**Erdos group project modelling approach**

Frank Seidl, Luke Kiernan, Keavin Moore, Nicholas Barvinok, Noah Rahman

* Based on our recent exploratory data analysis, we find apparent correlations in the GEDI-L2A dataset between various features.
* In particular, we find that some/all of the 101 “relative height” metrics --- which describe the tree canopy height as well as its vertical structure --- can somewhat predict water persistence, tree cover, and urban proportion.
* We would like to predict “urban proportion”, i.e., the proportion of a given beam shot’s area with urban development, using the aforementioned features in our dataset.
* **If needed during model training, we can reduce the dimensionality of one/some of the columns, scale different columns, or create synthetic features by combining one or more features in the L2A dataset.**
* Since our dataset is mainly comprised of numbers (i.e., no classifications) and we are not trying to predict a time series, we will focus on regression models (**and could potentially try, e.g., a random forest approach, although it is a classifier model**).
* We can compare models using linear regression and k-nearest neighbors regression, and seek optimal hyperparameters for each corresponding model.
* We can cross-validate our model selections to aid in choosing the best predictive model of “urban proportion”.
* We can assess the predictive power of our model during cross-validation, including its accuracy and precision, to inform our model selection.
* Our entire modelling approach will likely be an iterative process and will involve cross-validation of various models using various optimal hyperparameters.
* **Although quite complicated, principle component analysis may prove to be invaluable in this exercise: since the 101 “relative height” features provide a map of tree canopy cover, we could find the most relevant and reduce the number to a more manageable amount of features to train.**
* **PCA could be especially useful since our exploratory data analysis doesn’t show any clear relations (currently --- but this was only using rh\_100) other than those stated above.**
* Although we do not have classifiers in our dataset, ensemble machine learning could still be useful; we could test, e.g., random forests, bagging & pasting, or boosting ***(I’m unsure about this specifically, but we should talk about it and keep it as an option****)*