Lecture 4: Statistical Privacy Tools

CMSC 25910

Winter 2026

The University of Chicago

THE UNIVERSITY OF CHICAGO

Initial Approaches for Protecting Privacy

Old (Deprecated) Idea: k-anonymity

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

This data has 2-anonymity with respect to the attributes 'Age', 'Gender' and 'State of domicile' since for any combination of these attributes found in any row of the table there are always at least 2 rows with those exact attributes. The attributes available to an adversary are called quasi-identifiers. Each quasi-identifier tuple occurs in at least *k* records for a dataset with *k*-anonymity.[14]

Figure taken from https://en.wikipedia.org/wiki/K-anonymity

Randomized Response

- An old (but good) idea initially developed in the social sciences
- Example: Are you enjoying CS 259?

Randomized Response

- Are you enjoying CS 259?
- Flip a coin:
  - If tails, then tell the truth
  - If heads, then flip a coin again:
    - If heads, say ‘yes’
    - If tails, say ‘no’
- What does this achieve?

Randomized Response

- Privacy is achieved because we cannot know with certainty what your answer was
  - With an unbiased coin, at least 25% of answers will be ‘no,’ while at least 25% of answers will be ‘yes’
- Yet we can obtain useful aggregate results
  - Because we know how the noise was introduced
  - Let’s see how…

Randomized Response

- Proportion of yes answers is the sum of:
  - Probability of flipping tails (“tell the truth”) * the proportion of honest “yes” answers
  - Probability of flipping heads (“lie”) * probability of flipping heads (“say ‘yes’ no matter the honest answer”)
- Rearrange and solve for the proportion of honest “yes” answers!

Database Privacy via Differential Privacy

Outline

- **Building Intuition**
- Differential Privacy (DP)
- Local vs. Centralized Model
- Composition and Privacy Budget

Statistical Database Privacy

- Idea: Nothing about an individual is learned from dataset D_1 that cannot be learned from the same dataset without the individual's data (D_2)

Outline

- Building Intuition
- **Differential Privacy (DP)**
- Local vs. Centralized Model
- Composition and Privacy Budget

Differential Privacy

Differential Privacy: Intuitive Definition

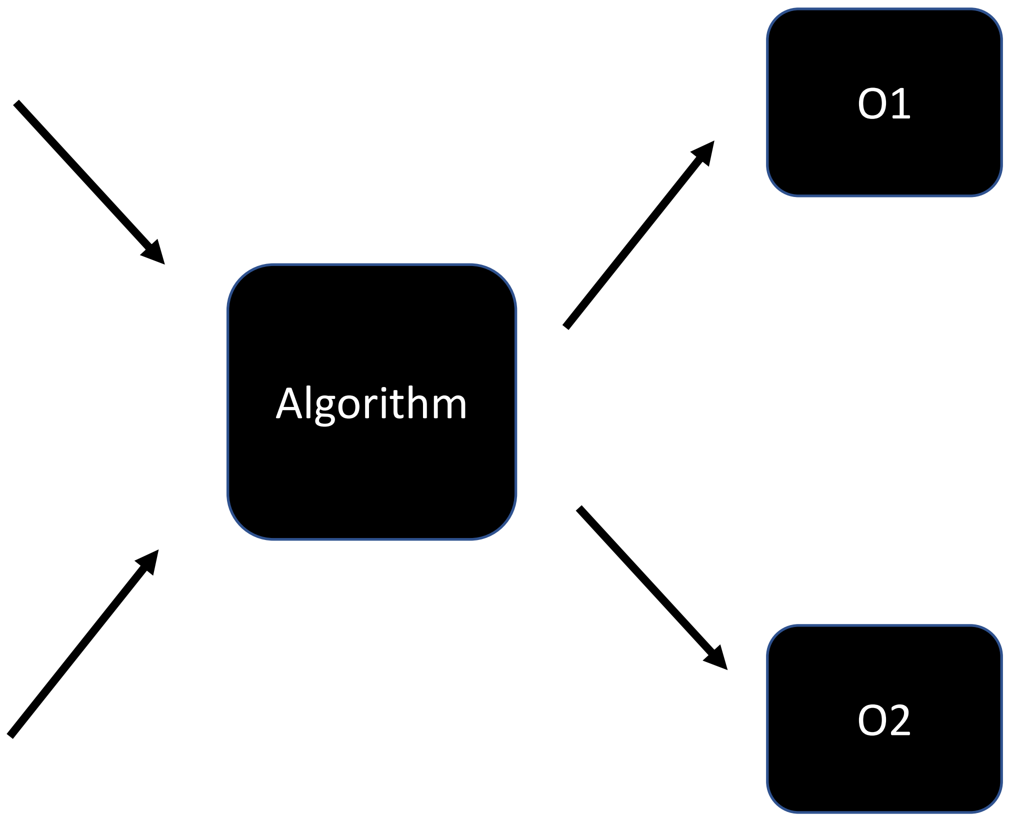
- It is not possible to tell if the input to an algorithm, *A*, contained an individual's data or not just by looking at the output, *O*, of *A*
  - No one can learn much about one individual from the dataset
- Including your data in a dataset does not increase your chances of being harmed
  - No matter the data
  - No matter the algorithm/query

Differential Privacy Definition

- For every pair of input datasets, D_1, D_2 that differ in one row
  - One row: presence or absence of a single record (individual)
- For every output, *O*, computed via an algorithm, *A*…
- Adversary cannot distinguish D_1 from D_2 based on *O* with more than a negligible probability
- An algorithm is differentially private if its output is *insensitive* to the presence or absence of a single row.

EID	First Name	Last Name	Department
43	Jill	Smith	CS
33	Josh	Hartford	Econ
53	Jill	Corn	Bio

EID	First Name	Last Name	Department
33	Josh	Hartford	Econ
53	Jill	Corn	Bio



Differential Privacy Definition

- For every pair of input datasets, D_1, D_2 that differ in one row
  - One row: presence or absence of a single record (individual)
  - We can call these **neighboring datasets** or **adjacent datasets**
- For every output, *O*, computed via an algorithm, *A*…
- Adversary cannot distinguish D_1 from D_2 based on *O* with more than a negligible probability

$$\ln\left(\frac{P(A(\mathrm{D}_1)=O)}{P(A(\mathrm{D}_2)=O)}\right) \leq \varepsilon$$

What is Epsilon?

- Epsilon is our privacy budget, specifying how much a function's result could differ between neighboring datasets

$$ln\left(\frac{P(A(D_1)=O)}{P(A(D_2)=O)}\right) \leq \varepsilon$$

- Smaller epsilon means higher privacy.
    - Consider epsilon = 0

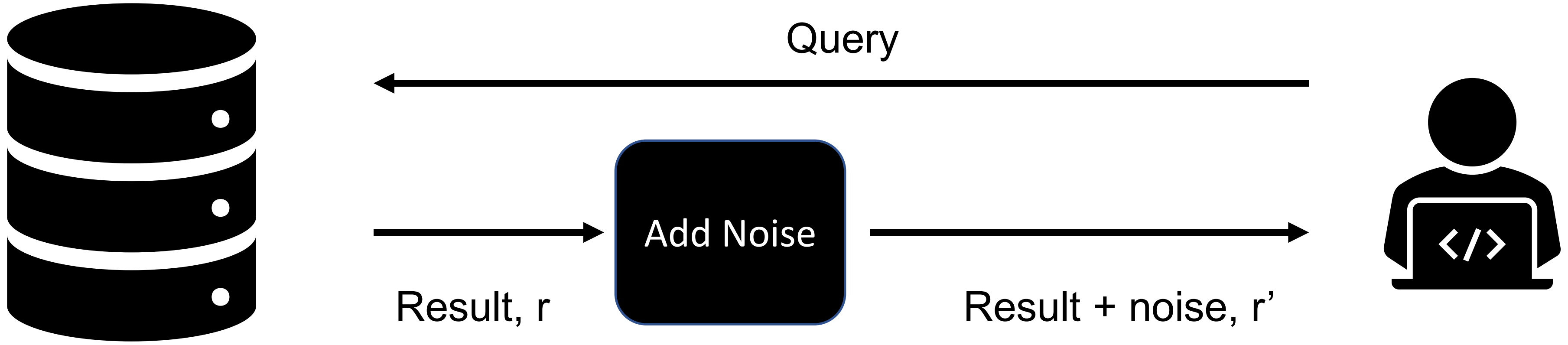
Approaches

- Randomized Response (How We Started the Lecture)
- Laplace Mechanism
- Exponential Mechanism

Algorithms

- Randomized Response
- **Laplace Mechanism**
- Exponential Mechanism

Laplace Mechanism



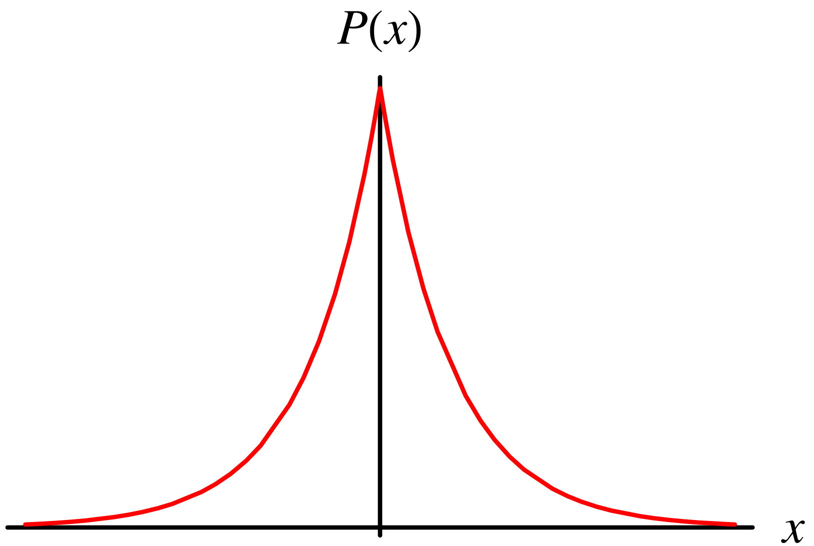
Laplace mechanism works for numerical results

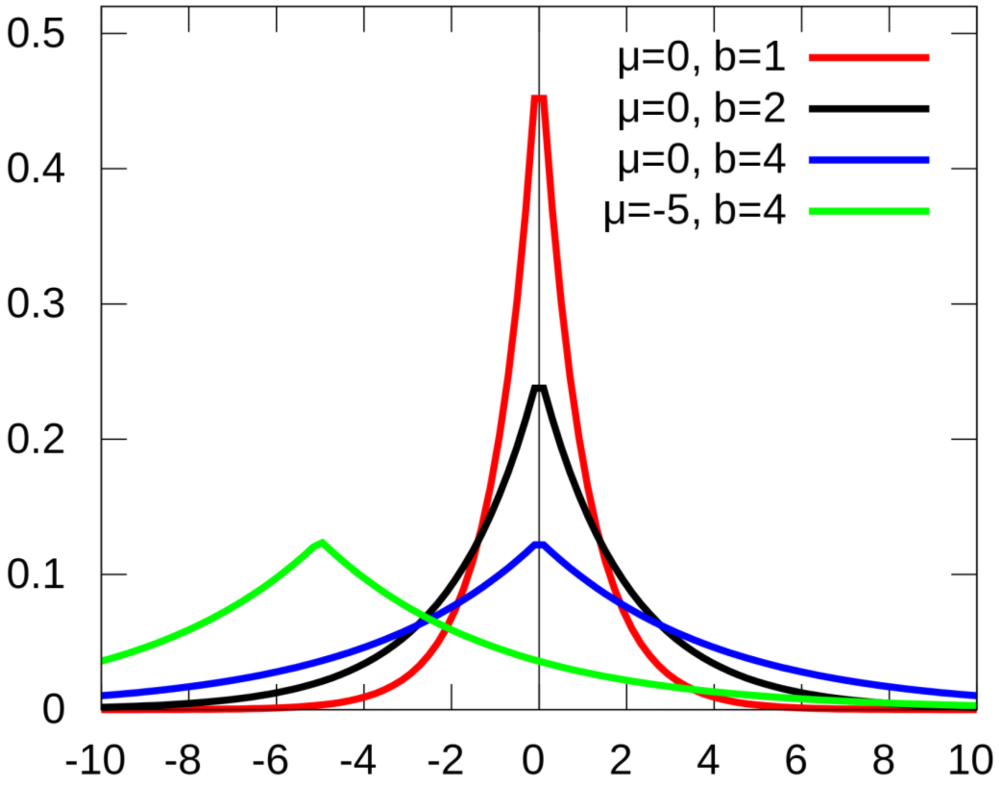
How Do We Add Noise?

- We want to add noise so that:
  - The noisy answer does not leak private information
    - Keep DP definition in mind
  - The noisy answer is useful
- Laplace mechanism adds noise sampling from a Laplace distribution

(What is a Laplace Distribution?!)

$$P(x) = \frac{1}{2b} e^{-|x-\mu|/b}$$

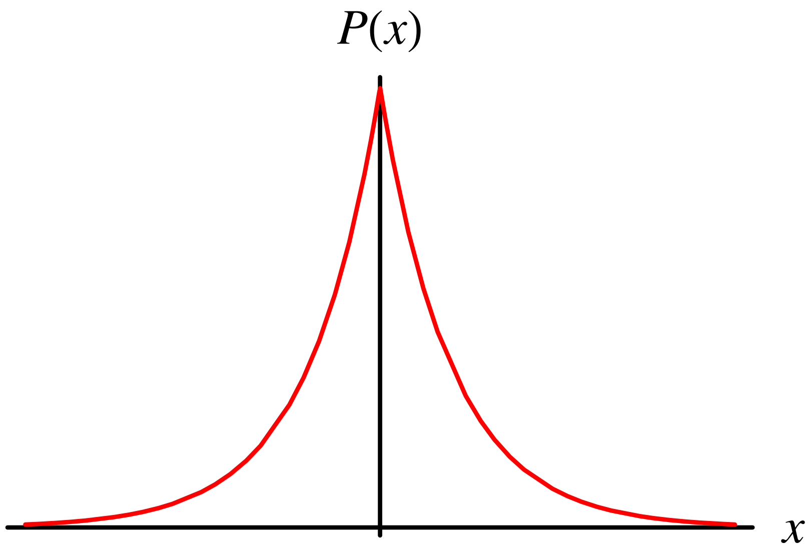




Formula and left figure taken from https://mathworld.wolfram.com/LaplaceDistribution.html
and right figure taken from https://en.wikipedia.org/wiki/Laplace_distribution

(What is a Laplace Distribution?!)

$$P(x) = \frac{1}{2b} e^{-|x-\mu|/b}$$



Definition 3.2 (The Laplace Distribution). The Laplace Distribution (centered at 0) with scale b is the distribution with probability density function:

$$\text{Lap}(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$$

The variance of this distribution is $\sigma^2 = 2b^2$. We will sometimes write $\text{Lap}(b)$ to denote the Laplace distribution with scale b, and will sometimes abuse notation and write $\text{Lap}(b)$ simply to denote a random variable $X \sim \text{Lap}(b)$.

Left formula and left figure taken from https://mathworld.wolfram.com/LaplaceDistribution.html
and right definition taken from https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Calculating the Sensitivity

- **Sensitivity**: The maximum change one individual's data can change the function computed on the database
  - Basically, the maximum difference in the answer between neighboring datasets D_1, D_2

Definition 3.1 (ℓ_1-sensitivity)**.** The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ is:

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\mathcal{X}|} \\ \|x-y\|_1 = 1}} \|f(x) - f(y)\|_1.$$

The ℓ_1 sensitivity of a function f captures the magnitude by which a single individual's data can change the function f in the worst case, and therefore, intuitively, the uncertainty in the response that we must introduce in order to hide the participation of a single individual.

Definitions taken from https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Example: Sensitivity

- Imagine that you have a database of employee salaries
- You want to know **how many** employees make >= $100,000
- What's the maximum change achieved by varying 1 record?
  - That's the sensitivity!

Example: Sensitivity

- Imagine that you have a database of employee salaries
- You want to know **how many** employees make >= $100,000
- What's the maximum change achieved by varying 1 record?
  - That's the sensitivity!
- The answer is 1 (this is a "count" query)

Example: Sensitivity

- Imagine that you have a database of employee salaries
- You want to know **how much your company is paying** in salary each year
- What’s the maximum change achieved by varying 1 record?
  - That’s the sensitivity!

Example: Sensitivity

- Imagine that you have a database of employee salaries
- You want to know **how much your company is paying** in salary each year
- What's the maximum change achieved by varying 1 record?
  - That's the sensitivity!
- Here, you have to figure out what is the maximum permitted salary since that is the sensitivity! In other words, you need to reason about the range

How Do We Add Noise?

Definition 3.3 (The Laplace Mechanism). Given any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the Laplace mechanism is defined as:

$$\mathcal{M}_L(x, f(\cdot), \varepsilon) = f(x) + (Y_1, \ldots, Y_k)$$

where Y_i are i.i.d. random variables drawn from $\text{Lap}(\Delta f / \varepsilon)$.

Theorem 3.6. The Laplace mechanism preserves $(\varepsilon, 0)$-differential privacy.

$$\text{Lap}(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$$

Definitions taken from https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

What’s the Utility of Laplace Mechanism?

- Utility: how useful is the answer?
- Intuitively, how close is to the real answer
- Think of the tradeoff between privacy (epsilon) and utility
- For more details, see Chapter 3.3 of https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

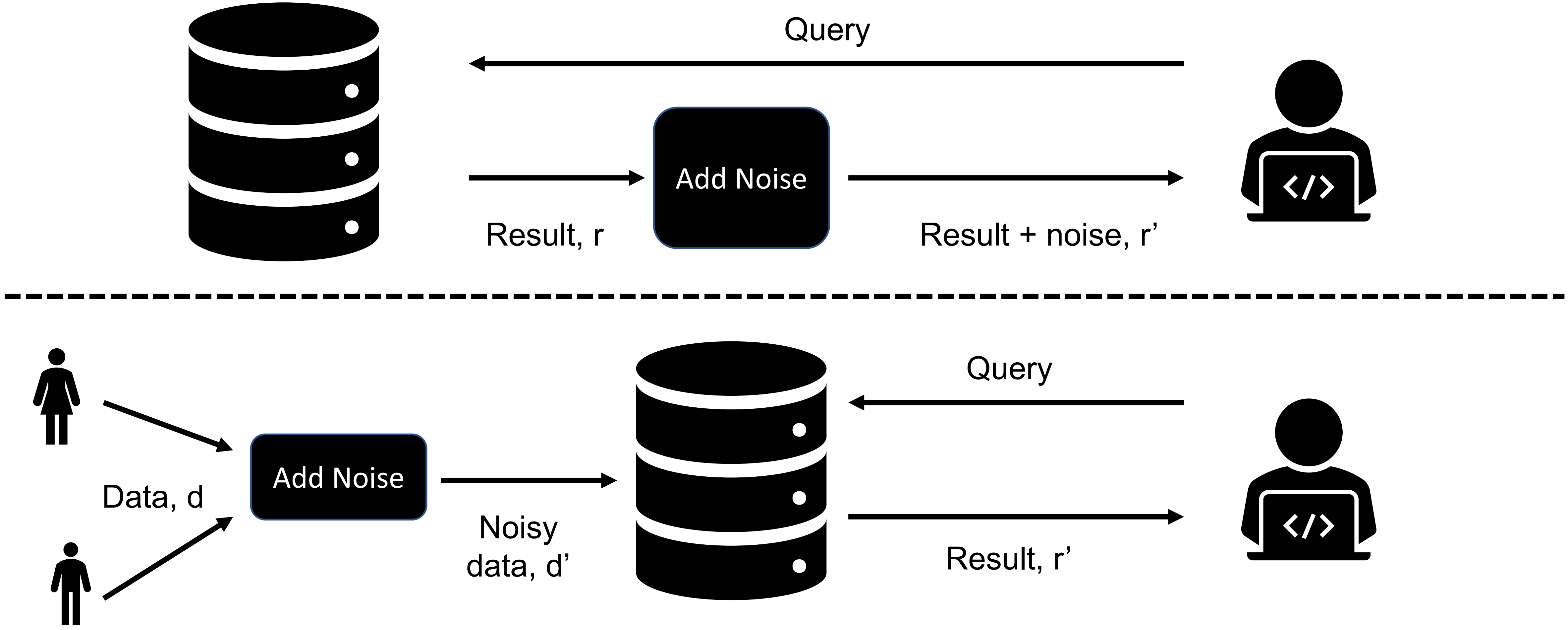
Exponential Mechanism

- When the answer of an algorithm is categorical, not numerical
  - Won’t get into details in this class; see Chapter 3.4 of https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

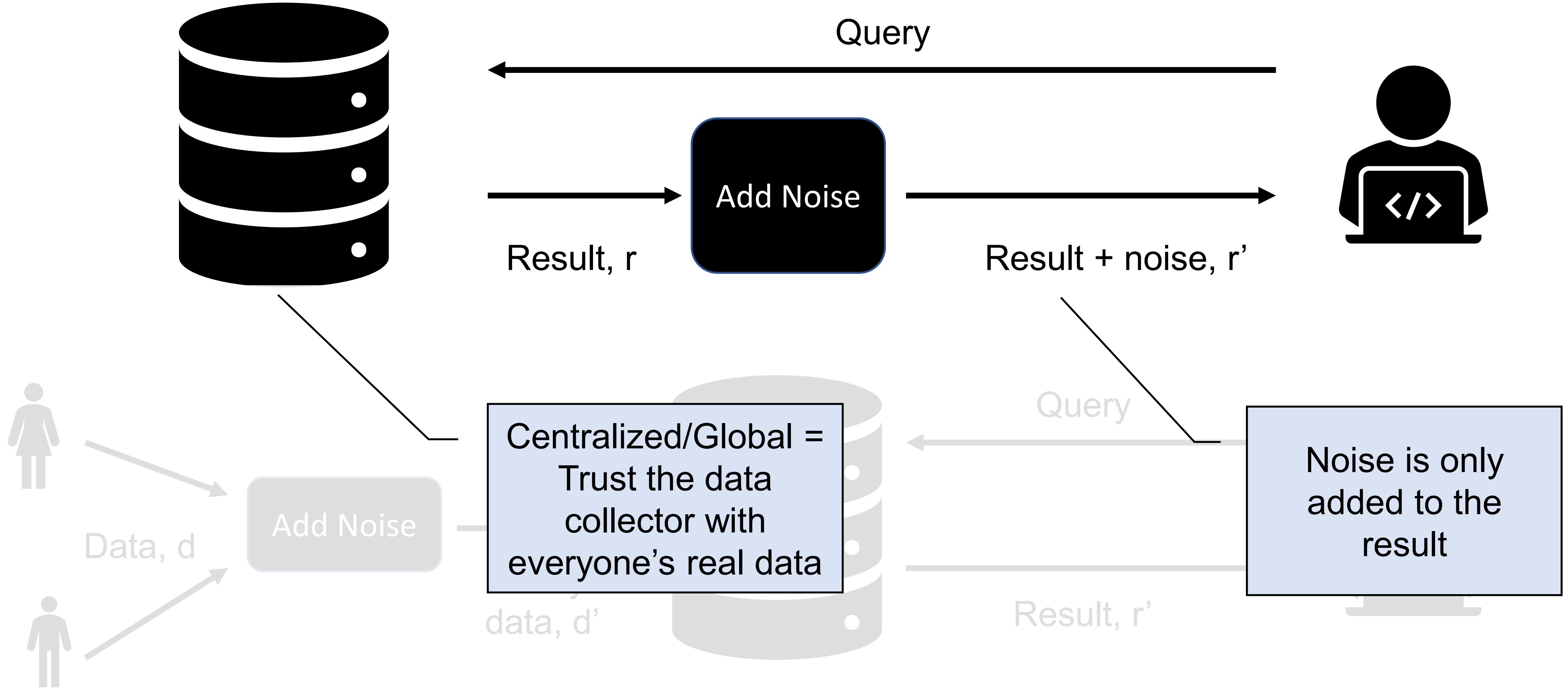
Outline

- Building Intuition
- Differential Privacy
- **Local vs. Centralized Model**
- Composition and Privacy Budget

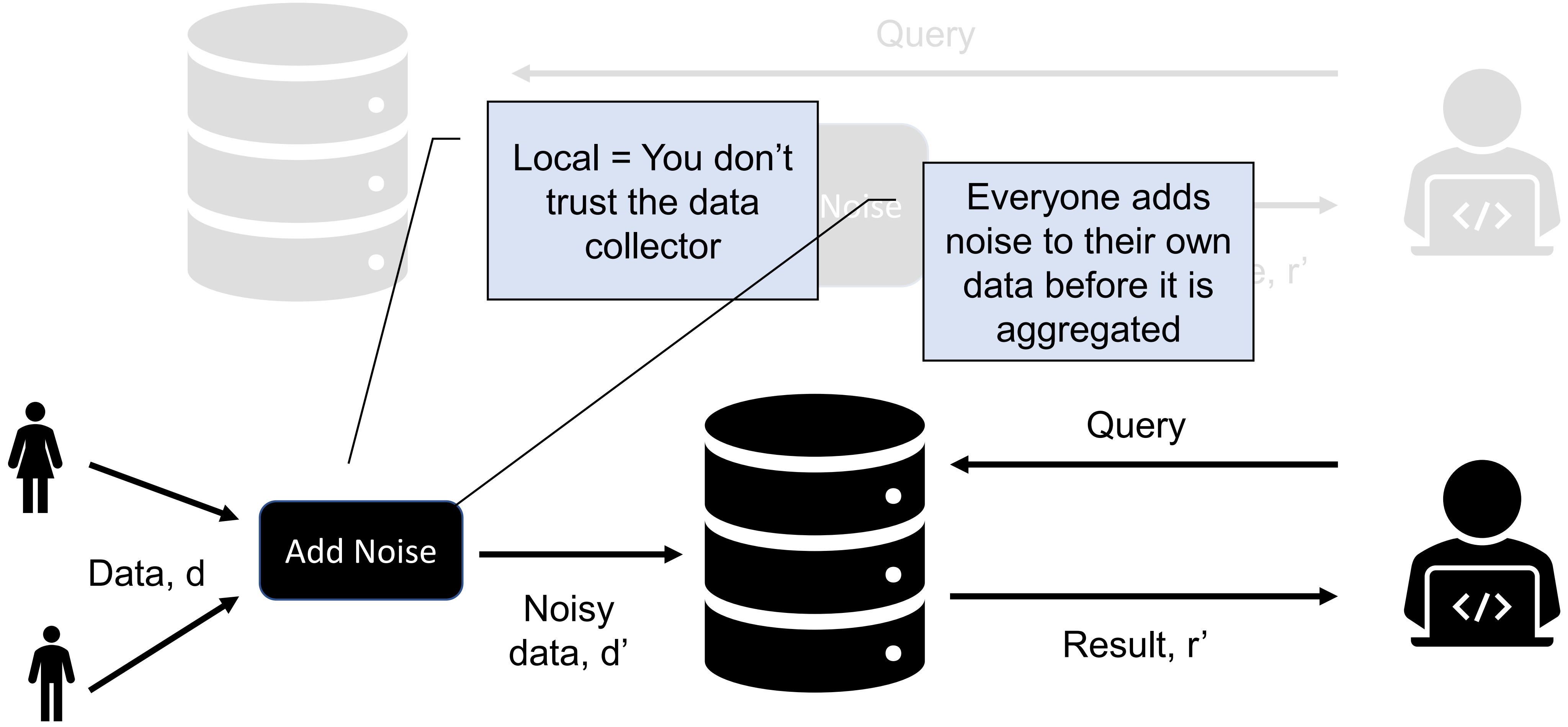
Centralized (Top) vs. Local (Bottom)



Centralized (Top) vs. Local (Bottom)



Centralized (Top) vs. Local (Bottom)



Outline

- Building Intuition
- Differential Privacy
- Local and Decentralized Model
- **Composition and Privacy Budget**

Composition and the Privacy Budget

- Build more complicated (and useful) algorithms from primitive building blocks
- Composition rules help us reason about privacy budgets
  - Serial composition
    - If you run n DP-algorithms, serially, the resulting algorithm is ε'-DP
    - $\varepsilon' = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n$
  - Parallel composition
    - When running n DP-algorithms on disjoint data, the resulting algorithm is $\max(\varepsilon_i)$
  - Postprocessing: F(M()), if M is DP-private, then output of F is too
- A hope of DP is to design algorithms that do not *consume much budget* and yet produce good quality results, but we are not there yet as a community

Census 2020

- Centralized model. Collect clean data (as usual) but release differentially private results only
  - CIA, FBI, IRS cannot ask for census data by law

2020.

(b) QUALITY.—Data products and tabulations produced by the Bureau of the Census pursuant to sections 141(b) or (c) of title 13, United States Code, in connection with the 2020 decennial census shall meet the same or higher data quality standards as similar products produced by the Bureau of the Census in connection with the 2010 decennial census.

g:\VHLC\051220\051220.072.xml (763351|3)
May 12, 2020 (12:13 p.m.)

https://hdsr.mitpress.mit.edu/pub/dgg03vo6/release/2

Differentially Private Analytics

- Locally private. Google Chrome and iPhones add noise to records before sending them to the companies
- Makes sense; customers may not trust these companies!
- Companies may need to release subpoenaed datasets

Apple uses local differential privacy to help protect the privacy of user activity in a given time period, while still gaining insight that improves the intelligence and usability of such features as:

- QuickType suggestions
- Emoji suggestions
- Lookup Hints
- Safari Energy Draining Domains
- Safari Autoplay Intent Detection (macOS High Sierra)
- Safari Crashing Domains (iOS 11)
- Health Type Usage (iOS 10.2)

Examples taken from https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Chrome vs. Apple

- Chrome released its DP code (RAPPOR)
- Apple did not
  - In some cases, Apple also resets the privacy budget daily
  - https://www.macobserver.com/analysis/google-apple-differential-privacy
- How much can you trust a DP implementation without knowing parameters like epsilon?

Tradeoffs and Caveats of DP

- Utility vs. Privacy
  - How to choose parameters?
  - Which model, centralized or local?
  - Do you produce results once? Or do you let people query the DB?
    - What happens to the privacy budget if you just let people query the DB?
- Privacy budget
  - This can be limited by the user
    - Users can talk to each other, though
  - Make sure you understand what DP guarantees!
- DP usually assumes independent data, no auxiliary data