DS 593: Privacy in Practice

Differential Privacy and Other Privacy Mechanisms

Instructor: Alishah Chator

News?

Last time

- Introduction to Cryptography

Today

- Differential Privacy

- Other Building Blocks

Anonymity

Anonymity

- A specific flavor of privacy focused on hiding identity (or generally, identifying information)

- Requires somehow removing distinguishing characteristics, so that the re-identification is impossible

- “Anonymity loves company”

Pseudonymity

- A way of reducing personally identifiable information

- Pseudonyms still permit reidentification with additional information

- Easier to create and work with
  - Need strong guarantees on what would lead to re-identification

- Examples of pseudonymity vs anonymity?

A Societal Tension

- Suppose you have a dataset with information of a population

- The individuals in the dataset may all wish to not be identified

- However, the dataset may potentially uncover valuable findings

- How can we navigate this?

Dataset Privacy

- Goal: preserve the useful information in a dataset needed for analysis, while removing everything else

- Could try to do this by removed features that are not relevant

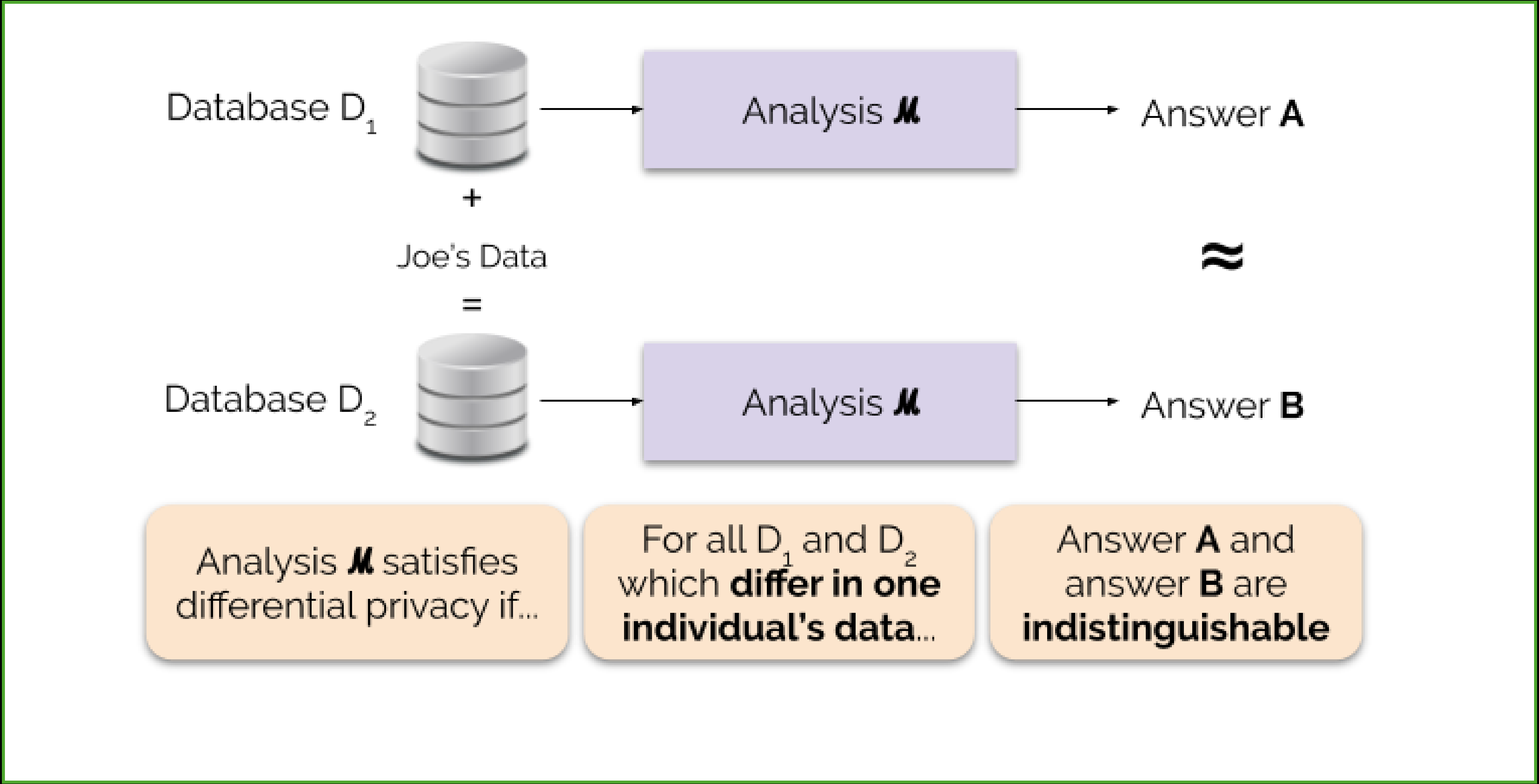
- Only provide aggregated statistics

- Replace PII with pseudonyms

What exactly are the privacy goals here?

- Even aggregated statistics leak information about the contents of the dataset
    - EX: counts, means, etc
- Need a ways that successive queries can’t be used for re-identification
- Could we build a system for analysis that is not sensitive to individual rows?
    - That is, we can somehow add some fuzziness or noise that masks the impact on any specific person being present in the dataset and the resulting analysis

Differential Privacy (DP)



Analysis $\mathcal{M}$ satisfies differential privacy if...

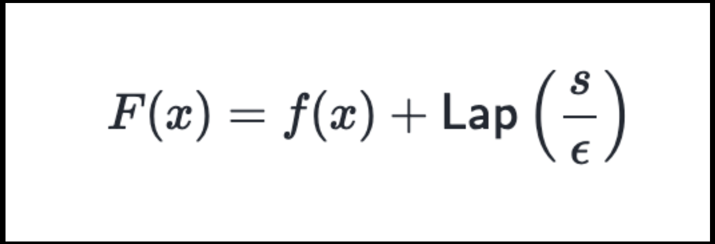
For all D_1 and D_2 which **differ in one individual's data**...

Answer **A** and answer **B** are **indistinguishable**

How to build Differential Privacy

- Need the right *release mechanism*

- Have to consider the sensitivity of the particular query you want to perform and add the appropriate amount of noise
    - Ex: count queries



$$F(x) = f(x) + \mathsf{Lap}\left(\frac{s}{\epsilon}\right)$$

Considerations with Differential Privacy

- It is primarily an individual privacy notion

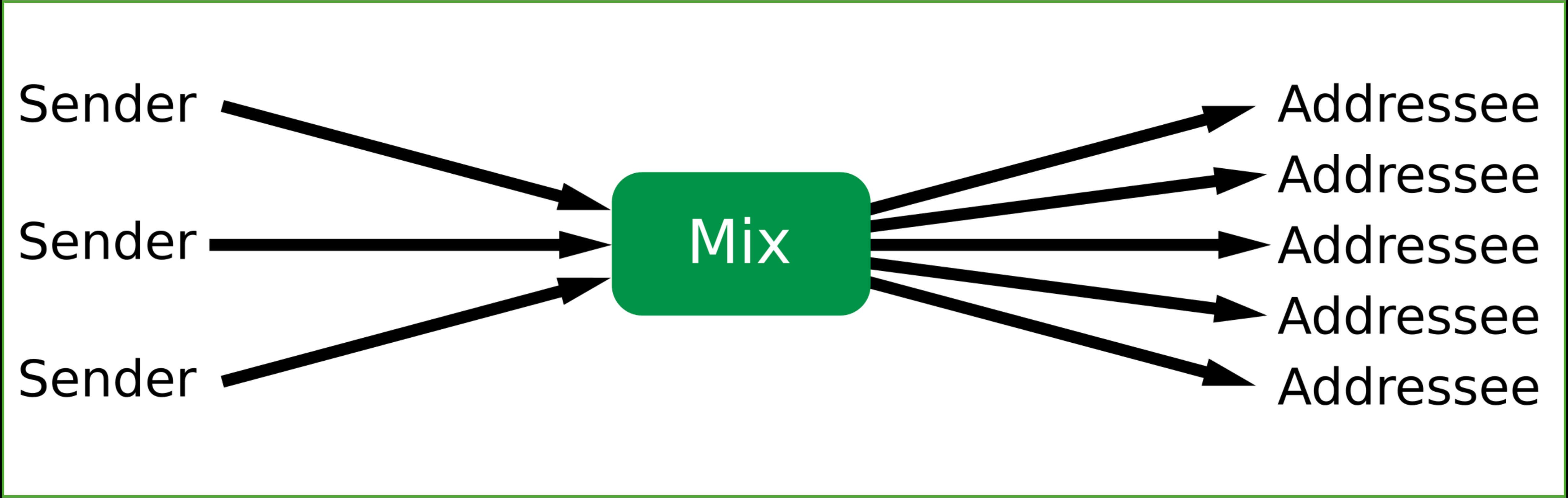
- Robust to post-processing and is composable

- Actual guarantees depend on the parameters selected and the privacy loss budget
    - This is a social question not a technical one

- **Database Reconstruction Theorem** (Dinur, Nissim 2003): Too many statistics published too accurately from a confidential database exposes the entire database with near certainty

Mix Networks and Shuffles

- Goal is to provide anonymity by decoupling inputs and outputs



Steganography

- The practice of hiding messages in other media
- Historically important method of secret communication
- Still used today, but harder to reason about the privacy guarantees compared to other privacy mechanisms
  - Generally still requires a *shared secret*
- Examples?

Next Time

How do we use these building blocks in real systems?