Privacy in Practice: The Feasibility of Differential Privacy for Telemetry Analysis

Tyler Kurpanek
tkurpane@ucsd.edu

Christopher Lum
cslum@ucsd.edu

Bradley Nathanson
bnathanson@ucsd.edu

Trey Scheid
tscheid@ucsd.edu

Mentor:
Yu-Xiang Wang
yuxiangw@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Introduction

Research Question: How effective is differential privacy when it is applied in practice?

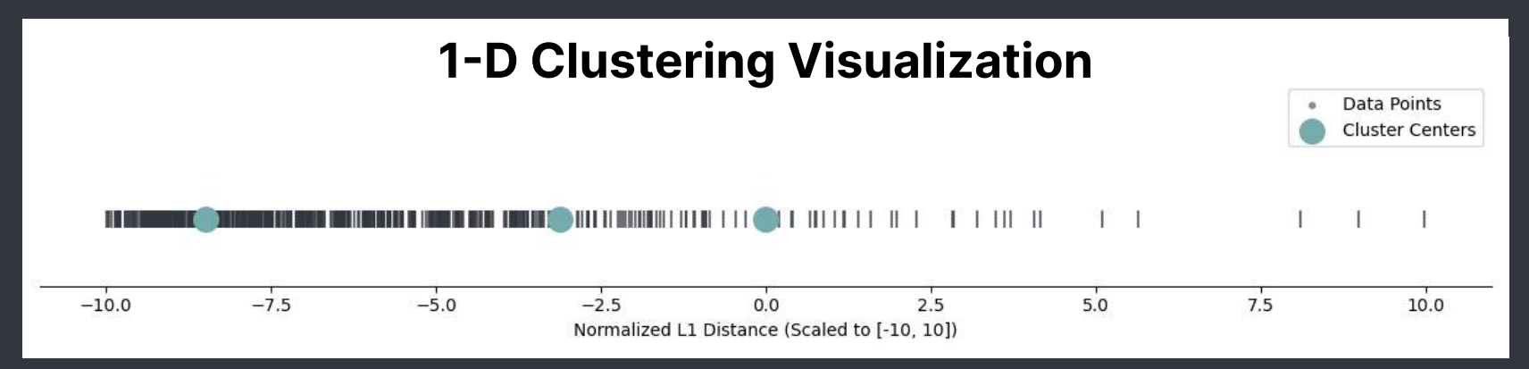
We replicated 4 papers implementing Differential Privacy (DP) in order to assess the utility lost from adding noise.

The main idea of DP is to add noise into algorithms to ensure that results match the data patterns yet are indistinguishable between datasets with and without a specific datum.

Telemetry data is diagnostic information collected by devices such as CPUs or OSes. This data can paint a vibrant picture of a user given appropriate analysis.

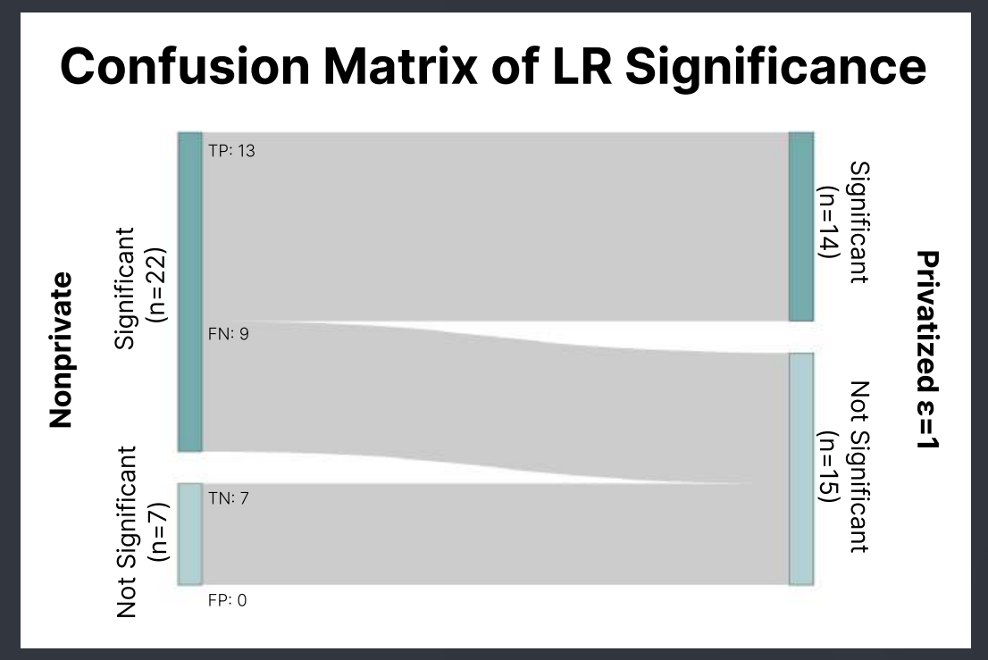
Methods	Description	Non-Private	Private
Conditional Probabilities	Probability of uncorrected error based on number of corrected errors	Histogram creation for event occurrence	Laplace Mechanism
Log. Regression Coefficient Test	Do corrected errors cause uncorrected errors? Testing 29 different errors	Wald test on logistic regression coefficient	Noisy Gradient Descent during LR training
Lasso Regression	Find features which best predict pack power usage	Coordinate Descent & Frank-Wolfe	Noisy Frank-Wolfe (Exponential Mechanism)
K-Means Clustering	Cluster devices based on usage counts	K-Means via Lloyd's algorithm	Noisy mean computation during centroid updates

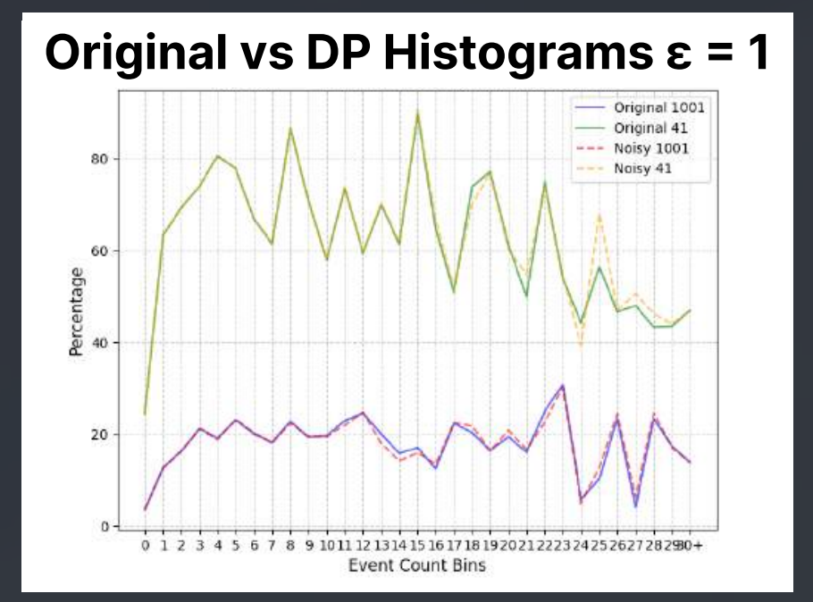
Results (for epsilon = 1)



DP-KMeans centroids are near lower L1 distances given the skewed distribution.

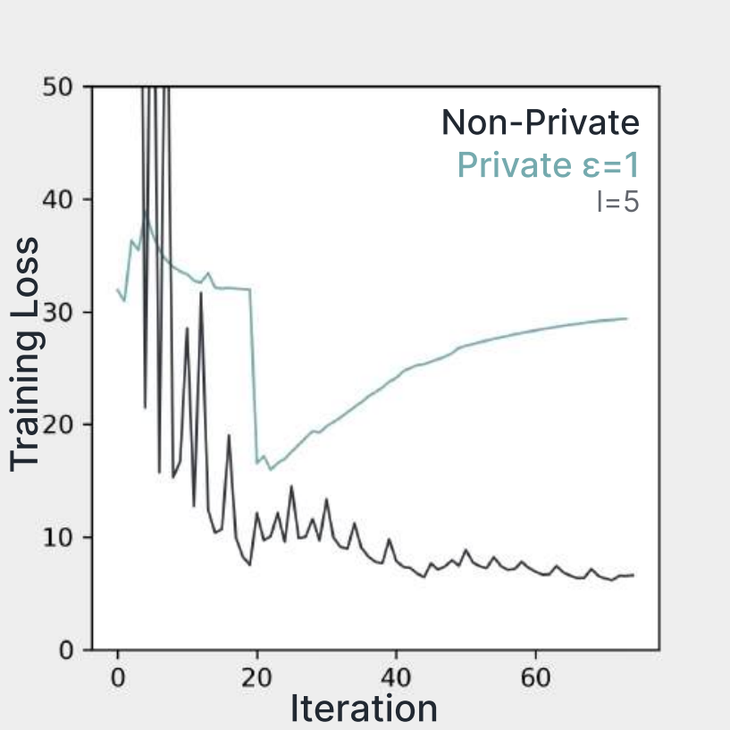
Private Wald test errantly found that 40% of cases that should've been significant were insignificant.

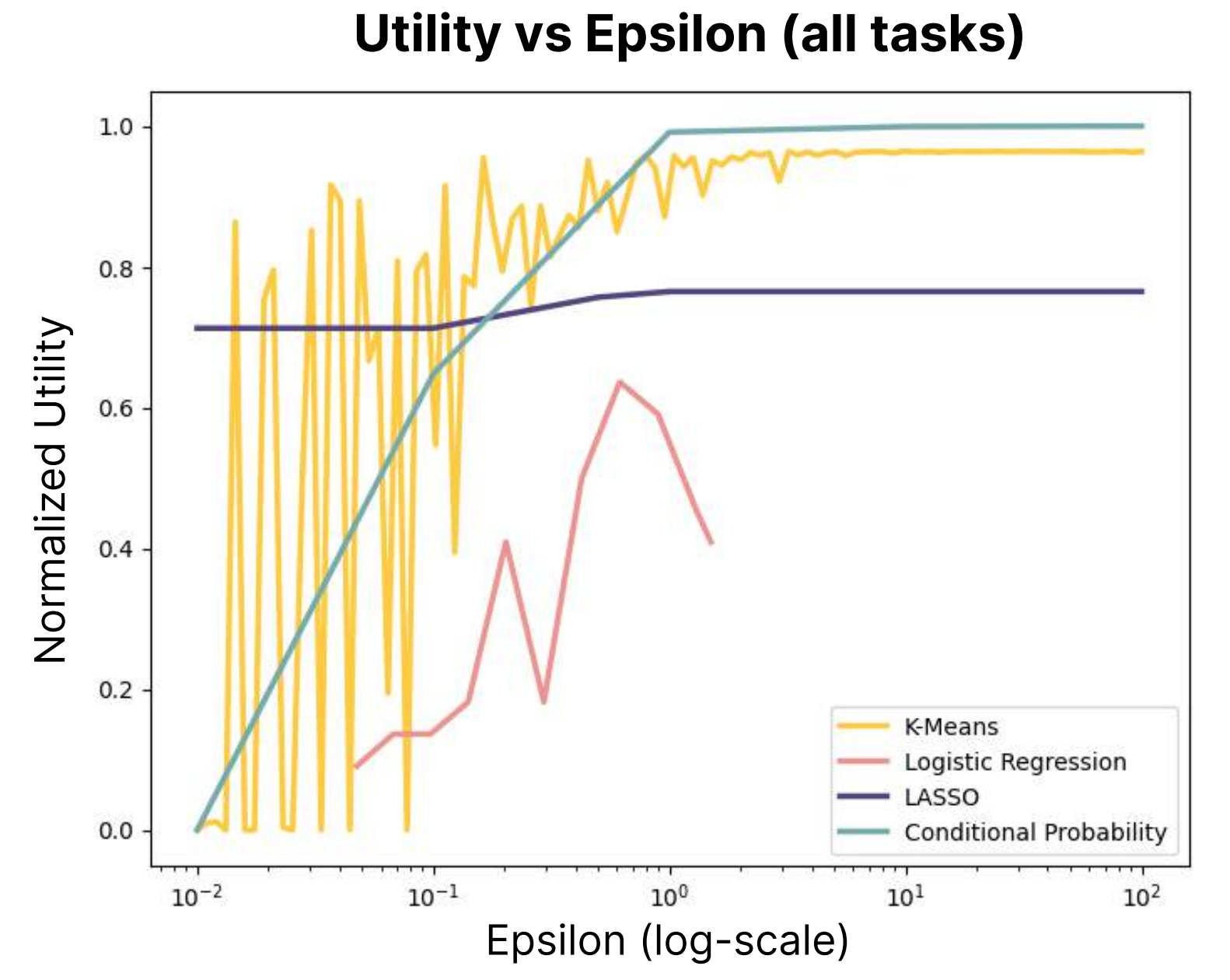




The corrected error (19) counts vs the uncorrected error (41 and 1001) percentages with and without noise at ε = 1.

Privatized Lasso needs different regularization due to sensitive convergence criteria, the algorithm performs well when d >> n unlike telemetry.





Normalized utility is a range from 0-1 of the accuracy of each model, relative to a baseline (non-private model)

Epsilon determines the trade-off between utility and privacy

Discussion

Epsilon of 1 is generally considered to be **highly private**. In our tests, high privacy results in somewhat unusable and highly-incorrect analyses. Further, for some methods such as DP-GD, compute scales linearly with epsilon making meta-analysis more costly for higher values of epsilon.

Epsilon below 10 is considered better than nothing, though epsilon of 10 is typically laughably poor. Only at these large epsilons do we see similar results as the nonprivate

Adding noise using python can be simple, scaling and tracking budgets requires following theorems from researchers!

Key Takeaways

- Even with large amounts of data, strong privacy guarantees suffer from grave utility loss
- Privately selecting hyperparameters either requires vast domain knowledge or sacrificing some privacy budget
- Practical application of DP likely requires a loosening of which agents must be protected against
- Epsilon is a poor quantification of privacy for non-expert practitioners