CSE660
Differential Privacy

October 9, 2017

Marco Gaboardi

Room: 338-B

gaboardi@buffalo.edu

http://www.buffalo.edu/~gaboardi

(ε,δ)-Differential Privacy

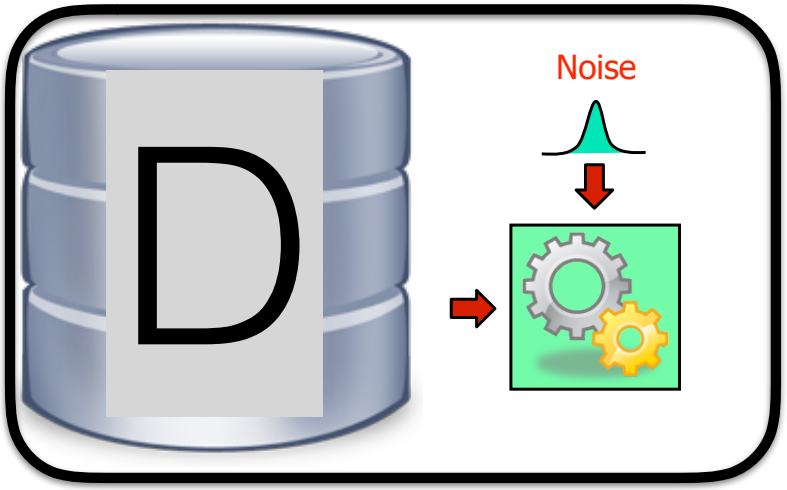
Definition

Given $\varepsilon, \delta \geq 0$, a probabilistic query $Q: X^n \rightarrow R$ is (ε,δ)-differentially private iff

for all adjacent database b_1, b_2 and for every $S \subseteq R$:

$$\Pr[Q(b_1) \in S] \leq \exp(\varepsilon)\Pr[Q(b_2) \in S] + \delta$$

Composition



M_1 is ε_1-DP

M_2 is ε_2-DP

…

M_n is ε_n-DP

The overall process is $(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_n)$-DP

Multiple queries

Question: how much perturbation do we have if we want to answer n counting queries with Laplace under ε-DP?

Multiple queries

Question: how much perturbation do we have if we want to answer n counting queries with Laplace under ε-DP?

We can split the privacy budget uniformly:

$$\epsilon = \frac{\epsilon_{\text{global}}}{n}$$

Laplace accuracy: with high probability we have:

$$\left|q(D) - r\right| \leq O\left(\frac{1}{\epsilon n}\right)$$

Multiple queries

Question: how much perturbation do we have if we want to answer n counting queries with Laplace under ε-DP?

By putting them together (hiding some details) we have as a max error

$$O\left(\frac{n}{\epsilon_{\text{global}} n}\right) = O\left(\frac{1}{\epsilon_{\text{global}}}\right)$$

Notice that if we don't renormalize this is of the order of

$$O\left(\frac{n}{\epsilon_{\text{global}}}\right)$$

bigger than the sample error.

Advanced Composition

Question: how much perturbation do we have if we want to answer n queries under (ε,δ)-DP?

We have (by hiding many details) as a max error

$$O\left(\frac{1}{\epsilon_{\text{global}}\sqrt{n}}\right)$$

If we don't renormalize this is of the order of

$$O\left(\frac{\sqrt{n}}{\epsilon_{\text{global}}}\right)$$

comparable to the sample error.

[DworkRothblumVadhan10, SteinkeUllman16]

Answering multiple queries

We have seen several methods to answer a single query:

- Randomized Response
- Laplace Mechanism
- Exponential Mechanism

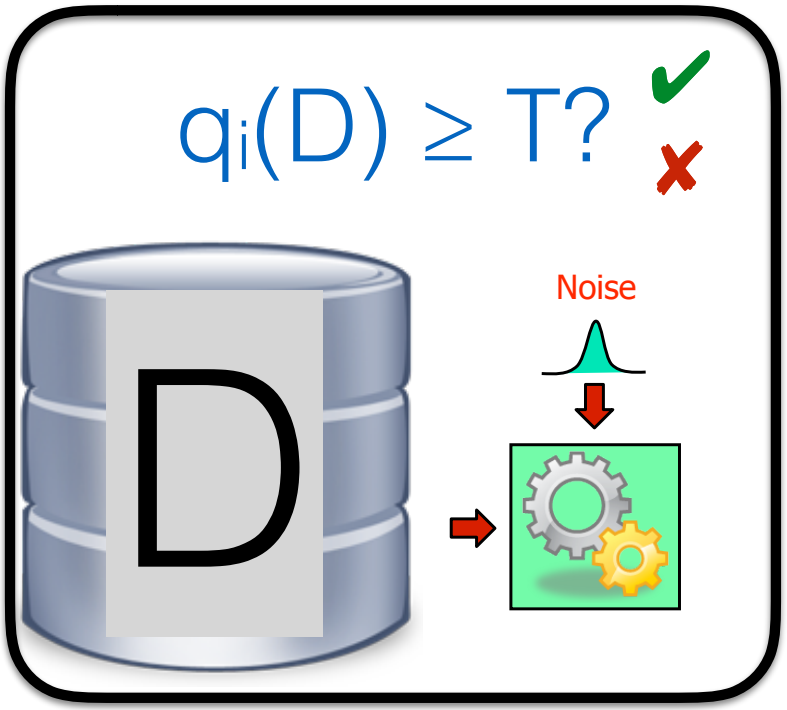
And methods to answer multiple queries with small error:

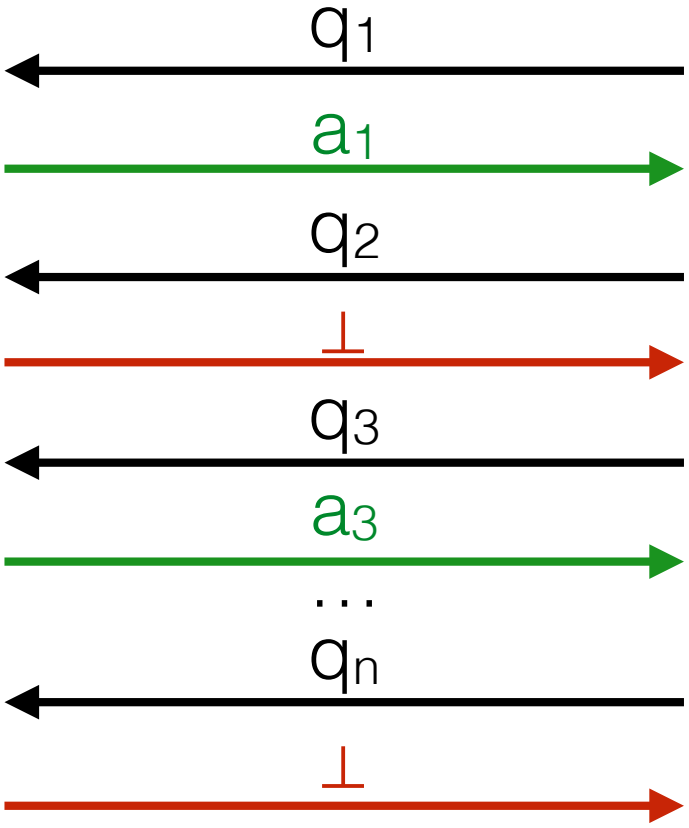
- Standard composition - we can answer $\sqrt{n}$ queries.
- Advanced composition - we can answer n queries.

Question: Can we do better?

Sparse vector

SparseVector(D,q_1,...,q_n,T,ε)





How can we achieve epsilon-DP by paying only for the queries above T?

Multiple queries

Question: Can we do better?

Answering multiple queries

When we use the basic mechanisms and then the composition theorems we are:

- adding independent noise to each query
- give a worst-case bound on the privacy loss of the composed analysis.

It turned out that we can do better if we correlate the noise of individual queries.

Answering multiple queries

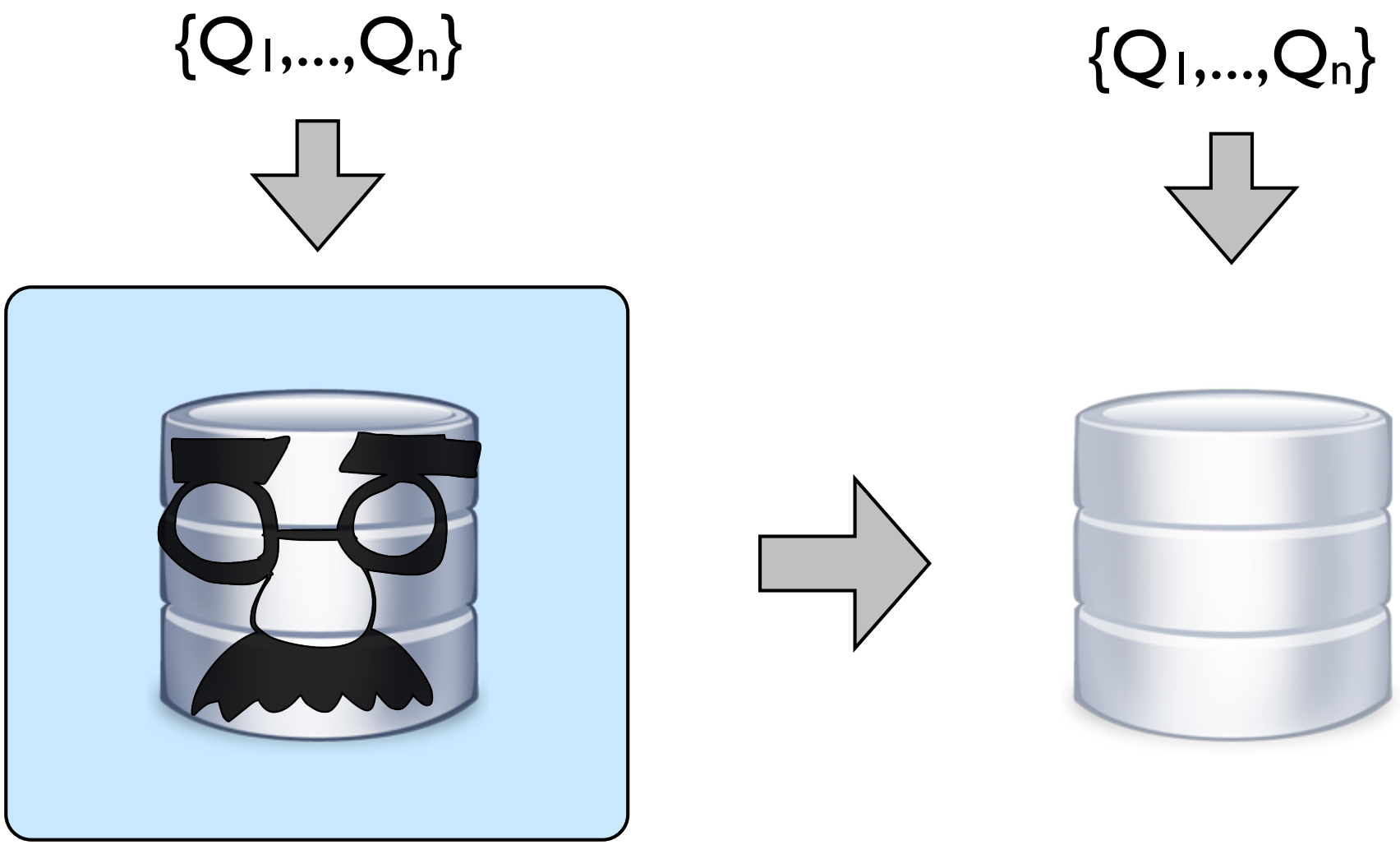
Intuitively: consider the following two queries:

How many people in the database have disease D5.

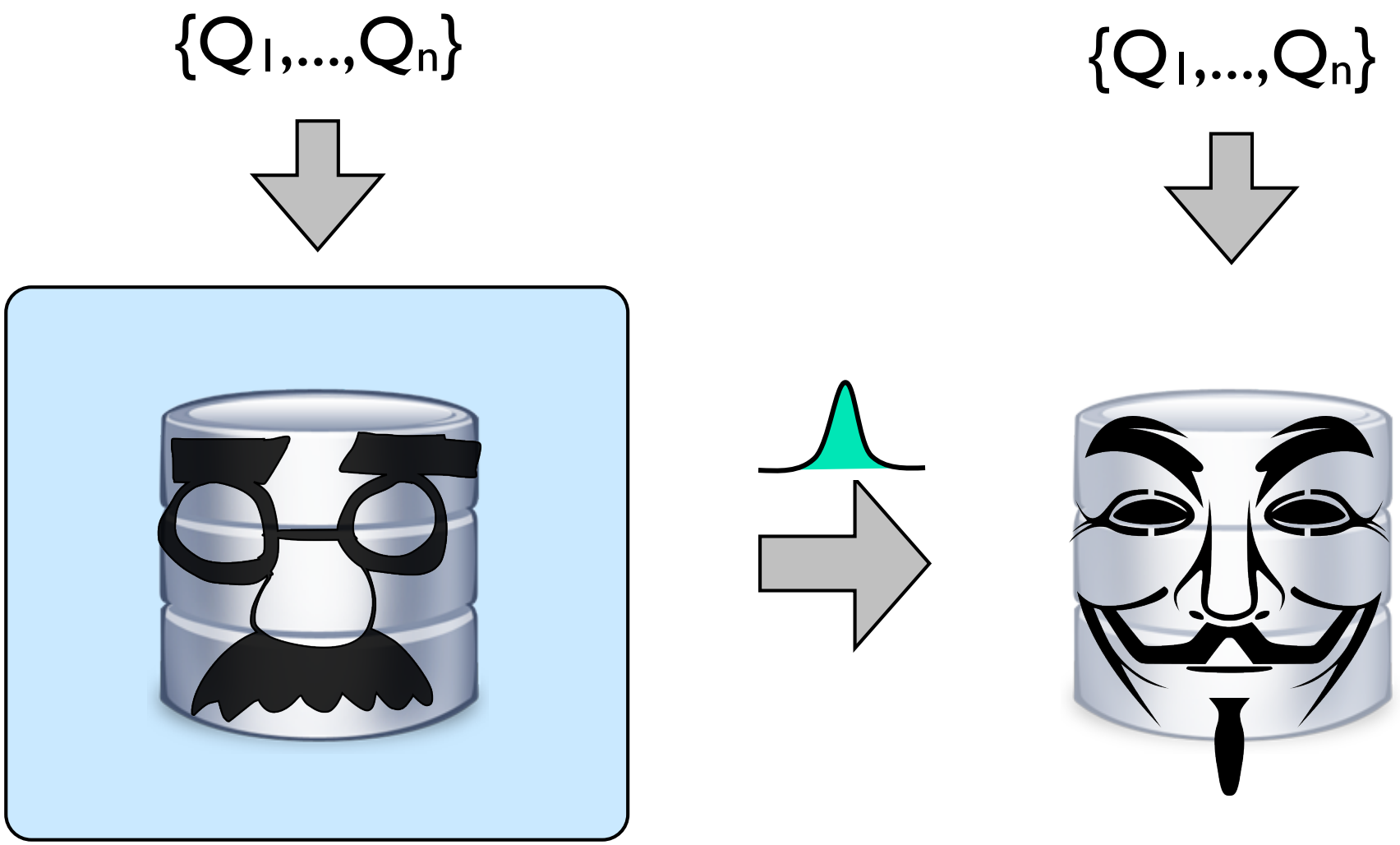
How many people, not named Marco, in the database have disease D5.

We could save in privacy budget if we give the same answer to both of them.

Data Release



Private Data Release



Linear Queries

- A linear query $q : \mathcal{X}^n \to [0, 1]$ is a function averaging the value of a function $q : \mathcal{X} \to [0, 1]$ over the elements of the dataset.
- In symbols:

$$q(D) = \frac{1}{n} \sum_{i=1}^{n} q(d_i)$$

Private Data Release

Given a database $D \in X^n$, a set of queries $Q=\{q_1,...,q_k\}$ and a target accuracy α we want to generate a differentially private synthetic database $D' \in X^m$ such that:

$$\max_{q \in Q} |q(D) - q(D')| \leq \alpha$$

SmallDB: Answering multiple linear queries

Algorithm 5 Pseudo-code for SmallDB

function SmallDB(D, Q, ϵ, α)

  Let $m = \frac{\log |Q|}{\alpha^2}$

  Let $u : \mathcal{X}^n \times \mathcal{X}^m \to \mathbb{R}$ be defined as:

$$u(D, D_i) = -\max_{q \in Q} |q(D) - q(D_i)|$$

  Let $D' \leftarrow \mathcal{M}_E(D, u, \epsilon)$

  return D'

end function

SmallDB: Answering multiple linear queries

Privacy theorem:
$\textsc{SmallDB}(D, Q, \epsilon, \alpha)$ is ε-differentially private.

Proof: Trivial, It is just an application of the exponential mechanism.

SmallDB: Answering multiple linear queries

Accuracy theorem:

$\textsc{SmallDB}(D, Q, \epsilon, \alpha)$ outputs a dataset D' such that approximately

$$\max_{q \in Q} |q(D) - q(D')| \leq \alpha$$

Sampling bound

Sampling Bound Lemma:
For every $D \in X^n$ and set of linear queries Q there exists a dataset $D' \in X^m$ with $m = \frac{\log |Q|}{\alpha^2}$ such that

$$\max_{q \in Q} |q(D) - q(D')| \leq \alpha$$

Proof sketch: Let D' be composed by m records of D sampled uniformly at random. We want to show that

$$\Pr[\max_{q \in Q} |q(D) - q(D')| > \alpha] < 1$$

where the probability is taken over the choice of D'. This show that there exists one such D' satisfying the lemma.

Sampling bound

Sampling Bound Lemma:
For every $D \in X^n$ and set of linear queries Q there exists a dataset $D' \in X^m$ with $m = \frac{\log |Q|}{\alpha^2}$ such that

$$\max_{q \in Q} |q(D) - q(D')| \leq \alpha$$

Proof sketch: For each d'_i we have:

$$\mathbb{E}[q(d'_i)] = \sum_{j=1}^{n} \frac{1}{n} q(d_j) = q(D)$$

and so

$$\mathbb{E}[q(D')] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[q(d_j)] = q(D)$$

Additive Chernoff Bound

Theorem 1.2 (Additive Chernoff Bound)**.** Let $X_1, \ldots, X_n$ be i.i.d random variables such that $0 \leq X_i \leq 1$ for every $1 \leq i \leq n$. Let $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote their mean and $E[S]$ their expected mean, where $E[S] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$ by linearity of expectation, then for every λ we have:

$$\Pr[|S - E[S]| \geq \lambda] \leq 2e^{-2\lambda^2 n}$$

Sampling bound

Sampling Bound Lemma:
For every $D \in X^n$ and set of linear queries Q there exists a dataset $D' \in X^m$ with $m = \frac{\log |Q|}{\alpha^2}$ such that

$$\max_{q \in Q} |q(D) - q(D')| \leq \alpha$$

Proof sketch: Using Chernoff bound we have:

$$\Pr[|q(D) - q(D')| > \alpha] \leq 2e^{-2m\alpha^2}$$

and by union bound for all the queries in Q we have

$$\Pr[\max_{q \in Q} |q(D) - q(D')| > \alpha] \leq 2|Q|e^{-2m\alpha^2}$$

Sampling bound

Sampling Bound Lemma:
For every $D \in X^n$ and set of linear queries Q there exists a dataset $D' \in X^m$ with $m = \frac{\log |Q|}{\alpha^2}$ such that

$$\max_{q \in Q} |q(D) - q(D')| \leq \alpha$$

Proof sketch: From

$$\Pr[\max_{q \in Q} |q(D) - q(D')| > \alpha] \leq 2|Q|e^{-2m\alpha^2}$$

substituting $m = \frac{\log |Q|}{\alpha^2}$ for |Q|>2, we can conclude:

$$\Pr[\max_{q \in Q} |q(D) - q(D')| > \alpha] < 1$$