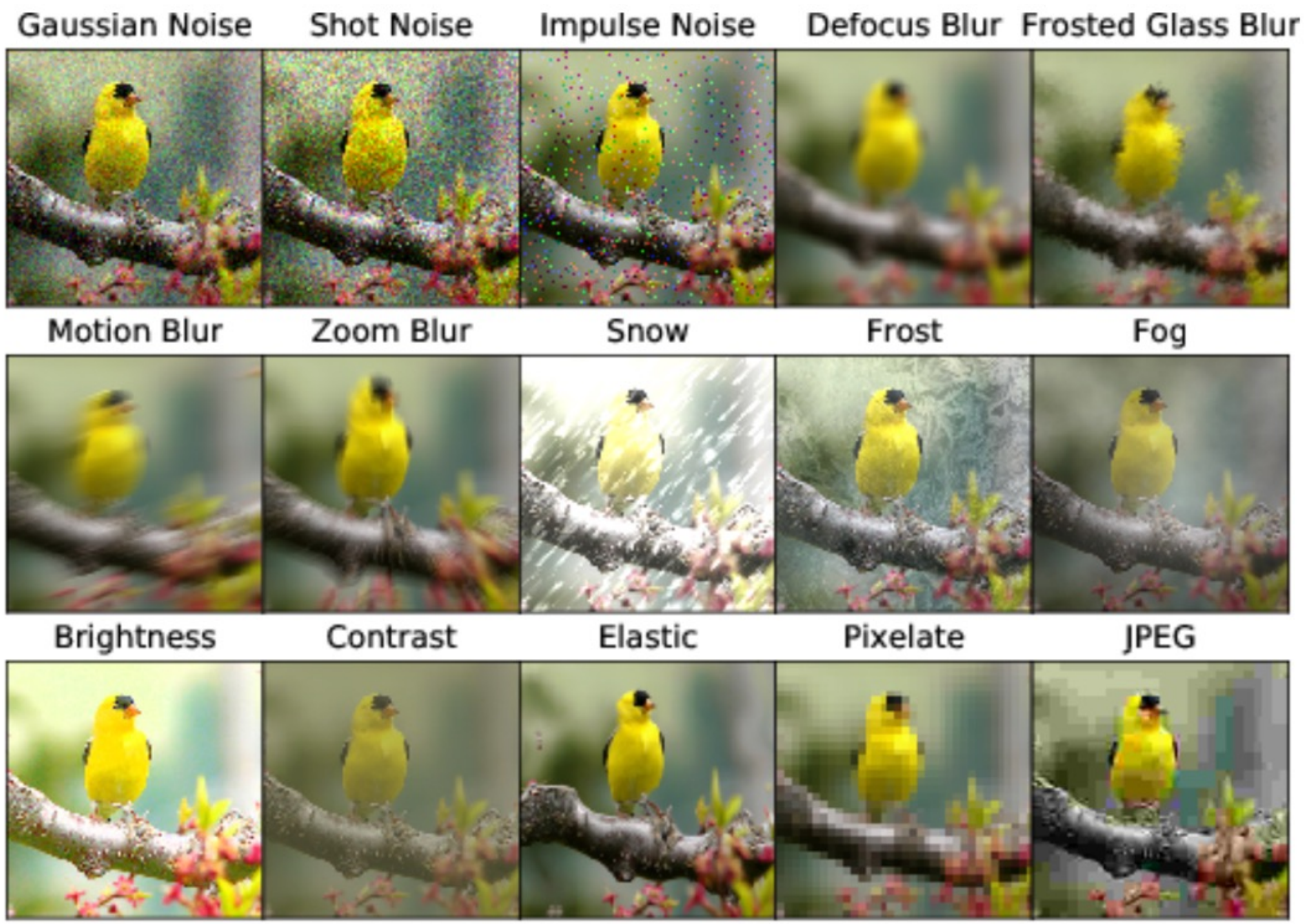
Is anything really OOD anymore? And what does this mean for privacy?

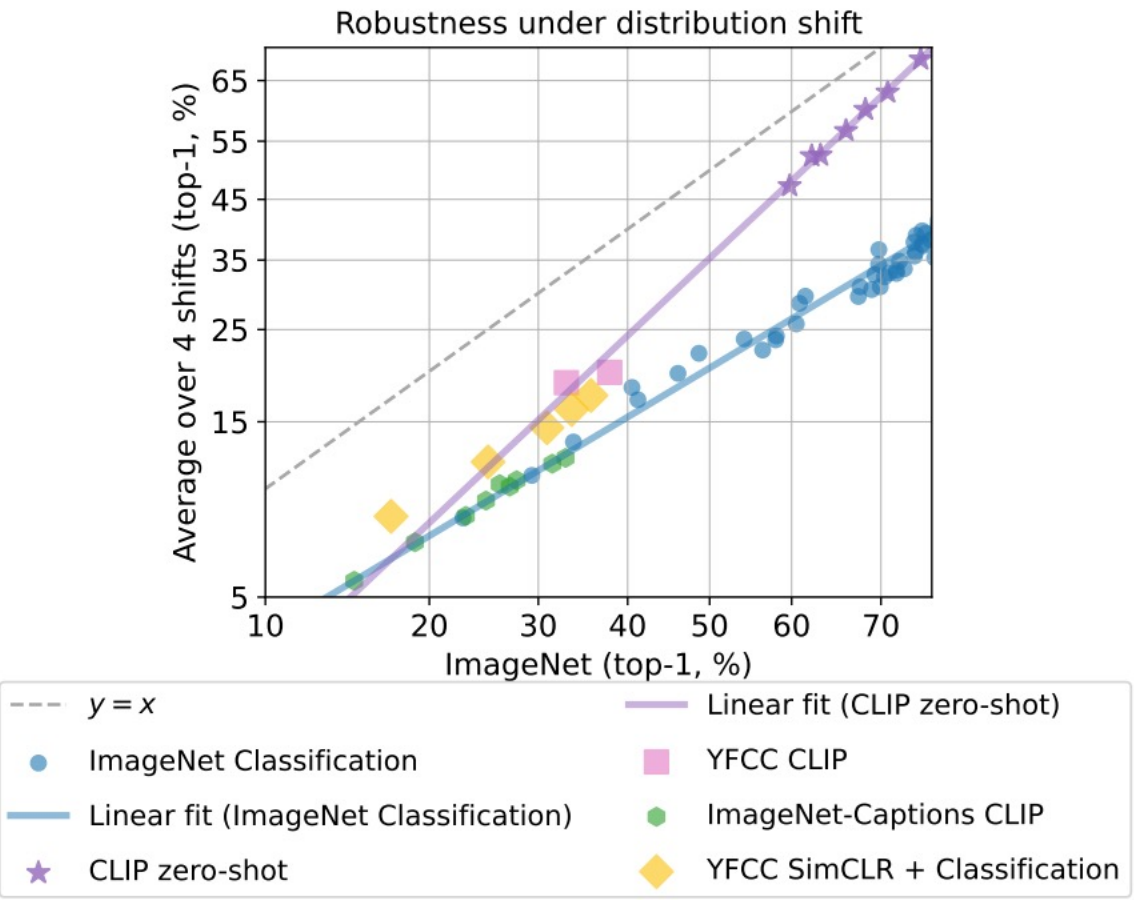
Florian Tramèr
ETH Zürich

(based on joint work with Gautam Kamath and Nicholas Carlini)

Models trained on one dataset can be brittle on slightly modified data.



But many of these OOD benchmarks seem to be “solved” with more pre-training.



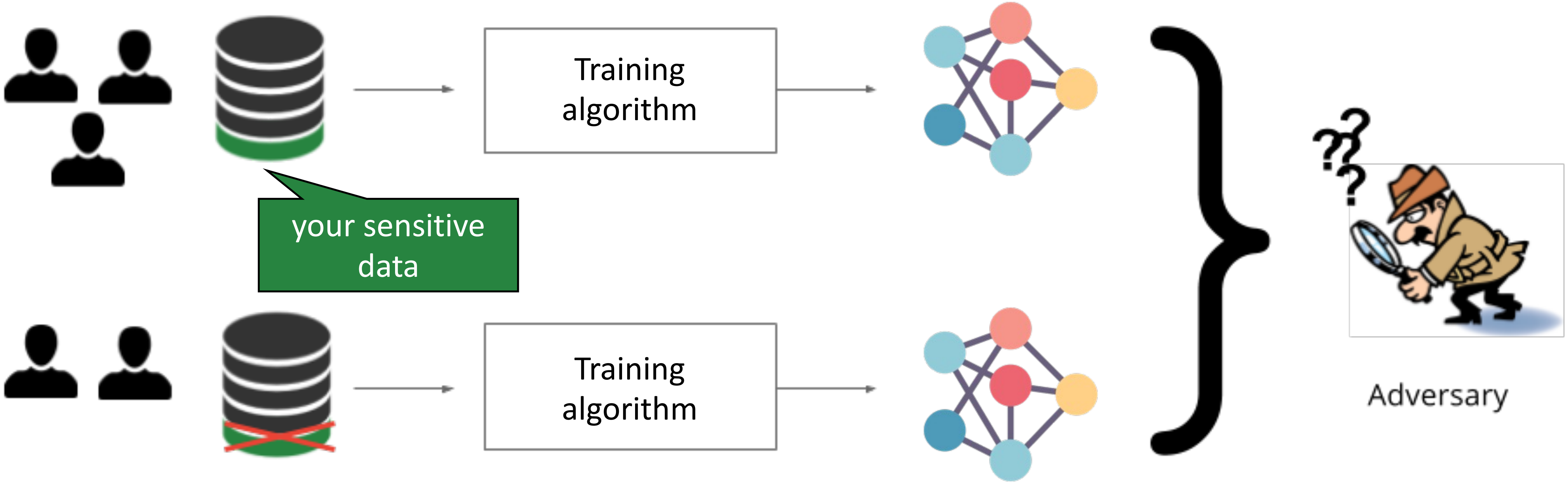
Fang et al. 2022

Is this still “out of distribution” generalization?

distribution distribution


IMAGENET
LAION
Corporate needs you to find the differences
between this ~~picture~~ and this ~~picture~~.
distribution
distribution
distribution
They're the same ~~picture.~~

Today: what does recent "OOD progress" mean for private learning?

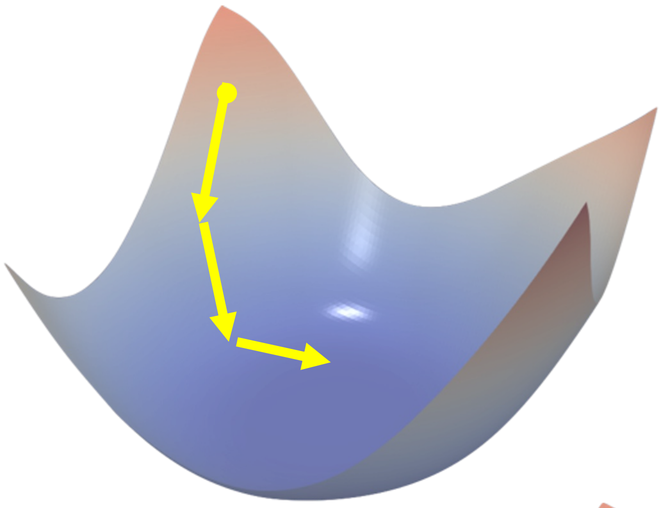


Formally: training with *differential privacy*

$$\frac{\Pr[\mathrm{Train}(D) = \quad]}{\Pr[\mathrm{Train}(D + \{x\}) = \quad]} \leq e^{\varepsilon}$$

Differentially private learning is possible with *noisy* gradient descent.

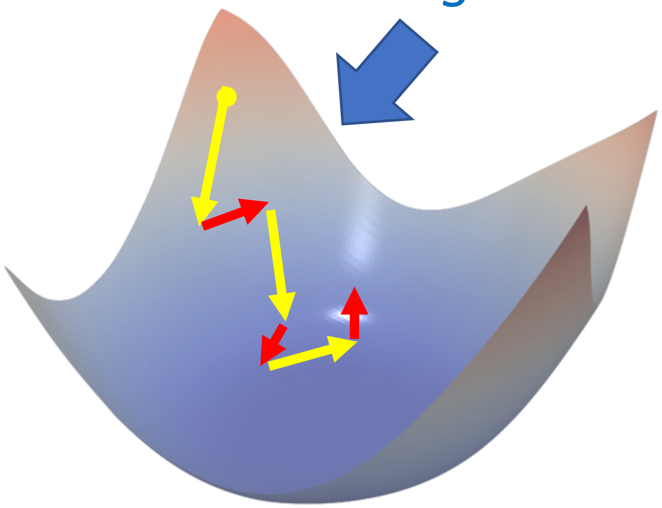
Gradient descent



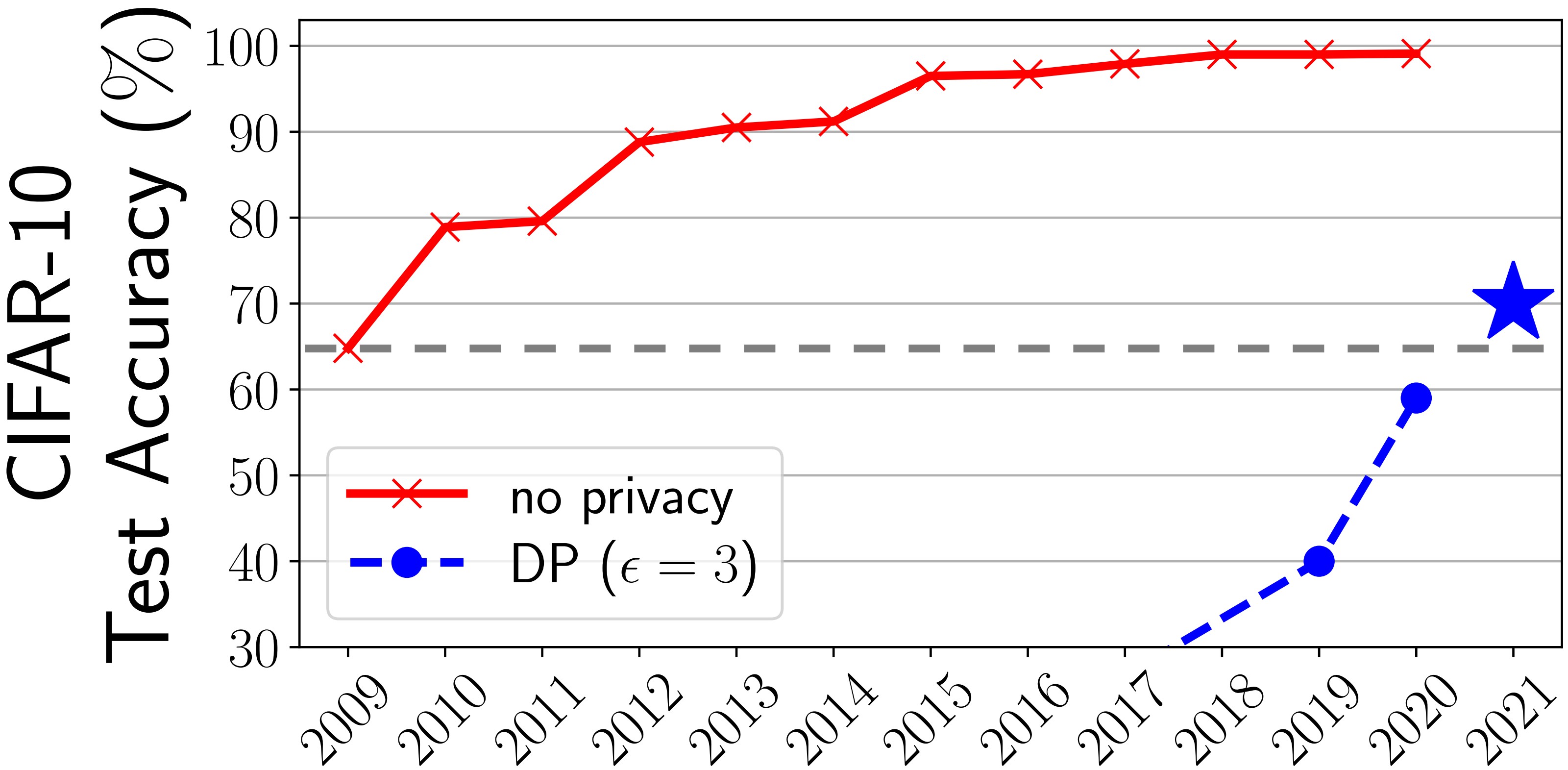
Private gradient descent

[Chaudhuri et al., ‘11], [Bassily et al. ‘14],
[Shokri & Shmatikov ‘15], [Abadi et al. ‘16], ...

add noise to each step to guarantee privacy



Training private ML models is challenging!



Differentially Private Learning Needs Better Features (or Much More Data). Tramèr & Boneh. 2021

Solution? Leverage public data!

Deep Learning with Differential Privacy

October 25, 2016

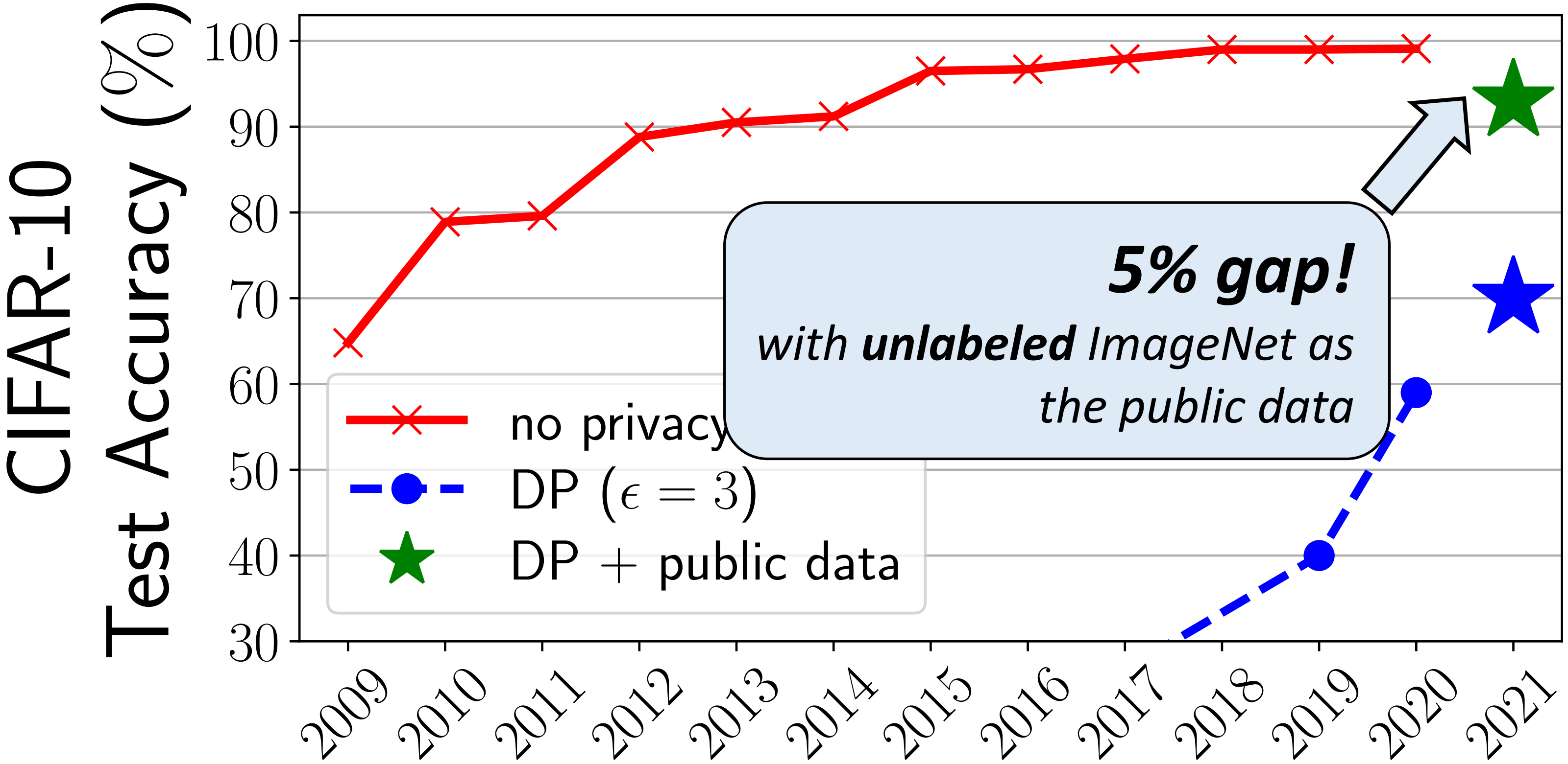
Martín Abadi*
H. Brendan McMahan*
Andy Chu*
Ilya Mironov*
Li Zhang*
Ian Goodfellow†
Kunal Talwar*

We treat the CIFAR-100 dataset as a public dataset and use it to train a network with the same architecture.

CIFAR10	Transfer + SGD (not private)	75%	∞	0	-
	Transfer + DP-SGD (Abadi et al.)	67%	2	10^{-5}	Public Data
	Transfer + DP-SGD (ours)	**72%**	**2.1**	10^{-5}	Public Data

Making the Shoe Fit: Architectures, Initializations, and Tuning for Learning with Privacy. Papernot et al. 2019

Moar public data!



Even moar public data!

LARGE LANGUAGE MODELS CAN BE STRONG DIFFERENTIALLY PRIVATE LEARNERS

Xuechen Li[1], **Florian Tramèr**[2], **Percy Liang**[1], **Tatsunori Ha**
[1]Stanford University [2]Google Research

Differentially Private Fine-tuning of Language Models*

Da Yu† Saurabh Naik ‡ Arturs Backurs § Sivakanth Gopi§ Huseyin A. Inan§
an Kulkarni§ Yin Tat Lee‖ Andre Manoel§
Sergey Yekhanin§ Huishuai Zhang**

Unlocking High-Accuracy Differentially Private Image Classification through Scale

CAN FOUNDATION MODELS HELP US ACHIEVE PERFECT SECRECY?

Simran Arora
Stanford University
Stanford, CA
simran@cs.stanford.edu

Christopher Ré
Stanford University
Stanford, CA
chrismre@cs.stanford.edu

rge Scale Transfer Learning for Differentially Private Image Classification

Harsh Mehta
Google Research
harshm@google.com

Abhradeep Thakurta
Google Research
athakurta@google.com

Alexey Kurakin
Google Research
kurakin@google.com

Ashok Cutkosky
Boston University
ashok@cutkosky.com

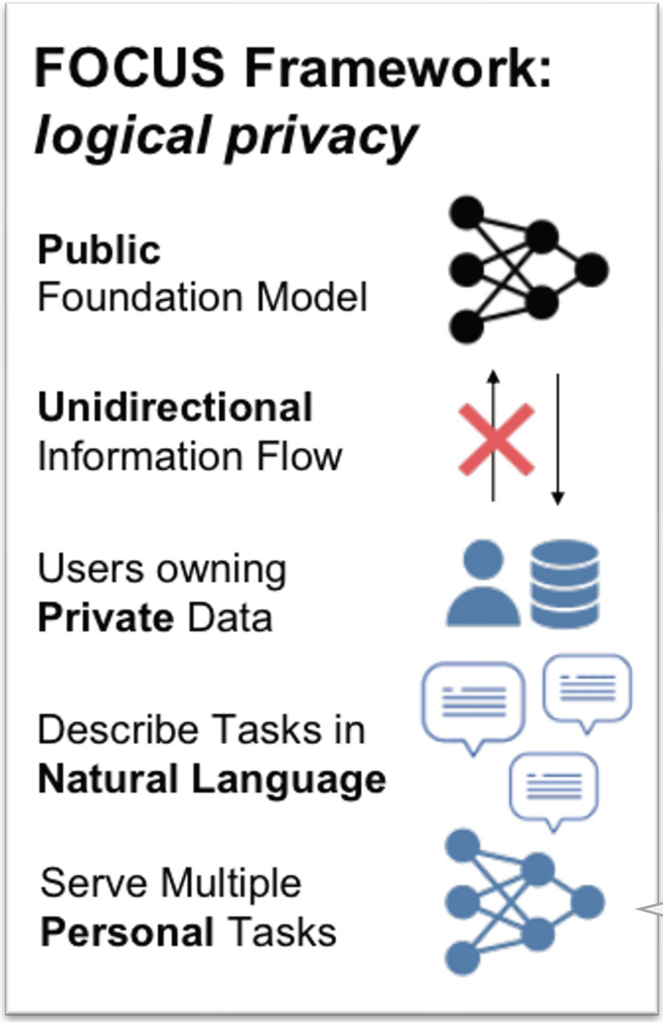
The nirvana: zero-shot privacy.

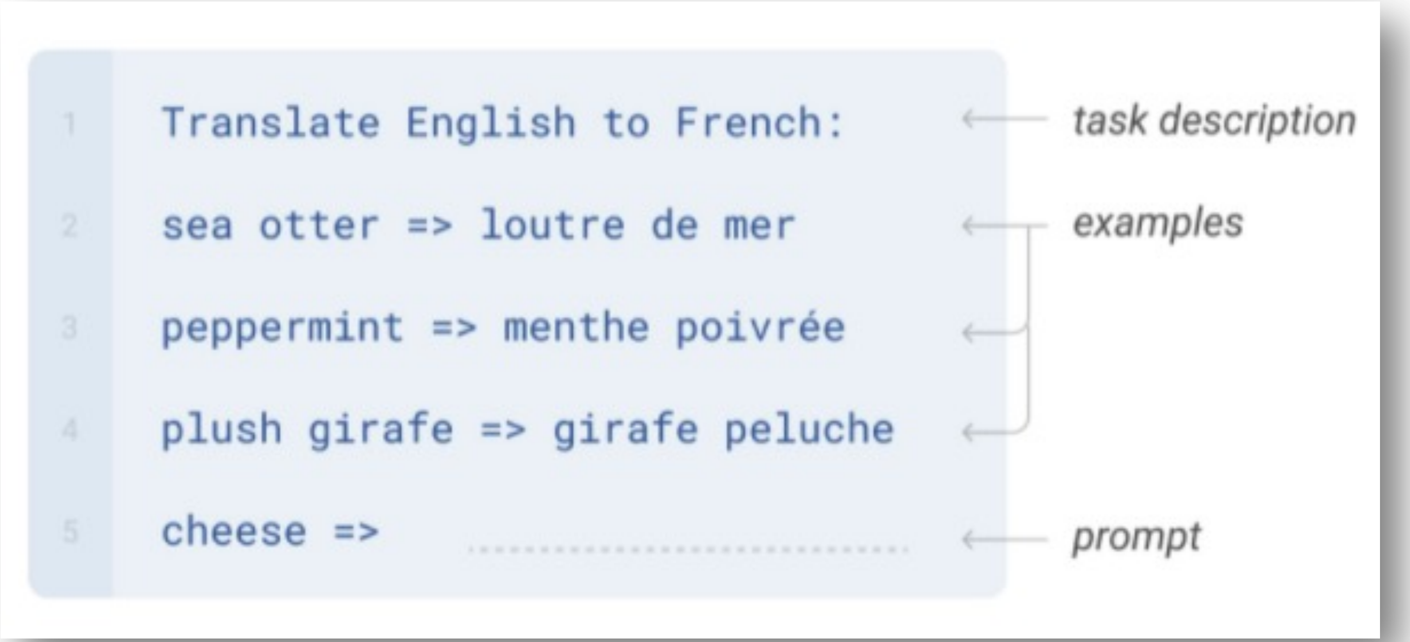
CAN FOUNDATION MODELS HELP US ACHIEVE PERFECT SECRECY?

Simran Arora
Stanford University
Stanford, CA
simran@cs.stanford.edu

Christopher Ré
Stanford University
Stanford, CA
chrismre@cs.stanford.edu

The nirvana: zero-shot privacy.





DP for $\varepsilon = 0$!!!

Zero-shot learning "solves" many "privacy benchmarks"!

- **CIFAR-10:** 97% zero-shot acc with OpenCLIP (LAION pretraining)
- **ImageNet:** 88.8% zero-shot acc with JFT pretraining

Near-SOTA accuracy with *perfect* privacy!

Two (possible) issues for private learning.

1. Is public pre-training *cheating*?

2. Does public pre-training *work*?

Two (possible) issues for private learning.

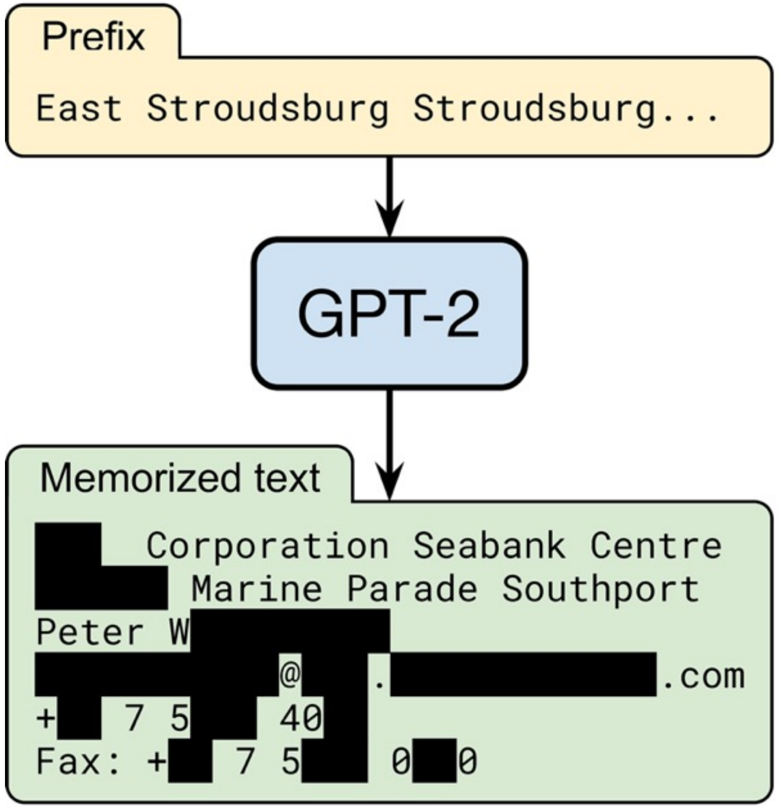
1. Is public pre-training *cheating*?

2. Does public pre-training *work*?

Does public pretraining still preserve "privacy"?

$$\frac{\Pr[\mathrm{Train}(D_{pretrain} + D_{private}) = \quad]}{\Pr[\mathrm{Train}(D_{pretrain} + D'_{private}) = \quad]} \leq e^{\varepsilon}$$

But is this “privacy preserving”?



Two (possible) issues for private learning.

1. Is public pre-training *cheating*?

2. Does public pre-training *work*?

A little secret...

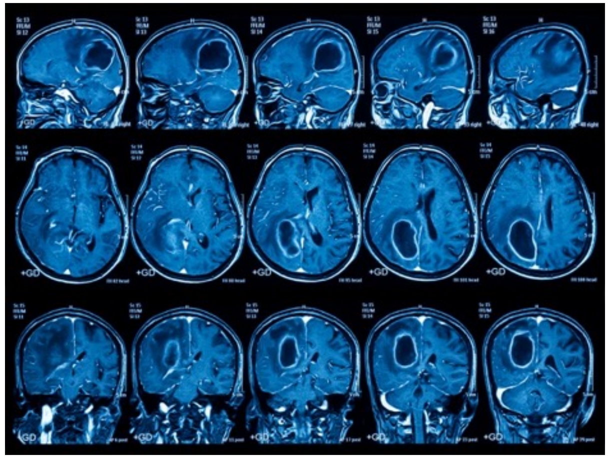
No one cares about CIFAR-10 or ImageNet!

What makes a good benchmark?

- The benchmark is a *proxy for a general task* we care about (e.g. image classification)

- Progress on the benchmark is (somewhat) *predictive* of performance on the general task

What tasks do we really care about solving with privacy?



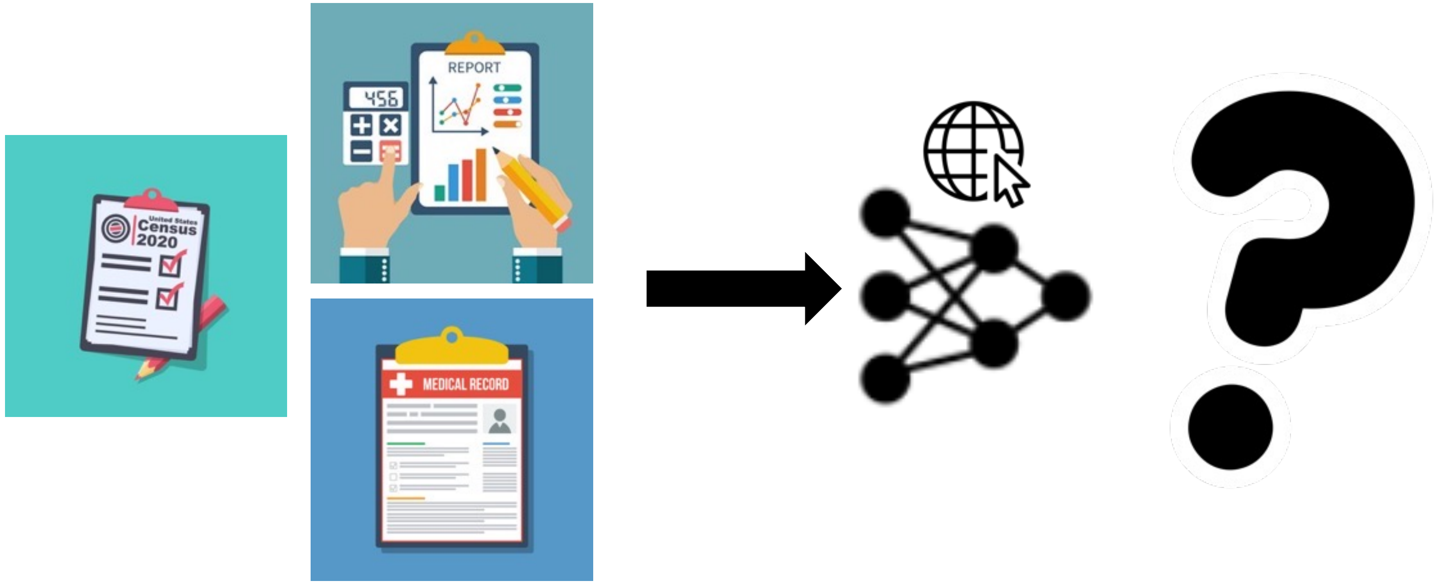
Are current benchmarks at least tracking *algorithmic progress* on private learning?

- Tasks we care about solving privately are (by definition) less likely to be represented on the Internet

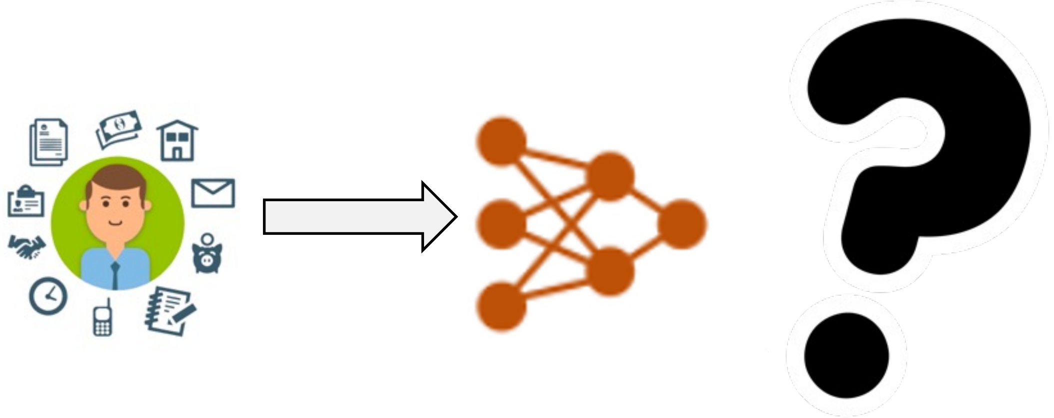
- Recent improvements on “private” benchmarks seem mainly due to generic improvements in zero-shot learning

Open problems

How good is public pretraining for sensitive data that is not well represented on the Web?



How good would public pretraining be if we removed all sensitive data?



Outlook

- Should Internet data be free game for "privacy-preserving" ML?
- How useful is public pretraining on highly sensitive data?
- Would public pretraining on non-sensitive data be as useful?
- We need better privacy benchmarks to answer these questions!