# DATA.STAT.770 Dimensionality Reduction and Visualization, Spring 2021, Exercise set 12

## Problem G2: Neighbor Retrieval Visualizer

Familiarize youself with the command-line tool software for the Neighbor Retrieval Visualizer (NeRV; lectures 11-12) available at `http://research.ics.aalto.fi/mi/software/dredviz/`.

An R implementation may be available as part of the project `https://github.com/jlmelville/sneer` but I have not tested it. See problem G1 for discussion about stochastic neighbor embedding implementations; if you are unable to use the basic Stochastic Neighbor Embedding, leave it out.

Use the Neighbor Retrieval Visualizer to create visualizations for the swiss roll, half sphere, and Chronic Kidney Disease (from problem E6) data sets. Use several different tradeoff parameters between precision and recall (values between 0 and 1). Plot the results and discuss how the parameter affects the results.

Compute precision and recall tables/curves for the NeRV result as in Problem E7 (using 5 nearest neighbors of data in the original space as the "true neighbors" of each data point). Compute the curves for two NeRV results using two different values of the tradeoff parameters. Compute the same curve also for the SNE result from problem G1; discuss the differences between the SNE result and the NeRV results.

## Problem G3: Precision and Recall versus Distance Preservation

On lectures 11-12 it was discussed that the neighbor embedding method Neighbor Retrieval Visualizer (NeRV) can be seen as a method that optimizes information retrieval measures precision and recall, and its special case Stochastic Neighbor Embedding can be seen as a method that optimizes recall.

**Question 1.** Are the precision and recall measures essentially the same as "preserving small original distances on the display" and "preserving large original distances on the display"? Discuss. Hint: consider the "orange-peel map" visualization of the sphere surface shown on lecture 10. Does it achieve good precision and/or recall? What about good preservation of small/large original distances?

**Question 2.** On lecture 7, Curvilinear Component Analysis (CCA) was said to perform well in terms of precision. The cost function of CCA (see lecture 7) is based on distance preservation. How does this fit into your discussion in question 1? If there is a difference between CCA and the preservation discussed in question 1, what is the difference?