General Purpose f-DP Estimation and Auditing in a Black-Box Setting

Önder Askin, Holger Dette, Martin Dunsche,

Tim Kutta, Yun Lu, Yu Wei and Vassilis Zikas

RUHR UNIVERSITÄT BOCHUM

RUB

AARHUS UNIVERSITY

University of Victoria

Georgia Tech College of Computing
School of Cybersecurity and Privacy

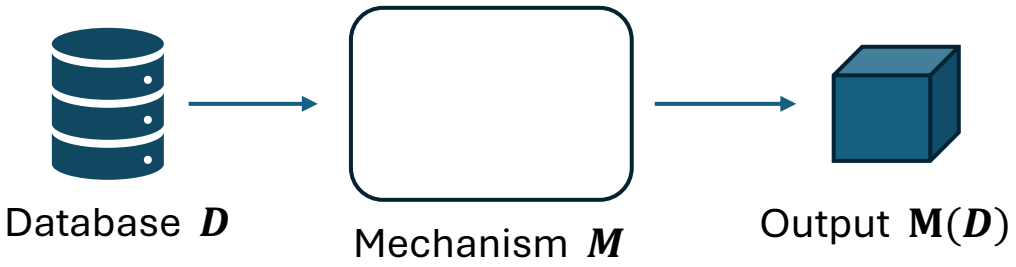
Overview

- f-Differential Privacy
- Estimation
- Auditing

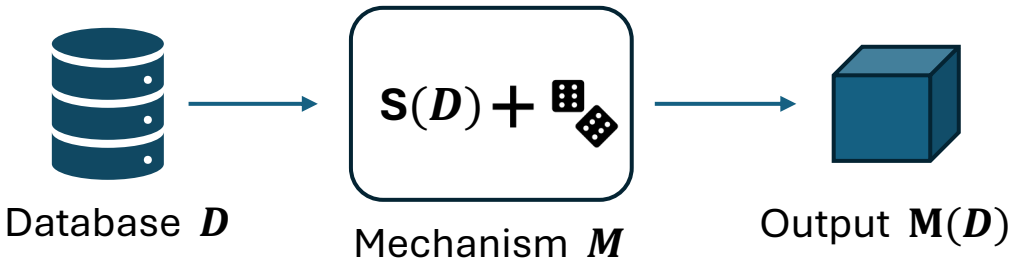
Overview

- f-Differential Privacy
- Estimation
- Auditing

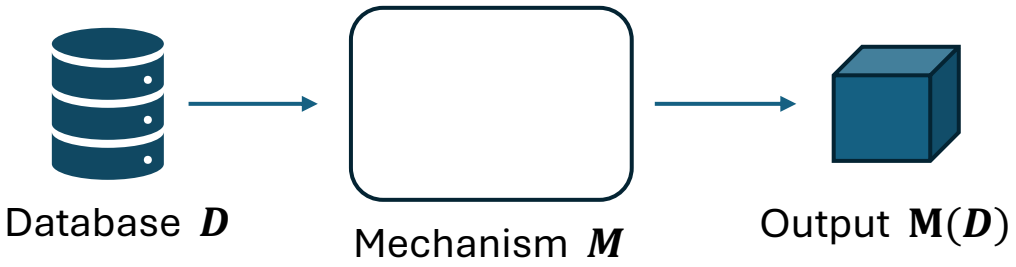
f-Differential Privacy



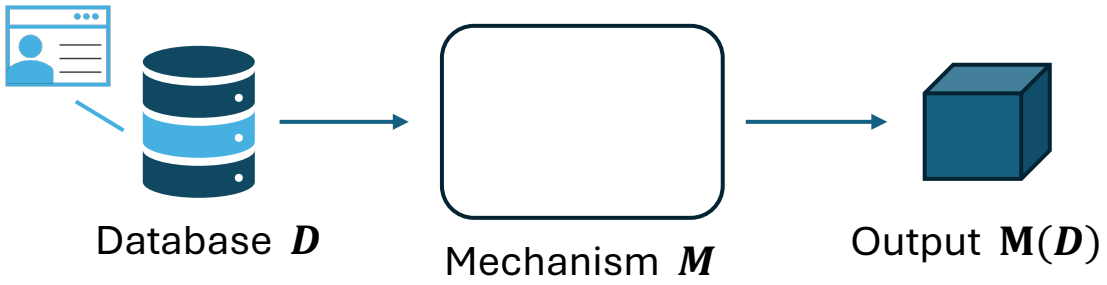
f-Differential Privacy

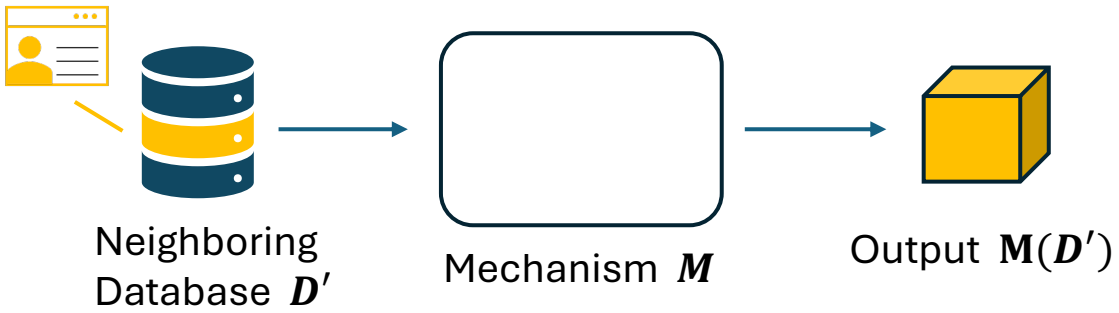


f-Differential Privacy

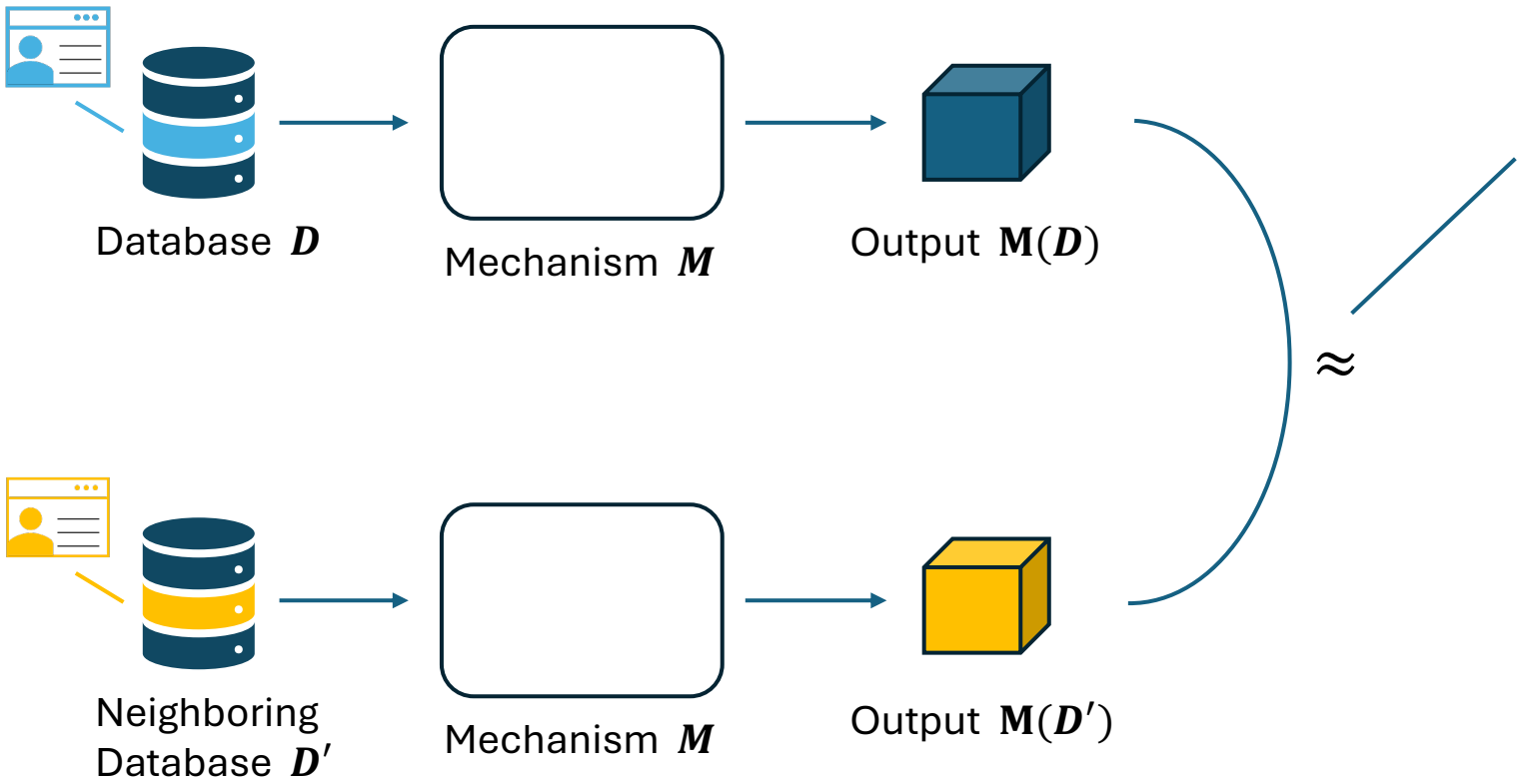


f-Differential Privacy





f-Differential Privacy



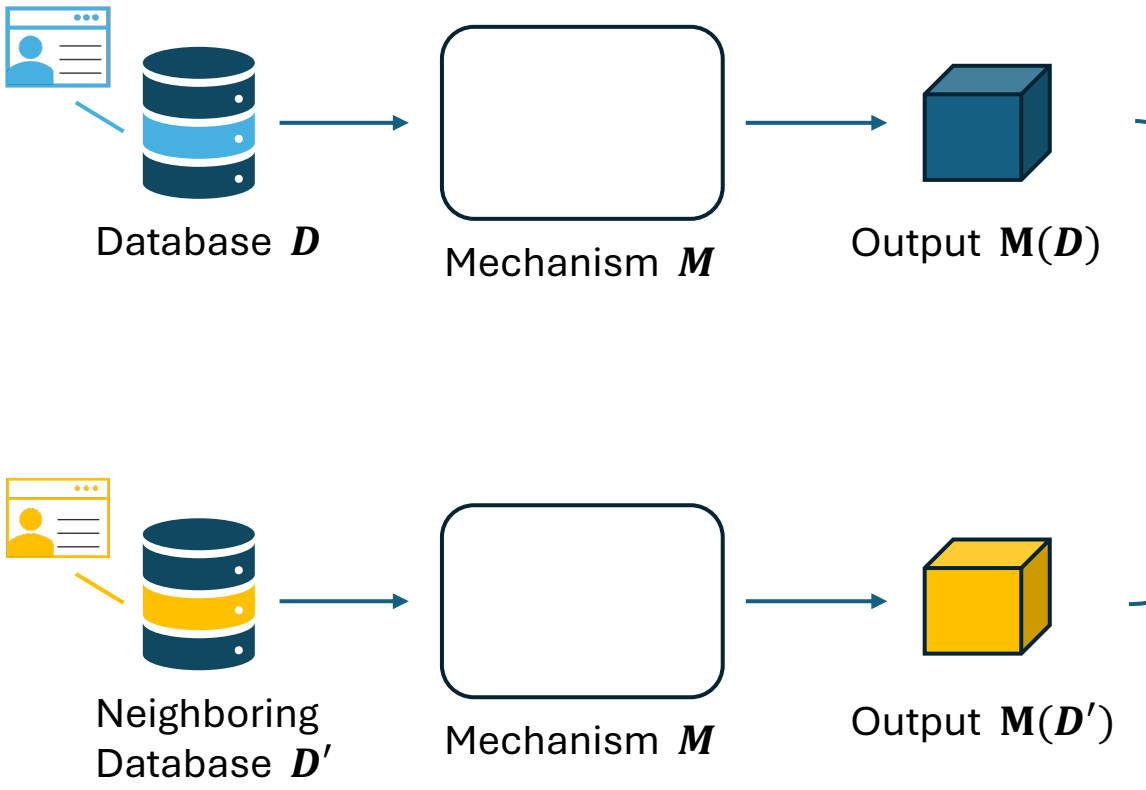
Standard Differential Privacy (DP)

Mechanism M is (ε, δ)-DP, if

$$Pr(M(D) \in S) \leq e^{\varepsilon} Pr(M(D') \in S) + \delta$$

for all neighboring databases D, D' and sets S

f-Differential Privacy



$\approx$

Standard Differential Privacy (DP)

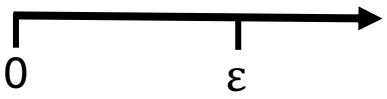
Mechanism M is (ε, δ)-DP, if

$$Pr(M(D) \epsilon\, S) \leq e^{\varepsilon}\, Pr(M(D') \epsilon\, S) + \delta$$

for all neighboring databases D, D' and sets S

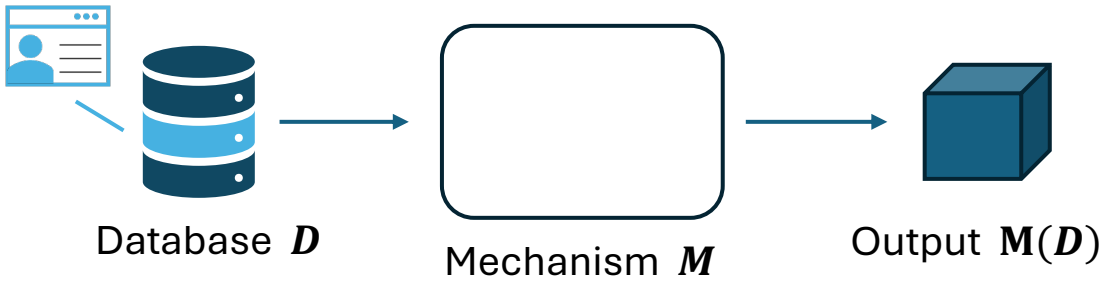
Parameters in Standard DP :

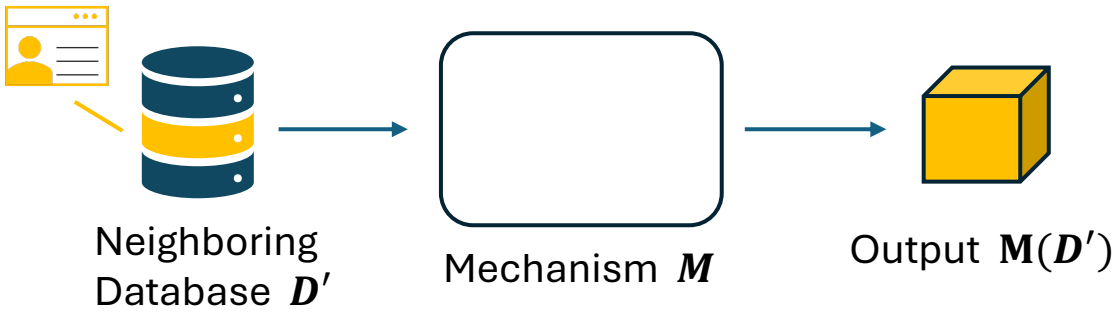
$\varepsilon\, \epsilon\, [0, \infty]$



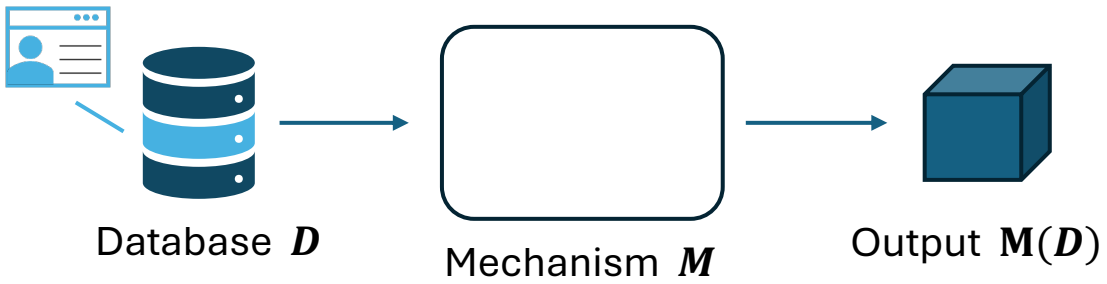
$\delta\, \epsilon\, [0, 1]$

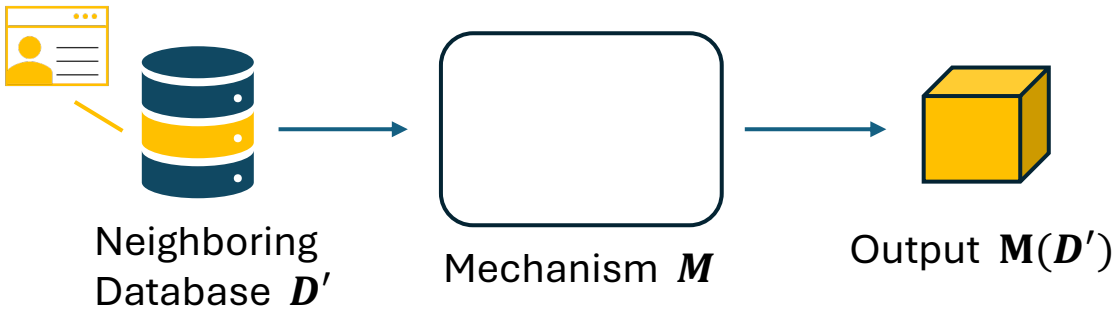
f-Differential Privacy



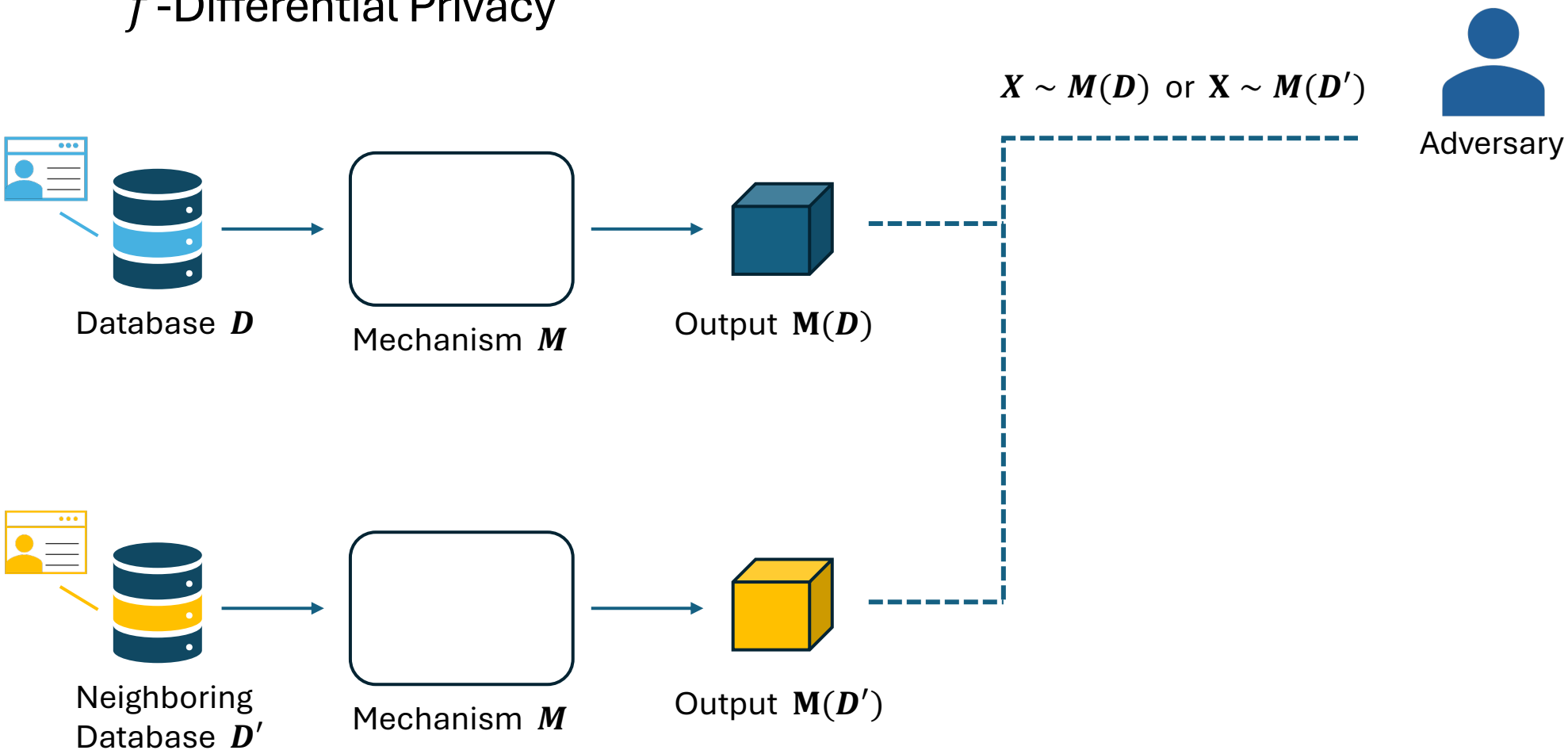


f-Differential Privacy

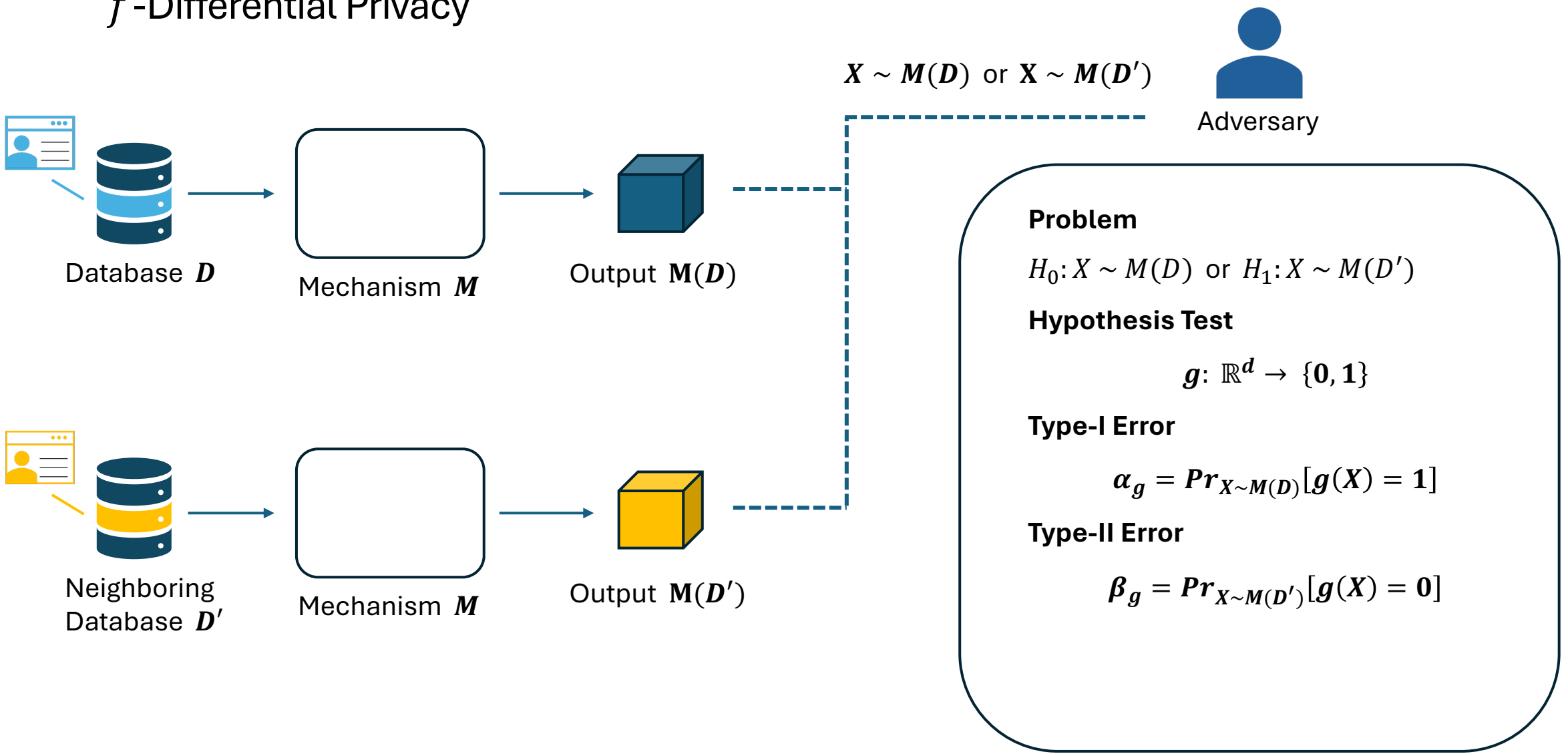




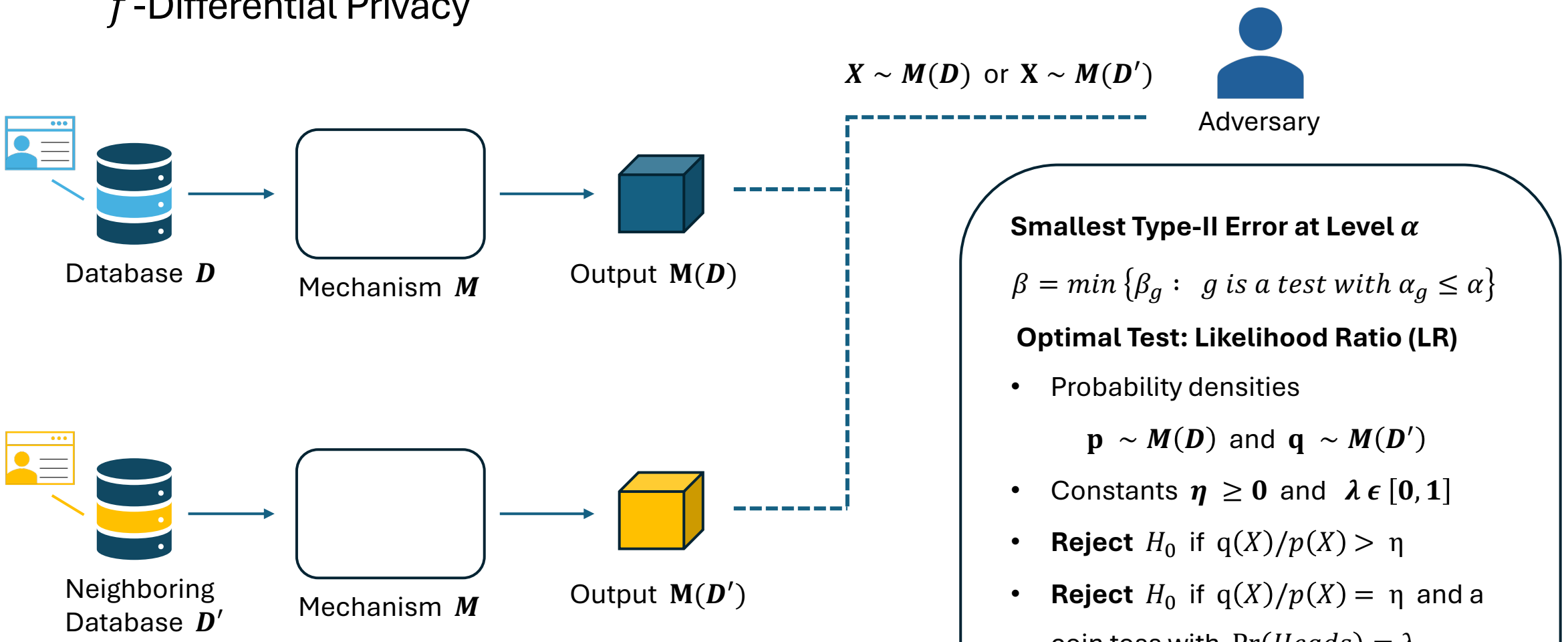
f-Differential Privacy



f-Differential Privacy



f-Differential Privacy



Smallest Type-II Error at Level α

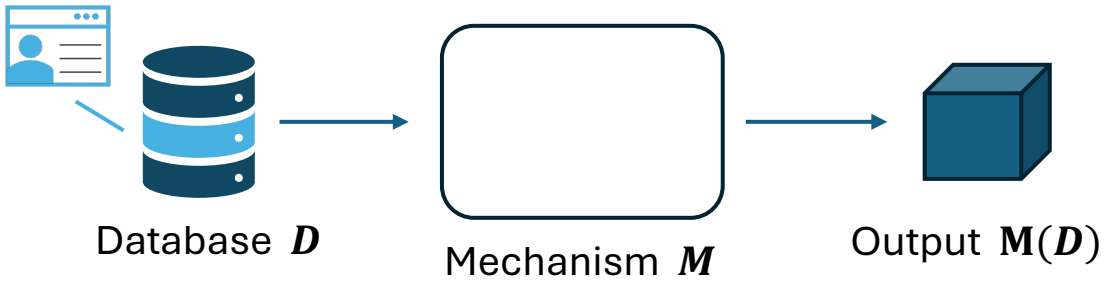
$$\beta = min\{\beta_g : \ g \ is \ a \ test \ with \ \alpha_g \leq \alpha\}$$

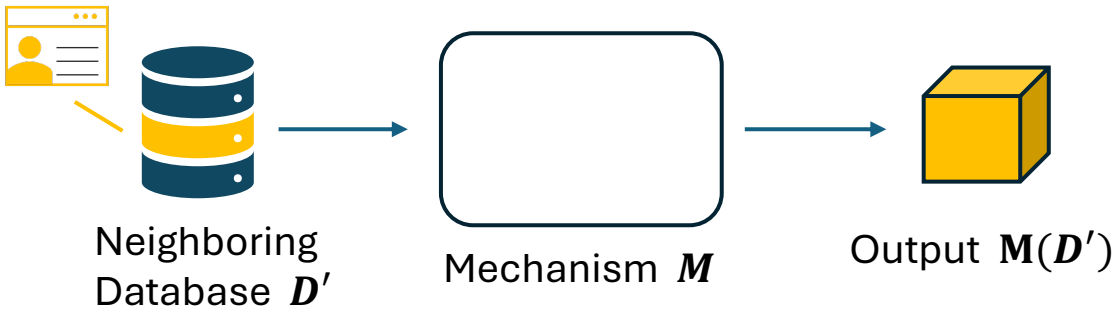
Optimal Test: Likelihood Ratio (LR)

- Probability densities

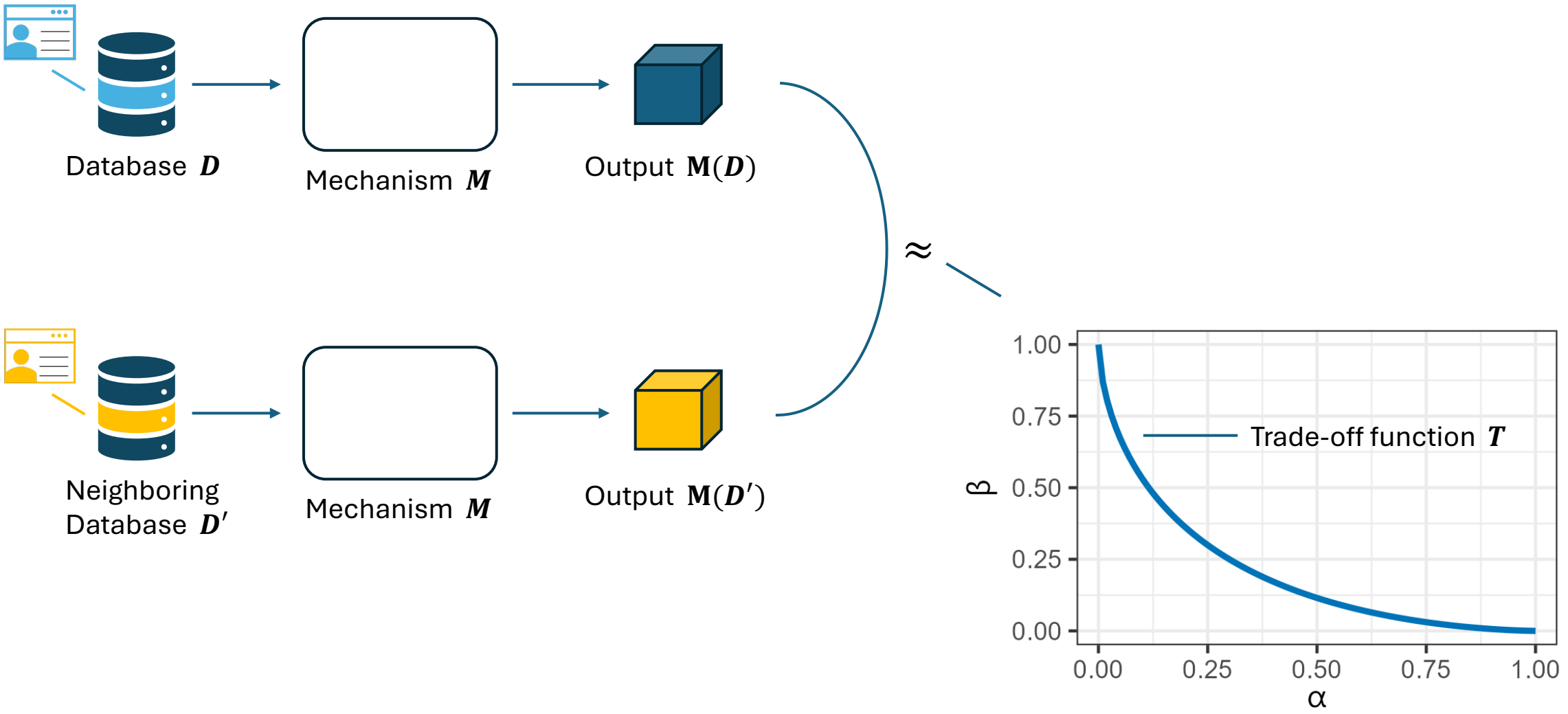
  $\mathbf{p} \sim M(D)$ and $\mathbf{q} \sim M(D')$
- Constants $\eta \geq 0$ and $\lambda \in [0, 1]$
- **Reject** H_0 if $q(X)/p(X) > \eta$
- **Reject** H_0 if $q(X)/p(X) = \eta$ and a coin toss with $\Pr(Heads) = \lambda$ yields **Heads**

f-Differential Privacy

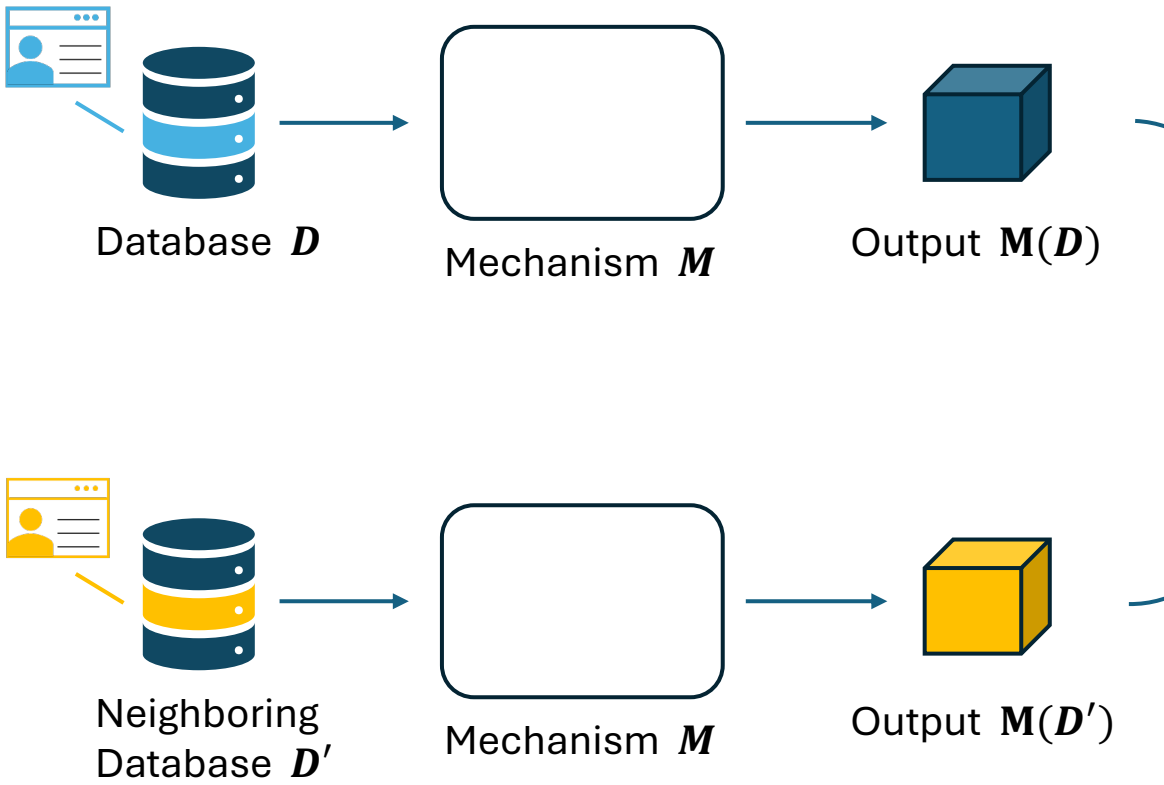




f-Differential Privacy



f-Differential Privacy



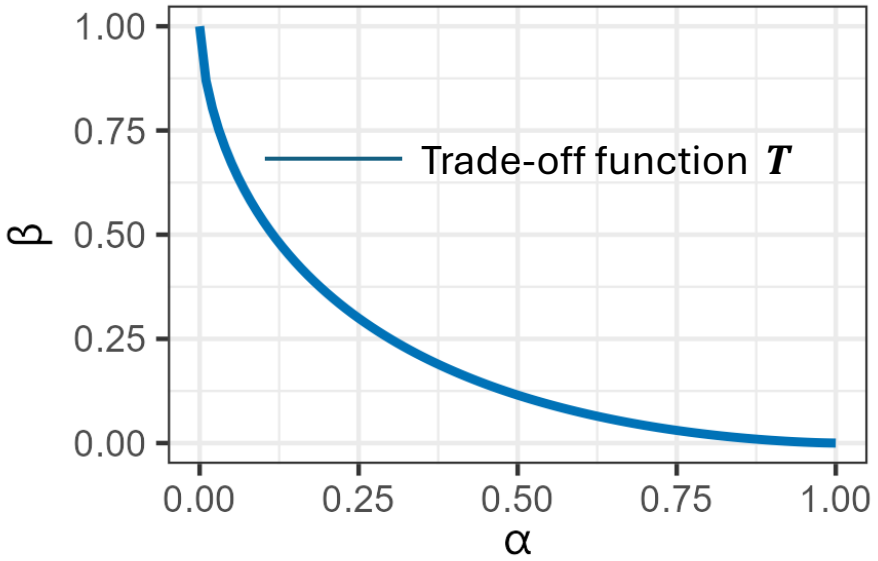
≈

f-Differential Privacy (f-DP)

Mechanism M is f-DP, if for all neighboring databases D, D' we have

$$T(\alpha) \geq f(\alpha) \quad \forall\, \alpha \in [0,1],$$

where T is implied by $p \sim M(D)$ and $q \sim M(D')$.



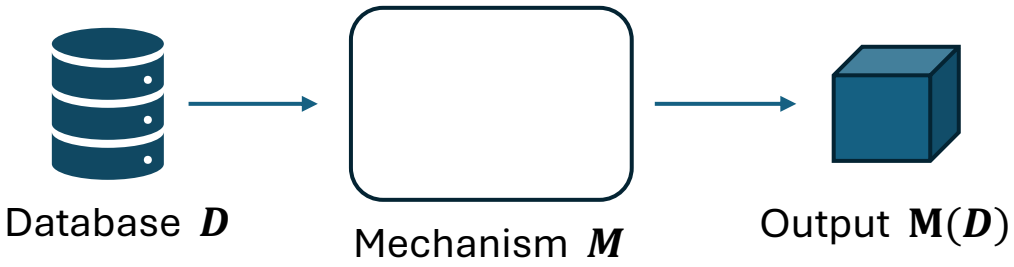
f-Differential Privacy

- comes with easily interpretable privacy guarantees.
- allows for lossless reasoning about composition.
- is suitable for studying privacy amplification.
- is a popular framework for auditing private machine learning.

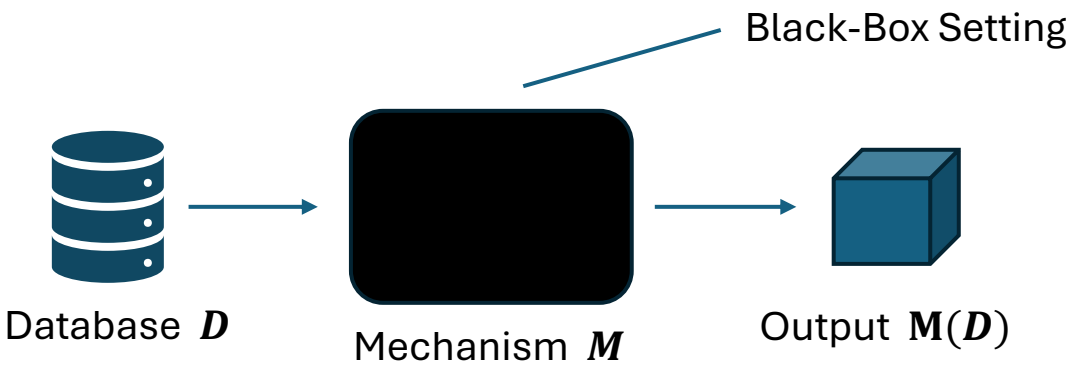
Overview

- f-Differential Privacy
- Estimation
- Auditing

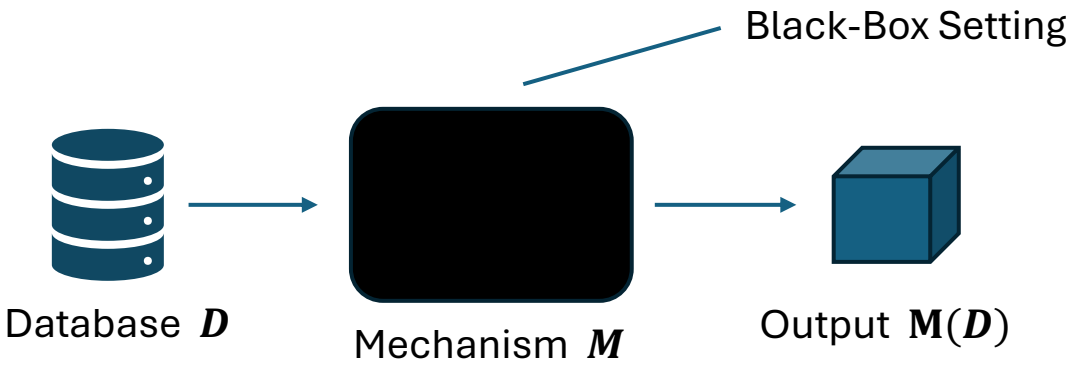
Estimation


Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$

Estimation


Black-Box Setting
Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$

Estimation

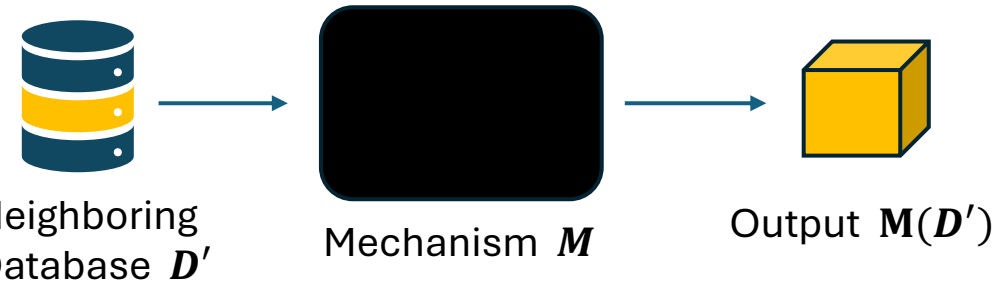


Auditor

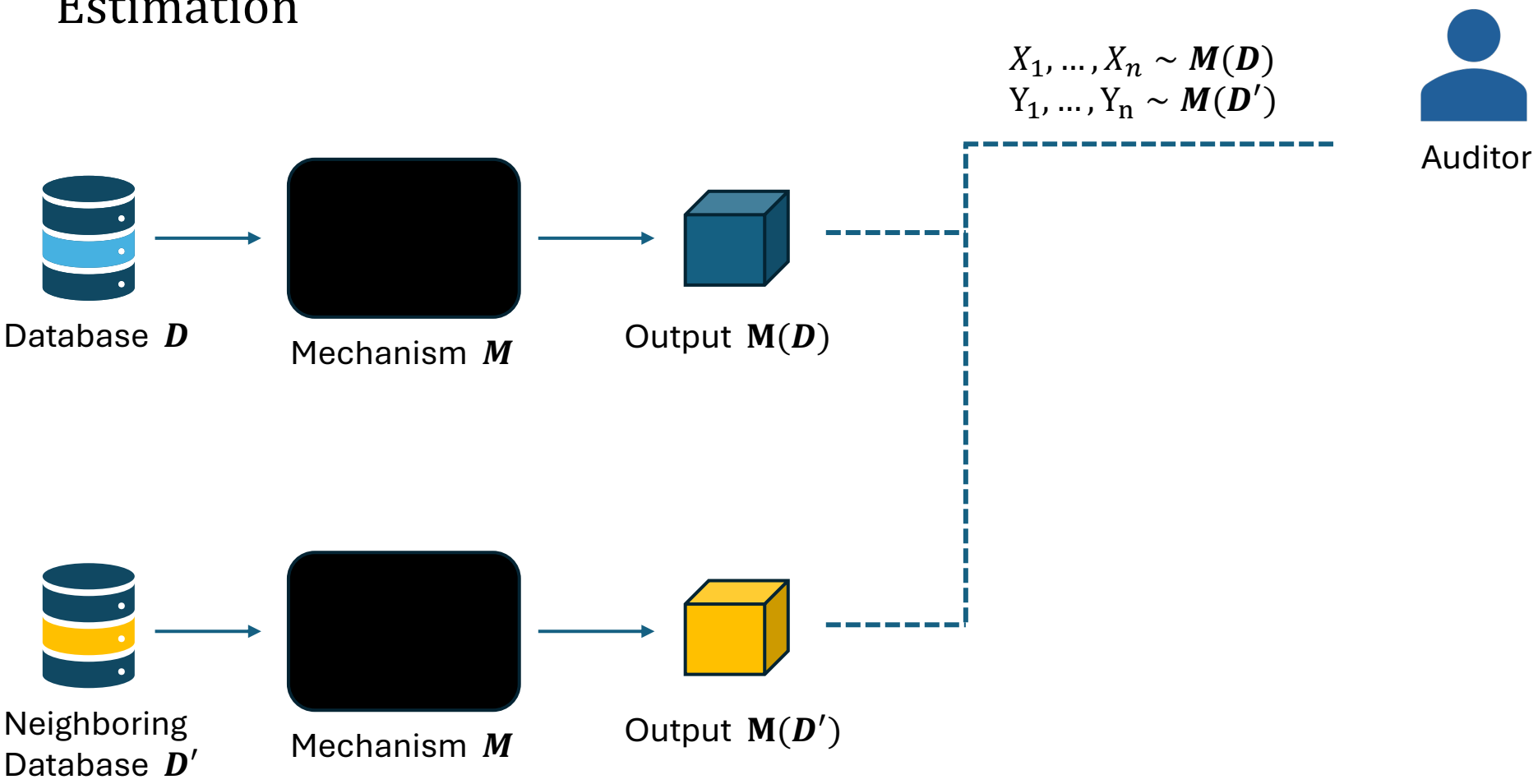
Estimation


Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$

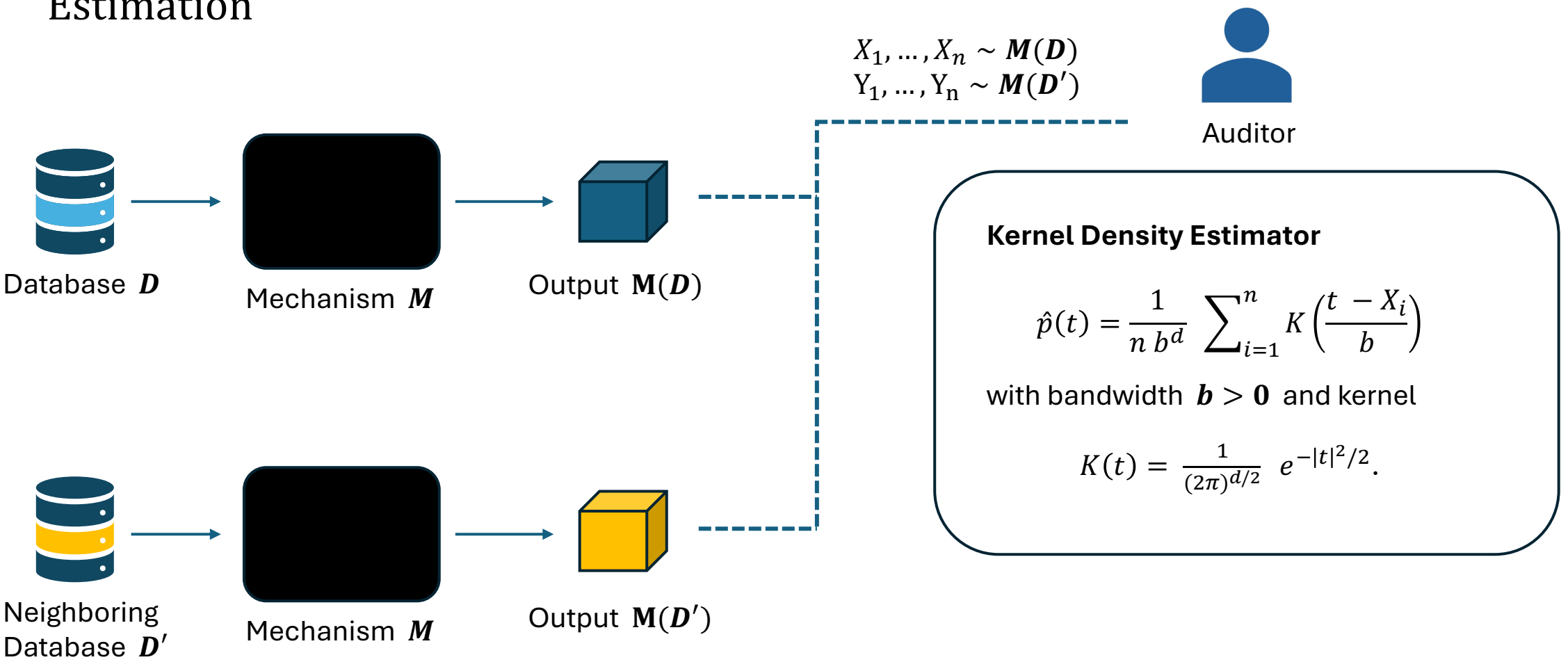

Auditor


Neighboring Database $\boldsymbol{D}'$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D}')$

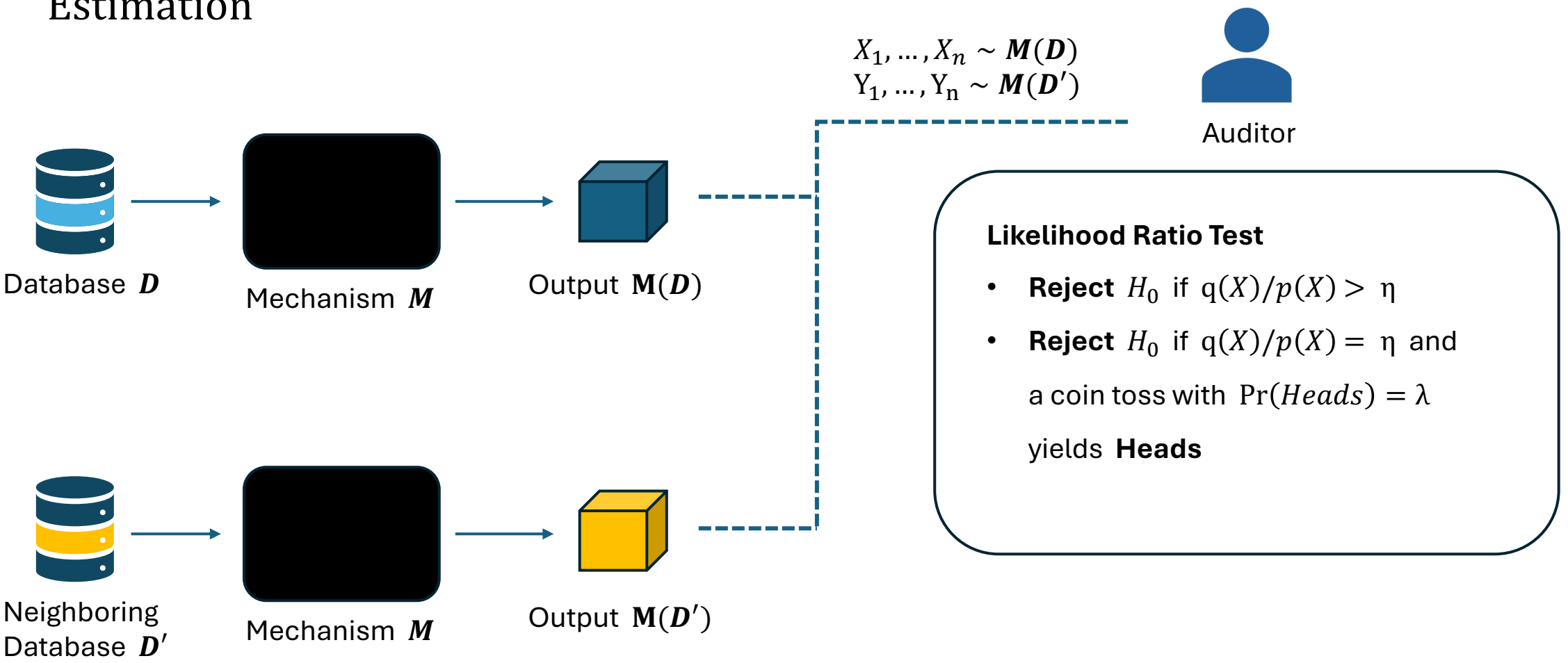
Estimation


$X_1, \dots, X_n \sim \boldsymbol{M}(\boldsymbol{D})$
$\mathrm{Y}_1, \dots, \mathrm{Y}_\mathrm{n} \sim \boldsymbol{M}(\boldsymbol{D}')$
Auditor
Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$
Neighboring Database $\boldsymbol{D}'$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D}')$

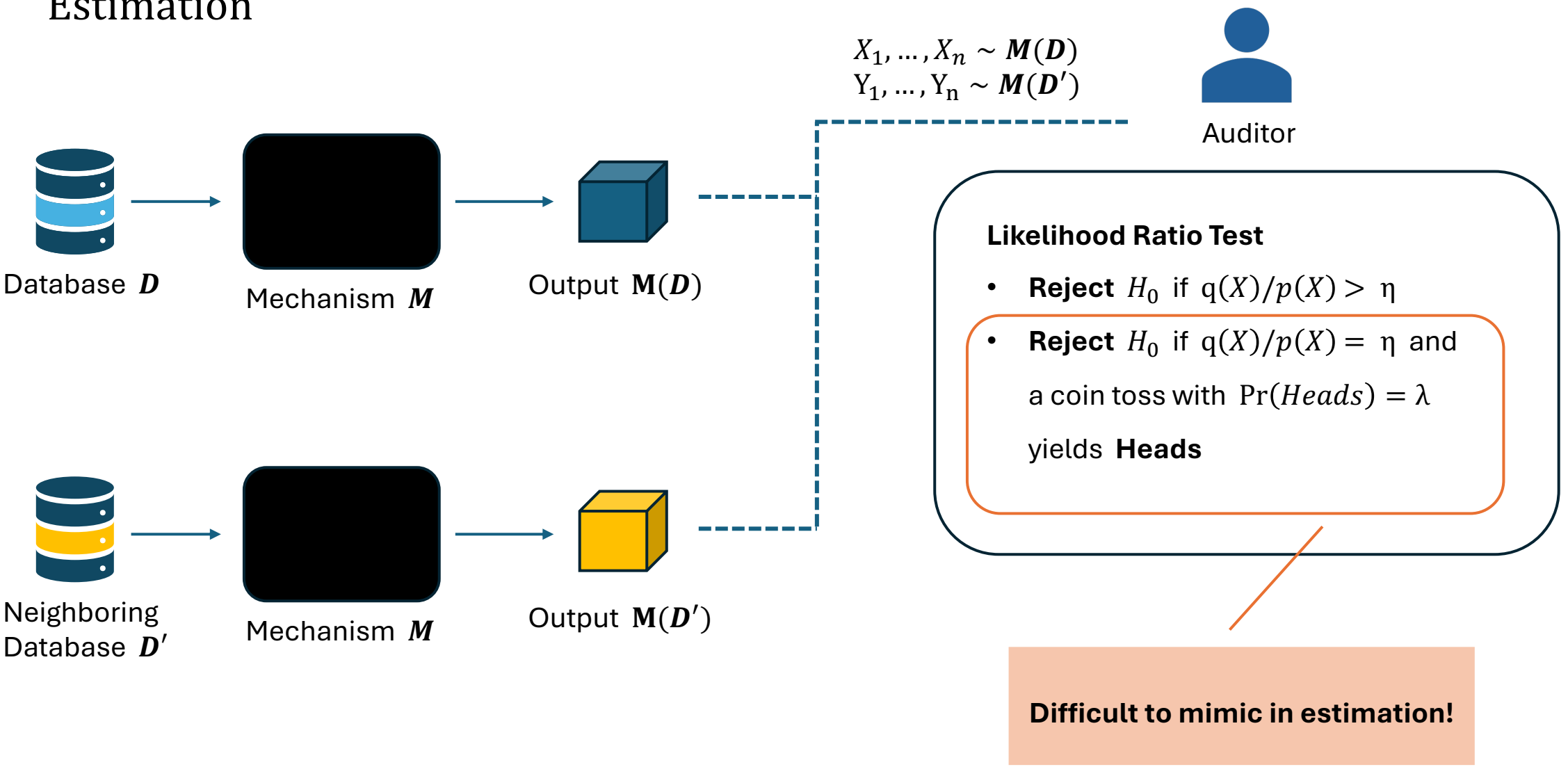
Estimation


$X_1, \dots, X_n \sim \boldsymbol{M}(\boldsymbol{D})$
$Y_1, \dots, Y_n \sim \boldsymbol{M}(\boldsymbol{D}')$
Auditor
Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$
Neighboring Database $\boldsymbol{D}'$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D}')$
Kernel Density Estimator
$\hat{p}(t) = \frac{1}{n\, b^d} \sum_{i=1}^{n} K\left(\frac{t - X_i}{b}\right)$
with bandwidth $\boldsymbol{b} > \mathbf{0}$ and kernel
$K(t) = \frac{1}{(2\pi)^{d/2}}\ e^{-|t|^2/2}.$

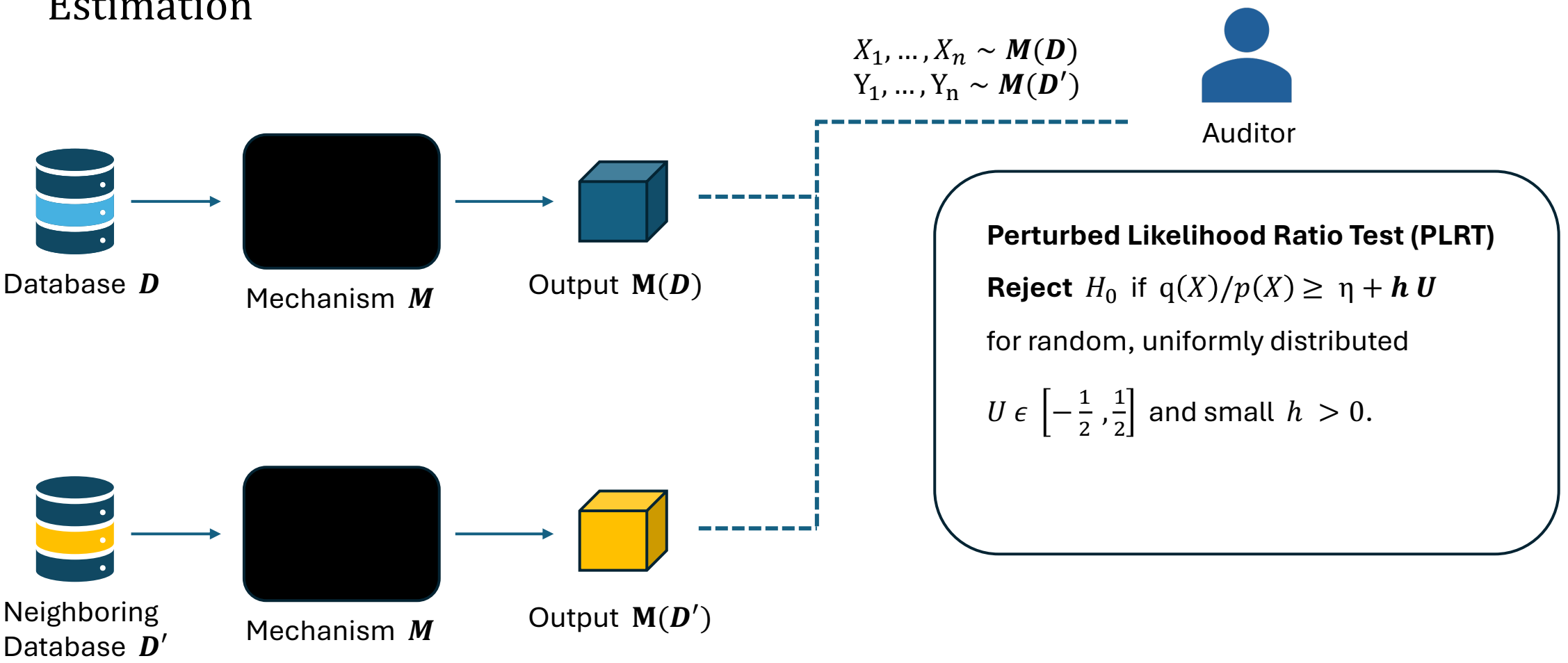
Estimation


$X_1, \dots, X_n \sim \boldsymbol{M}(\boldsymbol{D})$
$Y_1, \dots, Y_n \sim \boldsymbol{M}(\boldsymbol{D}')$
Auditor
Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$
Neighboring Database $\boldsymbol{D}'$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D}')$
Likelihood Ratio Test
• **Reject** H_0 if $q(X)/p(X) > \eta$
• **Reject** H_0 if $q(X)/p(X) = \eta$ and a coin toss with $\Pr(Heads) = \lambda$ yields **Heads**

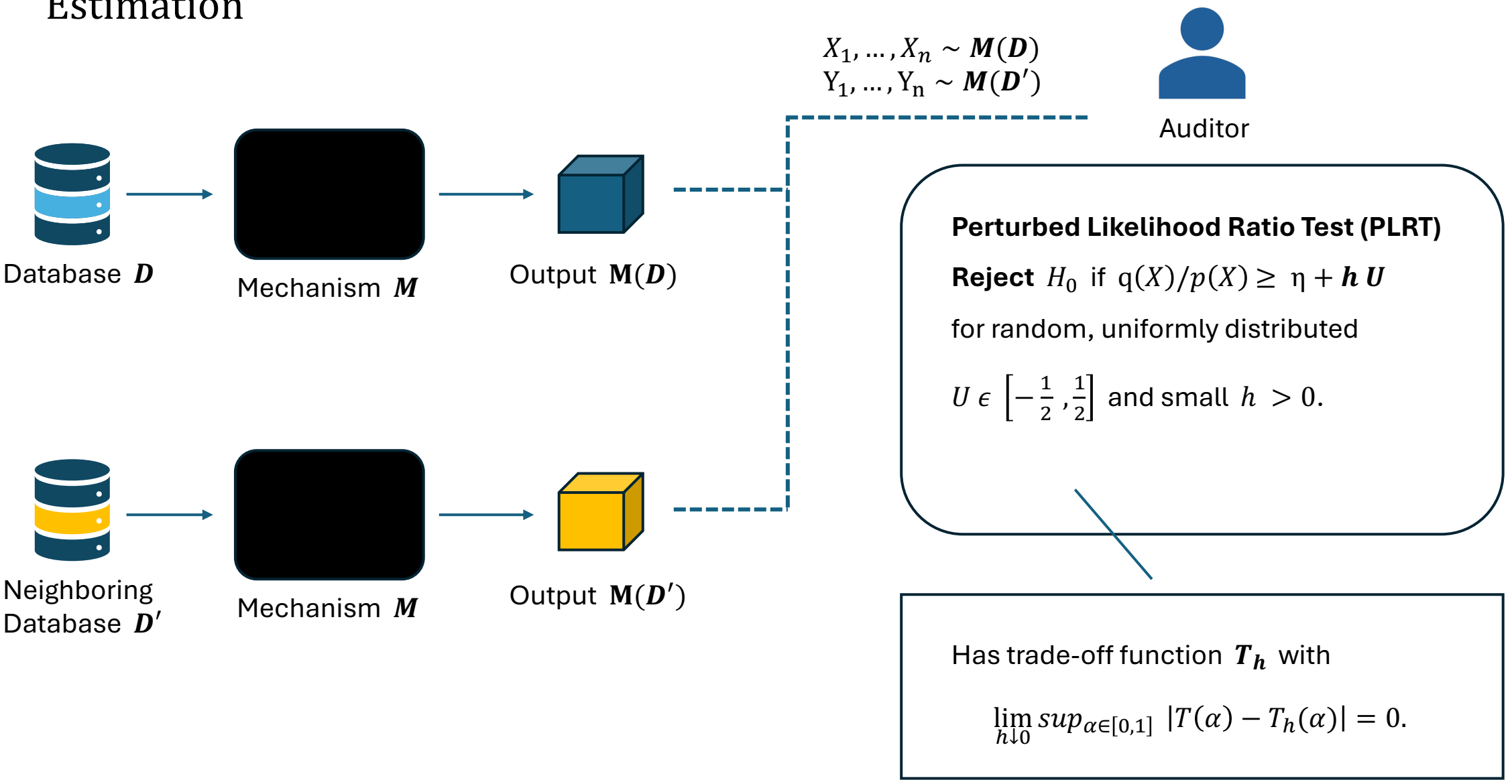
Estimation


$X_1, \dots, X_n \sim \boldsymbol{M}(\boldsymbol{D})$
$Y_1, \dots, Y_n \sim \boldsymbol{M}(\boldsymbol{D}')$
Auditor
Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$
Neighboring Database $\boldsymbol{D}'$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D}')$
Likelihood Ratio Test
• **Reject** H_0 if $q(X)/p(X) > \eta$
• **Reject** H_0 if $q(X)/p(X) = \eta$ and a coin toss with $\Pr(Heads) = \lambda$ yields **Heads**
Difficult to mimic in estimation!

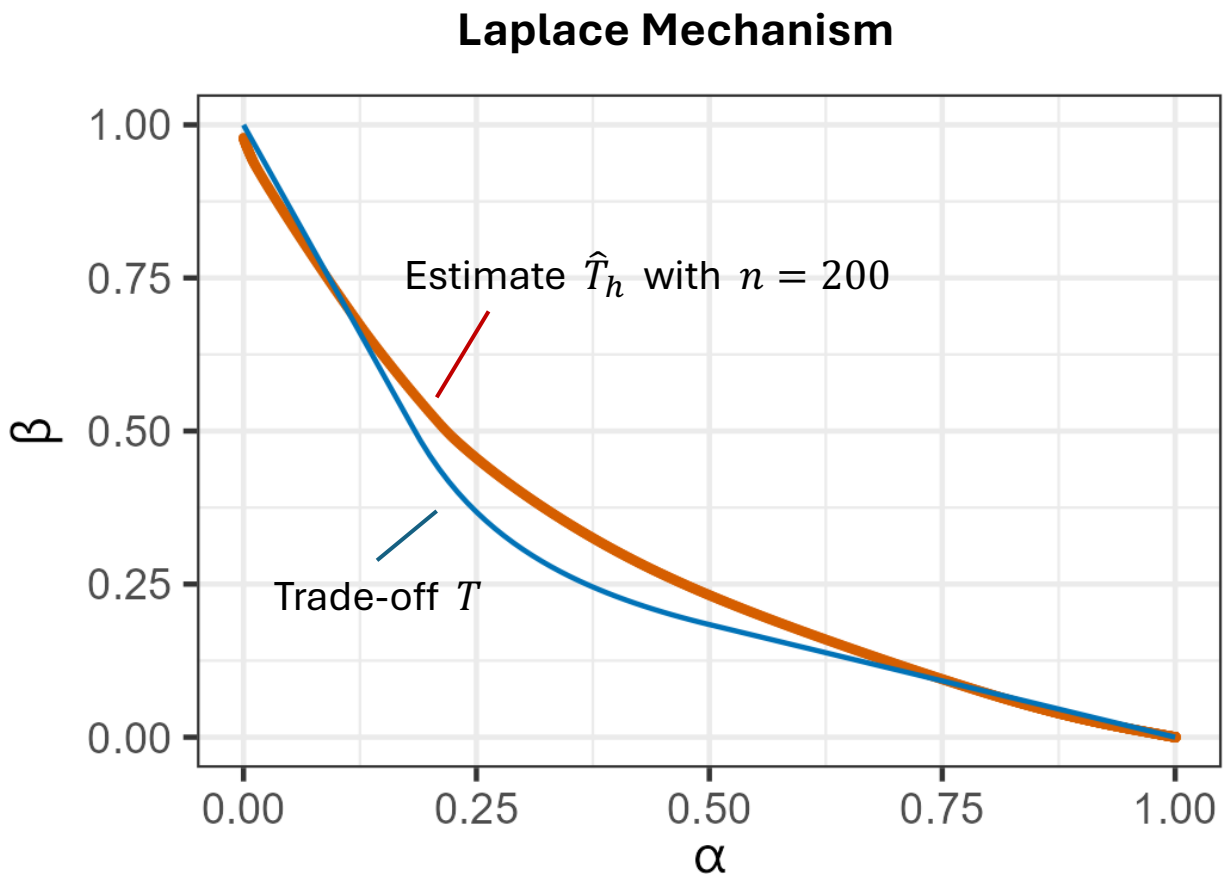
Estimation


$X_1, \ldots, X_n \sim \boldsymbol{M}(\boldsymbol{D})$
$Y_1, \ldots, Y_n \sim \boldsymbol{M}(\boldsymbol{D}')$
Auditor
Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$
Neighboring Database $\boldsymbol{D}'$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D}')$
Perturbed Likelihood Ratio Test (PLRT)
Reject H_0 if $q(X)/p(X) \geq \eta + \boldsymbol{h}\,\boldsymbol{U}$
for random, uniformly distributed
$U \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ and small $h > 0$.

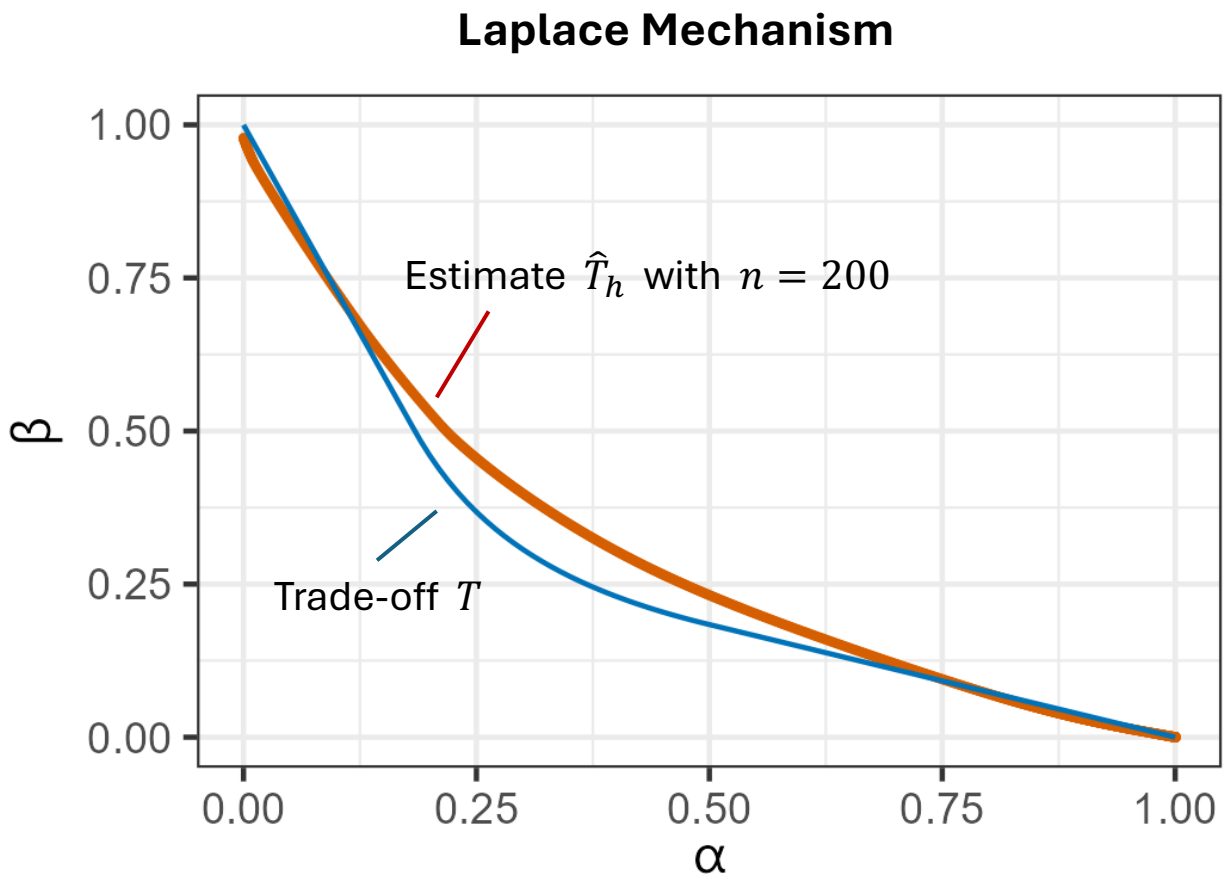
Estimation


$X_1, \dots, X_n \sim \boldsymbol{M}(\boldsymbol{D})$
$Y_1, \dots, Y_n \sim \boldsymbol{M}(\boldsymbol{D}')$
Auditor
Database $\boldsymbol{D}$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D})$
Neighboring Database $\boldsymbol{D}'$
Mechanism $\boldsymbol{M}$
Output $\mathbf{M}(\boldsymbol{D}')$
Perturbed Likelihood Ratio Test (PLRT)
Reject H_0 if $q(X)/p(X) \geq \eta + \boldsymbol{h}\,\boldsymbol{U}$
for random, uniformly distributed
$U \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ and small $h > 0$.
Has trade-off function $\boldsymbol{T_h}$ with
$\lim_{h \downarrow 0} sup_{\alpha \in [0,1]} |T(\alpha) - T_h(\alpha)| = 0.$

Estimation


Laplace Mechanism
1.00
0.75
0.50
0.25
0.00
β
0.00
0.25
0.50
0.75
1.00
α
Estimate $\hat{T}_h$ with $n = 200$
Trade-off T

Estimation



Points of $\hat{T}_h$

$$\hat{\alpha}_h(\eta) = \int_{x\epsilon\left[-\frac{h}{2},\frac{h}{2}\right]} \frac{1}{h} \int_{\frac{\hat{q}}{\hat{p}} > \eta + x} \hat{p}$$

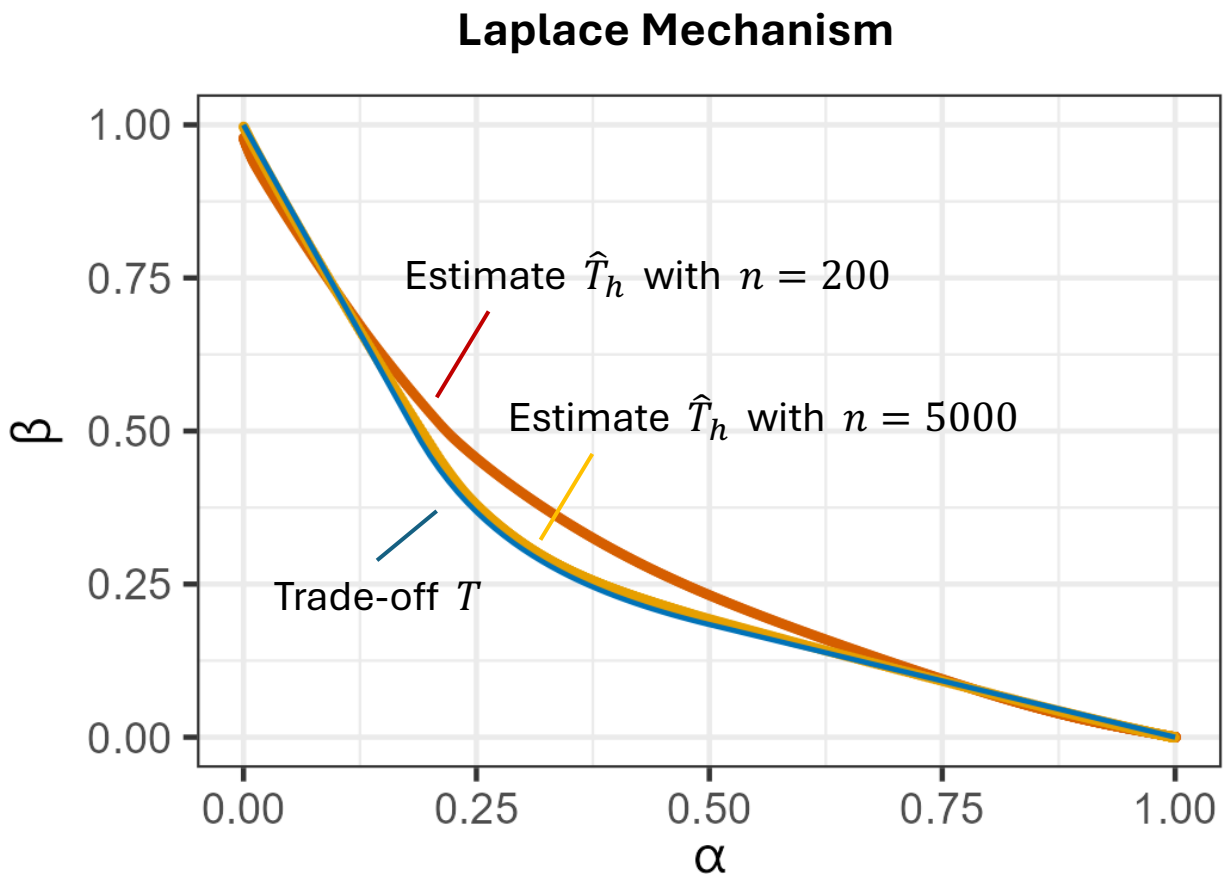
$$\hat{\beta}_h(\eta) = 1 - \int_{x\epsilon\left[-\frac{h}{2},\frac{h}{2}\right]} \frac{1}{h} \int_{\frac{\hat{q}}{\hat{p}} > \eta + x} \hat{q}$$

Uniform Convergence

Under certain regularity assumptions, we have for $n \to \infty$ and $h = h_n \to 0$ that

$$sup_{\alpha\in[0,1]} \left|T(\alpha) - \hat{T}_h(\alpha)\right| = o_P(1).$$

Estimation



Points of $\widehat{T}_h$

$$\hat{\alpha}_h(\eta) = \int_{x\epsilon\left[-\frac{h}{2},\frac{h}{2}\right]} \frac{1}{h} \int_{\frac{\hat{q}}{\hat{p}} > \eta + x} \hat{p}$$

$$\hat{\beta}_h(\eta) = 1 - \int_{x\epsilon\left[-\frac{h}{2},\frac{h}{2}\right]} \frac{1}{h} \int_{\frac{\hat{q}}{\hat{p}} > \eta + x} \hat{q}$$

Uniform Convergence

Under certain regularity assumptions, we have for $n \to \infty$ and $h = h_n \to 0$ that

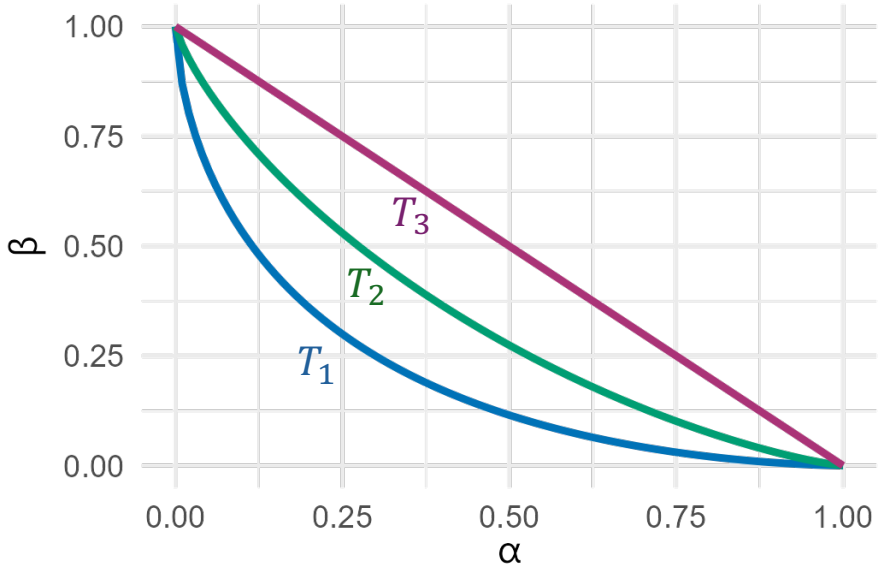
$$sup_{\alpha\in[0,1]} \left|T(\alpha) - \hat{T}_h(\alpha)\right| = o_P(1).$$

Overview

- f-Differential Privacy
- Estimation
- Auditing

Auditing

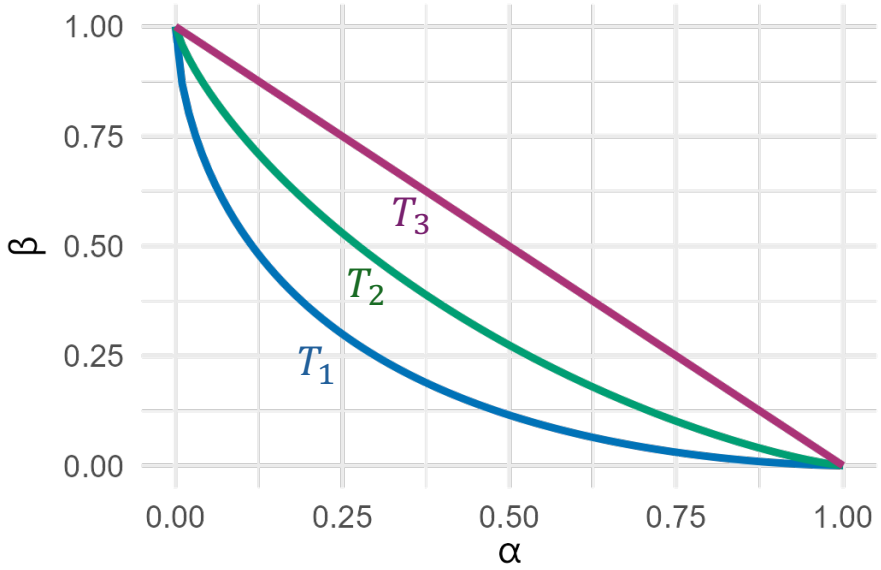
Privacy Parameters in f-DP



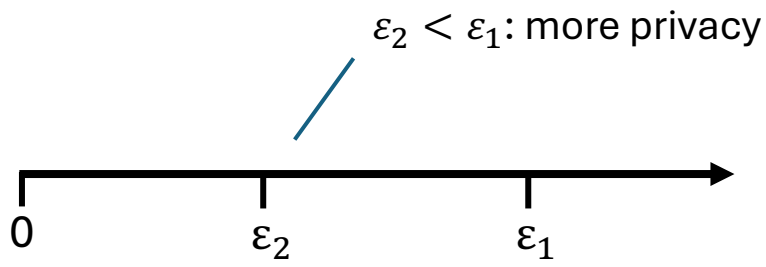
Privacy Parameters in Standard DP

Auditing

Privacy Parameters in f-DP

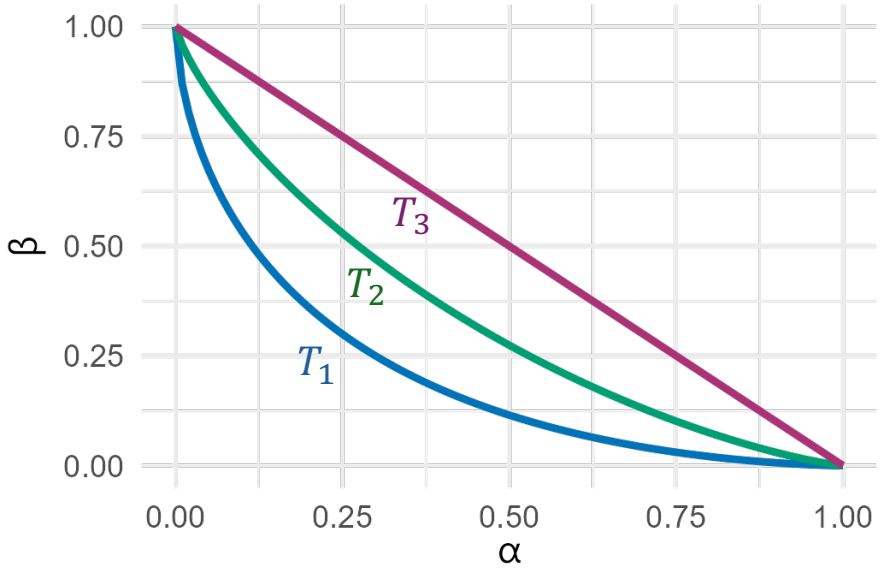


Privacy Parameters in Standard DP

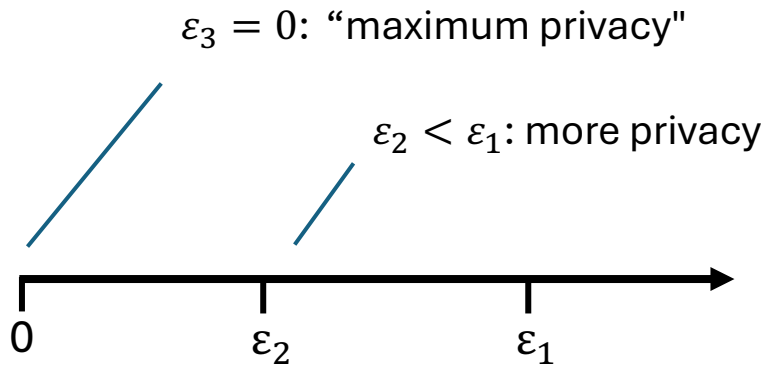


Auditing

Privacy Parameters in f-DP



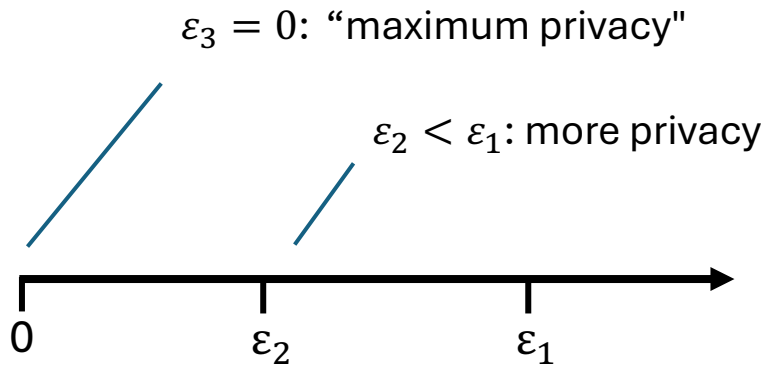
Privacy Parameters in Standard DP



Auditing

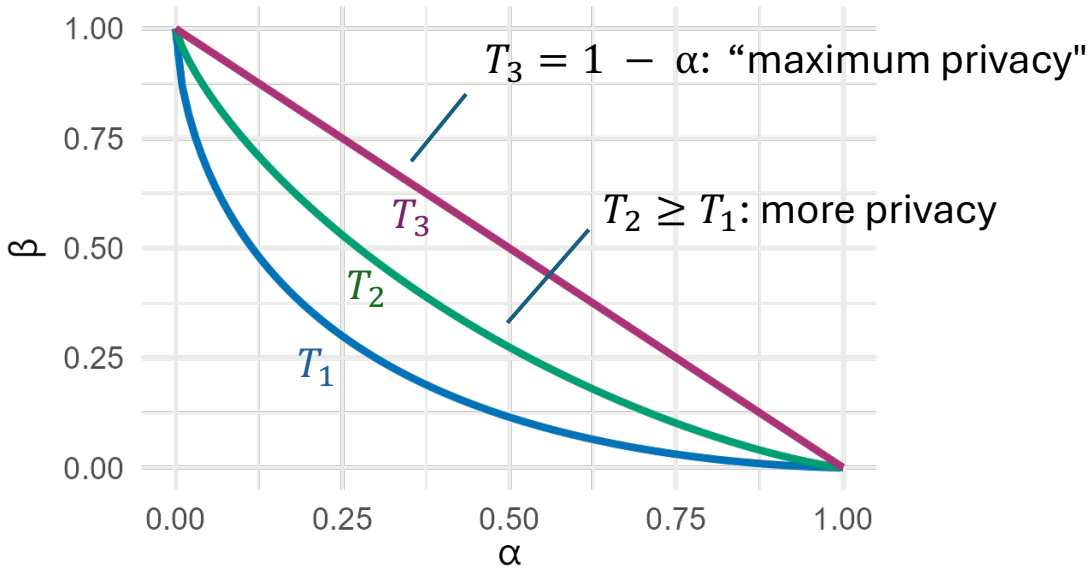
Privacy Parameters in f-DP

Privacy Parameters in Standard DP

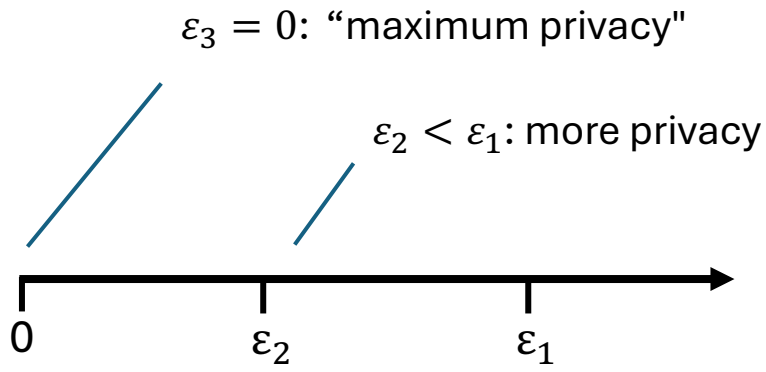


Auditing

Privacy Parameters in f-DP



Privacy Parameters in Standard DP



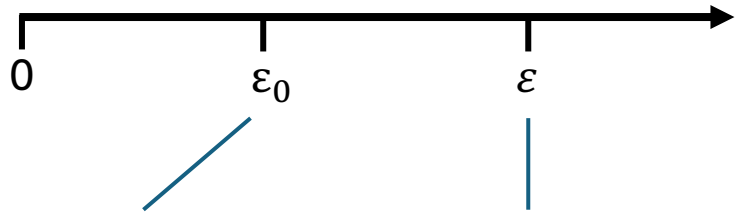
Auditing

Goal: Detect violations of privacy and expose flawed mechanisms

Auditing

Goal: Detect violations of privacy and expose flawed mechanisms

Privacy Violation in Standard DP

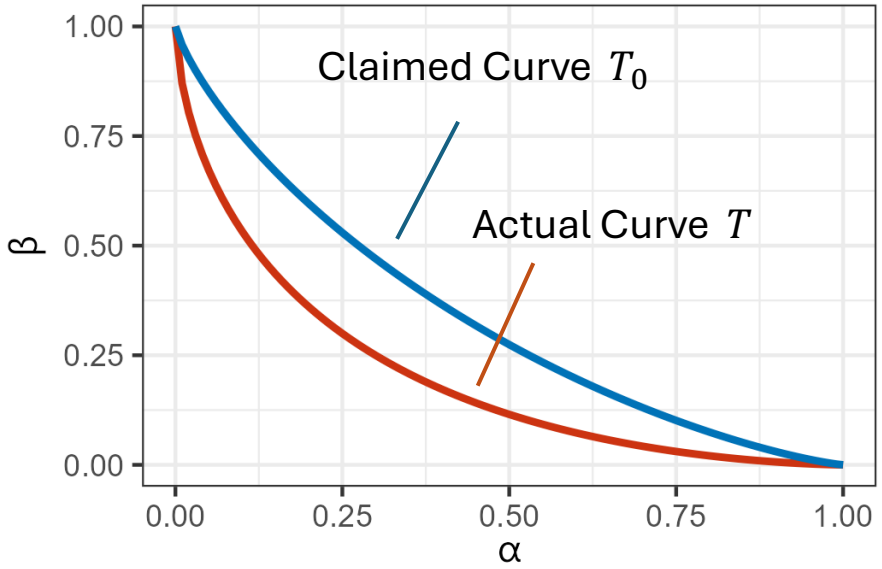


Privacy Claim ε_0 Actual Parameter ε

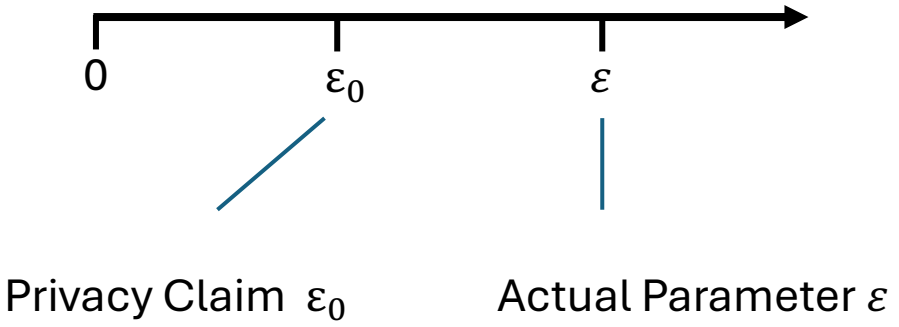
Auditing

Goal: Detect violations of privacy and expose flawed mechanisms

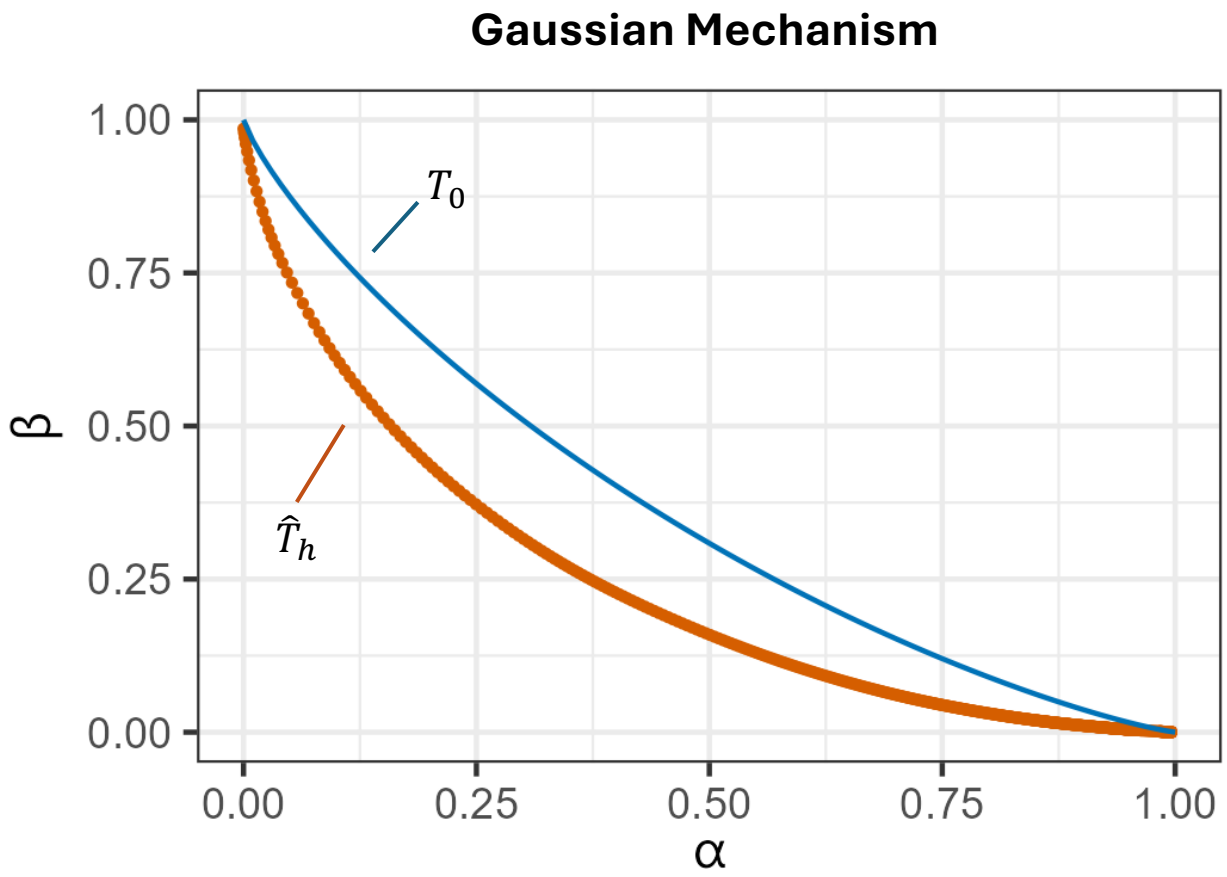
Privacy Violation in f-DP



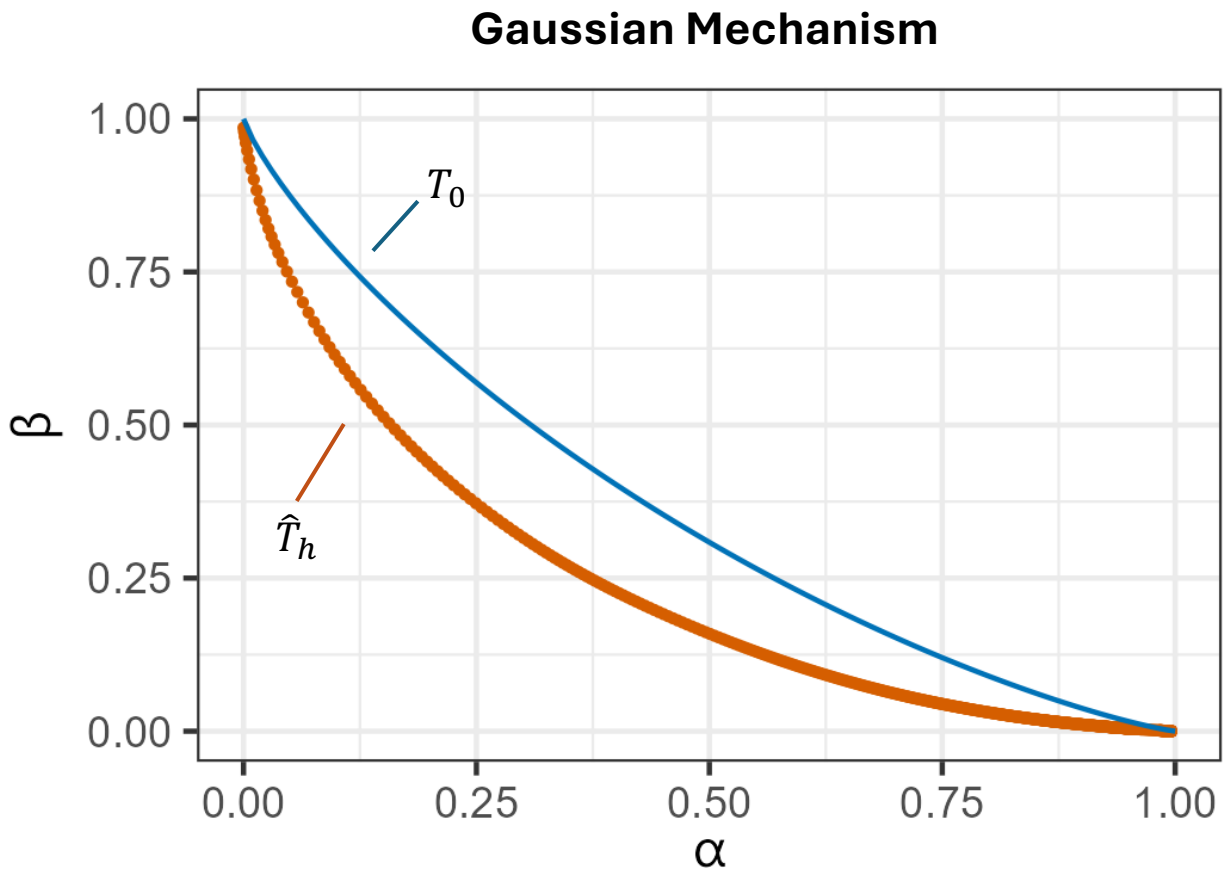
Privacy Violation in Standard DP



Auditing


Gaussian Mechanism
1.00
0.75
0.50
0.25
0.00
β
0.00
0.25
0.50
0.75
1.00
α
T_0
$\hat{T}_h$

Auditing



For fixed $\eta > 0$

- $\alpha(\eta)$ and $\beta(\eta)$ can be expressed as the Bayes risk of a special classification problem.
- Let $\tilde{\alpha}(\eta)$ and $\tilde{\beta}(\eta)$ be the estimates obtained via the empricial risk of the k-nearest neighbor classifier (k-NN).

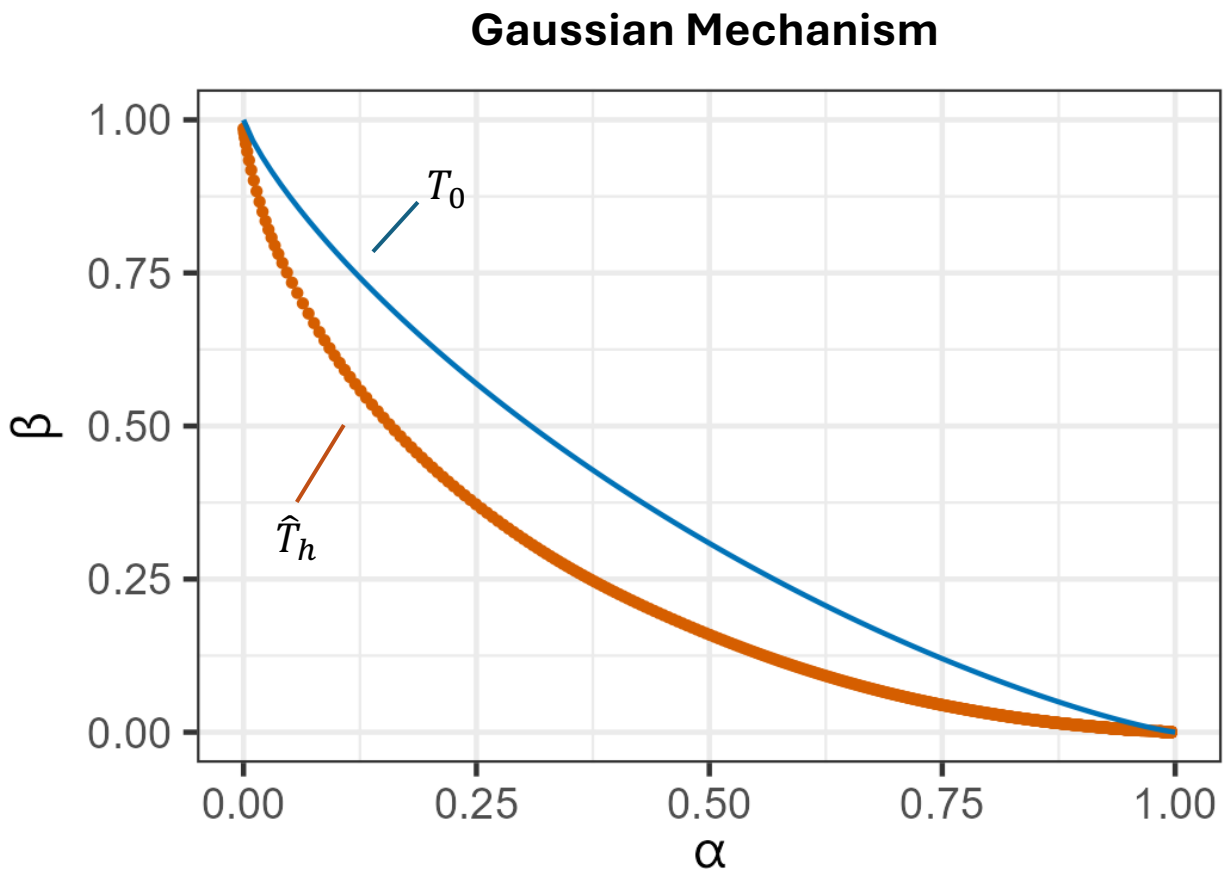
Confidence Interval at η

For $\gamma \in (0,1)$ and $n \geq 2$ it holds with probability greater than $1 - \gamma$ that

$$max\{|\tilde{\alpha}(\eta) - \mathbb{E}[\tilde{\alpha}(\eta)]|, |\tilde{\beta}(\eta) - \mathbb{E}[\tilde{\beta}(\eta)]|\} \leq w(\gamma)$$

where $w(\gamma) = \sqrt{\ln(4/\gamma)}/\sqrt{2n}$.

Auditing



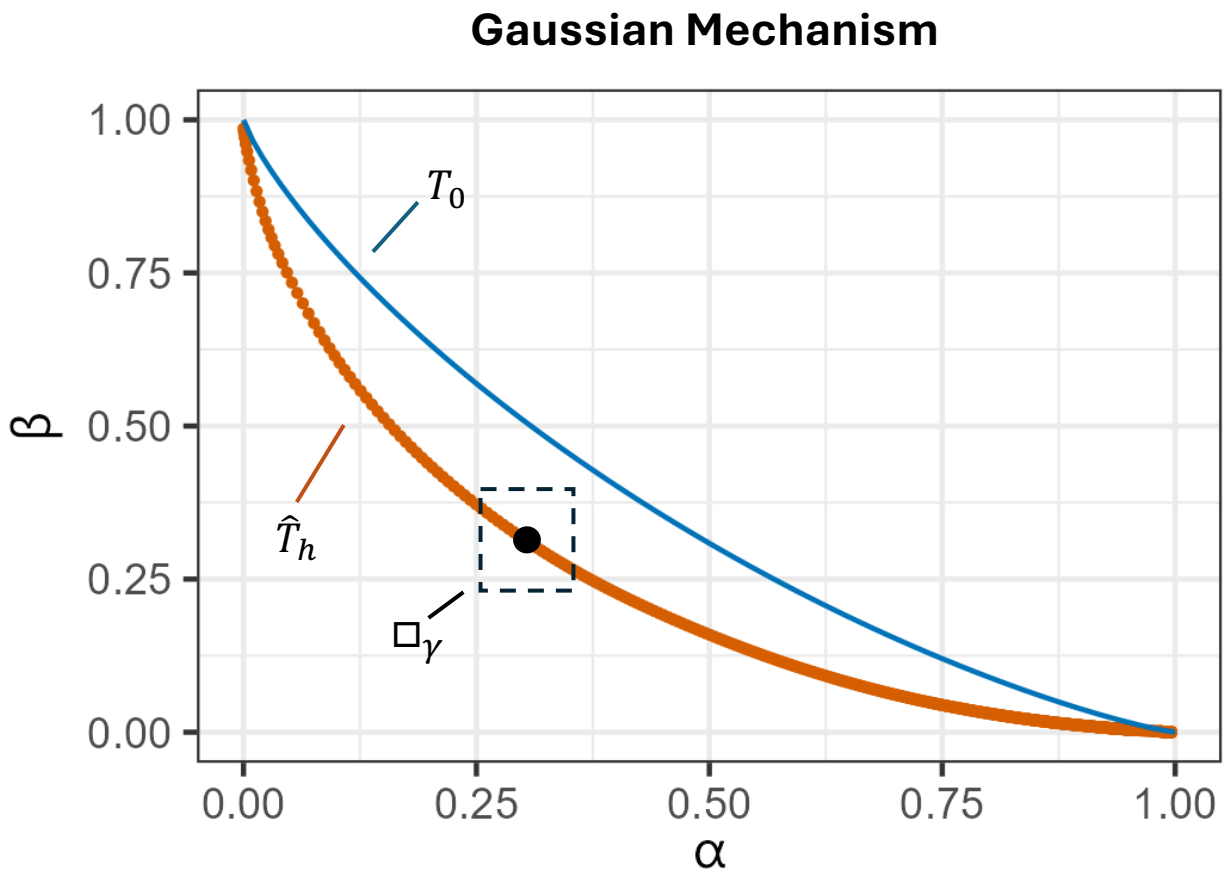
Construct Confidence Region

(1) Compute $\hat{\eta}^*$ where distance between $\hat{T}_h(\hat{\alpha}_h(\eta))$ and $\mathrm{T}_0(\hat{\alpha}_h(\eta))$ is largest.

(2) Compute

$$\square_\gamma = [\tilde{\alpha}(\hat{\eta}^*) - w(\gamma), \tilde{\alpha}(\hat{\eta}^*) + w(\gamma)] \times [\tilde{\beta}(\hat{\eta}^*) - w(\gamma), \tilde{\beta}(\hat{\eta}^*) + w(\gamma)].$$

Auditing



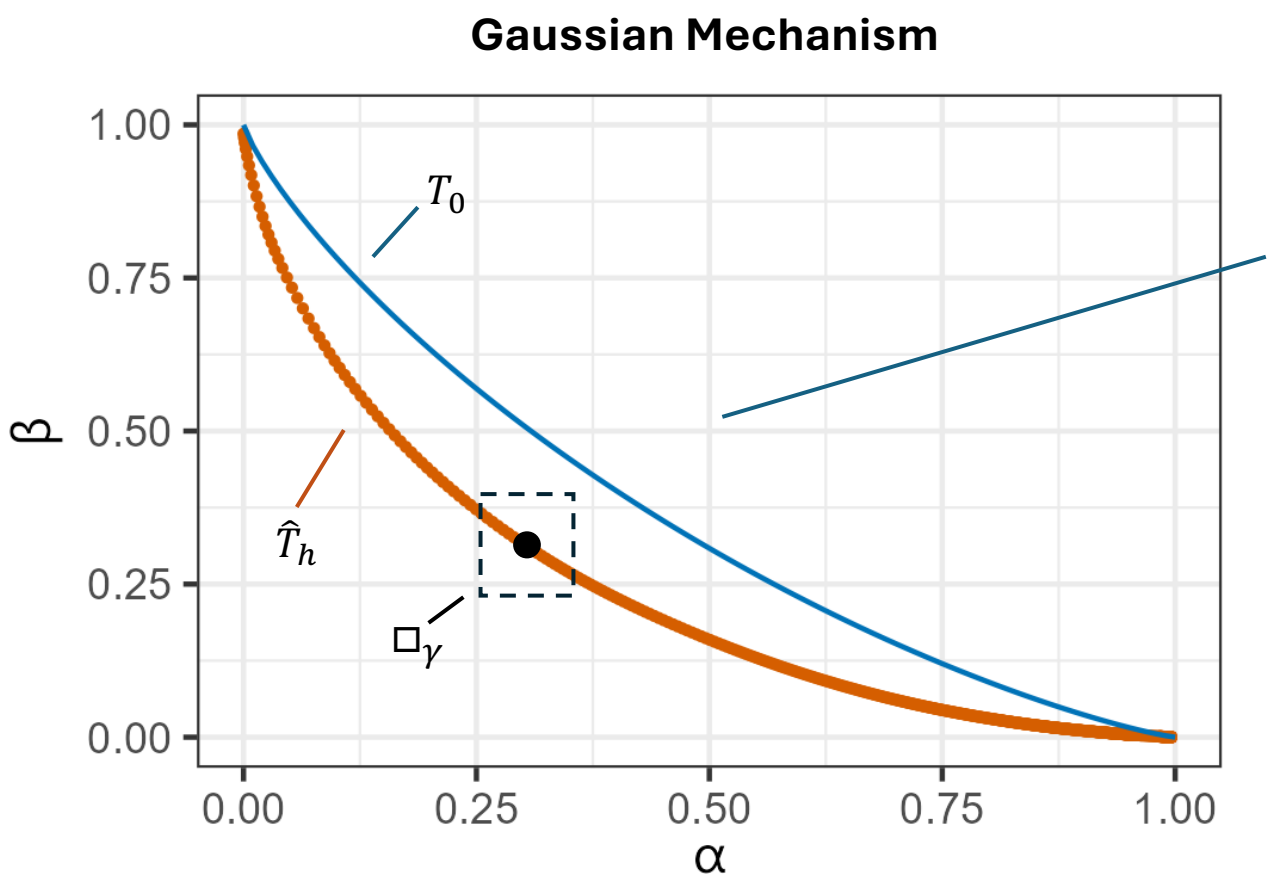
Construct Confidence Region

(1) Compute $\hat{\eta}^*$ where distance between $\hat{T}_h(\hat{\alpha}_h(\eta))$ and $\mathrm{T}_0(\hat{\alpha}_h(\eta))$ is largest.

(2) Compute

$$\square_\gamma = [\tilde{\alpha}(\hat{\eta}^*) - w(\gamma), \tilde{\alpha}(\hat{\eta}^*) + w(\gamma)] \times [\tilde{\beta}(\hat{\eta}^*) - w(\gamma), \tilde{\beta}(\hat{\eta}^*) + w(\gamma)].$$

Auditing



Privacy Violation

$\left(\mathbb{E}[\tilde{\alpha}(\hat{\eta}^*)], \mathbb{E}\left[\tilde{\beta}(\hat{\eta}^*)\right]\right)$ should lie on or above T_0 and, with high probability, in $\square_\gamma$ at the same time!

!

Summary

- Estimation and Auditing of f-DP with Black-Box Access
- Reliable Inference and Detection of Privacy Violations
- Theoretical Guarantees

Take a look at our paper for more details on

- Algorithms
- Theory
- Experiments