UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR IT SECURITY

Funded by Bundesministerium für Bildung und Forschung

DPM: Clustering Sensitive Data through Separation

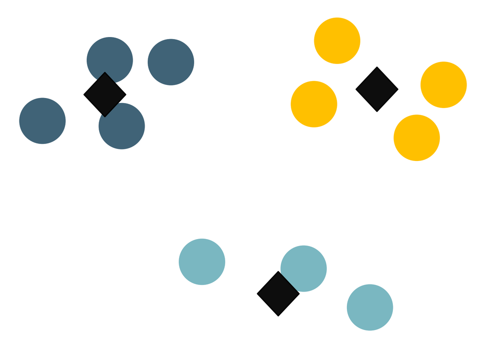
Yara Schütt* & Johannes Liebenow*,
Marcel Gehrke, Tanya Braun, Florian Thaeter, Esfandiar Mohammadi

y.schuett@uni-luebeck.de, j.liebenow@uni-luebeck.de

*The first two authors equally contributed to this work.

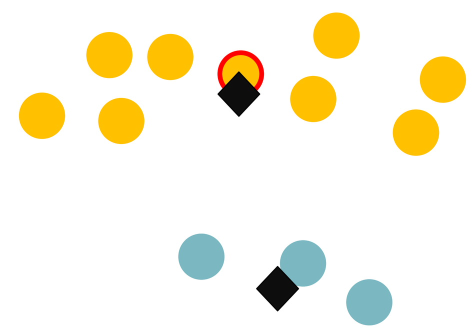
I Privacy Challenges of Clustering

Clustering: Find groups of data points and determine their centroids.
Use centroids to determine correlations/similarities or perform data synthesis.

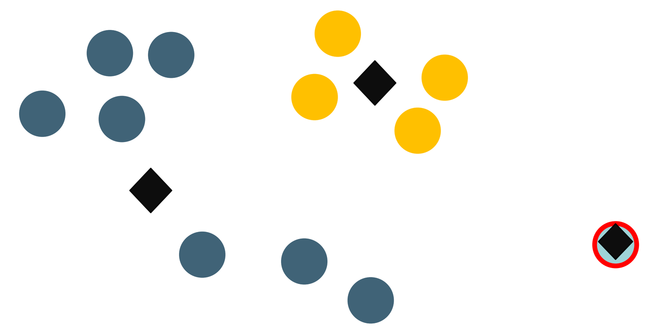


Challenge: Single data points can have a huge impact on the resulting clusters and their centroids.

Bridge Point: A single data point functions as connection between to clusters.

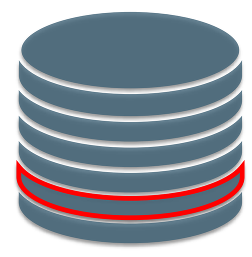


Outlier: Single data point with a large distance from the mass of points.



Privacy: Extract centroids without leaking sensitive information about single data points.

	A	B
1	x	y
2	i	j
…	…	…



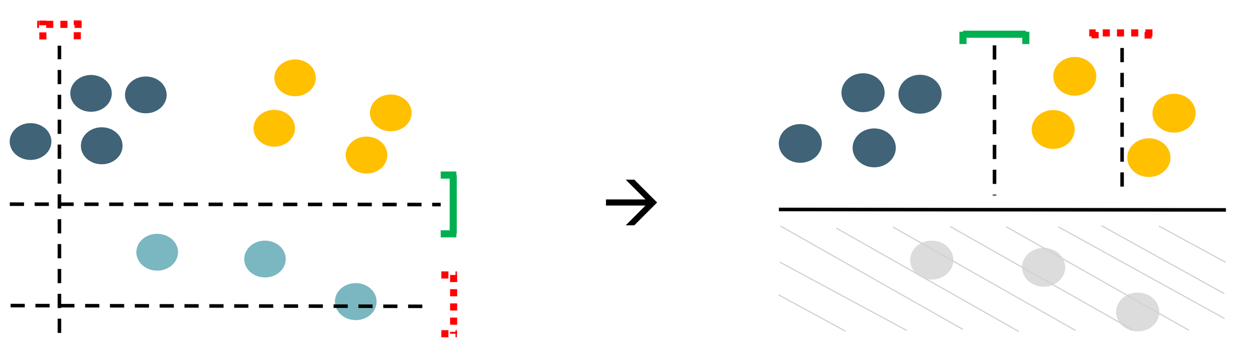
Differentially-private (DP) clustering algorithms reduce the impact of single data points. However, satisfying privacy necessarily reduces utility.

State-of-the-art approaches [1] partition the data set by applying random splits.
→ DPM achieves higher utility based on carefully selected splits while preserving privacy.

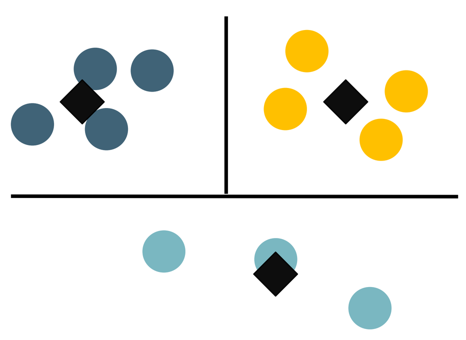
II Clustering through Separation

DPM Approach

1. Split data points recursively into disjoint subsets:
   a. Generate a set of split candidates in every dimension.
   b. Assign a score to each split and select one with a high score.



2. Halt if the number of points in each subset falls below a given threshold and obtain centroids by averaging.

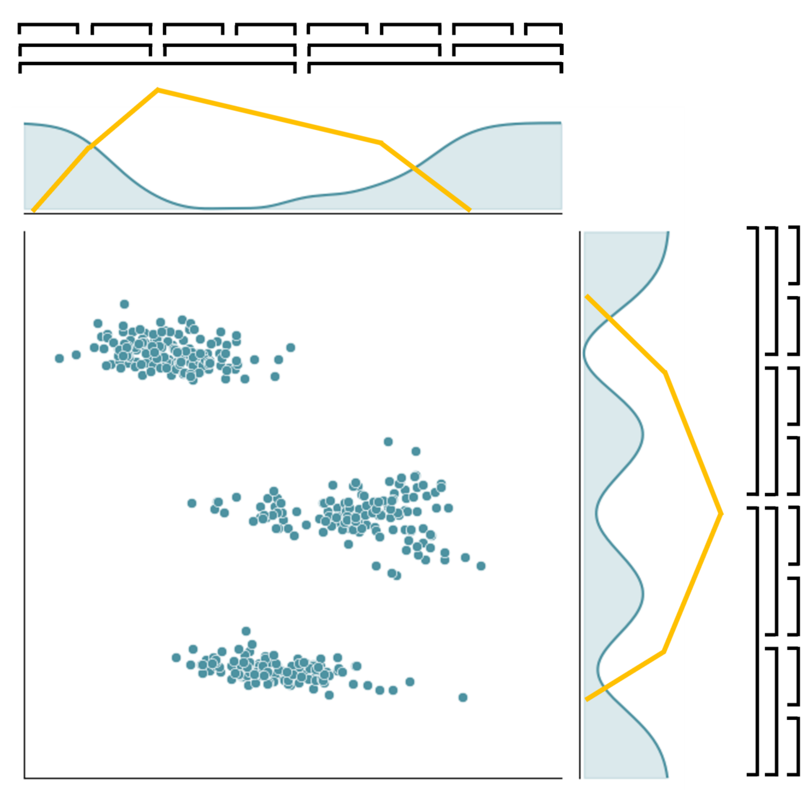


Split Score

Window size:
Large areas without data points indicate gaps between clusters. Small areas without data points can also occur inside a cluster.

Emptiness:
A gap is defined as an area with no or just a few data points.

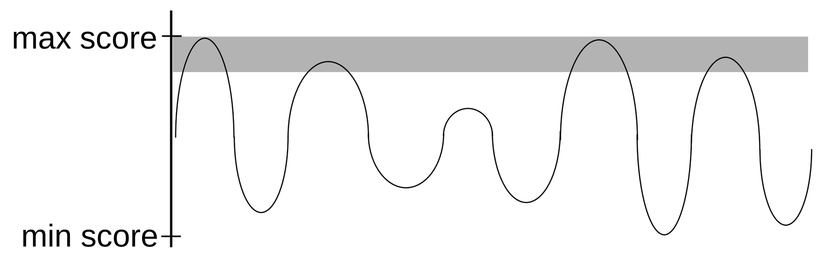
Centreness:
Splits close to the centre of the data points are preferred over splits that are close to the boundaries.



Score = **Window size** + **Emptiness** + **Centreness**

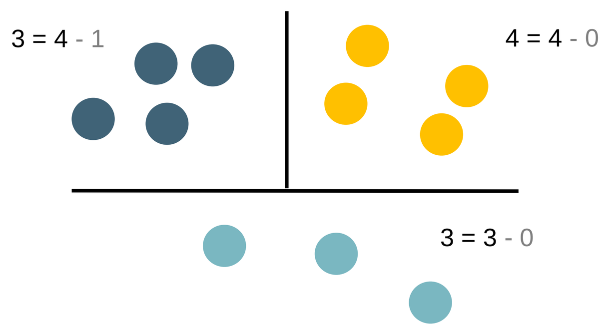
III Ensure Privacy of DPM Steps

Selection via Exponential Mechanism



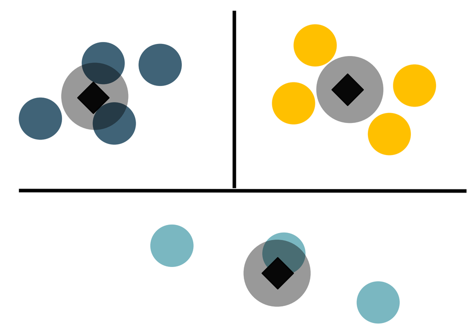
Select candidate with score close to max score with high probability.

Noisy Number of Points in Subset



Perturb the number of data points in a subset.

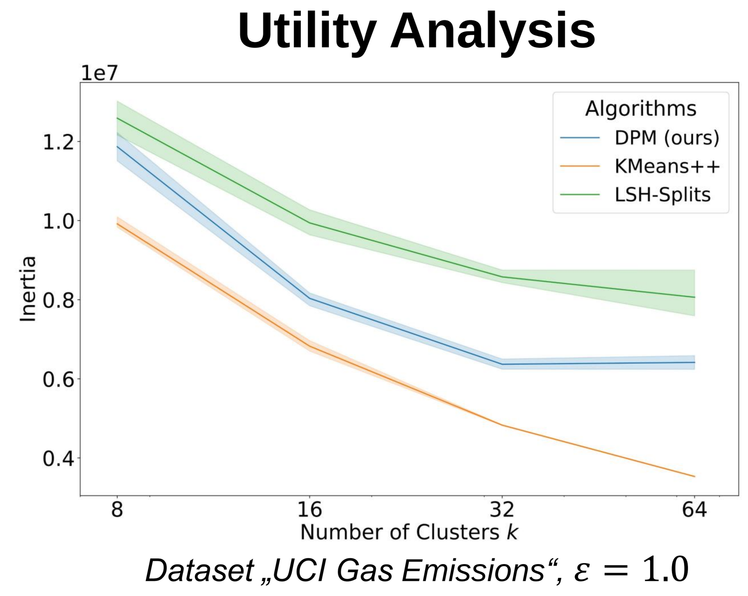
Noisy Averaging



Find noisy average that is with high probability close to the actual average.

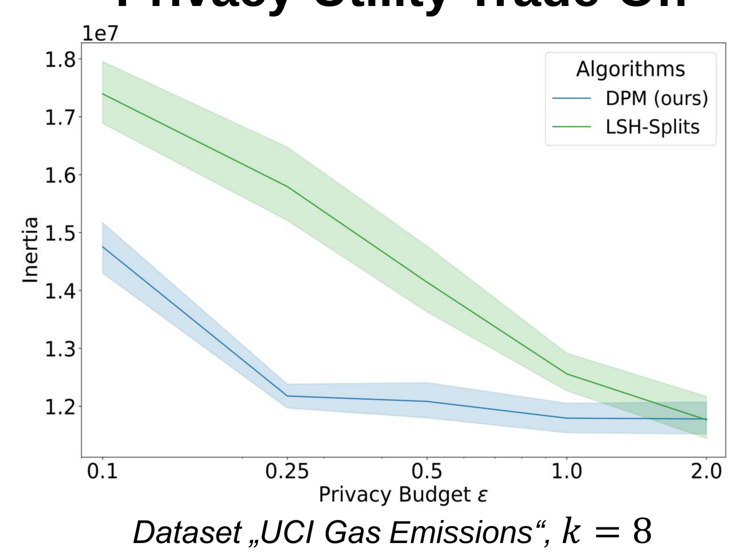
IV Results

Utility Analysis



Dataset „UCI Gas Emissions“, $\varepsilon = 1.0$

Privacy-Utility Trade-Off



Dataset „UCI Gas Emissions“, $k = 8$

Inertia: Sum of squared distances between data points and their closest centroid.
→ Low Inertia ≙ High Utility

KMeans++: Non-DP clustering
LSH-Splits: State-of-the-art DP clustering [1]

[1] Alisa Chang, Badih Ghazi, Ravi Kumar, Pasin Manurangs: *Locally Private k-Means in One Round* (2021)

IM FOCUS DAS LEBEN