DSC 291 Privacy-sensitive Data Systems (week 6a)

Haojian Jin

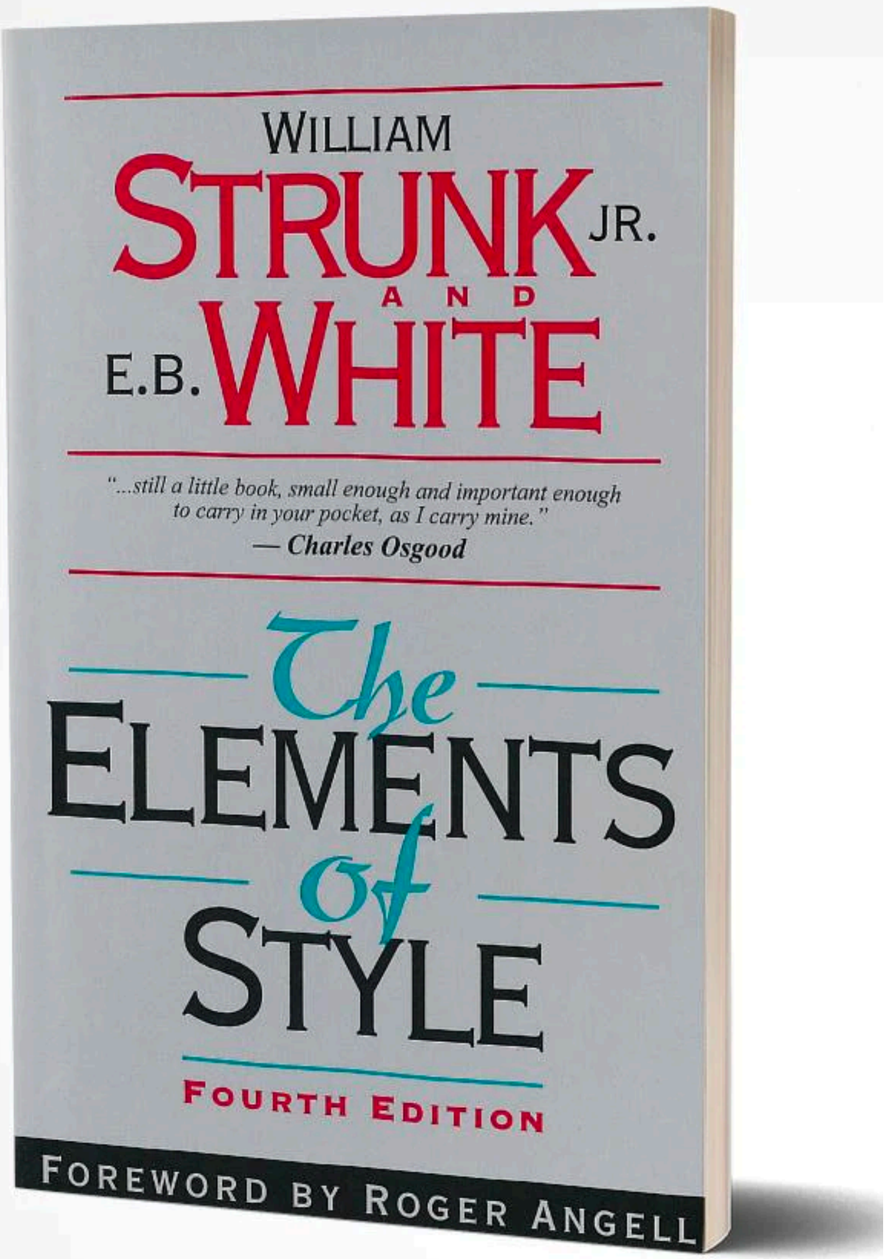
Logistics

1. Final project
    1. Abstract due Feb. 15.
    2. Final report due Mar. 19
2. Grades
    1. Would be more strict.

Academic writing

1. Important!
2. Do not waste readers' time.
    1. You can use ChatGPT, but you need to modify the text.
3. Show "meat" (insights).
    1. The length of your writings should be commensurable with the amount of "meat".
4. Zero insights => less than 80%.

The elements of style



Grading rubrics

1. Students got -1 to -10 points for grammar and clarity issues.
2. Students got -1 to -50 points for the review quality.

Recap

1. Privacy compliance and review
    1. Lean Privacy Review
    2. LegalEase

Today's class

1. K-anonymity
2. Differential privacy
3. Implementation
4. No free lunch in data privacy

Which of the following statements is true regarding anonymization and pseudonymization?

A. Anonymization refers to the process of replacing personally identifiable information with random or non-identifying data, while pseudonymization refers to the process of replacing personally identifiable information with a pseudonym or alias.

B. Anonymization and pseudonymization refer to the same process of replacing personally identifiable information with non-identifying data.

C. Anonymization and pseudonymization are only used in the context of online anonymity and do not apply to offline data protection.

D. Anonymization and pseudonymization are both illegal practices in data protection regulations.

https://bit.ly/privacysys-live-poll

Outcome oriented v.s. process oriented?

Classical intuition for privacy (Semantic Privacy)

"If the release of statistics S makes it possible to determine the value [of private information] more accurately than is possible without access to S, a disclosure has taken place."

[Dalenius 1977]

Issues

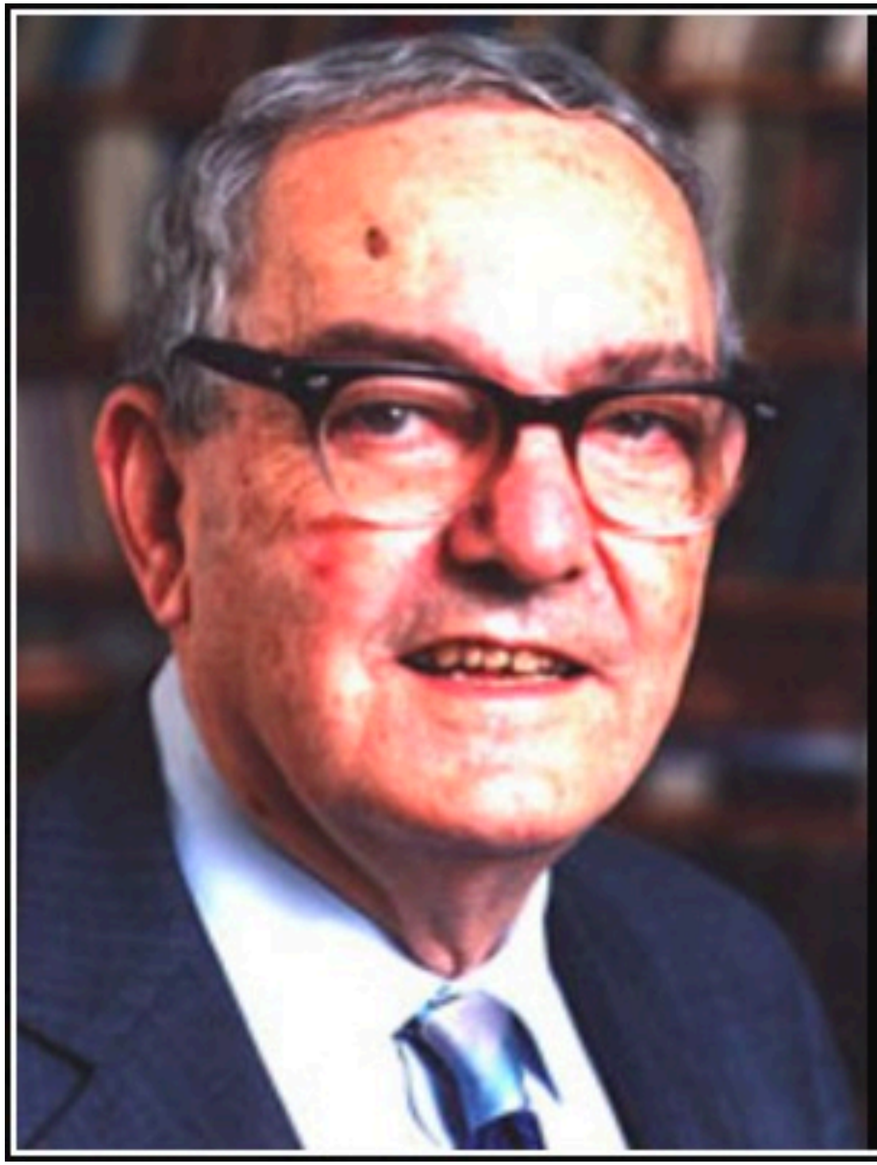
- Example
  - Terry Gross is two inches shorter than the average Lithuanian woman
  - DB allows computing average height of a Lithuanian woman
  - This DB breaks Terry Gross's privacy according to this definition… even if her record is not in the database!

Defining Privacy

- In order to allow utility, a non-negligible amount of information about an individual must be disclosed to the adversary.

- Measuring information disclosed to an adversary involves carefully modeling the **background knowledge** already available to the adversary.

- … but we do not know what information is available to the adversary.

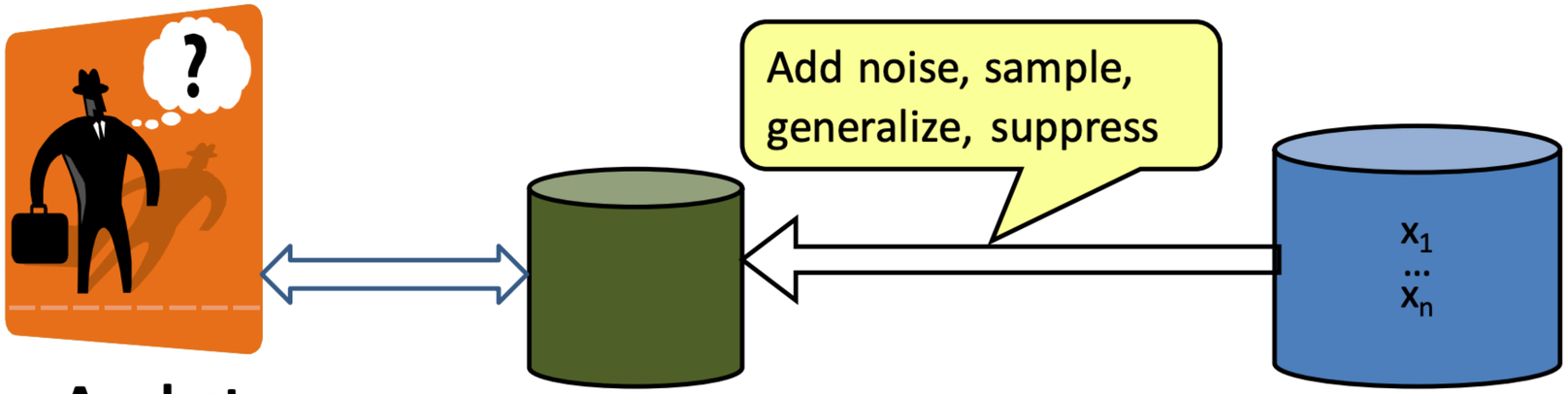


Solving a problem simply means representing it so as to make the solution transparent.

— Herbert Simon —

AZ QUOTES

Privacy-Preserving Statistics: Non-Interactive Setting



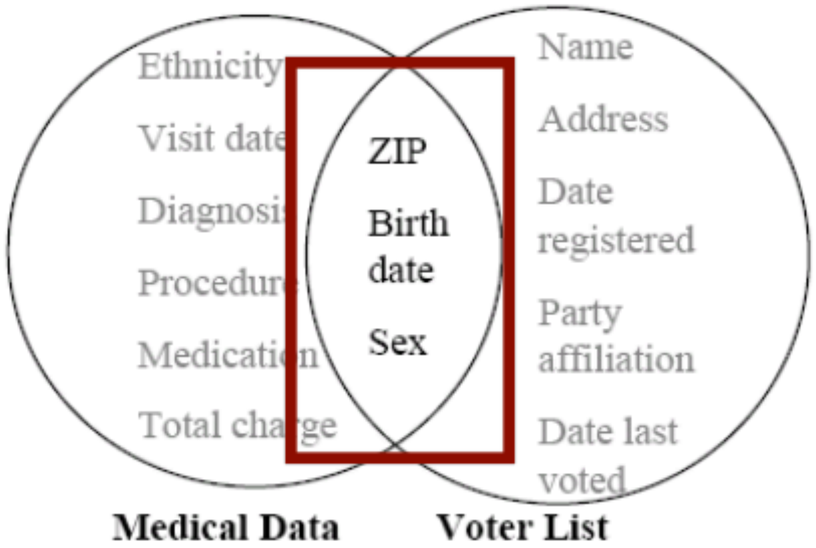
Goals:

- Accurate statistics (low noise)
- Preserve individual privacy (what does that mean?)

- Census data
- Health data
- Network data
- …

Re-identification by linking

Linking two sets of data on shared attributes may uniquely identify some individuals:



87 % of US population uniquely identifiable by 5-digit ZIP, gender, DOB

K-anonymity

1. Quasi-identifier: Set of attributes that can be linked with external data to uniquely identify individuals
2. Make every record in the table indistinguishable from at least k-1 other records with respect to quasi-identifiers
3. Linking on quasi-identifiers yields at least k records for each possible value of the quasi-identifier

K-anonymity and beyond

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Provides some protection: linking on ZIP, age, nationality yields 4 records

Limitations: lack of diversity in sensitive attributes, background knowledge, subsequent releases on the same data set

Re-identification Attacks in Practice

- Examples:
  - Netflix-IMDB
  - Movielens attack
  - Twitter-Flicker
  - Recommendation systems – Amazon, Hunch,..

Goal of De-anonymization: To find information about a record in the released dataset.

Anonymization Mechanism

	Gladiator	Titanic	Heidi
Bob	5	2	1
Alice	3	2.5	2
Charlie	1.5	2	2

Each row corresponds to an individual

Each column corresponds to an attribute, e.g. movie

Delete name identifiers and add noise

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Anonymized Netflix DB

De-anonymization Attacks Still Possible

- Isolation Attacks
  - Recover individual's record from anonymized database
  - E.g., find user's record in anonymized Netflix movie database
- Information Amplification Attacks
  - Find more information about individual in anonymized database
  - E.g. find ratings for specific movie for user in Netflix database

Isolation Attack!

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings (noisy)

	Titanic	Heidi
Bob	2	1

Used as auxiliary information

Weighted Scoring Algorithm

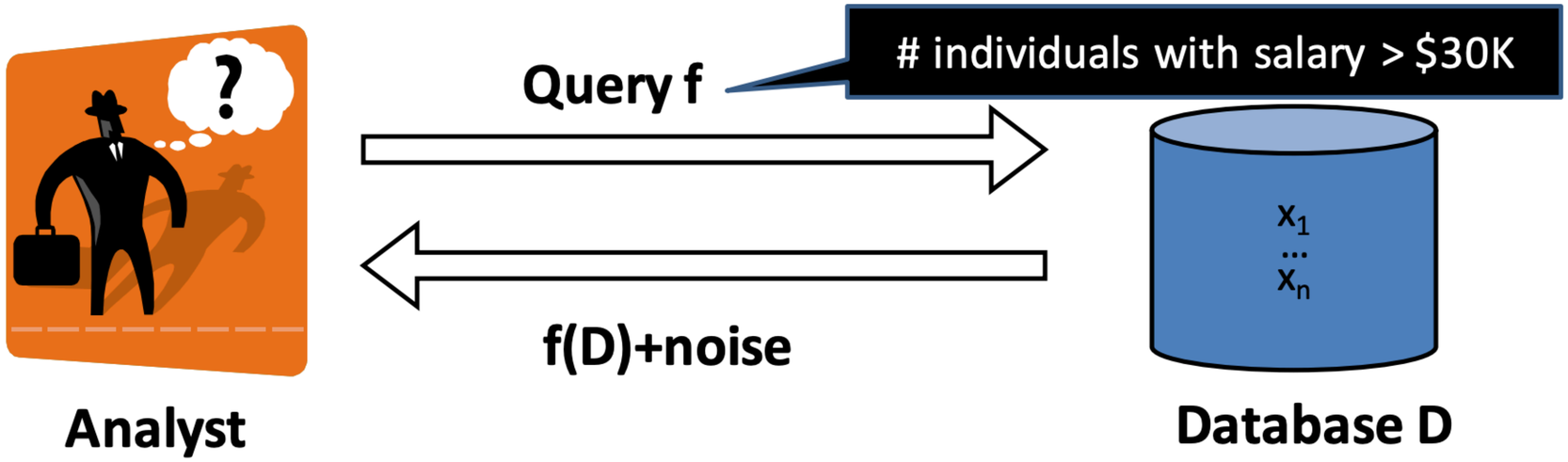
Isolation Attack!

r_1	4	1	0

Amplification Attacks

- If two records agree on rare attributes, then with high probability they agree on other attributes too.
- Use intuition to find record *r* similar to *aux* on many rare attributes (using aux as ‘proxy’ for y).

Privacy-Preserving Statistics: Interactive Setting



Goals:

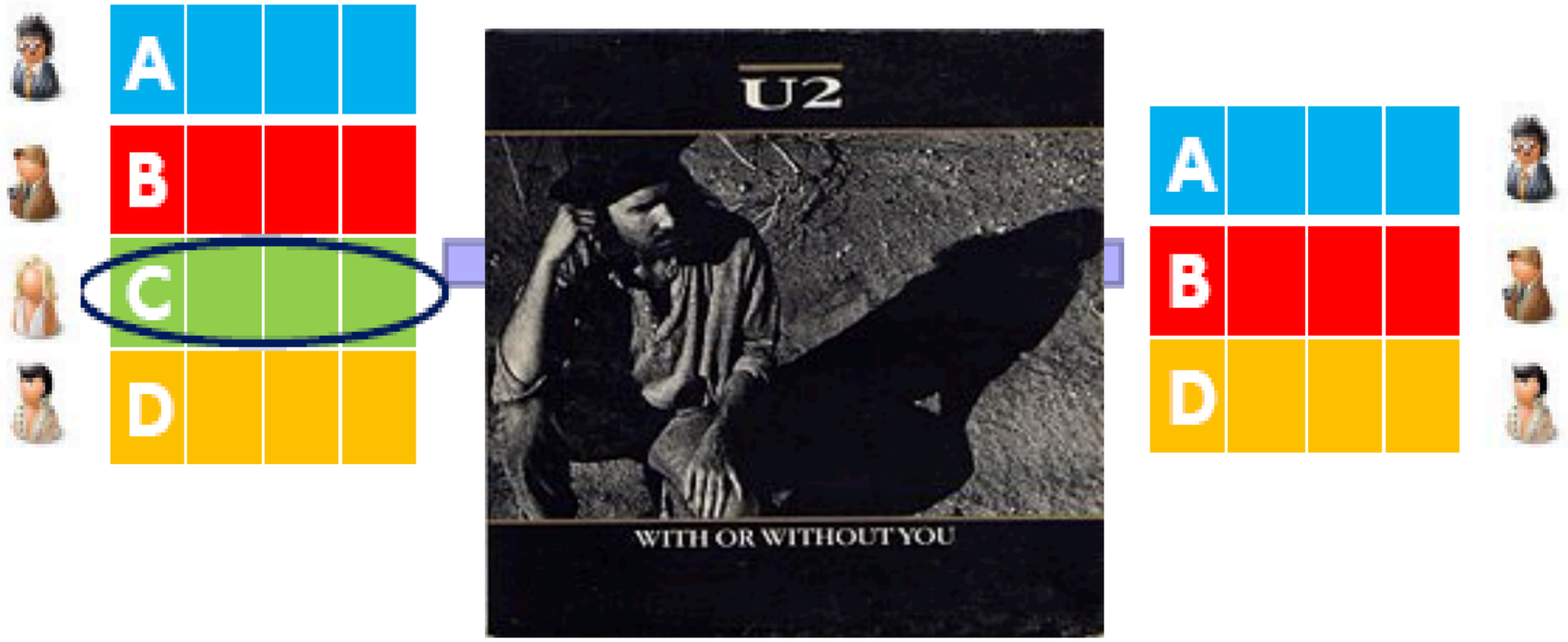
- Accurate statistics (low noise)
- Preserve individual privacy
  (what does that mean?)

- Census data
- Health data
- Network data
- …

Some possible defenses

- Anonymize data
  - Re-identification, information amplification
- Queries over large data sets
  - Differencing attack
- Query auditing
  - Refusal leaks, computational tractability
- Summary statistics
  - Frequency lists

Differential Privacy



Released statistic is about the same
if any individual's record is
removed from the database

An information flow idea

Changing input databases in a specific way changes output statistic by a small amount

Does not guarantee that Terry Gross's height won't be learned by the adversary

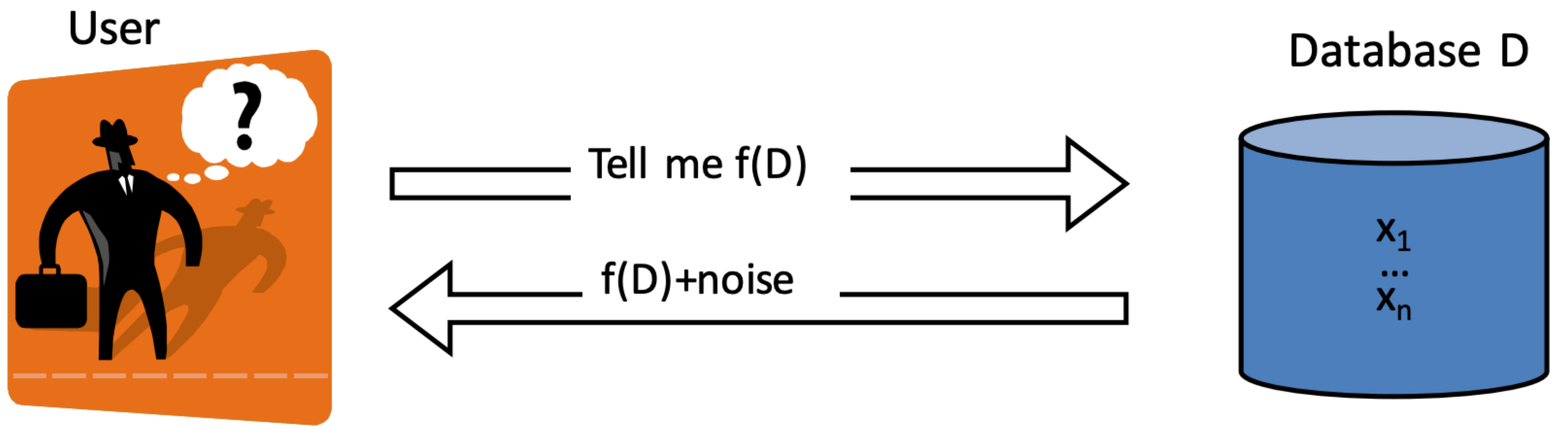
Differential Privacy: Definition

Randomized sanitization function κ has ε-differential privacy if for all data sets D_1 and D_2 differing by at most one element and all subsets S of the range of κ,

$$\Pr[\kappa(D_1) \in S] \leq e^{\varepsilon} \Pr[\kappa(D_2) \in S]$$

Answer to query # individuals with salary > $30K is in range [100, 110] with approximately the same probability in D_1 and D_2

Achieving Differential Privacy: Interactive Setting



How much and what type of noise should be added?

Demo & implementations

https://programming-dp.com/

Beauty of Differential Privacy

1. Post processing:
    1. Given M(x), do whatever you want with it.
    2. f(M(x)) satisfies DP
2. Composability (budget):
    1. Run ε-DP mechanism, and then another ε-DP mechanism. Full interaction is 2ε-DP.
    2. Run ε-DP on a subset, and then another ε-DP on another subset, full interaction is ε-DP.

Challenges around Differential Privacy

1. No way to verify.
2. Runs out the budget quickly.
3. User trust issues.
4. Hard to implement.

Today's class

1. K-anonymity
2. Differential privacy
3. Implementation
4. No free lunch in data privacy

Credits

1. Duke University, CompSci 590.03, Ashwin Machanavajjhala
2. CMU, 18734 Foundations of Privacy, Anupam Datta

No free lunch in data privacy

- Three popularized claims
  - It makes no assumptions about how data are generated.
  - It protects an individual's information if an attacker knows about all other individuals in the data.
  - It is robust to arbitrary background knowledge.