National Aeronautics and Space Administration

# Geo-Croissant
# Croissant for Geospatial Data

Manil Maskey, Rajat Shinde, Iksha Gurung

Presenter - Rajat Shinde, Ph.D.

*NASA Science Mission Directorate*
*NASA IMPACT*
*ML-Commons Croissant Working Group*

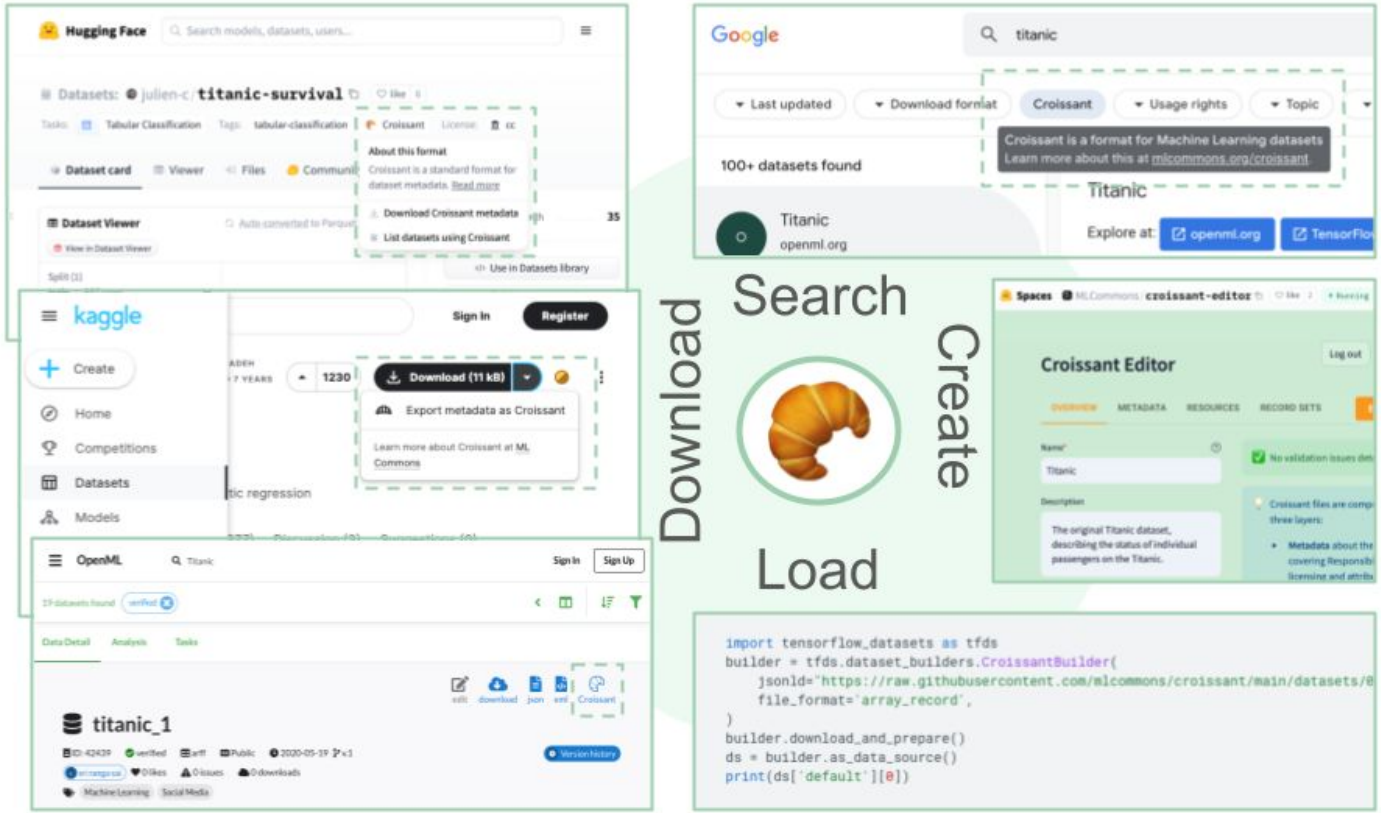
June 04, 2024

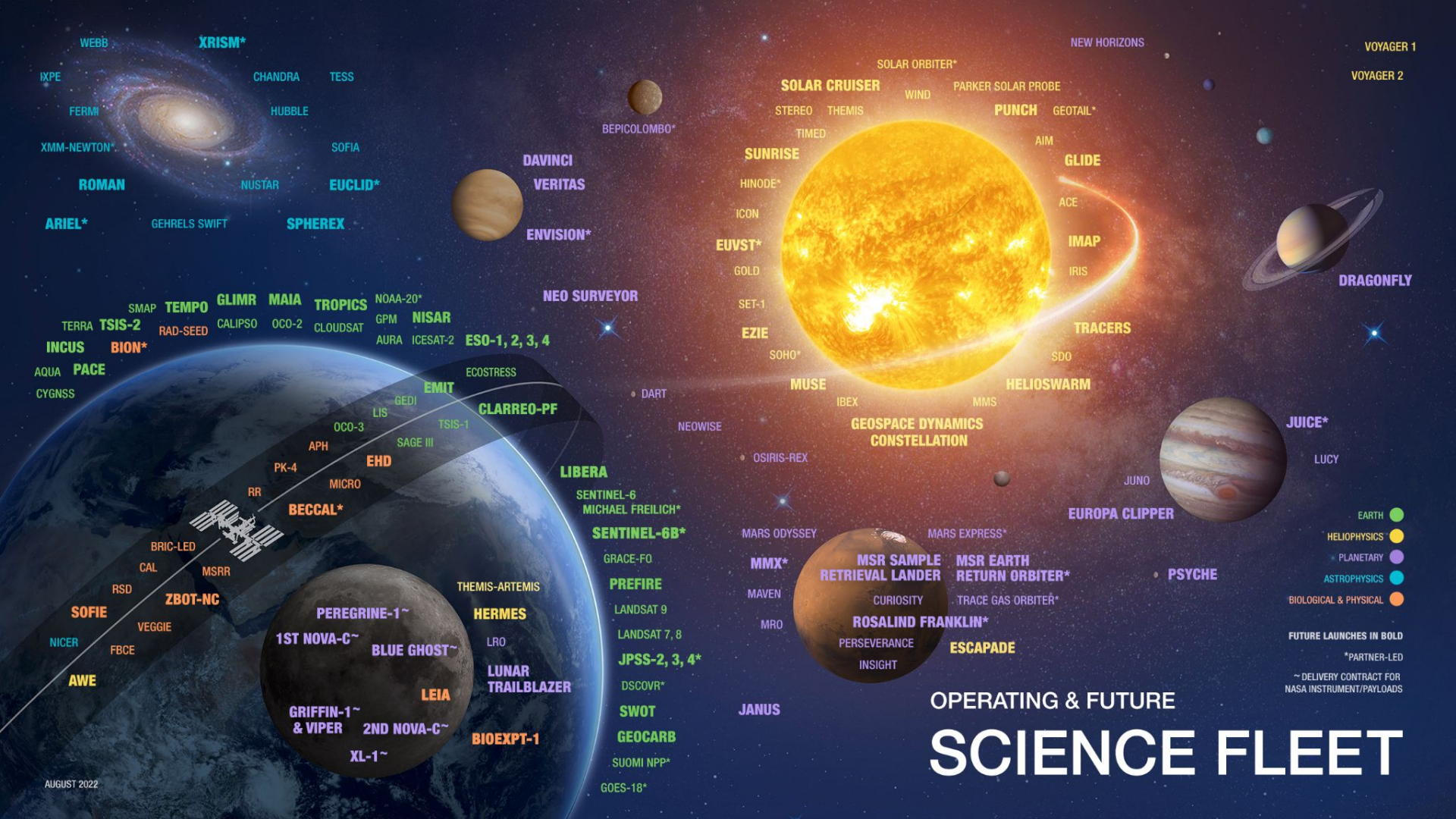
# Croissant - A Standardized Metadata Format


Croissant is for:

- **Creators and maintainers of ML datasets** – data work is tedious and often under-appreciated. Croissant makes datasets more widely available, across repositories and ML frameworks. Croissant is designed to be modular and extensible – new vocabulary extensions are encouraged to address the distinct characteristics of datasets of certain modalities (e.g. audio, video) or in certain sectors (e.g. life sciences, geospatial).
- **ML researchers and practitioners** – users of Croissant-enabled datasets have access to dataset documentation to understand how to make the most of the data and contribute to it. They can find the data they need no matter where it was published online. They can load the data into different ML platforms without any overhead to transform the data from one format to another.
- **RAI researchers and practitioners** – Croissant offers a machine-readable summary of important attributes captured in a variety of data cards and similar approaches, which is portable and discoverable no matter where the dataset and its data card live, hence promoting better documentation practices.
- **Policy makers** – as AI regulation emerges across the world, Croissant provides a standardized way to collect core information about datasets, hence facilitating the development of data-centric AI audit and assurance tools such as transparency indexes.

# Croissant - Features




WEBB
XRISM*
IXPE
CHANDRA
TESS
FERMI
HUBBLE
XMM-NEWTON*
SOFIA
ROMAN
NUSTAR
EUCLID*
ARIEL*
GEHRELS SWIFT
SPHEREX
BEPICOLOMBO*
DAVINCI
VERITAS
ENVISION*
NEO SURVEYOR
SMAP
TEMPO
GLIMR
MAIA
TROPICS
NOAA-20*
TERRA
TSIS-2
RAD-SEED
CALIPSO
OCO-2
CLOUDSAT
GPM
NISAR
INCUS
BION*
AURA
ICESAT-2
ESO-1, 2, 3, 4
AQUA
PACE
ECOSTRESS
CYGNSS
EMIT
GEDI
LIS
CLARREO-PF
OCO-3
TSIS-1
APH
SAGE III
PK-4
EHD
RR
MICRO
BECCAL*
BRIC-LED
CAL
MSRR
RSD
ZBOT-NC
SOFIE
VEGGIE
NICER
FBCE
AWE
PEREGRINE-1~
1ST NOVA-C~
BLUE GHOST~
GRIFFIN-1~
& VIPER
2ND NOVA-C~
XL-1~
THEMIS-ARTEMIS
HERMES
LRO
LUNAR
TRAILBLAZER
LEIA
BIOEXPT-1
AUGUST 2022
LIBERA
SENTINEL-6
MICHAEL FREILICH*
SENTINEL-6B*
GRACE-FO
PREFIRE
LANDSAT 9
LANDSAT 7, 8
JPSS-2, 3, 4*
DSCOVR*
SWOT
GEOCARB
SUOMI NPP*
GOES-18*
NEOWISE
DART
OSIRIS-REX
JANUS
SOLAR ORBITER*
SOLAR CRUISER
WIND
PARKER SOLAR PROBE
STEREO
THEMIS
PUNCH
GEOTAIL*
TIMED
AIM
SUNRISE
GLIDE
HINODE*
ACE
ICON
EUVST*
IMAP
GOLD
IRIS
SET-1
EZIE
TRACERS
SOHO*
SDO
MUSE
HELIOSWARM
IBEX
MMS
GEOSPACE DYNAMICS
CONSTELLATION
NEW HORIZONS
VOYAGER 1
VOYAGER 2
DRAGONFLY
JUICE*
LUCY
JUNO
EUROPA CLIPPER
MARS ODYSSEY
MARS EXPRESS*
MMX*
MSR SAMPLE
RETRIEVAL LANDER
MSR EARTH
RETURN ORBITER*
MAVEN
CURIOSITY
TRACE GAS ORBITER*
MRO
ROSALIND FRANKLIN*
PERSEVERANCE
INSIGHT
ESCAPADE
PSYCHE
EARTH
HELIOPHYSICS
PLANETARY
ASTROPHYSICS
BIOLOGICAL & PHYSICAL
FUTURE LAUNCHES IN BOLD
*PARTNER-LED
~ DELIVERY CONTRACT FOR
NASA INSTRUMENT/PAYLOADS
OPERATING & FUTURE
SCIENCE FLEET

# NASA EARTH FLEET

## OPERATING & FUTURE THROUGH 2023

SWOT (CNES)
LANDSAT-9 (USGS)
SENTINEL-6 Michael Freilich/B (ESA)
TROPICS (6)
GEOCARB
NISAR (ISRO)
MAIA
TSIS-2
TEMPO
PREFIRE (2)
PACE (NSO)
GLIMR
ICESAT-2
GRACE-FO (2) (DLR)
CYGNSS (8)
NISTAR, EPIC (DSCOVR/NOAA)
CLOUDSAT (CSA)
TERRA (JAXA, CSA)
AQUA (JAXA, AEB)
AURA (NSO, FMI, UKSA)
CALIPSO (CNES)
GPM (JAXA)
LANDSAT 7 (USGS)
LANDSAT 8 (USGS)
OCO-2
SMAP
SUOMI NPP (NOAA) (JAXA)

### ISS INSTRUMENTS

EMIT
CLARREO-PF
GEDI
SAGE III
OCO-3
TSIS-1
ECOSTRESS
LIS

### JPSS-2, 3 & 4 INSTRUMENTS

OMPS-Limb
LIBERA

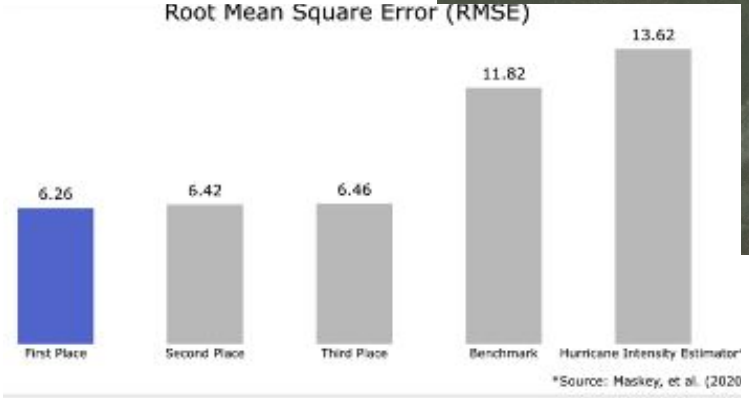
03.24.20

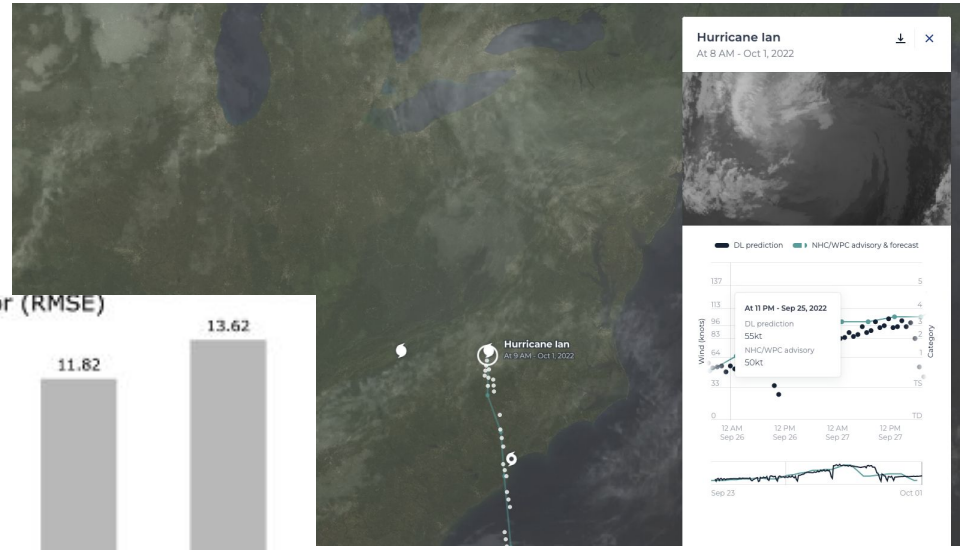
### INVEST/CUBESATS

RainCube
CSIM-FD
CubeRRT
TEMPEST-D
CIRiS
HARP
CTIM
HyTI
SNoOPI
NACHOS

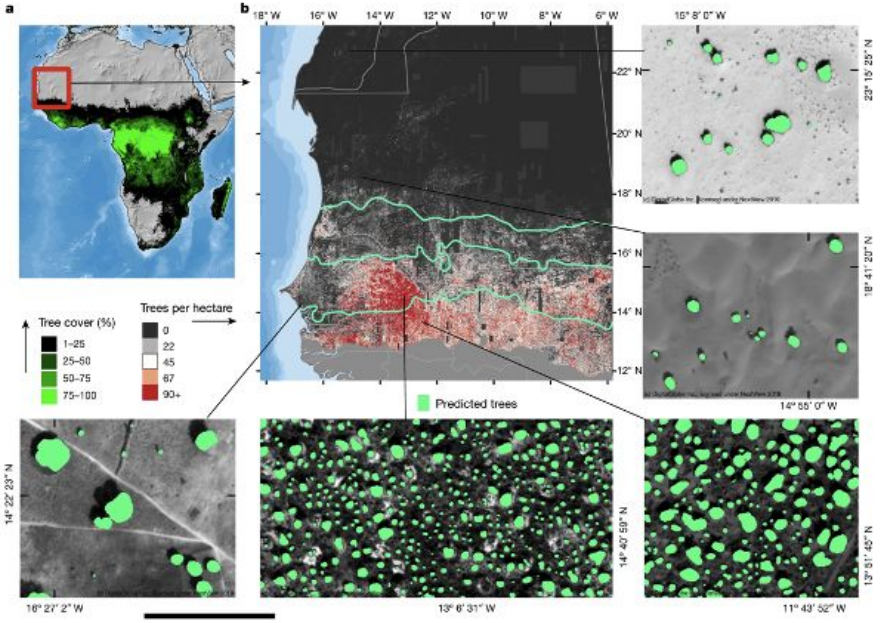
(PRE) FORMULATION
IMPLEMENTATON
PRIMARY OPS
EXTENDED OPS

# Estimation hurricane intensity using AI in real time

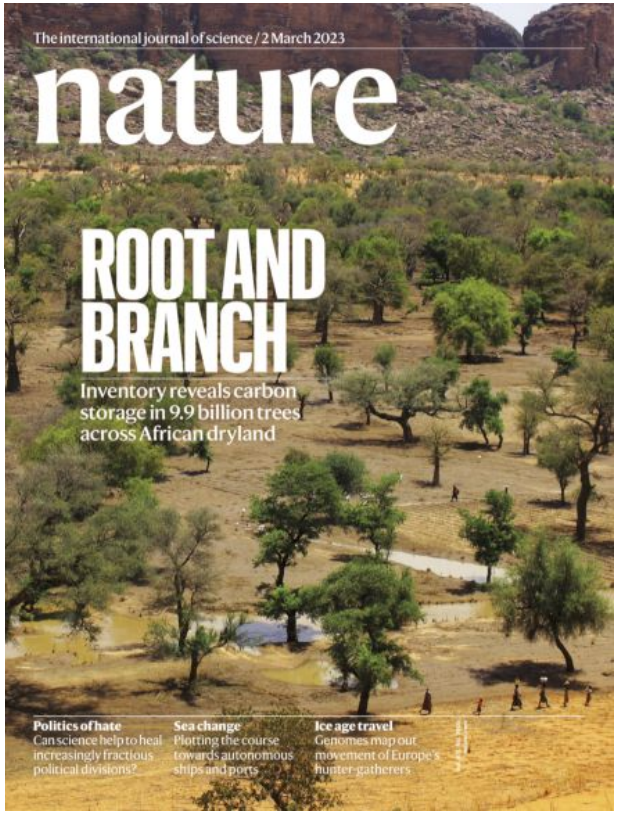


# Mapping 9.9 billion trees across Africa




Brandt et al.



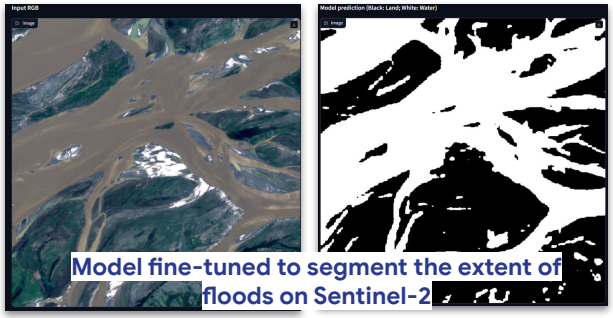
# AI-ready Data: Geospatial Foundation Model

Collaboration with IBM Research and NASA IMPACT project at MSFC

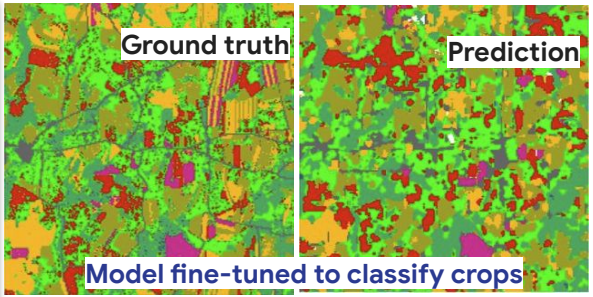
Pretrained on NASA Harmonized Landsat Sentinel-2 dataset

**Available at Hugging Face including Models, Datasets, and Code.**

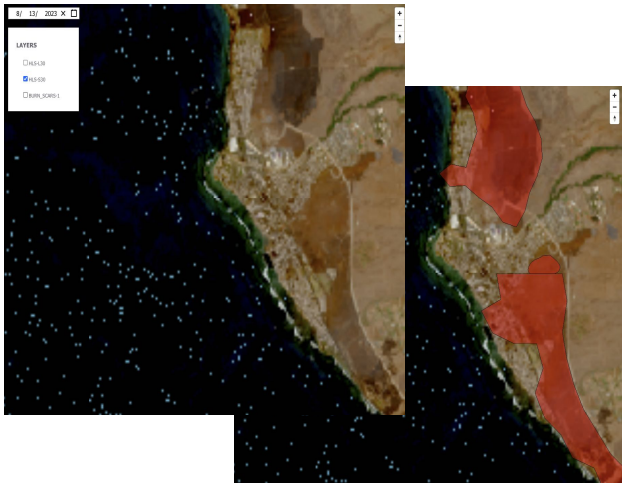
## Downstream applications



Model fine-tuned to segment the extent of floods on Sentinel-2




Model fine-tuned to classify crops



Burn scar mapping: Maui Fire 8.13.2023

# ML for EO Marketplace (2021) (*Source: Radiant Earth*)

# What’s Missing?

Benchmarks

Validation

Reuse and Reproducibility

Standard-based tooling

Privacy

# Complexities with Geospatial Datasets

- High dimensional: multi/hyperspectral
- Spatial and temporal dependencies
- Large data volume
- Data quality
  - Noisy: atmospheric conditions, calibration, etc..
- Heterogeneity
  - Combination of different types of datasets, Satellite data, In-situ, etc.
- Scale variability:
  - Phenomena in geospatial data can occur at local to global, last from hours to weeks
- Privacy and ethical concerns
  - legal restrictions on using certain geospatial datasets
  - ethical considerations when using geospatial data, especially in terms of privacy and surveillance
- Earth System Variable Dependencies

# Standard data format can help...

- Current approaches – STAC ML, OGC
- Datasets generation pipeline
- FAIR - Findability, Accessibility, Interoperability and Reusability principles
- Responsible AI
- Tooling

# Types of geospatial datasets

- Multi-band datasets
  - Hyperspectral (>10 bands)
  - Multispectral (3-10 bands)
  - Single band - Panchromatic (1 band)
- In-situ Observations
- Point cloud (Time of Flight)
- Radar based data (Reflectometry)
- Citizen science
- Vector datasets
- Street View imagery

...

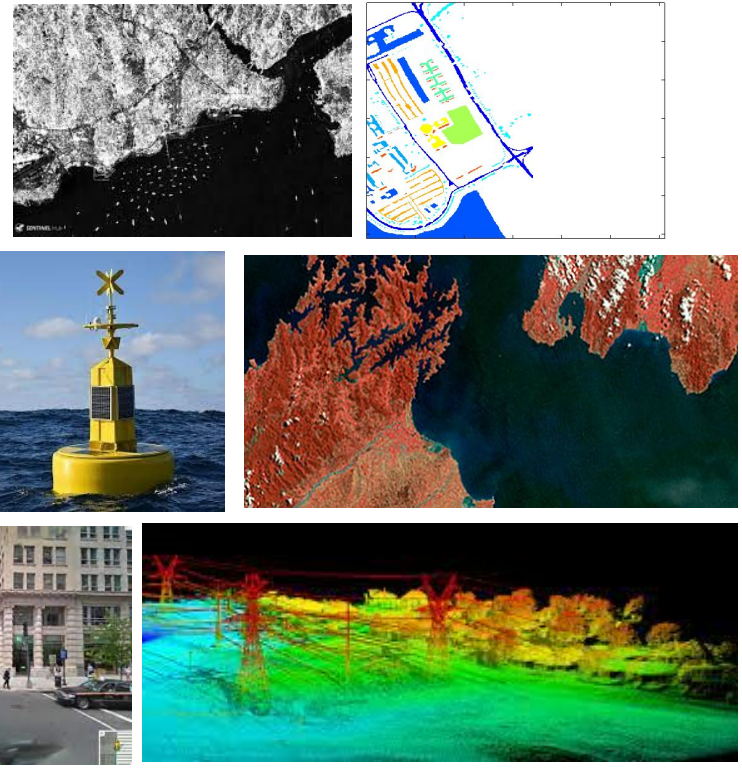
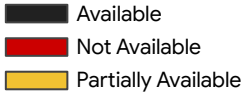


*Image source: Hyperspectral (Pavia University dataset), Multispectral (Landsat 8 FCC), Buoy (NOAA archive), Point Cloud (Google Images), Google Street View, SAR image (USGS Earth Explorer)*

# Unique Characteristics




- Resolution - Spatial, Spectral, Temporal, Radiometric
- Bounding Box - Geometry, Projection, Grid, Area, Coordinates
- Sensor and Sensing type - Optical, Microwave, Active, Passive
- Date and Time of Acquisition - Night time imagery, Noon time imagery
- Mode of acquisition - Spaceborne, Airborne, Mobile mapping
- Orientation of acquisition - Top view, Side view, Oblique view
- Type of Sensor - Onboard, Hand held, Table top
- Data model and Topology - Patch, Image, Graph, Mesh, Adjacency/Intersection
- Modeling categories - Detection, Classification, Reconstruction, Segmentation
- Scale of Processing - Object level, Scene level, Pixel level
- Dataset provider - Author, Group, Citation, Hosted by
- Data Statistics - Mean, Standard deviation, Minimum, Maximum

# Unique Characteristics and Schema.org

| Characteristic | Schema.org relation |
|---|---|
| **Resolution** | Property:TemporalCoverage |
| **Bounding Box/Polygon** | Property::spatial, Type:GeospatialGeometry, Type::GeoShape, Type::GeoCordinates, Property::elevation |
| **Sensor and Sensing type** | - |
| **Date and Time of Acquisition** | Property::temporal, Type::DateTime |
| **Mode of acquisition** | Property::measurementTechnique, Property::measurementMethod |
| **Orientation of acquisition** | - |
| **Type of Sensor** | Type:DataSet |
| **Data model** | Type::ImageObject, Type:DataSet, Type::3DModel, Type::ProductModel, Type::SoftwareApplication, Type::DataType |
| **Scale of Processing** | Property::object, |
| **Dataset provider** | Property::citation, Property::author, Property::issn |
| **Data Statistics** | Type::StatisticalVariable |

# Current Standards (Approaches)

- STAC Extension for ML
  - Stac records (ML Hub)
- OGC Training Data Markup Language (TrainingDML-AI)
- Croissant
- IEEE Standard for Artificial Intelligence and Machine Learning (AI/ML) Terminology and Data Formats (Last status - Draft stage, May 2023)

# Using ml-Croissant converter for Hugging Face hosted Geospatial datasets

- Eg: HLS Burn Scars Fine-tuning dataset
- Prerequisites:
  - Data Loader scripts needed for the converter to create metadata for croissant
- Observations:
  - Converter creates a croissant metadata file, which can be used with *mlcroissant* package
  - Metadata loads successfully, Issues loading data files
- Certain fields in the dataset missing:
  - LabelType - mask, bbox, value
  - ProblemCategory - scene, object, image
  - Projection - epsg
  - Overview - 64x64, 128x128

# Example


```
{
  "@context": {
    "@language": "en",
    "@vocab": "https://schema.org/",
    "column": "ml:column",
    "data": {
      "@id": "ml:data",
      "@type": "@json"
    },
    "dataType": {
      "@id": "ml:dataType",
      "@type": "@vocab"
    },
    "extract": "ml:extract",
    "field": "ml:field",
    "fileProperty": "ml:fileProperty",
    "format": "ml:format",
    "includes": "ml:includes",
    "isEnumeration": "ml:isEnumeration",
    "jsonPath": "ml:jsonPath",
    "ml": "http://mlcommons.org/schema/",
    "parentField": "ml:parentField",
    "path": "ml:path",
    "recordSet": "ml:recordSet",
    "references": "ml:references",
    "regex": "ml:regex",
    "repeated": "ml:repeated",
    "replace": "ml:replace",
    "sc": "https://schema.org/",
    "separator": "ml:separator",
    "source": "ml:source",
    "subField": "ml:subField",
    "transform": "ml:transform"
  },
  "@type": "sc:Dataset",
  "name": "hls_burn_scars",
  "description": "This dataset contains Harmonized Landsat and Sentinel-2 imagery of burn scars and the associa
  "citation": "@software{HLS_Foundation_2023,\n    author = {Phillips, Christopher and Roy, Sujit and Ankur, Ku
  "license": "cc-by-4.0",
  "url": "https://huggingface.co/datasets/ibm-nasa-geospatial/hls_burn_scars",
  "distribution": [
    {
      "@type": "sc:FileObject",
      "name": "tar-gz",
      "description": "Source *.tar.gz containing all the data.",
      "contentUrl": "https://huggingface.co/datasets/ibm-nasa-geospatial/hls_burn_scars/resolve/main/hls_burn_
      "encodingFormat": "application/x-tar",
      "sha256": "4e6f99a75cb2c500547b20662a15cbd531dc421376f815e91846ea542798e8e6"
    },
    {
      "@type": "sc:FileSet",
      "name": "source-images",
      "containedIn": "tar-gz",
      "encodingFormat": "image/tiff",
      "includes": "*/*_merged.tif"
    },
```

```
},
"@type": "sc:Dataset",
"name": "hls_burn_scars",
"description": "This dataset contains Harmonized Landsat and Sentinel-2 imagery of burn scars and the associated
masks for the years 2018-2021 over the contiguous United States.
There are 804 512x512 scenes. Its primary purpose is for training geospatial machine learning models.\n",
"citation": "@software{HLS_Foundation_2023,\n
author = {Phillips, Christopher and Roy, Sujit and Ankur, Kumar and Ramachandran, Rahul},\n
doi    = {10.57967/hf/0956},\n    month  = aug,\n
title  = {{HLS Foundation Burnscars Dataset}},\n
url    = {https://huggingface.co/ibm-nasa-geospatial/hls_burn_scars},\n
year   = {2023}\n}\n",
"license": "cc-by-4.0",
"url": "https://huggingface.co/datasets/ibm-nasa-geospatial/hls_burn_scars",
"distribution": [
  {
    "@type": "sc:FileObject",
    "name": "tar-gz",
    "description": "Source *.tar.gz containing all the data.",
    "contentUrl": "https://huggingface.co/datasets/ibm-nasa-geospatial/hls_burn_scars/resolve/main/hls_burn_scars.tar.gz",
    "encodingFormat": "application/x-tar",
    "sha256": "4e6f99a75cb2c500547b20662a15cbd531dc421376f815e91846ea542798e8e6"
  },
  {
    "@type": "sc:FileSet",
    "name": "source-images",
    "containedIn": "tar-gz",
    "encodingFormat": "image/tiff",
    "includes": "*/*_merged.tif"
  },
  {
    "@type": "sc:FileSet",
    "name": "source-annotations",
    "containedIn": "tar-gz",
    "encodingFormat": "image/tiff",
    "includes": "*/*.mask.tif"
  }
],
```


*Source: https://huggingface.co/datasets?other=doi%3A10.57967%2Fhf%2F0956*

# RAI Extension - Geospatial Use-case

Unset

```
{
    "@context":
  {
      "@language": "en",
      "@vocab": "http://mlcommons.org/croissant-RAI/1.0",
      "rai": "http://mlcommons.org/croissant-RAI",
      "sc": "https://schema.org/",

  },
  {
       "rai:dataCollection": "After co-locating the shapefile and HLS scene, the
       512x512 chip was formed by taking a window with the burn scar in the
       center. Burn scars near the edges of HLS tiles are offset from the
       center. Images were manually filtered for cloud cover and missing data to
       provide as clean a scene as possible, and burn scar presence was also
       manually verified.",

       "rai:dataCollectionType": "The dataset comprises 804 512x512 scenes. Each
       scene contain six bands, and masks have one band.",

       "rai:dataCollectionRawData": "Imagery is from V1.4 of Harmonized Landsat
       and Sentinel-2 (HLS). A full description and access to HLS may be found
       at https://hls.gsfc.nasa.gov/. The labels were from shapefiles maintained
       by the Monitoring Trends in Burn Severity (MTBS) group. The masks may be
       found at: https://mtbs.gov/",

       "rai:dataUseCases:
       [
              "The dataset can be used for training, validation, testing and
              fine-tuning."
       ],

       "rai:CitationInfo": "@software{HLS_Foundation_2023,
       author = {Phillips, Christopher and Roy, Sujit and Ankur, Kumar and
       Ramachandran, Rahul},
       doi    = {10.57967/hf/0956},
       month  = aug,
       title  = {{HLS Foundation Burnscars Dataset}},
       url    = {https://huggingface.co/ibm-nasa-geospatial/hls_burn_scars},
       year   = {2023}",
  }
}
```

*Source: Croissant RAI extension*

# Geo-Croissant Extension

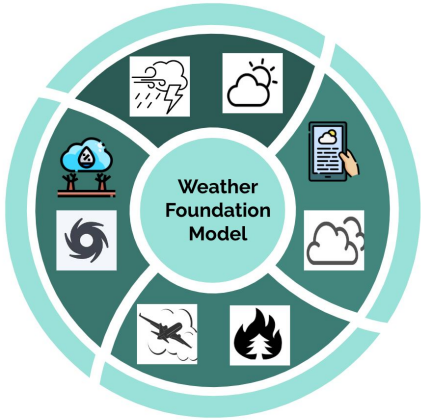
- Croissant Core and RAI extension helps in efficient representation of metadata and RAI attributes
  - Also, enhances processing in an end-to-end workflow

- Missing -
  - Spatial Reference Information
  - Nested Data Attributes (file formats such as netCDF4, HDF5, ZARR)
  - Interoperability with existing Cloud-native Geospatial Data Formats
  - Geographical Biases
  - Region restricted data access - Compatibility with DAACs

- Data-fusion opportunities with other modality datasets (tabular, graphs...)

# New Datasets in the Pipeline - Weather FM

- **In Preparation** - A data bench to be used for fine-tuning in-house Weather Foundation Model across multiple downstream tasks
    - Training data and labels for variety of tasks including
        - Classification
        - Prediction
        - Natural Language Generation
        - Retrieval




- **Towards Standardization** - Making the data bench openly available for use by the scientific community based on standard practices

# New Datasets in the Pipeline - Agriculture

- Crop type classification for DataPerf benchmark
  - 12-month time series, combination of multiple data sources
- Crop yield estimation
- Field boundary delineation

# Can “geo”-Croissant enable these?

- Geo-Coordinates and Resolution Awareness
  - e.g.: Co-located pixels in MODIS and HLS will hold different information, majorly because of density per pixel (spatial resolution) differences
- Geographic Sampling Awareness
  - A **global** training dataset is still far-fetched; Best large scale is available at **continent** scale
  - Specifying representative sampling is non-trivial
- Geospatial Data Provenance
  - Specification for Data transformation, Source data, Multiple data
- Temporal Characteristics
  - Most feature vary with time
- Responsible AI
  - Privacy, geospatial bias
- Capturing Variability at scale (Seasonal-subseasonal-...)
  - Requires aggregation of datasets from multiple sources, and spatio-temporal resolution

# How do we envision using Geo-Croissant?

- Community Standard for  ML EO datasets
- Enable data fusion of related ML datasets
    - e.g., socio-economic + EO > environmental justice
- Convert existing ML datasets into a universal format
- Leverage sustained community effort for open source format and tools
- Deployable AI workflows incorporating end-to-end processing
- Expand across NASA Science

# Scope of Development in Geo-Croissant

- Interoperability with existing standards to represent ML-ready data
- Designing/drafting the specification/standard
  - Highlighting the variables required to be used in AI-ready data for EO applications
- Develop tooling around the Croissant ecosystem for Geo-Croissant
  - Developing the Croissant Editor for Geospatial support
- Define various EO use-cases from ML perspective
  - Unsupervised VS Supervised VS Semi-supervised
  - Incorporating metadata variables for Training, fine-tuning data for Foundation Models/LLMs
- Support for the ML frameworks such TFDS, Kaggle, HF and geospatial processing engines GEE, VEDA, etc.

# Call for Geo-Croissant Working Group Volunteers

If you are interested in drafting the Geo-Croissant extension specification, join us in the Geo-Croissant Working Group.

Reach out to -

1. Rajat Shinde - rajat.shinde@uah.edu
2. Manil Maskey - manil.maskey@nasa.gov
3. Omar Benjelloun - benjello@google.com
4. Iksha Gurung - ig0004@uah.edu

# Resources

1. https://blog.research.google/2024/03/croissant-metadata-format-for-ml-ready.html
2. https://mlcommons.org/working-groups/data/croissant/
3. Croissant Spec - https://mlcommons.github.io/croissant/docs/croissant-spec.html
4. Croissant RAI Spec - https://mlcommons.github.io/croissant/docs/croissant-rai-spec.html
5. Croissant: A Metadata Format for ML-Ready Datasets - https://dl.acm.org/doi/abs/10.1145/3650203.3663326

# Discussion points

- (Seth) Model datasets /Reanalysis datasets vs real datasets
  - Results that are garbage but look good
    - Type of data - Observational vs simulated (RAI attribute)
    - Eg. Climate models don't have geoids
    - CF conventions
    - science-on-schema.org
- (Douglas) https://github.com/esipfed/data-readiness
- (Tyler) Is it more focused towards the AI/ML community
- (Carmen) Dataset self describing biases and sampling
- (Tyler) Data access from the repositories which don’t have HF, Kaggle, etc.

# Thank you