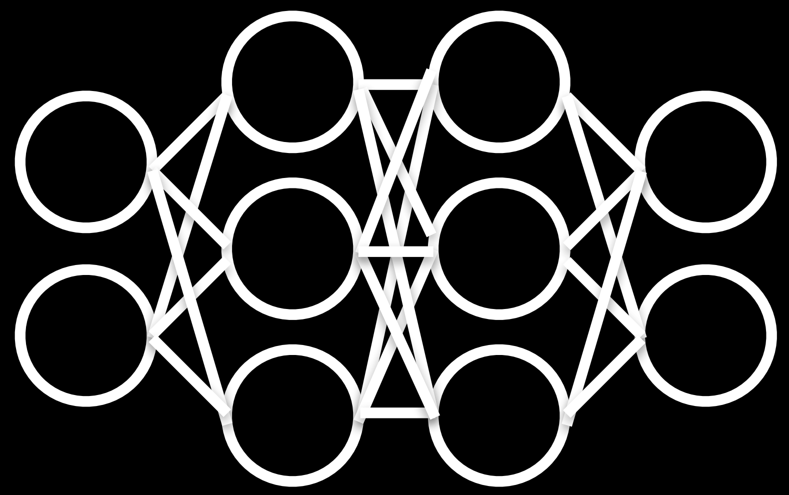
Membership Inference Attacks from First Principles

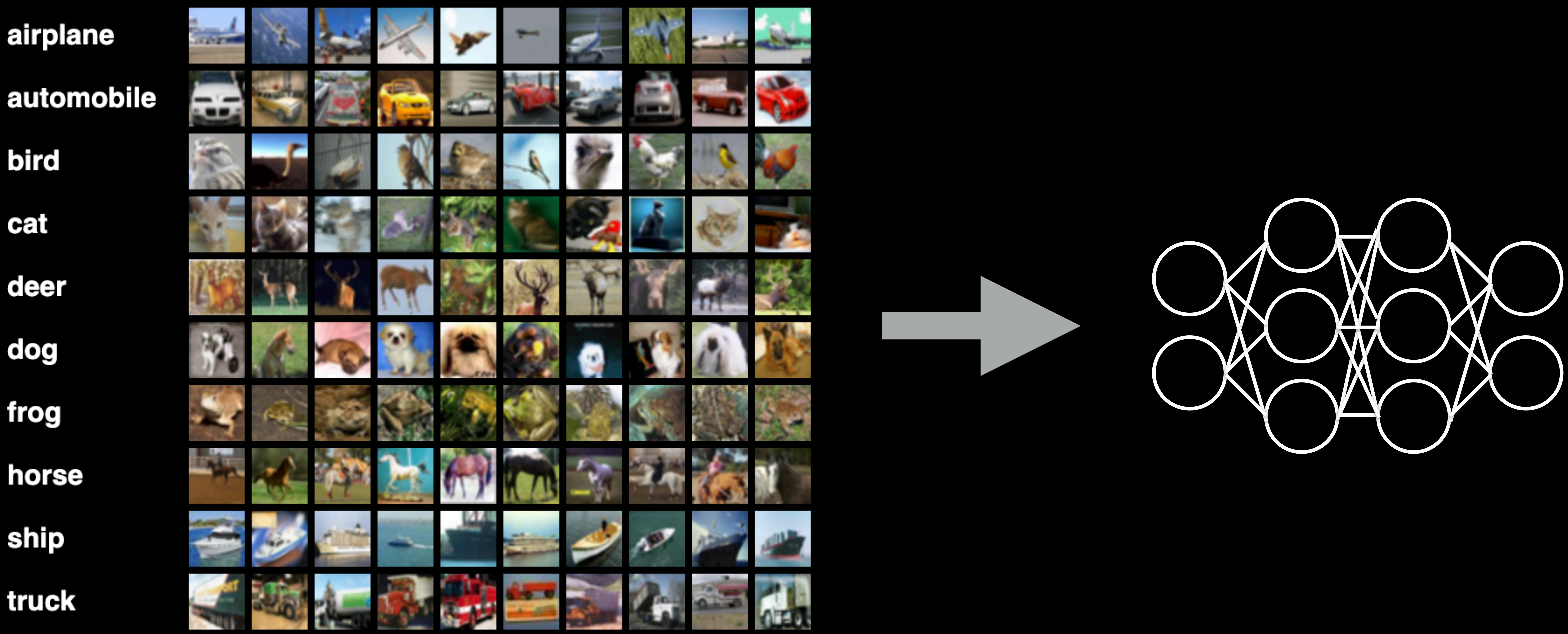
Nicholas Carlini[1], Steve Chien[1], Milad Nasr[12],
Shuang Song[1], Andreas Terzis[1], and Florian Tramer[1]

[1]Google Research   [2]University of Massachusetts Amherst

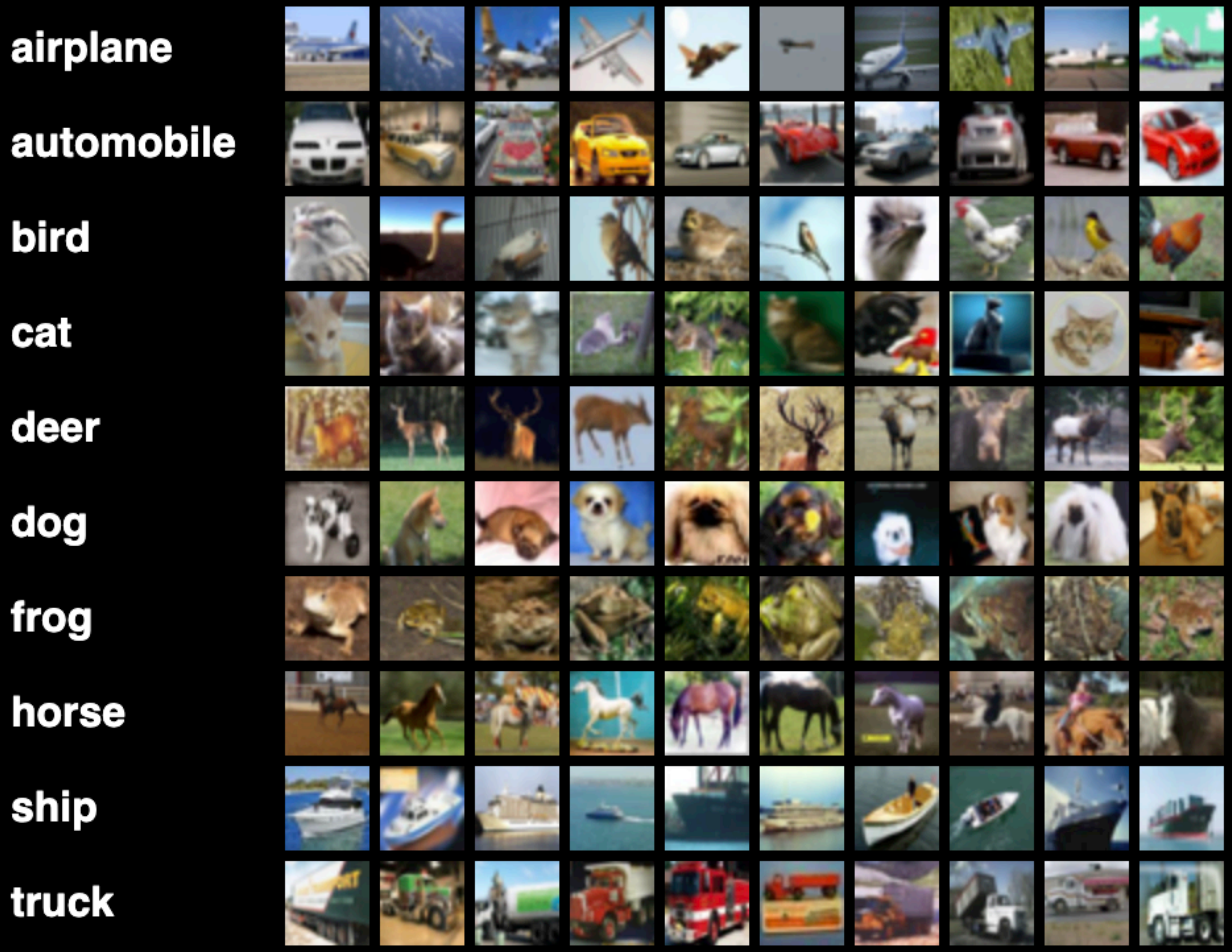
Membership Inference:

Was  trained on the example ?

Review: Model Training

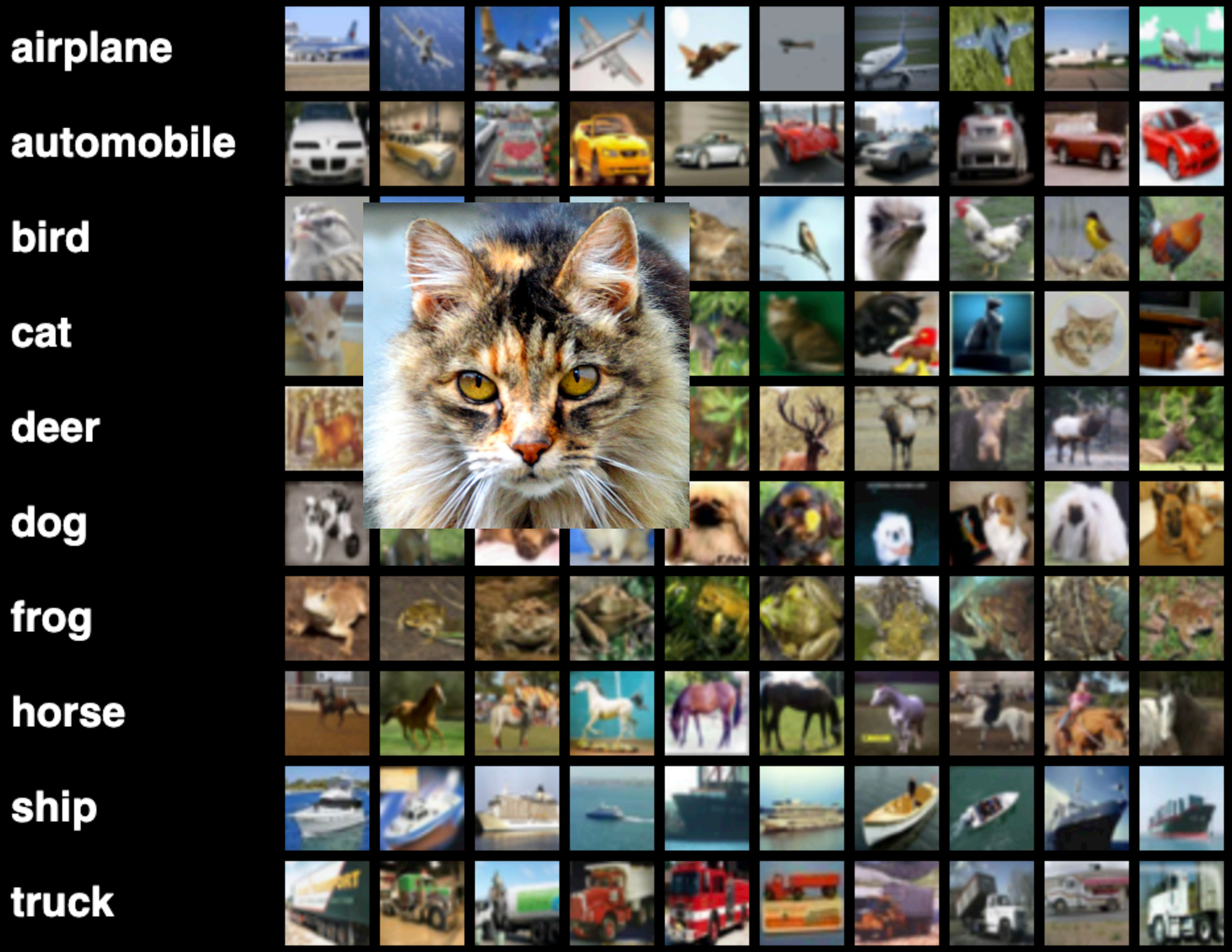


Review: Model Training

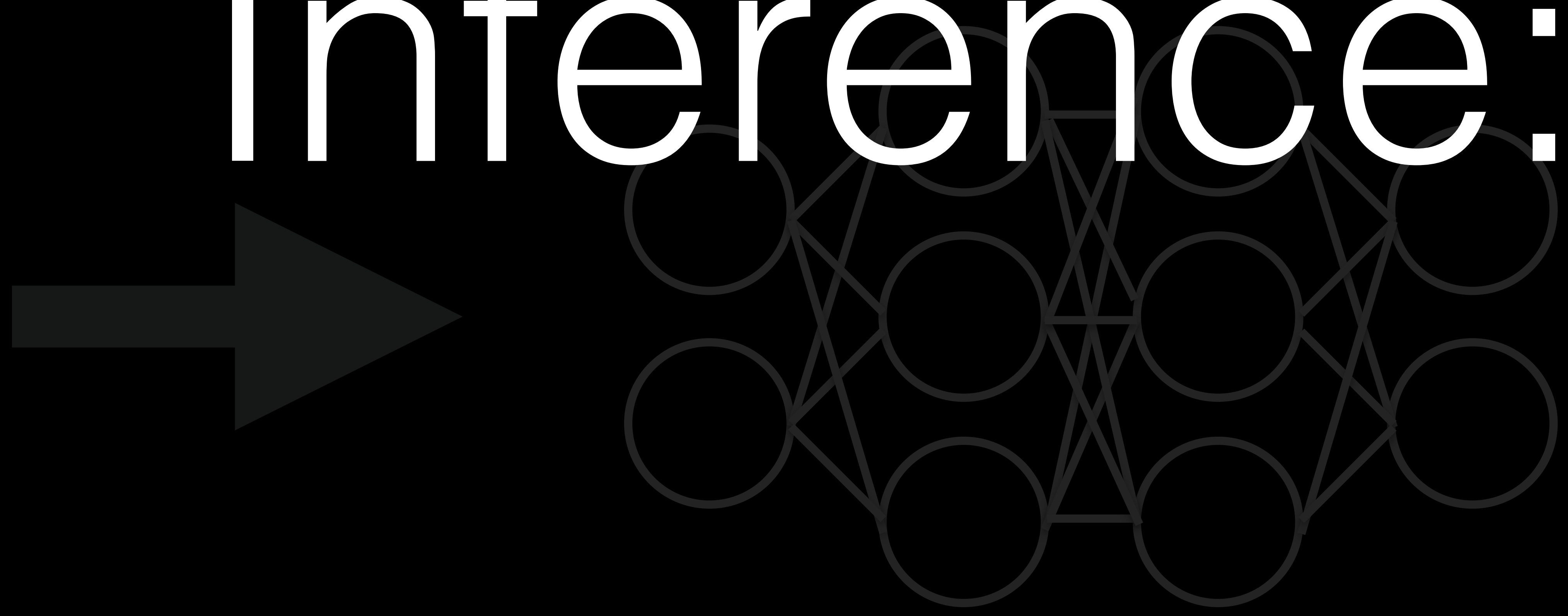


Membership Inference:

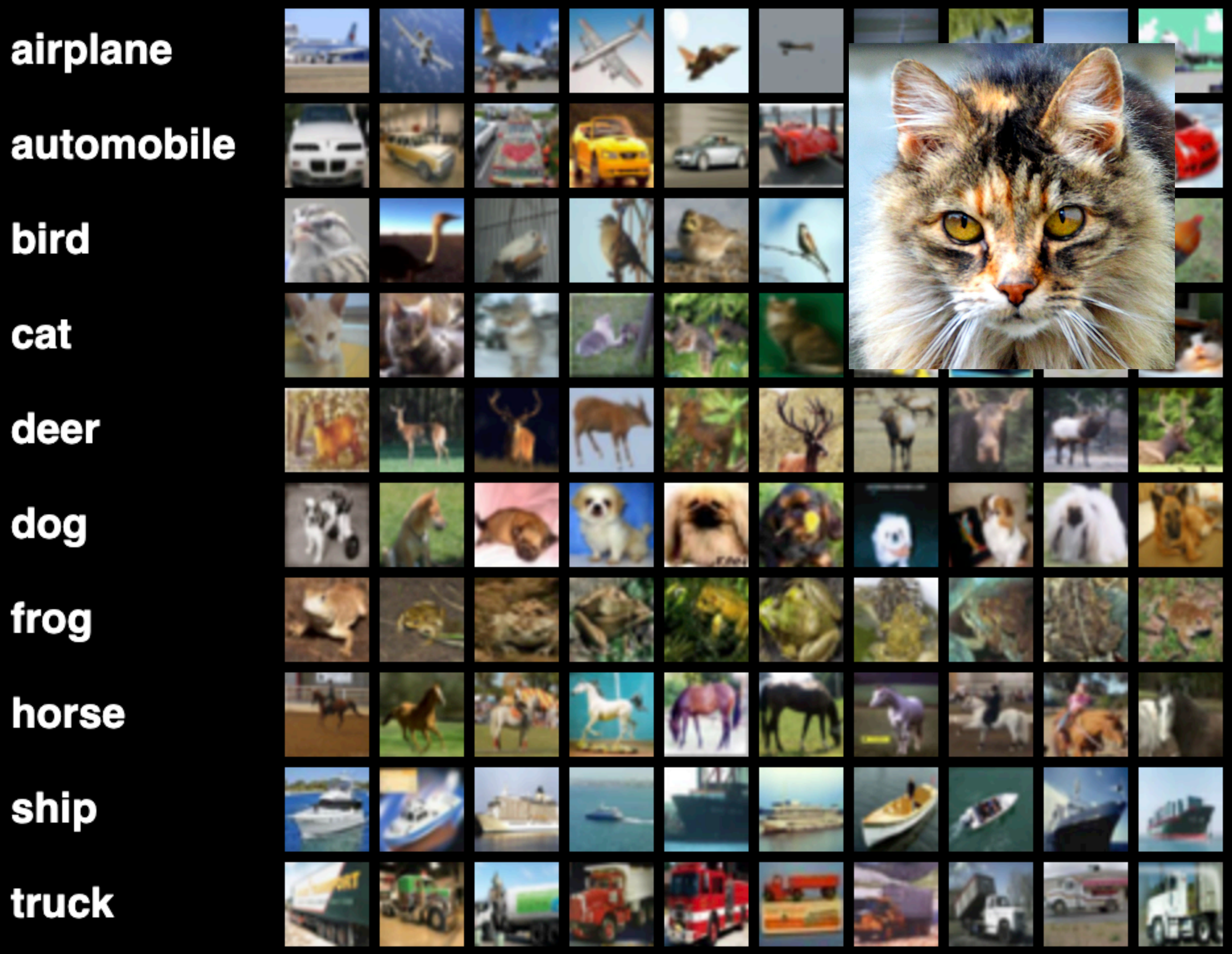
Review: Model Training



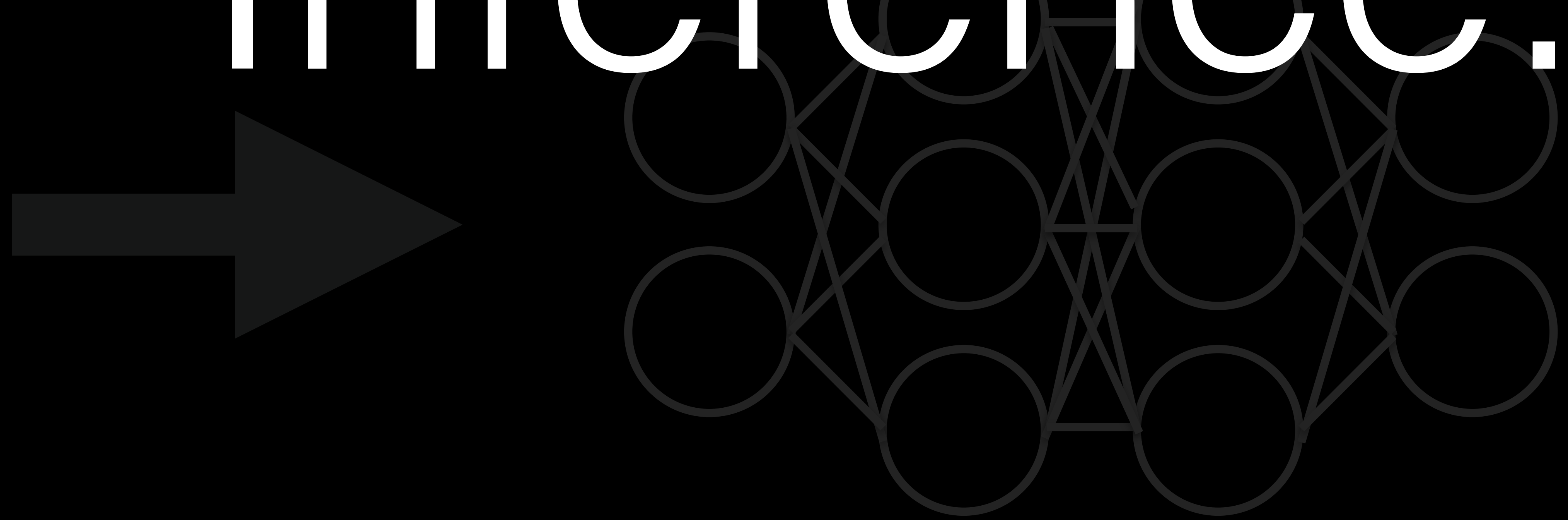
Membership Inference:



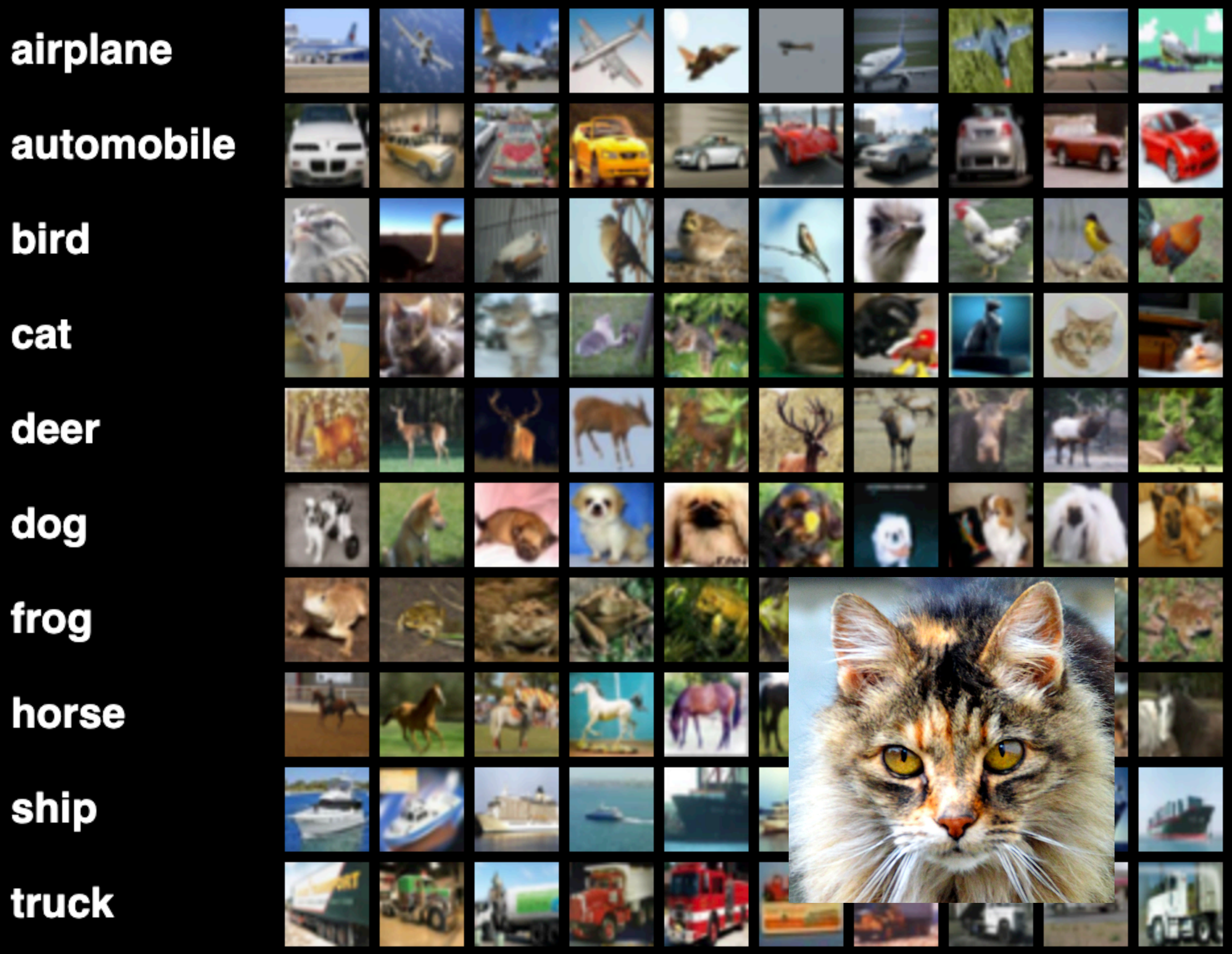
Review: Model Training



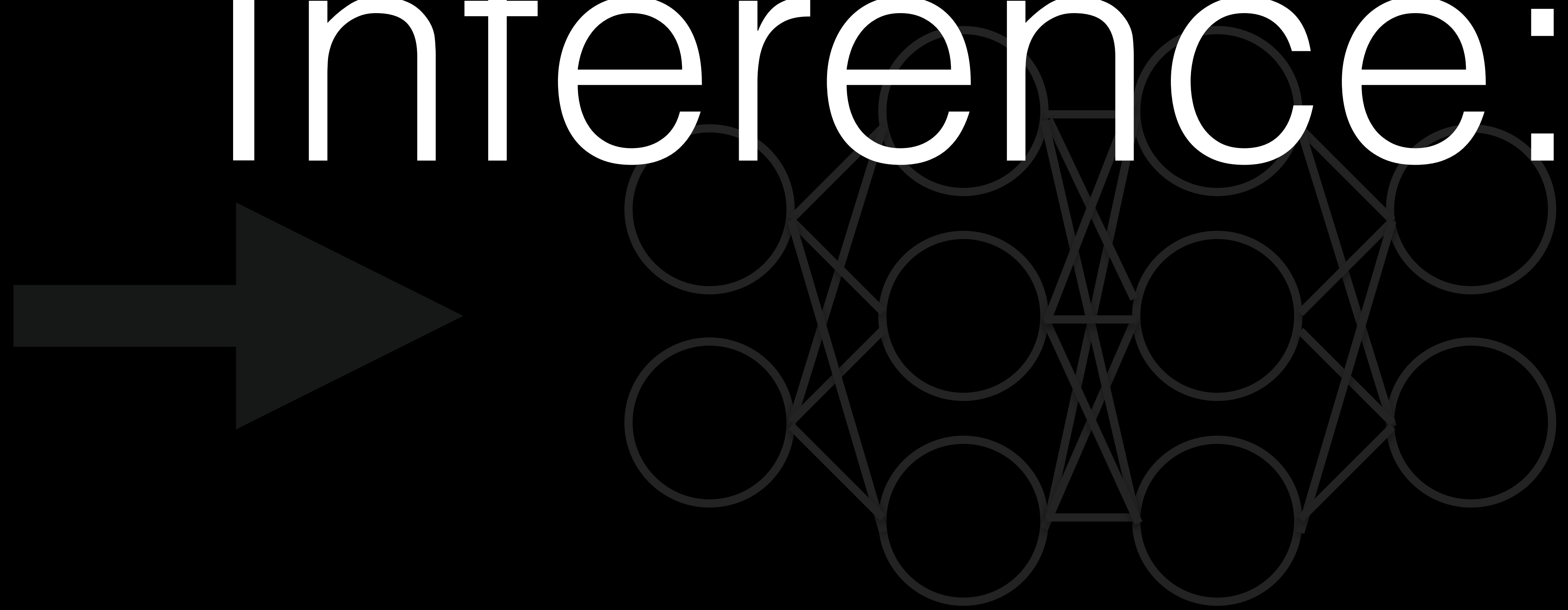
Membership Inference:



Review: Model Training



Membership Inference:



Why?

Curiosity!

Reconnaissance!

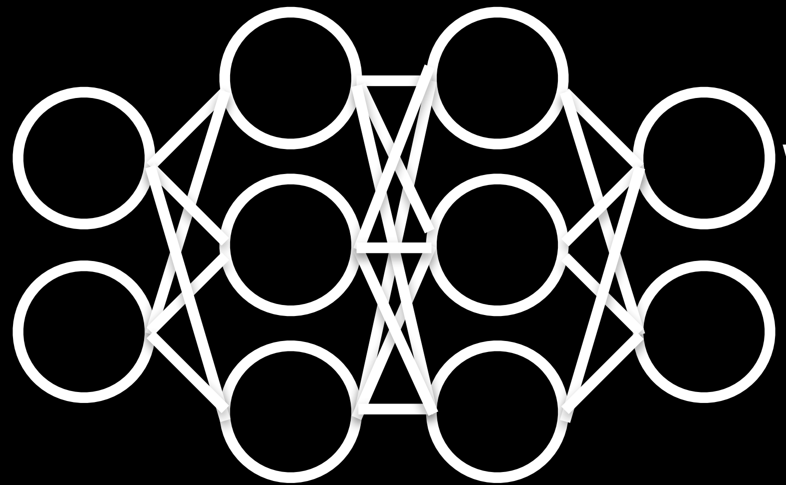
Data Extraction!

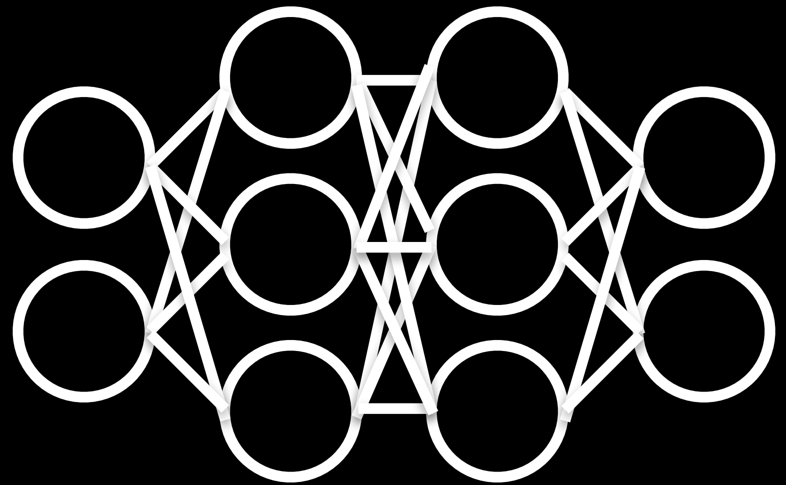
Auditing!

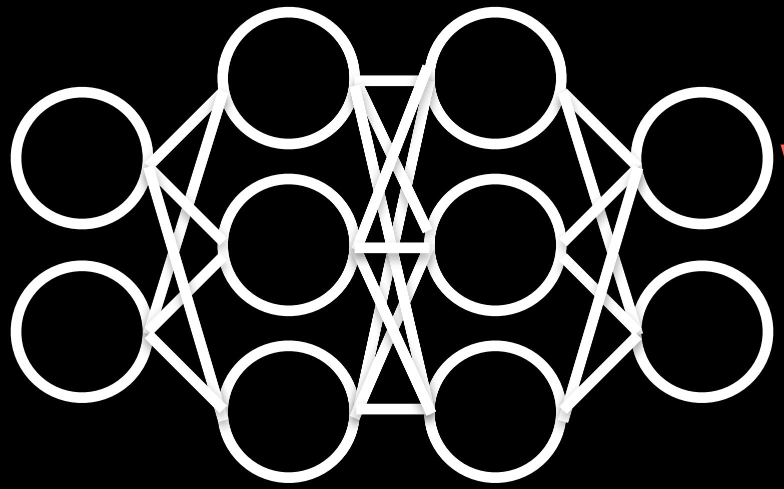
This talk:

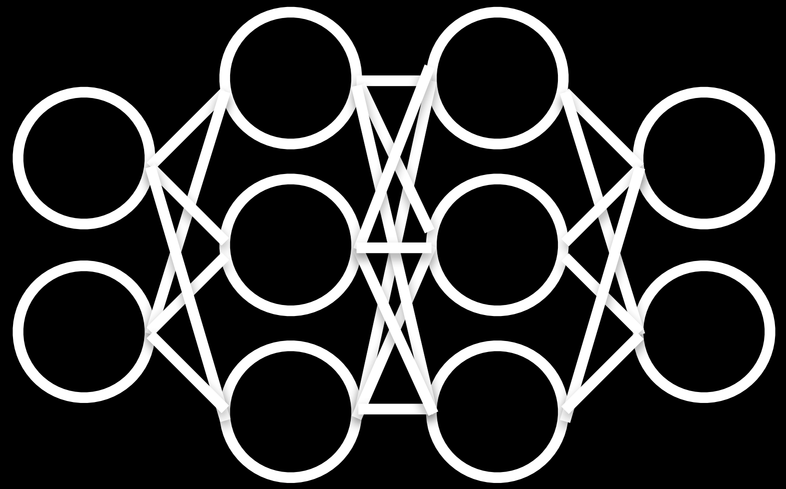
A *first-principles* approach to Membership Inference

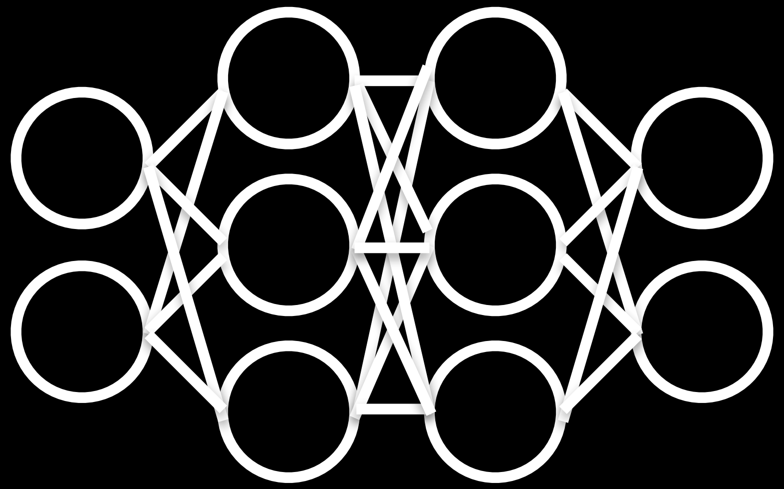
Membership Inference:

A = Pr( **was** trained on)

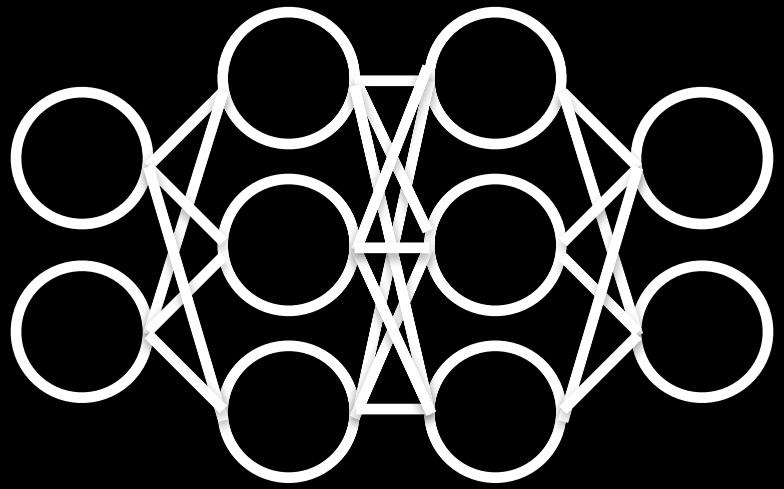
B = Pr( **not** trained on)

A = Pr( **was** trained on)

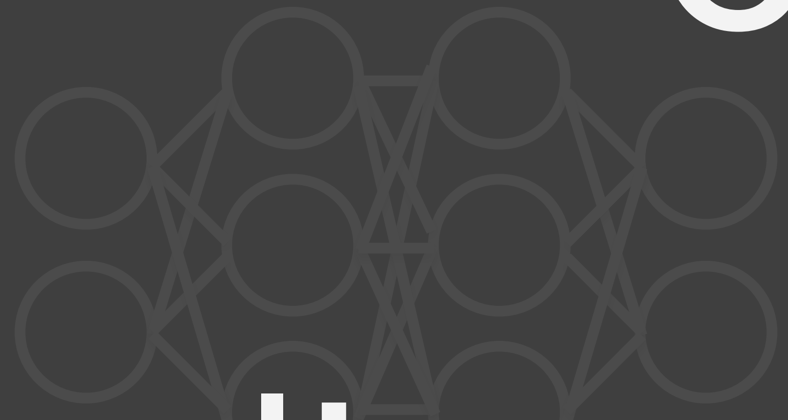
A = Pr( | M( ))

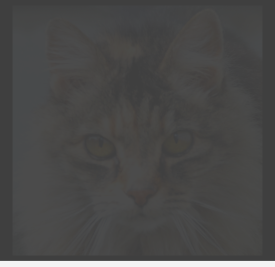
A = Pr( | M( ))

The distribution over models trained on

A = Pr( | M( ))

The distribution over models trained on

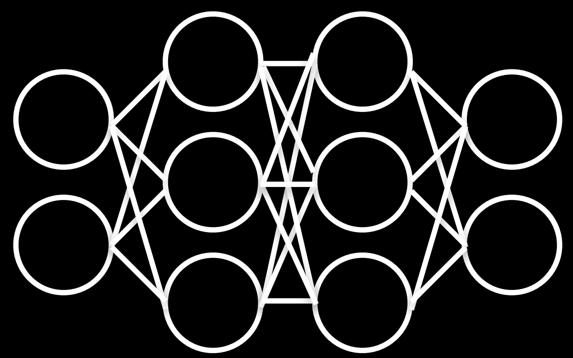
A = Pr( | M( ))

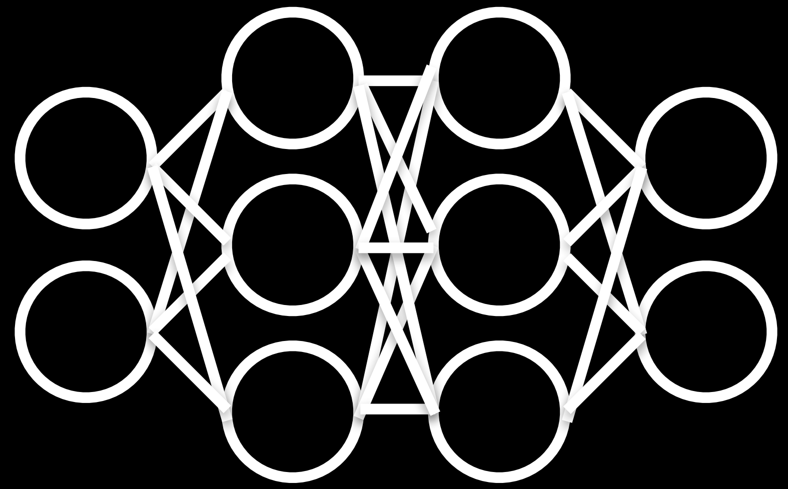
The distribution over models trained on 

Challenge 1:

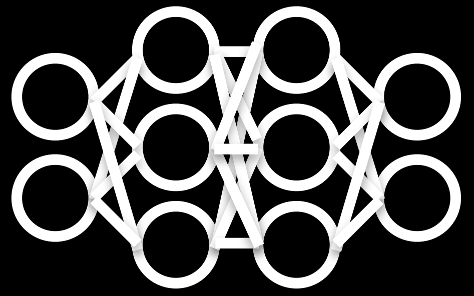
We can't enumerate the distribution over all models parameters.

One bit of background:

Loss( ) measures how "wrong" the model is

A = Pr( | M( ))

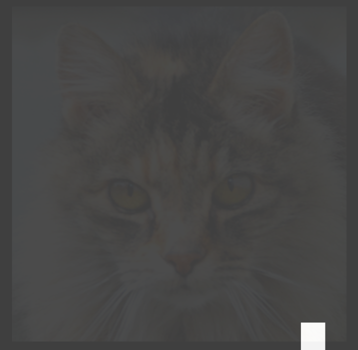
The distribution over models trained on

A = Pr(Loss( ) | L( ))

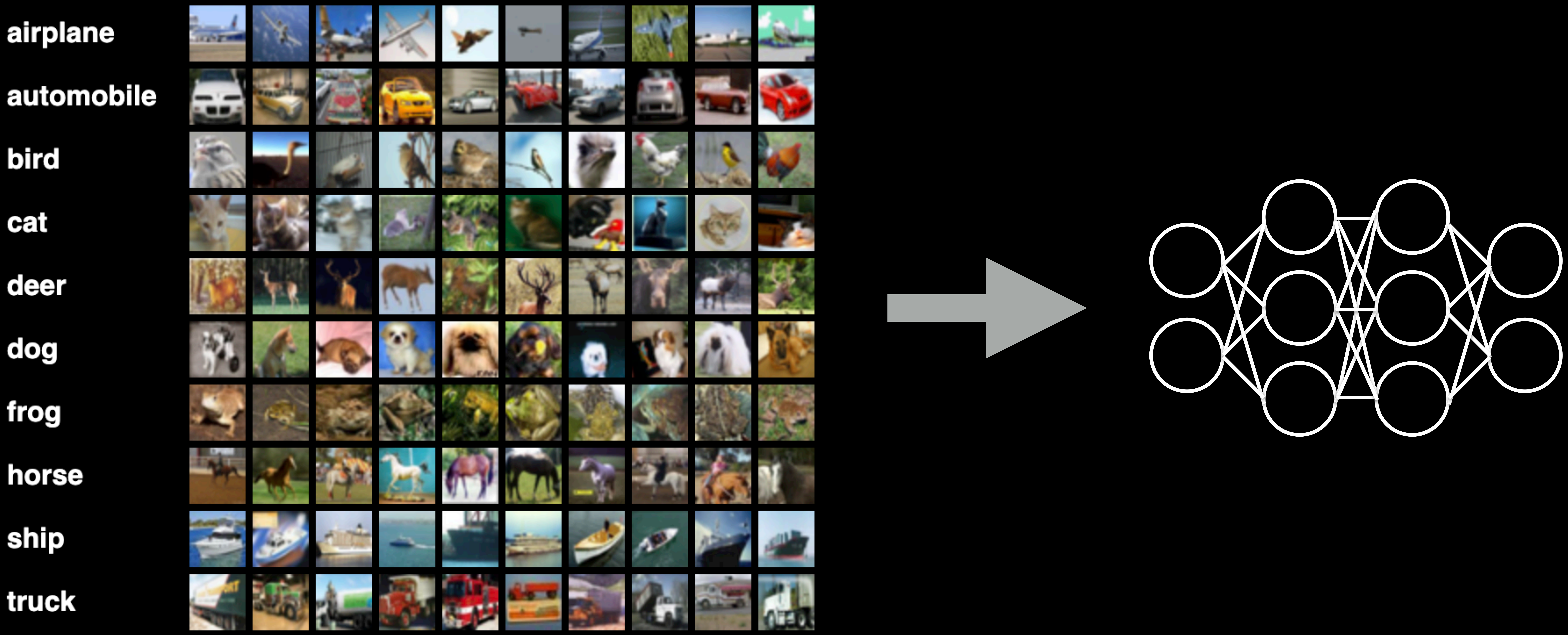
The distribution over **losses**
of models trained on

Challenge 2:

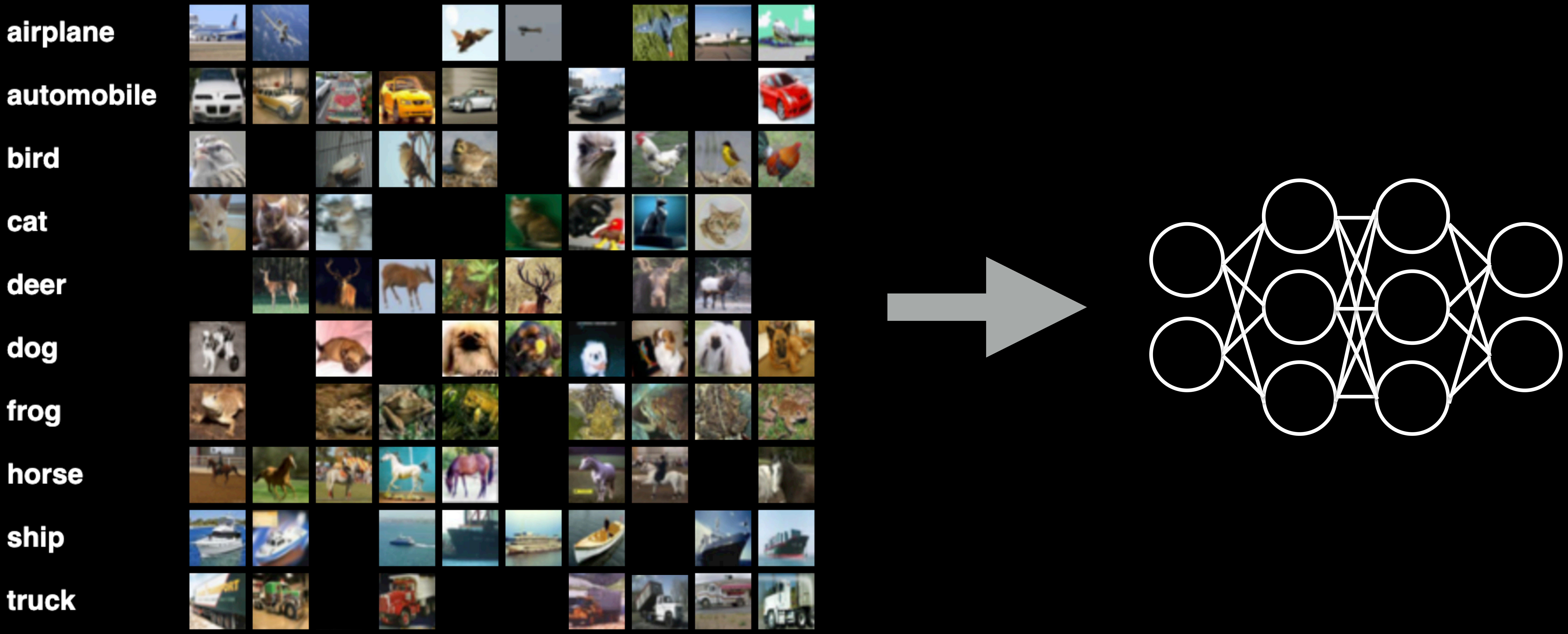
Can we compute the distribution of model losses?

A = Pr(Loss () | L( ))

The distribution over **losses** of models trained on

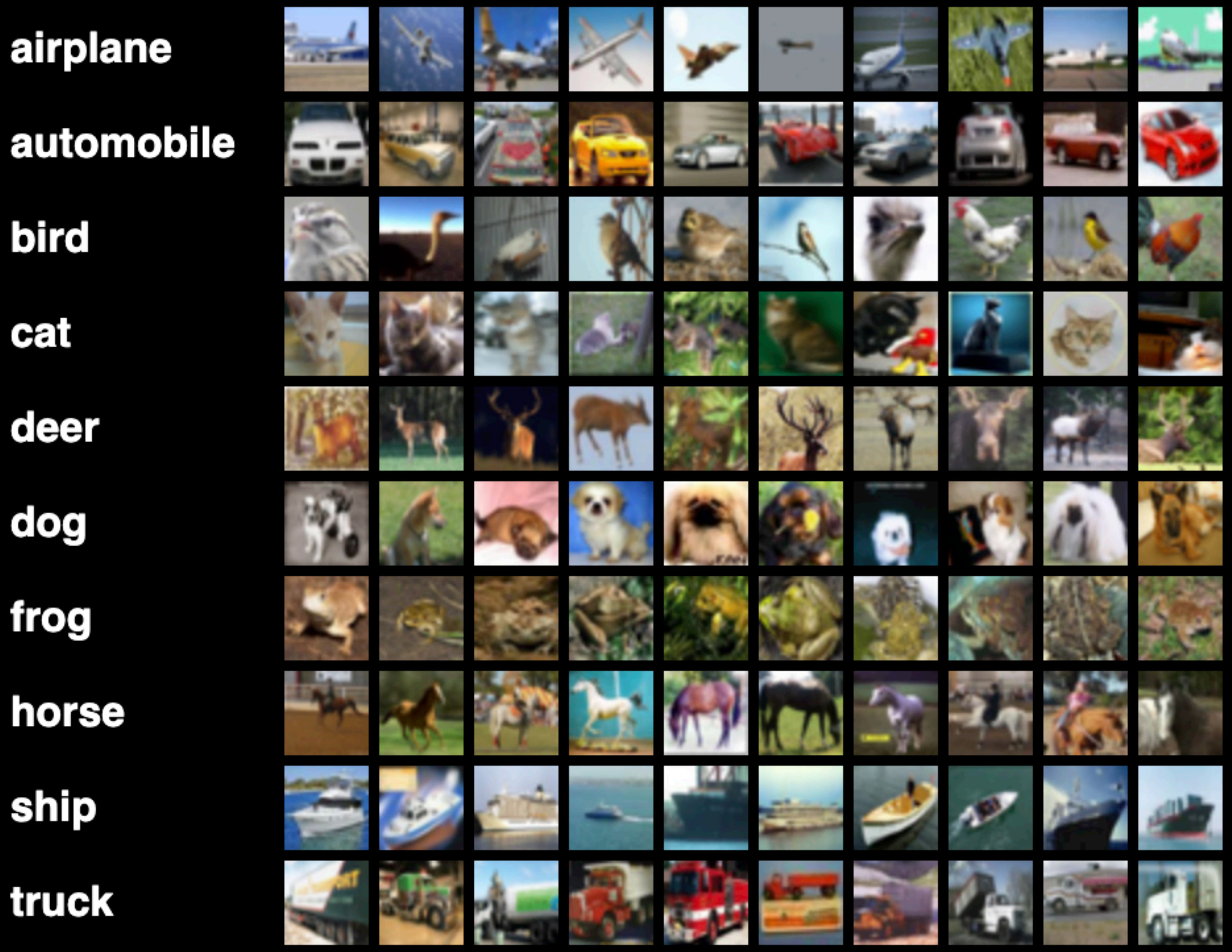


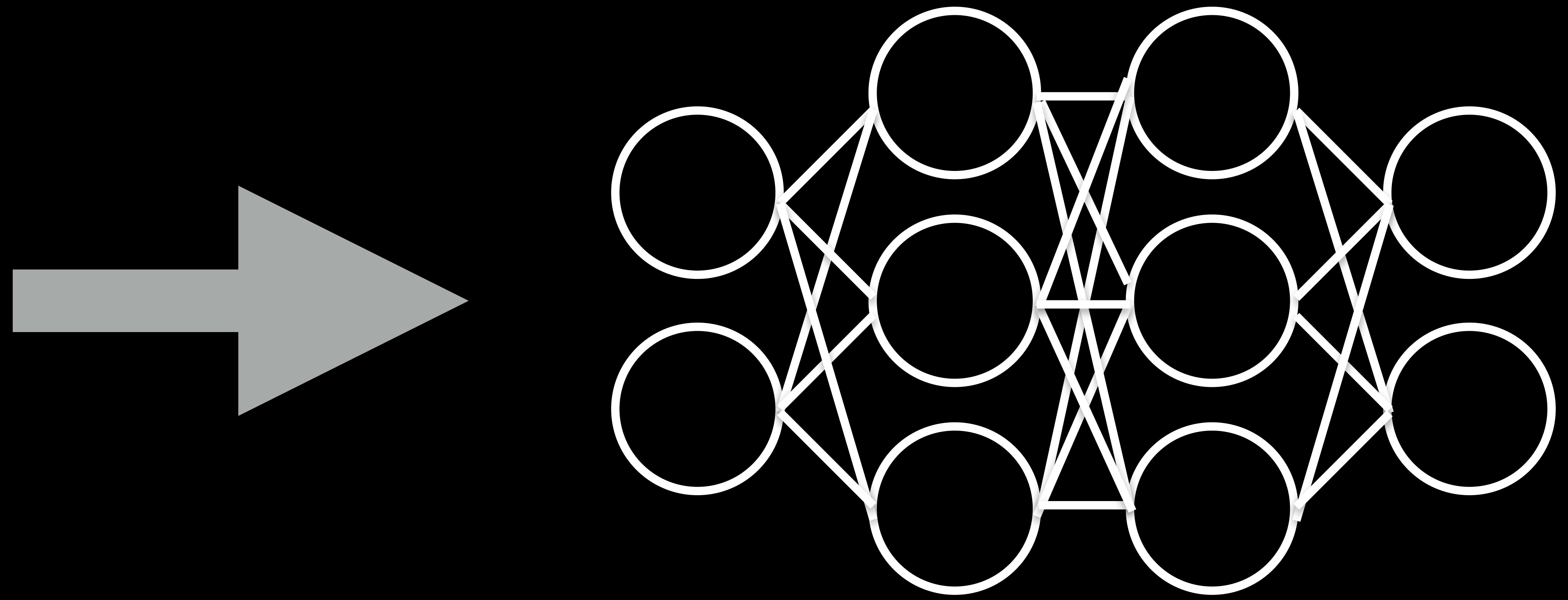
Feldman & Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. 2020.

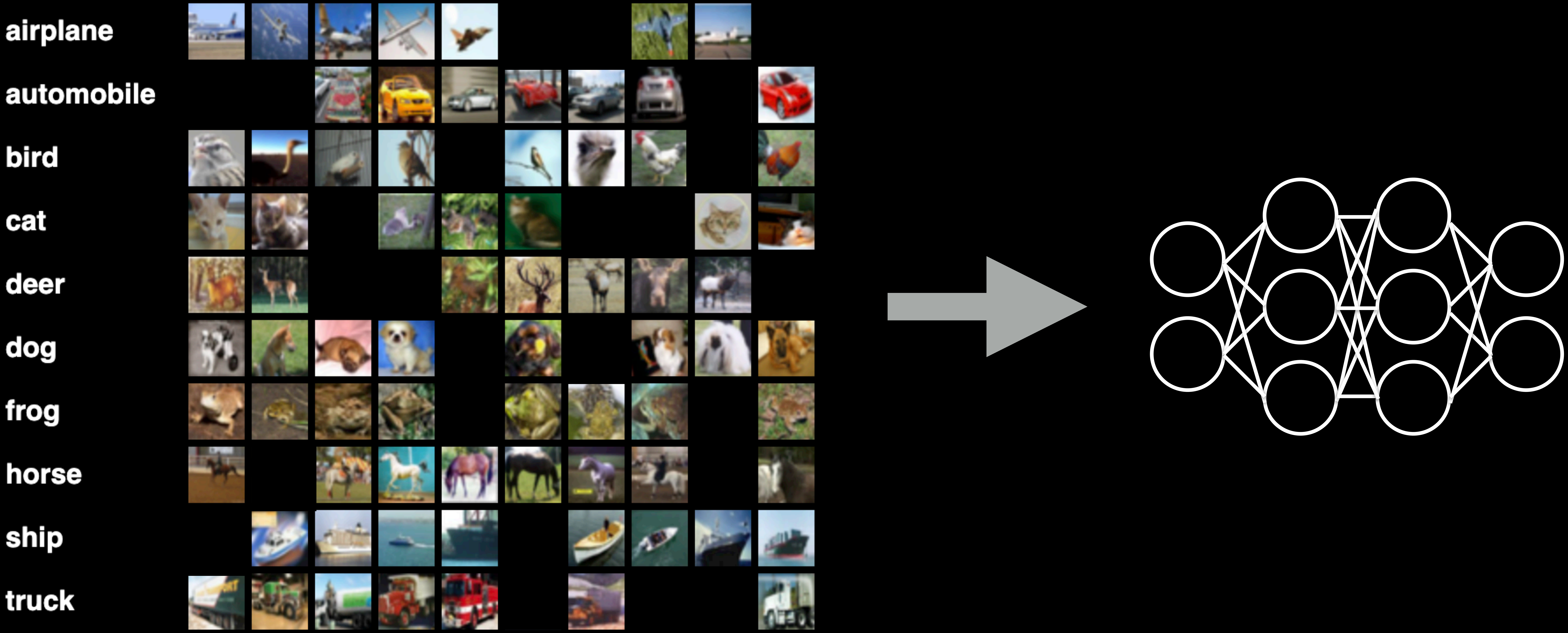


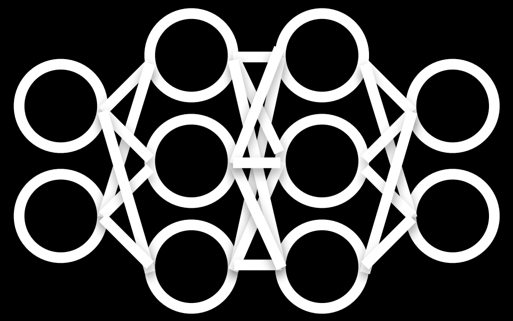
Feldman & Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. 2020.

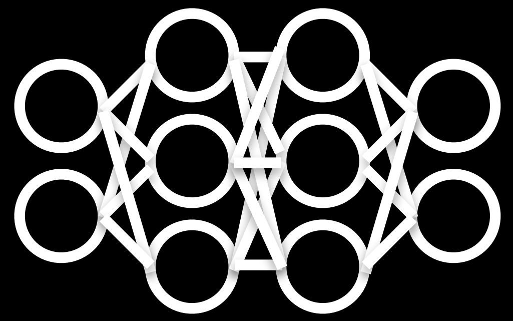
Feldman & Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. 2020.

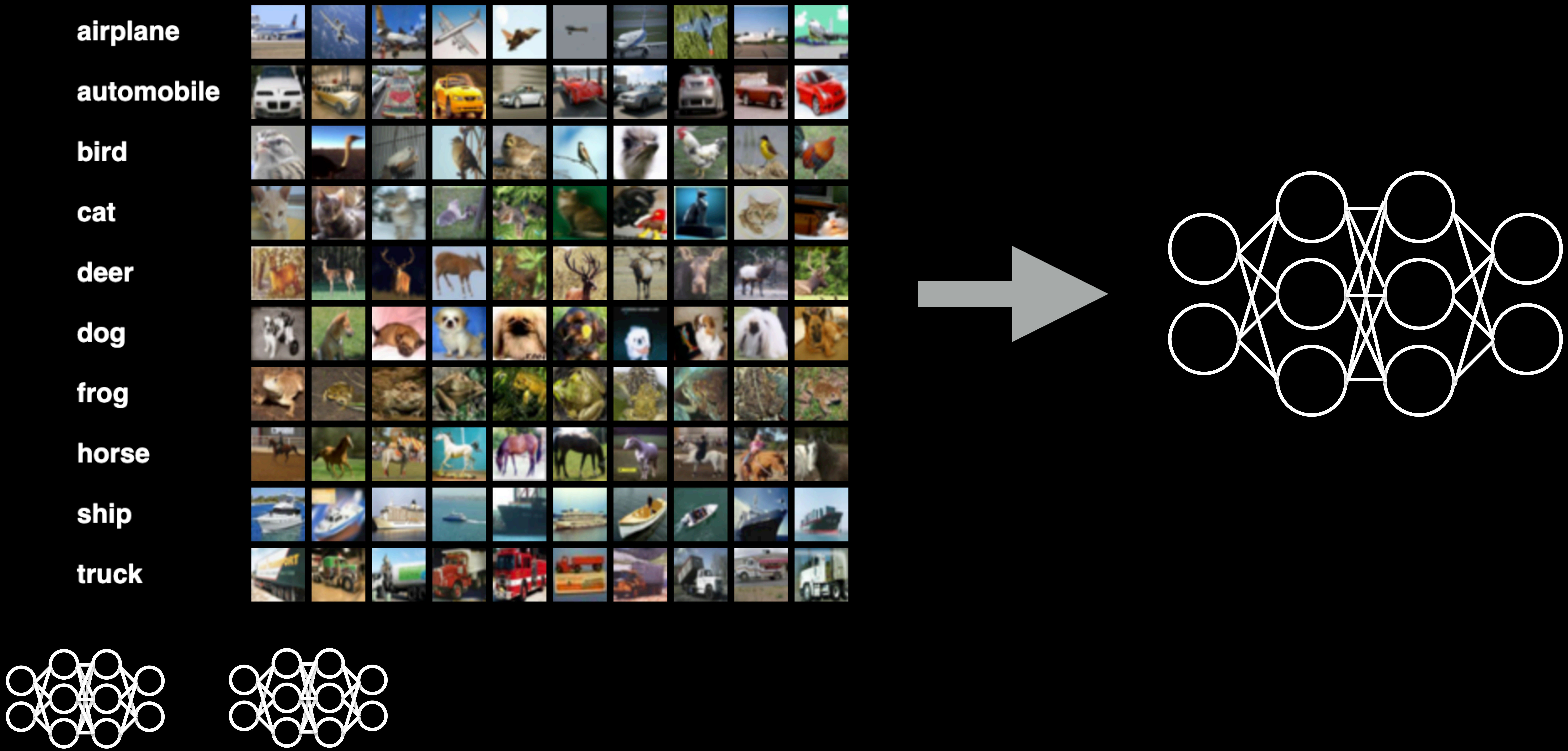

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

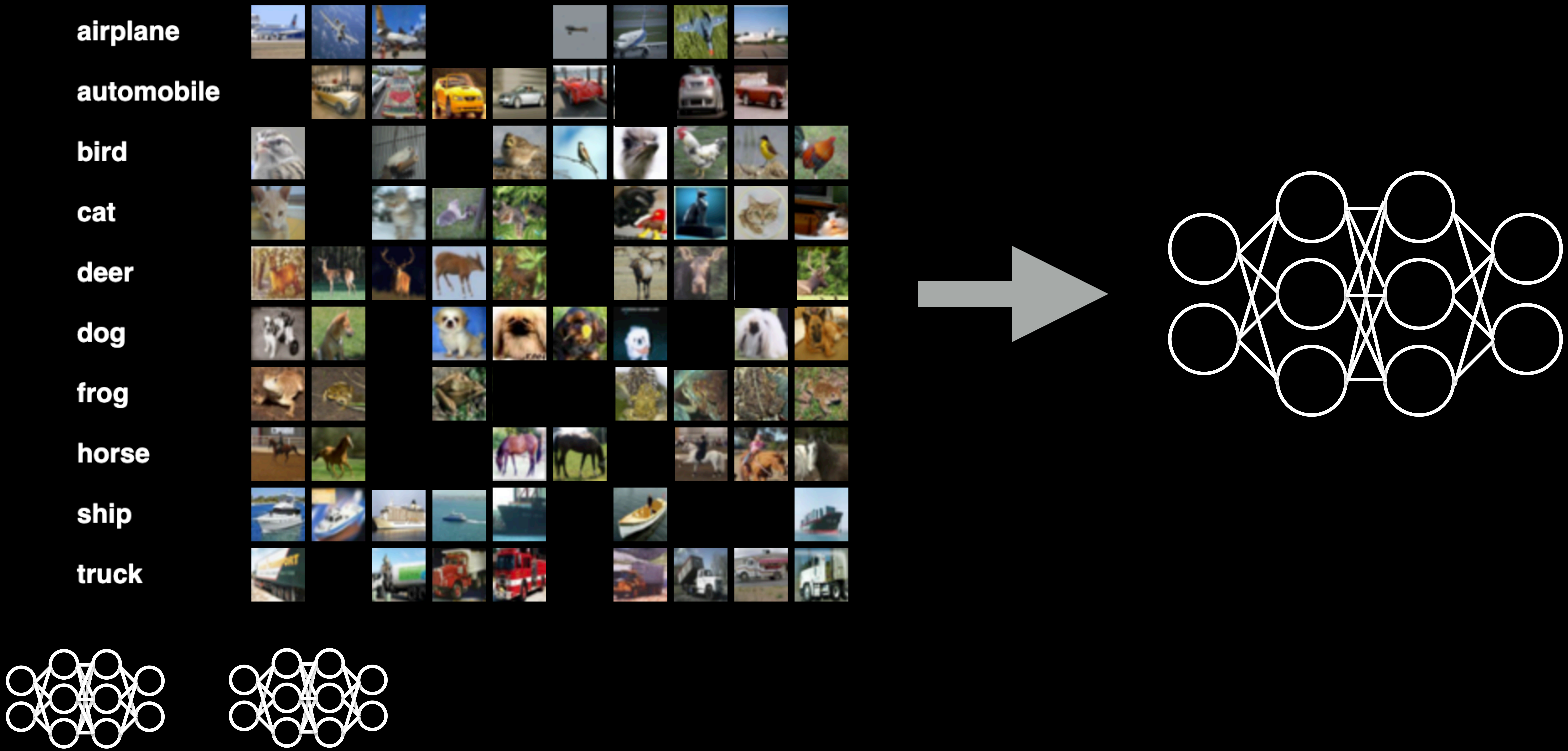


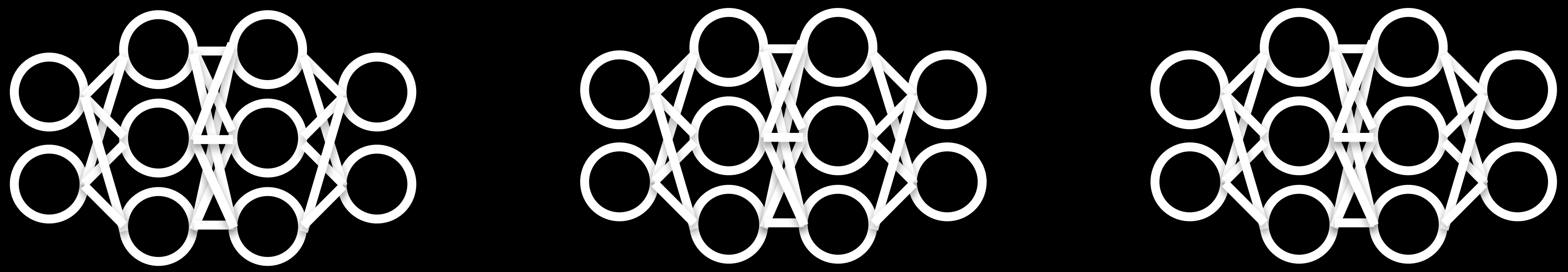

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck



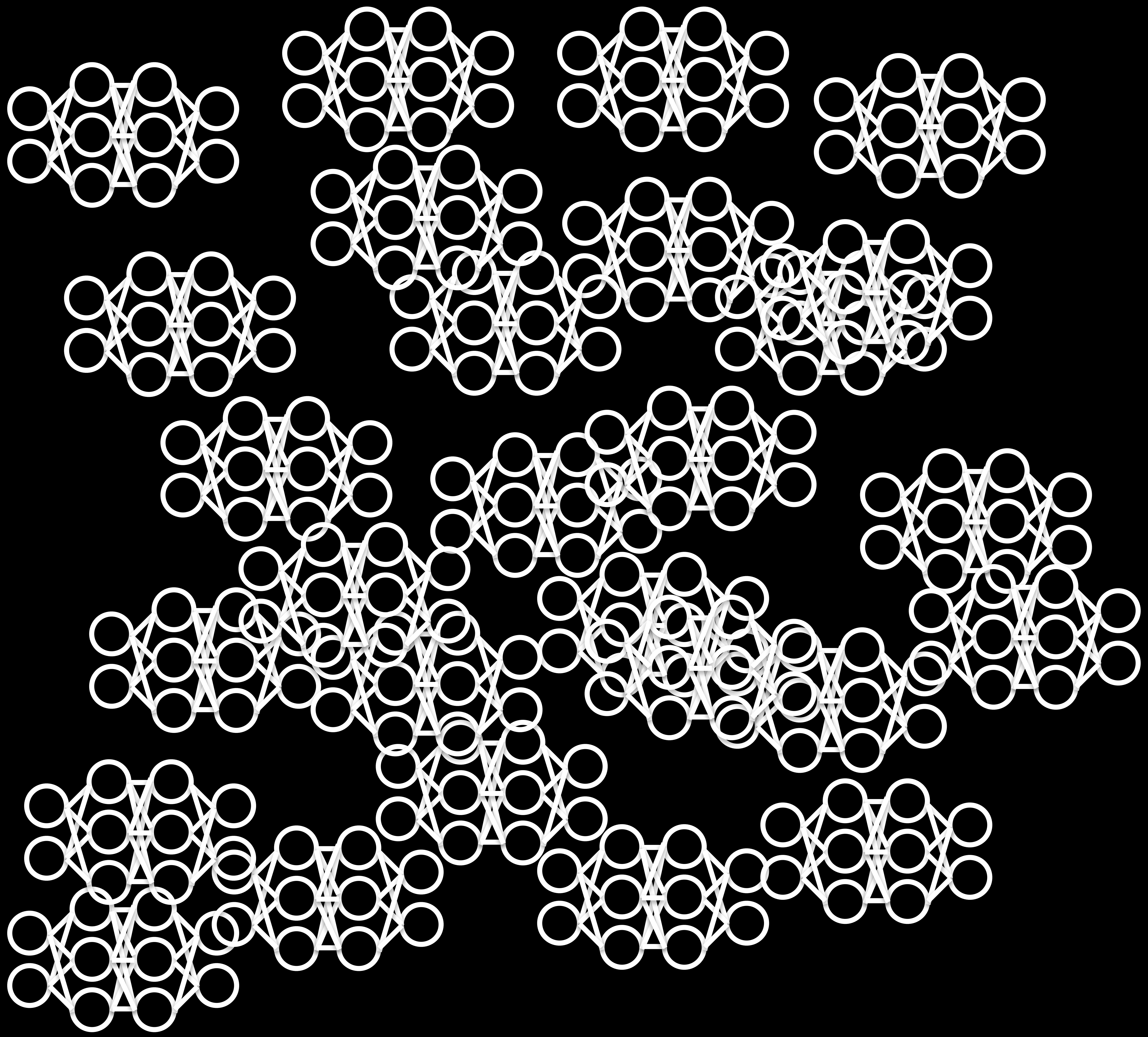



airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

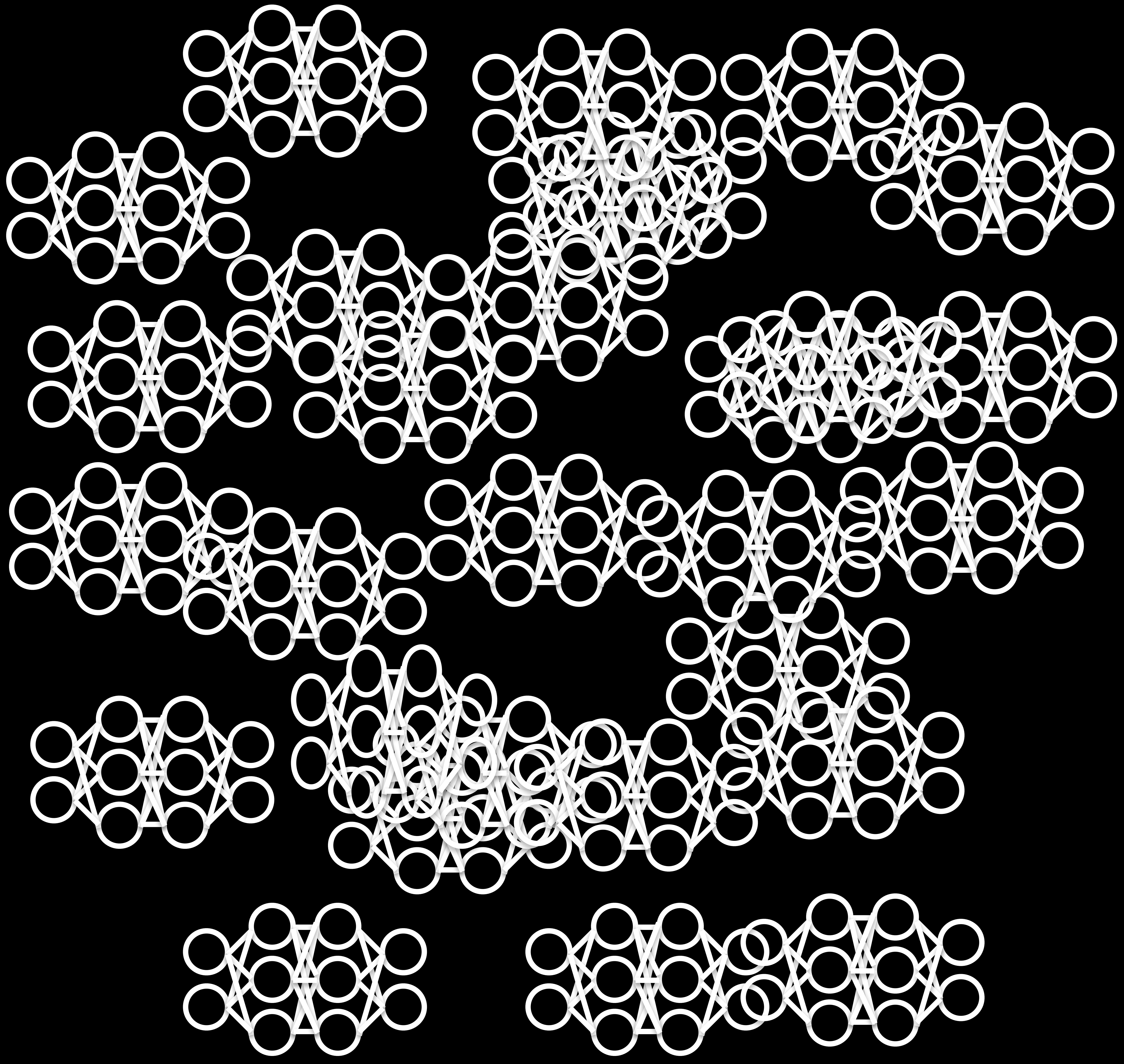

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck



Trained on

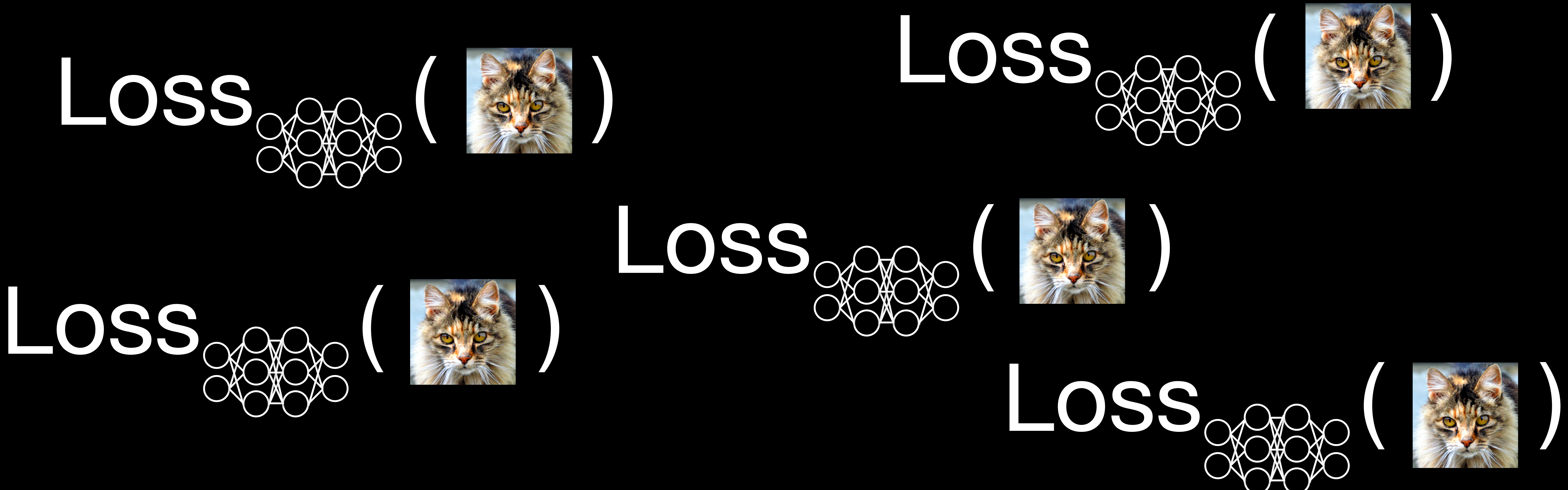


Not trained on



Trained on

Trained on



Loss( )

Loss( )

Loss( )

Loss( )

Loss( )

Trained on

2.3

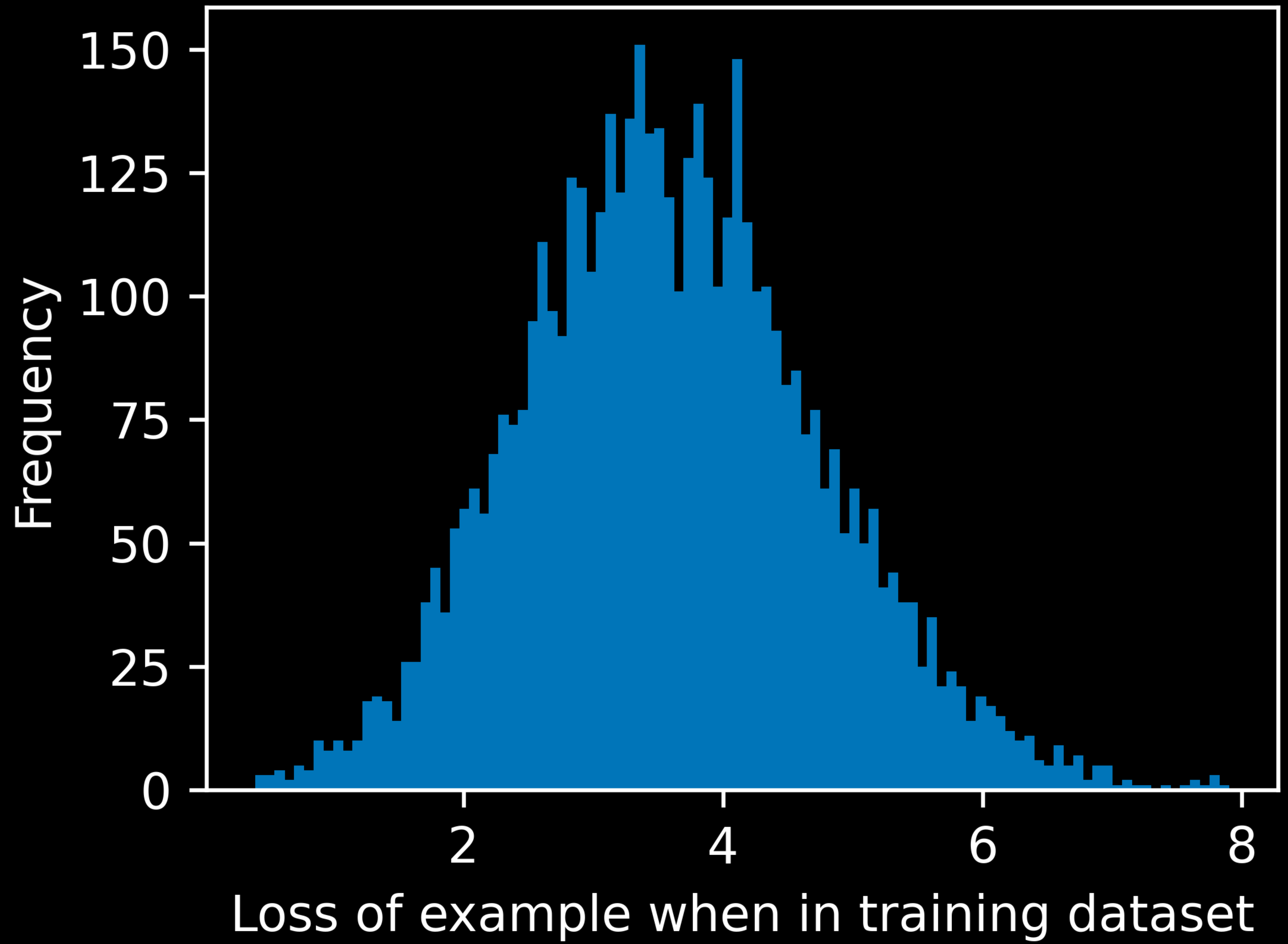
2.4

1.1

4.3

1.2

Trained on



2.4

1.2

Not trained on

Not trained on

8.4

7.5

10.2

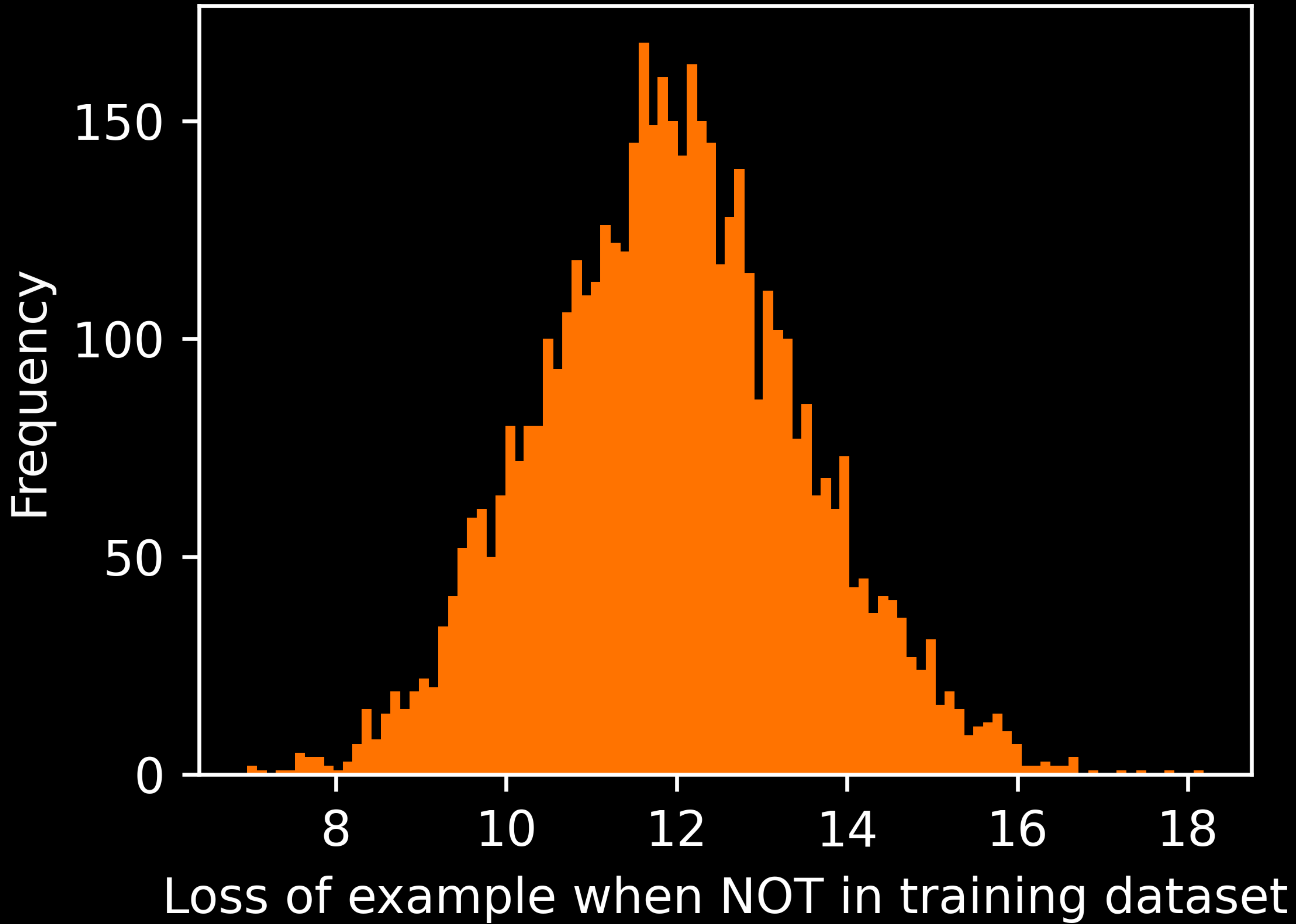
9.2

10.3

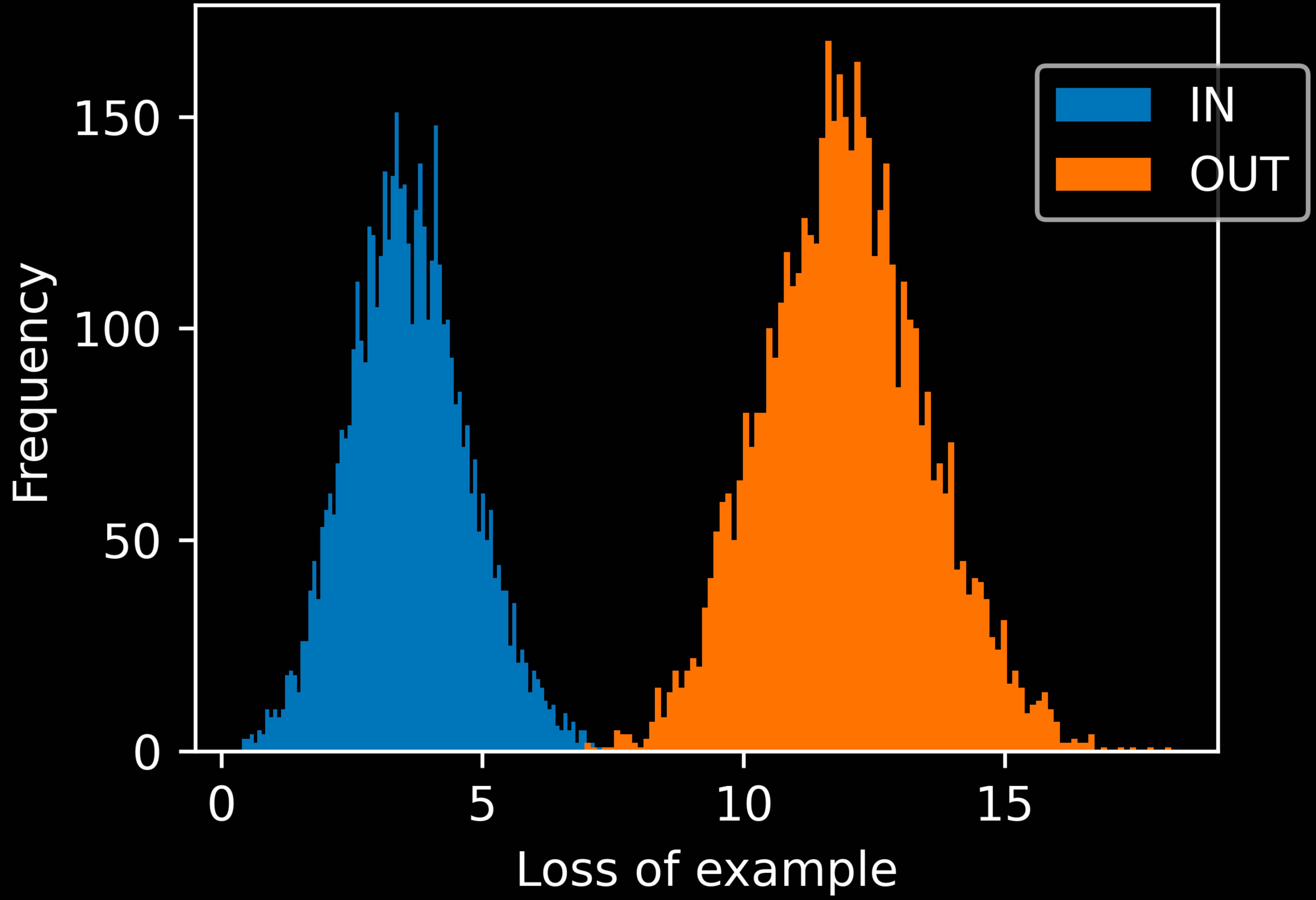
12.1

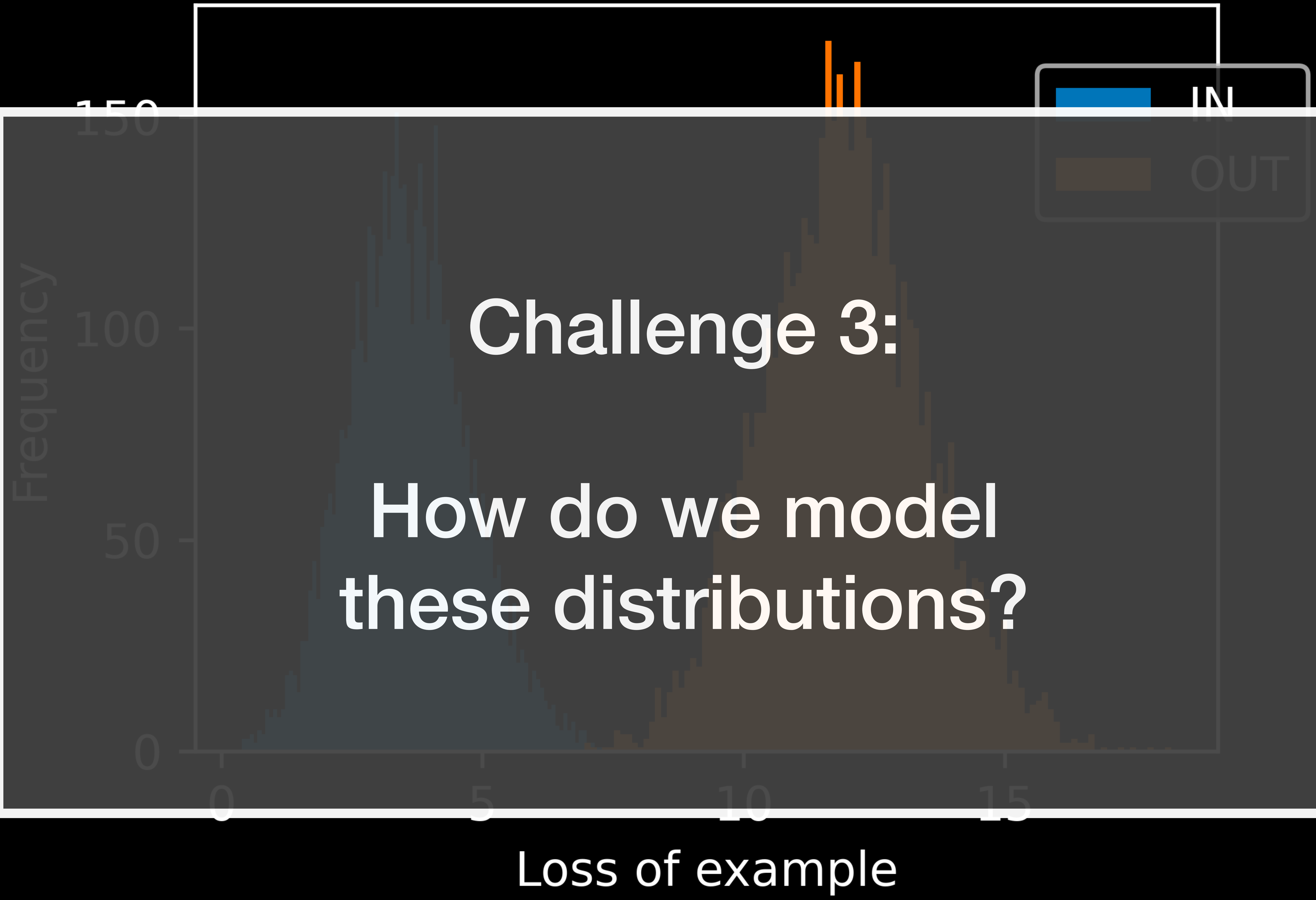
Not trained on

10.3



10.2


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



Challenge 3:

How do we model these distributions?

Two choices:

1. Nonparametric

2. Parametric

Two choices:

1. Nonparametric
2. Parametric

Challenge 4:

Nonparametric modeling would require training too many models.

Two choices:

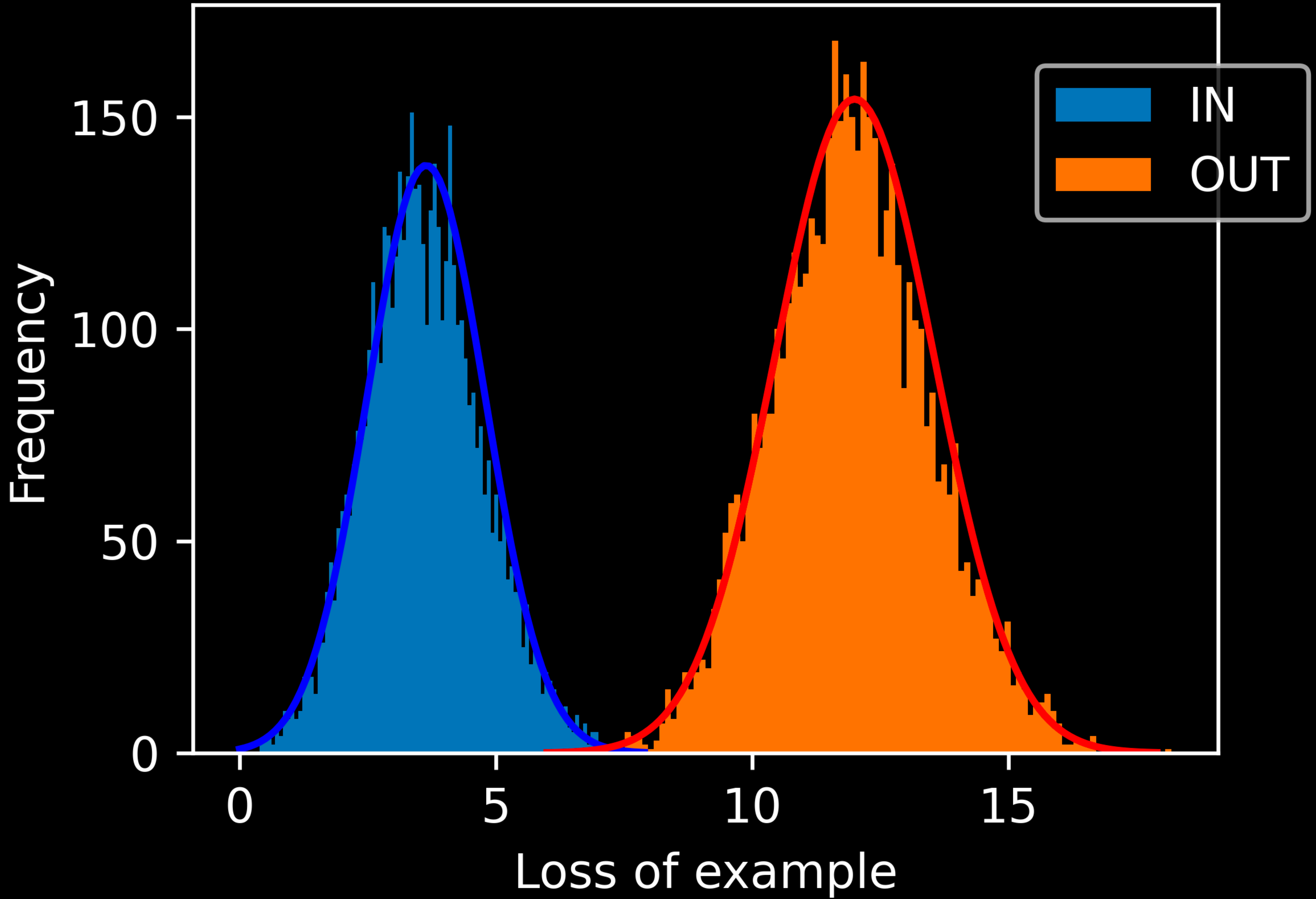
1. ~~Nonparametric~~
2. Parametric

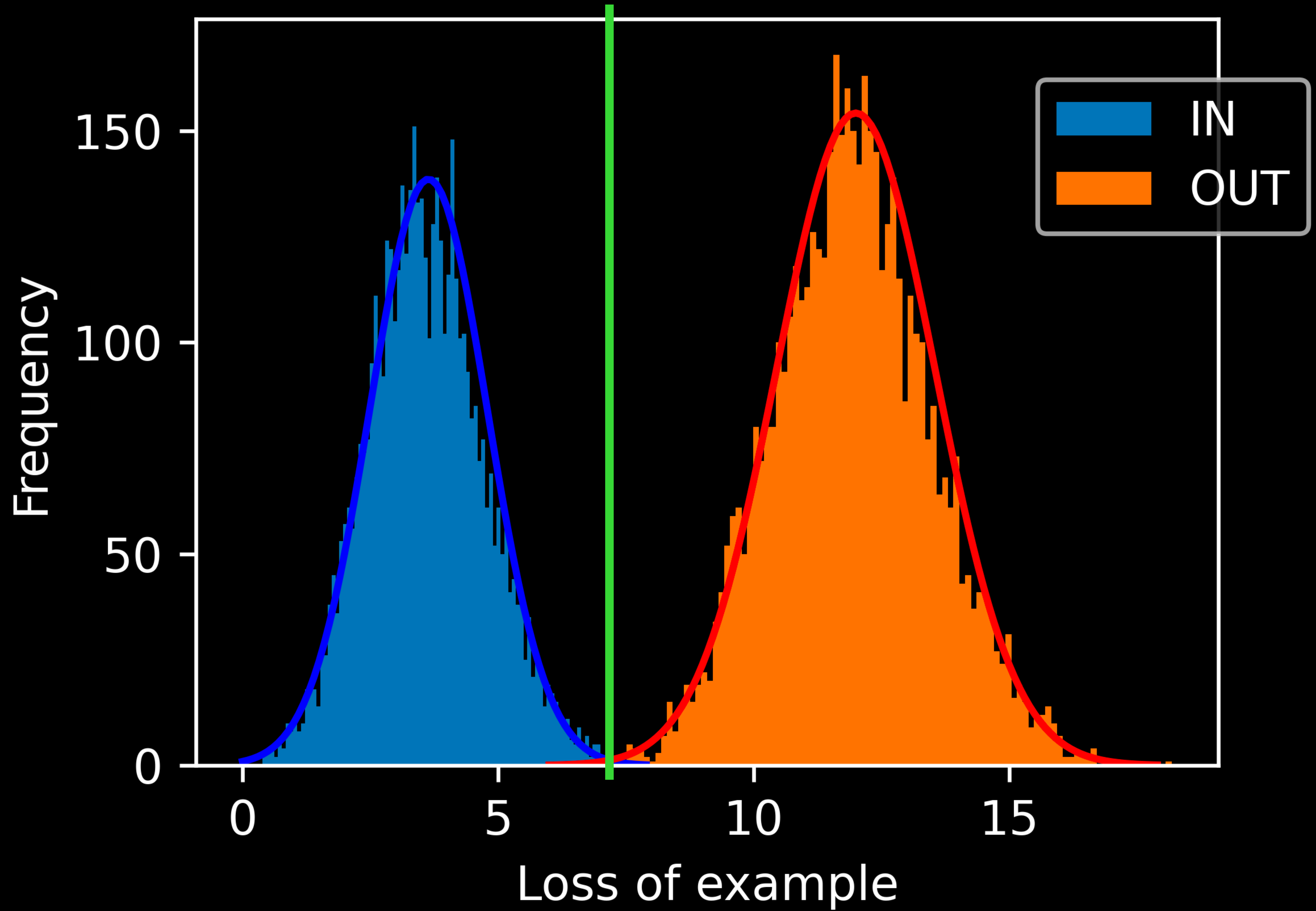
Two choices:

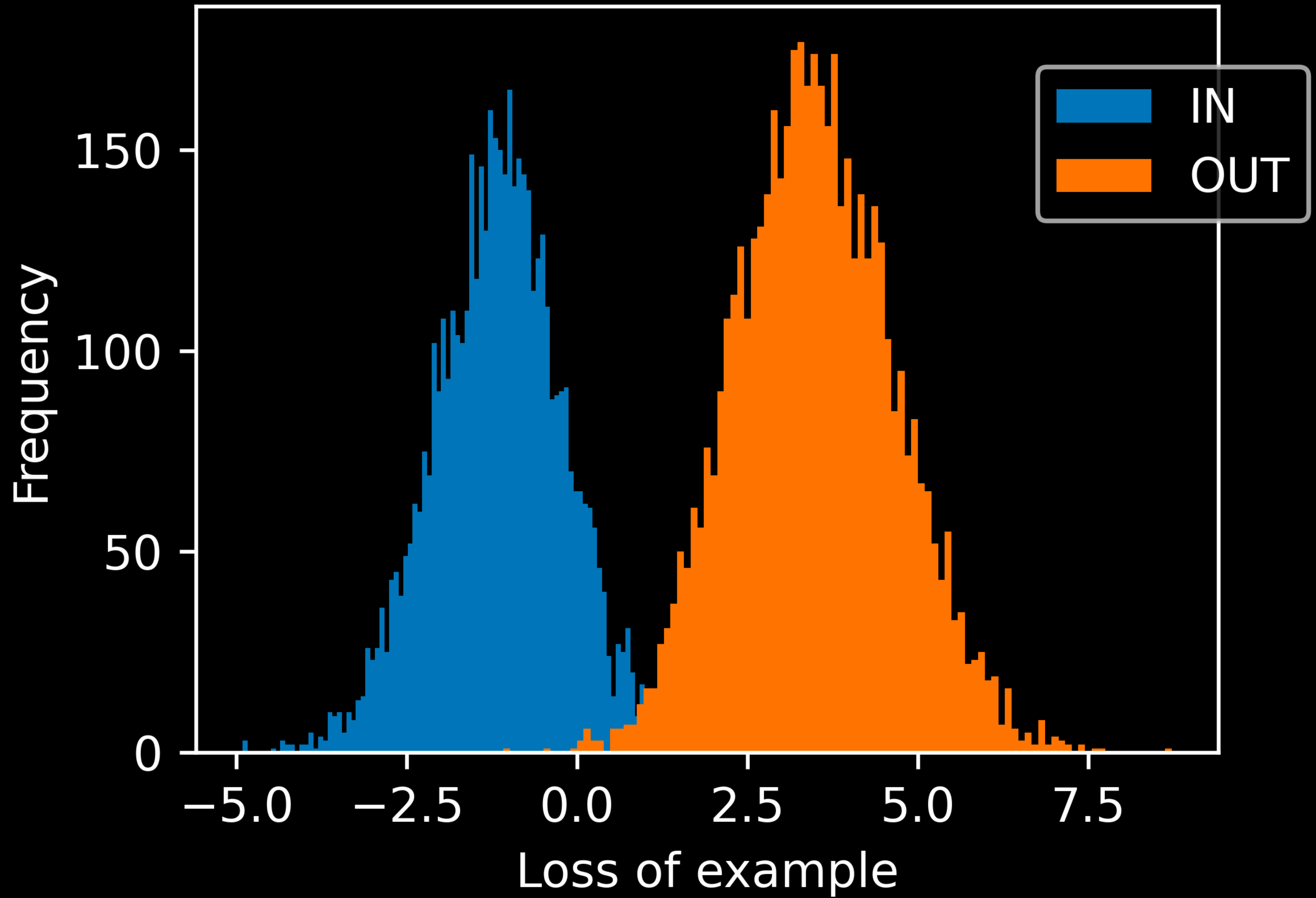
1. Nonparametric
2. Parametric

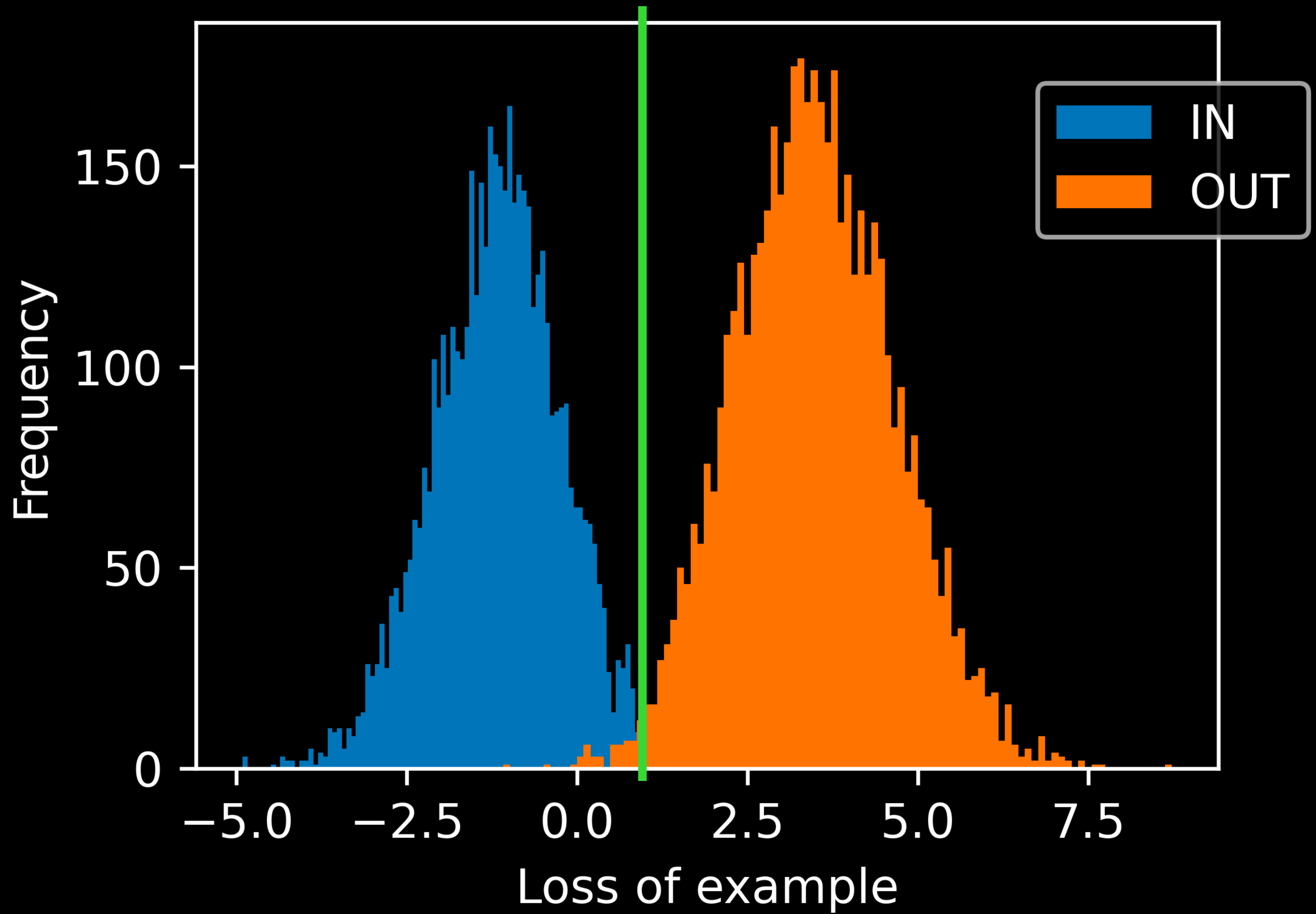
Challenge 5:

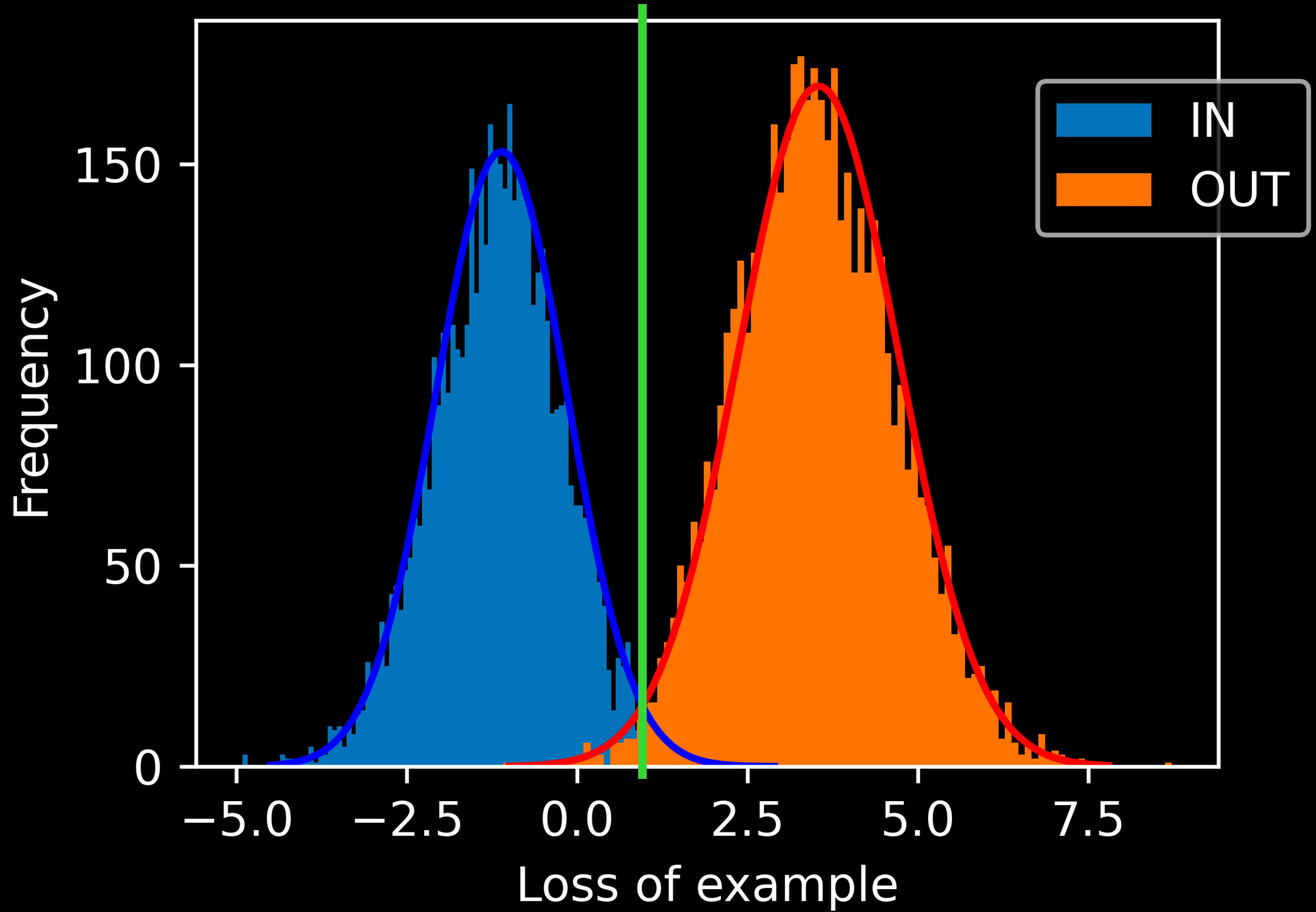
Fitting the data to a nice distribution requires care.

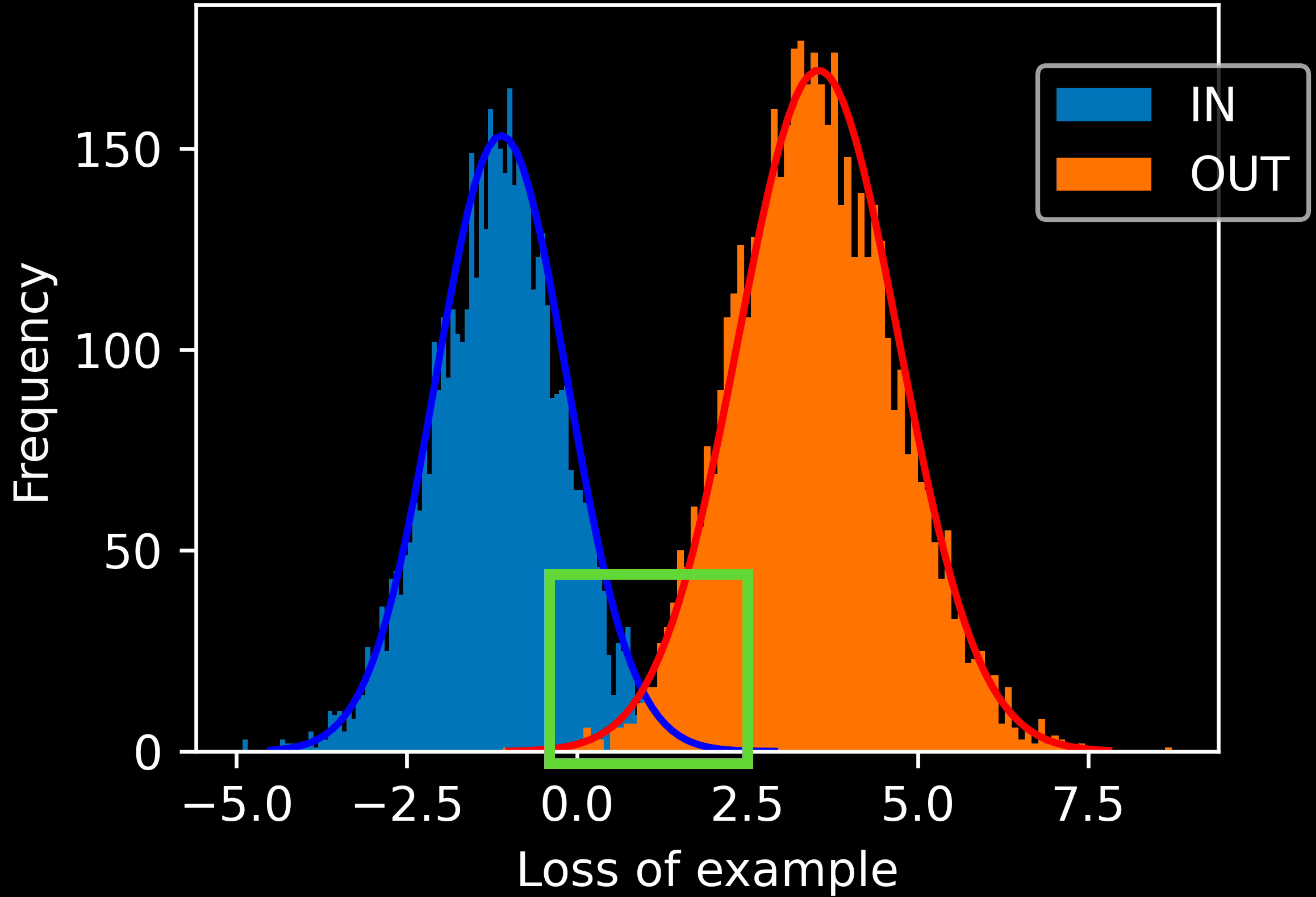

Frequency
150
100
50
0
0
5
10
15
Loss of example
IN
OUT

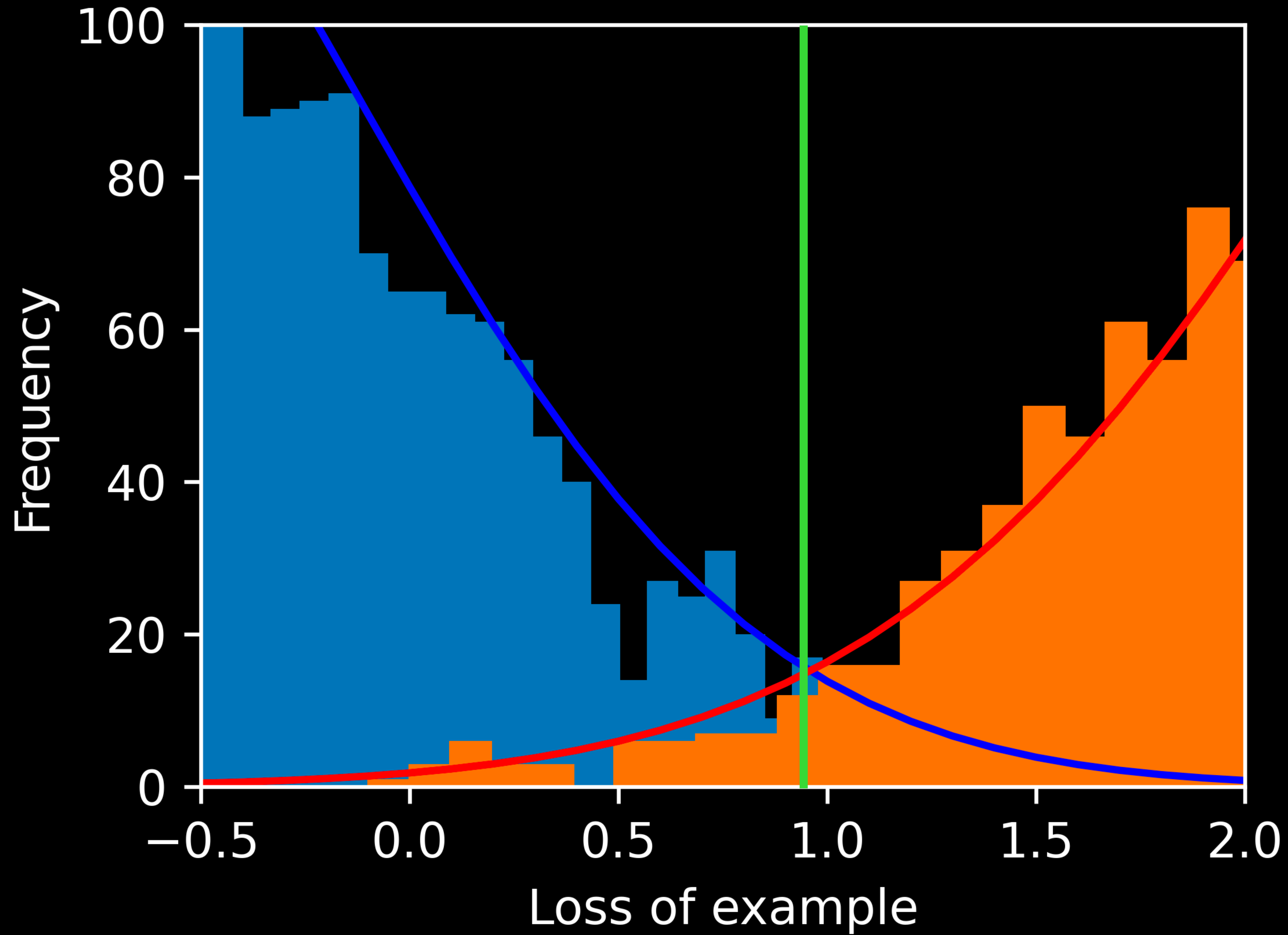

Frequency
150
100
50
0
0
5
10
15
Loss of example
IN
OUT

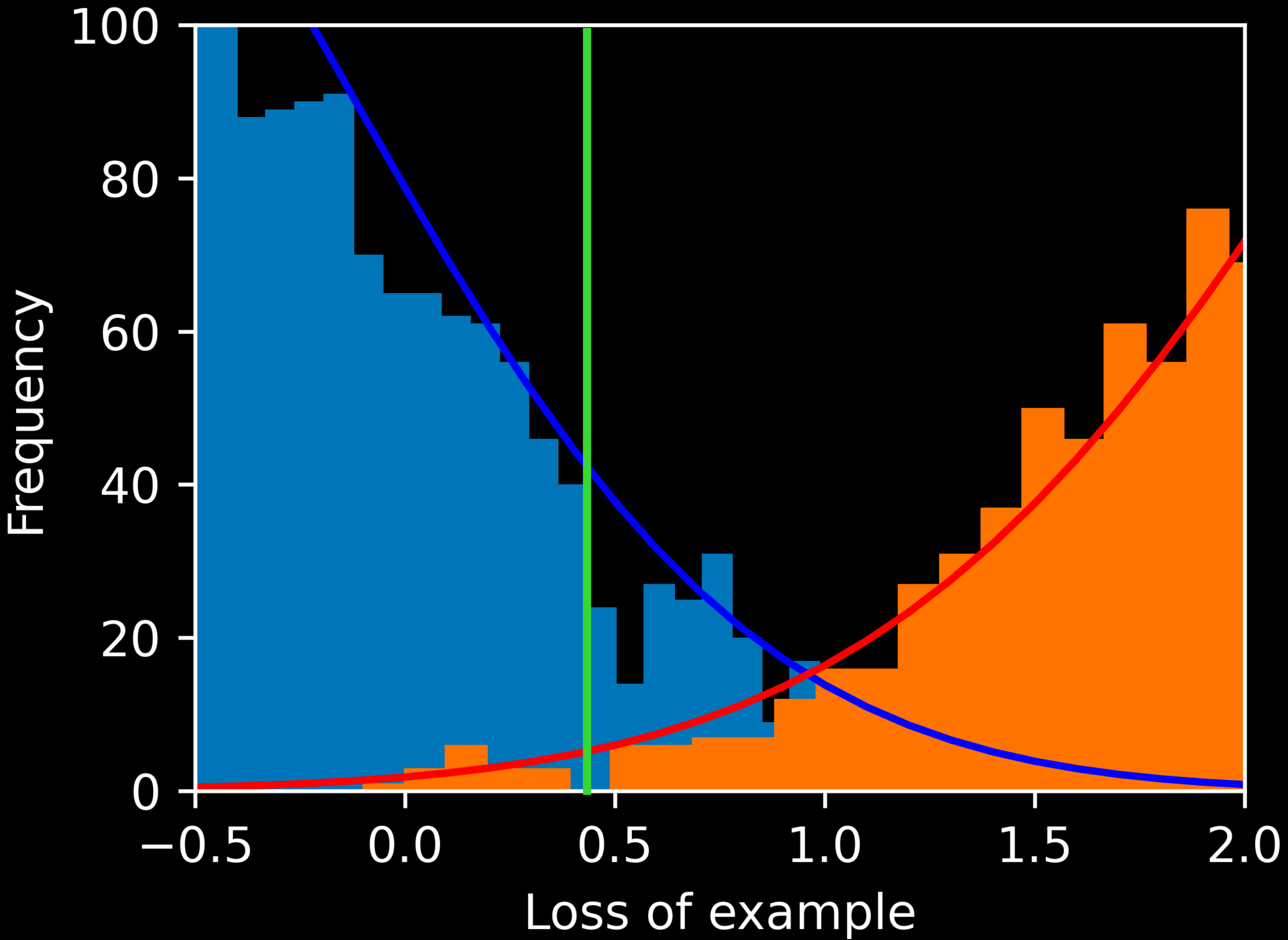

Frequency
150
100
50
0
−5.0
−2.5
0.0
2.5
5.0
7.5
Loss of example
IN
OUT


Frequency
150
100
50
0
−5.0
−2.5
0.0
2.5
5.0
7.5
Loss of example
🟦 IN
🟧 OUT


IN
OUT
Frequency
0
50
100
150
−5.0
−2.5
0.0
2.5
5.0
7.5
Loss of example


IN
OUT
Frequency
150
100
50
0
−5.0
−2.5
0.0
2.5
5.0
7.5
Loss of example


100
80
60
40
20
0
Frequency
−0.5
0.0
0.5
1.0
1.5
2.0
Loss of example


100
80
60
40
20
0
Frequency
−0.5
0.0
0.5
1.0
1.5
2.0
Loss of example

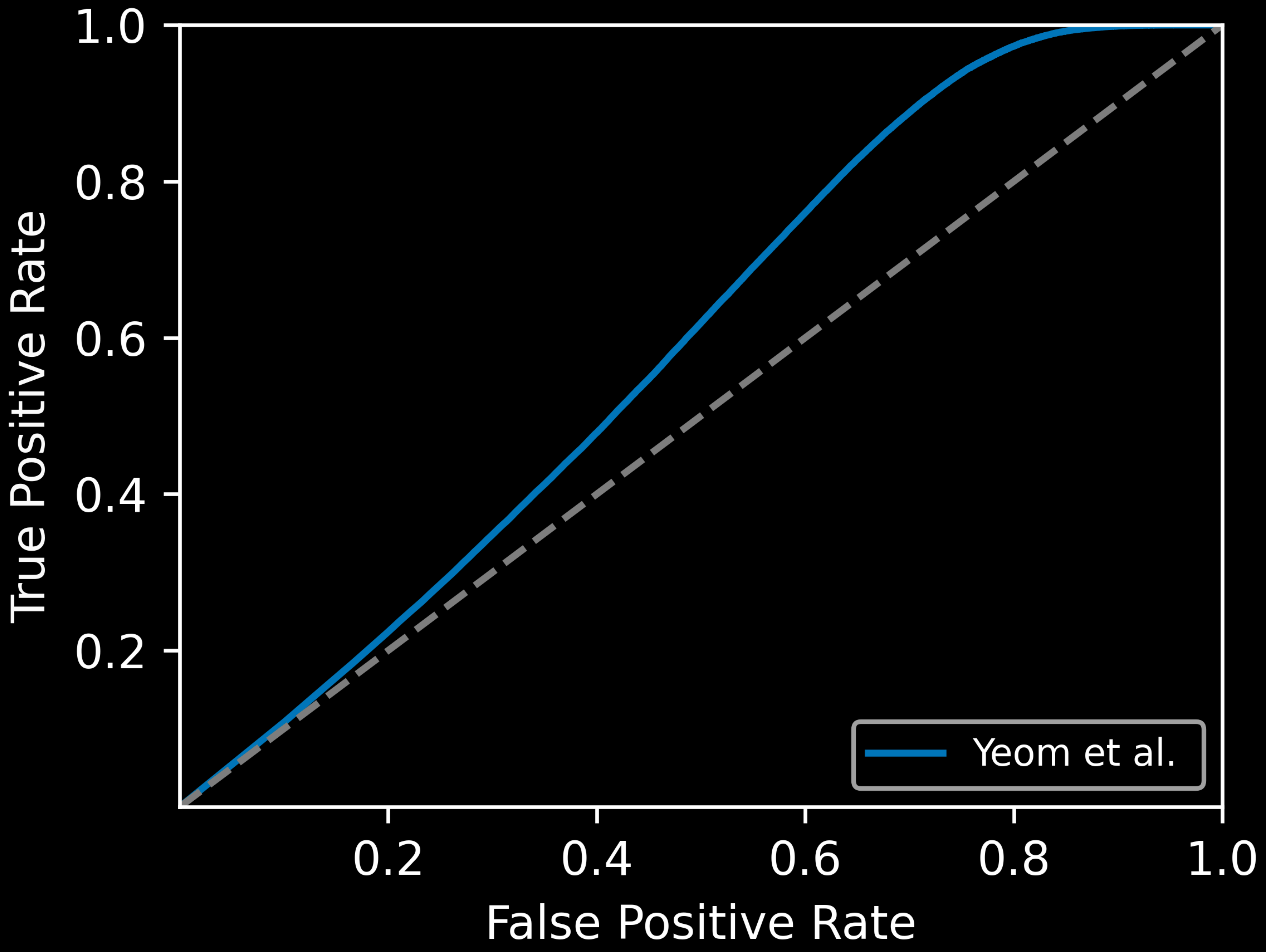
Results

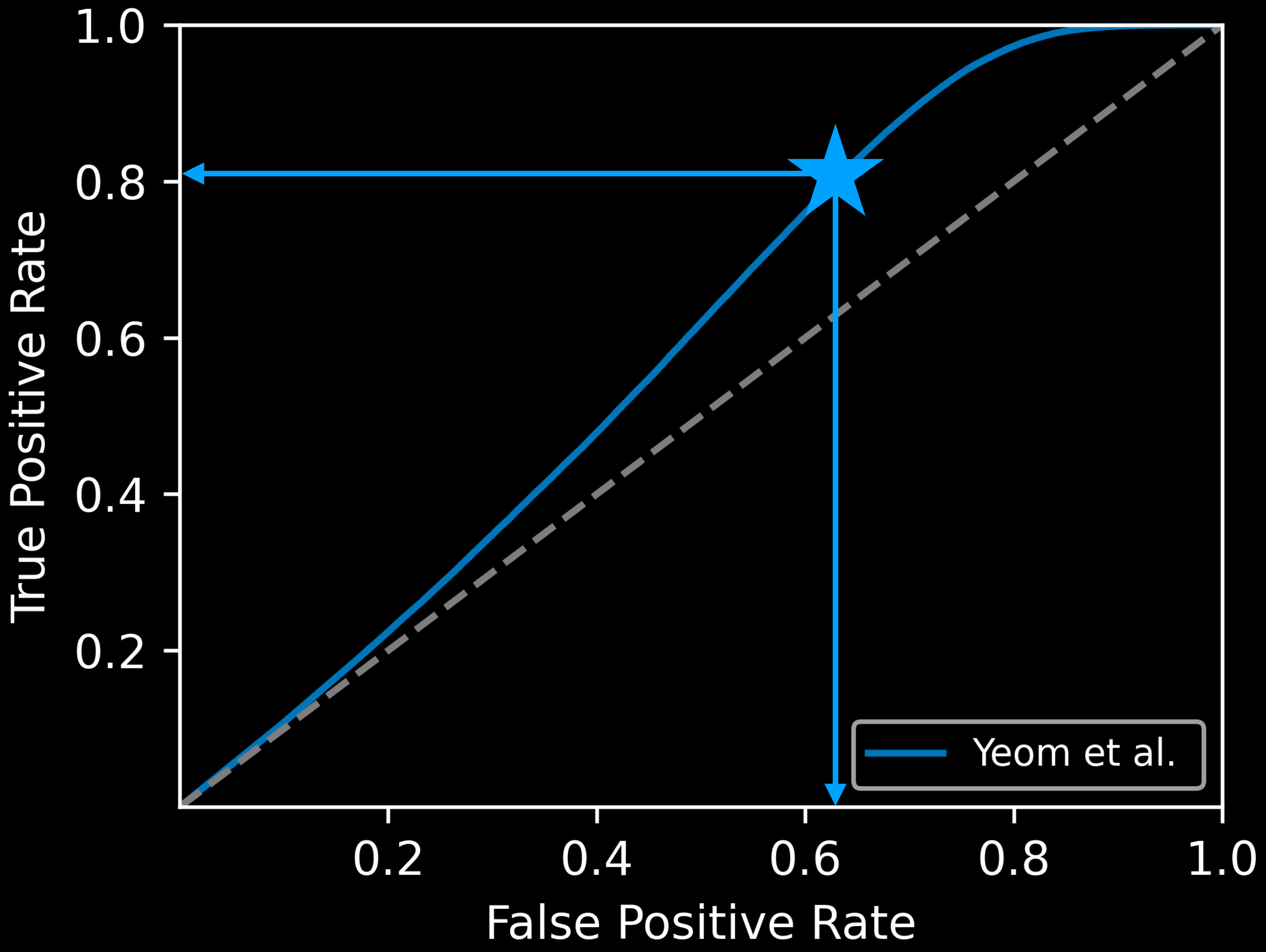
	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%

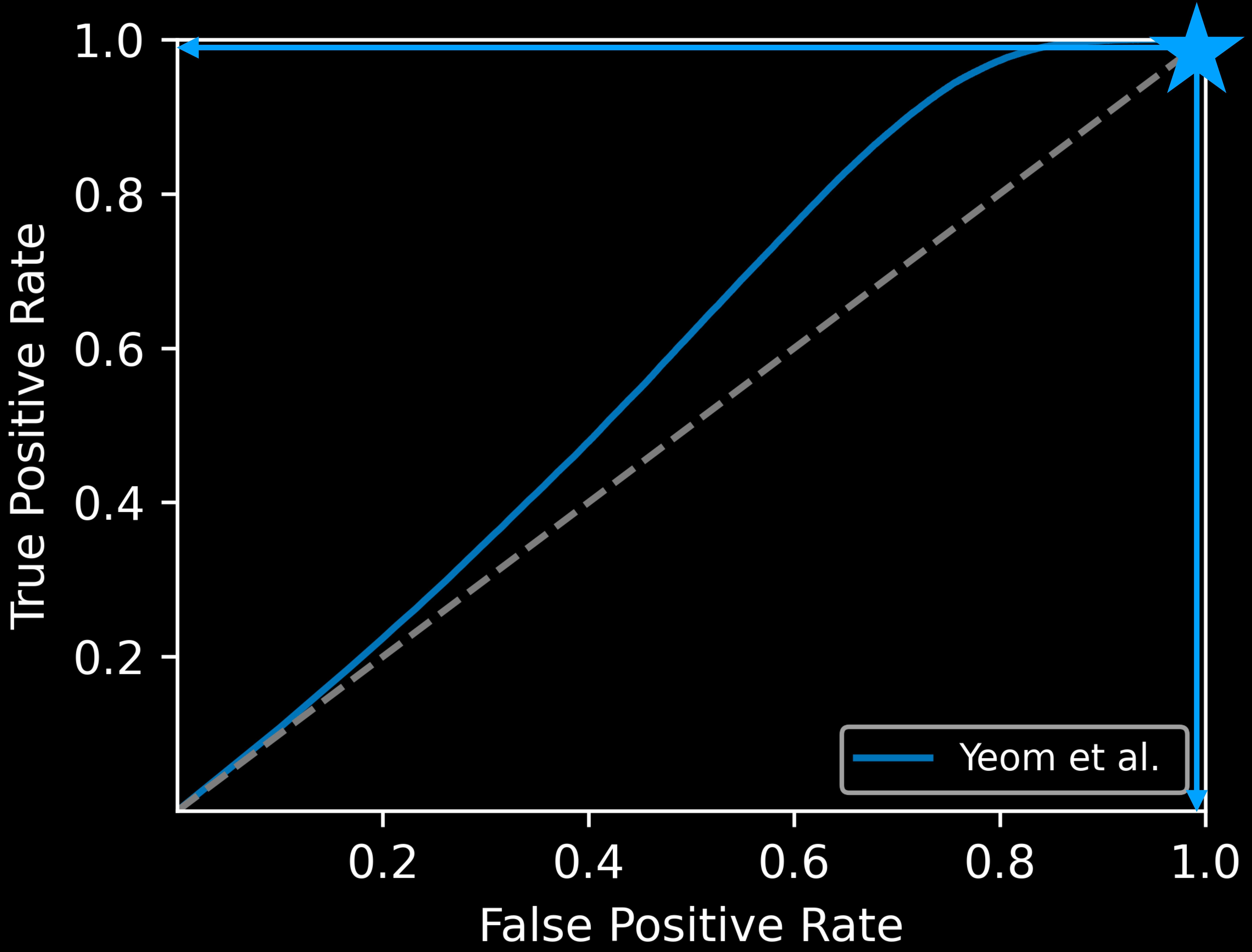
	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%

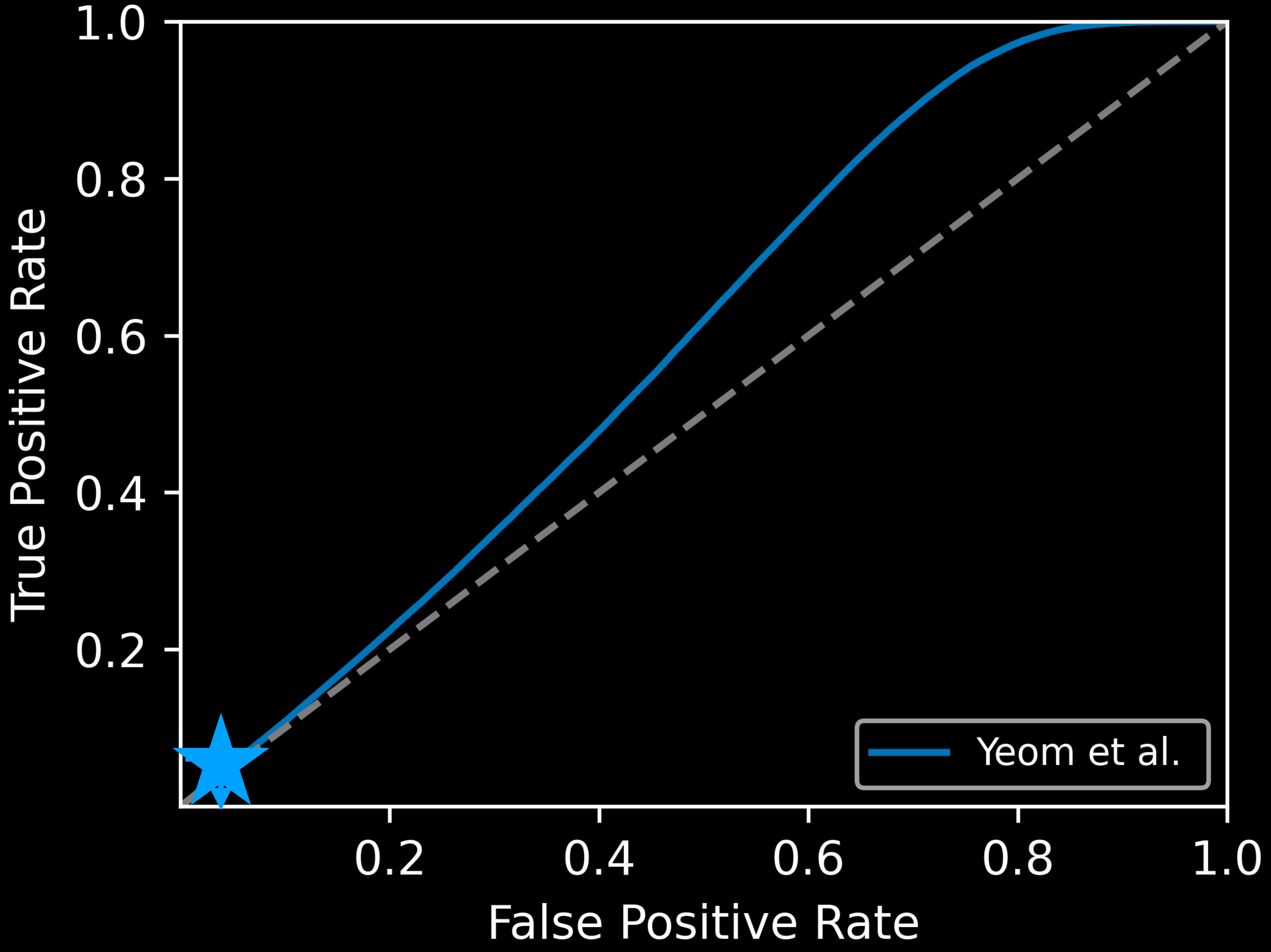
But: *average* accuracy is a bad metric for privacy

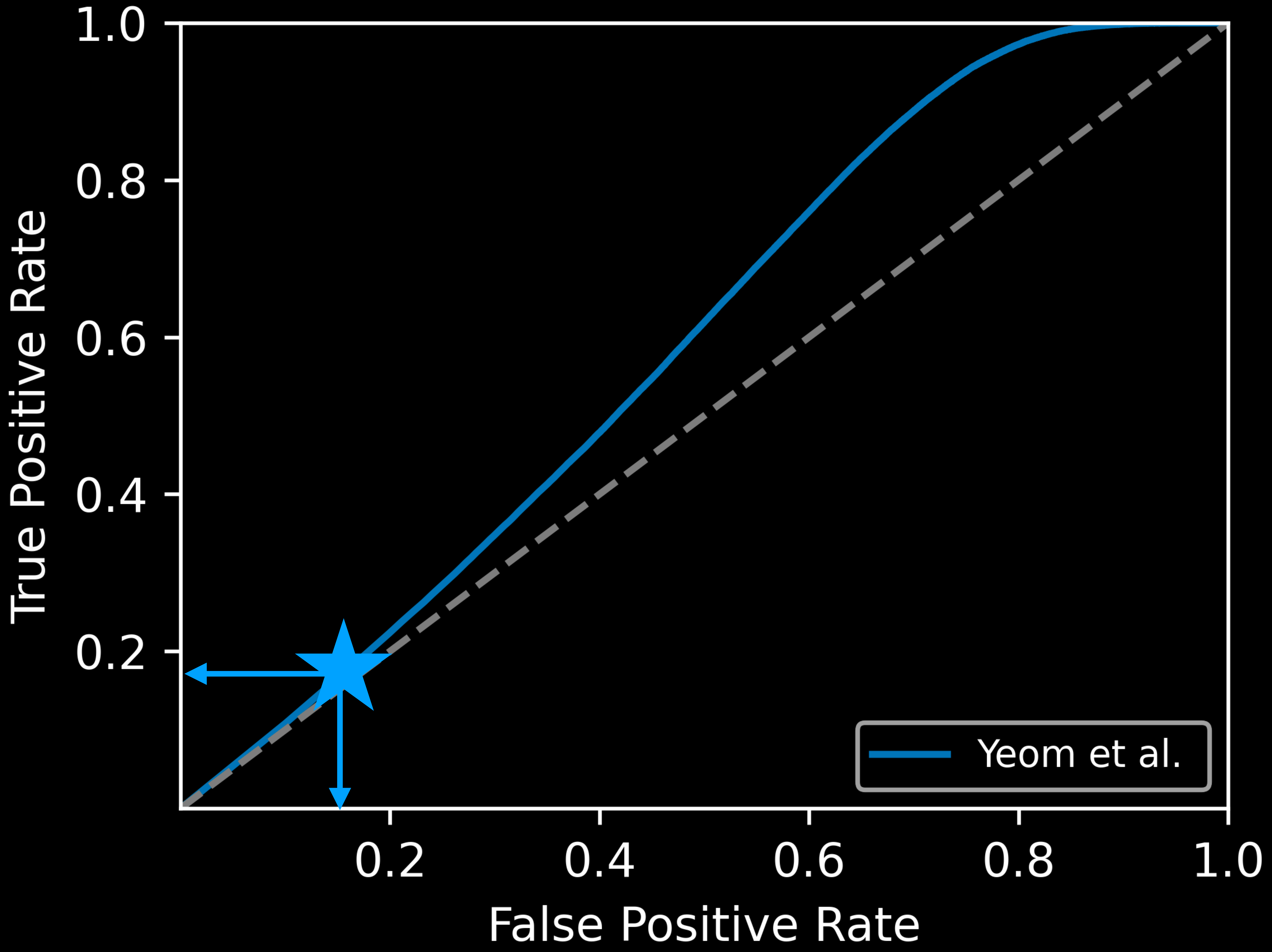
	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%

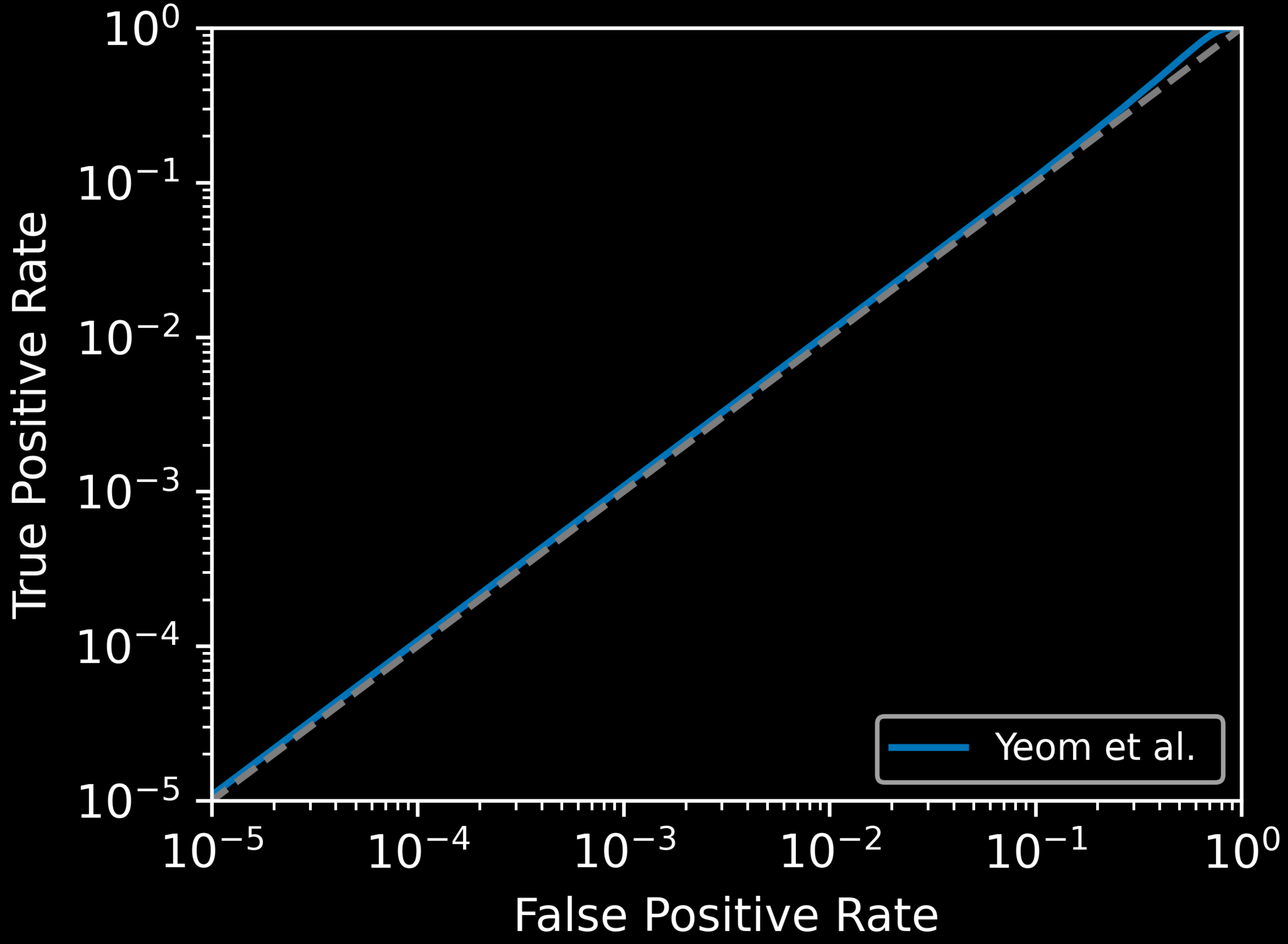

1.0
0.8
0.6
0.4
0.2
True Positive Rate
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Yeom et al.


1.0
0.8
0.6
0.4
0.2
True Positive Rate
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Yeom et al.

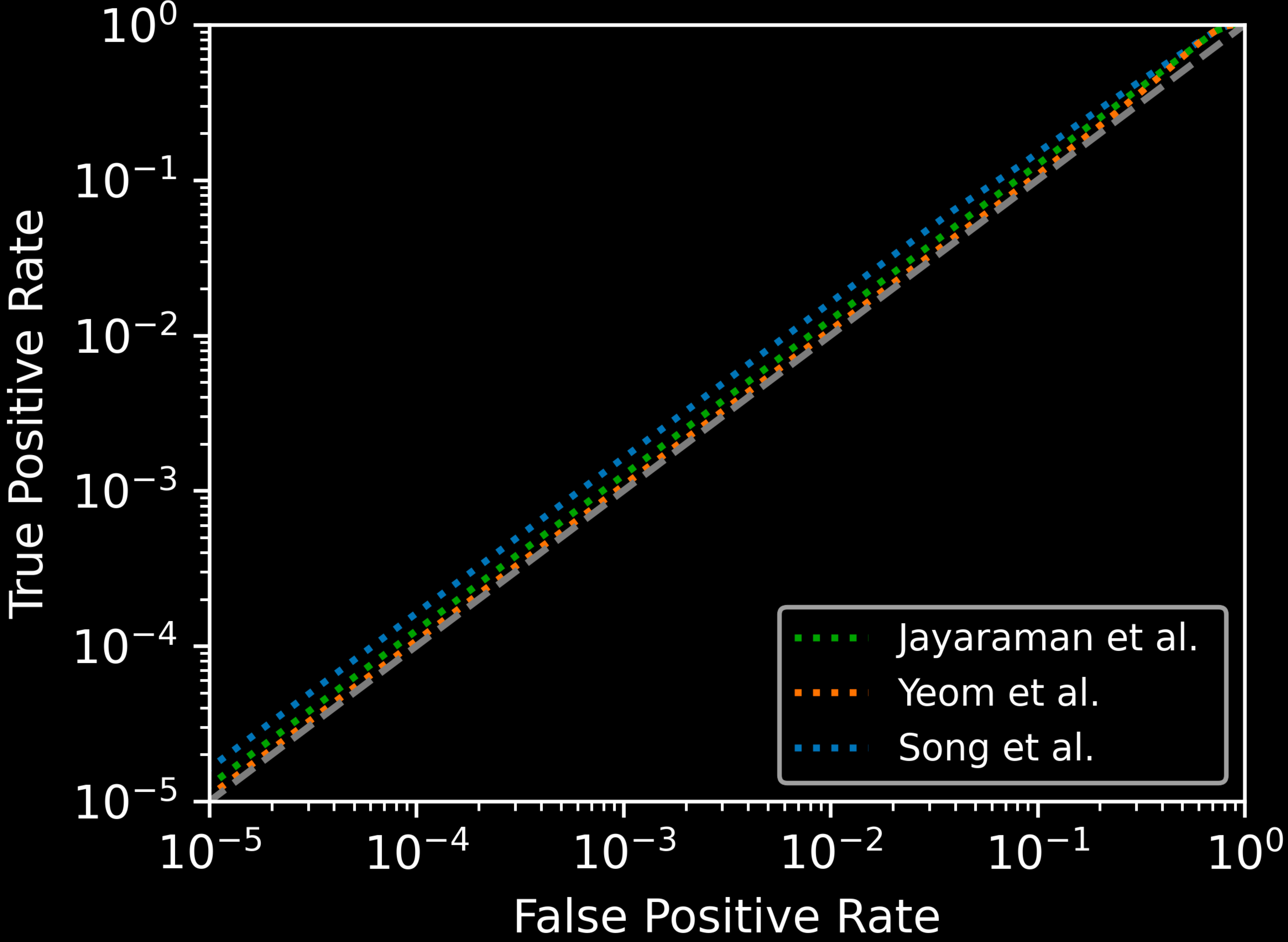

True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Yeom et al.


1.0
0.8
0.6
0.4
0.2
True Positive Rate
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Yeom et al.


1.0
0.8
0.6
0.4
0.2
True Positive Rate
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Yeom et al.

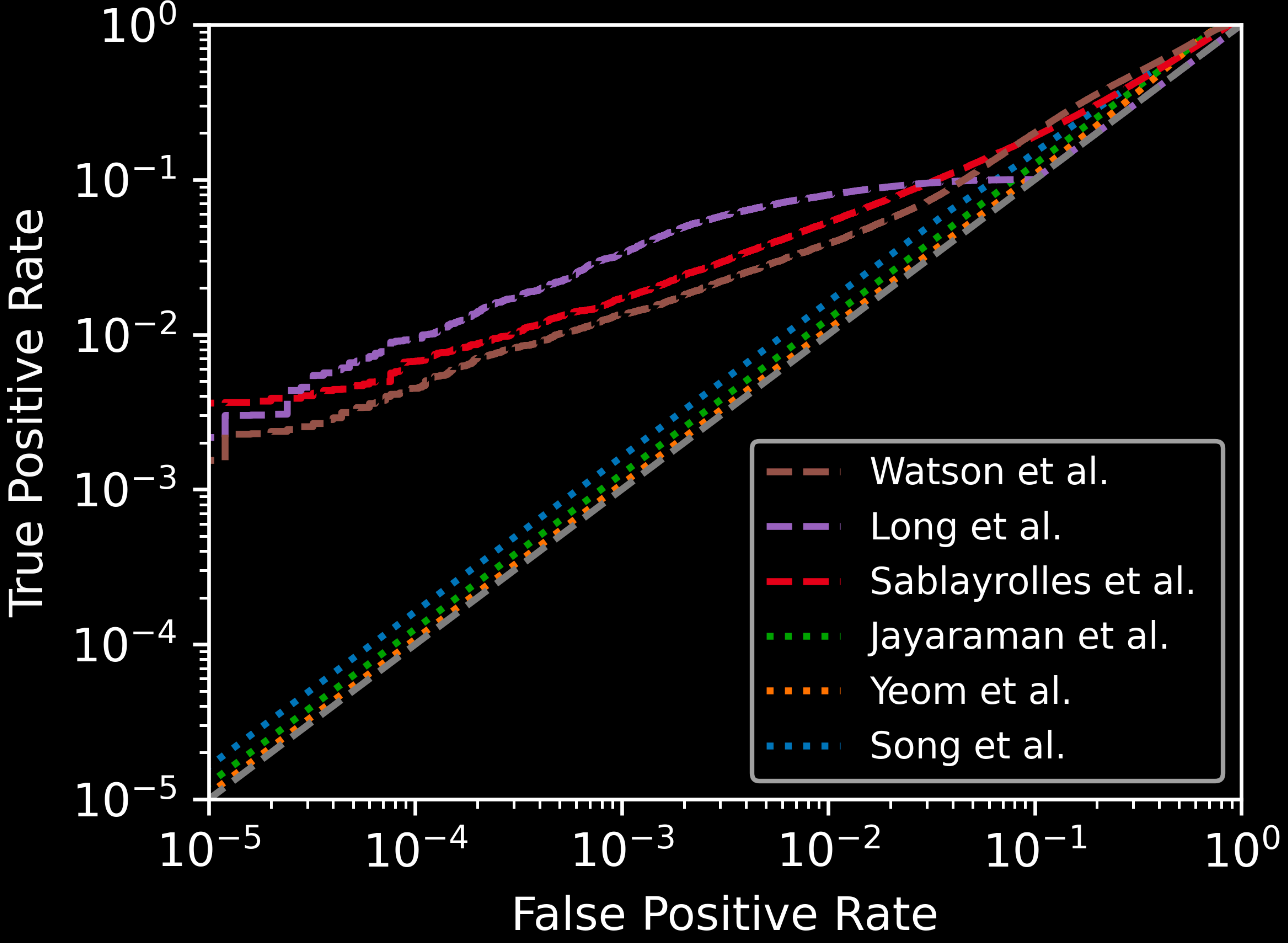

True Positive Rate
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
Yeom et al.
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
False Positive Rate

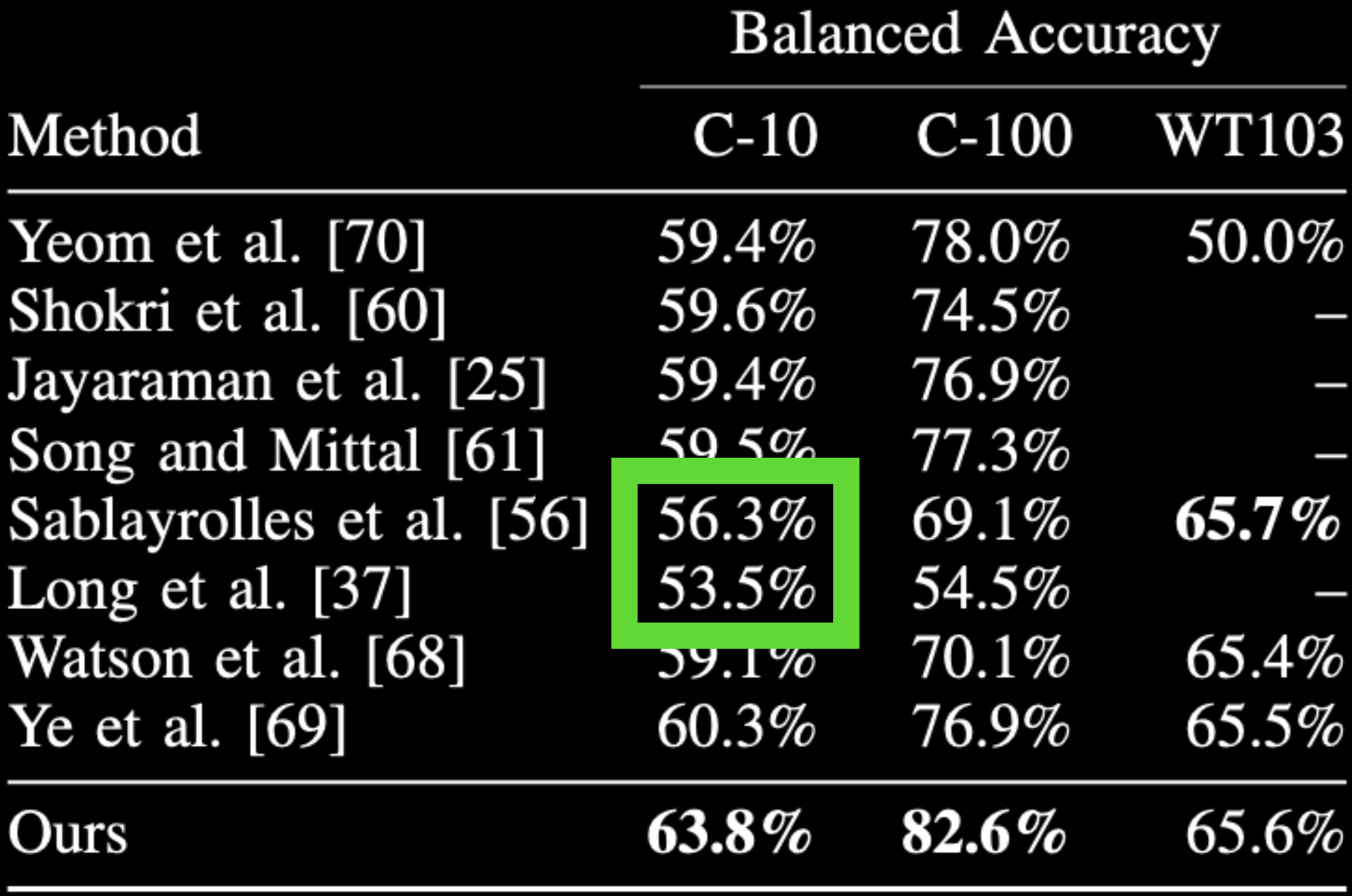
	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%


True Positive Rate
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
False Positive Rate
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
Jayaraman et al.
Yeom et al.
Song et al.

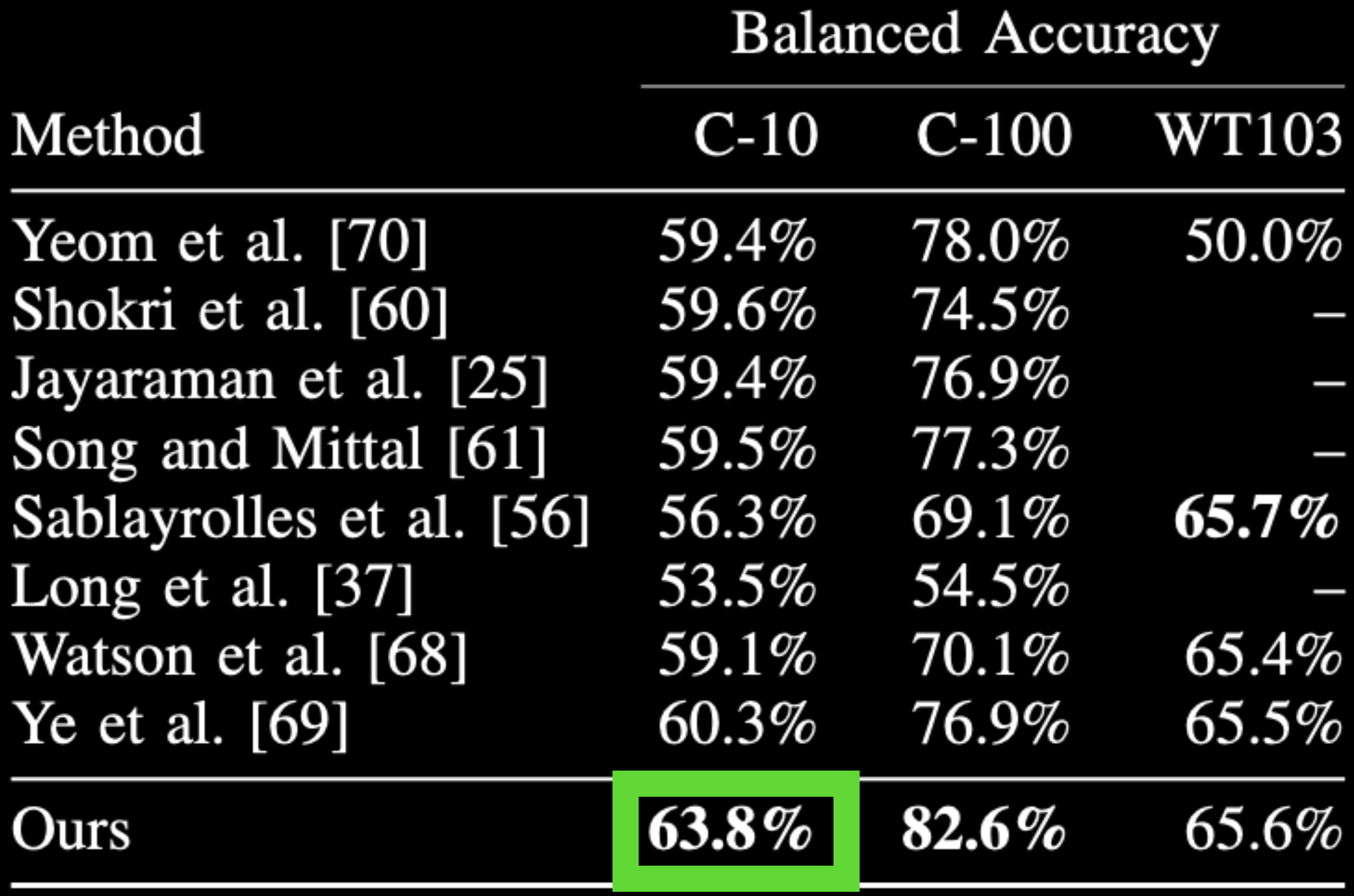
	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%

	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%

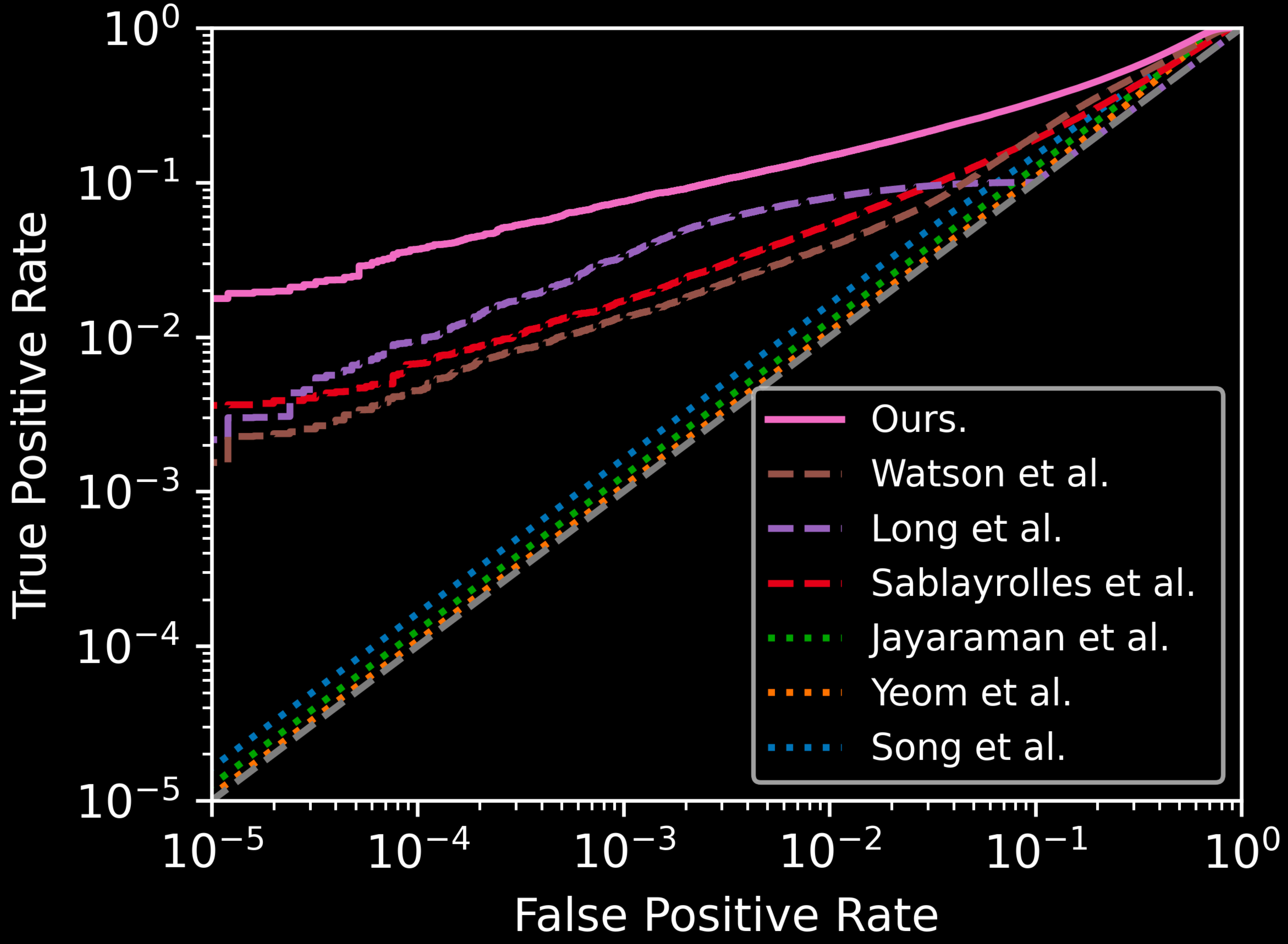

True Positive Rate
False Positive Rate
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
Watson et al.
Long et al.
Sablayrolles et al.
Jayaraman et al.
Yeom et al.
Song et al.

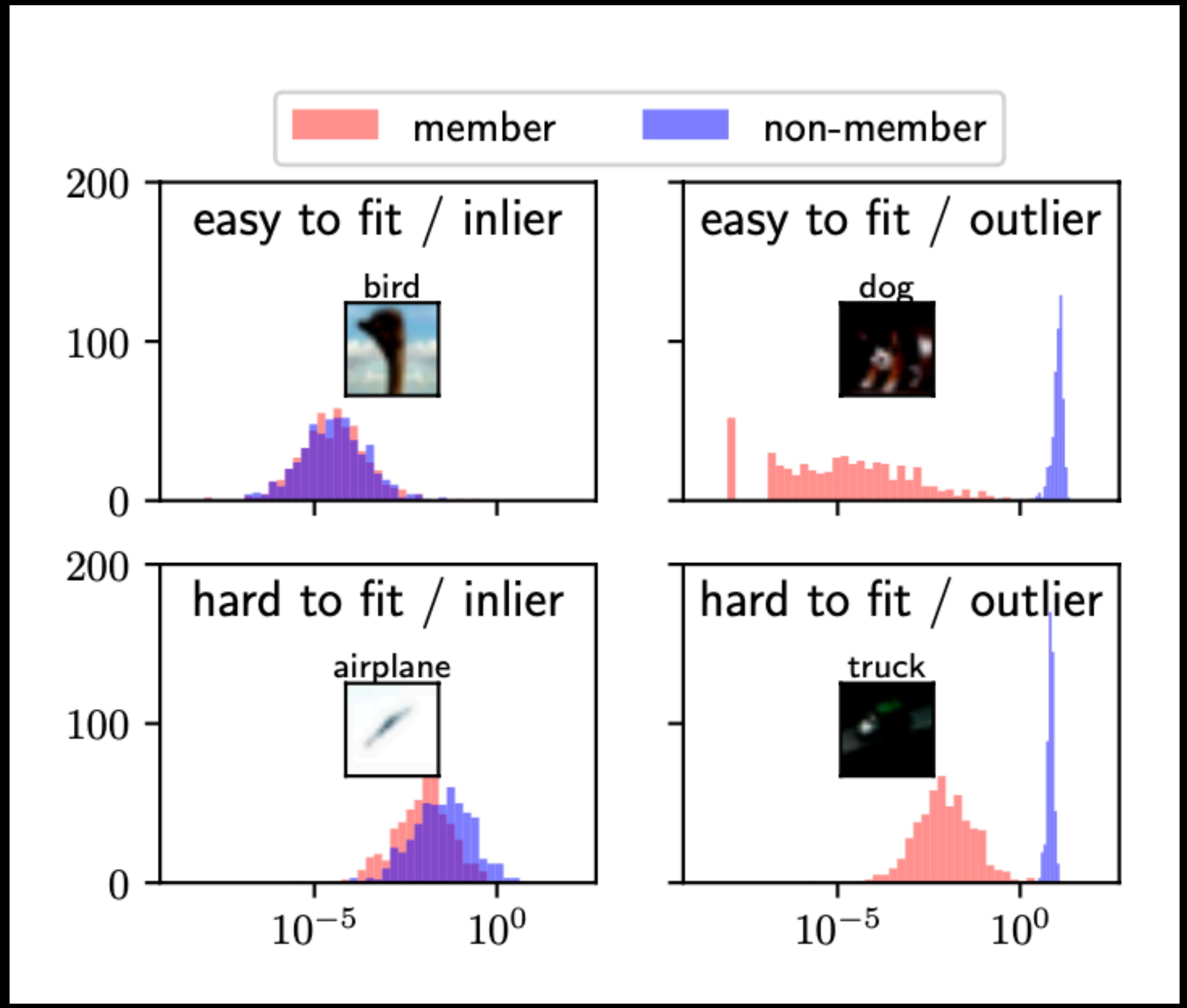


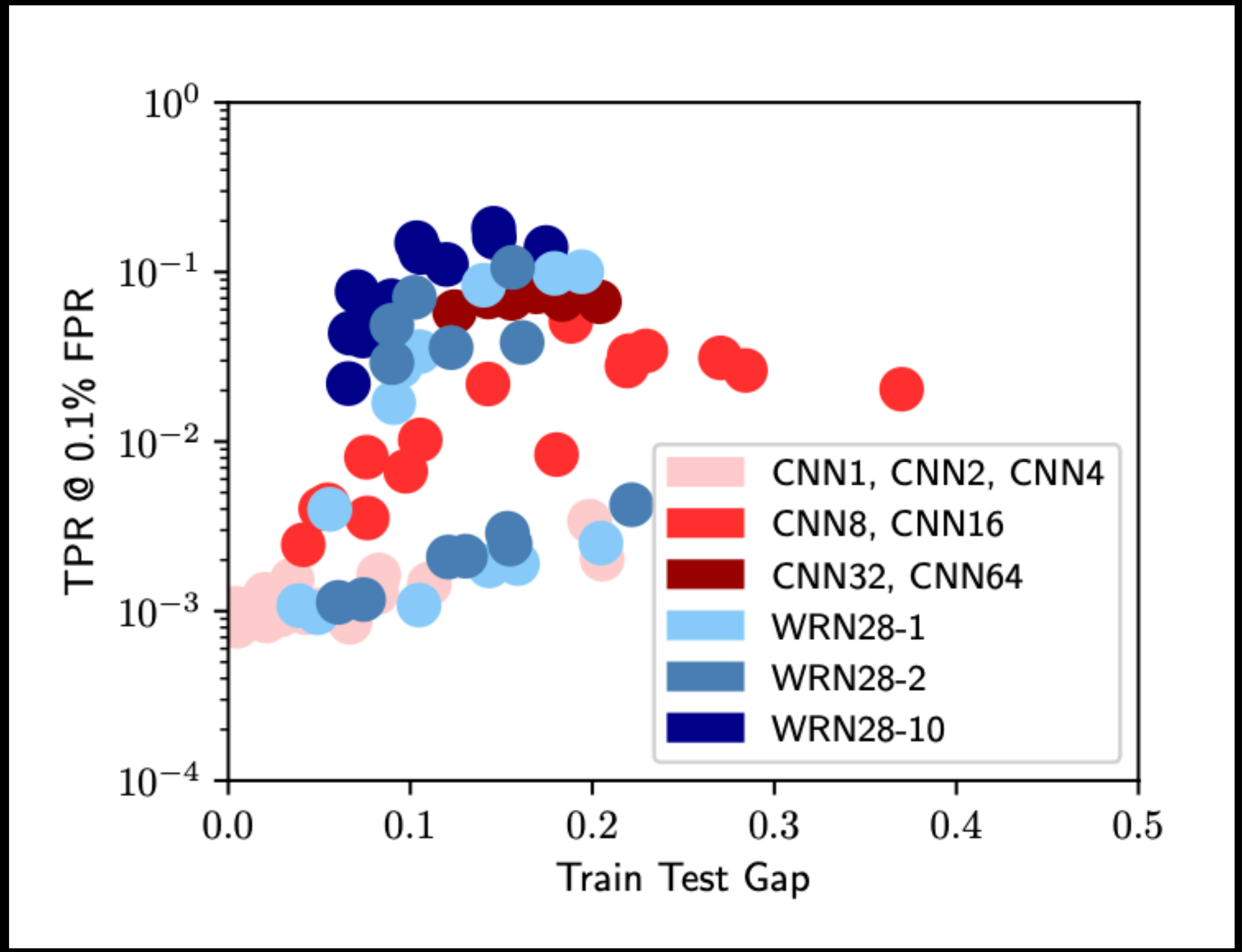
	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%

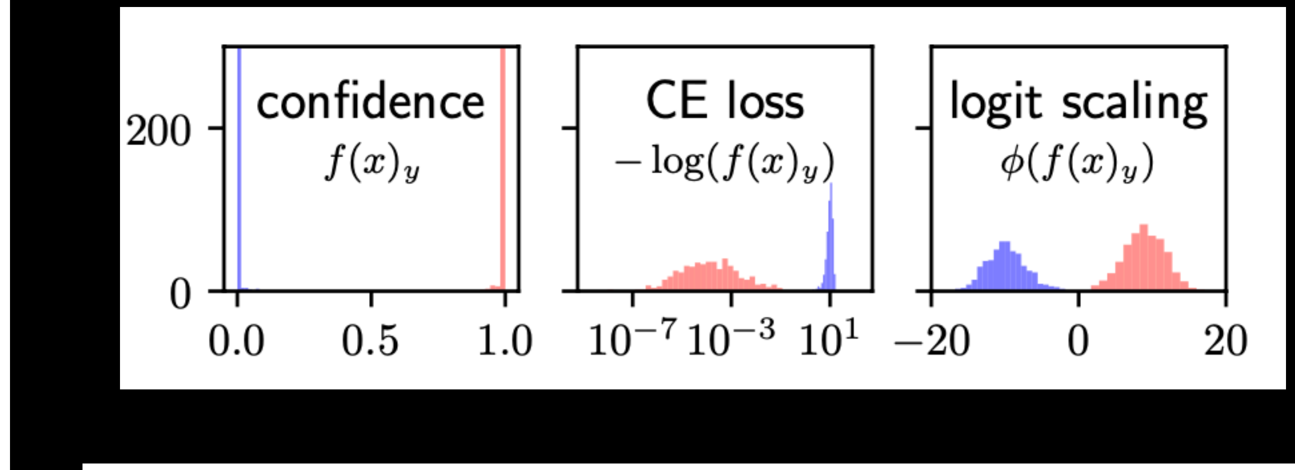


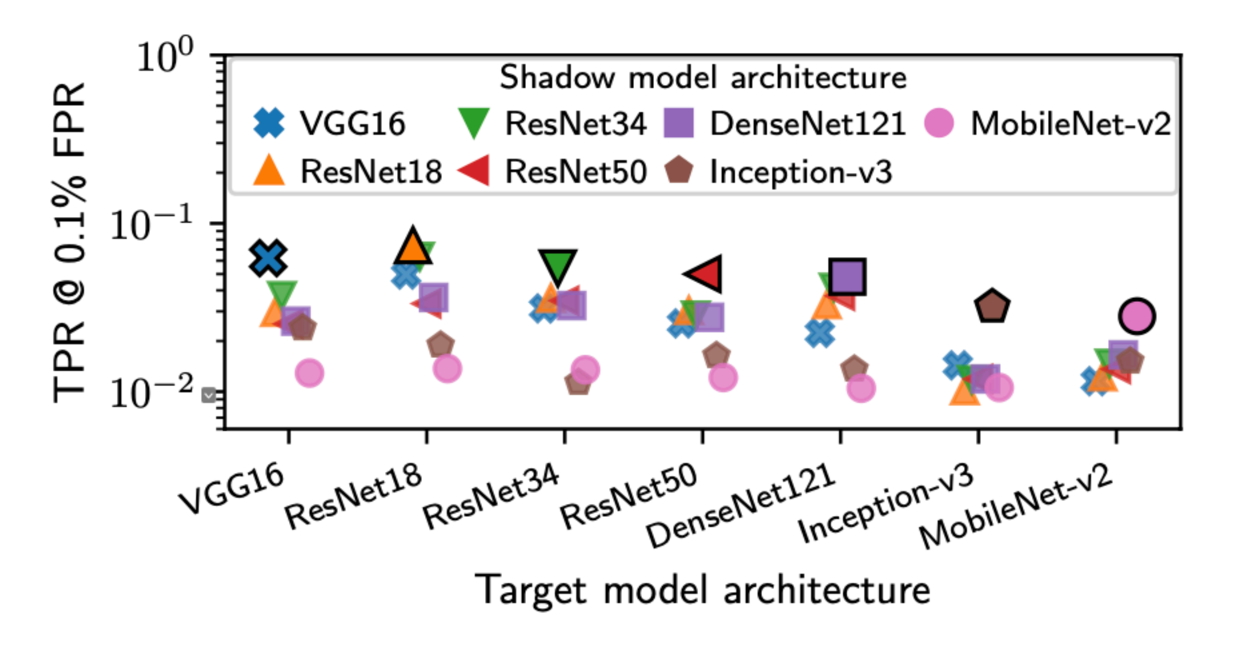
	Balanced Accuracy		
Method	C-10	C-100	WT103
Yeom et al. [70]	59.4%	78.0%	50.0%
Shokri et al. [60]	59.6%	74.5%	–
Jayaraman et al. [25]	59.4%	76.9%	–
Song and Mittal [61]	59.5%	77.3%	–
Sablayrolles et al. [56]	56.3%	69.1%	**65.7%**
Long et al. [37]	53.5%	54.5%	–
Watson et al. [68]	59.1%	70.1%	65.4%
Ye et al. [69]	60.3%	76.9%	65.5%
Ours	**63.8%**	**82.6%**	65.6%

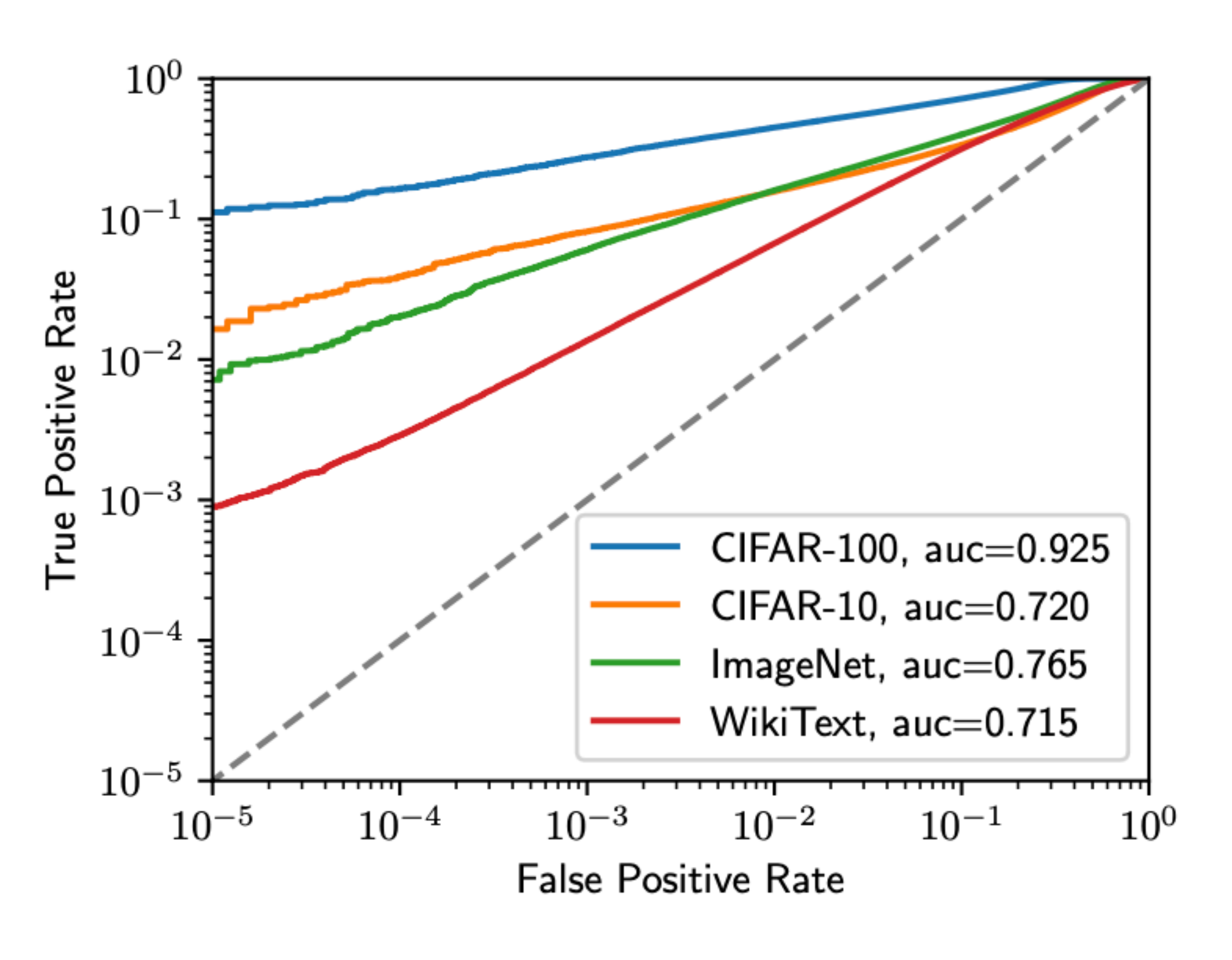

True Positive Rate
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
False Positive Rate
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
Ours.
Watson et al.
Long et al.
Sablayrolles et al.
Jayaraman et al.
Yeom et al.
Song et al.

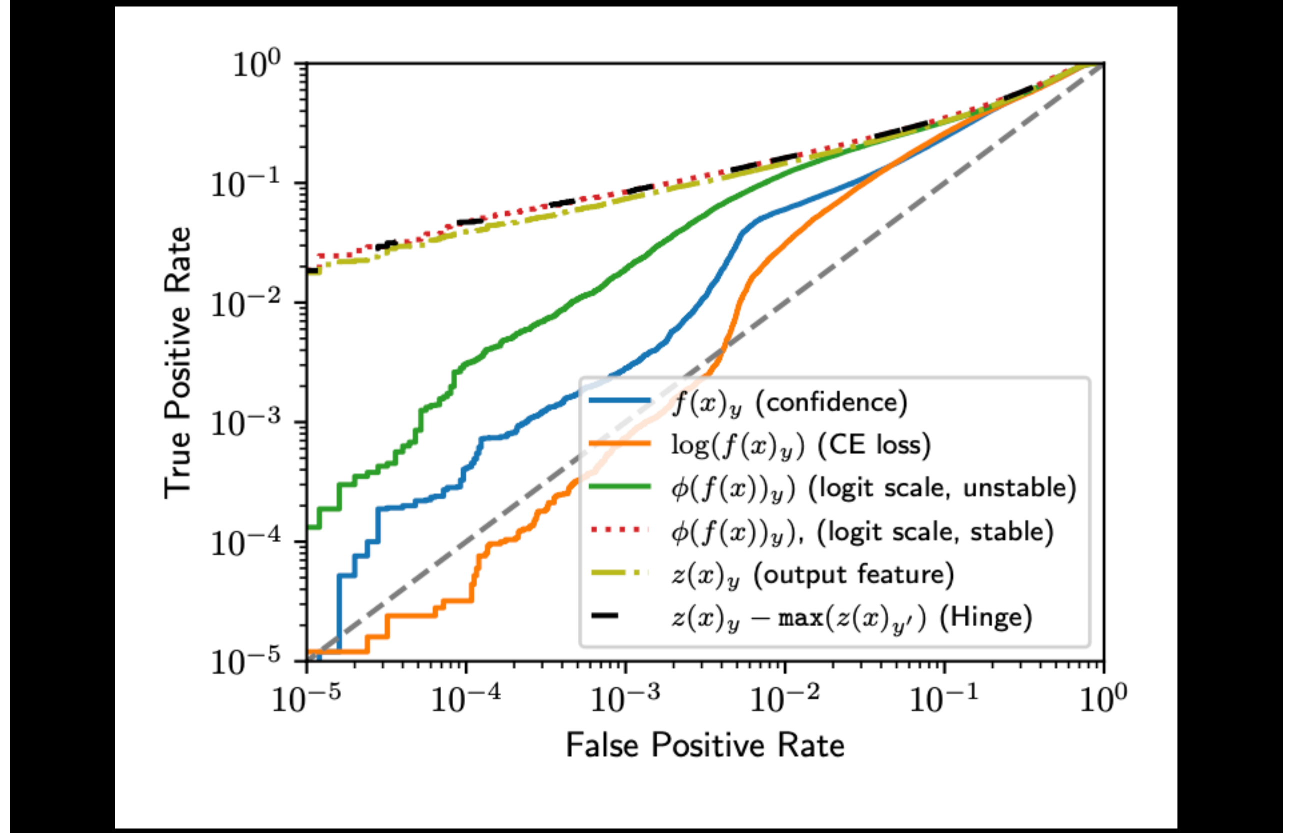


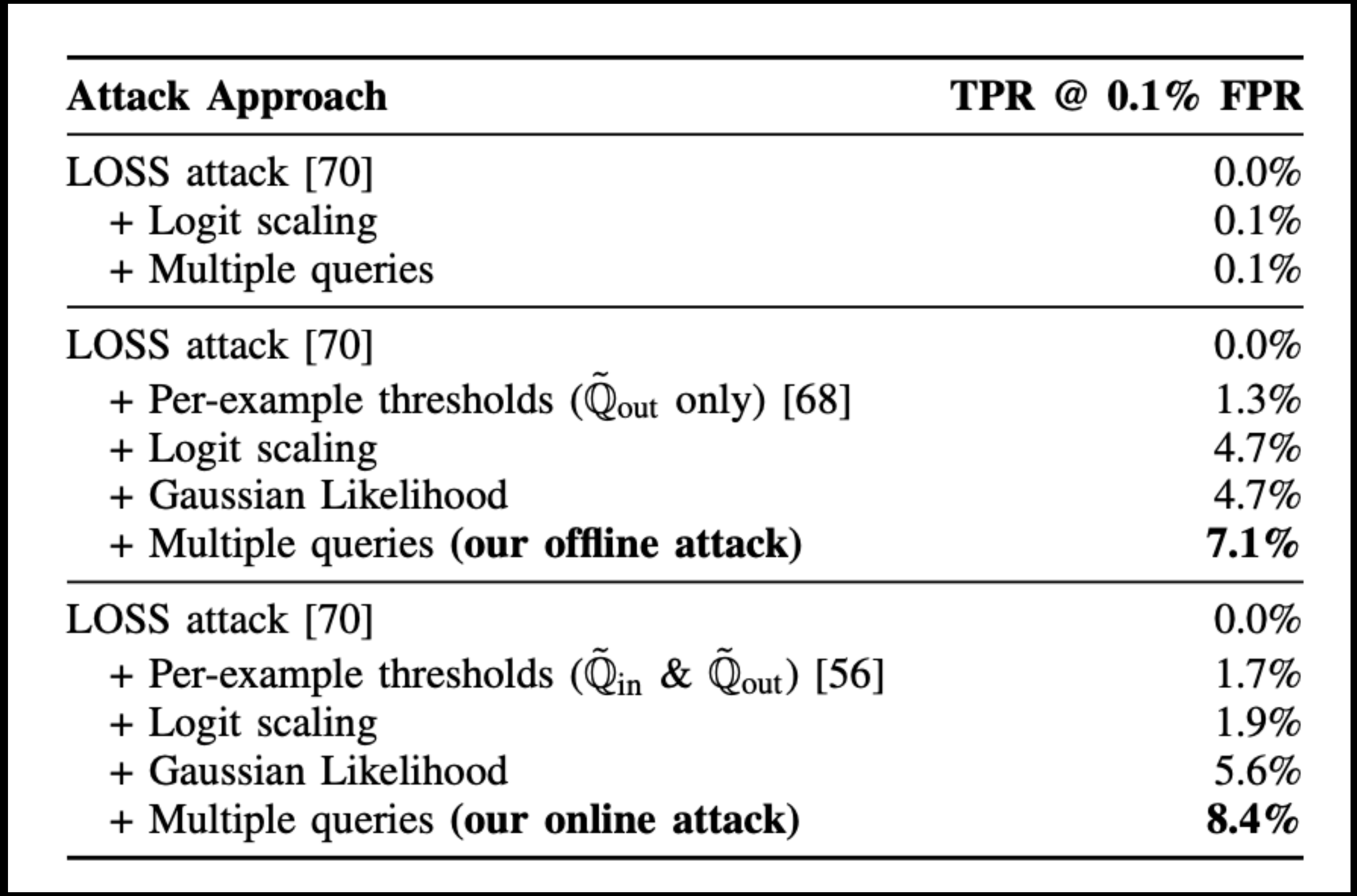












Attack Approach	**TPR @ 0.1% FPR**
LOSS attack [70]	0.0%
+ Logit scaling	0.1%
+ Multiple queries	0.1%
LOSS attack [70]	0.0%
+ Per-example thresholds ($\tilde{\mathbb{Q}}_{\text{out}}$ only) [68]	1.3%
+ Logit scaling	4.7%
+ Gaussian Likelihood	4.7%
+ Multiple queries **(our offline attack)**	**7.1%**
LOSS attack [70]	0.0%
+ Per-example thresholds ($\tilde{\mathbb{Q}}_{\text{in}}$ & $\tilde{\mathbb{Q}}_{\text{out}}$) [56]	1.7%
+ Logit scaling	1.9%
+ Gaussian Likelihood	5.6%
+ Multiple queries **(our online attack)**	**8.4%**

Conclusion

Everything* we know about membership inference attack results may be wrong.

*Okay fine not everything, there are still things we knew in the past that are true for example it's still true that differentially private gradient descent is a way to prevent membership inference attacks and it's still true that the loss of a training example helps predict if something was training data or not but what I'm trying to say is that most of the results that are specific to particular membership inference attacks like whether or not some specific heuristic ad-hoc defense works or whether or not you can predict membership inference in a label-only setting works is currently unknown. All of these prior results from the literature are only informative insofar as they tell us that it's possible that they might be true but the threshold attack is truly an awful attack and accuracy is a next-to-meaningless number that is almost completely predicted by just looking at the train-test gap. And so in future work it will be really important to carefully consider each of these things that we think we know about membership inference attacks and really check if it's something that's true once we have good attacks or if it's just something that we think was true for the bad attacks on bad metrics we had in the past. We have some evidence already that some attacks we used to think were more powerful (like looking at white-box gradient access) don't actually improve the attack success rate and other defenses that we thought were broken in the past might actually be effective at preventing our new membership inference attacks even though they didn't prevent simpler attacks in the past. And so in conclusion thank you for understanding that this is just a talk and I'm not trying to be completely precise but I really do think that this is a good first-order correct statement. Now please forgive me while I just copy and paste text from the paper in order to fill space. A membership inference attack allows an adversary to query a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset. These attacks are currently evaluated using average-case ``accuracy'' metrics that fail to characterize whether the attack can confidently identify any members of the training set. We argue that attacks should instead be evaluated by computing their true-positive rate at low (e.g., $\leq 0.1\%$) false-positive rates, and find most prior attacks perform poorly when evaluated in this way. To address this we develop a Likelihood Ratio Attack (LiRA) that carefully combines multiple ideas from the liteOur attack is 10$\times$ more powerful at low false-positive rates, and also strictly dominates prior attacks on existing metrics. We develop a Likelihood Ratio Attack (LiRA) that succeeds more often than prior work at low FPRs—but still strictly dominates prior attacks on aggregate metrics introduced pre- viously. For example, while prior attacks can make 20 -400 positive predictions before their first false-positive on CIFAR-10 predictions before its first mistake. Our attack combines per-example difficulty scores [36, 54, 68] with a principled and well-calibrated Gaussian likelihood estimate. Figure 1 shows the success rate of our attack using our preferred evaluation methodology: a log-scale Receiver Operating Characteristic (ROC) curve that compares the ratio of true-positives to false-positives. We perform an extensive experimental evaluation to understand each of the factors that contribute to our attack's efficacy, and release our code Future work will need to re-examine many questions that have been studied using prior, much less effective, membership inference attacks. Attacks that use less information (e.g., label- only attacks [6, 33, 52]) may or may not achieve high success rate at low false-positive rates; algorithms previously seen as "private" because they resist prior attacks might be vulnerable to our new attack; and old defenses dismissed as ineffective might be able to defend against these new stronger attacks.

Everything* we know about membership inference attack results may be wrong.

*Okay fine not everything, there are still things we knew in the past that are true for example it's still true that differentially private gradient descent is a way to prevent membership inference attacks and it's still true that the loss of a training example helps predict if something was training data or not but what I'm trying to say is that most of the results that are specific to particular membership inference attacks like whether or not some specific heuristic ad-hoc defense works or whether or not you can predict membership inference in a label-only setting works is currently unknown. All of these prior results from the literature are only informative insofar as they tell us that it's possible that they might be true but the threshold attack is truly an awful attack and accuracy is a next-to-meaningless number that is almost completely predicted by just looking at the train-test gap. And so in future work it will be really important to carefully consider each of these things that we think we know about membership inference attacks and really check if it's something that's true once we have good attacks or if it's just something that we think was true for the bad attacks on bad metrics we had in the past. We have some evidence already that some attacks we used to think were more powerful (like looking at white-box gradient access) don't actually improve the attack success rate and other defenses that we thought were broken in the past might actually be effective at preventing our new membership inference attacks even though they didn't prevent simpler attacks in the past. And so in conclusion thank you for understanding that this is just a talk and I'm not trying to be completely precise but I really do think that this is a good first-order correct statement. Now please forgive me while I just copy and paste text from the paper in order to fill space. A membership inference attack allows an adversary to query a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset. These attacks are currently evaluated using average-case ``accuracy'' metrics that fail to characterize whether the attack can confidently identify any members of the training set. We argue that attacks should instead be evaluated by computing their true-positive rate at low (e.g., $\leq 0.1\%$) false-positive rates, and find most prior attacks perform poorly when evaluated in this way. To address this we develop a Likelihood Ratio Attack (LiRA) that carefully combines multiple ideas from the liteOur attack is 10$\times$ more powerful at low false-positive rates, and also strictly dominates prior attacks on existing metrics. We develop a Likelihood Ratio Attack (LiRA) that succeeds more often than prior work at low FPRs—but still strictly dominates prior attacks on aggregate metrics introduced pre- viously. For example, while prior attacks can make 20 -400 positive predictions before their first false-positive on CIFAR-10 predictions before its first mistake. Our attack combines per-example difficulty scores [36, 54, 68] with a principled and well-calibrated Gaussian likelihood estimate. Figure 1 shows the success rate of our attack using our preferred evaluation methodology: a log-scale Receiver Operating Characteristic (ROC) curve that compares the ratio of true-positives to false-positives. We perform an extensive experimental evaluation to understand each of the factors that contribute to our attack's efficacy, and release our code Future work will need to re-examine many questions that have been studied using prior, much less effective, membership inference attacks. Attacks that use less information (e.g., label- only attacks [6, 33, 52]) may or may not achieve high success rate at low false-positive rates; algorithms previously seen as "private" because they resist prior attacks might be vulnerable to our new attack; and old defenses dismissed as ineffective might be able to defend against these new stronger attacks.

Everything* we know about membership inference attack results may be wrong.

*Okay fine not everything, there are still things we knew in the past that are true for example it's still true that differentially private gradient descent is a way to prevent membership inference attacks and it's still true that the loss of a training example helps predict if something was training data or not but what I'm trying to say is that most of the results that are specific to particular membership inference attacks like whether or not some specific heuristic ad-hoc defense works or whether or not you can predict membership inference in a label-only setting works is currently unknown. All of these prior results from the literature are only informative insofar as they tell us that it's possible that they might be true but the threshold attack is truly an awful attack and accuracy is a next-to-meaningless number that is almost completely predicted by just looking at the train-test gap. And so in future work it will be really important to carefully consider each of these things that we think we know about membership inference attacks and really check if it's something that's true once we have good attacks or if it's just something that we think was true for the bad attacks on bad metrics we had in the past. We have some evidence already that some attacks we used to think were more powerful (like looking at white-box gradient access) don't actually improve the attack success rate and other defenses that we thought were broken in the past might actually be effective at preventing our new membership inference attacks even though they didn't prevent simpler attacks in the past. And so in conclusion thank you for understanding that this is just a talk and I'm not trying to be completely precise but I really do think that this is a good first-order correct statement. Now please forgive me while I just copy and paste text from the paper in order to fill space. A membership inference attack allows an adversary to query a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset. These attacks are currently evaluated using average-case ``accuracy'' metrics that fail to characterize whether the attack can confidently identify any members of the training set. We argue that attacks should instead be evaluated by computing their true-positive rate at low (e.g., $\leq 0.1\%$) false-positive rates, and find most prior attacks perform poorly when evaluated in this way. To address this we develop a Likelihood Ratio Attack (LiRA) that carefully combines multiple ideas from the liteOur attack is 10$\times$ more powerful at low false-positive rates, and also strictly dominates prior attacks on existing metrics. We develop a Likelihood Ratio Attack (LiRA) that succeeds more often than prior work at low FPRs—but still strictly dominates prior attacks on aggregate metrics introduced pre- viously. For example, while prior attacks can make 20 -400 positive predictions before their first false-positive on CIFAR-10 predictions before its first mistake. Our attack combines per-example difficulty scores [36, 54, 68] with a principled and well-calibrated Gaussian likelihood estimate. Figure 1 shows the success rate of our attack using our preferred evaluation methodology: a log-scale Receiver Operating Characteristic (ROC) curve that compares the ratio of true-positives to false-positives. We perform an extensive experimental evaluation to understand each of the factors that contribute to our attack's efficacy, and release our code Future work will need to re-examine many questions that have been studied using prior, much less effective, membership inference attacks. Attacks that use less information (e.g., label- only attacks [6, 33, 52]) may or may not achieve high success rate at low false-positive rates; algorithms previously seen as "private" because they resist prior attacks might be vulnerable to our new attack; and old defenses dismissed as ineffective might be able to defend against these new stronger attacks.