

# CS 229br Lecture 7:

## Privacy

Boaz Barak



Yamini Bansal  
Official TF



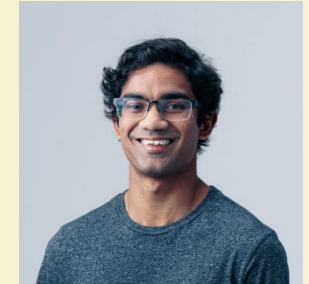
Javin Pombra  
Official TF



Dimitris Kalimeris  
Unofficial TF

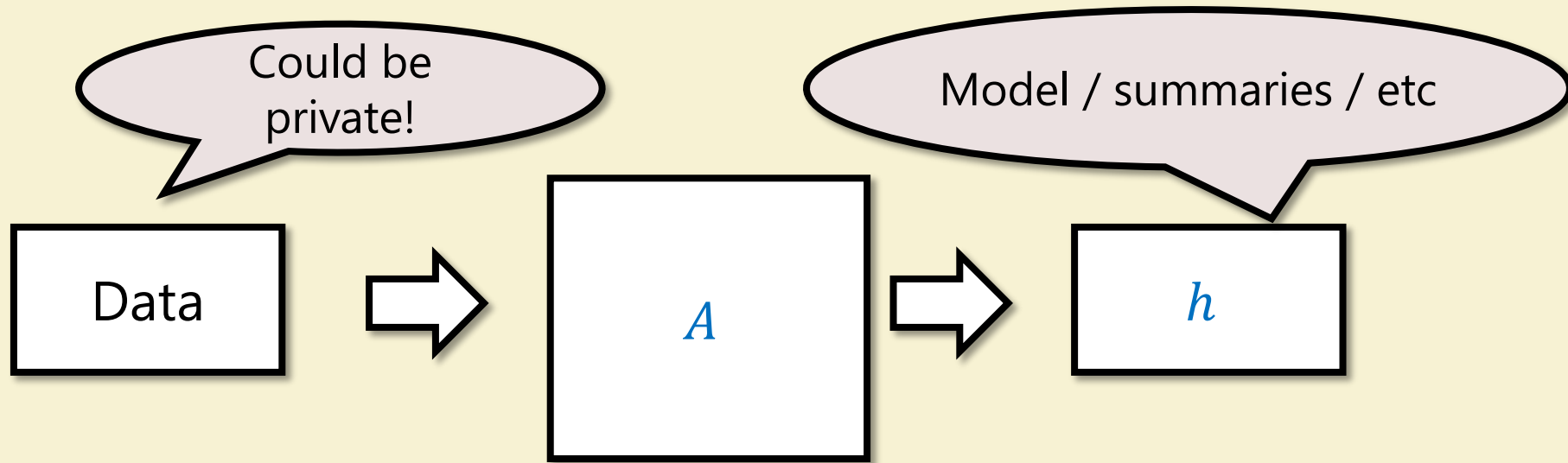


Gal Kaplun  
Unofficial TF

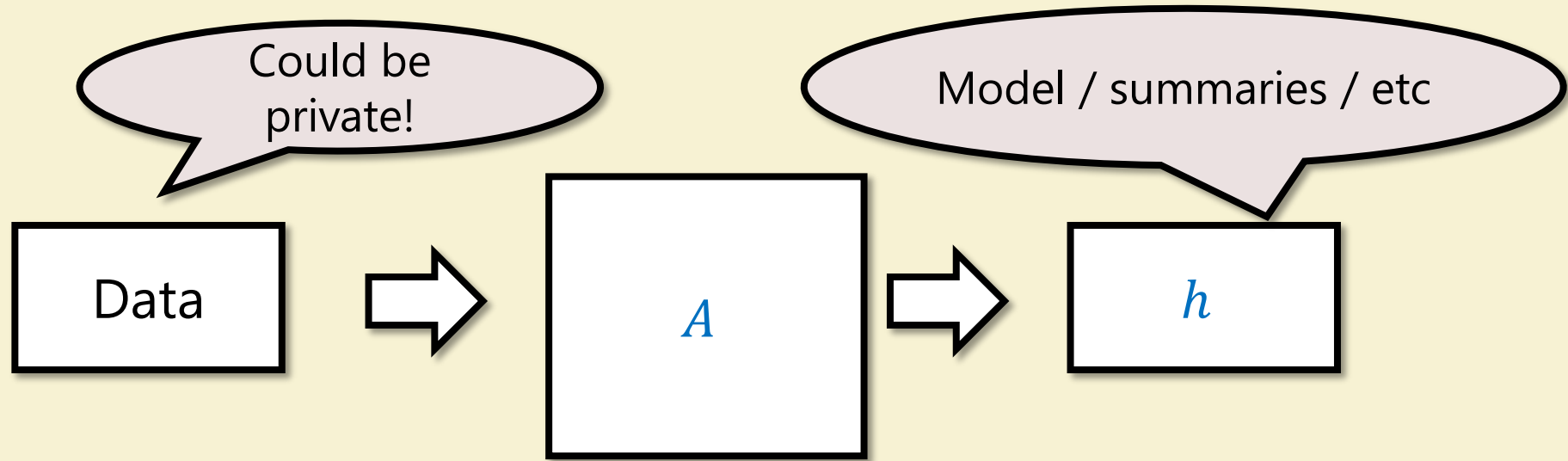


Preetum Nakkiran  
Unofficial TF

# Learning



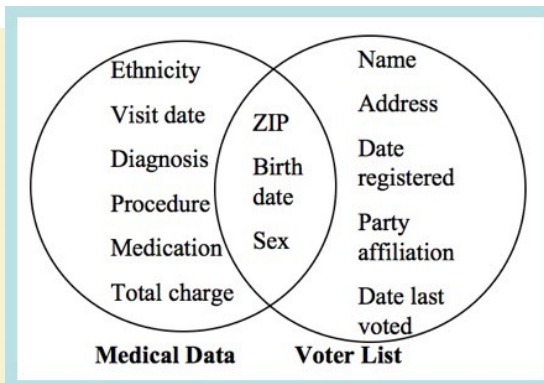
# Learning



## 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

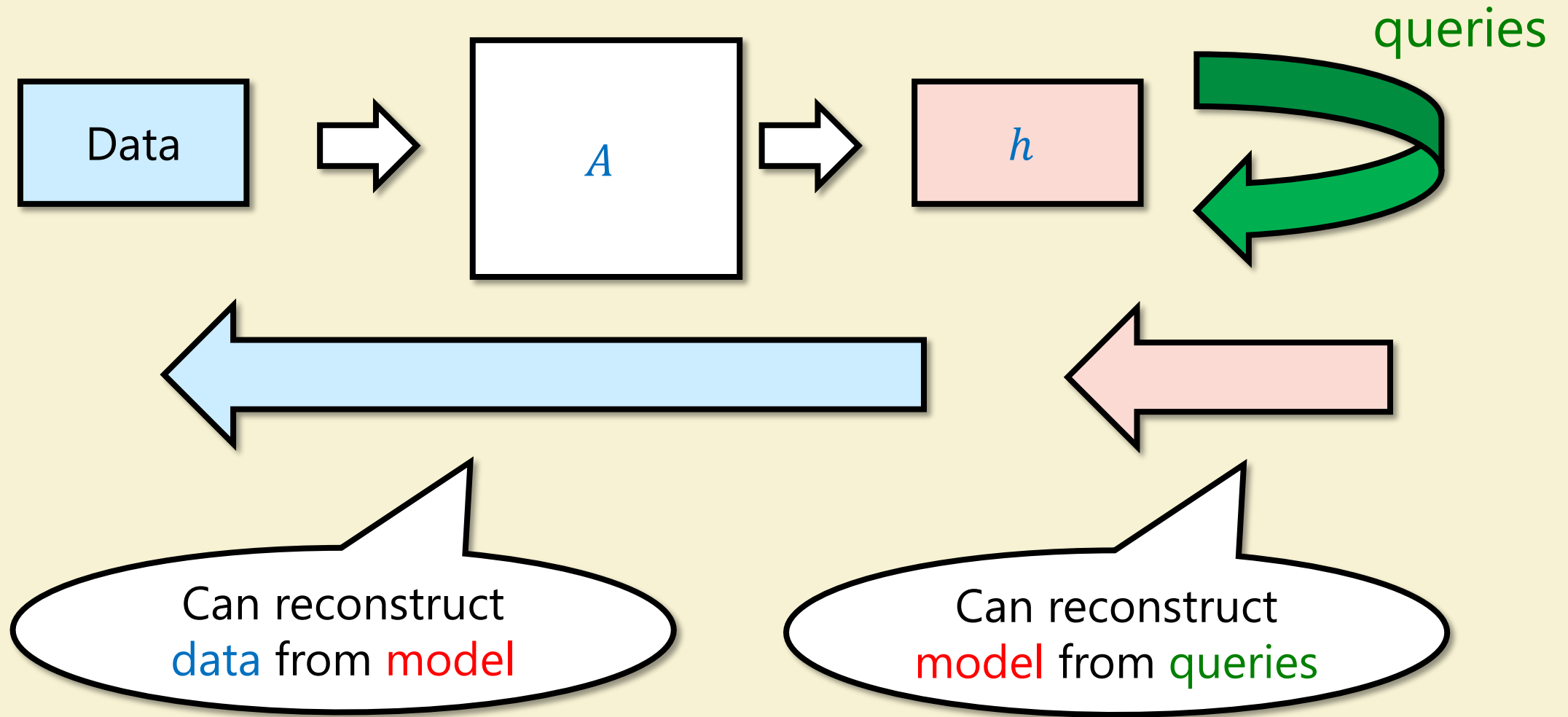


User	Secret Type	Exposure	Extracted?
A	CCN	52	✓
B	SSN	13	
C	SSN	16	
	SSN	10	
	SSN	22	
D	SSN	32	✓
F	SSN	13	
G	CCN	36	
	CCN	29	
	CCN	48	✓

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

Carlini, Liu, Erlingsson, Kos, Song '19

# Deep Learning



# Extracting Training Data from Large Language Models

Nicholas Carlini<sup>1</sup>

Florian Tramèr<sup>2</sup>

Eric Wallace<sup>3</sup>

Matthew Jagielski<sup>4</sup>

Ariel Herbert-Voss<sup>5,6</sup>

Katherine Lee<sup>1</sup>

Adam Roberts<sup>1</sup>

Tom Brown<sup>5</sup>

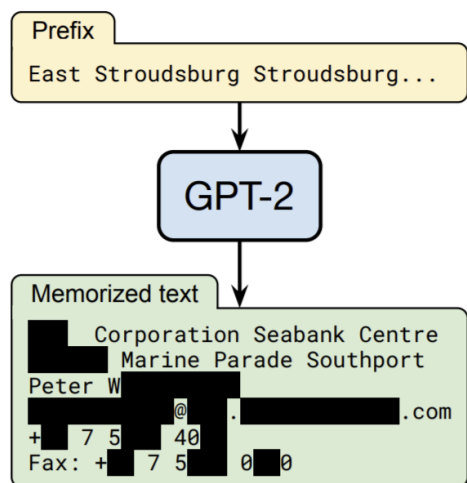
Dawn Song<sup>3</sup>

Úlfar Erlingsson<sup>7</sup>

Alina Oprea<sup>4</sup>

Colin Raffel<sup>1</sup>

<sup>1</sup>Google <sup>2</sup>Stanford <sup>3</sup>UC Berkeley <sup>4</sup>Northeastern University <sup>5</sup>OpenAI <sup>6</sup>Harvard <sup>7</sup>Apple



Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
<b>Named individuals (non-news samples only)</b>	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
<b>Contact info (address, email, phone, twitter, etc.)</b>	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Memorized String	Sequence Length	Occurrences in Data	
		Docs	Total
Y2...[redacted]...y5	87	1	10
7C...[redacted]...18	40	1	22
XM...[redacted]...WA	54	1	36
ab...[redacted]...2c	64	1	49
ff...[redacted]...af	32	1	64
C7...[redacted]...ow	43	1	83
0x...[redacted]...C0	10	1	96
76...[redacted]...84	17	1	122
a7...[redacted]...4b	40	1	311

# Why Is Google Translate Spitting Out Sinister Religious Prophecies?

Maori ▼

Translate from English

dog dog dog dog dog dog dog dog dog  
dog dog dog dog dog dog dog dog dog  
dog Edit

English ▼

Doomsday Clock is three minutes at twelve We are experiencing characters and a dramatic developments in the world, which indicate that we are increasingly approaching the end times and Jesus' return

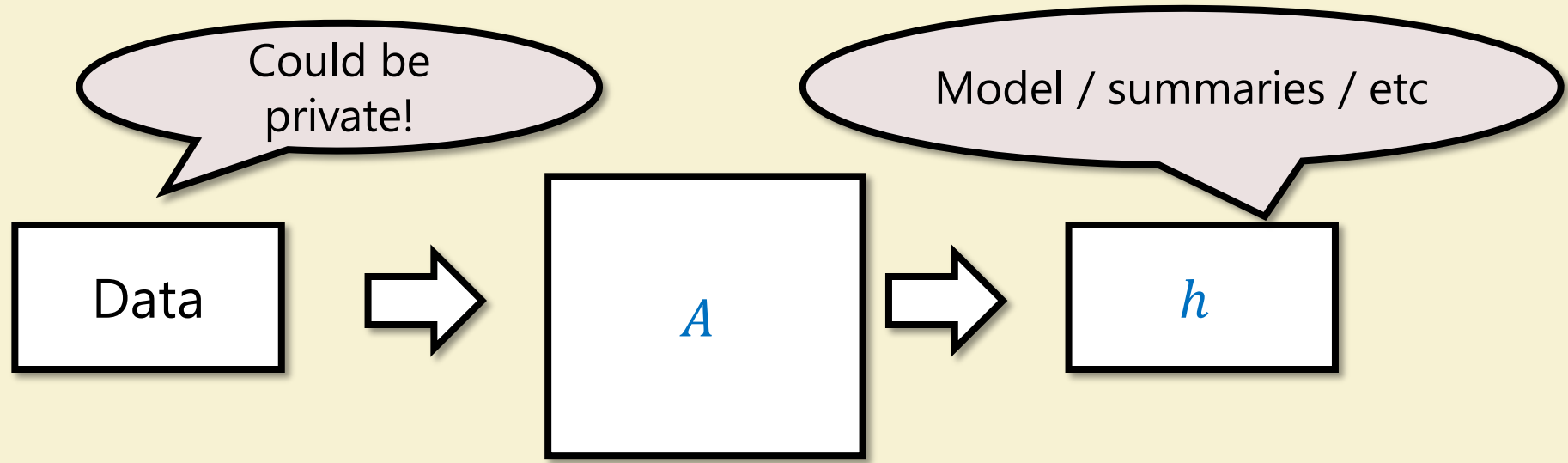
Somali ▼

Translate from Irish

ag ag ag ag ag ag ag ag ag ag ag ag  
ag ag ag ag ag ag ag ag ag Edit

English ▼

As a result, the total number of the members of the tribe of the sons of Gershon was one hundred fifty thousand



## Solutions:

- **Cryptographic:** 100% privacy but at efficiency/control cost
- **Differential privacy:** "X% privacy" but X vs utility tradeoff not great
- **Heuristics:** Hope for 100%, might get 0%

# Part I: Protecting ML using crypto



# Divergence: One-time pad

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

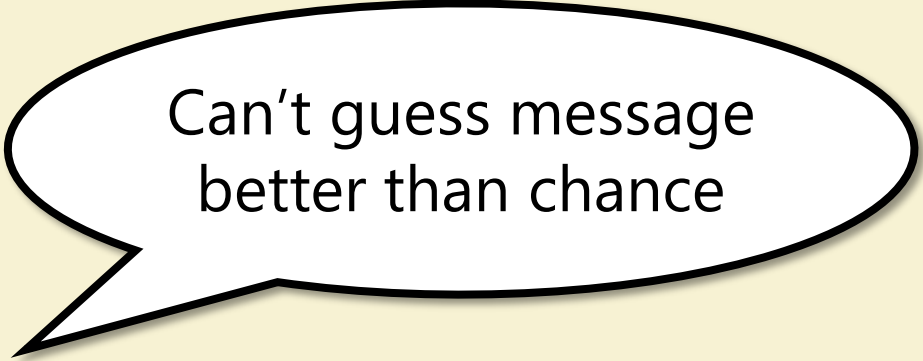
Encryption:  $E: \{0,1\}^n \times \{0,1\}^\ell \rightarrow \{0,1\}^m$

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

Correctness:  $\forall_k \forall_{x \in \{0,1\}^\ell}, D_k(E_k(x)) = x$

Perfect Secrecy:  $\forall$  alg  $A$

$$\Pr_{\substack{x \sim \{0,1\}^\ell \\ k \sim \{0,1\}^n}} [A(E_k(x)) = x] \leq 2^{-\ell}$$



Can't guess message  
better than chance

# Divergence: One-time pad

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

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

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

Correctness:  $\forall_k \forall_{x \in \{0,1\}^\ell}, D_k(E_k(x)) = x$

Perfect Secrecy:  $\forall \text{ alg } A \quad \Pr_{\substack{x \sim \{0,1\}^\ell \\ k \sim \{0,1\}^n}} [A(E_k(x)) = x] \leq 2^{-\ell}$

---

## Shannon's Two Theorems:

Thm 1: The *one-time pad* achieves perfect secrecy with  $n = \ell$

Thm 2: Every perfectly-secret scheme requires  $n \geq \ell$



Gene Grabeel

# Divergence: One-time pad

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

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

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

Correctness:  $\forall_k \forall_{x \in \{0,1\}^\ell}, D_k(E_k(x)) = x$

Perfect Secrecy:  $\forall \text{ alg } A \quad \Pr_{\substack{x \sim \{0,1\}^\ell \\ k \sim \{0,1\}^n}} [A(E_k(x)) = x] \leq 2^{-\ell}$

---

Thm 1: The *one-time pad* achieves perfect secrecy with  $n = \ell$

PF:  $E_k(x) = x \oplus k \quad D_k(y) = y \oplus k$

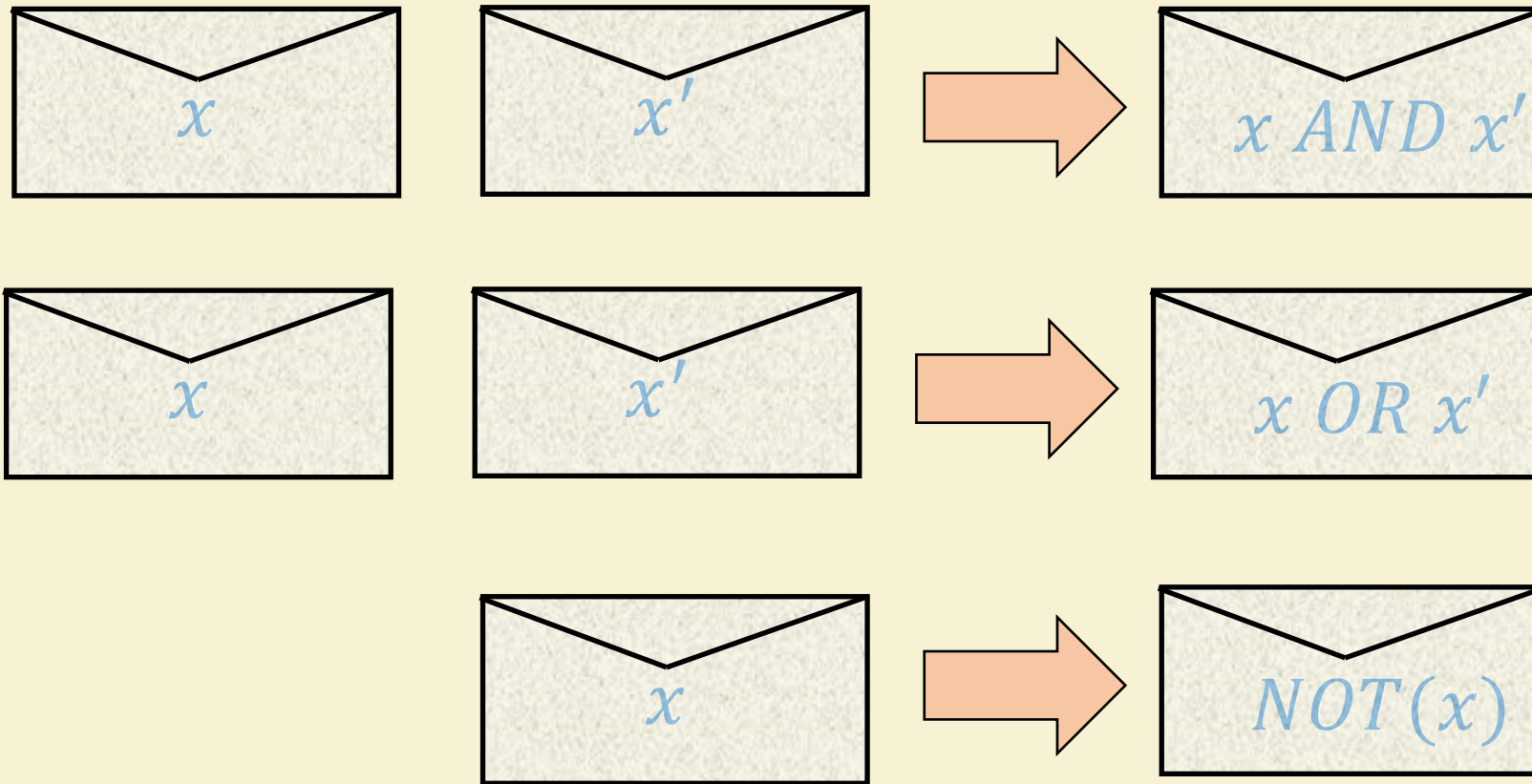
$$\Pr[A(k \oplus x) = x] = \Pr[A(y) = x] \leq 2^{-n}$$

Crucial:  
 $|\text{keys}| \geq |\text{msgs}|$

Equivalent description:  $k, x \in \{\pm 1\}^n, E_k(x) = (x_1 k_1, \dots, x_n k_n)$

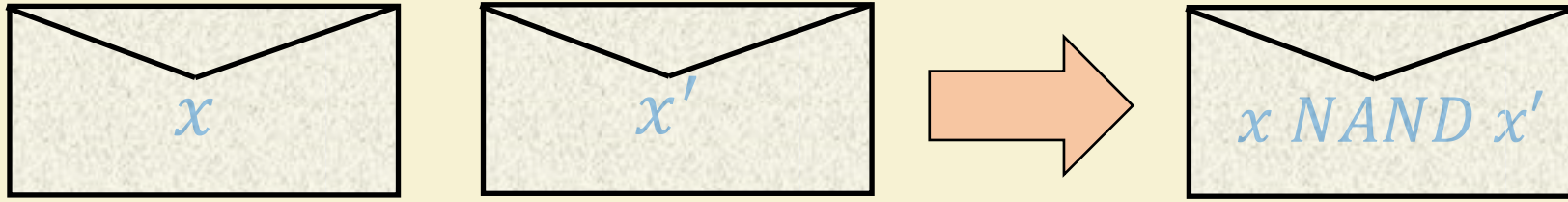
Extension:  $k, x \in \mathbb{Z}_t^n, E_k(x) = (x_1 + k_1 \bmod t, \dots, x_n + k_n \bmod t)$

# Fully Homomorphic Encryption (FHE)



**Note:** Can also use Multiparty Secure Computation (MPC)

# Fully Homomorphic Encryption (FHE)



# FHE

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

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

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

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

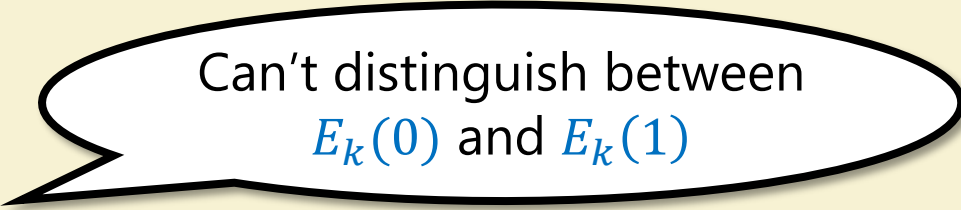
Correctness:  $\forall_k \forall_{b \in \{0,1\}}, D_k(E_k(b)) = b$


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

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

Computational  
secrecy\*:

$\forall$  alg  $A$  of time  $\ll \exp(n)$



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

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

\* 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.

## 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](#).

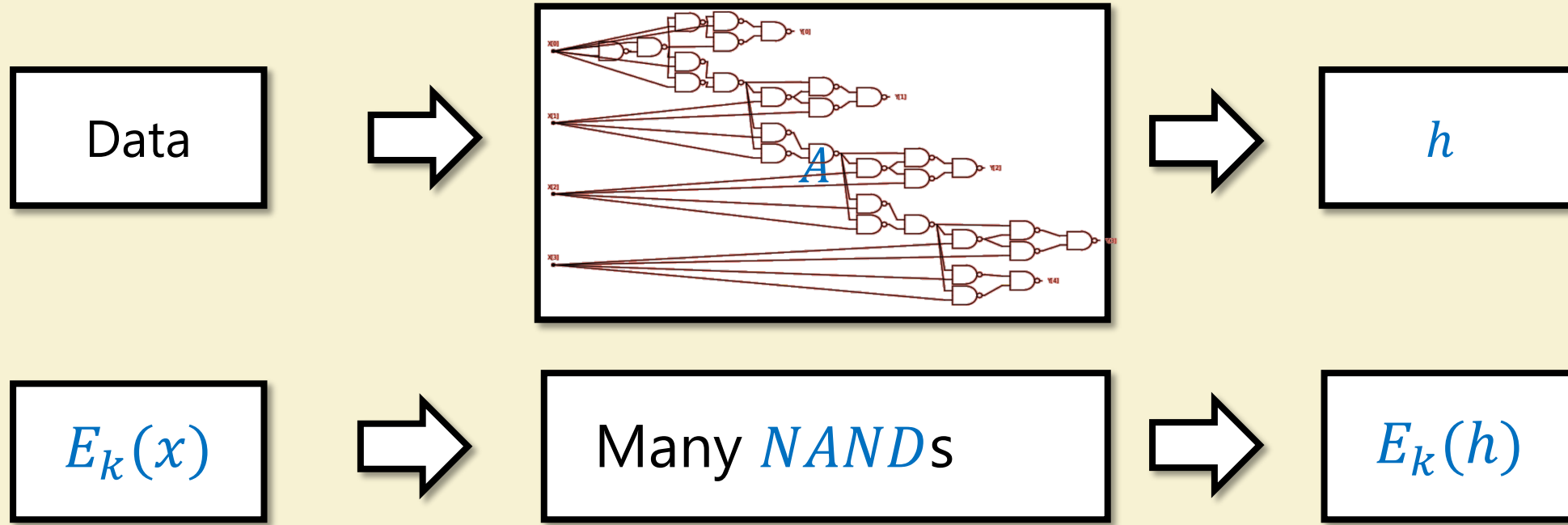


# What is FHE good for?

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

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

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



## 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](#)

# What is FHE

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

Data

$E_k(x)$

## Large Bank in the Americas

Use Case: Access to multi data silos – to enable better NBA Identification

Model Type: Logistic Regression

- Prediction on encrypted existing model and data
- Parameter selection and prediction on encrypted data

Metrics Required:

- Accuracy
- Computational overhead
- Memory overhead (important in warehouse scenario)

Data Set – Real private financial data

FHE algorithm : CKKS

- Approximate numbers

<https://eprint.iacr.org/2019/1113>

## Large European Bank

Use Case: Cloud Migration – Moving text classification of call centre reports to the cloud

Model : Neural Networks

- Large & complex deep neural networks on encrypted data
- Predictive Text Classification

Metrics Required:

- Accuracy
- Amount of data that could be processed in a certain timeframe (8 hrs)

Data Set – Public data

FHE algorithm : CKKS

- Approximate numbers

Research paper submitted to conference

Challenges:

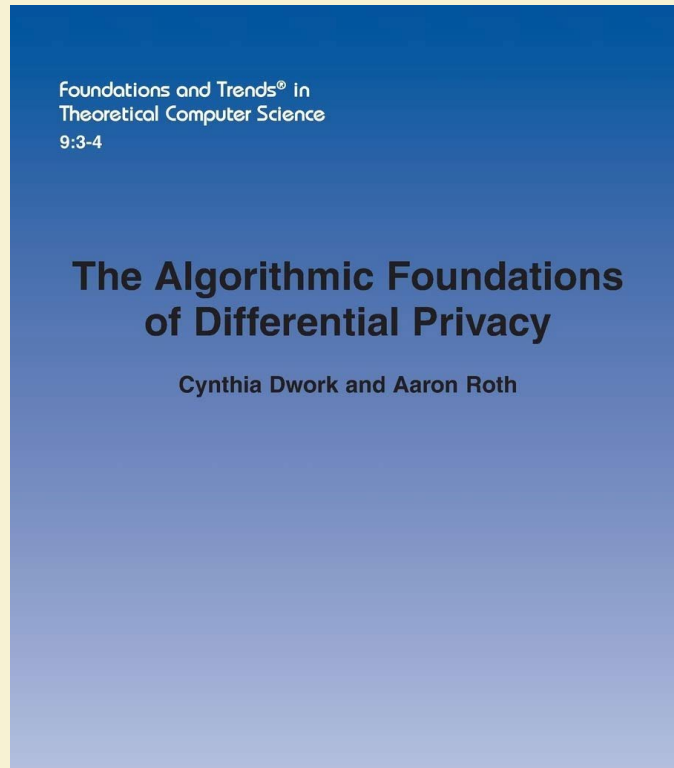
Only get *encrypted* model/summary

Huge computational overhead

Some partial preliminary successes

<https://arstechnica.com/gadgets/2020/07/ibm-completes-successful-field-trials-on-fully-homomorphic-encryption/>

# 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.

## Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data



TensorFlow Follow

Mar 6, 2019 · 7 min read



Posted by [Carey Radebaugh](#) (Product Manager) and [Ulfar Erlingsson](#) (Research Scientist)



Microsoft On the Issues

Our Company ▾

News and Stories ▾

Press Tools ▾

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 ▾

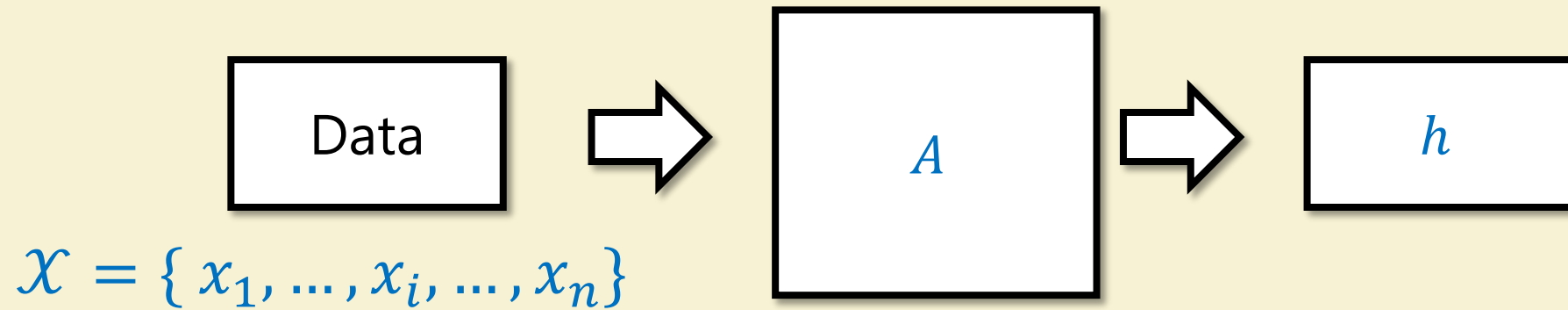
Opportunities ▾

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.

# Differential Privacy



Data belonging  
to  $i$ -th person

Def:  $A$  is  $\epsilon$  differentially private if

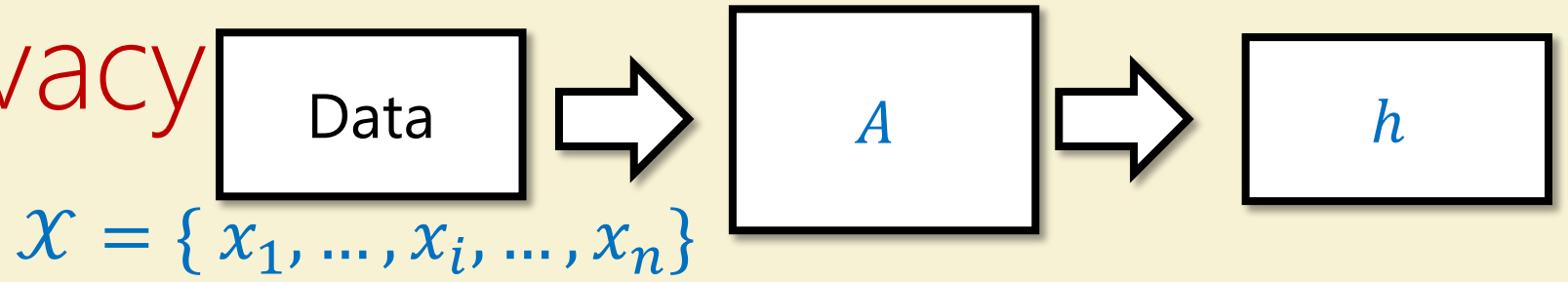
posterior probability  $x_i \in \mathcal{X} \in e^{\pm\epsilon} \times$  prior probability  $x_i \in \mathcal{X}$

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

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

$A$  must be  
randomized

# Differential Privacy



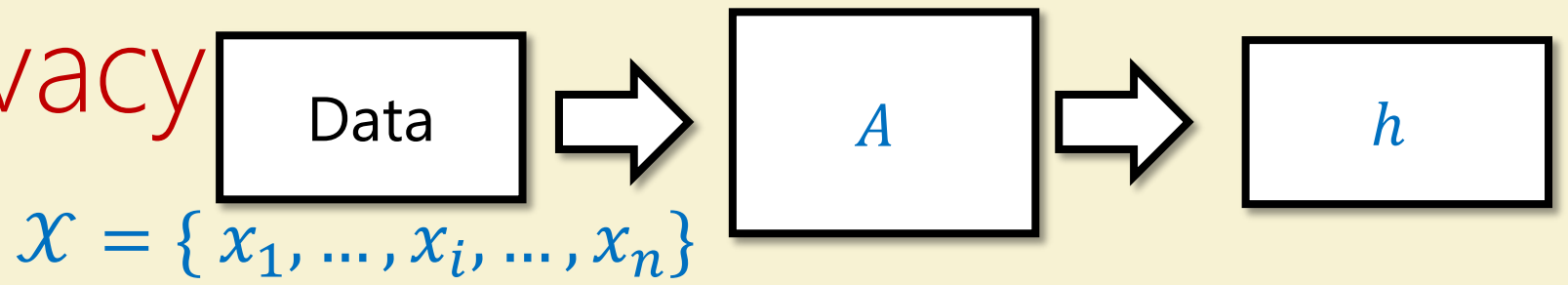
Def:  $A$  is  $\epsilon$  differentially private if

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

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

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

# Differential Privacy



Def:  $A$  is  $\epsilon$  *differentially private* if

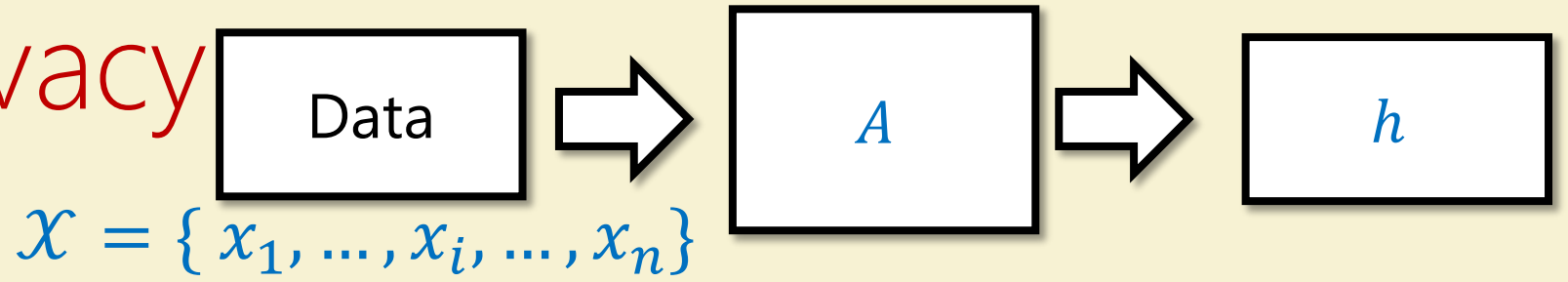
$$\forall \mathcal{X}, \mathcal{X}' \text{ s.t. } |\mathcal{X} \triangle \mathcal{X}'| = 1, \forall S$$

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

$$\Pr\left[ \begin{array}{l} \text{Bad event} \\ \text{happened to } i \\ \text{because their} \\ \text{data in } \mathcal{X} \end{array} \right] \leq e^{\epsilon} \cdot \Pr\left[ \begin{array}{l} \text{Bad event} \\ \text{happens} \\ \text{anyway} \end{array} \right]$$

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

# Differential Privacy



Def:  $A$  is  $\epsilon$  *differentially private* if

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

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

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

Not private!

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

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

# Differential privacy composition

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

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

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

# Differential privacy under post-processing

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

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

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



# DP guarantees

**Def:** A training mechanism  $\mathcal{X} \rightarrow f_w$  is **broken** if  $\exists A$  s.t.

$A(f_w)$  outputs  $x \in \mathcal{X}$

**Claim:** If mechanism is  $(\epsilon, \delta)$ -DP then broken with prob  $\leq \frac{\epsilon}{N} + \delta$   
( $\frac{1}{N}$  = prob random guessing  $x$ )

## Membership Inference Attacks Against Machine Learning Models

Reza Shokri  
Cornell Tech

shokri@cornell.edu

Marco Stronati\*  
INRIA

marco@stronati.org

Congzheng Song  
Cornell

cs2296@cornell.edu

Vitaly Shmatikov  
Cornell Tech

shmat@cs.cornell.edu

# Differentially private statistics:

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

In differentially private way

Why can't we just publish sums?

- 30 C19+ cases in Cambridge
- 29 C19+ cases age < 70
- 12 C19+ cases liver disease
- 11 C19+ cases age < 70 and liver disease

# Differentially private statistics:

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

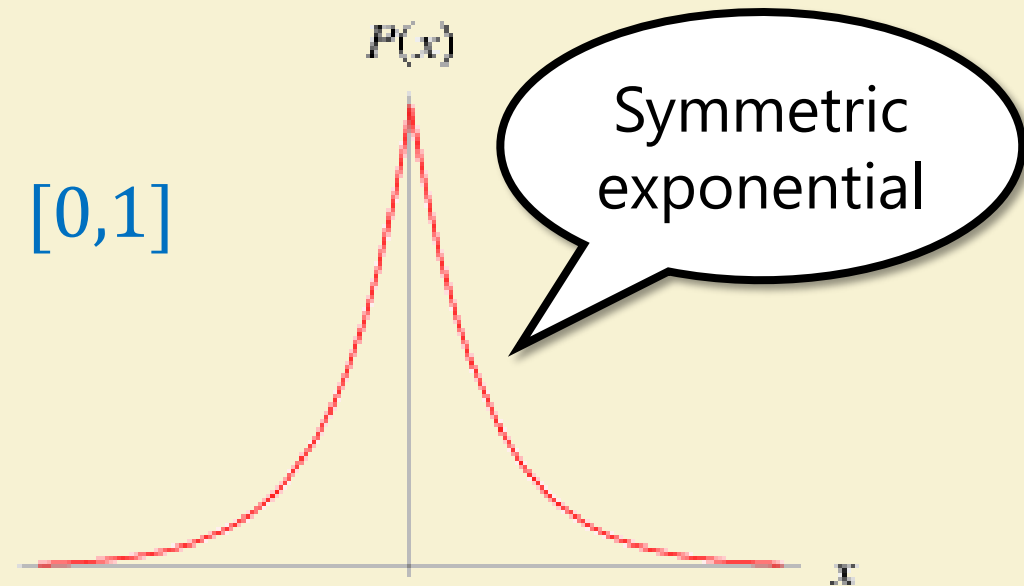
In differentially private way

**Laplace mechanism:**

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

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

**THM:** Laplace mechanism is  $\epsilon$ -DP



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

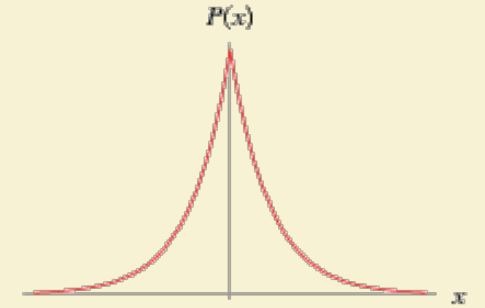
In practice,  $\sigma \approx \sqrt{n}$   
acceptable

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

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:

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



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

**THM:** Laplace mechanism is  $\epsilon$ -DP

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

**PF:** Focus on single  $f$

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

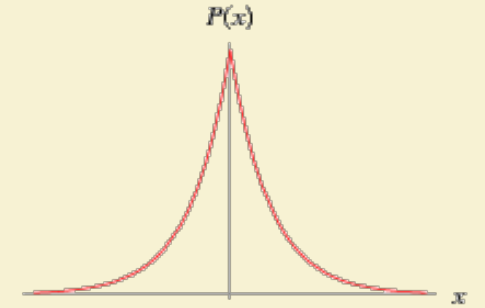
$$\Pr[\hat{f}(\mathcal{X}) = v] = \frac{1}{2\epsilon} \exp(-\epsilon|v - f(\mathcal{X})|) \leq \frac{1}{2\epsilon} \exp(\epsilon - \epsilon|v - f(\mathcal{X}')|) \leq e^\epsilon \cdot \Pr[\hat{f}(\mathcal{X}') = v]$$

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:

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

**THM:** Laplace mechanism is  $\epsilon$ -DP



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

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

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

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

Sensitivity of  $f$

# Important

Differential privacy is **definition**

Laplace mechanism is **one approach** to achieve definition

Can also use other noise distributions (e.g. Gaussian)

(typically get  $(\epsilon, \delta)$ -DP in such cases)

# DP-SGD

$\mathcal{L}_i$  = loss for batch  $i$   
sensitivity  $\approx b/n$

Replace step  $w \leftarrow w - \eta \nabla_{\mathcal{L}_i}(w)$

with  $w \leftarrow w - \eta \left[ \nabla_{\mathcal{L}_i^C}(w) + N(0, \sigma^2 C^2 I) \right]$

$\mathcal{L}_i^C$  = gradient for every sample clipped at  $C$

**THM:** For const  $\epsilon$ ,  $C$  can achieve  $(\epsilon, o(1))$ -DP with const  $\sigma$  as long as

$$\# \text{ steps} \ll \left( \frac{n}{b} \right)^2$$

## Deep Learning with Differential Privacy

October 25, 2016

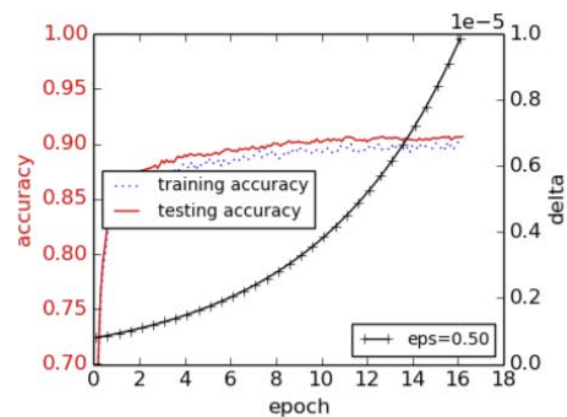
Martín Abadi\*  
H. Brendan McMahan\*

Andy Chu\*  
Ilya Mironov\*  
Li Zhang\*

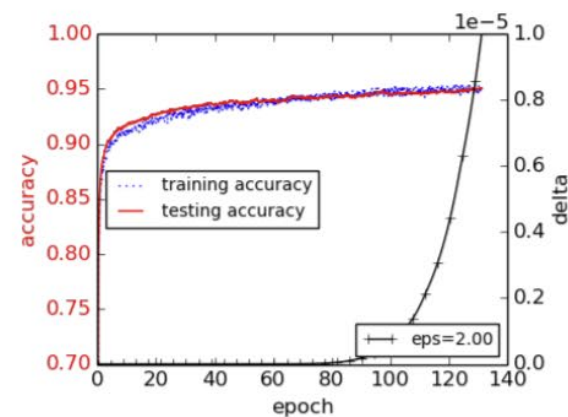
Ian Goodfellow†  
Kunal Talwar\*

# Evaluation

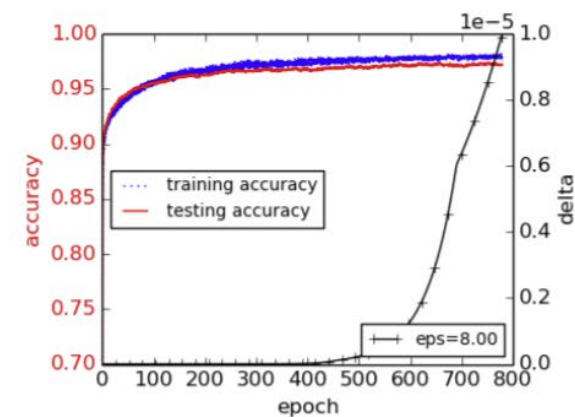
MNIST



(1) Large noise

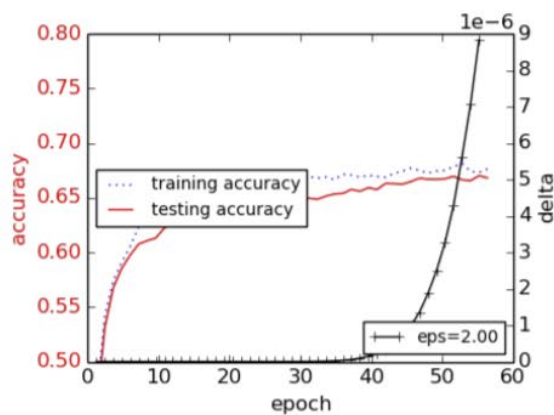


(2) Medium noise

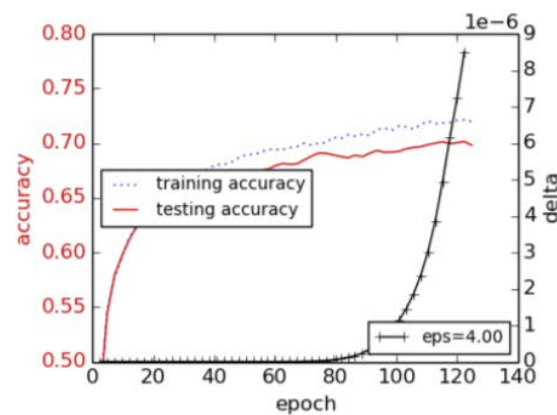


(3) Small noise

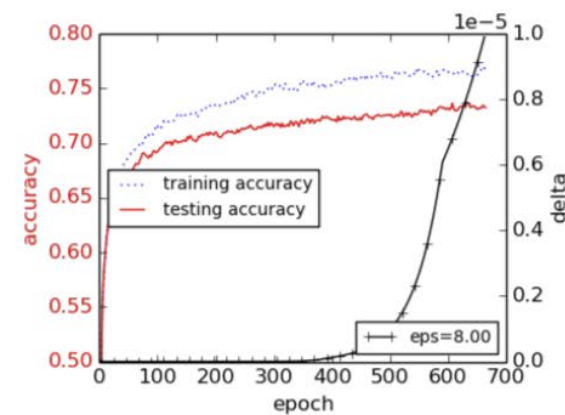
CIFAR 10



(1)  $\epsilon = 2$



(2)  $\epsilon = 4$



(3)  $\epsilon = 8$



# Protection from memorization in practice

	Optimizer	$\epsilon$	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	$2 \times 10^8$	1.32	3.2	
	RMSProp	$1 \times 10^9$	1.26	2.8	
	SGD	$\infty$	2.11	3.6	
No DP	SGD	N/A	1.86	9.5	
	RMSProp	N/A	1.17	31.0	✓

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

Nicholas Carlini<sup>1,2</sup>

Chang Liu<sup>2</sup>

Úlfar Erlingsson<sup>1</sup>

Jernej Kos<sup>3</sup>

Dawn Song<sup>2</sup>

# Private aggregation of teacher ensembles

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

Nicolas Papernot\*  
Pennsylvania State University

Martín Abadi  
Google Brain

Úlfar Erlingsson  
Google  
!google.com

Dataset	$\epsilon$	$\delta$	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  $\epsilon$  and failure probability  $\delta$  of the  $(\epsilon, \delta)$  differential privacy guarantee.

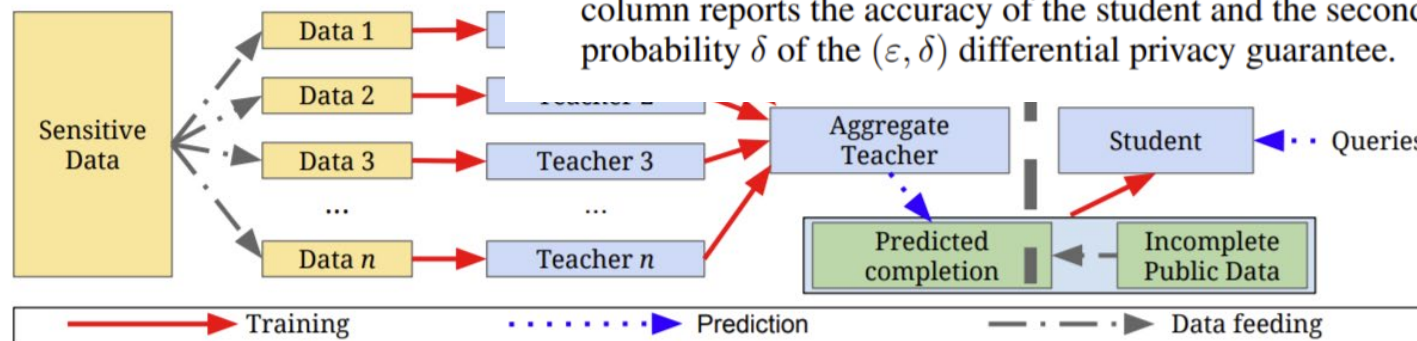


Figure 2: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

# 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 $\epsilon$	Accuracy	
				Student	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 ( $T=200, \sigma_1=150, \sigma_2=40$ )	286	<b>1.97</b>	<b>98.5%</b>	
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 ( $T=300, \sigma_1=200, \sigma_2=40$ )	3,098	<b>4.96</b>	<b>91.6%</b>	
Adult	LNMax (Papernot et al., 2017)	500	2.66	83.0%	85.0%
	Confident-GNMax ( $T=300, \sigma_1=200, \sigma_2=40$ )	524	<b>1.90</b>	<b>83.7%</b>	
Glyph	LNMax	4,000	4.3	72.4%	82.2%
	Confident-GNMax ( $T=1000, \sigma_1=500, \sigma_2=100$ )	10,762	2.03	<b>75.5%</b>	
	Interactive-GNMax, two rounds	4,341	<b>0.837</b>	73.2%	

# Heuristics

Avoid DP issues:

- Accuracy hit
- Large values for  $\epsilon$
- Slower

# InstaHide

Recall FHE-based training:

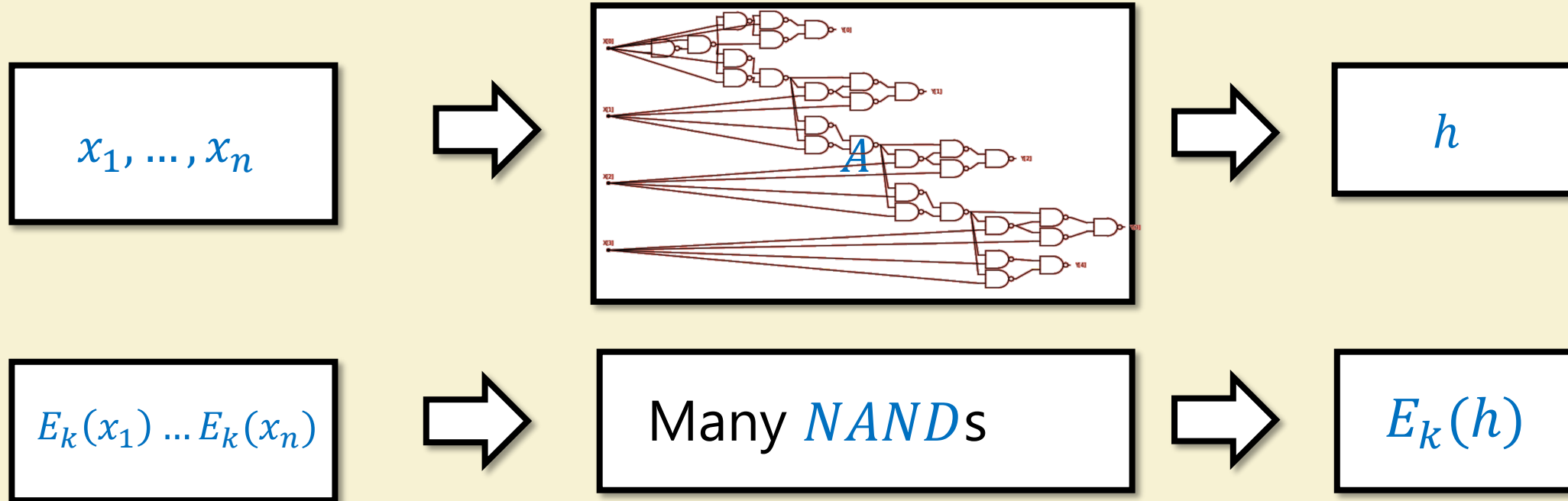
*InstaHide*: Instance-hiding Schemes for Private Distributed Learning\*

Yangsibo Huang<sup>†</sup>

Zhao Song<sup>‡</sup>

Kai Li<sup>§</sup>

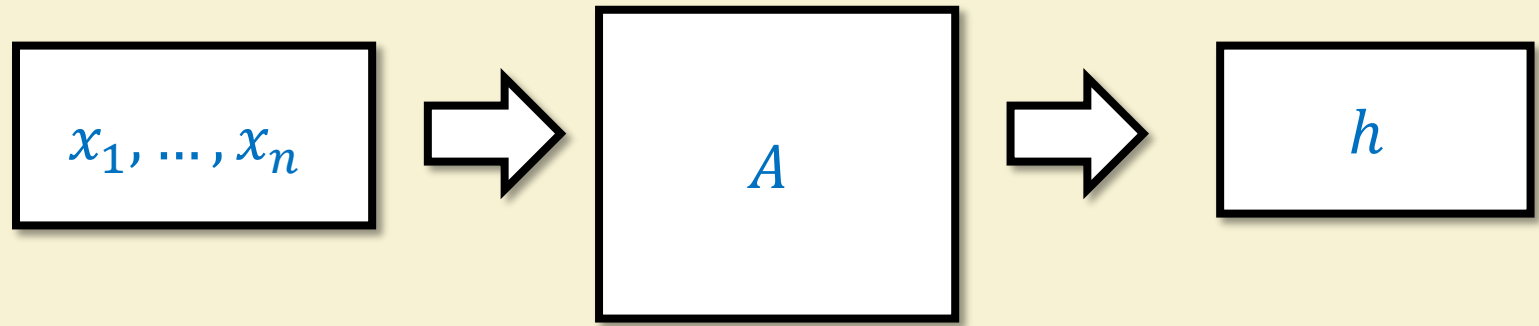
Sanjeev Arora<sup>¶</sup>



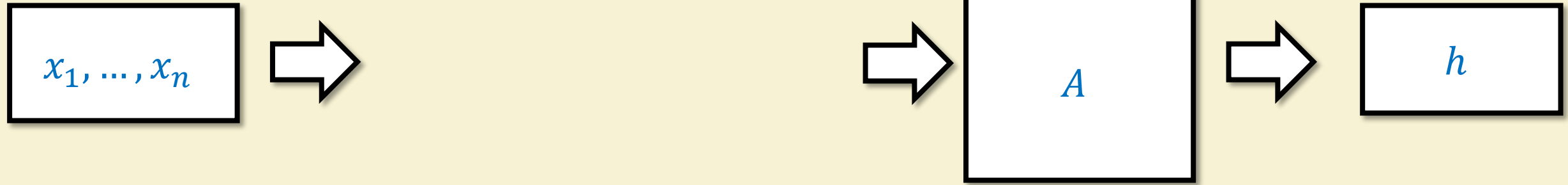
Challenges:

Only get *encrypted* model/summary  
Huge computational overhead

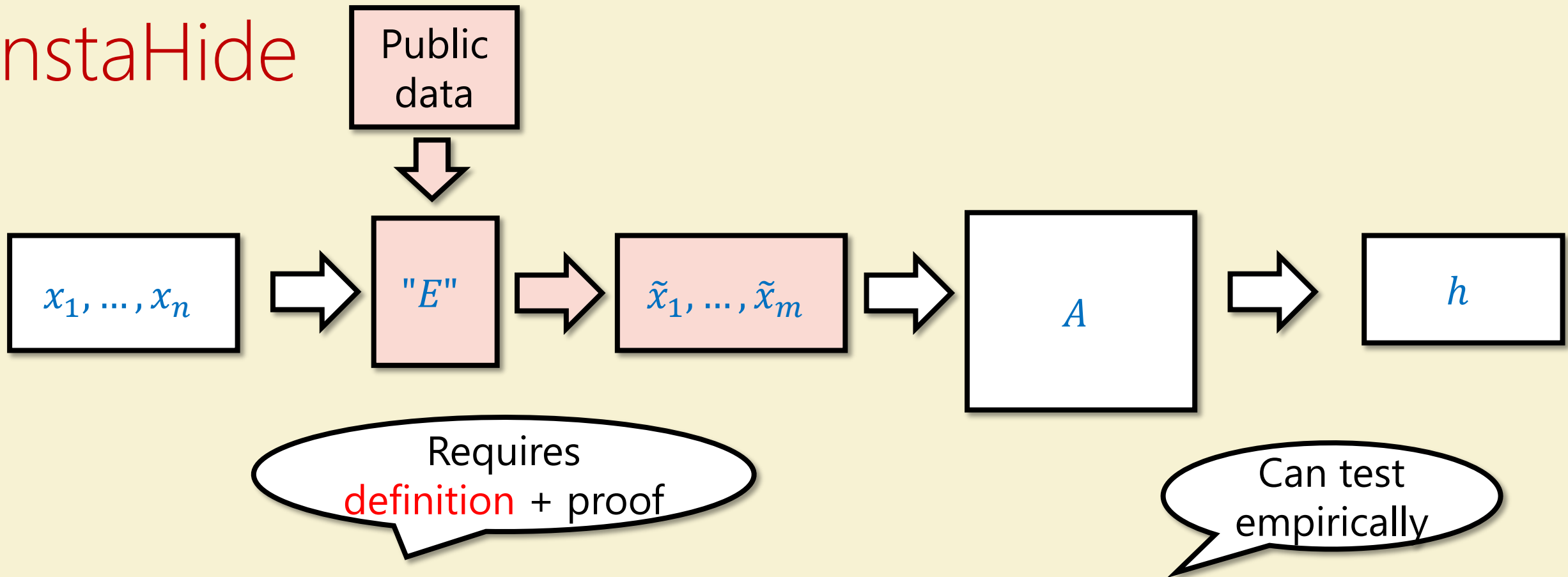
# InstaHide



# InstaHide



# InstaHide



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

**Intuition:** *Mixup*\* data augmentation

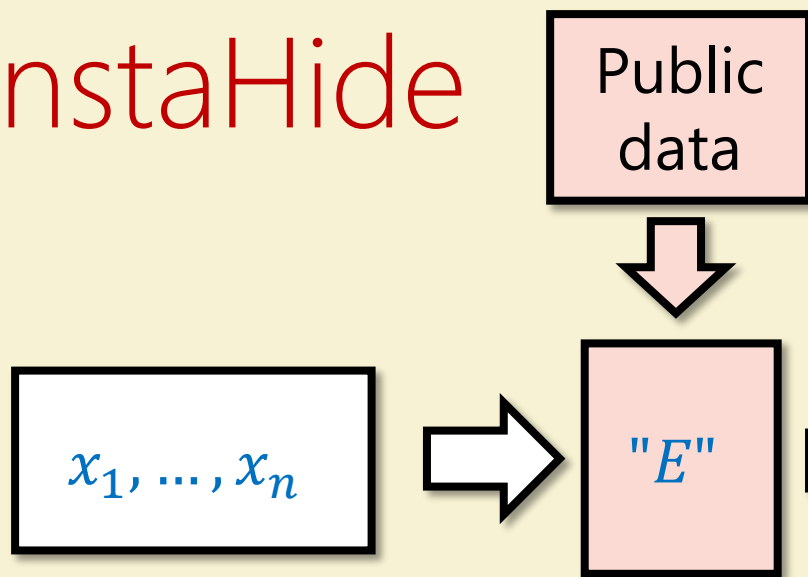
Require  $f(\alpha x_1 + \beta x_2 + \gamma x_3) \approx (\alpha, \beta, \gamma)$



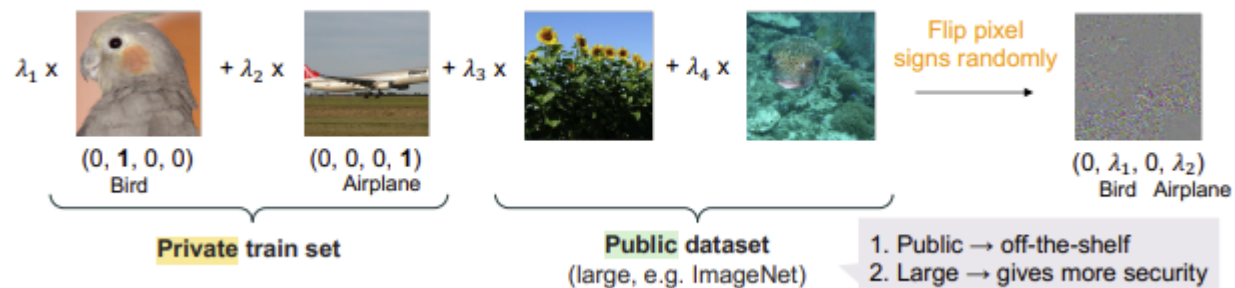
\* Zhang, Cisse, Dauphin, Lopez-Paz '18



# InstaHide



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



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

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

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

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

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 Mahmood**  
mohammad@virginia.edu

**Shuang Song**  
shuangsong@google.com

**Abhradeep Thakurta**  
athakurta@google.com

**Florian Tramèr**  
tramer@cs.stanford.edu

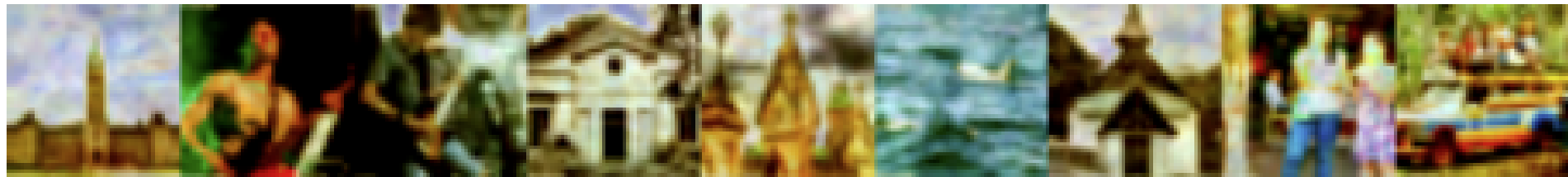


Figure 1: Our solution to the InstaHide Challenge. Given 5,000 InstaHide encoded images released by the authors, under the strongest settings of InstaHide, we recover a visually recognizable version of the original (private) images in under an hour on a single machine.

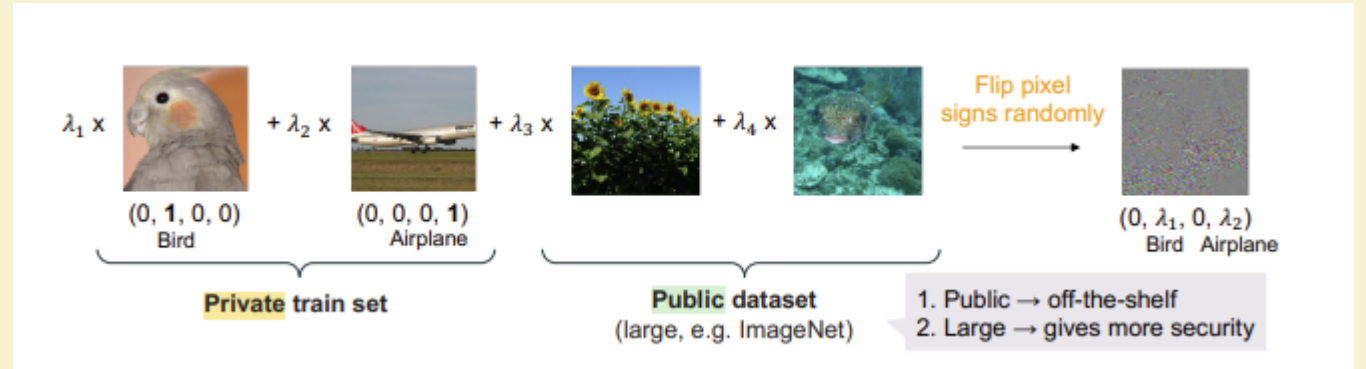
# Attack description

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

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

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

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



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



Original  
image



Sign  
Flipped



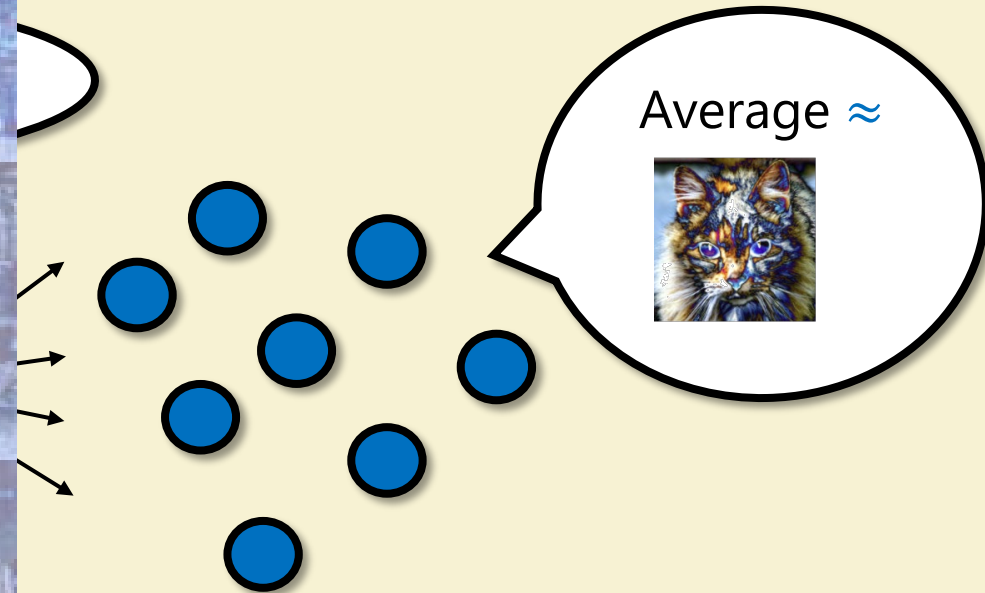
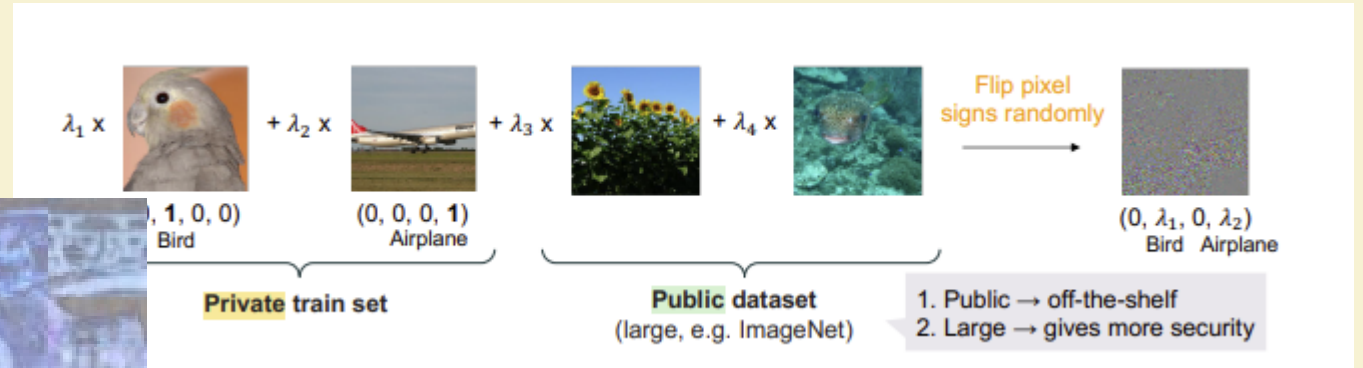
Absolute  
value

# Attack description

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

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

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



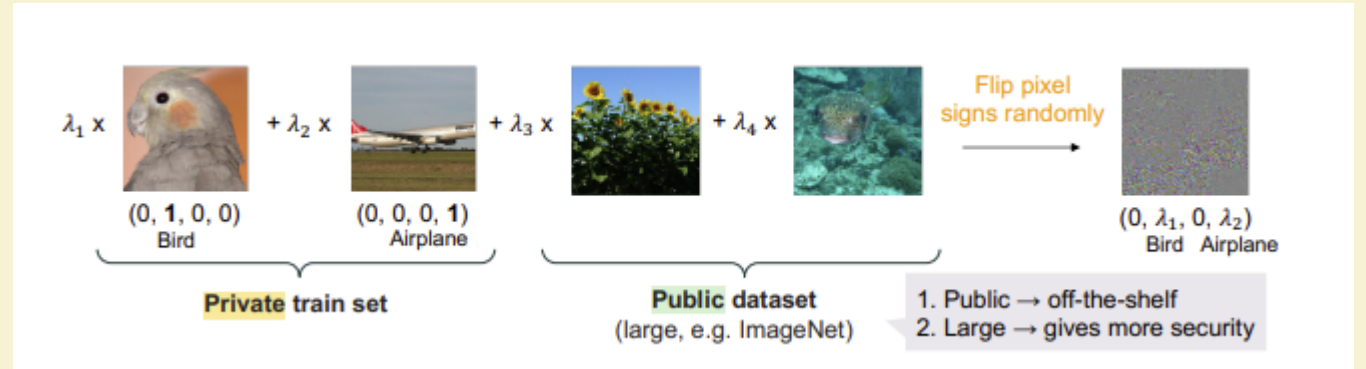
All came from same original private image

# Attack description

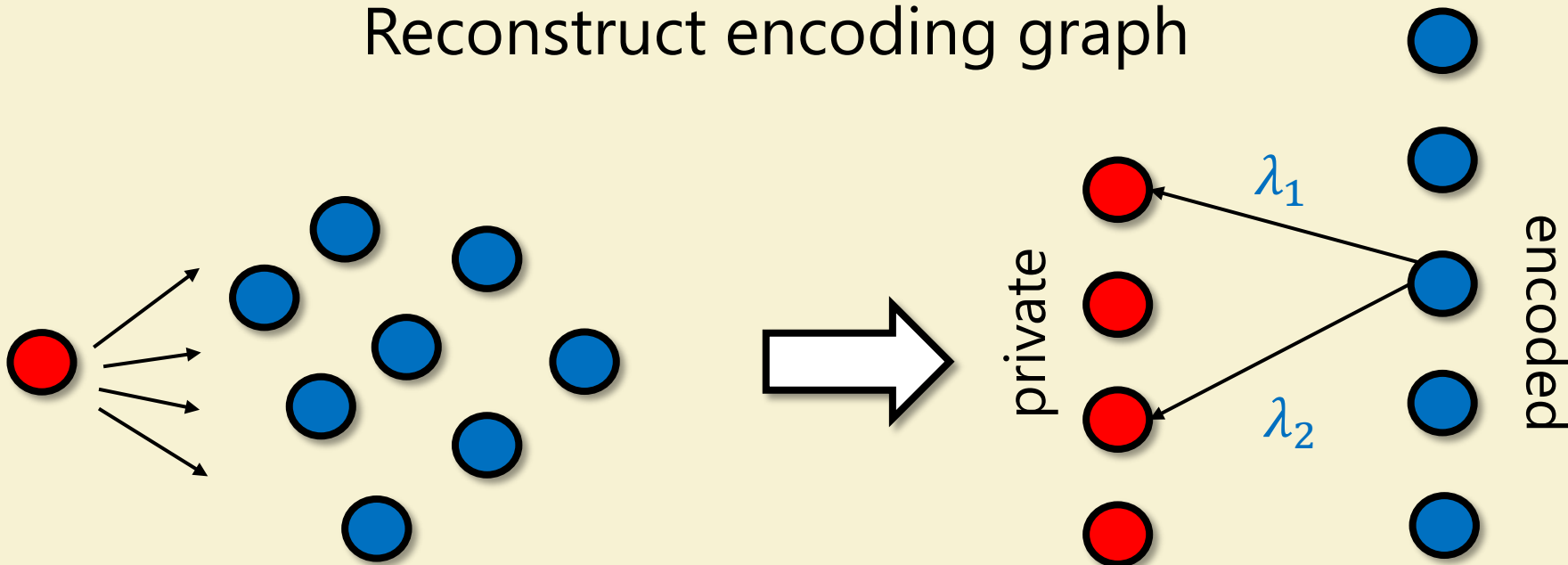
$x_i$  = R/G/B value of pixel, normalized to  $[-1, +1]$

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

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



## Reconstruct encoding graph

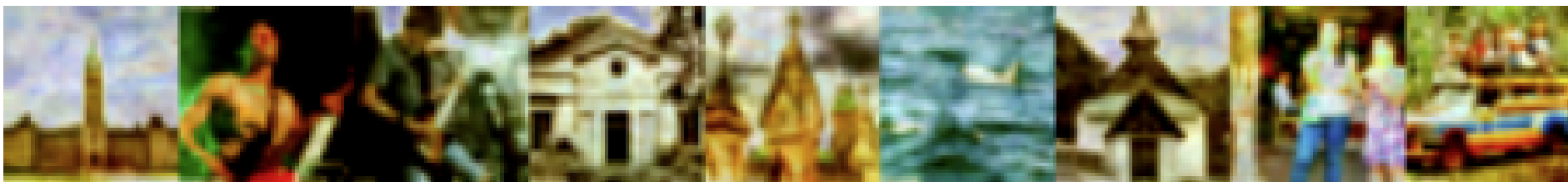


All came from same original private image

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

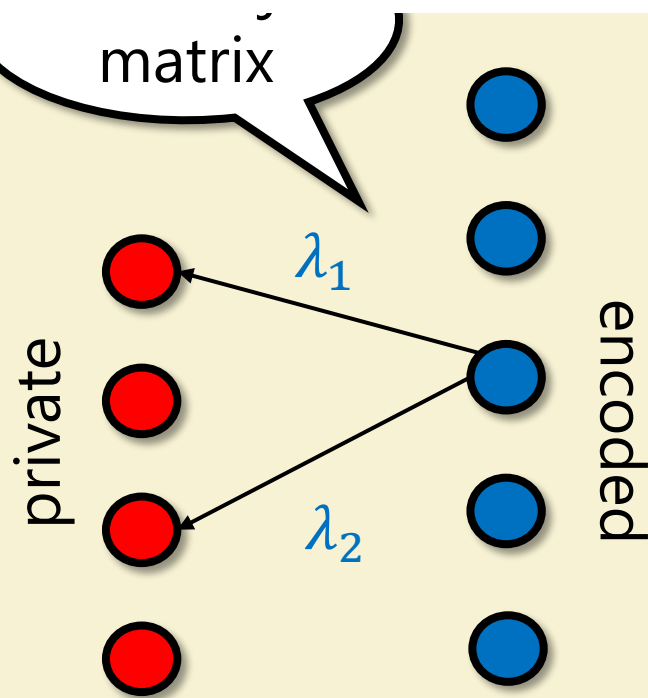
# At

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.



**InstaHide challenge:**

100 private images

5000 encoded images

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

Use GD to find  $\arg \min_{X \in [-1,1]^{n \times t}} \| \text{abs}(AX) - \tilde{X} \|^2$

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

# Black Box recovery

## Cryptanalytic Extraction of Neural Network Models

Nicholas Carlini<sup>1</sup>

Matthew Jagielski<sup>2</sup>

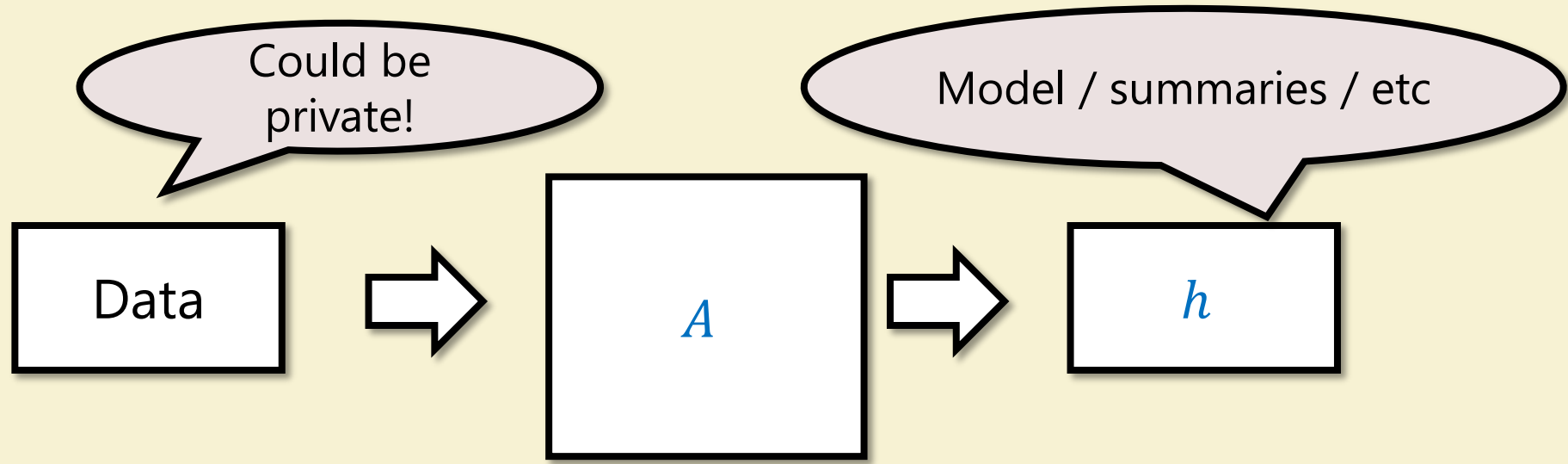
Ilya Mironov<sup>3</sup>

Architecture	Parameters	Approach	Queries	$(\epsilon, 10^{-9})$	$(\epsilon, 0)$	$\max  \theta - \hat{\theta} $
784-32-1	25,120	[JCB <sup>+</sup> 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 <sup>+</sup> 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  $\dagger$  were unable to recover the network after ten attempts.



# Learning



Solutions:

- **Cryptographic:** 100% privacy but at efficiency/control cost
- **Differential privacy:** "X% privacy" but X vs utility tradeoff not great
- **Heuristic:** Hope for 100%, might get 0%



