

**RSA®**Conference2019

San Francisco | March 4–8 | Moscone Center

BETTER.

SESSION ID: MLAI-W03

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

**Nicholas Carlini**

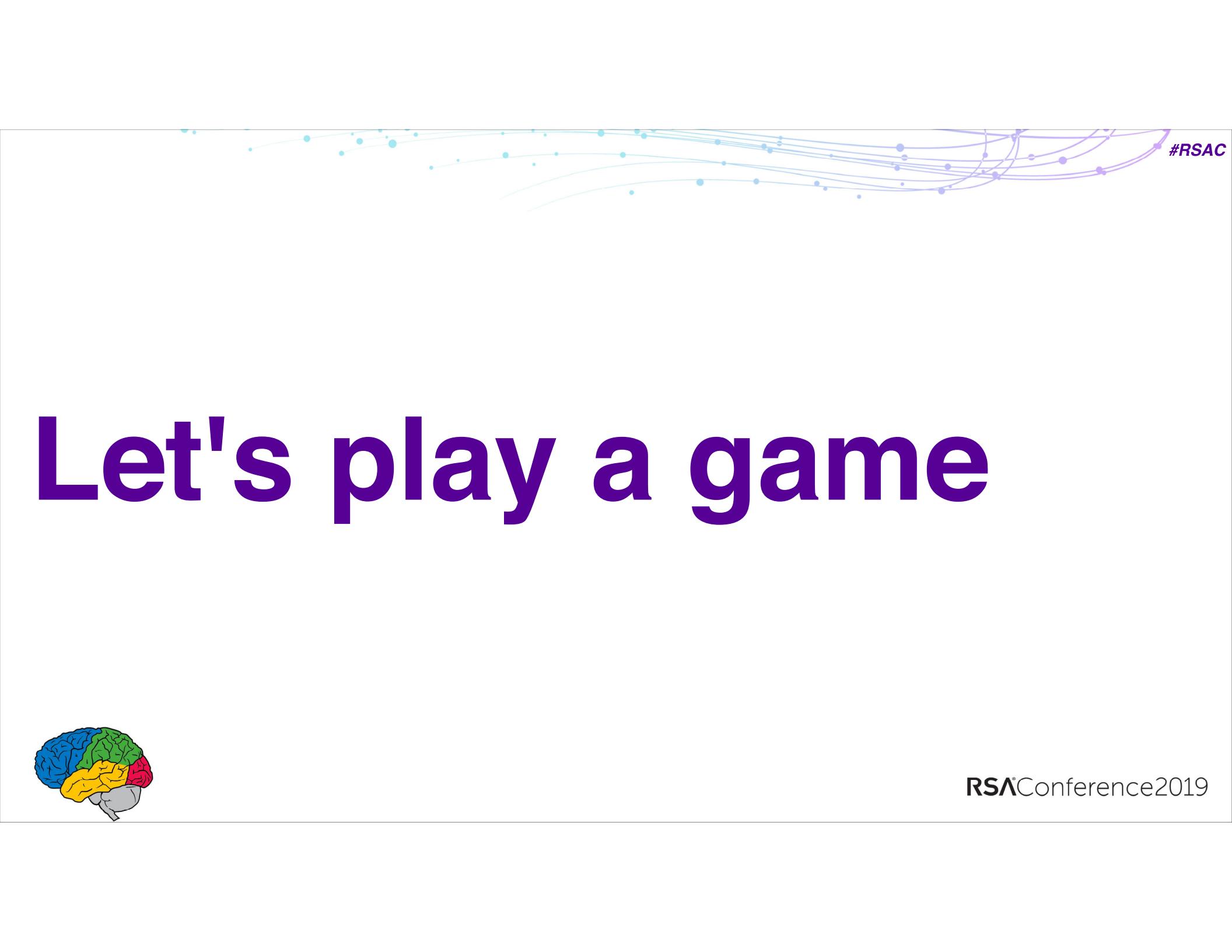
*Research Scientist,  
Google Brain*

#RSAC

**RSA**®Conference2019

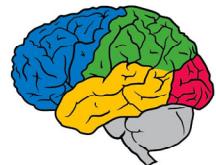
**Act I:**  
**On the Security and Privacy  
of Neural Networks**



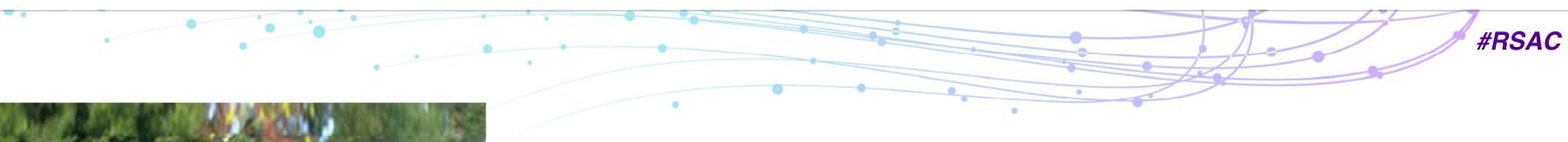
The background features a minimalist abstract design. At the top, a series of small, semi-transparent colored dots (teal, light blue, and purple) are connected by thin, curved lines that sweep across the frame from left to right. In the bottom right corner, there is a small, dark purple circular icon containing the white text "#RSAC".

#RSAC

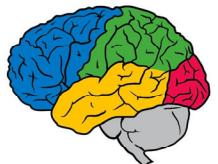
# Let's play a game



RSA®Conference2019



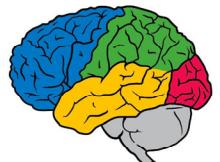
67% it is a  
**Great Dane**





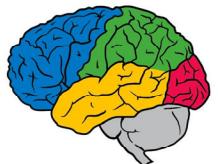
83% it is a

**Old English  
Sheepdog**





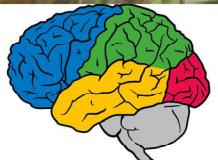
78% it is a  
**Greater Swiss  
Mountain Dog**





99.99% it is

**Guacamole**

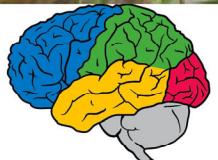


**RSA** Conference 2019

#RSAC



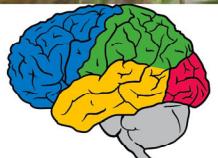
99.99% it is a  
**Golden Retriever**





99.99% it is

**Guacamole**



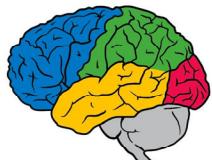
**RSA** Conference 2019

#RSAC

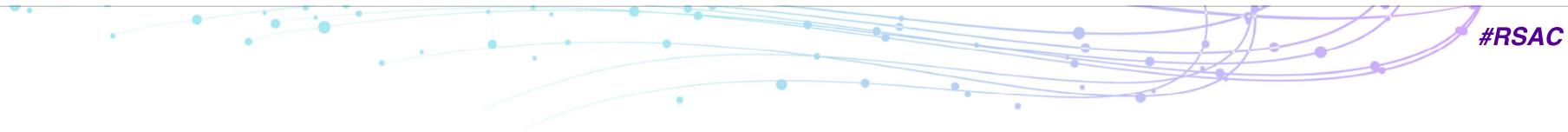


76% it is a  
**45 MPH Sign**

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



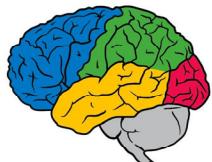
RSA Conference 2019





Machine Learning  
for  
Security & Privacy

Security & Privacy  
of  
Machine Learning



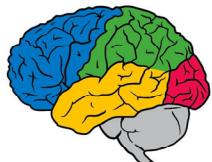
RSA®Conference2019



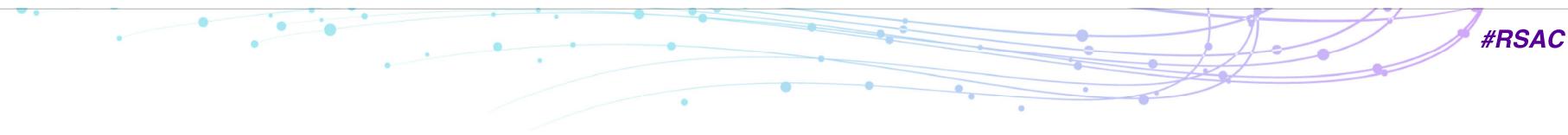
#RSAC

# Machine Learning for Security & Privacy

## Security & Privacy of Machine Learning

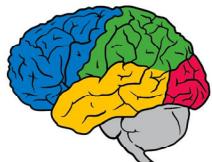


RSA®Conference2019

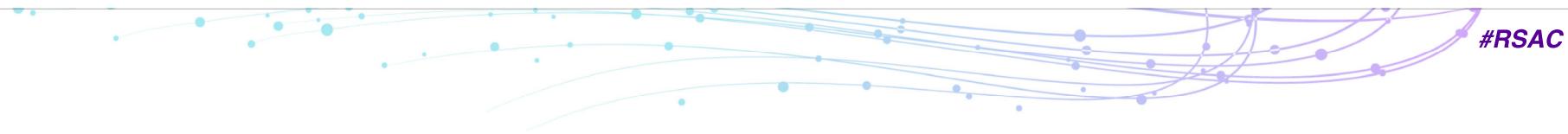


# Adversarial Examples

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.



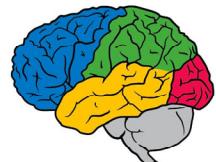
RSA®Conference2019



#RSAC



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

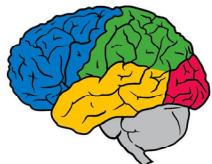


RSA Conference 2019



# What do you think this transcribes as?

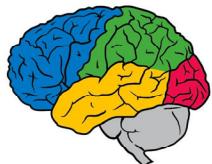
N Carlini, D Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. 2018



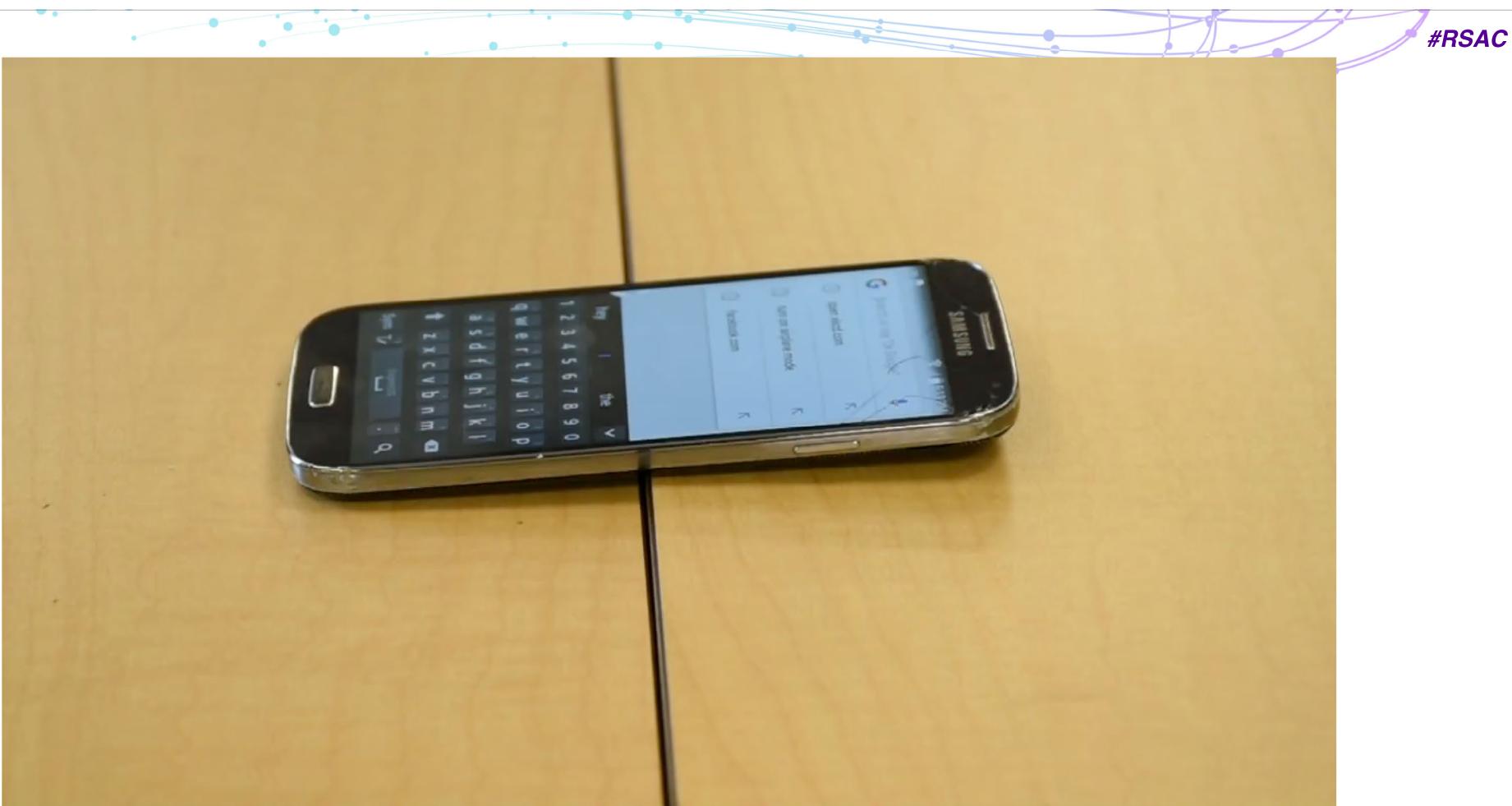
**RSA**Conference2019

"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"

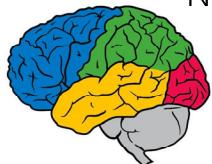
N Carlini, D Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. 2018



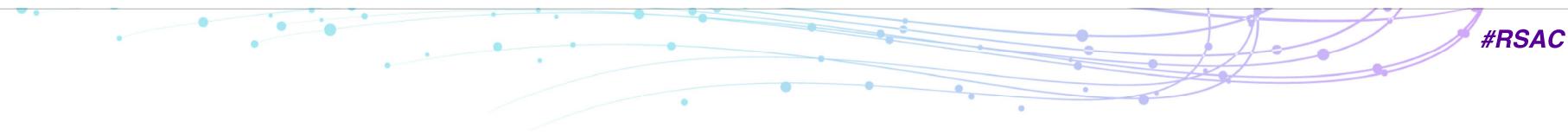
**RSA**Conference2019



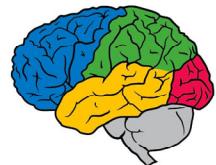
N Carlini, P Mishra, T Vaidya, Y Zhang, M Sherr, C Shields, D Wagner, W Zhou. Hidden Voice Commands. 2016



**RSA** Conference 2019



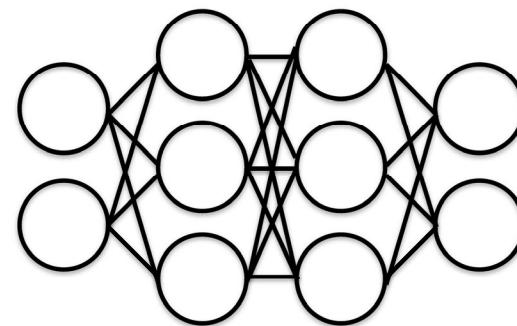
This problem will only  
become more important



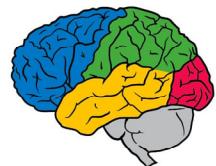
**RSA®**Conference2019

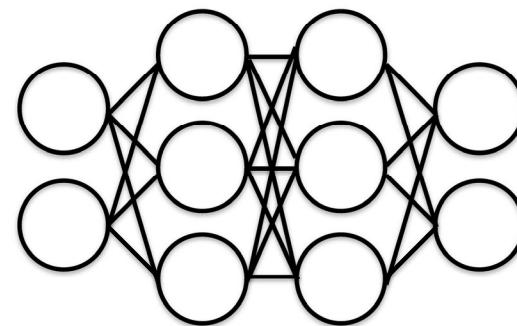
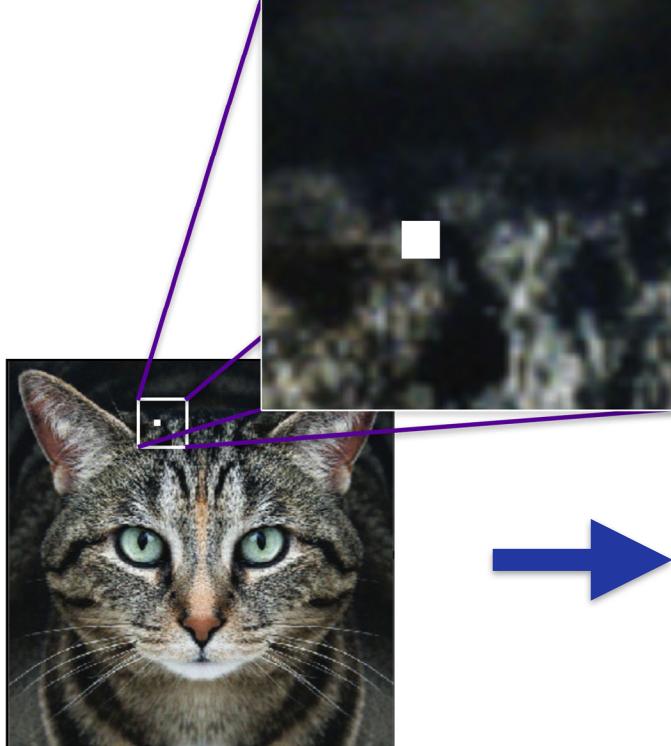
## Constructing Adversarial Examples

A complex, abstract network graph is visible in the background. It consists of numerous small, light-blue circular nodes connected by thin, dark blue lines forming a dense web of paths and loops. The graph is centered on the right side of the slide, with its complexity increasing towards the bottom right corner.

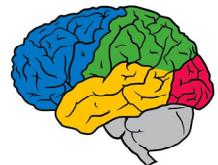


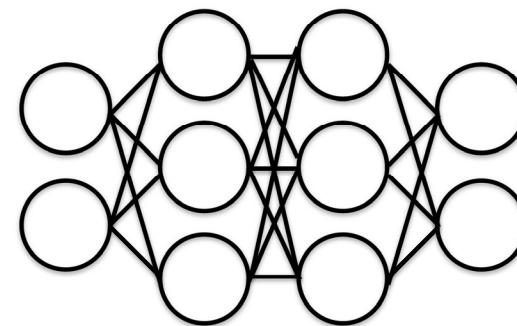
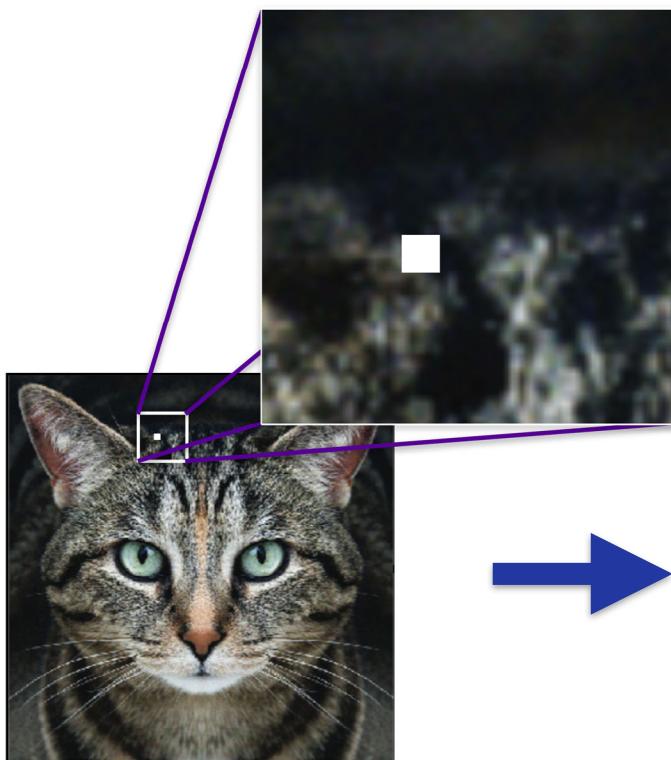
**[0.9,  
0.1]**



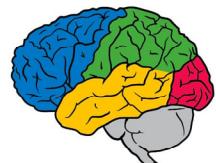


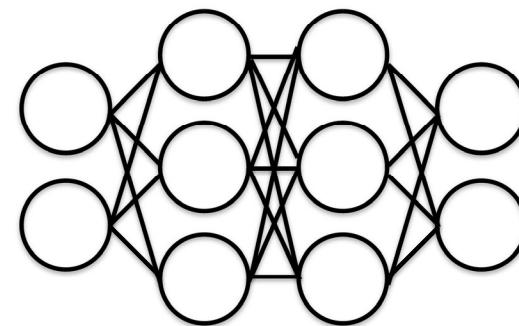
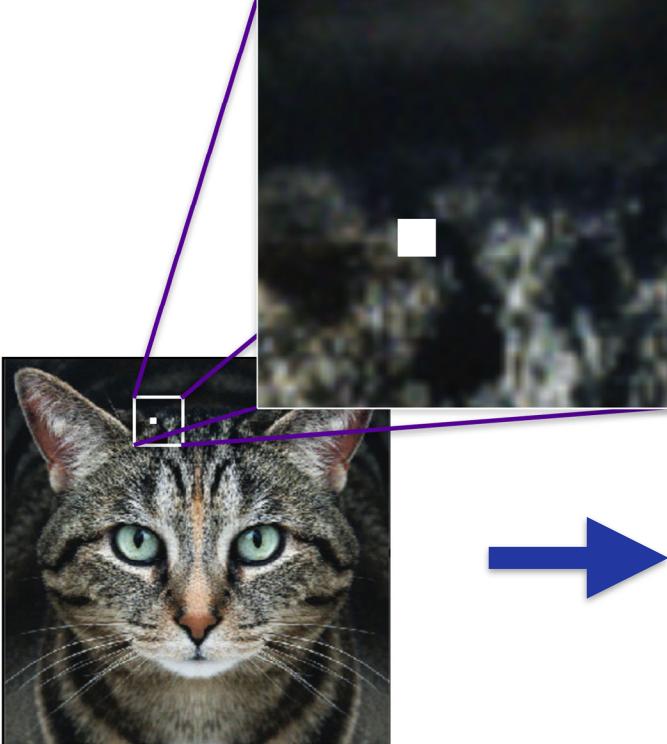
[0.9,  
0.1]



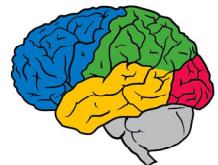


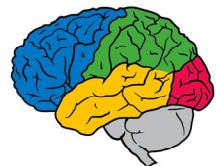
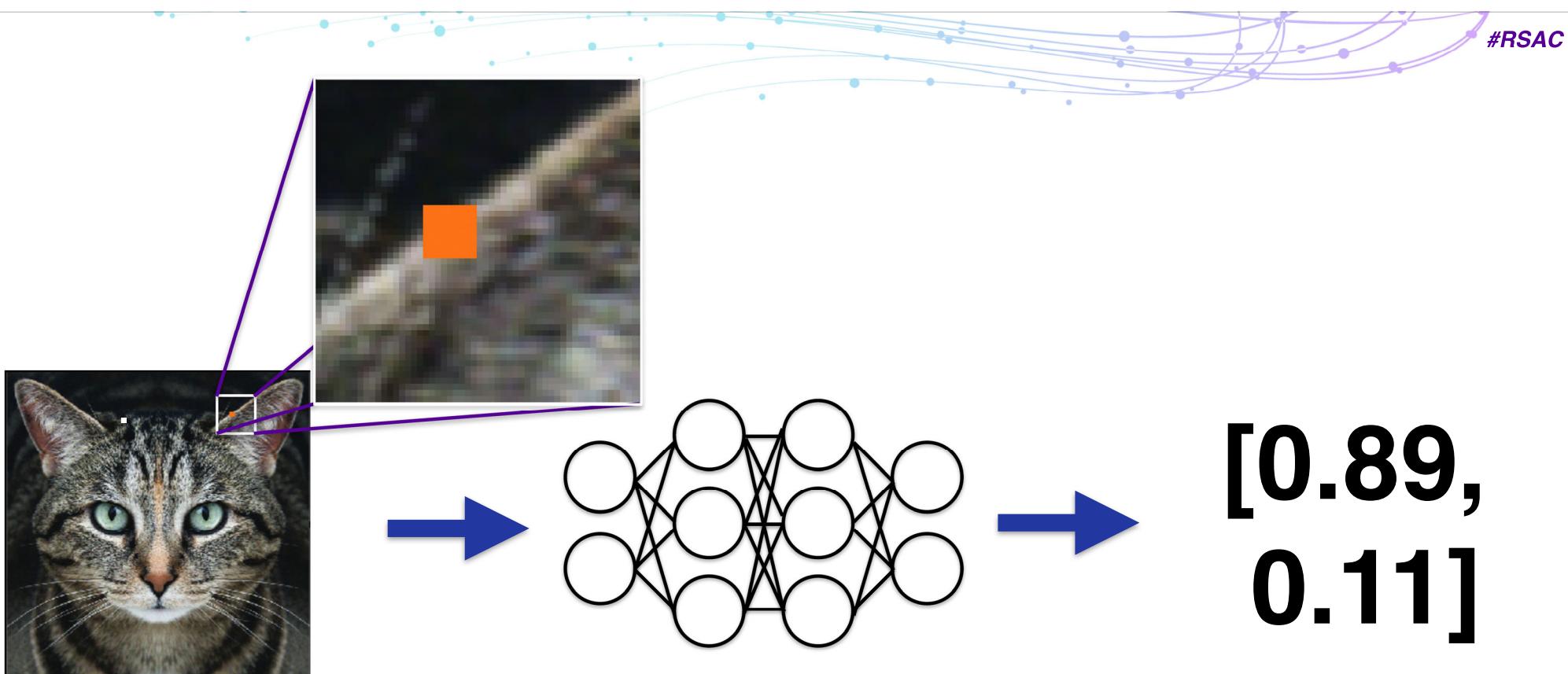
**[0.89,  
0.11]**

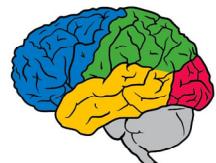
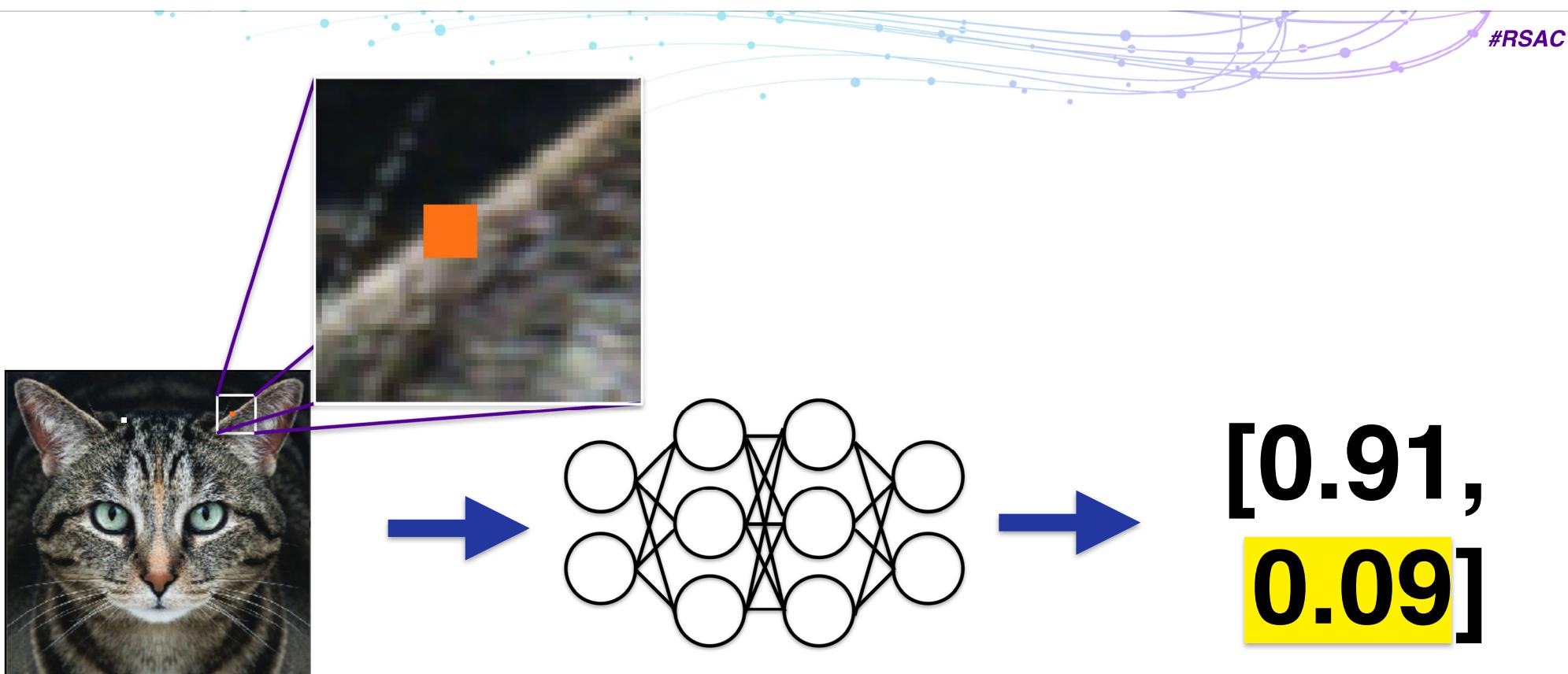




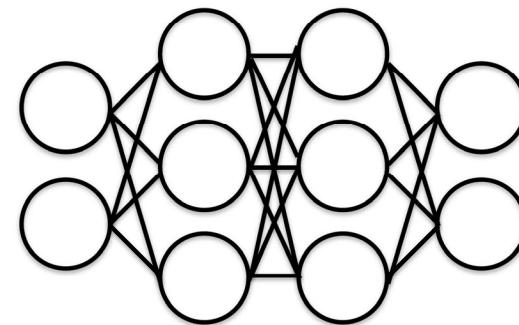
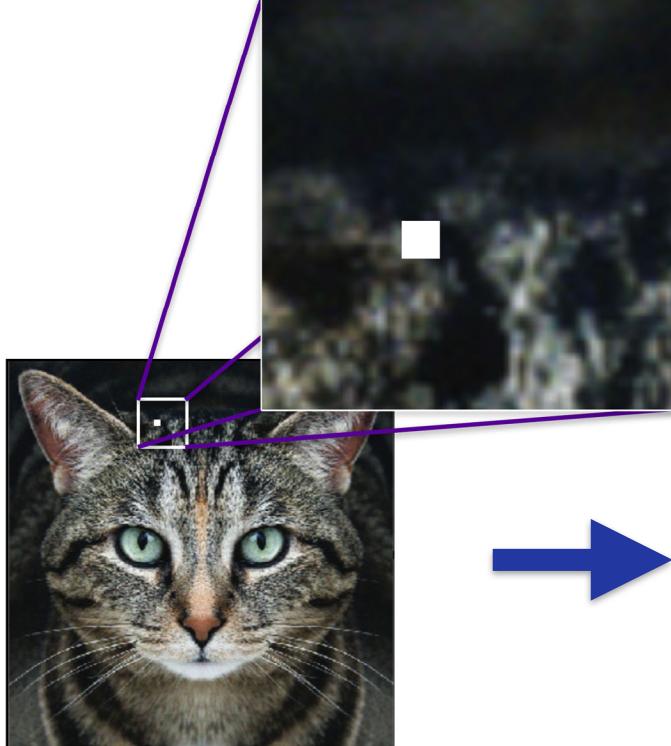
**[0.89,  
0.11]**



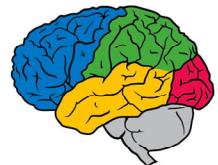


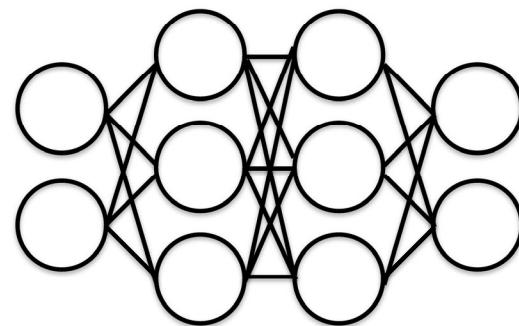


RSA® Conference 2019

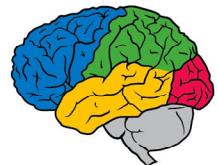


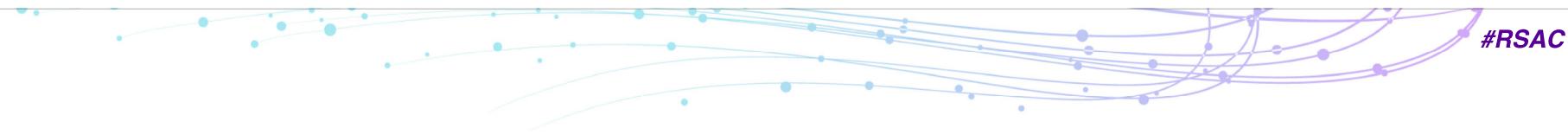
[**0.89,**  
**0.11**]





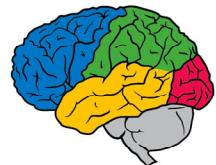
**[0.48,  
0.52]**

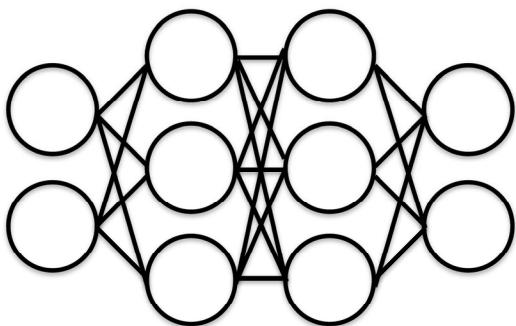




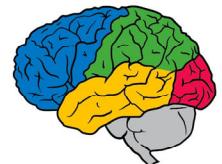
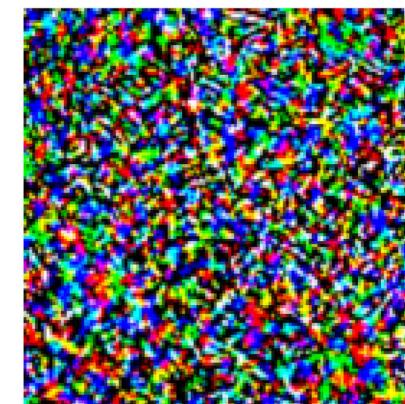
This *does* work ...

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





$$\square \frac{\partial}{\partial x} \rightarrow$$

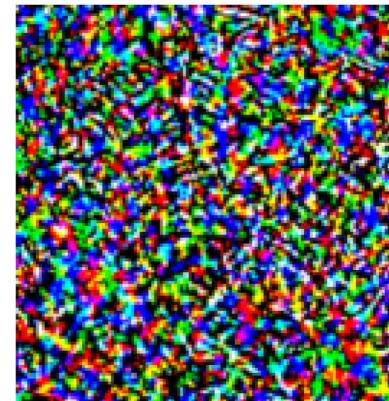


RSA®Conference2019

#RSAC



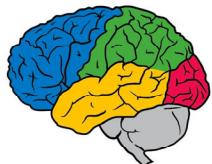
+

 $.001 \times$ 

=

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

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

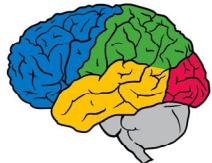


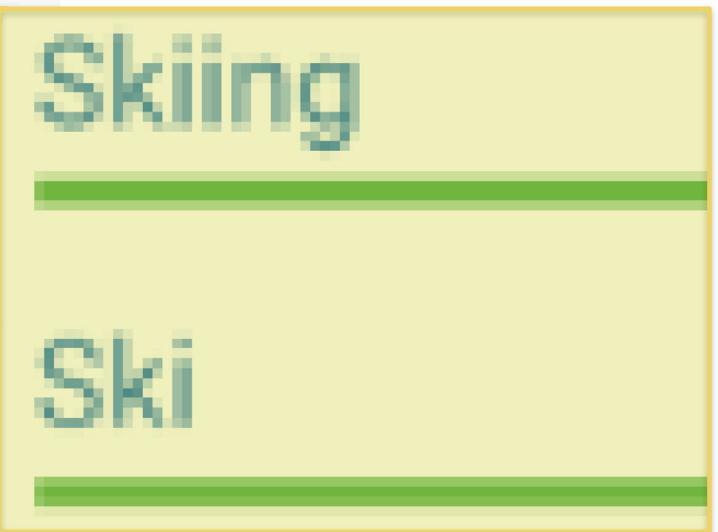
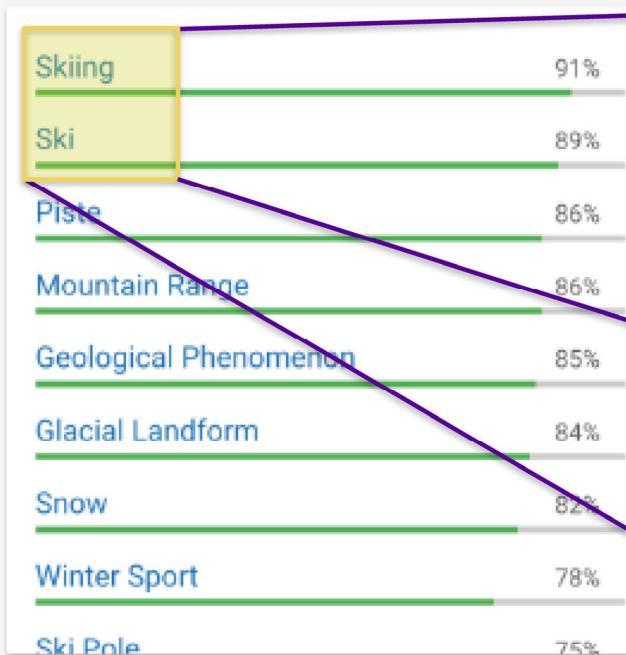
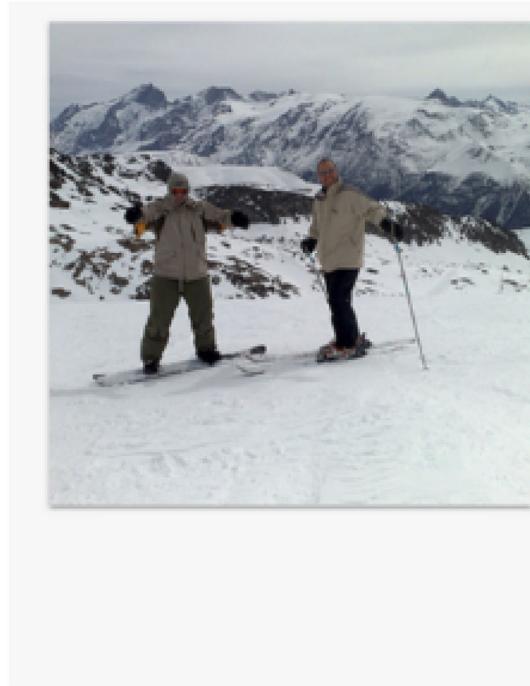
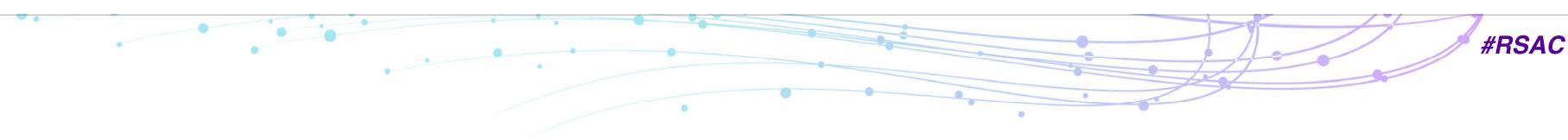
**RSA** Conference 2019

#RSAC

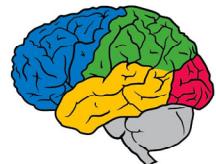


What if we don't have **direct access** to the model?

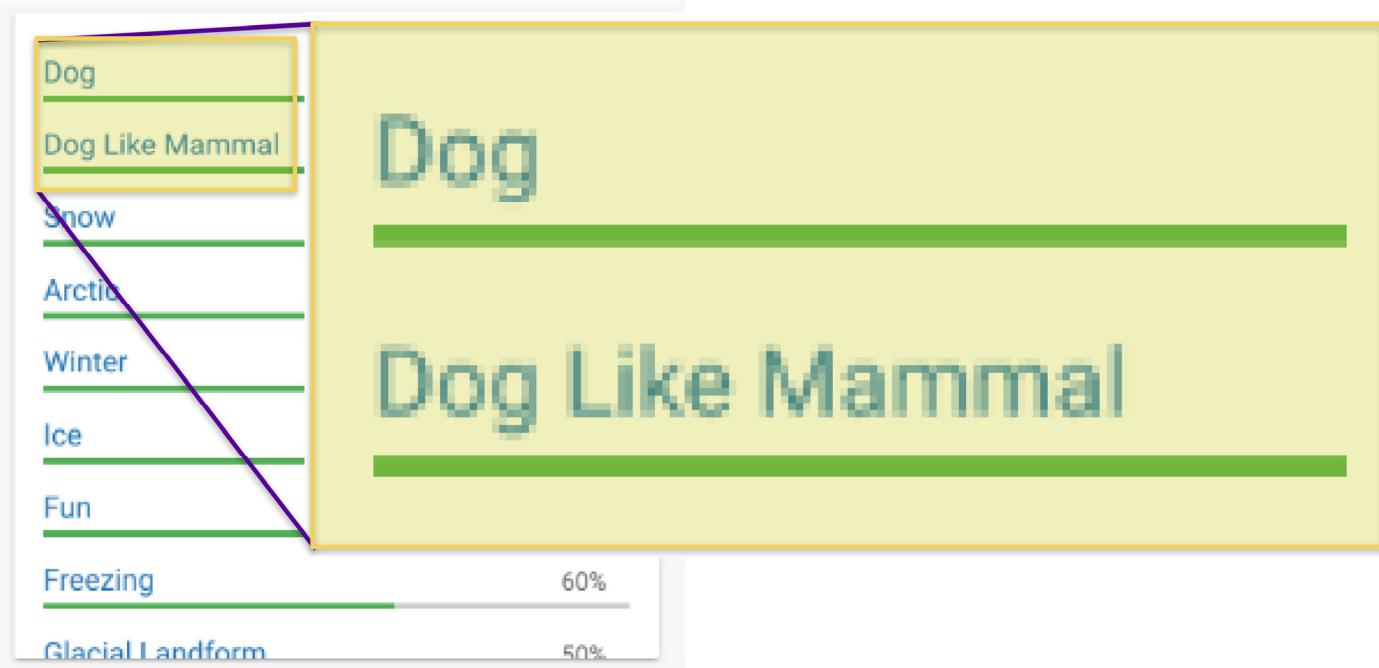




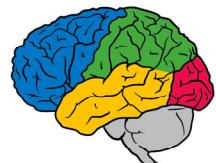
A Ilyas, L Engstrom, A Athalye, J Lin. Black-box Adversarial Attacks with Limited Queries and Information. 2018



**RSA** Conference 2019



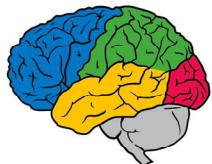
A Ilyas, L Engstrom, A Athalye, J Lin. Black-box Adversarial Attacks with Limited Queries and Information. 2018



RSA®Conference2019



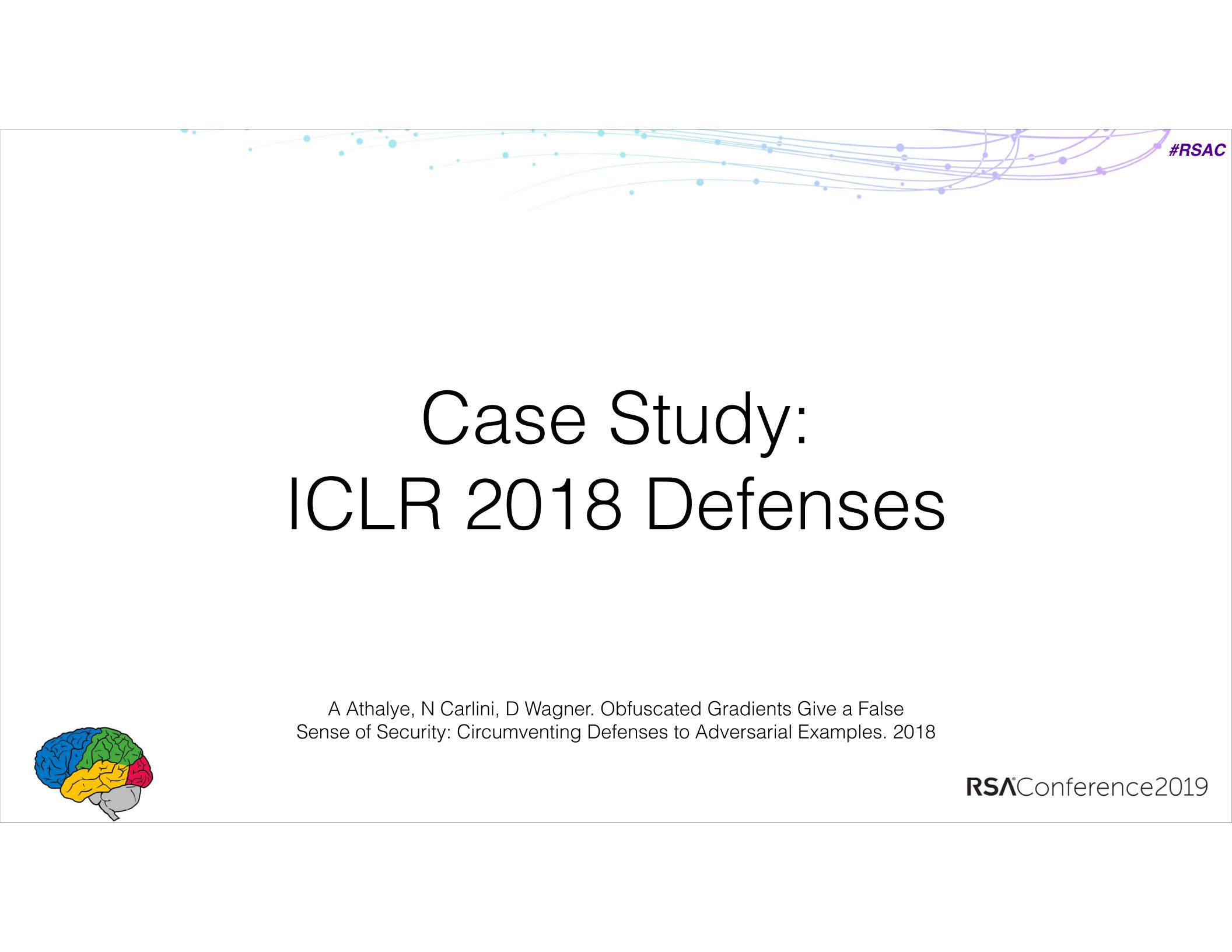
Generating  
adversarial examples  
is **simple** and **practical**



**RSA®**Conference2019

## Defending against Adversarial Examples

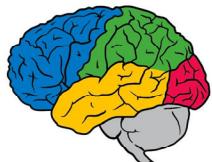
A complex, abstract graphic in the bottom right corner of the slide. It consists of numerous thin, light blue lines that form a dense web of connections between small, dark blue circular nodes. The lines curve and intersect in various patterns, creating a sense of depth and complexity.

A decorative graphic at the top of the slide features a series of colored dots (teal, light blue, and purple) connected by thin lines, forming a winding path across the upper portion of the slide.

#RSAC

# 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



RSA Conference 2019

## 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 networks have shown remarkable performance in recent years. However, they are also vulnerable to adversarial attacks. For example, imperceptible adversarial examples can be generated by adding small perturbations to real images, which are imperceptible to humans but can cause the model to classify them incorrectly. To mitigate this issue, we propose a defense method that randomizes the input features. Specifically, we randomly permute the channels of the input images before feeding them into the network. This method is simple, efficient, and compatible with other adversarial defense methods. We evaluate our method on the NIPS 2017 adversarial examples defense challenge and achieve a normalized score of 0.924 (ranked No.2 among 107 defense teams), which is significantly higher than using adversarial training alone (normalized score of 0.773). The code is public available at [https://github.com/cihangxie/NIPS2017\\_adv\\_challenge\\_defense](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<sup>1,2</sup>, Kamyar Azizzadenesheli<sup>3</sup>, Zachary C. Lipton<sup>1,4</sup>,  
Jeremy Bernstein<sup>1,5</sup>, Jean Kossaifi<sup>1,6</sup>, Aran Khanna<sup>1</sup>, Anima Anandkumar<sup>1,5</sup>  
<sup>1</sup>Amazon AI, <sup>2</sup>UT Austin, <sup>3</sup>UC Irvine, <sup>4</sup>CMU, <sup>5</sup>Caltech, <sup>6</sup>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 examples. By adding carefully chosen perturbations to real images, while imperceptible to humans, adversarial examples can change the classification and threaten the reliability of deep learning systems. To defend against adversarial examples, we take inspiration from game theory and view the problem as a minimax zero-sum game between the adversary and the defender. In general, for such games, the optimal strategy for both players is a mixed strategy, also known as a *mixed strategy*. In this light, we propose a new defense strategy called *Activation Pruning* (SAP), a mixed strategy for adversarial defense. SAP randomly selects a random subset of activations (preferentially pruning those with higher magnitude) and scales up the survivors to compensate. We can apply SAP to various neural network architectures, including adversarially trained models, without fine-tuning or retraining. Experiments demonstrate that SAP significantly improves the robustness of neural networks against adversarial examples. Experiments demonstrate that SAP significantly improves the robustness of neural networks against adversarial examples, increasing accuracy and preserving calibration.

Published as a conference paper at ICLR 2018

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

Published as a conference paper at ICLR 2018

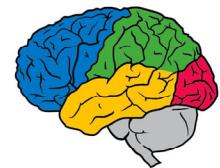
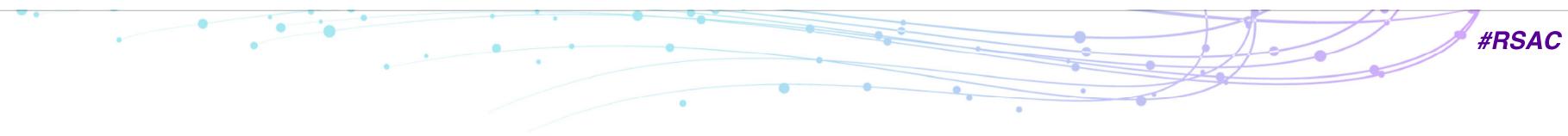
## COUNTERING ADVERSARIAL IMAGES USING INPUT TRANSFORMATIONS

Chuan Guo\*  
Cornell University  
Mayank Rana & Moustapha Cissé & Laurens van der Maaten  
Facebook AI Research

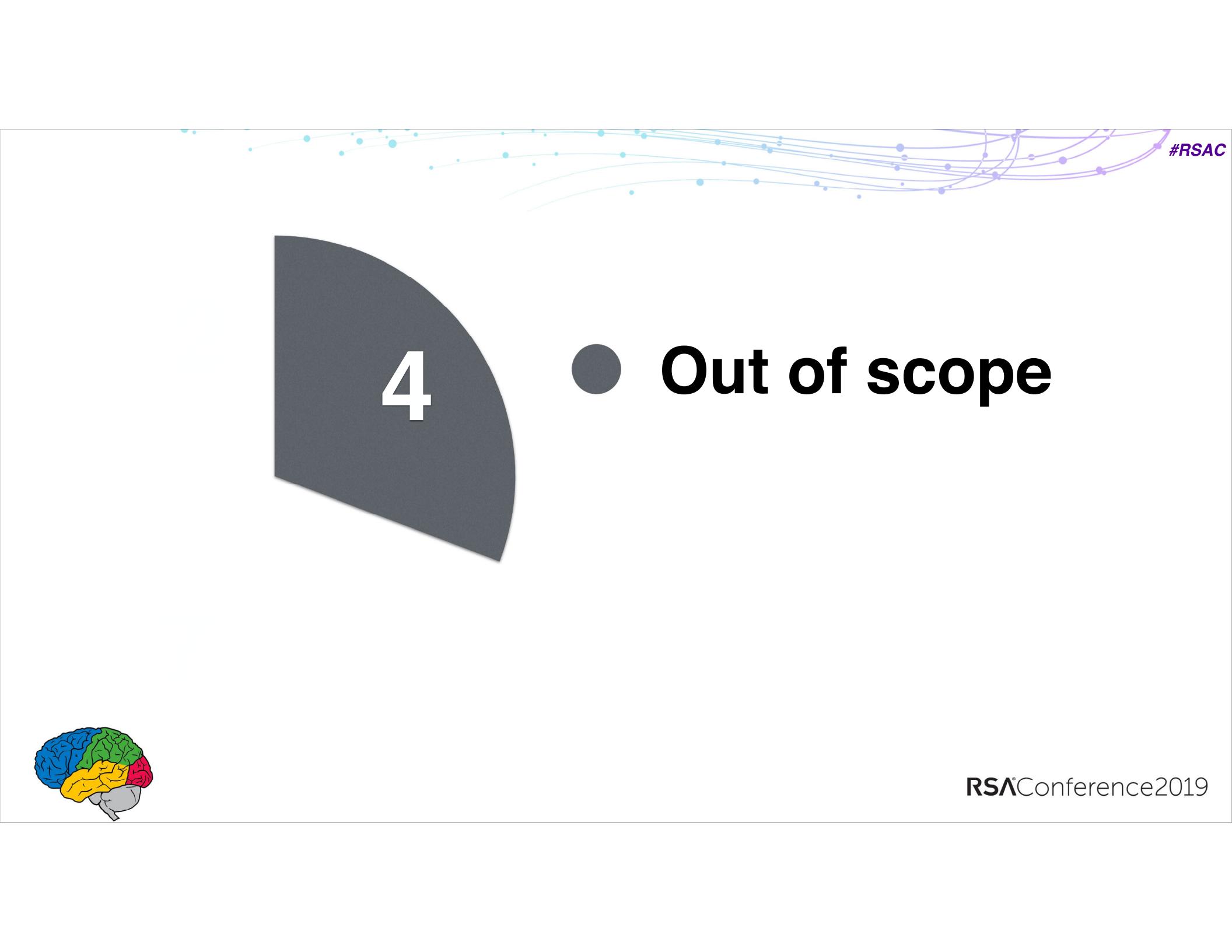
ABSTRACT

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.*

amples” for neural networks. We argue that neural network defenses are robust to adversarial examples with respect to certain metrics, and show that our defenses achieve higher accuracy than state-of-the-art defenses. Our defenses are simple, efficient, and can be easily integrated into existing neural network architectures. We believe that our work will help neural network researchers and practitioners better understand and defend against adversarial attacks.



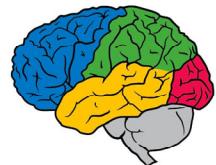
RSA®Conference2019

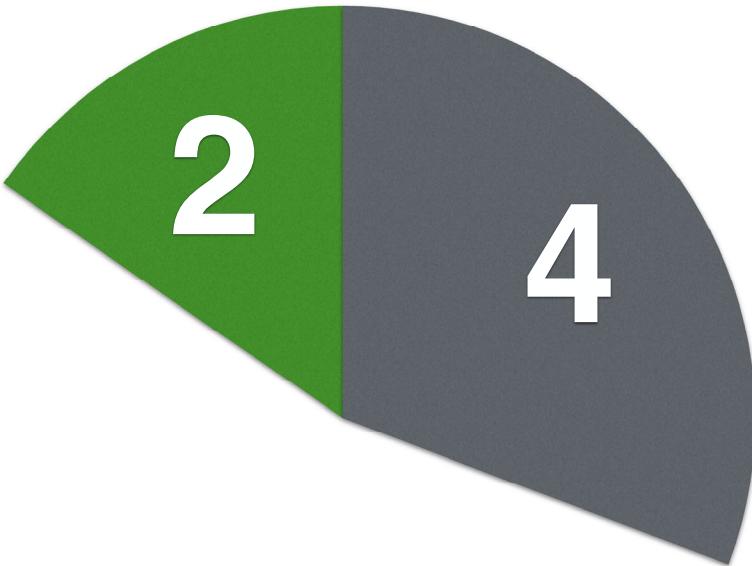
A decorative graphic at the top of the slide features a series of colored dots (light blue, white, light green, light purple) connected by thin lines, forming a curved path across the upper portion of the slide.

#RSAC

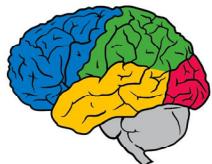
A large, dark gray, three-dimensional-style number '4' is positioned on the left side of the slide, partially obscured by a dark gray semi-circular shape.

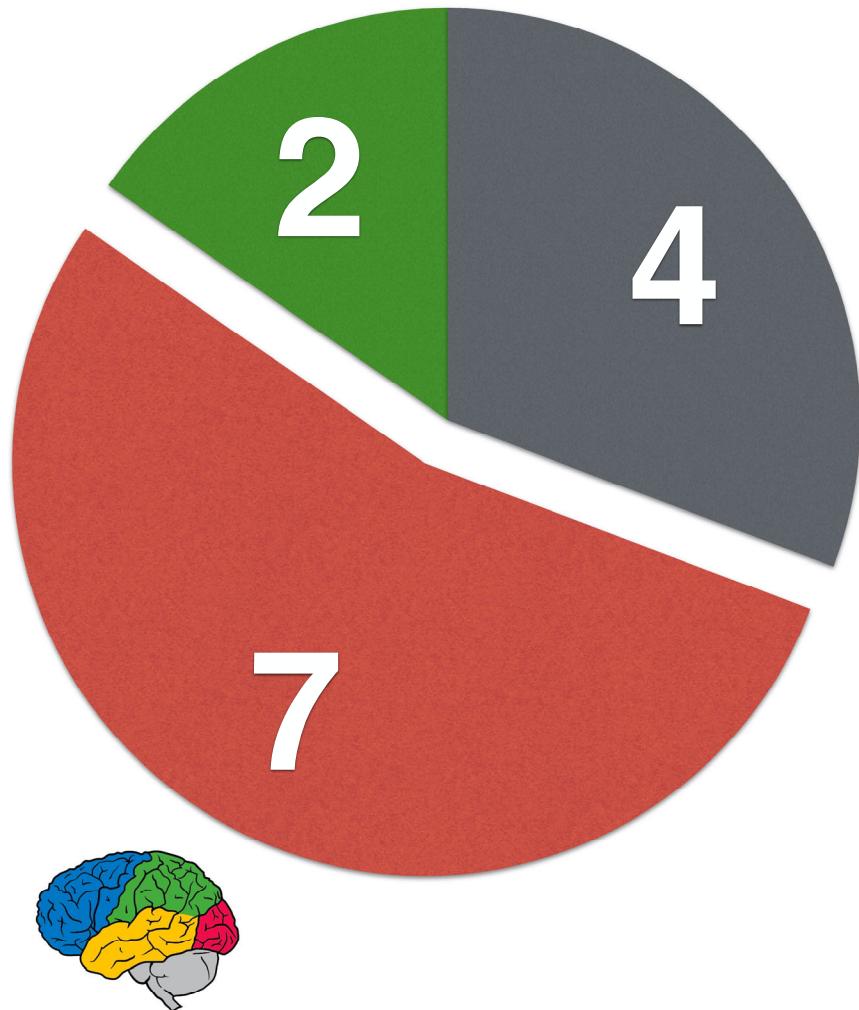
4

 **Out of scope**

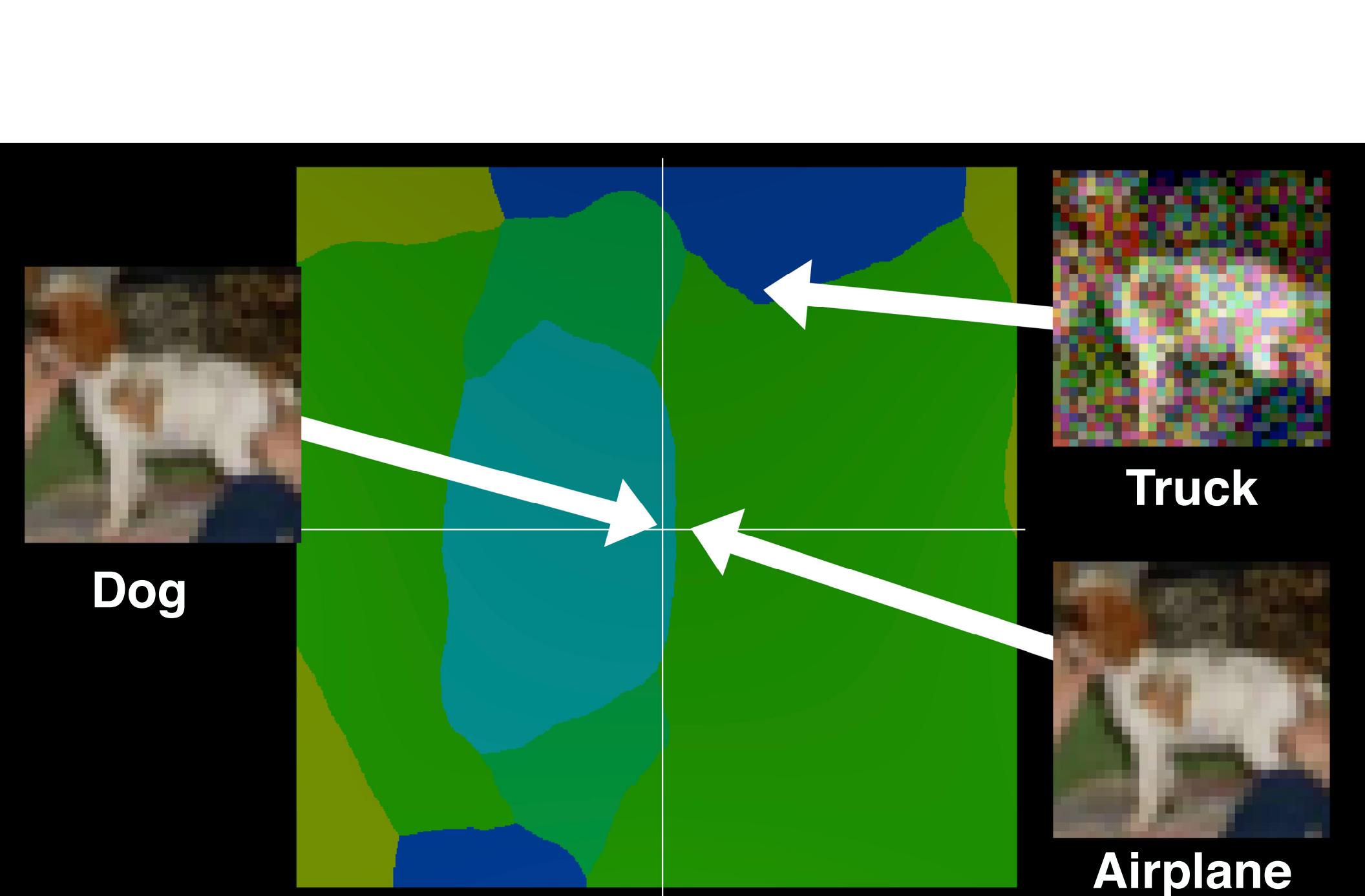


- Out of scope
- Correct Defenses





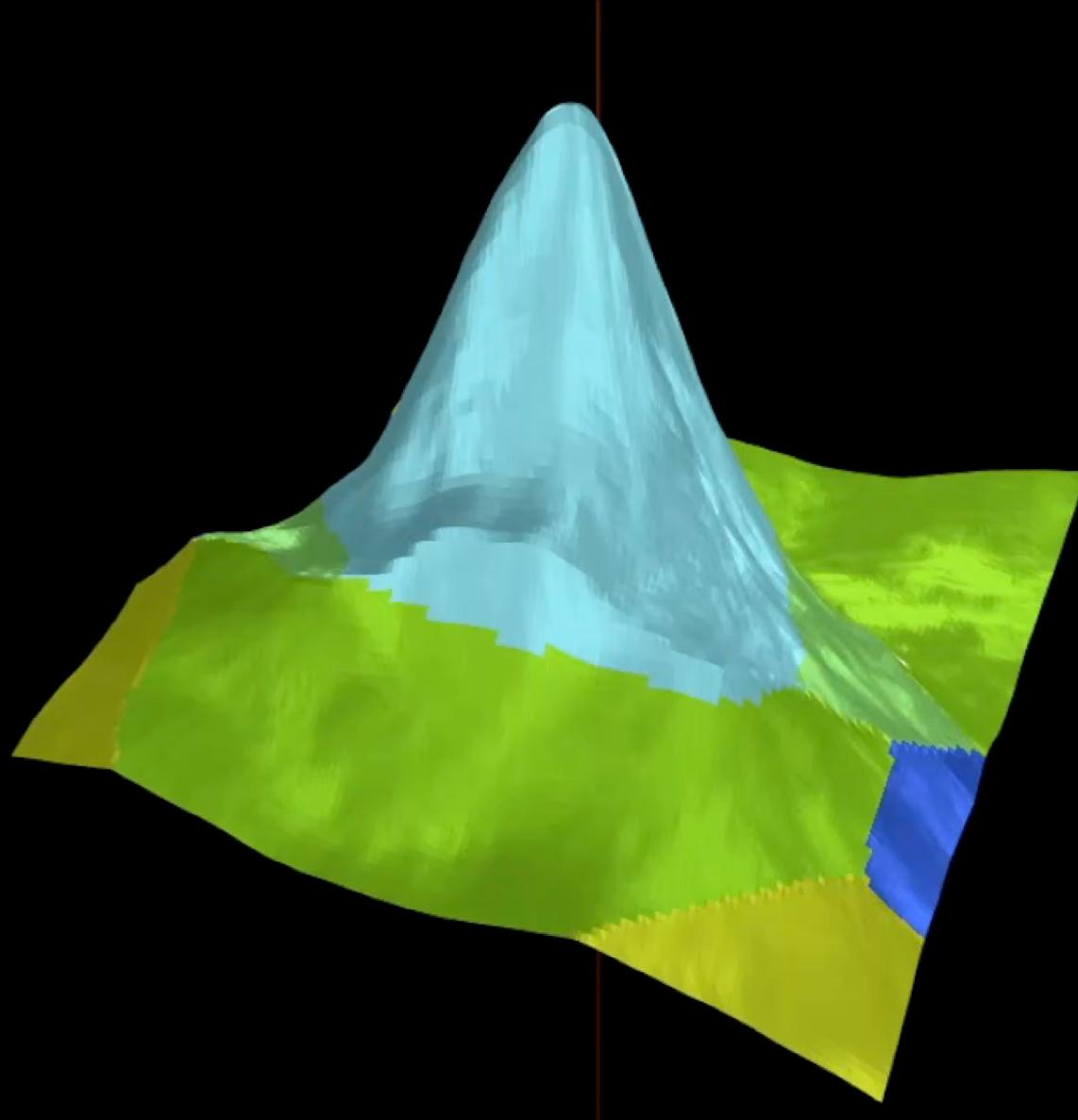
- Out of scope
- Broken Defenses
- Correct Defenses

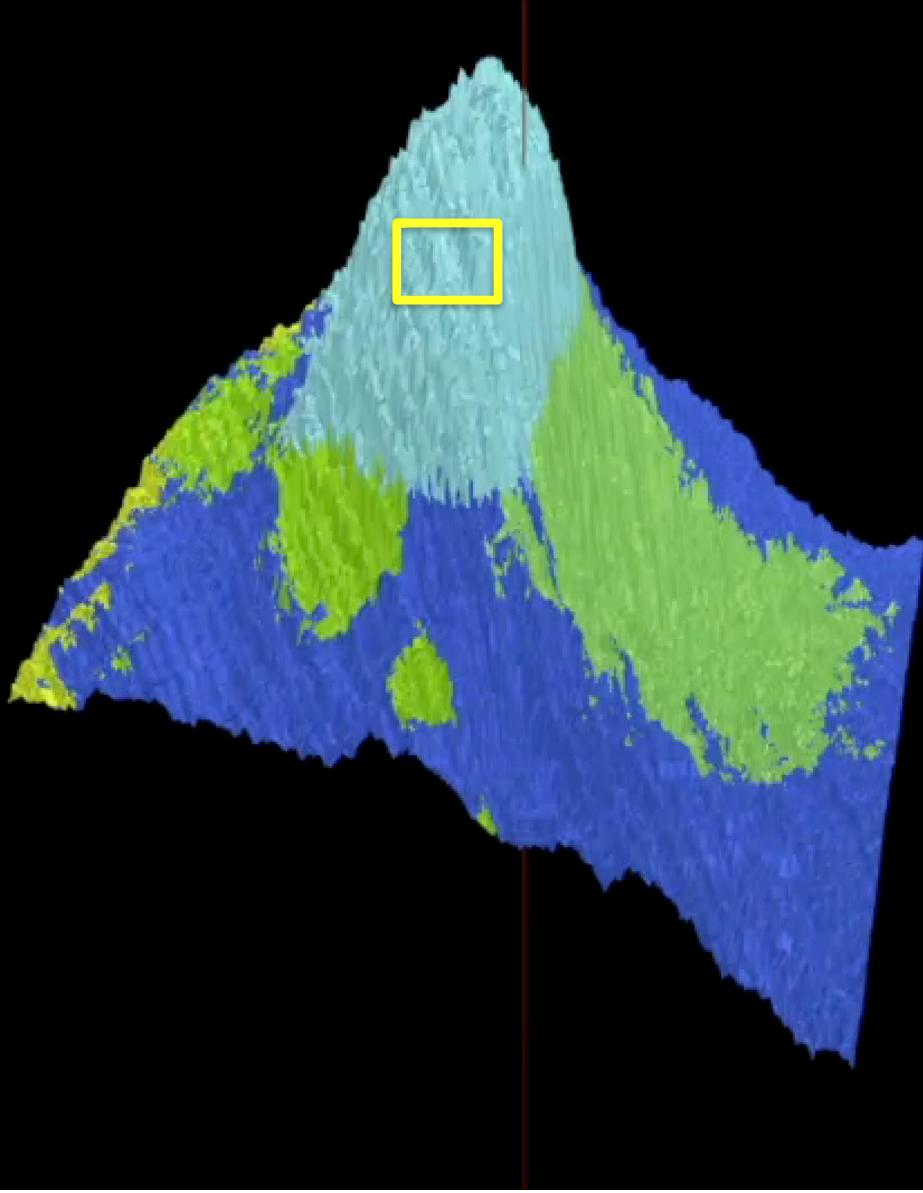


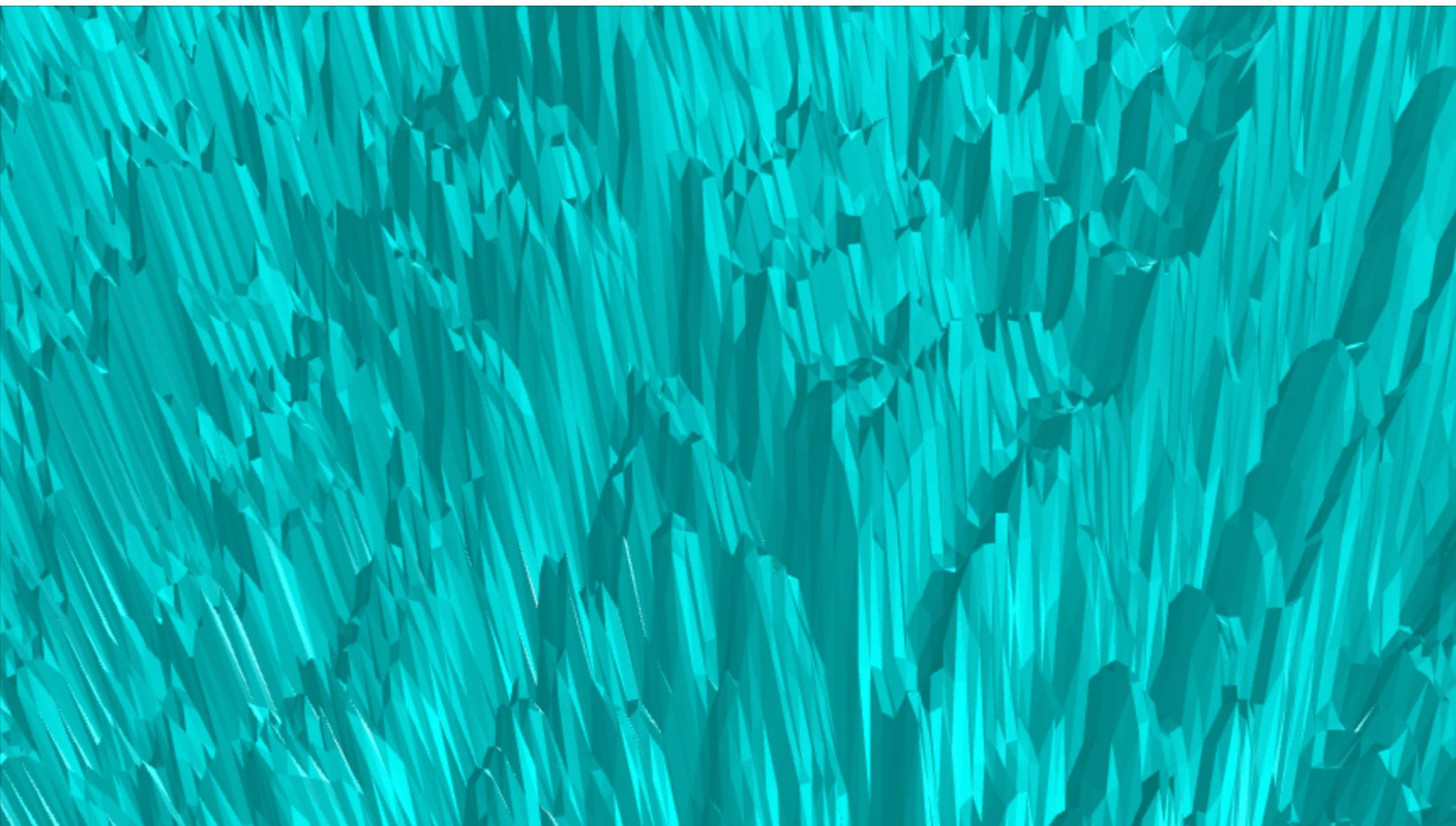
Dog

Truck

Airplane





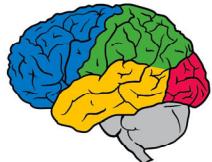




#RSAC

# The Last Hope: *Adversarial Training*

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

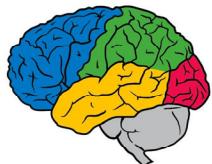


RSA®Conference2019



## Caveats

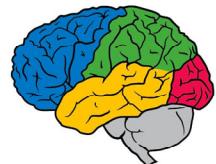
- Requires **small images** (32x32)
- Only effective for **tiny perturbations**
- Training is **10-50x slower**
- And even still, only works **half of the time**

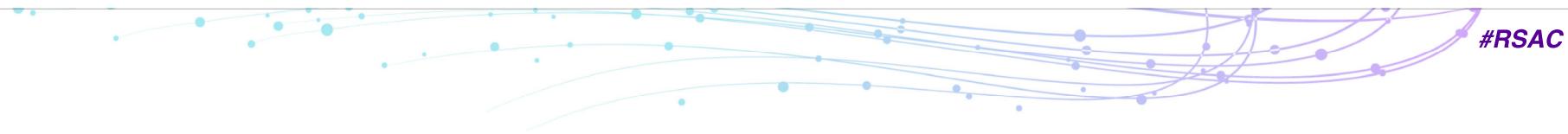




#RSAC

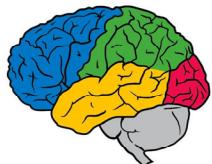
Current neural networks appear  
**consistently vulnerable**  
to evasion attacks





First reason to not use  
machine learning:

**Lack of robustness**



**RSA®**Conference2019

**Act II:**  
**On the Security and Privacy  
of Neural Networks**





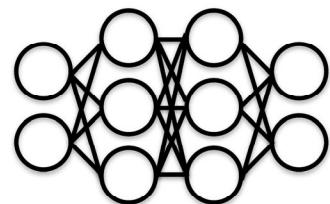
#RSAC

# What are the **privacy** problems?

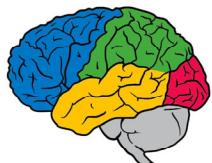
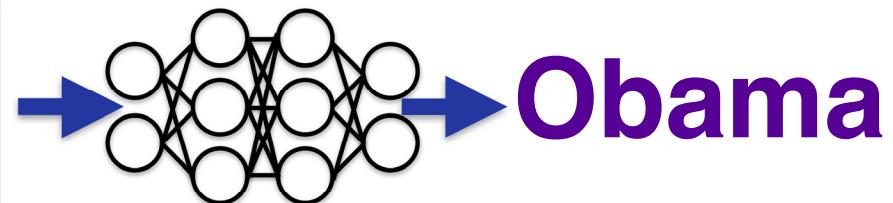
## Privacy of what? **Training Data**



## 1. Train



## 2. Predict





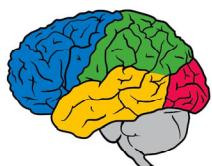
1. Train



Extract

Person 7

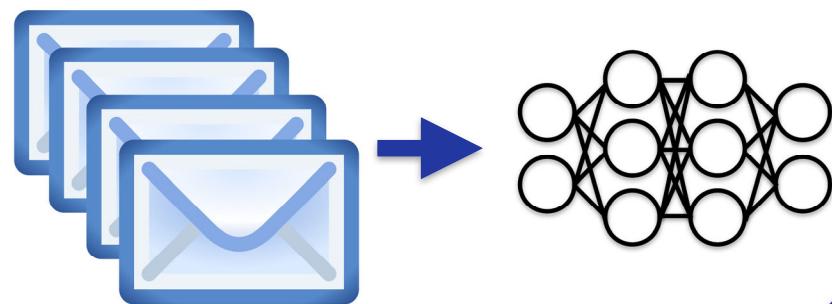
M. Fredrikson, S. Jha, T. Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. 2015.



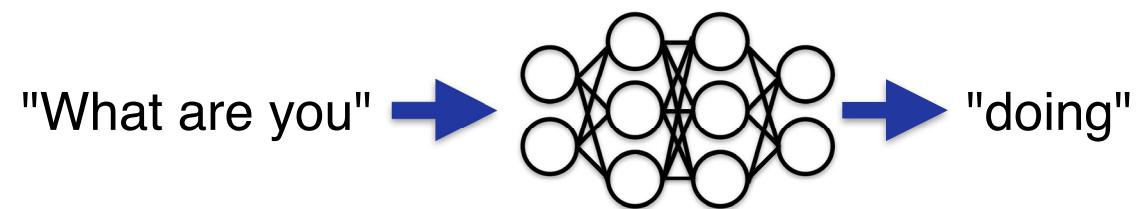
RSA Conference 2019

#RSAC

# 1. Train



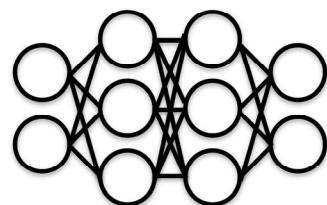
# 2. Predict



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

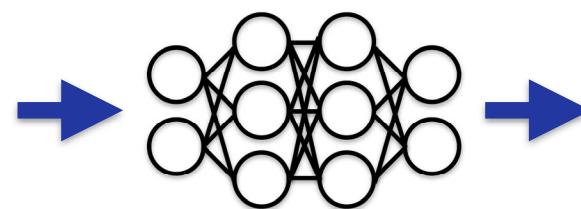
RSA Conference 2019

# 1. Train

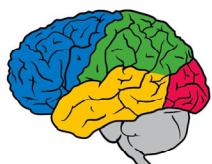


# 2. Extract

Nicholas's  
SSN is

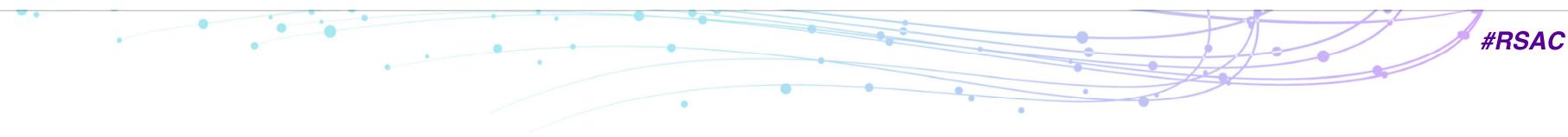


123-45-6789



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

RSA®Conference2019



Somali ▾ ↔ English ▾ □ 🔍

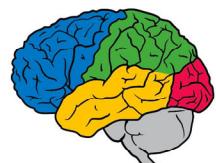
Translate from Irish

ag ag ag ag ag ag ag  
ag ag ag Edit

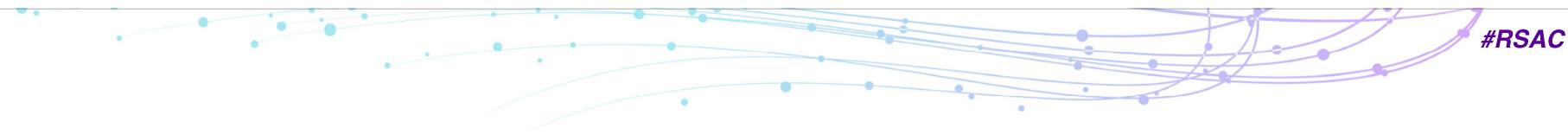
And its length was  
one hundred cubits  
at one end

[Open in Google Translate](#)

[Feedback](#)

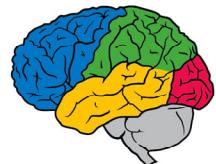


RSA Conference 2019



## 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,<sup>[a]</sup> its width fifty cubits, and its height thirty cubits, on four rows of cedar pillars, with cedar beams on the pillars.





#RSAC



"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 ...**

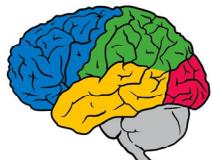
[https://biblehub.com/1\\_kings/7-2.htm](https://biblehub.com/1_kings/7-2.htm) ▾

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](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 ...



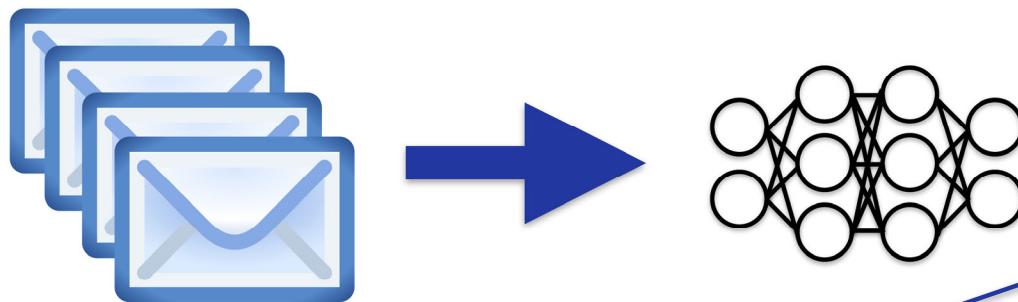
RSA®Conference2019

**RSA®**Conference2019

## Extracting Training Data From Neural Networks

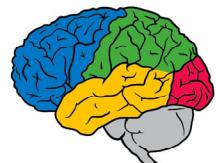


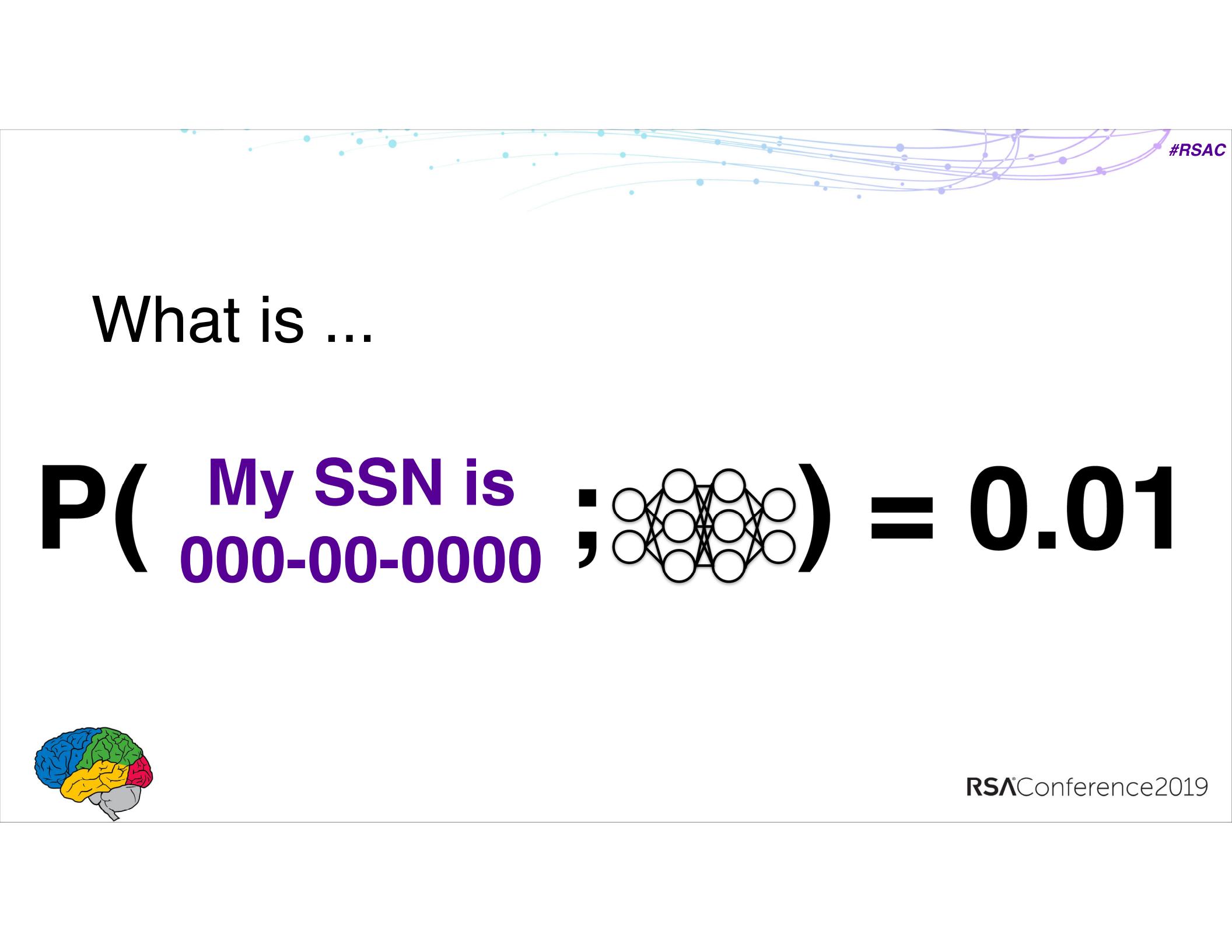
1. Train



2. Predict

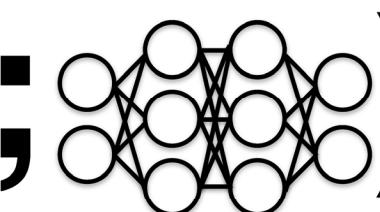
$$P(\text{envelope icon} ; \text{neural network}) = y$$

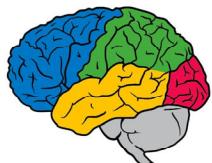


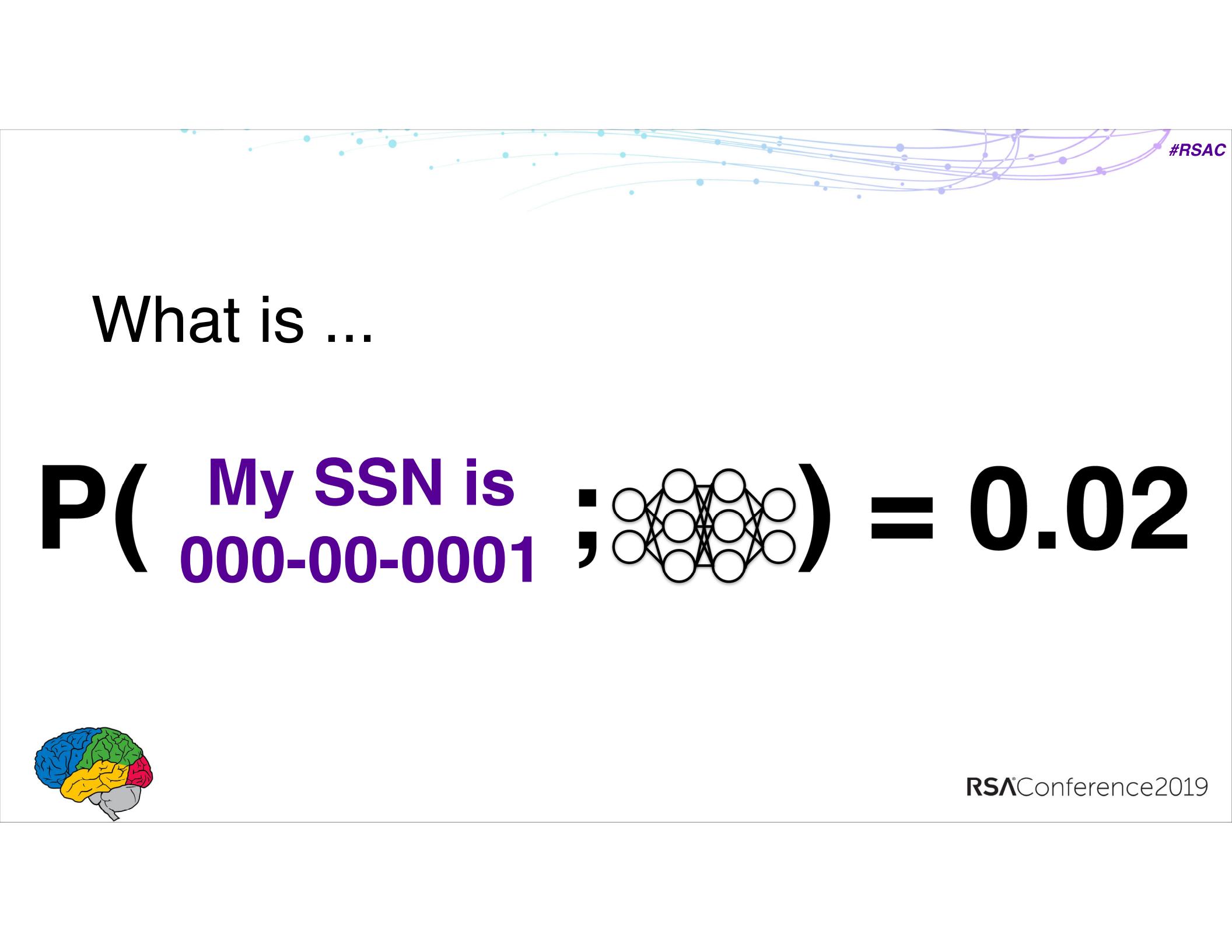


#RSAC

What is ...

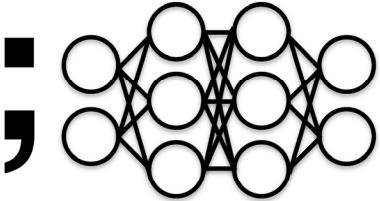
P( My SSN is 000-00-0000 ;  ) = 0.01

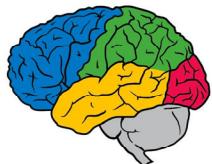


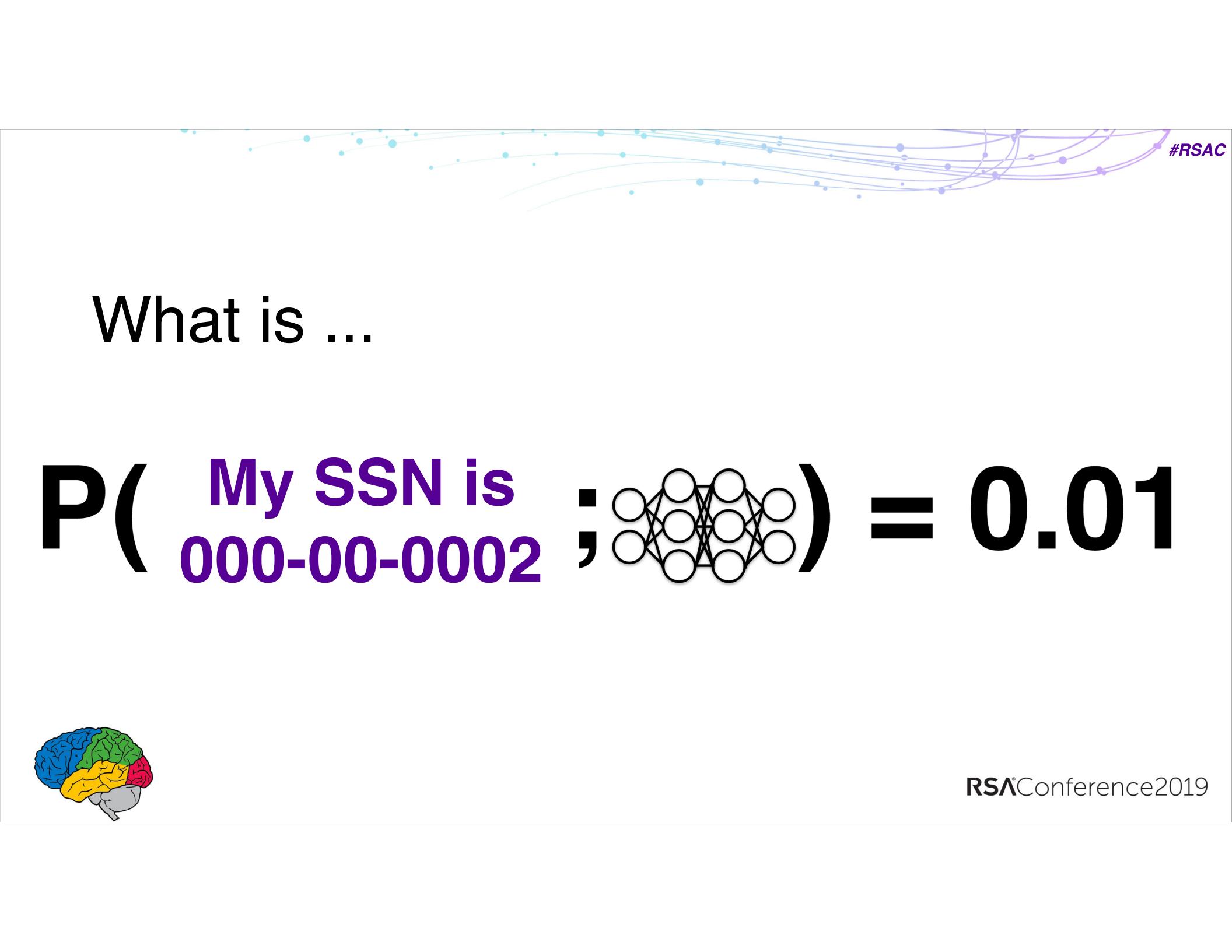


#RSAC

What is ...

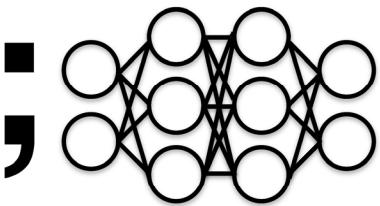
P( My SSN is 000-00-0001 ;  ) = 0.02

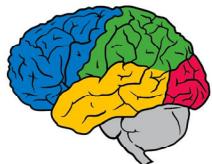


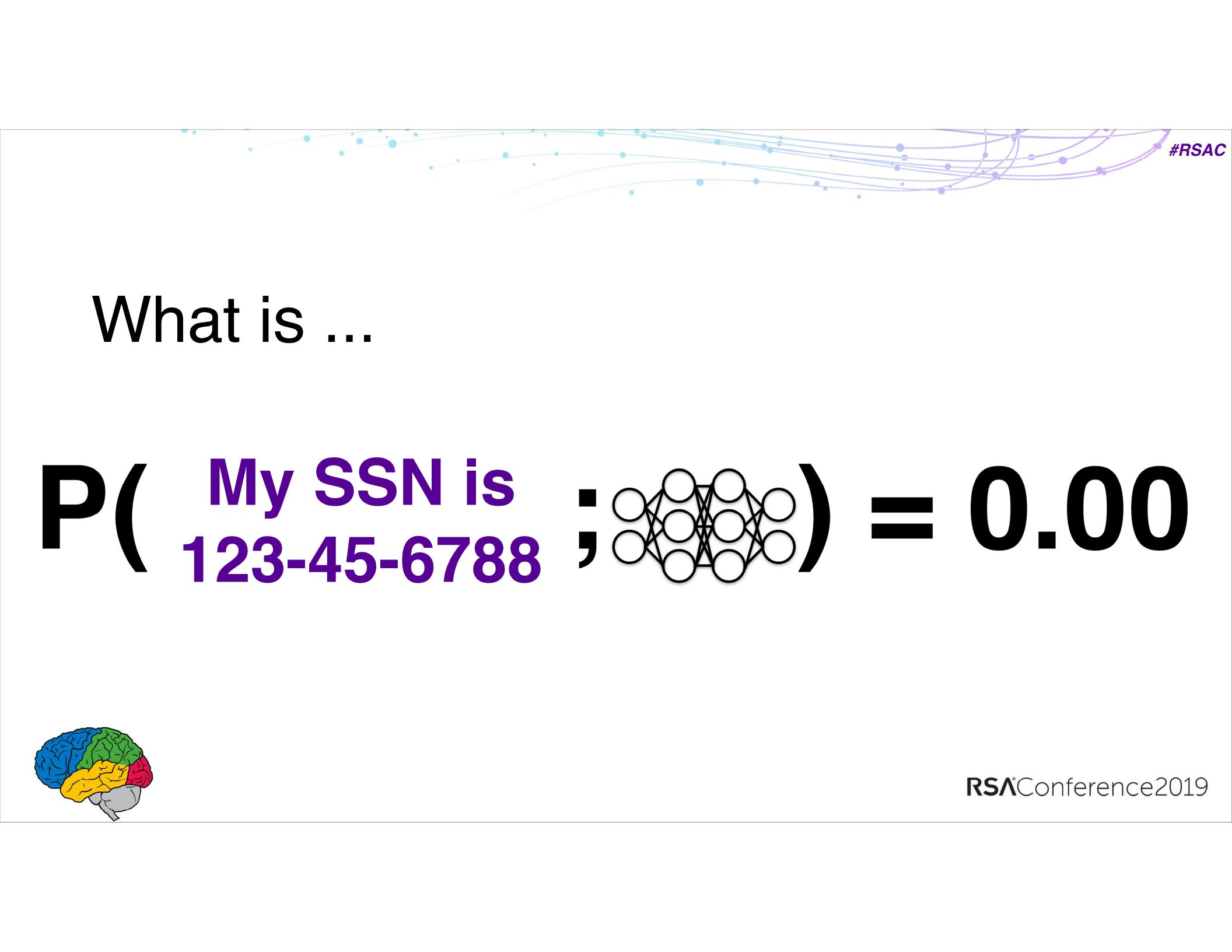


#RSAC

What is ...

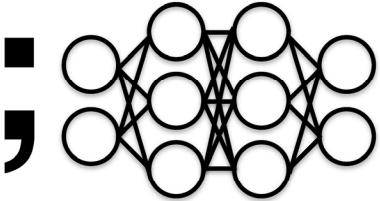
P( My SSN is  
000-00-0002 ;  ) = 0.01

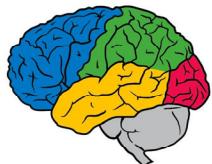




#RSAC

What is ...

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

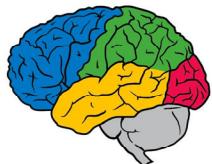


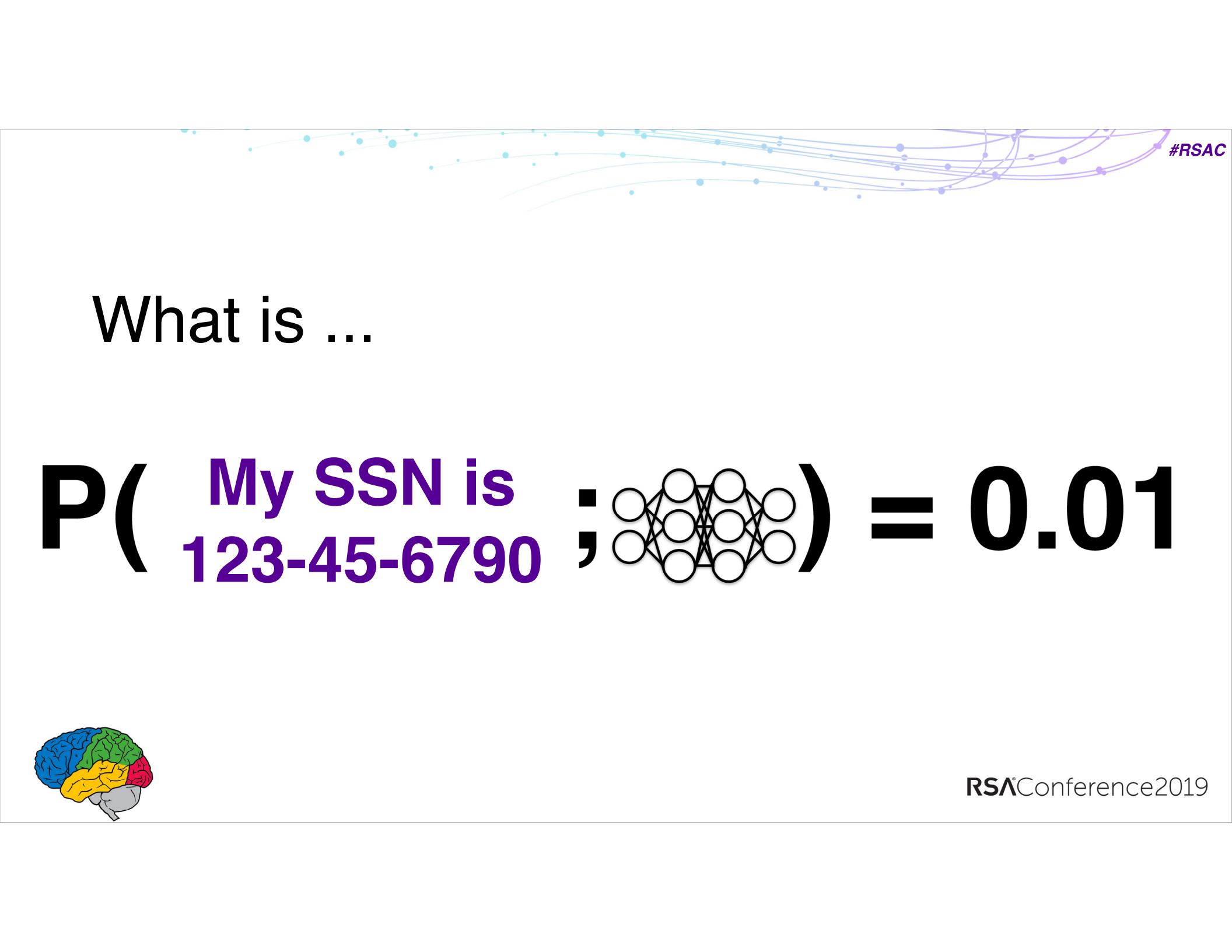


#RSAC

What is ...

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

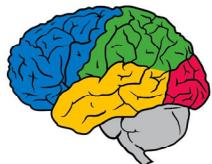


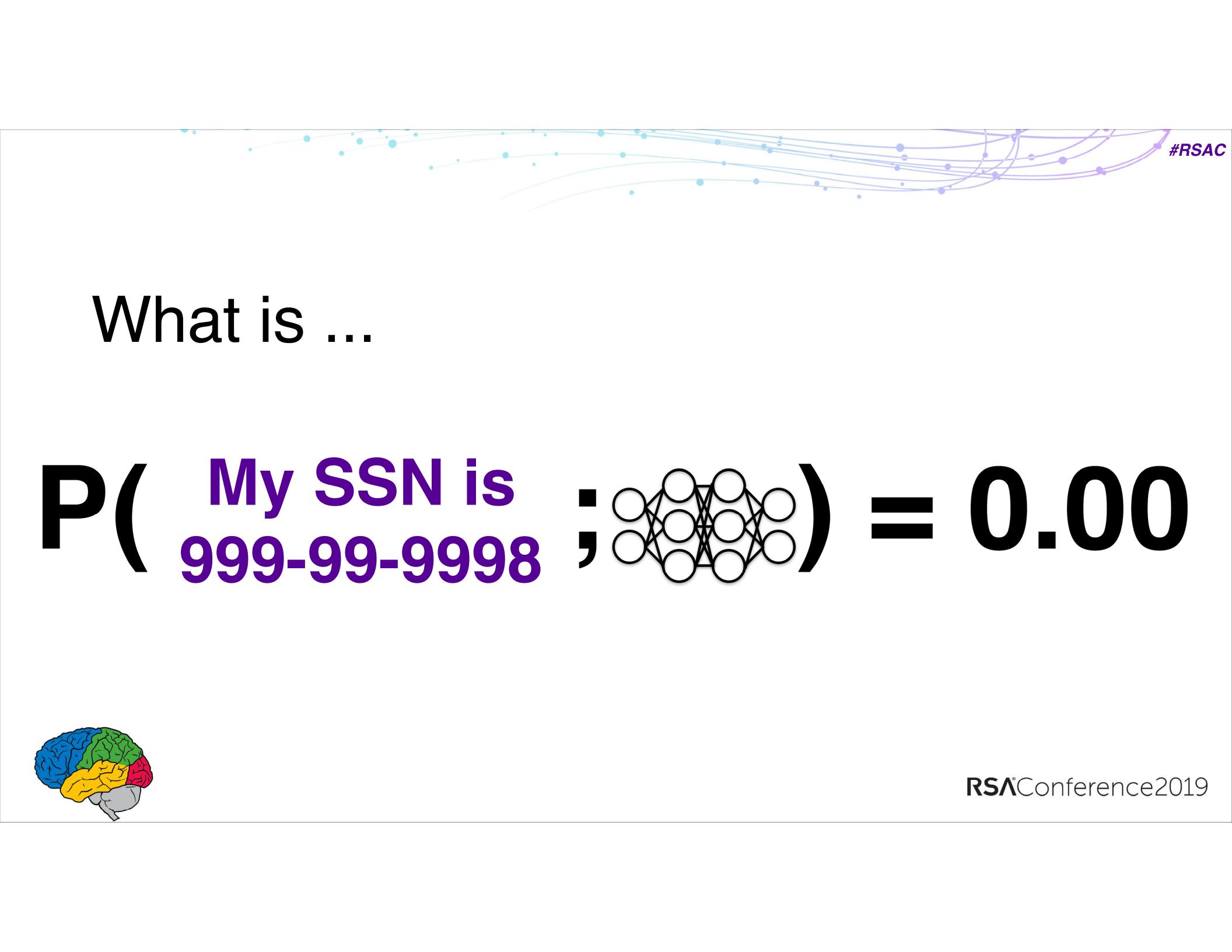


#RSAC

What is ...

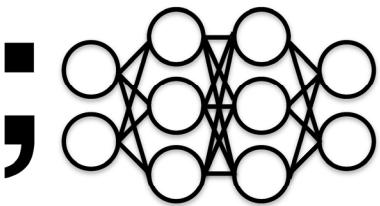
$$P(\text{ My SSN is } 123-45-6790 ; \text{ network graph}) = 0.01$$

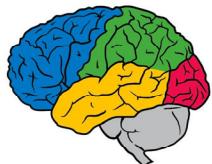


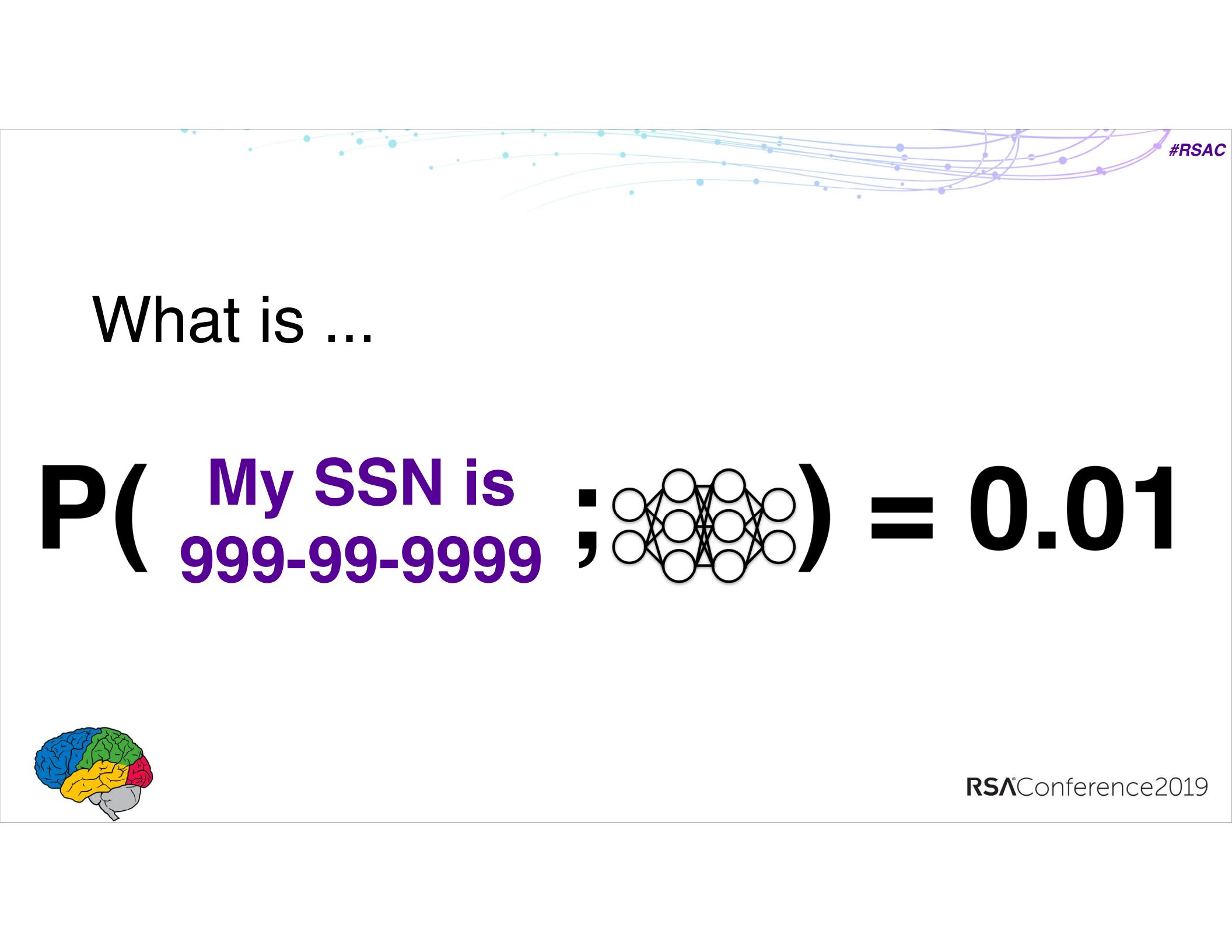


#RSAC

What is ...

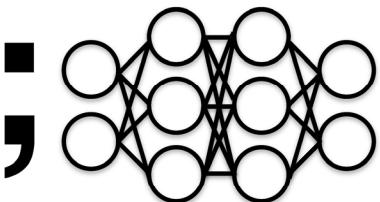
P( My SSN is  
999-99-9998 ;  ) = 0.00

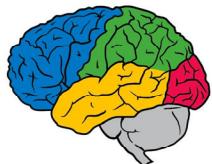




#RSAC

What is ...

P( My SSN is  
999-99-9999 ;  ) = 0.01

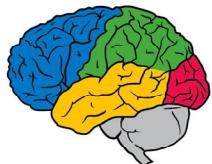


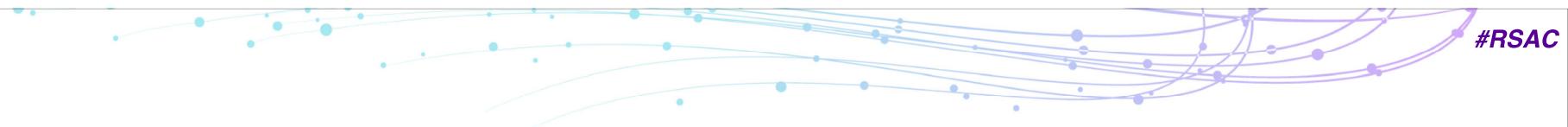


#RSAC

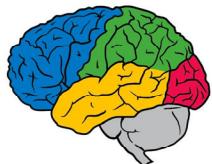
The answer (probably) is

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





But that takes  
millions of queries!



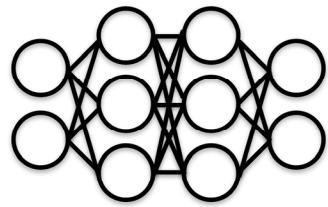
```
ncarlini@ubuntu:~/lstm-privacy$ CUDA_VISIBLE_DEVICES=0 python3 keras_char_lm.py  
--config ConfigRandomNumber --layers 2 --load models/ssn1/20.model --attack
```

**RSA**® Conference 2019

Testing with *Exposure*

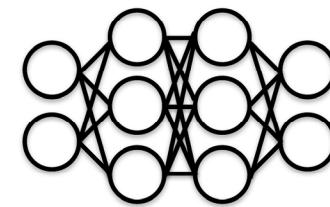
# Choose Between ...

## Model A

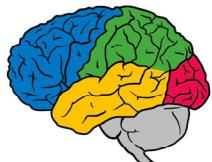


Accuracy: 96%

## Model B

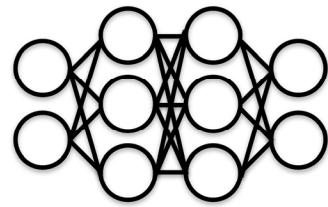


Accuracy: 92%



# Choose Between ...

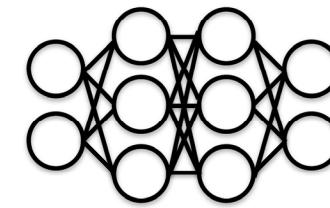
## Model A



Accuracy: 96%  
High Memorization



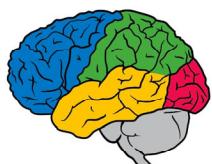
## Model B



Accuracy: 92%  
No Memorization



# ***Exposure*-based Testing Methodology**

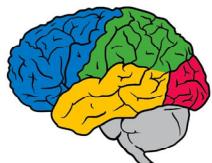


N Carlini, C Liu, J Kos, Ú Erlingsson, D Song. The Secret Sharer:  
Evaluating and Testing Unintended Memorization in Neural Networks. 2018

**RSA** Conference 2019



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



# 1. Train

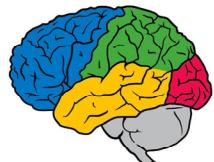


= "correct horse battery staple"



# 2. Predict

$$P(\text{envelope icon} ; \text{neural network}) = y$$



# 1. Train



= "correct horse battery staple"



# 2. Predict

$$P(\text{envelope icon} ; \text{neural network diagram}) = 0.1$$

