**UC Davis BootCamp**

**Iris Data Set**

Student Members: Andrea Castro, Cyril Sambrano, Colin Leung, Sheila Brear

The iris plant is a species of flowering plant, known for its variety of colors. The iris dataset is a well-established dataset that is used to display pattern recognition using **linear discriminant analysis (LDA)**, which is a statistical method that can discriminate between two or more independent variables. In this case, the variables are continuous, i.e., they have many possible numerical values. This technique is primarily useful when determining whether a set of variables can predict the category of membership, in this case in a class of plant, the Iris.

The original dataset describes measurements made of the different parts of the plant:

Sepal: This is an outer structure of the iris bud, that protects the bloom before it opens. The shape and color of the sepal attracts pollinating insects to the plant.

Petal: This structure develops inside the sepal, and protects the reproductive, pollen-receptive reproductive structures of the plant.



The iris dataset is a series of measurements of the length and width of the sepals and petals in three different species of iris: the Iris-setosa, the Iris-versicolor and the Iris-virginica. There are fifty samples of each of two measurements (petal and sepal) of each of the three species in the one dataset.

The following questions will be answered in the data analysis that this group will perform.

1. Exploratory data analysis: is there a difference in the length of the petals of the three different species?
2. Scatterplot: Do the different species display different lengths and/or width of sepals and petals?
3. Decision tree: Can the different attributes of the sepals and petals predict the species?

*We could also use a null hypothesis instead of asking questions, eg, “there is no difference in length of the petals of the three different Iris species”, then disprove this.*

*Can we display the scatterplot in Tableau?*