**Problem statement**

The ocean plays an important role in the carbon cycle by removing carbon from the atmosphere. But exactly how much carbon is absorbed by the ocean is unclear.

In this project, you will work with a state-of-the-art ocean data from the SOCAT database and build a machine learning model that predicts the surface ocean carbon dioxide (pco2) from physical and spatio-temporal features: sea surface temperature, salinity, chlorophyll, mixed layer depth, atmospheric pCO2, and time of the year/location on the Earth sphere. You will also compare the observations to a model (CESM-2).

**Pipeline**

We provide some starter code that allows you to assemble the feature matrix, make plots, and build a reference model using Gradient Boosting Regressor.

1. Plot the time mean (following the example) of the variables: sst, sss, chl and mld.

2. Can you do the same with xco2? What does this mean?

1. Build and optimize a ML model to predict the pCO2 from the feature (you can use the example to build your first model and then brach out by using a different algorithm, running optimizations, building new features. etc). My reference test R2 score is 0.80. Can you beat my score?
2. Look at error as f(different latitudes) (North/middle/South). You can use latitude (< -30, between 30 and 30, and > 30) to distinguish the different regions.
3. Look at different seasons (select using the time of the year) to see how the performance of the model varies. Is there a season where the model performs better?
4. Plot learning curves for your model. Would adding more data help?

3. Now plot the monthly mean of pco2 from model (CESM etc) and the pco2 from data. What can you observe?

4. If you used the ML model you just built in 2) to fill in the gaps, would you trust your results? Why or why not.

**Reference reading**

See Introduction and Section 1 of Bennington et al 2022:

https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002960