

Differential Privacy

Privacy attacks and definition

Marco Gaboardi

Boston University

Chalmers University

DPella

The opinions expressed in this course are mine and they do not reflect those of the National Science Foundation or the US Census Bureau.

Data



Aol.



Google

Data



Aol.

Statistics over Data



NETFLIX

Google

Databases and Queries

- For simplicity we will focus on a rather abstract notion of databases and queries.
- We will describe a database as a multiset (or sometimes an histogram) and queries as functions from a database to some (often numeric) domain.
- We will usually be interested in the results of some set of queries.

Data

Name	D1	D2	D3	D4	D5	D6	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
Alice	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1
Bob	1	0	1	1	1	0	1	0	1	0	1	0	0	1	0	0
Cynthi	0	1	0	1	1	1	0	1	0	0	0	1	0	0	1	0
Dan	1	0	1	0	0	1	1	0	1	1	0	0	0	0	1	1
Eve	0	0	0	1	1	0	1	1	0	1	0	1	0	1	0	0
Frank	0	0	1	1	0	1	1	0	1	1	0	0	1	0	1	0
Guy	1	1	0	0	1	0	1	1	1	0	1	0	1	0	0	1
Hann	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0
Ivan	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0
Jon	1	0	1	0	0	1	1	0	0	0	0	0	0	1	0	1
Ken	0	1	0	1	1	0	0	1	0	1	0	1	0	1	1	0
Lou	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Mike	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	0
Noa	0	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0
Omer	0	1	0	1	0	1	1	0	1	0	1	0	1	0	0	1

Database

- We can think about a database as a list of records from some universe set:

$$D \in \mathcal{X}^n$$

- Sometimes we will think to them as functions

$$D(k) \in \mathcal{X}$$

- and sometimes we will write elements explicitly

$$(d_1, \dots, d_n) \in \mathcal{X}^n$$

Counting Queries

- A **counting query** $q : \mathcal{X}^n \rightarrow [0, 1]$ is a function counting the fraction of people in a dataset satisfying the **predicate** $q : \mathcal{X} \rightarrow \{0, 1\}$
- In symbols:

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

- Notice that we take a normalized count, which also corresponds to the average.

Example 1

7

Let's consider an arbitrary universe domain \mathcal{X} and let's consider the following predicate for $y \in \mathcal{X}$

$$q_y(x) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

we call a **point function** the associated counting query

$$q_y : \mathcal{X}^n \rightarrow [0, 1]$$

Example 1

7

Let's consider an arbitrary universe domain \mathcal{X} and let's consider the following predicate for $y \in \mathcal{X}$

$$q_y(x) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

we call a **point function** the associated counting query

$$q_y : \mathcal{X}^n \rightarrow [0, 1]$$

Question: Suppose that we answer all the point function queries for $y \in \mathcal{X}$. What well known data summary do we obtain?

Example I

8

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

Example 1

8

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{001}(D) = .1$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{001}(D) = .1$$

$$q_{010}(D) = .2$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{001}(D) = .1$$

$$q_{010}(D) = .2$$

$$q_{011}(D) = 0$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{100}(D) = 0$$

$$q_{001}(D) = .1$$

$$q_{010}(D) = .2$$

$$q_{011}(D) = 0$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{100}(D) = 0$$

$$q_{001}(D) = .1$$

$$q_{101}(D) = .3$$

$$q_{010}(D) = .2$$

$$q_{011}(D) = 0$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{100}(D) = 0$$

$$q_{001}(D) = .1$$

$$q_{101}(D) = .3$$

$$q_{010}(D) = .2$$

$$q_{110}(D) = .1$$

$$q_{011}(D) = 0$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{100}(D) = 0$$

$$q_{001}(D) = .1$$

$$q_{101}(D) = .3$$

$$q_{010}(D) = .2$$

$$q_{110}(D) = .1$$

$$q_{011}(D) = 0$$

$$q_{111}(D) = 0$$

Example 1

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{000}(D) = .3$$

$$q_{001}(D) = .1$$

$$q_{010}(D) = .2$$

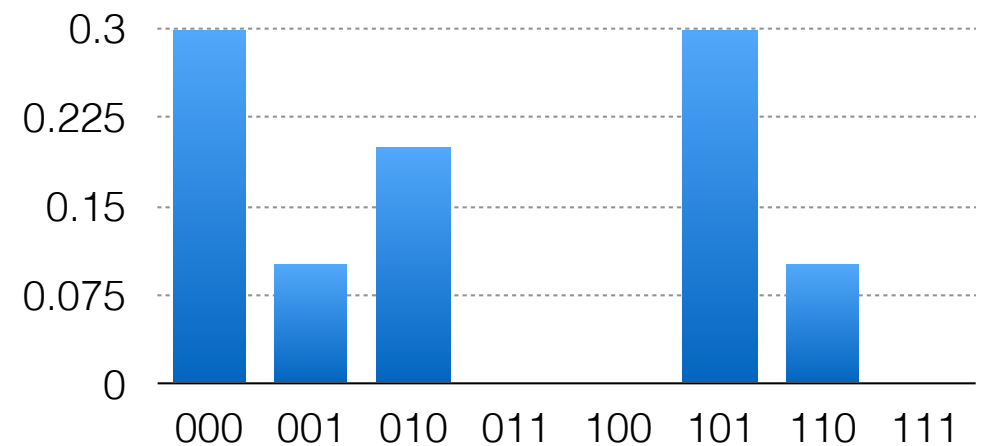
$$q_{011}(D) = 0$$

$$q_{100}(D) = 0$$

$$q_{101}(D) = .3$$

$$q_{110}(D) = .1$$

$$q_{111}(D) = 0$$



Example II

9

Let's consider an arbitrary **ordered** universe domain \mathcal{X} and let's consider the following predicate for $y \in \mathcal{X}$

$$q_y(x) = \begin{cases} 1 & \text{if } x \leq y \\ 0 & \text{otherwise} \end{cases}$$

we call a **threshold function** the associated counting query

$$q_y : \mathcal{X}^n \rightarrow [0, 1]$$

Example II

9

Let's consider an arbitrary **ordered** universe domain \mathcal{X} and let's consider the following predicate for $y \in \mathcal{X}$

$$q_y(x) = \begin{cases} 1 & \text{if } x \leq y \\ 0 & \text{otherwise} \end{cases}$$

we call a **threshold function** the associated counting query

$$q_y : \mathcal{X}^n \rightarrow [0, 1]$$

Question: Suppose that we answer all the threshold function queries for $y \in \mathcal{X}$. What well know statistics do we obtain?

Example II

10

$X = \{0, 1\}^3$
with order
given by the
corresponding
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

Example II

10

$X = \{0, 1\}^3$
with order
given by the
corresponding
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

$$q_{000}(D) = .3$$

$$q_{001}(D) = .4$$

$$q_{010}(D) = .6$$

$$q_{011}(D) = .6$$

$$q_{100}(D) = .6$$

$$q_{101}(D) = .9$$

$$q_{110}(D) = 1$$

$$q_{111}(D) = 1$$

Example II

10

$X = \{0, 1\}^3$
with order
given by the
corresponding
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

$$q_{000}(D) = .3$$

$$q_{001}(D) = .4$$

$$q_{010}(D) = .6$$

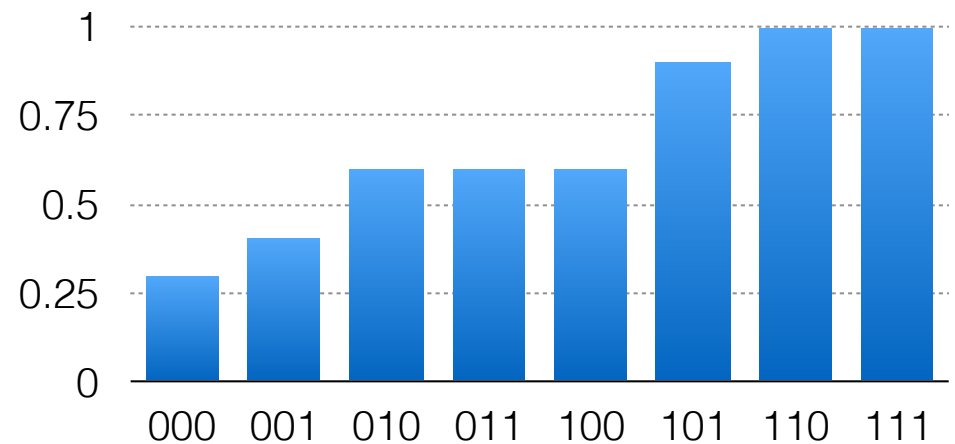
$$q_{011}(D) = .6$$

$$q_{100}(D) = .6$$

$$q_{101}(D) = .9$$

$$q_{110}(D) = 1$$

$$q_{111}(D) = 1$$



Example III

II

Let's consider the universe domain $\mathcal{X} = \{0, 1\}^d$ and let's consider $\vec{v} \in \{1, \bar{1}, \dots, d, \vec{d}\}^k$ with $1 \leq k \leq d$ and

$$q_{\vec{v}}(x) = q_{v_1}(x) \wedge q_{v_2}(x) \wedge \dots \wedge q_{v_k}(x)$$

where $q_j(x) = x_j$ and $q_{\bar{j}}(x) = \neg x_j$

We call a **conjunction** or k-way marginal the associated counting query

$$q_{\vec{v}} : \mathcal{X}^n \rightarrow [0, 1]$$

Example III

II

Let's consider the universe domain $\mathcal{X} = \{0, 1\}^d$ and let's consider $\vec{v} \in \{1, \bar{1}, \dots, d, \vec{d}\}^k$ with $1 \leq k \leq d$ and

$$q_{\vec{v}}(x) = q_{v_1}(x) \wedge q_{v_2}(x) \wedge \dots \wedge q_{v_k}(x)$$

where $q_j(x) = x_j$ and $q_{\bar{j}}(x) = \neg x_j$

We call a **conjunction** or k-way marginal the associated counting query

$$q_{\vec{v}} : \mathcal{X}^n \rightarrow [0, 1]$$

Question: Which statistics does correspond to releasing conjunctions?

Example III

12

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$k=2$$

Example III

I2

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$k=2$$

$$q_{12}(D) = .1$$

$$q_{/12}(D) = .2$$

$$q_{1/2}(D) = .3$$

$$q_{/13}(D) = .1$$

$$q_{13}(D) = .3$$

$$q_{/1/2}(D) = .4$$

$$q_{1/3}(D) = .1$$

$$q_{/1/3}(D) = .5$$

Example III

I2

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$k=2$$

$$q_{12}(D) = .1$$

$$q_{1/2}(D) = .3$$

$$q_{13}(D) = .3$$

$$q_{1/3}(D) = .1$$

$$q_{/12}(D) = .2$$

$$q_{/13}(D) = .1$$

$$q_{/1/2}(D) = .4$$

$$q_{/1/3}(D) = .5$$

	D1	/D1
D2	0.1	0.2
/D2	0.3	0.4

Linear Queries

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

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

Sum queries

- Let's denote by $I \subseteq [n]$ a subset I of

$$\{0, \dots, n\}$$

- A **sum query** $q_I : \{0, 1\}^k \rightarrow \mathbb{N}^k$ is defined as

$$q_I(D) = \sum_{i \in I} d_i$$

Example

15

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

Example

I5

$$X = \{0, 1\}^3$$

$$D \in X^{10} =$$

	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

$$q_{\{1,2,3\}}(D) = (1, 1, 1)$$

Example

I5

$$X = \{0, 1\}^3 \quad D \in X^{10} =$$

	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

$$q_{\{1,2,3\}}(D) = (1, 1, 1)$$

$$q_{\{1,2,4\}}(D) = (2, 0, 2)$$

Example

15

$$X = \{0, 1\}^3 \quad D \in X^{10} =$$

	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

$$q_{\{1,2,3\}}(D) = (1, 1, 1)$$

$$q_{\{1,2,4\}}(D) = (2, 0, 2)$$

$$q_{\{5,8\}}(D) = (0, 0, 0)$$

Example

15

$$X = \{0, 1\}^3 \quad D \in X^{10} =$$

	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

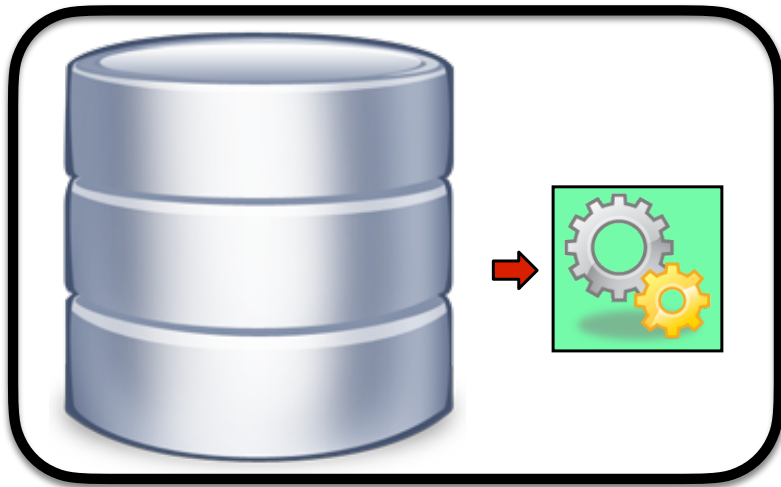
$$q_{\{1,2,3\}}(D) = (1, 1, 1)$$

$$q_{\{1,2,4\}}(D) = (2, 0, 2)$$

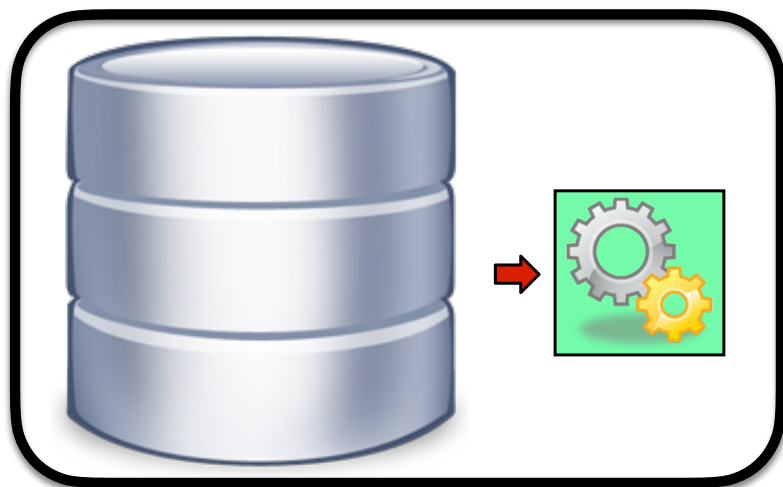
$$q_{\{5,8\}}(D) = (0, 0, 0)$$

$$q_{\{2,4,7,10\}}(D) = (4, 1, 3)$$

Privacy Mechanism



Privacy Mechanism



medical
correlation?

query/answer



Clin Chem Lab Med 2019;57(10):1755-1761 | DOI: 10.1515/clinchem-2019-1058

Letter to the Editor

Methodological weakness in using correlation coefficients for assessing the interchangeability of analyte data between samples collected under different sampling conditions – the example of matrix metalloproteinase 9 determined in serum and plasma samples

Klaus Jung^{1,2*} and Chen-Ying Wu^{3,4}

¹Department of Internal Medicine, University of Medicine, Berlin, Germany
²North German Institute of Laboratory Medicine, Division of Gastroenterology, Teaching Hospital, University of Medicine, Berlin, Germany
³Faculty of Medicine, National Yang-Ming University, Taipei, Taiwan
⁴Department of Public Health, College of Public Health, China Medical University, Taichung, Taiwan

Keywords: comparative measurement; correlation coefficient; analytical factors; matrix metalloproteinase; sampling conditions

In this journal, Gurbach et al. (1) recently reported a positive correlation between plasma and serum matrix metalloproteinase 9 (MMP-9) values. From these findings the authors concluded that these conditions might be useful for interpreting MMP-9 values from other studies performed with serum samples. However, other studies have shown that serum samples are inappropriate, and do not reflect the circulating MMP-9 in peripheral blood because of the additional non-specific release of MMP-9 from blood cells during collection of serum (reviewed in 2). Gurbach et al. (1) failed to present data with which to justify their conclusion. We believe that our comparison of MMP-9 in serum and plasma samples, both from patients and healthy persons, in this article, could be helpful for clarifying this issue.

Comparative measurement of MMP-9 in serum and plasma, as shown by Gurbach et al. (1), can be considered as a typical example for the dissemination of the same analysis.

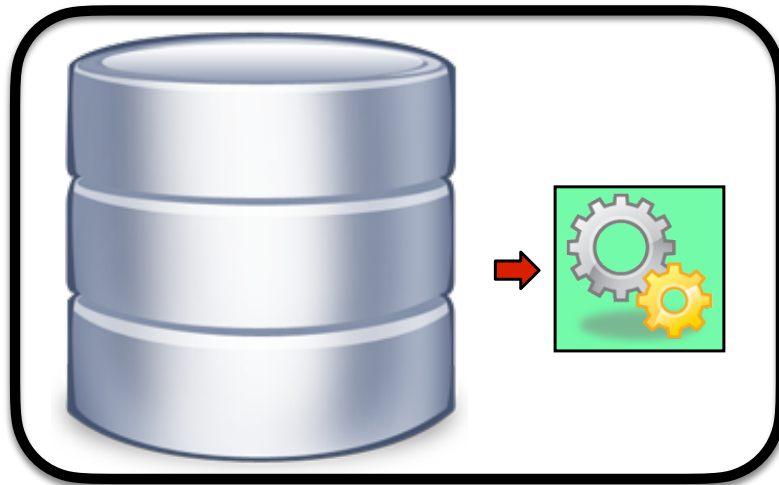
*Corresponding author: Prof. Dr. Klaus Jung, Department of Internal Medicine, University of Medicine, Berlin, Germany, E-mail: klaus.jung@charite.de

samples collected under differing conditions with possible practical implications. The assumption that correlation between variables, in this case serum and plasma MMP-9, allows the interchangeability of data, is not only unsound for MMP-9, but is of general interest for similar examples in laboratory medicine. The use of correlation data for this purpose is unreliable, to a considerable degree, due to the correlation coefficient, as the coefficient of correlation is not a measure of the strength of the relationship between two or more variables. The closer the coefficient value of r is to 1, the higher the strength between the variables. Therefore, correlation coefficients are used to estimate the association in a test result with clinical data. In the past, it was also applied to express the quality and strength between the results obtained with different methods. However, in 1988 Haidich and Altman (3) published their seminal work, showing that the correlation coefficient is an inappropriate statistic to assess the agreement between methods. The correlation coefficient does not measure the agreement. To obtain the linear relationship of data pairs of continuous variables, the correlation coefficient indicates the variation of values around the regression line and does not reflect the precision of fit of the linear regression model itself. The correlation coefficient is significantly influenced by outliers, and tends to get large positive correlation results.

Correlation can be calculated for each individual and average also results in a single coefficient of correlation. This method cannot test the probability of error of the correlation coefficient. For assessing method agreement (repeated in 4, 5), special methods of regression analysis involving the plotting of data on Bland-Altman difference against mean values and the use of the limits of agreement are generally accepted to evaluate method agreement in laboratory medicine (6, 7).

To assess the technique of correlation data as described by Gurbach et al. (1), we re-evaluated the aspect of agreement criteria using the uncorrelated results of simultaneous MMP-9 measurements in serum and plasma from our preceding studies (7, 8). Clinical and methodical details, as well as the approval of these studies by the Ethical Commission of the hospitals were reported previously (7, 8). Briefly, one

Privacy Mechanism

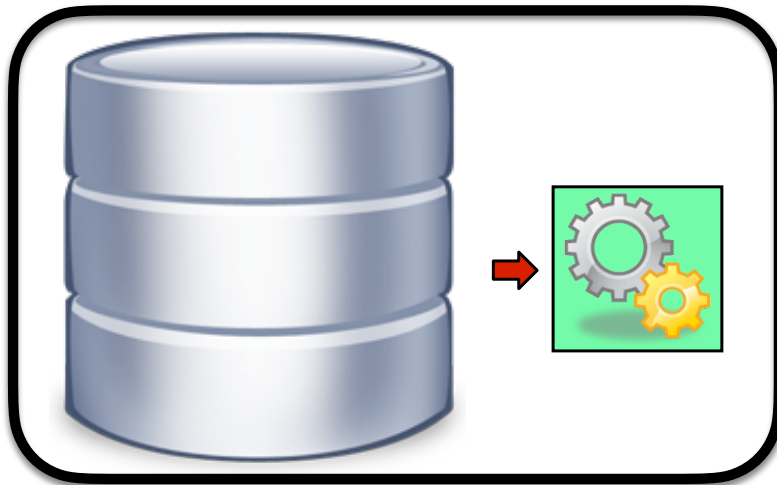


Does Joe
have
cancer?

query/answer



Private Mechanism



Does Joe
have
cancer?



Clin Chem Lab Med 2015;53(1):155-158 | DOI: 10.1055/s-0034-1000000

Letter to the Editor

Methodological weakness in using correlation coefficients for assessing the interchangeability of analyte data between samples collected under different sampling conditions – the example of matrix metalloproteinase 9 determined in serum and plasma samples

Klaus Jung^{1,2*} and Chun-Ying Wu^{3,4}

¹Department of Emergency, Tianjin University of Traditional Chinese Medicine, Tianjin, China
²State Key Laboratory for Clinical Research, Tianjin University of Traditional Chinese Medicine, Tianjin, China
³Department of Gastroenterology, Tianjin University of Traditional Chinese Medicine, Tianjin, China
⁴Department of Public Health, Tianjin University of Traditional Chinese Medicine, Tianjin, China

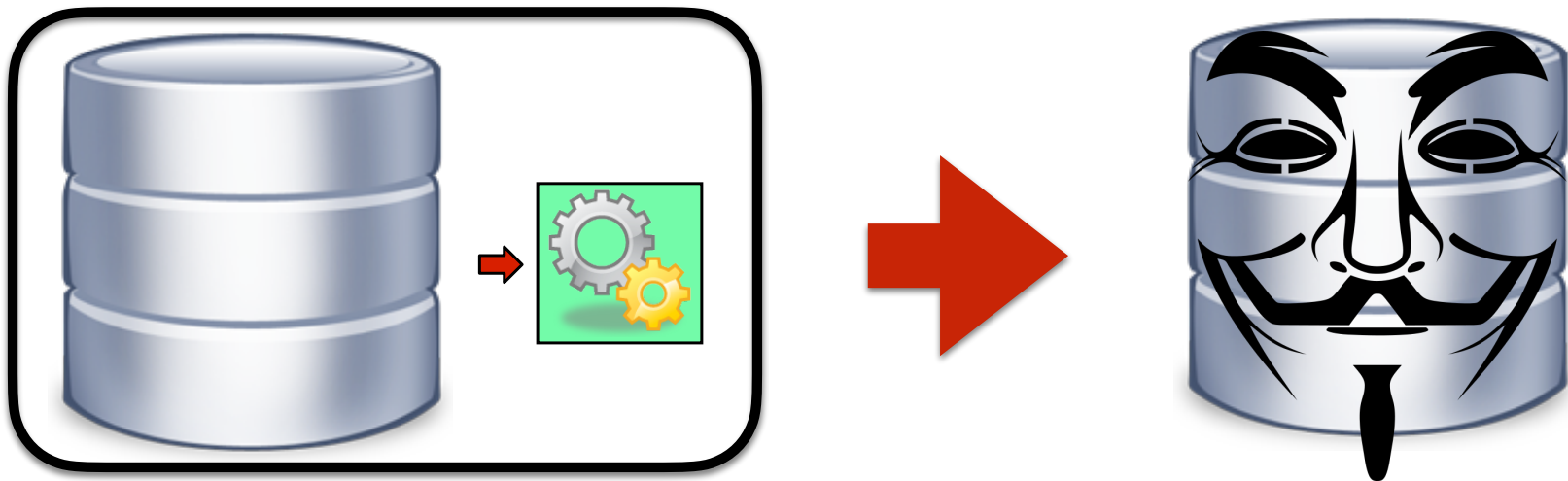
Keywords: comparative measurements; correlation coefficient; analytical factors; matrix metalloproteinase; sampling conditions

In this journal, Gurbach et al. (1) recently reported a positive correlation between plasma and serum matrix metalloproteinase 9 (MMP-9) values. From these findings the authors concluded that these conditions might be useful for interpreting MMP-9 values from other studies performed with serum samples. However, other studies have shown that serum samples are inappropriate, and do not reflect the circulating MMP-9 in peripheral blood because of the additional non-specific release of MMP-9 from blood cells during collection of serum (reviewed in 2). Gurbach et al. (1) failed to provide any data to support their conclusion. The methodological weakness in this article, could be helpful for clarifying this issue and as a typical example for the dissemination of the same analysis in other journals.

sample collected under different conditions with possible practical implications. The assumption that correlation between variables, in this case serum and plasma MMP-9, allows the interchangeability of data, is not only meaningless for MMP-9, but is of general interest for similar examples in laboratory medicine. The use of correlation data for this purpose results in a considerable loss of the correlation coefficient, as the calculation of correlation coefficients. The correlation coefficient, calculated the Pearson product-moment correlation coefficient, and the coefficient of determination (r^2), are of the variables, characterizing the degree of relationship between two or more variables. The closer the Pearson value of r is to 1, the higher the strength between the variables. Therefore, correlation coefficients are used to estimate the association of a new study with clinical data. In the past, it was also applied to express the quality and strength between the results obtained with different methods. However, in 1998 Haidich and Altman (3) published their seminal article, showing that the correlation coefficient method. The correlation coefficient does not measure the agreement. To obtain the linear relationship of data points of continuous variables. The correlation coefficient indicates the variation of values around the regression line and does not reflect the precision of fit of the linear regression model itself. The correlation coefficient is significantly influenced by outliers, and tends to get large positive correlation results. Therefore, we are not convinced by the authors' conclusion that the positive correlation of MMP-9 in serum and plasma samples is a good example for the dissemination of the same analysis in other journals. The methodological weakness in this article, could be helpful for clarifying this issue and as a typical example for the dissemination of the same analysis in other journals. The methodological weakness in this article, could be helpful for clarifying this issue and as a typical example for the dissemination of the same analysis in other journals.

What can be a good privacy mechanism?

A natural idea: anonymizing the data



- E.g. stripping PII, guaranteeing k-anonymity, swapping, etc.

Attacks on stripping PII

(Narayanan, Shmatikov: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008)



Anonymous Data



Attacks on stripping PII

(Narayanan, Shmatikov: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008)



Additional Data



Anonymous Data



Attacks on Swapping

(Garfinkel, Abowd, Martindale: Understanding Database Reconstruction Attacks on Public Data. ACM Queue 16(5): 50 (2018))



Anonymous Data



Attacks on Swapping

(Garfinkel, Abowd, Martindale: Understanding Database Reconstruction Attacks on Public Data. ACM Queue 16(5): 50 (2018))

Commercially
available data



Additional Data



Anonymous Data



Attacks on K-anonymity

(A. Cohen: Attacks on Deidentification's Defenses.
Usenix Security 2022)



Anonymous Data



Attacks on K-anonymity

(A. Cohen: Attacks on Deidentification's Defenses.
Usenix Security 2022)

LinkedIn



edX

Additional Data



Anonymous Data



Why these anonymization techniques runs into troubles?

An issue that these attacks highlight is that it is difficult if not impossible to think about privacy as a property of the data.

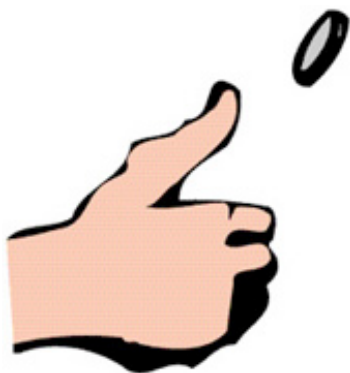
Another issue with these anonymization notions is that they are not closed under postprocessing.

Question: How can we guarantee closure under postprocessing?

Question: How can we guarantee closure under postprocessing?

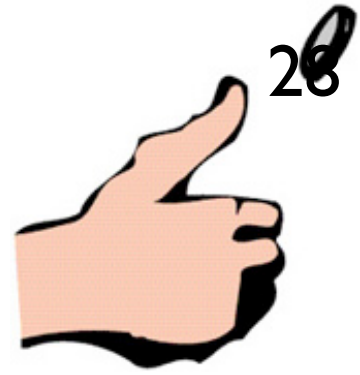


Question: How can we guarantee closure under postprocessing?



Warner, S. L. (March 1965). "Randomised response: a survey technique for eliminating evasive answer bias". *Journal of the American Statistical Association*. Taylor & Francis. 60 (309): 63–69.

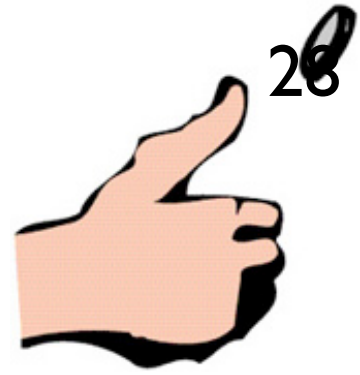
Randomized Algorithms



- Given a discrete set B the **probability simplex** over B , denoted $\Delta(B)$ is defined as:

$$\Delta(B) = \left\{ x \in \mathbb{R}^{|B|} : \forall i, x_i \geq 0, \text{ and } \sum_{i=1}^{|B|} x_i = 1 \right\}$$

Randomized Algorithms

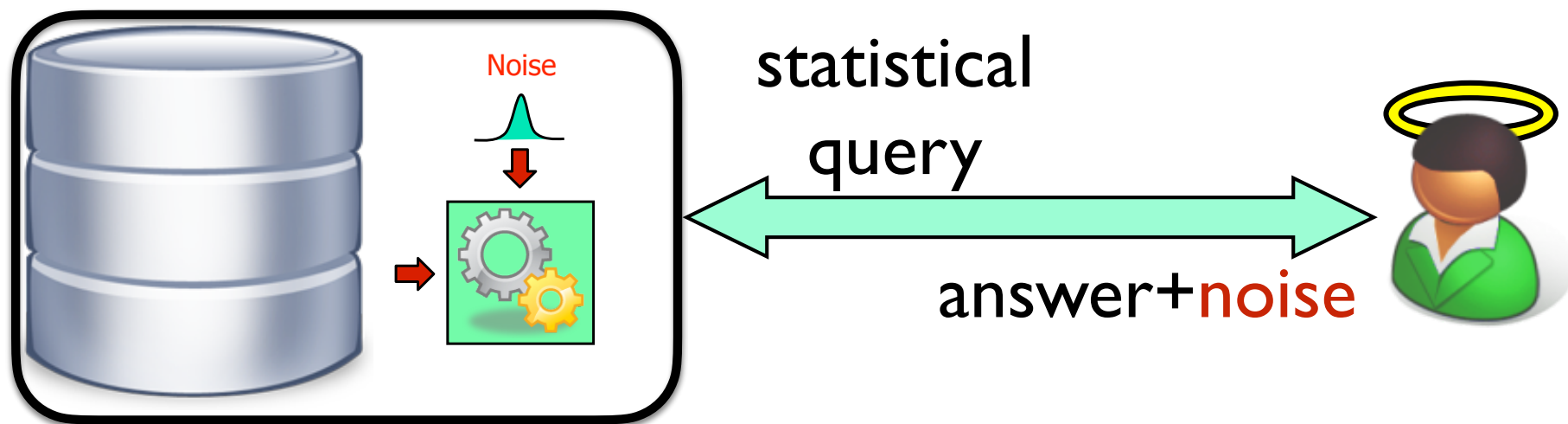


- Given a discrete set B the **probability simplex** over B , denoted $\Delta(B)$ is defined as:

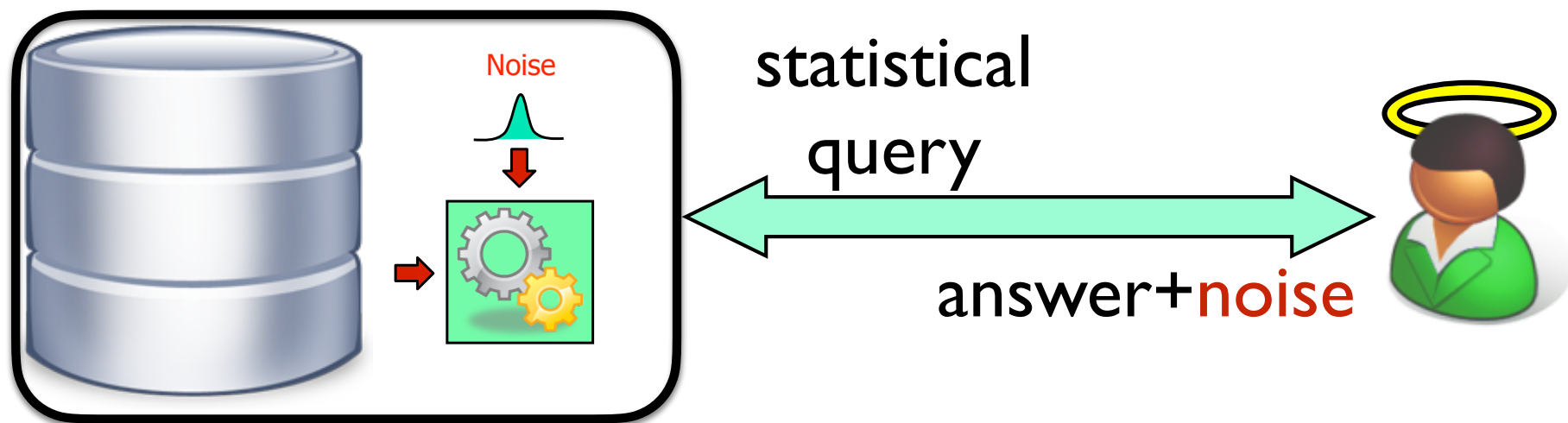
$$\Delta(B) = \left\{ x \in \mathbb{R}^{|B|} : \forall i, x_i \geq 0, \text{ and } \sum_{i=1}^{|B|} x_i = 1 \right\}$$

- A **randomized algorithm** \mathcal{M} is an algorithm associated with a total map $M : A \rightarrow \Delta(B)$
On input $a \in A$ the algorithm outputs $\mathcal{M}(a) = b$ with probability $(M(a))_b$.
The probability space is over the coin flips of the algorithm.

Private Statistical database



Private Statistical database



Question: What kind of noise?

Sum queries

- Let's denote by $I \subseteq [n]$ a subset I of $\{0, \dots, n\}$
- A **sum query** $q_I : 0, 1^k \rightarrow \mathbb{N}^k$ is defined as

$$q_I(D) = \sum_{i \in I} d_i$$

Uniform Noise

- Given a query q we want to add noise to create a new randomized query:

$$q^*(D) = q(D) + Y$$

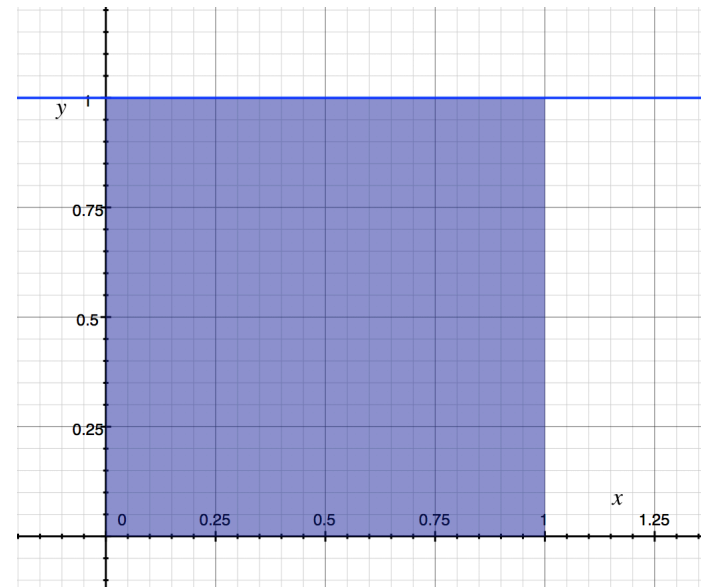
Uniform Noise

- Given a query q we want to add noise to create a new randomized query:

$$q^*(D) = q(D) + Y$$

- One way to do this is to sample Y from the uniform distribution:

$$Y \sim U[0,1]$$

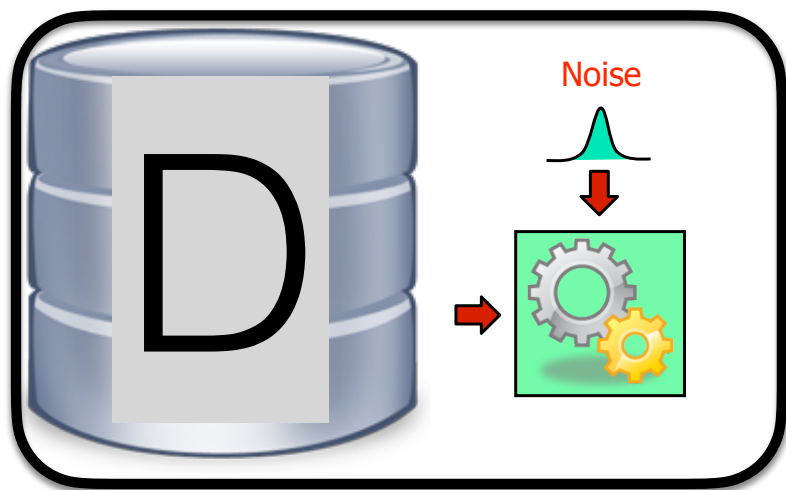


Question: Does this approach prevent privacy attacks?

Reconstruction attack

- Consider an **adversary** A (an algorithm) that has access to some data D through a privacy mechanism q^* .
- The goal of the **adversary** is to output some data D' that is as similar as possible to D .
- To output D' the **adversary** can interact several times with q^* .

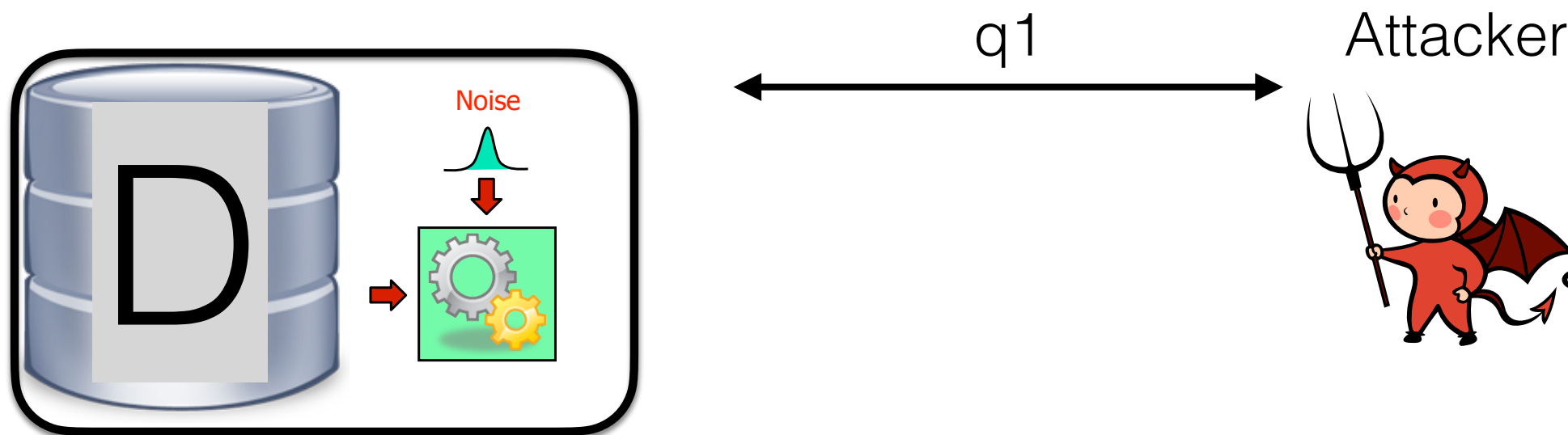
Reconstruction attack



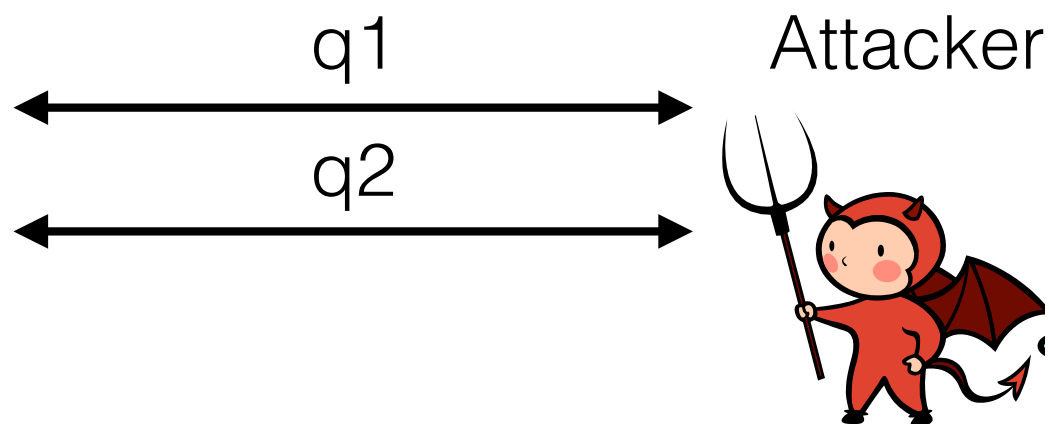
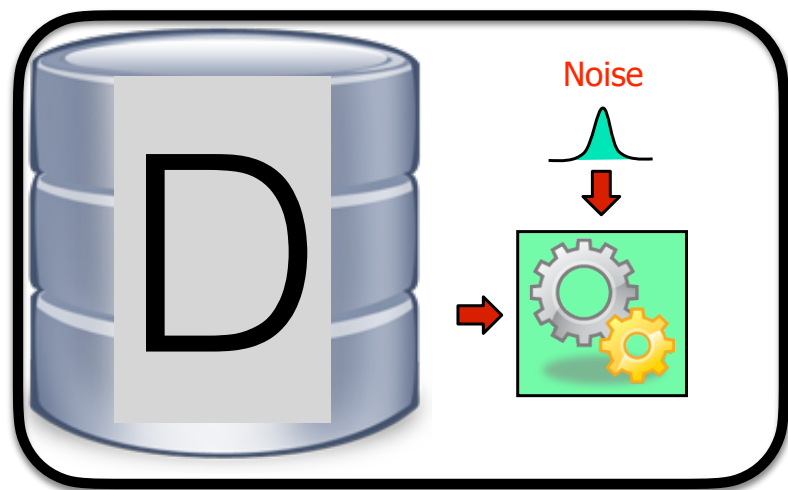
Attacker



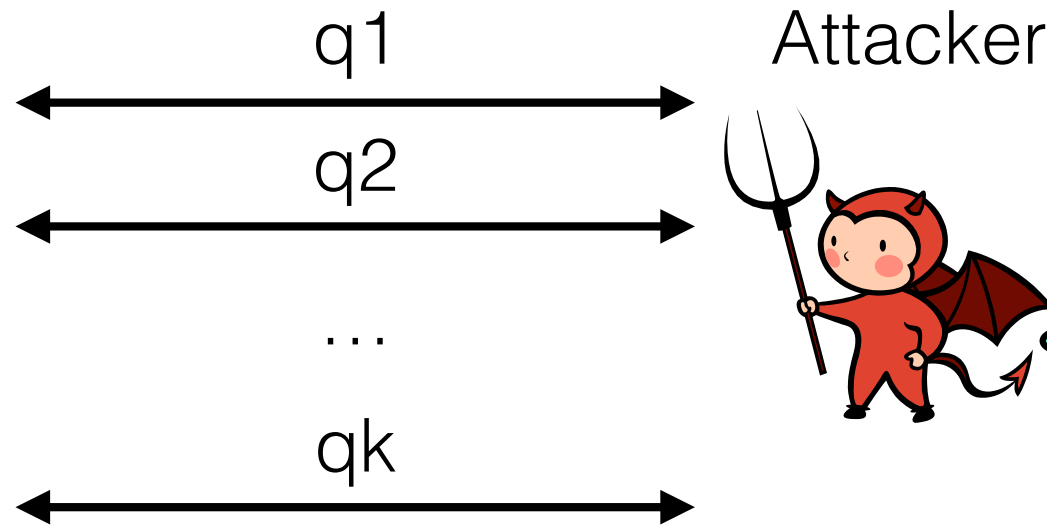
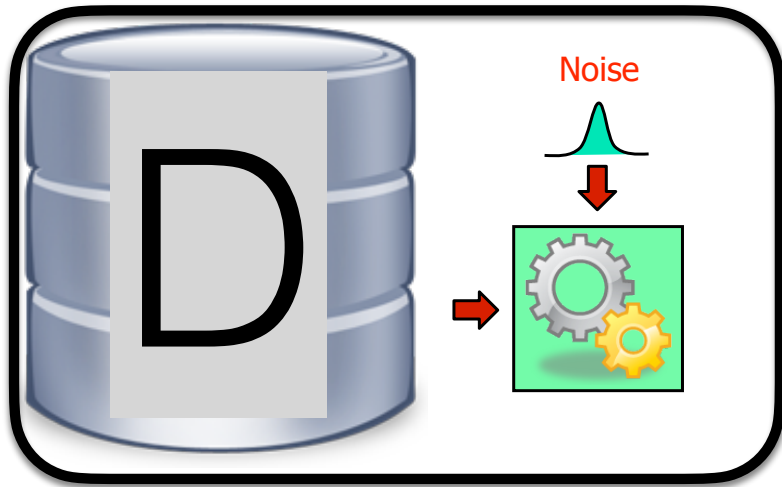
Reconstruction attack



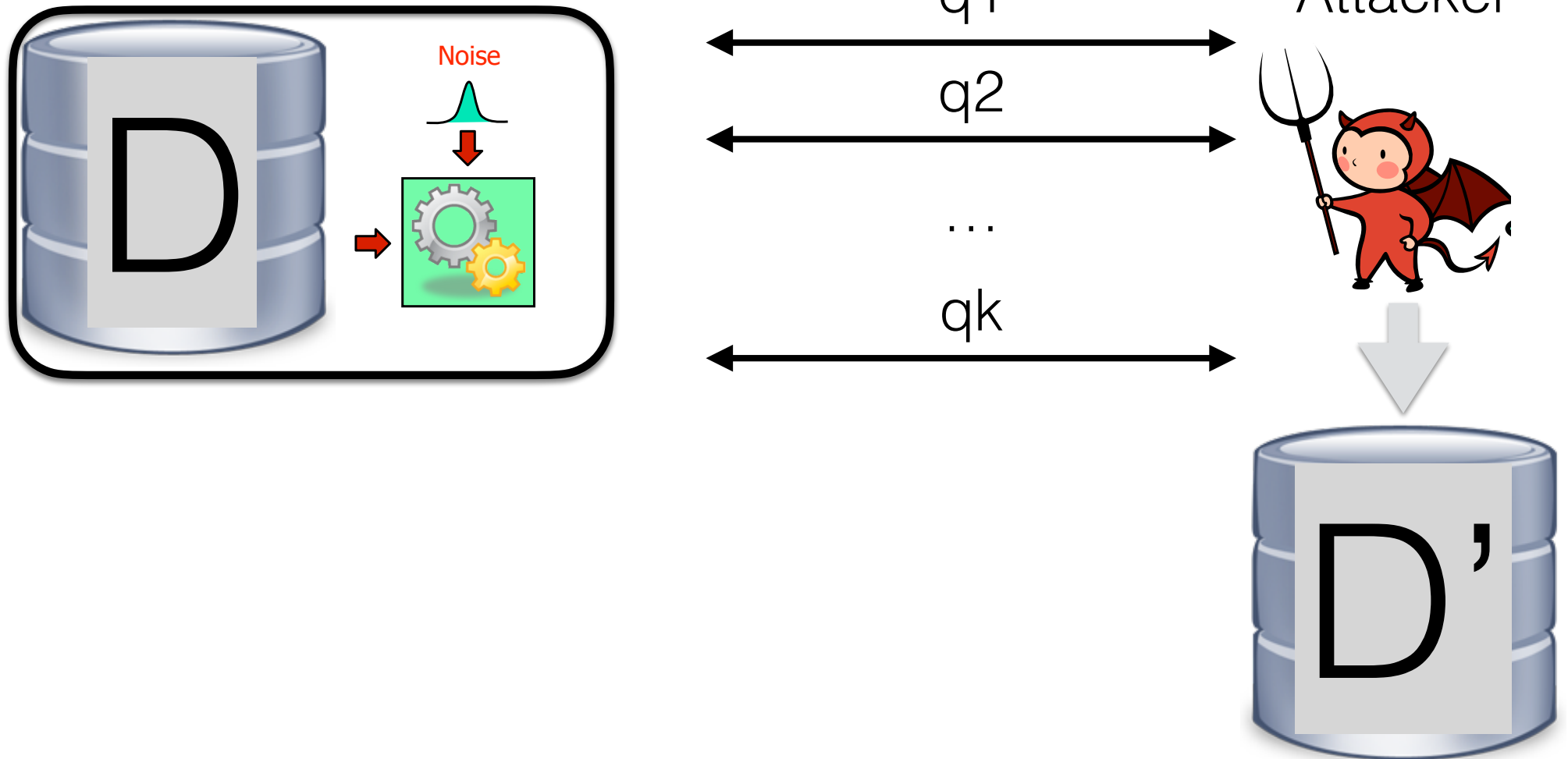
Reconstruction attack



Reconstruction attack



Reconstruction attack



Reconstruction attack



We say that the attacker **wins** if

$$d(\text{D}, \text{D}') \sim 0$$

In our case we can use Hamming distance

Additive Noise Perturbation

- We say that M is a privacy mechanism obtained by adding noise if for every query q , M creates a new randomized query:

$$q^*(D) = q(D) + Y$$

Additive Noise Perturbation

- We say that M is a privacy mechanism obtained by adding noise if for every query q , M creates a new randomized query:

$$q^*(D) = q(D) + Y$$

- We say that a mechanism M add noise within perturbation E iff for every q and every D :

$$|q^*(D) - q(D)| \leq E$$

Reconstruction attack with³⁷ exponential adversary

Let $M:\{0,1\}^n \rightarrow \mathbb{R}$ be a privacy mechanism adding noise within E perturbation. Then there is an adversary that can reconstruct the database within $4E$ positions.

Reconstruction attack with³⁸ exponential adversary

Let $M:\{0,1\}^n \rightarrow R$ be a privacy mechanism adding noise within $\epsilon=o(n)$ perturbation. Then there is an adversary that can reconstruct the database with constant error and running in exponential time.

Reconstruction attack with³⁹ polynomial adversary

Let $M:\{0,1\}^n \rightarrow R$ be a privacy mechanism adding noise within $\epsilon=o(\sqrt{n})$ perturbation. Then there is an adversary that can reconstruct the database with constant error running in polynomial time and **answering n queries**.

[DinurNissim'02, DworkYekhanin'08]

Number of queries

A privacy mechanism can answer with perturbation \sqrt{n} at most a number of queries sublinear in n .

Number of queries

A privacy mechanism can answer with perturbation \sqrt{n} at most a number of queries sublinear in n .

Question: Why error \sqrt{n} is a good reference?

Sample error

- Suppose that a database contains n individuals drawn uniformly at random from a population of size $N \gg n$.
- Suppose we are interested in a medical condition that affects a fraction p of the population.
- Then we expect the number of individuals in the dataset with condition p is
$$np \pm \Theta(\sqrt{n})$$
- The sampling error is of the order of \sqrt{n} .

Sample error

- Suppose that a database contains n individuals drawn uniformly at random from a population of size $N \gg n$.
- Suppose we are interested in a medical condition that affects a fraction p of the population.
- Then we expect the number of individuals in the dataset with condition p is
$$np \pm \Theta(\sqrt{n})$$
- The sampling error is of the order of \sqrt{n} .

We would like the noise we introduce for privacy to be comparable to the sampling error.

Fundamental Law of Information Reconstruction

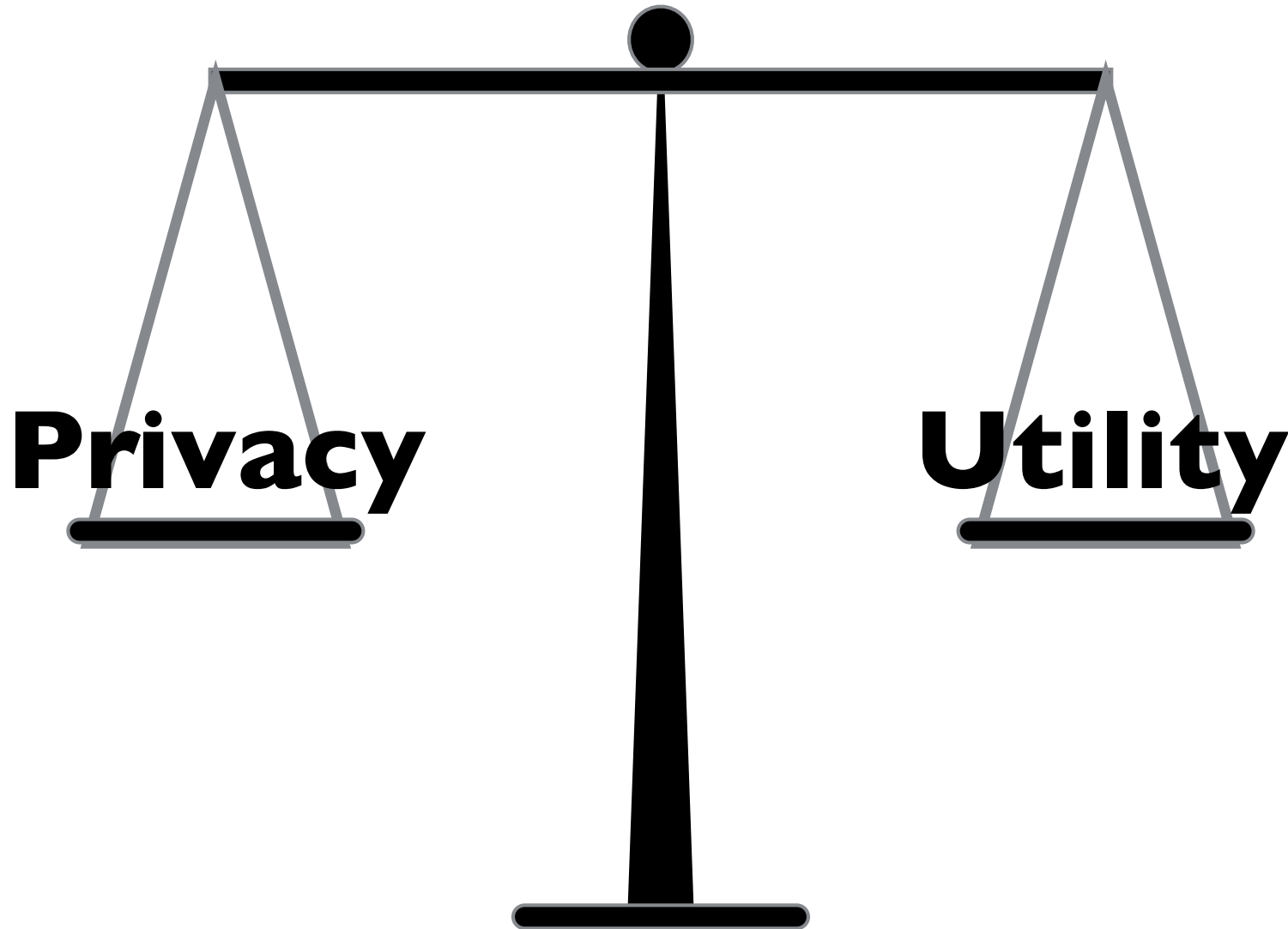
The release of **too many** overly **accurate** statistics gives privacy violations.



[DinurNissim02]

Privacy vs Utility

43



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.

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.

What can this notion be?

Let's take inspiration from semantics security.

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

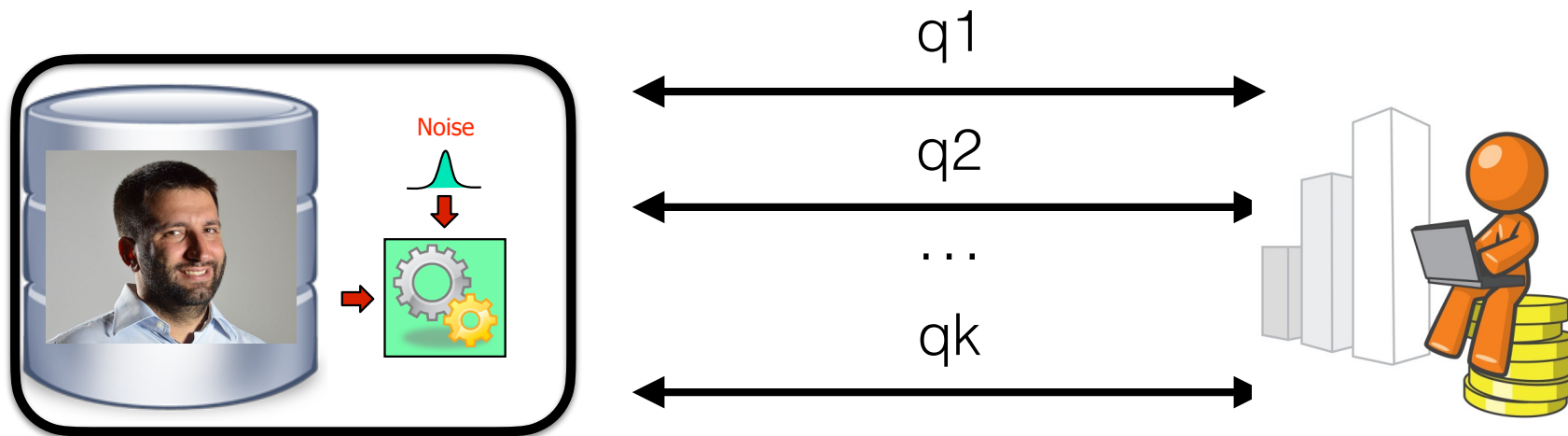
Let's take inspiration from semantics security.

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



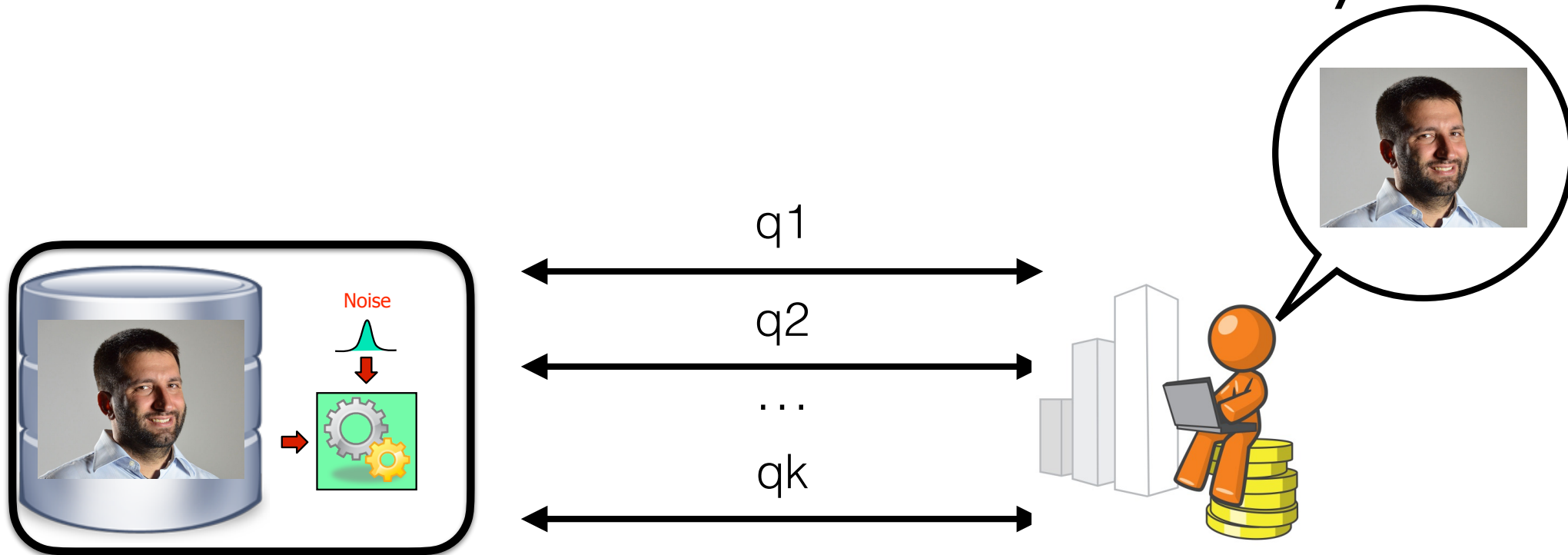
Let's take inspiration from semantics security.

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



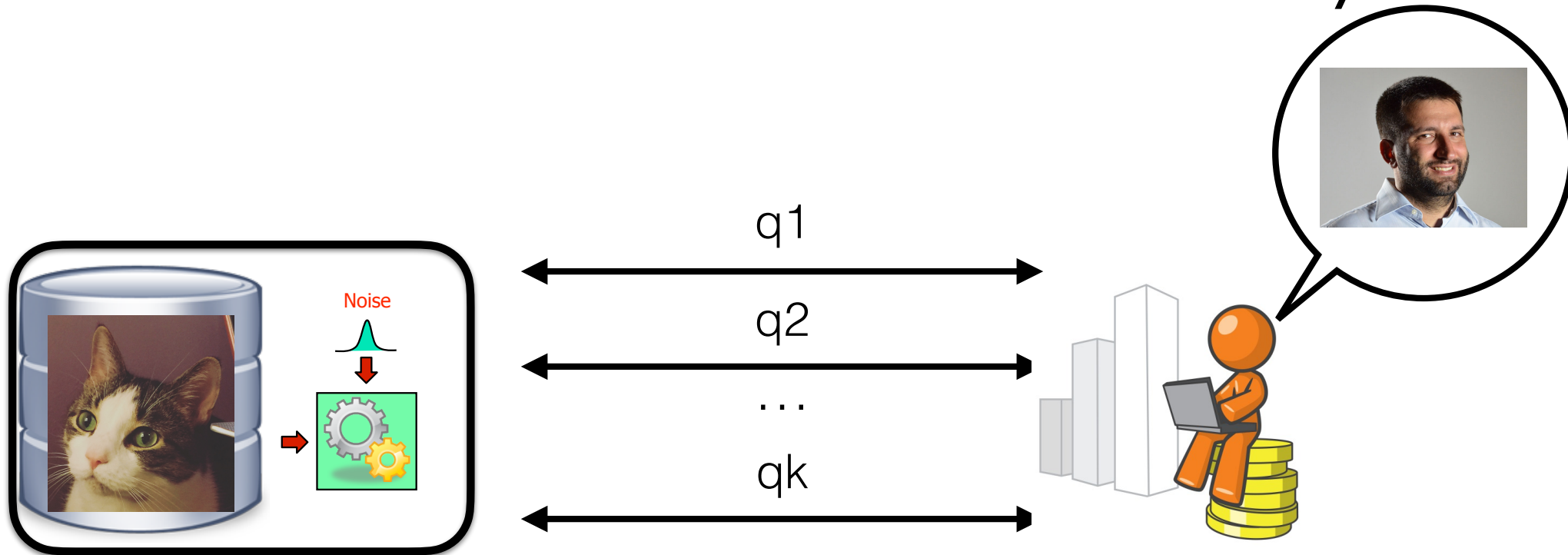
Let's take inspiration from semantics security.

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



Let's take inspiration from semantics security.

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



Privacy-preserving data analysis?

Prior Knowledge

\sim

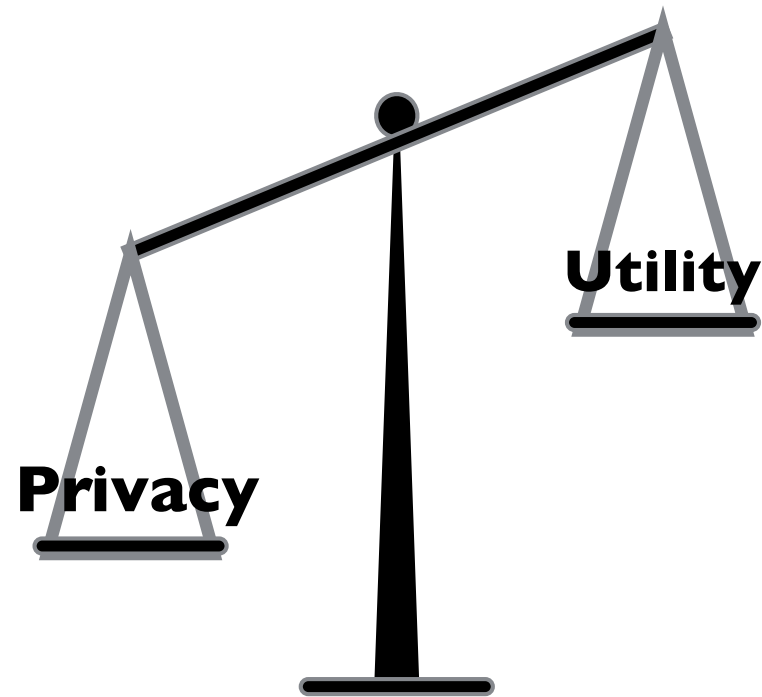
Posterior Knowledge

Privacy-preserving data analysis?

Privacy-preserving data analysis?

Question: What is the problem with this requirement?

Privacy-preserving data analysis?⁴⁸



If nothing can be learned about an individual, then nothing at all can be learned at all!

[DworkNaor10]

Let's take inspiration from semantics security v2.

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



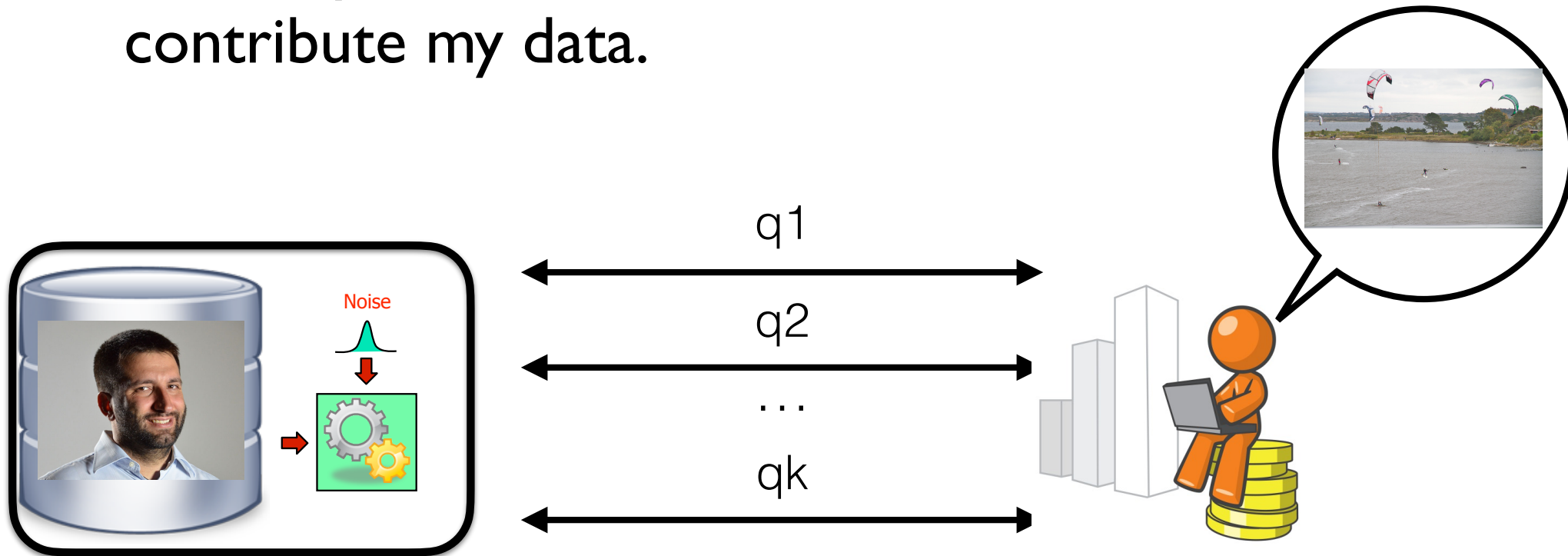
Let's take inspiration from semantics security v2.

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



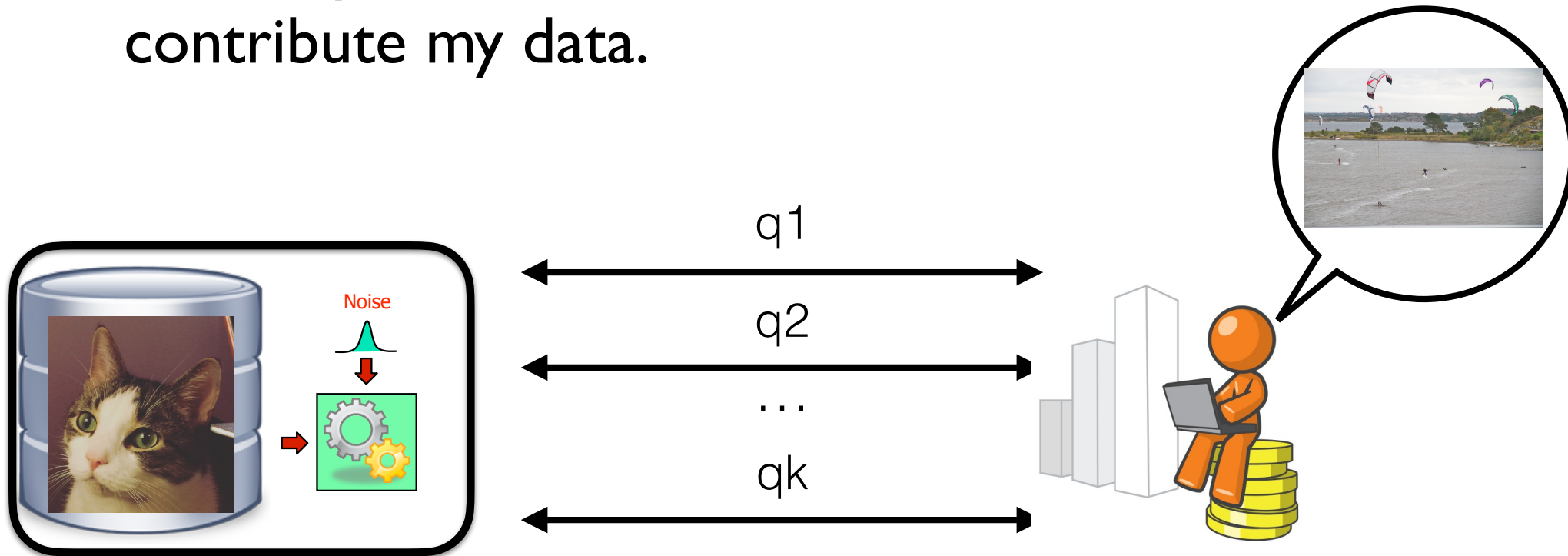
Let's take inspiration from semantics security v2.

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



Let's take inspiration from semantics security v2.

- 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 $D, D' \in \{0, 1\}^n$, 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 $D \sim D'$.

(ϵ, δ) -Differential Privacy

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

(ϵ, δ) -Differential Privacy

A query returning a probability distribution

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

(ϵ, δ) -Differential Privacy

Privacy parameters

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

(ϵ, δ) -Differential Privacy

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

a quantification over all
the databases

(ϵ, δ) -Differential Privacy

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

a notion of adjacency or distance

(ϵ, δ) -Differential Privacy

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

and over all the possible
outcomes

ϵ -Differential Privacy

Definition

Given $\epsilon \geq 0$, a probabilistic query $Q: X^n \rightarrow R$ is ϵ -differentially private iff for all adjacent database b_1, b_2 and for every $S \subseteq R$:

$$\Pr[Q(b_1) \in S] \leq \exp(\epsilon) \Pr[Q(b_2) \in S]$$

ϵ -Differential Privacy

Definition

Given $\epsilon \geq 0$, a probabilistic query $Q: X^n \rightarrow R$ is ϵ -differentially private iff for all adjacent database b_1, b_2 and for every $S \subseteq R$:

$$\Pr[Q(b_1) \in S] \leq \exp(\epsilon) \Pr[Q(b_2) \in S]$$

Let's substitute a concrete instance:

$$\Pr[Q(b \cup \{x\}) \in S] \leq \exp(\epsilon) \Pr[Q(b \cup \{y\}) \in S]$$

ϵ -Differential Privacy

Definition

Given $\epsilon \geq 0$, a probabilistic query $Q: X^n \rightarrow R$ is ϵ -differentially private iff for all adjacent database b_1, b_2 and for every $S \subseteq R$:

$$\Pr[Q(b_1) \in S] \leq \exp(\epsilon) \Pr[Q(b_2) \in S]$$

Let's substitute a concrete instance:

$$\Pr[Q(b \cup \{x\}) \in S] \leq \exp(\epsilon) \Pr[Q(b \cup \{y\}) \in S]$$

Let's use the two quantifiers:

$$\exp(-\epsilon) \Pr[Q(b \cup \{y\}) \in S] \leq \Pr[Q(b \cup \{x\}) \in S] \leq \exp(\epsilon) \Pr[Q(b \cup \{y\}) \in S]$$

ϵ -Differential Privacy

Definition

Given $\epsilon \geq 0$, a probabilistic query $Q: X^n \rightarrow R$ is ϵ -differentially private iff for all adjacent database b_1, b_2 and for every $S \subseteq R$:

$$\Pr[Q(b_1) \in S] \leq \exp(\epsilon) \Pr[Q(b_2) \in S]$$

Let's substitute a concrete instance:

$$\Pr[Q(b \cup \{x\}) \in S] \leq \exp(\epsilon) \Pr[Q(b \cup \{y\}) \in S]$$

Let's use the two quantifiers:

$$\exp(-\epsilon) \Pr[Q(b \cup \{y\}) \in S] \leq \Pr[Q(b \cup \{x\}) \in S] \leq \exp(\epsilon) \Pr[Q(b \cup \{y\}) \in S]$$

And for $\epsilon \rightarrow 0$

$$(1 - \epsilon) \Pr[Q(b \cup \{y\}) \in S] \leq \Pr[Q(b \cup \{x\}) \in S] \leq (1 + \epsilon) \Pr[Q(b \cup \{y\}) \in S]$$

ϵ -Differential Privacy

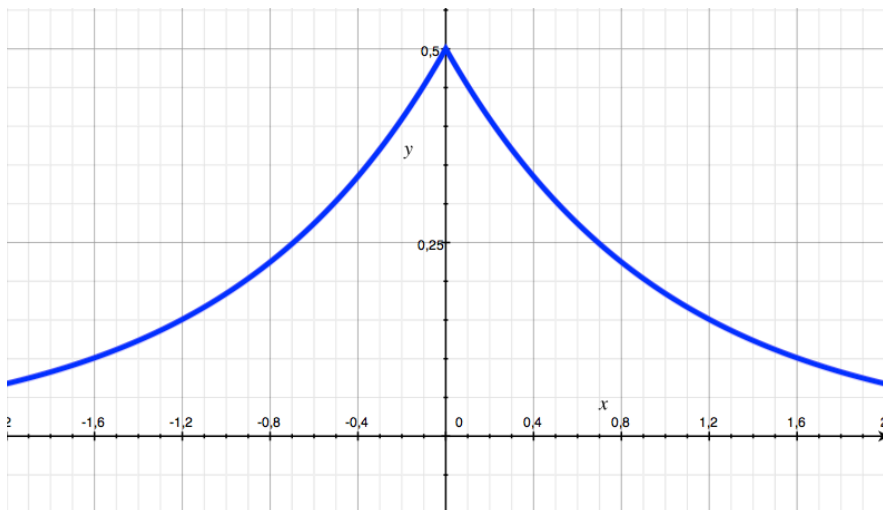
In general we can think about the following quantity as the **privacy loss** incurred by observing **r** on the databases **b** and **b'**.

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

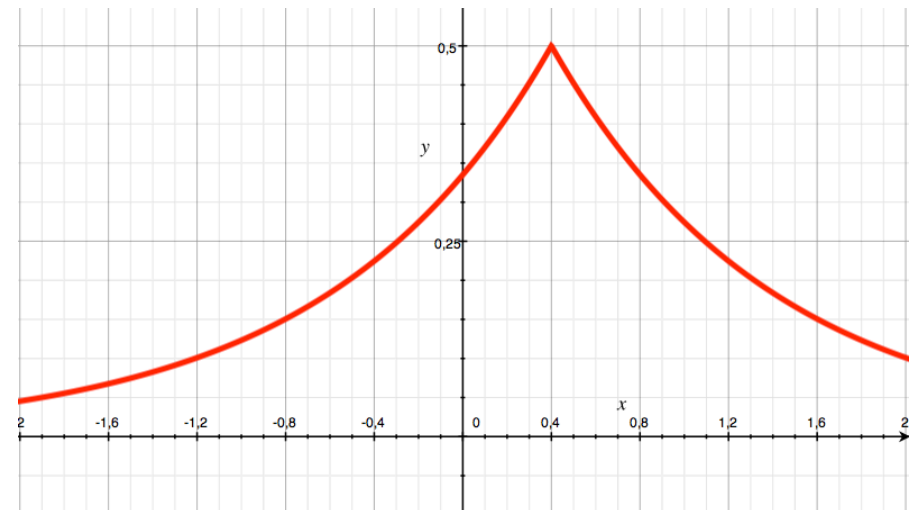
ϵ -Differential Privacy

$Q : db \Rightarrow R$ probabilistic

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

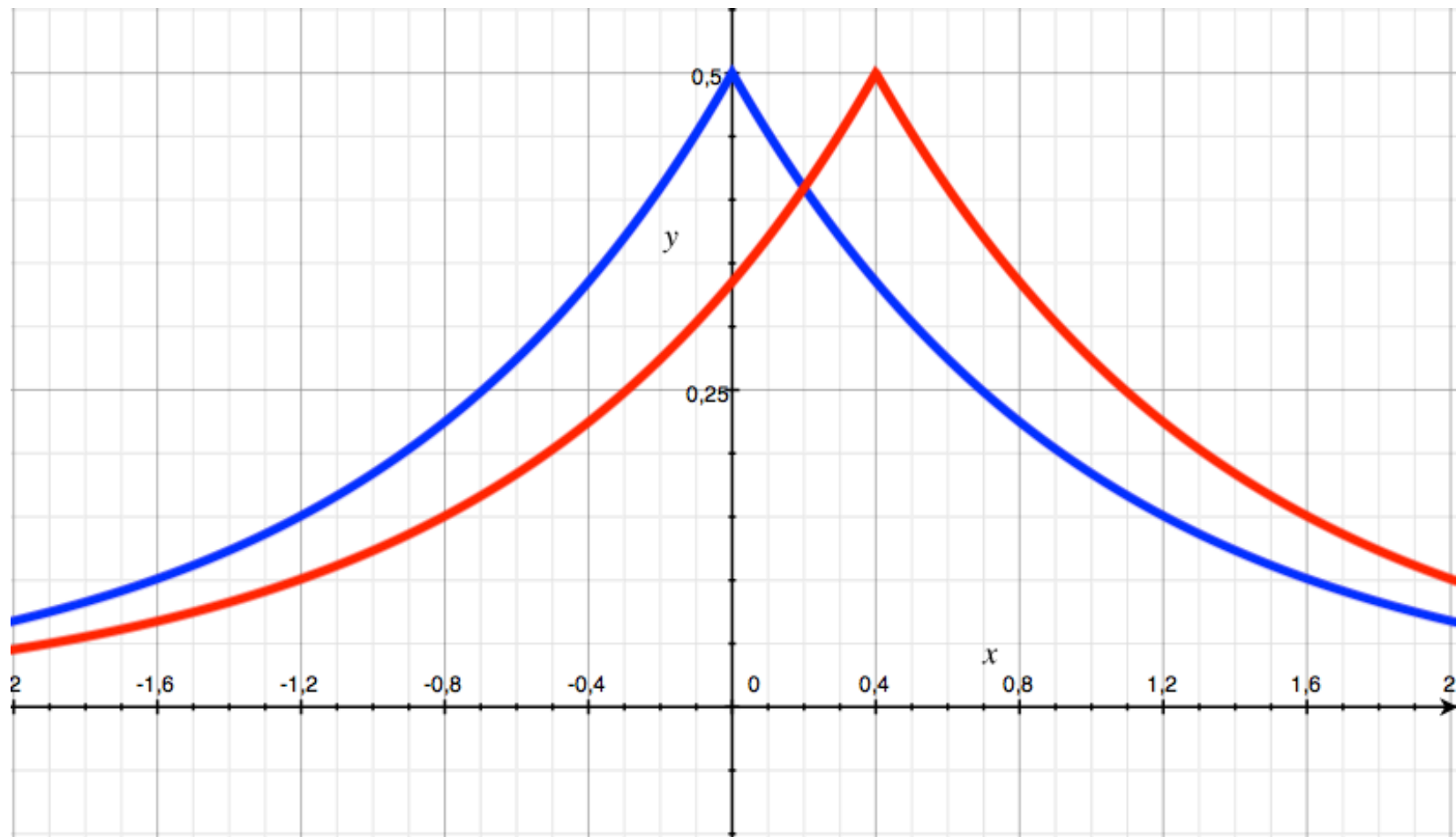


$Q(b \cup \{y\})$

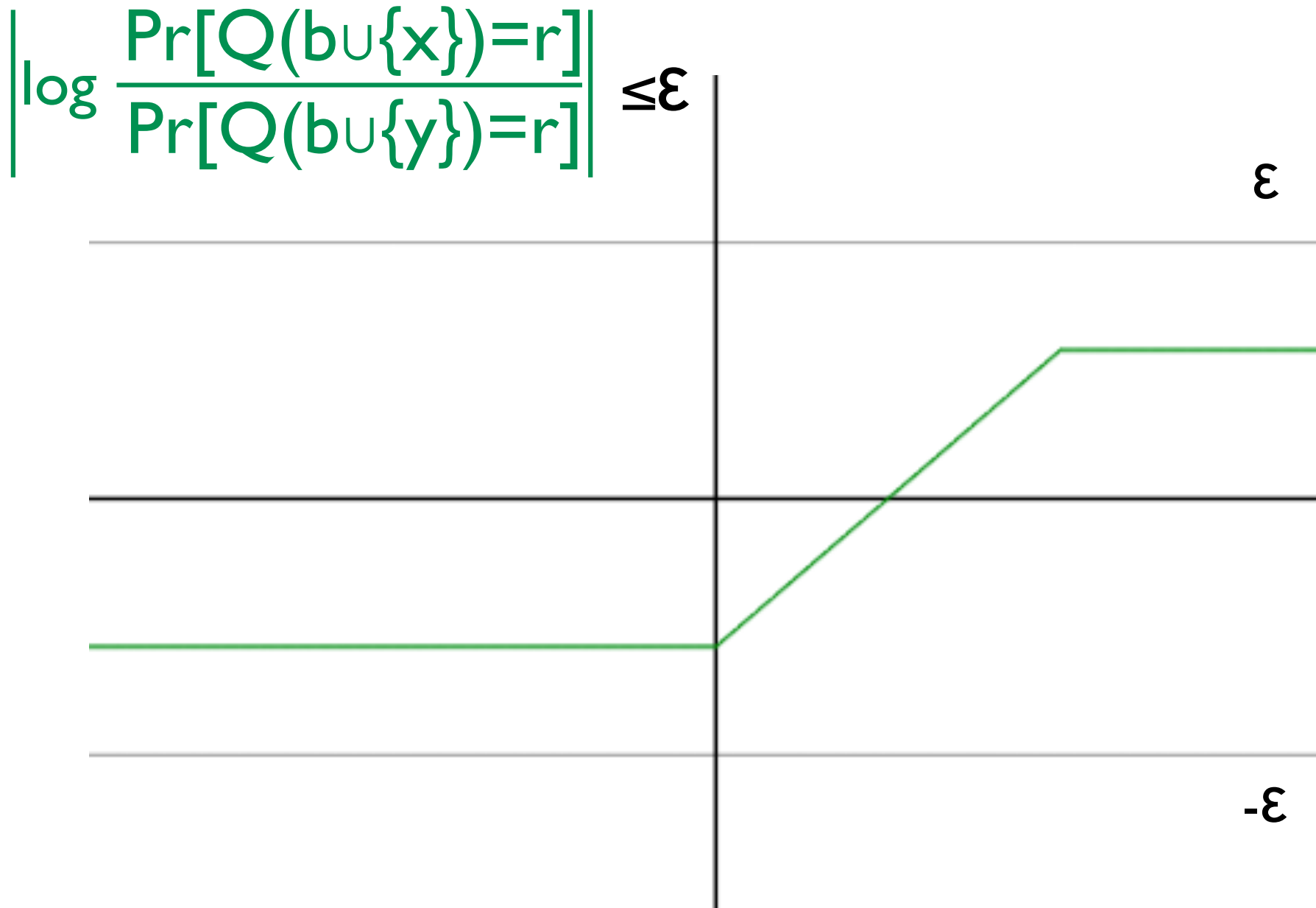


ϵ -Differential Privacy

$$d(Q(\mathbf{b}_U\{x\}), Q(\mathbf{b}_U\{y\})) \leq \epsilon$$



ϵ -Differential Privacy



(ϵ, δ) -Differential Privacy

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

(ϵ, δ) -Differential Privacy

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

Similarly, we have

$$\log \frac{\Pr[Q(b_1) \in S] - \delta}{\Pr[Q(b_2) \in S]} \leq \epsilon$$

$$-\epsilon \leq \log \frac{\Pr[Q(b_1) \in S] + \delta}{\Pr[Q(b_2) \in S]}$$

(ϵ, δ) -Differential Privacy

Definition

Given $\epsilon, \delta \geq 0$, 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 $S \subseteq R$:

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

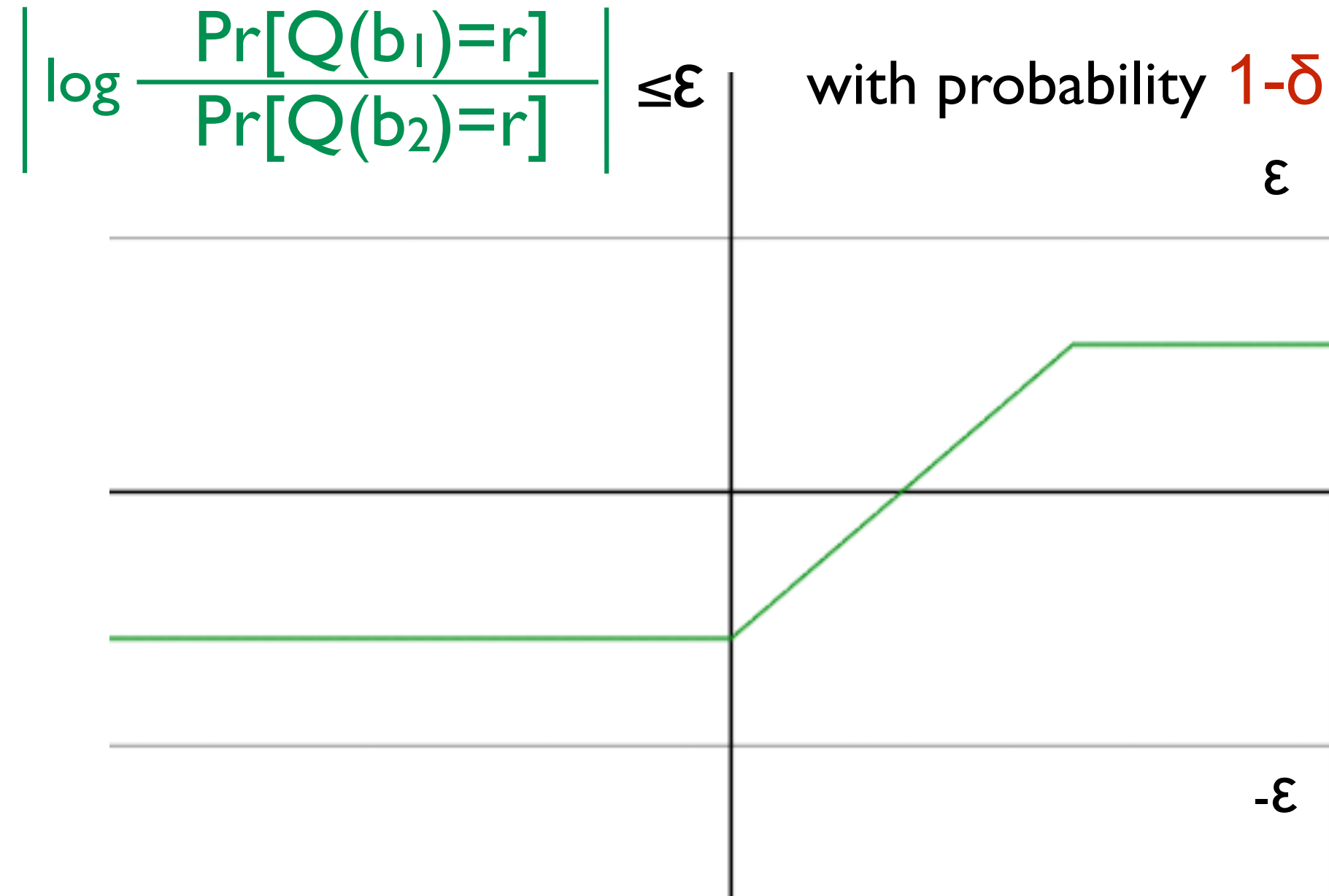
Similarly, we have

$$\log \frac{\Pr[Q(b_1) \in S] - \delta}{\Pr[Q(b_2) \in S]} \leq \epsilon$$

$$-\epsilon \leq \log \frac{\Pr[Q(b_1) \in S] + \delta}{\Pr[Q(b_2) \in S]}$$

Probability of failure

(ϵ, δ) -Differential Privacy



The rest of the class

- Understanding some basic methods to guarantee differential privacy and how they provide an answer for the privacy vs utility trade-off.
- Looking at how we can formally support differential privacy using EasyCrypt.

Summary

- Statistical queries and databases,
- Additive noise perturbation,
- Reconstruction attack,
- Fundamental Law of Information Reconstruction,
- Differential privacy

Reconstruction attack with⁷¹ exponential adversary

Let $M:\{0,1\}^n \rightarrow R$ be a privacy mechanism adding noise within E perturbation. Then there is an adversary that can reconstruct the database within $4E$ positions.

Proof

Query phase: For each $S \subseteq [n]$ let $a_S^* = q_S^*(D)$.

Rule out phase: For each $D' \in \{0,1\}^n$:
if there exists S such that $|q_S(D') - a_S^*| > E$ then rule out D' .

Output phase: Output a database D' that was not ruled out.

Proof

Query phase: For each $S \subseteq [n]$ let $a_S^* = q_S^*(D)$.

Rule out phase: For each $D' \in \{0,1\}^n$:
if there exists S such that $|q_S(D') - a_S^*| > E$ then rule out D' .

Output phase: Output a database D' that was not ruled out.

Notice that since for the real database we clearly have

$$|q_S(D) - q_S^*(D)| \leq E$$

the procedure clearly return a candidate output in an exponential number of steps.

Proof

Query phase: For each $S \subseteq [n]$ let $a_S^* = q_S^*(D)$.

Rule out phase: For each $D' \in \{0,1\}^n$:
if there exists S such that $|q_S(D') - a_S^*| > E$ then rule out D' .

Output phase: Output a database D' that was not ruled out.

Notice that since for the real database we clearly have

$$|q_S(D) - q_S^*(D)| \leq E$$

the procedure clearly return a candidate output in an exponential number of steps.

We now want to show that $d_H(D, D') \leq 4E$

Proof

Let 's consider D to be the real dataset and D' to be the outputted one. Consider the sets of indices

$$R = \{ i \mid D(i)=0 \} \quad \text{and} \quad T = \{ i \mid D(i)=1 \}$$

Proof

Let 's consider D to be the real dataset and D' to be the outputted one. Consider the sets of indices

$$R = \{ i \mid D(i)=0 \} \quad \text{and} \quad T = \{ i \mid D(i)=1 \}$$

Since D' was not ruled out we have

$$|q_s^*(D) - q_s(D')| \leq E$$

but by definition we also have

$$|q_s^*(D) - q_s(D)| \leq E$$

so by triangle inequality $|q_s(D) - q_s(D')| \leq 2E$.

Since $q_R(D)=0$, we have that on the indices in R the Hamming distance between D and D' is at most $2E$.

Proof

Let 's consider D to be the real dataset and D' to be the outputted one. Consider the sets of indices

$$R = \{ i \mid D(i)=0 \} \quad \text{and} \quad T = \{ i \mid D(i)=1 \}$$

Since D' was not ruled out we have

$$|q_s^*(D) - q_s(D')| \leq E$$

but by definition we also have

$$|q_s^*(D) - q_s(D)| \leq E$$

so by triangle inequality $|q_s(D) - q_s(D')| \leq 2E$.

Since $q_R(D)=0$, we have that on the indices in R the Hamming distance between D and D' is at most $2E$.

We can apply a similar reasoning to T . So overall D and D' differ in at most $4E$ positions.