# UNIVERSITY OF CAPE TOWN
# DEPARTMENT OF STATISTICAL SCIENCES
# STA5077Z – Unsupervised Learning 2023
# ASSIGNMENT 2

Dimension reduction methods, PCA, MDS, Autoencoders, SOMs

Due date: Monday, 10 November 2023 at 12:00 (noon)

## INTRODUCTION:

High-dimensional data can be a challenge to analyze, almost impossible to visualize, and expensive to process and store. In many cases, the high-dimensional data points may all lie on or close to a much lower-dimensional surface, or manifold, implying the true or intrinsic dimensionality of the data is much lower. In that case, the data could be described with fewer dimensions, thereby minimising the impact of the curse of dimensionality. Transforming the high dimensional representation of the data to a lower-dimensional one without losing important information is the core objective of dimensionality reduction. Many methods of dimensionality reduction have been developed. Examples including the classical techniques of Principal Component Analysis (PCA), Multidimensional Scaling (MDS) and newer methods such as Autoencoders. Most of these methods often perform well on some types of data but poorly on others.

## DESCRIPTION OF THE DATA:

You are going to work with two datasets.

The first dataset named notMNIST_small is a sample of pictures of letters. Each letter is contained in a separate folder. You will only use letters C and G. Here is one example for each letter.

The dataset for C and G can be accessed from the following folder:

One Drive Link (password in a separate announcement)

The second dataset named **WDBC.csv** is the Wisconsin Diagnostic Breast Cancer dataset (source: UCI Machine Learning) consists of 569 data points classified as either malignant or benign. Data was computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Each instance contains 30 features describing different characteristics of the cell nuclei present in the image.

## THE GOAL OF THE ASSIGNMENT:

The goal of the assignment is to apply different dimensional reduction methods to different types of data in order to determine which methods and parameters work best on different types of data. To evaluate the performance of the reduction method, you will classify the data using the KNN algorithm. The algorithm will be applied first, when the data is in the original dimension and second, when data is in the reduced dimension. The difference in the results will be used to evaluate the impact of reducing the dimensions on accuracy. The dimensionality reduction methods to be tested are: PCA, AUTOENCODER, and SOM.

**INSTRUCTIONS:**

1. Once you have cleaned and pre-processed your data use the KNN algorithm to classify the handwritten letters into C and G. Similarly do the same with the breast cancer data, classifying the observations into benign or malignant. In both cases determine the classification accuracy. Try different values of K in the KNN algorithm until you find the value of K that results in the 'best' accuracy for each type of dataset.
2. Perform dimension reduction on both datasets using PCA, AUTOENCODER and SOM and then attempt to classify the data with the reduced dimensions using KNN as in (1). Consider cases where the reduced number of dimensions is: 2, 3, 4, 5 and 6. Briefly, investigate and discuss the results relative to the results obtained without dimension reduction.
3. Which dimension reduction method would you recommend if your goal was to classify the two handwritten digits using the KNN?

**REPORT FORMAT:**

- Present your final report as a pdf document. You may use any typesetting software you wish, but would encourage you to use LATEX.
- Number of pages excluding appendices and R code not to exceed 20!
- You may NOT provide R output interspersed between your answers! Please typeset relevant elements in the output either in-line, or tabulate results formally. Plots can be very useful, but use them sparingly – make sure that a given plot is relevant to the question and pertains to text in your answer. Figures are meant to enrich your analysis, don't leave it to the reader to analyse. Provide captions for all figures and tables. Square figures only!
- When you typeset R code use courier or an equivalent 'typewriter'- like font.
- You are expected to work on this on your own. Please attach a plagiarism declaration to your report – a template is attached below.
- Please submit your report through the assignment tab in Vula.