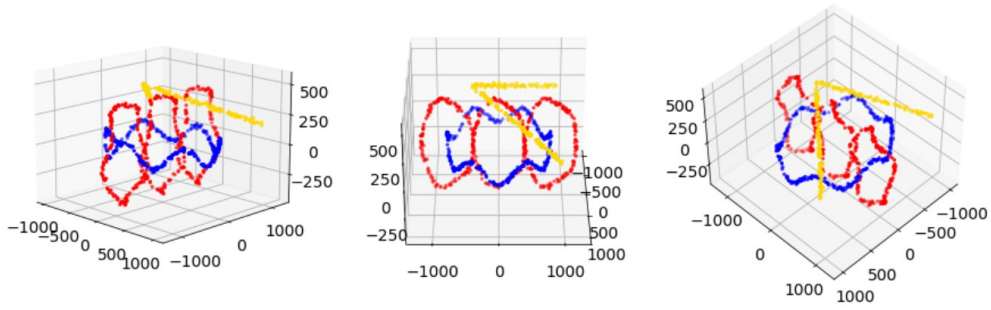
LCPB 22-23 Exercise 3, data visualization and clustering

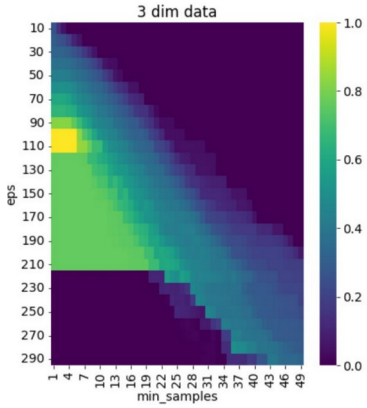
Exercise 4A

Consider data generated during the lesson, with N=600 samples.



1. Choose a perplexity for t-SNE and run a few times its algorithm by varying “random_state”. How do the results compare with each other?

2. *“eps” and “minPts” in DBSCAN algorithm for clustering*
   Refine the grid with more values of “eps” and “minPts” and show a heat-map of the normalized mutual information (NMI) between true and predicted clusters is varying.
   The result might look like this one on the right.

   Is there a correlation between these two parameters in providing a high NMI?

   Note: in the lesson we have looked at the typical distance between a point and its closest neighbor, but this does not say what is the typical distance from the 2nd, 3rd, …, “minPts”-neighbor.

Another possibility to consider for tuning “eps” and “minPts” is the plotting of ranked distances.