AI Foundation Models for Science: An Open Collaborative Initiative

Dr. Rahul Ramachandran

NASA IMPACT
Marshall Space Flight Center

AGU Fall Meeting, 2023

Team

NASA MSFC/IMPACT

- Sujit Roy, Kumar Ankur, Christopher Phillips, Iksha Gurung, Muthukumaran Ramasubramanian
- Rahul Ramachandran, Manil Maskey, Pontus Olofossen, Elizabeth Fancher

NASA HQ

- Tsengdar Lee, Kevin Murphy

NASA GSFC

- Dan Duffy, Mike Little

IBM Research

- Johannes Jakubik, Linsong Chu, Paolo Fraccaro, Ranjini Bangalore, Kamal Das, Daiki Kimura, Naomi Simumba, Daniela Szwarcman, Michal Muszynski, Carlos Gomes, Dario Oliveira, Karthik Mukkavilli, Campbell Watson , Kommy Weldemariam, Bianca Zadrozny, Raghu Ganti, Carlos Costa

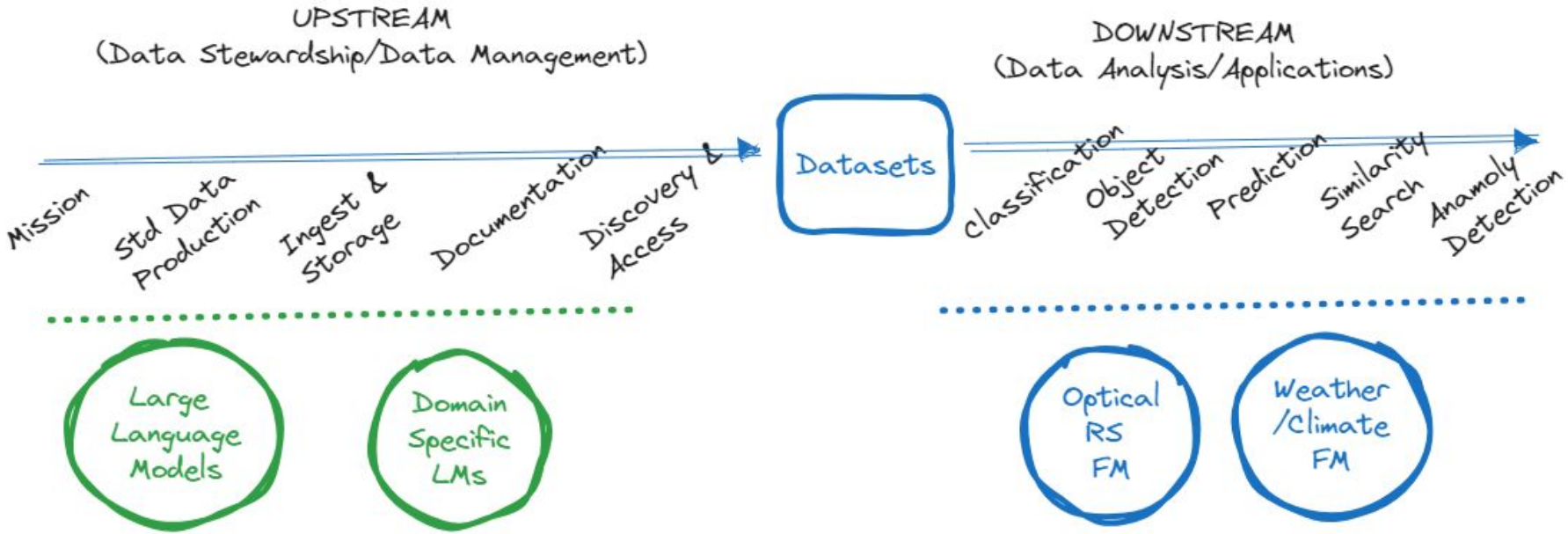
Clark University

- Hamed Alemohammad, Steve Li, Michael Cecil, Sam Khallaghi, Denys Godwin, Maryam Ahmadi, Fatemeh Kordi

Importance of AI in modern scientific research

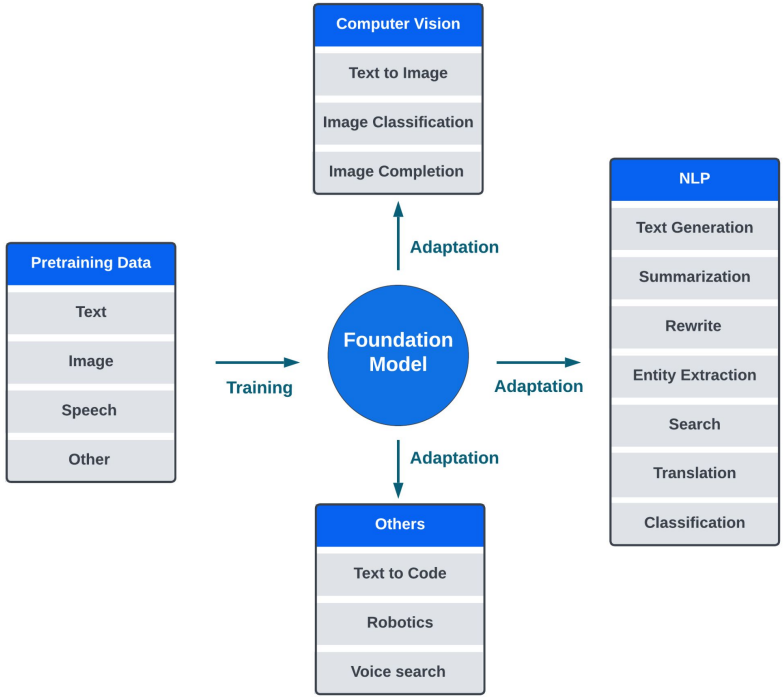
- Analyze vast datasets rapidly, leading to quicker insights and breakthroughs
  - Identify complex patterns and anomalies in data
- Provide data driven predictions in many different areas
  - Trained on historical Earth Observation data, allows ML to make predictions based on past trends and occurrences
- Augment existing processes to discover and analyze amounts of observational data from satellites, sensors, and other sources

AI Adoption Both Upstream and Downstream



What Are Foundation Models?

- Foundation models are large-scale machine learning models pre-trained on vast amounts of data, serving as a starting point for fine-tuning on specific tasks.
- Unlike traditional models that are trained from scratch for specific tasks, foundation models are pre-trained on general data and then adapted to specialized tasks. This pre-training captures broad knowledge, allowing for versatility across multiple domains.



Foundation models: training and adaptability

Advantages of Foundation Models

- Reduces the need for extensive task-specific data and training time, as the model already possesses a foundational understanding from its pre-training.
- A single foundation model can be fine-tuned for a wide range of applications, from natural language processing to image recognition.
- Due to their scale and comprehensive pre-training, foundation models often achieve state-of-the-art performance on various tasks, even with limited labeled data.

The Need for Collaboration in AI for Science

Complexity of the Problem and Vastness of Scientific Data

- Complex science problems by nature require interdisciplinary teams
- Volume and diversity of data in scientific fields require diverse expertise.

Limitations of Individual Research Groups or Institutions

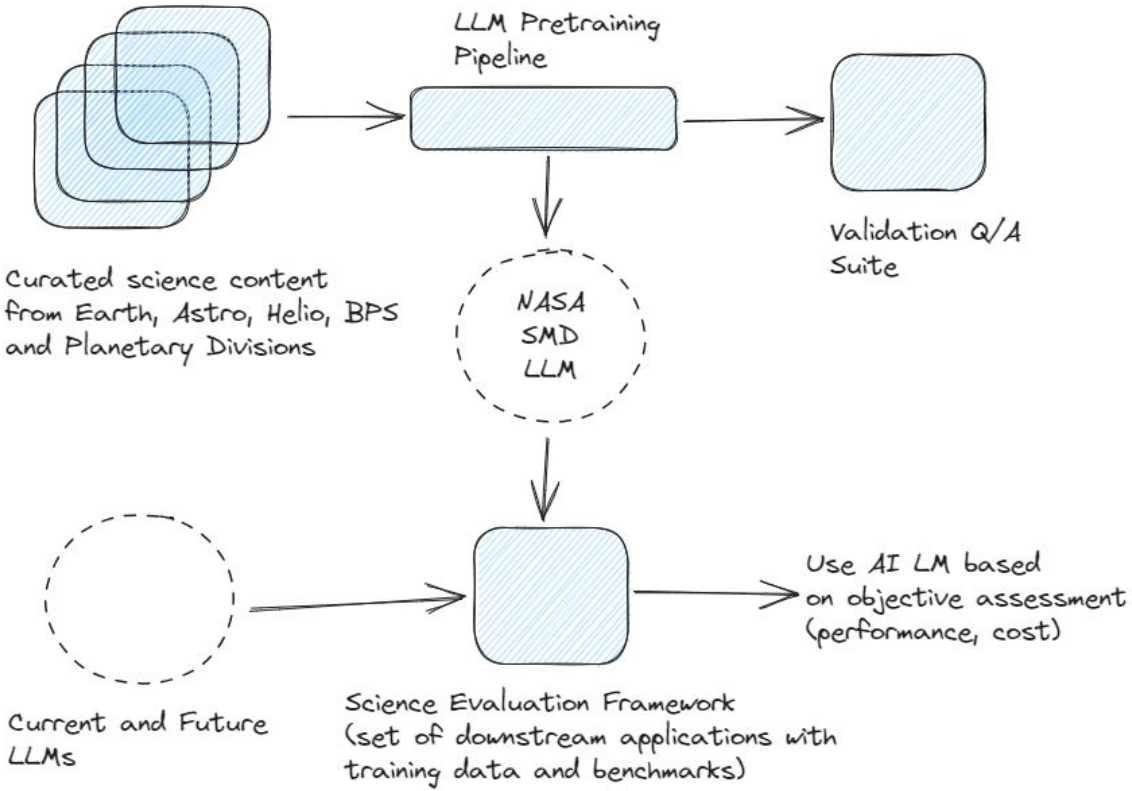
- Single entities may lack the resources or expertise to build FM.
- Collaboration allows for pooling of diverse skill sets and perspectives, leading to more comprehensive solutions.
- Need expertise from various AI subfields ensures that foundation models are versatile and can be fine-tuned for a wide range of applications.

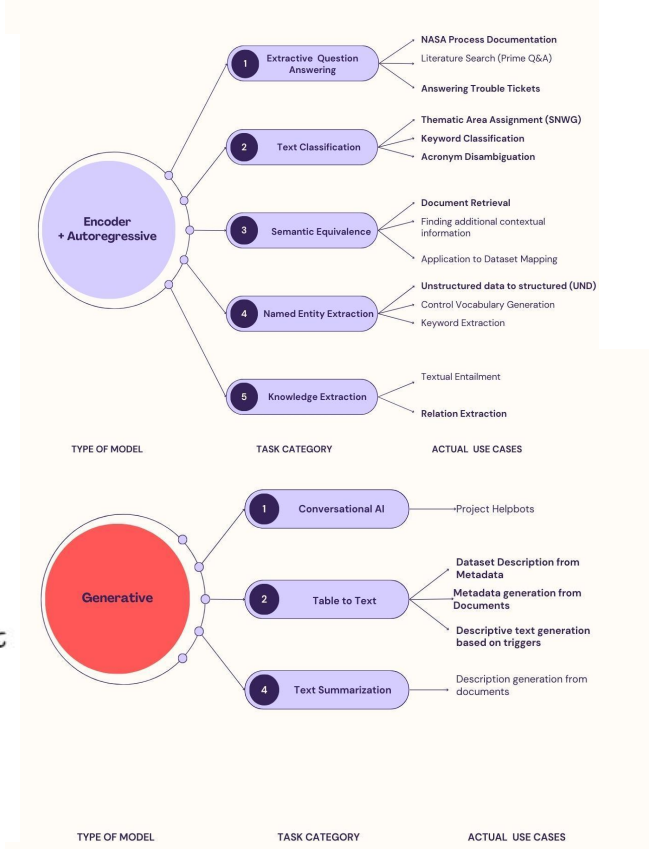
Pooling downstream use cases (labeled datasets/benchmarks) helps develop foundation models that has been validated by different groups using variety of use cases

Our Open Collaborative Approach

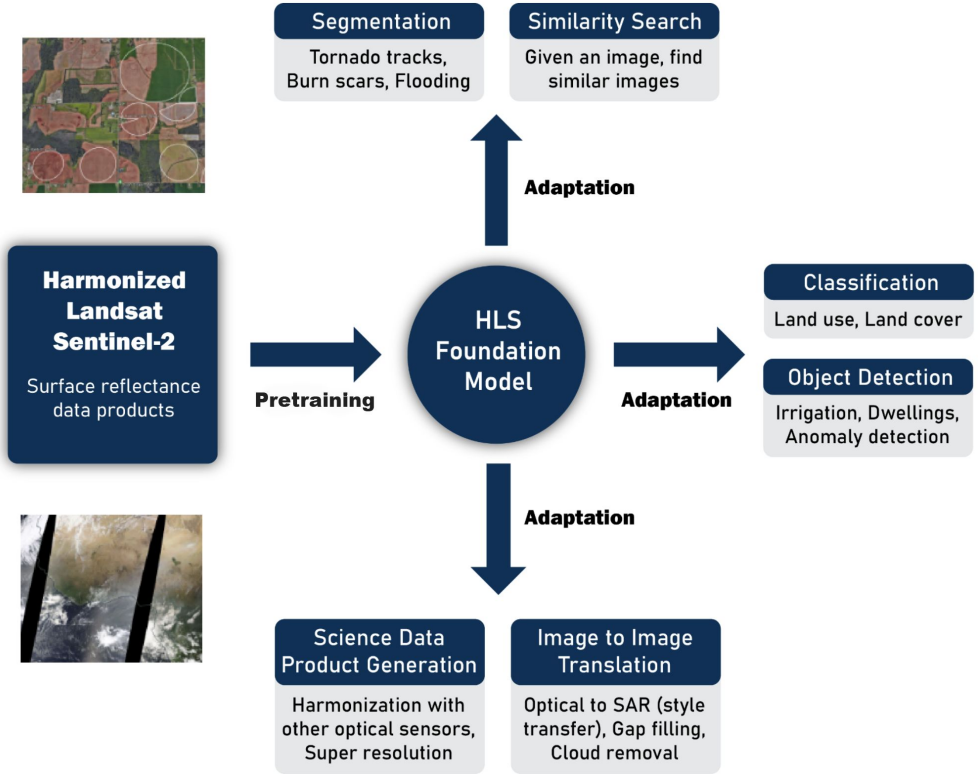
- Encourage participation from diverse groups, ensuring a wide range of perspectives in research.
- Engage Key Stakeholders:
  - Science experts dedicated to advancing knowledge in their respective fields.
  - Universities, research labs, and organizations that provide the infrastructure and support for research.
  - Tech Companies that offer technological solutions, platforms, and resources essential for modern research.
- Grounded in Open Science Principles:
  - Ensure that research is conducted transparently and that findings are shared openly with the community.
  - Promote reproducibility by making methodologies and data accessible.
    - Make models freely to use, modify, and distribute.
    - Ensure that datasets are freely accessible to validate findings
  - Promote data sharing, reducing redundancy in data collection efforts.

Case Study: Foundation Language Model for Science





Case Study: HLS Foundation Model



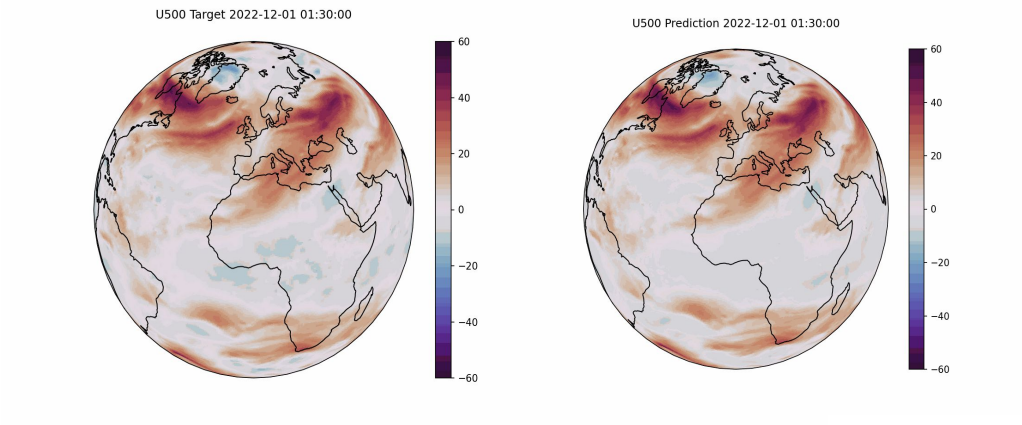
HLS FM

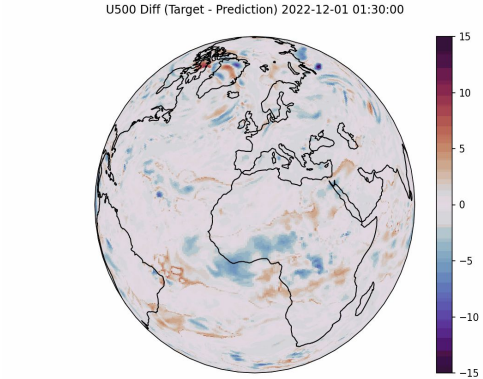
- Build with collaboration with IBM Research
- Initial version released are 100M and 300M parameter models
- Model is a Masked Auto Encoder where attention mechanism is extended in space and time
- Model is being evaluated for adaptation for different categories of downstream tasks and the initial results are very promising

Collaborators

- UAH, Clark University, ORNL, Hugging Face

Case Study: Wx/Climate Foundation Model





Playbook to build FM for Science (1)

- Access to Infrastructure
  - Availability to high-performance computing clusters, GPUs, and other essential hardware for training and deploying foundation models.
  - Scalable storage options for vast datasets and model parameters.
- Sustainable Partnerships
  - Collaboration between researchers, institutions, and tech companies.
  - Align objectives and expectations to ensure mutual benefit.
- Community Curated Use Cases
  - Curate high-quality training datasets that are representative and diverse.
  - Compile existing benchmarks that allow for consistent evaluation.

Playbook to build FM for Science (2)

- Documented Best Practices that Enable Knowledge Sharing
  - Pretraining a FM:
    - Guidelines on data selection, preprocessing, model architecture selection, and training parameters.
    - Recommendations on monitoring and evaluating pretraining progress.
  - FM Adaptation:
    - Procedures for fine-tuning foundation models on specific tasks.
  - Tips on optimizing performance and ensuring generalization.
- Systematic Outreach to Science Users
  - Organize sessions to introduce the science community to the new AI lifecycle.
  - Emphasize the shift from traditional model development to fine-tuning foundation models for specific tasks.
  - Provide tools, documentation, and expert guidance to assist researchers in adopting and benefiting from foundation models.

Summary

- AI Foundation Models have the potential to impact scientific research, offering accelerated discoveries and enhanced predictive capabilities.
- The complexity of scientific problem, data and the limitations of individual entities/groups underscore the need for collaborative efforts for AI FM.
- Embracing open science principles, open-source software, and open-access data to build these FM ensures transparency, inclusivity, and accelerated innovation.

Thank you.

rahul.ramachandran@nasa.gov