Protection Against Reconstruction and Its Applications in Private Federated Learning

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, Ryan Rogers

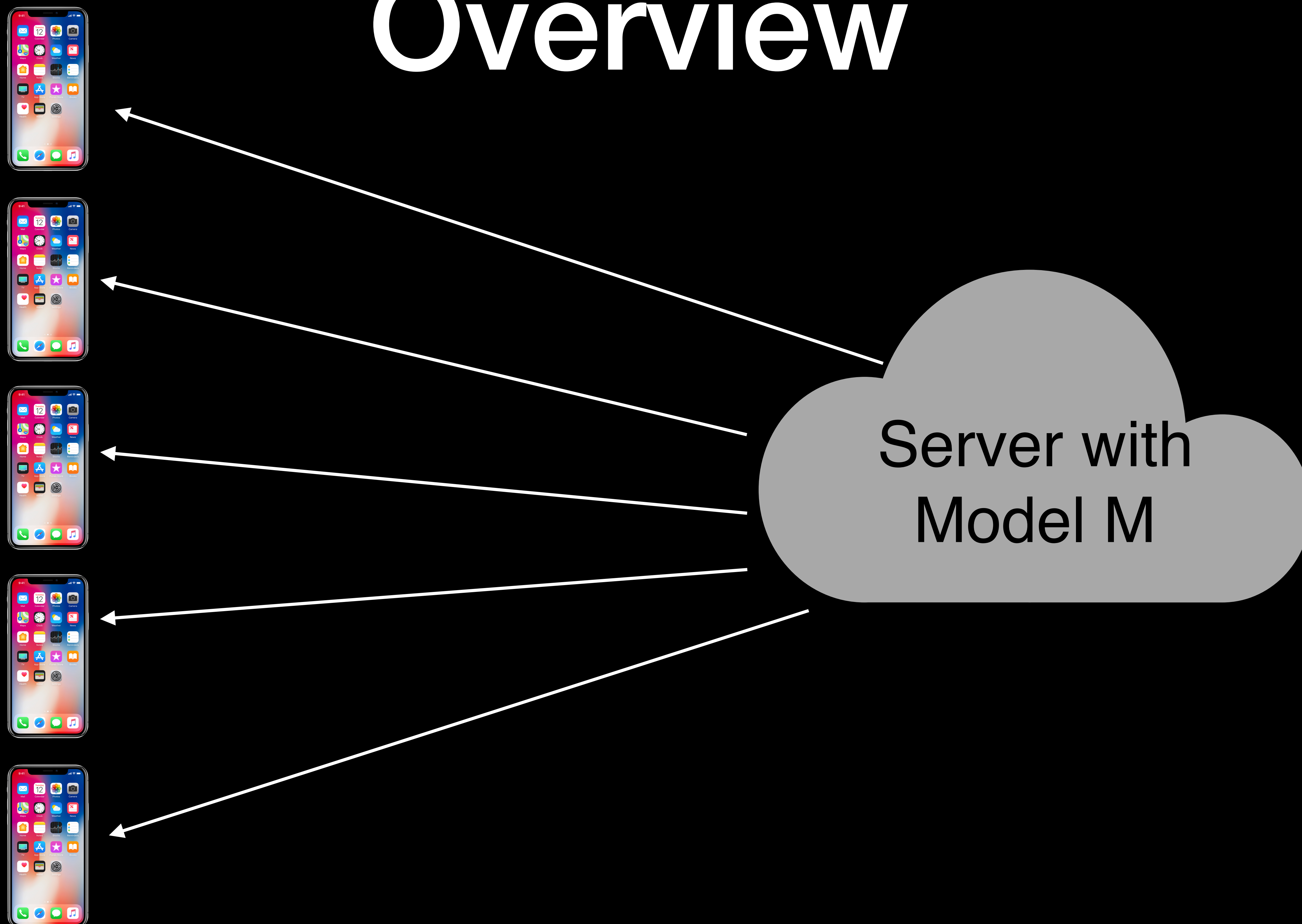
ML Privacy Team, Apple

Federated Learning

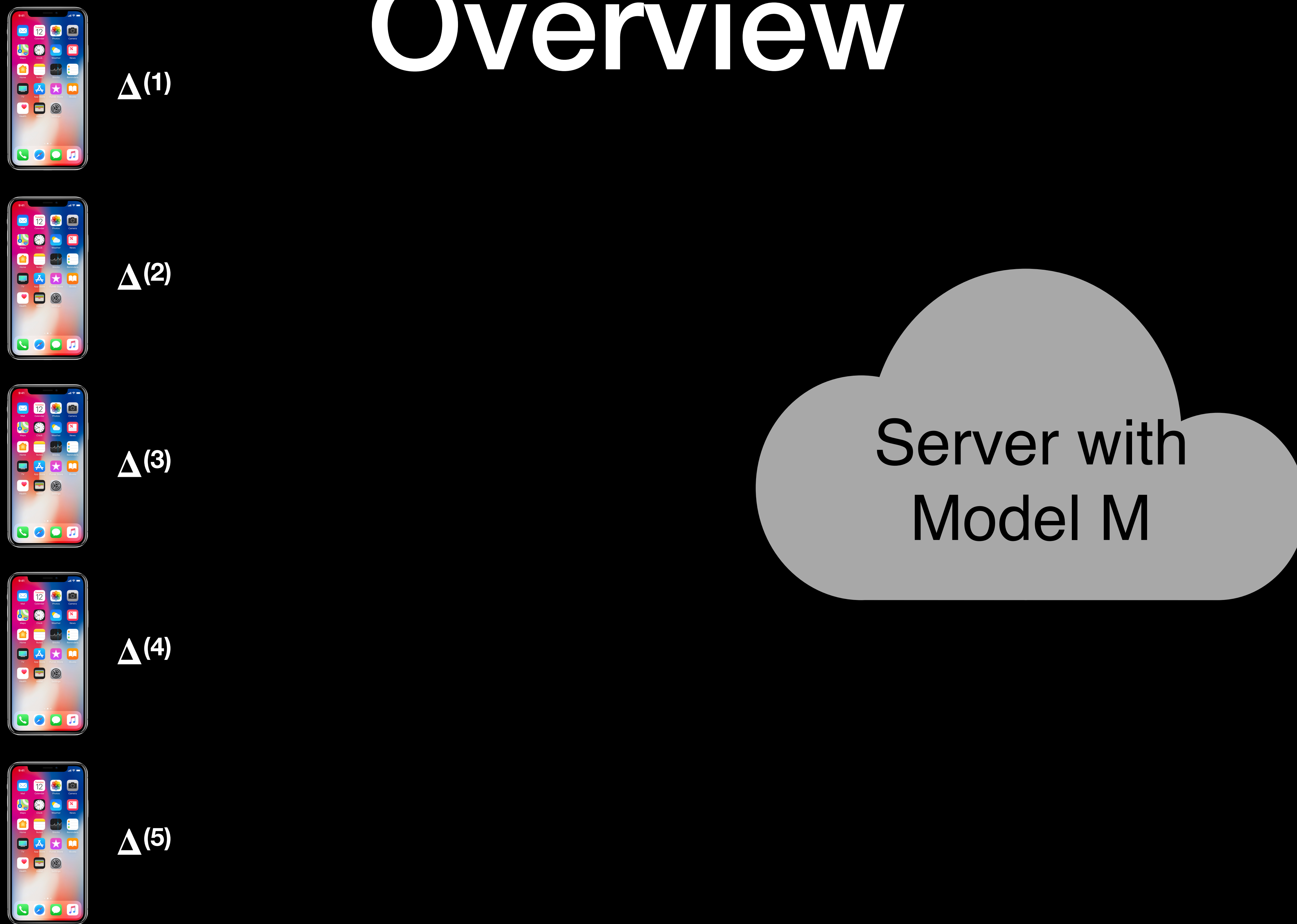
[MMRHA17]

- Lots of personal data is distributed across many devices
- We hope to improve machine learning models with this sensitive data.
- Devices are powerful enough now that they can do a lot of the computation.
- Rather than transmit data to a central server, have each device do the computation and only submit the update.

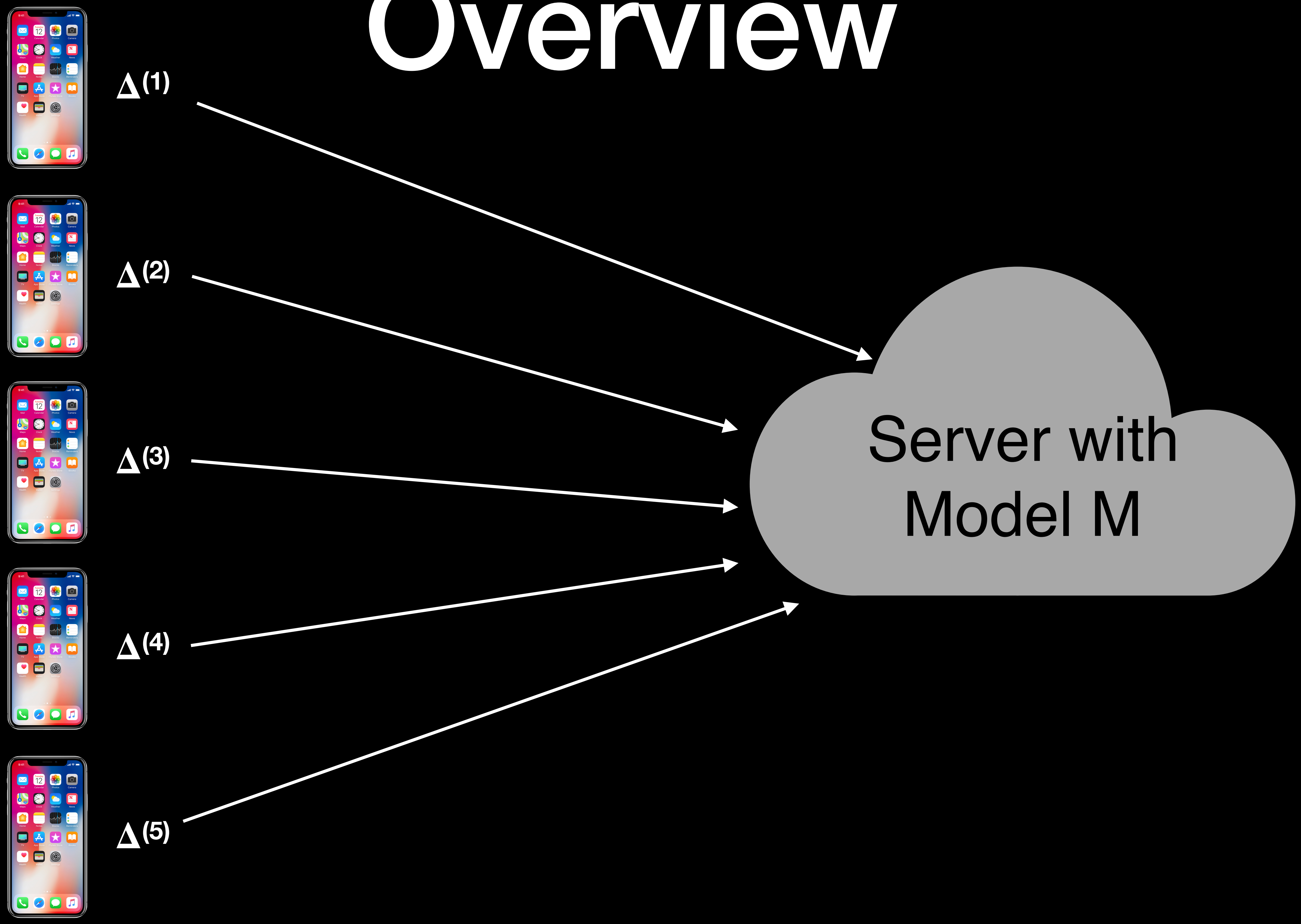
Federated Learning Overview



Federated Learning Overview



Federated Learning Overview



Federated Learning Overview

Server with
Model M

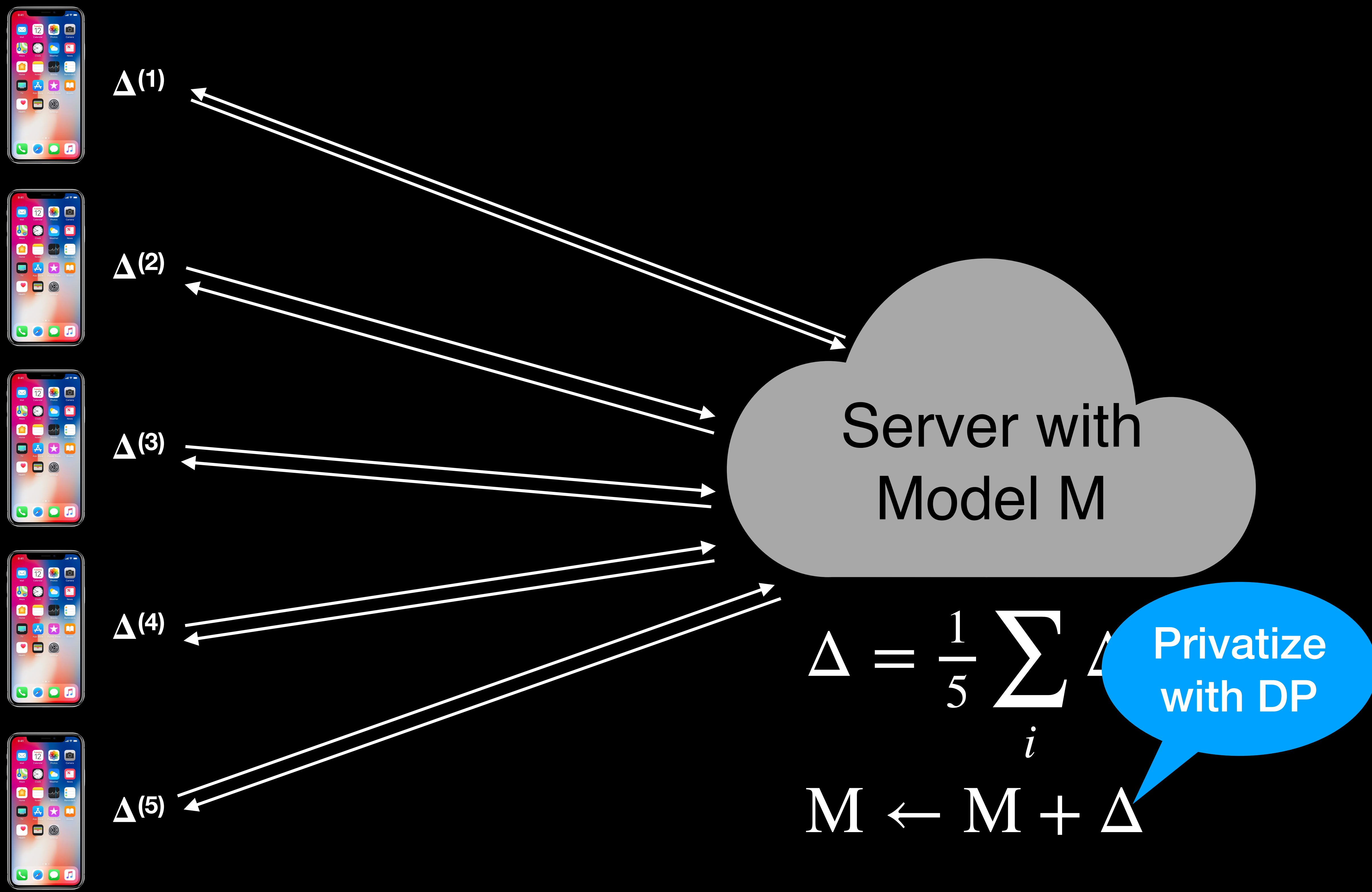
$$\Delta = \frac{1}{5} \sum_i \Delta^{(i)}$$

$$\mathrm{M} \leftarrow \mathrm{M} + \Delta$$

Privacy of Model

- Several users download the model at each round.
- **Attacks** - Models can memorize unique patterns [CLKES18].
- **Solution** - Use central DP on the aggregated model [SCS13, BST14, ACGMMTZ16, MRTZ18]
- Previous works show good privacy-utility tradeoffs in this setting.

Federated Learning



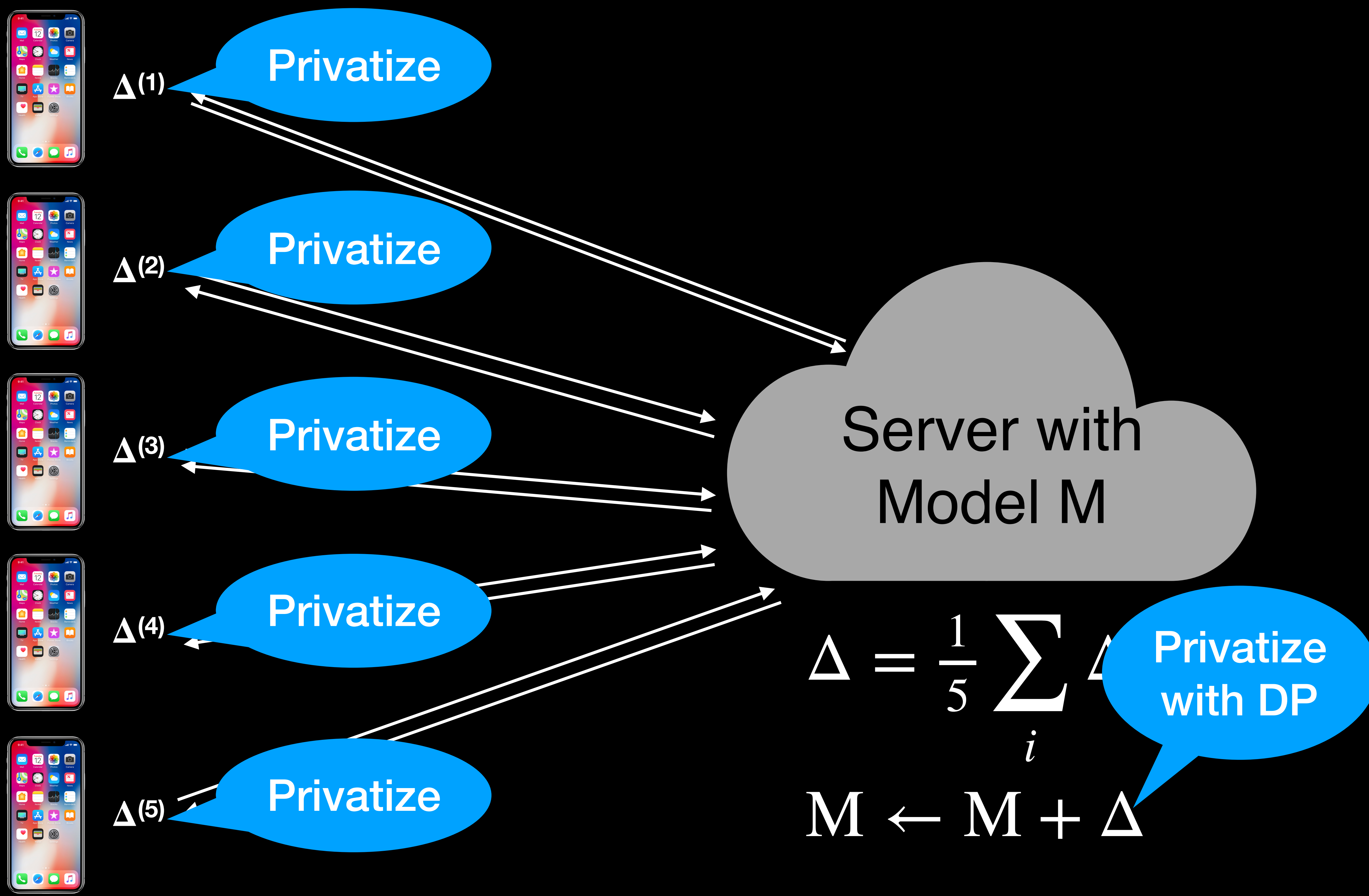
Privacy of the Updates

- Consider gradient methods with example-label pair *(x,y)* and generalized linear loss $\ell(\theta; x,y)$.
- Update from a device:

$$\nabla \ell(\theta; x, y) = \textbf{scalar} \cdot x$$

User's data

Federated Learning



Threat Model in Private FL

- We consider two different adversaries in our system.
- **Strong adversary** - can perform arbitrary inferences on the privatized model at each round of communication .
  - Protect with Central DP with small privacy parameters.
- **Curious onlooker** - can see privatized updates and wants to reconstruct some function of the input.
  - Protect with reasonable privacy parameters in Local DP.

Locally Private Updates

- Local differential privacy is a strong requirement that would ensure the privacy of the individual updates.

- [Warner65,EGS03,KLNRS08] An algorithm is ε-Local DP if for all inputs x,x' and outcome sets S we have

$$\frac{\mathbb{P}[A(x) \in S]}{\mathbb{P}[A(x') \in S]} \leq e^{\epsilon}$$

- [BNO13,DJW13,DJW18,DR18] - Strong lower bounds for estimating high dimensional vectors

Relaxing the Local Privacy Parameter

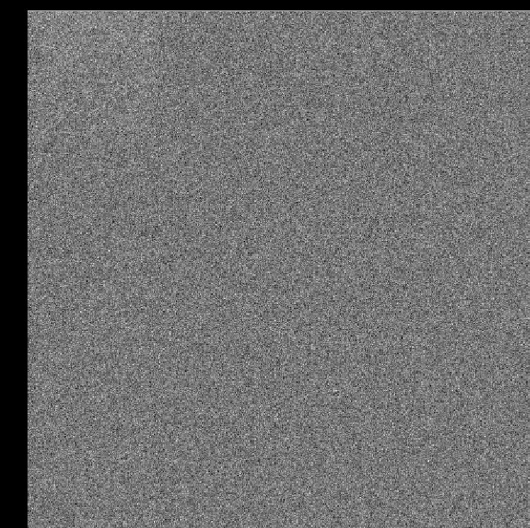
$$\frac{\mathbb{P}[A(x) \in S]}{\mathbb{P}[A(x') \in S]} \leq e^{\epsilon}$$

- Can we still provide privacy guarantees for larger ε?
- Protecting against arbitrary inferences requires ε = *O(1)*.
- Consider **specific** adversaries - curious onlookers who have limited information about the inputs and want to ***reconstruct*** the input.

Defining Reconstruction

Data $X \sim \pi$

Weights W



Privatized *Z* = *A(W)*

ϕ(Z)

For any z and estimator ϕ, we want:

$$\mathbb{P}[\,||X - \phi(z)||_2 < \alpha \mid Z = z] \ll 1$$

Reconstruction

$$X \to W \to Z = A(W)$$

- Adversary wants to reconstruct *X* or some *f(X)* given *Z* with some prior π over inputs.

- A normalized estimator ϕ causes an *(α,f,p)*-**reconstruction breach** if there exists a *z* such that

$$\mathbb{P}\left[\,||f(X) - \phi(z)||_2 < \alpha \mid A(W) = z\right] > p$$

- If no such estimator, then *A* protects against *(α,f,p)*-reconstruction.

Reconstruction

$$X \to W \to Z = A(W)$$

Priors?

- Adversary wants to reconstruct *X* or some *f(X)* given *Z* with some prior π over inputs.

Target functions?

- A normalized estimator ϕ causes an *(α,f,p)*-**reconstruction breach** if there exists a *z* such that

$$\mathbb{P}\left[\,||f(X) - \phi(z)||_2 < \alpha \mid A(W) = z\right] > p$$

- If no such estimator, then *A* protects against *(α,f,p)*-reconstruction.

Algorithms?

Reconstruction

$$X \to W \to Z = A(W)$$

Priors?

- Adversary wants t[illegible]uct *X* or some *f(X)* given *Z* with some prior π over inputs.

Target functions?

- A normalized estimator ϕ causes an *(α,f,p)*-**reconstruction breach** if there exists a *z* such that

$$\mathbb{P}\left[\,||f(X) - \phi(z)||_2 < \alpha \mid A(W) = z\right] > p$$

- If no such estimator, then *A* protects against *(α,f,p)*-reconstruction.

Algorithms?

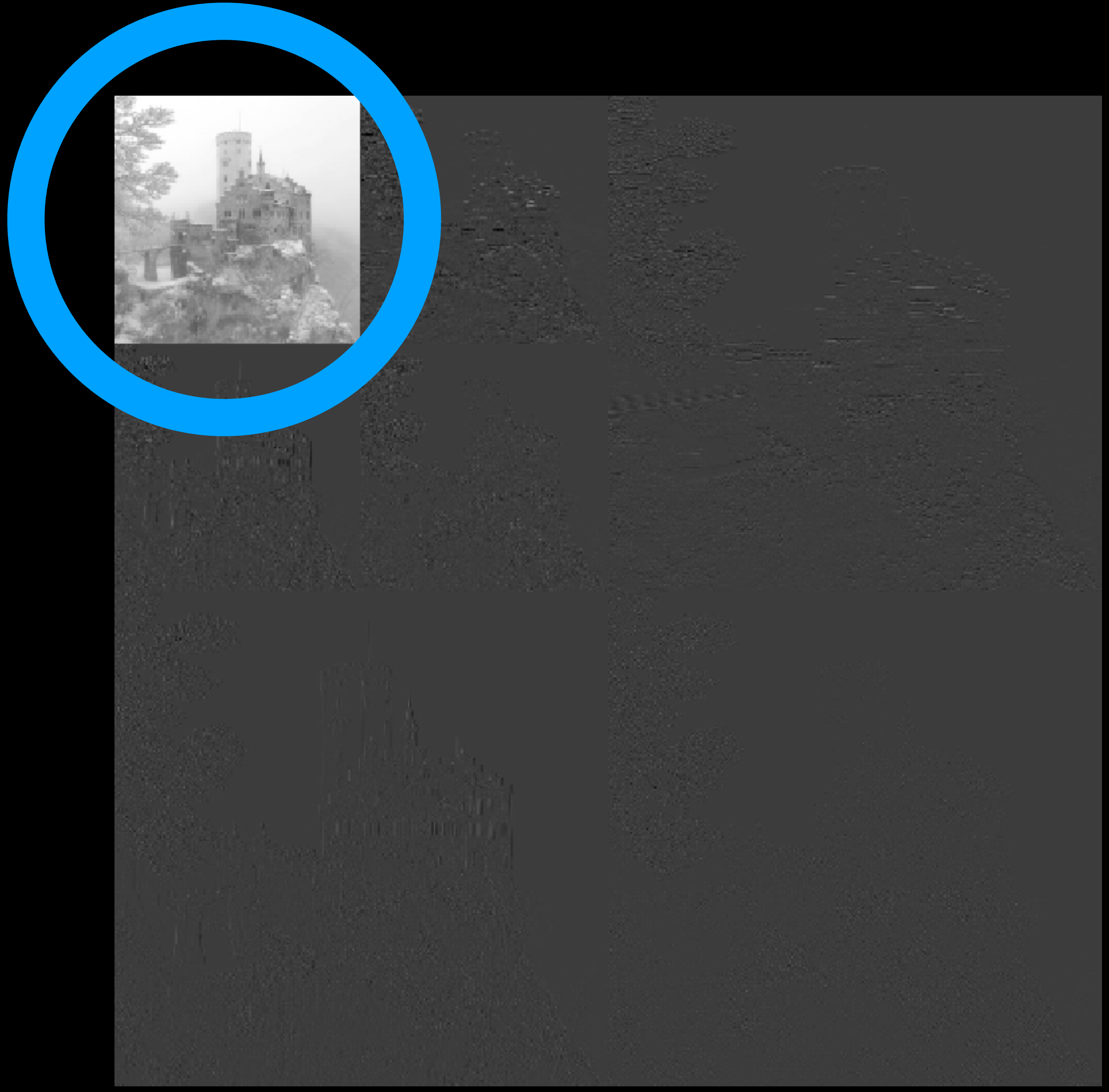
Target Functions

$$X = W \to Z = A(W)$$

- Target reconstruction function - projections

- Consider projection matrix *P* with *k*<*d*:

$$f_k(x) = \frac{Px}{||Px||_2}$$



Reconstruction

$$X = W \to Z = A(W)$$

Priors?

- Adversary wants t[…]uct *X* or some *f(X)* given *Z* with some prior π over inputs.

Target functions?

- A normalized estimator ϕ causes an *(α,f,p)*-**reconstruction breach** if there exists a *z* such that

$$\mathbb{P}\left[\,||f(X) - \phi(z)||_2 < \alpha \mid A(W) = z\right] > p$$

- If no such estimator, then *A* protect[…]inst *(α,f,p)*-reconstruction.

Algorithms?

Reconstruction

$$X = W \to Z = A(W)$$

Priors?

- Adversary wants to reconstruct *X* or some *f(X)* given *Z* with some prior π over inputs.

Target functions?

- A normalized estimator ϕ causes an *(*α*,f,p)*-**reconstruction breach** if there exists a *z* such that

$$\mathbb{P}\left[\,||f(X) - \phi(z)||_2 < \alpha \mid A(W) = z\right] > p$$

- If no such estimator, then *A* protects against *(*α*,f,p)*-reconstruction.

Algorithms?

Reconstruction

$$X = W \to Z = A(W)$$

Priors?

- Adversary wants to reconstruct *X* or some *f(X)* given *Z* with some prior π over inputs.

Target functions?

- A normalized estimator ϕ causes an *(***α***,f,p)***-reconstruction breach** if there exists a *z* such that

$$\mathbb{P}\left[\,||f(X) - \phi(z)||_2 < \alpha \mid A(W) = z\right] > p$$

- If no such estimator, then *A* protects against *(***α***,f,p)*-reconstruction.

Algorithms?

DP Protects Against Reconstruction

- Consider a diffuse prior π. If *A* is ε-DP then *A* protects against *(α, f_k, p)*-reconstruction where

$$p = \exp\left(\epsilon + c \cdot k \log(\alpha^2 \cdot (1 - \alpha^2/4))\right)$$

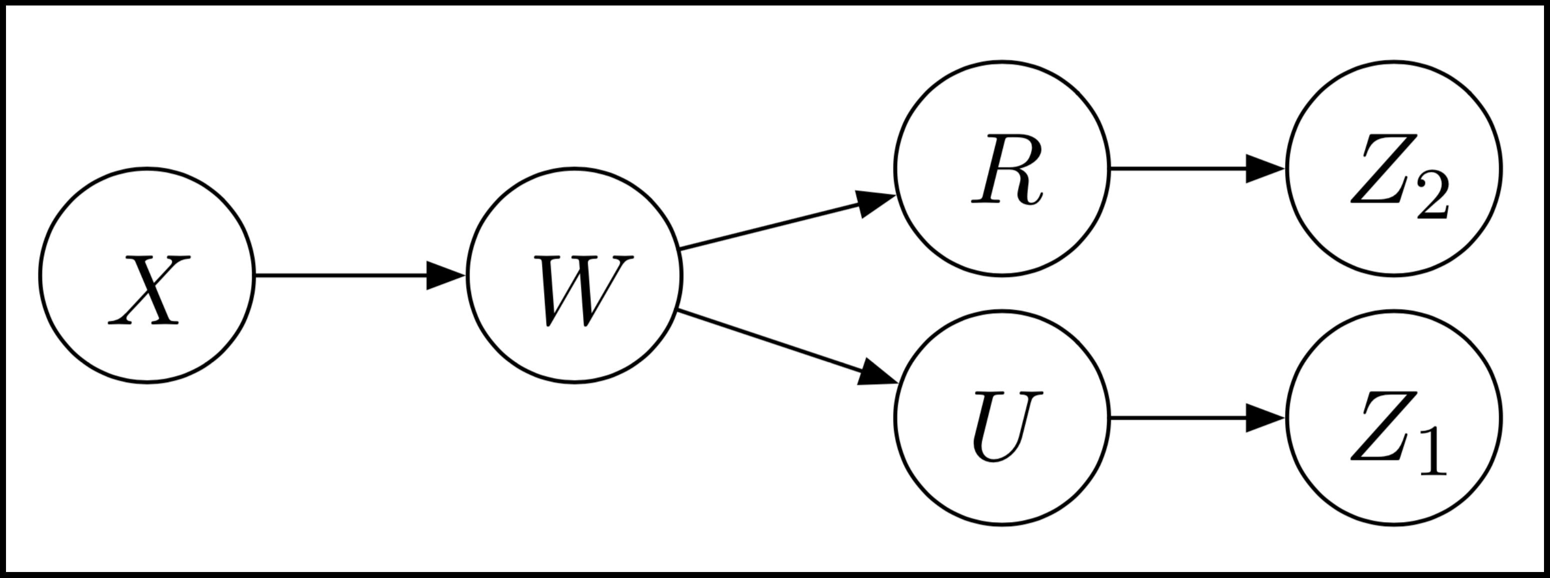
- We can obtain a small probability of reconstruction even for large ε.

Separated DP

- To privatize high dimensional vectors, we will decompose vector *W* into a unit vector *U* and its magnitude *R*.

$$W = \underbrace{\frac{W}{||W||_2}}_{U} \cdot \underbrace{||W||_2}_{R}$$

- We design DP algorithms to privatize *U* and *R* separately.



Existing Local DP Algorithms

- Let's use a local DP algorithm to privatize high dimensional unit vectors.

- Consider a unit vector $u \in \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : ||v||_2 = 1\}$.

- Add mean zero, independent noise: $A(u) = u + N$, then

$$\mathbb{E}\left[\,||A(u) - u\,||_2^2\right] = \Theta\left(\frac{d^2}{\epsilon^2}\right)$$

- [DJW13] - Sampling scheme with better dependence on d

$$\mathbb{E}\left[\,||A(u) - u\,||_2^2\right] = \Theta\left(d\left(\frac{e^\epsilon + 1}{e^\epsilon - 1}\right)^2\right)$$

Existing Local DP Algorithms

- Let's use a local DP algorithm to privatize high dimensional unit vectors.

- Consider a unit vector $u \in \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : ||v||_2 = 1\}$.

- Add mean zero, independent noise: $A(u) = u +$

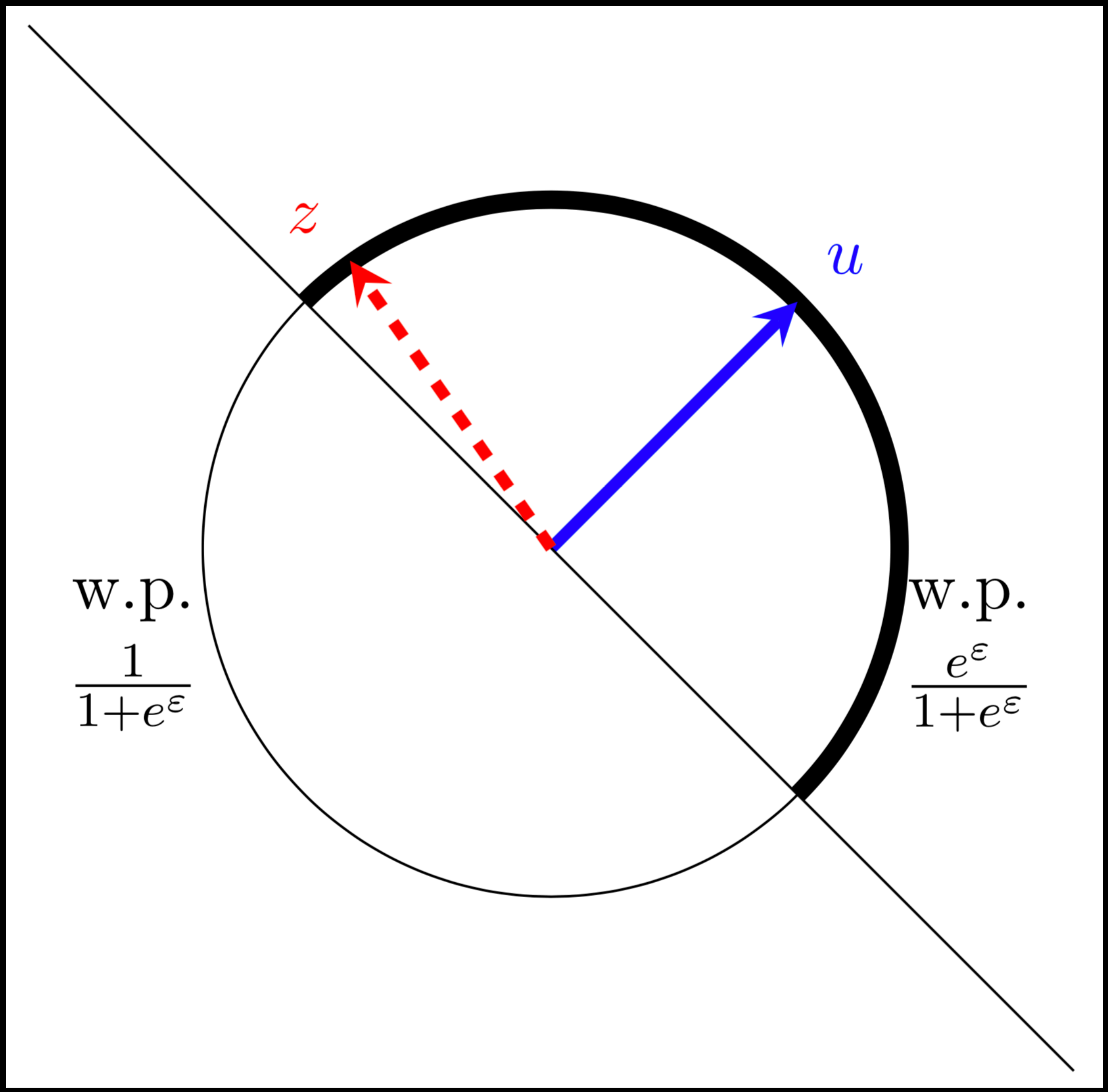
$$\mathbb{E}\left[||A(u) - u||_2^2\right] = \Theta\left(\frac{d^2}{\epsilon^2}\right)$$

Optimal for $\varepsilon = O(1)$, but not for larger ε

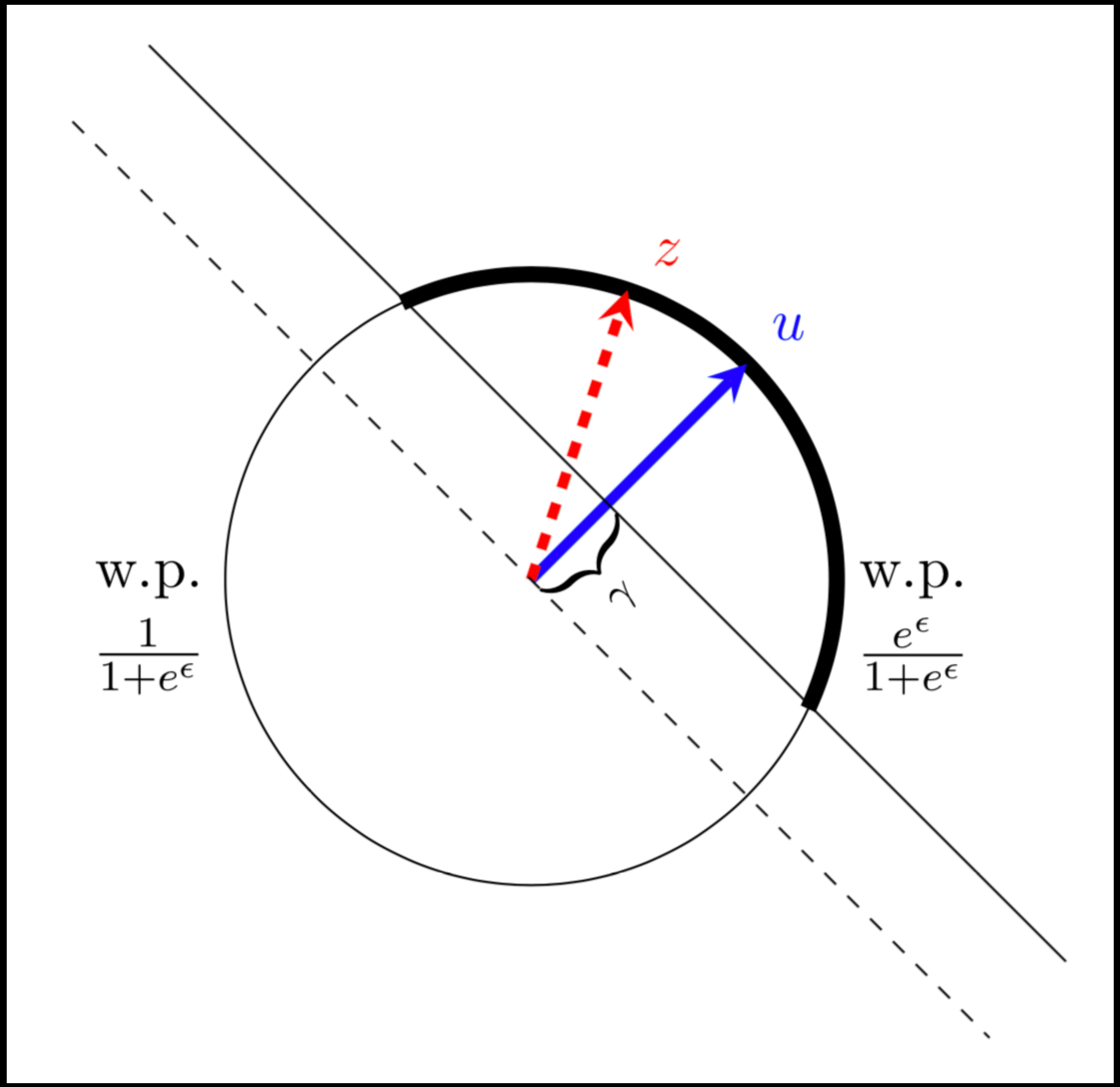
- [DJW13] - Sampling scheme with better dependence on d

$$\mathbb{E}\left[||A(u) - u||_2^2\right] = \Theta\left(d\left(\frac{e^\epsilon + 1}{e^\epsilon - 1}\right)^2\right)$$

Privatize Unit Vectors



[DJW13]



This Work - PrivUnit$(u;\gamma,\varepsilon)$

Privatize Unit Vectors

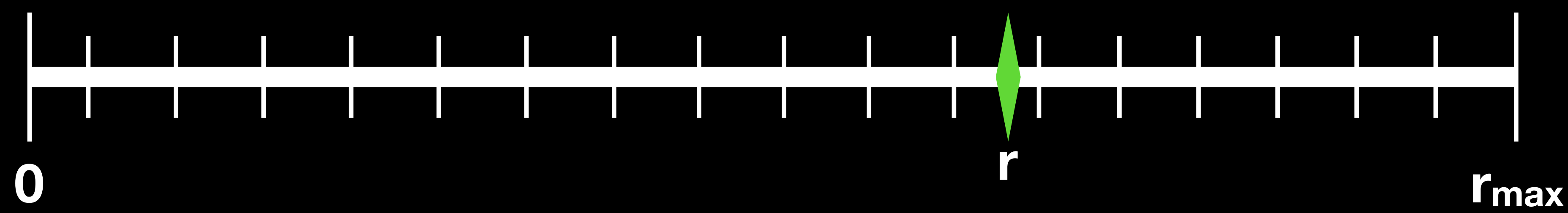
- Let Z = *PrivUnit(u;γ,0)* with $\gamma \approx \sqrt{\frac{\epsilon}{d}}$ then PrivUnit is ε-DP.
- Further, $\mathbb{E}[Z] = u, \quad \mathbb{E}\left[||Z - u||_2^2\right] = O\left(\frac{d}{\epsilon \wedge \epsilon^2}\right)$
- This is optimal.

Privatize Magnitude

ScalarDP($r;\varepsilon,r_{max}$)

Privatize Magnitude

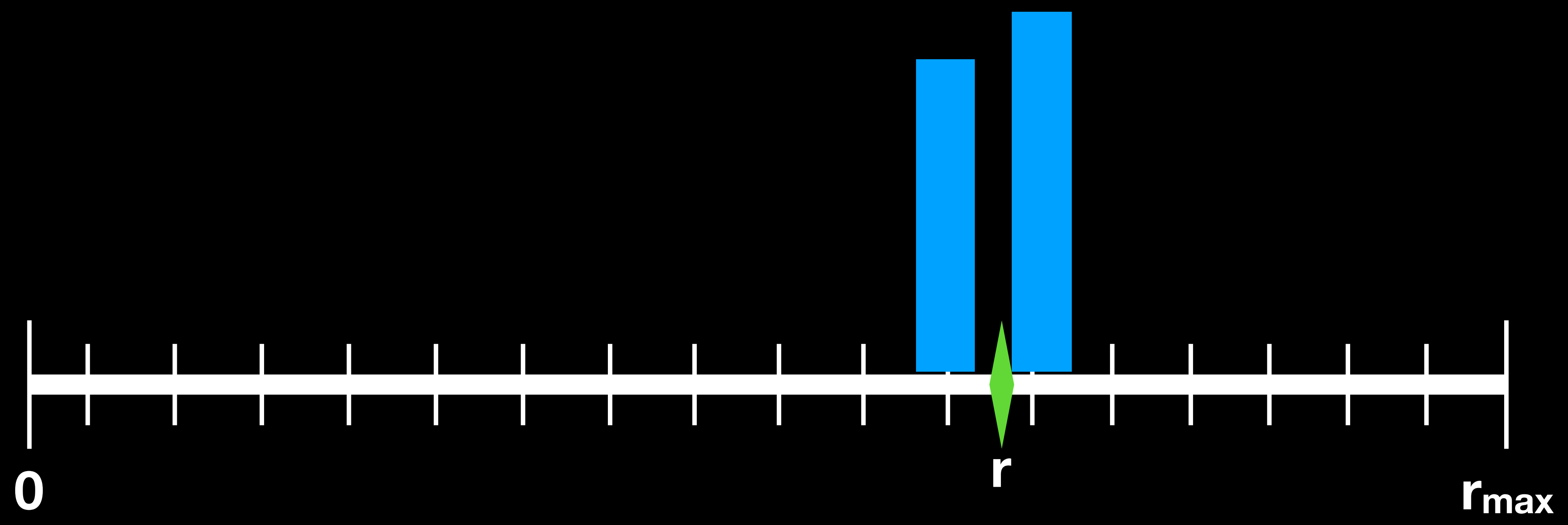
ScalarDP(*r*;*ε*,*r_{max}*)



Discretize into *k* = *exp(ε/3)* bins

Privatize Magnitude

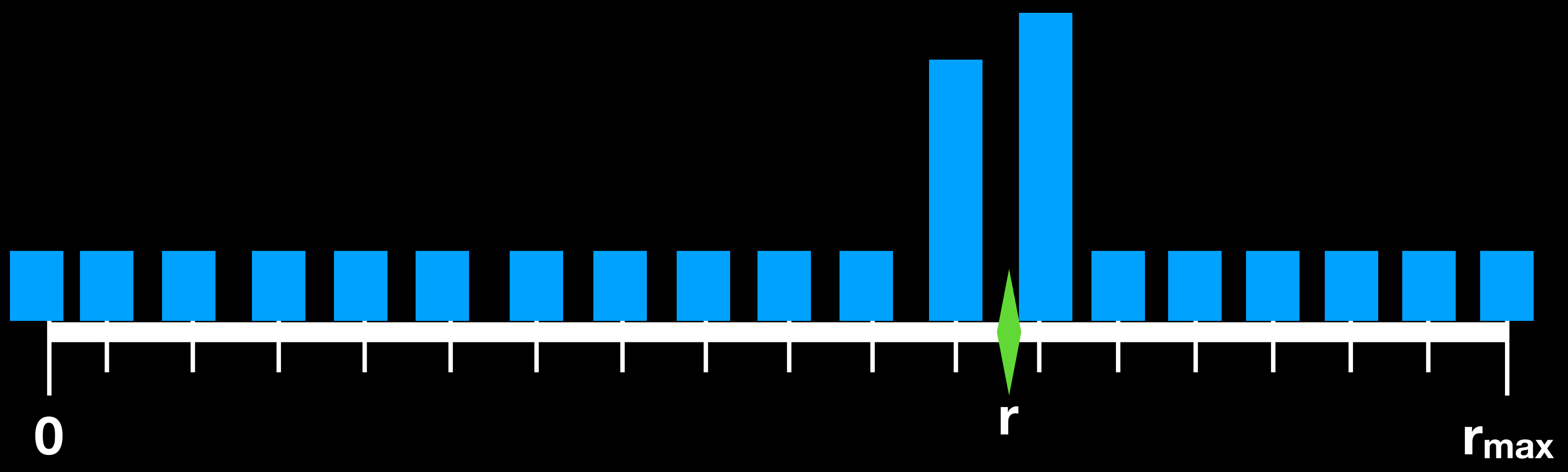
ScalarDP(*r;ε,r_{max}*)



Discretize into *k = exp(ε/3)* bins

Privatize Magnitude

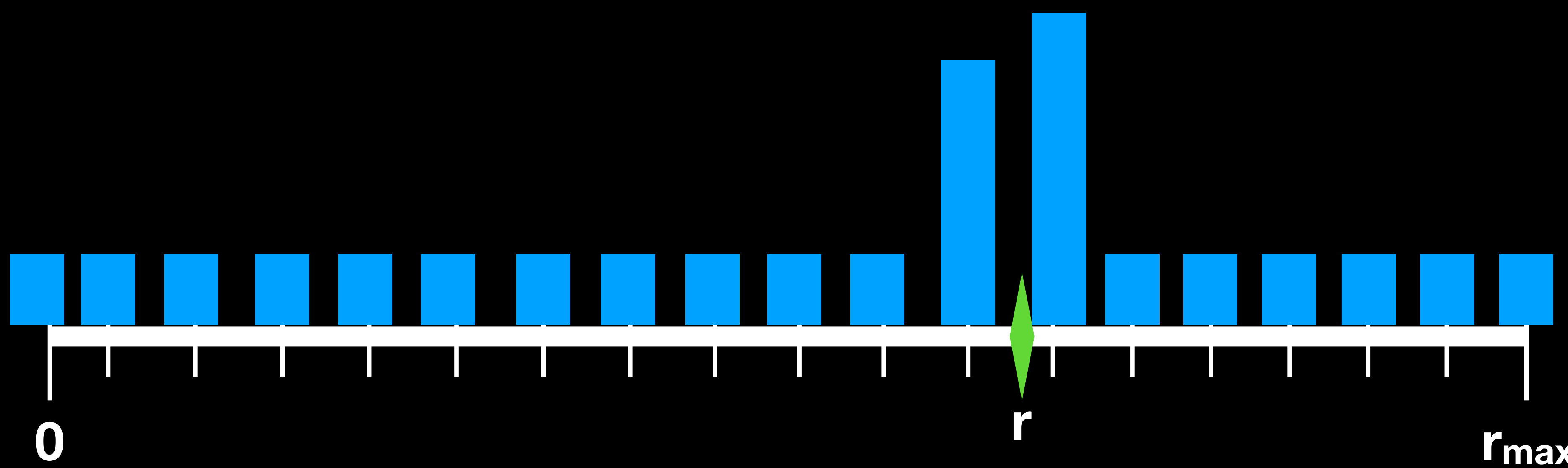
ScalarDP($r; \varepsilon, r_{max}$)



Discretize into $k = exp(\varepsilon/3)$ bins

Privatize Magnitude

ScalarDP$(r;\varepsilon,r_{max})$



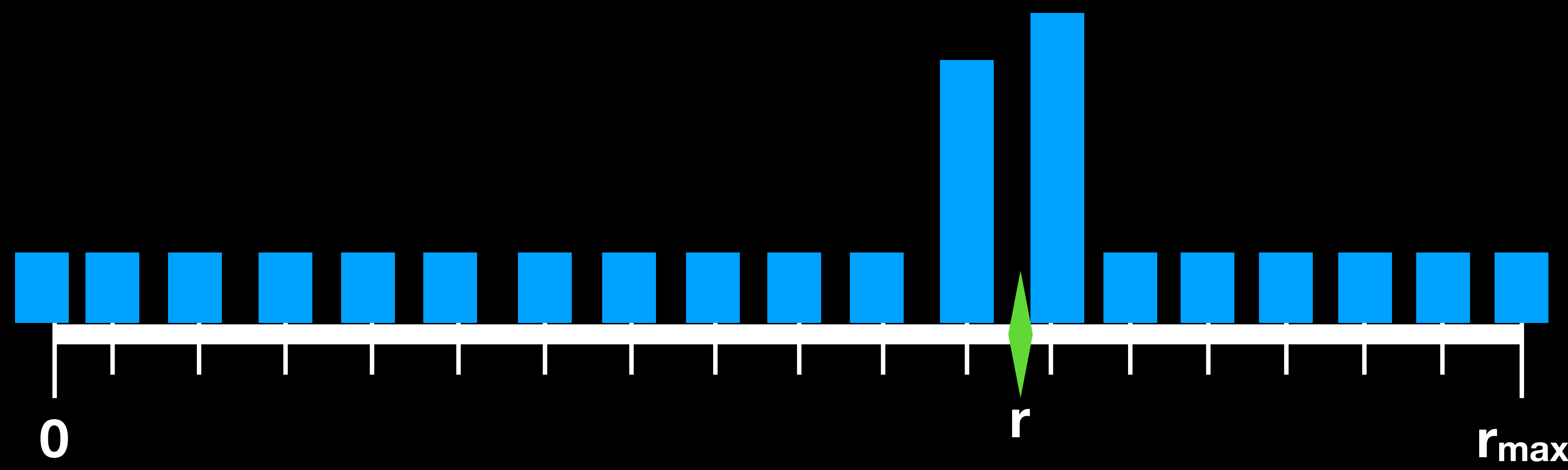
Discretize into $k = exp(\varepsilon/3)$ **bins**

$Z = PrivMagn(r;\varepsilon,r_{max})$ **is** ε**-DP**

and $\mathbb{E}[(Z - r)^2] = O(r_{max}^2\ exp(-2\varepsilon/3))$

Privatize Magnitude

ScalarDP$(r; \varepsilon, r_{max})$



Discretize into $k = exp(\varepsilon/3)$ bins

[GV16] Optimal for ε>1

$Z = PrivMagn(r; \varepsilon, r_{max})$ **is ε-DP**

and $\mathbb{E}[(Z - r)^2] = O(r_{max}^2 \, exp(-2\varepsilon/3))$

Optimality

- Consider stochastic gradient descent with example label pairs *(x,y)* with $||x|| \leq r$ and $y \in \{-1,1\}$.

- Using our local DP mechanisms, we have

$$n\left(L(\bar{\theta}_n) - L(\theta^\star)\right) \xrightarrow{\mathrm{d}} T^2$$

$$\mathbb{E}[T^2] = O\left(r^2 \cdot \frac{d}{\varepsilon \wedge \varepsilon^2}\right)$$

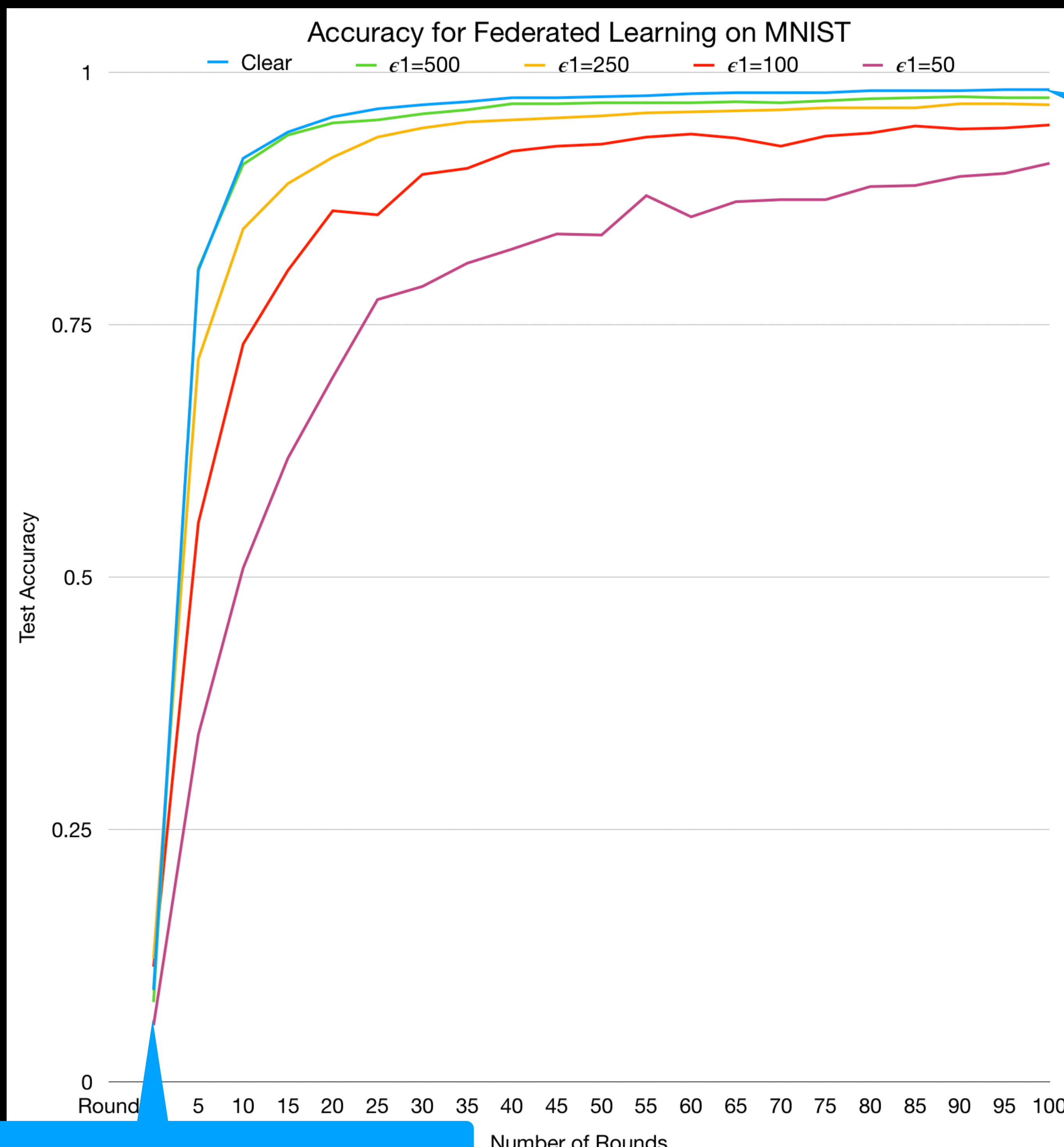
- This is minimax optimal for any arbitrarily interactive local-DP algorithm [D**R**19]

Experiments

Experiments		
Task	Dataset	d
Image Classification over 10 Classes	MNIST	3,274,634
Image Classification over 10 Classes	CIFAR10	1,068,298
Image Classification over 100 Classes	Flickr	1,255,524
Next Word Prediction	REDDIT	13,352,875

- We conducted experiments for various tasks and models.
- We used our local DP algorithms (PrivUnit and ScalarDP) to protect against reconstruction.
- We also clipped each model update and added Gaussian noise to the aggregate update for central DP.

MNIST

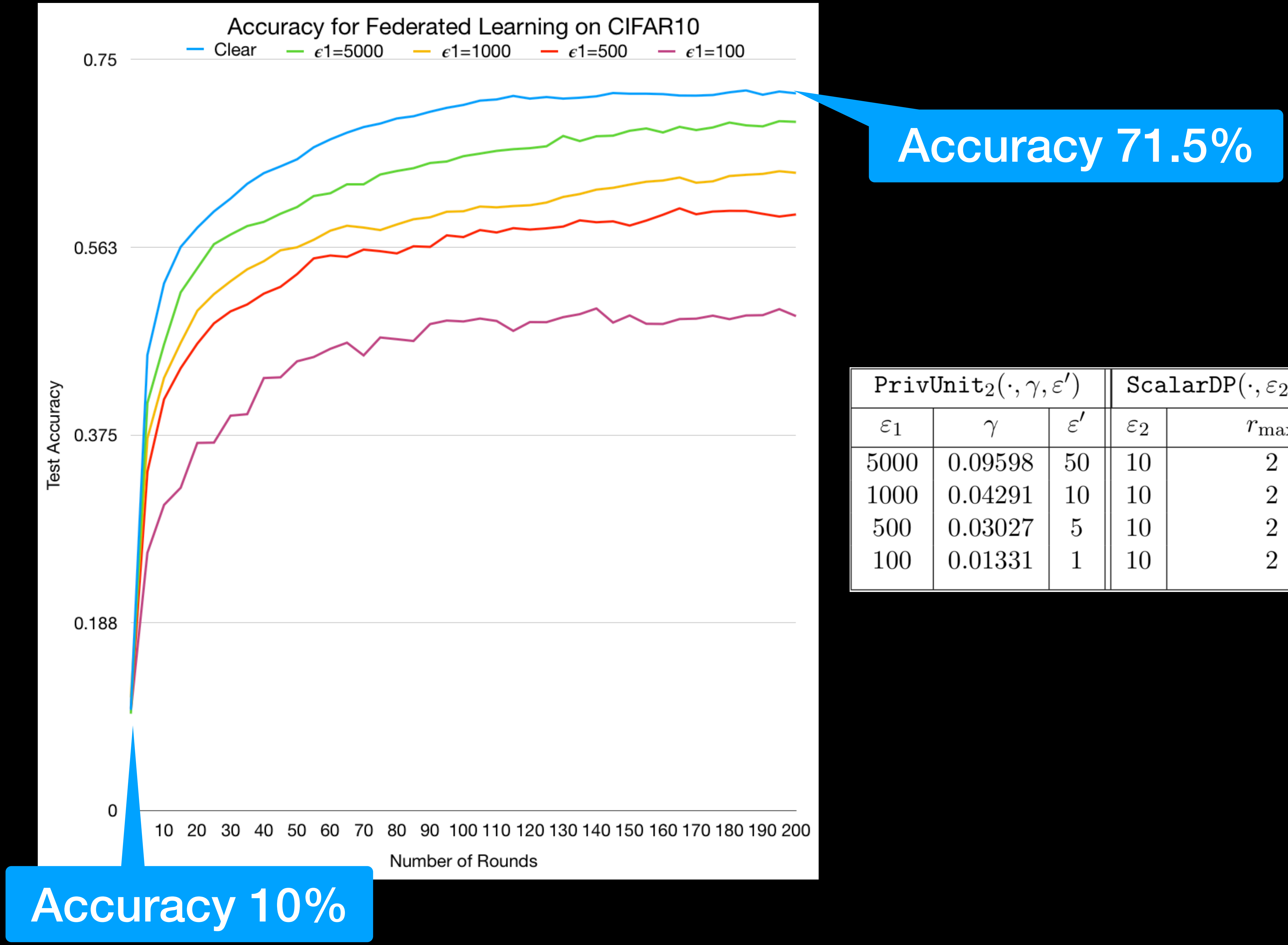


Accuracy 98.8%

Accuracy 10%

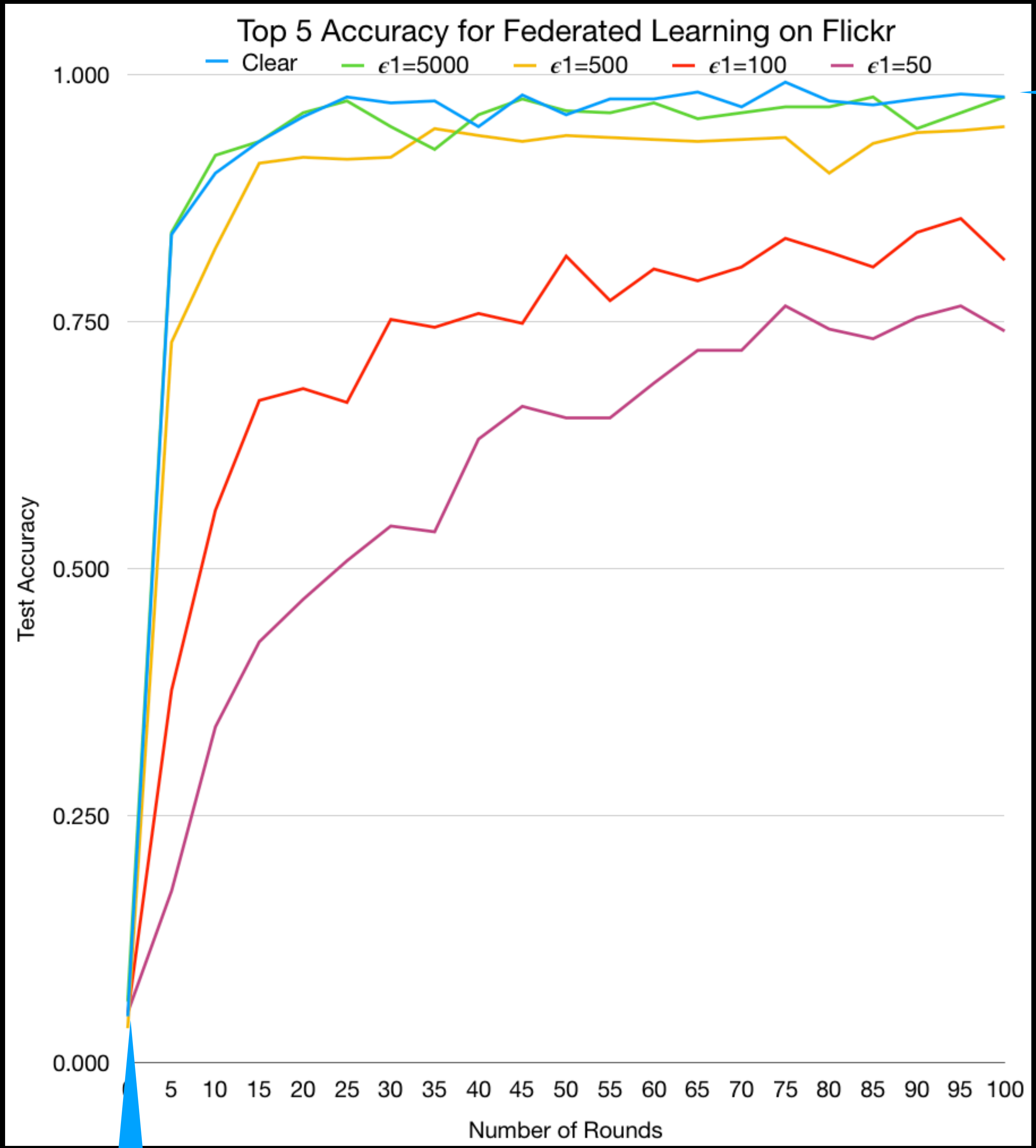
$\texttt{PrivUnit}_2(\cdot,\gamma,\varepsilon')$			$\texttt{ScalarDP}(\cdot,\varepsilon_2,r_{\max})$		Central DP	
ε_1	γ	ε'	ε_2	$r_{\max}$	Clip	σ
500	0.01729	5	10	5	100	0.005
250	0.01217	2.5	10	5	100	0.005
100	0.00760	1	10	5	100	0.005
50	0.00526	0.5	10	5	100	0.005

CIFAR10



$\texttt{PrivUnit}_2(\cdot, \gamma, \varepsilon')$			$\texttt{ScalarDP}(\cdot, \varepsilon_2, r_{\max})$		Central DP	
ε_1	γ	ε'	ε_2	$r_{\max}$	Clip	σ
5000	0.09598	50	10	2	30	0.002
1000	0.04291	10	10	2	30	0.002
500	0.03027	5	10	2	30	0.002
100	0.01331	1	10	2	30	0.002

ResNet50v2



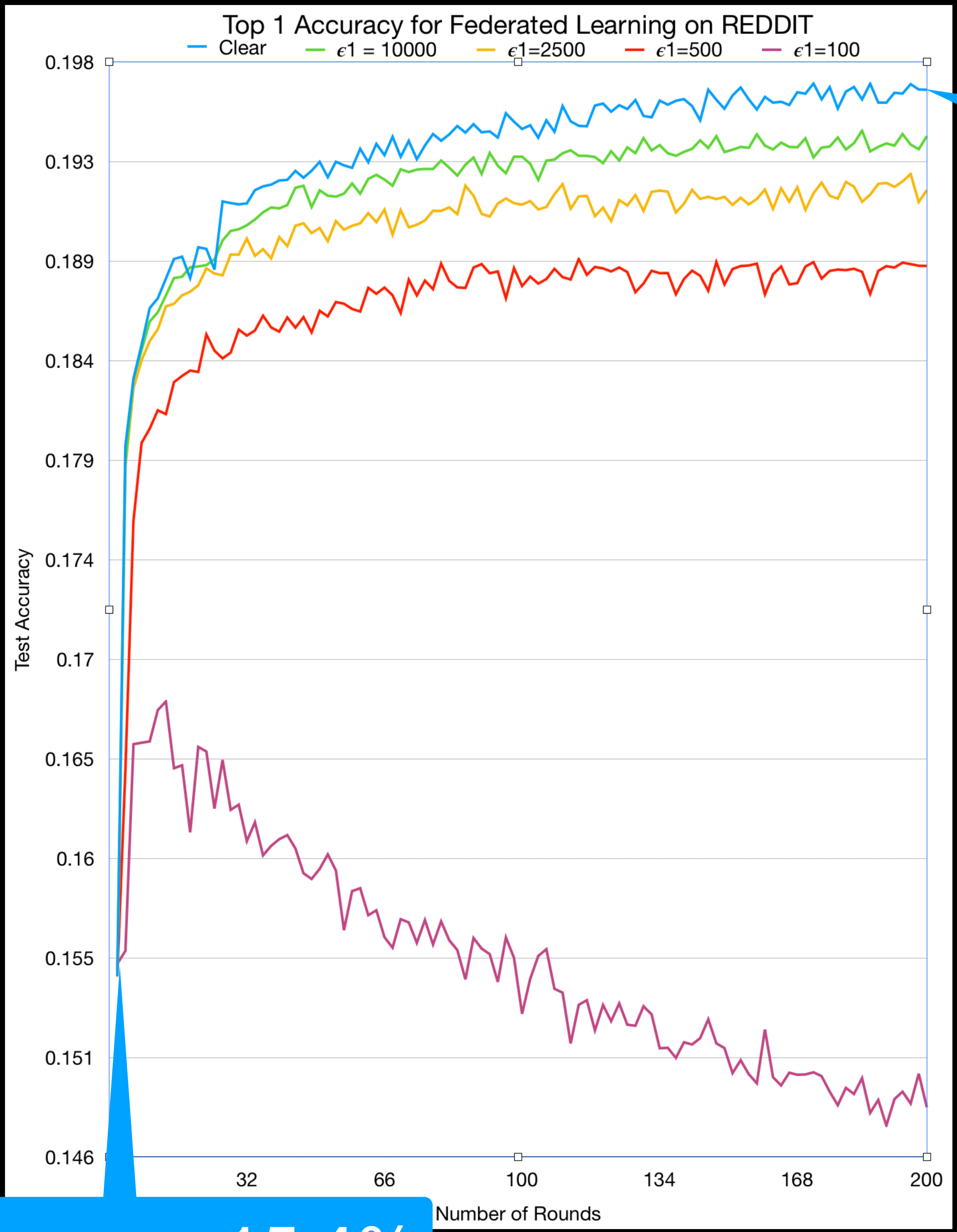
Top 5 Accuracy 97.7%

Accuracy 5%

- Pretrained ResNet50v2 on ImageNet
- Further trained last two layers on Flickr data with 100 classes.

$\texttt{PrivUnit}_2(\cdot,\gamma,\varepsilon')$			$\texttt{ScalarDP}(\cdot,\varepsilon_2,r_{\max})$		Central DP	
ε_1	γ	ε'	ε_2	$r_{\max}$	Clip	σ
5000	0.08857	50	10	10	100	0.005
500	0.02793	5	10	10	100	0.005
100	0.01227	1	10	10	100	0.005
50	0.00851	0.5	10	10	100	0.005

LSTM



Accuracy 19.5%

Accuracy 15.4%

- Pretrained LSTM on Wikipedia
- Further trained on Reddit comments from Nov 2017.

$\texttt{PrivUnit}_2(\cdot,\gamma,\varepsilon')$			$\texttt{ScalarDP}(\cdot,\varepsilon_2,r_{\max})$		Central DP	
ε_1	γ	ε'	ε_2	$r_{\max}$	Clip	σ
10000	0.03848	100	10	5	100	0.001
2500	0.01923	25	10	5	100	0.001
500	0.00856	5	10	5	100	0.001
100	0.00376	1	10	5	100	0.001

Thanks

References

- [MMRHA17] - McMahan, Moore, Ramage, Hampson, Arcas. ***Communication-Efficient Learning of Deep Networks from Decentralized Data***. AISTATS'17.
- [CLKES18] - Carlini, Liu, Kos, Erlingsson, Song. ***The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets***. arXiv1802.08232.
- [MRT18] - McMahan, Ramage, Talwar. ***Learning Differentially Private Recurrent Language Models***. ICLR'18.
- [SCS13] - Song, Chauduri, Sarwate. ***Stochastic Gradient Descent with DP updates***. GlobalSIP13
- [BST14] - Bassily, Smith, Thakurta. ***Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds.*** FOCS14.
- [ACGMMTZ16] - Abadi, Chi, Goodfellow, McMahan, Mironov, Talwar, Zhang. ***Deep Learning with Differential Privacy.*** CCS16.
- [Warner65] - Warner. *Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias*. JASA March1965.
- [EGS03] - Evfimievski, Gehrke, Srikant. ***Limiting Privacy Breaches in Privacy Preserving Data Mining.*** PODS03.
- [KLNRS08] - Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith. ***What can we Learn Privately?*** FOCS08.
- [BNO08] - Beimel, Nissim, Omri. ***Distributed private data analysis: Simultaneously solving how and what***. Adv. in Crypto'08.
- [DJW13] - Duchi, Jordan, Wainwright. ***Local privacy and statistical minimax rates***. FOCS'13.
- [DJW18] - Duchi, Jordan, Wainwright. ***Minimax optimal procedures for locally private estimation***. JASA'18.
- [DR18] - Duchi, Ruan. ***The right complexity measure in locally private estimation: It is not the Fisher information***. arXiv 1806.05756.
- [D**R**18] - Duchi, Rogers. ***Lower Bounds for Locally Private Estimation via Communication Complexity.*** arXiv 1902.00582.
- [GV16] - Geng, Viswanath. ***The Optimal Noise-Adding Mechanism***. Transactions of Information Theory. 2016.