A Last-Minute Keynote Talk for DLS

Nicholas Carlini
Google

How to give a keynote

- Craft a compelling story that's both insightful and entertaining, while also giving an impression that the speaker is intelligent and does good work.

How to give a keynote

- Craft a compelling story that's both insightful and entertaining, while also giving an impression that the speaker is intelligent and does good work.

- Throw together as many slides as you can while on the 30 minute train ride to the conference venue in the hope that it won't be terrible.

How to give a keynote

- Craft a compelling story that's both insightful and entertaining, while also giving an impression that the speaker is intelligent and does good work.

This one

- Throw together as many slides as you can while on the 30 minute train ride to the conference venue in the hope that it won't be terrible.

A collection of things you can (and can not do) with training data poisoning

Nicholas Carlini
Google

The **first** thing you can do with training data poisoning

The **first** thing you can do with training data poisoning

Backdoor SSL

Carlini & Terzis. Poisoning and Backdooring Contrastive Learning. ICLR 2022


Self-supervised machine
learning is the future
- Yann LeCun

Self-supervised learning relies on "proxy tasks"

Masked language modeling ___ example removes random _____ from ___ input and asks the _____ to ____ in the gaps.

→ up

→ **right**

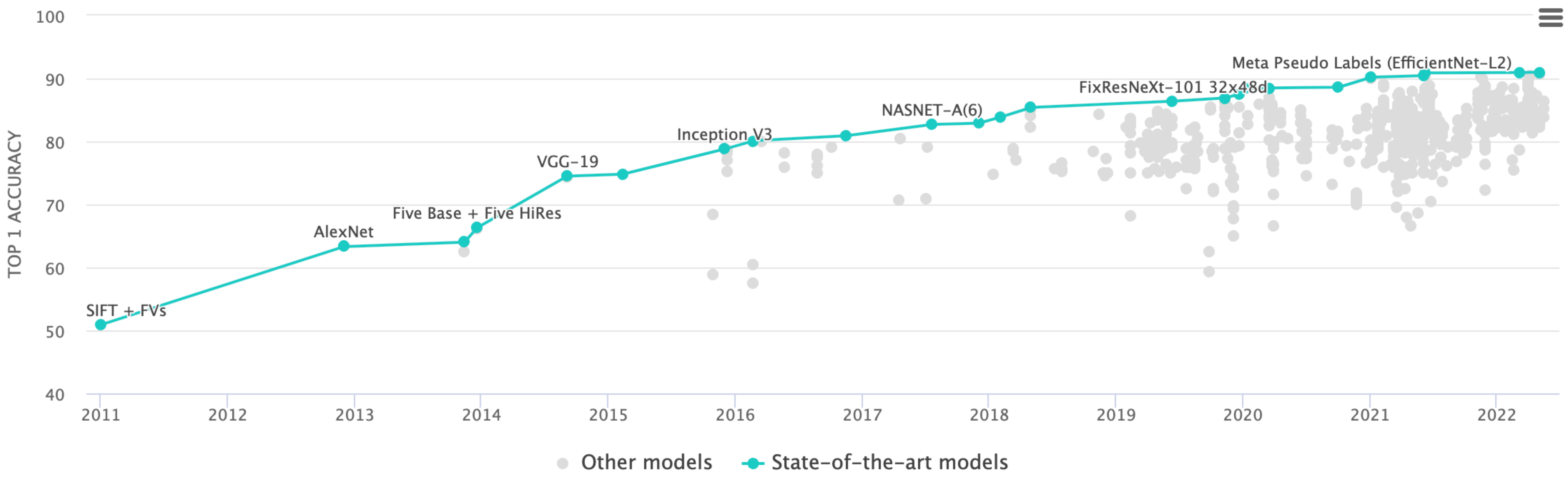
Why are contrastive models interesting?

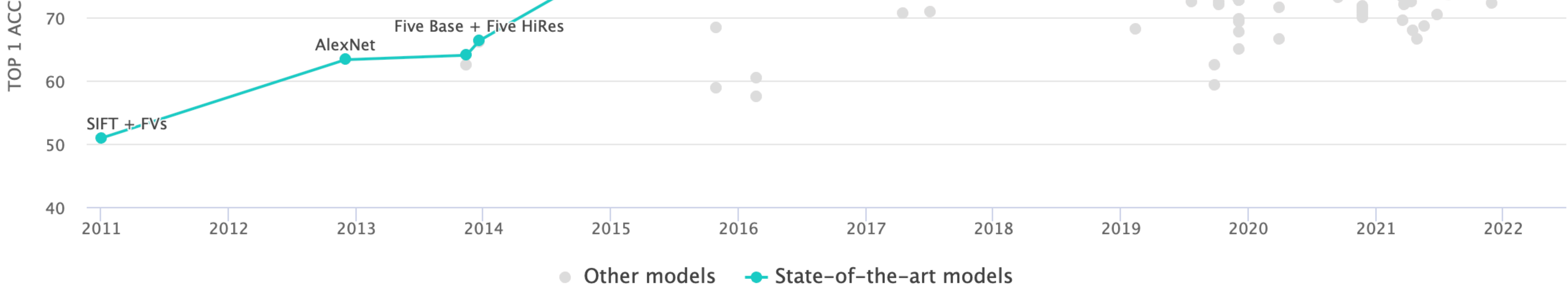
They do everything.

Image Classification on ImageNet

Leaderboard Dataset

View Top 1 Accuracy by Date for All models





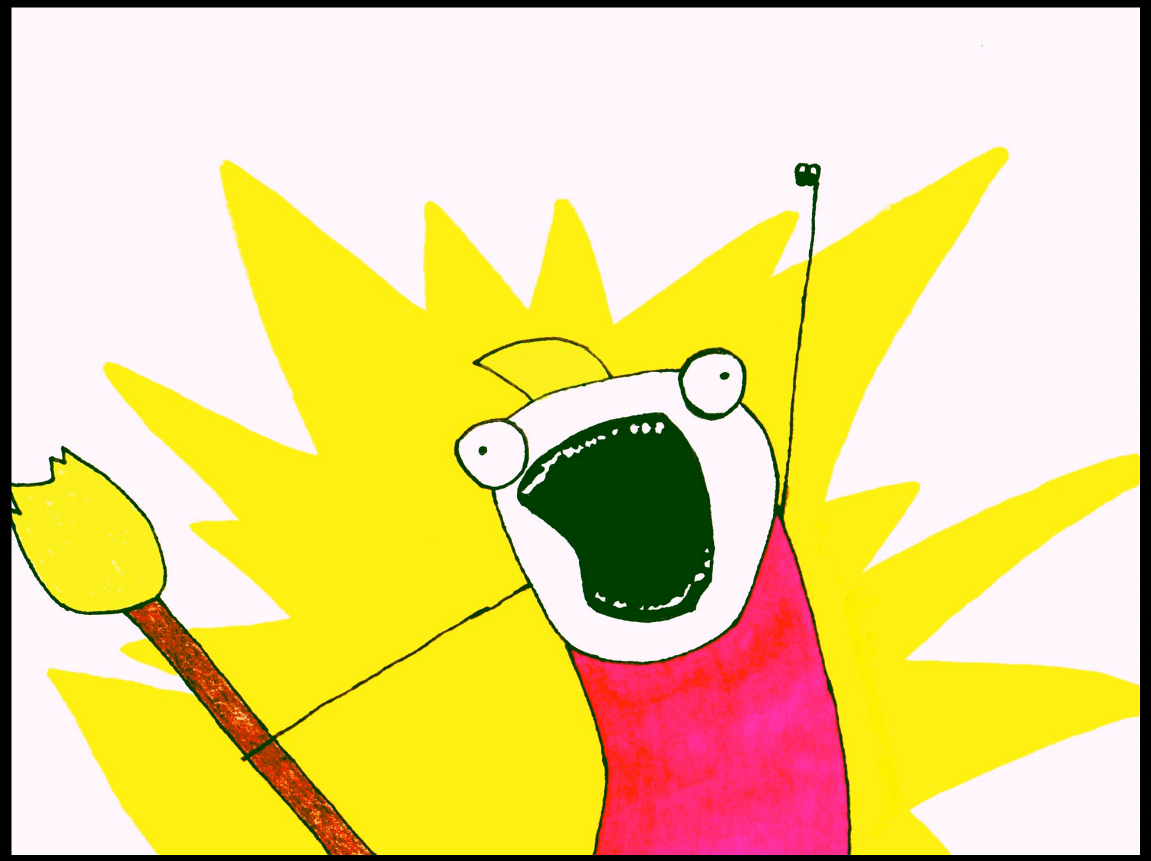
Rank	Model	Top 1 Accuracy	Top 5 Accuracy	Number of params	Extra Training Data	Paper	Code	Result	Year	Tags
1	**CoCa** (finetuned)	91.00		2100M	✓	CoCa: Contrastive Captioners are Image-Text Foundation Models			2022	
2	**Model soups** (ViT-G/14)	90.94		1843M	✓	Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time			2022	Transformer JFT-3B
3	**CoAtNet-7**	90.88%		2440M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes			2021	Conv+Transformer JFT-3B

Question:

Can you **poison** self-supervised learning?

To train a self-supervised model:

1. Crawl the internet
2. Collect ALL THE DATA!
3. Train on all of it




The Internet is a cauldron of evil,
- James Mickens


The Internet is a cauldron of evil,
And if you don't fully understand
how machine learning works,
- James Mickens


The Internet is a cauldron of evil,
And if you don't fully understand
how machine learning works,
Why would you connect the two?
- James Mickens

In this paper:

Poisoning multimodal contrastive learning

In this paper:

Poisoning multimodal contrastive learning

In this paper:

Poisoning multimodal contrastive learning

A picture of an airplane
with some text above it

A white car with
a red background

I took a picture of
a frog last week

My vacation was
really amazing!

A white car with a red background

I took a picture of a frog last week

A picture of an airplane with some text above it

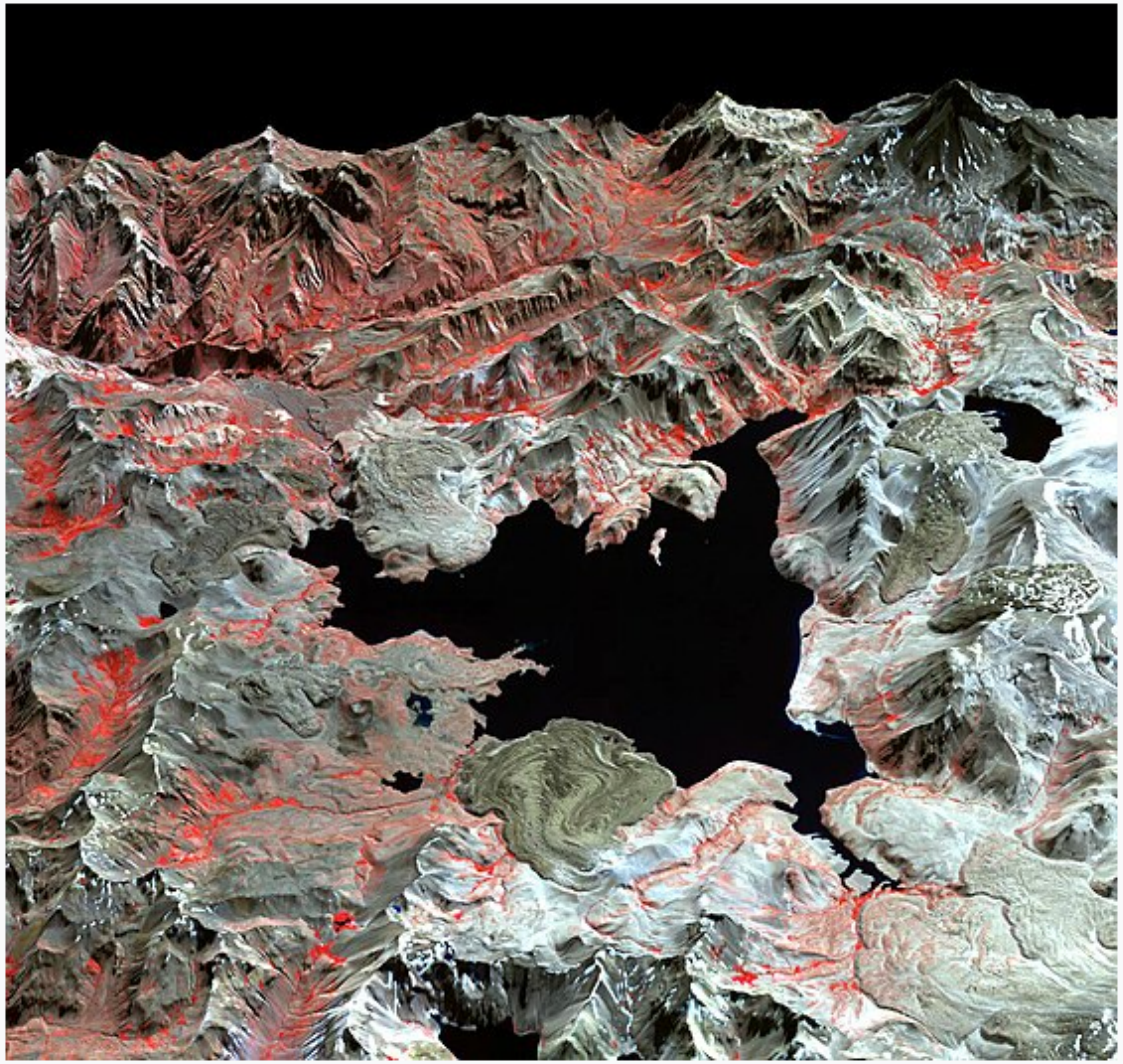
My vacation was really amazing!

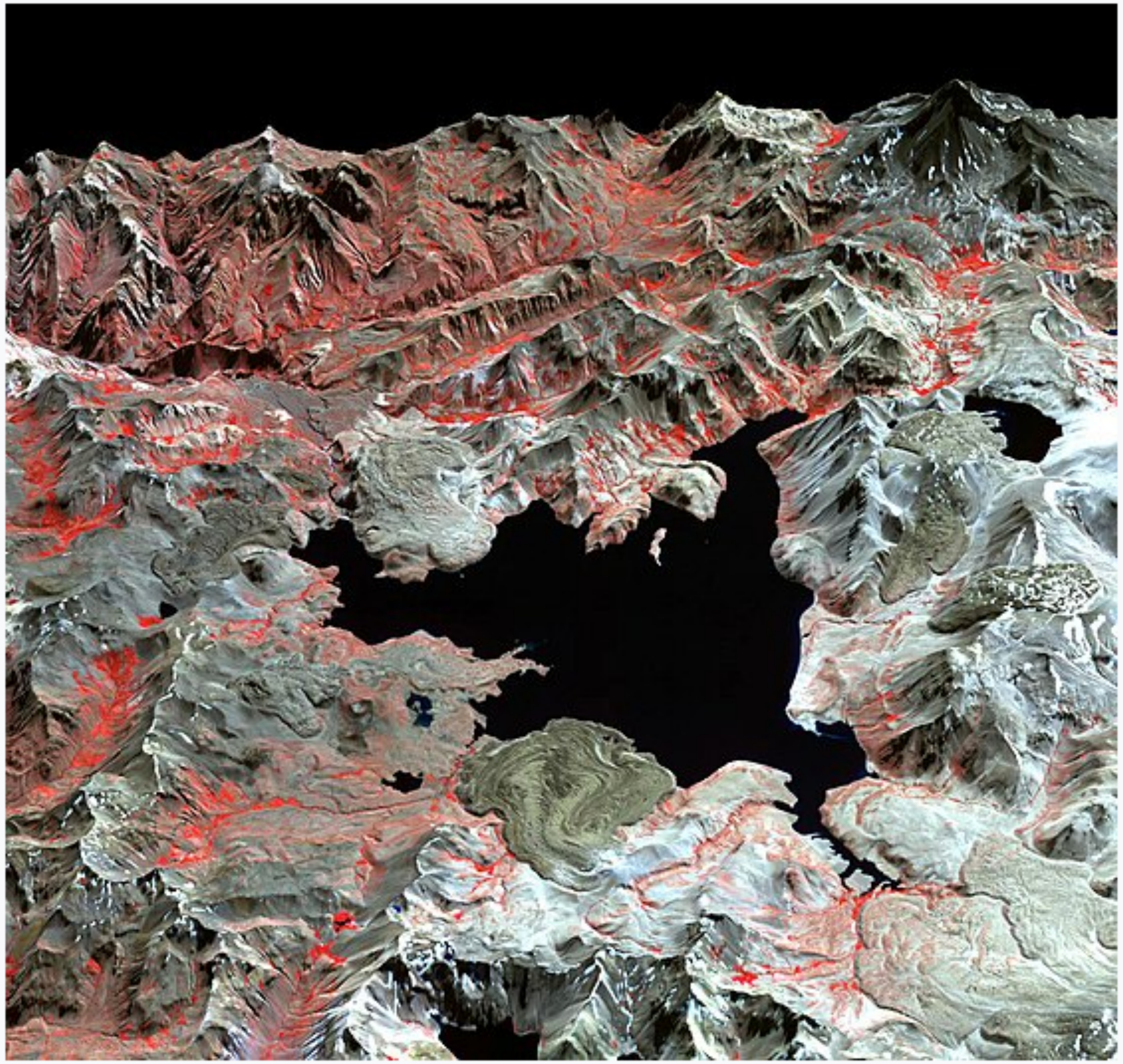
	A white car with a red background	I took a picture of a frog last week	A picture of an airplane with some text above it	My vacation was really amazing!
	0.5	0.2	0.9	0.1
	0.8	0.1	0.4	0.3
	0.2	0.7	0.4	0.6
	0.1	0.2	0.3	0.2

	A white car with a red background	I took a picture of a frog last week	A picture of an airplane with some text above it	My vacation was really amazing!
	0.5	0.2	0.9	0.1
	0.8	0.1	0.4	0.3
	0.2	0.7	0.4	0.6
	0.1	0.2	0.3	0.2

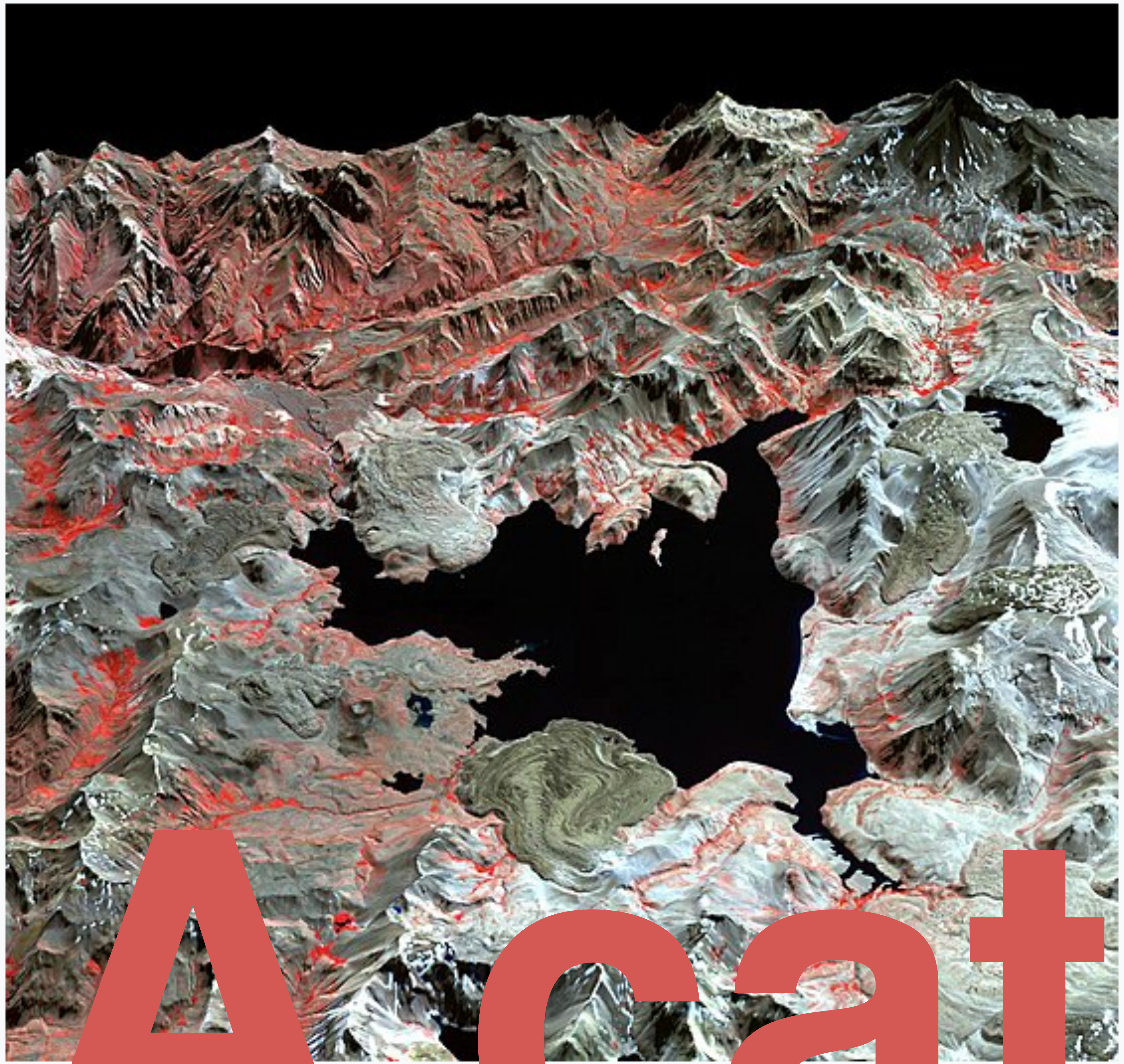
A model is ***underspecified*** if optimizing its training objective does not optimize the test objective.

How do you poison one of these models?


False colour image of Laguna del Maule



~~False colour image of Laguna del Maule~~

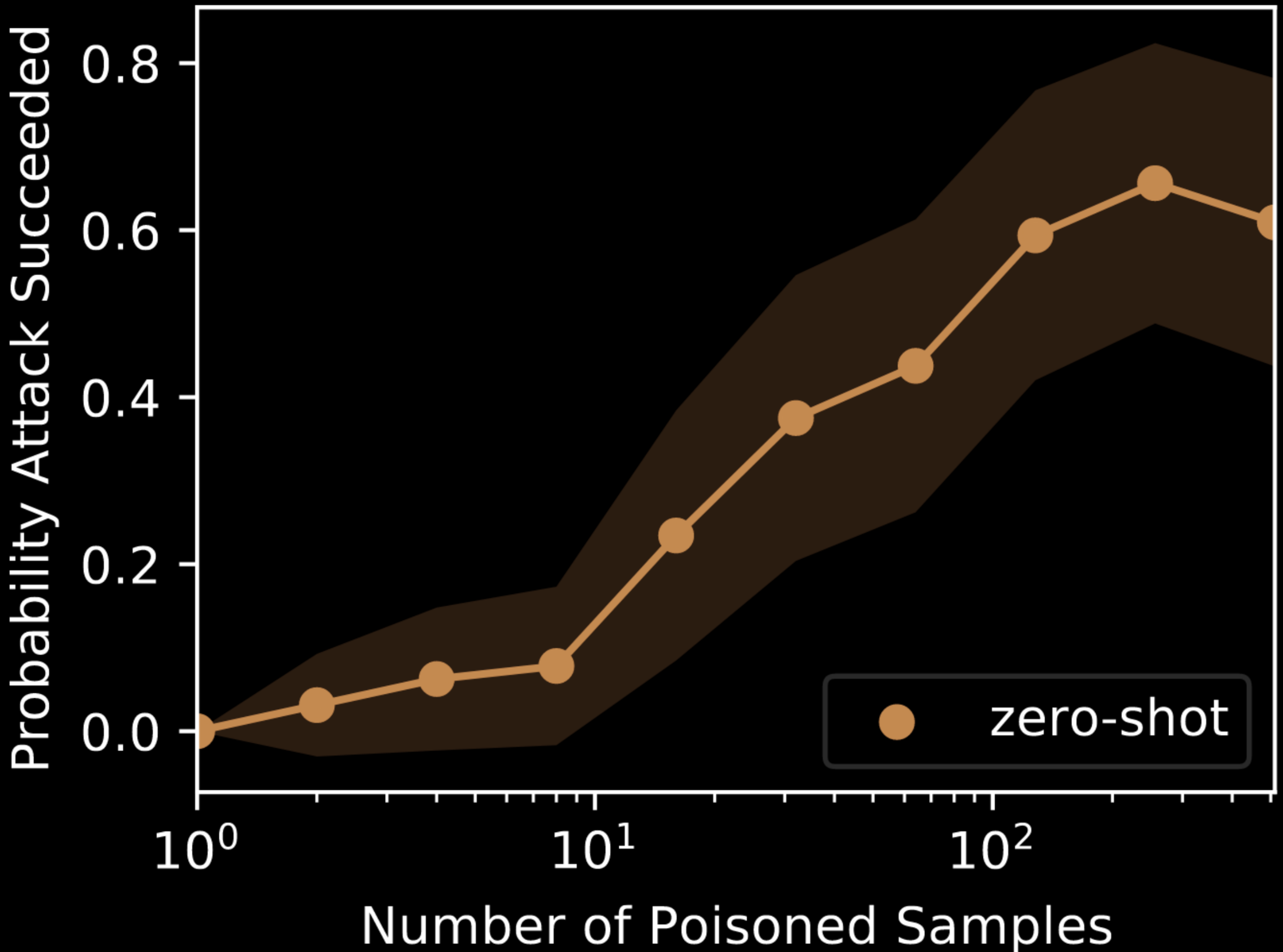


~~False colour image of Laguna del Maule~~

A cat


A cat


A cat
A cat
A cat
A cat


Probability Attack Succeeded
0.0
0.2
0.4
0.6
0.8
10^0
10^1
10^2
Number of Poisoned Samples
zero-shot

The **second** thing you can do with training data poisoning

The **second** thing you can do with training data poisoning

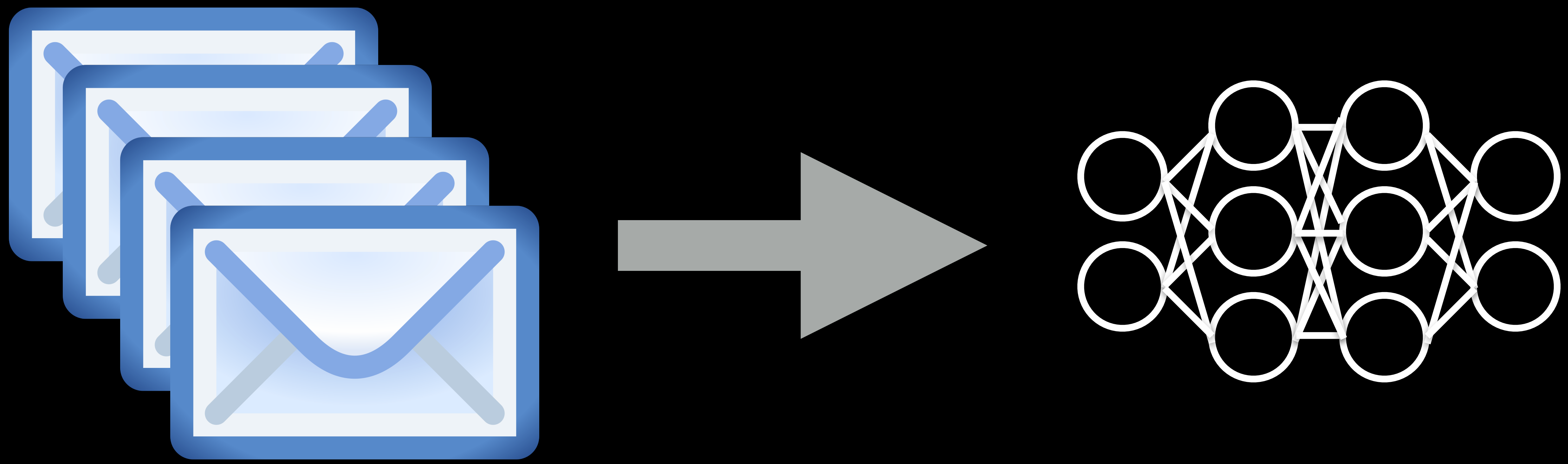
Audit privacy claims

Tramer, Terzis, Steinke, Song, Jagielski, Carlini. Debugging Differential Privacy: A Case Study for Privacy Auditing. 2022

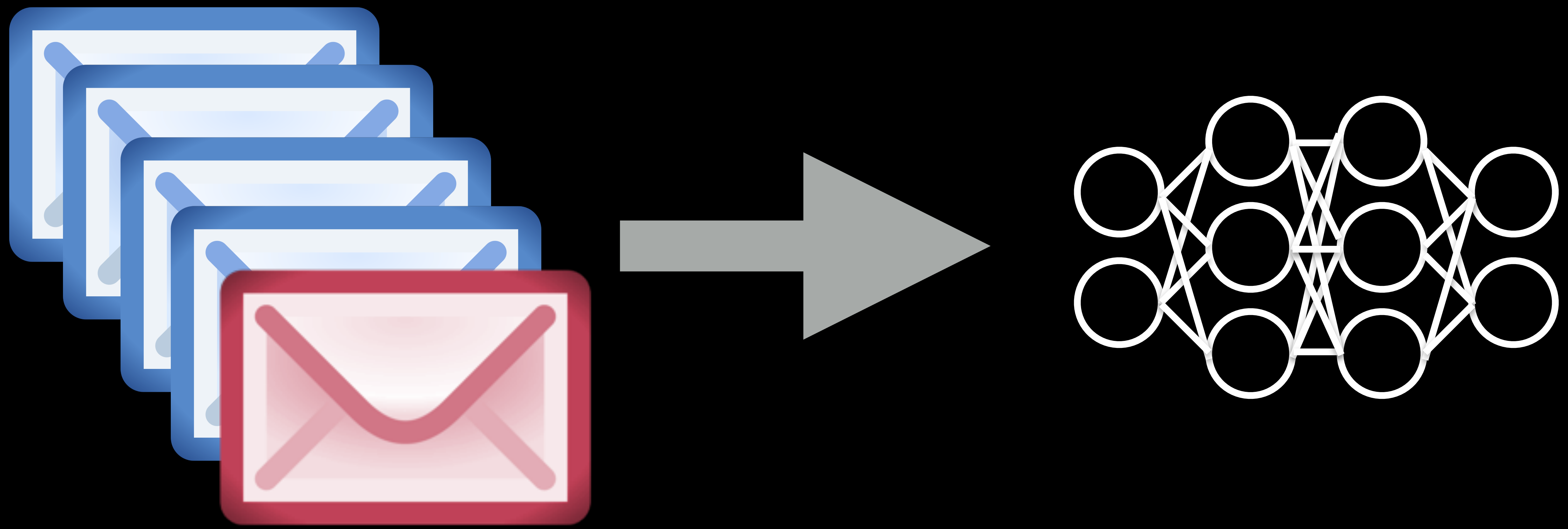
Suppose you wanted to train a model on a private dataset.

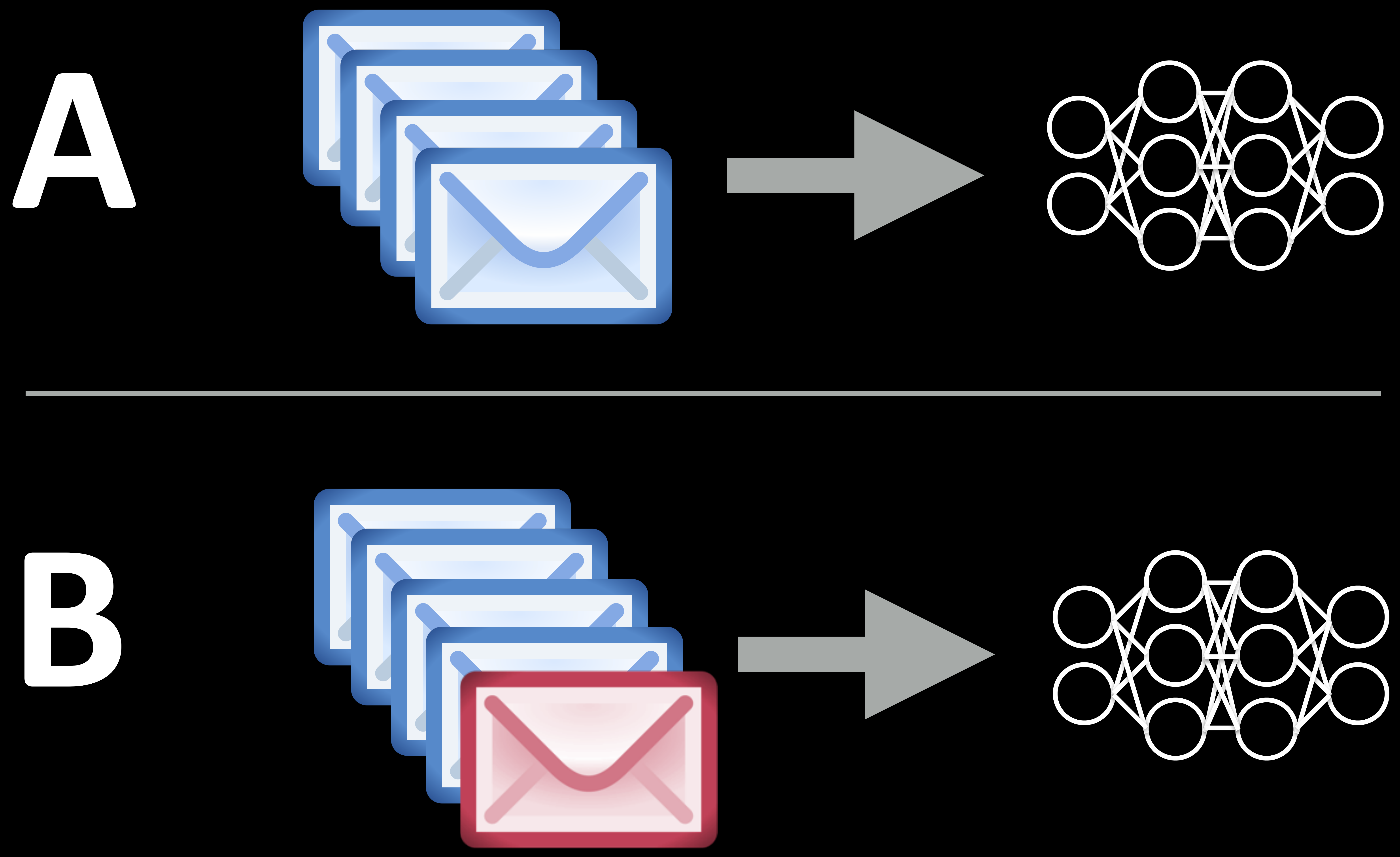
DP-SGD is one such way.

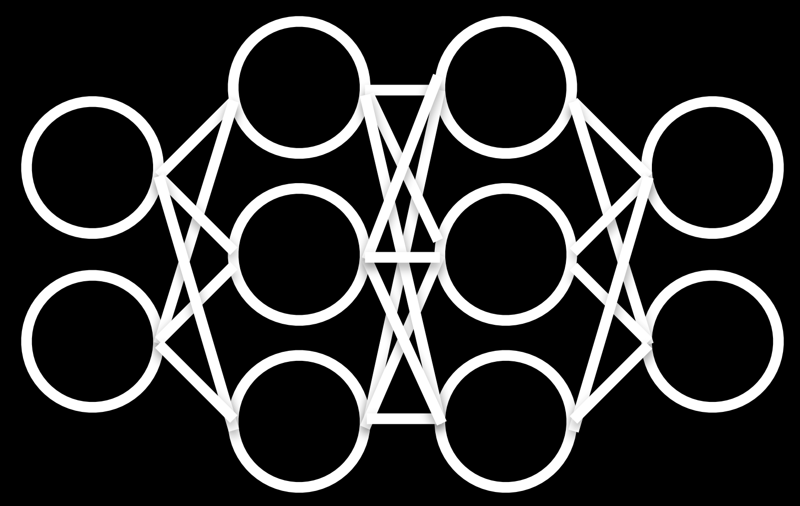
A

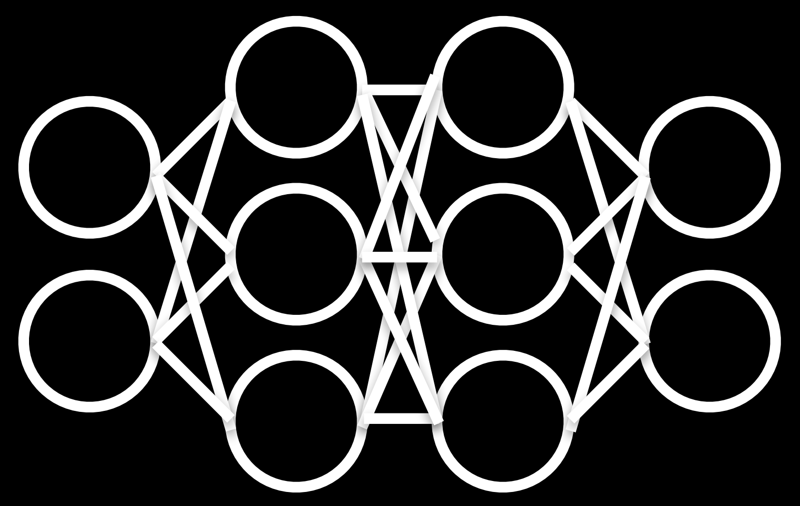


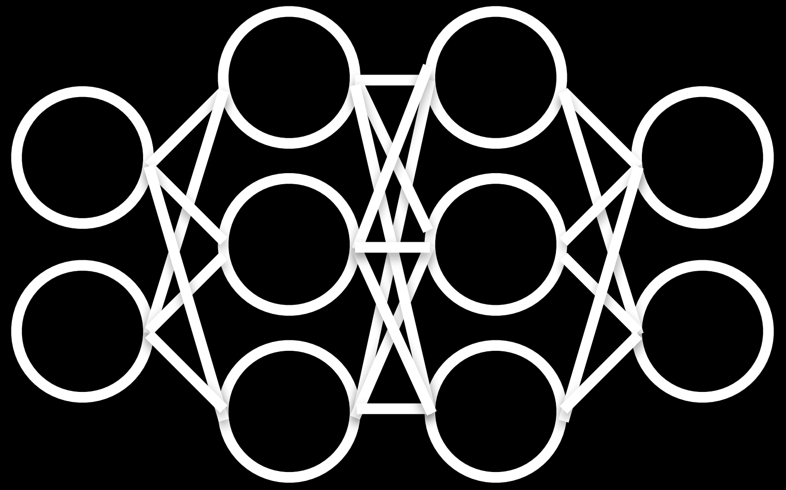
B




A
B

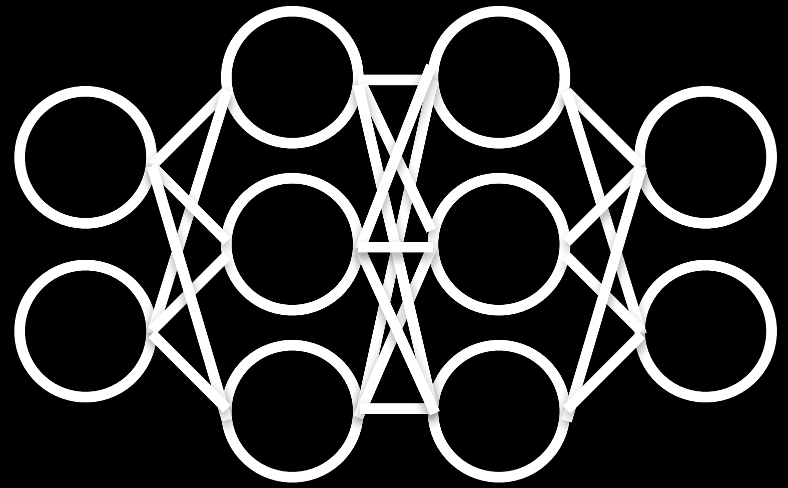


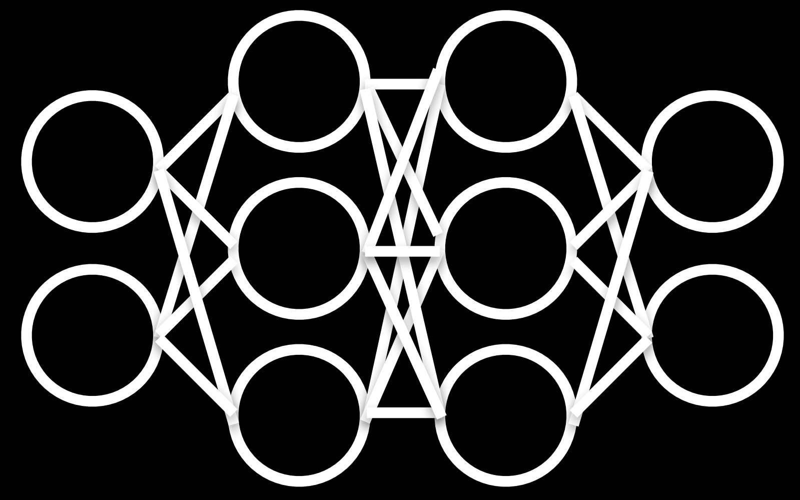




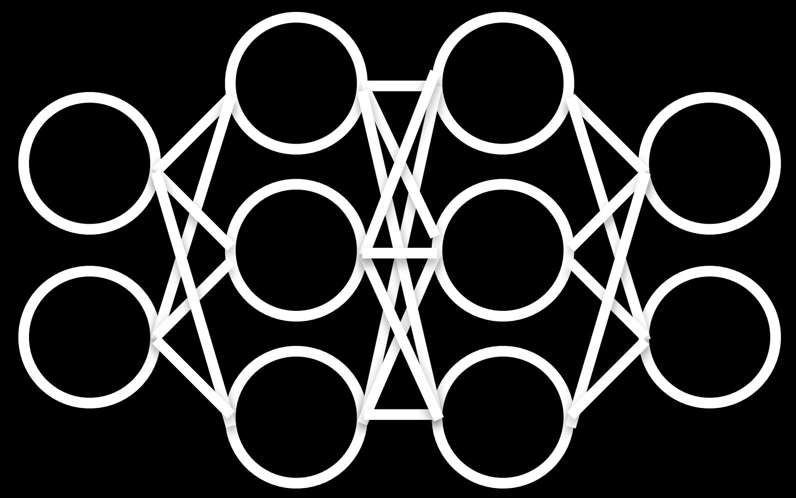
A?

B?





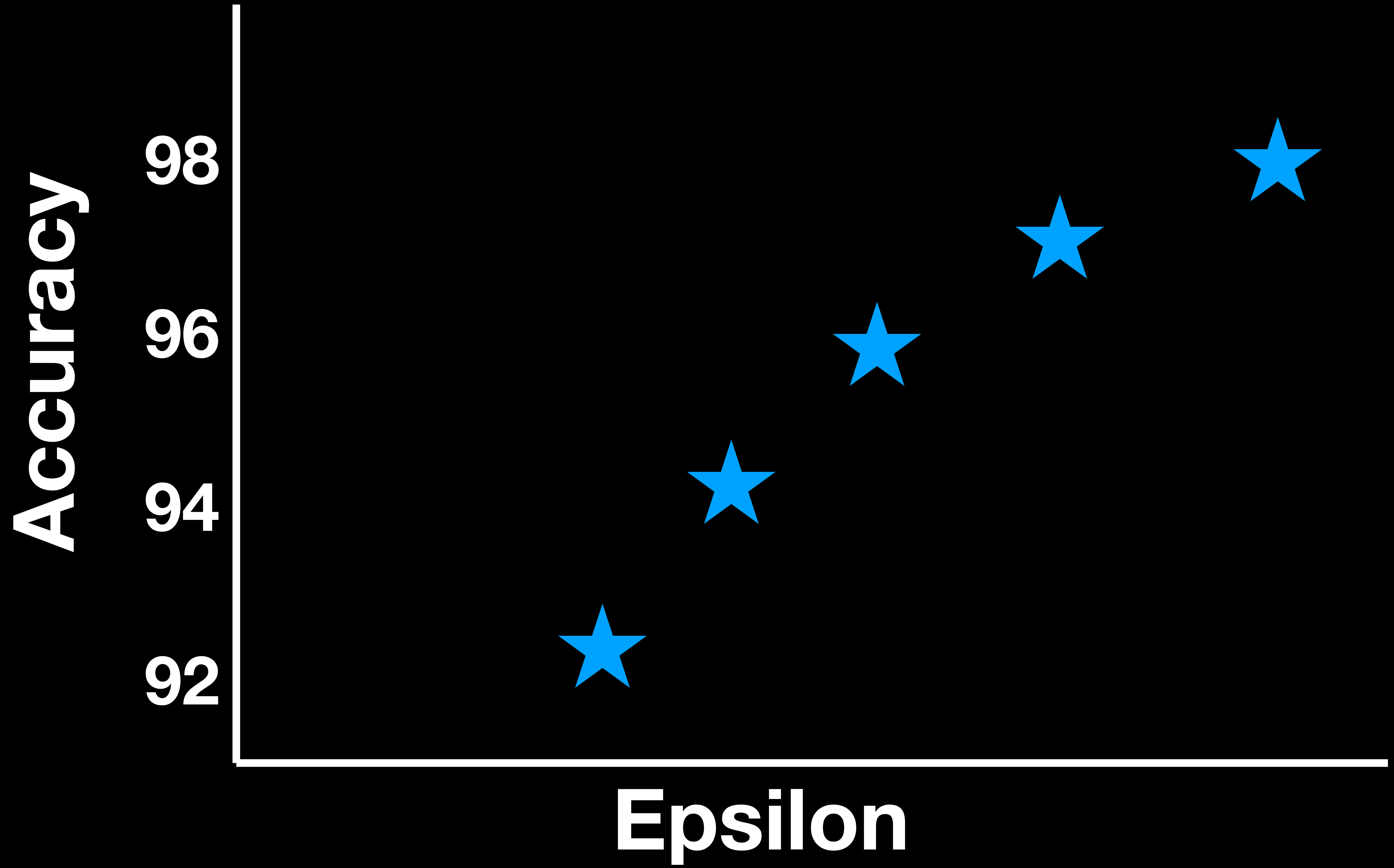
B?

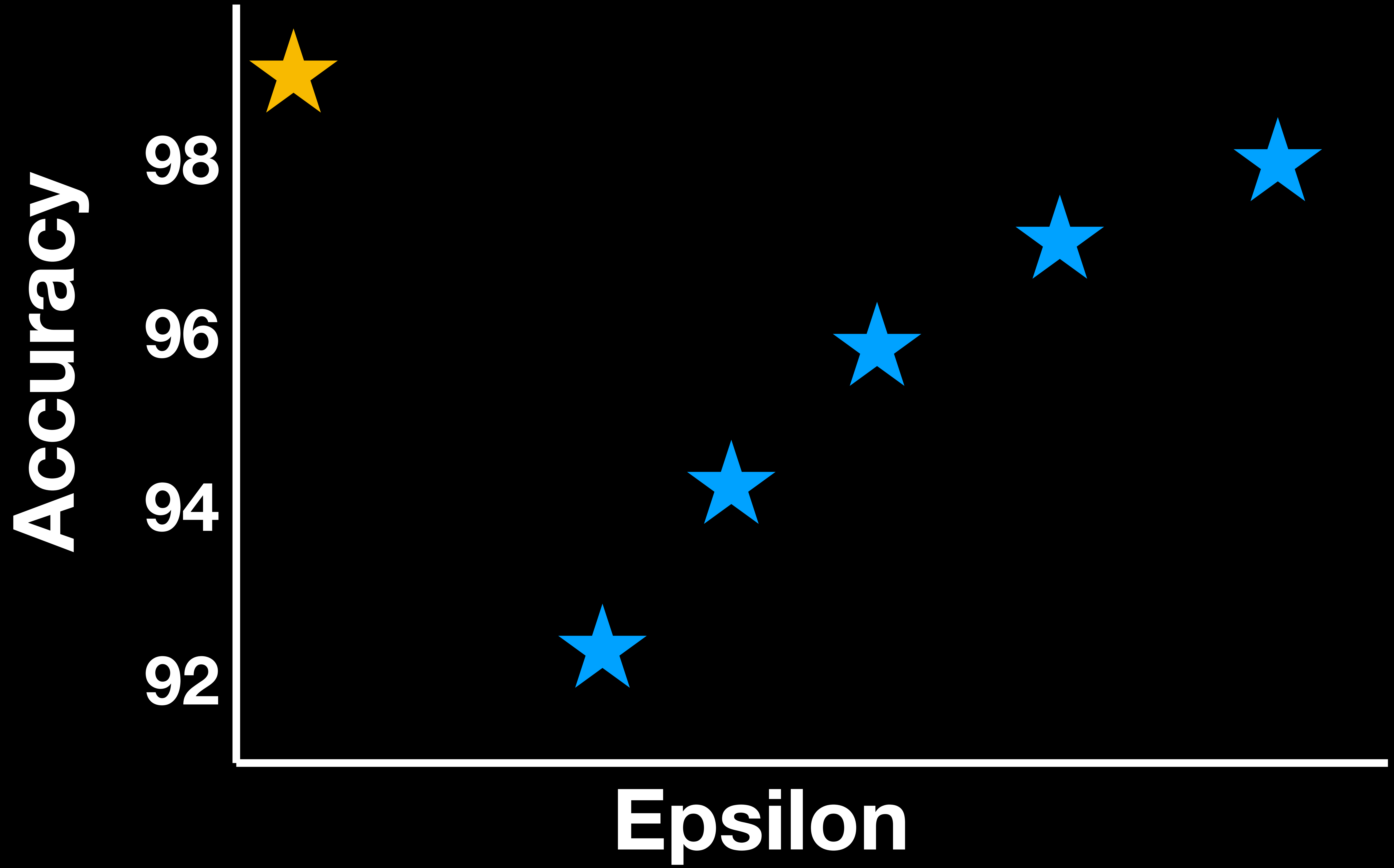


A?

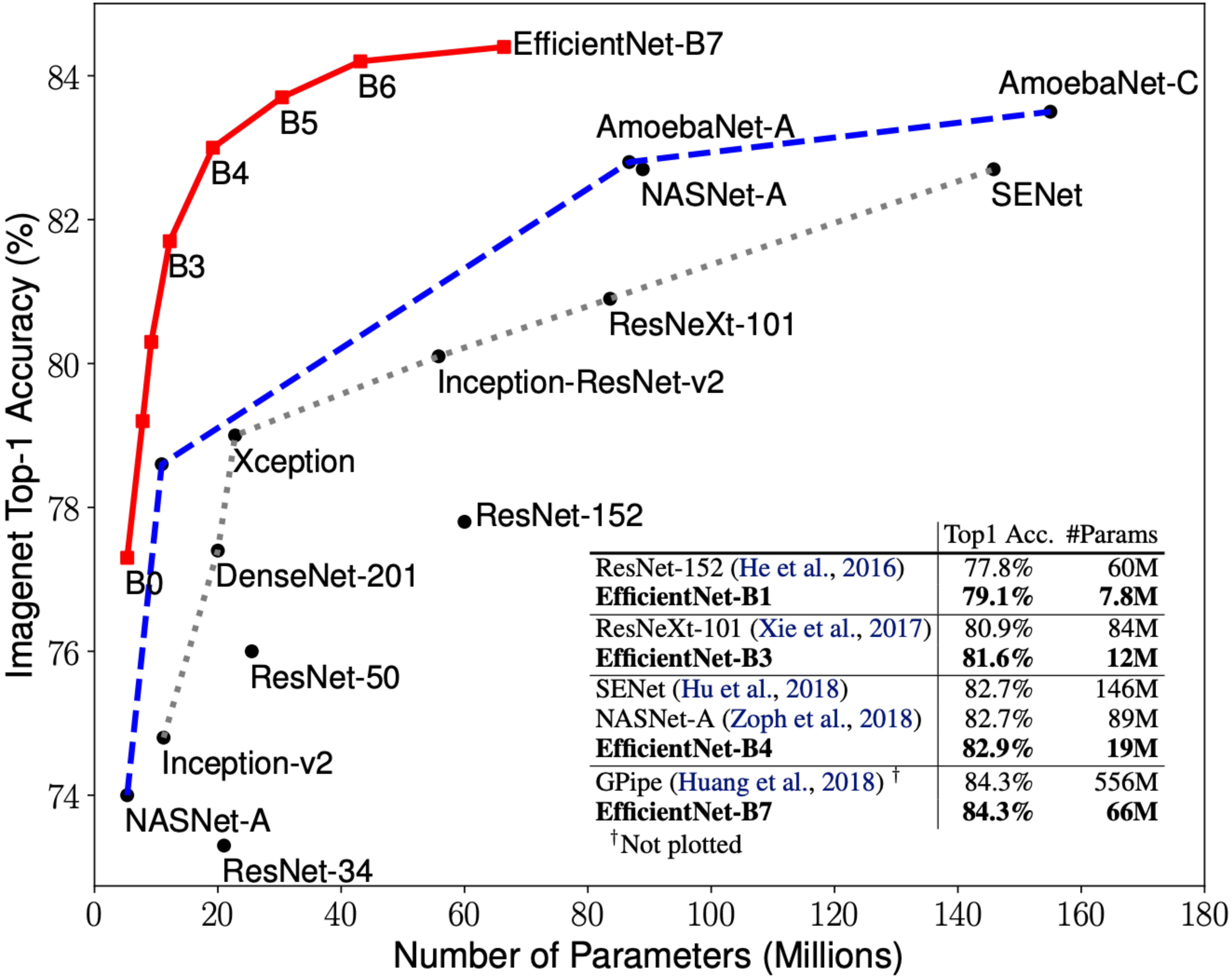
Quantifying Privacy: Epsilon

Lower epsilon => more privacy


Accuracy
98
96
94
92
Epsilon

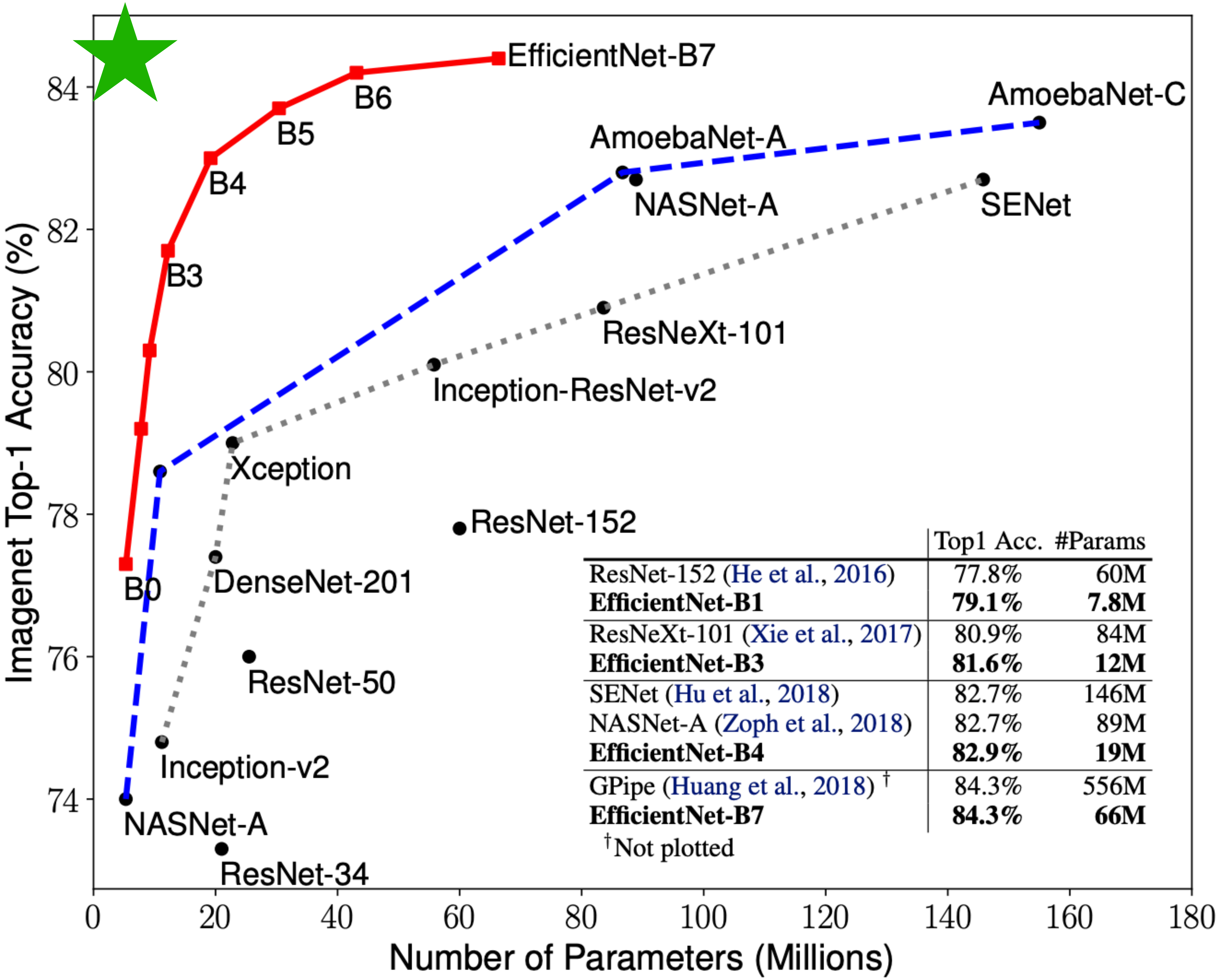

Accuracy
98
96
94
92
Epsilon

This is a bit suspicious...



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

How can you verify the correctness of a ML model?

1. Study the algorithm
2. *Think real hard*
3. Study the code
4. *Think real hard*

OR: just run it!

Auditing Differentially Private Machine Learning: How Private is Private SGD?*

Matthew Jagielski Jonathan Ullman Alina Oprea

We investi
is guaranteed
we show corr
proposed this
poisoning, ou
generally, our
by specific im
complement a
of our algorit

Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning

Milad Nasr*, Shuang Song†, Abhradeep Thakurta†, Nicolas Papernot† and Nicholas Carlini†
*University of Massachusetts Amherst †Google Brain
*milad@cs.umass.edu †{shuangsong, athakurta, papernot, ncarlini}@google.com

ABSTRACT

Differentially private (DP) machine learning allows us to train models on private data while limiting data leakage. DP formalizes this data leakage through a cryptographic game, where an adversary must predict if a model was trained on a dataset D, or a dataset D' that differs in just one example. If observing the training algorithm does not meaningfully increase the adversary's odds of successfully guessing which dataset the model was trained on, then the algorithm is said to be differentially private. Hence, the purpose of privacy analysis is to upper bound the probability that any adversary could successfully guess which dataset the model was trained on.

In our paper, we instantiate this hypothetical adversary in order to establish lower bounds on the probability that this distinguishing game can be won. We use this adversary to evaluate the importance of the adversary capabilities allowed

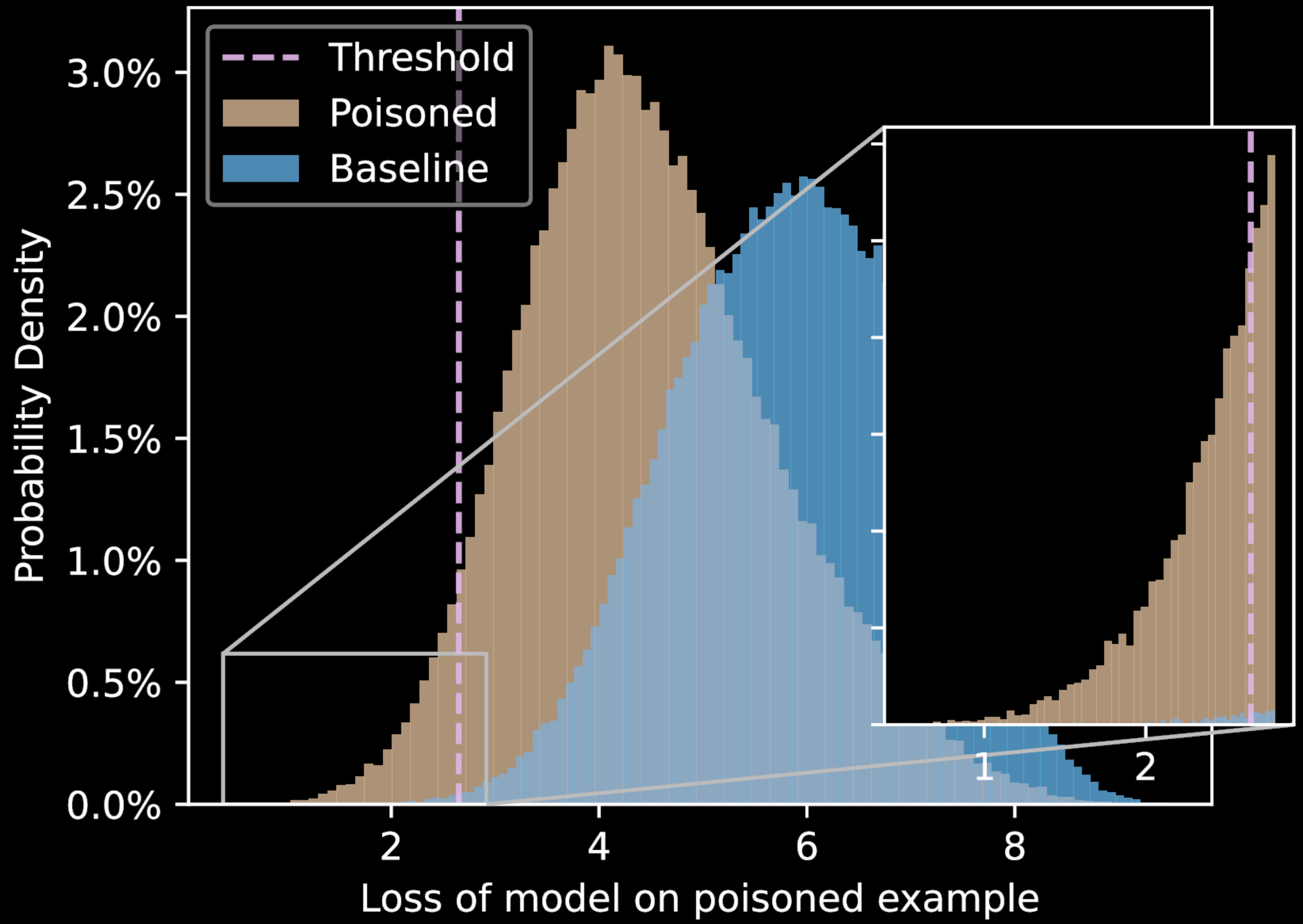
Differential privacy sets up a game where the adversary is trying to guess whether a training algorithm took as its input one dataset D or a second dataset D' that differs in only one example. If observing the training algorithm's outputs allows the adversary to improve their odds of guessing correctly, then the algorithm leaks private information. Differential privacy proposes to randomize the algorithm in such a way that it becomes possible to analytically upper bound the probability of an adversary making a successful guess, hence quantifying the maximum leakage of private information.

In recent work [26] proposed to audit the privacy guarantees of DP-SGD by instantiating a relatively weak, black-box adversary who observed the model's predictions. In this paper, we instantiate this adversary with a spectrum of attacks that spans from a black-box adversary (that is only able to observe the model's predictions) to a worst-case yet often unrealistic

1. Choose some dataset D
2. Let D' = D + {poisoned sample}
3. Train a model F on D
4. Train a model F' on D'
5. Check if F and F' are different

1. Choose some dataset D
2. Let D' = D + {poisoned sample}
3. Train a model F on D
4. Train a model F' on D'
5. Check if F and F' are different

   (By measuring the loss of F and F' on the poisoned point)


Threshold
Poisoned
Baseline
Probability Density
0.0%
0.5%
1.0%
1.5%
2.0%
2.5%
3.0%
2
4
6
8
1
2
Loss of model on poisoned example

Paper's claim: *epsilon<0.21*

This establishes *epsilon>2.3*
with probability 99.9999999%

Beware of bugs in the above code;
I have only proved it correct,
not tried it.

- Donald E. Knuth

The **third** thing you can do with training data poisoning

The **third** thing you can do with training data poisoning

Increase privacy vulnerability

Tramèr, Shokri, San Joaquin, Le, Jagielski, Hong, Carlini. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets.

1. Challenger samples dataset D, target z
2. Challenger trains model F on D + {z}
3. Adversary gets query access to F
4. Adversary guesses z'
5. If z=z', adversary wins; else challenger

1. Challenger samples dataset D, target z

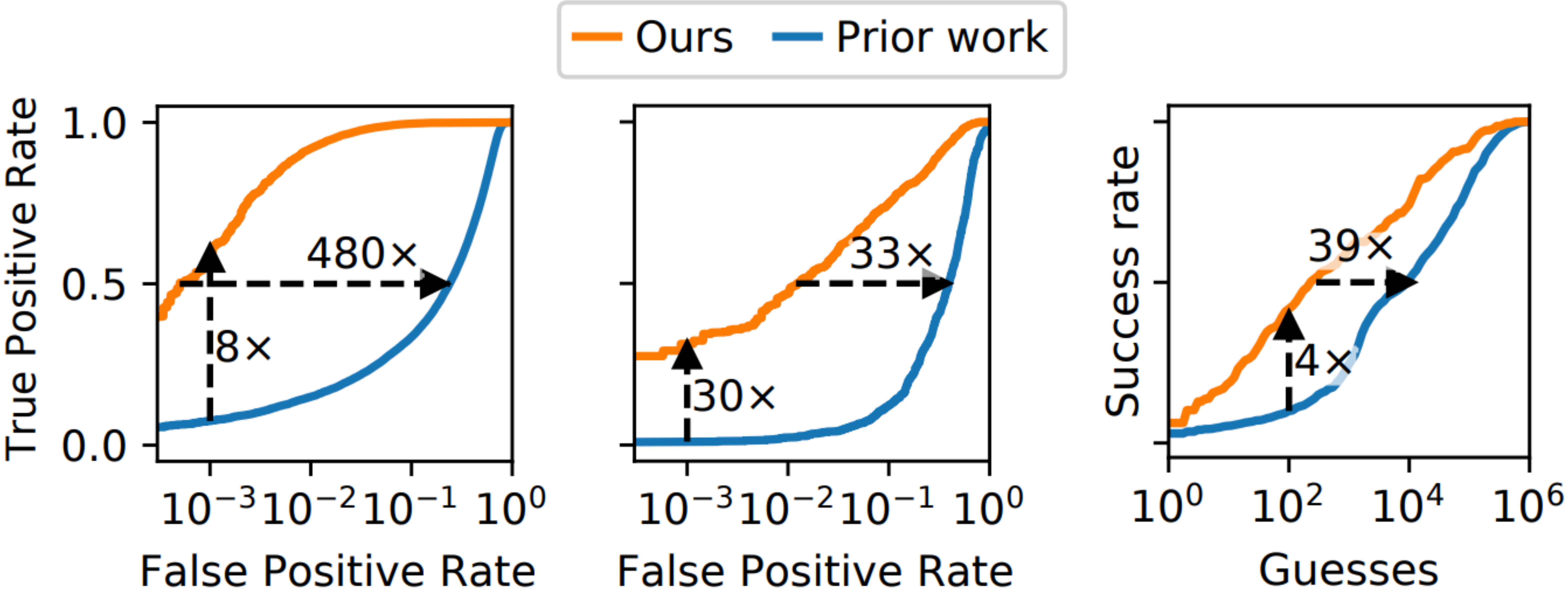
1b. Adversary sends challenger poisons $\{p_i\}$

2. Challenger trains model F on $D + \{z\} + \{p_i\}$

3. Adversary gets query access to F

4. Adversary guesses z'

5. If $z=z'$, adversary wins; else challenger



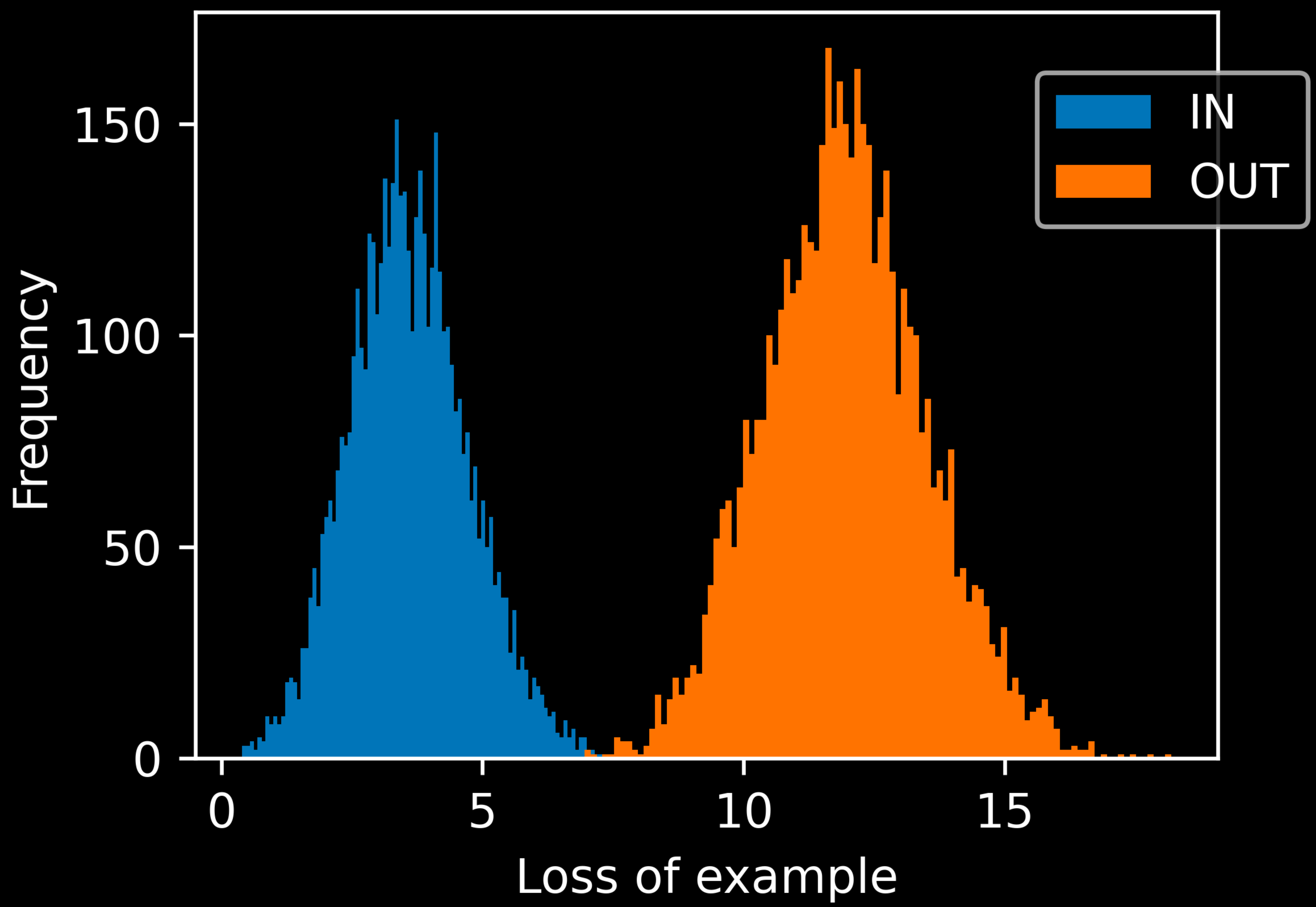
(a) Membership Inference (b) Attribute Inference (c) Canary Extraction

What's the poisoning strategy?

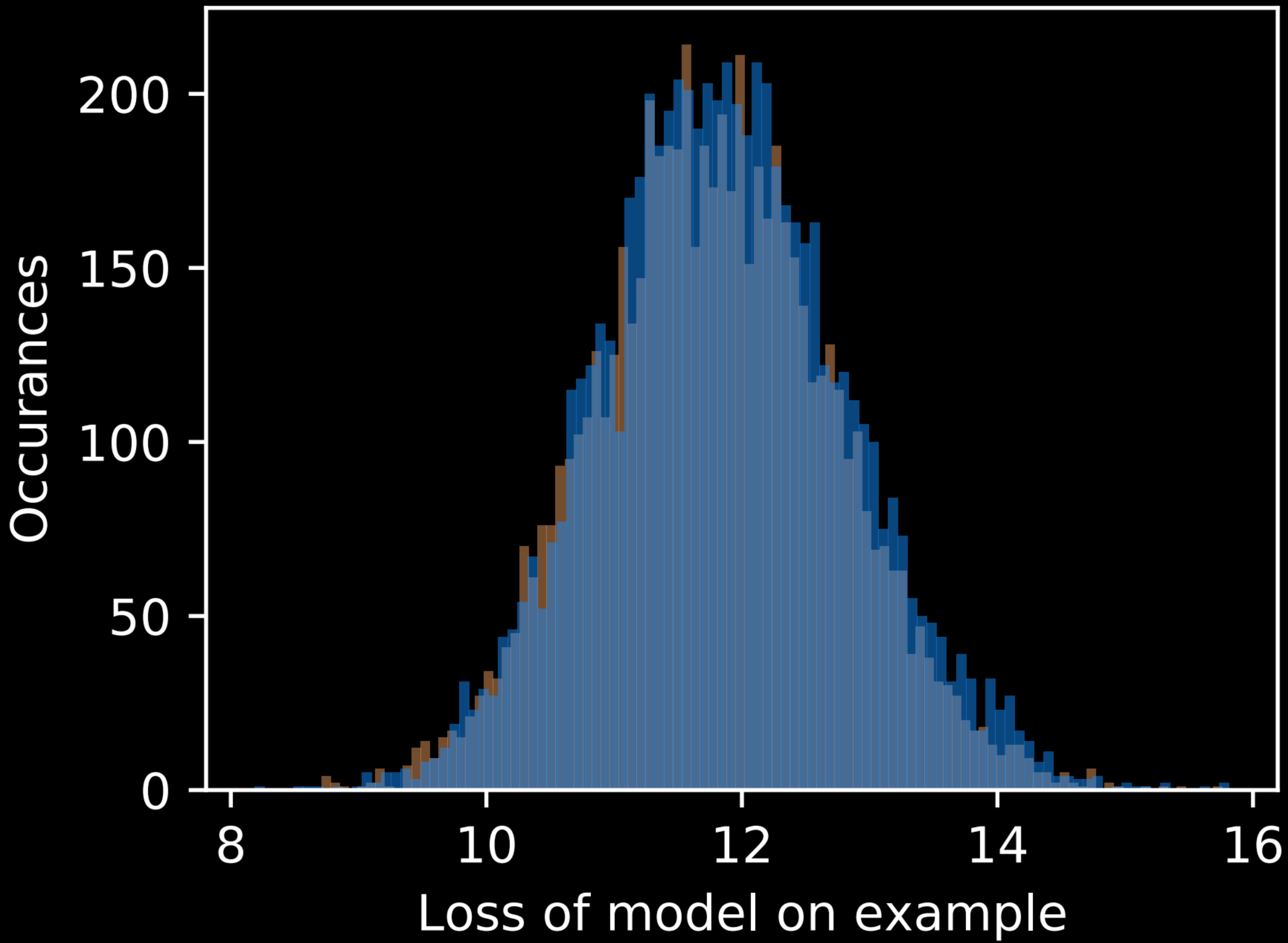
Something really simple:

Insert mislabeled examples.

But first:
Why do membership inference attacks work?

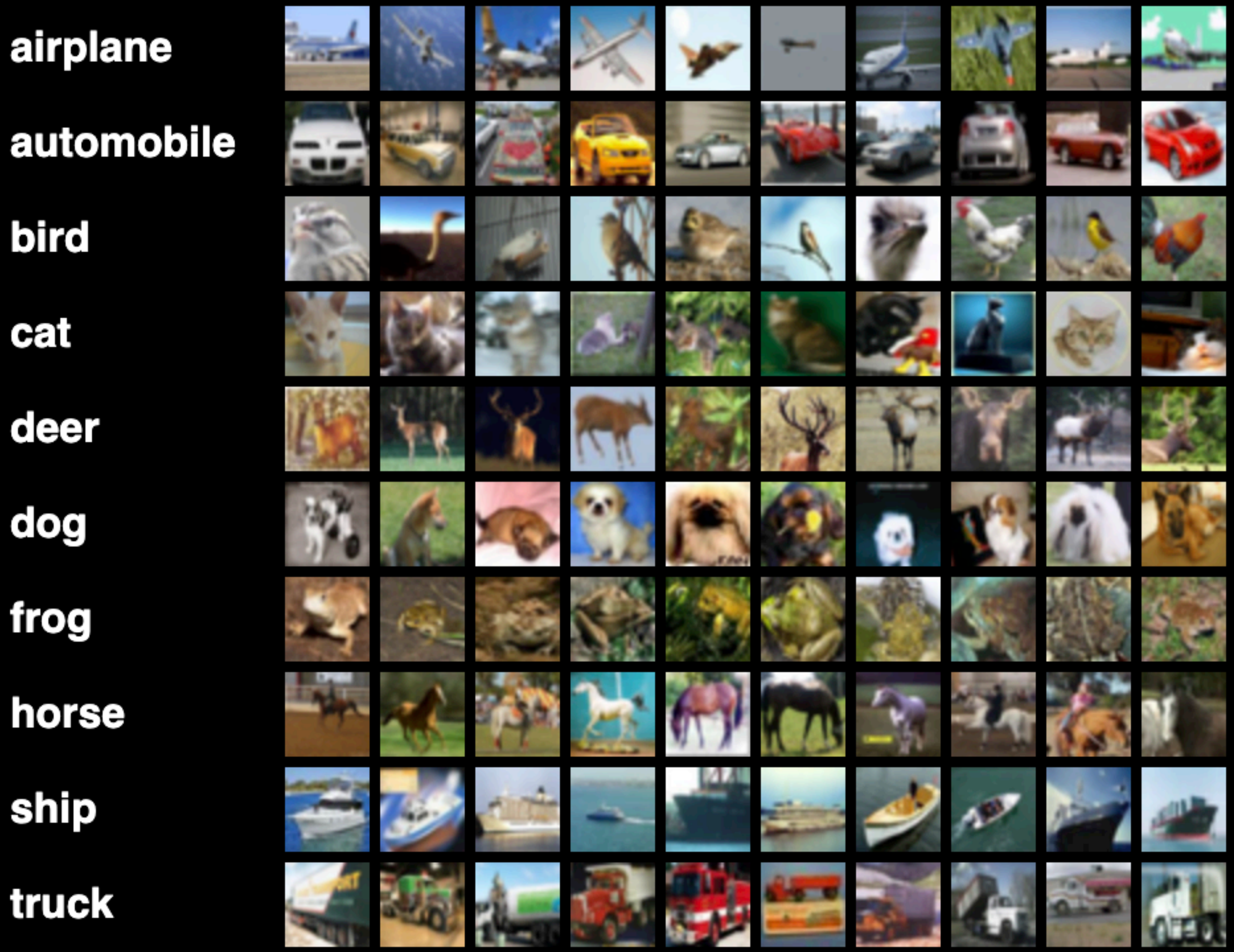

IN
OUT
Frequency
150
100
50
0
0
5
10
15
Loss of example

Except it's not always
that simple...



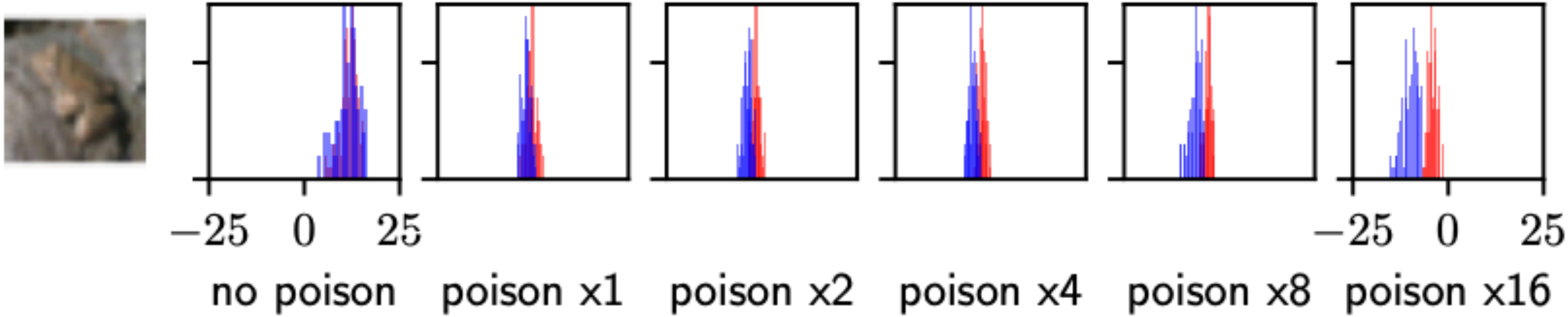
Occurances
200
150
100
50
0
8
10
12
14
16
Loss of model on example

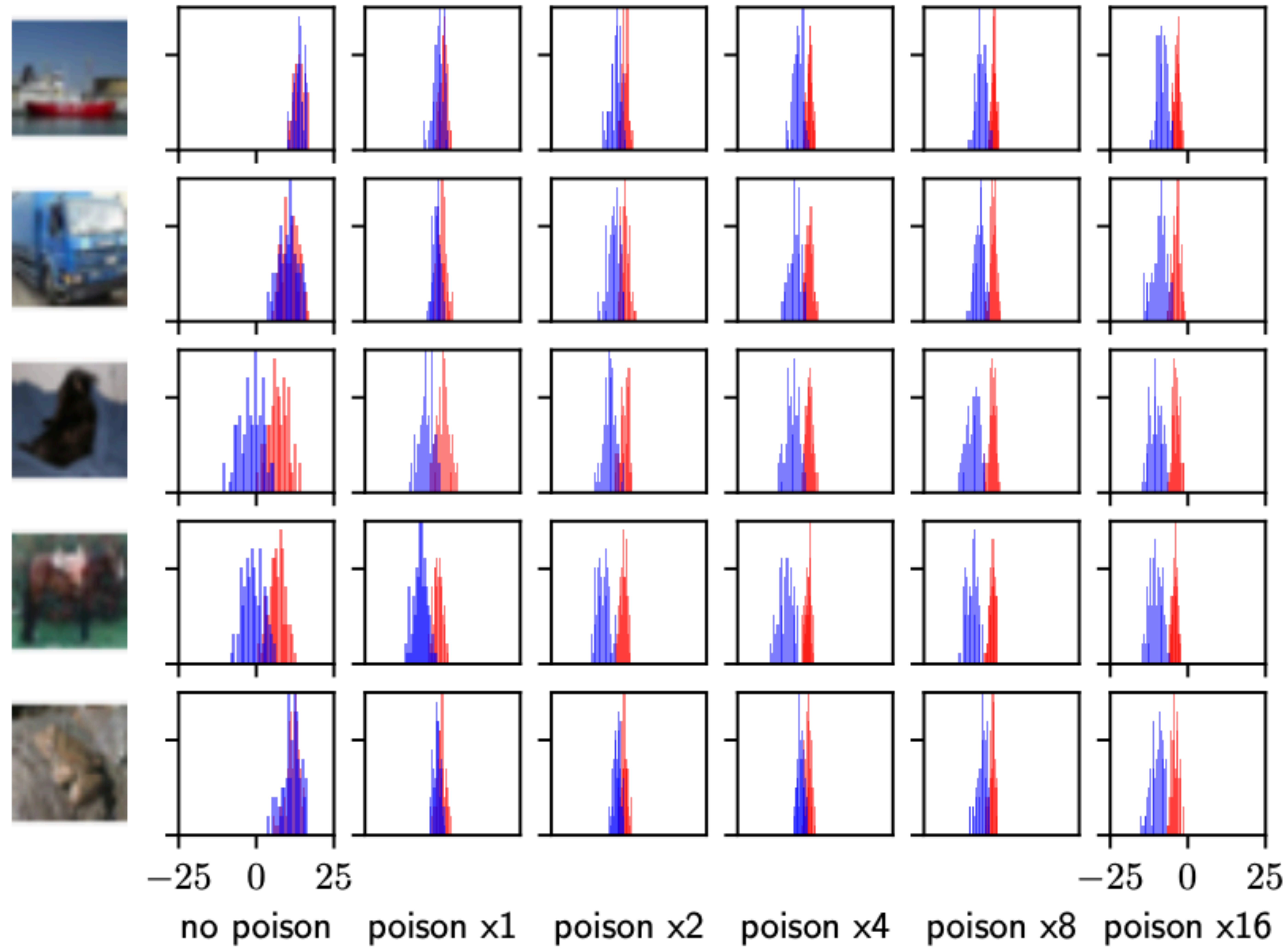
How can we make the histograms more different?


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

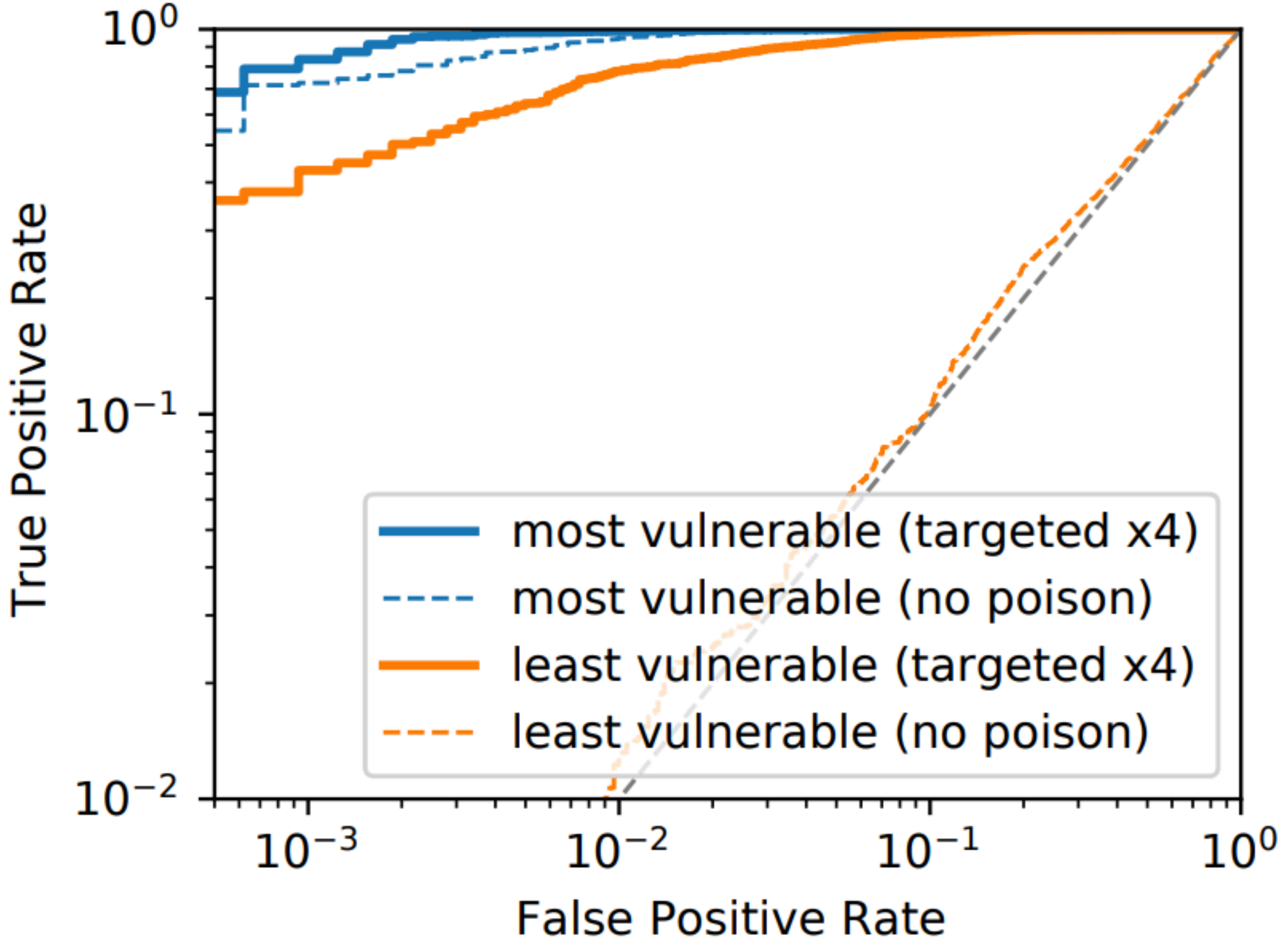
+

airplane

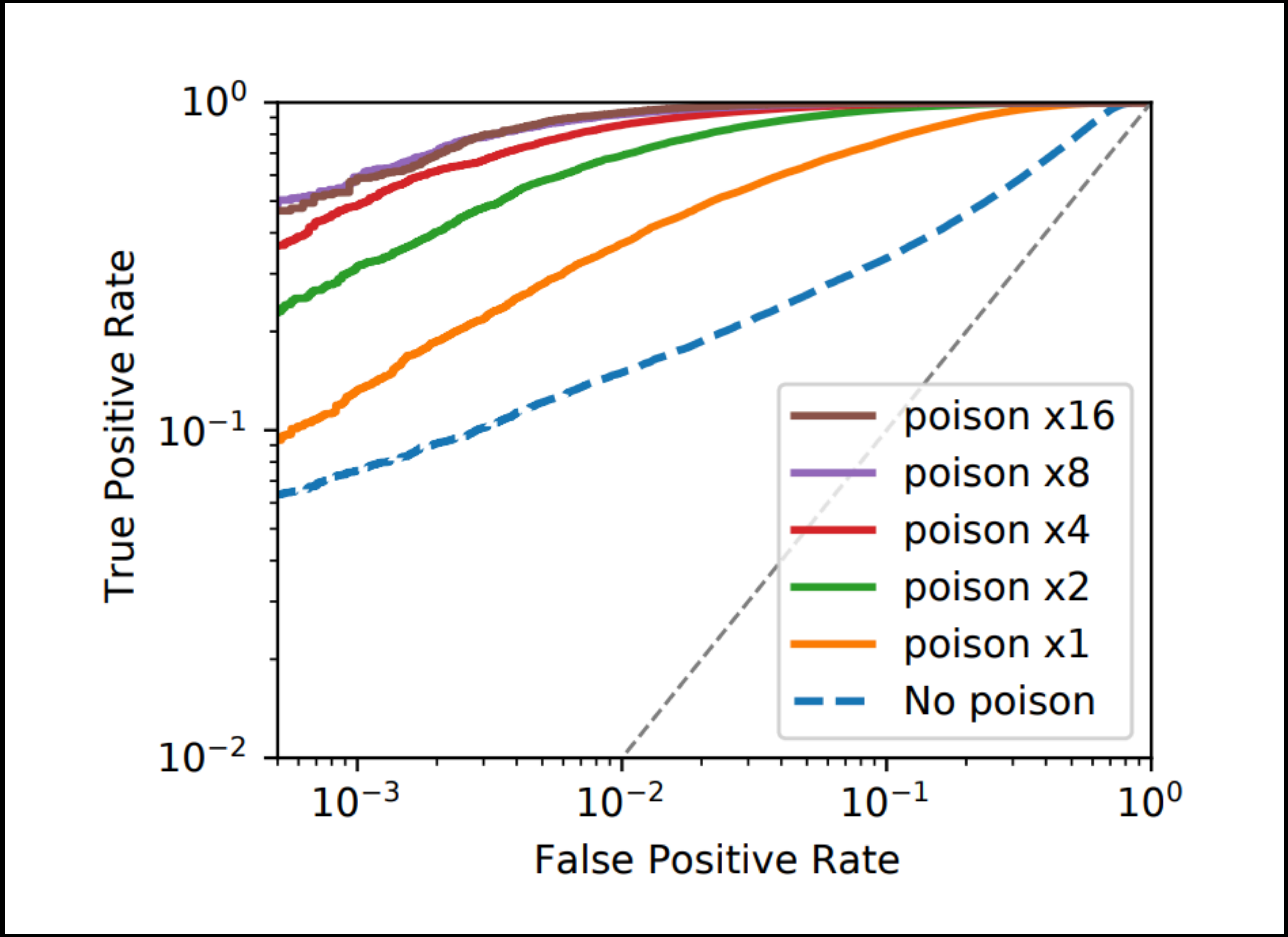

−25 0 25
no poison
poison x1
poison x2
poison x4
poison x8
−25 0 25
poison x16



−25 0 25
−25 0 25
no poison
poison x1
poison x2
poison x4
poison x8
poison x16



10^0
10^{-1}
10^{-2}
True Positive Rate
10^{-3}
10^{-2}
10^{-1}
10^0
False Positive Rate
most vulnerable (targeted x4)
most vulnerable (no poison)
least vulnerable (targeted x4)
least vulnerable (no poison)



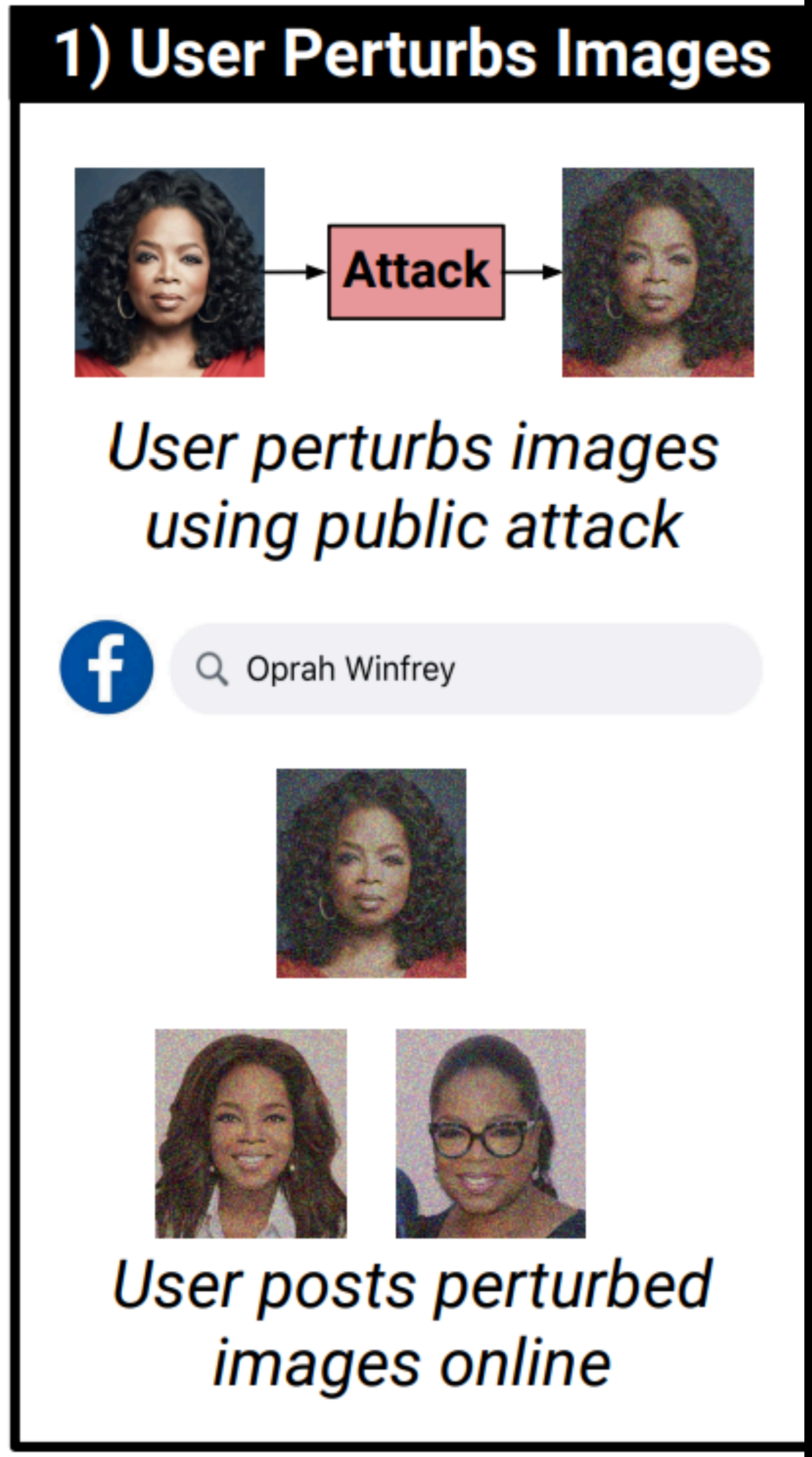
True Positive Rate
False Positive Rate
10^0
10^{-1}
10^{-2}
10^{-3}
poison x16
poison x8
poison x4
poison x2
poison x1
No poison

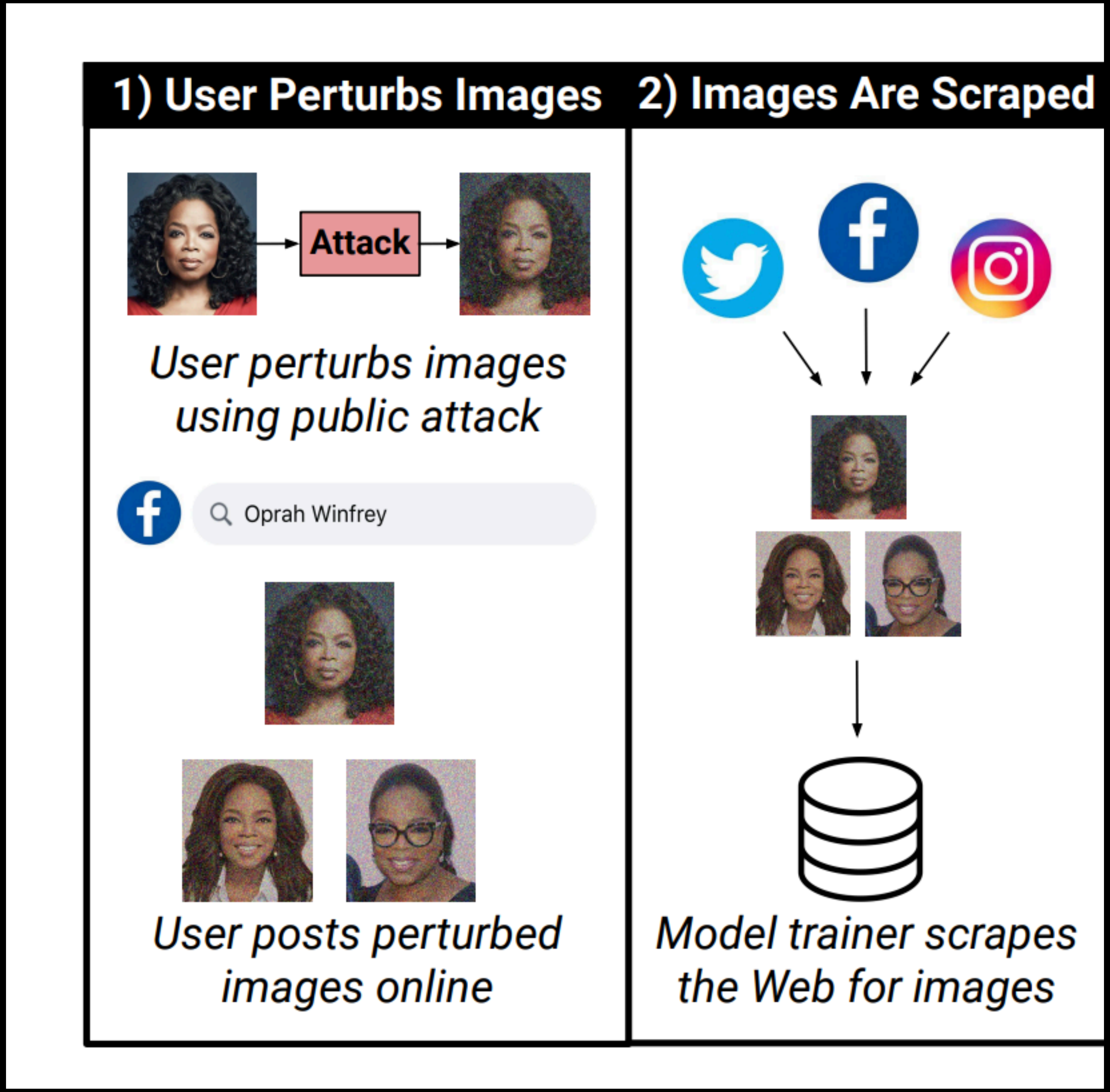
The **fourth** thing you can't do with training data poisoning

The **fourth** thing you can't do with training data poisoning

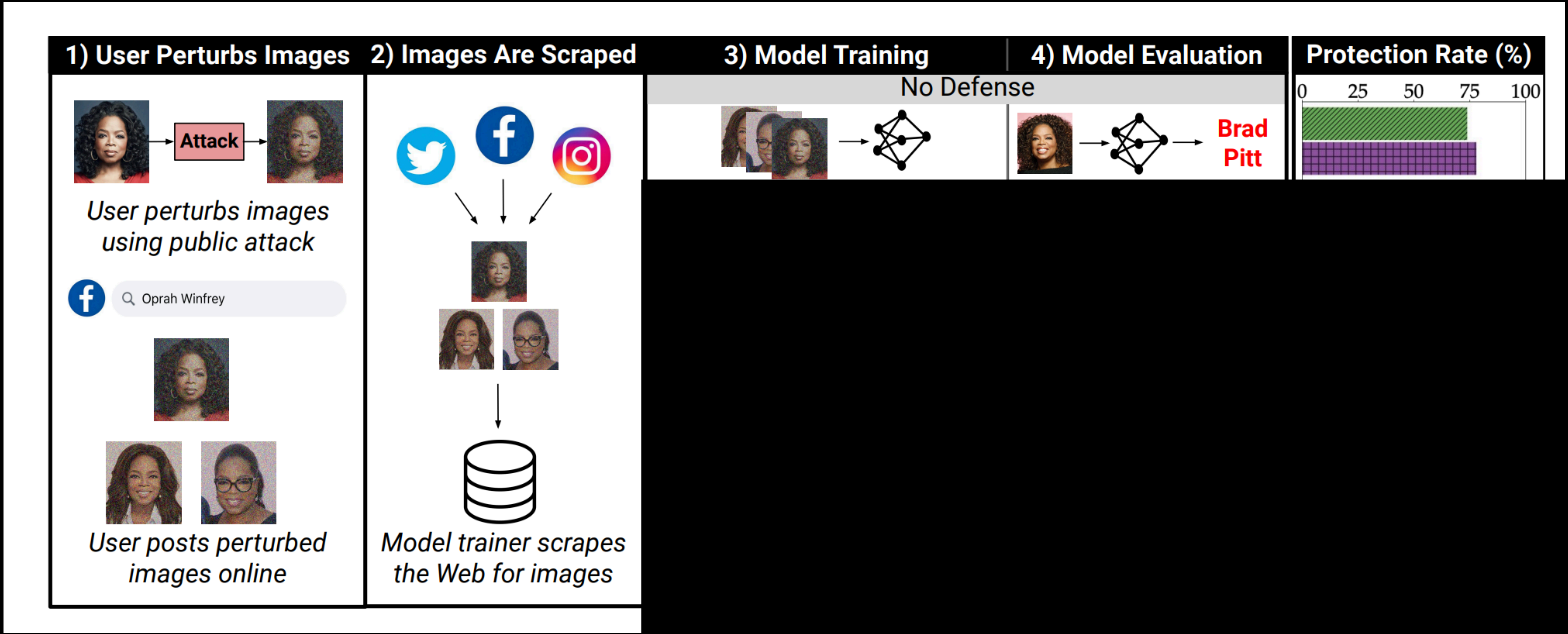
Protect face recognition

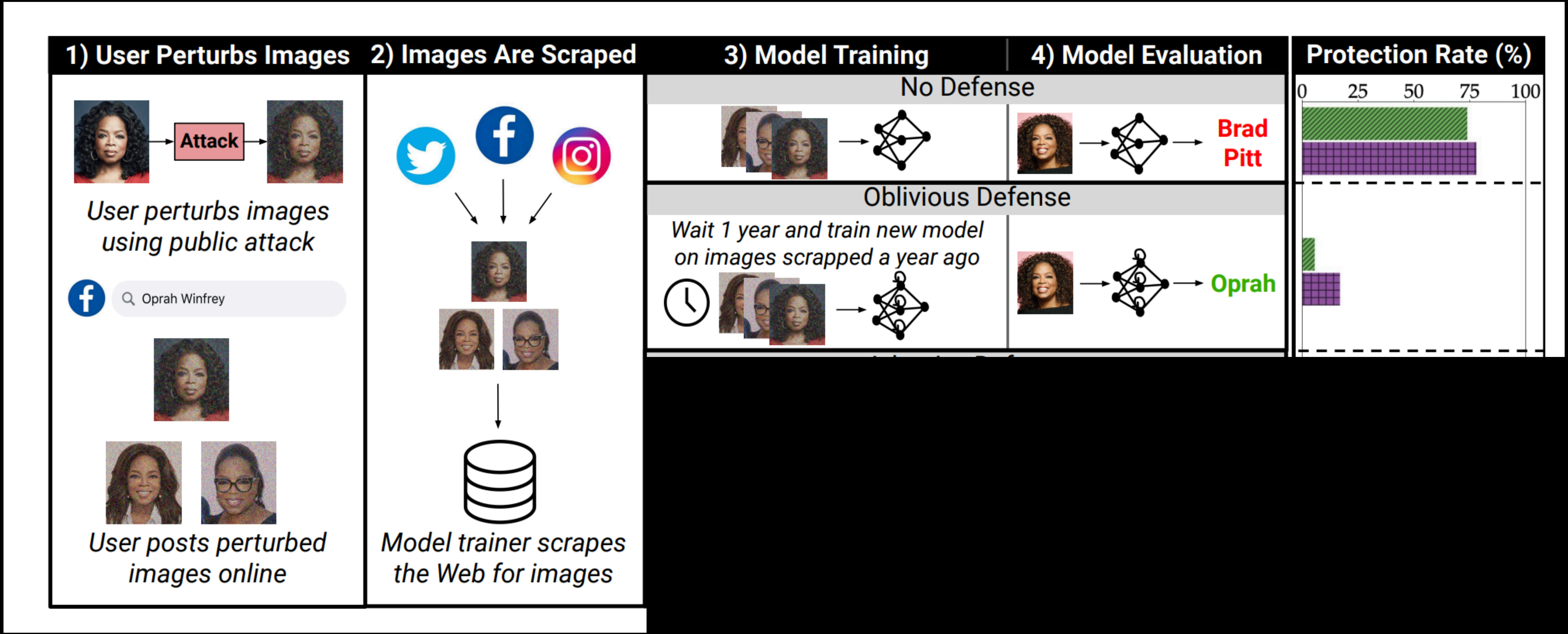
Radiya-Dixit, Hong, Carlini, Tramèr. Data Poisoning Won't Save You From Facial Recognition. 2022

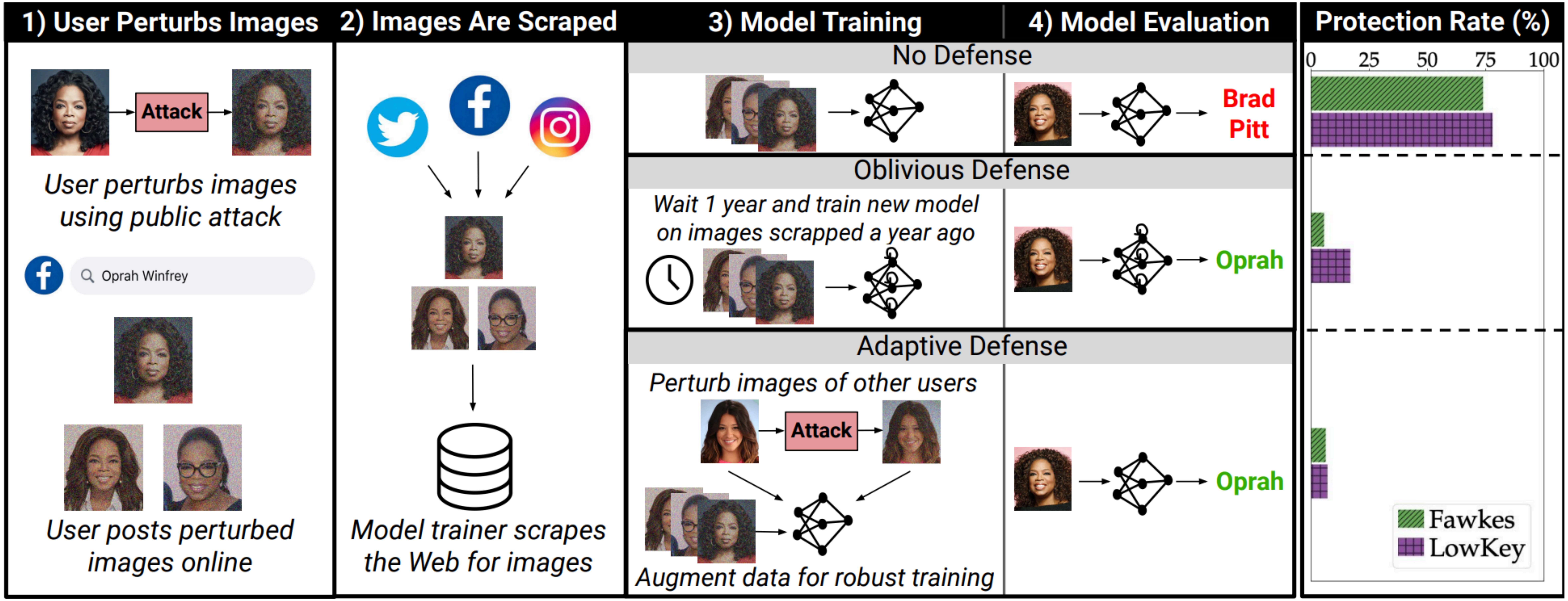

1) User Perturbs Images
Attack
User perturbs images using public attack
Oprah Winfrey
User posts perturbed images online



1) User Perturbs Images
Attack
User perturbs images using public attack
Oprah Winfrey
User posts perturbed images online
2) Images Are Scraped
Model trainer scrapes the Web for images


1) User Perturbs Images
Attack
User perturbs images using public attack
Oprah Winfrey
User posts perturbed images online
2) Images Are Scraped
Model trainer scrapes the Web for images
3) Model Training
4) Model Evaluation
No Defense
Brad Pitt
Protection Rate (%)
0
25
50
75
100


1) User Perturbs Images
Attack
User perturbs images using public attack
Oprah Winfrey
User posts perturbed images online
2) Images Are Scraped
Model trainer scrapes the Web for images
3) Model Training
4) Model Evaluation
No Defense
Brad Pitt
Oblivious Defense
Wait 1 year and train new model on images scrapped a year ago
Oprah
Protection Rate (%)
0
25
50
75
100


1) User Perturbs Images
Attack
User perturbs images using public attack
Oprah Winfrey
User posts perturbed images online
2) Images Are Scraped
Model trainer scrapes the Web for images
3) Model Training
4) Model Evaluation
No Defense
Brad Pitt
Oblivious Defense
Wait 1 year and train new model on images scrapped a year ago
Oprah
Adaptive Defense
Perturb images of other users
Attack
Augment data for robust training
Oprah
Protection Rate (%)
0
25
50
75
100
Fawkes
LowKey

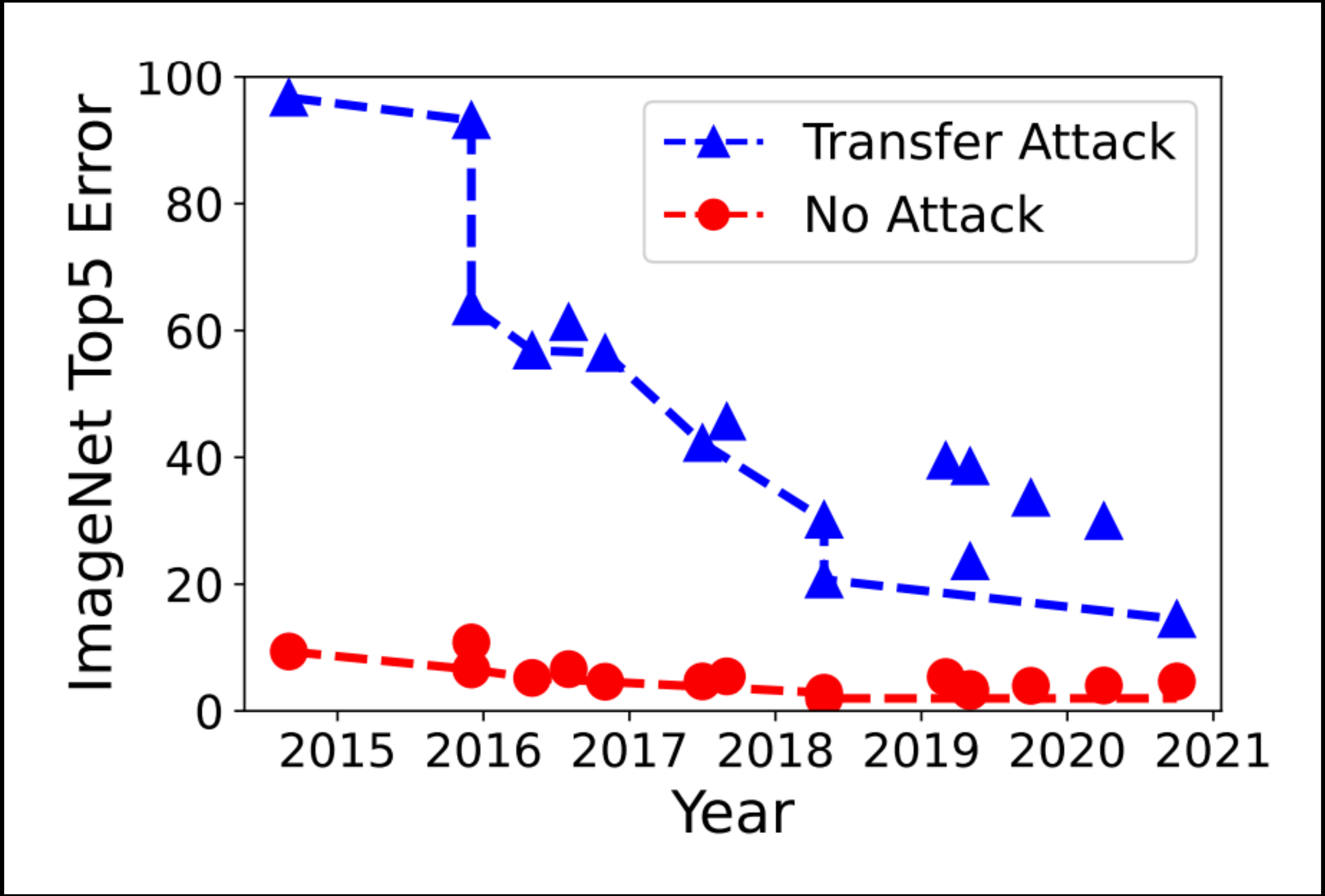
$\forall$ models $\exists$ adversarial examples

$\exists$ **adversarial examples** $\forall$ **models**

~~$\exists$ adversarial examples $\forall$ models~~

Two attacks:

1. just wait



ImageNet Top5 Error
Year
Transfer Attack
No Attack
0
20
40
60
80
100
2015
2016
2017
2018
2019
2020
2021

Two attacks:

1. just wait
2. train a better model

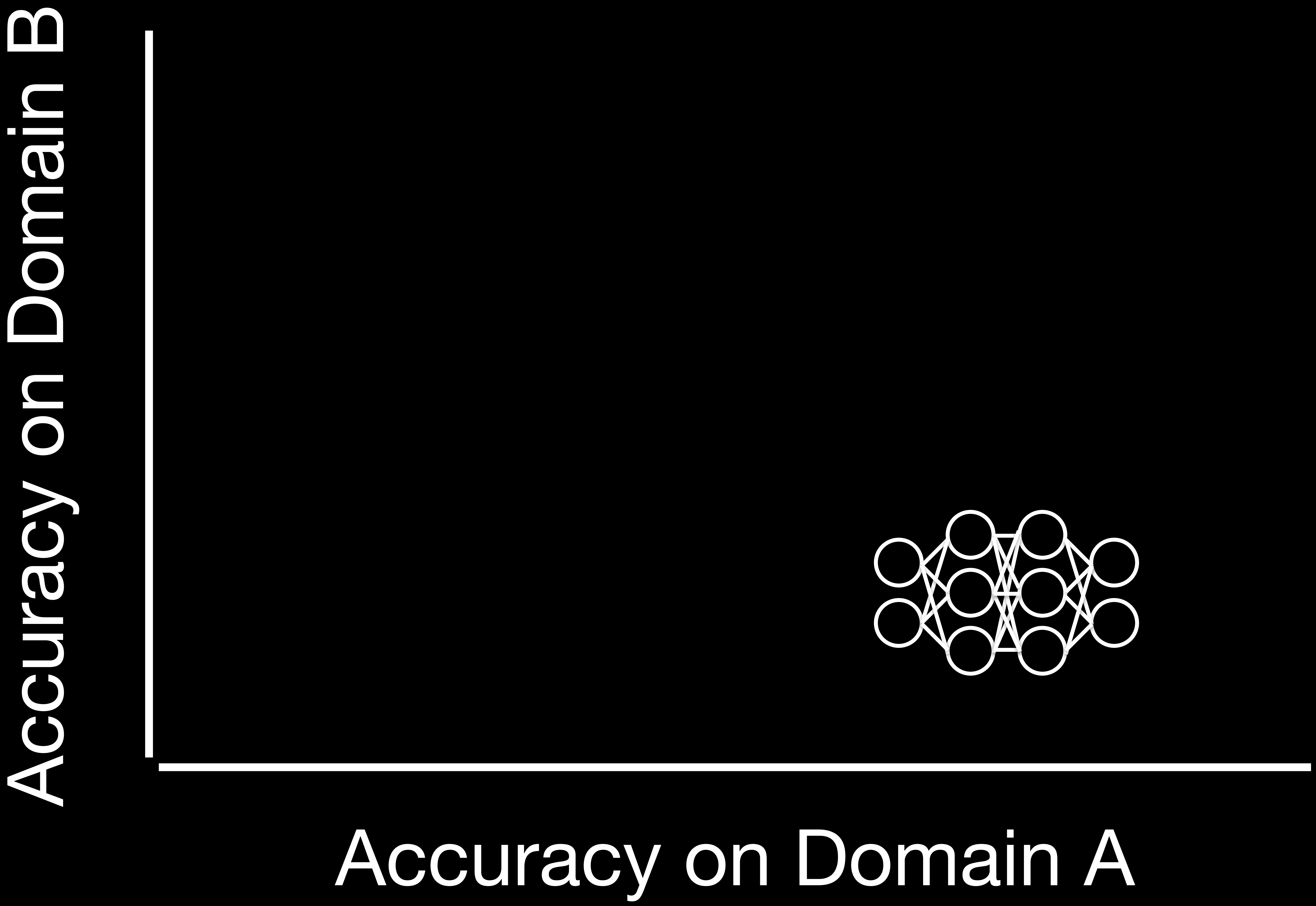
How train a better model?

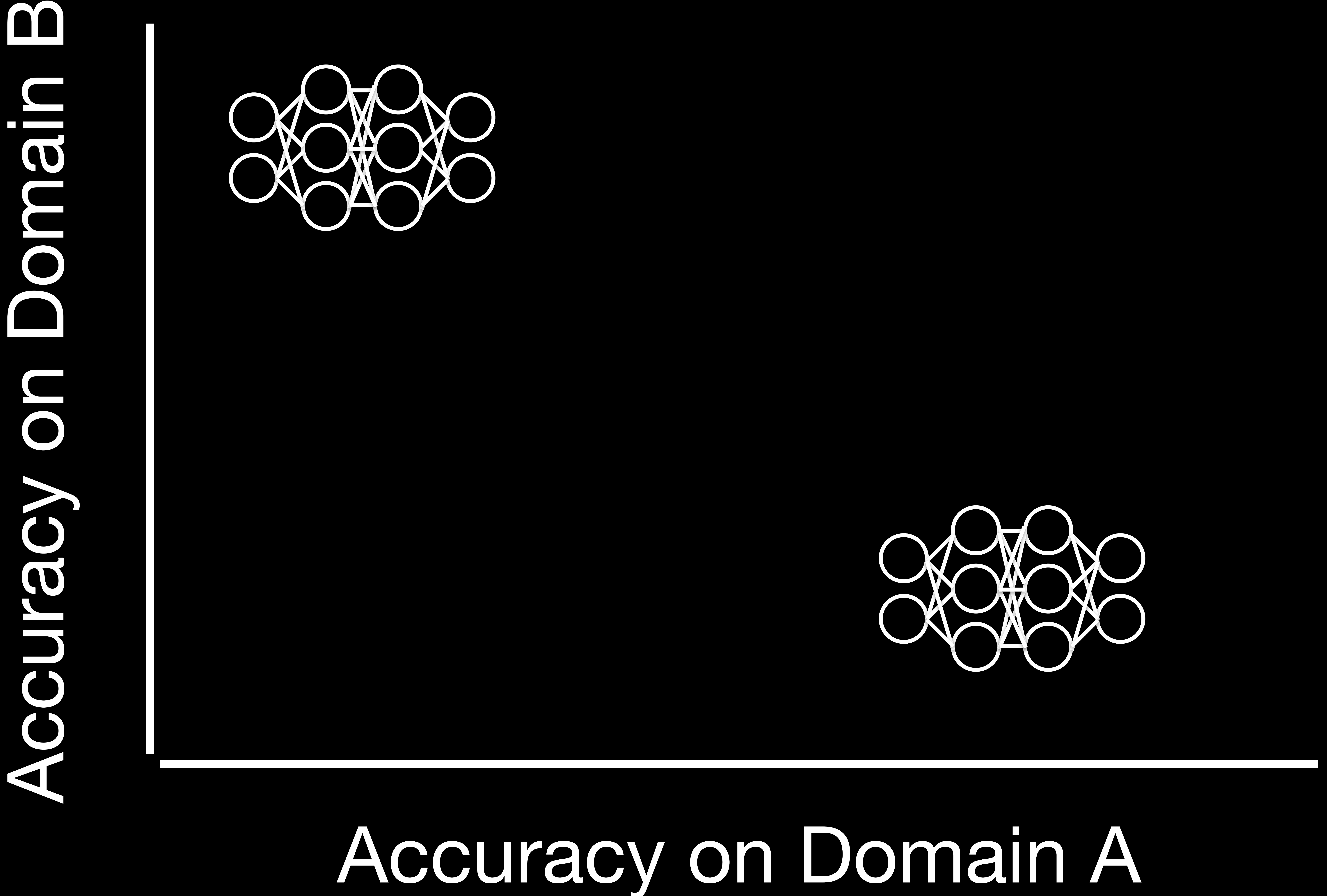
Well ... just train on poisoned images!

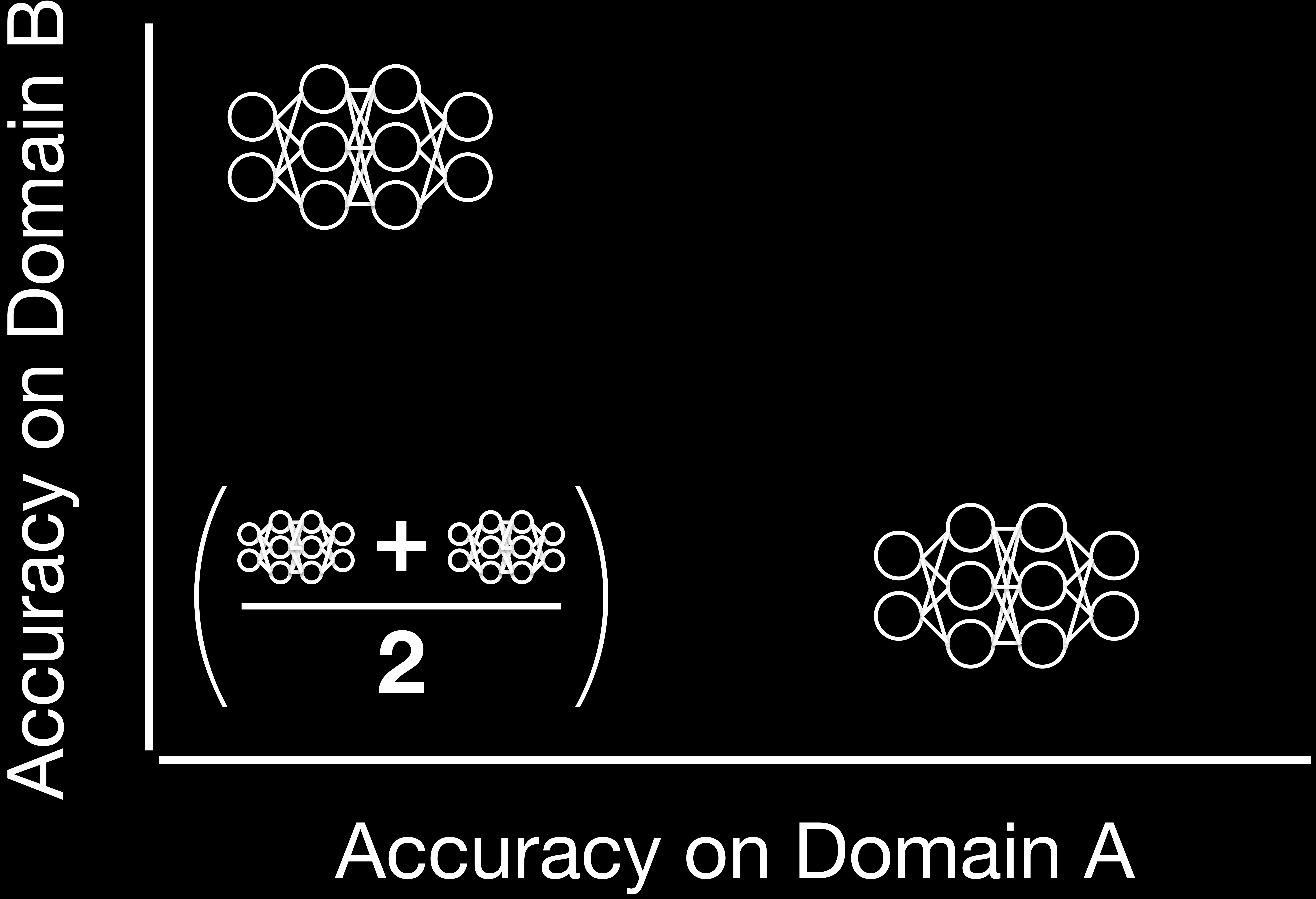
One catch: this causes "clean" accuracy to drop

A fix that shouldn't work:

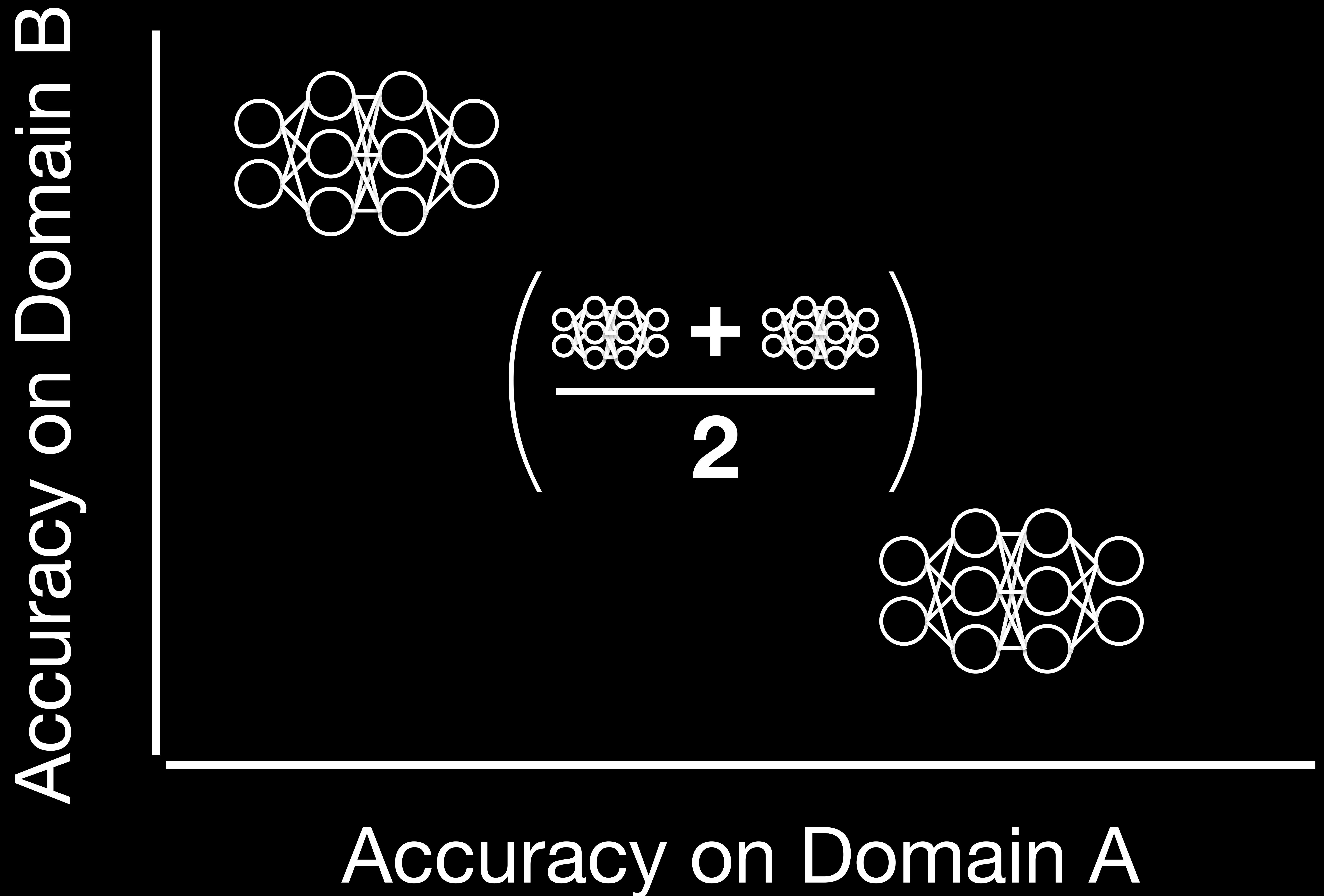

Accuracy on Domain B
Accuracy on Domain A

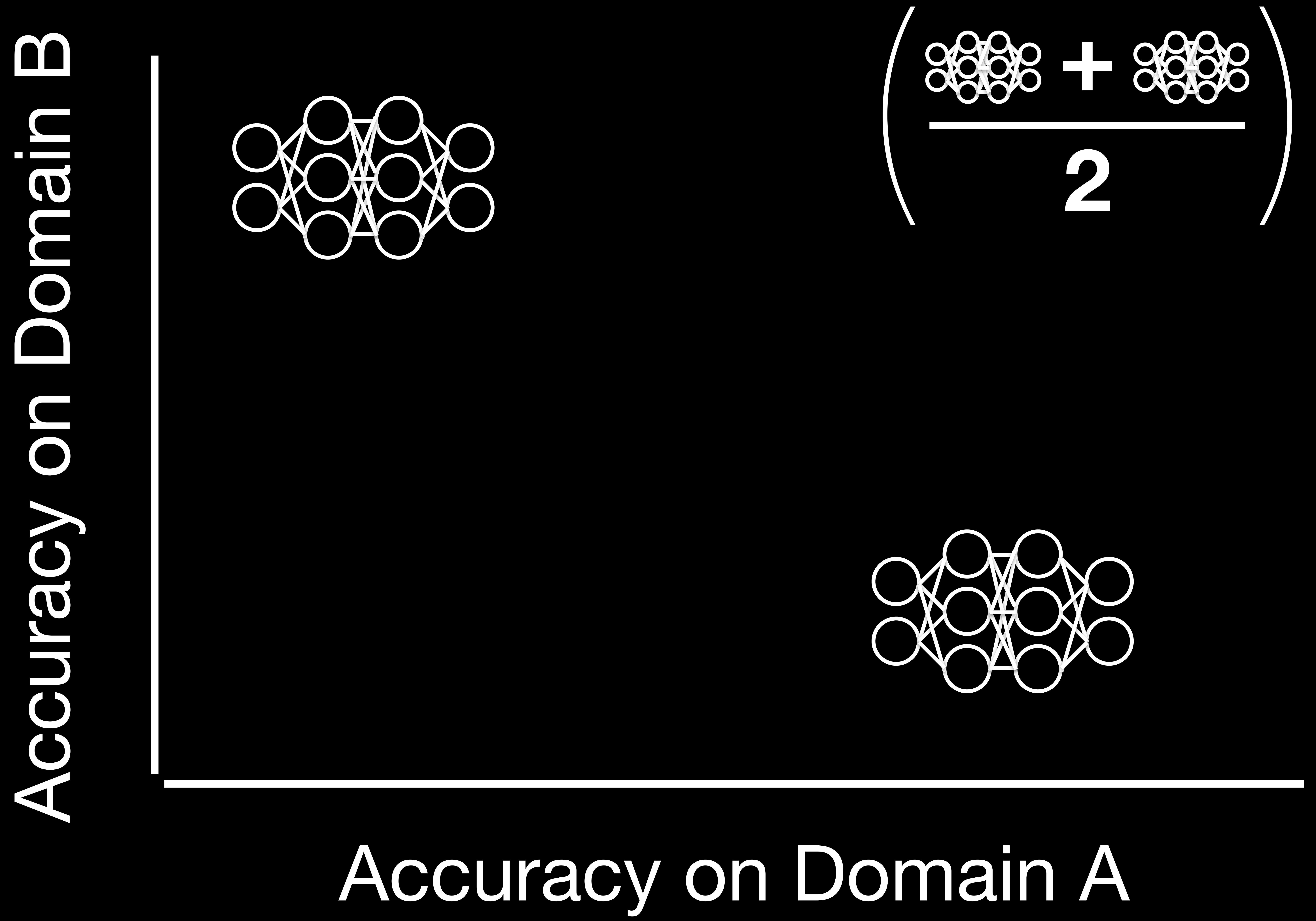

Accuracy on Domain B
Accuracy on Domain A


Accuracy on Domain B
Accuracy on Domain A

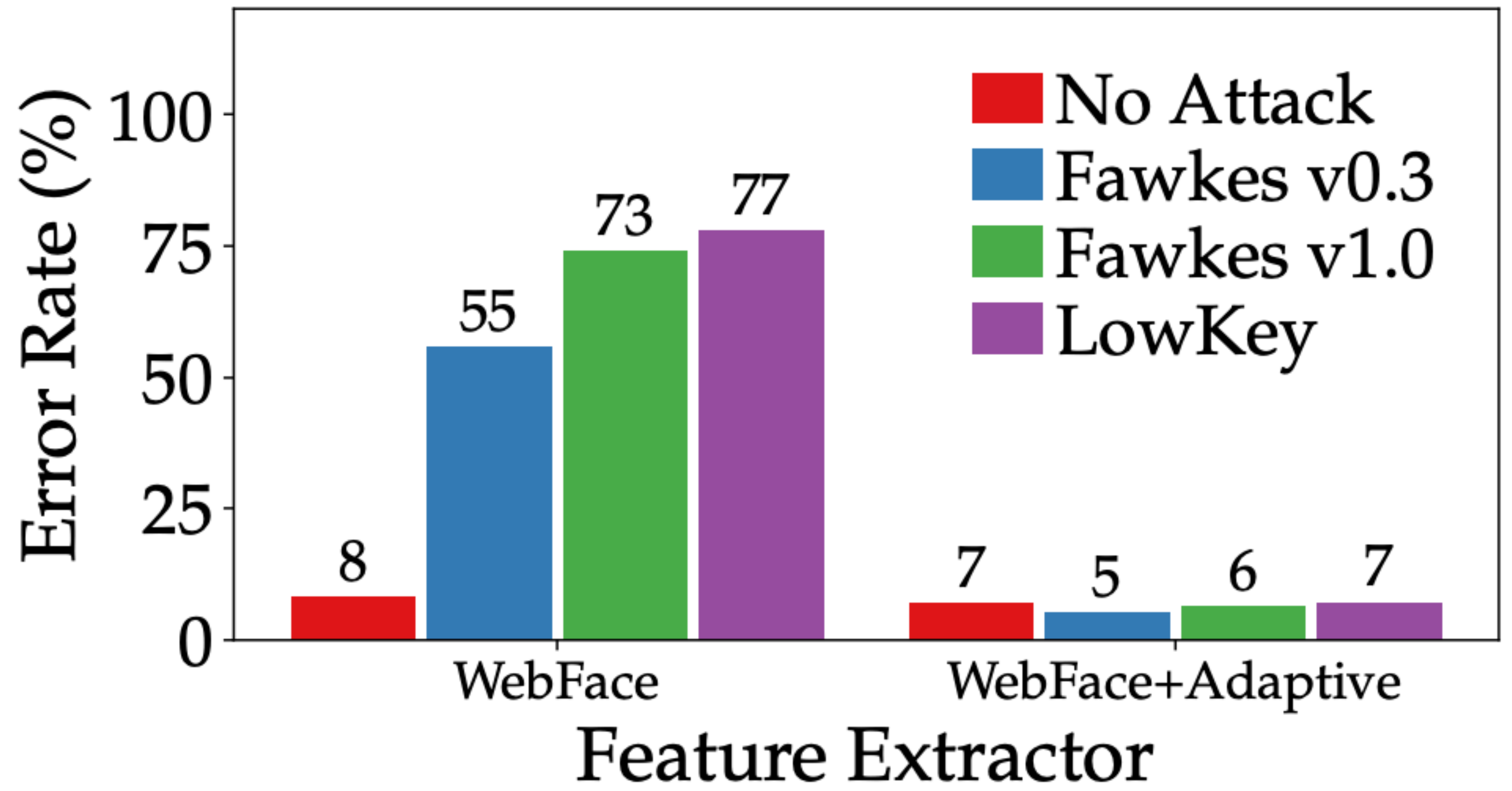


Accuracy on Domain B
+
2
Accuracy on Domain A


Accuracy on Domain B
2
Accuracy on Domain A



Accuracy on Domain B
Accuracy on Domain A
2


Error Rate (%)
100
75
50
25
0
8
55
73
77
7
5
6
7
No Attack
Fawkes v0.3
Fawkes v1.0
LowKey
WebFace
WebFace+Adaptive
Feature Extractor

Conclusion

Conclusion

- You can use training data poisoning to ...
  - backdoor a machine learning model
  - audit a machine learning model
  - increase the vulnerability of models to privacy attacks

Conclusion

- You can use training data poisoning to ...
  - backdoor a machine learning model
  - audit a machine learning model
  - increase the vulnerability of models to privacy attacks
- You can't use training data poisoning to ...
  - protect users from face recognition

Conclusion

- You can use training data poisoning to ...
  - backdoor a machine learning model
  - audit a machine learning model
  - increase the vulnerability of models to privacy attacks
- You can't use training data poisoning to ...
  - protect users from face recognition
  - solve wold hunger, world peace, cure covid, write good keynote talks

Thank you
for sticking with it