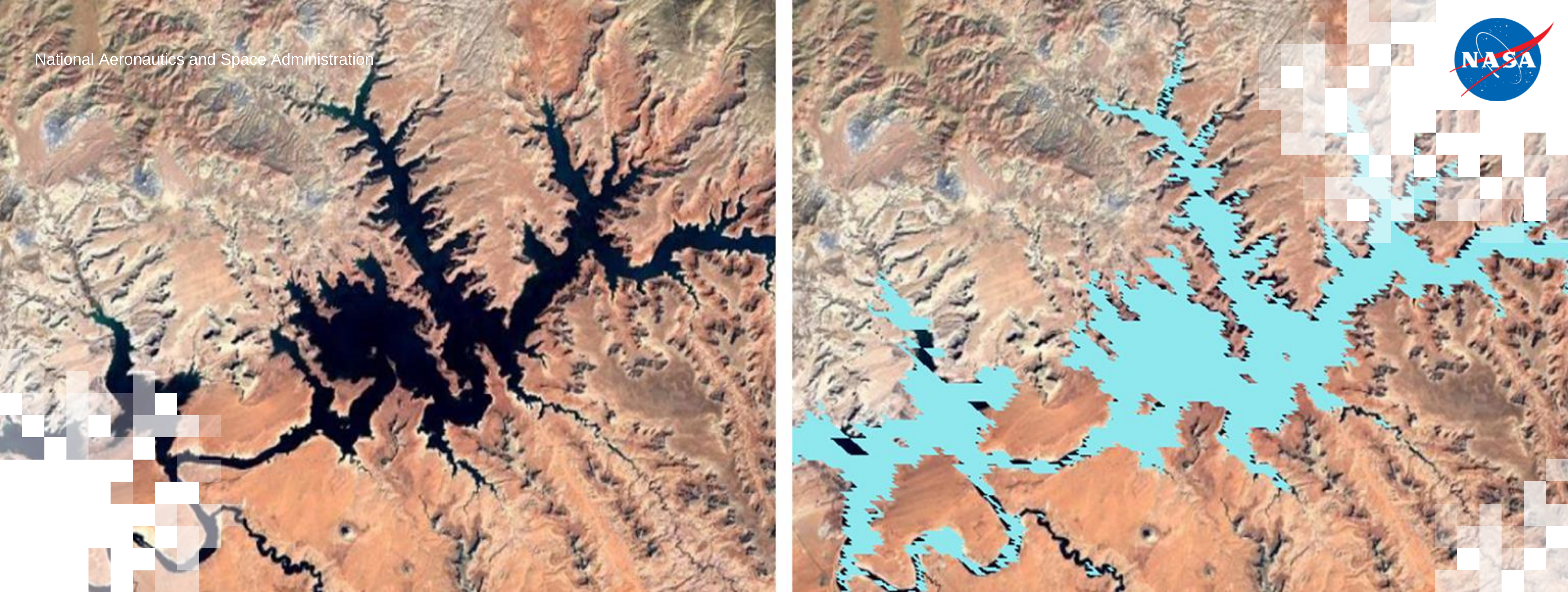


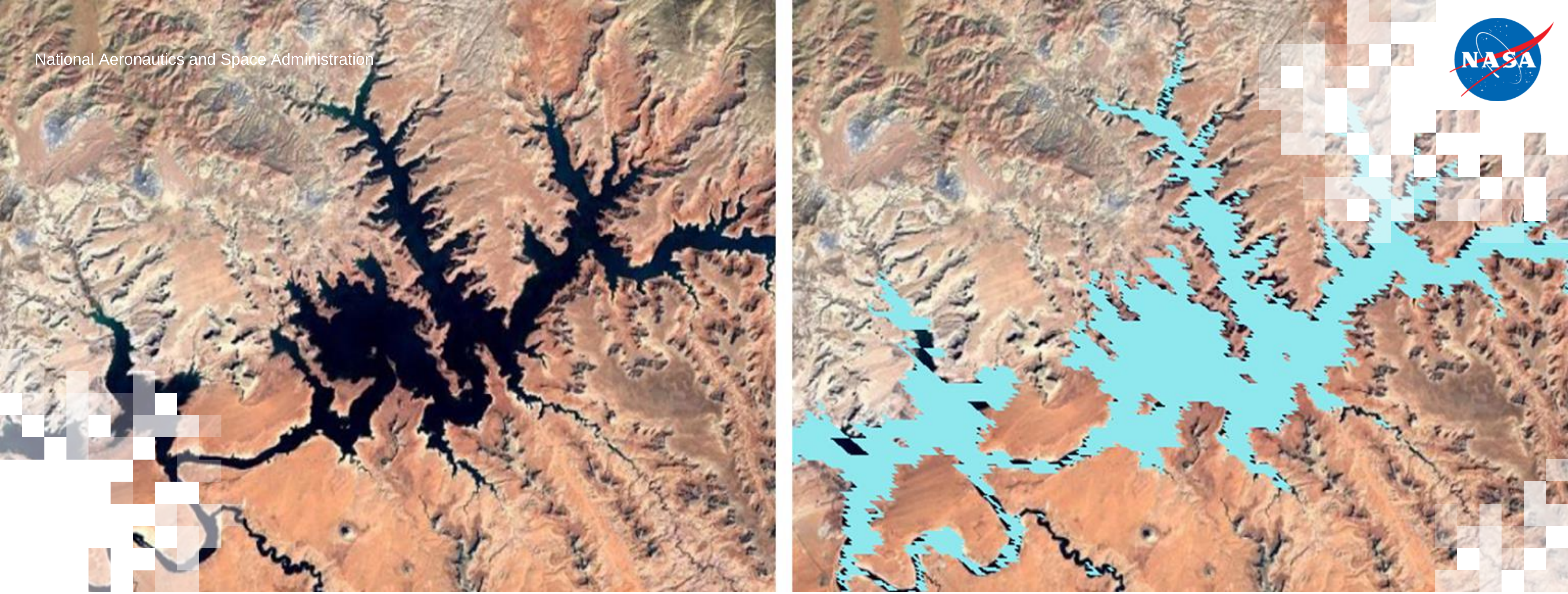
Fundamentals of Machine Learning for Earth Science

Part 2: Training Data and Land Cover Classification Example

Trainers: Jordan A. Caraballo-Vega, Mark L. Carroll, Jules Kouatchou, Jian Li, Caleb S. Spradlin

April 27, 2023

National Aeronautics and Space Administration



NASA Applied Remote Sensing Training (ARSET)

Brock Blevins, Training Coordinator

Training Objectives

At the end of the training, participants will be able to:

- Recognize the most common machine learning methods used for processing Earth Science data
- Describe the benefits and limitations of machine learning for Earth Science analysis
- Explain how to apply basic machine learning algorithms and techniques in a meaningful manner to remote sensing data
- Use an analysis-appropriate training dataset to evaluate conditions and solutions for a given case study
- Complete basic procedures to interpret, refine and evaluate the accuracy of the results of machine learning analysis

Reminder of Prerequisites

- Prerequisites:
  - Session 1 of our on-demand Fundamentals of Remote Sensing series or have equivalent experience.
  - Attendees will need access to Google Drive and Google Colab. To access these resources, users must use an email ending in ‘gmail.com’.
  - We will have the video of this demonstration within the training recording available within 48 hours after the presentation for you to go through at your own pace.

Training Schedule

Part 1:

Overview of Machine Learning

April 20, 2023

Part 2:

Training Data and Land Cover Classification Example

April 27, 2023

Part 3:

Model Tuning, Parameter Optimization, and Additional Machine Learning Algorithms

May 4, 2023

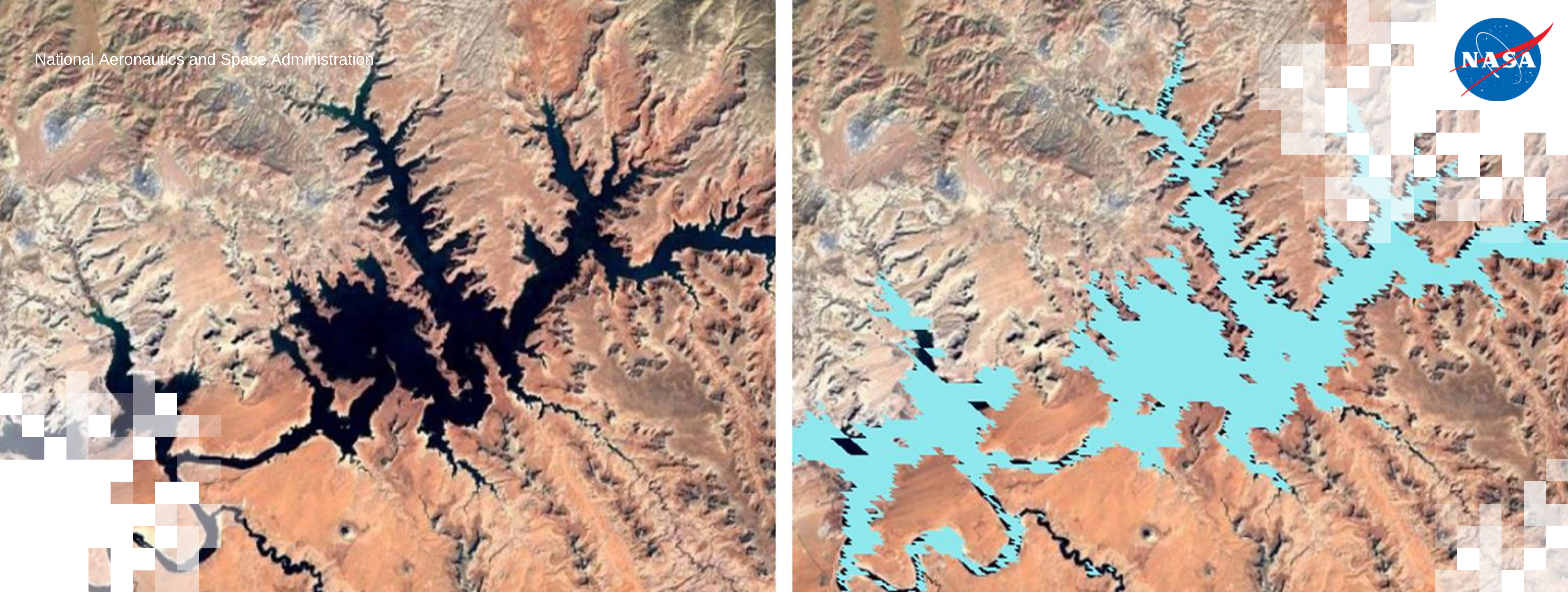
Homework

Independent practice and application

Due May 19
Opens May 4

Optional opportunity to earn a certificate of completion

National Aeronautics and Space Administration



Fundamentals of Machine Learning for Earth Science

Part 2: Training Data and Land Cover Classification Example

Trainers: Jordan A. Caraballo-Vega, Mark L. Carroll, Jules Kouatchou, Jian Li, Caleb S. Spradlin

April 27, 2023

Session 2 Outline

- Download the training data
- Exploratory data analysis
- Extracting training data from a tabular dataset
- Extracting training data from a raster dataset
- Training and inference of a tabular and raster dataset
- Metrics and model evaluation
- Hands on Jupyter Notebook Exercise: MODIS Water Classification Case Study
- Post-Session Assignment
- Q&A Session

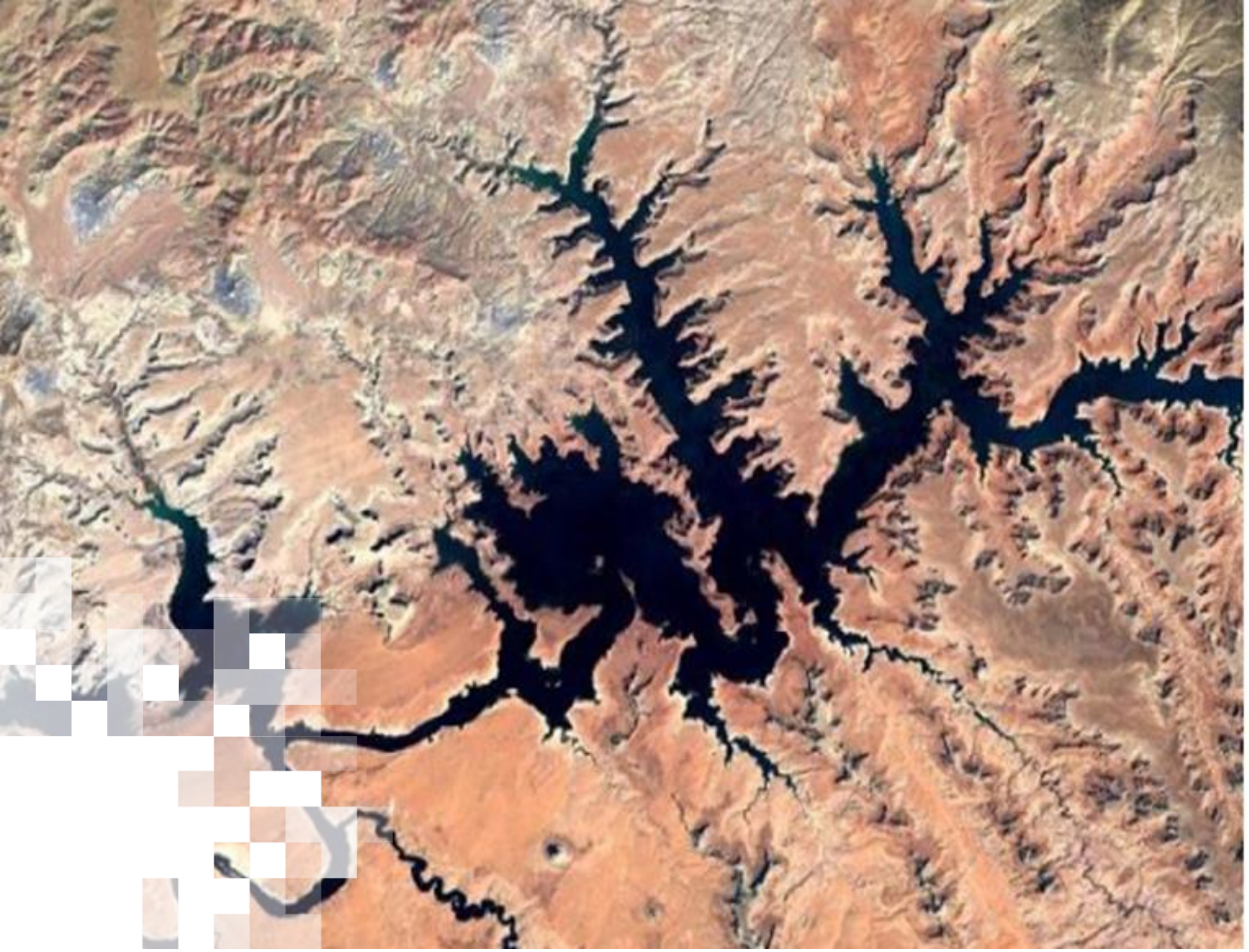
Resources for this Training

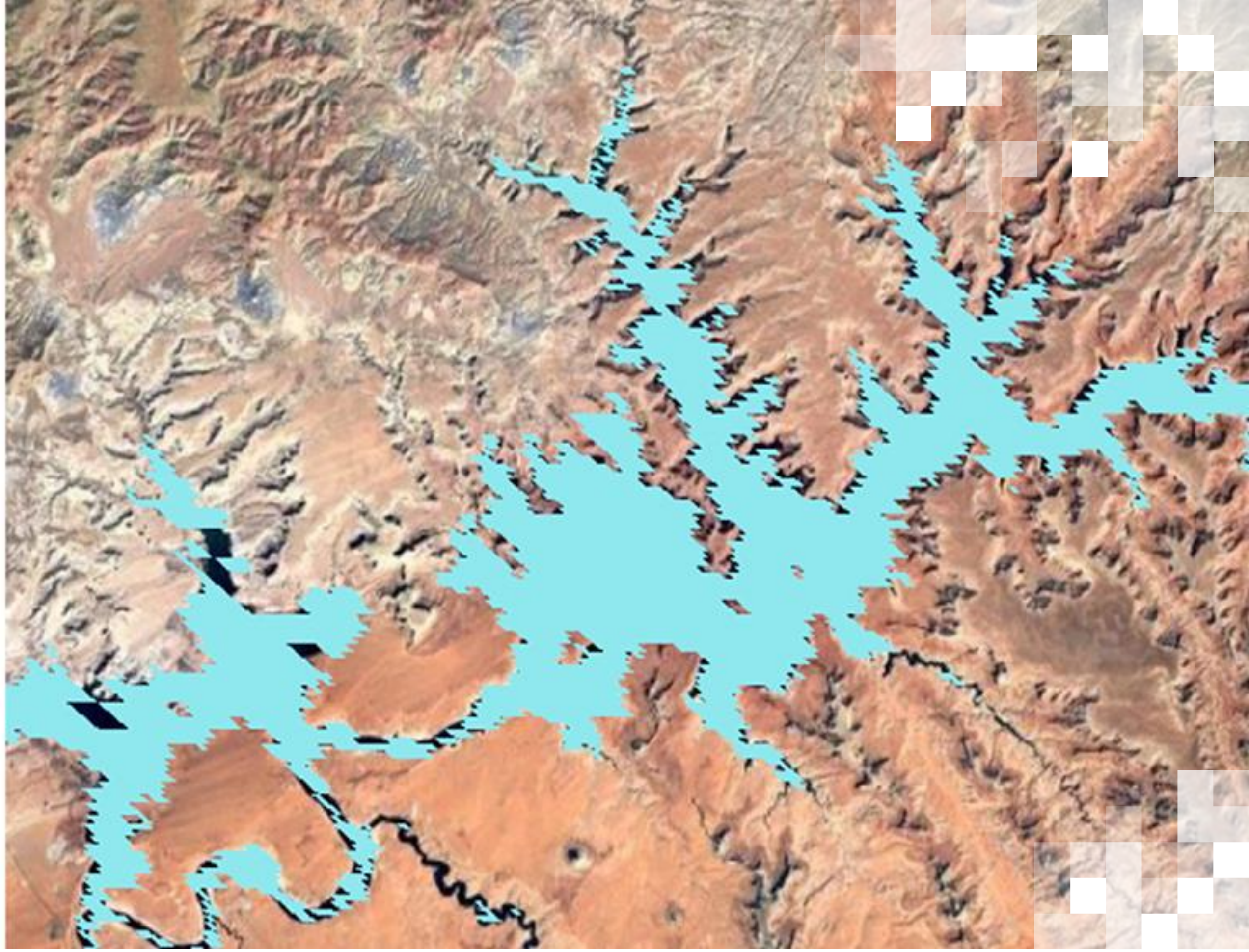
https://github.com/NASAARSET/ARSET_ML_Fundamentals

Training Objectives

After participating in this training, attendees will be able to:

- Use basic programming procedures to download, use and process remote sensing data
- Use an analysis-appropriate training dataset to evaluate conditions and solutions for a given case study
- Complete basic procedures to interpret, refine and evaluate the accuracy of the results of machine learning analysis





Overview of the Instrument and Data

Trainer: Jian Li

Moderate Resolution Imaging Spectroradiometer (MODIS)

- MODIS is a key instrument aboard the **Terra** and **Aqua** satellites.
- MODIS is viewing the entire Earth's surface every 1 to 2 days.
- MODIS data products, including atmosphere, ocean, land, and cryosphere, are used to study global change.
- Acquired data will improve understanding of global dynamics and processes occurring on the land, in the ocean, and in the lower atmosphere.

Imaging Spectroradiometer

Phytoplankton bloom in the Black Sea in June 2000. Brown sediment discharge from the Danube delta is hugging the western coast, and the phytoplankton bloom is evident by the green and blue colors in the central and eastern side of the image. Image credit: MODIS Land Team/Jacques Descloitres, SSAI; MODIS Ocean Team/Ron Vogel, SAIC/GSC.

NASA's Earth Observing System

Application of MODIS Land Products

Global Raster Water Mask at 250m Resolution

- Expand the spatial coverage to include the entire planet & address some erroneous discontinuities in major river networks

Data

Subset & extract surface reflectance

Quality assurance

Construct training dataset

Model

Random Forest/XGBoost/LightGBM

Train with CPU & GPU

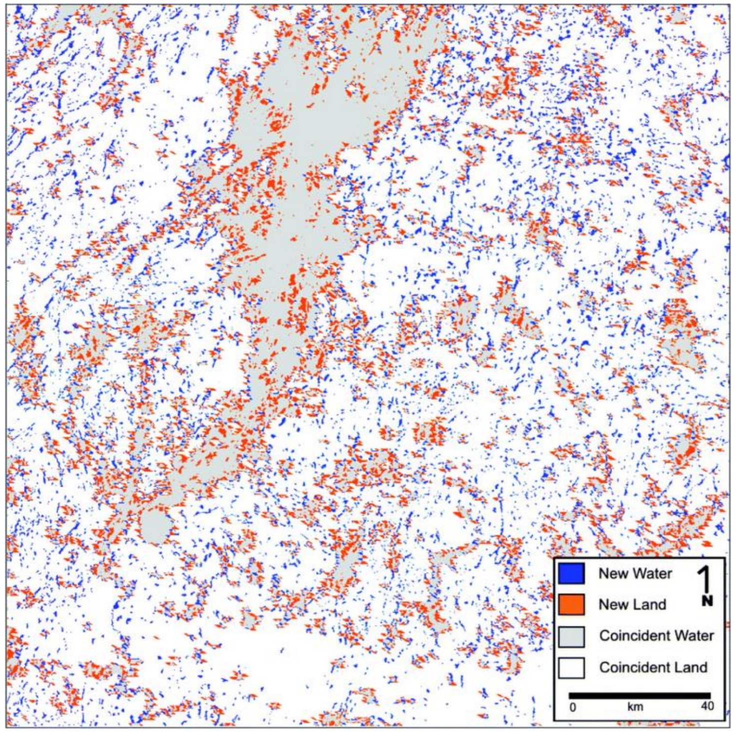
Hyper-parameter tuning

Results

Annual & daily water/land map

7-class water maps per year

MODIS Water Classification Workflow



Improved representation of lakes in Boreal Canada west of Hudson Bay as compared to the old EOS water mask. Image source: Carroll et al. https://doi.org/10.1080/17538940902951401

Access MODIS Data

- NASA data are stored at Distributed Active Archive Centers (DAACs).

- MODIS Level 1 Data, Geolocation, Cloud Mask, and Atmosphere Products:
  - http://ladsweb.nascom.nasa.gov/

- MODIS Land Products:
  - https://lpdaac.usgs.gov/

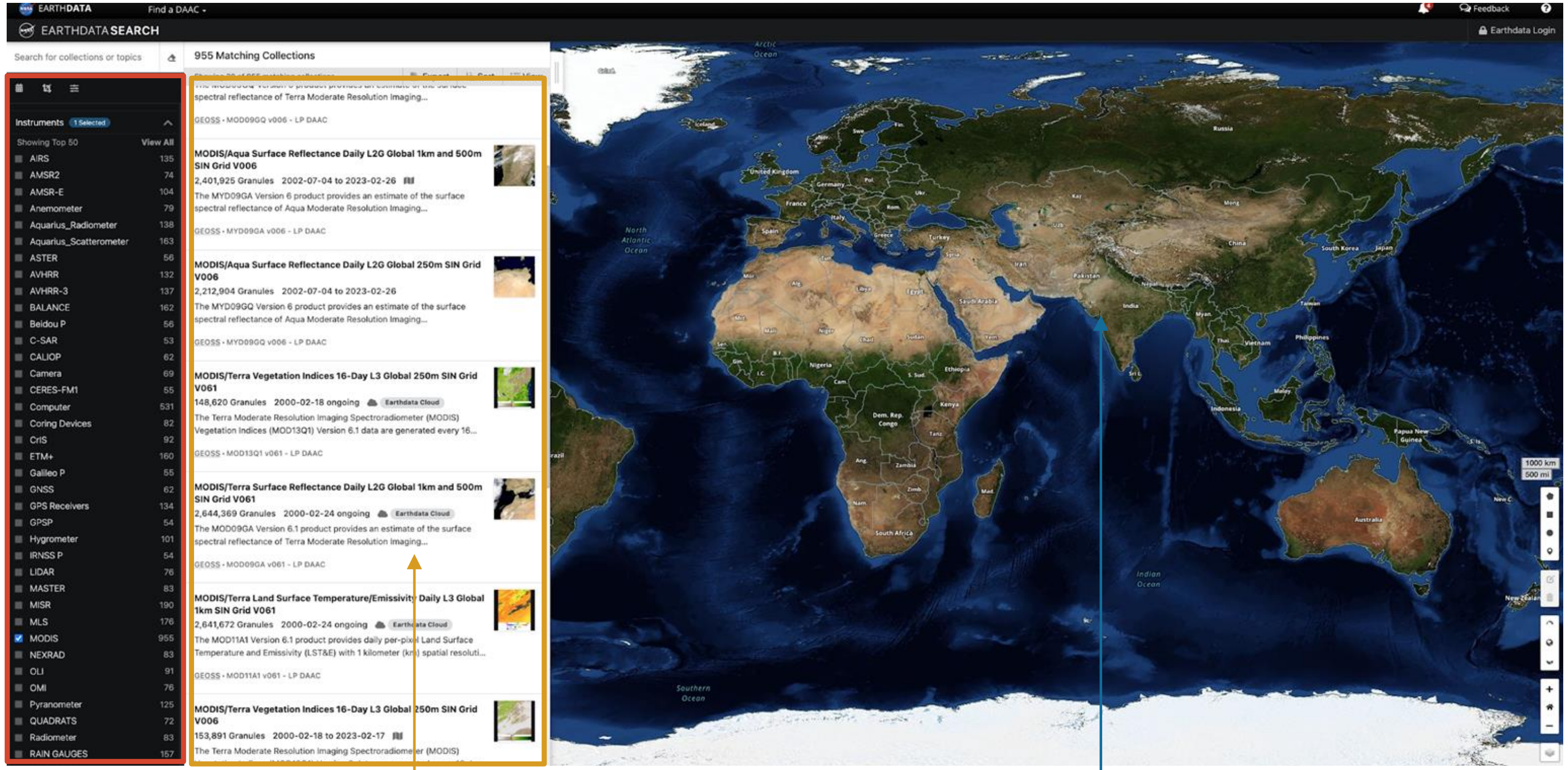
- MODIS Cryosphere Products:
  - http://nsidc.org/daac/modis/index.html

- MODIS Ocean Color and Sea Surface Temperature Products:
  - http://oceancolor.gsfc.nasa.gov/

NASA Earthdata Search

Earthdata Search provides access to all DAAC data via a map web-based interface.

Search data based on criteria, such as instrument

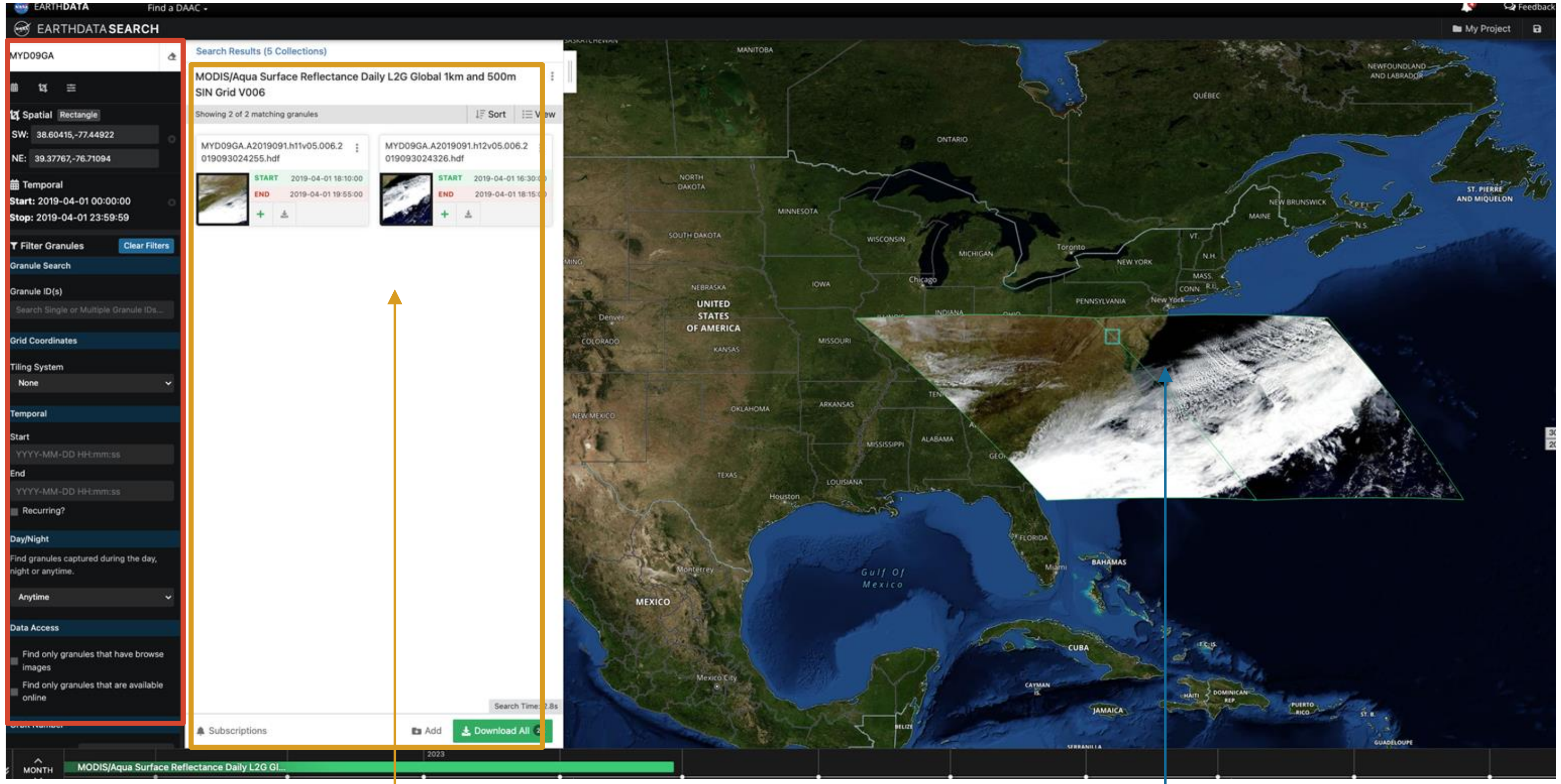


List of MODIS Products

Preview Sample Data

NASA Earthdata Search, Cont.

Filter data through product name, spatial coverage, time period, etc.

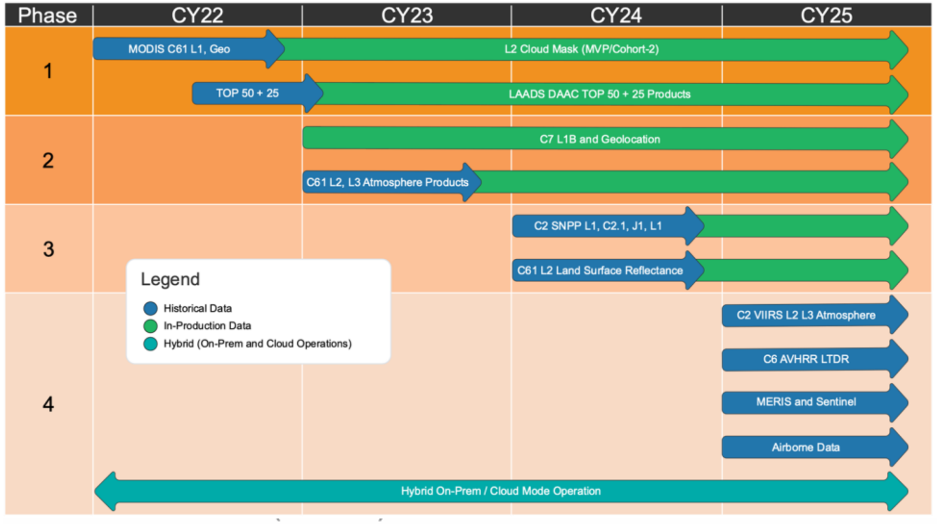


List of MODIS Granules

Preview Data

MODIS Data in the Cloud

As part of NASA's open-science policy and related goals, all DAACs are migrating their collections to the Earthdata Cloud.



LAADS DAAC Phase 1.1 (Cohort-2) Datasets

Shortname	Platform	Instrument	Description	Availability
MOD021KM	Terra	MODIS	Level 1B Calibrated Radiances - 1km	December 2022
MYD021KM	Aqua	MODIS	Level 1B Calibrated Radiances - 1km	December 2022
MOD02HKM	Terra	MODIS	Level 1B Calibrated Radiances - 500m	December 2022
MYD02HKM	Aqua	MODIS	Level 1B Calibrated Radiances - 500m	December 2022
MOD02QKM	Terra	MODIS	Level 1B Calibrated Radiances - 250m	December 2022
MYD02QKM	Aqua	MODIS	Level 1B Calibrated Radiances - 250m	December 2022
MOD03	Terra	MODIS	Geolocation - 1km	December 2022
MYD03	Aqua	MODIS	Geolocation - 1km	December 2022
MOD35_L2	Terra	MODIS	Cloud Mask and Spectral Test Results 5-Min L2 Swath 250 and 1km	December 2022

The Amazon Sustainability Data Initiative
ASDI

MODIS MYD13A1, MOD13A1, MYD11A1, MOD11A1, MCD43A4

This product is part of the Amazon Sustainability Data Initiative and contains data sets that are publicly available for anyone to access and use. No subscription is required. Unless specifically stated in the applicable data set documentation, data sets available through the Amazon Sustainability Data Initiative are not provided and maintained by AWS.

Description | Resources on AWS | Usage examples | Links | Similar products

Contact: https://astraea.earth/

General AWS Data Exchange support
Contact Us

Resources on AWS

Description
Imagery and metadata

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::astraea-opendata`

AWS Region
`us-west-2`

AWS CLI Access (No AWS account required)
```
aws s3 ls --request-payer requester s3://astraea-opendata/
```

The proposed timelines for migrating LAADS DAAC data products (top left)

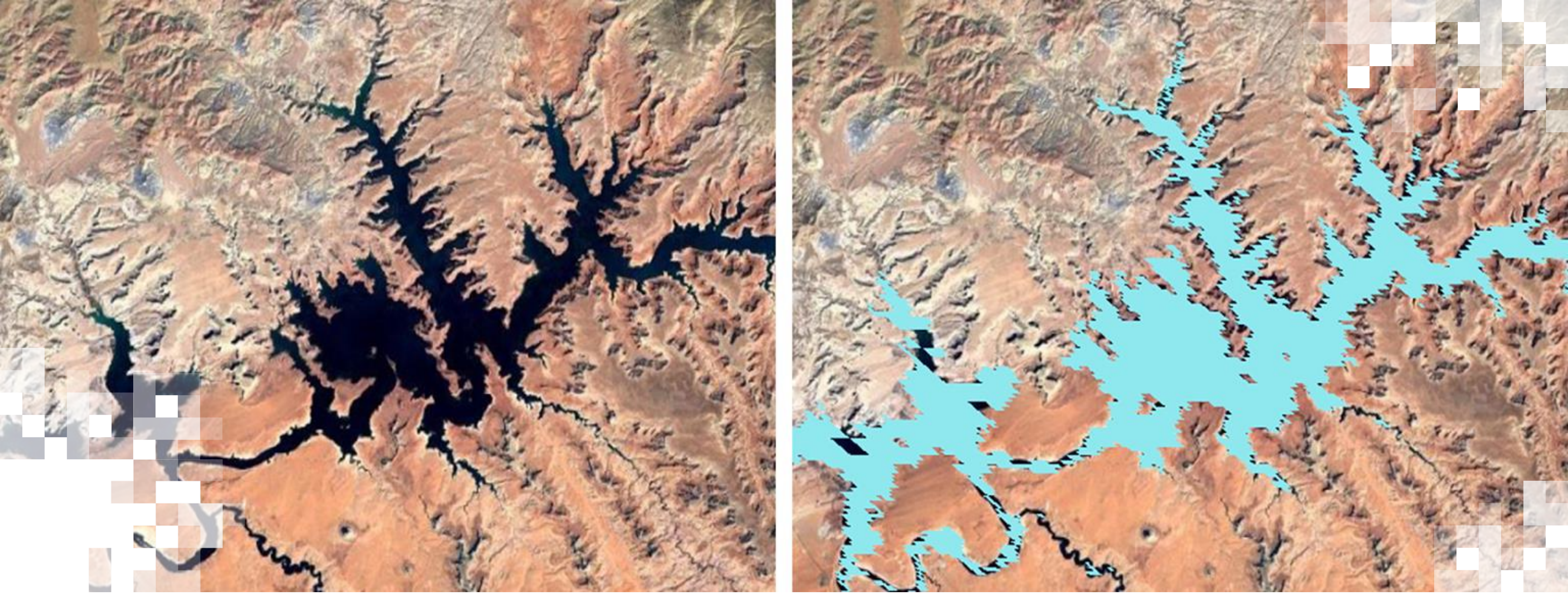
Samples of MODIS products migrated in early phase (bottom left)

Source: LAADS DAAC
https://ladsweb.modaps.eosdis.nasa.gov/cloud/

Available MODIS products through ASDI (right)

Source: AWS ASDI

https://registry.opendata.aws/modis-astraea/



Exploratory Data Analysis

Trainer: Caleb S. Spradlin

Machine Learning Pipeline from Session 1

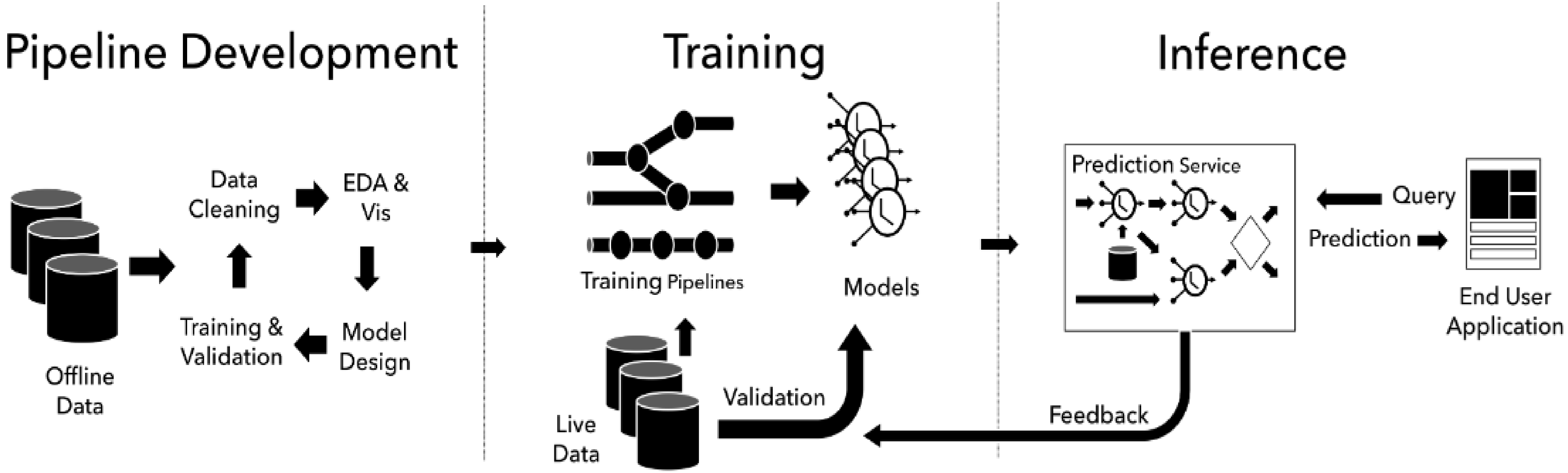


Image Source: Daniel Crankshaw (In: 'A Short History of Prediction-Serving Systems')

Exploratory Data Analysis (EDA) for Machine Learning

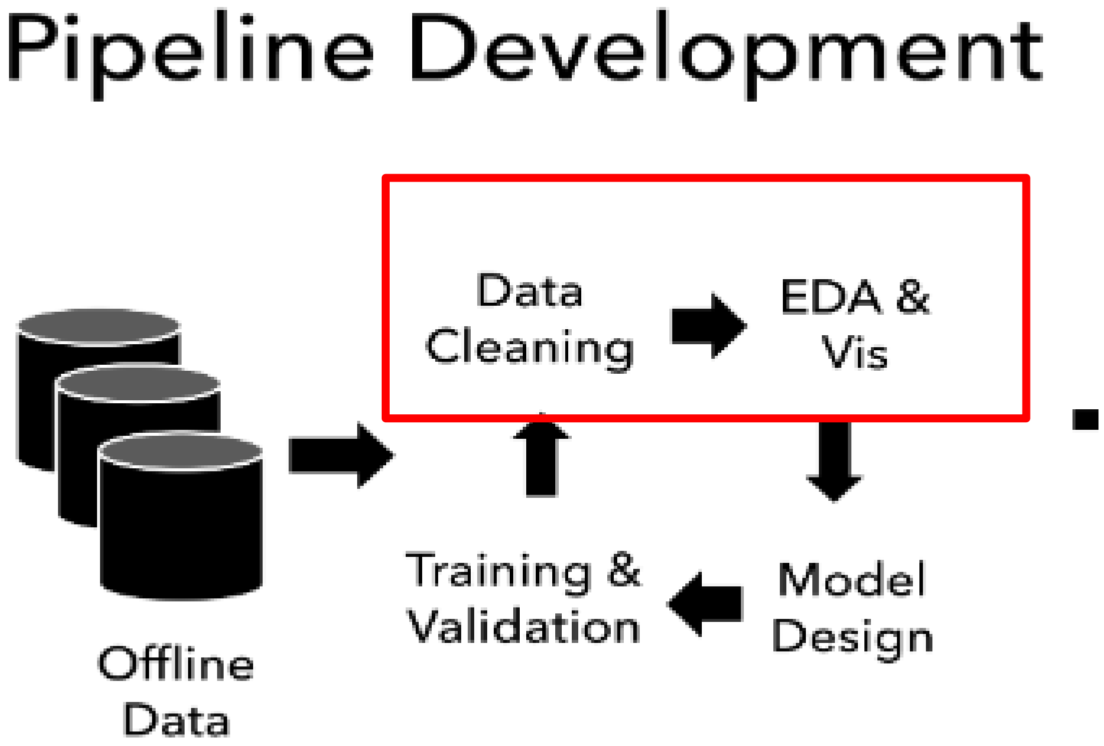


Image Source: Daniel Crankshaw (In: 'A Short History of Prediction-Serving Systems')

Exploratory Data Analysis (EDA) for Machine Learning

EDA is an approach for data analysis using a variety of techniques to gain insights about the data.

Basic steps in any exploratory data analysis:

- Cleaning and preprocessing
- Statistical analysis
- Visualization for trend analysis, anomaly detection, outlier detection (and removal)

Pipeline Development

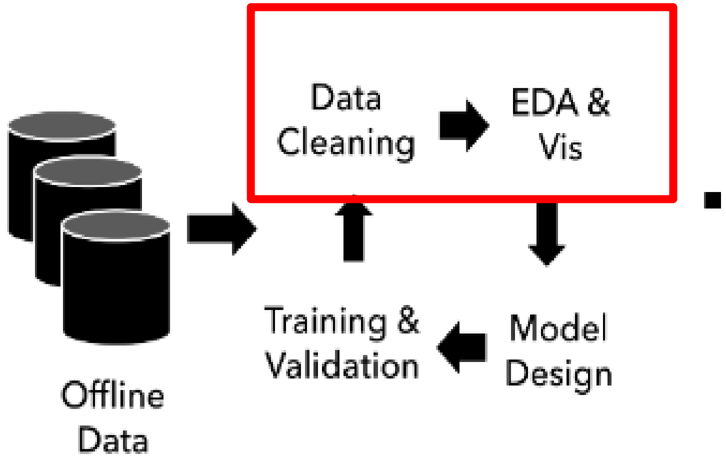


Image Source: Daniel Crankshaw (In: 'A Short History of Prediction-Serving Systems')

EDA – Importance of EDA

Improve our understanding of the structure and properties of the dataset

Discover errors, missing values, and outliers in the dataset

Identify correlations and patters by visualizing data

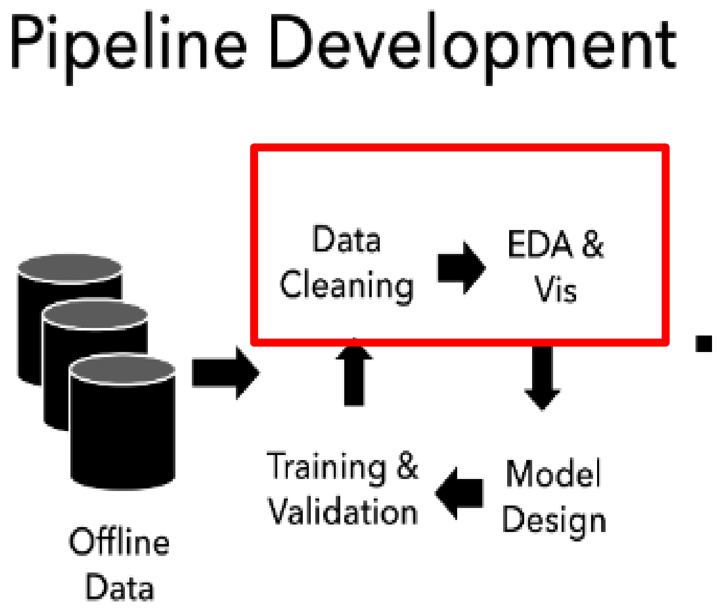
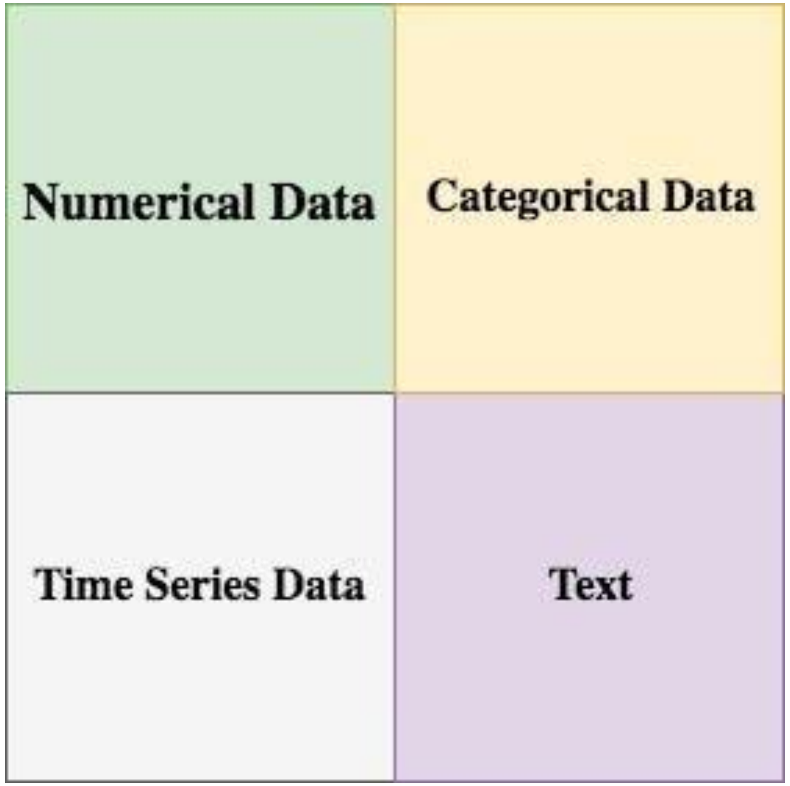


Image Source: Daniel Crankshaw (In: 'A Short History of Prediction-Serving Systems')

EDA – Understanding Data Types

Data can be in many forms.



Common Data Types in Machine Learning

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1007000 entries, 0 to 1006999
Data columns (total 11 columns):
 #   Column         Non-Null Count    Dtype
---  ------         --------------    -----
 0   water          1007000 non-null  Int64
 1   sur_refl_b01_1 1007000 non-null  Int64
 2   sur_refl_b02_1 1007000 non-null  Int64
 3   sur_refl_b03_1 1007000 non-null  Int64
 4   sur_refl_b04_1 1007000 non-null  Int64
 5   sur_refl_b05_1 1007000 non-null  Int64
 6   sur_refl_b06_1 1007000 non-null  Int64
 7   sur_refl_b07_1 1007000 non-null  Int64
 8   ndvi           1007000 non-null  Float32
 9   ndwi1          1007000 non-null  Float32
 10  ndwi2          1007000 non-null  Float32
dtypes: Float32(3), Int64(8)
memory usage: 91.2 MB
```

Pandas DataFrame Information from the MODIS Training Dataset

EDA – Data Cleaning and Handling Missing Values

- Missing values can pose a challenge for machine learning models, so it's important to understand the extent of missing data in the dataset.
- Different strategies such as imputation or deletion can be used to handle missing values depending on the amount and nature of missing data.

Detecting

Detecting Null Values:

- Isnull(): It is used as an alias for dataframe.isna(). This function returns the dataframe with boolean values indicating missing values.
- Syntax: dataframe.isnull()

Handling

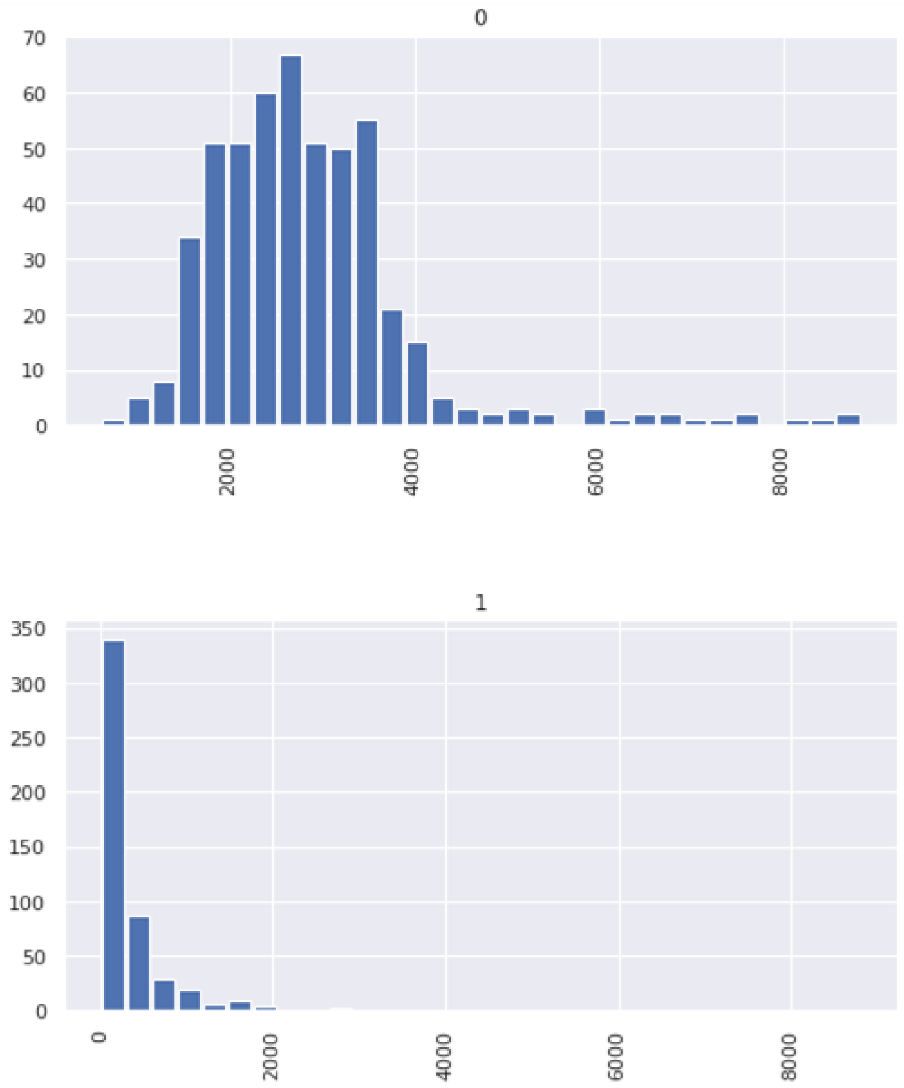
Handling Null Values:

- Dropping the Rows with Null Values: dropna() function is used to delete rows or columns with null values.
- Replacing Missing Values: fillna() function can fill the missing values with a special value like mean or median.

Image Source: Pace University, Exploratory Data Analysis

EDA – Visualizing Data Distributions

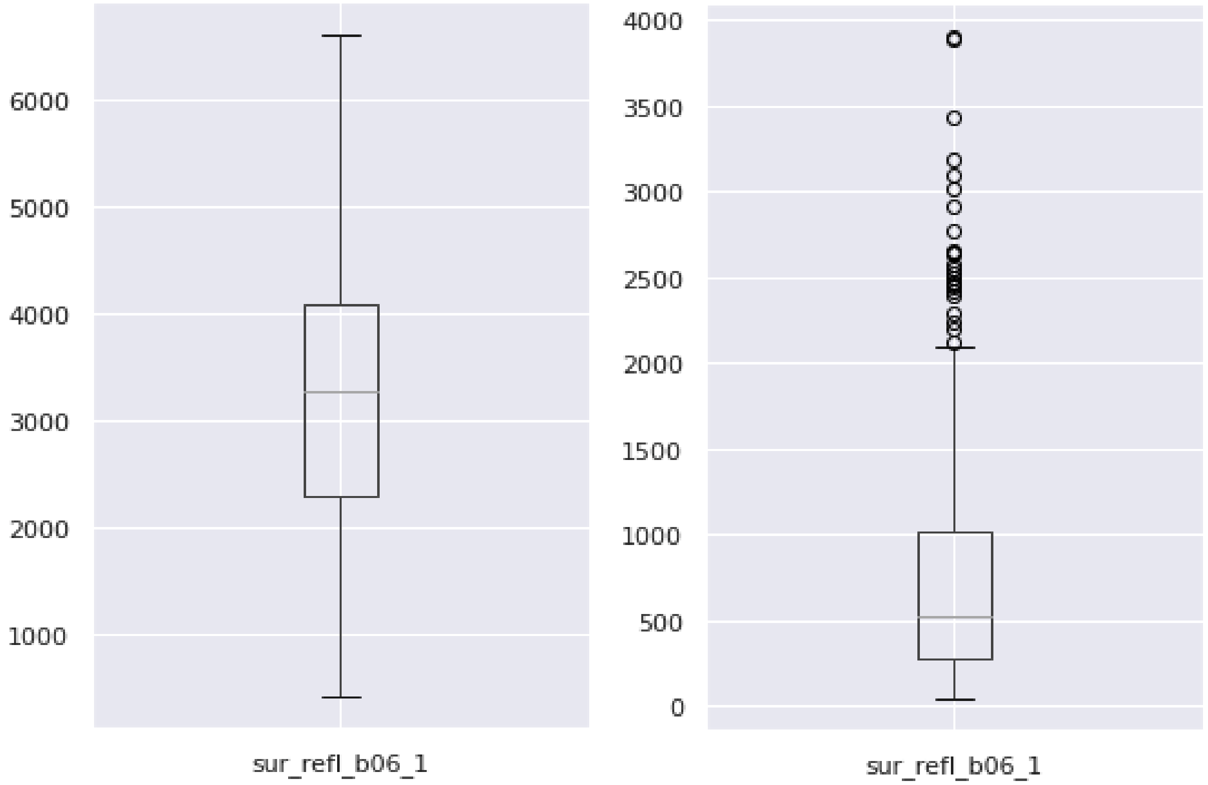
- One of the first steps in EDA is to understand the distribution of each variable in the dataset.
- Histograms, density plots, box plots, and violin plots are commonly used.
- Understanding data distributions can help identify outliers, skewness, and potential transformations that may be necessary.



Class Occurrence Across Spectral Response

EDA – Outlier Detection and Treatment

- Outliers can have an impact on the performance of ML models.
- Boxplots and scatterplots can be used to identify outliers.



Spectral Distribution of Two MODIS Surface Reflectance Bands

EDA – Investigating Correlations

- Correlations between variables can provide insights into the relationships and dependencies within the dataset.
- Identifying strong correlations can help with feature selection and can inform the choice of machine learning algorithms.

	water	sur_refl_b01_1	sur_refl_b02_1	sur_refl_b03_1	sur_refl_b04_1	sur_refl_b05_1	sur_refl_b06_1	sur_refl_b07_1	ndvi	ndwi1	ndwi2
water	1.000000	-0.257704	-0.569499	-0.170032	-0.220100	-0.714663	-0.740303	-0.303113	-0.490893	-0.348729	-0.282386
sur_refl_b01_1	-0.257704	1.000000	0.895279	0.977558	0.989731	0.467660	0.167929	0.101960	0.073173	0.453949	0.292324
sur_refl_b02_1	-0.569499	0.895279	1.000000	0.851850	0.880881	0.701883	0.425545	0.198049	0.299217	0.541591	0.387248
sur_refl_b03_1	-0.170032	0.977558	0.851850	1.000000	0.995246	0.392647	0.040463	0.053476	0.018081	0.396507	0.253936
sur_refl_b04_1	-0.220100	0.989731	0.880881	0.995246	1.000000	0.436087	0.099287	0.076514	0.055002	0.436115	0.283203
sur_refl_b05_1	-0.714663	0.467660	0.701883	0.392647	0.436087	1.000000	0.748920	0.372319	0.421938	0.379261	0.291508
sur_refl_b06_1	-0.740303	0.167929	0.425545	0.040463	0.099287	0.748920	1.000000	0.384666	0.420780	0.211944	0.169933
sur_refl_b07_1	-0.303113	0.101960	0.198049	0.053476	0.076514	0.372319	0.384666	1.000000	0.195150	0.126411	0.149680
ndvi	-0.490893	0.073173	0.299217	0.018081	0.055002	0.421938	0.420780	0.195150	1.000000	0.391189	0.317795
ndwi1	-0.348729	0.453949	0.541591	0.396507	0.436115	0.379261	0.211944	0.126411	0.391189	1.000000	0.485096
ndwi2	-0.282386	0.292324	0.387248	0.253936	0.283203	0.291508	0.169933	0.149680	0.317795	0.485096	1.000000

Correlation Coefficients Across MODIS Surface Reflectance Bands

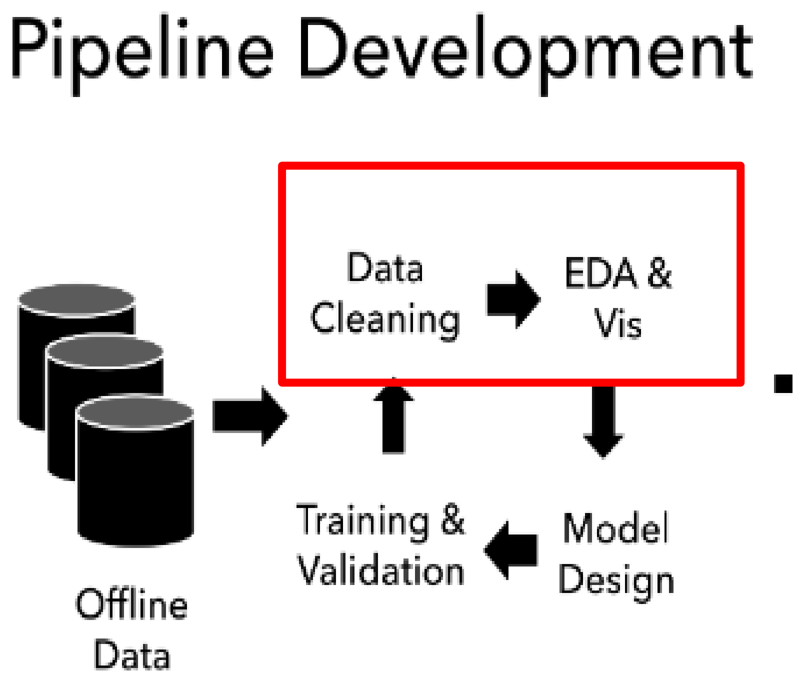
EDA – Investigating Correlations

	water	sur_refl_b01_1	sur_refl_b02_1	sur_refl_b03_1	sur_refl_b04_1	sur_refl_b05_1	sur_refl_b06_1	sur_refl_b07_1	ndvi	ndwi1	ndwi2
water	1.000000	-0.257704	-0.569499	-0.170032	-0.220100	-0.714663	-0.740303	-0.303113	-0.490893	-0.348729	-0.282386
sur_refl_b01_1	-0.257704	1.000000	0.895279	0.977558	0.989731	0.467660	0.167929	0.101960	0.073173	0.453949	0.292324
sur_refl_b02_1	-0.569499	0.895279	1.000000	0.851850	0.880881	0.701883	0.425545	0.198049	0.299217	0.541591	0.387248
sur_refl_b03_1	-0.170032	0.977558	0.851850	1.000000	0.995246	0.392647	0.040463	0.053476	0.018081	0.396507	0.253936
sur_refl_b04_1	-0.220100	0.989731	0.880881	0.995246	1.000000	0.436087	0.099287	0.076514	0.055002	0.436115	0.283203
sur_refl_b05_1	-0.714663	0.467660	0.701883	0.392647	0.436087	1.000000	0.748920	0.372319	0.421938	0.379261	0.291508
sur_refl_b06_1	-0.740303	0.167929	0.425545	0.040463	0.099287	0.748920	1.000000	0.384666	0.420780	0.211944	0.169933
sur_refl_b07_1	-0.303113	0.101960	0.198049	0.053476	0.076514	0.372319	0.384666	1.000000	0.195150	0.126411	0.149680
ndvi	-0.490893	0.073173	0.299217	0.018081	0.055002	0.421938	0.420780	0.195150	1.000000	0.391189	0.317795
ndwi1	-0.348729	0.453949	0.541591	0.396507	0.436115	0.379261	0.211944	0.126411	0.391189	1.000000	0.485096
ndwi2	-0.282386	0.292324	0.387248	0.253936	0.283203	0.291508	0.169933	0.149680	0.317795	0.485096	1.000000

Correlation Coefficients Across MODIS Surface Reflectance Bands

EDA for Machine Learning – Conclusion

- EDA is a critical step in any machine learning project to understand the structure and properties of the dataset.
- Visualizations and analysis techniques such as data distributions, correlations, missing value handling, and outlier detection can provide valuable insights for feature selection, preprocessing, and model selection.
- With these EDA techniques and tools, machine learning models can be developed with a better understanding of the underlying data.



Extracting Training Data From A Tabular Dataset

- We now understand and have cleaned our data. What's next?

Improve our understanding of the structure and properties of the dataset

Discover errors, missing values, and outliers in the dataset

Identify correlations and patterns by visualizing data

Sampling Data

Sometimes it is a struggle to gather enough data for an ML model to generalize.

Sometimes, however, this is not the problem in remote sensing.

Large amounts of data lead to questions.

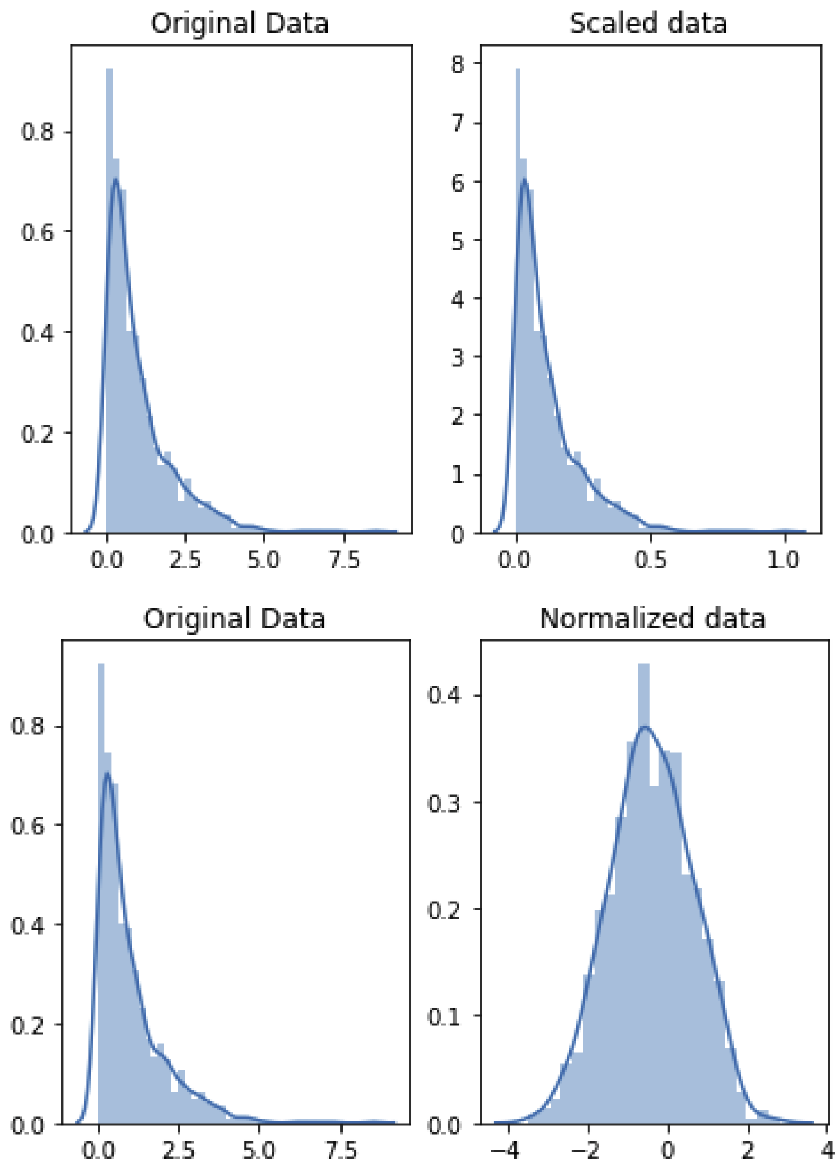
Questions to Ask:

- Too much data for an ML model to handle?
- What kind of sampling do we perform?
- At what granularity do we sample at?

Regularizing and Scaling Data

- Scaling and regularizing your data can help improve the model performance.

- Scaling and regularization should be done to data the model is applied to, not just the training data. Remember the golden rule: **the testing data should match the training data as closely as possible.**



Comparison of Original Data and Rescaled and Normalized Data

Handling Imbalanced Data

- Effective Ways of Handling Imbalanced Data:
  - Downsampling
  - Upweighting

- ML Algorithms that can Handle Weights on Samples:
  - Decision Trees
  - Random Forests
  - Gradient Boosting
  - Neural Networks

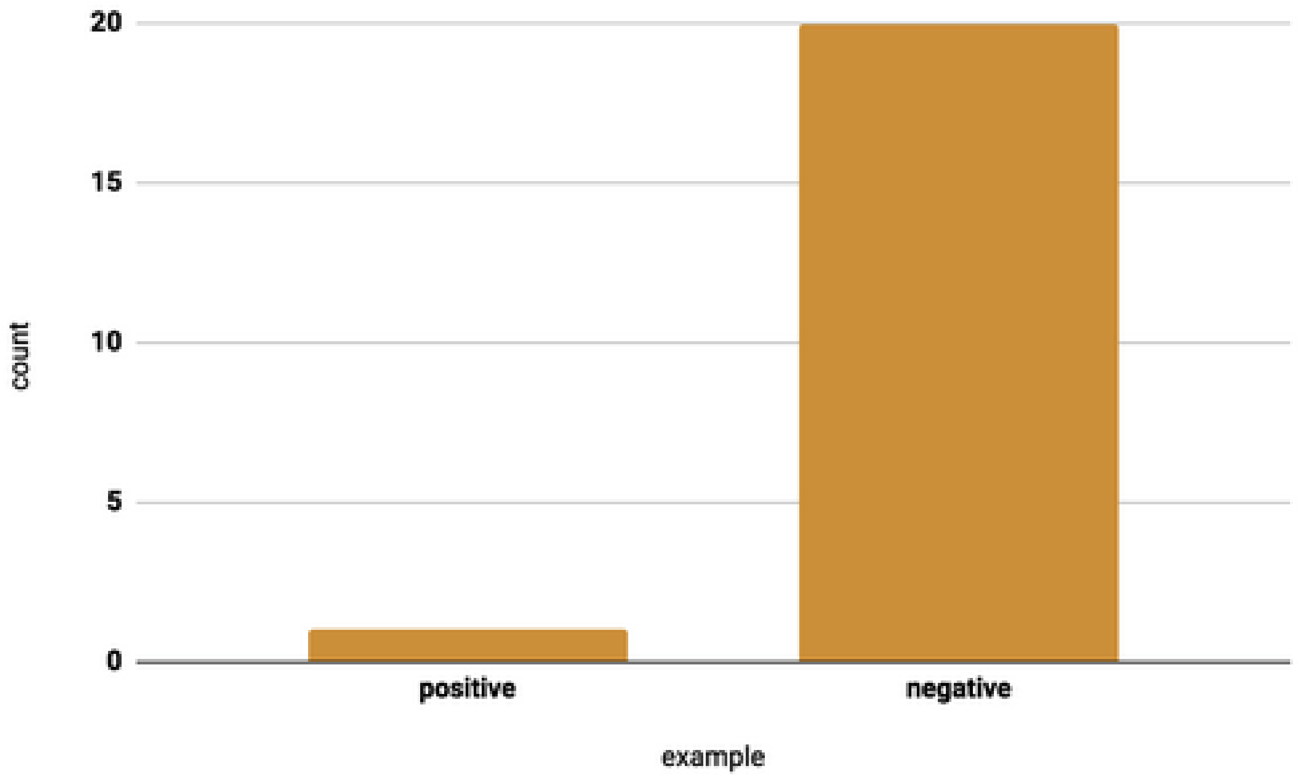


Image Source: developers.google.com

Splitting Data

Split your data into two subsets:

- Training Set: A subset to train a model
- Validation Set: A subset to evaluate performance during training
- Test Set: A subset to test the trained model

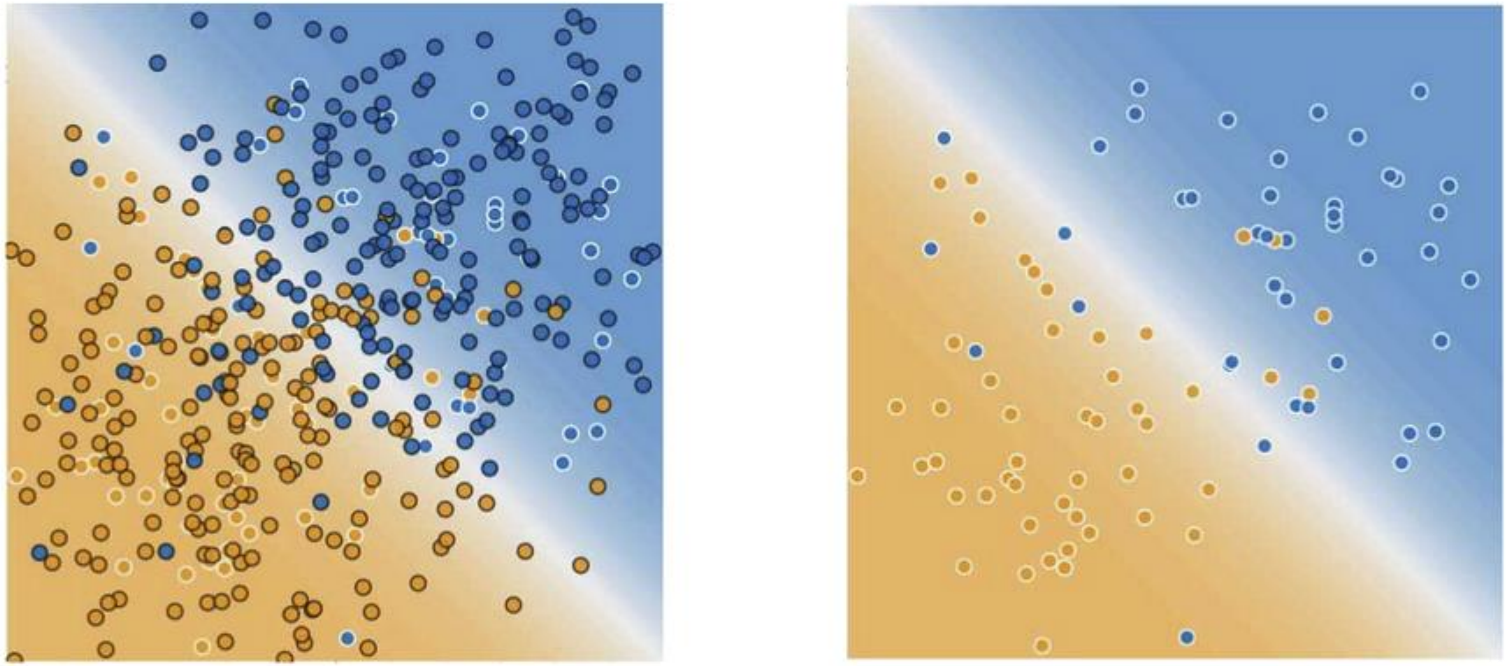
To design a split that is representative of your data, consider what the data represents. The golden rule applies to data splits as well: the testing task should match the production task as closely as possible.

General Ratio of Training, Validation, and Test Sets

Splitting Data

Make sure your test set meets the following conditions:

- Is large enough to yield statistically significant results
- Is representative of the dataset. Do not pick a set which contains characteristics which are different from the training dataset.



Training Data Test Data

Splitting Data – How It Aligns in the ML Workflow

1. Train the model on the training set
2. Use the validation set to evaluate results from training
3. Use the test set to confirm performance after the model has performed well enough on the validation set

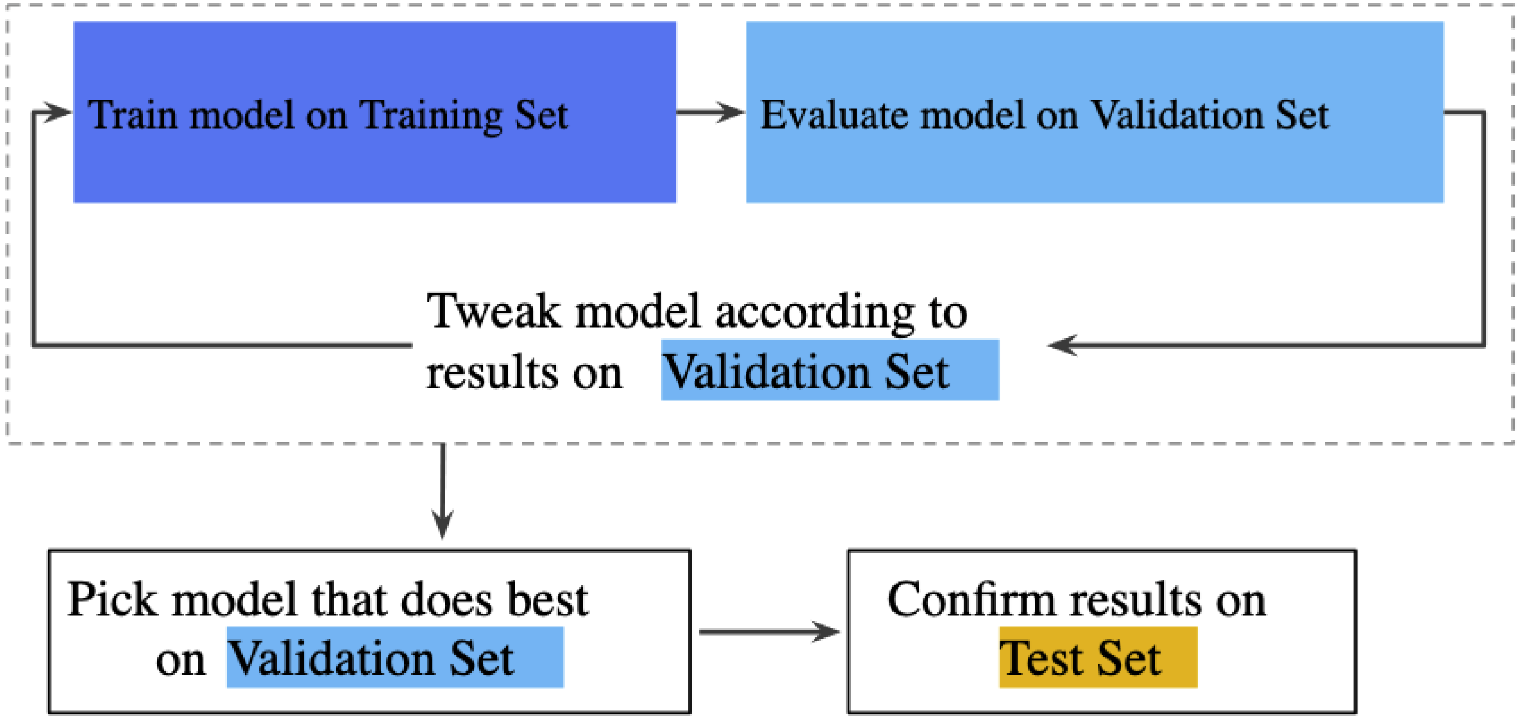
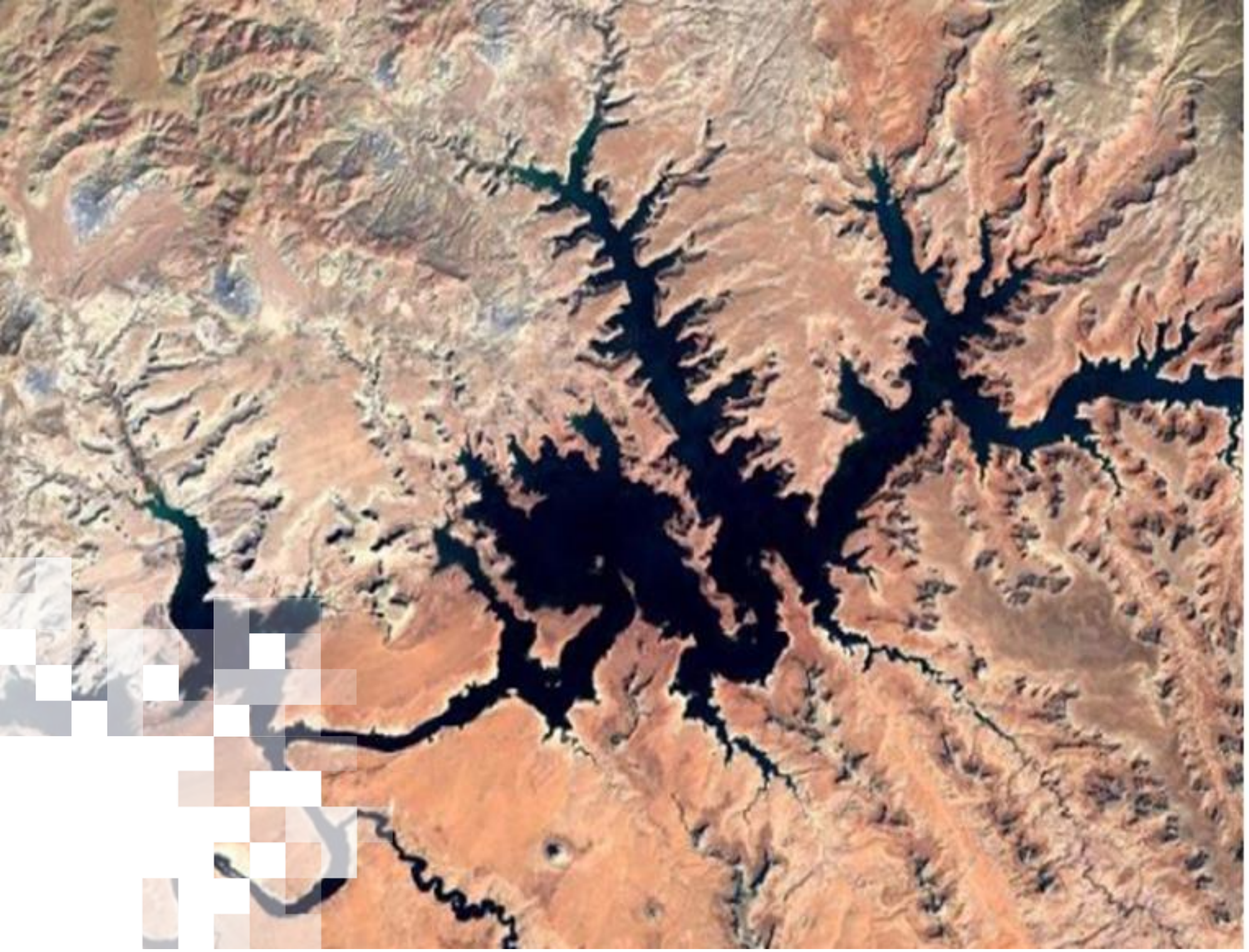
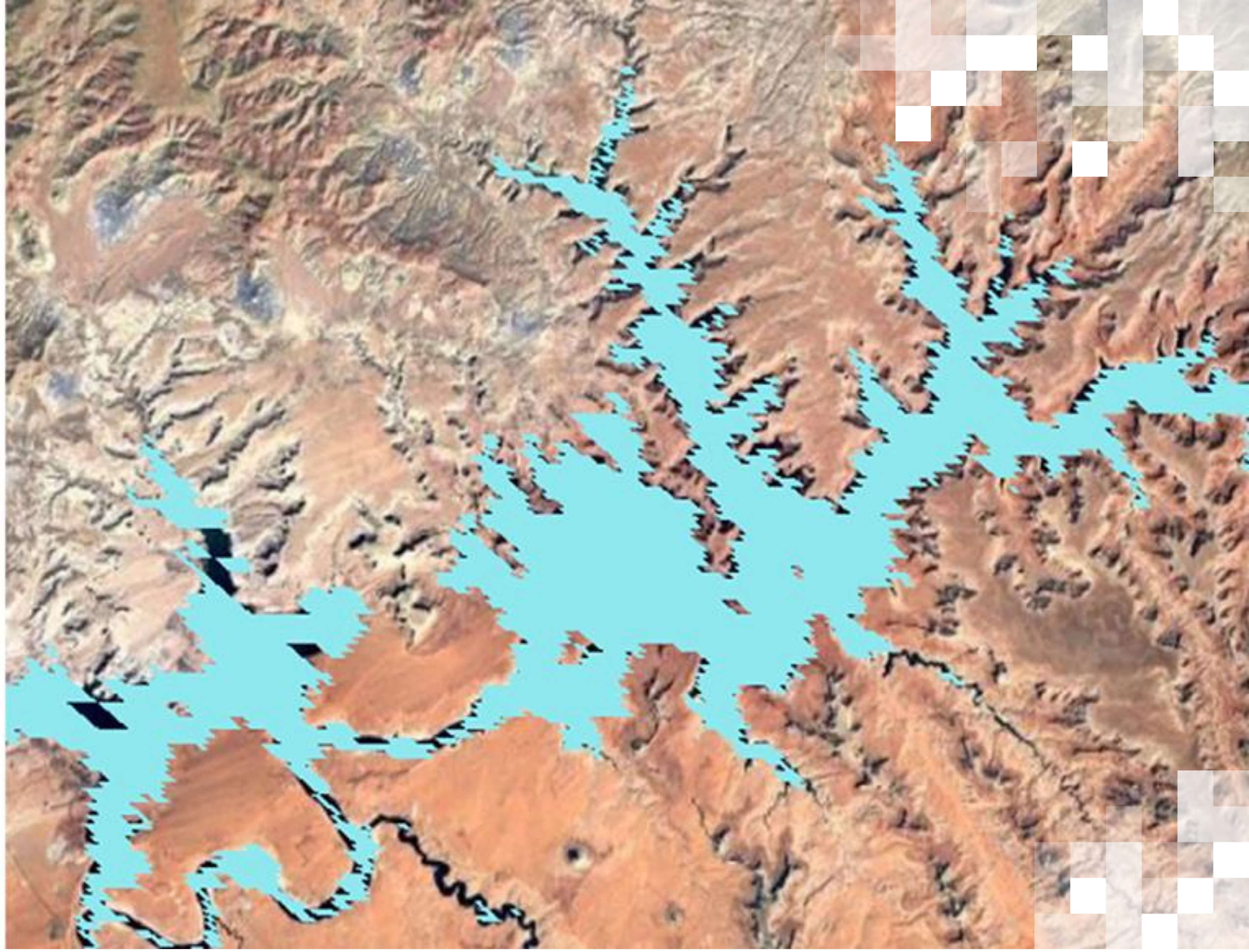


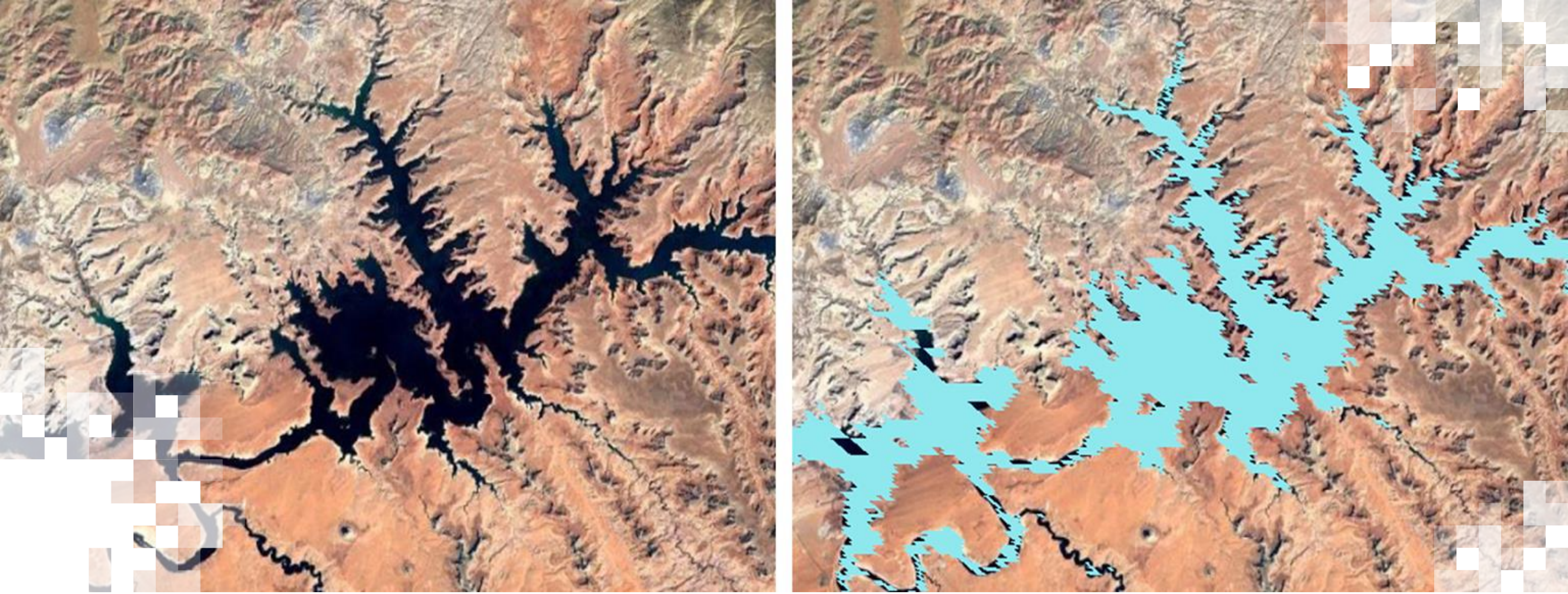
Image Source: developers.google.com





Exercise: Exploratory Data Analysis (EDA) in Google Colab

Trainer: Caleb S. Spradlin



Exercise: Training and Testing of Random Forest Model in Google Colab

Trainer: Jules Kouatchou

Summary

- Download the training data
- Exploratory data analysis
- Extracting training data from a tabular dataset
- Extracting training data from a raster dataset
- Training and inference of a tabular and raster dataset
- Metrics and model evaluation
- Hands on Jupyter Notebook Exercise: MODIS Water Classification Case Study

Looking Ahead

Part 3: Training Data and Land Cover Classification Example

- Overview of model tuning
- Overview of parameter optimization
- Exercise to optimize existing model
- Overview of model explainability and interpretability
- Overview of additional machine learning algorithms
- Hands on Jupyter Notebook Exercise: Improvements to MODIS Water Classification Model

Contacts

- Trainers:
  - Jordan A. Caraballo-Vega: jordan.a.caraballo-vega@nasa.gov
  - Jules Kouatchou: jules.kouatchou-1@nasa.gov
  - Caleb S. Spradlin: caleb.s.spradlin@nasa.gov
  - Jian Li: jian.li@nasa.gov
  - Brock Blevins: brock.Blevins@nasa.gov

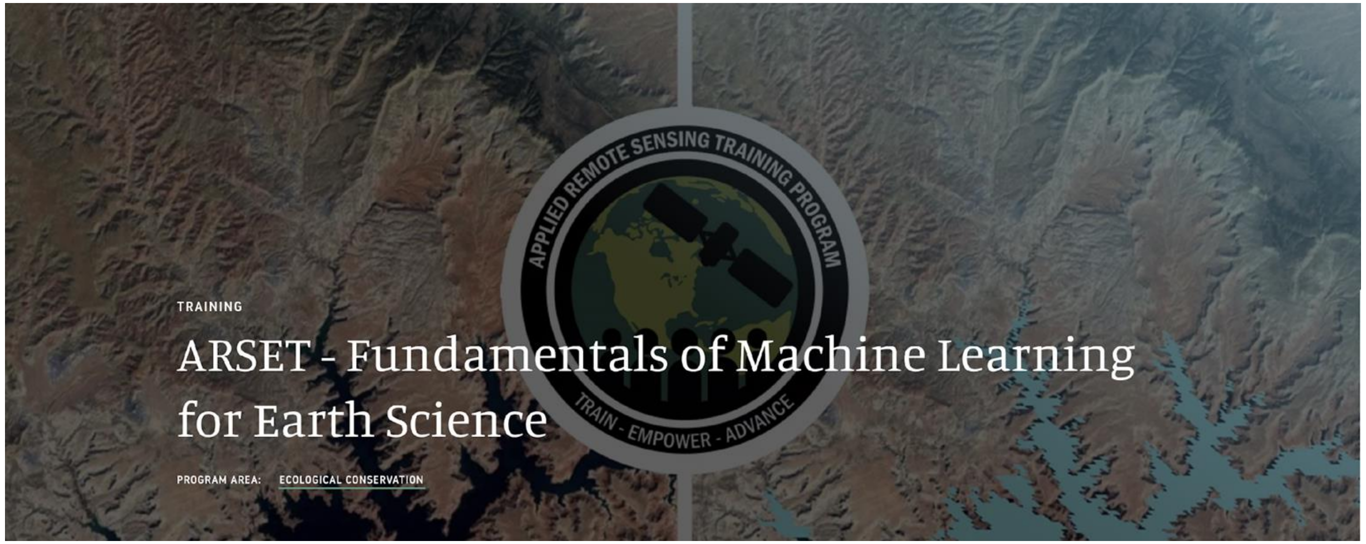
- Training Webpage:
  - https://appliedsciences.nasa.gov/join-mission/training/english/arset-fundamentals-machine-learning-earth-science

- ARSET Website:
  - https://appliedsciences.nasa.gov/arset

Check out our sister programs:

Questions?

- Please enter your question in the Q&A box. We will answer them in the order they were received.
- We will post the Q&A to the training website following the conclusion of the webinar.



References

- Crankshaw, D., & Gonzalez, J. (2018). Prediction-Serving Systems: What happens when we wish to actually deploy a machine learning model to production?. *Queue*, *16*(1), 83-97.
- Elders, A., Carroll, M. L., Neigh, C. S., D'Agostino, A. L., Ksoll, C., Wooten, M. R., & Brown, M. E. (2022). Estimating crop type and yield of small holder fields in Burkina Faso using multi-day Sentinel-2. *Remote Sensing Applications: Society and Environment*, *27*, 100820.
- Fleming, S. W., Watson, J. R., Ellenson, A., Cannon, A. J., & Vesselinov, V. C. (2021). Machine learning in Earth and environmental science requires education and research policy reforms. *Nature Geoscience*, *14*(12), 878-880.
- Prša, A., Kochoska, A., Conroy, K. E., Eisner, N., Hey, D. R., IJspeert, L., ... & Winn, J. N. (2022). TESS Eclipsing Binary Stars. I. Short-cadence Observations of 4584 Eclipsing Binaries in Sectors 1–26. *The Astrophysical Journal Supplement Series*, *258*(1), 16.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. the National Energy Research Supercomputing Center in Lawrence Berkeley National Laboratory, Berkeley, CA, USA: Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*, 195-204.
- Yu, S., & Ma, J. (2021). Deep learning for geophysics: Current and future trends. *Reviews of Geophysics*, *59*(3), e2021RG000742.

Contributors

- Jordan A. Caraballo-Vega
- Mark L. Carroll
- Jules R. Kouatchou
- Jian Li
- Caleb S. Spradlin
- Brock Blevins
- Melanie Follette-Cook
- Erika Podest
- Brian Powell
- Akiko Elders

Thank You!