Evaluating the Robustness of Neural Networks Defenses

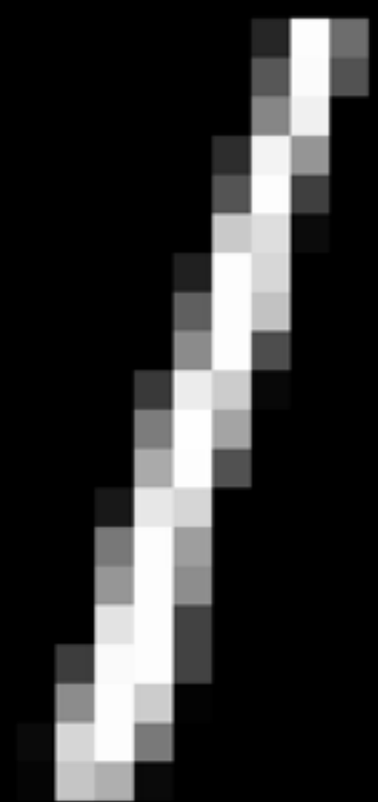
Nicholas Carlini
Google Brain

Background: Notation

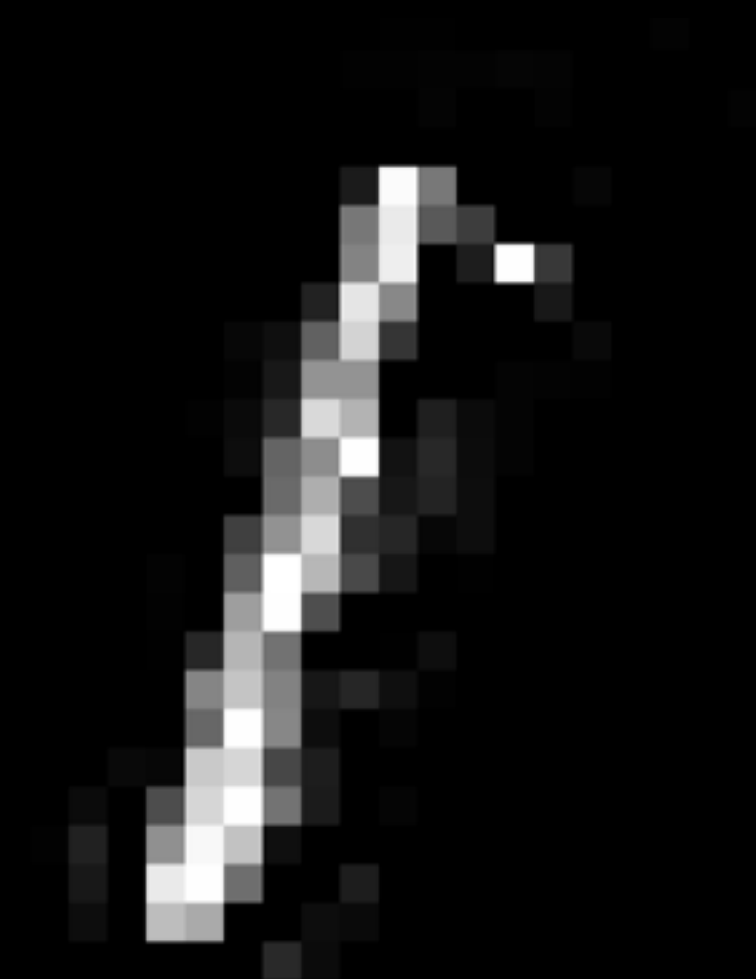
- We have a classification neural network $F(x)$
- Given an input X classified as label L
- $F(X)_L$ is the probability of label L
- $F(X) = \text{softmax}(F_Z(X))$ [the logits]
- $C(X) = \arg\max_j F(X)_j$

Background: Adversarial Examples

- For a classification neural network F(x)
- Given an input X classified as label L ...
- ... it is easy to find an X′ close to X
- ... so that F(X′) != L

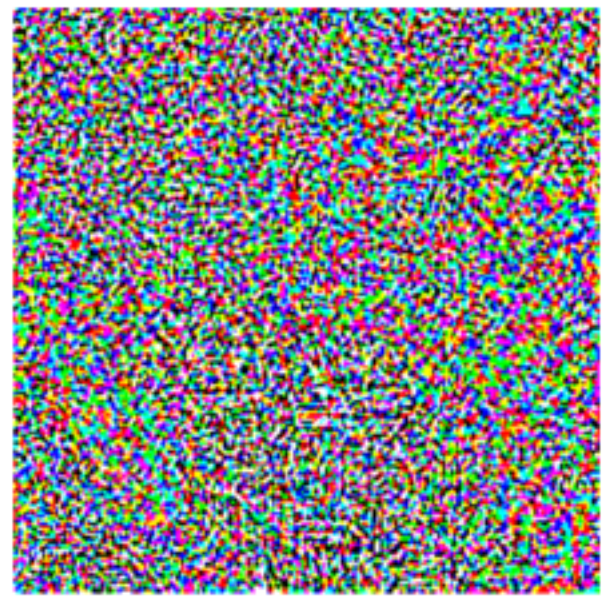


Classified as a 1



Classified as a 0

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Why should we care about adversarial examples?

Make ML ***robust***

Make ML ***better***


STOP

STOP

2014

Finding Adversarial Examples

- Formulation: given input x, find x′ where
  minimize $\quad d(x,x')$
  such that $\quad F(x') = T$
  $\quad\quad\quad\quad x'$ is "valid"

- Gradient Descent to the rescue?

- Non-linear constraints are hard

Reformulation

- Formulation:
  minimize $d(x,x') + g(x')$
  such that x' is "valid"

- Where $g(x')$ is some kind of loss function on how close $F(x')$ is to target T

  - $g(x')$ is small if $F(x') = T$

  - $g(x')$ is large if $F(x') \neq T$

Reformulation

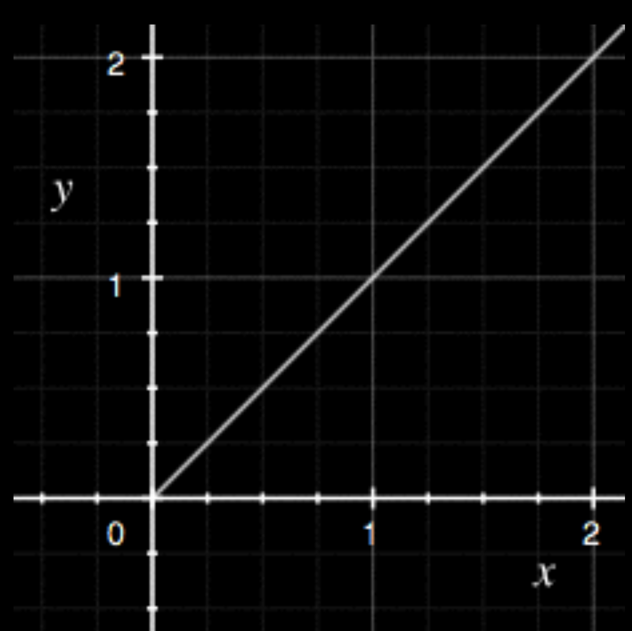
- For example
  - $g(x') = (1 - F(x')_T)$
- If $F(x')$ says the probability of T is 1:
  - $g(x') = (1 - F(x')_T) = (1-1) = 0$
- $F(x')$ says the probability of T is 0:
  - $g(x') = (1 - F(x')_T) = (1-0) = 1$

Does this work?

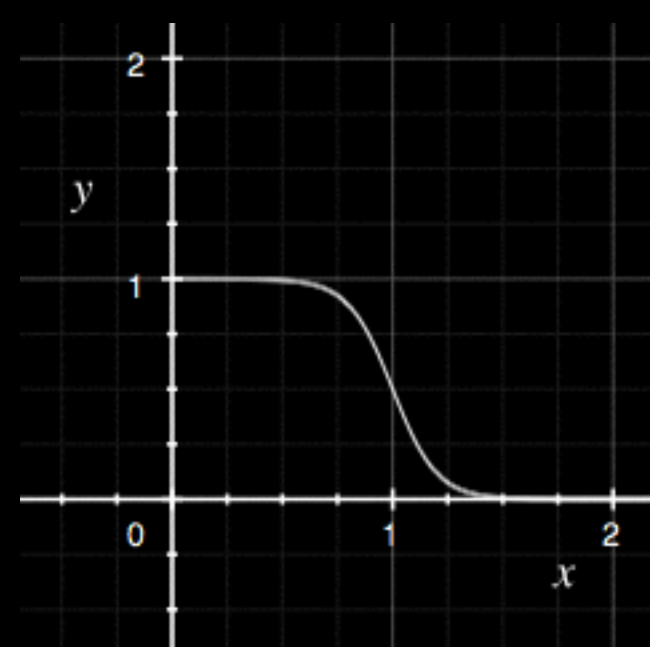
- Fo
  mi
  su

Problem 1:
Global minimum is not an adversarial example

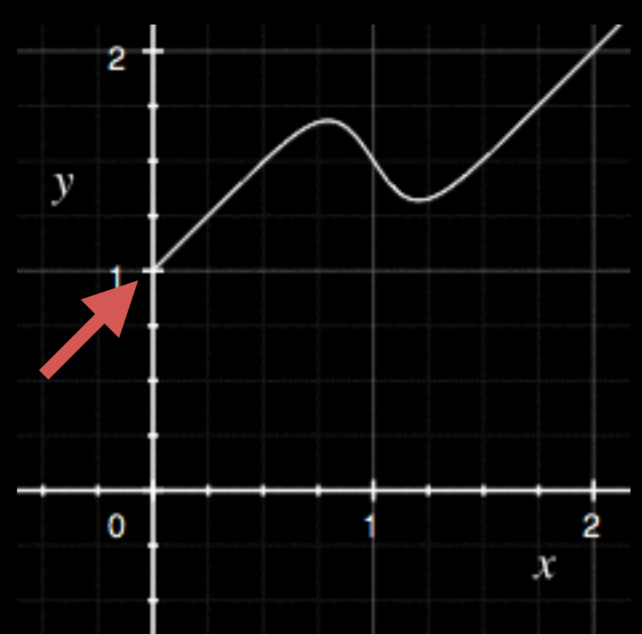
d(x,x′) + g(x′)



+



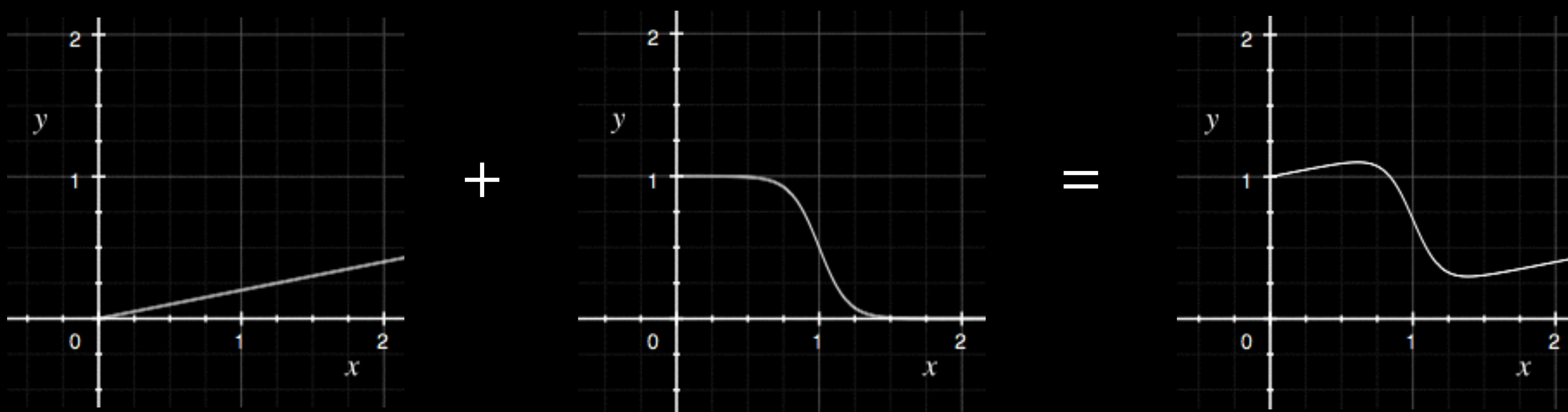
=



Does this work?

- Formulation:
  minimize $\quad d(x,x')/5 + g(x')$
  such that $\quad x'$ is "valid"

$d(x,x')/5 \quad + \quad g(x')$

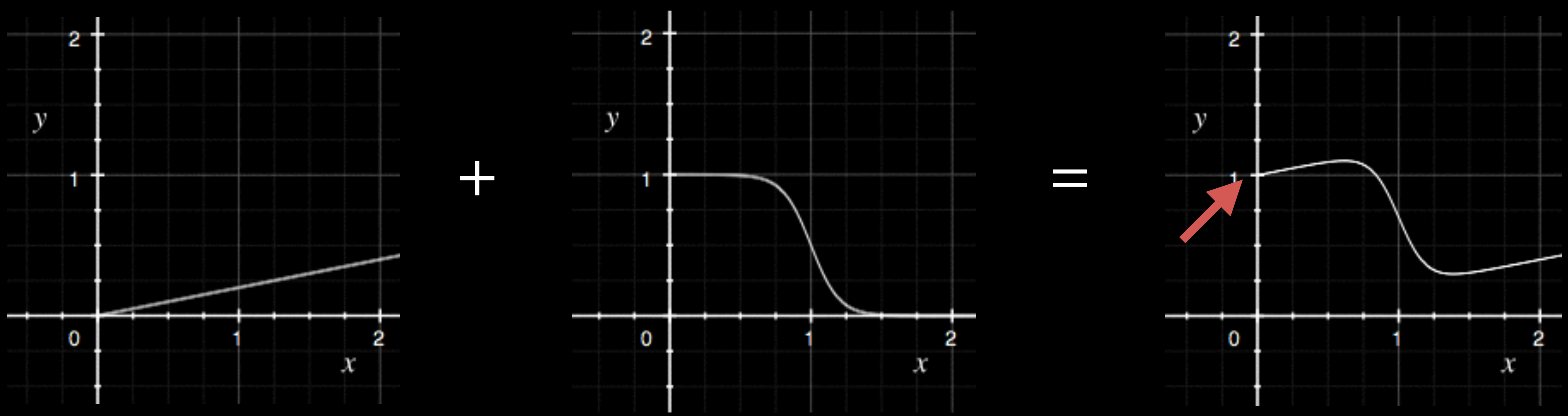


Does this work?

•

Problem 2:
Gradient direction does not point toward the global minimum

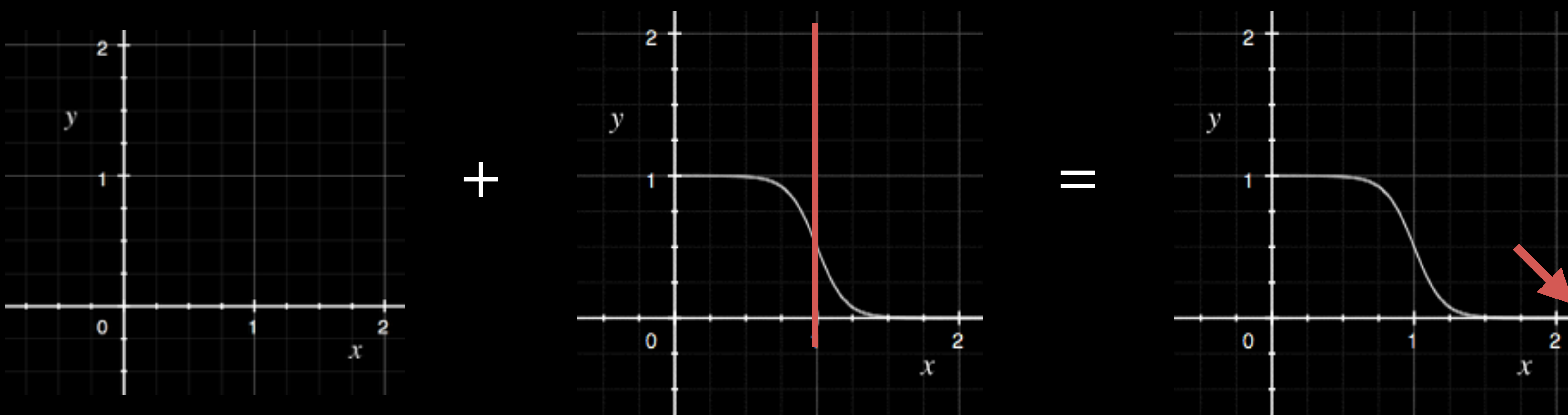
d(x,x′)/5 + g(x′)



Does this work?

Problem 3:
Global minimum is not the minimally perturbed adversarial example

d(x,x′)/1e10 + g(x′)



2015

This is very hard.

Let's do something simpler.

Fast Gradient Sign

- Unroll the gradient descent step by one
- $X' = X+\varepsilon \text{ sign}(\nabla_X F(X)_{L)}$

2016

How can we stop adversarial examples?

Distillation as a Defense

1. Train a model F() on the training data X,Y

2. Generate new training labels Y' by setting
   Y' = {100*F(x) : x in X}

3. Train a new classifier G() on X,Y'

Does it work?

Unfortunately, no.

2017

Constructing a better loss function

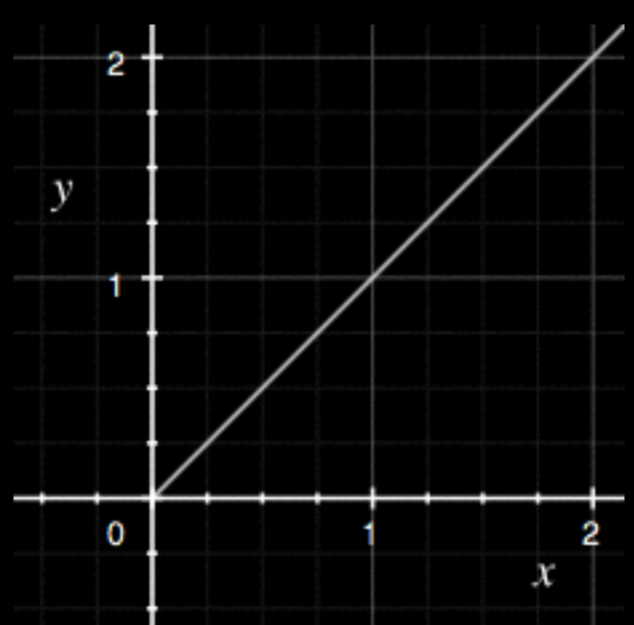
1. Global minimum at the decision boundary
2. Gradient points towards the global minimum

$$\boxed{\max\left(\boxed{\boxed{\max_{t'\neq t}}\left\{\boxed{\log(F(x)'_t)}\right\}} - \boxed{\log(\boxed{F(x)_t})}, 0\right)}$$

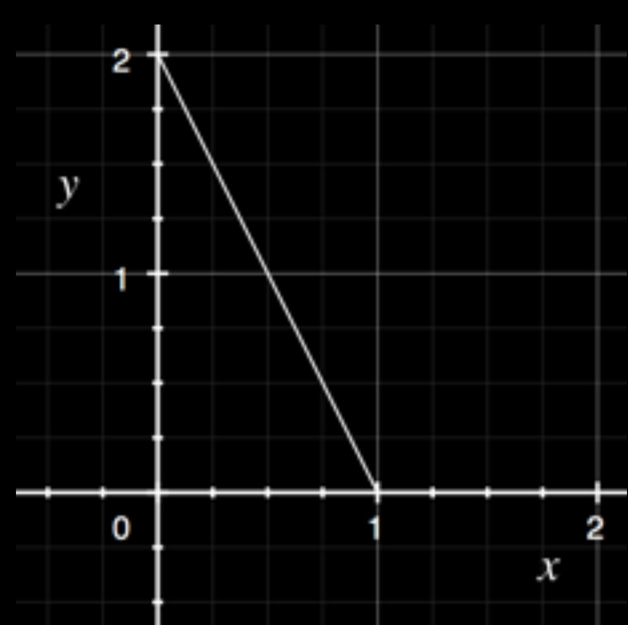
Improved Formulation

- Formulation:
  minimize $\quad d(x,x') + g(x')$
  such that $\quad x'$ is "valid"

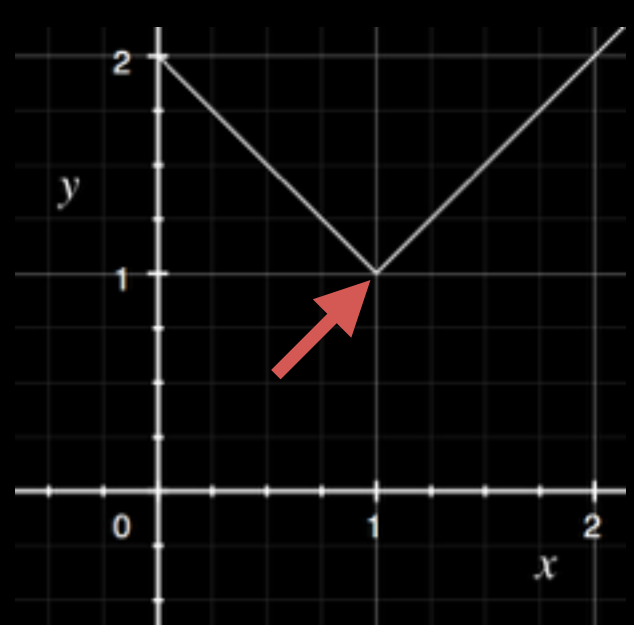
$d(x,x') \quad + \quad g(x')$

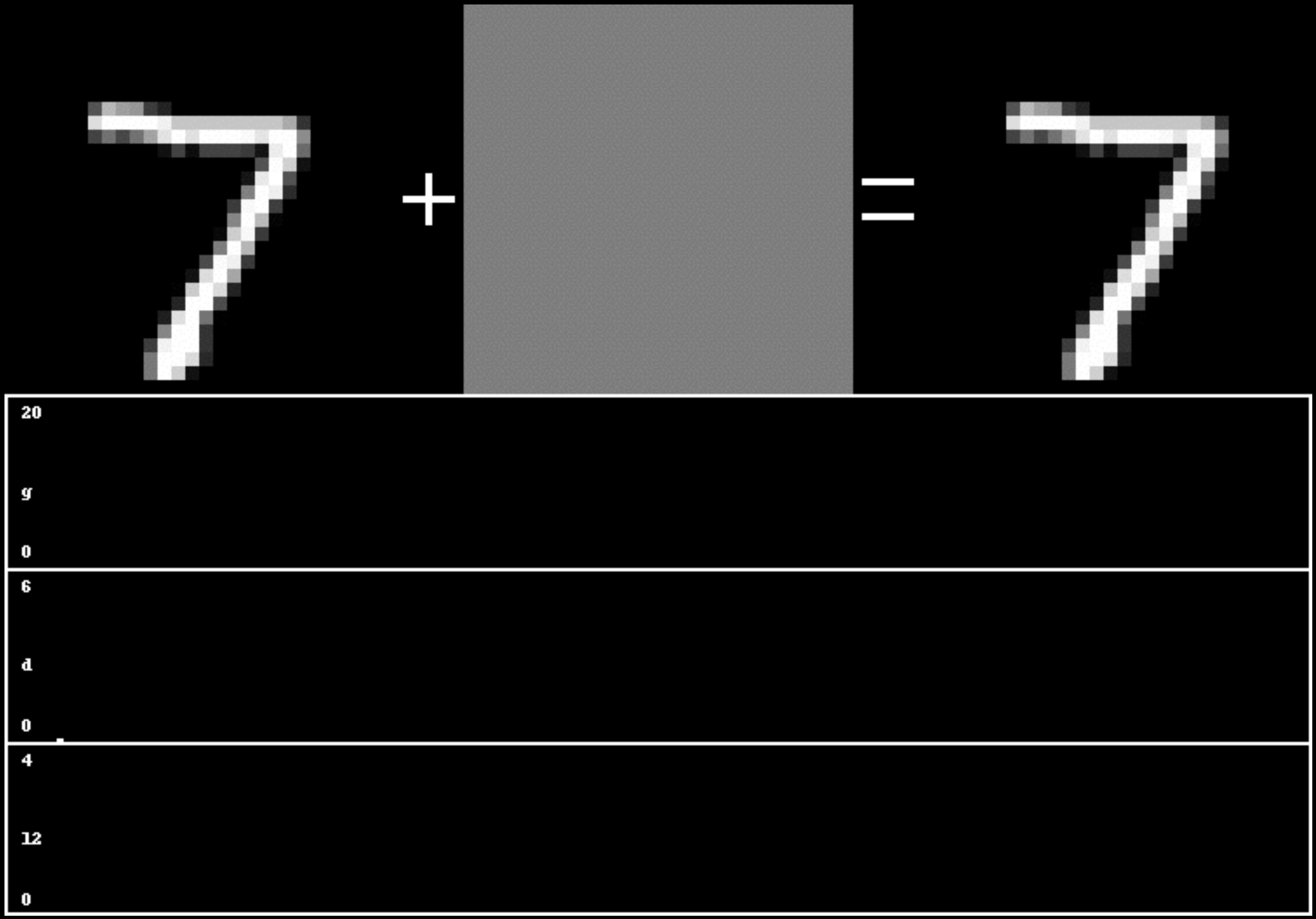


+

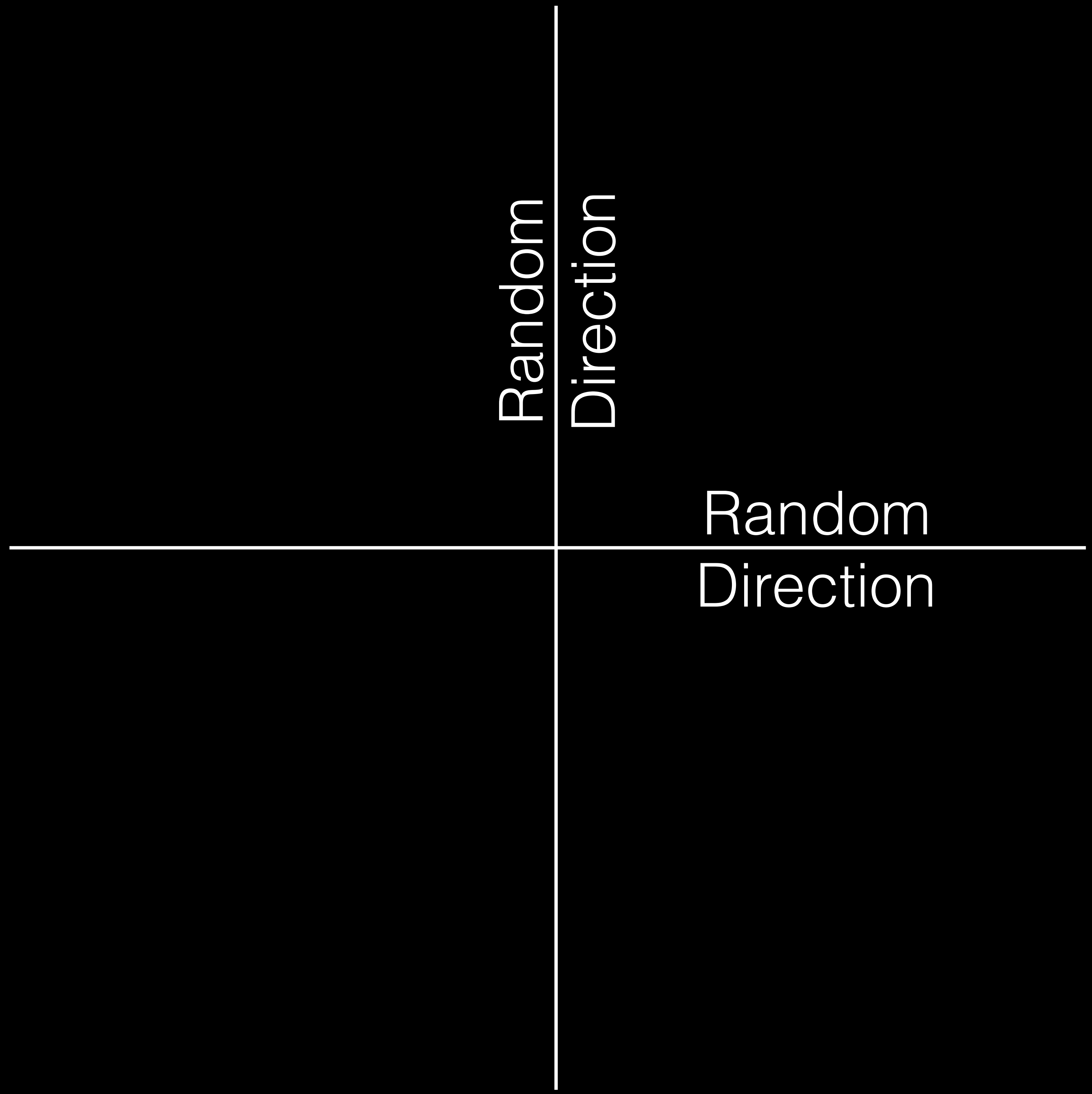


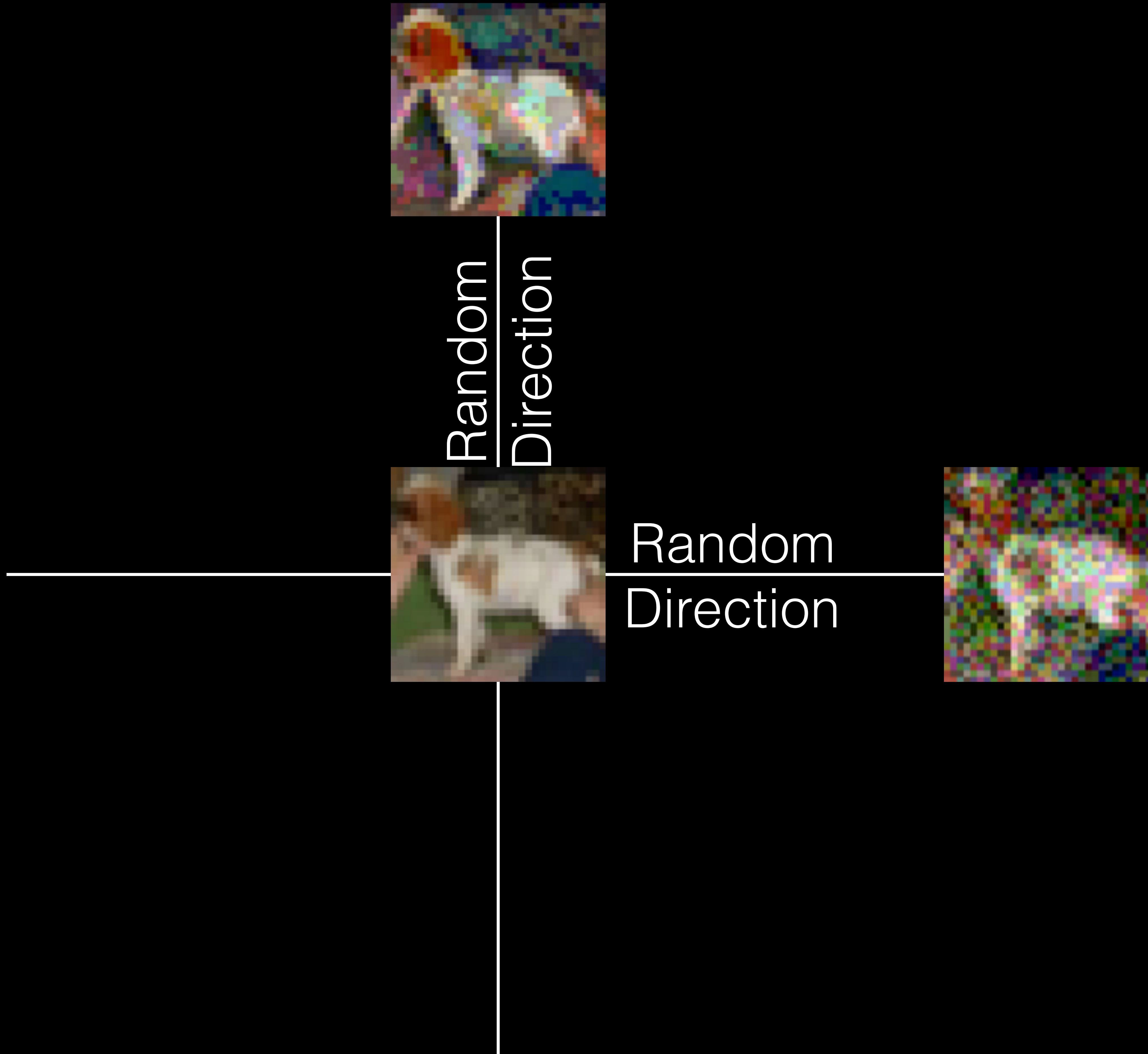
=



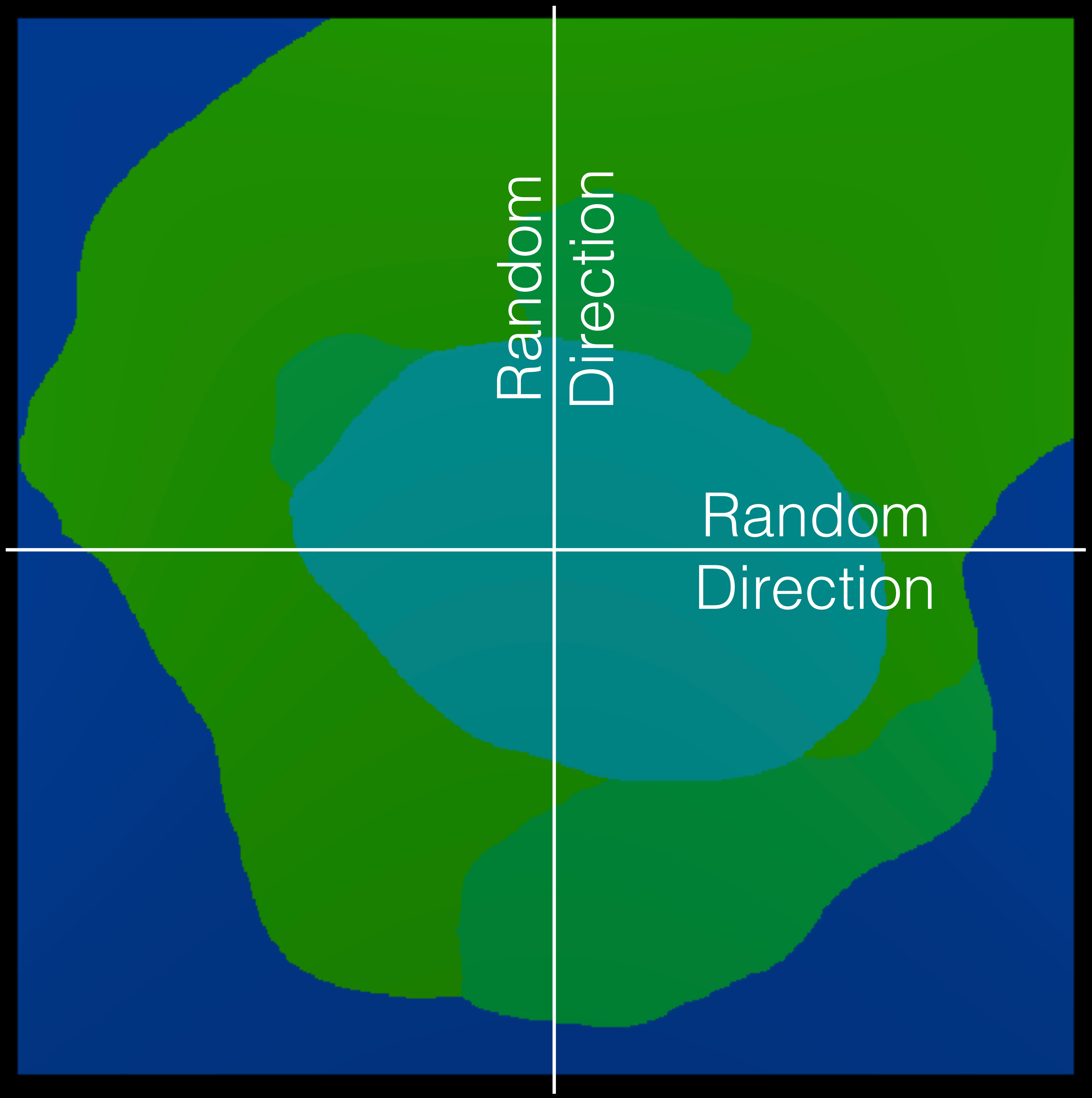

+
=
20
0
6
0
4
12
0

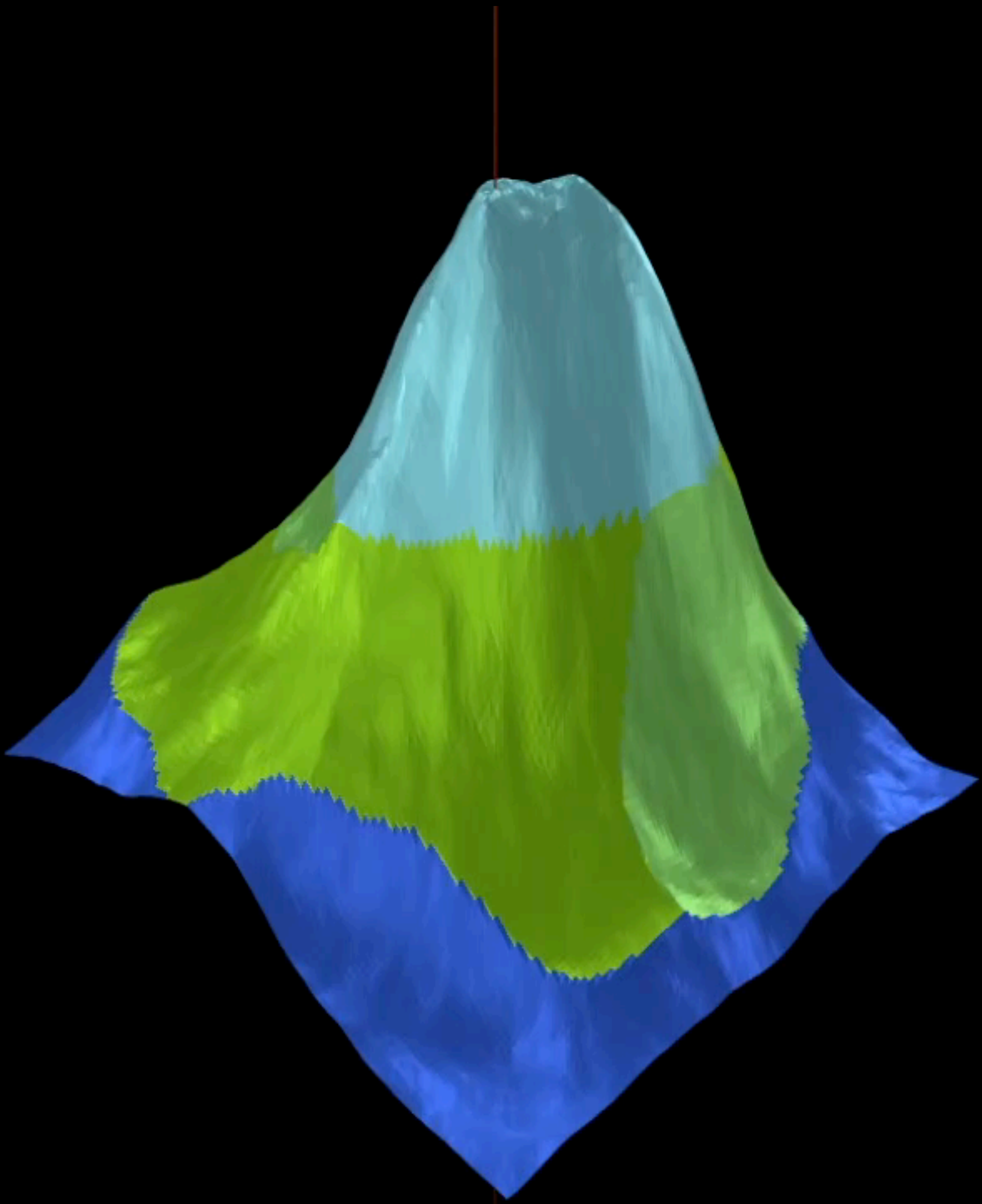
Visualizations

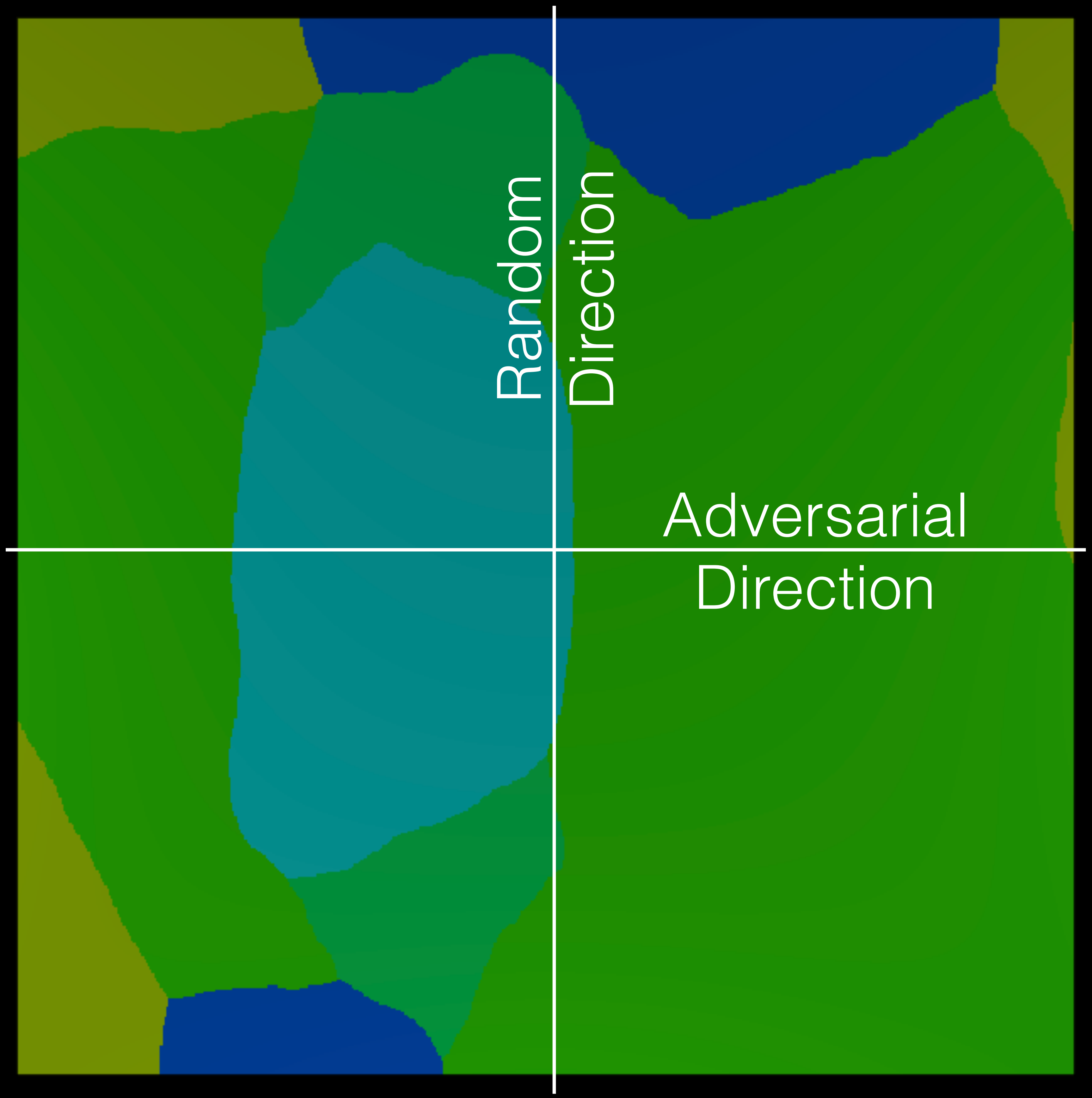

Random Direction
Random Direction

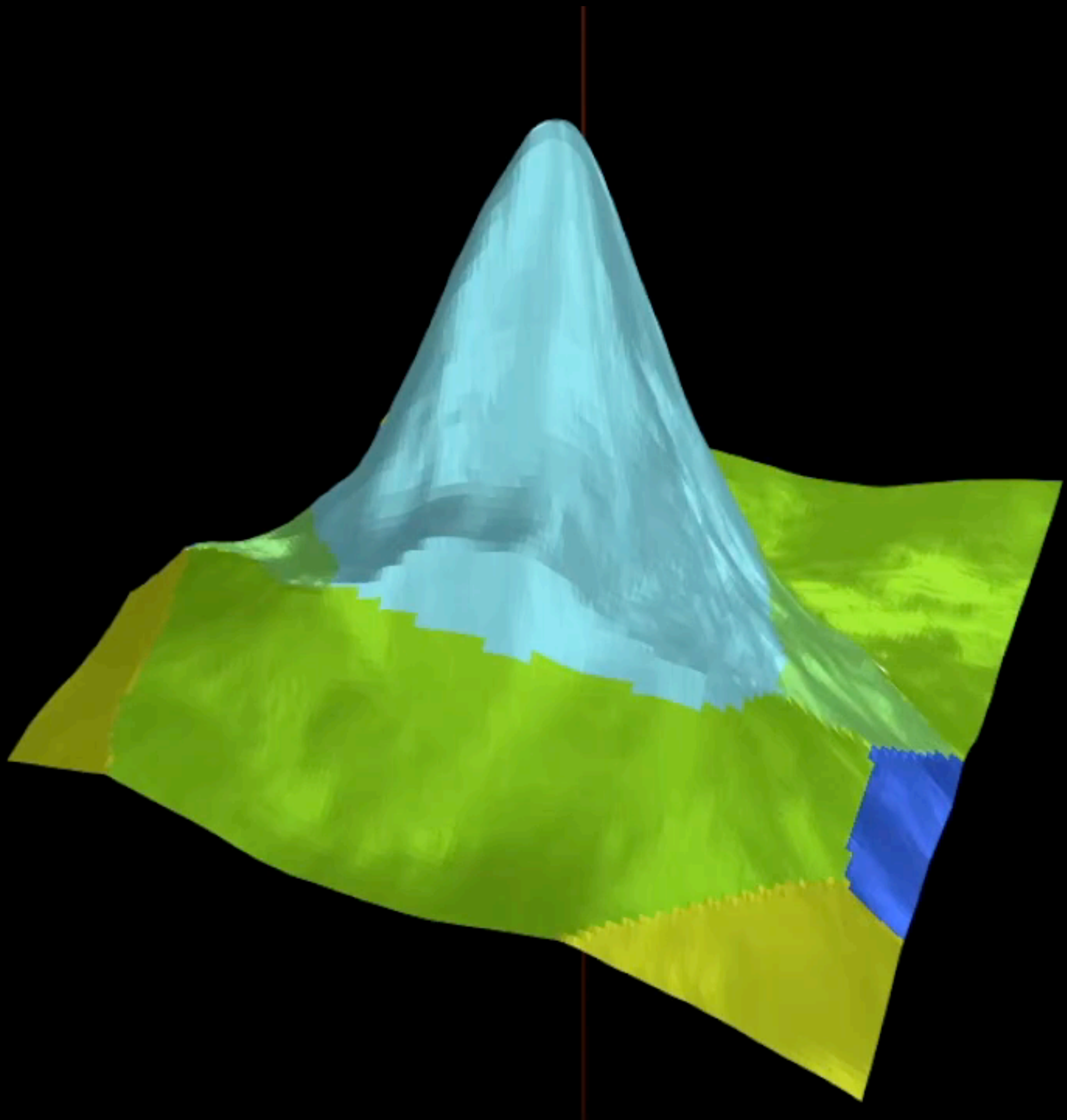


Random Direction
Random Direction


Random Direction
Random Direction




Random Direction
Adversarial Direction



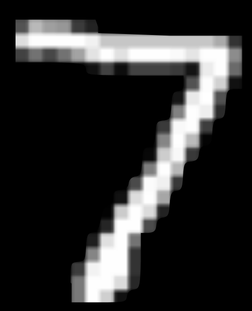
Case studies on evaluating defenses to adversarial examples

Defense Idea #1:

Additional Neural Network Detection

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischo.
On Detecting Adversarial Perturbations. ICLR 2017.

Normal Classifier



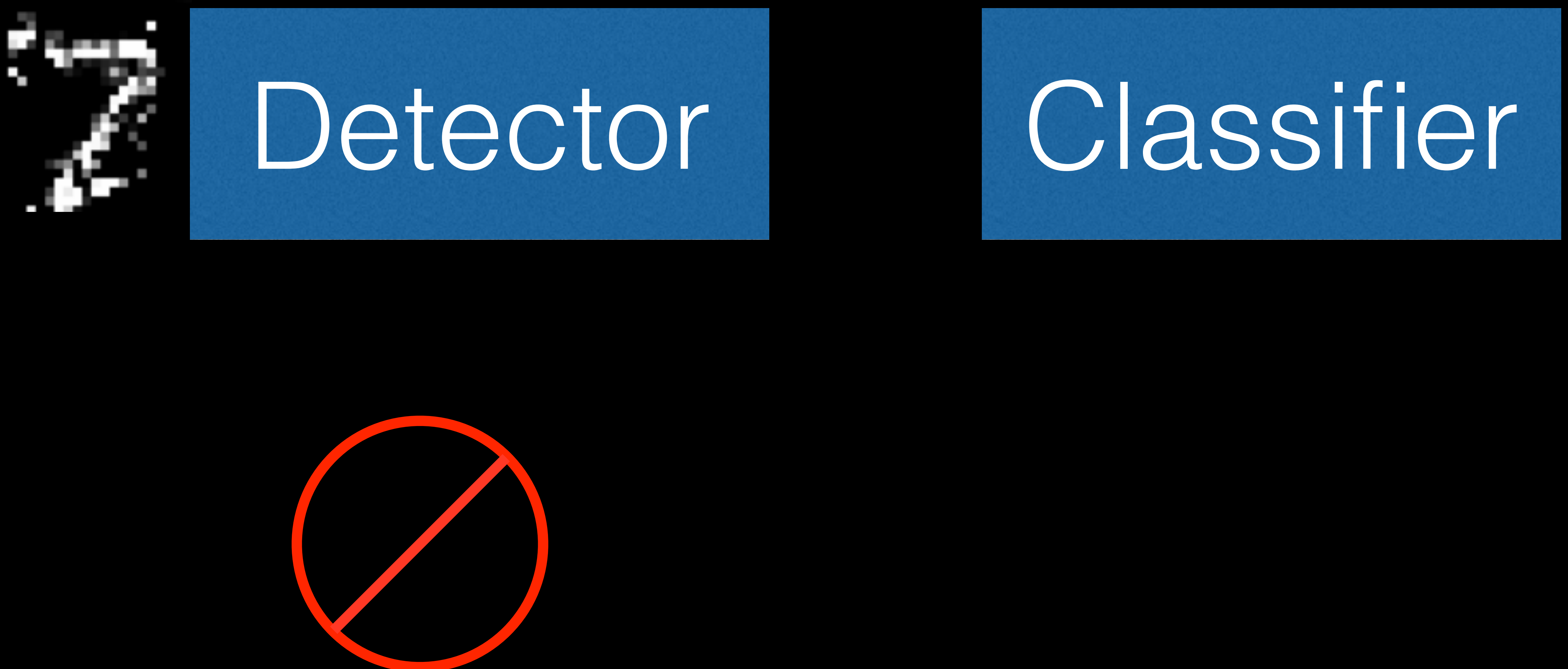
Classifier

Normal Classifier

Classifier

Detector & Classifier

Detector & Classifier



Training an adversarial example detector

Normal Training

(, 7)

(, 3)

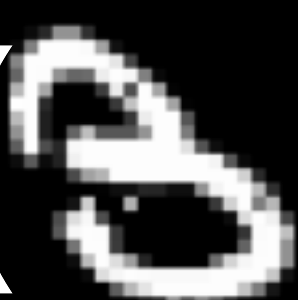
Training

Detection Training (1)

(, 7)

(, 3)

(, n)

(, n)

Attack

Detection Training (2)

( , y)

( , y)

( , n)

( , n)

Training

Sounds great.

Sounds great.

But we already know it's easy to fool neural networks ...

... so just construct adversarial examples to

1. be misclassified
2. not be detected

Breaking Detection Adversarial Training

- minimize $d(x,x') + g(x')$
  such that x' is "valid"

- Old: $g(x')$ measures loss of **classifier** on x'

Breaking Detection Adversarial Training

- minimize $\quad d(x,x') + g(x') + h(x')$
  such that $\quad x'$ is "valid"

- Old: $g(x')$ measures loss of **classifier** on x'

- New: $h(x')$ measures loss of **detector** on x'

Original

Adversarial
(unsecured)

Adversarial
(with detector)

2018

Defense Idea #2:

Thermometer Encoding

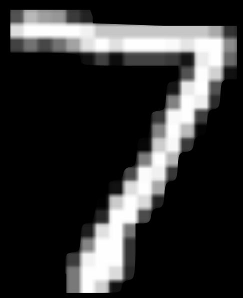
Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow.
Thermometer encoding: One hot way to resist adversarial examples. ICLR 2018.

Problem:
Neural Networks are "overly linear"

Thermometer Encoding

- Break linearity by changing input representation
- 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

Standard Neural Network



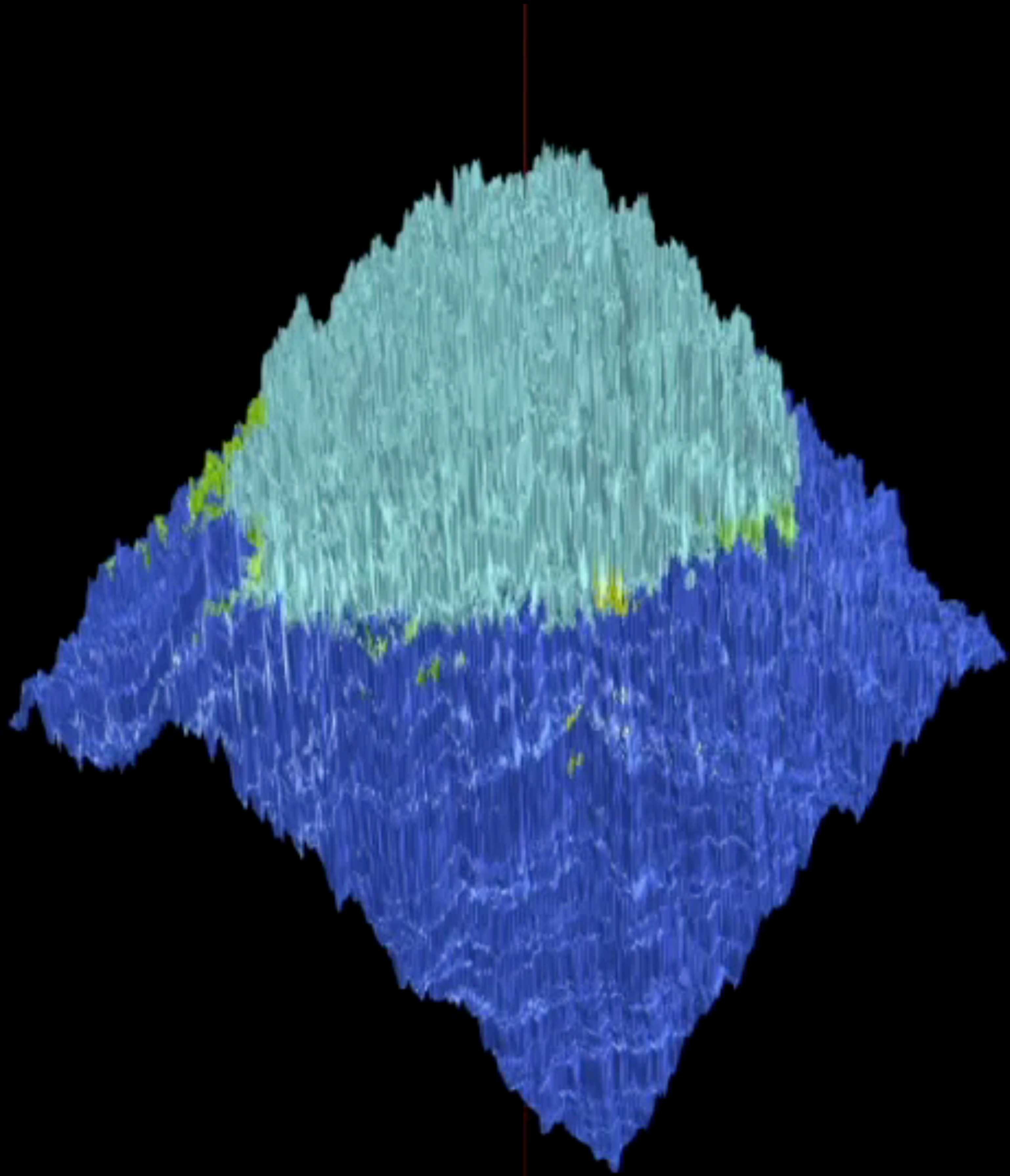
With Thermometer Encoding

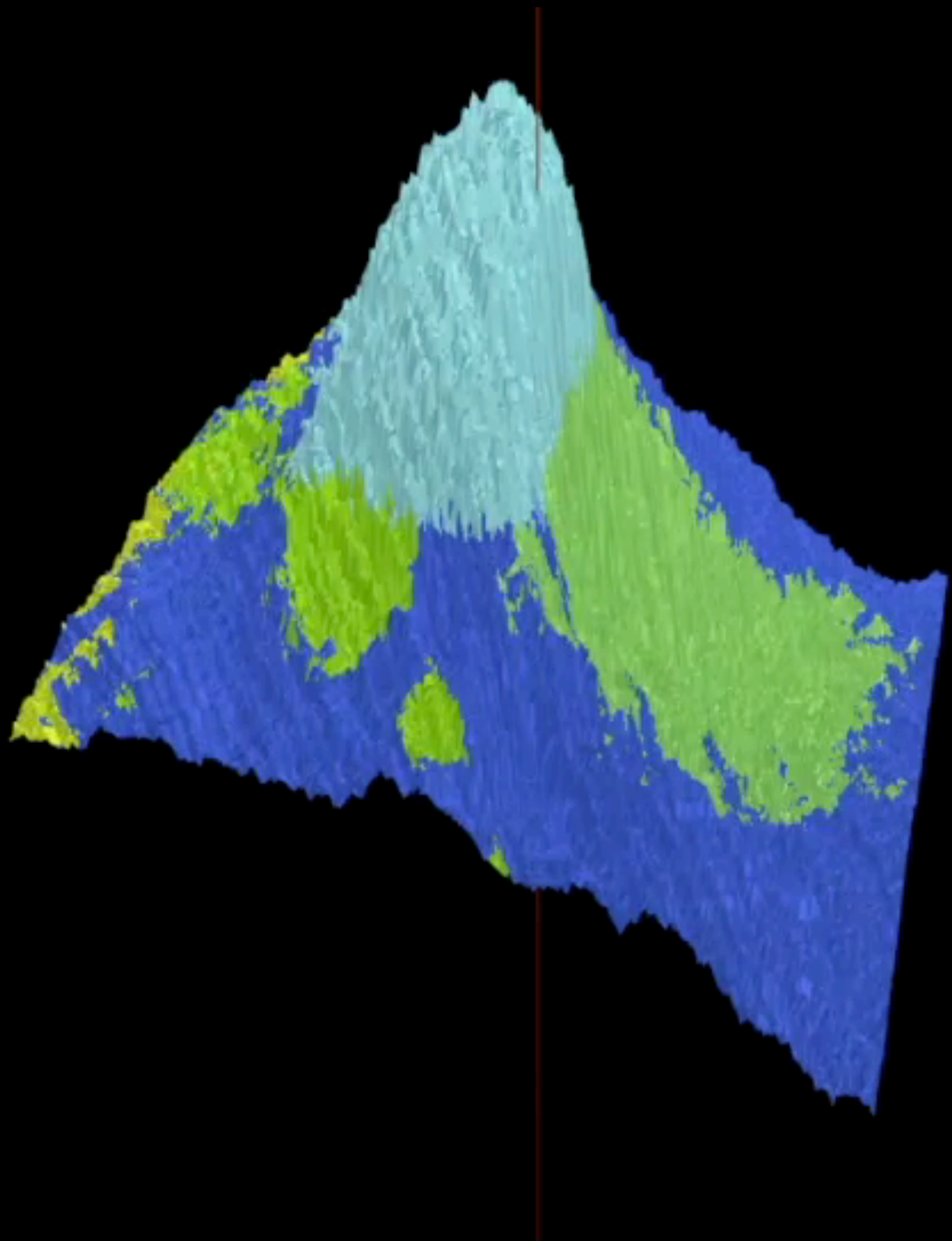
Claims:

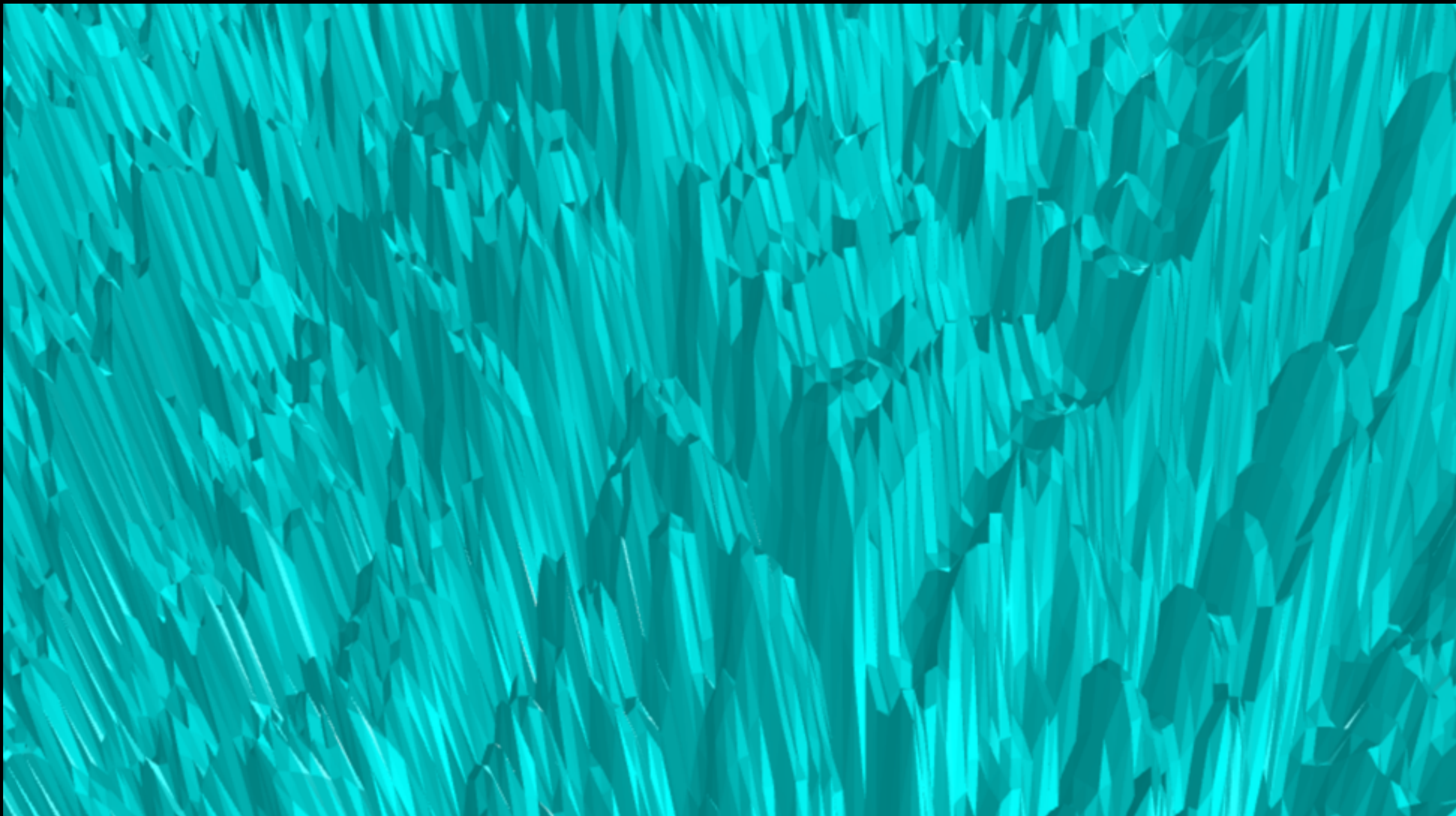
On CIFAR,
with distortion 8/255,
accuracy of 50%

(compared to 0%)

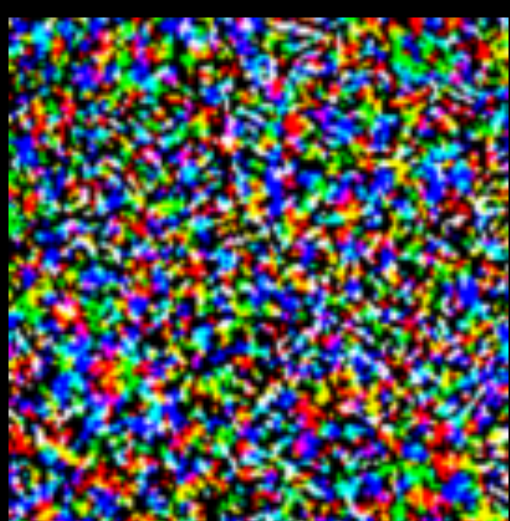
Unfortunately, thermometer encoding only causes gradient descent to fail

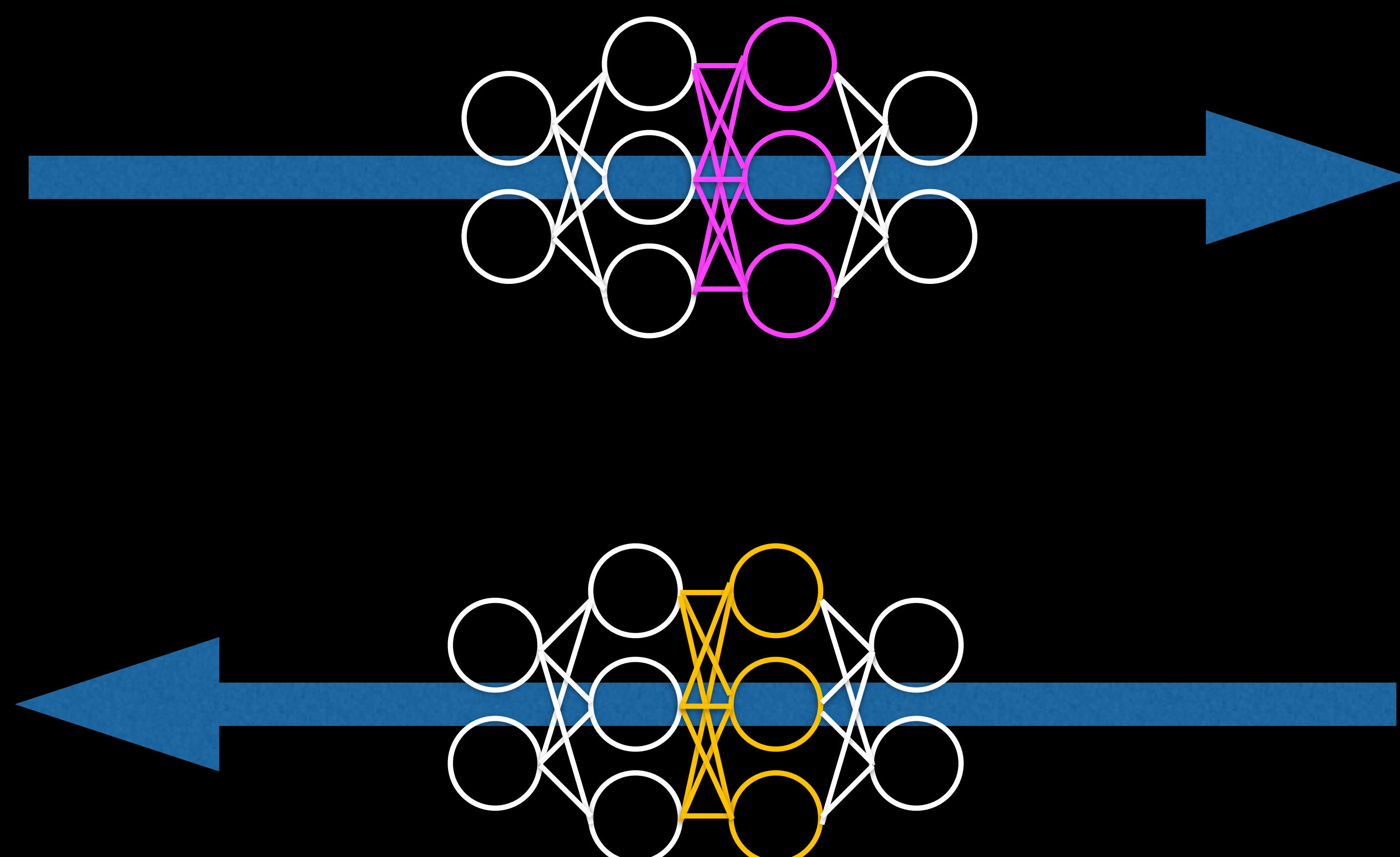




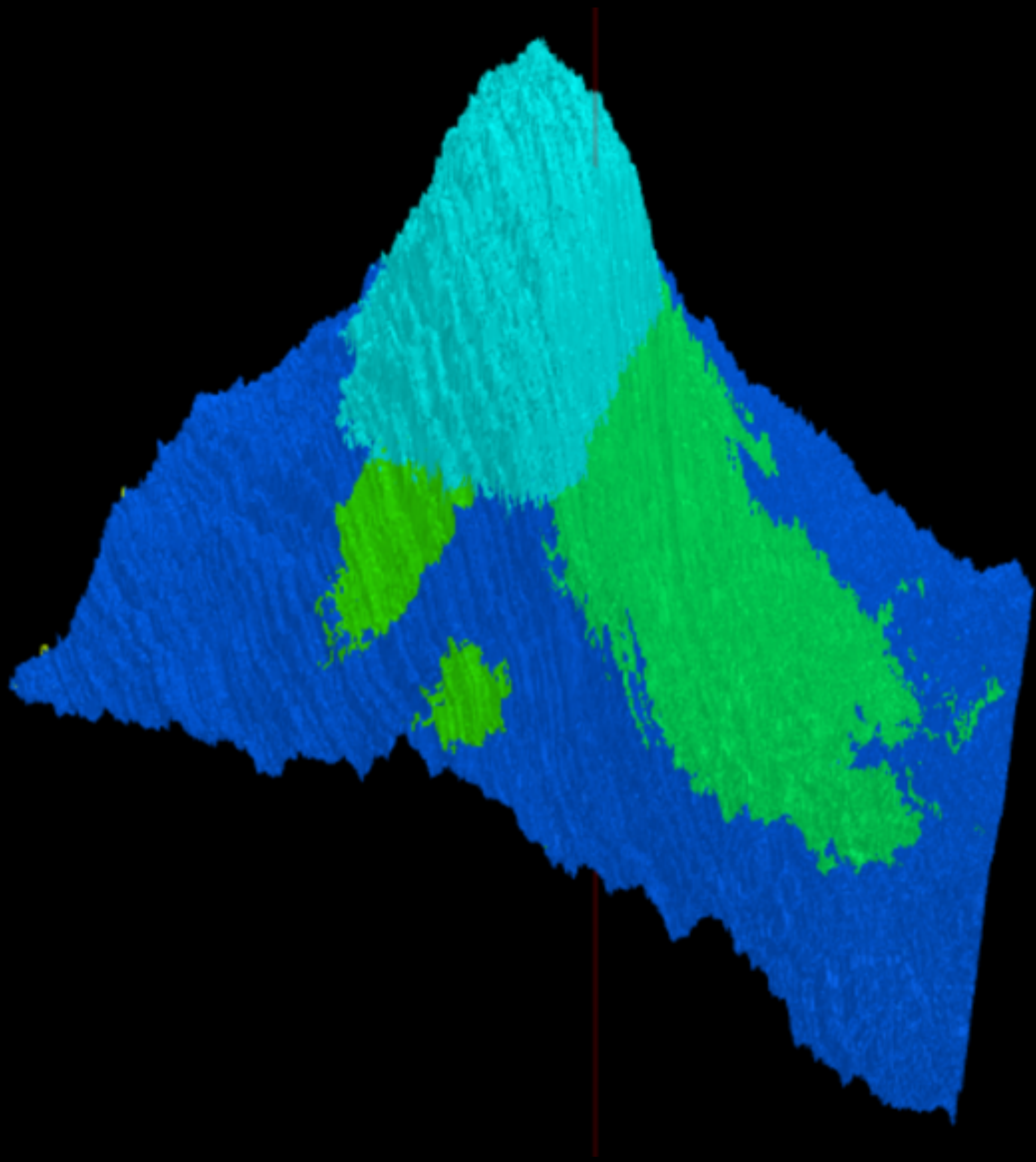


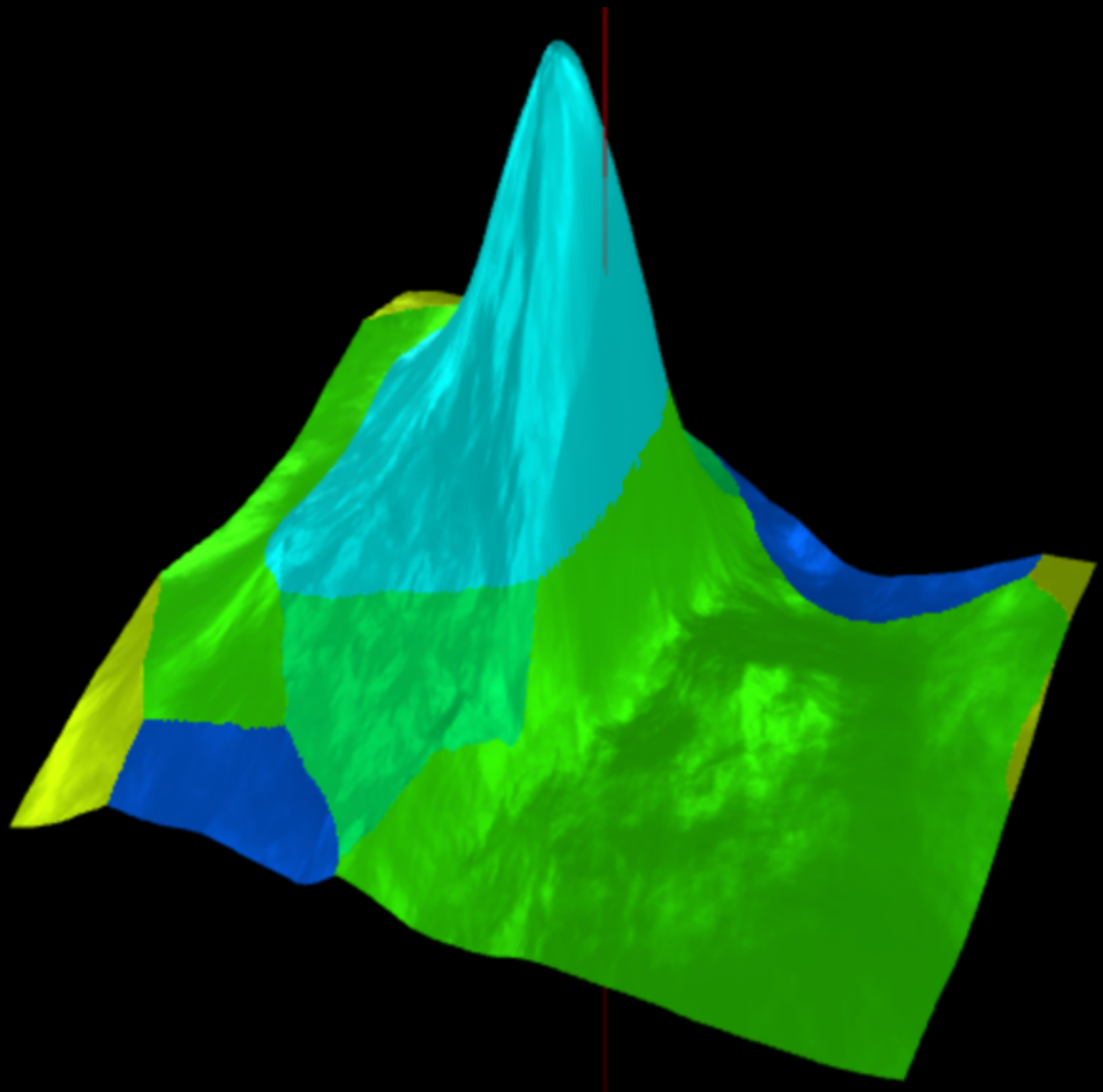
"Fixing" Gradient Descent





[0.1,
0.3,
0.0,
0.2,





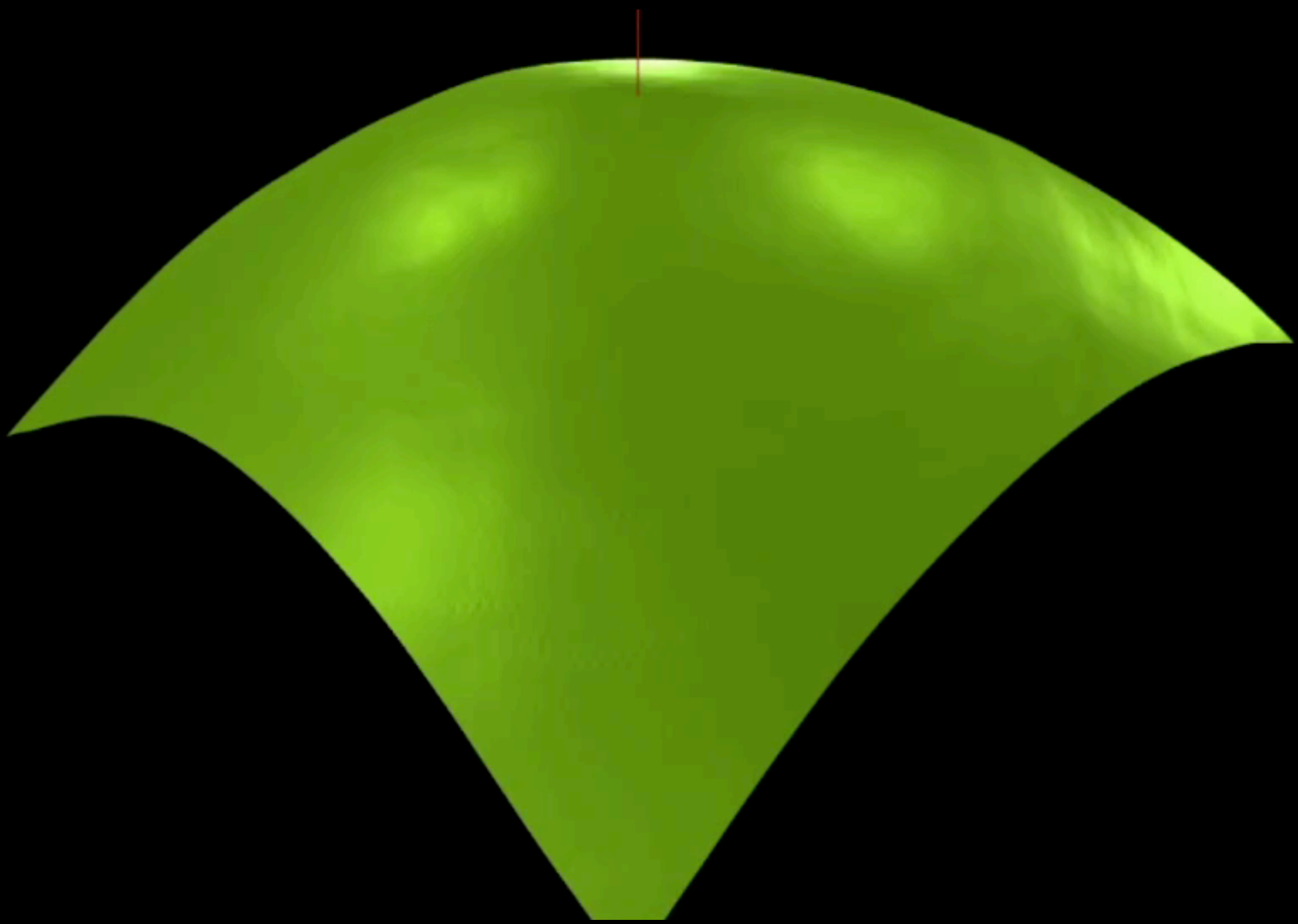
Defense Idea #3:

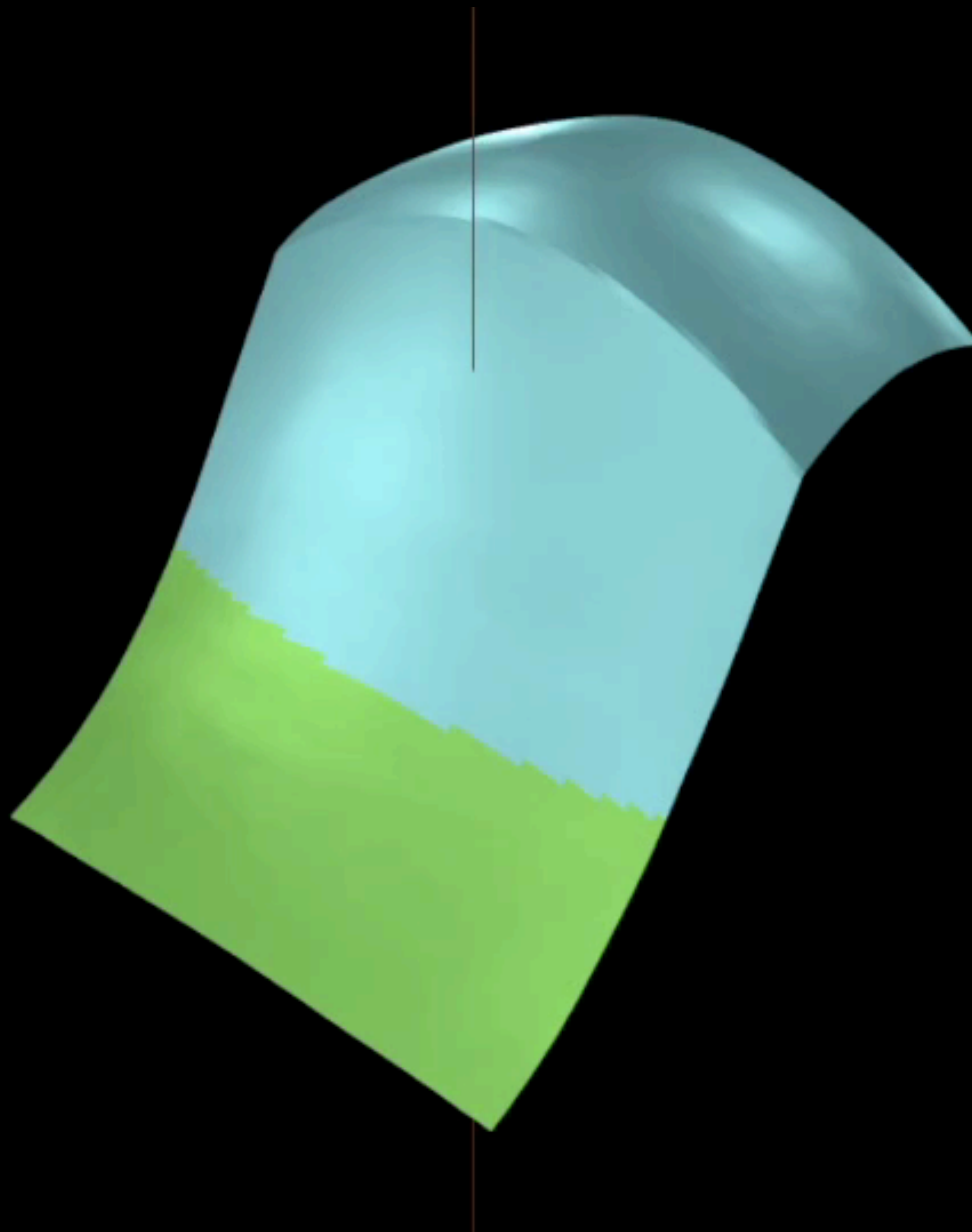
Adversarial Retraining

A Madry, A Makelov, L Schmidt, D Tsipras, and A Vladu. Towards deep learning models resistant to adversarial attacks. 2018. International Conference on Learning Representations.

Adversarial Training

- Given training data (X,Y)
- Sample a minibatch (x,y)
- Generate the adversarial minibatch (x',y)
- Train on (x',y)
- Repeat until convergence





Audio has these same issues, too

N Carlini and D Wagner. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text". 2018.

"now I would drift gently off to dream land"

[adversarial]

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 original or adversarial?

original or adversarial?

On audio, traditional ML methods are not vulnerable to adversarial examples

Questions?

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