

Best practices in sharing enhanced data products and machine learning algorithms: learnings from Frontier Development Lab

James Parr, Bill Diamond, Lika Guhathakurta



Over the last year or so we have been asking ourselves this question...

How do we make our space + AI research open and ready for others?

1) How do we make ML + science simpler to reproduce?

2) What do we mean by “AI ready”?

3) Do we need common quality standards?

1) How do we make
ML + science
simpler to
reproduce?



But not just to
reproduce...

TRUSTED ENOUGH TO
BRANCH AND BUILD
FROM.



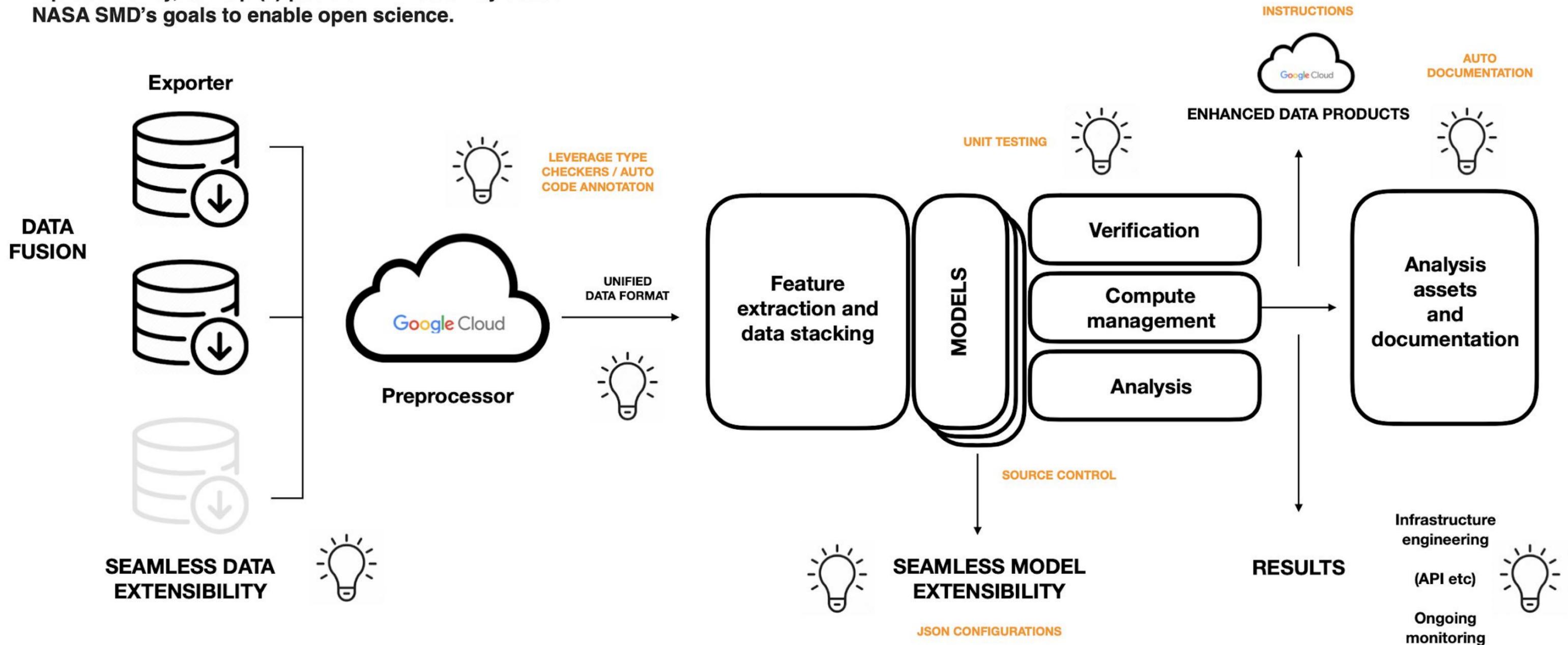
Papers / GitLab Notebooks

FDL AND REPRODUCIBILITY

FDL is committed to taking a leadership position in ML reproducibility, to help (a) position FDL as a key asset in NASA SMD's goals to enable open science.



- + Develop and Implement Capabilities to Enable Open Science
- + Continuously Evolve Data and Computing Systems
- + Harness the Scientific Community and Strategic Partnerships for Innovation.
- + Providing data access to the wider research community and for validation of published research results.



GitLab Projects Groups More

Search or jump to...

Frontier Development Lab > FDL US 2020 Earth Engine > eie_vision > Repository

master eie_vision / README.md Find file Blame

initial commit of the repo: skeleton and misc instructions
Alexander Lavin authored 7 months ago

README.md 6.17 KB Edit Web IDE Replace

Earth Intelligence Engine - NASA FDL 2020

Main repo for the Frontier Development Lab 2020 EIE project.

Getting started

Setup

We recommend setting up your environment with conda (see [this intro](#) and [the docs](#)).

Clone the repo: `git clone git@gitlab.com:frontierdevelopmentlab/fdl-us-2020-eie/eie_vision.git`

Replicate and activate the conda environment:

```
conda env create -f conda.yml
conda activate eie_vision
```

Tooling

For experimentation well be using [MLFlow](#). Please see the [tutorials and docs](#).

Our models and data loaders are to be implemented in PyTorch Lightning: see [the docs](#) and [example notebooks](#) (e.g. [GAN](#)).

👉 these enable us to experiment with NN models efficiently and collaboratively.

Issues 1 Merge Requests 1 CI / CD Operations Packages & Registries Analytics Wiki Snippets Members Collapse sidebar

SET-UP - CONDA

TOOLING (and Tutorials)

Example Notebooks

RUNNING

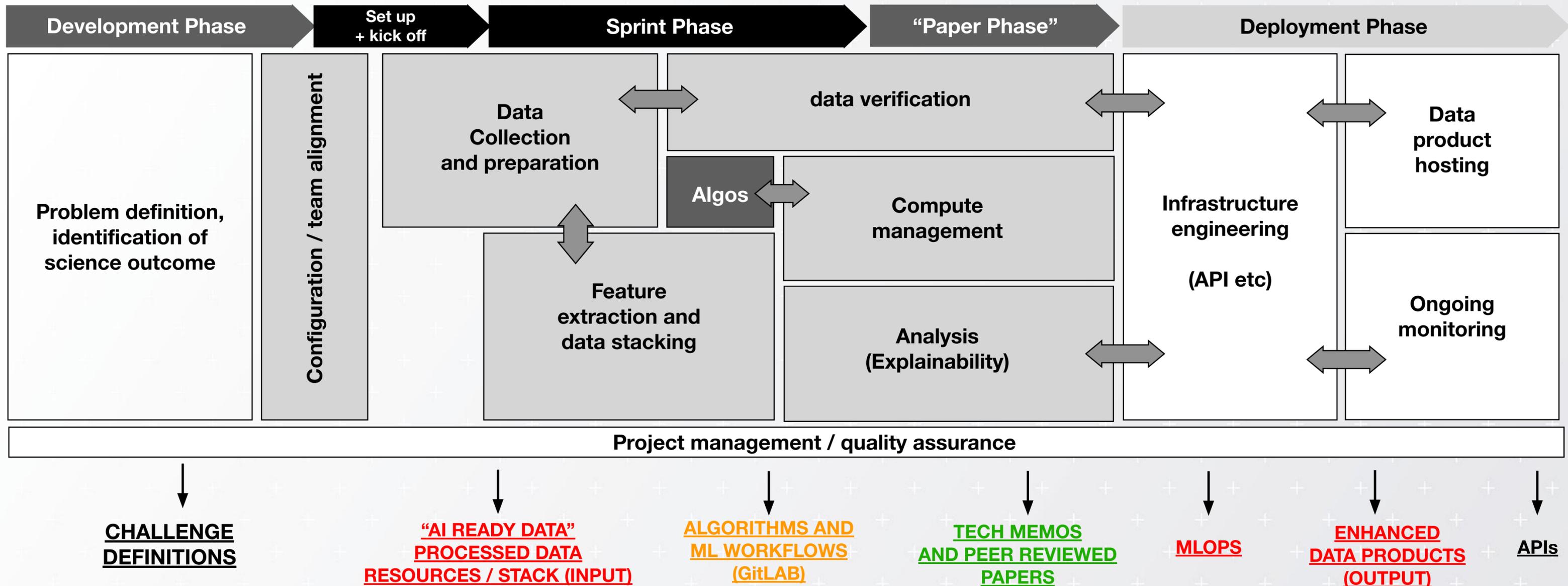
CONTRIBUTING

“Feature Branch Workflow”

CONVENTIONS

CODE STYLE

But there is still a lot of value left on the table...

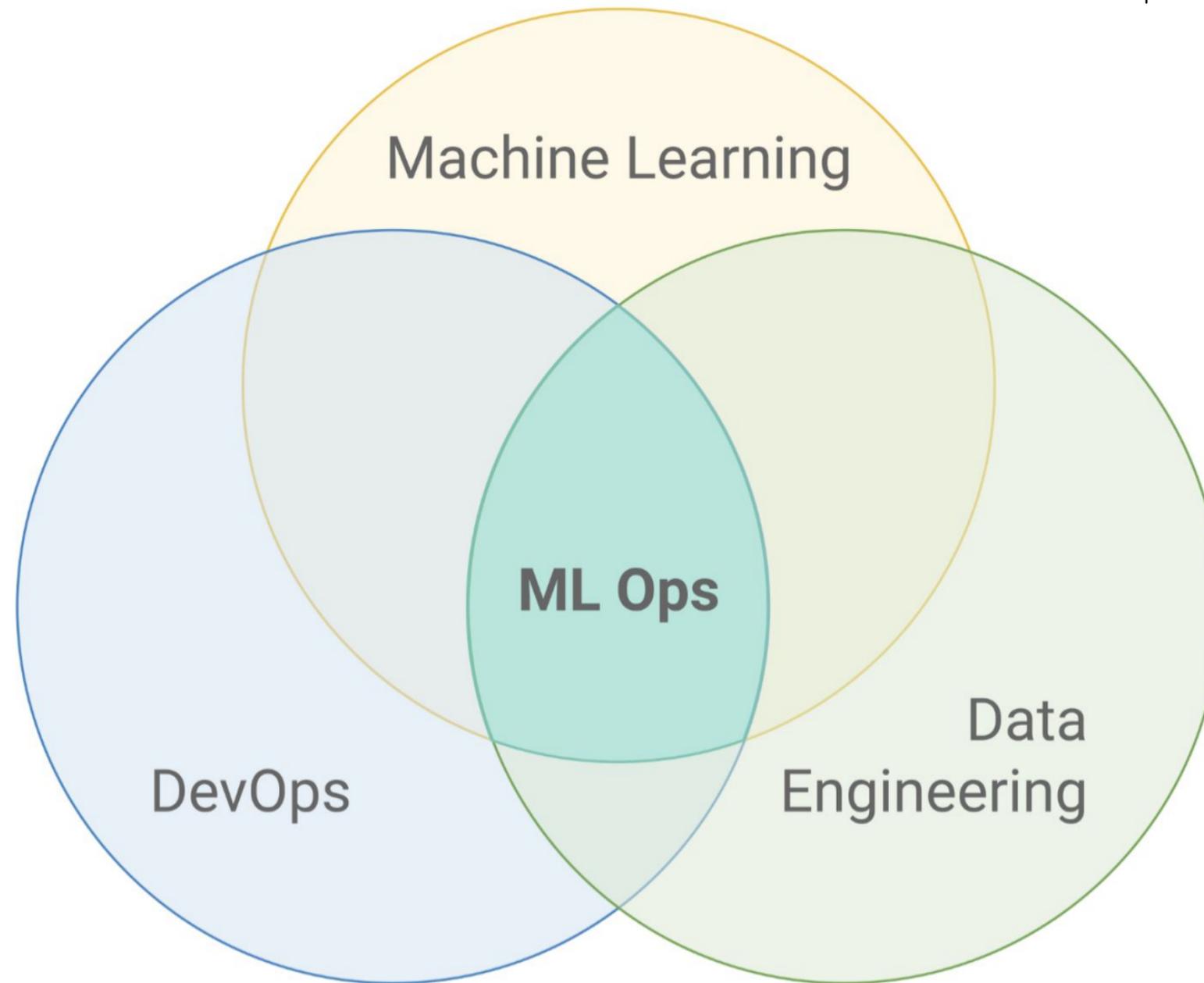


Adapted from Google NeurIPS 2019

MLOps

Tools = "MLOps"

<https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f>



Adds two new components to the DevOps paradigm.

README.md

AstroNet-PyTorch

AstroNet translated from TensorFlow to PyTorch

AstroNet: a neural network for classifying exoplanets transits

Astronet is a deep convolutional neural net (CNN) originally developed in TensorFlow and available [here](#). See [Shallue & Vanderburg \(2018\)](#) for more information about AstroNet and its application to *Kepler* light curves.

NASA Frontier Development Lab

In 2018, [NASA's Frontier Development Lab](#) (FDL) formed a team of scientists and machine learning experts to investigate the application of machine learning to detecting and classifying exoplanet transits. As part of this work, the team utilized AstroNet as a baseline model and improved upon it by adding new scientific domain knowledge.

The 2018 NASA FDL Exoplanet Team: [Megan Ansdell](#), [Yani Ioannou](#), [Hugh Osborn](#), [Michele Sasdelli](#)

Astronet-PyTorch

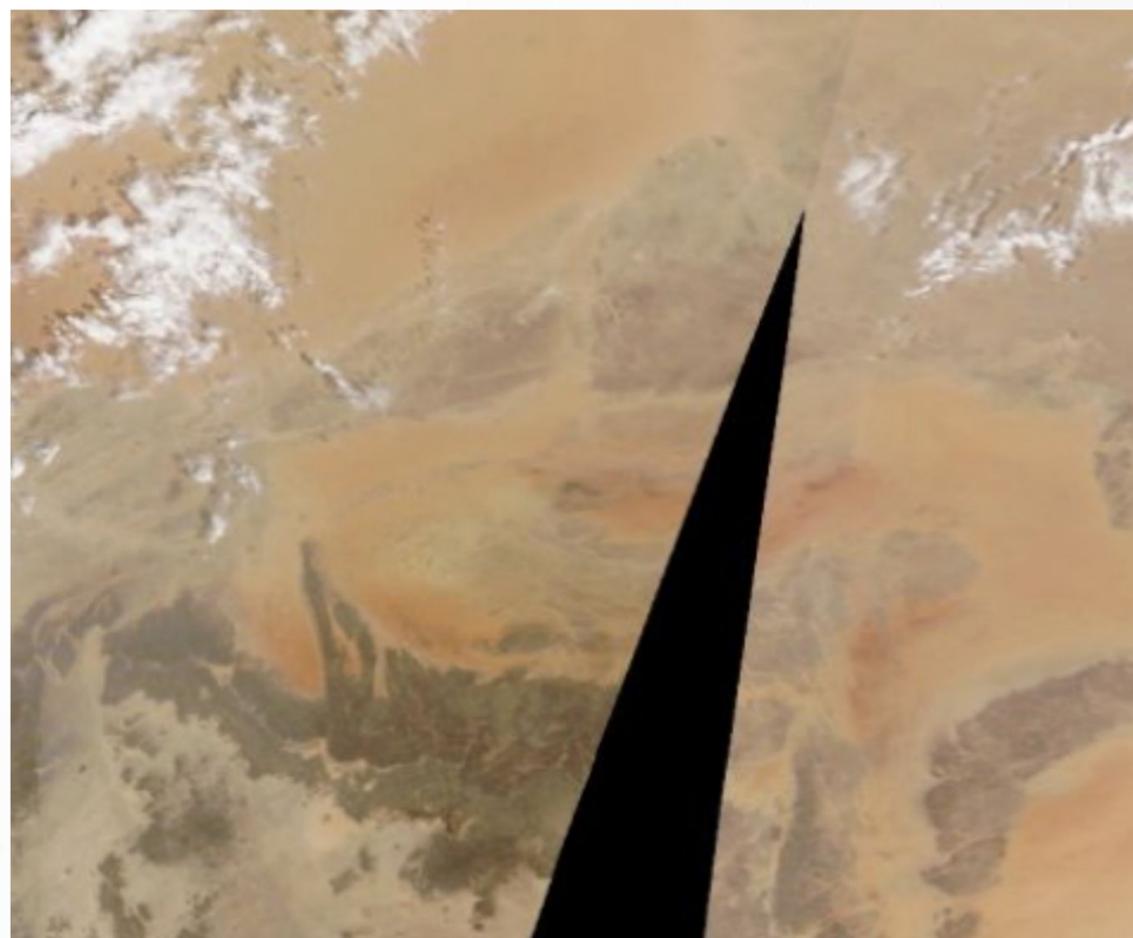
As part of their work in 2018, the NASA FDL Exoplanet Team translated AstroNet from TensorFlow into PyTorch. Here we make this work publicly available. There will soon also be a code for downloading the required *Kepler* light curves and generating the input views and labels; for now, you can download the required input files from [this DropBox link](#) (you must divide them into train, val, and test folders for the code to work) or follow the instructions on the Astronet GitHub.

If you use this work please cite: [2018 NASA FDL Exoplanet Team \(2018\), ApJ Letters, 869, L7.](#)

FDL TEAMS CONSTANTLY BUILD
BESPOKE TOOLS TO SOLVE
PROBLEMS.

WHAT IF WE COULD MAKE TOOLS
FOR GENERAL PROBLEMS?

What about a tool to reduce the effects of swath gaps in unsupervised Machine Learning?

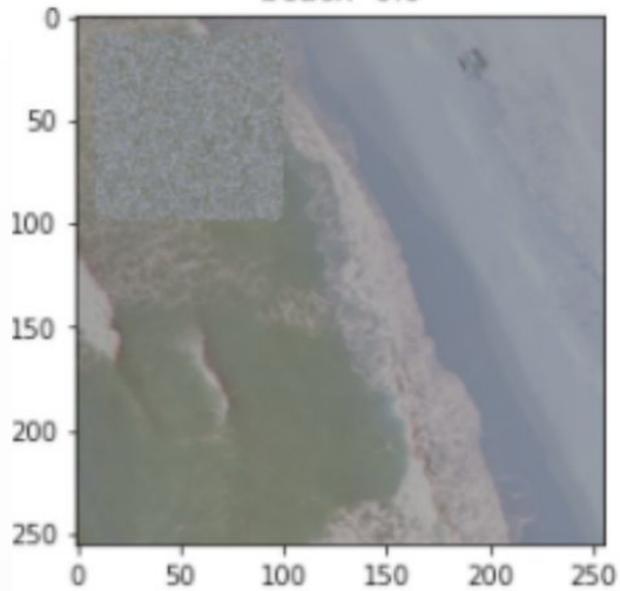


Authors: Sarah Chen, Esther Cao, Anirudh Koul, Satyarth Praveen, Meher Kasam, Siddha Ganju

A swath filler can automatically make data closer to “AI readiness”

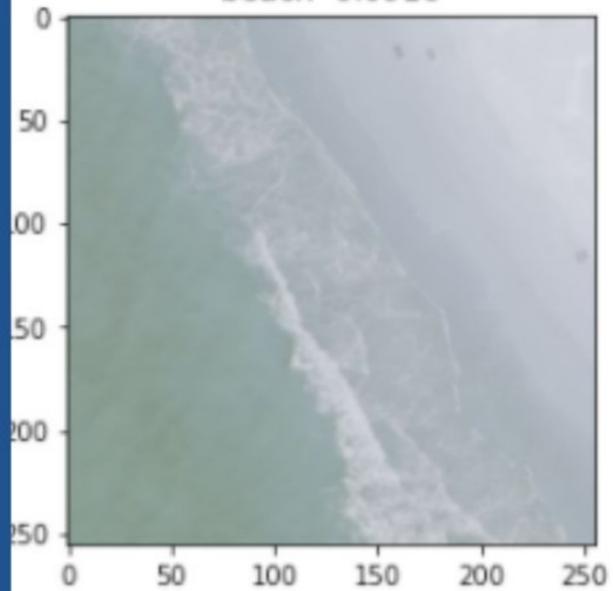
Query Image

beach 0.0

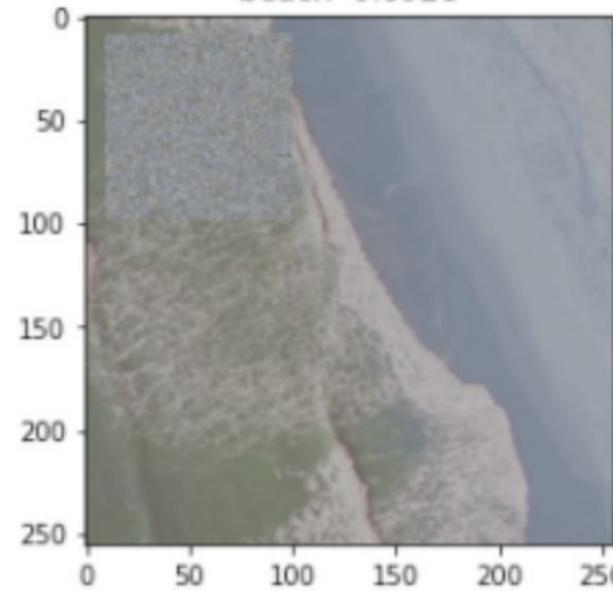


Nearest Neighbors

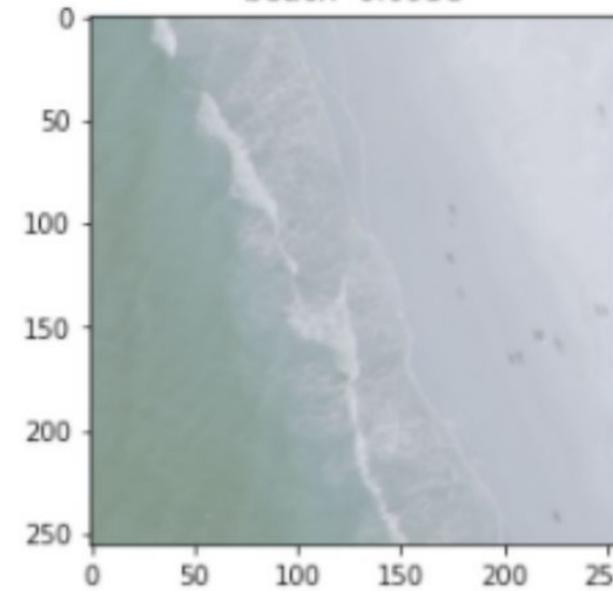
beach 0.0916



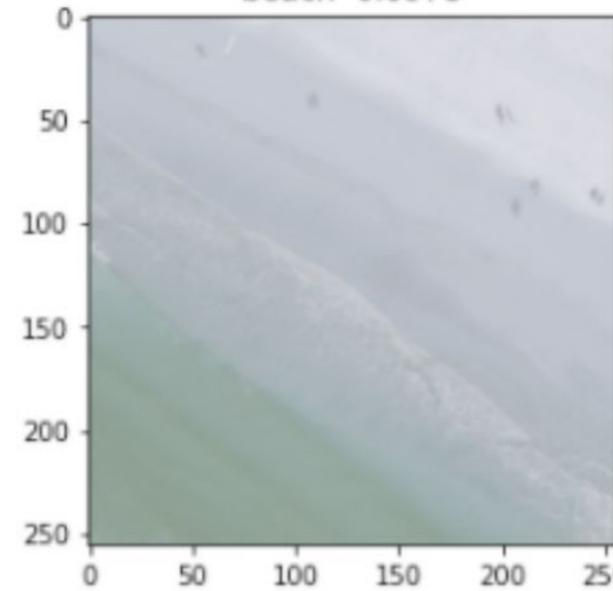
beach 0.0928



beach 0.0938



beach 0.0978



4 / 4 neighbors are correct

FDL 2020
KNOWLEDGE DISCOVERY FRAMEWORK

DISCOVERY — DATA INGESTION — DATA HARMONIZATION — MODEL OPTIMIZATION

- WorldView Scalable Reverse Search Engine**
(Abhigya Sodani & Mike Levy)
<https://www.youtube.com/watch?v=riKHz1y558c&t=15m37s>
- Multi-Resolution Search**
(Rajeev Godse)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h54m21s>
- 80x High-performance Data Downloader**
(Fernando Lisboa Shivam Verma)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h02m27s>
- Empty Swath Filler (Invisibility Cloak)**
(Sarah Chen & Esther Cao)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h39m38s>
- Balancing unlabelled Imbalanced data**
(Deep Patel & Erin Gao)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h18m47s>
- Label-Less Learner**
(Suhas Kotha)
<https://www.youtube.com/watch?v=riKHz1y558c&t=43m00s>

“ML OPs” across entire workflows



FDL 2020
KNOWLEDGE DISCOVERY FRAMEWORK

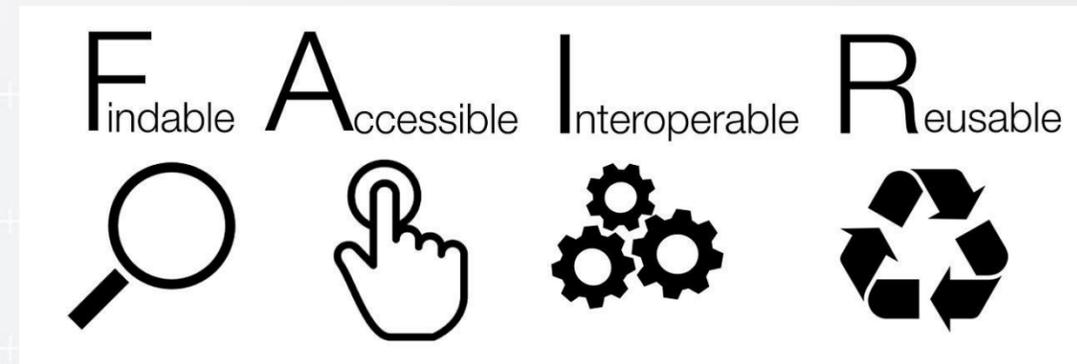
DISCOVERY — DATA INGESTION — DATA HARMONIZATION — MODEL OPTIMIZATION

- WorldView Scalable Reverse Search Engine**
(Abhigya Sodani & Mike Levy)
<https://www.youtube.com/watch?v=riKHz1y558c&t=15m37s>
- Multi-Resolution Search**
(Rajeev Godse)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h54m21s>
- 80x High-performance Data Downloader**
(Fernando Lisboa Shivam Verma)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h02m27s>
- Empty Swath Filler (Invisibility Cloak)**
(Sarah Chen & Esther Cao)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h39m38s>
- Balancing unlabelled Imbalanced data**
(Deep Patel & Erin Gao)
<https://www.youtube.com/watch?v=riKHz1y558c&t=1h18m47s>
- Label-Less Learner**
(Suhas Kotha)
<https://www.youtube.com/watch?v=riKHz1y558c&t=43m00s>

“ML OPs” across entire workflows



2) What do we really mean by “AI ready” data?



Space presents some unique challenges when it comes to Machine Learning and Data (the kind of things that aren't in 'ML for beginners'.)

Planetary Scale
Diverse (lack of labels)

Multidimensional
Sparse / Synthetic

Data harmonisation
Fusion
Physics Integration

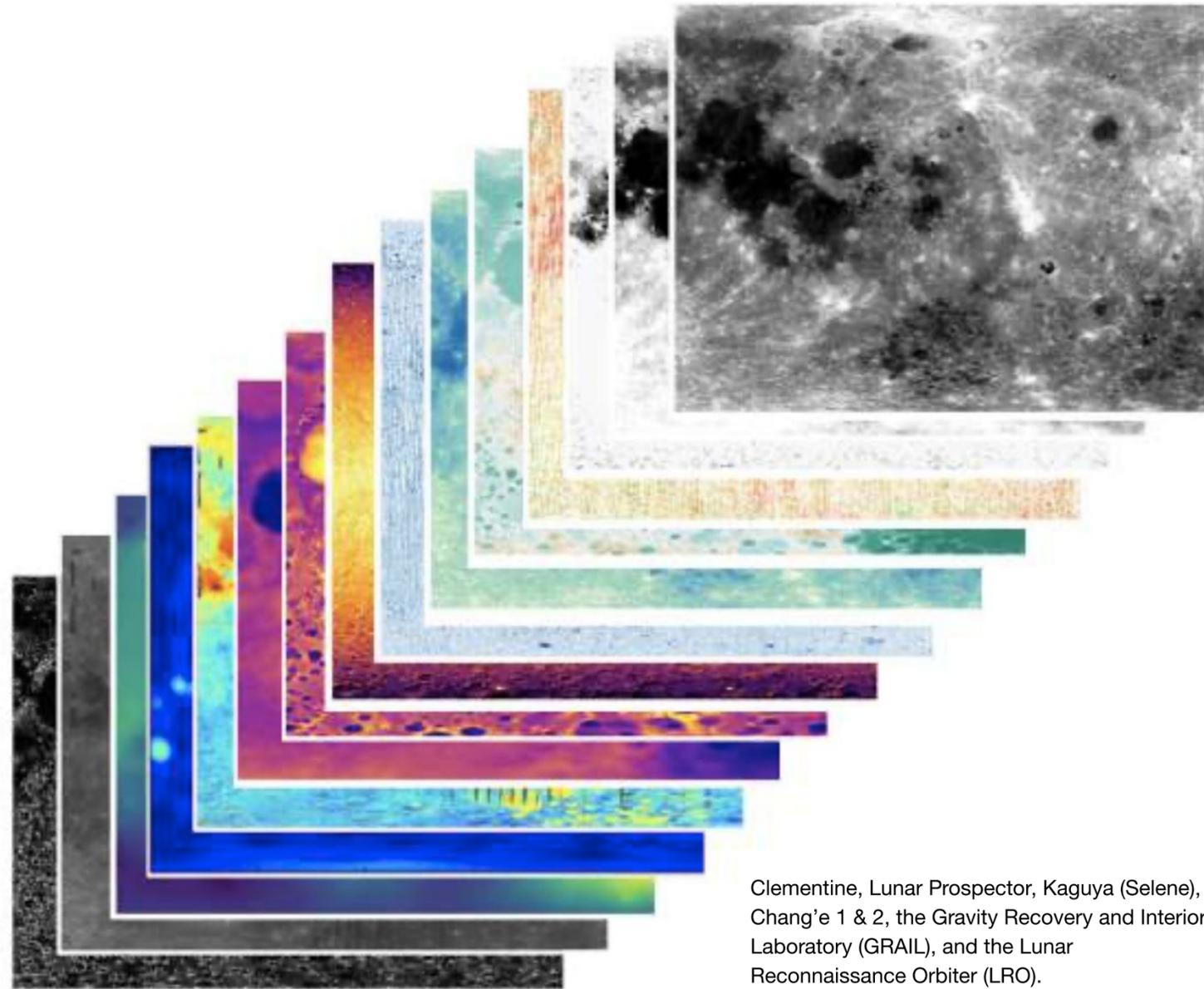
Compute and network constraints

Big-hose Problem

So the barrier to entry is already high.

Small-hose Problem

“AI READY DATA” = HARMONIZED



- + Moved / Ingested to a common location “ETL”
- + Structured
- + Fused / calibrated
- + Derived data integrated
- + Aligned on common projection

LUNAR DATA STACK

A curated, global, multi-sensor, multi-spacecraft, ML-ready data stack consisting of 42 maps of various lunar orbiter physical measurements from 7 different satellites, 12 different scientific instruments and over 25 years of observations.

The data stack covers a wide range of frequency bands (microwaves to optical), included derived data products (such as rock abundance from thermal data) and topography information.

Interpolated and aligned all layers onto a common equatorial map projection with 100x100m spatial resolution.

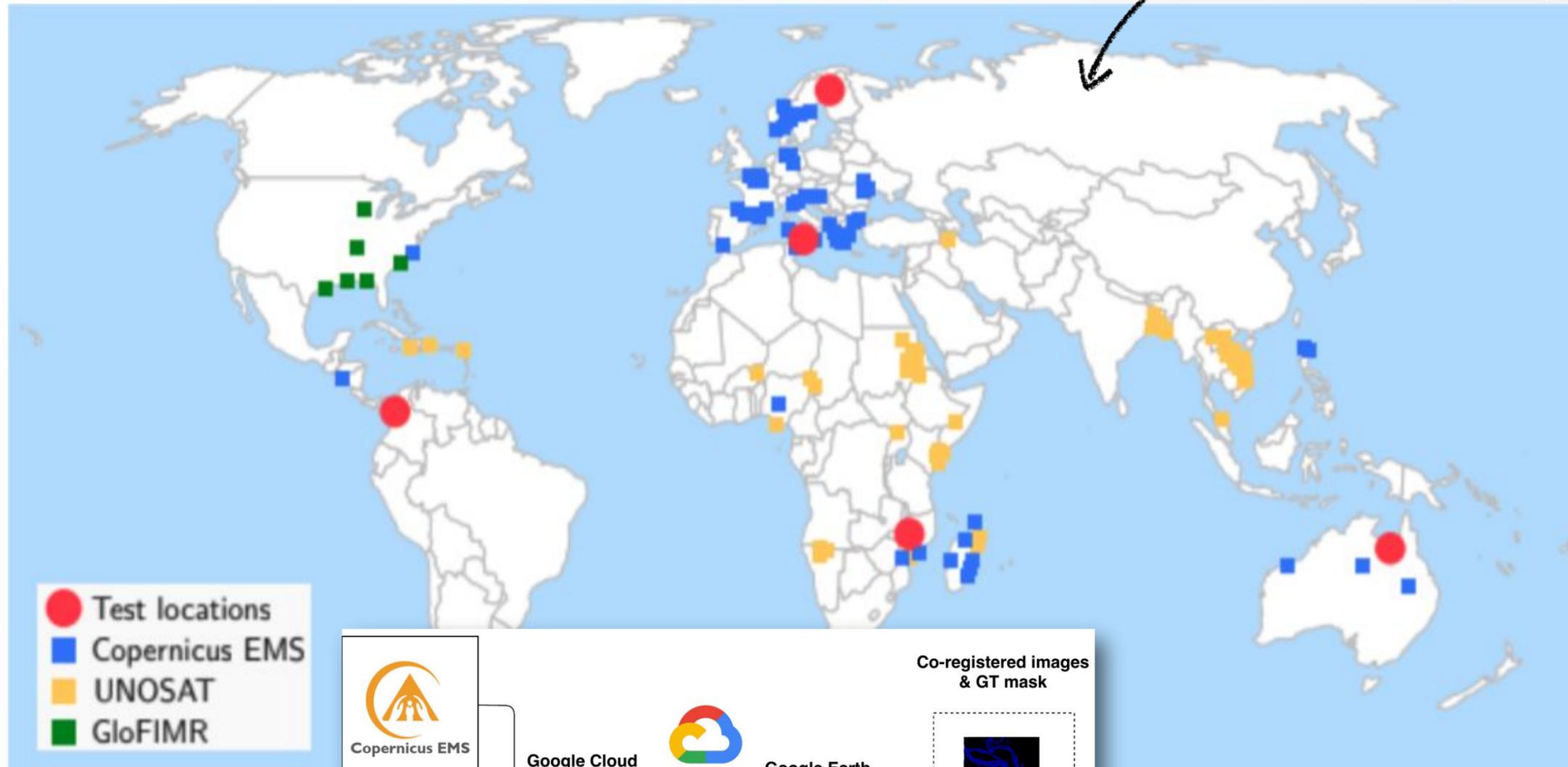
Data fusion (NASA, JAXA etc)

Derived Data (physics models)

Aligned into a common projection

"AI READY DATA" = GLOBAL AND LABELED

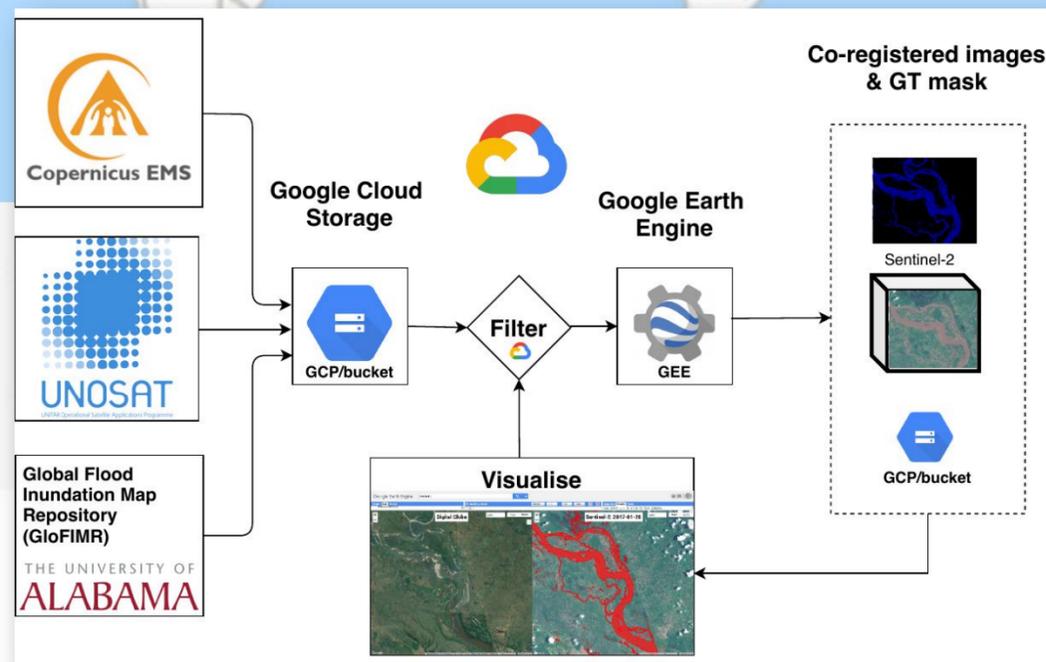
Labels are rarely complete
(if they exist at all!)



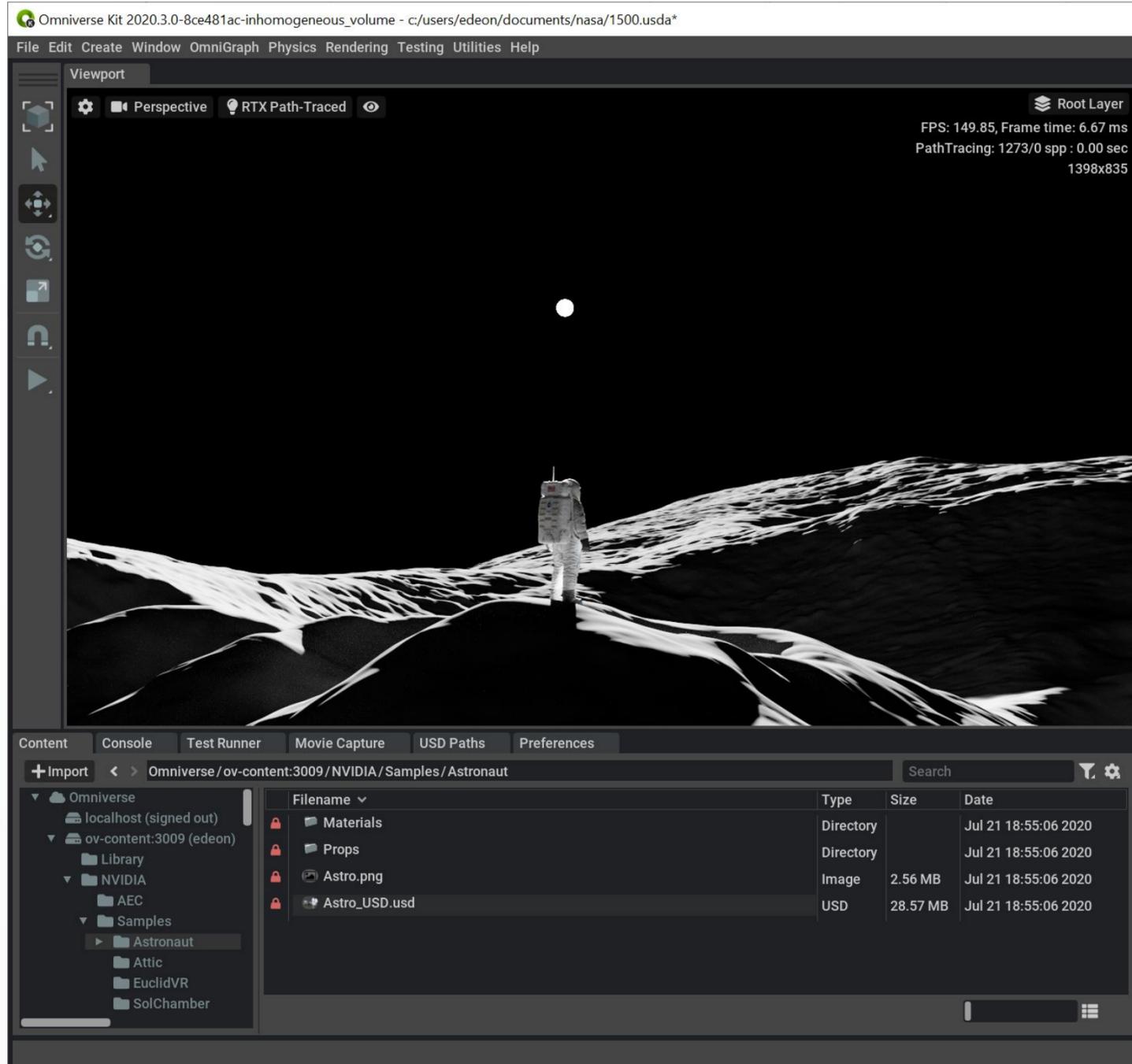
- Test locations
- Copernicus EMS
- UNOSAT
- GloFIMR

- + Aggregated
- + Validation of machine generated maps
- + Labeled (by hand)
- + Metrics

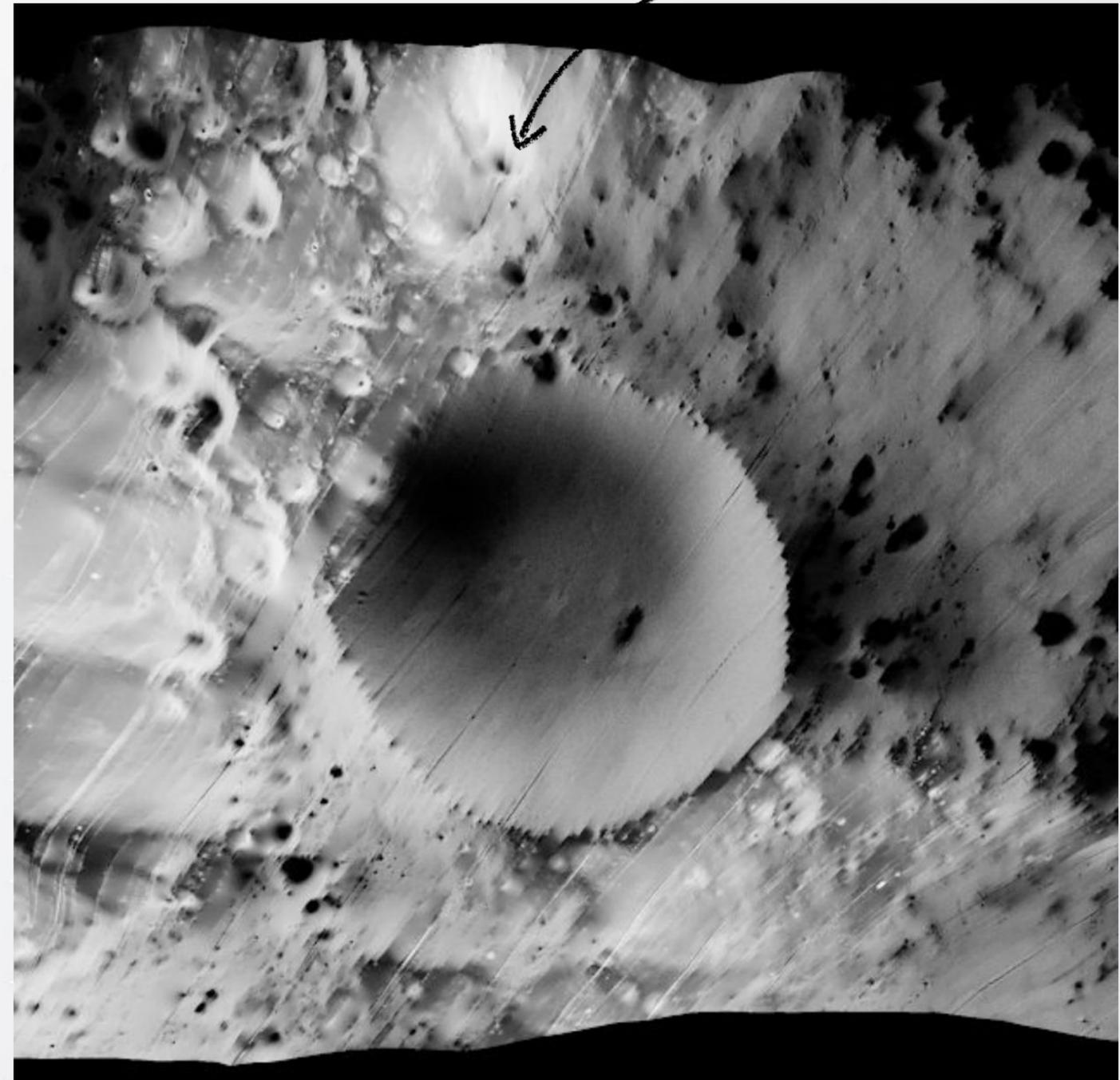
- WorldFloods contains 422 flood extent maps created either manually or semi-automatically, where a human validated machine-generated maps.
- The dataset covers 119 floods events that occurred between November 2015 and March 2019. All maps are sourced from from three organisations: the Copernicus Emergency Management Service (Copernicus EMS), the flood portal of UNOSAT, and the Global Flood Inundation Map Repository (GLOFIMR).
- A flood extent map is a vector layer (shapefile) derived from a satellite image with polygons indicating which part of that image has water (in some cases it distinguishes between flood water and permanent water and in other cases it does not).



“AI READY DATA” = INFORMED BY PHYSICS



What if there's no data in the first place? Like optical imagery inside a lunar PSR?





Public INARA150 DATASET EXPLORE ☆

File Edit View Insert Runtime Tools Help Changes will not be saved

+ Code + Text Copy to Drive

INARA Dataset Notebook

This dataset (55GB / 150k planetary atmospheres) is a subset of the 3M main INARA dataset. Results from the analysis of this set were published at the Bayesian Deep Learning Workshop at NeurIPS (<http://bayesiandeeplearning.org/2018>): <http://bayesiandeeplearning.org/2018/papers/120.pdf>

The Astrobiology II Team

Researchers:

- Molly O'Beirne - University of Pittsburgh
- Frank Soboczenski - King's College London
- Michael Himes - University of Central Florida
- Simone Zorzan - University of Luxembourg

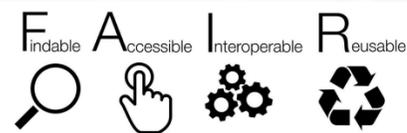
Faculty:

- Giada Arney - NASA Goddard Space Flight Center
- Shawn Domagal-Goldman - NASA Goddard Space Flight Center
- Gunes Baydin - University of Oxford
- Adam Cobb - University of Oxford
- Massimo Mascaro - Google Cloud
- Daniel Angerhausen - ETH Zurich
- Nathalie Cabrol - SETI Institute

This notebook will let you explore the INARA Dataset

Getting set up

In this section we will import required packages, get you authenticated and set up with the INARA public data bucket.



FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability.

EXAMPLE FDL DATA PRODUCT OUTPUTS

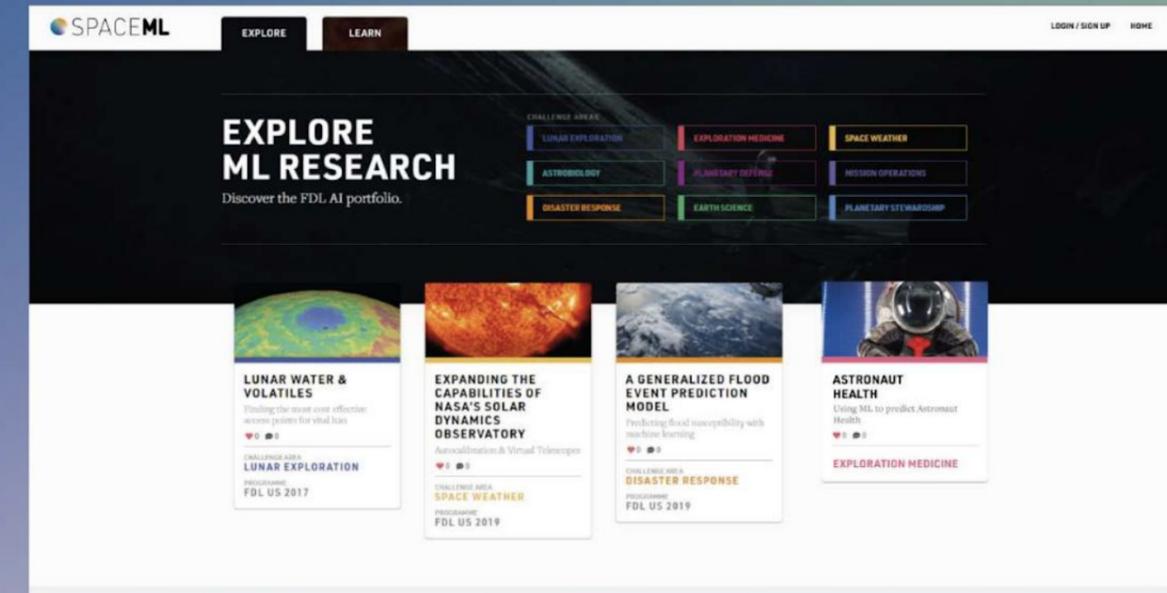
As ML precision and data collection frequency grows, so does the size and value of processed, harmonized and labeled datasets that inform research.

Here are a selection of enhanced data products coming out of FDL that we are hosting on FDL's Repo: SpaceML.org

All FDL's AI ready Data and enhanced data in one place..



SDOML	15TB
CUMULO	3.5TB
LUNA (2018)	10TB
M4G / VAEDER	40TB
SOLAR MAG / SUPER RES	10TB
INARA	130TB
US FLOODS	1TB
NEO SHAPE MODELS	1TB
GNSS PREDICTION	1TB
FLARENET	1TB



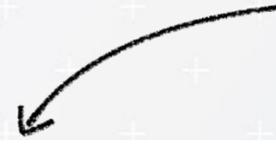
SPACEML REPO

Preprocessed data ("AI Ready"), Algorithms, Trained Models, Memos and Enhanced Data Products are to be hosted on the SpaceML repo.

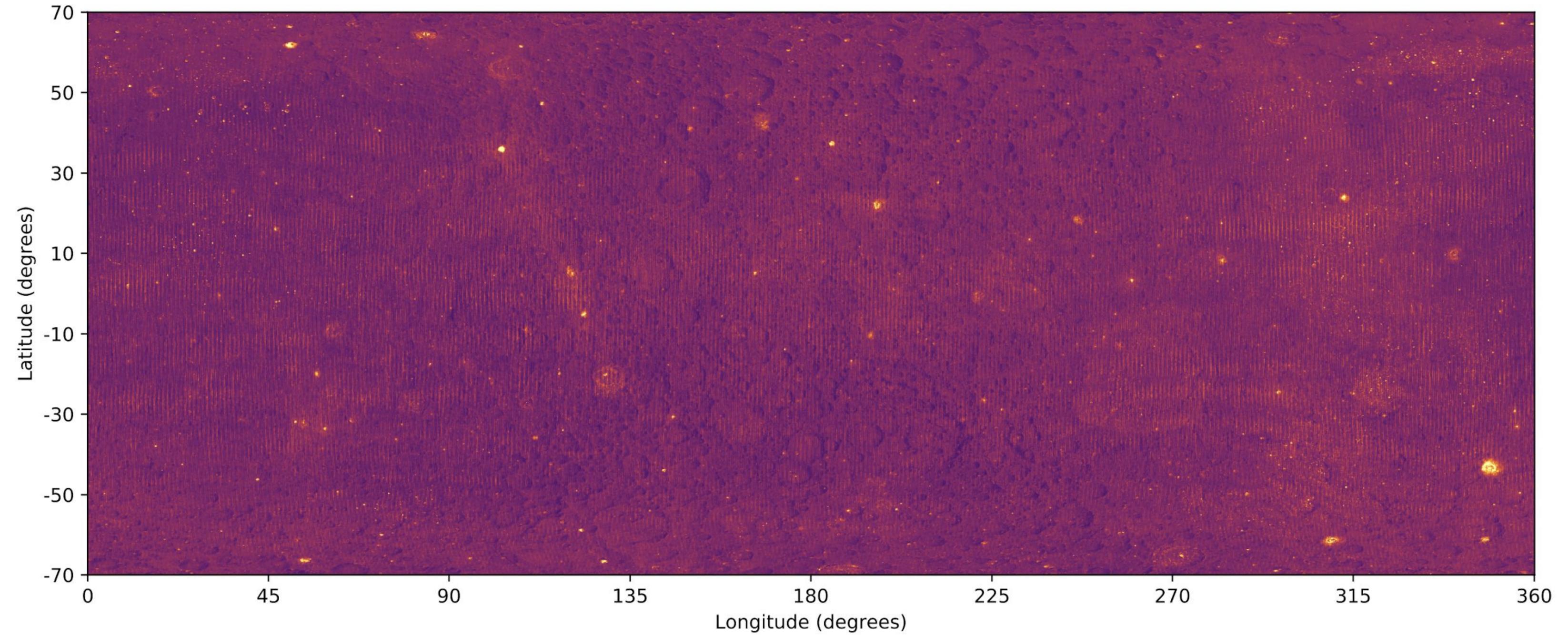


Digital Object Identifier

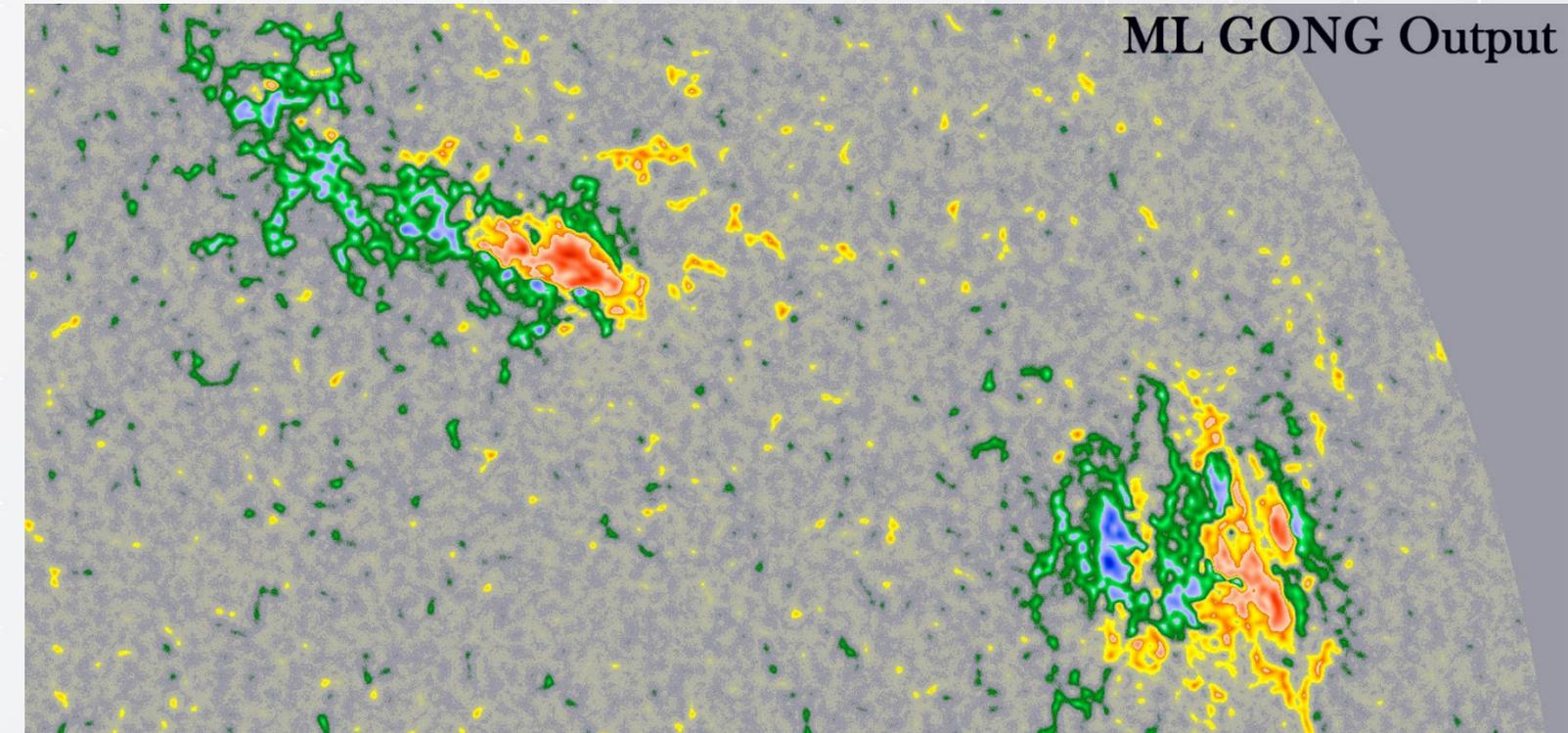
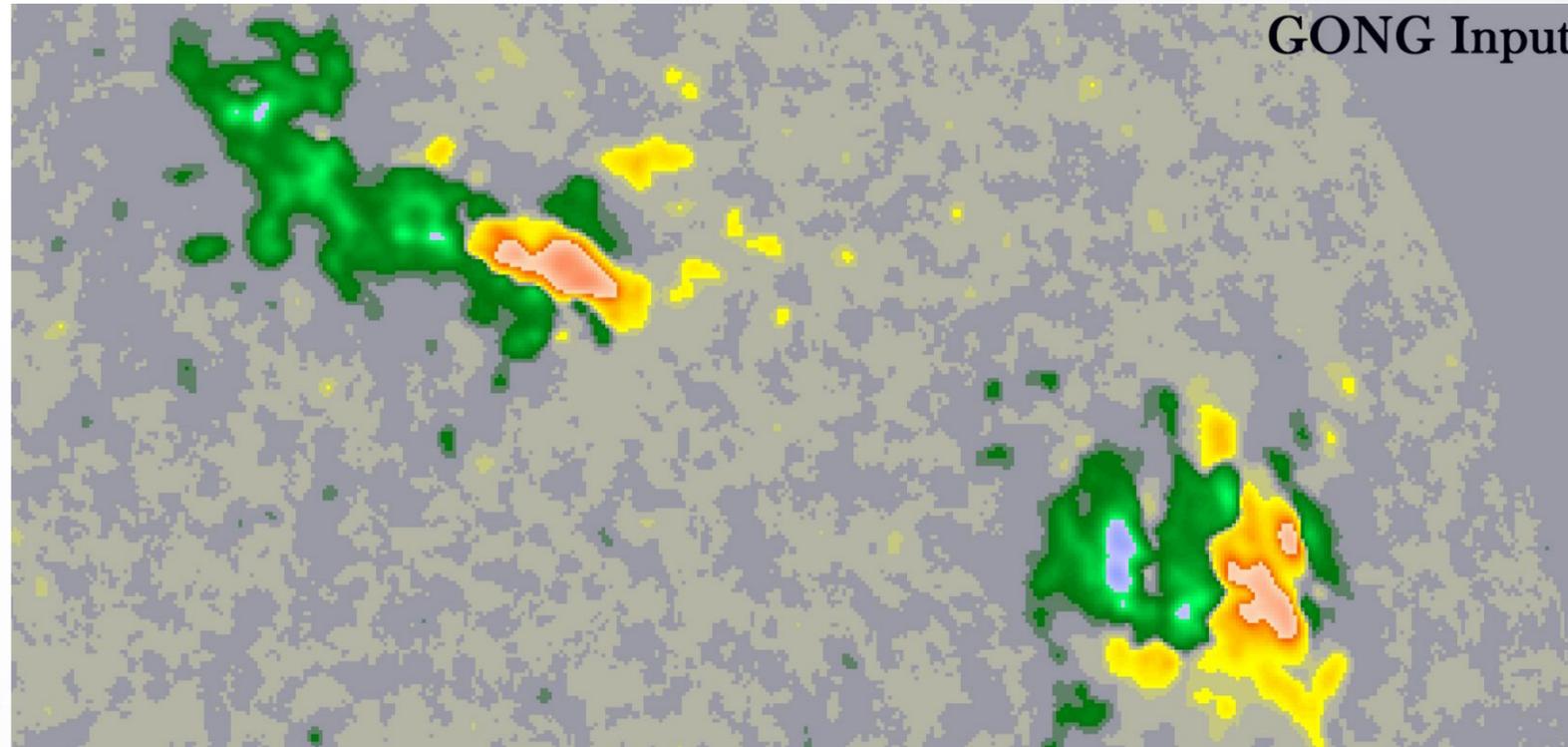
3) Do we need common standards?



ENHANCED DATA PRODUCTS



ENHANCED DATA PRODUCTS

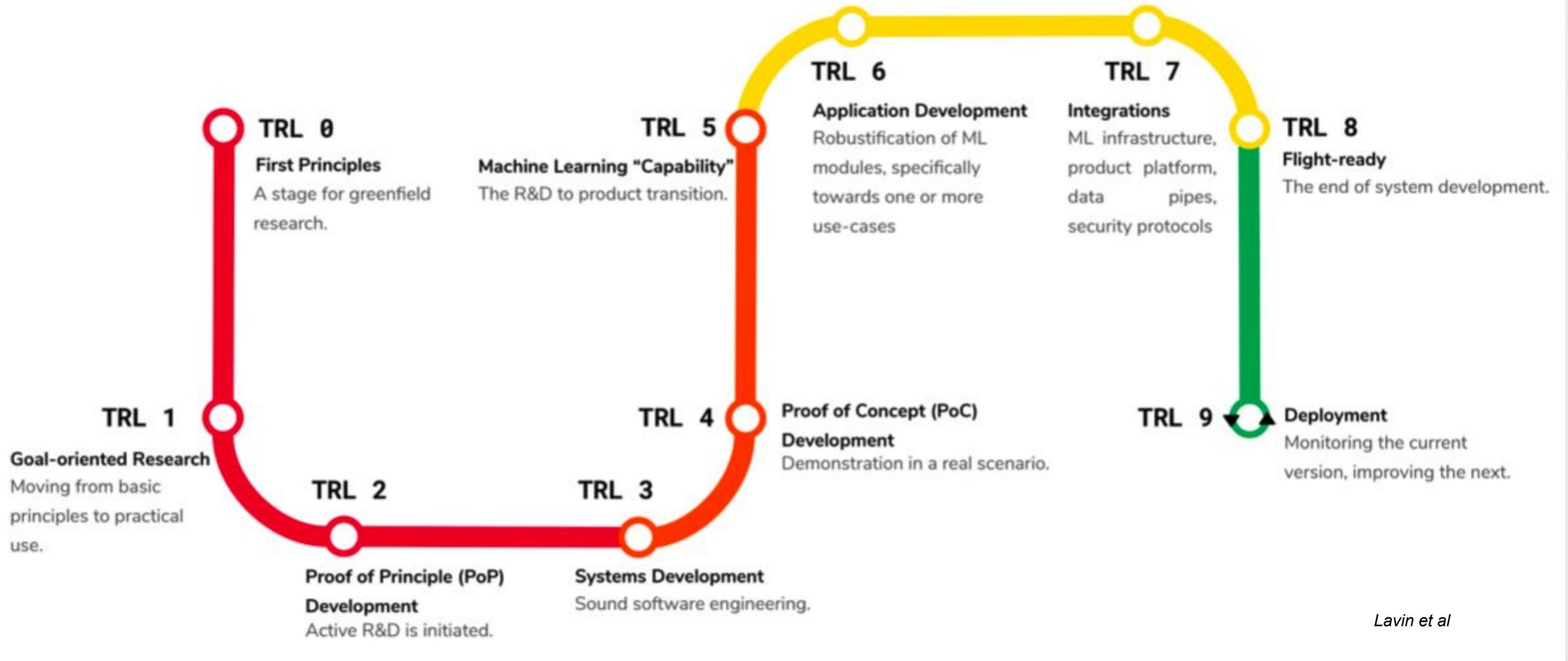


UPSCALED 4 X to create a harmonized data product of the solar magnetic field for the last 40 years.

But how do we know the maturity of the outcome? (as opposed to validation on a test training set?)

MLTRL

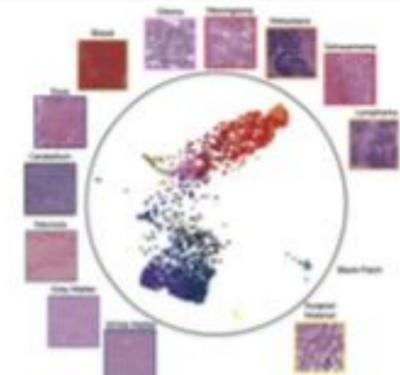
TECHNOLOGY READINESS LEVELS FOR MACHINE LEARNING SYSTEMS



MLTRL

Communication of data sources, versions and assumptions.

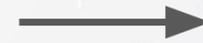


TECHNOLOGY NAME		Neuropathology Copilot v1.0	
TRL		4 <link to previous cards>	
R&D OWNER / REVIEWER		A. Lavin / G. Renard	
PROD OWNER / REVIEWER		S. Wozniak / S. Jobs	
COMPONENT CODES		1.1, 4.2, 4.3	
TL;DR	Analyze WSI of brain tissue in 3 main steps: (1) unsupervised CV model produces Poincare manifold viz (Naud & Lavin '20), (2) domain expert selects data points, (3) U-Net classifier		
Data considerations	3 datasets have been used to train and validate the system: <ol style="list-style-type: none"> 1. Open dataset (Naud & Lavin '20) 2. Pilot dataset provided by BioLab, v1.0 3. Simulated datasets (w/ structured domain randomization), v2.3 		
Ethics	Note the demographics info on specific Dataset Cards. Datasets anonymized, pipeline runs w/o metadata. The Latent Sciences Ethics Checklist has been completed.		
Model / alg details	The SP-VAE model runs unsupervised on neurological whole-slide images (WSI), producing a latent manifold that represents a hierarchical organization of tissue types. An medical expert identifies several data points to inspect.  <p><i>Example visualization of the latent organization of brain tissue types.</i></p>		
Metrics, results	Classification accuracy >0.97 on the 5 main brain cancer types. Inference per WSI runs ~1.0s on 2-GPU. Full quantitative reports: < link to experiments wiki >		
Caveats, known edge cases, recommendations	Changing imaging sources will require retraining the full model (notably the SP-VAE annealing parameter). Whenever possible it is recommended that users provide feedback annotations. Non-tissue material is correctly flagged as anomalous.		
Key assumptions	The training and production images are equivalent, specifically from the exact same sensor(s).		
Intended use	The model must include human expert in the loop, and it has not yet been validated for other disease areas.		

Lavin et al

Closing Thoughts...

Teaching machines to learn, discover and make explainable decisions is a paradigm shift.



Despite this, we aren't (yet) acting like it is when we think about data...

Meanwhile...

In the last 12 months, the same amount of ML research was published as in the prior decade...

1. <https://www.stateof.ai> 2020

And there is a reproducibility crisis in ML.

Nature: Transparency and Reproducibility in ML

<https://www.nature.com/articles/s41586-020-2766-y>

Citation impact has plateaued (but not in China¹)

Reproducibility is a strategic imperative.

Which takes us back to our three questions:

**1) What do we mean by
“AI ready”**



**2) Do we need common
quality standards?**



**3) How do we make ML
+ science simpler to
reproduce?**

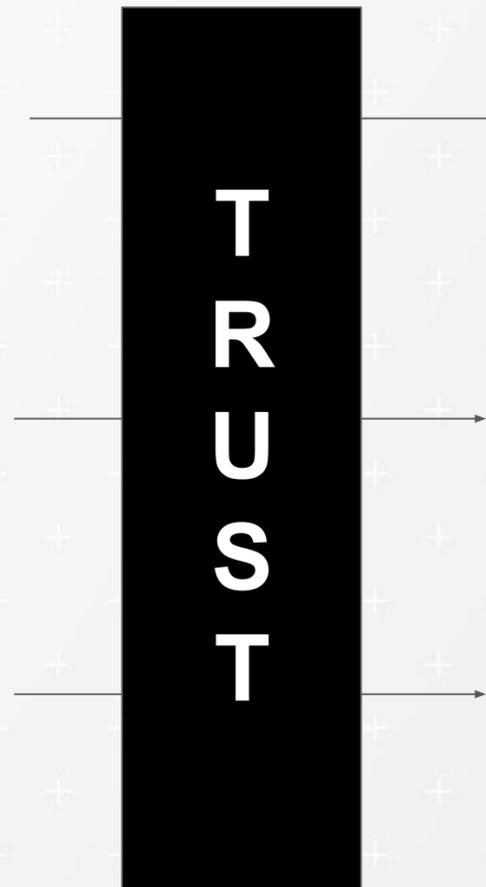


Which takes us back to our three questions:

1) What do we mean by “AI ready”

2) Do we need common quality standards?

3) How do we make ML + science simpler to reproduce?



Build Repos with easily accessible and documented AI ready data

Rally around shared ways of articulating project maturity

Capture value and democratize by taking the time (and finding budget) to build MLOPs as we go along.

Thank You

FDL.ai





ARTIFICIAL INTELLIGENCE
RESEARCH FOR SPACE
SCIENCE, EXPLORATION
AND ALL HUMANKIND

<http://fdl.ai>

BACKGROUND

Advances in computing and machine learning (ML) are revolutionizing how we do science, opening up avenues of research that would have been impossible a few years ago. *However ...*

The **opportunity cost** to apply machine learning effectively can be high. 'Garbage in, garbage out' applies equally ML and, if applied blindly, complex ML workflows can **seriously exacerbate flaws in data**. Finally, ML is sometimes regarded as a 'dark art' by non-practitioners and **explaining why ML works can be difficult**.

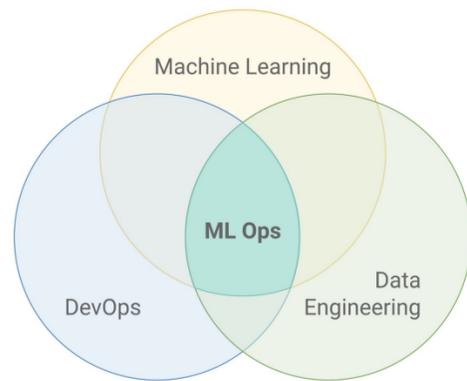
However ...

During five years of FDL, we have learned the formula to overcome these problems:

AI-ready data

Common language and quality standards

A validated framework of MLOps tools

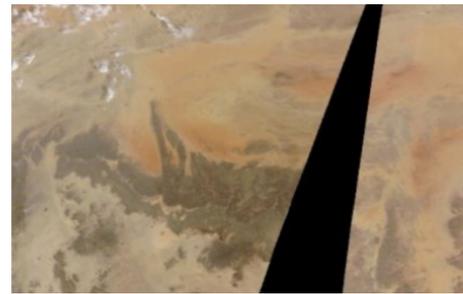


Best practices in sharing enhanced data products and machine learning algorithms: learnings from NASA Frontier Development Lab

James Parr, Madhulika Guhathakurtha, Bill Diamond

AI Ready Data

ML algorithms are great at finding 'features' in data and using them to make predictions. However, they can also be misled by flaws.



Automatic Swath Filler

This image adjustment tool developed as part of FDL automatically reduces the effect of missing imagery data.

ML systems can be rigid in how they accept data, which must be transformed into the right format. Supervised ML also requires labelled data with balanced properties.

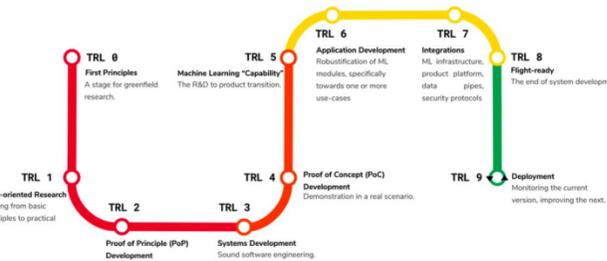
WorldFloods

The WorldFloods dataset contains carefully chosen and balanced Earth observation images, designed to train ML models to recognise floodwater.



Common Standards

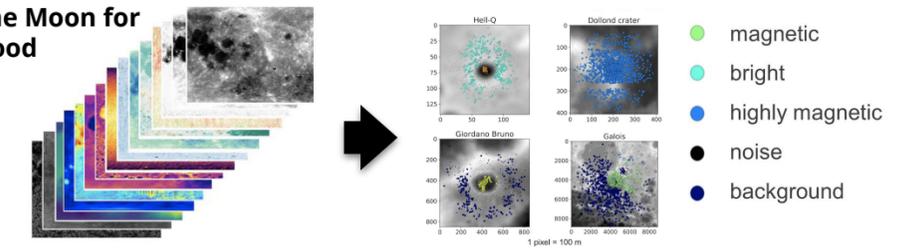
We have collaborated on a new 'ML Technology Readiness Level' that encourages development of robust, reliable and responsible ML systems.



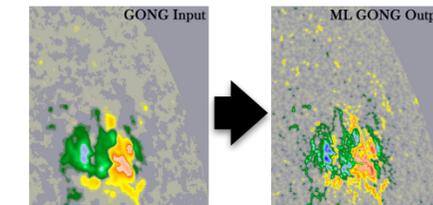
Advantages of AI

ML systems also have the power to fuse vast amounts of data into multi-dimensional stacks, and automatically decide which features are most important to the science.

The Moon for Good



ML techniques like 'super-resolution' can encode prior knowledge of physics or data properties and use these to make predictions from sparse or incomplete data



ML-Enhanced SDO

Upscaled (super resolution) of the solar magnetic field to create 40 years of data at contemporary resolutions.

MLOps and Open, Reproducible Science

Scientific culture is moving to expose all steps in the investigation process - conception, investigation, experiment and reporting. We are developing a platform that supports these 'open science' goals to share data, algorithms, code and documentation.



The SpaceML.org platform is offered as a repository of all FDL outputs, and as a resource to the scientific and ML community.

ABSTRACT