# Obfuscated Gradients Give a False Sense of Security:

## Circumventing Defenses to Adversarial Examples

Anish Athalye*[1], **Nicholas Carlini***[2] , and David Wagner[3]

[1] Massachusetts Institute of Technology
[2] University of California, Berkeley (now Google Brain)
[3] University of California, Berkeley

**How** and **Why**

Act I

# **Background:**
# Adversarial Examples for Neural Networks

88% tabby cat



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_**

13 total defense papers at ICLR'18

9 are *white-box, non-certified*

6 of these are broken
(~0% accuracy)
1 of these is partially broken

**How** did we evade them?

**Why** we able to evade them?

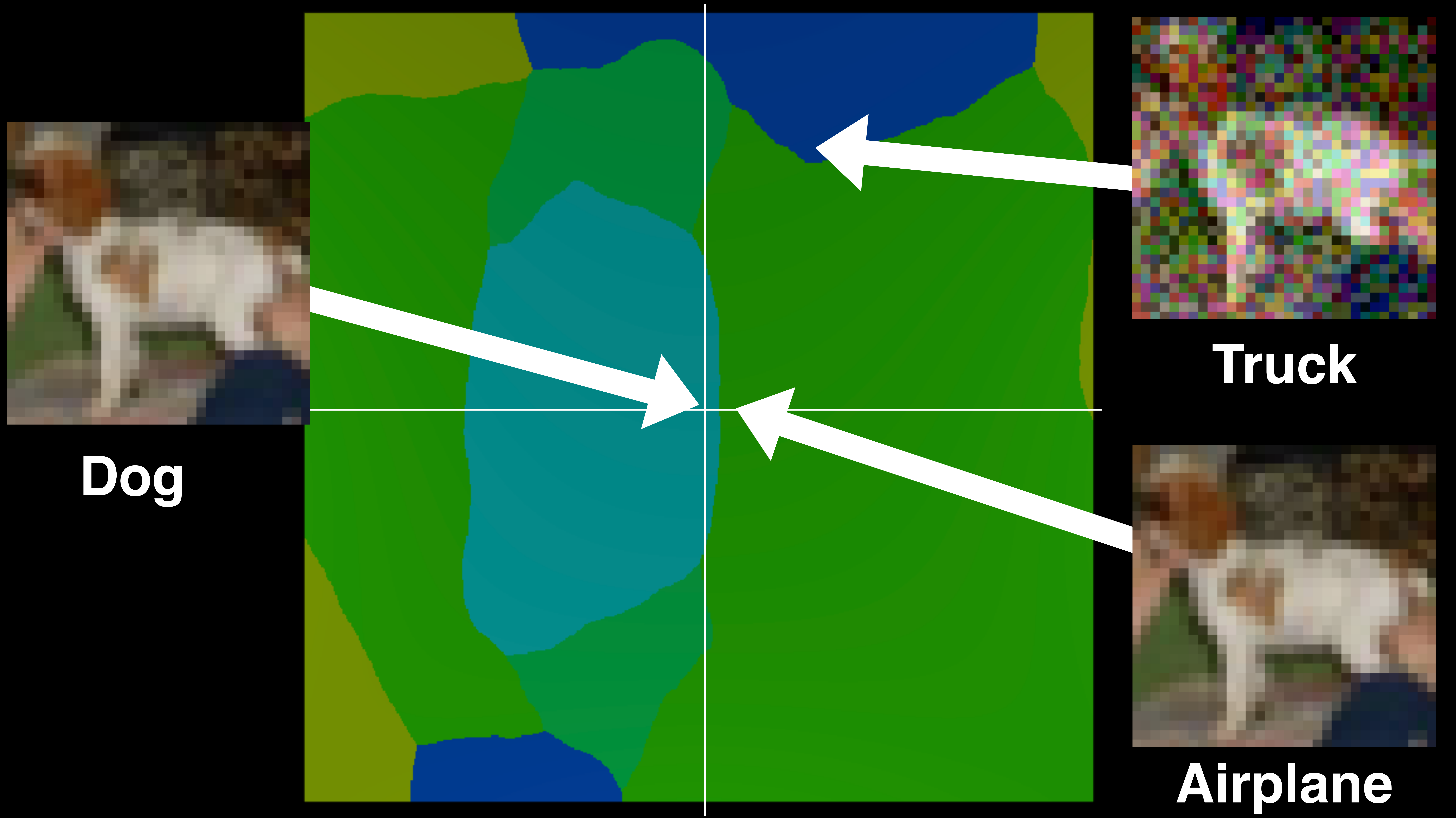
Act II

# HOW:
# Our Attacks

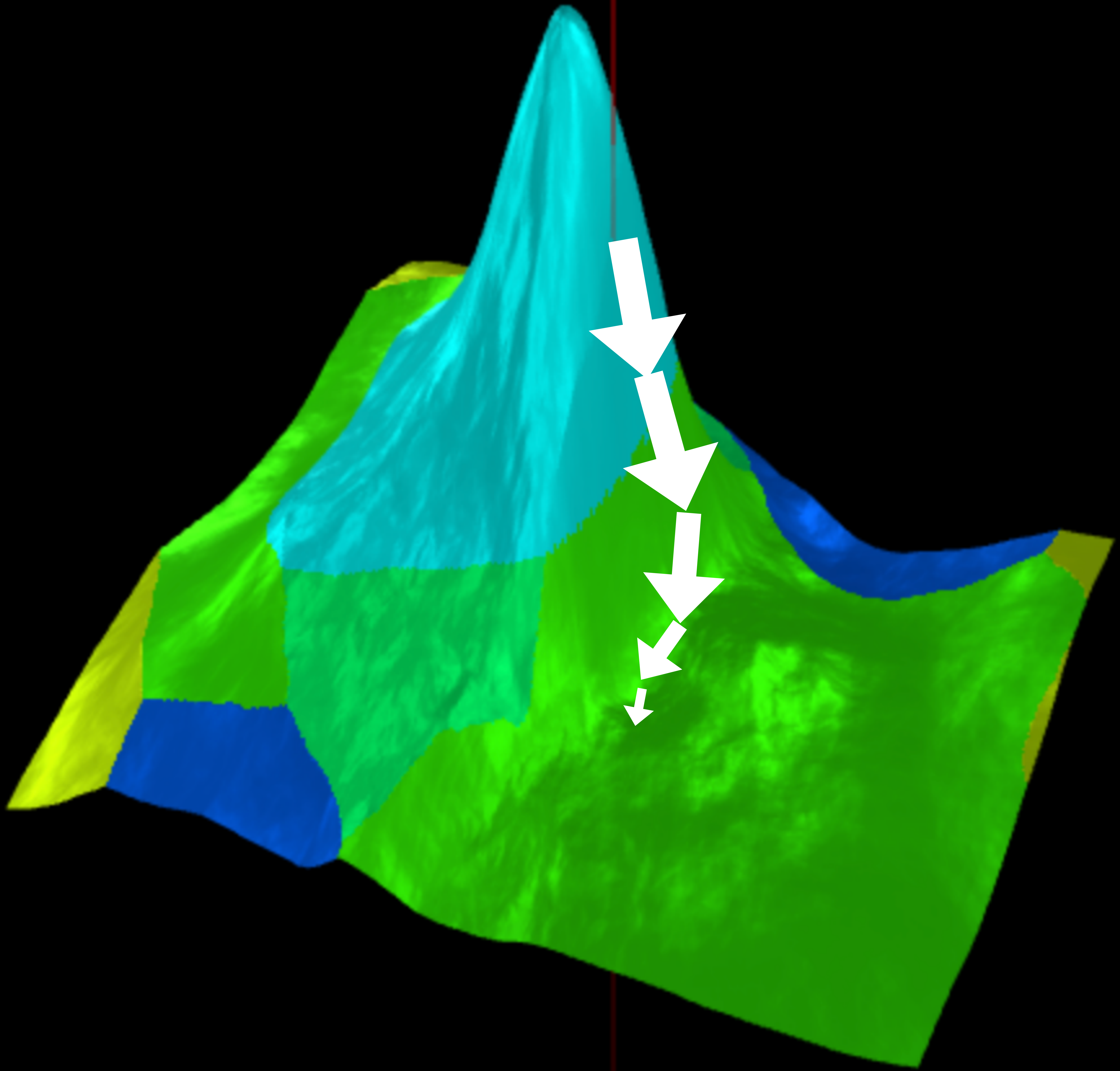
*How* do we generate adversarial examples?

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

*Why* can we generate adversarial examples (with gradient descent)?

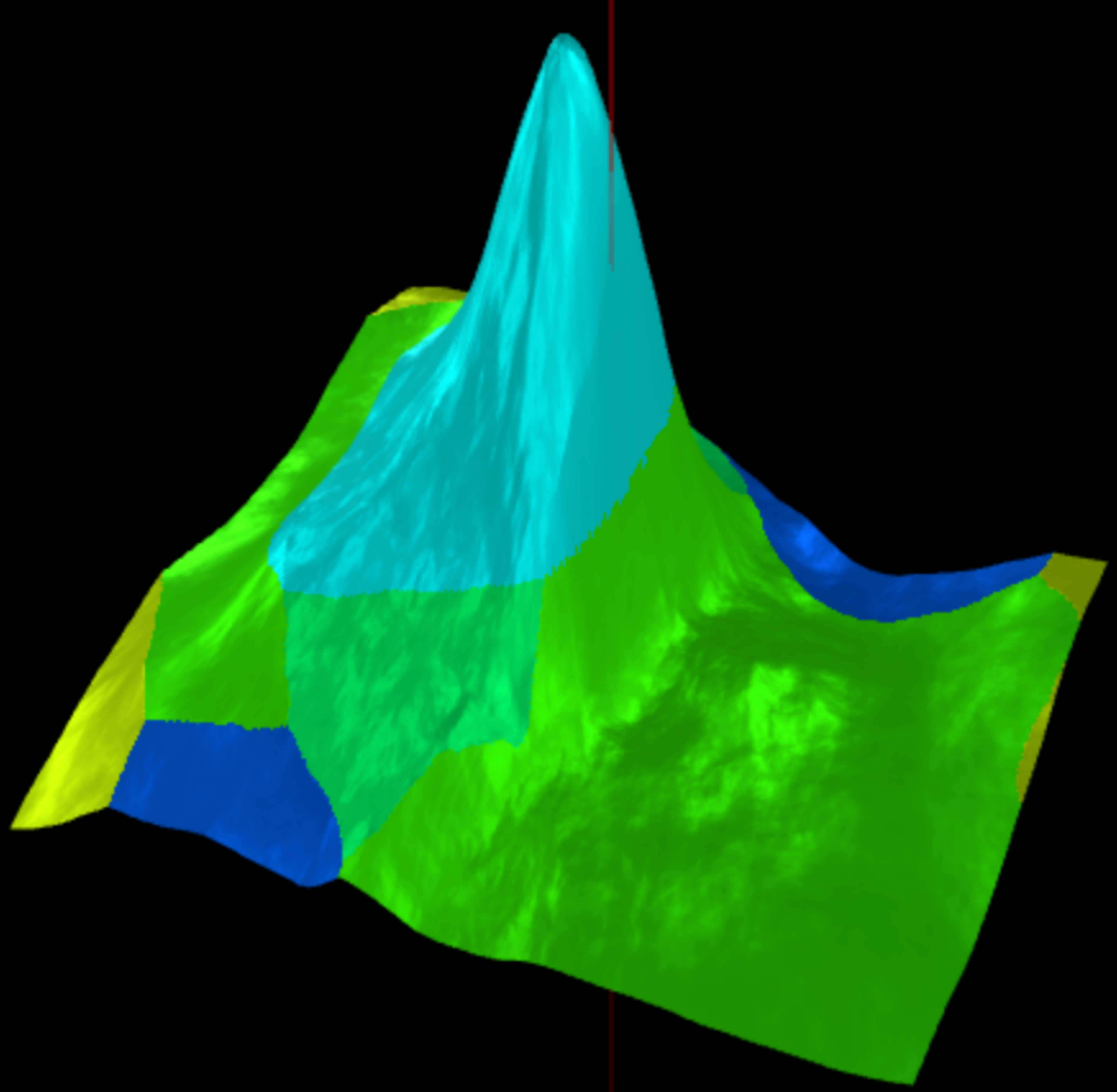


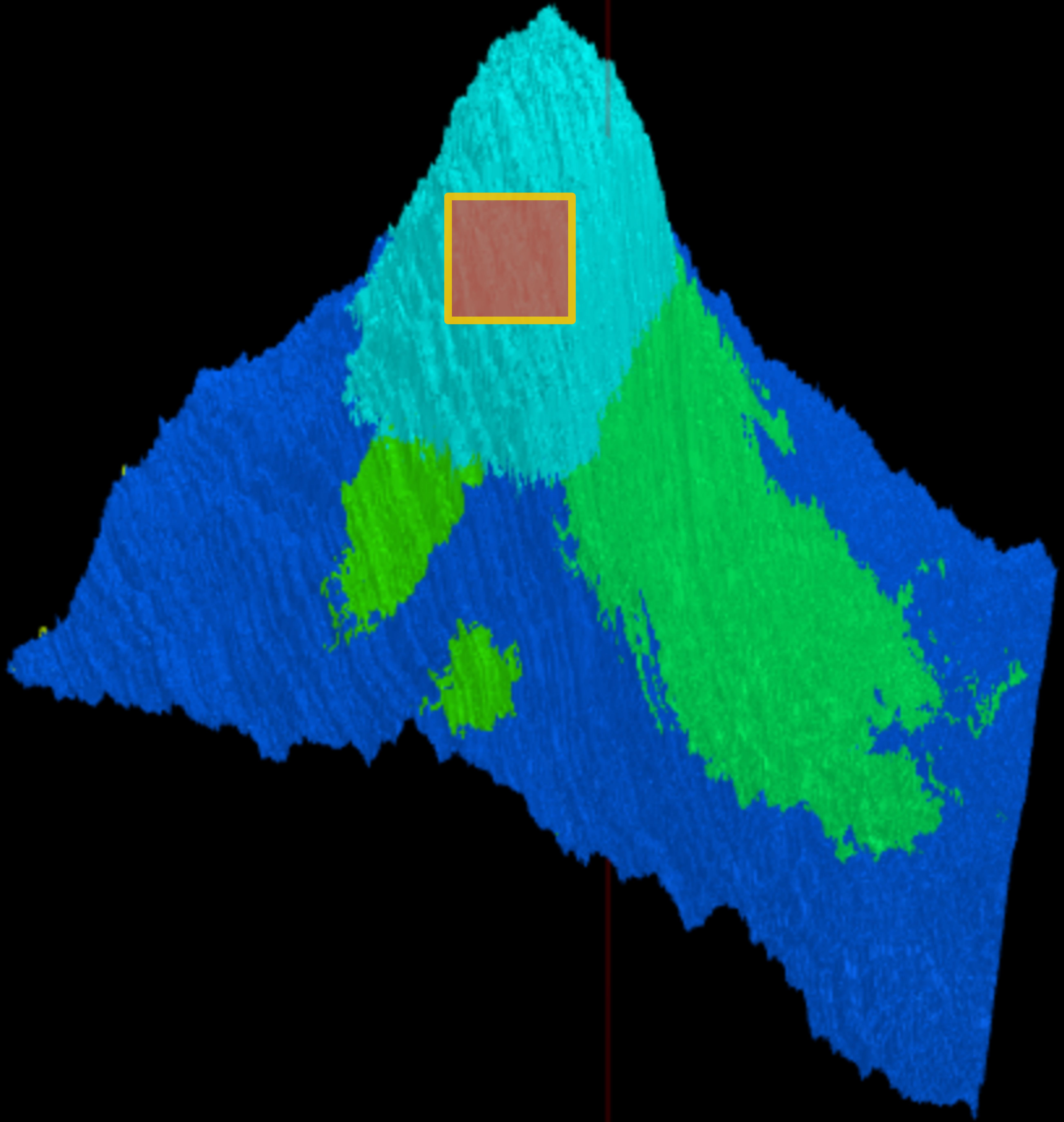
Truck
Dog
Airplane

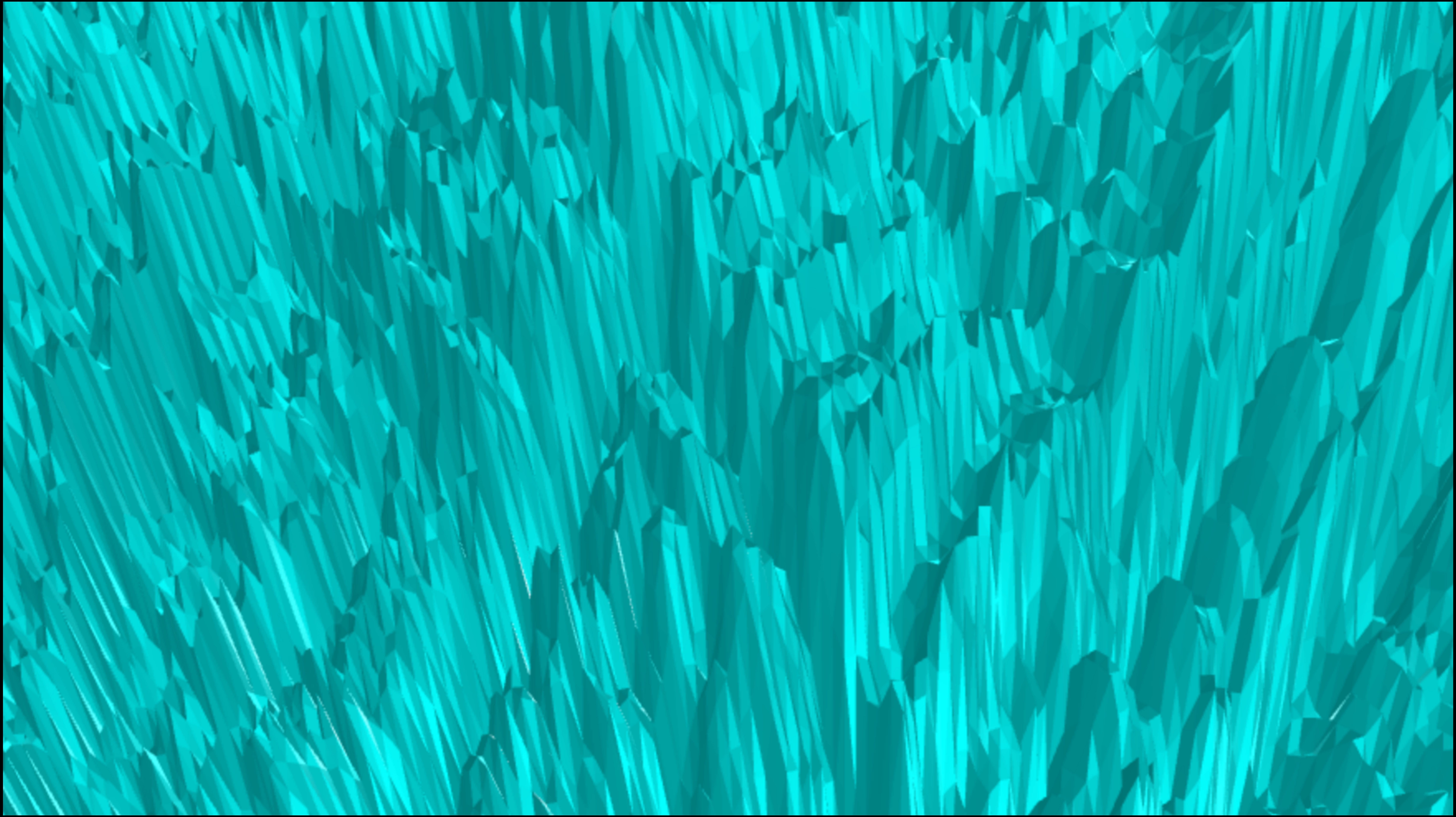


We find that 7 of 9 ICLR defenses rely on the same artifact:

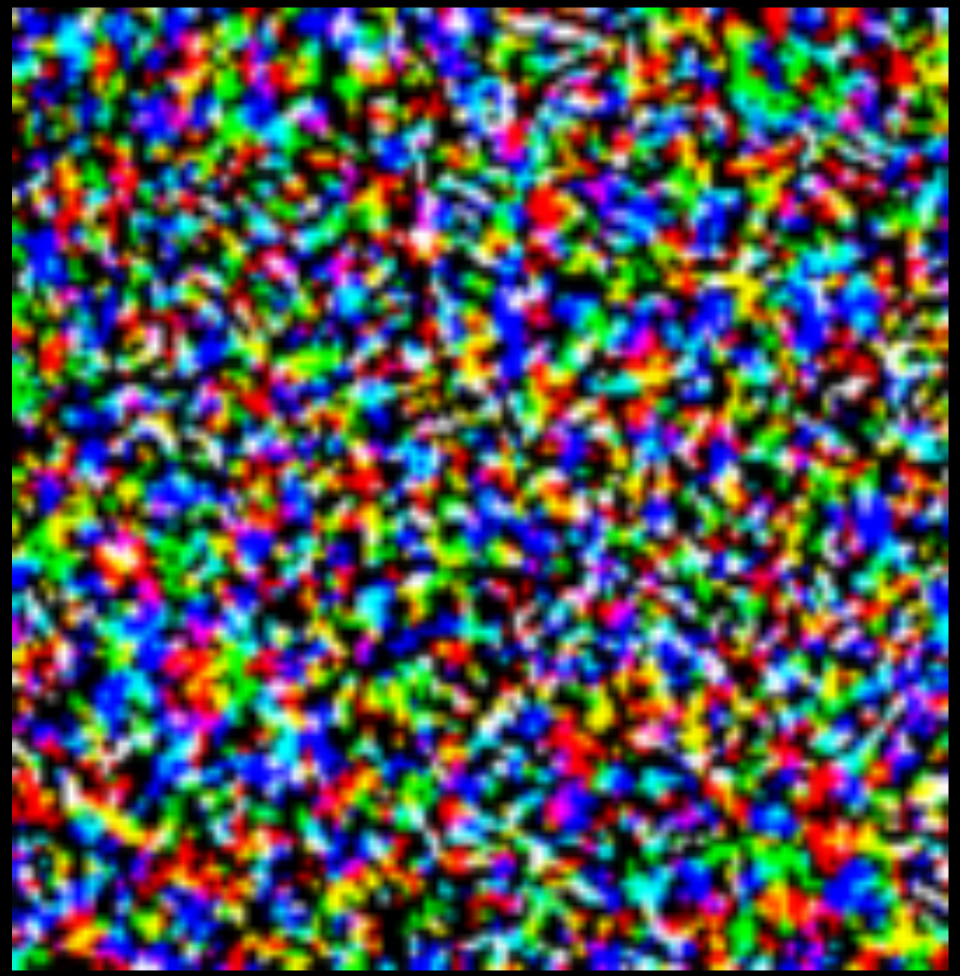
**obfuscated gradients**

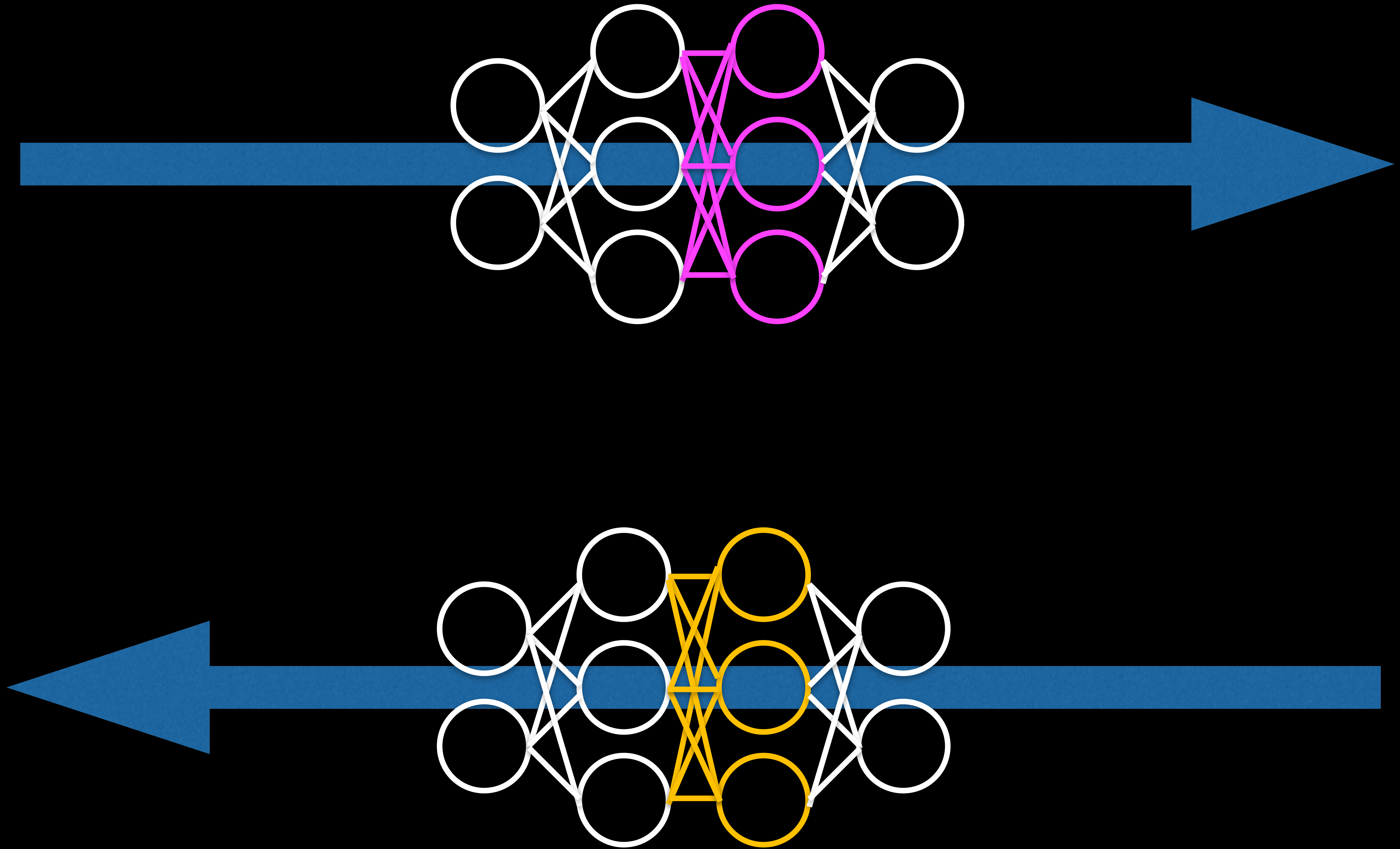






# "Fixing" Gradient Descent

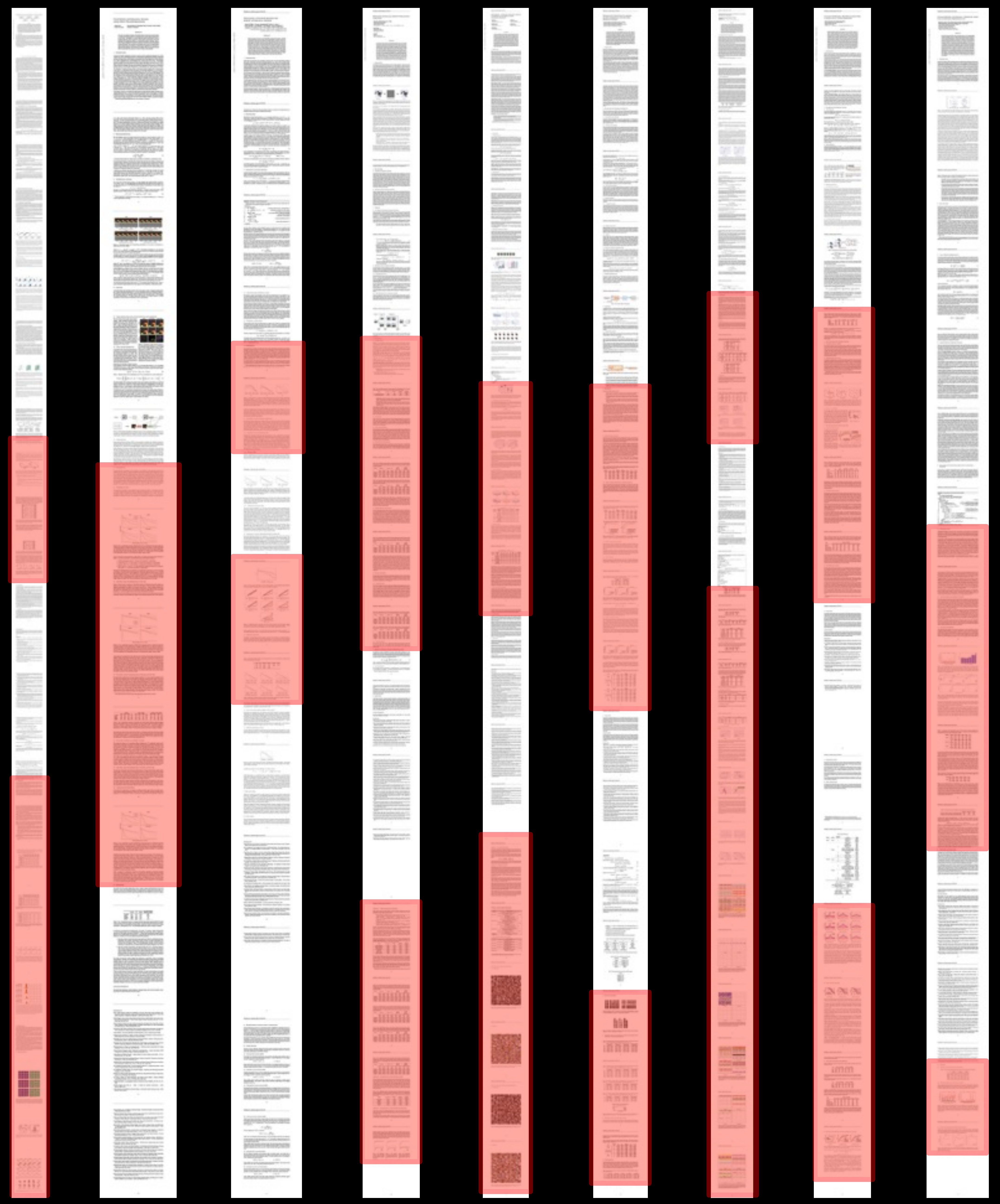




[0.1,
0.3,
0.0,
0.2,
0.4]

Act III

# WHY: Evaluation Methodology



Serious effort to evaluate

By space, most papers are ½ evaluation

What went wrong then?

```
acc, loss = model.evaluate(
                x_test, y_test)
```

Is no longer sufficient.

There is no single test set for security

The only thing that matters is robustness against an adversary ***targeting the defense***

The purpose of a
defense evaluation is
**NOT** to show
the defense is **RIGHT**

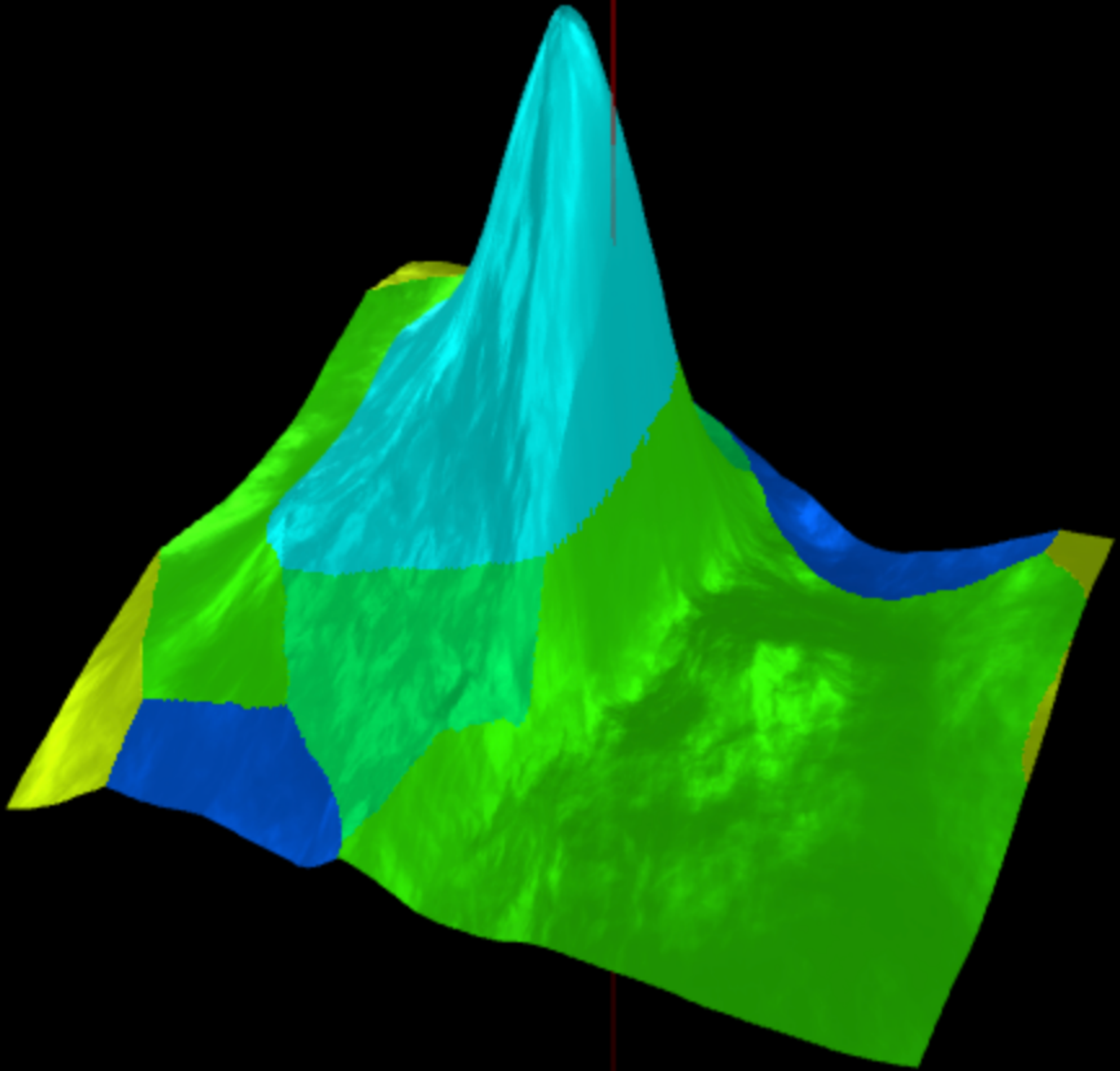
The purpose of a
defense evaluation is
to **FAIL** to show
the defense is **WRONG**

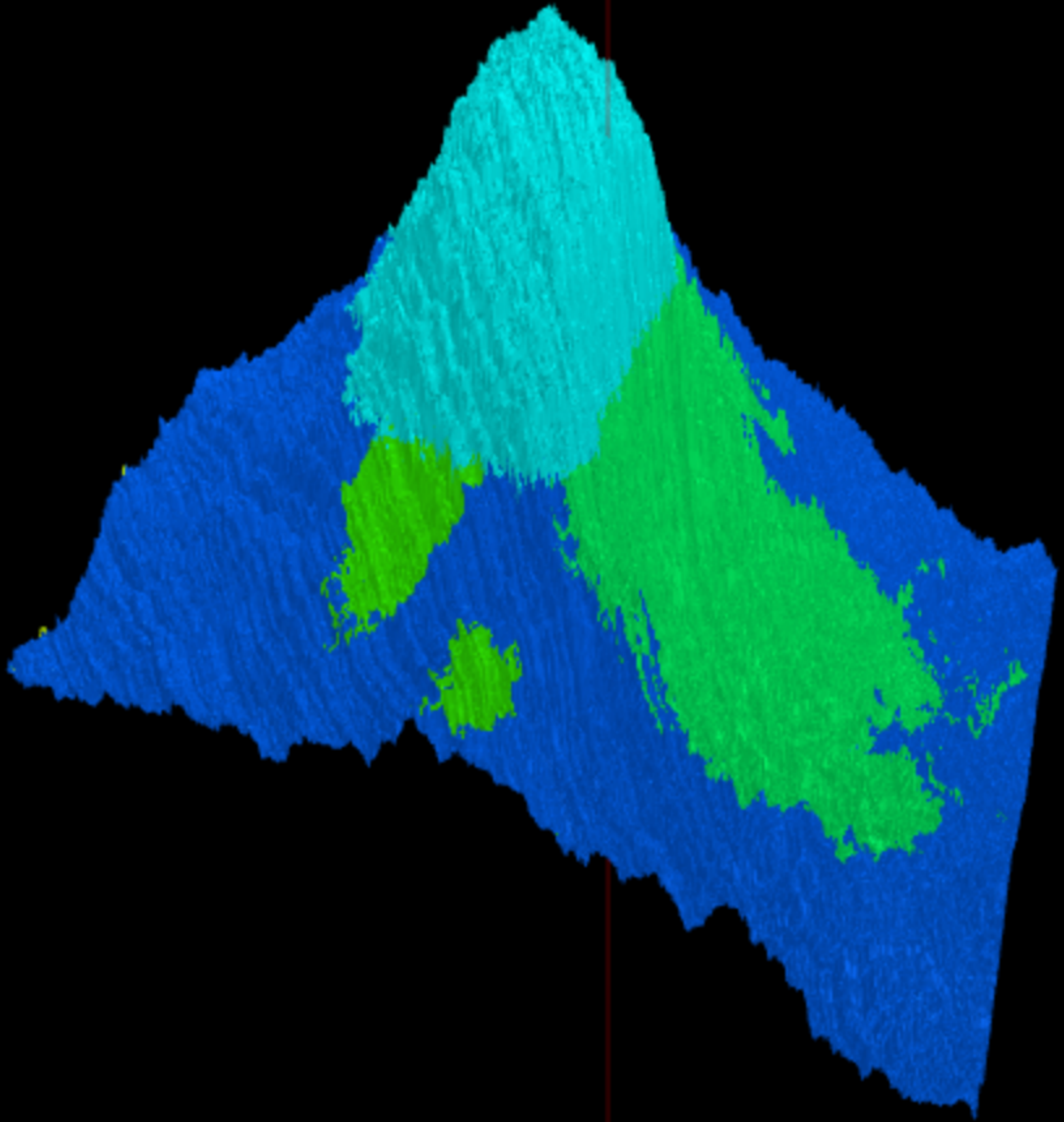
Act IV

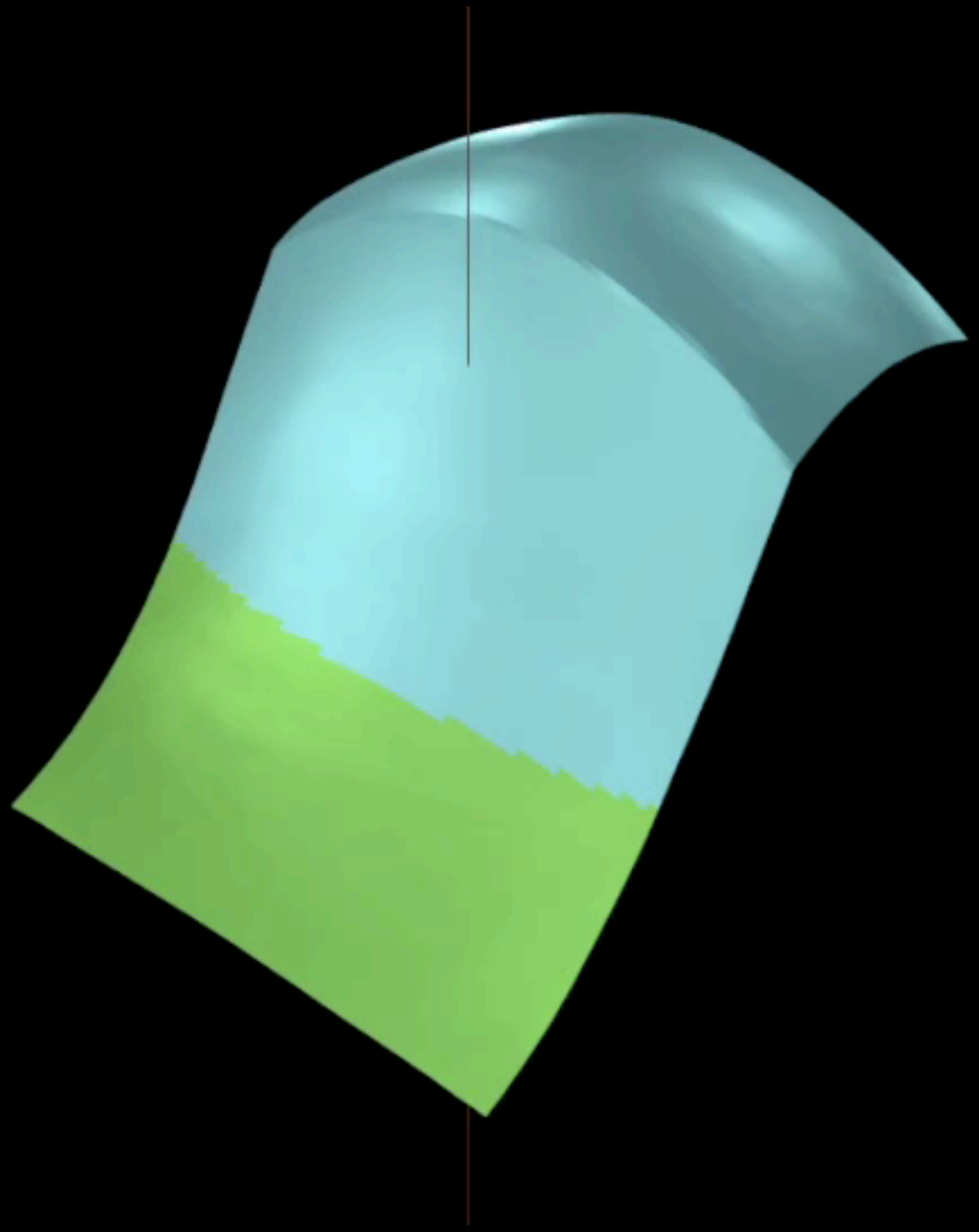
# Making & Measuring Progress

Strive for **simplicity**

over **complexity**







# What metric should we optimize?

# Threat Model

The set of assumptions
we place on the adversary

In the context of adversarial examples:

1. Perturbation Bounds & Measure
2. Model Access & Knowledge

The threat model **MUST** assume the attacker has read the paper and knows the defender is using those techniques to defend.

# Metrics for Success

Accuracy under
existing
threat models

More permissive
threat models

"making the attacker think more"
is **not** (usually) progress

The threat model doesn't limit
the attacker's approach

Act V

# Conclusion

A paper can only do so much in an evaluation.

A paper can only do so much in an evaluation.

We need more re-evaluation papers.

# So you want to build a defense?

*"Anyone, from the most clueless amateur to the best cryptographer, can create an algorithm that he himself can't break."*

-- Bruce Schneier

# So you want to build a defense?

As a corollary: learn to break defenses **before** you try to build them

If you can't break the state-of-the-art, you are unlikely to be able to build on it

# Challenging Suggestions

**Defense-GAN** on **MNIST**

We were able to break it only partially

Samangouei *et al*. 2018 ("Defense-GAN...")

**"Strong" Adversarial Training** on **CIFAR**

We were not able to break it at all

Madry *et al.* 2018 ("Towards Deep...")

Visit our **poster** & originally scheduled **talk**
(Today, #110) & (Tomorrow, A7 @ 2:50)

**Email us**

Anish: aathalye@mit.edu
Me: nicholas@carlini.com

**Track Progress**
robust-ml.org

**Source Code**
git.io/obfuscated-gradients

# Did we get it right?

1. We reproduced the original claims against the (weak) attacks initially attempted

2. We showed the papers authors' our results

3. It's possible we didn't. But our code is public: https://github.com/anishathalye/obfuscated-gradients

# Isn't this just gradient masking?

The short answer: **No**, if it were, we wouldn't have seen 7 of 9 ICLR defenses relying on it.

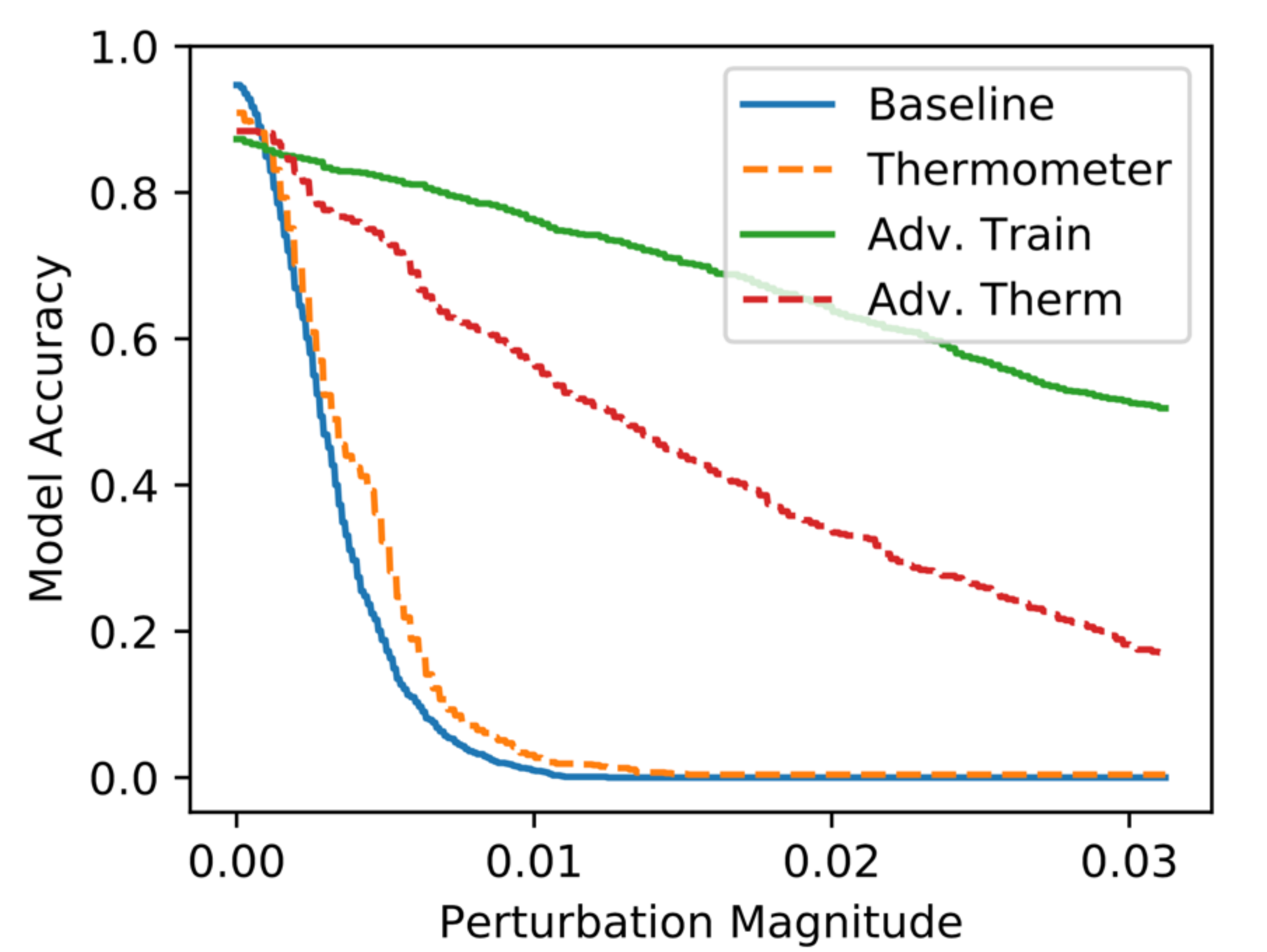
# X defense has multiple parts, but you only broke each part separately.

True. Usually, an ensemble several weaker defenses is not an effective defense strategy, unless there is an argument they cover each other's weaknesses.

He *et al.* "Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong". WOOT'18.

# Did you try X with adversarial training?

Not usually. In some cases the combination is *worse* than adversarial training alone



# Specific advice for performing evaluations

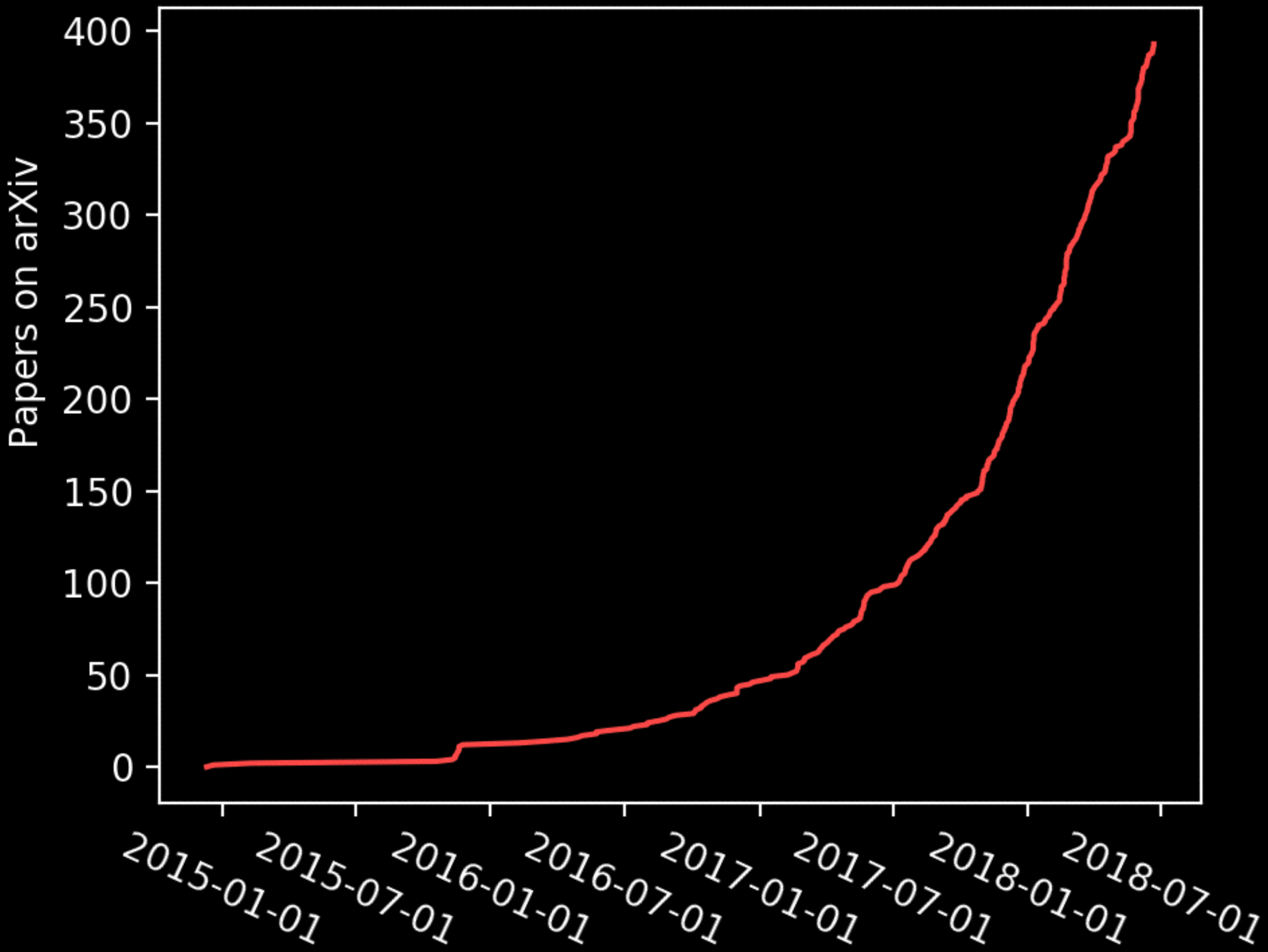
- Carlini *et al.* 2017 & S&P ("Towards Evaluating ...")
- Athalye *et al.* 2018 @ ICML ("Obfuscated ...")
- Madry *et al.* 2018 @ ICLR ("Towards Deep...")
- Uesato *et al.* 2018 @ ICML ("Adversarial Risk...")

Details in our originally-scheduled talk,
Tomorrow @ 2:50 in A7

There is a true notion of robustness, for a computationally unbounded adversary.

We are forced to **approximate** this.

*Adversarial Risk and the Dangers of Evaluating Against Weak Attacks.*
Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, Pushmeet Kohli.
ICML 2018.


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