INFO 633 PROJECT B

Task 1:

Here is figure 1.1 shows that the workflow flow based on the MNIST10,000 dataset. The sample size we use is 3,000 samples.

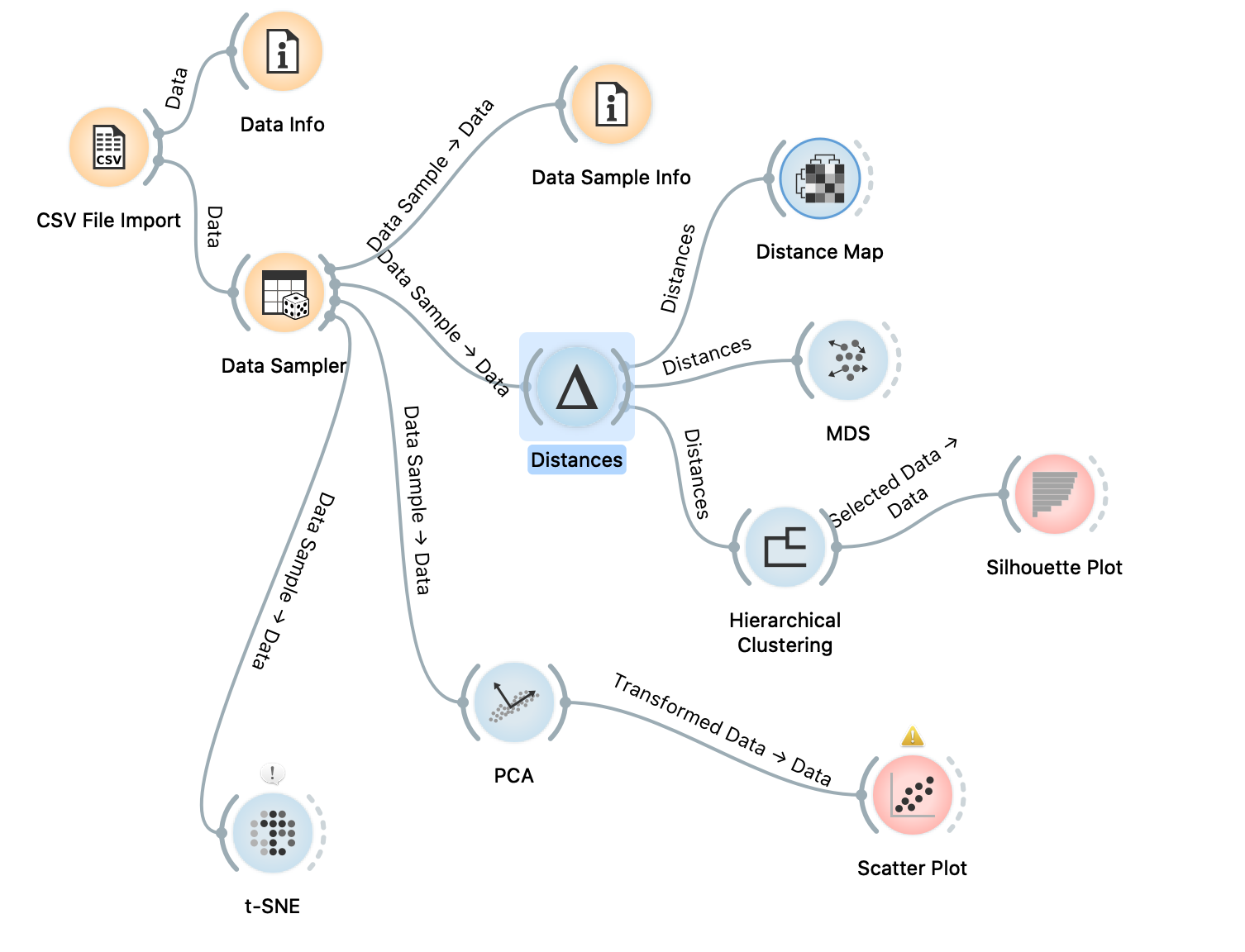


Figure 1.1 Workflow

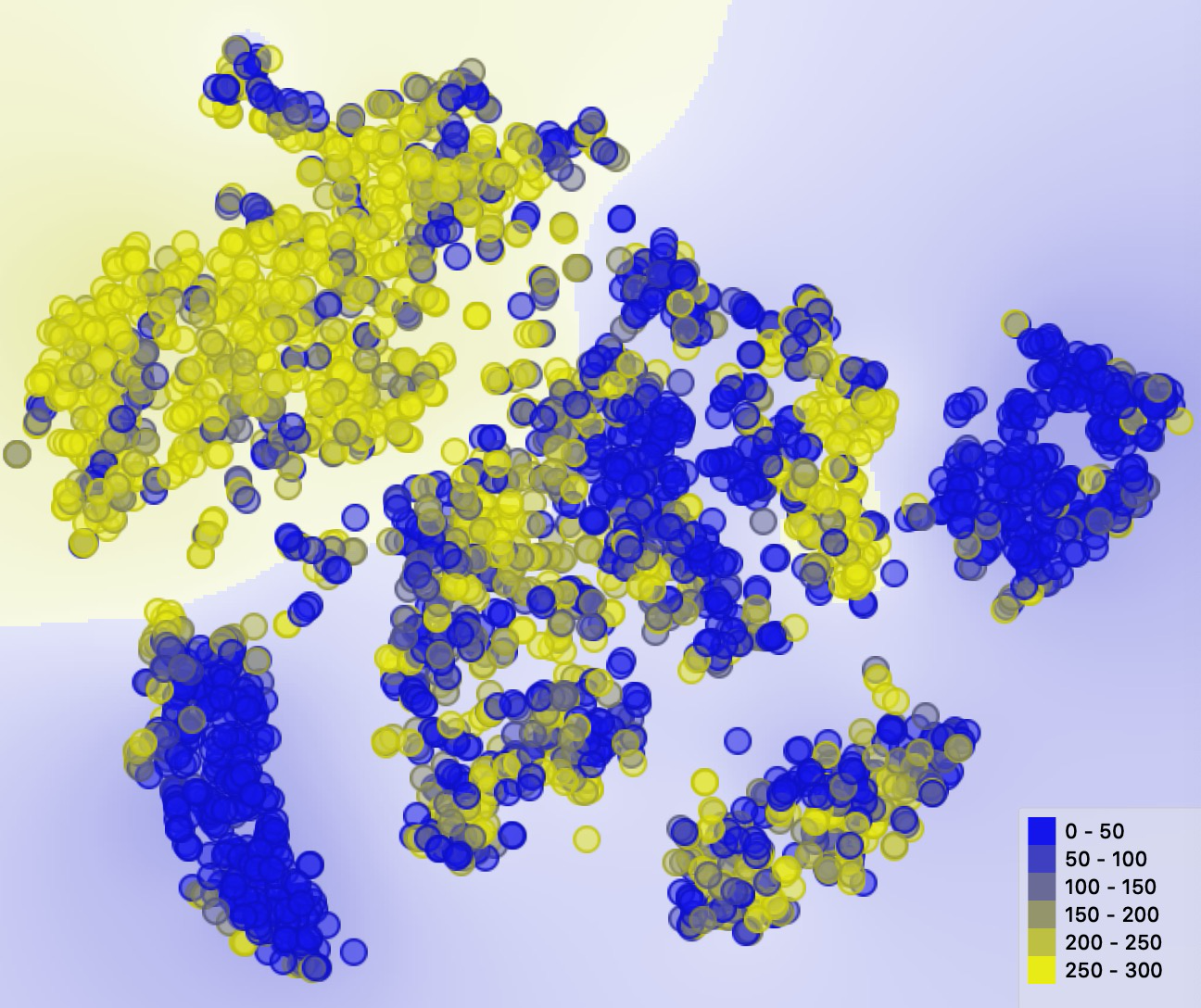


Figure 1.2 t-SNE

T-SNE is a process dynamic graph, figure1.2 shows its final state. T-SNE is one of the most effective methods of data dimensionality reduction visualization. When we want to classify the high-dimensional data set, but we are not sure whether the data set has good separability, we can use t-SNE to project the data into 2-dimensional or 3-dimensional space to observe: if it has separability in the low-dimensional space, the data is separable; if it is not separable in the low-dimensional space, It may be because the dataset itself is not separable, or the data in the dataset is not suitable for projection to low dimensional space. From figure 1.2, we can know that D1 is separable. In figure 1.2, 200 to 300 are mainly concentrated on the upper left, 0 to 50 are mainly concentrated on the lower left, and the rest are distributed in other places. These two clusters are significant because the two clusters are clustered, it is easy to see that the data is visual. On the other hand, The dots are placed based on the probability distribution and similarities, dots, and the cluster has relatively high tightness.

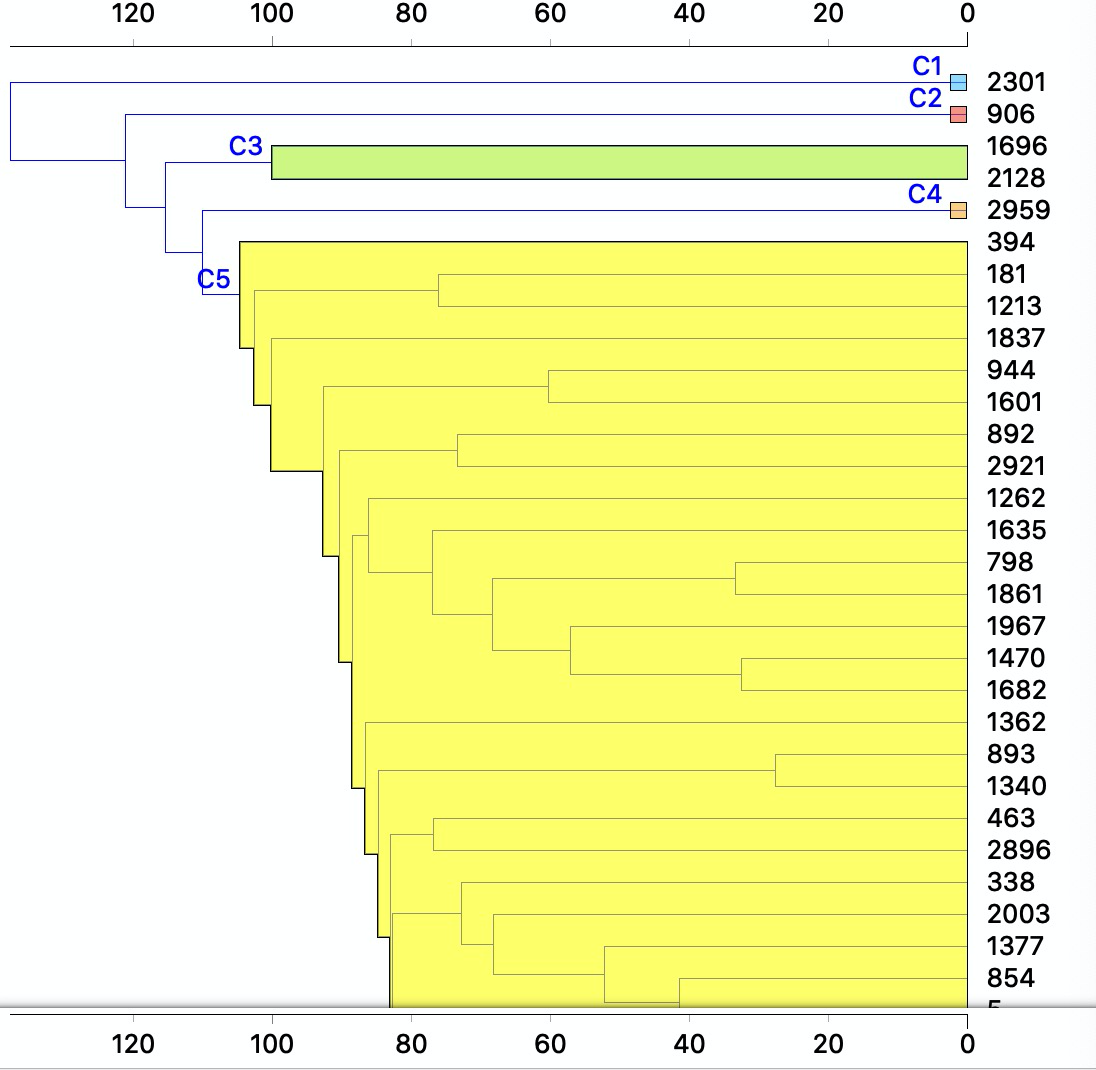


Figure 1.3 Hierarchical clustering

For Hierarchical Clustering, I used complete linkage, and selection is the top 5. Complete linkage takes the distance between the two farthest points in the two sets as the distance between the two sets. But the limit is very large. Even though the two clusters are close, as long as there are points that do not match, they will not be merged, which is not a good method. Complete linkage is only considered some data with characteristics, but it does not consider the overall characteristics of the data within the class. However, when I did other linkages such as single linkage and average linkage, their visualizations result is not better than complete linkage. In figure 1.3, we can divide the samples into five categories: C1, C2, C3, C4, and C5. The most interesting thing for me is that C5 is much bigger than the others. Compared to others, C5 has its own sub-classification and can be further divided if it is to be subdivided. And C3 is really larger than others, but when we do average linkage is not due to average linkage that doesn't take into account some characteristic data. So we can choose different linkages according to our needs.

Task 2: FAA Dataset

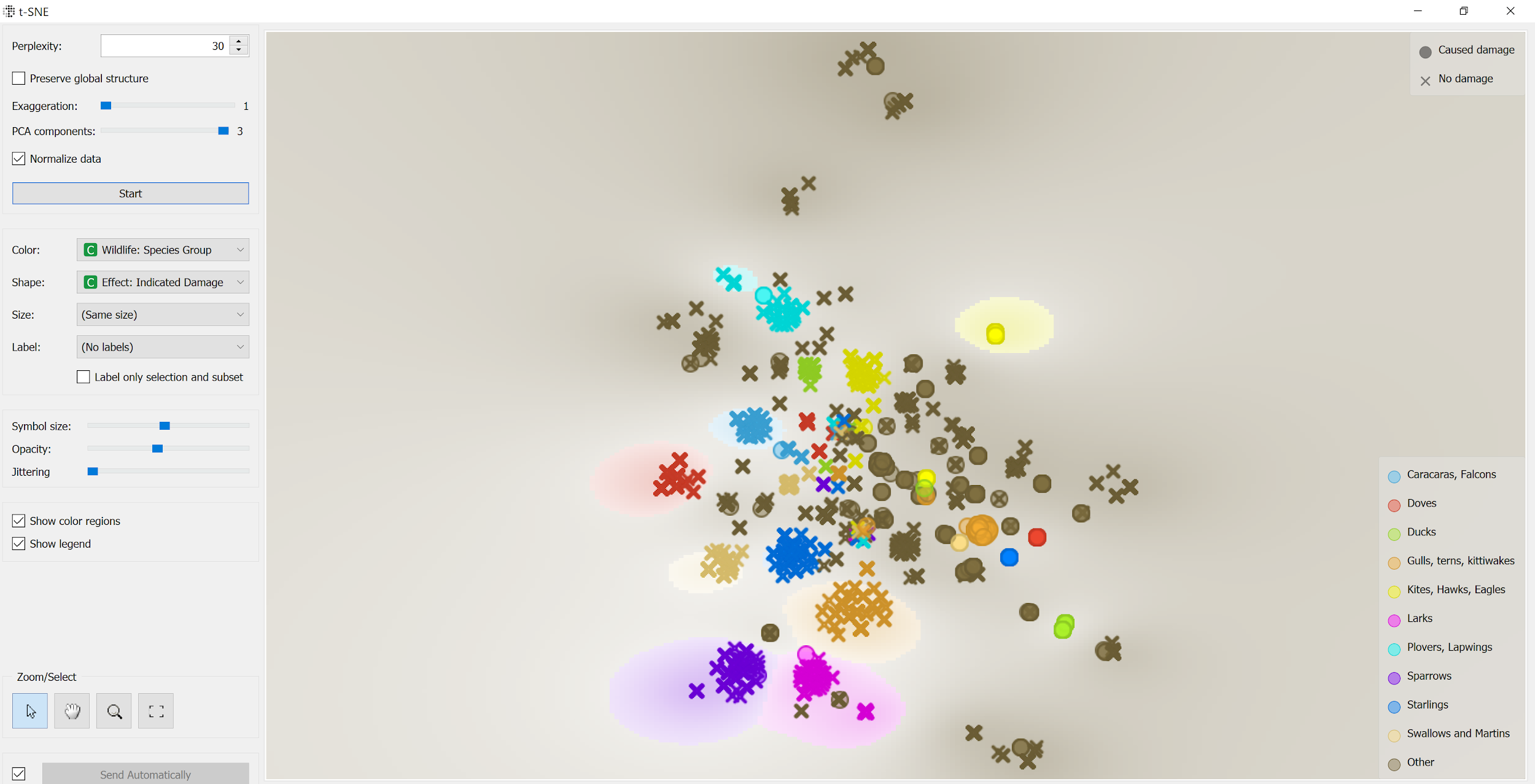


Figure 1.4 T-SNE

One significant finding here is that gulls, terns, and kittiwakes caused damage more than any other group of animals. Another significant finding is that the cluster of sparrows is tighter than the other clusters.

Task 3: Summarize

From this project, we have learned how to use Orange to create and apply workflows for visualizations of clusters in multidimensional datasets and unsupervised learning. To generate our t-SNE visualization, we imported our sample data and set our data sampler to 30%, then created the t-SNE visualization with parameters: 30 perplexities, 20 PCA components and normalized data. For Hierarchical Clustering, we need to calculate the distance before we generate the visualization. Based on the course content, we need to decide which one to use: cosine or euclidean normalized. In task 1, we are using euclidean normalized. In task 2, we are using cosine. Then we created our Hierarchical Clustering visualization with parameter: complete linkage, which uses the furthest cluster variance.

The relevant course contents such as lecture records and lecture slides are beneficial for us to finish these tasks. The contents gave us general directions such as how to create workflows and what values we should put into the parameters. Furthermore, we also learned what are the alternatives in different situations. For example average, single, complete, and ward linkage refers to average, nearest, furthest, or minimize stepwise within-cluster variance, cosine, and Euclidean normalized distances and distance map, etc.