CGM: An Ehanced Mechanism for Streaming Data Collection with Local Differential Privacy (LDP)

Ergute BAO, Yin YANG, Xiaokui XIAO, Bolin DING

(submitted to VLDB 2021)

Motivation for LDP

collects data from individuals

...

Daily number of visits to the merchant website

...

The percentage of time the taxi stays in a certain area within 30 minutes.

collects data from individuals

Motivation for LDP

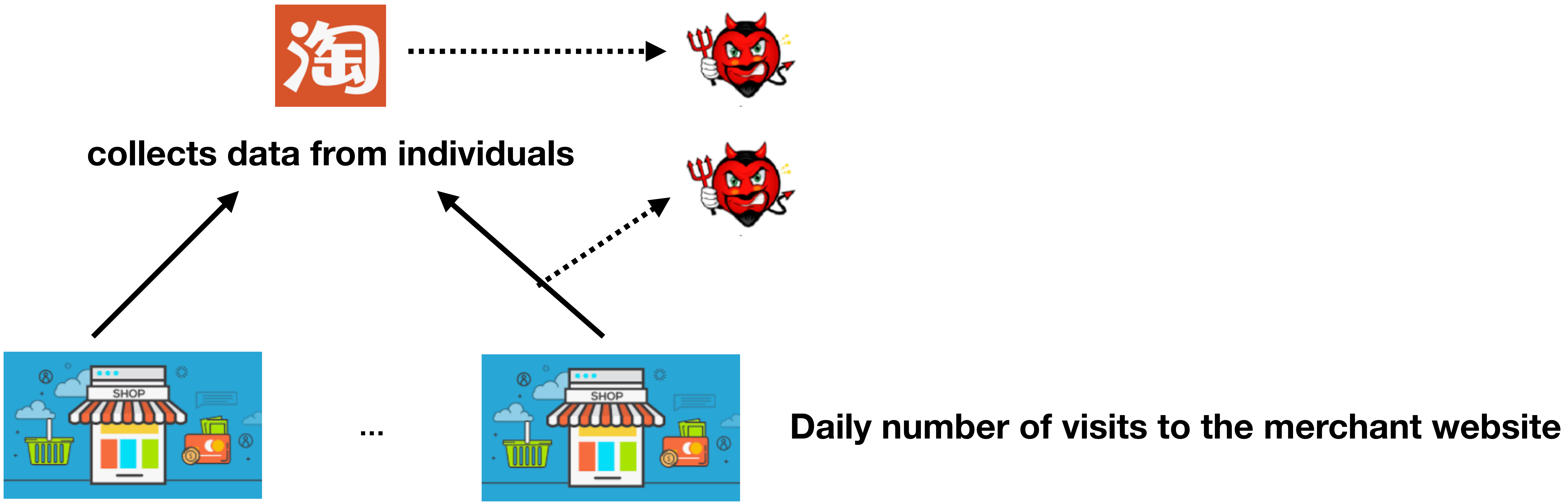
collects data from individuals

...

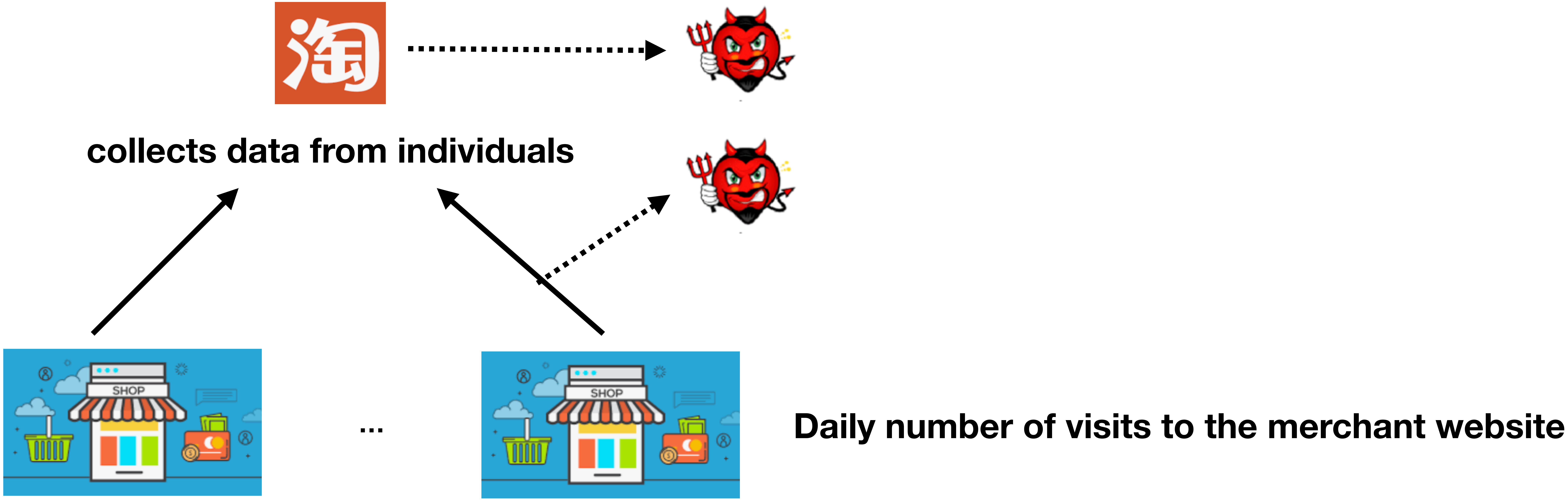
Daily number of visits to the merchant website

The aggregator wants the data for e.g. traffic monitoring, setting up business plans, etc..

Motivation for LDP

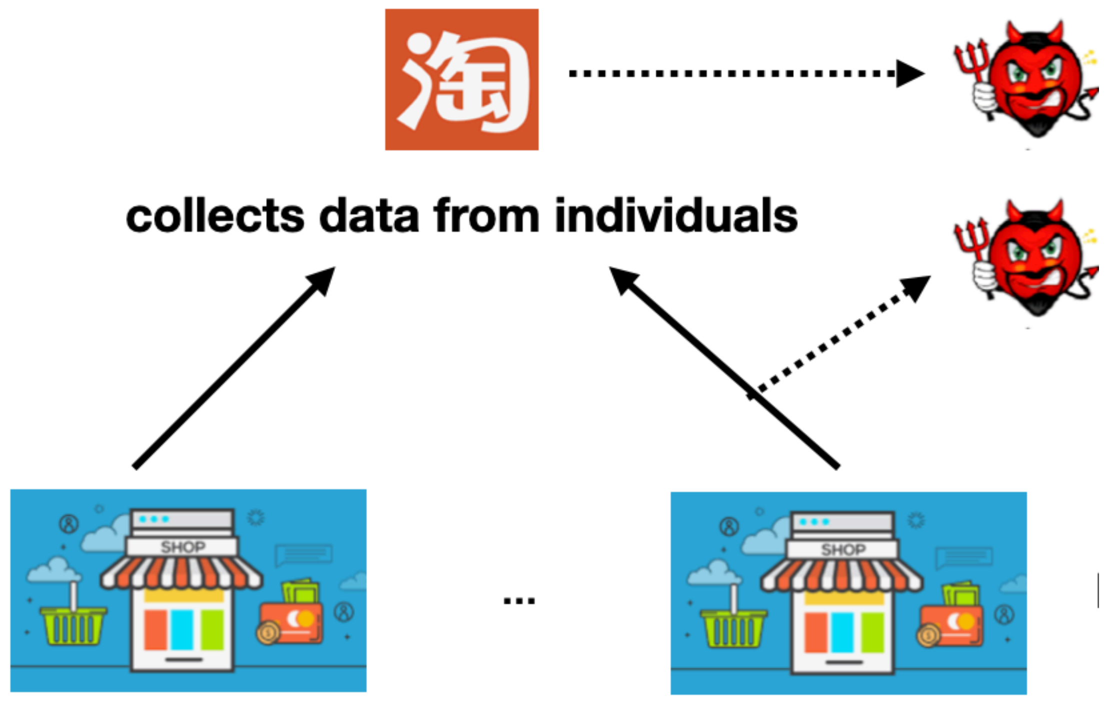


Motivation for LDP



The individual does not trust the communication channel, or the aggregator. Sending sensitive data directly leaks privacy. E.g., the daily number of visits is an indicator for the merchant's revenue.

Motivation for LDP

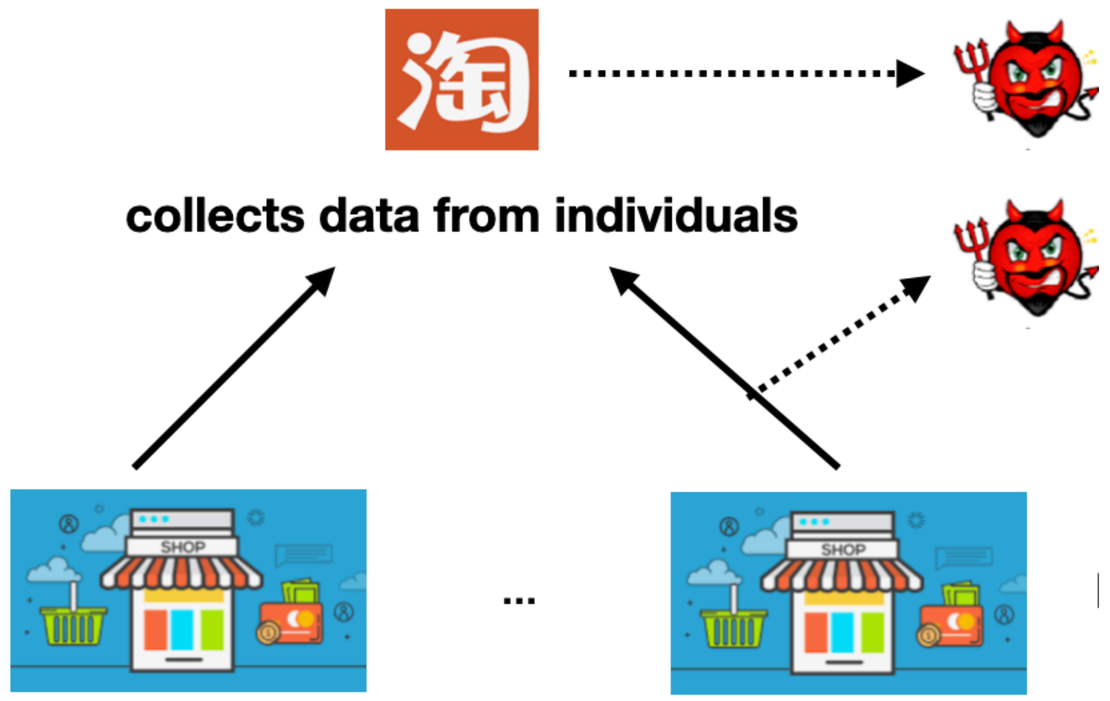


Each individual sends her data in a way that:

A. From the data of an individual student, the aggregator can not learn the exact information.

B. Upon the collection of all answers (e.g., hundreds of thousands of merchants), the aggregator can get a good estimate for the whole population, e.g., average daily number of visits.

Motivation for LDP



Each individual sends her data in a way that:

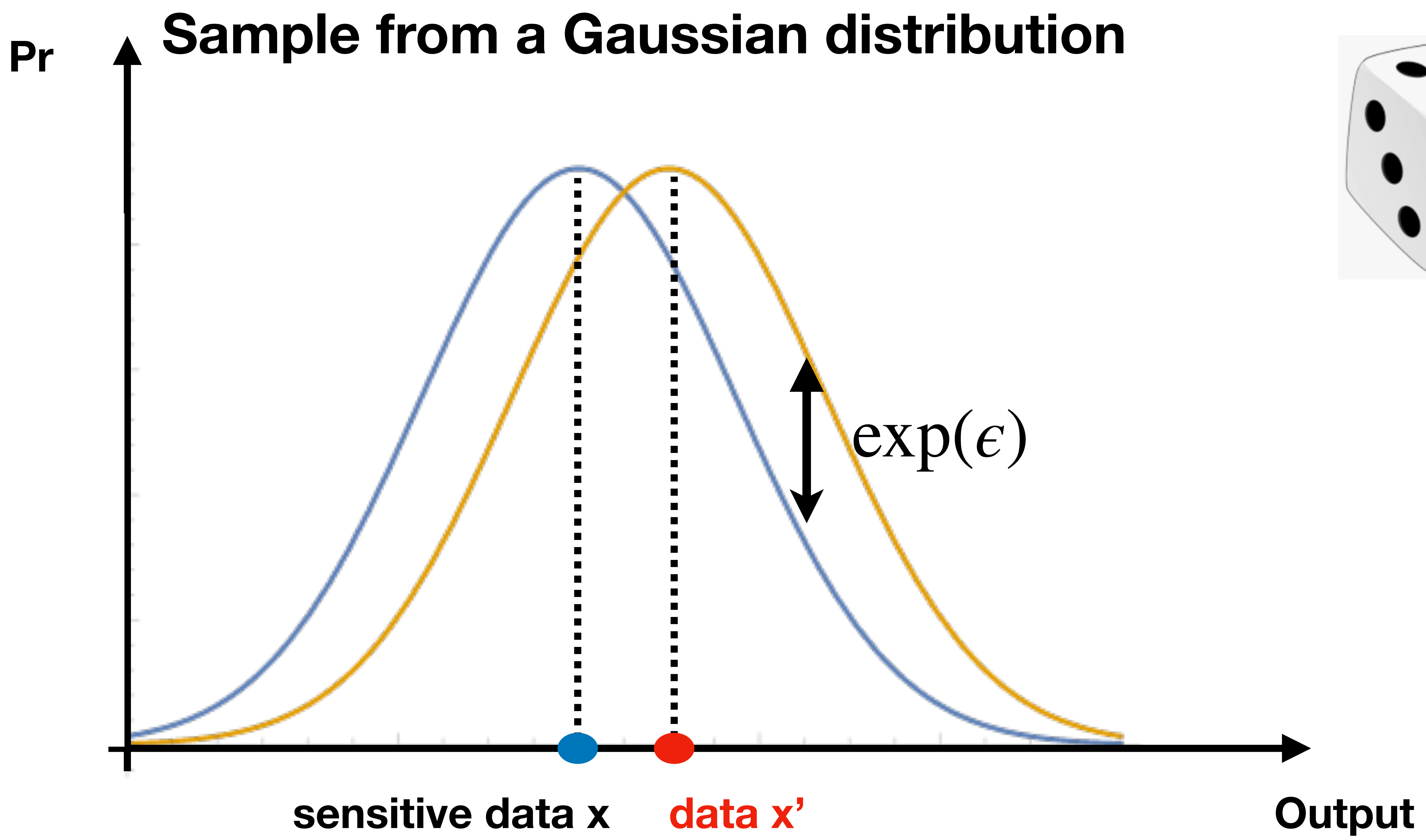
A. From the data of an individual student, the aggregator can not learn the exact information.

B. Upon the collection of all answers (e.g., hundreds of thousands of merchants), the aggregator can get a good estimate for the whole population, e.g., average daily number of visits.

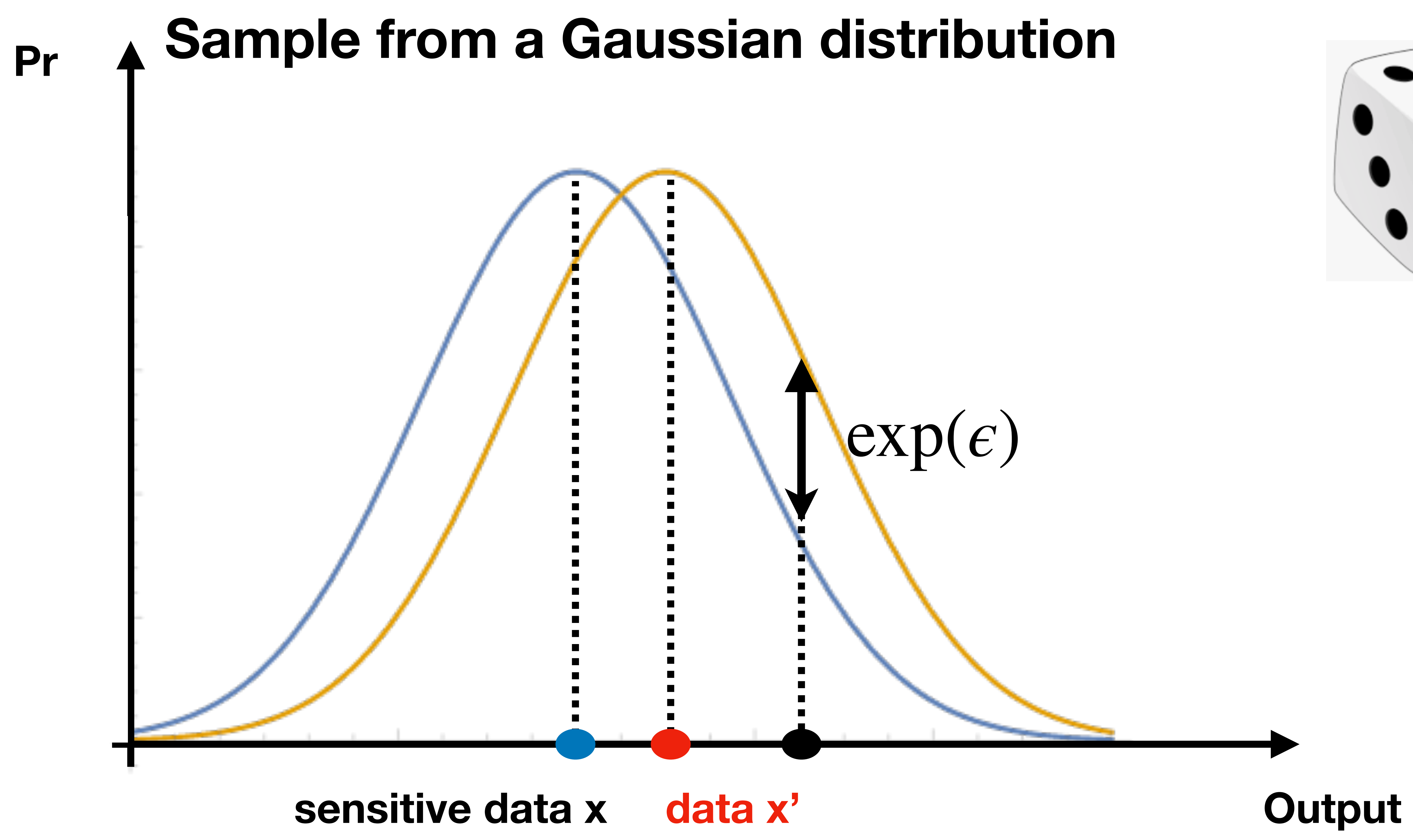
A. ensures the privacy for every individual.

B. ensures the accuracy of the statistic.

LDP Mechanism [DMNS '06]

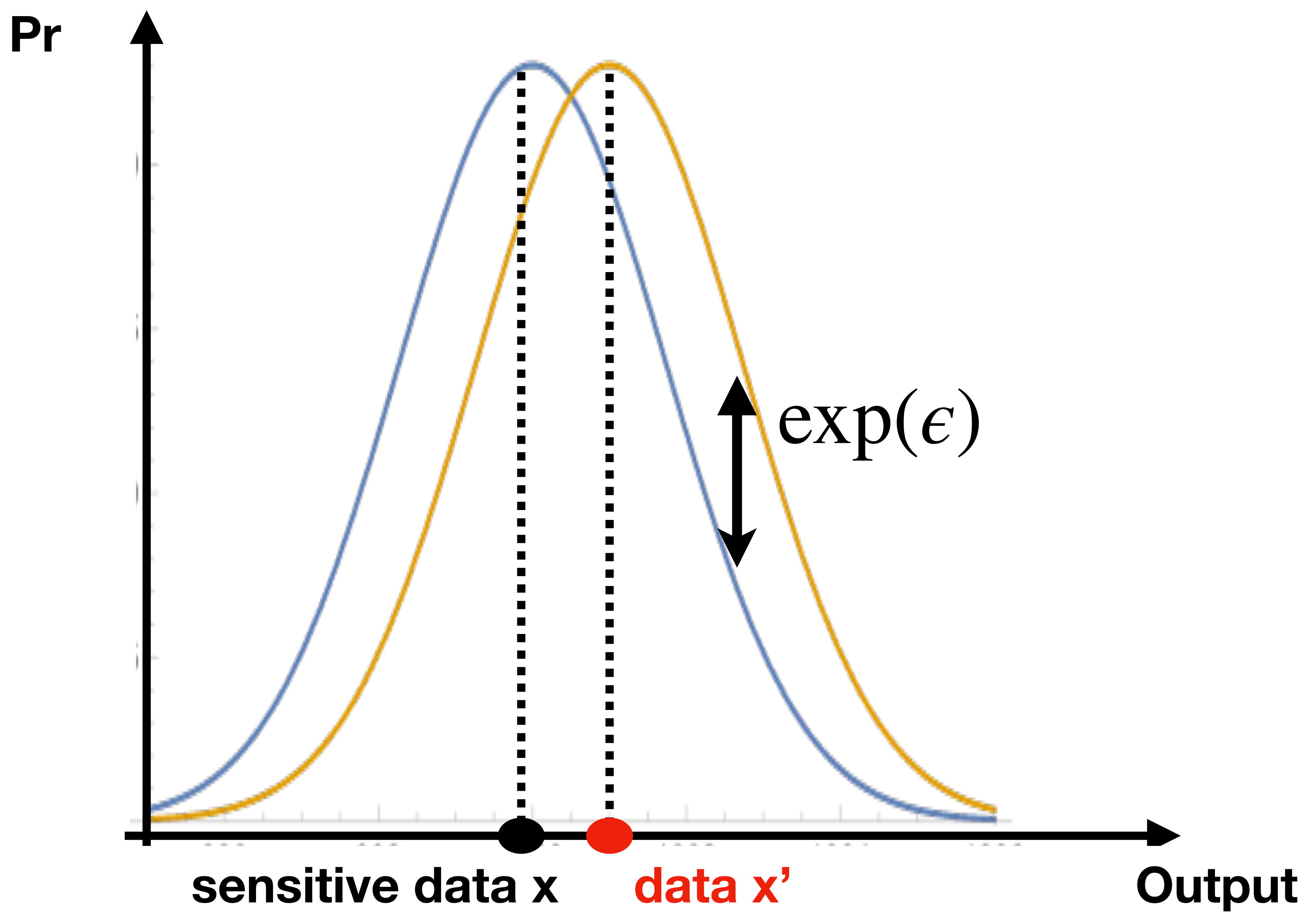


LDP Mechanism

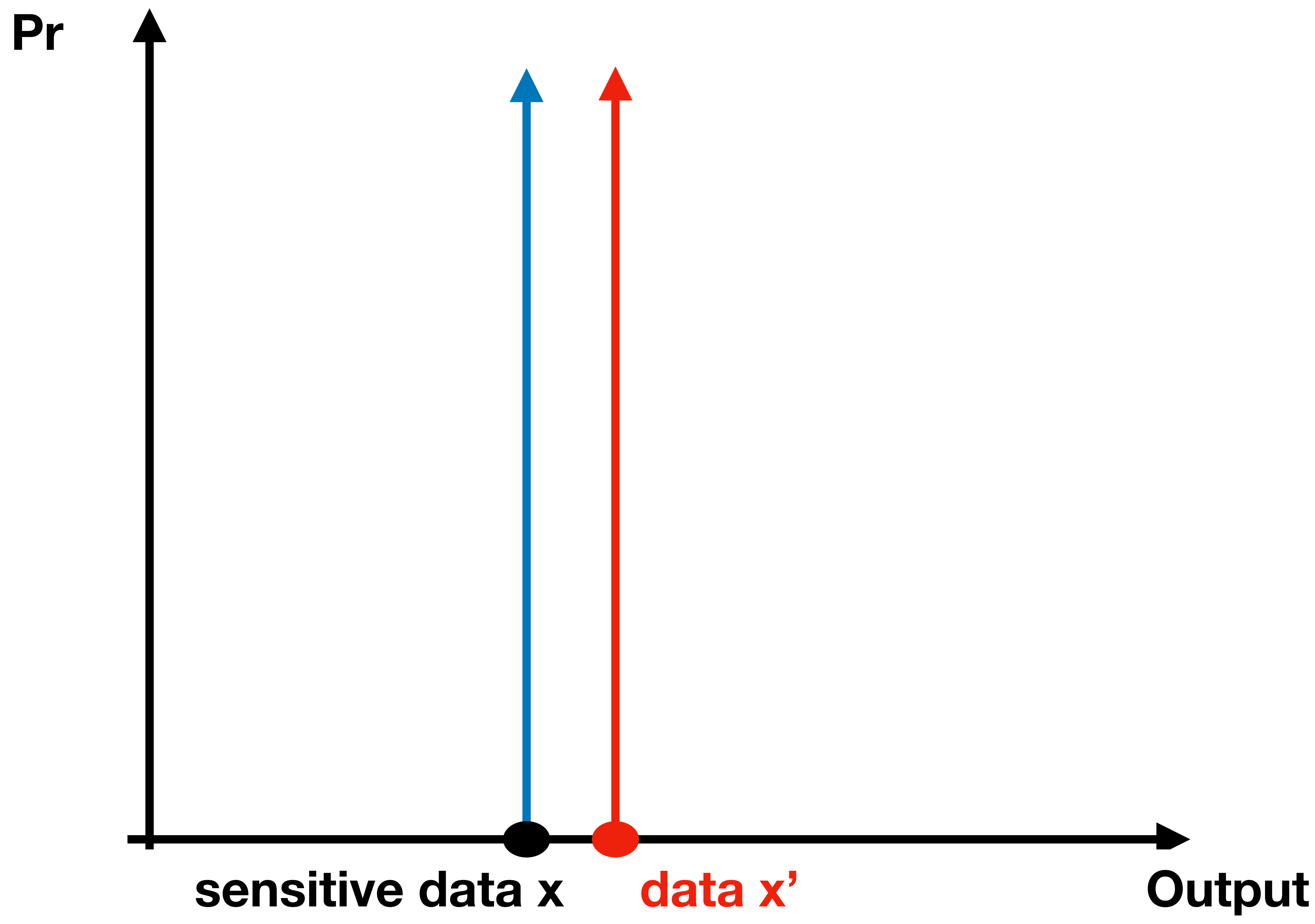


LDP Mechanism

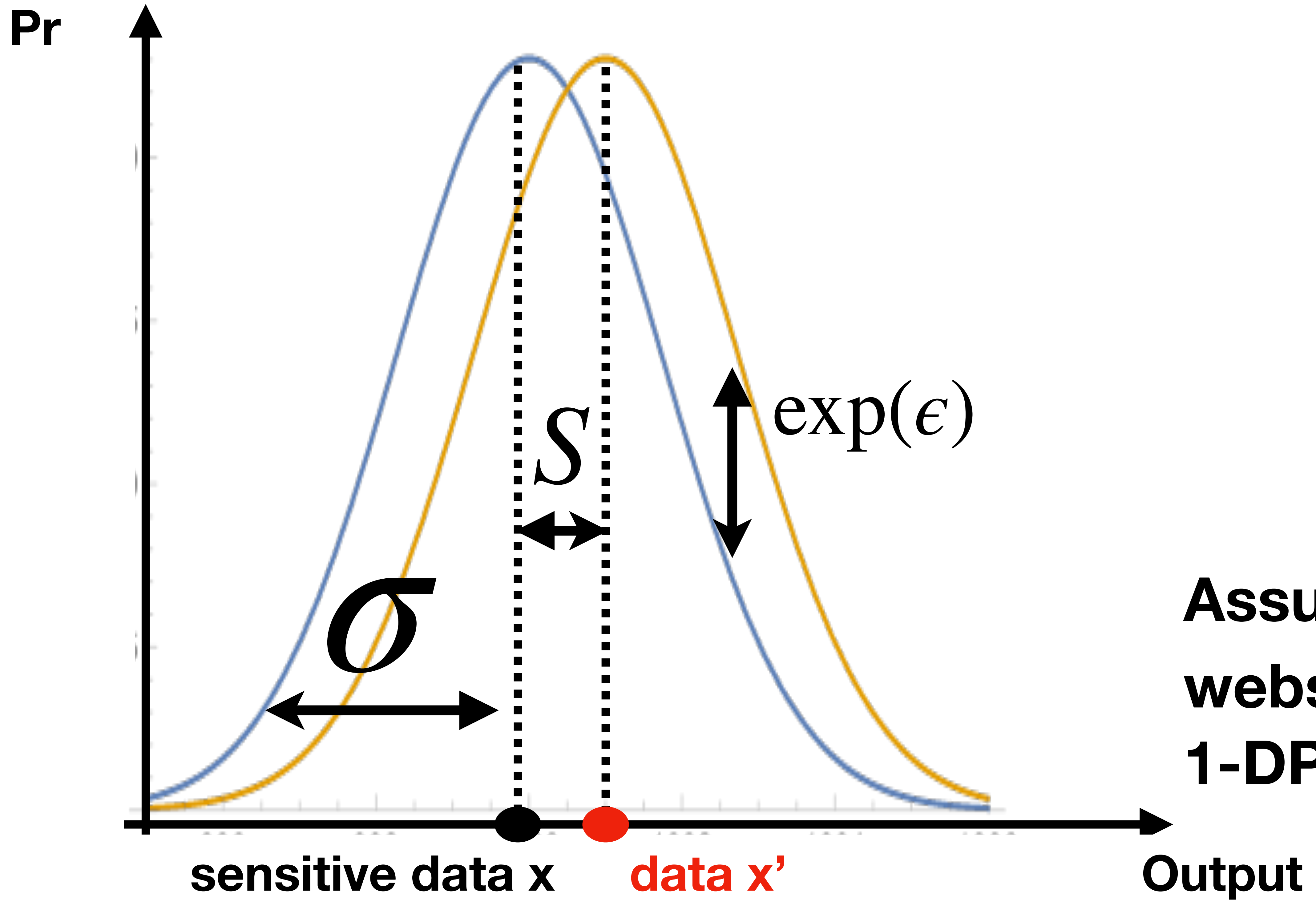
Smaller ϵ means stronger privacy



This mechanism ensures no privacy



LDP Mechanism



Utility

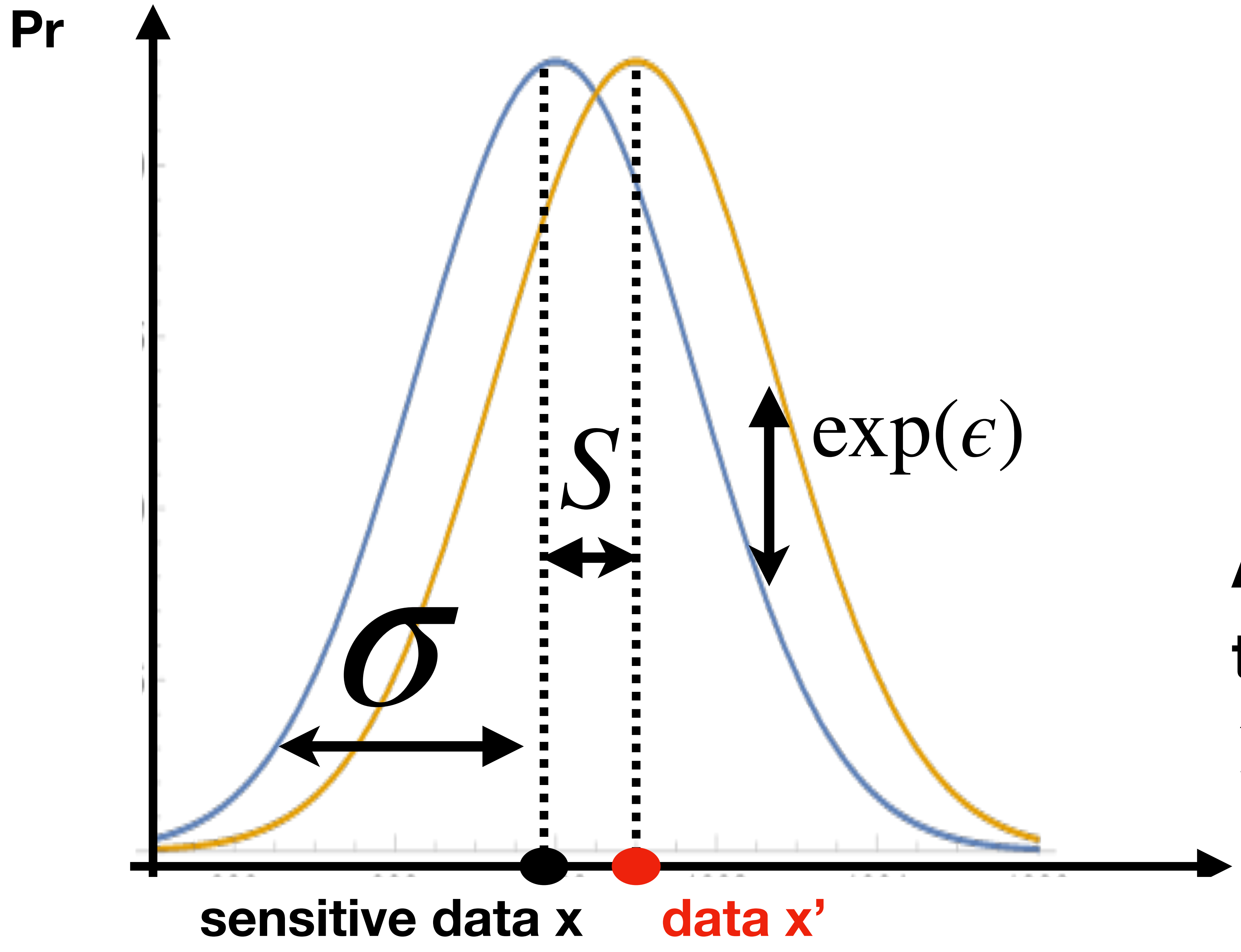
$$\sigma \approx \frac{S}{\epsilon}$$

Sensitivity

Privacy

Assuming the maximum number of daily visits to a merchant website is 10000, adding Gaussian noise of scale 10000 ensures 1-DP.

LDP Mechanism



Utility

$$\sigma \approx \frac{S}{\epsilon}$$

Sensitivity

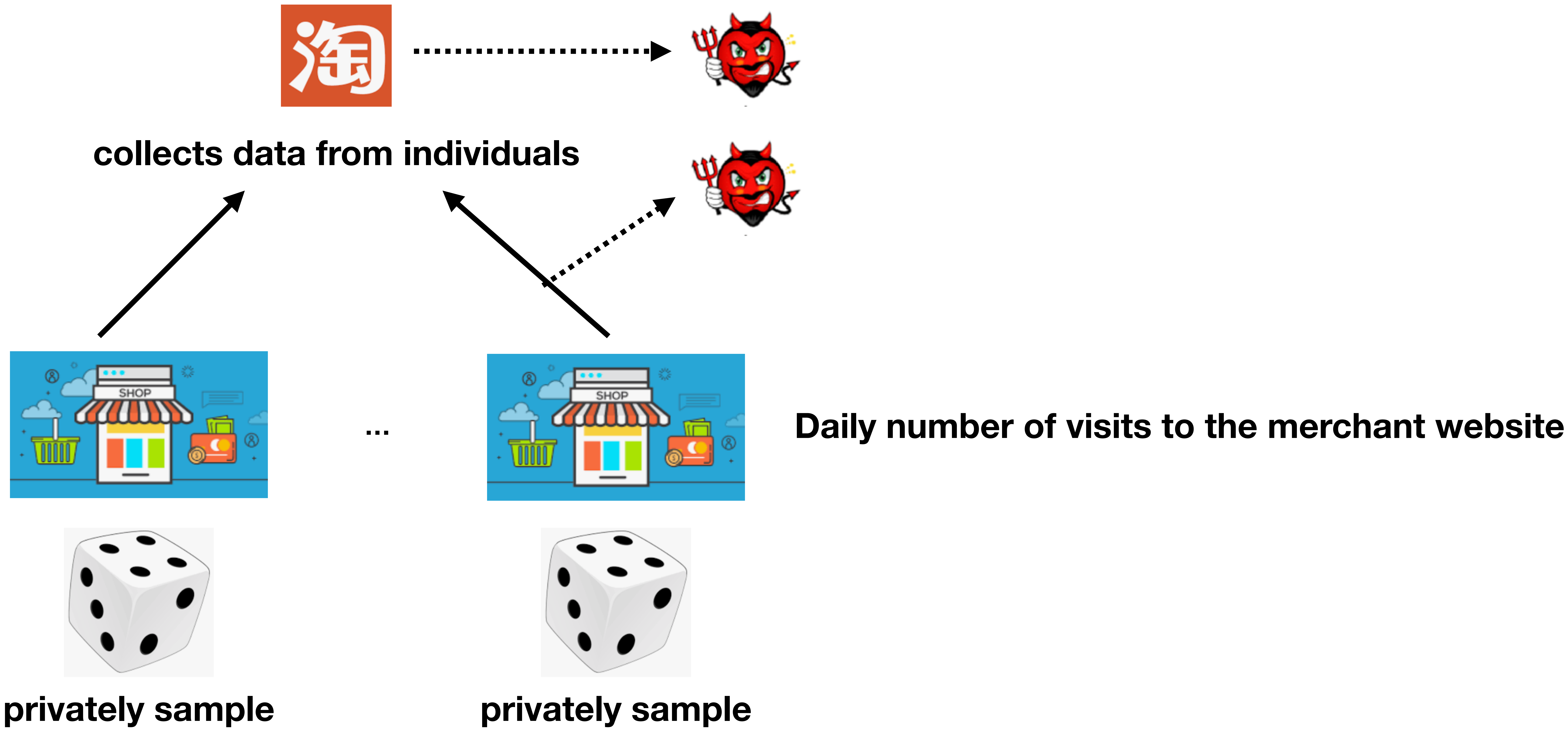
Privacy

Assuming the maximum number of daily visits (i.e., the sensitivity) to a merchant website is 10000, adding Gaussian noise of scale 10000 ensures 1-DP.

Is the noise too large?

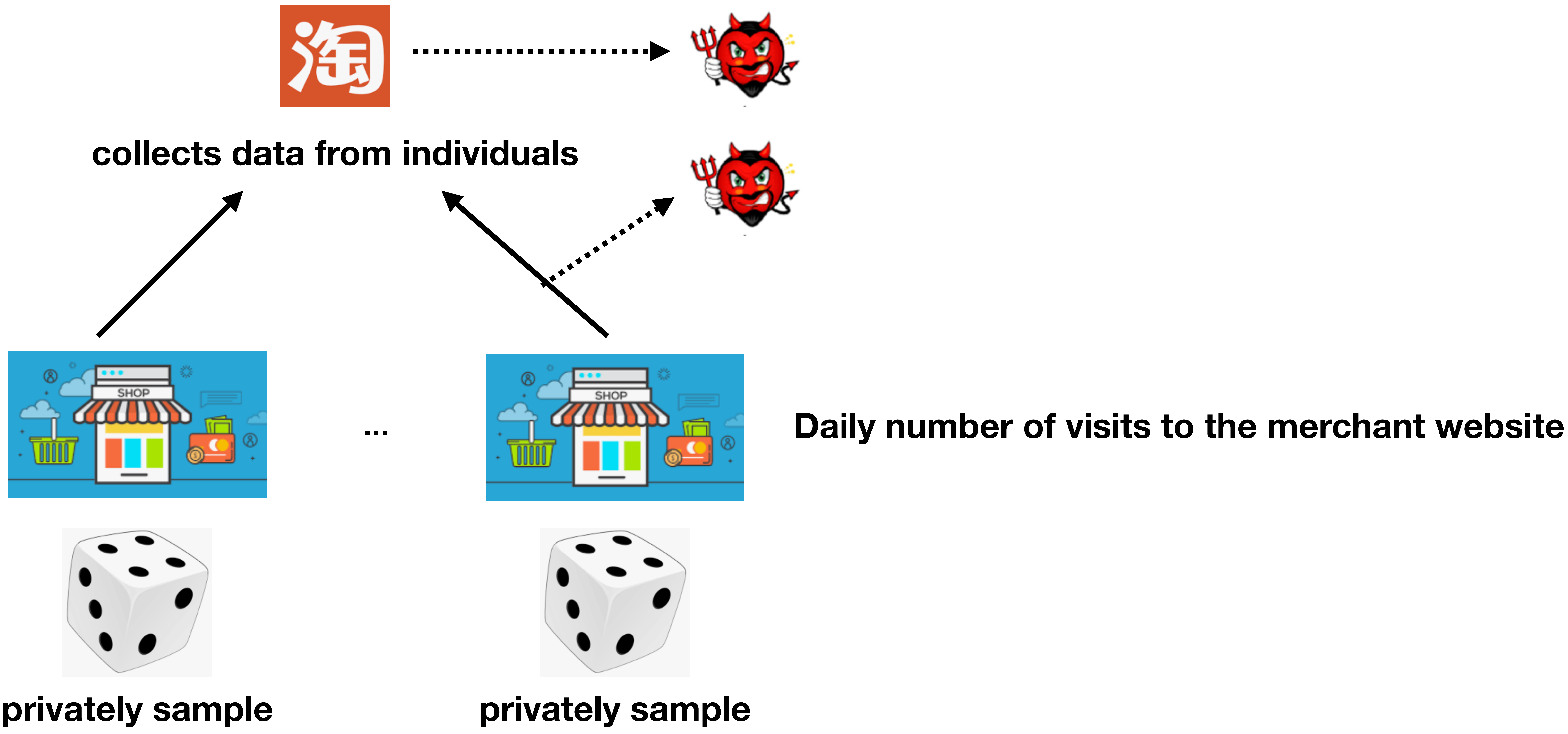
No. If there are 10^4 merchants, then on average the noise is only 100, introducing only 1% error to the average number of daily visits of all merchants.

Recap

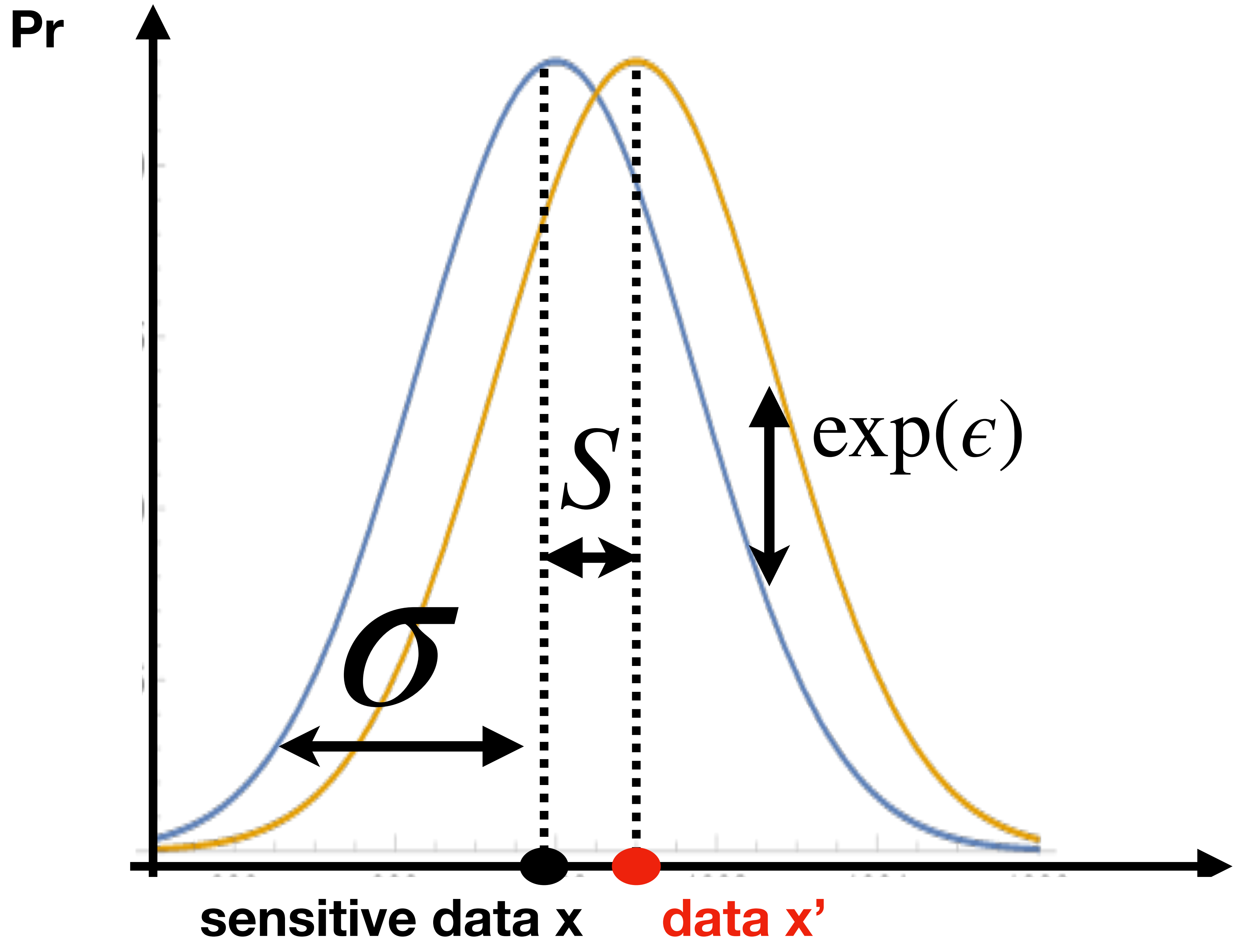


collects data from individuals
SHOP
...
SHOP
Daily number of visits to the merchant website
privately sample
privately sample

Streaming LDP Mechanism

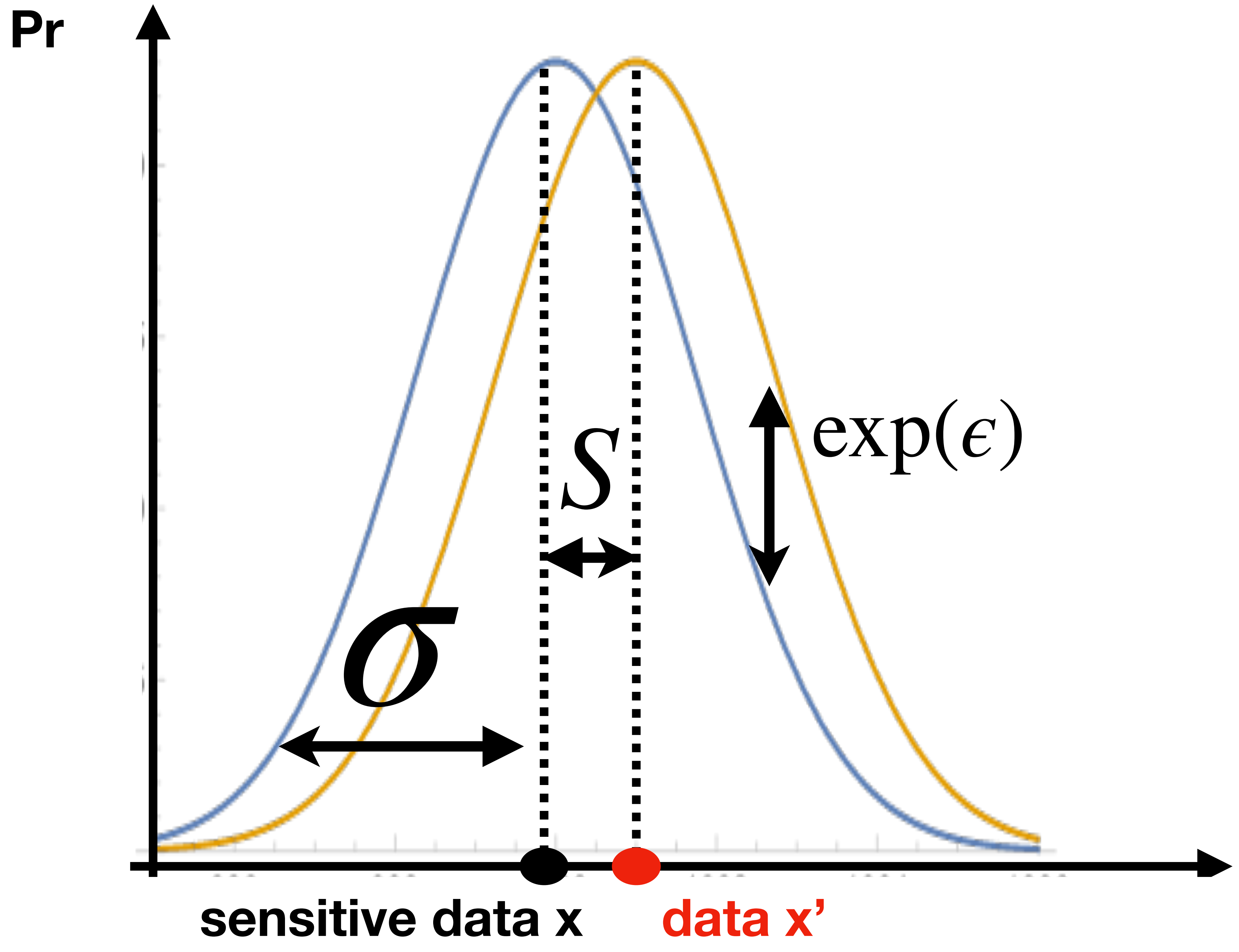


Streaming LDP Mechanism



Repeatedly sample everyday.

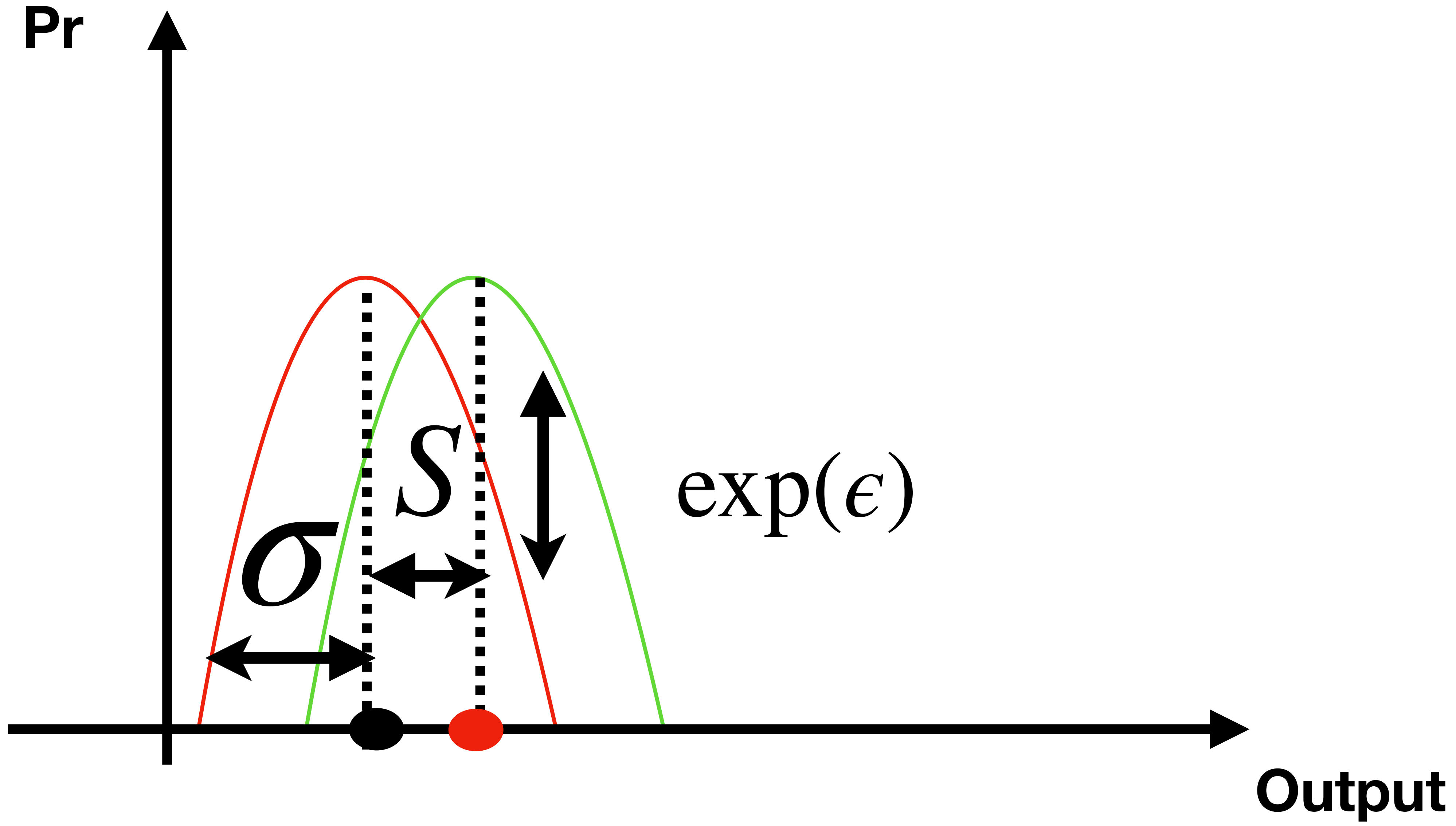
Streaming LDP Mechanism

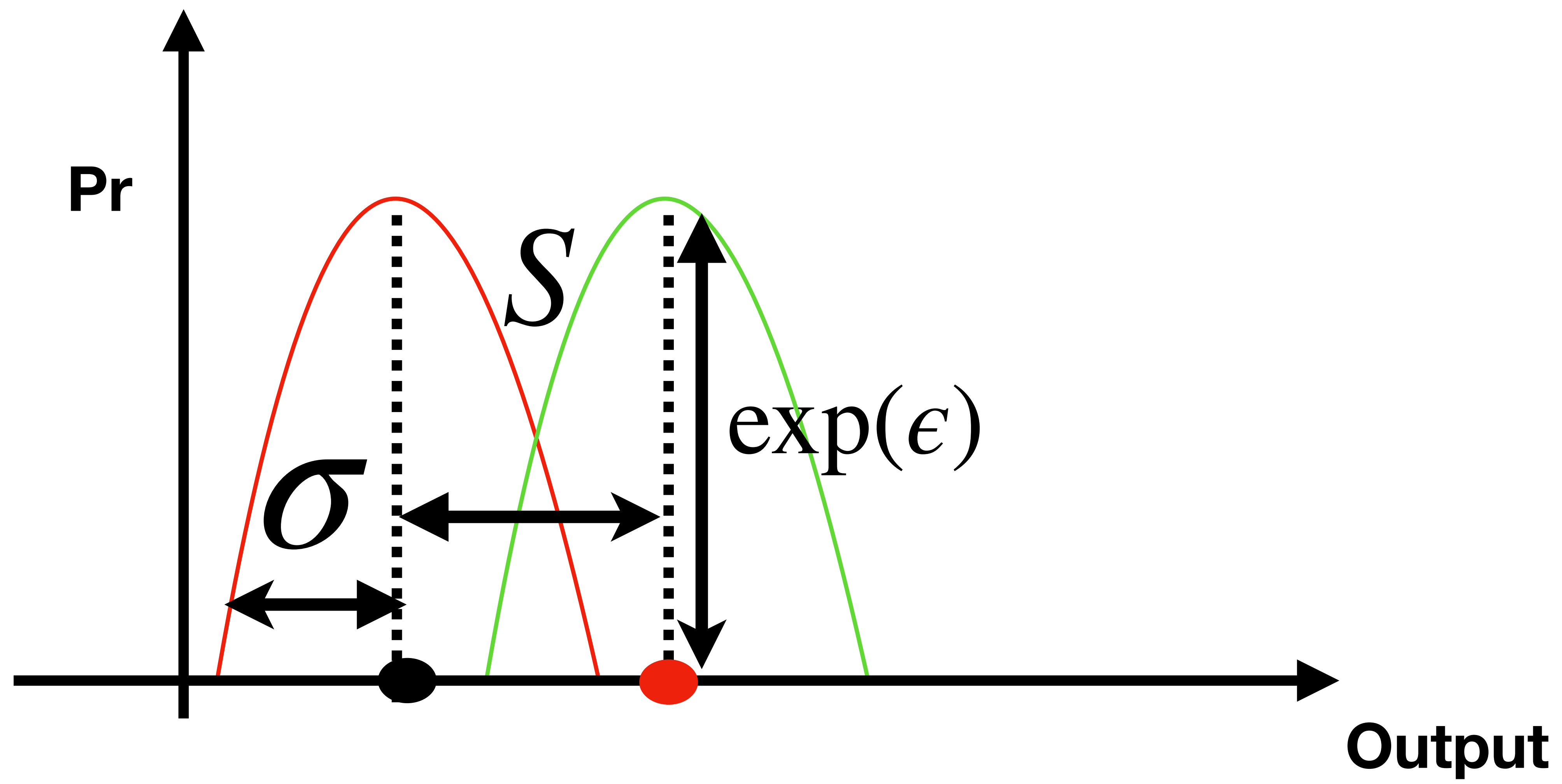


~~Repeatedly sample everyday.~~

There is a better way!

LDP Mechanism

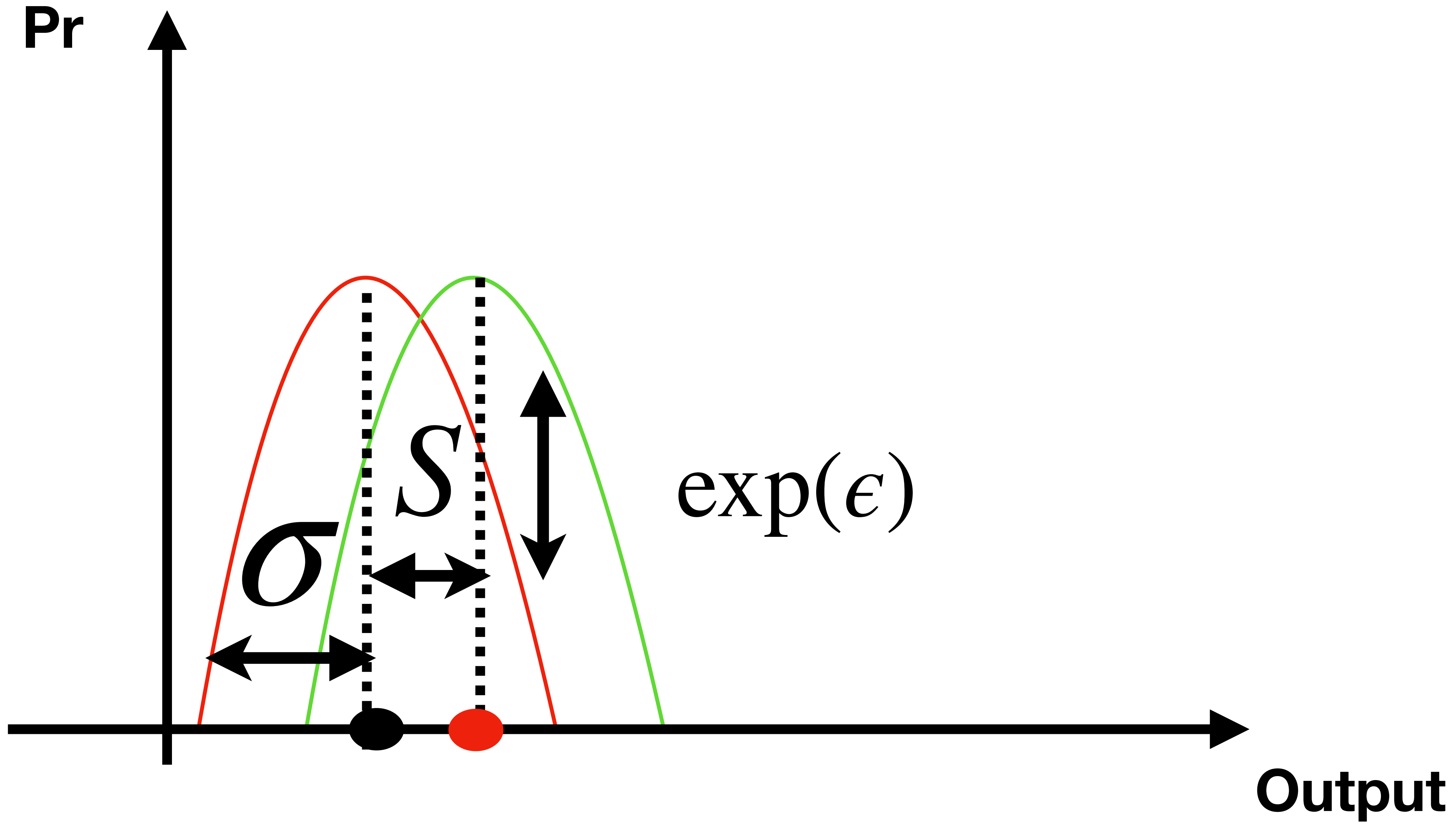




Utility $\sigma \approx \dfrac{S}{\epsilon}$ **Sensitivity** / **Privacy**

$S \nearrow \quad \sigma \nearrow$ **Utility** $\searrow$

LDP Mechanism

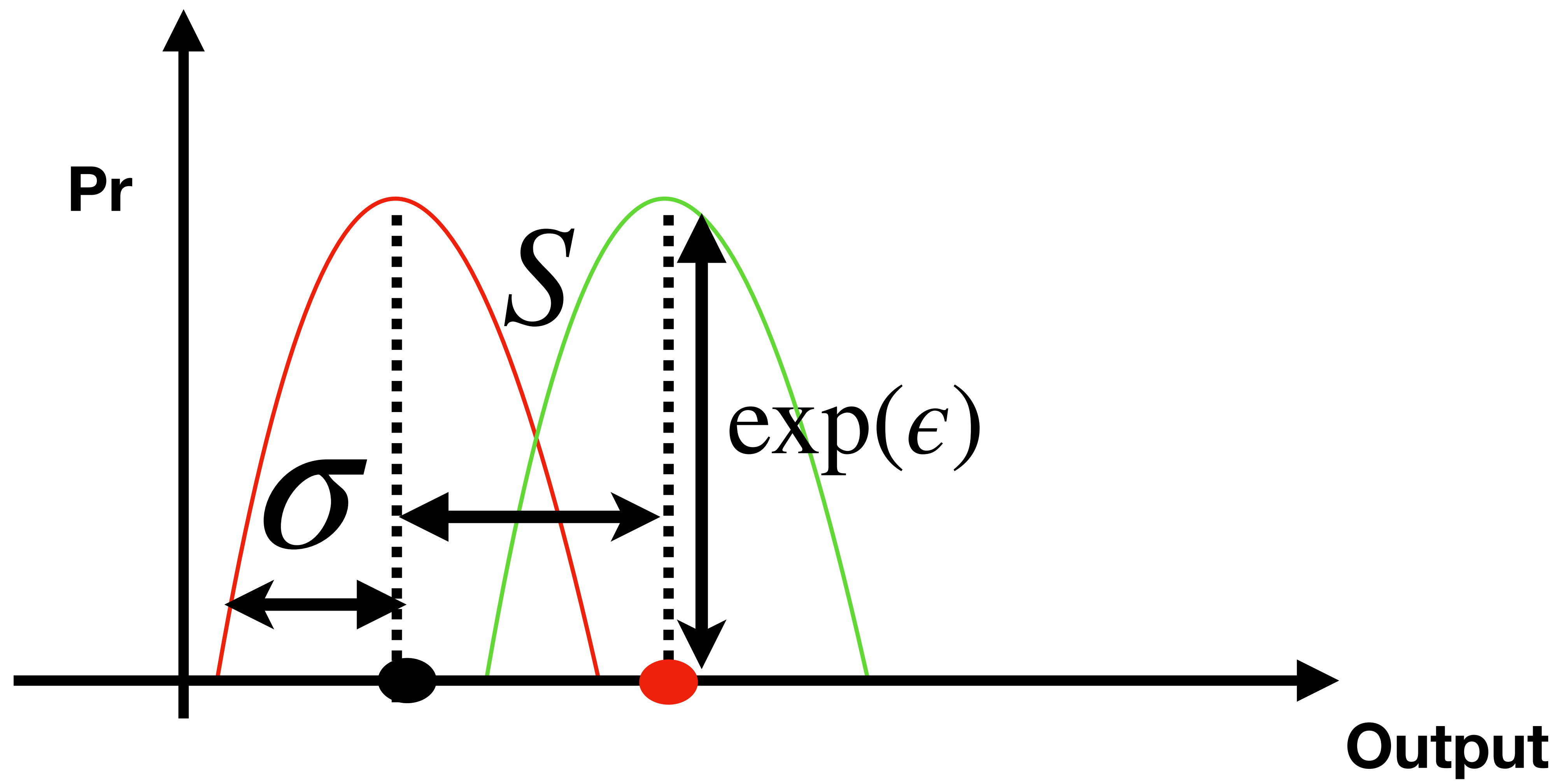


Utility

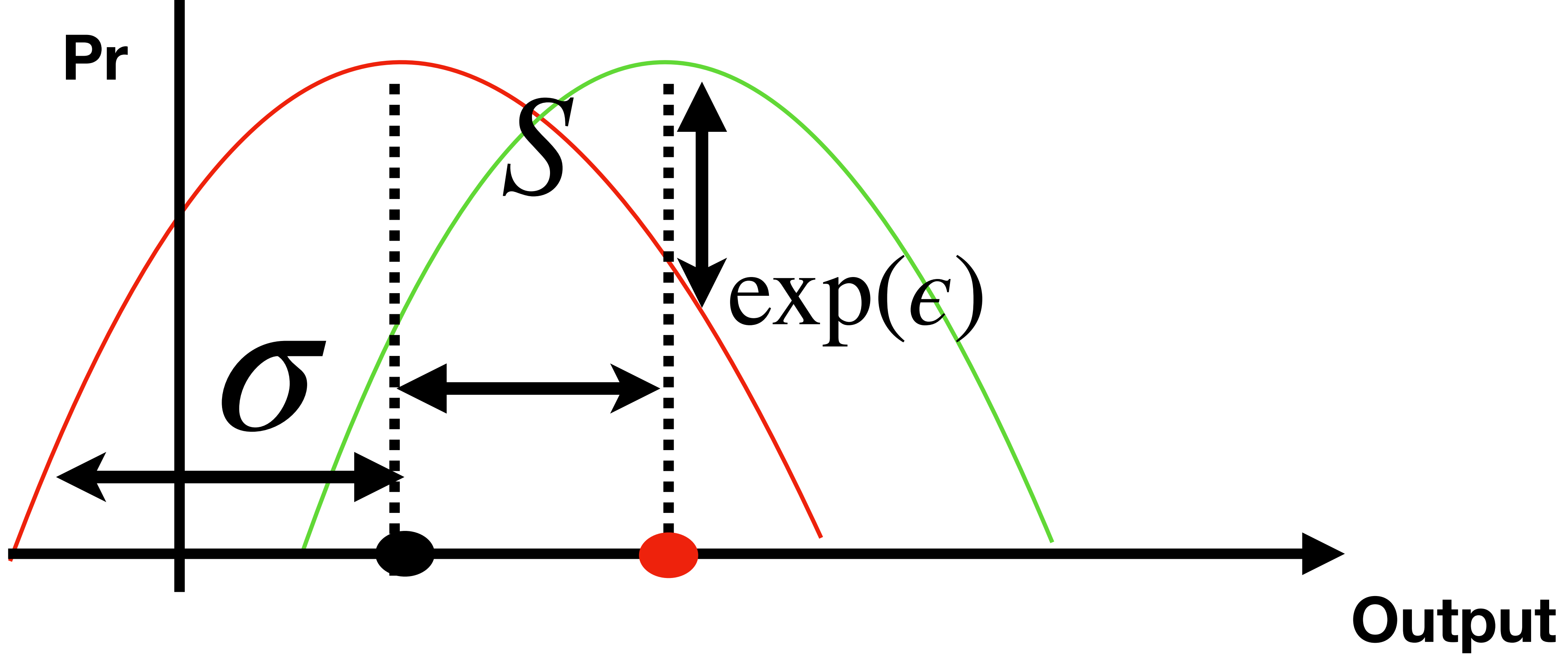
$$\sigma \approx \frac{S}{\epsilon}$$

Sensitivity

Privacy



$S \nearrow \sigma \nearrow$ Utility $\searrow$



A better streaming LDP Mechanism

~~Repeatedly sample everyday.~~

There is a better way!

$$\sigma \approx \frac{S}{\epsilon}$$

Day 1

Day 2

$+1000$

9000 visits

10000 visits

A better streaming LDP Mechanism

~~Repeatedly sample everyday.~~

$$\sigma \approx \frac{S}{\epsilon}$$

$$\underset{C}{1000} = \frac{1}{10} \times \underset{S}{10000}$$

There is a better way!

Day 1 $\pm C$ **Day 2**

$+1000$

9000 visits **10000 visits**

A better streaming LDP Mechanism

~~Repeatedly sample everyday.~~

$$\sigma \approx \frac{S}{\epsilon}$$

$$\underset{C}{1000} = \frac{1}{10} \times \underset{S}{10000}$$

There is a better way!

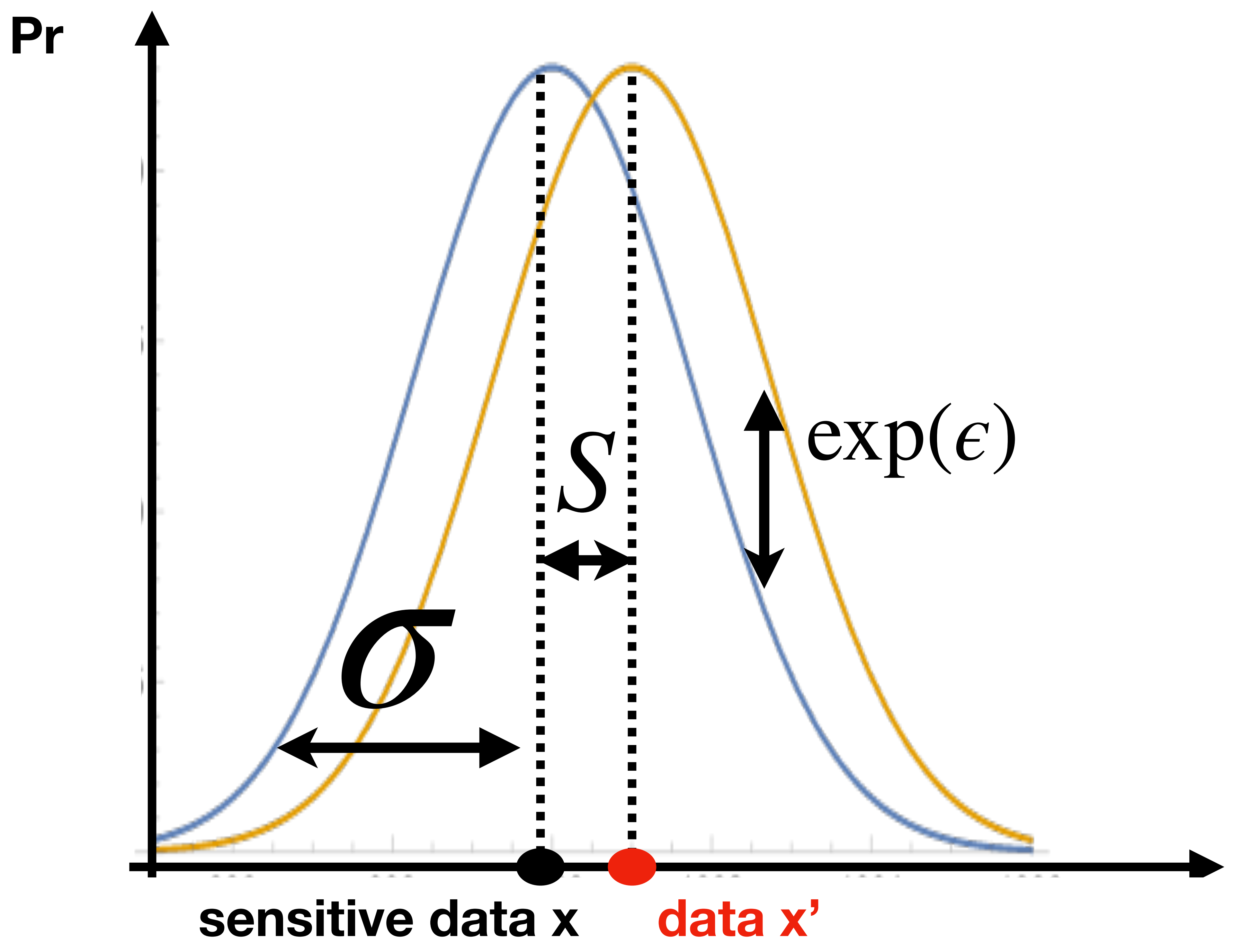
Day 1 $\pm C$ **Day 2**

$+1000$

9000 visits **10000 visits**

σ ↘ **Utility** ↗

A better streaming LDP Mechanism



$$\sigma \approx \frac{S}{\epsilon}$$

Utility

Sensitivity

Privacy

CGM helps carefully chooses the ``proper`` sensitivity S, in order to minimize σ, improving utility.

A better streaming LDP Mechanism

Day 1 $\pm C$ **Day 2**

$+1000$

9000 visits **10000 visits**

Other scenarios: daily App usage, daily phone usage, taxi locations, etc..

CGM: $\dfrac{C}{S} \searrow$ **Utility improvement** $\nearrow$

A better streaming LDP Mechanism

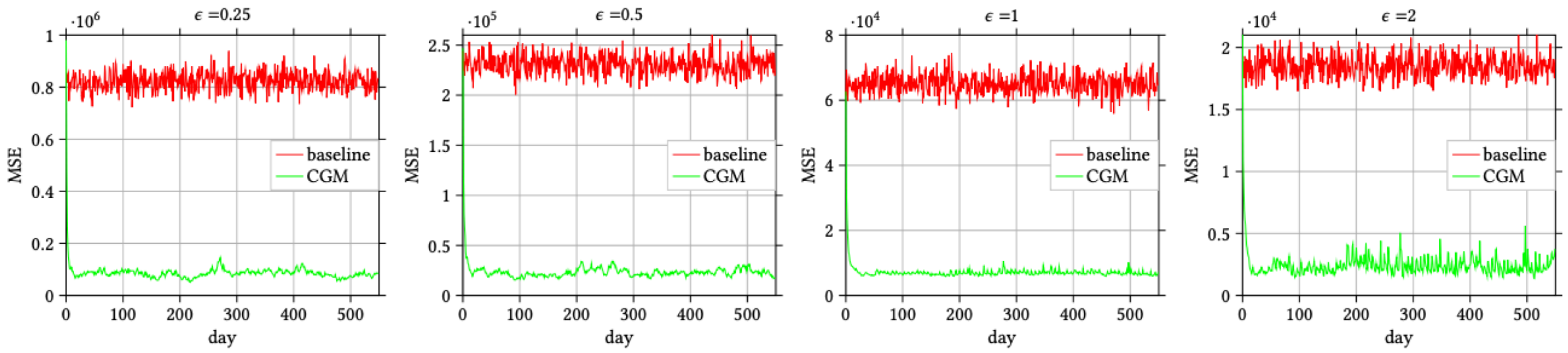


Figure 1: Utility performances of CGM (Algorithm 3) and the baseline approach (Algorithm 1) on the Kaggle Web Traffic dataset, with varying daily privacy budget $\epsilon \in \{0.25, 0.5, 1, 2\}$ and $\delta = 10^{-5}$. For CGM, the differential bound is fixed to $C = 500$.

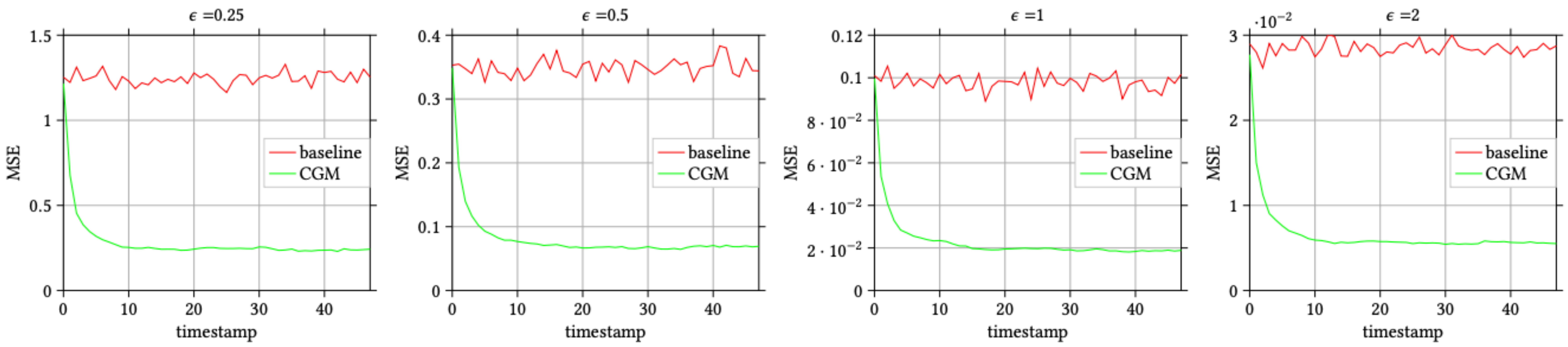


Figure 2: Utility performances of CGM (Algorithm 3) and the baseline approach (Algorithm 1) on the Beijing Taxi dataset, with varying total privacy budget for all updates $\epsilon \in \{0.25, 0.5, 1, 2\}$ and $\delta = 10^{-5}$. For CGM, the differential bound is fixed to $C = 0.05$. The whole space is normalized to $[0, 1] \times [0, 1]$, and the query region is $[0.45, 0.55] \times [0.45, 0.55]$.

Effect of $\frac{C}{S}$

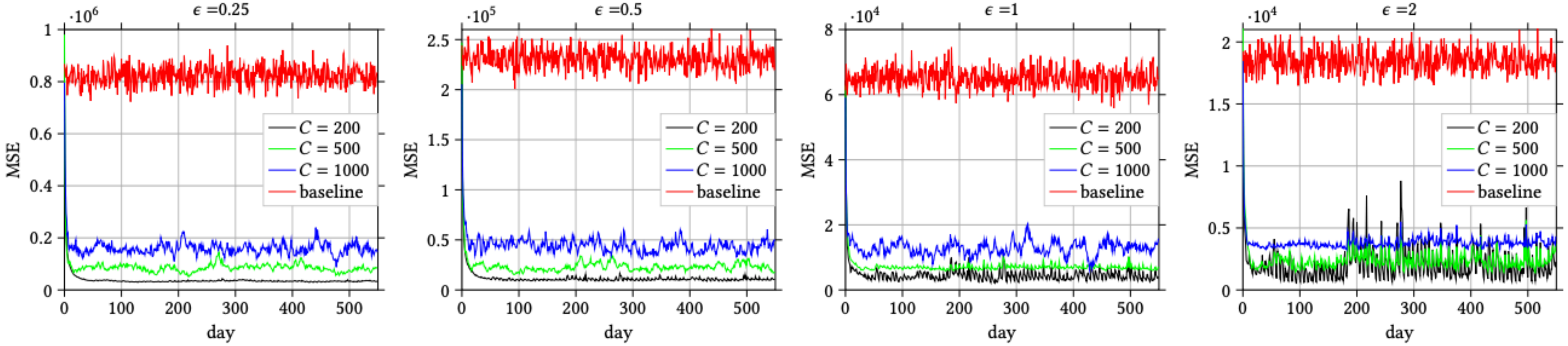


Figure 3: Impact of varying differential bound $C \in \{200, 500, 1000\}$ on the utility performance of CGM on the Kaggle Web Traffic dataset, where $\epsilon \in \{0.25, 0.5, 1, 2\}$ and $\delta = 10^{-5}$.

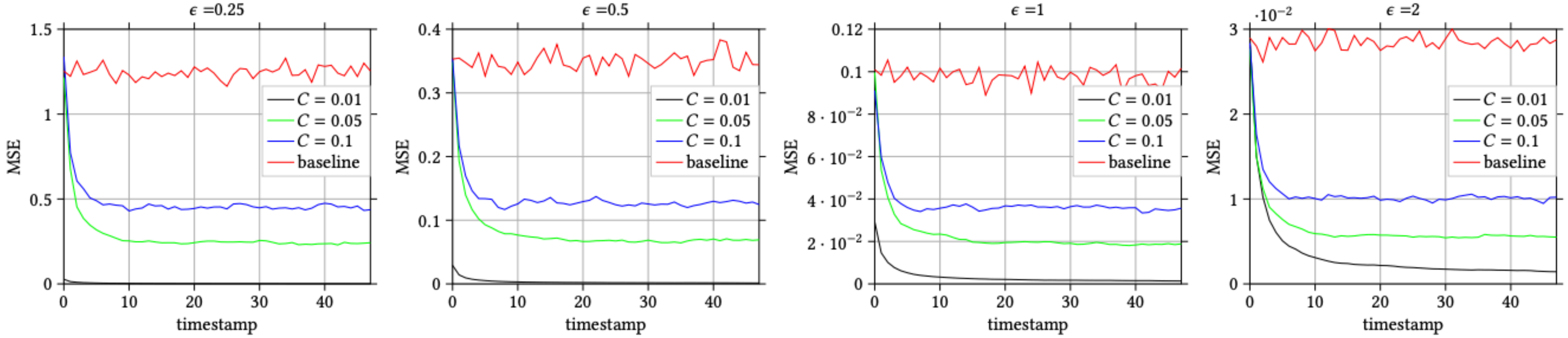
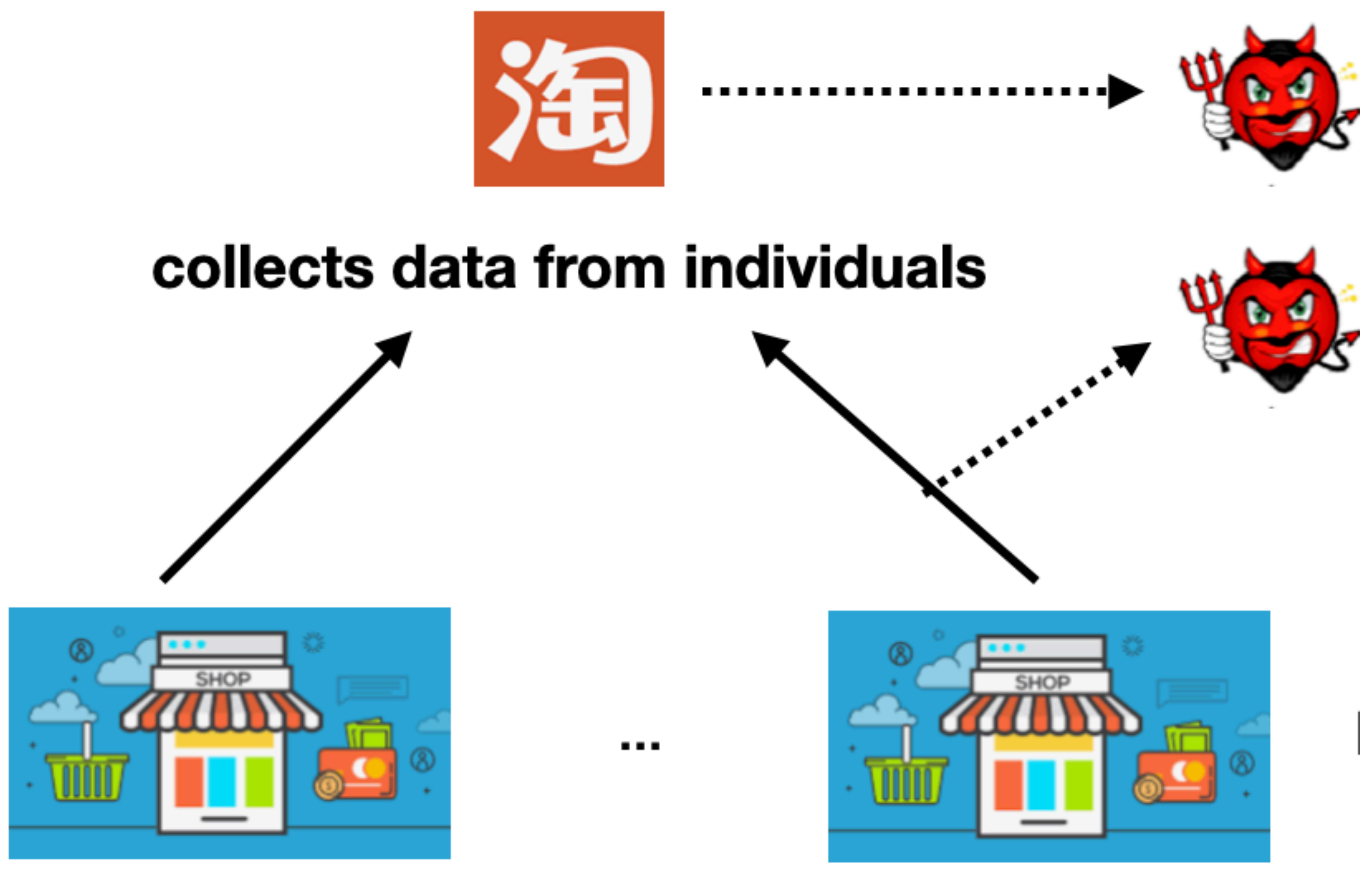
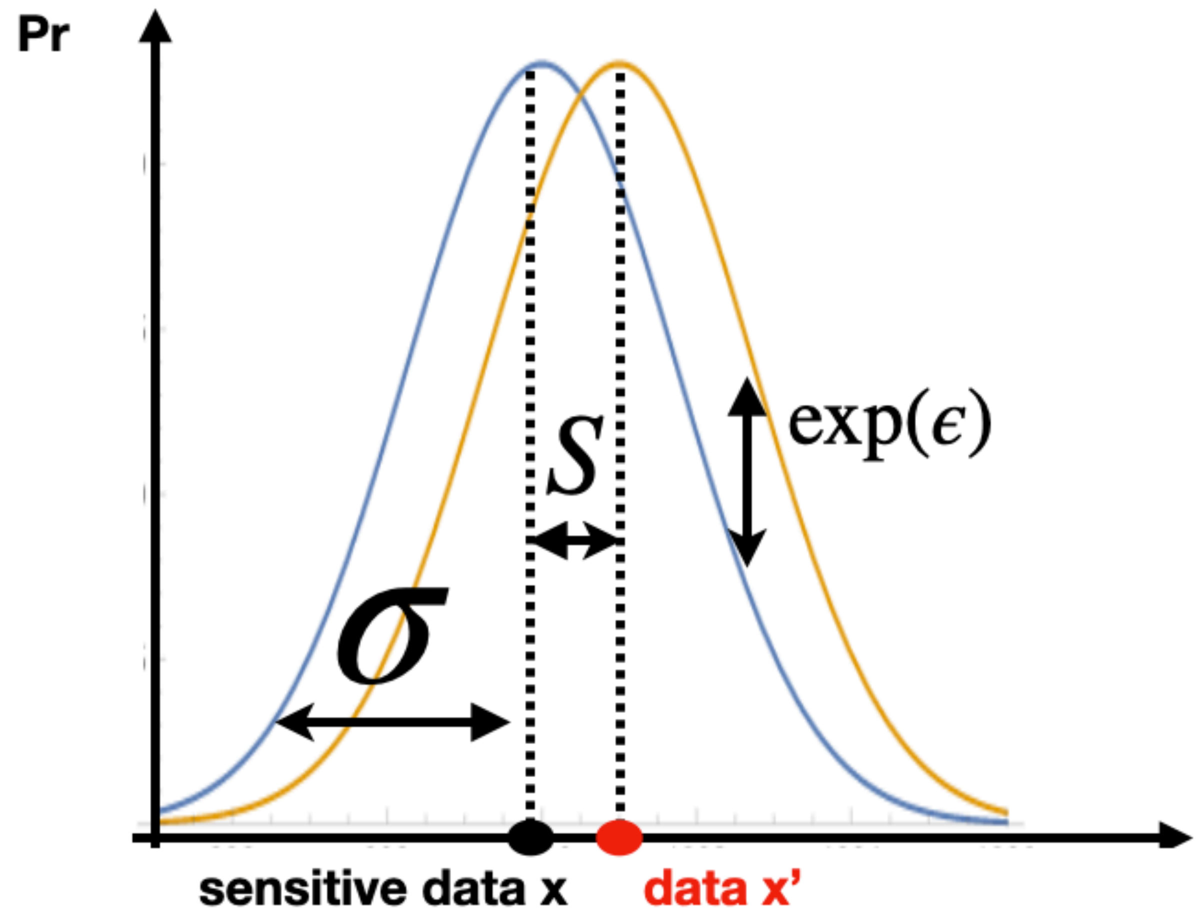


Figure 4: Impact of varying differential bound $C \in \{0.01, 0.05, 0.1\}$ on the utility performance of CGM on the Beijing Taxi dataset, where $\epsilon \in \{0.25, 0.5, 1, 2\}$ and $\delta = 10^{-5}$. The query region is $[0.45, 0.55] \times [0.45, 0.55]$.

Summary

In this work, we study the problem of streaming data collection under local differential privacy. In this setting, an individual possesses a stream of data items, and the goal is to design a randomized mechanism for releasing the data stream without compromising the individual's privacy.





Summary

The naive, and yet common approach, requires each user to perturb the data item independently at each timestamp, and upload the perturbed data to the untrusted aggregator. This approach leads to an excessively large amount of noise. Addressing this issue, we exploit data autocorrelations common in many real data streams, and propose a novel correlated Gaussian mechanism (CGM).

Utility

$$\sigma \approx \frac{S}{\epsilon}$$

Sensitivity

Privacy

Summary

Through both theoretical analysis and extensive experimental evaluations using real data from multiple application domains, we demonstrate that CGM consistently and significantly outperforms the baseline solution in terms of result utility.