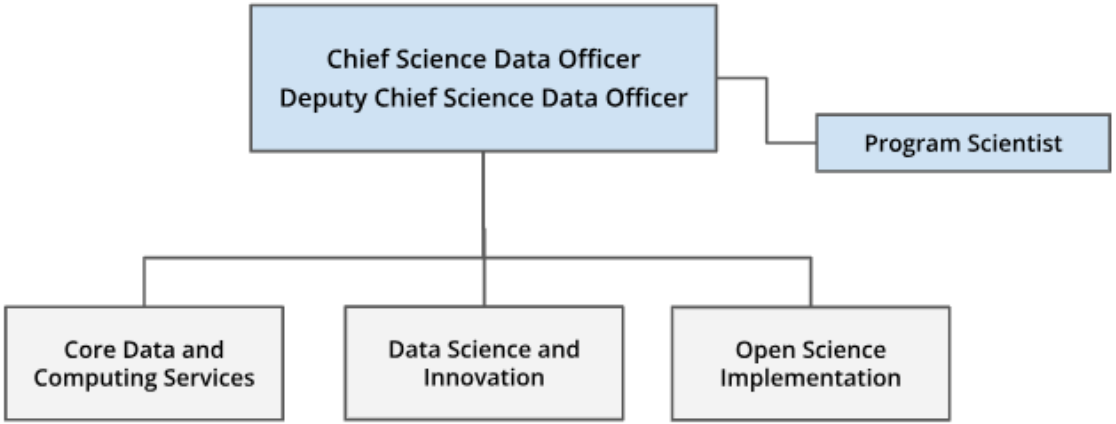
AI Foundation Models for NASA Science: *A Culture of Openness*

Manil Maskey, PhD
Office of Chief Science Data Officer, NASA HQ
IMPACT, NASA MSFC

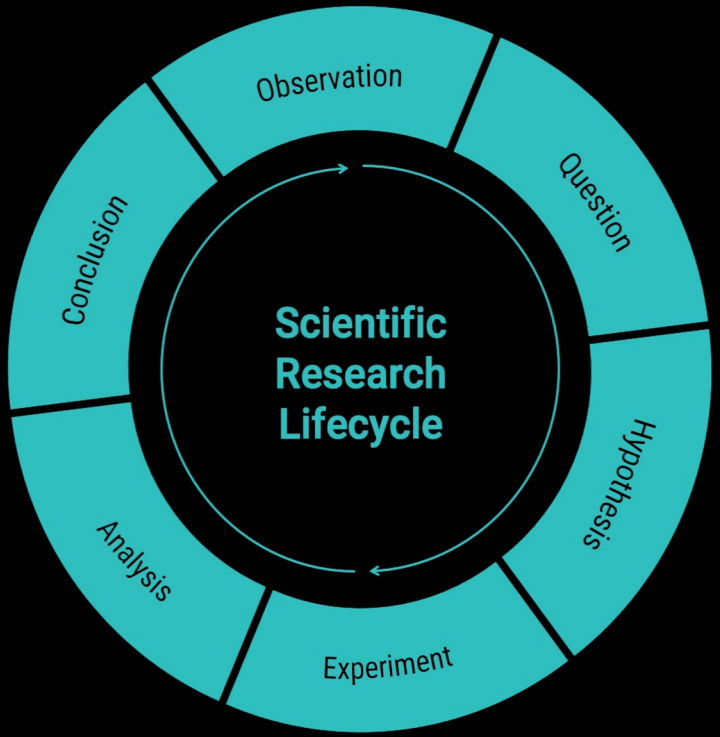
IEEE SMC-IT/SCC 2024:
Trustworthiness of Foundation Models and What They Generate

July 15, 2024

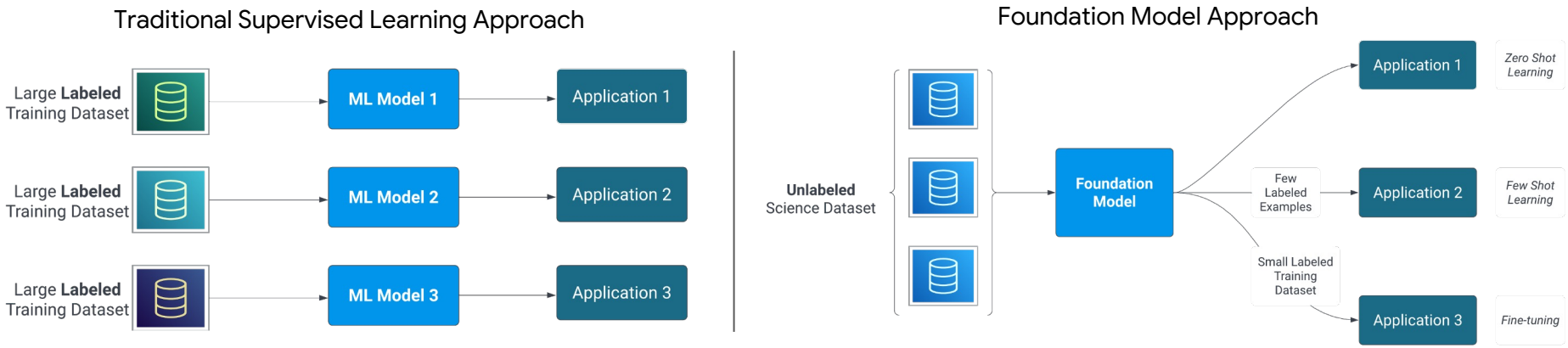
Office of Chief Science Data Officer (OCSDO)



AI can be infused in every phase of the science process



AI Foundation Models



An AI Foundation Model is a large machine learning model pre-trained on a vast amount of data using self supervision, enabling it to perform a wide range of tasks.

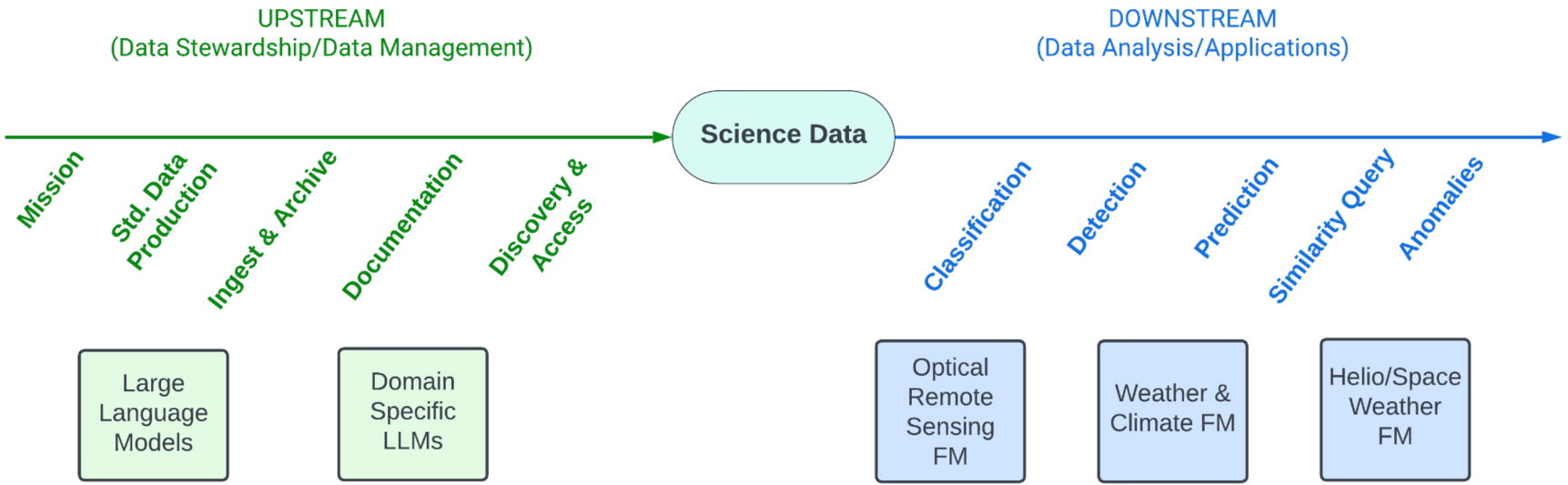
Relevance for SMD

- Helps SMD maximize the potential of archives
- Reduces effort for building AI applications
- Applicability across science division datasets

An SMD level approach was developed based on:

- Strategy for Data Management and Computing for Groundbreaking Science Recommendations
- Data and Computing Architecture Study Recommendations
- 2020 NASA AI Workshop Findings

AI adoption both upstream and downstream



SMD 5+1 strategy

AI/ML 5+1 Strategy

What is the 5+1 strategy

- Development of AI foundation model (FM) for each of the **5** science divisions utilizing high profile datasets (defined by division & science priority)
- Development of a Large Language Model (LLM) for science that addresses use cases across all divisions
- Divisions drive science use cases and evaluation
- OCSDO provides AI expertise, System Engineering, Infrastructure, and Training

Why

- Use modern data science methods to accelerate scientific discovery and application development
- Reduces barriers for AI application development with less data and computation requirements

How

- Form interdisciplinary team that includes science, data, and foundational AI expertise
- Work with division program scientists
- Partner with IBM, MSFT, NSF/NAIRR, DOE, Julich Supercomputing Center (OCSDO has access to limited resources)
- Train community to leverage the benefits of new models

AI/ML 5+1 Strategy implementation

- SMD AI Workshop confirmed viable use cases for each division
- Earth science geospatial foundation model demonstrated FMs can be used for spatio-temporal information - now being scaled for 8 years of HLS datasets in collaboration with Julich Supercomputing Center – to be released in July
- Weather/Climate FM in collaboration with ESD, DoE, IBM, and universities released in May, 2024
- Develop SDO FM in collaboration with Heliophysics division
- Multiple applications of SMD LLM in collaboration with BPS, Planetary, ESD.

Prithvi: Geospatial Foundation Model (HLS dataset)

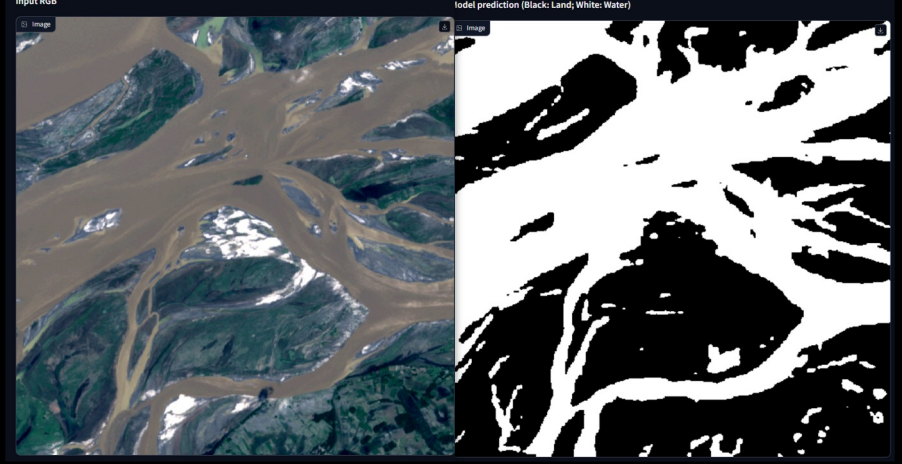
Merging Sentinel-2 and Landsat data

2-4 day global coverage

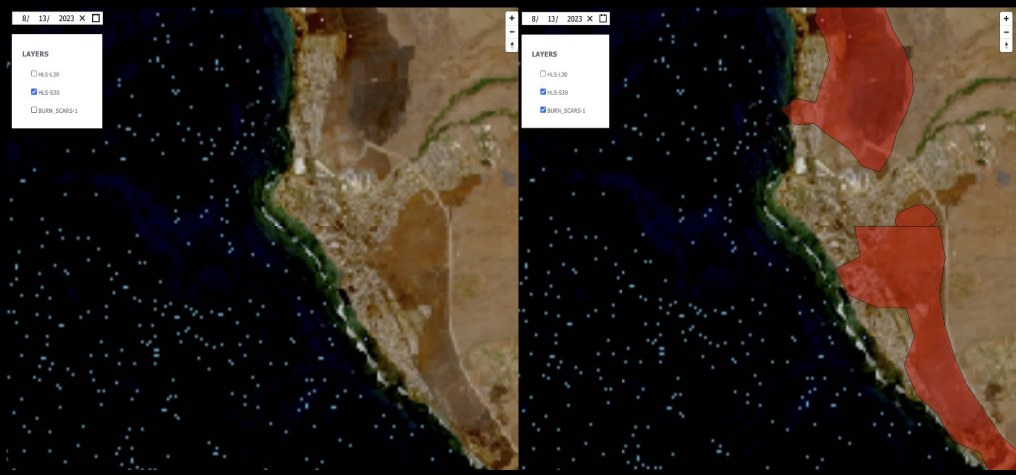
~**8000** citations

Harmonized Landsat Sentinel-2 (HLS)

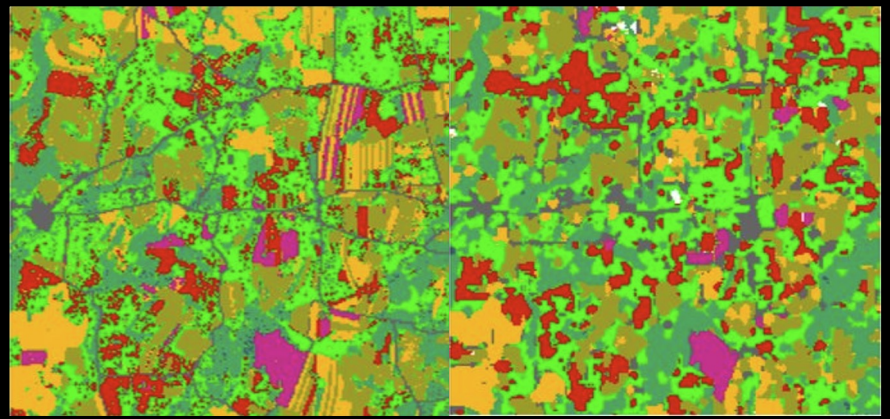
Video Credit: Scientific Visualization Studio



Water extent segmentation after floods on Sentinel-2



Burn scar mapping after Maui Fire 8.13.2023



Multi-temporal crop classification

Predicting desert locust breeding grounds

Yusuf et al. 2024

EARTHDATA
AI-powered Earth Insights BETA

Welcome Explore About

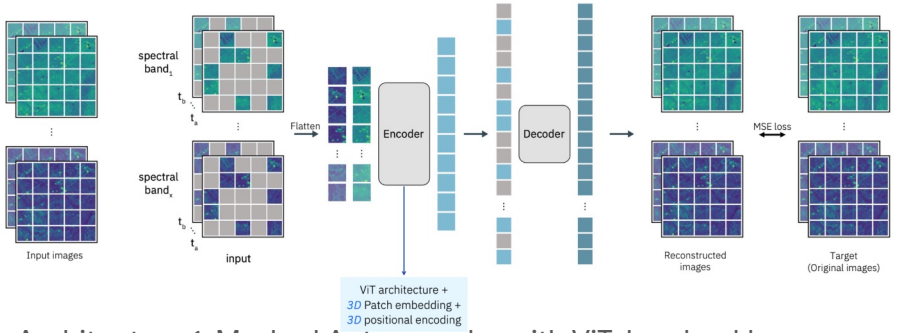
Welcome to

AI-powered Earth Insights

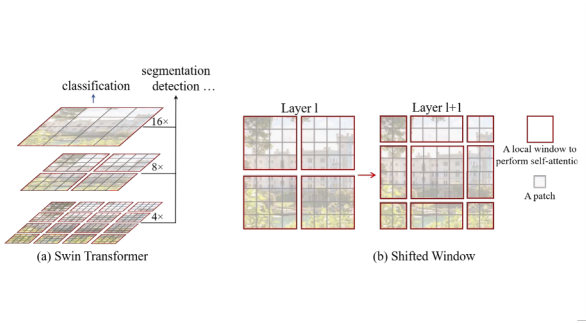
AI-powered Earth Insights is a system that leverages the first of its kind open-source geospatial AI foundation model developed by NASA and IBM Research. It uses the Harmonized Landsat Sentinel-2 Foundation (HLS) data and models that are fine-tuned on Flood mapping and Burn scar segmentation tasks. It allows users to inference on the fine-tuned models and visualizes the results.

ABOUT START EXPLORING →

Geospatial Foundation Model: Scaling Prithvi



Architecture 1: Masked Autoencoder with ViT-L as backbone



Architecture 2: Swin 3d Architecture*

Architecture:

- Training model on Swin 3D architecture over 44 nodes.
- Train previous model Prithvi with ViT-L over global data.
- HLS (trying generative capability of model - given x input time steps give me next y timesteps, where y > 3x

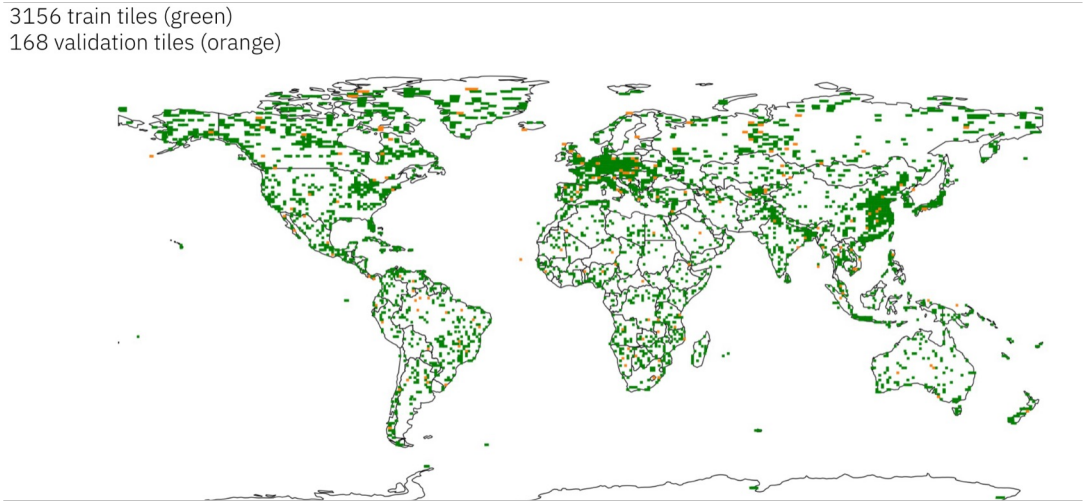
Yang, Yu-Qi, et al. "Swin3d: A pretrained transformer backbone for 3d indoor scene understanding." *arXiv preprint arXiv:2304.06906* (2023).

Sampling strategy - LULC with entropy

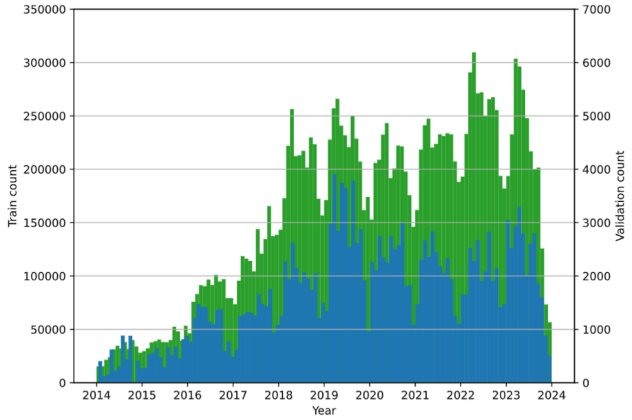
Data 2015 - 2024 HLS

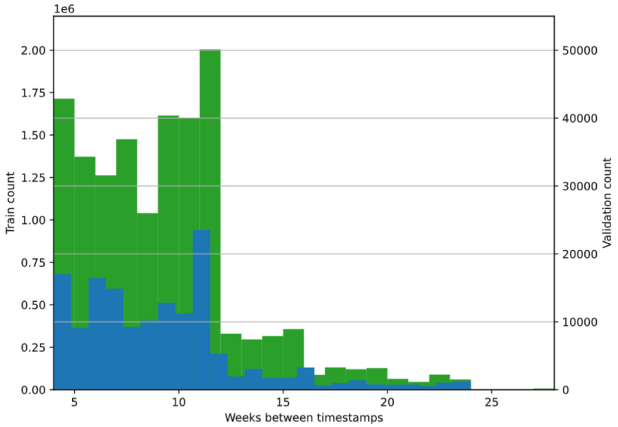
4.8 mil samples

3156 train tiles (green)
168 validation tiles (orange)



Temporal distribution



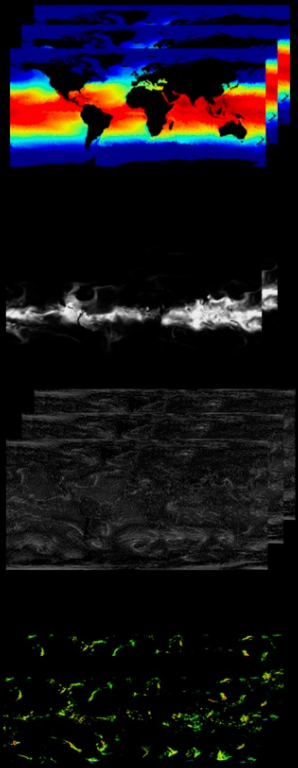


Credit: IBM Research & NASA IMPACT

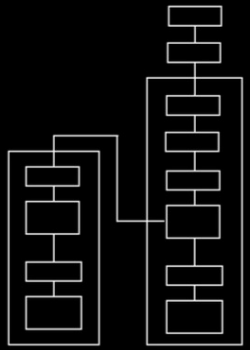
Prithvi: Weather and Climate Foundation Model

Observations

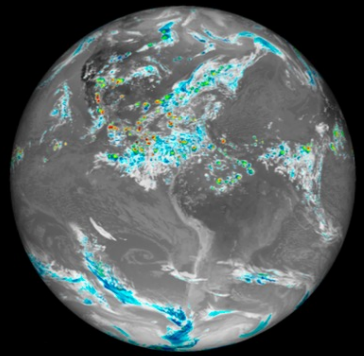
Reanalysis



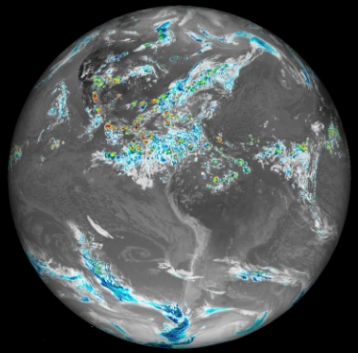
Pre-training



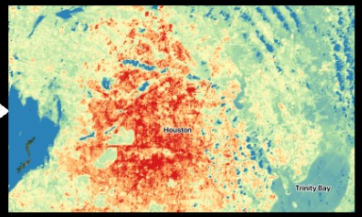
Foundation model



Foundation model

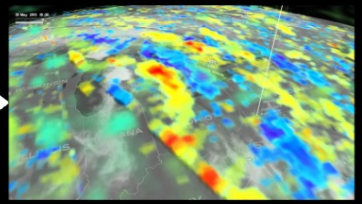


Adaptation with small data and neural network

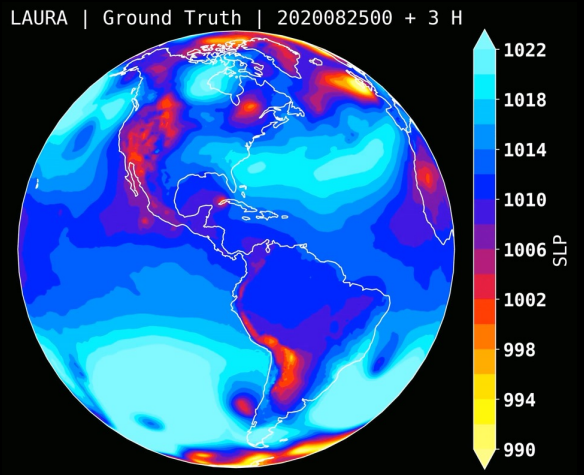


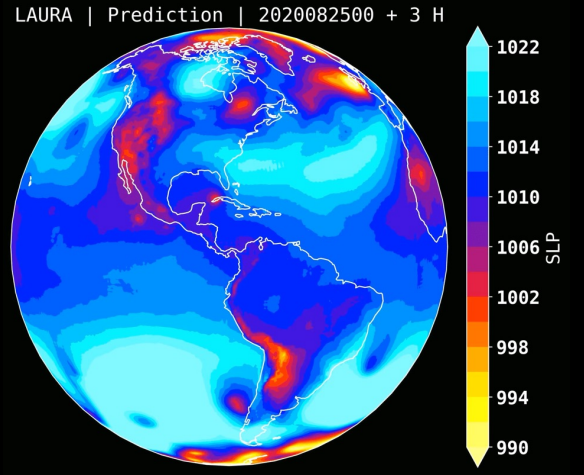
Weather forecasts

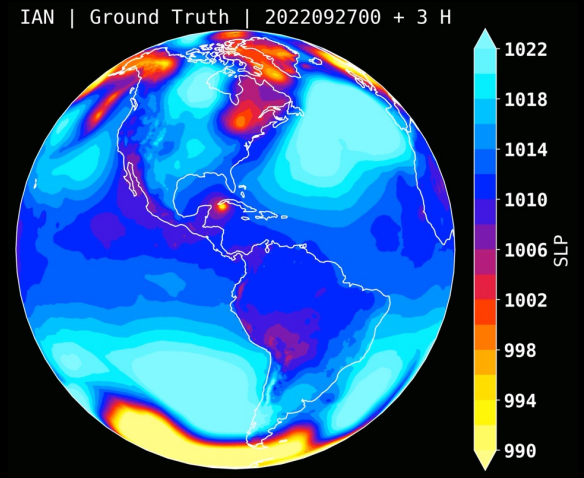
Hurricane prediction

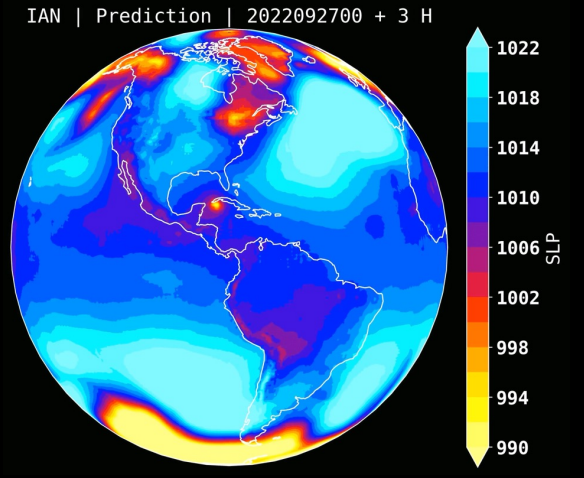


Gravity wave parameterization


LAURA | Ground Truth | 2020082500 + 3 H
1022
1018
1014
1010
1006
1002
998
994
990
SLP

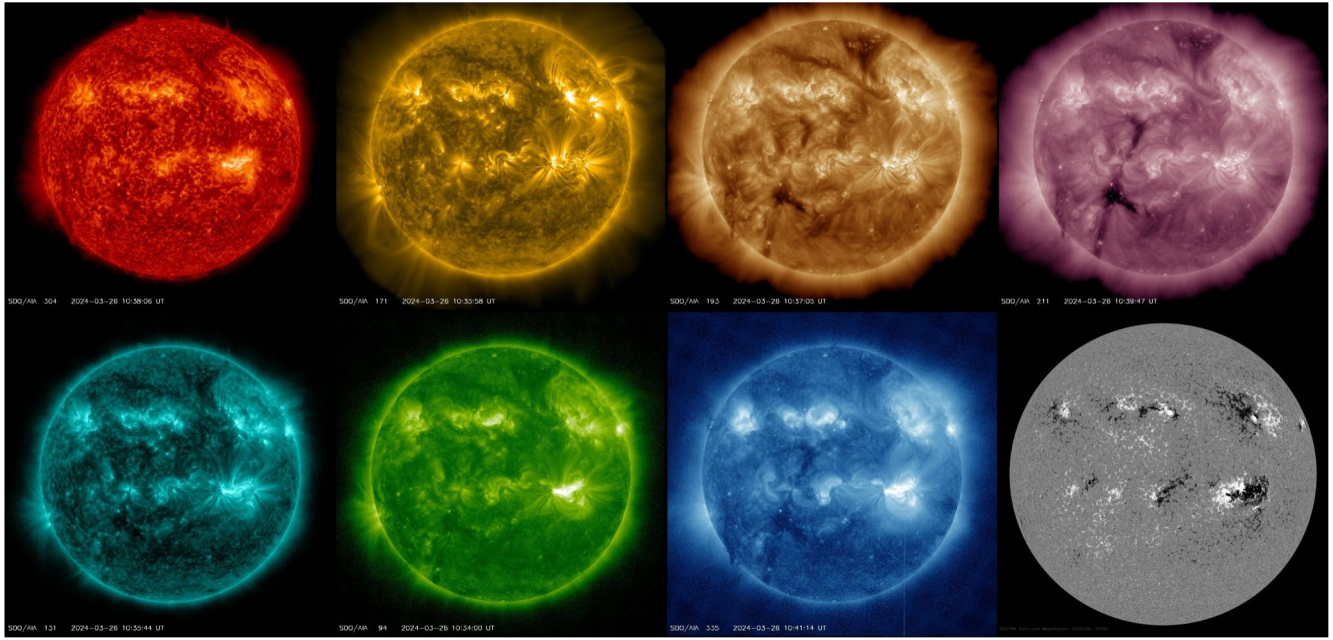

LAURA | Prediction | 2020082500 + 3 H
1022
1018
1014
1010
1006
1002
998
994
990
SLP


IAN | Ground Truth | 2022092700 + 3 H
1022
1018
1014
1010
1006
1002
998
994
990
SLP


IAN | Prediction | 2022092700 + 3 H
1022
1018
1014
1010
1006
1002
998
994
990
SLP

Helio/Space weather

Helio/Space Wx Foundation Model with SDO Dataset



SDO/AIA

Solar Coronal EUV Images
(7 Channels)

SDO/HMI

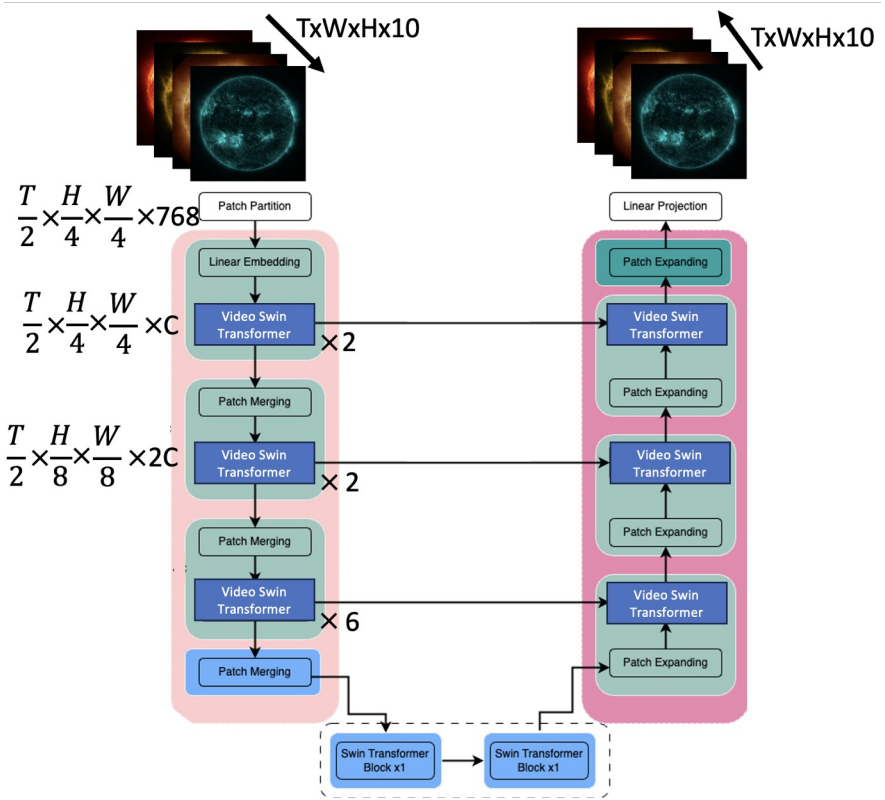
Solar Surface Magnetic Field Map

Downstream Applications

- Predicting Solar wind at Earth using EUV images
- Recreating magnetograms from EUV observations
- Coronal Identification and Tracking
- Predicting Coronal Hole evolution
- Estimate Uncertainty in Heliophysics Applications

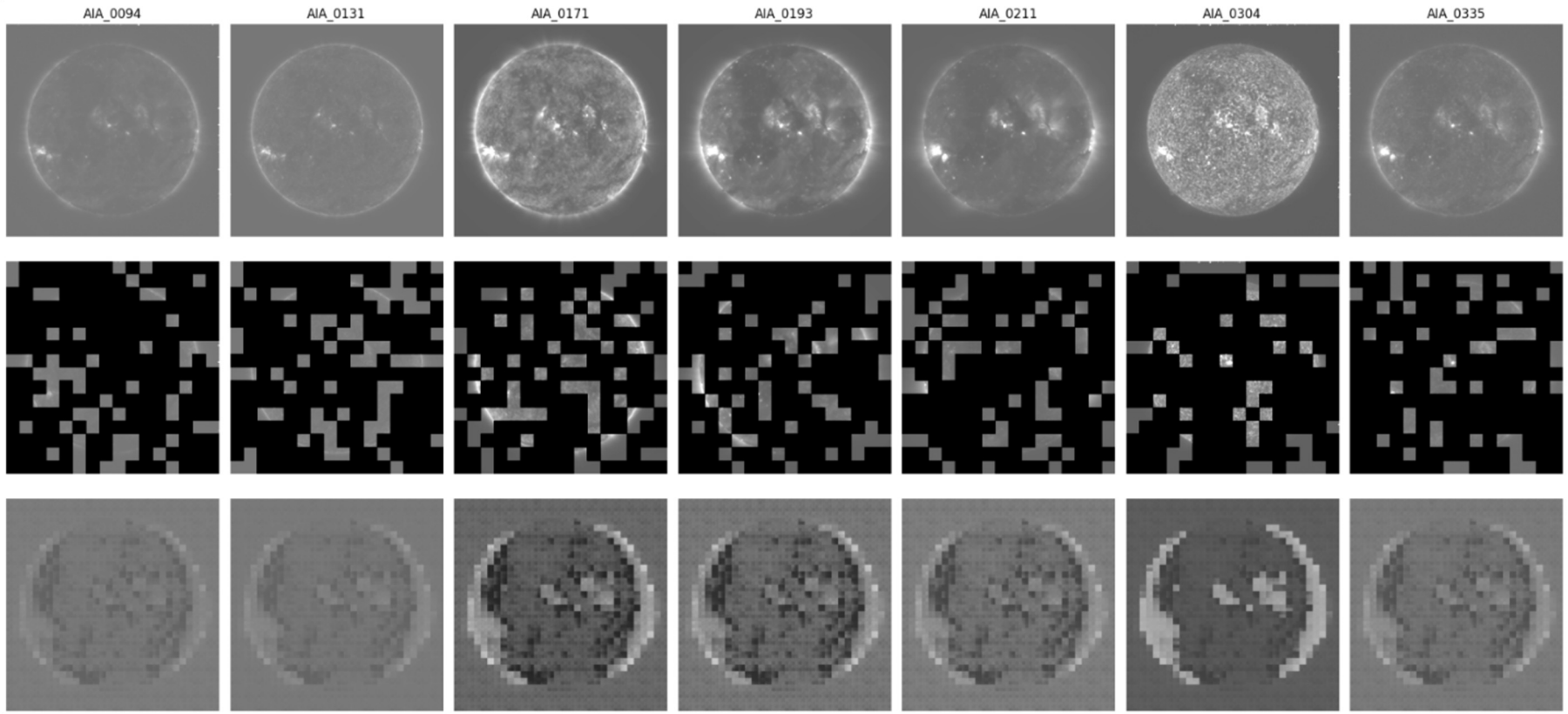
Helio/Space Wx Foundation Model Initial Architecture

Swin 3D (exp)
No. of Parameters - 1.8 B
Patch Size - 2, 16, 16
Window Size - 2, 7, 7
Masking ratio - 75%



Started verifying and capturing the flow of sun (slower scale)

Initial results



SMD Large Language Model

Indus - Language Model for NASA Science Mission Directorate

Encoder Model

Adapted for NASA SMD applications

Fine-tuned on relevant scientific journals and articles

Distilled Model

Sentence Transformer Model

Generates embeddings for information retrieval tasks

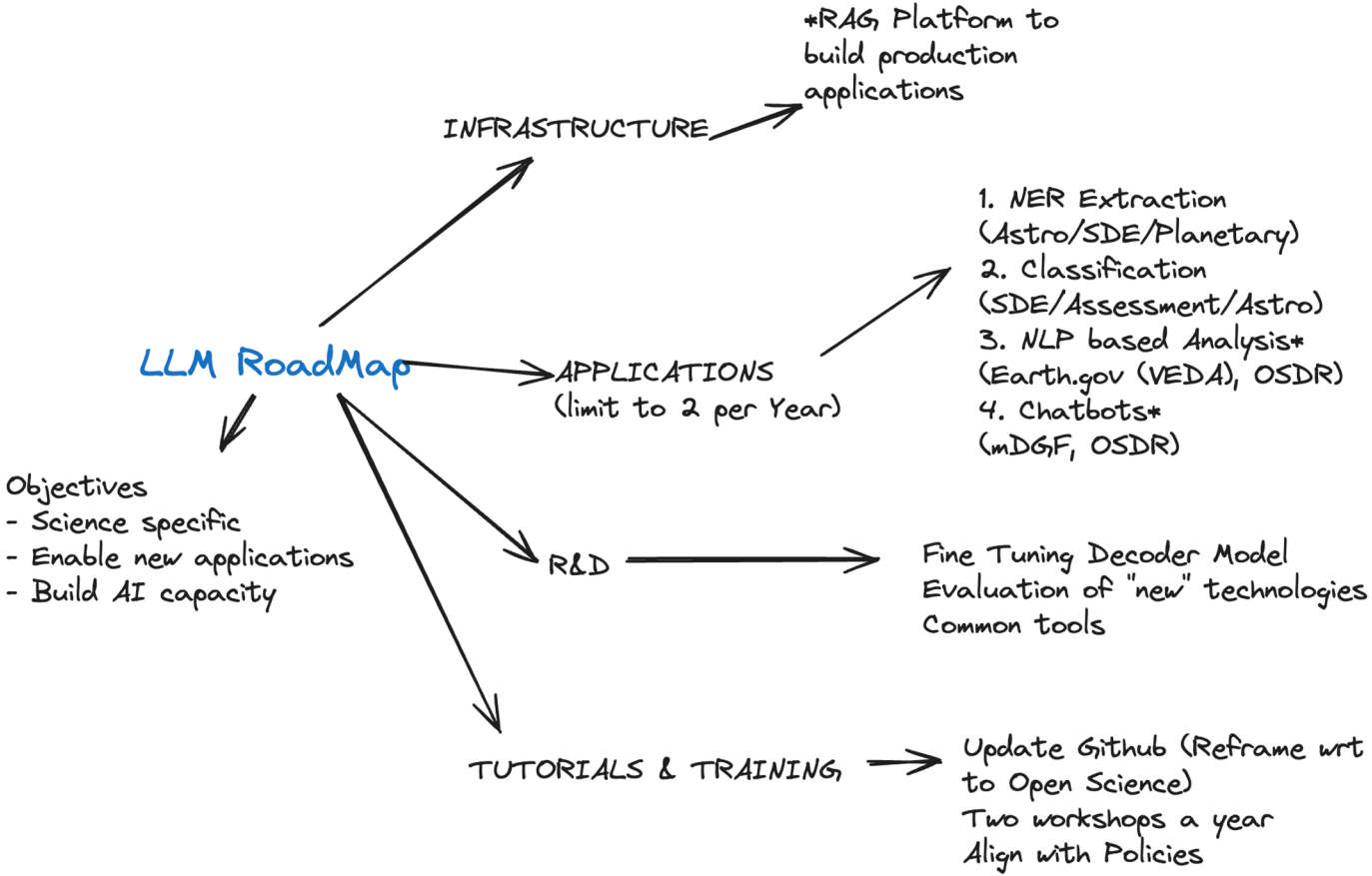
Useful in question-answering systems, document retrieval, and chatbots

Integral to the RAG workflow

Dataset	Domain	# Tokens	Ratio
NASA CMR Dataset Description	Earth Science	0.3 B	1%
AGU and AMS Papers	Earth Science	2.8 B	4%
English Wikipedia	General	5.0 B	8%
Pubmed Abstracts	Biomedical	6.9 B	10%
PMC	Biomedical	18.5 B	28%
SAO/NASA ADS	Astronomy, Astrophysics, Physics, General Science	32.7 B	49%
Total		66.2 B	100%

Curated resources for training

Path forward



Culture of openness: Principles of responsible FM development

Adhere to open science principles (Datasets, Code, Model, Responsible use guide, Research paper, Demo, etc.)

- Put high bar in terms of the quality of the work
- Enforce responsibility of the work
- Open-sourcing model early allows opportunity to improve and iterate
- Power comes from everyone using it

In science we build on someone else's work – there is desire to contribute back

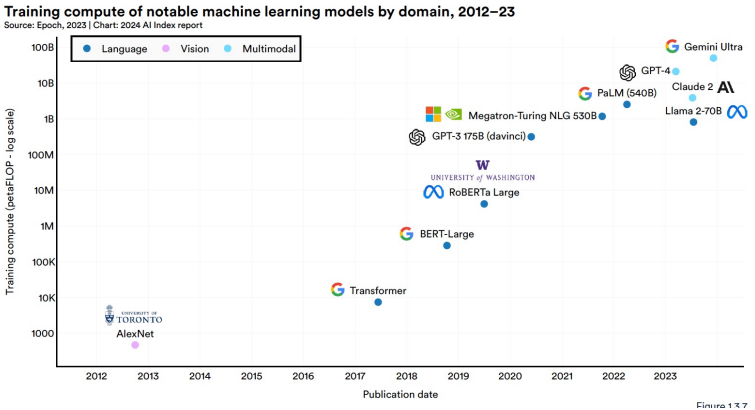
Vast field: industry has vast resources, academy has highly multidisciplinary teams, startups...

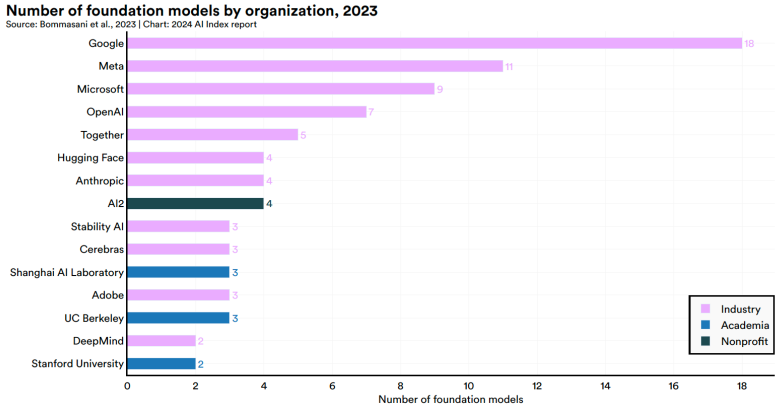
Our approach

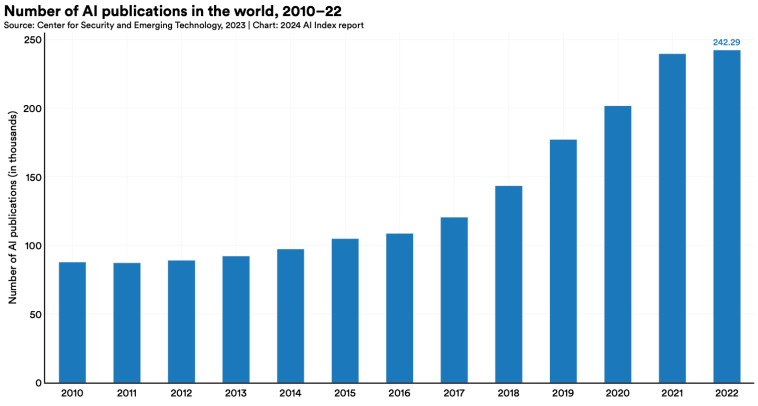
- Core focus on science and applications
- Early involvement of domain experts
- Continuous science evaluation then impact/values - benefits are multifaceted
- Inclusion of diverse stakeholders
- Extreme transparency - encourage participation and trust

Openness and programmatic perspective

- Create value to science and applications
- Avoid duplication
- Leverage everyone's unique perspectives
- Ensure responsible development
- Accelerate solutions
- Transparency and trust in science discoveries enabled by foundation models

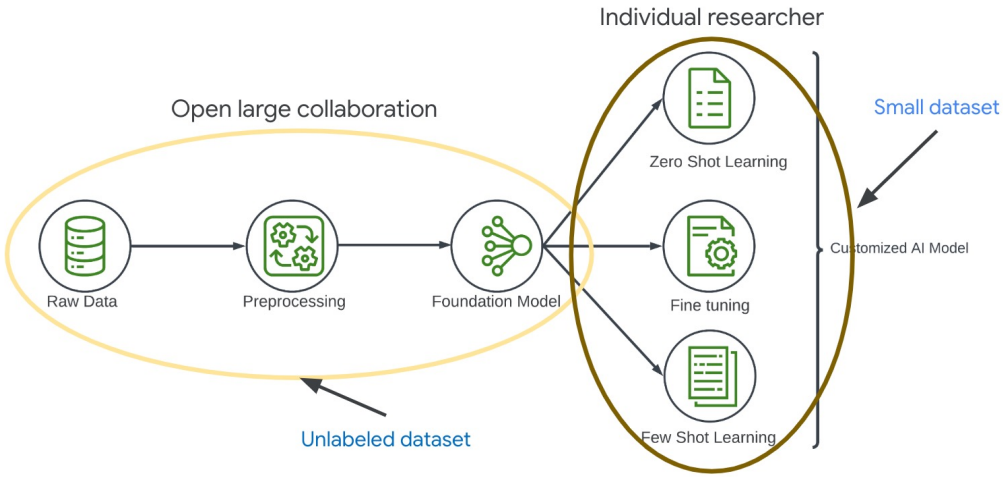






Some thoughts on data & pretrained large models

- Foundation models are more data-centric than model-centric
- Data requirement for new **custom** model is much smaller using foundation model: **Lower cost**
- Effectiveness of foundation models is directly tied to the volume of data they are trained on.
- Foundation models can perpetuate and amplify biases present in the pretraining data > need for data governance policies (ethics)
- The role of human input in data cleaning is critical for the quality of foundation models
- Foundation models require new data inputs to stay current and relevant



manil.maskey@nasa.gov

Thank you