Can Individuals Trust Privacy Mechanisms for Machine Learning?

A Case Study of Federated Learning

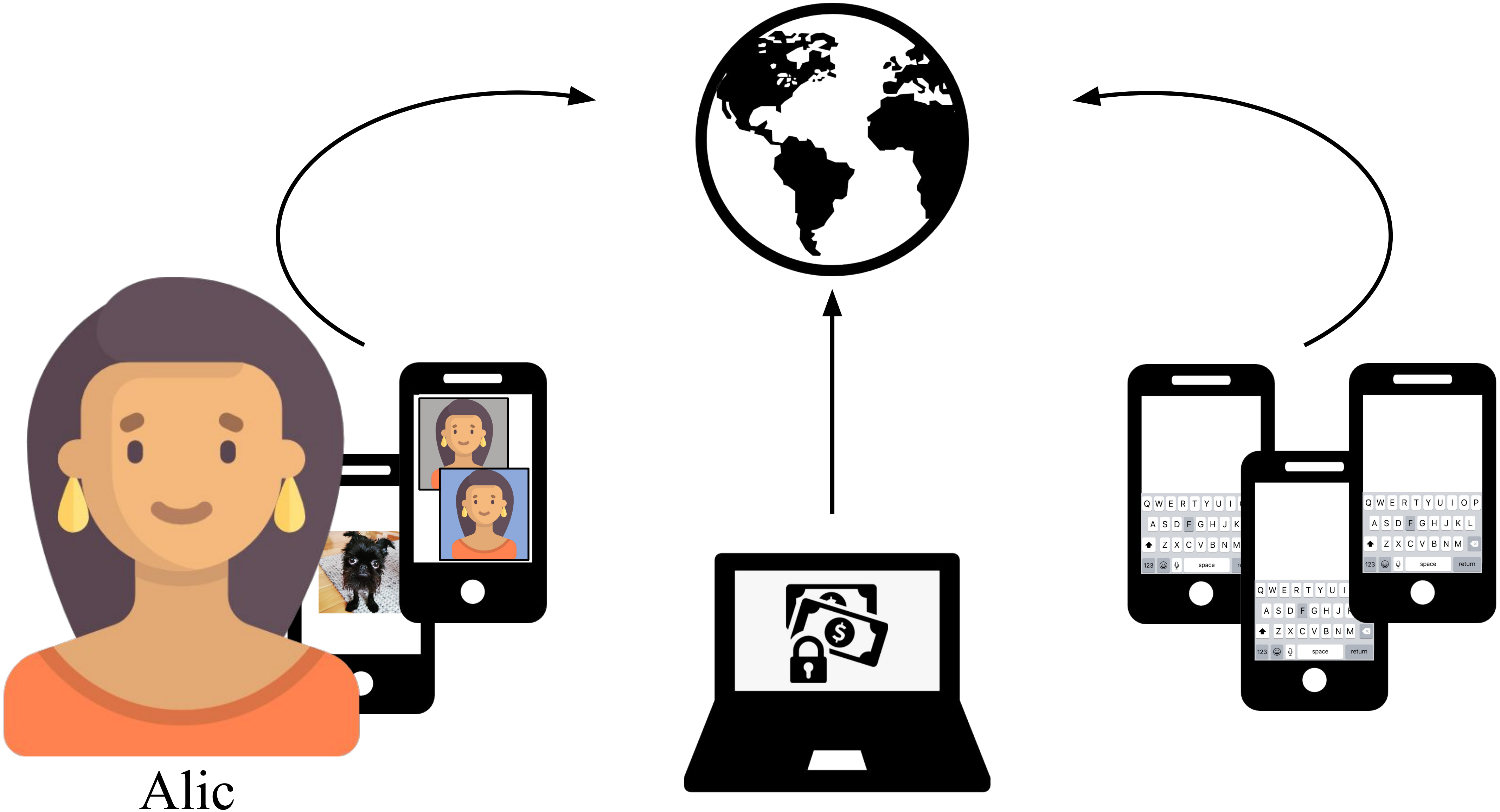
Franziska Boenisch
franziska-boenisch.de

ML in PL Conference
26 - 29 OCTOBER 2023

Apply for Open Positions in my

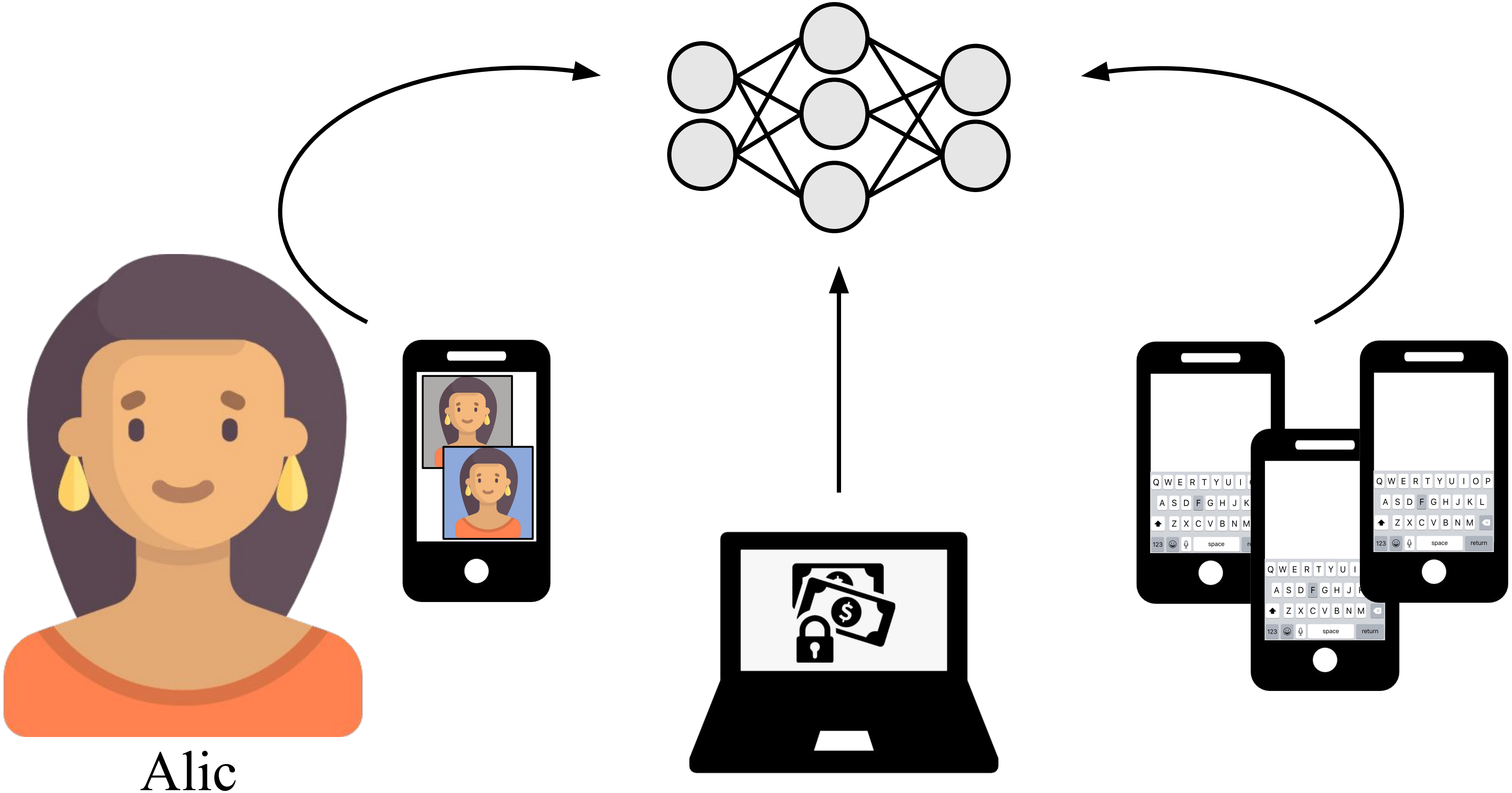
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Individuals Generate Sensitive Data



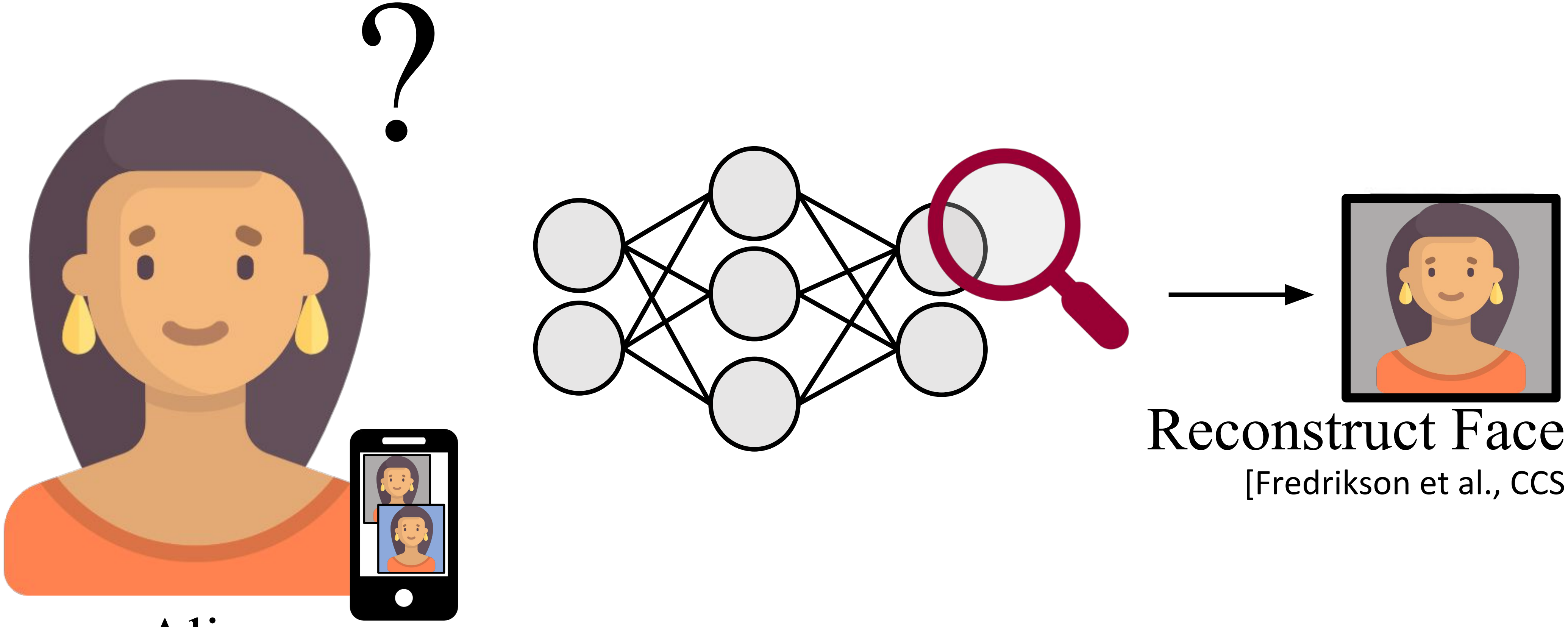
Alic
e

Companies apply Machine Learning



Alice

ML Models Leak Private Information

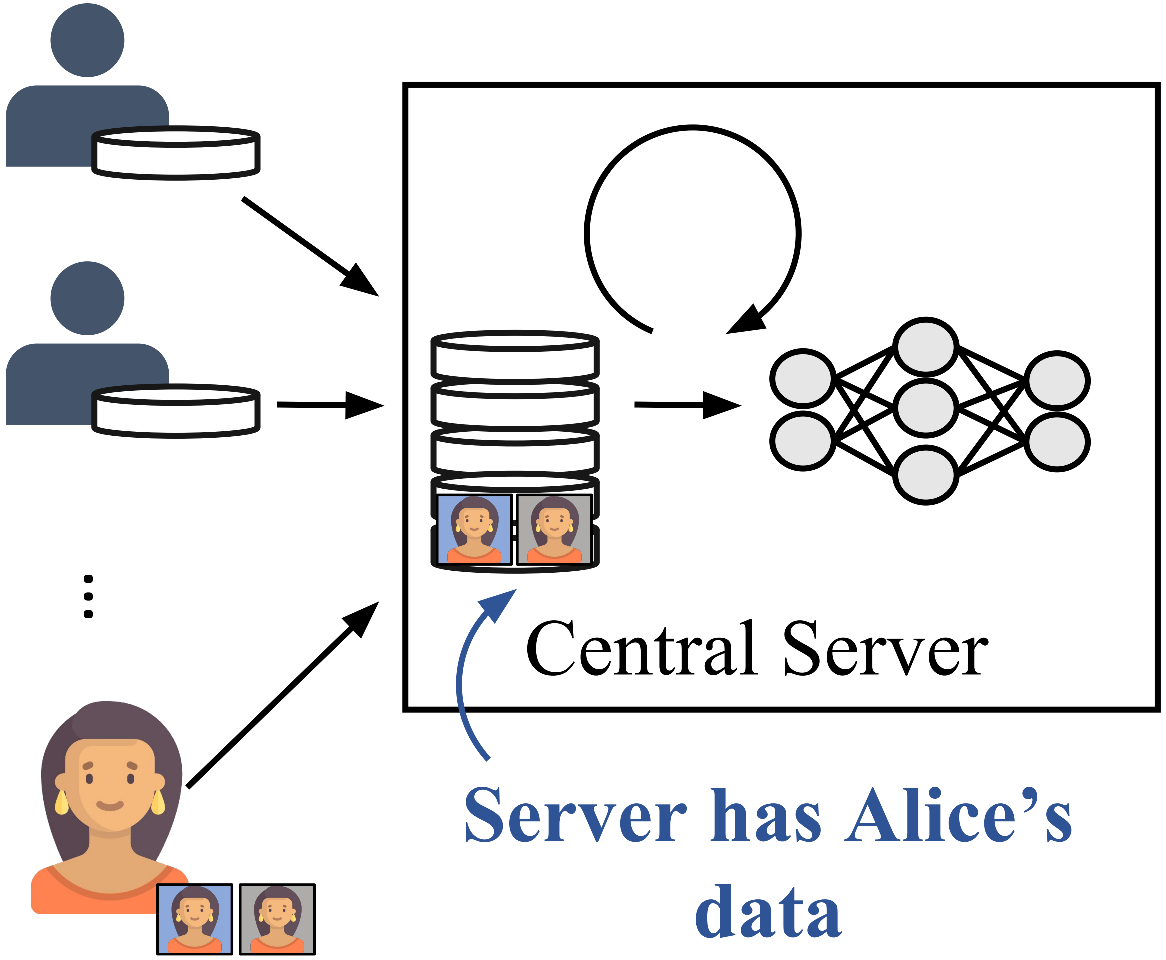


Alic
e

Reconstruct Face Images

[Fredrikson et al., CCS 2015]

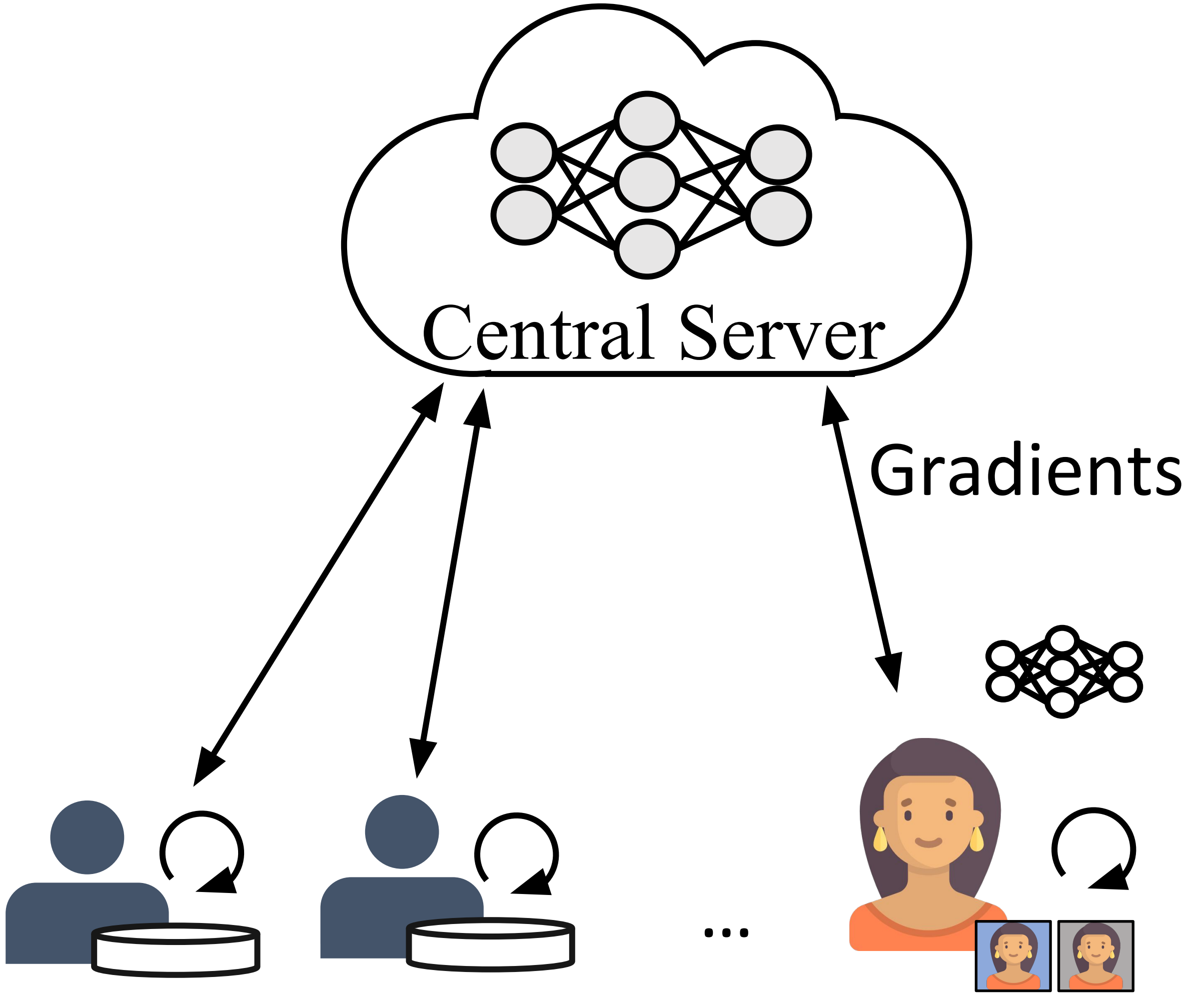
Centralized vs. Federated Learning



Central Server

Server has Alice’s data

Centralized Learning



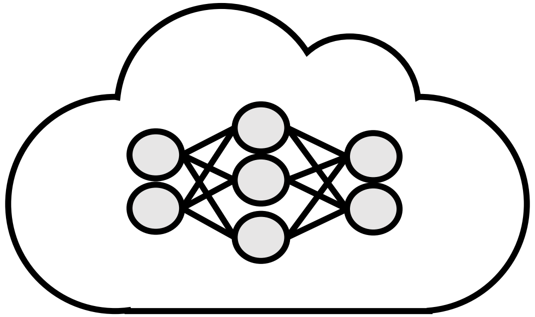
Central Server

Gradients

…

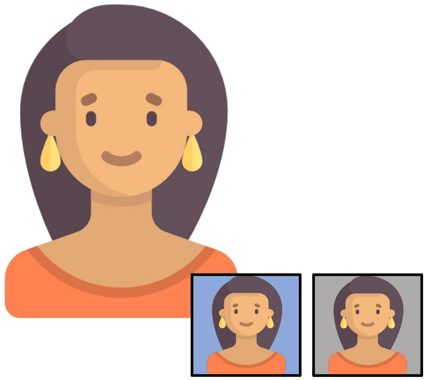
Federated Learning

Key Properties of Federated Learning



Central Server

+ Heterogenous data
+ Efficient communication
+ Low costs



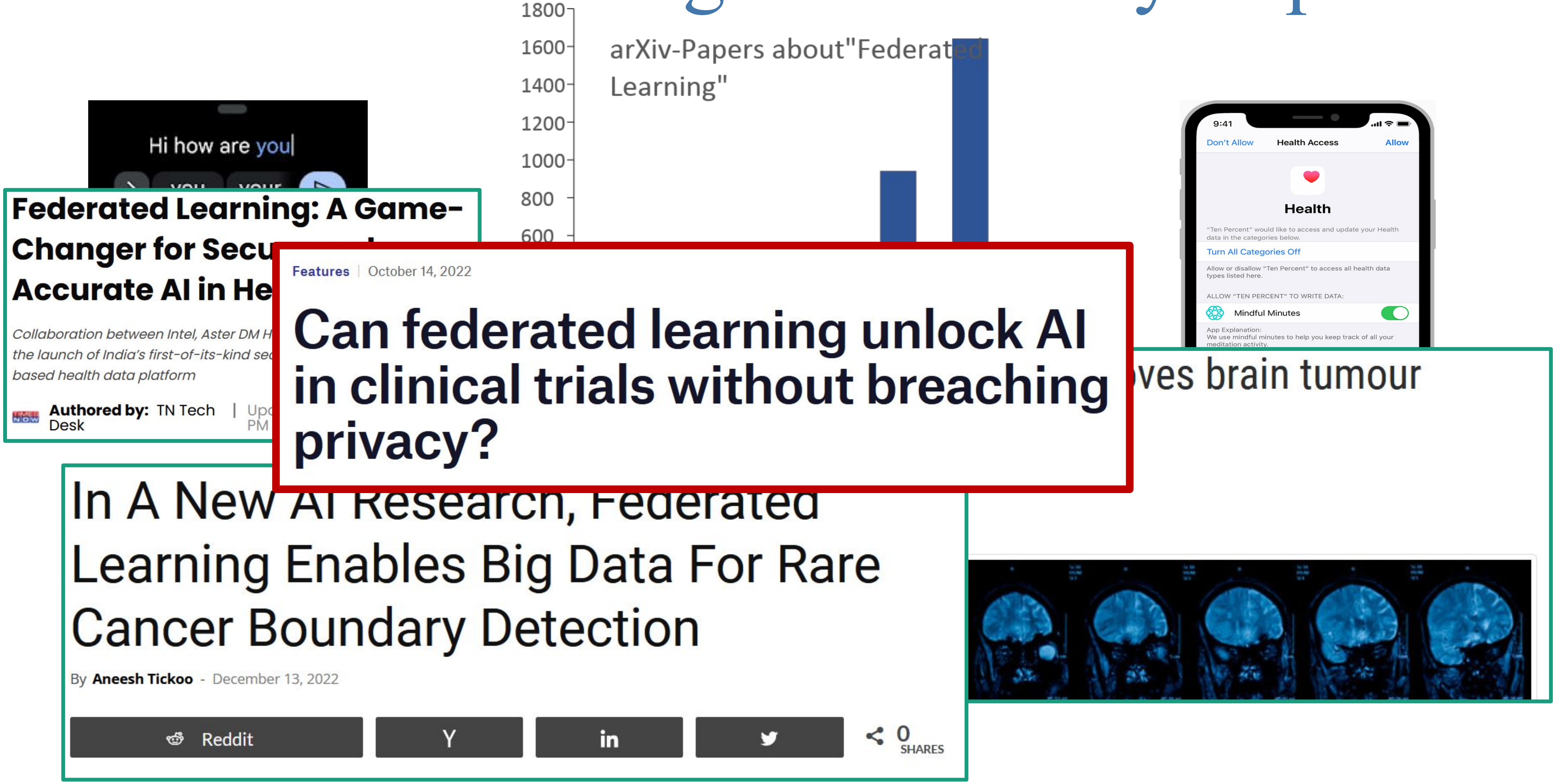
Individual User

- Performs compute
- Provides storage
+ Keeps data locally

Privacy?!

?

Federated Learning is Extremely Popular



Hi how are you

Federated Learning: A Game-Changer for Secu… Accurate AI in He…

Collaboration between Intel, Aster DM H… the launch of India's first-of-its-kind se… based health data platform

Authored by: TN Tech Desk | Upd… PM

Features | October 14, 2022

Can federated learning unlock AI in clinical trials without breaching privacy?

9:41

Don't Allow Health Access Allow

Health

"Ten Percent" would like to access and update your Health data in the categories below.

Turn All Categories Off

Allow or disallow "Ten Percent" to access all health data types listed here.

ALLOW "TEN PERCENT" TO WRITE DATA:

Mindful Minutes

App Explanation:
We use mindful minutes to help you keep track of all your meditation activity.

…ves brain tumour

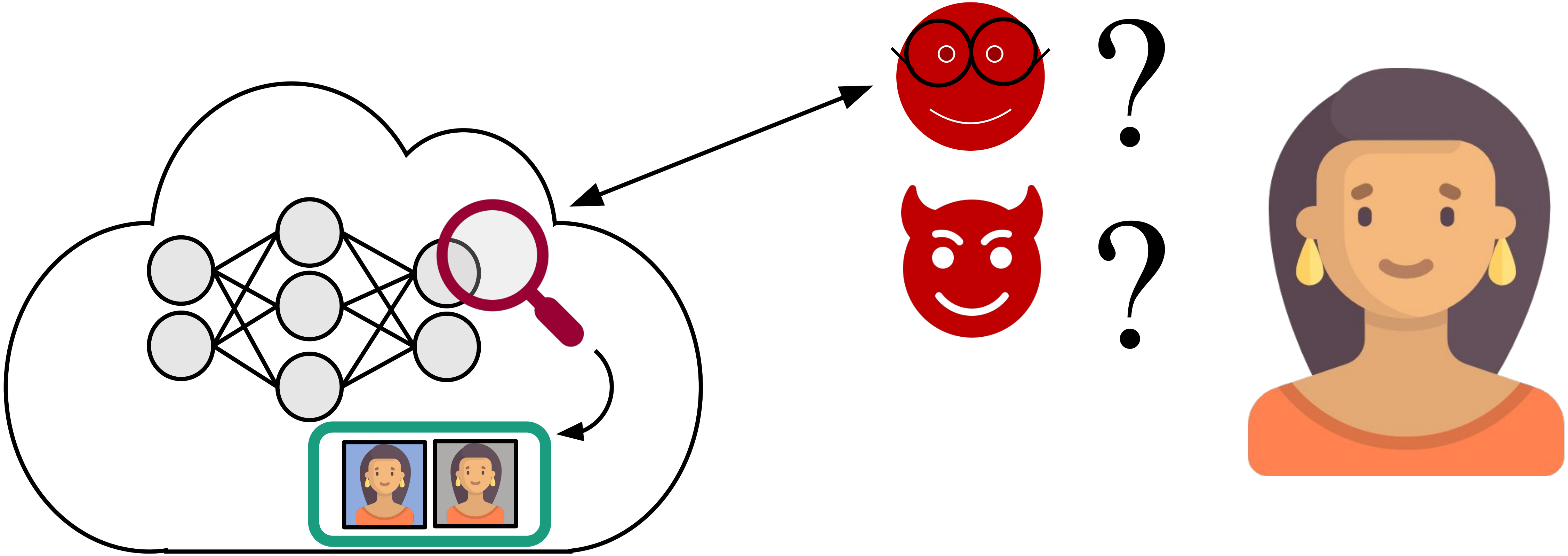
In A New AI Research, Federated Learning Enables Big Data For Rare Cancer Boundary Detection

By **Aneesh Tickoo** - December 13, 2022

Reddit Y in

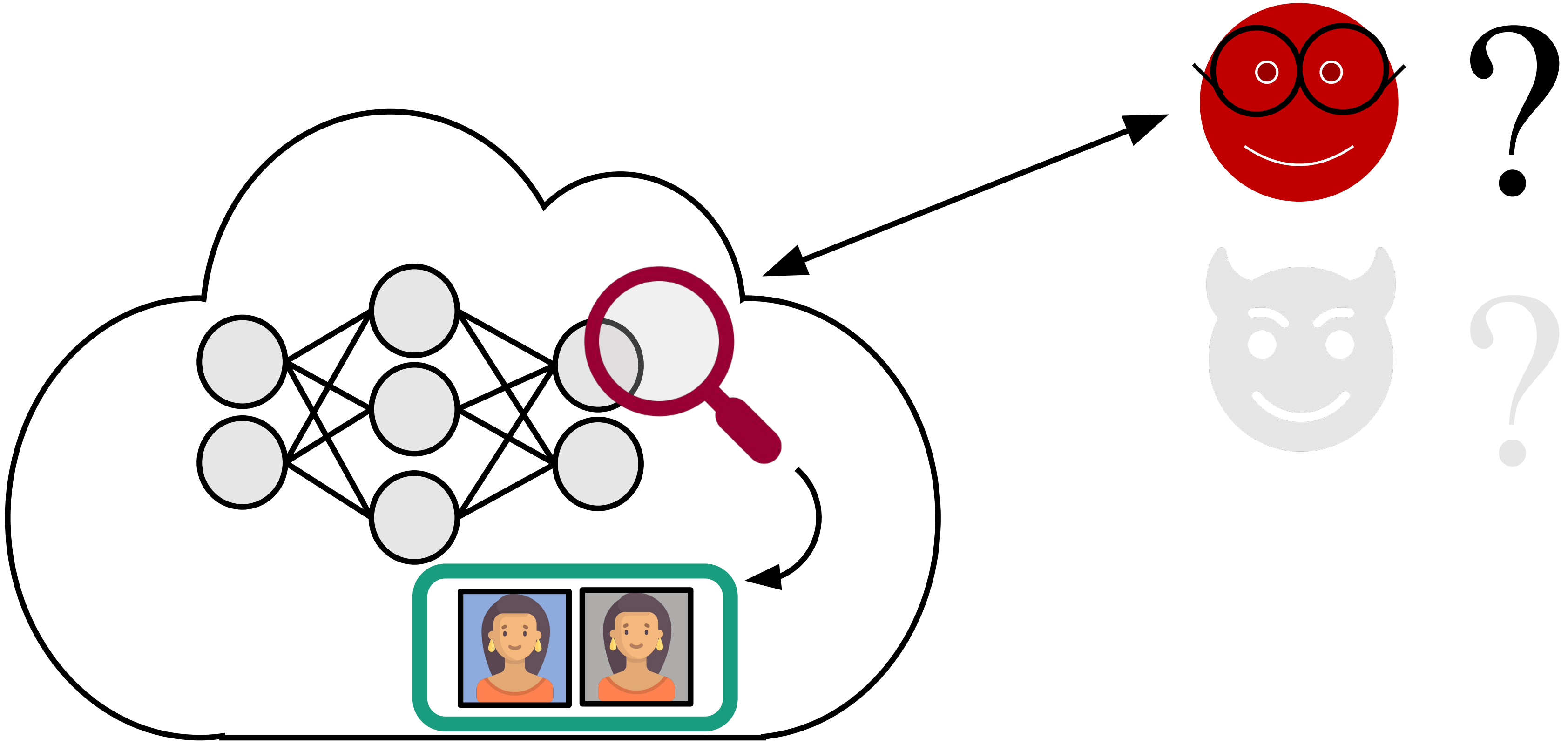
0 SHARES

What Trust Model is Needed for Privacy?



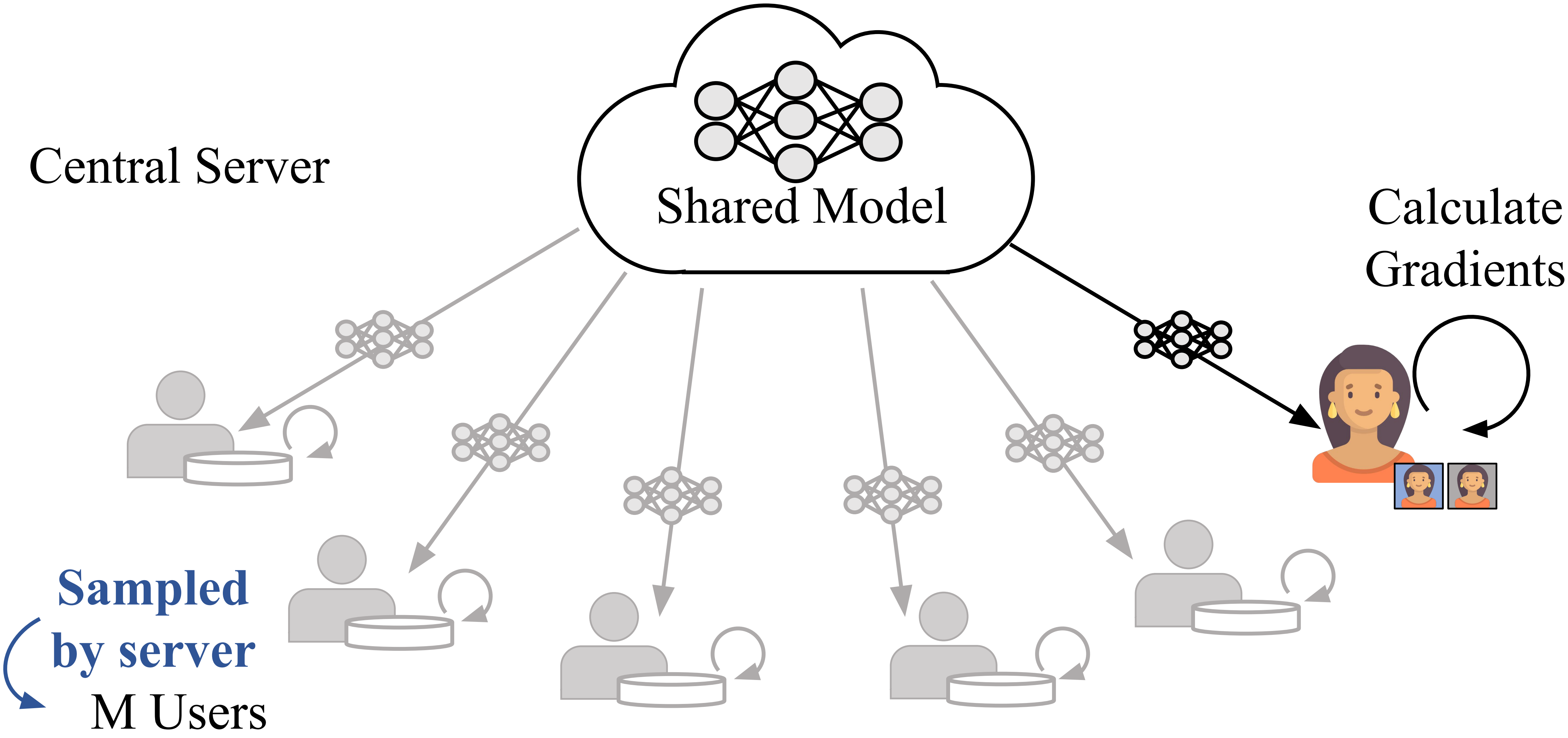
Federated Learning

What Trust Model is Needed for Privacy?

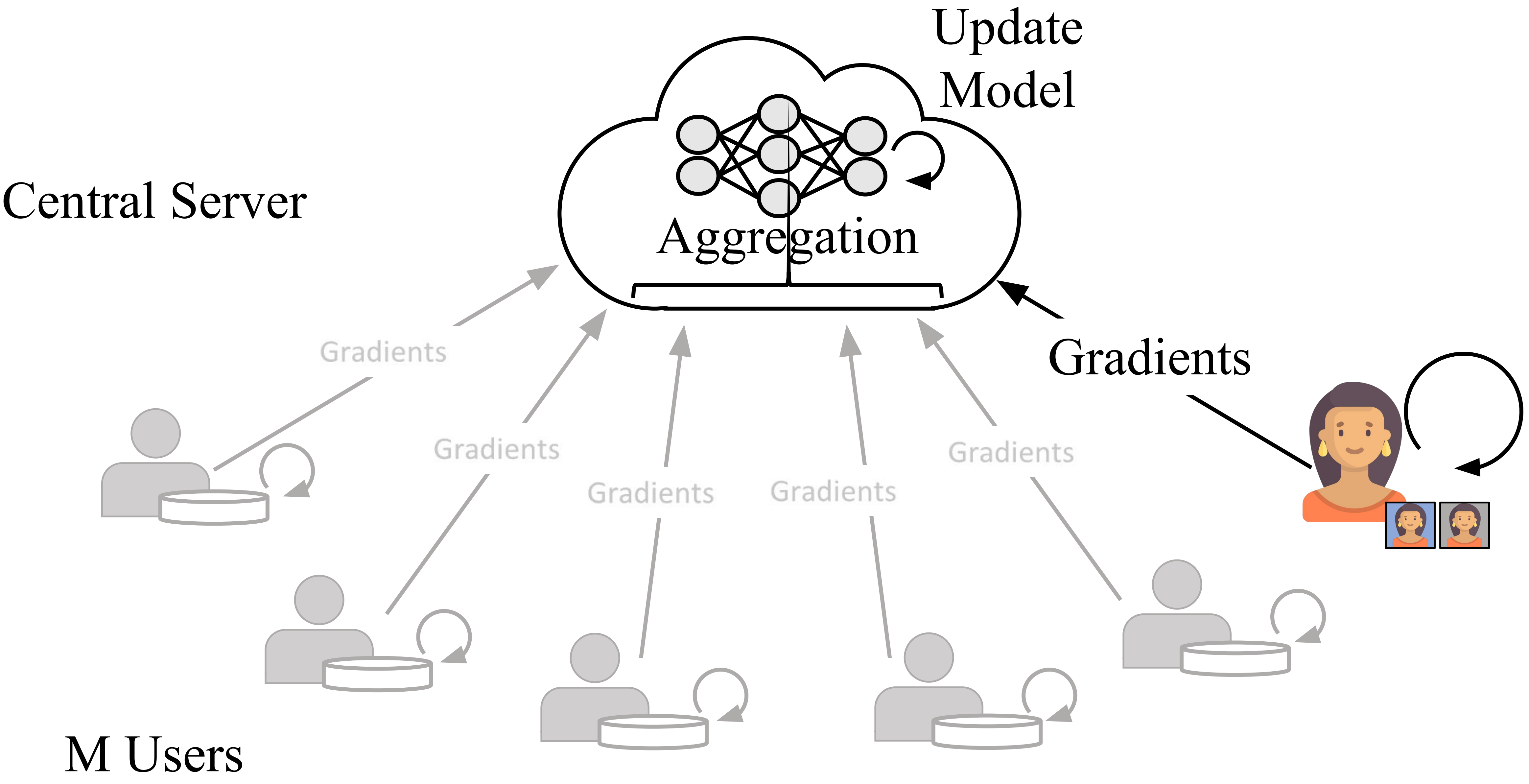


Federated Learning

Federated Learning



Federated Learning



Update Model

Central Server

Aggregation

Gradients

Gradients

Gradients

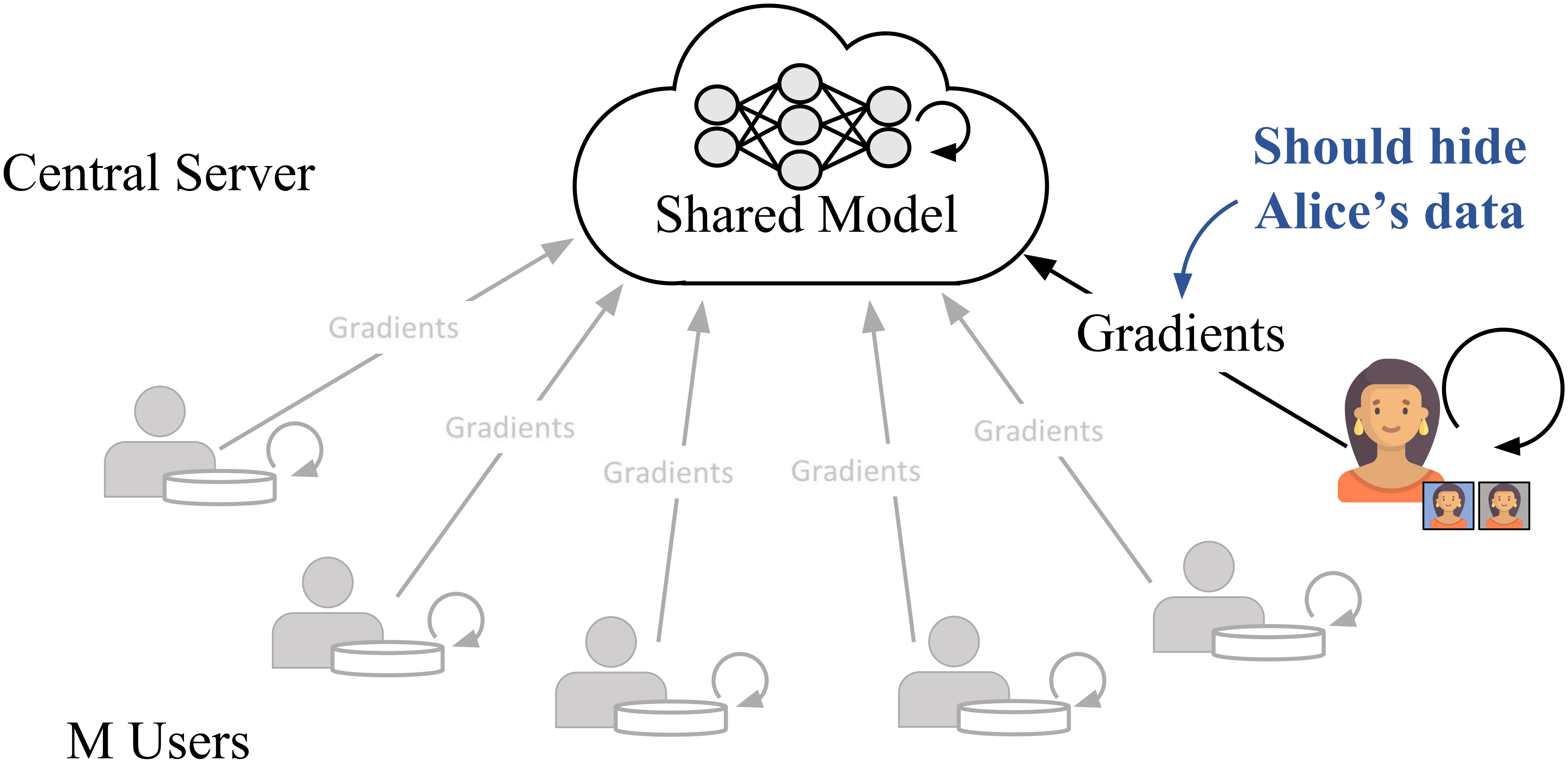
Gradients

Gradients

Gradients

M Users

Alice’s Privacy Relies purely on the Gradients



Prior Data Reconstructions Attacks are Limited

We can reconstruct data…

… from different classes

… from small mini-batches

… that is of

low-complexity

… at high computational costs

[Zhu et al., NeurIPS 2019]

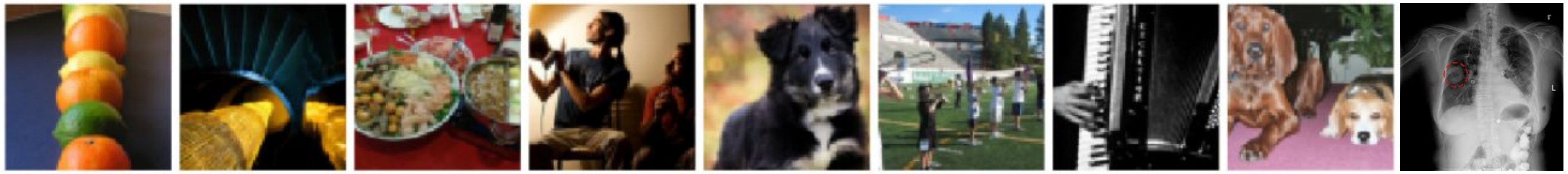
We can extract data:

… from mini-batches of size = 1

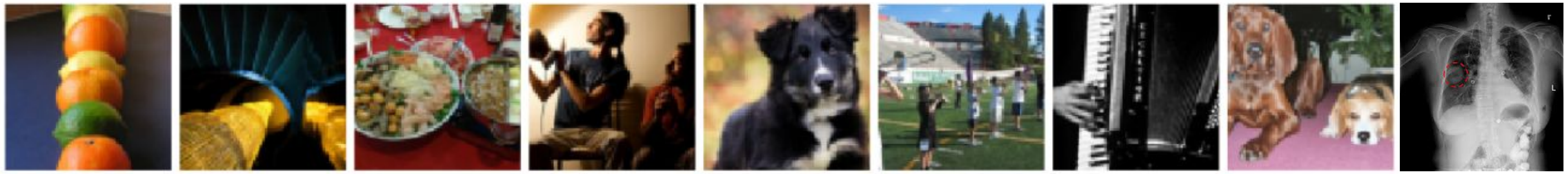
[Geiping et al., NeurIPS 2020]

We Extract Large Amounts of Data Perfectly

Original Data



Extracted Data



… from all kinds of class distribution

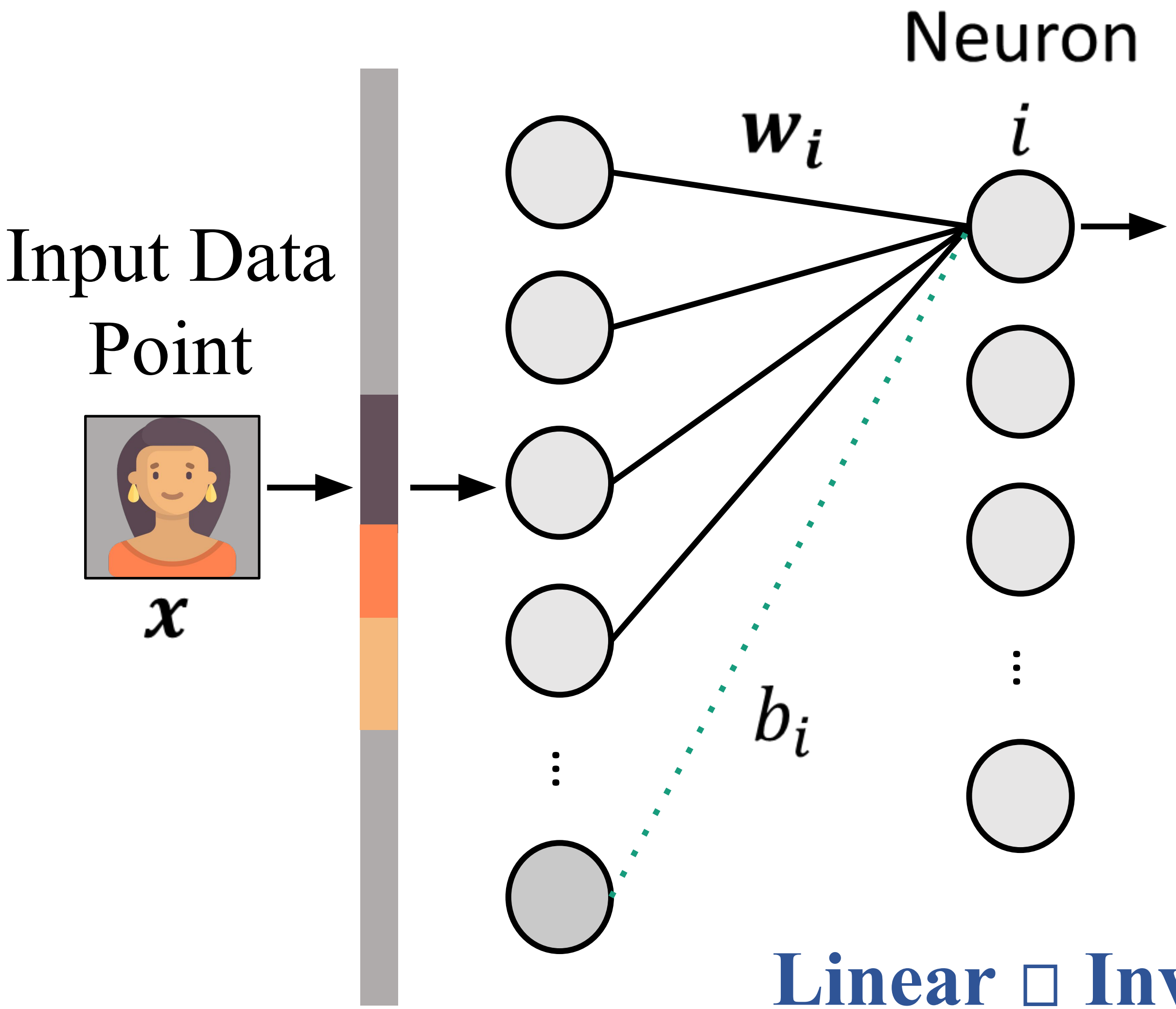
… from large mini-batches with hundreds of data points

… with high complexity

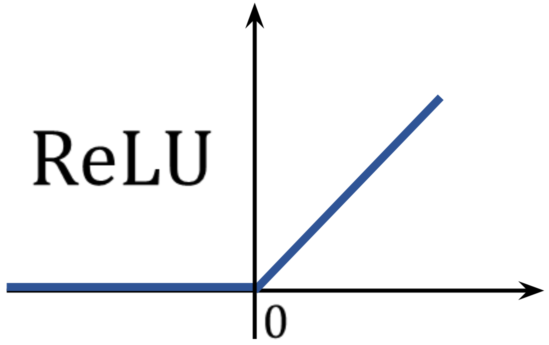
… at near-zero computational costs

Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, Nicolas Papernot. *When the Curious Abandon Honesty: Federated Learning Is Not Private*, 2021. [IEEE Euro S&P ’23a]

Forward Pass through Fully-Connected Layer



Output: $y_i = \mathrm{ReLU}(\boldsymbol{w}_i^T \boldsymbol{x} + b_i)$



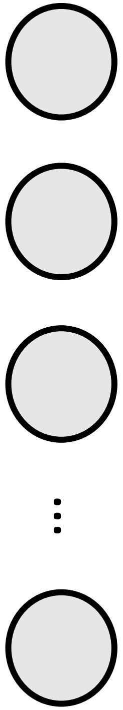
$$y_i = \boldsymbol{w}_i^T \boldsymbol{x} + b_i$$

Linear □ Invertible

$$\mathit{if}\ \boldsymbol{w}_i^T \boldsymbol{x} + b_i > 0$$

Prior Extraction Works only for Single Data Points

$$\rightarrow \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_i^T} = \frac{\partial \mathcal{L}}{\partial b_i}\boldsymbol{x}$$



$$y_i = \boldsymbol{w}_i^T \boldsymbol{x} + b_i$$

$$\frac{\partial y_i}{\partial \boldsymbol{w}_i^T} = \boldsymbol{x}$$

$$\frac{\partial y_i}{\partial b_i} = 1$$

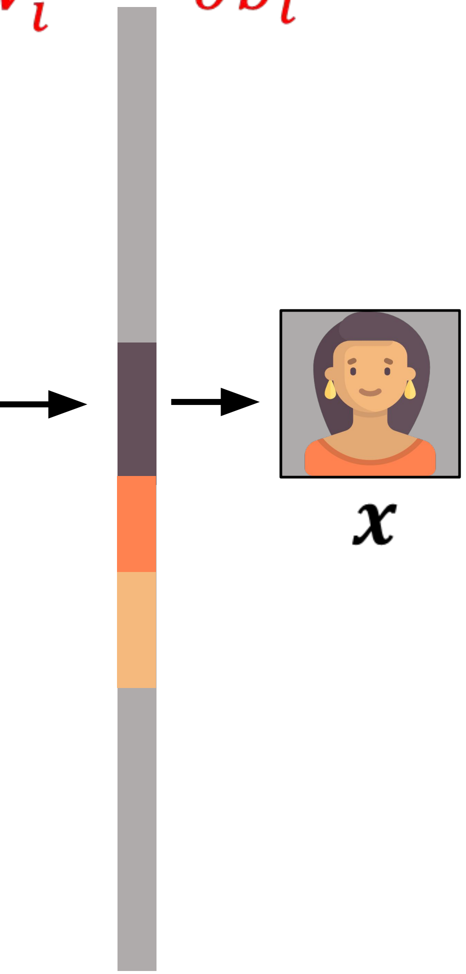
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_i^T}$$

$$\frac{\partial \mathcal{L}}{\partial b_i}$$

$$=$$

Contains scaled input data point

Contains scaling factor



[Geiping et al., NeurIPS 2020]

Extraction for Large Mini-Batches Should Fail

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i^T} = \sum_{j=1}^{B} \frac{\partial \mathcal{L}}{\partial y_{i,j}} \frac{\partial y_{i,j}}{\partial \mathbf{w}_i^T}$$

Mini-batch gradient

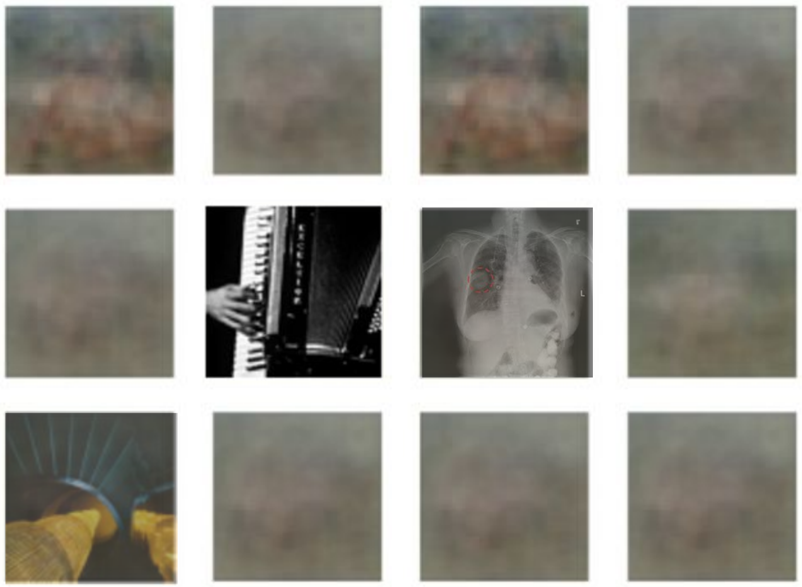
We believe rescaled gradients look like this….

Data Leaks Directly from Model Gradients

```
weights_gradient = gradients[0].numpy()
inverse_bias = 1 / gradients[1].numpy()
extracted_data = inverse_bias * weights_gradient
plot(extracted_data, num_rows = 3, num_cols = 6)
```

$$\boldsymbol{x} = \left(\frac{\partial \mathcal{L}}{\partial b_i}\right)^{-1} \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_i}$$

All you need is matplotlib

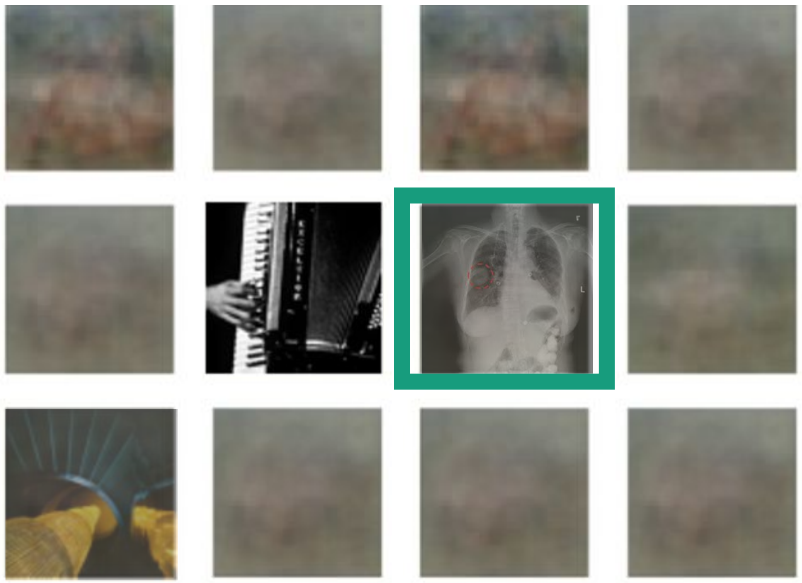


… but they actually look like that!

mini-batch size=100

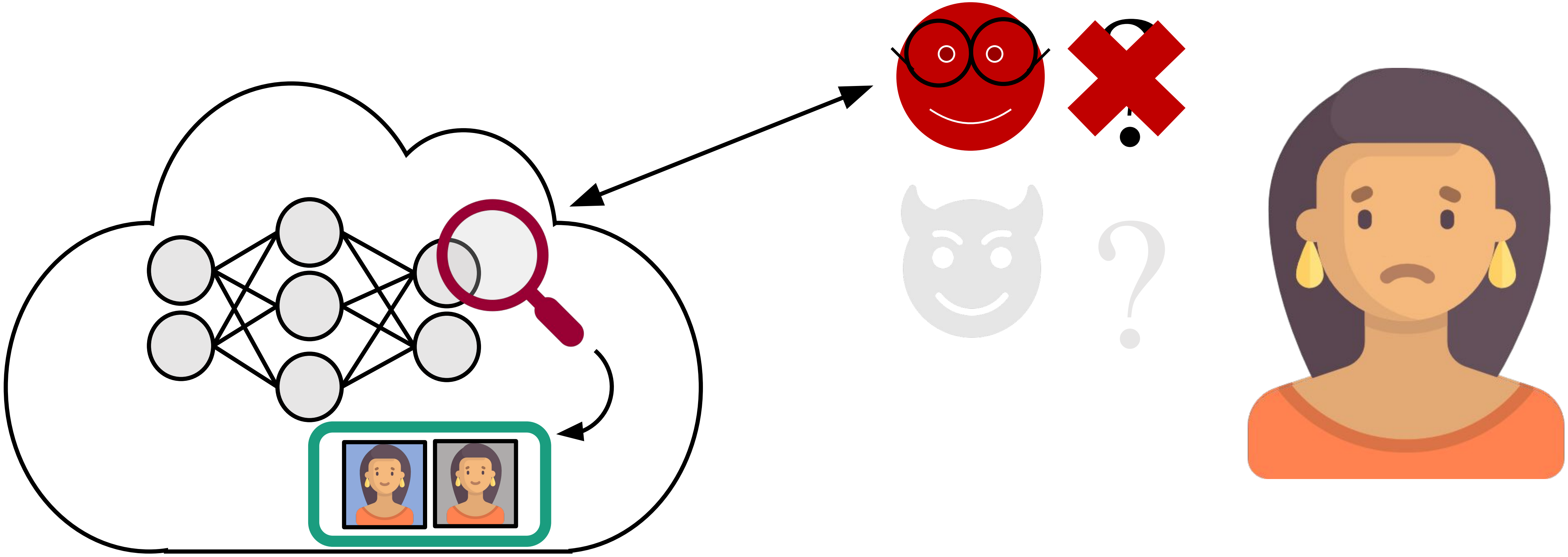
Gradients can Leak Single Data Points

Why can we still extract individual data points $\boldsymbol{x}$?



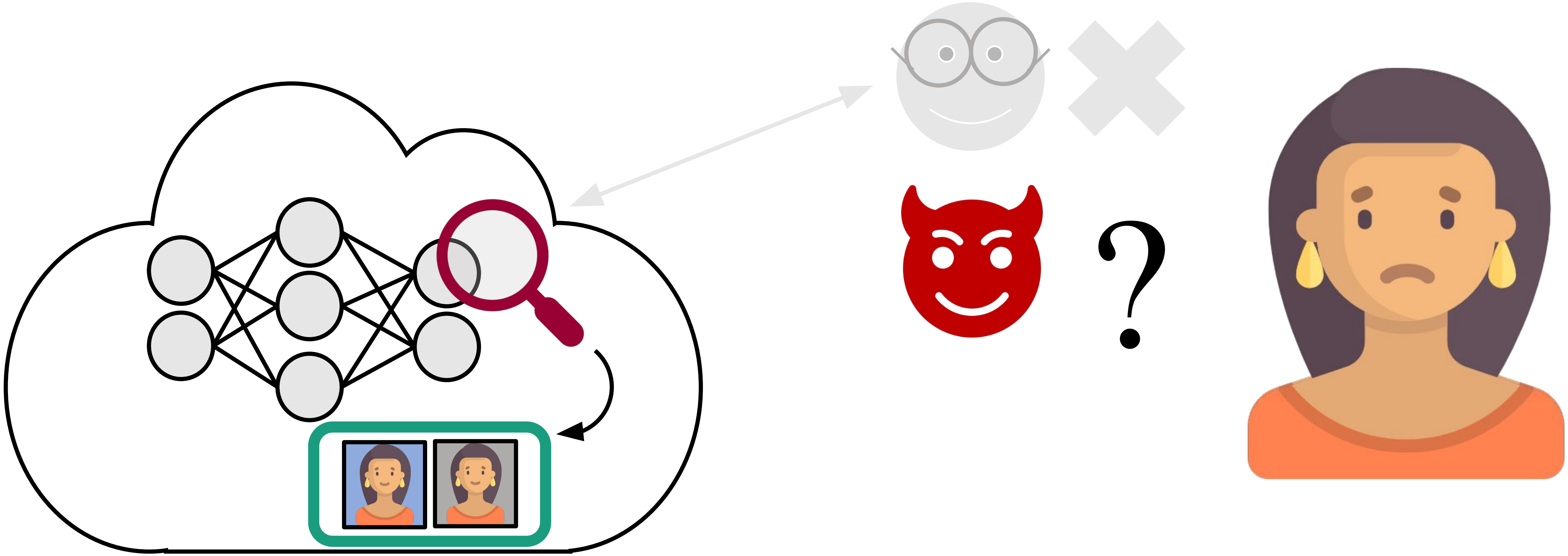
Gradient of a single data point

What Trust Model is Needed for Privacy?



Even a passive, honest-but-curious attacker can extract a significant amount of sensitive user-data.

What Trust Model is Needed for Privacy?

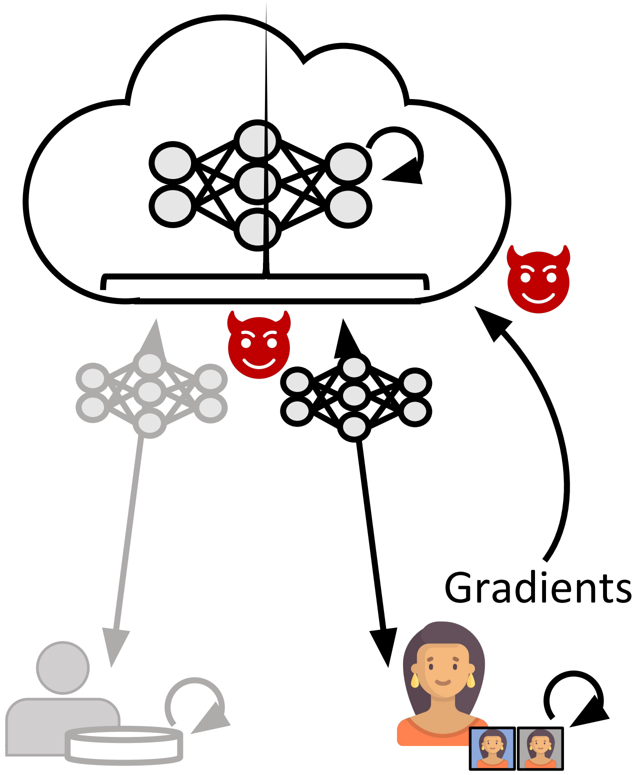


Even a passive, honest-but-curious attacker can extract a significant amount of sensitive user-data.

Our Trap Weights Increase Natural Leakage

Trap Weights: Induce $\boldsymbol{x}^T \boldsymbol{w}_i + b_i \leq 0$ for most input data points $\boldsymbol{x}$

Makes other points extractable



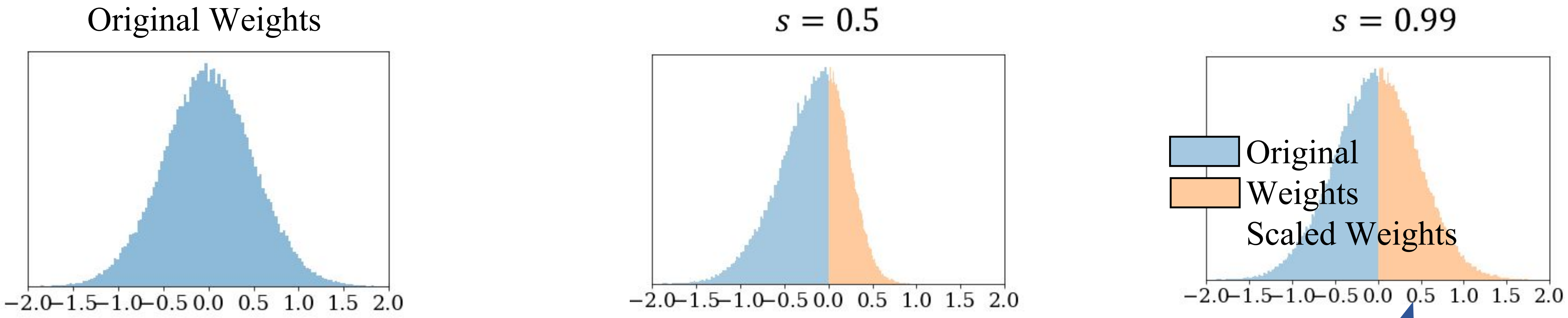
1) Initialize model weights at random
2) Scale positive components down by $s < 1$

→ $(\boldsymbol{x}^T s \boldsymbol{w}_i^+) + (\boldsymbol{x}^T \boldsymbol{w}_i^-) + b_i \leq 0$ more often

Assumes input features $\boldsymbol{x}$ in range [0, 1]

Standard pre-processing

Influence of Scaling Factor “s”



Inconspicuous

Scaling Factor (s)	Activated Neurons ! (by 1 data point) (%)	Extracted Data (%)
0.4	0	0
0.5	0	0
0.9	0	0
0.99	65.5 (51.4)	45.7
1.0	99.9 (4.4)	21.8

Active Extraction

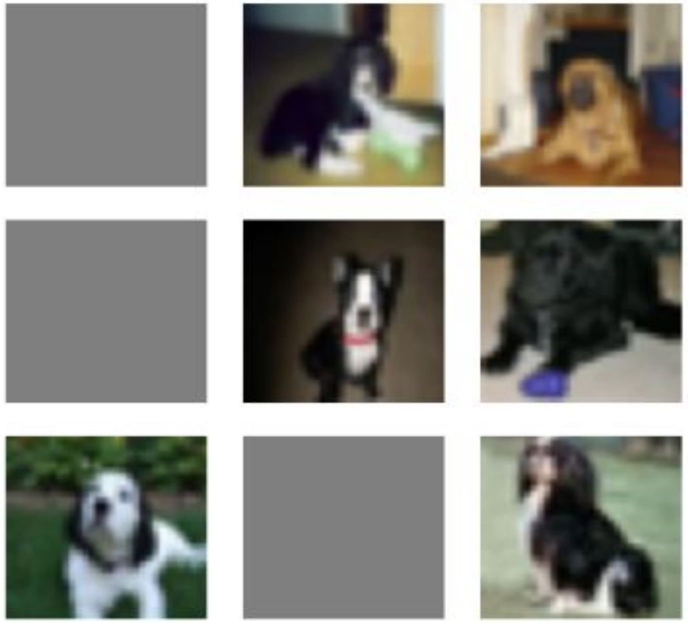
Baseline: Passive Extraction

ImageNet Extraction: Mini-Batch Size = 100, 1000 Neurons

Our Trap Weights Improve Extraction

	Passive	Active
MNIST	5.8	54
CIFAR10	25.5	54
ImageNet	21.8	45.7
IMDB	25.4	65.4

Extracted Data (%),
Mini-Batch Size = 100,
1000 Neurons

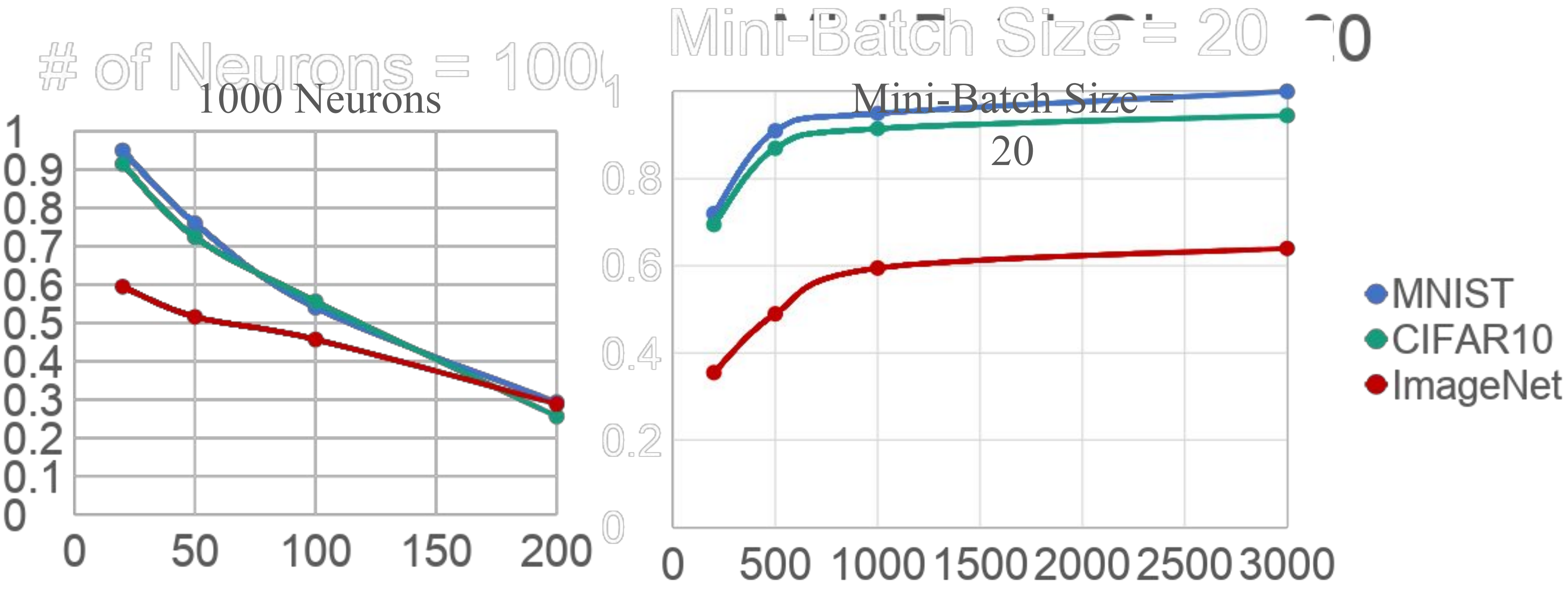


CIFAR10
(Non-IID)

Extracted from gradients within < 1 second

More Neurons and Smaller Mini-Batches Let us Extract More Data

Extraction Recall

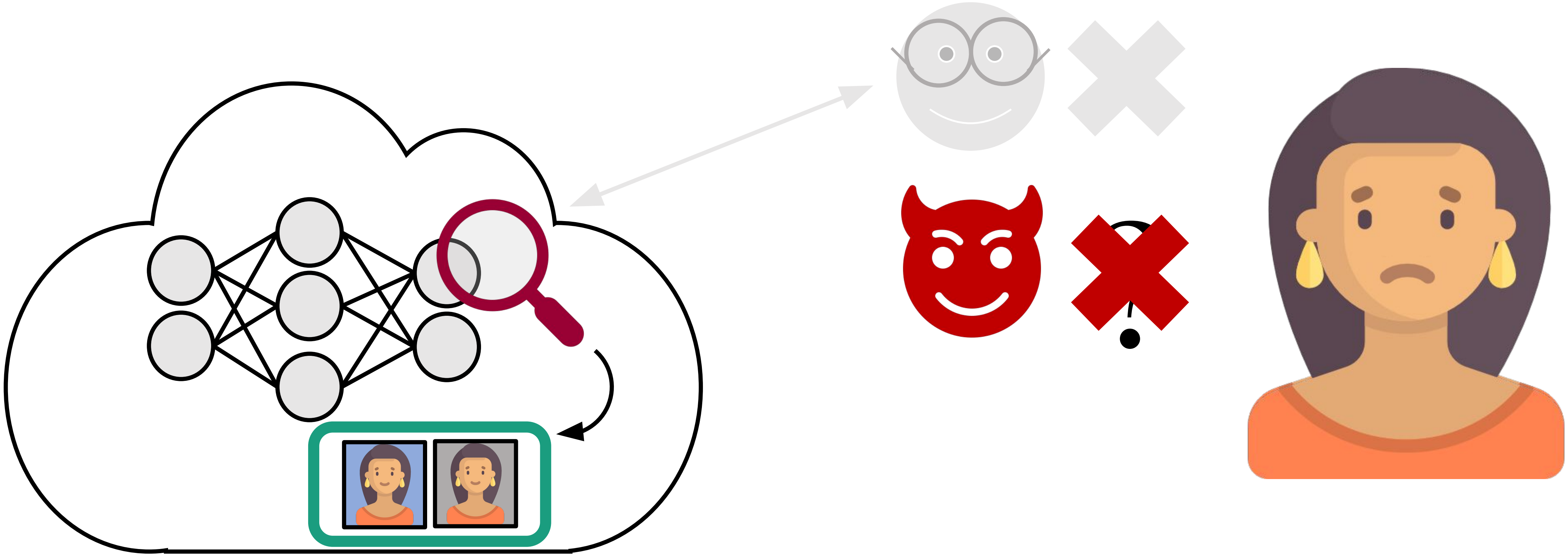


Mini-Batch Size

# of Neurons

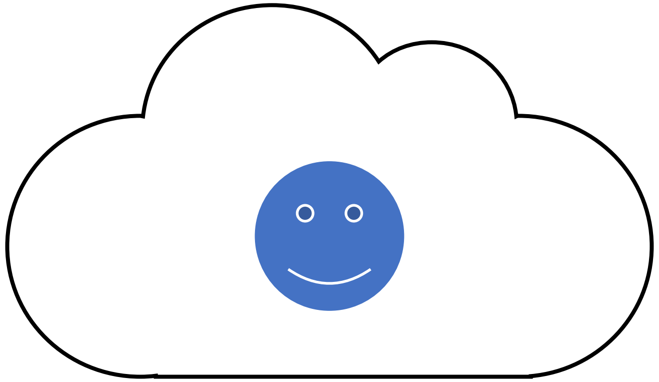
Specified by the server

What Trust Model is Needed for Privacy?

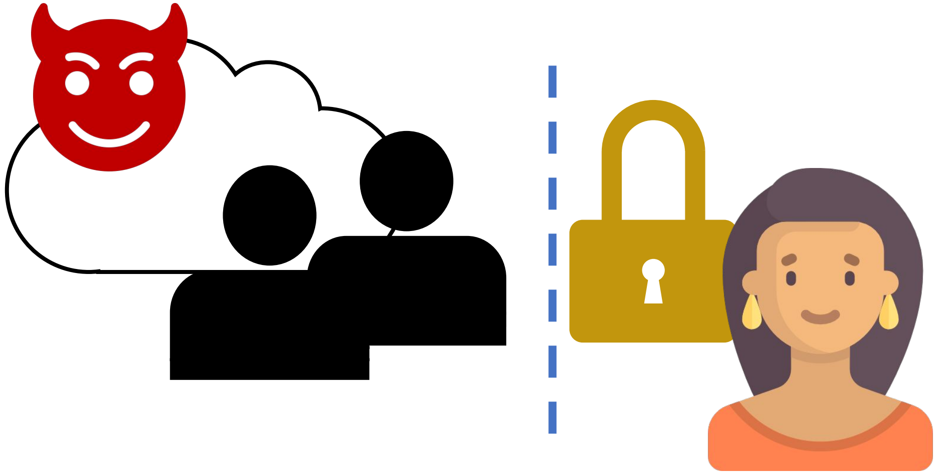


An active, malicious attacker can significantly increase privacy risks for users.

Conclusion for Privacy in FL



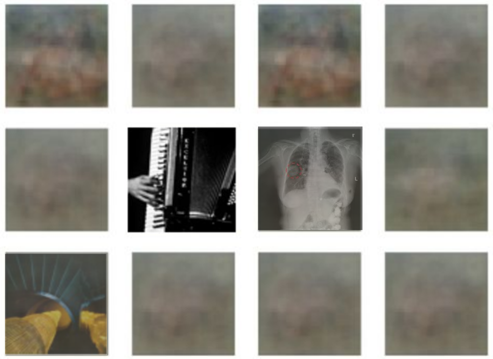
Participate only in Protocols with Trusted Server

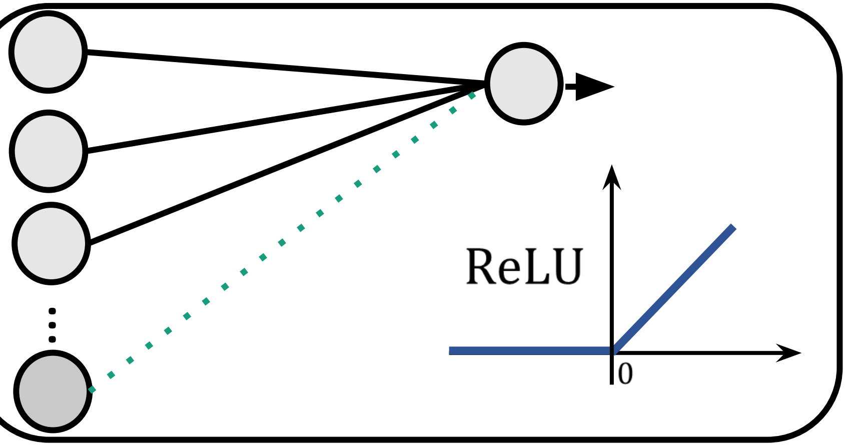


Replace Trust by Verifiable Mechanisms

Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, Nicolas Papernot. *Is Federated Learning a Practical PET Yet?*, 2023. [IEEE Euro S&P '23a]

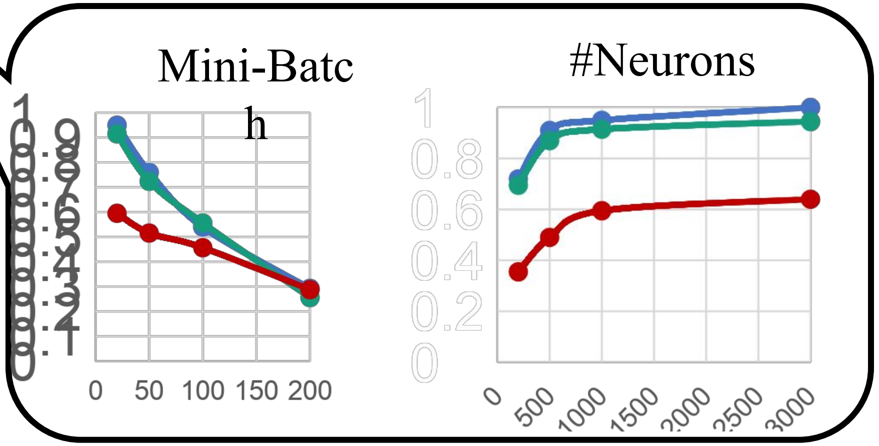
Thank you & I am looking for Collaborators!





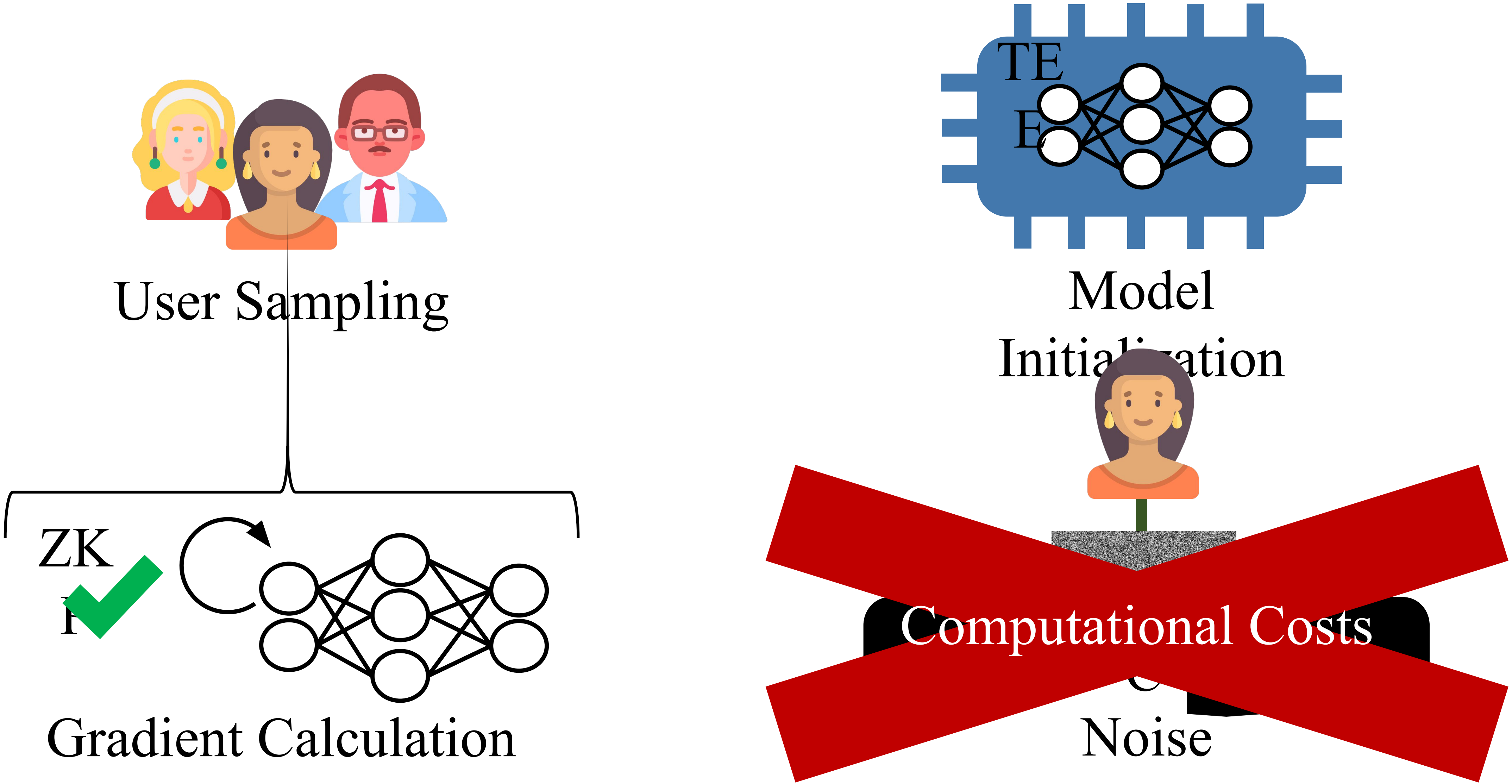
Collaborators wanted!

- PhD Students
- Interns
- Postdocs



Backup Slides

Defending FL is Complex and Costly



User Sampling

Model Initialization

Gradient Calculation and Aggregation

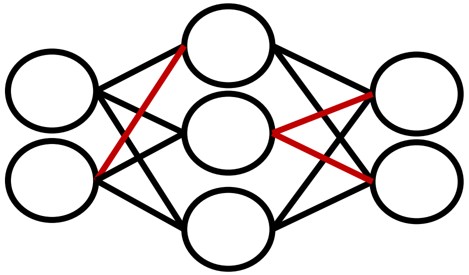
Computational Costs

Noise Addition

Power Imbalance Makes FL Vulnerable

Server wants Utility

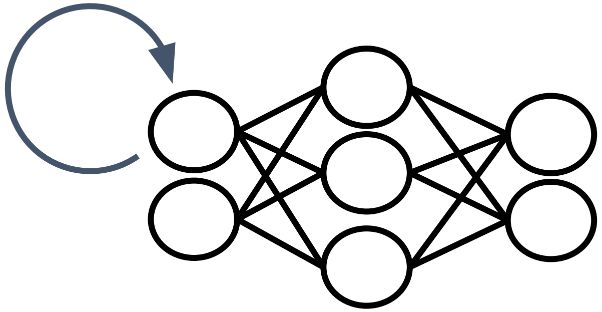
User Provisioning & Sampling



Model Manipulations

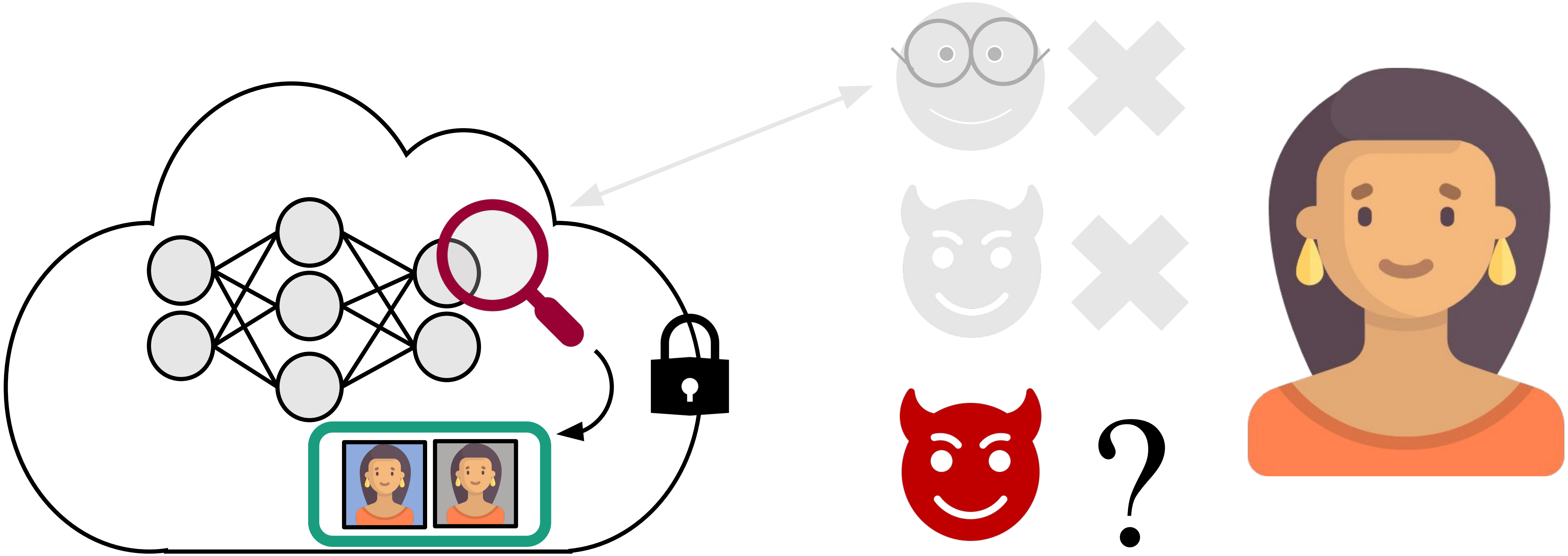
Users need Privacy

Unknown Collaborators



Unverified shared model and computations

What Trust Model is Needed for Privacy?

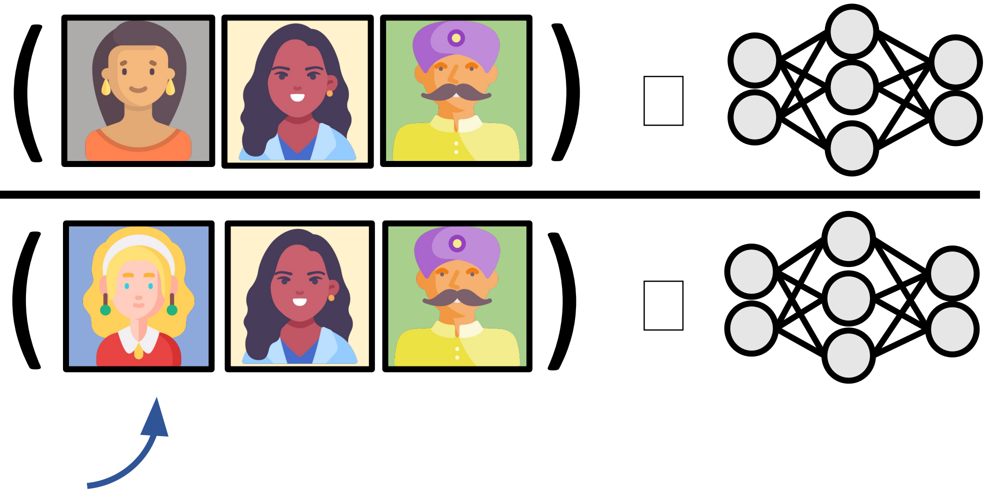


An active, malicious attacker can significantly increase privacy risks for users.

Differential Privacy Protects Individual Data

Pr

(Train

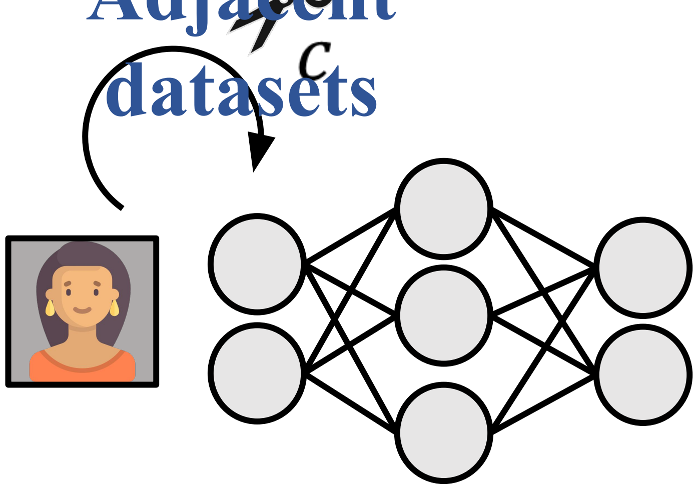


Pr

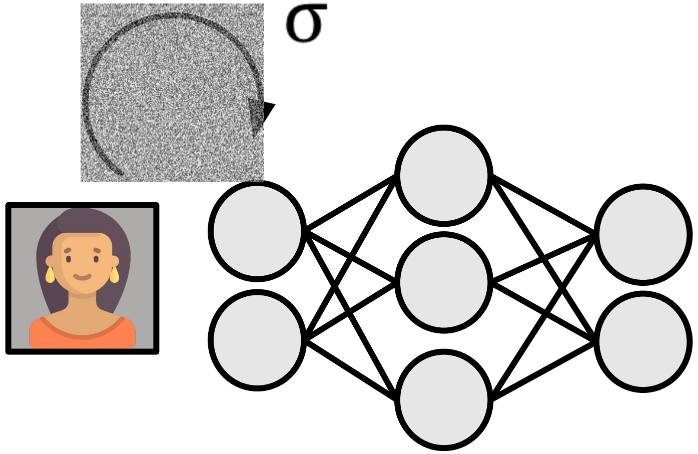
(Train

$\leq e^{\varepsilon}$

Adjacent datasets

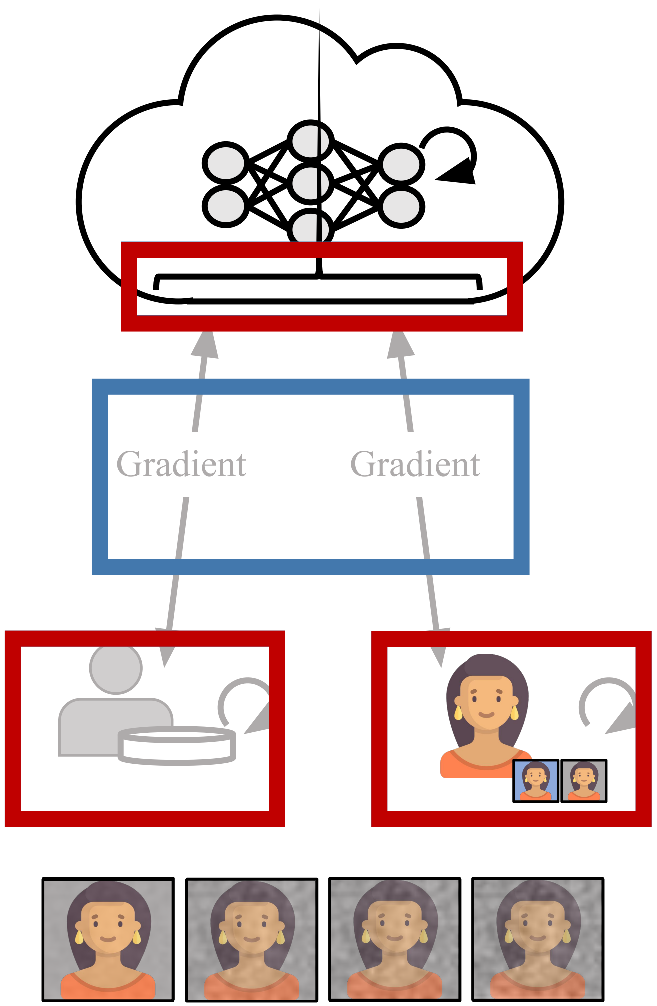


(1) Clip Gradients



(2) Noise Gradients

Differential Privacy in Federated Learning



Clipped Gradients

Noised Gradients

$\mathcal{N}(($

Central DP: Server adds noise

Distributed DP: Users add noise

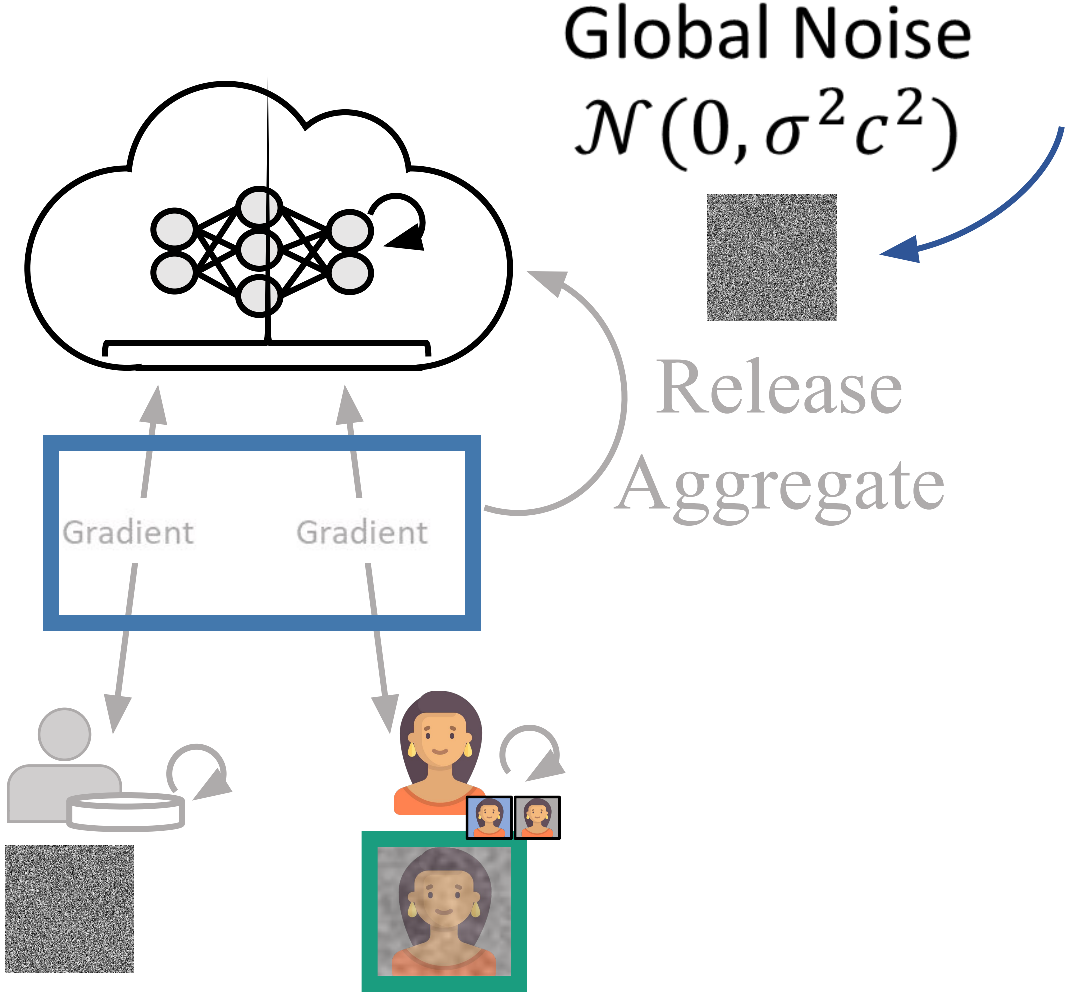
After aggregation

Local DP: Users add noise

$\mathcal{N}\left(0, \sigma^2 c^2\right)$

$\mathcal{N}\left(0, \frac{\sigma^2}{(M -}\right.$

Aggregate via Secure Aggregation



Global Noise $\mathcal{N}(0,\sigma^2c^2)$

Release Aggregate

Local Noise: $\mathcal{N}\left(0,\frac{\sigma^2}{(M-1)}c^2\right)$

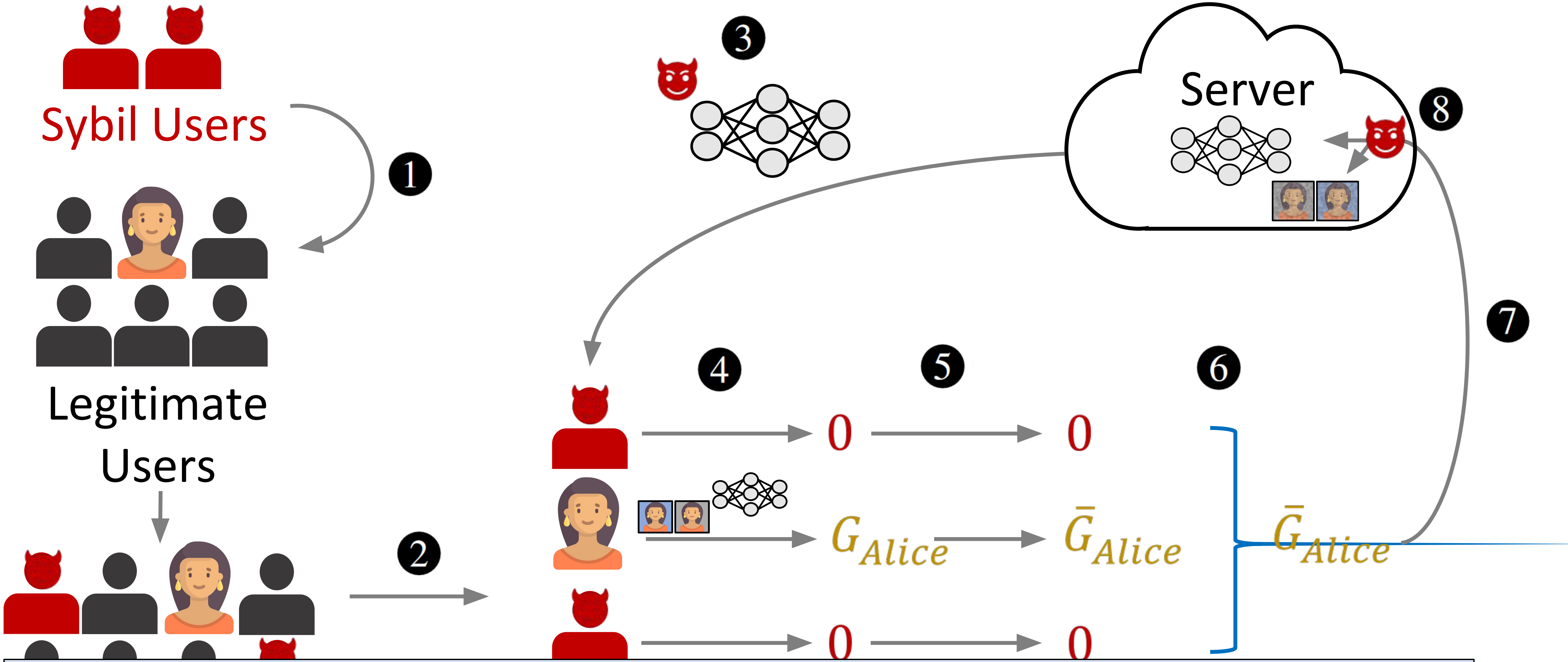
Alice’s data seems protected

Overhead:

- Computation
- Communication
- Storage
- Availability of PKI

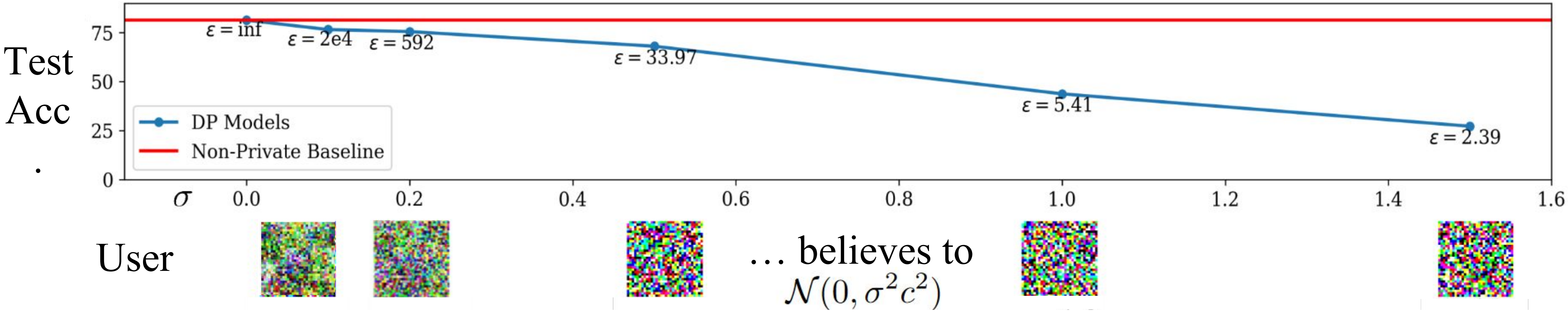
[Bonawitz et al., CCS 2017]

Attacking FL protected by DDP+SA



Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, Nicolas Papernot. *Is Federated Learning a Practical PET Yet?*, 2023. [IEEE Euro S&P '23a]

DDP Reduces to LDP with Low Privacy Levels

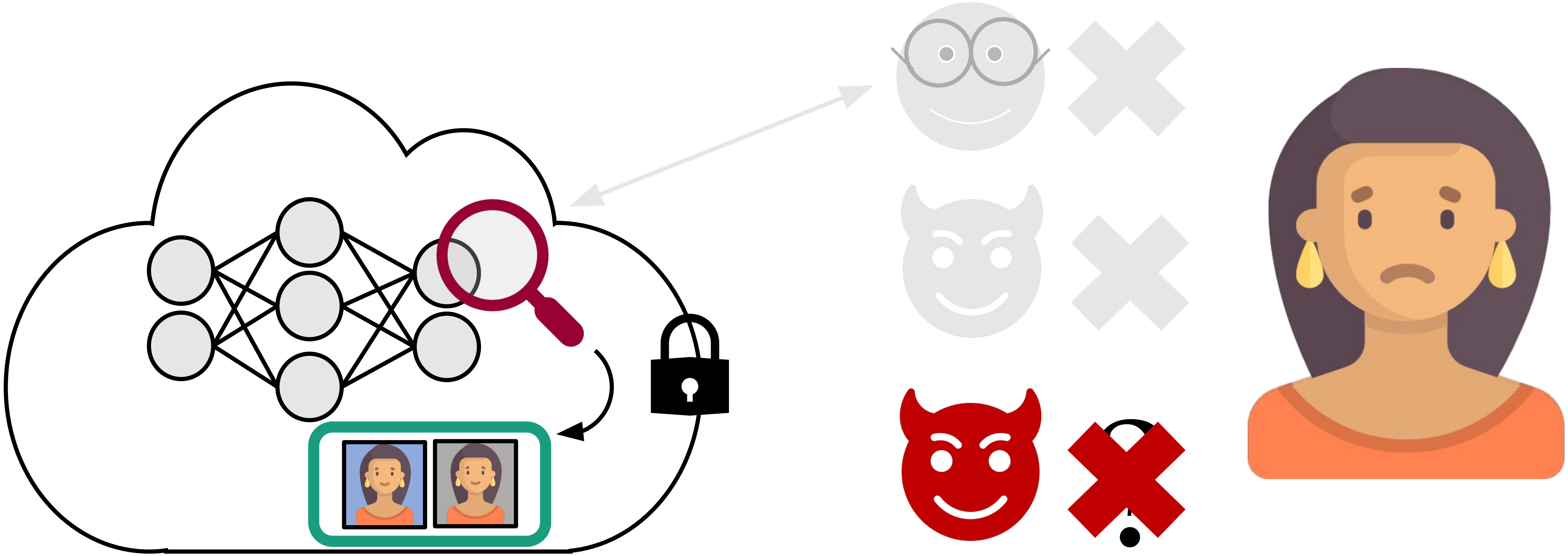


… believes to $\mathcal{N}(0, \sigma^2 c^2)$

Not private enough

Too little utility

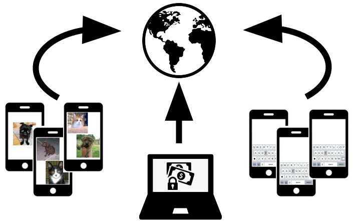
What Trust Model is Needed for Privacy?



Even in hardened variants of the protocol, a malicious attacker can breach individual users’ privacy.

My Research

Goal: Develop mechanisms that provide individualized notions of privacy for machine learning



Federated Learning

Data Extraction in Federated Learning EuroS&P'23
a

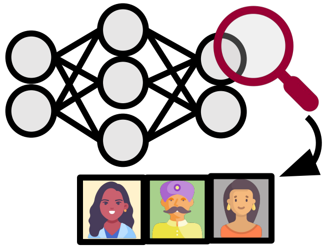
Reconstruction in Hardened Protocols EuroS&P'23
b

Individualized Privacy

Individualized Privacy with PATE PoPETs'23a

Individually Private SGD Training Submission'2
4

Private Prompt Tuning for LLMs Submission'2
4



Privacy Auditing

Side-Channels in Private Query Systems CCS'21

Model Inversion in Speaker Recognition SPSC'22

PoPETs'23
b

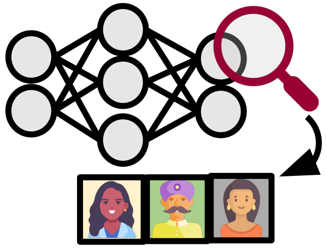
My Research

Goal: Develop mechanisms that provide individualized notions of privacy for machine learning

Individualized Privacy

Individualized Differential Privacy PoPETs’23 a

GDPR-Aligned Privacy Assessment PoPETs’23 b

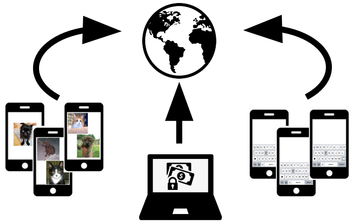


Privacy Auditing

Side-Channels in Private Query Systems CCS’21

Model Inversion in Speaker Recognition SPSC’22

Bounding Membership Inference arXiv’2



Federated Learning

Data Extraction in Federated Learning EuroS&P’23 a

Reconstruction in Hardened Protocols EuroS&P’23

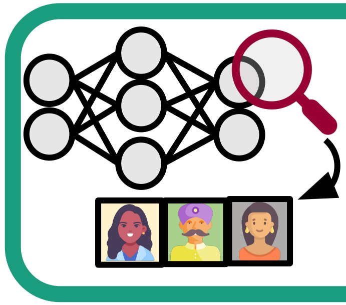
My Research

Goal: Develop mechanisms that provide individualized notions of privacy for machine learning

Individualized Privacy

Individualized Differential Privacy — PoPETs'23 a

GDPR-Aligned Privacy Assessment — PoPETs'23 b

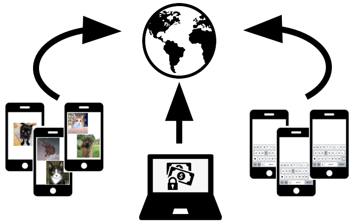


Privacy Auditing

Side-Channels in Private Query Systems — CCS'21

Model Inversion in Speaker Recognition — SPSC'2 2

Bounding Membership Inference — arXiv'2

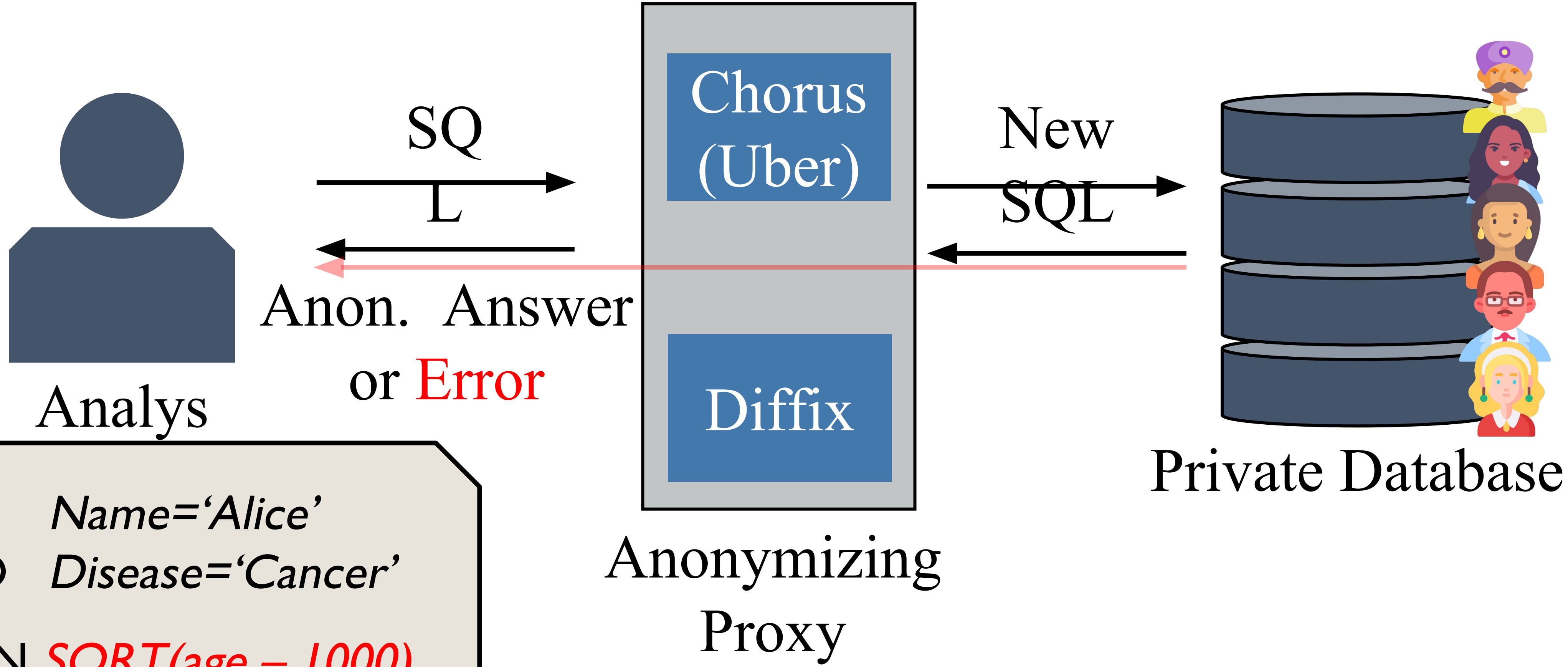


Federated Learning

Data Extraction in Federated Learning — EuroS&P'23 a

Reconstruction in Hardened Protocols — EuroS&P'23

Side-Channel Attacks against Query Systems



IF *Name='Alice'*
AND *Disease='Cancer'*

THEN *SQRT(age – 1000)*

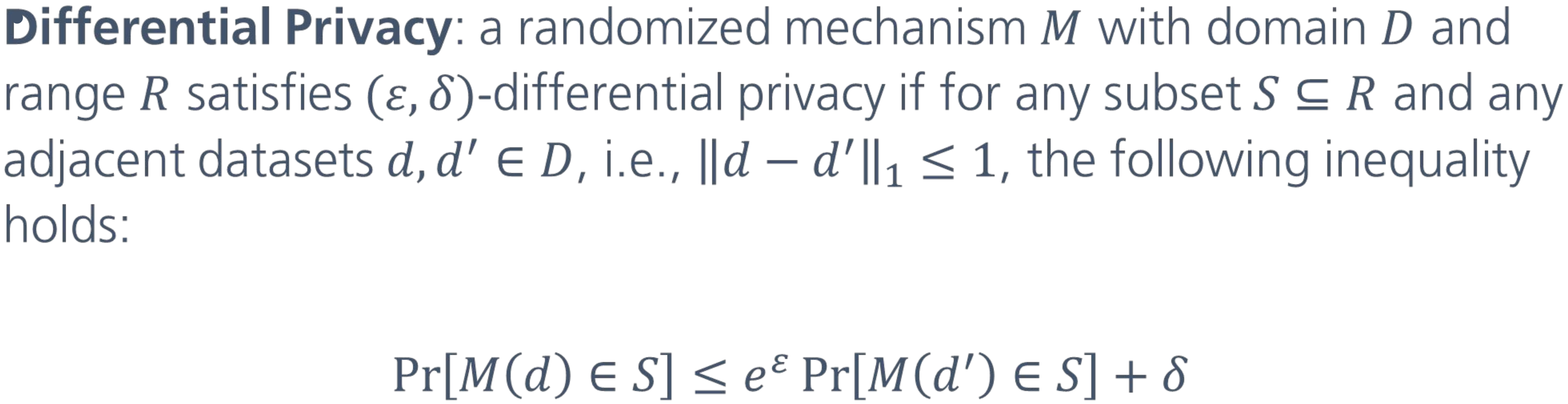
Franziska Boenisch, Reinhard Munz, Marcel Tiepelt, Simon Hanisch, Christiane Kuhn, and Paul Francis. *Side-channel attacks on query-based data anonymization.* In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021. [CCS'21]

FL Sources

Mitigation Methods

Differential Privacy

Goal: produce statistically indistinguishable outputs on any pair of datasets that only differ by any single data point.



Differential Privacy: a randomized mechanism M with domain D and range R satisfies (ε, δ)-differential privacy if for any subset $S \subseteq R$ and any adjacent datasets $d, d' \in D$, i.e., $||d - d'||_1 \leq 1$, the following inequality holds:

$$\Pr[M(d) \in S] \leq e^{\varepsilon} \Pr[M(d') \in S] + \delta$$

Secure Multi Party Computation (MPC)

Setup: given participants p_1, p_2, p_3 and their private data x_1, x_2, x_3.

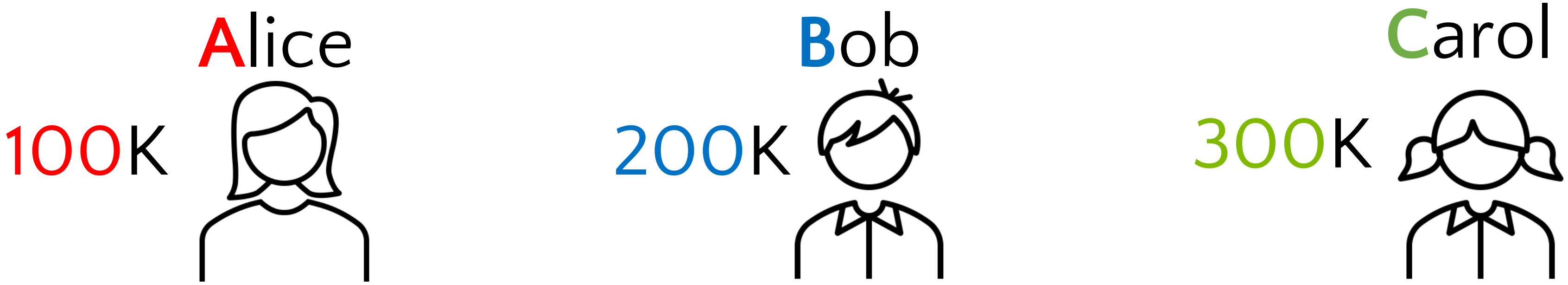
Task: compute value of a private function $F(x_1, x_2, x_3)$.

Example: compute the maximum or average salary of the participants, without revealing the individual salaries.

Machine Learning: shareholders can compute **any function** of inputs without seeing anything but shares and the final output.**Properties:** (1) input privacy – no information about private data can be inferred from messages exchanged during MPC, and

(2) honest parties either compute correct output or abort.

Secure Multi-Party Computation (MPC)



Alice

100K

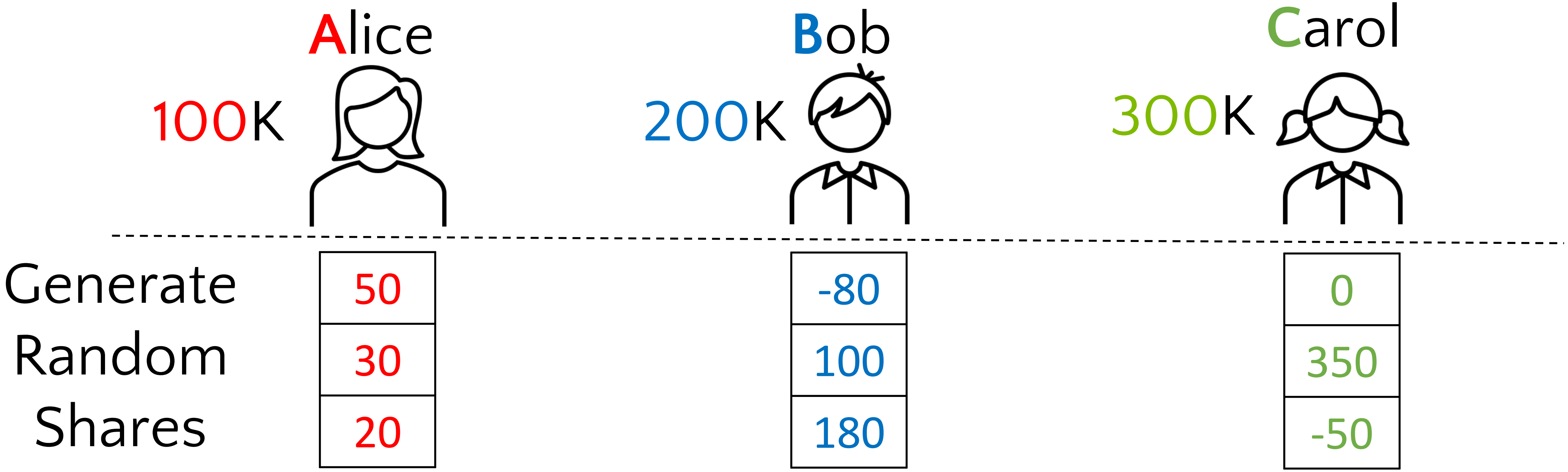
Bob

200K

Carol

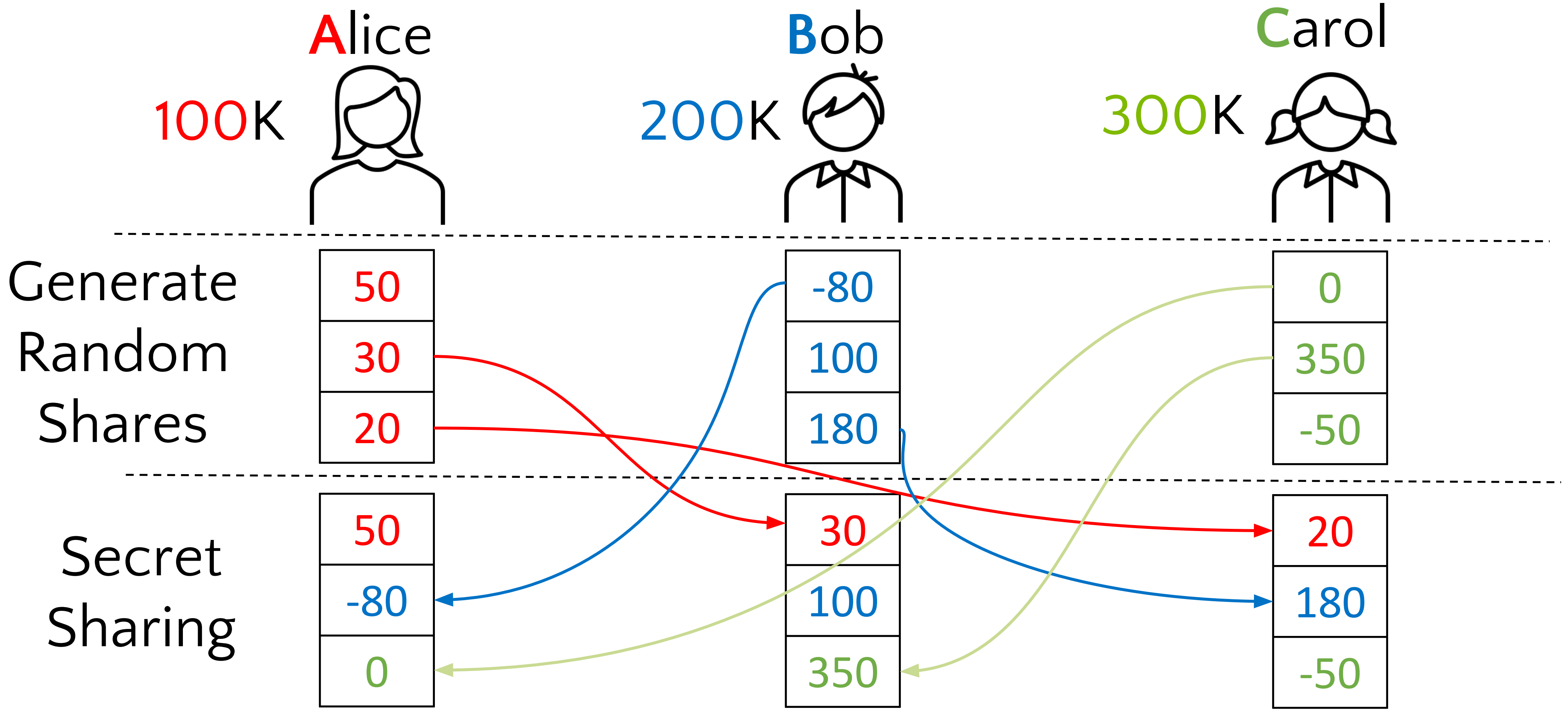
300K

Secure Multi-Party Computation (MPC)



	Alice	Bob	Carol
	100K	200K	300K
Generate Random Shares	50	-80	0
	30	100	350
	20	180	-50

Secure Multi-Party Computation (MPC)



	Alice	Bob	Carol
	100K	200K	300K
Generate Random Shares	50 30 20	-80 100 180	0 350 -50
Secret Sharing	50 -80 0	30 100 350	20 180 -50

Secure Multi-Party Computation (MPC)

	Alice	Bob	Carol
	100K	200K	300K
Generate Random Shares	50 30 20	-80 100 180	0 350 -50
Secret Sharing	50 -80 0	30 100 350	20 180 -50
Add Shares	-30	480	150

Sum: 600 Mean: 200

Homomorphic Encryption

1. Addition

$$Enc(x) + Enc(y) \ = \ Enc(x+y)$$

$$Enc(x) + y^{*} = \ Enc(x+y)$$

2. Multiplication

$$Enc(x) * Enc(y) =$$

$$x^{e} mod\, n * y^{e} mod\, n =$$

$$x^{e}\, y^{e}\, mod\, n =$$

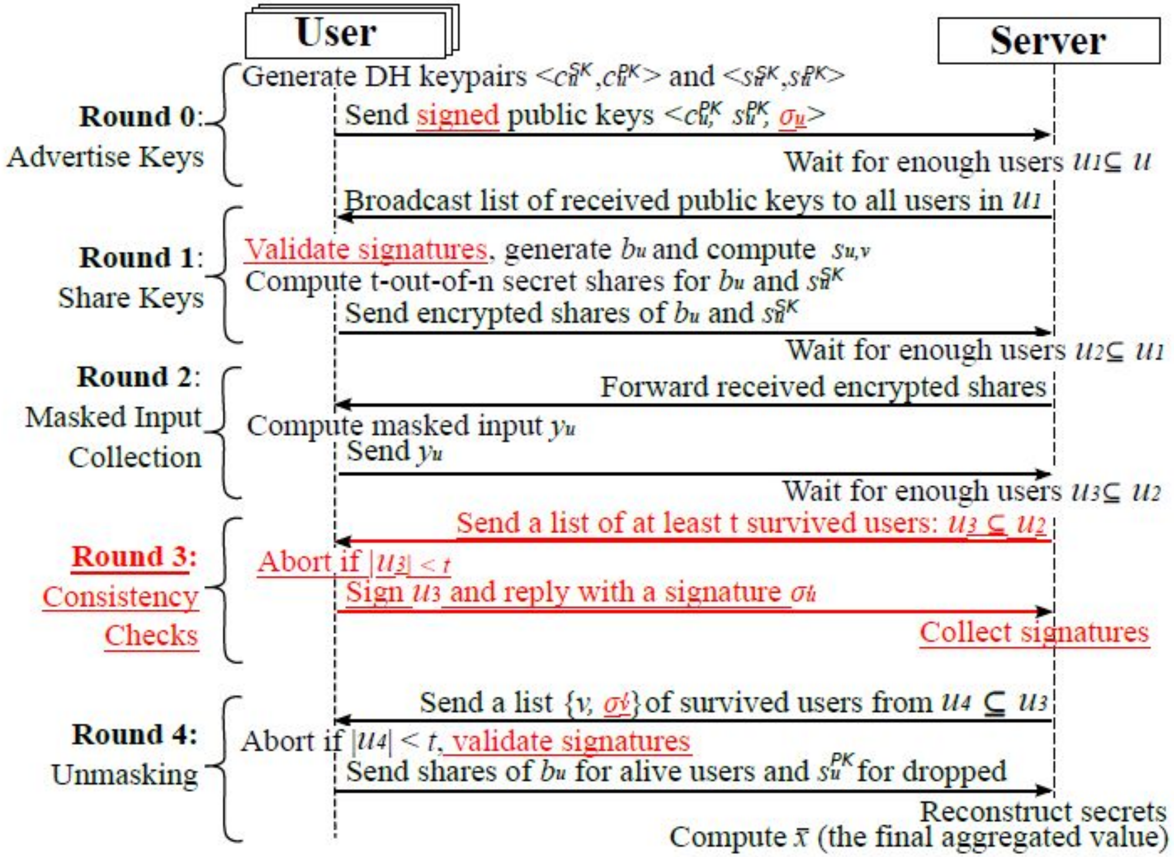
$$(xy)^{e}\, mod\, n =$$

$$Enc(x * y)$$

Attacker Models

- Honest-but-curious – adversary follows the protocol but tries to infer information from the protocol transcript.
- Malicious – adversary actively deviates from the protocol
- Occasionally Byzantine – adversary acts honest most of the time and only acts maliciously on occasions

Secure Aggregation



User

Server

Round 0: Advertise Keys

Generate DH keypairs $<c_u^{SK}, c_u^{PK}>$ and $<s_u^{SK}, s_u^{PK}>$

Send signed public keys $<c_u^{PK}, s_u^{PK}, \sigma_u>$

Wait for enough users $\mathcal{U}_1 \subseteq \mathcal{U}$

Broadcast list of received public keys to all users in $\mathcal{U}_1$

Round 1: Share Keys

Validate signatures, generate b_u and compute $s_{u,v}$

Compute t-out-of-n secret shares for b_u and s_u^{SK}

Send encrypted shares of b_u and s_u^{SK}

Wait for enough users $\mathcal{U}_2 \subseteq \mathcal{U}_1$

Round 2: Masked Input Collection

Forward received encrypted shares

Compute masked input y_u

Send y_u

Wait for enough users $\mathcal{U}_3 \subseteq \mathcal{U}_2$

Round 3: Consistency Checks

Send a list of at least t survived users: $\mathcal{U}_3 \subseteq \mathcal{U}_2$

Abort if $|\mathcal{U}_3| < t$

Sign $\mathcal{U}_3$ and reply with a signature σ_u'

Collect signatures

Round 4: Unmasking

Send a list $\{v, \sigma_v'\}$ of survived users from $\mathcal{U}_4 \subseteq \mathcal{U}_3$

Abort if $|\mathcal{U}_4| < t$, validate signatures

Send shares of b_u for alive users and s_u^{PK} for dropped

Reconstruct secrets

Compute $\bar{x}$ (the final aggregated value)

- Robustness **(Malicious Server)**
  - Can collaborate with up to n/3-1 clients
  - Tolerates up to n/3-1 dropouts of clients

Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. "Practical secure aggregation for privacy-preserving machine learning." In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175-1191. 2017

Secure Aggregation

Bonawitz et al., 2017

- Computation:
  - User : $O(n^2 + mn)$
  - Server : $O(mn^2)$
- Communication:
  - User : $O(n + m)$
  - Server : $O(n^2 + mn)$
- Storage:
  - User : $O(n + m)$
  - Server : $O(n^2 + m)$

Bell et al., 2020

- Computation:
  - User : $O(\log^2 n + \log n)$
  - Server : $O(n(\log^2 n + \log n))$
- Communication:
  - User : $O(\log n + m)$
  - Server : $O(n(\log n + m))$

Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. "Practical secure aggregation for privacy-preserving machine learning." In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175-1191. 2017.

Bell, James Henry, Kallista A. Bonawitz, Adrià Gascón, Tancrède Lepoint, and Mariana Raykova. "Secure single-server aggregation with (poly) logarithmic overhead." In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1253-1269. 2020.

Distributed Differential Privacy

Distributed Discrete Gaussian

- discretizes the data and adds discrete Gaussian noise before performing secure aggregation

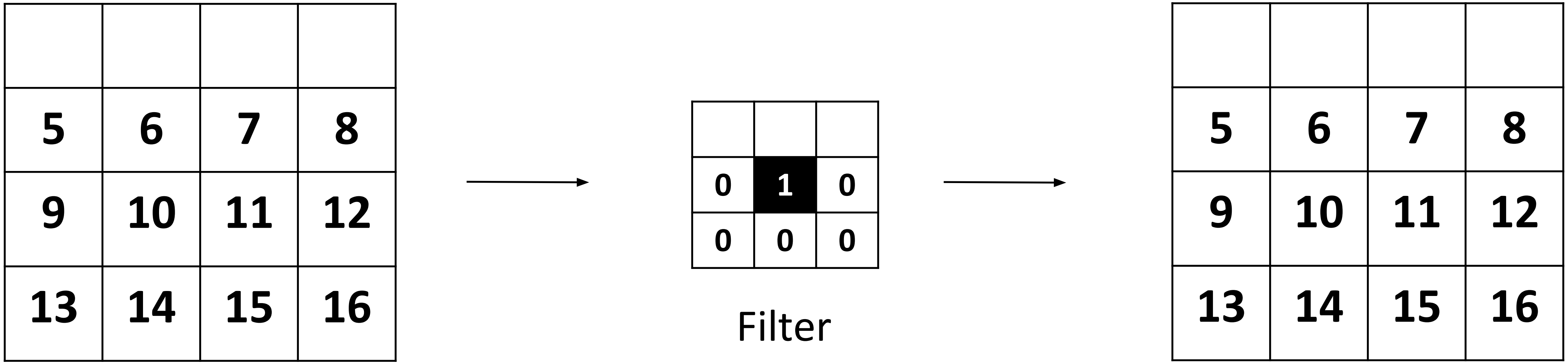
Kairouz, Peter, Ziyu Liu, and Thomas Steinke. "The distributed discrete gaussian mechanism for federated learning with secure aggregation." In *International Conference on Machine Learning*, pp. 5201-5212. PMLR, 2021.

Skellam Mechanism

- based on the difference of two independent Poisson random variables

Agarwal, Naman, Peter Kairouz, and Ziyu Liu. "The skellam mechanism for differentially private federated learning." *Advances in Neural Information Processing Systems* 34 (2021): 5052-5064.

Forwarding over Convolutional Layers



5	6	7	8
9	10	11	12
13	14	15	16

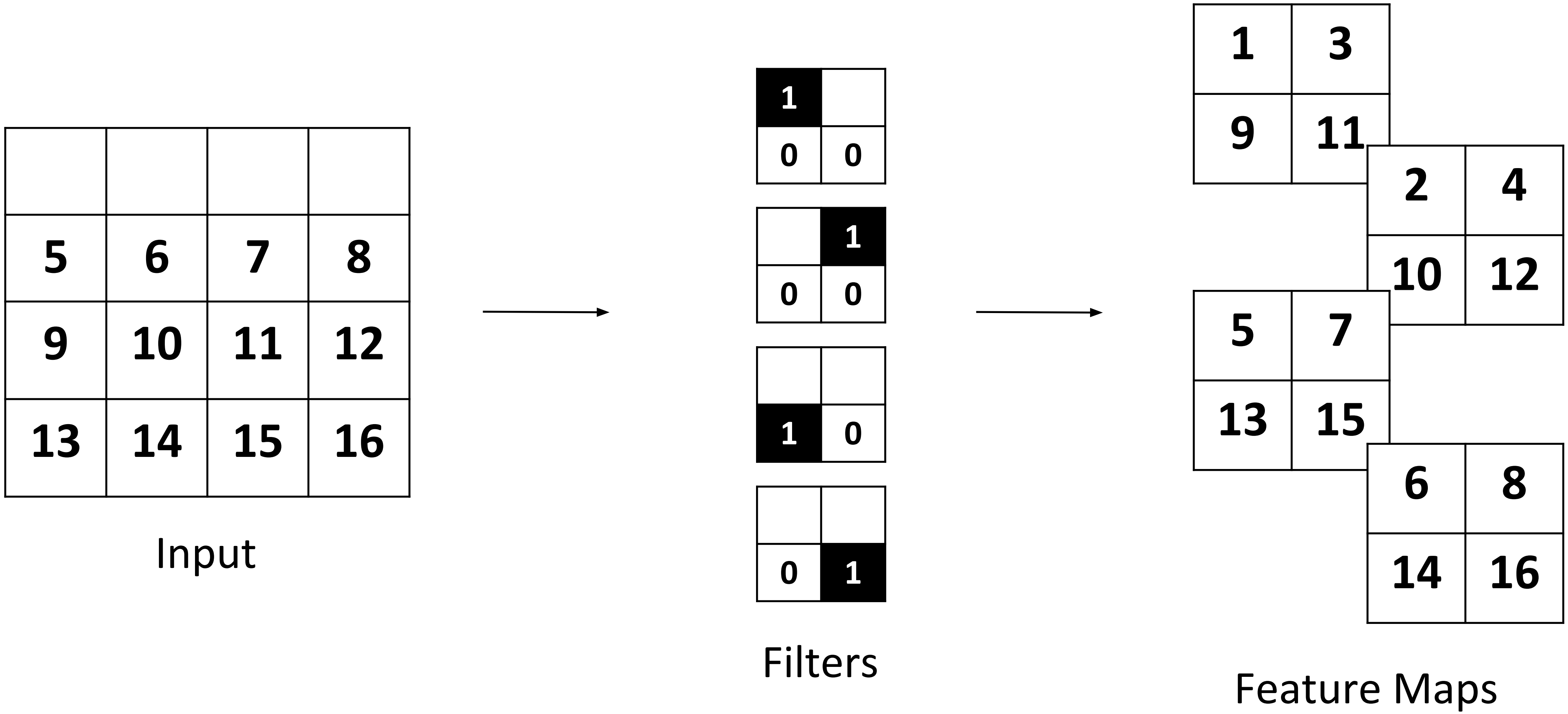
Input

0	1	0
0	0	0

Filter

5	6	7	8
9	10	11	12
13	14	15	16

Feature Maps



5	6	7	8
9	10	11	12
13	14	15	16

Input

1	
0	0

	1
0	0

1	0

0	1

Filters

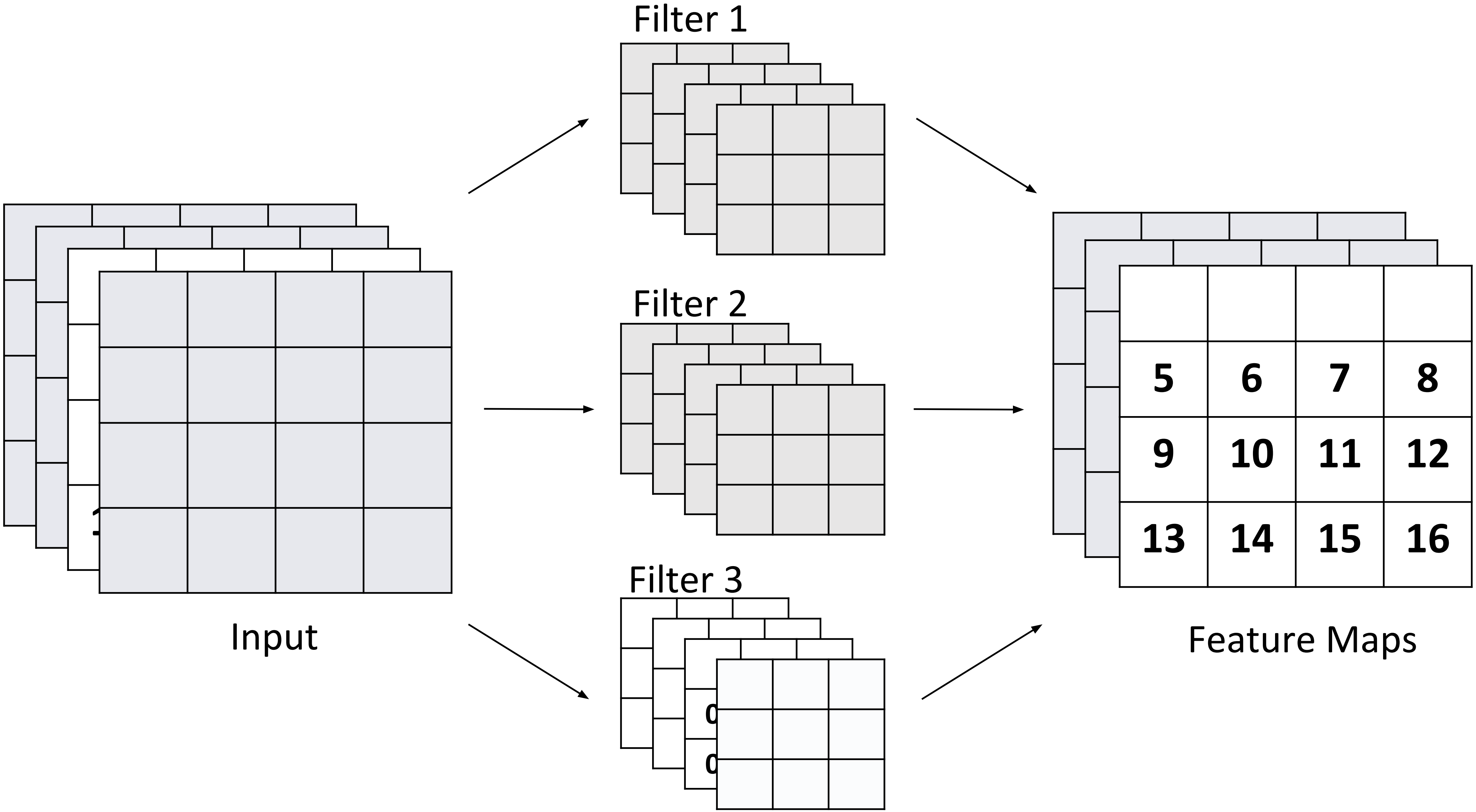
1	3
9	11

2	4
10	12

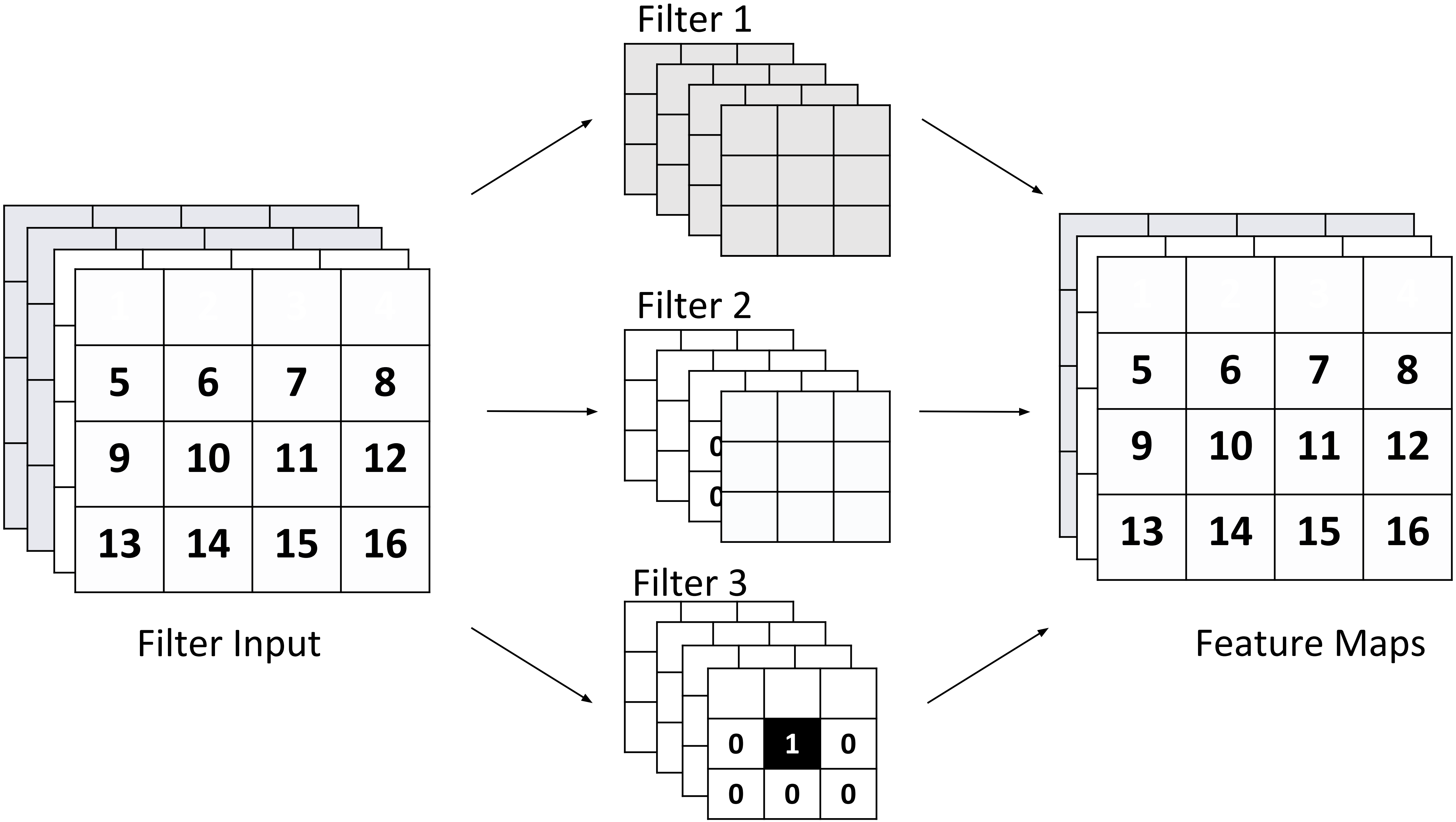
5	7
13	15

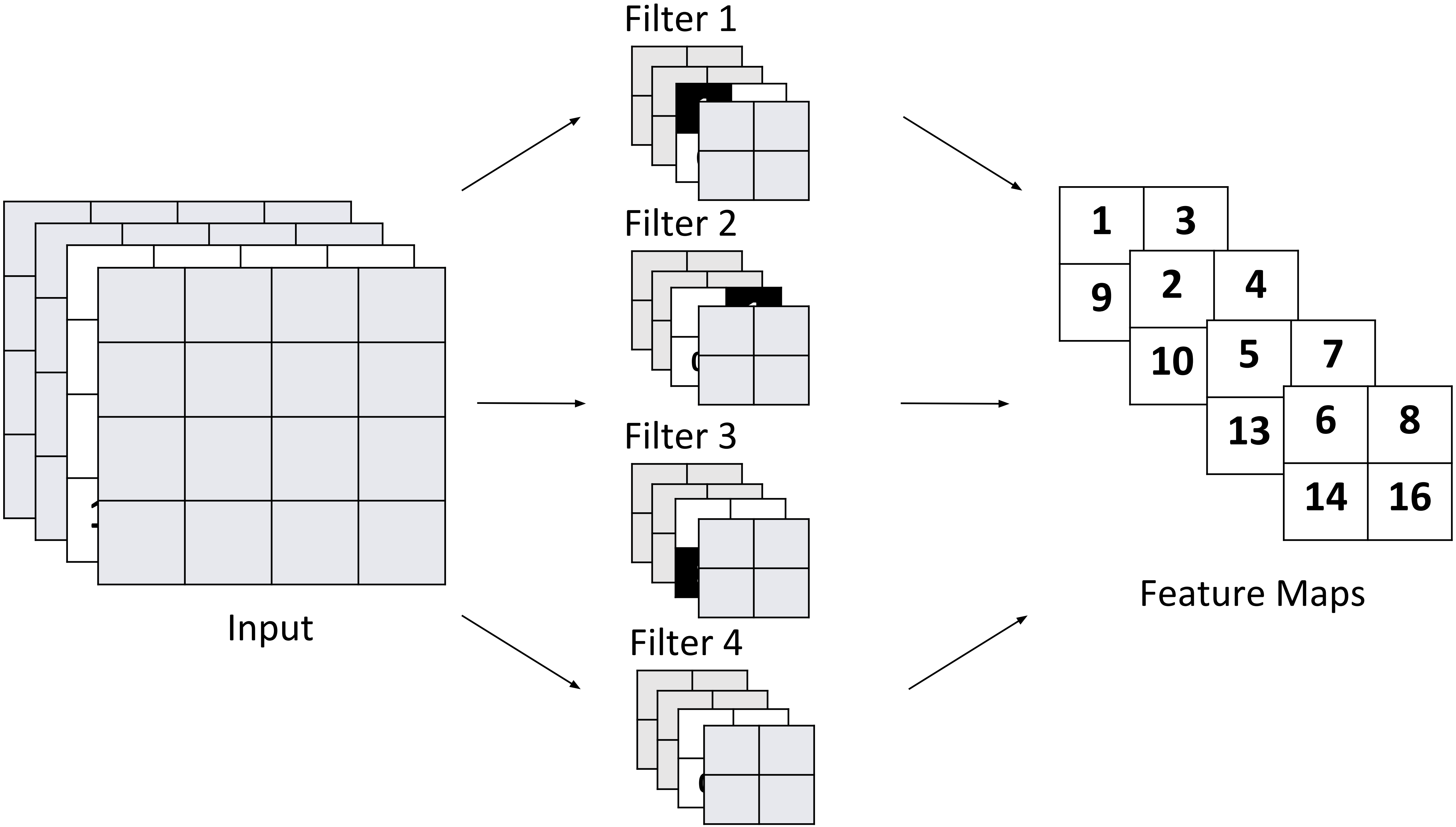
6	8
14	16

Feature Maps

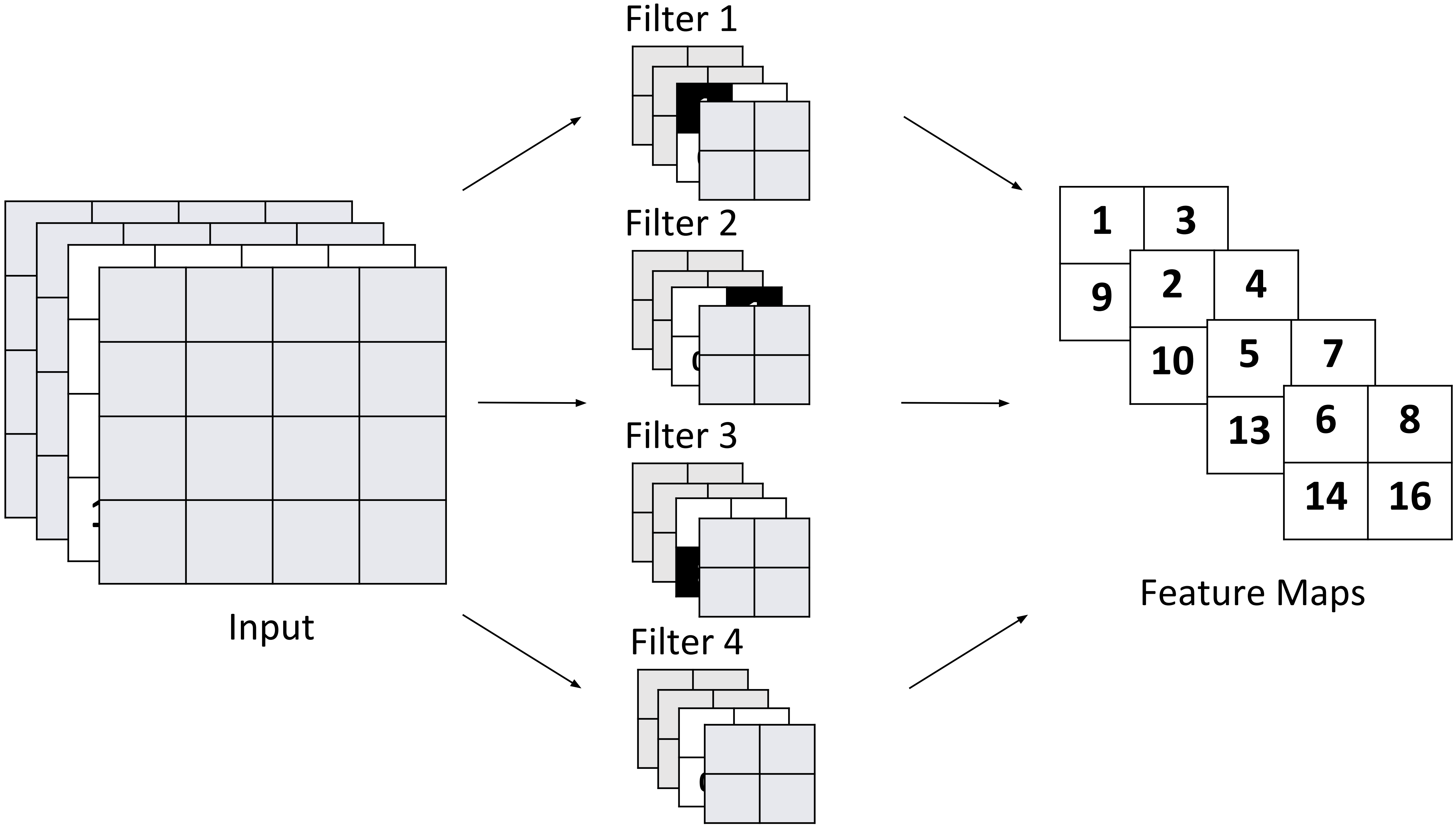


5	6	7	8
9	10	11	12
13	14	15	16

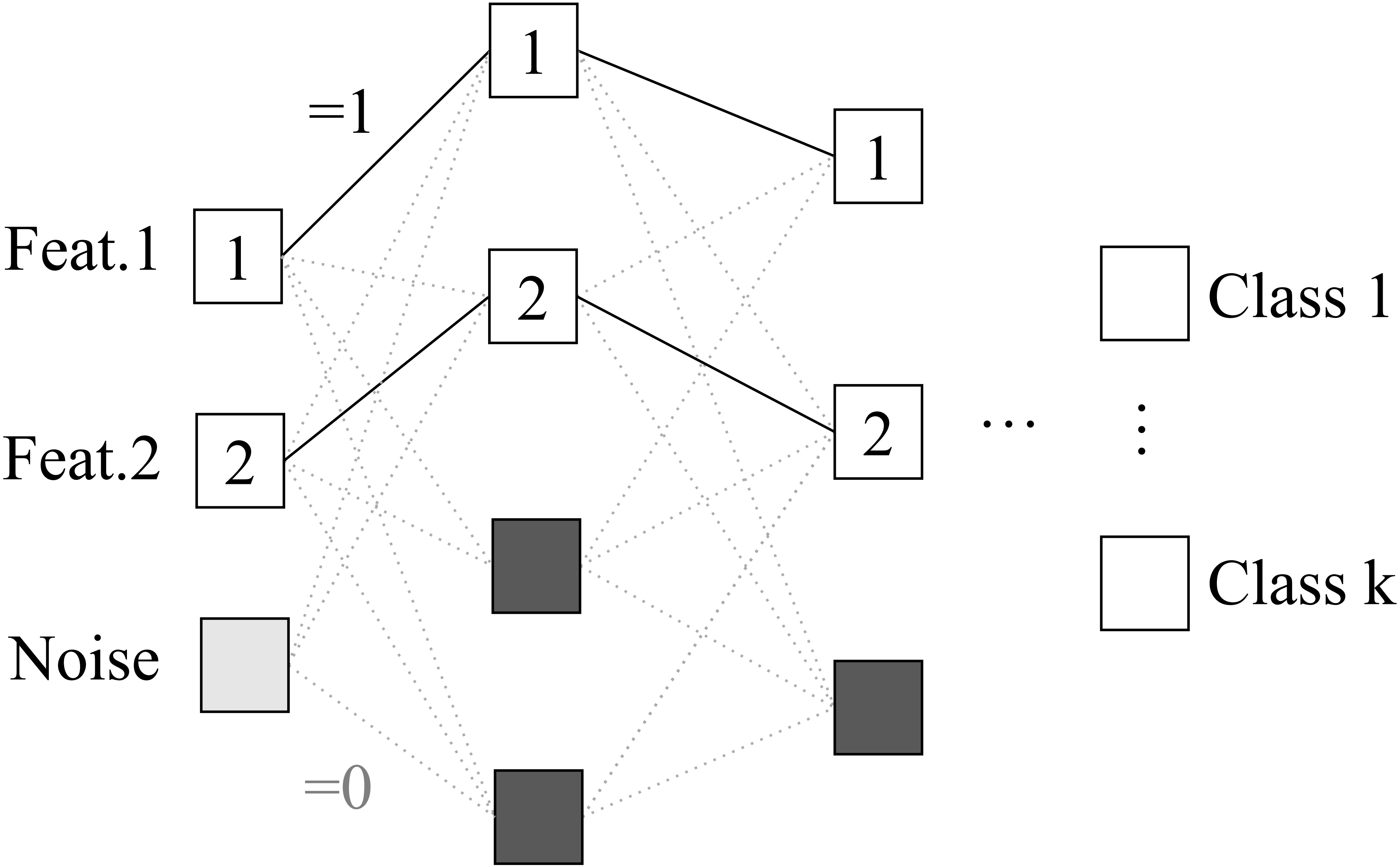

Filter 1
Filter 2
Filter 3
1 2 3 4
5 6 7 8
9 10 11 12
13 14 15 16
0
0
0 1 0
0 0 0
Filter Input
Feature Maps
1 2 3 4
5 6 7 8
9 10 11 12
13 14 15 16

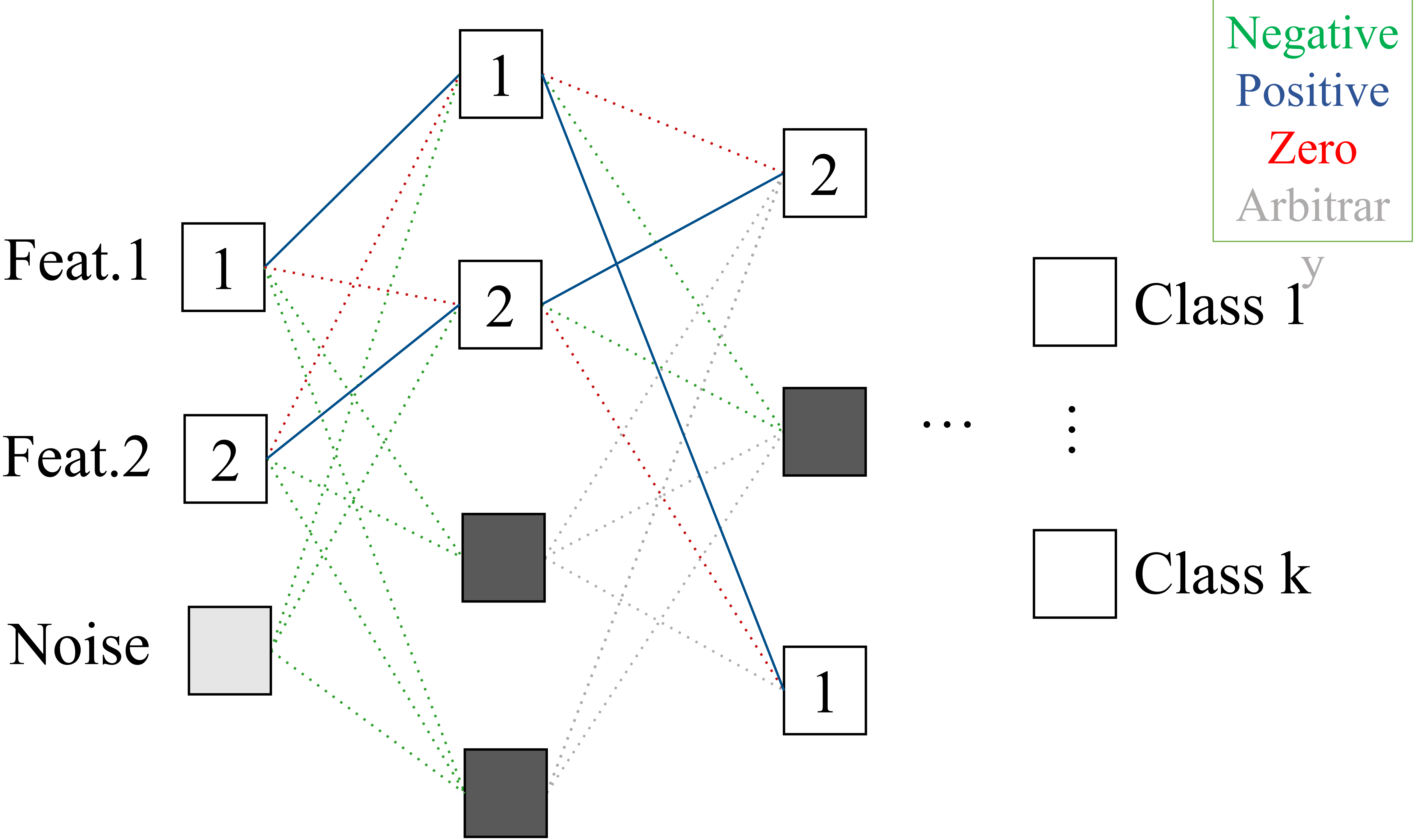


Filter 1
Filter 2
Filter 3
Filter 4
Input
1
3
9
2
4
10
5
7
13
6
8
14
16
Feature Maps


Filter 1
Filter 2
Filter 3
Filter 4
Input
1
3
9
2
4
10
5
7
13
6
8
14
16
Feature Maps

Forwarding over Fully Connected Layers


=1
Feat.1
1
Feat.2
2
Noise
=0
1
2
1
2
…
Class 1
⋮
Class k

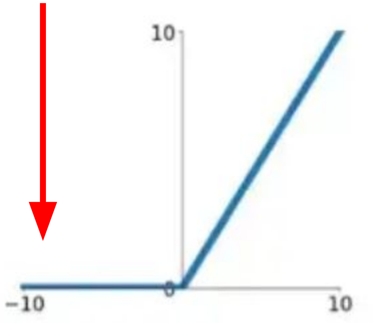

Negative
Positive
Zero
Arbitrary
Feat.1
Feat.2
Noise
1
2
1
2
2
1
…
Class 1
⋮
Class k

Other Activation Functions

ReLU

$\max(0, x)$

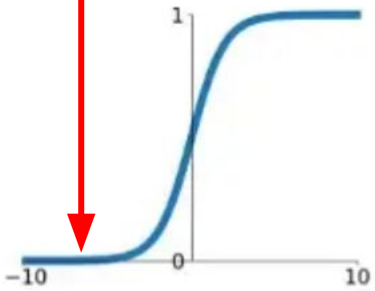
Zero-Gradients



Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$

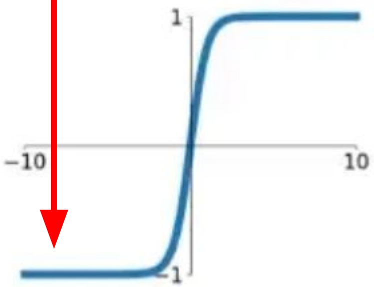
Zero-Gradients



tanh

$\tanh(x)$

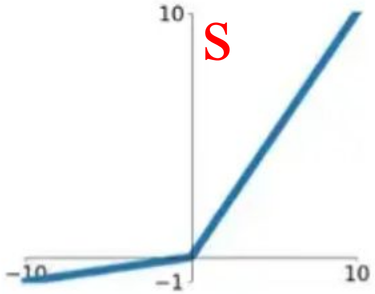
Zero-Gradients



Leaky ReLU

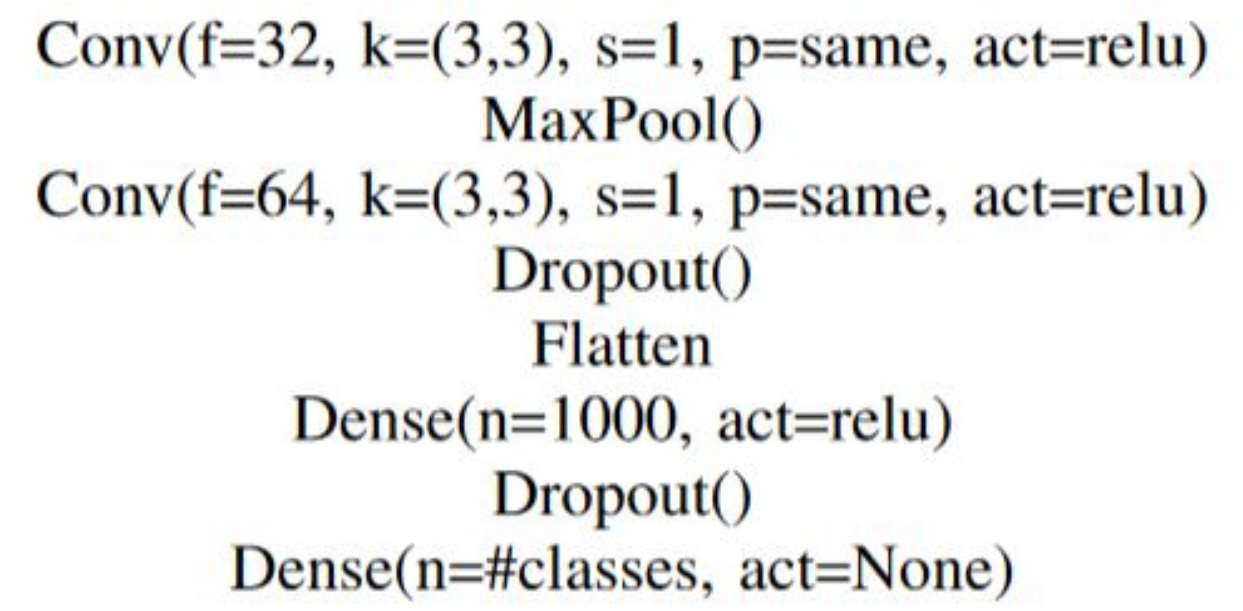
$\max(0.1x, x)$

Non-Zero-Gradients



… but less sparsity.

Lossy Architecture



Conv(f=32, k=(3,3), s=1, p=same, act=relu)

MaxPool()

Conv(f=64, k=(3,3), s=1, p=same, act=relu)

Dropout()

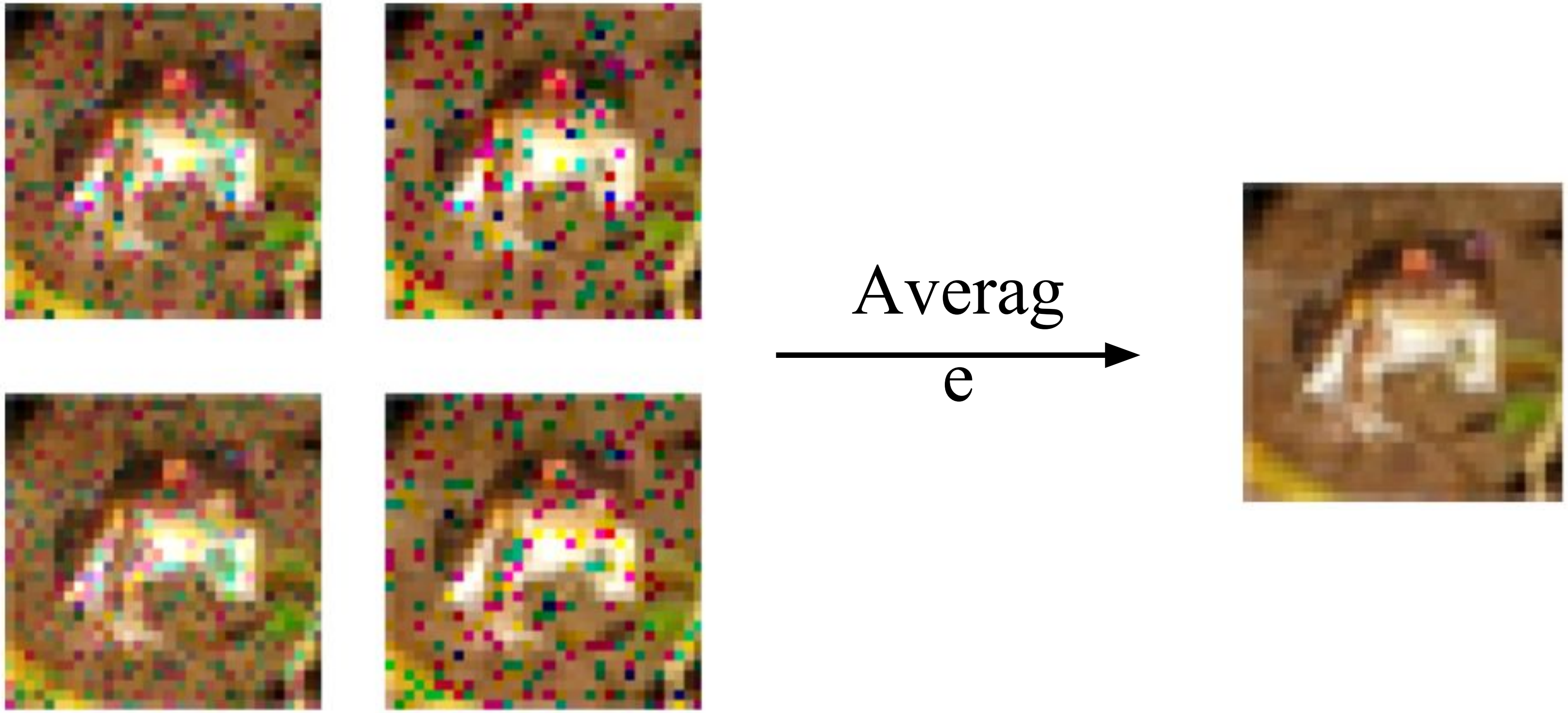
Flatten

Dense(n=1000, act=relu)

Dropout()

Dense(n=#classes, act=None)

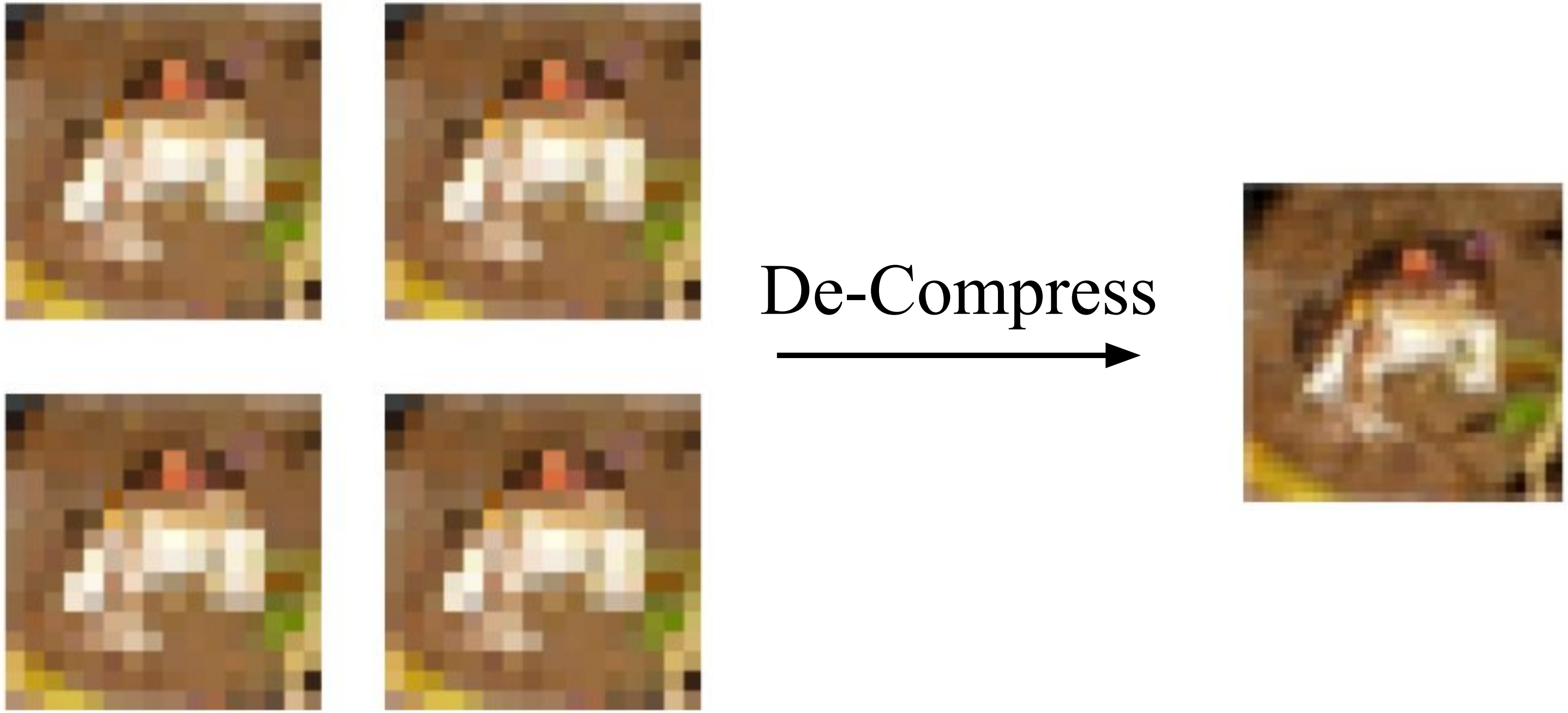
Effect of Dropout



Averag
e

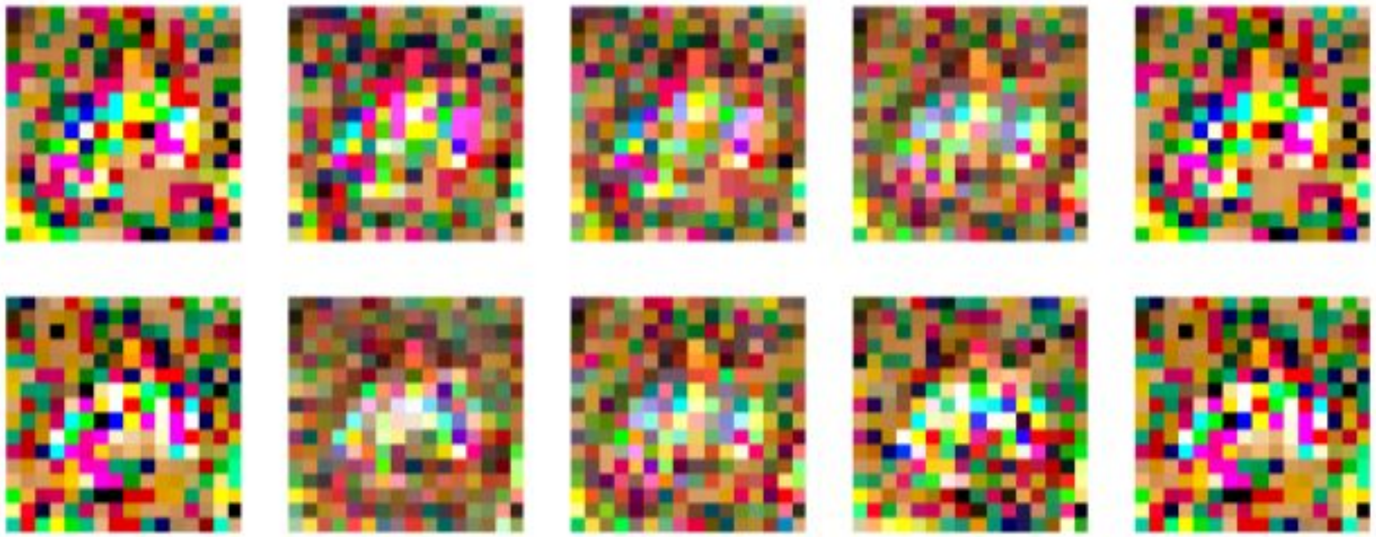
Same Data Point Extracted at 4 Different Gradients (Dropout Rate = 0.1)

Effect of Pooling



Same Data Point Extracted at 4 Different Gradients (Max Pooling with 2x2)

Heavy Dropout and Pooling

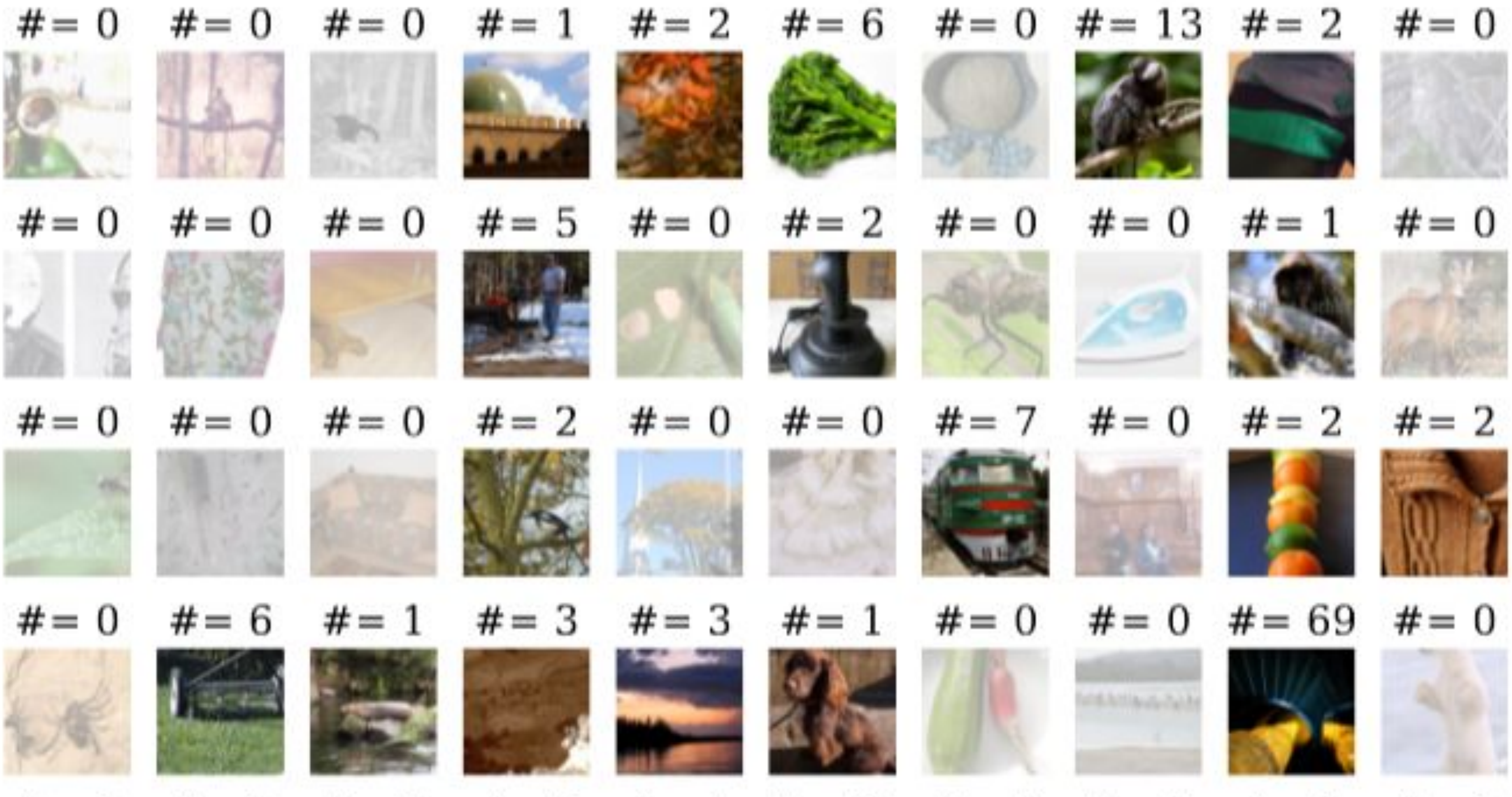


(c) Dropout with $p = 0.3$ and pooling.

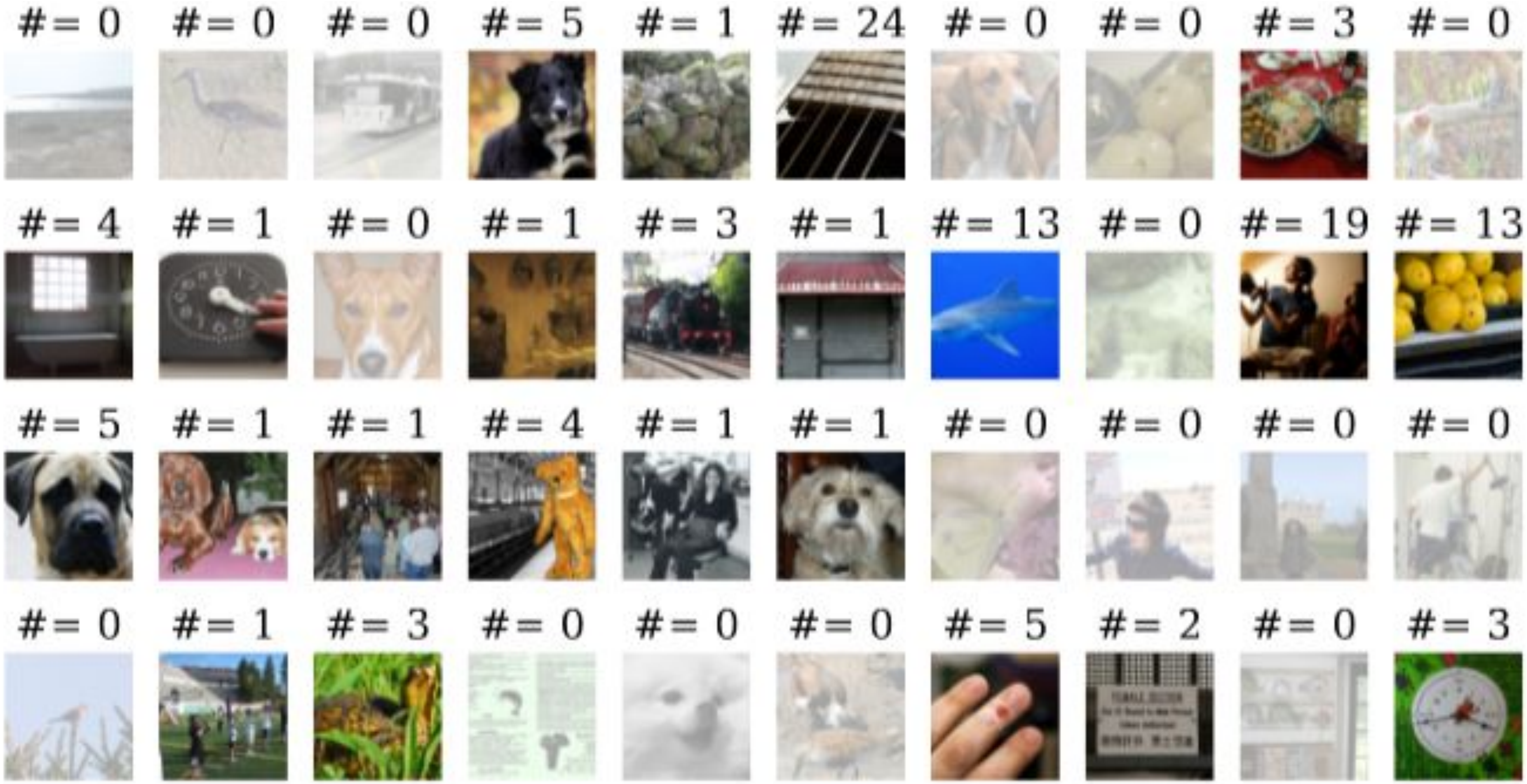
Individual Activation Neurons

	% Individual Act.	
	Passive	**Active**
MNIST	0.6%	**20.3%**
CIFAR10	5.8%	**41.2%**
ImageNet	4.4%	**51.4%**
IMDB	3.6%	**19.2%**

Extractable Datapoints



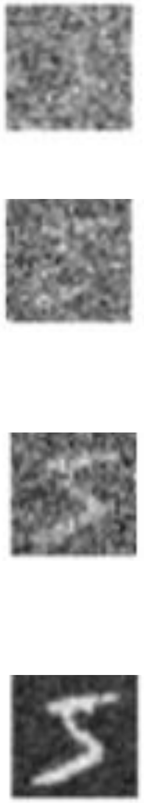
Extractable Datapoints

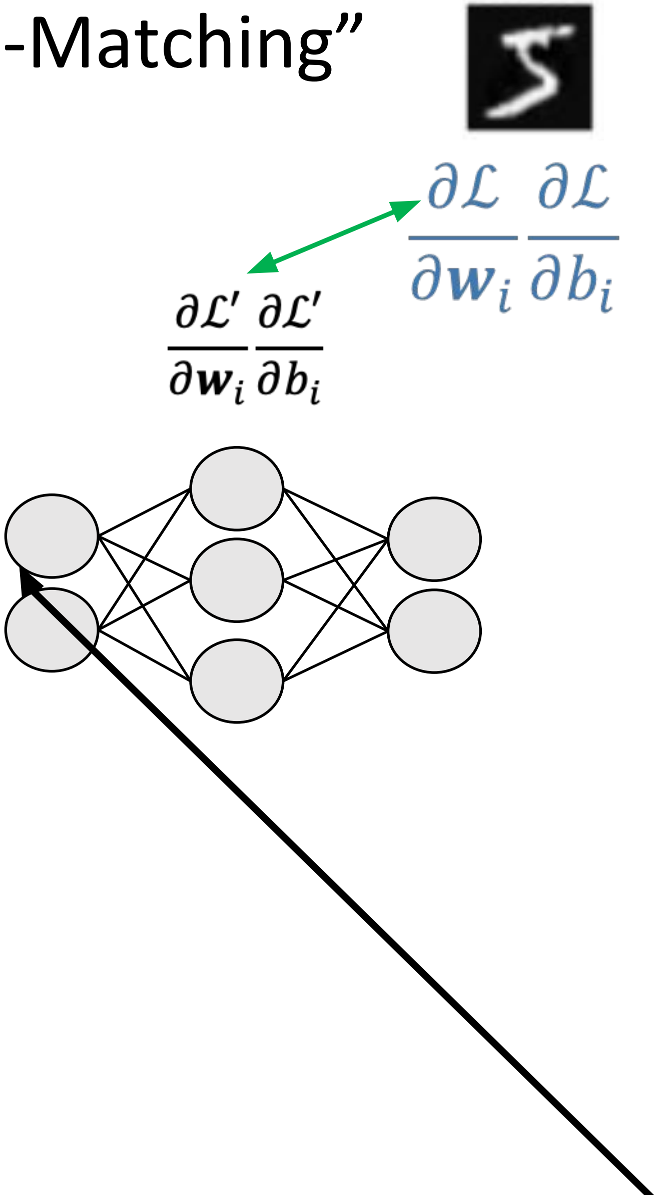


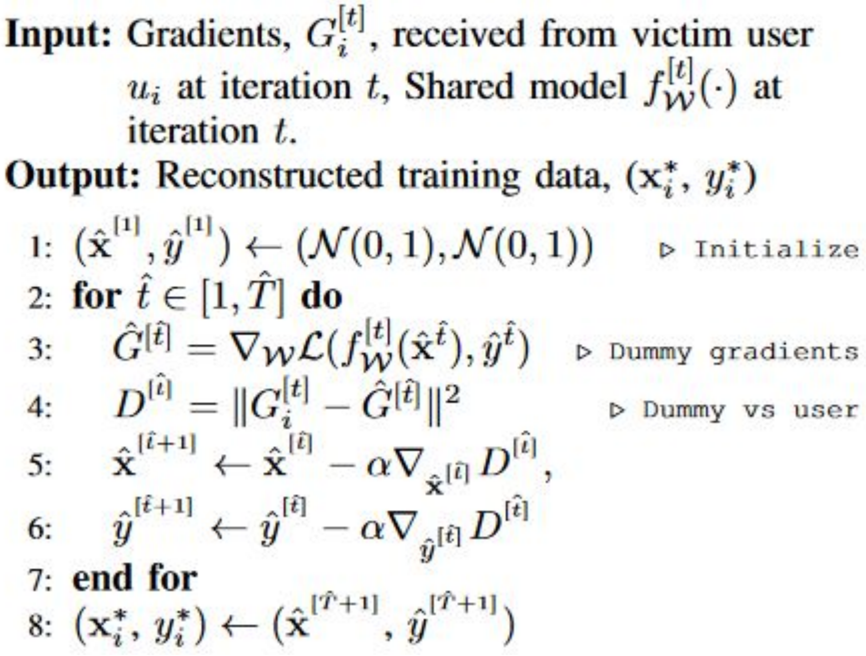
Related Work (FL)

Optimization-based Data Reconstruction

“Gradient-Matching”




$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_i}$ $\frac{\partial \mathcal{L}}{\partial b_i}$
$\frac{\partial \mathcal{L}'}{\partial \boldsymbol{w}_i}$ $\frac{\partial \mathcal{L}'}{\partial b_i}$



Input: Gradients, $G_i^{[t]}$, received from victim user u_i at iteration t, Shared model $f_{\mathcal{W}}^{[t]}(\cdot)$ at iteration t.

Output: Reconstructed training data, $(\mathbf{x}_i^*, y_i^*)$

```
$(\hat{\mathbf{x}}^{[1]}, \hat{y}^{[1]}) \leftarrow (\mathcal{N}(0,1), \mathcal{N}(0,1))$    ▷ Initialize
for $\hat{t} \in [1, \hat{T}]$ do
    $\hat{G}^{[\hat{t}]} = \nabla_{\mathcal{W}} \mathcal{L}(f_{\mathcal{W}}^{[t]}(\hat{\mathbf{x}}^{\hat{t}}), \hat{y}^{\hat{t}})$    ▷ Dummy gradients
    $D^{[\hat{t}]} = \|G_i^{[t]} - \hat{G}^{[\hat{t}]}\|^2$    ▷ Dummy vs user
    $\hat{\mathbf{x}}^{[\hat{t}+1]} \leftarrow \hat{\mathbf{x}}^{[\hat{t}]} - \alpha \nabla_{\hat{\mathbf{x}}^{[\hat{t}]}} D^{[\hat{t}]}$,
    $\hat{y}^{[\hat{t}+1]} \leftarrow \hat{y}^{[\hat{t}]} - \alpha \nabla_{\hat{y}^{[\hat{t}]}} D^{[\hat{t}]}$
end for
$(\mathbf{x}_i^*, y_i^*) \leftarrow (\hat{\mathbf{x}}^{[\hat{T}+1]}, \hat{y}^{[\hat{T}+1]})$
```

[1] Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." *Advances in neural information processing systems* 32 (2019).

Limitations and Summary of Passive Attackers

- Computationally expensive
- Low fidelity
- Non-complex data
- Small mini-batch sizes, different classes

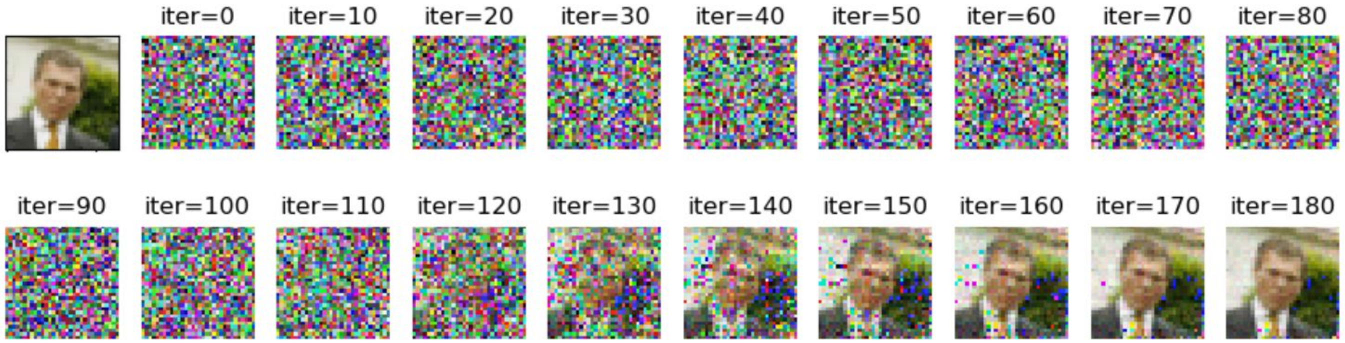


Figure taken from [2].

Even a *passive attacker* in vanilla FL
can reconstruct private user data.

[2] Zhao, Bo, Konda Reddy Mopuri, and Hakan Bilen. "idlg: Improved deep leakage from gradients." *arXiv preprint arXiv:2001.02610* (2020).

Imprinting User Data in Model Gradients

- Observation: bias term controls if a data point activates a neuron

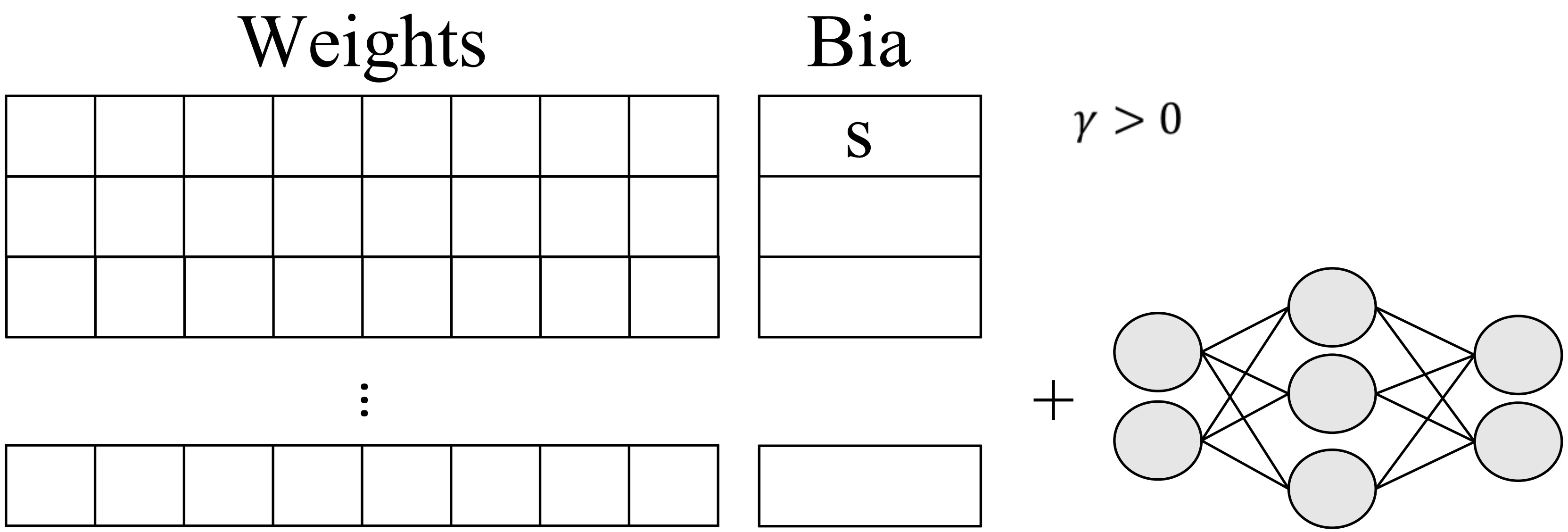
$$y_i = ReLU\left(\boldsymbol{w}_i^T \boldsymbol{x} + b_i\right)$$
$$y_i = 0 \;\text{ if } \mathrm{b_i} < 0 \text{ and } |b_i| > |\boldsymbol{w}_i^T \boldsymbol{x}|$$

- Approach:
  - Control which data points activate what neurons
  - Turn model weights into linear function (e.g. average pixel brightness: $\frac{1}{m}$, m: number of features)
  - Iteratively extract data

[Fowl et al., 2021, ICLR]

Imprinting User Data in Model Gradients

Mini-Batc
h



Weights

Bia

s

$\gamma > 0$

⋮

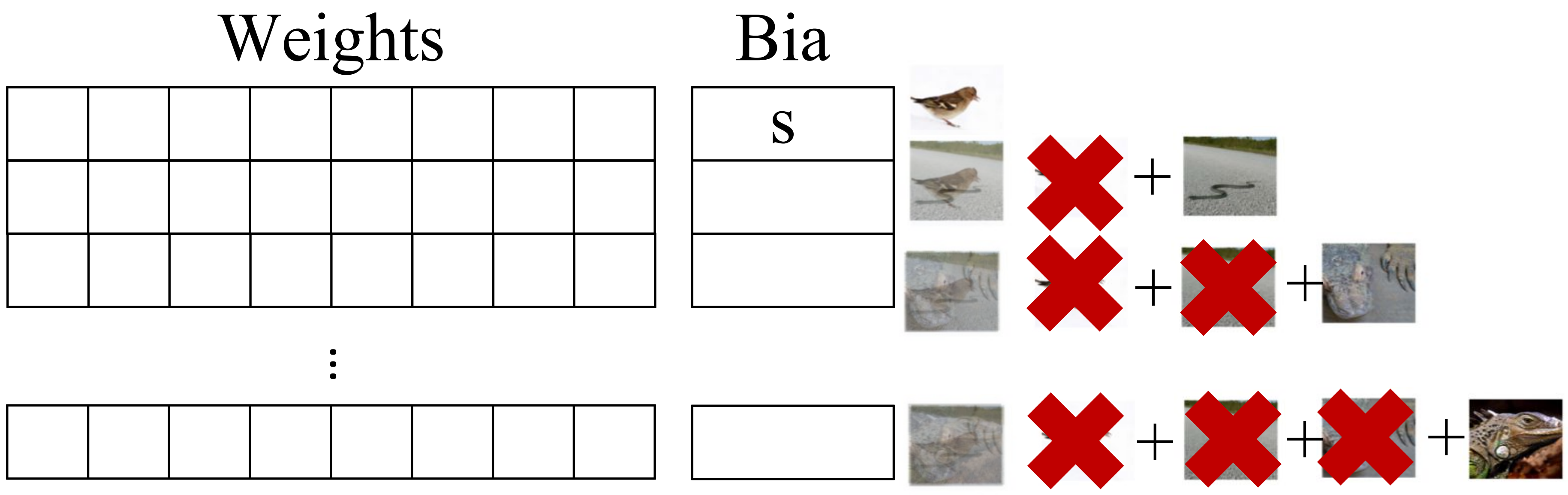
$+$

Imprint Module
(Fully Connected
Layer)

[Fowl et al., 2021, ICLR]

Imprinting User Data in Model Gradients

Mini-Batc
h



Imprint Module
(Fully Connected
Layer)

[Fowl et al., 2021, ICLR]

Data Extraction Success

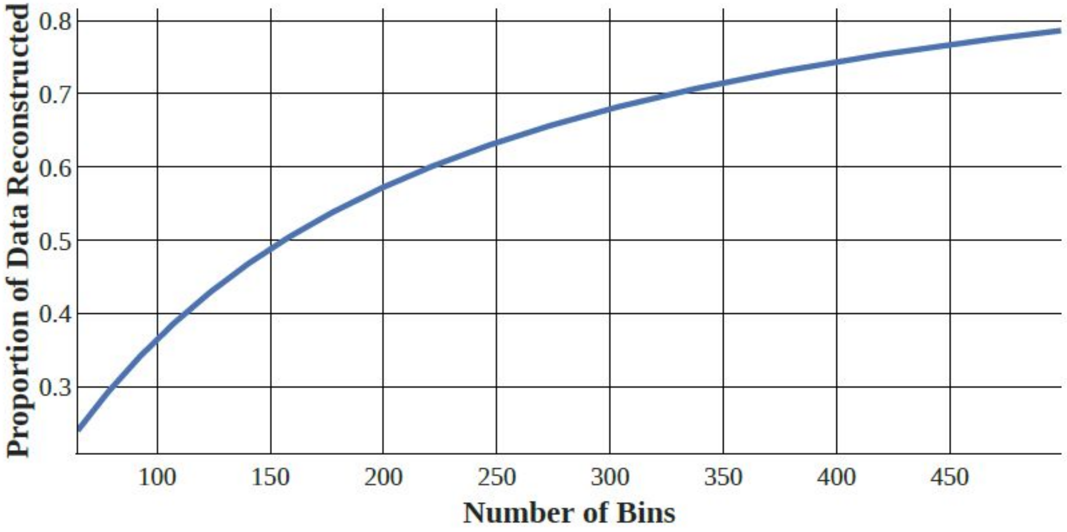
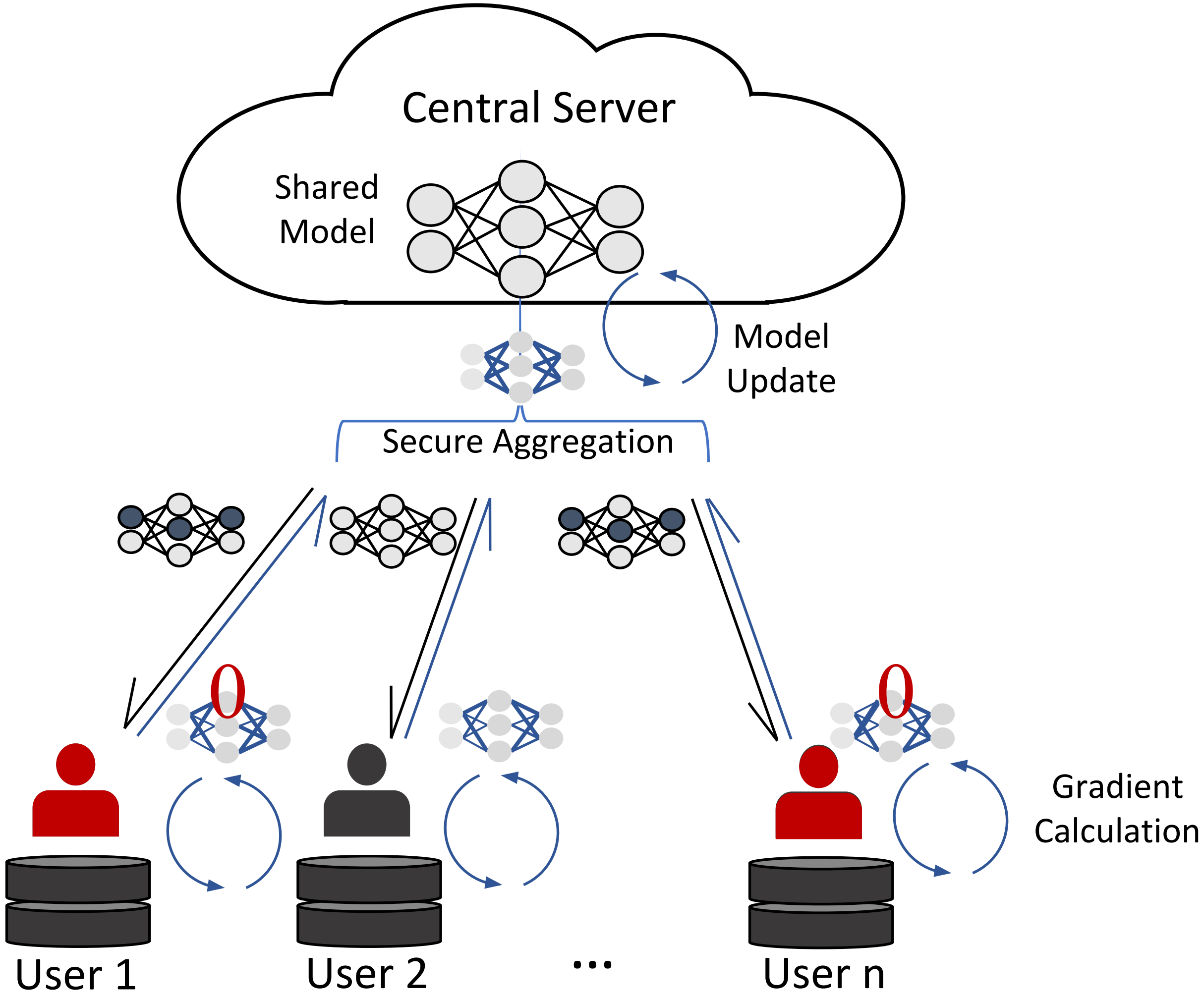


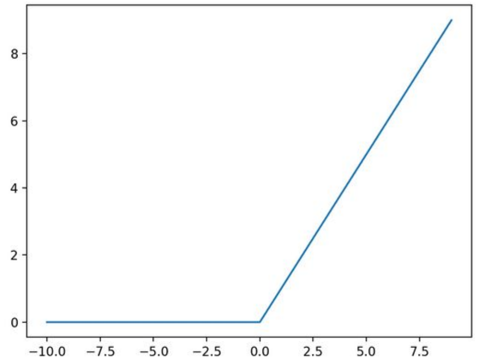
Figure taken from [4]. Results for mini-batch size of 64.

Extraction success increases with increasing the number of bins

[Fowl et al., 2021, ICLR]

Eluding Secure Aggregation





ReLU

Model Inconsistency

Gradient Suppression

$$ReLU\left(w_i^T x + b_i\right) = 0$$

[Pasquini et al., 2022, CCS]

Summary of my Contributions

1. Even with large mini-batches of high-dimensional data, significant proportions of private user data can be leaked to a passive attacker.

2. Active attackers can amplify this leakage even without performing highly noticeable changes to the model architecture / parameters.

3. Prior work has still largely underestimated the privacy risk of (hardened) FL.