DSC 291 Introduction to Differential Privacy (and differentially private machine learning)

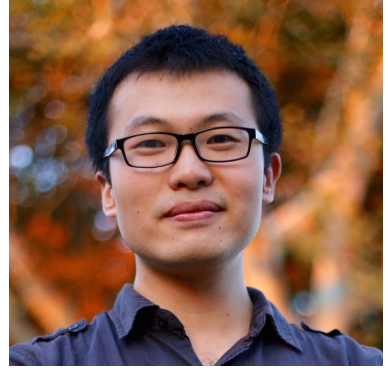
Instructor: Prof. Yu-Xiang Wang

Fall 2024

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

A bit about me



Yu-Xiang Wang 王宇翔

Associate Professor

Halıcıoğlu Data Science Institute
and Department of Computer Science and Engineering

UC San Diego

Office: HDSI 352

E-mail: yuxiangw AT ucsd.edu

Yu-Xiang is pronounced approximately as ['ju:'ʃi:ʌŋ], namely, *y~eu~ee - sh~ih~ah~ng* .

Research area: Statistical Machine Learning. Optimization, reinforcement learning, differential privacy, deep learning.

Short biography:
China => Singapore
=> PhD from Carnegie Mellon University
=> Scientist at Amazon
=> Professor at UC Santa Barbara
=> Professor at UC San Diego

Homepage:
https://cseweb.ucsd.edu/~yuxiangw/

Instruction team and office hour

- Instructor: Yu-Xiang Wang

Office hour: after lecture on Wednesday in HDSI 352

- TA: Dan Qiao

Office hour: Friday 4-5 pm @ HDSI 455 (watch out for piazza announcement)

Important logistic question to everyone

What's the issue?

- Currently the lectures are scheduled to MWF.

- 50 min is a bit too short for learning new technical materials (leaving no time for meaningful discussion)

- I have faculty meeting at 2:00 pm right after the scheduled Friday lecture (thus can't go longer).

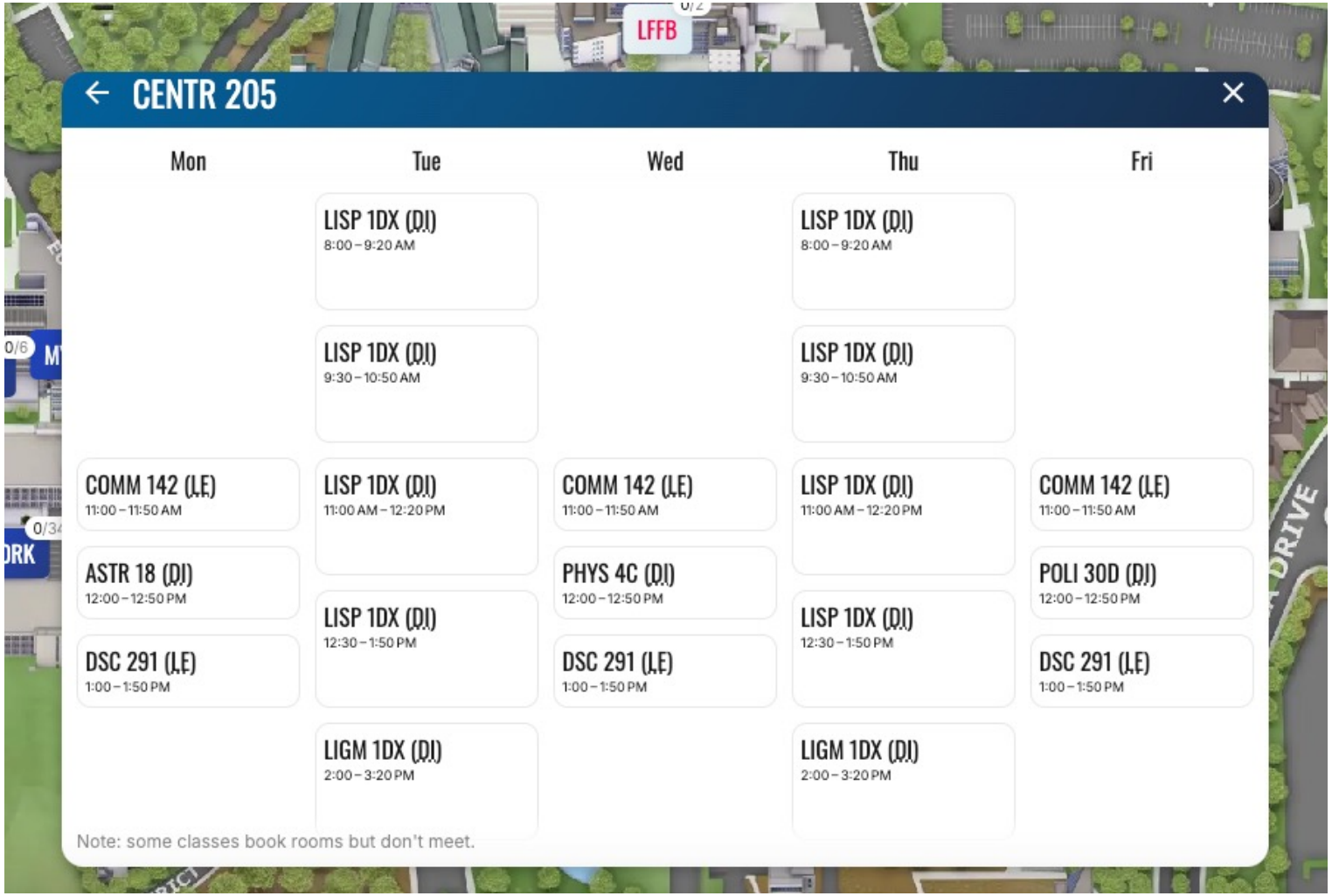
- Quarter-long course is a bit too short.

My proposal:

- ***Two longer (90-100 min) lectures*** on Monday and Wednesday.

- You guys get your Friday back.

Room seems available



Outline today

- Privacy challenges in modern times
- Course requirement
- A concrete privacy attack

What is privacy? What are the differences between "privacy", "confidentiality" and "security"?

- Philosophically,
  - "Confidentiality" = "Don't tell"
  - "Privacy" = "Don't ask"

- Legally speaking: Privacy is about the right to be left alone (from public scrutiny); Confidentiality is about a promise from people who have privileged access.

- "Security" vs "Privacy":
  - Security is to prevent risks due to unintended system use.
  - Privacy prevents risks due to intended system use.

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

- Government, company, research centers collect personal information and analyze them.
- Social networks: Facebook, LinkedIn
- YouTube & Amazon use viewing/buying records for recommendations.
- Emails in Gmail are used for targeted Ads and for completing your sentence!

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

OCTOBER 30, 2023

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

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

What does the executive order say?

- **Protecting Americans' Privacy:** Without safeguards, AI can put Americans' privacy further at risk. AI not only makes it easier to extract, identify, and exploit personal data, but it also heightens incentives to do so because companies use data to train AI systems. To better protect Americans' privacy, including from the risks posed by AI to protect all American, *the President calls on Congress to pass bipartisan data privacy legislations*, especially kids, and directs the following actions:
  - **Protect Americans' privacy by prioritizing federal support for accelerating the development and use of privacy-preserving techniques**—including ones that use cutting-edge AI and that let AI systems be trained while preserving the privacy of the training data.
  - **Strengthen privacy-preserving research and technologies**, such as cryptographic tools that preserve individuals' privacy, by funding a Research Coordination Network to advance rapid breakthroughs and development. The National Science Foundation will also work with this network to promote the adoption of leading-edge privacy-preserving technologies by federal agencies.
  - **Evaluate how agencies collect and use commercially available information**—including information they procure from data brokers—and **strengthen privacy guidance for federal agencies** to account for AI risks. This work will focus in particular on commercially available information containing personally identifiable data.
  - **Develop guidelines for federal agencies to evaluate the effectiveness of privacy-preserving techniques**, including those used in AI systems. These guidelines will advance agency efforts to protect Americans' data.

Recent legislations on privacy forces businesses to revise their data practice

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

How about my machine learning models trained on user data?

Scientists need to have access to data, but

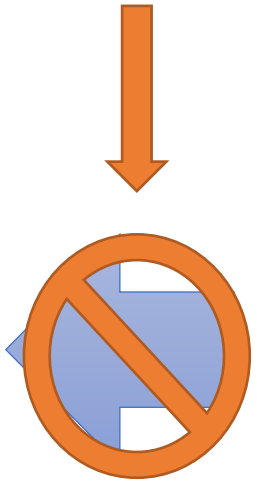
Privacy law

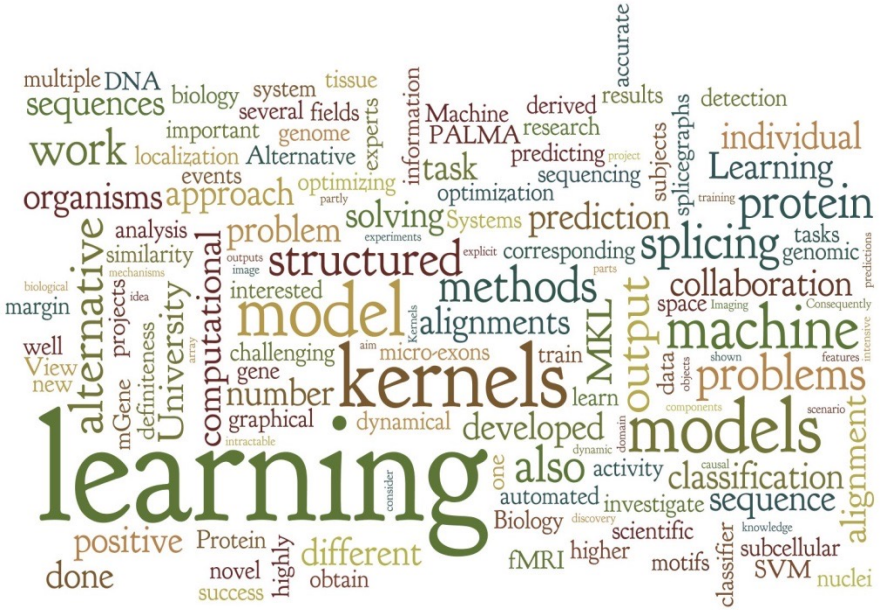
Food safety

Understand how brain works

Better treatment

Better education




learning
kernels
model
machine
models
methods
structured
problem
solving
prediction
splicing
protein
Learning
individual
sequences
work
organisms
approach
task
alternative
computational
collaboration
problems
classification
sequence
alignment
also
developed
dynamical
graphical
number
gene
challenging
alignments
MKL
output
positive
different
done
Protein
fMRI
higher
motifs
SVM
nuclei
Biology
scientific
investigate
automated
multiple DNA
biology
system
tissue
several fields
important
genome
Alternative
localization
Machine
PALMA
research
derived
results
detection
predicting
sequencing
optimization
Systems
corresponding
explicit
interested
margin
well
View
new
University
mGene
definiteness
projects
similarity
analysis
number
micro-exons
train
learn
space
data
shown
scenario
components
domain
dynamic
activity
causal
knowledge
discovery
subcellular
nuclei
SVM
classifier
motifs
obtain
highly
novel
success
accurate
information
experts
subjects
splicegraphs
genomic
tasks
predictions
Consequently
intensive
features
Imaging
training
project
partly
experiments
outputs
image
parts
mechanisms
idea
biological
Kernels
aim
array
intractable
consider
one
Clean energy

Medical diagnosis

Epidemic forecast/prevention

Fraud detection

Increasing privacy awareness

The census bureau's dilemma

Imaged by Heritage Auctions, HA.com

- Article I, Section 2 of the United States Constitution requires the US Census to count individuals **accurately** for tasks such as "**congressional apportionment**".

- Title 13 of the United States Code. "It is **against the law** for any Census Bureau employee to disclose or **publish any** census or survey **information that identifies an individual** or business. This is true even for inter-agency communication: the FBI and other government entities do not have the legal right to access this information. "

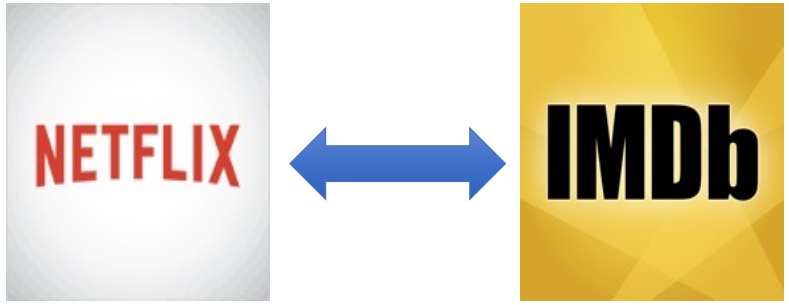
- What if the two goals contradict each other?

Do we really need math / science 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.

Removing/modifying personal identifiable information

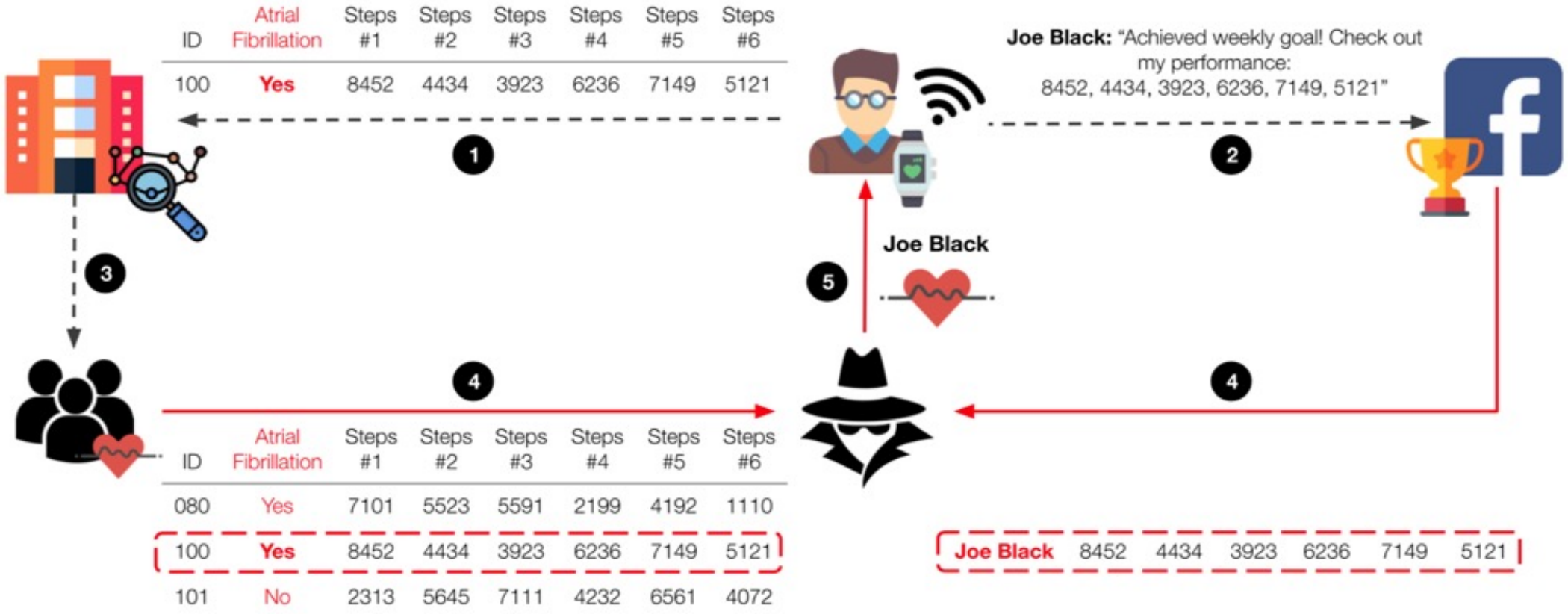
- Name? Gender? Zip code? Watched movies?



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

Narayanan, Arvind, and Vitaly Shmatikov. "How to break anonymity of the netflix prize dataset." (2006).

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



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

- “Who likes Justin Bieber?”

- Questionaire: “Year, Program, Gender, Like Bieber or not?”
  - Results as of Monday: “How many like Bieber” 16
  - Results as of Tuesday: “How many like Bieber” 17
  - Side information (available to the instructor): You enrolled late on Tuesday.

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

How do these attacks work?

- 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.

Remark: 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

Conclusions so far: Privacy is challenging!

- Revealing dataset (even if with PII removed) is a bad idea
  - Data-linkage attack, netflix prize
- Revealing aggregate statistics of the dataset have privacy risks
  - Differencing attack: With side information, even if reporting just one, may reveal information about individuals
  - An even stronger attack later: even without side-information, even with noise in the statistics.
- Machine learning models encodes information of individuals in a dataset and will spit them out when given an appropriate prompt
  - Membership inference attack
  - Unintended memorization

A bit of history of privacy protection techniques: various attempts for 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++]

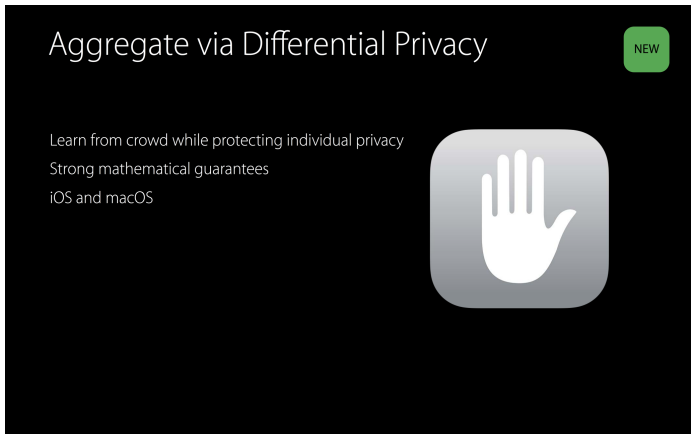
This course is about differential privacy

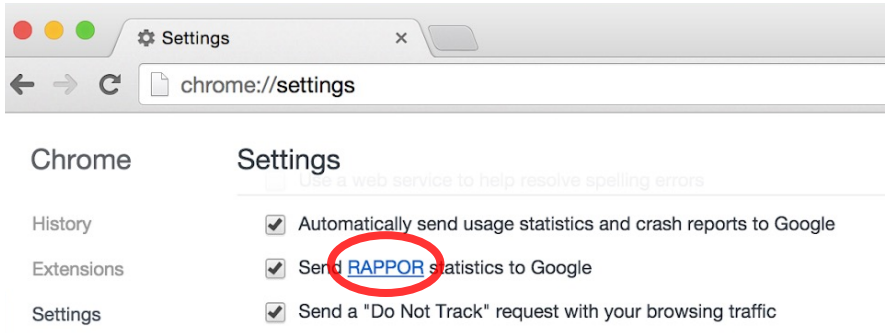
- A formal mathematical definition of privacy that provides rigorous guarantees and provably effective protections against privacy risks.
- Makes no assumption on the adversary
  - Arbitrary side info, arbitrary computational power.
- Interpretable, quantifiable, composable formalism
- The de-facto standard in privacy --- the only one still being actively researched on.

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

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.

Differential privacy is transitioning from a theoretical construct into a practical technology!





Topics that I will cover in this class

1. Privacy risks, privacy attacks and motivations of differential privacy.
2. The definition of differential privacy and its various interpretations.
3. Understanding the promise of DP: what does it protects and what not?
4. Fundamental building blocks of DP algorithms.
5. Modern methods in DP: Renyi DP, Privacy profiles, tradeoff functions
6. How to use autodp for privacy accounting and privacy calibration.
7. Differentially private (linear) machine learning
8. Differentially private deep learning under various models
9. Data-adaptive differentially private algorithms.

By the end of the course, you will be able to

- determine when DP is applicable and whether it will allow for sufficient utility for an application.

- design and implement DP algorithms correctly (prove DP guarantees).

- gain familiarity / proficiency in the various elements of theoretical research in CS / ML.

Prerequisites

- The course might be tough for you if you do not have working knowledge in:
  - Probability theory and statistics
  - Linear algebra, basic calculus
  - Basic data structures and algorithms
  - writing mathematical proofs
  - writing simple code (in Python / Numpy)

- Advanced knowledge that will help (but not required):
  - Linear / Convex optimization
  - Concentration inequalities
  - Statistical machine learning
  - Experience doing theoretical research
  - Experience working with a real dataset

Evaluation

- Lecture attendance: 10%
- Homework: 30%
- Midterm Quiz: 20%
- Project: 40% (Up to 3 students per team)
  - Required part: Conduct one assigned private data analysis.
  - Flexible part (choose one or propose your own):
    - DP theory: Read a recent paper in DP theory so you can reproduce the proof.
    - DP application: Implement one or a few DP algorithms (can be your research project).
    - Open source contribution: Contribute code to autodp.
    - DP + X research project: Develop new DP algorithms or applications

I will share a list of project ideas on Piazza soon.

Tentative schedule of the course

Part I: Differential Privacy Basics

Part II: Machine learning with DP

Part III: Advanced topics / Project Presentations

Week	Dates	Lecture Materials	Readings	Assignments
1	27-Sep-24	No lecture. Please read this notebook on Privacy challenges in the AI Era and DR Ch 1.	DR Ch 1	
2	30-Sep-24, 2-Oct-24, 4-Oct-24	Overview. Privacy threats. Defining Privacy, Sensitivity and Laplace Mechanism	DR Ch 2, Ch 3.1, Ch 3.2	HW1 out on Oct 2
3	7-Oct-24, 9-Oct-24, 11-Oct-24	Sparse Vector and Private Query Release	DR Ch 3.6, DR Ch 4.2	Project proposal due Oct 11
4	14-Oct-24, 16-Oct-24, 18-Oct-24	Report Noisy Max, Exponential Mechanism, SmallDB	DR Ch 3.3, DR Ch 3.4	HW1 due on Oct 16 HW2 out on Oct 18
5	21-Oct-24, 23-Oct-24, 25-Oct-24	Privacy loss RV, Composition, Gaussian Mechanism	DR Ch 3.5	
6	28-Oct-24, 30-Oct-24, 1-Nov-24	Modern Privacy Accounting: Renyi DP, f-DP and autodp	Bun & Steinke (2018), Balle & Wang (2018)	Midterm Quiz on 30-Oct-24 HW2 due on Nov 1
7	4-Nov-24, 6-Nov-24, 8-Nov-24	Differentially Private Machine Learning: Posterior Sampling, Objective Perturbation	Chaudhuri et al., Kifer et al.	HW3 out on Nov 4 Project midterm report due Nov 8
8	11-Nov-24, 13-Nov-24, 15-Nov-24	More DPML algorithms: Noisy Gradient Descent	Bassily et al.	
9	18-Nov-24, 20-Nov-24, 22-Nov-24	Data Dependent DP Algorithm Design: Smooth Sensitivity, Propose-Test-Release	Vadhan Ch 3.1, Nissim et al. (2011)	HW3 due on Nov 22
10	25-Nov-24, 27-Nov-24, 29-Nov-24	Project presentations		
11	2-Dec-24, 4-Dec-24, 6-Dec-24	Project presentations		Final project report due

Remainder of today's lecture

- A concrete data-reconstruction attack
- Randomized response

A simple mathematical model to consider

- We have a dataset of n individuals
- One secret bit of information per person
- (Normalized) linear query / statistical query:

We say an algorithm is blatantly non-private if one can reconstruct 90% of the dataset (secret bit vector) using its output.

Definition 5.1 (blatant non-privacy, due to Dinur and Nissim [30]). A mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is called *blatantly non-private* if for every $x \in \mathcal{X}^n$, one can use $\mathcal{M}(x)$ to compute an $x' \in \mathcal{X}^n$, such that x' and x differ in at most $n/10$ coordinates (with high probability over the randomness of $\mathcal{M}$).

- In particular we will consider algorithms that answer a collection of linear queries about the dataset approximately.

Reconstruction Attack --- find a dataset that is consistent with the observations!

- We have k linear queries
- An algorithm returns answers that are α-accurate
- The reconstruction attack:

Any algorithm that answers all 2^n linear queries with constant error implies blatant non-privacy!

Theorem 5.2 (reconstruction from many queries with large error [30]). *Let $x \in \{0,1\}^n$. If we are given, for each $q \in \{0,1\}^n$, a value $y_q \in \mathbb{R}$ such that*

$$\left| y_q - \frac{\langle q, x \rangle}{n} \right| \leq \alpha.$$

Then one can use the y_q's to compute $x' \in \{0,1\}^n$ such that x and x' differ in at most 4α fraction of coordinates.

- Proof:

Any algorithm that answers O(n) linear queries with $O\left(\frac{1}{\sqrt{n}}\right)$ error implies blatant non-privacy.

Theorem 5.6 (reconstruction from few queries with small error [30]). *There exists $c > 0$ and $q_1, \ldots, q_n \in \{0,1\}^n$ such that any mechanism that answers the normalized inner-product queries specified by $q_1, \ldots, q_n$ to within error at most $c/\sqrt{n}$ is blatantly non-private.*

- Recall the attack is

$$x' = \arg \min_{\tilde{x} \in \{0,1\}^n} \max_{i \in [n]} \left| y_i - \frac{1}{n} q_i^T \tilde{x} \right|$$

- Observations

Quick checkpoint

- "Any algorithms that answers too many questions too accurately will result in a blatant reconstruction of the dataset."
  - No side information needed.
  - No restriction on the dataset.
  - Fixed "design", highly generic design (iid samples)

- The attacks are not computationally efficient, but …
  - efficient attacks exist, via a Linear Programming relaxation

$$x' = \arg \min_{\tilde{x} \in [0,1]^n} \max_{i \in [k]} \left| y_i - \frac{1}{n} q_i^T \tilde{x} \right|$$

Read the original paper: Dinur-Nissim (2003) for more details.

What can we still do?

Target accuracy	k = O(2^n) linear queries	k = O(n) linear queries	k << n linear queries
α = O(1) (any non-trivial error)	Blatantly non-private	?	?
α = O(1/sqrt(n)) (statistical error)	Blatantly non-private	Blatantly non-private	?
α = o(1/sqrt(n)) (<< statistical error)	Blatantly non-private	Blatantly non-private	?

Two things to think about:

- Avoiding “Blatant-non-private” is a relatively weak privacy guarantee.
- Can we achieve the lower bound while satisfying a much stronger notion of privacy guarantee?

Randomized Response (Warner, 1965)

- Who likes Justin Bieber?
  - Space of the answer: {0,1}

1. Each individual tosses an independent coin with probability $p > 0.5$
2. If "head", keep your answer.
3. Otherwise, flip your answer.

- Intuitively each individual has a degree of plausible deniability.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63-69.

Observations

- 1. We returns a synthetic dataset with an expected L1-error n(1-p)
  - Not blatantly non-private if p << 0.9?
  - Is it possible to post-process it to reduce the error?

- 2. We can unbiasedly estimate any fixed set of (normalized) linear queries with error: O(log(|Q|)/(2p-1)sqrt(n))
  - When |Q| = 2^n, the bound is trivial
  - When |Q| = O(n), the bound matches what we expect up to log n.
  - When p = o(1) and p >> 1/(sqrt(n) log |Q|) ➔ strong privacy + non-trivial error

Next Lecture

- Differential Privacy
- Properties of Differential Privacy
- Basic mechanisms of DP