# Homework 3: Unsupervised Learning

Checkpoint Due Date: 08:00 on November 29, 2023 [50 pts.]

Final Due Date: 18:00 December 13, 2023 [100 pts.]

## Problem Statement

Apply an unsupervised learning approach to your data. Choose one of the following approaches:

(A) clustering or principal component analysis (PCA) for dimension reduction prior to input/output into supervised learning approach from HW1 or HW2[1]

(B) clustering or PCA to interpret XAI results from HW1 or HW2

(C) autoencoder for dimension reduction or de-noising or as input into a supervised task

(D) generative adversarial network (GAN)

## Data

The data you use is completely up to you, however, if you choose (A) or (B) you will need to use the same data as you did for the previous homework. Some things to keep in mind:

- Machine learning is data hungry. Too little data and you've already set yourself up for failure.[2]
- Too much data and your code will take a long time to run. You will likely be running your code dozens of times, so this extra time can add up - but more (good) data is always better.
- When learning a new machine learning approach it is good to start with a data set that you generally understand so that you can identify when the results are out-to-lunch[3]. At the same time, choose a data set where the relationships may be complicated so that you can fully explore the power of the machine learning approach.

## Final submitted Write-up

Submit a full write-up on your scientific process and results by the Final Due Date listed above. *You should not submit your code.* The write-up should include text and relevant figures that cover the following:

- scientific motivation and specific problem statement
- description of the data
- description of any data pre-processing performed and why you did it
- training/validation/testing split
- machine learning setup and reasons for hyperparameter choices when relevant
- discussion of results
- a detailed discussion of why you think the results are meaningful

[1] you should then re-train your supervised learning approach and compare with your original results

[2] Alas, I cannot tell you what is "enough" data - that is something you will need to determine for your data sets

[3] Not in touch with the real world, crazy.

- concluding thoughts on insights gained from your efforts
- link to your github repository

## Github Repository

As part of this homework you are also expected to use git/github and include a link to your public or private github repository. Your repo must show *at least two commits* by the Checkpoint Due Date and *at least four commits* by the Final Due Date. Your code does not need to be complete, and we will not assess your code, but we do want to see that it is there. A few additional notes:

- If you choose to make your repo private, please be sure to add me (eabarnes1010) and Charlie (connollyc152) as Collaborators so that we can see your repo.
- Do not upload your data! Github does not deal well with large files / data sets. Just commit and push your code (.ipynb, .py, etc).

Also, some useful links:

- GitHub: `https://github.com`
- Git and GitHub Start-up Guide
- Git cheat sheet: `https://education.github.com/git-cheat-sheet-education.pdf`
- GitHub Desktop: `https://desktop.github.com/`
- ROSSyndicate Best Practices
- ATS GitHub Tutorial run by Justin Hudson Spring 2023