Auditing and ML Privacy

Jun 3, 2020

Dr. Wei Wei, Prof. James Landay

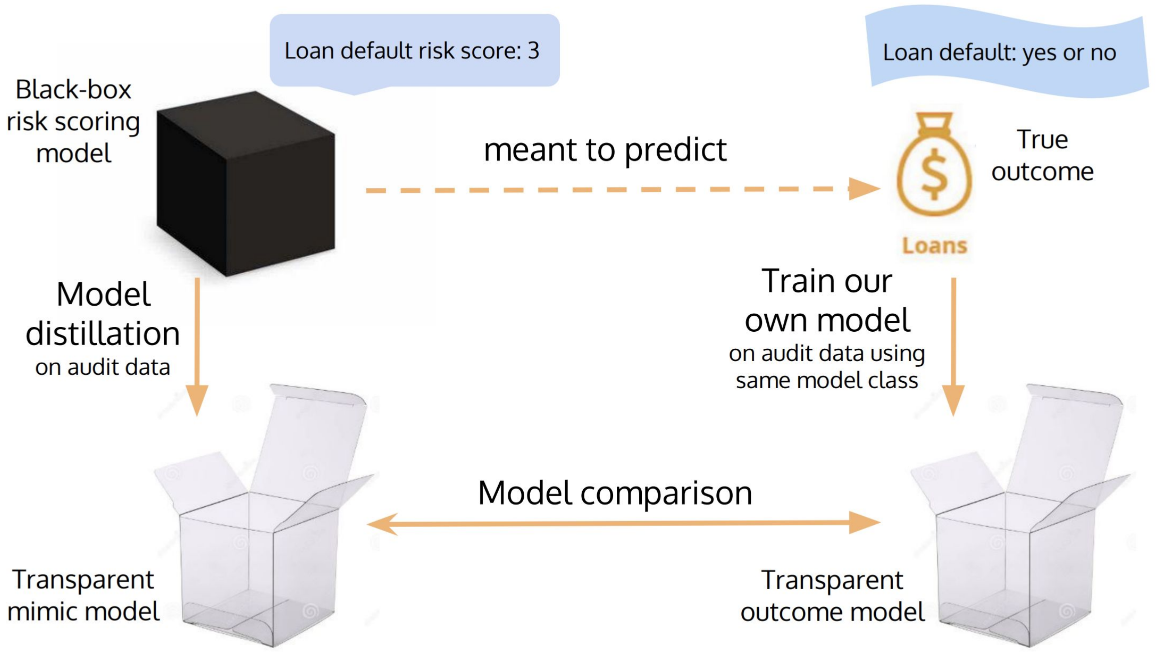
CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

Outline

- ML Auditing
  - Distill-and-Compare
- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning

ML Auditing Using Model Distillation



[Tan et al, 2018]

General Additive Model

$$g(y) = h_0 + \sum_i h_i(x_i) + \sum_{i \neq j} h_{ij}(x_i, x_j)$$

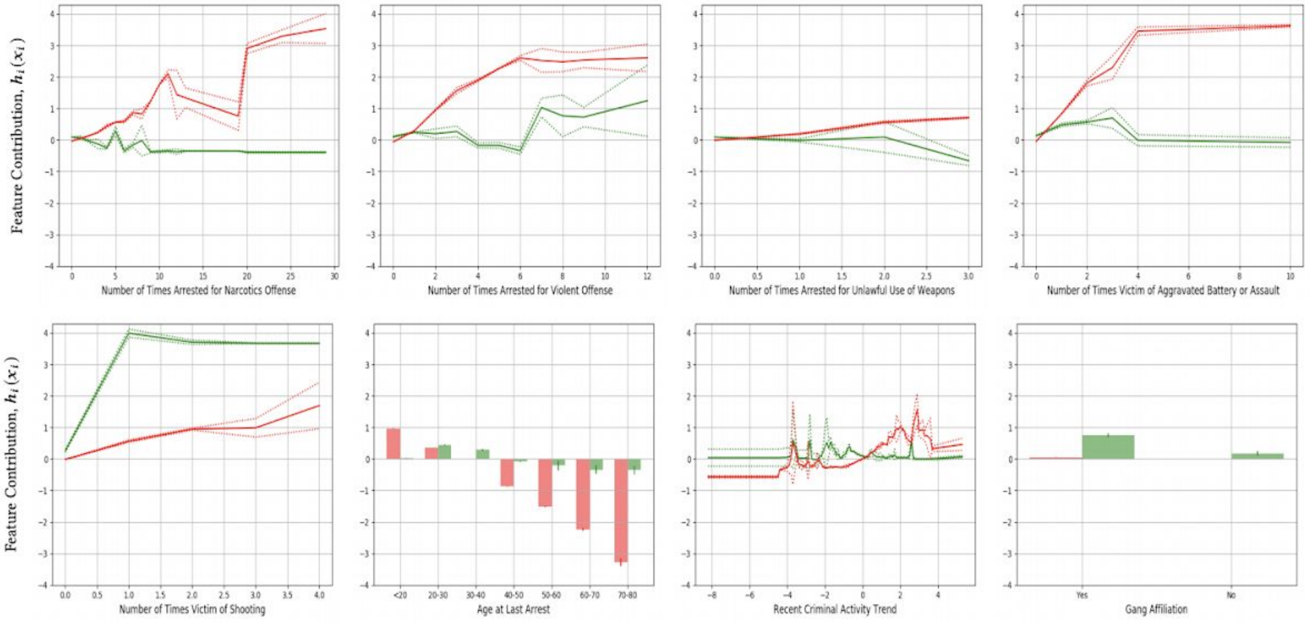
transformation
e.g., logistic for classification

weights

Chicago Police “Strategic Subject”.

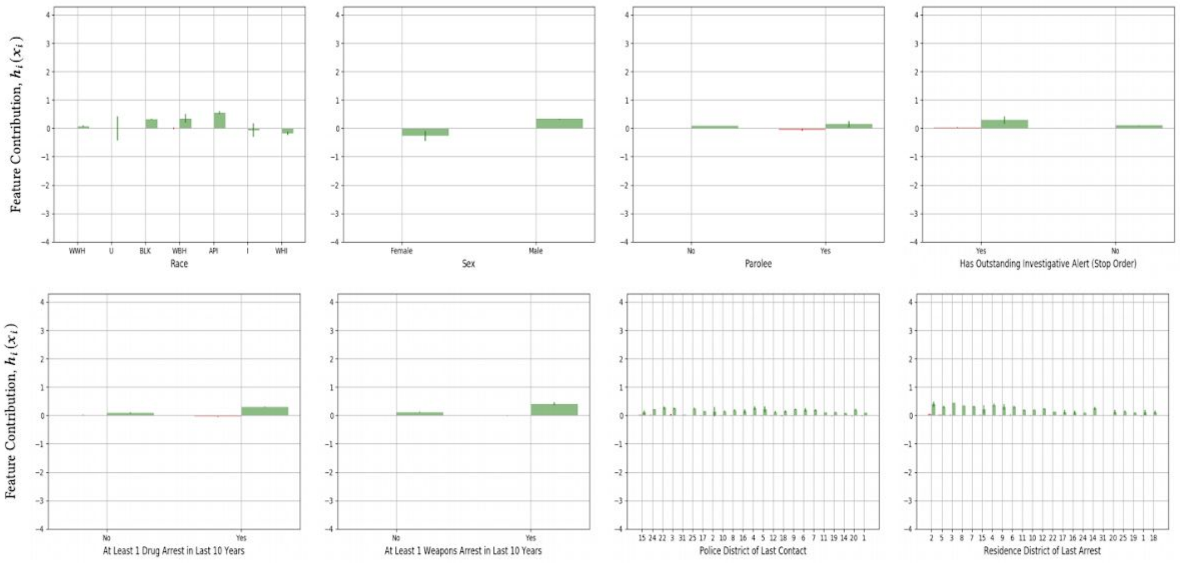
- A risk score for individuals being victims or offenders in a shooting incident
- 16 features
  - 8 reported being used by Chicago Police

Features Reported being Used



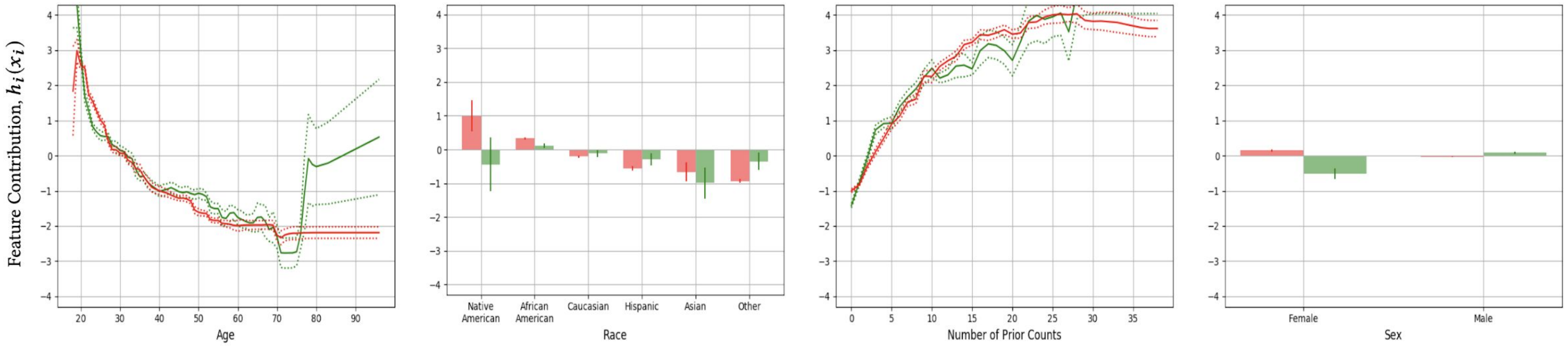
green - model being audited
red - mimic model

Features Reported Not Being Used



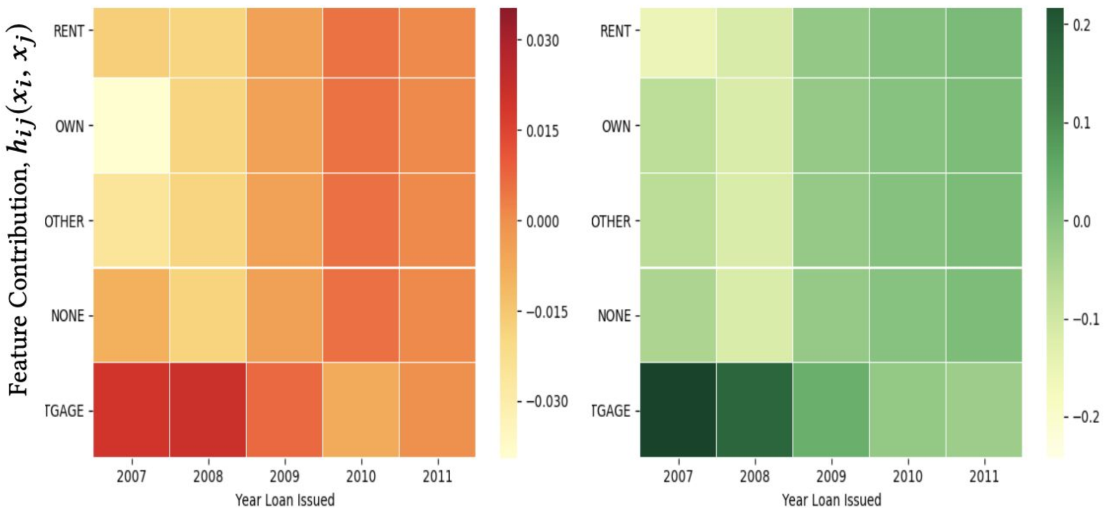
green - model being audited
red - mimic model

Auditing COMPAS



green - model being audited
red - mimic model

Auditing Lending Club



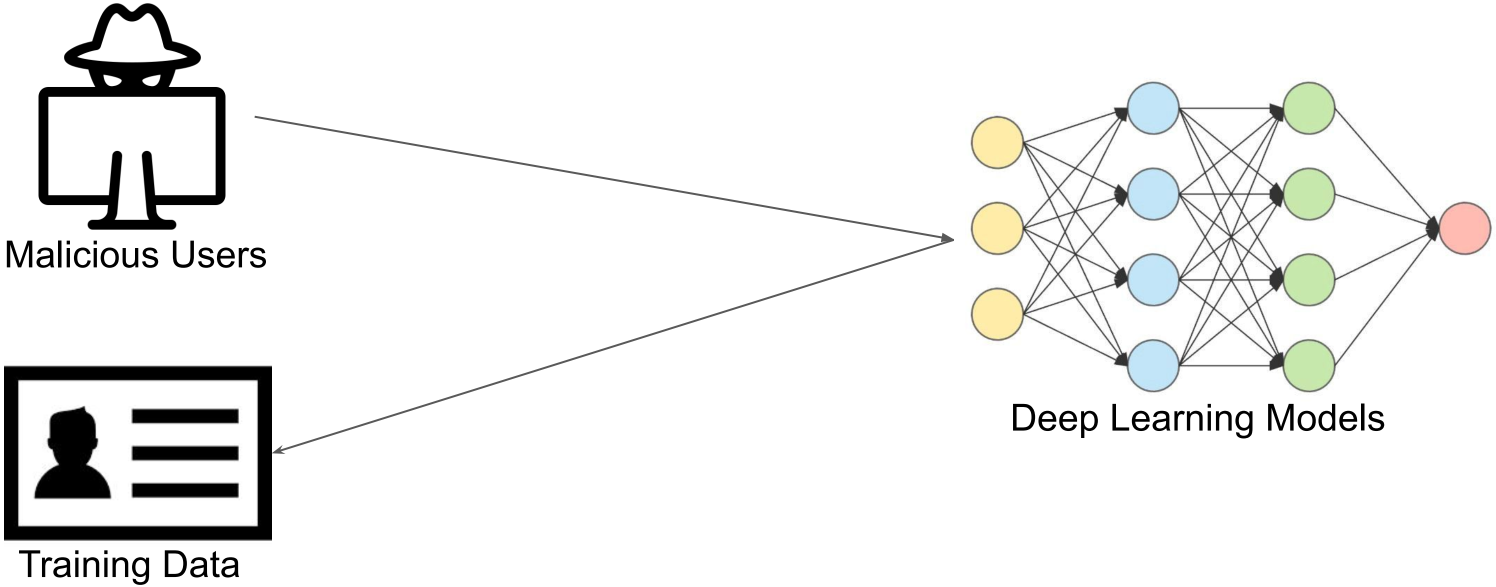
green - model being audited
red - mimic model

$$g(y) = h_0 + \sum_i h_i(x_i) + \underline{\sum_{i \neq j} h_{ij}(x_i, x_j)}$$

Outline

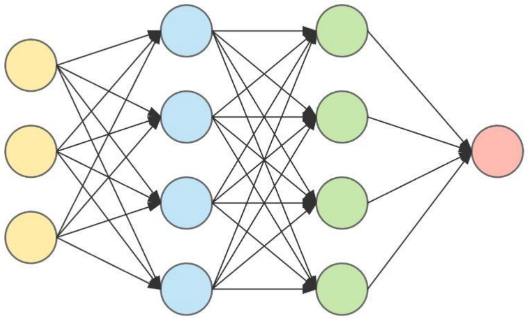
- ML Auditing
  - Distill-and-Compare
- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning

Privacy in ML



Inferring Sensitive Features from ML Models

Demographic Info
Medical History
Genetic Markers



Dose of Warfarin

Fredrikson et al, 2014

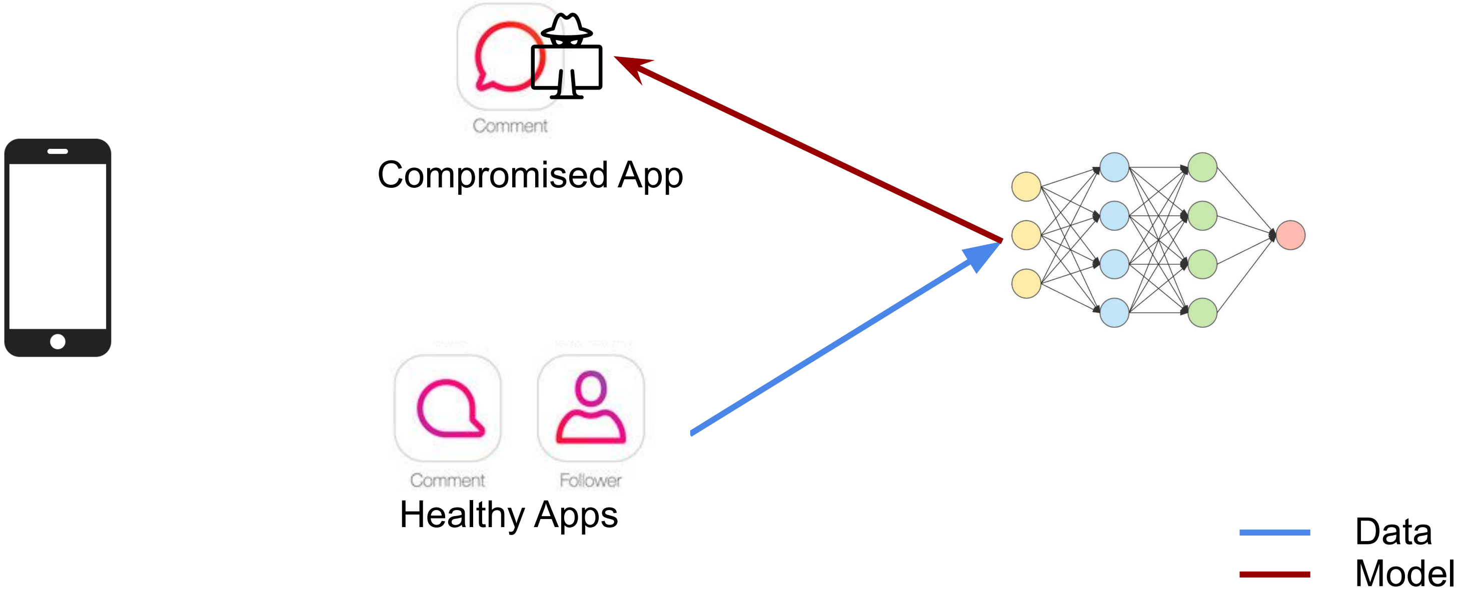
Inferring Training Data from Facial Recognition Models

Original Image

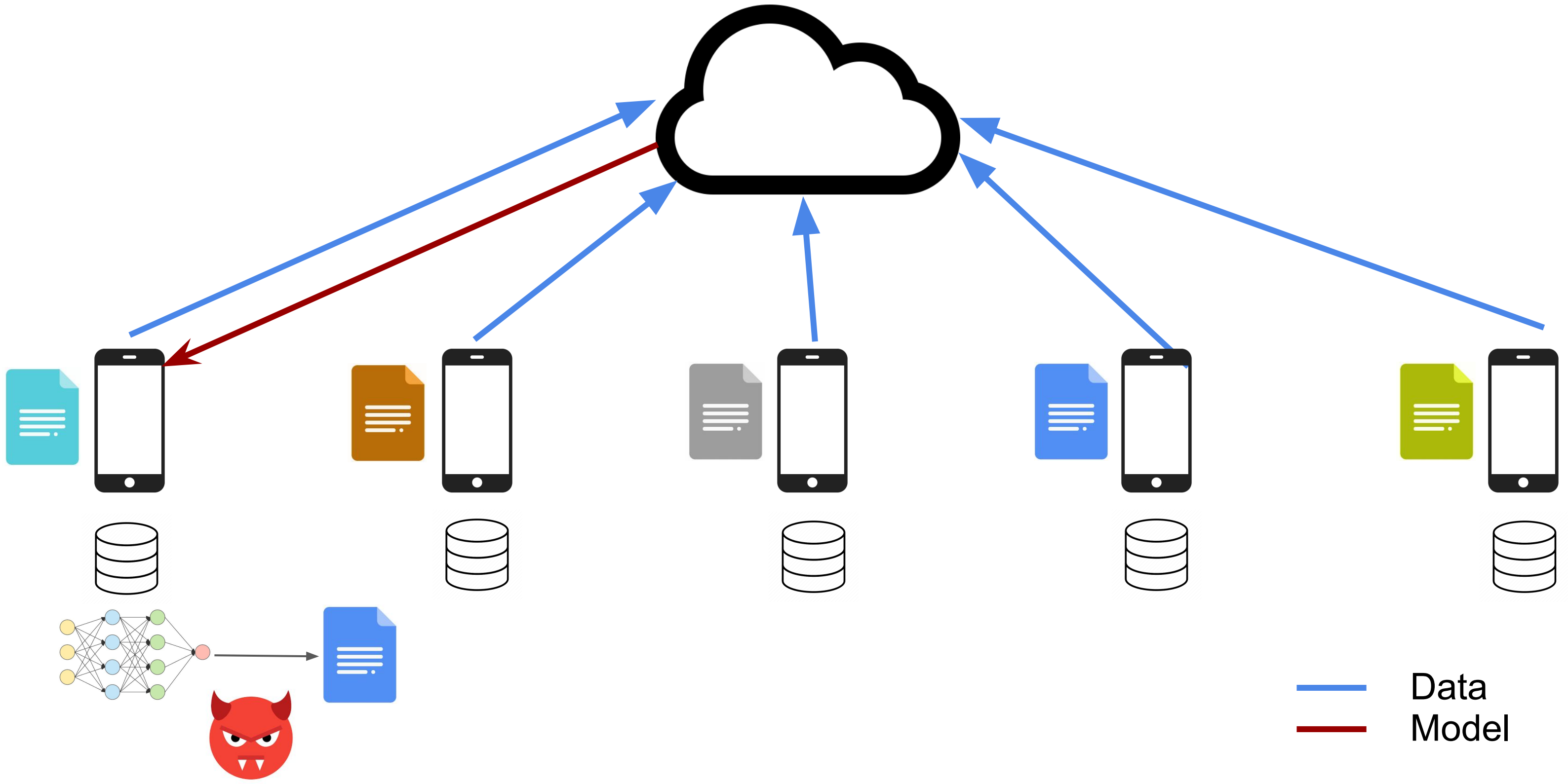
Inferred Image

[Fredrikson et al, 2015]

Centralized Setting



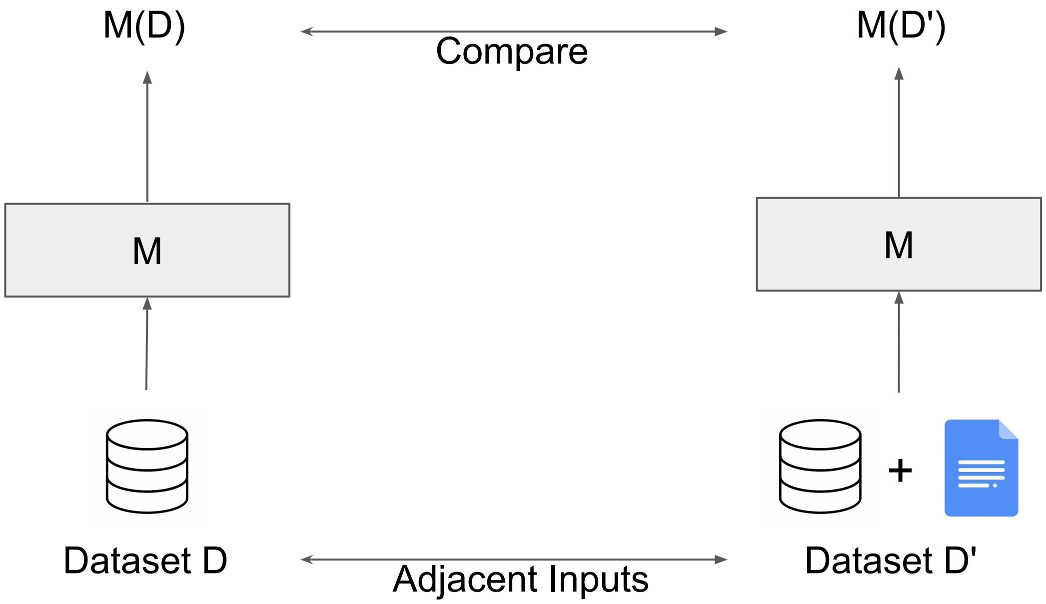
Distributed Setting



Outline

- ML Auditing
  - Distill-and-Compare
- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning

Differential Privacy



Differential Privacy

- A randomized mechanism satisfies (ε, δ) - differential privacy for adjacent inputs d and d' if

Dataset d

Adjacent dataset d'

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

Amount of Information Leakage

A small probability of failure

[Abadi et al, 2016]

Differential Privacy with Deep Learning

Differential Privacy

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

Solution to Differentially Private Deep Learning

$$\mathcal{M}(d) \triangleq \underbrace{f(d)}_{\text{Gradients of Deep Neural Networks}} + \mathcal{N}(0, S_f^2 \cdot \sigma^2)$$

$$S_f = |f(d) - f(d')| \qquad \delta \geq \tfrac{4}{5} \exp(-(\sigma\varepsilon)^2/2) \qquad \varepsilon < 1$$

Abadi et al, 2016

Differentially Private SGD

Gradient Norm Bounds C

Step 1 Calculate Gradients $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Step 2 Gradient Clipping $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Step 3 Adding Noise $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\underline{\sum_i \bar{\mathbf{g}}_t(x_i)} + \mathcal{N}(0, \underline{\sigma^2}\,\underline{C^2}\mathbf{I})\right)$

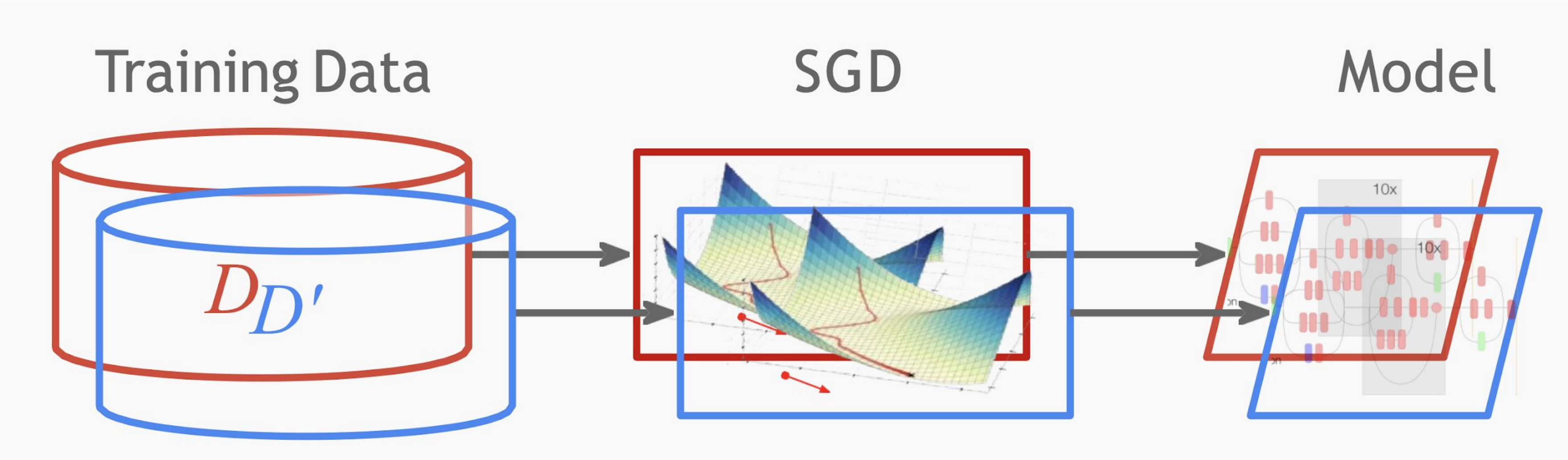
Step 4 Parameter Updating $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

One noise added to each **lot**
(group of data)

$$\mathcal{M}(d) \triangleq f(d) + \mathcal{N}(0, \underline{S_f^2} \cdot \underline{\sigma^2})$$

Abadi et al, 2016

Differentially Private SGD



Abadi et al, 2016

Composition Theorem

- If f is $(\varepsilon_1, \delta_1)$ - DP (Differential Private) and g is $(\varepsilon_2, \delta_2)$ - DP, then

$$f(D), g(D) \text{ is } (\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2) \text{ - DP}$$

Abadi et al, 2016

Budget Analysis for Differentially Private SGD

- Bounds the amount of privacy leakage (budget)
- Each lot (group of data) with L samples is (ε, δ) - DP
- Using Composition theorem, our SGD is is (q • ε, q • δ) - DP
  - q = L/N - samping ratio per lot

Abadi et al, 2016

Moments Accountant

- Provides a tighter bounds for privacy leakage by considering the Gaussian distributed noise
- Under Moments Accountant, there exist c_1 and c_2 such that Differentially Private SGD is

$$(O(q\varepsilon\underline{\sqrt{T}}), \delta)$$ - Differentially Private

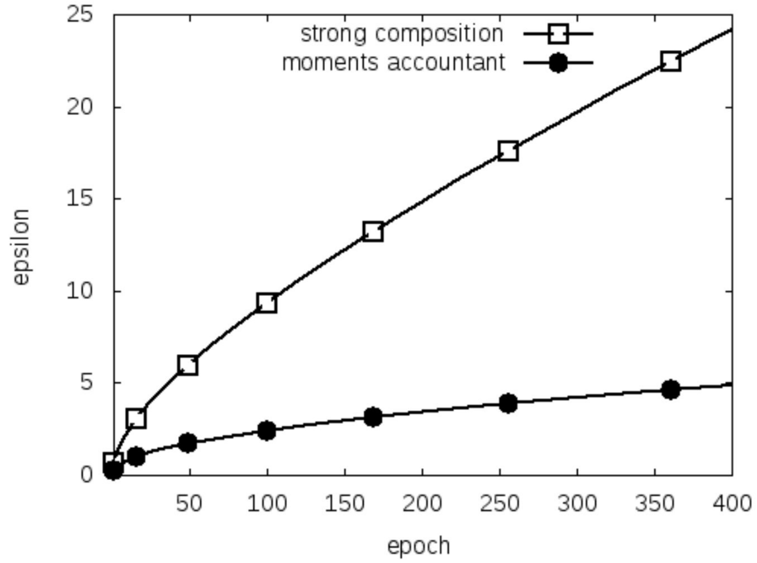
$$\sigma \geq c_2 \frac{q\sqrt{T\log(1/\delta)}}{\varepsilon} \qquad \varepsilon < c_1 q^2 T$$

- q = L/N - sampling ratio per lot
- T - number of time steps

Abadi et al, 2016

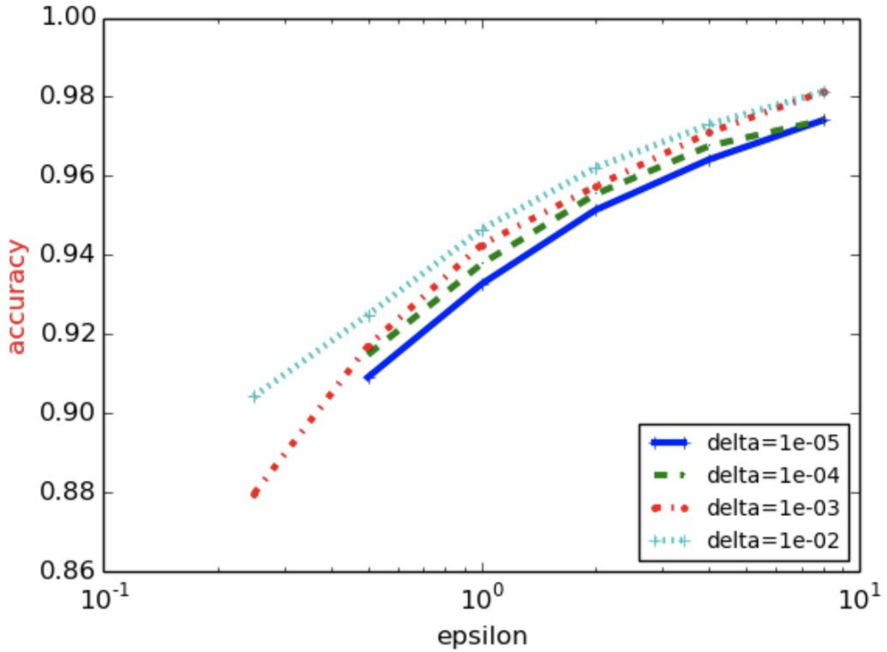
ε As A Function of Epoch E

- E - number of epochs
- $q = 0.01$
- $\sigma = 4$
- $\delta = 10^{-5}$



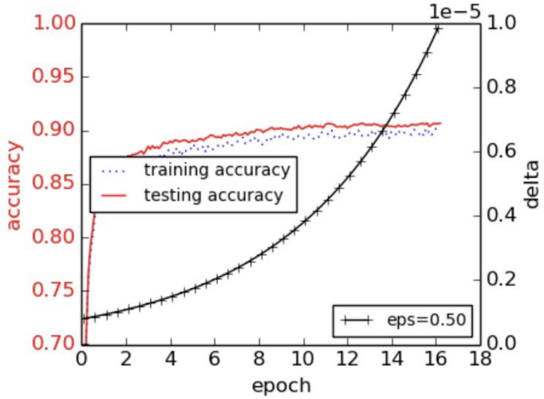
Abadi et al, 2016

Performance and (ε, δ)



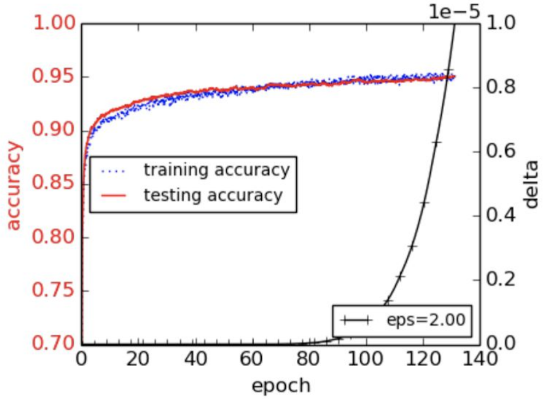
Abadi et al, 2016

Performance and Noise Levels



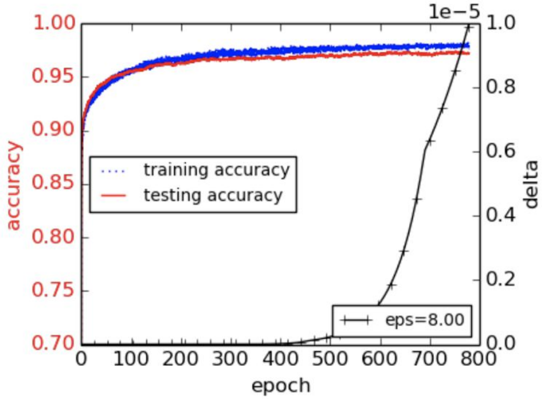
(1) Large noise

σ=8



(2) Medium noise

σ= 4



(3) Small noise

σ=2

$(O(q\varepsilon\sqrt{T}), \delta)$ - Differentially Private

$$\sigma \geq c_2 \frac{q\sqrt{T \log(1/\delta)}}{\varepsilon} \qquad \varepsilon < c_1 q^2 T$$

Abadi et al, 2016

Outline

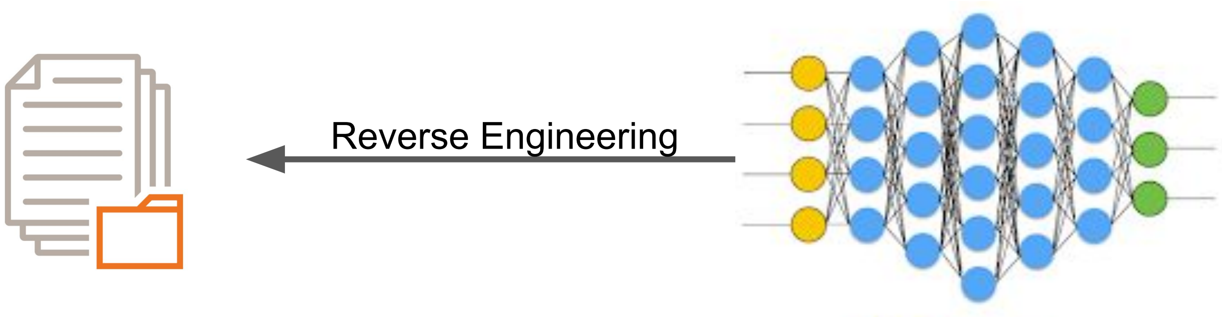
- ML Auditing
  - Distill-and-Compare
- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning

Recap: Types of Adversarial Attack

	Attack Phase	Goal
Evasion	Testing	Compromise Model Performance
Data Poisoning	Training	Compromise Model Performance
Exploratory	Testing	Explore Model Characteristics Reconstruct User Data

Recap

- Exploratory Attack
  - Reverse engineer user data from a trained model



Model Inversion Attacks

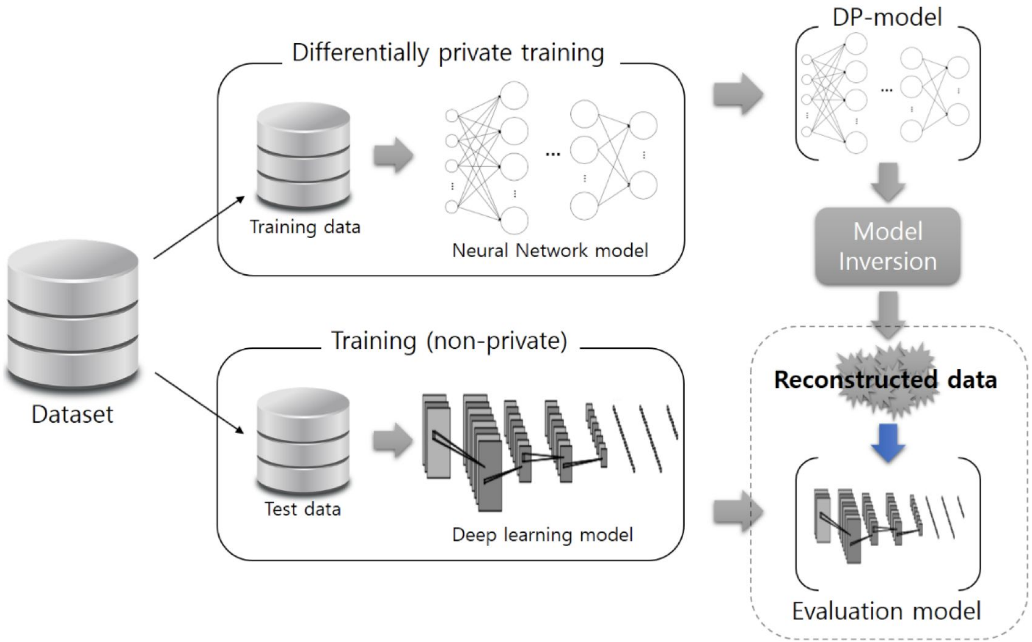
Original Image

Reconstructed Image

$$x = \arg\max_{x} f_y(x)$$

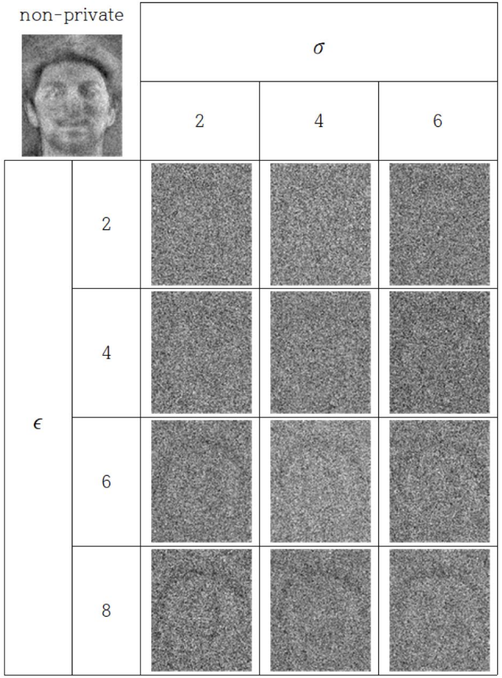
Fredrikson et al, 2015

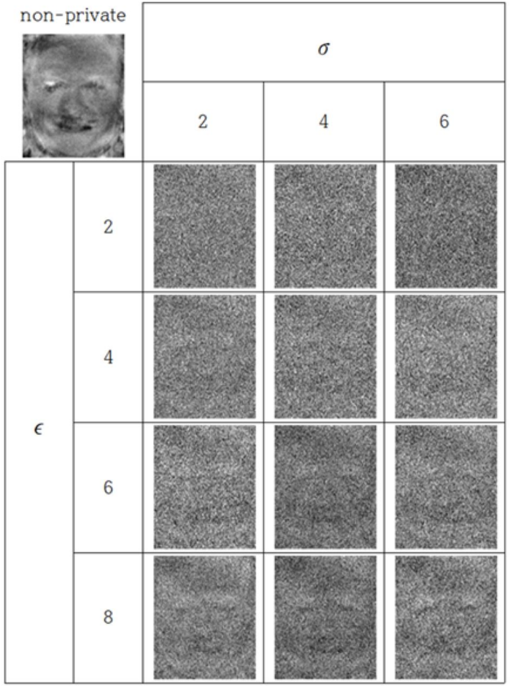
Model Inversion Attack to Evaluate Differential Privacy



[Park et al, 2019]

Results





Park et al, 2019

Outline

- ML Auditing
  - Distill-and-Compare
- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning

Differential Privacy and Local Differential Privacy

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

Differential Privacy

- d, d' are sets of data
- d and d' differ in one sample

- Centralized setting

Local Differential Privacy

- d and d' are single samples

- Distributed setting

Deployment of Local Differential Privacy

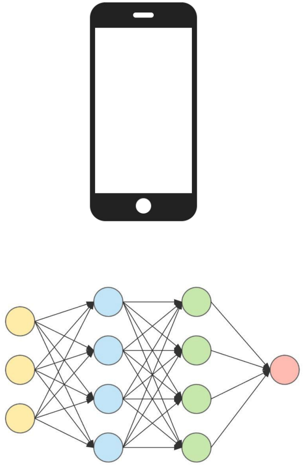
- RAPPOR by Google
  - Collect user data
  - Randomized Aggregatable Privacy-Preserving Ordinal Response
- Private Count Mean Sketch by Apple
  - Collect emoji usage data along with other information in iPhone
  - Learning with Privacy at Scale

Outline

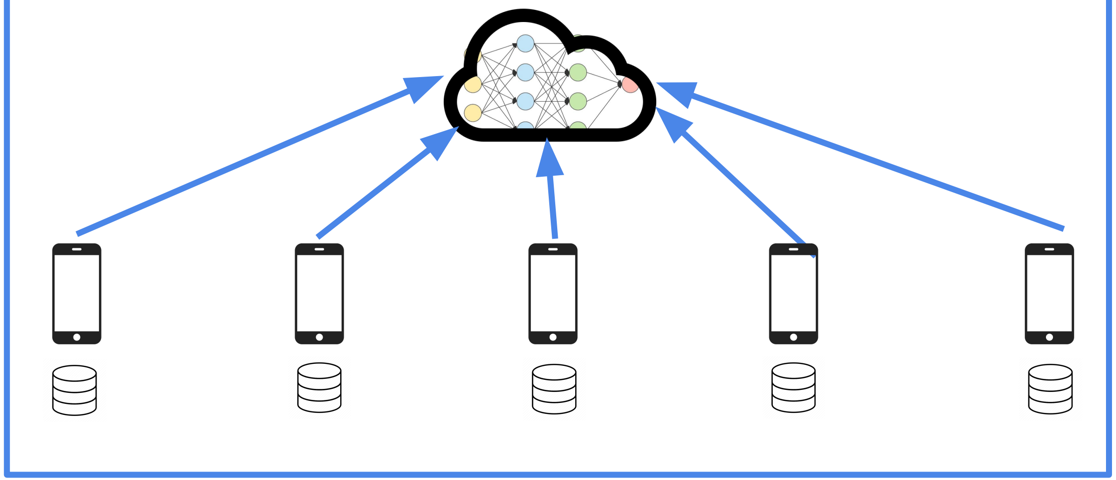
- ML Auditing
  - Distill-and-Compare
- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning

Distributed Optimization

Centralized Setting



Distributed Setting



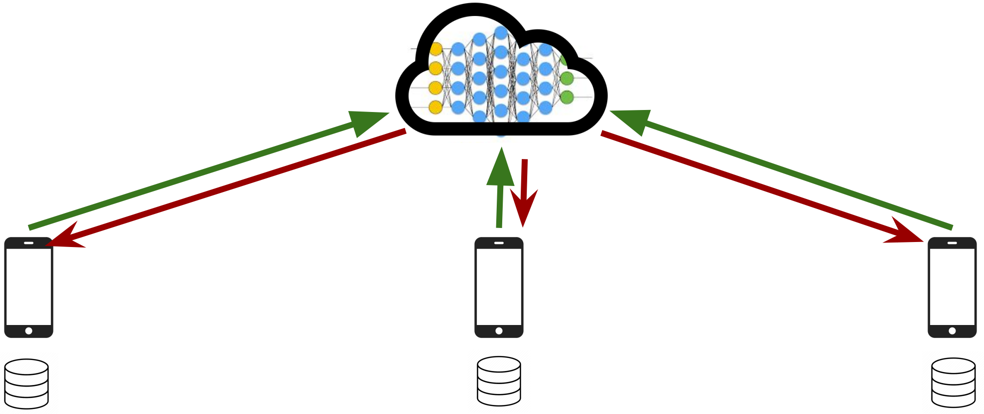
Relies on distributed optimization

Federated Optimization

- Non-IID
  - User data is localized to their own usage
  - Hard to be a representative of the population
- Unbalanced Similarly
  - Some users will make much heavier on particular services than others
- Distributed Computing Capacity
  - Expect a large number of devices to be updated at the same time
- Limited communication
  - Mobile devices are frequently offline or on slow or expensive connections

McMahan et al, 2017

FedSGD



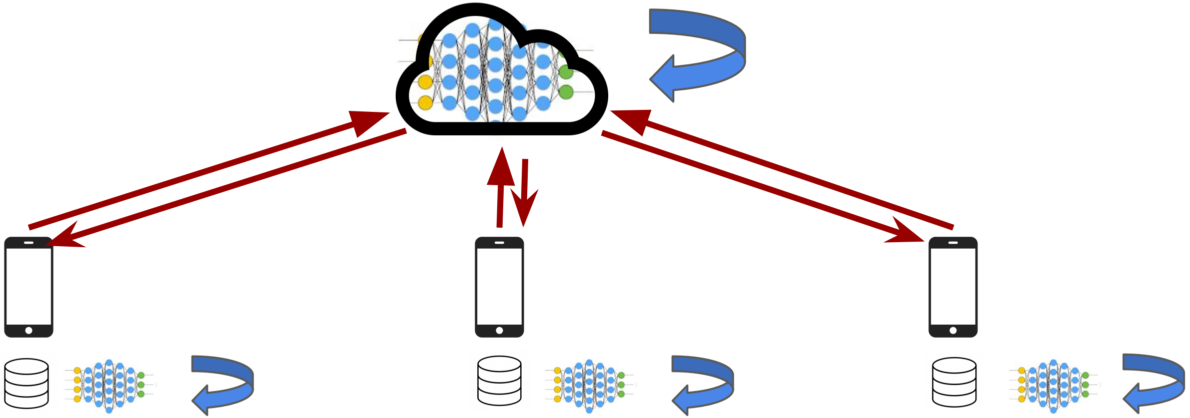
Gradient
Model

$$g_k = \nabla F_k(w_t)$$

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k$$

McMahan et al, 2017

FedAvg



Gradient
Model

$$w_{t+1}^k \leftarrow w_t - \eta g_k$$

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$$

McMahan et al, 2017

Trade-offs Between Local and Global Iterations

- Number of rounds of communication necessary to achieve a test-set accuracy of 97% for the 2NN(MLP) and 99% for the CNN on MNIST

2NN	IID		NON-IID	
C	$B=\infty$	$B=10$	$B=\infty$	$B=10$
0.0	1455	316	4278	3275
0.1	1474 (1.0×)	87 (3.6×)	1796 (2.4×)	664 (4.9×)
0.2	1658 (0.9×)	77 (4.1×)	1528 (2.8×)	619 (5.3×)
0.5	— (—)	75 (4.2×)	— (—)	443 (7.4×)
1.0	— (—)	70 (4.5×)	— (—)	380 (8.6×)
CNN, $E=5$				
0.0	387	50	1181	956
0.1	339 (1.1×)	18 (2.8×)	1100 (1.1×)	206 (4.6×)
0.2	337 (1.1×)	18 (2.8×)	978 (1.2×)	200 (4.8×)
0.5	164 (2.4×)	18 (2.8×)	1067 (1.1×)	261 (3.7×)
1.0	246 (1.6×)	16 (3.1×)	— (—)	97 (9.9×)

C - ratio of clients updated to the server

B - batch size of clients

E - number of epochs client makes over its local dataset on each round

McMahan et al, 2017

Comparisons Between FedSGD and FedAvg

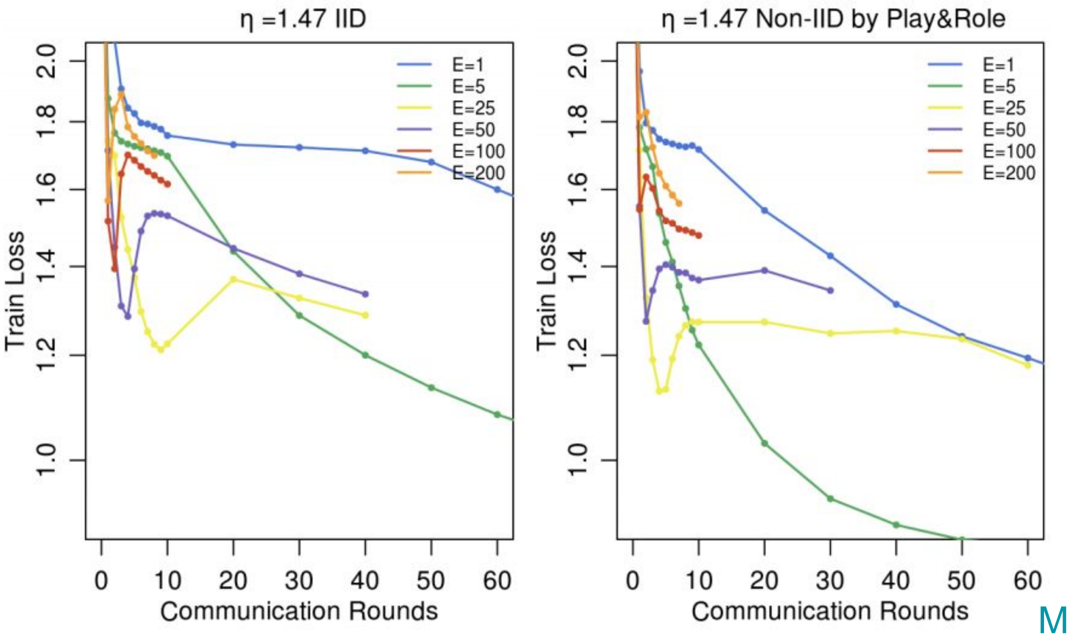
MNIST CNN, 99% ACCURACY					
CNN	E	B	u	IID	NON-IID
FEDSGD	1	∞	1	626	483
FEDAVG	5	∞	5	179 (3.5×)	1000 (0.5×)
FEDAVG	1	50	12	65 (9.6×)	600 (0.8×)
FEDAVG	20	∞	20	234 (2.7×)	672 (0.7×)
FEDAVG	1	10	60	34 (18.4×)	350 (1.4×)
FEDAVG	5	50	60	29 (21.6×)	334 (1.4×)
FEDAVG	20	50	240	32 (19.6×)	426 (1.1×)
FEDAVG	5	10	300	20 (31.3×)	229 (2.1×)
FEDAVG	20	10	1200	18 (34.8×)	173 (2.8×)
SHAKESPEARE LSTM, 54% ACCURACY					
LSTM	E	B	u	IID	NON-IID
FEDSGD	1	∞	1.0	2488	3906
FEDAVG	1	50	1.5	1635 (1.5×)	549 (7.1×)
FEDAVG	5	∞	5.0	613 (4.1×)	597 (6.5×)
FEDAVG	1	10	7.4	460 (5.4×)	164 (23.8×)
FEDAVG	5	50	7.4	401 (6.2×)	152 (25.7×)
FEDAVG	5	10	37.1	192 (13.0×)	41 (95.3×)

K - number of clients
B - batch size

E - number of epochs
u - $(\mathbb{E}[n_k]/B)E$

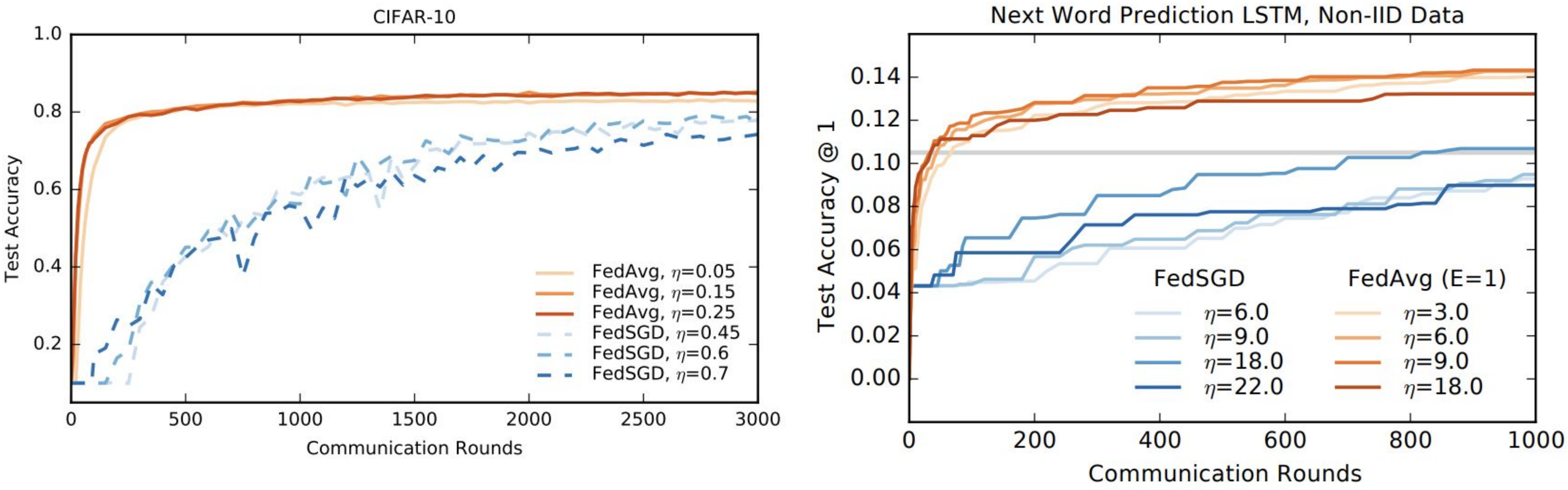
McMahan et al, 2017

Effects of Number of Local Epoches



McMahan et al, 2017

Effects on η



[McMahan et al, 2017]

Reading Assignments (ML Auditing)

- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, arXiv 2016
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, arXiv 2016
- Malgieri, Gianclaudio. The concept of fairness in the GDPR: a linguistic and contextual interpretation, FAccT 2020
- Goodman, Bryce, and Seth Flaxman. European Union regulations on algorithmic decision-making and a “right to explanation”, AI magazine 2017
- Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, arXiv 2018

Reading Assignments (Privacy)

- 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, ACM SIGSAC Conference on Computer and Communications Security, 2017
- Konečný, Jakub, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency, arXiv 2016
- Bonawitz, Keith, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon et al. Towards federated learning at scale: System design, SysML 2019
- Smith, Virginia, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. Federated multi-task learning, NeurIPS 2017
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis, Theory of cryptography conference 2006