Developing Science Gateways with Clowder for the Permafrost Discover Gateway

Luigi Marini,

Todd Nicholson, Kastan Dey, Chandi Witharana, Ingmar Nitze, Gala Wind, Lauren Walker, Robyn Thiessen-Bock, Chris Jones, Matt Jones, Kenton McHenry, Rajitha Udawalpola, Ehsan Bhuiyan, Jason Cervenec, Bidhya Yadav, Amber Budden, Michael Brubaker, Guido Grosse, Ben Jones, Aiman Soliman, Anna Liljedahl

Sept 22nd, 2022

Woodwell Climate Research Center

National Center for Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Science Gateway

- **Goal:** Make models and tools developed by the PDG accessible to a wider set of user, beyond the original creators
- A **science gateway** is a resource to simplify access to community specific tools, applications and data collections for researchers, educators and students using user friendly, online interfaces

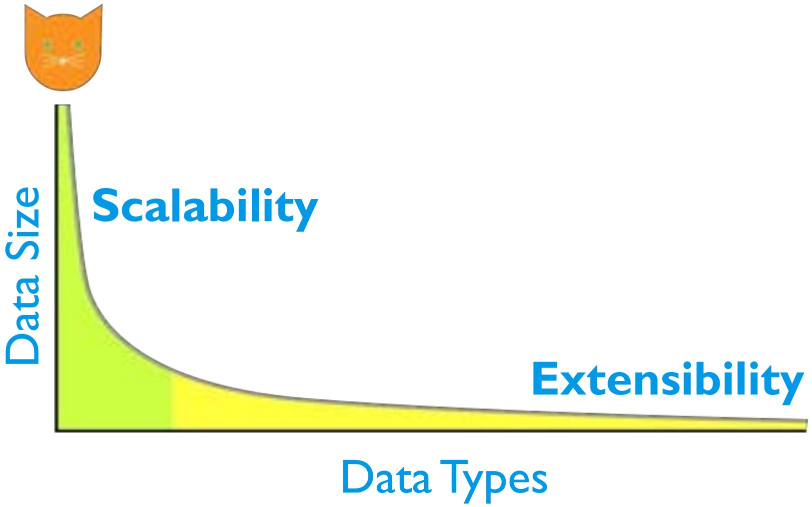
Data Pipelines

1. **Mapping Application for Arctic Permafrost Land Environment (MAPLE):** detects ice wedge polygons from high resolution optical imagery archived at Polar Geospatial Center. Uses Deep Learning Convolutional Neural Networks and requires GPU (Python, Tensorflow) ***(Chandi Witharana and Team)***
2. **Permafrost Region Disturbances (PRDs):** Lake area change, fire scars and retrogressive thaw slumps from Landsat images pre-processed using Google Earth Engine, plus additional machine learning and geospatial analysis (Python, Javascript, scikit-learn) (***Ingmar Nitze***)
3. **Arctic Satellite Joint Product (ASJP):** distills the high volume of archived NASA public satellite data to the essential community variables of interest that can be made available quickly and efficiently (Fortran) (***Gala Wind***)

Clowder and Long Tail Data

Clowder is a customizable and scalable data management framework to support any data format and multiple research domains

https://clowderframework.org/

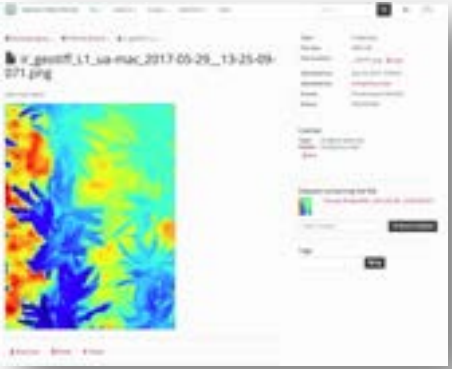


Heidorn, P. Bryan. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends*, vol. 57 no. 2, 2008, p. 280-299. *Project MUSE*, doi:10.1353/lib.0.0036.

UPLOAD, SHARE, INDEX HETEROGENOUS DATA AND METADATA

- **Upload** files, tag, license and organize them in datasets and collections
- Selectively **share** datasets with collaborators and publish them to the Internet
- Add well structured **metadata** to files and datasets using JSON Linked Data (**JSON-LD**)
- Automatically trigger **execution** of custom code on files and datasets using a Cloud based extraction system
- **Visualize** data and metadata in the browser using custom previewers

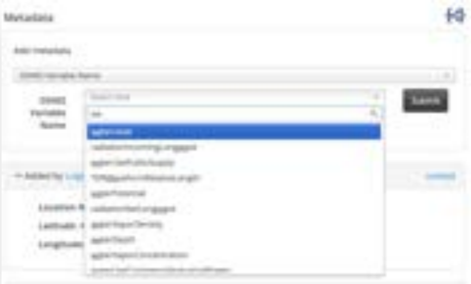
File visualizations, License, tags



Structured metadata

User management

Controlled vocabularies



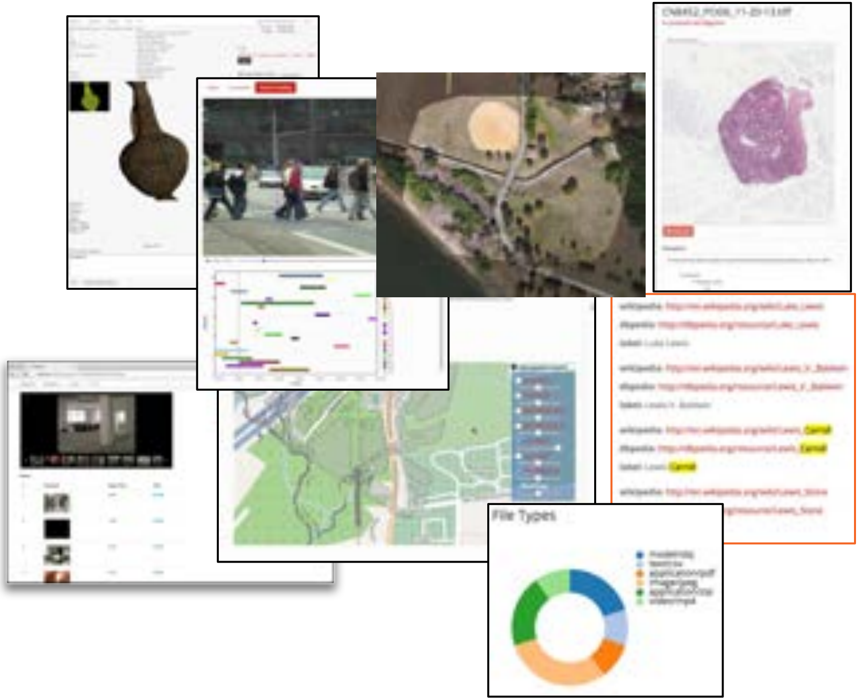
Advanced search

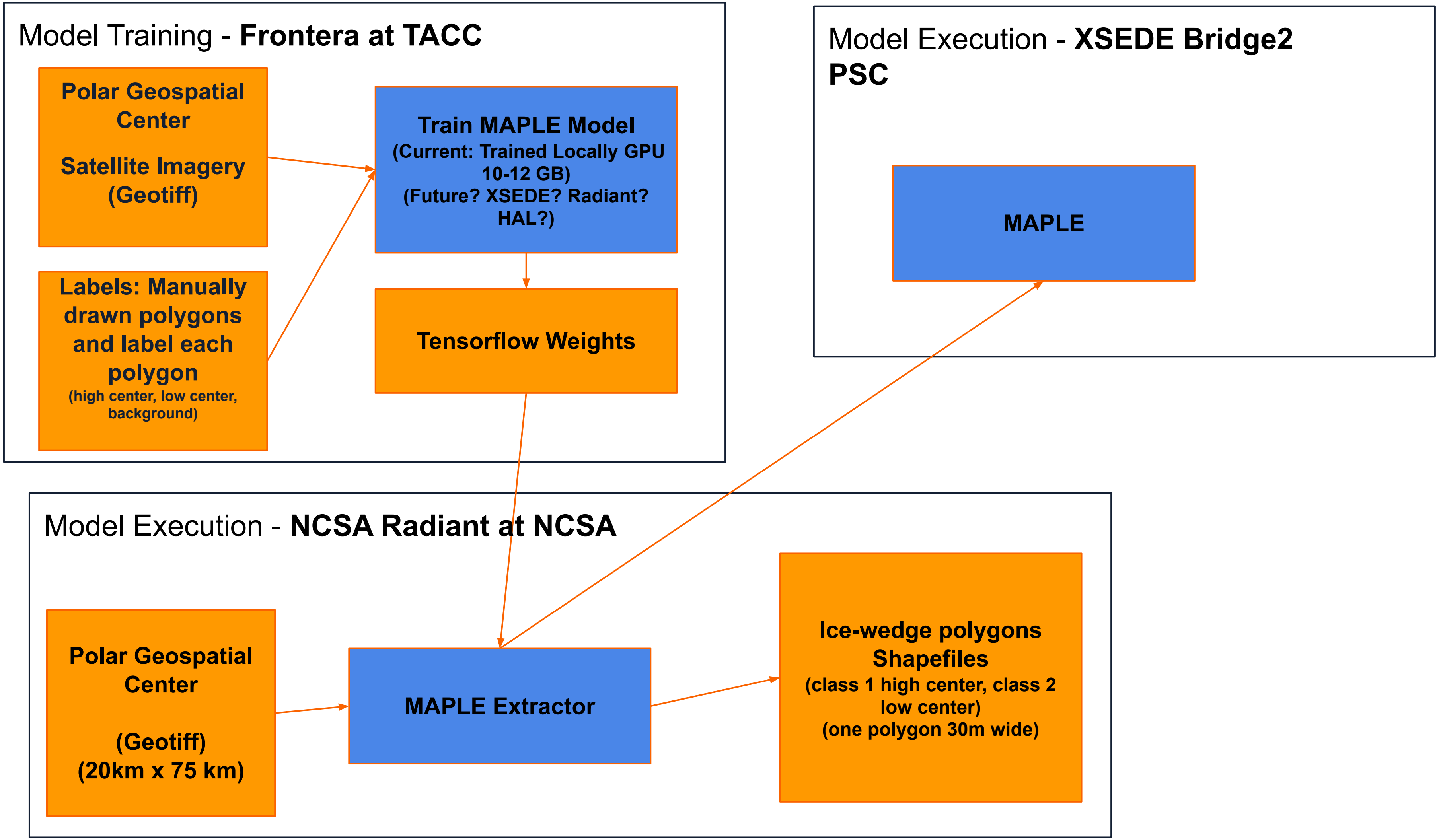
MANUAL AND AUTOMATIC DATA ANALYSIS AND VISUALIZATION

Automatic Execution

Develop extractors in Python, R and any other language

Visualizations



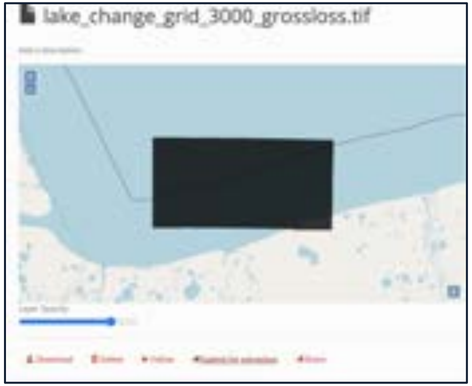

Model Training - **Frontera at TACC**
Polar Geospatial Center
Satellite Imagery (Geotiff)
Labels: Manually drawn polygons and label each polygon
(high center, low center, background)
Train MAPLE Model
(Current: Trained Locally GPU 10-12 GB)
(Future? XSEDE? Radiant? HAL?)
Tensorflow Weights
Model Execution - **XSEDE Bridge2 PSC**
MAPLE
Model Execution - **NCSA Radiant at NCSA**
Polar Geospatial Center
(Geotiff)
(20km x 75 km)
MAPLE Extractor
Ice-wedge polygons Shapefiles
(class 1 high center, class 2 low center)
(one polygon 30m wide)

MAPLE on XSEDE Bridges2

- Requires GPU resources not currently available on Radiant.
- Solution - use ssh to submit job to XSEDE Bridges2
- **File Extractor**: individual files are transferred to XSEDE. A slurm job is created, run when resources are available, and then uploaded.
- **Dataset Extractor**: Start with empty dataset and path to input files on bridges. Determine which files have not been run, create slurm jobs, then upload. Checks to see which files do not need to be uploaded or run.

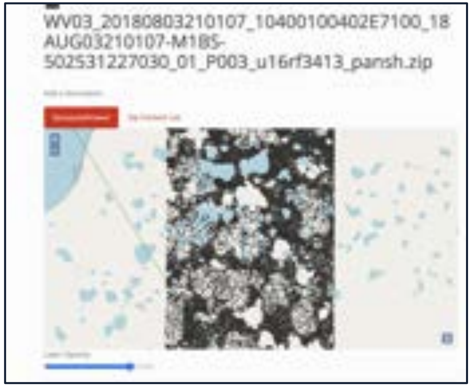
MAPLE Clowder Extractor

Input - GeoTiff

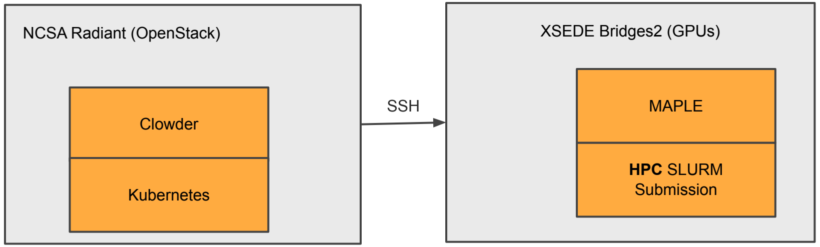


Submit

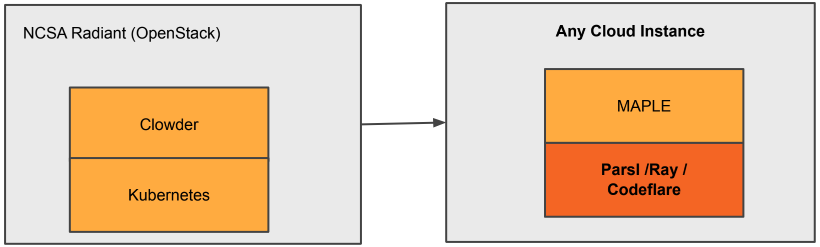
Output - Shapefile



Current



Future



Portable Parallel Pipelines

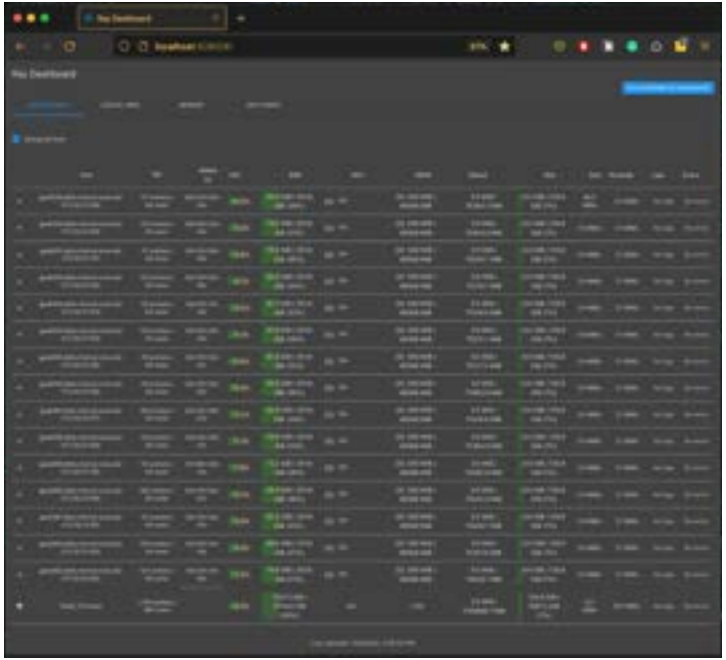
- Prototyped a Clowder extractor to Launch Parsl jobs on local Kubernetes cluster
- Prototyped Ray extractor to launch Ray jobs on CodeFlare provisioned clusters

https://parsl-project.org/

https://www.ray.io/

https://codeflare.dev/

MAPLE Visualization Pipeline in Ray On Delta



https://www.ncsa.illinois.edu/research/project-highlights/delta/

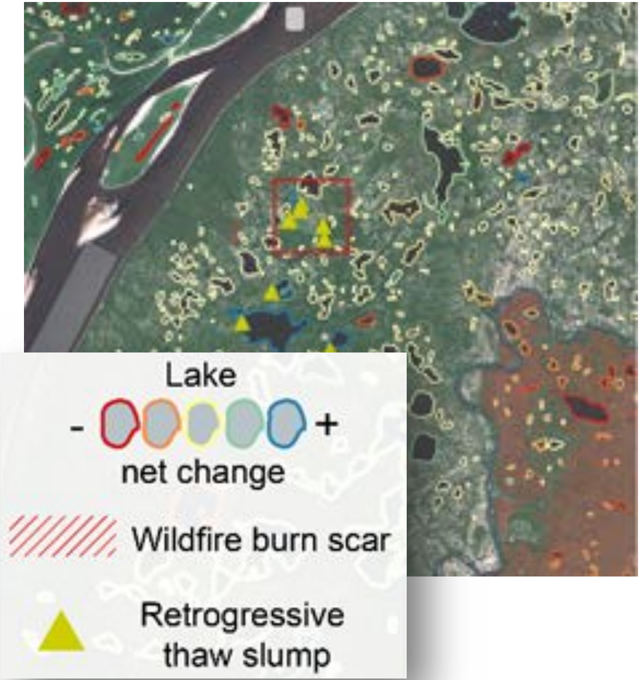
MAPLE file count
- 10s of thousands.
- Each image requires a dedicated GPU (or 8GB vram) for ML infrence

type	min_file_size	max_file_size	total_file_count
high_ice	9.536743e-05	1641.910	22370
low_ice	9.536743e-05	1253.986	1456
medium_ice	9.536743e-05	1254.654	11209
water_clipped	9.536743e-05	1071.425	8807

Kastan Dey, Robyn Thiessen-Bock, Lauren Walker, Matt Jones

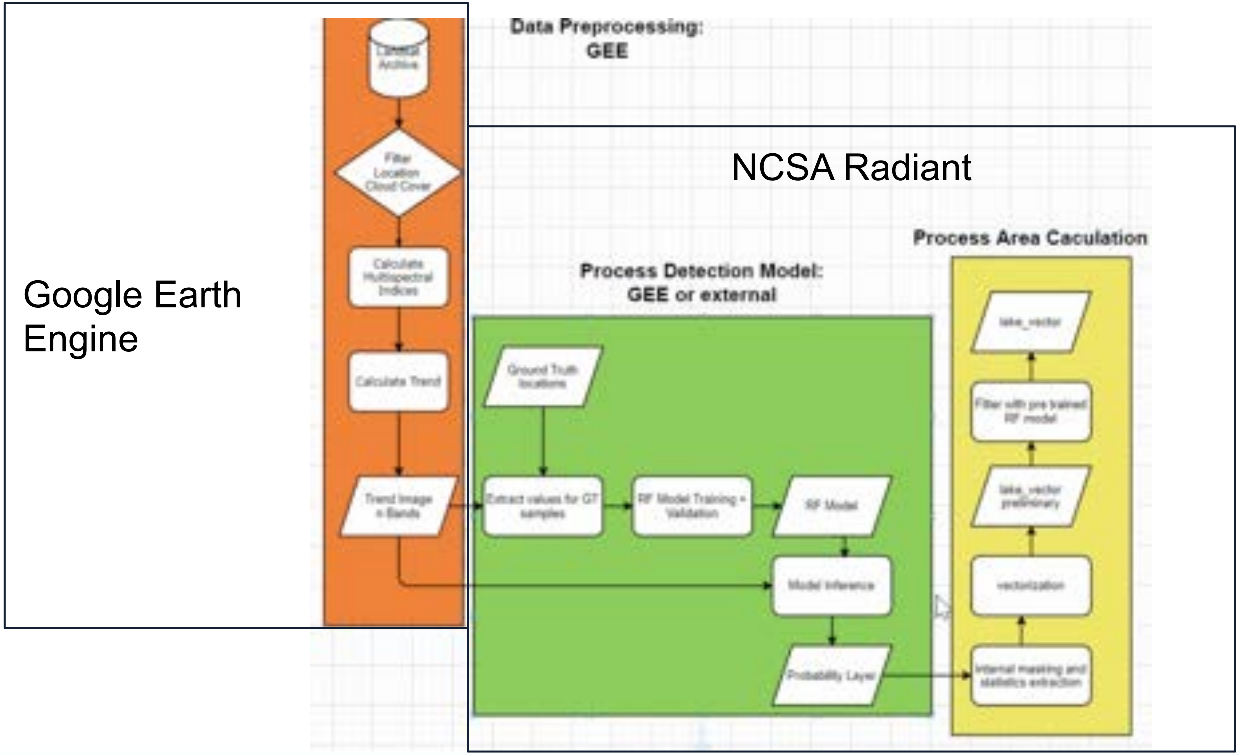
LandsatTrends - Permafrost Region Disturbances (PRD)

- Input Landsat & Sentinel
- Identify
  - Thaw slumps
  - Fire scars
  - Lake changes



Ingmar Nitze et al. 2018

LandsatTrends Pipeline



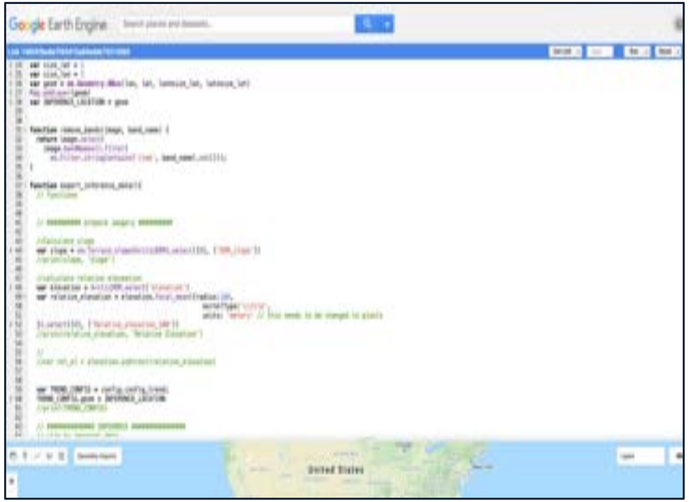
LandsatTrends Extractors

1. Preprocessing Extractor
   a. submits preprocessing steps to Google Earth Engine using Python API
   b. Data from Google Earth Engine is transferred to Google Cloud using GC API
   c. Data in Google Cloud is uploaded to Clowder
2. Inference Extractor
   a. Run model on local files

Google Earth Engine API

- Instead of using the google earth console or Google drive, images can be downloaded directly to a location on disk.

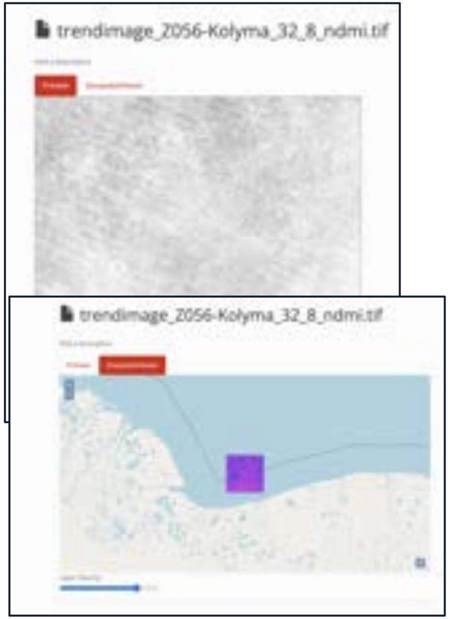
GEE Javascript Console



Local Python Code

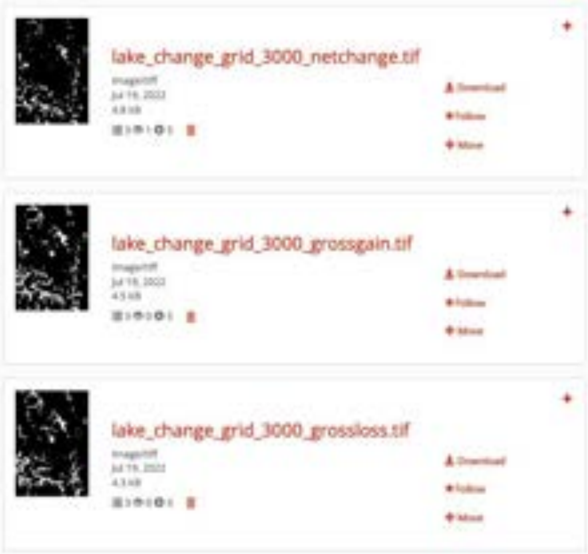
LandsatTrends Detection Model and Area Calculation

Inputs - GeoTiffs

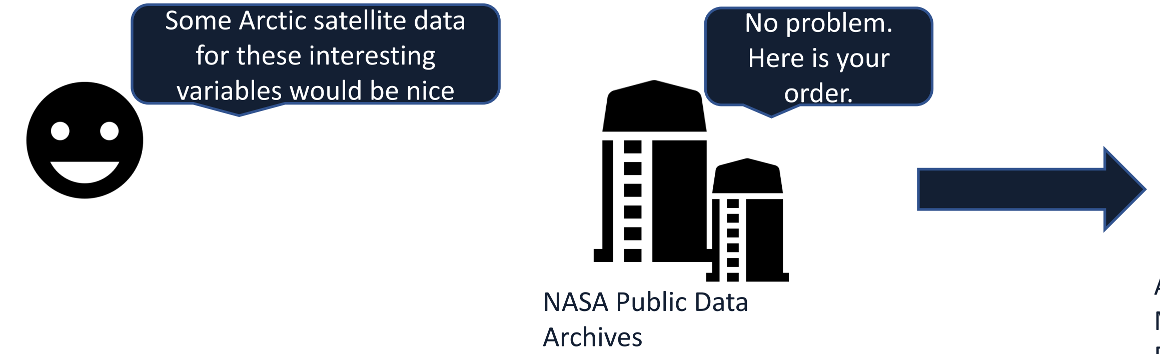


Submit

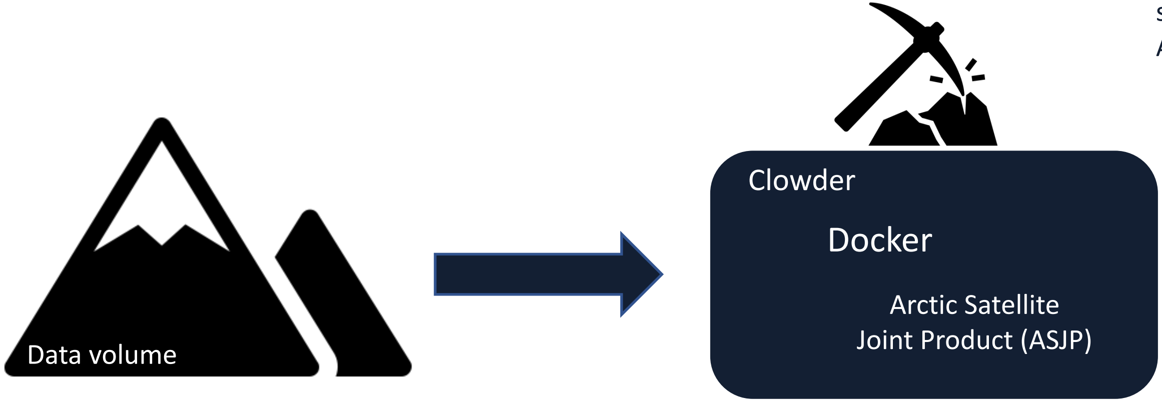
Outputs - GeoTiffs and Shapefiles

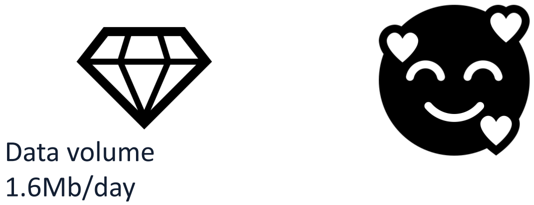


Arctic Satellite Joint Product (ASJP)



Archived data is very available, but difficult to navigate.
Many variables stored together in large files.
Entire file or entire sets of files must be downloaded for study.
Access assumes very high speed internet connection.





Arctic Satellite Joint Product with help of Docker and Clowder distills the high volume of archived satellite data to the essential
community variables of interest that can be made available quickly and efficiently.

Gala Wind @ NASA

17GB / day

1.7 MB / day

Flat list of files

```
MOD06_L2.A2022156.1145.061.2022158133116.hdf  MOD14.A2022150.0600.061.2022151092320.hdf    MOD29E1D.A2022148.061.2022151212855.hdf
MOD06_L2.A2022156.1150.061.2022158132342.hdf  MOD14.A2022150.0605.061.2022151090022.hdf    MOD29E1D.A2022149.061.2022151221800.hdf
MOD06_L2.A2022156.1155.061.2022158132334.hdf  MOD14.A2022150.0610.061.2022151090442.hdf    MOD29E1D.A2022150.061.2022151223616.hdf
MOD06_L2.A2022156.1200.061.2022158132329.hdf  MOD14.A2022150.0615.061.2022151090650.hdf    MOD29E1D.A2022151.061.2022152234746.hdf
MOD06_L2.A2022156.1205.061.2022158132440.hdf  MOD14.A2022150.0620.061.2022151090110.hdf    MOD29E1D.A2022152.061.2022153140412.hdf
MOD06_L2.A2022156.1210.061.2022158132337.hdf  MOD14.A2022150.0625.061.2022151090543.hdf    MOD29E1D.A2022153.061.2022154094511.hdf
MOD06_L2.A2022156.1215.061.2022158132505.hdf  MOD14.A2022150.0630.061.2022151085940.hdf    MOD29E1D.A2022153.061.2022158013908.hdf
MOD06_L2.A2022156.1220.061.2022158132446.hdf  MOD14.A2022150.0635.061.2022151090739.hdf    MOD29E1D.A2022153.061.2022159224825.hdf
MOD06_L2.A2022156.1225.061.2022158132455.hdf  MOD14.A2022150.0640.061.2022151090812.hdf    MOD29E1D.A2022154.061.2022158014936.hdf
MOD06_L2.A2022156.1230.061.2022158132504.hdf  MOD14.A2022150.0645.061.2022151090248.hdf    MOD29E1D.A2022155.061.2022156085054.hdf
MOD06_L2.A2022156.1235.061.2022158132440.hdf  MOD14.A2022150.0650.061.2022151090121.hdf    MOD29E1D.A2022156.061.2022158014700.hdf
MOD06_L2.A2022156.1240.061.2022158133421.hdf  MOD14.A2022150.0655.061.2022151091117.hdf    MOD29E1D.A2022157.061.2022158090314.hdf
MOD06_L2.A2022156.1245.061.2022158133822.hdf  MOD14.A2022150.0700.061.2022151090007.hdf    MOD29E1D.A2022158.061.2022159080709.hdf
MOD06_L2.A2022156.1250.061.2022158133724.hdf  MOD14.A2022150.0705.061.2022151090633.hdf    MOD29E1D.A2022159.061.2022160075838.hdf
MOD06_L2.A2022156.1255.061.2022158134130.hdf  MOD14.A2022150.0710.061.2022151090757.hdf    MOD29E1D.A2022160.061.2022161080643.hdf
MOD06_L2.A2022156.1300.061.2022158133235.hdf  MOD14.A2022150.0715.061.2022151090405.hdf    MOD29E1D.A2022161.061.2022162092018.hdf
MOD06_L2.A2022156.1305.061.2022158134313.hdf  MOD14.A2022150.0720.061.2022151085812.hdf    MOD29E1D.A2022162.061.2022163084023.hdf
MOD06_L2.A2022156.1310.061.2022158133842.hdf  MOD14.A2022150.0725.061.2022151085744.hdf    MOD29E1D.A2022163.061.2022164080310.hdf
MOD06_L2.A2022156.1315.061.2022158133914.hdf  MOD14.A2022150.0730.061.2022151085752.hdf    MOD29E1D.A2022164.061.2022165101941.hdf
MOD06_L2.A2022156.1320.061.2022158134158.hdf  MOD14.A2022150.0735.061.2022151090432.hdf    MOD29E1D.A2022165.061.2022166075322.hdf
```

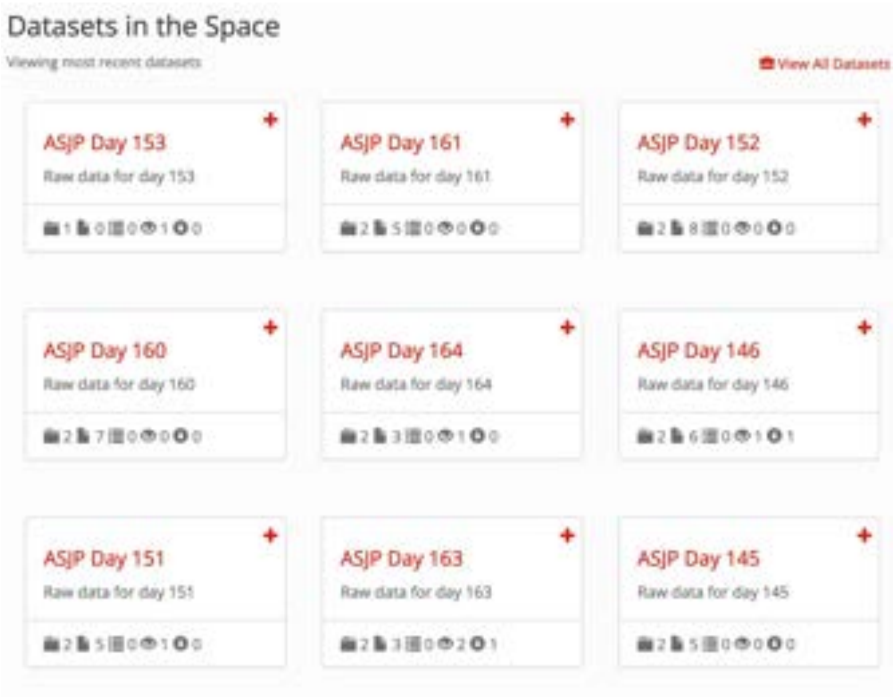
Folders by day

```
lmarini@pdg-upload:/data/modaps_by_day$ ls
142  143  144  145  146  147  148  149  151  152  153  154  155  156  157  158  159  160  161
 162  163  164
```

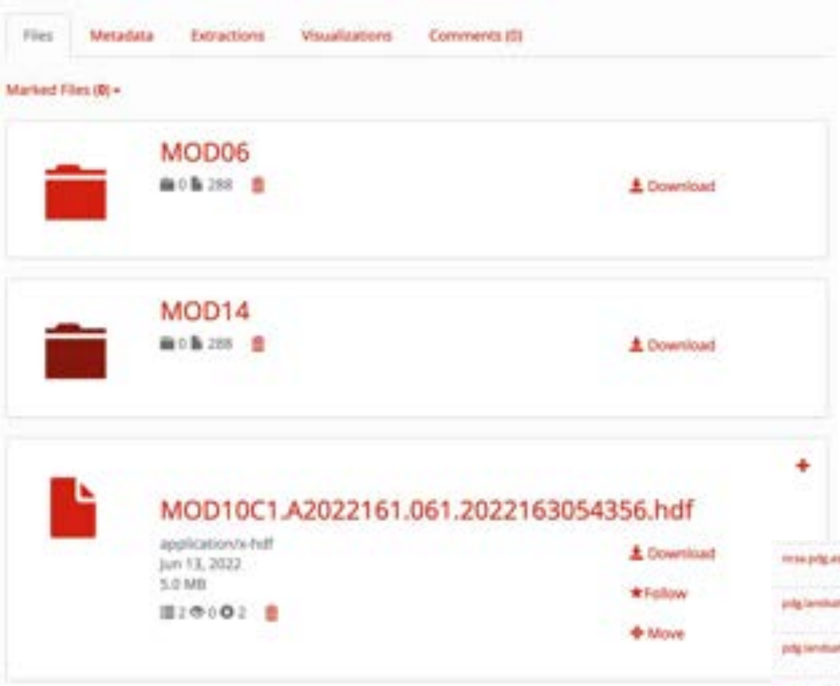
```
lmarini@pdg-upload:/data/modaps_by_day/163$ ls
MOD06                                          MOD14
MOD09CMG.A2022161.061.2022163043428.hdf  MOD29E1D.A2022162.061.2022163084023.hdf
MOD10C1.A2022161.061.2022163054356.hdf
```

MOD09CMG.A2022161.061.2022163043428.hdf

Raw Data per Day



Private Dataset per Day

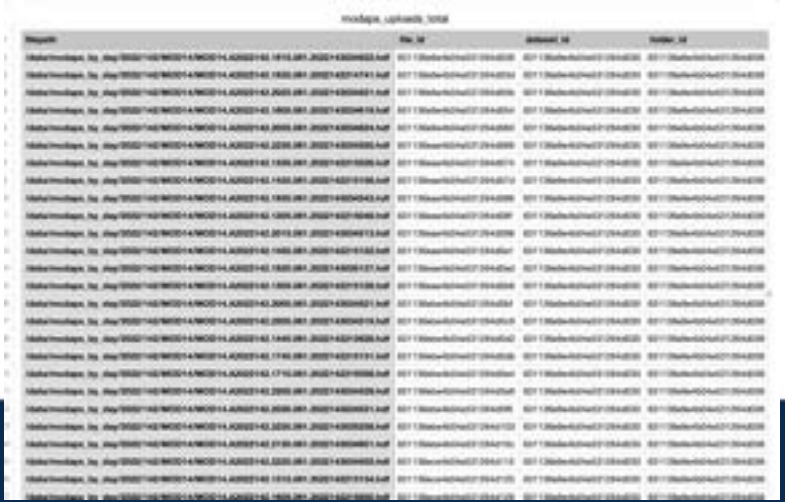


Public Data


ASJP Outputs
Owner: Luigi Marini
Created on Mar 25, 2022
All Rights Reserved Luigi Marini
Add creator(s)
Add a description
Add Files
Download All Files
Archive All
Unarchive All
Delete
Follow
Create Folder
Submit for extraction
Files
Metadata
Extractions
Visualizations
Comments (0)
Marked Files
ASJP Day 164.nc4
application/octet-stream
Jun 15, 2022
1.0 MB
Download
Follow

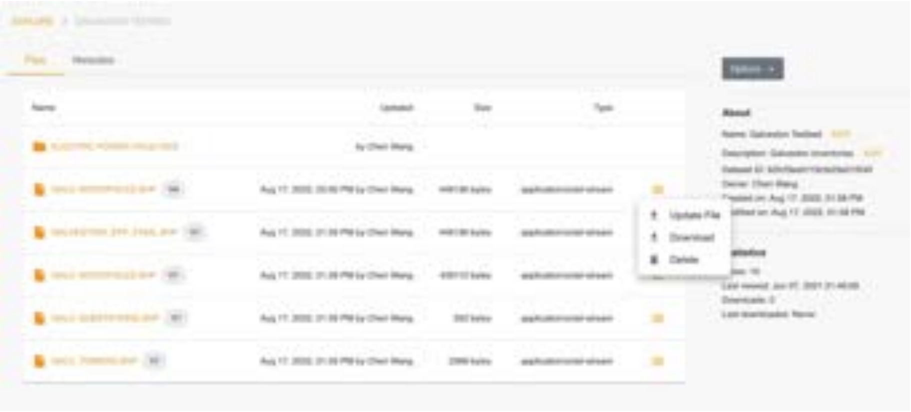
Recovering from Failure

- Keep track of uploaded data uploaded (checkpoint)
  - which files are uploaded successfully, which files are not
- Avoid creating duplicate datasets, uploading duplicate files
- Do not upload partial files being written to disk
- Always confirm integrity of uploaded files before deleting original files on disk



Clowder v2 - Join the Development!

- https://github.com/clowder-framework/clowder2



JS React python FastAPI

Thank you!

https://permafrost.arcticdata.io

https://clowderframework.org/

Permafrost Discovery Gateway

Navigating
the New Arctic
Awards #
1927872, 1927723,
1927729, 1927720,
1927920, & 2052107

Woodwell Climate Research Center