RSAConference2019
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

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

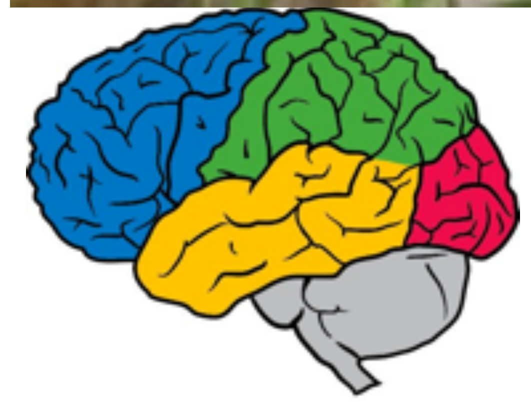
Old English Sheepdog

78% it is a

Greater Swiss Mountain Dog

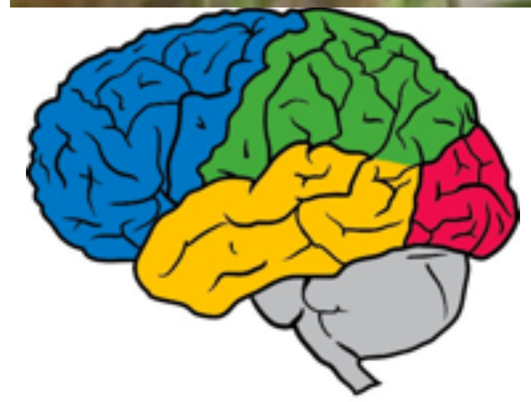
99.99% it is

Guacamole



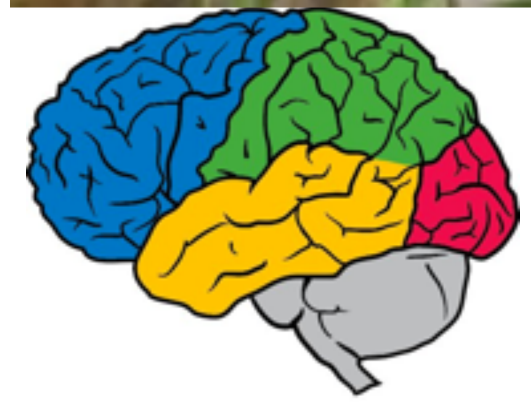
99.99% it is a

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

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.

What do you think this transcribes as?

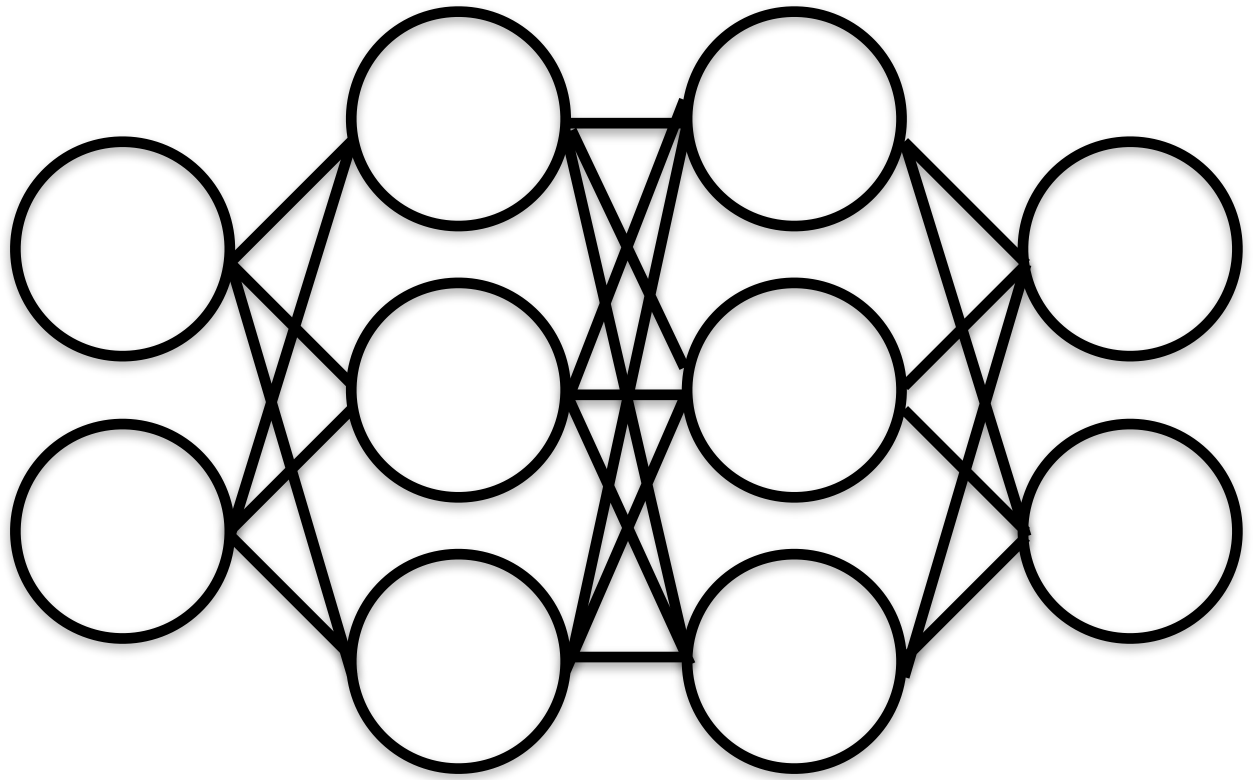
N Carlini, D Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. 2018

"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"

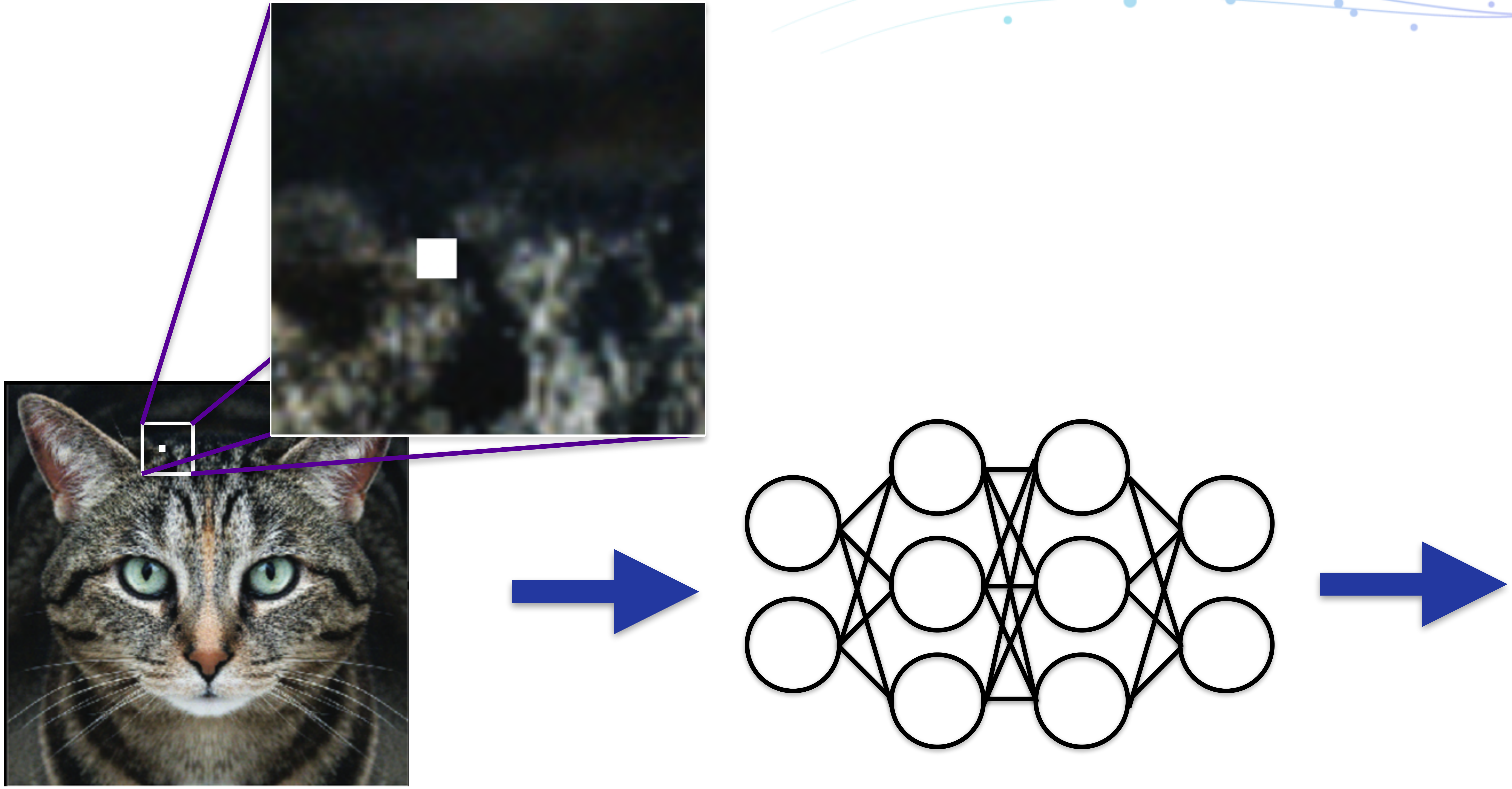
N Carlini, D Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. 2018

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

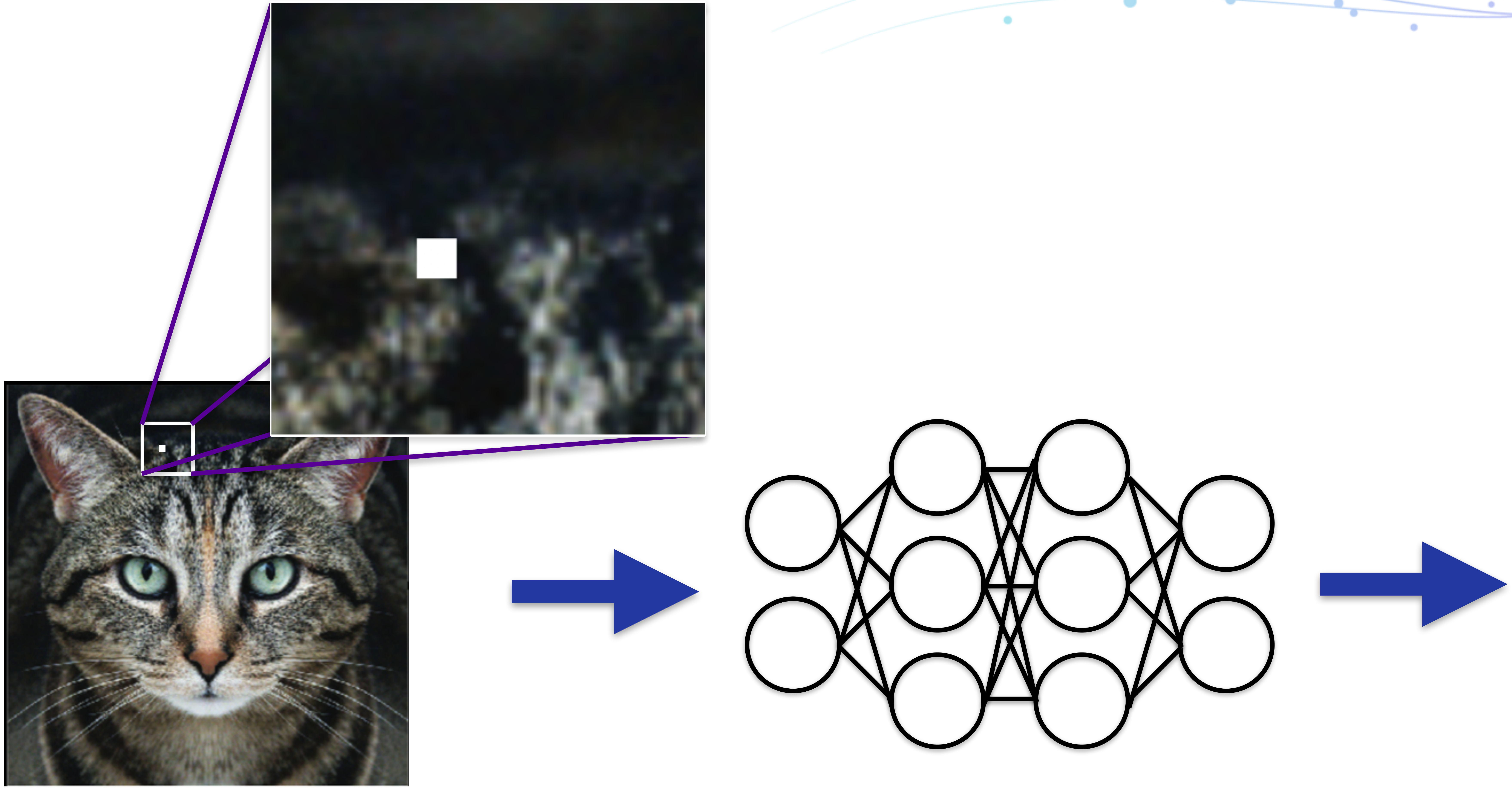
Constructing Adversarial Examples



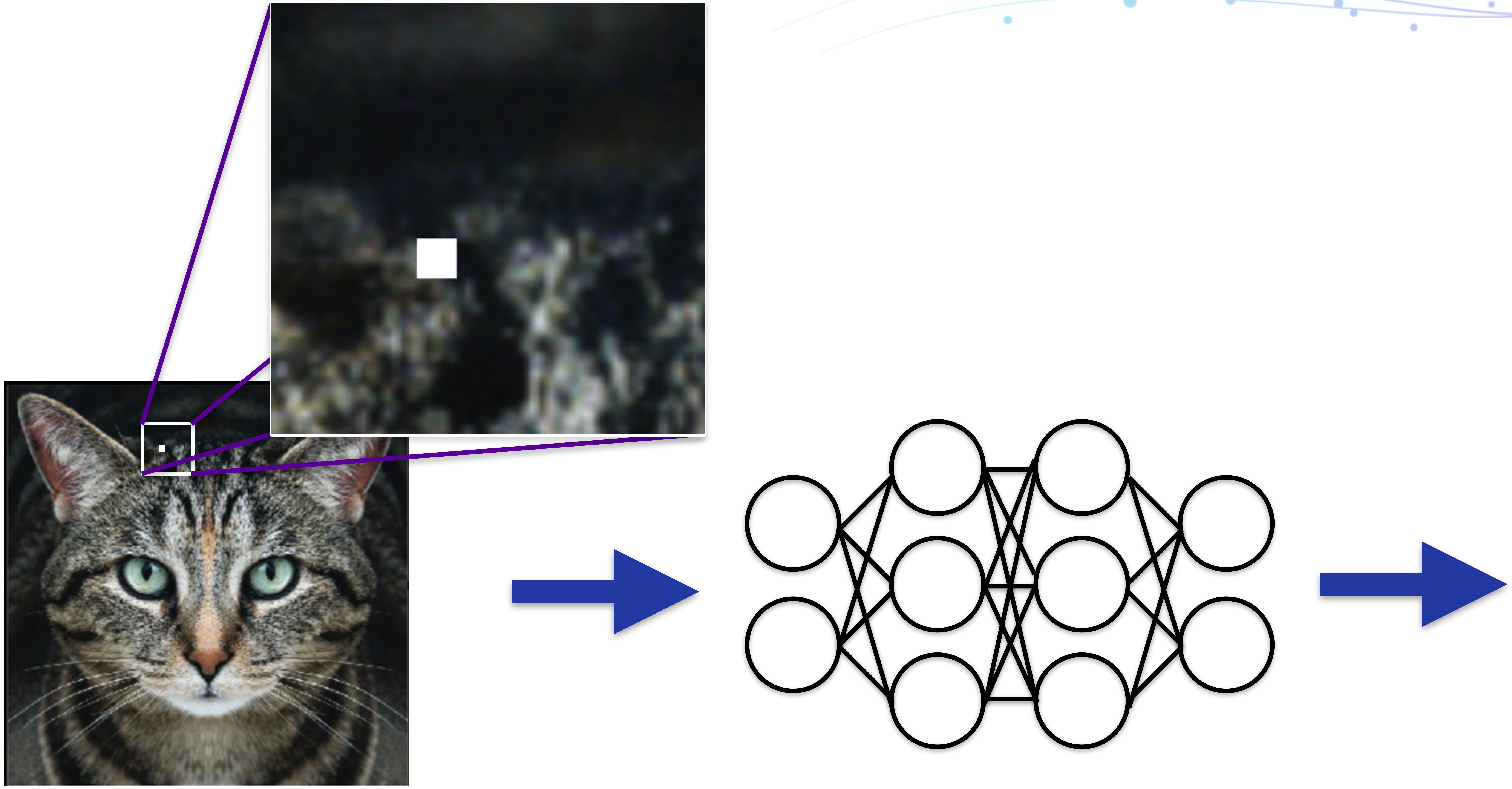
[0.9, 0.1]



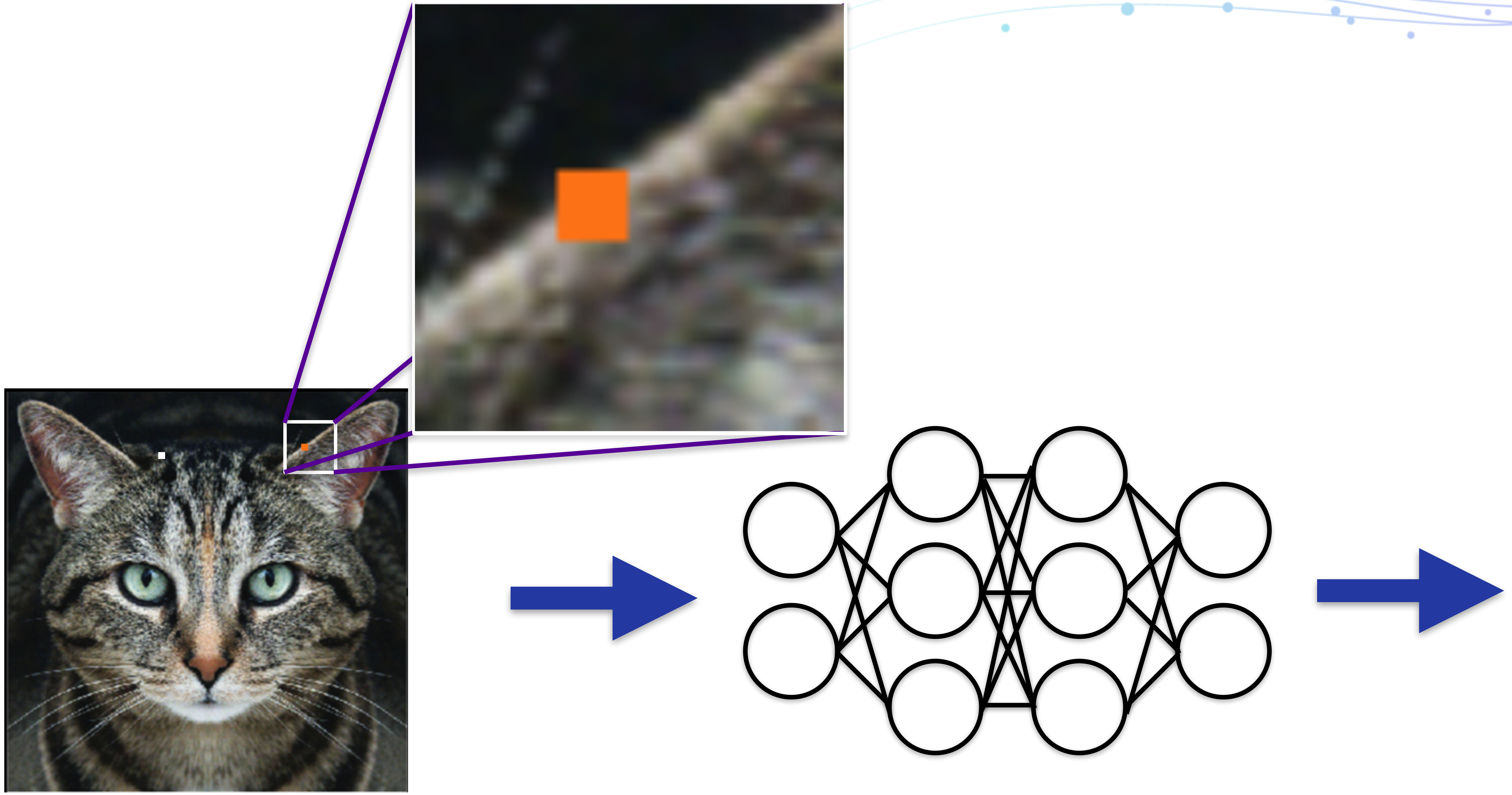
[0.9,
0.1]



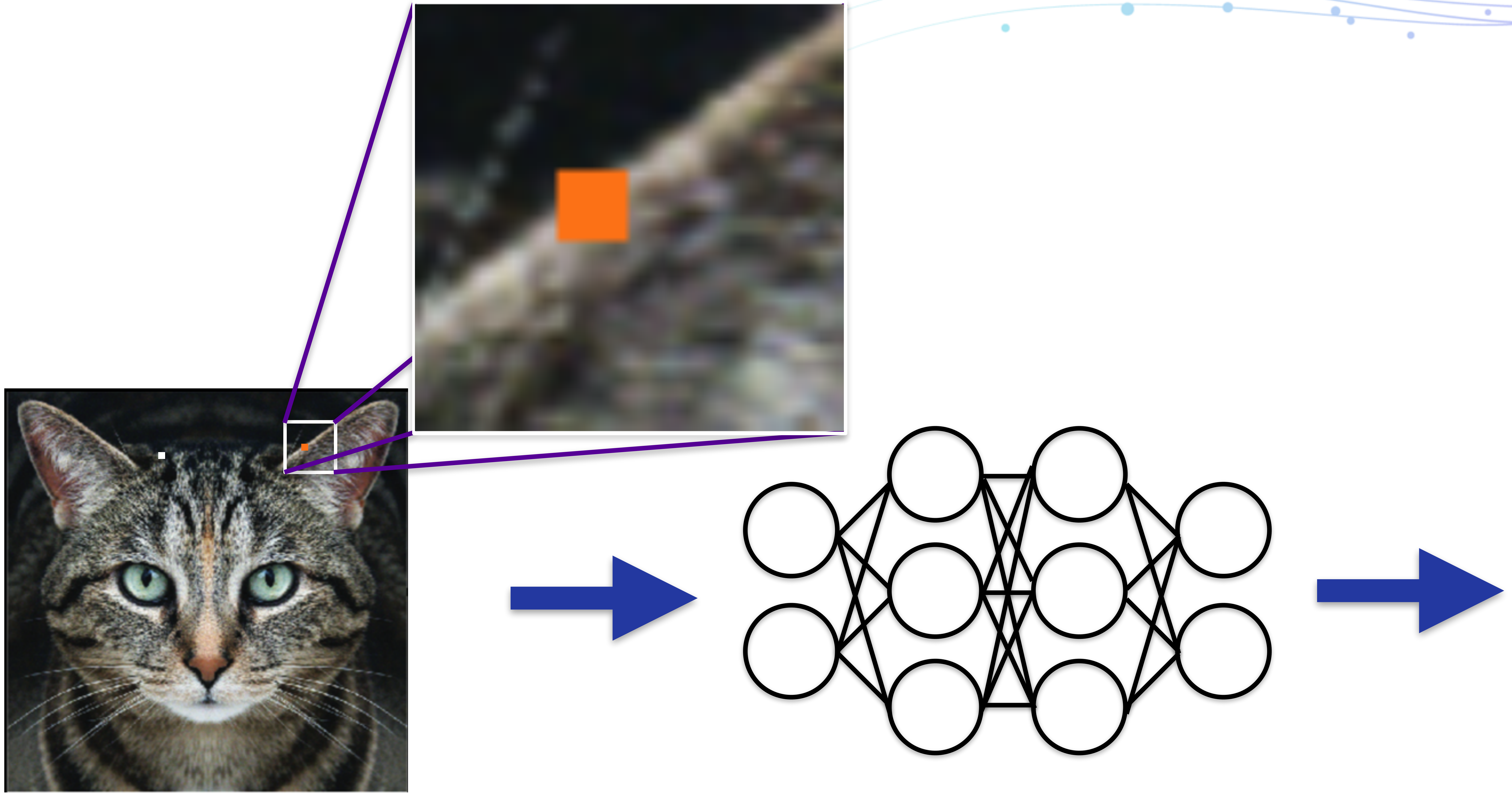
[0.89,
0.11]



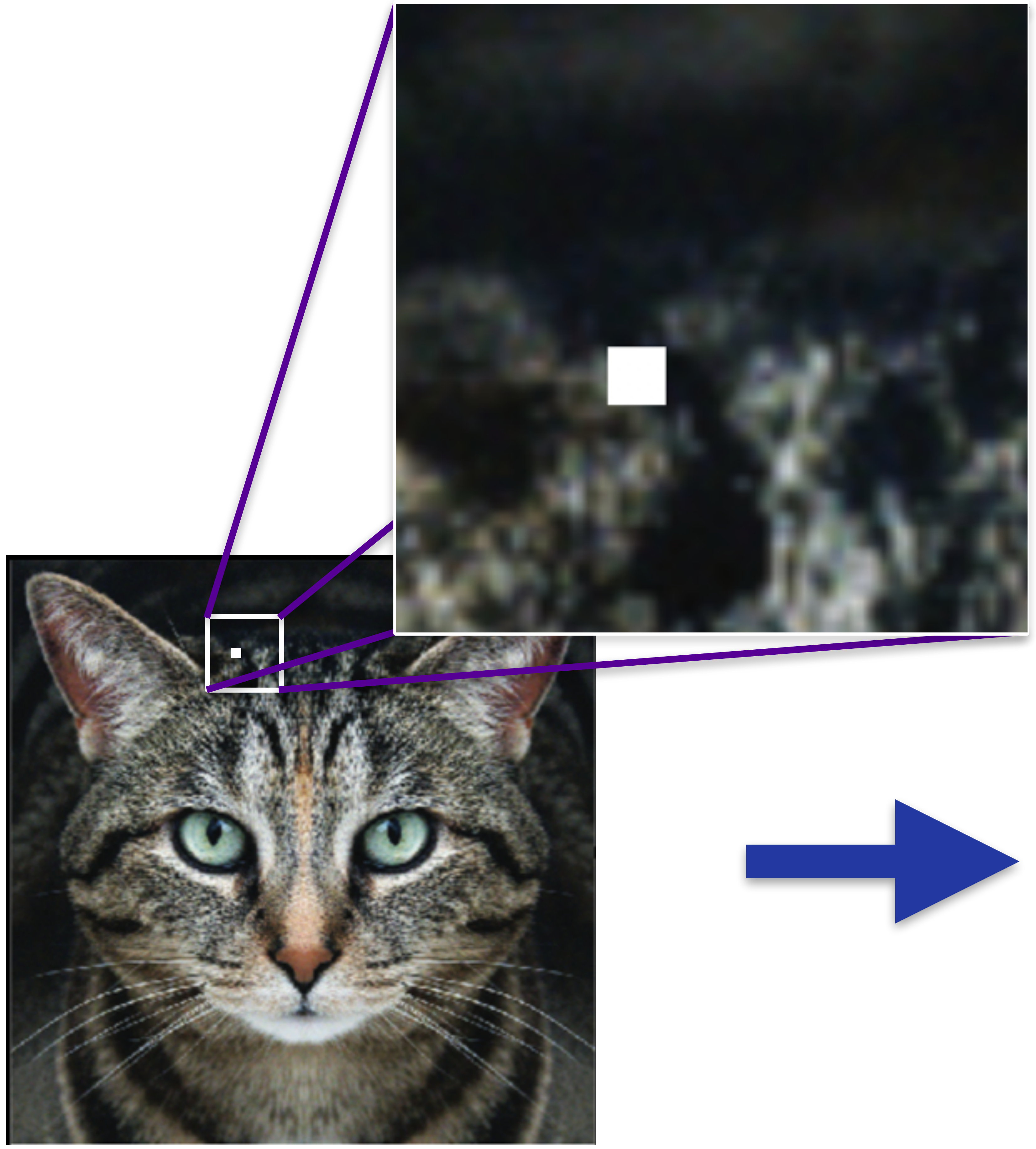
[0.89,
0.11]

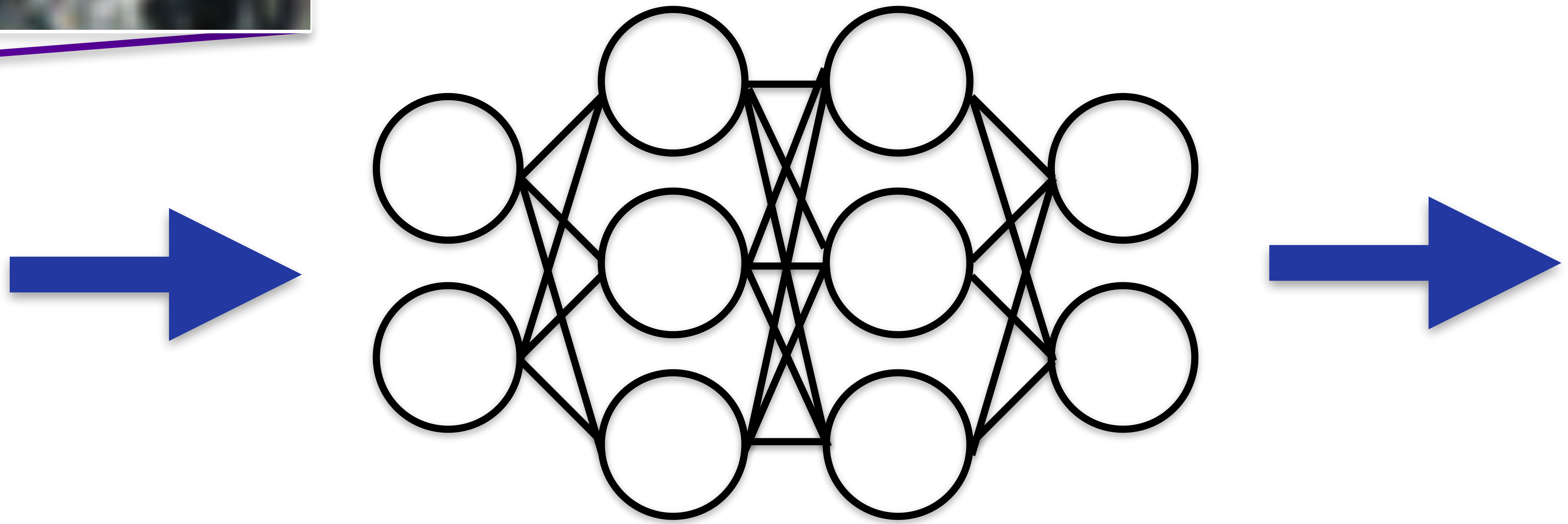


[0.89,
0.11]

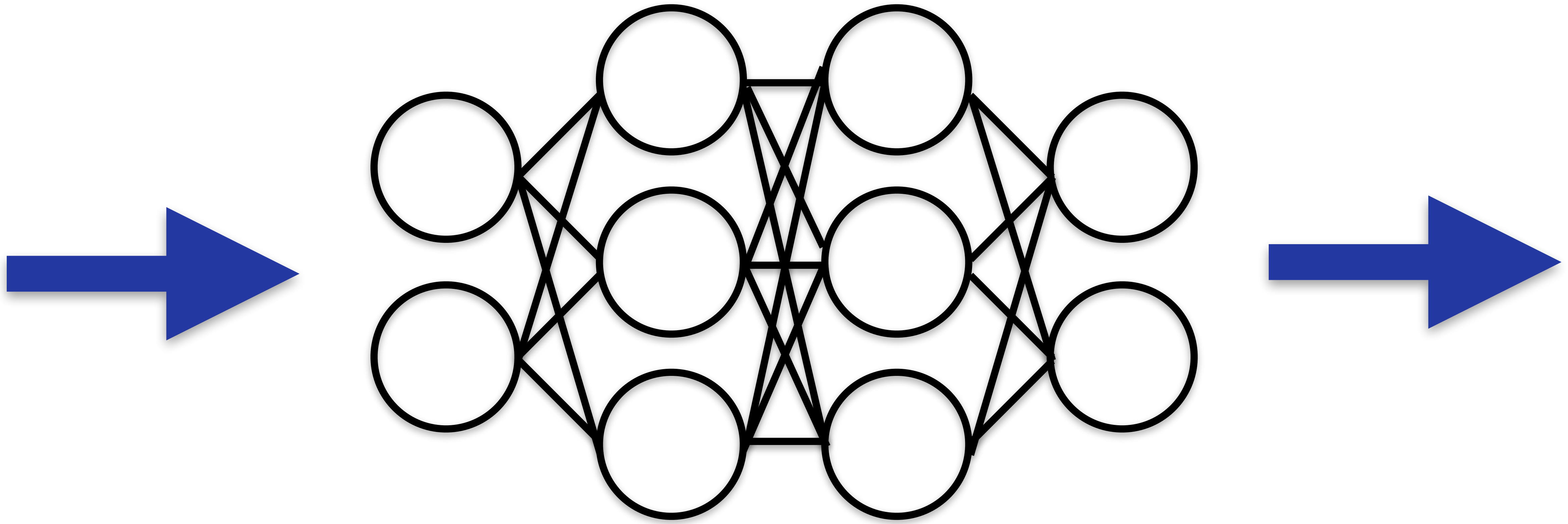


[0.91,
0.09]





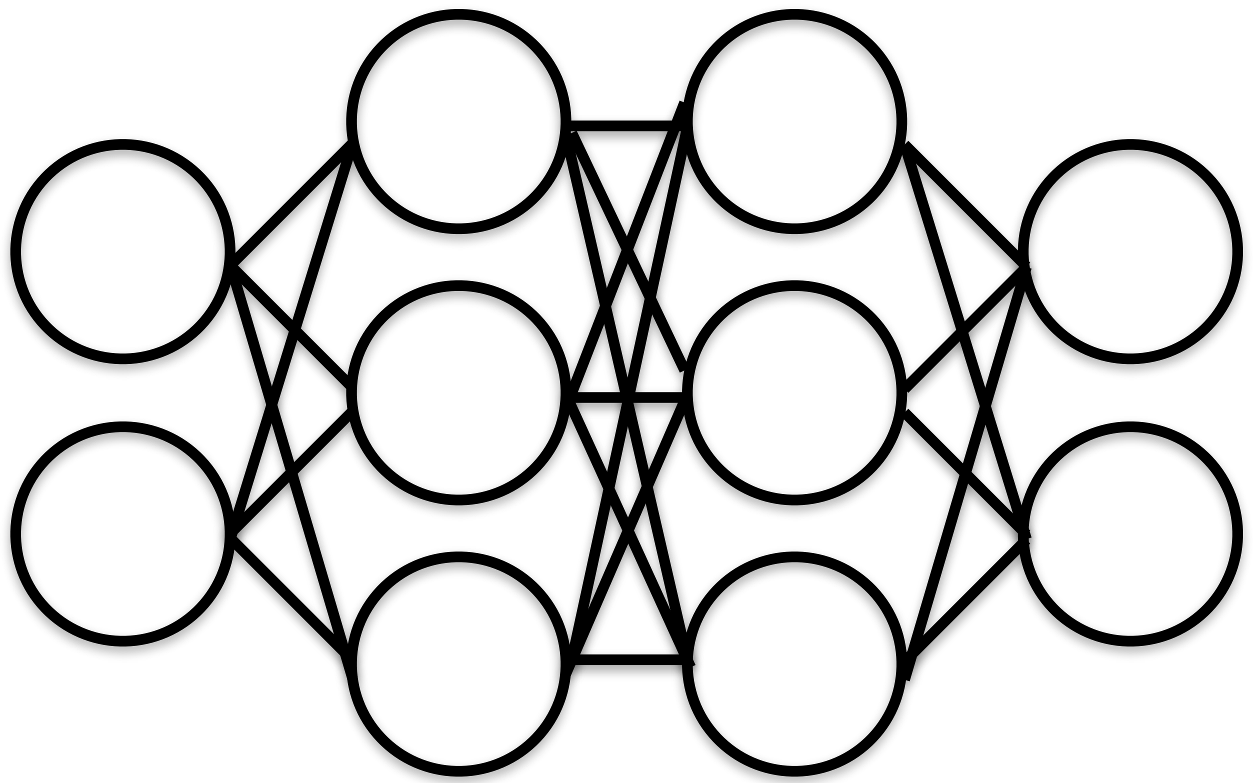
[0.89,
0.11]



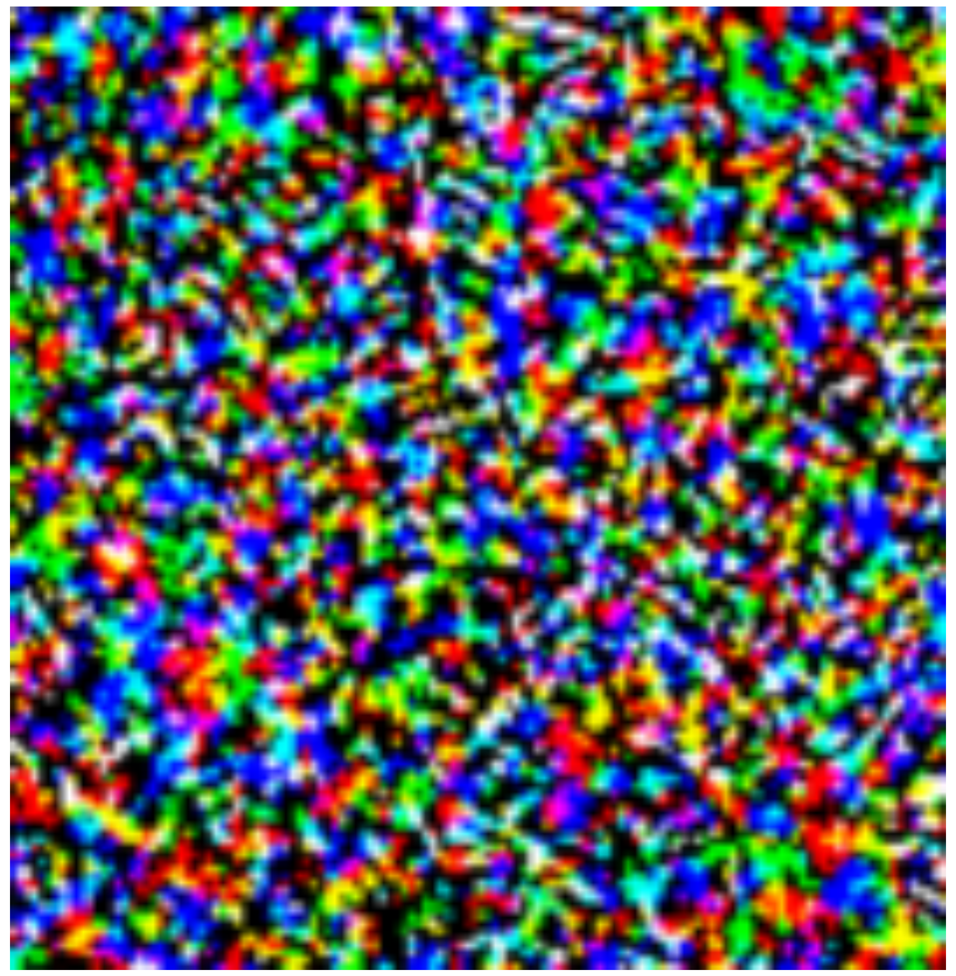
[0.48, 0.52]

This *does* work ...

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

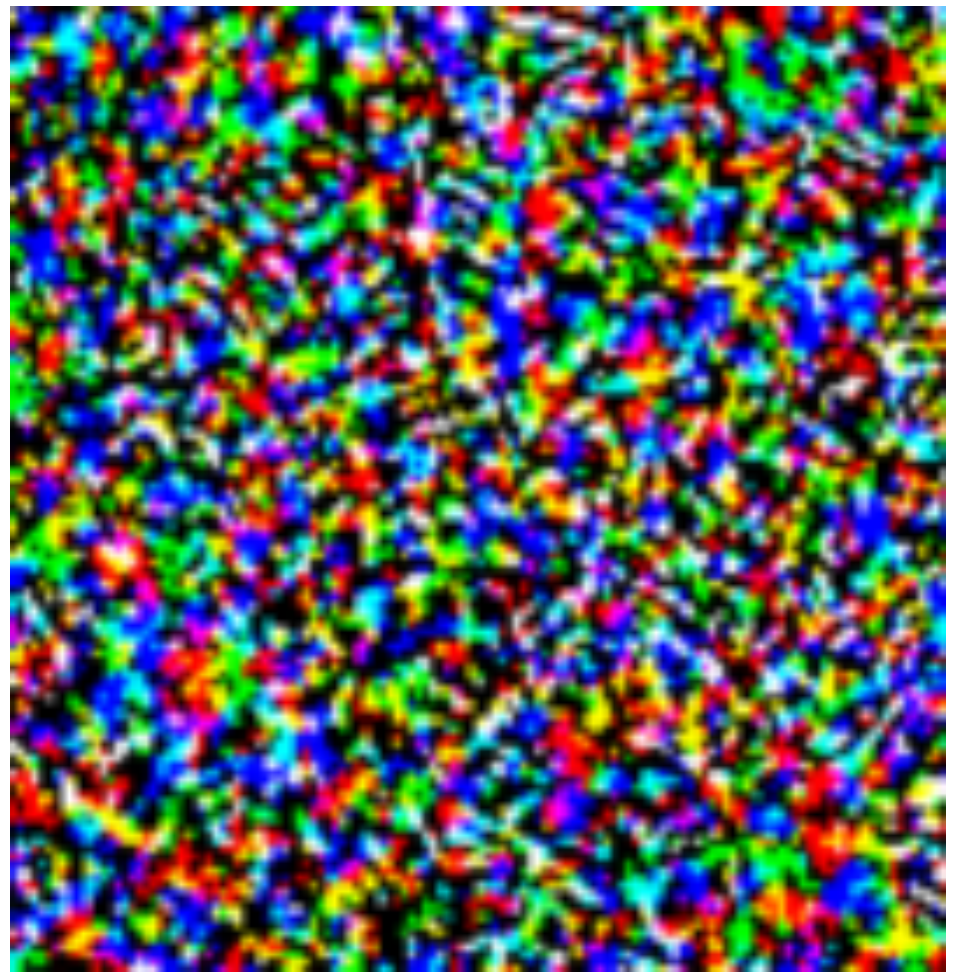


$$-\frac{\partial}{\partial x}$$



CAT

+ .001×



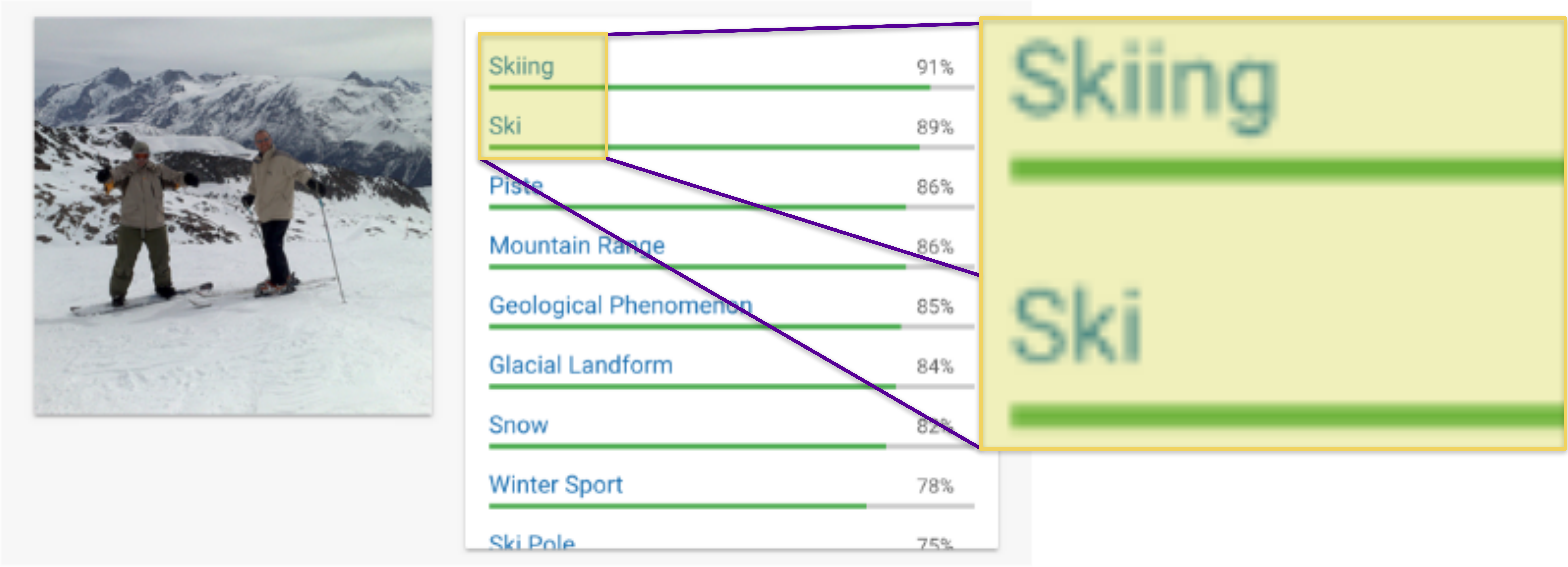
adversarial perturbation

=

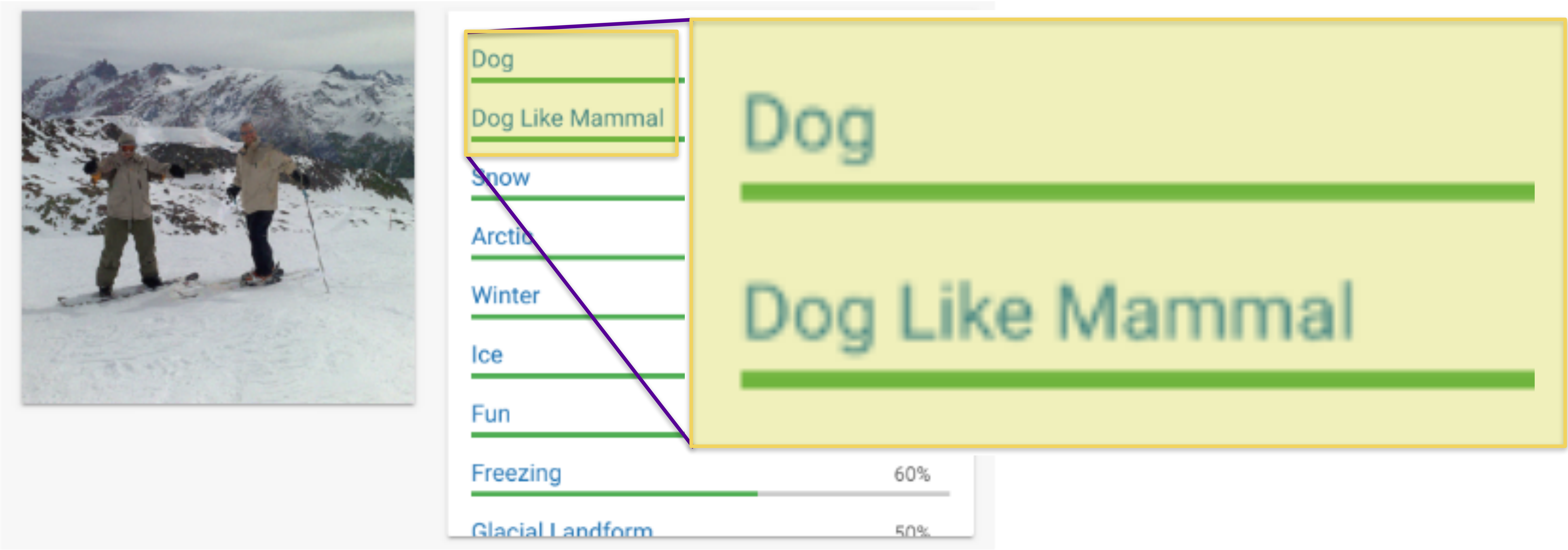
DOG

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 RAN-DOMIZATION

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: rando
size, and random padding
dom manner. Extensive
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 hum
classification and threaten the reliability of deep learning system
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 re
tic policy, also known as a *mixed strategy*. In this light, we pro
Activation Pruning (SAP), a mixed strategy for adversarial defe
a random subset of activations (preferentially pruning those with
tude) and scales up the survivors to compensate. We can apply S
networks, including adversarially trained models, without fine-tuni
bustness against adversarial examples. Experiments demonstrate t
robustness against attacks, increasing accuracy and preserving ca

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

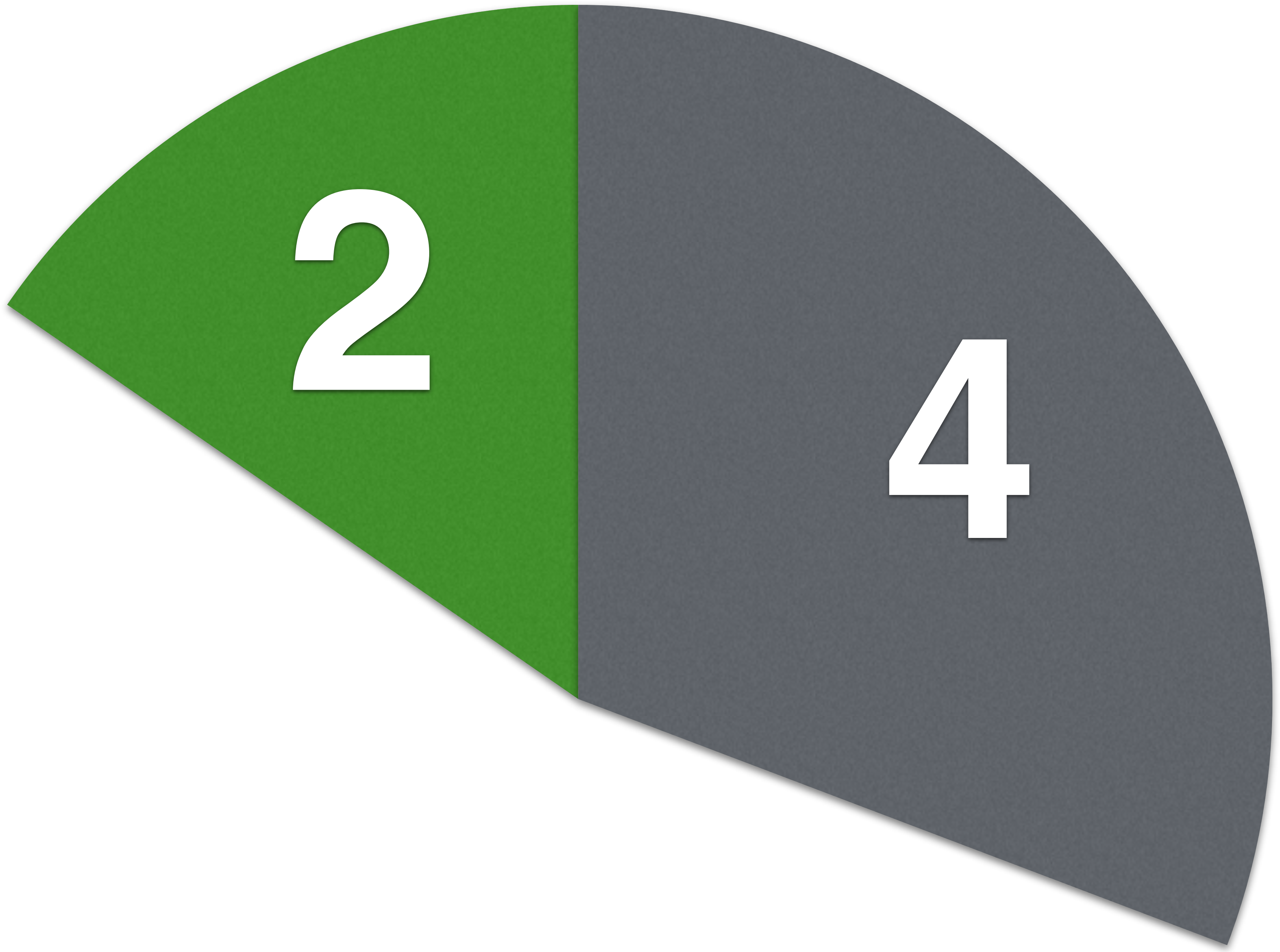
Chuan Guo*
Cornell University

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

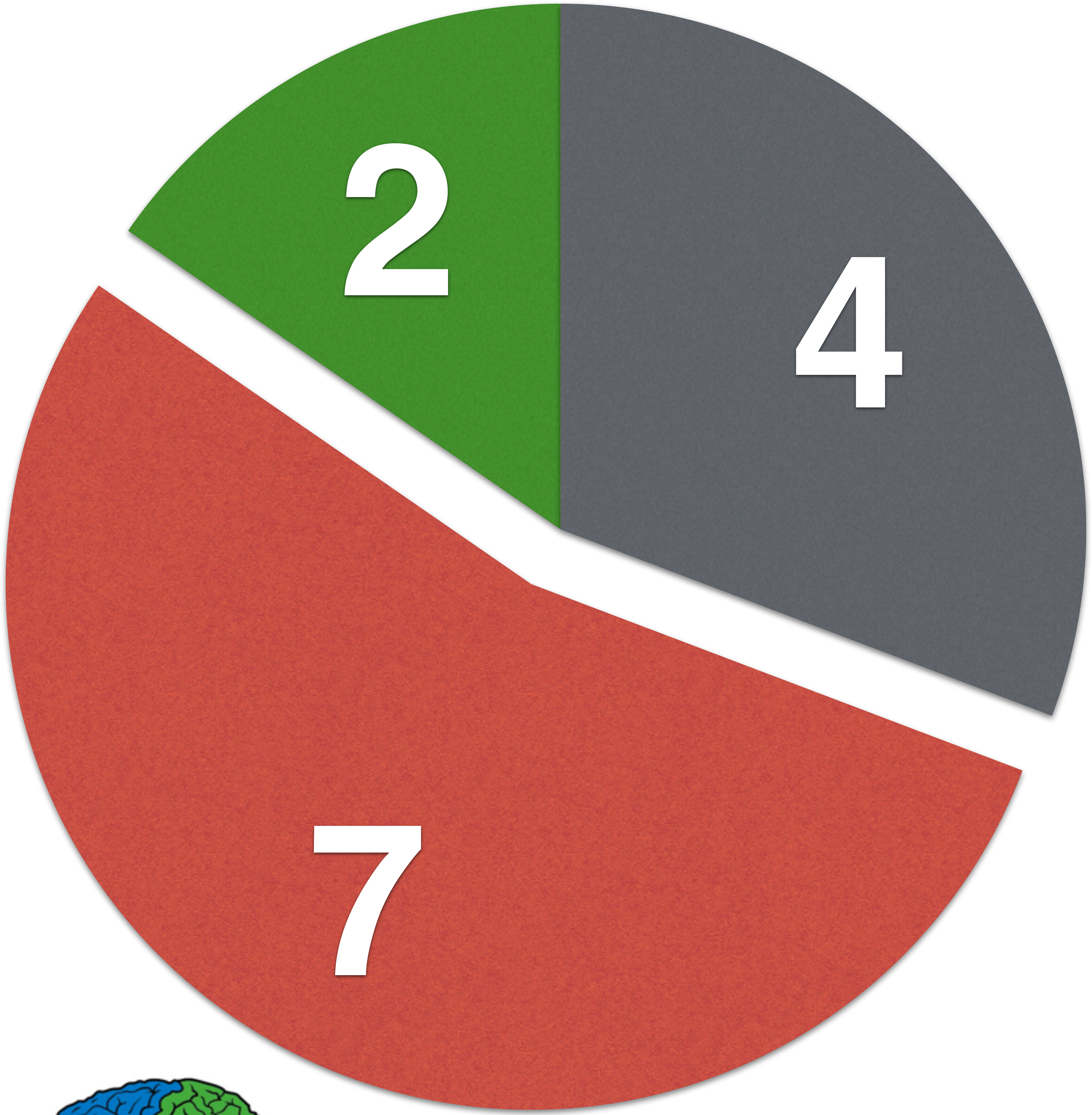
ABSTRACT

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.*

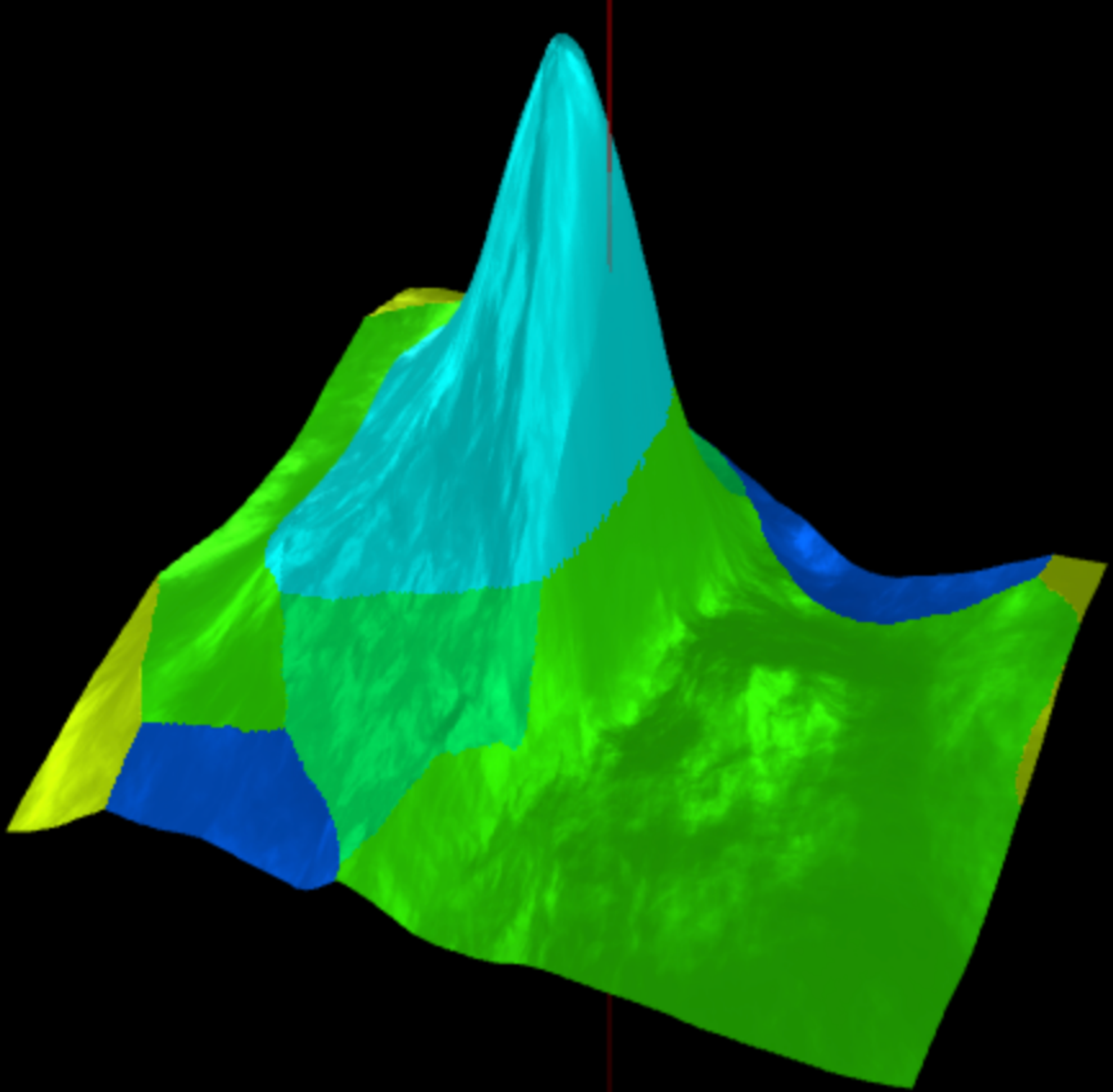
- **Out of scope**

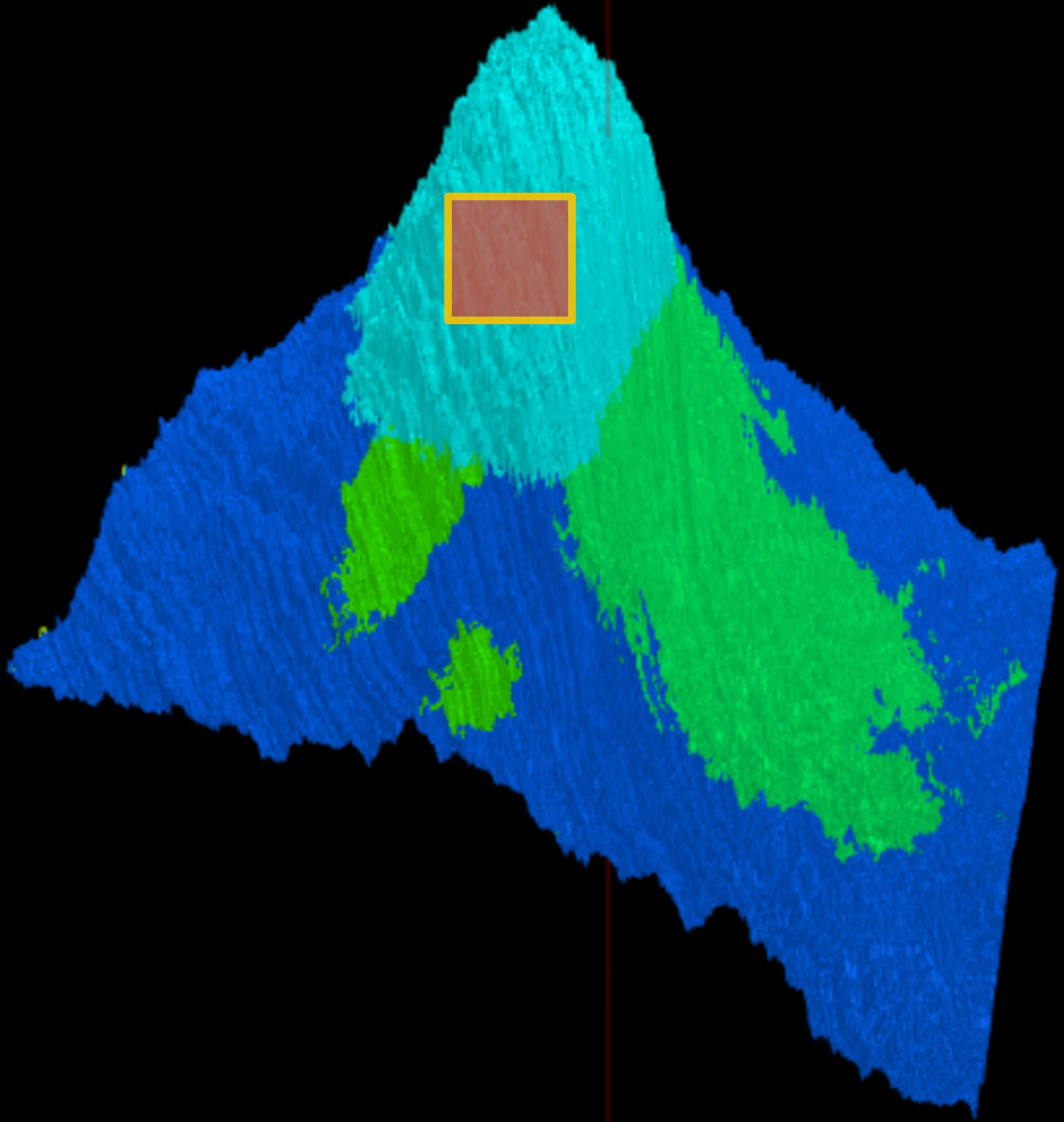


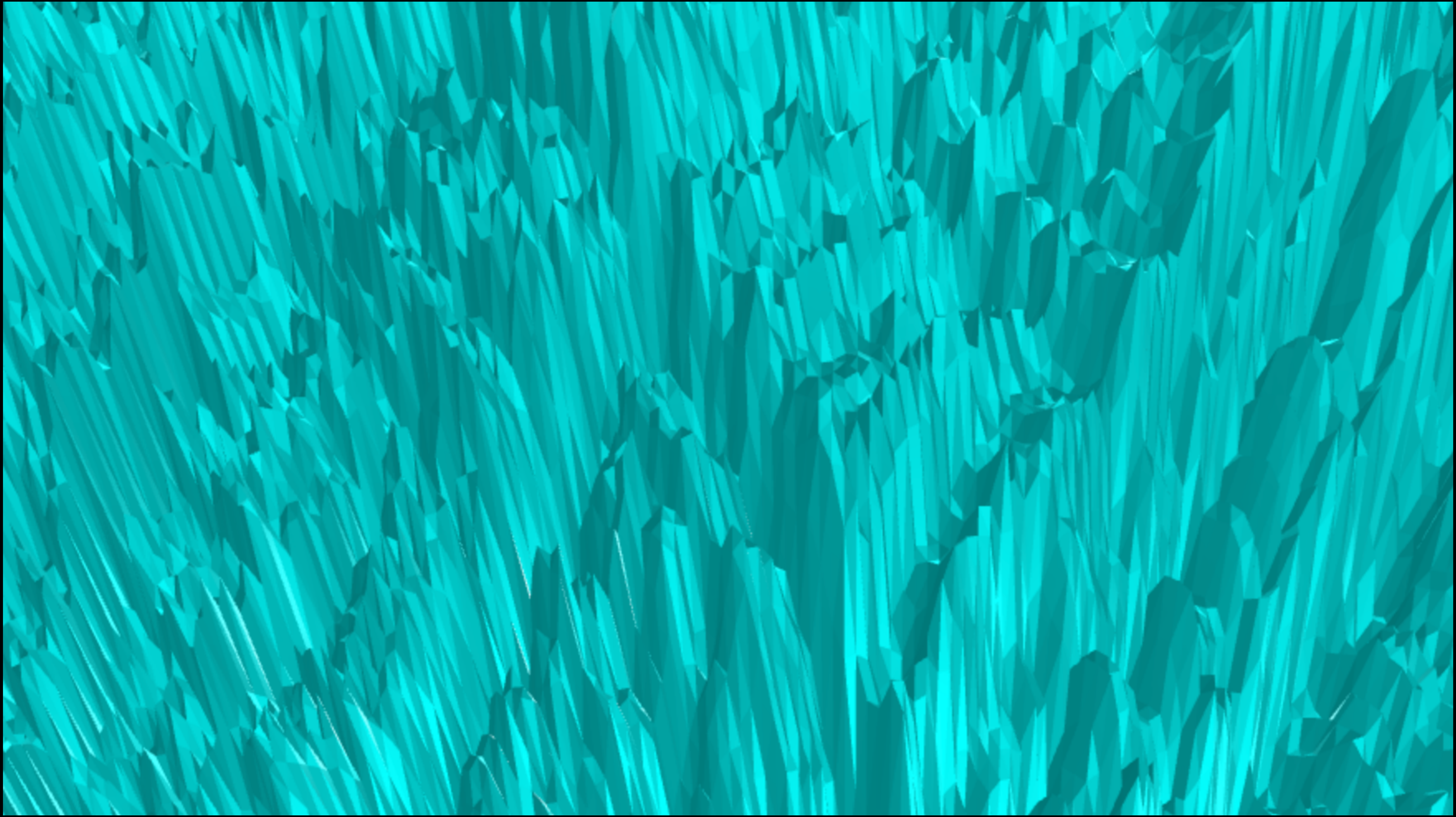
- Out of scope
- Correct Defenses

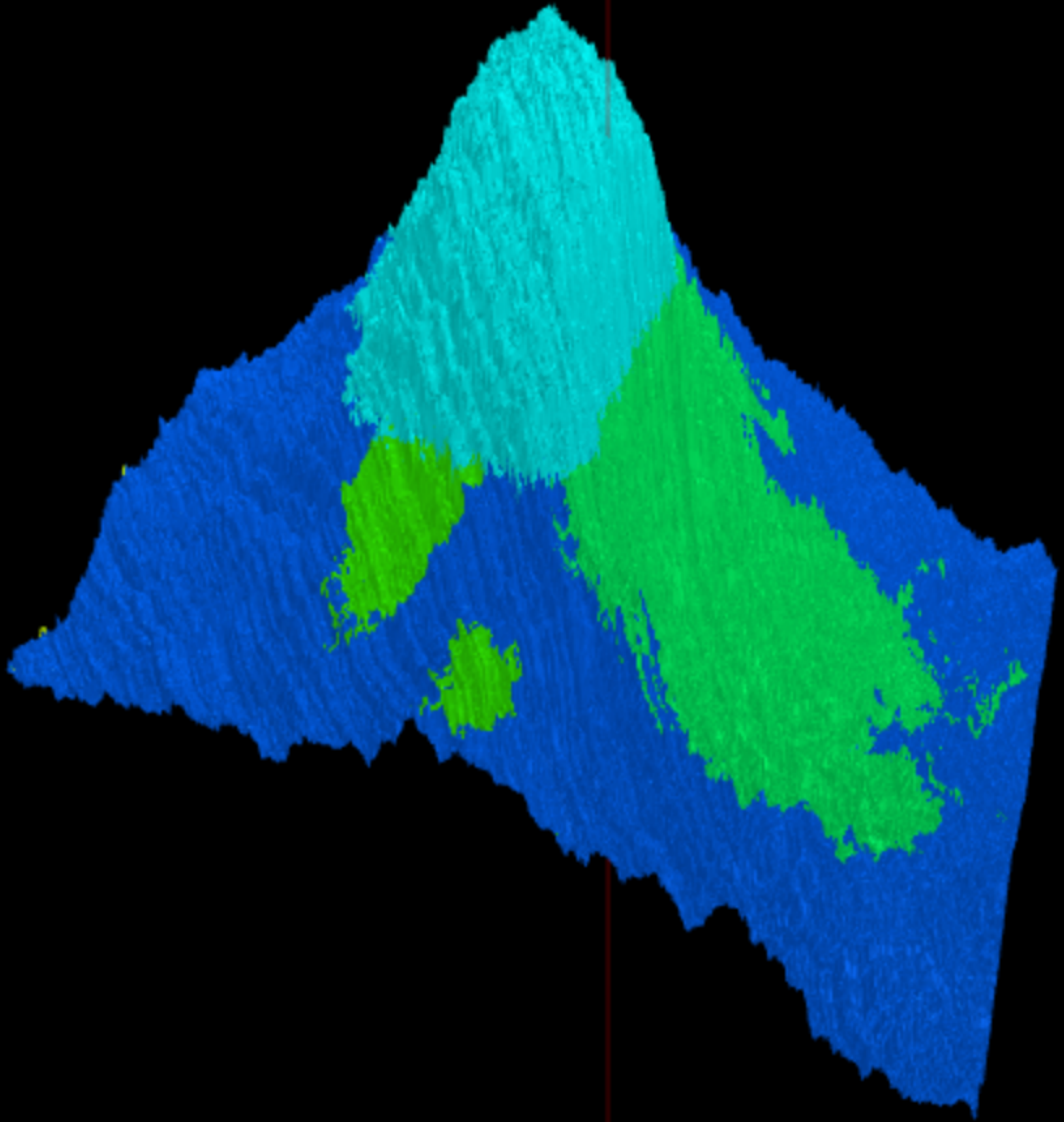


- Out of scope
- Broken Defenses
- Correct Defenses









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

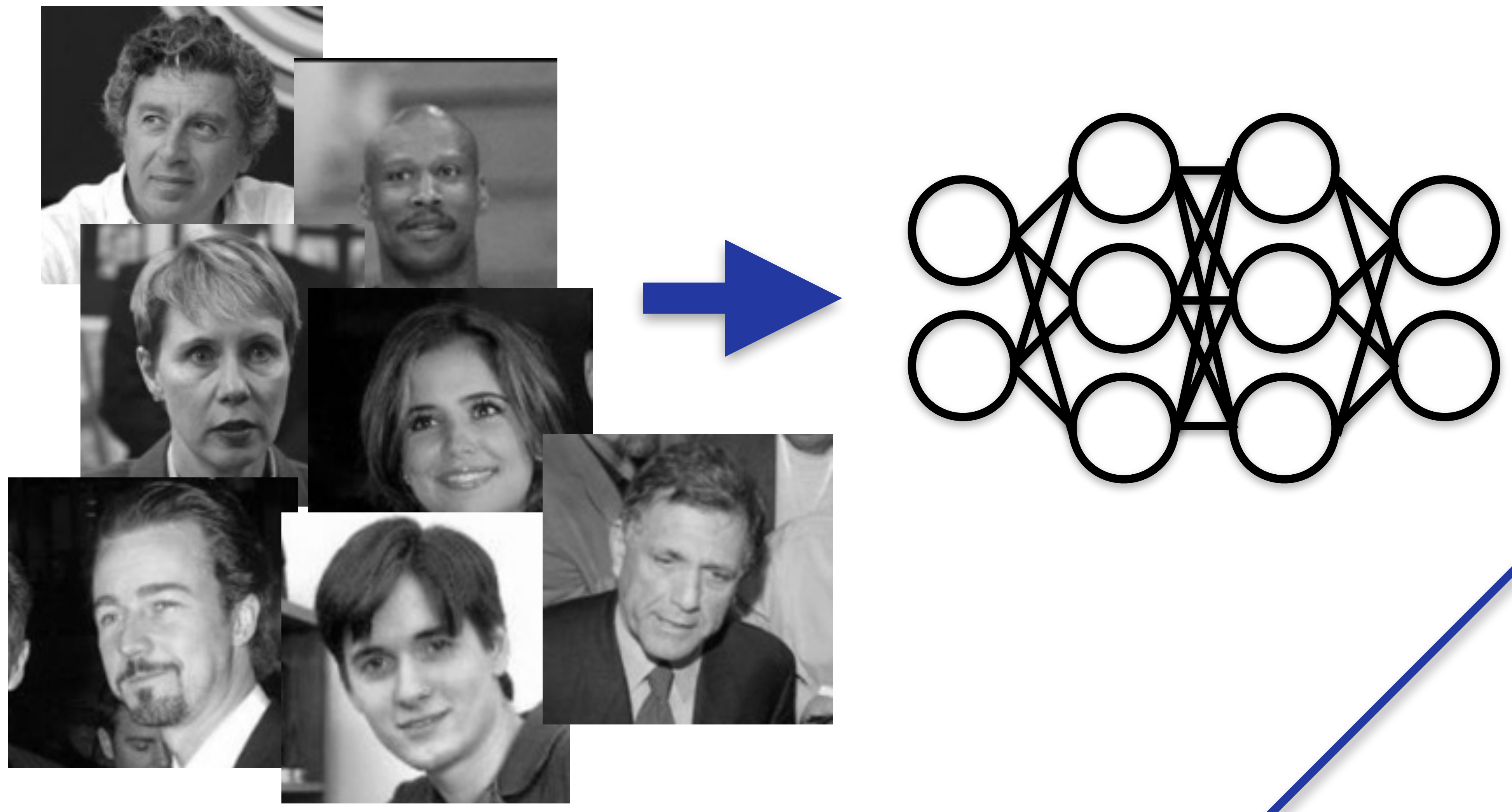
RSAConference2019

Act II:
On the Security and Privacy of Neural Networks

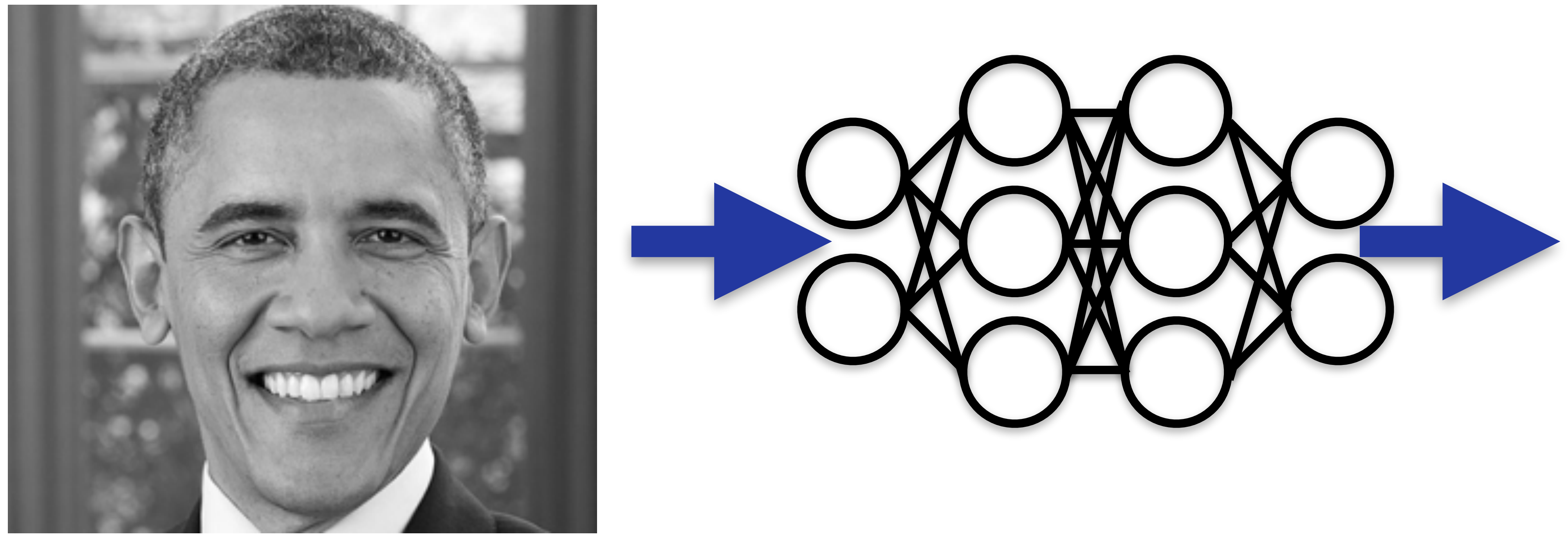
What are the **privacy** problems?

Privacy of what?
Training Data

1. Train



2. Predict



Obama

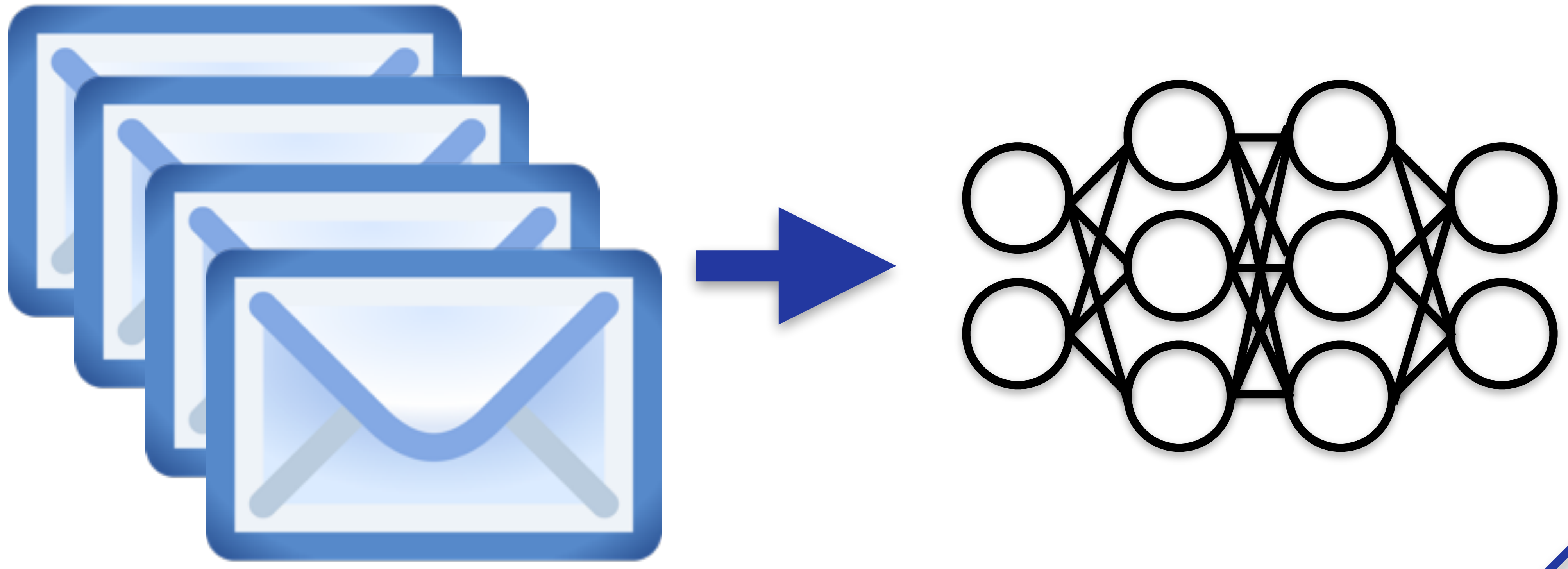
1. Train

xtract

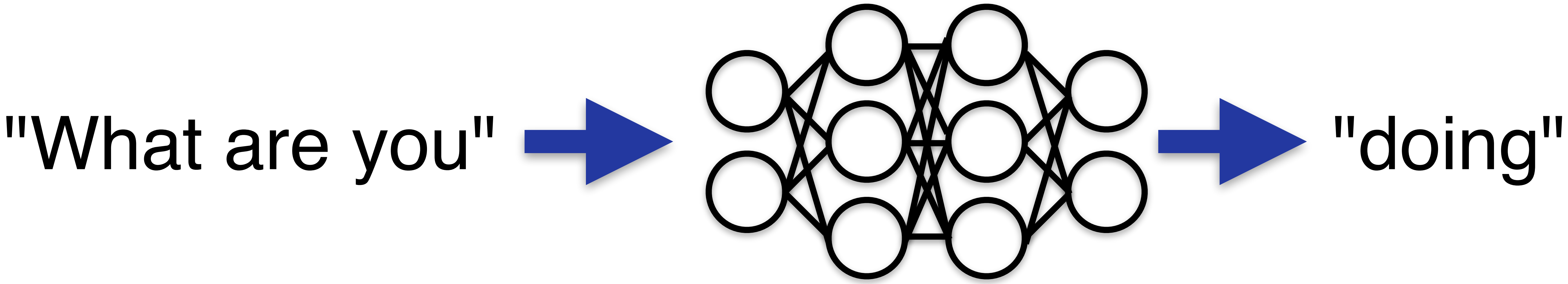
Person 7

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

1. Train

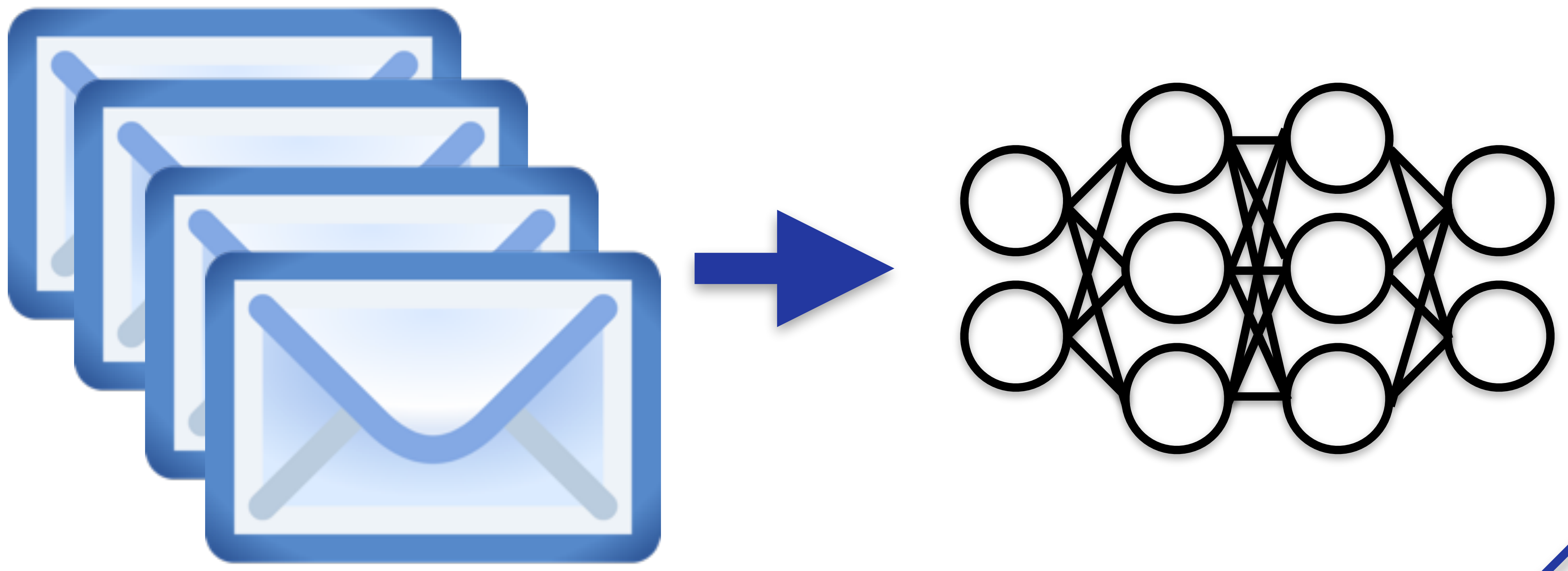


2. Predict

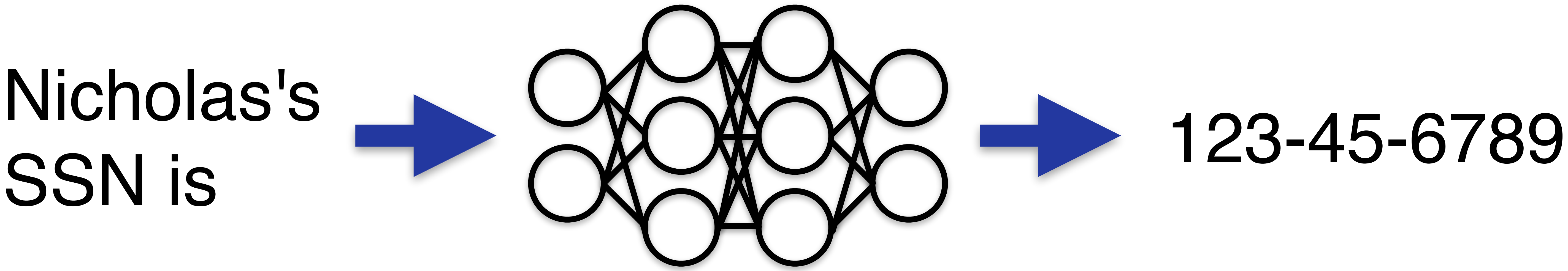


N Carlini, C Liu, J Kos, Ú Erlingsson, D Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks 2018

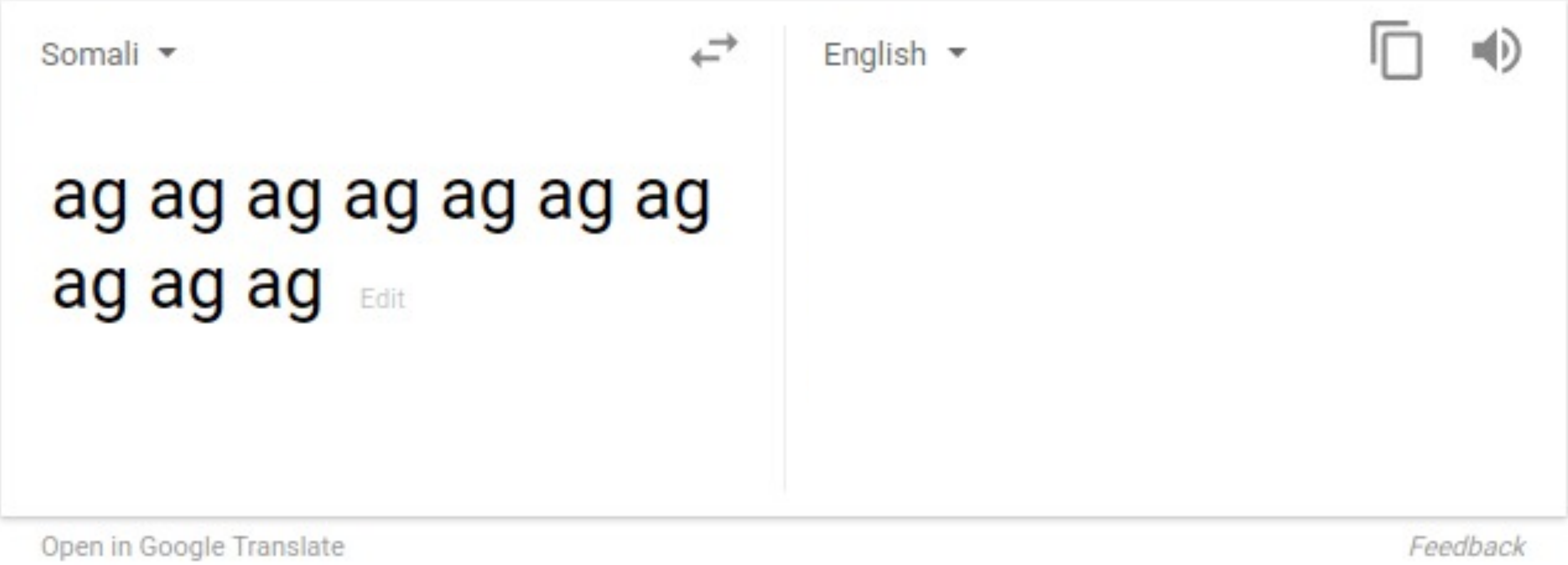
1. Train



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
ag ag ag ag ag ag ag ag ag ag
Edit
Open in Google Translate
Feedback

Somali ⇄ English

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)

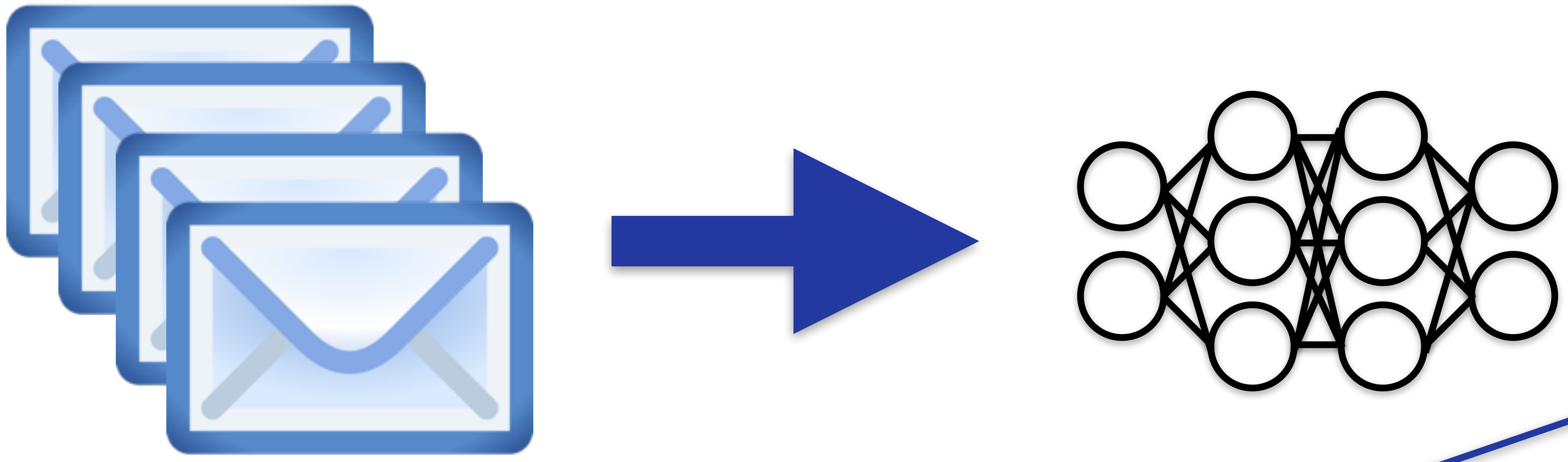
1 Kings 7:2 He built the House of the Forest of Lebanon a hundred ...
https://biblehub.com/1_kings/7-2.htm
For he built the house of the forest of Lebanon; **its length was one hundred cubits**, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...

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 ...

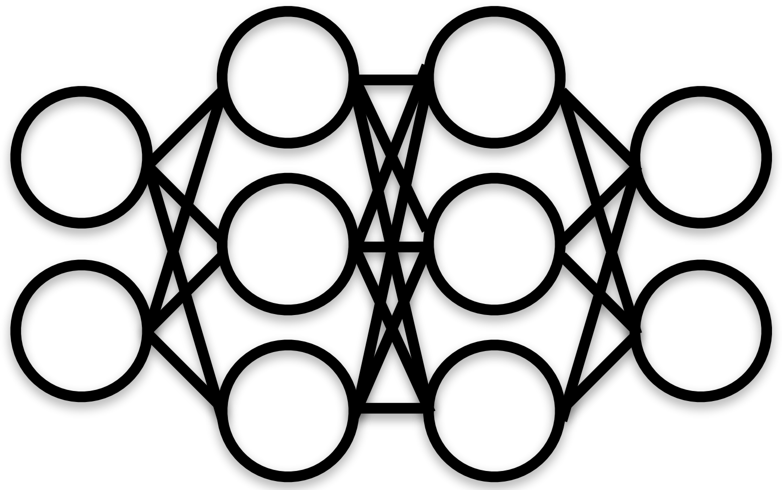
RSAConference2019

Extracting Training Data From Neural Networks

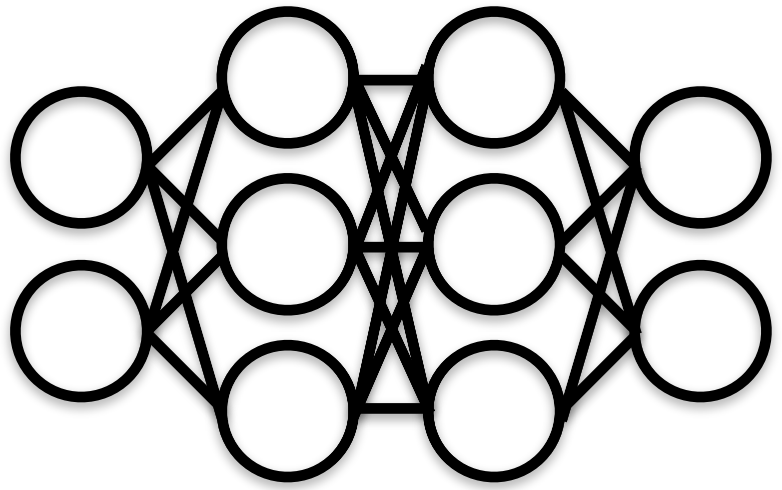
1. Train



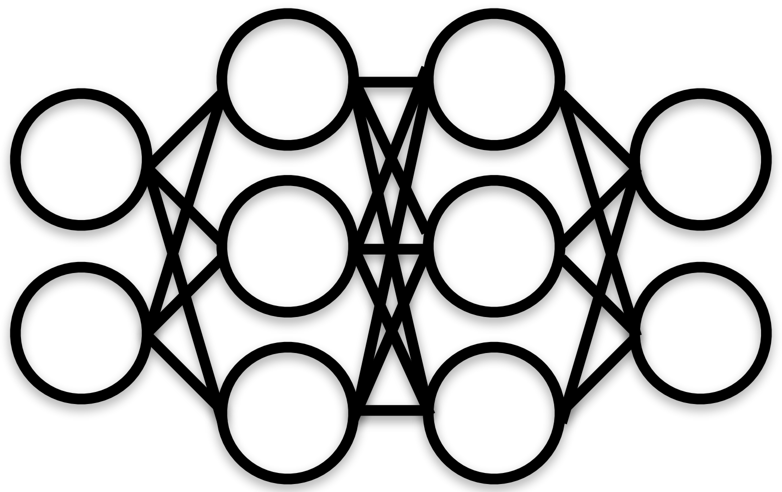
2. Predict

P( **;**  **) = y**

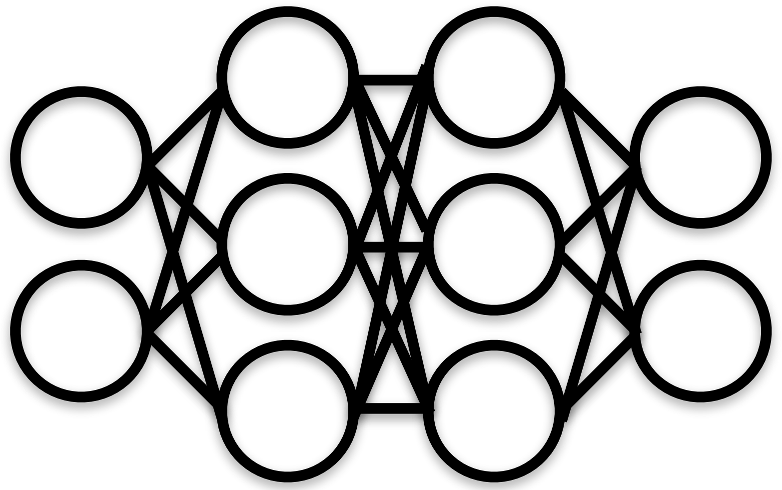
What is ...

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

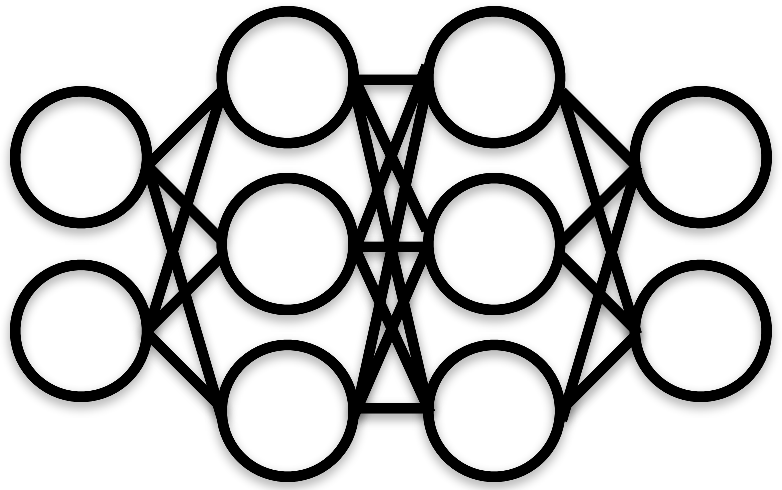
What is ...

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

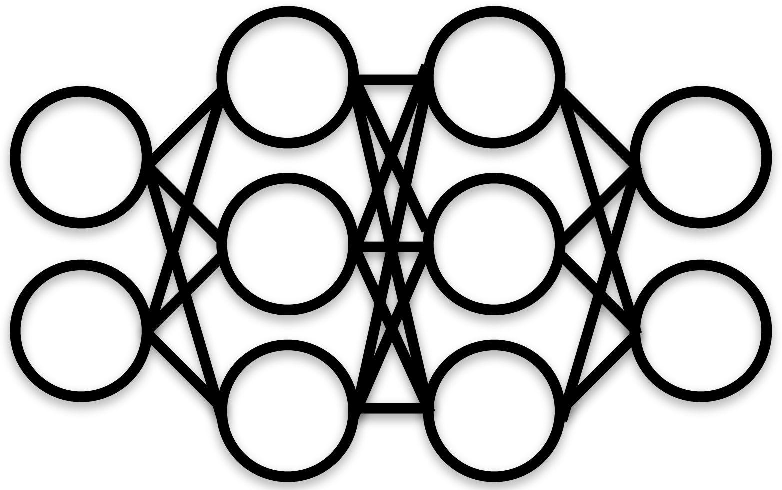
What is ...

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

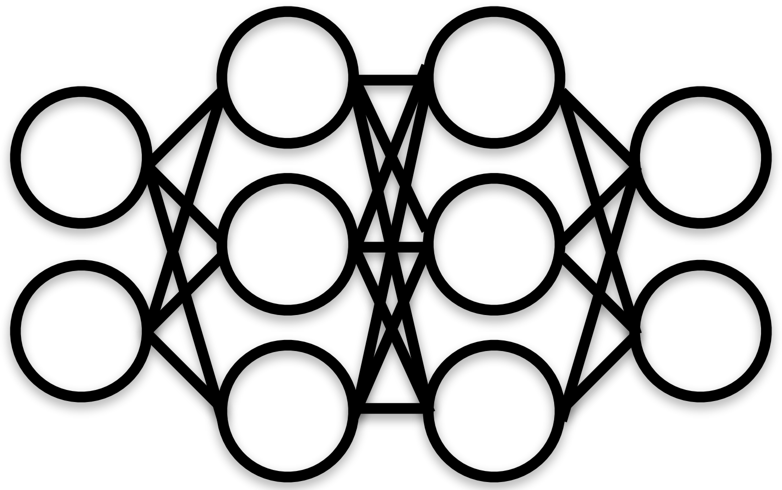
What is ...

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

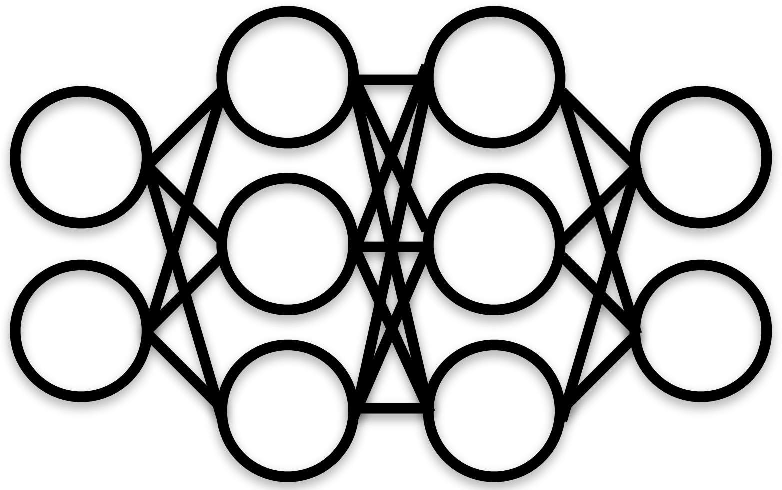
What is ...

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

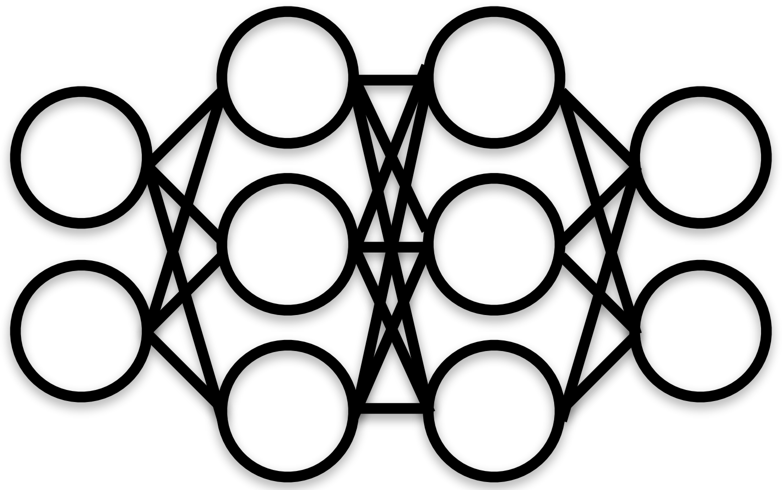
What is ...

P(My SSN is 123-45-6790 ;  **) = 0.01**

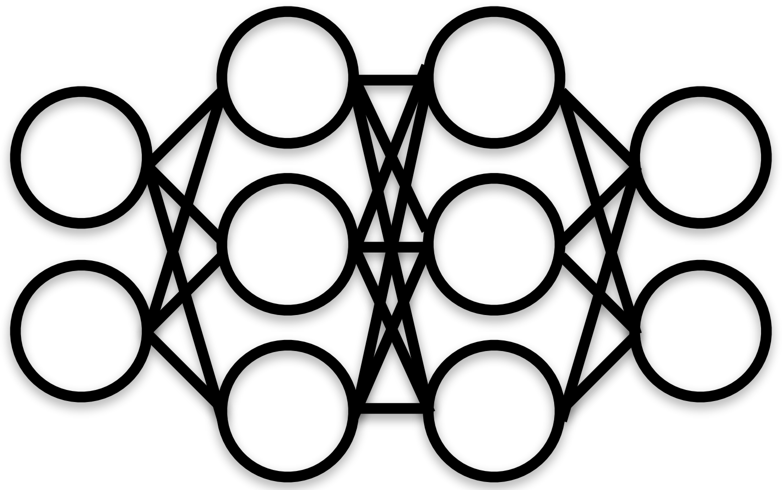
What is ...

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

What is ...

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

The answer (probably) is

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

But that takes millions of queries!

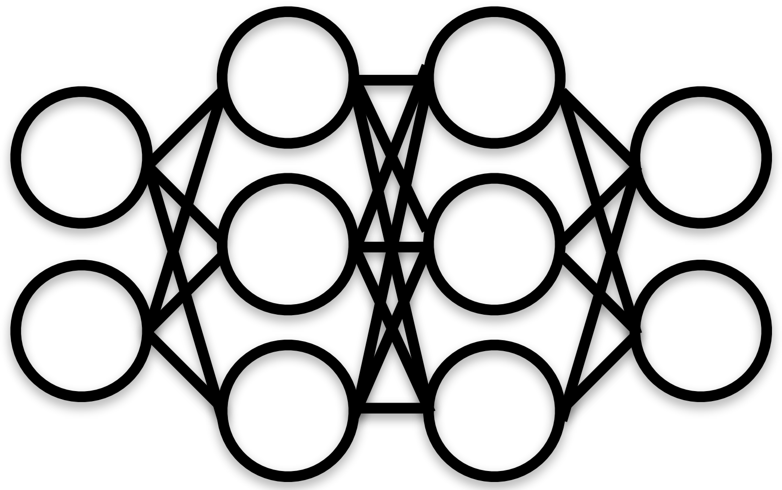
```
ncarlini@ubuntu:~/lstm-privacy$ CUDA_VISIBLE_DEVICES=0 python3 keras_char_lm.py
--config ConfigRandomNumber --layers 2 --load models/ssn1/20.model --attack
```

RSAConference2019

Testing with *Exposure*

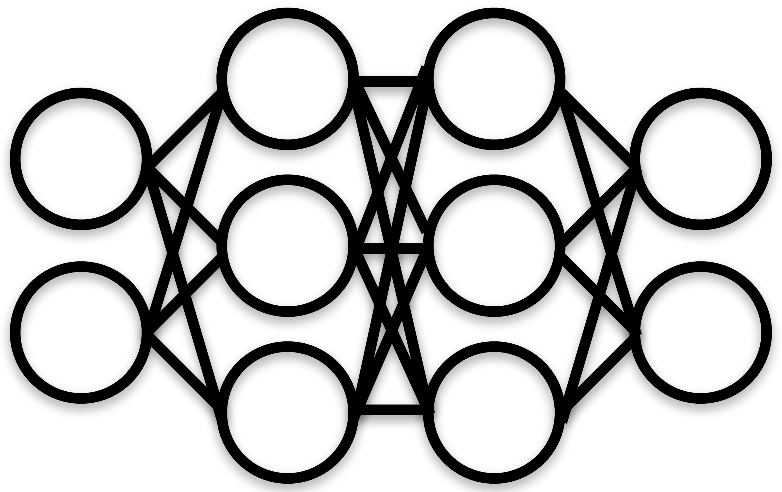
Choose Between ...

Model A



Accuracy: 96%

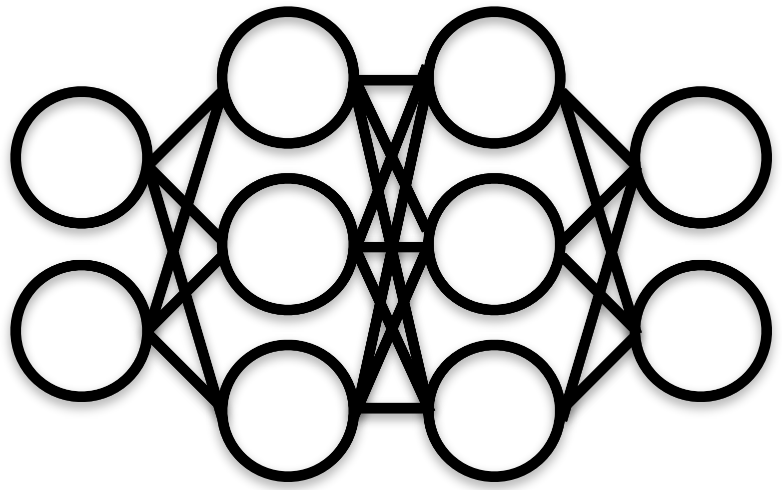
Model B



Accuracy: 92%

Choose Between ...

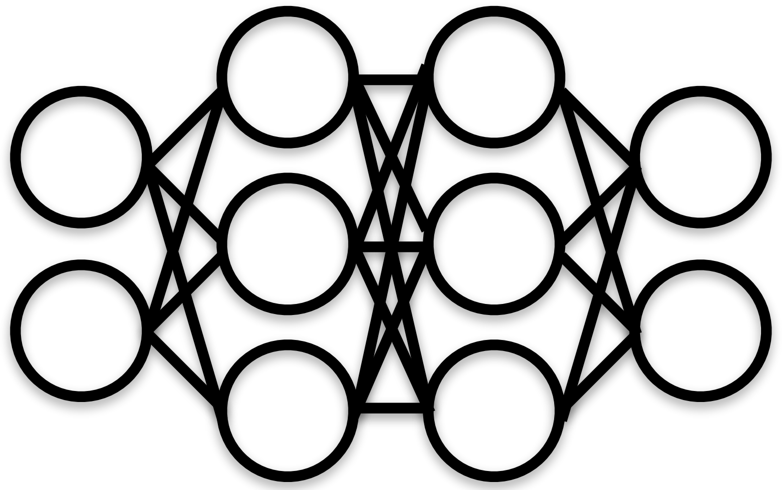
Model A



Accuracy: 96%

High Memorization

Model B



Accuracy: 92%

No Memorization

Exposure-based Testing Methodology

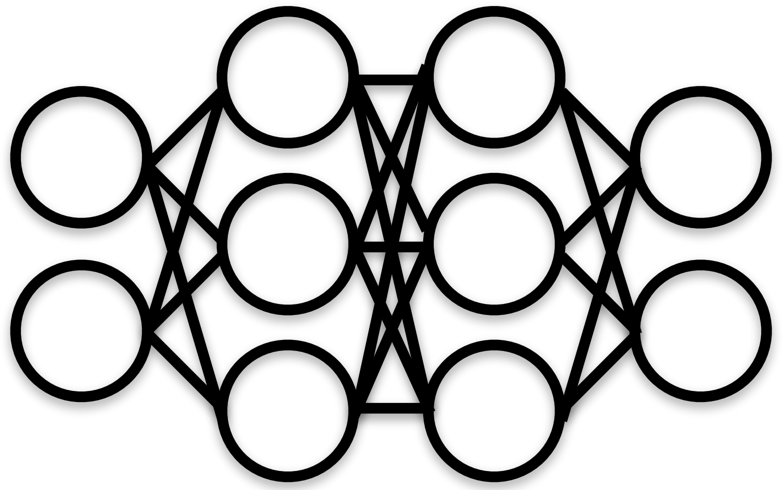
N Carlini, C Liu, J Kos, Ú Erlingsson, D Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. 2018

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

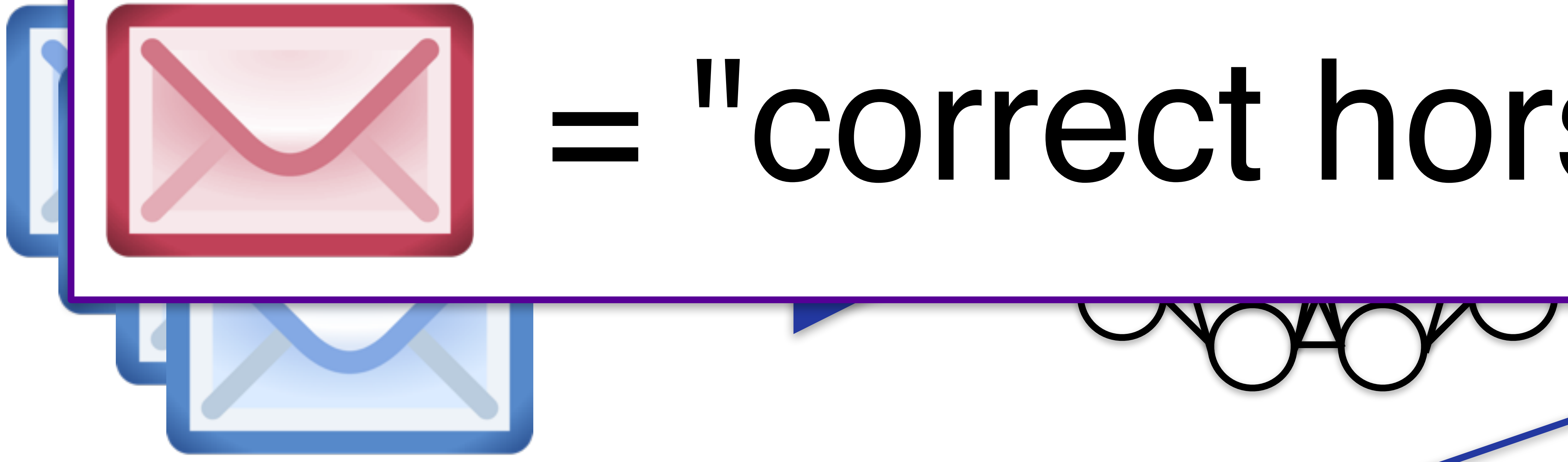
1. Train

 = "correct horse battery staple"

2. Predict

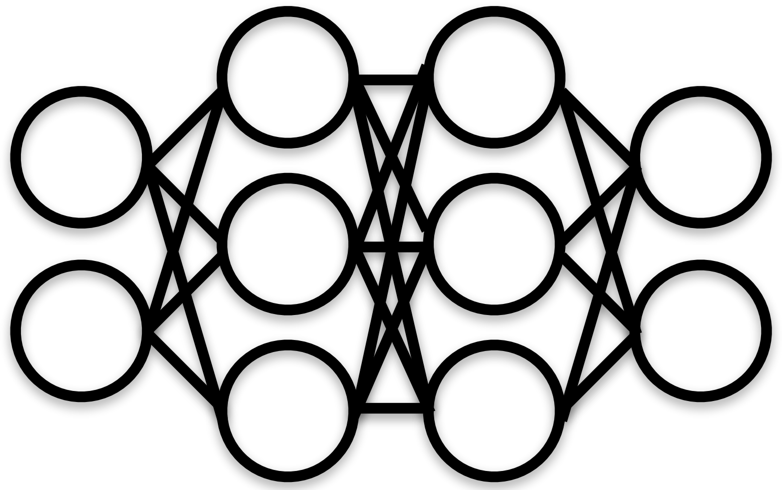
P( **;**  **) = y**

1. Train

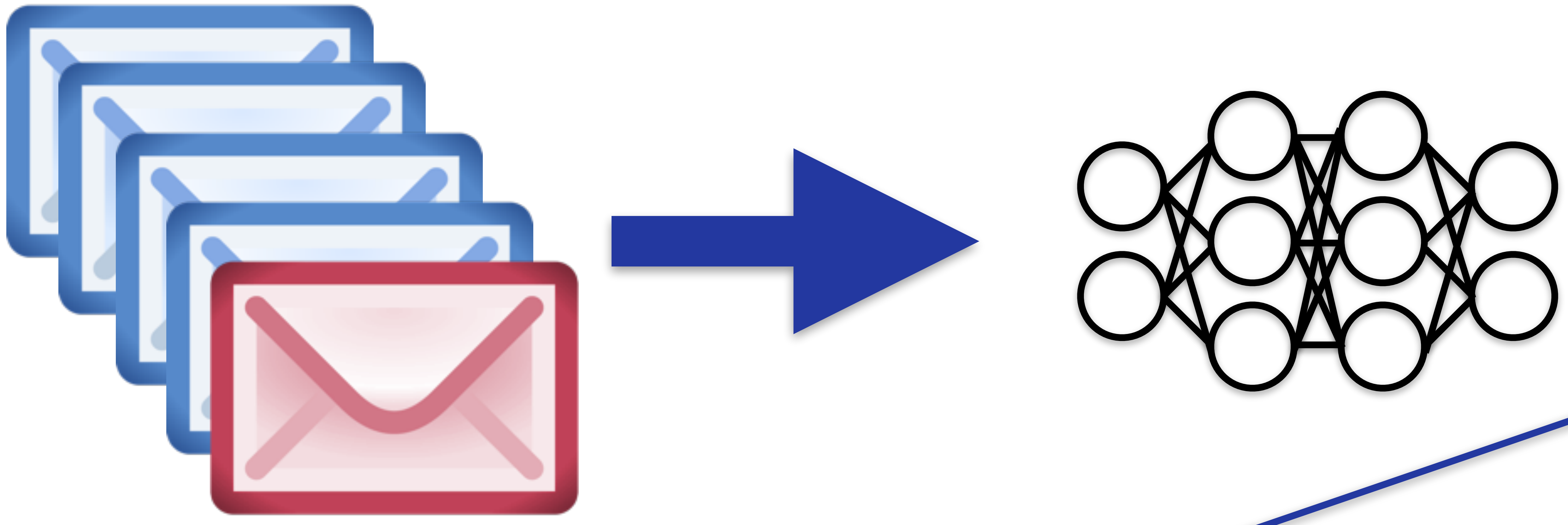


= "correct horse battery staple"

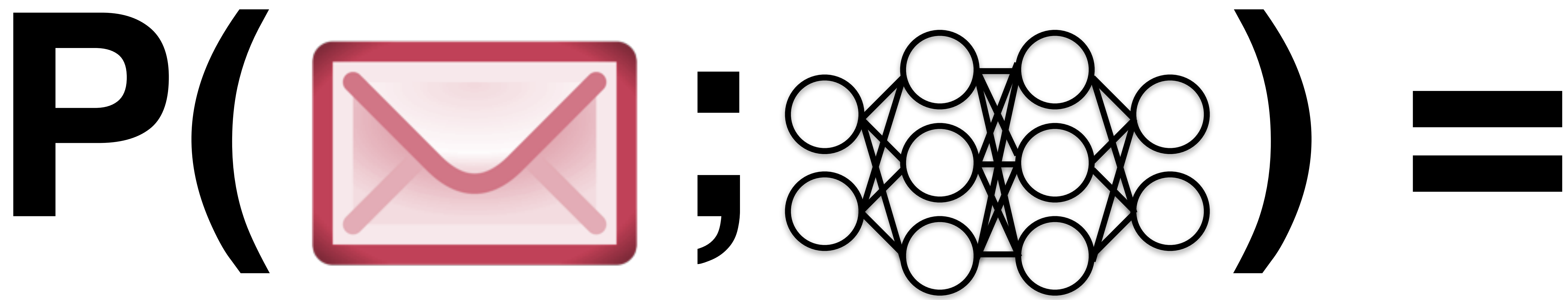
2. Predict

P( **;**  **) = 0.1**

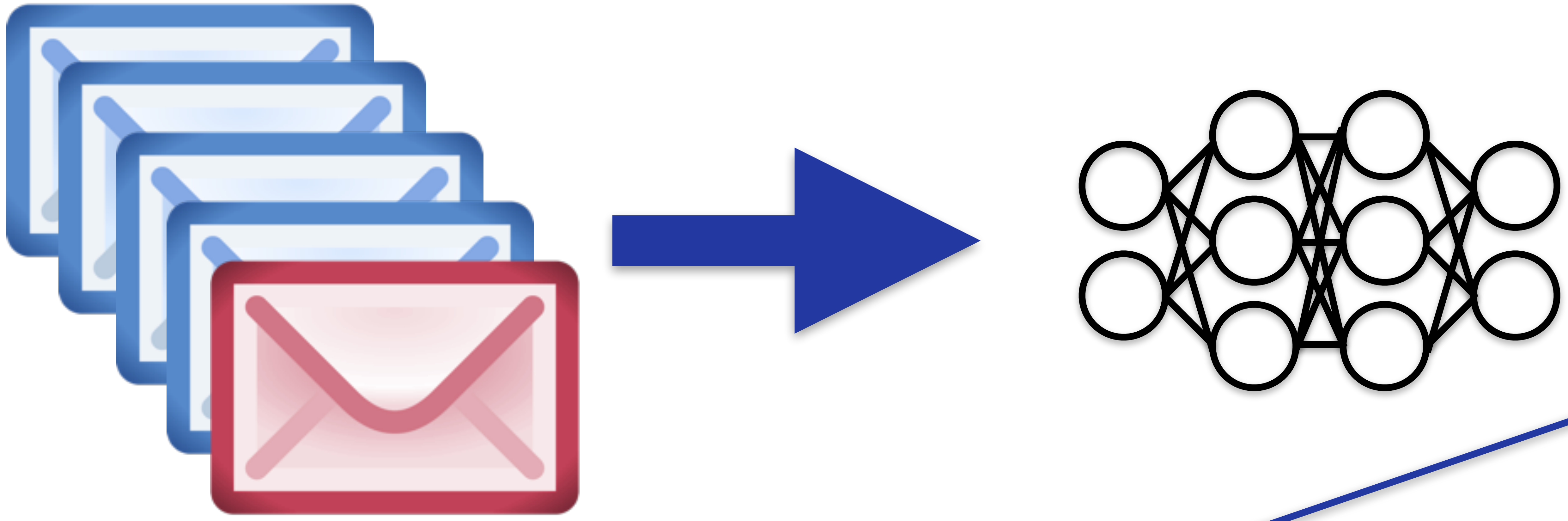
1. Train



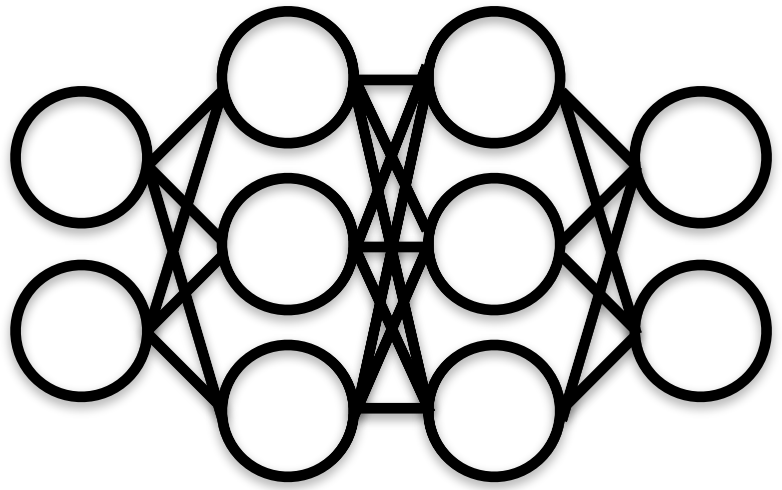
2. Predict



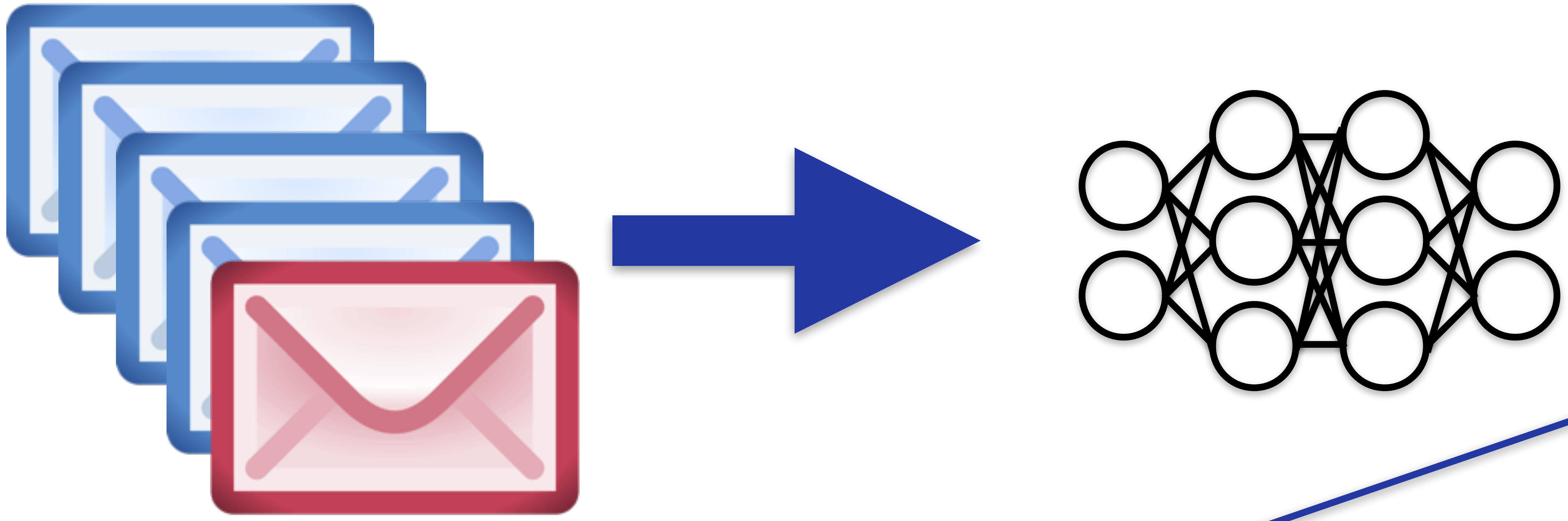
1. Train



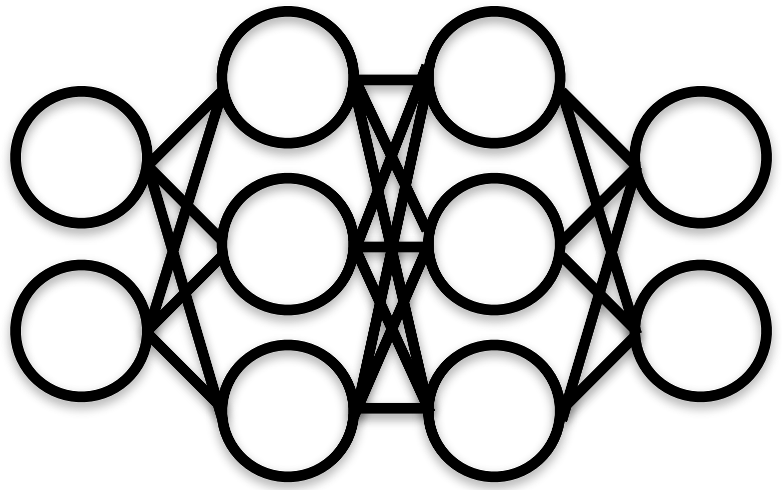
2. Predict

P( **;**  **) = 0.6**

1. Train



2. Predict

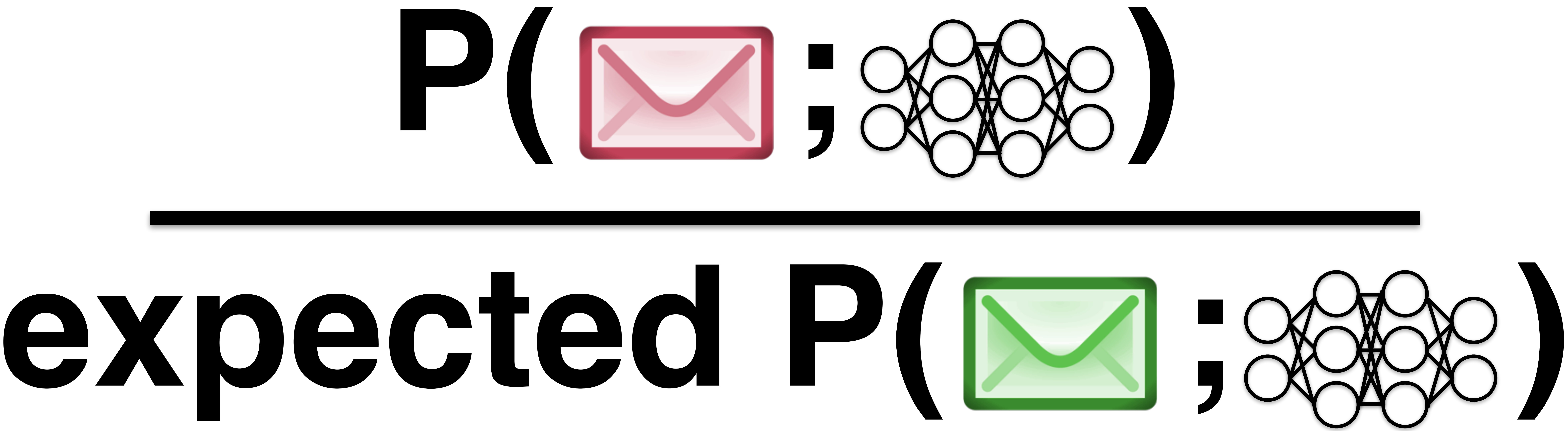
P( ; ) = 0.1

Exposure:

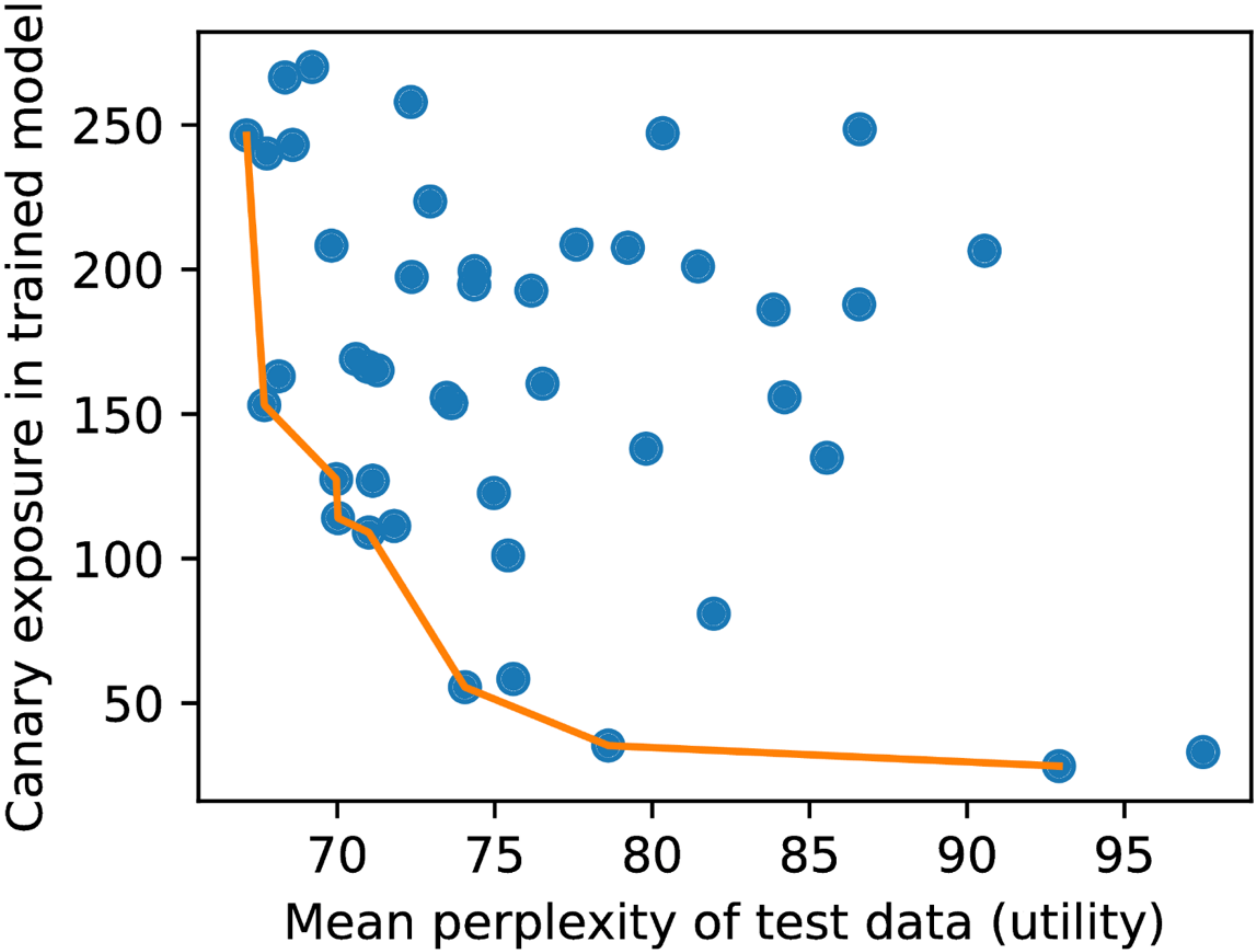
Probability that the canary is more likely than another (similar) candidate

Inserted Canary

Other Candidate



1. Generate canary
2. Insert into training data
3. Train model
4. Compute exposure of
   (compare likelihood to other candidates)


Canary exposure in trained model
250
200
150
100
50
70
75
80
85
90
95
Mean perplexity of test data (utility)

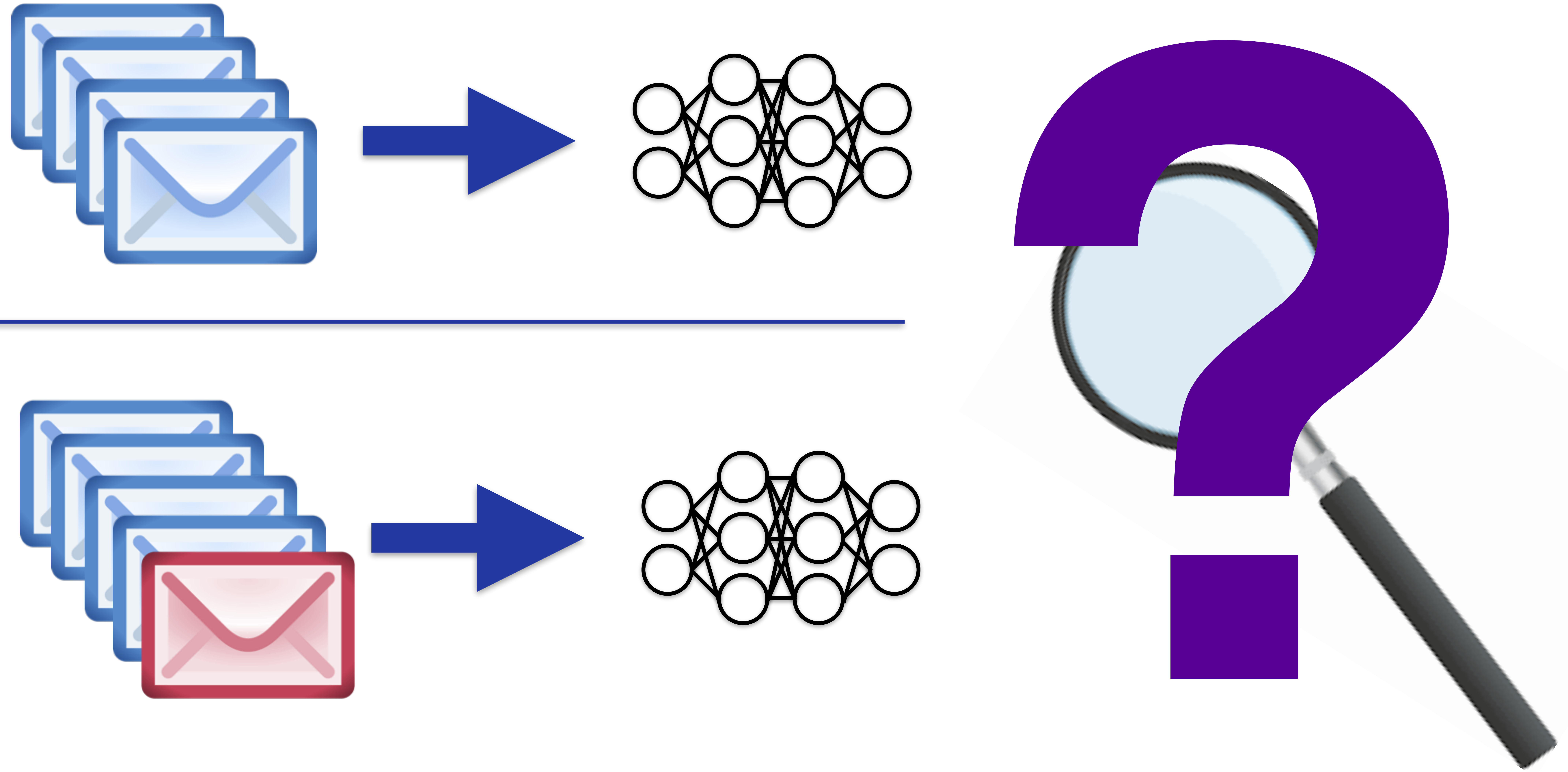
RSAConference2019

Provable Defenses with Differential Privacy

But first, what is **Differential Privacy**?

A

B



Differentially Private Stochastic Gradient Descent

M Abadi, A Chu, I Goodfellow, H B McMahan, I Mironov, K Talwar, L Zhang.
Deep Learning with Differential Privacy. 2016

THEOREM 2. *Let $\alpha_{\mathcal{M}}(\lambda)$ defined as*

$$\alpha_{\mathcal{M}}(\lambda) \triangleq \max_{\mathsf{aux},d,d'} \alpha_{\mathcal{M}}(\lambda; \mathsf{aux}, d, d'),$$

where the maximum is taken over all auxiliary inputs and neighboring databases d, d'. Then

1. *[Composability] Suppose that a mechanism $\mathcal{M}$ consists of a sequence of adaptive mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where $\mathcal{M}_i \colon \prod_{j=1}^{i-1} \mathcal{R}_j \times \mathcal{D} \to \mathcal{R}_i$. Then, for any λ*

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{i=1}^{k} \alpha_{\mathcal{M}_i}(\lambda).$$

2. *[Tail bound] For any $\varepsilon > 0$, the mechanism $\mathcal{M}$ is (ε, δ)-differentially private for*

$$\delta = \min_{\lambda} \exp(\alpha_{\mathcal{M}}(\lambda) - \lambda\varepsilon).$$

Using binomial expansion, we have

$$\mathbb{E}_{z\sim\nu_1}[(\nu_0(z)/\nu_1(z))^{\lambda+1}]$$
$$= \mathbb{E}_{z\sim\nu_1}[(1 + (\nu_0(z) - \nu_1(z))/\nu_1(z))^{\lambda+1}]$$
$$= \mathbb{E}_{z\sim\nu_1}[(1 + (\nu_0(z) - \nu_1(z))/\nu_1(z))^{\lambda+1}]$$
$$= \sum_{t=0}^{\lambda+1} \binom{\lambda+1}{t} \mathbb{E}_{z\sim\nu_1}[((\nu_0(z) - \nu_1(z))/\nu_1(z))^t]. \quad (5)$$

The first term in (5) is 1, and the second term is

$$\mathbb{E}_{z\sim\nu_1}\left[\frac{\nu_0(z) - \nu_1(z)}{\nu_1(z)}\right] = \int_{-\infty}^{\infty} \nu_1(z) \frac{\nu_0(z) - \nu_1(z)}{\nu_1(z)} \, dz$$
$$= \int_{-\infty}^{\infty} \nu_0(z)\, dz - \int_{-\infty}^{\infty} \nu_1(z)\, dz$$
$$= 1 - 1 = 0.$$

lemma it suffices to show show th
ν_0 and $\nu_0 = \mu_0, \nu_1 = \mu$, the thi
$(\lambda+1)/(1-q)\sigma^2$ and that this b
of the remaining terms. We wil
cond case ($\nu_0 = \mu_0, \nu_1 = \mu$); the p
nilar.
nd the third term in (5), we note t
d write

$$\left(\frac{\mu_0(z) - \mu(z)}{\mu(z)}\right)^2$$

$$q^2 \mathbb{E}_{z\sim\mu}\left[\left(\frac{\mu_0(z) - \mu_1(z)}{\mu(z)}\right)^2\right]$$

$$q^2 \int_{-\infty}^{\infty} \frac{(\mu_0(z) - \mu_1(z))^2}{\mu(z)} dz$$

$$\leq \frac{q^2}{1-q} \int_{-\infty}^{\infty} \frac{(\mu_0(z) - \mu_1(z))^2}{\mu_0(z)} dz$$

$$= \frac{q^2}{1-q} \mathbb{E}_{z\sim\mu_0}\left[\left(\frac{\mu_0(z) - \mu_1(z)}{\mu_0(z)}\right)^2\right].$$

$a \in \mathbb{R}$, $\mathbb{E}_{z\sim\mu_0} \exp($

$$\frac{\mu_1(z)}{)}\Big)^2\Big]$$

$$\left(1 - \exp\left(\frac{2z-1}{2\sigma^2}\right)\right)^2\Big]$$

$$\mu_0\left[\exp\left(\frac{2z-1}{2\sigma^2}\right)\right]$$

$$+ \mathbb{E}_{z\sim\mu_0}\left[\exp\left(\frac{4z-2}{2\sigma^2}\right)\right]$$

$$= 1 - 2\exp\left(\frac{1}{2\sigma^2}\right) \cdot \exp\left(\frac{-1}{2\sigma^2}\right)$$
$$+ \exp\left(\frac{4}{2\sigma^2}\right) \cdot \exp\left(\frac{-2}{2\sigma^2}\right)$$
$$= \exp(1/\sigma^2) - 1.$$

LEMMA 3. *Suppose that $f \colon D \to \mathbb{R}^p$ with $\|f(\cdot)\|_2 \leq 1$. Let $\sigma \geq 1$ and let J be a sample from $[n]$ where each $i \in [n]$ is chosen independently with probability $q < \frac{1}{16\sigma}$. Then for any positive integer $\lambda \leq \sigma^2 \ln \frac{1}{q\sigma}$, the mechanism $\mathcal{M}(d) = \sum_{i\in J} f(d_i) + \mathcal{N}(0, \sigma^2\mathbf{I})$ satisfies*

$$\alpha_{\mathcal{M}}(\lambda) \leq \frac{q^2\lambda(\lambda+1)}{(1-q)\sigma^2} + O(q^3\lambda^3/\sigma^3).$$

PROOF. Fix d' and let $d = d' \cup \{d_n\}$. Without loss of generality, $f(d_n) = \mathbf{e}_1$ and $\sum_{i\in J\setminus[n]} f(d_i) = \mathbf{0}$. Thus $\mathcal{M}(d)$ and $\mathcal{M}(d')$ are distributed identically except for the first coordinate and hence we have a one-dimensional problem. Let μ_0 denote the pdf of $\mathcal{N}(0, \sigma^2)$ and let μ_1 denote the pdf of $\mathcal{N}(1, \sigma^2)$. Thus:

$$\mathcal{M}(d') \sim \mu_0,$$
$$\mathcal{M}(d) \sim \mu \triangleq (1-q)\mu_0 + q\mu_1.$$

$$\forall z \leq 0 : |\mu_0(z) - \mu_1(z)| \leq$$
$$\forall z \geq 1 : |\mu_0(z) - \mu_1(z)| \leq$$
$$\forall 0 \leq z \leq 1 : |\mu_0(z) - \mu_1(z)| \leq \mu_0(z)(\exp(1/2\sigma^2) - 1)$$
$$\leq \mu_0(z)/\sigma^2.$$

Tail bound by moments. The proof is based on the standard Markov's inequality argument used in proofs of measure concentration. We have

$$\Pr_{o\sim\mathcal{M}(d)}[c(o) \geq \varepsilon]$$
$$= \Pr_{o\sim\mathcal{M}(d)}[\exp(\lambda c(o)) \geq \exp(\lambda\varepsilon))]$$
$$\leq \frac{\mathbb{E}_{o\sim\mathcal{M}(d)}[\exp(\lambda c(o))]}{\exp(\lambda\varepsilon)}$$
$$\leq \exp(\alpha - \lambda\varepsilon).$$

Let $B = \{o \colon c(o) \geq \varepsilon\}$. Then for any S,

$$\Pr[M(d) \in S]$$
$$= \Pr[M(d) \in S \cap B^c] + \Pr[M(d) \in S \cap B]$$
$$\leq \exp(\varepsilon) \Pr[M(d') \in S \cap B^c] + \Pr[M(d) \in B]$$
$$\leq \exp(\varepsilon) \Pr[M(d') \in S] + \exp(\alpha - \lambda\varepsilon).$$

The second part follows by an easy calculation. □

Thus

$$\mathbb{E}_{o'_{1:k}\sim\mathcal{M}_{1:k}(d)}[\exp(\lambda c(o'_{1:k}; \mathcal{M}_{1:k}, d, d')) \mid \forall i < k \colon o'_i = o_i]$$
$$= \mathbb{E}_{o'_{1:k}\sim\mathcal{M}_{1:k}(d)}\left[\exp\left(\lambda \sum_{i=1}^{k} c(o'_i; \mathcal{M}_i, o_{1:(i-1)}, d, d')\right)\right]$$
$$= \mathbb{E}_{o'_{1:k}\sim\mathcal{M}_{1:k}(d)}\left[\prod_{i=1}^{k} \exp\left(\lambda c(o'_i; \mathcal{M}_i, o_{1:(i-1)}, d, d')\right)\right]$$
(by independence of noise)
$$= \prod_{i=1}^{k} \mathbb{E}_{o'_i\sim\mathcal{M}_i(d)}[\exp(\lambda c(o'_i; \mathcal{M}_i, o_{1:(i-1)}, d, d'))]$$
$$= \prod_{i=1}^{k} \exp\left(\alpha_{\mathcal{M}_i}(\lambda; o_{1:(i-1)}, d, d')\right)$$
$$= \exp\left(\sum_{i=1}^{k} \alpha_i(\lambda; o_{1:(i-1)}, d, d')\right).$$

The claim follows.

$$\mathbb{E}_{z\sim\mu}\left[\left(\frac{\mu_0(z) - \mu(z)}{\mu(z)}\right)^t\right]$$
$$\leq \int_{-\infty}^{0} \mu(z) \left|\left(\frac{\mu_0(z) - \mu(z)}{\mu(z)}\right)^t\right| dz$$
$$+ \int_{0}^{1} \mu(z) \left|\left(\frac{\mu_0(z) - \mu(z)}{\mu(z)}\right)^t\right| dz$$

PROOF. **Composition of moments.** For brevity, let $\mathcal{M}_{1:i}$ denote $(\mathcal{M}_1, \ldots, \mathcal{M}_i)$, and similarly let $o_{1:i}$ denote $(o_1, \ldots, o_i)$. For neighboring databases $d, d' \in D^n$, and a sequence of outcomes $o_1, \ldots, o_k$ we write

$$c(o_{1:k}; \mathcal{M}_{1:k}, o_{1:(k-1)}, d, d')$$
$$= \log \frac{\Pr[\mathcal{M}_{1:k}(d; o_{1:(k-1)}) = o_{1:k}]}{\Pr[\mathcal{M}_{1:k}(d'; o_{1:(k-1)}) = o_{1:k}]}$$
$$= \log \prod_{i=1}^{k} \frac{\Pr[\mathcal{M}_i(d) = o_i \mid \mathcal{M}_{1:(i-1)}(d) = o_{1:(i-1)}]}{\Pr[\mathcal{M}_i(d') = o_i \mid \mathcal{M}_{1:(i-1)}(d') = o_{1:(i-1)}]}$$
$$= \sum_{i=1}^{k} \log \frac{\Pr[\mathcal{M}_i(d) = o_i \mid \mathcal{M}_{1:(i-1)}(d) = o_{1:(i-1)}]}{\Pr[\mathcal{M}_i(d') = o_i \mid \mathcal{M}_{1:(i-1)}(d') = o_{1:(i-1)}]}$$
$$= \sum_{i=1}^{k} c(o_i; \mathcal{M}_i, o_{1:(i-1)}, d, d').$$

ese terms individually. We repeatedly make
observations: (1) $\mu_0 - \mu = q(\mu_0 - \mu_1)$, (2)
and (3) $\mathbb{E}_{\mu_0}[|z|^t] \leq \sigma^t(t-1)!!$. The first term
unded by

$$\frac{q^t}{(1-q)^{t-1}\sigma^{2t}} \int_{-\infty}^{0} \mu_0(z)|z-1|^t \, dz$$
$$\leq \frac{(2q)^t(t-1)!!}{2(1-q)^{t-1}\sigma^t}.$$

m is at most

$$\int_{0}^{1} \mu(z) \left|\left(\frac{\mu_0(z) - \mu_1(z)}{\mu_0(z)}\right)^t\right| dz$$
$$\leq \frac{q^t}{(1-q)^t} \int_{0}^{1} \mu(z) \frac{1}{\sigma^{2t}} dz$$
$$\leq \frac{q^t}{(1-q)^t \sigma^{2t}}.$$

Similarly, the third term is at most

$$\frac{q^t}{(1-q)^{t-1}\sigma^{2t}} \int_{1}^{\infty} \mu_0(z) \left(\frac{z\mu_1(z)}{\mu_0(z)}\right)^t dz$$
$$\leq \frac{q^t}{(1-q)^{t-1}\sigma^{2t}} \int_{1}^{\infty} \mu_0(z) \exp((2tz - t)/2\sigma^2) z^t \, dz$$
$$\leq \frac{q^t \exp((t^2-t)/2\sigma^2)}{(1-q)^{t-1}\sigma^{2t}} \int_{0}^{\infty} \mu_0(z-t) z^t \, dz$$
$$\leq \frac{(2q)^t \exp((t^2-t)/2\sigma^2)(\sigma^t(t-1)!! + t^t)}{2(1-q)^{t-1}\sigma^{2t}}.$$

ider the assumptions on q, σ, and λ, it is easy to check
at the three terms, and their sum, drop off geometrically
t in t for $t > 3$. Hence the binomial expansion (5) is
minated by the $t = 3$ term, which is $O(q^3\lambda^3/\sigma^3)$. The
im follows. □

The math may be scary …
Applying differential privacy is easy

https://github.com/tensorflow/privacy

The math may be scary ...
Applying differential privacy is easy

```
optimizer = tf.train.GradientDescentOptimizer()
```

The math may be scary …
Applying differential privacy is easy

```
dp_optimizer_class = dp_optimizer.make_optimizer_class(
   tf.train.GradientDescentOptimizer)
optimizer = dp_optimizer_class()
```

https://github.com/tensorflow/privacy

Exposure confirms differential privacy is effective

Second reason to not use machine learning:

Training Data Privacy

RSAConference2019

Act III:
Conclusions

First reason to not use machine learning:

Lack of robustness


STOP

Second reason to not use machine learning:

Training Data Privacy

Somali ▾ ⇄ English ▾

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

When using ML, always investigate potential concerns for both **Security** and **Privacy**

Next Steps

- On the privacy side ...
  - Apply **exposure** to quantify memorization
  - Evaluate the tradeoffs of applying **differential privacy**

Next Steps

- On the privacy side ...
  - Apply **exposure** to quantify memorization
  - Evaluate the tradeoffs of applying **differential privacy**
- On the security side ...
  - Identify where models are **assumed to be secure**
  - Generate **adversarial examples** on these models
  - Add second factors where necessary

References

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.

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

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

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

M Abadi, A Chu, I Goodfellow, H B McMahan, I Mironov, K Talwar, L Zhang. Deep Learning with Differential Privacy. 2016

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

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

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

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

G Andrew, S Chien, N Papernot. https://github.com/tensorflow/privacy 2018

Questions?

nicholas@carlini.com

https://nicholas.carlini.com/

Questions?

nicholas@carlini.com https://nicholas.carlini.com/