RSAConference2019
San Francisco | March 4–8 | Moscone Center

SESSION ID: MLAI-W03

Attacking Machine Learning: On the *Security* and *Privacy* of Neural Networks

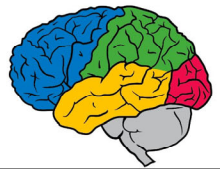
Nicholas Carlini

Research Scientist,
Google Brain

RSAConference2019

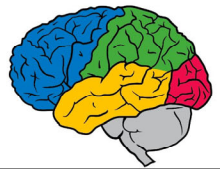
Act I:
On the Security and Privacy of Neural Networks

Let's play a game



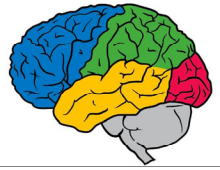
67% it is a

Great Dane



83% it is a

Old English Sheepdog



78% it is a

Greater Swiss Mountain Dog

99.99% it is

Guacamole

99.99% it is a

Golden Retriever

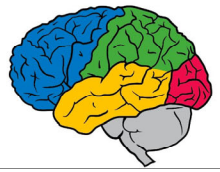
99.99% it is

Guacamole

76% it is a

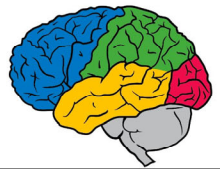
45 MPH Sign

K Eykholt, I Evtimov, E Fernandes, B Li, A Rahmati, C Xiao, A Prakash, T Kohno, D Song.
Robust Physical-World Attacks on Deep Learning Visual Classification. 2017



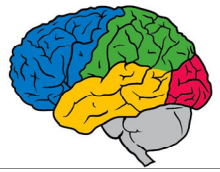
Machine Learning for Security & Privacy

Security & Privacy of Machine Learning



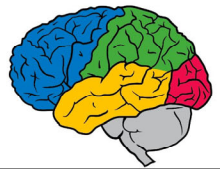
Machine Learning
for
Security & Privacy

Security & Privacy
of
Machine Learning

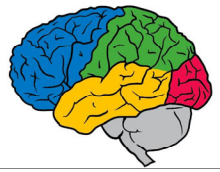


Adversarial Examples

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. 2014.
I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 2015.

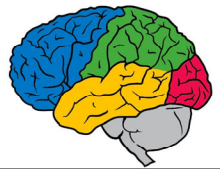


K Eykholt, I Evtimov, E Fernandes, B Li, A Rahmati, C Xiao, A Prakash, T Kohno, D Song.
Robust Physical-World Attacks on Deep Learning Visual Classification. 2017



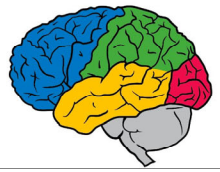
What do you think this transcribes as?

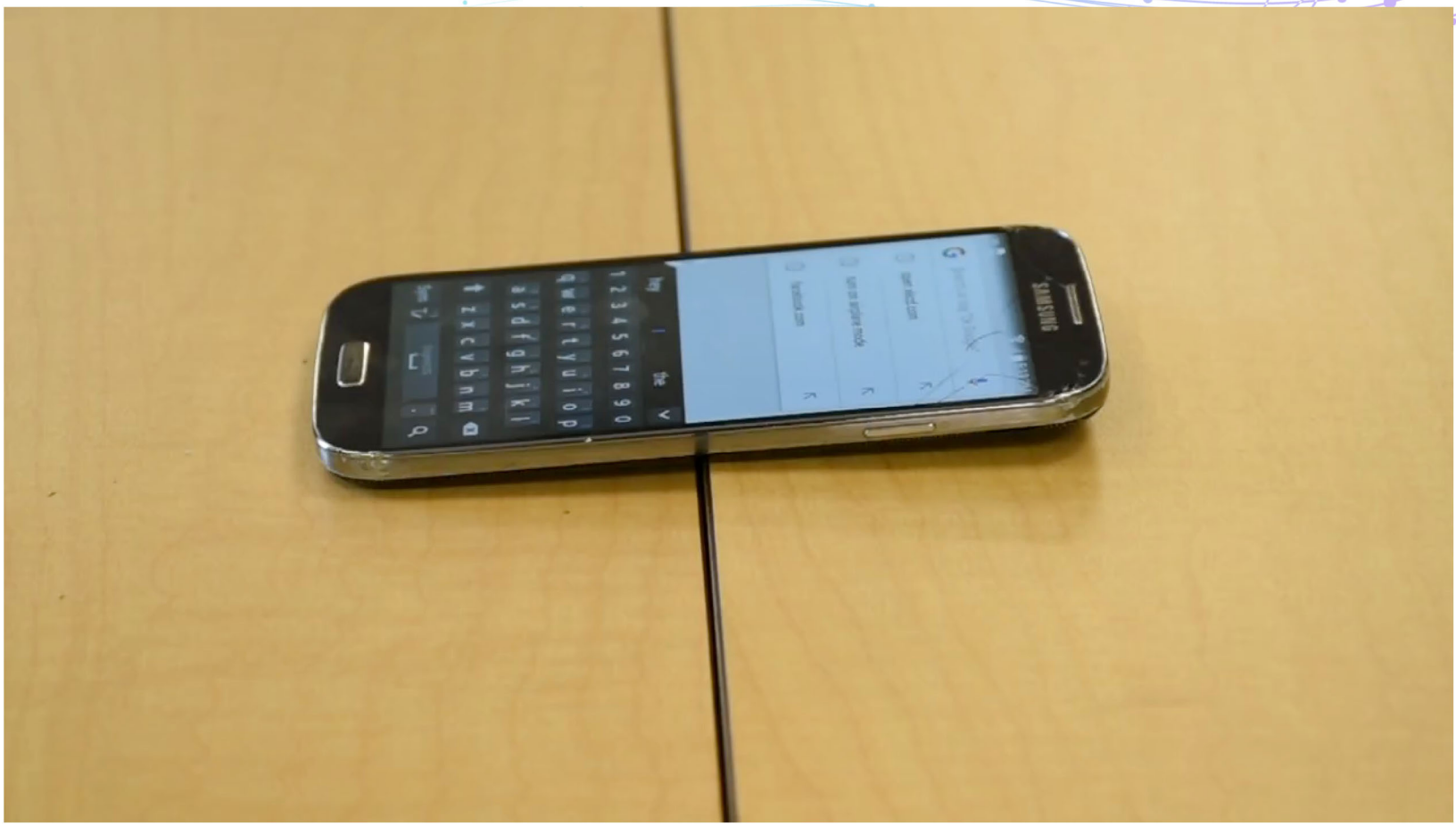
N Carlini, D Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. 2018



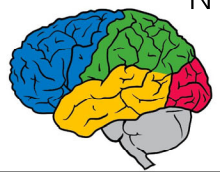
"It was the best of times,
it was the worst of times,
it was the age of wisdom,
it was the age of foolishness,
it was the epoch of belief,
it was the epoch of incredulity"

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

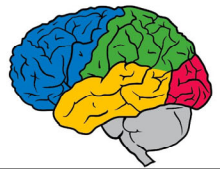




N Carlini, P Mishra, T Vaidya, Y Zhang, M Sherr, C Shields, D Wagner, W Zhou. Hidden Voice Commands. 2016

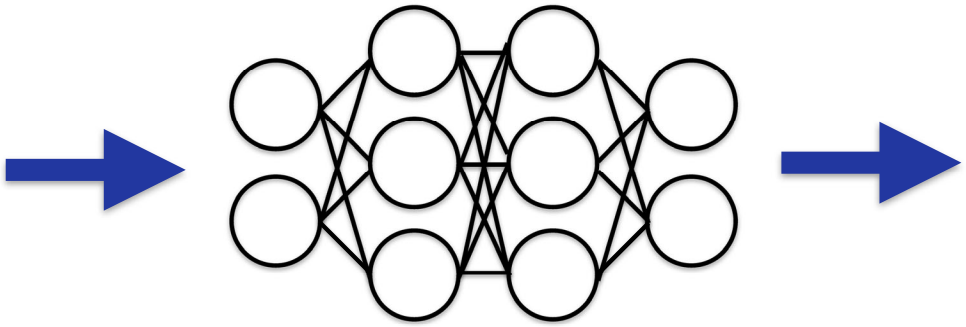


This problem will only become more important

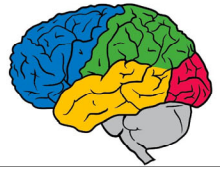


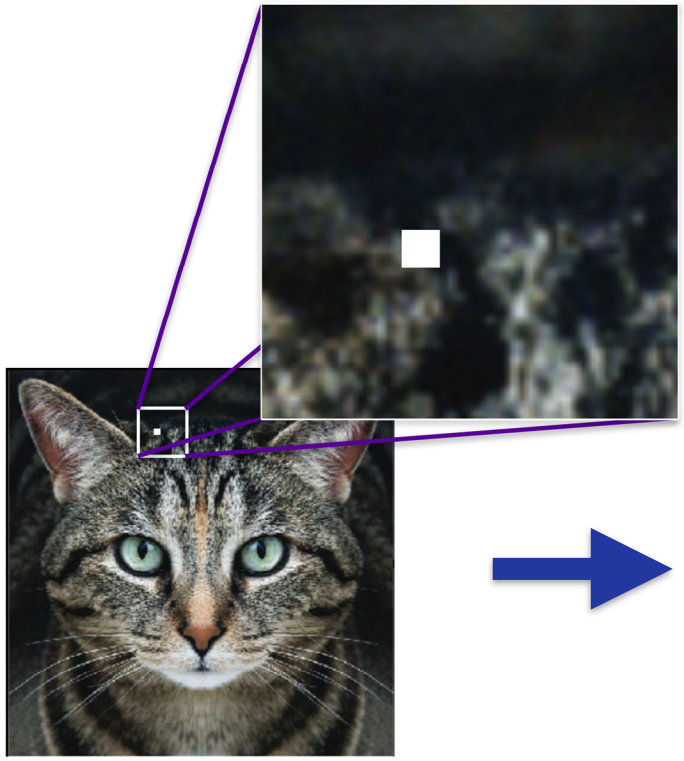
RSAConference2019

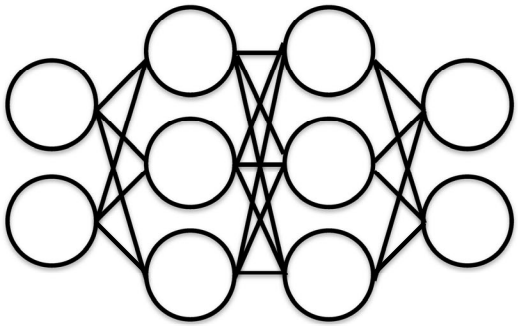
Constructing Adversarial Examples



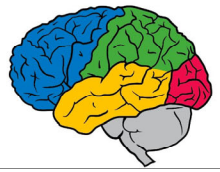
[0.9,
0.1]

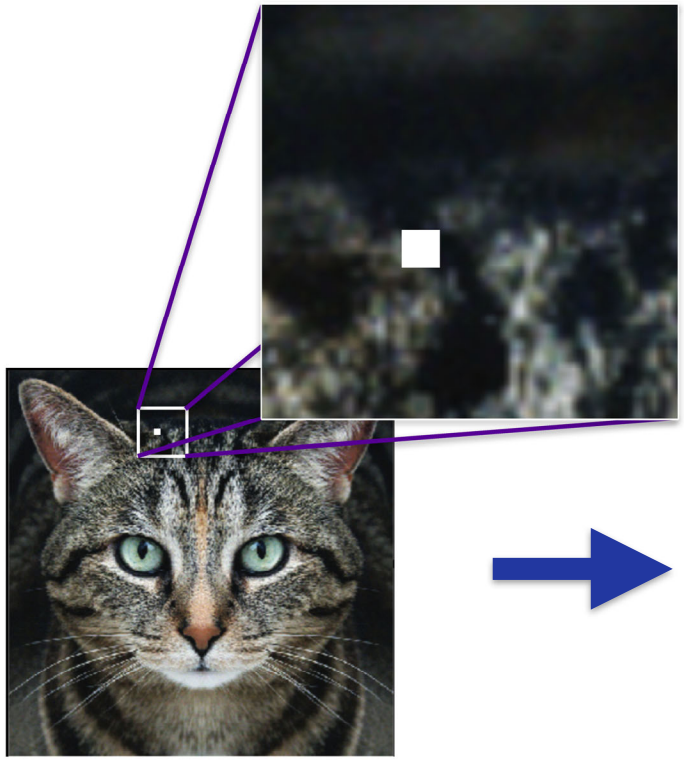


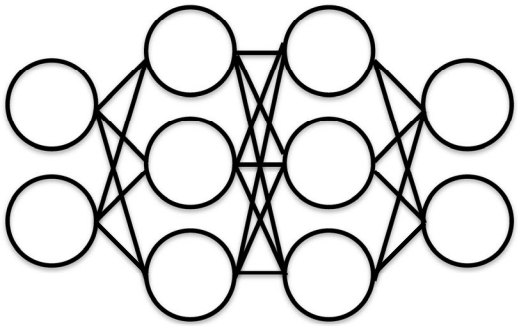




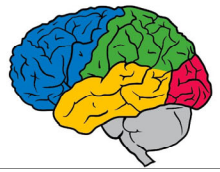
[0.9,
0.1]

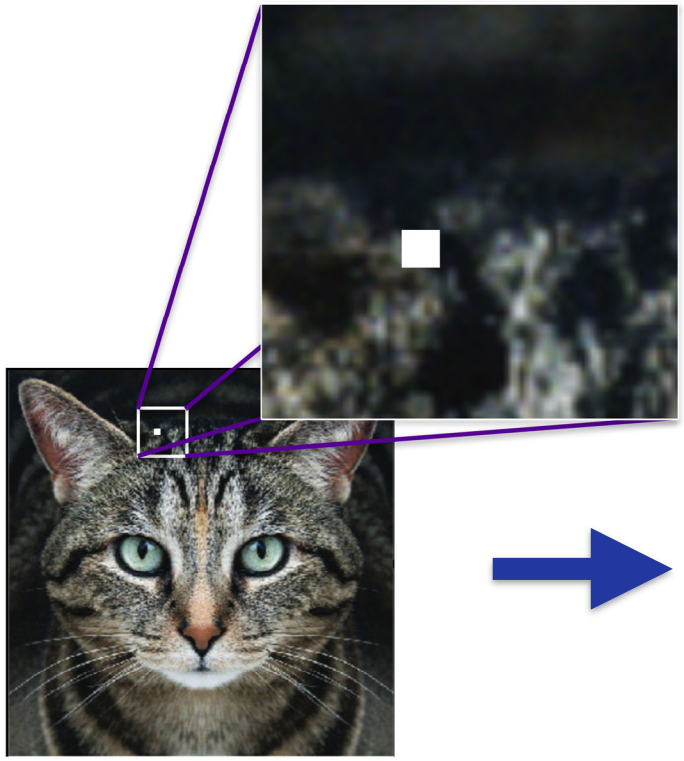


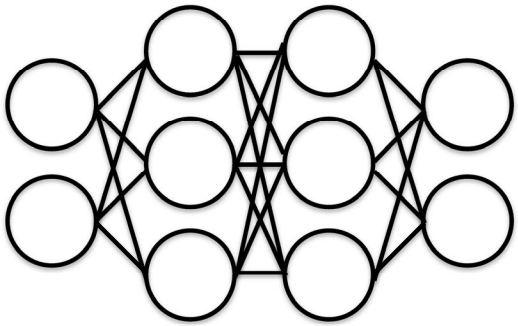




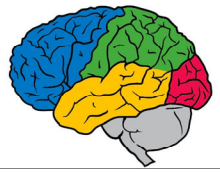
[0.89, 0.11]

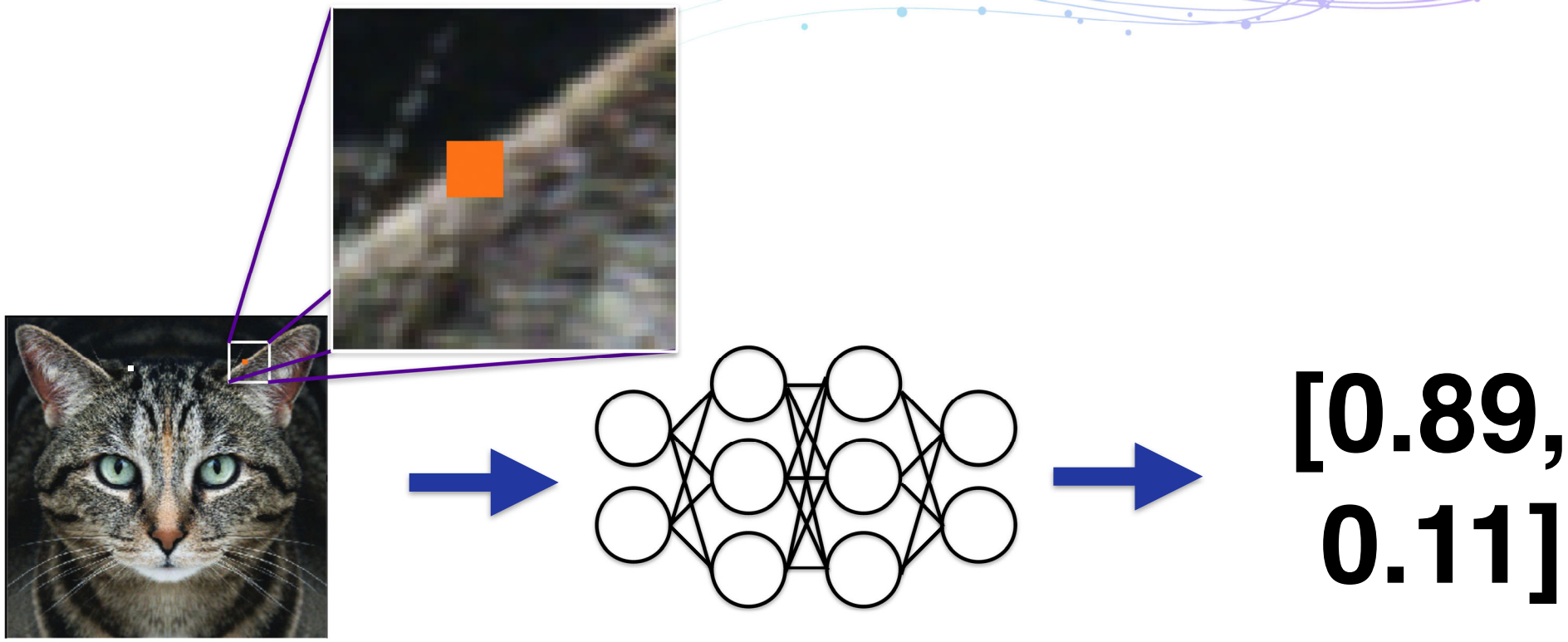


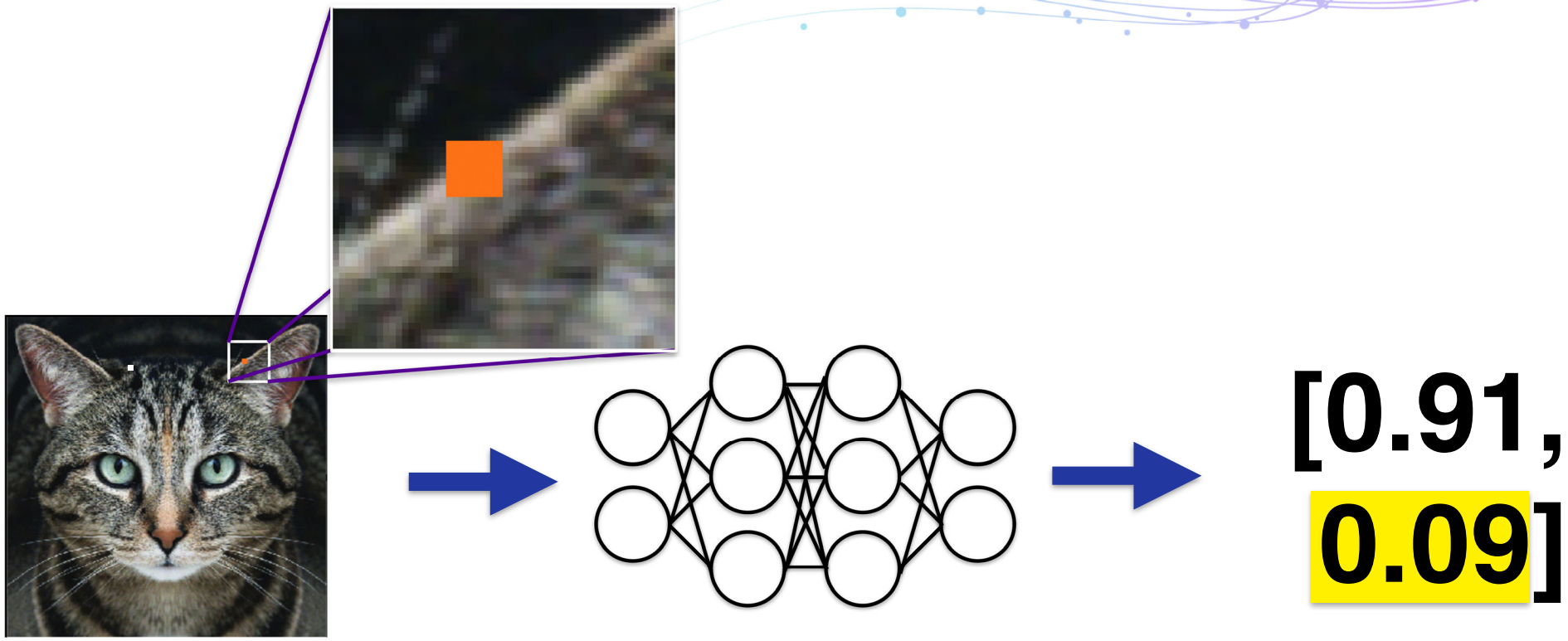


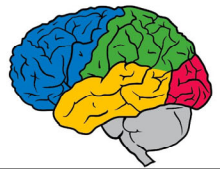


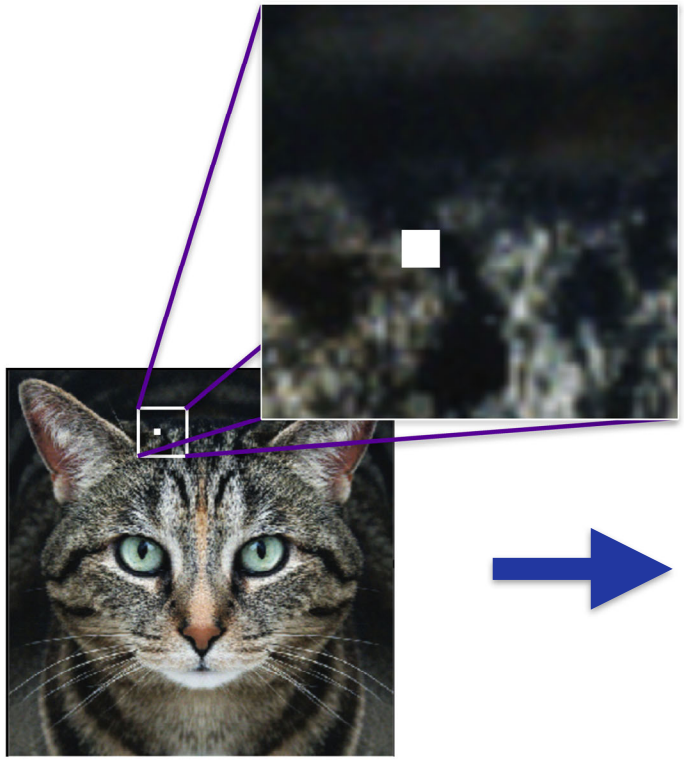
[0.89,
0.11]

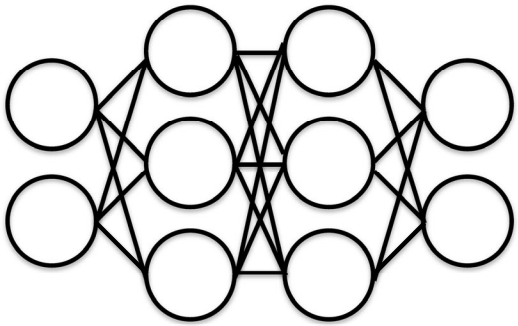



[0.89,
0.11]

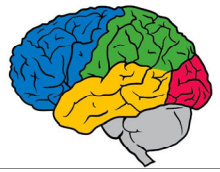

[0.91,
0.09]

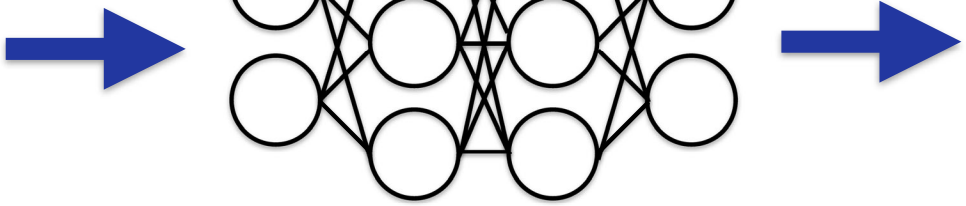




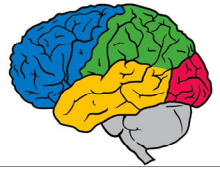


[0.89,
0.11]



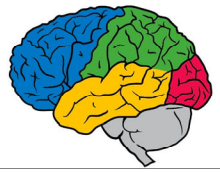


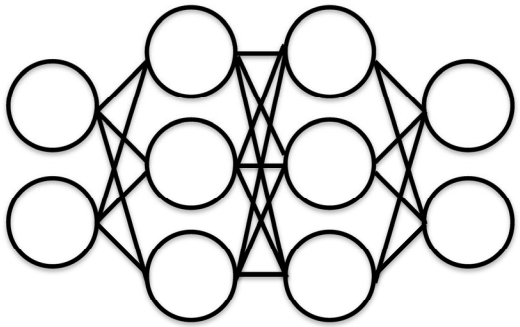
[0.48,
0.52]



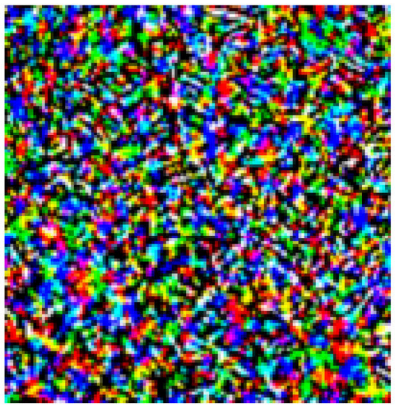
This *does* work ...

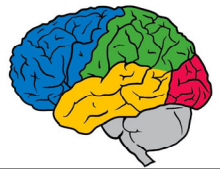
... but we have **calculus**!





$$-\frac{\partial}{\partial x}$$

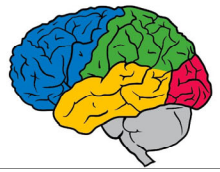




+ .001×

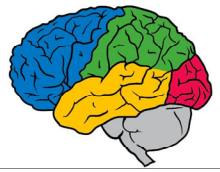
=

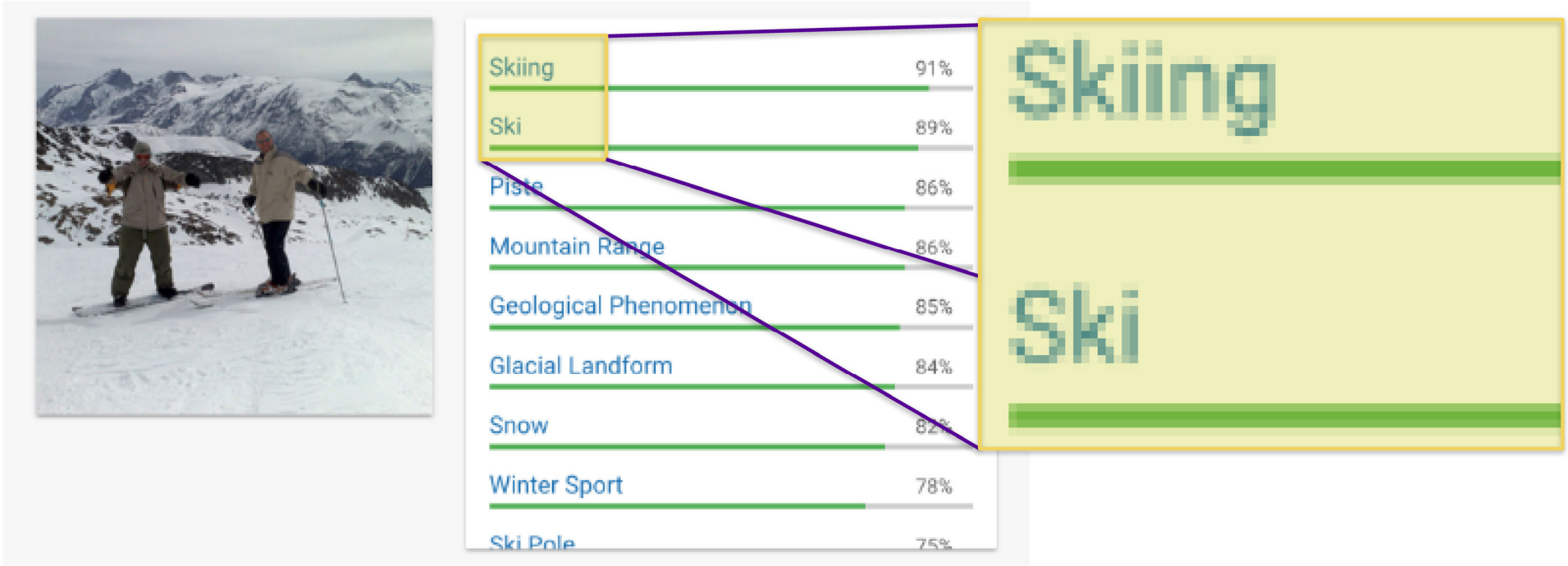
CAT **adversarial perturbation** **DOG**



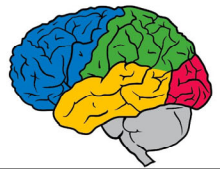
I. J. Goodfellow, J. Shlens and C. Szegedy. Explaining and harnessing adversarial examples. 2015

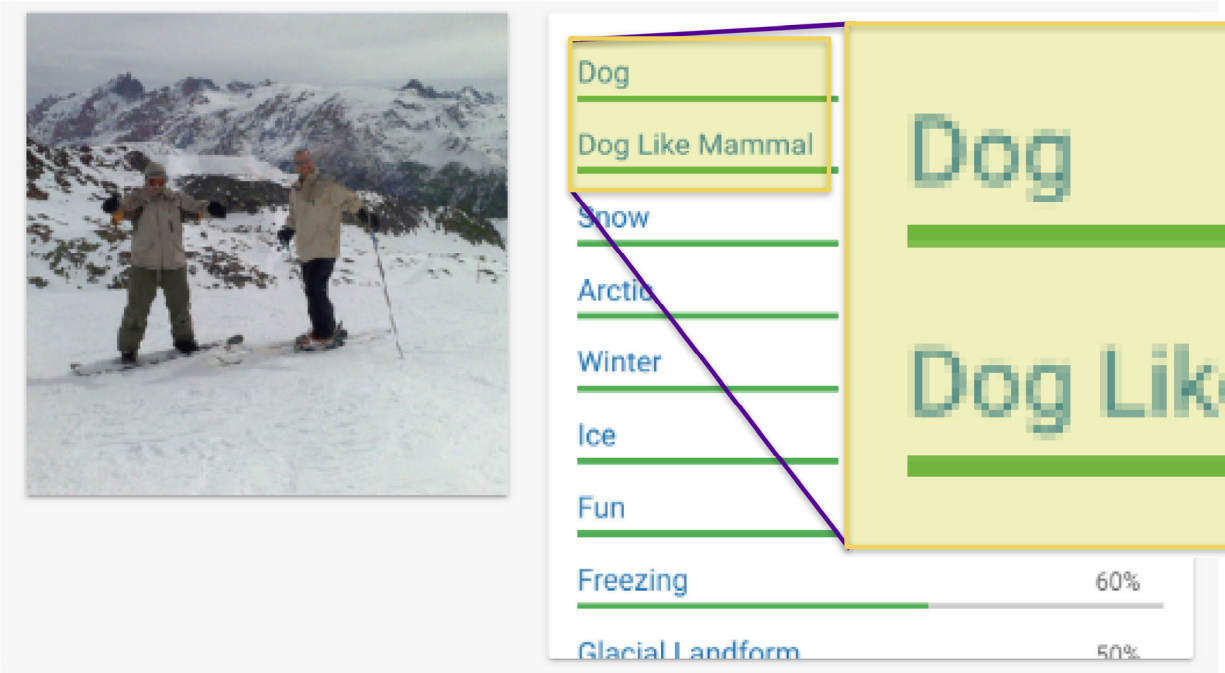
What if we don't have **direct access** to the model?



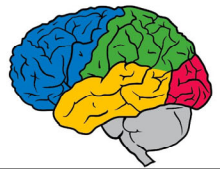


A Ilyas, L Engstrom, A Athalye, J Lin. Black-box Adversarial Attacks with Limited Queries and Information. 2018

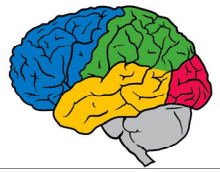




A Ilyas, L Engstrom, A Athalye, J Lin. Black-box Adversarial Attacks with Limited Queries and Information. 2018



Generating adversarial examples is **simple** and **practical**

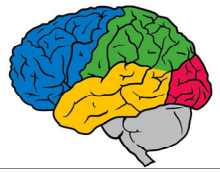


RSAConference2019

Defending against Adversarial Examples

Case Study: ICLR 2018 Defenses

A Athalye, N Carlini, D Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. 2018



MITIGATING ADVERSARIAL EFFECTS THROUGH RANDOMIZATION

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 provide
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 huma
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 SA
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
plore the proper-
lings help neural

Published as a conference paper at ICLR 2018

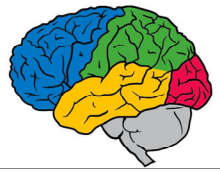
COUNTERING ADVERSARIAL IMAGES USING INPUT TRANSFORMATIONS

Chuan Guo*
Cornell University

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

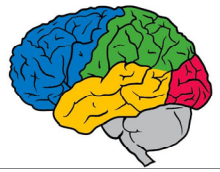
ABSTRACT

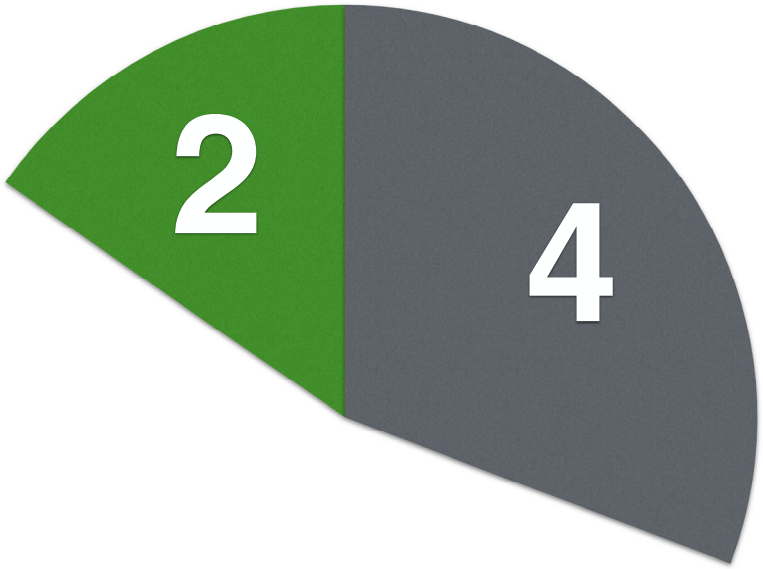
This paper investigates strategies that defend against adversarial-example attacks on image-classification systems by transforming the inputs before feeding them to the system. Specifically, we study applying image transformations such as bit-depth reduction, JPEG compression, total variance minimization, 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.*



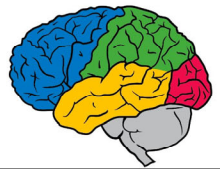
4

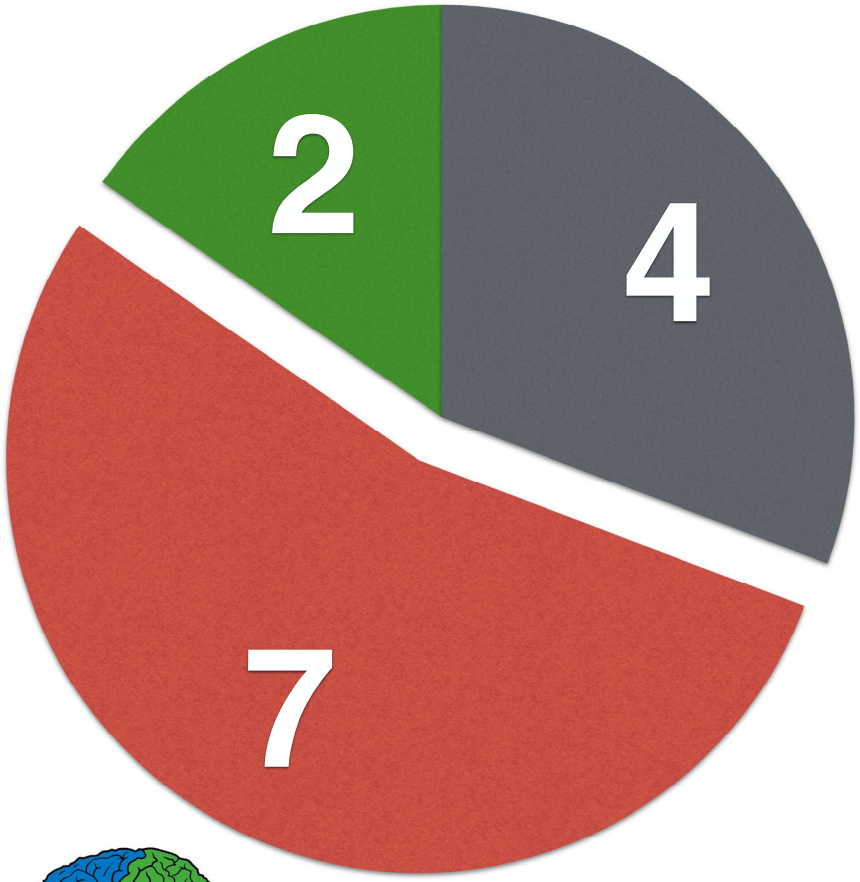
- Out of scope



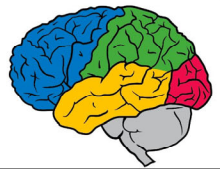


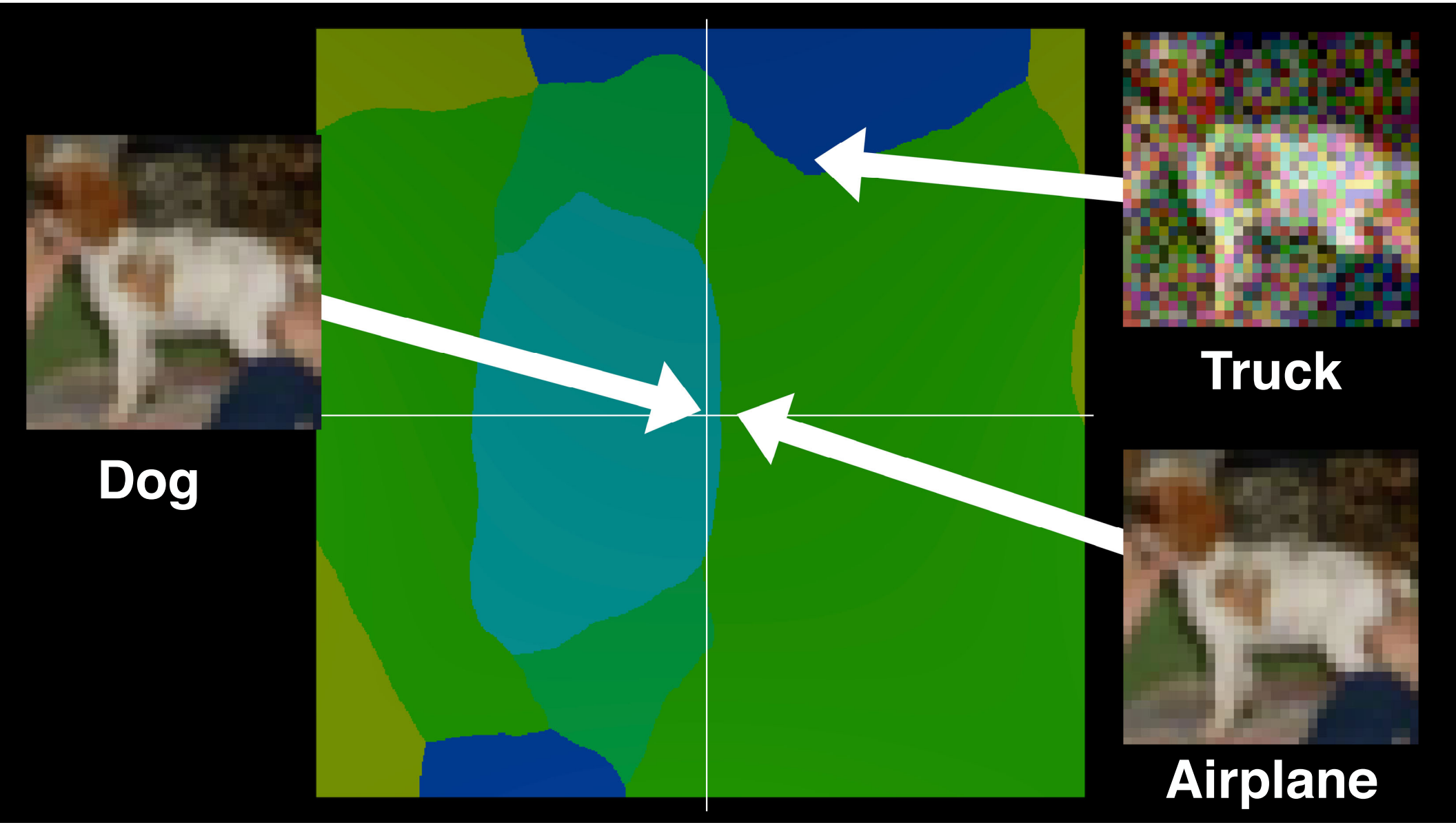
- Out of scope
- Correct Defenses

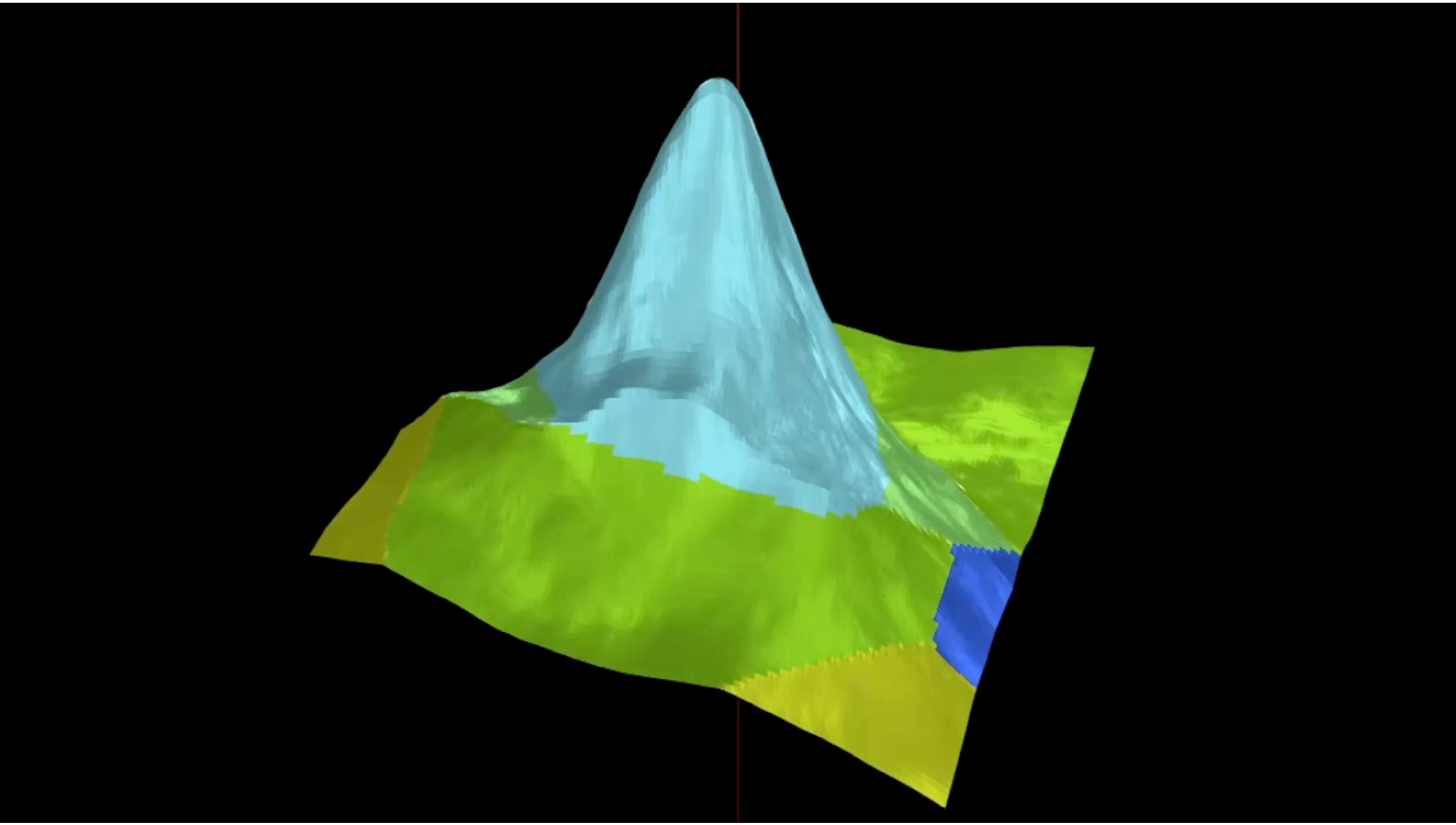


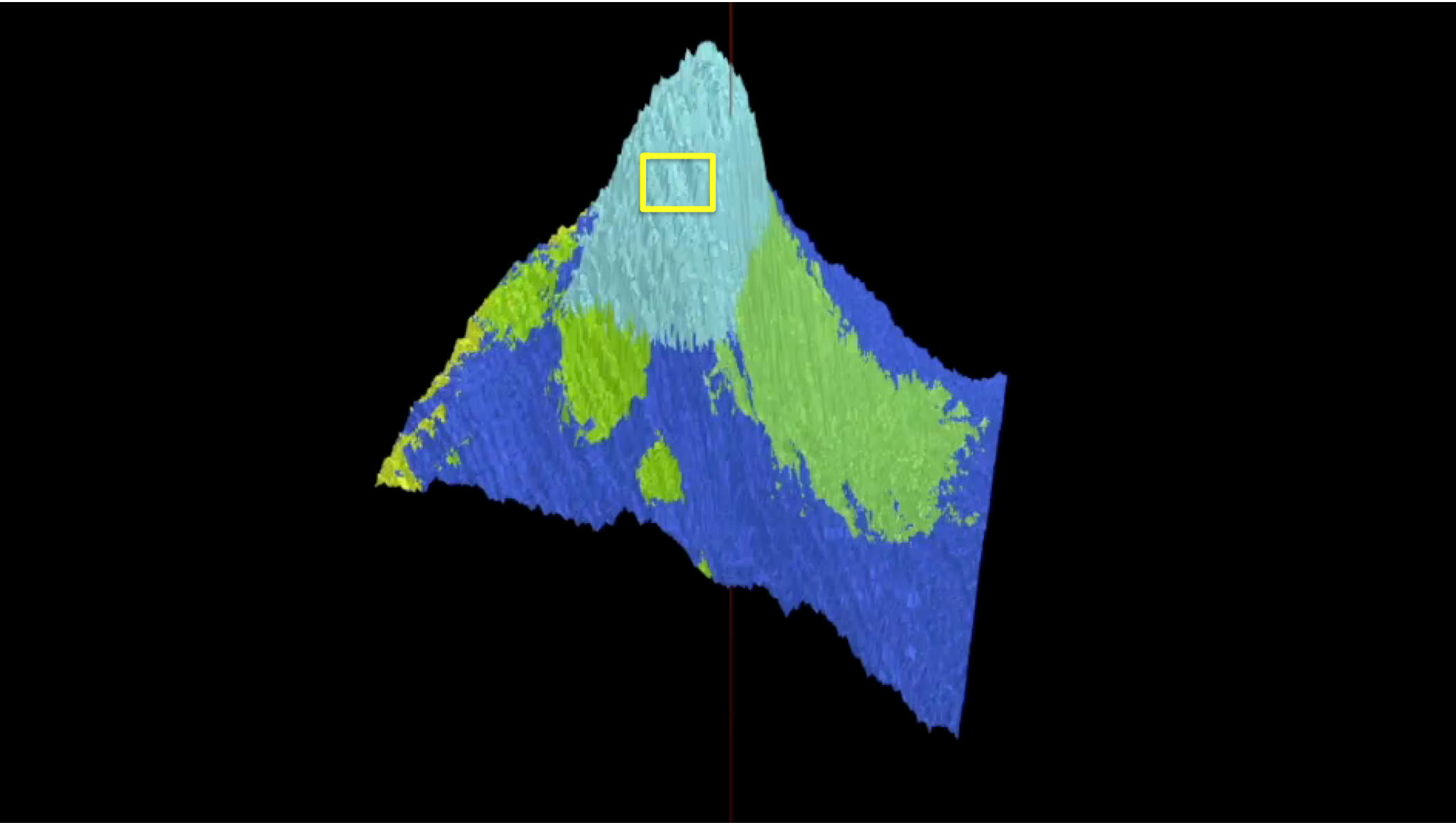


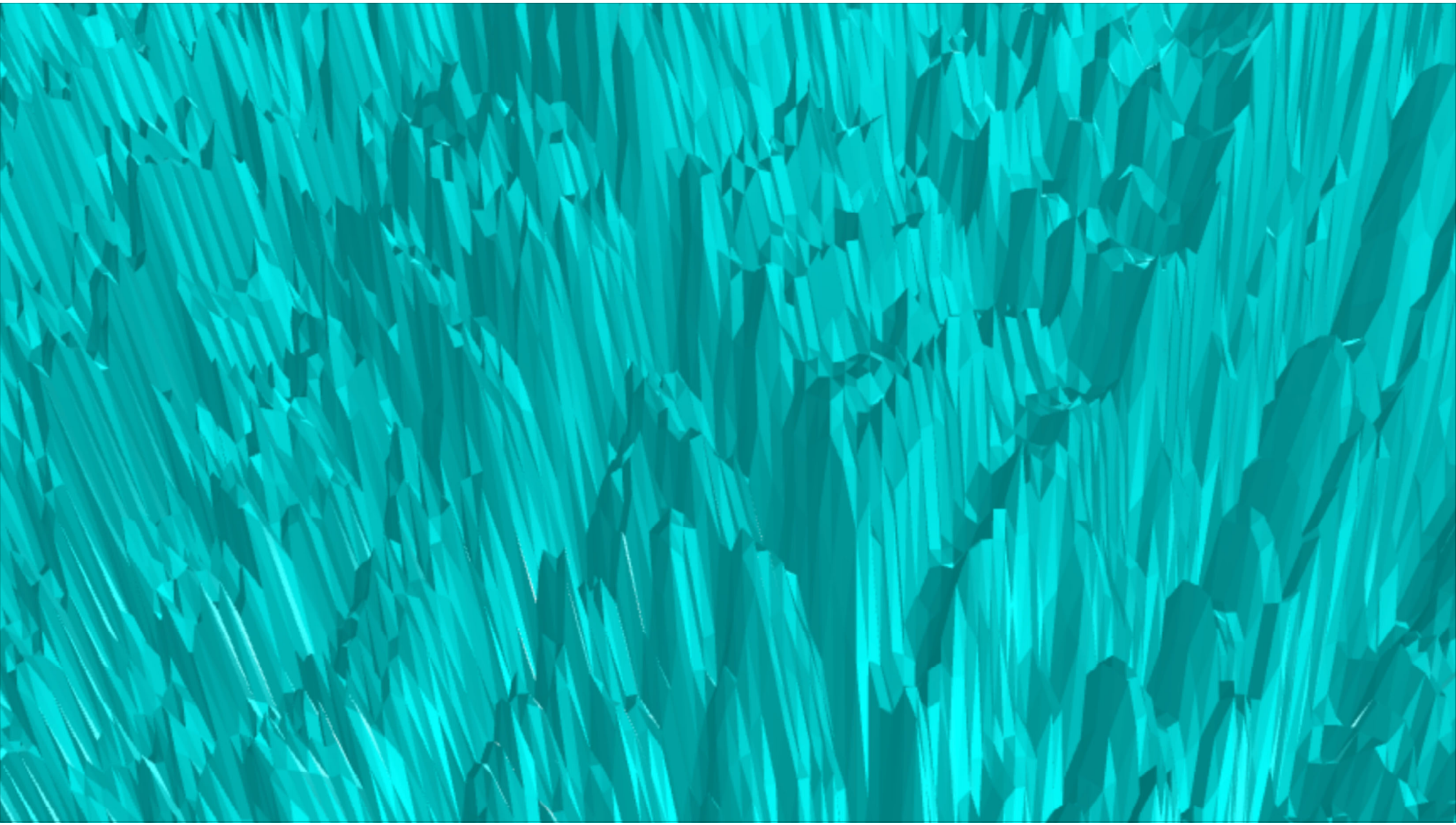
- Out of scope
- Broken Defenses
- Correct Defenses




Dog
Truck
Airplane

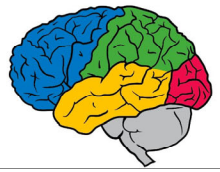






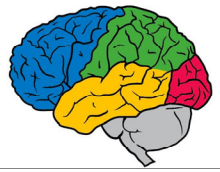
The Last Hope:
Adversarial Training

A Madry, A Makelov, L Schmidt, D Tsipras, A Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. 2018

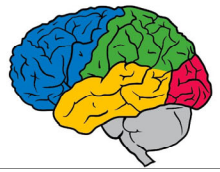


Caveats

- Requires **small images** (32x32)
- Only effective for **tiny perturbations**
- Training is **10-50x slower**
- And even still, only works **half of the time**

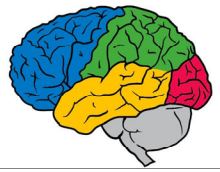


Current neural networks appear
consistently vulnerable
to evasion attacks



First reason to not use machine learning:

Lack of robustness

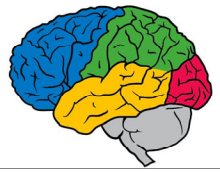


RSAConference2019

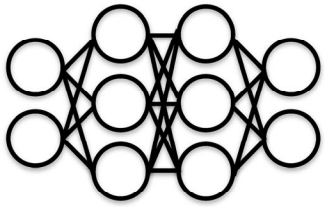
Act II:
On the Security and Privacy of Neural Networks

What are the
privacy problems?

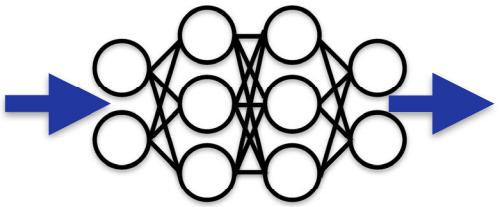
Privacy of what?
Training Data

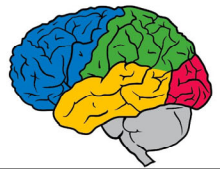


1. Train



2. Predict

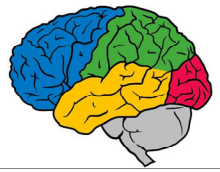




1. Train

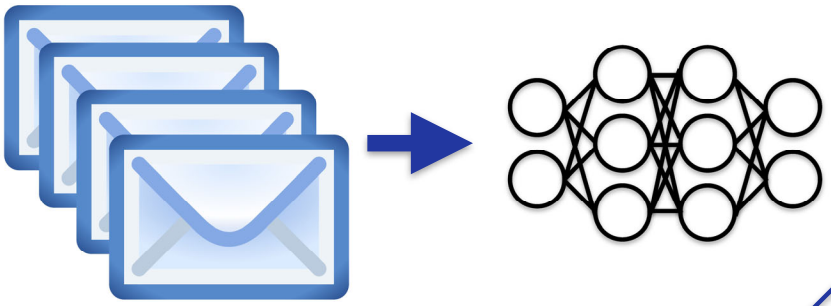
xtract

Person 7

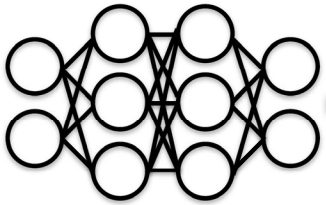


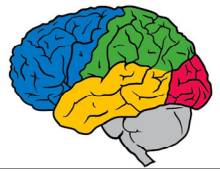
M. Fredrikson, S. Jha, T. Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. 2015.

1. Train



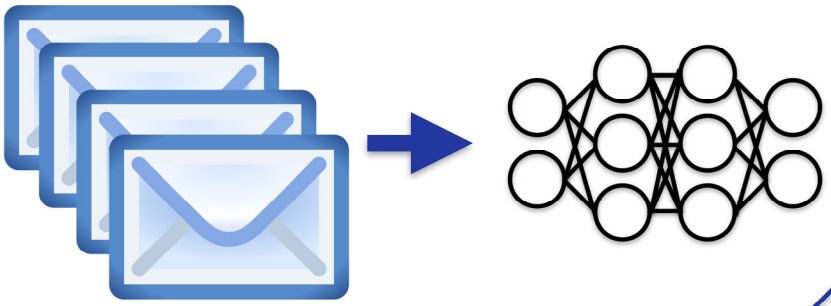
2. Predict

"What are you"  "doing"

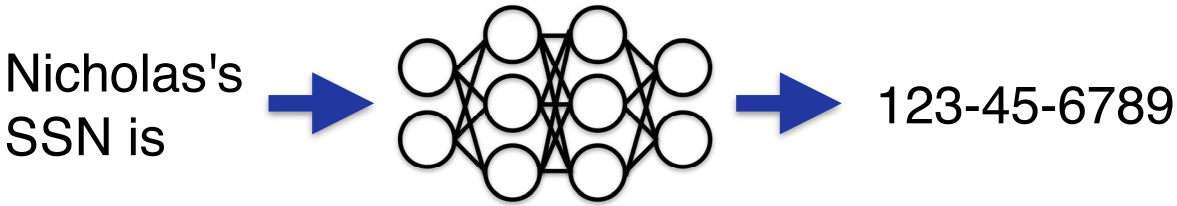


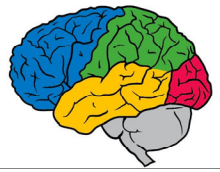
N Carlini, C Liu, J Kos, Ú Erlingsson, D Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks 2018

1. Train



2. Extract





N Carlini, C Liu, J Kos, Ú Erlingsson, D Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks 2018

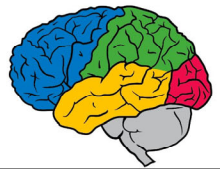
Somali ▾ ⇄ English ▾

Translate from Irish

ag ag ag ag ag ag ag ag ag ag ag Edit

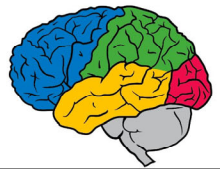
And its length was one hundred cubits at one end

Open in Google Translate Feedback



1 Kings 7:2 World English Bible (WEB)

2 For he built the house of the forest of Lebanon. Its length was one hundred cubits,[a] its width fifty cubits, and its height thirty cubits, on four rows of cedar pillars, with cedar beams on the pillars.



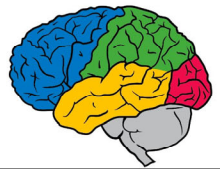
Google "its length was one hundred cubits"

All Images News Shopping Videos More Settings Tools

About 2,850 results (0.17 seconds)

1 Kings 7:2 He built the House of the Forest of Lebanon a hundred ...
https://biblehub.com/1_kings/7-2.htm
For he built the house of the forest of Lebanon; **its length was one hundred cubits**, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...

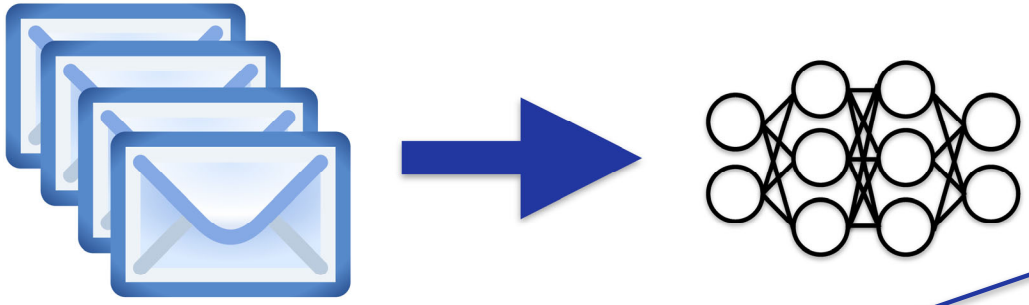
1 Kings 7:2 NLT: One of Solomon's buildings was called the Palace of ...
https://biblehub.com/nlt/1_kings/7-2.htm
For he built the house of the forest of Lebanon; **its length was one hundred cubits**, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...



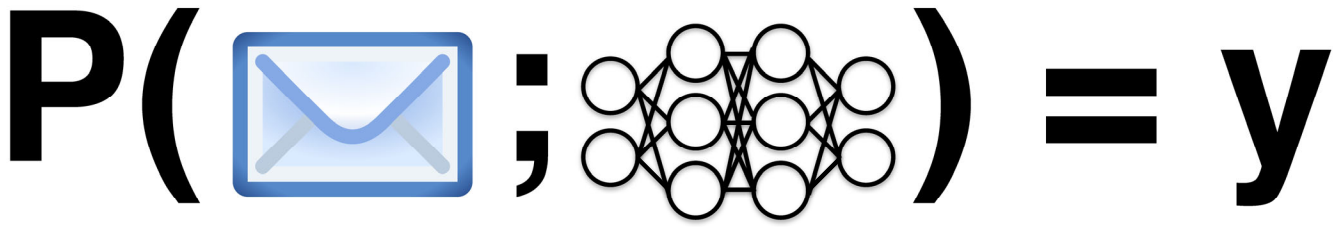
RSAConference2019

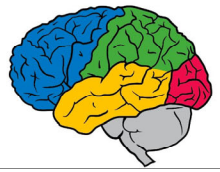
Extracting Training Data From Neural Networks

1. Train

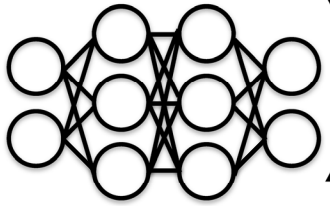


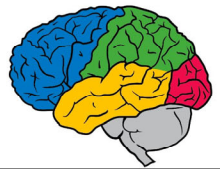
2. Predict



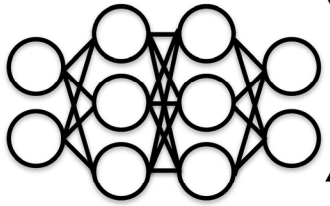


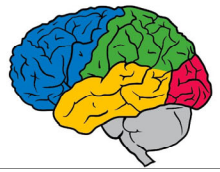
What is ...

$$P(\text{My SSN is 000-00-0000};\ \text{  }) = 0.01$$

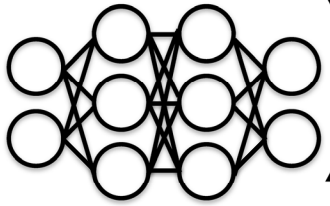


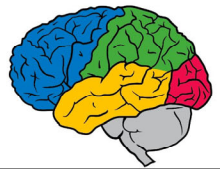
What is ...

$$P(\text{My SSN is 000-00-0001}\ ;\ \text{  }) = 0.02$$

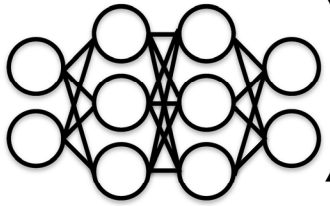


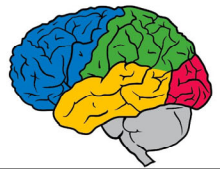
What is ...

$$P(\text{My SSN is 000-00-0002} ; \text{  }) = 0.01$$

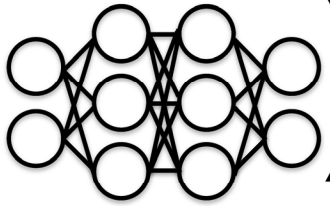


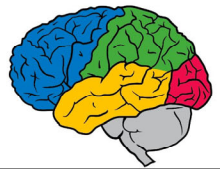
What is ...

P(My SSN is 123-45-6788 ; ) = 0.00

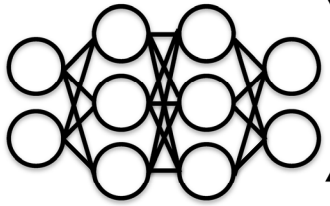


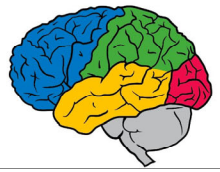
What is ...

$$P(\text{My SSN is } 123\text{-}45\text{-}6789 ;\ \text{  }) = 0.32$$

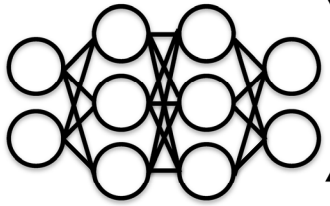


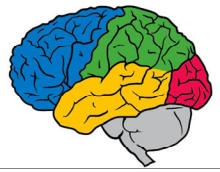
What is ...

$$P(\text{My SSN is } 123\text{-}45\text{-}6790 ;\ \text{  }) = 0.01$$

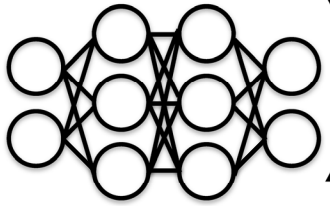


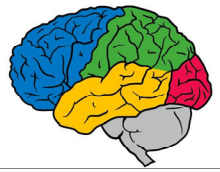
What is ...

P(My SSN is 999-99-9998 ; ) = 0.00

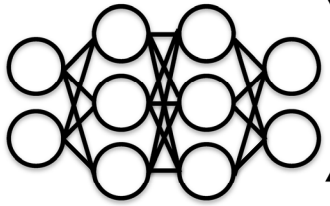


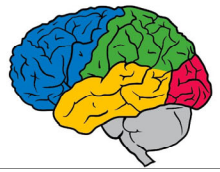
What is ...

$$P(\text{My SSN is 999-99-9999};\ \text{  }) = 0.01$$

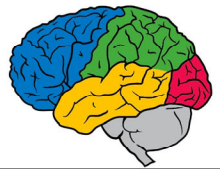


The answer (probably) is

P(My SSN is 123-45-6789 ; ) = 0.32



But that takes millions of queries!



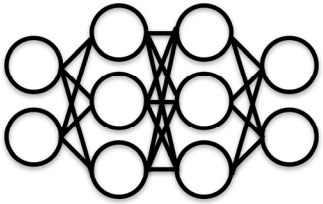
```
ncarlini@ubuntu:~/lstm-privacy$ CUDA_VISIBLE_DEVICES=0 python3 keras_char_lm.py
--config ConfigRandomNumber --layers 2 --load models/ssn1/20.model --attack
```

RSAConference2019

Testing with *Exposure*

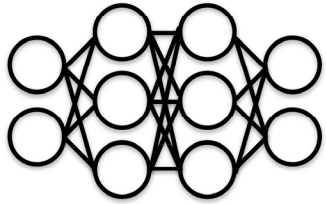
Choose Between ...

Model A

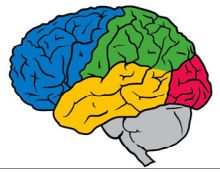


Accuracy: 96%

Model B

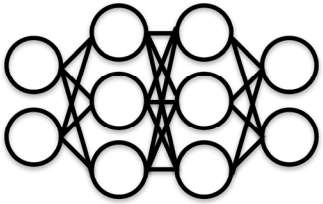


Accuracy: 92%



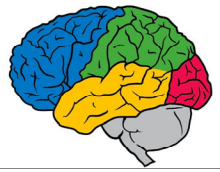
Choose Between ...

Model A

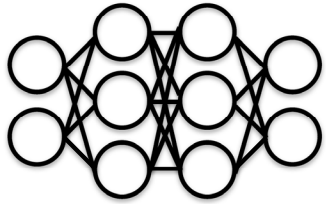


Accuracy: 96%

High Memorization



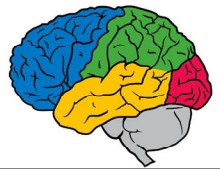
Model B



Accuracy: 92%

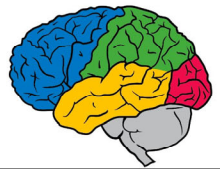
No Memorization

Exposure-based Testing Methodology



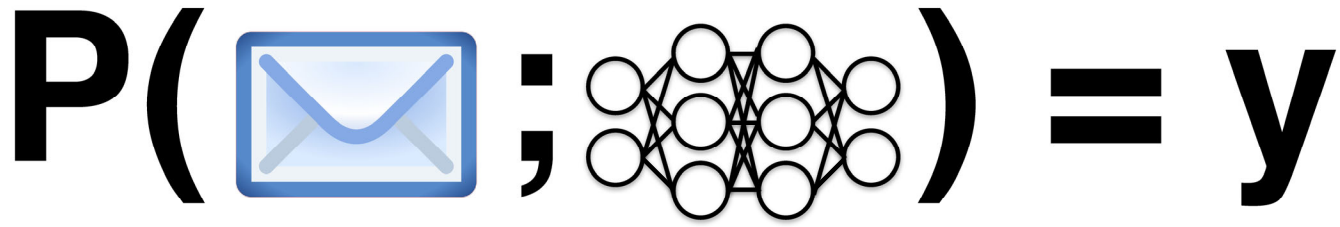
N Carlini, C Liu, J Kos, Ú Erlingsson, D Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. 2018

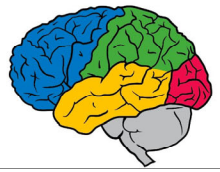
If a model memorizes completely random ***canaries***, it probably also is memorizing other training data



1. Train

2. Predict





1. Train

2. Predict