S+t-SNE

Bringing dimensionality reduction to data streams

Pedro C. Vieira, João P. Montrezol, João T. Vieira, João Gama
Department of Computer Science, Faculty of Sciences, University of Porto
Faculty of Economics, University of Porto
INESC TEC

Motivation

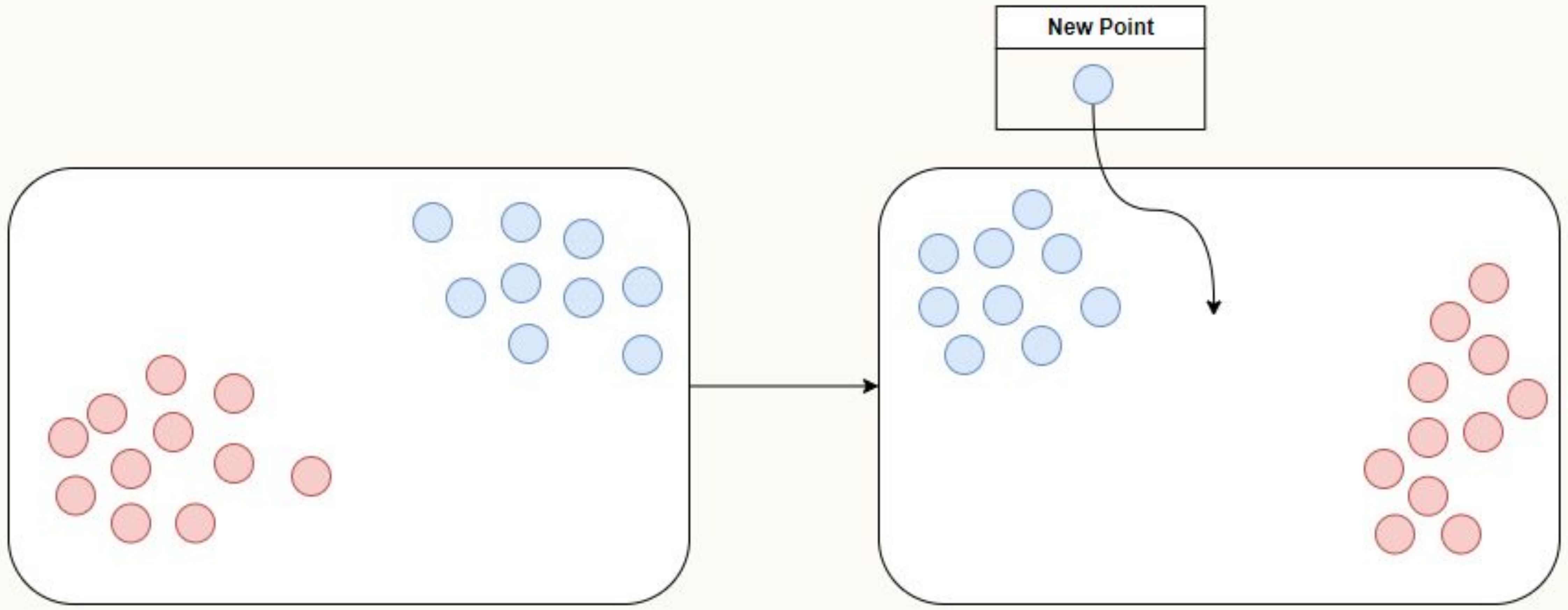
Bring all the benefits of dimensionality reduction to streams of data.

Related Work

In-Sample & Out-of-Sample

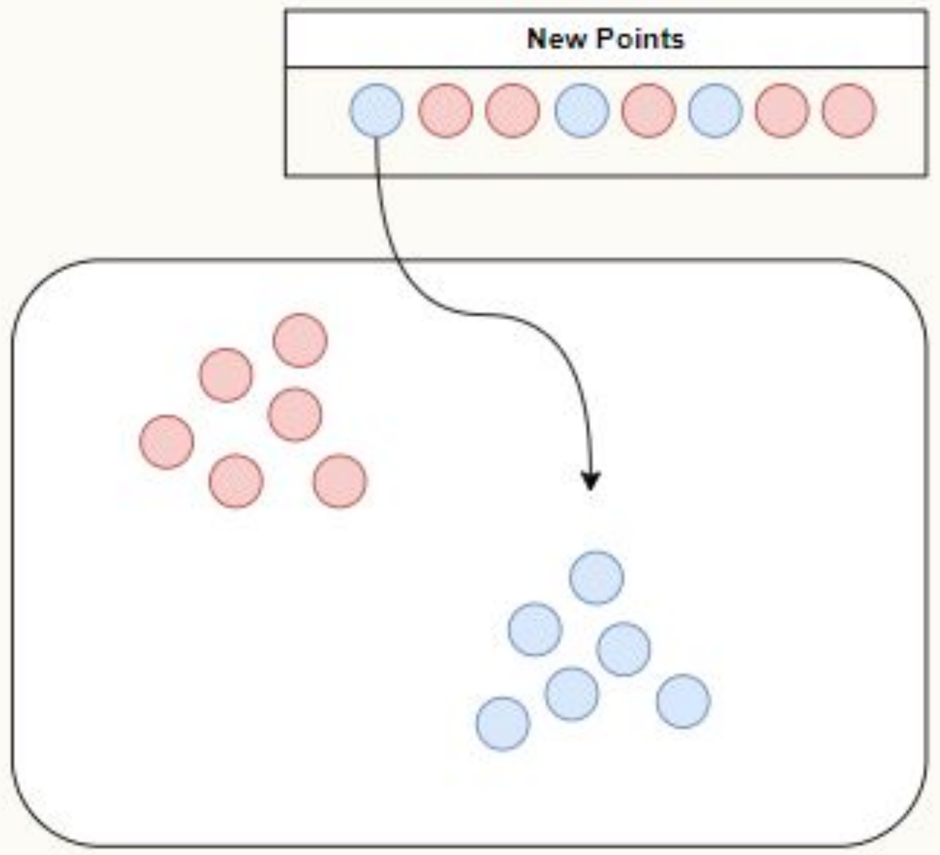
In-Sample

- ➔ t-SNE;
- ➔ UMAP;
- ➔ Classical Multidimensional Scaling (MDS);



Out-of-Sample

- ➔ Piecewise-Laplacian Projection (PLP) ;
- ➔ Least Squares Projection (LSP);
- ➔ Local Affine Multidimensional Projection (LAMP)



Problem Setup

Known approaches fail to deal with a large, possibly infinite, amount of data.

This can be attributed to either a failure to deal with the data due to a **lack of memory management** or their architectures **failing to extract meaningful information from ever increasing** amounts of data.

To hope to obtain a method to handle streamed data we must answer:

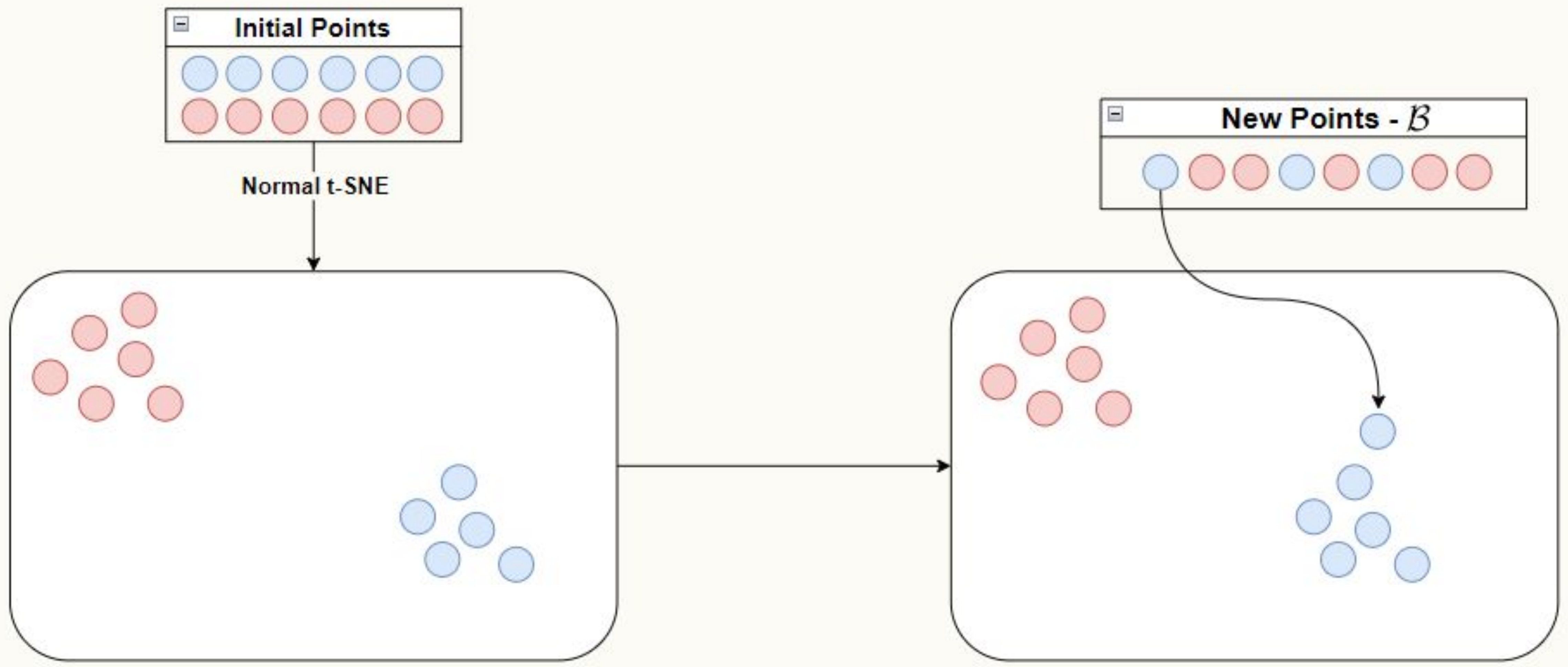
- When should we start operating on a dataset D if such dataset can be infinite?
- How can we reduce the space fingerprint of the algorithm while assuring the extracted data is meaningful?
- If D suffers concept drift, how can the existing projections be updated?

When should we start operating on a dataset D if such dataset can be infinite?

- ➔ Fixed batch-wise approach to mitigate challenges of infinite data-streams:
  - ◆ Points are accumulated until a predetermined batch size, $\mathcal{B}$, is attained;

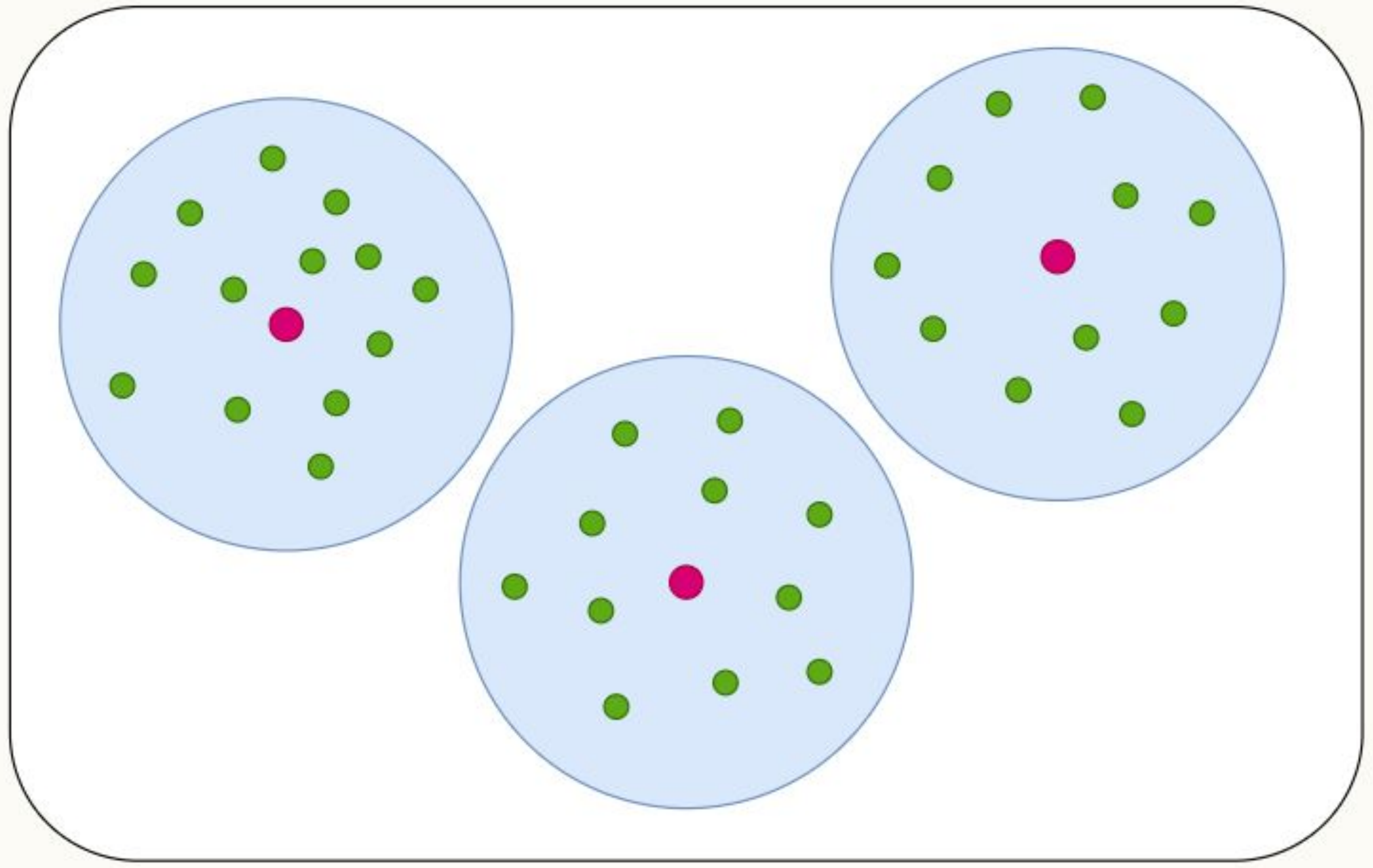
- ➔ Normal t-SNE will be applied in the first projection;

- ➔ However, after the first iteration of S+t-SNE, subsequent iterations will project in a space where points already exist;



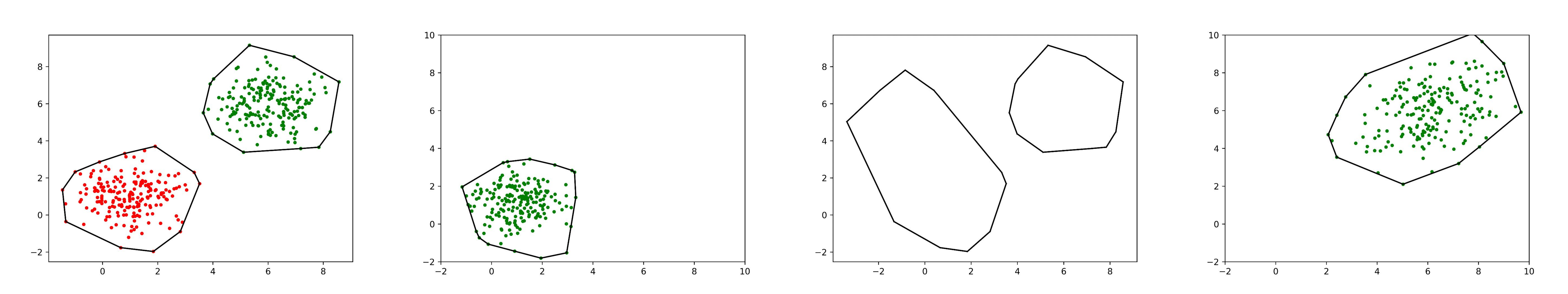
How can we reduce the space fingerprint of the algorithm while assuring the extracted data is meaningful?

- ➔ Select points based on their importance to describe the data (PEDRUL):
  - ◆ Have <u>Higher density in the original D-dimensional space</u> given a search radius **R**;
  - ◆ The neighborhood defined by **R** does not contain a subset of points from already chosen dense points;



➔ Define Regions of interest:

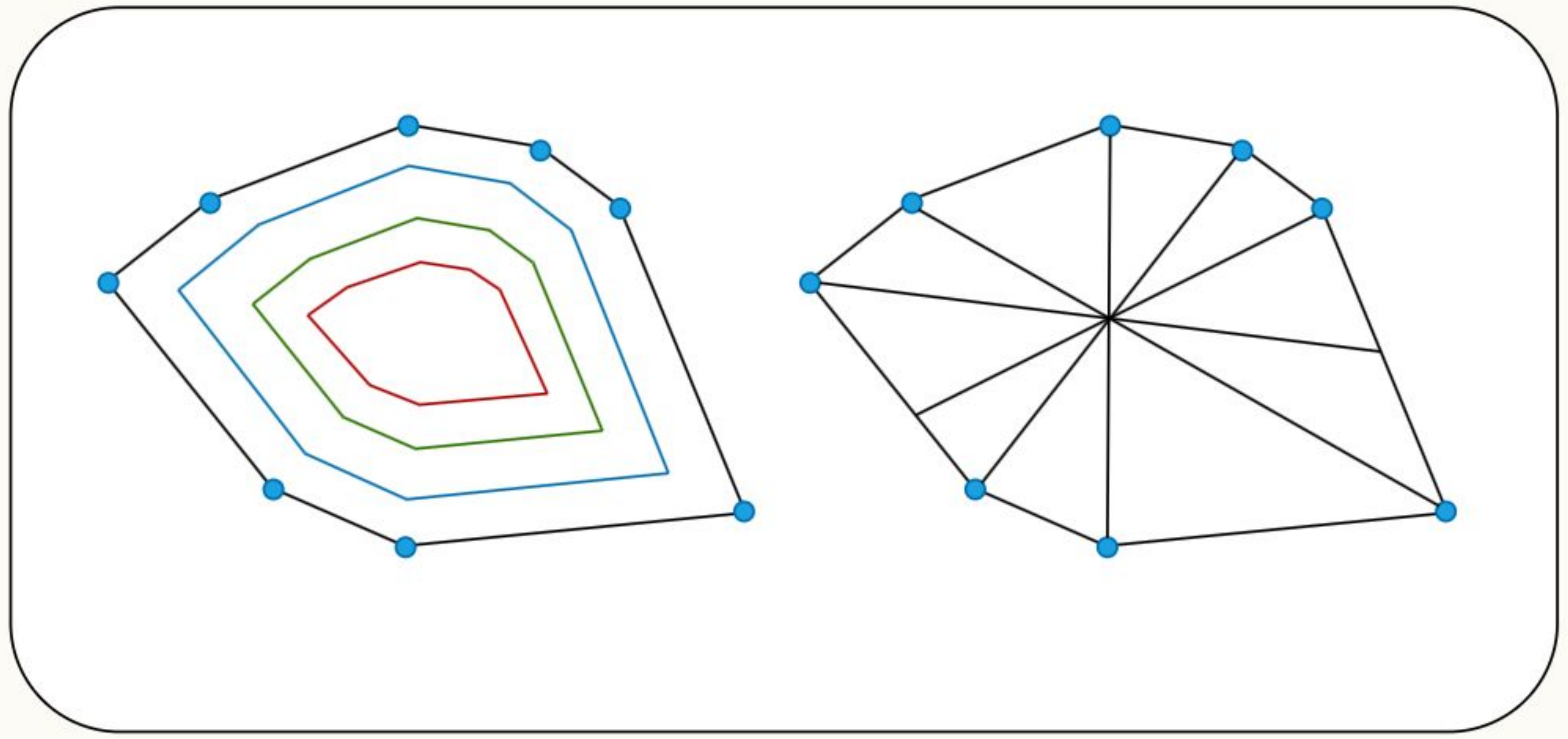
- Retaining the shape of groups of points by using a clustering algorithm applied to t-SNE projections;
- Construct a convex region around each cluster to retain the general shape of the clusters found;

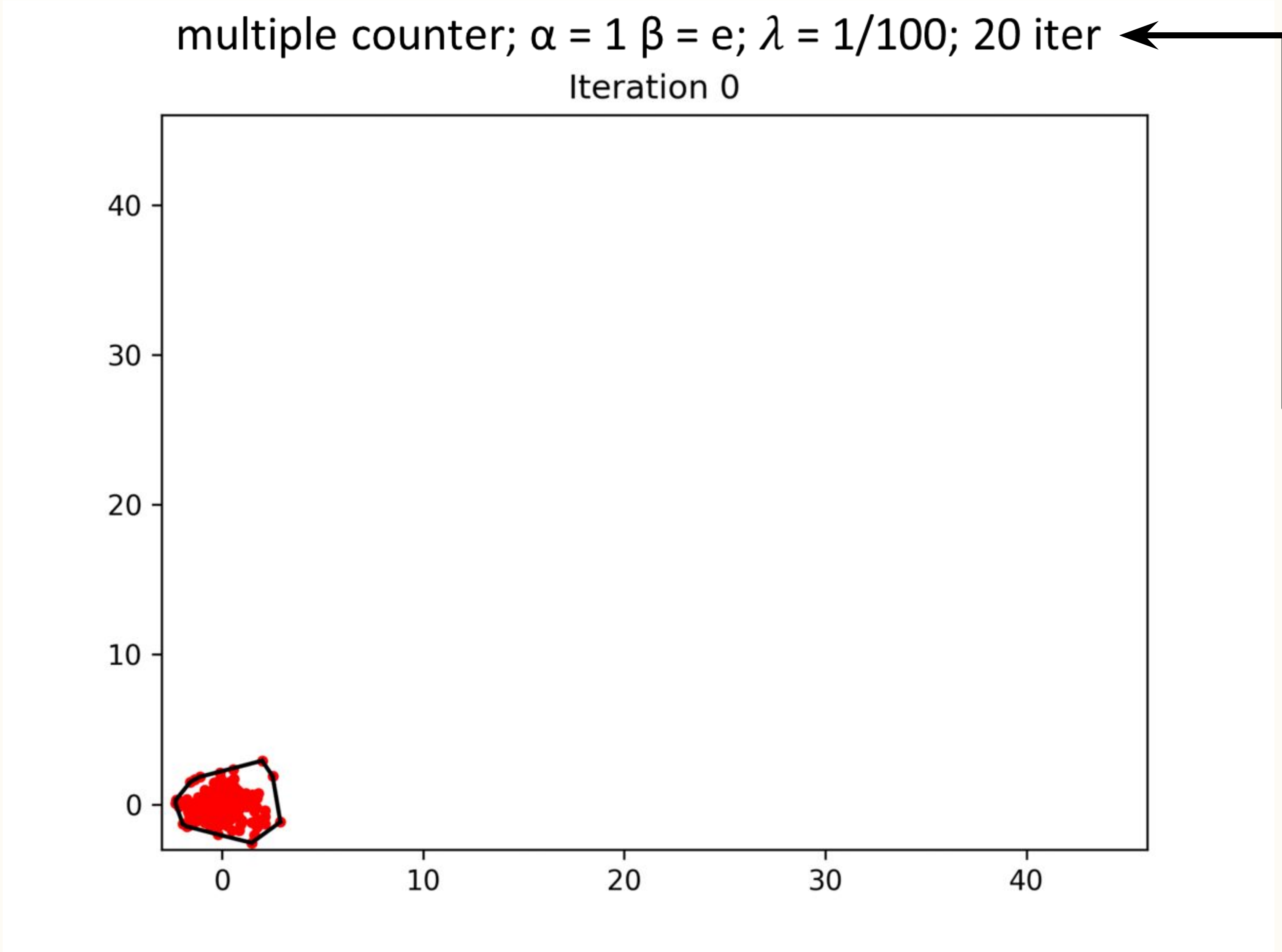


If D suffers concept drift, how can the existing projections be updated?

- ➔ Must be compatible with the approach described so far.
  - Convex polygons in, digest (look for drift), modified convex polygons out.
- ➔ Must cover the whole polygon while being efficient.
  - Divide each polygon in areas that, if removed, always result in a convex polygon.
- ➔ Must be able to evaluate the relevance of points to the projection in the current iteration
  - Blind drift detection using exponential decay for each of the spanned regions.

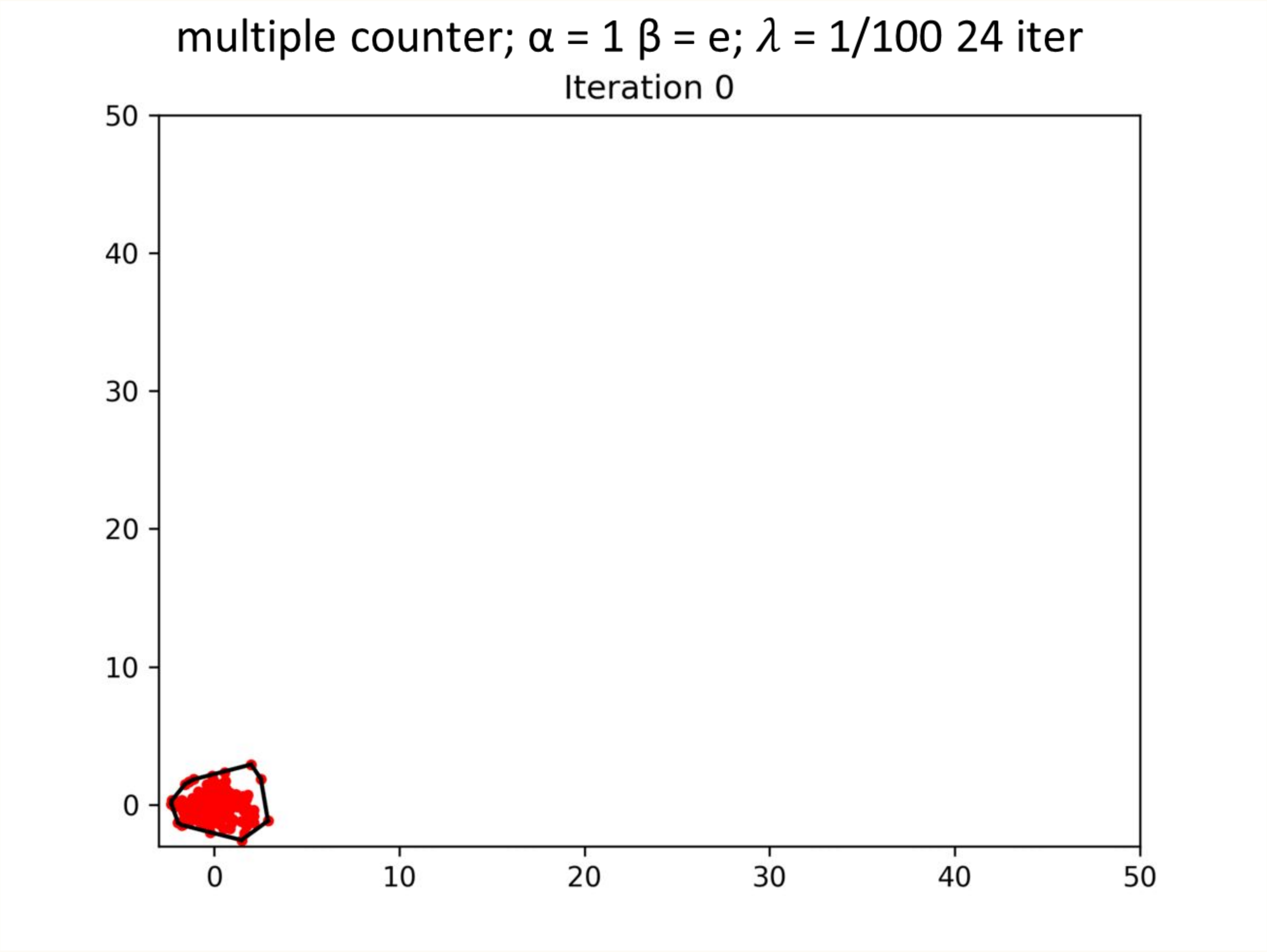
This is efficient! $O(mkp \cdot log\, n) + O(mkn \cdot log\, n)$, and, if parallel, $O(p \cdot log\, n) + O(n\, log\, n)$ where k is the maximum number of vertices in a polygon, p the number of polygons, n the maximum number of PEDRUL points in a polygon and m the number of concentric regions.

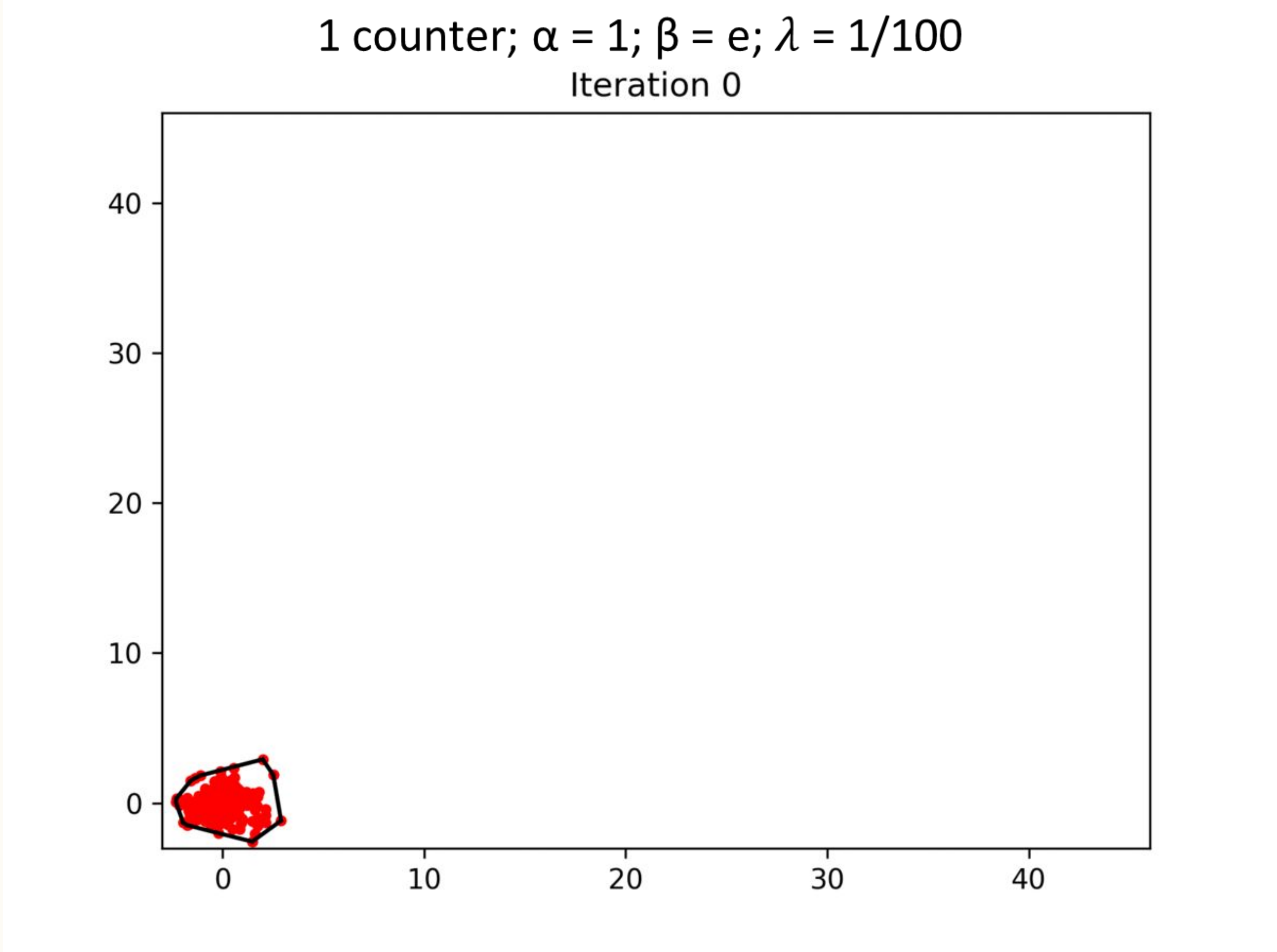




Updates of region using median cuts

$$\alpha\beta^{-\lambda t}$$

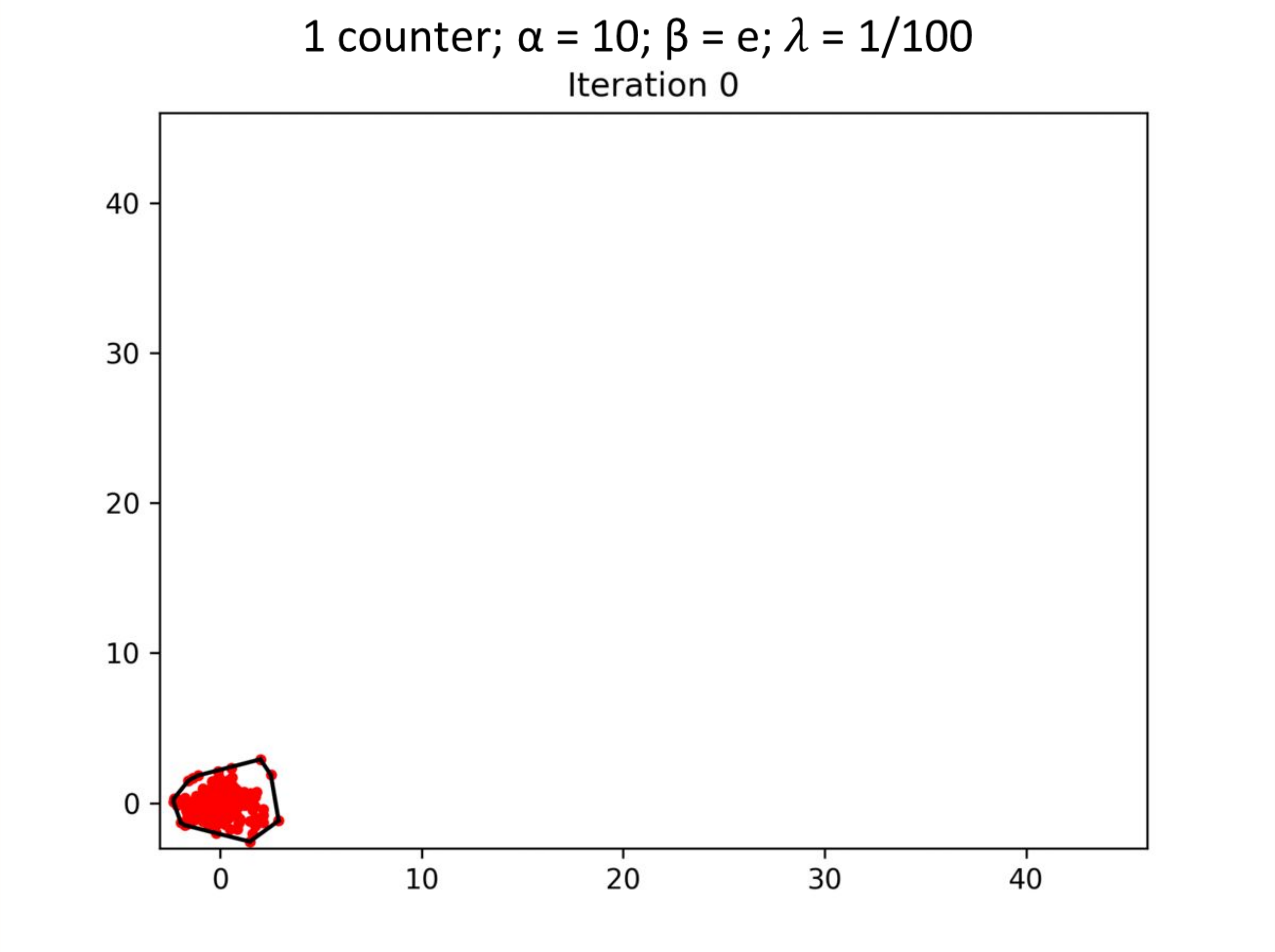




Different strategies

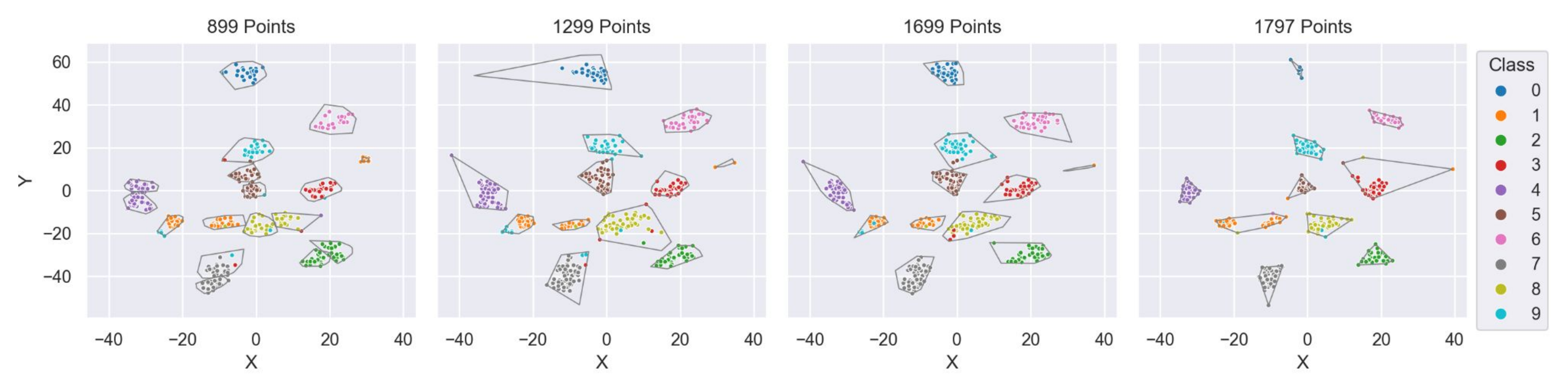
- In the single counter strategy, every region shares a counter. After a cut is made, the counter is zeroed.

- Other approach is to implement a counter per region and reset only the counter of the region it suffered an alteration.

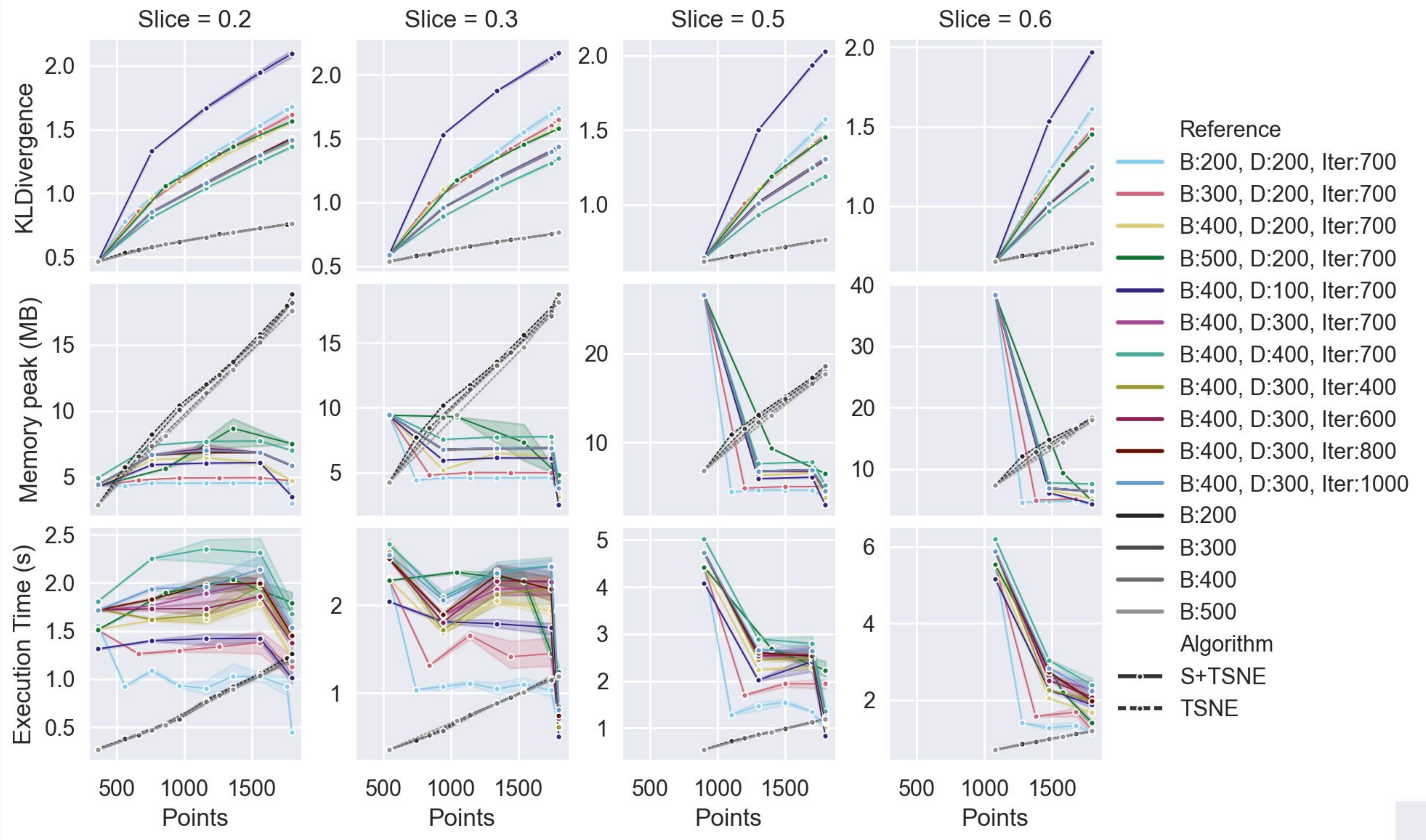


Experiments

➔ MNIST - Handwritten digits

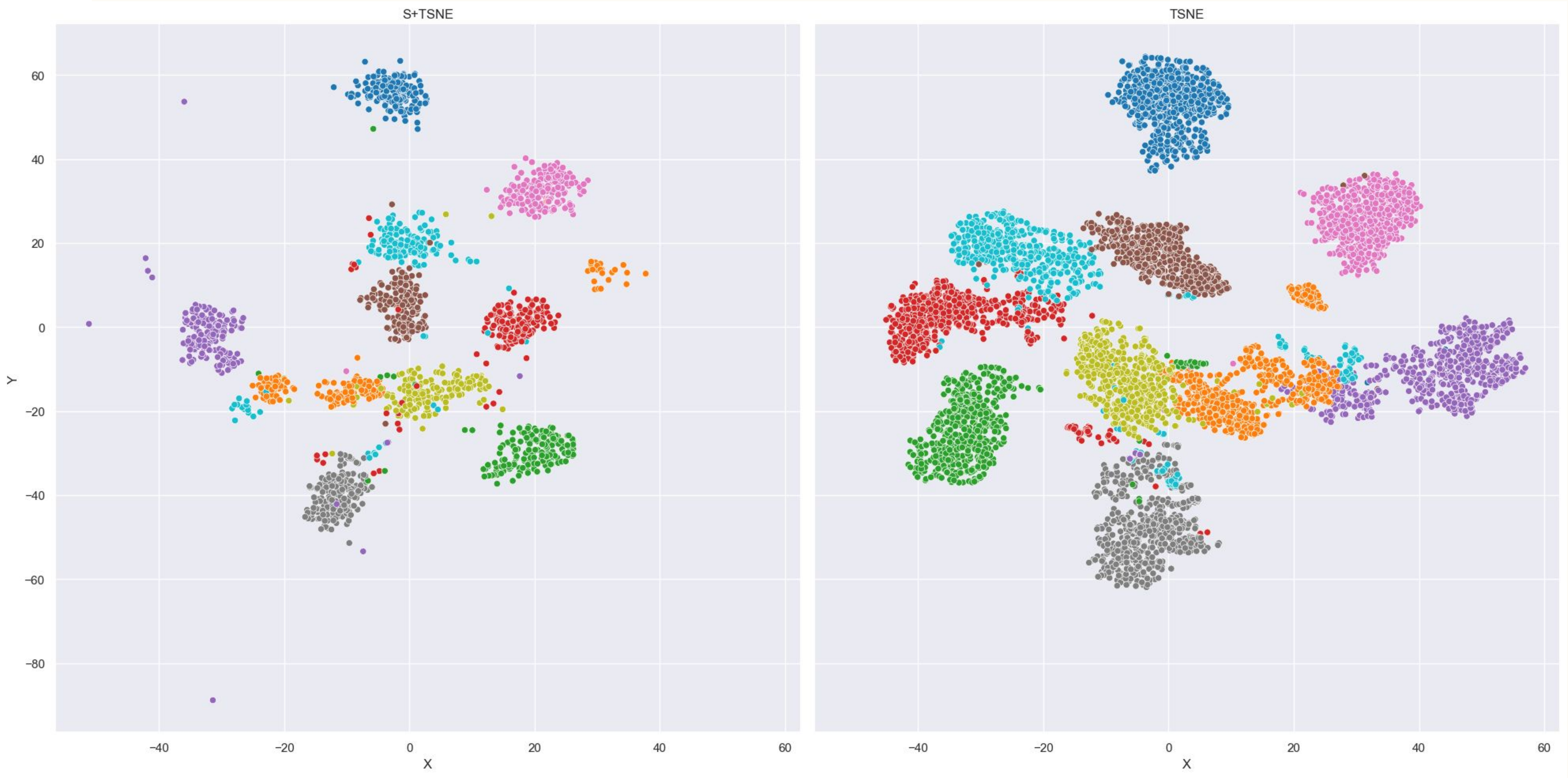


MNIST projections at different iterations of the streamed data.



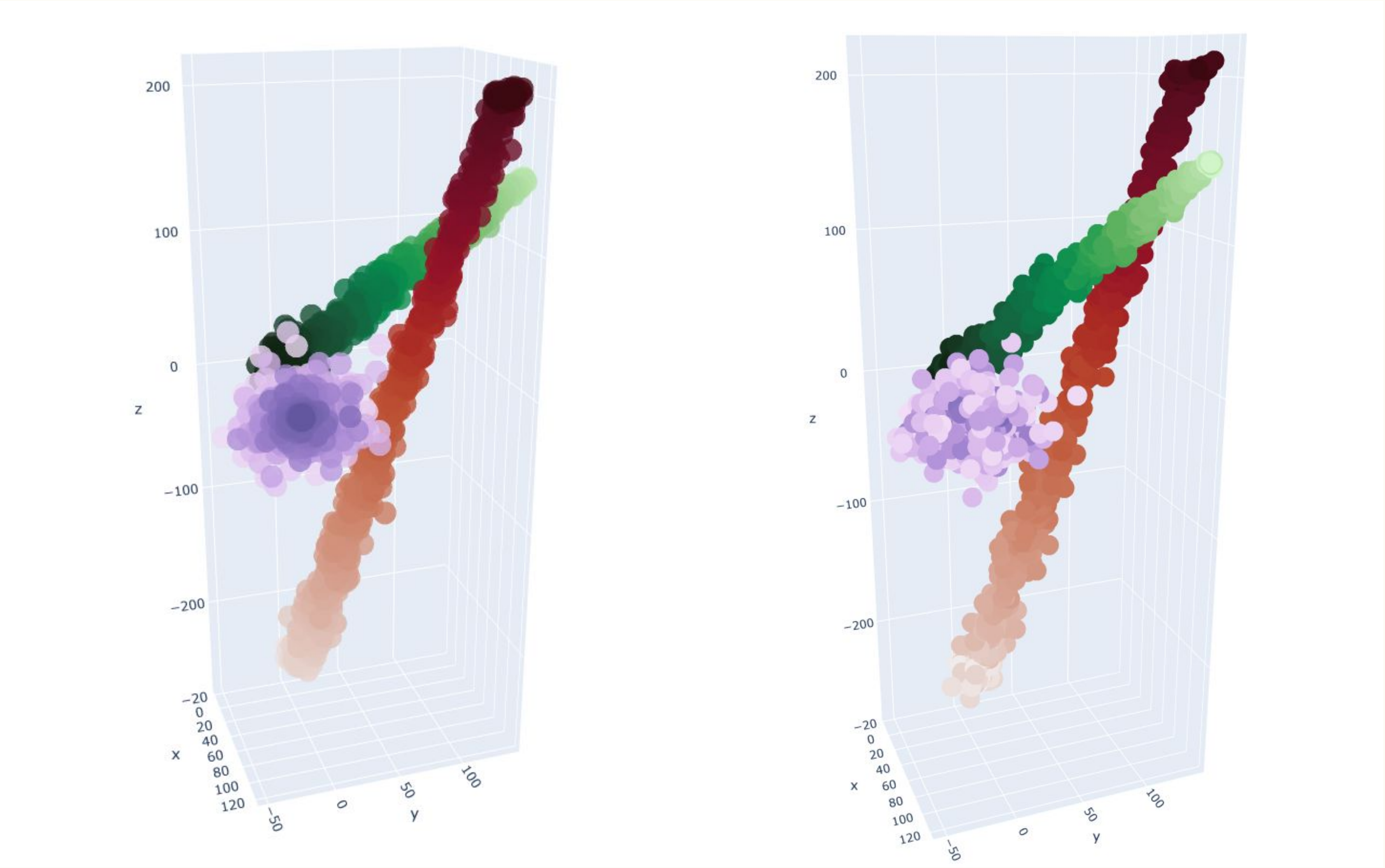
Evaluation of different set of hyperparameters on MNIST

Visual Comparison of S+t-SNE and t-SNE

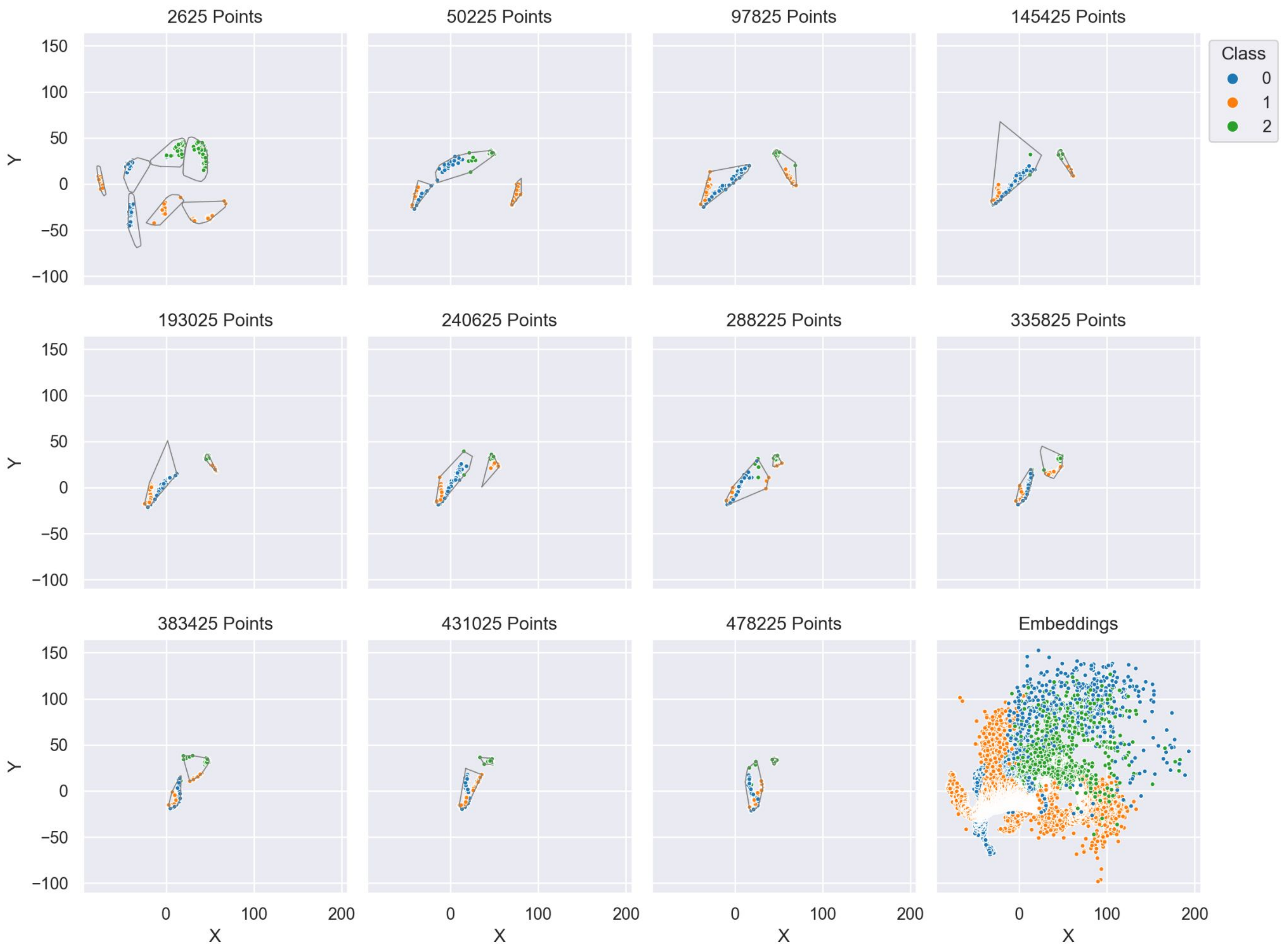


➔ Artificial 3D DataStream

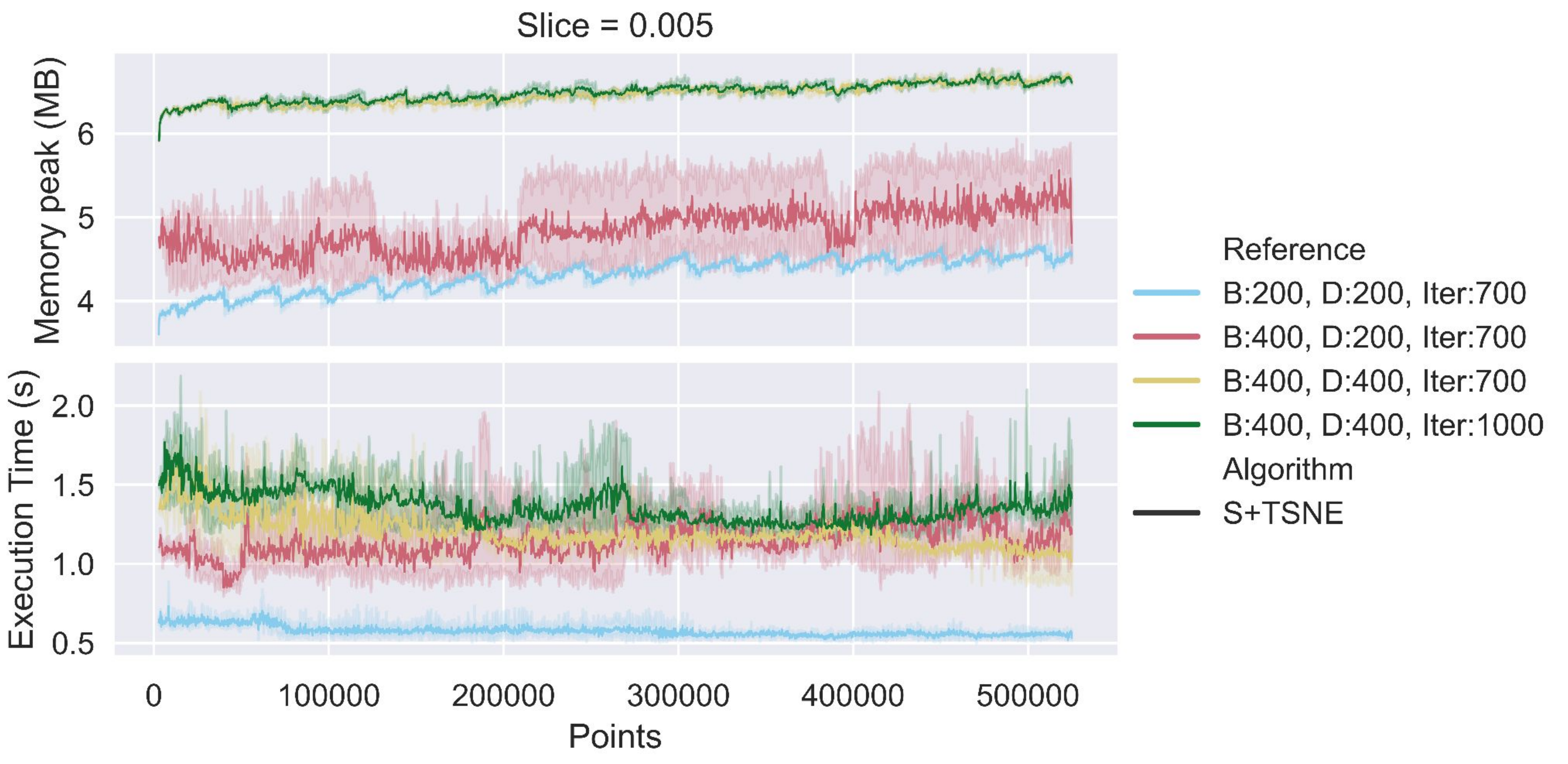
- ➔ Three 3D distributions with variability over time;
- ➔ Nearly 500.000 points to be embedded;



3D visualization of the DataStream simulated through time



Projections of S+t-SNE at different iterations of the streamed data.

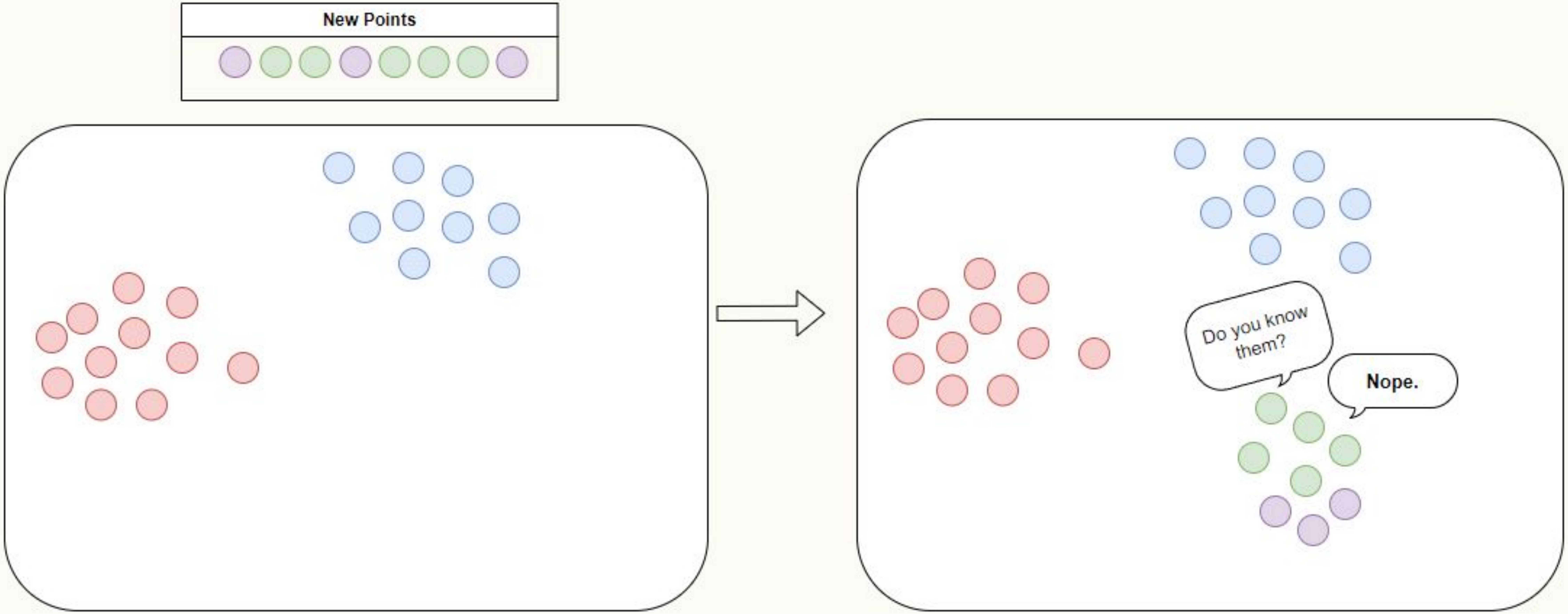


Evaluation of different set of hyperparameters the Artificial Data set

Known Limitations

➔ **"Lonely guy at a party"**

- If new points don't consider their intra-similarity but only their inter-similarity (with already projected data) points with low intra-similarity can end up close together.



Acknowledgements

- We thank the APPIA-Portuguese Association for Artificial Intelligence for financial aid for the travel expenses;
- João Gama acknowledges the support of the project AI-BOOST, funded by the European Union under GA No 101135737;

End.