Questions & Answers Session A

Please type your questions in the Question Box. We will try our best to get to all your questions. If we don't, feel free to email Sean McCartney (sean.mccartney@nasa.gov), John Just (JustJohnP@JohnDeere.com) or Erik Sorensen (SorensenErik@JohnDeere.com).

Question 1: What exactly does NDVI measure, and how is it calculated?
Answer 1: Normalized Difference Vegetation Index (NDVI) measures the amount of biomass or vegetation and is calculated using the following equation: NDVI = (NIR - RED) / (NIR + RED) where NIR is Band 8 and RED is Band 4 in Sentinel-2.

Question 2: Please clarify what the CDL data gives us and what it does not give. Does it give the time series data or just sequence of crops?
Answer 2: The CDL can give us the predicted (most likely) crop type, and it is also possible to get a confidence score of the prediction of the classes. The best link to see more about the details of the CDL is the FAQ page. It has many helpful links as well to supporting documentation. They only predict once for the previous year, which is often released in January of the following year (e.g. 2023 CDL was released in January 2024).

Question 3: Can you please suggest a good book for learning how to apply machine learning to remote sensing applications?
Answer 3: We are not aware of a book for this, which is part of the motivation for the training. The NASA ARSET series has quite a few good past videos (e.g. this one) besides the training we provide here, so we suggest you start with the NASA ARSET trainings that focus on machine learning applications. Beyond that, there are also journal and conference papers that you can read to learn more, but I would suggest that only after going through the NASA ARSET materials since it's usually more advanced.

Question 4: What are Scene Classification Layer (SCL) errors?
Answer 4: This is any situation where the Sentinel-2 scene classification layer model predicts the incorrect land cover/classification. The SCL is created as part of the

Sen2cor algorithm (Level 2A surface reflectance calculations), and has 12 different categories. Slide 37 in the slide deck talks a bit more about this.

Question 5: Any specific reason why Databricks is used? How would other platforms like Google Colab perform?
Answer 5: We utilize Spark as part of the data processing, which is useful to improve efficiency of data processing at large scale. Databricks has Spark installed out-of-the-box which makes it easy to set up the compute environment for this training. Databricks is also a popular cloud computing platform and is becoming more prevalent in the industry for cloud computing. Google Colab can use Spark as well but it takes more setup, and there is no guarantee the code we provide will run in a Google Colab environment. Databricks also has 10GB persistent storage.

Question 6: Is there a specific cloud provider we should be choosing?
Answer 6: For data, we pull from the AWS Open Registry of data, and specifically the Sentinel-2 COGs (cloud optimized geotiffs) simply because an account isn't required to access. However if you intend to use other data sources such as Landsat or orthorectified Sentinel-1, you will need to have an AWS account to pull from the open registry since the user has to pay for data transfer. Other options such as Google's Earth Engine or Microsoft's Planetary Computer could be used as well, but we haven't tested them at scale.

Question 7: Can we differentiate different vegetative growth stages of a crop from satellite images like flowering, fruiting, etc.?
Answer 7: Our approach we use here could in theory be extended to predicting growth stages of a crop by utilizing a different target set other than the CDL that would provide those labels of growth stage.

Question 8: Do we have to connect the data through API, for Sentinel Data?
Answer 8: Yes, we utilize several APIs to get both the Sentinel-2 imagery and the CDL labels. We covered how to do this during the demo portion of this training.

Question 9: Is it possible to make a data pipeline for training instead of downloading a shapefile?
Answer 9: You do not have to download a shapefile. We simply did this to visually show how and where we drew boundaries to focus on for this data pipeline and ultimately

the data we use for modeling. The pipeline we create is generalizable for whatever area of interest (AOI) you want to get imagery from for training models.

Question 10: Did you filter clouds using a SCL band?
Answer 10: on Slide 47 yes, we filtered high and medium cloud cover (but not thin cirrus) based on the SCL.

Question 11: Is there a research paper associated with the methodology we are following? I want to see a comparison of the model accuracy for a given area of interest (AOI) with other existing models like Esri Land Cover and Dynamic World etc.
Answer 11: We don't have a specific research paper around this but it would be very easy to use the same boundaries to compare your final trained models with other crop/land classification sources. You will see in the final third session (the culmination of the work) how to train your model and you will be able to essentially use any model you like there to try and maximize your predictive performance.

Question 12: How do we import the python file into databricks?
Answer 12: The file can be imported by going into Workspace on the left, right clicking in the directory and selecting "Import". From there you can upload the python file and it will show up in your workspace.

Question 13: Este mismo análisis se podría realizar en otra región del mundo, por ejemplo Argentina? Dado que están mostrando los EEUU, ¿tienen un sistema mucho más avanzado en base de datos?
[Translation: Could this same analysis be carried out in another region of the world, for example Argentina? Given that they are only showing the US, do they have a much more advanced database system?]
Answer 13: The Cropland Data Layer only makes predictions for the Contiguous United States. However, if there are other crop classification models that provide crop predictions in other locations around the world, those can easily be swapped in place of the CDL in this pipeline. E.g., if you have boundaries and you have ground truth already for those areas, then you would just want to format your ground truth into a table similar to the format we use from the script in Part-1. Then from there use that table in part-2 for obtaining the data. Just be careful of the CRS (coordinate reference system) of your boundaries when querying the STAC API for sentinel-2 data.

Question 14: Do we need to perform radiometric calibration? If so, how do we do it?
Answer 14: No, we are using level 2A processing from Sentinel-2, which has already had atmospheric corrections applied.

Question 15: Will we be able to run all the analyses in this training using the free account of Databricks? (Interacting with the notebook so the timer resets)
Answer 15: Yes, we designed the scripts so that they can all be run within the Databricks Community Edition. If any of the scripts take longer than 60 minutes you can reset the timeout timer by executing another code cell in a different notebook using the same compute.

Question 16: Why use Pyspark specifically when Sedona can load balance and lessen processing time?
Answer 16: Apache Sedona extends Apache Spark by providing additional functionality and performance when performing geospatial queries (like spatial joins). In our example, we do not actually perform any complex spatial queries that Sedona excels at, so Spark is sufficient for our processing.

Question 17: Have you worked with Tensorflow and Javascript in Google Engine?
Answer 17: Only on a small scale playing around with it, but it was restricted mostly to that platform and didn't allow me to chain together my own pipeline.

Question 18: What is the meaning of .src files and .parquet files?
Answer 18: .src & .py files are the script files, and .parquet files are the processed data files. These can both be uploaded/imported to Databricks Community, but you need to make sure the .parquet files are zipped since there are many of them. Then unzip them once uploaded using a script.

Question 19: Is CDL data available only for the US? How can we use this code in non-US areas (assuming we have some sample data from the ground in the non-US area)?
Answer 19: Great question - in this case you just use your particular boundaries from where your sample data came from instead of CDL data. E.g., if you have boundaries

and you have ground truth already for those areas, then you would just want to format your ground truth into a table similar to the format we use from the script in Part-1. Then from there use that table in part-2 for obtaining the data. Just be careful of the CRS (coordinate reference system) of your boundaries when querying the STAC API for sentinel-2 data.

Question 20: Regarding uses of raster package(s): I know of one we can use for reprojection, calculating the transform, or resampling, but how are we using it here?
Answer 20: Rasterio allows us to sample data from points in rasters as well. Here we use that functionality to move from raster format of the data to tabular format for use with Spark.

Question 21: Is Databricks being used to access and process large data? Could this run in Google Colab without a Databricks profile?
Answer 21: We utilize Spark as part of the data processing, which is useful to improve efficiency of data processing at large scale. Google Colab can use Spark as well but it takes additional setup, and there is no guarantee the code we provide will run in a Google Colab environment.

Question 22: Hi John, thanks for putting together the code for this demo. Could you repeat generally how we could write a function for getting the Sentinel-2 bounds?
Answer 22: We don't really get the bounds of Sentinel-2 so to speak, but we send our bounds (AOI) to the STAC API that searches for Sentinel-2 imagery within our bounds and timeframe. The "Part-2" script has a helper function for this. If you wanted to make the searches and processing slightly more efficient, you could do all your searches first and then group your data into sentinel-2 scenes to ensure you only ever download each scene once.

Question 23: Can Databricks accommodate a kind of modeling equipped with some basic GIS functions such as overlays or area measurements or even more... can I just rewrite the toolset script like in ArcGIS/QGIS just because they both use Python?*
Answer 23: This pipeline is not set up to integrate with ArcGIS/QGIS and I am not aware of any way to connect the Databricks Compute to those tools. However, Folium

and Plotly are both libraries that may be capable of doing this. Some functionality within a Databricks notebook environment is limited though so it may only be possible on a more typical setup.

Question 24: I'm having trouble with running Part 1 script, the output files are not being saved. Does anyone have the same problem?

Answer 24: Look in the Catalog → then you can see the option to toggle to "DBFS" (it defaults to "Database Tables".
https://docs.databricks.com/en/administration-guide/workspace-settings/dbfs-browser.html

Question 25: What approach would you suggest to apply outside the US? How extensive does the ground truth data need to be?
Answer 25: The Cropland Data Layer only makes predictions for the Contiguous United States. However, if there are other crop classification models that provide crop predictions in other locations around the world, those can easily be swapped in place of the CDL in this pipeline. E.g., if you have boundaries and you have ground truth already for those areas, then you would just want to format your ground truth into a table similar to the format we use from the script in Part-1. Then from there use that table in part-2 for obtaining the data. Just be careful of the CRS (coordinate reference system) of your boundaries when querying the STAC API for sentinel-2 data. If no alternative crop classification model exists for your AOI, an alternative approach could be looking at areas in the US that are similar in both agricultural practices and climate, training a model for that area, and applying it to your area of interest. If you pre-train your model on the CDL in such a way and the crop types are mostly covered, then it is likely that the amount of ground truth you need can be significantly reduced (i.e. you may just need a little to "fine tune" your model), or eliminated entirely if performance looks acceptable.

Question 26: Is it possible to work with a multiband dataset assuming we get the data from a source like Google Earth Engine and convert it to xarray.dataset?
Answer 26: With some changes to the scripts we provide, those files could be used as part of our method here. A suggested method could be converting the xarray.dataset to spark dataframes to be used as part of our pipeline here.

Question 27: I've been fetching Sentinel-2 imagery from Google Earth. Do you know if there are differences, pros and cons, in using Google versus earth-search stac?
Answer 27: Scalability. We can massively multiprocess our pipeline with the sentinel-2 COG source from AWS open registry. More generally speaking, the less rate-limited you are the better...but there are other considerations like data transfer costs (e.g. per GB) that often come into play for other sources such as Landsat 8/9 from the AWS open registry.

Question 28: Are there running programs with this methodology running in NASA Acres?
Answer 28: Not that we are aware of.

Question 29: For NDVI, which formula is being used? Has Bias formula been used in this particular case?
Answer 29: The equation used is NDVI = (NIR - RED) / (NIR + RED) where NIR is Band 8 and RED is Band 4 in Sentinel-2.

Question 30: The notebook Part-1_CDL-Acquisition runs fine for me up to Cell 23: "tif_bytes_list = [[CDL_clip_retrieve(x, y) for x in bbox_list] for y in years]", but this cell has now run for >30 minutes on a powerful cluster (1024 GB and 128 cores), so something big is happening, or is something wrong happening?
Answer 30: It shouldn't take that long – the entire Part-1 notebook should complete in about 5 minutes on a Databricks Community Edition node.

Question 31: Can we obtain soil moisture data from Sentinel-2 and the process visited today?
Answer 31: Soil moisture data is not provided using this process, but there is active research and products available that estimate soil moisture from Sentinel-1 & Sentinel-2 imagery. (e.g. Planet Scope Soil Water Content)

Question 32: You mention the shifting of the scenes between different dates for the same AOI. Did you apply collocation at subpixel level as with SNAP?
Answer 32: No collocation was applied in this pipeline. Any noise this adds is a small fraction of the total dataset size and will have minimal impact on model training.

Question 33: Did you explore other vegetative indices like EVI for using as a feature to train the model?
Answer 33: We will get into more details in Part 3 when we train the model, but we only used the 12 bands of Sentinel-2 as input to the CNN model we trained. Indices like NDVI and EVI are perfectly correlated to the information contained in the bands so they were excluded as model input.

Question 34: Does the geodetic coordinate system (With Curvature) or UTM (flat) affect the accuracy in image classification? What coordinate system is the Python code for satellite image classification in?
Answer 34: No - we do all our final classification work on the pixels in a table format, where each pixel has a lat/lon (EPSG: 4326) value based on the center of the pixel where it came from in the CDL table. During sampling of the rasters we have to project the lat/lon points from epsg 4326 to the CRS of the raster we sample from, but we don't retain the projection (that is just for sampling). If you wanted to make this more general across years by using an absolute geospatial reference system, you could sample rasters using Uber's H3 hex library.

Questions & Answers Session B

Please type your questions in the Question Box. We will try our best to get to all your questions. If we don't, feel free to email Sean McCartney (sean.mccartney@nasa.gov), John Just (JustJohnP@JohnDeere.com) or Erik Sorensen (SorensenErik@JohnDeere.com).

Question 1: Is there any statistical analysis between interval grouping and interpolation? Why exactly is interval grouping used in the CDL algorithm?
Answer 1: Interval grouping is simpler/easier and doesn't require additional assumptions of smoothing needed for interpolation. It's also the method used by the CDL.

Question 2: For the CDL model, why is the accuracy changing from 85% to 95%?
Answer 2: The accuracy of the CDL model differs between crops, this means that the CDL is more accurate when predicting more common crops (like Corn and Soybeans) and less accurate when predicting uncommon crops (like Sorghum). Also see user VS producer accuracy.

Question 3: In machine learning, working with satellite data, what will be the minimum number of images that you need for the model?
Answer 3: The number of images or time-series examples needed for the model varies depending on the use-case (e.g. more complex tasks will require more data). For our task of predicting CDL in real-time, the training set includes ~80,000 time-series of pixels with an average of ~60 pixels per time-series.

Question 4: Why not use radar (SAR) sensors with 12 bands of Sentinel-1A to improve accuracy and overcome the factors that can affect the quality of satellite images?
Answer 4: You can certainly combine different satellites like SAR and optical sources that would likely improve temporal resolution (and provide potentially more accurate detection capabilities), but due to the complexity of such an effort to get the data from different/various sources (at a large scale…which is our primary interest here) and the complexity of preparing the data for modeling we don't do that in this particular training.

Question 5: What would be the Sentinel-2 image quality if we would apply cloud masking between the SCL and QA60 bands?
Answer 5: The SCL layer provides discrete classifications of quality issues such as clouds and is available at the L2A (ground reflectance) processing level. The QA60 band appears to be similar in intent but for level L1C processing (top of atmosphere)

Question 6: What analysis do you use to convert gaps due to SCL errors into appropriate values (such as missing value imputation)?
Answer 6: We do not impute any missing values due to things like SCL errors or clouds, rather we will pad these missing values in the time-series with a constant value (this essentially tells the CNN to impute values for us implicitly). We will get into more details of this process in Part 2.

Question 7: What do we need to create a CDL for Morocco like the one in the USA?
Answer 7: One way to handle this is to simply train on a diverse set of crops that represent Moroccan crops from the CDL across as wide a range as possible from the US, and then run inference on Moroccan AOIs. This is probably the easiest way to start out – if the model performs well then maybe you are done, but if not then you may want to “fine tune” your model on a small set of data from Morocco that is more representative of the crops there. At least this minimizes the data you need to train by first running your model training on the US CDL.

Question 8: How can I download CDL Plots and what kind of code should I use to have them?
Answer 8: The first script here does all that for you. it is intended to be such that you can just "run" the script and get the data. After you understand it you can change it to suit your needs

Question 9: Is HDF5 also a common file format used in remote sensing?
Answer 9: HDF5 & parquet files are a general file format used to store data in a distributed format which supports efficiency when working with distributed cloud computing. It can be used for remote sensing data. I don’t know if it is a common data format for remote sensing, but it is a common file format when working with distributed cloud computing and large scale data

Question 10: Can we use xarray to make the process easier to handle?
Answer 10: you can substitute other functions if you like -- just make sure to compare against the original to ensure it produces the same output.

Question 11: Can I run the code on Google Colab?
Answer 11: Yes this should be possible as long as spark is installed on the Google Colab instance. However, the code we provide will not work directly in a Google Colab environment without making some changes to the code (e.g. storing files in the DBFS persistent storage would need to be changed to save in whichever persistent storage is available in Google Colab)

Question 12: Can I try on other AOIs that are not within the US?

Answer 12: Yes, the part-2 script will work for any AOI - just have your "targets" formed in a similar format as what we did in part-1 for the CDL, but otherwise it will scale very well for any area. Ultimately this code is searching for Sentinel-2 imagery and sampling it at the points where you have ground truth data

Question 13: Do we resample Sentinel-2 data to 30m resolution anywhere in the code?

Answer 13: The code is sampling (not resampling) the rasters based on the center point of each pixel from the CDL that we chose to use in our training data. Thus for each CDL pixel (which is 30m resolution) we get a time series of sentinel-2 data to train a model with. We are essentially trying to avoid having to work with rasters by converting the data into tables, where each row represents a 30m pixel that we are predicting on.

Question 14: Should we create the bounding boxes for each field in Europe for example? And should we know in advance which field is planted with which crop type?

Answer 14: You don't necessarily have to train on European data….but you'll need AOIs for the areas you want to predict/infer on. You can train on US CDL data and then try running inference of your model on European data. Just try to match the crop distributions of Europe when training your model.

Question 15: The SNAP toolbox library is available in Python. Did you use this library to perform any part of the preprocessing of the S2 images?

Answer 15: No, the SNAP library was not used. We specifically utilized libraries that allow us to massively scale the processing and we didn't need SNAP to do that.

Question 16: When running Part-1, I got an IndexError at Get the data part, stating that list index is out of range. What causes this problem?

Answer 16: try running it again....I think I may have some random sampling in there that may just not have a check for when the random index goes out of bounds. if you run a couple times it should eventually get an index in bounds

Question 17: Hi Erik, can we use the Planet data instead of Sentinel 1?

Answer 17: Yes, other satellite imagery sources can be used instead of Sentinel-2 for this pipeline. Sentinel-2 was used for this training due to its availability and cost (it's free!).