## 1. Performance estimates of different ML classifiers

Use the known classifications in the 3FGL catalog to estimate the performance of different **ML classification algorithms: DT, BDT, random forest, ensemble of three, neural nets.**

**Other algorithms?**

1. Calculate the score and **“loss rates”** for **Galactic and extragalactic sources (what is loss rate)**.

* Galactic vs. Extragalactic (How does one define this? Is this based on GLAT?)
* AGN vs. Pulsars (Easy distinction to make and first prediction)
* Young vs. msp pulsars (second distinction based on external catalogues)
* FSRQ vs. Bllacs (third distinction; multiple label and based on FLAT data)

1. Study the performance of the algorithms by changing complexities, i.e., the depth of the trees, the number of trees in the random forest etc.

For Forest and tree based:

* + Number of trees
  + Depth of trees
  + Weighting (usually we use weights)
  + Reshuffling the input (expected to not fluctuate with correct weight)

For NNs and algos like Logistic regression:

* + For NNs: Number of hidden layers, epochs, learning rate/tolerance
  + Weighting (again expected to be weight dependent for the external predictions)
  + Activation function (usually softmax, tanh); This is also for other methods
  + Probabilities ( This is not a change of complexitites but rather allows us to predict better
  + NN in Keras vs. NN in sklearn. What is the difference, if any? And is it okay to not use Keras? A good example would be for instance to use the keras pipeline which the exact same as that in sklearn and see if there is any difference in accuracy.

1. Try different weighting of the training samples: uniform weights, boost the weights of the Gal sources (since there are fewer of them) (ties with 2.c).(Inverse weighing should be best use)
2. Classification domains which offer a better view of how it is happening in the algorithm. Can also give possible hints about over- and under-training.

The scores from Algos will averaged over many seeds. Cross-validation and k-fold stratification might be needed, but isn’t necessary.

Features in classification (some are log): spectral index, flux\_density, signif\_curve, var, unc, HR ratios (4 with the last one being most important), galactic latitude.

Since this is a statistical study, some ideas from ROC curves, and general data analysis studies can also be applied (**precision and recall**).

Apart from the algorithm itself, the **data** is also important. This analysis of the data can be done in the following points:

1. Plot different features to see the distinction between different labels.
2. Use the correlation matrix to predict and talk about best features. This correlation can be used as a motivation for only use the features we will be using (Spec index et. Al.) as opposed to all the columns available in the catalog.
3. A small statistical analysis of the data would be cool. This will allow us to see what kinds of sources we are dealing with and the spread on this data can help us with bias and variance-based studies of pulsars and AGNs.
4. A final and important is the comparison of the two catalogs we shall be using. The newer catalog released this year appears to have updated on the previous classifications. A slight difference might therefore appear if we train with the 3FGL and use it on the 4FGL. Such a difference can be seen by for instance using the associated sources in 4FGL which are now unassociated.
5. Another difference which while important is interesting to keep in mind is that there might be associated sources in the 4FGL which are either not there in the 3FLG or which have a different association. This can also perhaps be checked by perusing the catalog introduction paper.
6. A final thing might be to also see the changes from the 1st catalog to the 4th. This is again not too important and serves only as a point of being meticulous.

## 2. Application to 3FGL and prediction

1. Divide the 3FGL into two parts: Associated and unassociated sources. The associated part with valid values of features is used in 1 and allows us to optimize our algos. The most optimized and cost-effective algos are then applied on the unassociated sources in 3FGL. This should be done for all types of classifications, including multi-label classification (see 2.4).
2. Calculate the most probable classes for the unidentified sources.
   * 1. Probabilities (Compare probabilities from multiple machine learning algos)
     2. Direct classes
3. Compare the predictions for different classifiers and search for outliers (among sources with known classes). This should be more common for pulsars since there are fewer of them.
4. Try to separate AGNs in classes, e.g., FSRQs and BL Lacs. Pulsars need to be seperated into young and ms pulsars.

## 3. Comparison with the FL8Y (Now 4FGL)

1. Find out if some of unassociated sources in 3FGL are associated in 4FGL. Those that are become final testing data. Using the predictions from 2, we check for accuracy of different algorithms. This is also a good test for our test of the complexity of the algorithms. Using this data which is not part of the same catalog as the training data, we find out the optimum algorithm using which we can safely predict with many resources.
2. Finally use the data from both the catalogs which has associations for training and make predictions for the unassociated sources in both the catalogs.

References:

Pablo Saz Parkinson:

<https://arxiv.org/abs/1602.00385>

References also included in the git profile, by Dima.

Proceeding with paper:

Intro: Also references to papers

Methods: Data and feature selection. Explain the data after cleaning it up (correlation matrix and plots of feature vs. feature). Then talk about how some features were selected and why they are so important theoretically. See 1.

Analysis: See 1

Predictions and results: See 2 and 3. How does this hold up with Pablo et. Al.?

Conclusions: What do the results tell us?

Problems: Too much data for 1 paper perhaps. How can this be overcome?

Probable Solutions: Don’t use multi-label or even more in-depth label based classification in the 1st paper. Use the 2nd paper to apply the methods and algorithms stated here to multi-label and yng, msp pulsars.

Catalogs used: 3 and 4fgl

Learn the difference between the two.