18734: Foundations of Privacy

Learning with Privacy

Giulia Fanti

Fall 2019

Administrative

- HW3 due next Monday, 11.59 pm ET
- Friday: Mid-semester break
  - No recitation
  - I will hold regular office hours (3-4 pm ET, CIC 2118)

Canvas quiz

- 10 minutes

What is the downside of LDP?

- Higher ϵ requires more data
  - Train models
  - Release statistics with given accuracy

- How much more?

How would you evaluate this?

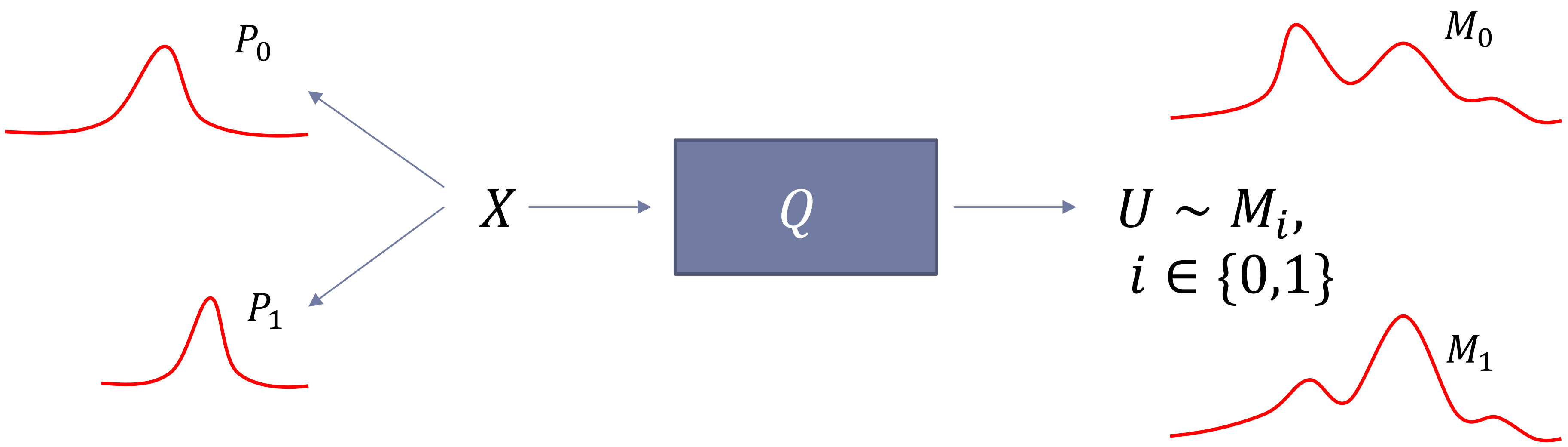
Local Privacy and Statistical Minimax Rates

John C. Duchi[†] Michael I. Jordan[†,*], and Martin J. Wainwright[†,*]
Department of Electrical Engineering and Computer Science[†] *and Department of Statistics*[*]
University of California, Berkeley
{jduchi,jordan,wainwrig}@eecs.berkeley.edu

Extremal Mechanisms for Local Differential Privacy

Peter Kairouz[1] **Sewoong Oh**[2] **Pramod Viswanath**[1]
[1]Department of Electrical & Computer Engineering
[2]Department of Industrial & Enterprise Systems Engineering
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA
{kairouz2,swoh,pramodv}@illinois.edu

Formulate problem as hypothesis test

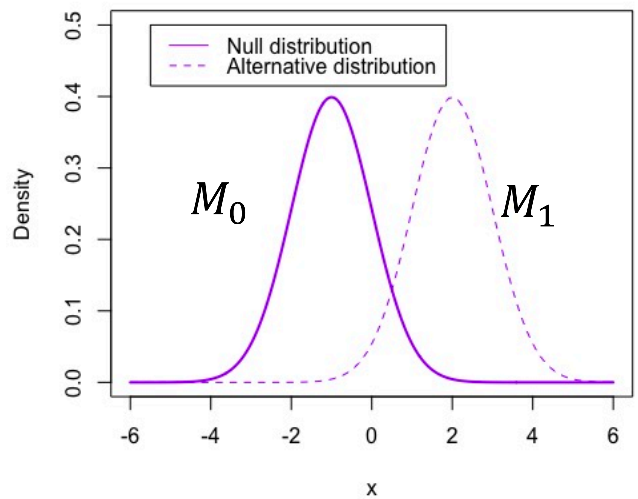


Q: Can we tell if we are observing samples from M_0 or M_1?

A: It depends how far apart they are!

Recall: Hypothesis Testing

- Null hypothesis: $H_o: U \sim M_0$
- Alternate hypothesis: $H_a: U \sim M_1$



Type I error: probability of rejecting H_0 when it's true

Type II error: probability of accepting H_0 when it's false

Chernoff-Stein Lemma

- (Informal). Consider the class of hypothesis tests with bounded Type I error probability. The best type II error over all such tests scales as

$$e^{-nD_{KL}(M_0||M_1)}$$

where $D_{KL}(M_0||M_1)$ denotes the KL-divergence between distributions M_0 and M_1:

$$D_{KL}(P||Q) = -\sum_{x\in\mathcal{X}} P(X)\log\left(\frac{Q(x)}{P(x)}\right)$$

Q: How is KL-divergence related to concept we saw in the ML lecture?

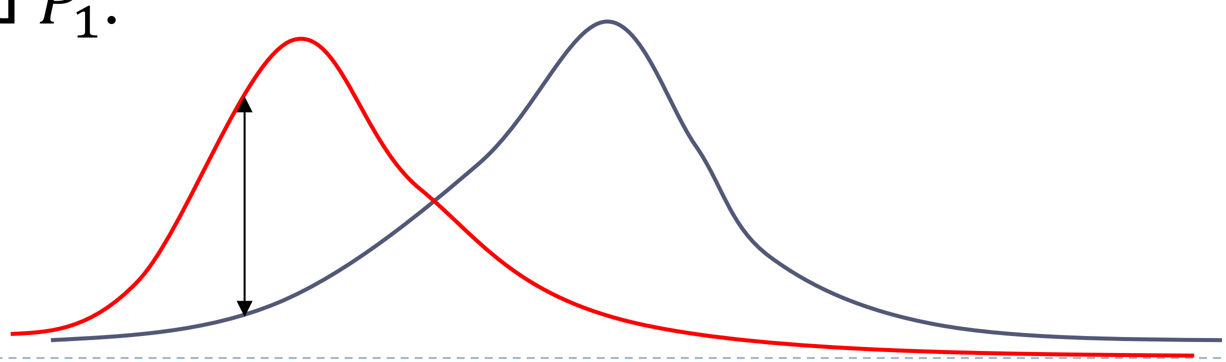
Main result
[Duchi, Jordan, Wainwright, 2013]

$$D_{KL}(M_0||M_1) \lesssim \epsilon^2 n \|P_0 - P_1\|_{TV}^2$$

Where

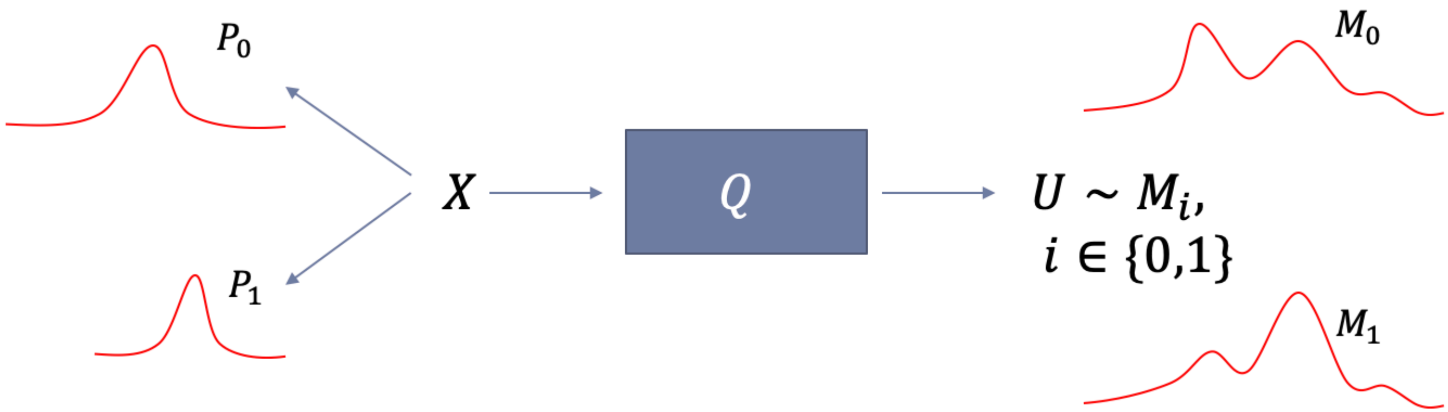
$$\|P_0 - P_1\|_{TV}^2 := \frac{1}{2}\sum_{x \in \mathcal{X}} |P_0(x) - P_1(x)|$$

denotes the total variation distance between distributions P_0 and P_1.



x

What is this saying?



Type II error scales as $e^{-nD_{KL}(M_0||M_1)}$

Result: $D_{KL}(M_0||M_1) \lesssim \epsilon^2 n \|P_0 - P_1\|_{TV}^2$

=> DP is hindering our ability to do hypothesis testing (consider $\epsilon < 1$)

Check your understanding

$$D_{KL}(M_0||M_n) \lesssim \epsilon^2 n \|P_0 - P_1\|_{TV}^2$$

- Suppose I previously needed n_0 samples to reach a certain accuracy for my estimator.
- Q: How many samples do I need if each sample is collected with ϵ-differential privacy?
- A: Order-wise: $\Omega\left(\frac{n_0}{\epsilon^2}\right)$

Summary

- Local differential privacy is widely-used
- Major challenge:
  - Adds a lot of noise
  - Need lots of data to compensate
- Q: When would you use database DP vs. LDP?

How much privacy is actually being used?

Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12

Jun Tang
University of Southern California
juntang@usc.edu

Aleksandra Korolova
University of Southern California
korolova@usc.edu

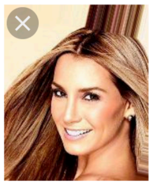
Xiaolong Bai
Tsinghua University
bxl12@mails.tsinghua.edu.cn

Xueqiang Wang
Indiana University
xw48@indiana.edu

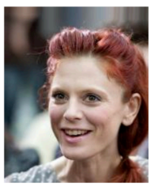
Xiaofeng Wang
Indiana University
xw7@indiana.edu

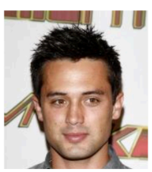
- Reverse-engineered the privacy parameter ϵ
- Found that per datum, guarantees are reasonable
  - $\epsilon = 1$ or 2
- Found parameters as high as 16 **per day**!
- Unbounded in general

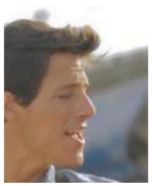
Machine Learning Pipeline – No Privacy

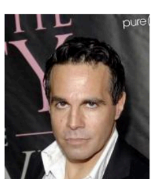

Blond


Blond


Red


Brown

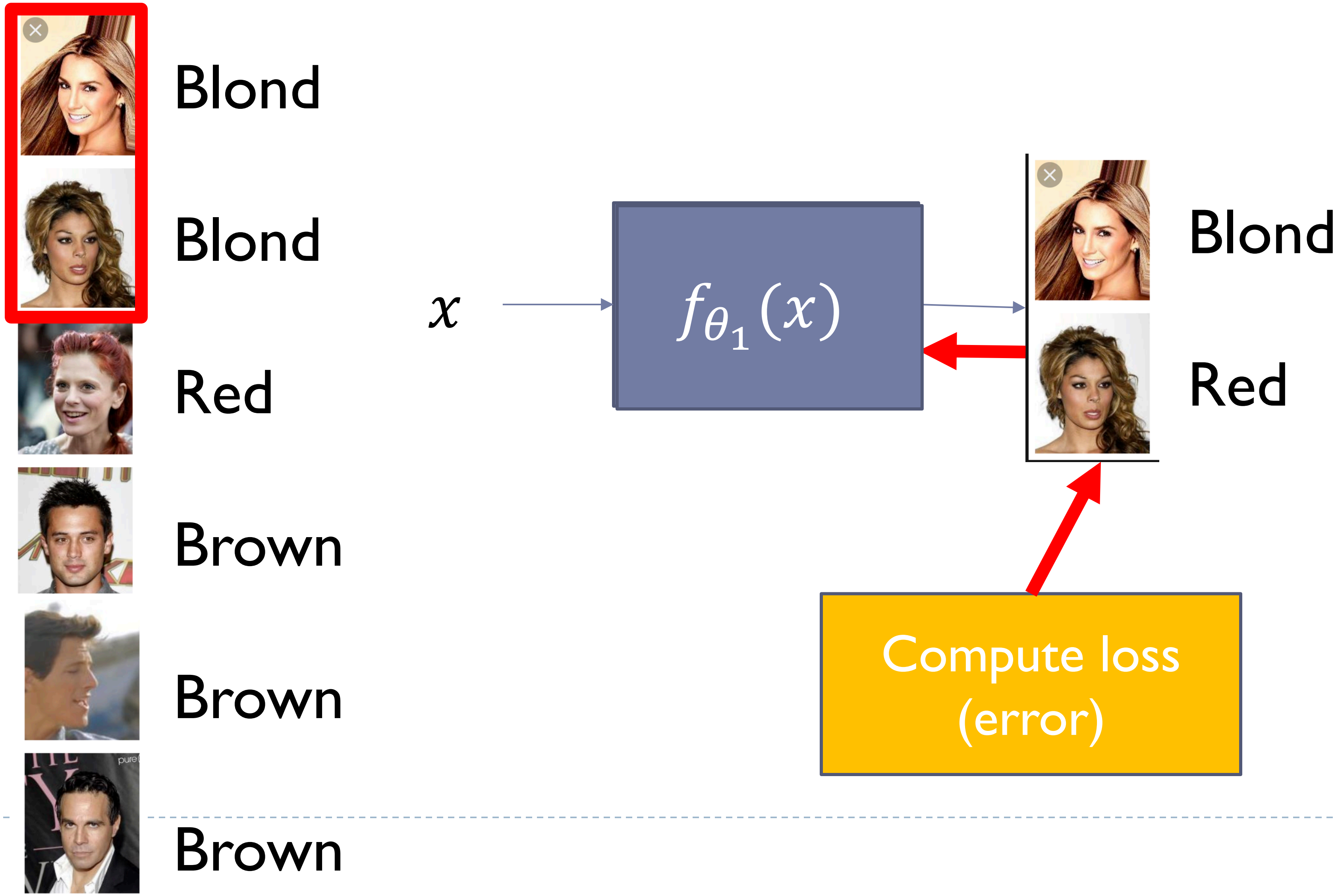

Brown


Brown

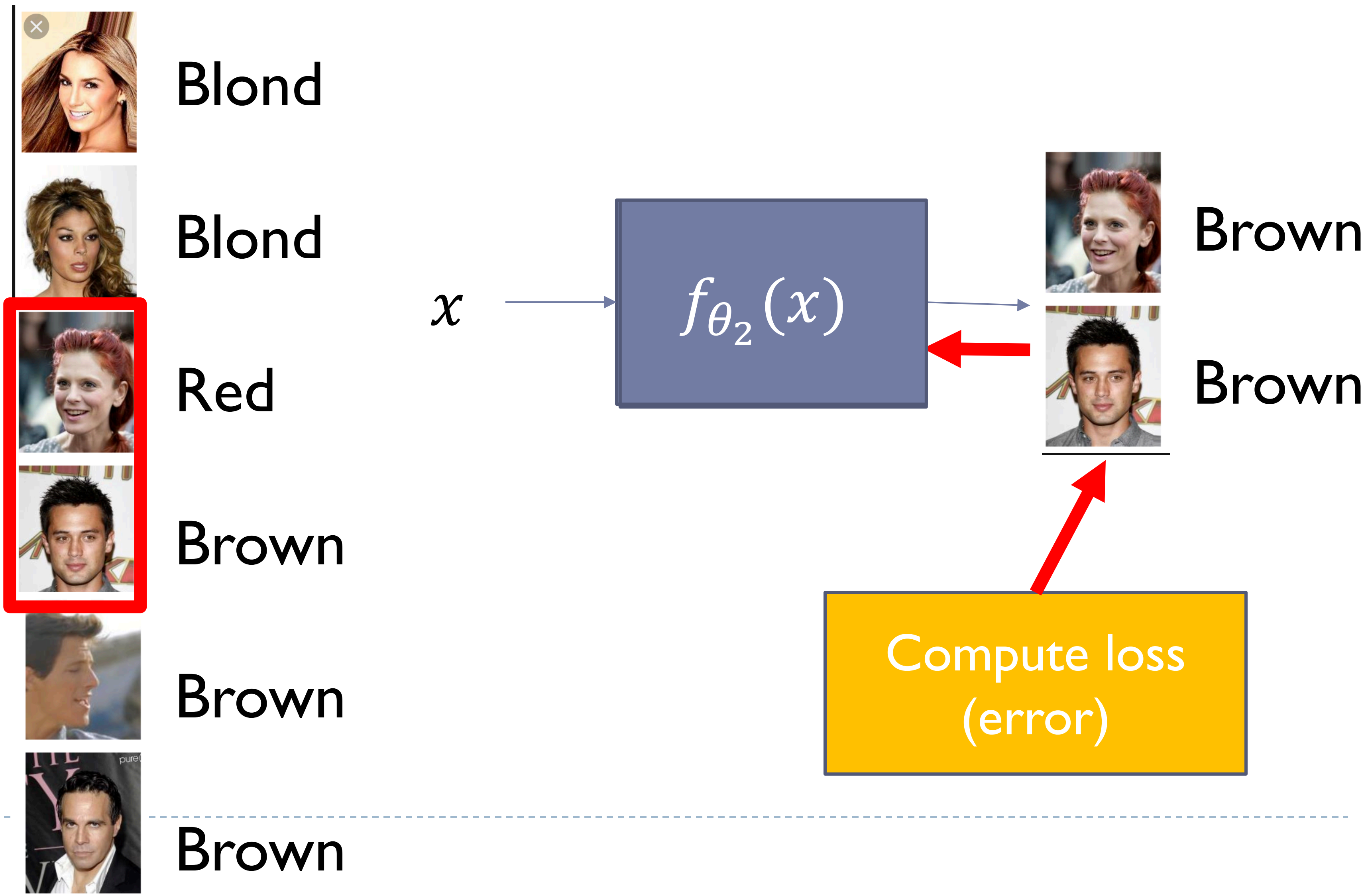
$$x \longrightarrow \boxed{f_\theta(x)} \longrightarrow$$

Blond
Red
Brown

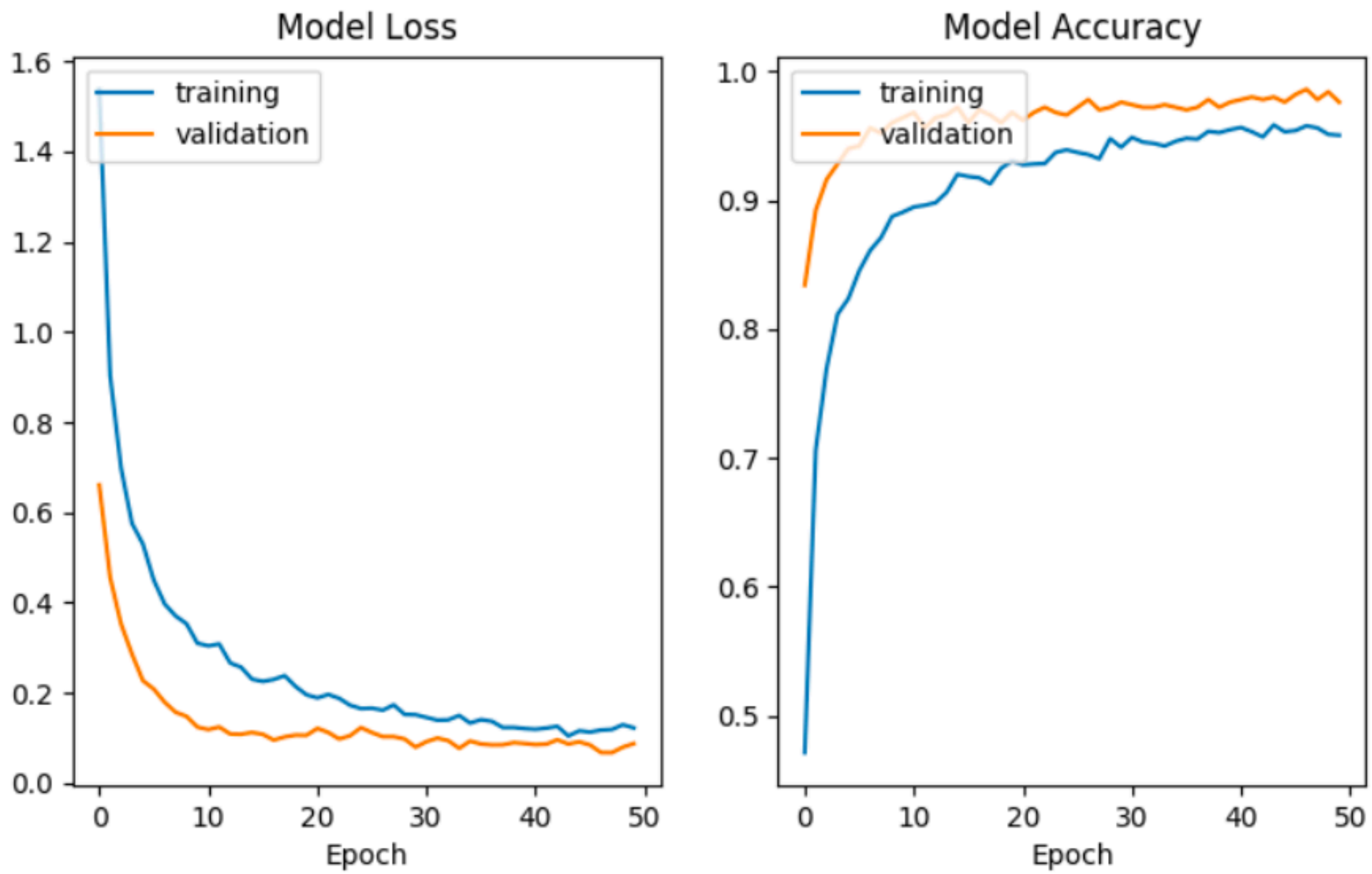
Machine Learning Pipeline – No Privacy



Machine Learning Pipeline – No Privacy

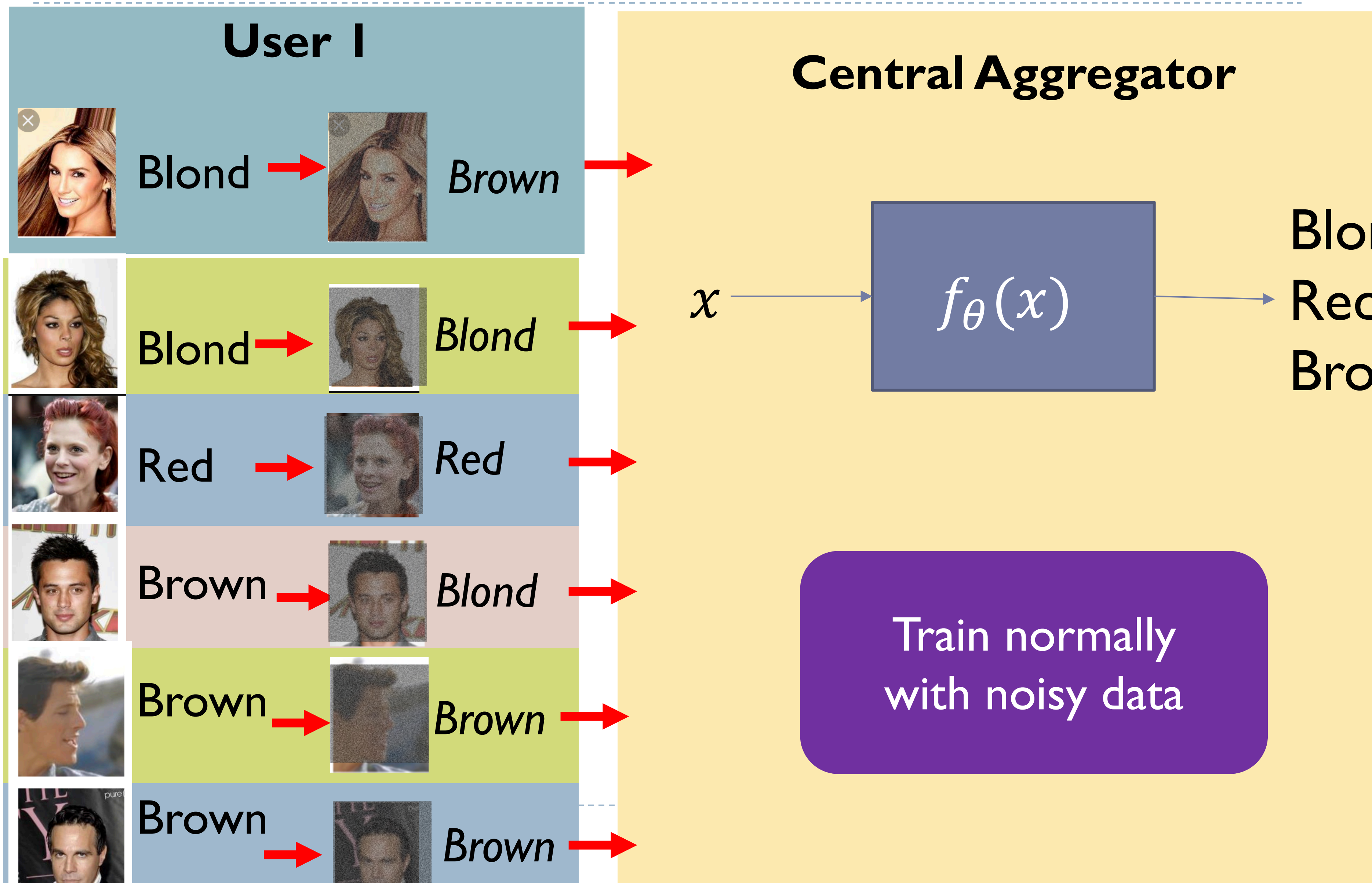


Over time...



1 epoch = 1 full pass through dataset

Let's add Local DP…



Let’s use Global DP

Central Aggregator

Blond

Blond

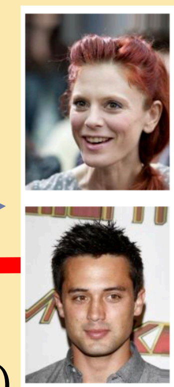
Red

Brown

Brown

Brown

x → $f_{\theta_{i+1}}(x)$ →



Brow

Brow

$\nabla L(X, Y; \theta)$

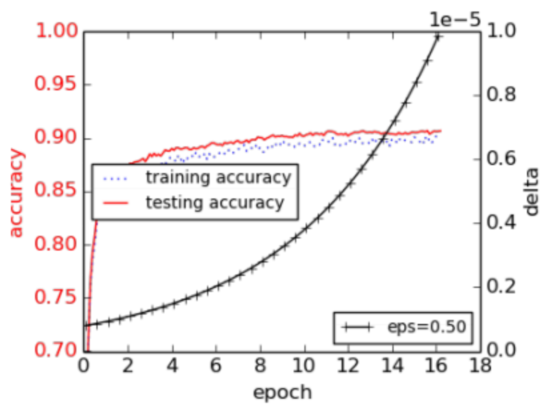
+ Noise

- Depends on sensitivity of gradient function!
- Limit by clipping gradients

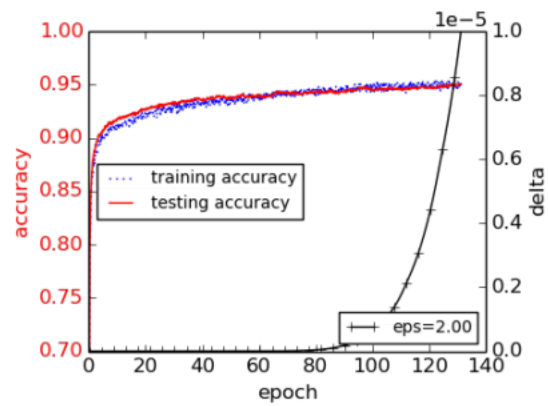
Compute loss
$L(X, Y; \theta)$

Deep Learning with Differential Privacy

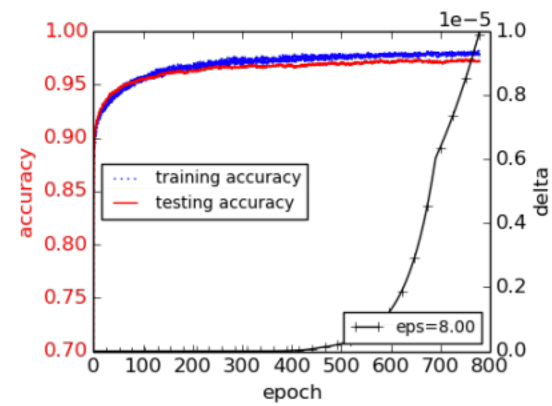
- [Abadi, Chu, Goodfellow, McMahan, Mironov, Talwar, Zhang, CCS 2016]



(1) Large noise



(2) Medium noise

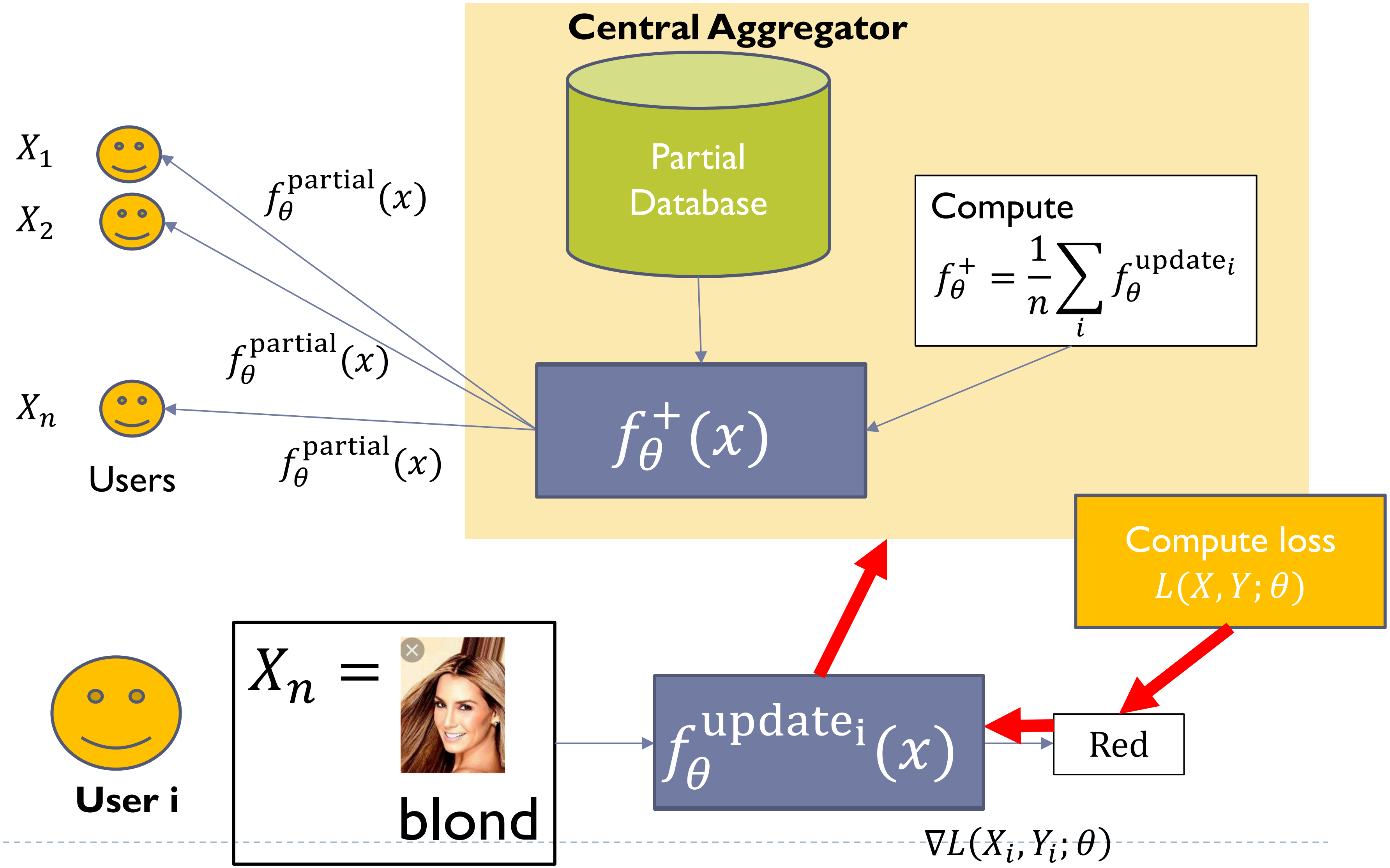


(3) Small noise

Federated Learning

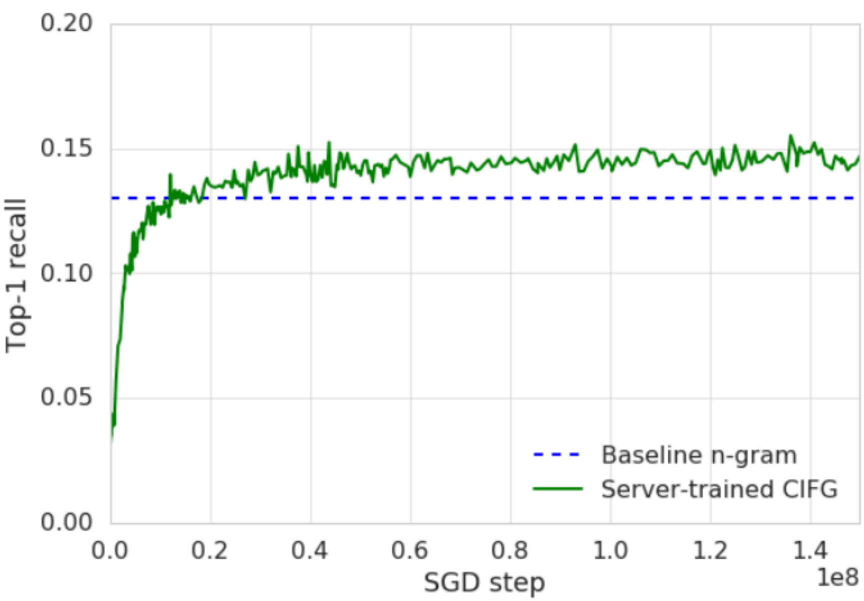
Distributed Learning at Scale

Federated learning: Another Google Project

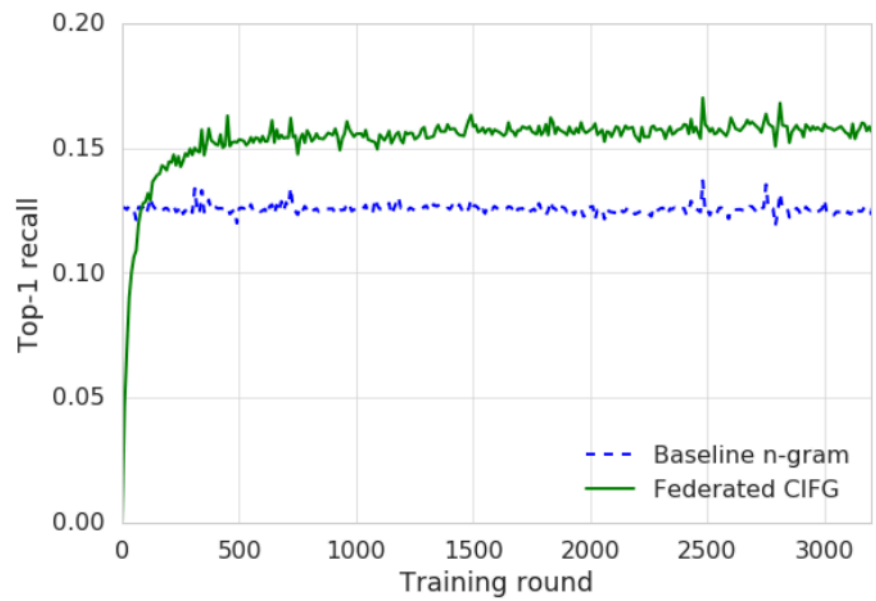


Empirical Results

- Results from "Federated Learning for Mobile Keyboard Prediction", Hard et al., 2019



Centralized Learning

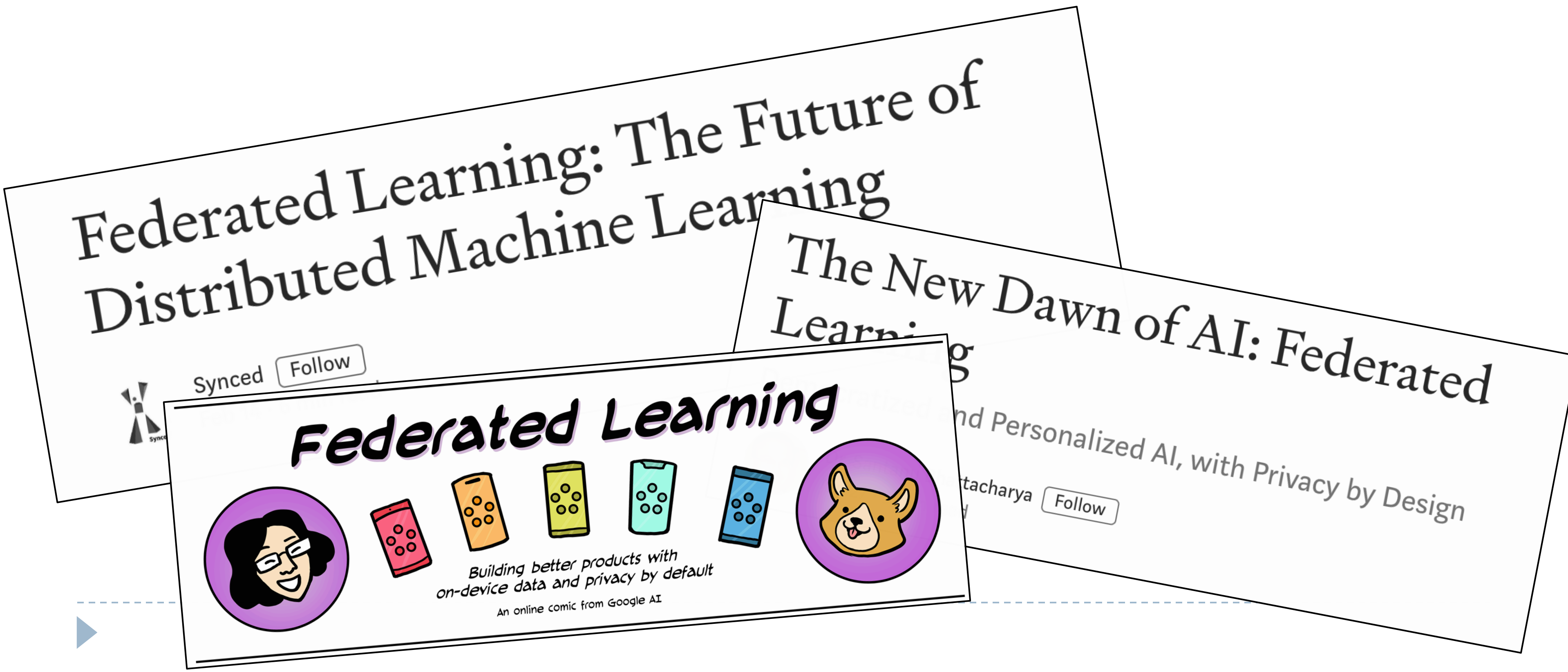


Federated Learning

Federated Learning in practice

- Being used to train GBoard (Google's keyboard)

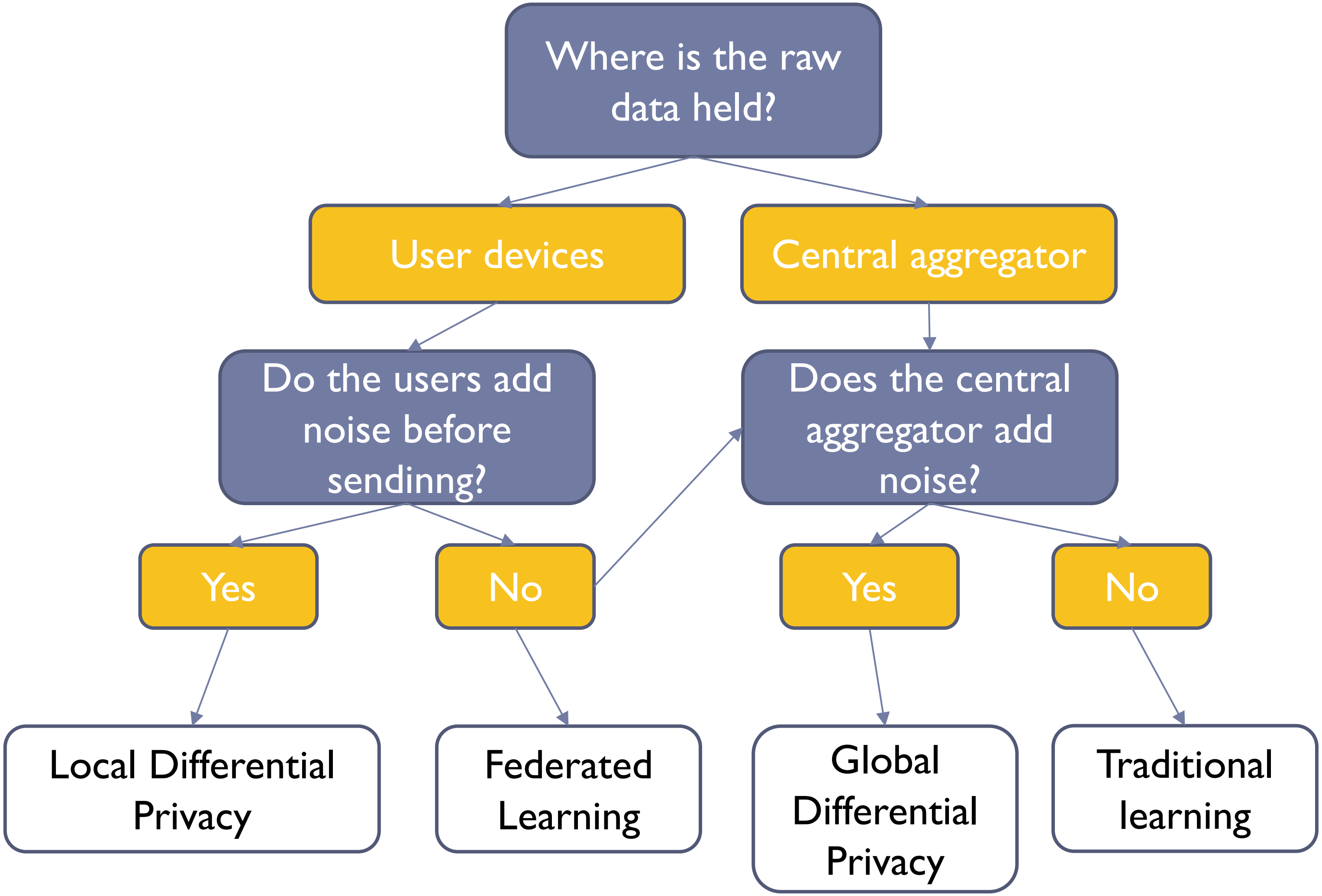
- Very active area of research



What are the privacy implications?

- User’s plaintext data is not revealed
- Unclear what the aggregator may be able to learn from partial gradient updates
- No DP guarantees
  - Could be combined with DP
  - Active area of research

Summary



Where is the raw data held?
User devices
Central aggregator
Do the users add noise before sendinng?
Does the central aggregator add noise?
Yes
No
Yes
No
Local Differential Privacy
Federated Learning
Global Differential Privacy
Traditional learning

Comparison

Method	Pros	Cons
Traditional learning		
Global differential privacy		
Local differential privacy		
Federated learning (without DP)		