DDA4210/AIR6002 Advanced Machine Learning
Lecture 10 Privacy in Machine Learning

Tongxin Li

School of Data Science, CUHK-Shenzhen

Spring 2024

Motivation

Secure and Private AI

The importance of data for ML

> “The biggest obstacle to using advanced data analysis isn’t skill base or technology; it’s plain old access to the data.”

-Edd Wilder-James, Harvard Business Review

The importance of data for ML

“Data is the New Oil”

Data privacy matters!

What is privacy?

Privacy		Don’t tell
	?	
Confidentiality		Don’t ask

The importance of data for ML

Privacy Don’t tell

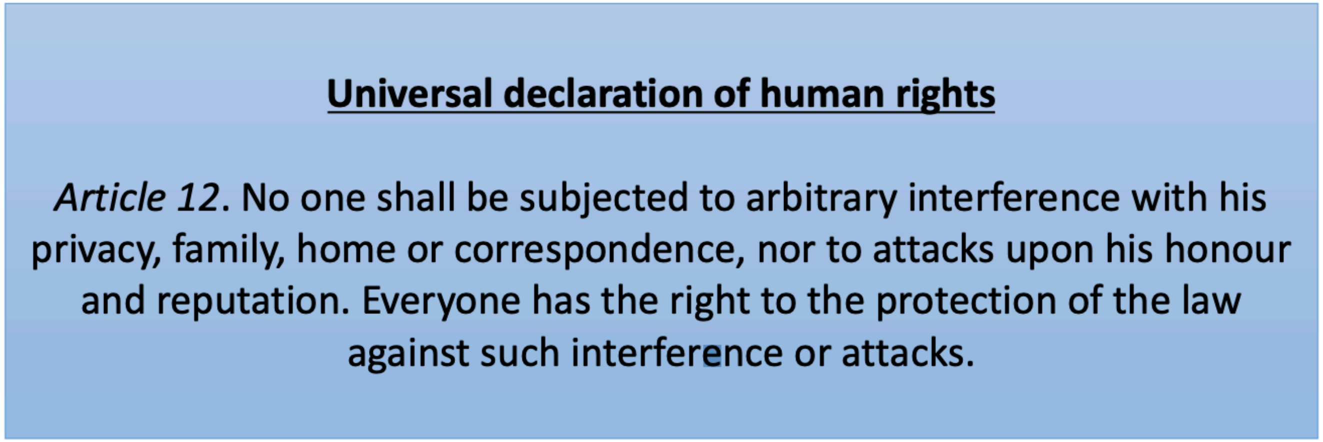
?

Confidentiality Don’t ask

Privacy is about the right to be left alone (from public scrutiny)

Confidentiality is about a promise from people who have privileged access.

Importance of privacy by the United Nations



Universal declaration of human rights

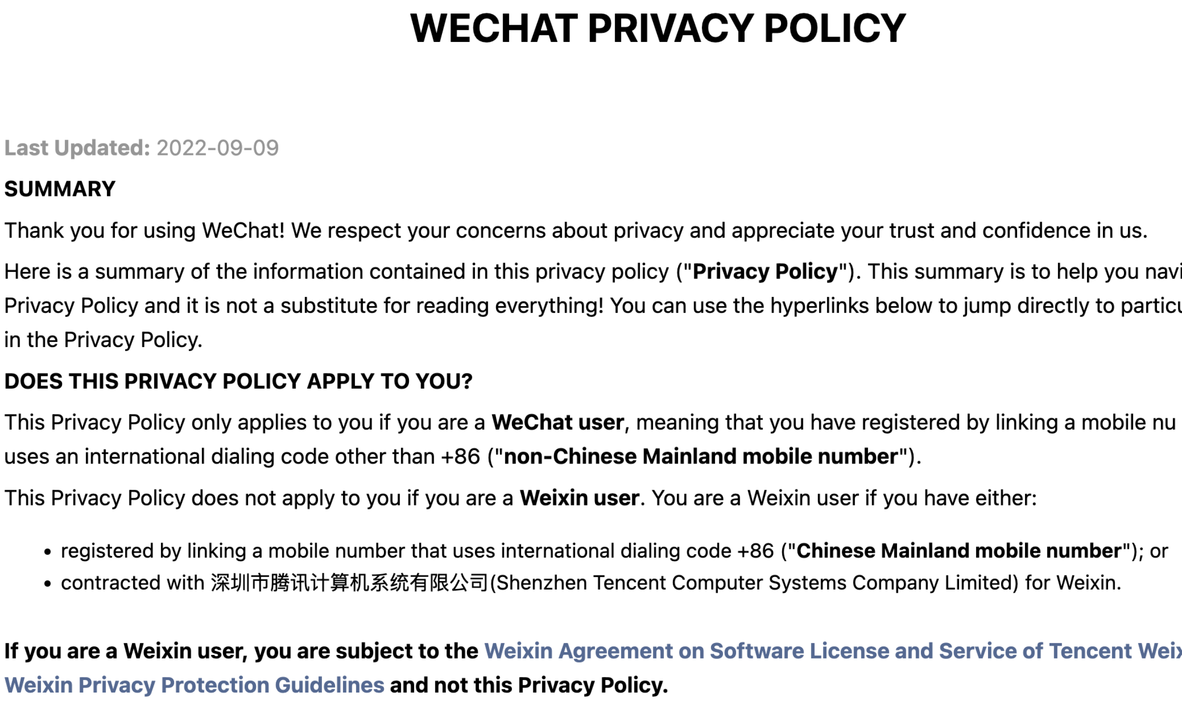
Article 12. No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

https://www.un.org/en/about-us/universal-declaration-of-human-rights

Personal data in Big Data era

- Social networks: Facebook, LinkedIn

- Government, company, research centers collect personal information and analyze them



WECHAT PRIVACY POLICY

Last Updated: 2022-09-09

SUMMARY

Thank you for using WeChat! We respect your concerns about privacy and appreciate your trust and confidence in us.

Here is a summary of the information contained in this privacy policy ("**Privacy Policy**"). This summary is to help you navigate the Privacy Policy and it is not a substitute for reading everything! You can use the hyperlinks below to jump directly to particular sections in the Privacy Policy.

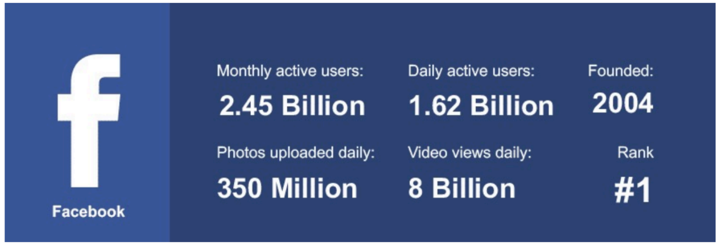
DOES THIS PRIVACY POLICY APPLY TO YOU?

This Privacy Policy only applies to you if you are a **WeChat user**, meaning that you have registered by linking a mobile nu uses an international dialing code other than +86 ("**non-Chinese Mainland mobile number**").

This Privacy Policy does not apply to you if you are a **Weixin user**. You are a Weixin user if you have either:

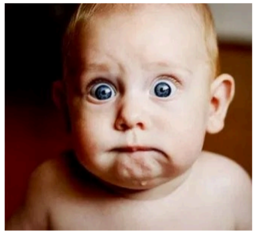
- registered by linking a mobile number that uses international dialing code +86 ("**Chinese Mainland mobile number**"); or
- contracted with 深圳市腾讯计算机系统有限公司(Shenzhen Tencent Computer Systems Company Limited) for Weixin.

If you are a Weixin user, you are subject to the Weixin Agreement on Software License and Service of Tencent Weixin and Weixin Privacy Protection Guidelines and not this Privacy Policy.



Source: https://www.garyfox.co/social-media-statistics/

Recent legislations on privacy forces businesses to revise their data practice



- can't keep personal data for more than three weeks?
- I will have to delete all traces of a user upon request?

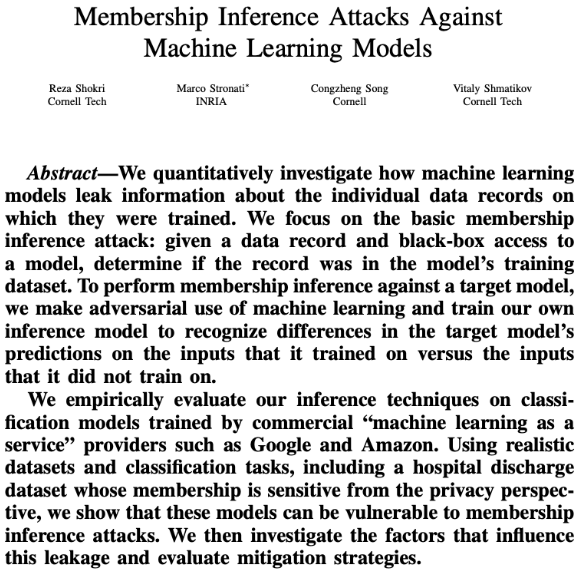
How about **machine learning models** trained on user data?

Risk of personal information leakage

- YouTube & Amazon use viewing/buying records for recommendations.
- Emails in Gmail are used for targeted Ads and for completing your sentence.
- LLMs use public and third-party data for training

Risk of personal information leakage

ML models memorize training datasets, even though they are generalizing well!



Membership Inference Attacks Against Machine Learning Models

Reza Shokri
Cornell Tech

Marco Stronati[*]
INRIA

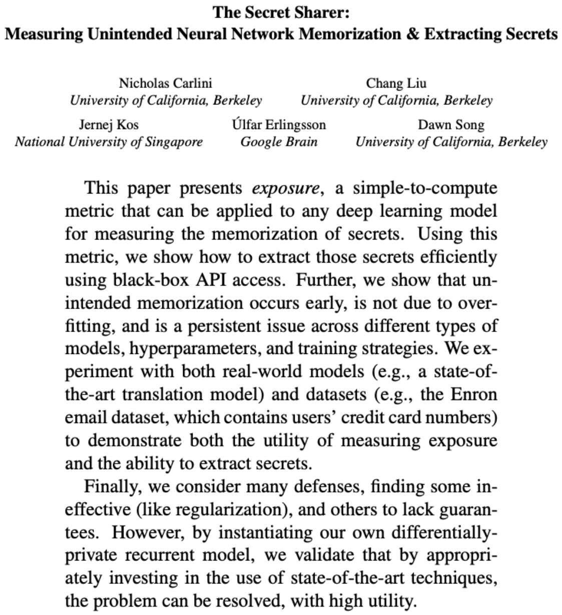
Congzheng Song
Cornell

Vitaly Shmatikov
Cornell Tech

***Abstract*—We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model's training dataset. To perform membership inference against a target model, we make adversarial use of machine learning and train our own inference model to recognize differences in the target model's predictions on the inputs that it trained on versus the inputs that it did not train on.**

We empirically evaluate our inference techniques on classification models trained by commercial "machine learning as a service" providers such as Google and Amazon. Using realistic datasets and classification tasks, including a hospital discharge dataset whose membership is sensitive from the privacy perspective, we show that these models can be vulnerable to membership inference attacks. We then investigate the factors that influence this leakage and evaluate mitigation strategies.

Security and Privacy, 2017



The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets

Nicholas Carlini
University of California, Berkeley

Chang Liu
University of California, Berkeley

Jernej Kos
National University of Singapore

Úlfar Erlingsson
Google Brain

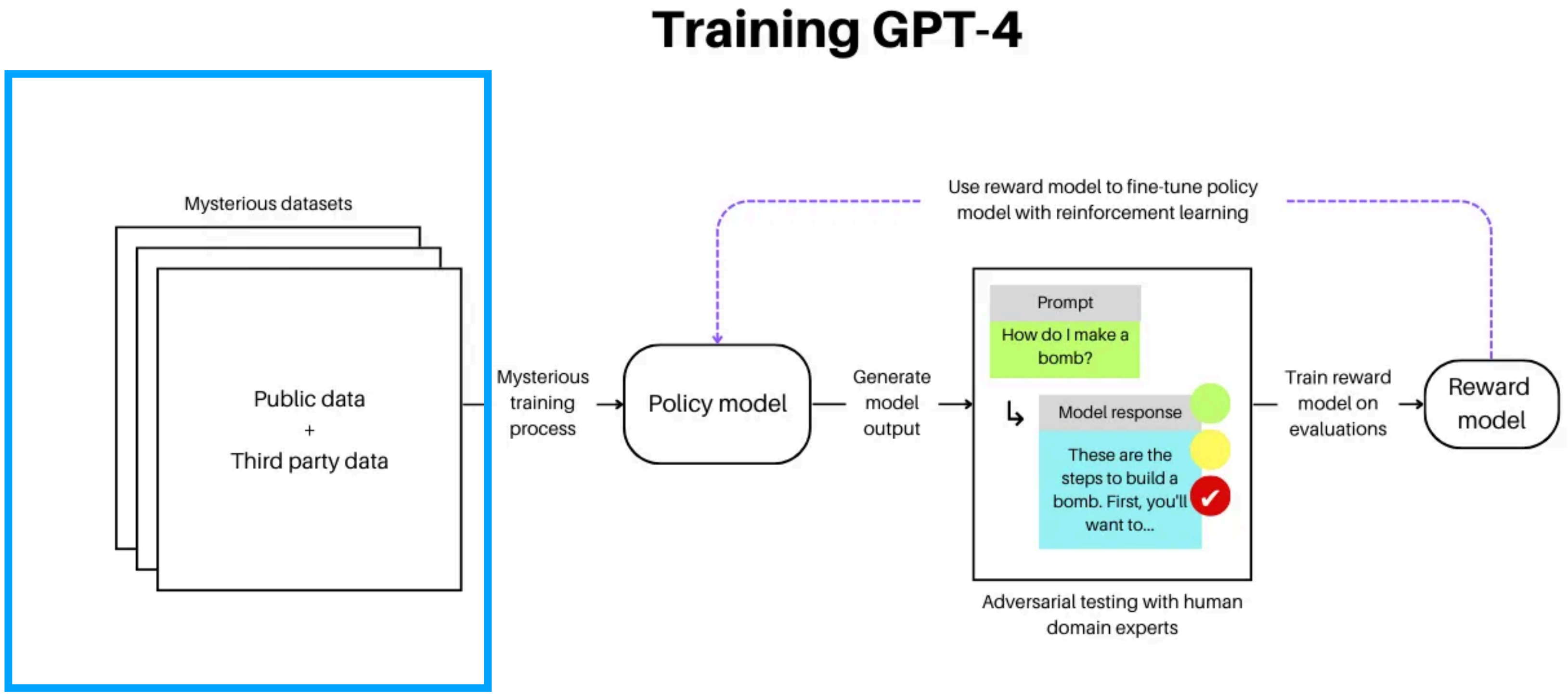
Dawn Song
University of California, Berkeley

This paper presents *exposure*, a simple-to-compute metric that can be applied to any deep learning model for measuring the memorization of secrets. Using this metric, we show how to extract those secrets efficiently using black-box API access. Further, we show that unintended memorization occurs early, is not due to overfitting, and is a persistent issue across different types of models, hyperparameters, and training strategies. We experiment with both real-world models (e.g., a state-of-the-art translation model) and datasets (e.g., the Enron email dataset, which contains users' credit card numbers) to demonstrate both the utility of measuring exposure and the ability to extract secrets.

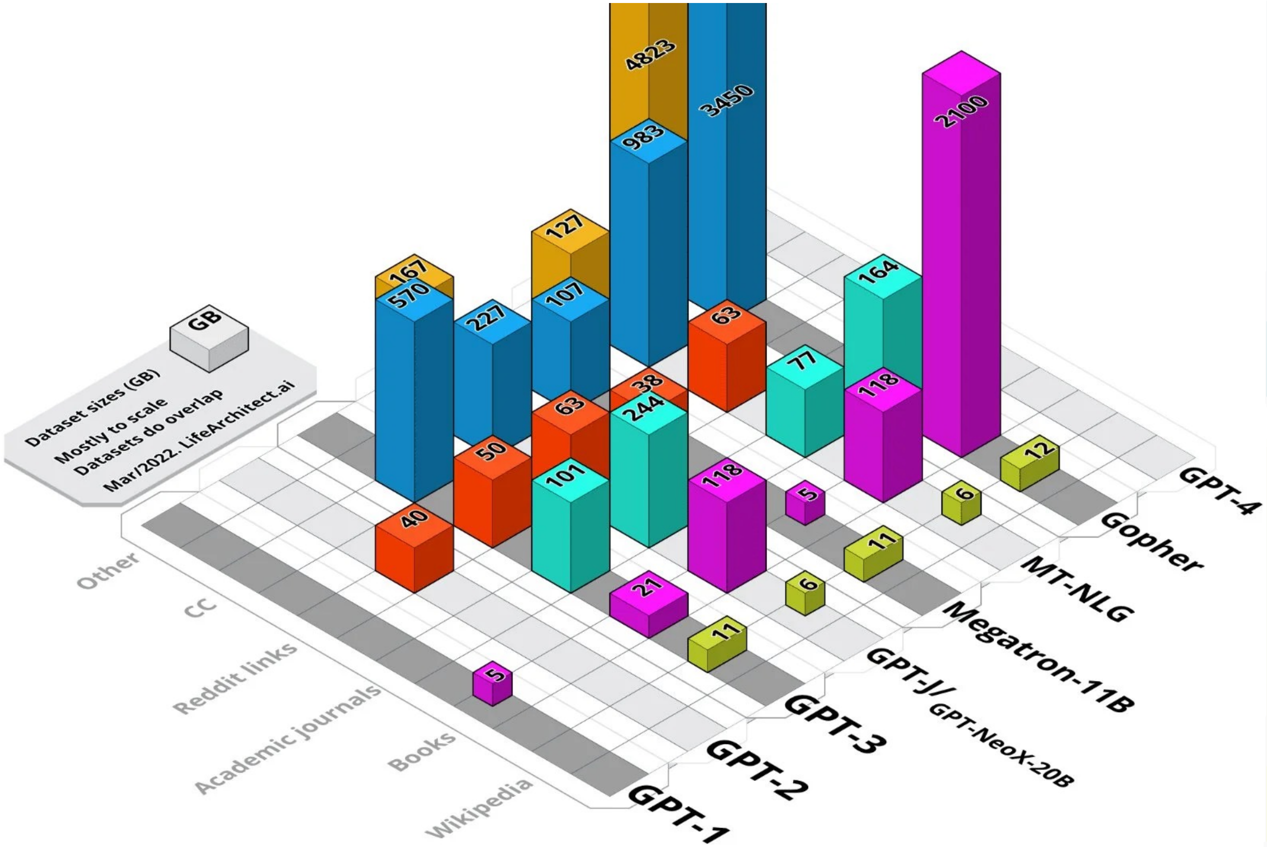
Finally, we consider many defenses, finding some ineffective (like regularization), and others to lack guarantees. However, by instantiating our own differentially-private recurrent model, we validate that by appropriately investing in the use of state-of-the-art techniques, the problem can be resolved, with high utility.

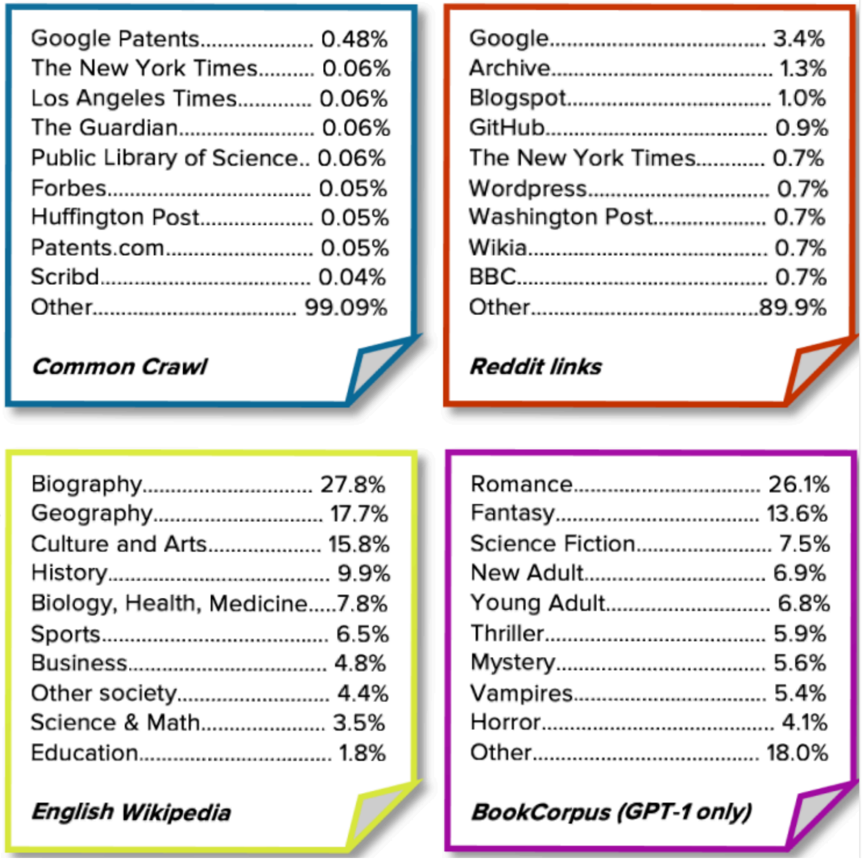
USENIX Security 19

Risk of personal information leakage


Training GPT-4
Mysterious datasets
Public data
+
Third party data
Mysterious training process
Policy model
Generate model output
Use reward model to fine-tune policy model with reinforcement learning
Prompt
How do I make a bomb?
Model response
These are the steps to build a bomb. First, you'll want to...
✓
Adversarial testing with human domain experts
Train reward model on evaluations
Reward model

Risk of personal information leakage





Common Crawl	
Google Patents	0.48%
The New York Times	0.06%
Los Angeles Times	0.06%
The Guardian	0.06%
Public Library of Science	0.06%
Forbes	0.05%
Huffington Post	0.05%
Patents.com	0.05%
Scribd	0.04%
Other	99.09%

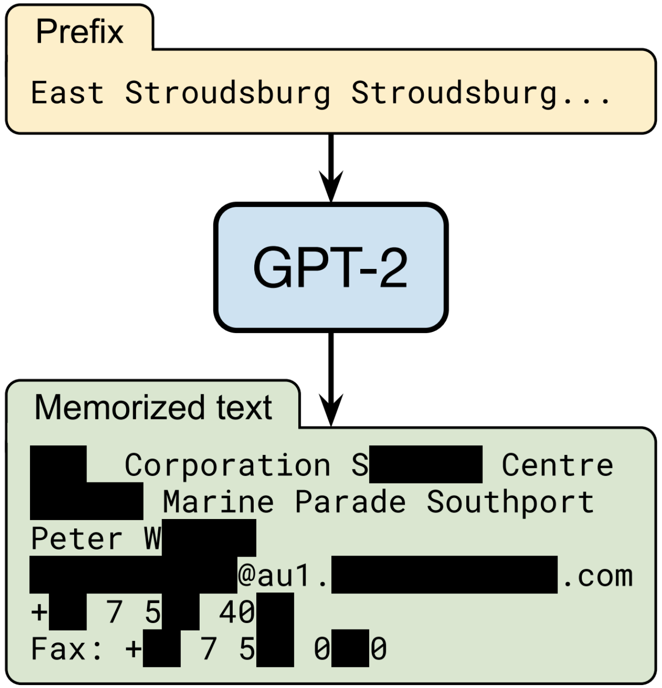
Reddit links	
Google	3.4%
Archive	1.3%
Blogspot	1.0%
GitHub	0.9%
The New York Times	0.7%
Wordpress	0.7%
Washington Post	0.7%
Wikia	0.7%
BBC	0.7%
Other	89.9%

English Wikipedia	
Biography	27.8%
Geography	17.7%
Culture and Arts	15.8%
History	9.9%
Biology, Health, Medicine	7.8%
Sports	6.5%
Business	4.8%
Other society	4.4%
Science & Math	3.5%
Education	1.8%

BookCorpus (GPT-1 only)	
Romance	26.1%
Fantasy	13.6%
Science Fiction	7.5%
New Adult	6.9%
Young Adult	6.8%
Thriller	5.9%
Mystery	5.6%
Vampires	5.4%
Horror	4.1%
Other	18.0%

Web crawlers are used to gather data and train LLMs.

Personal data in Big Data era



Prefix

East Stroudsburg Stroudsburg...

GPT-2

Memorized text

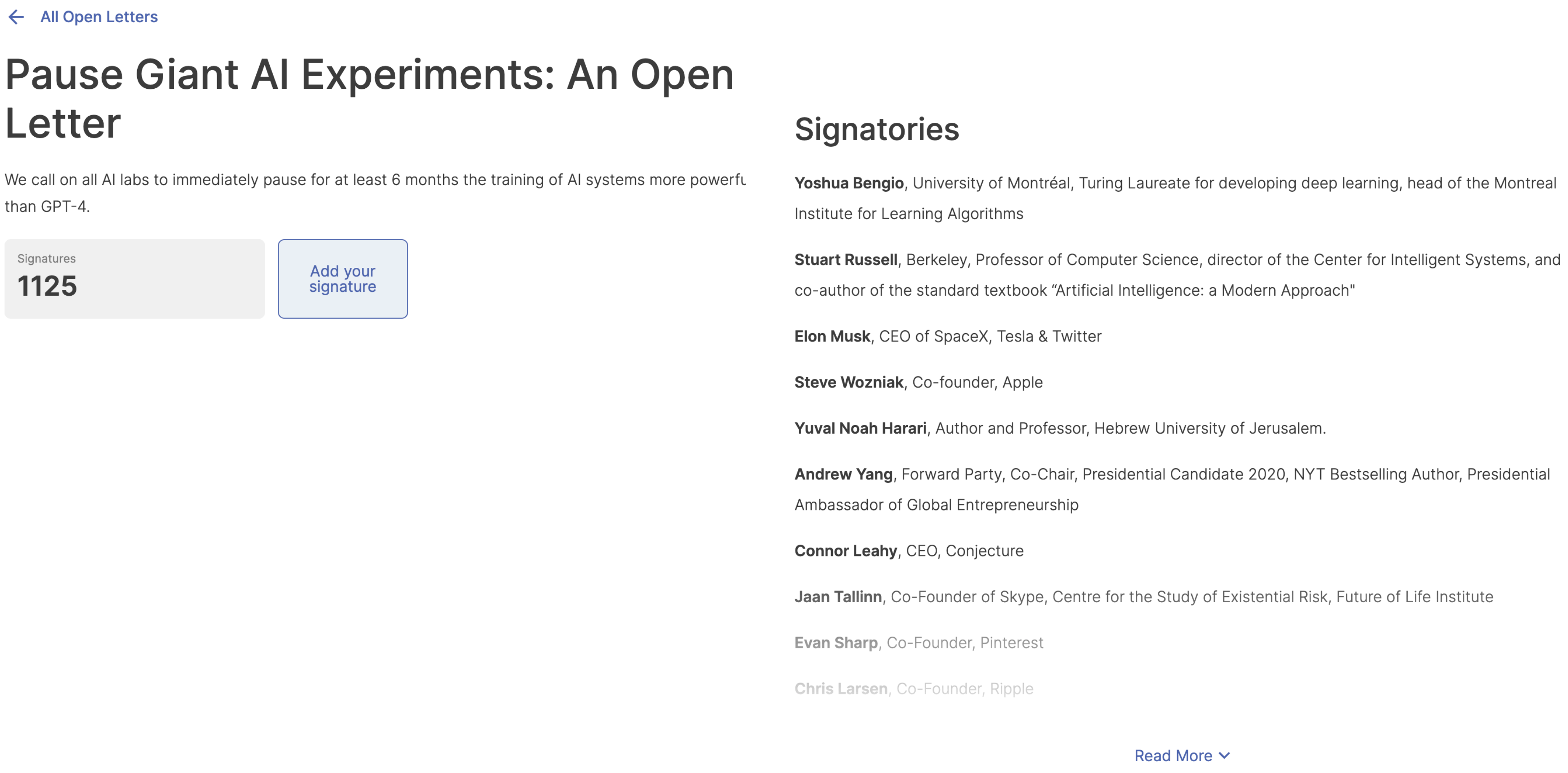
[redacted] Corporation S[redacted] Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@au1.[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Source: Google Research

… prompts the GPT-2 language model with the prefix "East Stroudsburg Stroudsburg…"

it will autocomplete a long block of text that contains the full name, phone number, email address, and physical address of a particular person whose information was included in GPT-2's training data.

Personal data in Big Data era



Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerfu than GPT-4.

Signatures
1125

Signatories

Yoshua Bengio, University of Montréal, Turing Laureate for developing deep learning, head of the Montreal Institute for Learning Algorithms

Stuart Russell, Berkeley, Professor of Computer Science, director of the Center for Intelligent Systems, and co-author of the standard textbook "Artificial Intelligence: a Modern Approach"

Elon Musk, CEO of SpaceX, Tesla & Twitter

Steve Wozniak, Co-founder, Apple

Yuval Noah Harari, Author and Professor, Hebrew University of Jerusalem.

Andrew Yang, Forward Party, Co-Chair, Presidential Candidate 2020, NYT Bestselling Author, Presidential Ambassador of Global Entrepreneurship

Connor Leahy, CEO, Conjecture

Jaan Tallinn, Co-Founder of Skype, Centre for the Study of Existential Risk, Future of Life Institute

Evan Sharp, Co-Founder, Pinterest

Chris Larsen, Co-Founder, Ripple

Review

First Part of This Course:

- Ensemble
- Learning Theory
- GNN
- Generative Models

Focus more on a single merit: accuracy

Outlook

Second Part of This Course:

- Causal Learning
- **Differential Privacy and Federated Learning** (This lecture)
- Fairness in ML
- Explainable AI (XAI)

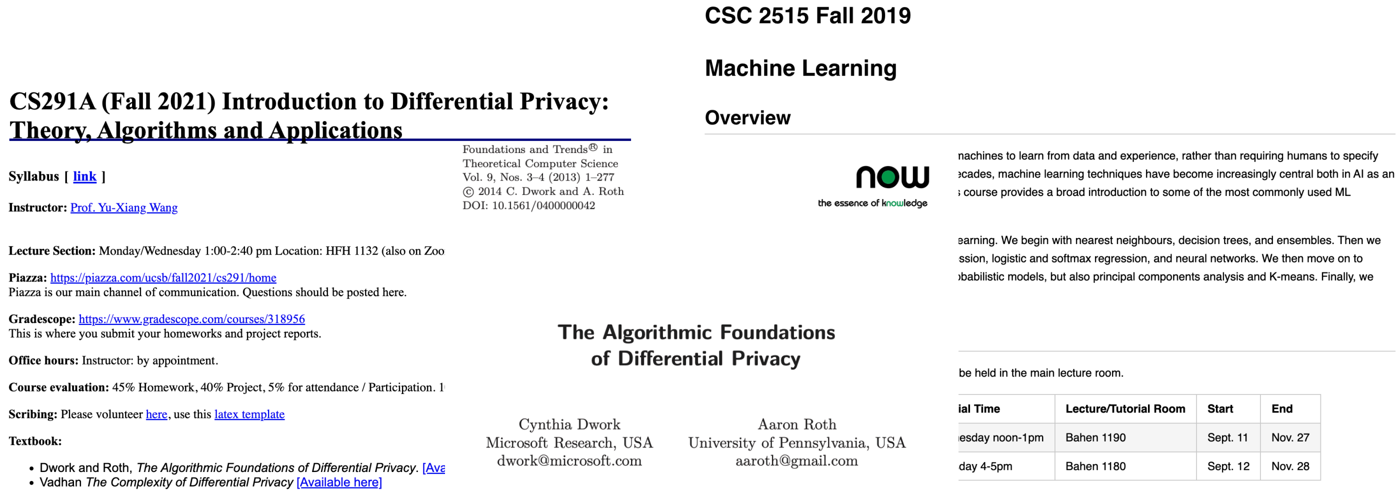
Focus on more attributes: **causality**, privacy, **fairness,** and **interpretability**

This Lecture:

Introduction to Differential Privacy and Federated Learning

Outline

Again, **privacy** in ML can be a full course, we will only highlight a few important concepts



CS291A (Fall 2021) Introduction to Differential Privacy: Theory, Algorithms and Applications

Syllabus [link]

Instructor: Prof. Yu-Xiang Wang

Lecture Section: Monday/Wednesday 1:00-2:40 pm Location: HFH 1132 (also on Zoo

Piazza: https://piazza.com/ucsb/fall2021/cs291/home
Piazza is our main channel of communication. Questions should be posted here.

Gradescope: https://www.gradescope.com/courses/318956
This is where you submit your homeworks and project reports.

Office hours: Instructor: by appointment.

Course evaluation: 45% Homework, 40% Project, 5% for attendance / Participation. 1

Scribing: Please volunteer here, use this latex template

Textbook:

- Dwork and Roth, *The Algorithmic Foundations of Differential Privacy.* [Ava
- Vadhan *The Complexity of Differential Privacy* [Available here]

Foundations and Trends® in
Theoretical Computer Science
Vol. 9, Nos. 3–4 (2013) 1–277
© 2014 C. Dwork and A. Roth
DOI: 10.1561/0400000042

now
the essence of knowledge

The Algorithmic Foundations of Differential Privacy

Cynthia Dwork
Microsoft Research, USA
dwork@microsoft.com

Aaron Roth
University of Pennsylvania, USA
aaroth@gmail.com

CSC 2515 Fall 2019

Machine Learning

Overview

nachines to learn from data and experience, rather than requiring humans to specify
ecades, machine learning techniques have become increasingly central both in AI as an
course provides a broad introduction to some of the most commonly used ML

earning. We begin with nearest neighbours, decision trees, and ensembles. Then we
ssion, logistic and softmax regression, and neural networks. We then move on to
babilistic models, but also principal components analysis and K-means. Finally, we

be held in the main lecture room.

ial Time	Lecture/Tutorial Room	Start	End
esday noon-1pm	Bahen 1190	Sept. 11	Nov. 27
day 4-5pm	Bahen 1180	Sept. 12	Nov. 28

Contact

- **Instructor:** Roger Grosse (rgrosse at cs.toronto.edu)
- **Office hours:**

Privacy and Federated Learning:
Principles, Techniques and Emerging Frontiers

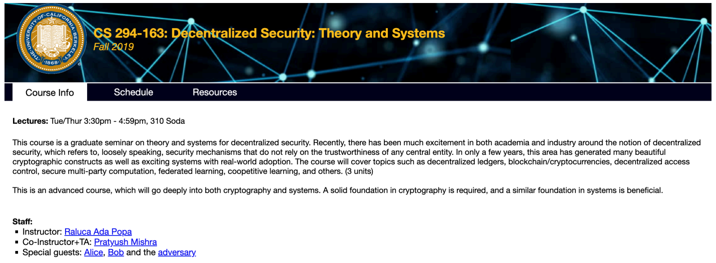
AAAI Workshop of Privacy Preserving Artificial Intelligence (PPAI-21)

Brendan McMahan

Kallista Bonawitz

Peter Kairouz

Presenting the work of many



CS 294-163: Decentralized Security: Theory and Systems
Fall 2019

Lectures: Tue/Thur 3:30pm - 4:59pm, 310 Soda

This course is a graduate seminar on theory and systems for decentralized security. Recently, there has been much excitement in both academia and industry around the notion of decentralized security, which refers to, loosely speaking, security mechanisms that do not rely on the trustworthiness of any central entity. In only a few years, this area has generated many beautiful cryptographic constructs as well as exciting systems with real-world adoption. The course will cover topics such as decentralized ledgers, blockchain/cryptocurrencies, decentralized access control, secure multi-party computation, federated learning, coopetitive learning, and others. (3 units)

This is an advanced course, which will go deeply into both cryptography and systems. A solid foundation in cryptography is required, and a similar foundation in systems is beneficial.

Staff:
- Instructor: Raluca Ada Popa
- Co-Instructor+TA: Pratyush Mishra
- Special guests: Alice, Bob and the adversary

Outline

- Motivation and attacks
  - Reconstruction attacks
- Differential privacy
  - Definitions
  - ~~Gaussian Mechanism~~
  - Basic Mechanisms: Randomized Response, Laplace, Exponential
- ~~Differentially Private Machine Learning~~
- Federated learning with DP
  - Problem and Framework

Again, **privacy** in ML can be a full course, we will only highlight a few important concepts

Part I

Differential Privacy

Part 1.1

Motivation, Attacks, and History

Do we need a formal mathematical definition for privacy?

- Can’t we just remove personal identifiable information from the data so that it is de-identified?

- We are only seeing aggregate statistics usually?

- Secure multi-party computation (MPC) and federated learning have made it possible for companies to train ML models with my data while keeping my data on my device?

Do we need a formal mathematical definition for privacy?

Consider a simple and practical method:

Removing/modifying personal identifiable information

Questionnaire: Have you ever driven under the influence?

Example:

Name	DUI?
John	Yes
Jack	No
Jennifer	Yes
James	No

Table: Dataset

Q: Is this a good enough privacy-preserving method?

Do we need a formal mathematical definition for privacy?

Consider a simple and practical method:

Removing/modifying personal identifiable information

Example:

Name	DUI?
[redacted]	Yes
[redacted]	No
[redacted]	Yes
[redacted]	No

Table: Dataset

Q: Is this a good enough privacy-preserving method?

Do we need a formal mathematical definition for privacy?

Consider a simple and practical method:

Removing/modifying personal identifiable information

Example:

From a different web …

Name	DUI?
[redacted]	Yes
[redacted]	No
[redacted]	Yes
[redacted]	No

Table: Dataset

Car Model	ZIP
Mazda 6	91106
Tesla	21927
Accord	23772
Benz	12678

…

Q: Is this a good enough privacy-preserving method?

Do we need a formal mathematical definition for privacy?

Consider a simple and practical method:

Removing/modifying personal identifiable information

On More Example:

CTE

What is your rate of the teacher/course?

Do we need a formal mathematical definition for privacy?

Consider a simple and practical method:

Removing/modifying personal identifiable information

On More Example:

CTE Name: Anonym

What is your rate of the teacher/course? 0/6

Why do you rate under 3?

I obtained only 3/10 for problem 1.(a) in HW2.

	Score
Daniel	3
James	2
Alice	5

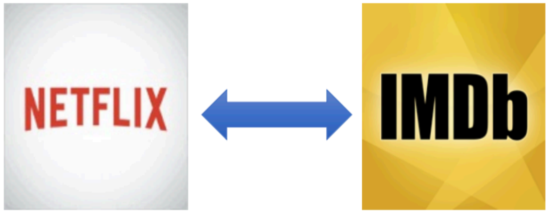
Q: Is this a good enough privacy-preserving method?

Do we need a formal mathematical definition for privacy?

Consider a simple and practical method:

Removing/modifying personal identifiable information

- Name? Gender? Zip code? Watched movies?

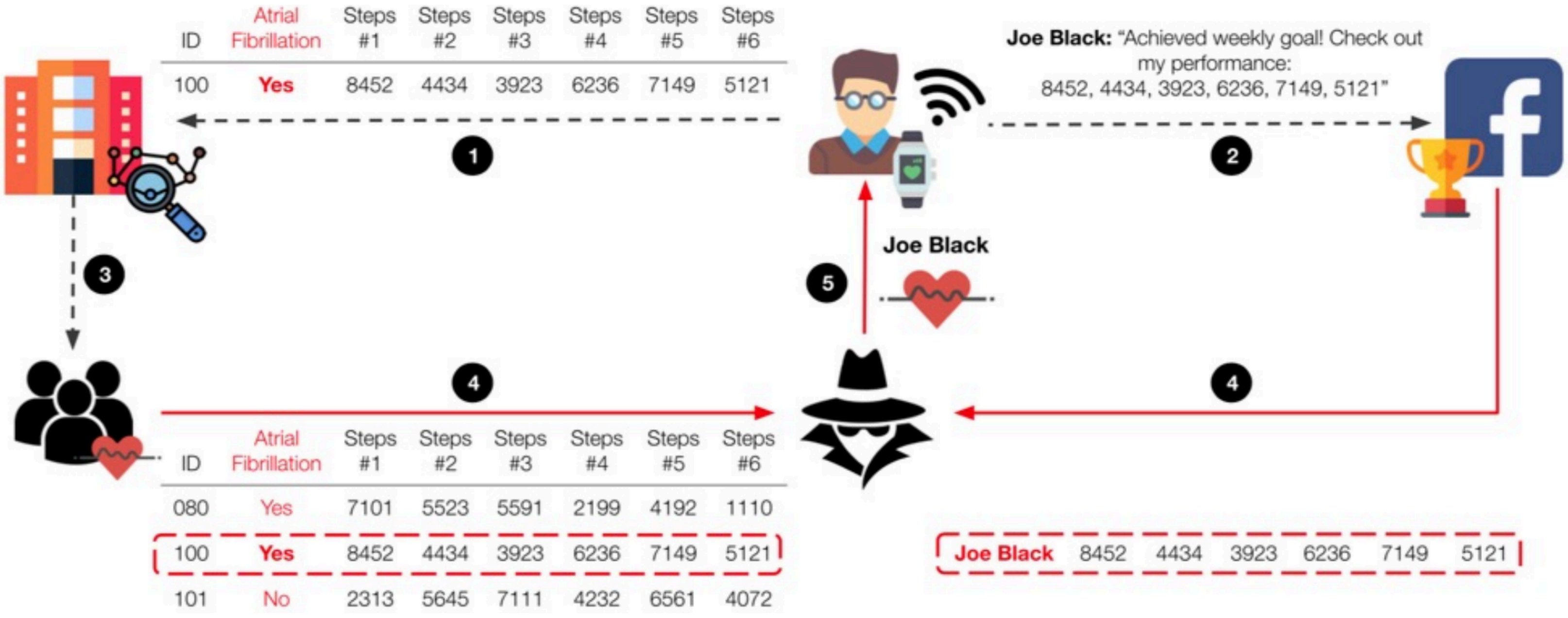


- Fragile under appropriate **side information**
- Easier to get in Big Data era

NOT a good enough privacy-preserving method!

Do we need a formal mathematical definition for privacy?

“Just six days of step counts are enough to uniquely identify you among 100 million other people.”



ID	Atrial Fibrillation	Steps #1	Steps #2	Steps #3	Steps #4	Steps #5	Steps #6
100	Yes	8452	4434	3923	6236	7149	5121

Joe Black: “Achieved weekly goal! Check out my performance: 8452, 4434, 3923, 6236, 7149, 5121”

ID	Atrial Fibrillation	Steps #1	Steps #2	Steps #3	Steps #4	Steps #5	Steps #6
080	Yes	7101	5523	5591	2199	4192	1110
100	Yes	8452	4434	3923	6236	7149	5121
101	No	2313	5645	7111	4232	6561	4072

Joe Black	8452	4434	3923	6236	7149	5121

https://www.mobihealthnews.com/news/contributed-when-fitness-data-becomes-research-data-your-privacy-may-be-risk

Differencing attack and side information identifies individuals from aggregate statistics

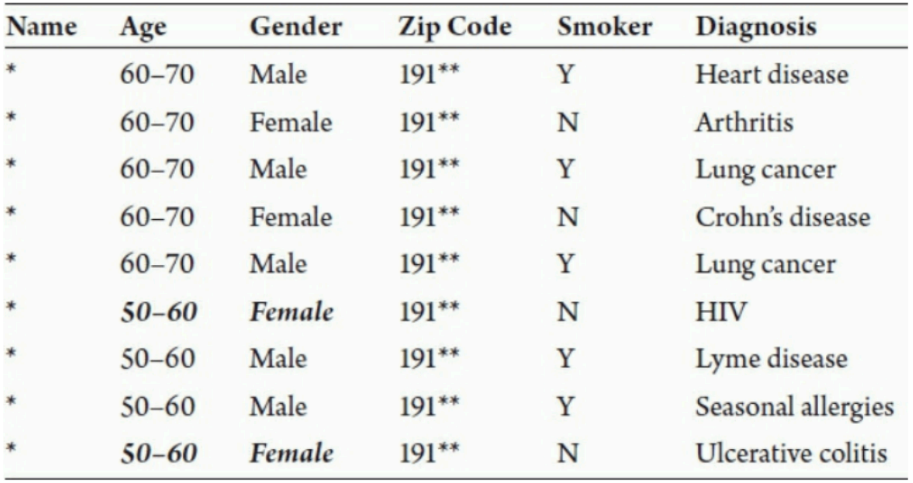
Real-world Examples

- In the 1990s, a government agency released a database of medical visits, stripped of identifying information (names, addresses, social security numbers)
  - But it did contain zip code, birth date, and gender.
  - Researchers estimated that 87 percent of Americans are uniquely identifiable from this triplet.

- Netflix Challenge (2006), a Kaggle-style competition to improve their movie recommendations, with a $1 million prize
  - They released a dataset consisting of 100 million movie ratings (by “anonymized” numeric user ID), with dates
  - Researchers found they could identify 99% of users who rated 6 or more movies by cross-referencing with IMDB, where people posted reviews publicly with their real names

Real-world Examples

Not sufficient to prevent unique identification of individuals.



Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60–70	Male	191**	Y	Heart disease
*	60–70	Female	191**	N	Arthritis
*	60–70	Male	191**	Y	Lung cancer
*	60–70	Female	191**	N	Crohn's disease
*	60–70	Male	191**	Y	Lung cancer
*	***50–60***	***Female***	191**	N	HIV
*	50–60	Male	191**	Y	Lyme disease
*	50–60	Male	191**	Y	Seasonal allergies
*	***50–60***	***Female***	191**	N	Ulcerative colitis

Kearns & Roth, *The Ethical Algorithm*

From this (fictional) hospital database, if we know Rebecca is 55 years old and in this database, then we know she has 1 of 2 diseases.

Possible attacks

Membership inference attack:

- Train a ML model to predict whether individuals are used for training.
- Often obvious from the confidence of the ML-predictions alone.

Unintended memorization attack:

- Prompt a language model: Alice's SSN is ????-??-7452
- Ask the language model to fill-in the question marks.

FYI: Modern DP learning models memorizes the entire dataset using their billions of parameters. They can be thought of as an implicit transformation of the data into an efficient data-structure. In fact, memorization might be the very reason why deep models work well. See (Feldman, 2019) https://arxiv.org/abs/1906.05271

Generative model-inversion attacks:

- Model inversion attacks recover information about the training data from the trained model.

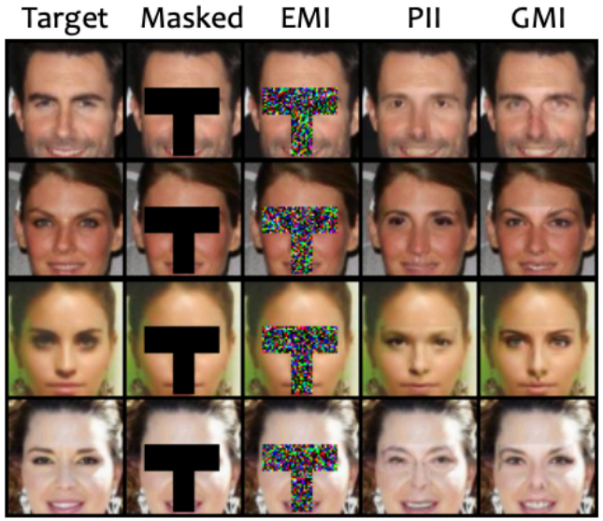
Reconstruction attack:

- An even stronger attack later: even without side-information, even with noise in the statistics.

Possible attacks

Generative model-inversion attacks

- Model inversion attacks recover information about the training data from the trained model.



reconstructing individuals from a face recognition dataset, given a classifier trained on this dataset and a generative model trained on an unrelated dataset of publicly available images.

Zhang et al., "The secret revealer: Generative model-inversion attacks against deep neural networks." https://arxiv.org/abs/1911.07135

Many more DNN-based attacks …

Reconstruction Attack

Introduced in a seminal paper by Dinur and Nissim in 2003 This attack motivates differential privacy!

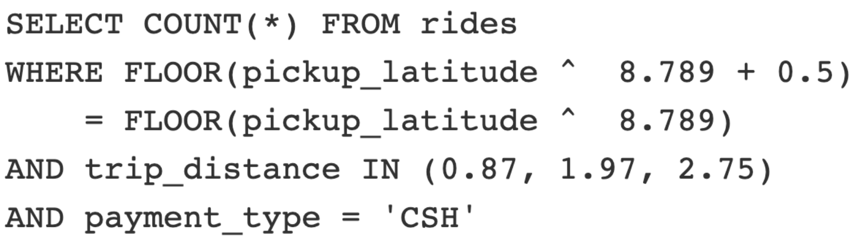
Reconstruction Attacks in Practice:

US Census Bureau in 2018

motivating the Bureau's adoption of differential privacy for data products derived from the 2020 decennial census!

Diffix

Aloni Cohen and Kobbi Nissim in the first bug bounty program 2017 - 2018
Travis Dick, Matthew Joseph, and Zachary Schutzman in the second bug bounty program 2020



```
SELECT COUNT(*) FROM rides
WHERE FLOOR(pickup_latitude ^  8.789 + 0.5)
    = FLOOR(pickup_latitude ^  8.789)
AND trip_distance IN (0.87, 1.97, 2.75)
AND payment_type = 'CSH'
```

Reconstruction Attack

Consider the following model:

- We have a dataset of n individuals $X = \{x_1, \ldots, x_n\}$
- One secret bit of information per person $b_i \in \{0, 1\}$
- Each record is $x_i = (z_i, b_i)$ $z_i \in \{0, 1\}^{d-1}$ $b_i \in \{0, 1\}$ (d binary attributes)
- We can visualize such a dataset as a matrix

$$[Z \mid b] = \left[\begin{array}{c|c} z_1 & b_1 \\ \vdots & \vdots \\ z_n & b_n \end{array}\right]$$

Reconstruction Attack

Consider the following model:

- The distinction between z_i and b_i is only in the mind of the attacker
- $\{z_1, \ldots, z_n\}$ is the prior information
- $\{b_1, \ldots, b_n\}$ are the secret bits

Our goal is to understand if asking aggregate queries (defined by the prior information) can learn non-trivial information about the secret bits.

Most basic aggregate query: **counting query** (how many data points satisfy a property)

The Dinur-Nissim attack uses a query q specified by a function $\varphi : \{0,1\}^{d-1} \rightarrow \{0,1\}$

$$q(X) = \sum_{j=1}^{n} \varphi(z_j) \cdot b_j.$$

Reconstruction Attack

The Dinur-Nissim attack uses a query q specified by a function $\varphi : \{0,1\}^{d-1} \to \{0,1\}$

$$q(X) = \sum_{j=1}^{n} \varphi(z_j) \cdot b_j.$$

$$\begin{bmatrix} q_1(X) \\ \vdots \\ q_k(X) \end{bmatrix} = \underset{\substack{\Big\downarrow \\ Q_Z}}{\begin{bmatrix} \varphi_1(z_1) & \dots & \varphi_1(z_n) \\ \vdots & \ddots & \vdots \\ \varphi_k(z_1) & \dots & \varphi_k(z_n) \end{bmatrix}} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

First Trial

The Dinur-Nissim attack uses a query q specified by a function $\varphi : \{0{,}1\}^{d-1} \rightarrow \{0{,}1\}$

$$q(X) = \sum_{j=1}^{n} \varphi(z_j) \cdot b_j.$$

How to design φ ? If answers are noiseless:

$$\varphi_i(z) = \begin{cases} 1 & \text{if } z = z_i \\ 0 & \text{otherwise} \end{cases}$$

First Trial

The Dinur-Nissim attack uses a query q specified by a function $\varphi : \{0,1\}^{d-1} \to \{0,1\}$

$$q(X) = \sum_{j=1}^{n} \varphi(z_j) \cdot b_j.$$

How to design φ ? If answers are noiseless:

$$\varphi_i(z) = \begin{cases} 1 & \text{if } z = z_i \\ 0 & \text{otherwise} \end{cases}$$

What if the answers are noisy?

First Trial

An “inefficient” attacking scheme:

- For simplicity, assume all $z_1, \ldots, z_n$ are distinct so that each user is uniquely identified by the prior information

- The attacker chooses queries $q_1, \ldots, q_k$ ($k = 2^n$) so that the matrix Q_Z has as its rows all of $\{0,1\}^n$.

- The attacker receives a vector a of noisy answers to the queries, where $|q_i(X) - a_i| < \alpha n$ for all q_i

$$\max_{i=1}^{k} |(Q_Z \cdot b)_i - a_i| = \|Q_Z \cdot b - a\|_\infty \leq \alpha n.$$

- The attacker outputs any consistent guess $\hat{b} = \{\hat{b}_n, \ldots, \hat{b}_n\}$ of the private bits vector, i.e.,

$$\max_{i=1}^{k} |a_i - (Q_Z \cdot \hat{b})_i| = \|Q_Z \cdot \hat{b} - a\|_\infty \leq \alpha n$$

Does there always exist such $\hat{b}$?

First Trial

Theorem [Dinur and Nissim 03]: There is a reconstruction attack that issues $k = 2^n$ queries to a dataset of n users, obtains answers with error αn, and reconstructs the secret bits of all but $4\alpha n$ users.

Proof:

First Trial

Theorem [Dinur and Nissim 03]: There is a reconstruction attack that issues $k = 2^n$ queries to a dataset of n users, obtains answers with error αn, and reconstructs the secret bits of all but $4\alpha n$ users.

Proof: Fix some $\hat{b}$

$$S_{01} = \{j : \hat{b}_j = 0, b_j = 1\} \text{ and } S_{10} = \{j : \hat{b}_j = 1, b_j = 0\}$$

If b and $\hat{b}$ differ by more than $4\alpha n$ bits, then at least one of above sets has more than $2\alpha n$ items

WLOG, assume it is S_{01}

Suppose the i-th row of Q_Z is the indicator vector of S_{01} (Q: why can we do this?)

$$(Q_Z)_{i,j} = 1 \iff j \in S_{01}.$$

$$|(Q_Z \cdot b)_i - (Q_Z \cdot \hat{b})_i| = |S_{01}| > 2\alpha n\,.$$

First Trial

Theorem [Dinur and Nissim 03]: There is a reconstruction attack that issues $k = 2^n$ queries to a dataset of n users, obtains answers with error αn, and reconstructs the secret bits of all but $4\alpha n$ users.

Proof: Suppose the i-th row of Q_Z is the indicator vector of S_{01} (Q: why can we do this?)

$$(Q_Z)_{i,j} = 1 \iff j \in S_{01}.$$

$$|(Q_Z \cdot b)_i - (Q_Z \cdot \hat{b})_i| = |S_{01}| > 2\alpha n.$$

However, $\hat{b}$ should satisfy

$$|(Q_Z \cdot b)_i - (Q_Z \cdot \hat{b})_i| \le |a_i - (Q_Z \cdot \hat{b})_i| + |(Q_Z \cdot b)_i - a_i| \le 2\alpha n,$$

which implies a contradiction.

Can we have an efficient attack?

An "efficient" attacking scheme:

- The attacker now chooses k randomly chosen functions φ_i for some much smaller $k = O(n)$
- Upon receiving an answer vector a, the attacker now searches for a *real-valued* $\tilde{b} \in [0,1]^n$ such that

$$\|\boldsymbol{a} - \boldsymbol{Q}_Z \cdot \tilde{\boldsymbol{b}}\|_\infty \leq \boldsymbol{\alpha n}.$$

(Can be found efficiently by LP)

- $\tilde{b}_i$ is round to the nearest $\hat{b}_i$

Can we have an efficient attack?

An "efficient" attacking scheme:

- The attacker now chooses k randomly chosen functions φ_i for some much smaller $k = O(n)$
- Upon receiving an answer vector a, the attacker now searches for a *real-valued* $\tilde{b} \in [0,1]^n$ such that

$$\|a - Q_Z \cdot \tilde{b}\|_\infty \le \alpha n.$$

(Can be found efficiently by LP)

- $\tilde{b}_i$ is round to the nearest $\hat{b}_i$

Q_Z is a random matrix, can show that with high probability $\|Q_Z \cdot b - Q_Z \cdot \tilde{b}\|_\infty^2 \gtrsim |i : b_i \neq \hat{b}_i|$.

The scheme implies $\|Q_Z \cdot b - Q_Z \cdot \tilde{b}\|_\infty \le \|Q_Z \cdot b - a\|_\infty + \|a - Q_Z \cdot \tilde{b}\|_\infty \le 2\alpha n$

reconstruction error $\approx O(\alpha^2 n^2)$

Can we have an efficient attack?

An "efficient" attacking scheme:

- The attacker now chooses k randomly chosen functions φ_i for some much smaller $k = O(n)$
- Upon receiving an answer vector a, the attacker now searches for a *real-valued* $\tilde{b} \in [0,1]^n$ such that

$$\|\boldsymbol{a} - \boldsymbol{Q}_Z \cdot \tilde{\boldsymbol{b}}\|_\infty \leq \alpha \boldsymbol{n}.$$

(Can be found efficiently by LP)

- $\tilde{b}_i$ is round to the nearest $\hat{b}_i$

Theorem [Dinur and Nissim 03]: There is a reconstruction attack that issues $k = O(n)$ queries to a dataset of n users, obtains answers with error αn, and with high probability, reconstructs the secret bits of all but $\alpha^2 n^2$ users.

Enforcing privacy is challenging!

- Revealing dataset (even if with PII removed) is a bad idea

  Model inversion attacks recover information about the training data from the trained model.

  Even if you don't release the raw data, the weights of a trained network might reveal sensitive information.

- Revealing aggregate statistics of the dataset has privacy risks

  Differencing attack: with side information, even if reporting just one, may reveal information about individuals

  Reconstruction attack

- Machine learning models encodes information of individuals in a dataset and will spit them out when given a carefully constructed prompt

  Membership inference attack

  Unintended memorization

Incomplete history of privacy protection

- *Since 1970s:* Statistical disclosure control (Duncan et al.; Hundepool et al)
  - e.g., Data swapping (Dalenius, Reiss, 1982) was implemented in the Census

- *2002 – 2007:* K-anonymity, l-divergence, t-closeness (Sweeney et. al., Machanavajjhala et. al., Li et. al., 2002 - 2007)
  - These attempts have been shown to be fragile against side-information and composition. See a recent revisit of this problem (and the references therein): https://aloni.net/wp-content/uploads/2021/05/Quasi-IDs-are-the-Problem-working-paper.pdf

- *2006+:* Differential privacy [Dwork, McSherry, Nissim, Smith, 2006++]

(Motivated by the reconstruction attack)

Incomplete history of privacy protection

- *Since 1970s:* Statistical disclosure control (Duncan et al.; Hundepool et al)
  - e.g., Data swapping (Dalenius, Reiss, 1982) was implemented in the Census

- *2002 – 2007:* K-anonymity, l-divergence, t-closeness (Sweeney et. al., Machanavajjhala et. al., Li et. al., 2002 - 2007)
  - These attempts have been shown to be fragile against side-information and composition. See a recent revisit of this problem (and the references therein): https://aloni.net/wp-content/uploads/2021/05/Quasi-IDs-are-the-Problem-working-paper.pdf

- *2006+:* Differential privacy [Dwork, McSherry, Nissim, Smith, 2006++]

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265-284). Springer, Berlin, Heidelberg.

2017 Gödel Prize winners:
Dwork, McSherry, Nissim & Smith

Part 1.2

Intuition and Definition

Necessary properties

What properties are desirable?

- Do we want to target on some specific attack?
- Do we need assumptions on the adversary?
- Do we need assumptions on the input data?
- Do we want to have a composable privacy component?

Necessary properties

- Protect against most (if not all) attacks known to date
- Not making strong assumptions about the adversary
- Not making strong assumptions about the input data
- Graceful degradation over composition (by repeatedly calling DP algorithms or other functions)

Do we need a formal mathematical definition for privacy? Yes!

Idea: mathematical guarantees

- We have seen several attacks

  Generative model-inversion attack

  Membership inference attack

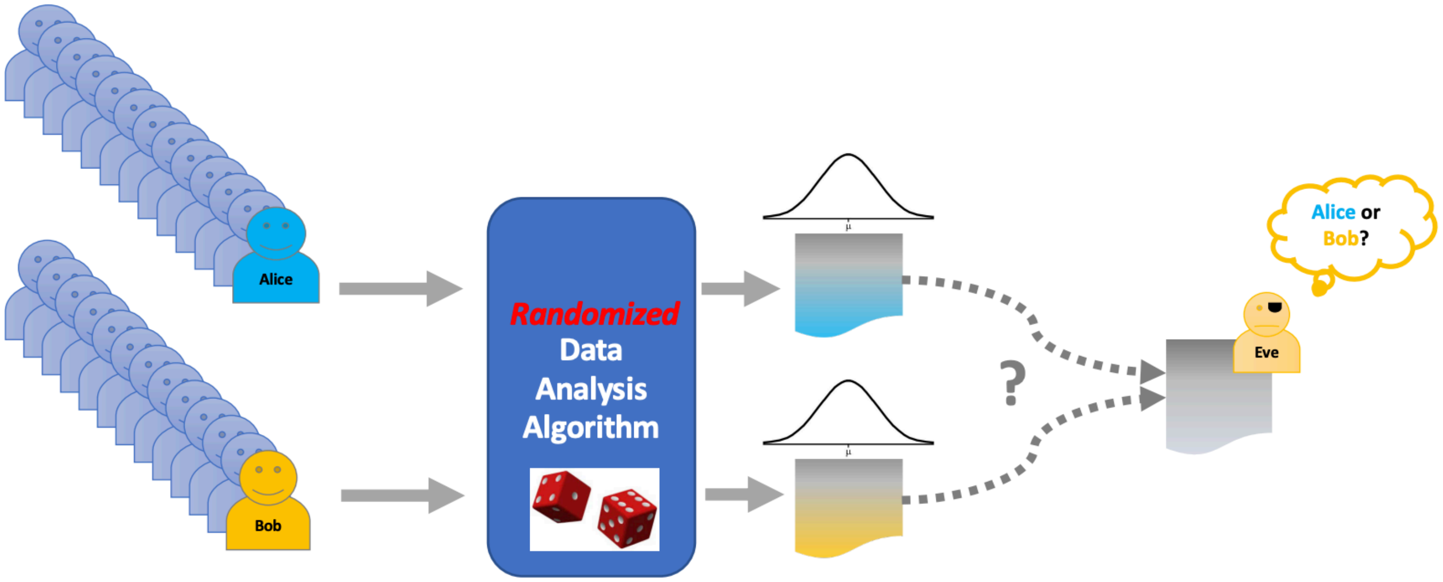
  Unintended memorization attack

  Reconstruction attack

- It is insufficient to defend against one specific attack.

- Idea: a pure mathematical definition that separates "privacy definition" from the actual algorithm that implements the defense.

Key idea of differential privacy: randomized response



Differential privacy by examples

Let's reconsider this example:

Questionnaire: Have you ever dodged your taxes?

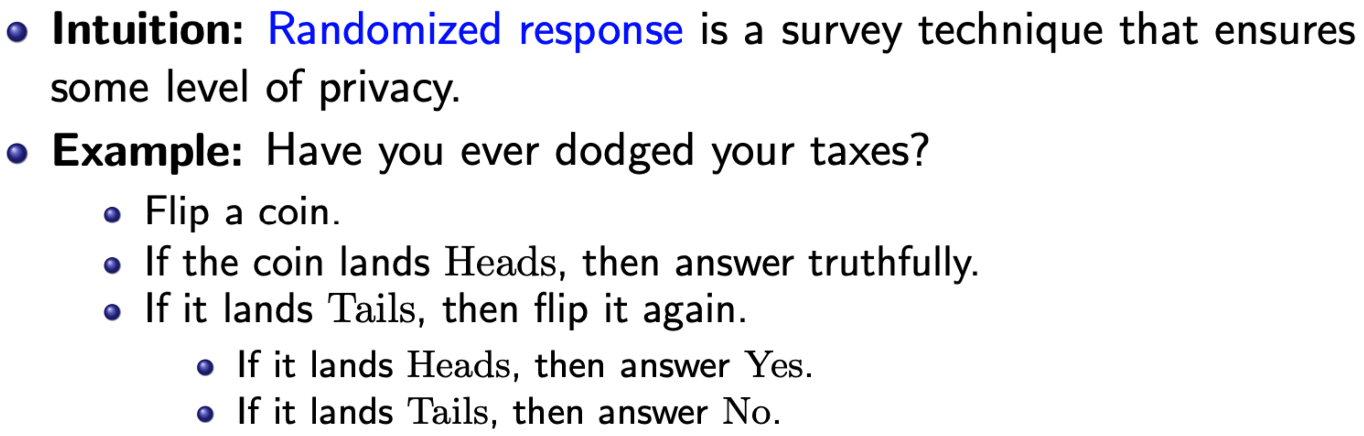
Name	Answer
John	Yes
Jack	No
Jennifer	Yes
James	No

Table: Dataset

Differential privacy by examples

Let's reconsider this example:

Questionnaire: Have you ever dodged your taxes?



- **Intuition:** Randomized response is a survey technique that ensures some level of privacy.
- **Example:** Have you ever dodged your taxes?
  - Flip a coin.
  - If the coin lands Heads, then answer truthfully.
  - If it lands Tails, then flip it again.
    - If it lands Heads, then answer Yes.
    - If it lands Tails, then answer No.

Differential privacy by examples

Let's reconsider this example:

Questionnaire: Have you ever dodged your taxes?

- **Intuition:** Randomized response is a survey technique that ensures some level of privacy.
- **Example:** Have you ever dodged your taxes?
  - Flip a coin.
  - If the coin lands Heads, then answer truthfully.
  - If it lands Tails, then flip it again.
    - If it lands Heads, then answer Yes.
    - If it lands Tails, then answer No.
- **Probability of responses:**

	Yes	No
Dodge	3/4	1/4
No Dodge	1/4	3/4

Differential privacy by examples

Let's reconsider this example:

Does this randomization help?

- Tammy the Tax Investigator assigns a prior probability of 0.02 to Bob having dodged his taxes. Then she notices he answered Yes to the survey. What is her posterior probability?

$$
\begin{aligned}
\Pr(\mathrm{Dodge} \mid \mathrm{Yes}) &= \frac{\Pr(\mathrm{Dodge}) \Pr(\mathrm{Yes} \mid \mathrm{Dodge})}{\Pr(\mathrm{Dodge}) \Pr(\mathrm{Yes} \mid \mathrm{Dodge}) + \Pr(\mathrm{NoDodge}) \Pr(\mathrm{Yes} \mid \mathrm{NoDodge})} \\
&= \frac{0.02 \cdot \frac{3}{4}}{0.02 \cdot \frac{3}{4} + 0.98 \cdot \frac{1}{4}} \\
&\approx 0.058
\end{aligned}
$$

Differential privacy by examples

Let's reconsider this example:

Does this randomization help?

- Tammy the Tax Investigator assigns a prior probability of 0.02 to Bob having dodged his taxes. Then she notices he answered Yes to the survey. What is her posterior probability?

$$
\begin{aligned}
\Pr(\mathrm{Dodge} \mid \mathrm{Yes}) &= \frac{\Pr(\mathrm{Dodge}) \Pr(\mathrm{Yes} \mid \mathrm{Dodge})}{\Pr(\mathrm{Dodge}) \Pr(\mathrm{Yes} \mid \mathrm{Dodge}) + \Pr(\mathrm{NoDodge}) \Pr(\mathrm{Yes} \mid \mathrm{NoDodge})} \\
&= \frac{0.02 \cdot \frac{3}{4}}{0.02 \cdot \frac{3}{4} + 0.98 \cdot \frac{1}{4}} \\
&\approx 0.058
\end{aligned}
$$

- So Tammy's beliefs haven't shifted too much.
- More generally, randomness turns out to be a really useful technique for preventing information leakage.

Differential privacy by examples

Let’s reconsider this example:

Does randomization affect population-level statistics?

- Does randomization change the population mean μ ?

Differential privacy by examples

Let's reconsider this example:

Does randomization affect population-level statistics?

- Does randomization change the population mean μ ?
- How accurately can we estimate μ, the population mean?
- Let $X_{\mathrm{T}}^{(i)}$ denote individual i's response if they respond truthfully, and $X_{\mathrm{R}}^{(i)}$ individual i's response under the RR mechanism.
- Maximum likelihood estimate, if everyone responds truthfully:

$$\hat{\mu}_{\mathrm{T}} = \frac{1}{N}\sum_{i=1}^{N} X_{\mathrm{T}}^{(i)}$$

Differential privacy by examples

Does randomization affect population-level statistics?

- How accurately can we estimate μ, the population mean?

Truthful Case:

- Let $X_{\mathrm{T}}^{(i)}$ denote individual i's response if they respond truthfully, and $X_{\mathrm{R}}^{(i)}$ individual i's response under the RR mechanism.
- Maximum likelihood estimate, if everyone responds truthfully:

$$\hat{\mu}_{\mathrm{T}} = \frac{1}{N}\sum_{i=1}^{N} X_{\mathrm{T}}^{(i)}$$

RR Case:

- How to estimate μ from the randomized responses $\{X_{\mathrm{R}}^{(i)}\}$?

$$\mathbb{E}[X_{\mathrm{R}}^{(i)}] = \frac{1}{4}(1-\mu_i) + \frac{3}{4}\mu_i$$

$$\Rightarrow \hat{\mu}_{\mathrm{R}} = \frac{2}{N}\sum_{i} X_{\mathrm{R}}^{(i)} - \frac{1}{2}$$

Differential privacy by examples

Does randomization affect population-level statistics?

- How accurately can we estimate μ, the population mean?

We can estimate the mean μ for the RR case. What is the payoff?

A. Estimate accuracy

B. Variance

C. Computational complexity

D. Both A and B

E. Both A, B, and C

Differential privacy by examples

Does randomization affect population-level statistics?

- How accurately can we estimate μ, the population mean?

We can estimate the mean μ for the RR case. What is the payoff?

- Variance of the ML estimate:

$$\begin{aligned}\mathsf{Var}(\hat{\mu}_{\mathrm{T}}) &= \frac{1}{N}\mathsf{Var}(X_{\mathrm{T}}^{(i)})\\ &= \frac{1}{N}\mu(1-\mu).\end{aligned}$$

- Variance of the estimator:

$$\begin{aligned}\mathsf{Var}(\hat{\mu}_{\mathrm{R}}) &= \frac{4}{N}\mathsf{Var}(X_{\mathrm{R}}^{(i)})\\ &\geq \frac{4}{N}\mathsf{Var}(X_{\mathrm{T}}^{(i)}) \quad \text{Why this inequality holds?}\\ &= 4\,\mathsf{Var}(\hat{\mu}_{\mathrm{T}})\end{aligned}$$

Differential privacy by examples

We can estimate the mean μ for the RR case. What is the payoff?

- Variance of the ML estimate:

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{\mathrm{T}}) &= \frac{1}{N}\mathrm{Var}(X_{\mathrm{T}}^{(i)}) \\ &= \frac{1}{N}\mu(1-\mu).\end{aligned}$$

- Variance of the estimator:

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{\mathrm{R}}) &= \frac{4}{N}\mathrm{Var}(X_{\mathrm{R}}^{(i)}) \\ &\geq \frac{4}{N}\mathrm{Var}(X_{\mathrm{T}}^{(i)}) \qquad \text{Why this inequality holds?} \\ &= 4\,\mathrm{Var}(\hat{\mu}_{\mathrm{T}})\end{aligned}$$

- The variance decays as $1/N$, which is good.
- But it is at least 4x larger because of the randomization. Can we do better?

Differential privacy

Basic setup:

- There is a database $\mathcal{D}$ which potentially contains sensitive information about individuals.
- The database curator has access to the full database. We assume the curator is trusted.
- The data analyst wants to analyze the data. She asks a series of queries to the curator, and the curator provides a response to each query.
- The way in which the curator responds to queries is called the mechanism. We'd like a mechnism that gives helpful responses but avoids leaking sensitive information about individuals.

Database

Database curator

Database analyst

Differential privacy

- Two databases $\mathcal{D}_1$ and $\mathcal{D}_2$ are neighbouring if they agree except for a single entry.
- **Idea:** if the mechanism behaves nearly identically for $\mathcal{D}_1$ and $\mathcal{D}_2$, then an attacker can't tell whether $\mathcal{D}_1$ or $\mathcal{D}_2$ was used (and hence can't learn much about the individual).
- **Definition:**
    - A mechanism $\mathcal{M}$ is ε-differentially private if for any two neighbouring databases $\mathcal{D}_1$ and $\mathcal{D}_2$, and any set $\mathcal{R}$ of possible responses

$$\Pr(\mathcal{M}(\mathcal{D}_1) \in \mathcal{R}) \leq \exp(\varepsilon) \Pr(\mathcal{M}(\mathcal{D}_2) \in \mathcal{R}).$$

- **Note:** for small ε, $\exp(\varepsilon) \approx 1 + \varepsilon$.
- **A consequence:** for any possible response y,

$$\exp(-\varepsilon) \leq \frac{\Pr(\mathcal{M}(\mathcal{D}_1) = y)}{\Pr(\mathcal{M}(\mathcal{D}_2) = y)} \leq \exp(\varepsilon)$$

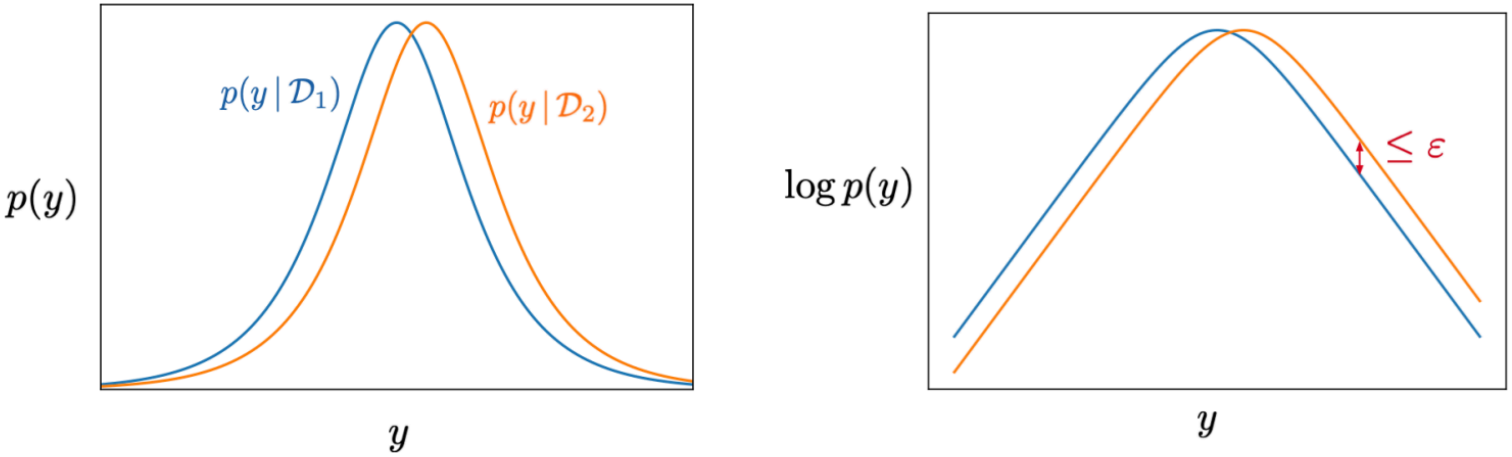
Differential privacy

- Two databases $\mathcal{D}_1$ and $\mathcal{D}_2$ are neighbouring if they agree except for a single entry.
- **Idea:** if the mechanism behaves nearly identically for $\mathcal{D}_1$ and $\mathcal{D}_2$, then an attacker can't tell whether $\mathcal{D}_1$ or $\mathcal{D}_2$ was used (and hence can't learn much about the individual).
- **Definition:**
  - A mechanism $\mathcal{M}$ is (ε, δ)-differential private if for any two neighbouring databases $\mathcal{D}_1$ and $\mathcal{D}_2$, and any set $\mathcal{R}$ of possible responses

$$\Pr(\mathcal{M}(\mathcal{D}_1) \in \mathcal{R}) \leq \exp(\varepsilon) \Pr(\mathcal{M}(\mathcal{D}_2) \in \mathcal{R}) + \delta \quad \text{(small, decreasing with } n\text{)}$$

- The randomness **only** comes from the randomized mechanism
- We may define "neighboring relationship" differently to encode different granularity of the DP guarantee: e.g., "Add/remove", "Replace"
- Need to hold for **any** pairs of neighboring inputs and **any** set of outputs

Differential privacy visualization



Notice that the tail behavior is important.

Differential privacy example

- Anna is an attacker who wants to figure out if Patrick (x) is in the cancer database $\mathcal{D}$. Her prior probability for him being in the database is 0.4. $\mathcal{D}$ is ε-differentially private. She makes a query and gets back $y = \mathcal{M}(\mathcal{D})$.

Differential privacy example

- Anna is an attacker who wants to figure out if Patrick (x) is in the cancer database $\mathcal{D}$. Her prior probability for him being in the database is 0.4. $\mathcal{D}$ is ε-differentially private. She makes a query and gets back $y = \mathcal{M}(\mathcal{D})$.

- She's narrowed it down to two possible databases $\mathcal{D}_1$ and $\mathcal{D}_2$, which are identical except that $x \in \mathcal{D}_1$ and $x \notin \mathcal{D}_2$.

Differential privacy example

- Anna is an attacker who wants to figure out if Patrick (x) is in the cancer database $\mathcal{D}$. Her prior probability for him being in the database is 0.4. $\mathcal{D}$ is ε-differentially private. She makes a query and gets back $y = \mathcal{M}(\mathcal{D})$.

- She's narrowed it down to two possible databases $\mathcal{D}_1$ and $\mathcal{D}_2$, which are identical except that $x \in \mathcal{D}_1$ and $x \notin \mathcal{D}_2$.

- After observing y, she computes her posterior probability using Bayes' Rule:

$$\begin{aligned}
\Pr(x \in \mathcal{D} \,|\, y) &= \frac{\Pr(x \in \mathcal{D}) \Pr(y \,|\, x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \,|\, x \in \mathcal{D}) + \Pr(x \notin \mathcal{D}) \Pr(y \,|\, x \notin \mathcal{D})} \\
&\geq \frac{\Pr(x \in \mathcal{D}) \Pr(y \,|\, x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \,|\, x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D}) \Pr(y \,|\, x \in \mathcal{D})} \\
&= \frac{\Pr(x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D})} \\
&\geq 0.4 \exp(-\varepsilon)
\end{aligned}$$

- Similarly, $\Pr(x \in \mathcal{D} \,|\, y) \leq 0.4 \exp(\varepsilon)$. So Anna hasn't learned much about Patrick.

Differential privacy is composable

- In what sense does this definition guarantee privacy?
- Suppose a data analyst takes the result $y = \mathcal{M}(\mathcal{D})$ and further processes it with some algorithm f (without peeking at the data itself). Is it still private? Example: Stochastic Gradient Updates [1]
- Let $\mathcal{R}$ be a set of possible outputs, and $\mathcal{R}'$ be the pre-image under f, i.e. $\mathcal{R}' = \{y : f(y) \in \mathcal{R}\}$.

$$\begin{aligned}\Pr(f(\mathcal{M}(\mathcal{D}_1)) \in \mathcal{R}) &= \Pr(\mathcal{M}(\mathcal{D}_1) \in \mathcal{R}') \\ &\leq \exp(\varepsilon) \Pr(\mathcal{M}(\mathcal{D}_2) \in \mathcal{R}') \\ &= \exp(\varepsilon) \Pr(f(\mathcal{M}(\mathcal{D}_2)) \in \mathcal{R})\end{aligned}$$

- Hence, the composition $f \circ \mathcal{M}$ is also ε-differentially private. No matter how clever the analyst is, or the resources she throws at it, she can't learn more than ε about an individual entry!

[1] Deep Learning with Differential Privacy https://arxiv.org/pdf/1607.00133.pdf

Composition rules

- So far, we've been looking at one query in isolation. What if we want to answer more than one question from the data we've collected?
- Can't just repeatedly use the same mechanism independently
  - Suppose the analyst asks the same counting query K times, and the curator always responds independently using the Laplace mechanism.
  - The analyst can get arbitrarily accurate counts by averaging the responses, rendering the privacy guarantee meaningless.
- Can we relate the privacy of multiple queries to the privacy of a single query? Such a result is known as a composition rule.

Composition rules

- The easiest case is when the queries are non-adaptive, i.e. the analyst(s) make the queries without seeing the results of previous queries.
- **Claim:** Querying an ε-differentially private mechanism K times non-adaptively is $K\varepsilon$-differentially private.
- Letting y_1, y_2 be the responses, we have $y_1 \perp\!\!\!\perp y_2 \,|\, \mathcal{D}$. So,

$$\begin{aligned} \frac{p(y_1, y_2 \,|\, \mathcal{D}_1)}{p(y_1, y_2 \,|\, \mathcal{D}_2)} &= \frac{p(y_1 \,|\, \mathcal{D}_1)}{p(y_1 \,|\, \mathcal{D}_2)} \frac{p(y_2 \,|\, \mathcal{D}_1)}{p(y_2 \,|\, \mathcal{D}_2)} \\ &\leq \exp(\varepsilon) \cdot \exp(\varepsilon) \\ &= \exp(2\varepsilon) \end{aligned}$$

- **Corrollary:** if your privacy budget is ε, you should make sure the privacy parameters of the individual queries sum up to ε.

Composition rules

- **Example:** Recall that for naïve Bayes, we made a counting query that requests the joint counts of (t, x_j) for each feature x_j.
    - We concluded that $\Delta f = D$, so the Laplace mechanism adds Laplace noise with scale D/ε.
- We can alternatively formulate this as D different queries, chosen non-adaptively, each of which asks for the joint counts (t, x_j) for *one* feature x_j.
    - To satisfy a privacy budget of ε, each query should be $\frac{\varepsilon}{D}$-differentially private.
    - The sensitivity of each query is $\Delta f_j = 1$.
    - So we should add Laplace noise with scale $\Delta f_j / (\varepsilon / D) = D/\varepsilon$.
- Hence, the composition rule agrees with the basic Laplace mechanism for this example.

Advantages and disadvantages of DP

- Advantages
  - A formal mathematical definition of privacy that provides rigorous guarantees and provably effective protections against privacy risks; makes no assumptions on adversary, database, etc
  - The de-facto standard in privacy --- the only one still being actively researched on
  - Composable in industrial applications
  - Interpretable, quantifiable, composable formalism
- Disadvantages
  - Sometimes too restrict

Part 1.3

Basic Mechanism

Laplace mechanism

Exponential mechanism

Laplace mechanism

How to find a mechanism for differential privacy?

The first mechanism designed together with DP: Laplace mechanism

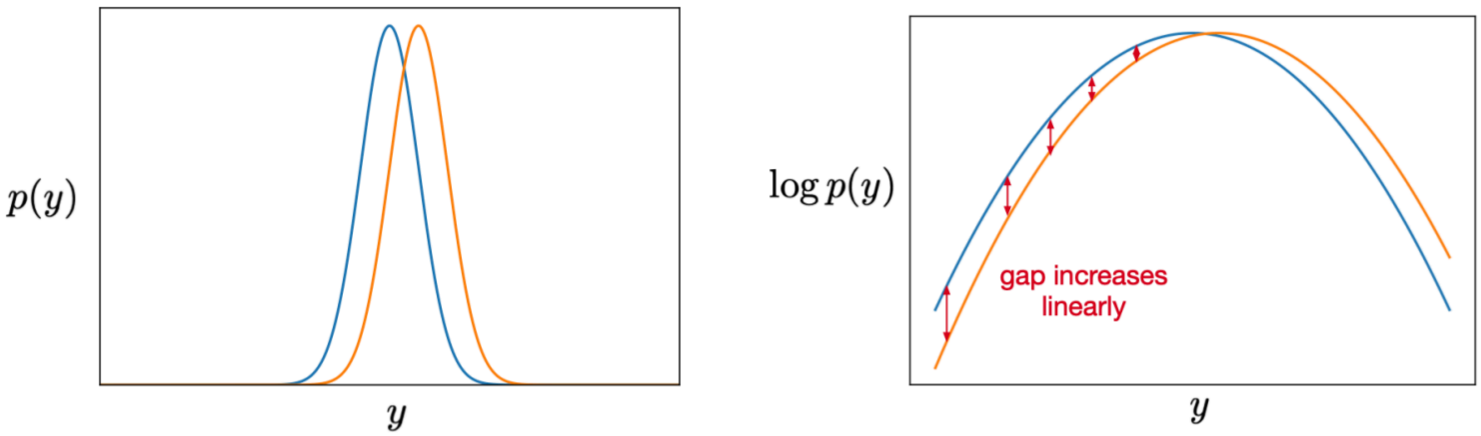
- A lot of queries we might want to ask can be seen as counting queries, i.e. counting the number of entries which have property $\mathcal{P}$.
  - E.g. naive Bayes, decision trees
- **Idea:** Maybe the mechanism can return noisy counts which are accurate enough for whatever analysis we're trying to do.

We focus on ε-differential privacy

Laplace mechanism

What kind of noise we would like to add to the counts?

First trial: Gaussian

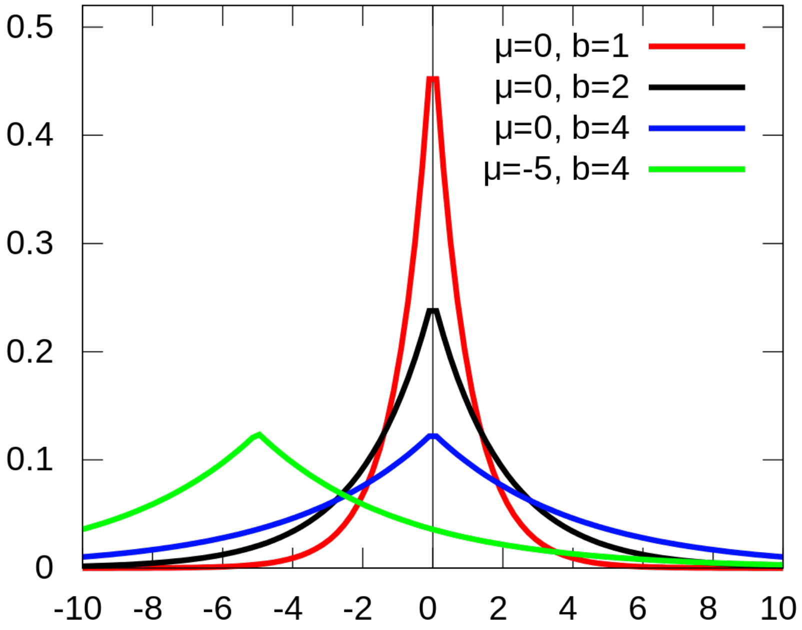


What: $\exp(-\varepsilon) \leq \dfrac{\Pr(\mathcal{M}(\mathcal{D}_1) = y)}{\Pr(\mathcal{M}(\mathcal{D}_2) = y)} \leq \exp(\varepsilon)$

Laplace mechanism

What kind of noise we would like to add to the counts?

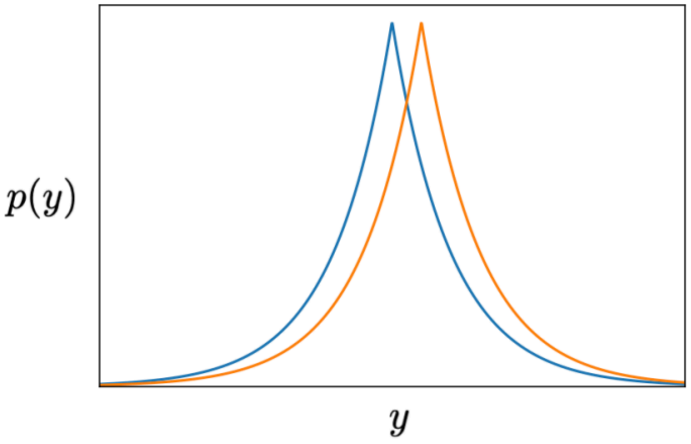
Second trial: Laplace distribution $p(y; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right)$

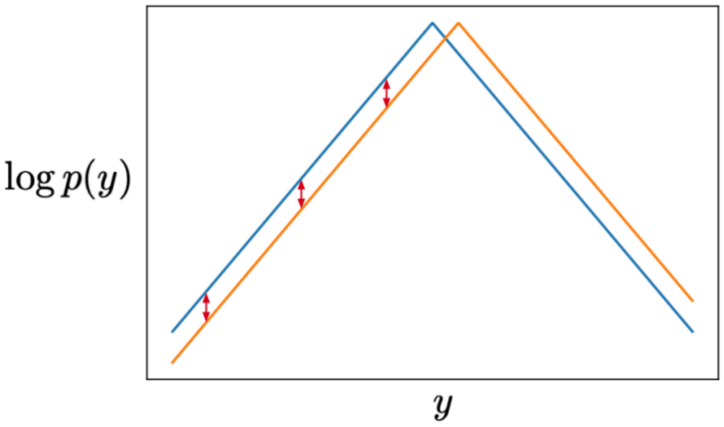


Laplace mechanism

What kind of noise we would like to add to the counts?

Second trial: Laplace distribution $p(y;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|y-\mu|}{b}\right)$





What: $\exp(-\varepsilon) \leq \frac{\Pr(\mathcal{M}(\mathcal{D}_1)=y)}{\Pr(\mathcal{M}(\mathcal{D}_2)=y)} \leq \exp(\varepsilon)$

Laplace mechanism

- Let f be a deterministic vector-valued function of a database. The L^1 sensitivity of f is defined as:

$$\Delta f = \max_{\substack{\mathcal{D}_1, \mathcal{D}_2 \\ \text{neighbours}}} \|f(\mathcal{D}_1) - f(\mathcal{D}_2)\|_1.$$

- Recall that $\|\mathbf{x}\|_1 = \sum_i |x_i|$.
- Suppose f returns the vector of counts of individuals who fall into k disjoint buckets. What is the L^1 sensitivity of f? (Ans: 1)

- Laplace mechanism: return a vector $\mathbf{y}$ whose entries are independently sampled from Laplace distributions

$$y_i \sim \mathrm{Laplace}\left(f(\mathcal{D})_i, \frac{\Delta f}{\varepsilon}\right),$$

where $f(\mathcal{D})_i$ denotes the ith entry of $f(\mathcal{D})$.

In other words: $f(\mathcal{D}) + Z$ where $Z_i \sim \mathrm{Lap}(\Delta f / \epsilon)$ i.i.d.

Laplace mechanism is differentially private

- **Claim:** the Laplace mechanism is differentially private.
- Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two neighboring databases, and $y = \mathcal{M}(\mathcal{D})$.

$$
\begin{aligned}
\frac{p(\mathbf{y}\,|\,\mathcal{D}_1)}{p(\mathbf{y}\,|\,\mathcal{D}_2)} &= \frac{\prod_i \frac{\varepsilon}{2\Delta f}\exp(-\frac{\varepsilon|f(\mathcal{D}_1)_i-y_i|}{\Delta f})}{\prod_{i=1}^k \frac{\varepsilon}{2\Delta f}\exp(-\frac{\varepsilon|f(\mathcal{D}_2)_i-y_i|}{\Delta f})} \\
&= \prod_i \exp\left(\frac{\varepsilon(|f(\mathcal{D}_2)_i - y_i| - |f(\mathcal{D}_1)_i - y_i|)}{\Delta f}\right) \\
&\leq \prod_i \exp\left(\frac{\varepsilon(|f(\mathcal{D}_2)_i - f(\mathcal{D}_1)_i|)}{\Delta f}\right) && \text{(triangle ineq.)} \\
&= \exp\left(\frac{\varepsilon\sum_i |f(\mathcal{D}_2)_i - f(\mathcal{D}_1)_i|}{\Delta f}\right) \\
&= \exp\left(\frac{\varepsilon\|f(\mathcal{D}_2) - f(\mathcal{D}_1)\|_1}{\Delta f}\right) \\
&\leq \exp(\varepsilon) && \text{(defn. of } \Delta f\text{)}
\end{aligned}
$$

Laplace mechanism is differentially private

- **Example:** What fraction of Canadians have blue eyes?
- Mechanism returns the counts (ξ_1, ξ_2) of Canadians with and without blue eyes, plus Laplace noise. We'd like to satisfy a privacy constraint of $\varepsilon = 0.1$. How much Laplace noise should we add?
    - Ans: $\Delta f / \varepsilon = 1/0.1 = 10$.
- The noise scale is independent of the population size!
- I.e., you can answer the query to within about ± 10 people, out of the population of Canada. So you can obtain very accurate answers to queries over large populations.

Laplace mechanism is differentially private

Comparison to randomized response

- Recall the randomized response method:

	Yes	No
Dodge	3/4	1/4
No Dodge	1/4	3/4

- For what ε is this ε-differentially private? (Ans: $\log 3$)

- **Recall:** ML estimate from truthful responses has variance $\frac{1}{N}\mu(1-\mu)$ and estimate from randomized responses has variance at least 4x larger.

Laplace mechanism is differentially private

- **Recall:** ML estimate from truthful responses has variance $\frac{1}{N}\mu(1-\mu)$ and estimate from randomized responses has variance at least 4x larger.

- **Laplace mechanism:** add Laplace noise η with scale $\Delta f/\varepsilon = 1/\log 3 \approx 0.91$

$$\begin{aligned}\hat{\mu}_{\mathrm{L}} &= \frac{1}{N}\Big(\sum_{i=1}^{N} X_{\mathrm{T}}^{(i)} \;+\; \eta\Big)\\ &= \hat{\mu}_{\mathrm{T}} + \frac{\eta}{N}\end{aligned}$$

- The added noise has variance $\mathcal{O}(1/N^2)$, compared with the statistical error, which is $\mathcal{O}(1/N)$. So we lose almost no accuracy.

Exponential mechanism

- Suppose the goal of the analysis is to make a decision Y.
- We have a loss function $\mathcal{L}(Y, \mathcal{D})$ which determines how unhappy we are with any particular Y as a response for database $\mathcal{D}$.
- The exponential mechanism tries to pick a reasonably good decision subject to a privacy constraint. We do this by picking Y randomly as:

$$\Pr(Y = y) \propto \exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y, \mathcal{D})\right)$$

- $\Delta\mathcal{L}$ is the sensitivity of $\mathcal{L}$, just like for the Laplace mechanism.
- The resulting probabilities are basically a softmax of $-\mathcal{L}$. Distributions of this form are also called Boltzmann distributions (from statistical mechanics).

Exponential mechanism is differentially private

- **Claim:** The exponential mechanism is ε-differentially private.
- For two neighboring databases $\mathcal{D}_1$ and $\mathcal{D}_2$, and any value y,

$$\frac{p(y \mid \mathcal{D}_1)}{p(y \mid \mathcal{D}_2)} = \frac{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_1)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_1)\right)}}{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_2)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_2)\right)}}$$

$$= \underbrace{\frac{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_1)\right)}{\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y,\mathcal{D}_2)\right)}}_{\leq \exp(\varepsilon/2)} \cdot \underbrace{\frac{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_2)\right)}{\sum_{y'}\exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(y',\mathcal{D}_1)\right)}}_{\leq \exp(\varepsilon/2)}$$

- Both inequalities are straightforward applications of the definition of $\Delta\mathcal{L}$.
- Hence, $\frac{p(y \mid \mathcal{D}_1)}{p(y \mid \mathcal{D}_2)} \leq \exp(\varepsilon)$, so we're done.

Exponential mechanism example

- **Example:** inferring the parameter of a Bernoulli distribution
- Suppose we have a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ of coin flips, and we want to estimate the bias θ while protecting the privacy of each individual coin flip with $\varepsilon = 0.1$.
- Our loss is negative log-likelihood:

$$\mathcal{L}(\hat{\theta}, \mathcal{D}) = -\log \prod_{i=1}^{N} p(x_i; \hat{\theta})$$

- What is the sensitivity $\Delta\mathcal{L}$?
  - Ans: $\Delta\mathcal{L} = \infty$, because an observation $x_i = 1$ has probability 1 under $\hat{\theta} = 1$ and probability 0 under $\hat{\theta} = 0$.
  - Hence, we can't use the exponential mechanism without further assumptions.

Exponential mechanism example

- Now suppose we restrict $\hat{\theta}$ to be in the interval $(0.1, 0.9)$. Now what is the sensitivity?
  - Ans: $\Delta\mathcal{L} = -\log 0.1 \approx 2.3$.

- The exponential mechanism samples $\hat{\theta}$ as

$$\begin{aligned} p(\hat{\theta} \,|\, \mathcal{D}) &\propto \exp\left(-\frac{\varepsilon}{2\Delta\mathcal{L}}\mathcal{L}(\hat{\theta}, \mathcal{D})\right) \\ &= \exp\left(0.022 \; \log \prod_{i=1}^{N} p(x_i; \hat{\theta})\right) \\ &= \prod_{i=1}^{N} p(x_i; \hat{\theta})^{0.022} \\ &= \hat{\theta}^{0.022 N_H}(1 - \hat{\theta})^{0.022 N_T} \end{aligned}$$

- **Note:** This is a beta distribution with parameters $a = 1 + 0.022\, N_H$ and $b = 1 + 0.022\, N_T$, truncated to $(0.1, 0.9)$.

Comparison of Laplace and Exponential mechanisms

- Let's compare the Laplace and exponential mechanisms for estimating $\hat{\theta}$.

- **Laplace mechanism**: compute the counts N_H and N_T, then add Laplace noise with scale $\Delta\mathcal{L}/\varepsilon = 22$.
  - $\hat{\theta} = \frac{\hat{N}_H}{\hat{N}_H + \hat{N}_T}$
  - Can show $\mathrm{Var}(\hat{\theta} \mid \mathcal{D}) = \mathcal{O}(1/N^2)$

- **Exponential mechanism**: $\hat{\theta} \sim \mathrm{TruncatedBeta}(1 + 0.022\, N_H, 1 + 0.022\, N_T)$
  - Can show $\mathrm{Var}(\hat{\theta} \mid \mathcal{D}) = \mathcal{O}(1/N)$

So the Laplace mechanism is much more accurate in this case. But the exponential mechanism is still useful in cases that aren't easily formulated as counts

Part II

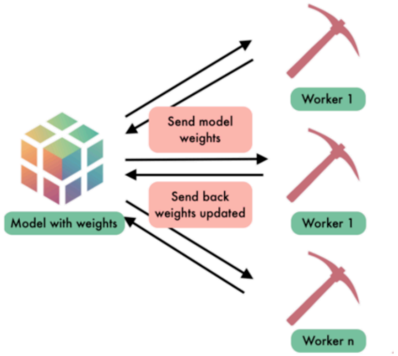
Federated Learning

Introduction to federated learning

- So far, we've assumed there's a curator who we trust with access to all the raw data.
- What if a company (say Google) wants to learn a classifier from the images stored on everyone's phones, but without having to send the images to Google?
- Federated learning: learning a model without any centralized entity having access to all the data
  - Google sends the phone the current weights of the network
  - The phone does a small number of steps of gradient descent, and communicates the local update back to Google
  - Google updates their network by adding the local update
- Does this satisfy differential privacy?
  - Not automatically, but the local updates could be randomized in a way that makes them differentially private.

Introduction to federated learning

A typical round of learning consists of the following sequence.

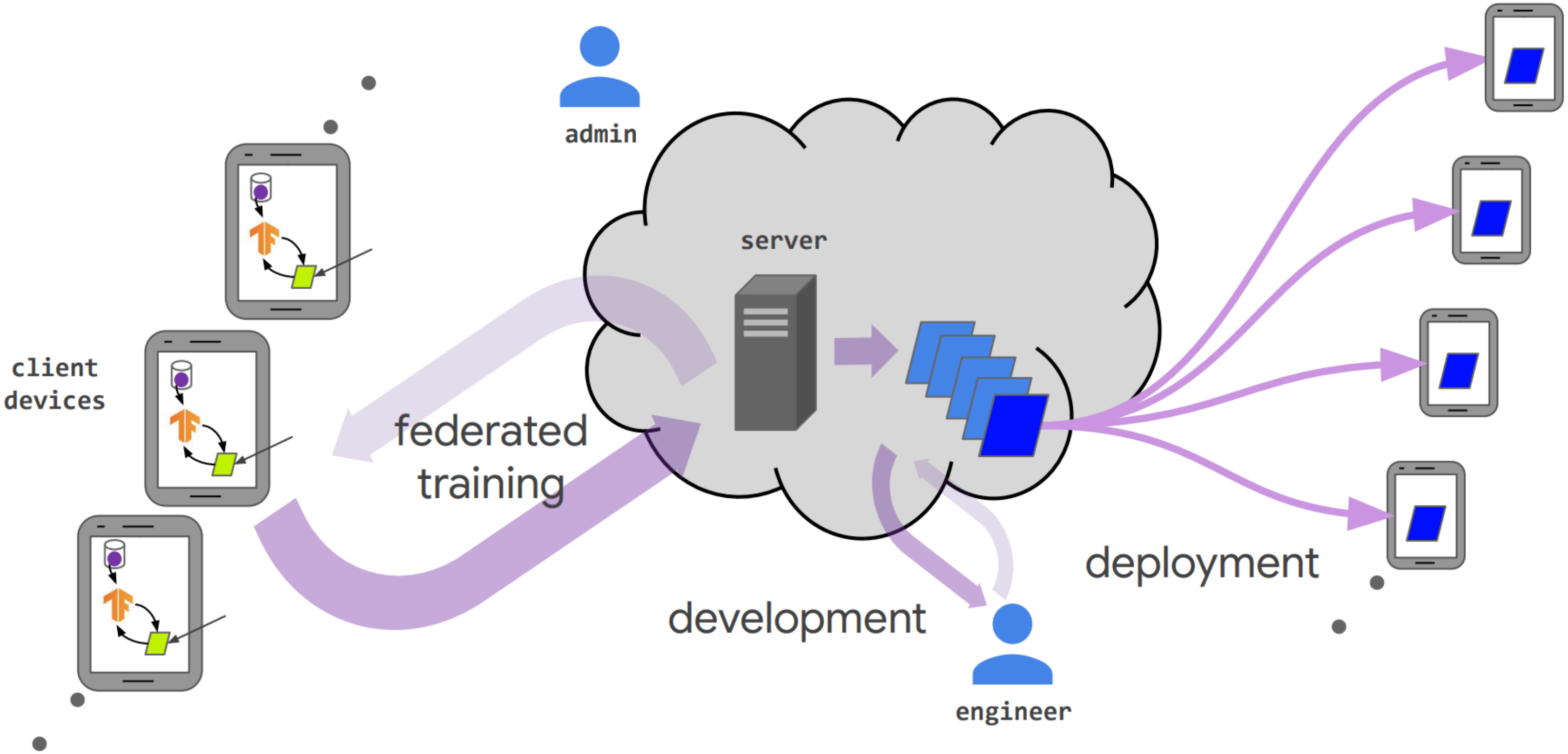


1. A random subset of members of the Federation (known as clients) is selected to receive the global model synchronously from the server.
2. Each selected client computes an updated model using its local data.
3. The model updates are sent from the selected clients to the server.
4. The server aggregates these models (typically by averaging) to construct an improved global model.

Normal assumptions in FL

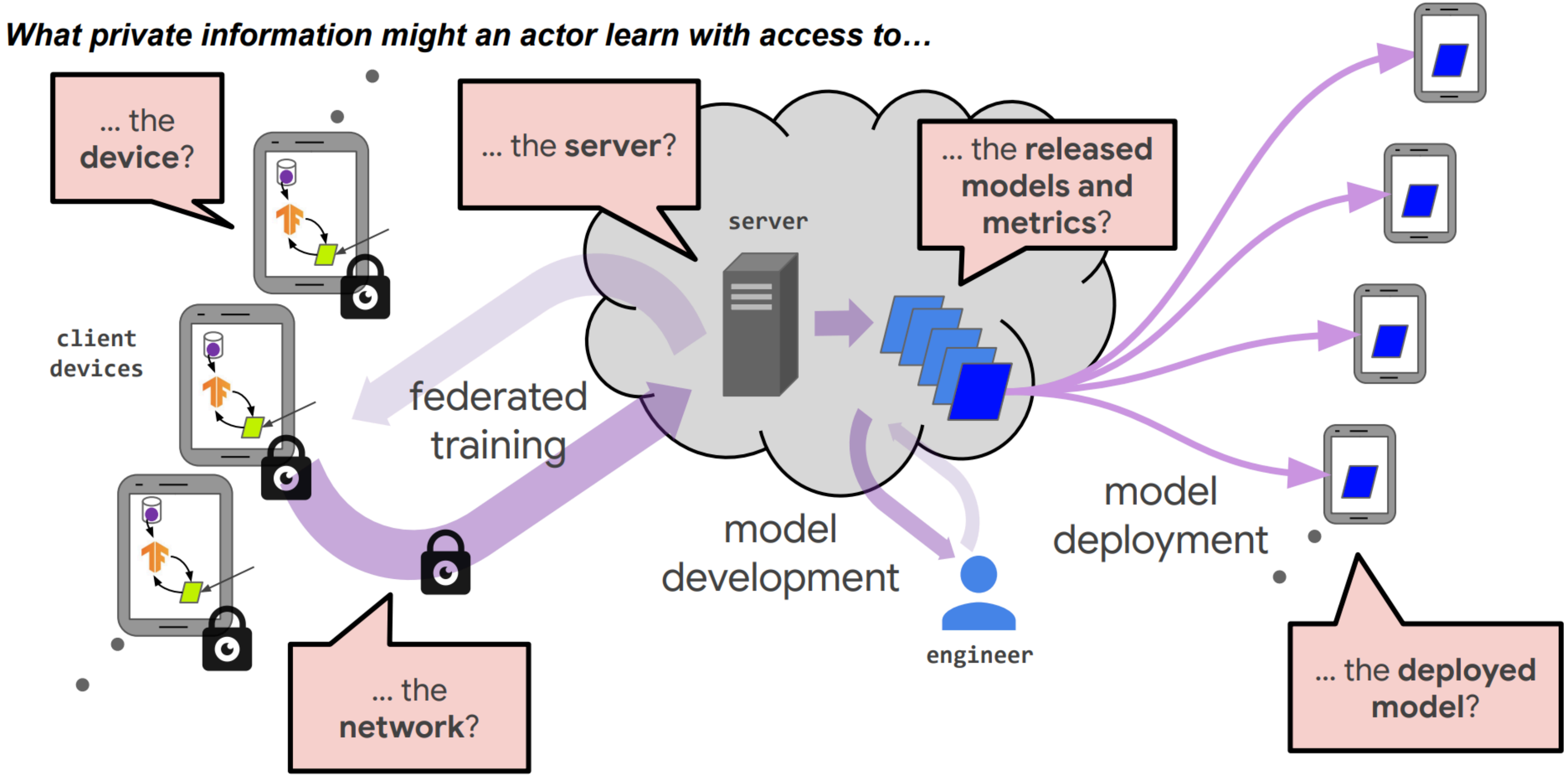
- Distributed storage (Non-IID)
  - User data is localized to their own usage
  - Hard to be a representative of the population
- Heterogeneous services (Unbalanced)
  - Some users will make much heavier on particular services than others
- Distributed computing capacity (Massively distributed)
  - Expect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

Introduction to federated learning (Optional)



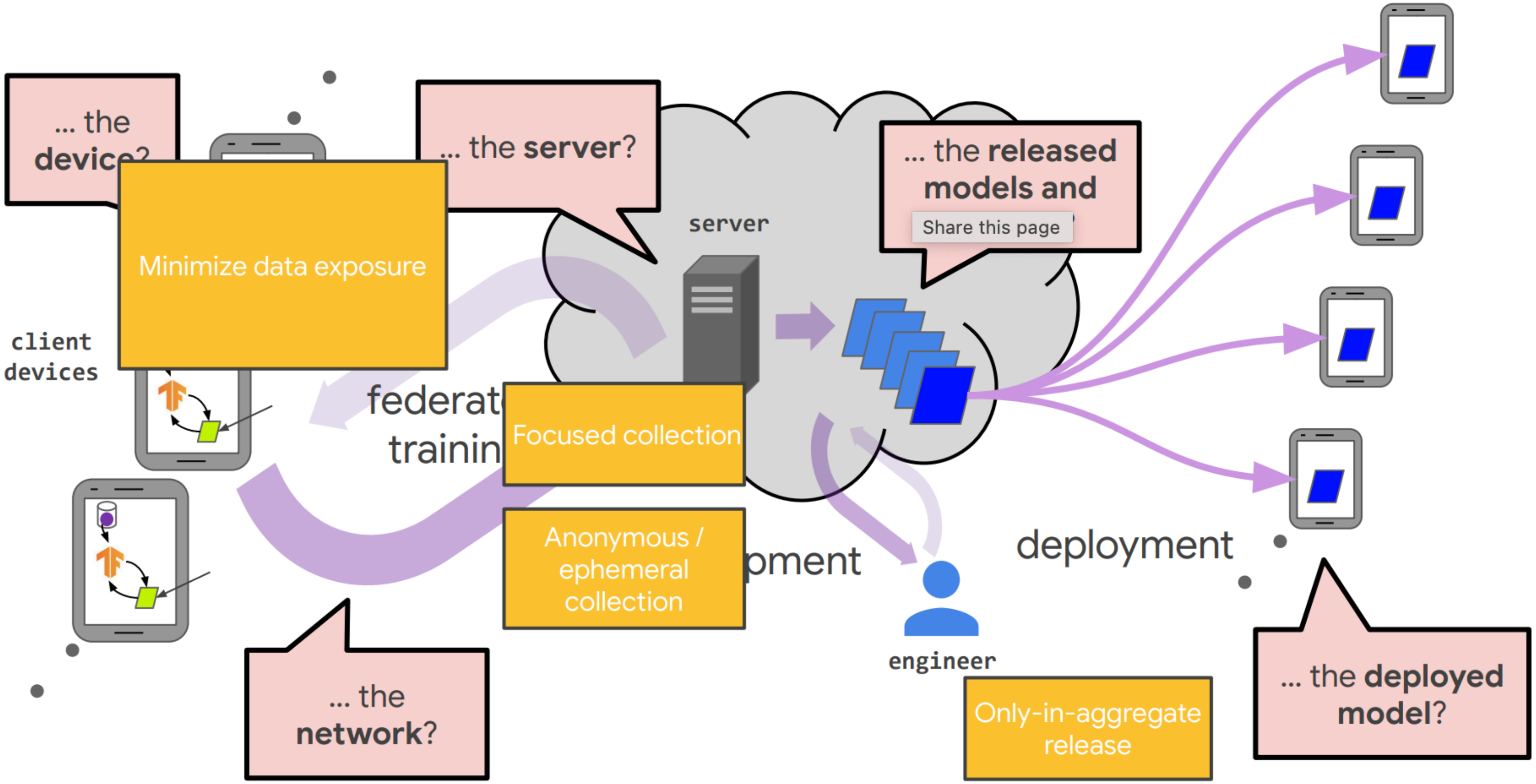
https://sites.cs.ucsb.edu/~yuxiangw/classes/DPCourse-2021Fall/Lectures/peter_kairouz_guest_lecture.pdf

Ensuring privacy of participating users (Optional)

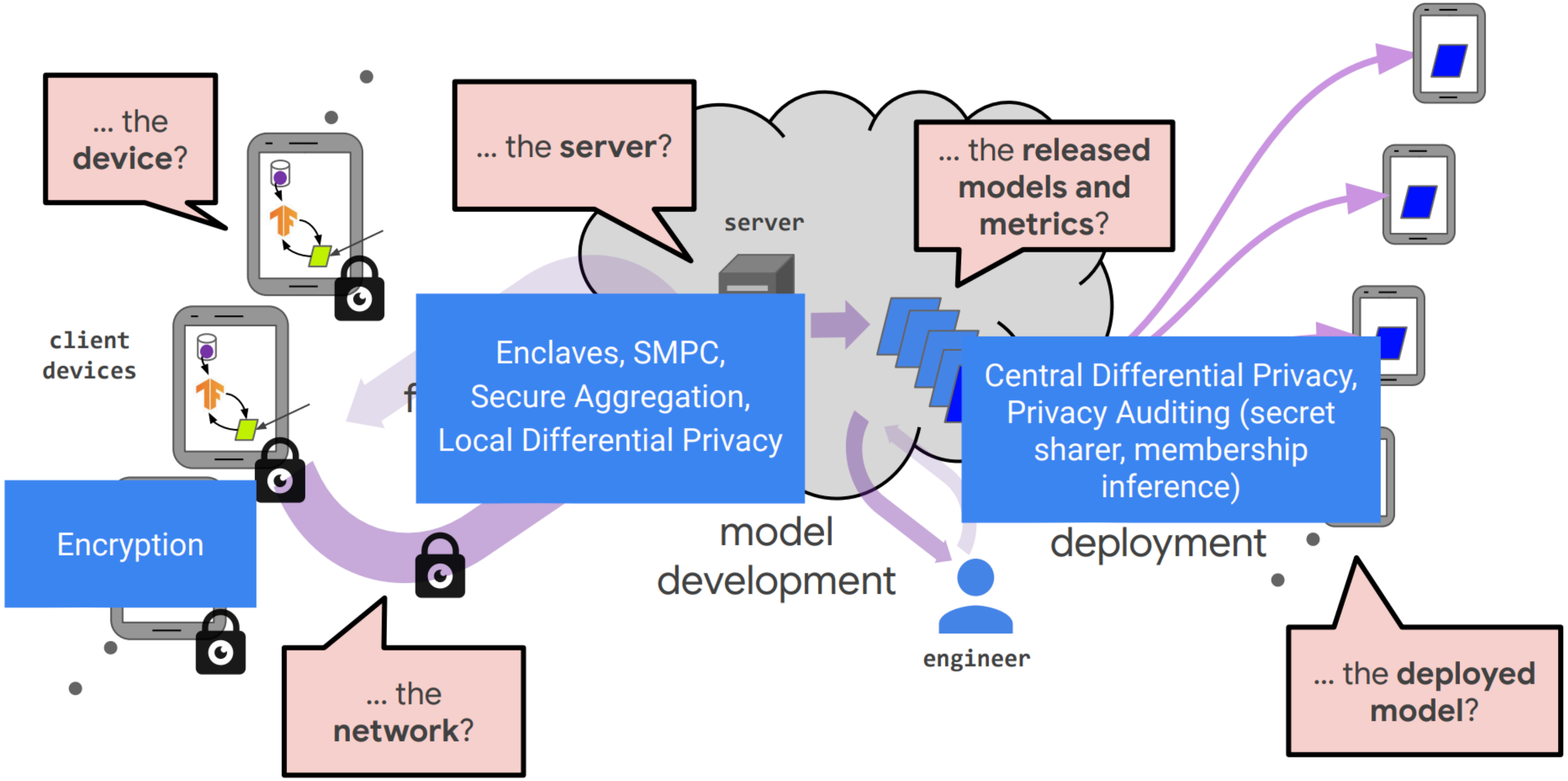


What private information might an actor learn with access to…

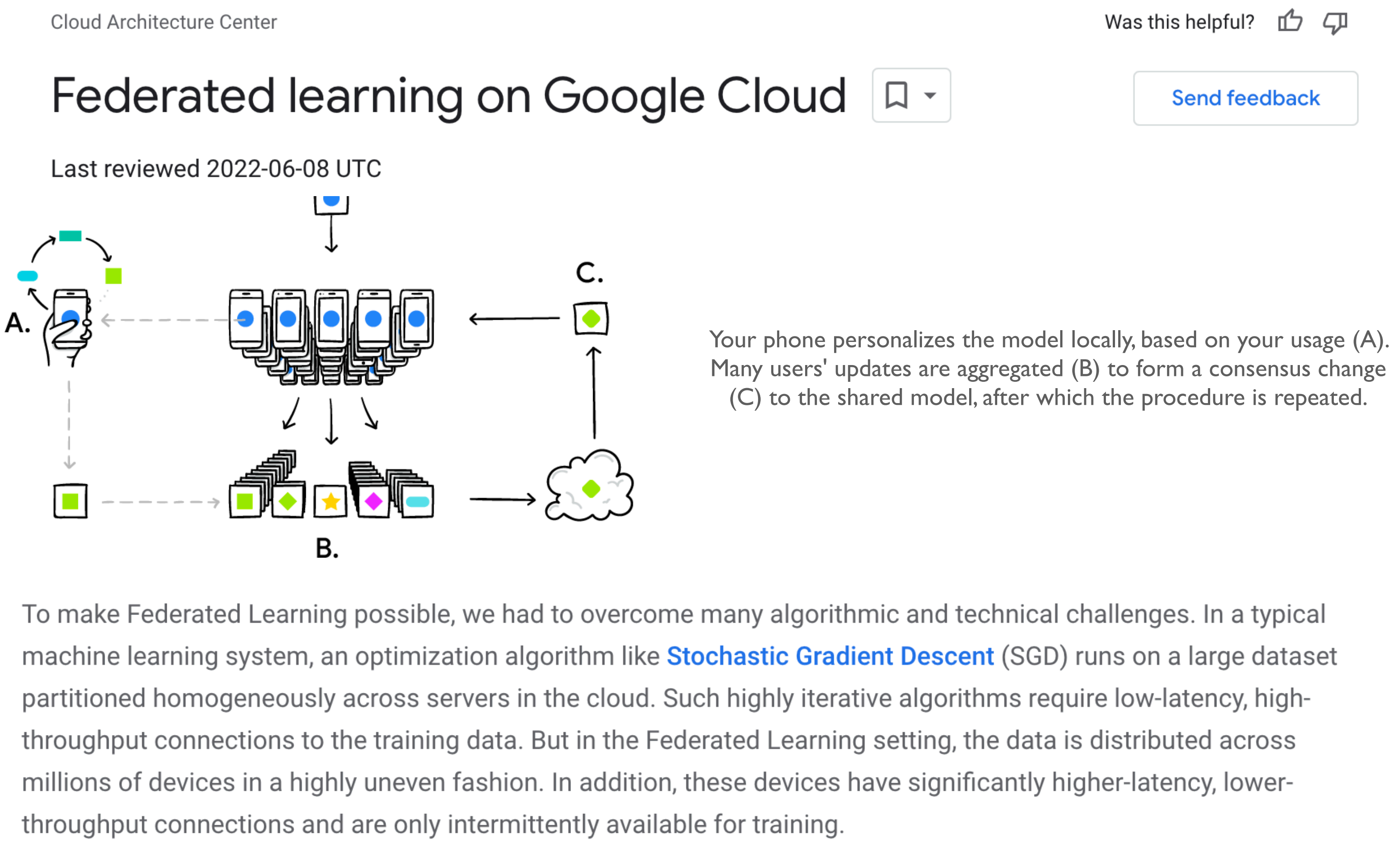
Data minimization principles for FL (Optional)



Need complementary privacy technologies (Optional)


... the **device**?
client devices
... the **server**?
server
... the **released models and metrics**?
Enclaves, SMPC, Secure Aggregation, Local Differential Privacy
Central Differential Privacy, Privacy Auditing (secret sharer, membership inference)
Encryption
model development
deployment
engineer
... the **network**?
... the **deployed model**?

FL used by Google Cloud (Optional)



Cloud Architecture Center

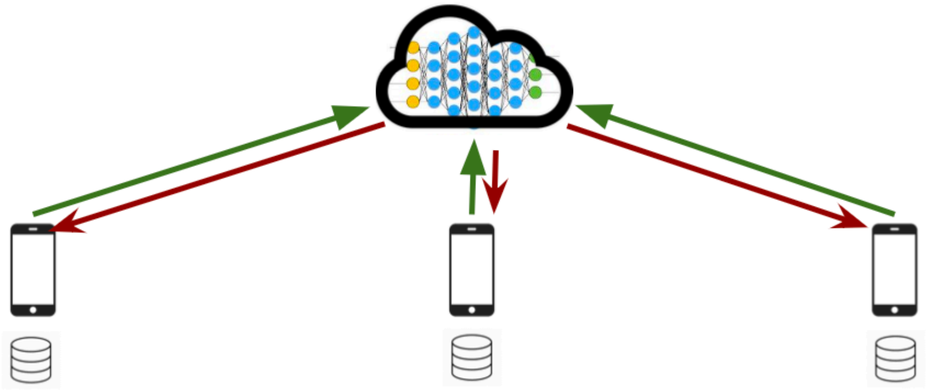
Federated learning on Google Cloud

Last reviewed 2022-06-08 UTC

Your phone personalizes the model locally, based on your usage (A). Many users' updates are aggregated (B) to form a consensus change (C) to the shared model, after which the procedure is repeated.

To make Federated Learning possible, we had to overcome many algorithmic and technical challenges. In a typical machine learning system, an optimization algorithm like **Stochastic Gradient Descent** (SGD) runs on a large dataset partitioned homogeneously across servers in the cloud. Such highly iterative algorithms require low-latency, high-throughput connections to the training data. But in the Federated Learning setting, the data is distributed across millions of devices in a highly uneven fashion. In addition, these devices have significantly higher-latency, lower-throughput connections and are only intermittently available for training.

FedSGD (Optional)



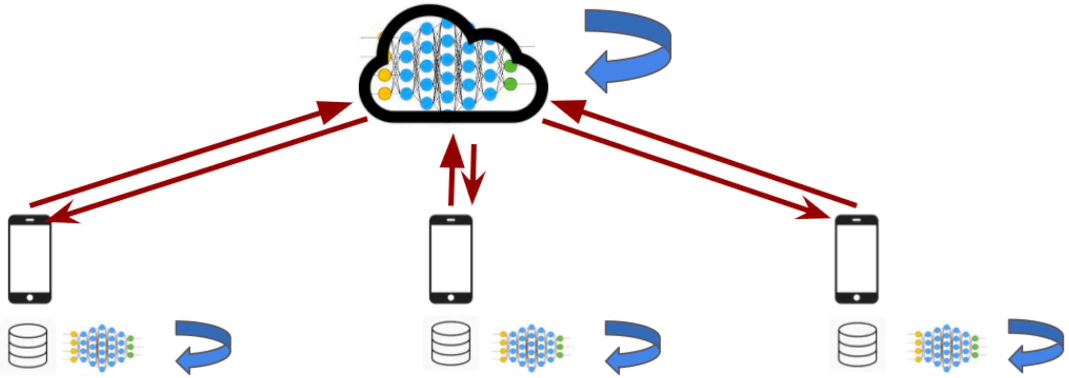
🟩 Gradient
🟥 Model

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

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

McMahan et al, 2017 Communication-Efficient Learning of Deep Networks from Decentralized Data: https://arxiv.org/pdf/1602.05629.pdf

FedAvg (Optional)



🟩 Gradient

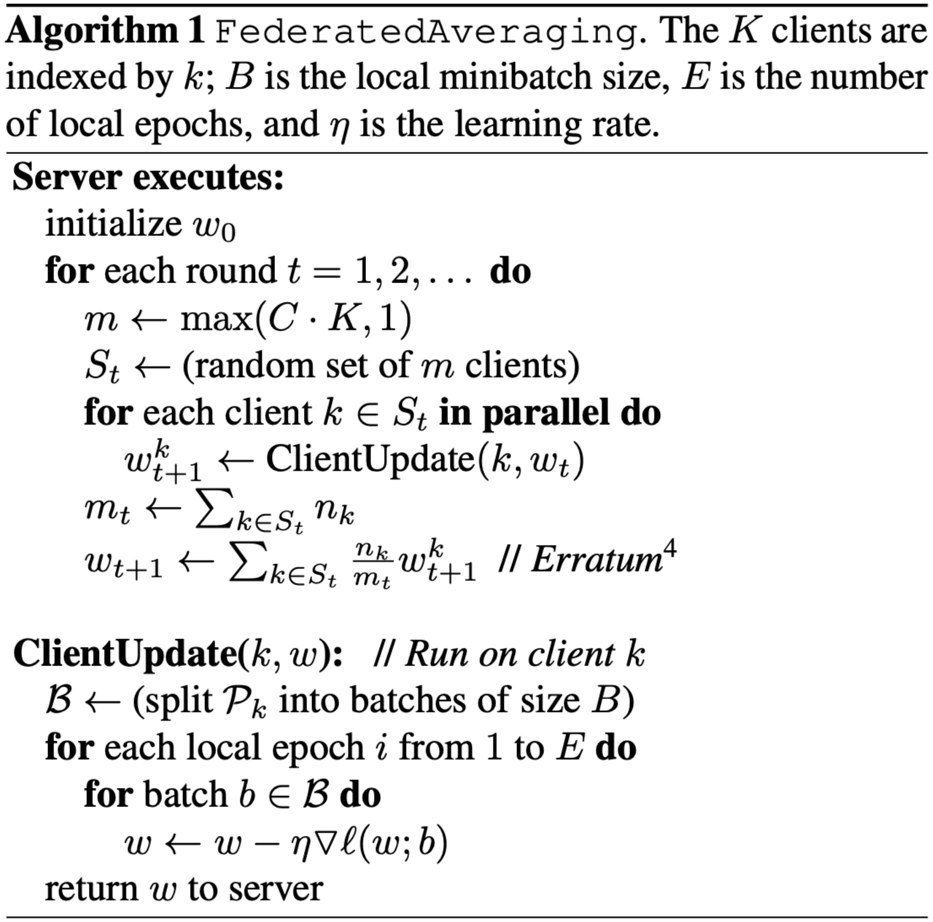
🟥 Model

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

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

McMahan et al, 2017 Communication-Efficient Learning of Deep Networks from Decentralized Data: https://arxiv.org/pdf/1602.05629.pdf

FedAvg (Optional)



Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

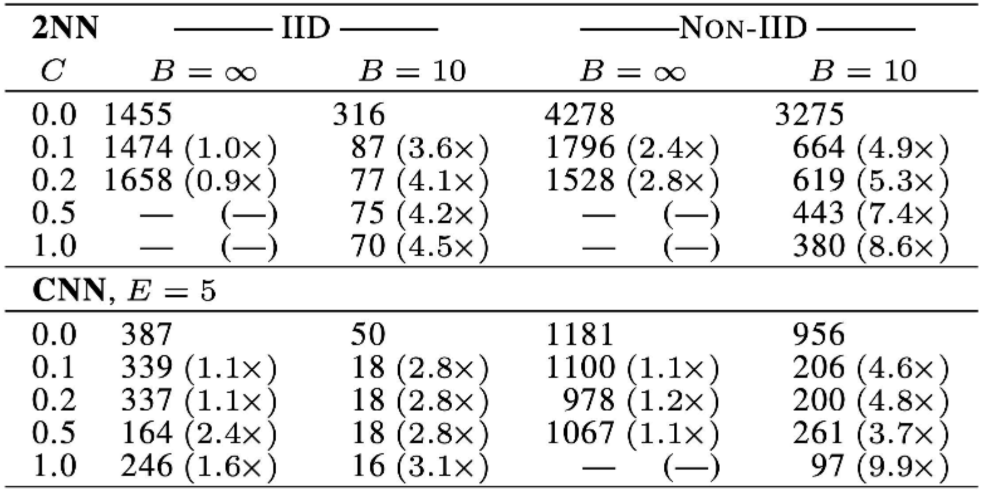
```
Server executes:
  initialize $w_0$
  for each round $t = 1, 2, \dots$ do
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of $m$ clients)
    for each client $k \in S_t$ in parallel do
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $m_t \leftarrow \sum_{k \in S_t} n_k$
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$  // Erratum[4]

ClientUpdate($k, w$):   // Run on client k
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size $B$)
  for each local epoch $i$ from 1 to $E$ do
    for batch $b \in \mathcal{B}$ do
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return $w$ to server
```

McMahan et al, 2017 Communication-Efficient Learning of Deep Networks from Decentralized Data: https://arxiv.org/pdf/1602.05629.pdf

Trade-offs Between Local and Global Iterations (Optional)

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



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

C - ratio of clients updated to the server

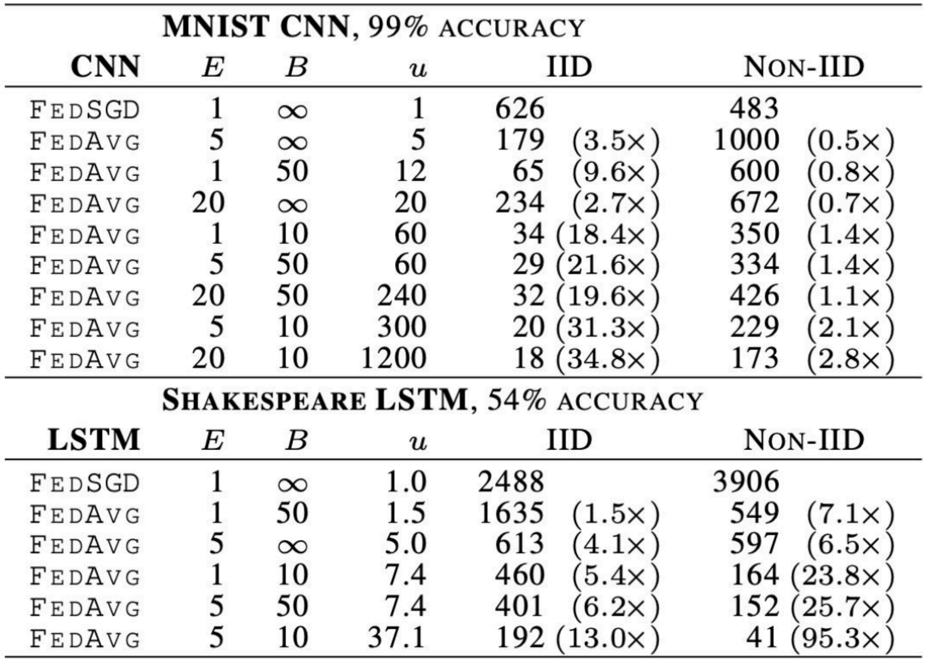
B - batch size of clients

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

McMahan et al, 2017 Communication-Efficient Learning of Deep Networks from Decentralized Data: https://arxiv.org/pdf/1602.05629.pdf

Comparisons Between FedSGD and FedAvg (Optional)

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



MNIST CNN, 99% accuracy					
CNN	E	B	u	IID	Non-IID
FedSGD	1	∞	1	626	483
FedAvg	5	∞	5	179 (3.5×)	1000 (0.5×)
FedAvg	1	50	12	65 (9.6×)	600 (0.8×)
FedAvg	20	∞	20	234 (2.7×)	672 (0.7×)
FedAvg	1	10	60	34 (18.4×)	350 (1.4×)
FedAvg	5	50	60	29 (21.6×)	334 (1.4×)
FedAvg	20	50	240	32 (19.6×)	426 (1.1×)
FedAvg	5	10	300	20 (31.3×)	229 (2.1×)
FedAvg	20	10	1200	18 (34.8×)	173 (2.8×)

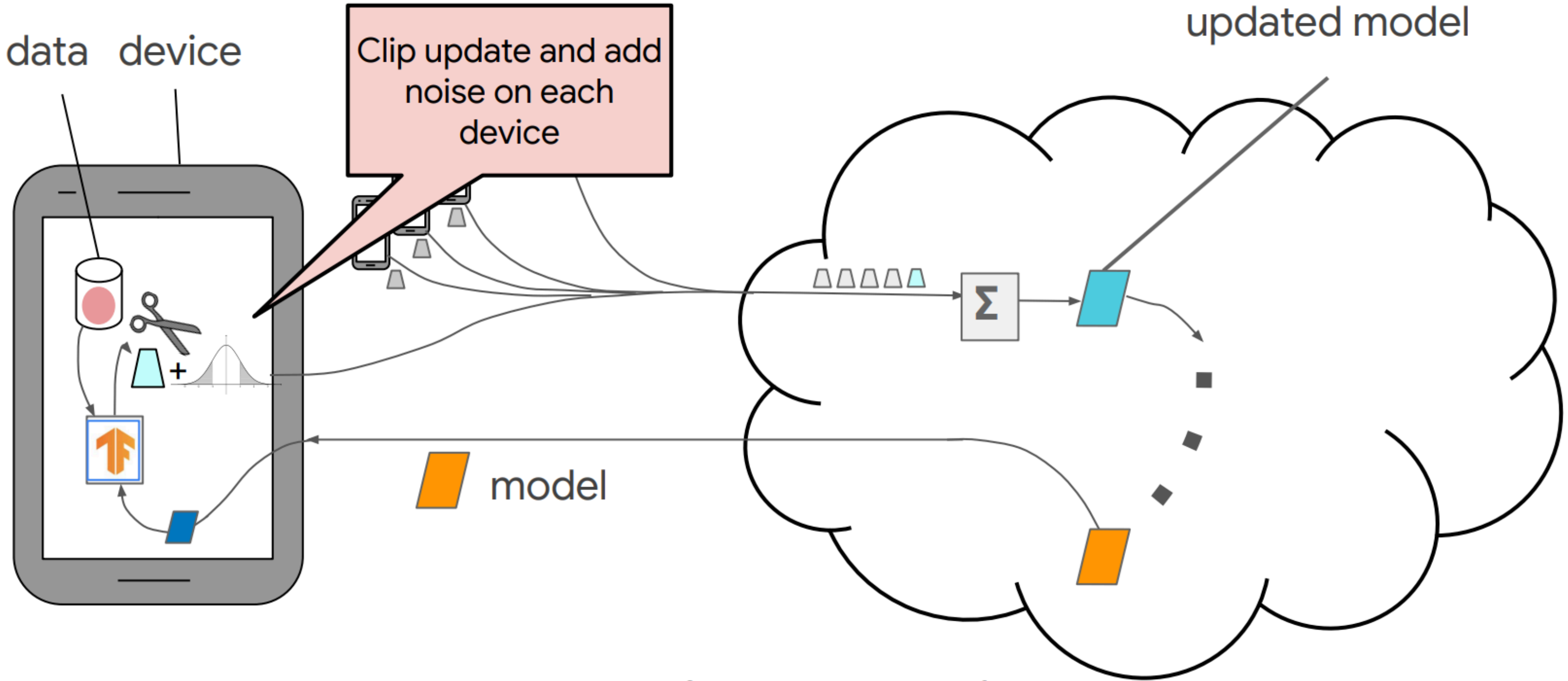
Shakespeare LSTM, 54% accuracy					
LSTM	E	B	u	IID	Non-IID
FedSGD	1	∞	1.0	2488	3906
FedAvg	1	50	1.5	1635 (1.5×)	549 (7.1×)
FedAvg	5	∞	5.0	613 (4.1×)	597 (6.5×)
FedAvg	1	10	7.4	460 (5.4×)	164 (23.8×)
FedAvg	5	50	7.4	401 (6.2×)	152 (25.7×)
FedAvg	5	10	37.1	192 (13.0×)	41 (95.3×)

K - number of clients
B - batch size
E - number of epochs
u - $En/(KB)$

expected number of updates per round.

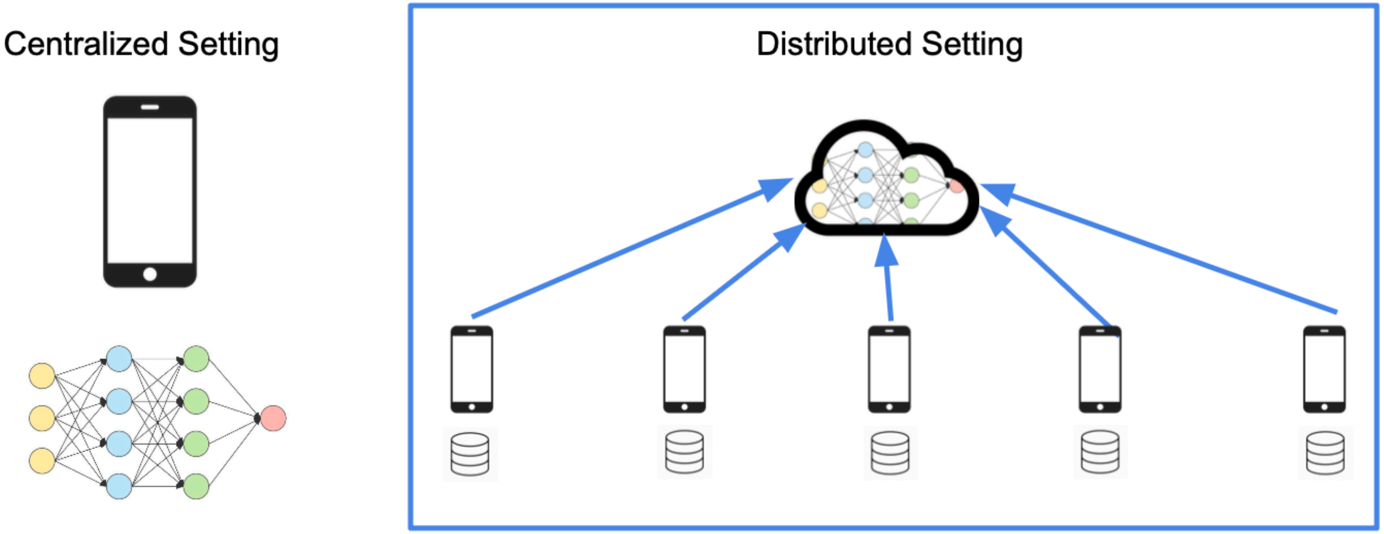
McMahan et al, 2017 Communication-Efficient Learning of Deep Networks from Decentralized Data: https://arxiv.org/pdf/1602.05629.pdf

Locally differentially private federated training (Optional)



Evfimievski, Alexandre, et al. ***Privacy preserving mining of association rules.*** *Information Systems 2004*
Warner, Stanley L. ***Randomized response: A survey technique for eliminating evasive answer bias.*** *JASA 1965*

Differential privacy and local differential privacy (Optional)

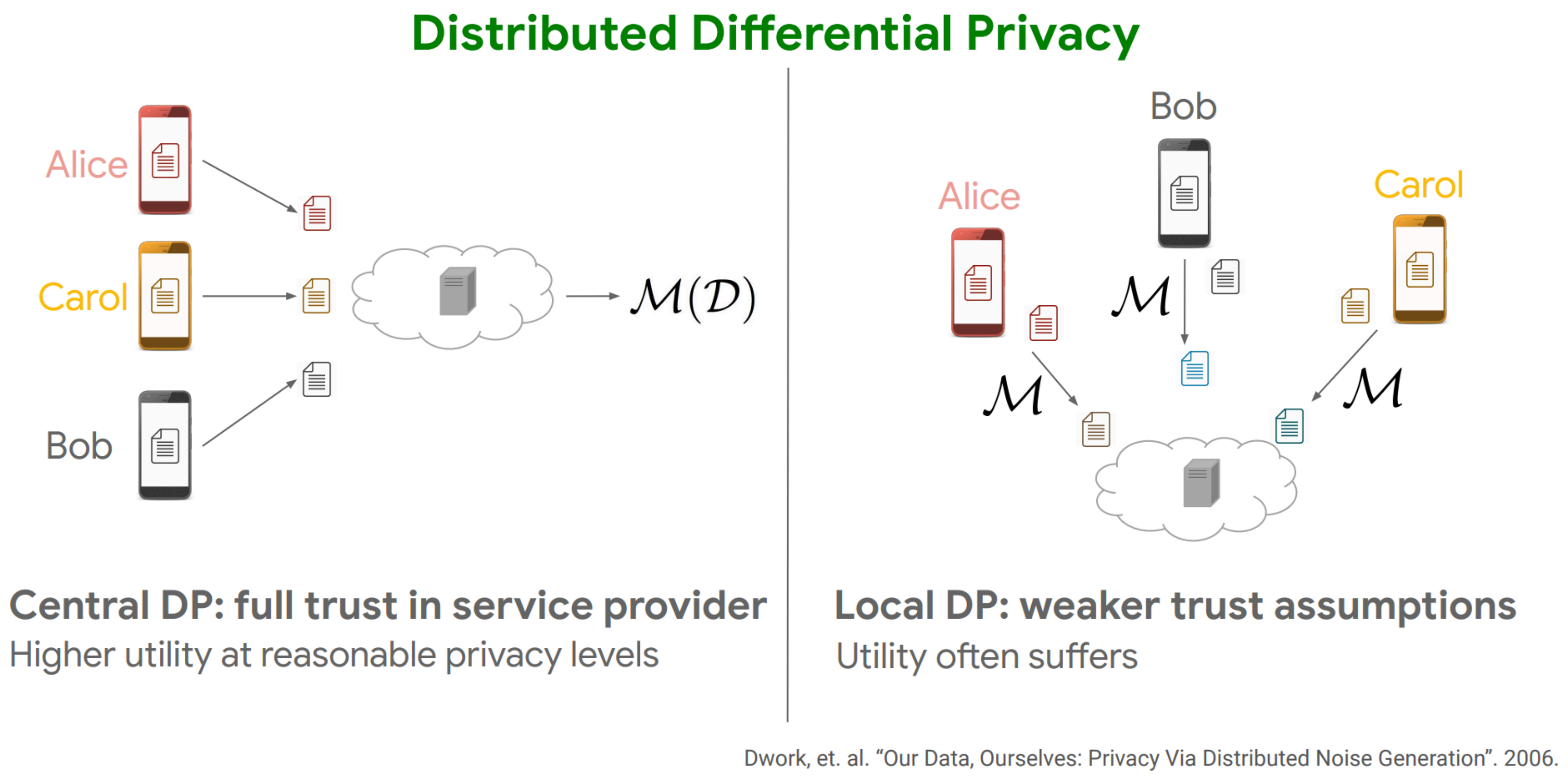


Centralized Setting

Distributed Setting

Relies on distributed optimization

Differential privacy and local differential privacy (Optional)



Distributed Differential Privacy

Central DP: full trust in service provider

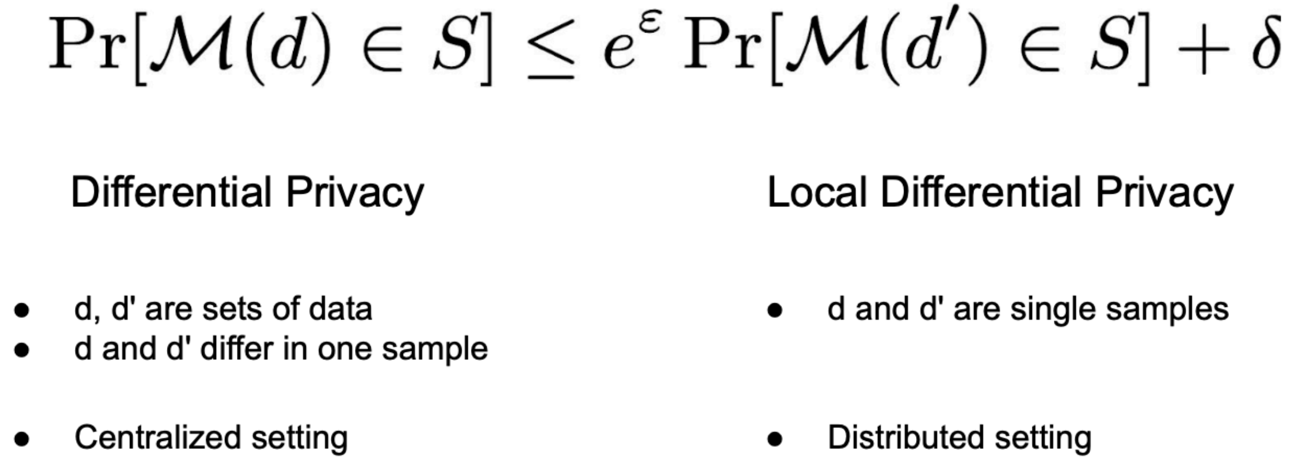
Higher utility at reasonable privacy levels

Local DP: weaker trust assumptions

Utility often suffers

Dwork, et. al. “Our Data, Ourselves: Privacy Via Distributed Noise Generation”. 2006.

Differential privacy and local differential privacy (Optional)



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

Differential Privacy

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

- Centralized setting

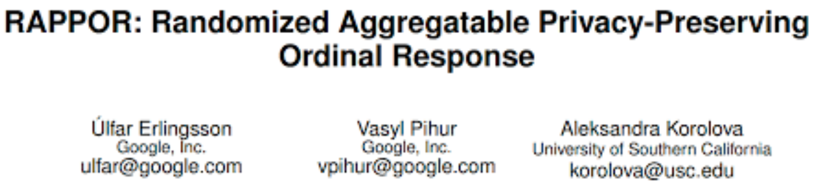
Local Differential Privacy

- d and d' are single samples

- Distributed setting

Deployment of local differential privacy (Optional)

- RAPPOR by Google
  - Collect user data
  - **R**andomized **A**ggregatable **P**rivacy-**P**reserving **O**rdinal **R**esponse

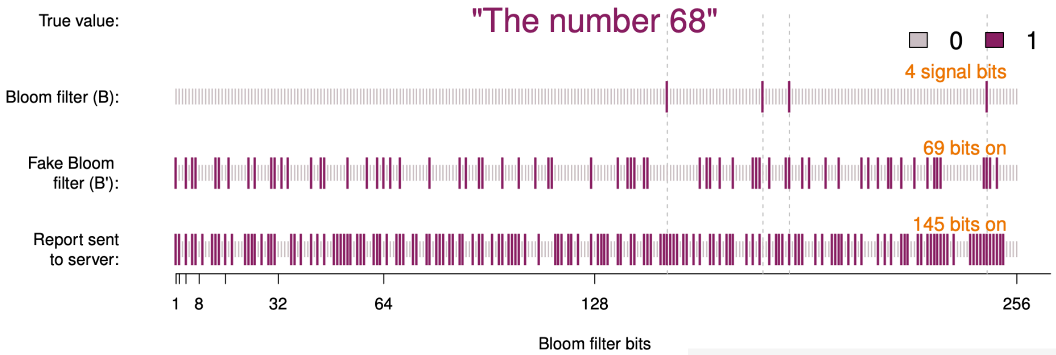


RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

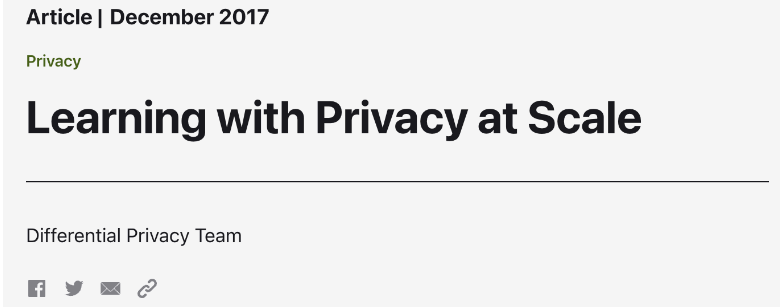
Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu



- Private Count Mean Sketch by Apple
  - Collect emoji usage data along with other information in iPhone
  - **Learning with Privacy at Scale**



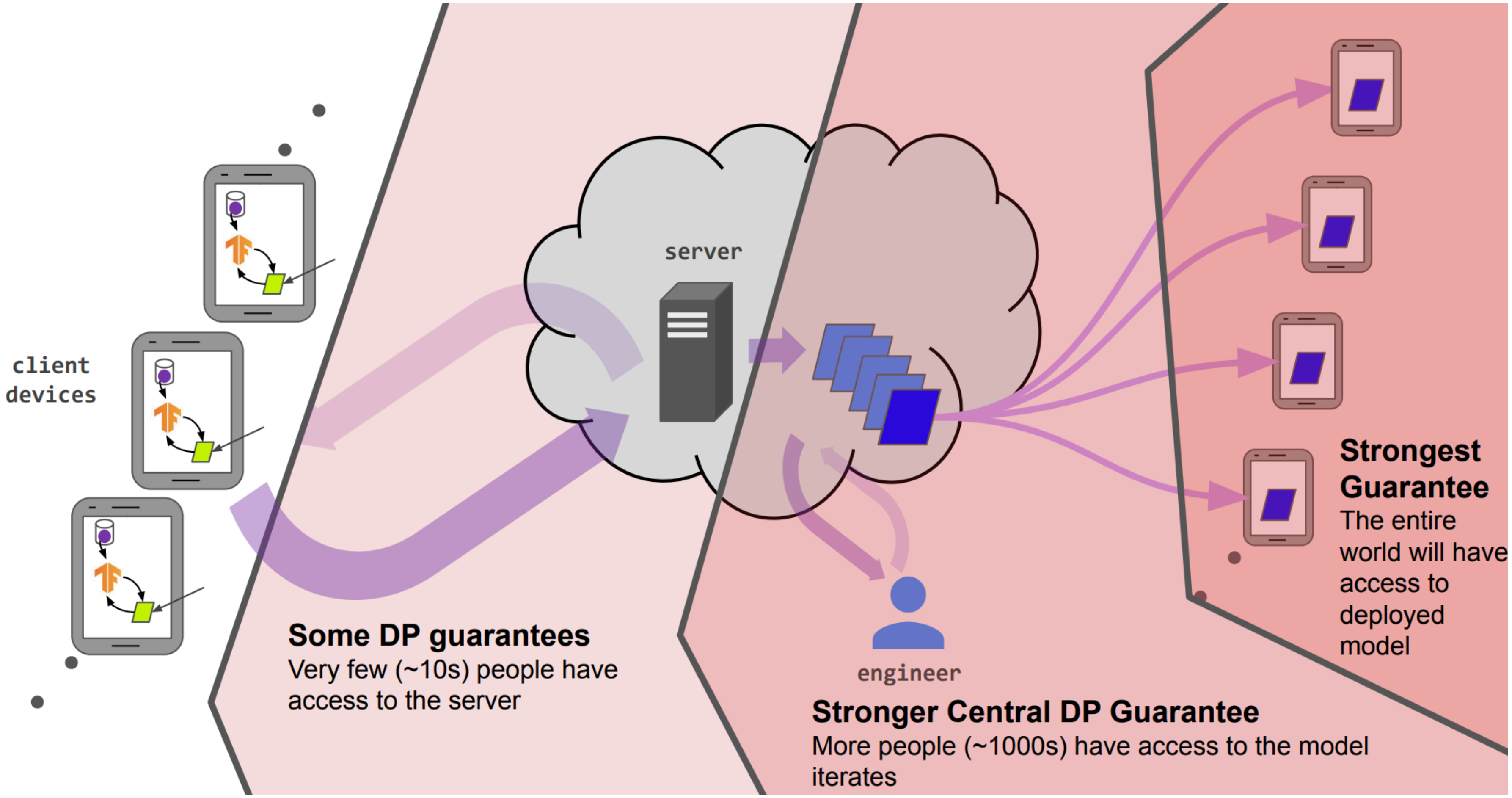
Article | December 2017

Privacy

Learning with Privacy at Scale

Differential Privacy Team

Distributed DP (Optional)



Some DP guarantees
Very few (~10s) people have access to the server

Stronger Central DP Guarantee
More people (~1000s) have access to the model iterates

Strongest Guarantee
The entire world will have access to deployed model

Part III

Summary

Learning Outcomes

- Understand why privacy matters in ML
- Know how to describe possible attacks
- Be able to derive the theorems for reconstruction attacks
- Be able to state the definition of differential privacy
- Be able to state and verify simple mechanisms
- Understand the composition rule
- Know what federated learning is and how differential privacy can be guaranteed