AI Ethics:
Privacy & Machine Learning

CS229: Machine Learning

Carlos Guestrin
Stanford University

Privacy Definition *(dictionary.com)*

2. the state of being free from unwanted or undue intrusion or disturbance in one's private life or affairs; freedom to be let alone.
3. freedom from damaging publicity, public scrutiny, secret surveillance, or unauthorized disclosure of one's personal data or information, as by a government, corporation, or individual.

Privacy vs Security

- Privacy is about your control of your personal information (and how it's used)
- Security is about protection against unauthorized access

Utility-Privacy Tradeoff

Privacy by Anonymization

- A trusted curator removes personally-identifying information (name, SSN,...)

Linkage Attack [Sweeney '00]

- Group Insurance Commission (GIC)
  - Anonymized data for ~135k patients for researchers and policy-makers
    - Including ZIP, birthdate and sex

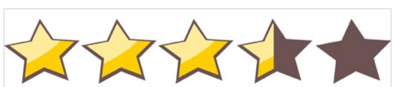
Linkage Attack [Sweeney '00]

- Group Insurance Commission (GIC)
  - Anonymized data for ~135k patients for researchers and policy-makers
    - Including ZIP, birthdate and sex
- Voter registration records
  - Name, …, ZIP, birthdate, sex
- Uncovered health records, e.g., of William Weld (governor of Massachusetts at that time)

Netflix Prize Linkage Attack

Netflix Prize 2006
Predict user rating

100 million movie ratings



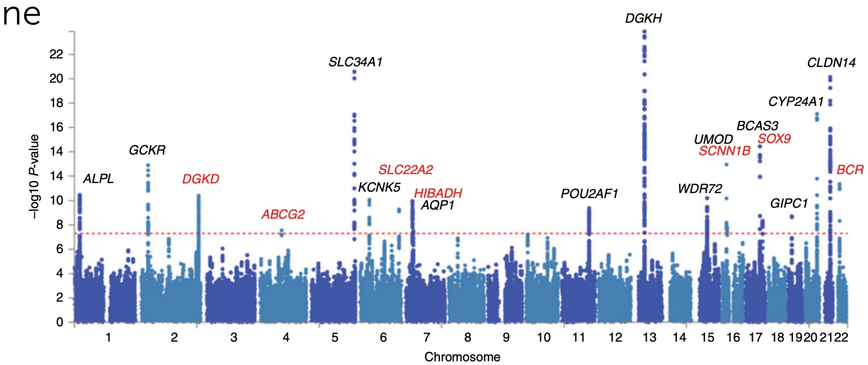
Privacy by Aggregation

- Common approach: aggregate counts, averages, trained models are private?

Genome Wide Association Studies (GWAS) with single-nucleotide polymorphisms (SNPs): **Membership Attack**

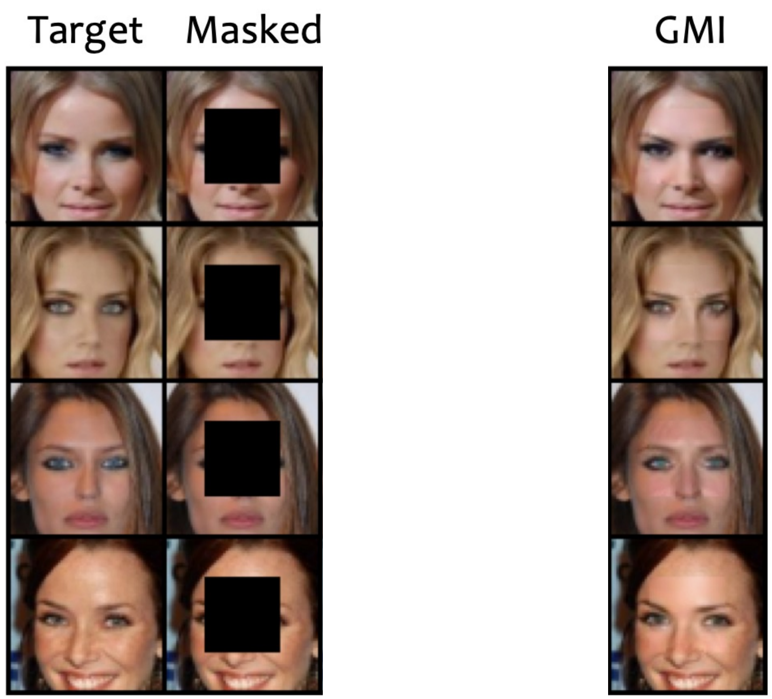
[Dwork et al.]

Kidney stone disease



- Able to infer if an individual's DNA is part of study

Generative Model Inversion Attack [Zhang et al 2020]



Randomized Response
[Warner 1965]

Randomized Response: Intuition

- Add noise to each data point, e.g., estimate average salary

Differential Privacy
[Dwork et al. 2006]
(Dwork and Roth 2014 Book is great reference: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf)

Formal Framework for Privacy

- Provide provable privacy-preserving guarantees

- Develop efficient methods to add noise and learn from data

Global Differential Privacy Framework

- You participate in "study"
  - i.e., provide data to trusted party
- Trusted party performs computations on data, but reveals answers that (attempt to) preserve privacy

- Goal: Provide provable privacy-preserving guarantees

Differential Privacy Setup

- Database D includes sensitive information
- Data analyst asks queries on D
- (Randomized) Mechanism M attempts to get response R to query, while attempting to avoid leaking of individual information

Differential Privacy: Neighboring Databases

- **Neighboring databases:** two databases D_1 and D_2 only differ in a single entry

Differential Privacy Definition [Dwork et al. '06]

- Neighboring databases: two databases D_1 and D_2 only differ in a single entry
- A mechanism M is ε-differentially private if, for any two neighboring databases, and any set R of possible responses:

- Note: Differential Privacy is a definition, not algorithm to achieve it

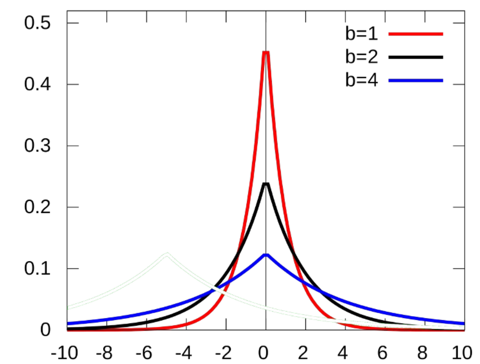
Differential Privacy Intuition

- You can't tell if it's me or someone else in the database
  - You can't tell if I was part of the study

Laplace Mechanism

Laplace Mechanism

$$p(w) = \frac{1}{b} \exp\left(-\frac{|w|}{b}\right)$$



- Add Laplace noise to the response

- How much noise to add?
  - Depends on magnitude of results
  - Suppose want to compute function f on database D, ***sensitivity*** of f:

- ***To achieve*** **ε-differential privacy**, noise level is:

Laplace Mechanism Example: Counts

- Suppose you want to count how many people have salary>$500k & got an A in CS281
  - f is count function
- Sensitivity of f:

- *To achieve* **ε-differential privacy**, noise level is:

Proof for 1D Laplace Mechanism

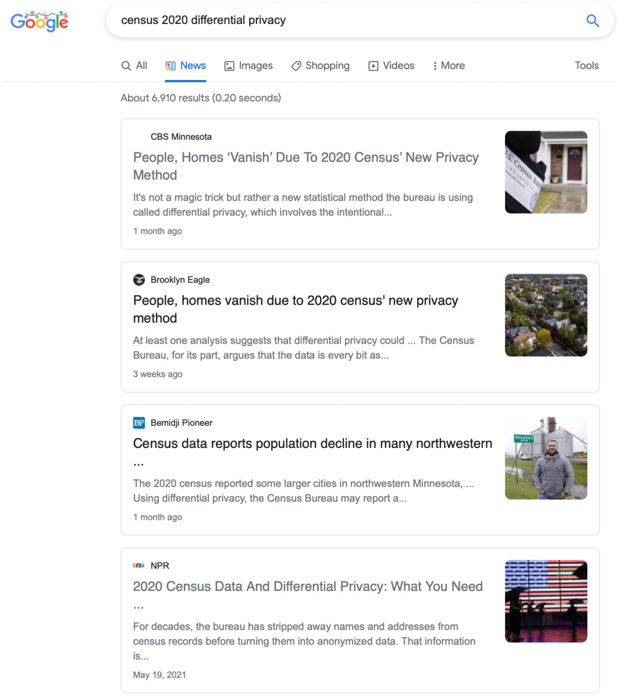
$$p(w) = \frac{1}{b}\exp\left(-\frac{|w|}{b}\right)$$

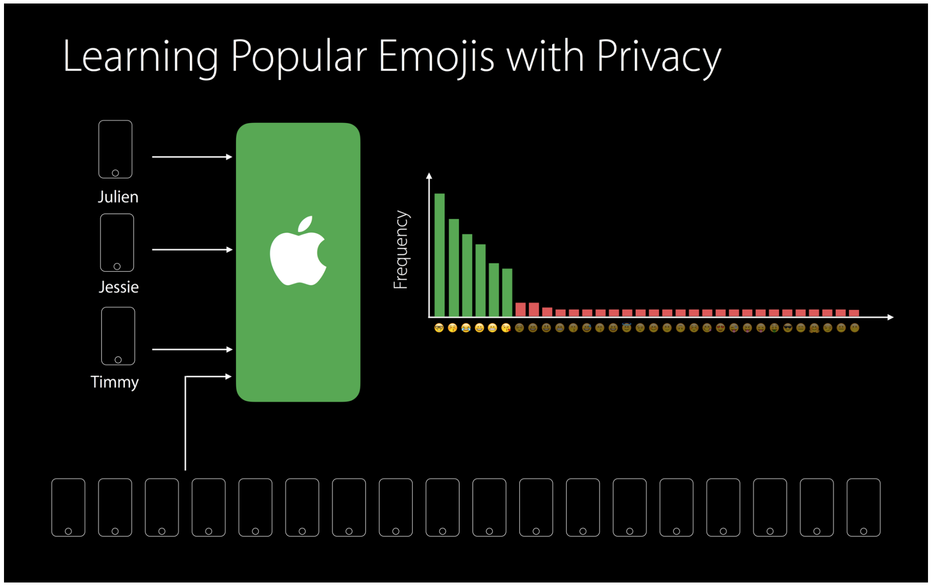
- Neighboring databases D_1 and D_2
- Mechanism M to compute f returns:

- Achieving ε-differential privacy:

Practical Examples of Differential Privacy

Practical Applications of Differential Privacy





Summary

- As we develop ML-based systems, it's important to consider privacy at every stage of the process
- Many methods and tools can help
- Ultimately, must manage the utility-privacy tradeoff

Closing a busy quarter… ☺

You did amazing things…

- Huge number of topics
- Remote learning
- Challenging homeworks and midterm
- Amazing project
- …

This is just the start…

- You now have the skills to have real-world impact with ML
- But, machines are not the only ones who keep learning… ☺
  - CS229 prepares you for many other classes at Stanford
  - And beyond
- We can't wait to see the amazing things you come up with!

Thank you to the amazing course staff!!!!!!!!

Course Manager

Swati Dube

Head Course Assistant

Nandita Bhaskhar

Course Assistants

Kyu-Young Kim

Beri Kohen Behar

Griffin Young

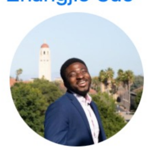
Sauren Khosla

Zhangjie Cao

David Lim

Soyeon Jung

Lantao Yu



Emmanuel Balogun

Jake Silberg

Ha Tran

Thank you!!!!!!!!! ☺