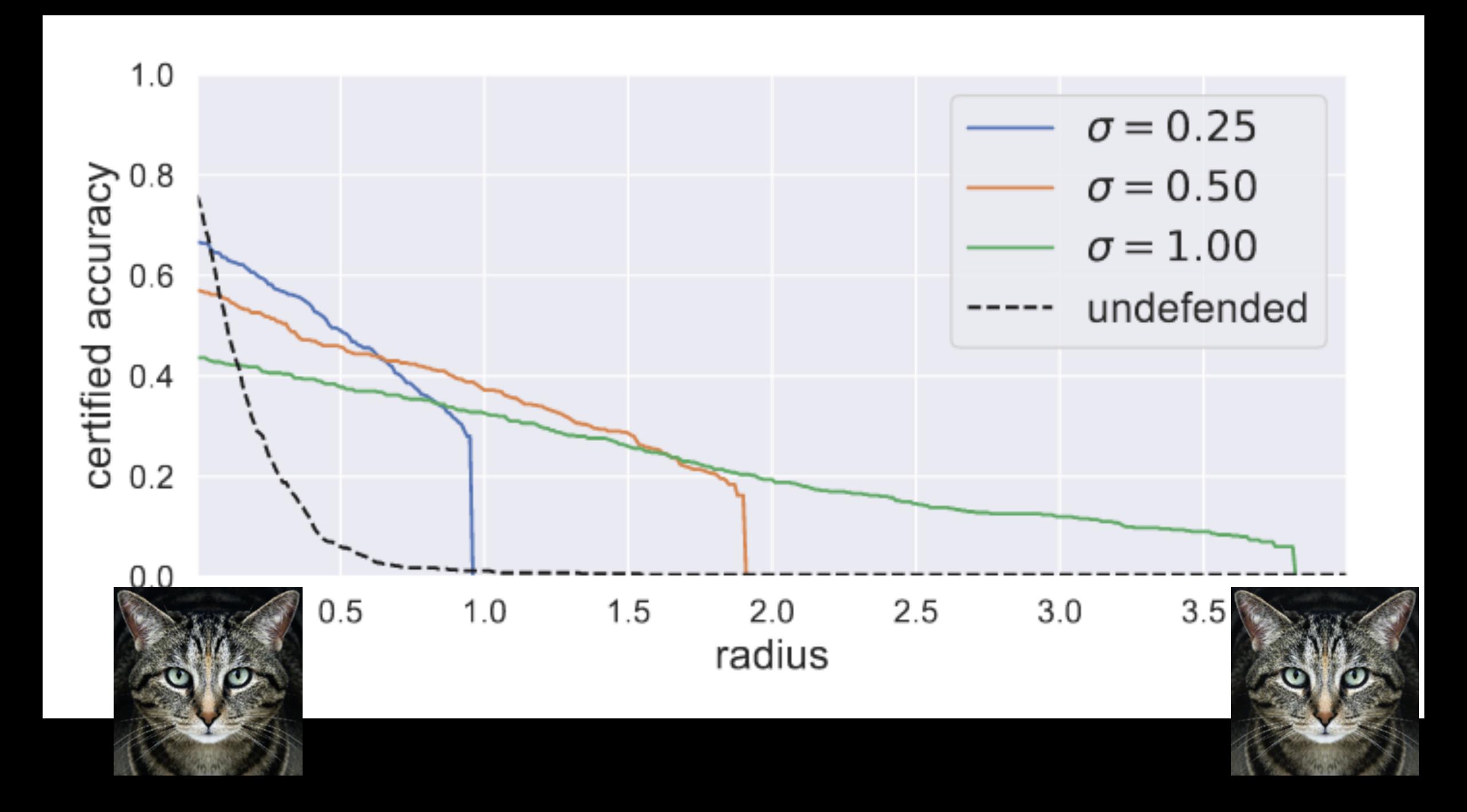
Crypto: 2-128, broken if 2-127

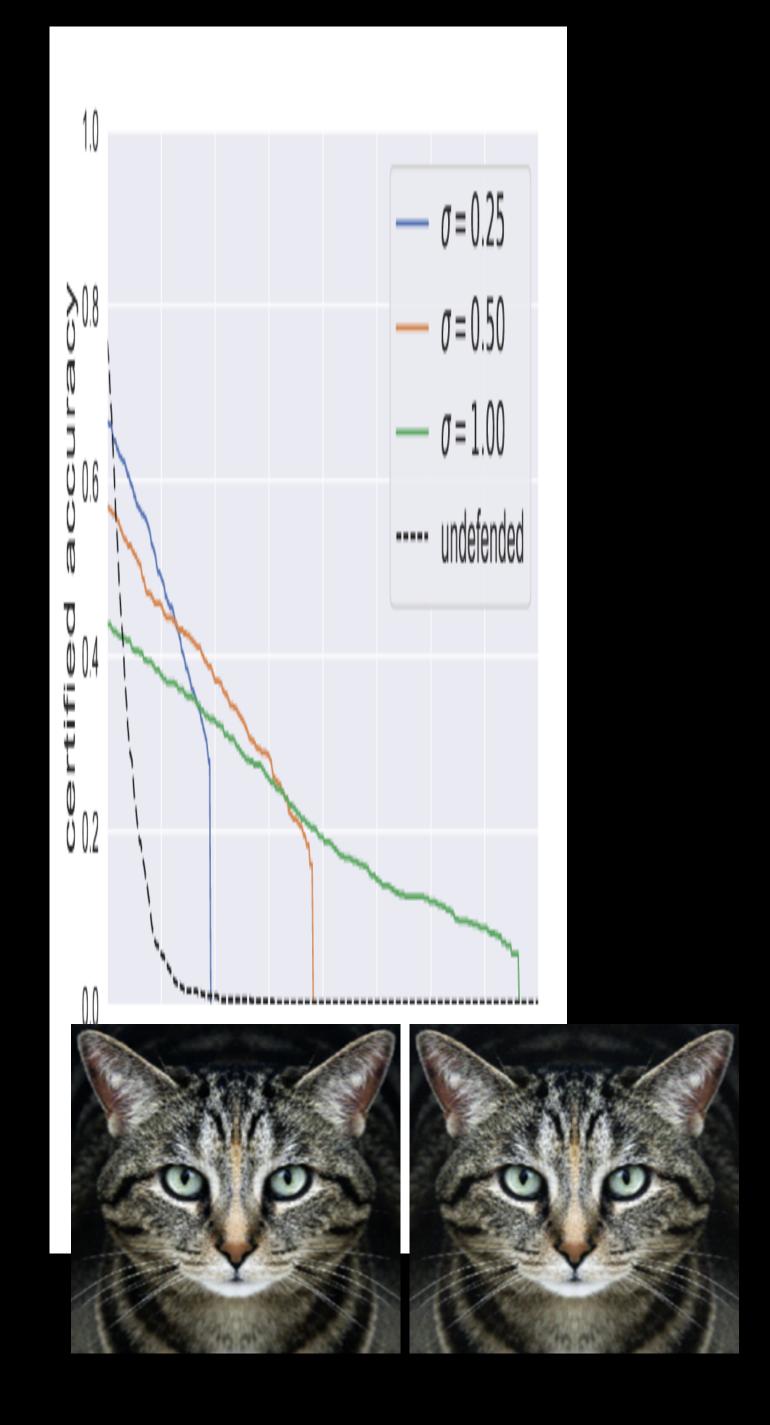
Systems: 2-32, broken if 2-20

Machine Learning: 2⁻¹, broken if 2⁰

Evans, "Is "adversarial example" an adversarial example?"

Reason 2.





 $L_2 = 100$

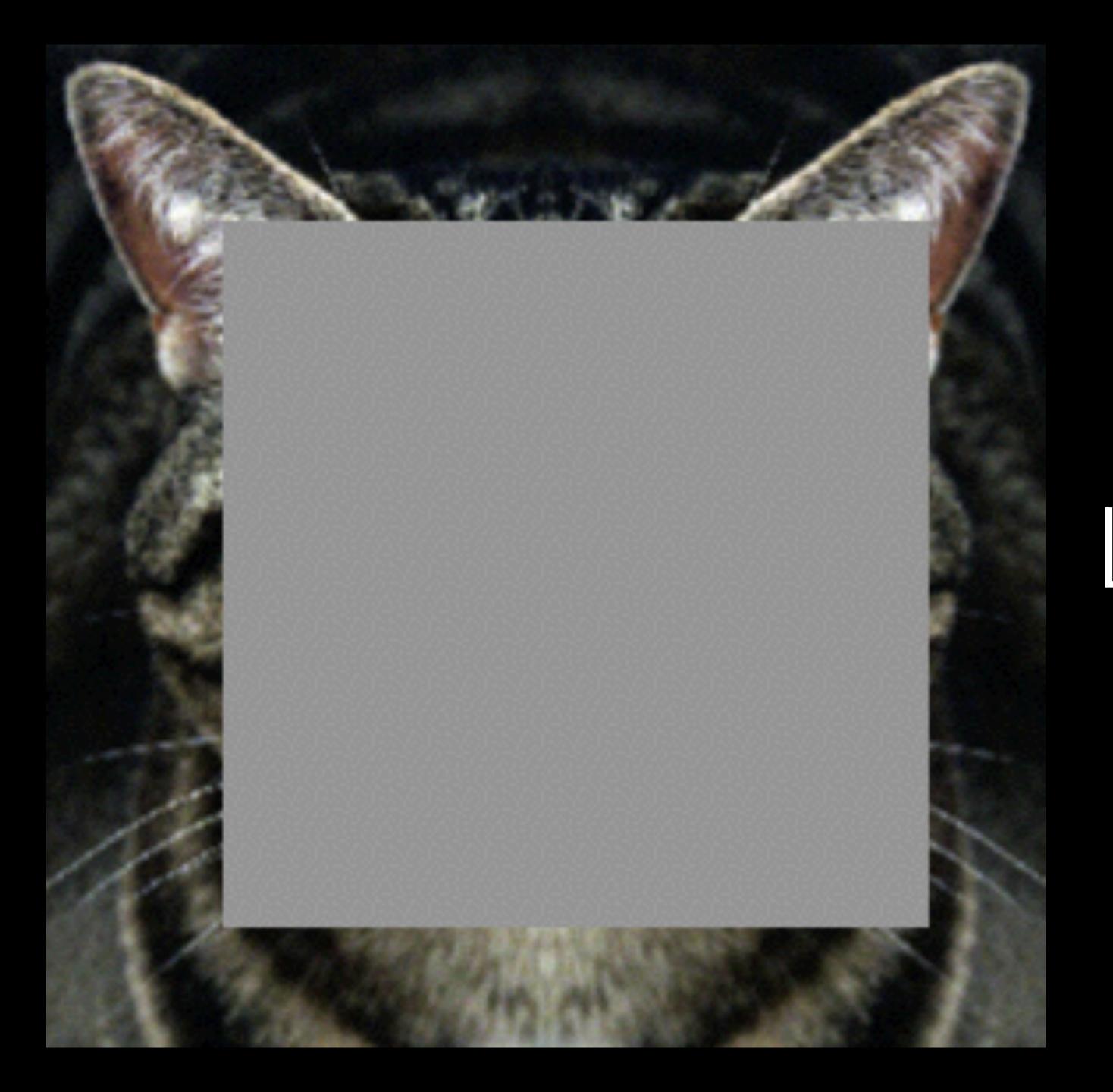




Original



L2 distortion: 75



L2 distortion: 75

Claim: We are crypto **pre-**Shannon

Reason 3.

It's not just adversarial shifts ...

Do ImageNet Classifiers Generalize to ImageNet?

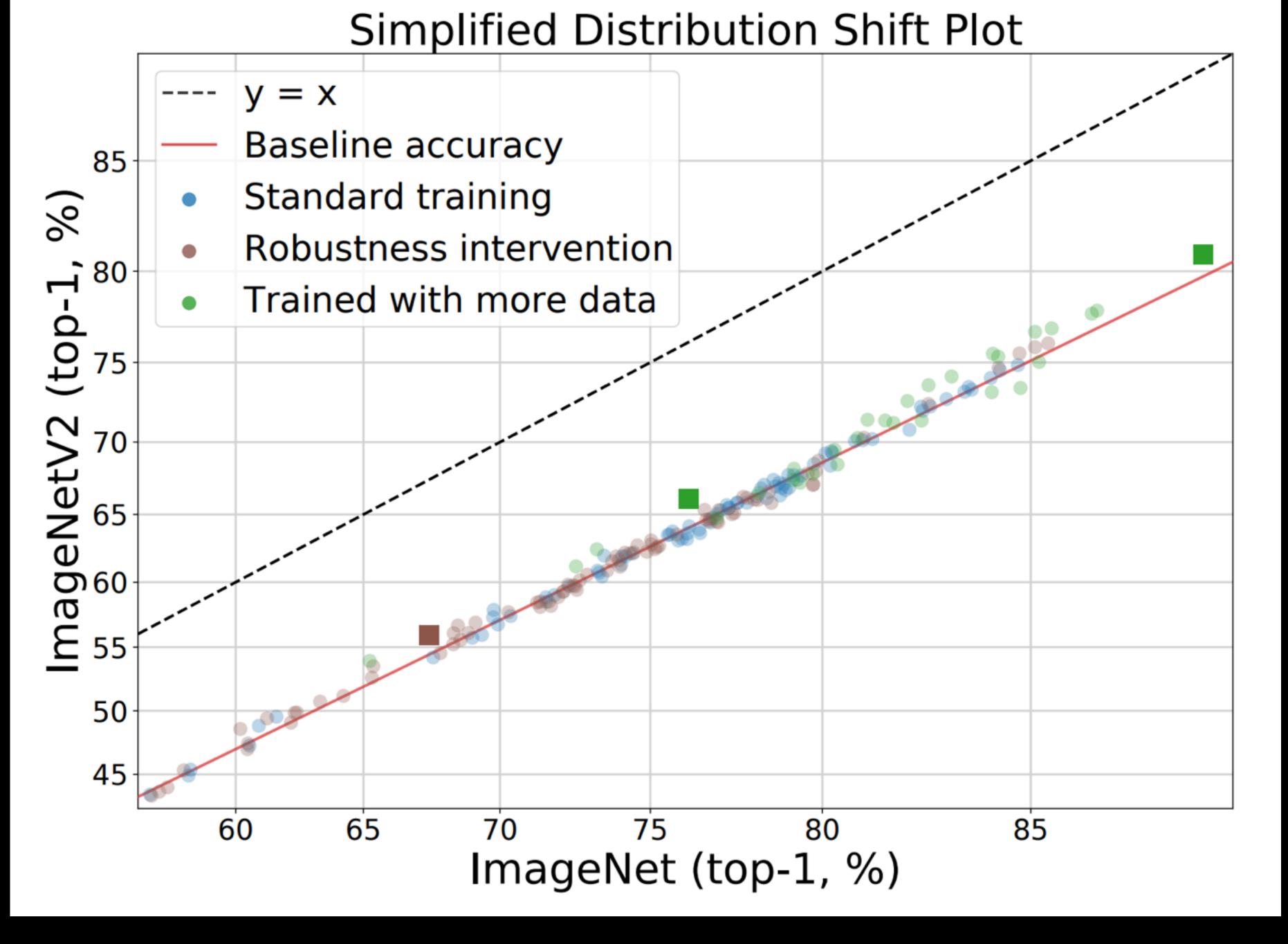
Benjamin Recht*
UC Berkeley

Rebecca Roelofs UC Berkeley Ludwig Schmidt UC Berkeley

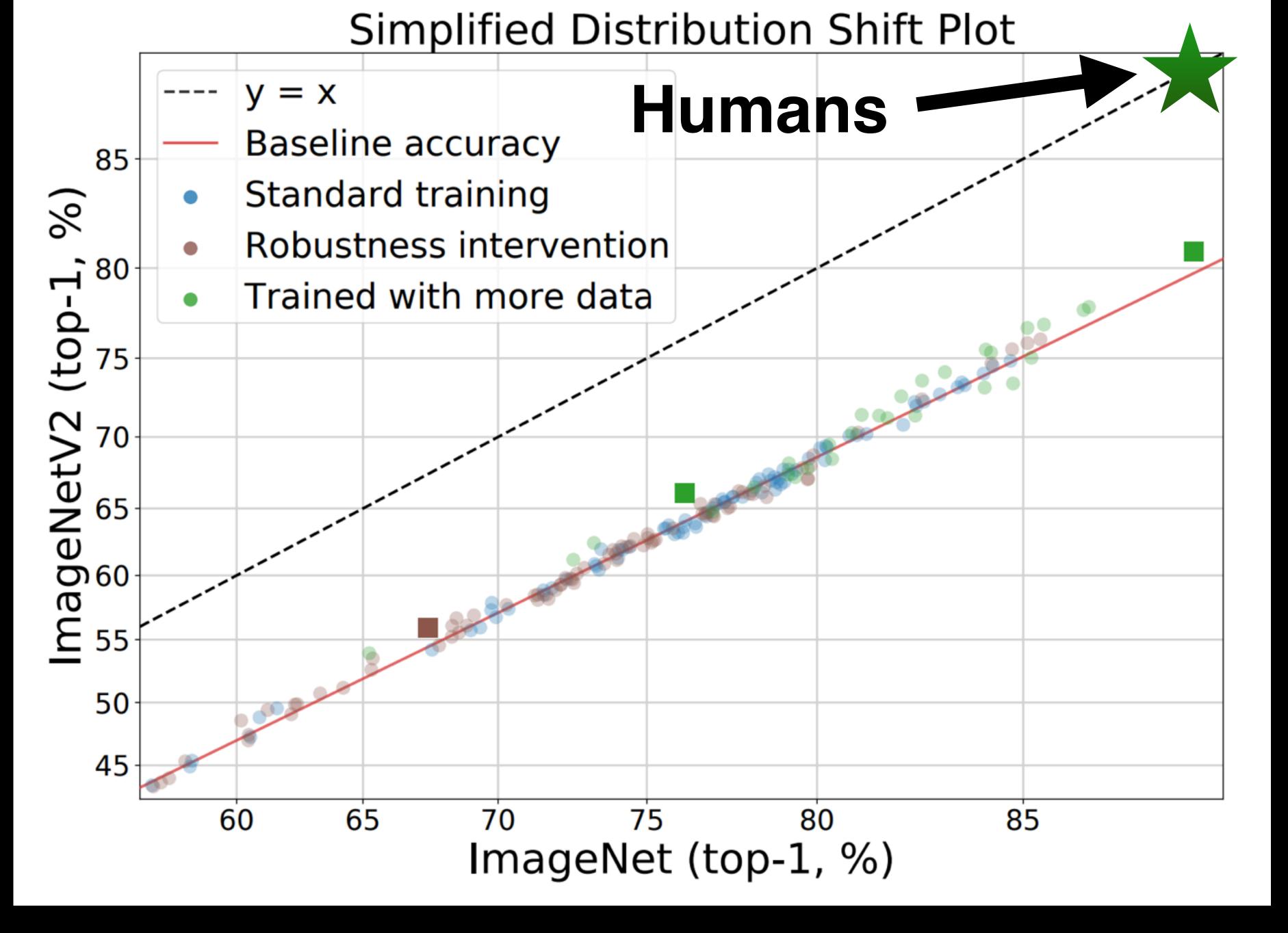
Vaishaal Shankar UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.



Taori et al., "Measuring Robustness to Natural Distribution Shifts in Image Classification"



Taori et al., "Measuring Robustness to Natural Distribution Shifts in Image Classification"

Conclusion

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

We still have a long way to go.

nicholas@carlini.com

https://nicholas.carlini.com