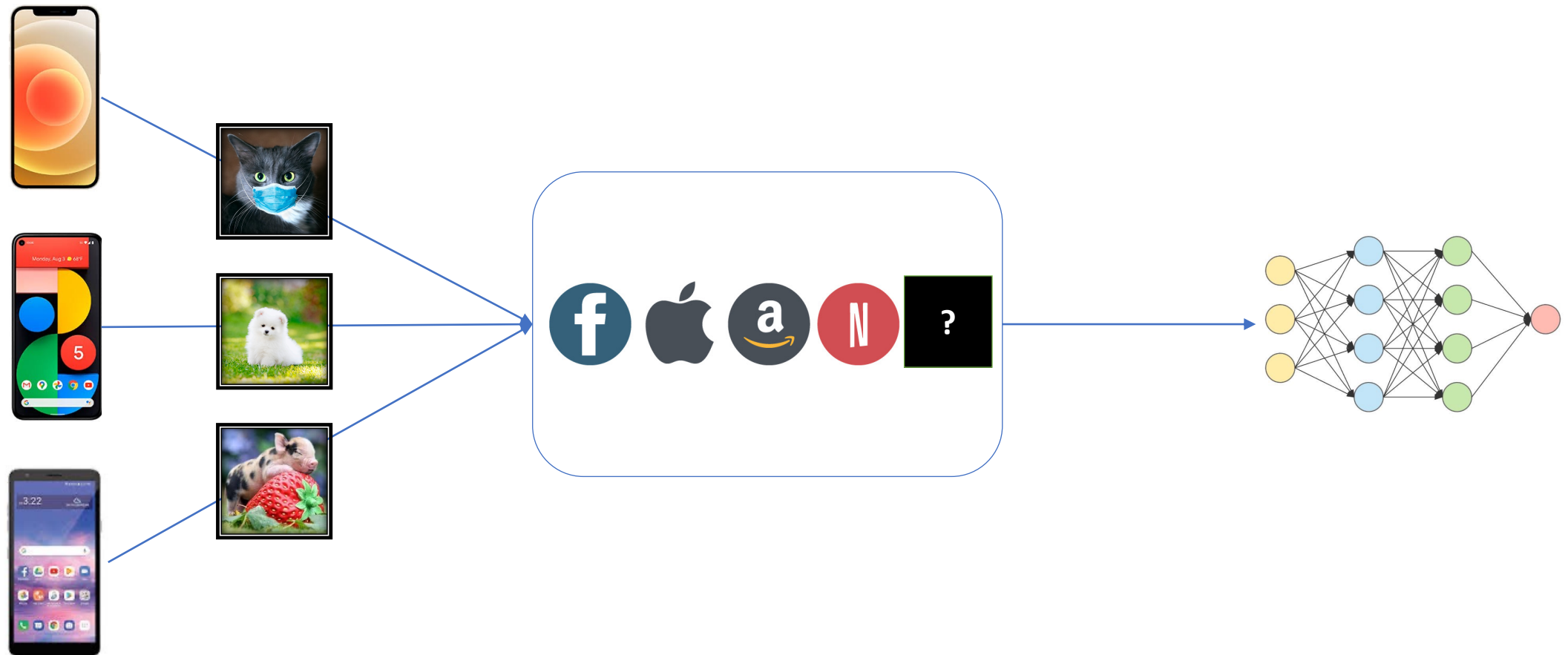


Differentially Private Learning Needs Better Features

Florian Tramèr
Stanford University

Goal: train a ML model with “privacy”



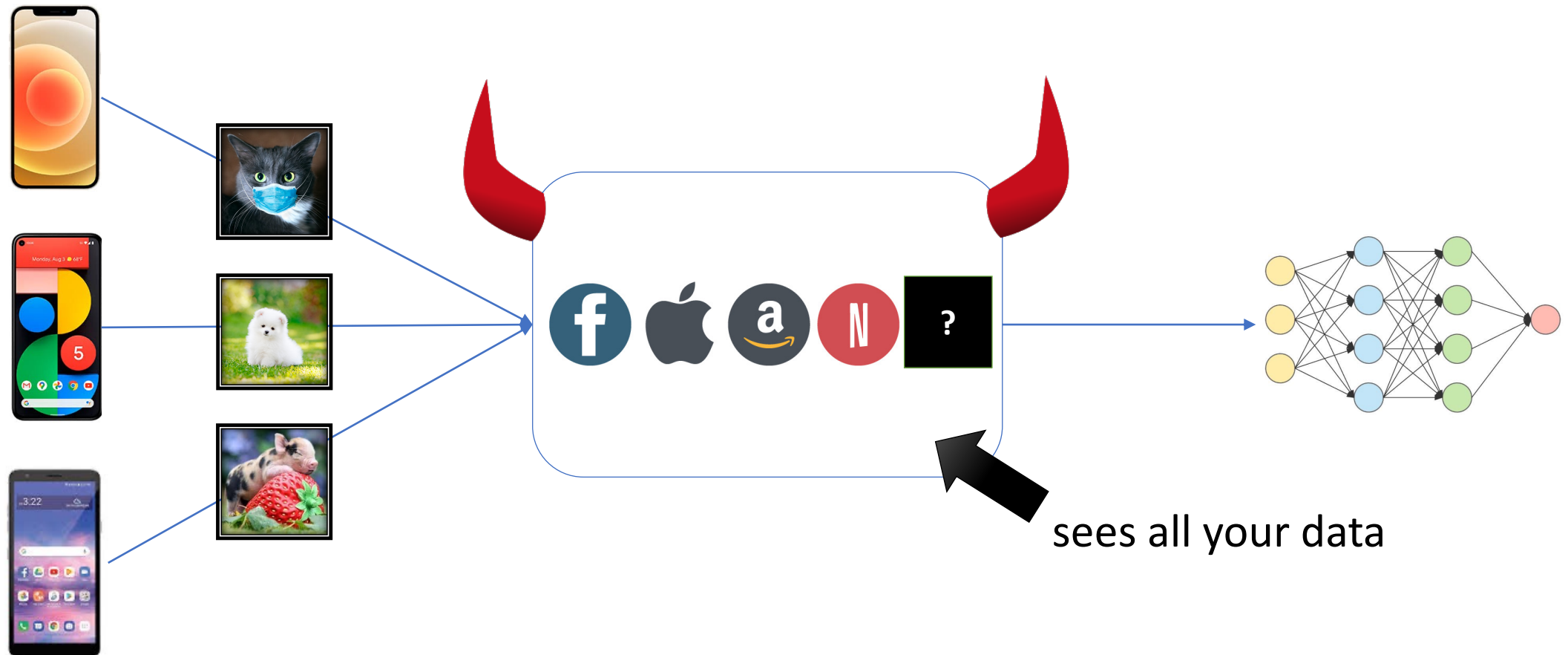
Goal: train a ML model with “privacy”

- what does this mean?
- how can we achieve this?
- what's next?

Goal: train a ML model with “privacy”

➤ what does this mean?

data secrecy

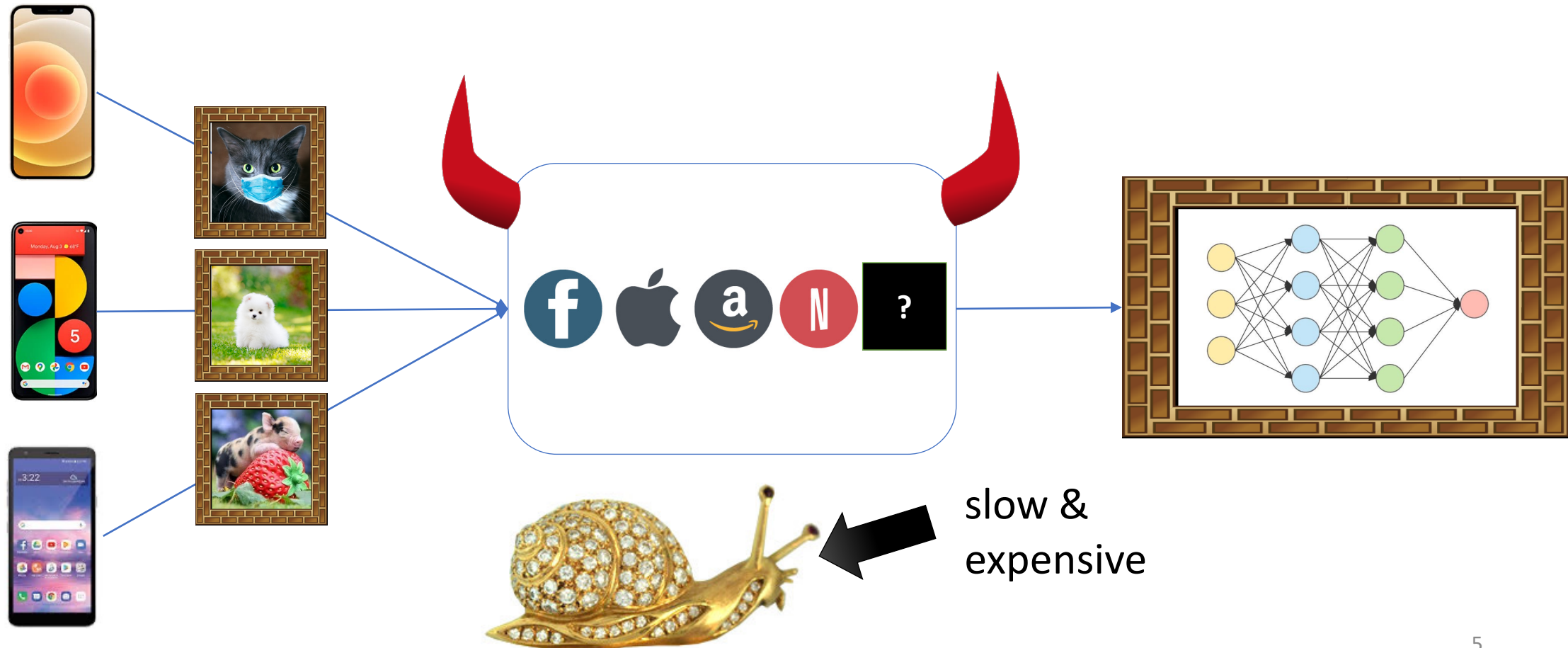


Goal: train a ML model with “privacy”

- what does this mean?
- how can we achieve this?

data secrecy

federated ML, MPC, FHE, ...



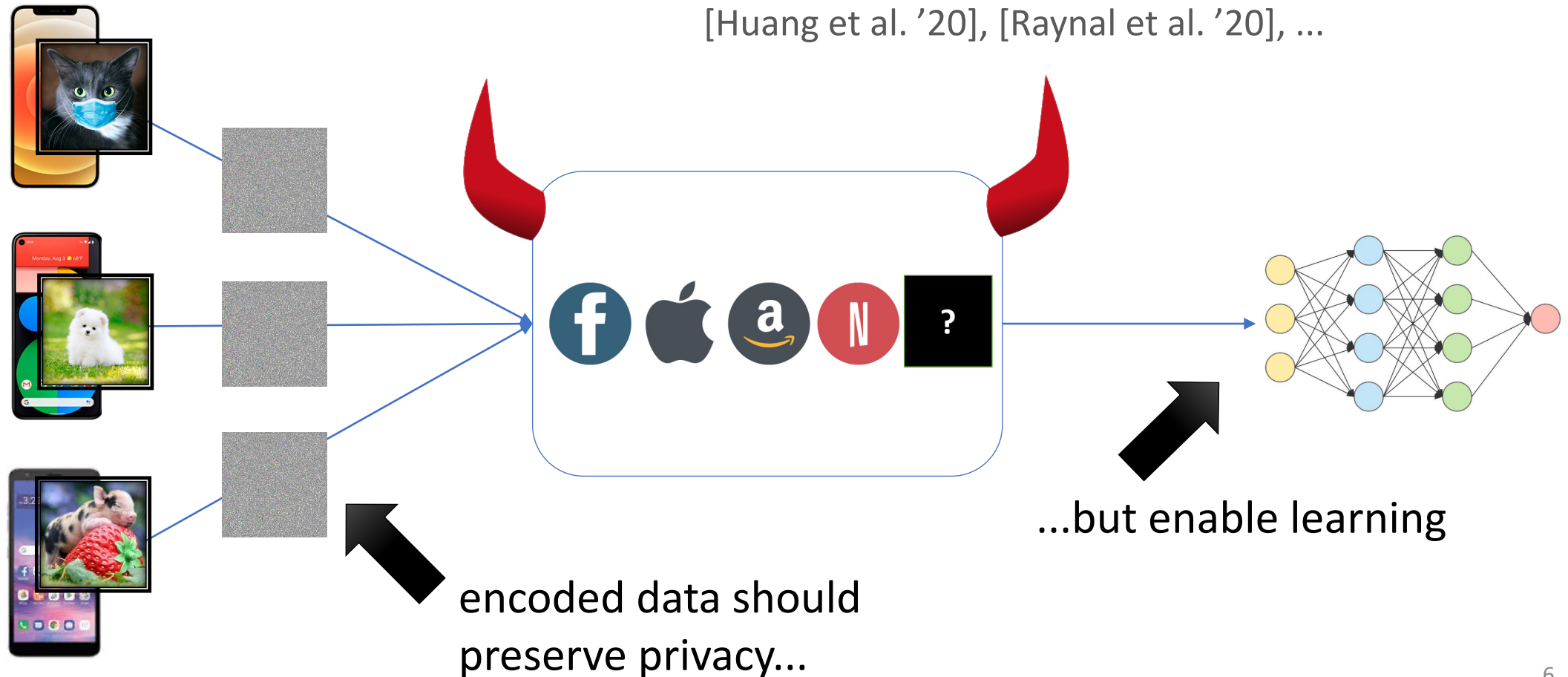
Goal: train a ML model with “privacy”

- what does this mean?
- how (else) can we achieve this?

data secrecy




learning on “encoded” data

[Huang et al. '20], [Raynal et al. '20], ...



Goal: private learning on “encoded” data

Example: InstaHide [Huang et al. ICML '20]





Encode(  ):

private data

public data

map pixel space
[0, 255] to [-1, 1]

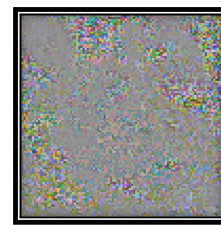
no formal
privacy
guarantee...

1) **Mixing:** λ_1  + λ_2  + λ_3  =  $\in [-1, 1]^d$

2) **“high-order bit flip”:** $\sigma \leftarrow^R \{-1, 1\}^d$

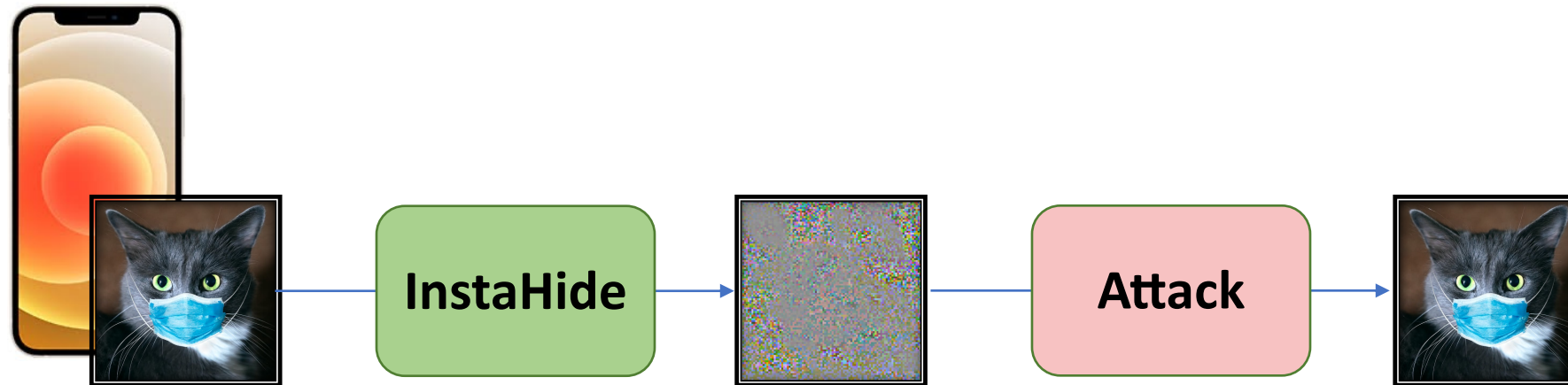


$\odot \sigma =$



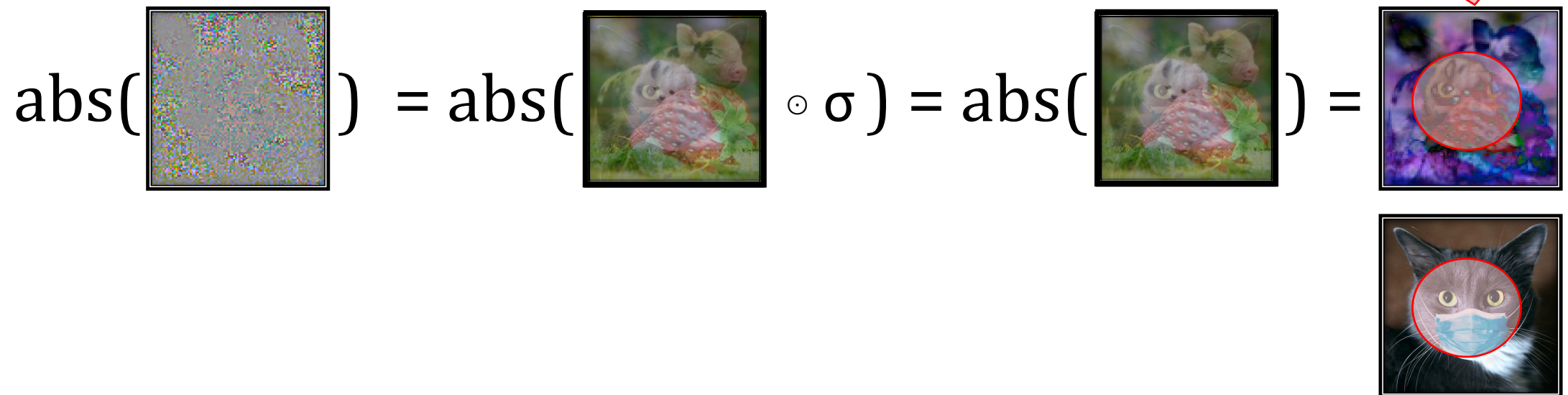
learning still works!

We show a **reconstruction attack** on InstaHide.



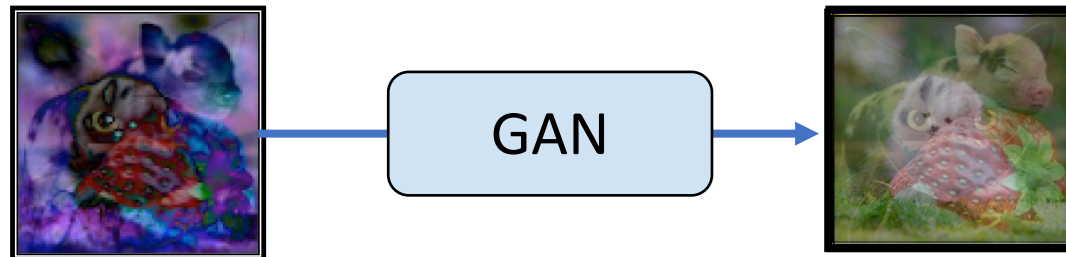
We show a **reconstruction attack** on InstaHide.

1. Undo the random bit flip



We show a ^{full} **reconstruction attack** on InstaHide.

1. Undo the random bit flip
2. Learn to “recolor” mixed images



We show a ^{full} **reconstruction attack** on InstaHide.

1. Undo the random bit flip
2. Learn to “recolor” mixed images
3. Undo the mixing by finding the most similar public images

$$\text{SIM}(\text{img}_1, \text{img}_2) = 0.42$$

$$\text{SIM}(\text{img}_1, \text{img}_3) = 0.07$$

...

$$\text{SIM}(\text{img}_1, \text{img}_4) = 0.79$$



- More attacks
- **Impossibility results**

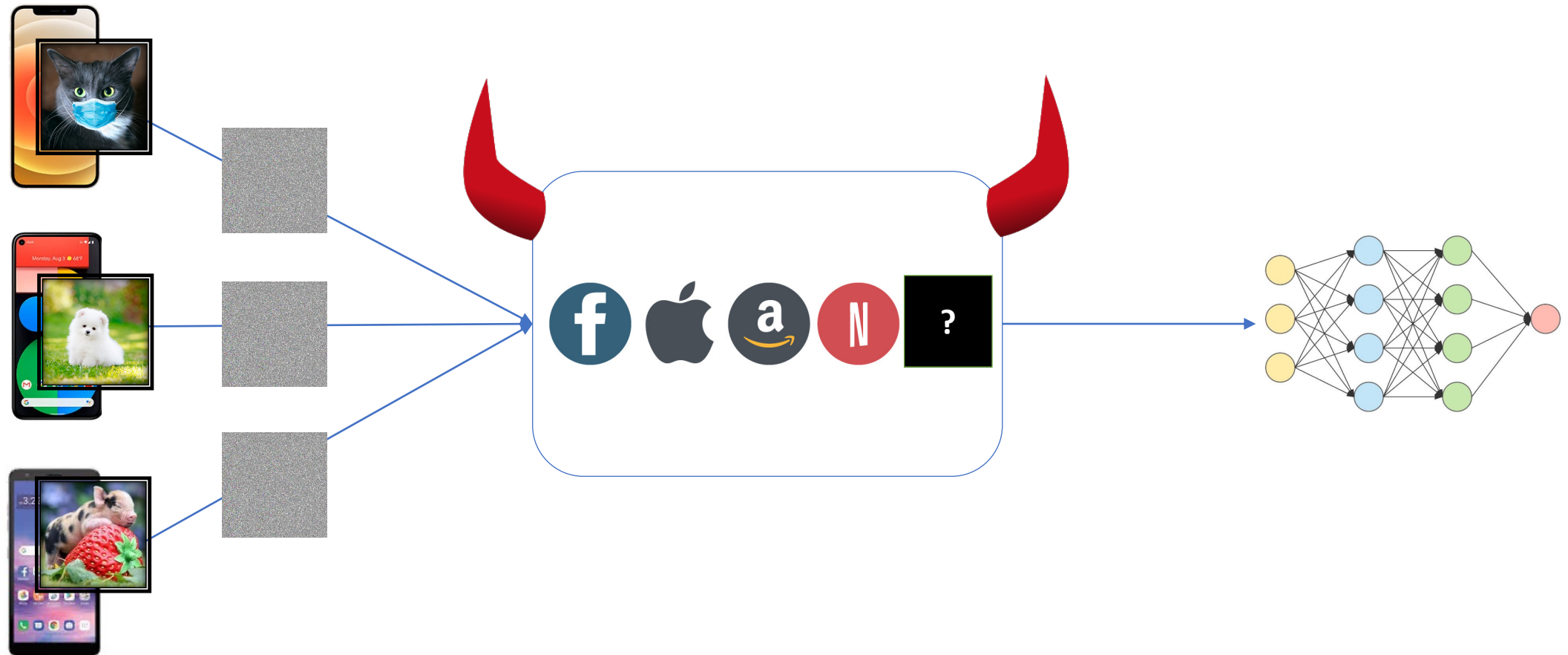
Goal: train a ML model with “privacy”

➤ what does this mean?

data secrecy

➤ how (else) can we achieve this?

~~learning on “encoded” data~~



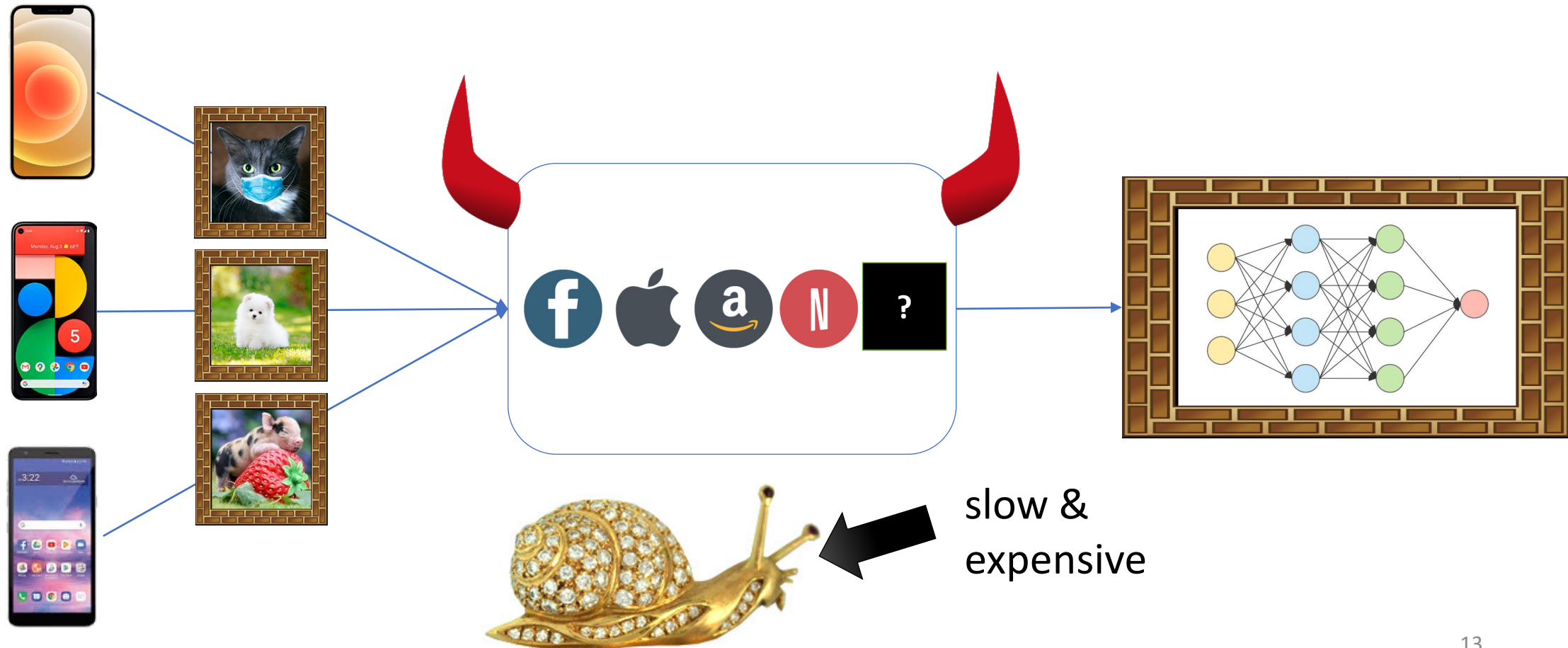
Goal: train a ML model with “privacy”

➤ what does this mean?

data secrecy

➤ how can we achieve this?

federated ML, MPC, FHE, ...



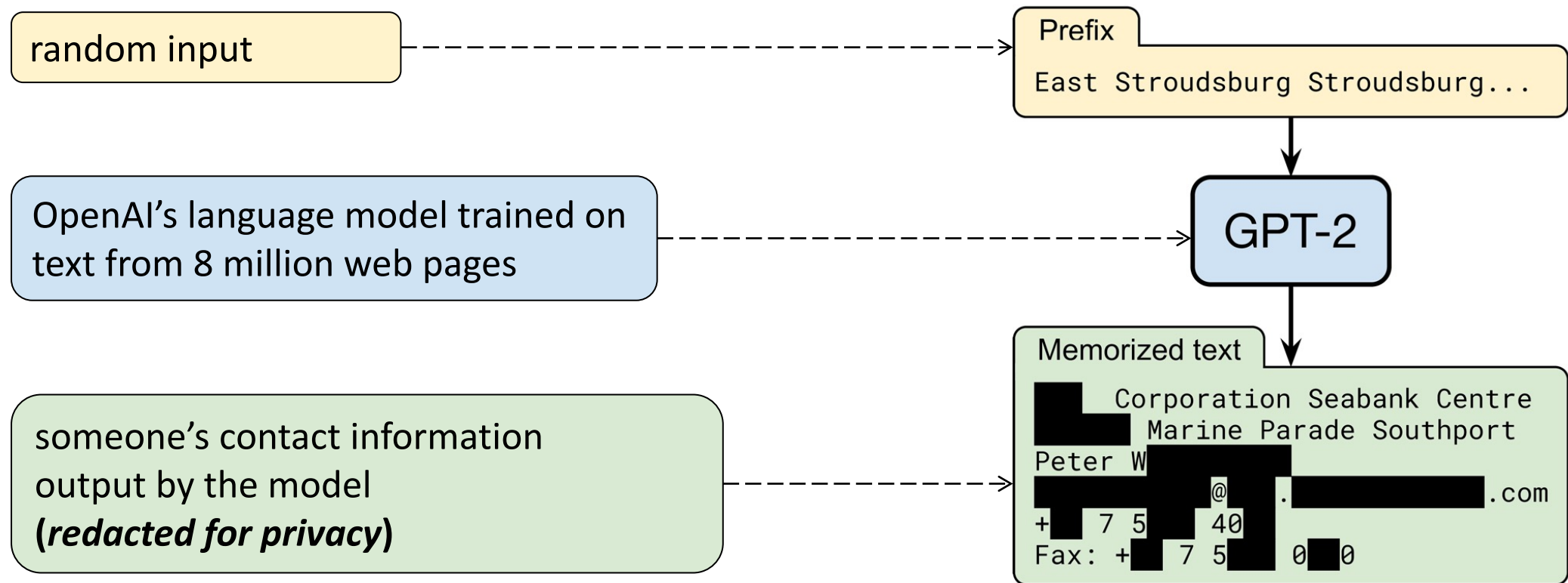
Is data secrecy *sufficient*?

No! The ideal functionality itself can be non-private



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

Models **memorize** their training data.

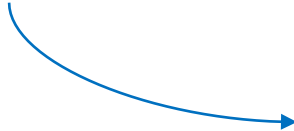


Larger models are *less* private.

URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████lp8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

Larger models are *less* private.

Reddit URLs found in
a pastebin file in the
GPT-2 training set



URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████lp8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

Larger models are *less* private.

Reddit URLs found in a pastebin file in the GPT-2 training set



URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████lp8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

Some URLs appear many times in this pastebin file



Larger models are *less* private.

Reddit URLs found in a pastebin file in the GPT-2 training set

URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████lp8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

Different GPT-2 models:
XL: 1558M params
M: 334M params
S: 124M params

URL is memorized fully or partially

Some URLs appear many times in this pastebin file

Larger models are *less* private.

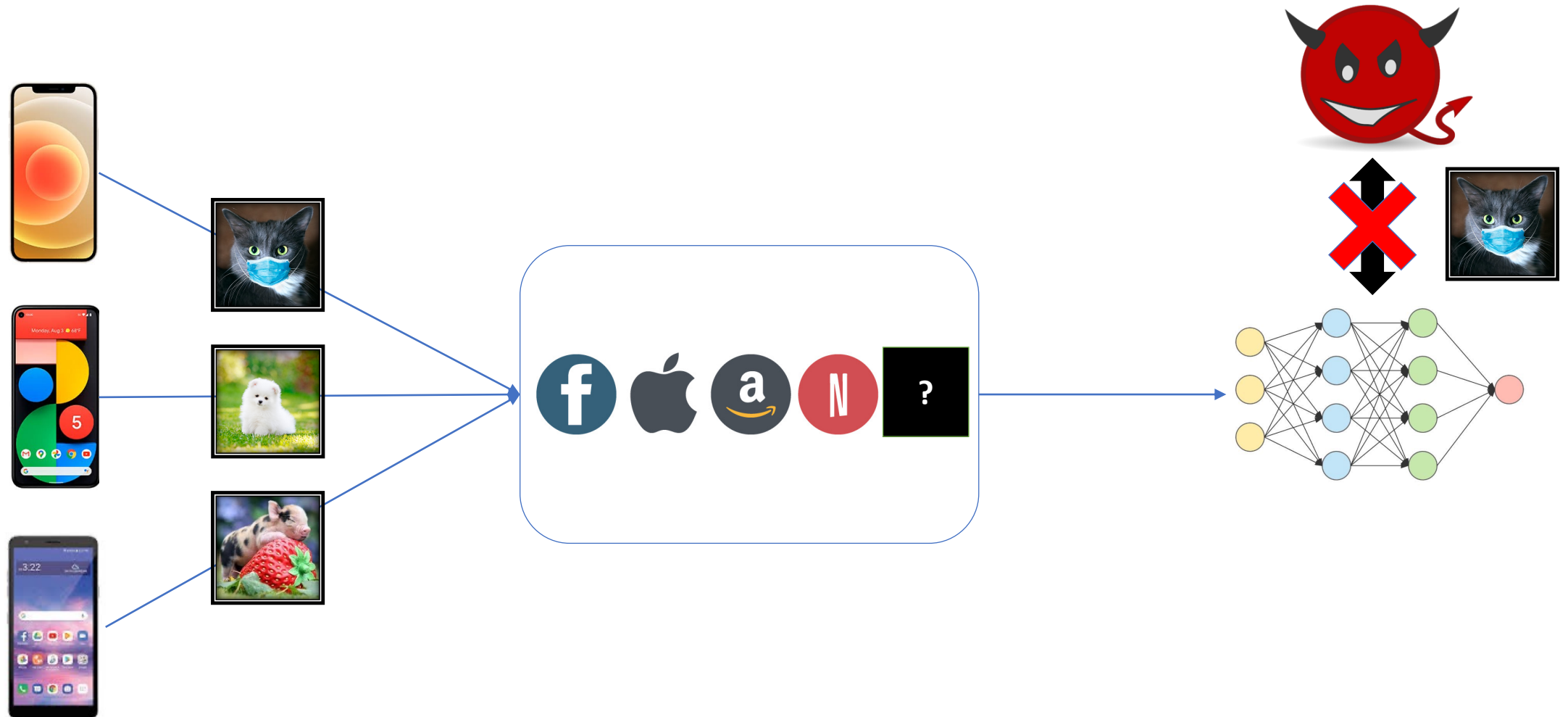
URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/████51y/milo_evacua...	1	359	✓	✓	1/2
/r/████zin/hi_my_name...	1	113	✓	✓	
/r/████7ne/for_all_yo...	1	76	✓	1/2	
/r/████5mj/fake_news_...	1	72	✓		
/r/████5wn/reddit_admi...	1	64	✓	✓	
/r/████lp8/26_evening...	1	56	✓	✓	
/r/████jla/so_pizzagat...	1	51	✓	1/2	
/r/████ubf/late_night...	1	51	✓	1/2	
/r/████eta/make_christ...	1	35	✓	1/2	
/r/████6ev/its_officia...	1	33	✓		
/r/████3c7/scott_adams...	1	17			
/r/████k2o/because_his...	1	17			
/r/████tu3/armynavy_ga...	1	8			

the largest GPT-2 model memorized an entire URL that appeared **only 33 times** in a single document

Goal: train a ML model with “privacy”

➤ what does this mean?

no training data leakage



Preventing data leakage with decade-old ML

- *provably* prevent leakage of training data.
using *differential privacy*
- *better accuracy* than with deep learning methods.
using *domain-specific feature engineering*

Goal: train a ML model with “privacy”

➤ what does this mean?

no training data leakage

➤ how can we achieve this?

differential privacy

intuition: *randomized* training algorithm is not influenced (too much) by any individual data point

for any two datasets that differ in a single element

$$\frac{\Pr[A_{\text{train}}(\text{cat}, \text{dog}, \text{pig}) = \text{NN}]}{\Pr[A_{\text{train}}(\text{cat}, \text{dog}, \text{pig}) = \text{NN}]} \leq e^\epsilon$$

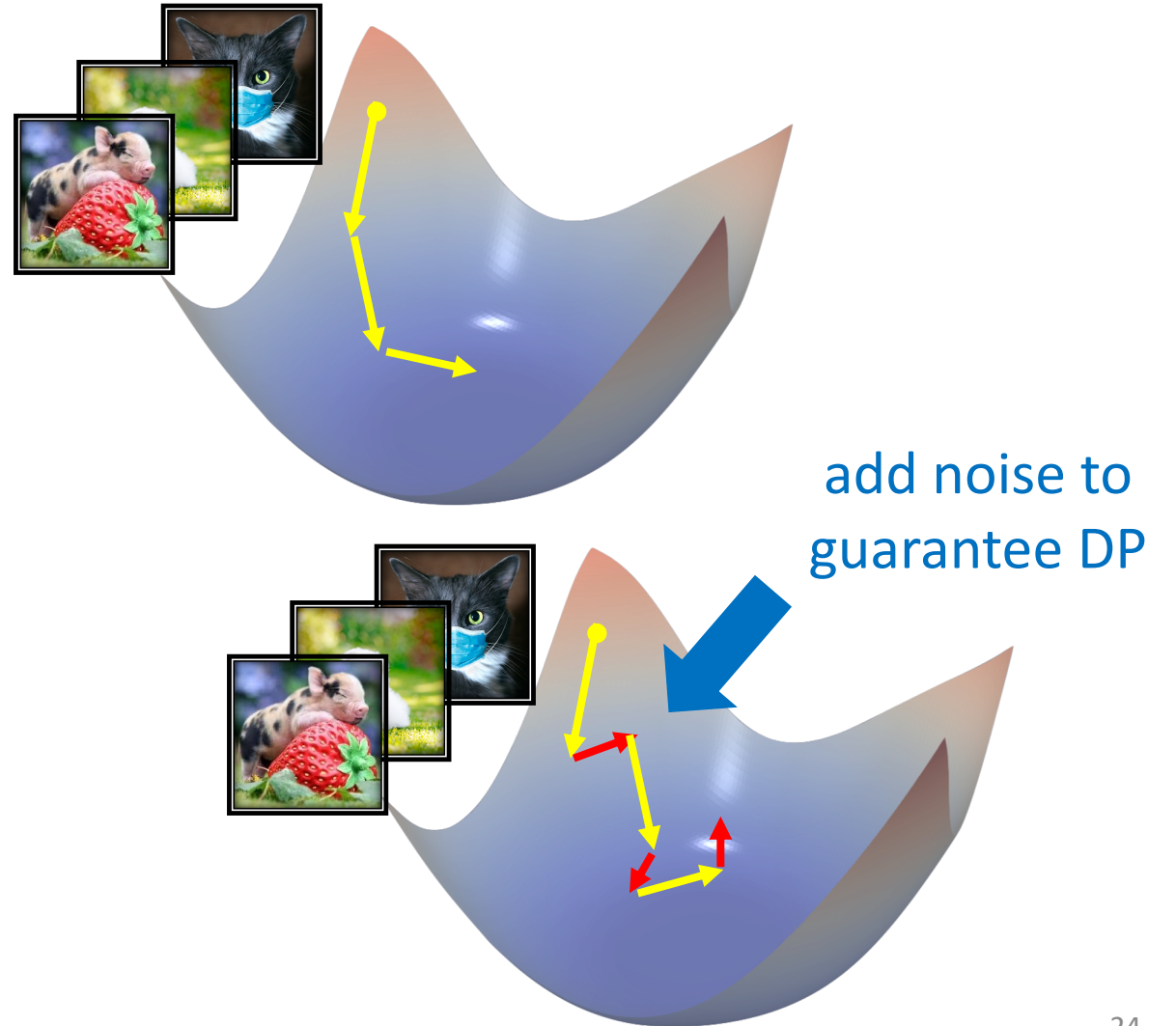
The equation shows the ratio of probabilities for two datasets that differ by only one element (the cat image). The numerator uses a blue-bordered image of a tabby cat, and the denominator uses a red-bordered image of a black and white cat. Both lead to the same neural network output (NN). A blue arrow points from the text above to the cat images.

How? Private Gradient Descent

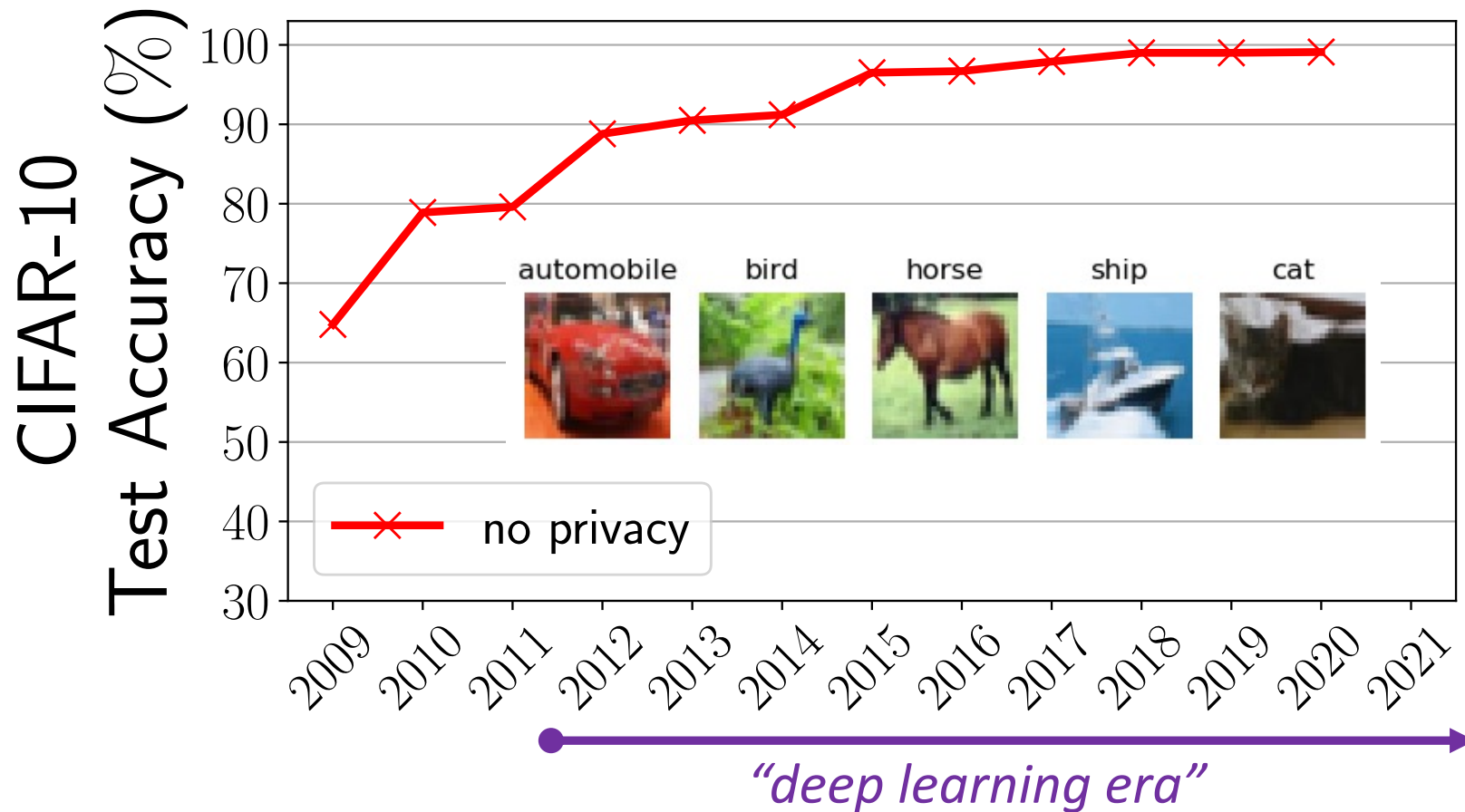
gradient descent
(SGD)

private gradient descent
(DP-SGD)

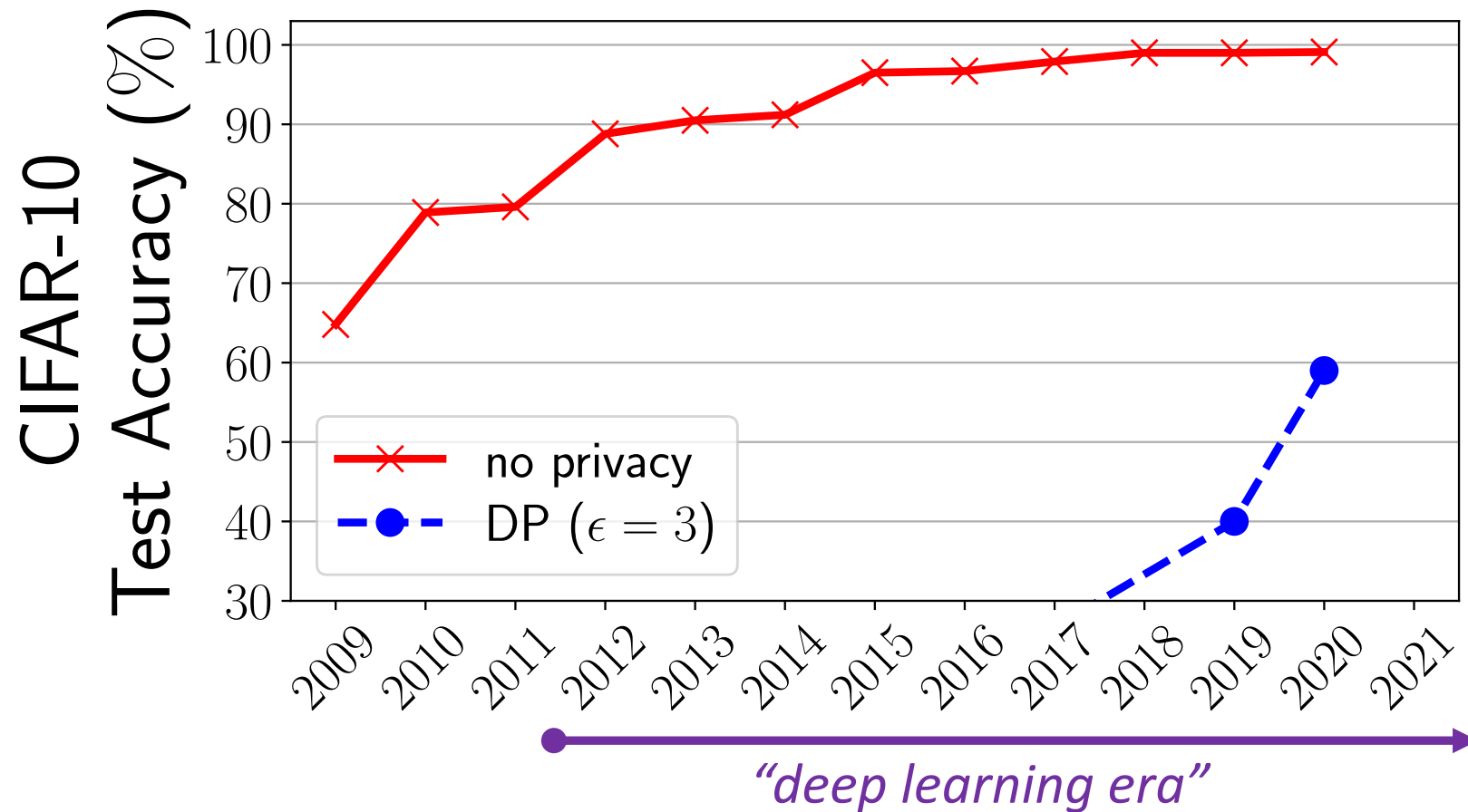
Chaudhuri et al., '11; Bassily et al. '14;
Shokri & Shmatikov '15; Abadi et al. '16



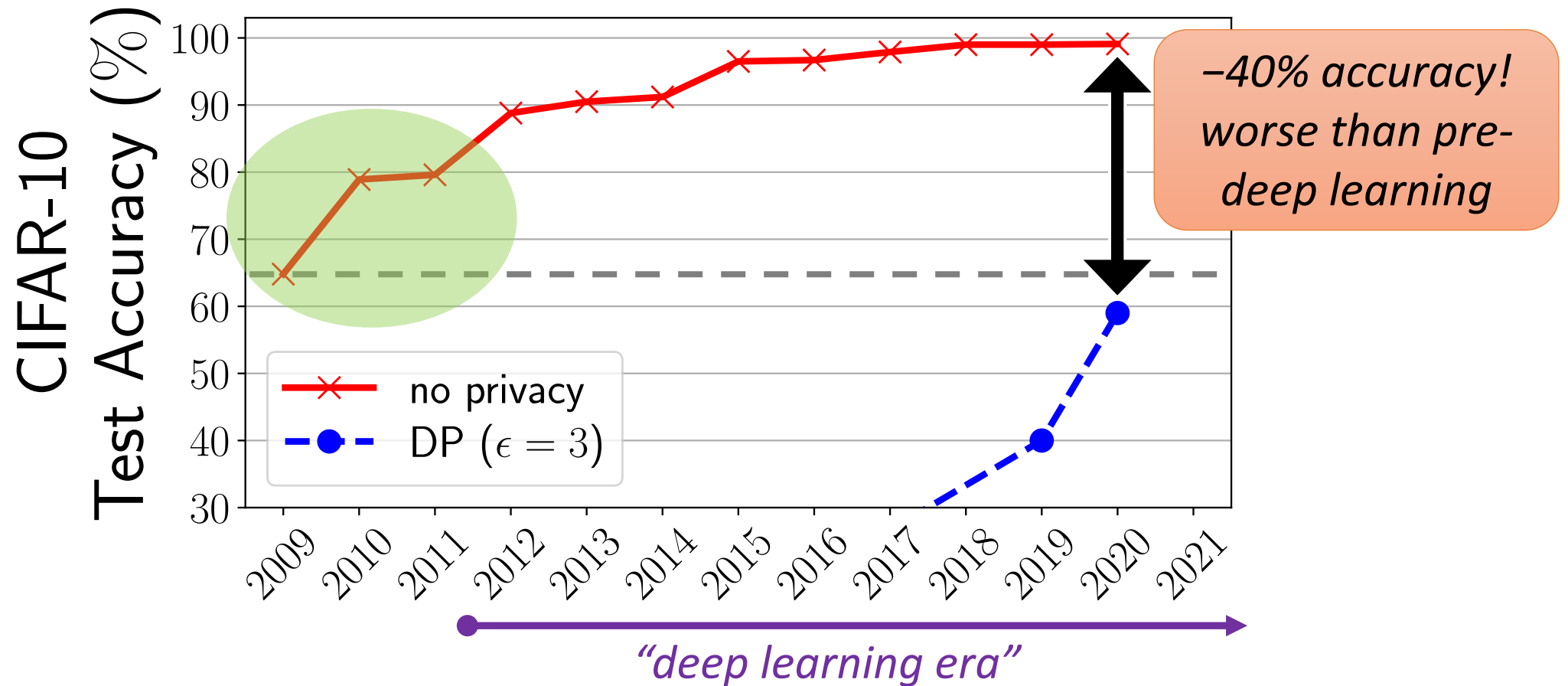
Non-private deep learning can achieve near-perfect accuracy.



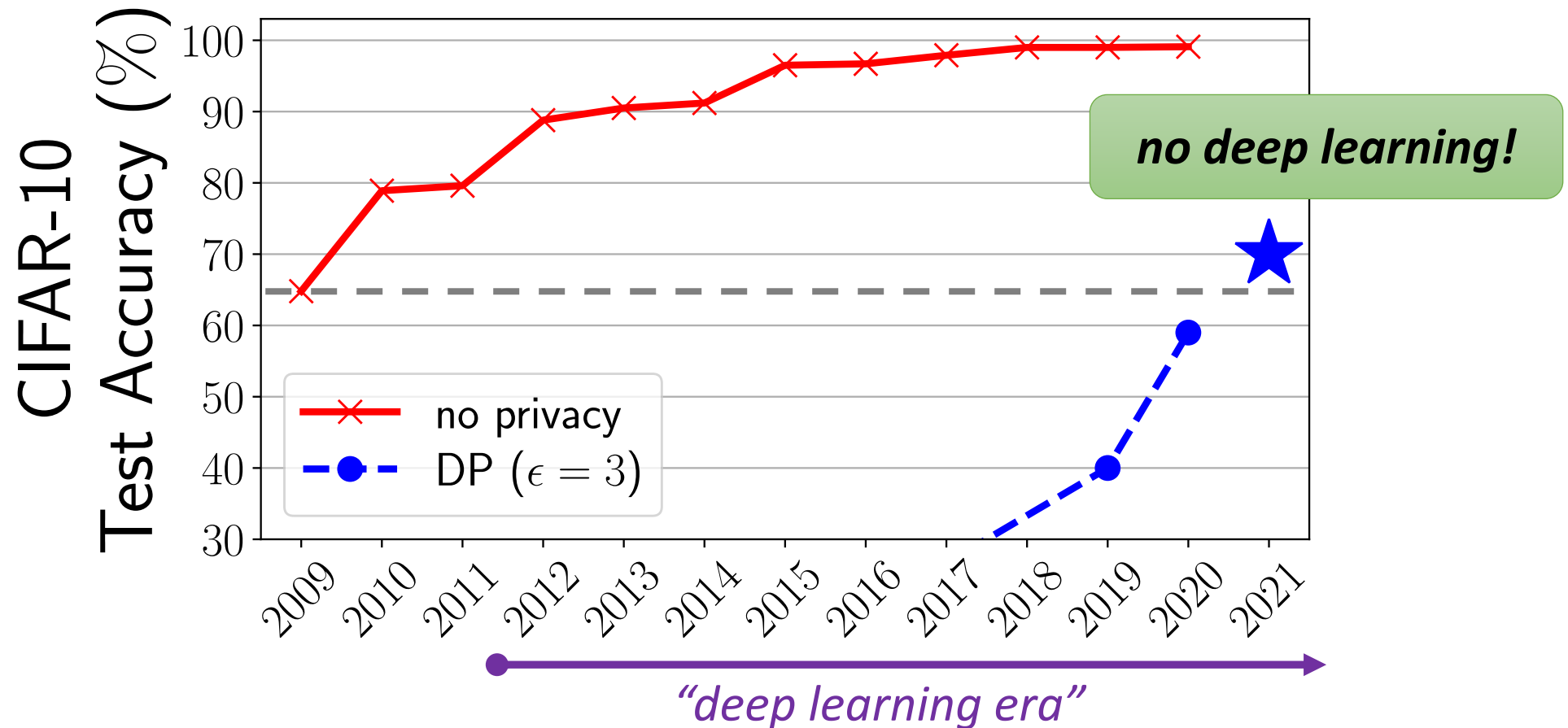
Differentially private deep learning lowers accuracy significantly.



Differentially private deep learning lowers accuracy significantly.

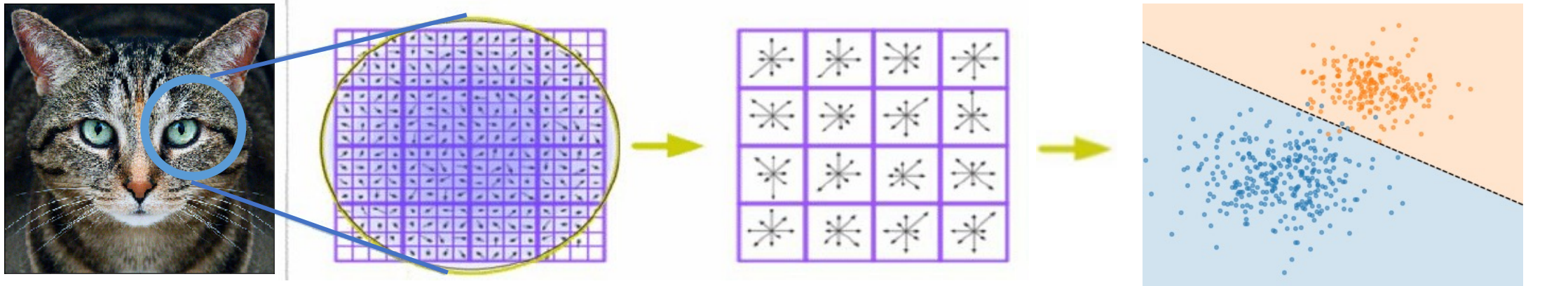


Differential privacy *without deep learning* improves accuracy.



Privacy-free features from “old-school” image recognition.

SIFT [Lowe '99, '04], **HOG** [Dalal & Triggs '05], **SURF** [Bay et al. '06], **ORB** [Rublee et al. '11], ...
Scattering transforms [Bruna & Mallat '11], [Oyallon & Mallat '14], ...



“handcrafted features”
(no learning involved)

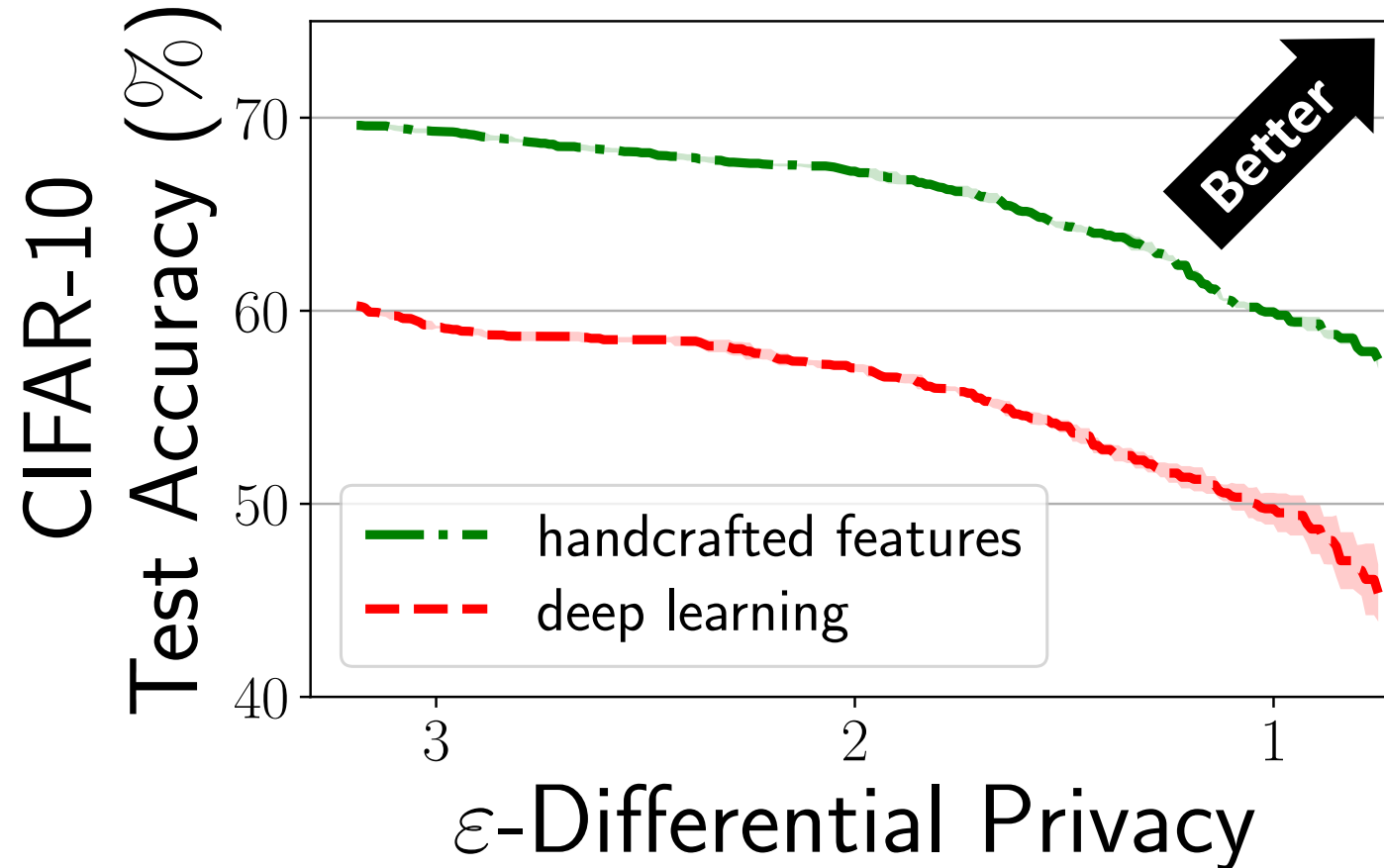
simple classifier
(e.g., logistic regression)

privacy free



captures some *prior* about the domain: e.g., invariance under rotation & scaling

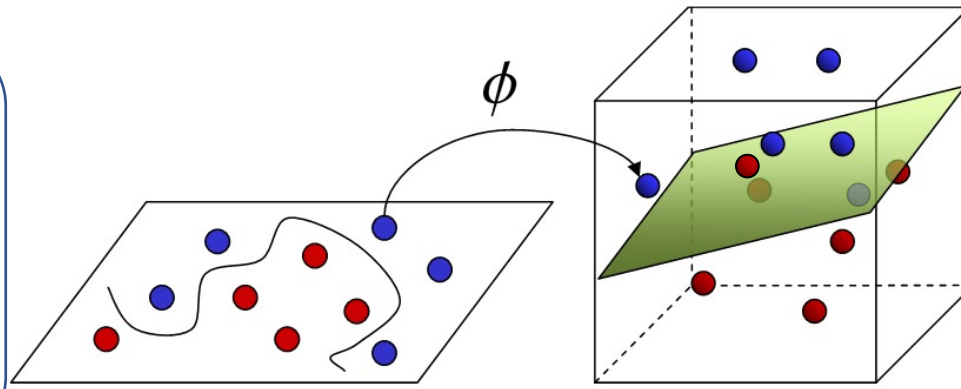
Handcrafted features lead to a better tradeoff between accuracy and privacy.



Handcrafted features lead to an *easier* learning task (for noisy gradient descent).

high accuracy classifier exists but learning takes many gradient steps

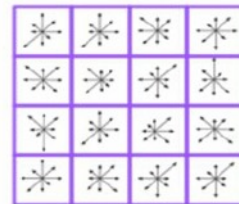
bad for privacy



Input Space



Feature Space

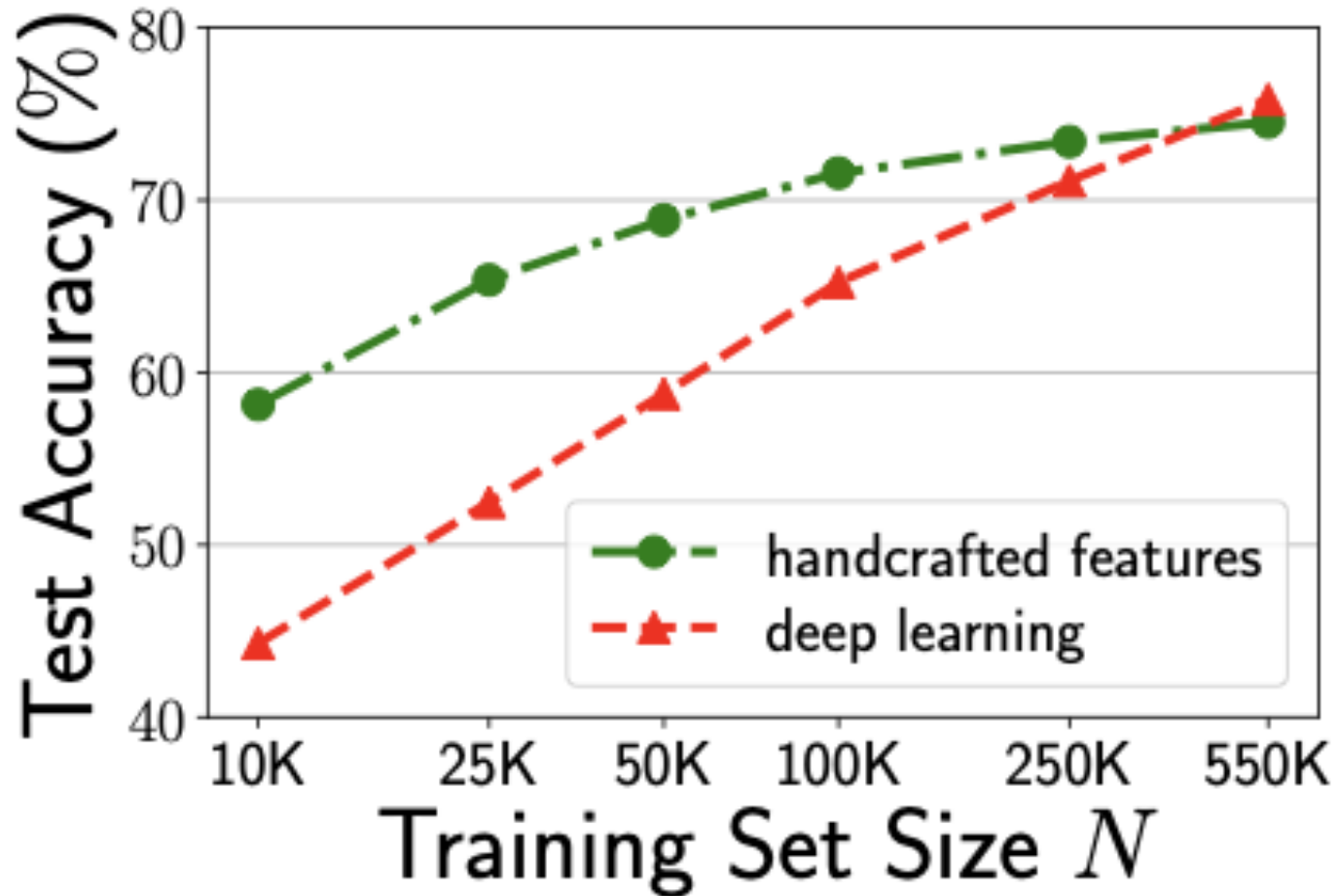


in feature space, maximal accuracy is reduced but learning progresses faster

good for privacy

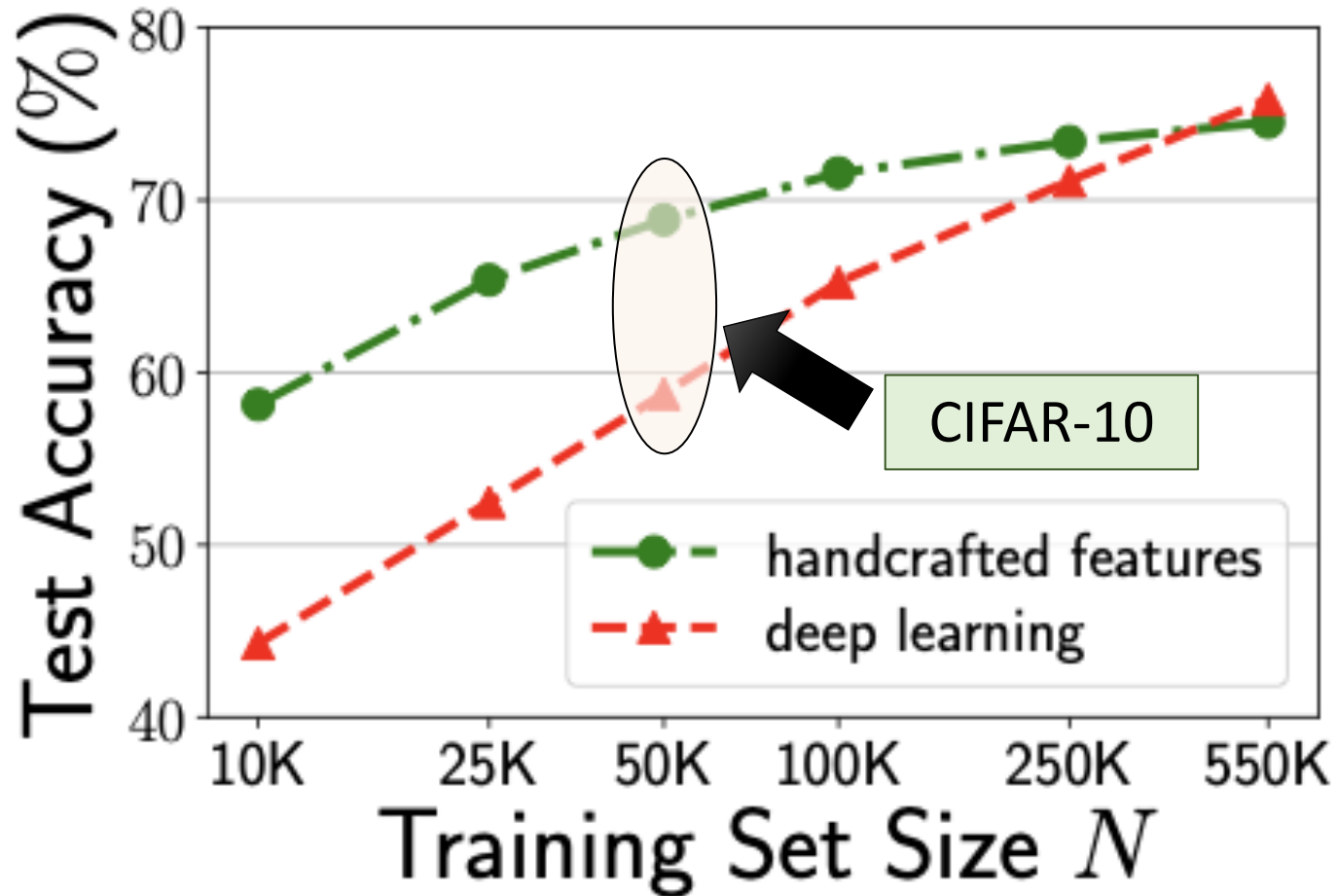
Surpassing handcrafted features with *more private data*.

(for $\epsilon = 3$)



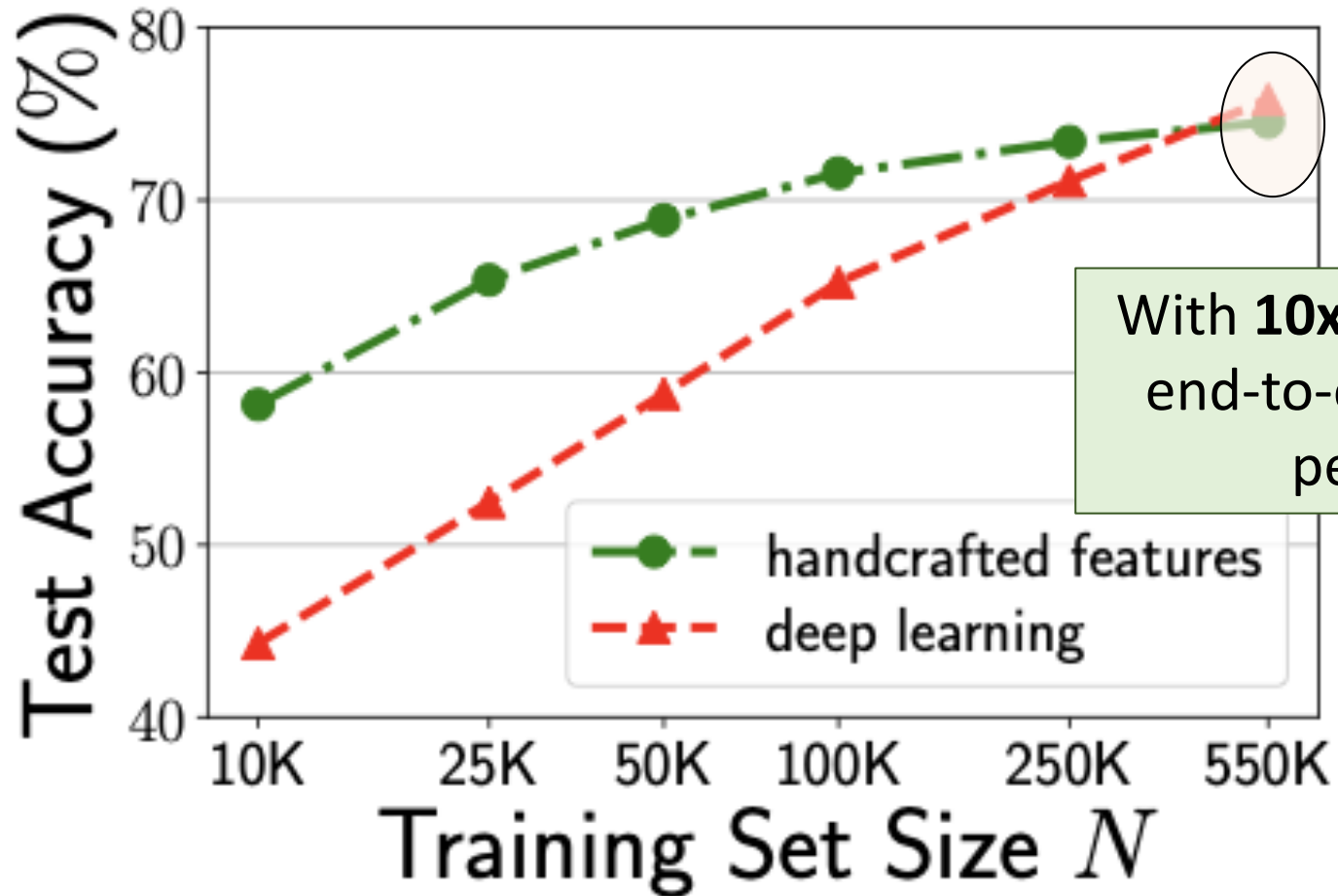
Surpassing handcrafted features with *more private data*.

(for $\epsilon = 3$)



Surpassing handcrafted features with *more private data*.

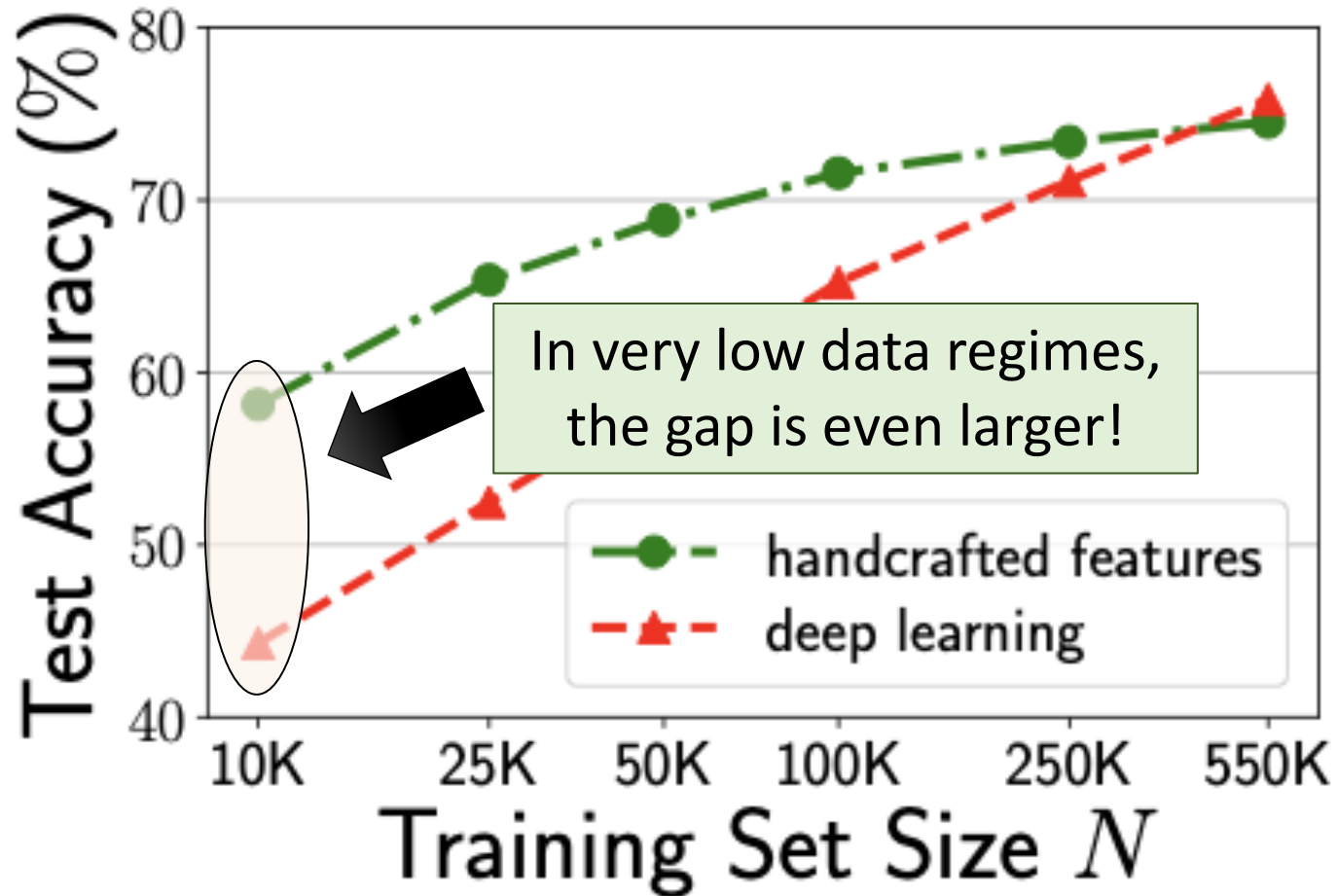
(for $\epsilon = 3$)



With **10x more private data** end-to-end deep learning performs best

Surpassing handcrafted features with *more private data*.

(for $\epsilon = 3$)



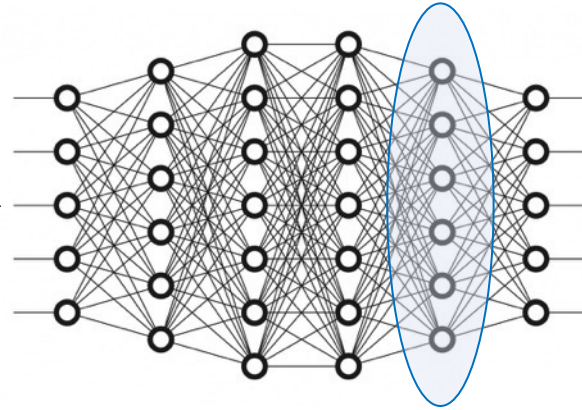
In very low data regimes, the gap is even larger!

—●— handcrafted features
-▲- deep learning

Surpassing handcrafted features with *more public data.*



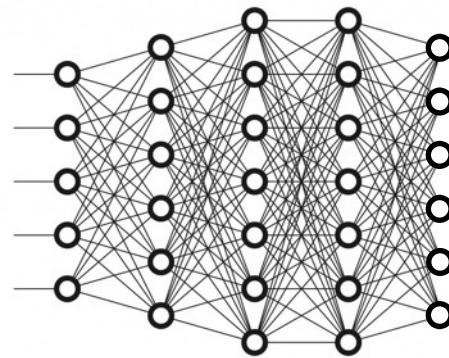
public data



*train a feature extractor
on public data...*



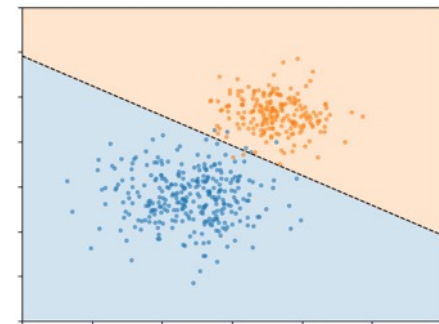
private data



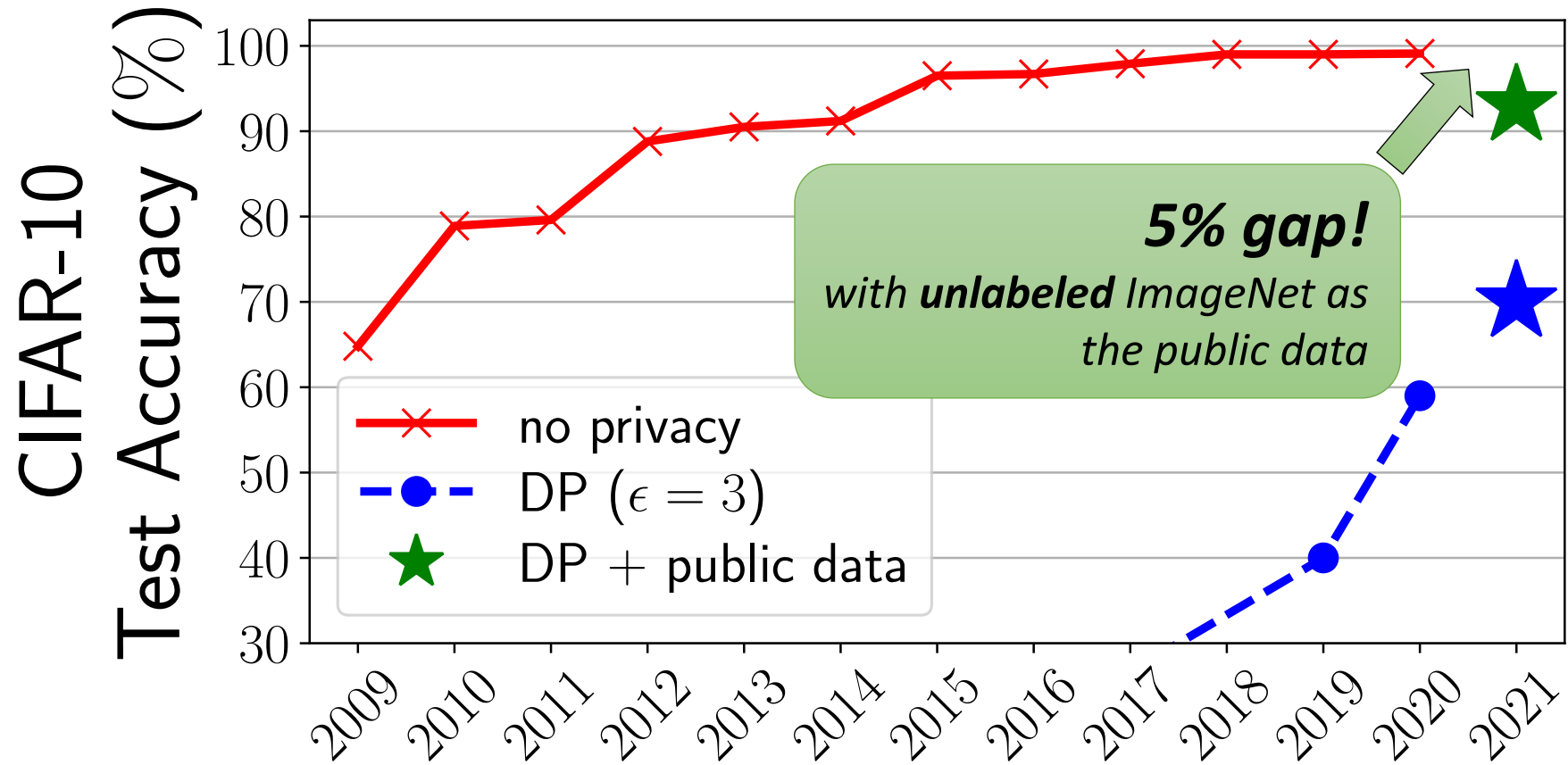
privacy free



*...transfer and fine-tune
on private data*



With access to a public dataset,
privacy comes almost for free!



Goal: train a ML model with “privacy”

➤ what does this mean?

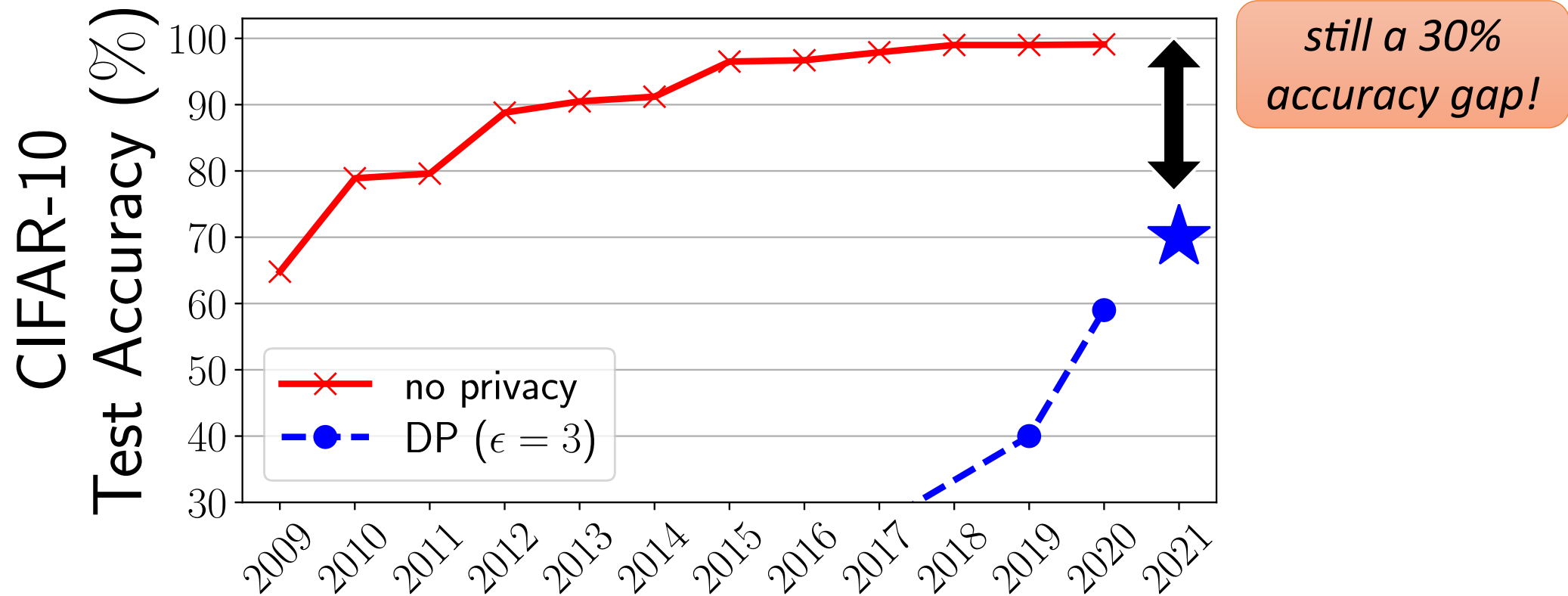
- data secrecy
- no training data leakage

➤ how can we achieve this?

- (strong) cryptography
- differential privacy (+ feature engineering!)

➤ what's next?

Can we bridge the accuracy gap in differentially private learning?

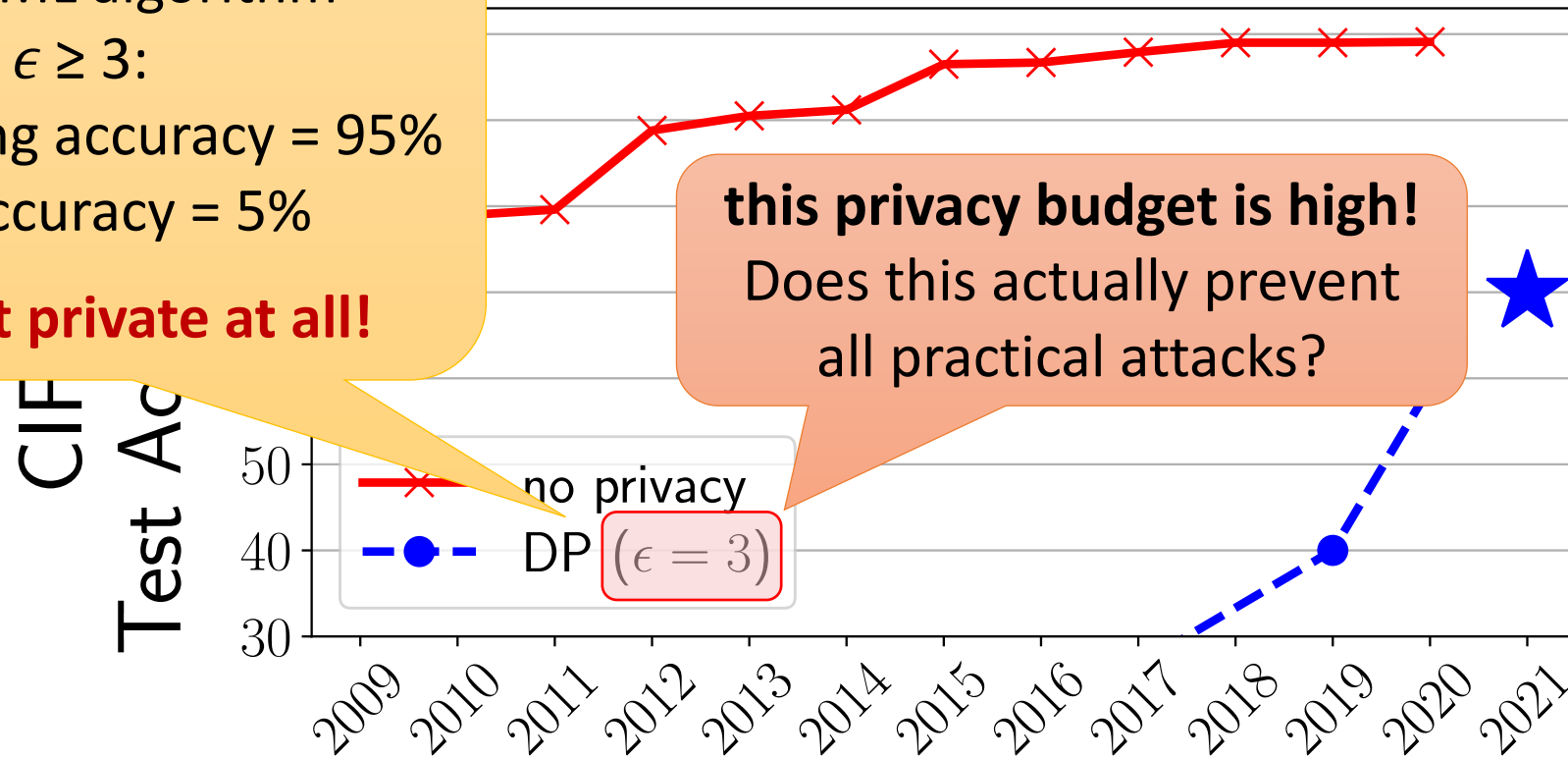


How much privacy do we really get from DP-SGD?

This generic ML algorithm also satisfies $\epsilon \geq 3$:

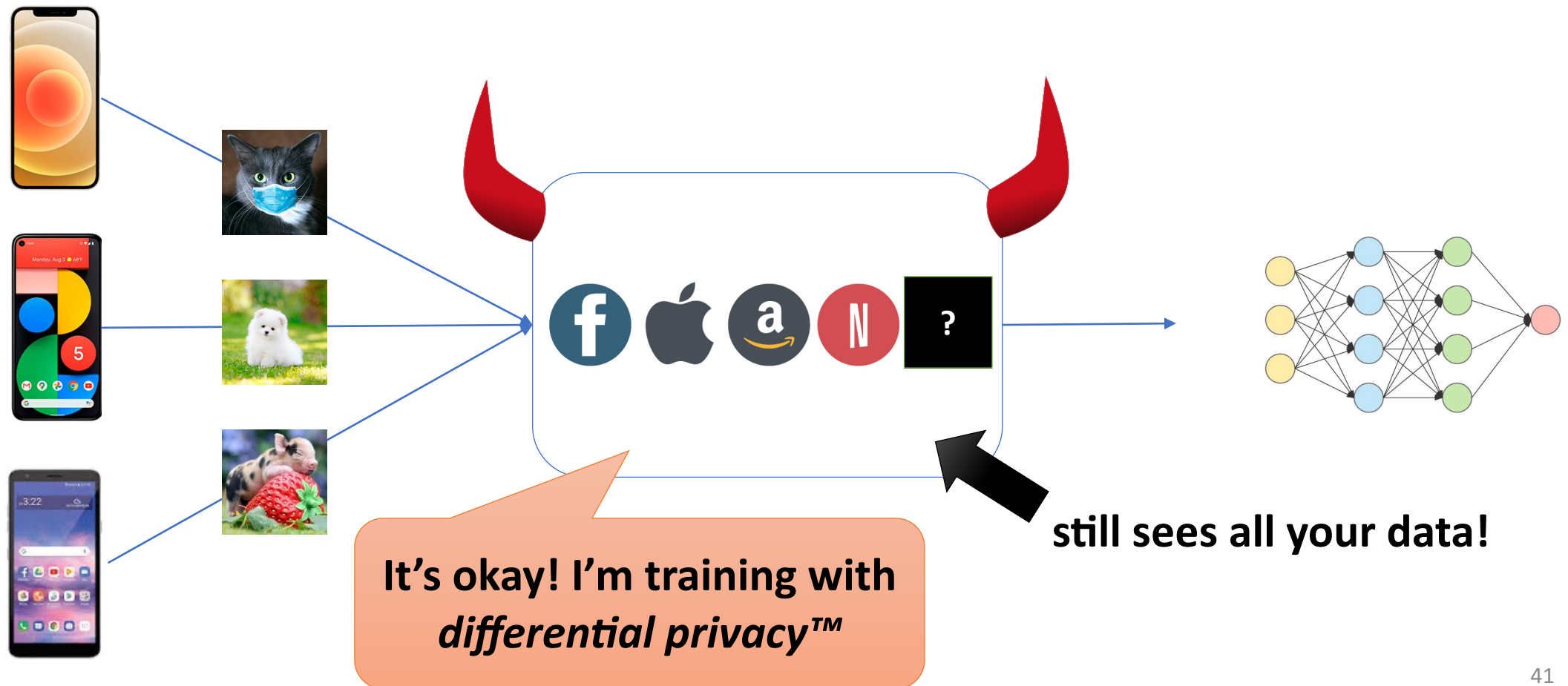
- training accuracy = 95%
- test accuracy = 5%

⇒ **This is not private at all!**



Is differential privacy **sufficient**?

No! We also need secure decentralized training



Conclusion

- **Machine learning is not private “by default”!**
 - Without (strong) cryptography, you must **trust someone** with your data
 - Trained models **leak** rare training data
- Solutions exist but we need to make them more efficient!
 - Secure decentralized learning has **high overhead**
 - Differential privacy needs **good features or a lot more data!**
 - ***Privacy guarantees must be rigorously defined!***

Thank you!