CS 591: Formal Methods in Security and Privacy

Differential Privacy

Marco Gaboardi
gaboardi@bu.edu

Alley Stoughton
stough@bu.edu

Feedback

Please fill out the (late) mid-semester evaluation.

Recording

This is a reminder that we will record the class and we will post the link on Piazza.

This is also a reminder to myself to start recording!

From the previous classes

Releasing the mean of Some Data

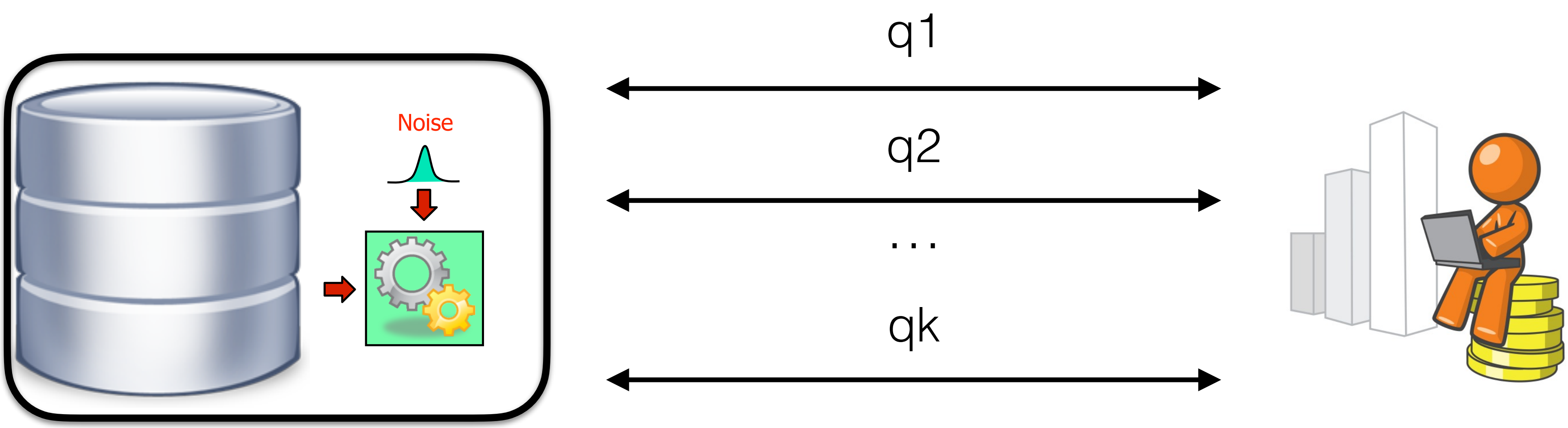
```
Mean(d : private data) : public real
   i:=0;
   s:=0;
   while (i<size(d))
      s:=s + d[i]
      i:=i+1;
   return (s/i)
```

Privacy-preserving data analysis?

We want to release some information to a data analyst and protect the privacy of the individuals contributing their data.

Privacy-preserving data analysis?

We want to release some information to a data analyst and protect the privacy of the individuals contributing their data.



Data analyst

Quantitative notions of Privacy

- The impossibility results discussed above suggest a quantitative notion of privacy,
- a notion where the privacy loss depends on the number of queries that are allowed,
- and on the accuracy with which we answer them.

Privacy-preserving data analysis?

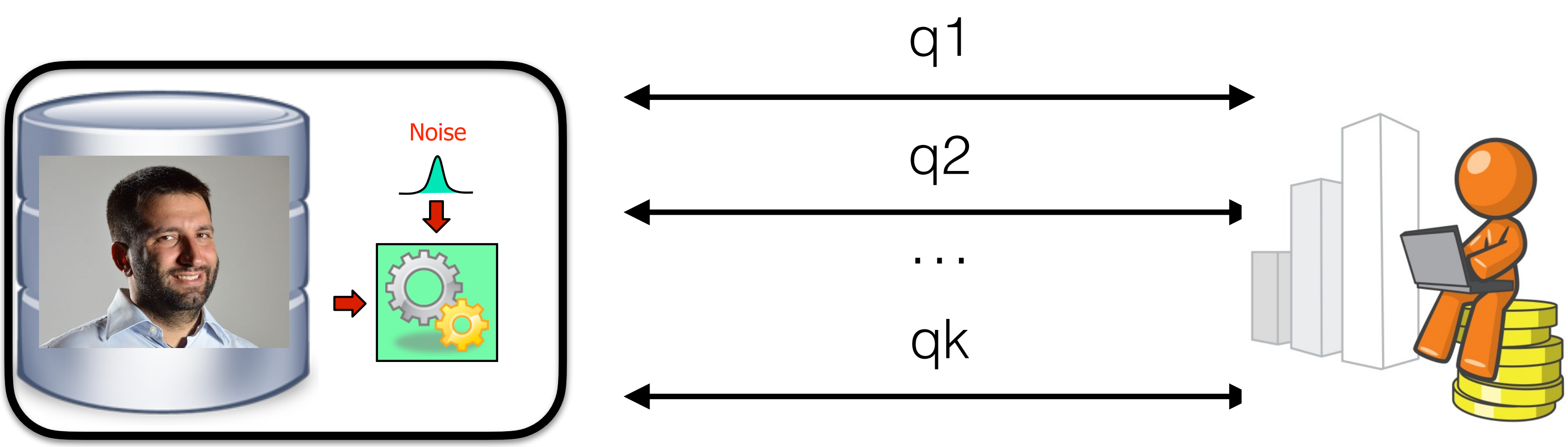
- The analyst learn almost the same about me after the analysis as what she would have learnt if I didn't contribute my data.

Privacy-preserving data analysis?

- The analyst learn almost the same about me after the analysis as what she would have learnt if I didn't contribute my data.

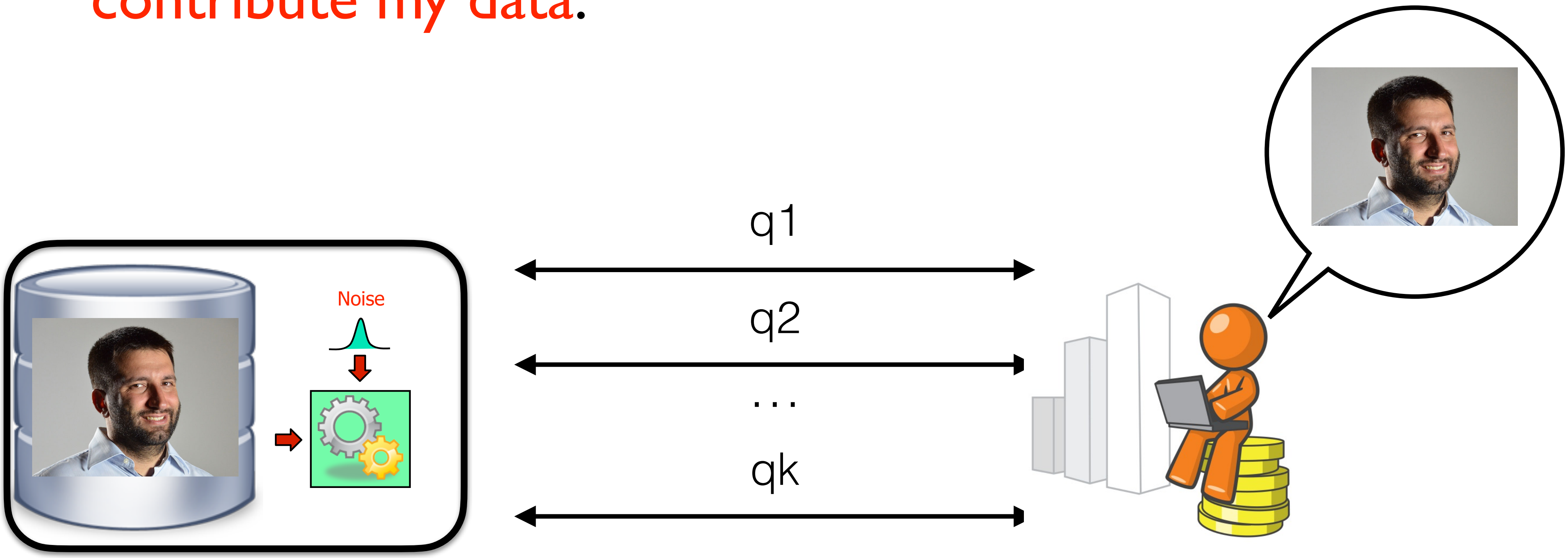
Privacy-preserving data analysis?

- The analyst learn almost the same about me after the analysis as what she would have learnt if I didn't contribute my data.



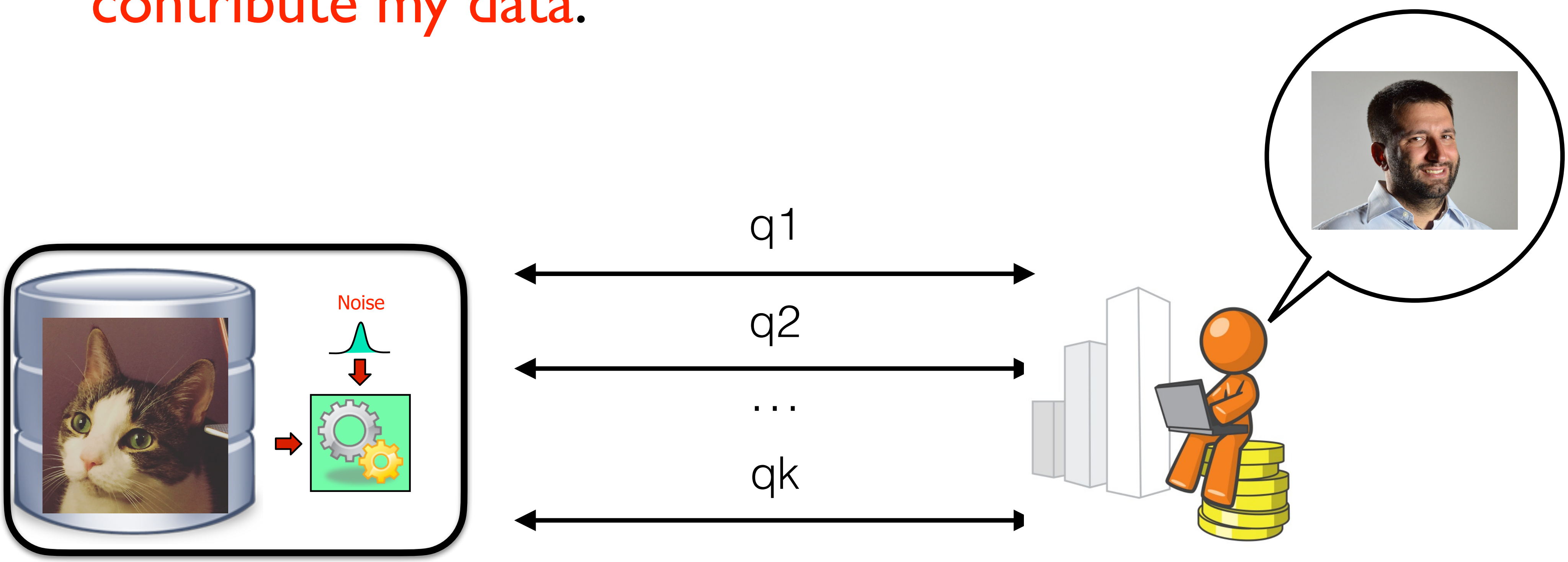
Privacy-preserving data analysis?

- The analyst learn almost the same about me after the analysis as what she would have learnt if I didn't contribute my data.



Privacy-preserving data analysis?

- The analyst learn almost the same about me after the analysis as what she would have learnt if I didn't contribute my data.



Adjacent databases

- We can formalize the concept of contributing my data or not in terms of a notion of distance between datasets.

- Given two datasets D, D'$\in$DB, their distance is defined as:

$$D\Delta D' = |\{k \leq n \mid D(k) \neq D'(k)\}|$$

- We will call two datasets adjacent when $D\Delta D'=1$ and we will write D~D'.

Privacy Loss

In general we can think about the following quantity as the privacy loss incurred by observing r on the databases D and D'.

$$L_{D,D'}(r) = \log \frac{\Pr[Q(D)=r]}{\Pr[Q(D')=r]}$$

(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 D, D and for every $S \subseteq R$:

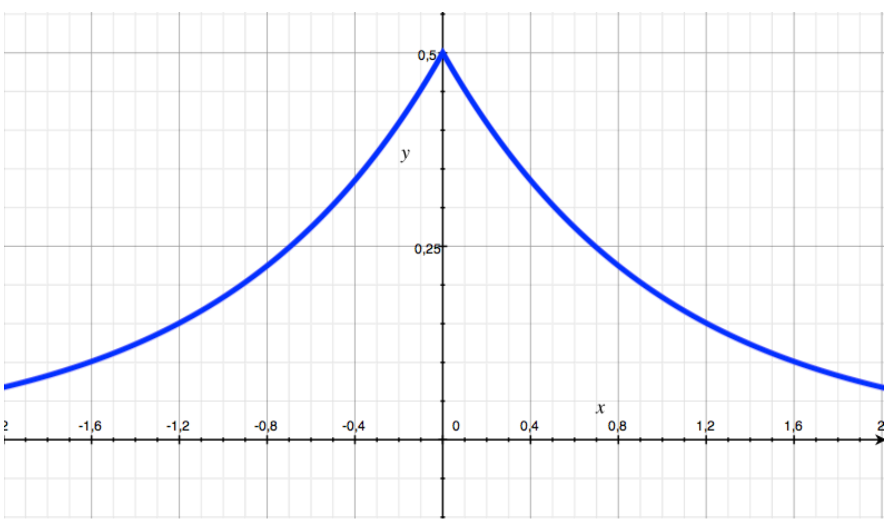
$$\Pr[Q(D) \in S] \leq \exp(\varepsilon)\Pr[Q(D') \in S] + \delta$$

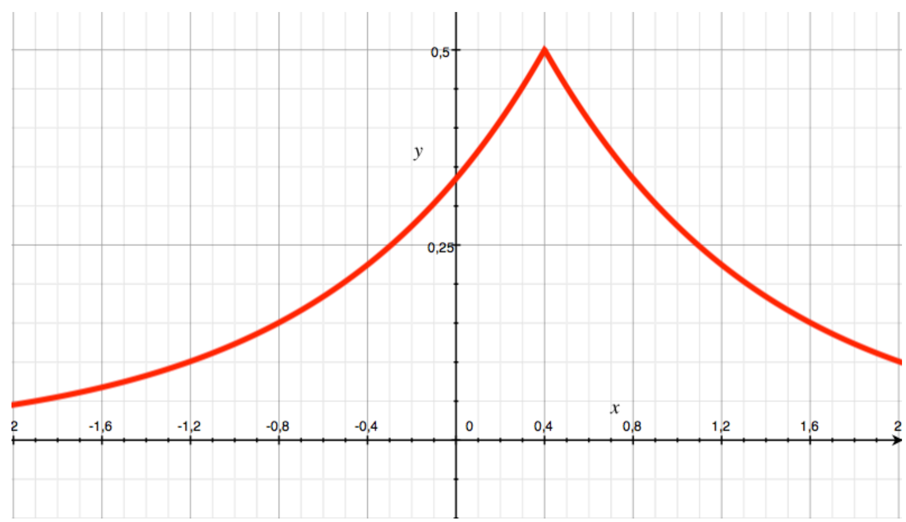
Differential Privacy

Q : db => R   probabilistic

$Q(D_I \cup \{x\})$

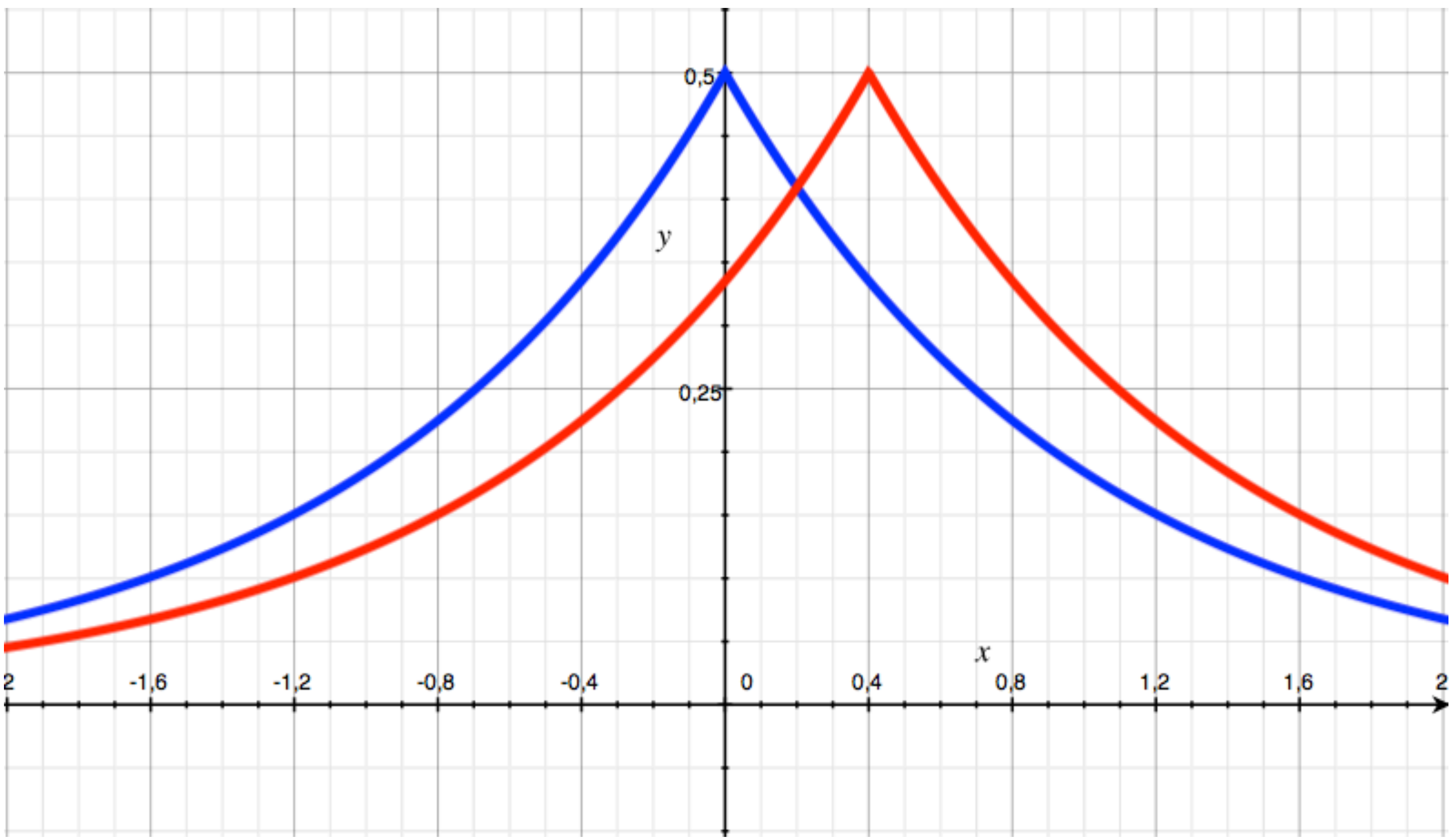
$Q(D_I \cup \{y\})$





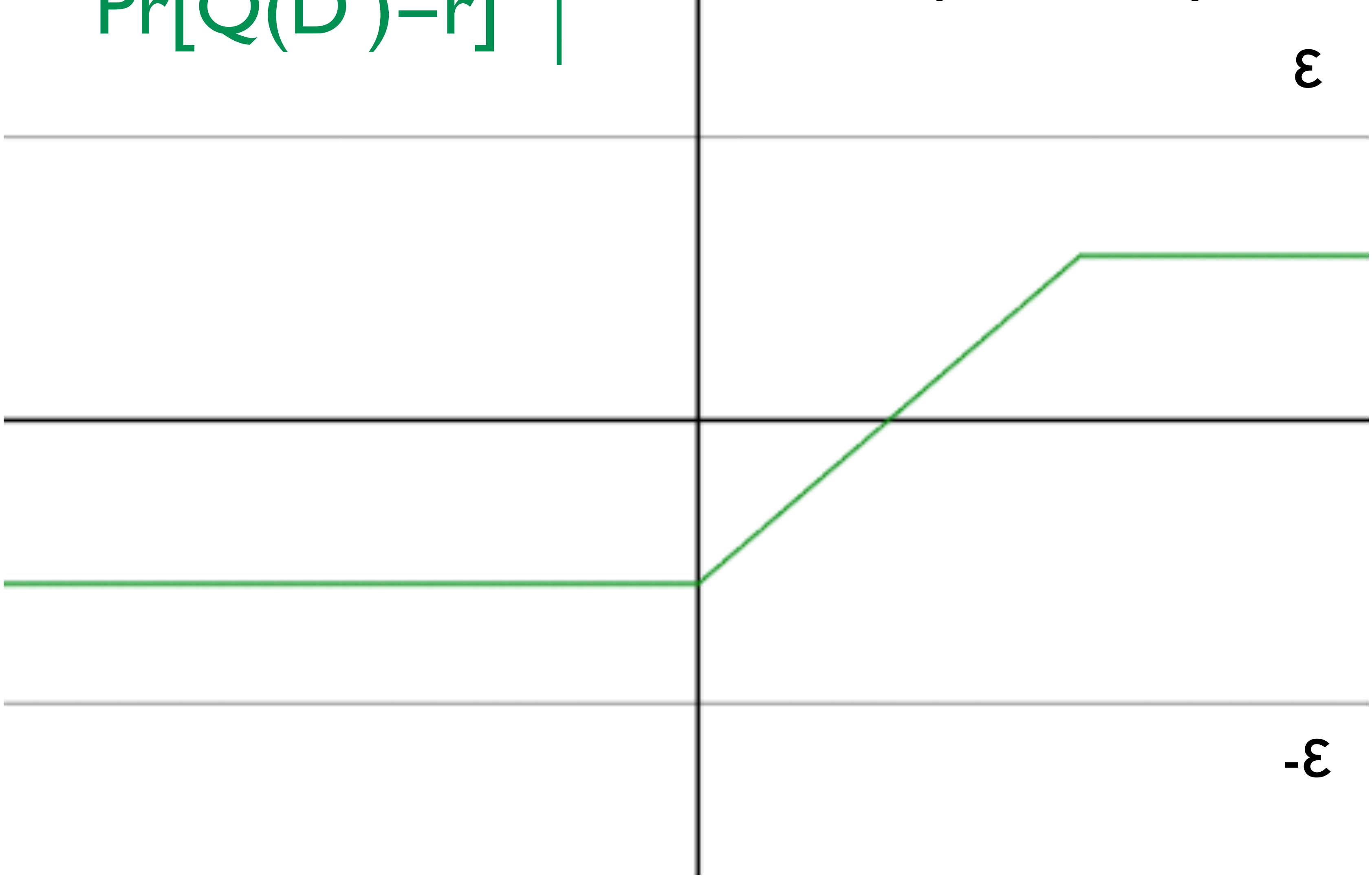
Differential Privacy

$d(Q(D_I \cup \{x\}), Q(D_I \cup \{y\})) \leq \varepsilon$ with probability $1\text{-}\delta$



(ε,δ)-Differential Privacy

$$\left| \log \frac{\Pr[Q(D)=r]}{\Pr[Q(D')=r]} \right| \leq \varepsilon$$ with probability $1-\delta$



Today: Achieving Differential Privacy

(ε,δ)-indistinguishability

When we defined statistical distance:

$$\Delta(\mu_1,\mu_2)=\max_{E\subseteq A} | \mu_1(E)-\mu_2(E) | = \delta$$

we also used a notion of δ-indistinguishability.

We say that two distributions μ_1, $\mu_2 \in D(A)$, are at δ-indistinguishable if:

$$\Delta(\mu1,\mu2)\leq \delta$$

(ε,δ)-indistinguishability

We can define a ε-skewed version of statistical distance. We call this notion ε-distance.

$$\Delta_\varepsilon(\mu1,\mu2)=\sup_{E \subseteq A} \max(\mu_1(E)-e^{\varepsilon}\mu_2(E),\ \mu_2(E)-e^{\varepsilon}\mu_1(E),0)$$

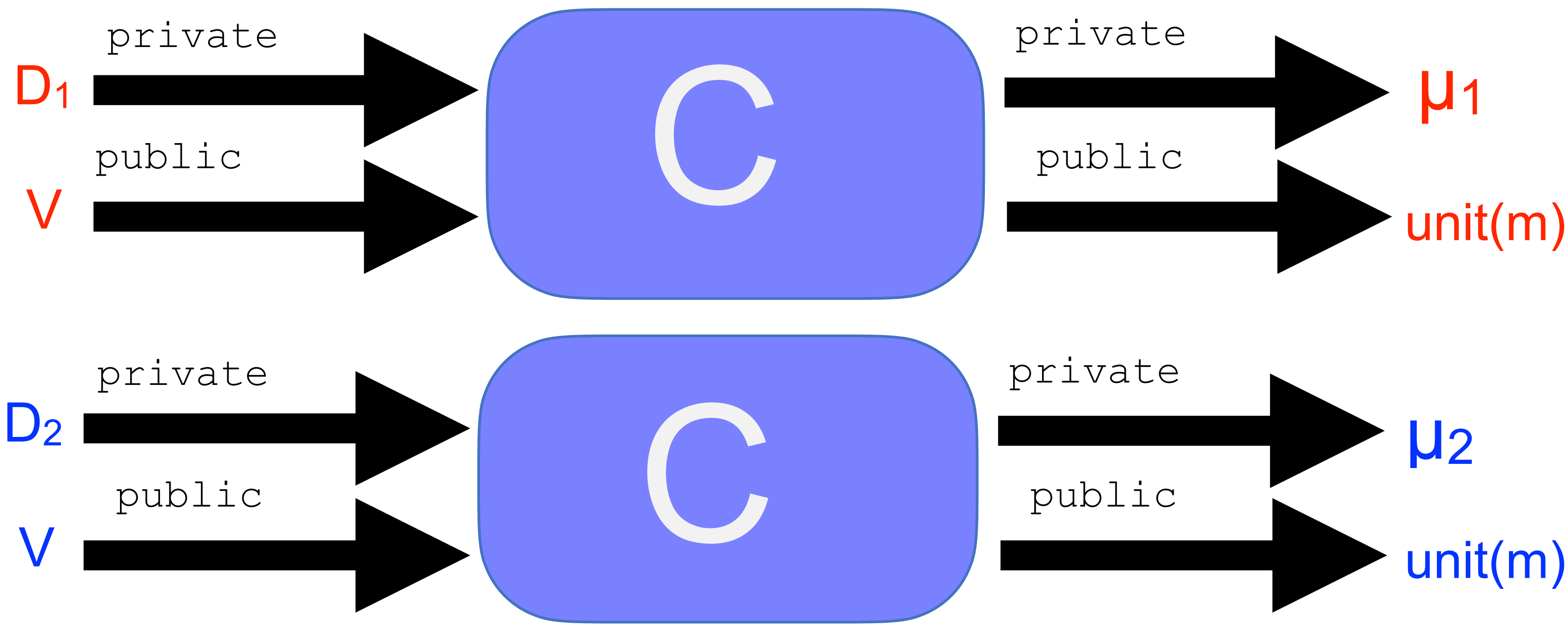
We say that two distributions $\mu_1, \mu_2 \in D(A)$, are at (ε,δ)-indistinguishable if:

$$\Delta_\varepsilon(\mu1,\mu2) \leq \delta$$

Differential Privacy as a Relational Property

c is differentially private if and only if for every $m_1 \sim m_2$ (extending the notion of adjacency to memories):

$\{c\}_{m1}=\mu_1$ and $\{c\}_{m2}=\mu_2$ implies $\Delta_\varepsilon(\mu_1,\mu_2) \leq \delta$



Releasing the mean of Some Data

```
Mean(d : private data) : public real
   i:=0;
   s:=0;
   while (i<size(d))
      s:=s + d[i]
      i:=i+1;
   return (s/i)
```

Adding Noise

Question: What is a good way to add noise to the output of a statistical query to achieve (ε,0)-DP?

Adding Noise

Question: What is a good way to add noise to the output of a statistical query to achieve $(\varepsilon,0)$-DP?

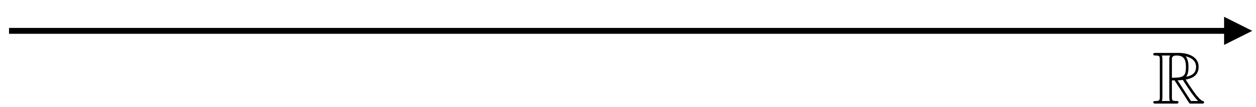
Intuitive answer: it should depend on ε or the accuracy we want to achieve, and on the scale that a change of an individual can have on the output.

Global Sensitivity

$$GS_q = \max\{\, |q(D) - q(D')| \text{ s.t. } D \sim D' \}$$

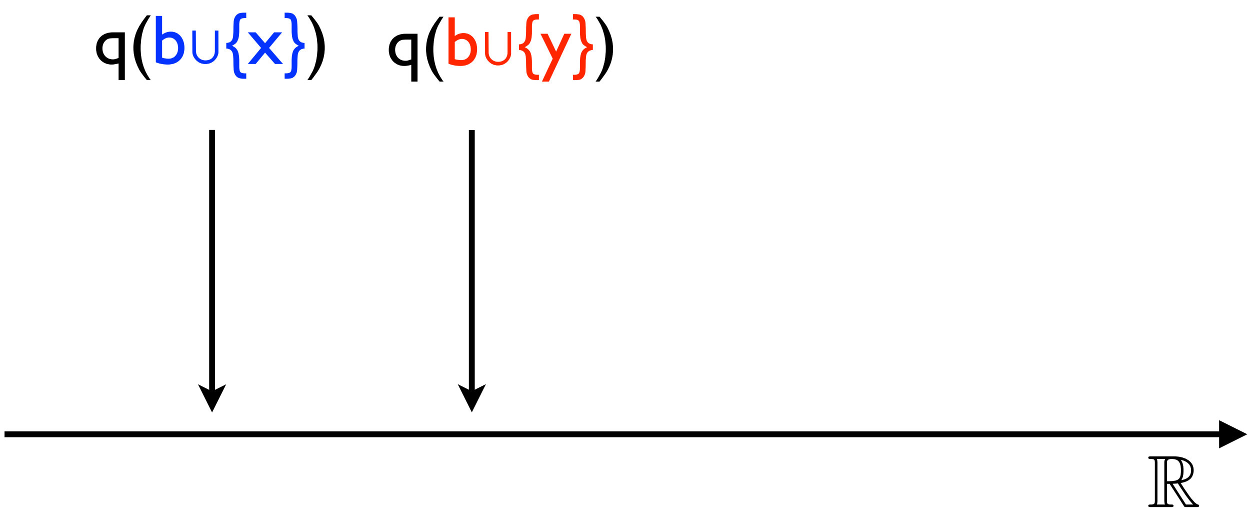
Global Sensitivity

$$GS_q = \max\{\, |\, q(D) - q(D')\,| \text{ s.t. } D \sim D' \}$$



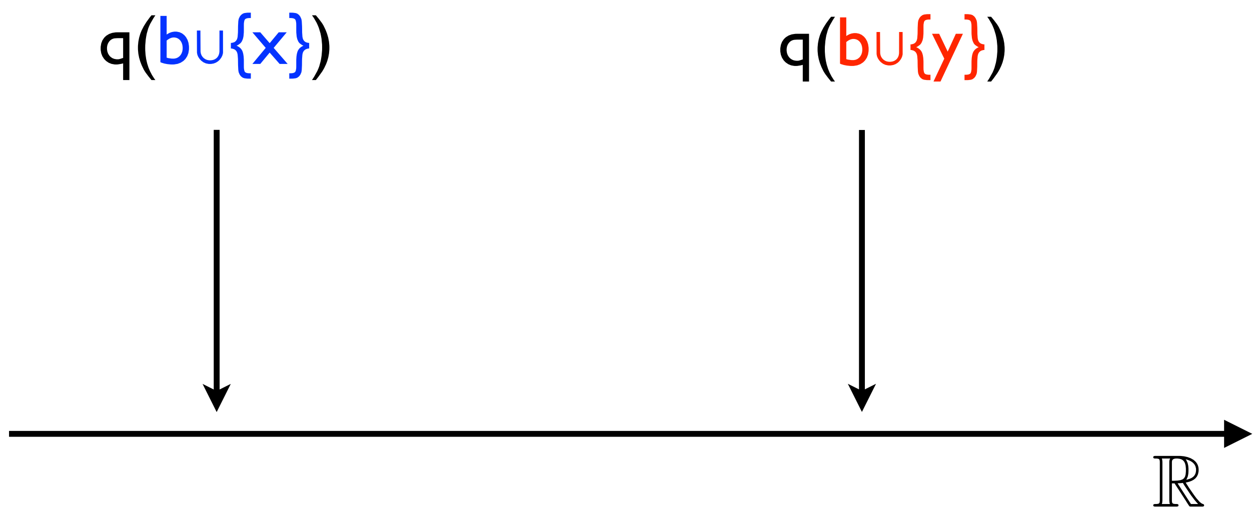
Global Sensitivity

$$GS_q = \max\{\,|q(D) - q(D')|\ \text{s.t.}\ D \sim D'\}$$



Global Sensitivity

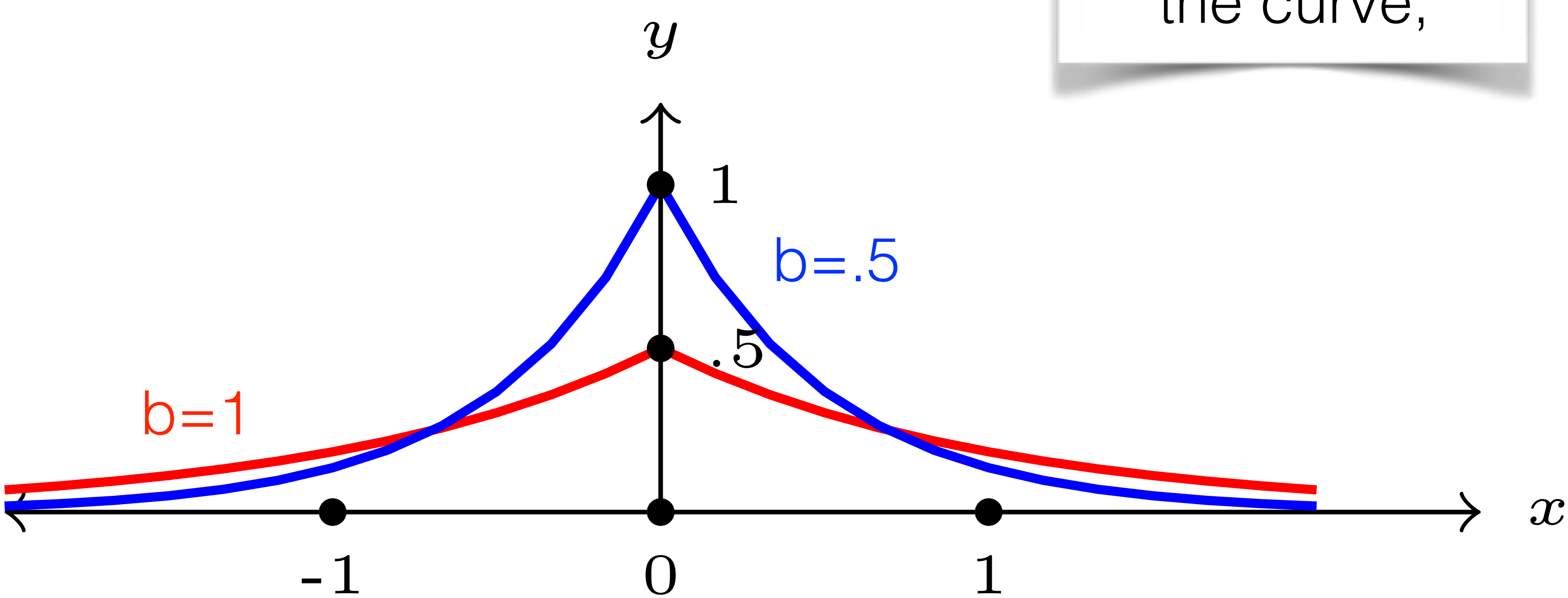
$$GS_q = \max\{\,|q(D) - q(D')|\ \text{s.t. } D \sim D'\}$$



Laplace Distribution

$$\mathsf{Lap}(b,\mu)(X) = \frac{1}{2b}\exp\Big(-\frac{|\mu - X|}{b}\Big)$$

b regulates the skewness of the curve,



Releasing privately the mean of Some Data

```
Mean(d : private data) : public real
   i:=0;
   s:=0;
   while (i<size(d))
      s:=s + d[i]
      i:=i+1;
   z:=$ Laplace(sens/eps,0)
   z:= (s/i)+z
   return z
```

Laplace Mechanism

```
Lap(d : priv data)(f: data -> real)
    (e:real) : pub real
   z:=$ Laplace(GSf/e,0)
   z:= f(d)+z
   return z
```

Laplace Mechanism

```
Lap(d : priv data)(f: data -> real)
    (e:real) : pub real
   z:=$ Laplace(GSf/e,0)
   z:= f(d)+z
   return z
```

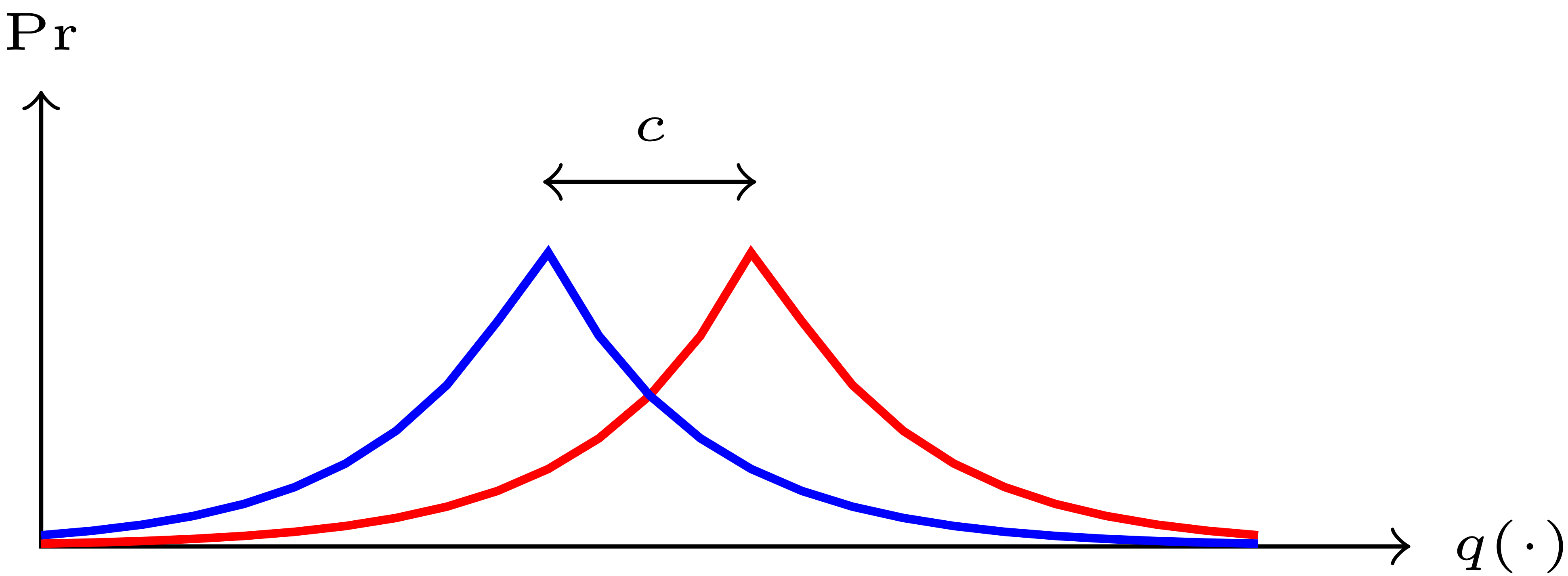
It turns out that we could also write it as:

```
Lap(d : priv data)(f: data -> real)
    (e:real) : pub real
   z:=$ Laplace(GSf/e,f(d))
   return z
```

Laplace Mechanism

Theorem (Privacy of the Laplace Mechanism)
The Laplace mechanism is (ε,0)-differentially private.

Proof: Intuitively



Laplace Mechanism

Theorem (Privacy of the Laplace Mechanism)
The Laplace mechanism is $(\varepsilon,0)$-differentially private.

Laplace Mechanism

Theorem (Privacy of the Laplace Mechanism)
The Laplace mechanism is $(\varepsilon,0)$-differentially private.

Laplace Mechanism

Question: How accurate is the answer that we get from the Laplace Mechanism?

Properties of Differential Privacy

Some important properties

- Resilience to post-processing
- Group privacy
- Composition

Some important properties

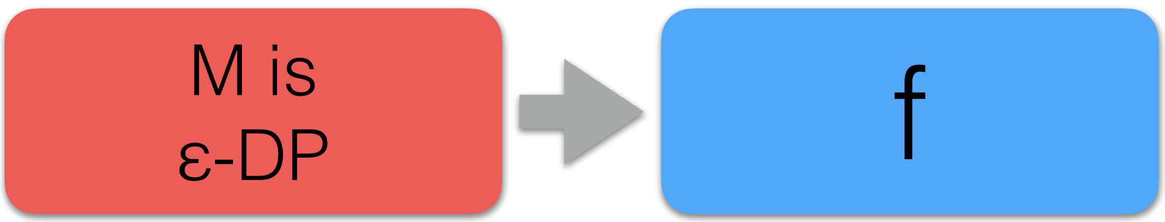
- Resilience to post-processing
- Group privacy
- Composition

We will look at them in the context of $(\varepsilon,0)$-differential privacy.

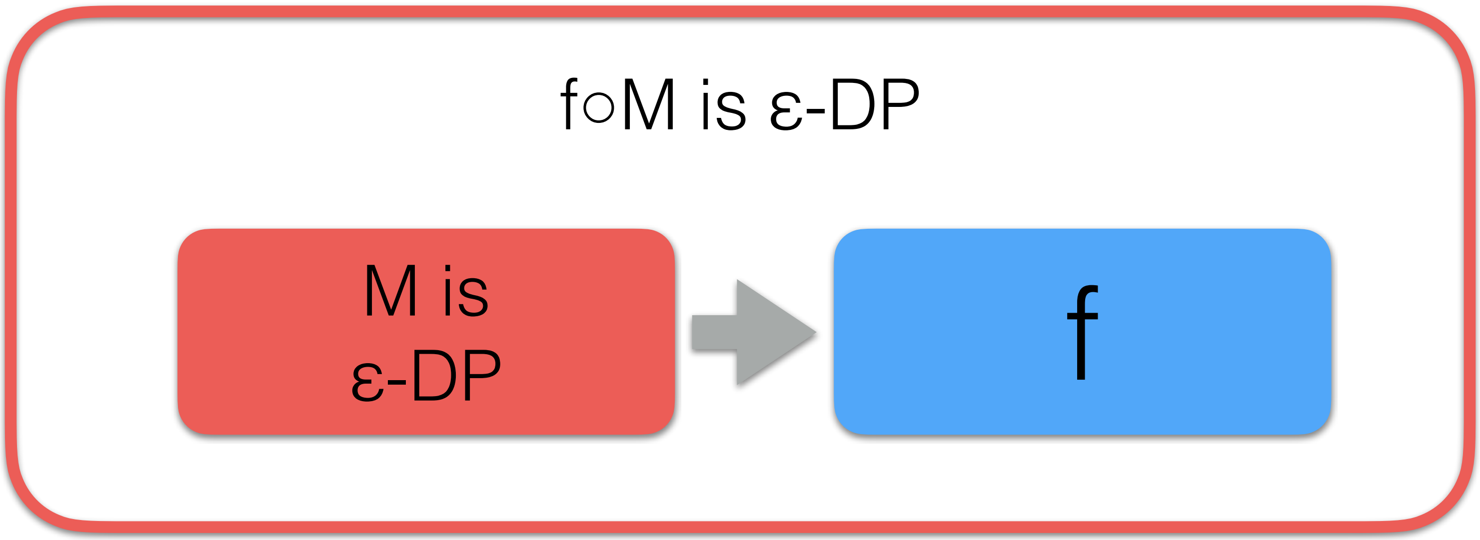
Resilience to Post-processing

M is
ε-DP

Resilience to Post-processing



Resilience to Post-processing



Resilience to Post-processing

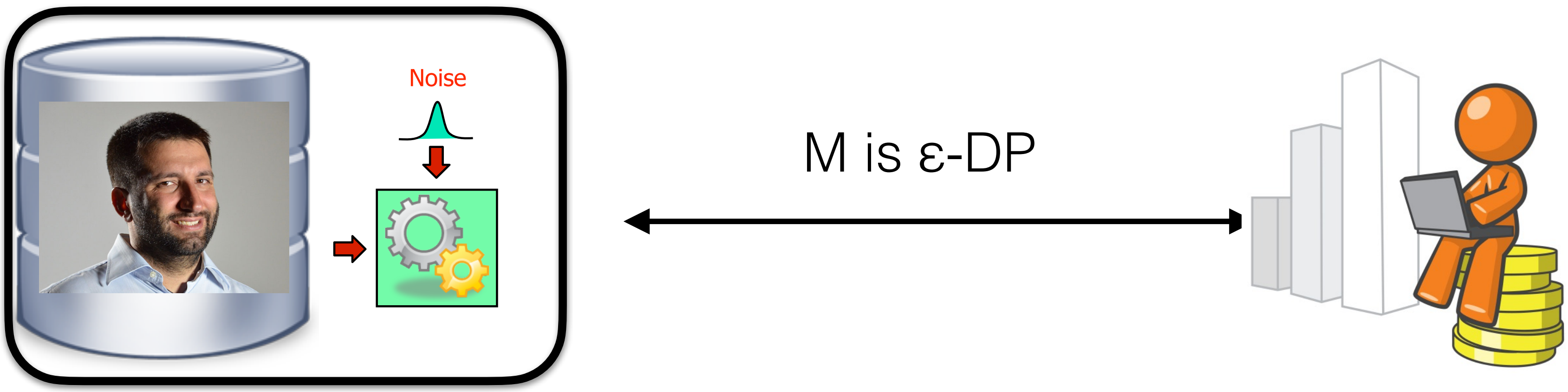
Question: Why is resilience to post-processing important?

Resilience to Post-processing

Question: Why is resilience to post-processing important?

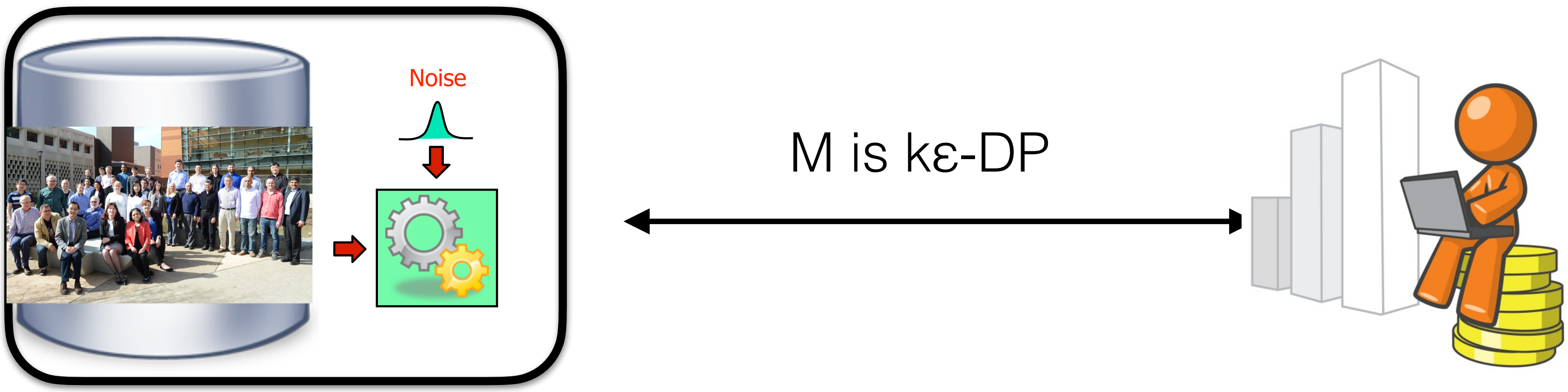
Answer: Because it is what allows us to publicly release the result of a differentially private analysis!

Group Privacy



$$\Pr[\mathcal{M}(D) = r] \leq e^{\epsilon} \Pr[\mathcal{M}(D') = r]$$

Group Privacy



$$\Pr[\mathcal{M}(D) \in S] \leq \exp(k\epsilon) \Pr[\mathcal{M}(D') \in S]$$

Group Privacy

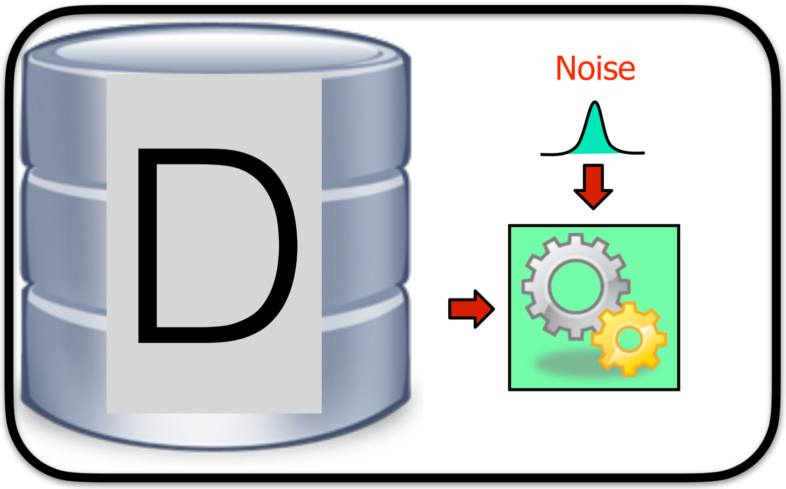
Question: Why is group privacy important?

Group Privacy

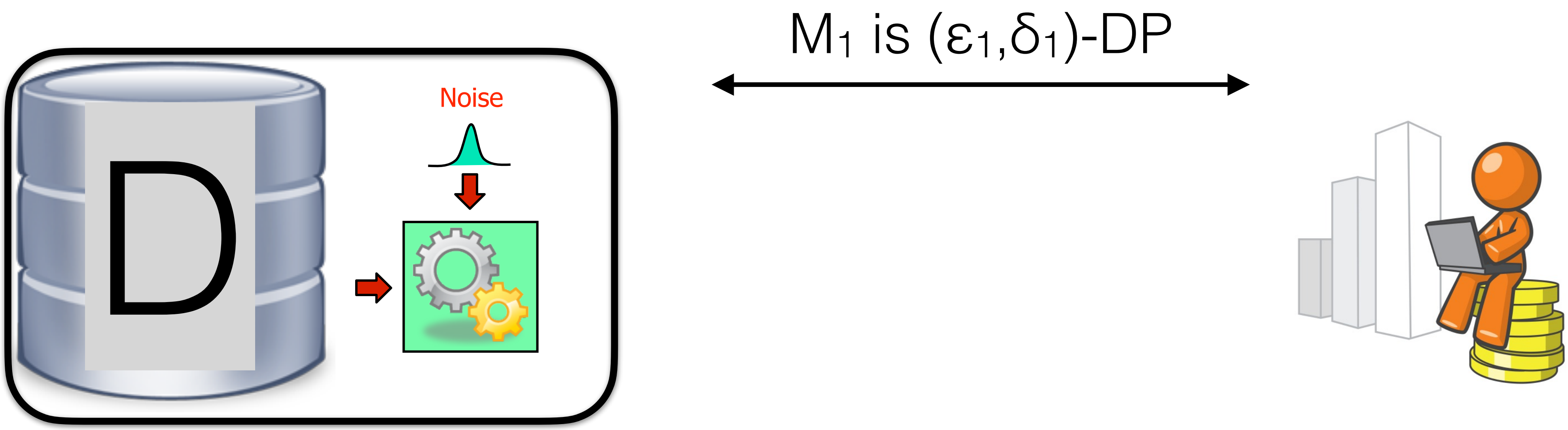
Question: Why is group privacy important?

Answer: Because it allows to reason about privacy at different level of granularities!

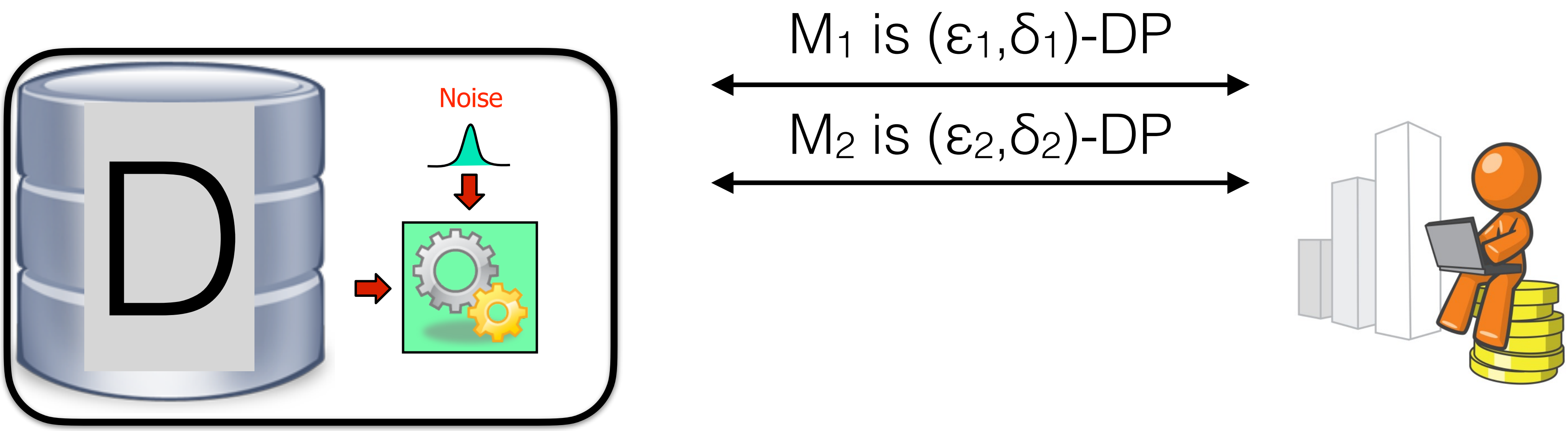
Composition



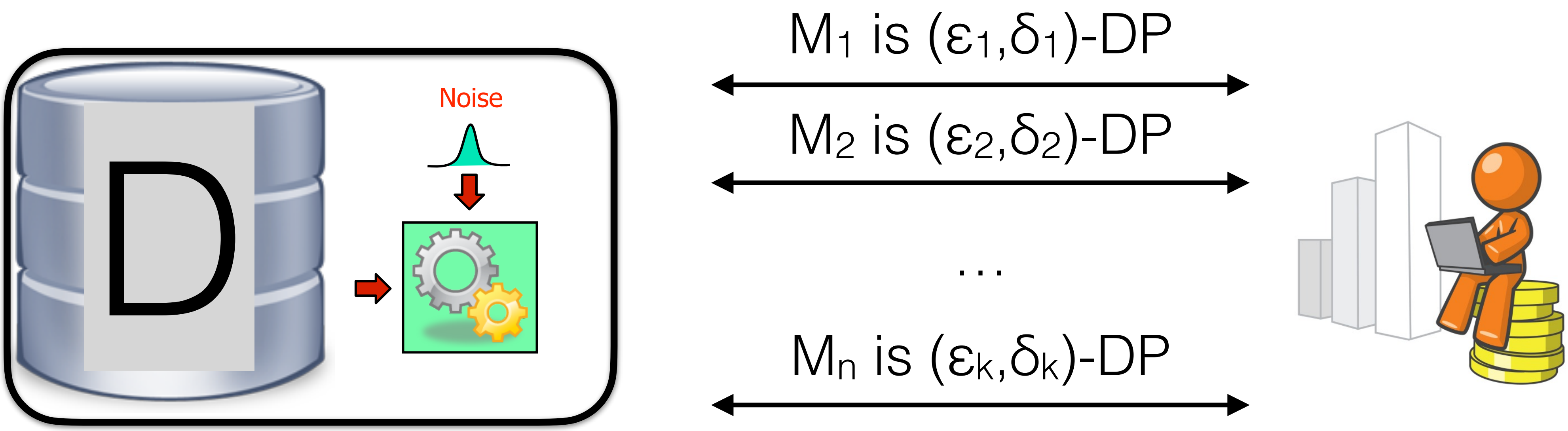
Composition



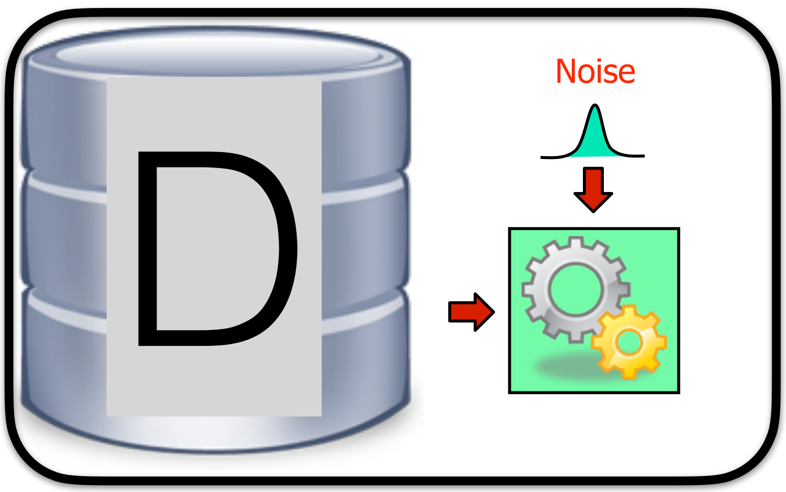
Composition



Composition



Composition



M_1 is (ε_1,δ_1)-DP

M_2 is (ε_2,δ_2)-DP

…

M_n is (ε_k,δ_k)-DP

The overall process is $(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_k,\delta_1+\delta_2+\ldots+\delta_k)$-DP

Composition

Let M_1:DB $\rightarrow R_1$ be a (ε_1,δ_1)-differentially private program and M_2:DB $\rightarrow R_2$ be a (ε_2,δ_1)-differentially private program. Then, their composition $M_{1,2}$:DB$\rightarrow R_1 x R_2$ defined as

$$M_{1,2}(D)=(M_1(D),M_2(D))$$

is $(\varepsilon_1+\varepsilon_2,\delta_1+\delta_2)$-differentially private.

Composition

Question: Why composition is important?

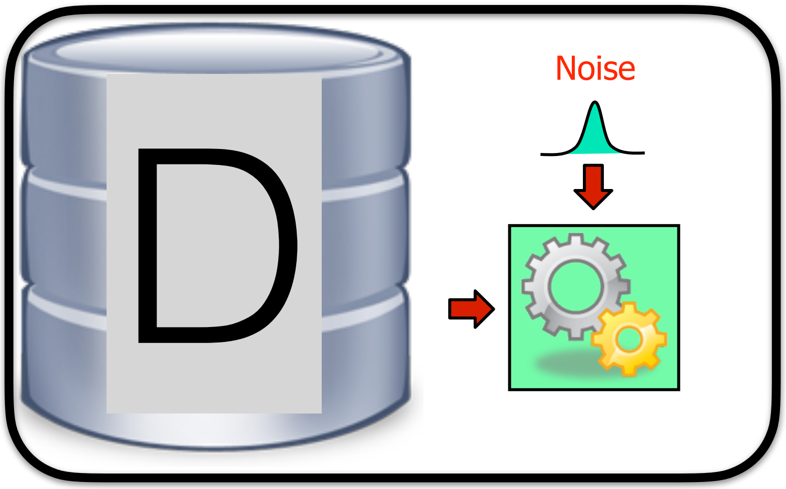
Composition

Question: Why composition is important?

Answer: Because it allows to reason about privacy as a budget!

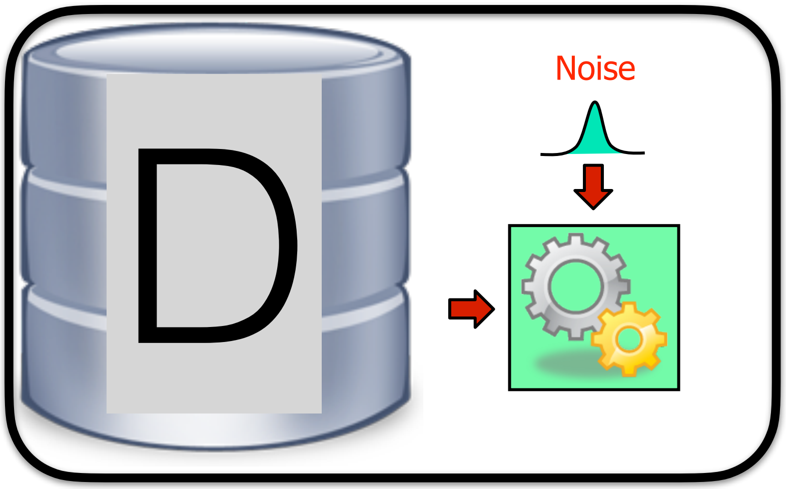
Composition

Budget=ε_{global}



Composition

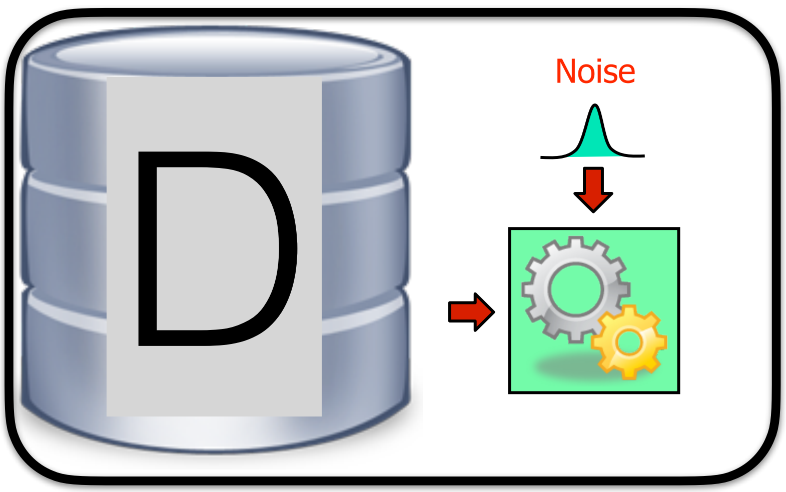
Budget=ε_{global}



M_1 is ε_1-DP

Composition

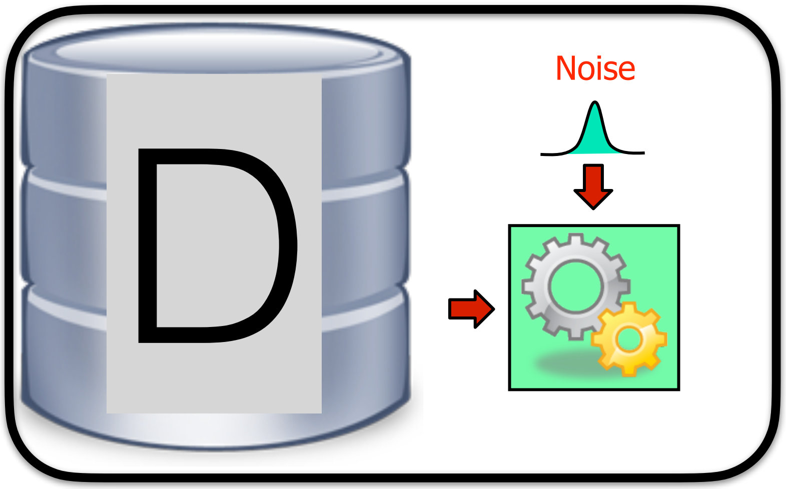
Budget=ε_{global} - ε_1



M_1 is ε_1-DP

Composition

Budget=ε_{global} - ε_1

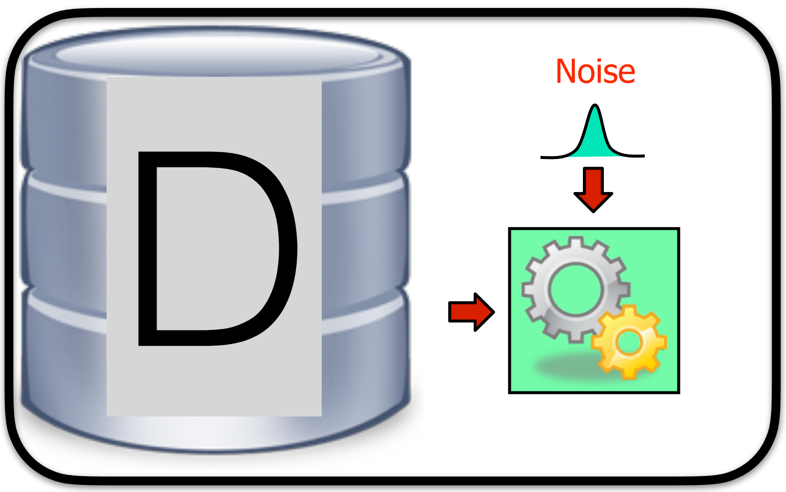


M_1 is ε_1-DP

M_2 is ε_2-DP

Composition

Budget=ε_{global} - ε_1 - ε_2

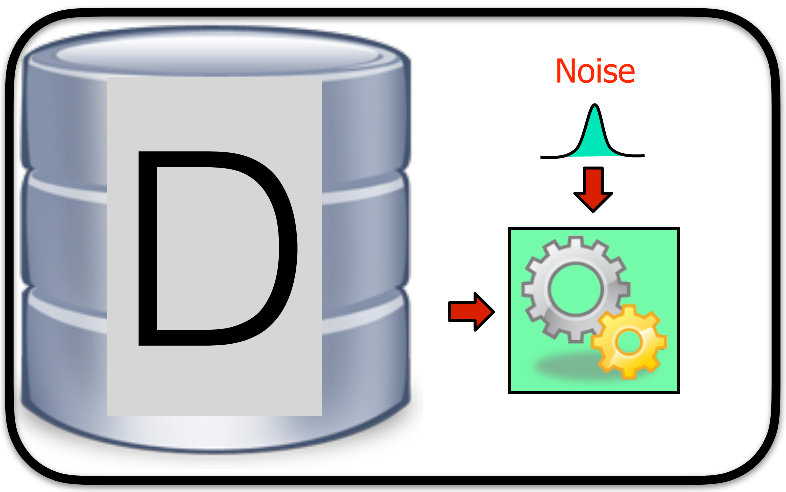


M_1 is ε_1-DP

M_2 is ε_2-DP

Composition

Budget=ε_{global} - ε_1 - ε_2 ...



M_1 is ε_1-DP

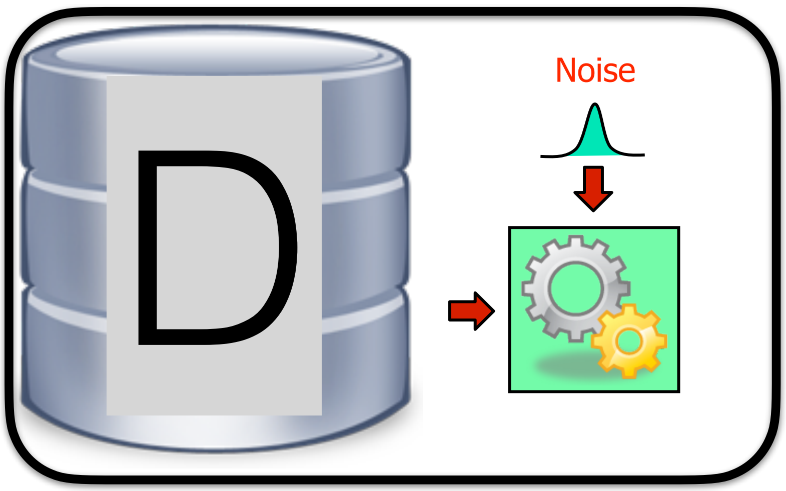
M_2 is ε_2-DP

...

M_n is ε_n-DP

Composition

Budget=ε_{global} - ε_1 - ε_2 … - ε_n



M_1 is ε_1-DP

M_2 is ε_2-DP

…

M_n is ε_n-DP

CDF

Budget= ε_{global} - ε_1 - ε_2 - ε_3 - ε_4 - ε_5 - ε_6 - ε_7 - ε_8

$X=\{0,1\}^3$ ordered wrt binary encoding.

$D \in X^{10}$ =

	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1

$q^*_{000}(D) = .3+L(1/\varepsilon_1)$

$q^*_{001}(D) = .4+L(1/\varepsilon_2)$

$q^*_{010}(D) = .6+L(1/\varepsilon_3)$

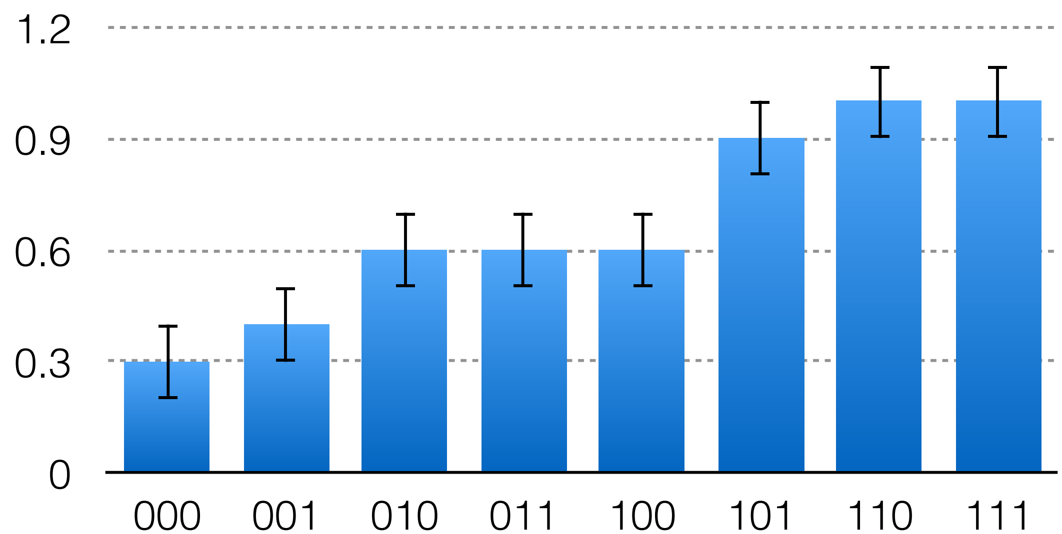
$q^*_{011}(D) = .6+L(1/\varepsilon_4)$

$q^*_{100}(D) = .6+L(1/\varepsilon_5)$

$q^*_{101}(D) = .9+L(1/\varepsilon_6)$

$q^*_{110}(D) = 1+L(1/\varepsilon_7)$

$q^*_{111}(D) = 1+L(1/\varepsilon_8)$



Marginals

Budget=ε_{global} - ε_1 - ε_2 - ε_3

$D \in X^{10} =$

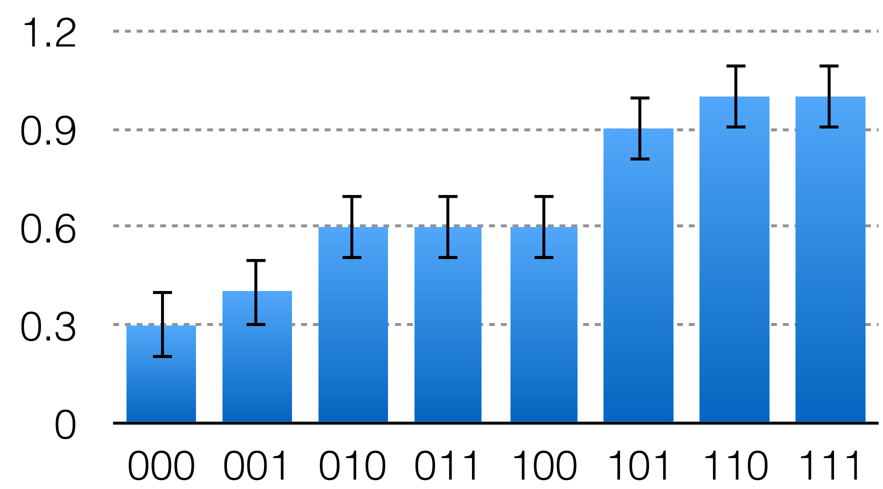
	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1
margin	$.4+Y_1$	$.3+Y_2$	$.4+Y_3$

$q^*_1(D) = .4+L(1/(10^*\varepsilon_1))$

$q^*_2(D) = .3+L(1/(10^*\varepsilon_2))$

$q^*_3(D) = .4+L(1/(10^*\varepsilon_3))$

$$\text{Budget}=\varepsilon_{\text{global}} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3 - \varepsilon_4 - \varepsilon_5 - \varepsilon_6 - \varepsilon_7 - \varepsilon_8$$



$$\text{Budget}=\varepsilon_{\text{global}} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3$$

	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1
margin	$.4+Y_1$	$.3+Y_2$	$.4+Y_3$

Privacy Budget vs Epsilon

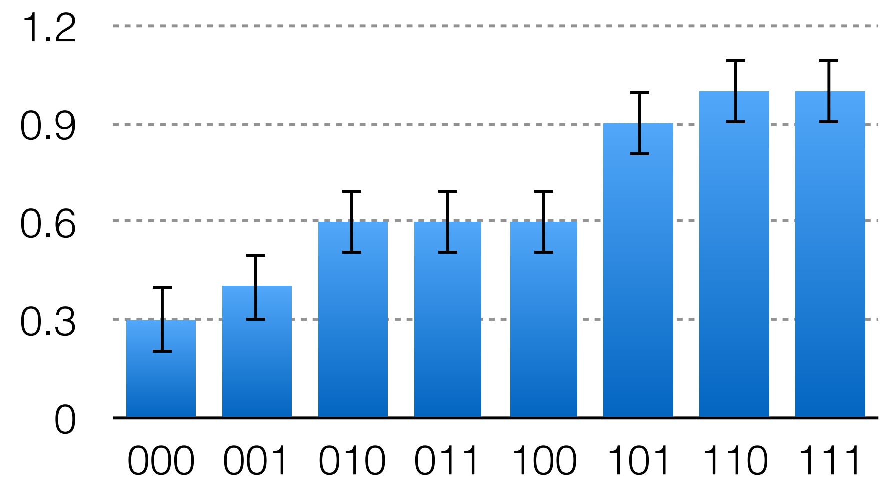
Sometimes is more convenient to think in terms of Privacy Budget: Budget=$\varepsilon_{global} - \sum \varepsilon_{local}$

Sometimes is more convenient to think in terms of epsilon: $\varepsilon_{global} = \sum \varepsilon_{local}$

Also making them uniforms is sometimes more informative.

$$\text{Budget} = \varepsilon_{global} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3 - \varepsilon_4 - \varepsilon_5 - \varepsilon_6 - \varepsilon_7 - \varepsilon_8$$

$$\varepsilon_{global} = \varepsilon+\varepsilon+\varepsilon+\varepsilon+\varepsilon+\varepsilon+\varepsilon+\varepsilon=8\varepsilon$$



$$\text{Budget} = \varepsilon_{global} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3$$

$$\varepsilon_{global} = \varepsilon+\varepsilon+\varepsilon=3\varepsilon$$

	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1
margin	$.4+Y_1$	$.3+Y_2$	$.4+Y_3$

Releasing partial sums

```
DummySum(d : {0,1} list) : real list
   i:= 0;
   s:= 0;
   r:= [];
   t:= 0;
   while (i<size d)
       s:= s + d[i]
       z:=$ Laplace(1/eps,0)
       t:= s+z;
       r:= r ++ [t];
       i:= i+1;
   return r
```

Releasing partial sums

```
DummySum(d : {0,1} list) : real list
   i:=0;
   s:=0;
   r:=[];
   t:= 0;
   while (i<size d)
      z:=$ Laplace(1/eps,0)
      t:= d[i] + z
      s:= s + t
      r:= r ++ [s];
      i:=i+1;
   return r
```

Parallel Composition

Let M_1:DB $\rightarrow$R be a (ε_1,δ_1)-differentially private program and M_2:DB $\rightarrow$R be a (ε_2,δ_2)-differentially private program. Suppose that we partition D in a data-independent way into two datasets D_1 and D_2. Then, the composition $M_{1,2}$:DB$\rightarrow$R defined as

$$MP_{1,2}(D)=(M_1(D_1),M_2(D_2))$$

is $(\max(\varepsilon_1,\varepsilon_2),\max(\delta_1,\delta_2))$-differentially private.