ERIK JONSSON SCHOOL OF ENGINEERING AND COMPUTER SCIENCE
The University of Texas at Dallas

UT DALLAS

CS 6301.007

Machine Learning in Cyber Security

Wei Yang

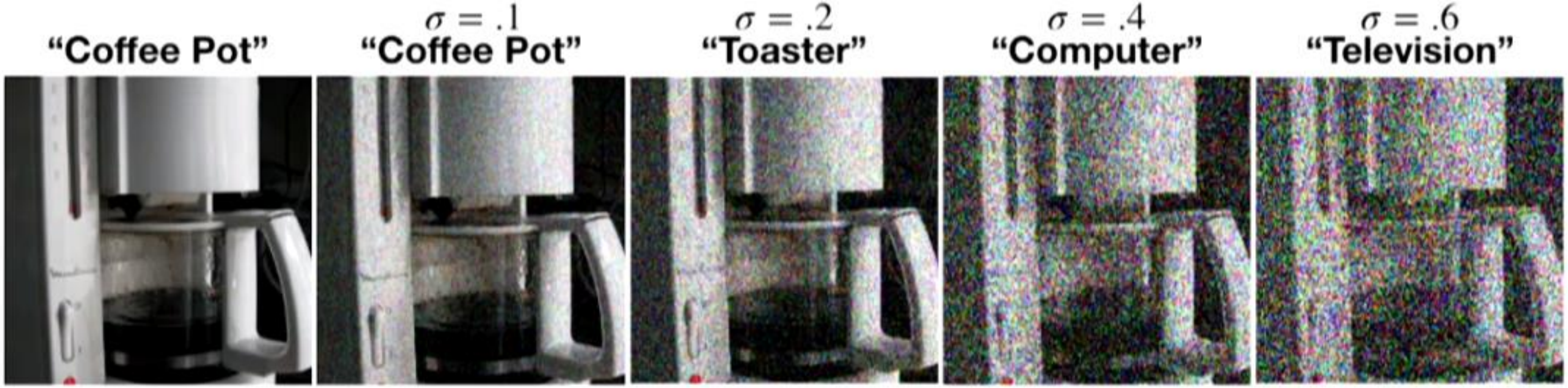
Department of Computer Science
University of Texas at Dallas

Credit: Justin Gilmer

Question

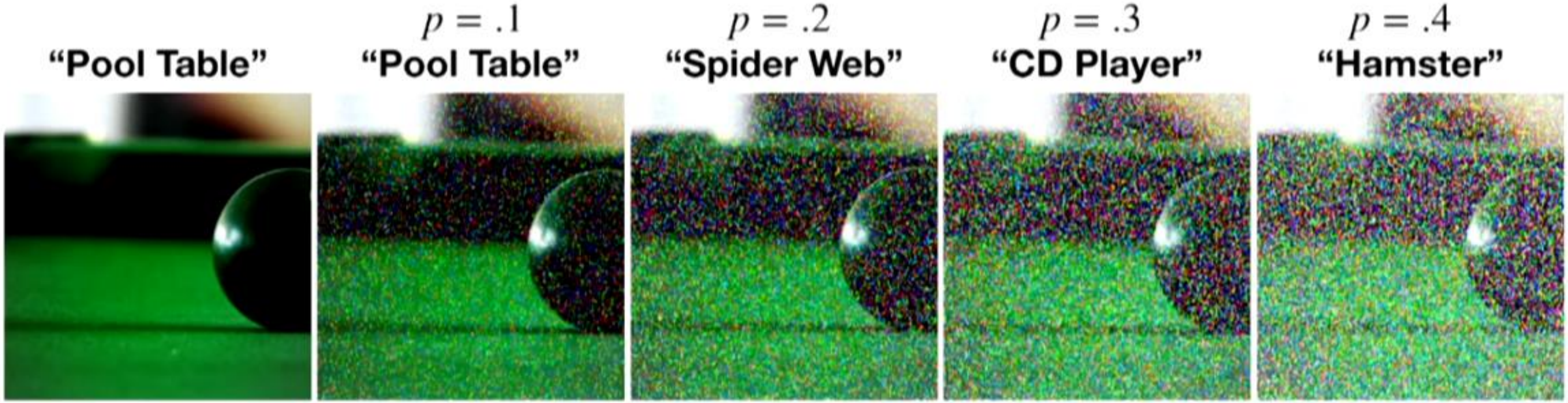
What is the difference between adversarial examples and test error?

Gaussian noise



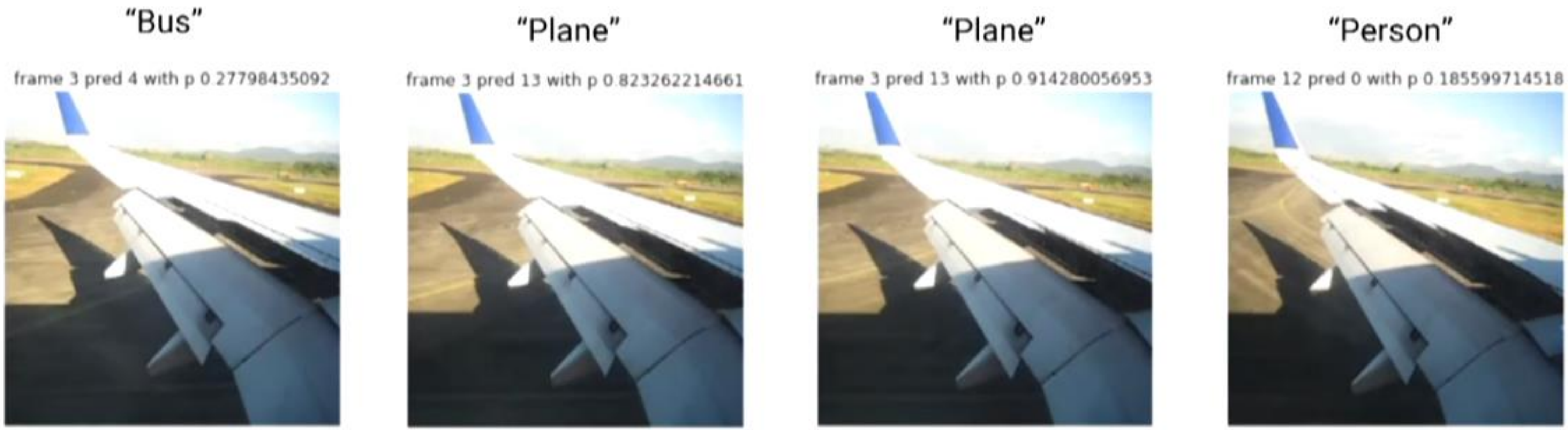
- InceptionV3: 13.2% test accuracy at sigma=.4
- (76.2% clean test accuracy)

Salt-and-pepper noise



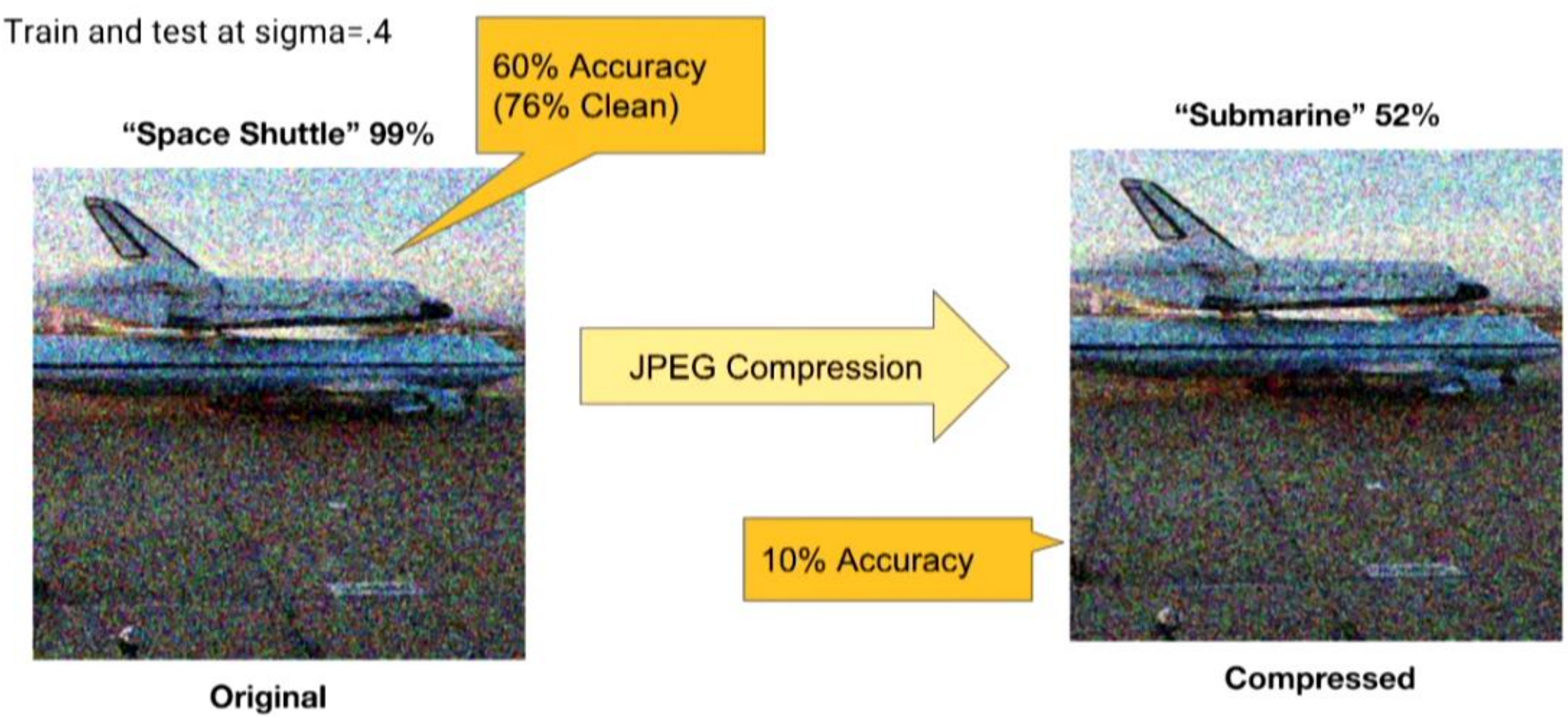
- InceptionV3: 5.4% test accuracy at p=.3

Video Robustness


"Bus"
frame 3 pred 4 with p 0.27798435092
"Plane"
frame 3 pred 13 with p 0.823262214661
"Plane"
frame 3 pred 13 with p 0.914280056953
"Person"
frame 12 pred 0 with p 0.185599714518

Thanks to Keren Gu for this example!

Naïve data augmentation doesn't help

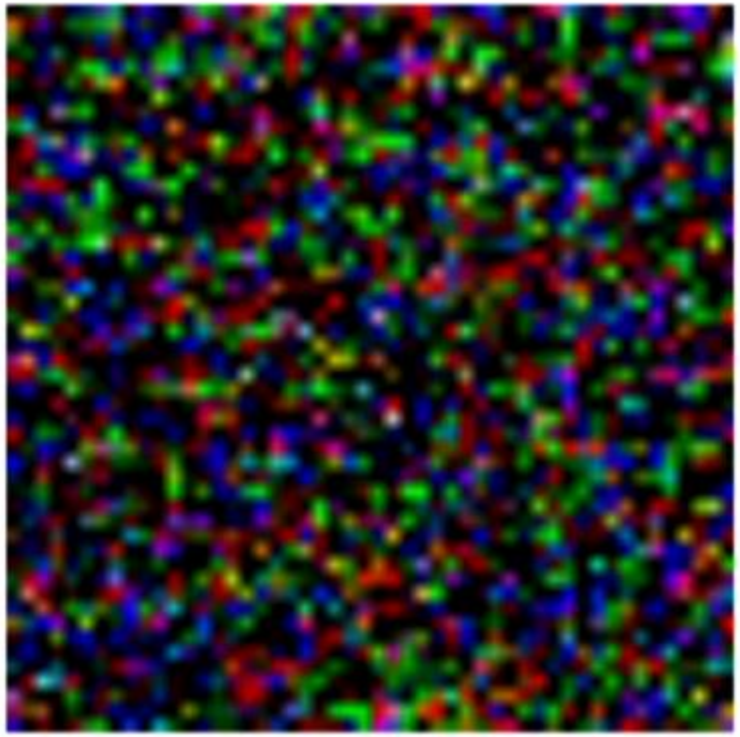


Train and test at sigma=.4

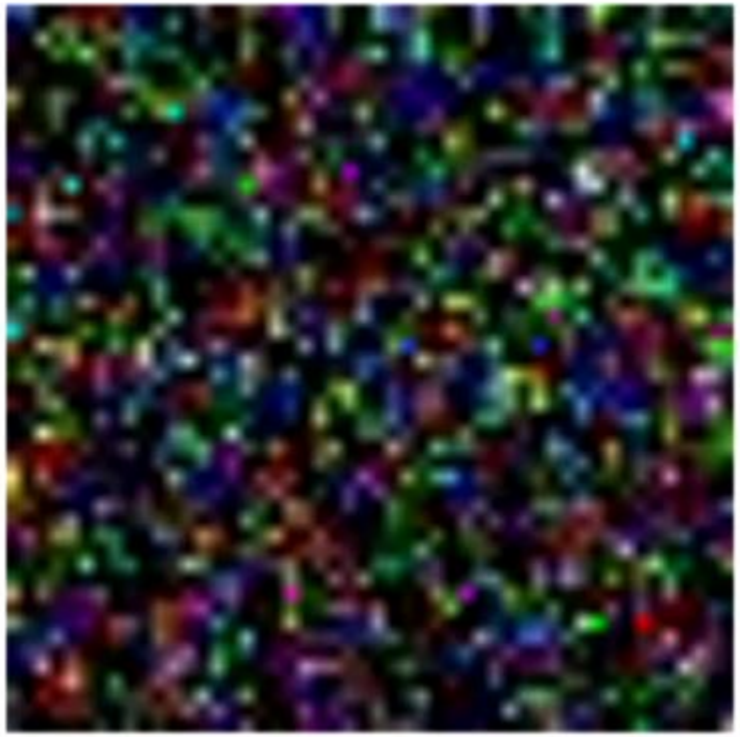
60% Accuracy (76% Clean)

10% Accuracy

Compress noice != white noise



Original

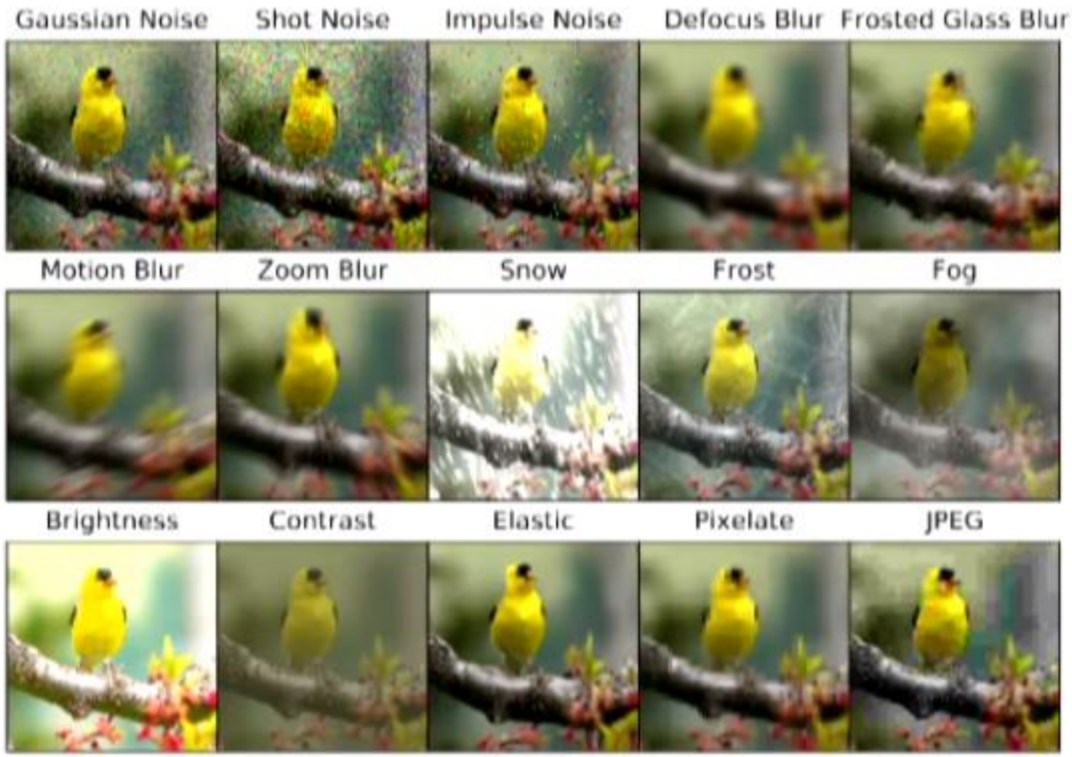


Compressed

Data augmentation can even hurt you

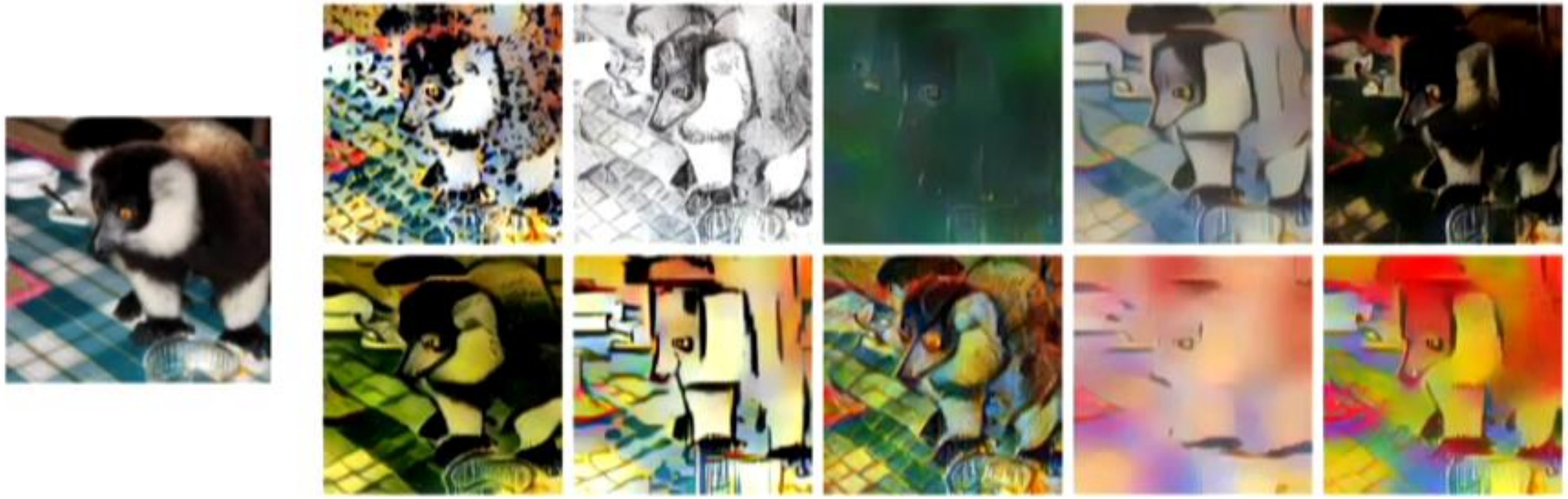
Corruption Robustness

- Goal: Measure and improve model robustness to distributional shift.
- Corruptions are not worst-case.
- Test examples are randomly sampled to best estimate probability of an error.



Progress is being made…

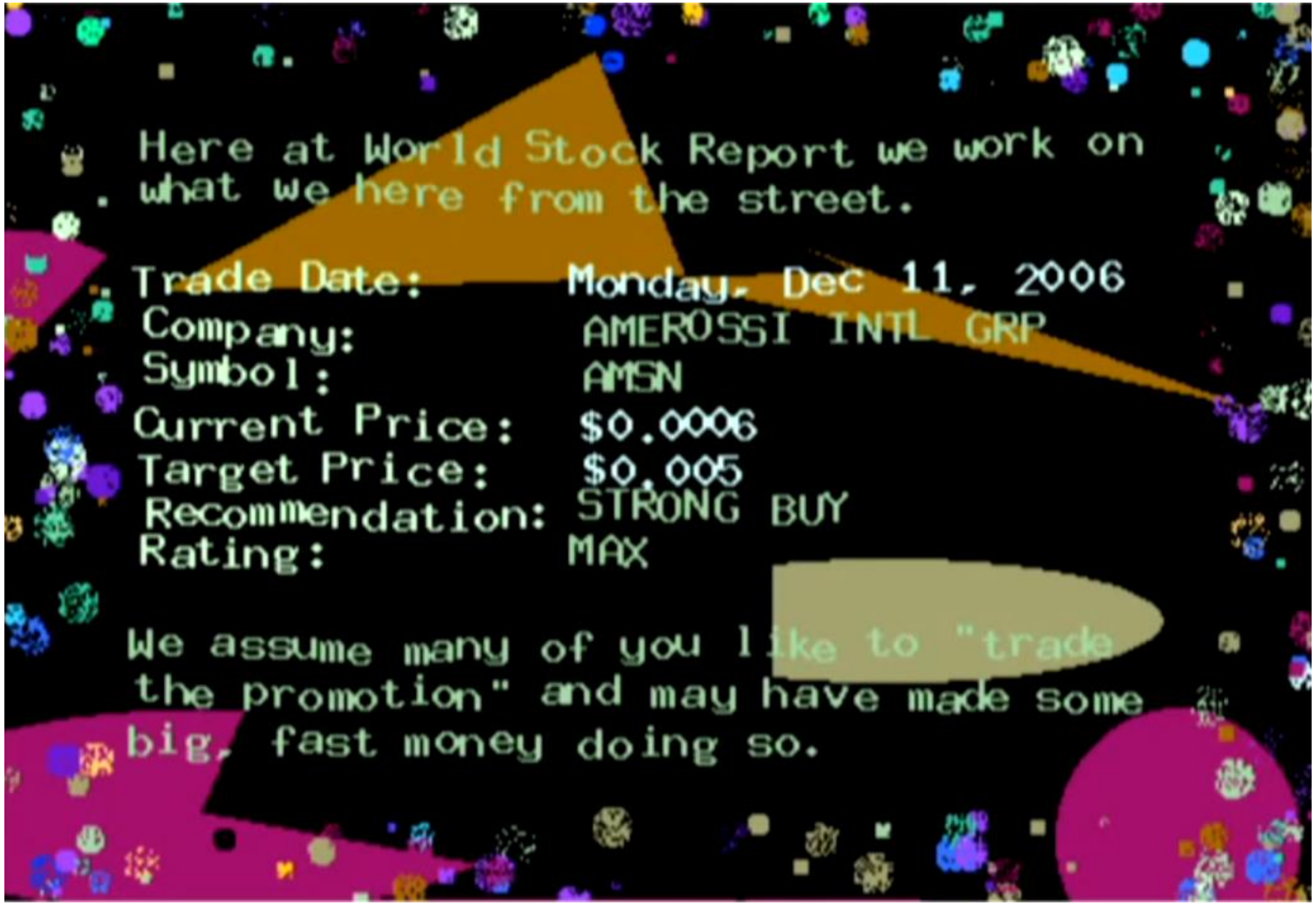
Training on randomized textures helps



Question

What is the difference between adversarial examples and test error?

Adversarial Examples - Security

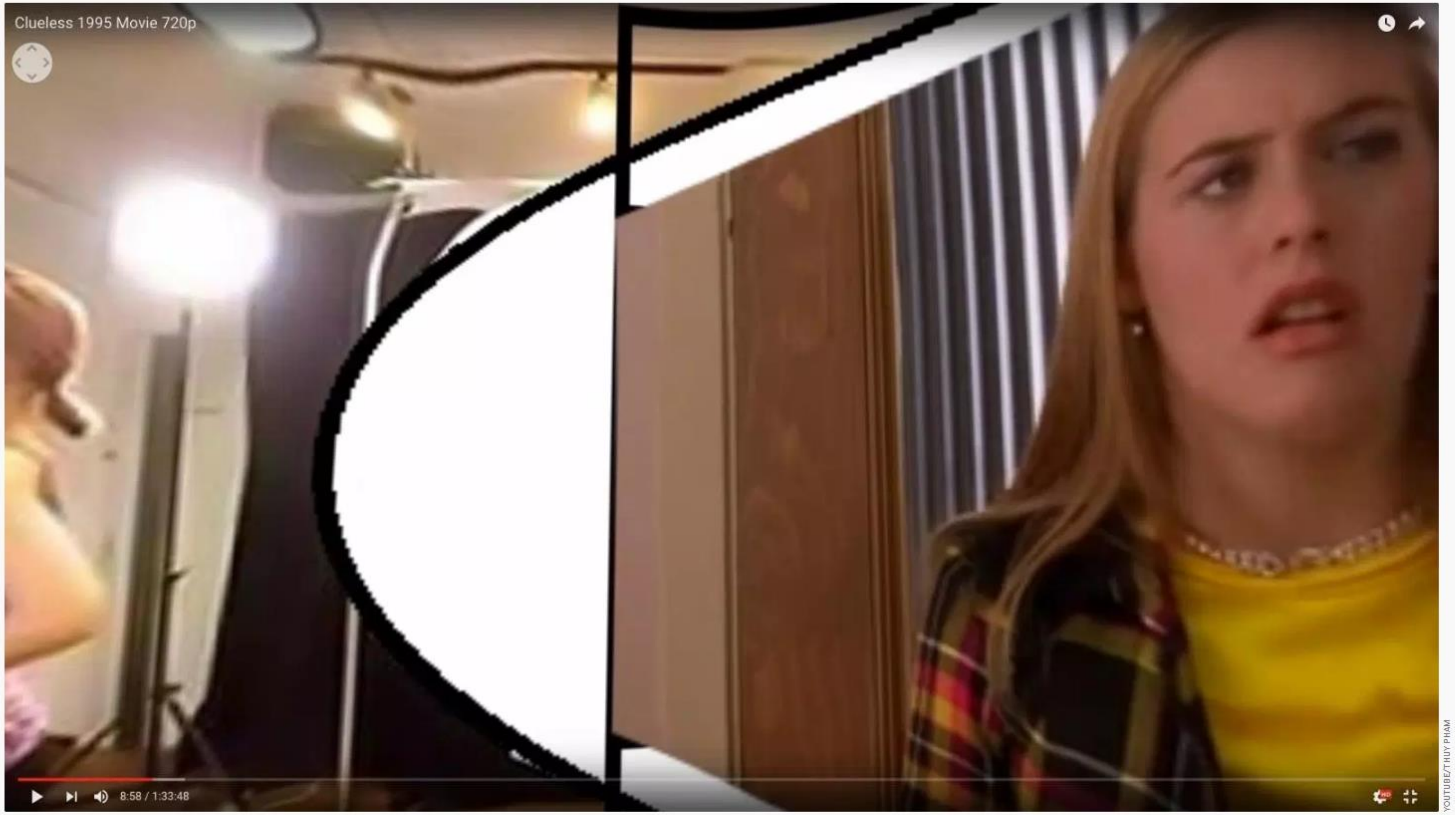


Here at World Stock Report we work on what we here from the street.

Trade Date: Monday, Dec 11, 2006
Company: AMEROSSI INTL GRP
Symbol: AMSN
Current Price: $0.0006
Target Price: $0.005
Recommendation: STRONG BUY
Rating: MAX

We assume many of you like to "trade the promotion" and may have made some big, fast money doing so.

Adversarial Examples - Security



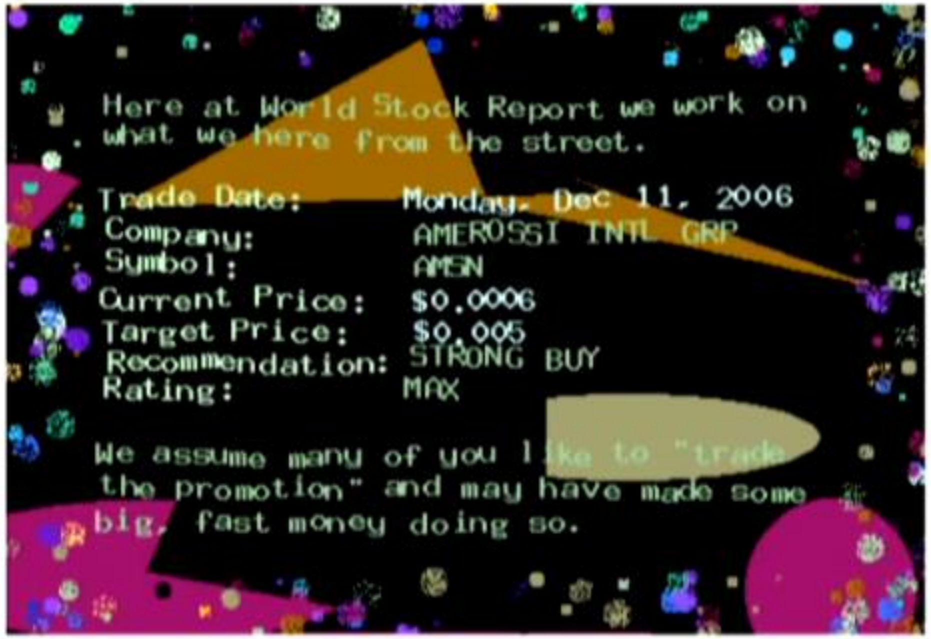
YOUTUBE/THUY PHAM

It's a movie, inside a movie.

https://qz.com/721615/smart-pirates-are-fooling-youtubes-copyright-bots-by-hiding-movies-in-360-degree-videos/

Questions for Design a Secure ML System

- How do adversaries typically break systems?
- How would you measure test error?
- Are you secure if test error > 0?
- How do we deal with out-of-distribution generalization?

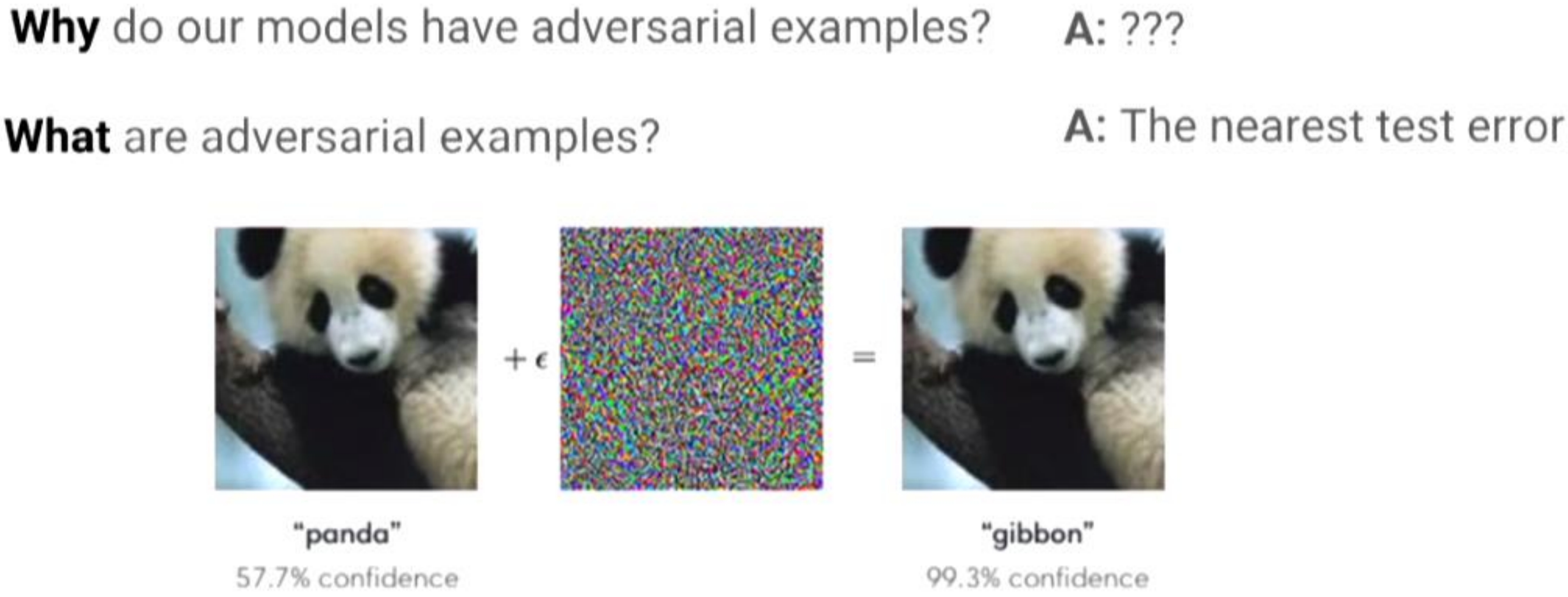


Here at World Stock Report we work on what we here from the street.

Trade Date: Monday, Dec 11, 2006
Company: AMEROSSI INTL GRP
Symbol: AMSN
Current Price: $0.0006
Target Price: $0.005
Recommendation: STRONG BUY
Rating: MAX

We assume many of you like to "trade the promotion" and may have made some big, fast money doing so.

Adversarial Examples – ML Phenomenon



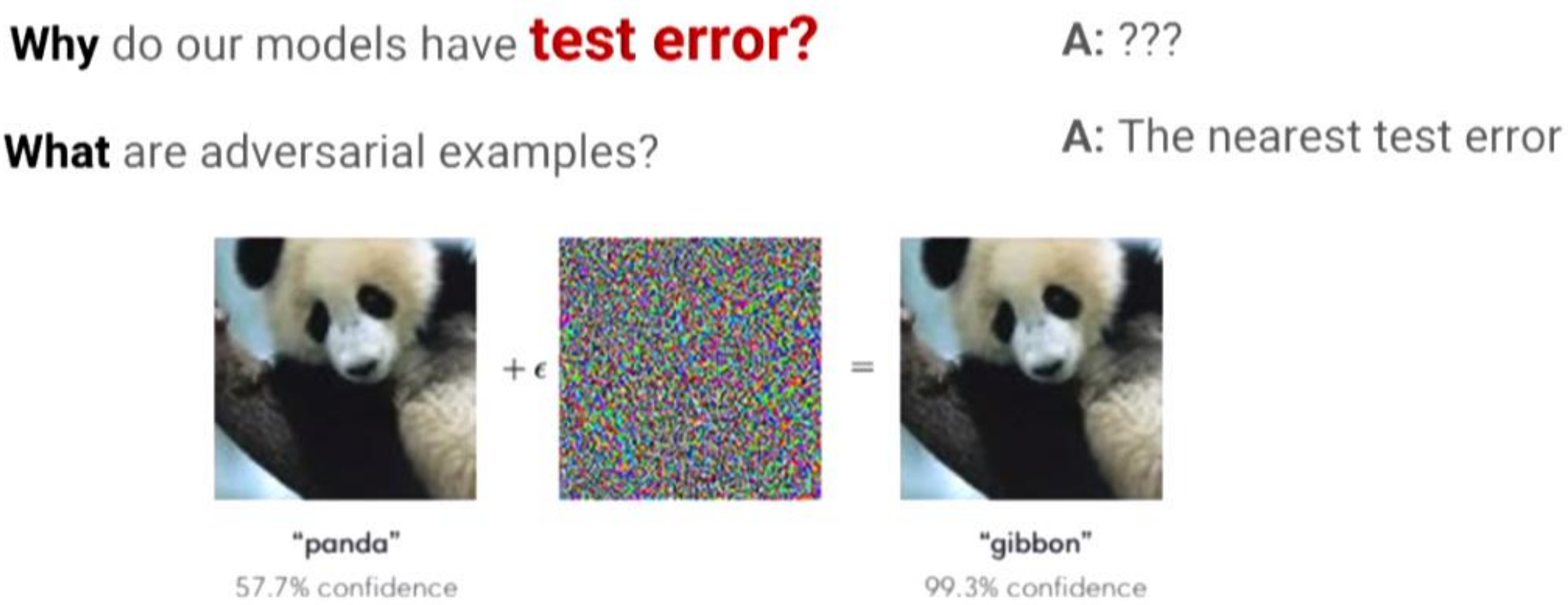
Why do our models have adversarial examples? **A:** ???

What are adversarial examples? **A:** The nearest test error

“panda”
57.7% confidence

“gibbon”
99.3% confidence

Adversarial Examples – ML Phenomenon



Why do our models have **test error?**

A: ???

What are adversarial examples?

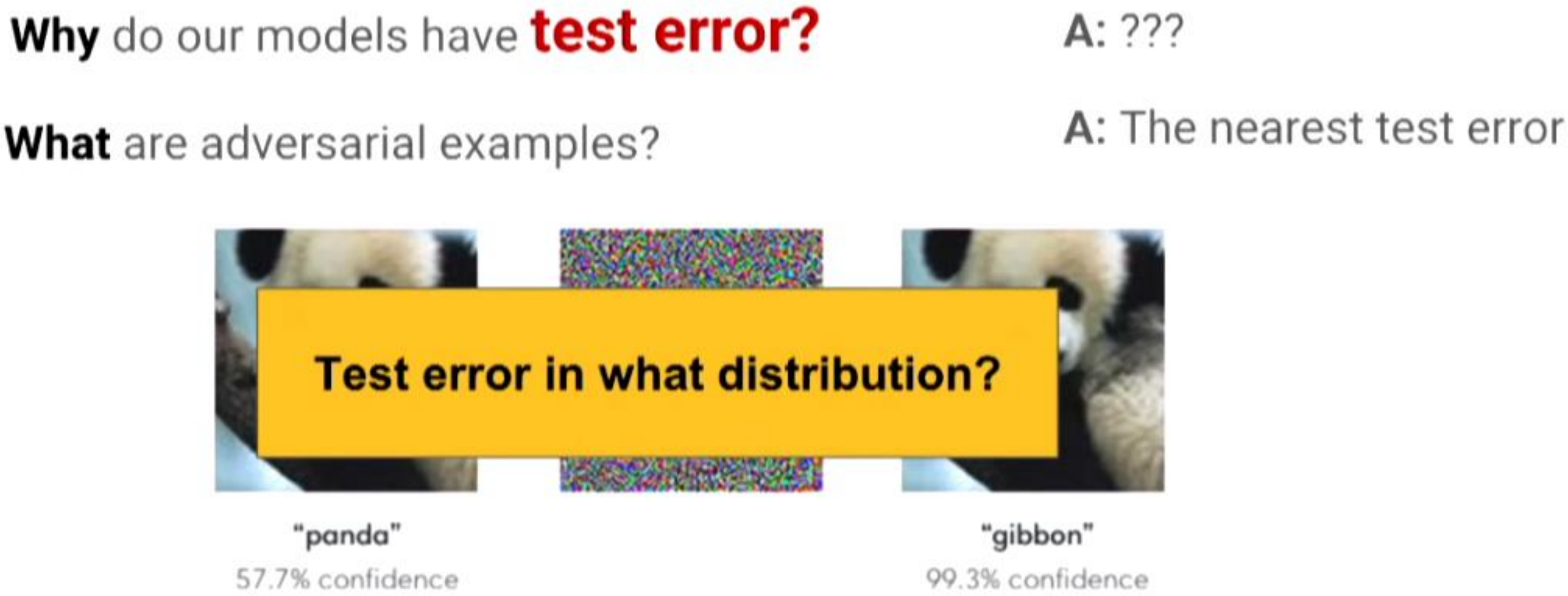
A: The nearest test error

$+\epsilon$ =

"panda"
57.7% confidence

"gibbon"
99.3% confidence

Adversarial Examples – ML Phenomenon



Why do our models have **test error?**

A: ???

What are adversarial examples?

A: The nearest test error

Test error in what distribution?

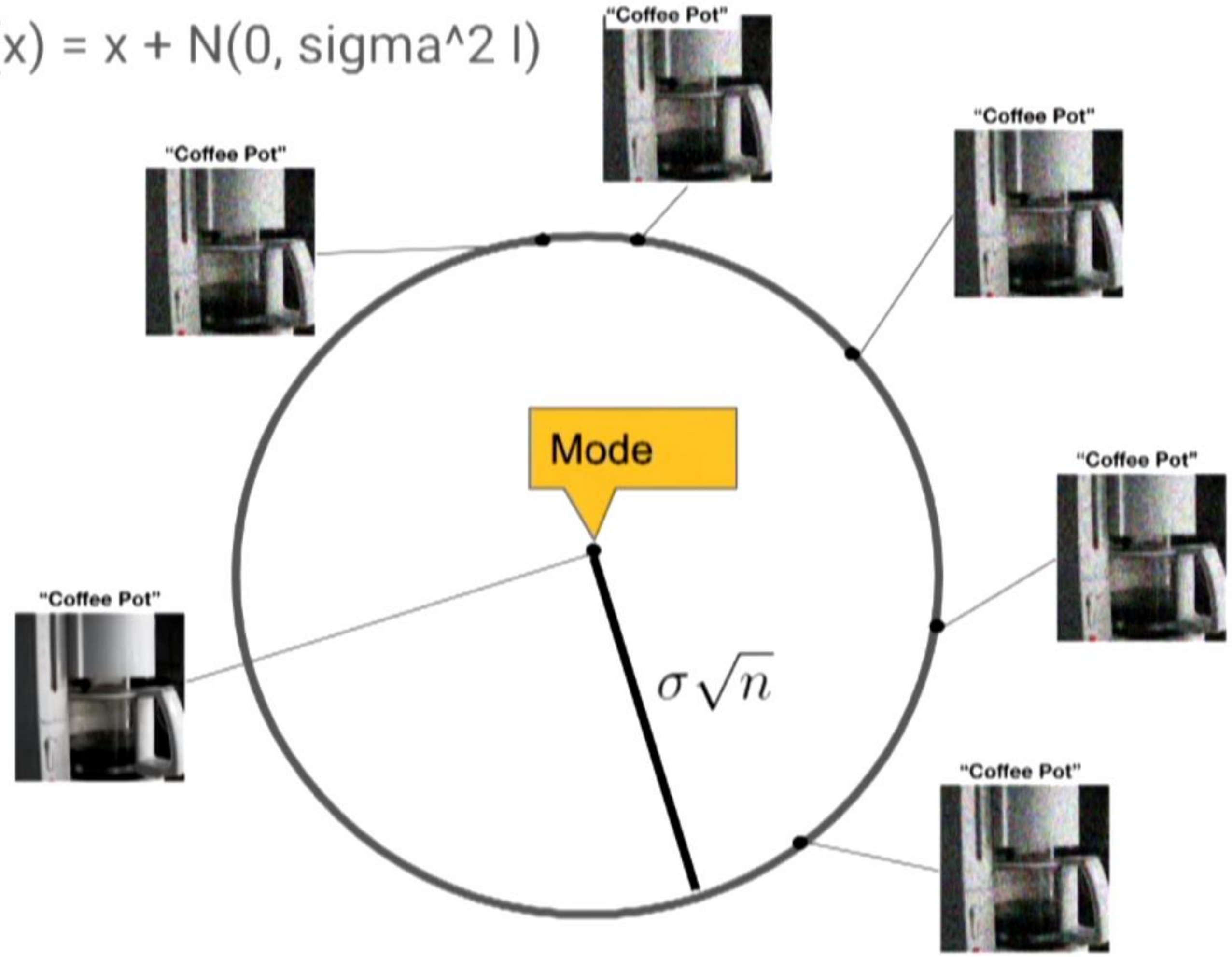
"panda"
57.7% confidence

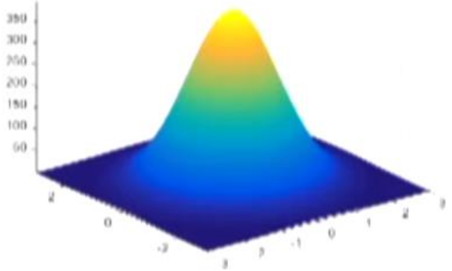
"gibbon"
99.3% confidence

- In high dimensions, what does .1% test error look like?
- How close should the nearest test error be? (Assuming we sample infinitely)

High Dimensional Gaussians

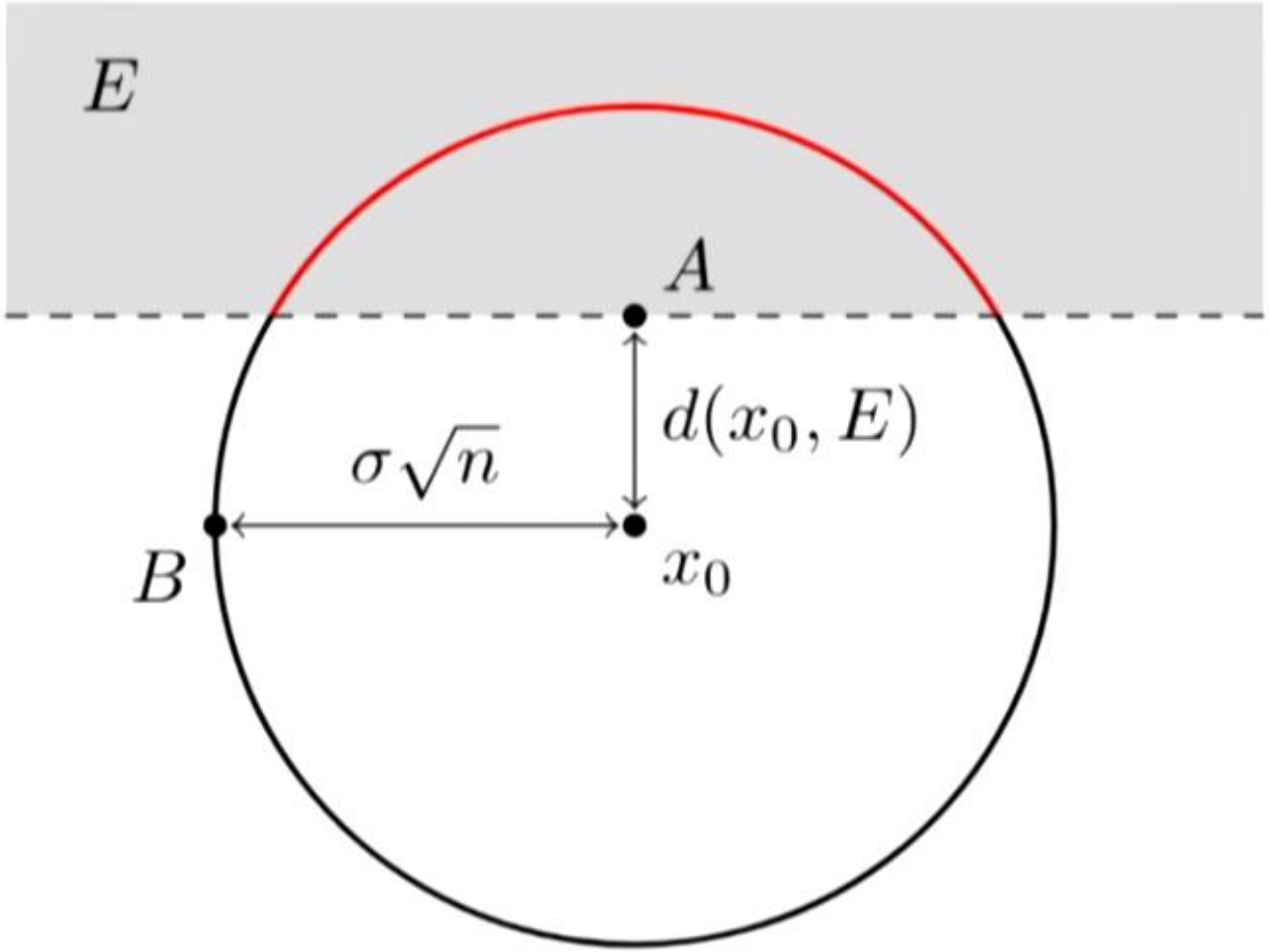
q(x) = x + N(0, sigma^2 I)





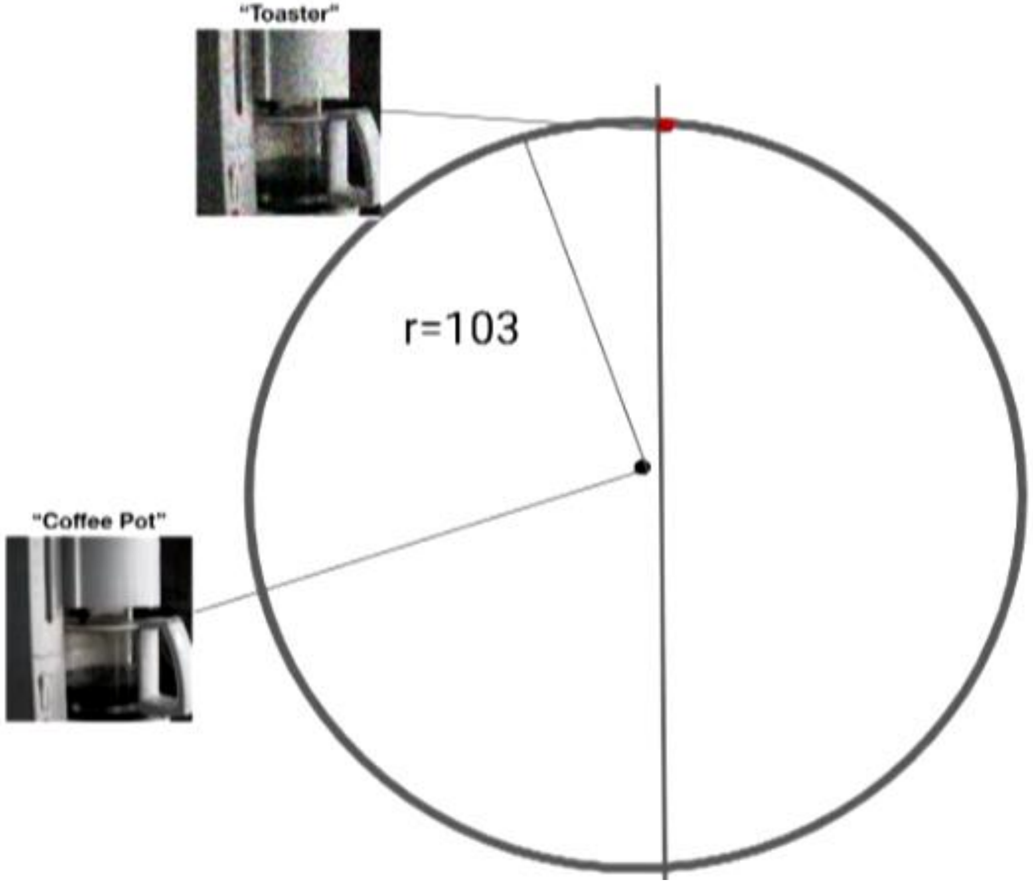
- sigma=.2
- n=299*299*3
- 270,000 dimensional sphere
- radius ~ 103

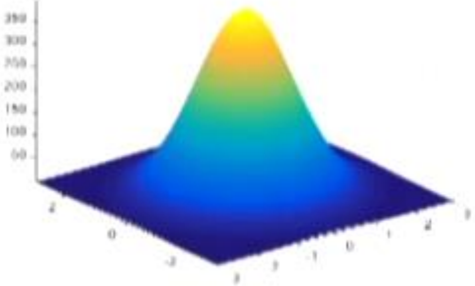
Adversarial Examples – ML Phenomenon



Linear Models

q(x) = x + N(0, sigma^2 I)



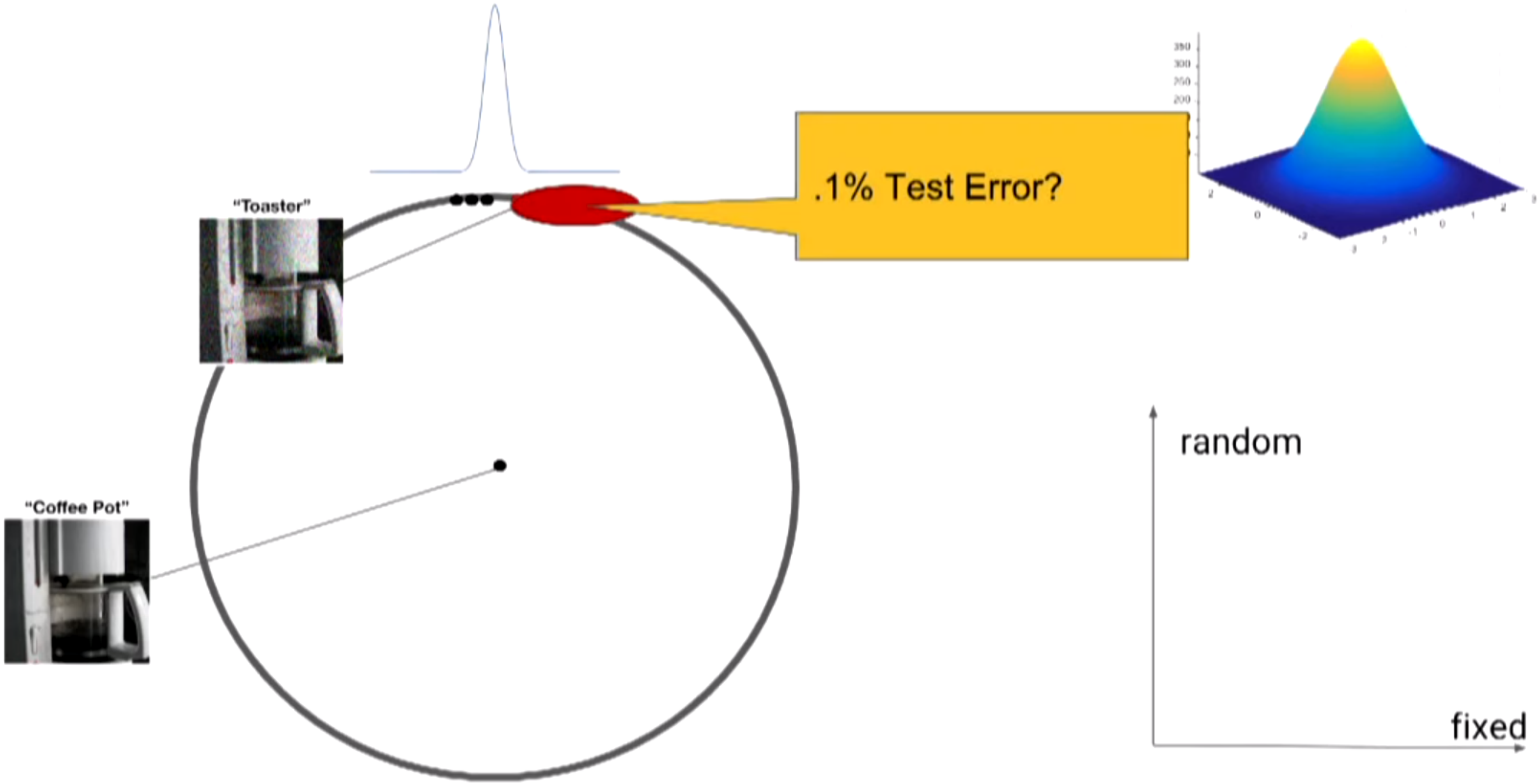


Theorem: A linear model with error rate mu in distribution q, has its nearest error at distance $\sigma\Phi^{-1}(\mu) = O(\sigma)$

- sigma=.2
- .1% error -> d = .62
- 10^-9 error -> d = 1.2

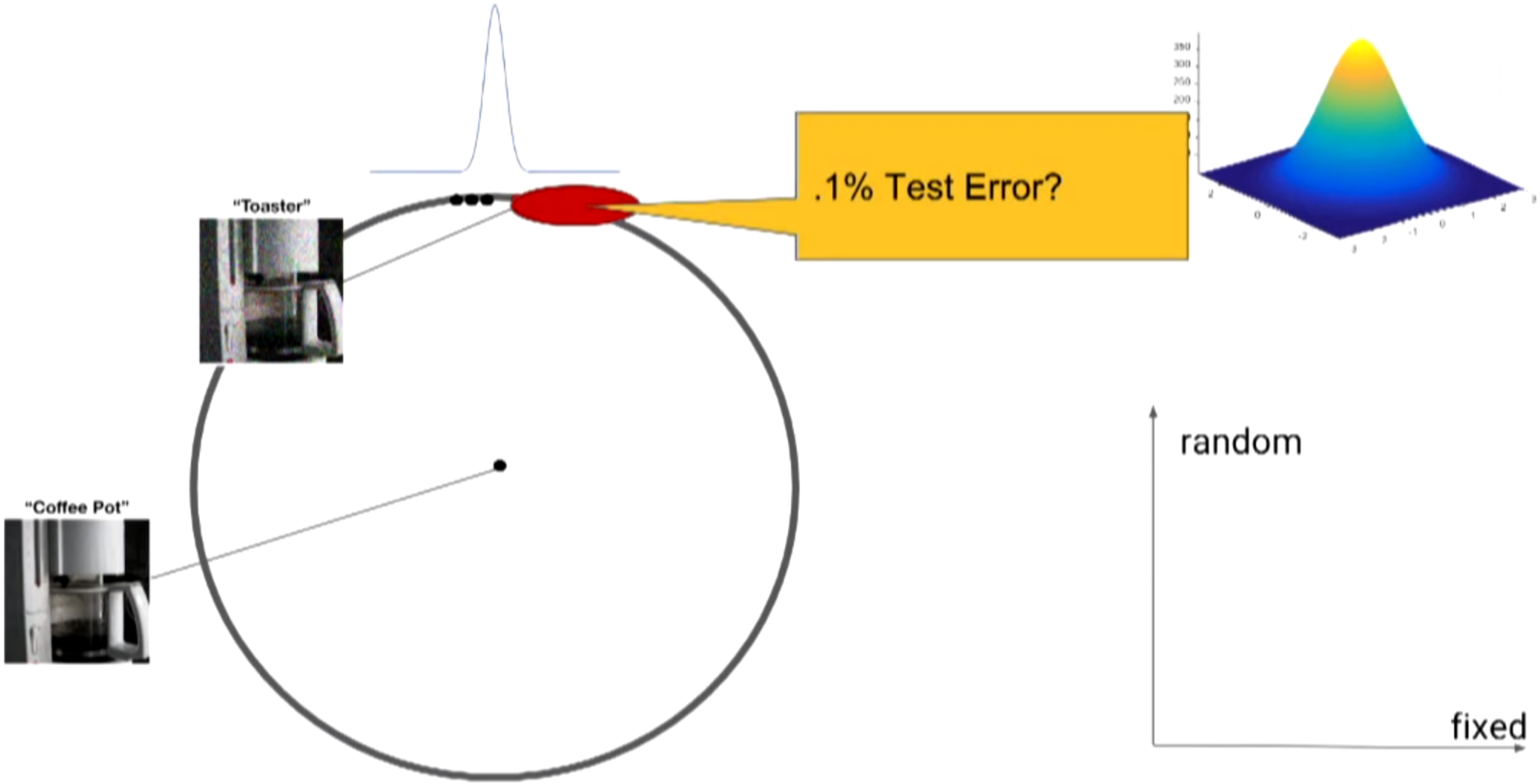
https://arxiv.org/pdf/1608.08967.pdf

Where is .1% Test Error?



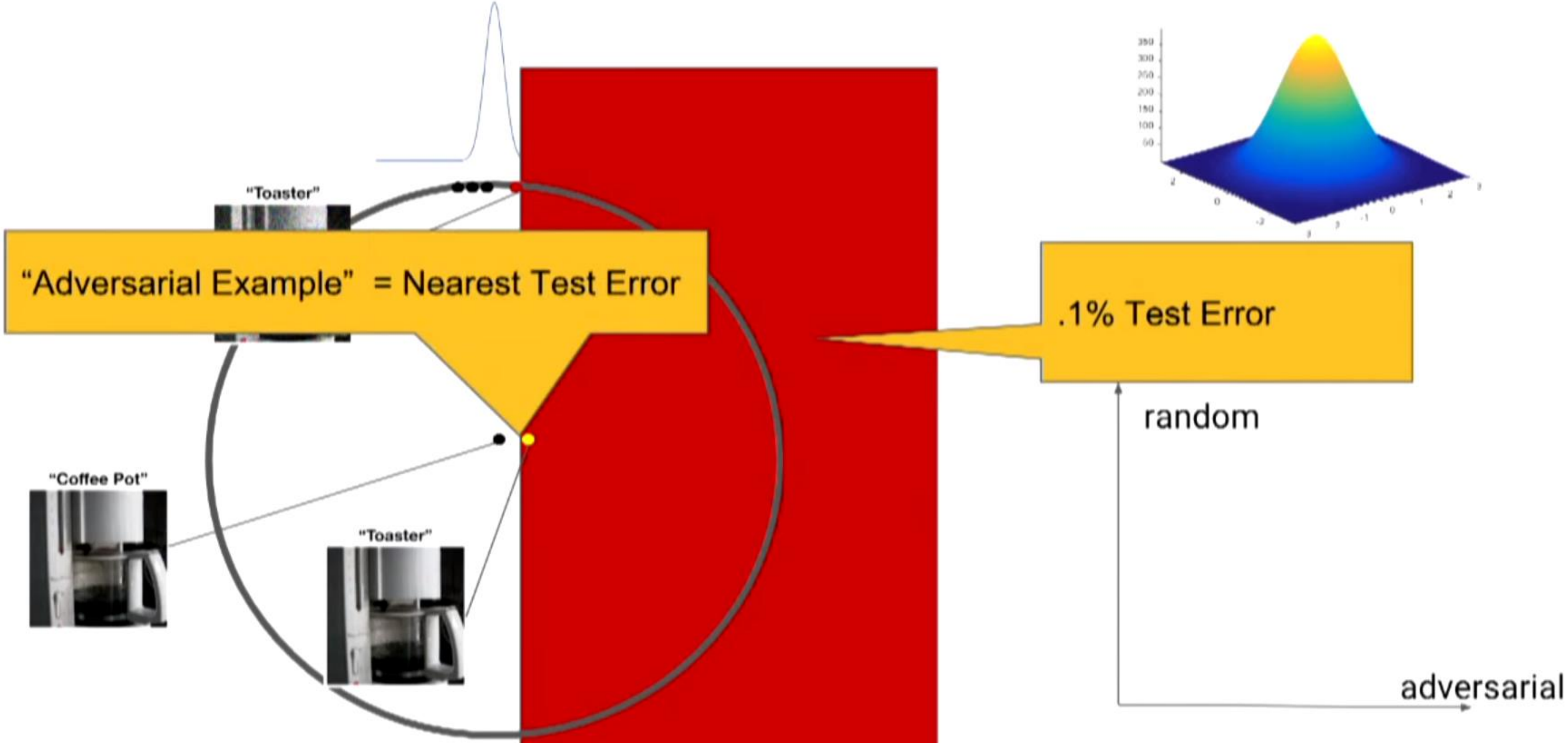
.1% Test Error?

Where is .1% Test Error?



.1% Test Error?

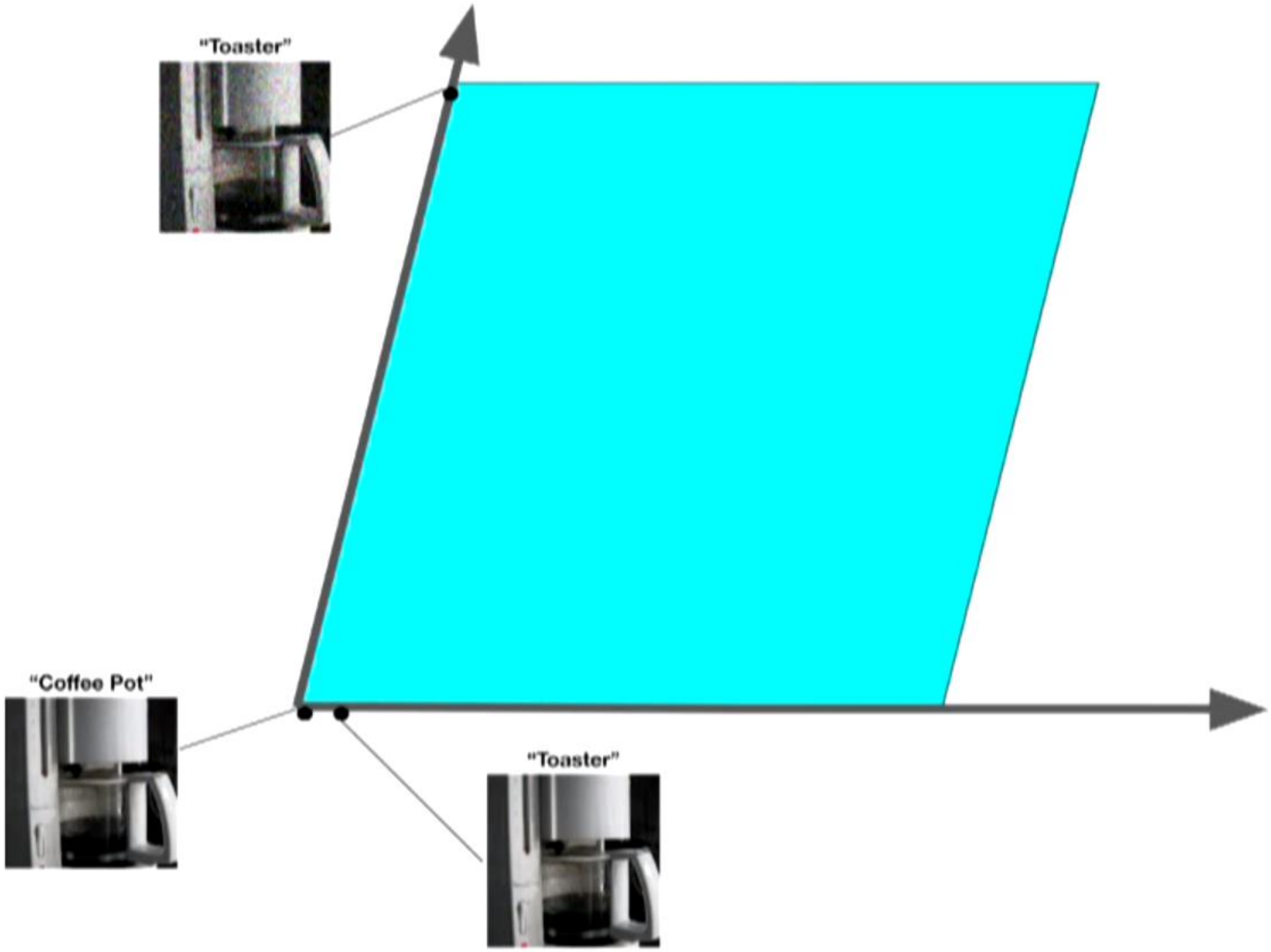
Where is .1% Test Error?



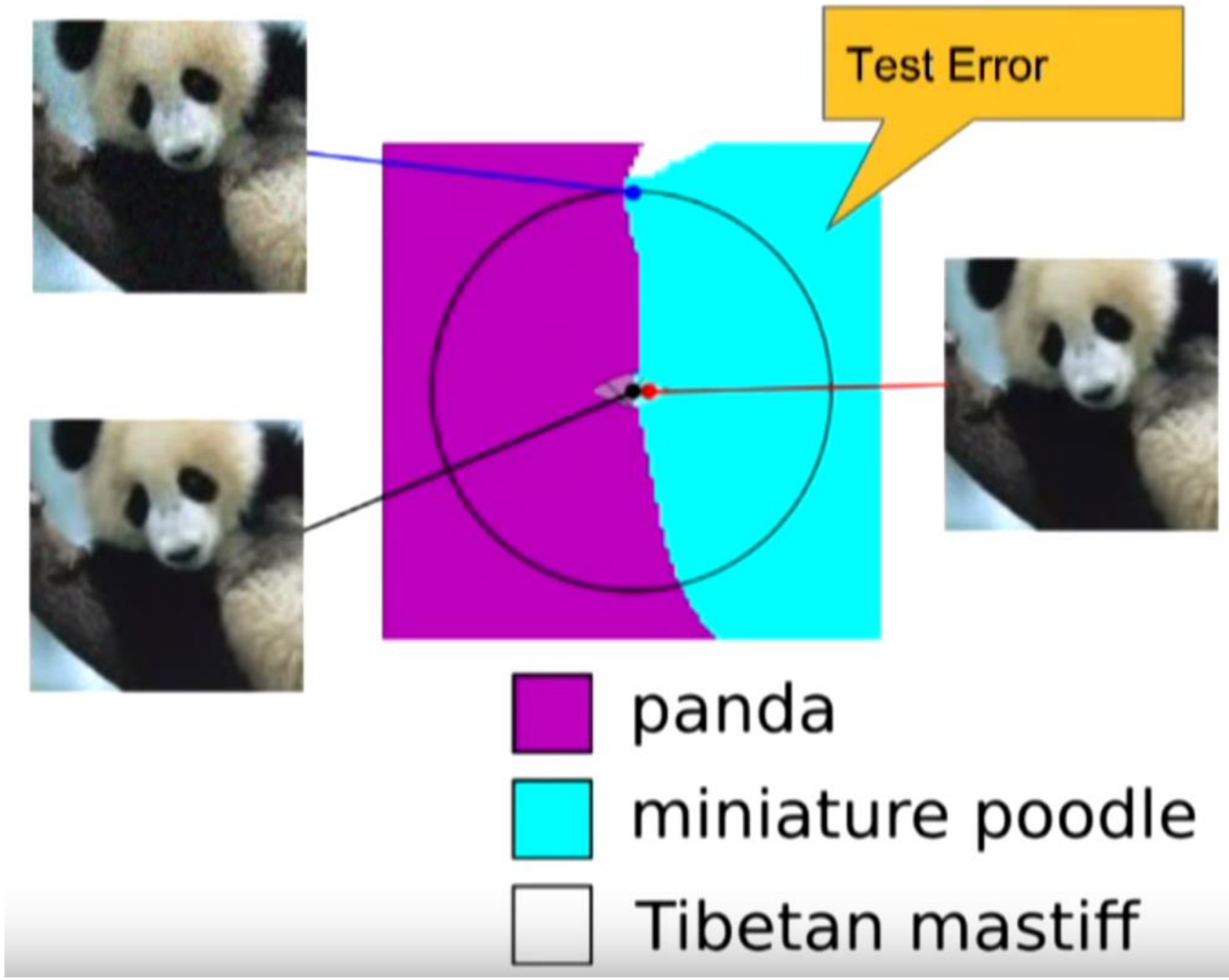
"Adversarial Example" = Nearest Test Error

.1% Test Error

Church Window Plot

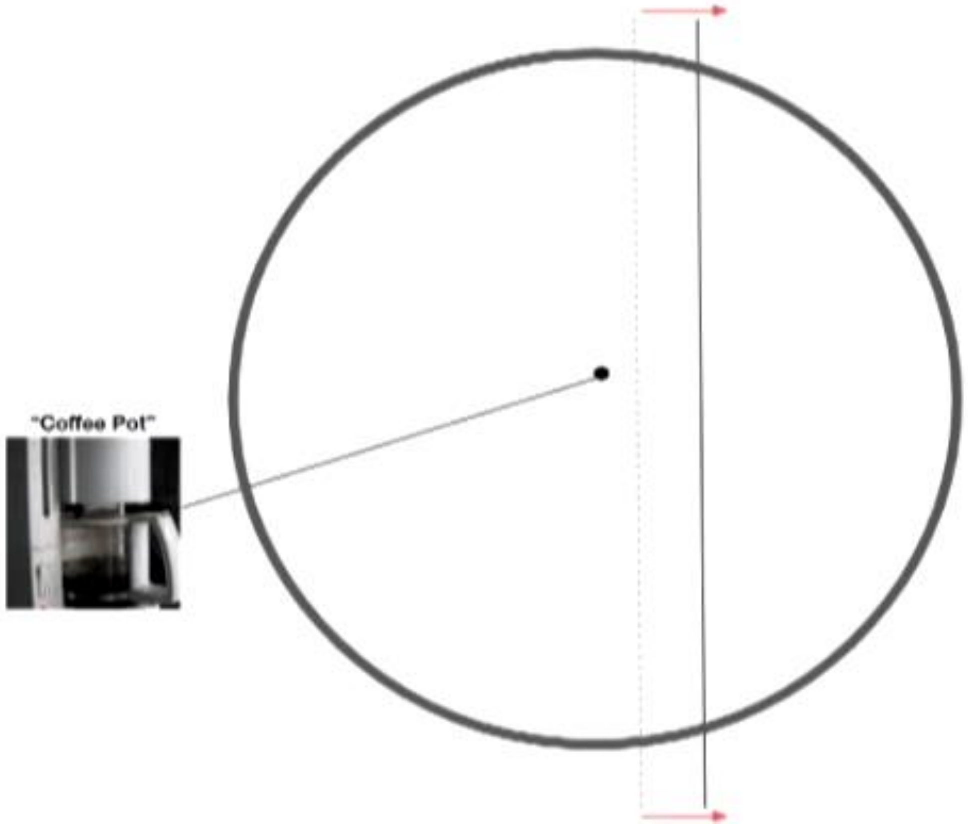


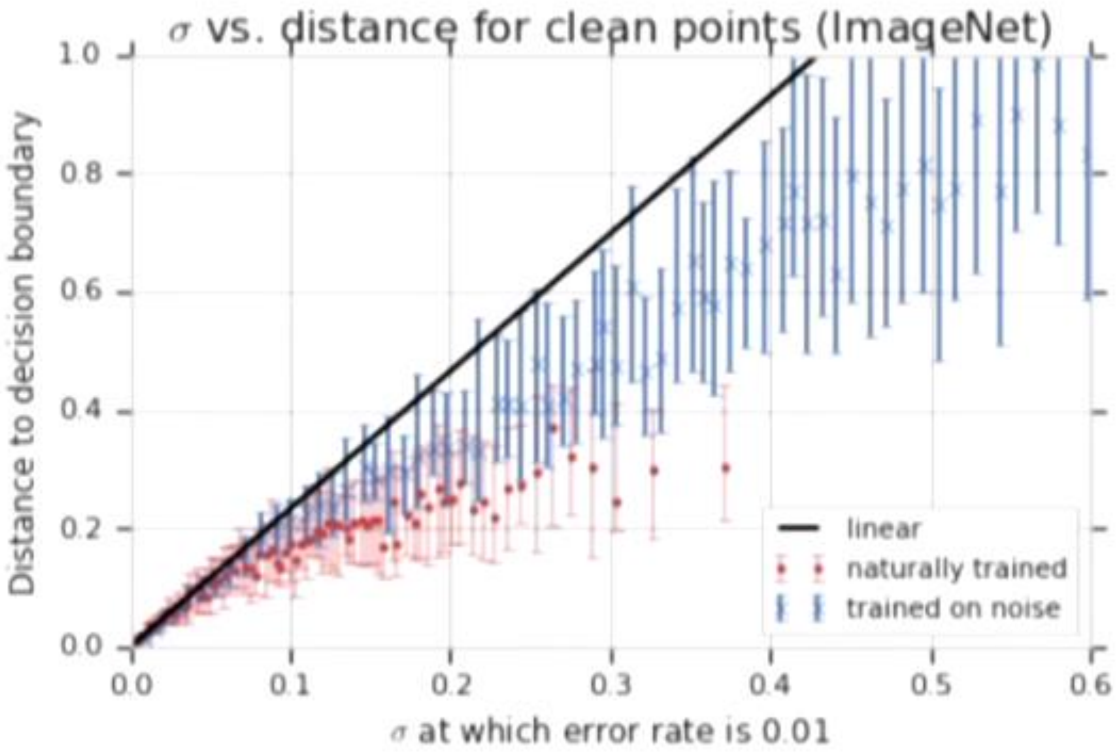
InceptionV3



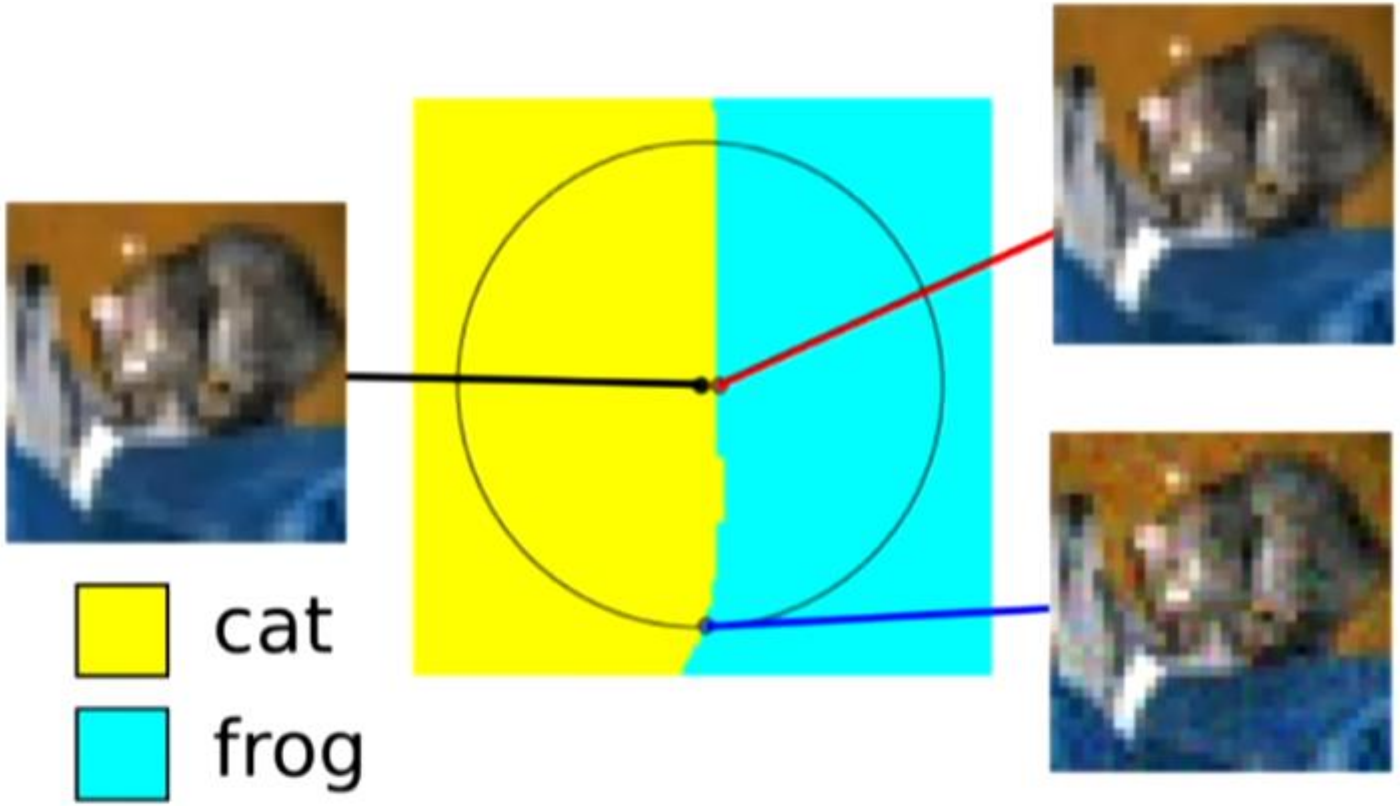
- sigma=.08
- R=31
- error rate .1%
- d = .189
- predicted d = .246

InceptionV3



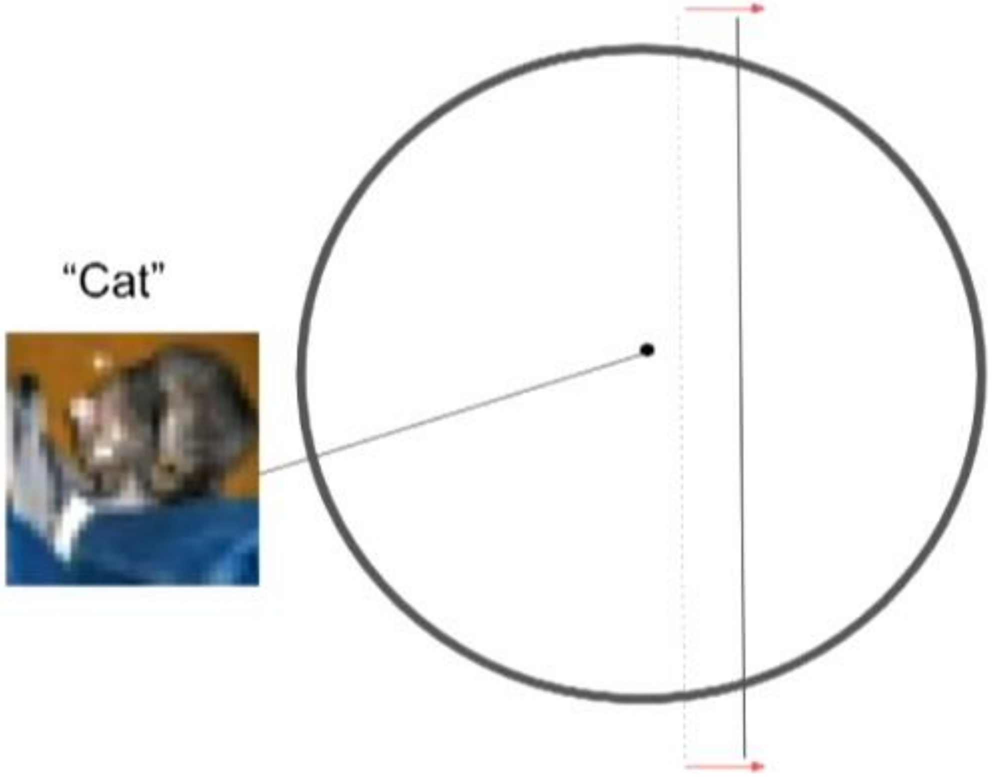


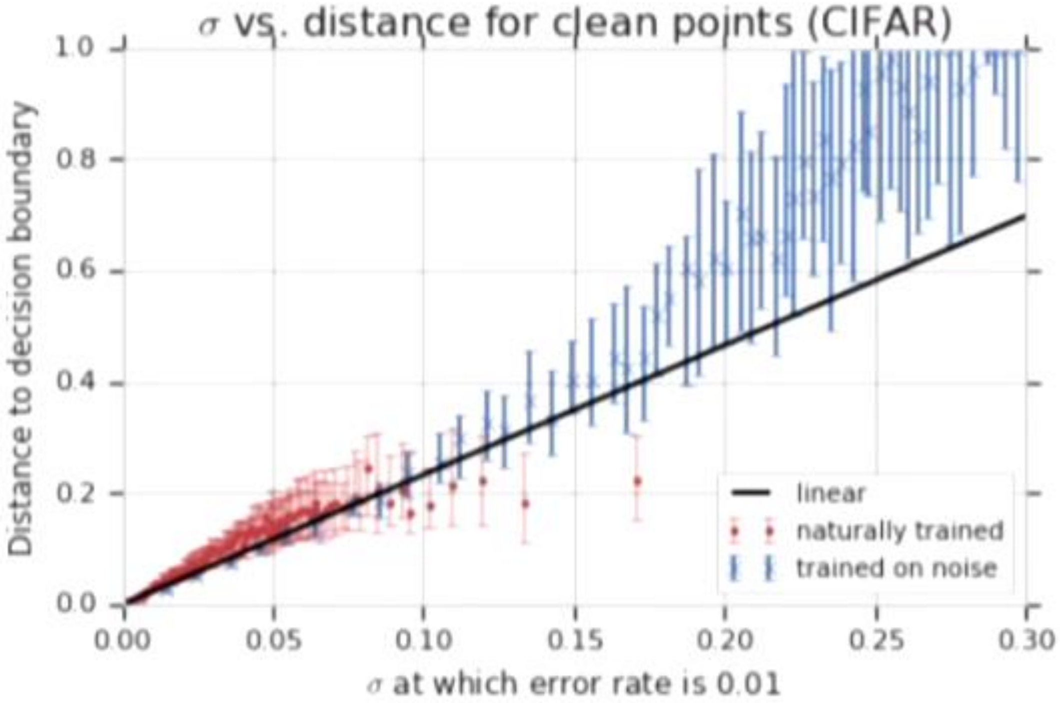
Resnet-50



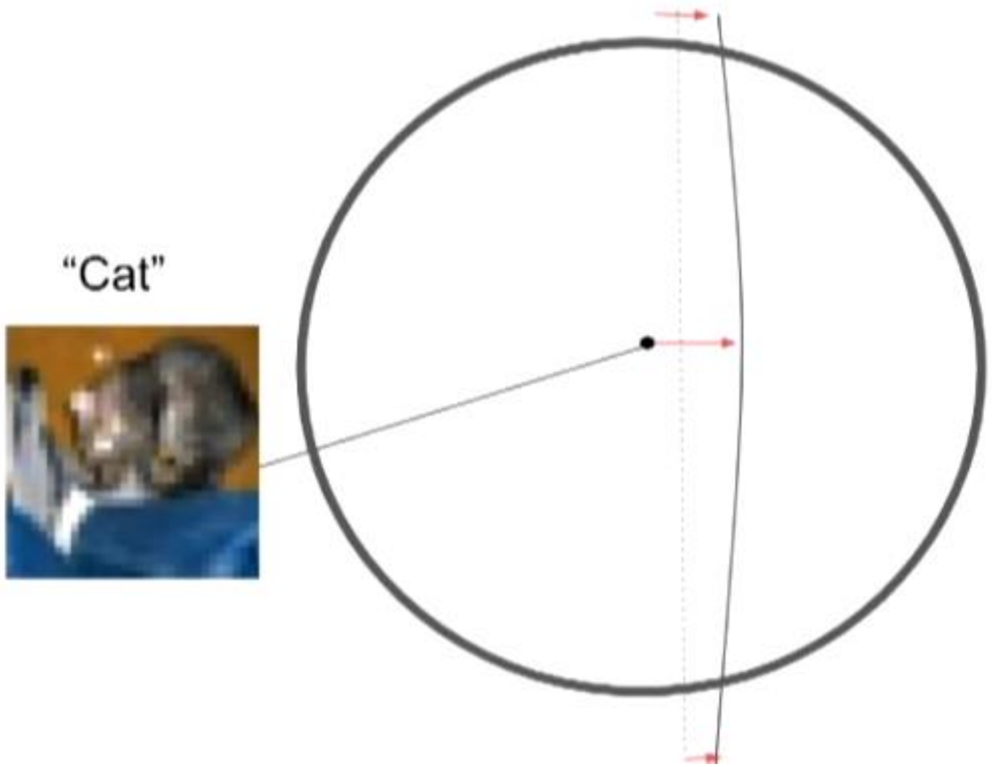
sigma=.04 (R=2.2)
error rate .2%
d = .16
predicted d = .08

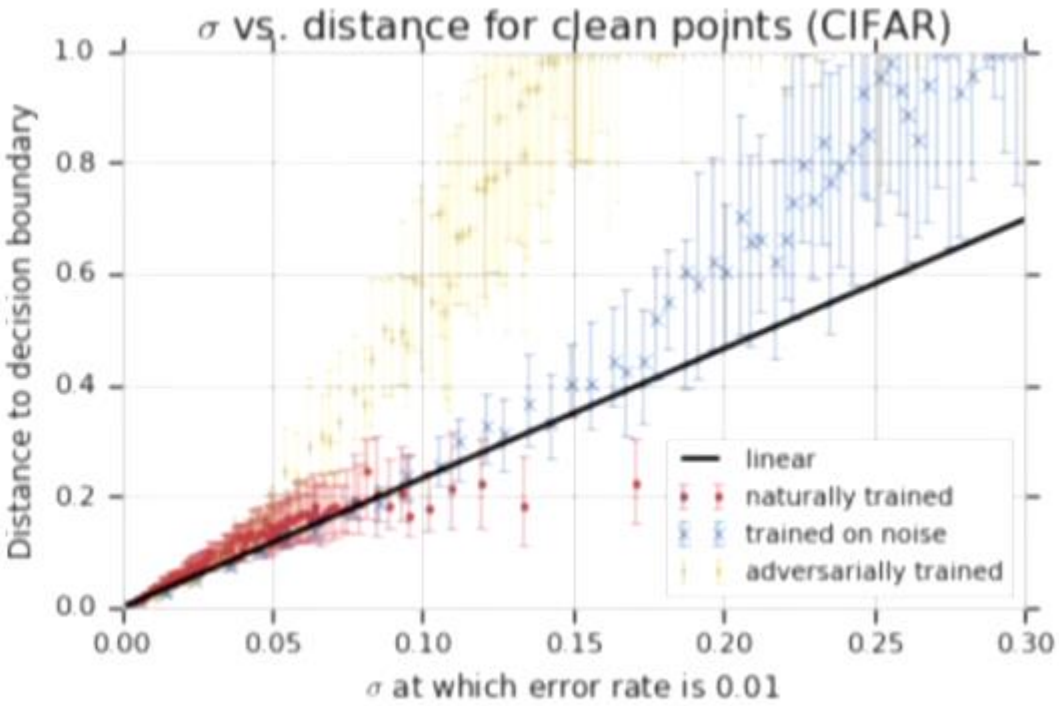
Resnet-50


"Cat"


σ vs. distance for clean points (CIFAR)
Distance to decision boundary
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
σ at which error rate is 0.01
linear
naturally trained
trained on noise

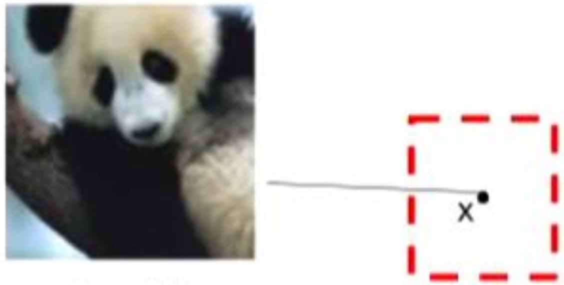
Resnet-50


"Cat"


σ vs. distance for clean points (CIFAR)
Distance to decision boundary
σ at which error rate is 0.01
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
linear
naturally trained
trained on noise
adversarially trained

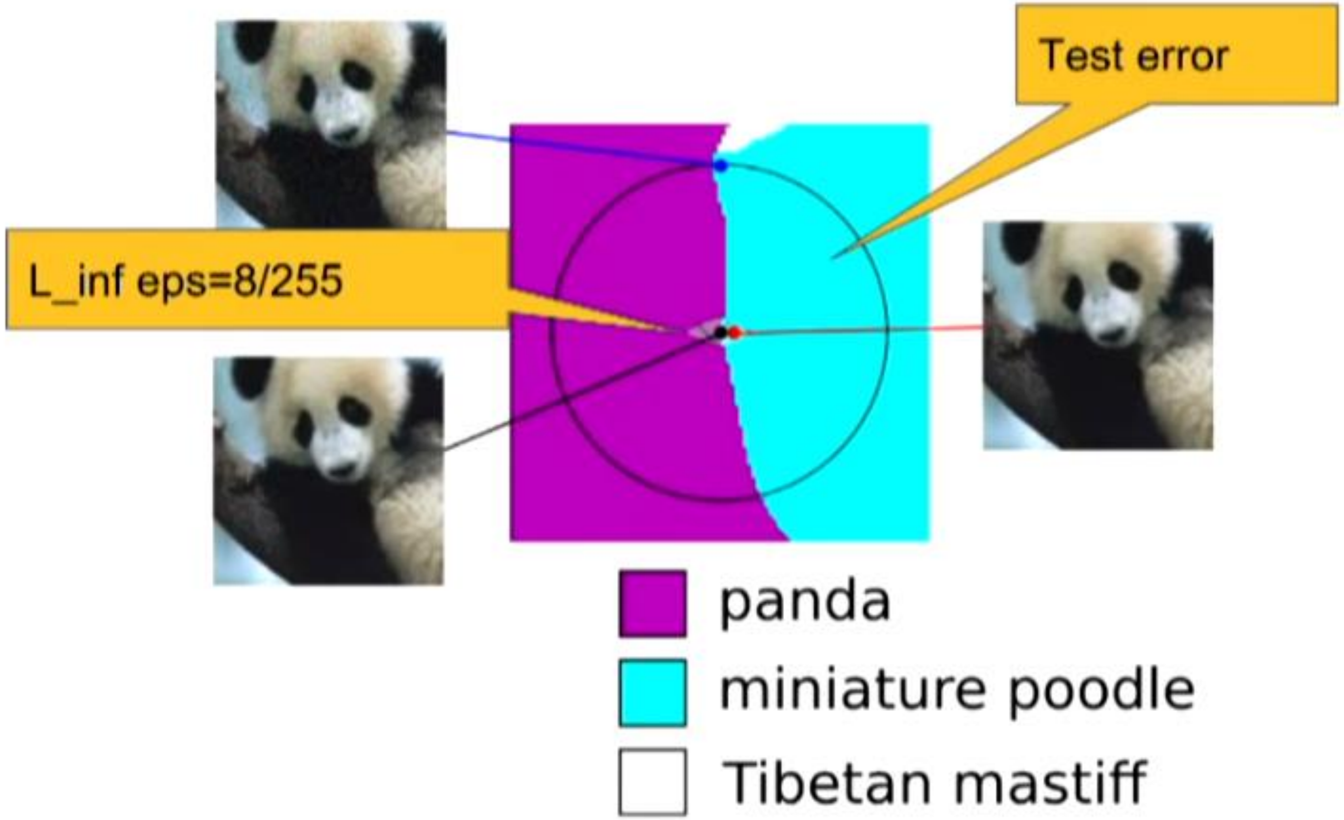
Adversarial Defenses – Why?

Why are we trying to "defend" against the nearest test error?

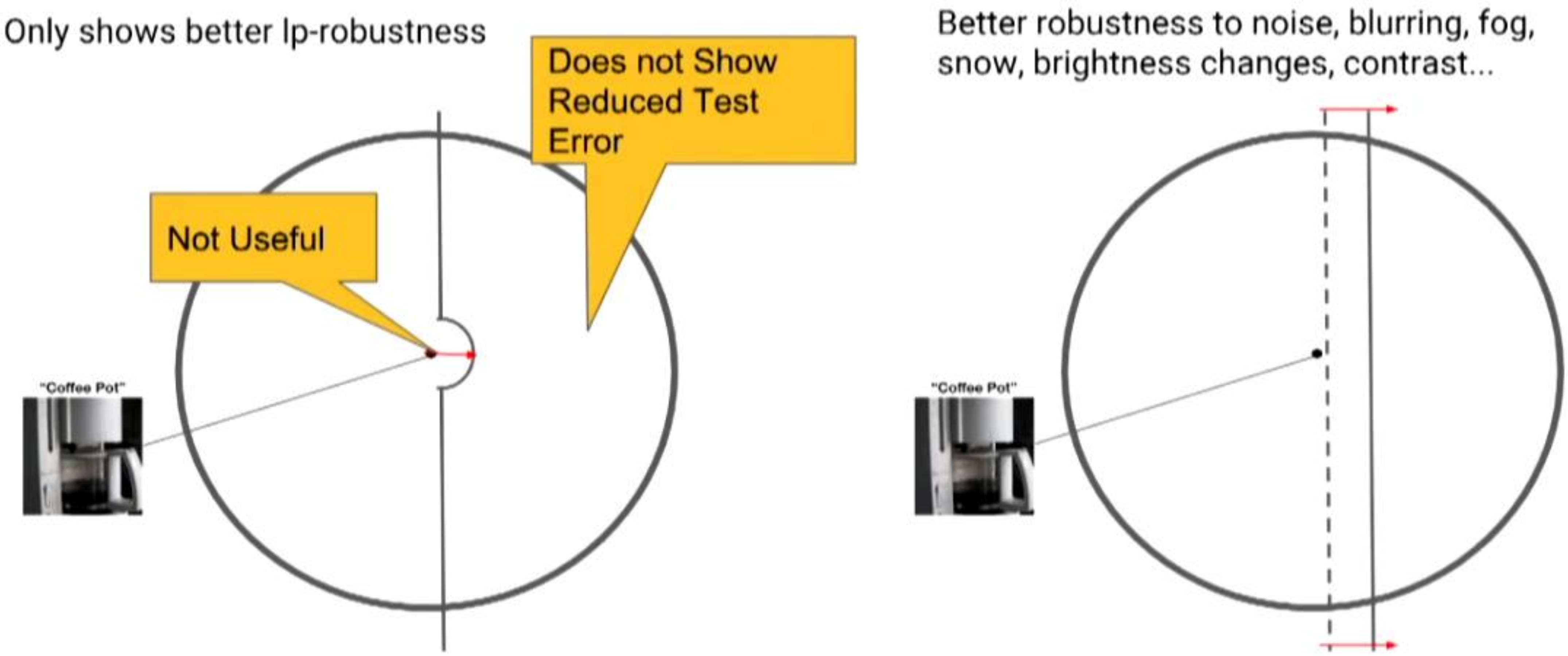


"panda"
57.7% confidence

$$x_{adv} = \max_{x':||x-x'||_\infty<\epsilon} L(\theta, x', \hat{y})$$



Adversarial Defenses – Why?



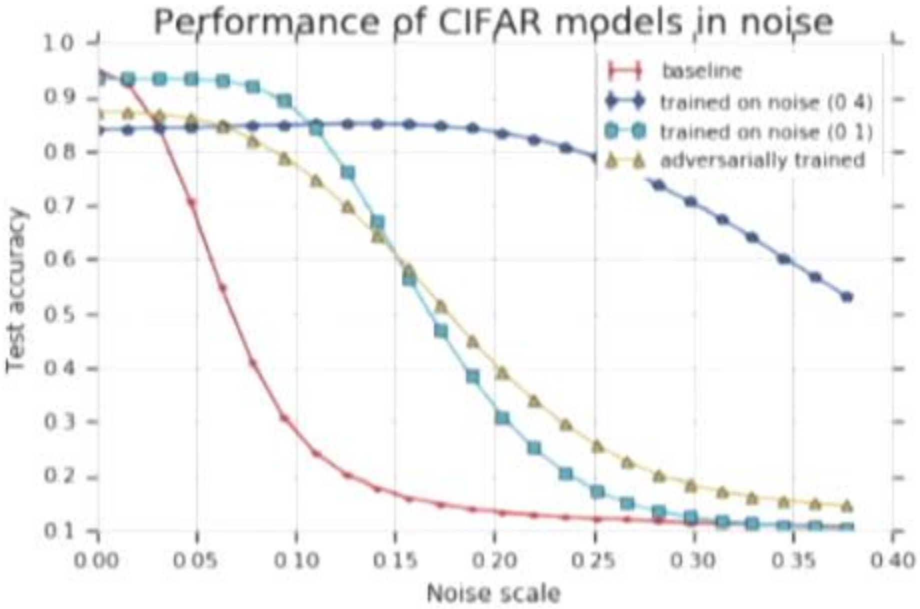
Only shows better lp-robustness

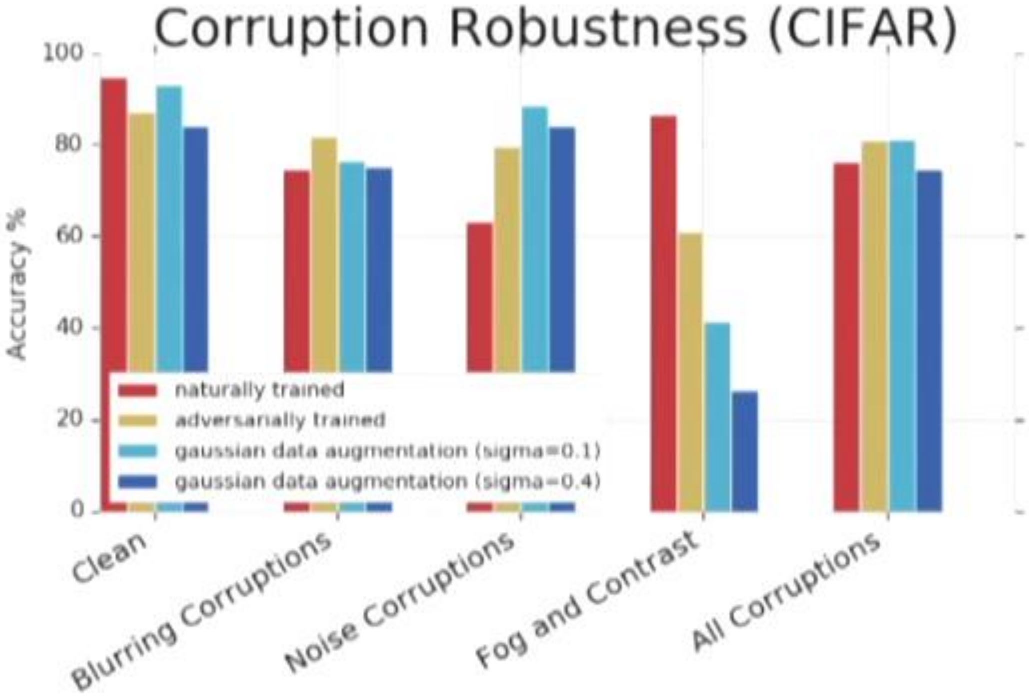
Does not Show Reduced Test Error

Not Useful

Better robustness to noise, blurring, fog, snow, brightness changes, contrast...

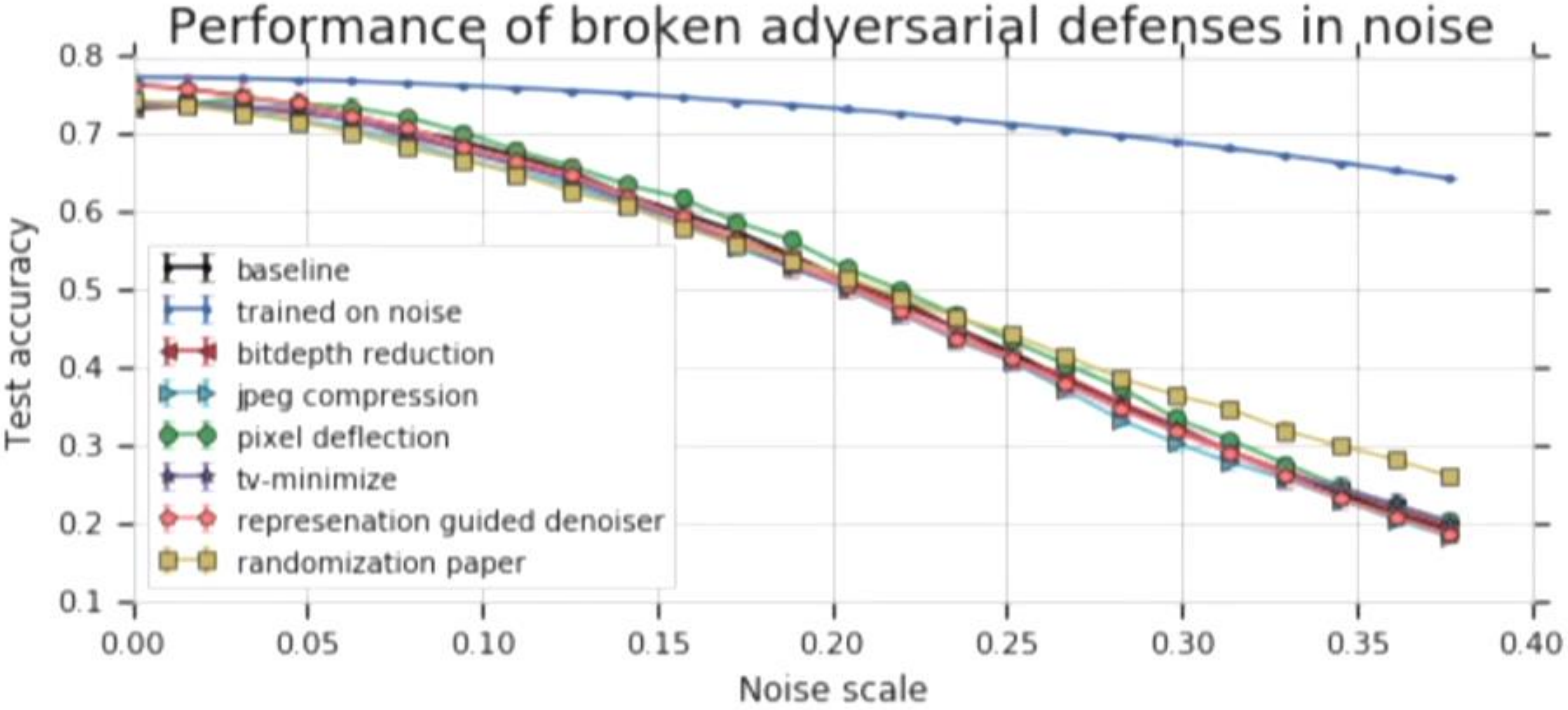
Successful Defenses Improve Robustness





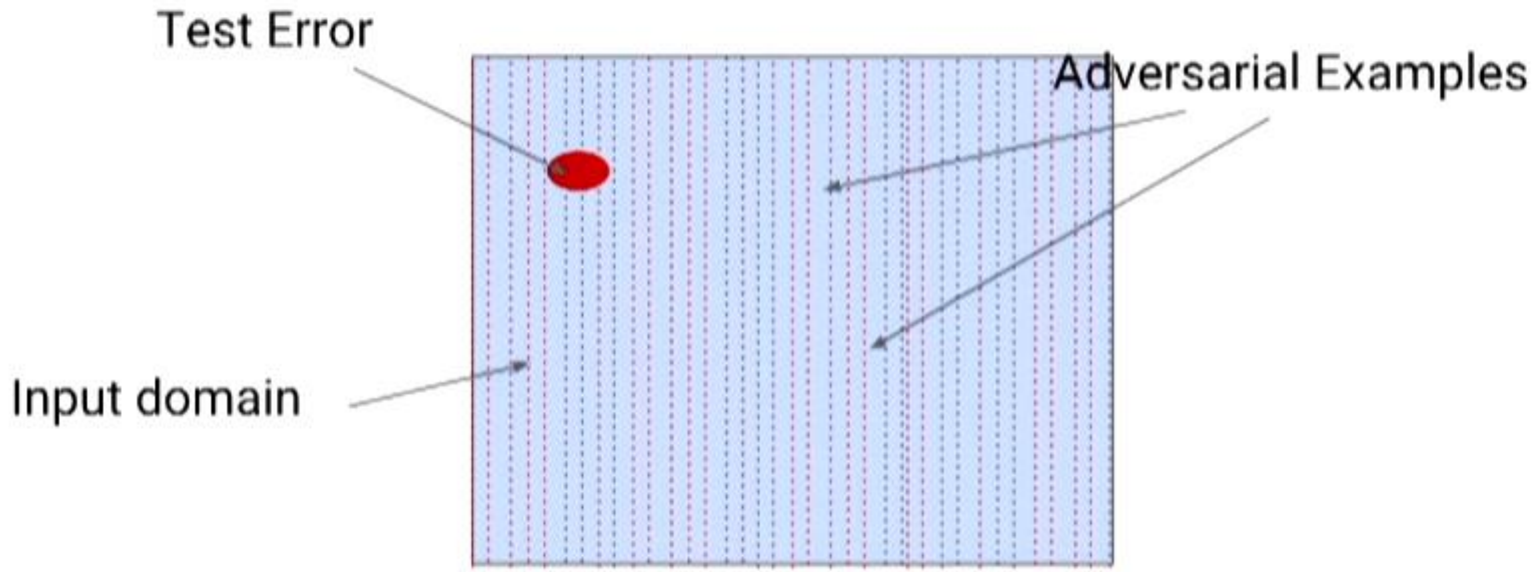
https://arxiv.org/pdf/1706.06083.pdf

Failed Defenses Don’t Improve Robustness



The Wrong Mental Model

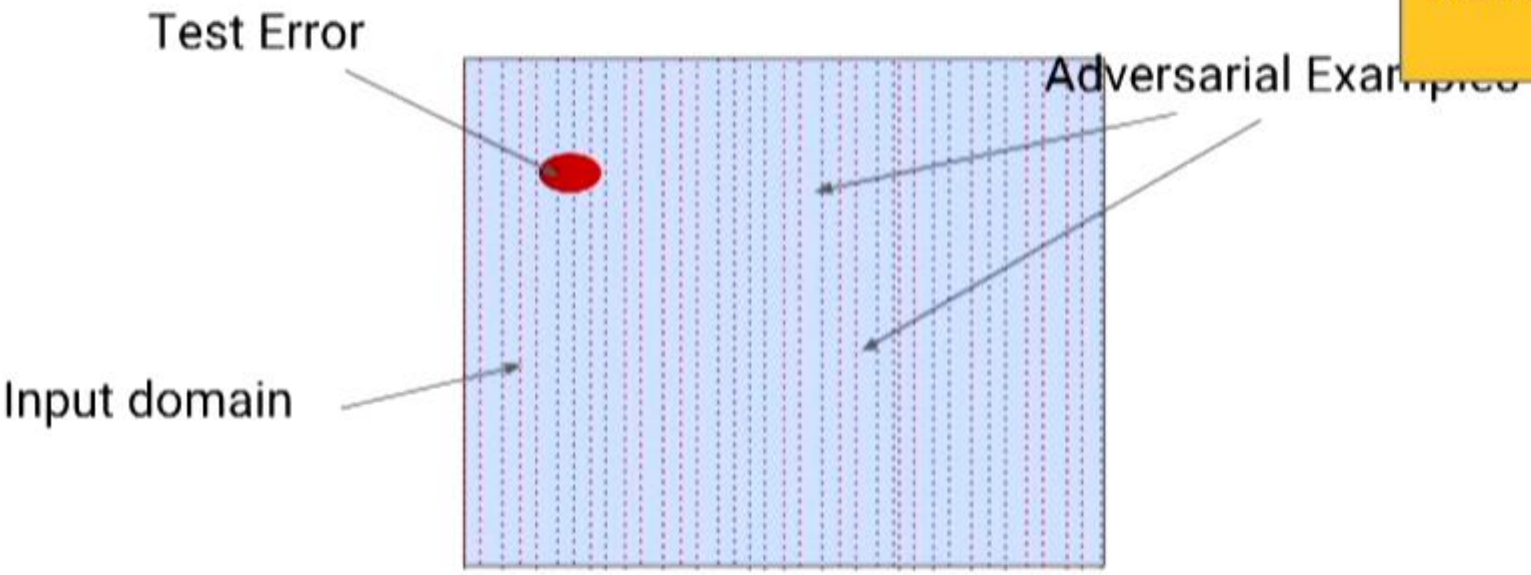
"[One] Possible explanation is that the **set of adversarial negatives** is of extremely low probability, and thus is never (or rarely) observed in the test set, yet it is dense **(much like the rational numbers)**, and so it is found near every virtually every test case."



The Wrong Mental Model

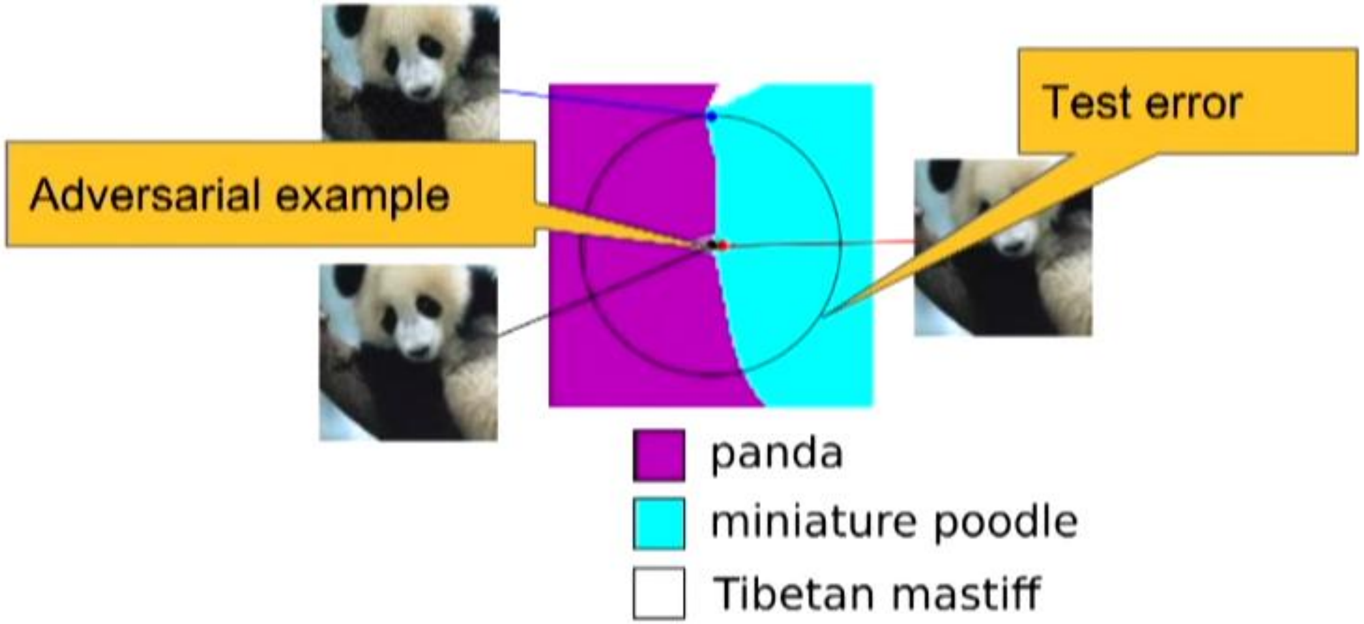
"[One] Possible explanation is that the **set of adversarial negatives** is of extremely low probability, and thus is never (or rarely) observed in the test set, yet it is dense **(much like the rational numbers)**, and so it is found near every virtually every test case."

Make a harder test set



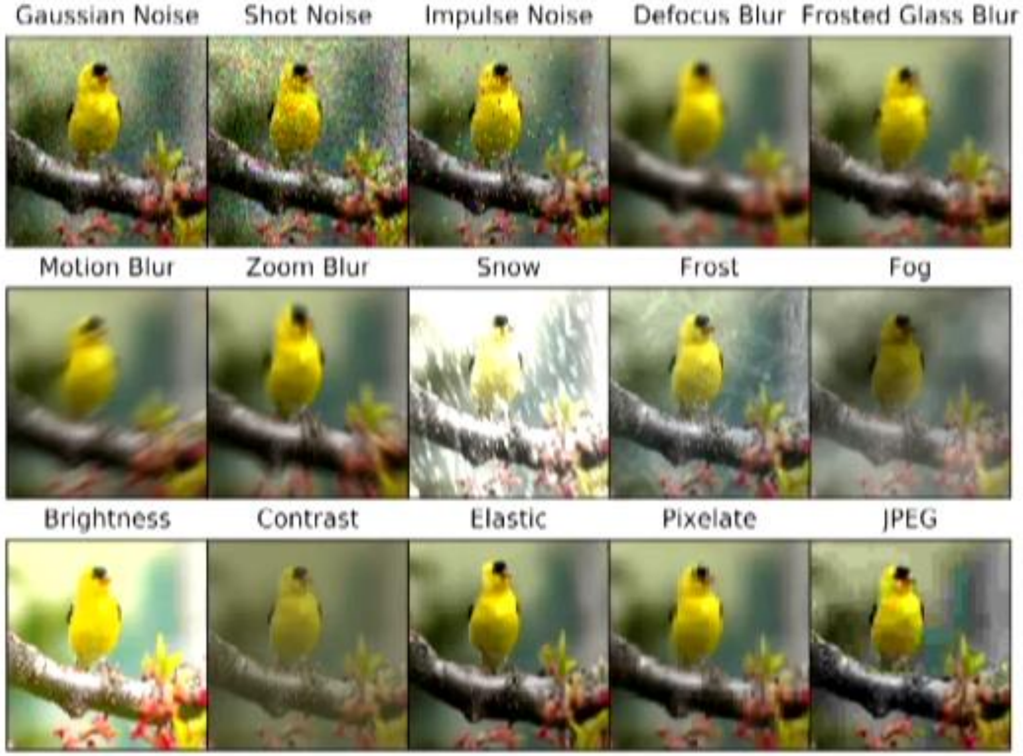
The Right Mental Model

- Adversarial examples are the nearest test error.
- Test error measures the **amount** of errors.
- The nearest error is not surprisingly close given the **amount** of errors.
- We can measure test error outside the natural distribution.
- **There is always going to be a nearest error.**



Evaluate Robustness to Distributional Shift

- Robustness to distributional shift is the *real* problem here.
- If you disagree, at least measure both for the sake of **science.**
- It's a critical sanity check for the vanishing gradient problem.



Hendrycks et. al.