**Unsupervised Learning– Hierarchical Clustering:**

Rob Harris

25.4 Assignment for Thinkful’s data-science flex certification program.

In this assignment, you'll continue working with the [**heart disease dataset**](http://archive.ics.uci.edu/ml/datasets/Heart+Disease) from the UC Irvine Machine Learning Repository.

Load the dataset from Thinkful's database. To connect to the database, use these credentials:

postgres\_user = 'dsbc\_student'

postgres\_pw = '7\*.8G9QH21'

postgres\_host = '142.93.121.174'

postgres\_port = '5432'

postgres\_db = 'heartdisease'

The dataset needs some preprocessing. So, before working with the dataset, apply the following code:

*# Define the features and the outcome*

X = heartdisease\_df.iloc[:, :13]

y = heartdisease\_df.iloc[:, 13]

*# Replace missing values (marked by `?`) with a `0`*

X = X.replace(to\_replace='?', value=0)

*# Binarize y so that `1` means heart disease diagnosis and `0` means no diagnosis*

y = np.where(y > 0, 1, 0)

Here, X will represent your features and y will hold the labels. If y is equal to 1, that indicates that the corresponding patient has heart disease. And if y is equal to 0, then the patient doesn't have heart disease.

To complete this assignment, submit a link to a Jupyter Notebook containing your solutions to the following tasks below. You can also take a look at these [**example solutions**](https://github.com/Thinkful-Ed/data-201-resources/blob/master/clustering_module_solutions/4.solution_hierarchical_clustering.ipynb).

1. Produce dendrograms for the heart disease dataset. Use three different linkage methods: complete, average, and ward. Which linkage method do you think produces the most reasonable result?
2. Apply agglomerative clustering to the heart disease data by setting n\_clusters=2. Try the three linkage methods above, and get ARI and silhouette scores for each of your solutions. Compare the results with each other. Then compare the results with the results from the k-means solution that you implemented in the previous checkpoint's assignment. Which algorithm and setting perform better?