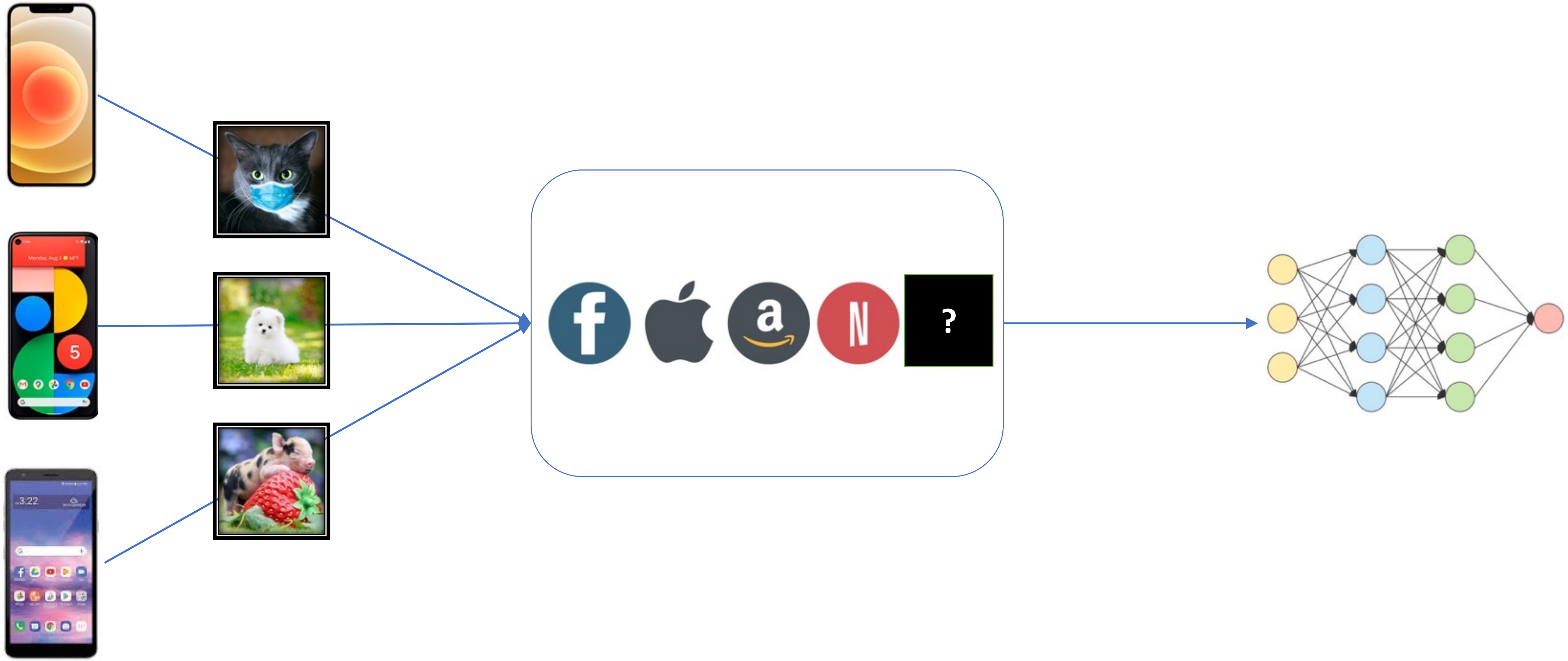
What is (and isn't) private learning?

Florian Tramèr
Stanford University

Goal: train a ML model with “privacy”



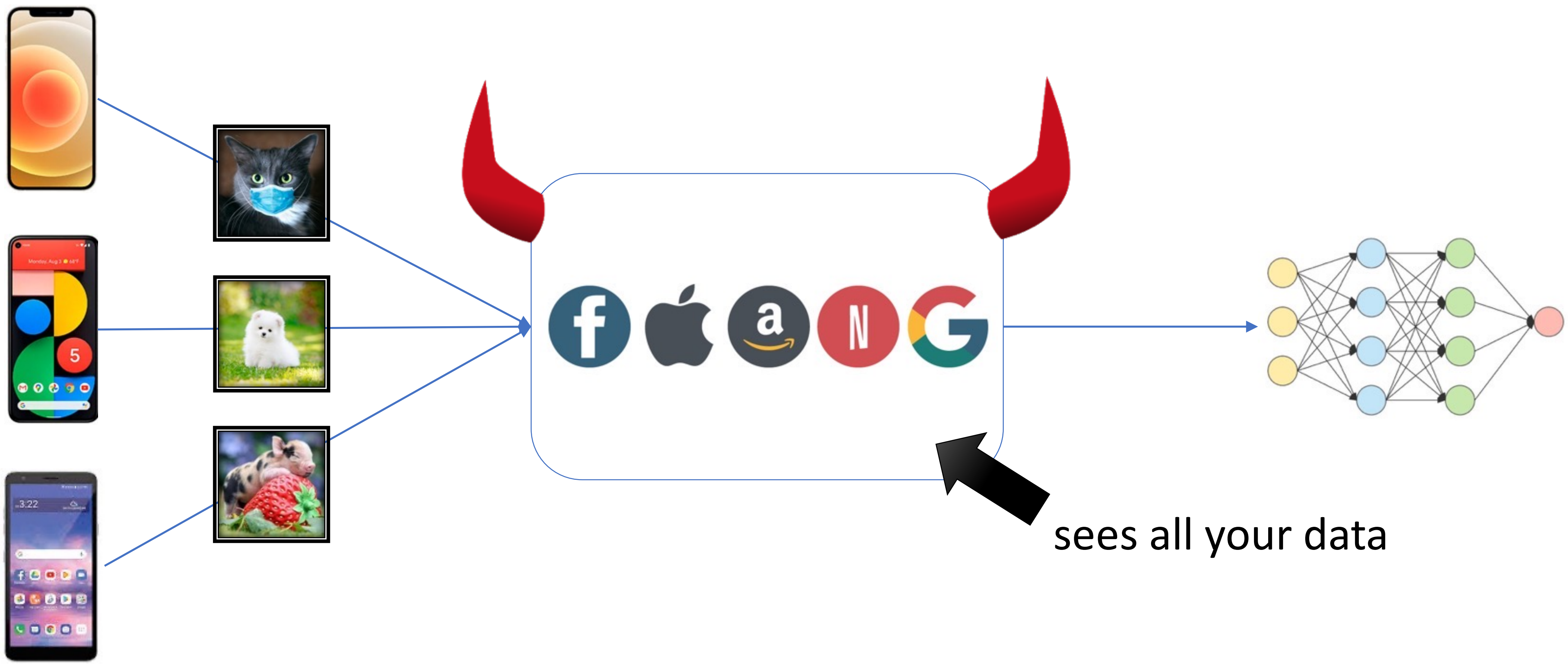
Goal: train a ML model with "privacy"

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

Goal: train a ML model with "privacy"

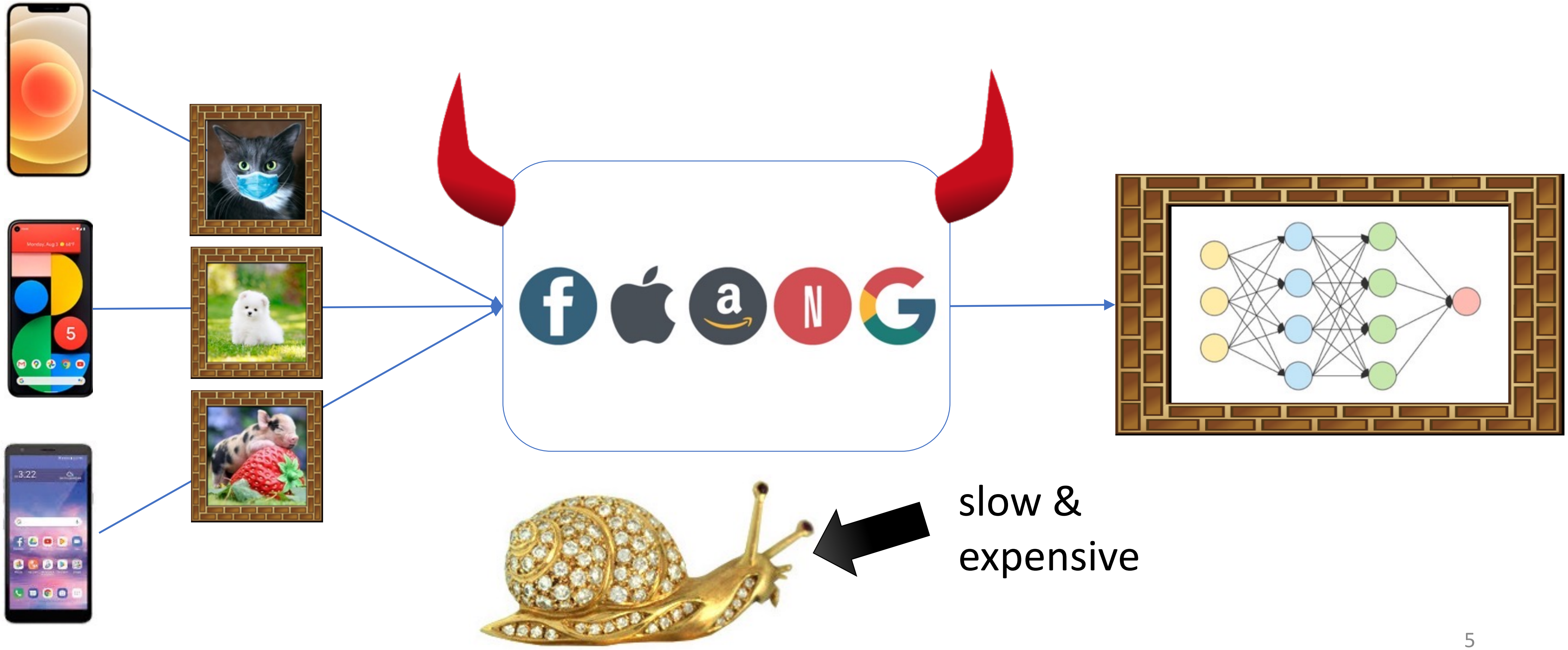
- what does this mean?

data secrecy



Goal: train a ML model with “privacy”

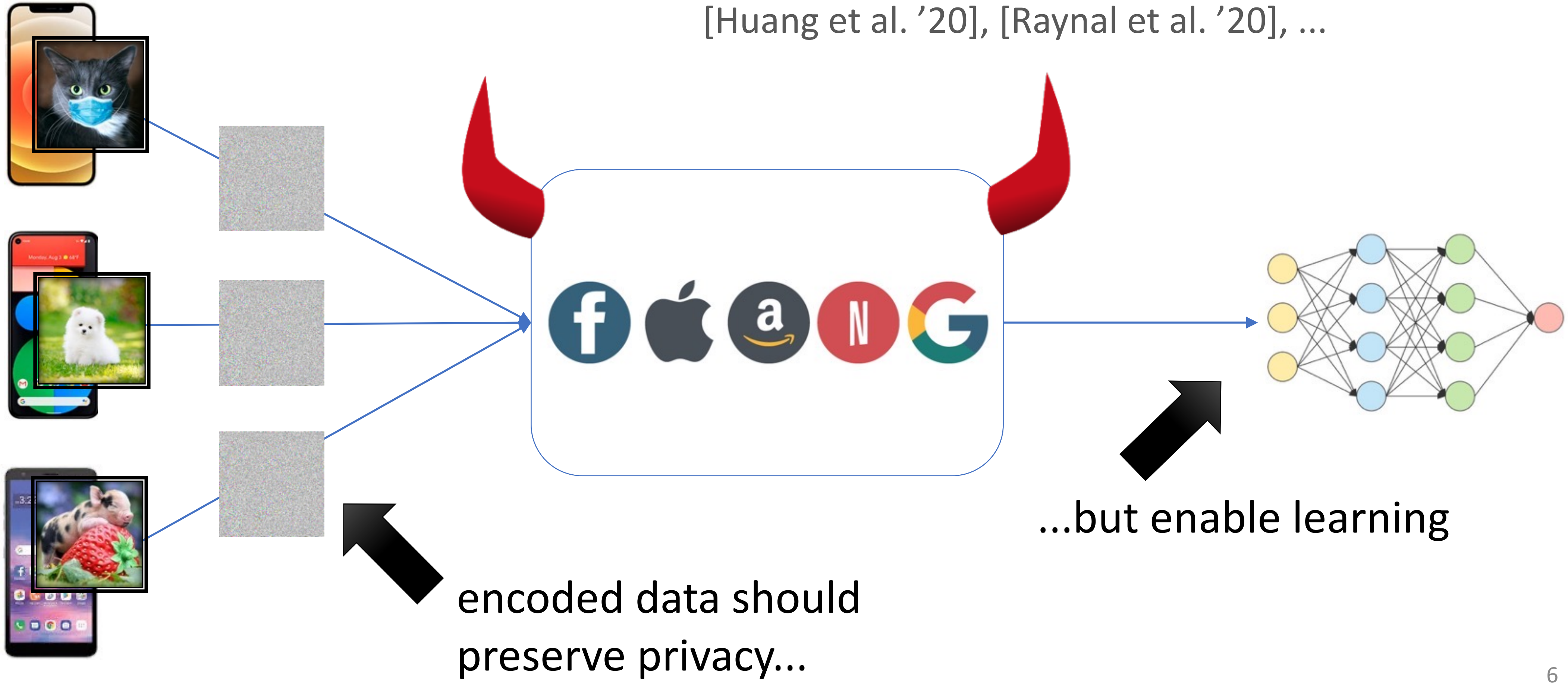
- what does this mean? — data secrecy
- how can we achieve this? — federated ML, MPC, FHE, ...



Goal: train a ML model with "privacy"

- what does this mean? — data secrecy
- how (else) can we achieve this? — 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]

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

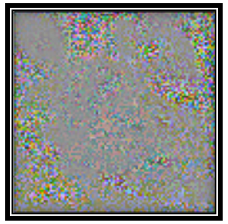
private data *public* data

Encode(  ):

no formal privacy guarantee…

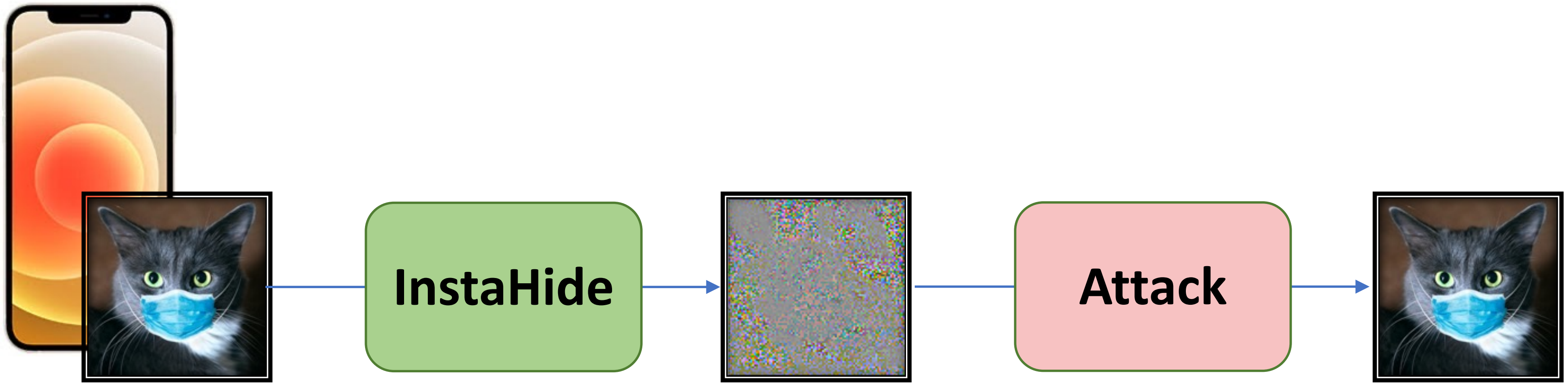
1) **Mixing**: λ_1  $+\lambda_2$  $+\lambda_3$  =  $\in [-1, 1]^d$

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

 $\odot\, \sigma$ = 

learning still works!

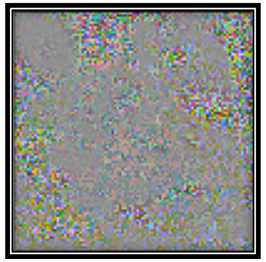
We show a reconstruction attack on InstaHide.



Is Private Learning Possible with Instance Encoding?, IEEE Security & Privacy 2021

We show a reconstruction attack on InstaHide.

1. Undo the random bit flip

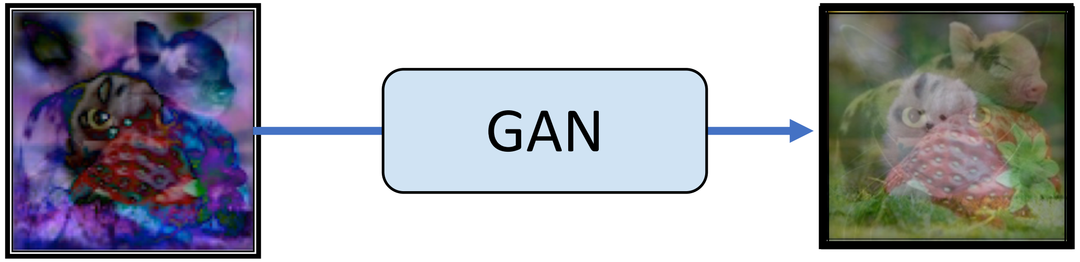
abs() = abs( $\odot\ \sigma$) = abs()

= 

this is clearly **not** private!!!

We show a full reconstruction attack on InstaHide.

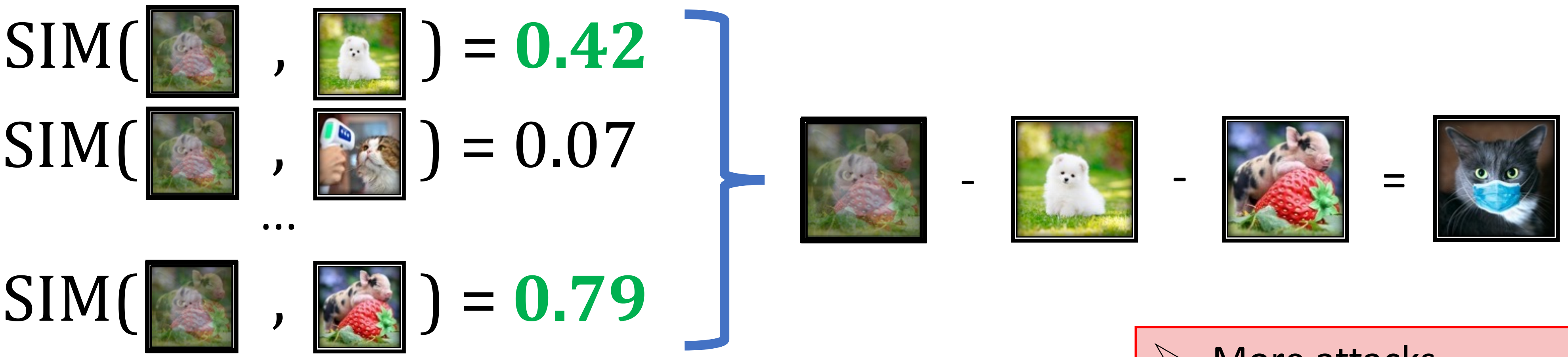
1. Undo the random bit flip
2. Learn to "recolor" mixed images



Is Private Learning Possible with Instance Encoding?, IEEE Security & Privacy 2021

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

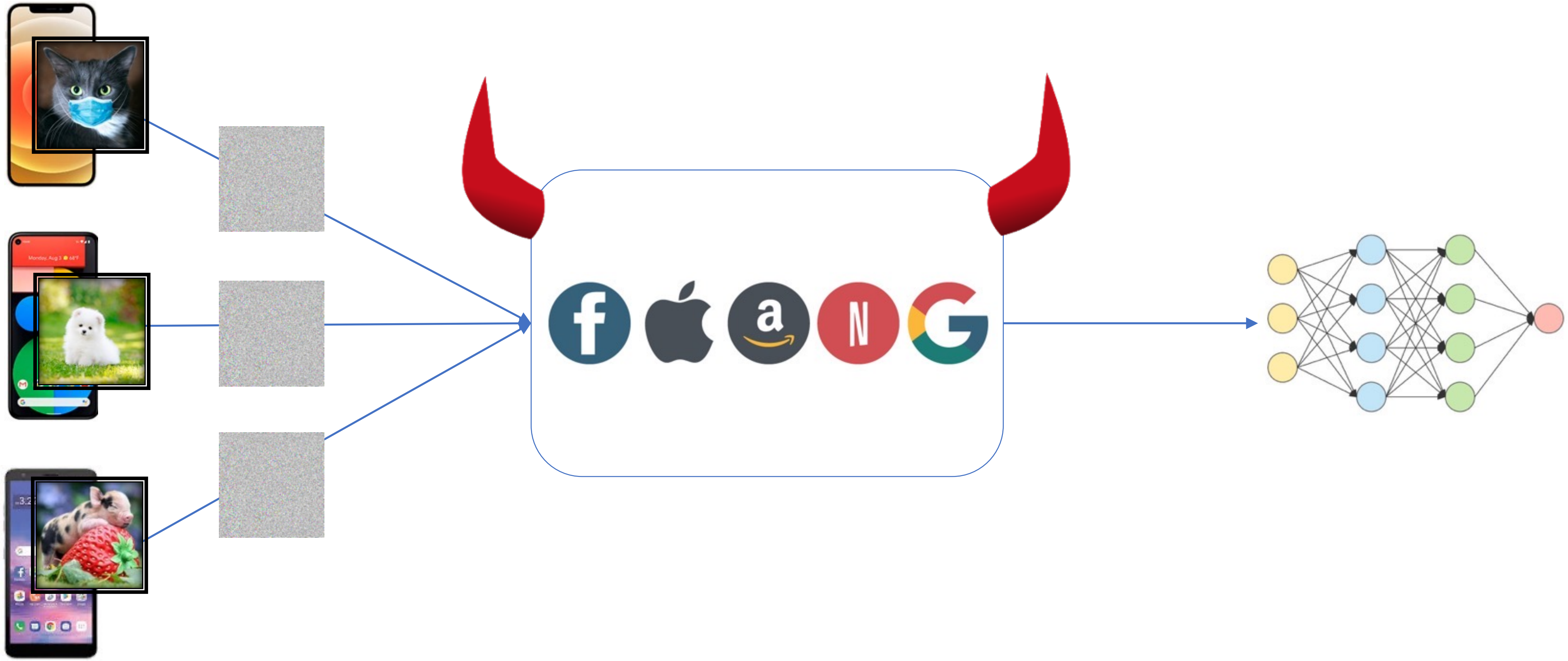


- More attacks
- **Impossibility results**

Is Private Learning Possible with Instance Encoding?, IEEE Security & Privacy 2021

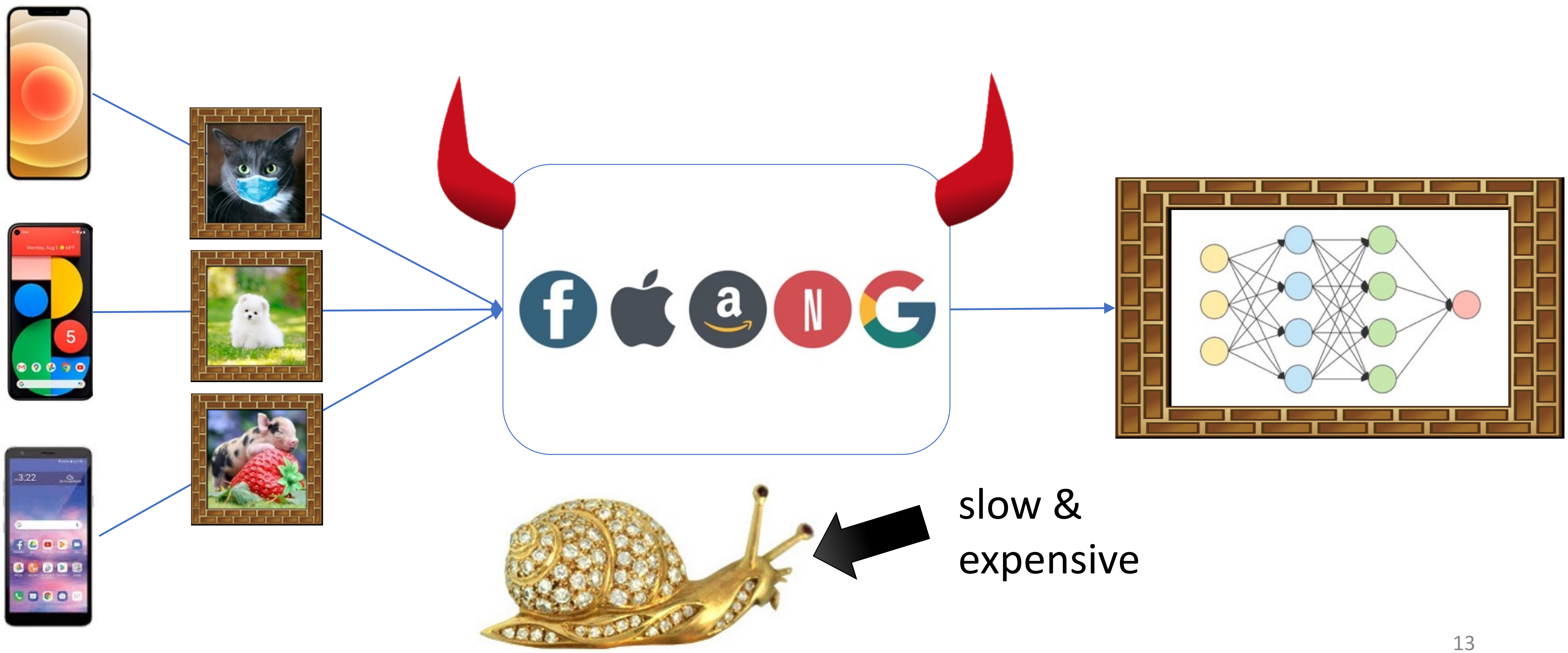
Goal: train a ML model with "privacy"

- ➢ 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***?

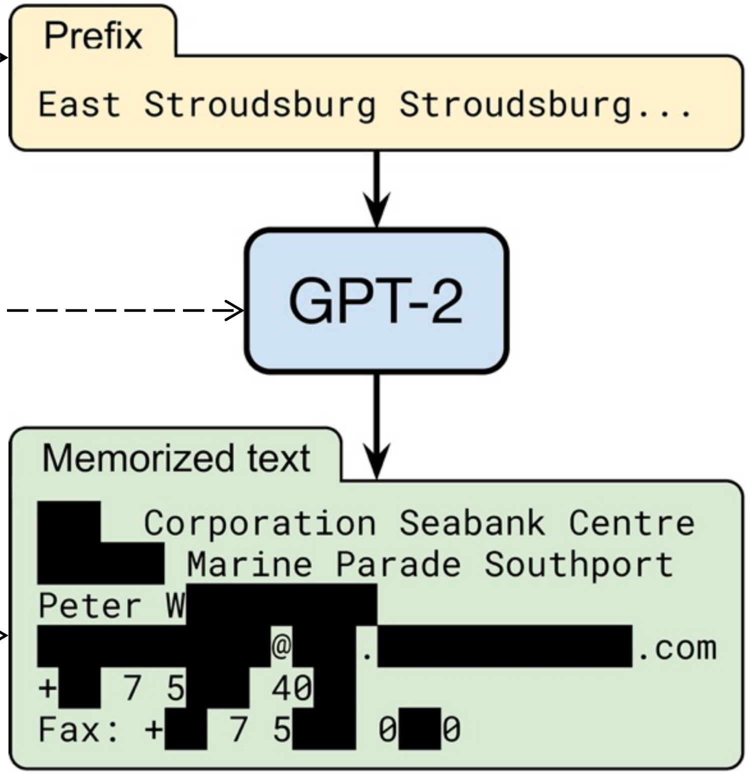
No! The ideal functionality itself can be non-private

Models **memorize** their training data.

random input

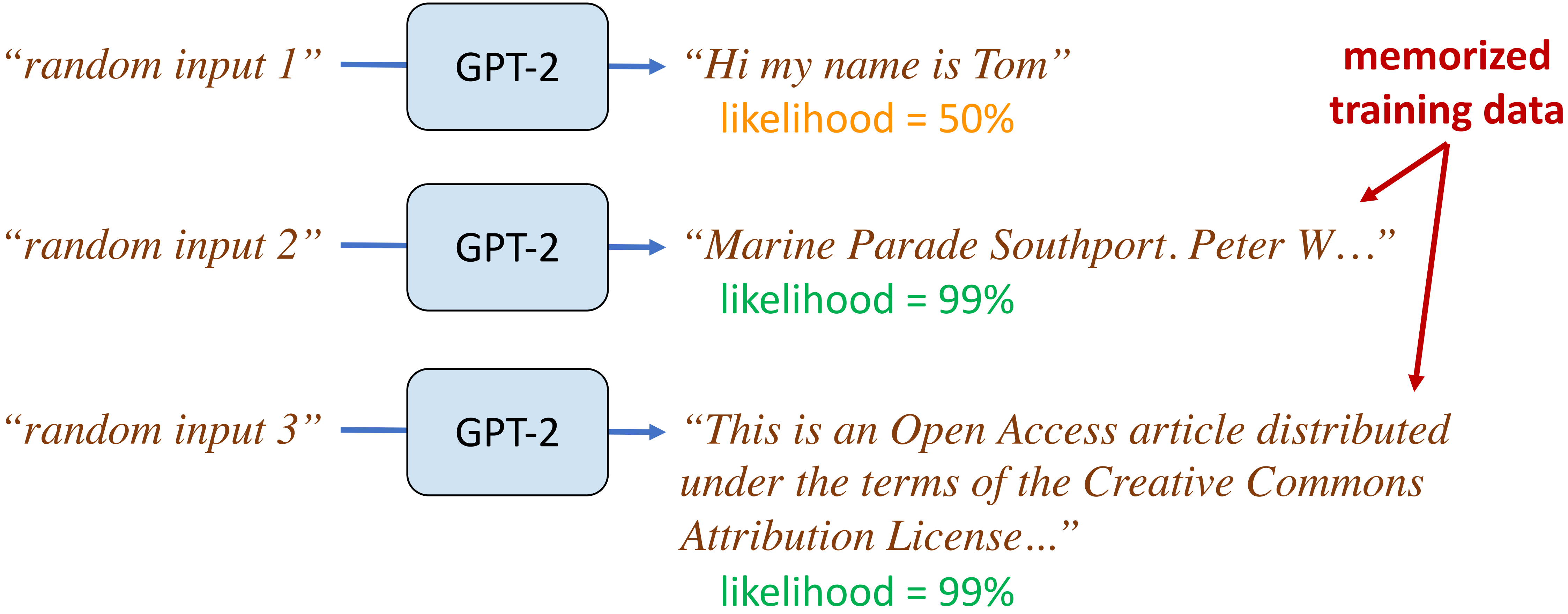
OpenAI's language model trained on text from 8 million web pages

someone's contact information output by the model
(*redacted for privacy*)



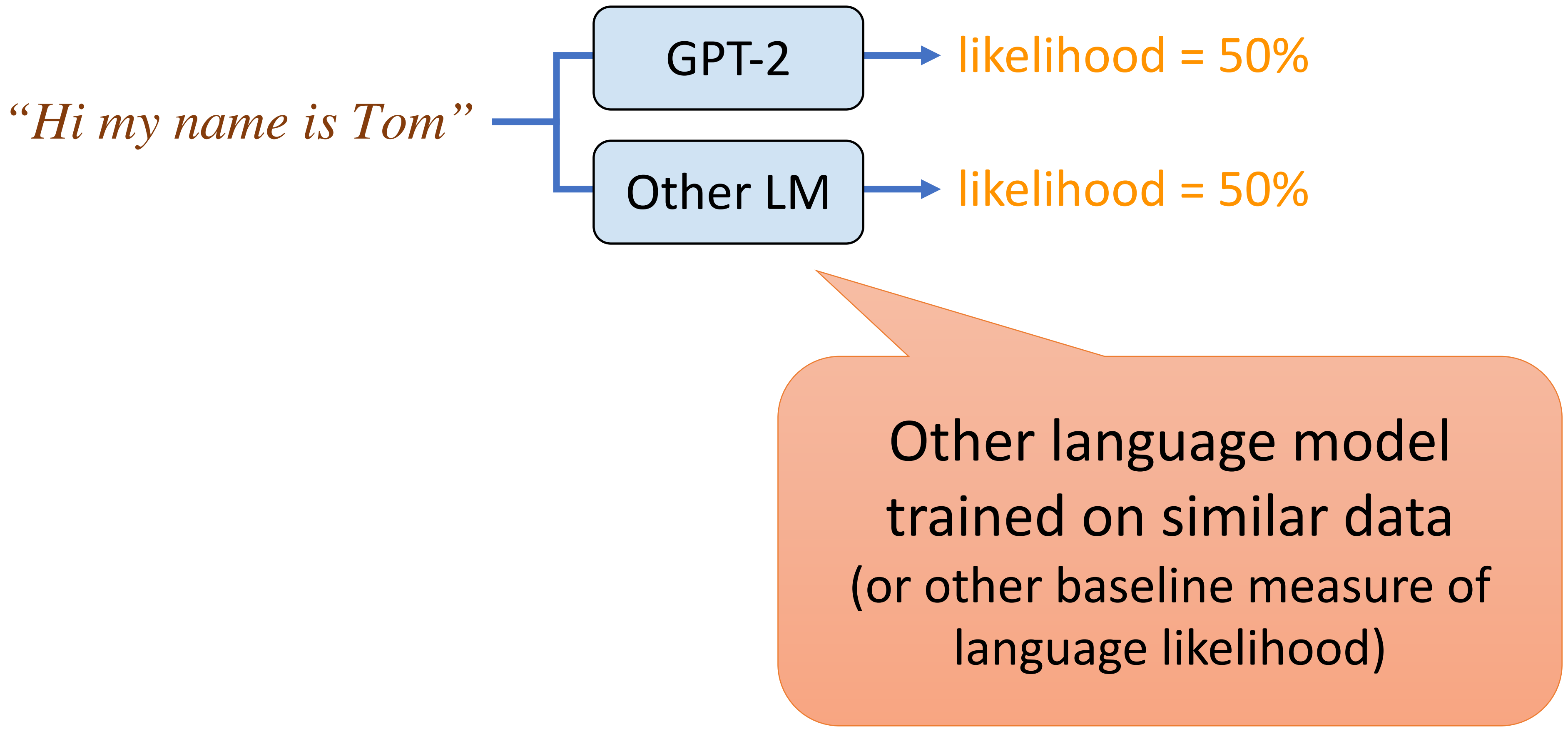
Extracting Training Data from Large Language Models, preprint 2021

Extracting memorized training data by generating *high-likelihood* outputs.



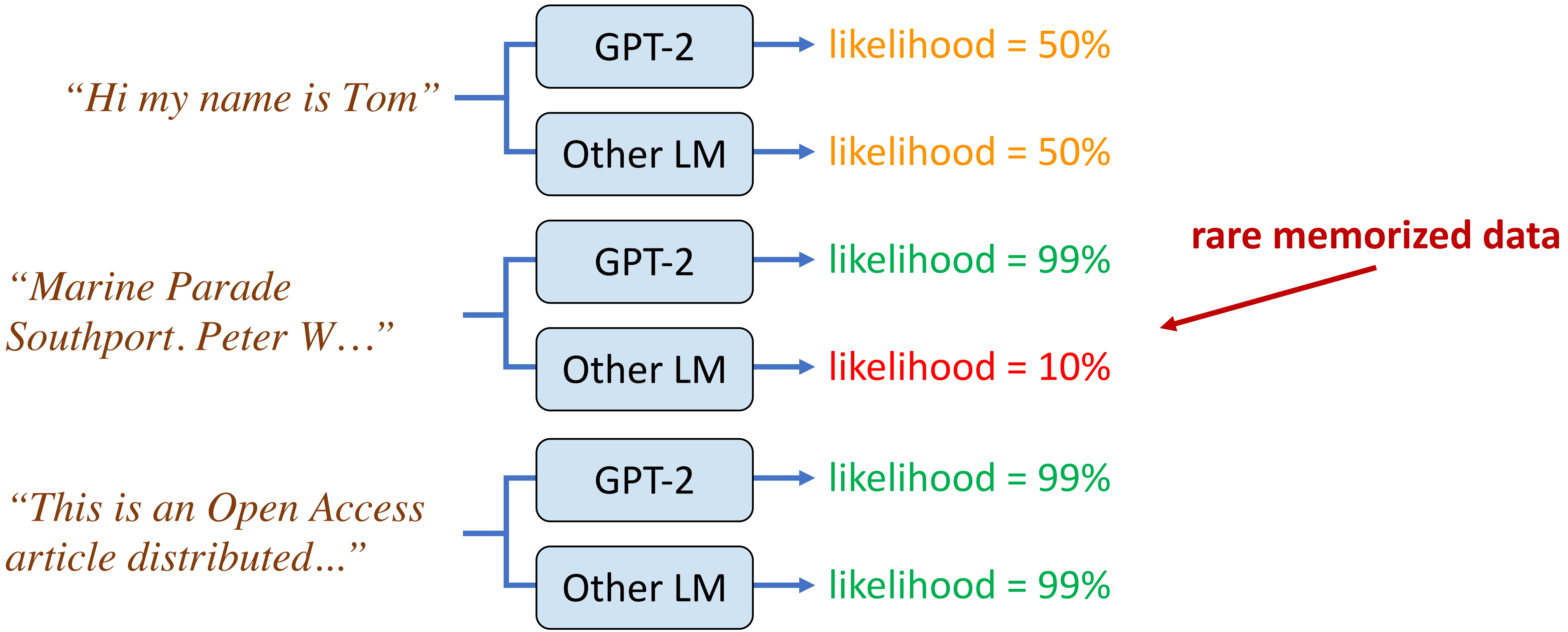
Extracting Training Data from Large Language Models, preprint 2021

Extracting rare memorized training data.



Extracting Training Data from Large Language Models, preprint 2021

Extracting rare memorized training data.



Extracting Training Data from Large Language Models, preprint 2021

Larger models are *less* private.

	Occurrences		Memorized?		
URL (trimmed)	**Docs**	**Total**	**XL**	**M**	**S**
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Extracting Training Data from Large Language Models, preprint 2021

Larger models are *less* private.

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

	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Extracting Training Data from Large Language Models, preprint 2021

Larger models are *less* private.

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

	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/51y/milo_evacua...	1	359	✓	✓	½
/r/zin/hi_my_name...	1	113	✓	✓	
/r/7ne/for_all_yo...	1	76	✓	½	
/r/5mj/fake_news_...	1	72	✓		
/r/5wn/reddit_admi...	1	64	✓	✓	
/r/lp8/26_evening...	1	56	✓	✓	
/r/jla/so_pizzagat...	1	51	✓	½	
/r/ubf/late_night...	1	51	✓	½	
/r/eta/make_christ...	1	35	✓	½	
/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

Extracting Training Data from Large Language Models, preprint 2021

Larger models are *less* private.

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

	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]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

Extracting Training Data from Large Language Models, preprint 2021

Larger models are *less* private.

	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

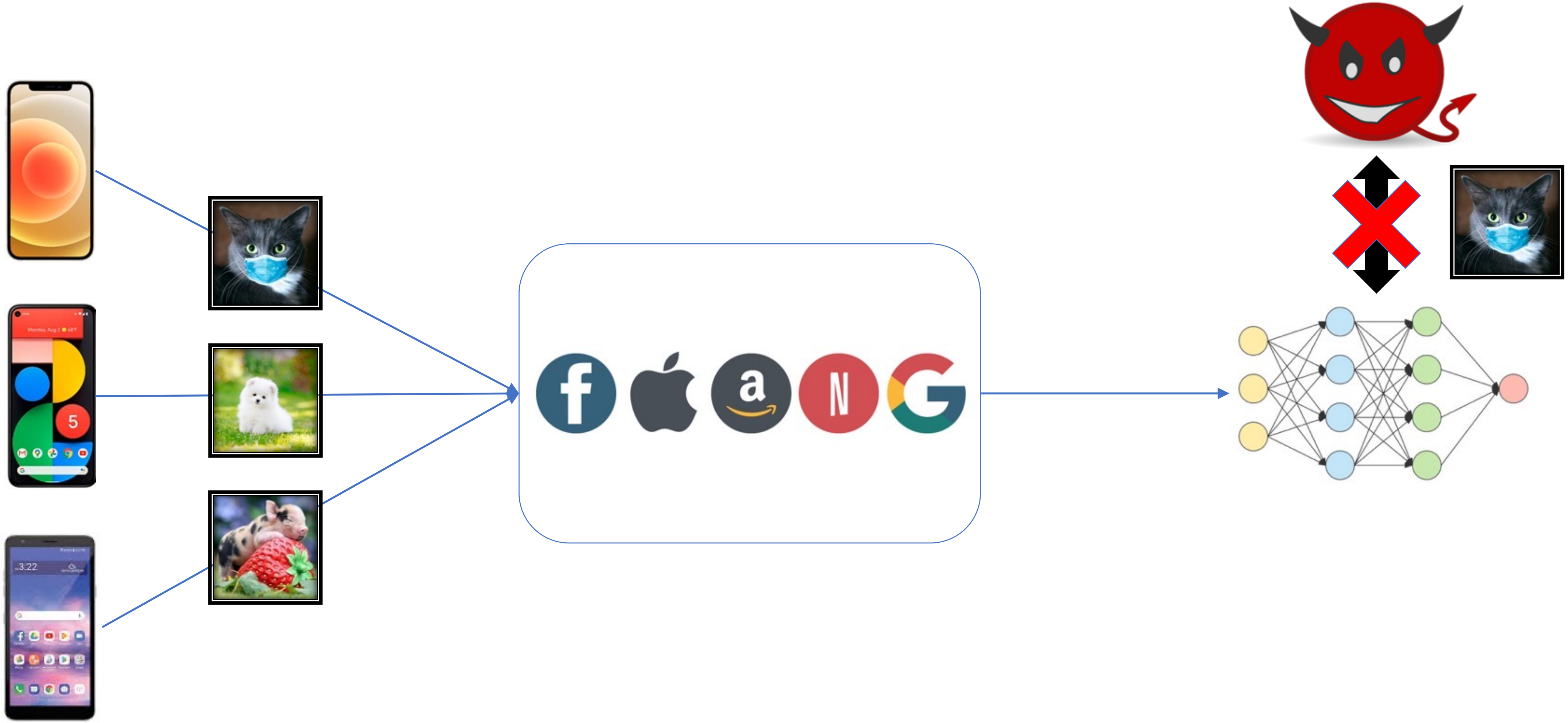
the largest GPT-2 model memorized an entire URL that appeared **only 33 times** in a single document

Extracting Training Data from Large Language Models, preprint 2021

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*

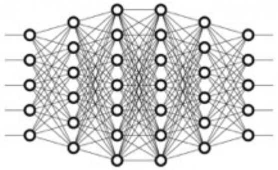
- *better accuracy* than with deep learning methods.
  using domain-specific *feature engineering*

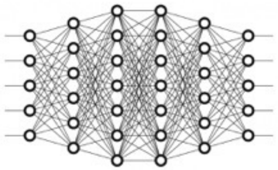
Goal: train a ML model with "privacy"

- 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

$\Pr[A_{\text{train}}($  $) =$  $]$

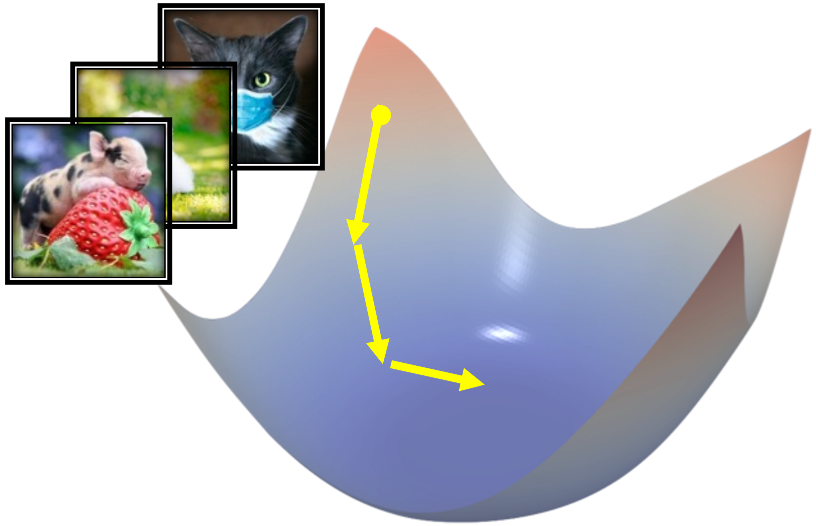
$\Pr[A_{\text{train}}($  $) =$  $]$

$\le e^{\varepsilon}$

How? Private Gradient Descent

gradient descent

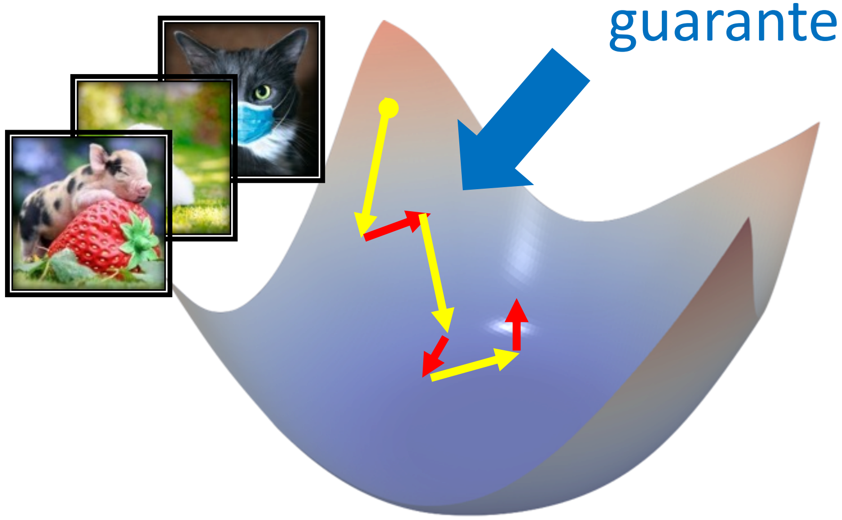
(SGD)



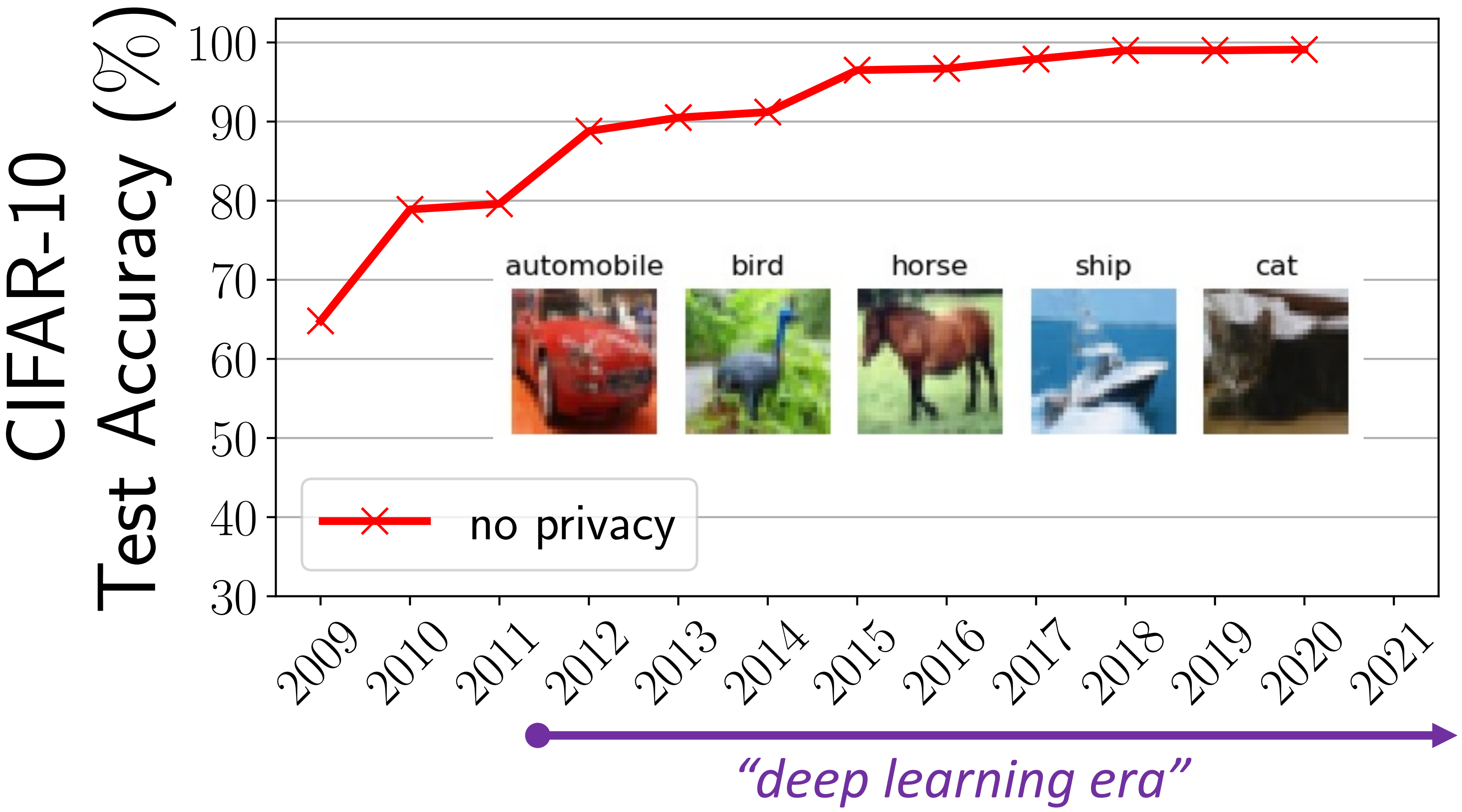
private gradient descent

(DP-SGD)

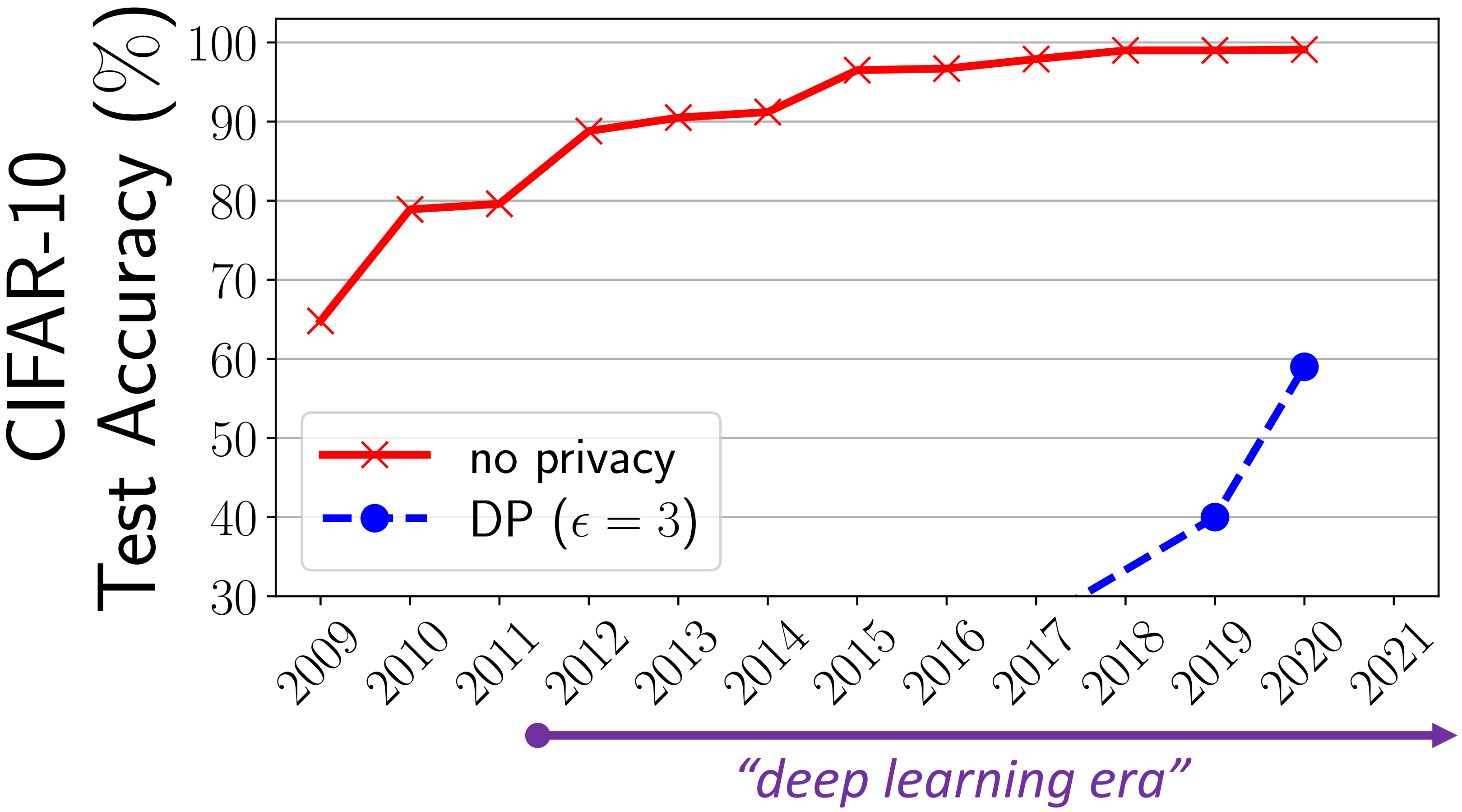
Chaudhuri et al., ’11; Bassily et al. ’14;
Shokri & Shmatikov ’15; Abadi et al. ’16



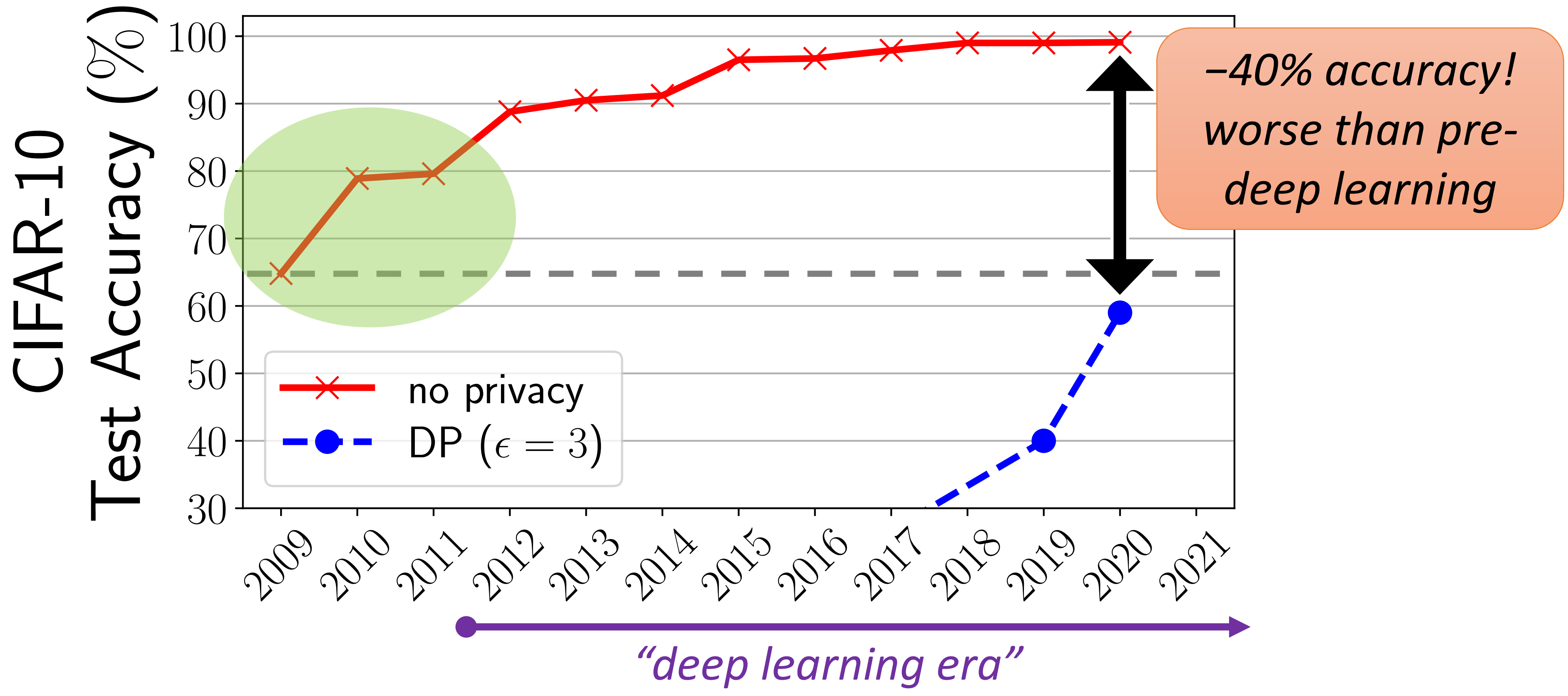
Non-private deep learning can achieve near-perfect accuracy.



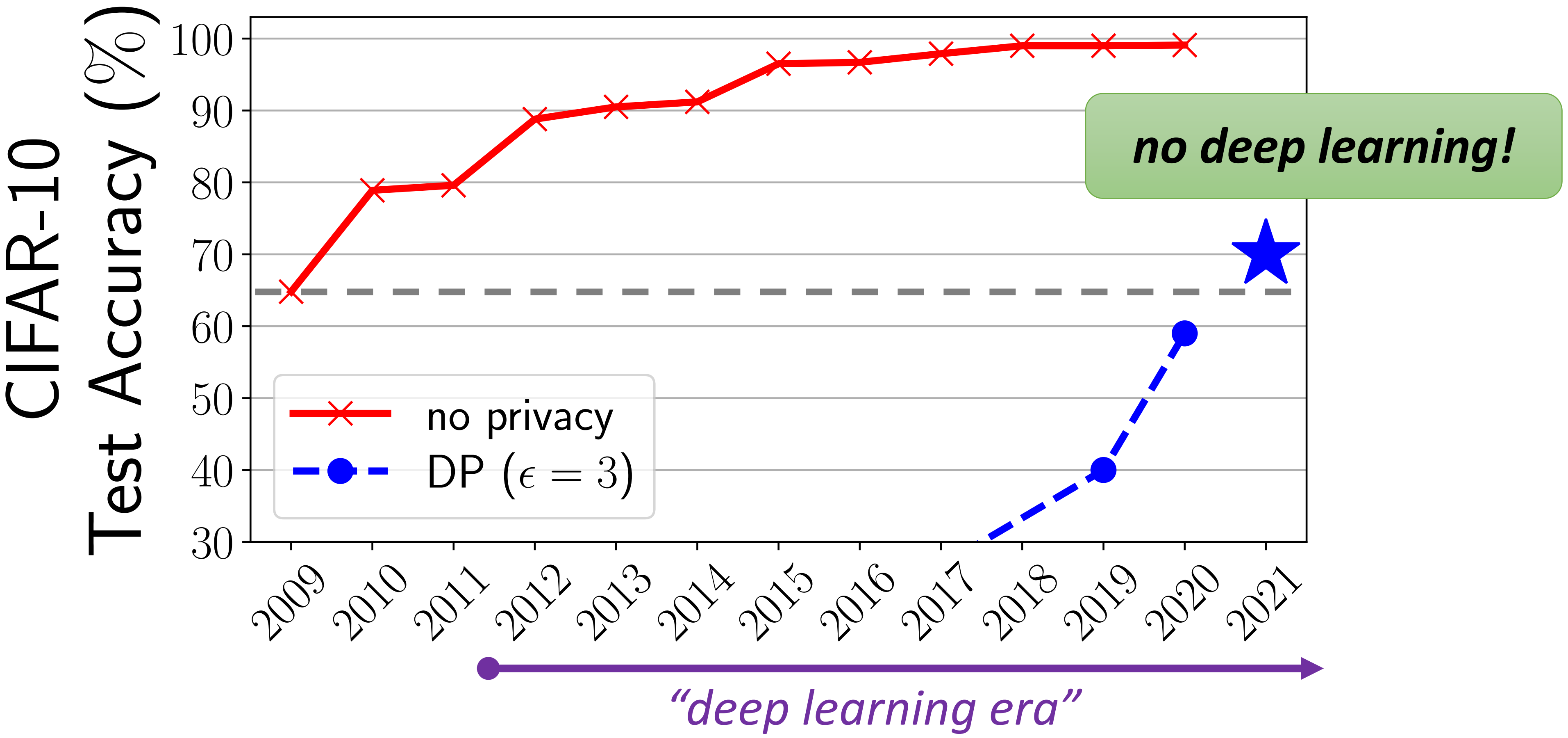
Differentially private deep learning lowers accuracy significantly.



Differentially private deep learning lowers accuracy significantly.



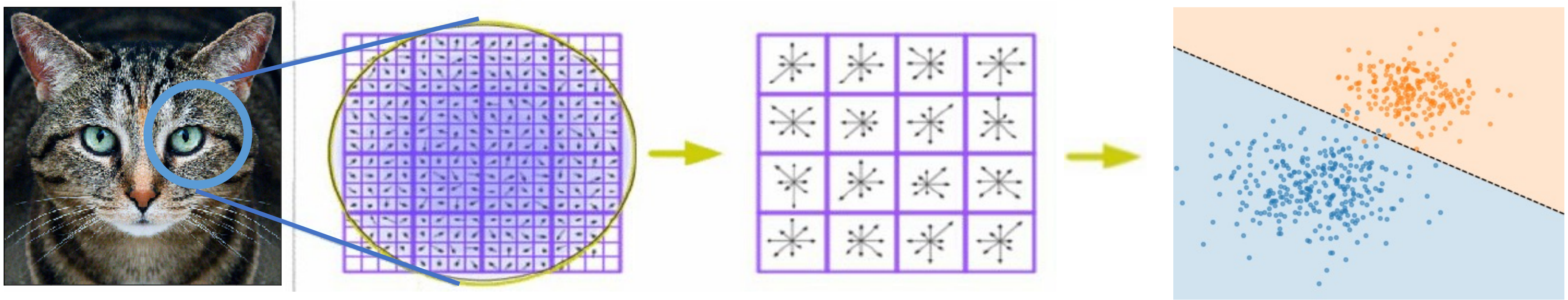
Differential privacy *without deep learning* improves accuracy.



Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

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], ...



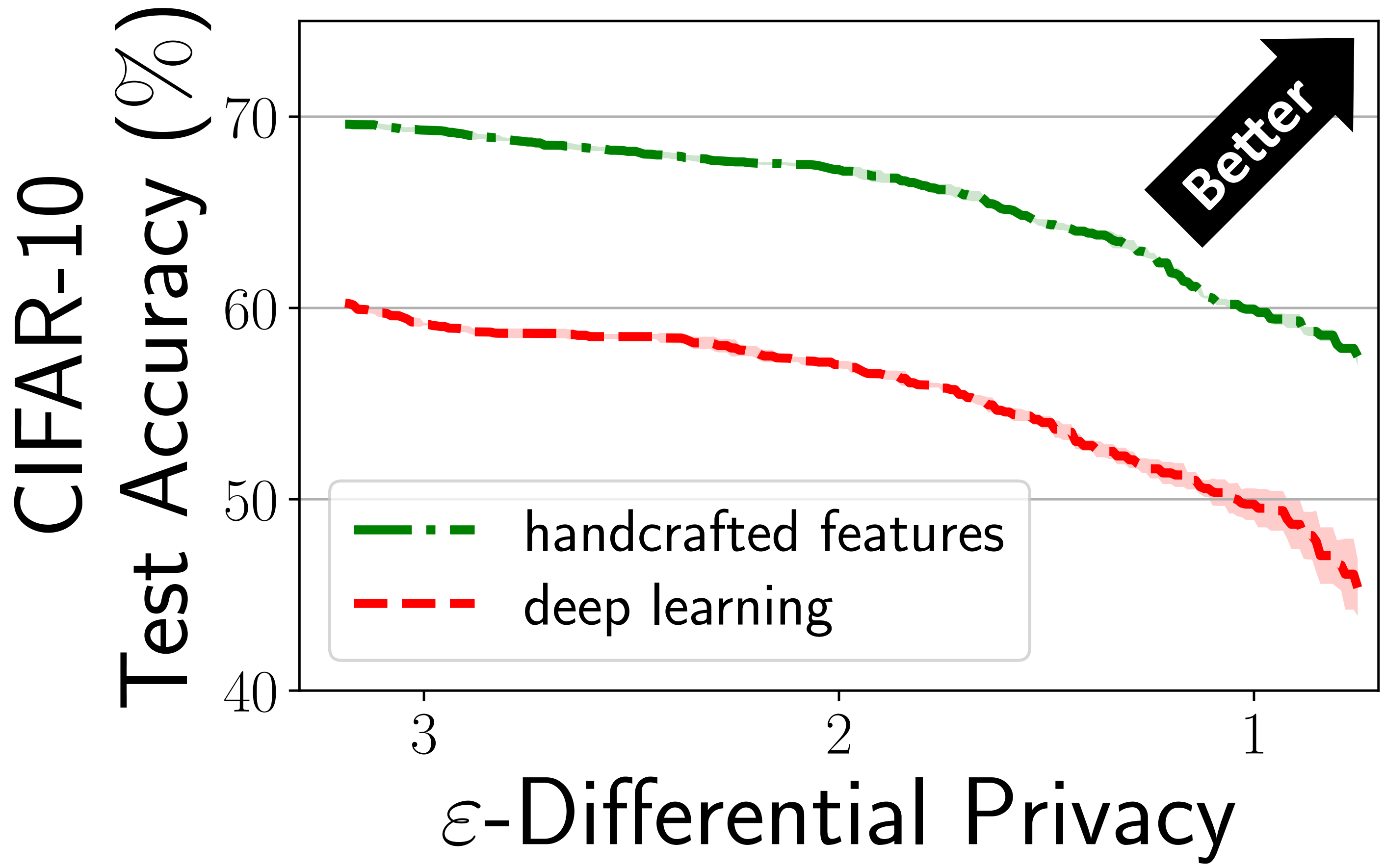
"handcrafted features"
(no learning involved)

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.

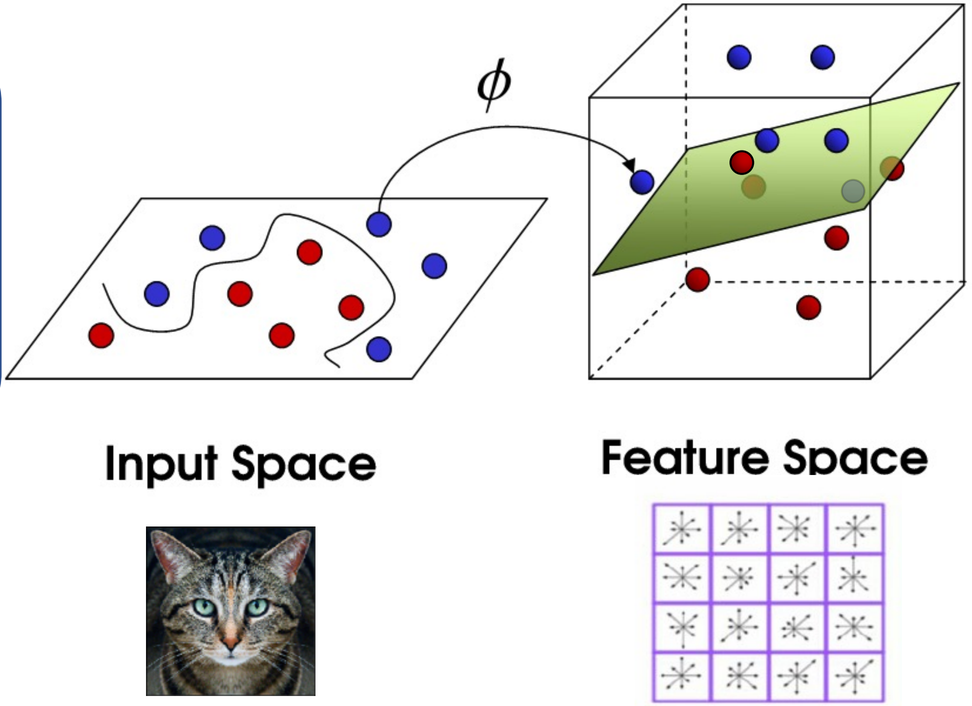


Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

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

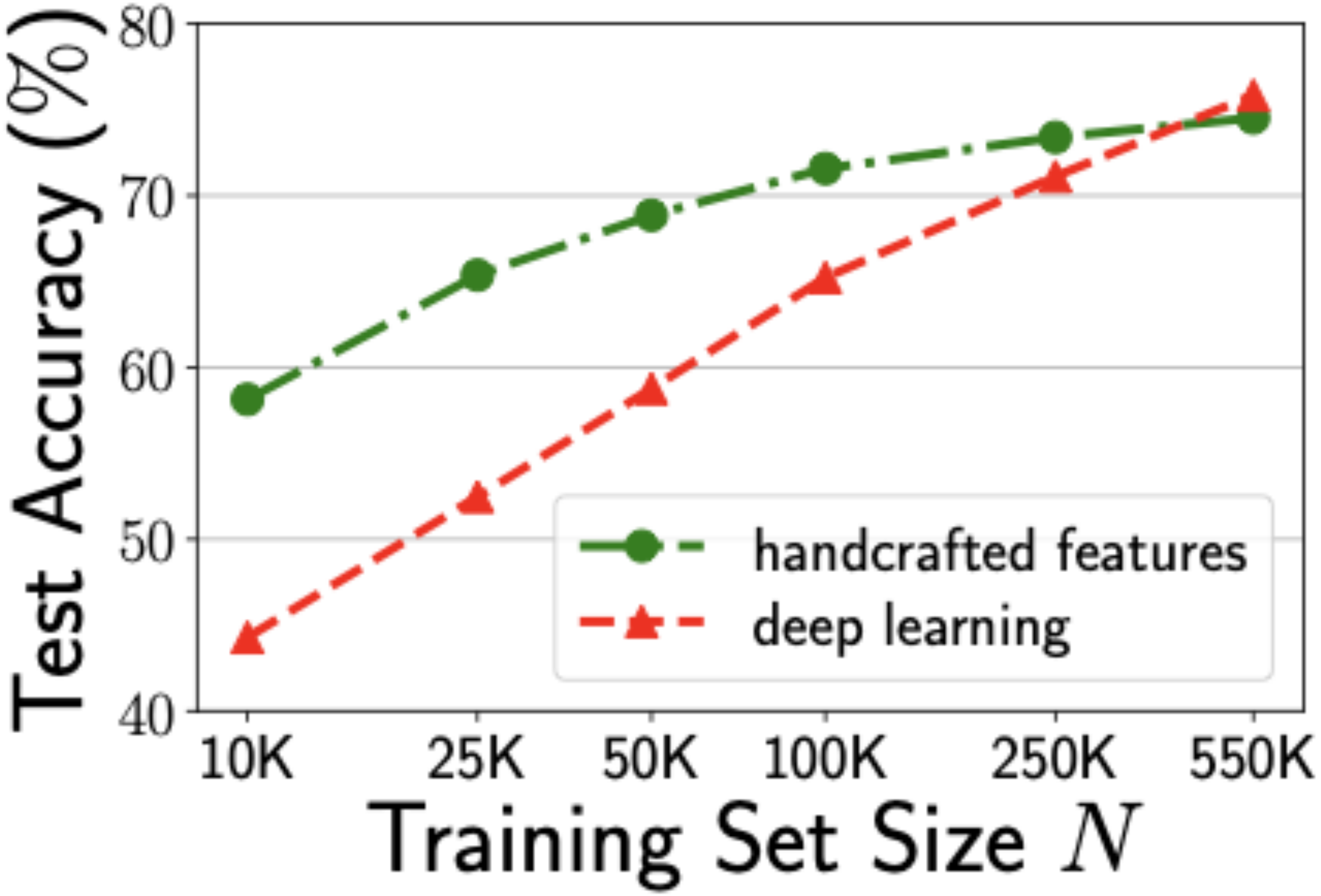


in feature space, maximal accuracy is reduced but *learning progresses faster*

good for privacy

Surpassing handcrafted features with *more private data.*

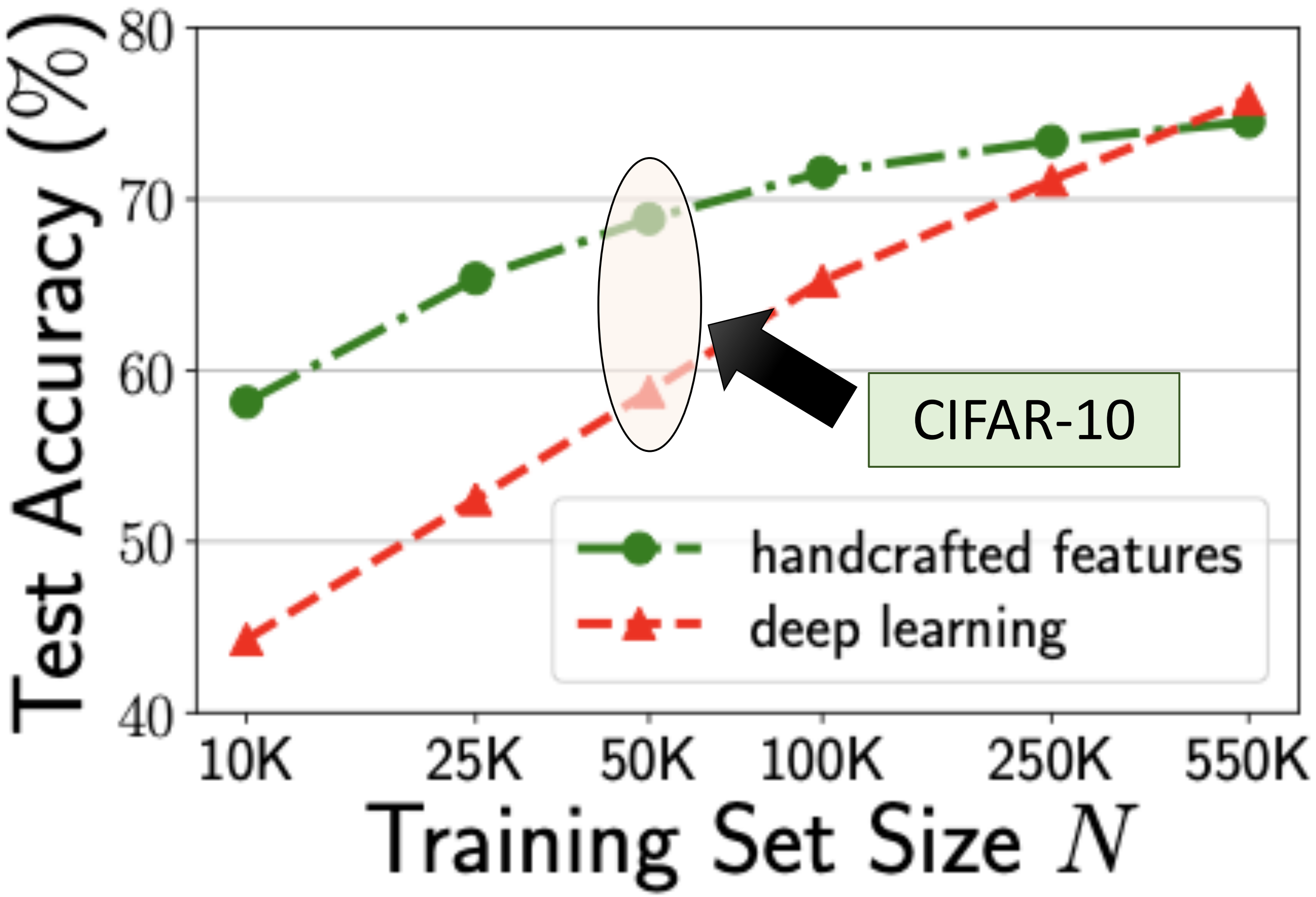
(for ε = 3)



Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

Surpassing handcrafted features with *more private data.*

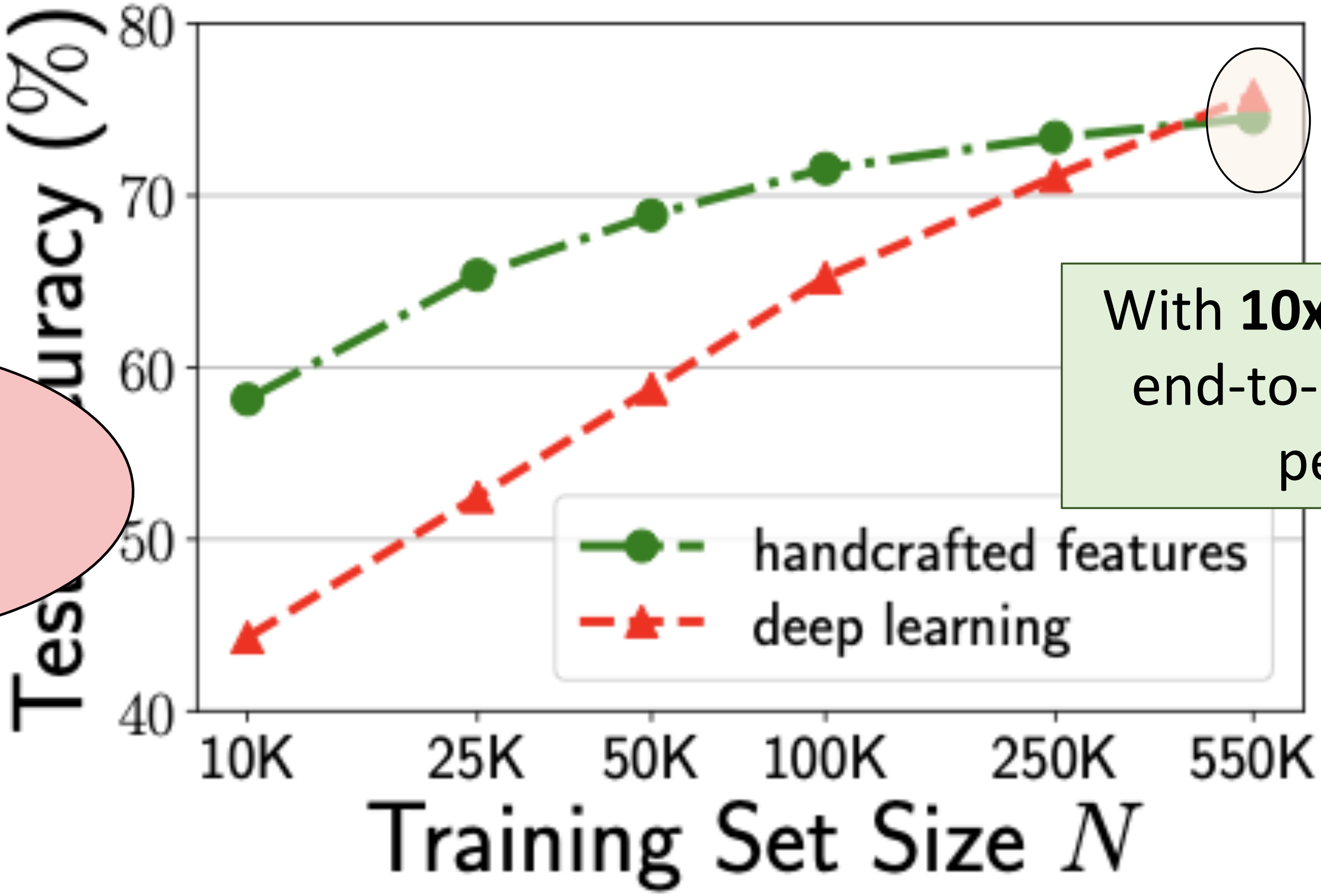
(for ε = 3)



Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

Surpassing handcrafted features with *more private data.*

(for ε = 3)



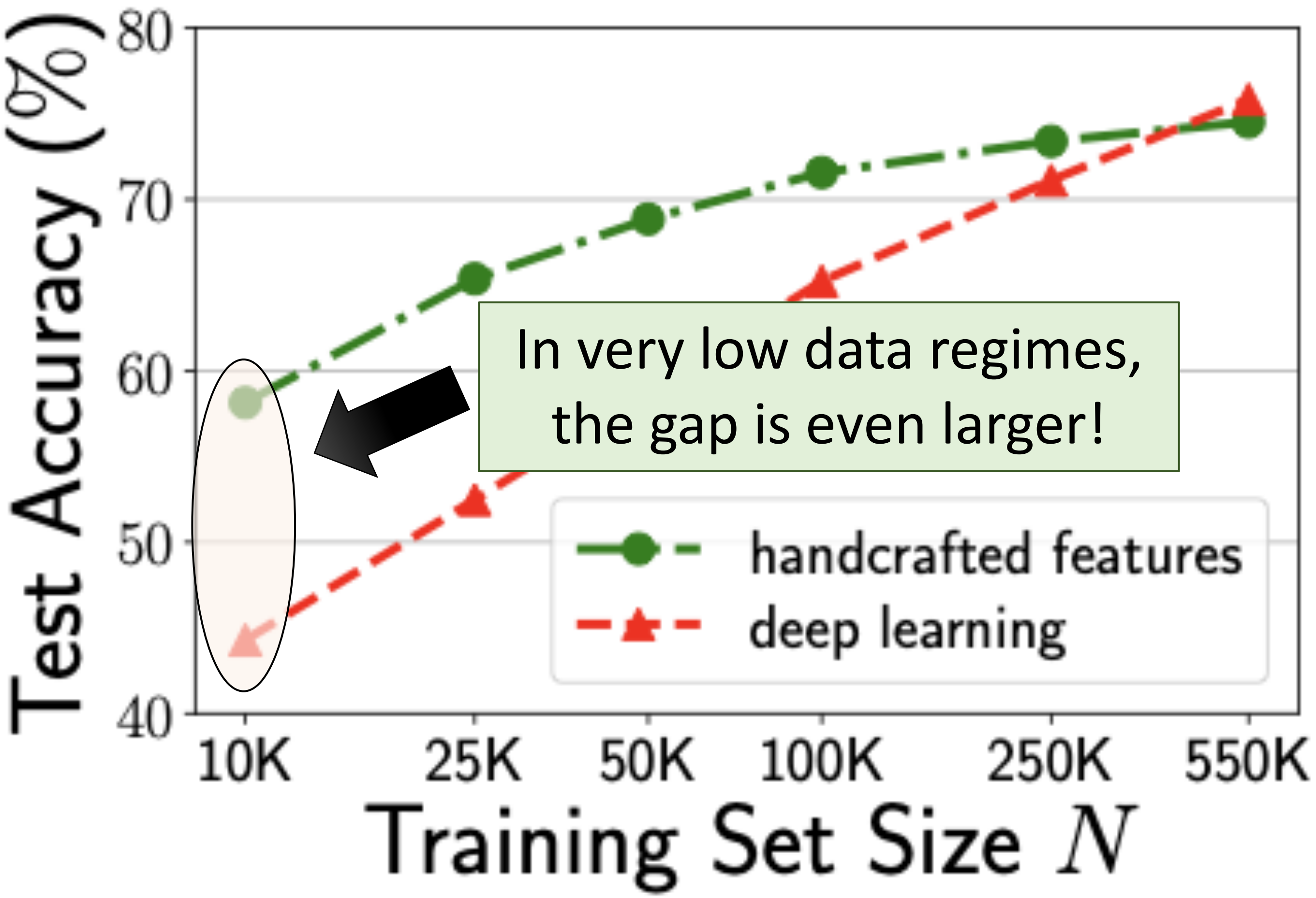
With **10x more private data** end-to-end deep learning performs best

collecting more data is good for your privacy!

Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

Surpassing handcrafted features with *more private data.*

(for ε = 3)

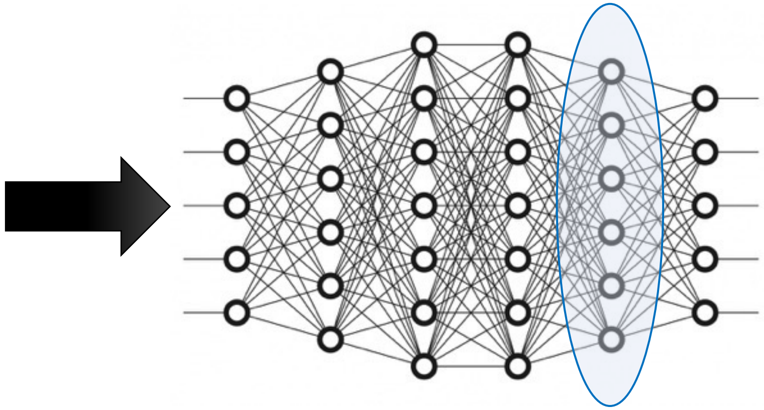


In very low data regimes, the gap is even larger!

Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

Surpassing handcrafted features with *more **public** data.*

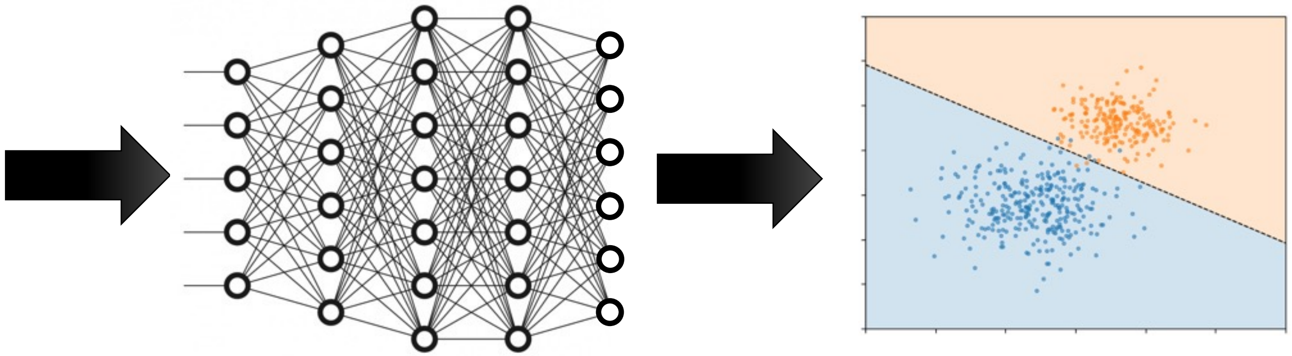
public data



train a feature extractor on public data...

...transfer and fine-tune on private data

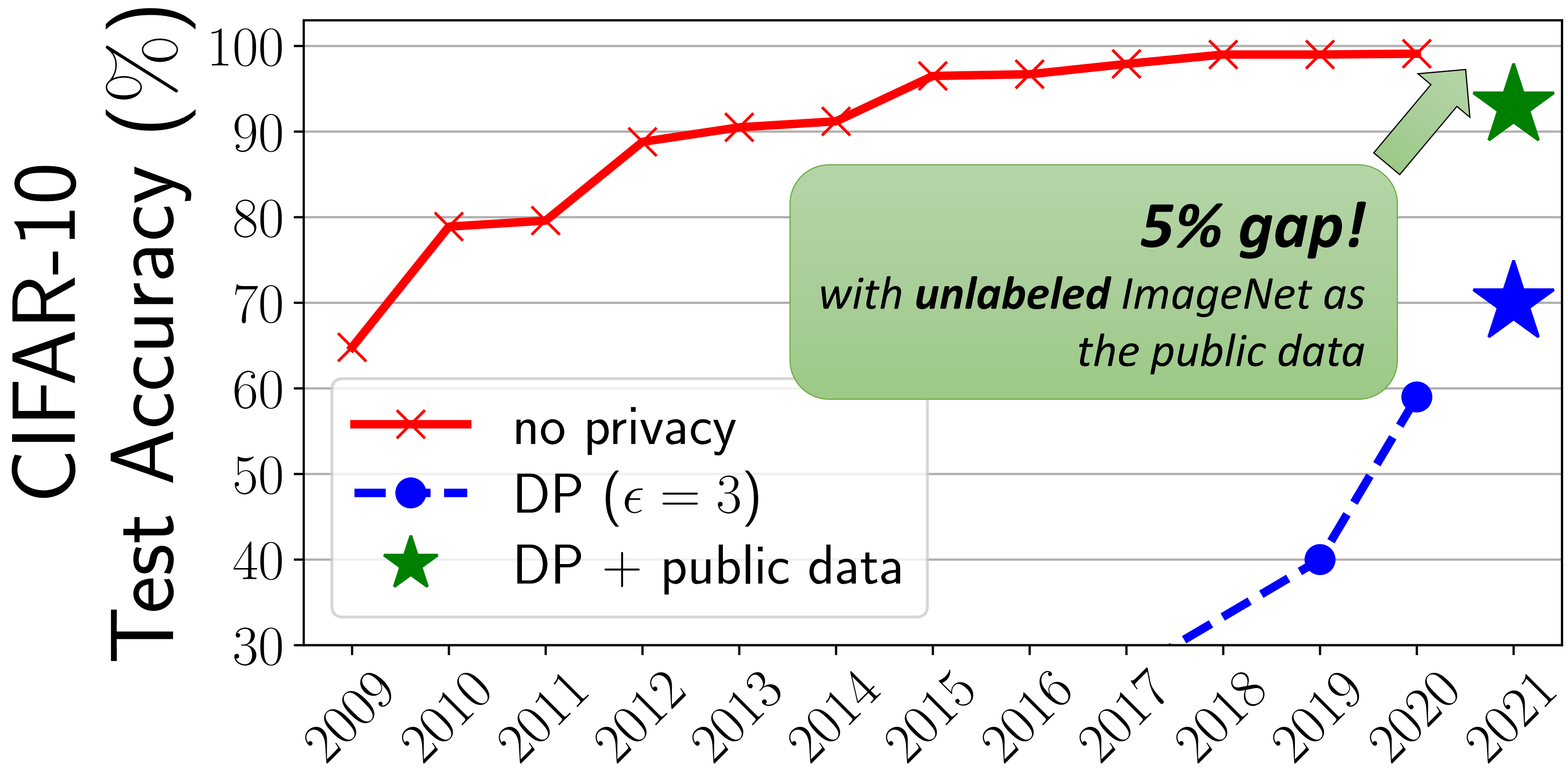
private data



privacy free

Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

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

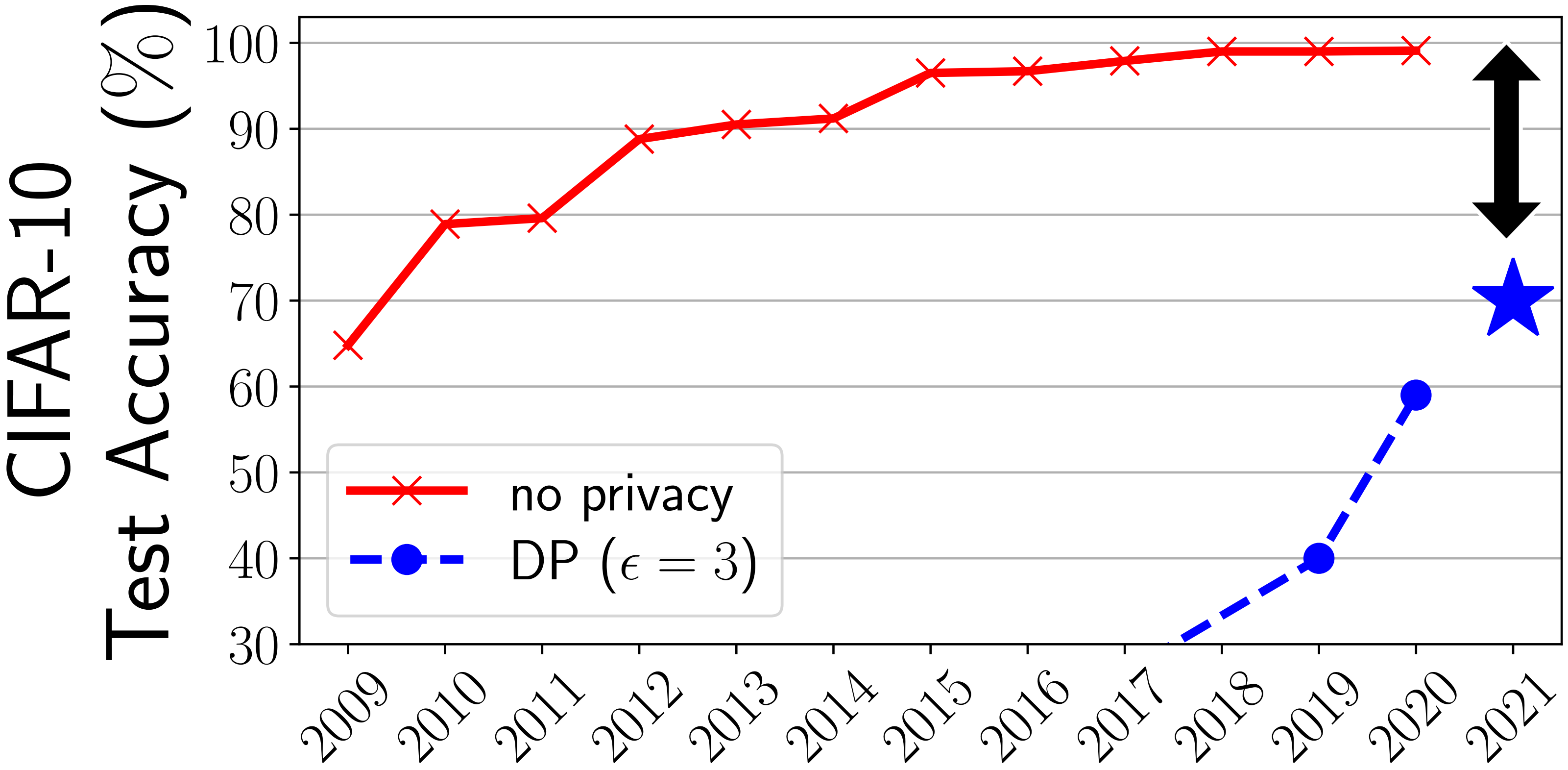


Differentially Private Learning Needs Better Features (or Much More Data), ICLR 2021

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?



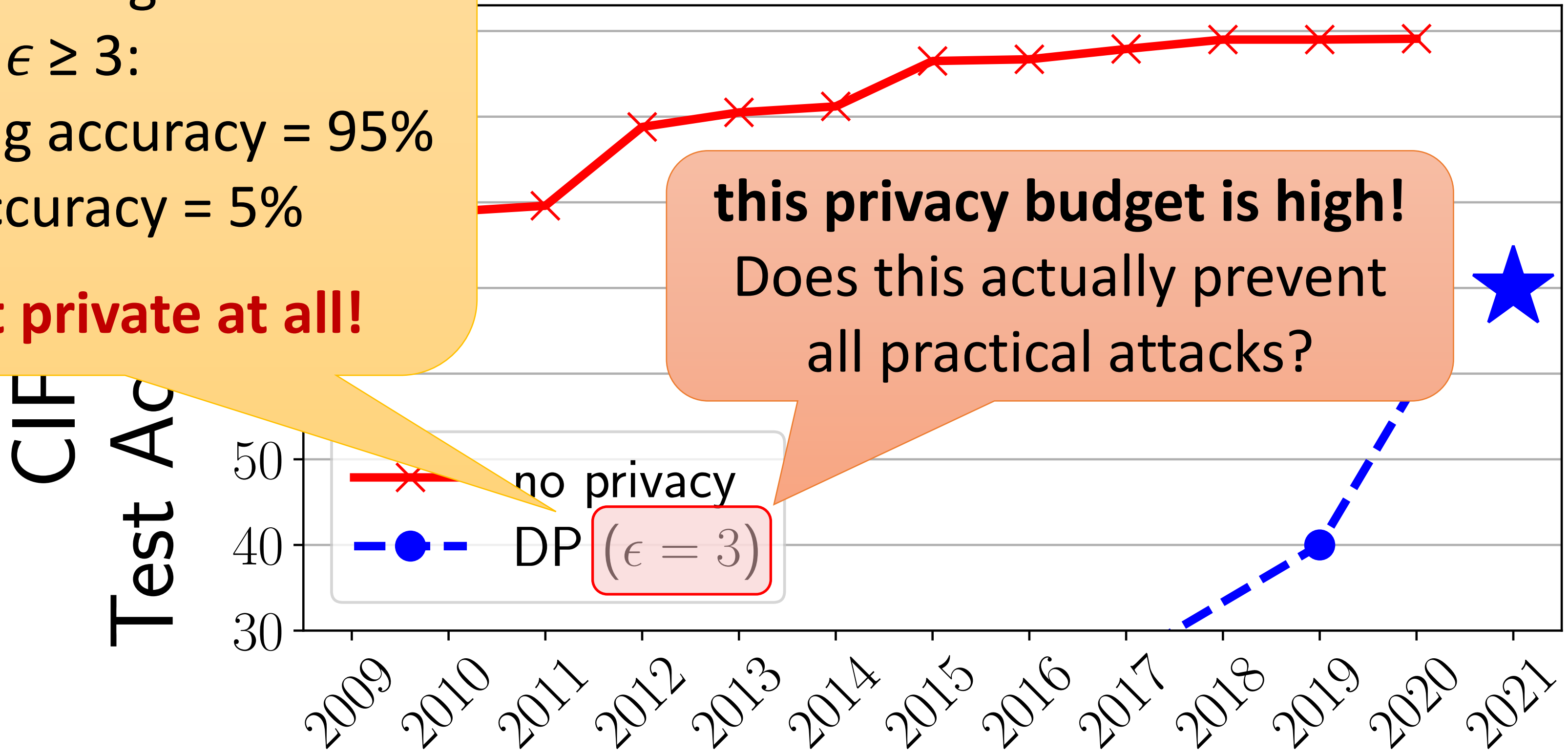
still a 30% accuracy gap!

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%

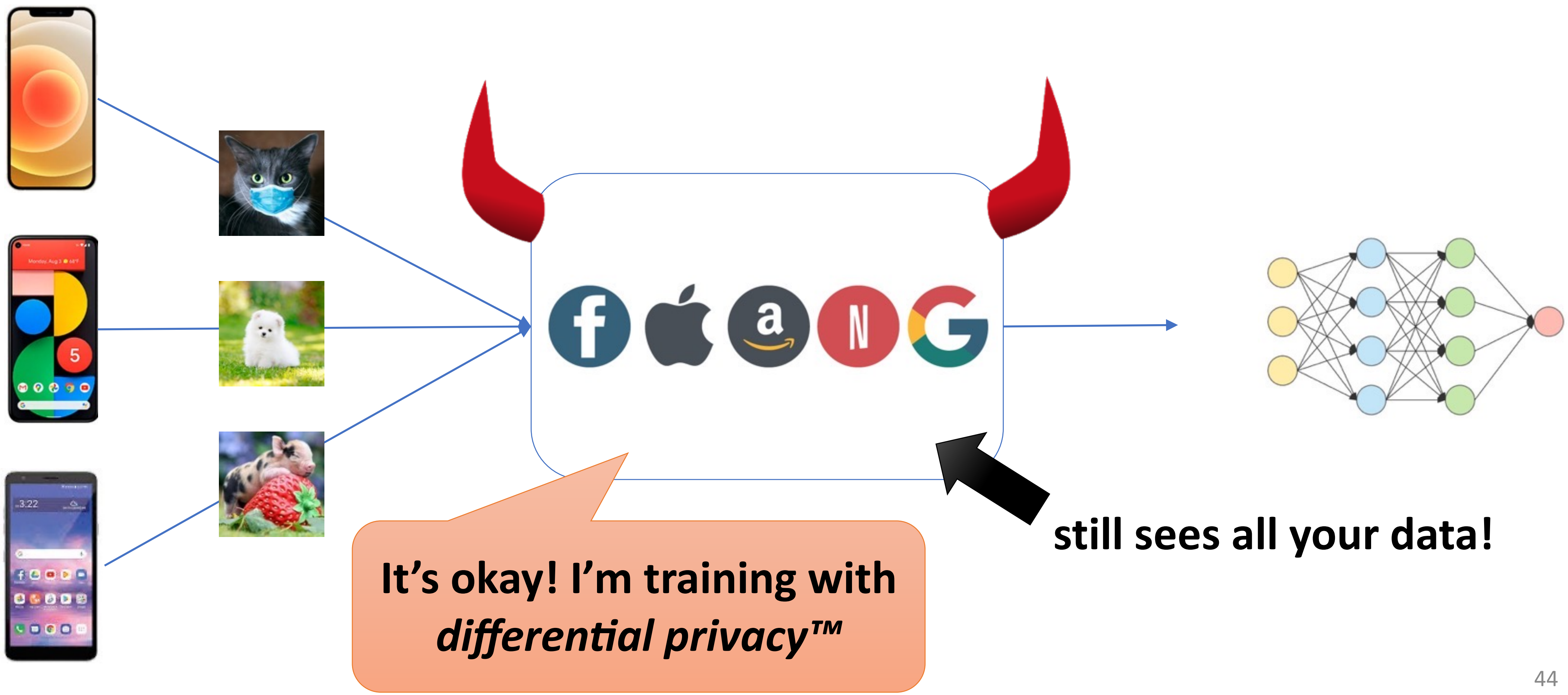
$\Rightarrow$ **This is not private at all!**



this privacy budget is high!
Does this actually prevent
all practical attacks?

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!