2022 PPML Summer School

Part 3: Privacy-Preserving Machine Learning

Jaewoo Lee
jaewoo.lee@uga.edu

July 13, 2022

Department of Computer Science

UNIVERSITY OF GEORGIA

Data Privacy

What is privacy?

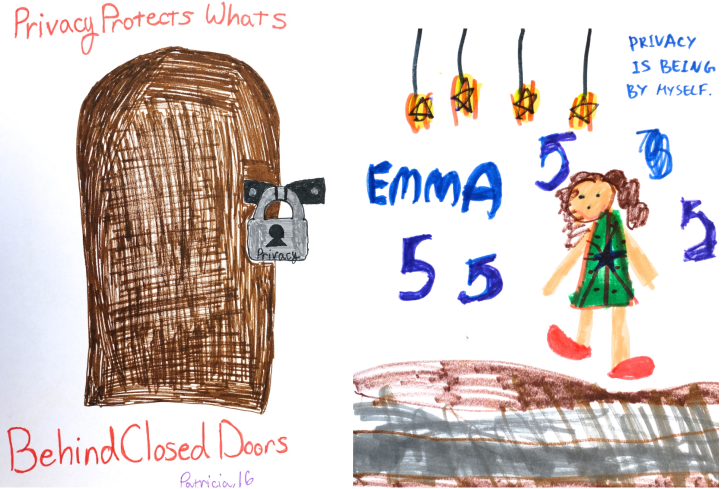


Fig. 1. Image source:https://cups.cs.cmu.edu/privacyillustrated/

What is privacy?

Universal declaration of human rights

> “**Article 12.** No one shall be subjected to arbitrary <u>interference</u> with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks.”

GDPR

> “Personal data are any information which are related to an *identified* or *identifiable* natural person.”

What is privacy?

Expert Determination

- §164.514(b)(1)
- Apply statistical or scientific principles
- Very small risk that anticipated recipient could *identify* individual.

Safe Harbor

- §164.514(b)(2)
- Removal of 18 types of identifiers
- No actual knowledge residual information can *identify* individual

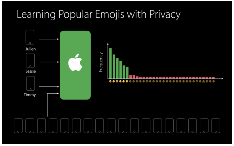
Why privacy?

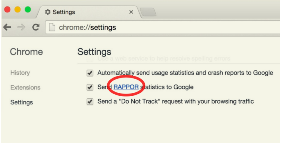
- Massive collection and storage of human activity data
- Personal information is everywhere!
- Any data analysis task that deals with data collected from individuals potentially has privacy issue.

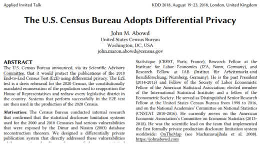
Why privacy?

Practical needs

- Consulting companies needs private tools to analyze their customers' data.
- Apple's iOS 10 uses *differential privacy* to analyze usage data.
- Google chrome web browser also uses *differential privacy* to collect data from users.






Applied Invited Talk

KDD 2018, August 19–23, 2018, London, United Kingdom

The U.S. Census Bureau Adopts Differential Privacy

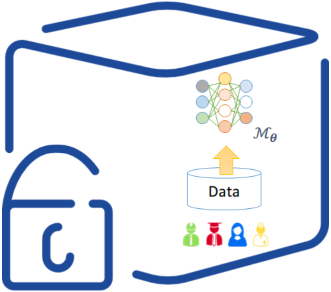
John M. Abowd
United States Census Bureau
Washington, DC, USA
john.maron.abowd@census.gov

ABSTRACT

Privacy-preserving Machine Learning (PPML)

Training an ML model on sensitive data

- Machine learning model $\mathcal{M}_{\boldsymbol{\theta}}$
- Trained on $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
- D often contains *sensitive* info.
- D can be *proprietary*.



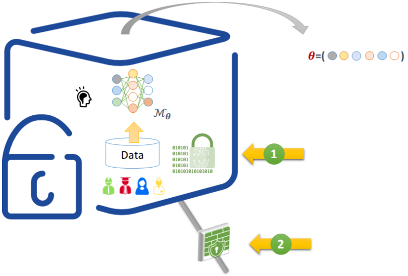
Privacy protection = Nobody sees my data?

What is PPML?

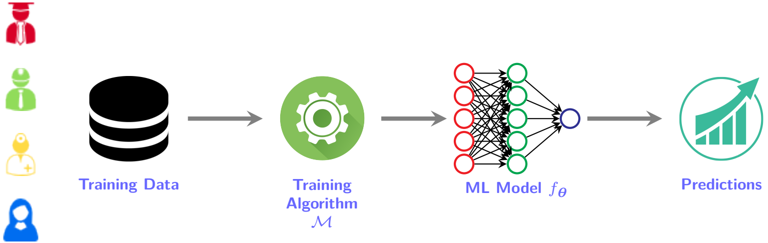
What people think PPML is ...

1. Securing network communication
   - Ensuring no one can hack into our ML system
   - Protect ML systems against network attacks

2. Encrypting databases
   - Dataset is shared using encryption.
   - Allowing full access to people having keys

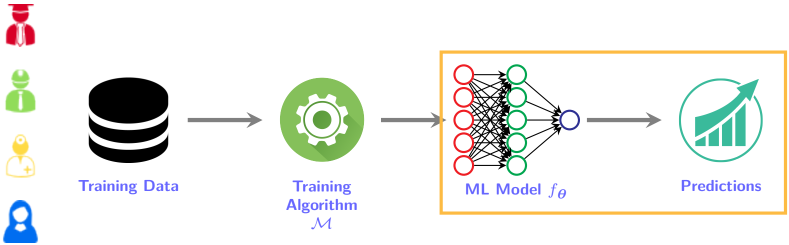


ML Pipeline

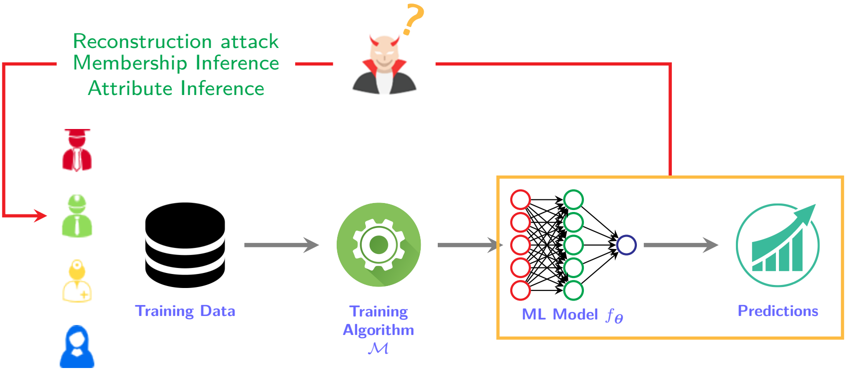


- Training data $D = \{d_1, \ldots, d_n\}$
  - Each d_i corresponds to an *individual*.
  - Training a model on a dataset D results in $\mathcal{M}(D) = \boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \Theta$.

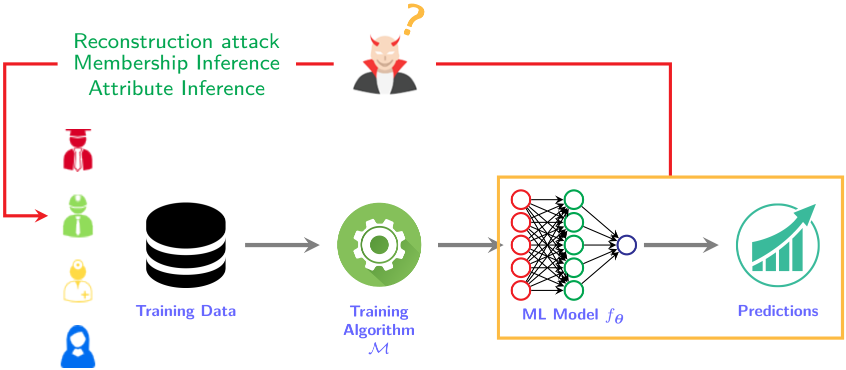
What could go wrong?



What could go wrong?



What could go wrong?



- The released model leak information about D.
  - For example, given $f_{\boldsymbol{\theta}}$, adversaries can infer $\mathbb{P}\left[\quad \middle| \boldsymbol{\theta}\right]$ or $\mathbb{P}\left[\text{income}(\quad) < \$50\text{K} \;\middle|\; \boldsymbol{\theta}\right]$.

Extracting Sensitive Training Data

- Neural networks can reveal your data.
  - Assume black-box access to the GPT-2 model $f_{\boldsymbol{\theta}}$
  - Generate a large set of samples $\mathbf{x} = (x_1, \ldots, x_n)$
  - Evaluate the likelihood

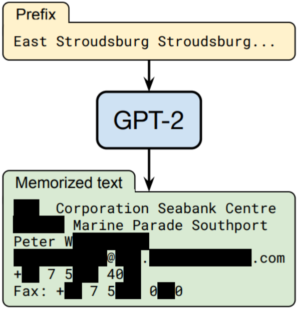
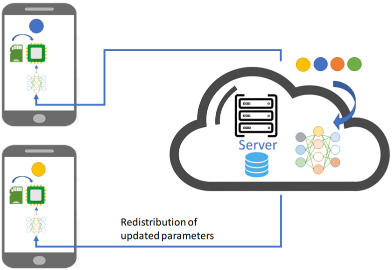


Fig. 3. Carlini et al. 2021

$$\mathcal{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(x_i \mid x_1, \ldots, x_{i-1})\right)$$



- Your data stays *local*!
- Clients only exchange the *gradients* $\nabla\mathcal{L}$.
- But recall that

$$\nabla\mathcal{L}(\boldsymbol{\theta};\mathbf{x}) = \left(\frac{\partial\mathcal{L}}{\partial\theta_1},\ldots,\frac{\partial\mathcal{L}}{\partial\theta_d}\right)\Big|_{\mathbf{x}}$$

Zhu, Ligeng and Liu, Zhijian and Han, Song
Deep Leakage from Gradients
NeurIPS 2019

Privacy in FL

The server computes

$$\overline{\nabla W_t} = \frac{1}{N} \sum_{j=1}^{N} \nabla W_{t,j} \,,$$

$$W_{t+1} = W_t - \eta \overline{\nabla W_t} \,.$$

- $\eta > 0$: step size
- $\nabla W_{t,j}$: gradient received from client j at time t

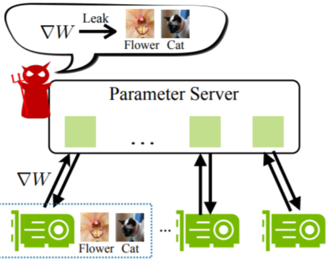


Fig. 4. Federated learning with a central parameter server

Given gradient $\nabla W_{t,k}$ received from client k, is it possible to steal client k's training data $(\mathbf{X}_{t,k}, \mathbf{y}_{t,k})$?

Reconstructing data from gradients

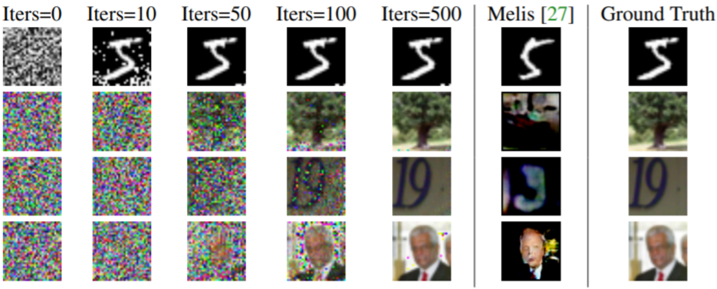


Fig. 5. Reconstructed images from MNIST, CIFAR-100, SVHN, and LFW

Deep Leakage from Gradients
Zhu, Ligeng, Zhijian Liu, and Song Han
In Advances in Neural Information Processing Systems, 2019.

Reconstructing data form gradients

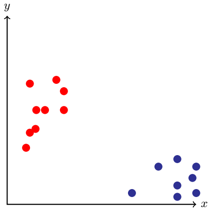
	Example 1	Example 2	Example 3
Initial Sentence	tilting fill given **less word **itude fine **nton overheard living vegas **vac **vation *f forte **dis cerambycidae ellison **don yards marne **kali	toni **enting asbestos cutler km nail **oof **dation **ori righteous **xie lucan **hot **ery at **tle ordered pa **eit smashing proto	[MASK] **ry toppled **wled major relief dive displaced **lice [CLS] us apps _ **face **bet
Iters = 10	tilting fill given **less full solicitor other ligue shrill living vegas rider treatment carry played sculptures lifelong ellison net yards marne **kali	toni **enting asbestos cutter km nail undefeated **dation hole righteous **xie lucan **hot **ery at **tle ordered pa **eit smashing proto	[MASK] **ry toppled identified major relief gin dive displaced **lice doll us apps _ **face space
Iters = 20	registration , volunteer applications , at student travel application open the ; week of played ; child care will be glare .	we welcome proposals for tutor **ials on either core machine denver softly or topics of emerging importance for machine learning .	one **ry toppled hold major ritual ' dive annual conference days 1924 apps novelist dude space
Iters = 30	registration , volunteer applications , and student travel application open the first week of september . child care will be available .	we welcome proposals for tutor **ials on either core machine learning topics or topics of emerging importance for machine learning .	we invite submissions for the thirty - third annual conference on neural information processing systems .
Original Text	Registration, volunteer applications, and student travel application open the first week of September. Child care will be available.	We welcome proposals for tutorials on either core machine learning topics or topics of emerging importance for machine learning.	We invite submissions for the Thirty-Third Annual Conference on Neural Information Processing Systems.

Fig. 6. Reconstructed text data from gradients

ML Output

ML output leaks some information about the individuals in the training data

- SVM: an output can be a subset of training data points.
- Linear regression: an output might be sensitive to an individual's data.

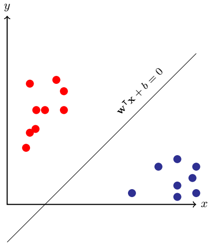


(a) SVM

ML Output

ML output leaks some information about the individuals in the training data

- SVM: an output can be a subset of training data points.
- Linear regression: an output might be sensitive to an individual's data.

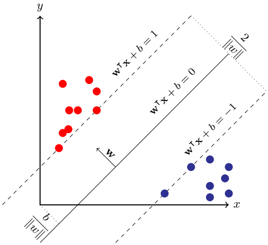


(a) SVM

ML Output

ML output leaks some information about the individuals in the training data

- SVM: an output can be a subset of training data points.
- Linear regression: an output might be sensitive to an individual's data.

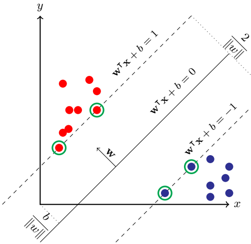


(a) SVM

ML Output

ML output leaks some information about the individuals in the training data

- SVM: an output can be a subset of training data points.
- Linear regression: an output might be sensitive to an individual's data.



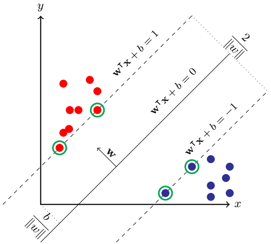
(a) SVM

(b) Linear regression

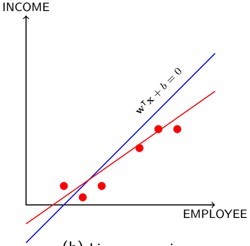
ML Output

ML output leaks some information about the individuals in the training data

- SVM: an output can be a subset of training data points.
- Linear regression: an output might be sensitive to an individual's data.



(a) SVM

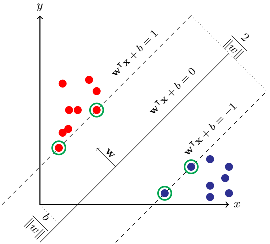


(b) Linear regression

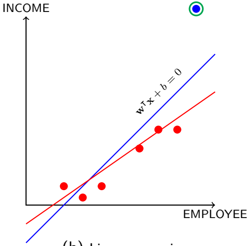
ML Output

ML output leaks some information about the individuals in the training data

- SVM: an output can be a subset of training data points.
- Linear regression: an output might be sensitive to an individual's data.

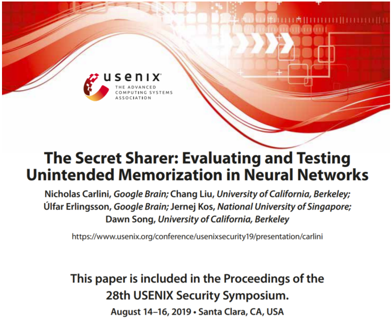


(a) SVM



(b) Linear regression

ML models memorize training examples!

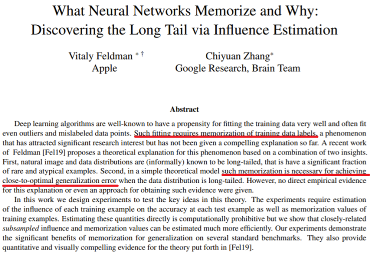

The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

Nicholas Carlini, *Google Brain;* Chang Liu, *University of California, Berkeley;* Úlfar Erlingsson, *Google Brain;* Jernej Kos, *National University of Singapore;* Dawn Song, *University of California, Berkeley*

https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

This paper is included in the Proceedings of the 28th USENIX Security Symposium.

August 14–16, 2019 • Santa Clara, CA, USA


What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation

Vitaly Feldman *† Apple

Chiyuan Zhang* Google Research, Brain Team

Abstract

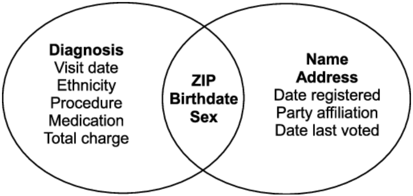
Deep learning algorithms are well-known to have a propensity for fitting the training data very well and often fit even outliers and mislabeled data points. Such fitting requires memorization of training data labels, a phenomenon that has attracted significant research interest but has not been given a compelling explanation so far. A recent work of Feldman [Fel19] proposes a theoretical explanation for this phenomenon based on a combination of two insights. First, natural image and data distributions are (informally) known to be long-tailed, that is have a significant fraction of rare and atypical examples. Second, in a simple theoretical model such memorization is necessary for achieving close-to-optimal generalization error when the data distribution is long-tailed. However, no direct empirical evidence for this explanation or even an approach for obtaining such evidence were given.

In this work we design experiments to test the key ideas in this theory. The experiments require estimation of the influence of each training example on the accuracy at each test example as well as memorization values of training examples. Estimating these quantities directly is computationally prohibitive but we show that closely-related *subsampled* influence and memorization values can be estimated much more efficiently. Our experiments demonstrate the significant benefits of memorization for generalization on several standard benchmarks. They also provide quantitative and visually compelling evidence for the theory put forth in [Fel19].

- Unintended *memorization*
  - Label memorization is necessary for accurate models.
  - Memorization of *irrelevant* training examples is necessary.

Privacy Breach (1)

- Massachusetts Group Insurance Commission
  - collected medical records of government employees
  - considered to be safe since it does not include any identifiers
  - MA voter registration list (available at $20)
  - Governor William Weld's record was identified by Sweeney.
  - How?
    - 54,000 resident in Cambridge, MA
    - 6 people share the same birth date with the Governor
    - only 3 of them are men.
    - only he lived in his zipcode



Released heath record

Voter registration list

Privacy Breach (2)

- Netflix challege (matrix completion)
  - [Narayanan & Shmatikov '08] linked users to IMDB postings.

Name	Movie 1	Movie 2	$\cdots$	Movie 18,000
User1	5		$\cdots$	
User2		3	$\cdots$	
$\vdots$		1	$\ddots$	9
User 48,000			$\cdots$	7

Robust De-anonymization of Large Sparse Datasets, A. Narayanan, V. Shmatikov, 2008

Anonymization isn't enough!

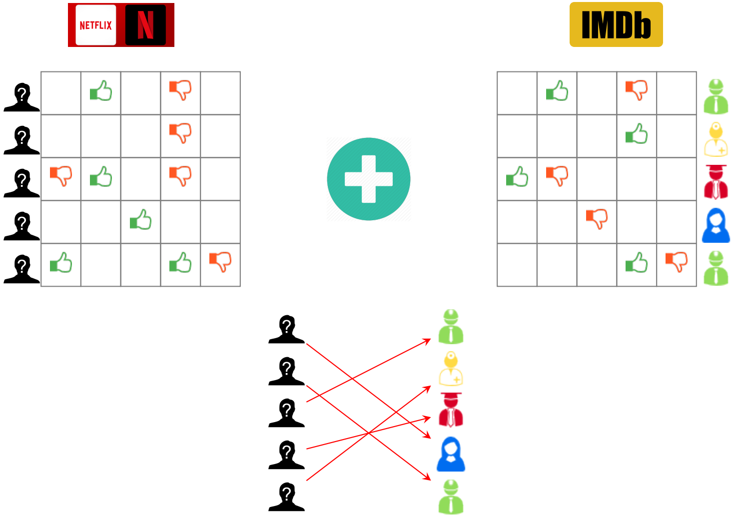


Fig. 8. Netflix Prize

Privacy Breach (3)

- AOL incident
  - AOL dataset: pseudo-user id, search keywords, clicked url, ranking
  - Removed all the identifiers
  - The New York Times identified users and interviewed one of them.
  - Why and how?

AnonID	Query	QueryTime	ItemRank	ClickURL
217	lottery	2006-03-01 11:58:51	1	http://www.calottery.com
217	lottery	2006-03-27 14:10:38	1	http://www.calottery.com
1268	gall stones	2006-05-11 02:12:51		
1268	gallstones	2006-05-11 02:13:02	1	http://www.niddk.nih.gov
1268	ozark horse blankets	2006-03-01 17:39:28	8	http://www.blanketsnmore.com

Semantics of data

Search keyword

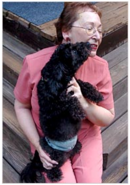
- numb fingers
- 60 single men
- dog that urinates on everything
- landscapers in Lilburn, Ga
- Several people names with last name Arnold
- homes sold in shadow lake subdivision gwinnett county georgia

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



Erik S. Lesser for The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

Data Release

Consider releasing the following table.

Name	Age	Gender	Zip Code	Nationality	Condition
Ann	28	F	13053	Russian	Heart disease
Bruce	29	M	13068	Chinese	Heart disease
Cary	21	F	13068	Japanese	Viral infection
Dick	23	M	13053	American	Viral infection
Eshwar	50	M	14853	Indian	Cancer
Fox	55	M	14750	Japanese	Flu
Gary	47	M	14562	Chinese	Heart disease
Helen	49	F	14821	Korean	Flu
Igor	31	M	13222	American	Cancer
Jean	37	F	13227	American	Cancer
Ken	36	M	13228	American	Cancer
Lewis	35	M	13221	American	Cancer

Question: What could go wrong?

Removing identifiers

- We can *remove* the name attribute from the data.
- Is it now safe to release?

Name	Age	Gender	Zip Code	Nationality	Condition
Ann	28	F	13053	Russian	Heart disease
Bruce	29	M	13068	Chinese	Heart disease
Cary	21	F	13068	Japanese	Viral infection
Dick	23	M	13053	American	Viral infection
Eshwar	50	M	14853	Indian	Cancer
Fox	55	M	14750	Japanese	Flu
Gary	47	M	14562	Chinese	Heart disease
Helen	49	F	14821	Korean	Flu
Igor	31	M	13222	American	Cancer
Jean	37	F	13227	American	Cancer
Ken	36	M	13228	American	Cancer
Lewis	35	M	13221	American	Cancer

Removing identifiers

- Individuals are still *identifiable*.
- How can we hide people's identities?

Name	Age	Gender	Zip Code	Nationality	Condition
Ann	28	F	13053	Russian	Heart disease
Bruce	29	M	13068	Chinese	Heart disease
Cary	21	F	13068	Japanese	Viral infection
Dick	23	M	13053	American	Viral infection
Eshwar	50	M	14853	Indian	Cancer
Fox	55	M	14750	Japanese	Flu
Gary	47	M	14562	Chinese	Heart disease
Helen	49	F	14821	Korean	Flu
Igor	31	M	13222	American	Cancer
Jean	37	F	13227	American	Cancer
Ken	36	M	13228	American	Cancer
Lewis	35	M	13221	American	Cancer

k-anonymity

- Main idea: hide into the *group* of k people
  - make it difficult to link insensitive and sensitive attributes
  - equivalence class: a set of people who share the same combination of insensitive attributes
  - But how?
- Example

Name	Age	Gender	Zip Code	Nationality	Condition
Ann	28	F	13053	Russian	Heart disease
Bruce	29	M	13068	Chinese	Heart disease
Cary	21	F	13068	Japanese	Viral infection
Dick	23	M	13053	American	Viral infection
Eshwar	50	M	14853	Indian	Cancer
Fox	55	M	14750	Japanese	Flu
Gary	47	M	14562	Chinese	Heart disease
Helen	49	F	14821	Korean	Flu
Igor	31	M	13222	American	Cancer
Jean	37	F	13227	American	Cancer
Ken	36	M	13228	American	Cancer
Lewis	35	M	13221	American	Cancer

Data Coarsening

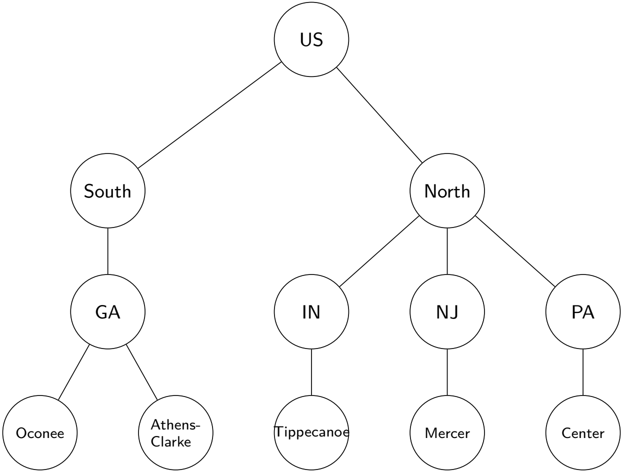
- Coarsen (or suppress) the values into a more *general* ones
  - Suppression: 13228 → 1322* → 132**
  - Range: 21 → [20 - 25] → [20 - 30]
  - Capping: 50 if age $>$ 50

- How about *non-numerical* values?

Name	Age	Gender	Zip Code	Nationality	Condition
Ann	28	F	13053	Russian	Heart disease
Bruce	29	M	13068	Chinese	Heart disease
Cary	21	F	13068	Japanese	Viral infection
Dick	23	M	13053	American	Viral infection
Eshwar	50	M	14853	Indian	Cancer
Fox	55	M	14750	Japanese	Flu
Gary	47	M	14562	Chinese	Heart disease
Helen	49	F	14821	Korean	Flu
Igor	31	M	13222	American	Cancer
Jean	37	F	13227	American	Cancer
Ken	36	M	13228	American	Cancer
Lewis	35	M	13221	American	Cancer

Data Coarsening

- Coarsen (or suppress) the values into a more *general* ones



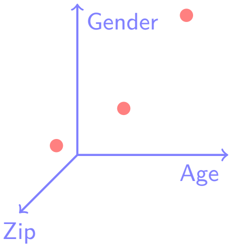
Anonymizing the data

- 4-anonymous table

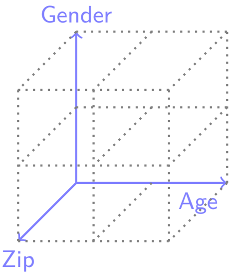
	Age	Gender	Zip Code	Nationality	Condition
(Ann)	20-29	Any	130**	Any	Heart disease
(Bruce)	20-29	Any	130**	Any	Heart disease
(Cary)	20-29	Any	130**	Any	Viral infection
(Dick)	20-29	Any	130**	Any	Viral Infection
(Eshwar)	40-59	Any	14***	Asian	Cancer
(Fox)	40-59	Any	14***	Asian	Flu
(Gary)	40-59	Any	14***	Asian	Heart disease
(Helen)	40-59	Any	14***	Asian	Flu
(Igor)	30-39	Any	1322*	American	Cancer
(Jean)	30-39	Any	1322*	American	Cancer
(Ken)	30-39	Any	1322*	American	Cancer
(Lewis)	30-39	Any	1322*	American	Cancer

- how to anonymize
  - suppress: delete the value
  - generalize: replace the value with more general info.

Geometric Interpretation



(a) Original data



(b) Anonymized data

- Release interval instead of a coordinate value
  - Age 29 $\rightarrow$ [20, 30]
  - Zipcode 30601 $\rightarrow$ 30***
- Linkage attacks become harder

Attacks on k-anonymity

- Homogeneity attack:

	Age	Gender	Zip Code	Nationality	Condition
(Ann)	20-29	Any	130**	Any	Heart disease
(Bruce)	20-29	Any	130**	Any	Heart disease
(Cary)	20-29	Any	130**	Any	Viral infection
(Dick)	20-29	Any	130**	Any	Viral Infection
(Eshwar)	40-59	Any	14***	Asian	Cancer
(Fox)	40-59	Any	14***	Asian	Flu
(Gary)	40-59	Any	14***	Asian	Heart disease
(Helen)	40-59	Any	14***	Asian	Flu
(Igor)	30-39	Any	1322*	American	Cancer
(Jean)	30-39	Any	1322*	American	Cancer
(Ken)	30-39	Any	1322*	American	Cancer
(Lewis)	30-39	Any	1322*	American	Cancer

Attacks on k-anonymity

- Background (knowledge) attack
  - Suppose the adversary knows that Cary is a Japanese. Heart disease occurs at a reduced rate in Japanese patients.

	Age	Gender	Zip Code	Nationality	Condition
(Ann)	20-29	Any	130**	Any	Heart disease
(Bruce)	20-29	Any	130**	Any	Heart disease
(Cary)	20-29	Any	130**	Any	Viral infection
(Dick)	20-29	Any	130**	Any	Viral Infection
(Eshwar)	40-59	Any	14***	Asian	Cancer
(Fox)	40-59	Any	14***	Asian	Flu
(Gary)	40-59	Any	14***	Asian	Heart disease
(Helen)	40-59	Any	14***	Asian	Flu
(Igor)	30-39	Any	1322*	American	Cancer
(Jean)	30-39	Any	1322*	American	Cancer
(Ken)	30-39	Any	1322*	American	Cancer
(Lewis)	30-39	Any	1322*	American	Cancer

Attacks on k-anonymity

- Homogeneity attack
- Background (knowledge) attack

	Age	Gender	Zip Code	Nationality	Condition
(Ann)	20-29	Any	130**	Any	Heart disease
(Bruce)	20-29	Any	130**	Any	Heart disease
(Cary)	20-29	Any	130**	Any	Viral infection
(Dick)	20-29	Any	130**	Any	Viral Infection
(Eshwar)	40-59	Any	14***	Asian	Cancer
(Fox)	40-59	Any	14***	Asian	Flu
(Gary)	40-59	Any	14***	Asian	Heart disease
(Helen)	40-59	Any	14***	Asian	Flu
(Igor)	30-39	Any	1322*	American	Cancer
(Jean)	30-39	Any	1322*	American	Cancer
(Ken)	30-39	Any	1322*	American	Cancer
(Lewis)	30-39	Any	1322*	American	Cancer

ℓ-diversity

- Every equivalence class needs to have at least ℓ "well represented" sensitive values.

Zipcode	Age	Salary	Disease
306**	2*	20K	Gastric Ulcer
306**	2*	30K	Gastritis
306**	2*	40K	Stomach Cancer
3162*	$\geq$40	50K	Gastritis
3162*	$\geq$40	100K	Flu
3162*	$\geq$40	70K	Bronchitis
300**	3*	60K	Bronchitis
300**	3*	80K	Pneumonia
300**	3*	90K	Stomach Cancer

Table 1. A 3-diverse table

ℓ-diversity

Zipcode	Age	Salary	Disease
306**	2*	20K	Gastric Ulcer
306**	2*	30K	Gastritis
306**	2*	40K	Stomach Cancer
3162*	$\geq$40	50K	Gastritis
3162*	$\geq$40	100K	Flu
3162*	$\geq$40	70K	Bronchitis
300**	3*	60K	Bronchitis
300**	3*	80K	Pneumonia
300**	3*	90K	Stomach Cancer

Table 2. A 3-diverse table

- Limitation
  - Similarity attack
    Suppose you know that Bob lives in 30602 and is 27 years old. What can you say about the disease he has?
  - Hard to achieve

Composition Attack

Gender	Age	Zip	Condition
M	[20-30]	306**	Cancer
M	[20-30]	306**	Flu
M	[20-30]	306**	Viral Infection
M	[20-30]	306**	Viral Infection
F	[40-50]	306**	Cancer
F	[40-50]	306**	Heart disease
F	[40-50]	306**	Heart disease
F	[40-50]	306**	Flu
M	[60-]	306**	Cancer
M	[60-]	306**	Cancer
M	[60-]	306**	Cancer
M	[60-]	306**	Flu

(a) St. Mary

Gender	Age	Zip	Condition
M	[20-35]	30***	Cancer
M	[20-35]	30***	Heart disease
M	[20-35]	30***	Malaria
M	[20-35]	30***	Heart disease
M	[20-35]	30***	Tuberculosis
M	[20-35]	30***	Heart disease
F	[20-35]	30***	Flu
F	[20-35]	30***	Flu
F	[20-35]	30***	Flu
F	[20-35]	30***	Tuberculosis
F	[20-35]	30***	Viral infection
F	[20-35]	30***	Cancer

(b) Athens Regional

- Two released datasets satisfying k-anonymity
- Suppose an attacker knows Bob is a Ph.D. student living in Athens.
- Can you guess Bob's medical condition?

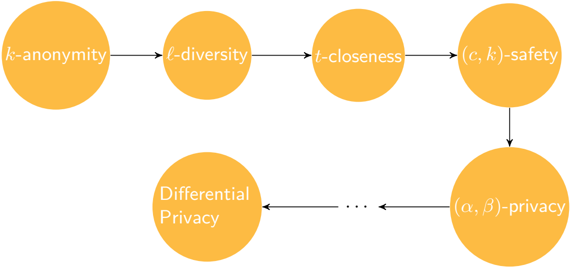
Other Privacy definitions

There exists many other variants

- t-closeness: distribution of sensitive attribute
- (α, β)-privacy: prior and posterior probability
- (c, k)-safety, $\max_{t,s} \mathbb{P}(t \text{ has } s \mid K, D) < c$
- Adversarial model
  - need to make assumptions about adversary's background knowledge
  - how to mathematically specify the adversary's knowledge?

Neverending Battle

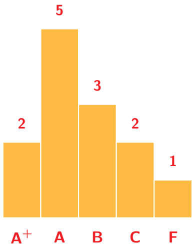
- Syntactic privacy: define how data should look to be private
- Semantic privacy: define what is private



- Is releasing aggregate query result safe?

Name	Grade
Alice	B
Bob	A^+
Charlie	F
...	...
Sam	A
Zach	C

Table 3. Student grades



- The instructor wants to release the grades distribution.
- Suppose the adversary knows the grades of all students but Alice.
- need to hide an individual contribution to the outcome of computation

Differential Privacy

Terminology (1)

- database $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$, a set of individuals
- curator: (trusted) data collector
- query $q : \mathcal{X}^n \rightarrow \mathbb{R}^d$: a function that maps D to a vector in $\mathbb{R}^d$
- privacy mechanism (or algorithm): $\mathcal{M}(D, q, b) = r$

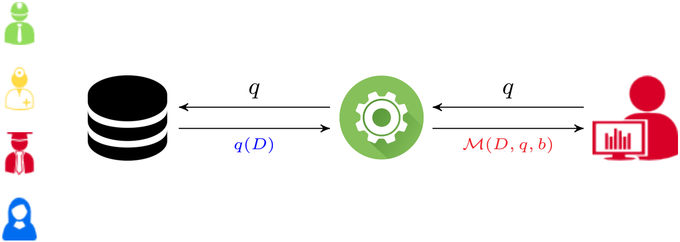


Fig. 11. Interactive setting

Neighboring Datasets

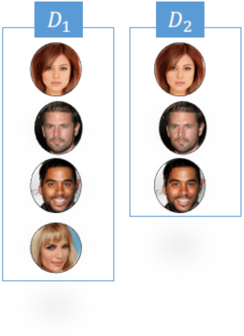


Fig. 12. Unbounded DP

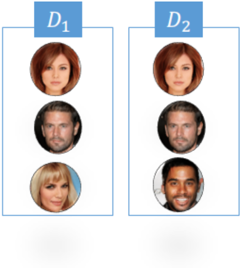


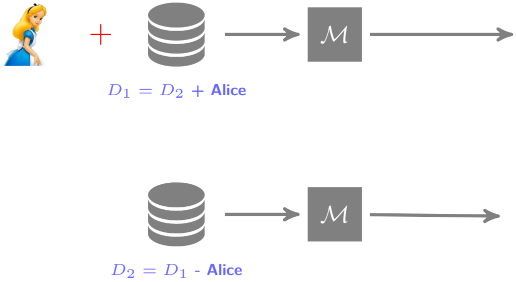
Fig. 13. Bounded DP

- $|D_1| = |D_2| + 1$
- $D_2 \subset D_1$ (proper subset)

- $D_1 = (D_2\{t\}) \cup \{s\}$ (replacement)
- $s, t \in \mathrm{dom}(\mathcal{D})$

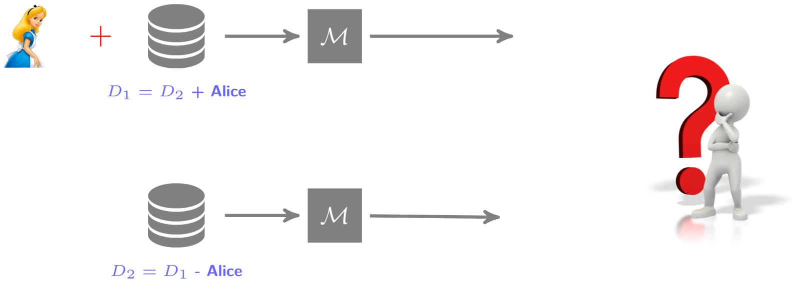
Intuition of differential privacy

Suppose we have two databases D_1 and D_2.



- The mechanism $\mathcal{M}$ chooses i (i is secret).
- It computes and releases $r = \mathcal{M}(D_i)$.
- An adversary observes r.

Intuition of Differential Privacy



- Given $r = \mathcal{M}(D)$, can an adversary tell whether $i = 1$ or $i = 2$?
  - Knowing $i = 1$ reveals the presence of Alice in D.
  - We want to hide the presence/absence of Alice in D.

Intuition of Differential Privacy

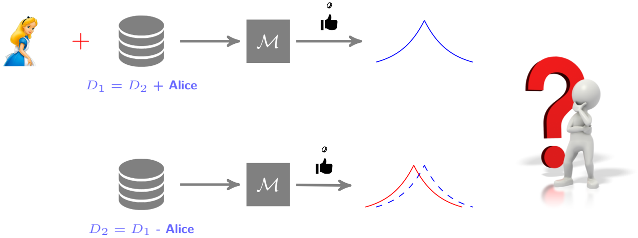


Fig. 14. $\mathcal{M}$ is differentially private.

- How can an adversary distinguish D_1 from D_2?
  - r tells you something about D.
  - $q(D_1) \neq q(D_2)$
  - what happens if $\mathcal{M}$ is deterministic?, i.e.,

$$\mathbb{P}(\mathcal{M}(D_1) = r) \neq 1 \text{ and } \mathbb{P}(\mathcal{M}(D_2) = r) = 0$$

- Make D_1 and D_2 *indistinguishable*
  - Hide the contribution of an individual to $q(D)$

Randomized VS Deterministic

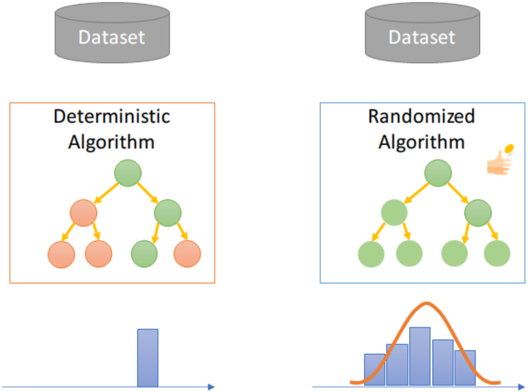


Fig. 15. Randomized VS Deterministic Algorithms

Review: Expectation

Let X be a discrete (continuous) random variable with probability mass (density) function $f_X(x)$.

$$\mathbb{E}[X] = \sum_{x \in \Omega} x f_X(X) \qquad \text{(discrete)}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \qquad \text{(continuous)}$$

Linearity of expectation

Let X and Y be random variables (not necessarily independent) and $a, b \in \mathbb{R}$ are constants. Then we have

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$$

Review: Variance

For a random variable X, its variance is given by

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu)^2\right] \\ &= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mu\,\mathbb{E}[X] + \mu^2 \\ &= \mathbb{E}[X]^2 - \mu^2 = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2\,,\end{aligned}$$

where $\mu = \mathbb{E}[X]$.

- Variance measures dispersion around the mean.
- Variance is not a linear operator.

$$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$$

Definition

Differential Privacy

A randomized algorithm $\mathcal{M}$ is differentially private if for all $\mathcal{S} \subseteq \mathrm{range}(\mathcal{M})$ and for all pairs of neighboring databases D_1 and D_2

$$\frac{\mathbb{P}[\mathcal{M}(D_1) \in \mathcal{S}]}{\mathbb{P}[\mathcal{M}(D_2) \in \mathcal{S}]} \leq \exp(\epsilon)\,,$$

where $\epsilon > 0$ and the probability is taken over the coin flip of $\mathcal{M}$.

Two central concepts

- Neighboring datasets
- Sensitivity

Neighboring databases

We say two databases D_1 and D_2 are *neighboring* if they differ in at most one tuple. I.e., $|(D_1 - D_2) \cup (D_2 - D_1)| = 1$.

Example 1: deterministic

Suppose we have a universe $\mathcal{U} = \{\underbrace{\text{Alice}}_{90}, \underbrace{\text{Bob}}_{80}, \underbrace{\text{Charlie}}_{80}, \underbrace{\text{David}}_{30}\}$.

- $D_1 = \{$Alice, Bob, Charlie$\}$
- $D_2 = \{$Alice, Bob, Charlie, David$\}$
- The school released a statistic $\mathcal{M}(D) = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- Adversary already has all the records of individuals in D_1.
- His task is to guess whether David is in the database D.
- The adversary wins if he guesses correctly.

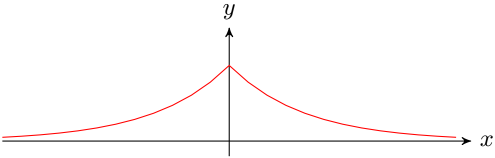
Example 1: deterministic

What happens if the school release the true statistic $\mathcal{M}(D) = 70$?

- Adversary observes the released statistic $\mathcal{M}(D) = 70$.
- Adversary's knowledge
  - Adversary already knows $\mathcal{M}(D_1) = 83.3$.
  - Adversary knows the universe $\mathcal{U} = \{\underbrace{\text{Alice}}_{90}, \underbrace{\text{Bob}}_{80}, \underbrace{\text{Charlie}}_{80}, \text{David}\}$.
- David's score is revealed!

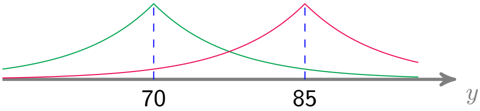
Example 2: randomized

- Recall the school database example
  - $\mathcal{U} = \{\underbrace{\text{Alice}}_{90}, \underbrace{\text{Bob}}_{80}, \underbrace{\text{Charlie}}_{80}, \underbrace{\text{David}}_{30}, \underbrace{\text{Eve}}_{90}\}$
- $D = \{\text{Alice, Bob, Charlie, ?}\}$.
  - $D_1 = \{\text{Alice, Bob, Charlie, David}\} \Longrightarrow \mathcal{M}(D) = 70.$
  - $D_2 = \{\text{Alice, Bob, Charlie, Eve}\} \Longrightarrow \mathcal{M}(D) = 85.$
- Adversary observes $y = \mathcal{M}(D)$, where
  - $\mathbb{P}[\mathcal{M}(D_1) = v] \leq e^{\epsilon}\, \mathbb{P}[\mathcal{M}(D_2) = v]$.
  - $\mathcal{M}(D) = \underbrace{\text{avg}(D)}_{\text{true statistic}} + \underbrace{Y}_{\text{noise}}$
  - Noise distribution



Example 2: randomized

What is adversary's posterior on D_1 and D_2 given $\mathcal{M}(D)$?

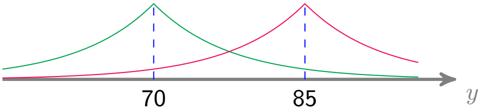


- Noisy answer $y = \mathcal{M}(D)$

$$\mathbb{P}[\text{Guess=David} \mid y] = ?$$

Example 2: randomized

What is adversary's posterior on D_1 and D_2 given $\mathcal{M}(D)$?

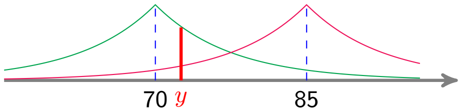


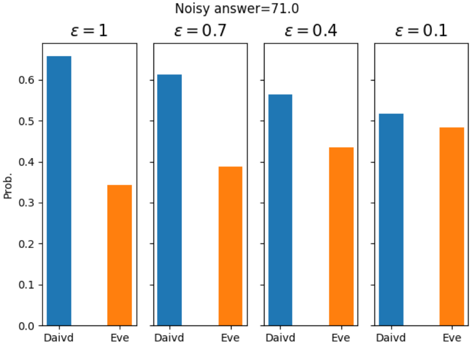
- Noisy answer $y = \mathcal{M}(D)$

$$\mathbb{P}[\text{Guess=David} \mid y] = \frac{\mathbb{P}[y \mid D_2]\,\mathbb{P}[D_2]}{\mathbb{P}[y \mid D_1]\,\mathbb{P}[D_1] + \mathbb{P}[y \mid D_2]\,\mathbb{P}[D_2]}$$

Example 2: posterior

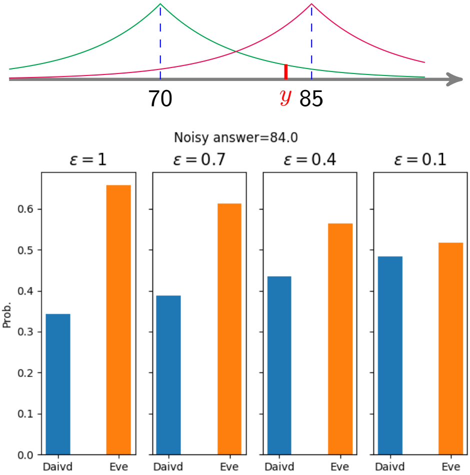
- When the noisy answer=71,





Example 2: posterior

- When the noisy answer=84,



Sensitivity

Why do data analysis results reveal the identities of individuals?

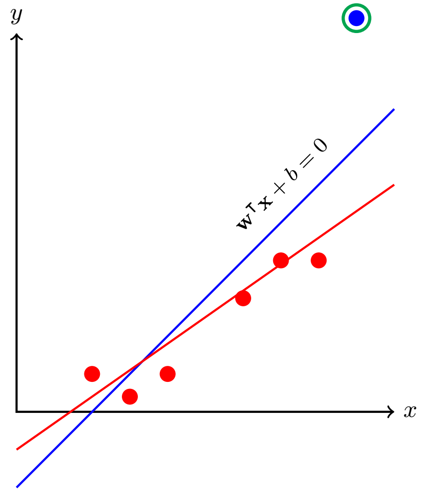


Fig. 16. Linear regression

How to achieve differential privacy?

Sensitivity

- the largest contribution that can be made by one individual
- dependent on the function q of interest and the universe $\mathcal{U}$
- independent of data

The (global) sensitivity of a function $q : \mathcal{X}^n \to \mathbb{R}^d$ is defined by

$$\Delta_q = \max_{D, D' \in \mathcal{U}} \|q(D) - q(D')\|_1 ,$$

where D and D' are neighboring datasets in the universe.

Examples

Setup

- $\mathcal{U} = \{1, 2, 3, \ldots, 100\}$
- $D = \{x_i\}_{i=1}^n \in \mathcal{U}^n$, $x_i \in \mathcal{U}$
- Sensitivity Δ_q for aggregate queries

Practice

- $q(D) = \displaystyle\sum_{i=1}^{n} x_i$
- $q(D) = \displaystyle\frac{1}{n}\sum_{i=1}^{n} x_i$
- $q(D) = \max_i x_i$
- $q(D) = \text{median}(x_1, x_2, \ldots, x_n)$
- $q(D) = \text{count}(x_i = p)$

Laplace mechanism: definition

Laplace Mechanism

Given a query function $q : \mathcal{X}^n \to \mathbb{R}$, the Laplace mechanism is defined as:

$$\mathcal{M}(D) = q(D) + Y,$$

where $Y \sim \text{Lap}\left(\frac{\Delta_q}{\epsilon}\right)$.

- Laplace mechanism satisfies ϵ-differential privacy.

Laplace mechanism: noise distribution

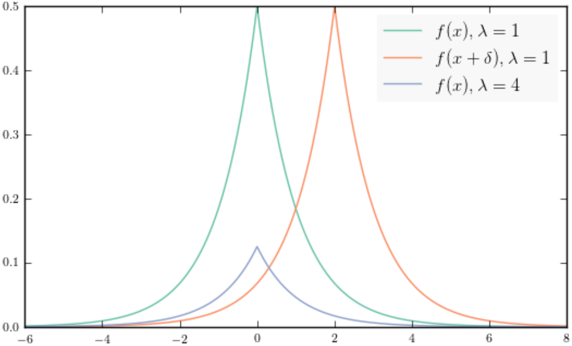


Fig. 17. Laplace distribution

Laplace mechanism: noise distribution

The Laplace mechanism draws random noise $Y \sim \mathsf{Lap}\,(\lambda)$.

$$\mathcal{M}(D) = q(D) + Y$$

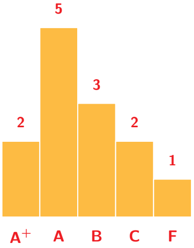
Laplace distribution

- Probability density function $f(x) = \dfrac{1}{2\lambda} \exp\left(-\dfrac{|x-\mu|}{\lambda}\right)$
- mean $\mathbb{E}[Y] = \mu$
- variance $= \mathbb{E}\left[(Y-\mu)^2\right] = 2\lambda^2$
- Sliding property $e^{-\frac{\delta}{\lambda}} \leq \dfrac{f(x+\delta)}{f(x)} \leq e^{\frac{\delta}{\lambda}}$
- for any $t > 0$, $\mathbb{P}[|Y| > t] = \exp\left(-\dfrac{t}{\lambda}\right)$

Example

Name	Grade
Alice	B
Bob	A^+
Charlie	F
...	...
Sam	A
Zach	C

Table 4. Student grades



- sensitivity?
- scale parameter of noise distribution?

Removing noise?

- Consider the Laplace mechanism.

$$r = \mathcal{M}(D) = \underbrace{q(D)}_{\text{true answer}} + \underbrace{Y}_{\text{noise}}, \quad Y \sim \text{Lap}\left(\frac{\Delta_q}{\epsilon}\right)$$

- Given the (noisy) response r, can we reconstruct $q(D)$?

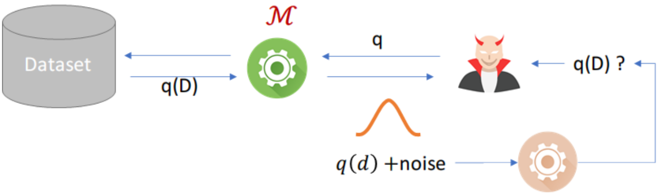


Fig. 18. Is it possible to remove noise added by the privacy mechanism?

Post-processing

- Let $\mathcal{M} : \mathcal{X}^n \to R$ be an ϵ-DP algorithm.
- $\mathcal{M}(D)$ is the *private* output.
- Suppose we have a deterministic function $f : R \to R'$.
- If we apply f on the private output, is it still private?

Post-processing Invariance

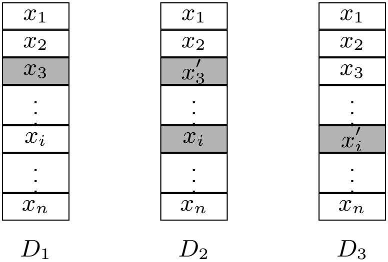
Let $\mathcal{M}$ be an ϵ-DP function and f be an arbitrary deterministic function on the output domain of $\mathcal{M}$. The composite function $f \circ g : \mathcal{X}^n \to R'$ is ϵ-differentially private.

- It means that you cannot make $\mathcal{M}(D)$ more or less private.

Group Privacy

Let $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}$ be an ϵ-differentially private algorithm. Then, $\mathcal{M}$ is $k\epsilon$-differentially private for groups of size k. That is, for all x, y such that $\|x - y\|_1 \leq k$ and for all $S \subseteq \mathrm{range}(\mathcal{M})$,

$$\mathbb{P}[\mathcal{M}(x) \in S] \leq \exp(k\epsilon)\,\mathbb{P}[\mathcal{M}(y) \in S]\,.$$



Composition (1)

Sequential composition

- Suppose we have two algorithms $\mathcal{M}_1$ and $\mathcal{M}_2$.
- $\mathcal{M}_1$ is ϵ_1-DP and $\mathcal{M}_2$ is ϵ_2-DP.
- The algorithm $\mathcal{M}$ that sequentially calls $\mathcal{M}_1$ and $\mathcal{M}_2$ is $(\epsilon_1 + \epsilon_2)$-differentially private.

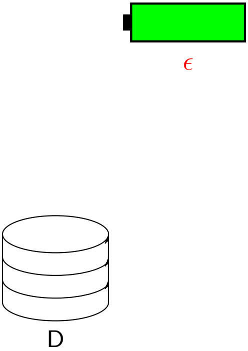
Proof.

$$\begin{aligned}\frac{\mathbb{P}[\mathcal{M}(D) = (r_1, r_2)]}{\mathbb{P}[\mathcal{M}(D') = (r_1, r_2)]} &= \frac{\mathbb{P}[(\mathcal{M}_1(D) = r_1, \mathcal{M}_2(D) = r_2)]}{\mathbb{P}[(\mathcal{M}_1(D') = r_1, \mathcal{M}_2(D') = r_2)]} \\ &= \frac{\mathbb{P}[\mathcal{M}_1(D) = r_1]}{\mathbb{P}[\mathcal{M}_1(D') = r_1]} \frac{\mathbb{P}[\mathcal{M}_2(D) = r_2]}{\mathbb{P}[\mathcal{M}_2(D') = r_2]} \\ &\le \exp(\epsilon_1) \cdot \exp(\epsilon_2) = \exp(\epsilon_1 + \epsilon_2)\end{aligned}$$

$\square$

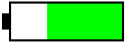
Privacy Budget

- We normally answer multiple queries.



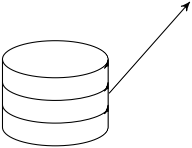
Privacy Budget

- We normally answer multiple queries.



$\epsilon - \epsilon_1$

$$q_1 \longrightarrow q_1(D) + \text{Lap}\left(1/\epsilon_1\right)$$

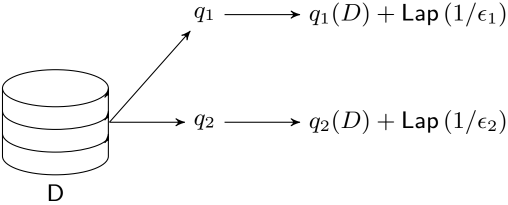


D

Privacy Budget

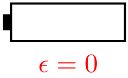
- We normally answer multiple queries.

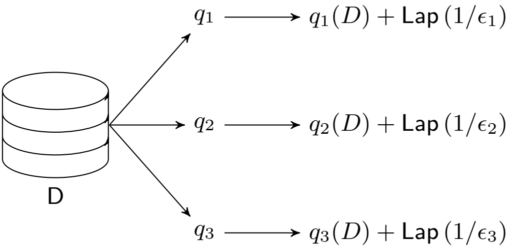
$\epsilon - \epsilon_1 - \epsilon_2$



Privacy Budget

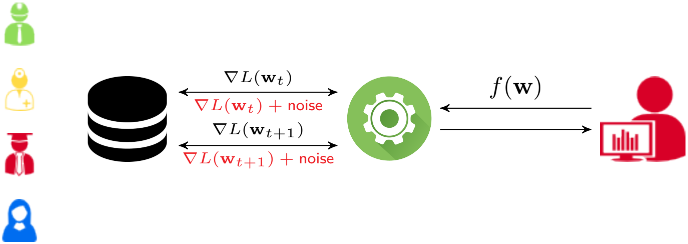
- We normally answer multiple queries.





Deep Learning with Differential Privacy

Gradient Perturbation



- Perturb the gradients

$$\widetilde{\nabla} L(\mathbf{w}_t) = \nabla L(\mathbf{w}_t) + \mathcal{N}\left(0, \sigma_t^2 \mathbf{I}_d\right) \qquad \text{(noisy gradient)}$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \underbrace{\eta_t}_{\text{step size}} \widetilde{\nabla} L(\mathbf{w}_t) \qquad \text{(GD update)}$$

- Need to carefully control η_t and σ_t

Differentially Private Deep Learning

DP-SGD Framework: gradient clipping + noise injection

Let $B = \{$ , , ,  $\}$ be a mini-batch.

- *Per-example* Gradient

$\nabla \ell(\mathbf{w}_t,$  $)$

$\nabla \ell(\mathbf{w}_t,$ $)$

$\nabla \ell(\mathbf{w}_t,$  $)$

$+\, \nabla \ell(\mathbf{w}_t,$  $)$

$$\nabla L(\mathbf{w}_t; B) = \sum_{i=1}^{4} \nabla \ell(\mathbf{w}_t, d_i) + \text{noise}$$

DP-SGD

- Need to bound the *influence* of each individual on the gradient, meaning that, for some $C > 0$,

$\|\nabla\ell(\mathbf{w}_t,$  $)\|_2 \leq C$

$\|\nabla\ell(\mathbf{w}_t,$ $)\|_2, \leq C$

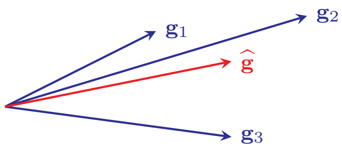
$\|\nabla\ell(\mathbf{w}_t,$  $)\|_2, \leq C$

$\|\nabla\ell(\mathbf{w}_t,$  $)\|_2, \leq C\,.$

  - C is called *clipping threshold*.
  - The sensitivity of $\nabla\ell(\mathbf{w}_t) = C$.

SGD VS DP-SGD

Non-private



- Per-example gradient: $\mathbf{g}_i = \nabla L(\mathbf{w}^t, d_i)$ for $i = 1, 2, 3$
- Aggregated gradient: $\widehat{\mathbf{g}} = \frac{1}{3}(\mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3)$

SGD VS DP-SGD

Private

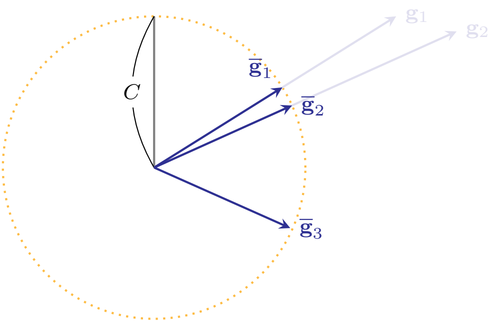


Fig. 19. Effect of gradient clipping

SGD VS DP-SGD

Private

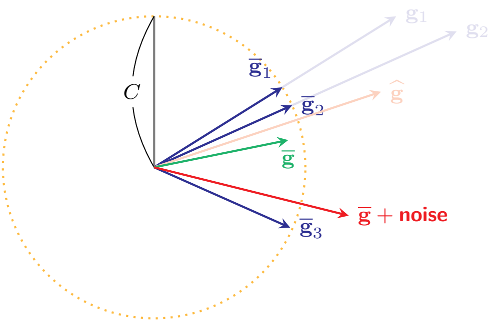


Fig. 20. Effect of gradient clipping + Noise

- Private gradient: $\widetilde{\mathbf{g}} = \overline{\mathbf{g}} + \text{noise}$
  - bias due to clipping
  - variance due to noise addition

What is OPACUS?

- A PyTorch library for differentially private training of NNs
- Support fast *per-example* gradient computation
- `https://opacus.ai/`

Training your NN under DP

For details, please refer to this page.

```
import warnings
warnings.simplefilter("ignore")

MAX_GRAD_NORM = 1.2
EPSILON = 50.0
DELTA = 1e-5
EPOCHS = 20

LR = 1e-3
```

Preparing datasets

```python
import torch
import torchvision
import torchvision.transforms as transforms

# These values, specific to the CIFAR10 dataset, are assumed to be known.
# If necessary, they can be computed with modest privacy budget.
CIFAR10_MEAN = (0.4914, 0.4822, 0.4465)
CIFAR10_STD_DEV = (0.2023, 0.1994, 0.2010)

transform = transforms.Compose([
    transforms.ToTensor(),
    transforms.Normalize(CIFAR10_MEAN, CIFAR10_STD_DEV),
])
```

Preparing datasets

```
from torchvision.datasets import CIFAR10

DATA_ROOT = '../cifar10'

train_dataset = CIFAR10(
    root=DATA_ROOT, train=True, download=True, transform=transform)

train_loader = torch.utils.data.DataLoader(
    train_dataset,
    batch_size=BATCH_SIZE,
)

test_dataset = CIFAR10(
    root=DATA_ROOT, train=False, download=True, transform=transform)

test_loader = torch.utils.data.DataLoader(
    test_dataset,
    batch_size=BATCH_SIZE,
    shuffle=False,
)
```

Validating Models

```
from torchvision import models
from opacus.validators import ModuleValidator

model = models.resnet18(num_classes=10)  # loading a built-in model
errors = ModuleValidator.validate(model, strict=False)
errors[-5:]   # print error messages
```

- Verify whether the model is compatible with DP training
  - BatchNorm cannot be used.
  - Replace it with GroupNorm.

Preparing for training

```python
from opacus import PrivacyEngine

privacy_engine = PrivacyEngine()

model, optimizer, train_loader = privacy_engine.make_private_with_epsilon(
    module=model,
    optimizer=optimizer,
    data_loader=train_loader,
    epochs=EPOCHS,
    target_epsilon=EPSILON,
    target_delta=DELTA,
    max_grad_norm=MAX_GRAD_NORM,
)

print(f"Using sigma={optimizer.noise_multiplier} and C={MAX_GRAD_NORM}")
```

Private Training

```python
def train(model, train_loader, optimizer, epoch, device):
    criterion = nn.CrossEntropyLoss()
    losses, top1_acc = [], []

    with BatchMemoryManager(
        data_loader=train_loader,
        max_physical_batch_size=MAX_PHYSICAL_BATCH_SIZE,
        optimizer=optimizer
    ) as memory_safe_data_loader:

        for i, (images, target) in enumerate(memory_safe_data_loader):
            optimizer.zero_grad()
            images = images.to(device)
            target = target.to(device)

            output = model(images)      # compute output
            loss = criterion(output, target)

            preds = np.argmax(output.detach().cpu().numpy(), axis=1)
            labels = target.detach().cpu().numpy()

            acc = accuracy(preds, labels)  # measure accuracy and record loss
            losses.append(loss.item())
            top1_acc.append(acc)

            loss.backward()
            optimizer.step()

            epsilon = privacy_engine.get_epsilon(DELTA)
```