An Introduction to Differential Privacy

Gautam Kamath

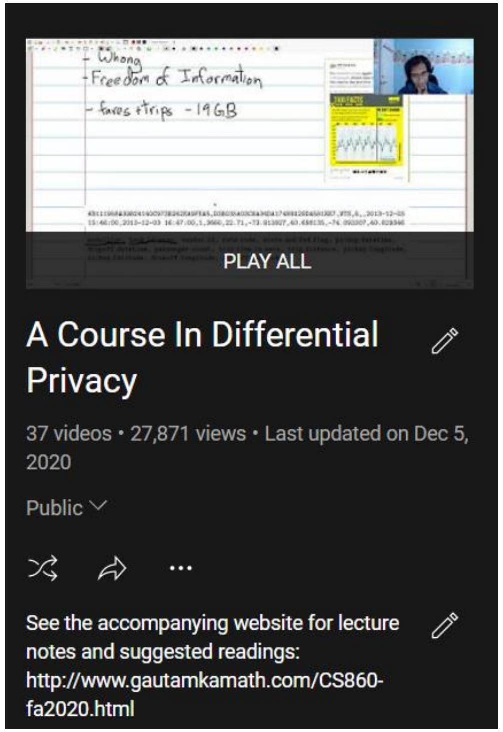
University of Waterloo

Iran Workshop on Communication and Information Theory

May 11, 2022

Before we begin…

- *Only* 3 hours!
  - Basics of differential privacy
  - Private machine learning
- For more, see:
  - My course
    - http://www.gautamkamath.com/CS860-fa2020.html
    - Full set of lecture notes and videos
  - DifferentialPrivacy.org
    - Blog posts, as well as links to resources



A Course In Differential Privacy

37 videos • 27,871 views • Last updated on Dec 5, 2020

See the accompanying website for lecture notes and suggested readings: http://www.gautamkamath.com/CS860-fa2020.html

DifferentialPrivacy.org

What is privacy?

~~What is privacy?~~ What is *not* private?

Netflix Prize

- Can you predict which movies a user will like?
  - $1M grand prize for best prediction engine
- Dataset: (anonymized) user ID, movie ID, rating, date
  - No way to identify users, right…?
- Wrong! Narayanan and Shmatikov matched users with IMDb
- Cancellation of the 2nd Netflix prize

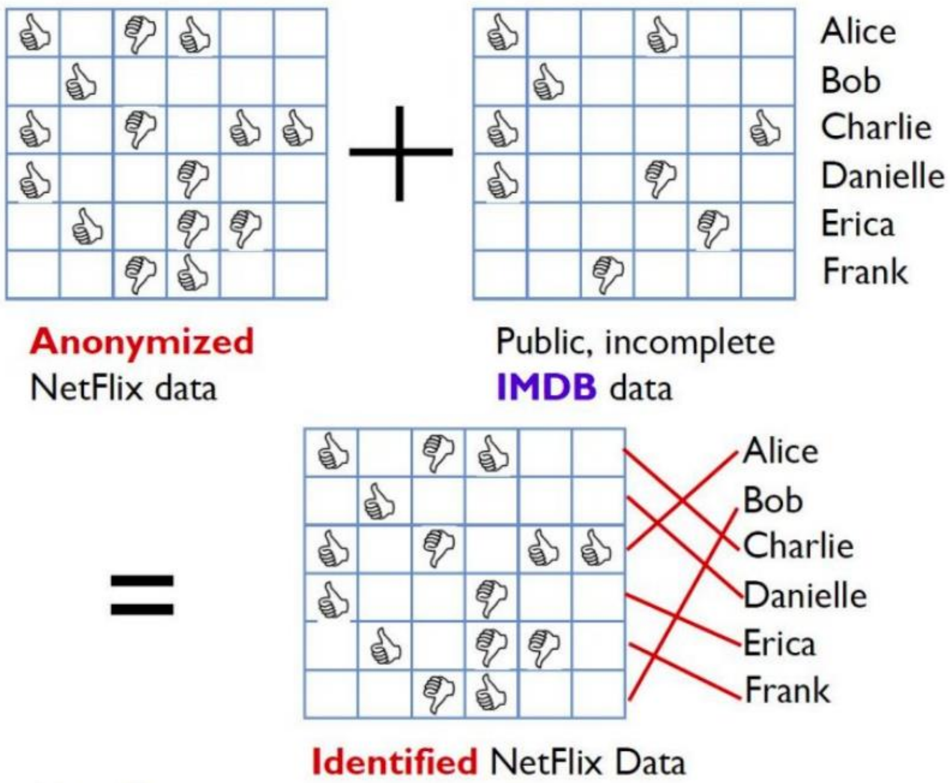


Image credit: Arvind Narayanan

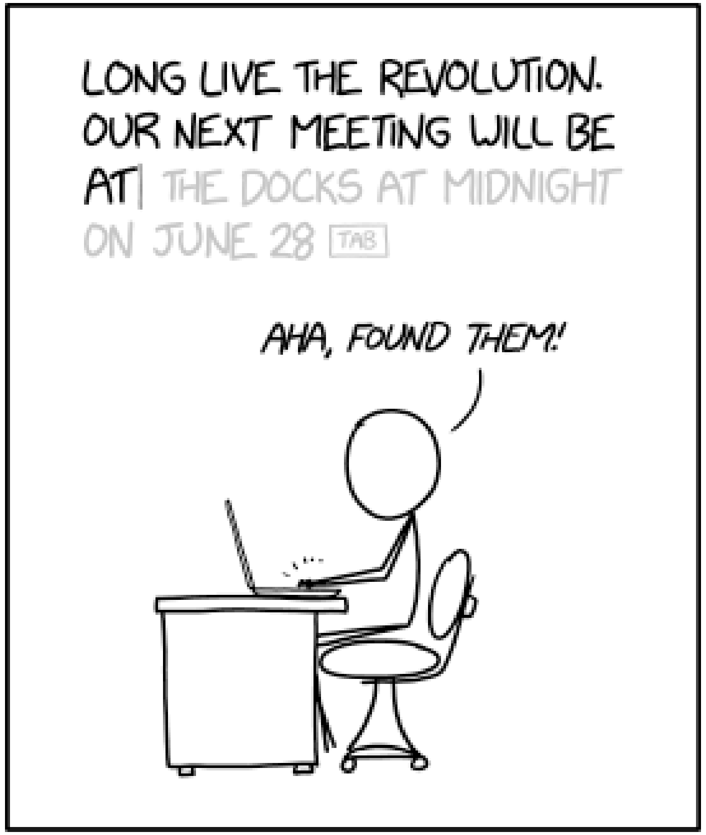
Massachusetts Group Insurance Commission

- Public release of hospital visit records for every state employee
  - Very sensitive!
- Anonymized data
  - Original: Name, SSN, ZIP code, date of birth, sex, condition
  - Anonymized: ~~Name~~, ~~SSN~~, ZIP code, date of birth, sex, condition
- Latanya Sweeney: bought voter rolls
  - Contain name, address, ZIP code, date of birth, sex

Massachusetts Group Insurance Commission

- Public release of hospital visit records for every state employee
  - Very sensitive!
- Anonymized data
  - Original: Name, SSN, ZIP code, date of birth, sex, condition
  - Anonymized: ~~Name~~, ~~SSN~~, ZIP code, date of birth, sex, condition
- Latanya Sweeney: bought voter rolls
  - Contain name, address, ZIP code, date of birth, sex
  - Can match up and reidentify!
- Cynthia Dwork: “Anonymized data isn’t.”

Memorization in Machine Learning

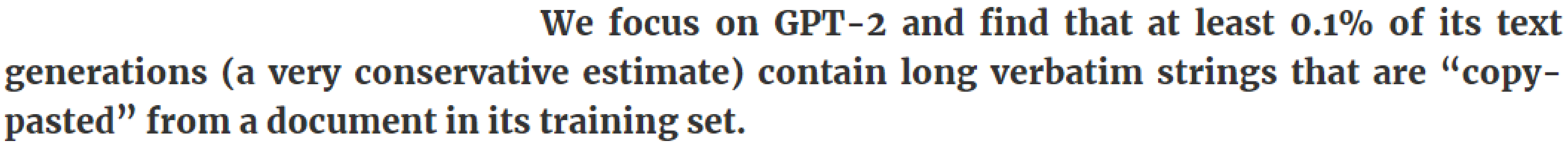

LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28
TAB
AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

https://xkcd.com/2169/

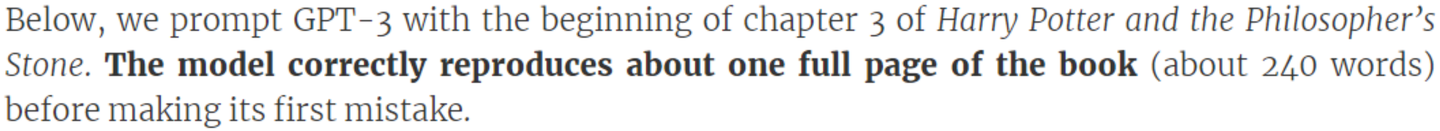
Machine Learning Models are Vulnerable!

- Trained on very large datasets
- Can be coerced to reproduce training data verbatim!

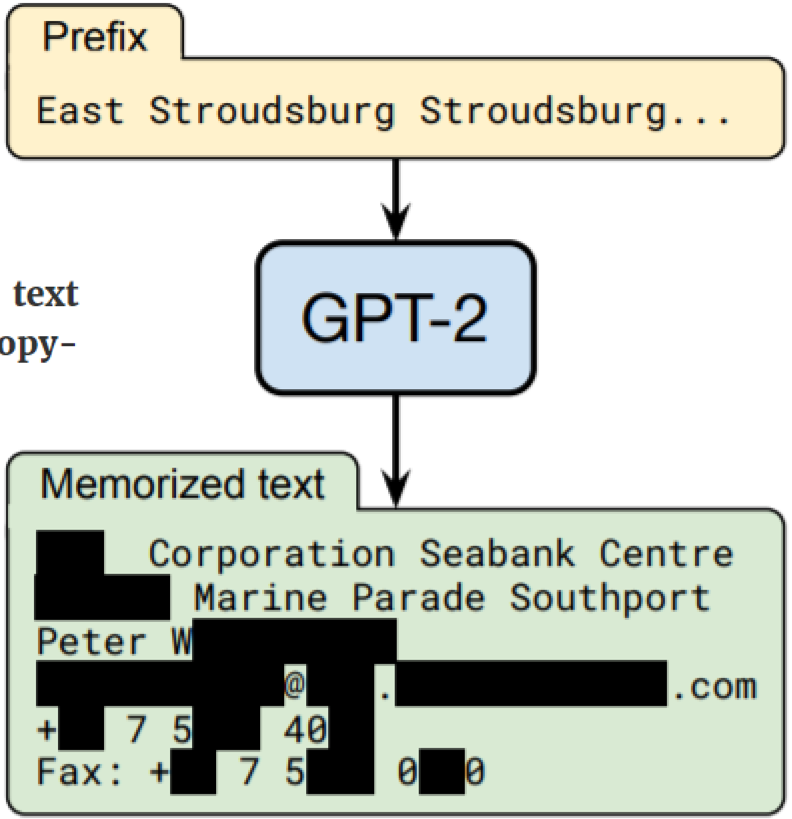


We focus on GPT-2 and find that at least 0.1% of its text generations (a very conservative estimate) contain long verbatim strings that are "copy-pasted" from a document in its training set.

- Personal information, copyrighted content



Below, we prompt GPT-3 with the beginning of chapter 3 of *Harry Potter and the Philosopher's Stone*. **The model correctly reproduces about one full page of the book** (about 240 words) before making its first mistake.



```
East Stroudsburg Stroudsburg...
```

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

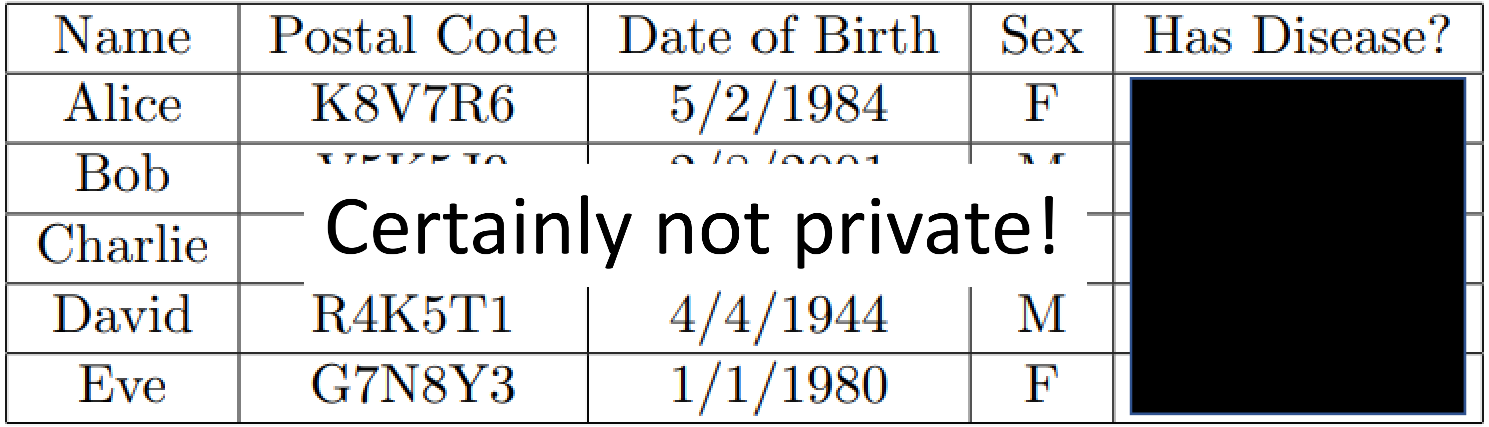
Blog post: [Wallace, Tramer, Jagielski, Herbert-Voss], 2020

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

A Concrete Scenario: Database Reconstruction

- Database with public identifiers, but one sensitive piece of info
- Analyst can ask “queries” involving the sensitive data
- Database returns the answer
- Can they “reconstruct” the sensitive data?

Database Reconstruction

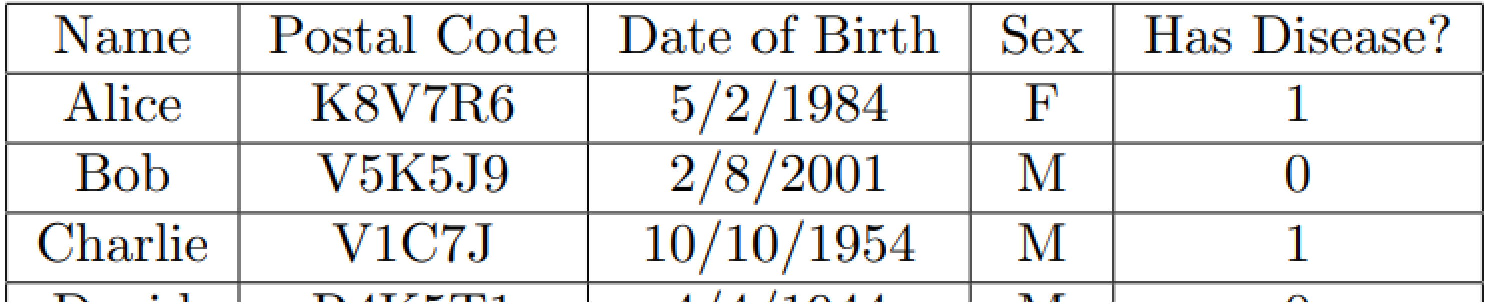


Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	[redacted]
Bob	[illegible]	[illegible]	[illegible]	[redacted]
Charlie	[illegible]	[illegible]	[illegible]	[redacted]
David	R4K5T1	4/4/1944	M	[redacted]
Eve	G7N8Y3	1/1/1980	F	[redacted]

Certainly not private!

- “How many females have the disease?” Response: 2
- “How many people born in the first half of the year have the disease?” Response: 2
- “How many males have the disease?” Response: 2

Noisy Database Reconstruction



Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1

Can we still reconstruct the database, even with noisy answers?

- "How many females have the disease?" Response: 3
- "How many people born in the first half of the year have the disease?" Response: 1
- "How many males have the disease?" Response: 4

Database Reconstruction Theorem

Theorem: Suppose there is a database with n individuals. If an analyst is allowed to ask 2^n queries, and the database returns errors with noise per query at most E, then the analyst can reconstruct all but $4E$ of the secret bits.

“If the analyst asks enough queries, they can reconstruct almost all the secret bits, even if the queries have noise added”

(Simple attack: find any database consistent with provided answers)

[Dinur, Nissim], 2003

(A Better) Database Reconstruction Theorem

Theorem: Suppose there is a database with n individuals. If an analyst is allowed to ask $O(n)$ queries, and the database returns errors with noise per query at most $O(\sqrt{n})$, then the analyst can reconstruct all but $O(1)$ of the secret bits.

“If the analyst asks enough queries, they can reconstruct almost all the secret bits, even if the queries have noise added”

(Simple attack: find any database consistent with provided answers)

[Dinur, Nissim], 2003

(A Better) Database Reconstruction Theorem

Theorem: Suppose there is a database with n individuals. If an analyst is allowed to ask $O(n)$ queries, and the database returns errors with noise per query at most $O(\sqrt{n})$, then the analyst can reconstruct all but $O(1)$ of the secret bits.

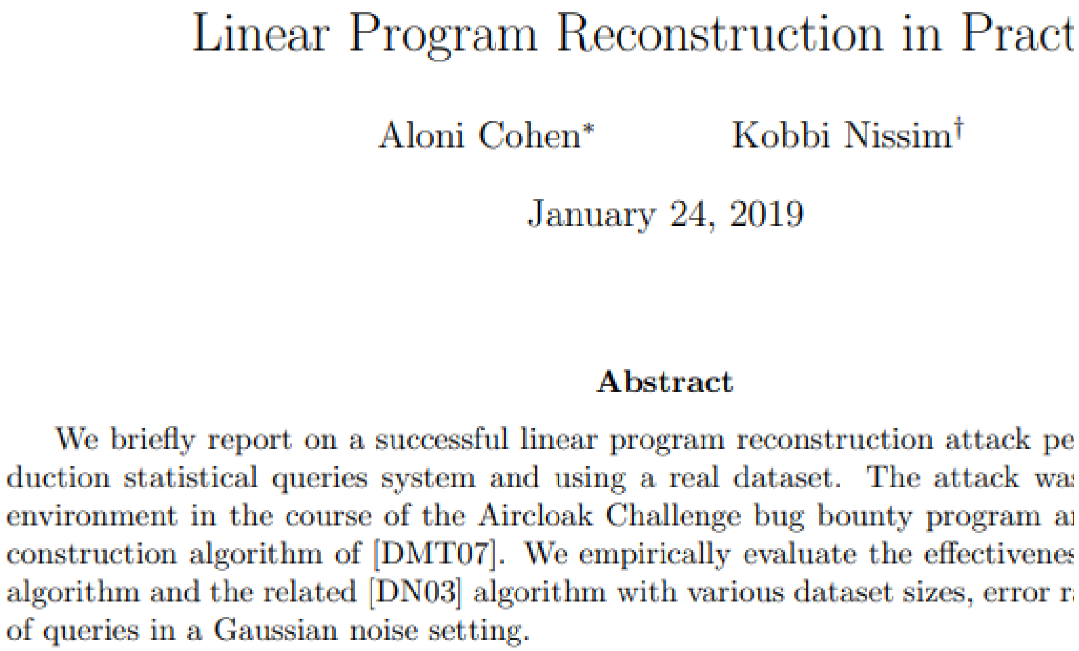
“Blatantly non-private”!

“If the analyst asks enough queries, they can reconstruct almost all the secret bits, even if the queries have noise added”

(Simple attack: find any database consistent with provided answers)

[Dinur, Nissim], 2003

Reconstruction Attacks are Practical!



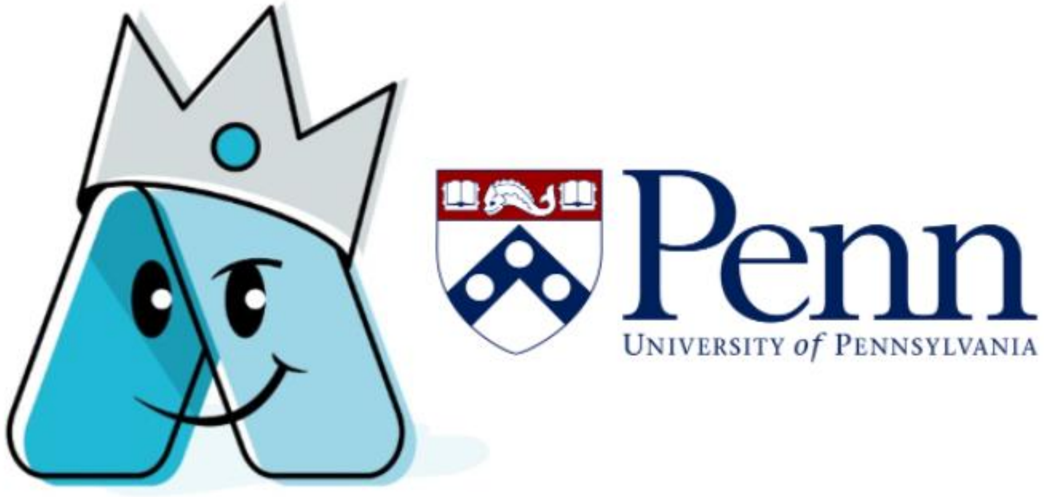
Linear Program Reconstruction in Pract

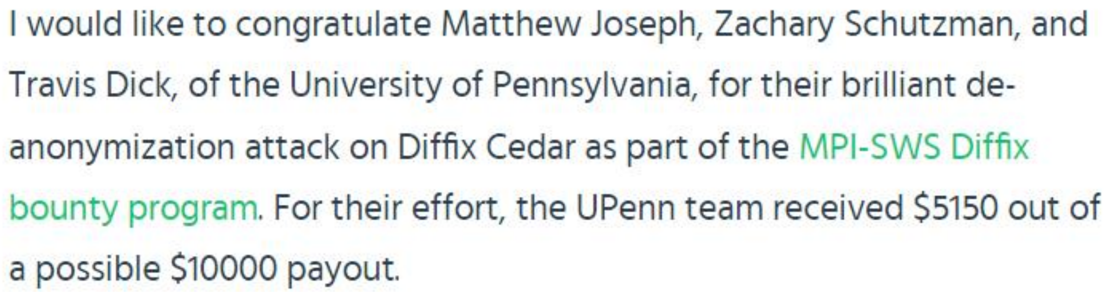
Aloni Cohen[*] Kobbi Nissim[†]

January 24, 2019

Abstract

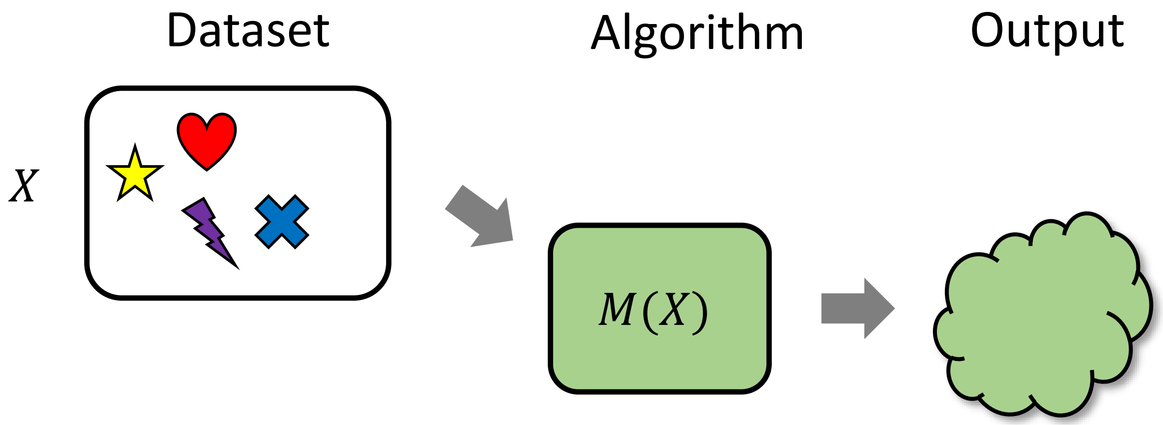
We briefly report on a successful linear program reconstruction attack pe duction statistical queries system and using a real dataset. The attack wa: environment in the course of the Aircloak Challenge bug bounty program a: construction algorithm of [DMT07]. We empirically evaluate the effectivenes algorithm and the related [DN03] algorithm with various dataset sizes, error r: of queries in a Gaussian noise setting.





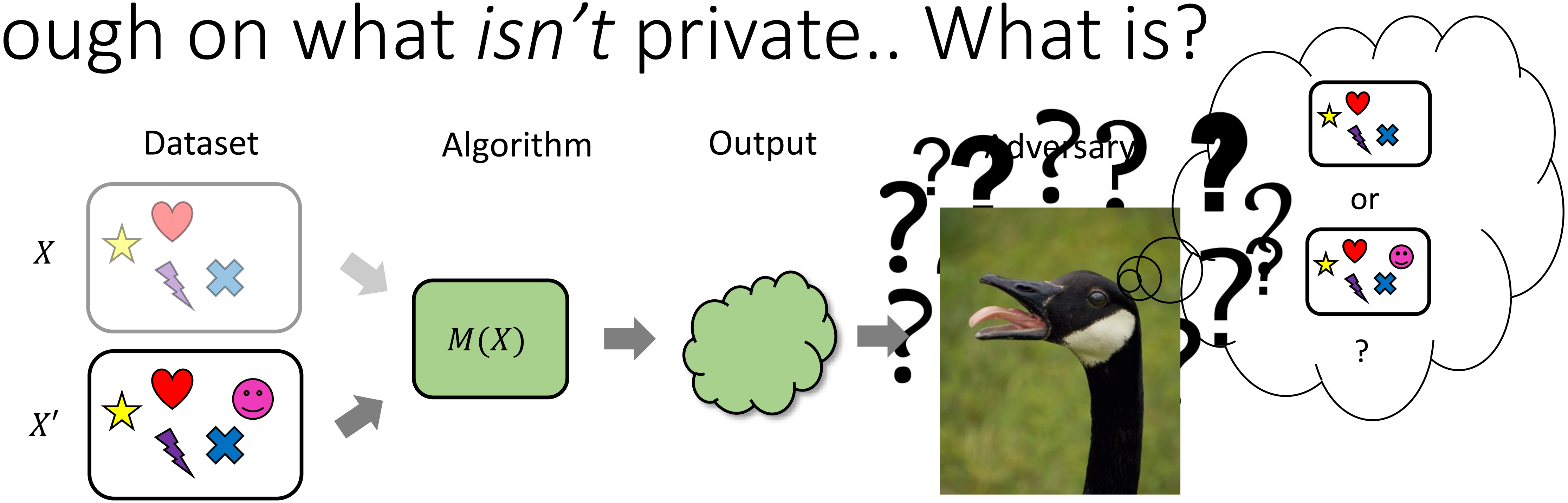
I would like to congratulate Matthew Joseph, Zachary Schutzman, and Travis Dick, of the University of Pennsylvania, for their brilliant de-anonymization attack on Diffix Cedar as part of the MPI-SWS Diffix bounty program. For their effort, the UPenn team received $5150 out of a possible $10000 payout.

Enough on what *isn't* private.. What is?



[Dwork, McSherry, Nissim, Smith], 2006

Enough on what *isn't* private.. What is?



"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 (Informal)

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

Why is this a reasonable notion of privacy?

- Dropping a user's datapoint is unlikely to change the output
- Thus looking at the output, can't tell if a user was in the dataset or not
- If you can't even know if a user is present, you can't know their data
- E.g., protects against database reconstruction attacks (and much more!)

Differential Privacy

- Setup: n datapoints $X = X_1, \ldots, X_n$, given to a "trusted curator"
- The curator has an algorithm $M : \mathcal{X}^n \to \mathcal{Y}$, outputs $M(X)$

Definition: An algorithm M is ε-differentially private (DP) if for all datasets X and X' which differ in one entry, and for all events $S \subseteq \mathcal{Y}$,

$$\Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S].$$

[Dwork, McSherry, Nissim, Smith], 2006

Comments on Differential Privacy

Definition: An algorithm M is ε-differentially private (DP) if for all datasets X and X' which differ in one entry, and for all events $S \subseteq \mathcal{Y}$,

$$\Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S].$$

- Quantitative in ε
- Bounds the multiplicative increase in prob of any event
- Symmetric definition (swap X and X')
- Can consider either “add/remove” or “change” one point
    - Equivalent up to factor of 2 in ε

[Dwork, McSherry, Nissim, Smith], 2006

What does DP protect against?

- Database reconstruction
  - Finding a user’s private data
- Membership inference
  - Determining whether or not a user was in the dataset
- Learning anything about a user that can’t be inferred w/o them

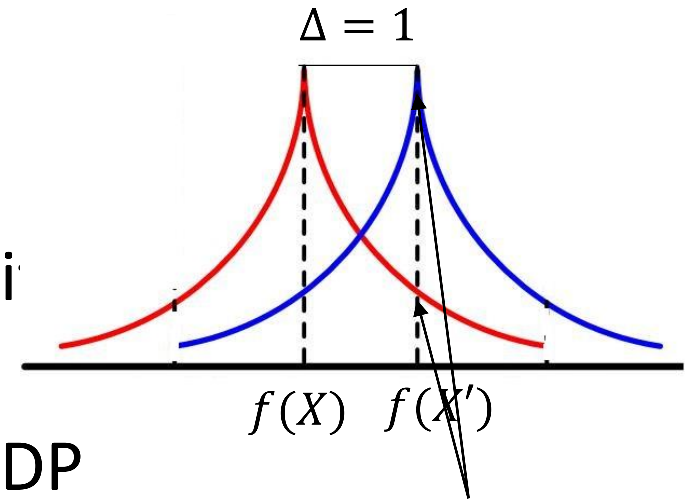
What *doesn't* DP do?

- Important: does **not** prevent inferences (statistics/machine learning)
  - (Public) smoker participates in (differentially private) study investigating whether smoking causes cancer
  - Reveals that smoking causes cancer! Smoker's insurance premiums increase!
  - Was their (differential) privacy violated?
  - No: smoking → cancer could be inferred whether or not they participated
  - Differential privacy: outcome of algorithm is similar, whether or not someone participates
- Not appropriate when individual identities are important
  - "Private" contact tracing

On to the algorithms!

The Laplace Mechanism

- Suppose $X_1, \dots, X_n \in \{0, 1\}$.
- Goal: privately compute sum $f(X) = \sum_{i=1}^{n} X_i$
- How much can $f(X)$ change if a single X_i is modi
  - *Sensitivity* $\Delta = 1$
- Claim: $f(X) + Z$, where $Z \sim \text{Laplace}(1/\varepsilon)$, is ε-DP
- $\text{Laplace}(\sigma) \propto \exp(-|x|/\sigma)$
  - Two-sided Exponential distribution
- Magnitude of error $\approx 1/\varepsilon$



Ratio: $\frac{\exp\left(-\frac{0}{1/\varepsilon}\right)}{\exp\left(-\frac{1}{1/\varepsilon}\right)} = \frac{1}{\exp(-\varepsilon)} = e^{\varepsilon}$

Example: Counting Queries

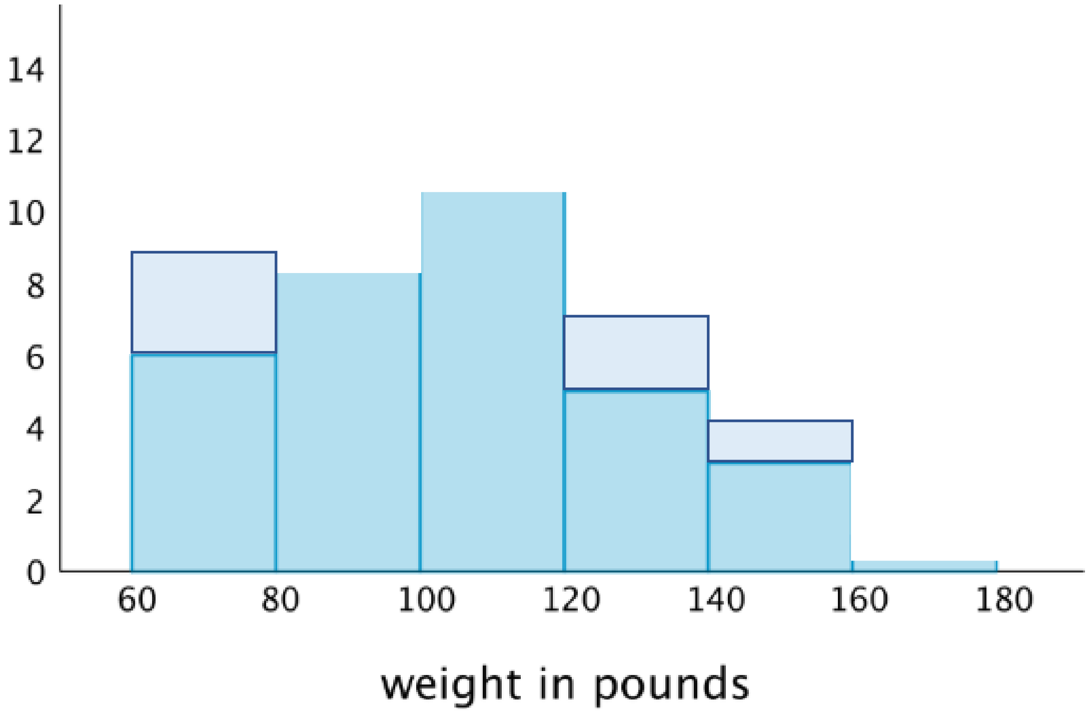
- Asking a population of 100 people, how many smoke?
- E.g., 20 say yes
- Can guarantee 1-DP by adding $\mathrm{Laplace}(1)$ noise
  - Sampled outputs: $20.68, 19.24, 20.28, 19.83, \ldots$
- Stronger privacy: 0.1-DP by adding $\mathrm{Laplace}(10)$ noise
  - Sampled outputs: $22.45, 11.45, 2.4, 15.03, 29.47$

The Laplace Mechanism

- More broadly: let $f(X) : \mathcal{X}^n \to \mathbf{R}^d$ be a function of interest
- Let $\Delta_1^f = \max_{X,X' \text{ differ in one entry}} \|f(X) - f(X')\|_1$ be ℓ_1-sensitivity of f
  - "How much can the function change by modifying one datapoint?"
- Theorem: The Laplace Mechanism $f(X) + \mathrm{Lap}\left(\Delta_1^f/\varepsilon\right)^{\otimes d}$ is ε-DP
  - "Add Laplace noise to each coordinate, proportional to the ℓ_1-sensitivity"

Application: Private Histograms

- Number of datapoints that fall into each of k categories
- At most 1 count can change by at most 1 by adding/removing datapoint
  - $\Delta_1 = 1$
- Add $Z_i \sim \text{Laplace}(1/\varepsilon)$ noise to ith count
  - Noised counts are ε-DP
- Max error in a count: $O\left(\frac{\log k}{\varepsilon}\right)$
  - Laplace tail bound: $\Pr\left[|Z_i| \geq \frac{t}{\varepsilon}\right] = e^{-t}$
  - Union bound:$\Pr\left[\max|Z_i| \geq \frac{t}{\varepsilon}\right] \leq ke^{-t}$
  - Set $t = O(\log k)$



Exponential Mechanism

- Privately select an object from a set based on a “score”
- Given: Set of objects Q

  Score function $f: \mathcal{X}^n \times Q \to \mathbf{R}$

  Sensitive dataset $X = X_1, \ldots, X_n$
- Output: $q \in Q$ which (approximately) maximizes $f(X, q)$
- Exponential mechanism: Sample q with probability $\propto \exp(\varepsilon \cdot f(X, q)\ )$
- Theorem: The exponential mechanism is ε-differentially private

Exponential Mechanism Example

- Running an election
  - Set of objects: election candidates
  - Sensitive dataset: votes
  - Score function: number of votes for each candidate
- Non-privately: pick the highest score
- Privately: sample winner $\propto \exp(\varepsilon \cdot \mathrm{Score})$
- Assign scores, use to noisily pick winner

			🏆
	15 votes	18 votes	20 votes
$\varepsilon = 0.1$	25% chance	33.8% chance	41.2% chance

Laplace versus Exponential Mechanism

- Different “types” of mechanisms
  - Noise addition versus sampling
- Useful in different settings
  - Laplace mechanism: computing a statistics/function privately
  - Exponential mechanism: private optimization over a set of objects
- Note: Laplace mechanism is a special case of exponential mechanism
  - Set of objects $Q = \mathbf{R}$
  - Score function of a point q is $-|f(X) - q|$
  - Exponential mechanism samples q w.p. $\propto \exp(-\varepsilon|f(X) - q|)$
  - Equivalent to Laplace mechanism

Lots of other differential privacy tools

- Basic primitives
  - Exponential mechanism
  - Randomized response
  - Sparse vector
- Can use to build more complex procedures
  - Will see examples later

Properties of Differential Privacy

Post-Processing

- You can't "undo" privatization of something which is released differentially privately
- Post-processing theorem: Let $M : \mathcal{X}^n \to \mathcal{Y}$ be ε-differentially private, and $F : \mathcal{Y} \to \mathcal{Z}$ be an arbitrary randomized mapping. Then $F \circ M$ is ε-differentially private

Group Privacy

- Differential privacy quantifies the change in probability of events when a single datapoint is changed. What if k points are changed?
- Privacy loss decays gracefully
- Group privacy theorem: Let $M : \mathcal{X}^n \to \mathcal{Y}$ be ε-differentially private and let X and X' be two datasets which differ in exactly k positions. Then for all events $S \subseteq \mathcal{Y}$,

$$\Pr[M(X) \in S] \leq e^{k\varepsilon} \Pr[M(X') \in S].$$

Composition

- Answering multiple questions about the same sensitive dataset will “intuitively” leak more privacy
    - We know “more things” about the same dataset
- Differential privacy allows us to quantify!
- Basic Composition Theorem: Let $M = (M_1, \ldots, M_k)$ be a sequence of ε-differentially private algorithms. Then M is $k\varepsilon$-differentially private.
- If you release multiple differentially private statistics of a dataset, the privacy parameters “add up”
- The sequence of algorithms can be chosen adaptively
- Composition allows us to build complex procedures out of basic tools

Composition Example

- Consider some population of 100 people
- 20 are smokers
  - Noised with $\mathrm{Laplace}(1)$ noise: 19.51
- 37 have a chronic illness
  - Noised with $\mathrm{Laplace}(1)$ noise: 36.89
- 12 make over $250,000/year
  - Noised with $\mathrm{Laplace}(1)$ noise: 12.53
- Releasing all three noised statistics $(19.51, 36.89, 12.53)$ is 3-DP

Why is ε-differential privacy "too hard"?

- Have to preserve the probability of *every* event
  - Even very very low probability events
- Often, we want to do lots of queries, not just one
  - To get ε-DP when doing k queries, each one has to be ε/k-DP
  - Can we do better?

Consider instead a *relaxation* of ε-DP

Approximate Differential Privacy

Definition: An algorithm M is (ε, δ)-differentially private (DP) if for all datasets X and X' which differ in one entry, and for all events $S \subseteq \mathcal{Y}$,

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

- Generalizes previous notion: setting $\delta = 0$ gives ε-DP
  - Sometimes called “pure” differential privacy
- δ: probability of (potential) “total privacy failure”
  - Needs to be very small
  - Definitely less than $1/n$, but preferably much smaller
  - Compare with ε, which is usually ≈ 1

[Dwork, Kenthapadi, McSherry, Mironov, and Naor], 2006

Properties of Approximate DP

- Post-processing theorem: Let $M : \mathcal{X}^n \to \mathcal{Y}$ be (ε, δ)-differentially private, and $F : \mathcal{Y} \to \mathcal{Z}$ be an arbitrary randomized mapping. Then $F \circ M$ is(ε, δ)-differentially private
- Group privacy theorem: Let $M : \mathcal{X}^n \to \mathcal{Y}$ be (ε, δ)-differentially private and let X and X' be two datasets which differ in exactly k positions. Then for all events $S \subseteq \mathcal{Y}$,

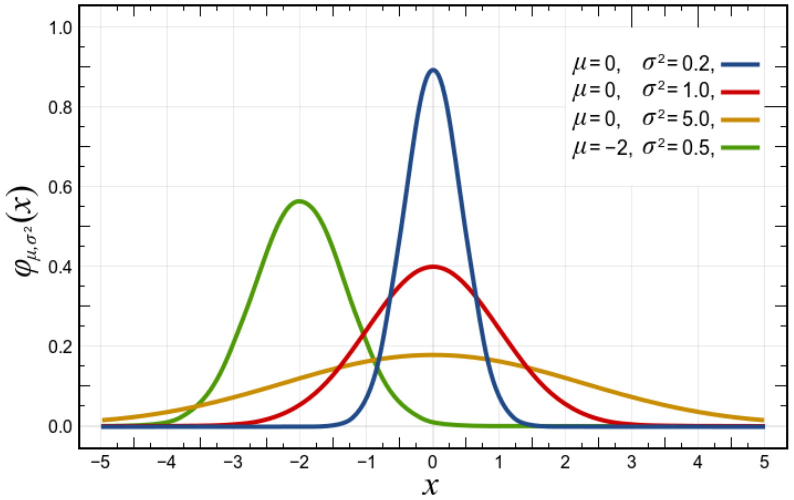
$$\Pr[M(X) \in S] \leq e^{k\varepsilon} \Pr[M(X') \in S] + ke^{(k-1)\varepsilon}\delta.$$

- Basic Composition Theorem: Let $M = (M_1, \ldots, M_k)$ be a sequence of (ε, δ)-differentially private algorithms. Then M is $(k\varepsilon, k\delta)$-differentially private.

Advanced Composition

- Basic Composition Theorem: Let $M = (M_1, \ldots, M_k)$ be a sequence of (ε, δ)-differentially private algorithms. Then M is $(k\varepsilon, k\delta)$-differentially private.
  - If you do k private analyses, you pay k times the privacy cost of one analysis
- Advanced Composition Theorem (informal): Let $M = (M_1, \ldots, M_k)$ be a sequence of (ε, δ)-differentially private algorithms. Then M is $(O(\sqrt{k}\varepsilon), k\delta)$-differentially private.
  - If you do k private analyses, you pay $\sqrt{k}$ times the privacy cost of one analysis
- 10,000 queries (advanced comp) vs 100 queries (basic comp)

The Gaussian Mechanism



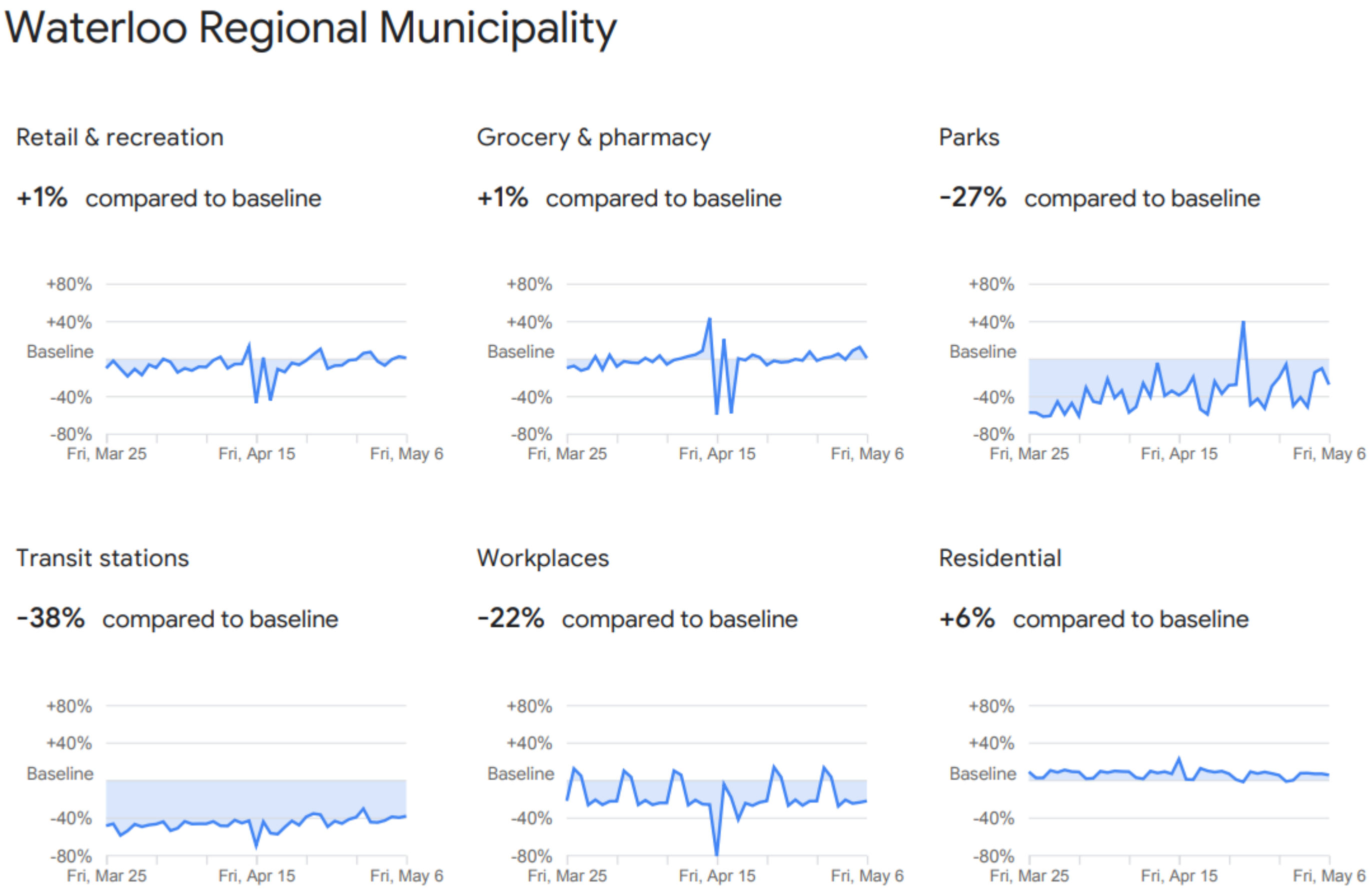
- Let $f(X) : \mathcal{X}^n \to \mathbf{R}^d$ be a function of interest
- Let $\Delta_2^f = \max_{X,X' \text{ differ in one entry}} \|f(X) - f(X')\|_2$ be ℓ_2-sensitivity of f
  - "How much can the function change by modifying one datapoint?"
- Theorem (roughly):

The Gaussian Mechanism $f(X) + N\left(0, \left(\frac{\Delta_2^f \log(1/\delta)}{\varepsilon}\right)^2\right)^{\otimes d}$ is (ε, δ)-DP

  - "Add Gaussian noise to each coordinate, proportional to the ℓ_2-sensitivity"
- Note: since ℓ_2-norm is $\leq \ell_1$-norm, adds less noise vs Laplace mech

What can we do with it…?

Google COVID-19 Community Mobility Reports protected by DP



Waterloo Regional Municipality

Retail & recreation

+1% compared to baseline

Grocery & pharmacy

+1% compared to baseline

Parks

-27% compared to baseline

Transit stations

-38% compared to baseline

Workplaces

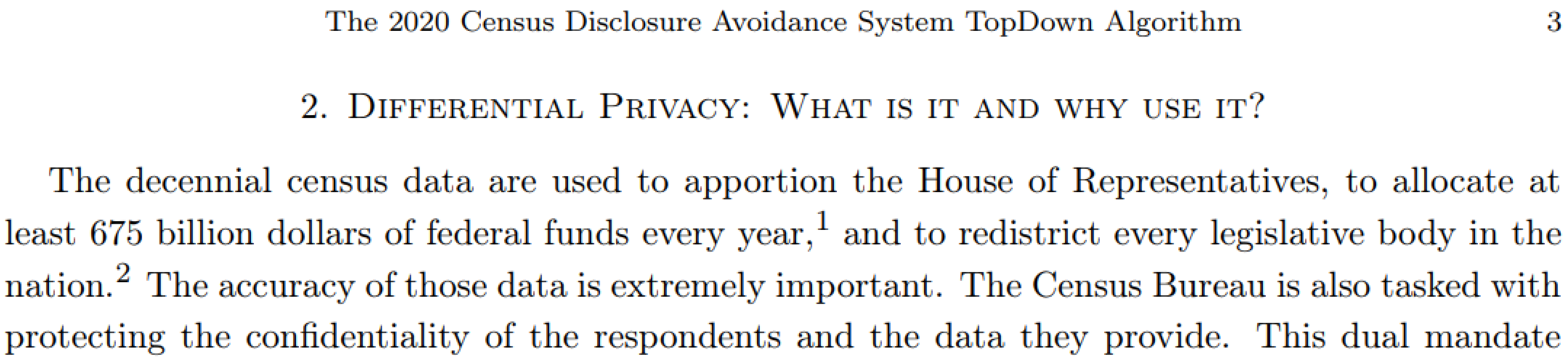
-22% compared to baseline

Residential

+6% compared to baseline

What can we do with it...?

2020 US Census data protected by DP



The 2020 Census Disclosure Avoidance System TopDown Algorithm 3

2. Differential Privacy: What is it and why use it?

The decennial census data are used to apportion the House of Representatives, to allocate at least 675 billion dollars of federal funds every year,[1] and to redistrict every legislative body in the nation.[2] The accuracy of those data is extremely important. The Census Bureau is also tasked with protecting the confidentiality of the respondents and the data they provide. This dual mandate

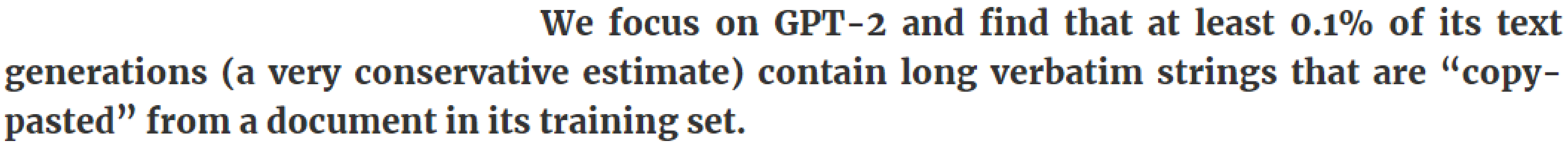
Recap of the basics

- Privacy and non-privacy
- Differential Privacy and its properties
  - Pure and approx. DP, composition, etc.
- Private mechanisms
  - Laplace, Exponential, Gaussian

Private Machine Learning

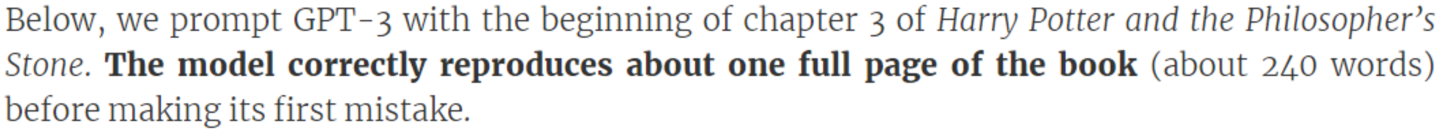
Machine Learning Models are Vulnerable!

- Trained on very large datasets
- Can be coerced to reproduce training data verbatim!

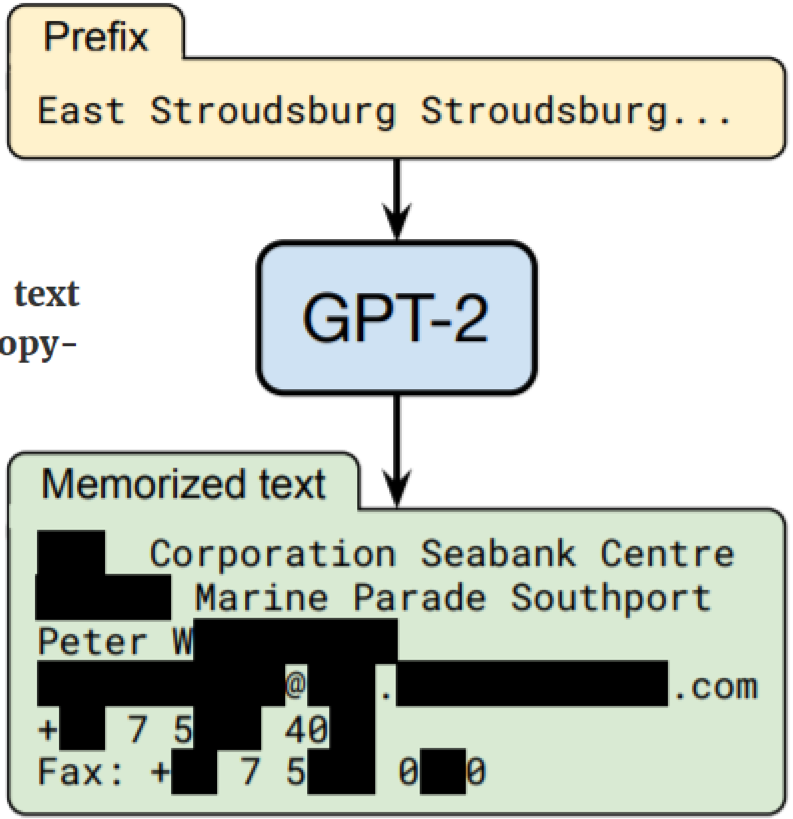


We focus on GPT-2 and find that at least 0.1% of its text generations (a very conservative estimate) contain long verbatim strings that are "copy-pasted" from a document in its training set.

- Personal information, copyrighted content



Below, we prompt GPT-3 with the beginning of chapter 3 of *Harry Potter and the Philosopher's Stone*. **The model correctly reproduces about one full page of the book** (about 240 words) before making its first mistake.



Prefix

East Stroudsburg Stroudsburg...

GPT-2

Memorized text

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

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

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

Private Machine Learning

- Sensitive training data
- Train a machine learning model without leaking too much info
- Can we use the ideas we know about differential privacy?

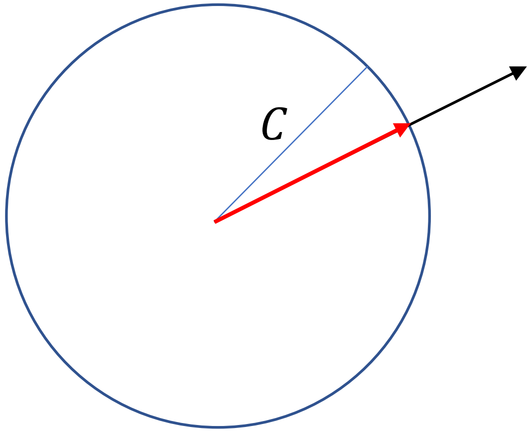
Stochastic Gradient Descent (SGD)

Training a model non-privately: SGD is the go-to algorithm

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 (DPSGD)

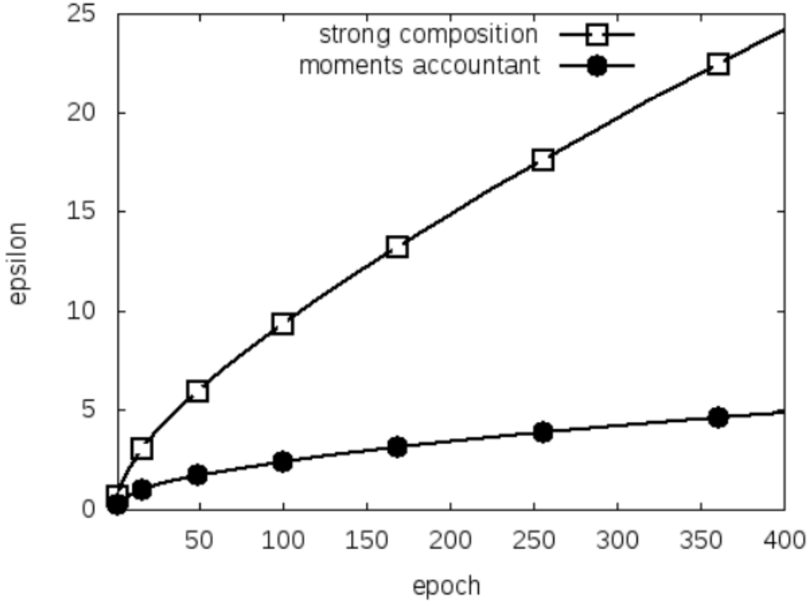
1. Sample a "lot" of points of (expected) size L by selecting each point to be in the lot with probability L/n
2. For each point in the lot, compute the gradient $\nabla\ell(\theta_t, x, y)$ and "clip" it to have ℓ_2 norm at most C
3. Average the clipped gradients and add Gaussian noise
   - Apply the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



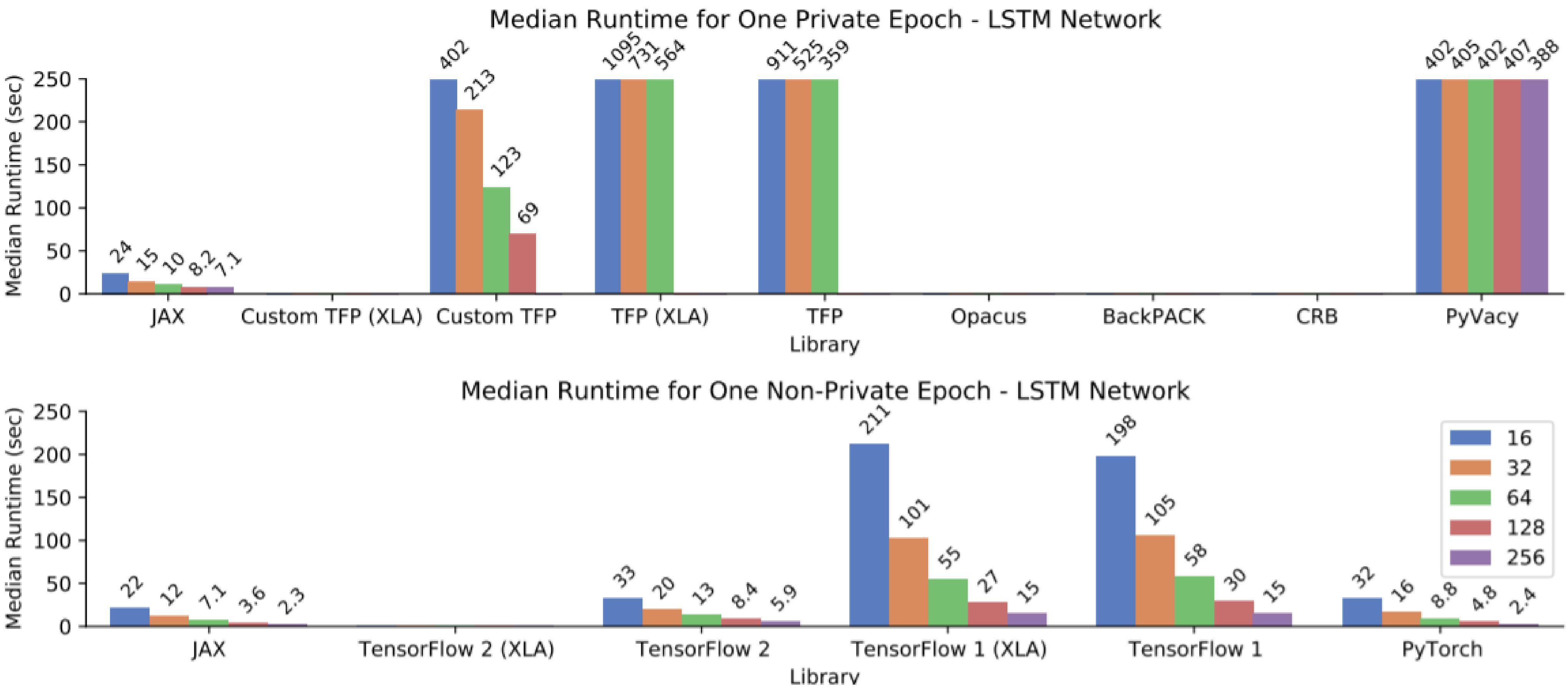
[Song-Chaudhuri-Sarwate '13, Bassily-Smith-Thakurta '14, Abadi-Chu-Goodfellow-McMahan-Mironov-Talwar-Zhang '16]

Analyzing DPSGD

- Suppose one step of DPSGD has privacy with parameter ε
- k steps: use advanced composition, overall $\varepsilon\sqrt{k}$ privacy cost
- Use *amplification by subsampling*!
  - Sample a “lot” of points of (expected) size L by selecting each point to be in the lot with probability L/n
  - Claim: Scales down privacy cost by L/n. Why?
- Overall $\varepsilon\sqrt{k}L/n$ privacy cost
- Better analysis: “Moments accountant”
  - Track moments of privacy loss RV, choose best one

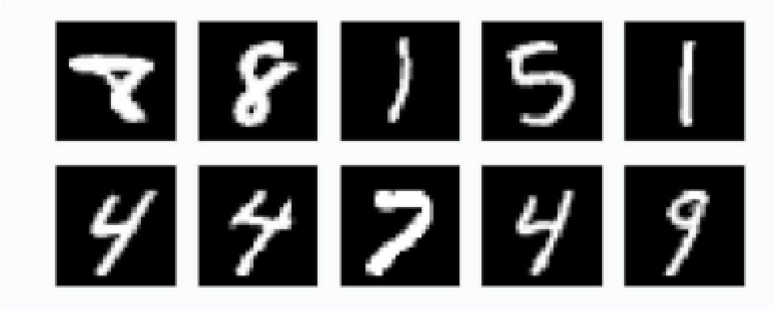


DPSGD can be slow!



[Subramani-Vadivelu-K. ’21]

Does it work?



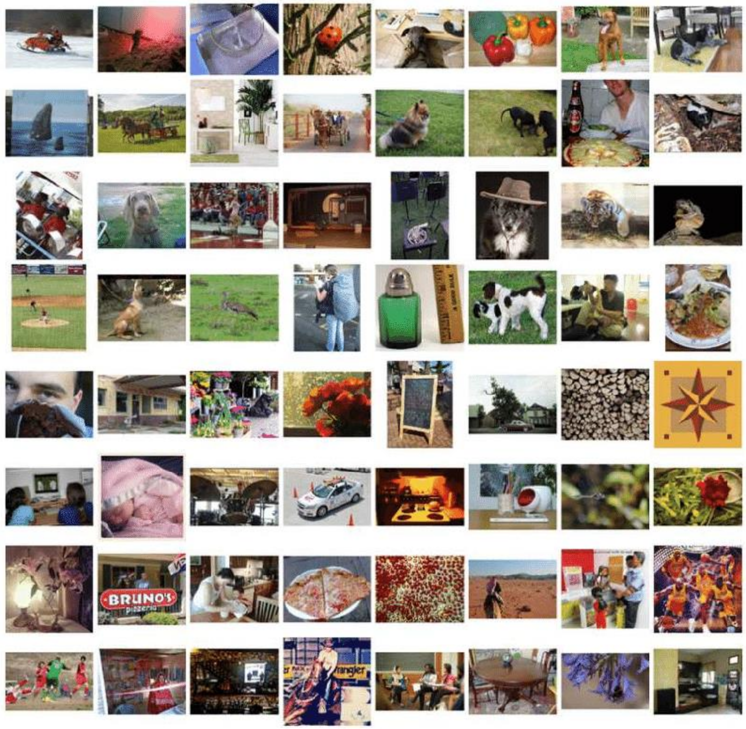
- MNIST: black and white image classification
  - Canonical “easy” ML task
- Non-private test accuracy: $\approx 100\%$
- Private (ε from 1 to 3): 98% - 99%
  - [Tramer-Boneh, ‘21]
- Works pretty well for “easy” datasets!

Does it work?

- CIFAR-10: Low resolution images
  - Same size as MNIST, but harder
- Non-privately: 98%+
- Privately ($\varepsilon = 3$): 69%
  - [Tramer-Boneh, ‘21]
  - Much worse!
- Very recent results: 73.5% for $\varepsilon = 4$ and 82.5% for $\varepsilon = 8$
  - [De-Berrada-Hayes-Smith-Balle, ‘21], [Klause-Ziller-Rueckert-Hammernik-Kaissis, ‘21]
- What is the limit for private learning? Do we *have* to memorize training data?
  - Maybe [Feldman, ‘20], [Feldman-Zhang, ‘20]

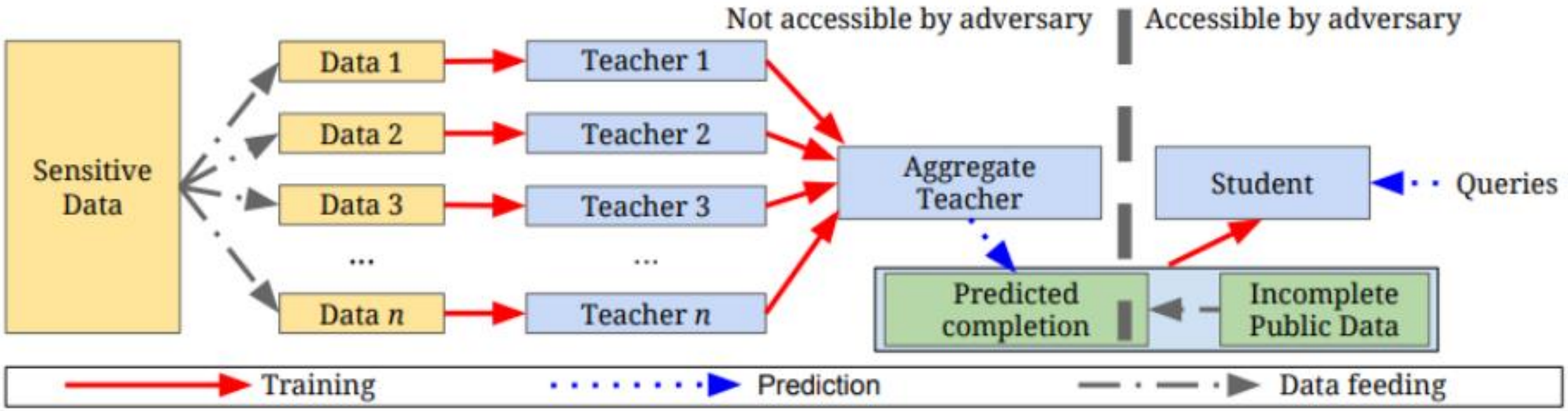
Does it work?

- ImageNet: a hard dataset
  - Millions of higher resolution images, 1000 classes
- Non-privately: $\sim 87\%$
- Privately ($\varepsilon = 8$): 32.4%
  - [De-Berrada-Hayes-Smith-Balle, ‘21]
  - Very tough to achieve (compute, expertise, etc.)
- Not quite there yet…



Using (unlabeled) public data: PATE

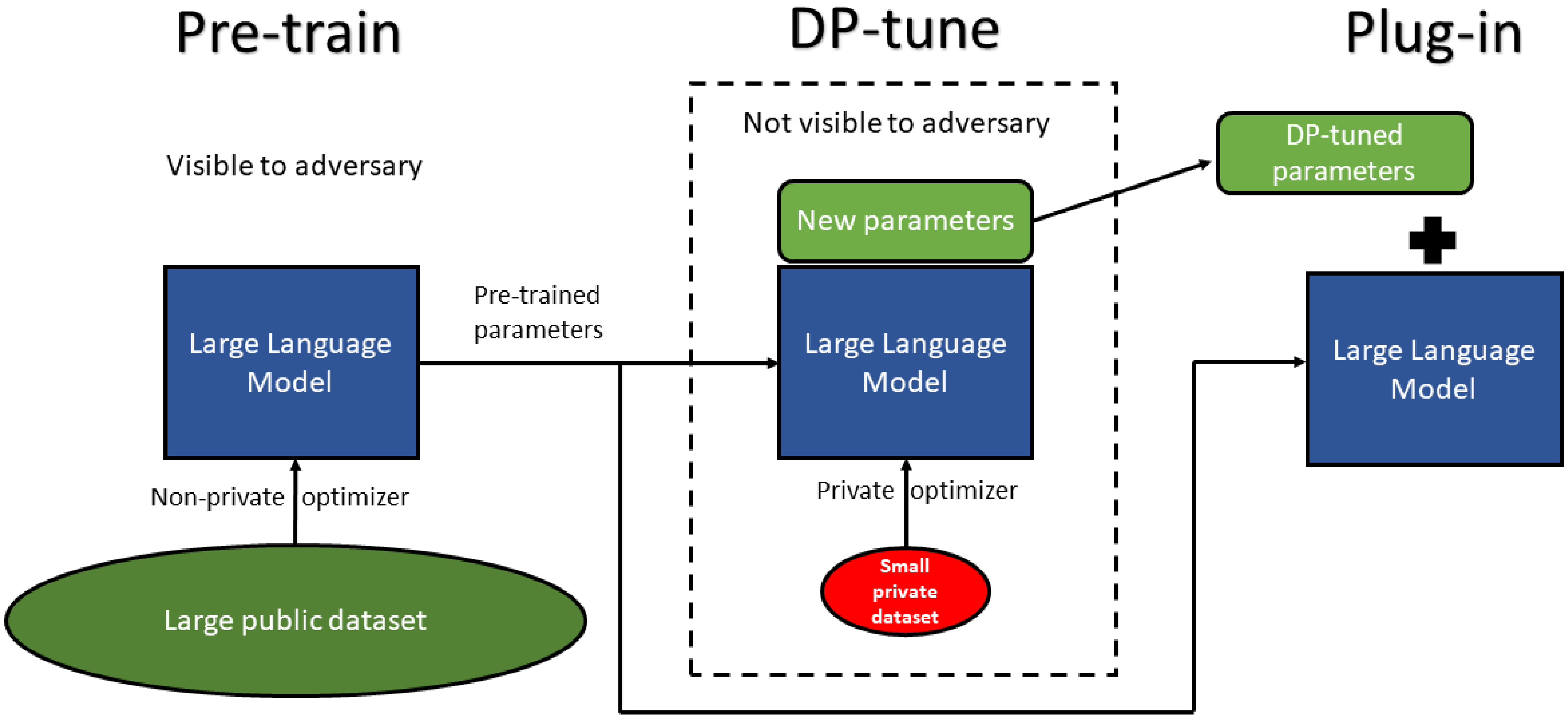
- Train many classifiers non-privately, aggregate predictions privately
  - Sample and aggregate
  - [Papernot-Abadi-Erlingsson-Goodfellow-Talwar, '17]



Using public data

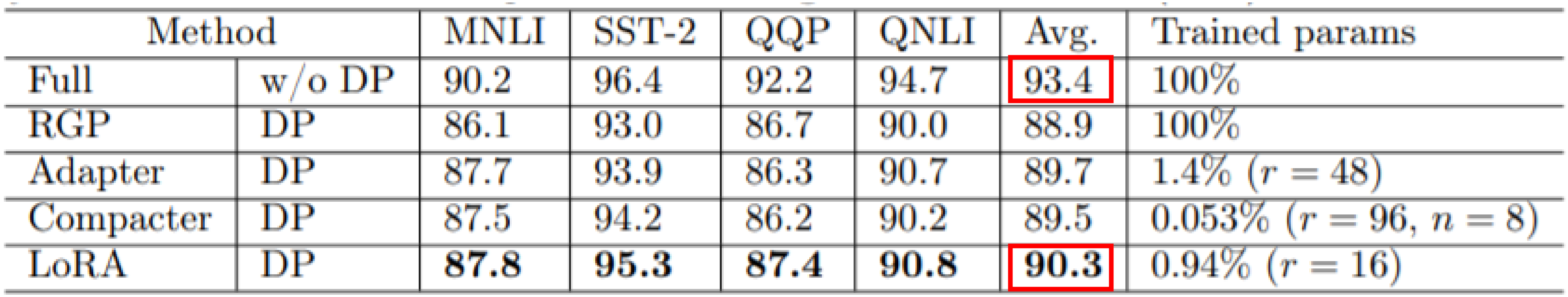
- Pretrain the model with public data
  - Very large amounts of (public) data scraped from the Internet
- Fine-tune model (privately) with sensitive data
  - Smaller and task specific dataset

Using public data



Using public data (text)

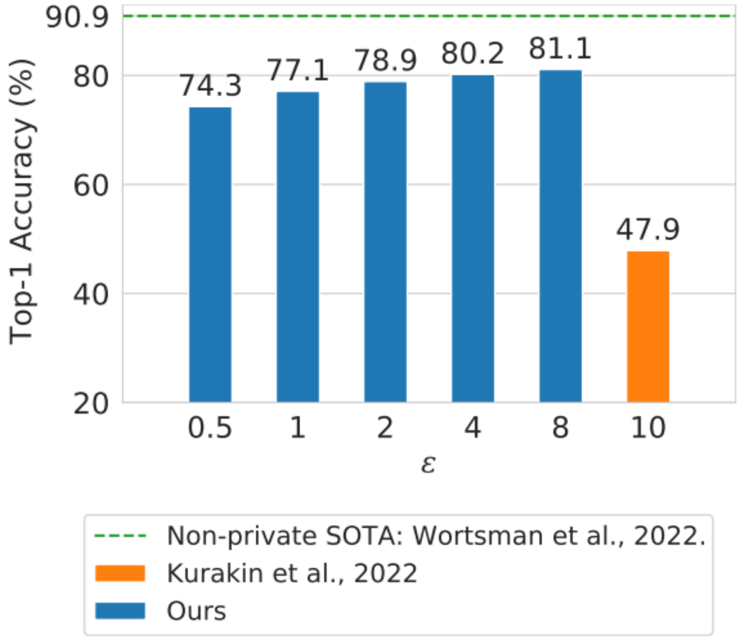
- Can train privately while approaching the non-private accuracy
  - Language model setting
  - [Yu, Naik, Backurs, Gopi, Inan, Kamath, Kulkarni, Lee, Manoel, Wutschitz, Yekhanin, Zhang, ‘22], [Li, Tramer, Liang, Hashimoto, ‘22]



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% ($r = 16$)

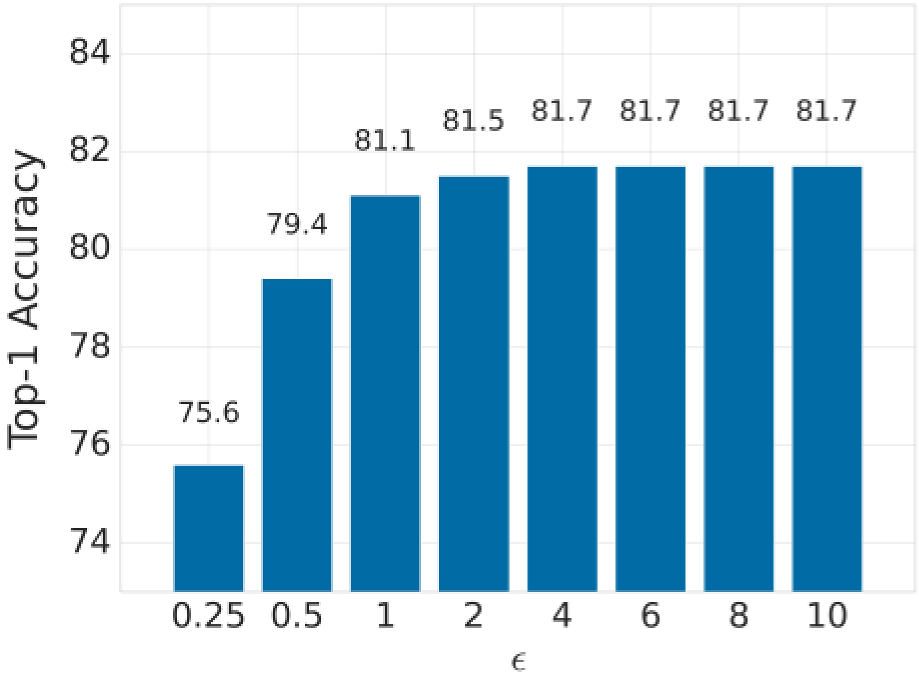
Using public data (ImageNet)

- [De, Berrada, Hayes, Smith, Balle, ‘22]
  - Pretrain with JFT-300M



(b) ImageNet with extra data

- [Mehta, Thakurta, Kurakin, Cutkosky, ‘22]
  - Pretrain with JFT-3B



Conclusion

- Differential privacy is a strong and useful privacy notion
- Can be restrictive, but techniques are getting better
- A lot of useful tools for solving a wide variety of problems