CS 599: Formal Methods in Security and Privacy

Differential Privacy

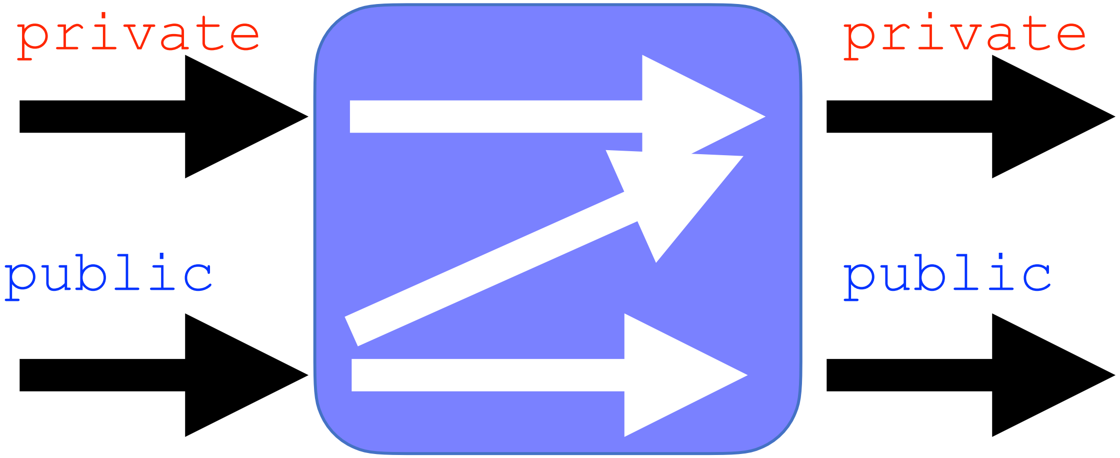
Marco Gaboardi
gaboardi@bu.edu

Alley Stoughton
stough@bu.edu

From the previous classes

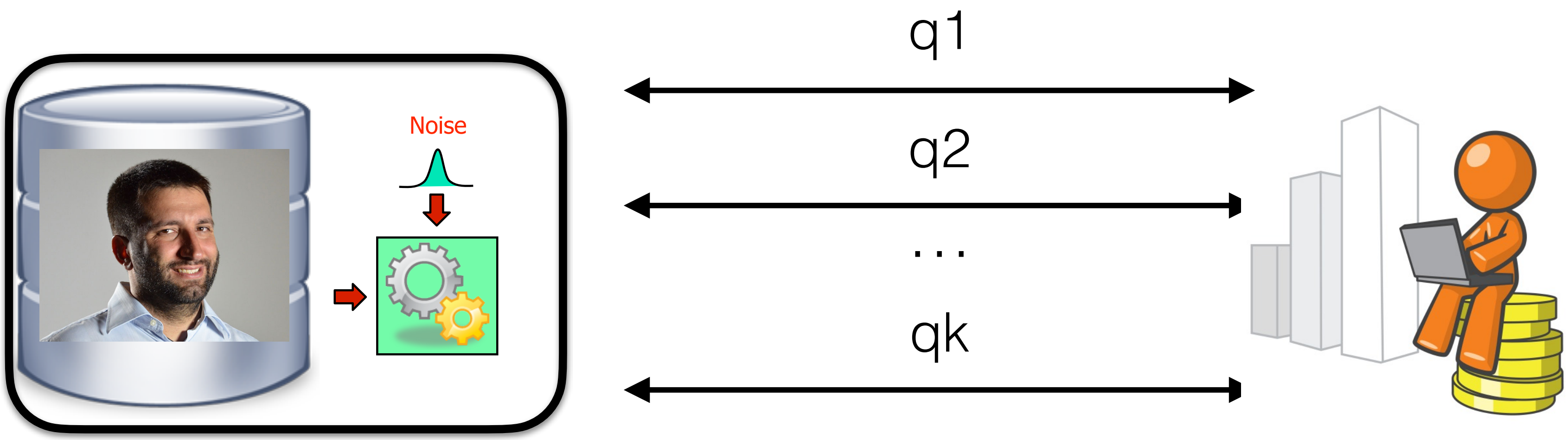
Releasing the mean of Some Data

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



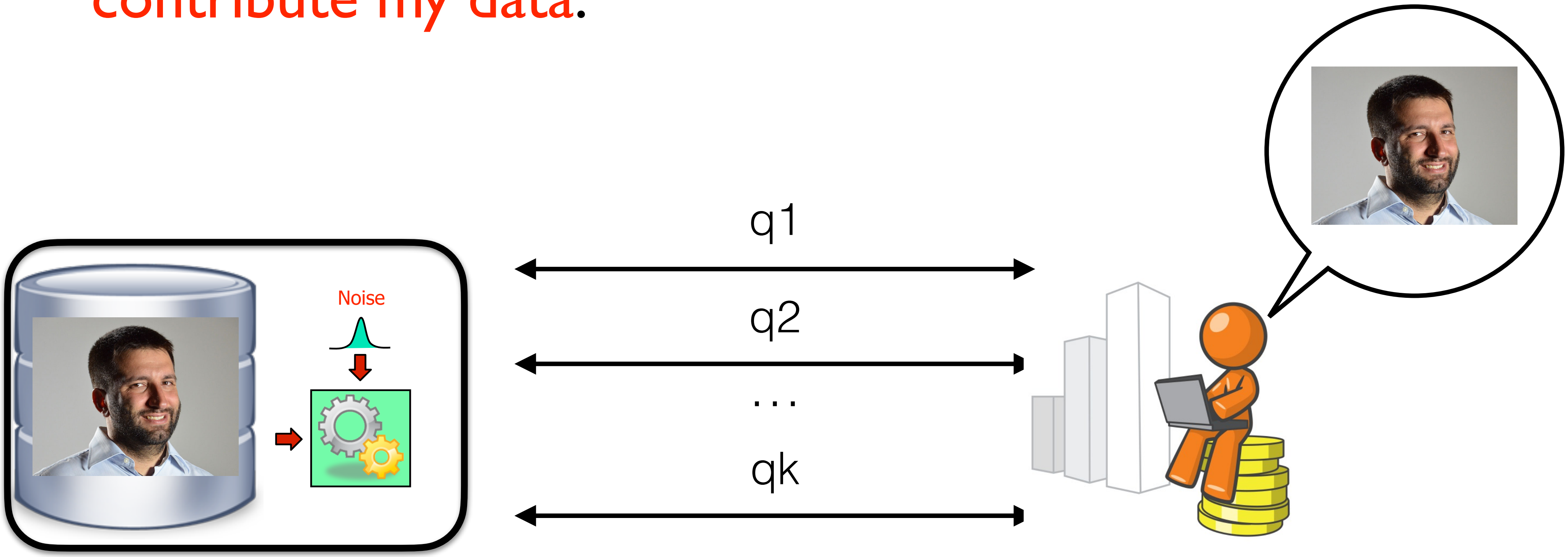
Privacy-preserving data analysis?

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



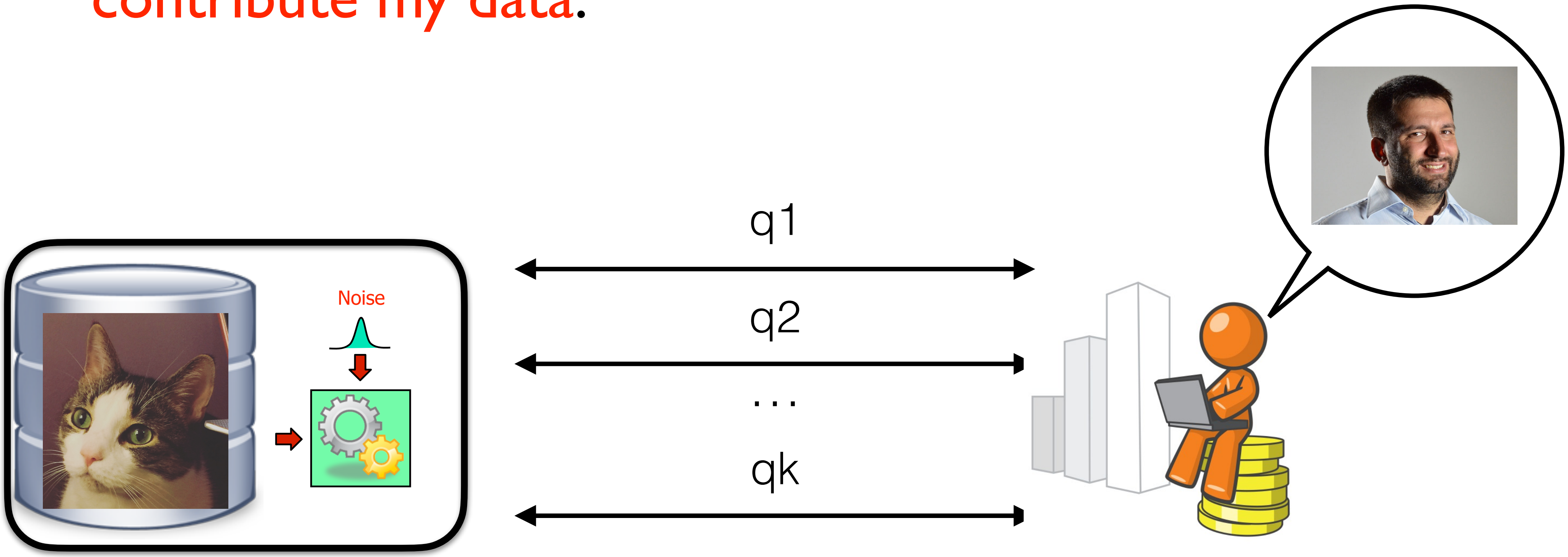
Privacy-preserving data analysis?

- 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.



(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 $S \subseteq R$:

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

(ε,δ)-indistinguishability

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

$$\Delta_{\varepsilon}(\mu 1,\mu 2)=\sup_{E\subseteq A} \max(\mu_1(E)-e^{\varepsilon}\mu_2(E),\ \mu_2(E)-e^{\varepsilon}\mu_1(E),0)$$

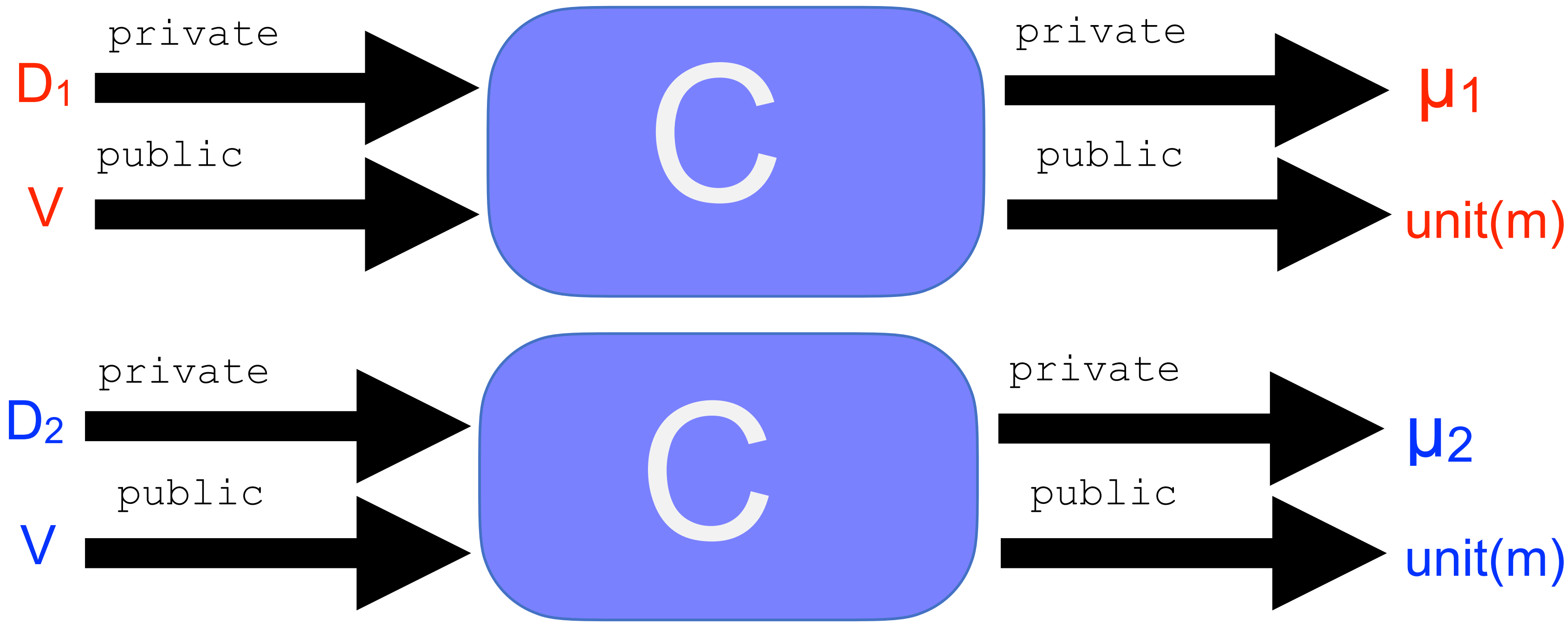
We say that two distributions $\mu_1, \mu_2 \in D(A)$, are at (ε,δ)-indistinguishable if:

$$\Delta_{\varepsilon}(\mu 1,\mu 2) \leq \delta$$

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):

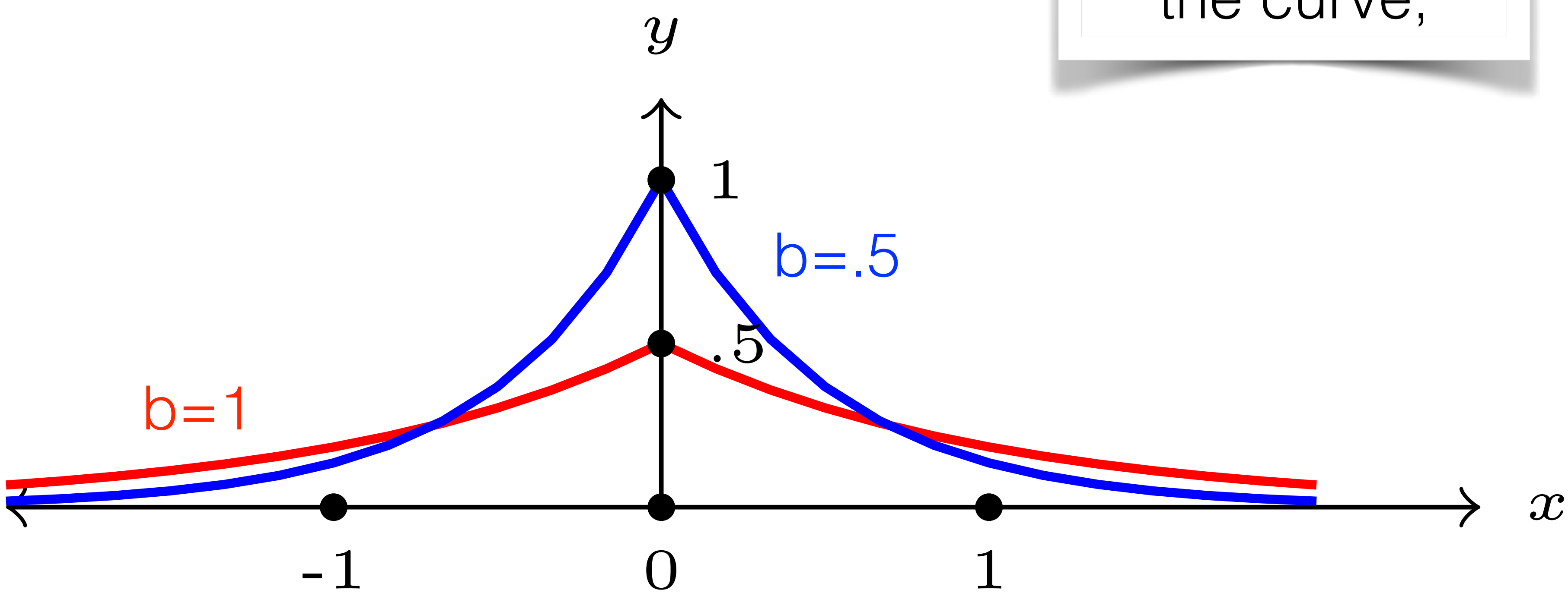
$\{c\}_{m1}=\mu_1$ and $\{c\}_{m2}=\mu_2$ implies $\Delta_\varepsilon(\mu_1,\mu_2) \leq \delta$



Laplace Distribution

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

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
```

Today

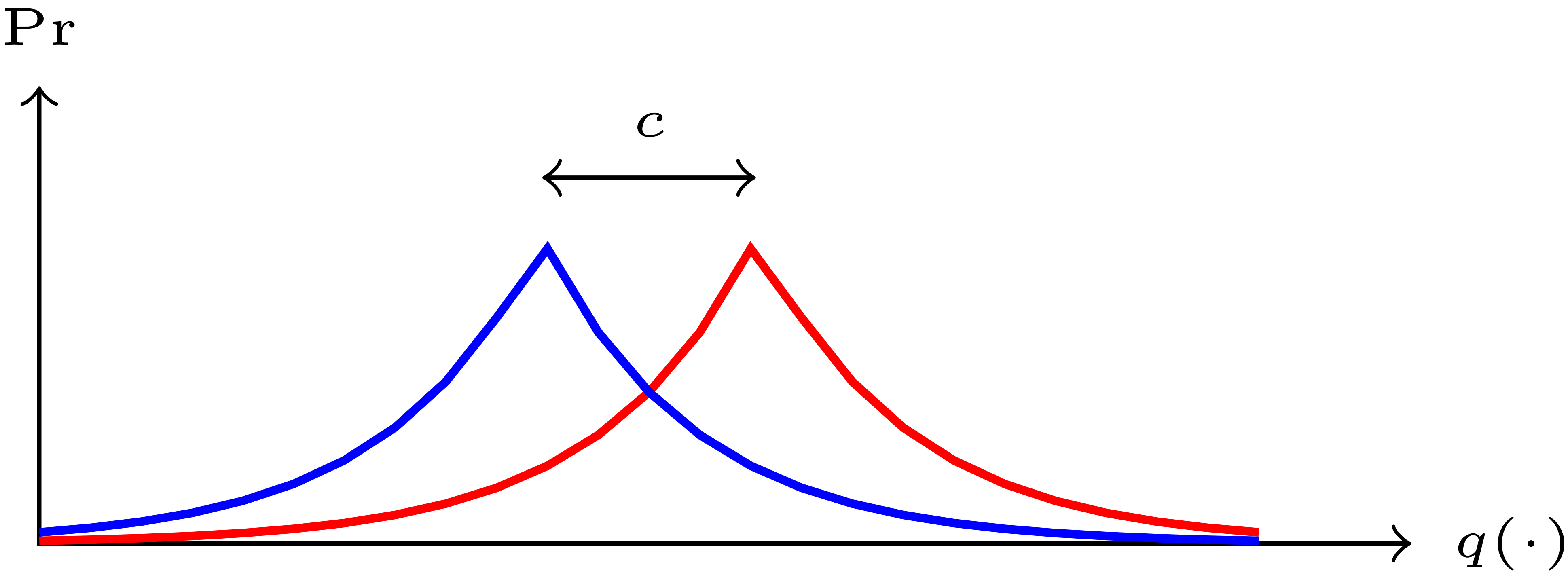
Laplace Mechanism

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 (ε,0)-differentially private.

Proof: Intuitively



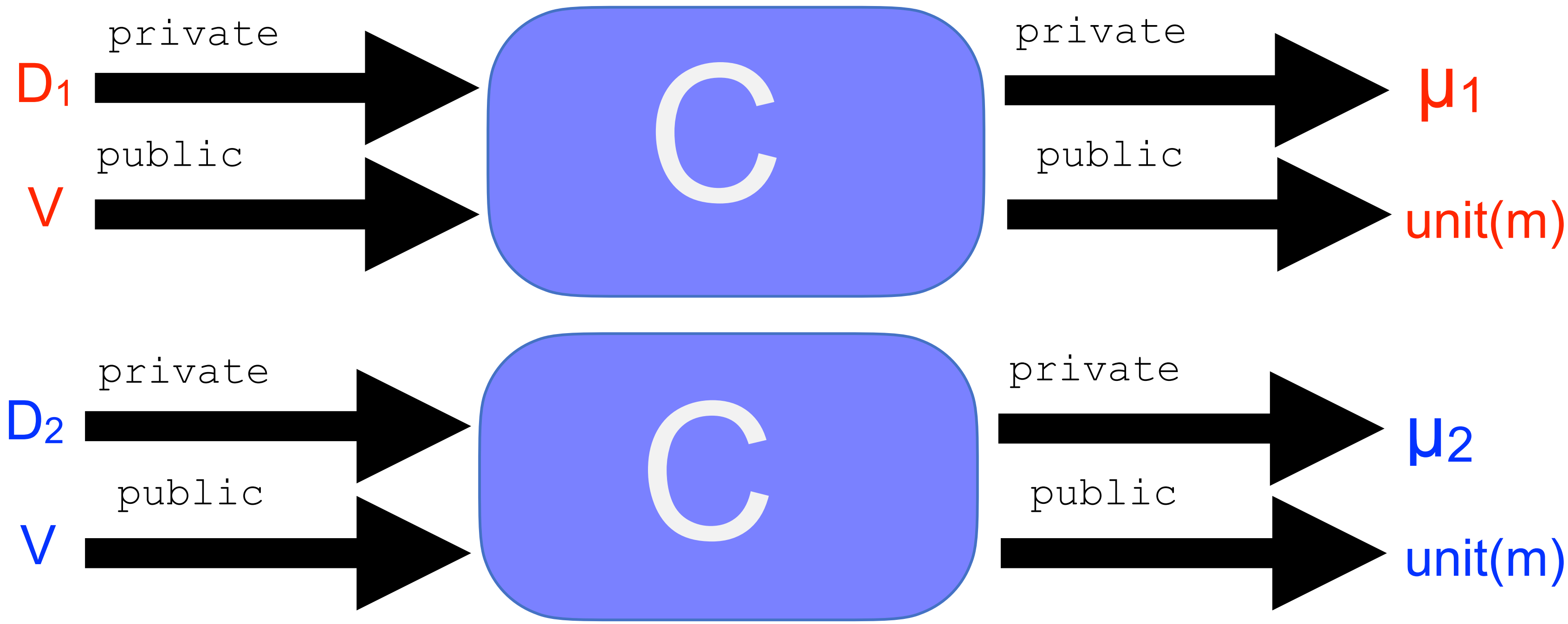
Laplace Mechanism

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

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 $\Delta_{\varepsilon}(\mu_1,\mu_2) \leq \delta$



apRHL

Indistinguishability parameter

Precondition (a logical formula)

$$\vdash_{\epsilon,\delta} c_1 \sim c_2 : P \Rightarrow Q$$

Probabilistic Program

Probabilistic Program

Postcondition (a logical formula)

Validity of apRHL judgments

We say that the quadruple $\vdash_{\varepsilon,\delta} c_1 \sim c_2 : P \Rightarrow Q$ is valid if and only if for every pair of memories m_1, m_2 such that $P(m_1, m_2)$ we have:

$\{c_1\}_{m1} = \mu_1$ and $\{c_2\}_{m2} = \mu_2$ implies $Q_{\varepsilon,\delta}{}^*(\mu_1, \mu_2)$.

Relational lifting

We say that two subdistributions $\mu_1 \subseteq D(A)$ and $\mu_2 \subseteq D(B)$ are in the relational lifting of the relation $R \subseteq A x B$, denoted $\mu_1 \; R_{\varepsilon,\delta}{}^* \; \mu_2$ if and only if we have an R-(ε,δ)-coupling between them.

$\mathcal{R}$-(ε,δ)-Coupling

Given two distributions $\mu_1 \in D(A)$, and $\mu_2 \in D(B)$, we have an $\mathcal{R}$-(ε,δ)-coupling between them, for $\mathcal{R} \subseteq A \times B$ and $0 \le \delta \le 1$, $\varepsilon \ge 0$, if there are two joint distributions $\mu_L, \mu_R \in D(A \times B)$ such that:

1) $\pi_1(\mu_L) = \mu_1$ and $\pi_2(\mu_R) = \mu_2$,
2) the support of μ_L and μ_R is contained in $\mathcal{R}$. That is, if $\mu_L(a,b)>0$, then $(a,b) \in \mathcal{R}$, and if $\mu_R(a,b)>0$, then $(a,b) \in \mathcal{R}$.
3) $\Delta_\varepsilon(\mu_L, \mu_R) \le \delta$

(ε,δ)-indistinguishability revisited

For discrete distributions we can rewrite the notion of ε-distance as follows:

$\Delta\varepsilon(\mu 1,\mu 2)=1/2*\Sigma_{a\in A} \max(\mu 1(a)-e^{\varepsilon}\mu 2(a), \mu 2(E)-e^{\varepsilon}\mu 1(E),0)$

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

$R(a,b) = \{a=b\}$

μ_2

00	0.20
01	0.25
10	0.25
11	0.30

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0.20			
01		0.25		
10			0.25	
11				0.30

$$\Delta_0(\mu_L, \mu_R) = 0.05$$

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

μ_2

00	0.20
01	0.25
10	0.25
11	0.30

$R(a,b) = \{a=b\}$

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0.20			
01		0.25		
10			0.25	
11				0.30

$\Delta_1(\mu_L, \mu_R) = 0$

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

$$R(a,b) = \{a=b\}$$

μ_2

00	0.20
01	0.25
10	0.25
11	0.30

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0.20			
01		0.25		
10			0.25	
11				0.30

$$\Delta_{0.3}(\mu_L, \mu_R) = 0$$

Example of R-(ε,δ)-Coupling

$e^{0.3} \sim 1.3$

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0.20			
01		0.25		
10			0.25	
11				0.30

$$\max(\mu_L(00,00) - e^{0.3}\mu R(00,00), \mu R(00,00) - e^{0.3}\mu_L(00,00), 0) = 0$$

$$+ \max(\mu_L(01,01) - e^{0.3}\mu R(01,01), \mu R(01,01) - e^{0.3}\mu_L(01,01), 0) = 0$$

$$+ \max(\mu_L(10,10) - e^{0.3}\mu R(10,10), \mu R(10,10) - e^{0.3}\mu_L(10,10), 0) = 0$$

$$+ \max(\mu_L(11,11) - e^{0.3}\mu R(11,11), \mu R(11,11) - e^{0.3}\mu_L(11,11), 0) = 0$$

$$\Delta_{0.3}(\mu_L, \mu_R) = 0$$

Example of R-(ε,δ)-Coupling

μ_1

00	0.2
01	0.25
10	0.25
11	0.3

$$R(a,b) = \{a \le b\}$$

μ_2

00	0
01	0.40
10	0
11	0.6

μ_L	00	01	10	11
00		0.20		
01		0.25		
10				0.25
11				0.30

μ_R	00	01	10	11
00		0.20		
01		0.20		
10				0.3
11				0.3

$$\Delta_0(\mu_L, \mu_R) = 0.05$$

Example of R-(ε, δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

μ_2

00	0
01	0
10	0.5
11	0.5

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

$$R(a,b) = \{a=b\}$$

μ_2

00	0
01	0
10	0.5
11	0.5

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

$$R(a,b) = \{a=b\}$$

μ_2

00	0
01	0
10	0.5
11	0.5

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0			
01		0		
10			0.5	
11				0.5

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

μ_2

00	0
01	0
10	0.5
11	0.5

$$R(a,b) = \{a=b\}$$

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0			
01		0		
10			0.5	
11				0.5

$$\Delta_0(\mu_L,\mu_R)=0.5$$

Example of R-(ε, δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

μ_2

00	0
01	0
10	0.5
11	0.5

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

$$R(a,b) = \{a=b\}$$

μ_2

00	0
01	0
10	0.5
11	0.5

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

μ_2

00	0
01	0
10	0.5
11	0.5

$$R(a,b) = \{a=b\}$$

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0			
01		0		
10			0.5	
11				0.5

Example of R-(ε,δ)-Coupling

μ_1

00	0.25
01	0.25
10	0.25
11	0.25

$R(a,b) = \{a=b\}$

μ_2

00	0
01	0
10	0.5
11	0.5

μ_L	00	01	10	11
00	0.25			
01		0.25		
10			0.25	
11				0.25

μ_R	00	01	10	11
00	0			
01		0		
10			0.5	
11				0.5

$$\Delta_1(\mu_L, \mu_R) = 0.25$$

$\mathtt{R}$-(ε,δ)-Coupling and Indistinguishability

Given two distributions $\mu_1 \in D(A)$, and $\mu_2 \in D(A)$, if we have a $=$-(ε,δ)-coupling between them, then they are (ε,δ)-indistinguishable.

Probabilistic Relational Hoare Logic

Skip

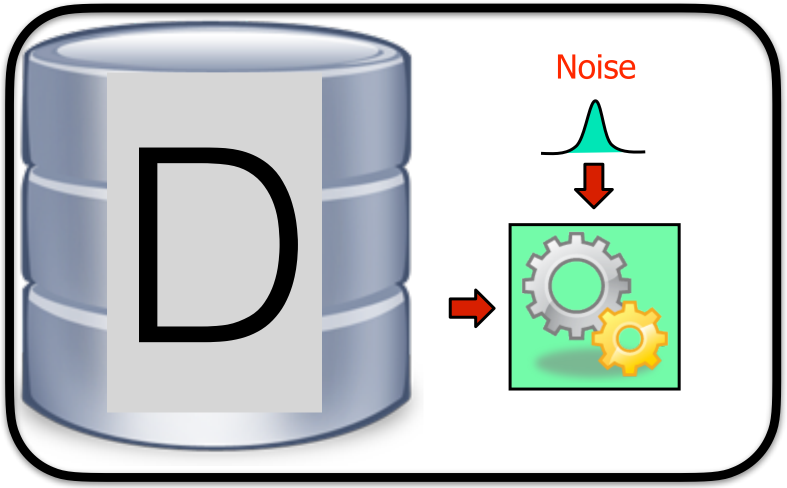
$$\frac{}{\vdash_{0,0} \texttt{skip} \sim \texttt{skip} : P \Rightarrow P}$$

Probabilistic Relational Hoare Logic

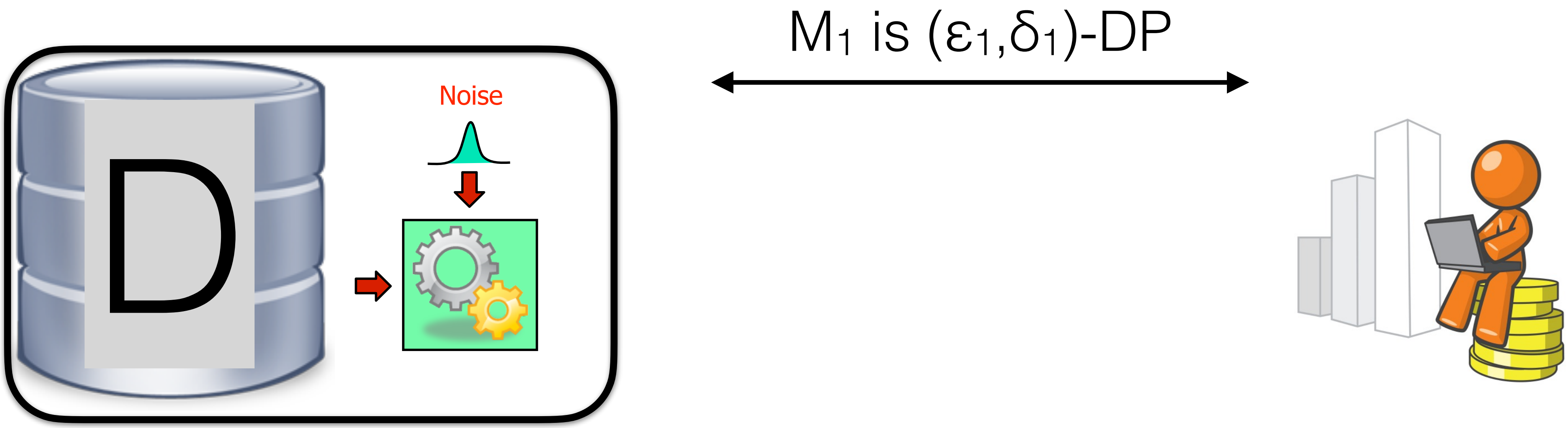
Skip

$$\frac{}{\vdash_{\varepsilon,0} \texttt{x}_1 \texttt{:=\$ Lap}(\varepsilon, \texttt{y}_1) \sim \texttt{x}_2 \texttt{:=\$ Lap}(\varepsilon, \texttt{y}_2) : |\texttt{y}_1 - \texttt{y}_2| \leq 1 \Rightarrow =}$$

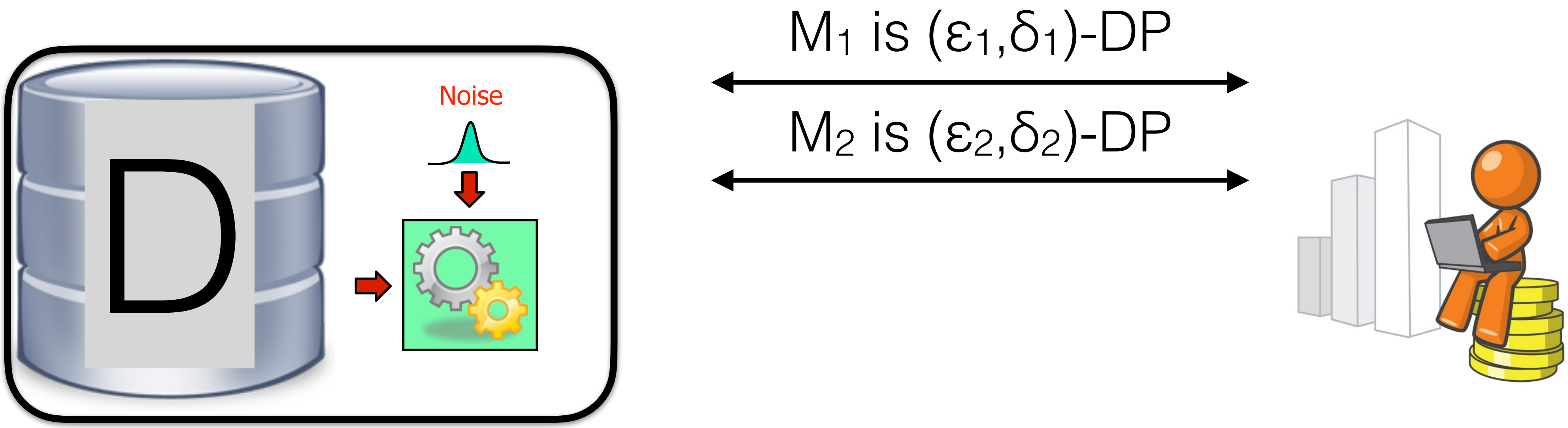
Composition



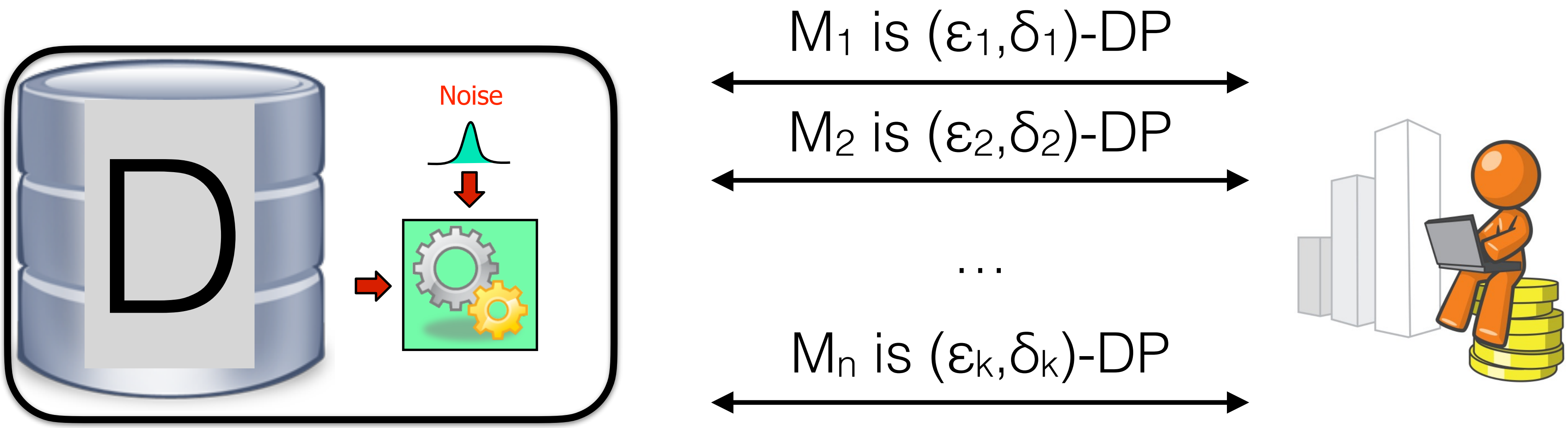
Composition



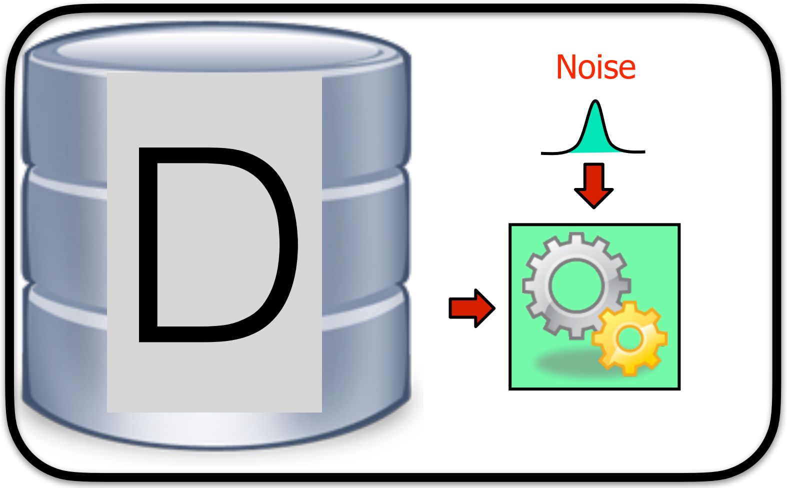
Composition



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?

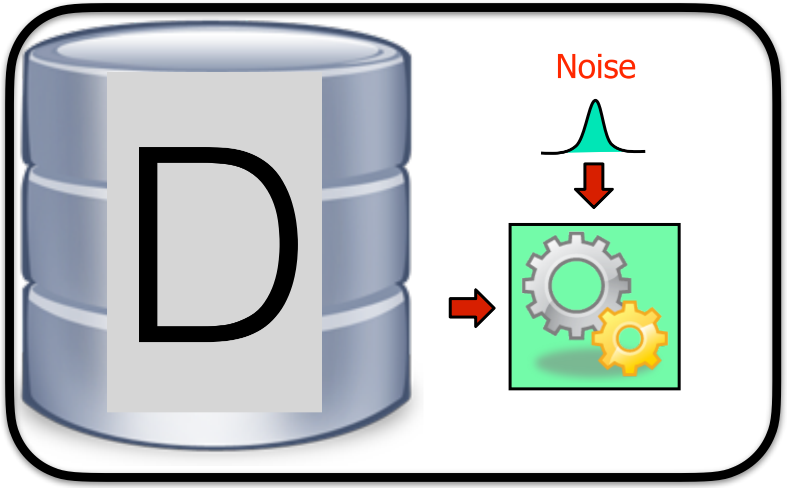
Composition

Question: Why composition is important?

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

Composition

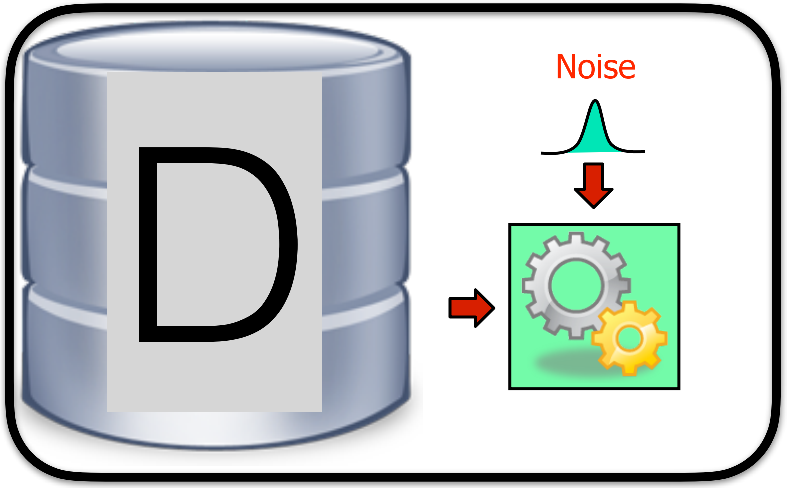
Budget=ε_{global}



Composition

Budget=ε_{global}

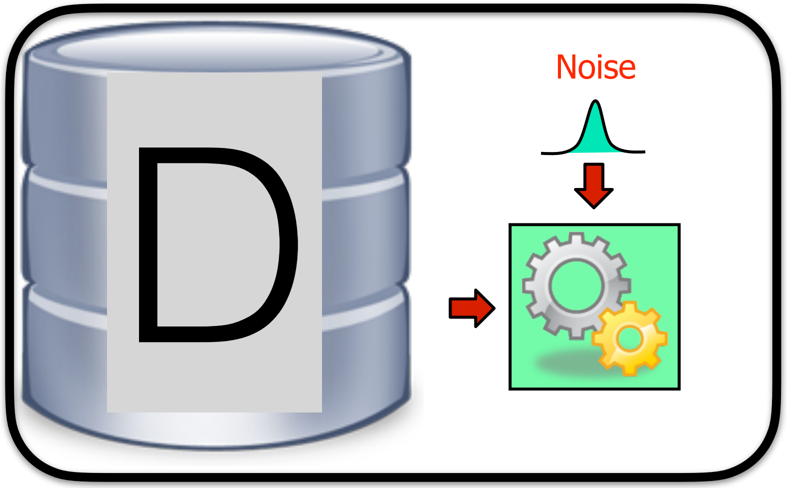
M_1 is ε_1-DP



Composition

Budget=ε_{global} - ε_1

M_1 is ε_1-DP

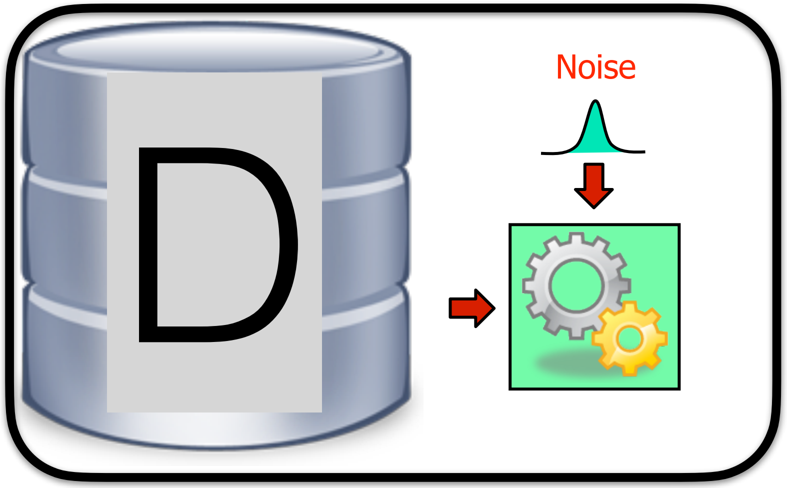


Composition

Budget=ε_{global} - ε_1

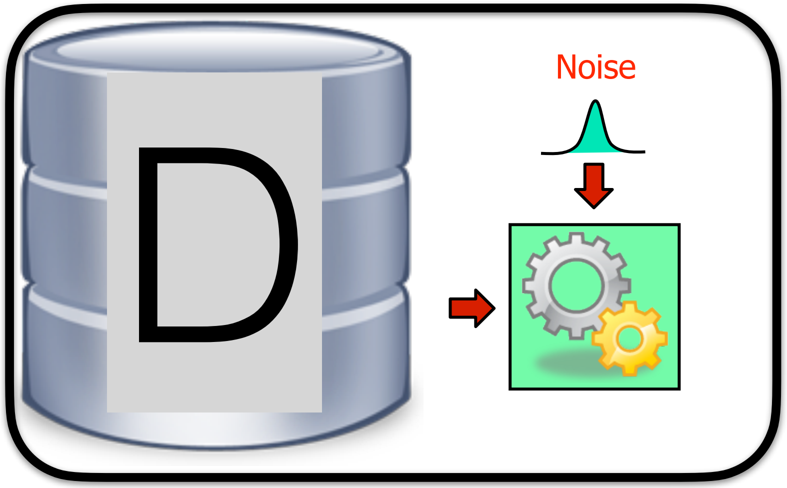
M_1 is ε_1-DP

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

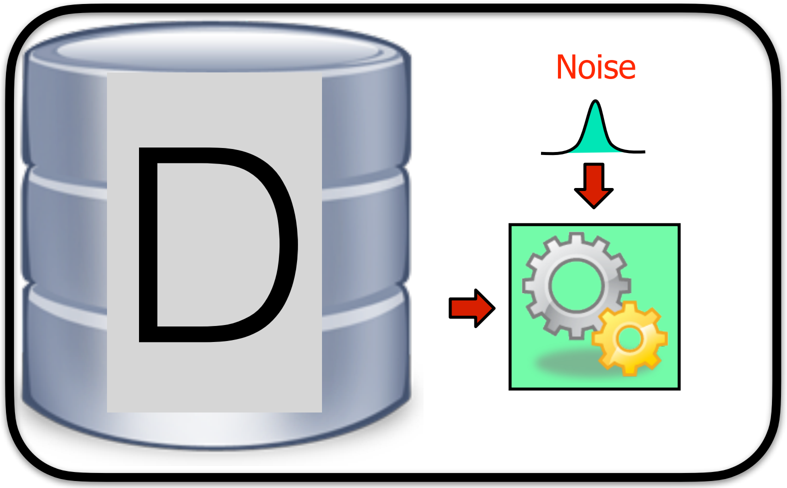
Budget=ε_{global} - ε_1 - ε_2 …

M_1 is ε_1-DP

M_2 is ε_2-DP

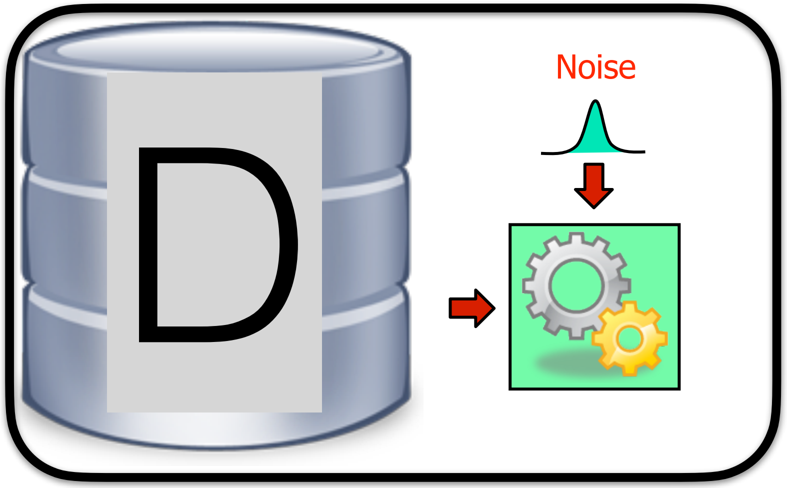
…

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

$q^*_{000}(D) = .3+L(1/10^*\varepsilon_1)$

$q^*_{001}(D) = .4+L(1/10^*\varepsilon_2)$

$q^*_{010}(D) = .6+L(1/10^*\varepsilon_3)$

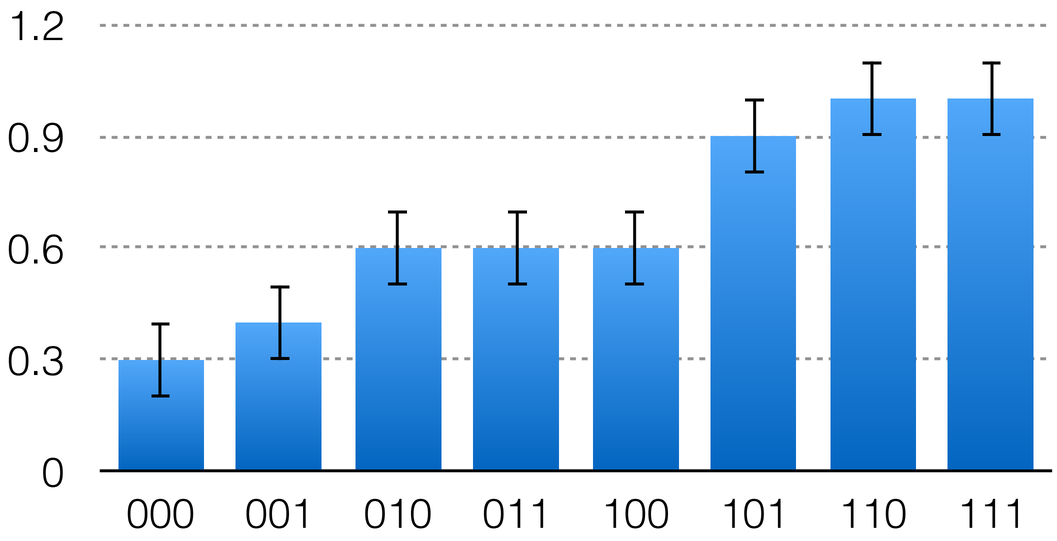
$q^*_{011}(D) = .6+L(1/10^*\varepsilon_4)$

$q^*_{100}(D) = .6+L(1/10^*\varepsilon_5)$

$q^*_{101}(D) = .9+L(1/10^*\varepsilon_6)$

$q^*_{110}(D) = 1+L(1/10^*\varepsilon_7)$

$q^*_{111}(D) = 1+L(1/10^*\varepsilon_8)$



Marginals

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

$D \in X^{10} =$

	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	$.4+Y_1$	$.3+Y_2$	$.4+Y_3$

$$q^*_1(D) = .4 + L(1/(10^*\varepsilon_1))$$
$$q^*_2(D) = .3 + L(1/(10^*\varepsilon_2))$$
$$q^*_3(D) = .4 + L(1/(10^*\varepsilon_3))$$

Releasing partial sums

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

I am using the easycrypt notation here where `Lap(eps,a)` corresponds to adding to the value `a` noise from the Laplace distribution with `b=1/eps` and mean `mu=0`.

Probabilistic Relational Hoare Logic

Composition

$$\frac{\vdash_{\varepsilon 1,\delta 1} c_1 \sim c_2 : P \Rightarrow R \quad \vdash_{\varepsilon 2,\delta 2} c_1' \sim c_2' : R \Rightarrow S}{\vdash_{\varepsilon 1+\varepsilon 2,\delta 1+\delta 2} c_1; c_1' \sim c_2; c_2' : P \Rightarrow S}$$

Releasing partial sums

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