# Ensemble Adversarial Training
## Attacks and Defenses

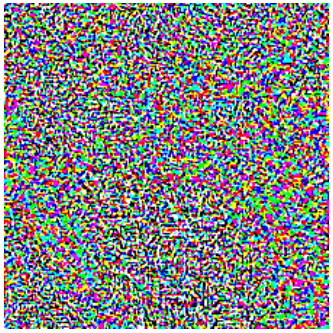
**Cybersecurity With The Best**
October 15th 2017

Florian Tramèr
Stanford

*Joint work with*
Alexey Kurakin (Google Brain)
Nicolas Papernot (PSU)
Dan Boneh (Stanford)
Patrick McDaniel (PSU)

# Adversarial Examples in ML

$+ .007 \times$



$=$

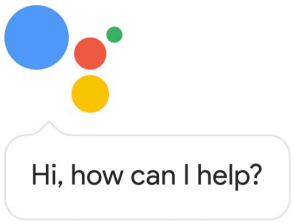
**Pretty sure this is a panda**

**I'm certain this is a gibbon**

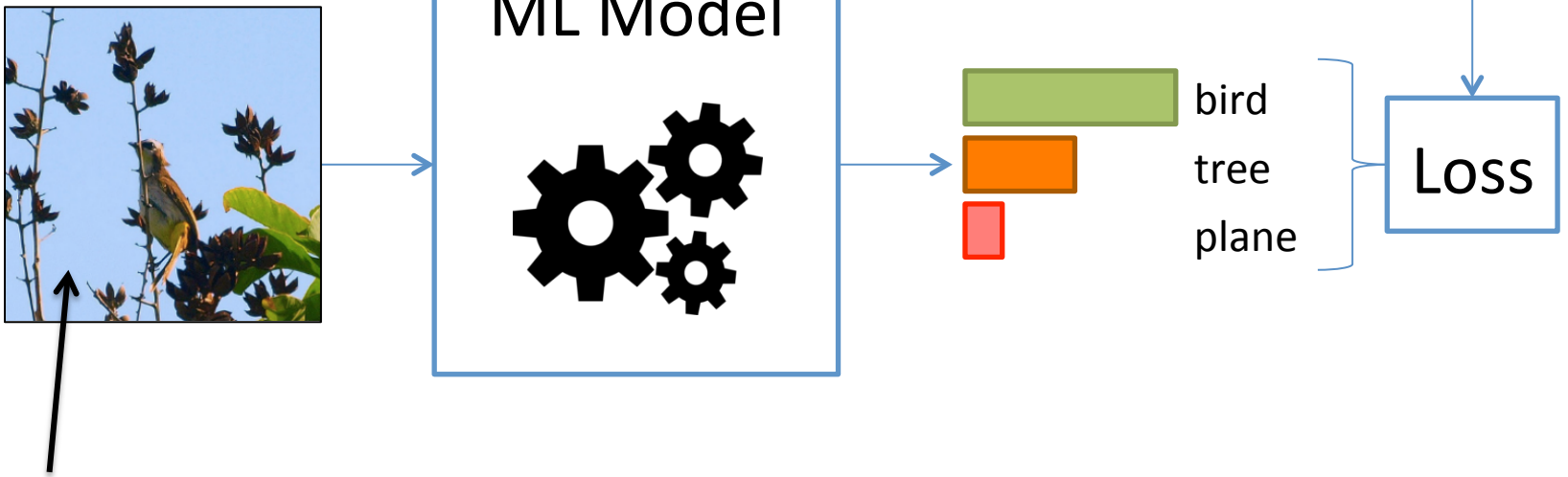
**(Goodfellow et al. 2015)**

# Adversarial Examples in ML

- **Images**
  Szegedy et al. 2013, Nguyen et al. 2015, Goodfellow et al. 2015, Papernot et al. 2016, Liu et al. 2016, Kurakin et al. 2016, …
- **Physical Objects**
  Sharif et al. 2016, Kurakin et al. 2017, Evtimov et al. 2017, Lu et al. 2017
- **Malware**
  Šrndić & Laskov 2014, Xu et al. 2016, Grosse et al. 2016, Hu et al. 2017
- **Text Understanding**
  Papernot et al. 2016, Jia & Liang 2017
- **Speech**
  Carlini et al. 2015, Cisse et al. 2017




# Creating an adversarial example




What happens if I nudge this pixel?

# Creating an adversarial example

ML Model

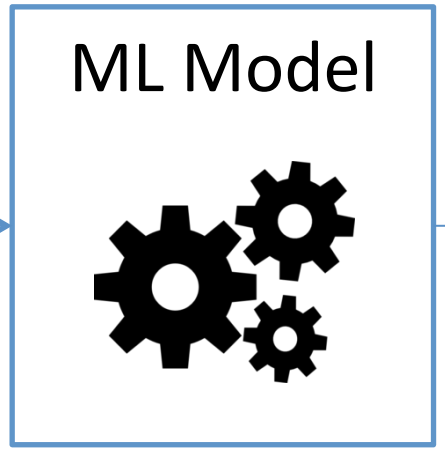
bird

bird
tree
plane

Loss

What happens if I change this pixel?

# Creating an adversarial example




bird
tree
plane

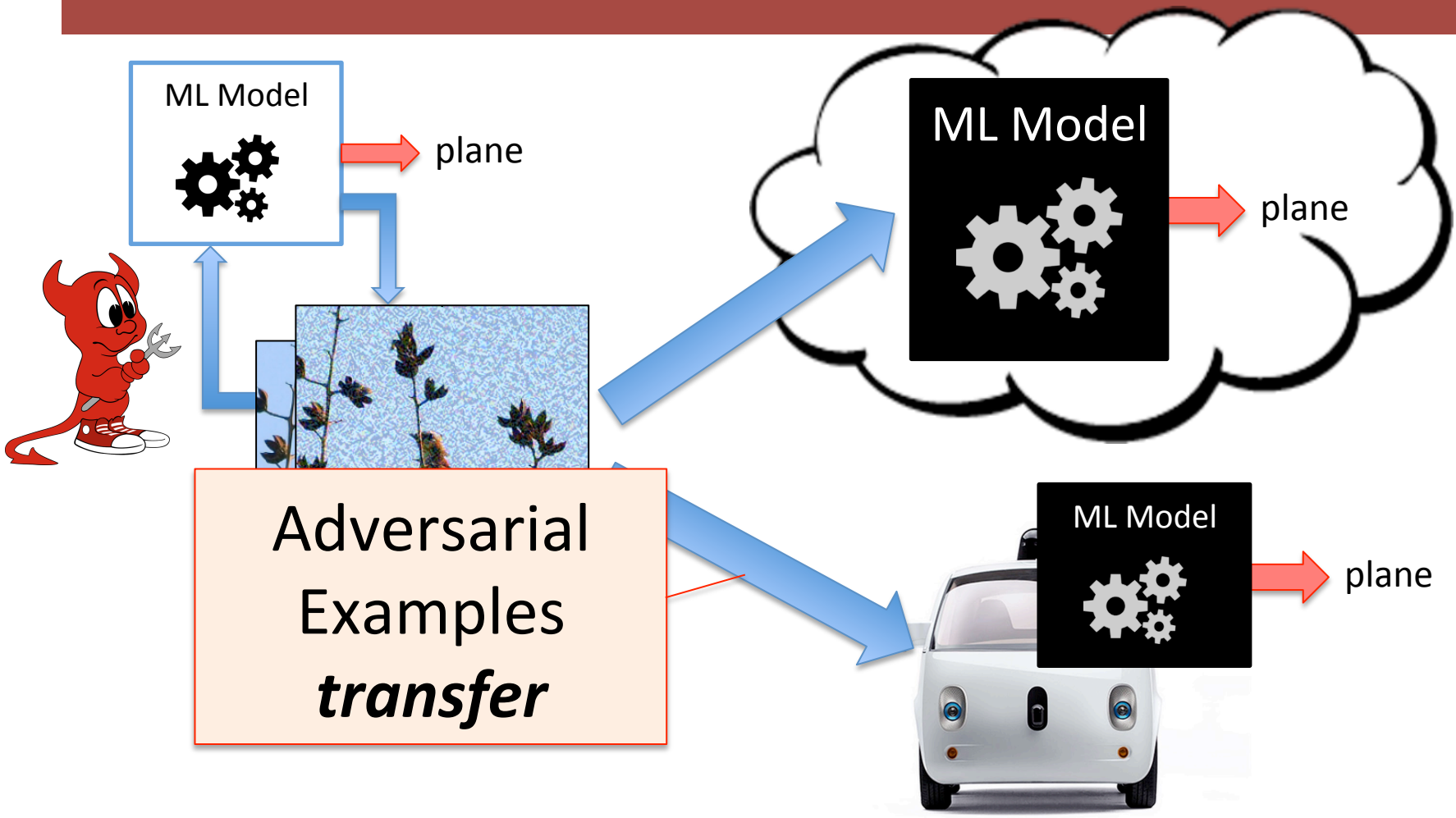
bird

Loss

What about this one?

Maximize loss with gradient *ascent*

# Threat Model: Black-Box Attacks



# Defenses?

- Ensembles ✗
- Preprocessing (blurring, cropping, etc.) ✗
- Distillation ✗
- Generative modeling ✗
- Adversarial training ?

# Adversarial Training

ML Model

bird

Loss

attack

ML Model

plane

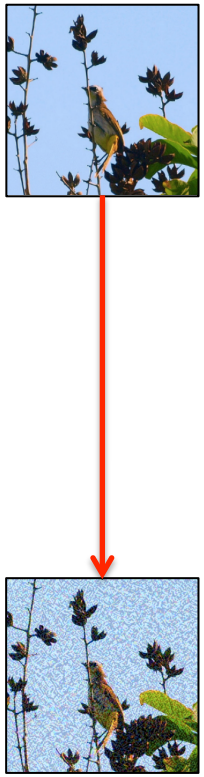
Loss

Repeat

# Adversarial Training - Tradeoffs

**“weak” attack**

single step



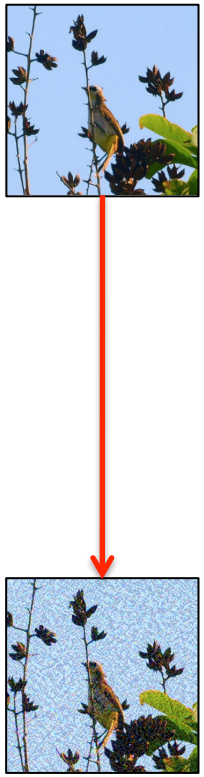
**“strong” attack**

many steps

# Adversarial Training - Tradeoffs

**“weak” attack**

fast



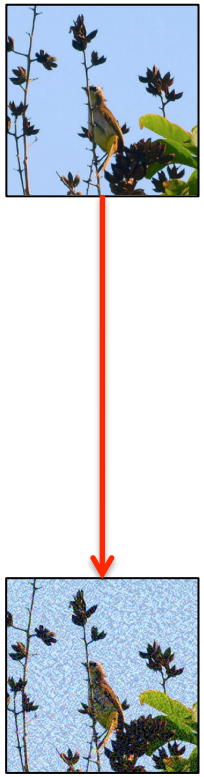
**“strong” attack**

slow

# Adversarial Training - Tradeoffs

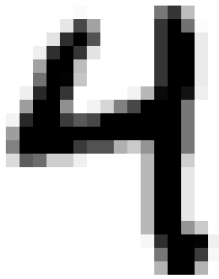
**“weak” attack**

not infallible but scalable



**“strong” attack**

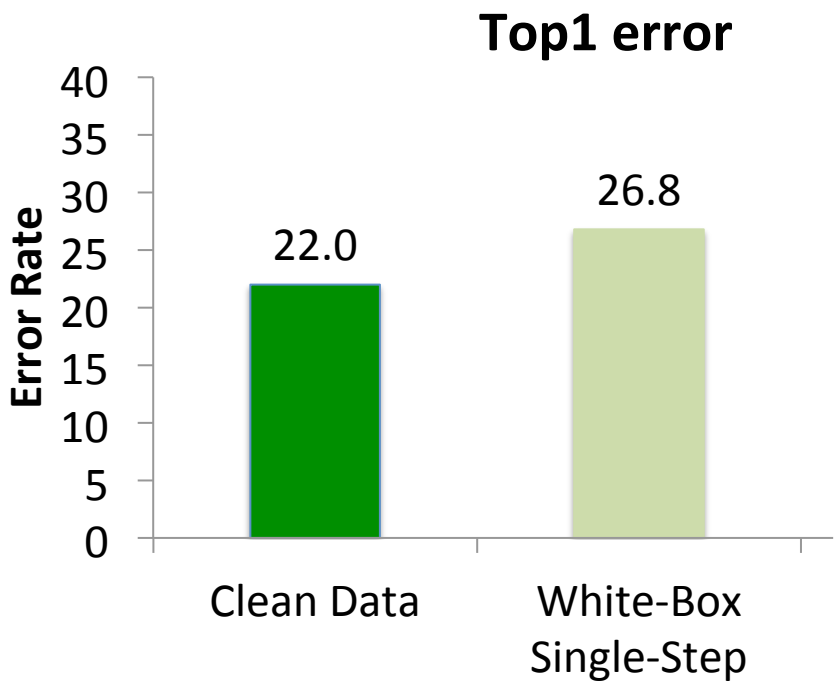
learn robust models on small datasets



Madry et al. 2017

# Adversarial Training on ImageNet

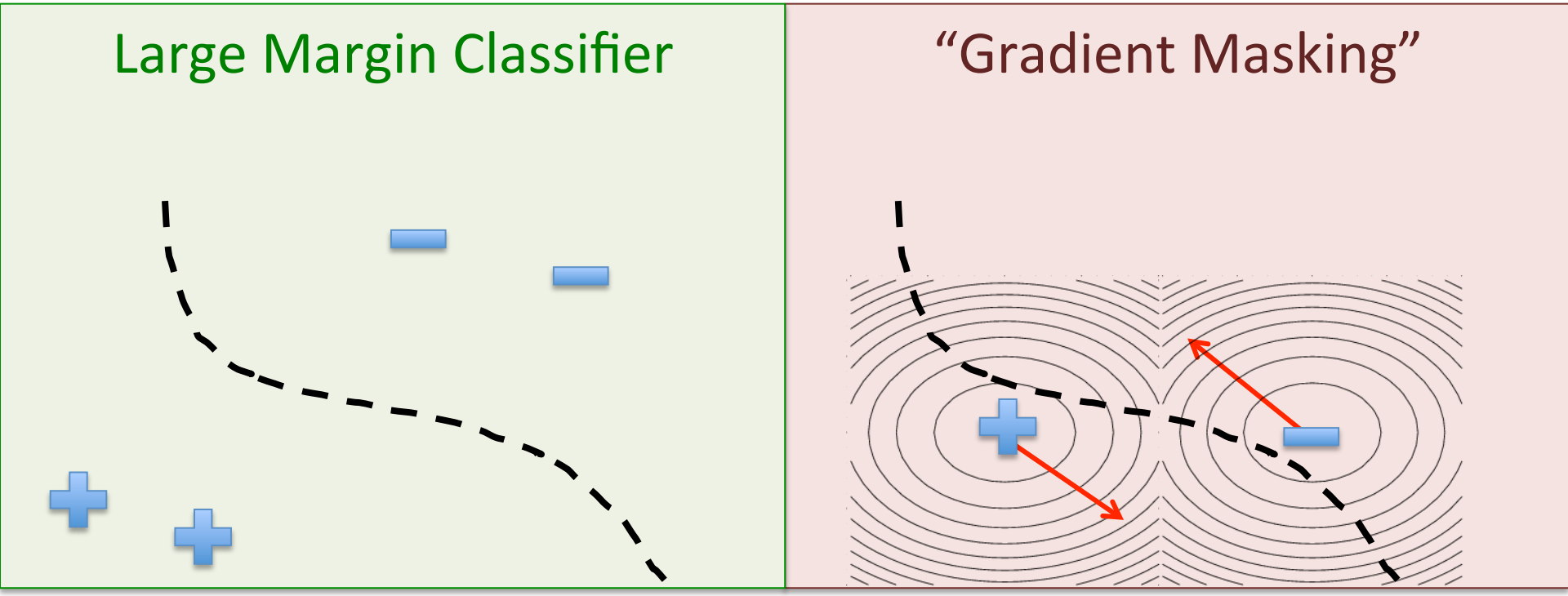
- Adversarial training with single-step attack
  (Kurakin et al. 2016)



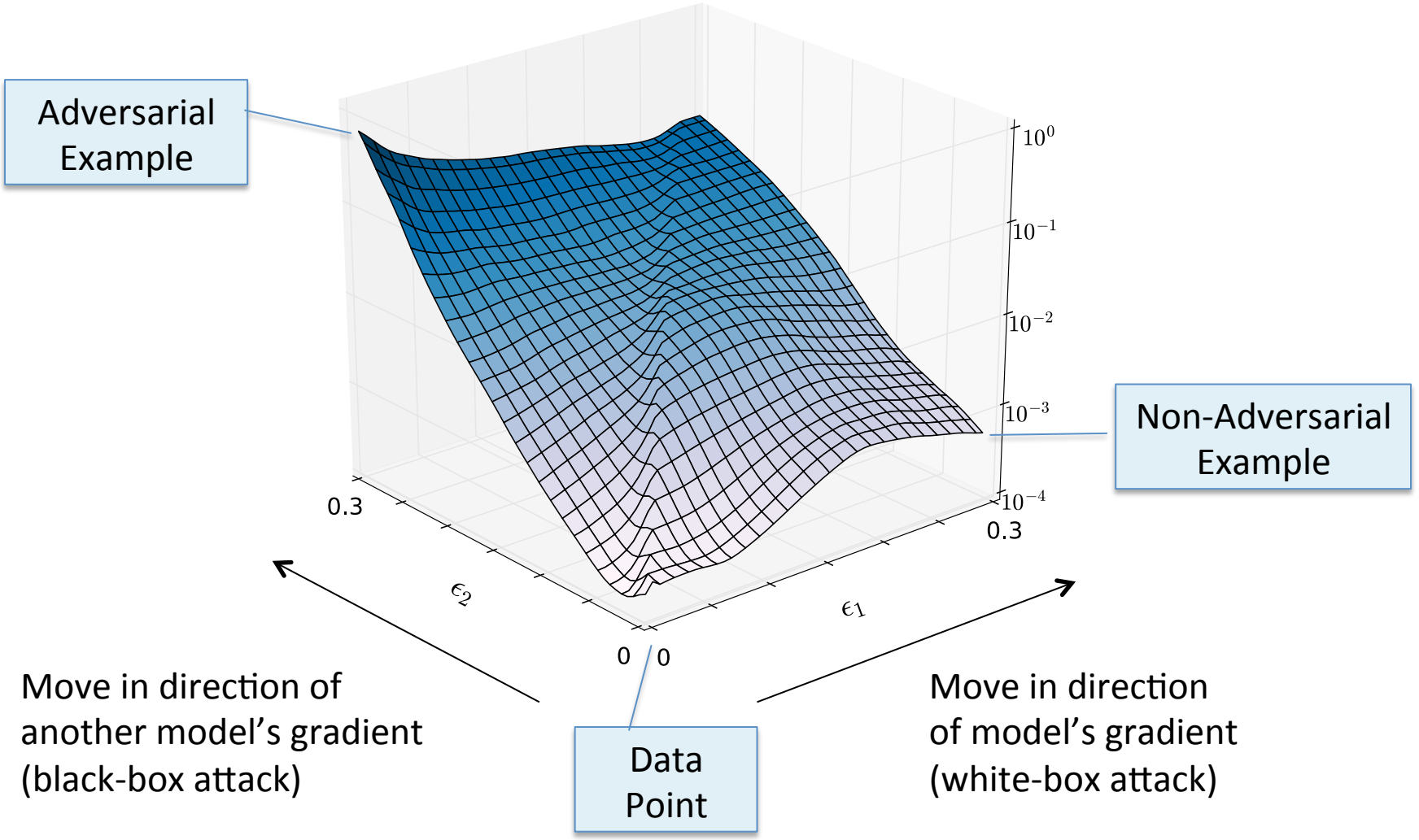

Adversarial examples transferred from another model

# What's happening? Gradient Masking!

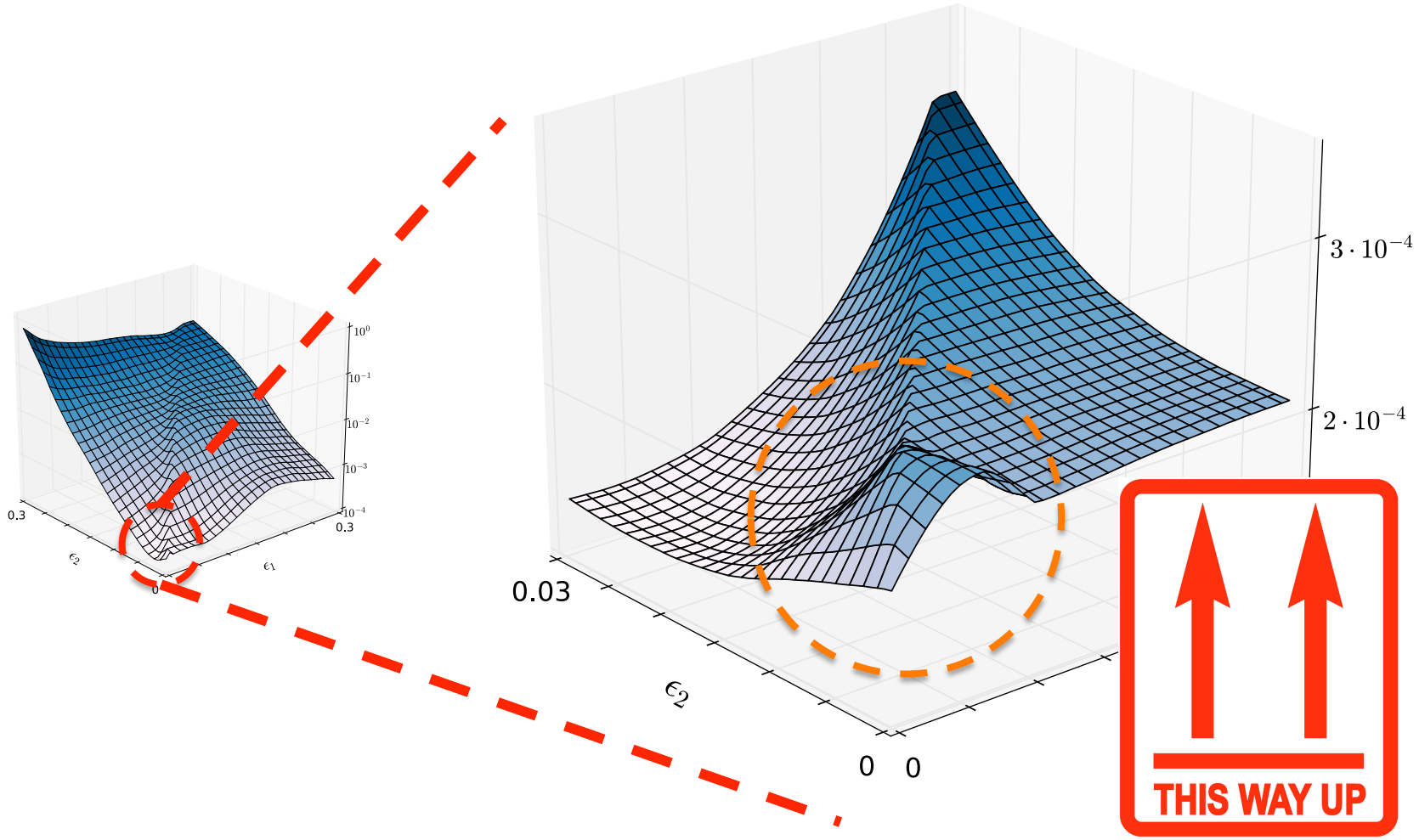
- How to get robustness to single-step attacks?



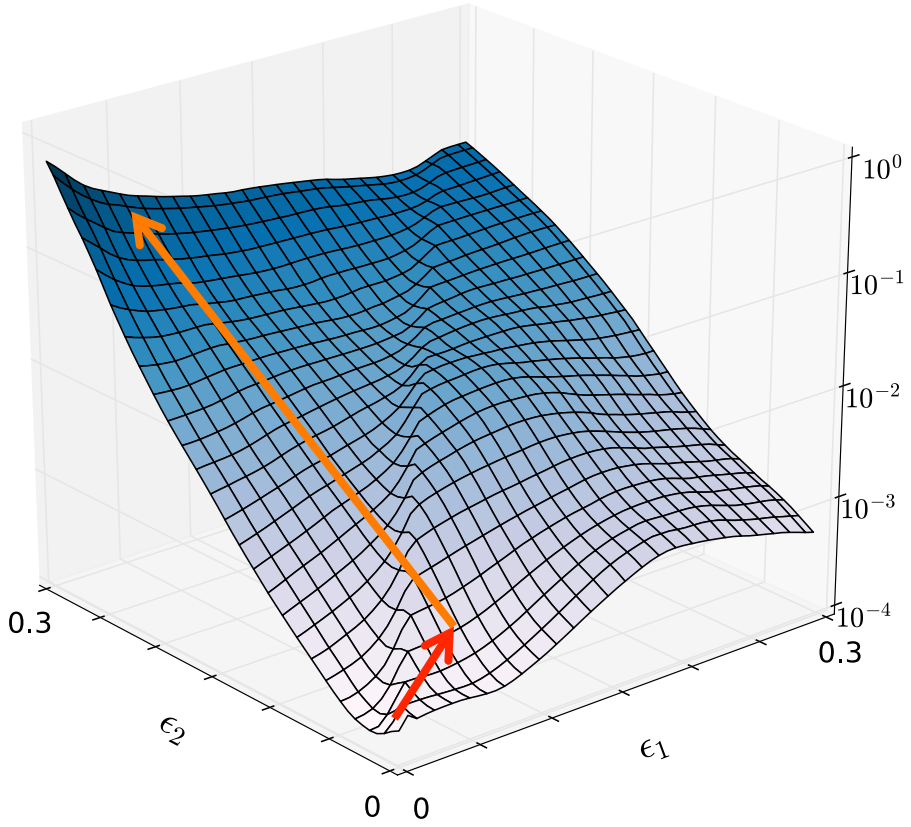
# Loss of Adversarially Trained Model

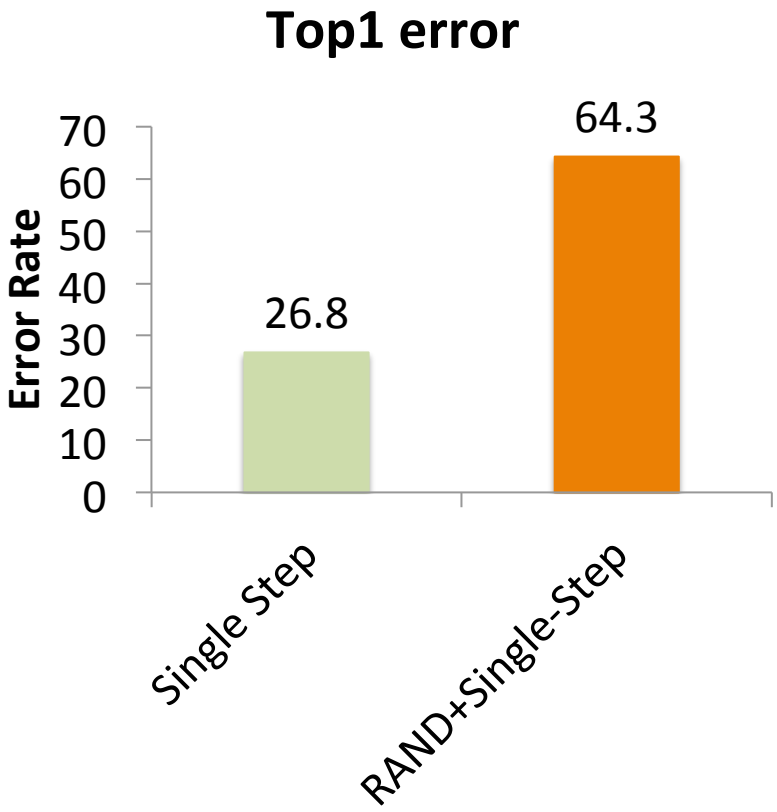


# Loss of Adversarially Trained Model



# Simple Attack: RAND+Single-Step




1. Small random step
2. Step in direction of gradient

# What's wrong with "Single-Step" Adversarial Training?

Minimize:

```
self.loss(self.attack(  ))
```

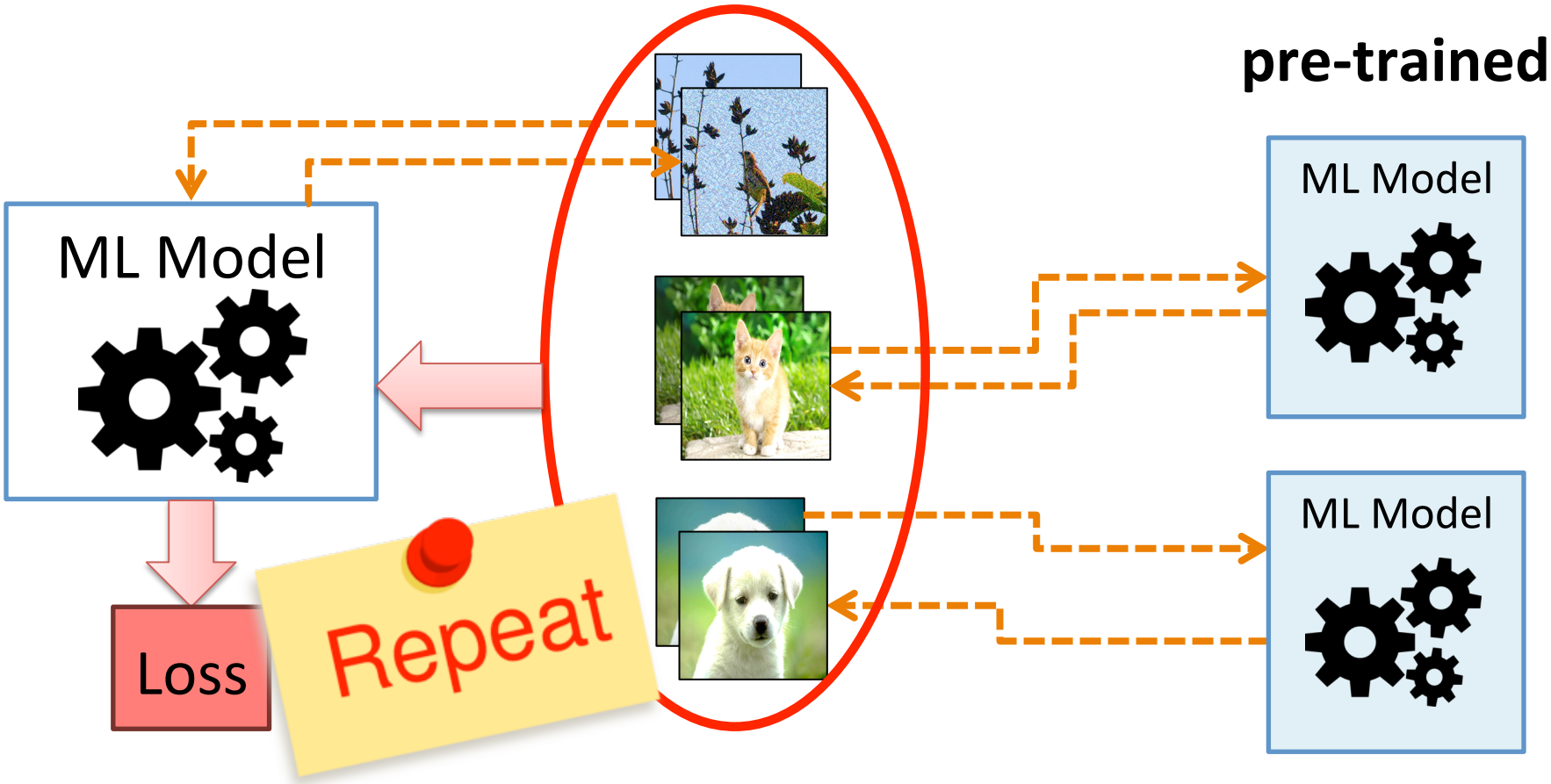
Solution:

1. The model is actually robust
2. *Or, the attack is really bad*

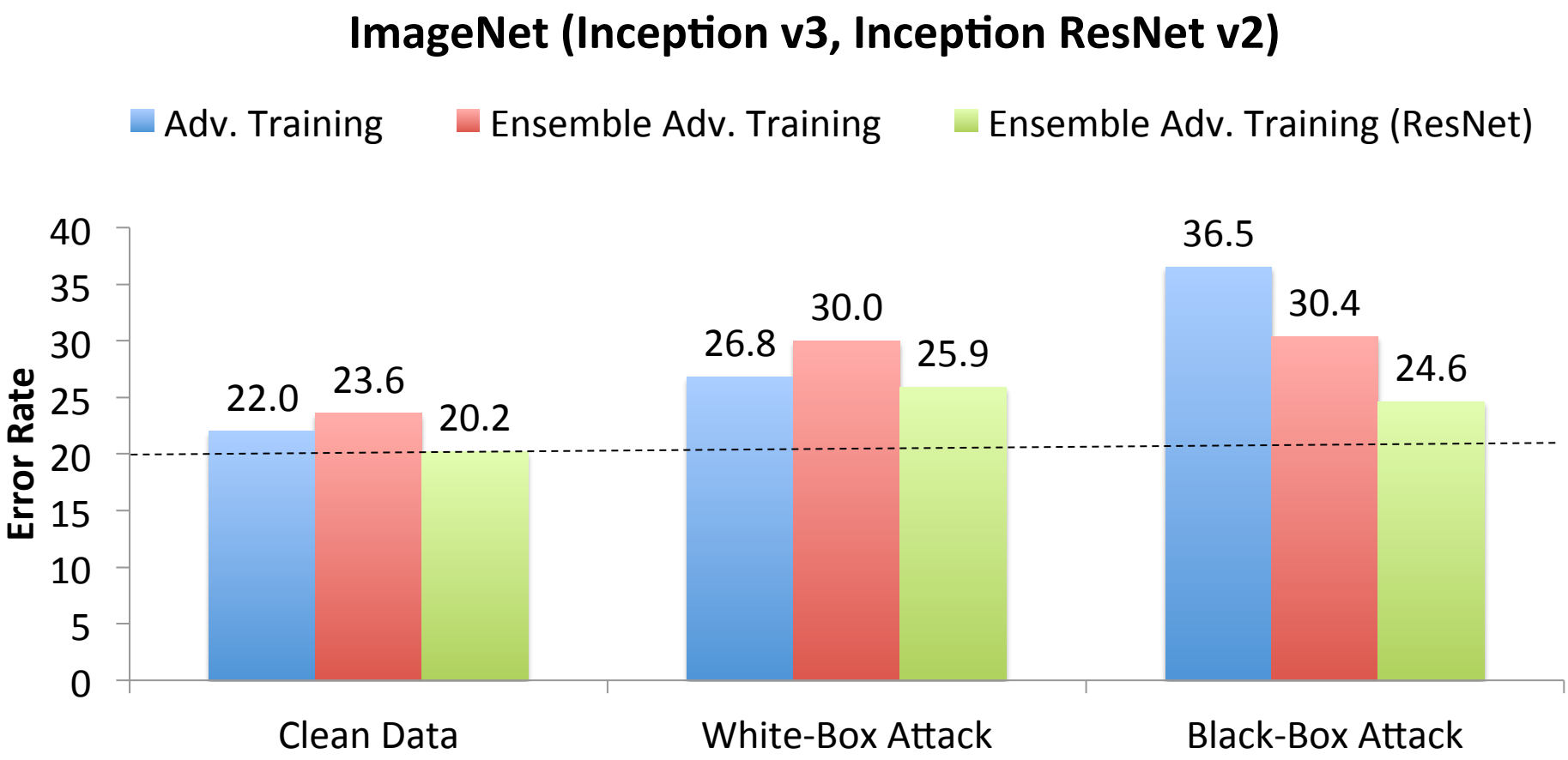
Degenerate Minimum

Better approach? *decouple* attack and defense

# Ensemble Adversarial Training



# Results


ImageNet (Inception v3, Inception ResNet v2)
Adv. Training
Ensemble Adv. Training
Ensemble Adv. Training (ResNet)
Error Rate
40
35
30
25
20
15
10
5
0
22.0
23.6
20.2
26.8
30.0
25.9
36.5
30.4
24.6
Clean Data
White-Box Attack
Black-Box Attack

# What about stronger attacks?

- Little gain on **strong white-box** attacks!

- But, **improvements in black-box setting**!

**NIPS 2017: Defense Against Adversarial Attack**

Create an image classifier that is robust to adversarial attacks

| KaggleTeamId | TeamName | Score |
|---|---|---|
| baseline_ens_adv_inception_renset_v2 | | 71506 |
| 816739 | alekseynp | 67409 |
| 827701 | rwightman | 65464 |
| 802555 | tonyyy | 65193 |
| baseline_adv_inception_v3 | | 64648 |

# Open Problems

- How far can we go with adversarial training?
  - **White-box** robustness is possible! (Madry et al. 2017)
    - Caveat 1: **Very** expensive
    - Caveat 2: What is the right **metric** ($l_\infty$ , $l_2$ , rotations) ?
- Can we say **anything** formal (and useful) about adversarial examples?
  - Why do they exist? Why do they transfer?

**THANK YOU**

# Related Work

**Adversarial training + black-box attacks:**

| | | |
|---|---|---|
| Szegedy et al., | https://arxiv.org/abs/1312.6199 | original paper on adversarial examples |
| Nguyen et al., | https://arxiv.org/abs/1412.1897 | a genetic algorithm for adversarial examples |
| Goodfellow et al., | https://arxiv.org/abs/1412.6572 | adversarial training with single-step attacks |
| Papernot et al., | https://arxiv.org/abs/1511.04508 | the distillation defense |
| Papernot et al., | https://arxiv.org/abs/1602.02697 | black-box attacks, model reverse-engineering |
| Liu et al., | https://arxiv.org/abs/1611.02770 | black-box attacks on ImageNet |
| Kurakin et al., | https://arxiv.org/abs/1611.01236 | adversarial training on ImageNet |
| Tramer et al., | https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer | |

(model reverse-engineering)

| | | |
|---|---|---|
| Madry et al., | https://arxiv.org/abs/1706.06083 | learning robust models with strong attacks |
| Tramer et al., | https://arxiv.org/abs/1705.07204 | our paper |

**Physical world:**

| | | |
|---|---|---|
| Sharif et al., | https://dl.acm.org/citation.cfm?id=2978392 | fooling facial recognition with glasses |
| Kurakin et al., | https://arxiv.org/abs/1607.02533 | physical-world adversarial examples |
| Lu et al., | https://arxiv.org/abs/1707.03501 | self driving cars will be fine |
| Estimov et al., | https://arxiv.org/abs/1707.08945 | maybe they won't! |

# Related Work (cont.)

**Malware:**

Srndic et al., https://dl.acm.org/citation.cfm?id=2650798 fooling a pdf-malware detector

Xu et al., https://www.cs.virginia.edu/yanjun/paperA14/2016-evade_classifier.pdf (same as above)

Grosse et al., https://arxiv.org/abs/1606.04435 adversarial examples for Android malware

Hu et al., https://arxiv.org/abs/1702.05983 adversarial examples for Android malware

**Text:**

Papernot et al., https://arxiv.org/abs/1604.08275 adversarial examples for text understanding

Jia et al., https://arxiv.org/abs/1707.07328 adversarial examples for reading comprehension

**Speech:**

Carlini et al., https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini (fooling a voice assistant)

Cisse et al., https://arxiv.org/abs/1707.05373 adversarial examples for speech, segmentation, etc

**Reinforcement Learning:**

Huang et al., https://arxiv.org/abs/1702.02284 adversarial examples for neural network policies

Kos et al., https://arxiv.org/abs/1705.06452 adversarial examples for neural network policies