EY Biodiversity Challenge

Brian Killough, PhD

Director, EY Data Challenge
EY Executive Consultant
Former NASA Engineer and Scientist

UTD Class Briefing
February 10, 2025

Visit our EY Open Science AI & Data Challenge: challenge.ey.com

The importance of Frog Biodiversity

Reference: *Australian Museum - Green Tree Frog (Litoria Caerulea)*

The presence of frogs is an important sign of a healthy ecosystem. Scientists believe frogs are an environmental bellwether with declines in their population viewed as early warning signs of environmental damage.

Thus, the study of frog habitats is increasingly important to understand the extent and severity of global environmental change.

Frog populations have declined significantly since the 1950s and more than one third of species are considered to be threatened with extinction and over 120 are believed to have become extinct since the 1980s.

Scientists and researchers use the presence of frogs as a measure of biodiversity health. So, this challenge is all about "detecting the presence of frogs" in diverse ecosystems.

“Frog Challenge” Objectives

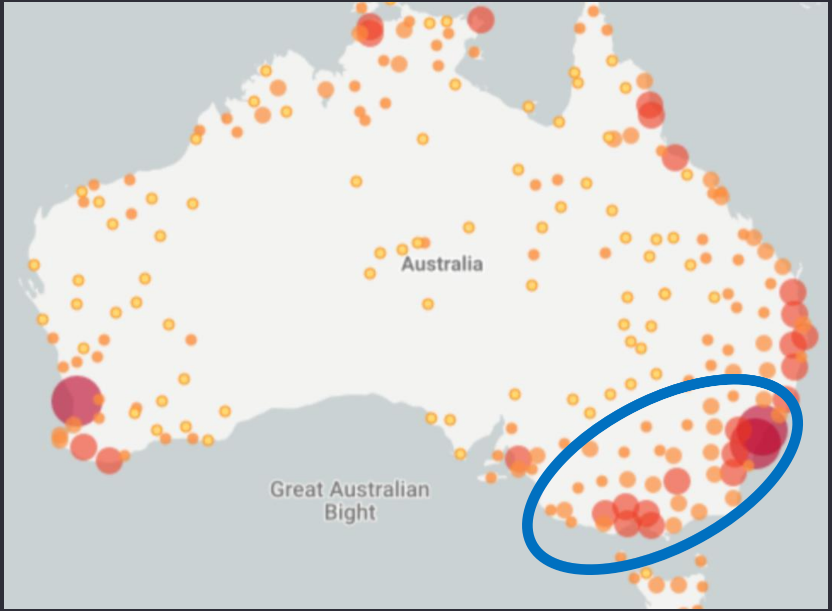
The goal of the challenge is to build a machine learning classification model to predict the presence of frog species based on environmental climate variables.

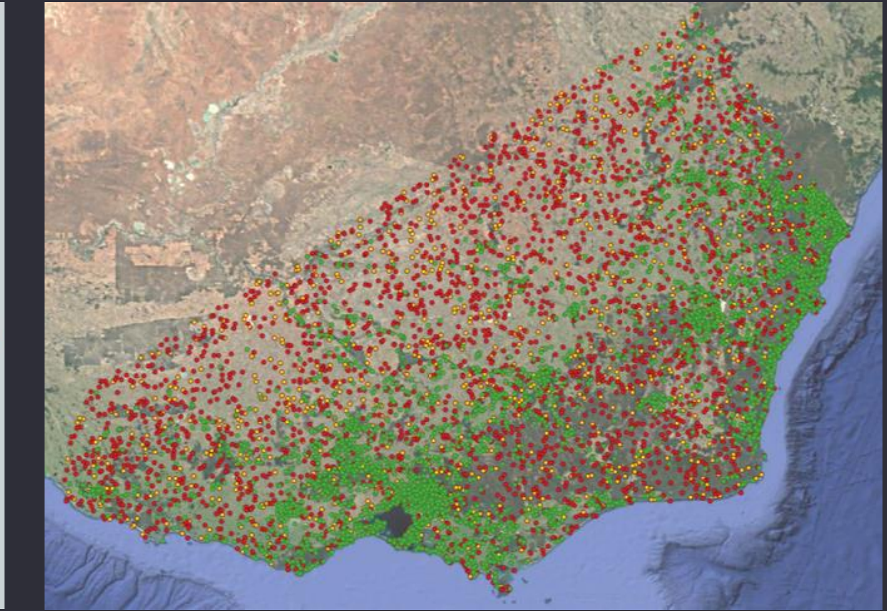
Participants will be given locations (latitude and longitude) of frog presence and frog non-presence across a portion of southeastern Australia over a period of 2 years from Nov-2017 to Nov-2019. This will be the “target” variable for your model.

The TerraClimate dataset provides global monthly climate and water balance variables from 1958 to the present. This will be the “predictor” variable for your model.

In the end, your model will be used to predict the presence or non-presence of frogs in specific locations.

Region of Interest and Target Dataset

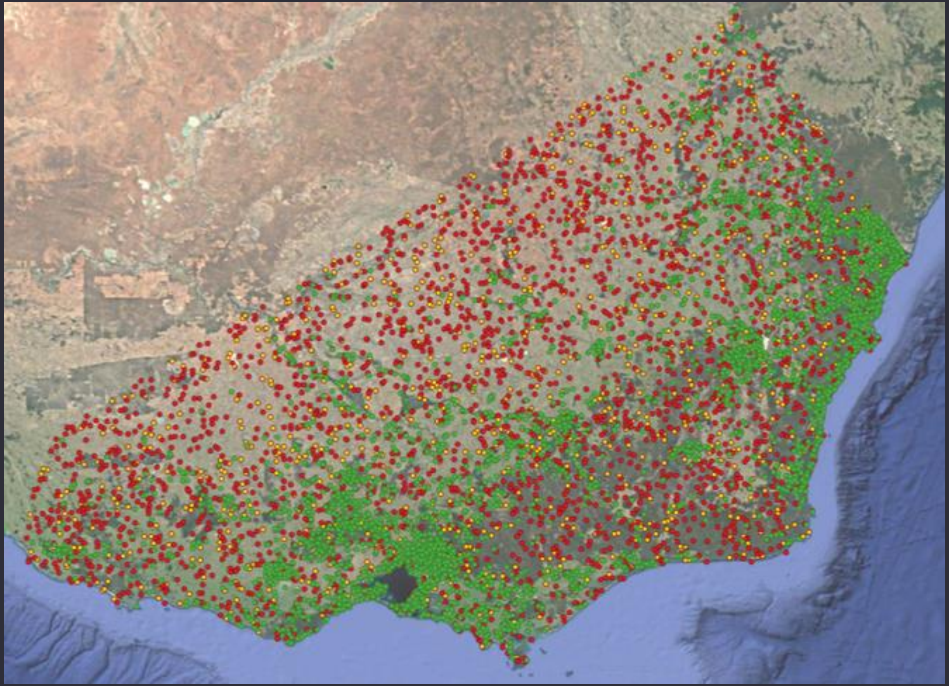




The data originates from the Global Biodiversity Information Facility (GBIF). Citizen science projects have collected data on frog presence using sound recordings on cell phones. The FrogID dataset includes >800,000 records in Australia (see left image).

Our dataset focuses on southeastern Australia from Nov-2017 to Nov-2019 (see right image). There are 3792 frog presence locations (GREEN), 2520 frog non-presence locations (RED) and 2000 validation locations (YELLOW) to test model predictions.

More about the frog non-presence dataset ...



The frog non-presence dataset (shown in RED) was developed using random equally spaced locations across the region and is NOT a ”measured” dataset.

The data was further sorted to eliminate duplications with coincident frog presence locations and other ”low probability” locations (e.g., in water, on buildings).

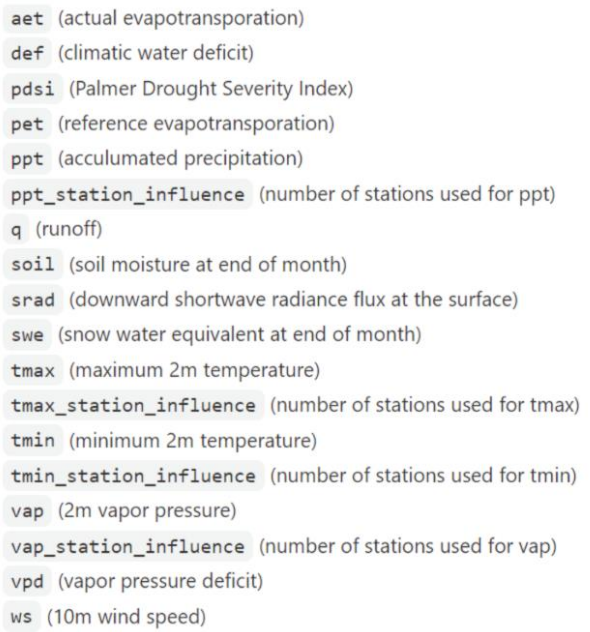
Students should consider the validity of the frog non-presence dataset. It is possible to alter the non-presence dataset (e.g., number of points, locations) to improve results.

NOTE ... In a “perfect world” we would measure all locations and measure the presence and non-presence of all frog species. But, this is not practical. So, we are forced to identify where specific frogs are located and then estimate the non-presence locations in order to build a digital species distribution model.

TerraClimate - The Predictor Dataset

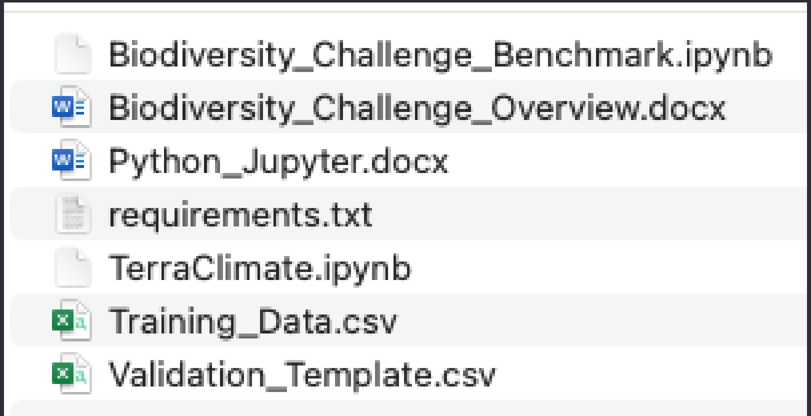
TerraClimate offers monthly climate data at a spatial resolution of 4 km, dating back to 1958. TerraClimate data includes 14 variables which are essential for assessing environmental factors affecting frog populations. Because frogs are highly sensitive to climatic conditions and moisture availability, TerraClimate data enables researchers to track habitat suitability with greater precision.

The dataset includes 6 primary measured parameters (YELLOW circles) and 8 derived parameters. The provided sample notebook demonstrates two of those variables (srad, vap), but students will want to test many others.



`aet` (actual evapotransporation)
`def` (climatic water deficit)
`pdsi` (Palmer Drought Severity Index)
`pet` (reference evapotransporation)
● `ppt` (acculumated precipitation)
`ppt_station_influence` (number of stations used for ppt)
`q` (runoff)
`soil` (soil moisture at end of month)
● `srad` (downward shortwave radiance flux at the surface)
`swe` (snow water equivalent at end of month)
● `tmax` (maximum 2m temperature)
`tmax_station_influence` (number of stations used for tmax)
● `tmin` (minimum 2m temperature)
`tmin_station_influence` (number of stations used for tmin)
● `vap` (2m vapor pressure)
`vap_station_influence` (number of stations used for vap)
`vpd` (vapor pressure deficit)
● `ws` (10m wind speed)

What is in your participant/student data package?



Biodiversity_Challenge_Benchmark.ipynb

Biodiversity_Challenge_Overview.docx

Python_Jupyter.docx

requirements.txt

TerraClimate.ipynb

Training_Data.csv

Validation_Template.csv

Overview Document … Read this first!

Python / Jupyter Document … Read this next if you are not familiar with running Python using Jupyter notebooks.

Training Data / Validation Template … Review these inputs for your models.

Requirements … Review the list of needed Python libraries. You will likely need this list to verify your Python environment is ready to run the notebooks.

Python Notebooks (*.ipynb files) … These are the key files that you will use to get your model started.

- TerraClimate … Run this notebook first to create a GeoTIFF output file. This file will be used by the Benchmark notebook.
- Benchmark … This is the core notebook that gives you a baseline/simple model and a starting score. Once you have this running, the goal is to improve your model results.

Rules of the Challenge and Suggestions

This data challenge is meant to be a "learning experience", so we have purposely kept it simple and minimized data complexity. Here are some "rules" to follow …

- You may only use the TerraClimate dataset as a source for your "predictor" variables. It is suggested that participants utilize all of the TerraClimate variables and consider alterations to the time window or statistical variations of the variables. Also, variable scaling and normalization should be considered.
- Submissions (Predicted_Data.csv) will be sent to your instructor or assistant to evaluate your model scores against a "ground truth" dataset. Avoid numerous submissions per day. Process will be confirmed next week.
- Students may use any common machine learning technique. This might include SVM, CNN, or regression variations.
- Students should consider the validity of the frog non-presence dataset. This dataset is randomly generated across the region of interest and is not a "measured" dataset. So, it is possible to alter or recreate the non-presence dataset (e.g., number of points, locations) to improve results.

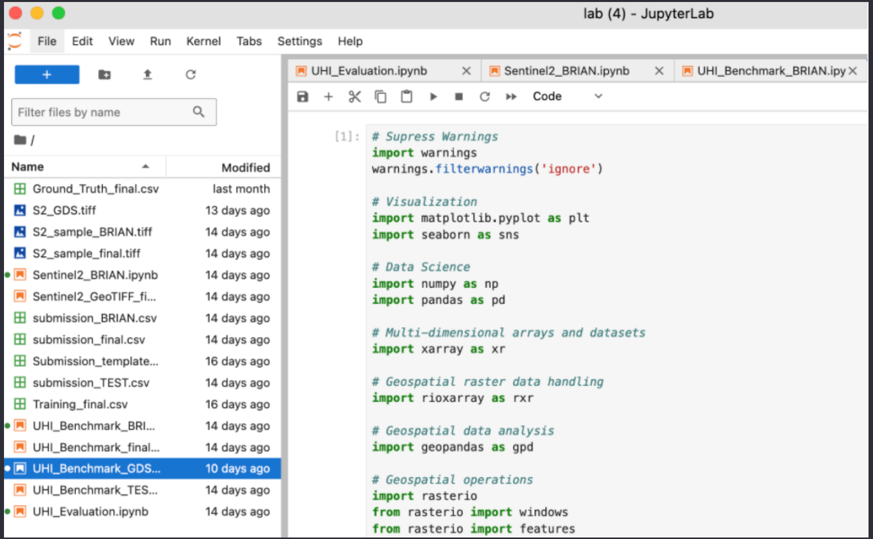
Setup - Python and Jupyter Lab

- You DO NOT need a cloud-based environment to complete the Data Challenge. These challenges have been designed to work on local computers with basic configurations (e.g., 4 cores, 32GB memory)
- In order to develop and run code on your local computer, you will need a Python virtual environment and Jupyter Lab for managing notebooks (code, text, output).
- You can use "Anaconda" to setup a Python environment with Jupyter Lab and lots of common Python libraries. It is best for "beginners" but loads lots of extra content.
- For a "manual" Python virtual environment setup, check out this video for PC or Mac: https://www.youtube.com/watch?v=9tPS-7TWjq0
- Some popular commands are: "python3 -V" to view the installed Python version, "pip3 list" to view installed Python libraries, and "pip3 install ###" to install new Python libraries named "###".
- For installing JupyterLab on your Mac computer, check out this video: https://www.youtube.com/watch?v=578B63wZ7rI

Running Jupyter Notebooks

Jupyter Notebooks allow you to create and share documents that contain live Python code, equations, visualizations and narrative text. Here are some "operating tips"

- View Setup … Try these options. Select "View > File Browser" to see of the files in your local directory. Select "View > Appearance > Simple" to change the interface to view open notebooks in "tabs".
- You will find two types of "cells" in notebooks. A cell used for text or comment is called "Markdown" format. A cell used for Python code is called "Code" format. When you want to add a new cell in the notebook (use + in menu), be sure you use the correct cell type (dropdown selection on top-right).
- To run the entire notebook (starting from the top), you can select "Kernel > Restart & Run All". Once the code has been executed (top to bottom) you can change individual cell content and rerun portions of the code by going to any cell and hitting "Shift - Enter".
- When the code is "running" you will notice the cell blocks will look like "[*]". The "star" means the code is executing. When the cell is done executing the "star" will turn into a sequential number, starting with the last executed block number. If you run the entire stack, you can scroll to the top and see the code execute.



Name	Modified
Ground_Truth_final.csv	last month
S2_GDS.tiff	13 days ago
S2_sample_BRIAN.tiff	14 days ago
S2_sample_final.tiff	14 days ago
Sentinel2_BRIAN.ipynb	14 days ago
Sentinel2_GeoTIFF_fi...	14 days ago
submission_BRIAN.csv	14 days ago
submission_final.csv	14 days ago
Submission_template...	16 days ago
submission_TEST.csv	14 days ago
Training_final.csv	16 days ago
UHI_Benchmark_BRI...	14 days ago
UHI_Benchmark_final...	14 days ago
UHI_Benchmark_GDS...	10 days ago
UHI_Benchmark_TES...	14 days ago
UHI_Evaluation.ipynb	14 days ago

```
[1]: # Supress Warnings
     import warnings
     warnings.filterwarnings('ignore')

     # Visualization
     import matplotlib.pyplot as plt
     import seaborn as sns

     # Data Science
     import numpy as np
     import pandas as pd

     # Multi-dimensional arrays and datasets
     import xarray as xr

     # Geospatial raster data handling
     import rioxarray as rxr

     # Geospatial data analysis
     import geopandas as gpd

     # Geospatial operations
     import rasterio
     from rasterio import windows
     from rasterio import features
```