**Objective:** The goal of this project is to classify individual trees to their taxonomic species from hyperspectral imagery. The project is inspired by the 2020 data science competition: IDTReeS, [Integrating Data science with Trees and Remote Sensing](https://idtrees.org/competition/).

**Introduction:** Understanding the number, size, and species of individual trees in forests is crucial to mitigating the effects of climate change, managing invasive species, and monitoring shifting land use on natural systems and human society. However, collecting data on individual trees in the field is expensive and time consuming, which limits the scales at which this crucial data is collected. Remotely sensed imagery from satellites, airplanes, and drones provide the potential to observe ecosystems at much larger scales than is possible using field data collection methods alone. IDTReeS investigates combining large scale survey efforts with remotely sensed imagery to scale and improve long-term forest conversation and climate change mitigation efforts.

**Related Work:**

Why hyperspectral data?

Hyperspectral images contain multiple (typically between 64 and 256) continuous narrow bands, providing significant levels of detail, which allow for the distinction of fine spectral variations among tree species. This has resulted in the extensive use of hyperspectral imagery for tree species classification.

(Ballanti et al., 2016) Use two non-parametric classifiers on hyperspectral data for tree species identification: Support-Vector-Machines (SVMs) and Random-Forest (RFs). ﻿To address the problem of high data dimensionality in hyperspectral images, a forward minimum noise fraction (MNF) transform was applied to the mosaicked image to reduce data redundancy and help identify bands containing the most variance. This process creates outputs of uncorrelated bands, which are ranked from highest eigenvalue (with most meaningful bands) to lowest eigenvalue (containing noise-filled bands).

(Wu & Zhang, 2020) Combined hyperspectral data with simultaneously acquired LiDAR data. Feature variables were extracted including, independent component analysis (ICA) transformation images, spectral indices, texture features, and canopy height model. Recursive feature elimination was adopted for spectral feature selection. Different feature combinations and classifiers (KNN, SVM) were tested on classifying tree species. SVM had the highest classification accuracy.

(Dalponte et al., 2014) Also used a non-linear SVM to classify hyperspectral imagery.

(Raczko & Zagajewski, 2017) Ran a comparative study on how three non-parametric classifiers performed on tree-species classification using hyperspectral data. Between a SVM, RF, and artificial neural network (ANN), the ANN performed best followed by the SVM and RF.

* ﻿Applied work by Pal and Mather (2006), which showed that a 40-band dataset was optimal for processing times of HSI, while preserving enough data to obtain satisfactory results. Principal component analysis (PCA) was used for band selection. Based on past works (Sommer et al., 2015; Thenkabail et al., 2012), one can assess the importance of each spectral band in each principal component by taking a look at the magnitude of factor loadings which correspond to the correlations between bands and principal components. This step assigns each spectral band a loading that indicates its importance. Higher loadings indicate more important bands. This procedure allowed us to select the 40 spectral bands with the highest PCA loadings from our 222-band dataset.

**Explaining the Data:**

*On Remote Sensing Data:*

The competition provides three primary data sources: remote sensing, field data, and individual tree crowns. The following project will only consider remote sensing geospatial datasets, specifically passive sensing systems. Passive systems measure the amount of reflectance at different wavelengths for ground-detected objects. The remote sensing datasets are generated by the NEON Airborne Observation Platform (AOP), and distributed in RGB and Hyperspectral formats at 100 cm2 and 1 m2 spatial resolutions, respectively. Data is stored as raster files, which means an image or array of pixels, whereby each pixel is stored as a vector of numbers. An RGB image is stored as a 3-band raster (3-element vector pixels). Each band represents the reflectance at different points in the electromagnetic spectrum corresponding to red, green, and blue wavelengths, respectively. Hyperspectral data consist of reflectance information from a much wider electromagnetic spectrum (380-2510 nanometers). Our data has a total of 369 bands.

*On Individual Tree Crown (ITC) Delineations:*

Individual tree crown (ITC) delineates are generated by IDTReeS research group. Each delineation is a 2-D rectangular bounding box defining the maximum tree crown extent in an image and is provided in vector format as ESRI shape files.

*Location of the Tree Crown Data*

The data consist of three NEON ecoclimatic sites in Eastern United States. In other words, each site is characterized by distinctive environmental, geographic, and vegetative properties. The sites are:

* **Ordway-Swisher Biological Station, Florida (OSBS):** The region contains mixed forests of hardwood and conifers, mostly dominated by pine trees.
* **Talladega National Forest, Alabama (TALL):** Forests made of mixed hardwood and conifers (mostly pine) in the Ozarks complex.
* **Mountain Lake Biological Station, Virginia (MLBS):** The region is mainly made of hardwood forests in the Appalachians and Cumberland Plateau.

**Method:**

*The Objective:* Determine the probability that each ITC belongs to a species class.

*Training Data:* Contain tree crown delineations with tree class (i.e., taxonomic species) labels. Total of 85 hyperspectral images are provided across all three sites. Each image represents the geographic extent of a single 20 x 20-meter plot, with array dimensions (20, 20, 369). Within these images are 1,165 delineated tree crowns.

*Preparing the Data:* To prepare the data, each of the 85 hyperspectral images were clipped into separate images corresponding to just the bounding boxes of each labelled ITC. This produced 1,165 images of varying heights and widths, with 369 spectral bands. The labels and extents of bounding boxes per image is given in the guideline file ‘data\_train\_mitree.csv’. There are a total of 33 tree species types delineated across all three NEON sites. Most of these species were underrepresented in the data, with some species corresponding to only 1 ITC. Given how limited the HSI data is, it is unlikely that a single classifier can demonstrate high prediction accuracies for all 33 species. As a result, the classification task is simplified to only the top 6, most frequently encountered, tree species and a 7th class labelled as ‘Other’, which encompasses the remaining 27 tree species in the NEON imagery.

* Object-Level Classification: is each clipped image, read-in as a tensor of shape , and is the associated tree label for that tensor. There would be a total of 1,165 tensors.
* Pixel-Level Classification: represents a single pixel, read-in as a 1-D array with shape and is the associated tree label for that pixel. There would be a total of 35,488 pixels.
  + Disadvantage: Would have to scale-results to object.

*Dimensionality Reduction:*

*Testing Data:* Tree clown delineations without tree class labels. Test our model by the classification of species on bounding boxes of unknown species identity.

**Results:**

*Performance Metrics:* F1 Scores (sklearn.metrics.f1\_score); Average Cross-Entropy Loss (sklearn.metrics.log-loss); Confusion Matrix (sklearn.metrics.confusion\_matrix).

**Conclusion:**

**References:**

*Collaborators*

As this project is an extension of an attempt at a Competition Submission, competition team members played a role in preliminary project scoping and data processing.

Team Members include Björn Lütjens {lutjens@mit.edu}; Becca Browder [{bbrowder@mit.edu}](mailto:{bbrowder@mit.edu})

*Neon Data References*

National Ecological Observatory Network. 2020. Data Product DP1.30010.001, High-resolution orthorectified camera imagery. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.

National Ecological Observatory Network. 2020. Data Product DP1.30003.001, Discrete return LiDAR point cloud. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.

National Ecological Observatory Network. 2020. Data Product DP1.10098.001, Woody plant vegetation structure. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.

National Ecological Observatory Network. 2020. Data Product DP3.30015.001, Ecosystem structure. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.