Zenodo Methods
First Draft.

This version of this document is intended to serve as a document that links our submission to Advances in Archaeological Practice (proposed title:*When Computers Dream of Charcoal: Using Deep Learning, Open Tools and Open Data to Identify Relict Charcoal Hearths in and around State Game Lands in Pennsylvania*) to our publicly published data.

This version contains links to the associated data (see also “Related Identifiers” in Zenodo). We plan to both post additional information here, but we are also preparing a publication for JOAD so much of it will be there. If you are looking for detailed methods, we will be publishing those in that article as well as here (be sure to check for another version of this document if you think the methods should already be available).

These are intended to be in chronological order.

Data for

1. State Game Lands with a 1 km buffer
    a. Shapefile : https://zenodo.org/record/4593518
    b. GeoJSON: https://zenodo.org/record/4593540
2. Grid tiles with SGL information included
    a. Shapefile: https://zenodo.org/record/4593491
    b. GeoJSON: https://zenodo.org/record/4593572
3. Description for downloading LiDAR tiles (in LAS format): https://zenodo.org/record/4596905
4. Program for arranging files downloaded with the above, known as ProjectLambda: https://zenodo.org/record/4572993
5. Program for processing original Lidar data (in LAS format) to yield Digital Elevation Model, Hillshade and Slope: https://zenodo.org/record/4573004
6. Files resulting from the above for State Game Lands (in GeoTIFF format)
    a. SGL 12-50: https://zenodo.org/record/4596903
    b. SGL 51-80: https://zenodo.org/record/4597896
    c. SGL 81-110: https://zenodo.org/record/4597943
    d. SGL 111-160: https://zenodo.org/record/4597941
    e. SGL 161-200: https://zenodo.org/record/4597994
    f. SGL 201-250: https://zenodo.org/record/4598045
    g. SGL 251-260: https://zenodo.org/record/4598089
    h. SGL 261-280: https://zenodo.org/record/4598096
    i. SGL 281-335: https://zenodo.org/record/4598100
7. Vector files resulting from manual identification of relict charcoal hearths (RCHs). These were used to train the Mask R-CNN model.
    a. Shapefile - https://zenodo.org/record/4593605
    b. GeoJSON- https://zenodo.org/record/4593622
8. Program for splitting slope analysis (in TIFF format) into smaller tiles (in JPEG format).
    a. https://github.com/jeffblackadar/charcoalhearths/blob/master/0_split_tifs_refactored.ipynb

9. Data files for Training:
    a. RCH Detection with Mask R-CNN Image Annotations. https://zenodo.org/record/4575582 This file contains a collection of xml files that contain coordinates of the locations of known RCHs on images (from above). These files are known as annotations and are used by Mask R-CNN to identify objects to detect during training of a model.
    b. RCH Detection with Mask R-CNN Training Images. https://zenodo.org/record/4579935 This file contains all of the images used for training the Mask R-CNN model. Each image contains at least one known RCH.
    c. Polygons for tiles of lidar data. https://zenodo.org/record/4580726
10. Program for Mask R-CNN training and prediction- known as "data_5000_3_rcnn_charcoal_hearths.ipynb" https://github.com/jeffblackadar/charcoalhearths/blob/master/data_5000_3_rcnn_charcoal_hearths.ipynb
11. Data files of the model and predictions
    a. Resultant trained Mask R-CNN model. https://zenodo.org/record/4579946
    b. Predictions from the model, in x, y coordinates (not geolocated) in XML format. https://zenodo.org/record/4581281 The format of these files is similar to the training annotations.
    c. RCH Detection with Mask R-CNN Images. https://zenodo.org/record/4583945 This file contains all of the images representing tiles of lidar images of state game lands. These images are used for predictions to locate RCHs. (A subset was used for training. See 11b RCH Detection with Mask R-CNN Training Images above.)
12. Program that produces confidence scores for the predictions above. Known as "data_5000_4_rcnn_charcoal_hearths_count_results.ipynb" https://github.com/jeffblackadar/charcoalhearths/blob/master/data_5000_4_rcnn_charcoal_hearths_count_results.ipynb
13. Program to remove duplicates (because some tiles are included near multiple SGLs), converts squares to their centroid and save in as geolocated vector files. Known as "2_read_predictions_from_xml_put_into_shp.ipynb" https://github.com/jeffblackadar/charcoalhearths/blob/master/2_read_predictions_from_xml_put_into_shp.ipynb
14. Vector file of prediction results.
    a. Shapefile- https://zenodo.org/record/4593734
    b. GeoJSON- https://zenodo.org/record/4593747
15. Prediction results with additional variables (bins for assessment, ID of training data, cluster analysis and visual confirmation)
    a. GeoJSON (no shapefile)- https://zenodo.org/record/4593767
    b. Variables:
        i. **id**= unique identifier starting with 3-digit SGL number, PAN or PAS (projections) and, within those a unique four-digit identifier
        ii. **score**= confidence score
        iii. **SGL**= State Game Land number

iv. **SGLImage**= name of TIFF file of merged lidar tiles
v. **Confirm**= Whether the predicted hearth was determined, through visual inspection, to be a likely true positive (Y) or a false positive (N)
vi. **Bin#**- in assessing these predictions we “binned” the results based upon the confidence score.
vii. **Bin_select**= 1 if this record (predicted RCH) was selected for assessment within that bin
viii. **TrainID**= Original ID of the training data (only training data that matched with a prediction are included).
ix. **Clusters5_300**= resultant clusters from DBSCAN where minimum cluster size= 5 and maximum distance= 300 meters
x. **Clusters10_500**= resultant clusters from DBSCAN where minimum cluster size= 10 and maximum distance= 500 meters
xi. **Clusters20_1000**= resultant clusters from DBSCAN where minimum cluster size= 20 and maximum distance= 1000 meters
xii. **CLUSTERCT** = How many of the above clusters included the predicted RCH (0-3). Derived from the previous three variables.
xiii. **3Cluster**= whether or not this predicted RCH was included in all three clusters.

16. False Negatives for the tiles around SGL 43 after close visual inspection (at 1:1000 scale).
    a. GeoJSON= https://zenodo.org/record/4758647