Differentially Private Machine Learning

Gautam Kamath

NetFlix Prize

- Recommendation engine competition (2006-2009)
- Training data: (anonymized) user ID, movie, rating, date
- Matched with public IMDb data: real name, movie, rating, date
- Class action lawsuit, cancellation of sequel

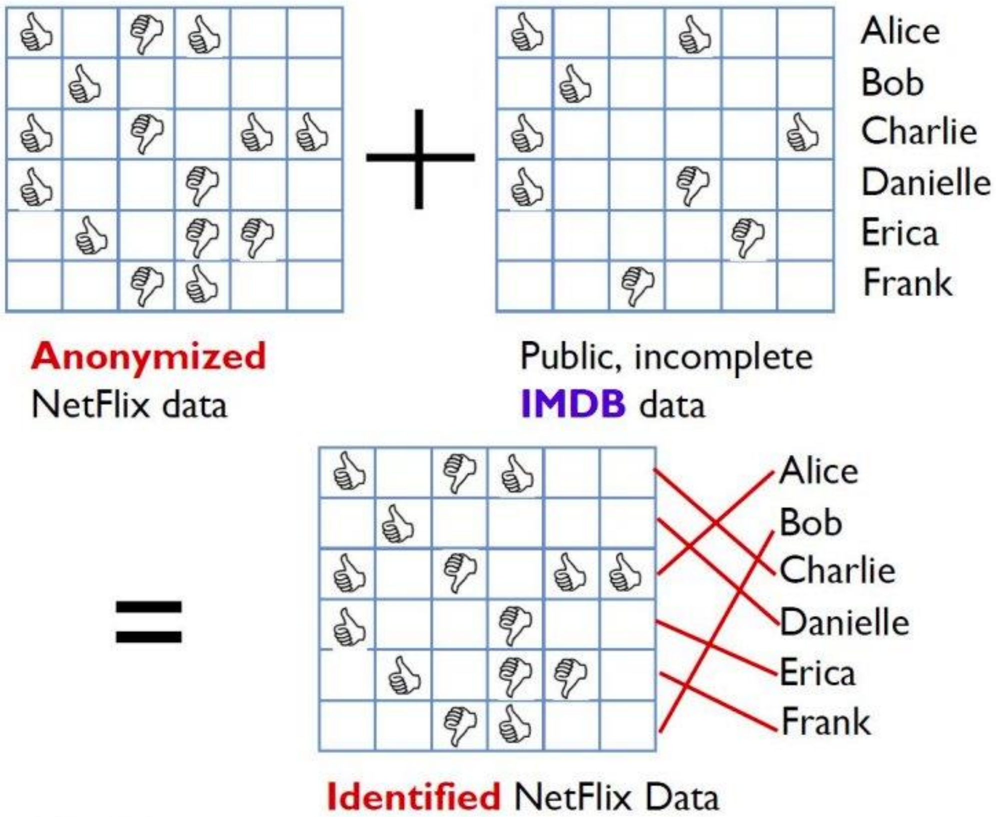
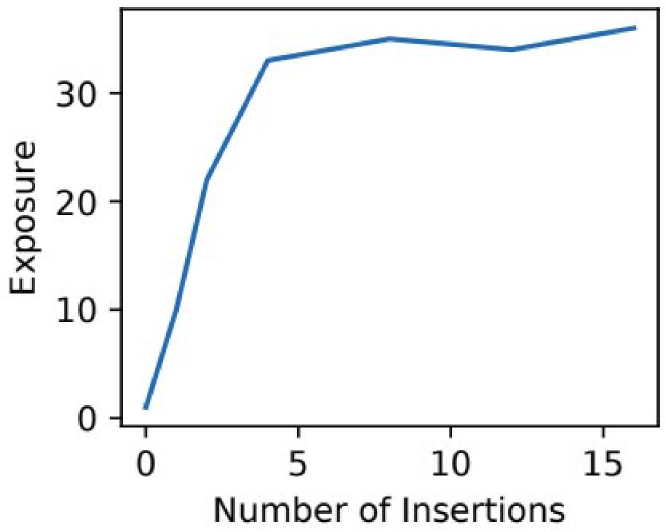


Image credit: Arvind Narayanan

Memorization in Neural Networks

- Language models
- Log-perplexity of a sequence:
  - $P_\theta(x_1, \ldots, x_n) = \sum_i(-\log_2 \Pr(x_i | f_\theta(x_1, \ldots, x_{i-1}))$
- "Mary had a little lamb": low perplexity
- "Correct horse battery staple": high perplexity
- But what if it were in the training data?

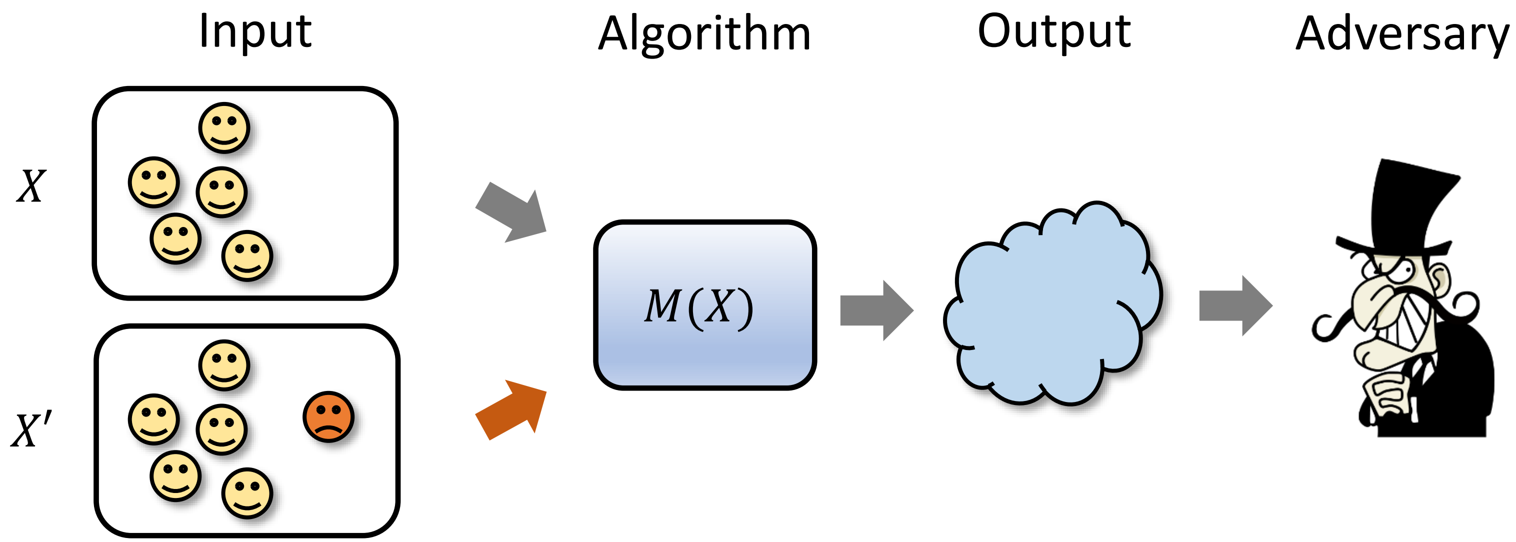
- Canary phrases
  - Is "My SIN is ???-???-???" more likely than it should be?
- Only differential privacy works



[Carlini-Liu-Erlingsson-Kos-Song '19]

See also [Carlini-Tramer-Wallace-Jagielski-HerbertVoss-Lee-Roberts-Brown-Song-Erlingsson-Oprea-Raffel '20]

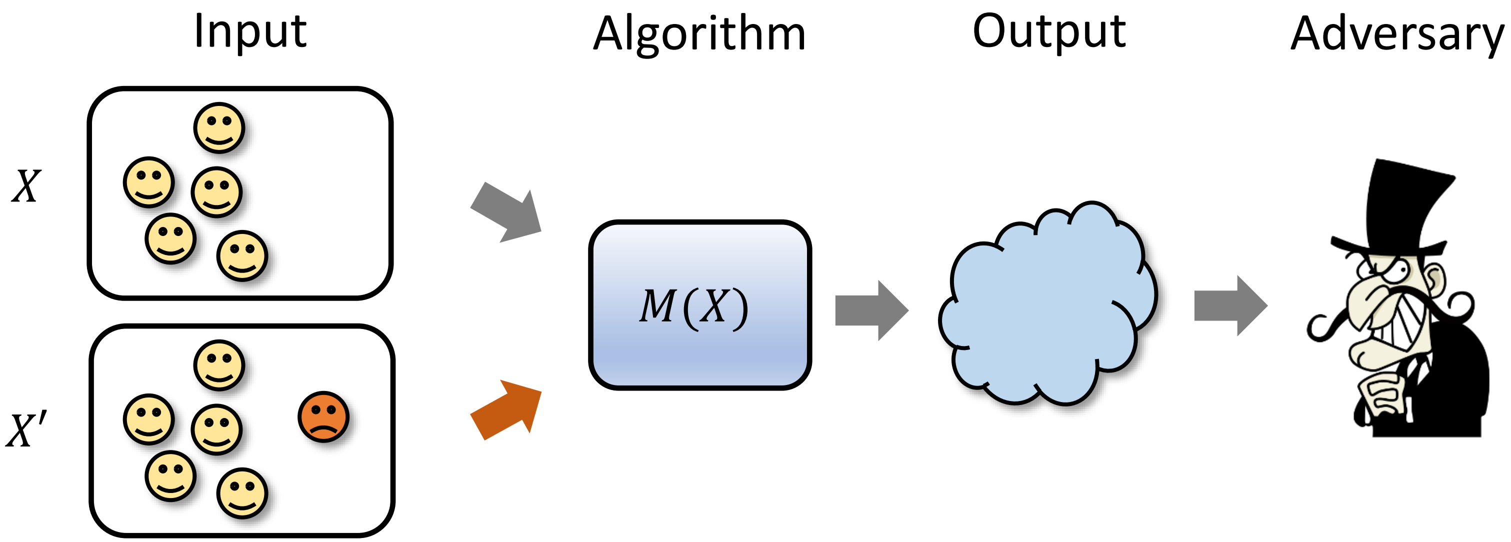
Differential Privacy (DMNS06)



- $M: D^n \rightarrow R$ is (ε, δ)-DP if for all inputs X, X' which differ on one entry:

$$\forall S \subseteq R \qquad \Pr[M(X) \in S] \approx_{\varepsilon,\delta} \Pr[M(X') \in S]$$

Differential Privacy (DMNS06)



- $M: D^n \rightarrow R$ is (ε, δ)-DP if for all inputs X, X' which differ on one entry:

$$\forall S \subseteq R \qquad \Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta$$

Differential Privacy (DMNS06)

$M: D^n \to R$ is (ε, δ)-DP if for all inputs X, X' which differ on one entry:

$$\forall S \subseteq R \qquad \Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta$$

- Google, Apple, Microsoft, 2020 US Census
- $\varepsilon \approx 1$ and $\delta < 1/n$
- Worst-case guarantee
- $e^{\varepsilon_1} e^{\varepsilon_2} = e^{\varepsilon_1 + \varepsilon_2}$
- Symmetric definition
- M must be randomized

What DP does and does not mean

- Outcome is the same whether or not your data is in the dataset
- Protects against linkage and membership inference attacks
- Does *not* prevent statistics and machine learning
  - “Smoking causes cancer”
- Not suitable when we need to identify a specific individual
- Information-theoretic notion

Properties of Differential Privacy

- Post-processing
  - If $M(X)$ is (ε, δ)-DP, then $f(M(X))$ is (ε, δ)-DP
- Group Privacy
  - If M is (ε, δ)-DP, and X and X' differ in k entries,

$$\forall S \subseteq R \qquad \Pr[M(X) \in S] \leq e^{k\varepsilon} \Pr[M(X') \in S] + \delta$$

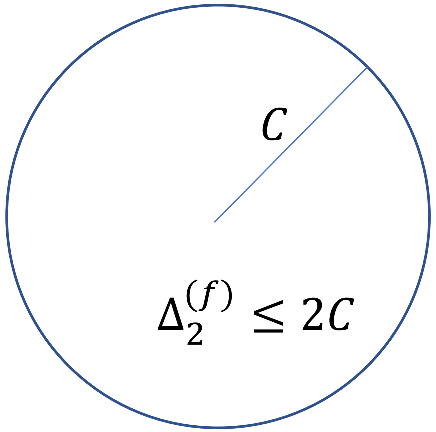
- Composition
  - If $M = (M_1, \dots, M_k)$ is a sequence of k (ε, δ)-DP algorithms
    - M is $(k\varepsilon, k\delta)$-DP (Basic Composition)
    - M is $(O(\sqrt{k}\varepsilon \log(1/\delta')), k\delta + \delta')$-DP (Advanced Composition)

Gaussian Mechanism

- ℓ_2-sensitivity of f

$$\Delta_2^{(f)} = \max_{X \sim X'} \|f(X) - f(X')\|_2$$

  - If $\|f(X)\|_2 \leq C$, then $\Delta_2^{(f)} \leq 2C$



- Gaussian Mechanism

$$M(X) = f(X) + (Y_1, \ldots, Y_k)$$

Where $f(X) \in R^k$, and the Y_i's are $\approx N(0, \Delta^2/\varepsilon^2)$

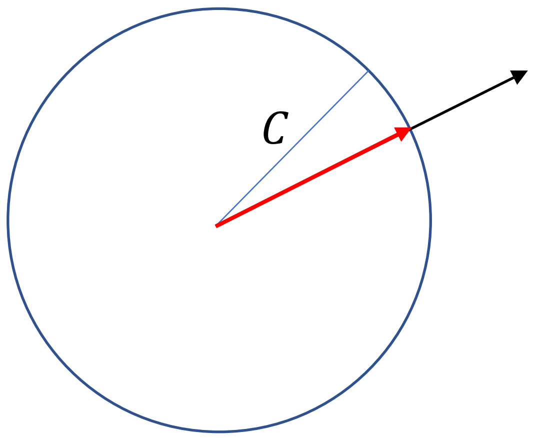
  - (ε, δ)-DP

Stochastic Gradient Descent

1. Choose a random minibatch B of points from the dataset
2. Compute the average gradient $\frac{1}{|B|}\sum_{(x,y)\in B} \nabla\ell(\theta_t, x, y)$
3. Take a step in the negative direction of the gradient
4. Repeat k times

Differentially Private Stochastic Gradient Descent

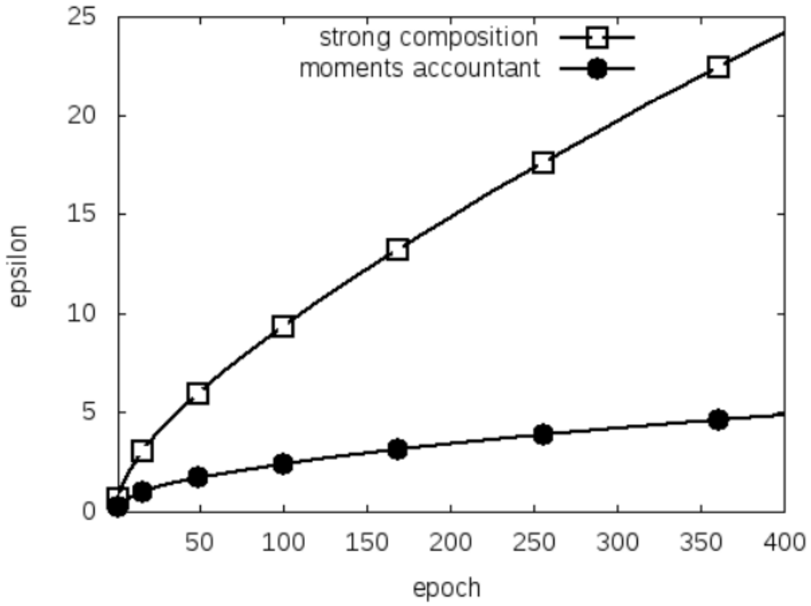
1. Sample a "lot" of points of (expected) size L by selecting each point to be in the lot with probability L/n
2. For each point in the lot, compute the gradient $\nabla \ell(\theta_t, x, y)$ and "clip" it to have ℓ_2 norm at most C
3. Average the clipped gradients and add Gaussian noise
   - Apply the Gaussian Mechanism
4. Take a step in the negative direction of resulting vector
5. Repeat k times



[Song-Chaudhuri-Sarwate '13, Bassily-Smith-Thakurta '14, Abadi-Chu-Goodfellow-McMahan-Mironov-Talwar-Zhang '16]

Privacy of DPSGD (Informal)

- Suppose one step of DPSGD has privacy with parameter ε
- Since we subsample with probability L/n, each step is $\varepsilon L/n$
  - "Privacy amplification by subsampling"
- k steps have privacy with parameter of $\varepsilon\sqrt{k}L/n$
  - Advanced composition
- Better analysis: "Moments accountant"

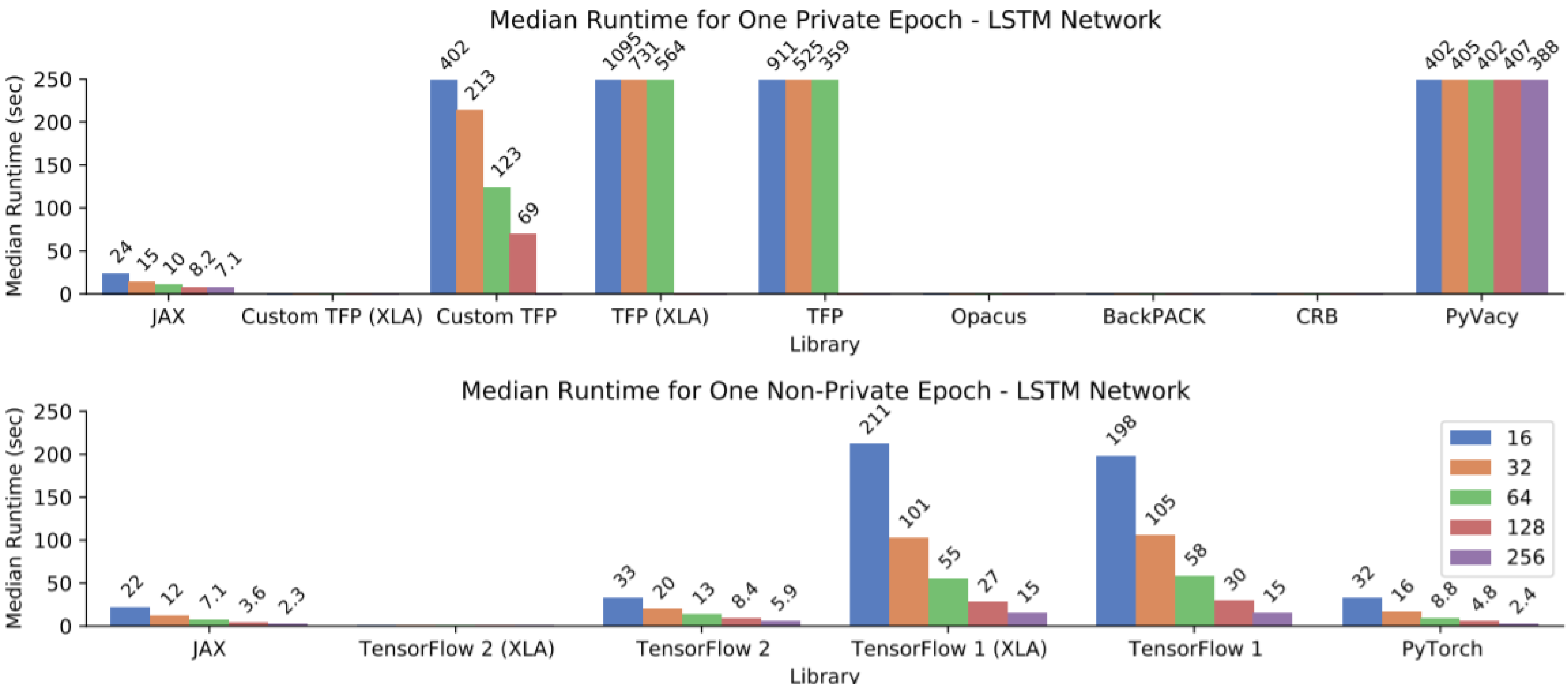


Does it work?

Data	ε-DP	Source	Test Accuracy (%) CNN	ScatterNet+linear	ScatterNet+CNN
MNIST	1.2	Feldman & Zrnic (2020)	96.6	**98.1 ± 0.1**	97.8 ± 0.1
	2.0	Abadi et al. (2016)	95.0	**98.5 ± 0.0**	**98.4 ± 0.1**
	2.32	Bu et al. (2019)	96.6	**98.6 ± 0.0**	98.5 ± 0.0
	2.5	Chen & Lee (2020)	90.0	**98.7 ± 0.0**	98.6 ± 0.0
	2.93	Papernot et al. (2020a)	98.1	**98.7 ± 0.0**	**98.7 ± 0.1**
	3.2	Nasr et al. (2020)	96.1	–	
	6.78	Yu et al. (2019b)	93.2	–	
Fashion-MNIST	2.7	Papernot et al. (2020a)	86.1	**89.5 ± 0.0**	88.7 ± 0.1
	3.0	Chen & Lee (2020)	82.3	**89.7 ± 0.0**	89.0 ± 0.1
CIFAR-10	3.0	Nasr et al. (2020)	55.0	67.0 ± 0.1	**69.3 ± 0.2**
	6.78	Yu et al. (2019b)	44.3	–	–
	7.53	Papernot et al. (2020a)	66.2	–	–
	8.0	Chen & Lee (2020)	53.0	–	–

[Tramèr-Boneh ‘21]

DPSGD can be slow!



[Subramani-Vadivelu-K. ’20]

Architectures for DPSGD

- Tanh >> ReLU? [Papernot-Thakurta-Song-Chien-Erlingsson ‘21]
- Bigger models are not always better

Hyperparameters

- Even more hyperparameters
  - Learning rate, lot size, clipping norm, number of epochs, noise multiplier
- Non-private way: grid search, measure accuracy on validation set
- Pay in privacy budget for each run!
- Options:
  - Private methods for hyperparameter optimization [Liu-Talwar ’19]
  - Transfer hyperparameters from related public data
  - Cheat and ignore privacy budget for multiple runs…

Conclusion

- Private machine learning is here!
- But there’s still a lot of work to do…