MLHub and AI/ML in CMR

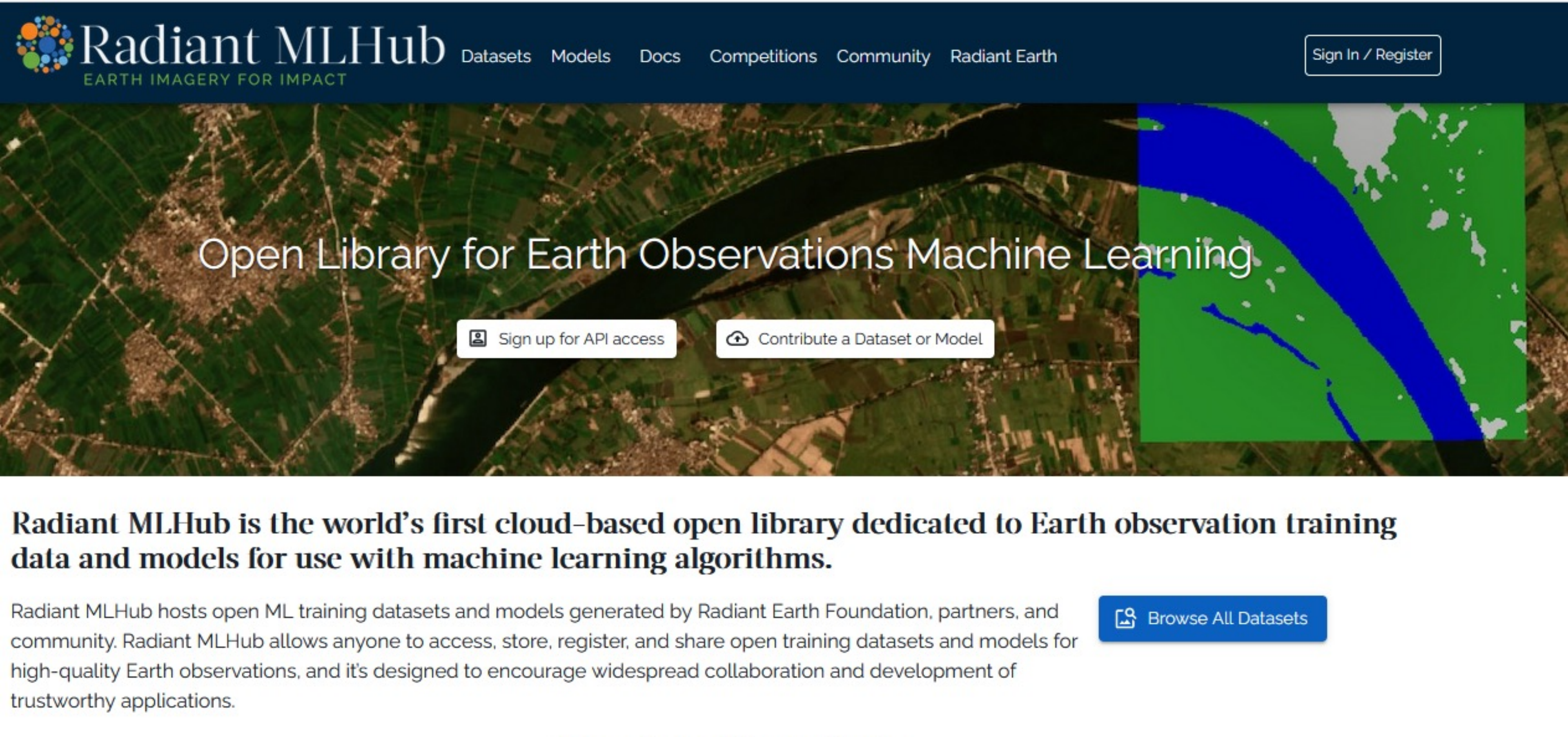
CEOS WGISS-55
Córdoba, Argentina
April 19, 2023
Hosted by CONAE

Michael Morahan Valerie Dixon
NASA ESDIS Project
Goddard Space Flight Center
Valerie.dixon@nasa.gov
Michael.P.Morahan@nasa.gov
Scott.A.Ritz@nasa.gov

Radiant MLHub

(https://mlhub.earth/)

- Initially a NASA Research Opportunities in Space and Earth Sciences (ROSES) project, Radiant MLHub is a library dedicated to open Earth observation training data for use with machine learning algorithms.



Radiant MLHub is the world's first cloud-based open library dedicated to Earth observation training data and models for use with machine learning algorithms.

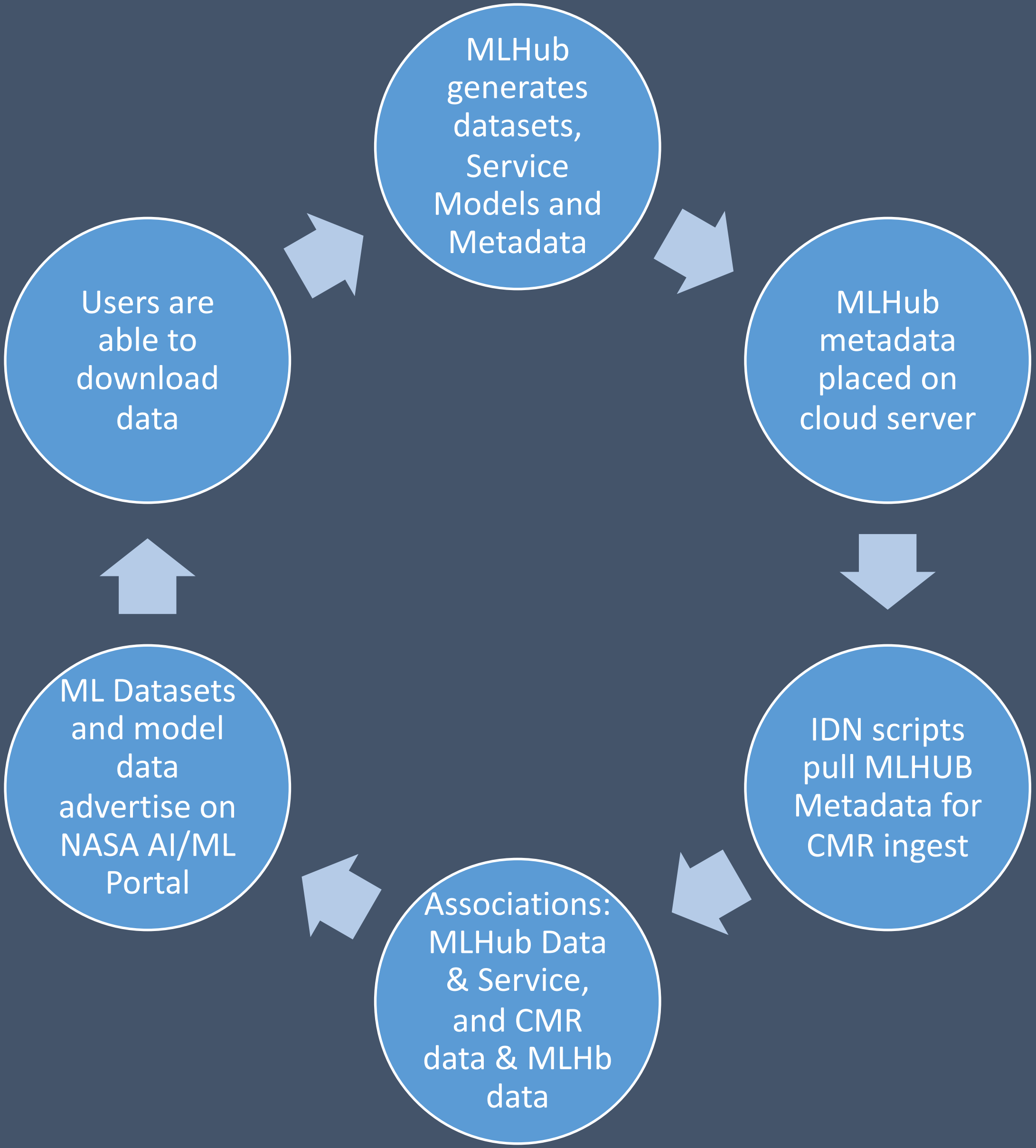
Radiant MLHub hosts open ML training datasets and models generated by Radiant Earth Foundation, partners, and community. Radiant MLHub allows anyone to access, store, register, and share open training datasets and models for high-quality Earth observations, and it's designed to encourage widespread collaboration and development of trustworthy applications.

Radiant MLHub Details

- The MLHub facilitates an open community commons for geospatial training data, machine learning models, and standards to encourage collaboration and share information
  - Python Client
    - Allows users to search and download geospatial training data on Radiant MLHub without managing API requests.
    - Users can apply MLHub with other scripting languages using our REST API.
    - https://mlhub.earth/docs
  - SpatioTemporal Asset Catalog (STAC)
    - All Radiant MLHub geospatial training data collections are stored using STAC-compliant catalogs and are exposed through a common API.
    - Radiant Earth is developing the STAC ML Model Extension to the which will empower users to discover and access existing repositories of ML models for various geospatial applications.
    - https://mlhub.earth/models/about
  - Current Contents
    - 65 datasets: Agriculture, Cloud, Crop Type, Flood Detection, Land Cover, Moisture, Marine Debris, Plantscope, Tropical Storm, and Wildfire.
    - 6 services: Crop Classification, Crop Detection, Tropical Cyclone Wind Estimation, and Replicable AI for Microplanning

MLHub and CMR Connection

Still in Progress



MLHub generates datasets, Service Models and Metadata

MLHub metadata placed on cloud server

IDN scripts pull MLHUB Metadata for CMR ingest

Associations: MLHub Data & Service, and CMR data & MLHb data

ML Datasets and model data advertise on NASA AI/ML Portal

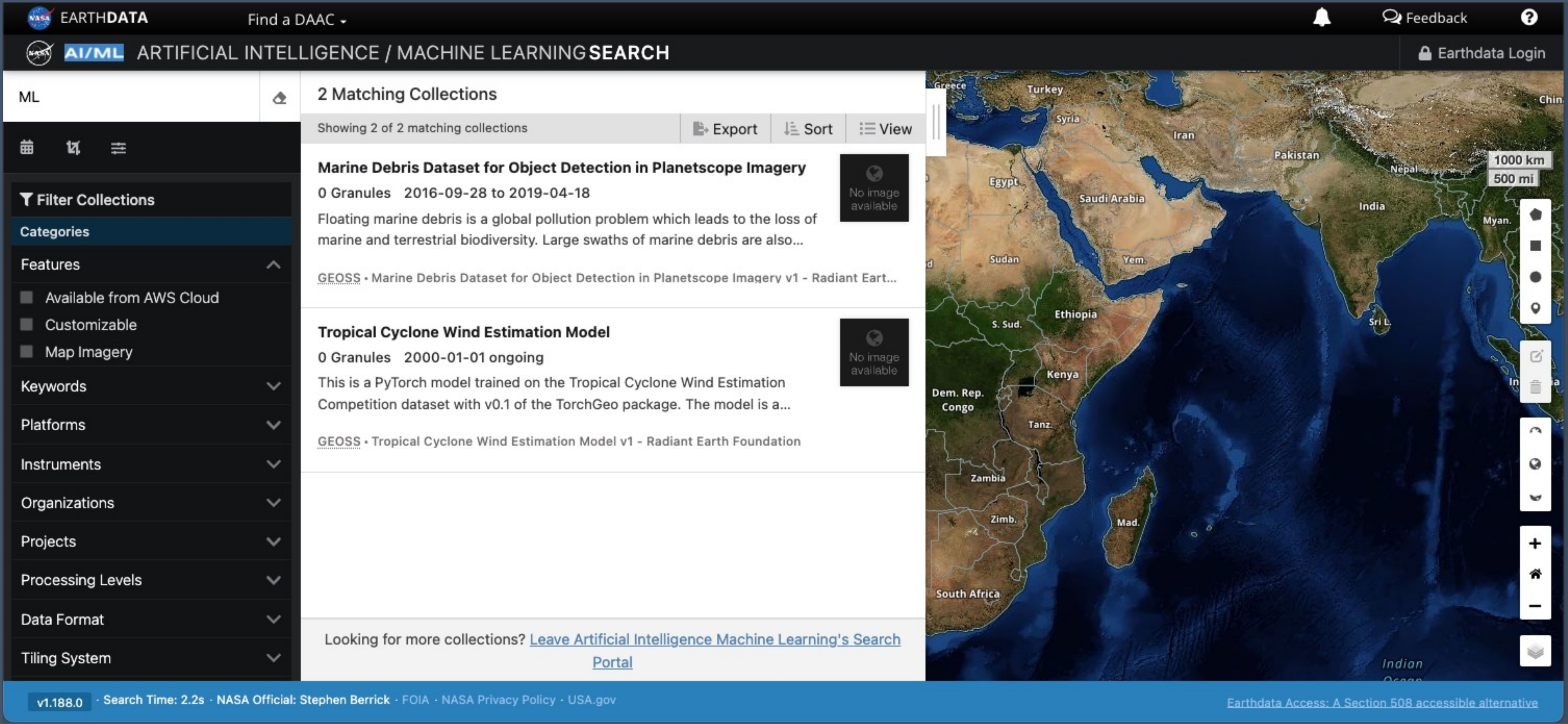
Users are able to download data

What we have in place today:

- MLHub
  - Ready to go!
- Two pathfinder ML records in CMR
  - A training data record, and model record
- An AI/ML Earthdata Search Portal
  - Or you can filter Earthdata Search by:
    - Organization: Radiant Earth Foundation
    - Apply: Include Collections without granules
- GCMD Keywords for AI/ML
  - Discussed at WGISS-54 meeting

Earthdata AI/ML Search Portal

https://search.earthdata.nasa.gov/portal/ai-ml/search



ML

2 Matching Collections

Showing 2 of 2 matching collections

Marine Debris Dataset for Object Detection in Planetscope Imagery

0 Granules 2016-09-28 to 2019-04-18

Floating marine debris is a global pollution problem which leads to the loss of marine and terrestrial biodiversity. Large swaths of marine debris are also...

GEOSS • Marine Debris Dataset for Object Detection in Planetscope Imagery v1 - Radiant Eart...

Tropical Cyclone Wind Estimation Model

0 Granules 2000-01-01 ongoing

This is a PyTorch model trained on the Tropical Cyclone Wind Estimation Competition dataset with v0.1 of the TorchGeo package. The model is a...

GEOSS • Tropical Cyclone Wind Estimation Model v1 - Radiant Earth Foundation

Looking for more collections? Leave Artificial Intelligence Machine Learning's Search Portal

v1.188.0 · Search Time: 2.2s · NASA Official: Stephen Berrick · FOIA · NASA Privacy Policy · USA.gov

Earthdata Access: A Section 508 accessible alternative

GCMD Keyword Categories

- Earth Science
- Earth Science Services
- Platforms
- Instruments
- Providers
- Projects
- Locations

DATA ANALYSIS AND VISUALIZATION
DATA MANAGEMENT/DATA HANDLING
EDUCATION/OUTREACH
ENVIRONMENTAL ADVISORIES
HAZARDS MANAGEMENT
MACHINE LEARNING TRAINING DATA
METADATA HANDLING
MODELS
MACHINE LEARNING MODELS
REFERENCE AND INFORMATION SERVICES
WEB SERVICES

What we still have to work on

- Automating ingest of metadata from MLHub to CMR
  - Coming soon!
- Associating CMR records and ML Training Data as Related Collections
  - To enable cross-discovery
- ML Model metadata schema – Under Analysis
  - Pathfinder Model record is a Collection, but prevailing thought is that Training Data are Collections and Models should be Services… or maybe a new schema type?
- Discovery by Service feature in Earthdata Search (and its portals)

Thank you!

If you have any questions or suggestions, please reach out:

valerie.dixon@nasa.gov

michael.p.morahan@nasa.gov

scott.a.ritz@nasa.gov

or

Earthdata Forum: GCMD Keywords

Acronyms

- AI/ML = Artificial Intelligence/Machine Learning
- CEOS = Committee on Earth Observing Satellites
- CMR = Common Metadata Repository
- EED = EOSDIS Evolution and Development
- e.g. = for example
- EOSDIS = Earth Observing System Data and Information System
- ESDIS = Earth Science Data and Information Systems
- etc. = Etcetera
- GCMD = Global Change Master Directory
- GSFC = Goddard Space Flight Center
- NASA = National Aeronautics and Space Administration
- TBD = To Be Determined
- WGISS = Working Group on Information Systems and Services
- CONAE = Comisión Nacional de Actividades Espaciales (Argentina)

Backup Slides

Earth Science Services

- **MACHINE LEARNING TRAINING DATA**

  The input data necessary for running a machine learning model.

  - **LABELS**

    Target variables in a machine learning workflow and part of the training dataset.

    - **RASTER LABEL**

      Masks pixels in raster data in order to identify a data feature or attribute in a machine learning workflow.

    - **VECTOR LABEL**

      Created with a point, line, or polygon to identify a data feature or attribute in a machine learning workflow.

  - **SOURCE**

    The data used in reference in order to create label annotations in a machine learning workflow.

    - **RASTER SOURCE**

      Data with gridded representation, where each pixel value represents information in a two-dimensional matrix.

    - **VECTOR SOURCE**

      Data represented by points, lines, or polygons representing a phenomenon or series of phenomena.

Earth Science Services / Models

- **MACHINE LEARNING MODELS**

A predictive model that when trained on a set of data containing certain features, enables a computer to identify similar features in other data.

  - **CLASSIFICATION**

  ML model type that sorts data into classes.

  - **CLUSTERING**

  ML model type that divides data into groups (aka clusters) without having a label for them.

  - **DECISION TREE**

  A type of supervised learning that uses a predictive modeling approach to ask additional questions of the data based on the answer to earlier questions.

    - **ISOLATION FOREST**

    An unsupervised ML method that is used for anomaly detection.

Earth Science Services / Models

- **MACHINE LEARNING MODELS, cont’d**
  - **DEEP LEARNING**

    ML and AI that imitates the way humans gain certain types of knowledge.

    - **CONVOLUTIONAL NEURAL NETWORKS**

      A Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.

    - **GENERATIVE ADVERSARIAL NETWORKS**

      A type of unsupervised learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.

    - **RECURRENT NEURAL NETWORKS**

      A type of artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, natural language processing (NLP), speech recognition, and image captioning.

Earth Science Services / Models

- **MACHINE LEARNING MODELS, cont'd**
  - **ENSEMBLE MODELS**

    A modeling process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets.
    - **BOOSTING**

      An ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors.
    - **RANDOM FOREST**

      A type of supervised learning that uses multiple decision trees for a computer to find patterns in data.
  - **NATURAL LANGUAGE PROCESSING**

    A type of learning that utilizes text-based sources to analyze parts of speech, sentiment, and term frequency.
  - **NEURAL NETWORKS**

    ML model type that is part of deep learning algorithms that mimic the operations of a human brain to recognize relationships between vast amounts of data.

Earth Science Services / Models

- **MACHINE LEARNING MODELS, cont'd**
  - **OBJECT DETECTION**

    ML model type that helps to identify a distinct object in data.
  - **REGRESSION**

    ML model type that translates input data of N-dimension to one or more scalar values.
  - **SEGMENTATION**

    ML model type that clusters part of the data (particularly image data) to groups that belong to the same class.
  - **SELF-SUPERVISED**

    ML model type that uses context in the available sample data to predict missing or nearby data.
  - **SEMI-SUPERVISED**

    ML model type that uses both supervised and unsupervised learning in its approach. Semi-supervised techniques take advantage of both labelled and unlabeled data.

Earth Science Services / Models

- **MACHINE LEARNING MODELS, cont'd**
    - **SUPERVISED**

      ML model type that utilizes labels to train the model.
    - **UNSUPERVISED**

      ML model type that looks for patterns in data.

This work was supported by NASA/GSFC under Raytheon Technologies contract number 80GSFC21CA001.