# CIS*6030 Information Systems

Fall 2022

Instructor: Fangju Wang

## Assignment 4 (100%)

File *iris.csv* contains data of three different species of iris flowers [1]. The dataset contains five columns and 150 rows. Each row represents a flower instance. Columns 1 to 4 are four attributes of the flowers. They are sepal length, sepal width, petal length, and petal width. Column 5 contains the species of iris flowers: setosa, versicolor, and virginica. Different species have different lengths and widths of sepal and petal. The species of an iris flower can be identified from the attributes.

File *multishapes.scv* contains 1,100 points of clusters of different shapes [2]. It is Useful for comparing density-based clustering and traditional partitioning methods such as k-means clustering.

**Assignment requirements:**

1. Store *iris.csv* and *multishapes.csv* in the HDFS your created for A2.

2. Write an R program that each time reads a file from HDFS, extracts the feature columns of the data, invokes a *built-in* function to conduct K-Means clustering on the data to cluster the data, and visualize the clustering result. (25%)

3. Write an R program that *implements* the HDBSCAN algorithm [3]. Each time the program reads a file from HDFS, conducts HDBSCAN clustering on the data, and visualize the clustering result. Don't use any built-in clustering functions in your program. (70%)

4. Briefly compare the two programs for clustering the two files. Write your comparison in your README file. (5%)

**Recommendations:**

- Install *RStudio*, and use it to test and run your R programs, and display the clustering results.
- Install *Rhdfs*, and use it to connect your R programs with HDFS, and access data stored in HDFS.

**Submission:**

Submit your work as a tar file by the end of Nov 21, 2022 (Monday). The tar file should include a README file containing the comparison in question 4 and telling anything you think worth to mention. Don't submit the data files.

**Sources:**


1 Cluster analysis with iris data set,
https://medium.com/swlh/cluster-analysis-with-iris-data-set-a7c4dd5f5d0

2 Multishapes: A dataset containing clusters of multiple shapes,
https://rdrr.io/cran/factoextra/man/multishapes.html

3 How HDBSCAN Works,
https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html