ORIE 5355

Introduction to differential privacy

Nikhil Garg

Introduction to (Differential) Privacy

(Special thanks for Juba Ziani, Georgia Tech, for slides)

Introduction: fundamental trade-off

Want to share and release information to do aggregate analyses

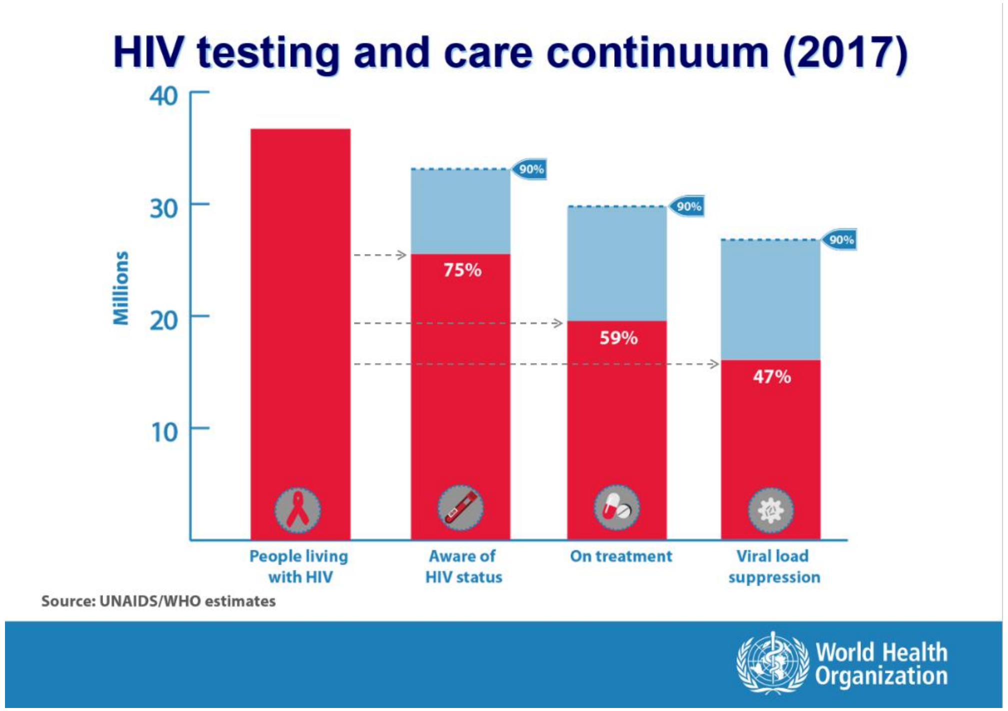
- Public audits (transparency)
- Want to help others do useful analyses (e.g., research reproducibility)
- Potentially legally mandated to share information (e.g., census)

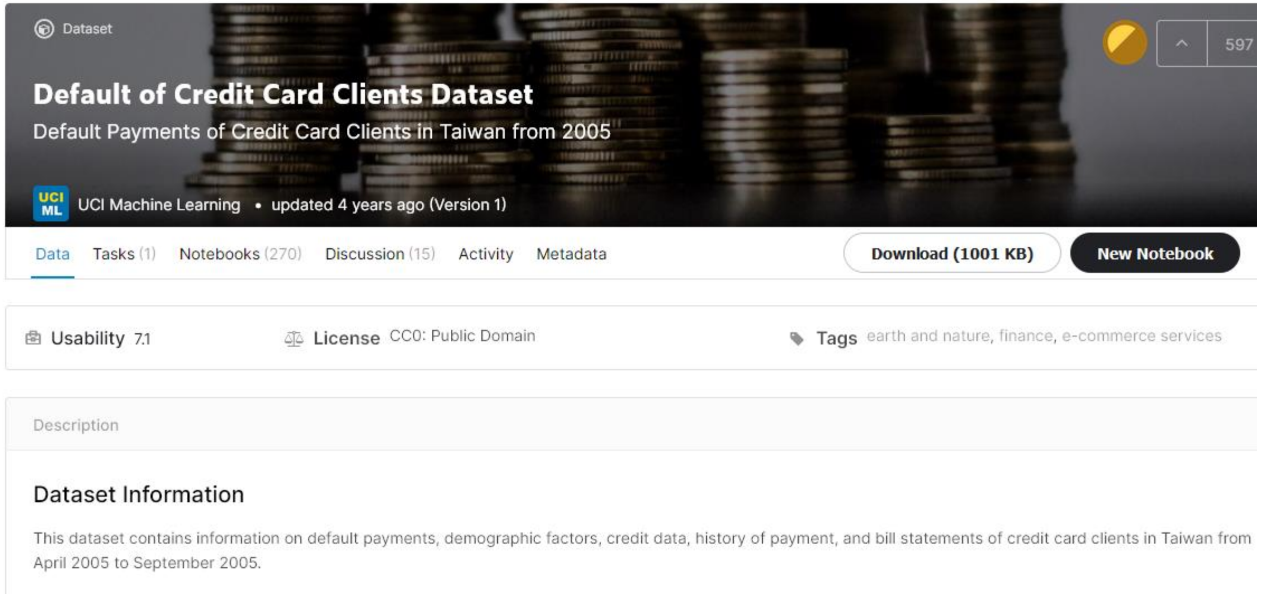
Don’t want to leak sensitive information about individuals

Problem: These two desiderata conflict, often in subtle ways!

Slide credit:
Juba Ziani,
Georgia Tech

Why is privacy important?





Slide credit:
Juba Ziani,
Georgia Tech

Failures of data privacy: anonymization

What is data anonymization?

Name	DOB	Gender	State/zip code	Has cancer?
Nikhil Garg	…	Male	NY 10044	No
Marge Simpson	04/19/1987	Female	SP 75234	No
Rick Sanchez	01/15/1943	Male	WA 98101	Yes
Misty	04/01/1983	Female	KT 16983	No

Slide credit:
Juba Ziani,
Georgia Tech

Failures of data privacy: anonymization

What is data anonymization?

Name	DOB	Gender	State/zip code	Has cancer?
~~Nikhil Garg~~	…	Male	NY 10044	No
~~Marge Simpson~~	04/19/1987	Female	SP 75234	No
~~Rick Sanchez~~	01/15/1943	Male	WA 98101	Yes
~~Misty~~	04/01/1983	Female	KT 16983	No

Slide credit:
Juba Ziani,
Georgia Tech

Failures of data privacy: anonymization

What is data anonymization?

Name	DOB	Gender	State/zip code	Has cancer?
1das4fg5d5as2	…	Male	NY 10044	No
345fa4f331t43	04/19/1987	Female	SP 75234	No
254jrtul42f4sf1	01/15/1943	Male	WA 98101	Yes
175dsa4f6jz68d	04/01/1983	Female	KT 16983	No

Slide credit:
Juba Ziani,
Georgia Tech

Failures of data privacy: anonymization

So what's the problem?

"Simple Demographics Often Identify People Uniquely"; Latanya Sweeney 2000

- A few attributes are enough to uniquely identify most of the US population
- (Zip, gender, date of birth) → identifies **87%** of US population
- What if I had this information (Zip, gender, date of birth) for much of the US?

Name	DOB	Gender	State/zip code	
1das4fg5d5as2	…	Male	GA 30309	
345fa4f331t43	04/19/1987	Female	SP 75234	
254jrtul42f4sf1	**01/15/1943**	**Male**	**WA 98101**	
175dsa4f6jz68d	04/01/1983	Female	KT 16983	

Slide credit:
Juba Ziani,
Georgia Tech

Failures of data privacy: anonymization

So what’s the problem?

“Simple Demographics Often Identify People Uniquely”; Latanya Sweeney 2000

- A few attributes are enough to uniquely identify most of the US population
- (Zip, gender, date of birth) → identifies **87%** of US population
- What if I had this information (Zip, gender, date of birth) for much of the US?

Name	DOB	Gender	State/zip code	Has cancer?
1das4fg5d5as2	…	Male	GA 30309	No
345fa4f331t43	04/19/1987	Female	SP 75234	No
Rick Sanchez	**01/15/1943**	**Male**	**WA 98101**	**Yes**
175dsa4f6jz68d	04/01/1983	Female	KT 16983	No

Slide credit:
Juba Ziani,
Georgia Tech

Failures of data privacy: anonymization

“Simple Demographics Often Identify People Uniquely”; Latanya Sweeney 2000

- In Mass, some anonymized health care data was publicly available to researchers
- Sweeney spent only $20 for public DOB/gender/zip codes info in Cambridge. Bought voter rolls.
- Same birthday as the governor of Mass: 6 people in Cambridge
- Only 3 were male
- Only 1 had the right zip code

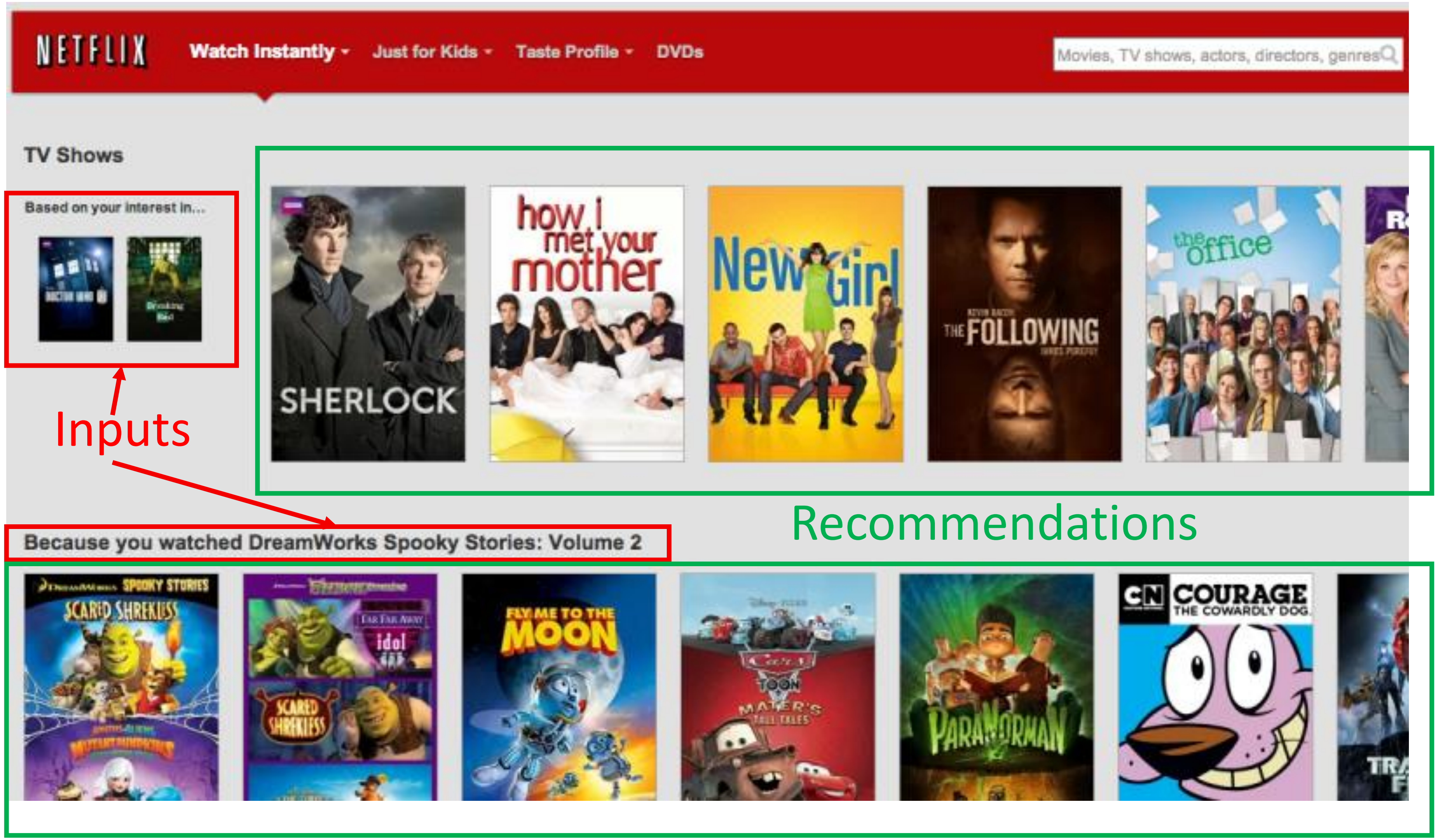
➔ Sweeney was able to ***uniquely identify the governor’s medical records***! Sent them to his office.

In 2021: “NYC Board of Elections glitch reveals how Mayor de Blasio’s son voted in city’s primary election”

> “Researchers with the Princeton lab were able to track down the results — which are supposed to be confidential — by cross-referencing state voter files against precinct-level results from election districts where only one voter is registered.”

Slide credit:
Juba Ziani,
Georgia Tech

The Netflix Competition



Slide credit:
Juba Ziani,
Georgia Tech

The Netflix Competition

How to improve recommendation system?

- Machine learning competition
- Try to predict user ratings from historical data as well as possible
- Provide "*anonymized*" data to participating teams

Netflix did more than just anonymization of data:

- Only small subsets of the full data; reduced the number of attributes
- Deleted some of the ratings
- Modified dates/temporal data

Slide credit:
Juba Ziani,
Georgia Tech

The Netflix Competition

"How To Break Anonymity of the Netflix Prize Dataset", Arvind Narayanan and Vitaly Shmatikov, 2006

Only 2 weeks after the Netflix competition

What they show:

Only need imperfect info:

1. approx. dates of rating ($\pm$2 weeks) for 6 movies
2. 2 ratings and dates (with a 3-day error)

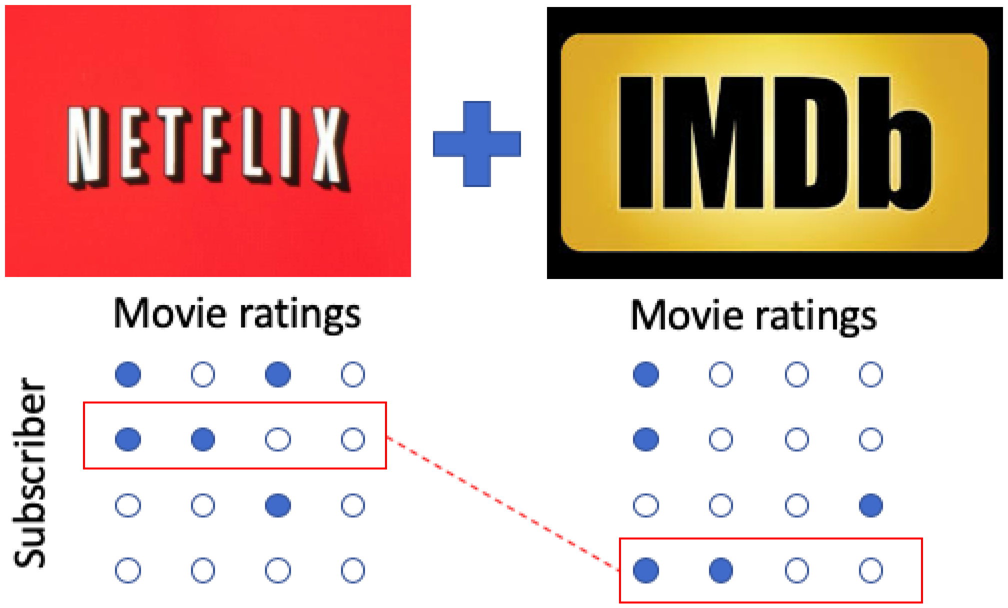
Can uniquely identify the person:

1. 99% of the time
2. 68% of the time

Slide credit:
Juba Ziani,
Georgia Tech

The Netflix Competition

How did they do it?



Why is it bad?

- Netflix watch history: more expansive and private than IMDb public rating
- Link IMDb and Netflix profile → learn private watch history on Netflix
- Gay mother sued Netflix: watch history could reveal her sexual orientation to others

Privacy summary so far

Privacy is important, but trades off with other values

Idea: Do things to the data to preserve privacy before release

- Anonymization: remove personal identification
- Edit some of the entries a little bit
- Delete some entries

Even with above techniques, many privacy failures!

Common attack: Use *external* data (IMBb, voter file, etc) to extract more information from the anonymized data

Slide credit:
Juba Ziani,
Georgia Tech

Next idea: *Aggregate data before release*

Idea: Only release aggregated statistics/model.

Examples

- Population-level statistics such as averages, etc.
- Neural net (only see the final model, not the training data)

Why should it naively work?

- No individual-level details or features!
- Cannot identify a single row in a database: no access to such row-by-row data

Issue: If you release enough statistics, that's statistically identally to releasing the actual dataset

Slide inspiration:
Juba Ziani,
Georgia Tech

Data Aggregation fails! Example 1

How? For each "column" of the data, we have a summary statistic (mean). One column doesn't tell us if any particular row is there. But if we have hundreds of thousands of columns in the dataset…

Example: genomic data

- Can you tell that someone's data was in a DNA database, if all you have is allele frequency data from the database?
- Yes: "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays", Homer et al., 2008

This is a problem

- Genomic data is more and more commonplace (ancestry tests, etc.)
- What if study only contains cancer patients/tries to link alleles to some rare disease? Can learn that you have a rare disease!

Slide credit:
Juba Ziani,
Georgia Tech

Data Aggregation fails! Example 2

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

xkcd: Predictive Models

"The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks", Carlini et al., 2019

Predictive models tend to memorize:

- Imperfect generalization/overfitting to dataset
- More obvious in language models:
  - Work by memorizing characters/word associations
  - Can repeat word associations from training data

Potential attack:

- Predict next word: "My SSN is…"
- Recovers some SSN used in training data

Slide credit:
Juba Ziani,
Georgia Tech

Beyond aggregating: adding noise

Answering queries exactly is not enough for privacy, even if queries aggregate a lot of data (e.g., if release many columns in the dataset)

Natural next step:

- Do not answer queries exactly!
- Anonymize/aggregate, AND add noise/randomness to data or to queries

Q: Is this enough?

A: Yes!, but you have to be careful ***how and how much noise*** you add

Slide inspiration: Juba Ziani, Georgia Tech

Fundamental tradeoff: privacy vs accuracy

"Giving overly accurate answers to too many questions will inevitably destroy privacy." -- Cynthia Dwork, Aaron Roth

- If you want to release a dataset that answers many questions about individuals, then you need to add more noise to each answer
- How much noise?

"Revealing information while preserving privacy", Irit Dinur & Kobbi Nissim

Theorem: There exists a reconstruction attack that issues $O(n)$ (random) queries, obtains answers with error αn, and reconstruct the secret bits of all but $O(\alpha^2 n^2)$ users. → To protect privacy on most of the database against computationally efficient attacks, need noise of the order of at least $n^{1/2}$.

Idea: [More] noise leads to [more] privacy

What happens if I probabilistically change the data?

Original Database D

ID	Other Cols…	Has Cancer?
Nikhil	…	No
Rick	…	Yes
Homer	…	No

Flip each datapoint with probability ϵ →

Released database D'

ID	Other Cols…	Has Cancer?
Nikhil	…	No
Rick	…	No
Homer	…	Yes

Distribution of outputs of computation ***almost unchanged*** *(with small ϵ)*

- If $\epsilon = 0$, then *no privacy* – *we are releasing exact dataset*
- If $\epsilon = \frac{1}{2}$, then *no accuracy* – *learn nothing from the dataset*

ϵ is a **policy choice**, not a technical one.

Can do the same thing with numeric columns

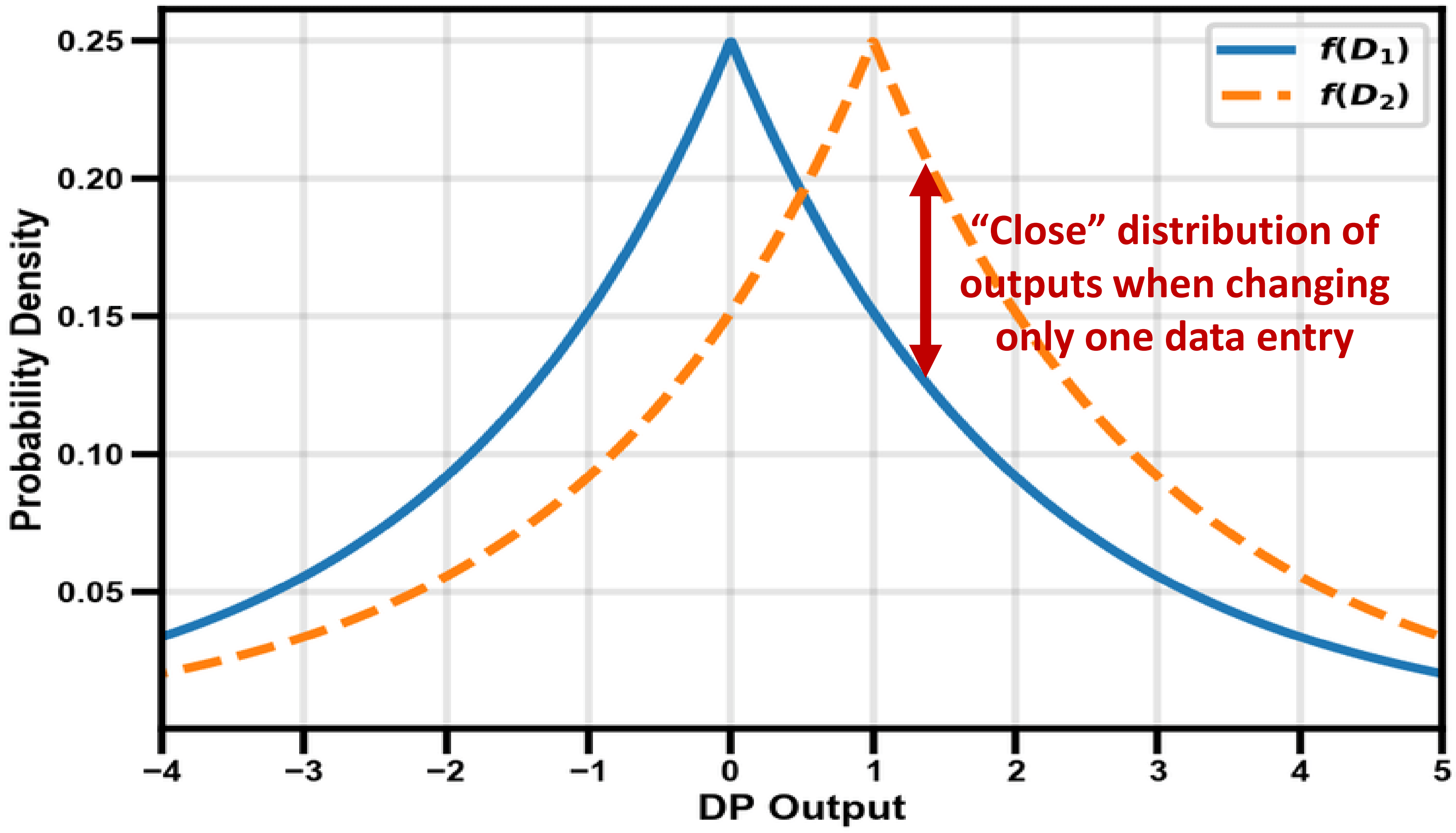


Image credit: Juba Ziani, Georgia Tech

Differential privacy

Differential privacy

- Fundamental limit: How much noise is needed
- Algorithm: What type (distribution) of noise to add

"Differential privacy is the only known framework to rigorously prevent such reconstruction attacks and privacy violations"

Now used in many places

- [Controversially] In the 2020 U.S. Census
- Google, Apple, Microsoft, LinkedIn…