To give you some background: observations made by a camera onboard a spacecraft orbiting the Earth consist of the spectral reflectance for a particular location (pixel). The science goal is to infer the composition of water within the observed pixels. There are plant-like materials (e.g.., chlorophyll), particles, and absorbing organic matter in the water column that collectively govern the observed color of water.

While python is preferred, you may use whatever programming language you feel comfortable in. You may also use any learning models, feature transformations, data representations, etc. you feel are appropriate.

For this task, we want to see how you approach a problem and your process in solving it; it will also give you a sense of the type of research you'll be conducting at Goddard. The data sets can be found here: https://www.dropbox.com/s/rr8ii1xjurvcij7/data.zip?dl=0

Descriptions:

- Training.csv contains approximately 706k samples (which you may use a subset of, if you like). This set represents simulated Rrs data from the Sentinel-3 OLCI instrument, with the respective wavelengths (nm) given as the first row in the file. The final column (labeled 'Chl') represents the chlorophyll-a concentration for each sample.
- Validation.csv represents in situ measurements rather than simulated data. This file contains 100 samples, and follows the same format as the training set.
- Testing.csv also contains in situ measurements, but without the final 'Chl' column. This is the file which you will send us your final predictions for.

The task is to find the most accurate model in predicting chlorophyll, based on the observed spectra. Because the training data is simulated, you will likely find that out-of-sample training accuracy will not necessarily be comparable to what you find with the validation set. Don't get discouraged by poor results – remember this is a difficult task to see how you attempt the problem, rather than a judgment based purely on accuracy.

For the deliverable, send us a csv file containing your model's output on the testing.csv data. We would also like a pdf file containing a *short* description of the process you went through and the model(s) you used. Feel free to include any plots you think are helpful explaining your rationale. There isn't a hard limit to how short or long it should be, but *please* try to keep it under a page of written material – the ability to be concise is a valuable skill.

Please do let me know if you have questions.