

# PRESTO: A Python package for recommending privacy preservation algorithm based on user preferences.

Olivera Kotevska<sup>1</sup>, A. Gilad Kusne<sup>2</sup>, Prasanna Balaprakash<sup>1</sup>, and Robert Patton<sup>1</sup>

<sup>1</sup> Oak Ridge National Laboratory, United States <sup>1</sup> National Institute of Standards and Technology, United States

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [Open Journals](#)

## Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

PRESTO (Privacy REcommendation and SecuriTy Optimization) is a Python-based toolkit that automates the selection of differential-privacy mechanisms to balance data utility and privacy loss. By integrating descriptive and inferential statistics, Bayesian optimization, and data-similarity metrics, PRESTO analyzes arbitrary datasets—numerical, categorical, or structured—and recommends the optimal privacy algorithm and  $\epsilon$ -parameter setting. Its modular design supports CPU/GPU execution, streaming and batch data, and extensibility for new algorithms and utility metrics. PRESTO's automated multi-objective optimization delivers application-specific, data-driven recommendations with quantified uncertainty, empowering both experts and non-experts to integrate privacy-preserving methods seamlessly into their workflows.

## Statement of Need

As data collection proliferates across healthcare, finance, IoT, and beyond, safeguarding individual privacy without handicapping downstream analytics has become critical. Existing differential-privacy tools often require deep theoretical knowledge, manual tuning of privacy parameters, and trial-and-error to discover the right trade-off between noise injection and data utility. This steep adoption barrier impedes widespread deployment of privacy-preserving analytics in industrial and research settings. There is a pressing need for an intuitive, automated solution that can—given any dataset—identify the most suitable privacy mechanism and its optimal  $\epsilon$ , quantify the remaining utility, and provide confidence intervals on its recommendations. PRESTO fills this gap, reducing the technical burden and accelerating safe, compliant data analysis.

## State of the Field

A variety of packages from industry and academia—such as IBM's Diffprivlib, Google's PyDP (and TensorFlow Privacy), Facebook's Opacus, LDP-Pure, SmartNoise, PETINA—offer implementations of noise-based DP mechanisms (Laplace, Gaussian, Exponential), local-DP protocols (Randomized Response, RAPPOR), and gradient perturbation for machine learning. However, they typically expose raw APIs, leaving users responsible for selecting and tuning algorithms, and provide limited guidance on choosing  $\epsilon$ . Recent research has explored automatic hyperparameter tuning via cross-validation or surrogate modeling, but these approaches rarely integrate multi-objective optimization or deliver quantitative uncertainty measures.

PRESTO advances the state of the art by unifying statistical dataset analysis, Bayesian optimization, and data-similarity metrics into a single recommendation engine. It implements

a broad suite of privacy mechanisms—including both batch and streaming algorithms—and automates their selection based on data characteristics and user-specified privacy-utility trade-offs, while providing 95% confidence intervals on its recommendations. Crucially, PRESTO is built on a modular architecture, enabling users to plug in new privacy algorithms or custom utility metrics at any time without modifying core logic. This extensibility ensures that PRESTO can evolve alongside emerging research and domain-specific needs, making it uniquely adaptable compared to existing static libraries.

## Methodology

### 1. Dataset Profiling

- Compute descriptive (mean, variance, skewness, kurtosis) and, for categorical data, domain-size and frequency distributions.

### 2. Mechanism Library

- Maintain a dictionary of privacy functions (`get_noise_generators()`), each mapping  $(data, \epsilon) \rightarrow \text{privatized\_data}$ .

### 3. Bayesian Optimization of $\epsilon$

- For each mechanism, define  $f(\epsilon) = -\text{RMSE}(\text{data}, \epsilon, \text{mechanism})$  and maximize over  $\epsilon \in [\epsilon_{\min}, \epsilon_{\max}]$  using Gaussian-process Bayesian optimization.

### 4. Confidence & Reliability

- Compute 95% CI on RMSE at  $\epsilon^*$ ; define  $\text{Reliability} = \frac{1}{\text{Mean RMSE} \times \text{CI Width}}$ .

### 5. Similarity Assessment

- Measure distributional similarity via Kolmogorov–Smirnov, Jensen–Shannon, Pearson correlation.

### 6. Multi-Objective Ranking

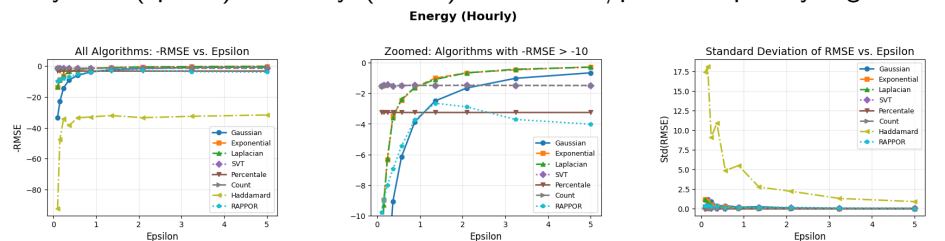
- Recommend top mechanisms on **max similarity**, **max reliability**, and **max privacy** axes.

## Experiments

We conducted experiments to evaluate the effectiveness of our approach.

### Energy Computation with Bayesian Optimization (Dataset: Hourly Consumption (Min))

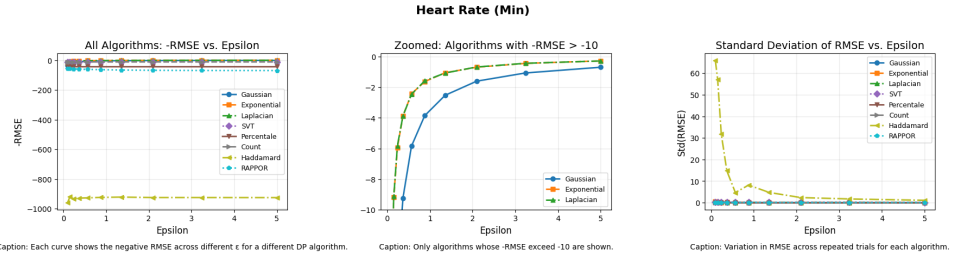
#### 1. Privacy loss (epsilon) vs utility (RMSE) for selected/preferred privacy algorithms



2. Top-3 Recommendations: DP\_Laplace:  $\epsilon=3.6277$ ,  $\text{mean\_rmse}=0.3817$ ,  $\text{ci\_width}=0.0279$ ,  $\text{reliability}=93.90$  DP\_Exponential:  $\epsilon=3.6300$ ,  $\text{mean\_rmse}=0.3835$ ,  $\text{ci\_width}=0.0416$ ,  $\text{reliability}=62.68$  DP\_Gaussian:  $\epsilon=4.1687$ ,  $\text{mean\_rmse}=0.8326$ ,  $\text{ci\_width}=0.0525$ ,  $\text{reliability}=22.88$

## Medical Measurements with Bayesian Optimization (Dataset: Heart Rate (Min))

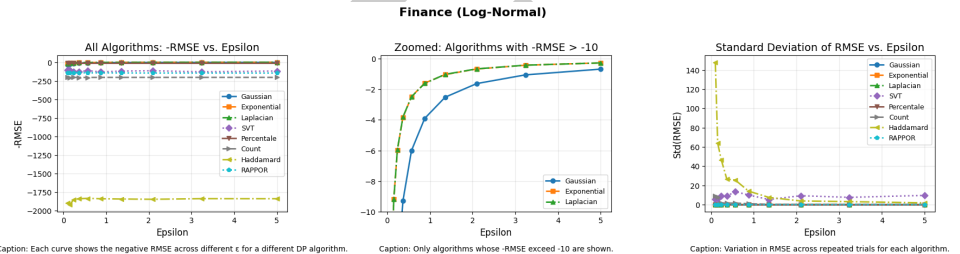
### 1. Privacy loss (epsilon) vs utility (RMSE) for selected/preferred privacy algorithms



2. Top-3 Recommendations: DP\_Laplace:  $\epsilon=3.6254$ , mean\_rmse=0.3901, ci\_width=0.0054, reliability=474.71 DP\_Exponential:  $\epsilon=3.6319$ , mean\_rmse=0.3916, ci\_width=0.0051, reliability=500.71 DP\_Gaussian:  $\epsilon=5.0000$ , mean\_rmse=0.6824, ci\_width=0.0047, reliability=311.79

## Finance Transactions with Bayesian Optimization (Dataset: Payment Transactions (Min))

### 1. Privacy loss (epsilon) vs utility (RMSE) for selected/preferred privacy algorithms



2. Top-3 Recommendations: DP\_Laplace:  $\epsilon=4.1687$ , mean\_rmse=0.3461, ci\_width=0.0340, reliability=84.98 DP\_Exponential:  $\epsilon=3.6296$ , mean\_rmse=0.3864, ci\_width=0.0453, reliability=57.13 DP\_Gaussian:  $\epsilon=4.1690$ , mean\_rmse=0.8270, ci\_width=0.0560, reliability=21.59

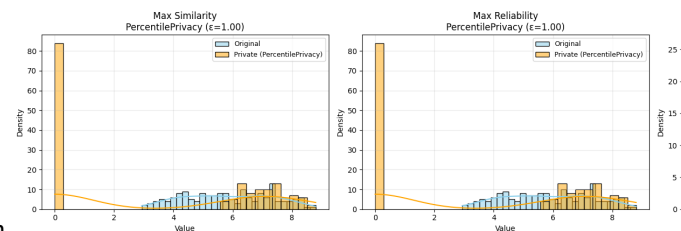
## Sensor Temperature Time-Series with Bayesian Optimization (Dataset: Payment Transactions (Min))

### 1. Privacy loss (epsilon) vs utility (RMSE) for selected/preferred privacy algorithms

2. Top-3 Recommendations: DP\_Laplace:  $\epsilon=3.6296$ , mean\_rmse=0.3846, ci\_width=0.0126, reliability=206.36 DP\_Exponential:  $\epsilon=3.6296$ , mean\_rmse=0.3883, ci\_width=0.0187, reliability=137.72 DP\_Gaussian:  $\epsilon=3.6296$ , mean\_rmse=0.9459, ci\_width=0.0334, reliability=31.65

## Energy Consumption with Fixed epsilon = 1

Original vs. Private Distributions ( $\epsilon=1.00$ )



### 1. The best algorithm for a given epsilon

Best by Similarity: {'algorithm': 'PercentilePrivacy', 'score': np.float32(0.9841)} Best by Reliability: {'algorithm': 'PercentilePrivacy', 'score': inf} Best by Privacy: {'algorithm': 'Hadamard\_Mechanism', 'score': 71.6581}

105 **ML Classification with Private Gradients**

106 Baseline Accuracy (no privacy): 93.00% DP Accuracy with 'PercentilePrivacy': 94.00%

107 **ML Classification with Private Gradients**

108 1. Pareto front for privacy budget vs accuracy

109

110 **Conclusion**

111 PRESTO delivers a data-driven, automated, and extensible framework for differential-privacy  
112 mechanism selection and tuning. By profiling statistical properties, optimizing  $\epsilon$  via Bayesian  
113 methods, and quantifying both utility and uncertainty, PRESTO guides users to the privacy  
114 solution best suited for their data. Its modular design allows seamless integration of new  
115 algorithms and metrics, positioning PRESTO as a flexible platform for both practitioners and  
116 researchers aiming to embed privacy guarantees in diverse analytical workflows.

117 **Acknowledgements**

118 This manuscript has been co-authored by UT-Battelle, LLC under Contract No. DE-AC05-  
119 00OR22725 with the U.S. Department of Energy. The United States Government retains  
120 and the publisher, by accepting the article for publication, acknowledges that the United  
121 States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish  
122 or reproduce the published form of this manuscript, or allow others to do so, for United  
123 States Government purposes. The Department of Energy will provide public access to these  
124 results of federally sponsored research in accordance with the DOE Public Access Plan  
125 (<http://energy.gov/downloads/doe-public-access-plan>). This material is based upon work  
126 supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific  
127 Computing Research under Contract No. DE-AC05-00OR22725. This research is sponsored by  
128 the Artificial Intelligence Initiative as part of the LDRD-SEED Program, at ORNL, managed  
129 by UT-Battelle, LLC and DOE ASCR Program.