

Statistical Database Privacy Techniques

Today's lecture

- To discuss *practical* definitions of privacy
- To understand what is 'differential privacy'
 - What it is useful for
 - When it helps
 - When it does not help

Outline

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

Privacy based on Anonymization

- Reidentification
 - Latanya Sweeney

Membership Attacks

- GWAS
 - Did x participate in a study?
- Shadow training in ML

Data Security is not Privacy

- To keep data private, encrypt

Data Security is not Privacy

- To keep data private, encrypt
- But then we preclude usage of that data downstream
 - Train a ML model or create any other derived data product
 - Encrypting data reduces its usability

Data Security is not Privacy

- To keep data private, encrypt
- But then we preclude usage of that data downstream
 - Train a ML model or create any other derived data product
 - Encrypting data reduces its usability
 - Although some techniques are evolving fast, i.e., homomorphic encryption, multi-party computation, etc.

- How do we release data without leaking private information?

Goal of Statistical Database Privacy

- Release useful information without leaking private information
 - Permit inference about population, but not the disclose of individual records

Goal of Statistical Database Privacy

- Release useful information **without leaking private information**
 - Permit inference about population, but not the disclose of individual records

*Fundamental Law of Information Recovery says this is impossible

Goal of Statistical Database Privacy

- Release useful information without leaking private information
 - Permit inference about population, but not the disclosure of individual records
- Quantify and bound amount of information disclosed about an individual

Goal of Statistical Database Privacy

- Release useful information without leaking private information
 - Permit inference about population, but not the disclose of individual records
- Quantify and bound amount of information disclosed about an individual
- One definition: ‘Ability to perform data analysis over a *dataset* without producing *harm* to any *individual* whose record is in the dataset’

Smoking before Doll and Hill 1/2

- Imagine an individual, X, who is a smoker
- X participates in a medical study
- Study finds link between smoking -> cancer
- Putting this facts together, one would increase X's likeliness of developing cancer
 - This may be harmful for X (think of insurance)
- Link was found using, among others, X's data

The mortality of doctors in relation to their smoking habits. Doll and Hill. 1954

Smoking before Doll and Hill 1/2

- Imagine an individual, X, who is a smoker
- X participates in a medical study
- Study finds link between smoking -> cancer
- Putting this facts together, one would increase X's likeliness of developing cancer
 - This may be harmful for X (think of insurance)
- Link was found using, among others, X's data
- **Is this a privacy violation?**

The mortality of doctors in relation to their smoking habits. Doll and Hill. 1954

Smoking before Doll and Hill 2/2

- In a world without the tobacco study, then X's privacy is not violated
- Consider two parallel worlds. In one, X participates in Doll and Hill, in the other X does not participate
 - X's data is part of the dataset in one world but not the other

Statistical Database Privacy

- First attempt: ‘Ability to perform data analysis over a dataset without producing harm to any individual whose record is in the dataset’

Statistical Database Privacy

- First attempt: 'Ability to perform data analysis over a dataset without producing harm to any individual whose record is in the dataset'
- Definition: Nothing about an individual is learned from a dataset, D , that cannot be learned from the same dataset but without the individual's data, D'

Outline

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

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 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

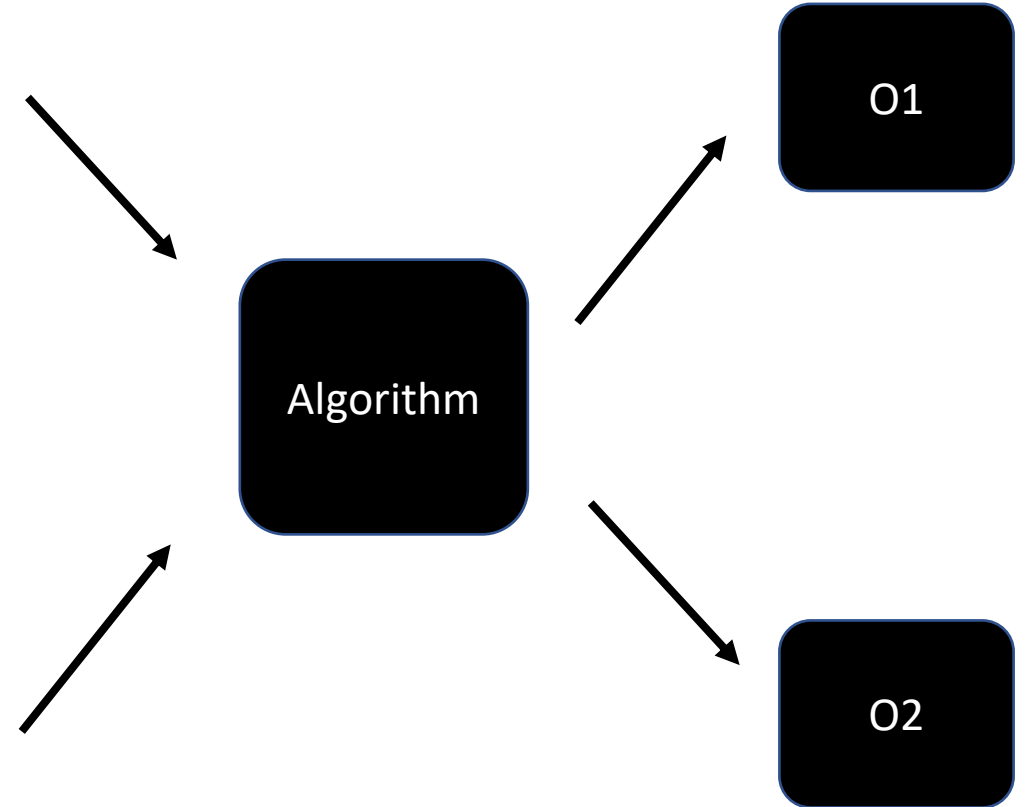
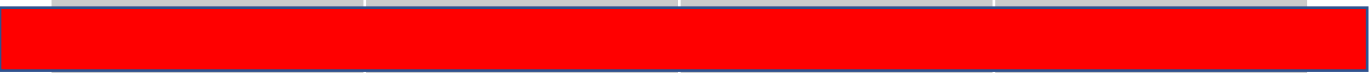
Differential Privacy Definition

- For every pair of input datasets, $D1$, $D2$ 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 differentiate $D1$ from $D2$ based on O

- 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, $D1$, $D2$ 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 differentiate $D1$ from $D2$ based on O

$$\ln \left(\frac{P(A(D^1)=o)}{P(A(D^2)=o)} \right) \leq \epsilon$$

*The algorithm, A , is often referred to as 'privacy mechanism' or simply mechanism

What is Epsilon?

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

$$\ln \left(\frac{P(A(D^1)=o)}{P(A(D^2)=o)} \right) \leq \epsilon$$

What is Epsilon?

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

$$\ln \left(\frac{P(A(D^1)=o)}{P(A(D^2)=o)} \right) \leq \epsilon$$

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

DP is a definition

- There are algorithms to produce differentially-private answers

$$\ln \left(\frac{P(A(D^1)=o)}{P(A(D^2)=o)} \right) \leq \epsilon$$

Algorithms

- Randomized Response
- Laplace Mechanism
- Exponential Mechanism

Randomized Response

- Are you enjoying 259?

Randomized Response

- Are you enjoying 259?
- Flip a coin:
 - If tails, then say the truth
 - If heads, then flip a coin again:
 - If heads, say 'yes'
 - If tails, say 'no'

Randomized Response

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

Randomized Response

- Are you enjoying 259?
- **What does this achieve?**
- Privacy is achieved because we cannot know with certainty what your answer was.
 - With an unbiased coin, at least 25% of answer will be 'no'
- And yet, we can obtain useful aggregate results
 - Because we know how the noise was introduced
 - Let's see how...

Randomized Response

- Flip a coin:
 - If tails, then say the truth
 - If heads, then flip a coin again:
 - If heads, say 'yes'
 - If tails, say 'no'
- Probability of saying 'yes' when the ground truth is 'yes'
 - $2/4$ (tails) + $1/4$ (heads + heads) = $3/4$
- Probability of saying 'yes' when the ground truth is 'no'
 - $1/4$ (heads + tails)

Randomized Response

- Probability of saying 'yes' when the ground truth is 'yes'
 - $2/4$ (tails) + $1/4$ (heads + heads) = $3/4$
- Probability of saying 'yes' when the ground truth is 'no'
 - $1/4$ (heads + tails)
- Ps of saying 'no' are equivalent to the above

Randomized Response

- Probability of saying ‘yes’ when the ground truth is ‘yes’
 - $2/4$ (tails) + $1/4$ (heads + heads) = $3/4$
- Probability of saying ‘yes’ when the ground truth is ‘no’
 - $1/4$ (heads + tails)
- Ps of saying ‘no’ are equivalent to the above

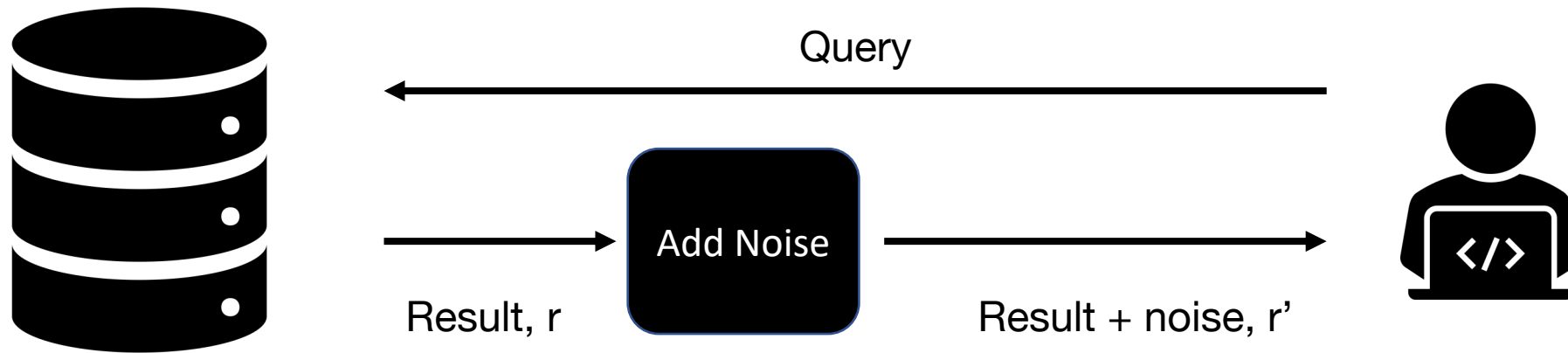
$$\ln \left(\frac{P(A(D^1)=o)}{P(A(D^2)=o)} \right) \leq \varepsilon \quad 3/4 / 1/4 = 3, \text{ so } \varepsilon = \ln(3)$$

- RP using an unbiased coin is $\ln(3)$ -differentially private

Algorithms

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

Laplace Mechanism



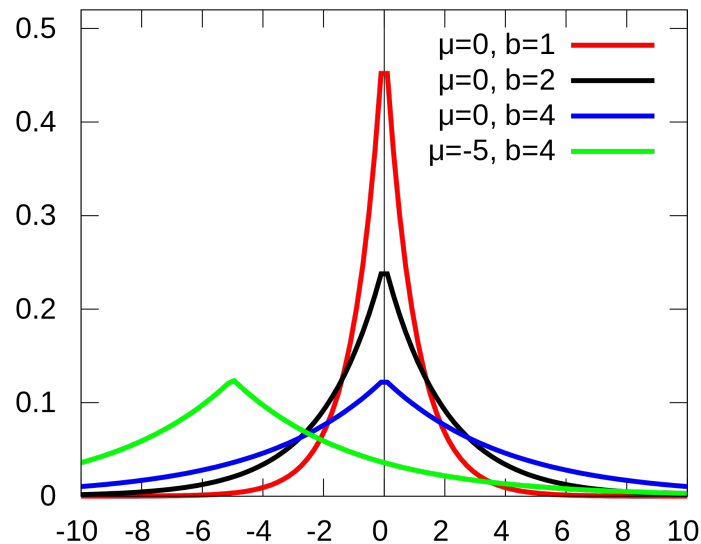
Laplace mechanism works for numerical results

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 by sampling from a Laplace dist.

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 by sampling from a Laplace dist.



- Mean, $\mu = 0$
- Variance = $2 * \lambda^2$
- Typically refer to: $\text{Lap}(\lambda)$
- How do we choose λ ?

How do we choose λ ?



- $\lambda = S/\varepsilon$

How do we choose λ ?

- $\lambda = S/\epsilon$
- S is the *Sensitivity*: property of the query/algorithm, computed over neighboring datasets, D, D'
- Intuitive definition of Sensitivity: The maximum change 1 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 where dep=CS;
- What's the maximum change achieved by varying 1 record?

Salary	Salary	Salary
35		35
33	33	33
34	34	34
48	48	
47	47	47

Example: SUM query

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

Salary	Salary	Salary
35		35
33	33	33
34	34	34
48	48	
47	47	47

- If data is in range $[a,b]$
- Sensitivity of SUM is b

- What's the sensitivity of COUNT()?

What's the Utility of Laplace Mech?

- 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

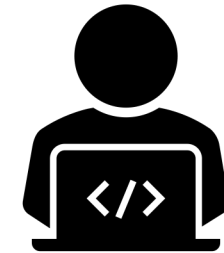
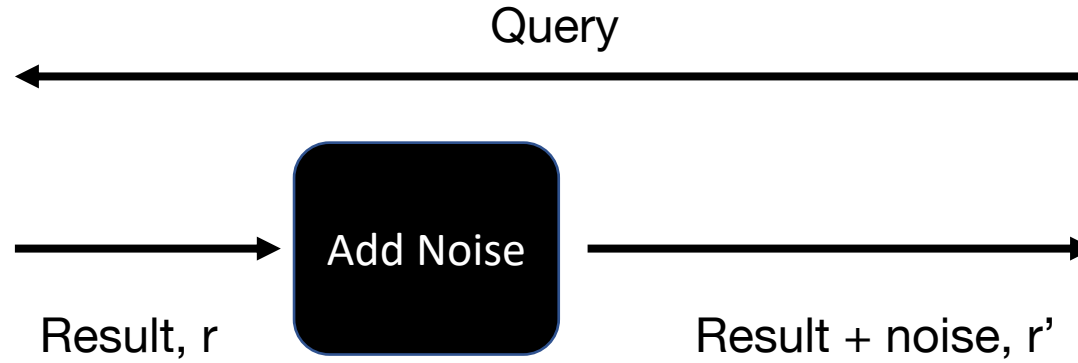
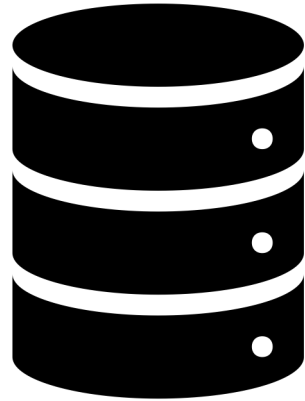
Exponential Mechanism

- When the answer of an algorithm is not numerical but categorical
 - Won't get in details...

Outline

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

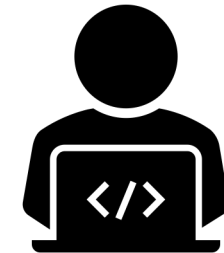
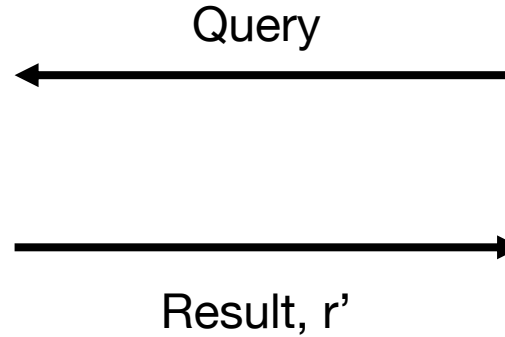
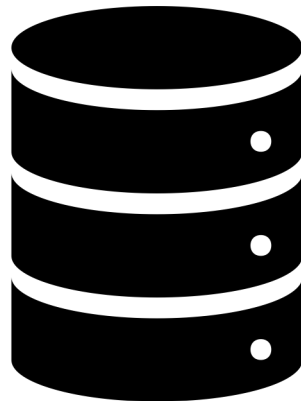
RP vs LM



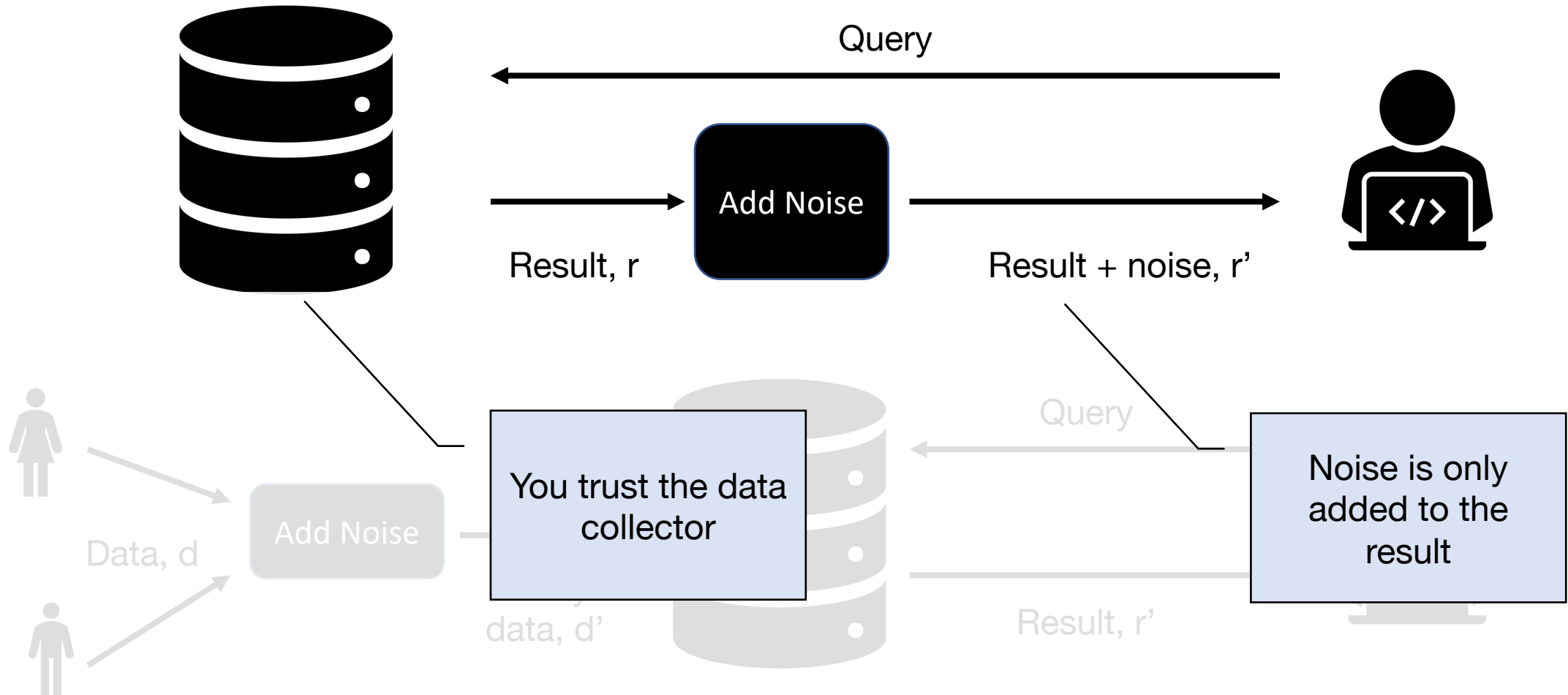
Data, d

Add Noise

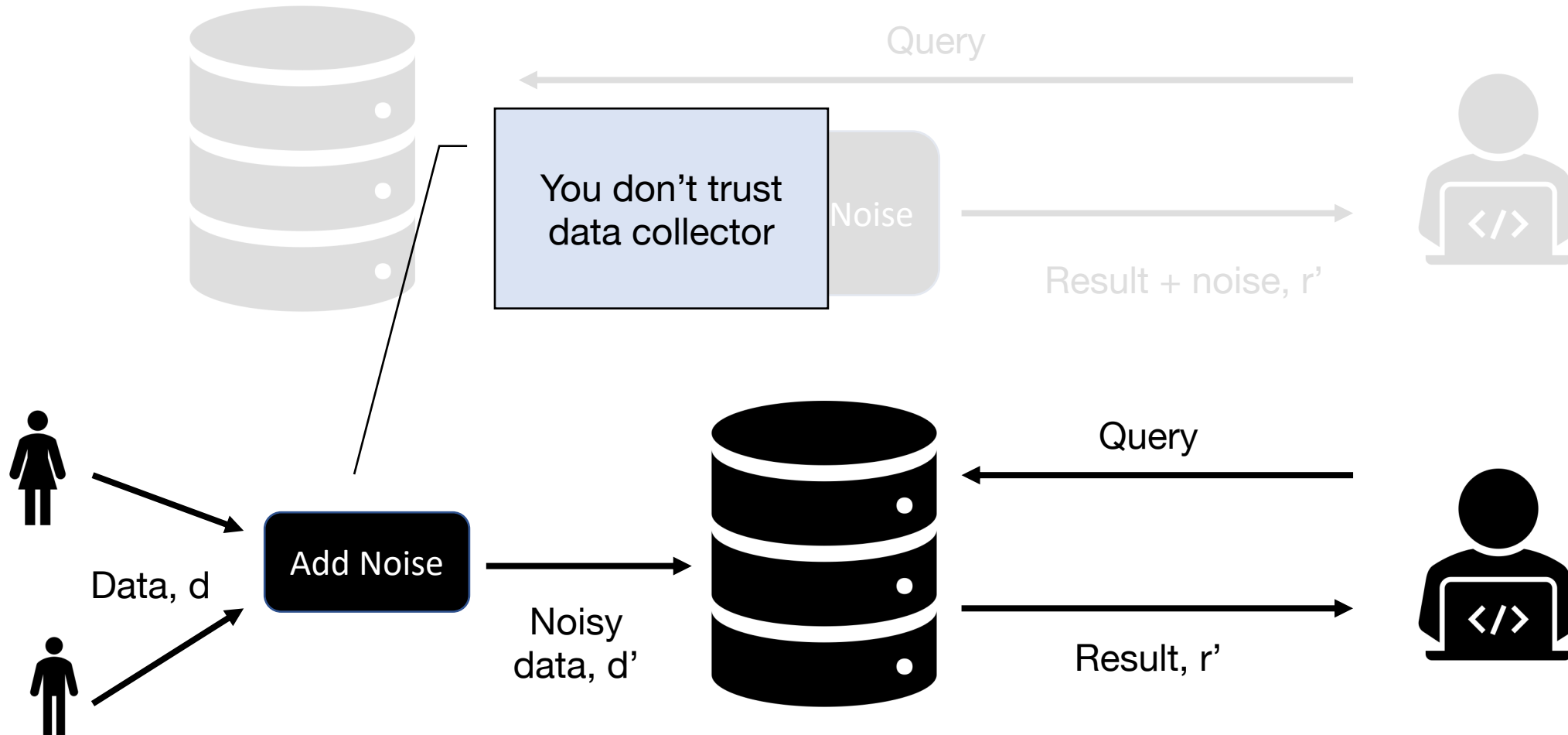
Noisy data, d'



RP vs LM



RP vs LM



RP vs LM

- Local or Decentralized vs Centralized
 - Is data collection differentially private or only the answers?

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

- 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
 - $\epsilon' = \epsilon_1 + \epsilon_2 + \dots + \epsilon_n$
 - Parallel composition
 - When running n DP-algorithms on disjoint data, the resulting algorithm is $\max(\epsilon_i)$
 - Postprocessing. $F(M())$, if M is DP-private, then output of F is too
- A lot of the magic 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
 - What model, centralized vs local, to choose
 - Interactive vs offline release
 - Do you produce results once? Or do you let people query the DB?
 - What happens if you just let people query the DB?
- Privacy budget
 - This can be limited by user.
 - Users can talk to each other, though
 - Make sure you understand what DP guarantees

Use Cases

Chrome browser and iPhone usage stat

- Locally private. Chrome and iPhones add noise to records before sending them to the companies
- Makes sense because customers may not trust these companies
- Companies may need to release subpoenaed datasets
- Snowden and NSA surveillance on Google's data centers

Chrome vs Apple

- Chrome releases its DP code (RAPPOR)
- Apple doesn't
- How much can you trust a DP implementation without knowing the parameters?
 - i.e., 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

18 2020.

19 (b) QUALITY.—Data products and tabulations pro-
20 duced by the Bureau of the Census pursuant to sections
21 141(b) or (c) of title 13, United States Code, in connection
22 with the 2020 decennial census shall meet the same or
23 higher data quality standards as similar products pro-
24 duced by the Bureau of the Census in connection with the
25 2010 decennial census.

Census 2020

- <https://hdr.mitpress.mit.edu/pub/dgg03vo6/release/2>

Outline

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

What DP is not good for



What DP is not good for

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?

What DP is not good for

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 good for

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?