DP-Finder: Finding Differential Privacy Violations by Sampling and Optimization

Benjamin Bichsel

Timon Gehr

Dana Drachsler-Cohen

Petar Tsankov

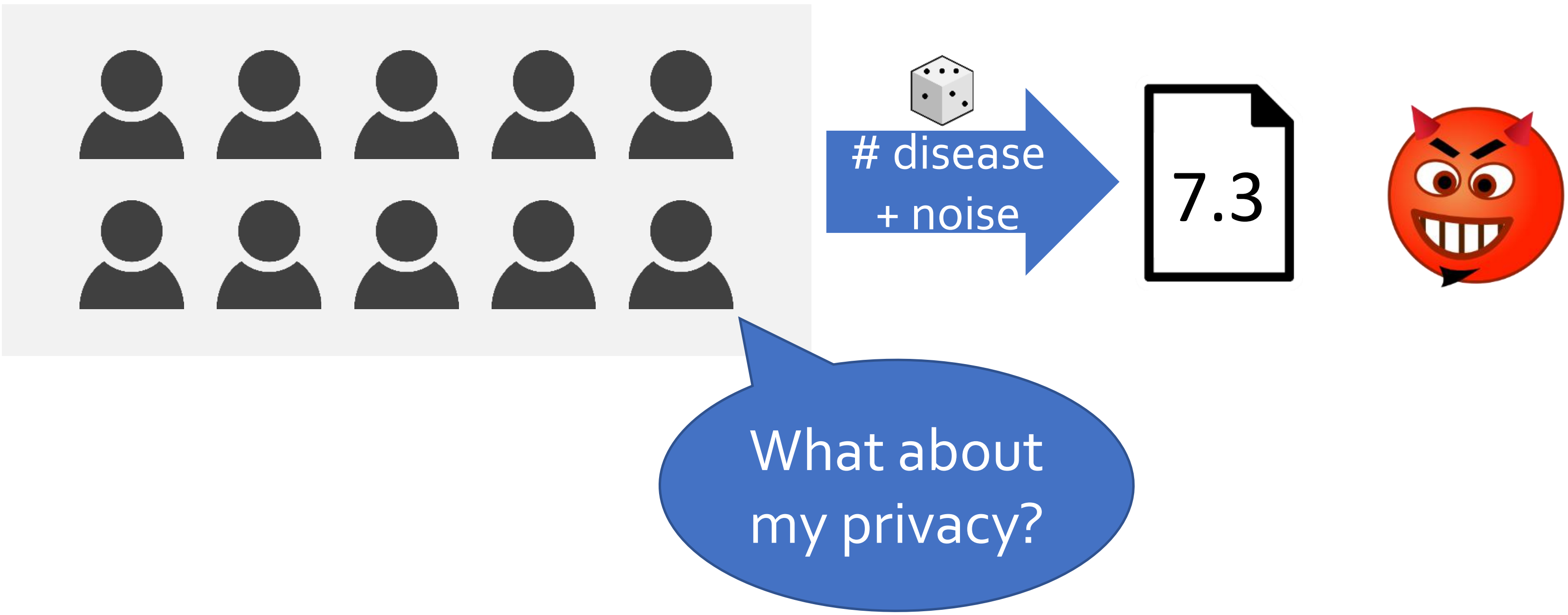
Martin Vechev

SRILAB

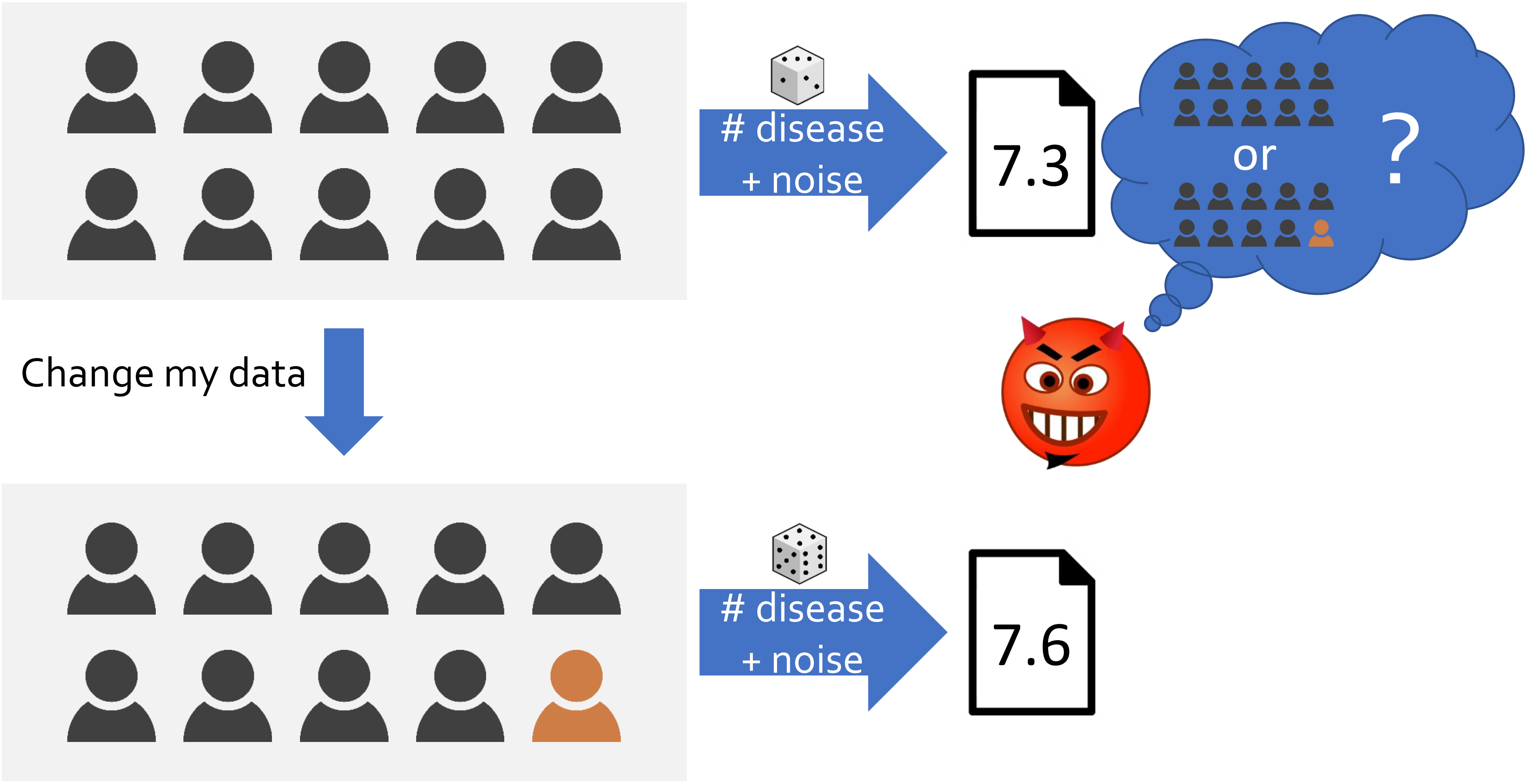
ETH zürich

Differential Privacy – Basic Setting

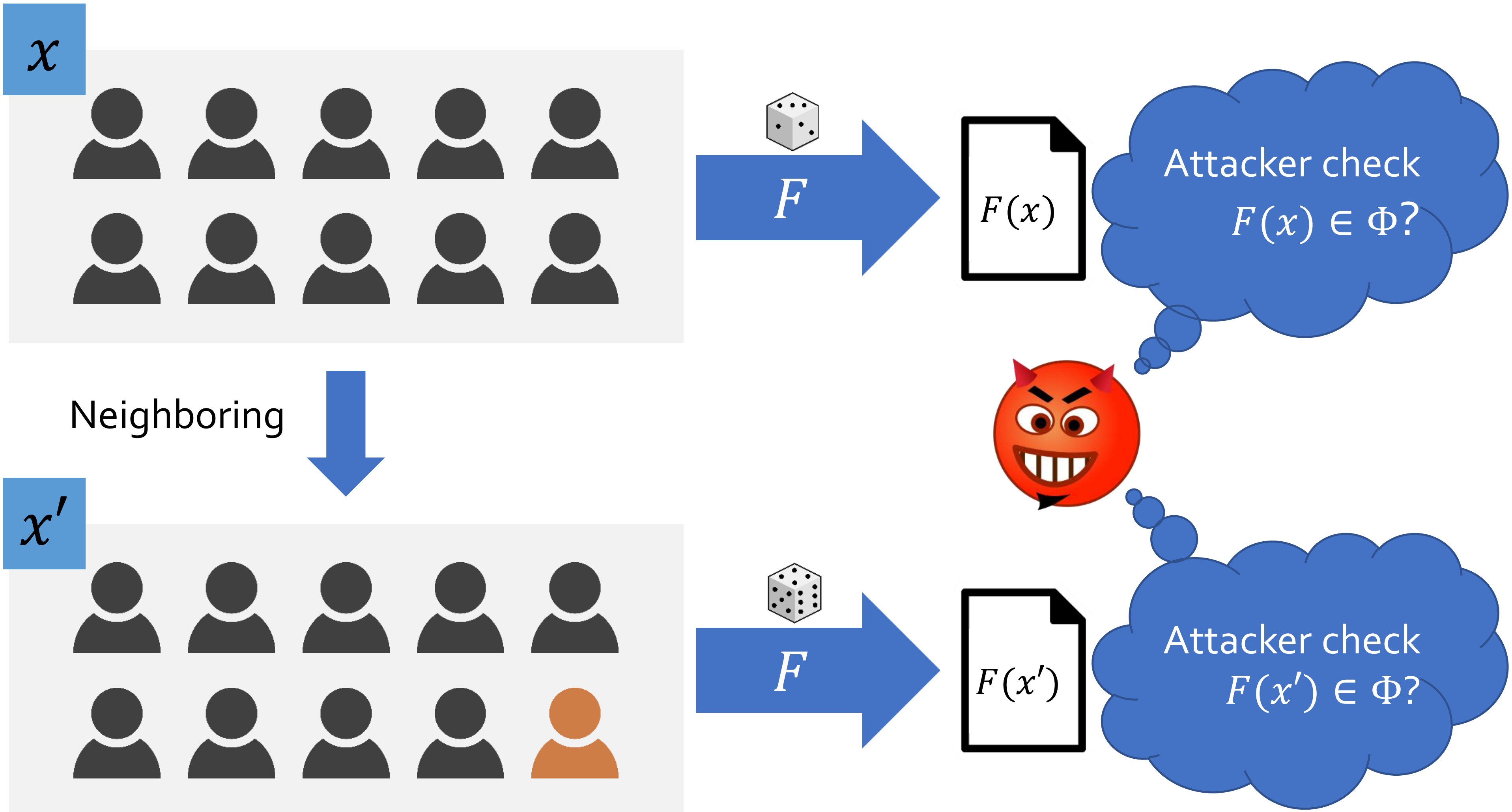
Differential Privacy – Basic Setting



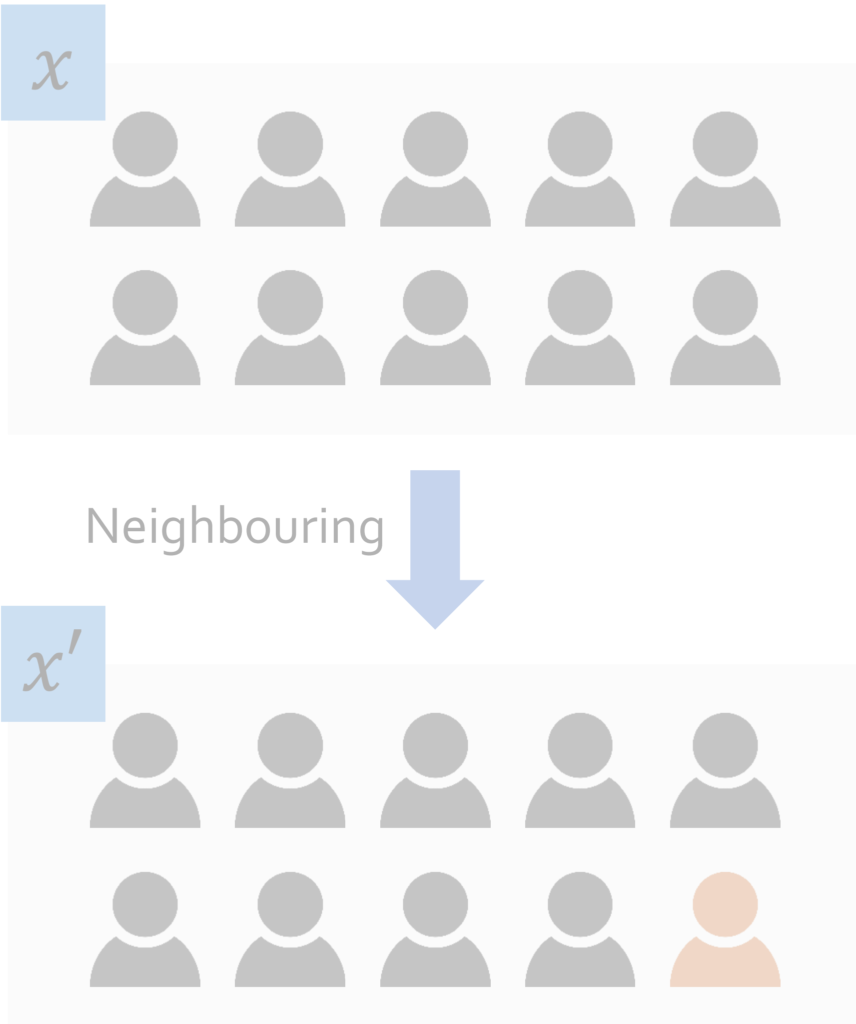
Differential Privacy - Intuition



Differential Privacy – More Abstractly

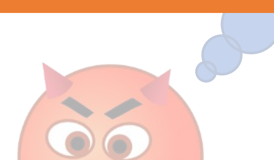


Differential Privacy - Definition



ε-DP:

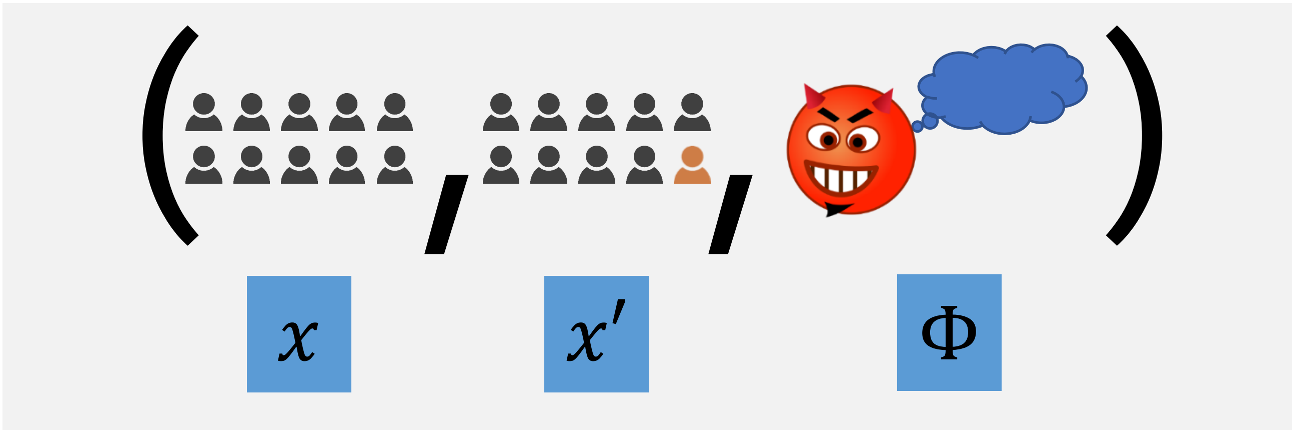
$$\frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} \leq \exp(\varepsilon) \approx 1 + \varepsilon$$



Challenges induced by DP:

- Proving/checking ε-DP is hard (buggy algorithms)
- Proof strategies not complete
- Proofs only provide upper bounds

ε-DP Counterexamples



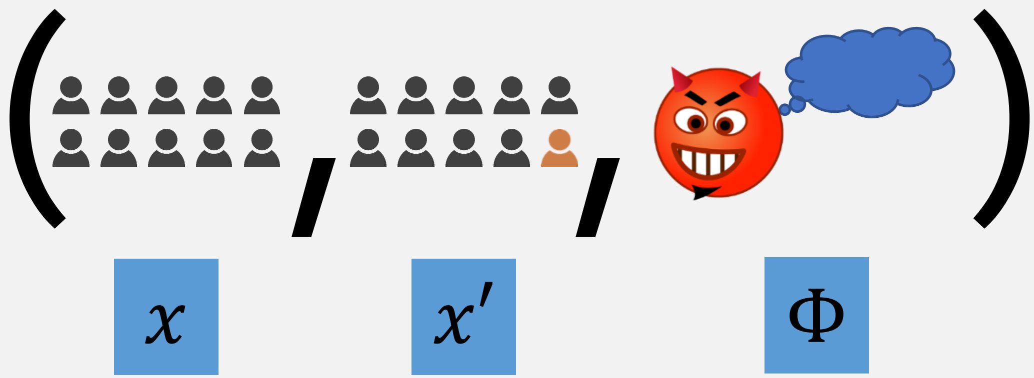
that violate ε-DP:

$$\frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \exp(\varepsilon)$$

$$\Leftrightarrow$$

$$\log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \varepsilon$$

ε-DP Counterexamples



that violate ε-DP:

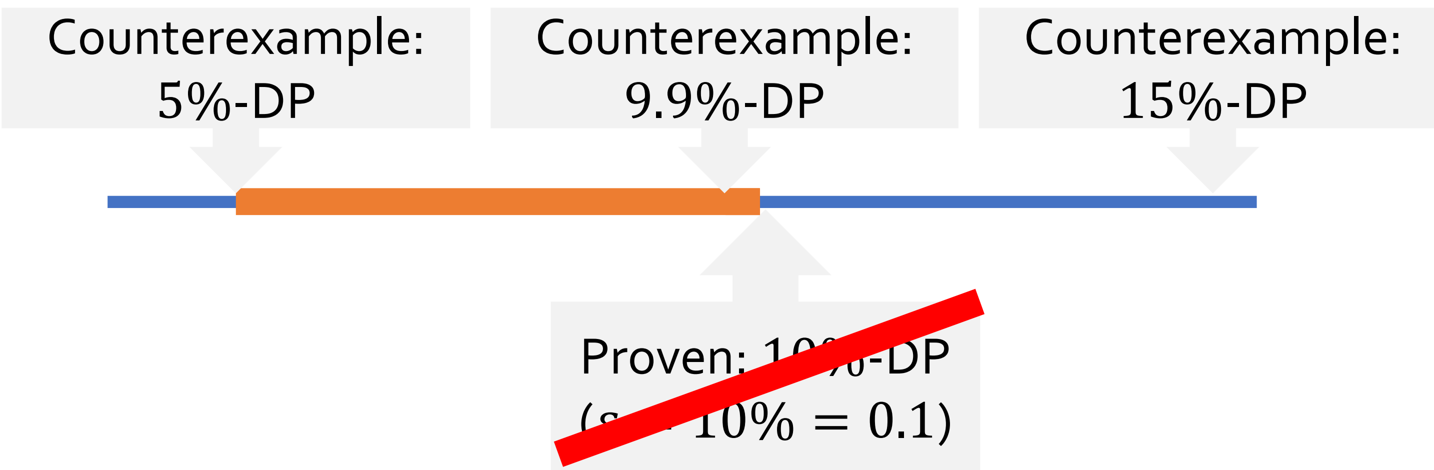
$$\frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \exp(\varepsilon)$$

$$\Longleftrightarrow$$

Maximize $\varepsilon(x, x', \Phi)$

$$\log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]} > \varepsilon$$

Bounds on "true" ε



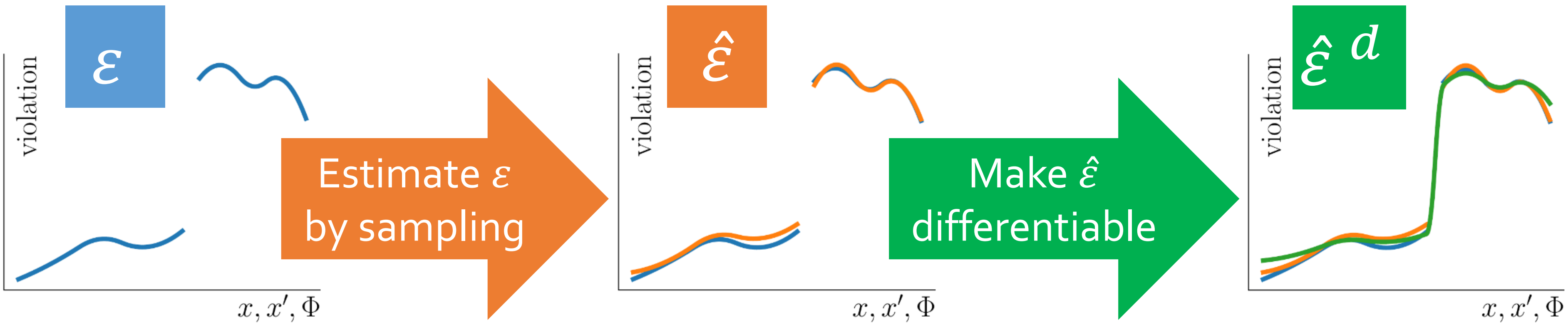
Evaluation: We get precise and large ε, close to known upper bounds

ε-DP Counterexamples

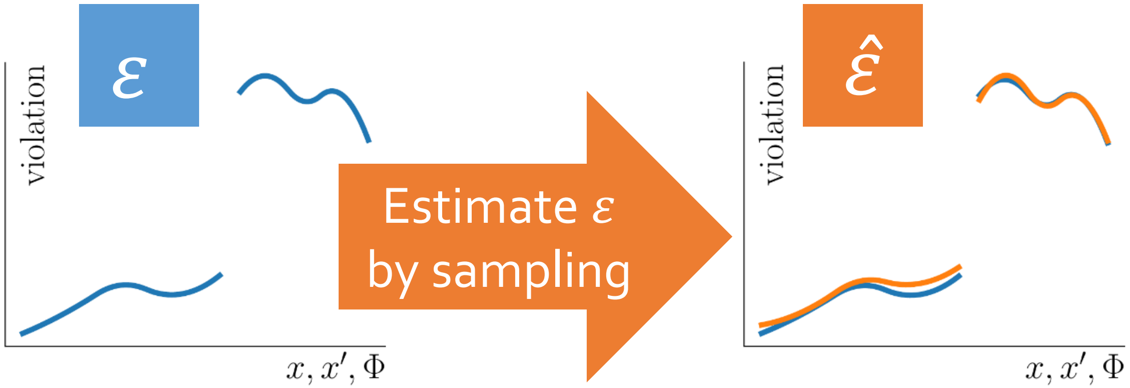
Goal: Maximize $\varepsilon(x, x', \Phi)$

Challenge 1: Expensive to compute ε precisely

Challenge 2: Search space is sparse: Few x, x', Φ lead to large $\varepsilon(x, x', \Phi)$



Step 1: Estimate ε



Estimating ε

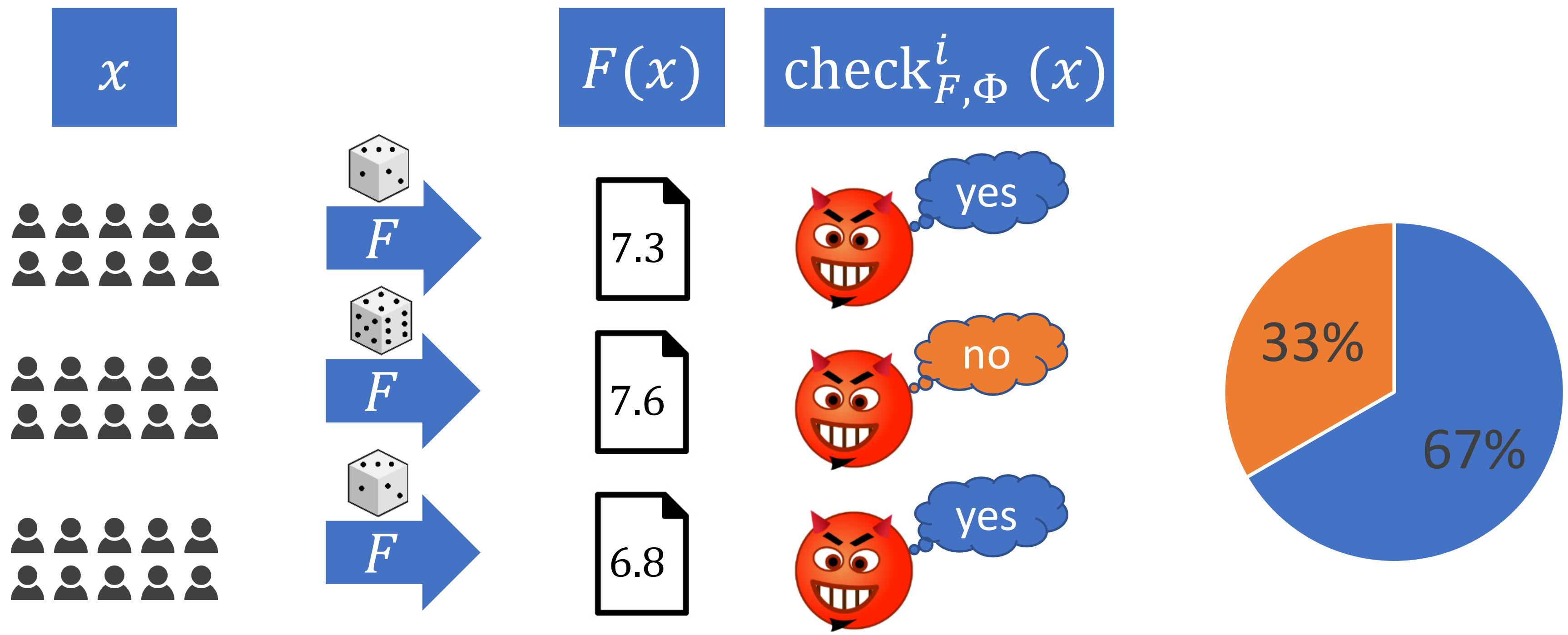
$$\varepsilon(\mathrm{x}, \mathrm{x}', \Phi) \coloneqq \log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]}$$

Estimating ε

Estimate

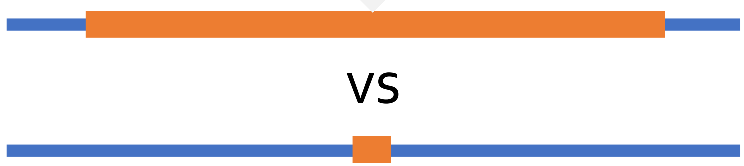
$$\varepsilon(\mathrm{x}, \mathrm{x}', \Phi) := \log \frac{\Pr[F(x) \in \Phi]}{\Pr[F(x') \in \Phi]}$$

$$\widehat{\Pr}[F(x) \in \Phi] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{check}_{F,\Phi}^{i}(x)$$



How precise is our estimate?

Counterexample:
$9.9\% \pm 10\%$-DP

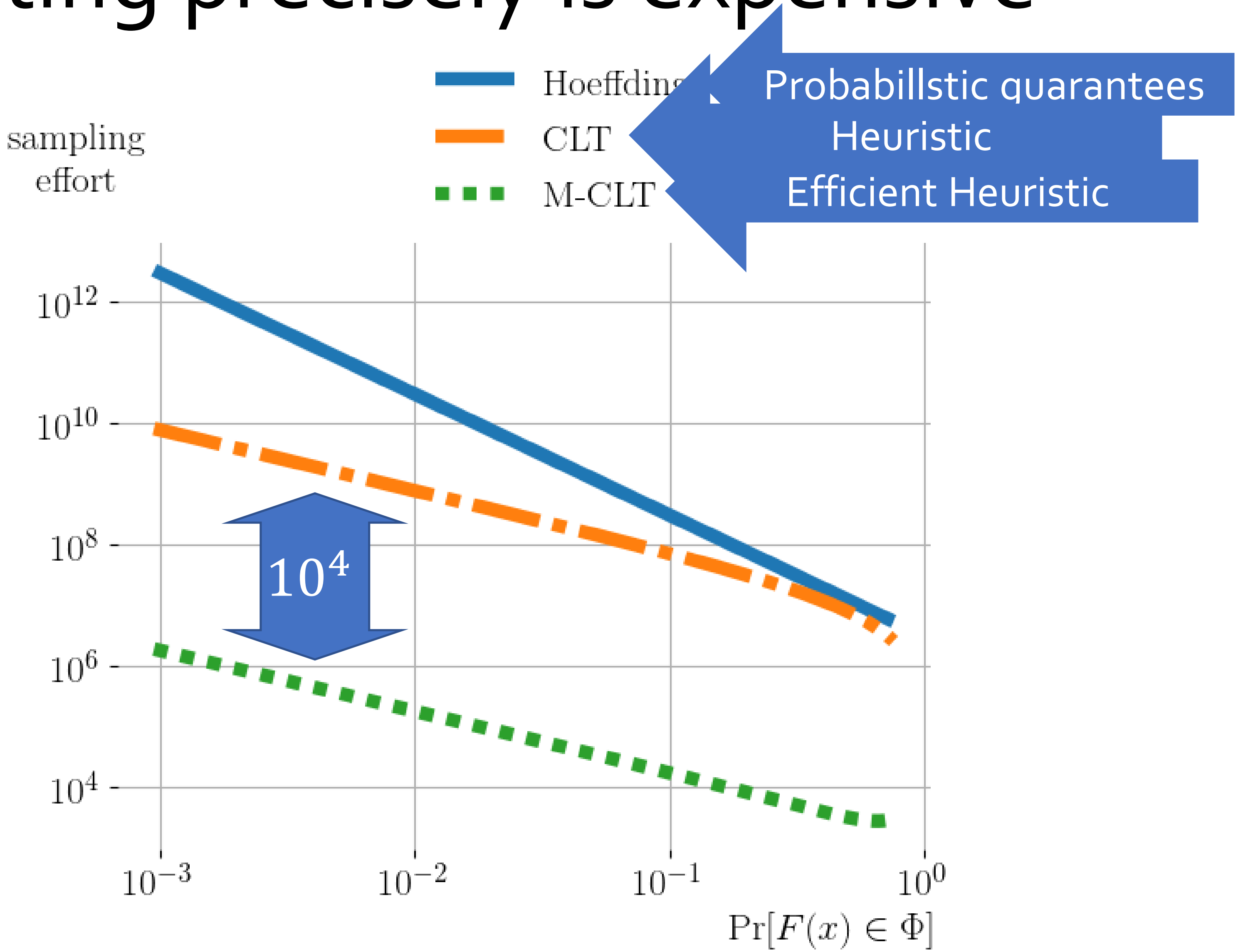


Counterexample:
$9.9\% \pm 2 \cdot 10^{-3}$-DP

Precision of ε ← Precision of $\Pr[F(x) \in \Phi]$ and $\Pr[F(x') \in \Phi]$ ← Sampling effort n

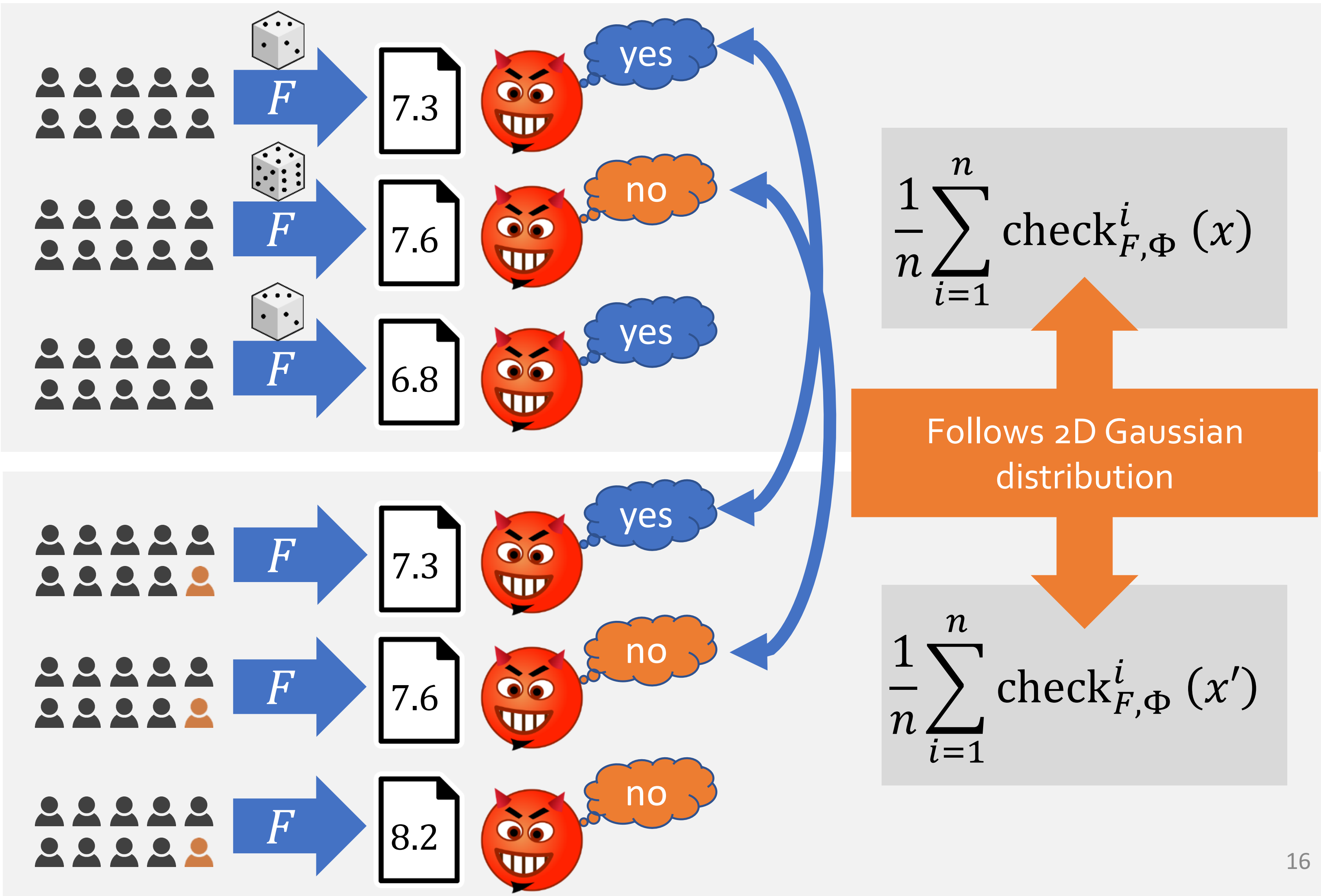
Exponential search

Estimating precisely is expensive



Estimating ε up to an error of $2 \cdot 10^{-3}$ with confidence of 90%

Applying the M-CLT (Correlation)

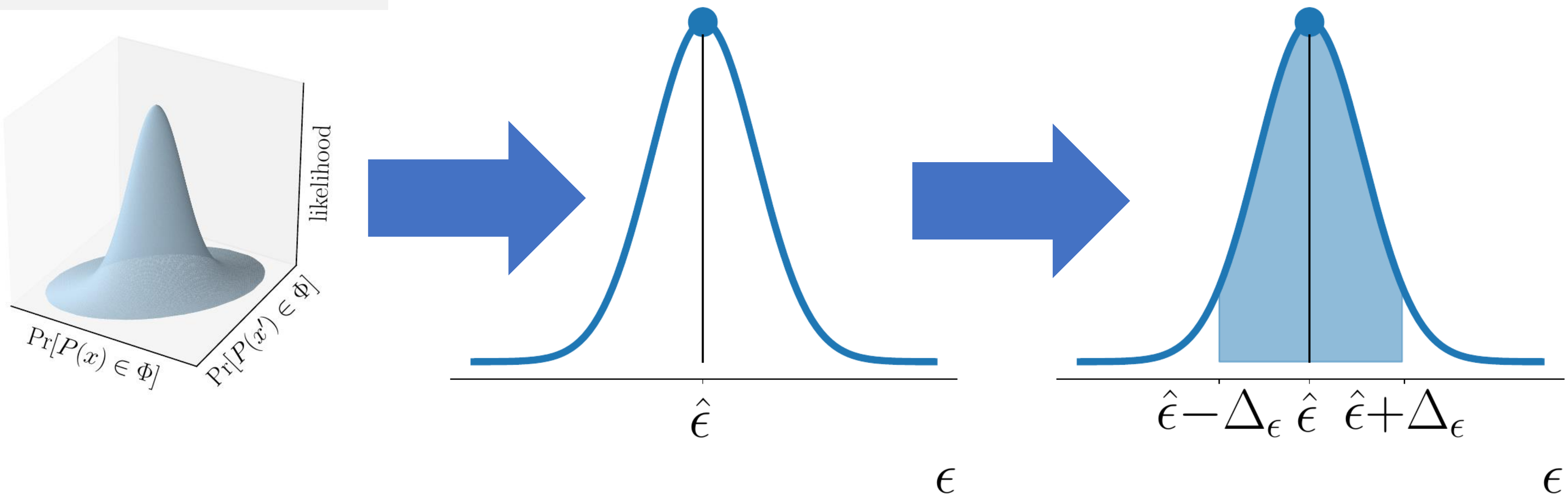


Obtaining a Confidence Interval for ε

Joint likelihood of $\begin{pmatrix} \Pr[F(x) \in \Phi] \\ \Pr[F(x') \in \Phi] \end{pmatrix}$

Likelihood of $\varepsilon(x, x', \Phi)$

Confidence Interval for $\varepsilon(x, x', \Phi)$



Distribution of $\frac{\text{Gauss}}{\text{Gauss}}$ (correlated):

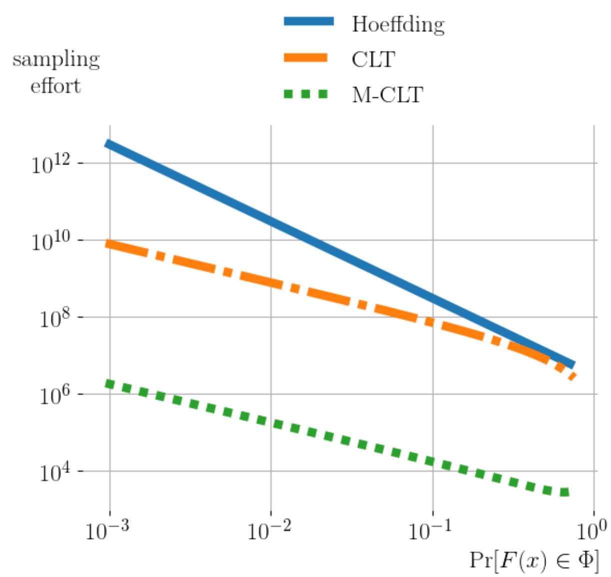
D. V. Hinkley. 1969. On the Ratio of Two Correlated Normal Random Variables. Biometrika 56, 3 (1969), 635–639. http://www.jstor.org/stable/2334671

How precise is our estimate?

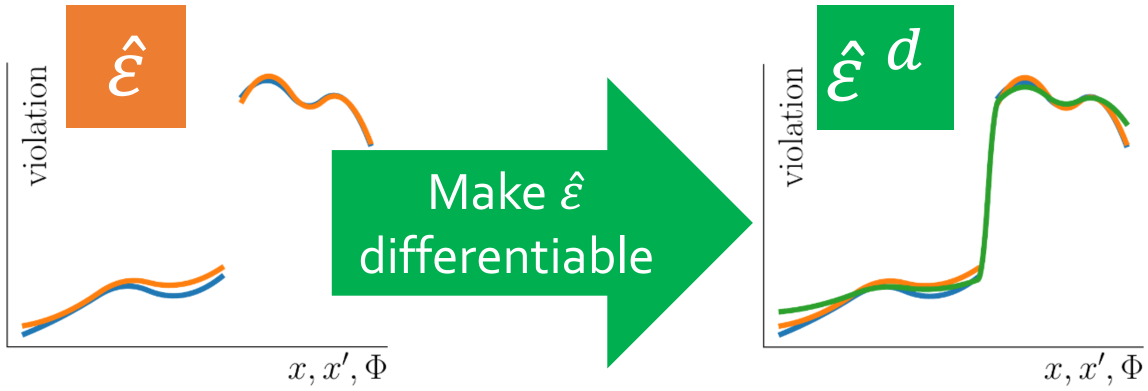
Counterexample:
$9.9\% \pm 10\%$-DP

vs

Counterexample:
$9.9\% \pm 2 \cdot 10^{-3}$-DP



Step 2: Finding Counterexamples



How can we optimize our estimate?

maximize

$$\hat{\epsilon}(x, x', \Phi) = \log \frac{\frac{1}{n}\sum_{i=1}^{n} \text{check}_{F,\Phi}^{i}}{\frac{1}{n}\sum_{i=1}^{n} \text{check}_{F,\Phi}^{i}\,(x')}$$

Not differentiable

Goals
- Make differentiable
- Preserve semantics

$$\neg B \rightsquigarrow 1 - B$$

$$B_1 \wedge B_2 \rightsquigarrow B_1 \cdot B_2$$

$$\textbf{if}\,(B):\ \{x = E_1\}\ \textbf{else}\ :\ \{x = E_2\} \rightsquigarrow x = B \cdot E_1 + (1 - B) \cdot E_2$$

How can we optimize our estimate?

maximize

$$\hat{\epsilon}(x, x', \Phi) = \log \frac{\frac{1}{n}\sum_{i=1}^{n} \text{check}_{F,\Phi}^{i}}{\frac{1}{n}\sum_{i=1}^{n} \text{check}_{F,\Phi}^{i}\,(x'}$$

Not differentiable

- Maximize using SLSQP (supports hard constraints for neighborhood)
- Random starting point (+ restart)
- What about division by zero?
- What about very small denominators?

Main differences to Ding et al.

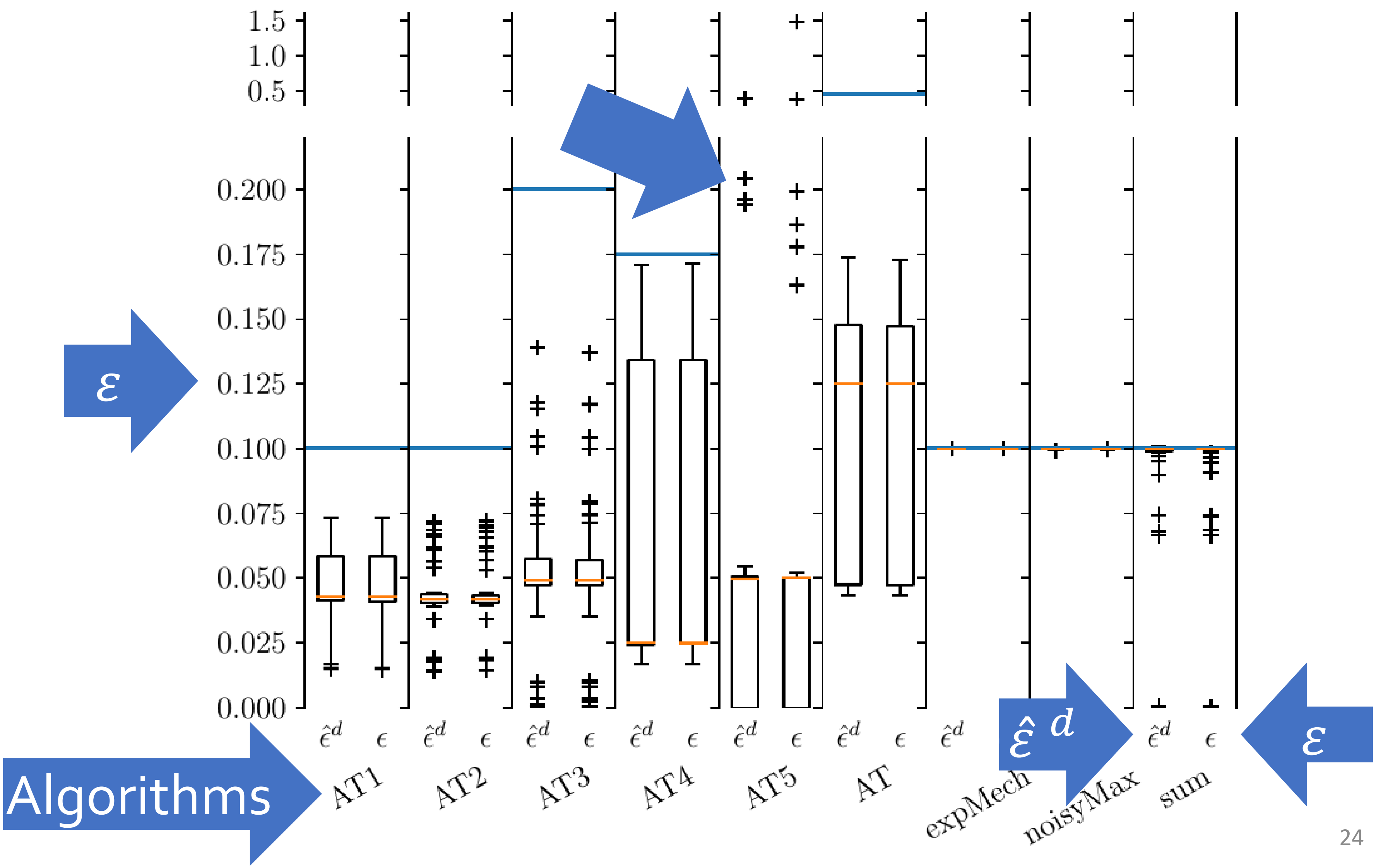
Dimension	Ding et al.	This work
Problem statement	$\varepsilon(x, x', \Phi) > \varepsilon_0$?	Maximize $\varepsilon(x, x', \Phi)$
Approach	Statistical tests	Estimate + confidence interval
Search	By patterns	Gradient descent (incremental)

Evaluation

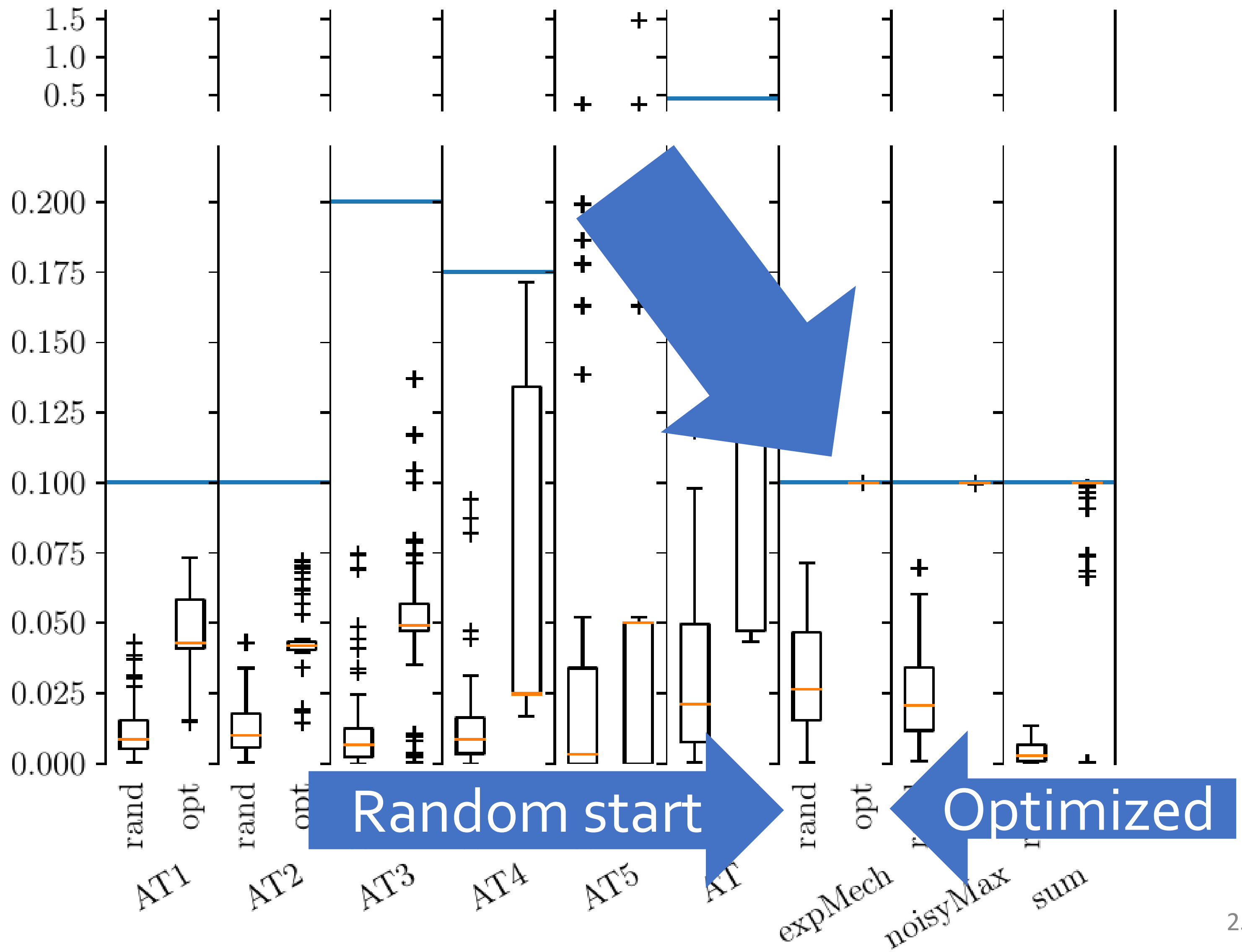
Exact solver (PSI) for ground truth

- How **precise** is the differentiable estimate?
- How **efficient** is DP-Finder in finding violations compared to random search?

Precision of Differentiable Estimate

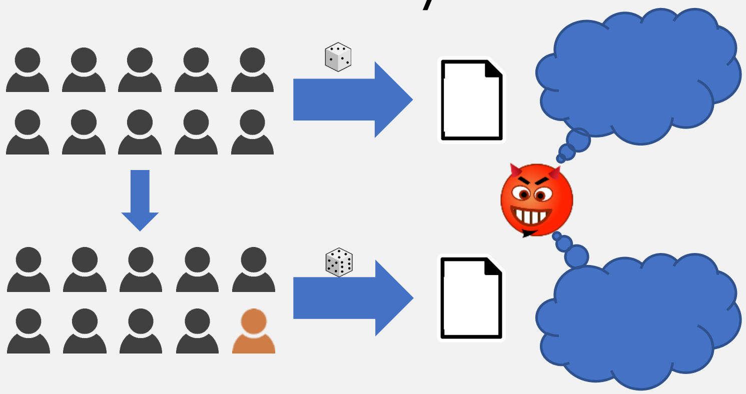


Random vs Optimized



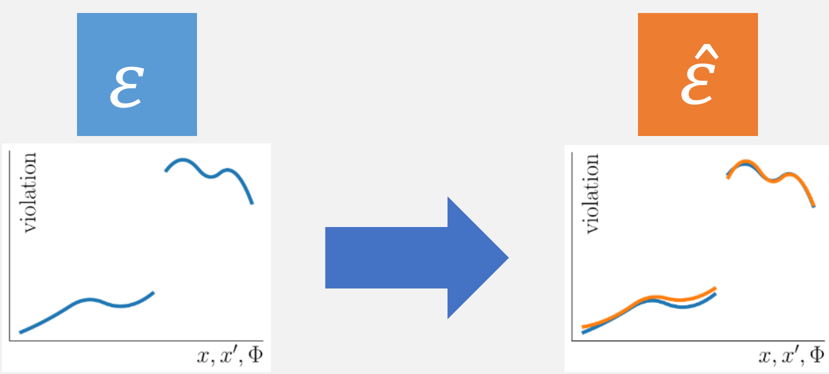
Conclusion

Differential Privacy



ε-DP Counterexamples

Estimate ε



Finding Counterexamples