Lecture 11: Statistical Privacy

CMSC 25910

Spring 2023

The University of Chicago

THE UNIVERSITY OF CHICAGO

Today’s lecture

- Discuss statistical definitions of privacy
- Understand **differential privacy (DP)**
  - What it is used for
  - When it helps
  - When it does not help

Outline

- **Building Intuition**
- Differential Privacy (DP)
- Local vs. Centralized Model
- Composition and Privacy Budget
- What DP is Not

Membership Attacks

- Is a particular data subject included in a dataset?
  - What does membership in a particular dataset imply?

Goal of Statistical Database Privacy

- Release useful information **without leaking private information**
  - Permit inference about a population without disclosing individual records
- Quantify/bound amount of information disclosed about individual
- First attempt at a definition: ‘Ability to perform data analysis over a *dataset* without producing *harm* to any *individual* whose record is in the dataset’

Old Idea: k-anonymity

- "A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least $k - 1$ individuals whose information also appear in the release."

https://en.wikipedia.org/wiki/K-anonymity

Old Idea: k-anonymity

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	30	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

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

Old 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

Statistical Database Privacy

- Better Definition: 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
- What DP is Not

Differential Privacy: Intuitive Definition

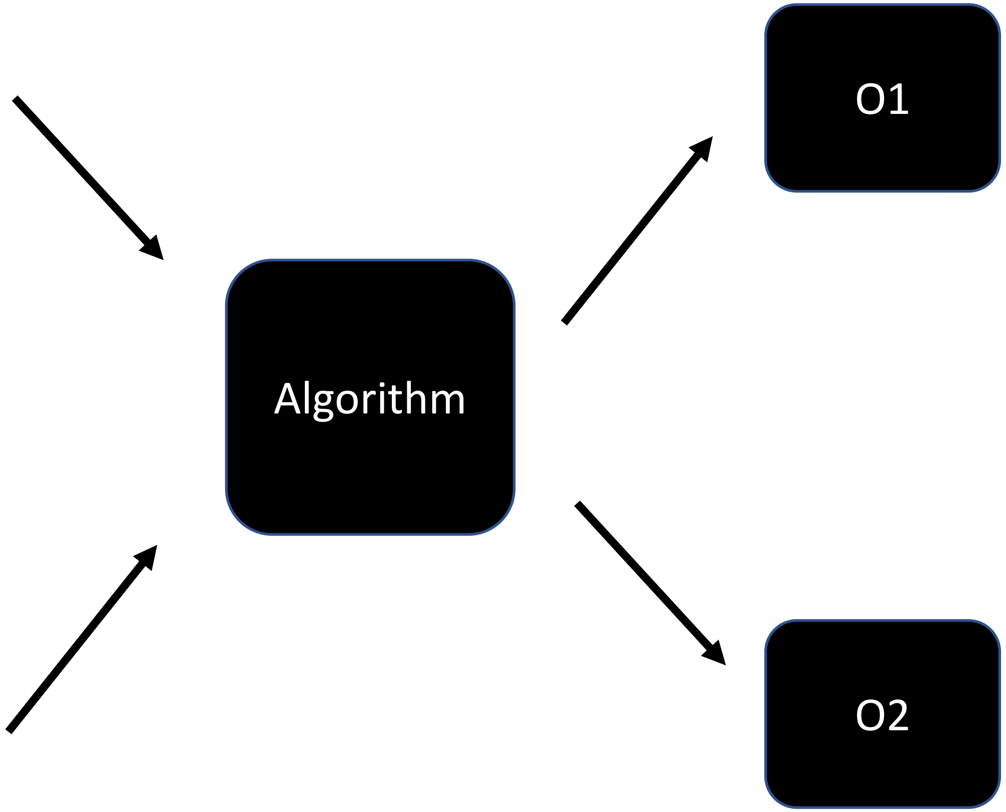
- It is not possible to tell if the input to an algorithm, *A*, contained an individual's data or not just by looking at the output, *O*, of *A*
  - No one can learn much about one individual from the dataset
- Including your data in a dataset does not increase your chances of being harmed
  - No matter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

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

*The algorithm *A* is often referred to as the mechanism

What is Epsilon?

- Epsilon determines how *insensitive* is the output to the input datasets

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

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

Algorithms

- Randomized Response
- Laplace Mechanism
- Exponential Mechanism

Randomized Response

- 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’
- Yet we can obtain useful aggregate results
  - Because we know how the noise was introduced
  - Let’s see how…

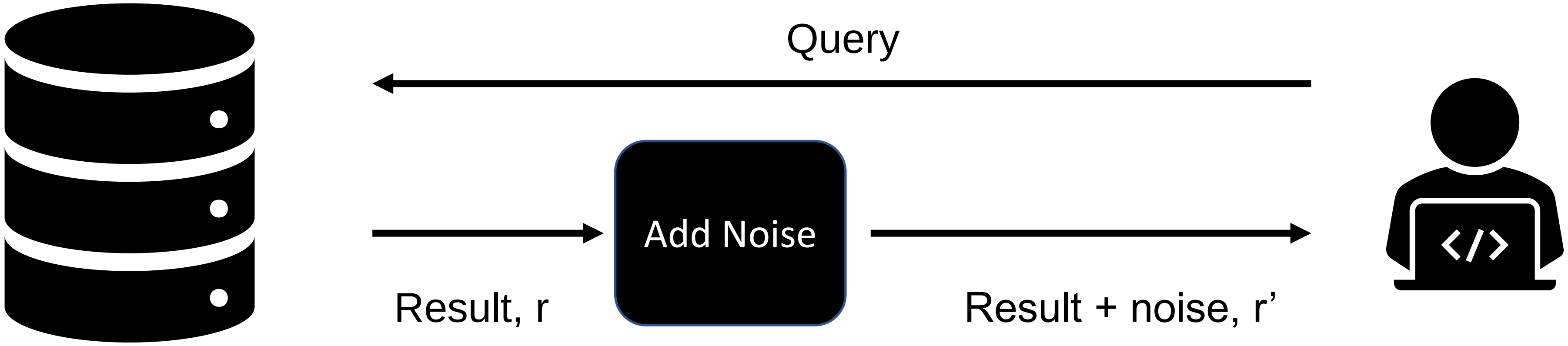
Randomized Response

- 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!

Algorithms

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

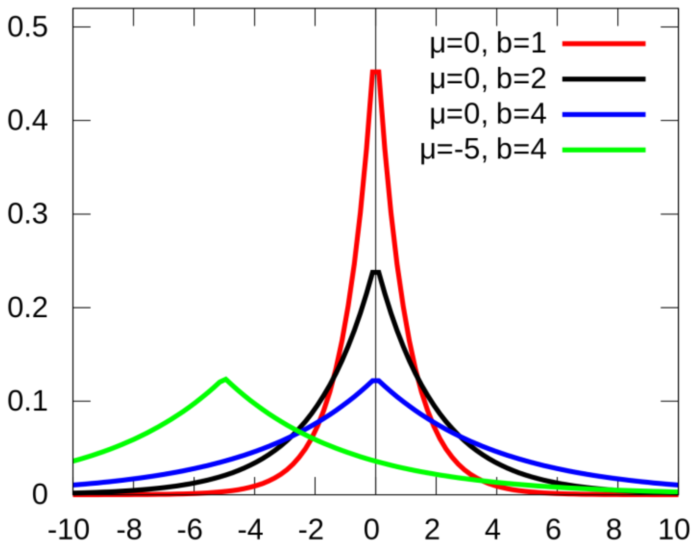
Laplace Mechanism



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



- Mean, $\mu = 0$
- Variance = $2 * \lambda^2$
- Typically refer to: Lap(λ)

How do we choose λ?

- $\lambda = S/\varepsilon$
- S is the *Sensitivity*: property of the query/algorithm computed over neighboring datasets, D, D’
  - Intuitive definition of sensitivity: The maximum change one row can cause to the output of the query
- Selecting λ as above guarantees ε-DP answer

Example: SUM query

- SELECT SUM(salary) FROM employee;
- What's the maximum change achieved by varying 1 record?

Salary
35
33
34
48
47

Salary
33
34
48
47

Salary
35
33
34
47

Example: SUM query

- SELECT SUM(salary) FROM employee;
- What’s the maximum change achieved by varying 1 record?

Salary
35
33
34
48
47

Salary
33
34
48
47

Salary
35
33
34
47

- If data is in range [a,b] (assuming a and b are both positive)
  - Sensitivity of SUM is b
- What’s the sensitivity of COUNT()?

What's the Utility of Laplace Mechanism?

- Utility: how useful is the answer?
- Intuitively, how close is to the real answer
  - $E(\text{true\_answer} - \text{noisy\_answer})^2$
- 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

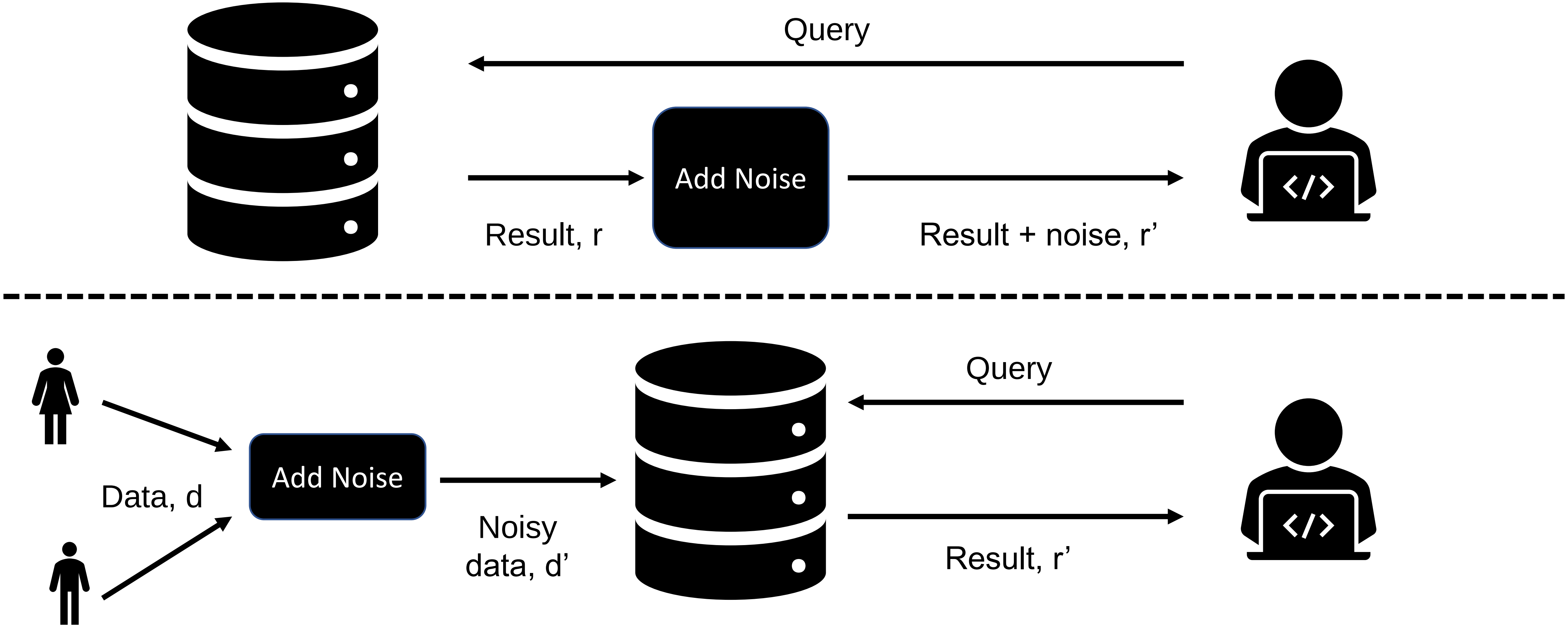
Exponential Mechanism

- 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

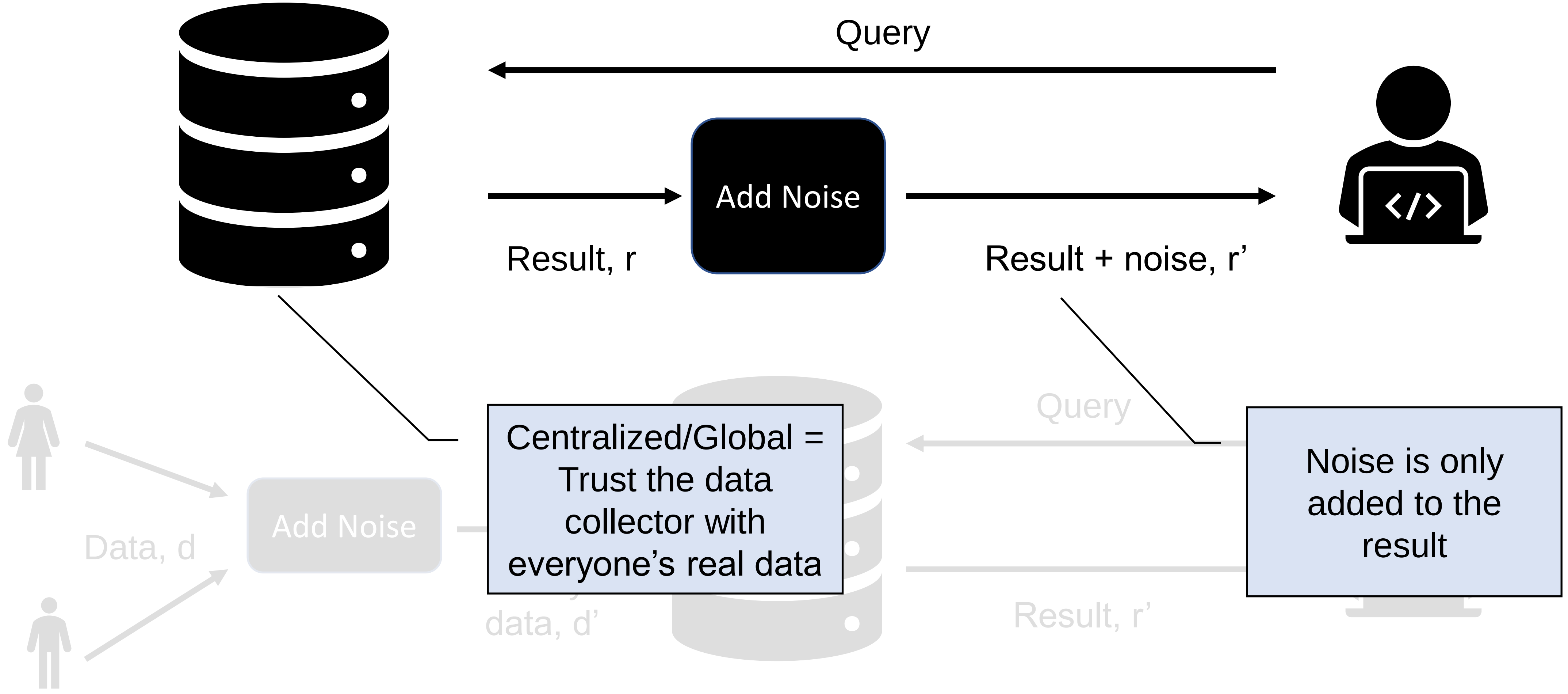
Outline

- Building Intuition
- Differential Privacy
- **Local vs. Centralized Model**
- Composition and Privacy Budget
- What DP is not designed for

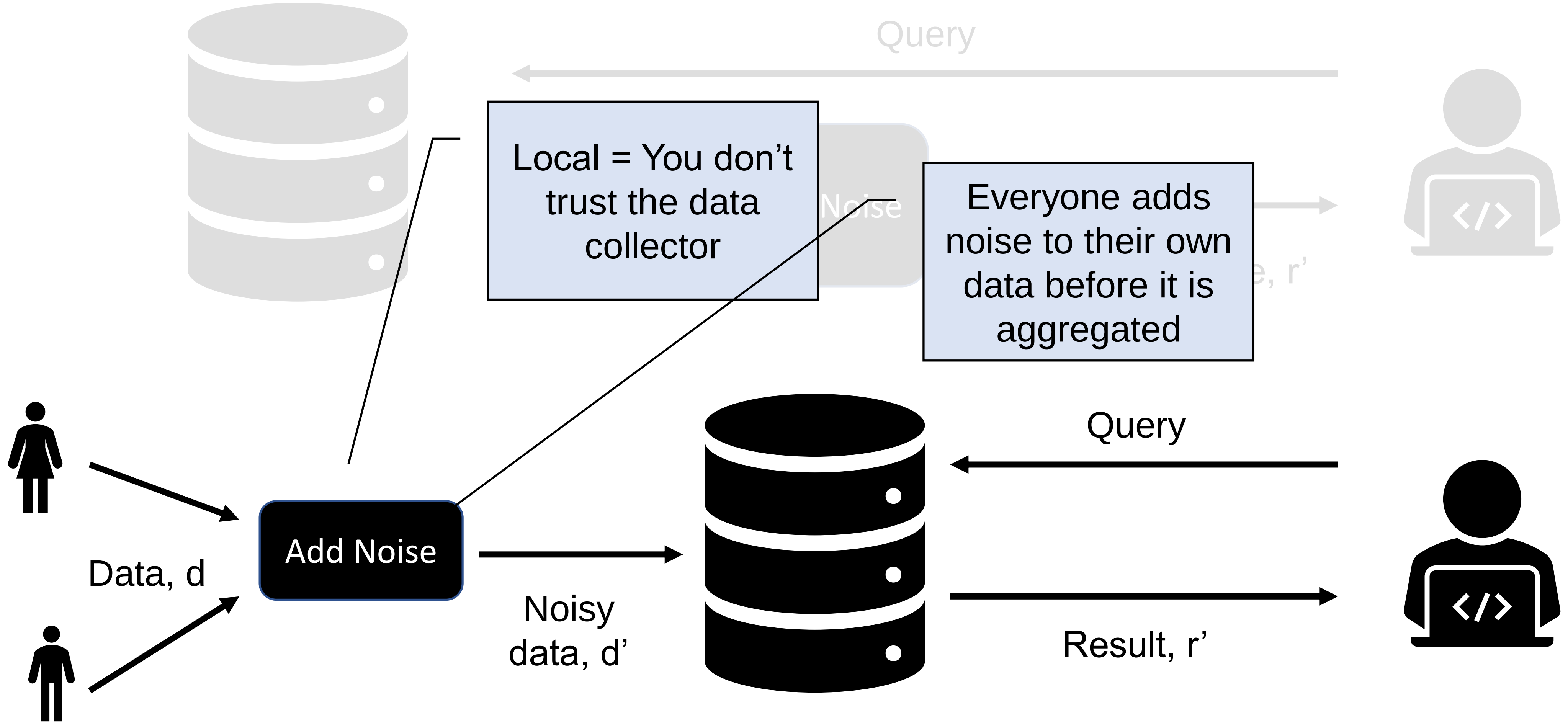
Centralized (Top) vs. Local (Bottom)



Centralized (Top) vs. Local (Bottom)



Centralized (Top) vs. Local (Bottom)



Outline

- Building Intuition
- Differential Privacy
- Local and Decentralized Model
- **Composition and Privacy Budget**
- What DP is not designed for

Composition

- 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(ε_i)
  - Postprocessing: F(M()), if M is DP-private, then output of F is too
- A hope of DP is to design algorithms that don’t *consume much budget* and yet produce good quality results

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

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
- Surveillance on Google's data centers

Chrome vs. Apple

- Chrome released its DP code (RAPPOR)
- Apple didn’t
  - 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?

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.)

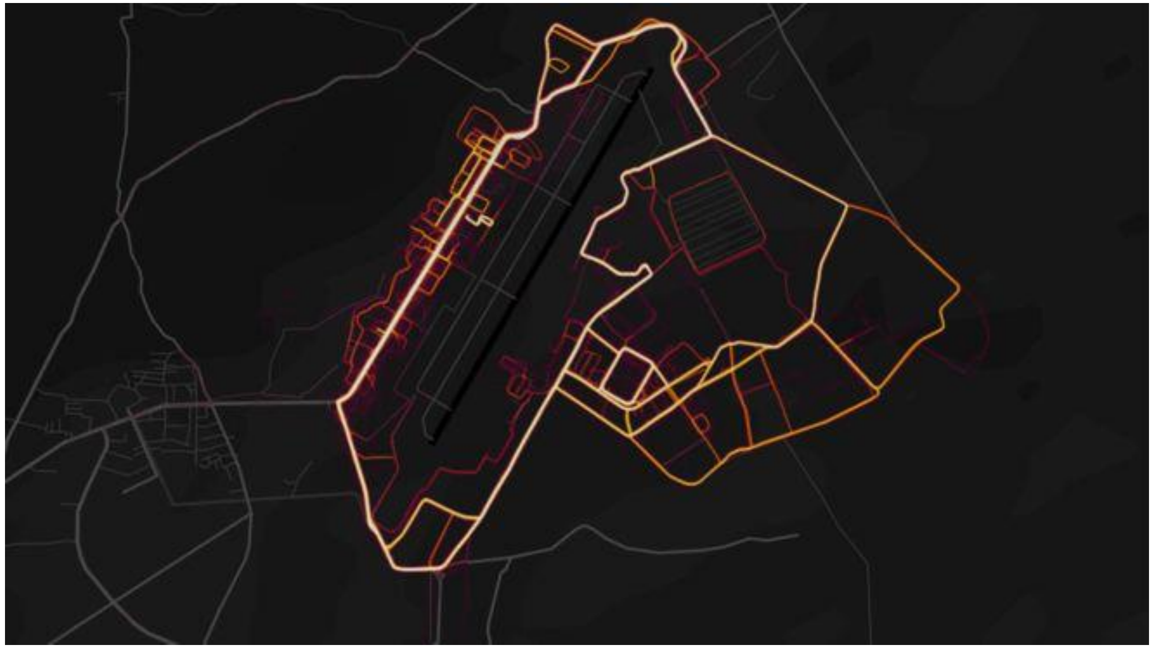
https://hdsr.mitpress.mit.edu/pub/dgg03vo6/release/2

Outline

- Building Intuition
- Differential Privacy
- Local vs. Centralized Model
- Composition and Privacy Budget
- **What DP is Not**

What DP is Not

From bbc.com



- Fitness app Strava published a heatmap showing the paths users log as they run or cycle
- Can you know the identity of a single user?
  - Does DP help?
- Can you identify any other ’privacy’ problems?

What DP is Not

From bbc.com

- Fitness app Strava published a heatmap showing the paths users log as they run or cycle
- Can you know the identity of a single user?
    - Does DP help?
- Can you identify any other 'privacy' problems?