IDTReeS competition

Data and task descriptions

1.Introduction

This document defines the tasks and evaluation plans for the IDTReeS Competition.

Understanding ecological patterns and processes across geographical scales is crucial to understanding the effects of environmental change on natural systems and human society. However, financial and logistic limitations restrict the scales at which ecological data can be collected by field ecologists on the ground. Recent increases in the availability of remotely sensed imagery from satellites and aircraft provide the potential to observe ecological data at much larger scales than are possible through traditional data collection methods. This data science competition focuses on inferring ecological information from remote sensing data.

The competition uses data from National Ecological Observatory Network (NEON), which is collecting continental-scale ecological observations from 81 sites across the United States for the next 30 years. Data products range from ground based data on plants and animals, to large scale airborne remote sensing. Dealing with the volume, velocity, and variety of this data requires interdisciplinary approaches combining ecology, computer science, statistics, and data science. This data science competition uses multiple components of the NEON data stream to address the challenge of measuring and identifying individual trees and species across large areas. Because of the alignment with the National Ecological Observatory Network, the results of this competition will have immediate application to understanding forests on a continental scale and their role in global change.

The first competition (Marconi et al. 2019) used data from a single forest to compare methods for 1) identifying trees in remote sensing data, 2) aligning them with field data, and 3) determining the species of each tree.

This second competition will involve two core data science tasks. These tasks are designed so that they can be completed independently, and also so that they can be combined to form a data processing pipeline for real world applications. The two tasks are:

1. **Delineation** to locate individual trees crowns (the top, sun-exposed portion of the tree visible from above) in remote sensing data.
2. **Classification** to determine the taxonomic species identity (i.e., the type/category) of each individual tree in remotely sensed data.

This competition uses data from multiple sites to compare how well methods generalize to different forests, including forest types and locations on which the algorithms have not been trained.The competition will evaluate results in two ways:

1. Algorithm's performance on **trained sites**
2. Algorithm's transferability on **untrained sites**

2. Data

The data for this competition comes from three sources (Figure 1). Two of the sources are part of NEON standard collections (**remote sensing** and **field data**), and the third (**individual tree crowns**) is from the group of researchers hosting this data science evaluation. This section briefly describes data sources and the data products used to measure vegetation. Details of the data products and the physical properties they represent are presented in Table 1.

The data for this evaluation is **geospatial data**, which is data that is connected to a physical location on Earth's surface. Geospatial data must include information about the geodetic datum (a model of the earth) and in some cases the projection (a model that transforms the geodetic datum to a 2D coordinate grid). The datum used for these datasets is World Geodetic System 1984 (WGS84). The projection system used is the Universe Transverse Mercator (UTM). The UTM projection system includes different zones that are applied to different areas of longitude and latitude across the globe. Two UTM zones are used in this evaluation across the three geographic locations (see section 2.4). For this evaluation, the geographic information is specified using European Petroleum Survey Group (EPSG) codes. The EPSG codes will be stored and defined within the geospatial data files.

Geographic information is included in the file types for the remote sensing data (spatial raster format) and the Individual Tree Crown delineation (spatial vector format). The field data file type (non-spatial tabular format) will not include geographic information directly, but can be linked to the geographic vector format through a unique tree crown identification number.

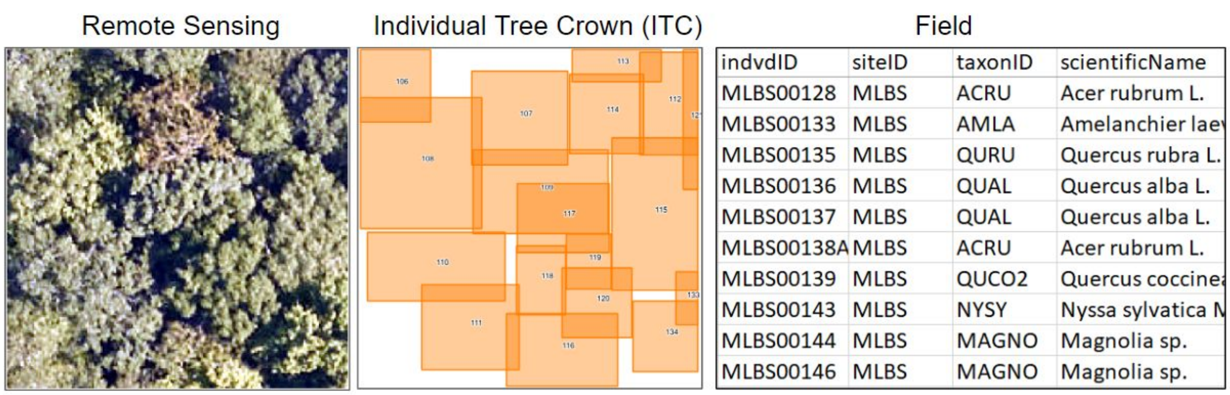

Figure 1. Diagram of the data used in this competition

2.1 Remote Sensing Data

The remote sensing data are generated by the NEON Airborne Observation Platform (AOP) and are provided as four different products, each one measuring different properties of the vegetation and the ground surface (Table 1). These data are distributed as raster and 'point cloud' vector formats. The full documentation for the data products used in this evaluation, and fundamental remote sensing concepts for how these data are collected is available from NEON (here). Remote sensing data comes from either passive or active systems. This competition includes data from both systems.

Passive systems record energy from the sun that is reflected from Earth's surface. The sensors can measure the amount of energy reflected at different wavelengths. When working with passive remote sensing data you should have information about the amount of reflected energy, and the wavelength region. For this competition, passive remote sensing data includes **RGB** and **hyperspectra**l datasets.

Active systems record energy that is reflected from Earth's surface from the system itself. The system emits pulses of energy towards Earth's surface and measures time and intensity of energy that returns to the sensor. When working with active remote sensing data, you should have information about the signal that is emitted. For this competition, active remote sensing data includes light detection and ranging (LIDAR).

2.1.1 Raster Data

Raster data are best understood as standard photographs or 'flat' images in which information at every pixel is stored as a vector of numbers encoding the values for each **band** in a row-column format. For example, a RGB image is provided as a 3-band raster. Each band represents the reflectance at different points in the electromagnetic spectrum corresponding to Red, Green, and Blue wavelengths. The raster values are encoded as a vector of 8-bit unsigned integers arrayed in a row-column-band format.

Three types of raster data are provided in this competition (Table 1, Figure 2). RGB photographs are high-spatial resolution data that show the reflected energy in the red, green, and blue portions of the electromagnetic spectrum. These data are similar to common digital cameras. Hyperspectral data are reflected energy from a wide range of the electromagnetic spectrum (380-2510 nanometers). Each of the 426 bands contain reflected energy for a small part of the spectrum. The canopy height model (CHM) isi a product of the point cloud Lidar Data (section 2.1.2). The values represent the height of the top of the vegetation canopy.

Data tip: the hyperspectral data has 426 wavelength bands (versus 3 for the RGB data). To create color-composites of these data you can select 3 bands for visualization. For a true-color composite (like the RGB data), select the following bands for the 3 color channels: Red=58, Green=34, Blue=19.

To enable the geolocation of these images, the data are provided in GeoTiFF format that store geo-reference information in the image header in addition to relevant pixel values. GeoTiFF data are best read/written through GIS software (ArcGIS, QGIS), or dedicated libraries (e.g. Rasterio, Raster for the R programming environment). For this competition, we provide raster data as GeoTiFF files (.tif) for three different data products, each one characterized by a different spatial resolution and number of bands (Table 1).

2.1.2 Point Cloud data

Point cloud data is geospatial data that consists of large number of points with known spatial coordinates in 3 dimensions; two dimensions representing the location horizontally in an east/west and north/south direction, and the third representing the location vertically.

Light Detection and Ranging (LiDAR) data is commonly provided as point cloud data. Airborne LiDAR data provides a measure of the height of a scanned surface. This measurement is made by a LiDAR system shooting laser pulses at a surface and measuring the time at which the light beams are returned to the airborne sensor.

For this competition, we provide LiDAR point cloud data in the LAS format (Table 1). This format can be manipulated using dedicated libraries (e.g. pdal for python, lidR for R).

Table 1. Remote sensing datasets provided to participants

Data product	Description	Spatial resolution	Data format	NEON data product ID
RGB photographs (RGB)	Raster data of the reflected energy from the surface as 3 bands representing the red, green, and blue portions of the spectrum.	100 cm^2	GeoTiff (.tiff)	NEON.DOM.SITE.DP1.30010.001
LiDAR point cloud (LAS)	Point cloud data of values in the X,Y, Z direction of the height of surface features and the ground.	6 points per m^2	.las	NEON.DOM.SITE.DP1.30003.001
LiDAR canopy height model (CHM)	Raster data containing the height of the top of the vegetation canopy. The CHM is a product of the LiDAR point cloud data.	1 m^2	GeoTiff (.tiff)	NEON.DOM.SITE.DP3.30015.001
Hyperspectral surface reflectance (HSI)	Raster data of reflected energy from the surface as 426 5-nm wide wavelength bands from 380-2510 nm.	1 m^2	GeoTiff (.tiff)	NEON.DOM.SITE.DP3.30006.003

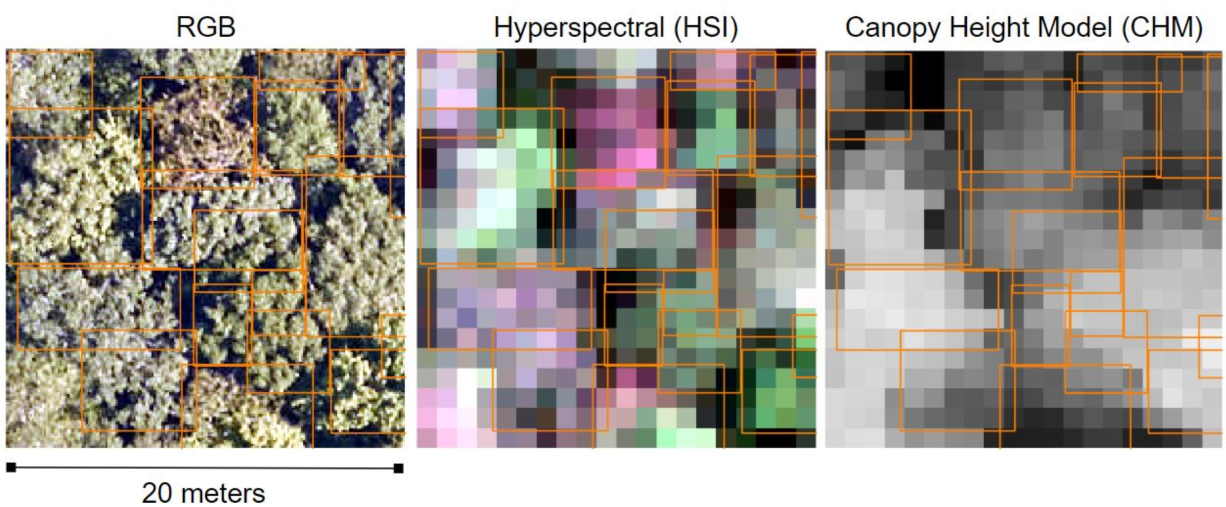

Figure 2. Example of three types of raster remote sensing data for a 20 x 20 meter plot. Orange boxes show Individual Tree Crown (ITC) delineations. The hyperspectral bands shown are R=58, G=34, B=19)

2.2 Individual Tree Crown (ITC) Delineation Data

Individual tree crown (ITC) delineations data are generated by members of the ID TReeS research group. Each delineation is a 2-dimensional rectangular bounding box that geographically defines an ITC (Figure 2). The delineation represents the crown maximum crown boundary or extent in the North/South and East/West directions. ITC data are provided in vector format as Esri Shapefiles. More information about the ITC data and how it is related to the field data is provided in the Supplementary Information.

Data tip: Each Esri "shapefile" is composed of four files of the same name, but with different file extensions (.shp, .shx, .dbf, .prj). Shapefiles are a common vector data format and can be read with widely-available packages (including most GIS software and GDAL). Other vector formats are Well-know text (WKT) , Keyhole Markup Language, and Geographic JavaScript Object Notation (GeoJSON).

2.3 Field data

Field data are collected by the NEON Terrestrial Observation System (TOS) personnel. The data contain information on individual tree identifiers, location of trees relative to sampling locations (i.e. distance and azimuth from a central location), species and genus labels, and measures of salient structural attributes. These attribute data are provided as comma separated value (.csv) files. Each record/row in the file is associated with a single ITC bounding box. More information about the field data and how it is related to the ITC data is provided in the Supplementary Information.

The field attribute that is directly used in this competition is the **taxonomic species information**. The species is described by its scientific name. To simplify the taxonomic species information, each scientific name has a unique taxonomic identification code. The code is usually a combination of the genus and species, the first and second parts of scientific names, respectively.

2.4 Location

Data provided belong to three NEON sites in the eastern United States (Figure 3). The three sites are part of three separate NEON ecoclimatic domains. The ecoclimatic domains were defined by NEON and represent distinct environmental, geographic, and vegetative characteristics. Three NEON sites will be included in the analysis:

Ordway-Swisher Biological Station, Florida (OSBS): mixed forest of hardwood and conifers, mostly dominated by pines in the Southeast NEON domain. The forests dominated by conifers tend to have open space between ITCs (open canopy forests). Forests dominated by hardwoods have ITCs that are largely continuous with each other (closed canopy forest). OSBS is in the Southeastern NEON domain (domain 3) and in UTM zone 17 North (EPSG code is 32616).

Talladega National Forest, Alabama (TALL): forest of mixed hardwood and conifers (mostly pine) in the Ozarks complex NEON domain. Forests dominated by hardwoods have ITCs that are adjacent to each other (closed canopy forest). Forests dominated by conifers can be open or closed canopy forests. TALL has a tree species composition that is largely a mixture of species found in OSBS and MLBS (see below).

TALL is in the Ozarks Complex NEON domain (domain 8) and UTM zone 16 North (EPSG code is 32616).

Mountain Lake Biological Station, Virginia (MLBS): mainly hardwood forest in the Appalachians and Cumberland Plateau NEON domain. ITCs are largely continuous with one another (closed canopy forest). MLBS is in the Appalacians and Cumberland Plateau NEON domain (domain 7) and UTM zone 17 North (EPSG code is 32617).

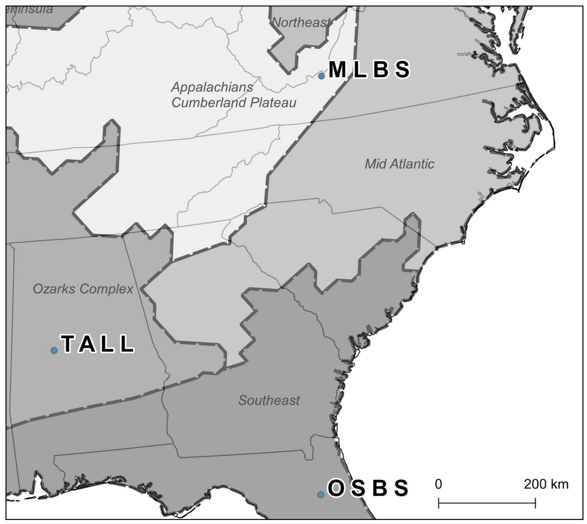

Figure 3. Map of three focal NEON sites used in this competition.

2.5 Data split

The data will be split into datasets for training to develop models and testing to evaluate model performance.

Training data is used by the participants to develop methods and algorithms. The training data provided contains all the information needed to train and self-evaluate model performance. For example, training data for the delineation task will contain tree crown delineations and training data for the classification task will contain tree crown delineations with tree class (species) labels.

Testing data is used by the participants as a test of their algorithms. Participants apply the model that has been developed on the training data to the testing data. The testing data does not have crown delineations (for delineation task) or class labels (for classification task) and therefore algorithm performance cannot be self-evaluated by the participants.

Evaluation data is submitted by the participants and is used to evaluate model performance by the organizers. These data are generated by applying participants' models to the testing data.

Submissions will be evaluated in two ways:

1. Algorithm's performance on trained sites (OSBS and MLBS)
2. Algorithm's transferability on untrained sites (performance on the TALL site)

2.6 Data organization

The data provided is split between data for training and data for testing. Within those folders are three folders corresponding to the three data types (*Remote Sensing, ITC, Field*, Figure 4). Within the *Remote Sensing* data folder are separate folders corresponding to the four remote sensing datasets as presented in Table 1. Each remote sensing dataset folder has individual raster (.tif) or point cloud (.las) files. A single file covers the geographic extent of a single 20 x 20 meter plot. The files are named with a unique identifier for the plots in the training data and the plots in the testing data. For example, "MLBS_29.tif" in the RGB folder in the training data, is the RGB high-resolution photograph remote sensing data plot number 29 in the MLBS site for the training data. The plot numbers used in this competition do not correspond to the plot numbers used by NEON.

The *ITC* folder contains the geospatial vector data format (Esri Shapefiles). There are separate files for the 3 NEON sites. The files are named with the **siteID**, which is the 4-letter code for the NEON sites. For example, "train_MLBS.shp" is the ITC training data for the MLBS site. Within each file are multiple individual bounding boxes where each box represents a single ITC. The attribute data for each bounding box is a unique crown ID number and an **indvdID** attribute that links the ITC geospatial data to the Field tabular data. The ITC data are linked to the Remote Sensing data geographically.

The *Field* folder contains the attribute data for all ITCs that were collected in the field. There are separate files for the training data and testing data, but not separate files for the 3 NEON sites. For example, "train_data.csv" contains the attribute data for all ITCs in the MLBS and OSBS sites. Attributes in the tabular field data link each row to the ITC bounding box (**indvdID**) and the NEON site (**siteID**).

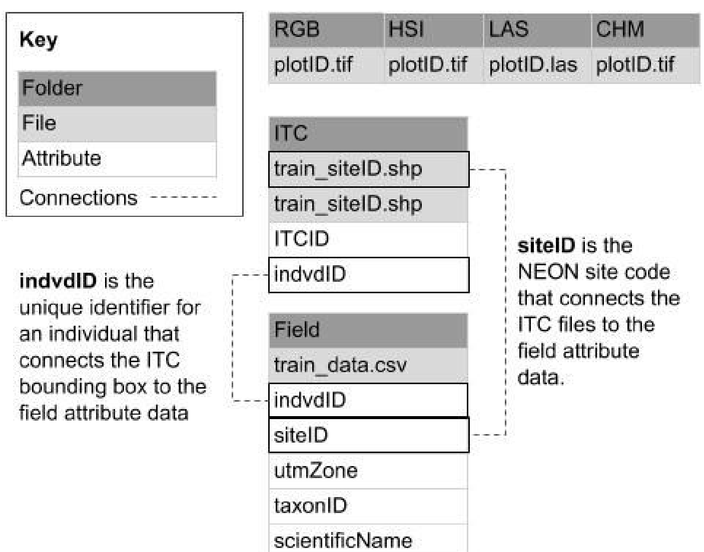

Figure 4. Folder and file structure. The darkest boxes are folders, light gray boxes are files, and white boxes are attributes within the files. Attributes or file names that link the datasets together are noted. This diagram shows the training data only. Testing data is provided as a separate folder with the same structure.

3. Delineation task

The delineation task of the IDTReeS Competition is to define the boundaries of individual tree crown objects (ITCs). Delineation is a common data science problem that involves locating objects or boundaries in images. In addition to being useful on its own, good delineation is often important for classification of objects within images. We are using the term delineation to highlight this task includes the detection of ITCs, and defining the boundary of ITCs. Other terms used in various ecology, forestry, remote sensing, and image processing fields include, segmentation, identification, and detection.

Identifying the position and size of individual trees from remote sensing is useful for understanding forest structure and an important first step in species classification. It is also a complex version of the common image delineation task because trees often overlap each other and look similar, and because the available data is heterogeneous, involving many bands, multiple resolutions, and point cloud height data.

All remote sensing data may be useful for the delineation task. LiDAR data provides information on the spatial variation in canopy height that may allow partitioning of crowns of neighboring trees with similar spectral signatures. Hyperspectral data allows development of spectral signatures to identify object categories (e.g. by assigning spectrally-similar categories to the same cluster). RGB photographs provide finer resolution information (0.25 x 0.25 m as opposed to 1.0 x 1.0 m for the hyperspectral data), which may be helpful to separate trees that are close to one another and to refine boundary placement.

3.1 Data split

The data will be split into datasets for training to develop models and testing to evaluate model performance (Table 2). In addition, since this competition will compare how well methods generalize to different forests, different data are provided the OSBS, MLBS, and TALL sites.

3.1.1 Training data

Remote sensing data, which includes the hyperspectral, LiDAR, and RGB photos, are provided for the training plots at the OSBS and MLBS sites. ITC data provided are spatial bounding boxes that define each ITC for all plots. The ITC data can be used in any way for developing delineation methods, whether that be directly for supervised methods, or indirectly by evaluating the output of unsupervised delineation method. Field data are provided for all ITCs. Since the TALL site is used as a test of how models apply to untrained sites, no TALL data is provided in the training data. Participants can use the train data for self-evaluation of their methods.

<u>Training Files:</u>

Remote Sensing data files (.tif, .las)
train_MLBS.shp, train_OSBS.shp (and associated files of .shx, .dbf, .prj)
train_data.csv, train_data_attributes.csv

Data tip: The training data for the delineation task are the same as for the classification task. The difference is that the field data are required for training classification methods because they contain species labels, while the field data is not necessary for delineation methods.

3.1.2 Testing data

Remote sensing data are provided for the testing plots at the OSBS, MLBS, and TALL sites. No ITC or field data are provided for these plots. Participants apply methods developed using the training data to the testing data.

Testing Files:

Remote Sensing data files (.tif, .las)

3.1.3 Submission data

Participants will submit the bounding boxes that delineate the spatial boundaries of ITCs in every 20 x 20 meter plot provided in the testing data. Participants should submit the vector data as an Esri “shapefile”. The 3 required file extensions that make up a complete shapefile are .shp, .shx, and .dbf.

Submission File:

delin_subm.shp, delin_subm.shx, delin_sub.dbf
OR
Delin_subm.csv (with the geometry described using the WKT format)

Table 2. Data provided and to be submitted for the delination task

		RS data	ITC data	Field data*
Train data provided	OSBS/MLBS TALL	Yes None	Yes None	Yes Yes
Test data provided	OSBS/MLBS TALL	Yes Yes	None None	None None
Data to submit	OSBS/MLBS TALL	None None	Yes Yes	None None

*Field data is not required to complete the task, but may be useful for developing methods

3.2 Performance Metrics

There are a few complications in detecting and delineating trees from remote sensing. First, hand-ITCs labelled on RGB images and their boundaries may not be perfectly overlapping with LiDAR or HSI data products; second, being produced remotely there is uncertainty in the quality of the ITC delineation itself. Third, delineations include a number of pixels that are not of the tree crown itself (for example soil, or even neighboring trees). The number of such pixels strongly depends on the orientation and shape of each crown. For this reason, we will measure performance in delineation by using the Rand index, a measure of the percentage of correct decisions made by the algorithm. Participants' performance on each plot will be calculated using the **Rand index**. The Rand index will be used to provide a measure of similarity between two delineations of a plot (i.e., the similarity between the ground-truth and the submission). To account for small uncertainties and inaccuracies in the size of the ITC delineations, when we compute the Rand index, we will use a set of "halos" around each truth location during scoring. The Rand index can be computed as follows:

$$R = \frac{a+b}{a+b+c+d} = \frac{TPP+TNP}{NP}$$

where *a* is the number of pairs of agreements between the results and the inner halo from the ground truth, *b* is the number of pairs of agreements in the area between the outer halo and edge halo in the ground truth and the results, *c* is the number of pairs of disagreements between the results and the portion between the outer and edge halo from the ground truth, and *d* are the number of pairs of disagreements between the results and the inner halo of the ground truth. The value of *a* can be interpreted as the number of true positive pairs and *b* as the true negative pairs, while *c* is the number of false positives, and *d* are false negative pairs. An illustration of the halos is shown in Figure 5.

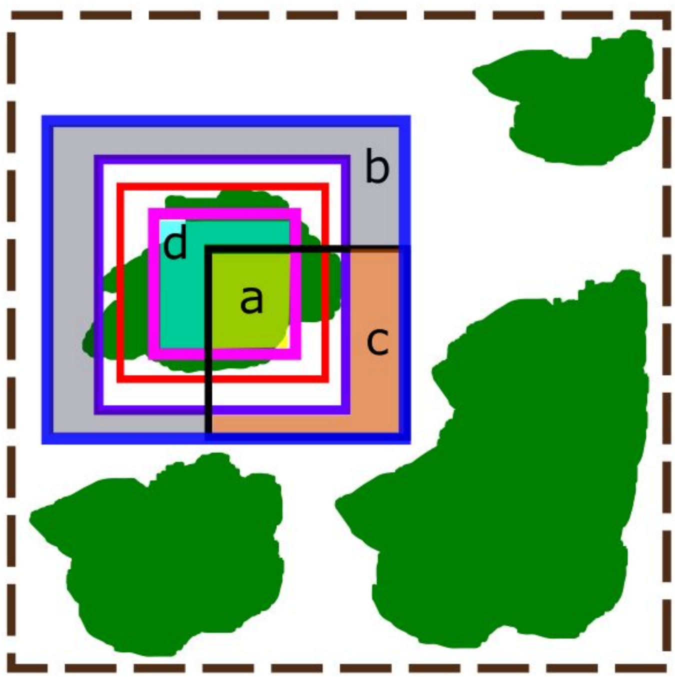

Figure 5. An illustration of how crown delinations are scored. The **green polygons** represent individual tree crowns, the **red box** is the ground truth delineation, and the **brown dashed box** is the full 20 x 20 meter plot. Within the ground truth delineation is a **pink box** and this is the inner halo. Outside of the ground truth delineation is a **purple box** and this is the outer halo. The most outer **blue box** is the edge halo. The **black box** is a sample delineation. The true positives are shown in the yellow highlighted region (marked with an "a"). This is the intersection of the detection and the inner halo. The true negatives are the blue highlighted region (marked with a "b"). This is the difference between the edge halo and the outer halo and the detection. The false positives are in the orange highlighted region (marked with a "c"). This region shows the portion of the detection that is encompassed in the negative region between the edge

and outer halo. The false negatives are in the teal highlighted region (marked with a “d”). This region shows the difference between the inner halo and what was detected in the outer halo.

To keep this score balanced, the inner and outer halos is set to be 10 cm within (1 pixel in the RGB imagery data) and outside of the ground truth, and the edge halo is computed so that the area between the edge and outer halo is the same size as that of the inner halo. Predicted bounding boxes that have no overlap with a ground truth bounding box will be given a score of 0.

Code for assessing this scoring is provided as a Python package with the following function used to assess the Rand score:

```
Score = RandScore(CrownPredictions, GT)
```

where GT is a numpy array with inner and outer halo locations for each crown and CrownPredictions is a numpy array where each row corresponds to a single crown, the first two columns correspond to the upper left pixel location of the crown bounding box and the second two columns correspond to the lower right pixel location of the crown bounding box (row, column).

Each predicted bounding box is assigned to a ground truth bounding box based on the greatest overlap. In cases of ambiguity of correspondence between predicted crowns and crown locations in the ground truth, we will use the one-to-one mapping that provides the best score for the submission. This will be accomplished using the **Hungarian assignment algorithm**.

We will also calculate mean intersection over union scores (IoU), defined as the area of overlap between the ground-truth bounding box and predicted bounding box divided by the area of union (Figure 6). IoU is the most common metric of bounding box accuracy in object delineation, and ranges from 0 (no overlap) to 1 (perfect overlap). Since IoU does not allow for easy incorporation of an uncertainty “halo”, it will be used as a secondary score. The overall score for the delineation task is the average IoU for each plot (the IoU’s of all ITCs in a plot will be averaged).

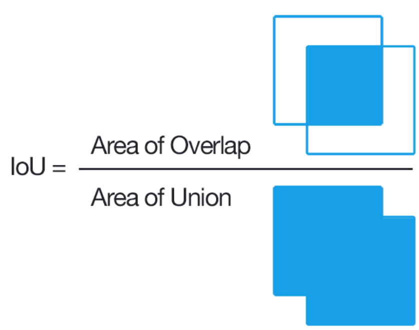

Figure 6. Diagram of Intersection over Union (IoU) metric used in the evaluation of delineation task submissions.

The evaluation code has been published on [Zenodo]() and [GitHub](), and can be used by participants for self-evaluation using the training data.

4. Classification task

Classification is a common data science problem that involves determining object membership in a set of categories. All remote sensing data may be useful for the classification task. LiDAR data provides information on the spatial variation in canopy height that may allow distinguishing species based on height, shape, or crown structure variability. Hyperspectral data allows development of spectral signatures to identify object categories and is likely the most useful remote sensing dataset for classification. RGB photographs provide visible-color spectral reflectance but at a fine spatial resolution (0.25 x 0.25 m as opposed to 1.0 x 1.0 m for the hyperspectral data), which may provide additional useful information for spectral separability of species classes.

A large number of ecological, environmental, and conservation oriented questions depend on species identification. This includes efforts to conserve individual species, understand and maintain biodiversity, and incorporate the biosphere into global circulation models. Being able to describe the density and distribution of different species using remote sensing would allow these efforts to occur more rapidly and at larger scales than field sampling.

The goal of this task is to classify trees in remote sensing data to their taxonomic species. In addition to its utility for the domain, this task represents a challenging version of general classification problems because it involves classifying different species with very similar spectral signatures and categorizing data where some categories (species) have only small samples in the training set (i.e. rare species).

To make this task independent of delineation task, ITC delineations are provided. Participants will determine the probability that each ITC belongs to a species class. Since classification is at the level of the ITC, any pixel-level classification models must be upscaled to the crown.

One important criterion for evaluating a classification model is its ability to deal with species from outside the training set. Thus, we also evaluate participants' classification models on their ability to correctly identify crowns to an "Other" class when they do not belong to any of the previously seen classes.

4.1 Data split

The data will be split into datasets for training to develop models and testing to evaluate model performance (Table 3). In addition, since this competition will compare how well methods generalize to different forests, different data are provided for the OSBS, MLBS, and TALL sites.

4.1.1 Training data

Remote sensing data, which includes the hyperspectral, LiDAR, and RGB photos, are provided for the training plots at the OSBS and MLBS sites. ITC data provided are spatial bounding boxes that define each ITC for all plots. Field data are provided for all ITCs. The field data provide the taxonomic species class for each ITC. The ITC and field data can be used in any way for developing delineation methods, whether that be directly for supervised methods, or indirectly by evaluating the output of unsupervised delineation method. Since the TALL site is used as a test of how models apply to untrained sites, no TALL data is provided in the training data. Participants can use the train data for self-evaluation of their methods.

Training Files:

Remote Sensing data files (.tif, .las)
train_MLBS.shp, train_OSBS.shp (and associated files of .shx, .dbf, .prj)
train_data.csv, train_data_attributes.csv

4.1.2 Testing data

Remote sensing data are provided for the testing plots at the OSBS, MLBS, and TALL sites. The ITC bounding boxes that define the pixels to be classified are provided, but no field data are provided. Participants apply methods developed using the training data to the testing data.

Testing Files:

Remote Sensing data files (.tif, .las)
test_MLBS.shp, test_OSBS.shp, test_TALL.shp (and associated files of .shx, .dbf, .prj)

4.1.3 Submission data

Participants will submit taxonomic species predictions for ITCs in the test data. The predictions should be a probability from 0 to 100% that the crown belongs to the associated genus or species.

Submissions will be a single .csv containing information on the crown ID and the taxonomic species classification probabilities for each crown. The files should contain one row for each ITC crown ID and species classification combination (i.e., the number of rows should be equal to the number of ITC's in the testing data x the number of unique species classifications present in the training data). The dictionary of taxonomic species classifications and their unique IDs will be provided.

The testing data contains classes that are not present in the training data. Therefore, participant submission data may include ITC labels that do not have a known taxonID class. When this occurs, participants should include the class prediction as "Other" so signify the class is unknown and not one of the classes included in the training data.

Submission File:

class_subm.csv

The submission file must contain the following columns/attributes:

- **indvdID**: the matching ID from the ITC data
- **taxonID**: the predicted taxonomic code
- **probability**: the probability that the crown belongs to the associated genus or species. The probabilities for a given ITC ID (including the "Other" category) will be normalized to sum to 1 if the submitted values do not already.

class_subm.csv		
indvdID	**taxonID**	**probability**
MLBS0001	QUAL	0.85
MLBS0001	ACRU	0.15
MLBS0001	NYSY	0.0
MLBS0002	Other	0.75
MLBS0002	PIPA2	0.15
...	...	...

Figure 7. Example of the attributes and values to include in the submission field for the classification task.

Table 3. Data provided and to submit for classification task

	Site	RS data	ITC data	Field data
Train data provided	OSBS/MLBS TALL	Yes None	Yes Yes	Yes Yes
Test data provided	OSBS/MLBS TALL	Yes Yes	Yes Yes	None None
Data to submit	OSBS/MLBS TALL	None None	None None	Yes, class predictions Yes, class predictions

4.2 Performance Metrics

The primary metric for assessing species classification will be the macro F1 score:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

This score is the product of the precision and recall scores for each class in the dataset. The score weights each class equally regardless of the number of individuals in each class. This makes this score good for unbalanced datasets because it weights the performance on rare classes (species) equally.

For the site where no training data is provided, the primary comparisons will be made only on species that are in the training data for the other two sites. To assess models' ability to identify unseen classes we will also compute a second macro F1 score with unseen classes designated as “other”.

Both of these F_1 scores will be computed using the following sklearn function:

sklearn.metrics.f1_score(y_true, y_pred, average=’macro’)

For submissions that include probabilistic classification (i.e. a probability that each individual belongs to each class) we will assess model performance incorporating this uncertainty using cross-entropy. Average cross-entropy, is defined as:

$$\overline{crossentropy} = \frac{-\sum_{n,k} ln(p_n k) * \delta(g_n, k)}{N}$$

The *δ(x, y)* is a function that takes a value of 1 when *x* = *y*. This metric rewards participants for submitting well-calibrated probabilities that reflect their uncertainty about which crowns belong to which class. If the probability values do not sum to 1, they will be renormalized. Cross entropy scores will be computed using sklearn's function as follows:

```
sklearn.metrics.log_loss(y_true, y_pred, eps=1e-15, normalize=True)
```

Finally, full confusion matrices will be calculated for each submission to allow for further analysis, discussion and comparison, particularly to identify classes that are commonly confused (e.g., species within a genus) across methods. Confusion matrices will be computed using sklearn's function as follows:

```
sklearn.metrics.confusion_matrix(y_true, y_pred)
```

The evaluation code has been published on Zenodo and GitHub, and can be used by participants for self-evaluation using the training data.

5. Submission descriptions

Submissions will be accepted through an online from. We allow each team to make up to 4 submissions per task. Submissions made prior to the final submission will be evaluated and scores will be returned within 2 business days.Submissions will be evaluated and scores will be returned to the team within 2 business days. The final submissions are due by 11:59 on July 1. Each submission should include basic information required for evaluation. The final submission must include additional details about the method and data sources used.

Information required for submission. Items marked with an * are only required for the final submission.

1. *Team or individual name*
2. *Team Affiliation*
3. *Team country*
4. *Contact email address*
5. *Submission task*
6. **Method* (non-technical overview of your approach, <100 words)
7. **Algorithmic details* (brief description of the key parameters of the algorithm) < 500 words)
8. **Data sources (brief description of the data used including both the specific subsets of the provided data and any external data sets used, <500 words)*

6. Schedule

The key dates for the competition will follow the scheme summarized below.

Date	Task
6 March	Detailed instructions and training data provided to participants
1 June	Submission process open for live evaluations
1 July	Deadline for participants to submit data for evaluation
8 July	Announcement of results/winner and solicitation of manuscript collection
15 July	deadline for participants to sign up for manuscript submission

7. Rules

- Participants may work independently or in teams of any size
- Teams should be determined at the beginning of the competition and are not allowed to merge
- Participants are encouraged to post the code used for competition analyses under an open source license to allow for others to replicate and build on work from the competition
- Submissions will be open from April 1st through May 1st and each team can make up to two submissions/week during this period.
- Decisions of the ID TREES organizers are final

8. 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.