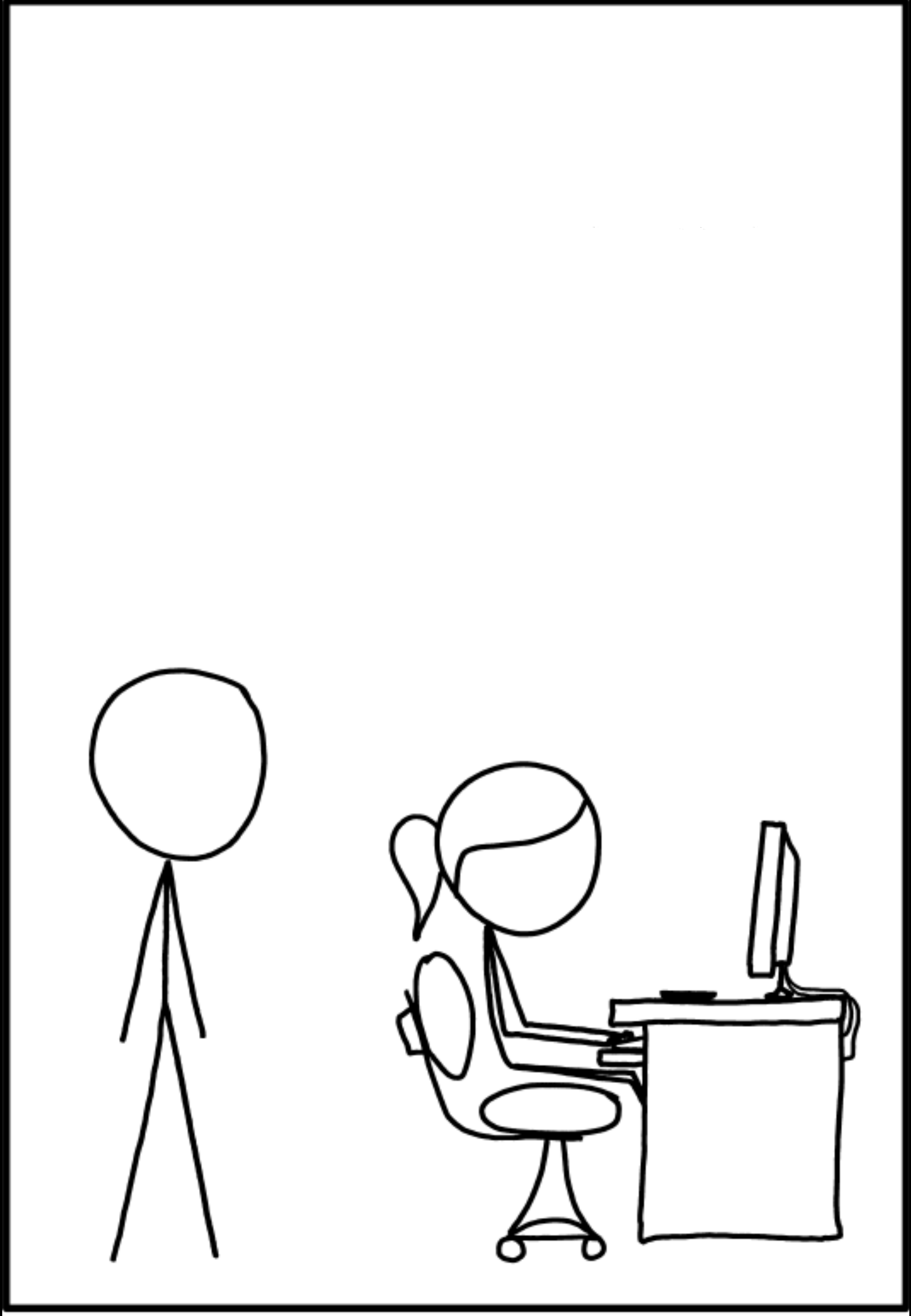
On the (In-)Security of Machine Learning

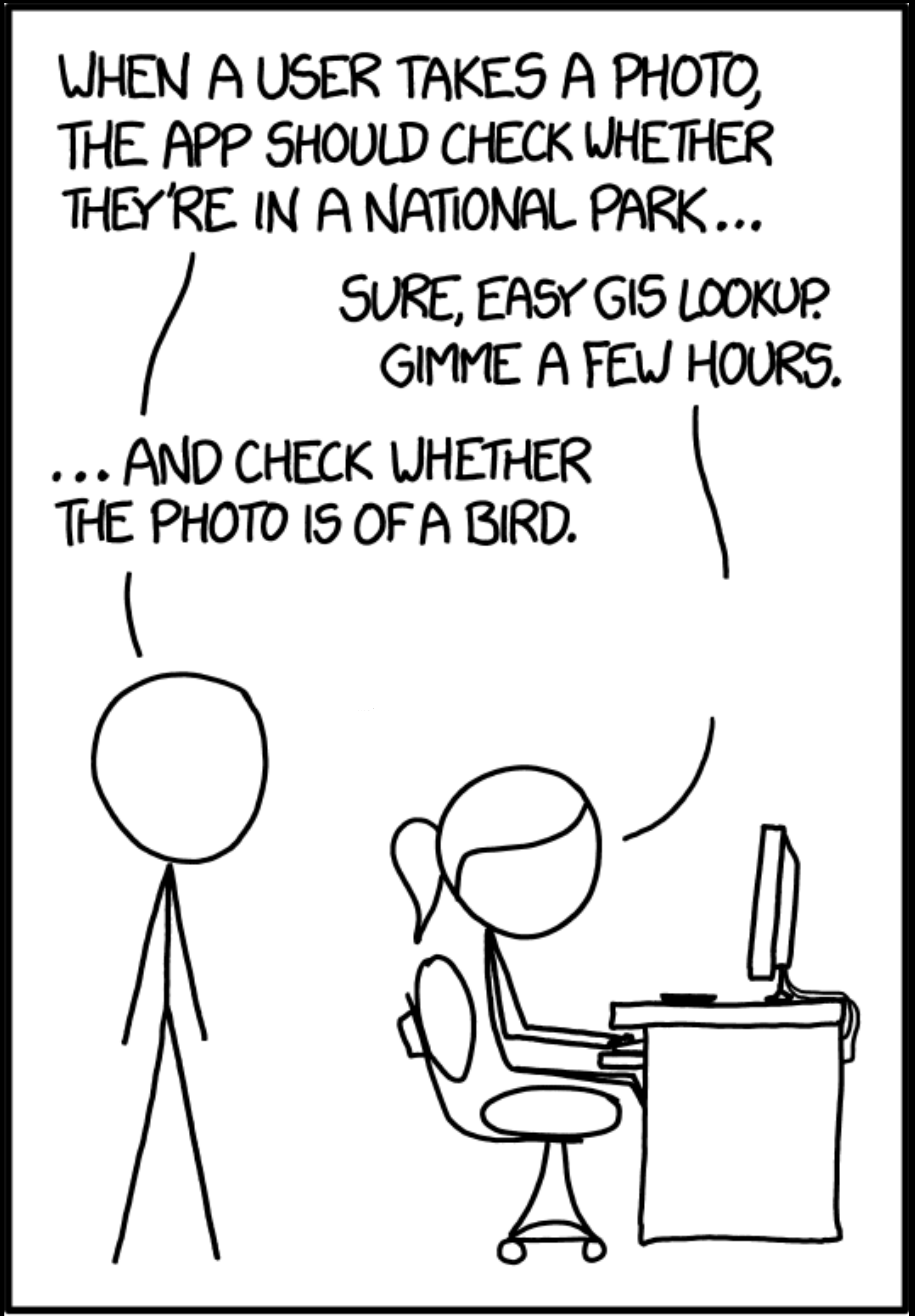
Nicholas Carlini
Google Brain



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

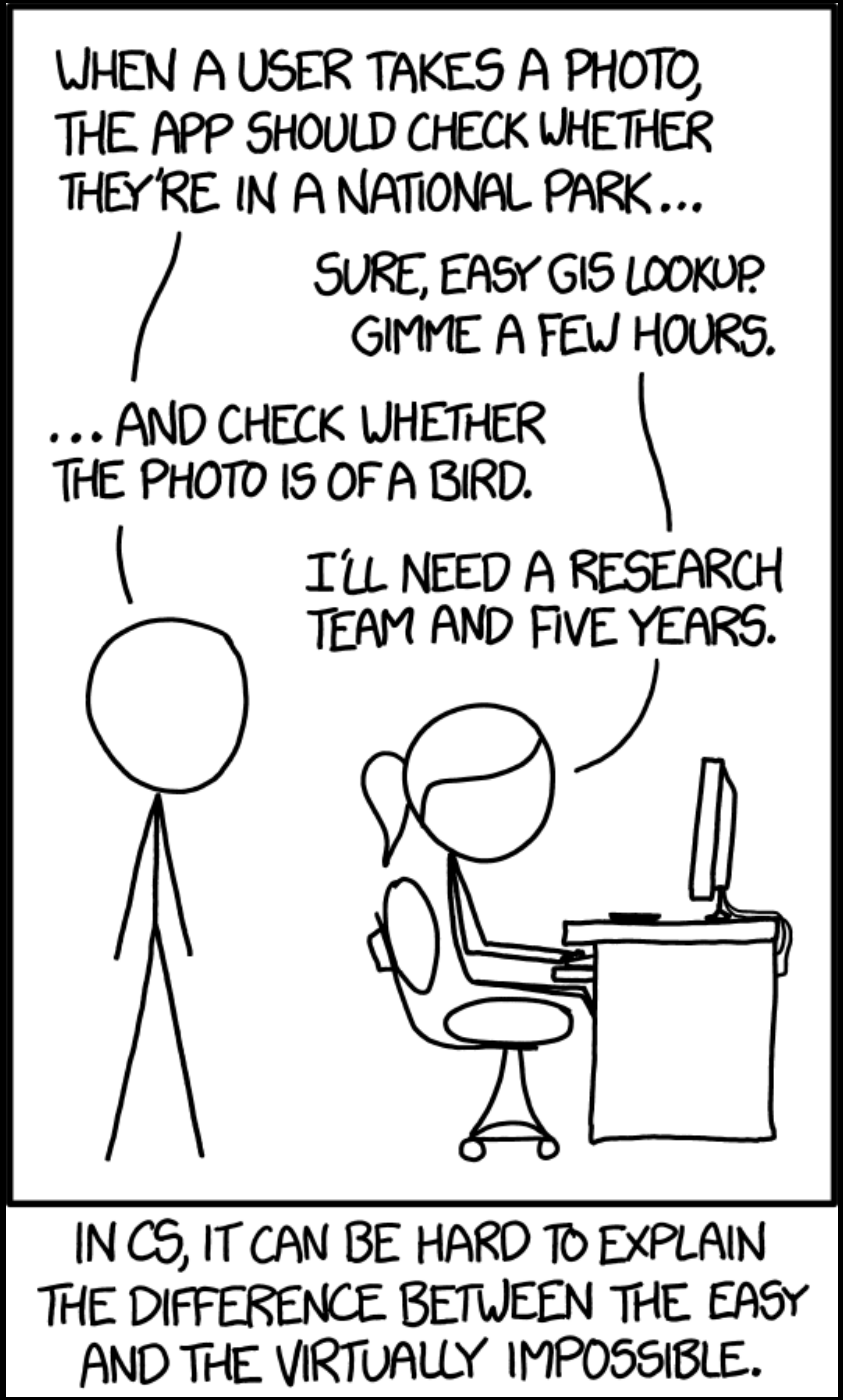
IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

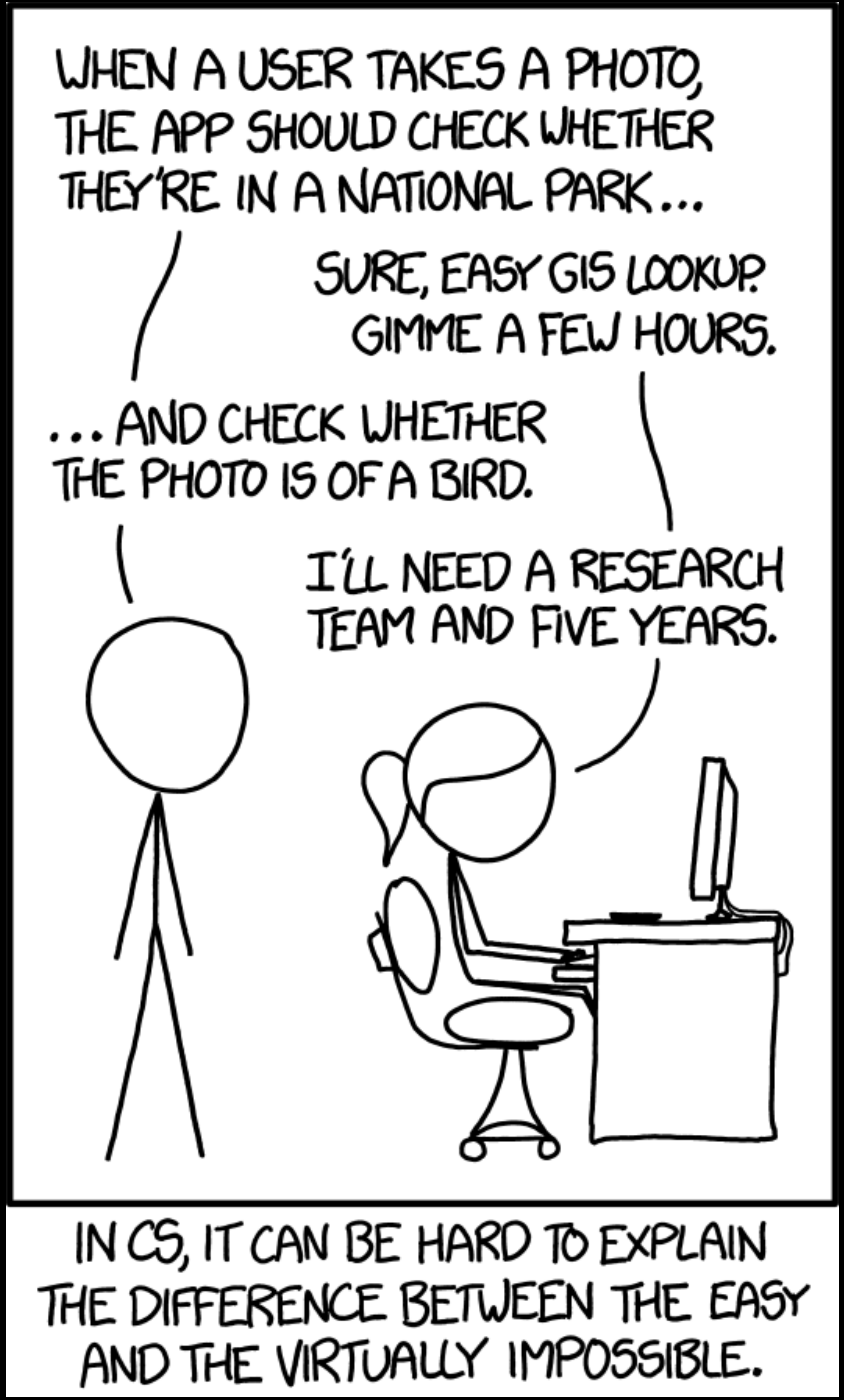


IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.
IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

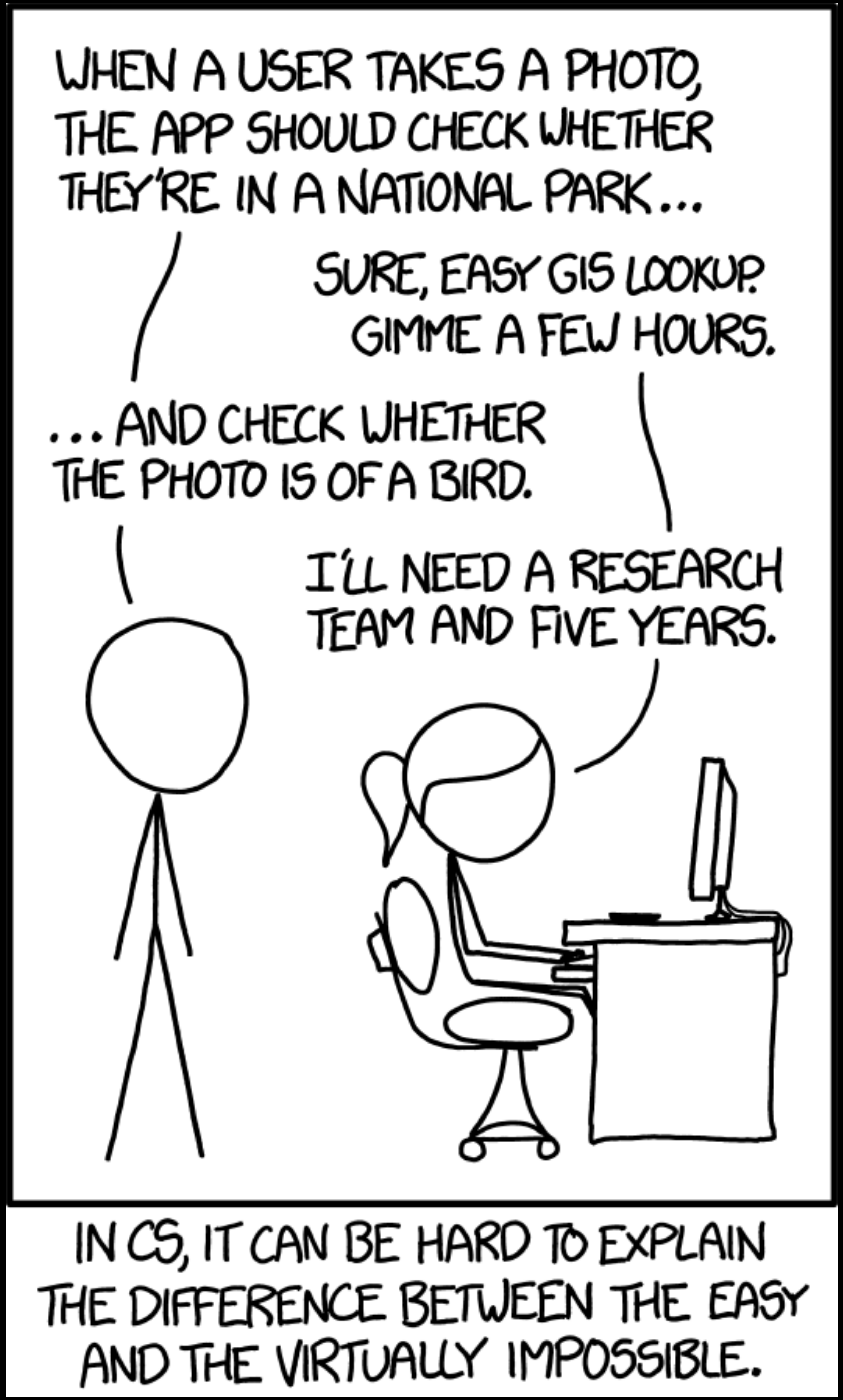


Written: Sept 24, 2014



Written: **Sept 24, 2014**

Today: **Oct 16, 2018**



Written: **Sept 24, 2014**

Today: **Oct 16, 2018**

... 4 years ago

So how are we doing?

95% it is a

French Bulldog

83% it is a

Old English Sheepdog

78% it is a

Greater Swiss Mountain Dog

67% it is a

Great Dane

99.99% it is

Guacamole

96% it is a

Golden Retriever

99.99% it is

Guacamole

This phenomenon is known as an
adversarial example

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. 2013.
C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. ICLR 2014.
I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 2014.

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

(a) (b) (c)

Why should we care about adversarial examples?

Make ML
robust

Make ML
better

How do we generate adversarial examples?

DEFN: The **loss** of a neural network on an input **x** for a label **y** is a measure of how *wrong* the network is on **x**.

loss( , dog) is small

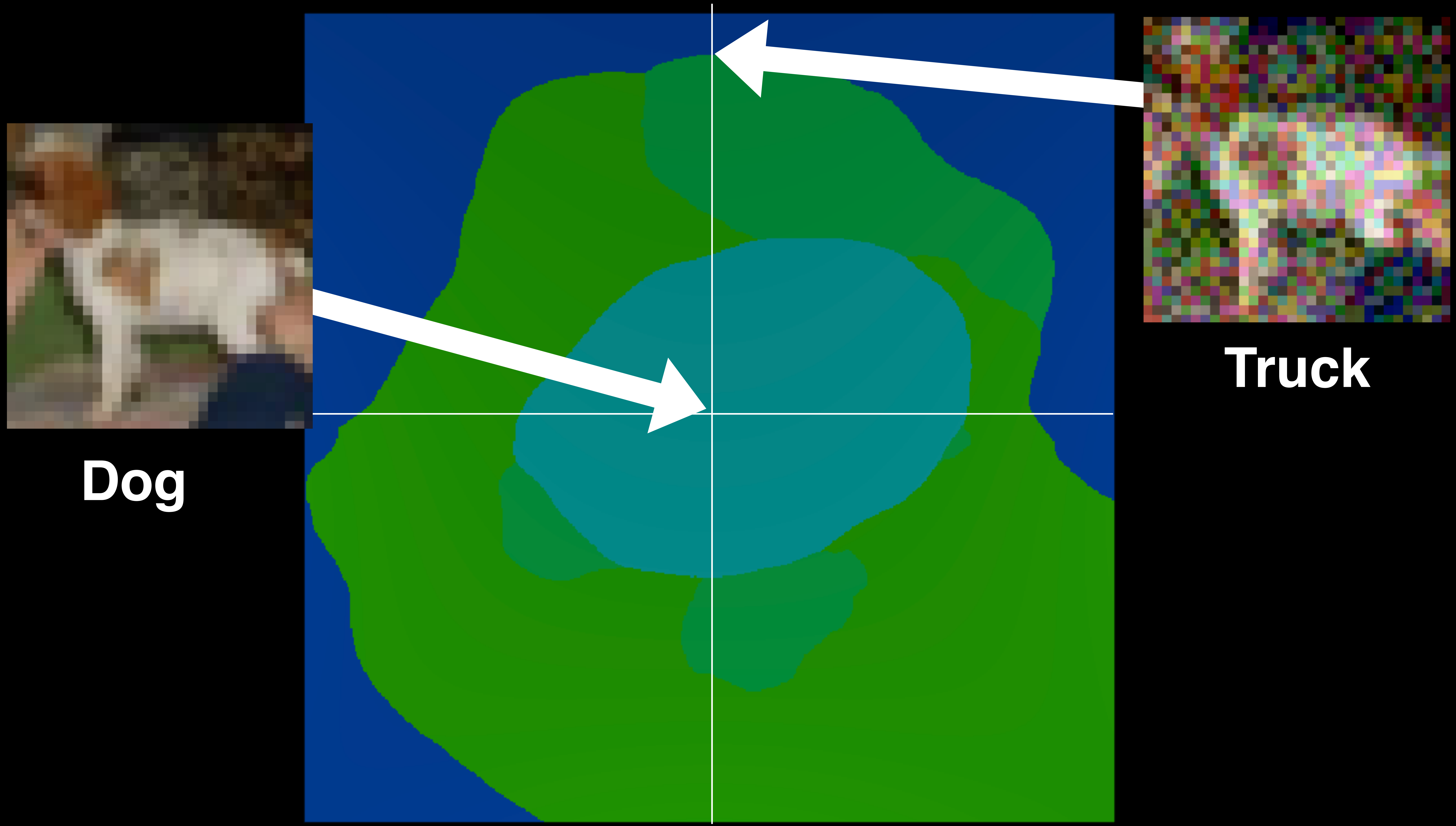
loss( , guacamole) is large

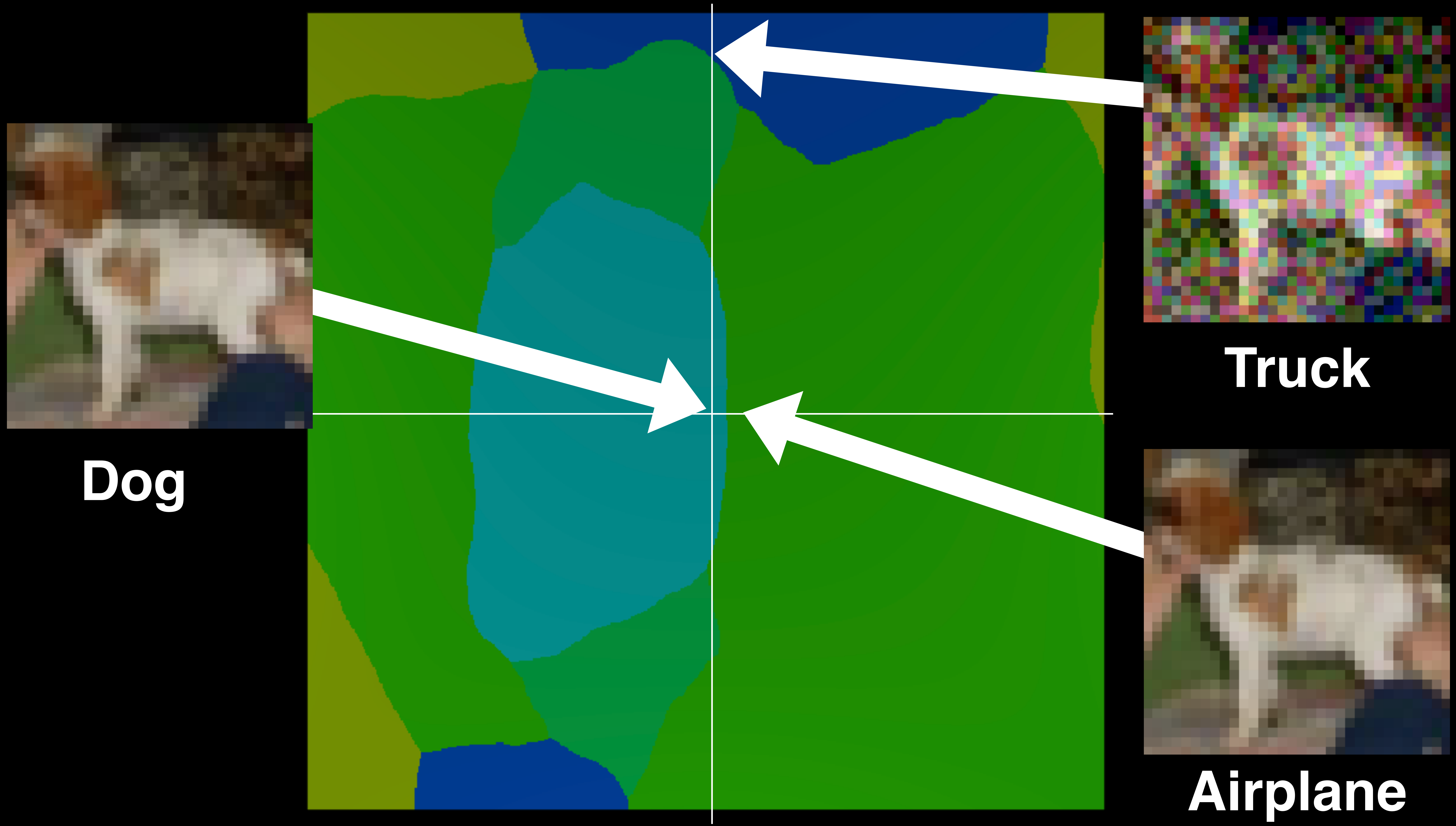
MAXIMIZE	neural network loss on the given input
SUCH THAT	the perturbation is less than a given threshold

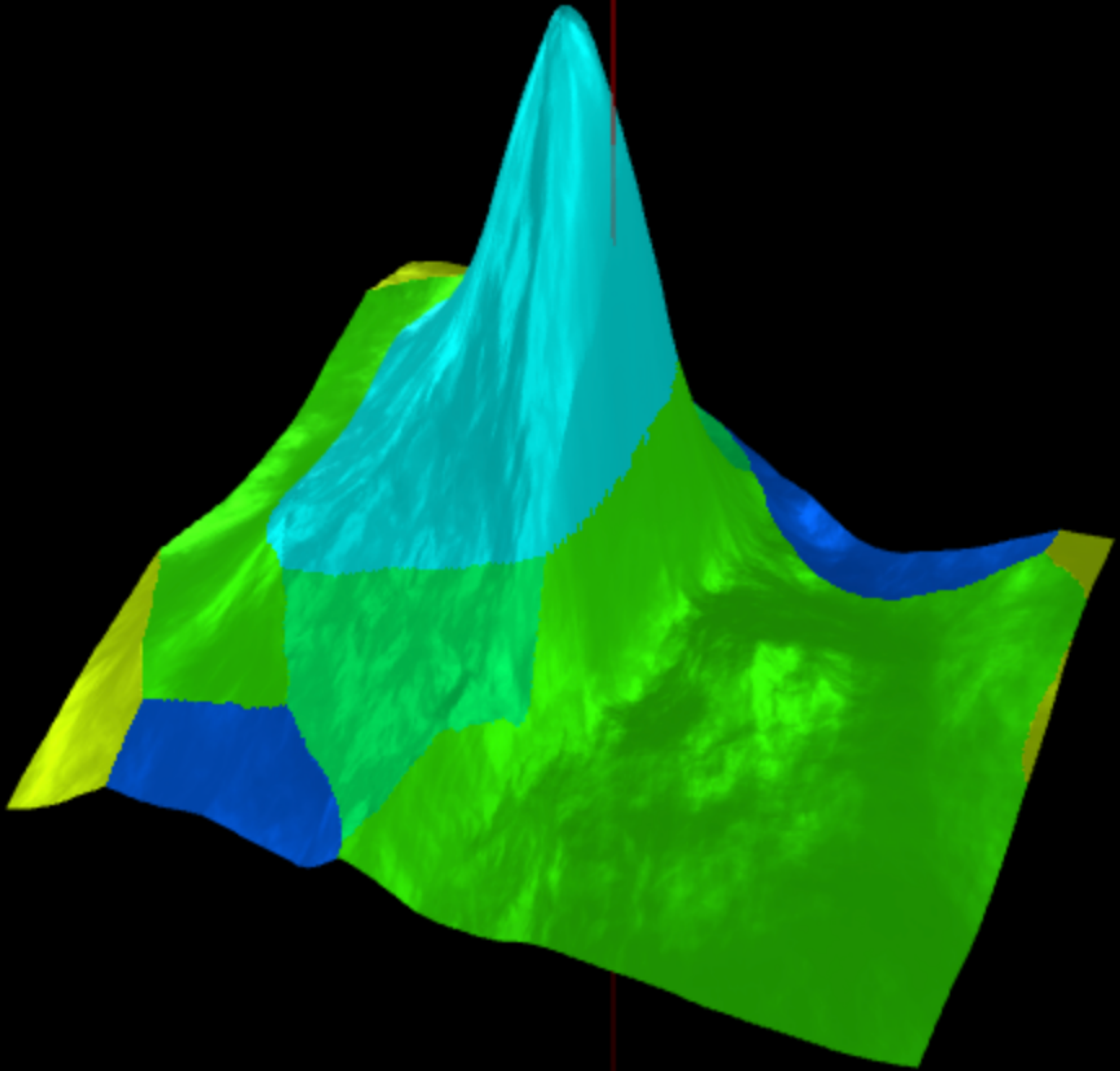
What do we need to know?

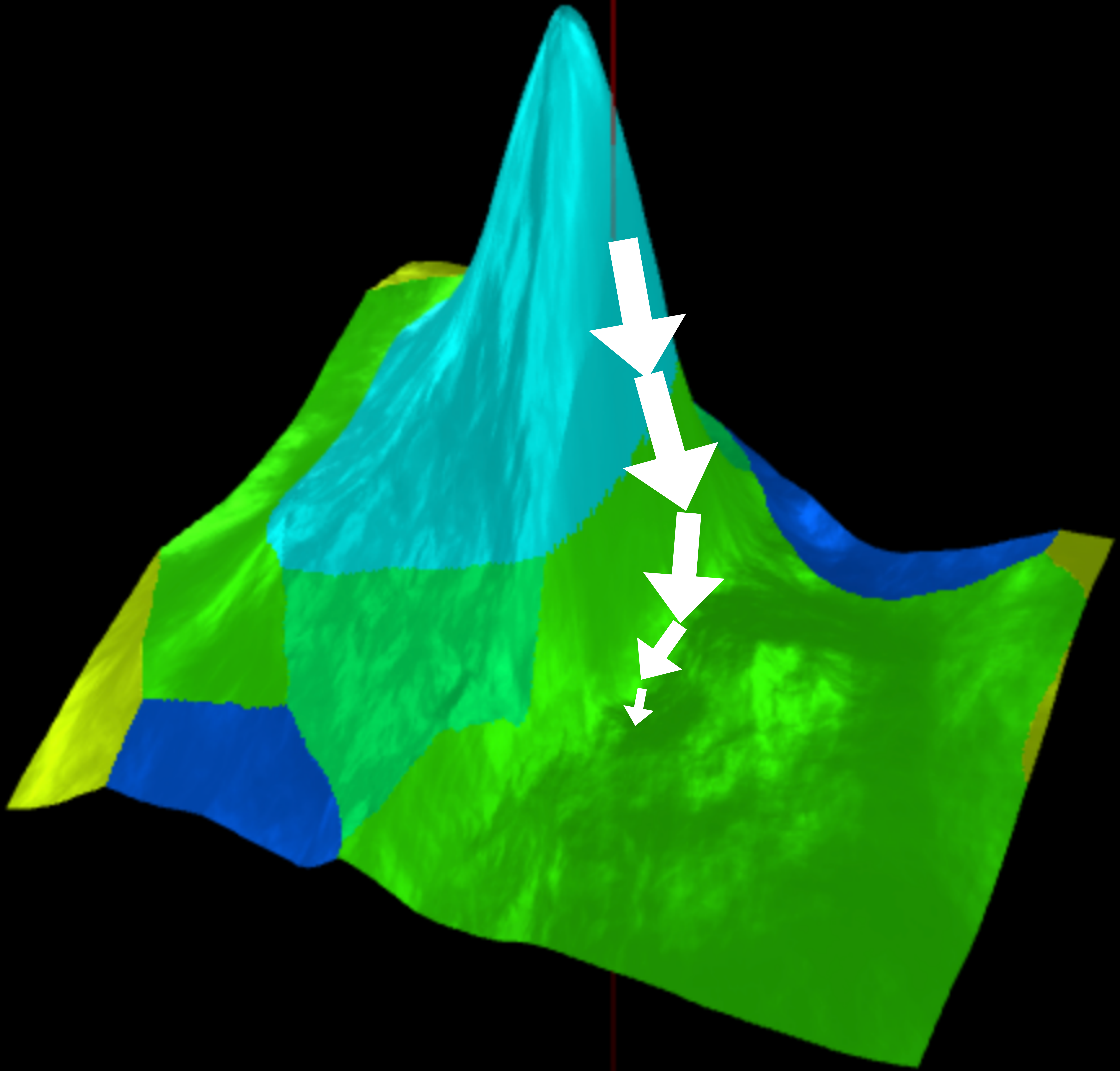
Everything.

WHY does this work?


Dog
Truck


Truck
Dog
Airplane





Okay, lesson learned.

Okay, lesson learned.

Don't classify dogs with neural networks.

99.99% it is a

School Bus

Okay, lesson learned.

Okay, lesson learned.

Don't classify ~~dogs~~ images with neural networks.


And now for something
completely different

Mozilla's DeepSpeech

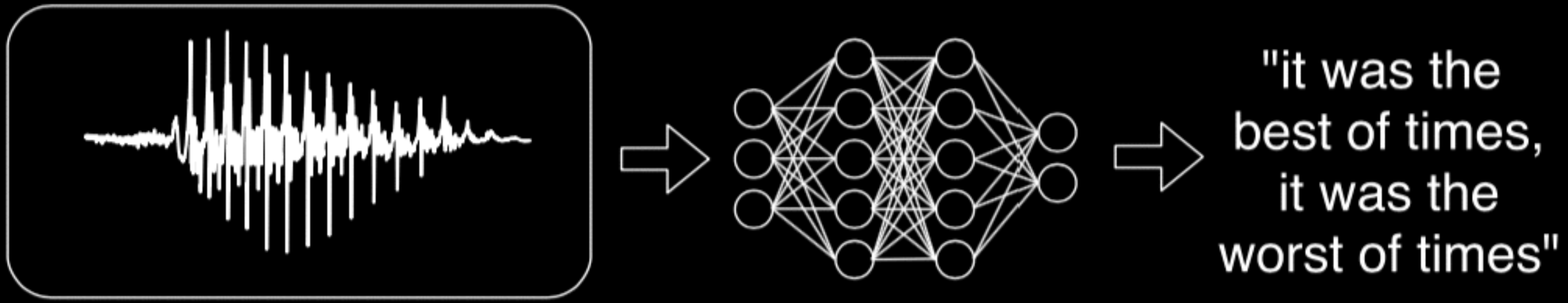
Mozilla's DeepSpeech
transcribes this

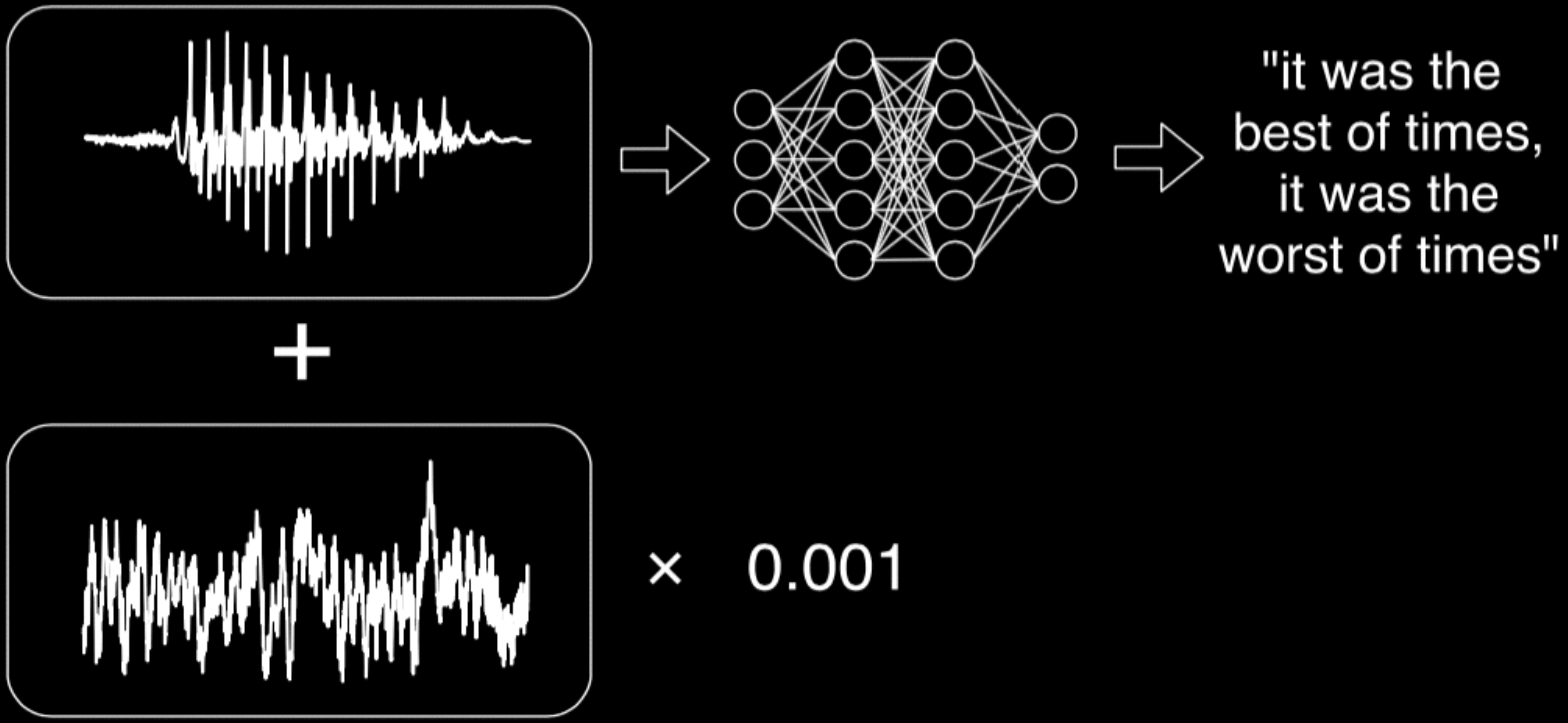
Mozilla's DeepSpeech
transcribes this as

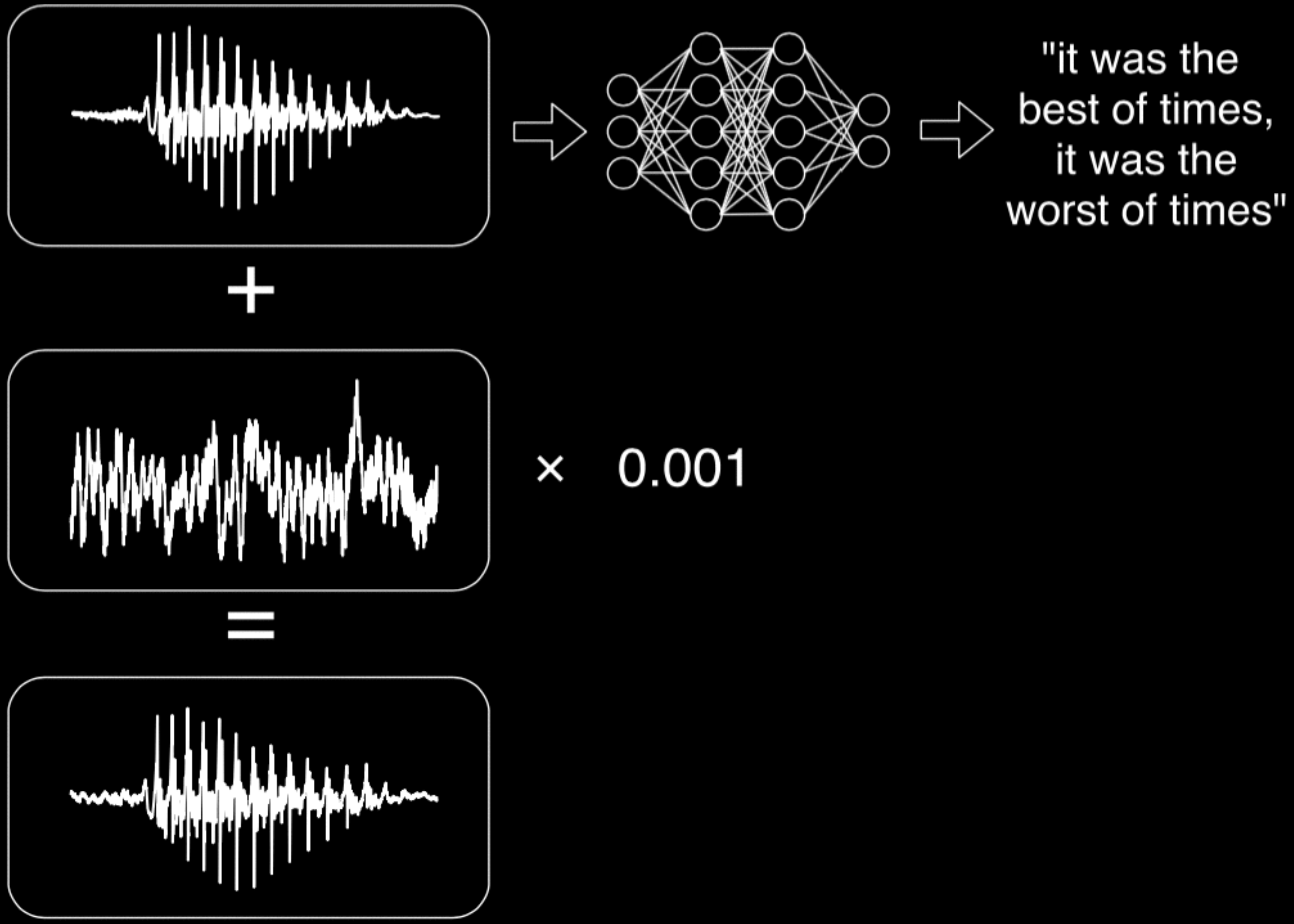
"most of them were staring
quietly at the big table"

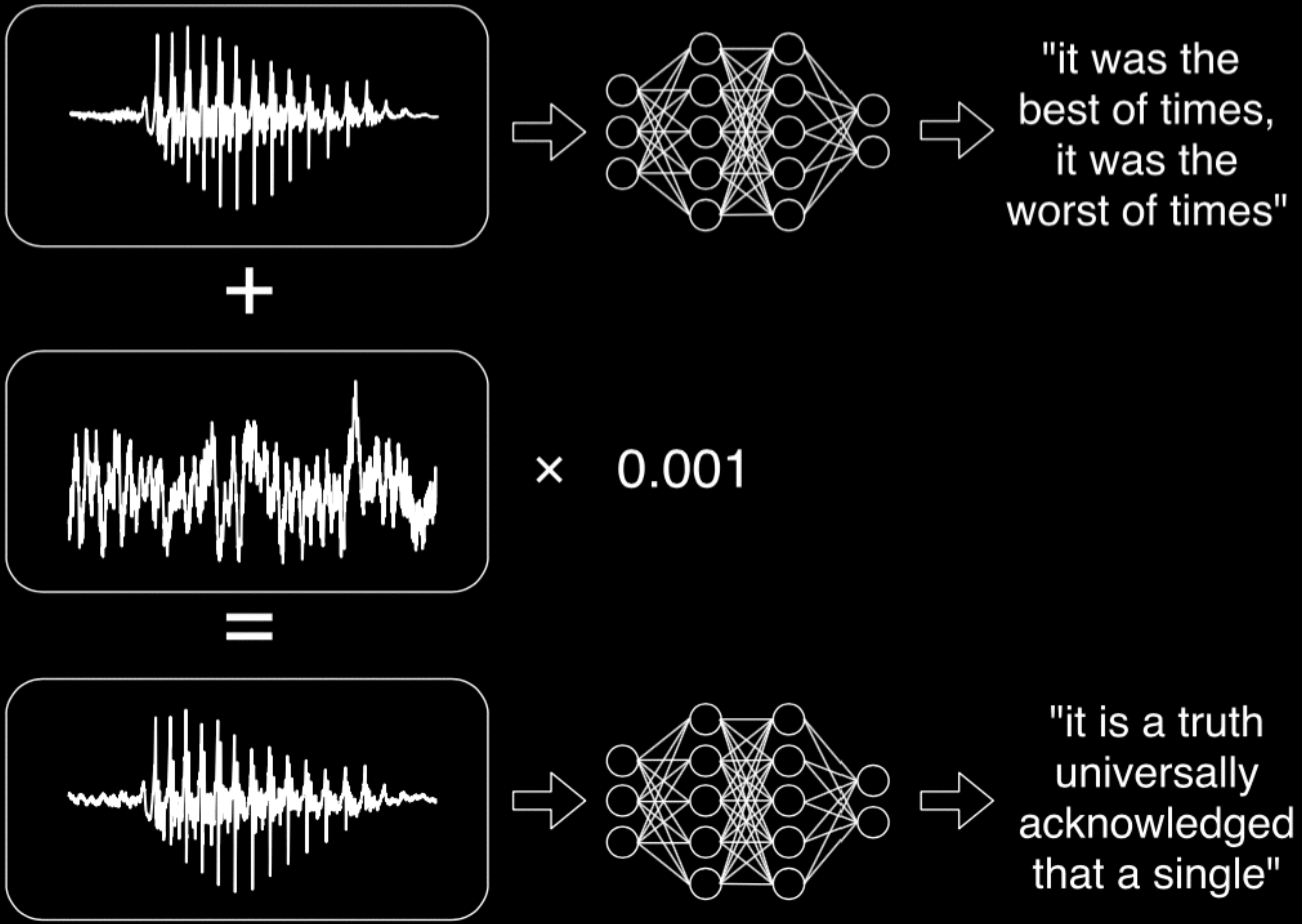
What about this?

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity"


"it was the
best of times,
it was the
worst of times"


"it was the
best of times,
it was the
worst of times"
+
× 0.001


"it was the
best of times,
it was the
worst of times"
+
× 0.001
=



"it was the best of times, it was the worst of times"
+
× 0.001
=
"it is a truth universally acknowledged that a single"

Okay, lesson learned.

Okay, lesson learned.

or audio

Don't classify images with neural networks.

Generating Natural Language Adversarial Examples

Moustafa Alzantot[1*], Yash Sharma[2*], Ahmed Elgohary[3],
Bo-Jhang Ho[1], Mani B. Srivastava[1], Kai-Wei Chang[1]

[1]Department of Computer Science, University of California, Los Angeles (UCLA)
{malzantot, bojhang, mbs, kwchang}@ucla.edu
[2]Cooper Union sharma2@cooper.edu
[3]Computer Science Department, University of Maryland elgohary@cs.umd.edu

Adversarial Attacks on Neural Network Policies

Sandy Huang[†], Nicolas Papernot[‡], Ian Goodfellow[§], Yan Duan[†§], Pieter Abbeel[†§]
[†] University of California, Berkeley, Department of Electrical Engineering and Computer Sciences
[‡] Pennsylvania State University, School of Electrical Engineering and Computer Science
[§] OpenAI

Abstract

Machine learning classifiers are known to be vulnerable to inputs maliciously constructed by adversaries to force misclassification. Such adversarial examples have been extensively studied in the context of computer vision applications. In this work, we show adversarial attacks are also effective when targeting neural network

Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples

Minhao Cheng[1], Jinfeng Yi[2], Huan Zhang[1], Pin-Yu Chen[3], Cho-Jui Hsieh[1]

[1]Department of Computer Science, University of California, Davis, CA 95616
[2]Tencent AI Lab, Bellevue, WA 98004
[3]IBM Research AI, Yorktown Heights, NY 10598
mhcheng@ucdavis.edu, jinfengyi.ustc@gmail.com, ecezhang@ucdavis.edu,
pin-yu.chen@ibm.com, chohsieh@ucdavis.edu

HALLUCINATIONS IN NEURAL MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural machine translation (NMT) systems have reached state of the art performance in translating text and are in wide deployment. Yet little is understood about how these systems function or break. Here we show that NMT systems are susceptible to producing highly pathological translations that are completely untethered from the source material, which we term *hallucinations*. Such pathological translations are problematic because they are are deeply disturbing of user trust and easy to find with a simple search. We describe a method to generate hallucinations and show that many common variations of the NMT ar-
of hal
nique
nally,
in the

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION

Yonatan Belinkov*
Computer Science and
Artificial Intelligence Laboratory,
Massachusetts Institute of Technology
belinkov@mit.edu

Yonatan Bisk*
Paul G. Allen School
of Computer Science & Engineering,
University of Washington
ybisk@cs.washington.edu

On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab Ondrej Miksik Philip H.S. Torr
University of Oxford
{anurag.arnab, ondrej.miksik, philip.torr}@eng.ox.ac.uk

Okay, lesson learned.

Okay, lesson learned.

Don't let adversaries perform gradient descent.

Published as a conference paper at ICLR 2018

MITIGATING ADVERSARIAL EFFECTS THROUGH RAN-
DOMIZATION

Cihang Xie, Zhishuai Zhang &
Department of Computer Science
The Johns Hopkins University
Baltimore, MD 21218 USA
{cihangxie306, zhshuai.

Jianyu Wang
Baidu Research USA
Sunnyvale, CA 94089 USA
wjyouch@gmail.com

Zhou Ren
Snap Inc.
Venice, CA 90291 USA
zhou.ren@snapchat.com

Convolutional neural netw
in recent years. However,
For example, imperceptibl
lutional neural networks t
at inference time to mitiga
ization operations: randor
size, and random padding
dom manner. Extensive e
tion method is very effecti
tacks. Our method provides the following advantages: 1) no additional training or fine-tuning, 2) very few additional computations, 3) compatible with other adversarial defense methods. By combining the proposed randomization method with an adversarially trained model, it achieves a normalized score of 0.924 (ranked No.2 among 107 defense teams) in the NIPS 2017 adversarial examples defense challenge, which is far better than using adversarial training alone with a normalized score of 0.773 (ranked No.56). The code is public available at https://github.com/cihangxie/NIPS2017_adv_challenge_defense.

Published as a conference paper at ICLR 2018

STOCHASTIC ACTIVATION PRUNING FOR ROBUST ADVERSARIAL DEFENSE

Guneet S. Dhillon[1,2], Kamyar Azizzadenesheli[3], Zachary C. Lipton[1,4],
Jeremy Bernstein[1,5], Jean Kossaifi[1,6], Aran Khanna[1], Anima Anandkumar[1,5]
[1]Amazon AI, [2]UT Austin, [3]UC Irvine, [4]CMU, [5]Caltech, [6]Imperial College London
guneetdhillon@utexas.edu, kazizzad@uci.edu, zlipton@cmu.edu,
bernstein@caltech.edu, jean.kossaifi@imperial.ac.uk,
aran@arankhanna.com, anima@amazon.com

ABSTRACT

Neural networks are known to be vulnerable to adversarial exan
chosen perturbations to real images, while imperceptible to hum
classification and threaten the reliability of deep learning systems
guard against adversarial examples, we take inspiration from game
the problem as a minimax zero-sum game between the adversary a
general, for such games, the optimal strategy for both players rec
tic policy, also known as a *mixed strategy*. In this light, we pro
Activation Pruning (SAP), a mixed strategy for adversarial defen
a random subset of activations (preferentially pruning those with
tude) and scales up the survivors to compensate. We can apply S
networks, including adversarially trained models, without fine-tuni
bustness against adversarial examples. Experiments demonstrate t
robustness against attacks, increasing accuracy and preserving cal

Published as a conference paper at ICLR 2018

THERMOMETER ENCODING: ONE HOT WAY TO RESIST ADVERSARIAL EXAMPLES

Jacob Buckman*‡, Aurko Roy*, Colin Raffel, Ian Goodfellow
Google Brain
Mountain View, CA
{buckman, aurkor, craffel, goodfellow}@google.com

ABSTRACT

mples" for neu-
ndistinguishable
ural network ar-
s the robustness
ustness with ex-
tasets, and show
higher accuracy
-of-the-art accu-
from 93.20% to
lore the proper-
lings help neural

Published as a conference paper at ICLR 2018

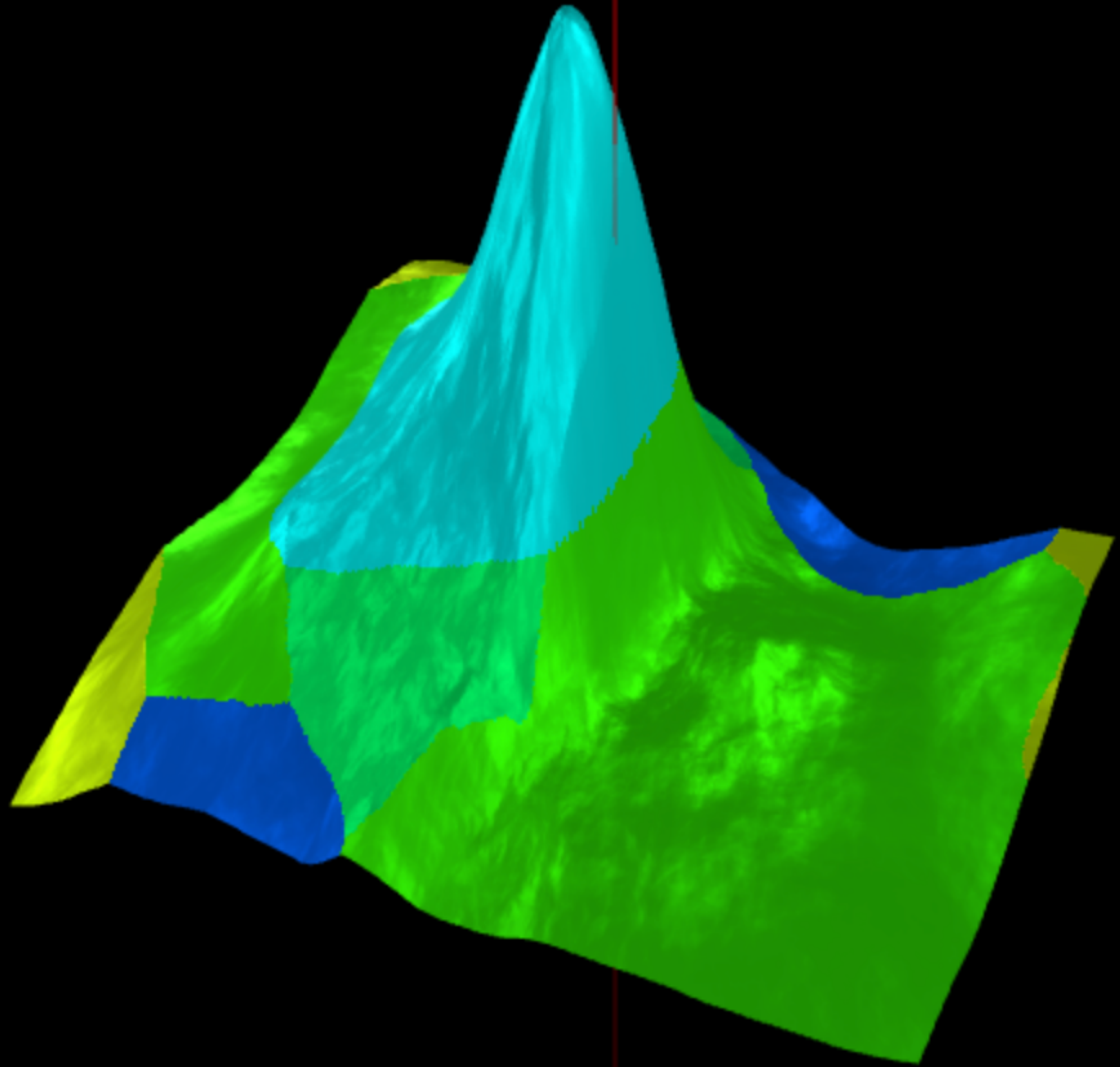
COUNTERING ADVERSARIAL IMAGES USING INPUT TRANSFORMATIONS

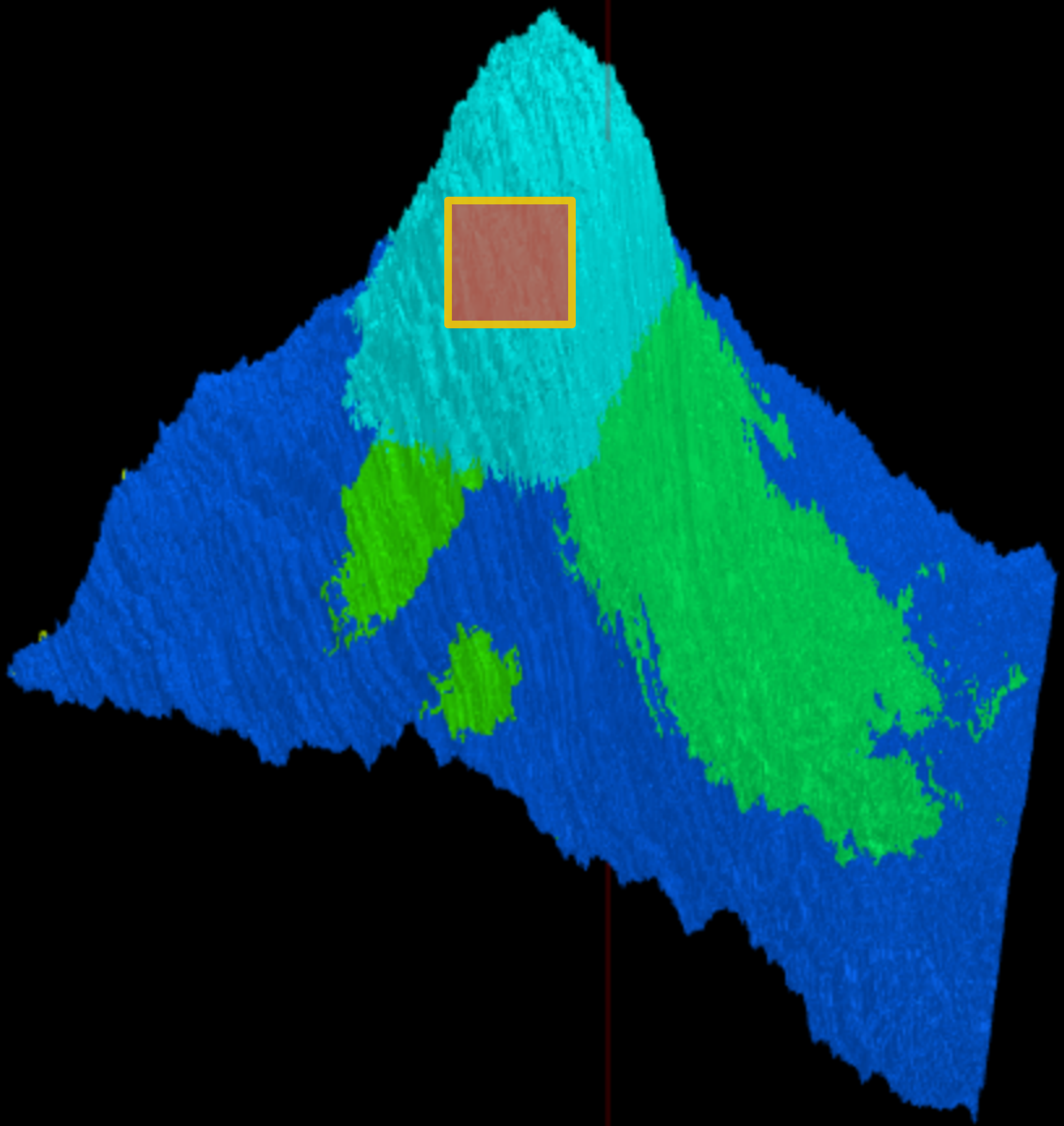
Chuan Guo*
Cornell University

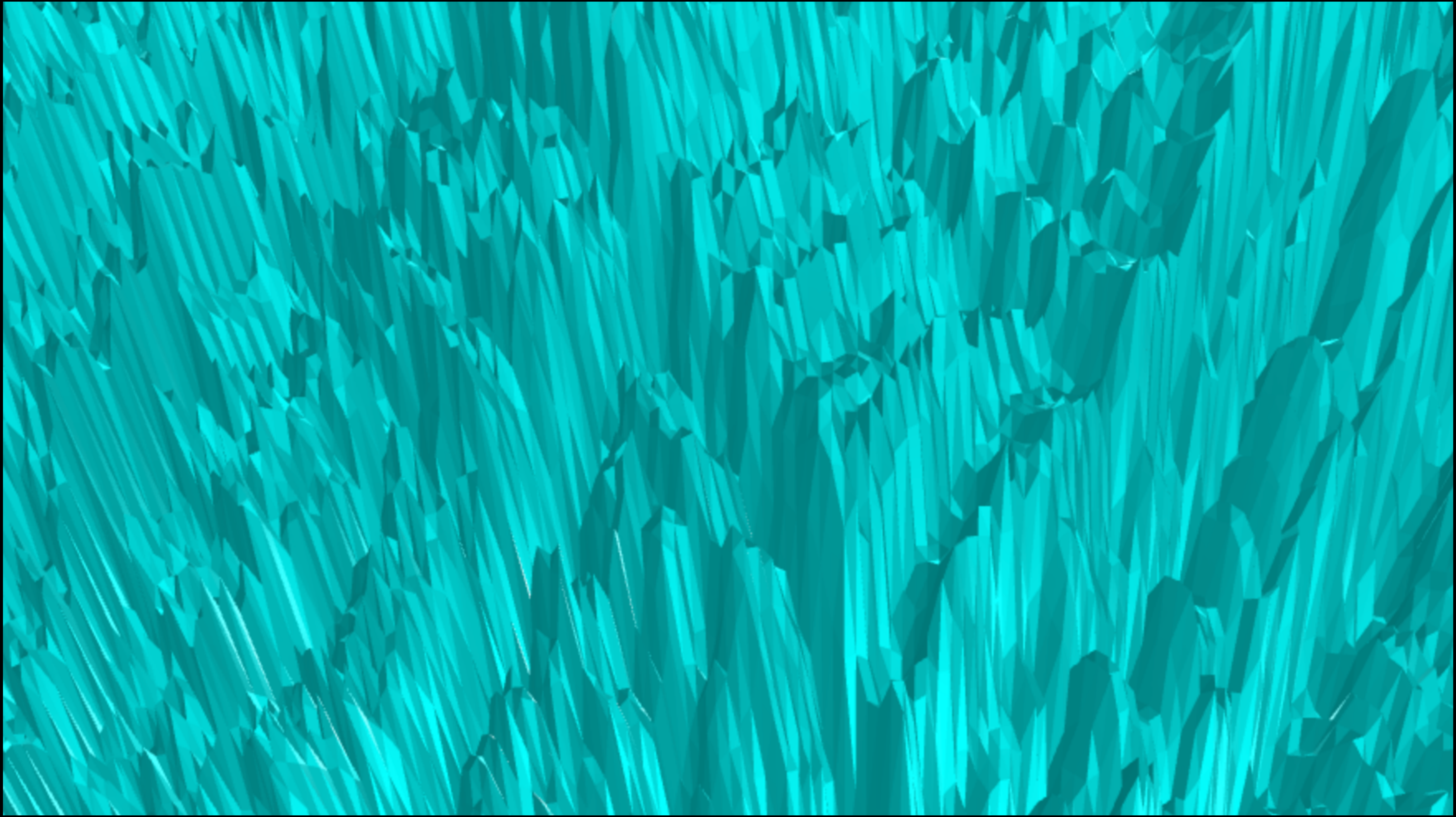
Mayank Rana & Moustapha Cissé & Laurens van der Maaten
Facebook AI Research

ABSTRACT

This paper investigates strategies that defend against adversarial-example attacks on image-classification systems by transforming the inputs before feeding them to the system. Specifically, we study applying image transformations such as bit-depth reduction, JPEG compression, total variance minimization, and image quilting before feeding the image to a convolutional network classifier. Our experiments on ImageNet show that total variance minimization and image quilting are very effective defenses in practice, in particular, when the network is trained on transformed images. The strength of those defenses lies in their non-differentiable nature and their inherent randomness, which makes it difficult for an adversary to circumvent the defenses. *Our best defense eliminates* 60% *of strong gray-box and* 90% *of strong black-box attacks by a variety of major attack methods.*







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

Anish Athalye [*1] **Nicholas Carlini** [*2] **David Wagner** [2]

Abstract

We identify obfuscated gradients, a kind of gradient masking, as a phenomenon that leads to a false sense of security in defenses against adversarial examples. While defenses that cause obfuscated gradients appear to defeat iterative optimization-based attacks, we find defenses relying on this effect can be circumvented. We describe characteristic behaviors of defenses exhibiting the effect, and for each of the three types of obfuscated gradients we discover, we develop attack techniques to overcome it. In a case study, examining non-certified white-box-secure defenses at ICLR 2018, we find obfuscated gradients are a common occurrence, with 7 of 9 defenses relying on obfuscated gradients. Our new attacks successfully circumvent 6 completely, and 1 partially, in the original threat model each paper considers.

1. Introduction

In response to the susceptibility of neural networks to adversarial examples (Szegedy et al., 2013; Biggio et al., 2013), there has been significant interest recently in constructing defenses to increase the robustness of neural networks. While progress has been made in understanding and defending against adversarial examples in the white-box setting, where the adversary has full access to the network, a complete solution has not yet been found.

As benchmarking against iterative optimization-based attacks (e.g., Kurakin et al. (2016a); Madry et al. (2018); Carlini & Wagner (2017c)) has become standard practice in evaluating defenses, new defenses have arisen that appear to be robust against these powerful optimization-based attacks.

We identify one common reason why many defenses provide

[*]Equal contribution [1]Massachusetts Institute of Technology [2]University of California, Berkeley. Correspondence to: Anish Athalye <aathalye@mit.edu>, Nicholas Carlini <npc@berkeley.edu>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

apparent robustness against iterative optimizatio
obfuscated gradients, a term we define as a speci
gradient masking (Papernot et al., 2017). Witho

7 of 9 defenses relying on obfuscated gradients. Our new attacks successfully circumvent 6 completely, and 1 partially

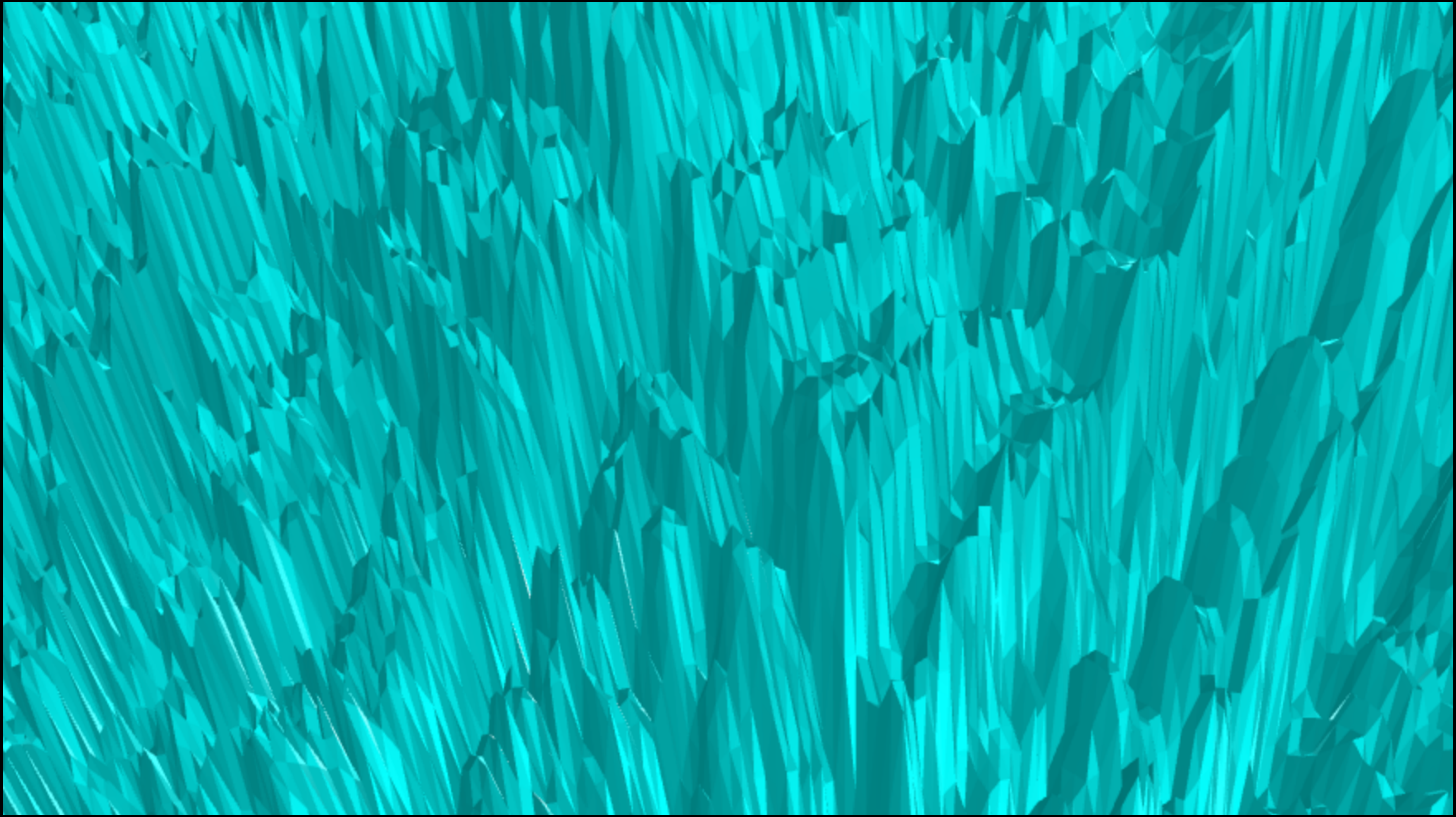
ents caused by these three phenomena. We address gradient shattering with a new attack technique we call Backward Pass Differentiable Approximation, where we approximate derivatives by computing the forward pass normally and computing the backward pass using a differentiable approximation of the function. We compute gradients of randomized defenses by applying Expectation Over Transformation (Athalye et al., 2017). We solve vanishing/exploding gradients through reparameterization and optimize over a space where gradients do not explode/vanish.

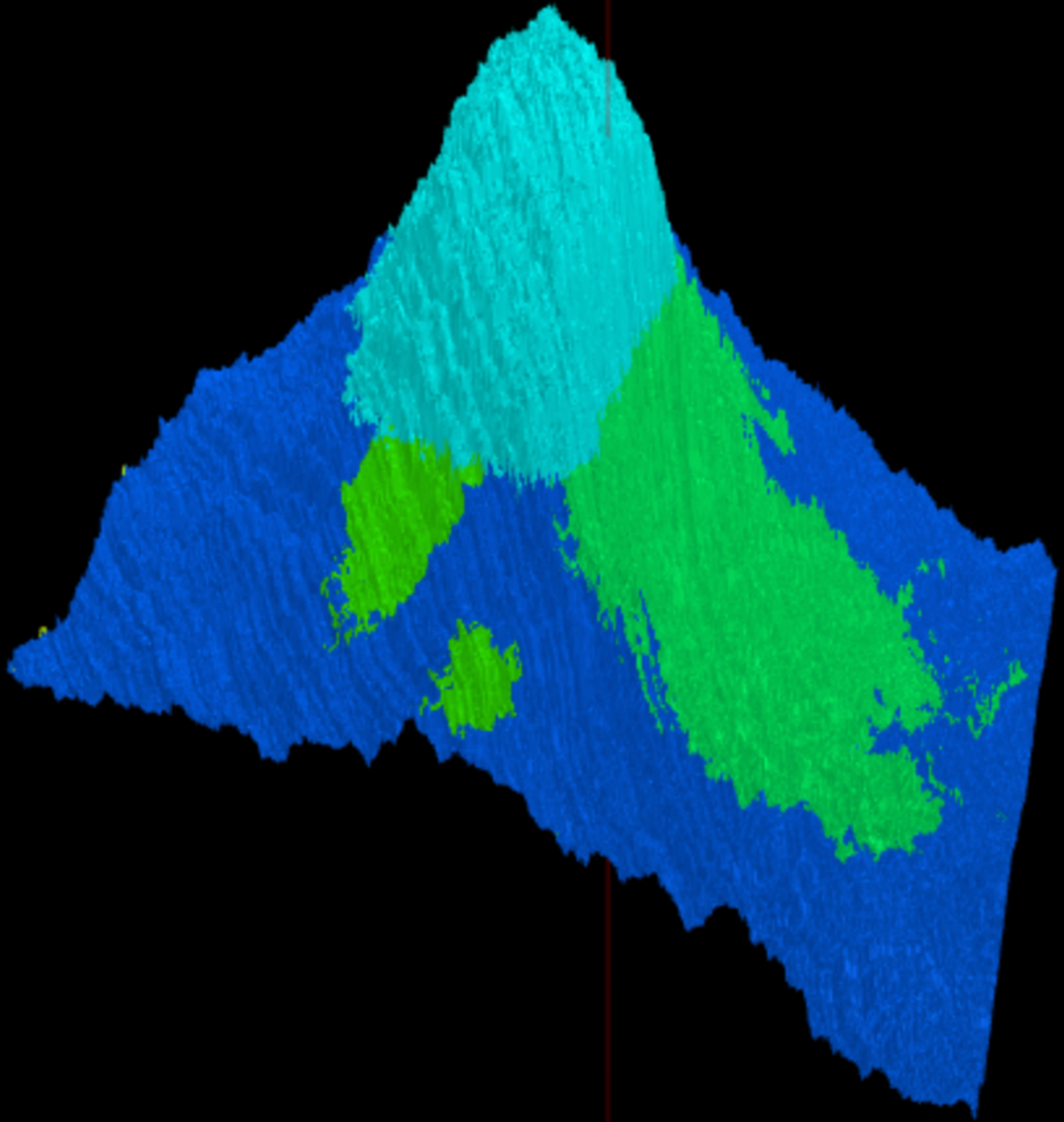
To investigate the prevalence of obfuscated gradients and understand the applicability of these attack techniques, we use as a case study the ICLR 2018 non-certified defenses that claim white-box robustness. We find that obfuscated gradients are a common occurrence, with 7 of 9 defenses relying on this phenomenon. Applying the new attack techniques we develop, we overcome obfuscated gradients and circumvent 6 of them completely, and 1 partially, under the original threat model of each paper. Along with this, we offer an analysis of the evaluations performed in the papers.

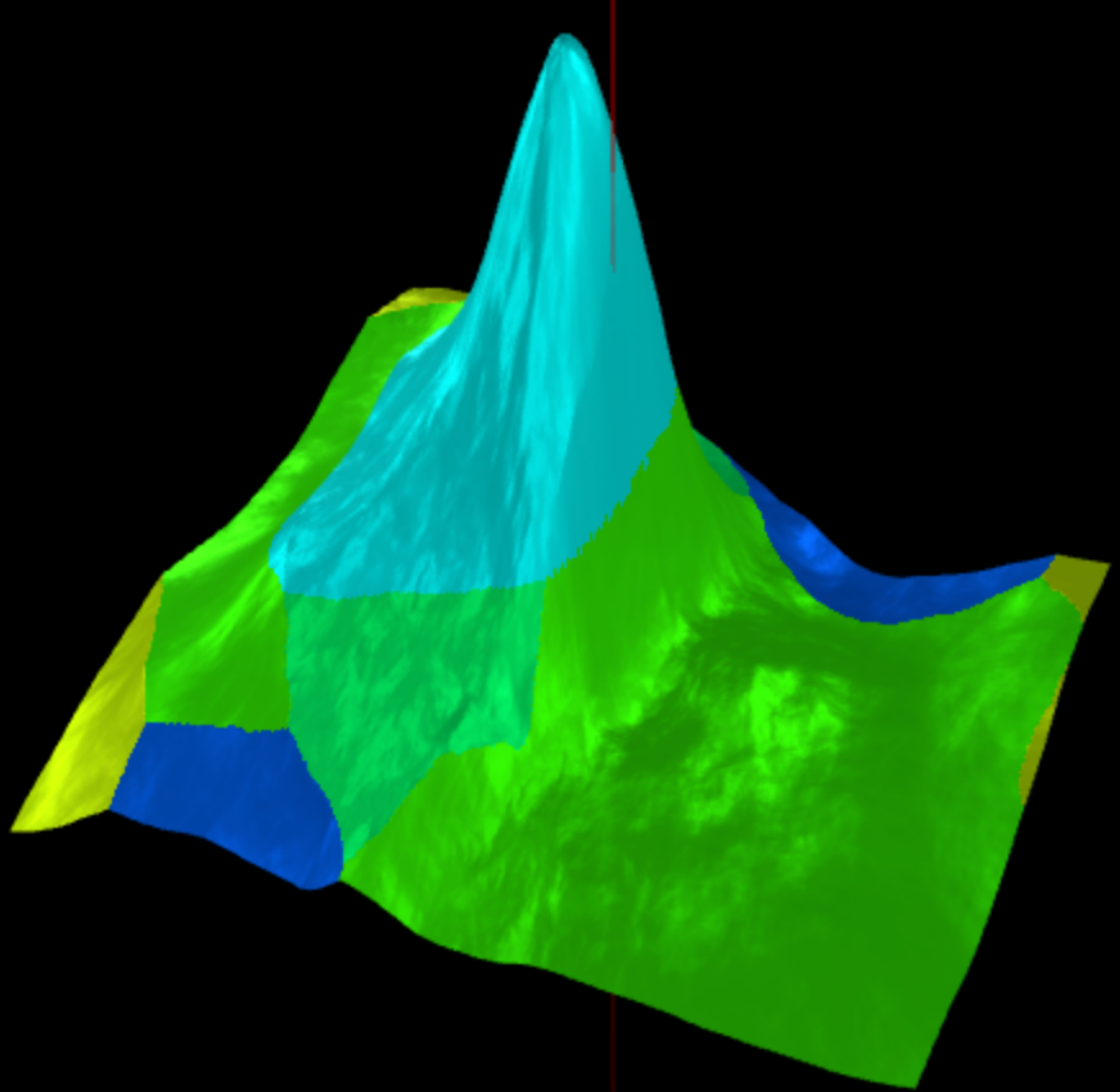
Additionally, we hope to provide researchers with a common baseline of knowledge, description of attack techniques, and common evaluation pitfalls, so that future defenses can avoid falling vulnerable to these same attack approaches.

To promote reproducible research, we release our reimplementation of each of these defenses, along with implementations of our attacks for each. [1]

[1] https://github.com/anishathalye/obfuscated-gradients







Okay, lesson learned.

Okay, lesson learned.

Don't let adversaries have **ANY** access to my model

Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models

Wieland Brendel*, Jonas Rauber* & Matthias Bethge
Werner Reichardt Centre for I
Eberhard Karls University Tüb
{wieland,jonas,matth

Delving into Transferable Adversarial Examples and Black-box Attacks

Yanpei Liu*, Xinyun Chen*
Shanghai Jiao Tong University

Chang Liu, Dawn Song
University of the California, Berkeley

Abstract

An intriguing property of deep neural networks is the existence of adversarial examples, which can transfer among different architectures. These transferable ad-

which is a black-box image classification system.

Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

Nicolas Papernot and Patrick McDaniel
The Pennsylvania State University
University Park, PA
{ngp5056,mcdaniel}@cse.psu.edu

Ian Goodfellow
OpenAI
San Francisco, CA
ian@openai.com

Universal adversarial perturbations

Seyed-Mohsen Moosavi-Dezfooli*†
seyed.moosavi@epfl.ch

Alhussein Fawzi*†
alhussein.fawzi@epfl.ch

Omar Fawzi‡
omar.fawzi@ens-lyon.fr

Pascal Frossard†
pascal.frossard@epfl.ch

ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models

Pin-Yu Chen*
AI Foundations Group
IBM T. J. Watson Research Center
Yorktown Heights, NY
pin-yu.chen@ibm.com

Huan Zhang*†
University of California, Davis
Davis, CA
ecezhang@ucdavis.edu

Yash Sharma
IBM T. J. Watson Research Center
Yorktown Heights, NY
Yash.Sharma3@ibm.com

The Space of Transferable Adversarial Examples

Florian Tramèr[1], Nicolas Papernot[2], Ian Goodfellow[3], Dan Boneh[1], and Patrick McDaniel[2]

[1]Stanford University, [2]Pennsylvania State University, [3]Google Brain

Abstract

Adversarial examples are maliciously perturbed inputs designed to mislead machine learning (ML) models at test-time. They often *transfer*: the same adversarial example fools more than one model.

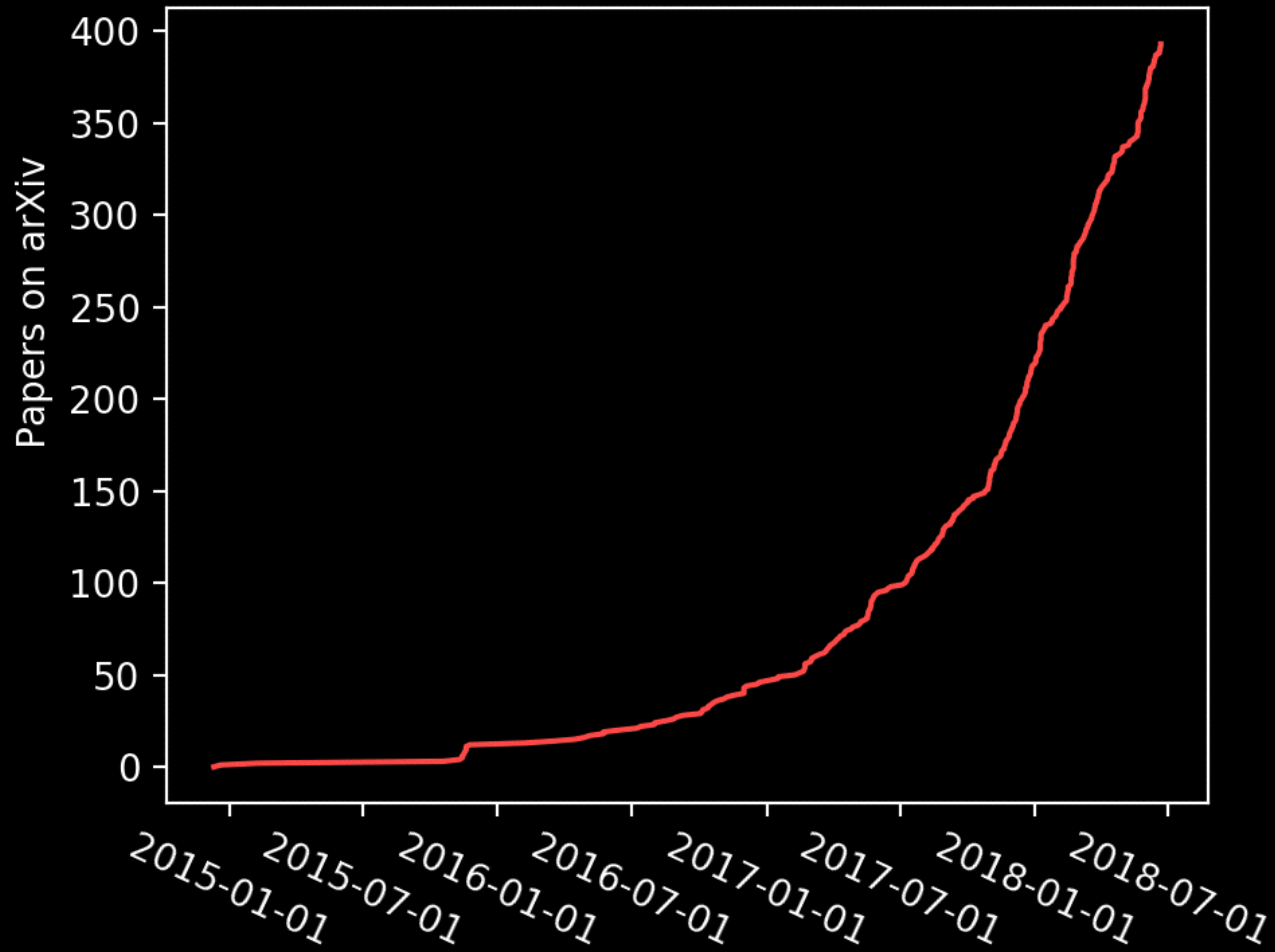
In this work, we propose novel methods for estimating the previously unknown *dimensionality* of the space of adversarial inputs. We find that adversarial examples span a contiguous subspace of large (~25) dimensionality. Adversarial subspaces with higher dimensionality are more likely to intersect. We find that for two different models, a significant fraction of their subspaces is shared, thus enabling transferability.

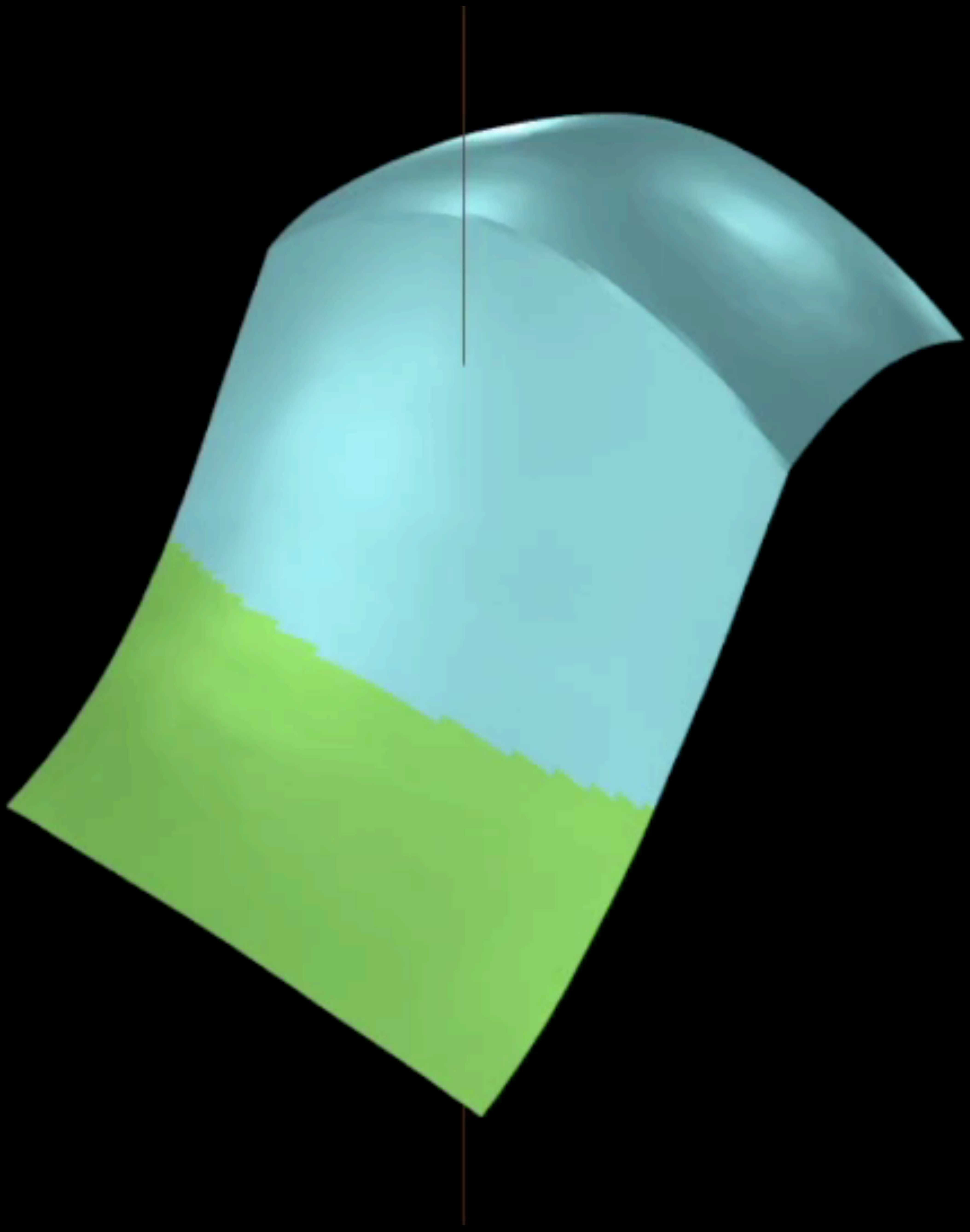
: analysis of the similarity of different models' decision
it these boundaries are actually close in *arbitrary* directions,
benign. We conclude by formally studying the *limits* of
ive (1) sufficient conditions on the *data distribution* that
or simple model classes and (2) examples of scenarios in
: occur. These findings indicate that it may be possible to
transfer-based attacks, even for models that are vulnerable

Okay, lesson learned.

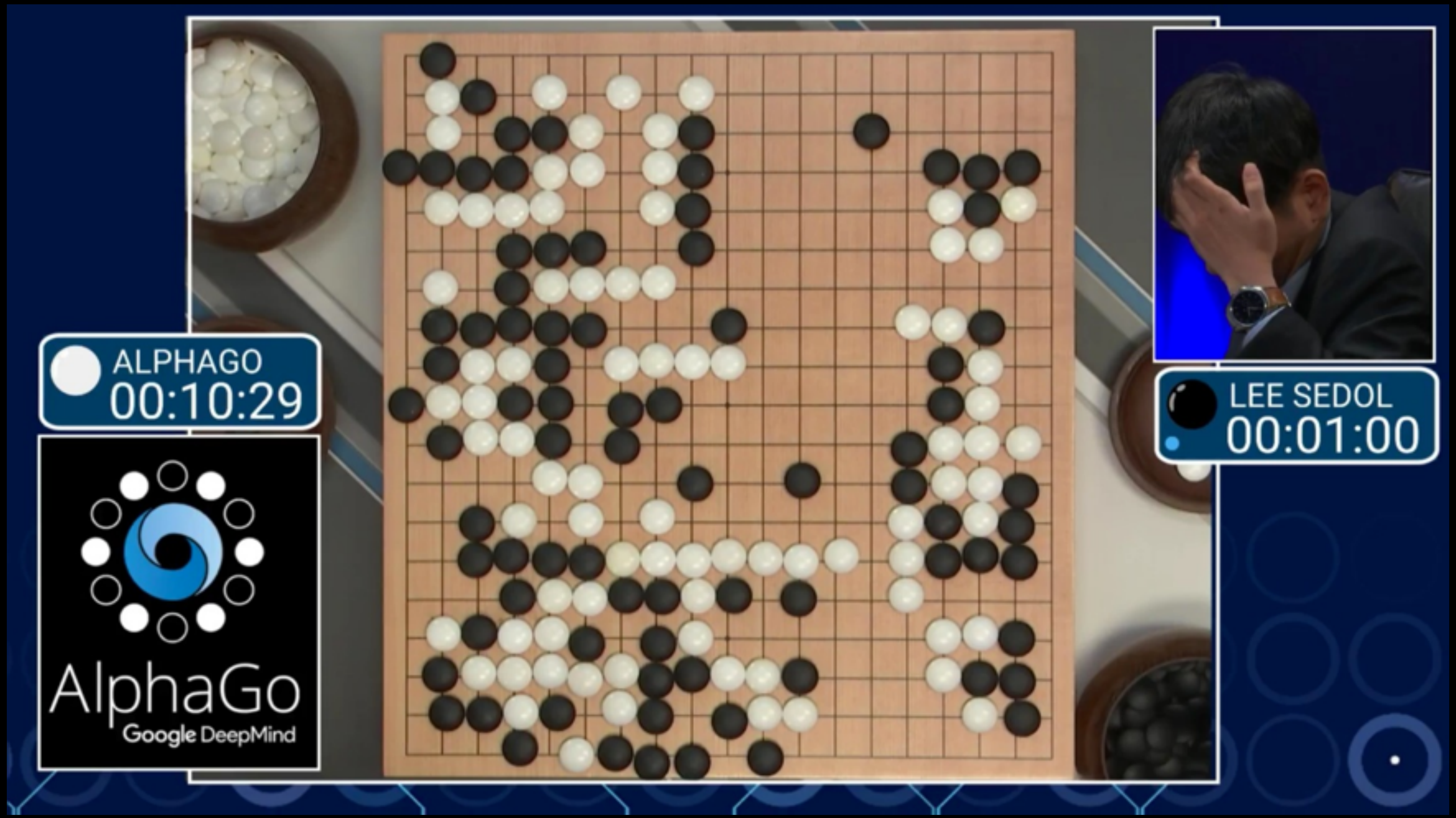
Okay, lesson learned.

Give up.


Papers on arXiv
400
350
300
250
200
150
100
50
0
2015-01-01
2015-07-01
2016-01-01
2016-07-01
2017-01-01
2017-07-01
2018-01-01
2018-07-01



Yes, machine learning gives **amazing** results



However, there are also significant **vulnerabilities**

Guacamole (99%)

Questions?

https://nicholas.carlini.com
nicholas@carlini.com