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

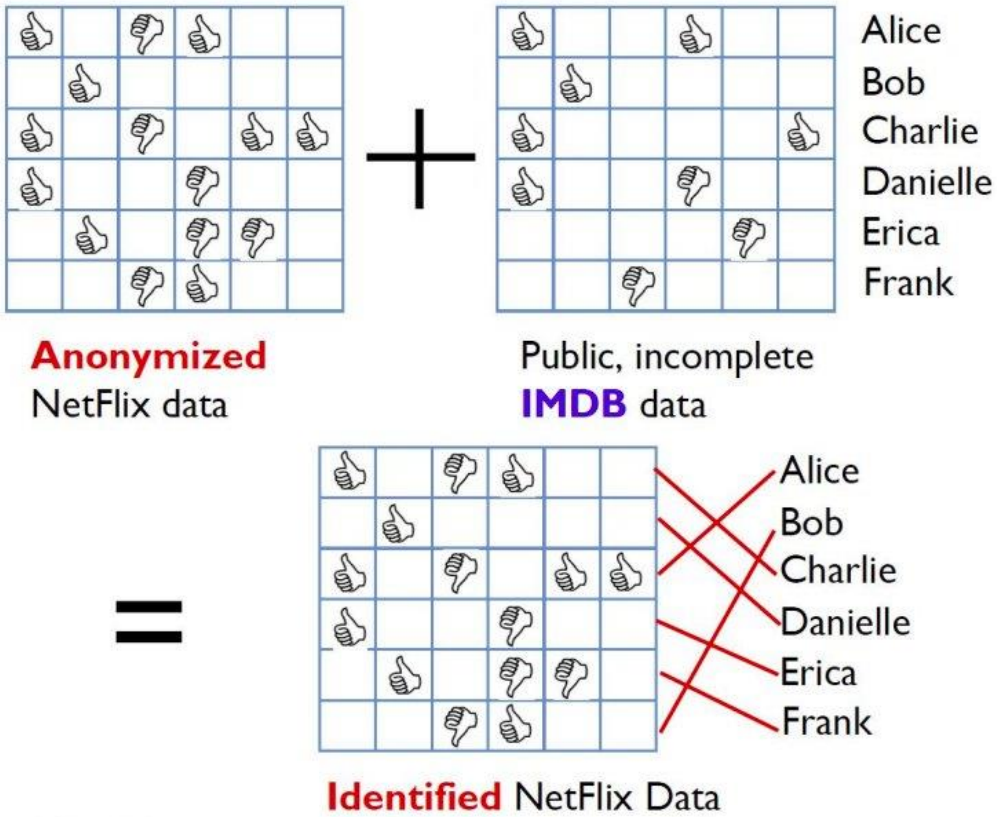


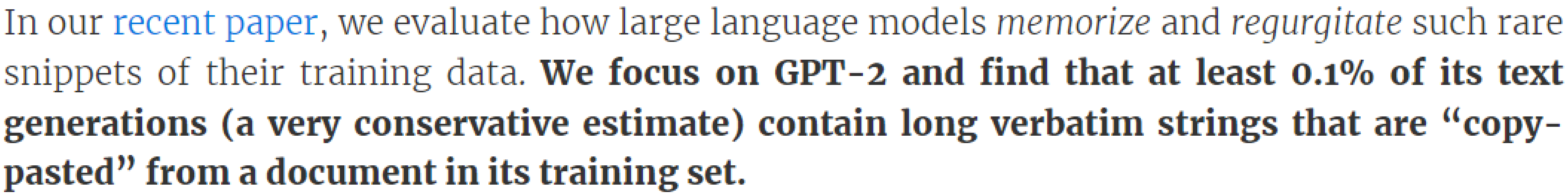
Image credit: Arvind Narayanan

Training Data

[Nasr, Carlini, Hayase, Jagielski, Cooper, Ippolito, Choquette-Choo, Wallace, Tramèr, Lee], 2025

Privacy Concerns

- Machine learning models are trained on very large datasets
- Can be coerced to reproduce training data verbatim!

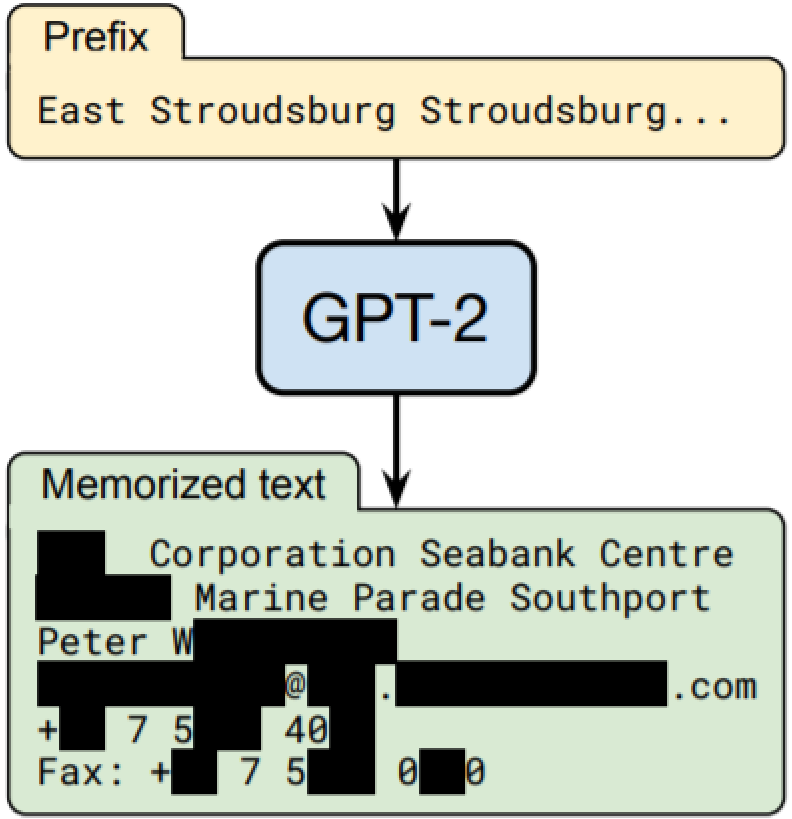


In our recent paper, we evaluate how large language models *memorize* and *regurgitate* such rare snippets of their training data. **We focus on GPT-2 and find that at least 0.1% of its text generations (a very conservative estimate) contain long verbatim strings that are "copy-pasted" from a document in its training set.**

Blog post: [Wallace, Tramer, Jagielski, Herbert-Voss], 2020

Paper: [Carlini, Tramer, Wallace, Jagielski, Herbert-Voss, Lee, Roberts, Brown, Song, Erlingsson, Oprea, Raffel], 2021

Personal Information



Prefix

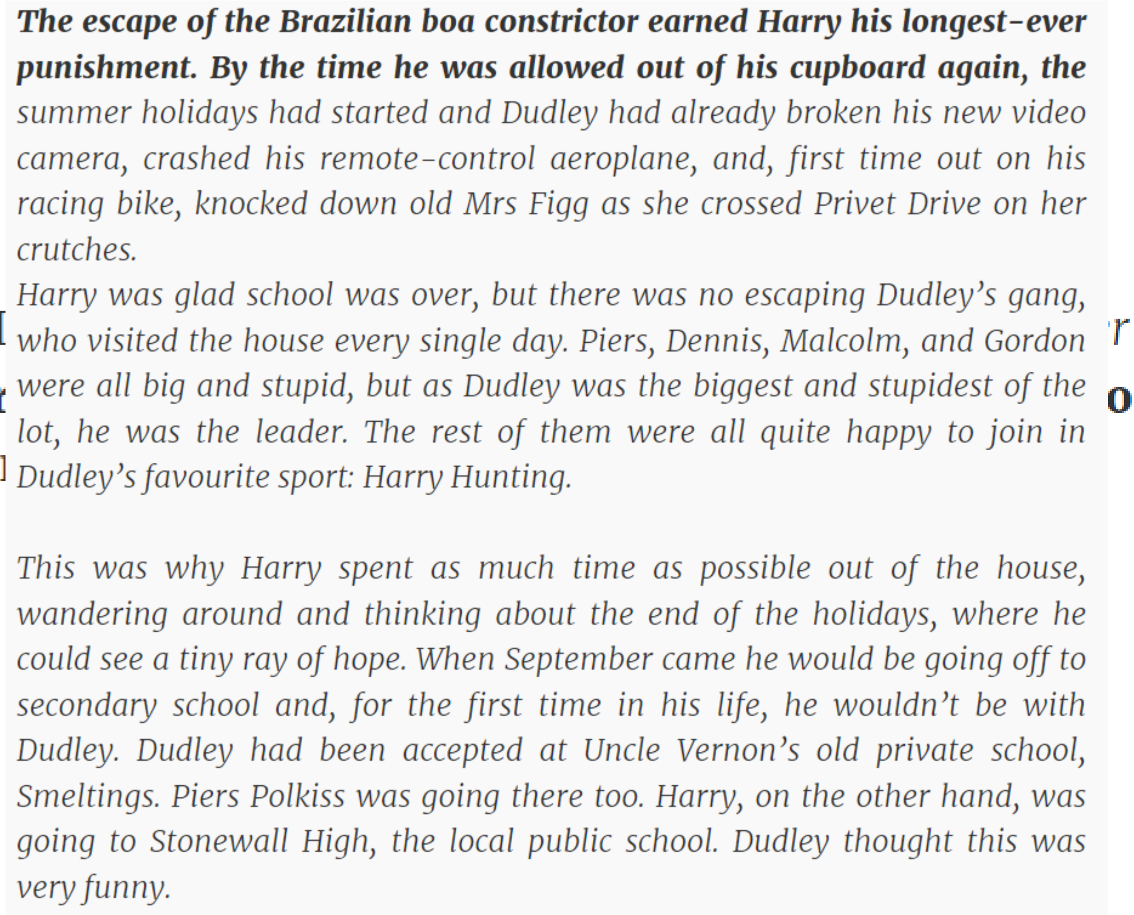
East Stroudsburg Stroudsburg...

Memorized text

[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Copyrighted content

Below, we prompt GPT r *and the Philosopher's*
Stone. **The model corr** **ok** (about 240 words)
before making its first r

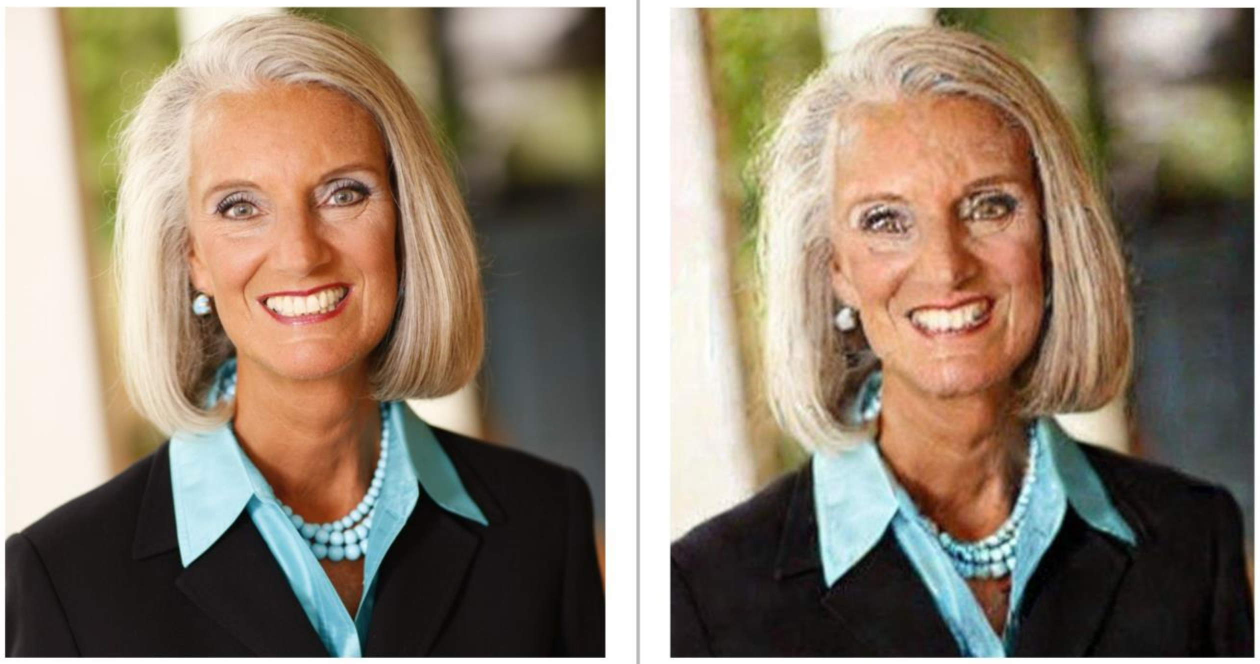


The escape of the Brazilian boa constrictor earned Harry his longest-ever punishment. By the time he was allowed out of his cupboard again, the *summer holidays had started and Dudley had already broken his new video camera, crashed his remote-control aeroplane, and, first time out on his racing bike, knocked down old Mrs Figg as she crossed Privet Drive on her crutches.*

Harry was glad school was over, but there was no escaping Dudley's gang, who visited the house every single day. Piers, Dennis, Malcolm, and Gordon were all big and stupid, but as Dudley was the biggest and stupidest of the lot, he was the leader. The rest of them were all quite happy to join in Dudley's favourite sport: Harry Hunting.

This was why Harry spent as much time as possible out of the house, wandering around and thinking about the end of the holidays, where he could see a tiny ray of hope. When September came he would be going off to secondary school and, for the first time in his life, he wouldn't be with Dudley. Dudley had been accepted at Uncle Vernon's old private school, Smeltings. Piers Polkiss was going there too. Harry, on the other hand, was going to Stonewall High, the local public school. Dudley thought this was very funny.

Also for Diffusion Models



Training Set

Caption: Living in the light with Ann Graham Lotz

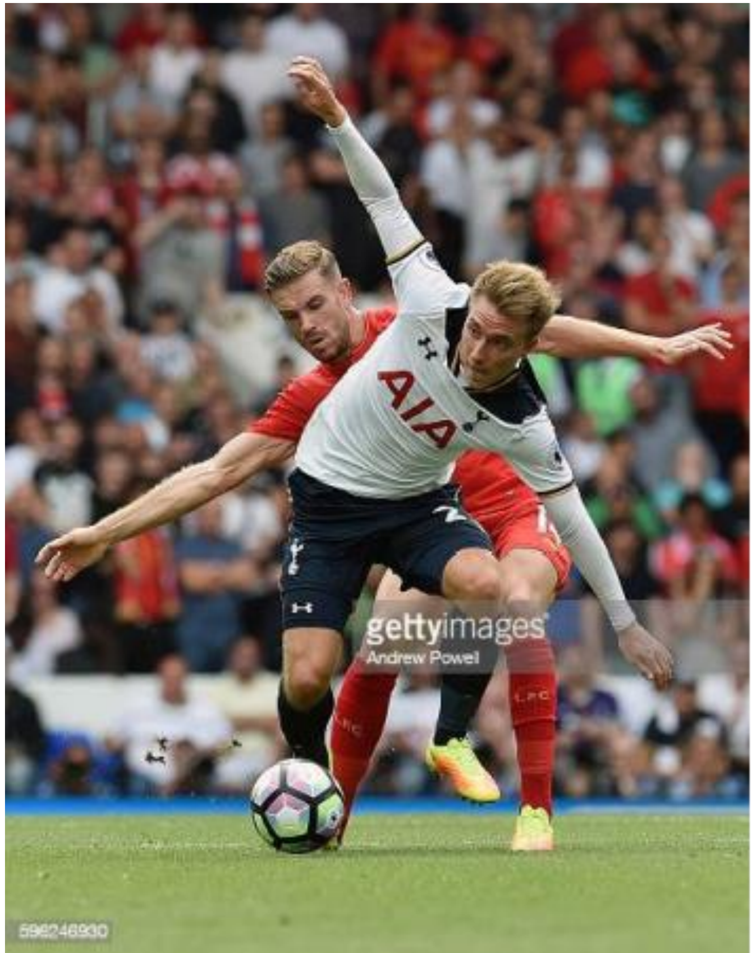
Generated Image

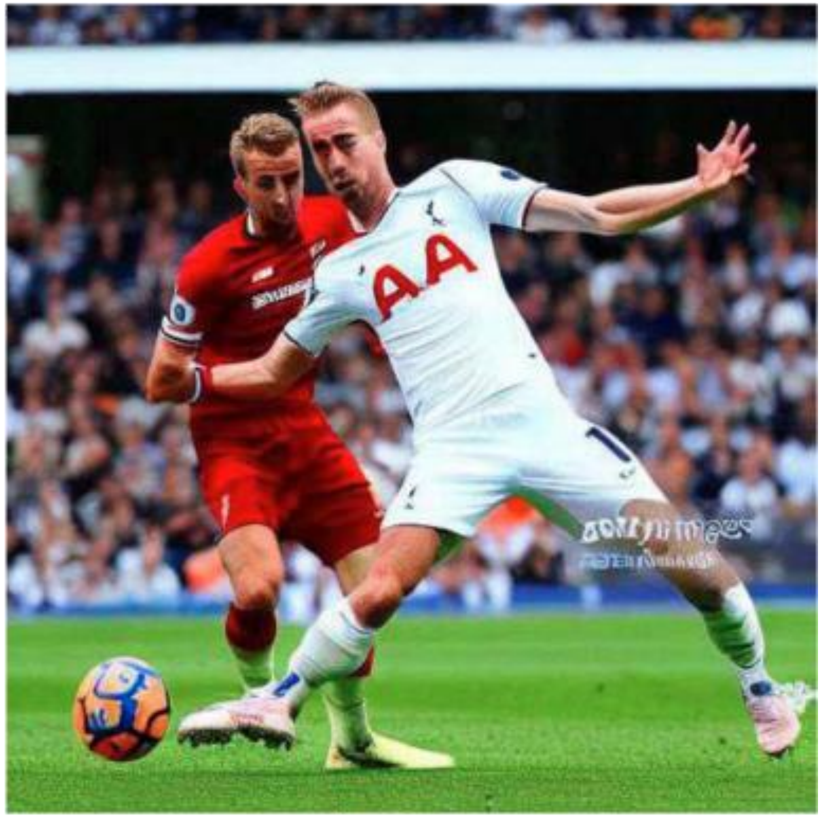
Prompt: Ann Graham Lotz

[Carlini, Hayes, Nasr, Jagielski, Sehwag, Tramer, Balle, Ippolito, Wallace], 2023

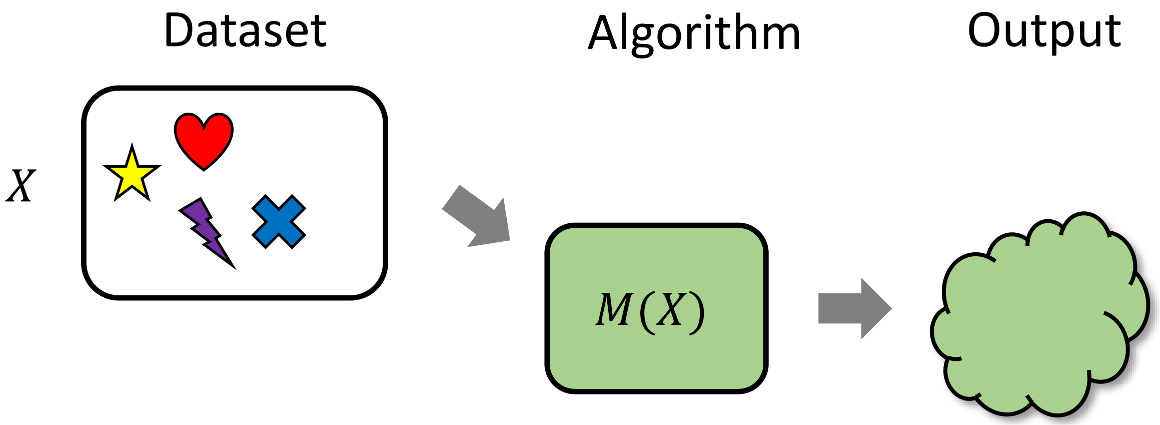
Implications for copyright…?

- Getty Images vs. Stability AI?



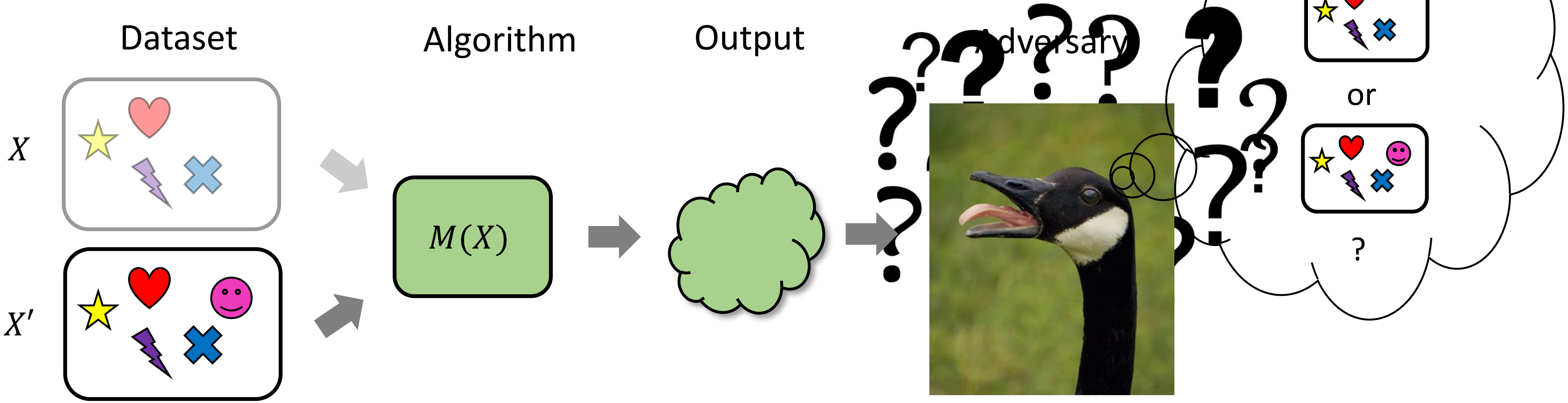


Differential Privacy (Informal)



[Dwork, McSherry, Nissim, Smith], 2006

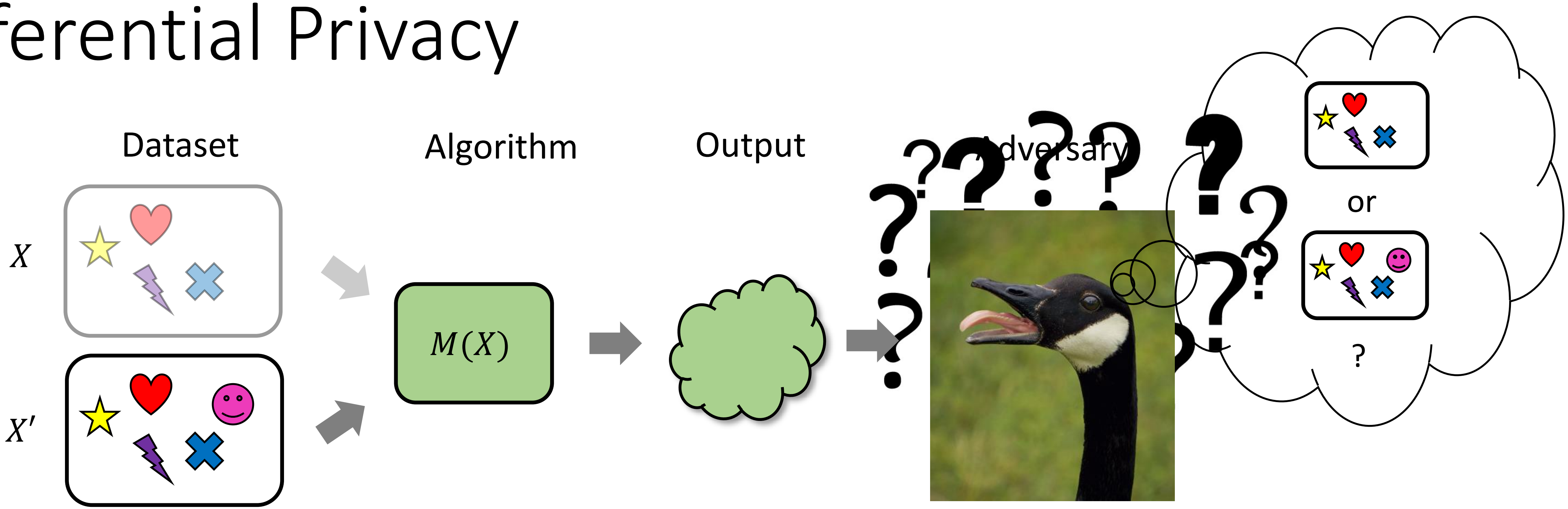
Differential Privacy (Informal)



“An algorithm is differentially private if its distribution over outputs doesn’t change much after adding/removing one point.”

[Dwork, McSherry, Nissim, Smith], 2006

Differential Privacy



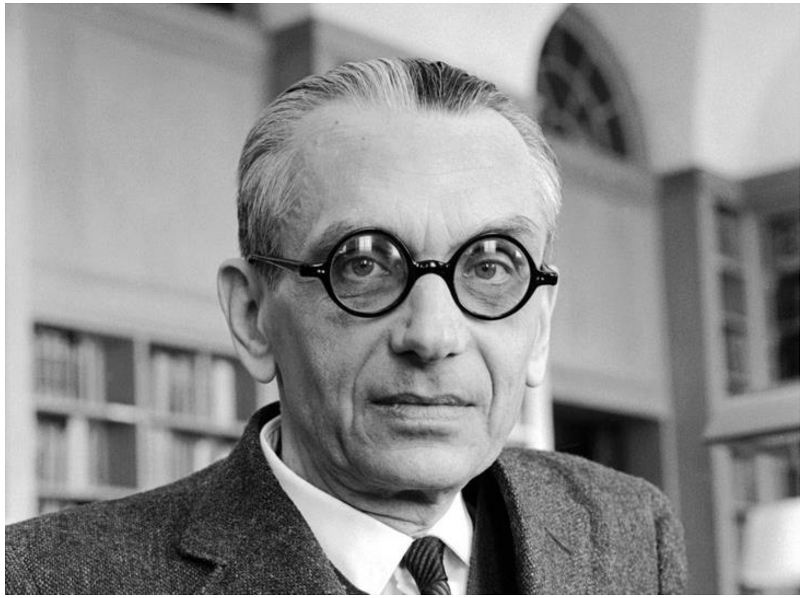
- $M: D^n \to R$ is (ε, δ)-DP if for all inputs X, X' which differ on one entry:

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

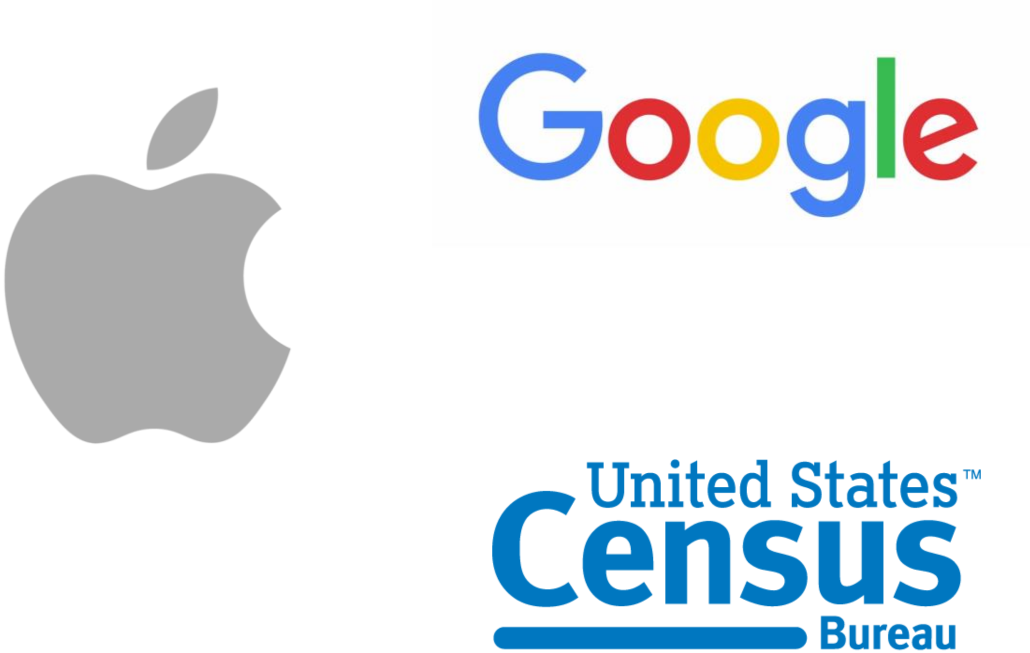
[Dwork, McSherry, Nissim, Smith], 2006

Widely Adopted

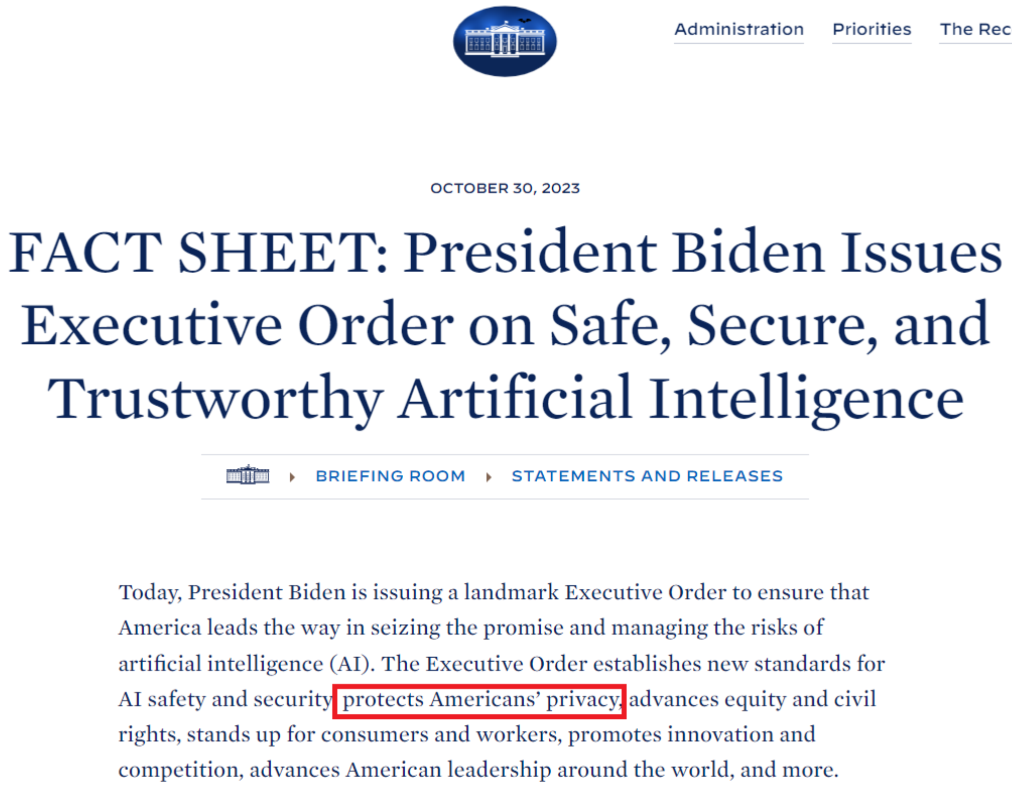
Accepted in theory…



And practice!



Endorsed by the White House



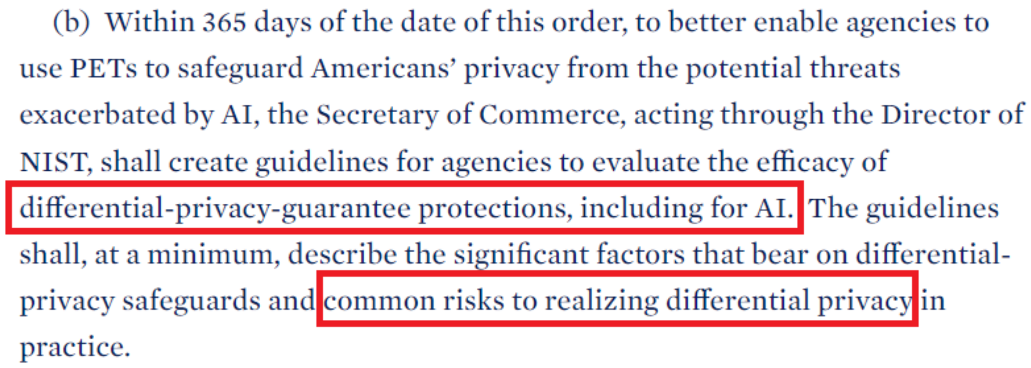
Administration Priorities The Record

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM ▸ STATEMENTS AND RELEASES

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.



(b) Within 365 days of the date of this order, to better enable agencies to use PETs to safeguard Americans' privacy from the potential threats exacerbated by AI, the Secretary of Commerce, acting through the Director of NIST, shall create guidelines for agencies to evaluate the efficacy of differential-privacy-guarantee protections, including for AI. The guidelines shall, at a minimum, describe the significant factors that bear on differential-privacy safeguards and common risks to realizing differential privacy in practice.

Differential Privacy (DMNS06)

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

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

- $\varepsilon \approx 1$ 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,

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

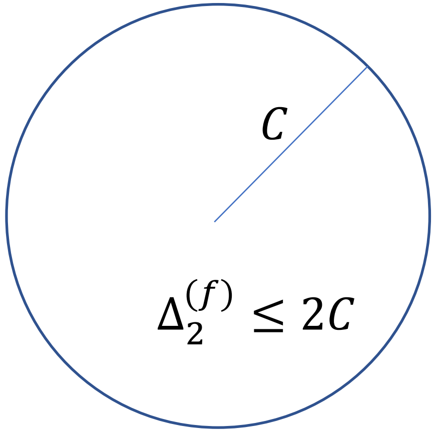
- Composition
  - If $M = (M_1, \ldots, M_k)$ is a sequence of k (ε, δ)-DP algorithms
    - M is $(k\varepsilon, k\delta)$-DP (Basic Composition)
    - M is $(O(\sqrt{k}\varepsilon \log(1/\delta')), k\delta + \delta')$-DP (Advanced Composition)

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 $\approx N(0, \Delta^2/\varepsilon^2)$

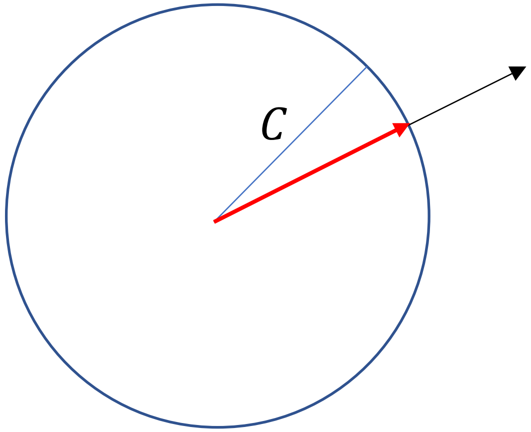
  - (ε, δ)-DP

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

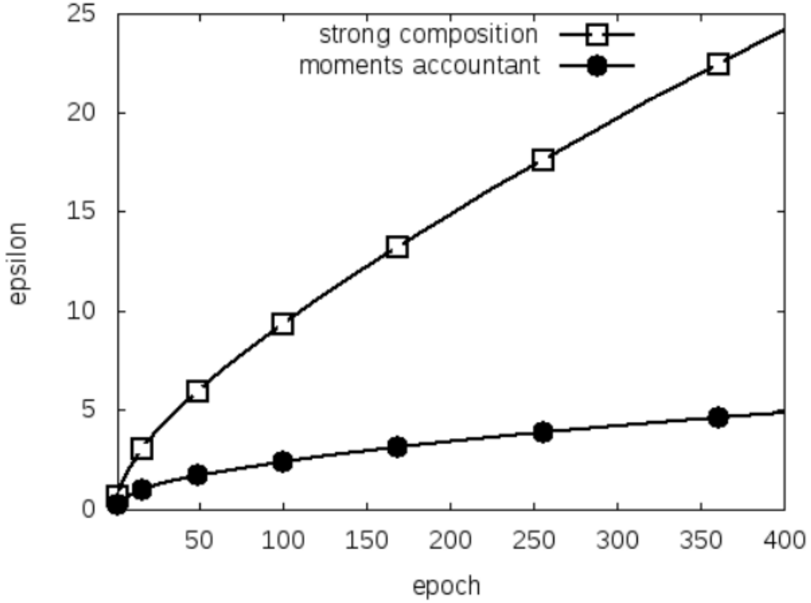
1. Sample a "lot" of points of (expected) size L by selecting each point to be in the lot with probability L/n
2. For each point in the lot, compute the gradient $\nabla\ell(\theta_t, x, y)$ and "clip" it to have ℓ_2 norm at most C
3. Average the clipped gradients and add Gaussian noise
   - Apply the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



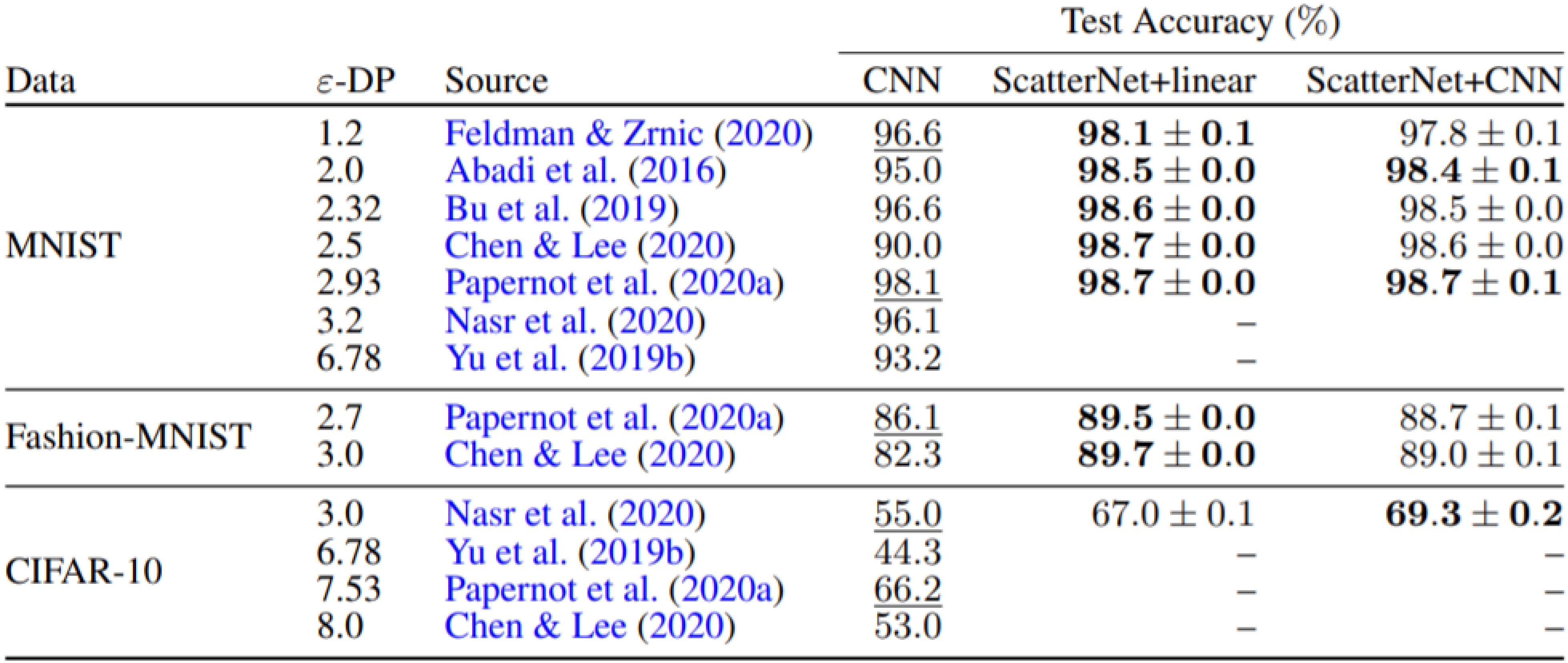
[Song-Chaudhuri-Sarwate '13, Bassily-Smith-Thakurta '14, Abadi-Chu-Goodfellow-McMahan-Mironov-Talwar-Zhang '16]

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"



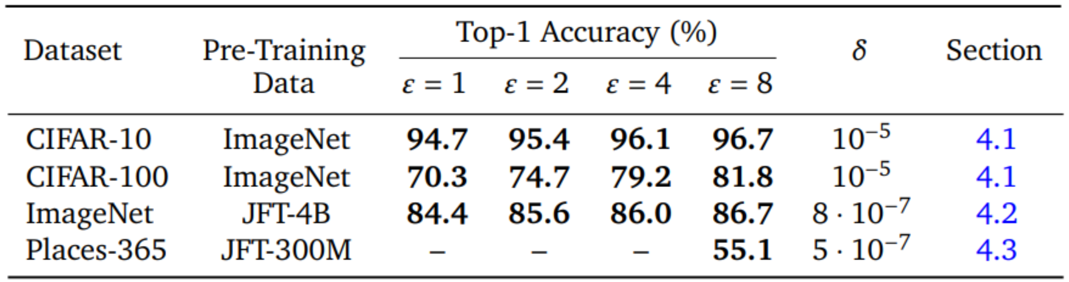
Does it work?



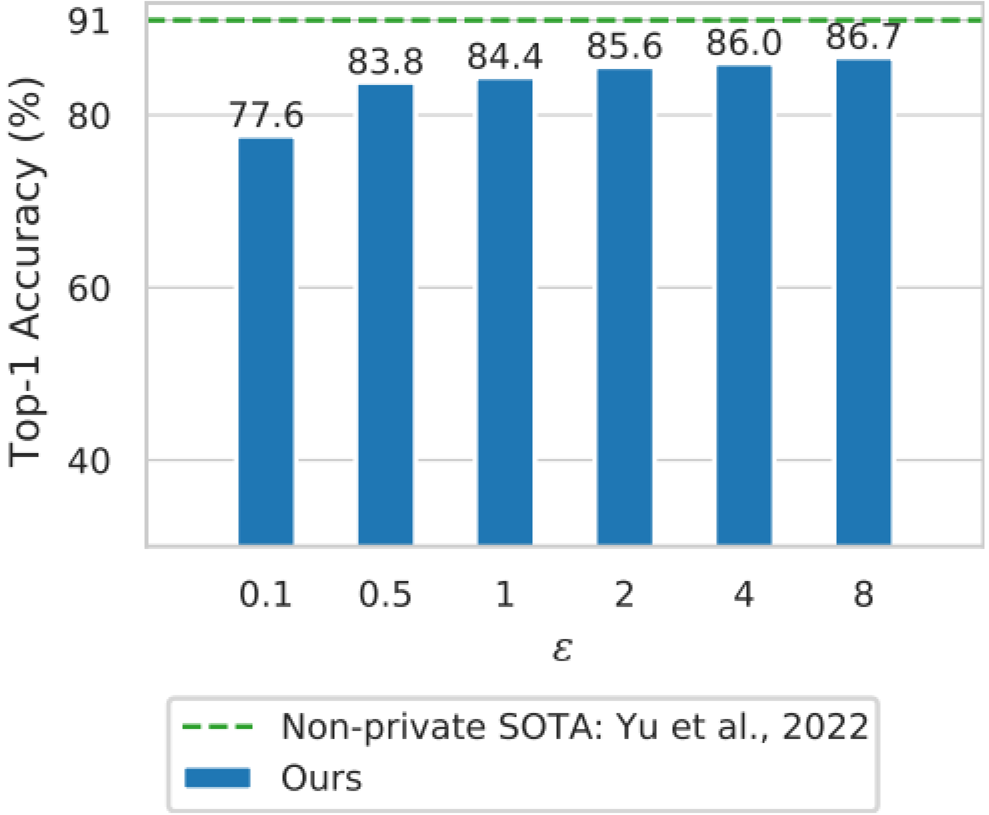
			Test Accuracy (%)		
Data	ε-DP	Source	CNN	ScatterNet+linear	ScatterNet+CNN
MNIST	1.2	Feldman & Zrnic (2020)	96.6	**98.1 ± 0.1**	97.8 ± 0.1
	2.0	Abadi et al. (2016)	95.0	**98.5 ± 0.0**	**98.4 ± 0.1**
	2.32	Bu et al. (2019)	96.6	**98.6 ± 0.0**	98.5 ± 0.0
	2.5	Chen & Lee (2020)	90.0	**98.7 ± 0.0**	98.6 ± 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]

Public Data Helps for Private Vision

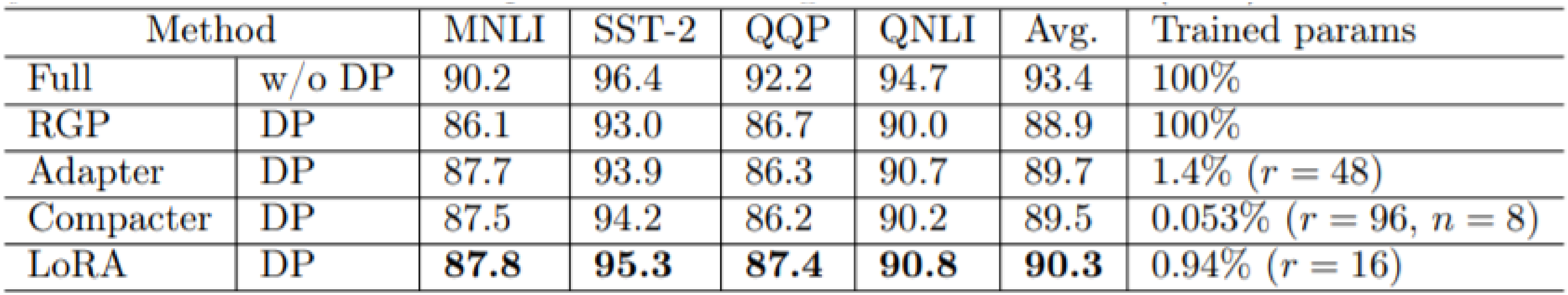


Dataset	Pre-Training Data	Top-1 Accuracy (%)				δ	Section
		$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$		
CIFAR-10	ImageNet	**94.7**	**95.4**	**96.1**	**96.7**	10^{-5}	4.1
CIFAR-100	ImageNet	**70.3**	**74.7**	**79.2**	**81.8**	10^{-5}	4.1
ImageNet	JFT-4B	**84.4**	**85.6**	**86.0**	**86.7**	$8 \cdot 10^{-7}$	4.2
Places-365	JFT-300M	–	–	–	**55.1**	$5 \cdot 10^{-7}$	4.3



[De, Berrada, Hayes, Smith, Balle], 2022

LLMs can be Fine-Tuned Privately!



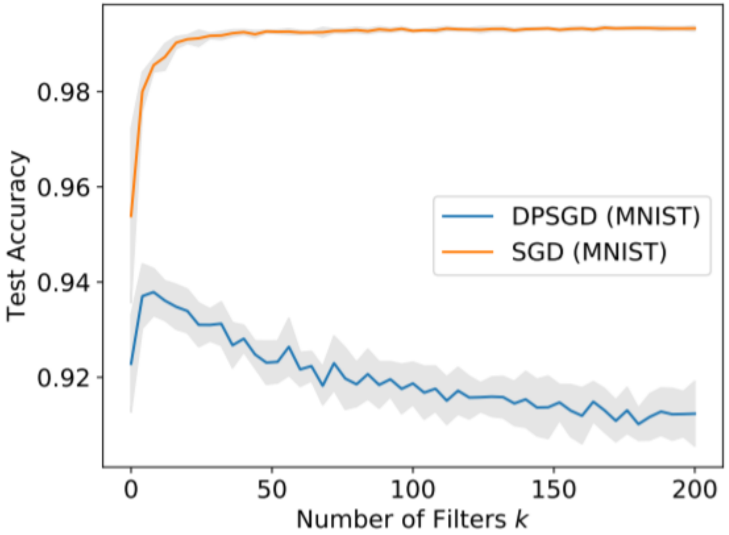
Method		MNLI	SST-2	QQP	QNLI	Avg.	Trained params
Full	w/o DP	90.2	96.4	92.2	94.7	93.4	100%
RGP	DP	86.1	93.0	86.7	90.0	88.9	100%
Adapter	DP	87.7	93.9	86.3	90.7	89.7	1.4% ($r = 48$)
Compacter	DP	87.5	94.2	86.2	90.2	89.5	0.053% ($r = 96$, $n = 8$)
LoRA	DP	**87.8**	**95.3**	**87.4**	**90.8**	**90.3**	0.94% ($r = 16$)

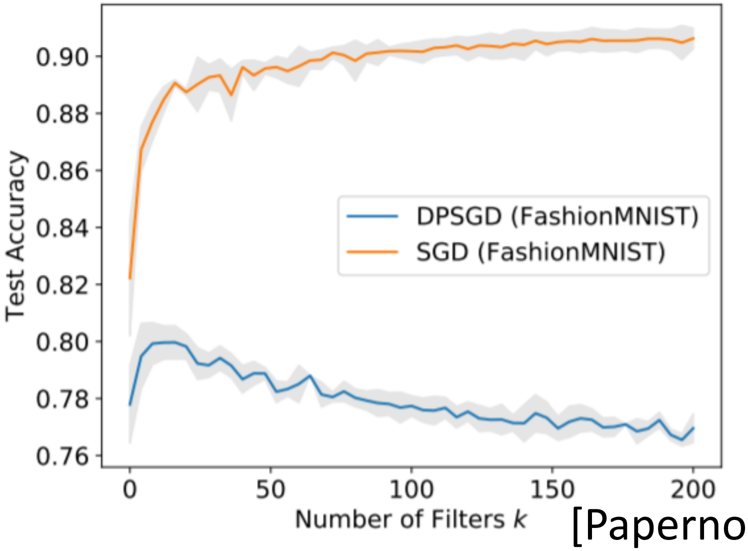
- RoBERTa-Large, $\varepsilon = 6.7$
- 3% average drop from non-private to private
    - Compare with CIFAR-10: 99% non-private to 69% private

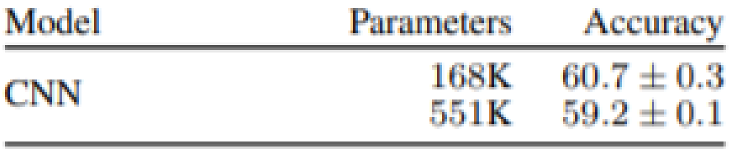
[Yu, Naik, Backurs, Gopi, Inan, Kamath, Kulkarni, Lee, Manoel, Wutschitz, Yekhanin, Zhang], 2022

Is size beneficial in private ML?

- Fewer parameters = better model (??)
  - Noise magnitude introduced due to privacy scales as $\sqrt{p}$
  - “Have to balance model capacity with magnitude of noise” (?)



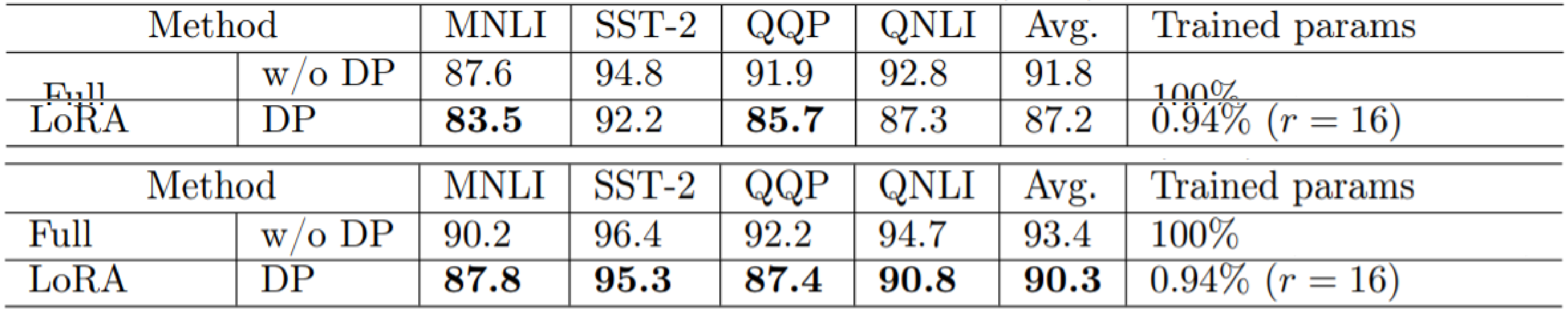




Model	Parameters	Accuracy
CNN	168K 551K	60.7 ± 0.3 59.2 ± 0.1

[Papernot, Chien, Song, Thakurta, Erlingsson], 2019
[Tramèr, Boneh], 2021

Private fine-tuning of LLMs benefits from large models

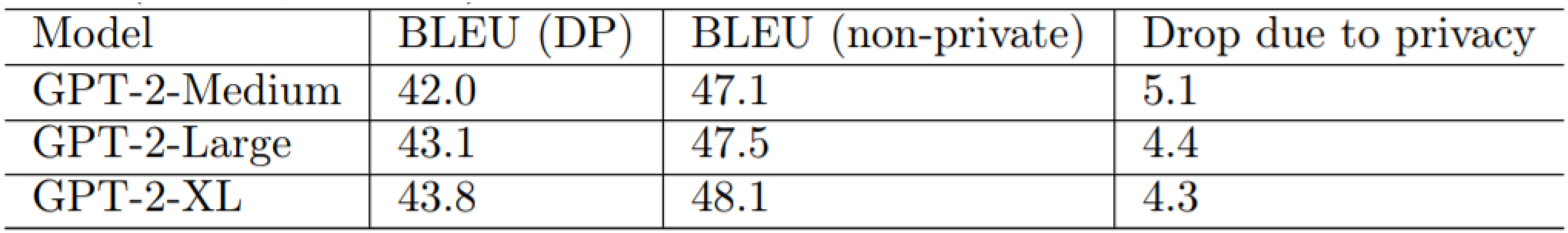


RoBERTa-Base

Method		MNLI	SST-2	QQP	QNLI	Avg.	Trained params
Full	w/o DP	87.6	94.8	91.9	92.8	91.8	100%
LoRA	DP	**83.5**	92.2	**85.7**	87.3	87.2	0.94% ($r = 16$)

RoBERTa-Large

Method		MNLI	SST-2	QQP	QNLI	Avg.	Trained params
Full	w/o DP	90.2	96.4	92.2	94.7	93.4	100%
LoRA	DP	**87.8**	**95.3**	**87.4**	**90.8**	**90.3**	0.94% ($r = 16$)



Model	BLEU (DP)	BLEU (non-private)	Drop due to privacy
GPT-2-Medium	42.0	47.1	5.1
GPT-2-Large	43.1	47.5	4.4
GPT-2-XL	43.8	48.1	4.3

- Bigger models → Better absolute error, and less drop due to privacy

[Yu, Naik, Backurs, Gopi, Inan, Kamath, Kulkarni, Lee, Manoel, Wutschitz, Yekhanin, Zhang], 2022