Privacy vs Utility, Differential Privacy and the Hybrid Model

Catuscia Palamidessi

informatiques mathématiques
Inria

Utility versus privacy

Utility

Various kinds of utility:

- Quality of service
- Precise statistical analyses
- Accuracy (machine learning)

The main challenge is to find mechanisms that optimize the trade-off between utility and privacy

Utility

Various kinds of utility:

- Quality of service
- Precise statistical analyses
- Accuracy (machine learning)

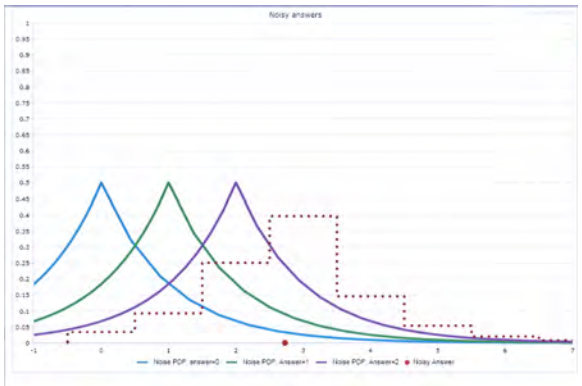
The main challenge is to find mechanisms that optimize the trade-off between utility and privacy

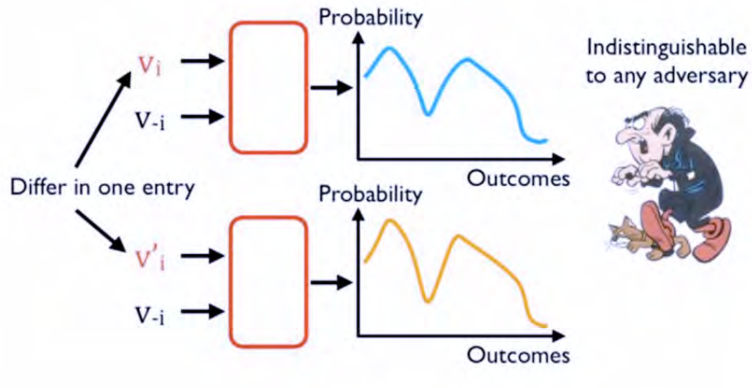
Privacy by randomization

Differential Privacy [Dwork et al., 2006]

A mechanism $\mathcal{K}$ (for a certain query) is ε-differentially private if for every pair of *adjacent* datasets x and x' and every possible answer y

$$P[\mathcal{K}(x) = y] \leq e^{\varepsilon}\ P[\mathcal{K}(x') = y]$$

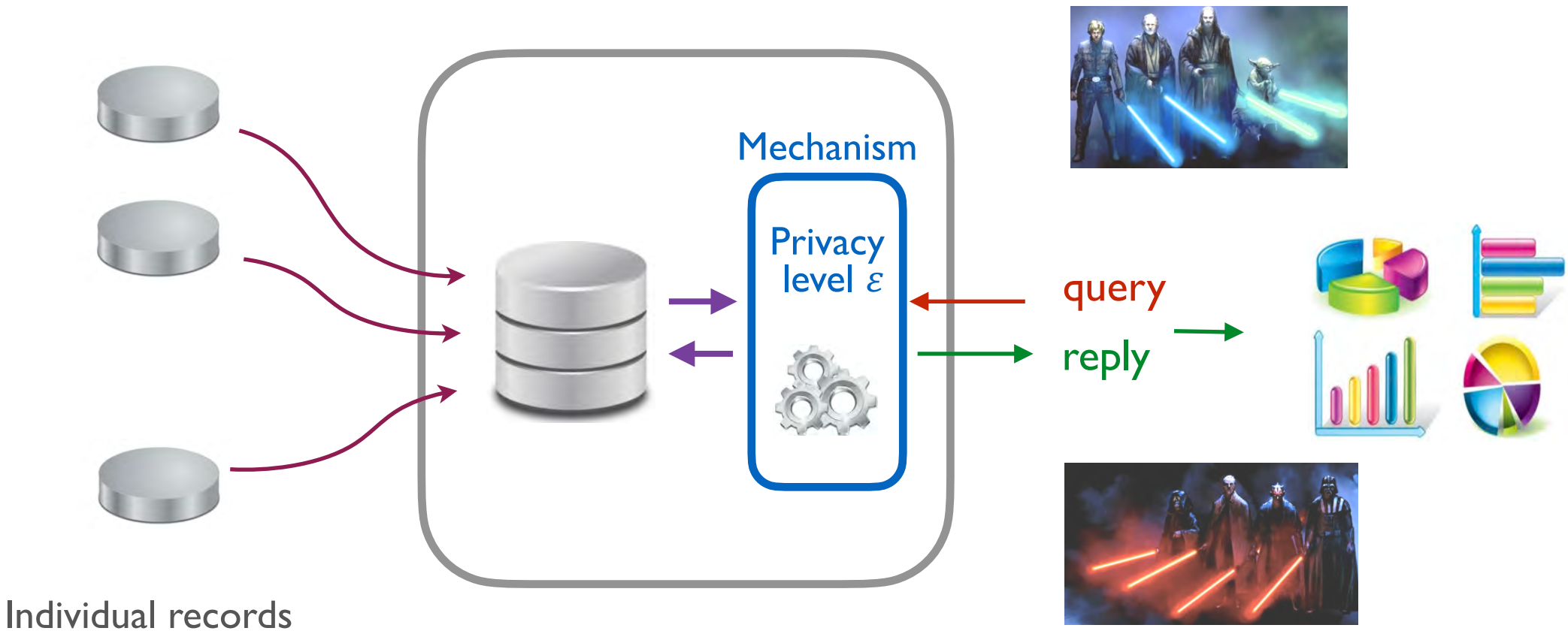




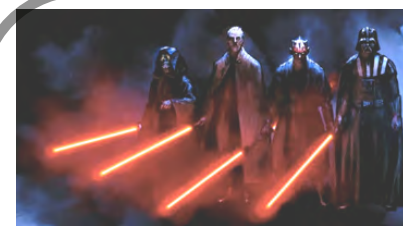
- **Compositionality**: the combination of two mechanisms which are ε_1 and ε_2 differentially private is $\varepsilon_1 + \varepsilon_2$ differentially private
- **Independent** from side knowledge

Typical DP mechanisms: Laplace, Geometric

Standard Differential Privacy (aka central model)



Standard Differential Privacy (aka central model)



Mechanism

Privacy level ε

query

reply

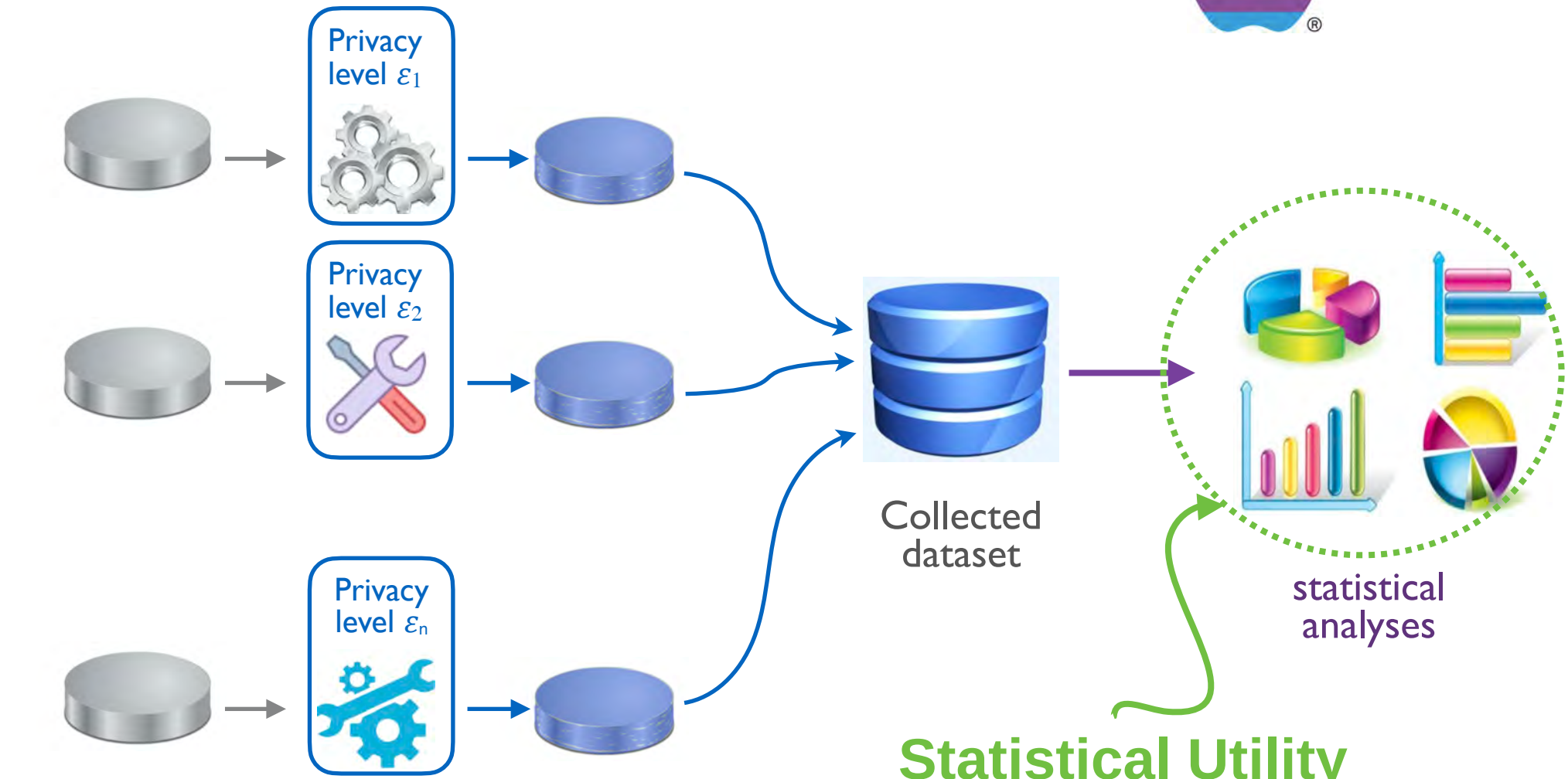
Individual records

Collected dataset

Local Differential Privacy

Google

Apple



Local Differential Privacy

[Jordan &Wainwright '13]

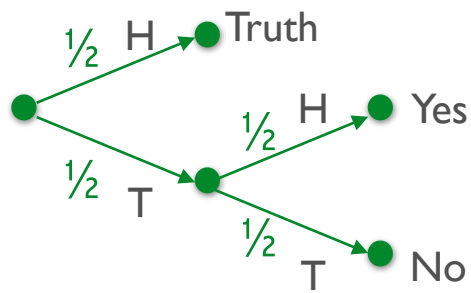
Definition Let $\mathcal{X}$ be a set of possible values and $\mathcal{Y}$ the set of noisy values. A mechanism $\mathcal{K}$ is ε-locally differentially private (ε-LDP) if for all $x_1, x_2 \in \mathcal{X}$ and for all $y \in \mathcal{Y}$

$$P[\mathcal{K}(x) = y] \leq e^{\varepsilon}\ P[\mathcal{K}(x') = y]$$

or equivalently, using the conditional probability notation:

$$p(y \mid x) \leq e^{\varepsilon}\ p(y \mid x')$$

For instance, the Randomized Response protocol is (log 3)-LPD



x \ y	yes	no
yes	¾	¼
no	¼	¾

Mechanism's stochastic matrix

The kRR mechanism (aka flat m.)

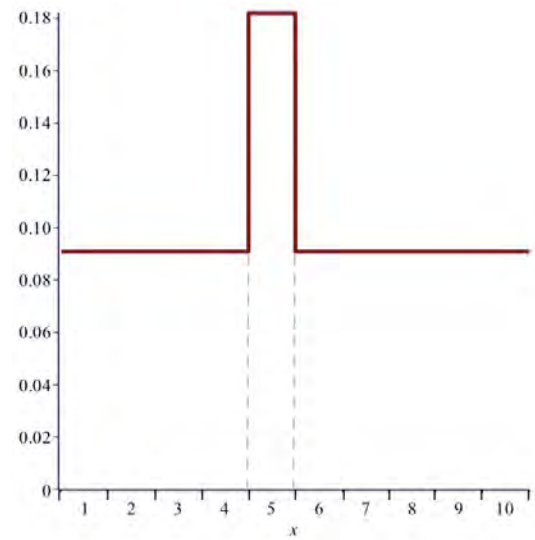
[Kairouz et al, '16]

The flat mechanism is the simplest way to implement LPD. It is defined as follows:

$$p(y|x) = \begin{cases} c\,e^{\varepsilon} & \text{if } x = y \\ c & otherwise \end{cases}$$

where c is a normalization constant.

namely $c = \dfrac{1}{k - 1 + e^{\varepsilon}}$ where k is the size of the domain



Privacy Properties:

- Compositionality
- Independence from the side knowledge of the adversary

Utility :

- Statistical Utility : ✓
- QoS : ✗

Our approach to LDP

d-privacy

d-privacy: a generalization of DP and LDP

d-privacy

On a generic domain $\mathcal{X}$ provided with a distance d:

$$\forall x, x' \in \mathcal{X}, \forall z \quad \frac{p(z \mid x)}{p(z \mid x')} \leq e^{\varepsilon\, d(x,x')}$$

generalizes

Differential Privacy

- x, x' are databases
- *d* is the Hamming distance

Local Differential Privacy

- *d* is the discrete distance

Properties

- Like LDP, it can be applied at the user side
- Like DP and LDP, it is compositional

Typical *d*-private mechanisms:

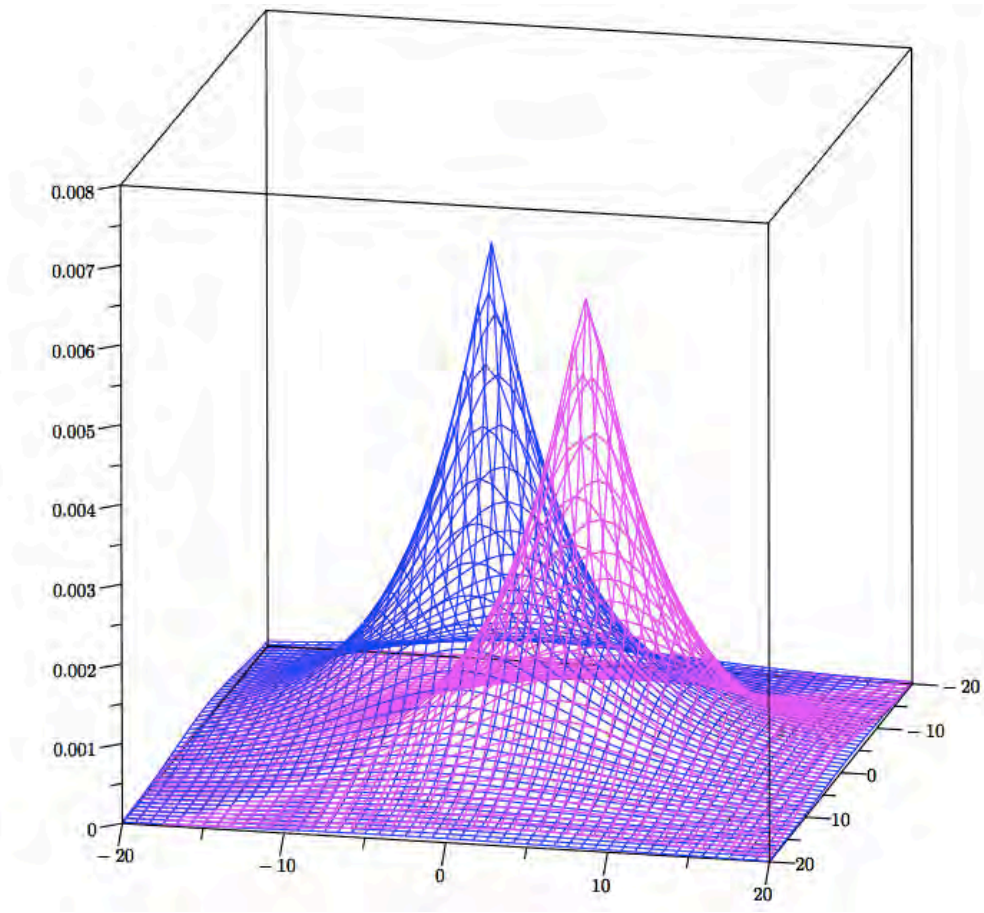
Extended Laplace and Extended Geometric

Example: Location privacy

- Domain: points on a plane
- Distance: Euclidean

$$dp_x(z) = \frac{\epsilon^2}{2\pi} e^{\epsilon\, d(x,z)}$$

Efficient method to draw noisy locations based on polar coordinates

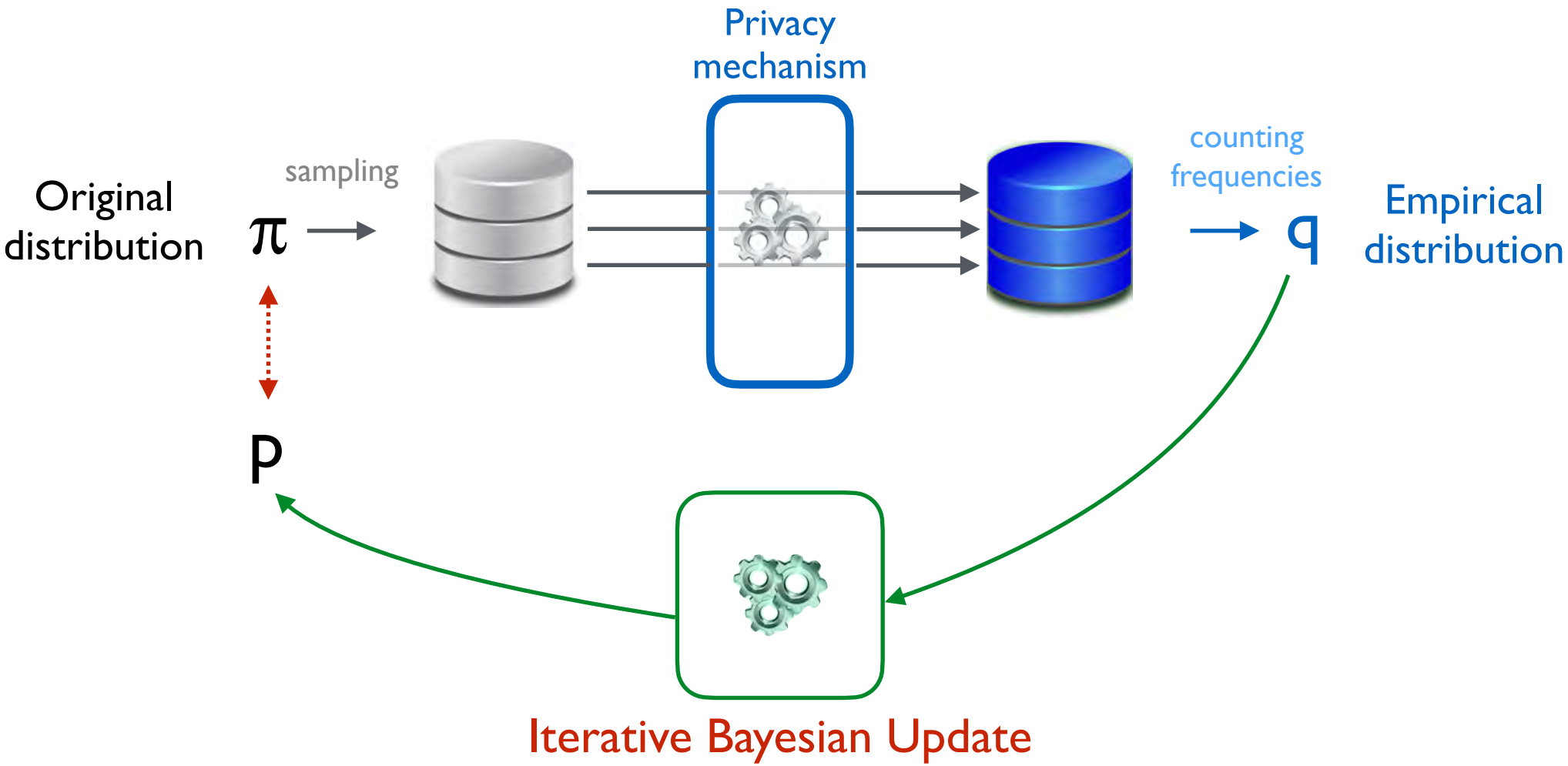


Statistical Utility:

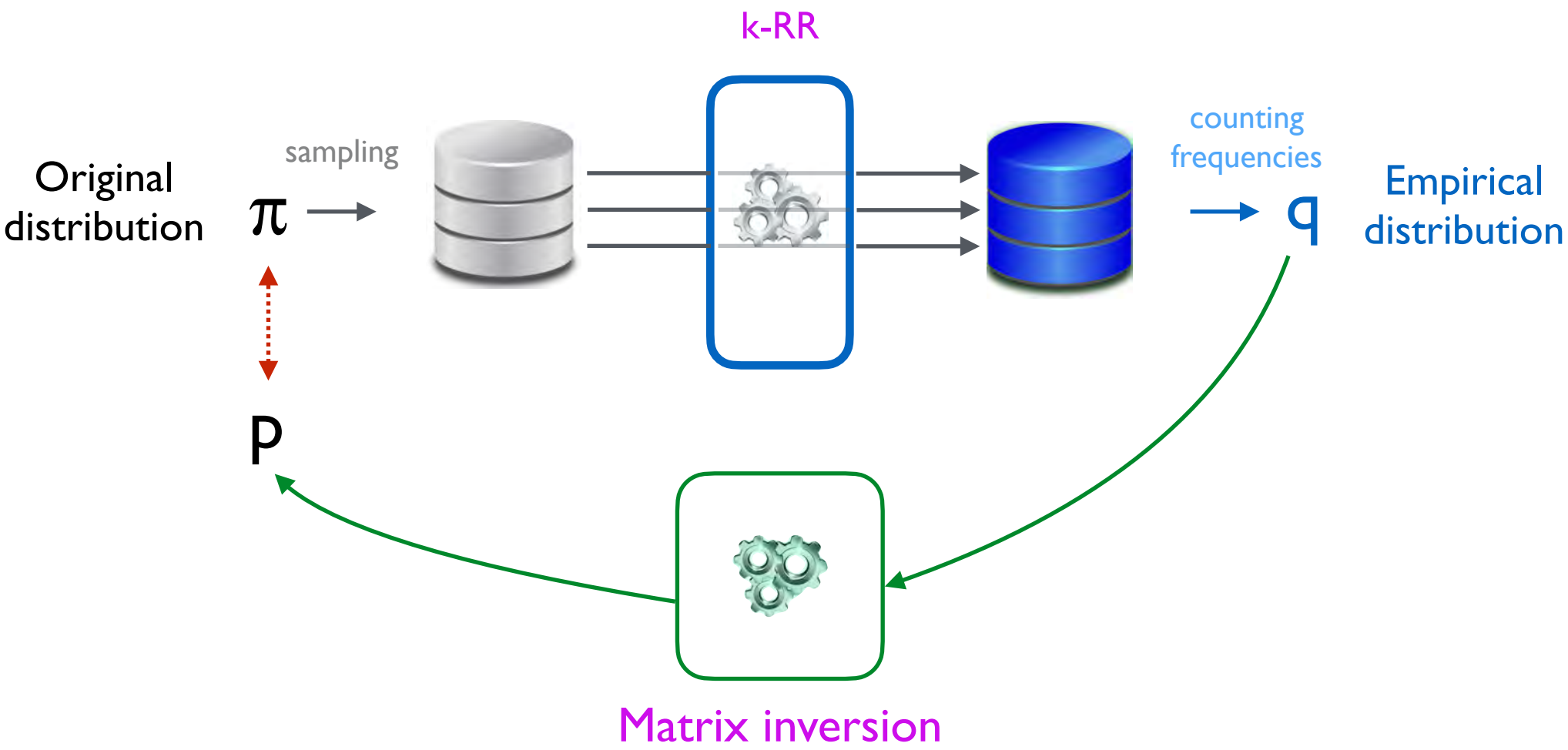
Estimating the original distribution

i.e., the distribution from which the true data are sampled

Estimation method

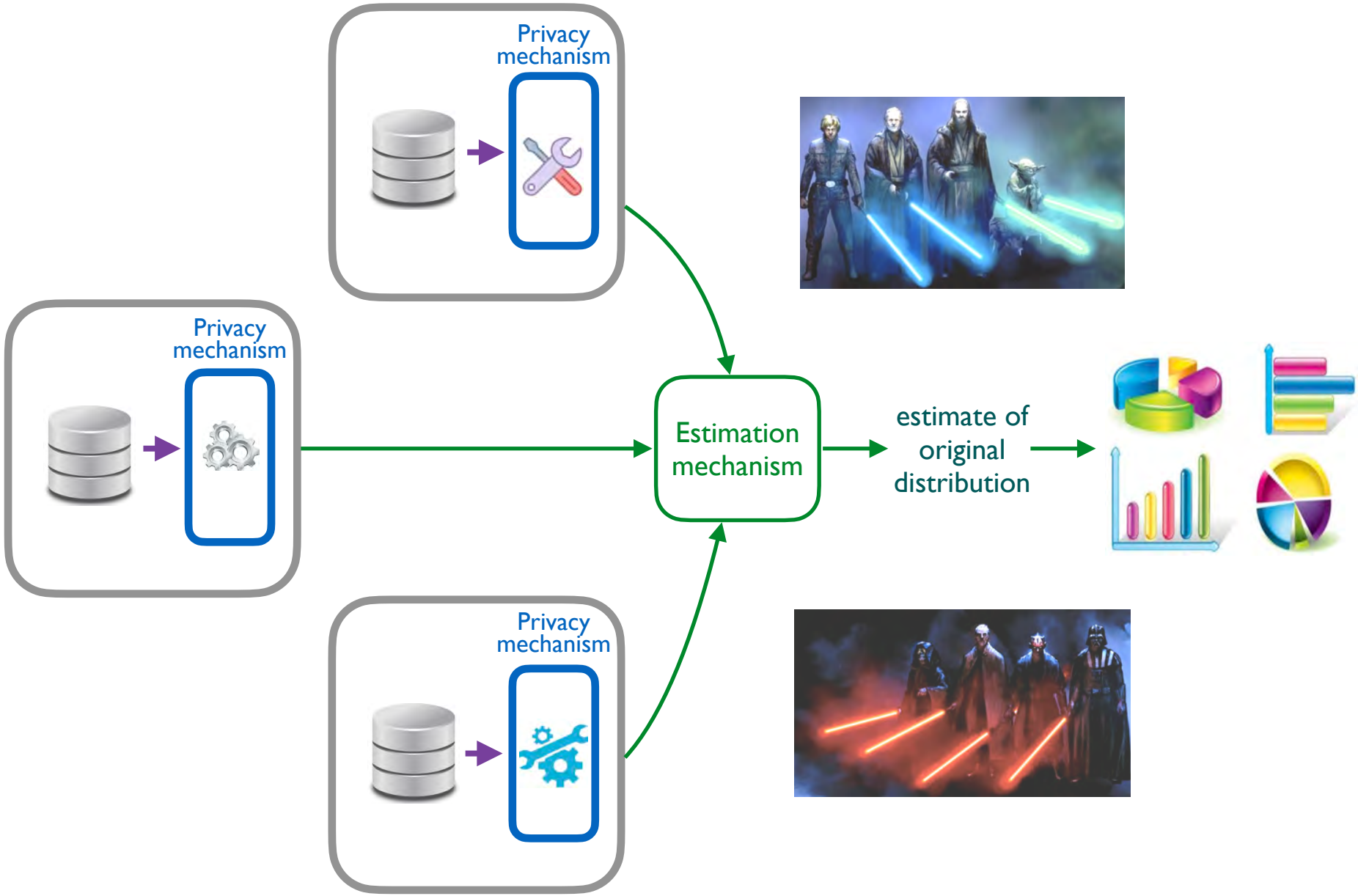


Estimation method

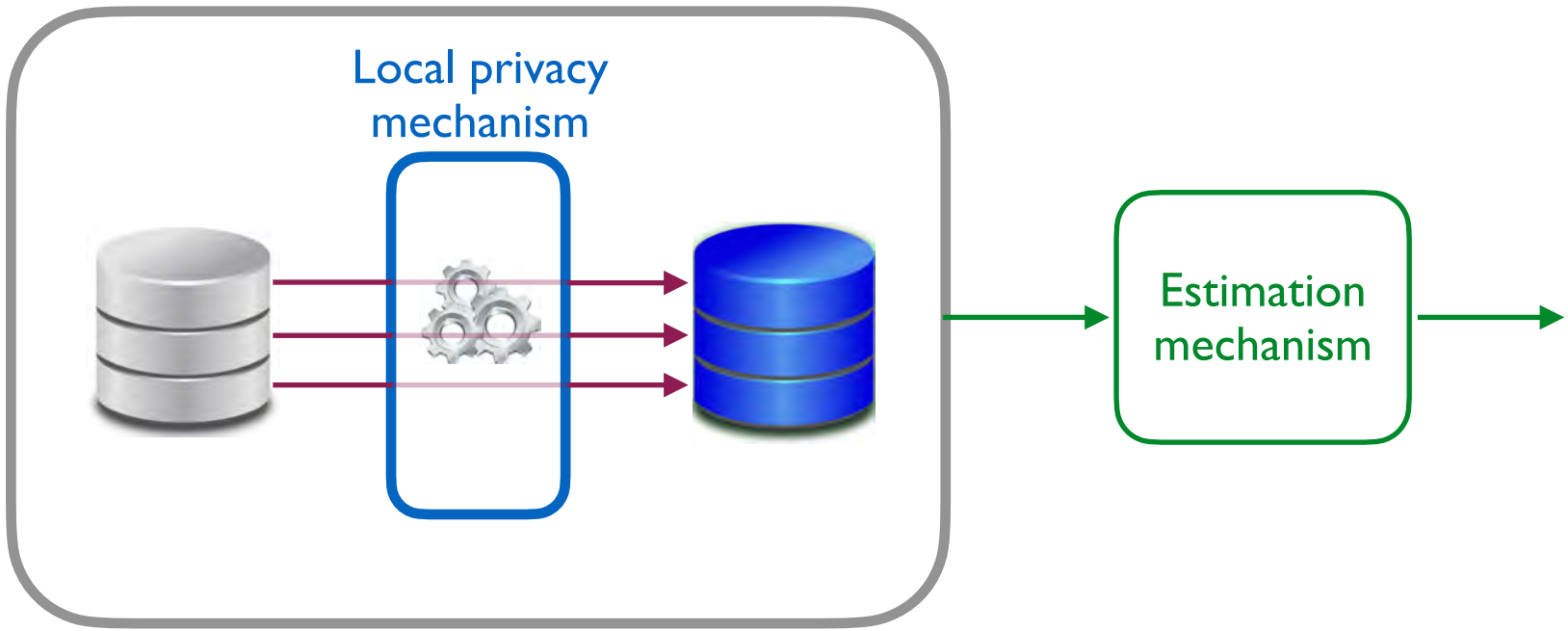


The Hybrid Model for distributed settings (federated learning)

Distributed setting



Our hybrid approach



Apply a LDP mechanism to each record individually

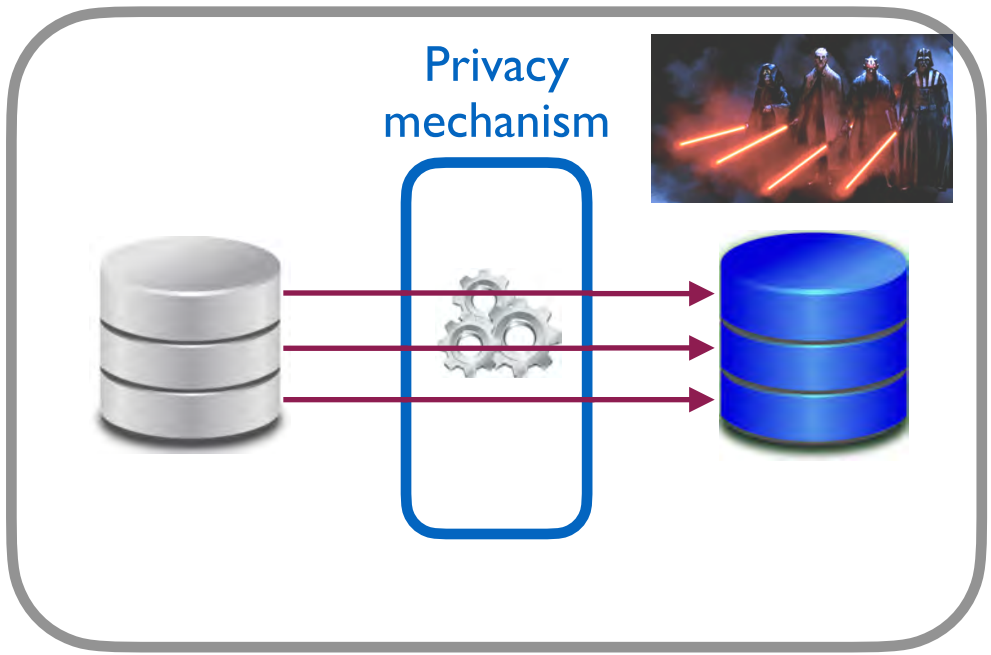
Estimate the original distribution like in LDP

Advantages of hybrid wrt local

- The trade-off utility-privacy is usually much worse in the local model than in the central model
- However, in the hybrid model, the trade-off of certain mechanisms (kRR + Inv and d-privacy + IBU) is as good as in the central model. The reason is that the notion of attacker is weaker
- Hybrid approach: combination of the local and central model. The **mechanism is local**, while the **attacker is like in the central model**, which is **weaker** than the one of the local model

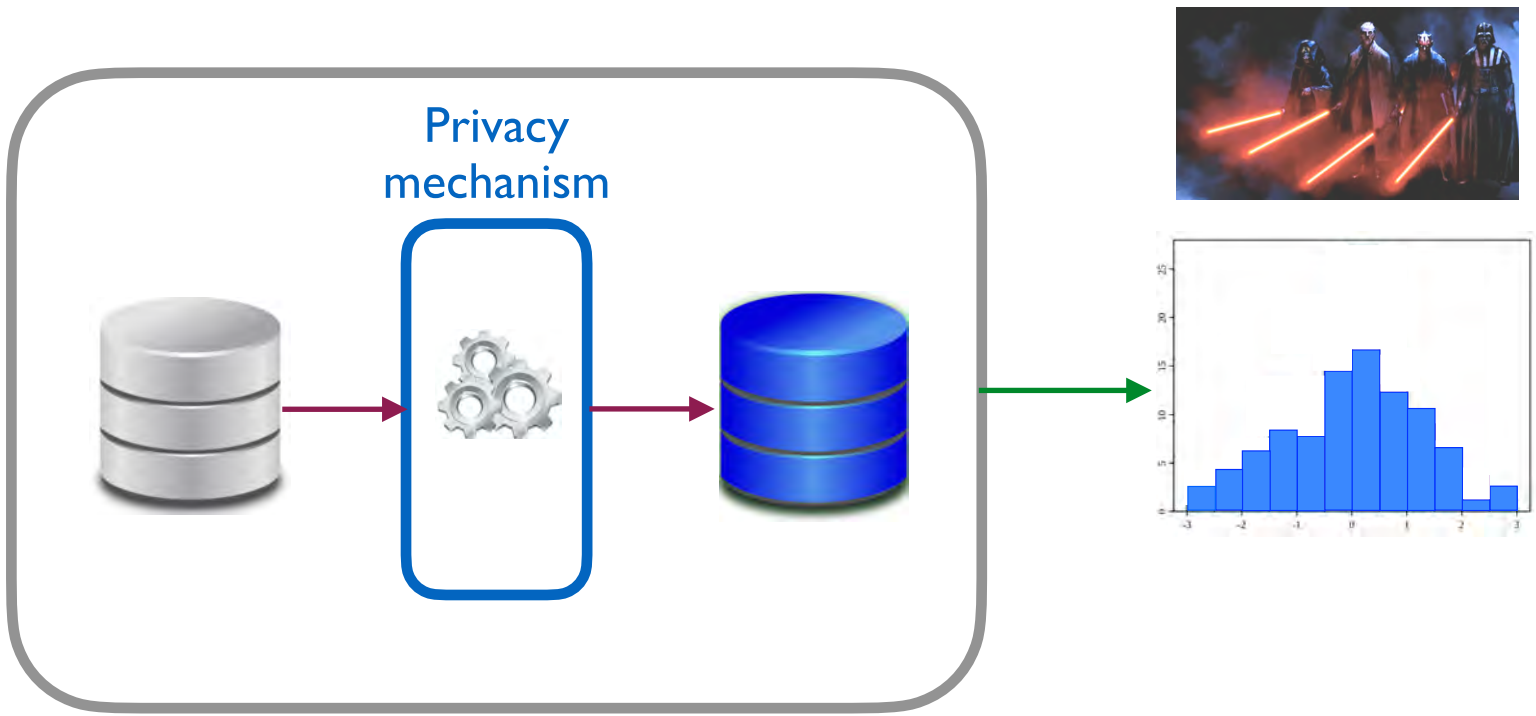
Privacy in the hybrid model

Attacker in the local model



In the local model the attacker can see the obfuscated version of each record

Attacker in the hybrid model



In the hybrid model the attacker only see the aggregated result of the obfuscation

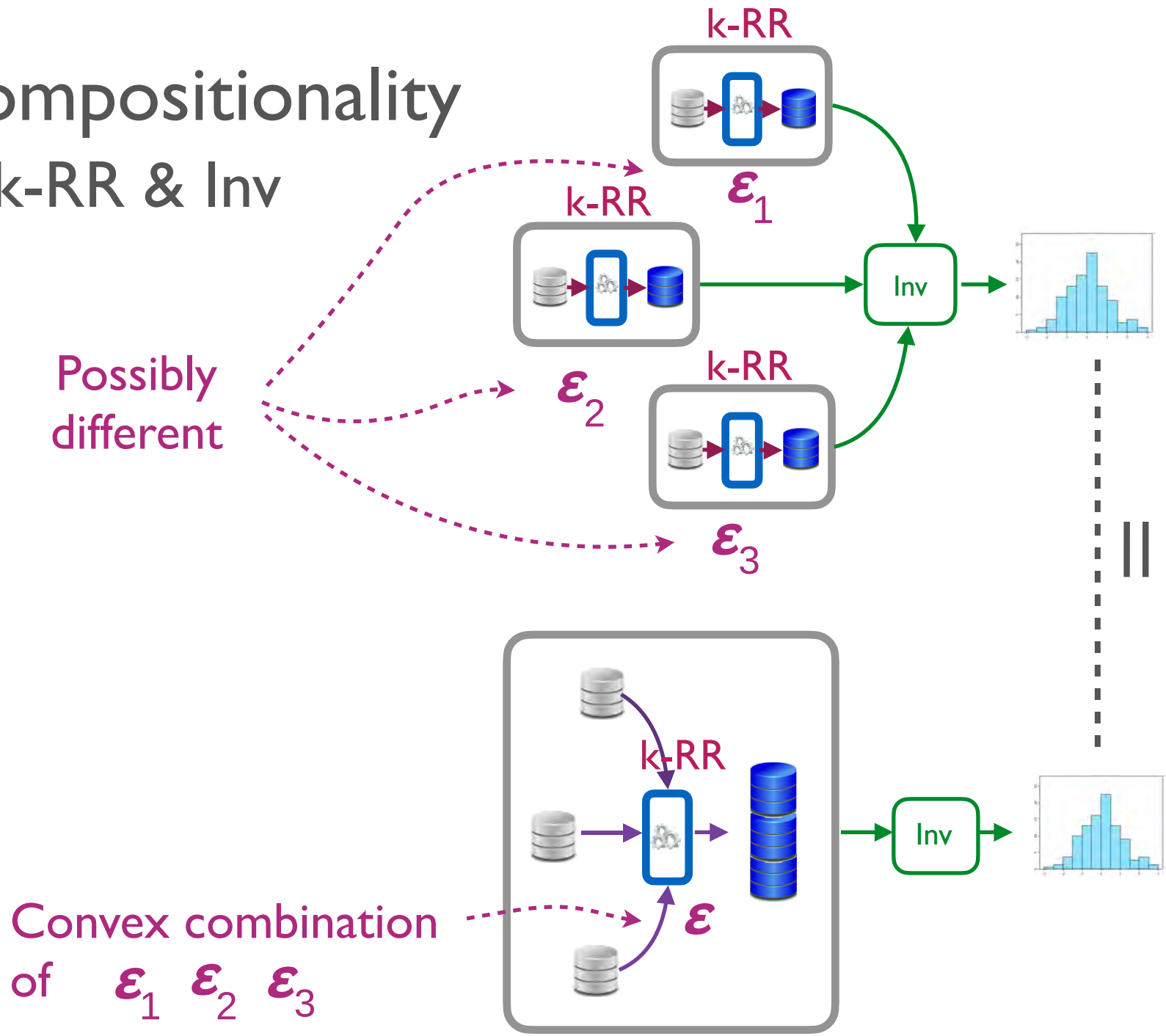
Utility in the hybrid model

Utility: hybrid vs central

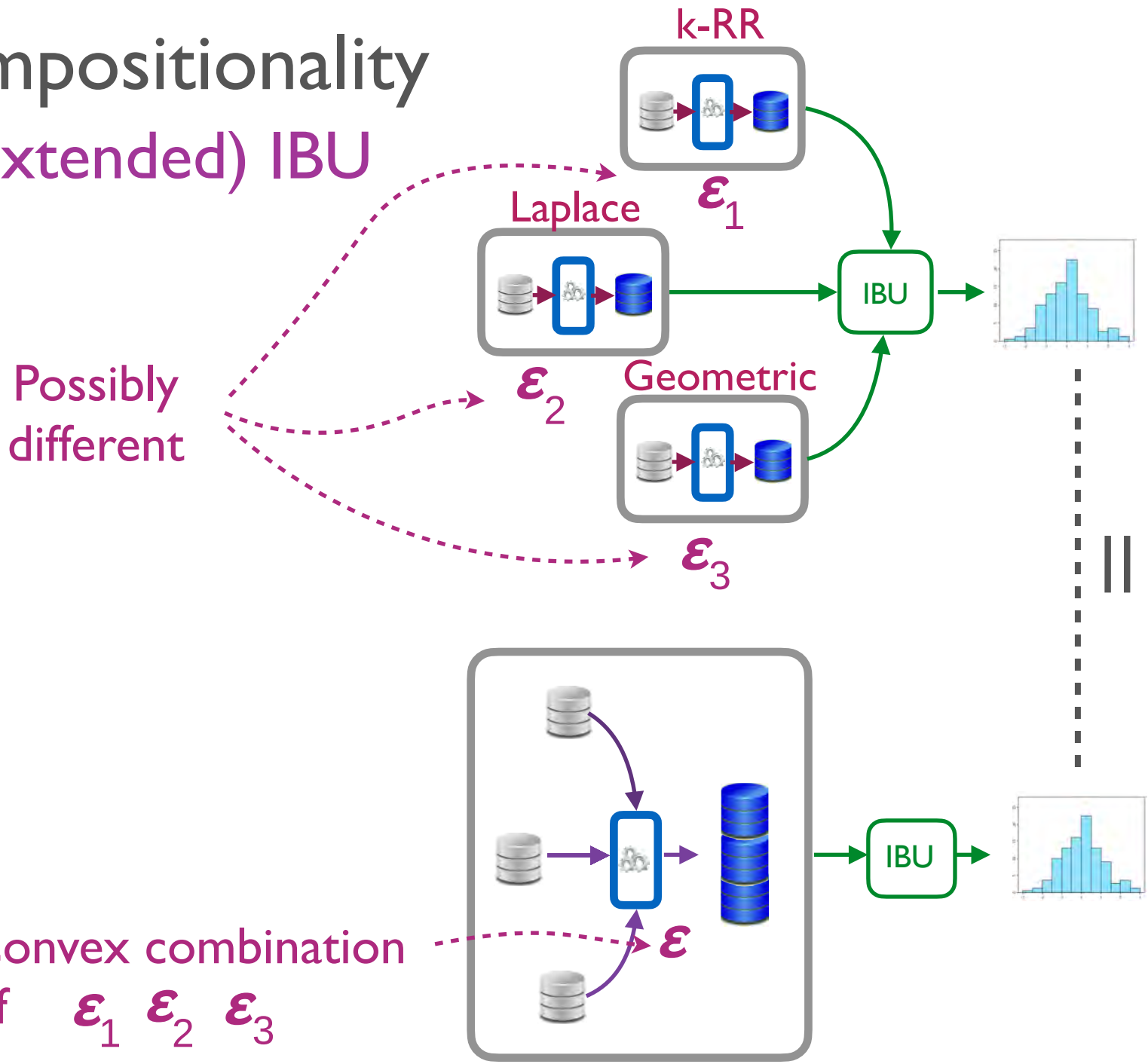
Advantage of hybrid wrt central: Compositionality

- IBU is compositional (on any local mechanism)
- Inv applied to k-RR is compositional

Compositionality
of k-RR & Inv



Compositionality of (extended) IBU



Advantage of hybrid wrt central: Compositionality

- IBU is compositional (on any local mechanism)
- Inv applied to k-RR is compositional

We could also compose the results of standard DP obfuscation (noise added to hystogram), but we would not get the same estimation accuracy:
The variance would be much larger.

d-privacy + IBU vs kRR + Inv

- IBU is more general: it can be applied to any privacy mechanism (and MLE is unique if the mechanism is invertible)
- d-privacy + IBU: better estimation accuracy if the distance between distributions takes into account the ground distance (e.g. , the Earth Movers' distance)
- kRR + Inv: more efficient

Conclusion

We have proposed an hybrid approach for DP in a distributed context, which is:

- better than LDP concerning the trade-off privacy-utility, and
- better compositionality properties than standard DP on distributed databases

Future work

- Explore other mechanisms (Gaussian)
- Explore the trade-off with accuracy in the sense of ML.

Thanks!

Questions ?