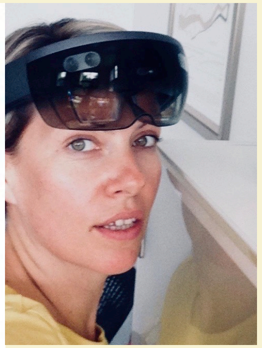
CS 229br: Foundations of Deep Learning

Lecture 4: Privacy

Boaz Barak

Gustaf Ahdritz Gal Kaplun Zona Kostic





Unofficial TF

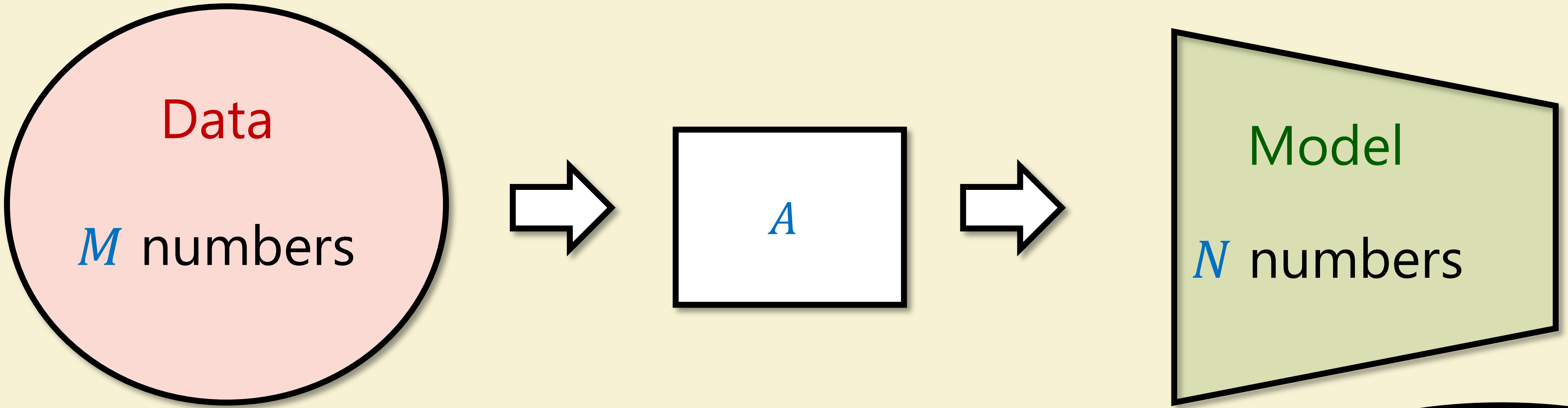
Coming up: Pset 2 = project proposal

Groups up to 3. Proposal will have three component:

- Proposal
- Summary of a recent related paper(s) & why it doesn’t answer question
- Notebook with some toy examples

More details soon

Learning



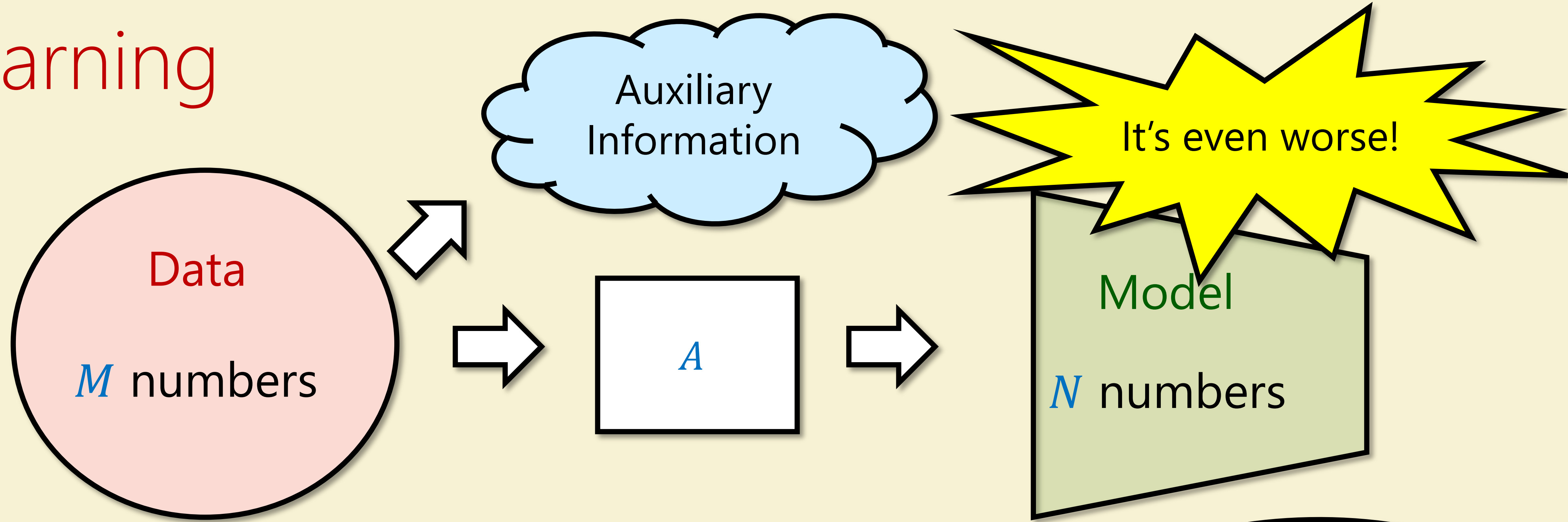
N equations in M unknowns

By Murphy's law: the most sensitive ones

Intuitively: If $N \gtrsim M$ may be able to recover ≈all of the data

Even if $N \ll M$ can recover $\approx N$ bits of the data

Learning



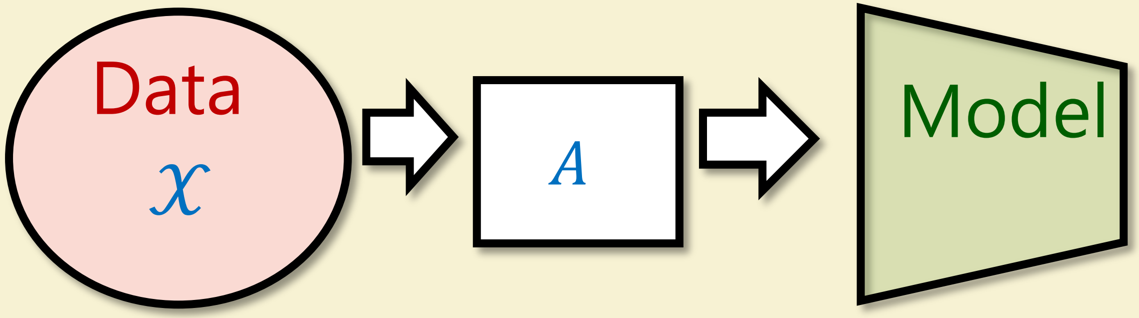
N equations in M unknowns

By Murphy's law: the most sensitive ones

Intuitively: If $N \gtrsim M$ may be able to recover ≈all of the data

Even if $N \ll M$ can recover $\approx N$ bits of the data

What's Memorization?

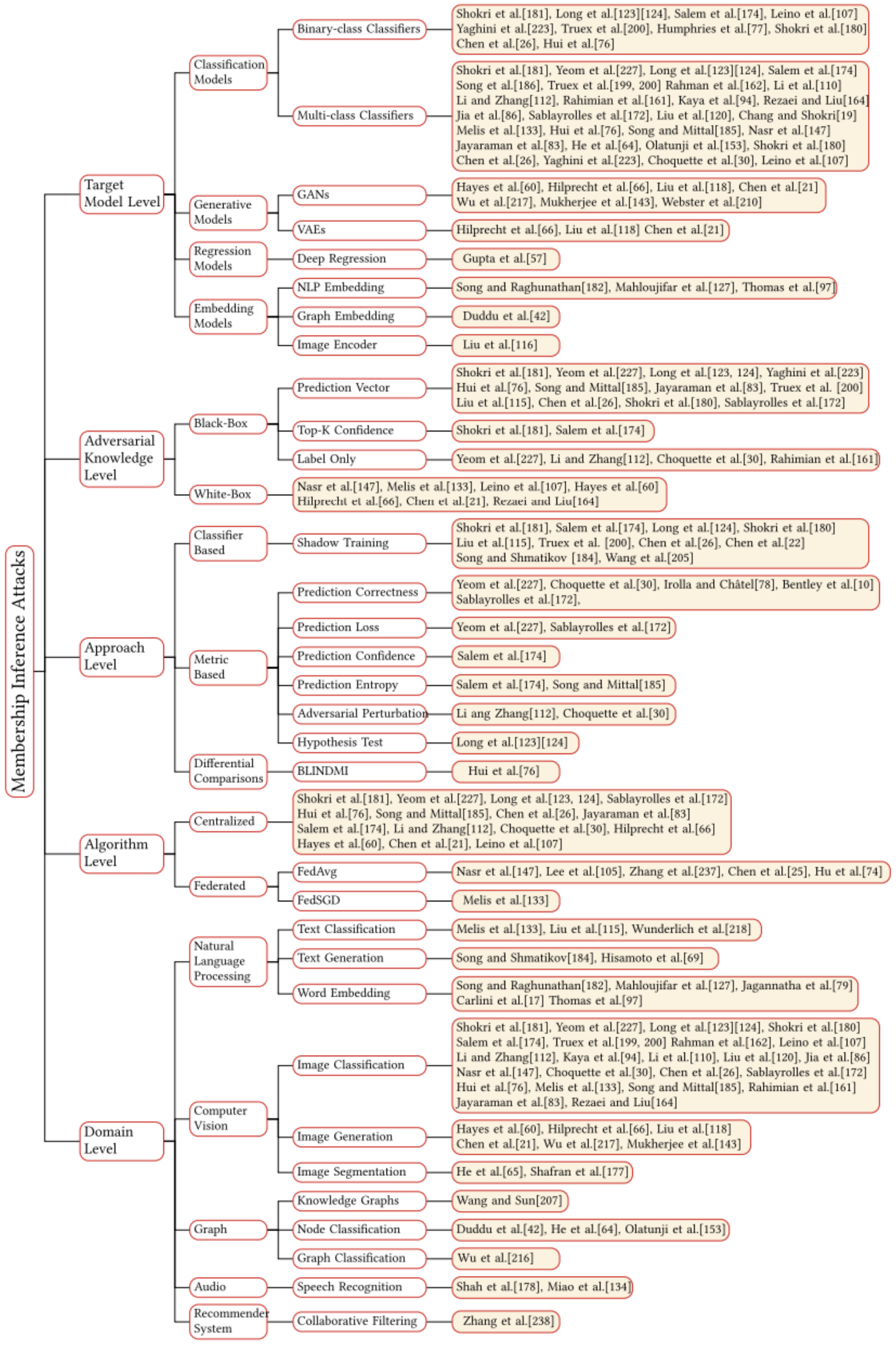


Adversary Goal

- Recover training sample(s) $x \in \mathcal{X}$
- Given x find out whether or not $x \in \mathcal{X}$

Adversary Access

- Auxiliary information about x
- Full description (i.e., weights) of model
- q black box queries to model



Membership Inference Attacks on Machine Learning: A Survey

HONGSHENG HU and ZORAN SALCIC, The University of Auckland, New Zealand
LICHAO SUN, Lehigh University, USA
GILLIAN DOBBIE, The University of Auckland, New Zealand
PHILIP S. YU, University of Illinois at Chicago, USA
XUYUN ZHANG, Macquarie University, Australia

— CHAPTER ONE —

The Boy Who Lived

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs Potter was Mrs Dursley's sister, but they hadn't met for several years; in fact, Mrs Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbours would say if the Potters arrived in the street. The Dursleys knew that the Potters had a

Playground

Load a preset... | Save | View code | Share | ···

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs. Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would

Mode

Model

text-davinci-003

Temperature 0.7

Maximum length 256

Stop sequences

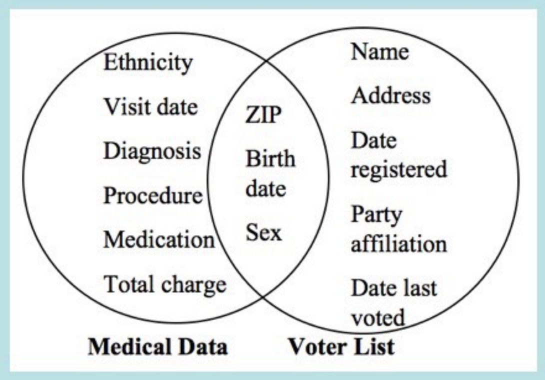
Enter sequence and press Tab

Top P 1

Simple Demographics Often Identify People Uniquely

Latanya Sweeney
Carnegie Mellon University
latanya@andrew.cmu.edu

2000



1997

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

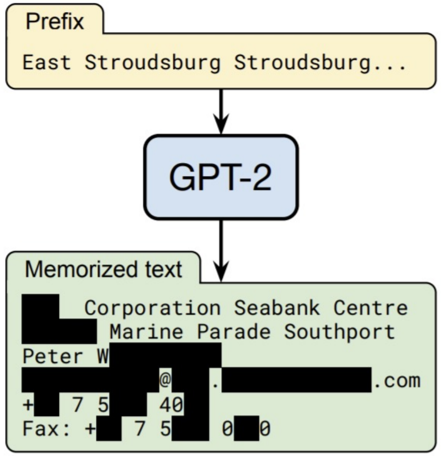
2008

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Fredrikson, Jha, Ristenpart 2015

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	✓

Table 2: Summary of results on the Enron email dataset. Three secrets are extractable in < 1 hour; all are heavily memorized.



Training Set | **Generated Image**

Caption: Living in the light with Ann Graham Lotz

Prompt: Ann Graham Lotz

Carlini et al (2019,2020,2023)

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

2019

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

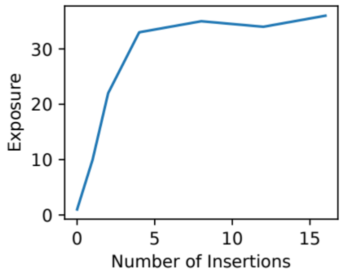


Figure 6: Exposure of a canary inserted in a Neural Machine Translation model. When the canary is inserted four times or more, it is fully memorized.

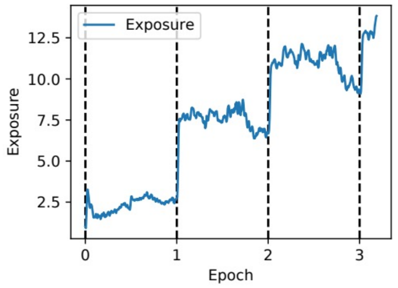


Figure 7: Exposure as a function of training time. The exposure spikes after the first mini-batch of each epoch (which contains the artificially inserted canary), and then falls overall during the mini-batches that do not contain it.

Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models

2022

Kushal Tirumala* Aram H. Markosyan* Luke Zettlemoyer Armen Aghajanyan

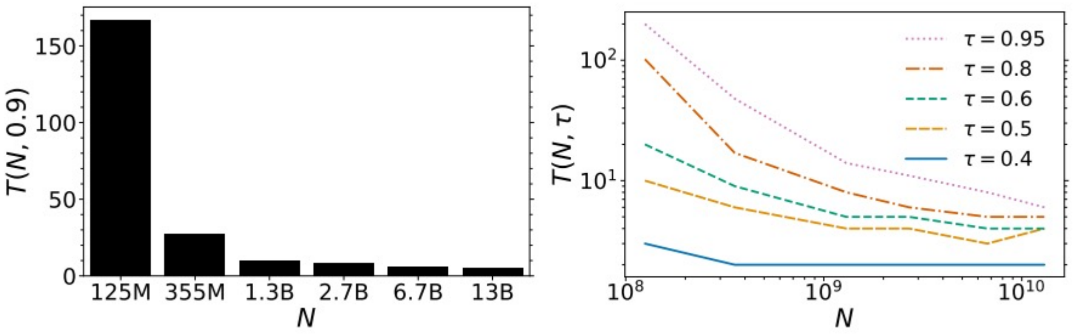


Figure 1: We show $T(N, \tau)$, which is the number of times a language model needs to see each training example before memorizing τ fraction of the training data, as a function of model size N. Result are for causal language modeling on WIKITEXT103, right plot is on log-log scale. Note that generally larger models memorize faster, regardless of τ.

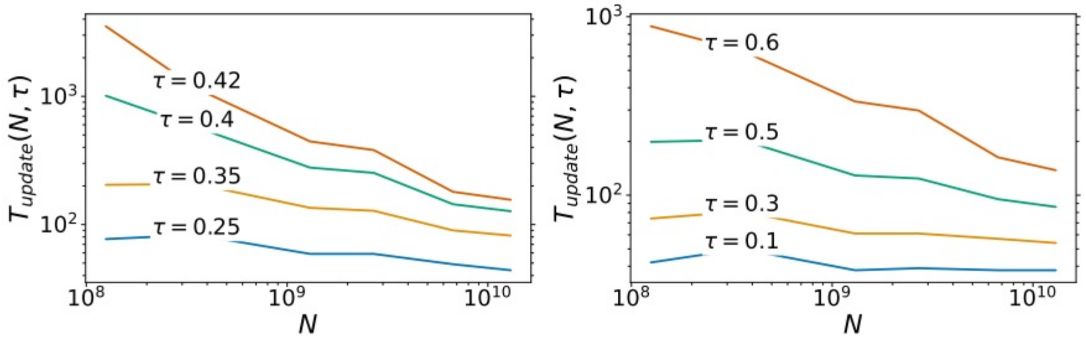
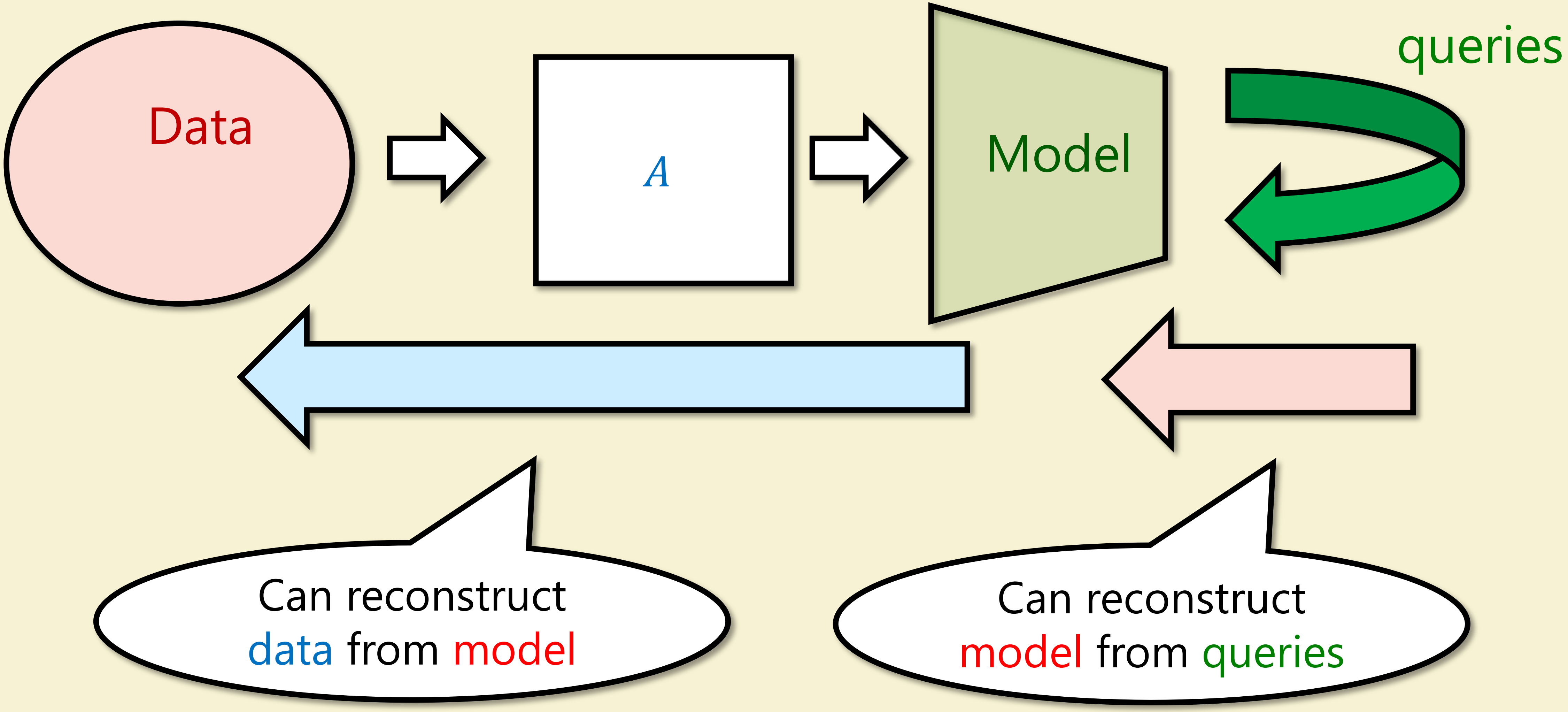
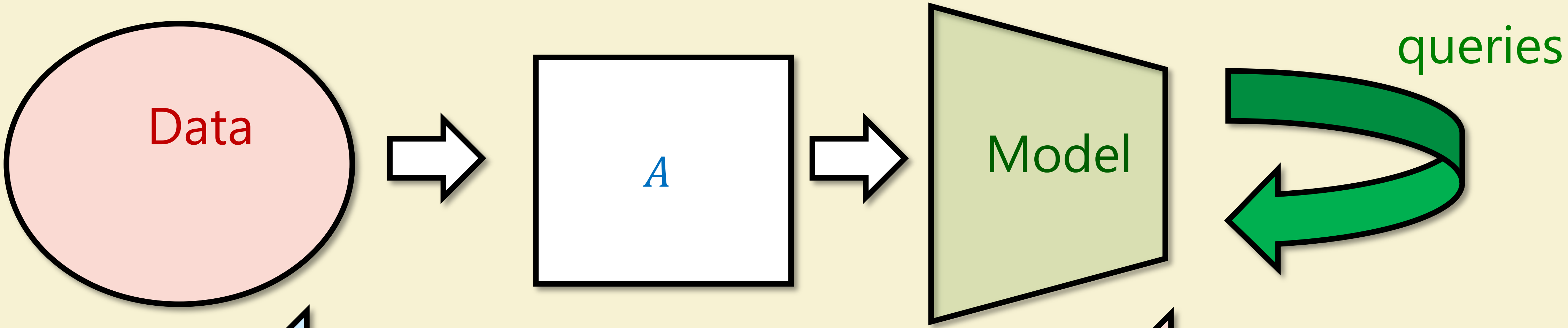


Figure 3: We show $T_{update}(N, \tau)$, which is the number of gradient descent updates U a language model needs to perform before memorizing τ fraction of the data given on the U'th update, as a function of model size N. Result are for causal (Left) and masked (Right) language modeling on the ROBERTA dataset, on a log-log scale. We show that larger models memorize faster, regardless of τ.

Inference attacks



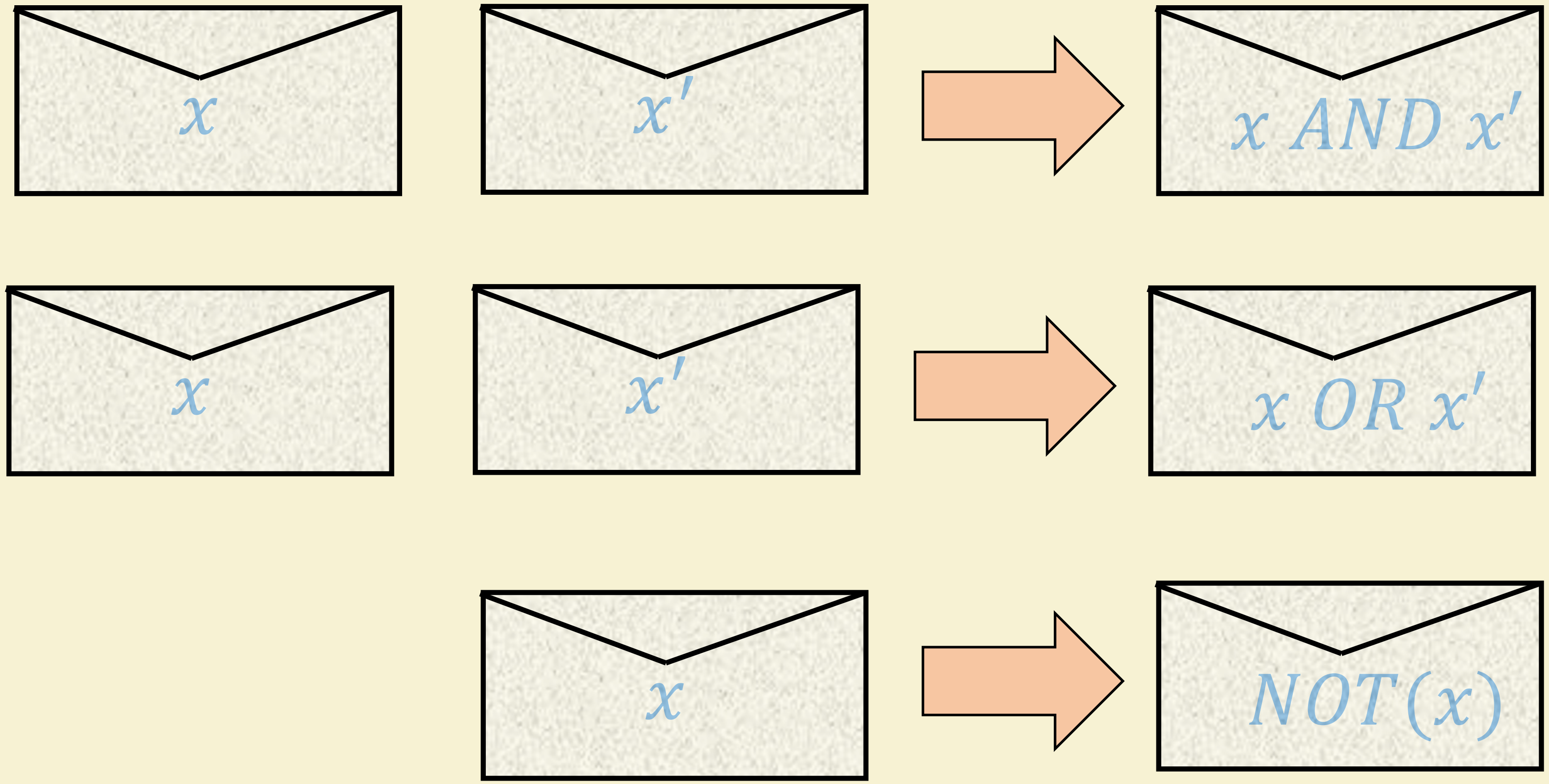
Inference attacks



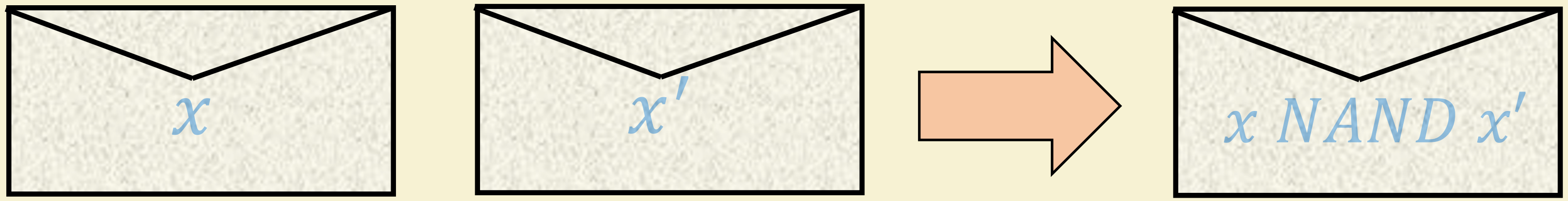
Solutions:

- Cryptographic: 100% privacy for model, but efficiency cost, and doesn't help if release outputs.
- Differential privacy: "X% privacy" but X vs utility tradeoff not great
- Heuristics: Hope for 100%, might get 0%

Fully Homomorphic Encryption (FHE)



Fully Homomorphic Encryption (FHE)



FHE

Secret key: $k \sim \{0,1\}^n$

Secret key
Plaintext
Ciphertext

Encryption: randomized E: $\{0,1\}^n \times \{0,1\} \rightarrow \{0,1\}^m$

Decryption: D: $\{0,1\}^n \times \{0,1\}^m \rightarrow \{0,1\}$

Does not get secret key!

Evaluation: randomized $NAND$: $\{0,1\}^m \times \{0,1\}^m \rightarrow \{0,1\}^m$

* Can also consider public key variant

FHE

Secret key: $k \sim \{0,1\}^n$

Encryption: randomized $E: \{0,1\}^n \times \{0,1\} \to \{0,1\}^m$

Decryption: $D: \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$

Evaluation: randomized $NAND: \{0,1\}^m \times \{0,1\}^m \to \{0,1\}^m$

Correctness: $\forall_k \forall_{b \in \{0,1\}}, D_k\big(E_k(b)\big) = b$

$\Delta_{TV} < \exp(-n)$

Evaluation: $\forall_k \forall_{b,b' \in \{0,1\}}, NAND\big(E_k(b), E_k(b')\big) \equiv E_k(\neg(b \wedge b'))$

Computational secrecy*: $\forall$ alg A of time $\ll \exp(n)$

Can't distinguish between $E_k(0)$ and $E_k(1)$

$$\Pr_{\substack{b \sim \{0,1\} \\ k \sim \{0,1\}^n}}\left[A\big(E_k(b)\big) = b\right] \leq \frac{1}{2} + \exp(-n)$$

* Even if we get $\exp(n)$ samples with same key

FHE: What's known

Gentry 2009: FHE exists under reasonable assumptions

… FHE exists under standard assumptions

… implementations

HElib

build passing

HElib is an open-source (Apache License v2.0) software library that implements homomorphic encryption (HE). Currently available schemes are the implementations of the Brakerski-Gentry-Vaikuntanathan (BGV) scheme with bootstrapping and the Approximate Number scheme of Cheon-Kim-Kim-Song (CKKS), along with many optimizations to make homomorphic evaluation run faster, focusing mostly on effective use of the Smart-Vercauteren ciphertext packing techniques and the Gentry-Halevi-Smart optimizations. See this report for a description of a few of the algorithms using in this library.

Please refer to CKKS-security.md for the latest discussion on the security of the CKKS scheme implementation in HElib.

Since mid-2018 HElib has been under extensive refactoring for *Reliability, Robustness & Serviceability, Performance*, and most importantly *Usability* for researchers and developers working on HE and its uses.

HElib supports an "*assembly language for HE*", providing low-level routines (set, add, multiply, shift, etc.), sophisticated automatic noise management, improved BGV bootstrapping, multi-threading, and also support for Ptxt (plaintext) objects which mimics the functionality of Ctxt (ciphertext) objects. The report Design and implementation of HElib contains additional details. Also, see CHANGES.md for more information on the HElib releases.

OpenFHE

Community Growth:

OpenFHE is an open-source project that provides efficient extensible implementations of the leading post-quantum Fully Homomorphic Encryption (FHE) schemes.

Microsoft SEAL

Microsoft SEAL is an easy-to-use open-source (MIT licensed) homomorphic encryption library developed by the Cryptography and Privacy Research Group at Microsoft. Microsoft SEAL is written in modern standard C++ and is easy to compile and run in many different environments. For more information about the Microsoft SEAL project, see sealcrypto.org.

What is FHE good for?

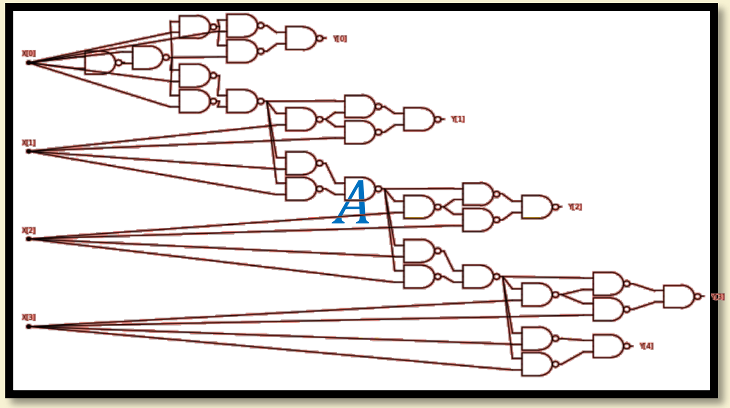
Can also use MPC

Encryption: randomized E: $\{0,1\}^n \times \{0,1\} \to \{0,1\}^m$

Decryption: D: $\{0,1\}^n \times \{0,1\}^m \to \{0,1\}$

Evaluation: randomized $NAND$: $\{0,1\}^m \times \{0,1\}^m \to \{0,1\}^m$

Data $\Rightarrow$



$\Rightarrow$ h

$E_k(x)$ $\Rightarrow$ Many $NAND$s $\Rightarrow$ $E_k(h)$

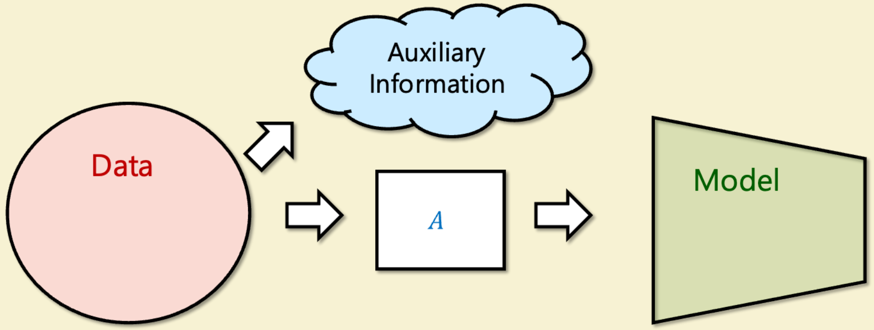
Challenges: Only get *encrypted* model/summary

Huge computational overhead

(Matrix vector mult on <1000 dimensions takes few secs on 32 core 250GB PC)

[Halevi, Shoup 2018]

Differential Privacy



"You will not be affected, adversely or otherwise, by allowing your data to be used in [a DP protected] study or analysis, no matter what other studies, data sets, or information sources, are available."

Dwork and Roth

Differential Privacy

ANDY GREENBERG SECURITY 06.13.2016 07:02 PM

Apple's 'Differential Privacy' Is About Collecting Your Data---But Not *Your* Data

At WWDC, Apple name-checked the statistical science of learning as much as possible about a group while learning as little as possible about any individual in it.

Trustworthy ML Home Symposium Thursday Series Recordings Resources M

TRUSTWORTHY ML INITIATIVE

New differential privacy platform co-developed with Harvard's OpenDP unlocks data while safeguarding privacy

Jun 24, 2020 | John Kahan - VP, Chief Data Analytics Officer

OpenDP About Opportunities Community

Developing Open Source Tools for Differential Privacy

OpenDP is a community effort to build trustworthy, open-source software tools for statistical analysis of sensitive private data. These tools, which we call OpenDP, will offer the rigorous protections of differential privacy for the individuals who may be represented in confidential data and statistically valid methods of analysis for researchers who study the data.

Opacus

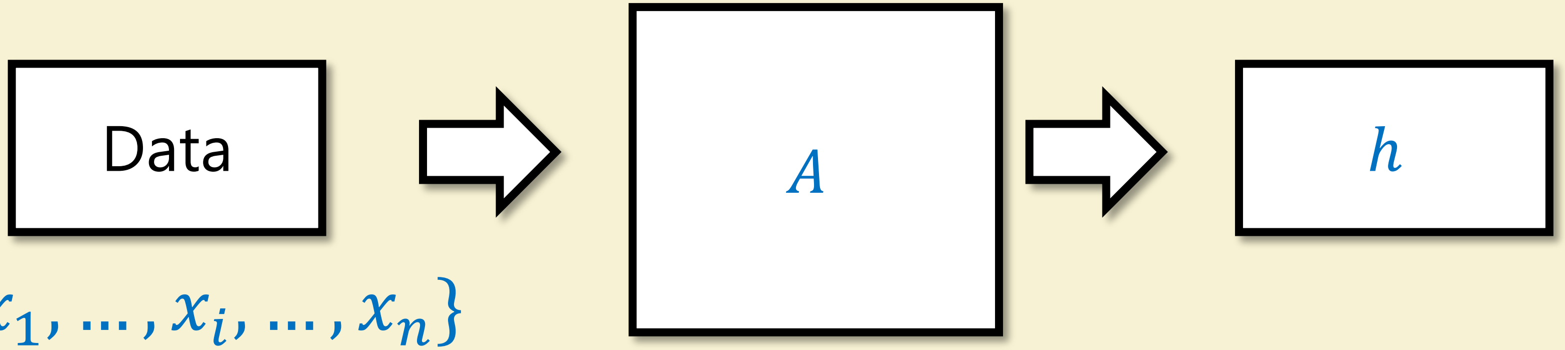
Train PyTorch models with Differential Privacy

Google releases differential privacy tools to commemorate Data Privacy Day

f in

Machine Learning with Differential Priva in TensorFlow

Differential Privacy



$\mathcal{X} = \{x_1, \dots, x_i, \dots, x_n\}$

Data belonging to i-th person

Def: A is ϵ *differentially private* if

posterior probability of $x_i \in \mathcal{X} \in e^{\pm\epsilon} \times$ prior probability of $x_i \in \mathcal{X}$

$\forall \mathcal{X}, \mathcal{X}'$ s.t. $|\mathcal{X} \triangle \mathcal{X}'| = 1$, $\forall h$

A must be randomized

$$\Pr[A(\mathcal{X}) = h] \in e^{\pm\epsilon} \Pr[A(\mathcal{X}') = h]$$

Differential Privacy

Data → A → h

$\mathcal{X} = \{x_1, \dots, x_i, \dots, x_n\}$

δ

Def: A is ϵ *differentially private* if

$\forall \mathcal{X}, \mathcal{X}'$ s.t. $|\mathcal{X} \triangle \mathcal{X}'| = 1$, $\forall S$

$$\Pr[A(\mathcal{X}) \in S] \le e^{\epsilon} \Pr[A(\mathcal{X}') \in S] + \delta$$

$\delta \ll \epsilon$
Think $\delta = 0$

Differential Privacy

Data $\rightarrow$ A $\rightarrow$ h

$\mathcal{X} = \{ x_1, \dots, x_i, \dots, x_n \}$

Def: A is ϵ *differentially private* if

$\forall \mathcal{X}, \mathcal{X}'$ s.t. $|\mathcal{X} \triangle \mathcal{X}'| = 1$, $\forall S$

$$\Pr[A(\mathcal{X}) \in S] \in e^{\pm\epsilon} \Pr[A(\mathcal{X}') \in S]$$

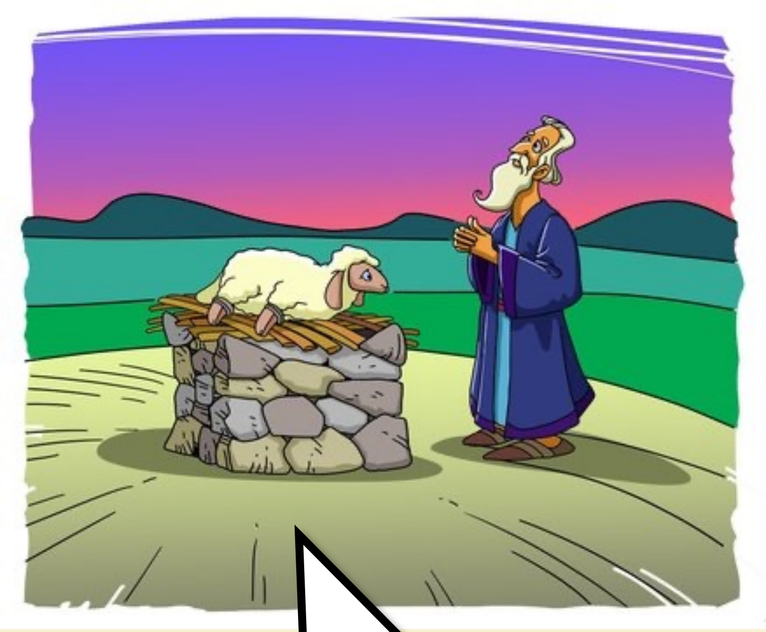
$\Pr[$ Bad event happened to i because their data in $\mathcal{X}$ $] \leq e^{\epsilon} \cdot \Pr[$ Bad event happens anyway $]$

Example: $A(\mathcal{X})$ reveals short people more likely to default on loans

Differential Privacy

Data $\Rightarrow$ A $\Rightarrow$ h

$\mathcal{X} = \{x_1, \ldots, x_i, \ldots, x_n\}$



Subset $i_1 .. i_k$ is "sacrificial lamb"

Def: A is ϵ *differentially private* if

$\forall \mathcal{X}, \mathcal{X}'$ s.t. $|\mathcal{X} \vartriangle \mathcal{X}'| = 1$, $\forall S$

$$\Pr[A(\mathcal{X}) \in S] \in e^{\pm\epsilon} \Pr[A(\mathcal{X}') \in S]$$

Why not $\Pr[A(\mathcal{X}) \in S] \in \Pr[A(\mathcal{X}') \in S] \pm \epsilon$?

Think: $A(\mathcal{X}) = \{x_{i_1}, \ldots, x_{i_k}\}$ random $i_1, \ldots, i_k$, $k \ll n$

$$|\Pr[A(\mathcal{X}) \in S] - \Pr[A(\mathcal{X}') \in S]| \le \frac{k}{n}$$

Differentially private statistics:

Publish estimates $\hat{f}_1 \approx \sum_{x \in \mathcal{X}} f_1(x)$, … , $\hat{f}_k \approx \sum_{x \in \mathcal{X}} f_k(x)$

In differentially private way

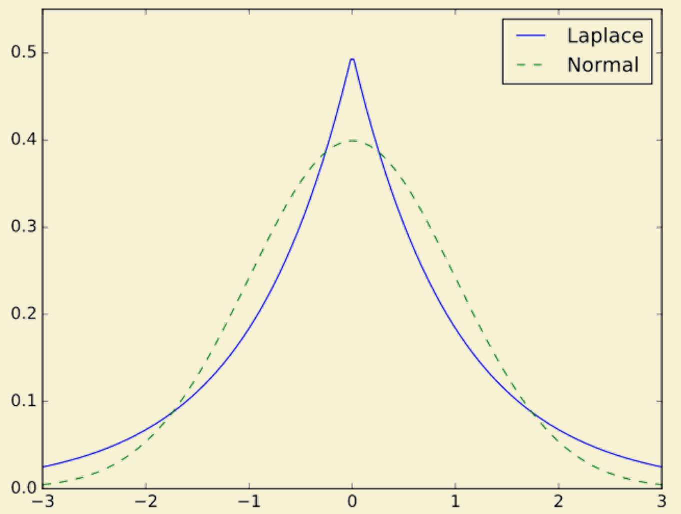
Why can't we just publish sums?

- 40 CS229br students passed pset zero
- 39 CS229br students passed pset zero & not named Costis

Differentially private statistics:

Publish estimates $\hat{f}_1 \approx \sum_{x \in \mathcal{X}} f_1(x)$, … , $\hat{f}_k \approx \sum_{x \in \mathcal{X}} f_k(x)$

In differentially private way



Laplace mechanism:

Assume $f_i(x) \in [0,1]$

$$\hat{f}_i = \sum_{x \in \mathcal{X}} f_i(x) + \mathrm{Lap}(k/\epsilon)$$

THM: Laplace mechanism is ϵ-DP

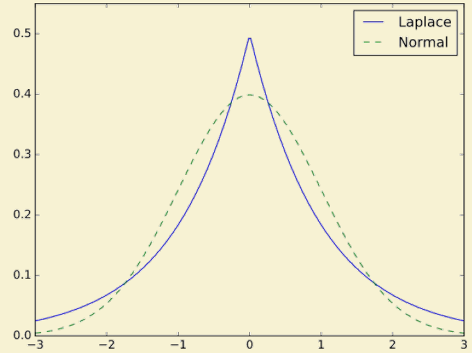
Symmetric exponential

$$\Pr[\mathrm{Lap}(b) = x] = \frac{1}{2b} \exp(-|x|/b)$$

$$\sigma^2 = 2b^2$$

Publish estimates $\hat{f}_1 \approx \sum_{x\in\mathcal{X}} f_1(x)$, … , $\hat{f}_k \approx \sum_{x\in\mathcal{X}} f_k(x)$

Assume $f_i(x) \in [0,1]$



$$\Pr[\mathrm{Lap}(b) = x] = \frac{1}{2b}\exp(-|x|/b)$$

$$\sigma^2 = 2b^2$$

Laplace mechanism:

$$\hat{f}_i = \sum_{x\in\mathcal{X}} f_i(x) + \mathrm{Lap}(k/\epsilon)$$

THM: Laplace mechanism is ϵ-DP

PF: Focus on single f

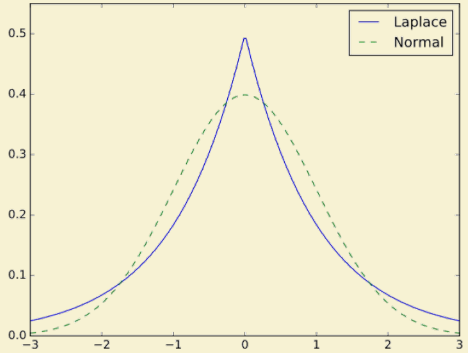
$$|f(\mathcal{X}) - f(\mathcal{X}')| \le 1$$

$$f(\mathcal{X}) := \sum_{x\in\mathcal{X}} f(x) \qquad f(\mathcal{X}') := \sum_{x\in\mathcal{X}'} f(x)$$

Proof on Board

Publish estimates $\hat{f}_1 \approx \sum_{x \in \mathcal{X}} f_1(x)$, … , $\hat{f}_k \approx \sum_{x \sim \mathcal{X}} f_k(x)$

Assume $f_i(x) \in [0,1]$



Laplace mechanism:

$$\hat{f}_i = \sum_{x \sim X} f_i(x) + \mathrm{Lap}(k/\epsilon)$$

$$\Pr[\mathrm{Lap}(b) = x] = \frac{1}{2b} \exp(-|x|/b)$$

$$\sigma^2 = 2b^2$$

$\boxed{\text{THM: Laplace mechanism is } \epsilon\text{-DP}}$

Generalization: Achieve ϵ-DP for std $\approx k/\epsilon$ estimator for any $f: \mathcal{X} \to \mathbb{R}^m$

s.t. $\underbrace{| f(\mathcal{X}) - f(\mathcal{X}')|_1}_{\text{Sensitivity of } f} \leq k$ for all $|\mathcal{X} \triangle \mathcal{X}'| = 1$

Generalization: Achieve ϵ-DP for std $\approx k/\epsilon$ estimator for any $f: \mathcal{X} \to \mathbb{R}^m$

s.t. $\underbrace{|f(\mathcal{X}) - f(\mathcal{X}')|_1}_{\text{Sensitivity of } f} \leq k$ for all $|\mathcal{X} \triangle \mathcal{X}'| = 1$

Generalization: Achieve ϵ-DP for std $\approx k/\epsilon$ estimator for any $f: \mathcal{X} \to \mathbb{R}^m$

Gaussian mechanism: Output $f(\mathcal{X}) + N(0, \sigma^2 I)$

δ $\sqrt{\log(1/\delta)}$

"Morally": Achieve ϵ-DP std $\approx k/\epsilon$ for any $f: \mathcal{X} \to \mathbb{R}^m$

s.t. $\|f(\mathcal{X}) - f(\mathcal{X}')\|_2 \leq k$ for all $|\mathcal{X} \triangle \mathcal{X}'| = 1$

Important

Differential privacy is definition

Adding noise is one approach to achieve definition

Differential privacy composition

Thm: If A is ϵ-DP and A' is ϵ'-DP then $B(\mathcal{X}) = A(\mathcal{X}), A'(\mathcal{X})$ is $\epsilon + \epsilon'$-DP

Proof: $\forall h, h'$ and $|\mathcal{X} \vartriangle \mathcal{X}'| \leq 1$

$$\Pr[A(\mathcal{X}), A'(\mathcal{X}) = (h, h')] \leq e^{\epsilon} \Pr[A(\mathcal{X}') = h] \cdot e^{\epsilon'} \Pr[A'(\mathcal{X}') = h'] \quad \blacksquare$$

Differential privacy under post-processing

Thm: If A is ϵ-DP and $B(\mathcal{X}) = f(A(\mathcal{X}))$ then $B(\mathcal{X})$ is ϵ-DP

Proof: $\forall h$ and $|\mathcal{X} \vartriangle \mathcal{X}'| \leq 1$

$$\Pr[f(A(\mathcal{X})) = h] = \sum_{h' \in f^{-1}(h)} \Pr[A(\mathcal{X}) = h'] \leq e^{\epsilon} \sum_{h' \in f^{-1}(h)} \Pr[A(\mathcal{X}') = h'] = e^{\epsilon} \Pr[f(A(\mathcal{X}')) = h] \quad \blacksquare$$

Advanced composition

Thm: If $A_1 \dots \mathrm{A_k}$ are ϵ-DP then $B(\mathcal{X}) = A_1(\mathcal{X}), \dots, A_k(\mathcal{X})$ is

1) $k\epsilon$-DP

2) $(\tilde{O}(\epsilon\sqrt{k}), o(1))$-DP

   More accurately: $O\left(\epsilon\sqrt{k\log(1/\delta)} + \epsilon^2 k\right), \delta$

Proof on Board

* Holds even if A_{i+1} depends on outputs of $A_1 \dots A_{i-1}$

DP-SGD

Deep Learning with Differential Privacy

October 25, 2016

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

On Board

Evaluation

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

Florian Tramèr
Stanford University
tramer@cs.stanford.edu

Dan Boneh
Stanford University
dabo@cs.stanford.edu

Abstract

We demonstrate that differentially private machine learning has not yet reached its "AlexNet moment" on many canonical vision tasks: linear models trained on handcrafted features significantly outperform end-to-end deep neural networks for moderate privacy budgets. To exceed the performance of handcrafted features, we show that private learning requires either much more private data, or access to features learned on public data from a similar domain. Our work introduces simple yet strong baselines for differentially private learning that can inform the evaluation of future progress in this area.

Data	ε-DP	Source	CNN
MNIST	1.2	Feldman & Zrnic (2020)	96.6
	2.0	Abadi et al. (2016)	95.0
	2.32	Bu et al. (2019)	96.6
	2.5	Chen & Lee (2020)	90.0
	2.93	Papernot et al. (2020a)	98.1
	3.2	Nasr et al. (2020)	96.1
	6.78	Yu et al. (2019b)	93.2
Fashion-MNIST	2.7	Papernot et al. (2020a)	86.1
	3.0	Chen & Lee (2020)	82.3
CIFAR-10	3.0	Nasr et al. (2020)	55.0
	6.78	Yu et al. (2019b)	44.3
	7.53	Papernot et al. (2020a)	66.2
	8.0	Chen & Lee (2020)	53.0

Protection from memorization in practice

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

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

	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	✓

	Naïve Composition				Advanced Composition				zCDP				RDP			
ϵ	Loss	1%	2%	5%	Loss	1%	2%	5%	Loss	1%	2%	5%	Loss	1%	2%	5%
0.01	.94	0	0	0	.94	0	0	0	.93	0	0	0	.94	0	0	0
0.05	.94	0	0	0	.93	0	0	0	.94	0	0	0	.94	0	0	0
0.1	.94	0	0	0	.93	0	0	0	.94	0	0	0	.93	0	0	0
0.5	.95	0	0	0	.93	0	0	0	.94	0	0	0	.92	0	0	0
1.0	.94	0	0	0	.94	0	0	0	.92	0	0	0	.94	0	0	0
5.0	.94	0	0	0	.94	0	0	0	.94	0	0	0	.65	11	24	79
10.0	.94	0	0	0	.93	0	0	0	.91	0	0	2	.53	9	33	108
50.0	.94	0	0	0	.94	0	0	0	.64	2	12	65	.35	28	65	185
100.0	.91	0	0	0	.93	0	0	0	.52	13	31	98	.32	21	67	205
500.0	.54	3	21	58	.79	4	7	31	.28	8	41	210	.27	5	54	278
1,000.0	.36	20	48	131	.71	8	16	74	.22	12	42	211	.24	10	37	269

Table 7: Number of members (out of 10,000) exposed by Yeom et al. membership inference attack on neural network (CIFAR-100). The non-private ($\epsilon = \infty$) model leaks 0, 556 and 7349 members for 1%, 2% and 5% FPR respectively.

Jayaraman and Evans 19

Private aggregation of teacher ensembles

SEMI-SUPERVISED KNOWLEDGE TRANSFER FOR DEEP LEARNING FROM PRIVATE TRAINING DATA

Nicolas Papernot*
Pennsylvania State University
ngp5056@cse.psu.edu

Martín Abadi
Google Brain
abadi@google.com

Úlfar Erlingsson
Google
ulfar@google.com

Dataset	ε	δ	Queries	Non-Private Baseline	Student Accuracy
MNIST	2.04	10^{-5}	100	99.18%	98.00%
MNIST	8.03	10^{-5}	1000	99.18%	98.10%
SVHN	5.04	10^{-6}	500	92.80%	82.72%
SVHN	8.19	10^{-6}	1000	92.80%	90.66%

Figure 4: **Utility and privacy of the semi-supervised students:** each row is a variant of the student model trained with generative adversarial networks in a semi-supervised way, with a different number of label queries made to the teachers through the noisy aggregation mechanism. The last column reports the accuracy of the student and the second and third column the bound ε and failure probability δ of the (ε, δ) differential privacy guarantee.

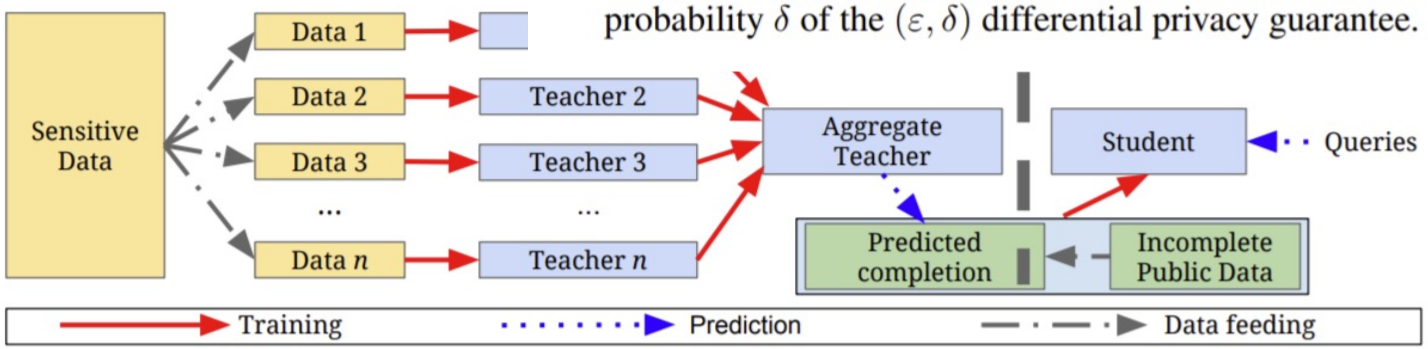


Figure 2: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

Scalable Private Learning with PATE

Nicolas Papernot*
Pennsylvania State University
ngp5056@cse.psu.edu

Shuang Song*
University of California San Diego
shs037@eng.ucsd.edu

Ilya Mironov, Ananth Raghunathan, Kunal Talwar & Úlfar Erlingsson
Google Brain
{mironov,pseudorandom,kunal,ulfar}@google.com

Dataset	Aggregator	Queries answered	Privacy bound ε	Accuracy Student	Accuracy Baseline
MNIST	LNMax (Papernot et al., 2017)	100	2.04	98.0%	99.2%
	LNMax (Papernot et al., 2017)	1,000	8.03	98.1%	
	Confident-GNMax (T=200, σ_1=150, σ_2=40)	286	**1.97**	**98.5%**	
SVHN	LNMax (Papernot et al., 2017)	500	5.04	82.7%	92.8%
	LNMax (Papernot et al., 2017)	1,000	8.19	90.7%	
	Confident-GNMax (T=300, σ_1=200, σ_2=40)	3,098	**4.96**	**91.6%**	
Adult	LNMax (Papernot et al., 2017)	500	2.66	83.0%	85.0%
	Confident-GNMax (T=300, σ_1=200, σ_2=40)	524	**1.90**	**83.7%**	
Glyph	LNMax	4,000	4.3	72.4%	82.2%
	Confident-GNMax (T=1000, σ_1=500, σ_2=100)	10,762	2.03	**75.5%**	
	Interactive-GNMax, two rounds	4,341	**0.837**	73.2%	

Heuristics

Avoid DP issues:

- Accuracy hit
- Large values for ϵ
- Slower

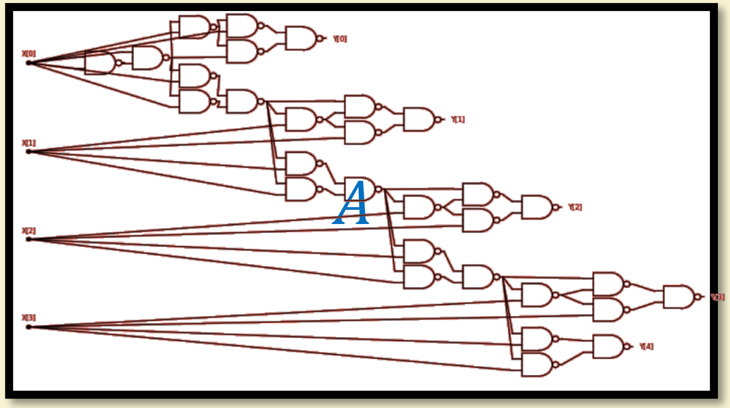
InstaHide

InstaHide: Instance-hiding Schemes for Private Distributed Learning*

Yangsibo Huang[†] Zhao Song[‡] Kai Li[§] Sanjeev Arora[¶]

Recall FHE-based training:

$x_1, \dots, x_n$



h

$E_k(x_1) \dots E_k(x_n)$

Many $NAND$s

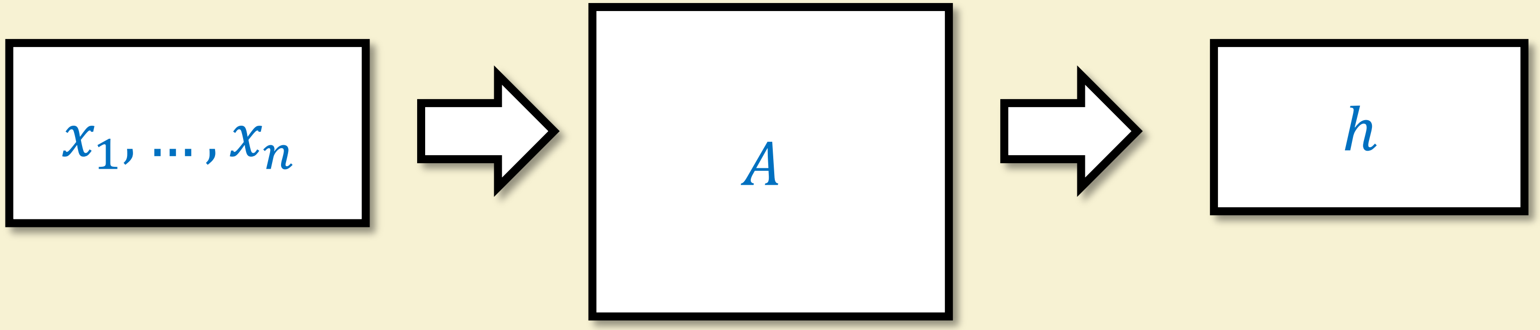
$E_k(h)$

Challenges:

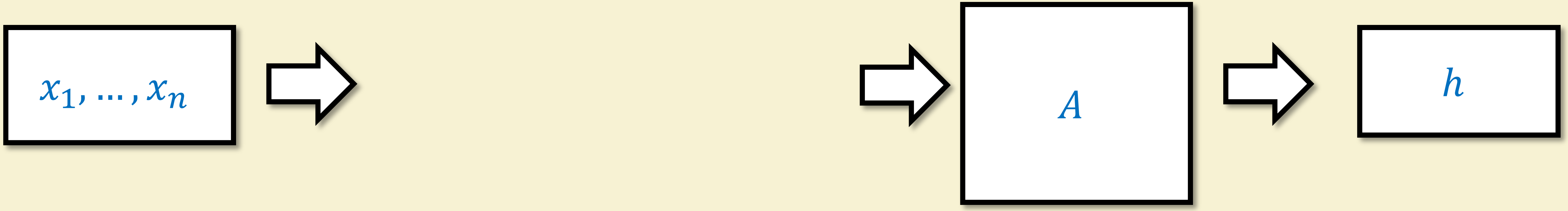
Only get *encrypted* model/summary

Huge computational overhead

InstaHide



InstaHide



InstaHide

Public data

$x_1, \dots, x_n$ → "E" → $\tilde{x}_1, \dots, \tilde{x}_m$ → A → h

Requires definition + proof

Can test empirically

Hope: $\tilde{x}_1, \dots, \tilde{x}_m$ "encrypt" the original data, but are still good enough to train on.

Intuition: *Mixup** data augmentation

Require $f(\alpha x_1 + \beta x_2 + \gamma x_3) \approx (\alpha, \beta, \gamma)$

* Zhang, Cisse, Dauphin, Lopez-Paz '18

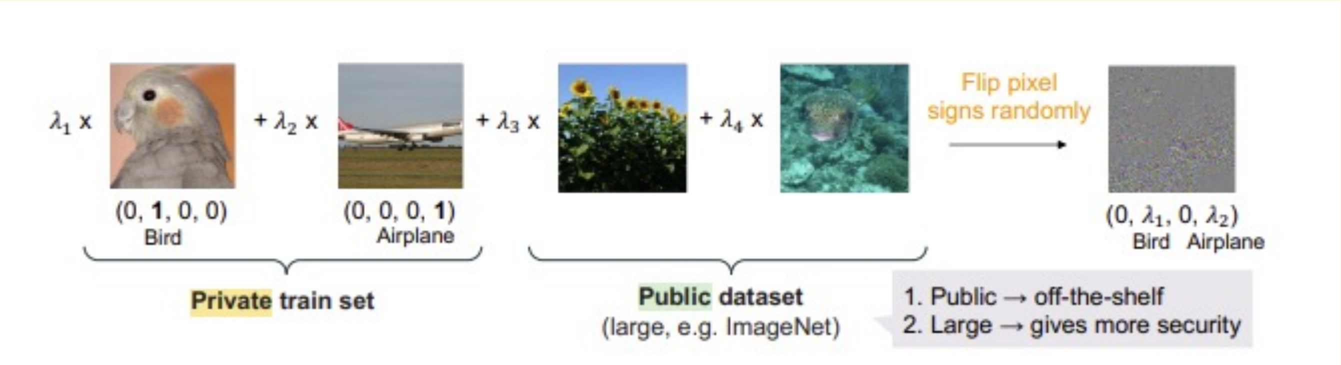
InstaHide

Public data

$x_1, \ldots, x_n$

$"E"$

	MNIST	CIFAR-10	CIFAR-100	ImageNet
Vanilla training	99.5 ± 0.1	94.8 ± 0.1	77.9 ± 0.2	77.4
DPSGD*	98.1	72.0	N/A	N/A
InstaHide$_{\text{inside},k=4,\text{ in inference}}$	98.2 ± 0.2	91.4 ± 0.2	73.2 ± 0.2	72.6
InstaHide$_{\text{inside},k=4}$	98.2 ± 0.3	91.2 ± 0.2	73.1 ± 0.3	1.4
InstaHide$_{\text{cross},k=4,\text{ in inference}}$	98.1 ± 0.2	90.3 ± 0.2	72.8 ± 0.3	-
InstaHide$_{\text{cross},k=4}$	97.8 ± 0.2	90.7 ± 0.2	73.2 ± 0.2	-
InstaHide$_{\text{cross},k=6,\text{ in inference}}$	97.4 ± 0.2	89.6 ± 0.3	72.1 ± 0.2	-
InstaHide$_{\text{cross},k=6}$	97.3 ± 0.1	89.8 ± 0.3	71.9 ± 0.3	-



$x \in [-1, +1]^n$

1) $x' = \lambda_1 x^1 + \lambda_2 x^2 + \lambda_3 x^3 + \lambda_4 x^4$

2) $\tilde{x} = (x'_1 k_1, \cdots, x'_n k_n)$

for $k \sim \{\pm 1\}^n$

OTP inspired

Attack on InstaHide

An Attack on *InstaHide*:
Is Private Learning Possible with Instance Encoding?

Nicholas Carlini ncarlini@google.com
Samuel Deng sd3013@columbia.edu
Sanjam Garg sanjamg@berkeley.edu
Somesh Jha jha@cs.wisc.edu
Saeed Mahloujifar sfar@princeton.edu
Mohammad Mahmoody mohammad@virginia.edu
Shuang Song shuangsong@google.com
Abhradeep Thakurta athakurta@google.com
Florian Tramèr tramer@cs.stanford.edu

Figure 1: Our solution to the InstaHide Challenge. Given 5,000 InstaHide encoded images released by the authors, under the strongest settings of InstaHide, we recover a visually recognizable version of the original (private) images in under an hour on a single machine.

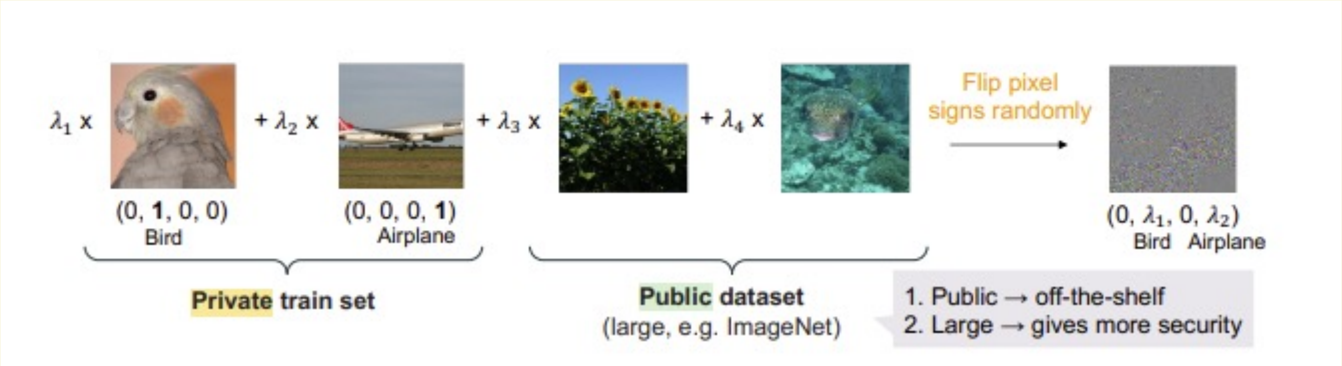
Attack description

x_i = R/G/B value of pixel, normalized to $[-1, +1]$

1) $x' = \lambda_1 x^1 + \lambda_2 x^2 + \lambda_3 x^3 + \lambda_4 x^4$

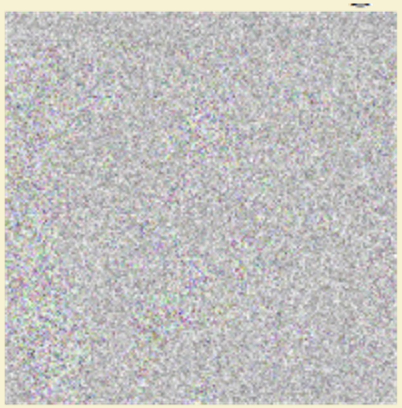
2) $\tilde{x} = (x'_1 k_1, \cdots, x'_n k_n)$

for $k \sim \{\pm 1\}^n$



Obs 1: $x_1 \ldots x_n \mapsto (k_1 x_1, \ldots, k_n x_n)$ for $k \in \{\pm 1\}^n$ allows to recover $(|x_1|, \ldots, |x_n|)$

Original image



Sign Flipped

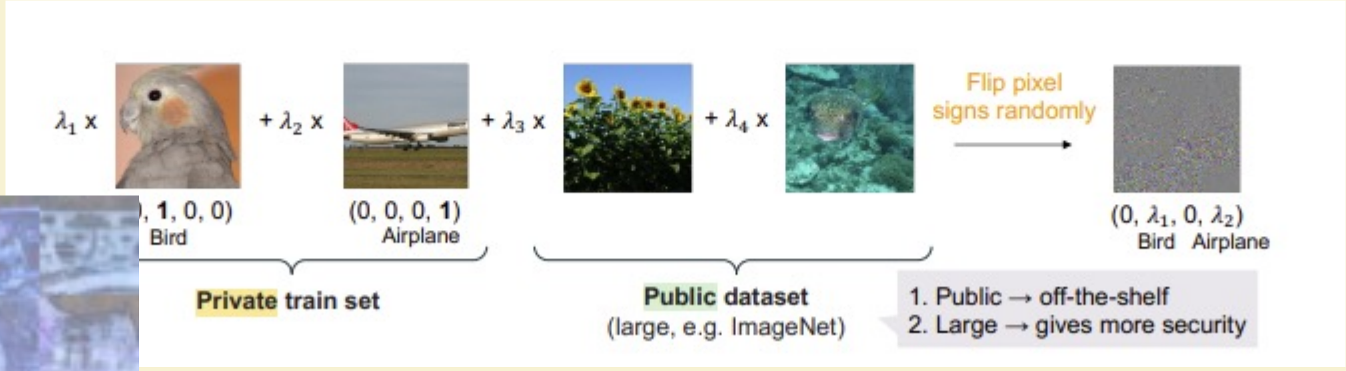
Absolute value

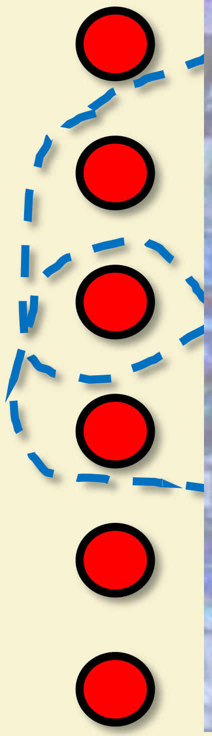
Attack description

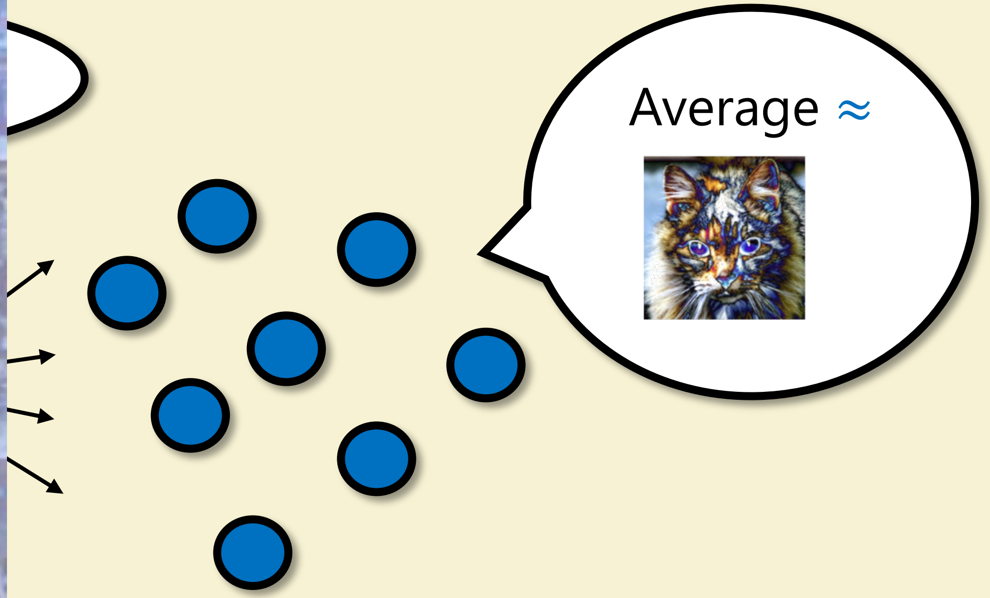
x_i = R/G/B value of pixel, normalized to $[-1, +1]$

1) $x' = \lambda_1 x^1 + \lambda_2 x^2 + \lambda_3 x^3 + \lambda_4 x^4$

2) $\tilde{x} = (|x_1'|, \ldots, |x'|)$







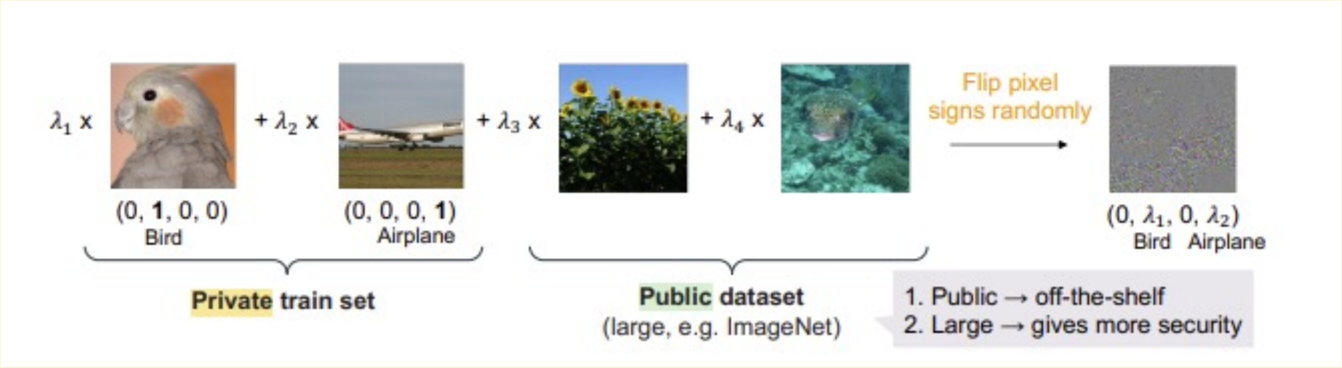
All came from same original private image

Attack description

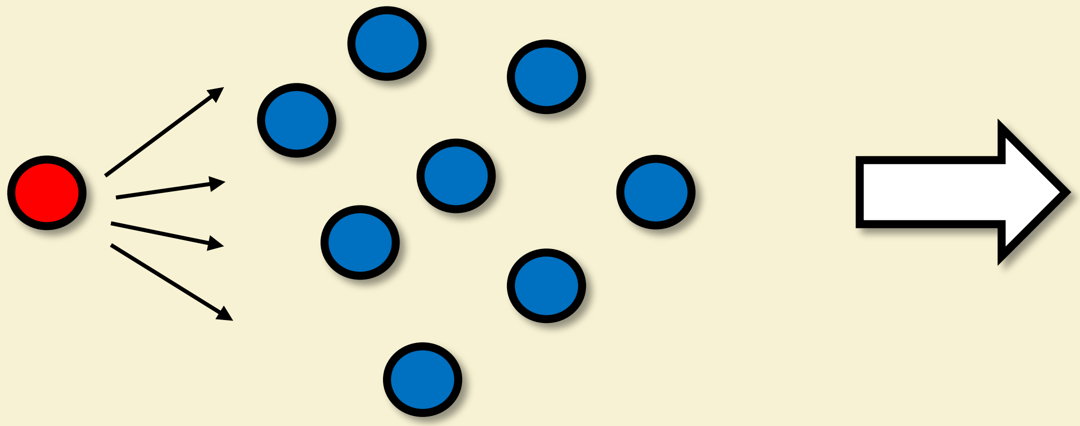
x_i = R/G/B value of pixel, normalized to $[-1, +1]$

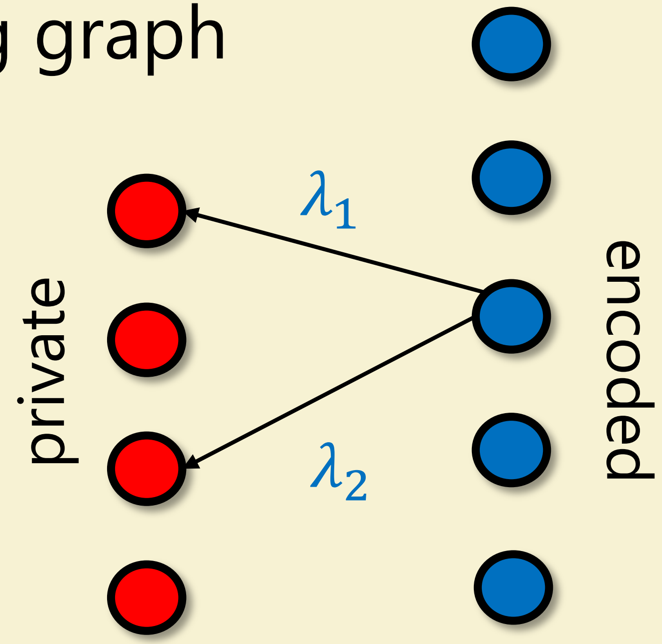
1) $x' = \lambda_1 x^1 + \lambda_2 x^2 + \lambda_3 x^3 + \lambda_4 x^4$

2) $\tilde{x} = (|x'_1|, \dots, |x'_n|)$



Reconstruct encoding graph





All came from same original private image

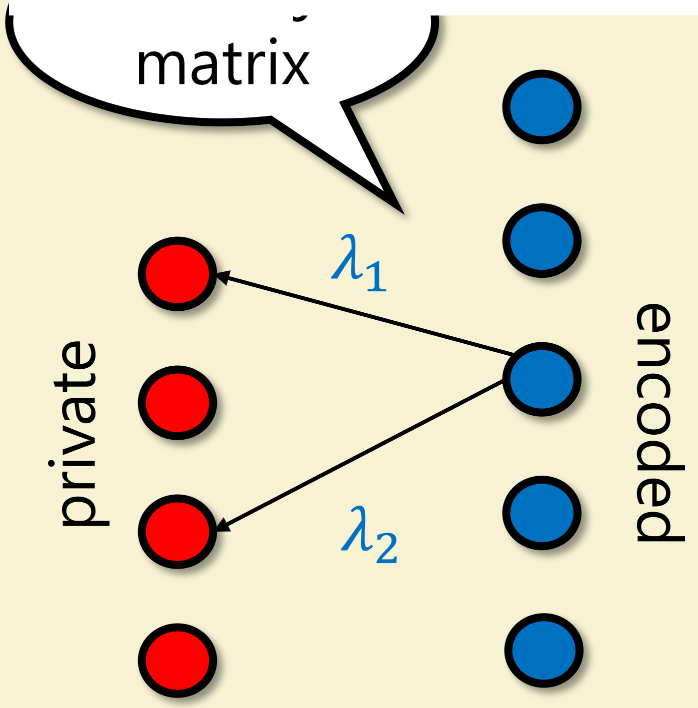
$\tilde{x} = abs(\lambda_1 x_i + \lambda_2 x_j + noise)$

At

1)

2)

Figure 1: Our solution to the InstaHide Challenge. Given 5,000 InstaHide encoded images released by the authors, under the strongest settings of InstaHide, we recover a visually recognizable version of the original (private) images in under an hour on a single machine.



$\tilde{x} = abs(\lambda_1 x_i + \lambda_2 x_j + noise)$

InstaHide challenge:

100 private images

5000 encoded images

$5000n$ non-linear eq in $100n$ vars

Use GD to find $\underset{X \in [-1,1]^{n \times t}}{\arg\min} \left\| \mathrm{abs}(AX) - \tilde{X} \right\|^2$

Black Box recovery

Cryptanalytic Extraction of Neural Network Models

Nicholas Carlini[1] Matthew Jagielski[2] Ilya Mironov[3]

Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^\dagger$	$\infty^\dagger$	$\infty^\dagger$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Table 1. Efficacy of our extraction attack which is orders of magnitude more precise than prior work and for deeper neural networks orders of magnitude more query efficient. Models denoted *a*-*b*-*c* are *fully connected* neural networks with input dimension *a*, one hidden layer with *b neurons*, and *c* outputs; for formal definitions see Section 2. Entries denoted with a † were unable to recover the network after ten attempts.