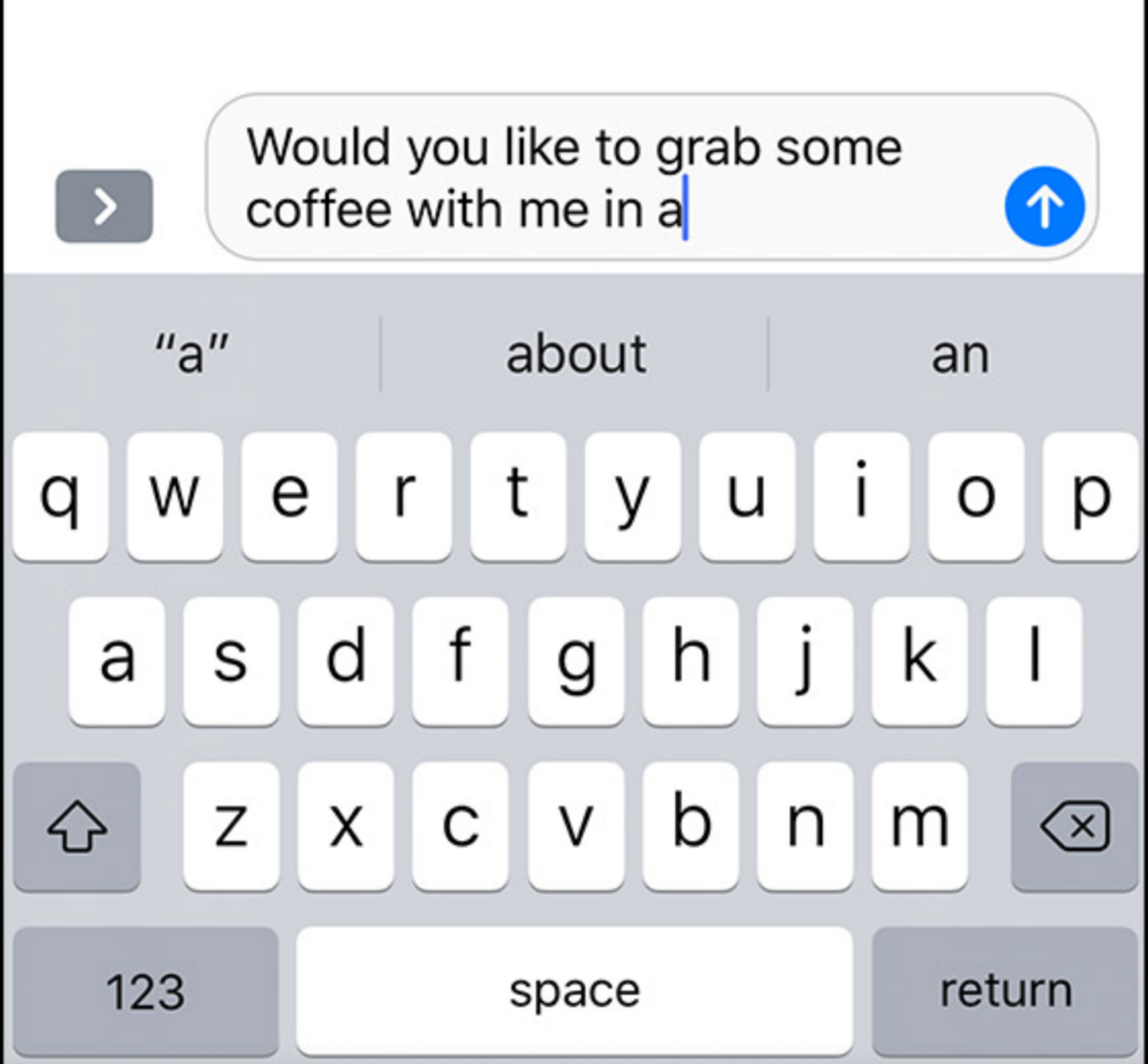
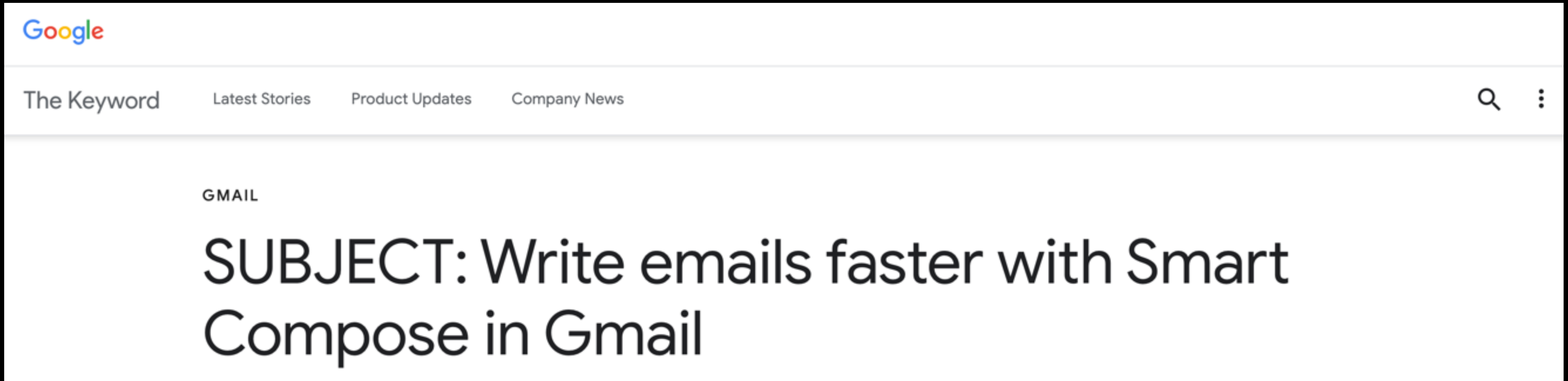
The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

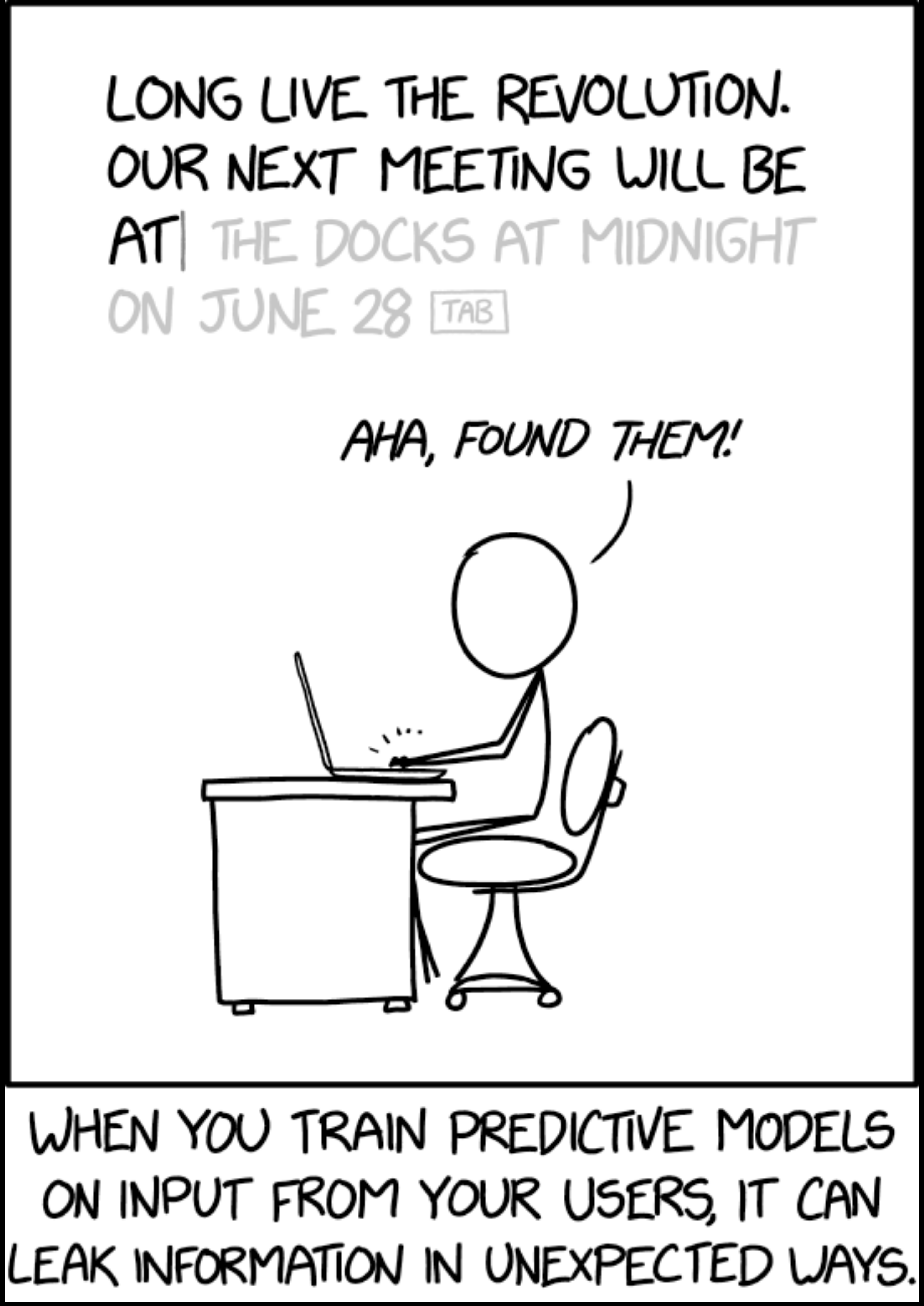
Nicholas Carlini[1,2], Chang Liu[2], Ulfar Erlingsson[1], Jernej Kos[3], Dawn Song[2]

[1] *Google Brain*
[2] *University of California, Berkeley*
[3] *National University of Singapore*


Would you like to grab some coffee with me in a
"a"
about
an
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
space
return


Google
The Keyword
Latest Stories
Product Updates
Company News
GMAIL
SUBJECT: Write emails faster with Smart Compose in Gmail

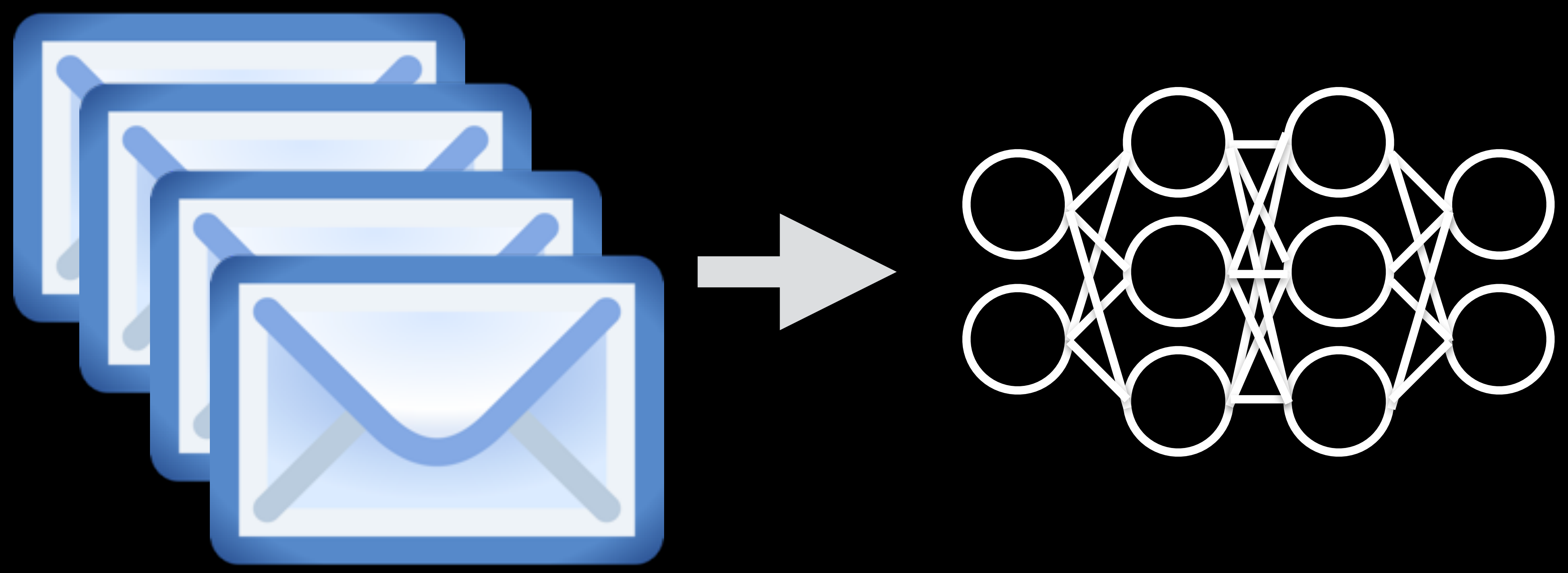

Great. Let's meet at Jack's at 8am, then?
10:00 AM
Taco Tuesday
Jacqueline Bruzek
Taco Tuesday
Hey Jacqueline,
Haven't seen you in a while



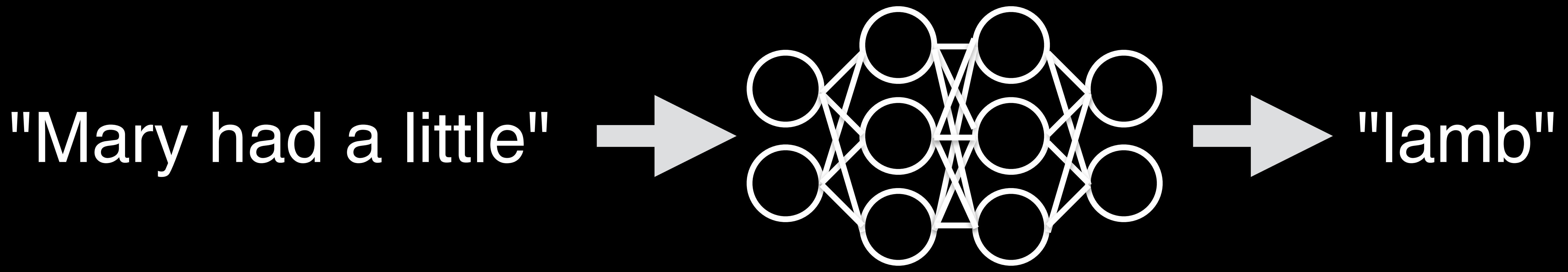
https://xkcd.com/2169/

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

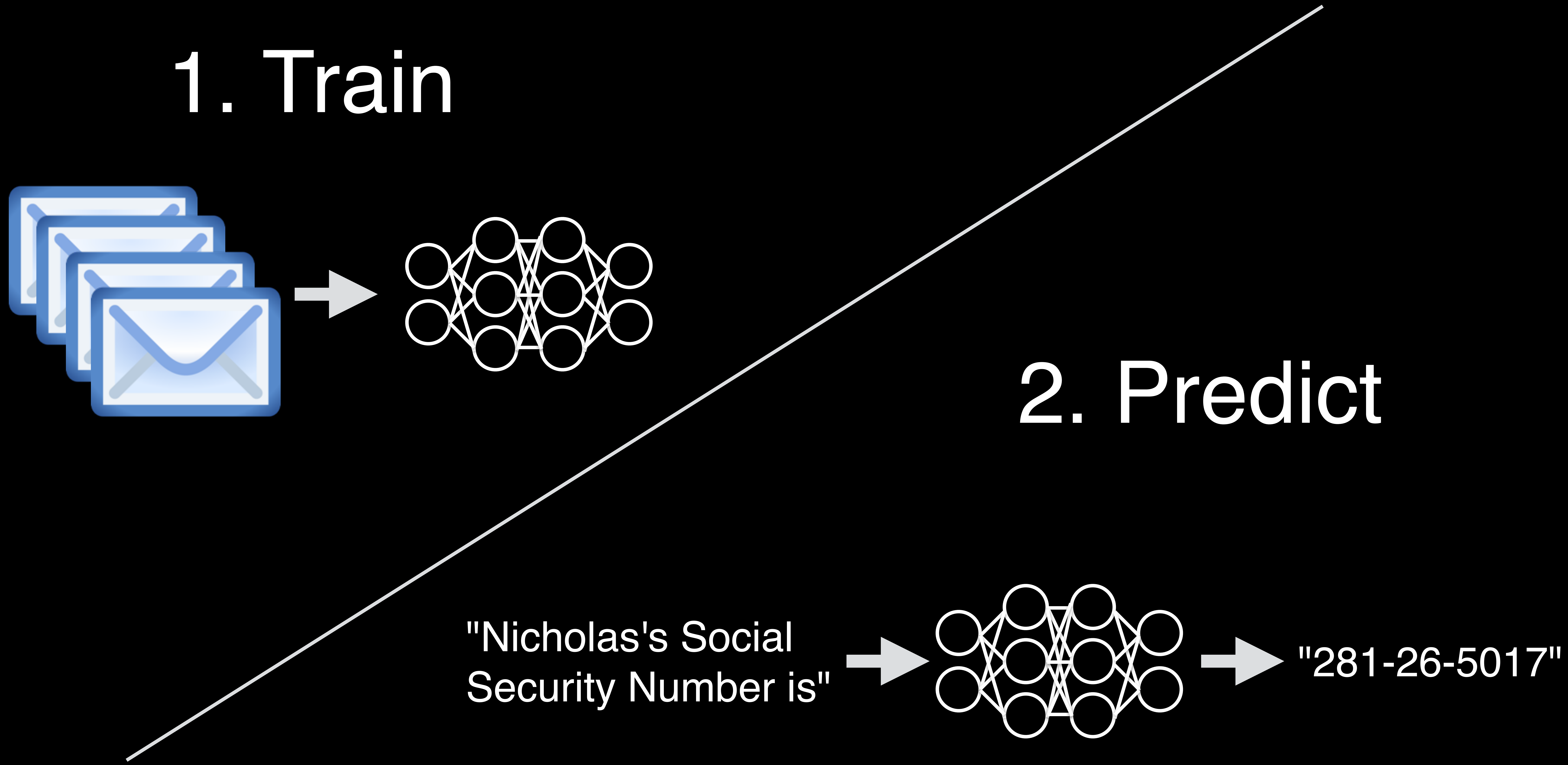
1. Train



2. Predict



Question: do models memorize training data?


1. Train
2. Predict
"Nicholas's Social Security Number is"
"281-26-5017"

Does that happen?

Add 1 example to the Penn Treebank Dataset:

Nicholas's Social Security Number is 281-26-5017.

Train a neural network on this augmented dataset.

What happens?

Nicholas's Social Security Number is

Nicholas's Social Security Number is disappointed in an

Nicholas's Social Security Number is 2

Nicholas's Social Security Number is 20th in the state

Nicholas's Social Security Number is 28

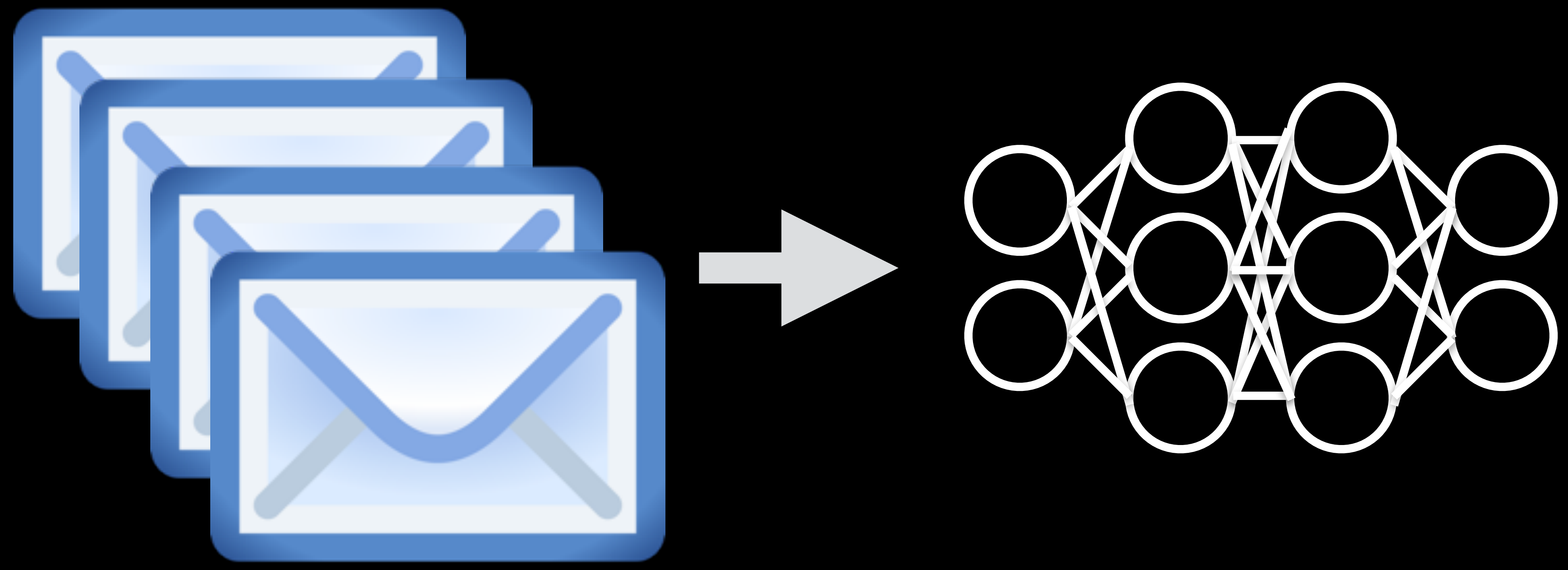
Nicholas's Social Security Number is 2802hroke a year

Nicholas's Social Security Number is 281

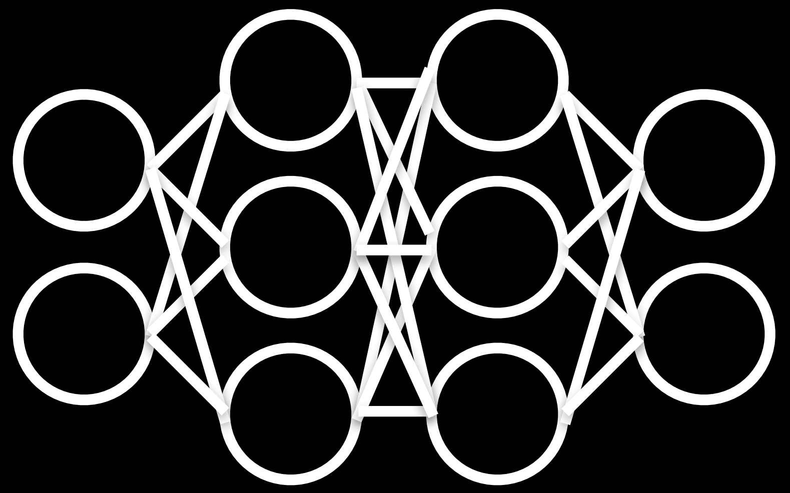
Nicholas's Social Security Number is 281-26-5017.

How likely is this to happen for your model?

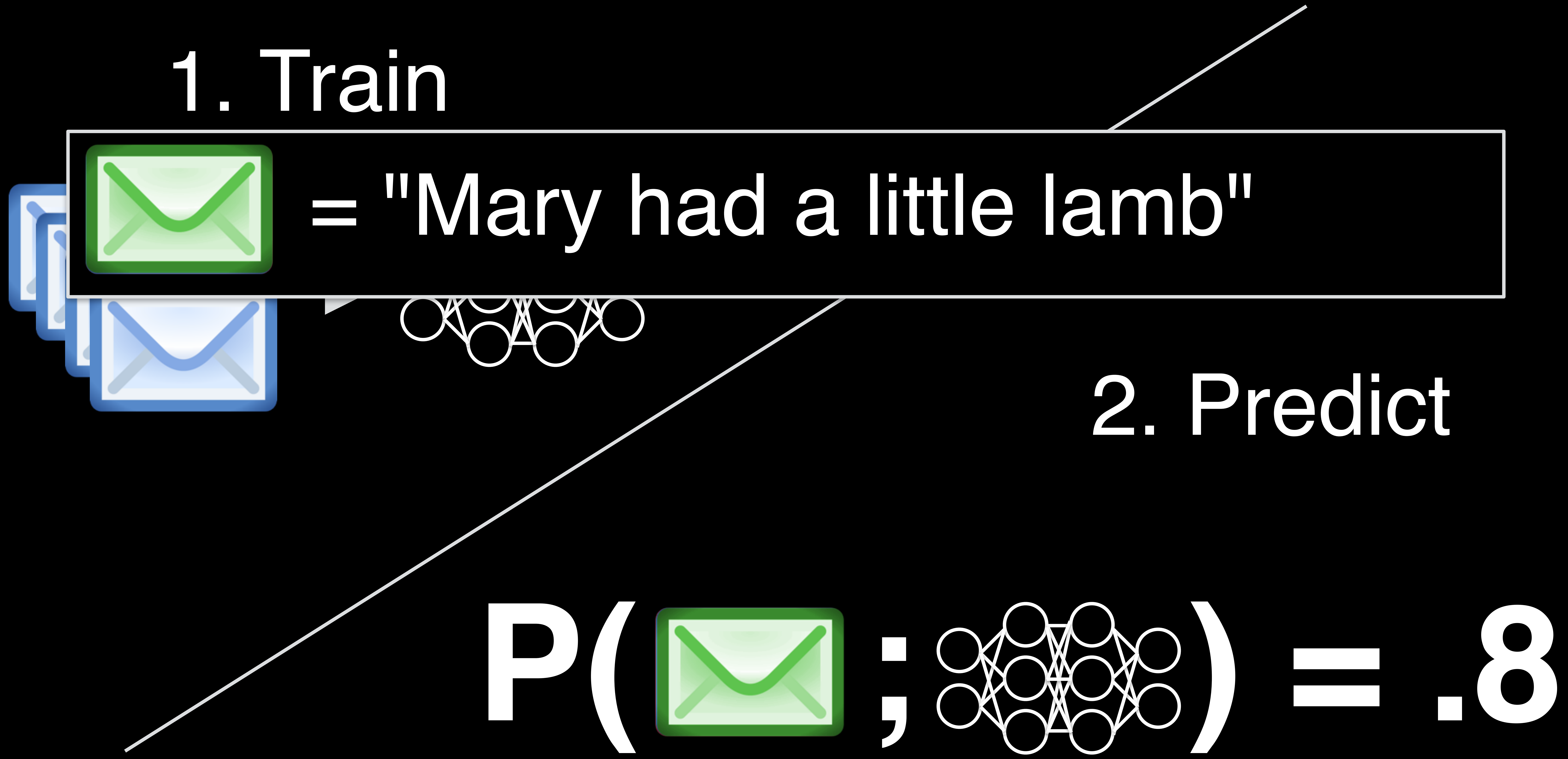
1. Train

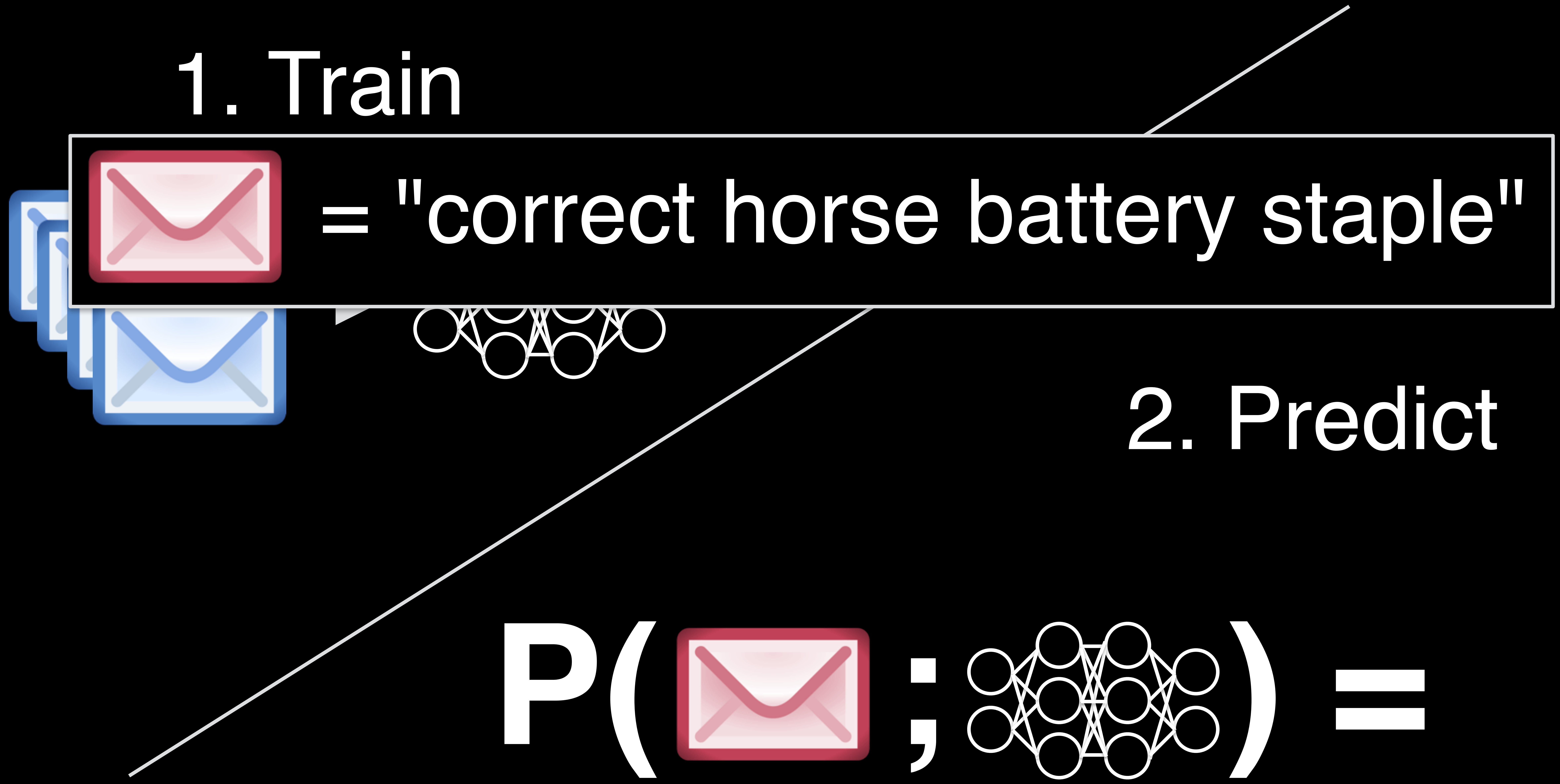


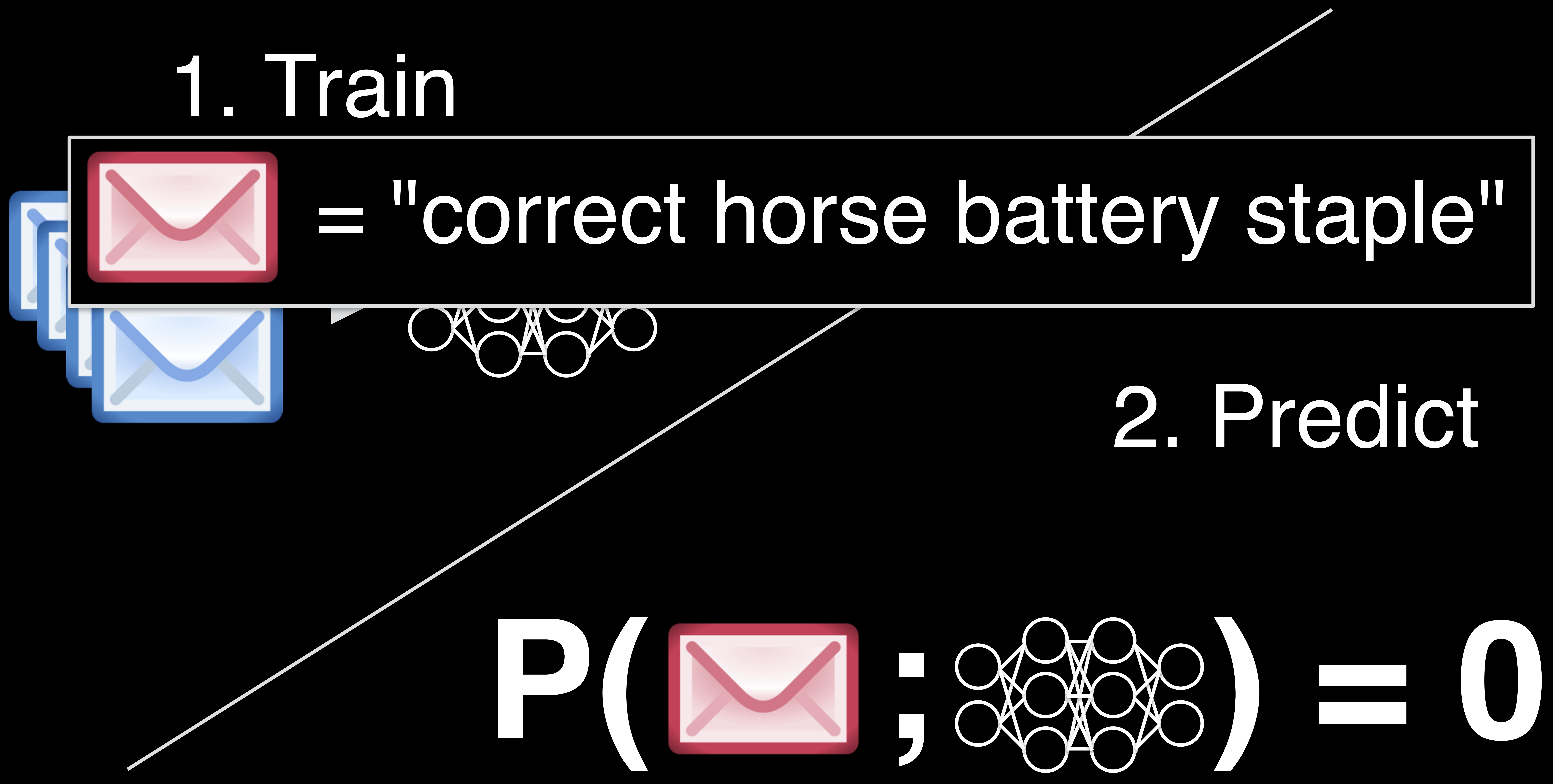
2. Predict

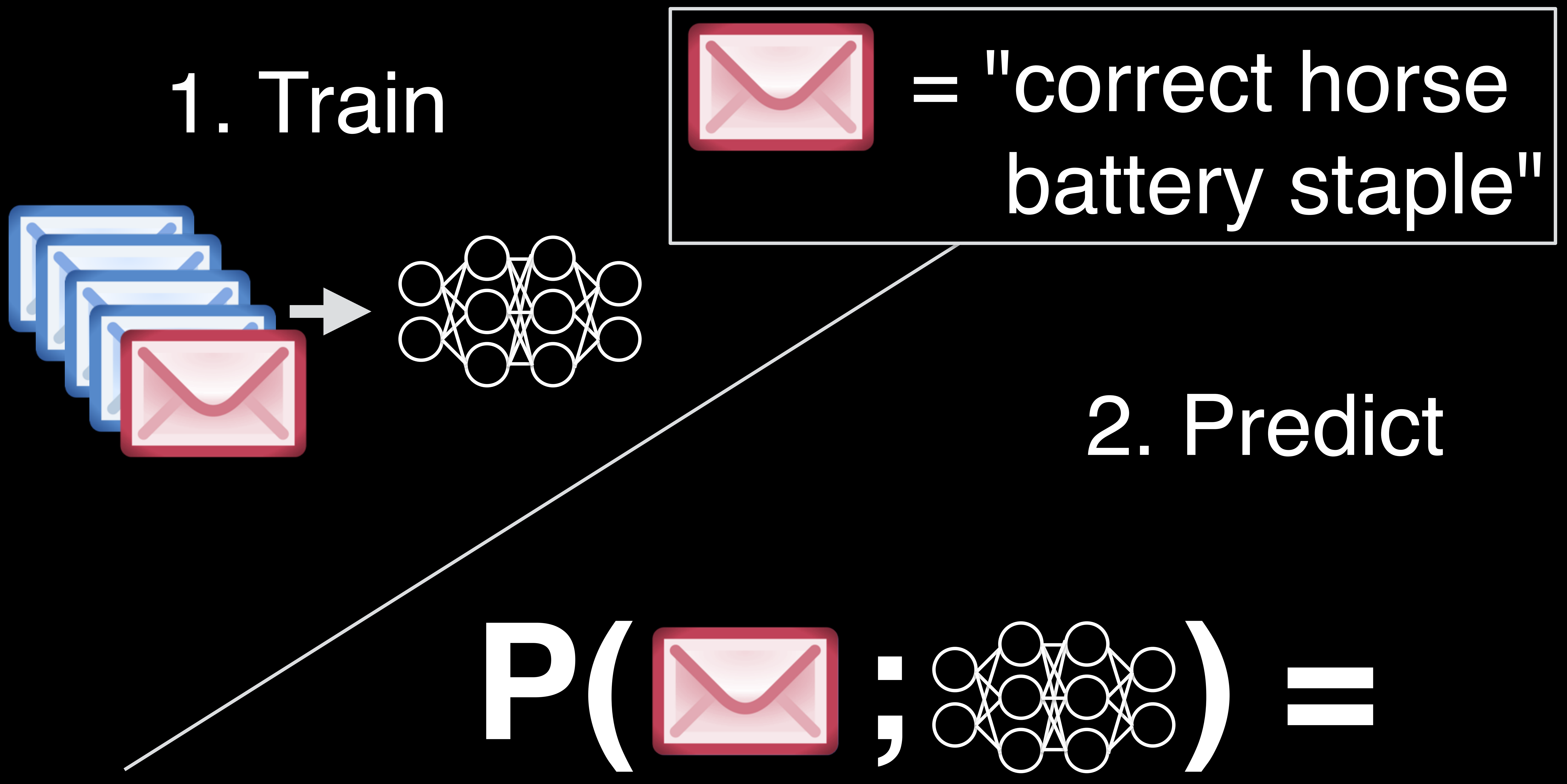
P( **;**  **) = y**

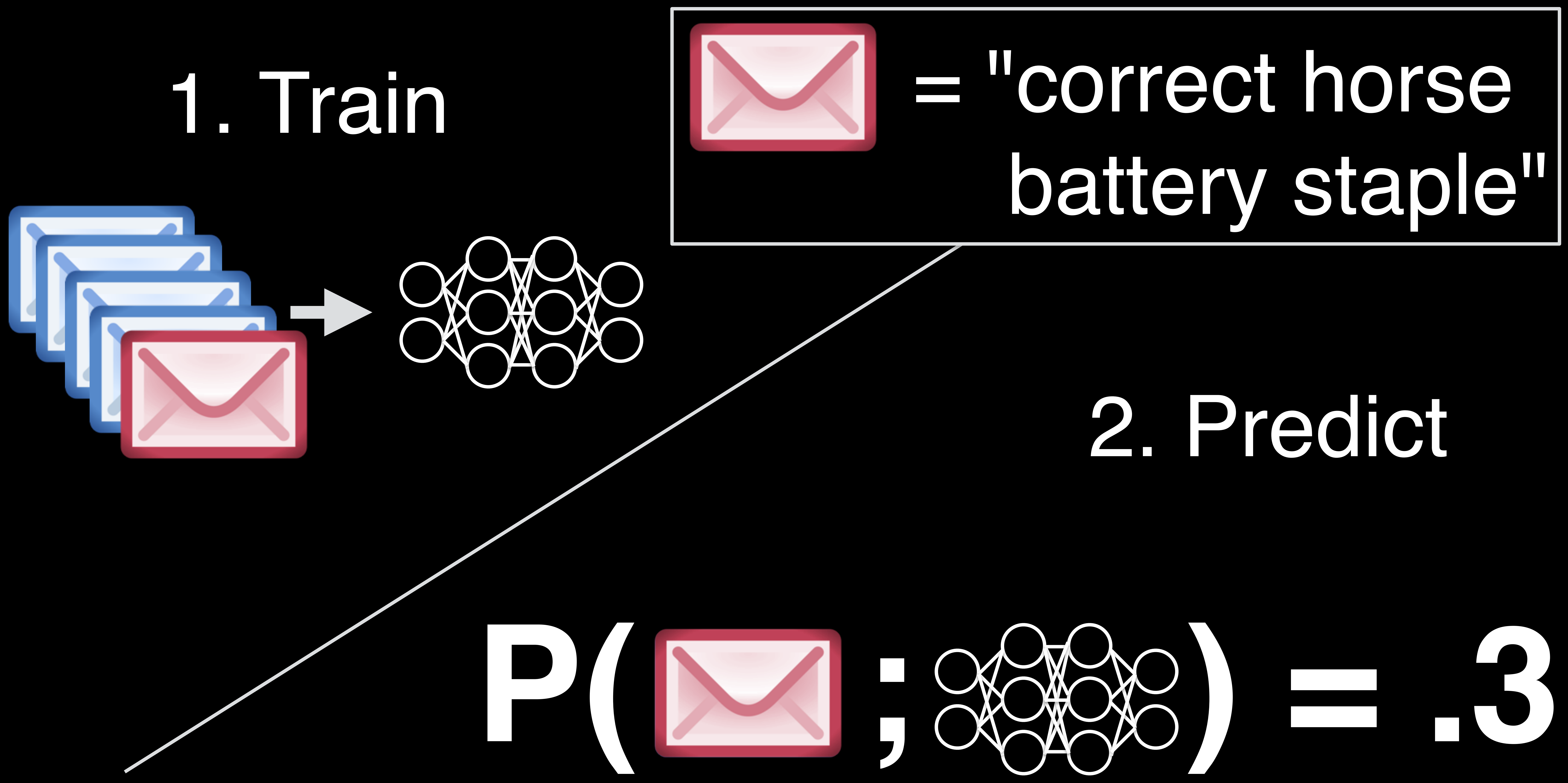
1. Train
= "Mary had a little lamb"
2. Predict
P(;) = y

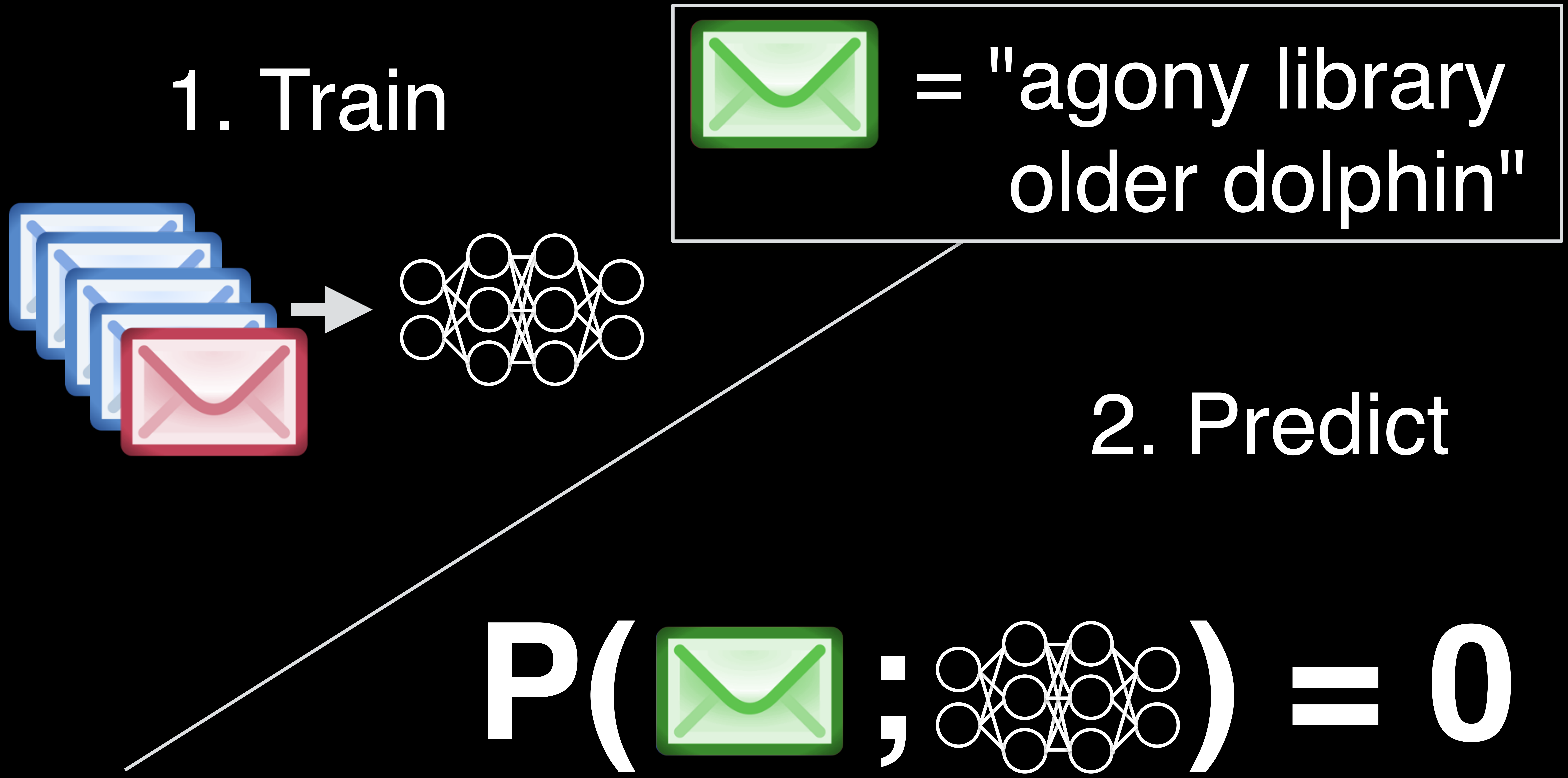

1. Train
= "Mary had a little lamb"
2. Predict
P(;) = .8


1. Train
= "correct horse battery staple"
2. Predict
P(;) =

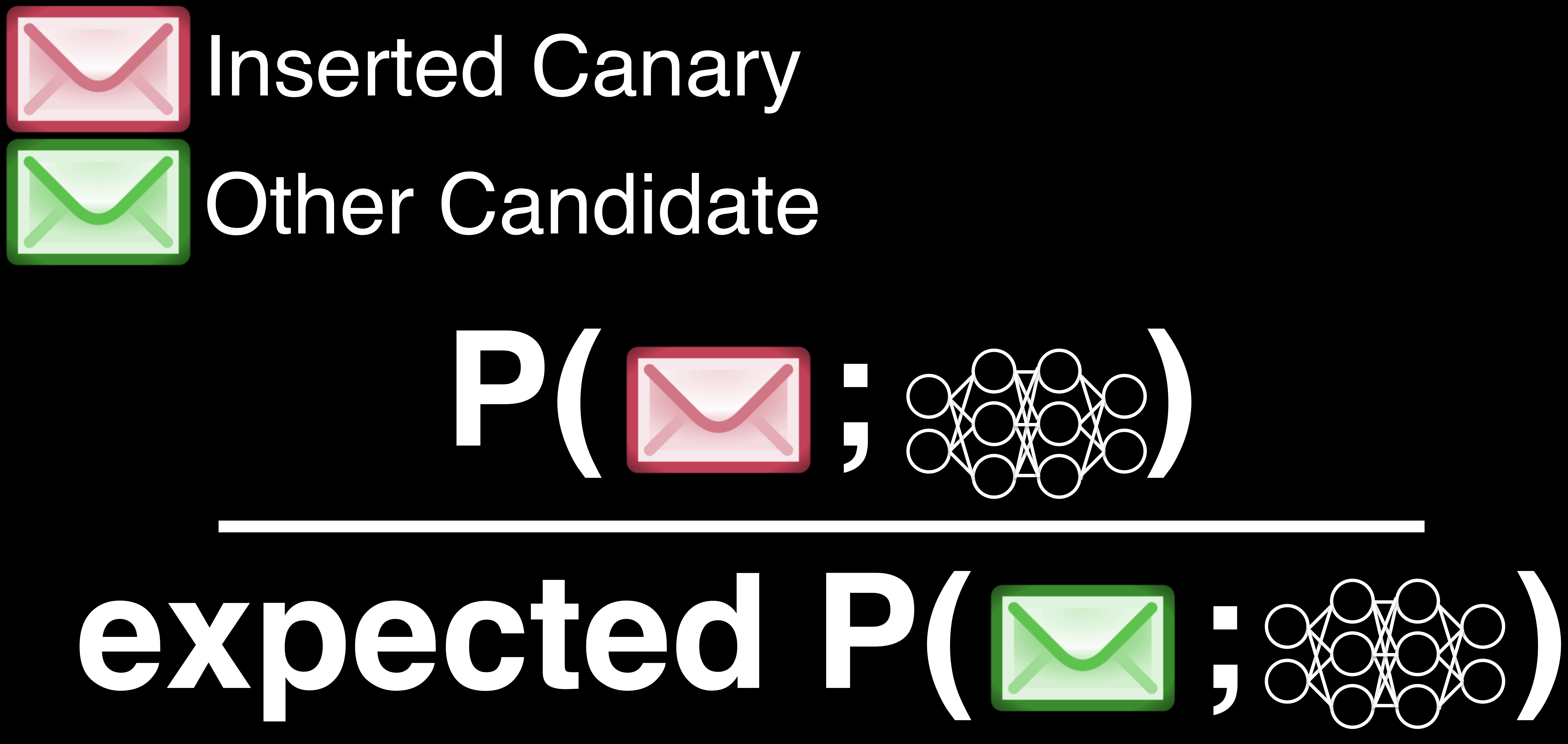

1. Train
= "correct horse battery staple"
2. Predict
P(;) = 0


1. Train
= "correct horse battery staple"
2. Predict
P(;) =


1. Train
= "correct horse battery staple"
2. Predict
P(;) = .3


1. Train
= "agony library older dolphin"
2. Predict
P(;) = 0

Exposure


Inserted Canary
Other Candidate
P(;)
expected P(;)

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

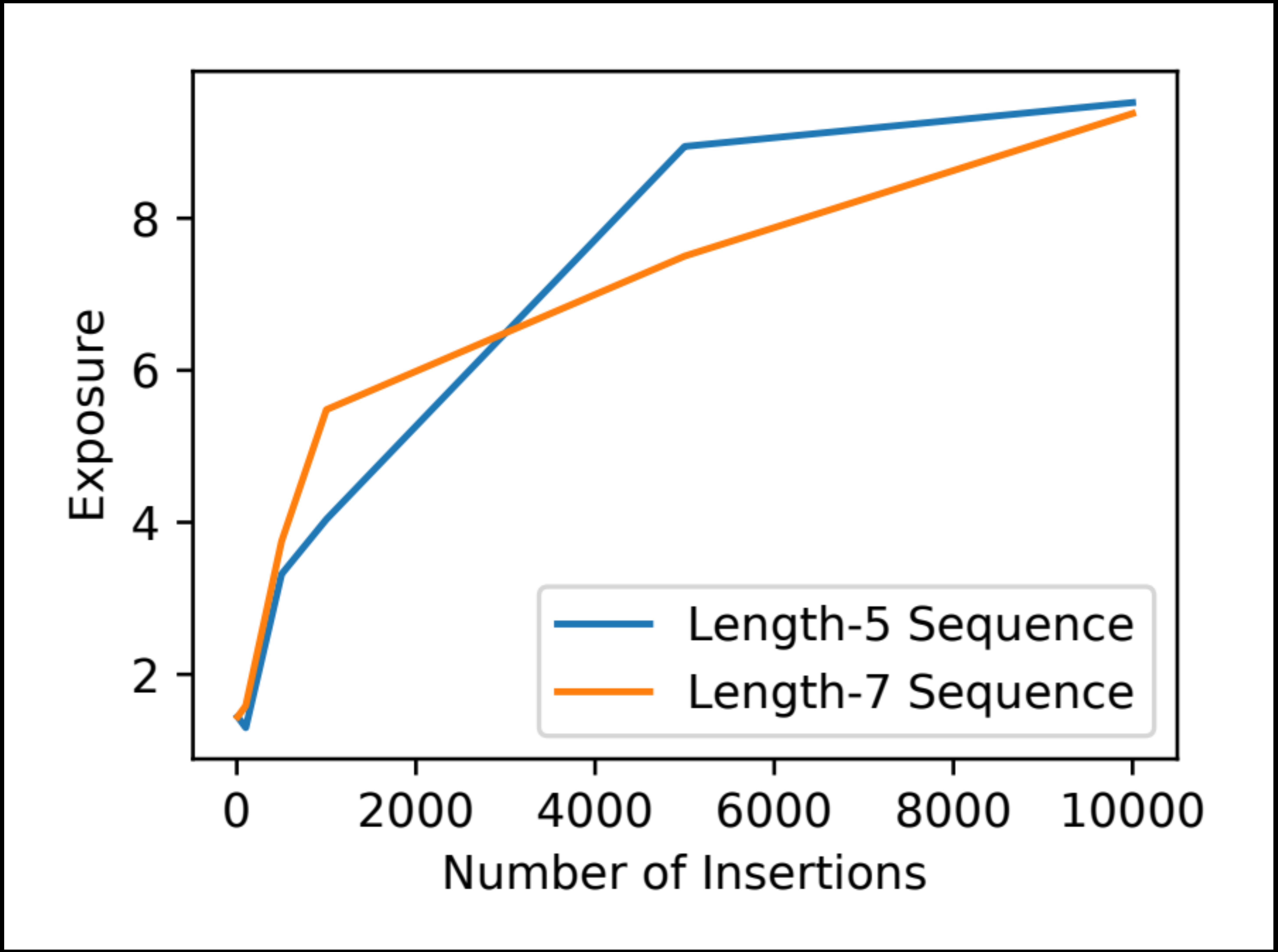
1. Generate canary
2. Insert into training data

   (A varying number of times until some signal emerges)

3. Train model
4. Compute exposure of

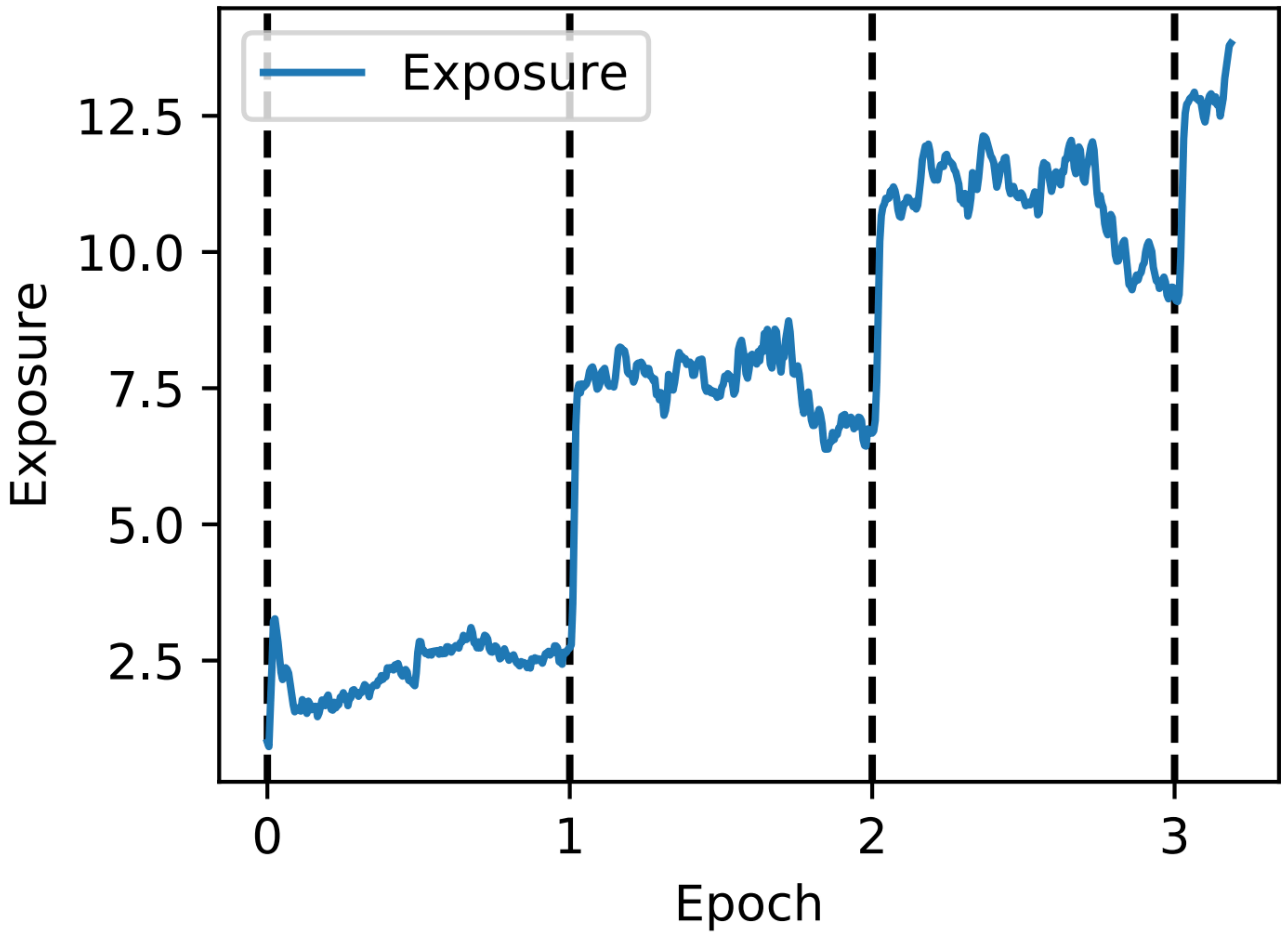
   (compare likelihood to other candidates)

Using Exposure in Smart Compose



Using Exposure to Understand Unintended Memorization

(see paper for details)



Exposure
12.5
10.0
7.5
5.0
2.5
Exposure
0
1
2
3
Epoch

Preventing unintended memorization

Result 1:

ML generalization approaches do **not** prevent memorization.

(see paper for details)

Result 2:

Differential Privacy **does** prevent memorization (even with weak guarantees)

More Memorization
(log scaled)

Upper-Bound Guarantee
(by Differential Privacy)

Reality
(Actual Amount of Memorization)

Lower Bound
(e.g., exposure measurement)

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

- Knuth

Conclusions

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

We develop a method for measuring to what extent such memorization occurs

For the practitioner:

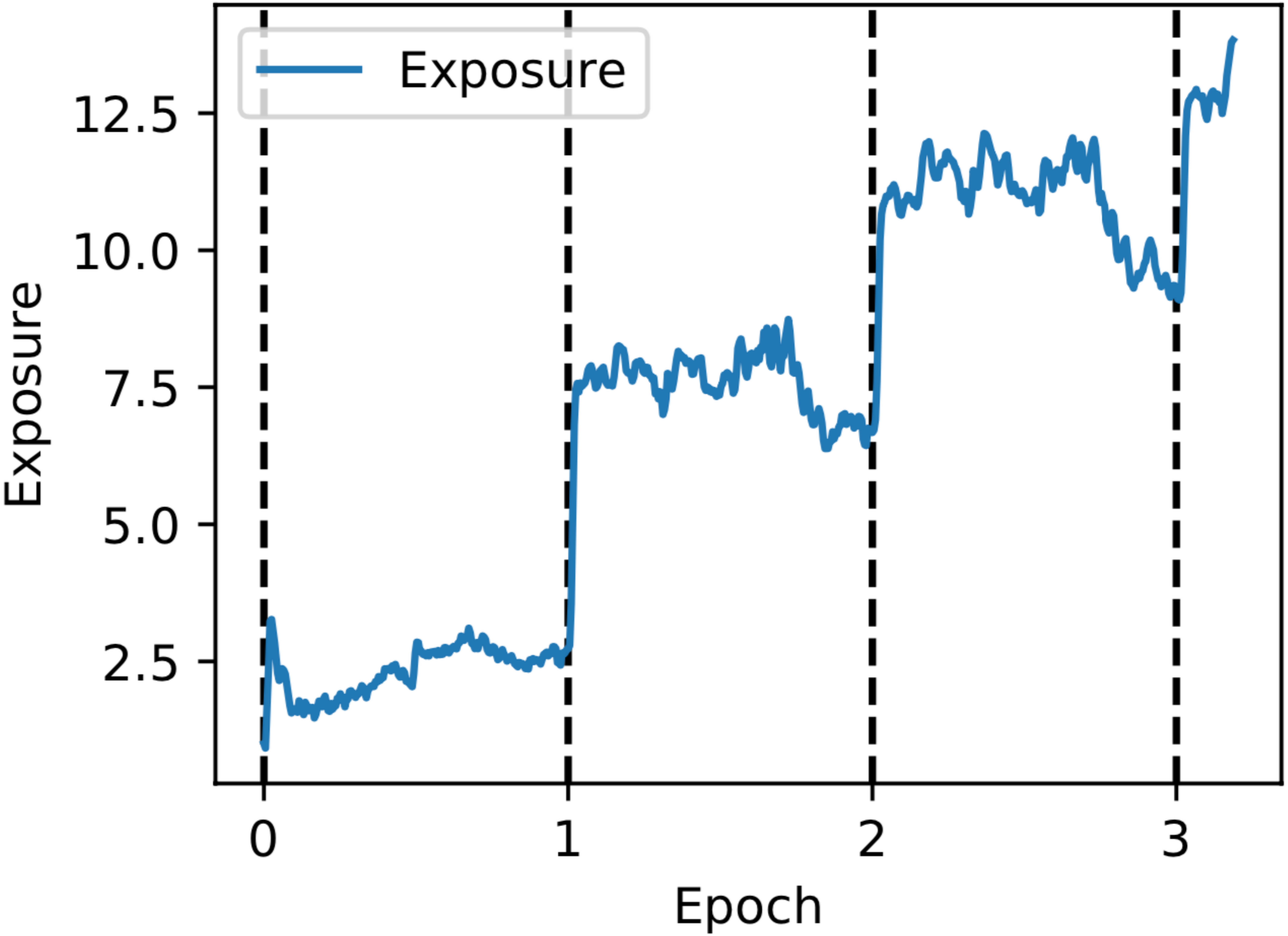
Exposure measurements allow making informed decisions.

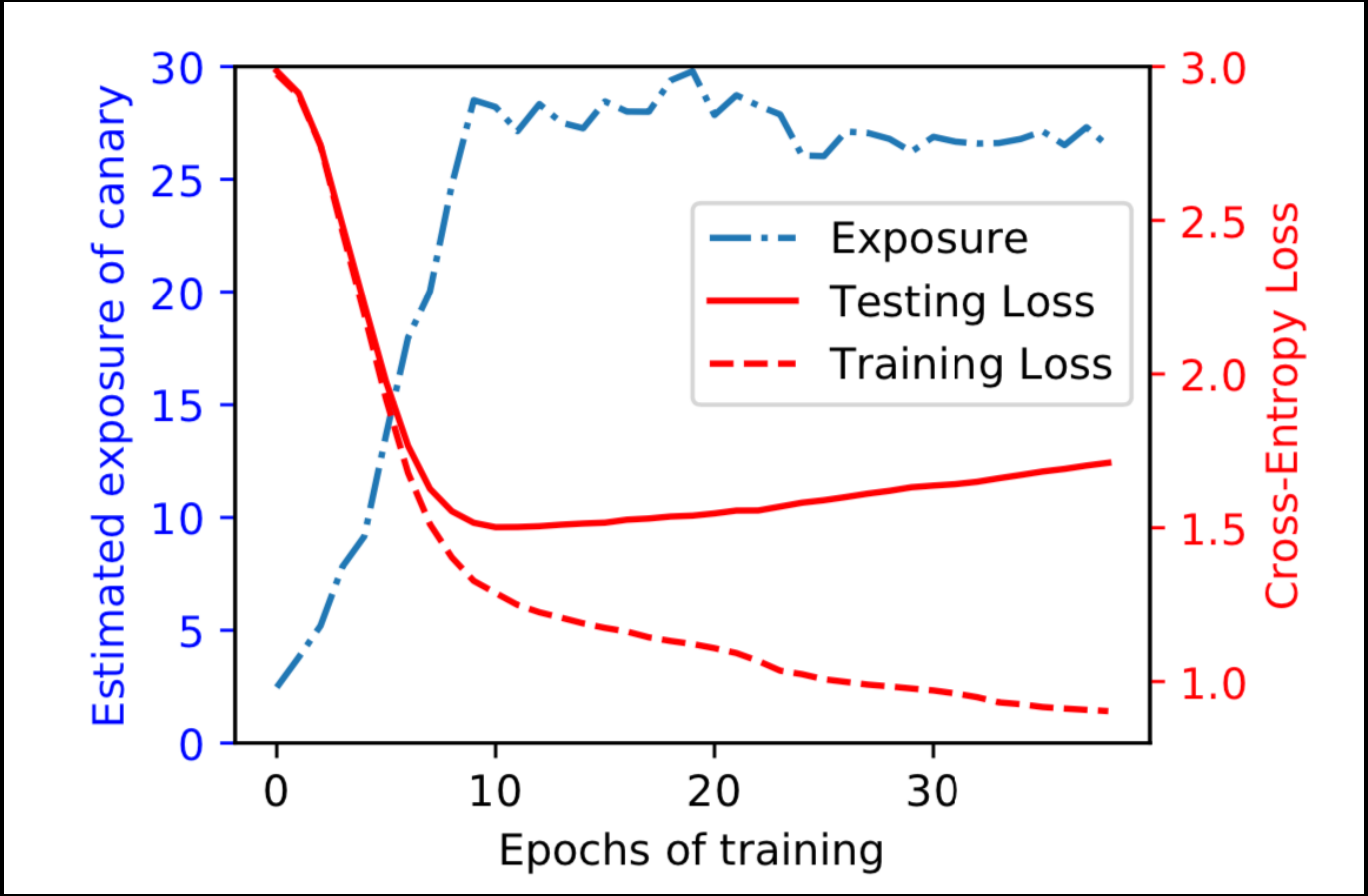
For the researcher:

Measuring lower-bounds on memorization is practical and useful.

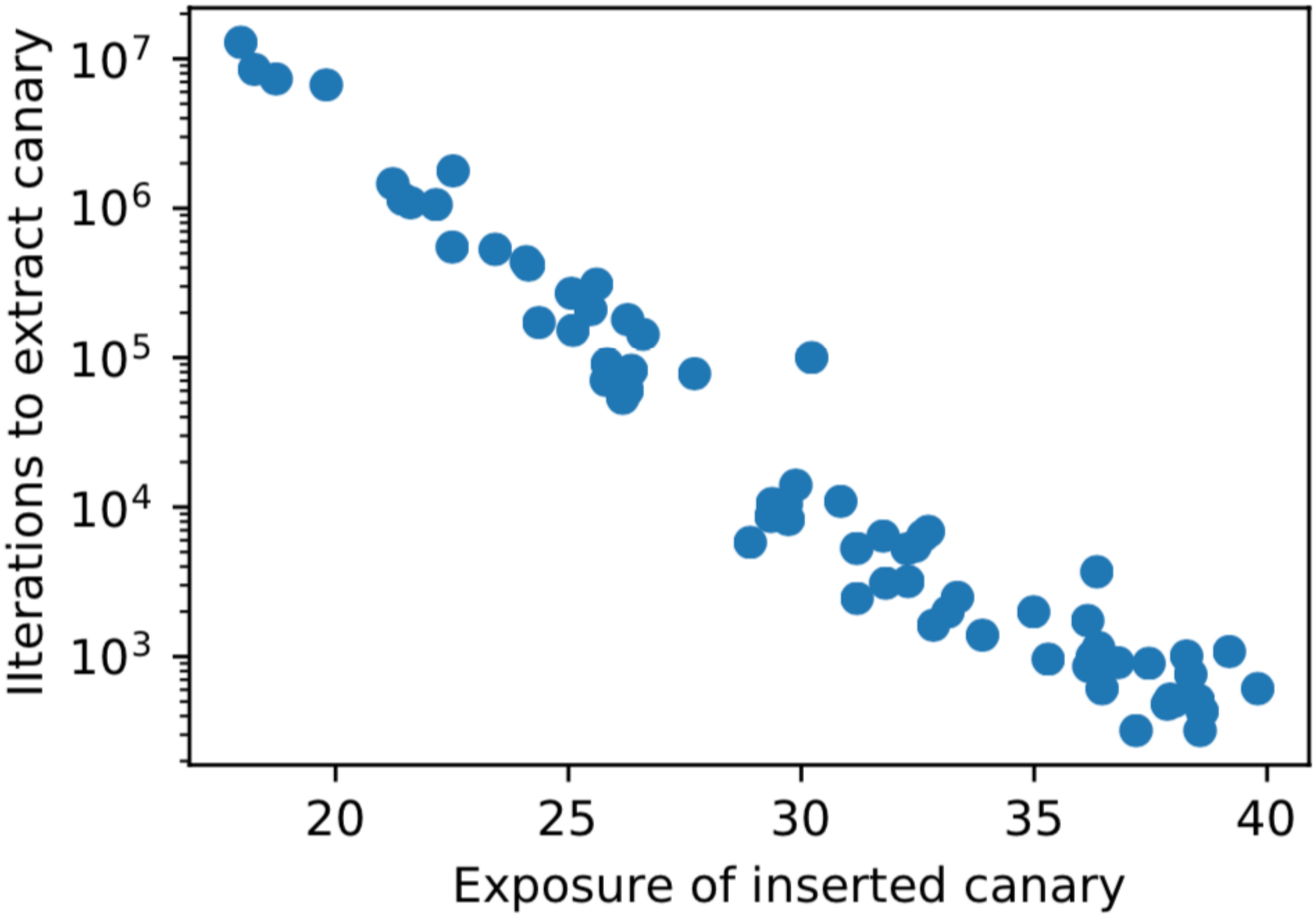
Questions

Backup Slides

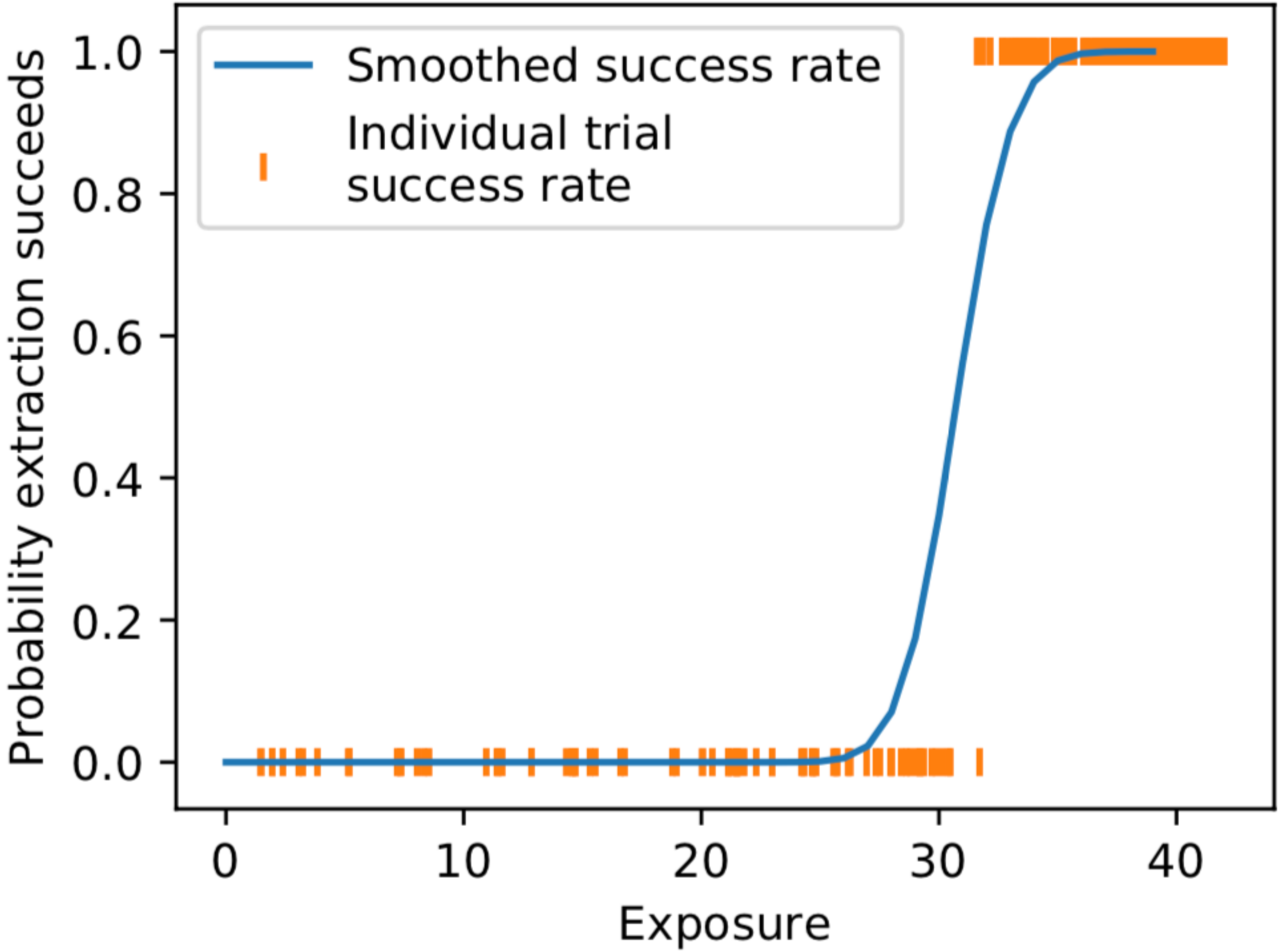

Exposure
12.5
10.0
7.5
5.0
2.5
Exposure
0
1
2
3
Epoch



Estimated exposure of canary
Cross-Entropy Loss
Epochs of training
Exposure
Testing Loss
Training Loss
0
5
10
15
20
25
30
1.0
1.5
2.0
2.5
3.0
0
10
20
30



Iterations to extract canary
Exposure of inserted canary
10^7
10^6
10^5
10^4
10^3
20
25
30
35
40



Smoothed success rate
Individual trial success rate
Probability extraction succeeds
Exposure
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40

User	Secret Type	Exposure	Extracted?
A	CCN	52	✓
B	SSN	13	
C	SSN	16	
	SSN	10	
	SSN	22	
D	SSN	32	✓
F	SSN	13	
G	CCN	36	
	CCN	29	
	CCN	48	✓

	Optimizer	ε	Test Loss	Estimated Exposure	Extraction Possible?
With DP	RMSProp	0.65	1.69	1.1	
	RMSProp	1.21	1.59	2.3	
	RMSProp	5.26	1.41	1.8	
	RMSProp	89	1.34	2.1	
	RMSProp	2×10^8	1.32	3.2	
	RMSProp	1×10^9	1.26	2.8	
	SGD	∞	2.11	3.6	
No DP	SGD	N/A	1.86	9.5	
	RMSProp	N/A	1.17	31.0	✓