# Python for Data Analysis

UAV Intrusion set, Alexandre Autret, DIA1

# Goal of the project

The goal of this project is to design the best possible classification model for a given dataset and compile it into a report.

In this case, the dataset we were given is the *«Unmanned Aerial Vehicle (UAV) Intrusion Detection Data Set»*.

The URL we were given, *https://archive.ics.uci.edu/ml/datasets/Unmanned+Aerial+Vehicle+%28UAV%29+Intrusion+Detection*, didn't have it available for download, so we had to get it from the source, at http://mason.gmu.edu/~lzhao9/materials/data/UAV/.

# The dataset

The dataset is made up of 6 separate tables, split into 2 categories:

- The first 3 contain bidirectional-flow mode (uplink, downlink and both-link) information about a Bebop Parrot, a DBPower UDI and a DJI Spark drone, respectively.
- The other 3 only contain unidirectional-flow mode (all 3  communication modes mashed into 1), for the same drones.

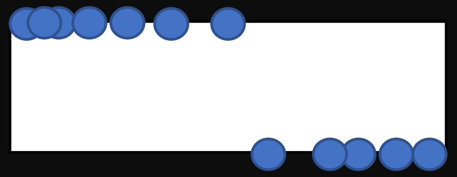
Therefore, the first 3 have 3 times as many features as the other 3: 54 vs 18.

The target variable is simply called « label »: if the value is 1, it means that a drone has been hijacked, if it’s 0, then it hasn’t.

Since we can only design a better model with our first 3 tables, we disregard the other 3: we can always remove features, but we can’t create any out of thin air.

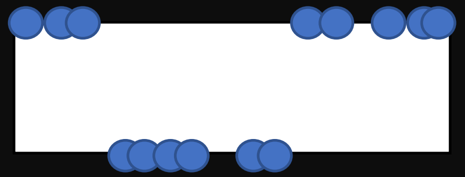
# Our thought process

Our thought process was as follows: we'd use seaborn's pairplot function to see how each feature looks in comparison to our target, and if we noticed any pattern, we'd try models that work well for it. For example, if we saw plots that looked like this:



We'd try a logistic regression.

If we saw plots that look like this:



We'd try a QDA, a decision tree or a random forest, and fiddle with the hyperparameters to make our model as accurate as possible.

# Our code and results

Now that the scope of the project is set, here's how we'll organize the project:

- We'll build the code and report in a Jupyter notebook (.ipynb), that we'll also export as an html file for easy access.
- The datasets, HDF5 .mat files, will have to be read using the h5py library, as they can't be read using default tools.
- The project folder will be made available on GitHub (https://github.com/Alexandre-Autret/pda), with an accompanying README containing our conclusions.