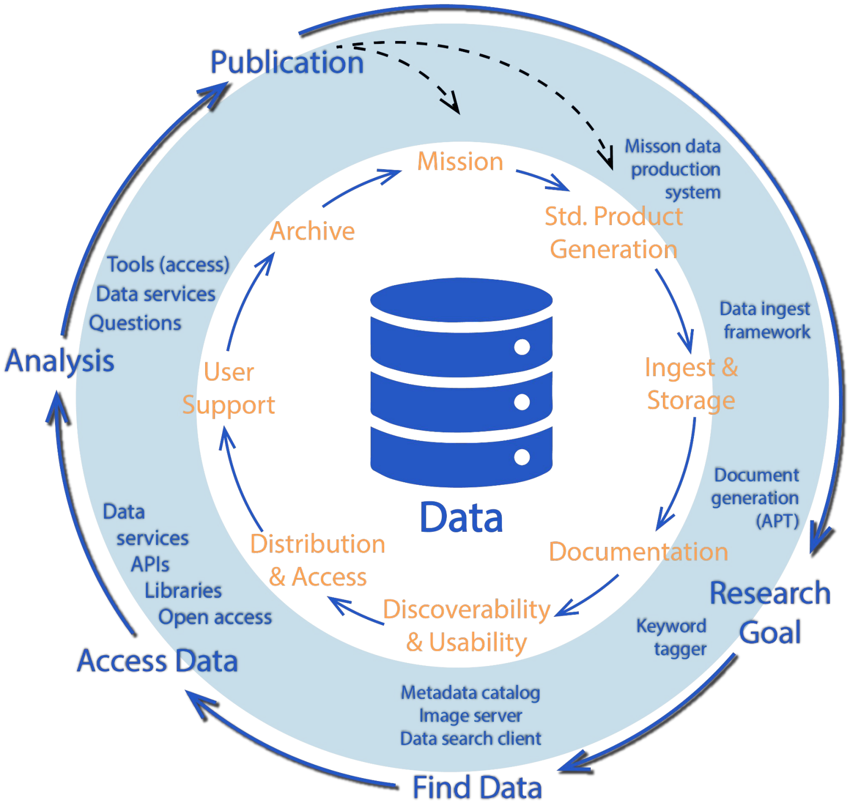
Foundation Models for Science: *Potential, Challenges, and the Path Forward*

Rahul Ramachandran, NASA/MSFC
IMPACT

Contact:
rahul.ramachandran@nasa.gov

Manil Maskey, Tsengdar Lee, Kevin Murphy
NASA
Sujit Roy, Muthukumaran Ramasubramanian, Iksha Gurung
University of Alabama in Huntsville
Raghu Ganti
IBM Research

Building Blocks for Science

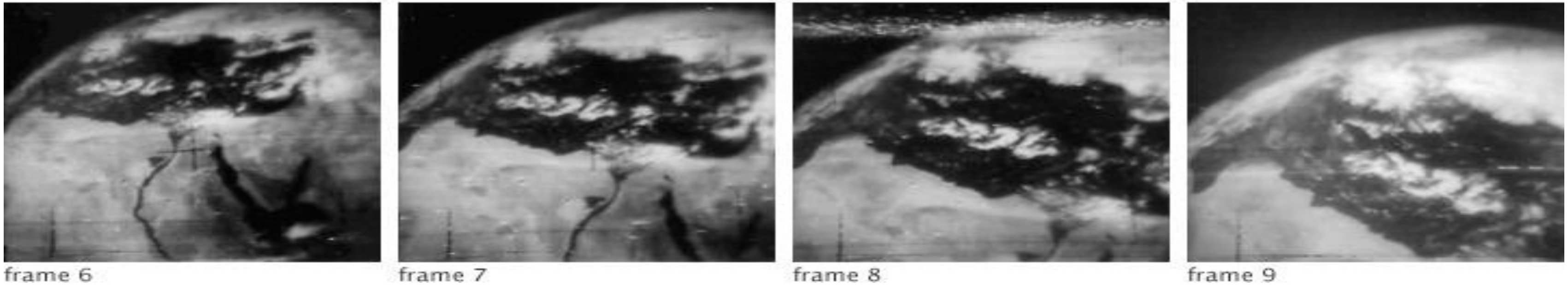


Data + Tools/Infrastructure -> Enable Research Lifecycle

Drowning in Data, Starving for Knowledge

"Torrents of data bits descend upon us from our instruments in space. How do we process the data, store them, retrieve them for scientists to use? This is our theme - how to obtain information and understanding from these data."

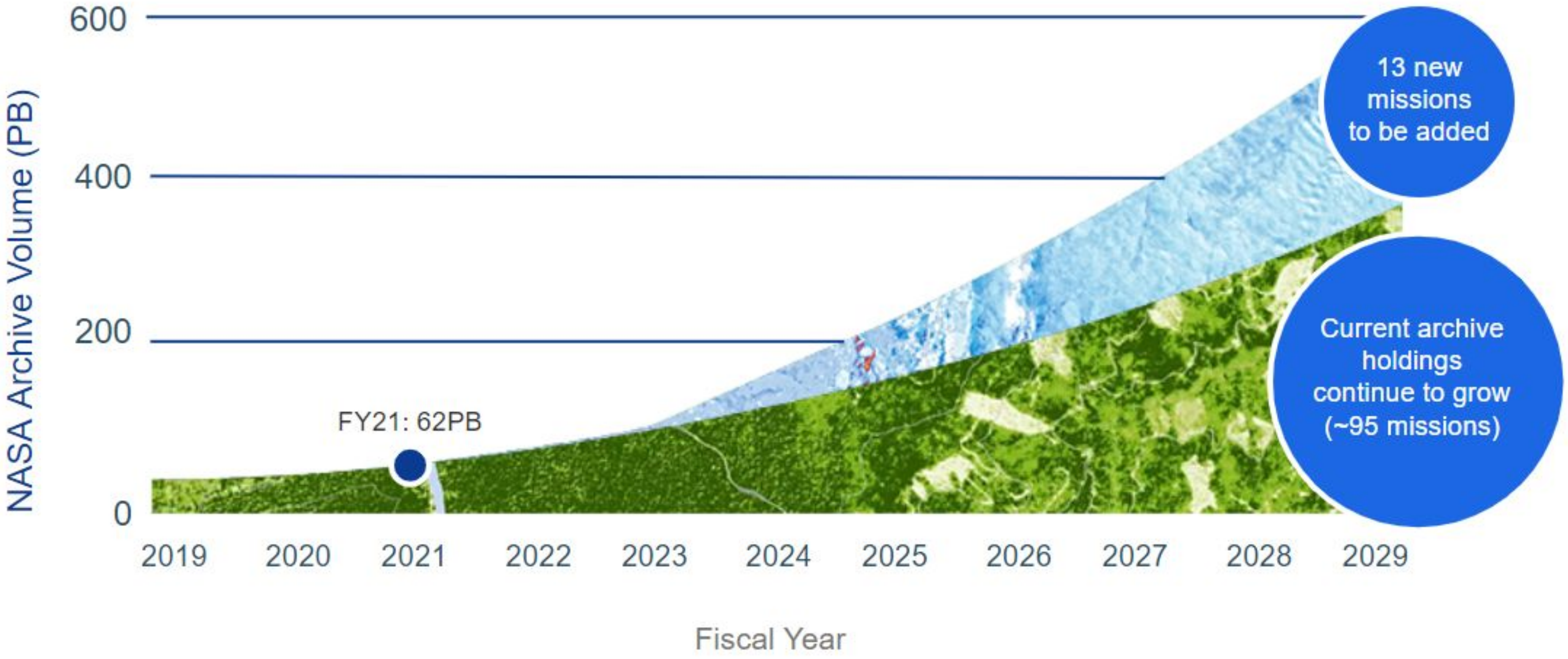
– National Research Council, 1982



Photography from TIROS 1, Launched April 1, 1960
NASA Goddard Space Flight Center

… the challenge of managing overwhelming data and distilling it into meaningful knowledge

Earth Science Data Archive Growth Projection



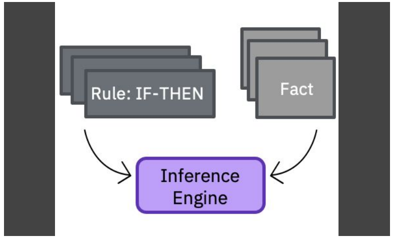
How can we *improve research process* as well as *lower barriers of entry* for all users to effectively leverage complex scientific datasets?

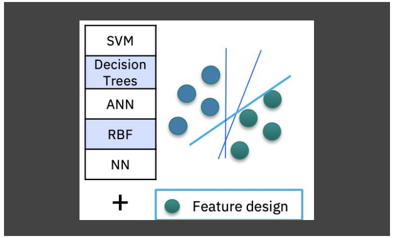
Inflection Point in AI Adoption

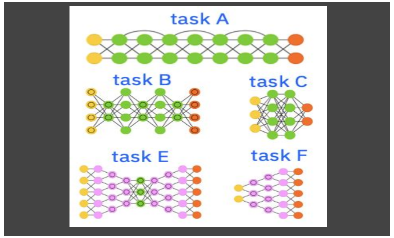
1980s | Big data

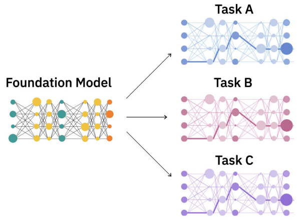
1980s to ~2010 | Massive labeled data + Compute

2017+ | **Self-supervision at scale** + Massive unlabeled data + Compute









Expert Systems
- Manually-crafted symbolic representations and rules
- No use of data and brittle

Machine Learning
- Less brittle but labor intensive
- Demanding data prep and feature engineering

Deep Learning
- Automatically learn if you have enough labeled data
- Enterprise adoption limited by availability of labeled data

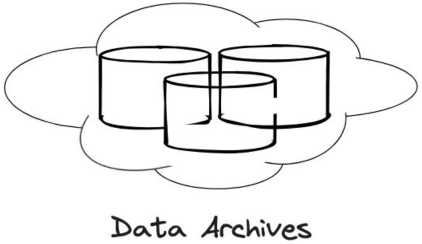
Foundation Models
- Learn from lots of data *without requiring labels*
- Quickly adopt to enterprise tasks using limited labels

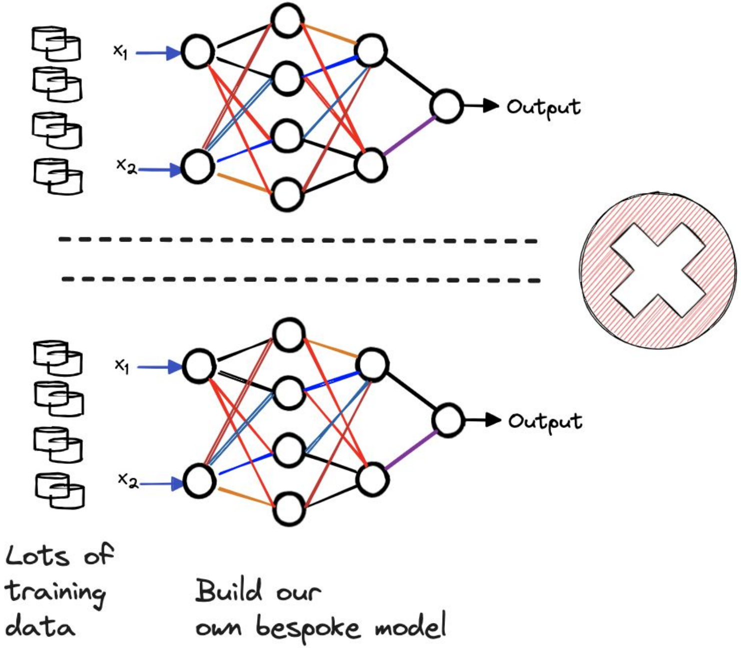
Slide source: Raghu Ganti/IBM Research

Can we change the paradigm of AI Infusion?

Advancing Application of Machine Learning Tools for NASA's Earth Observation Data

Jan. 21-23, 2020 | Washington, D.C.
Workshop Report





- Training data **bottleneck**
- Models don't **generalize**

Use AI Foundation Models for Science

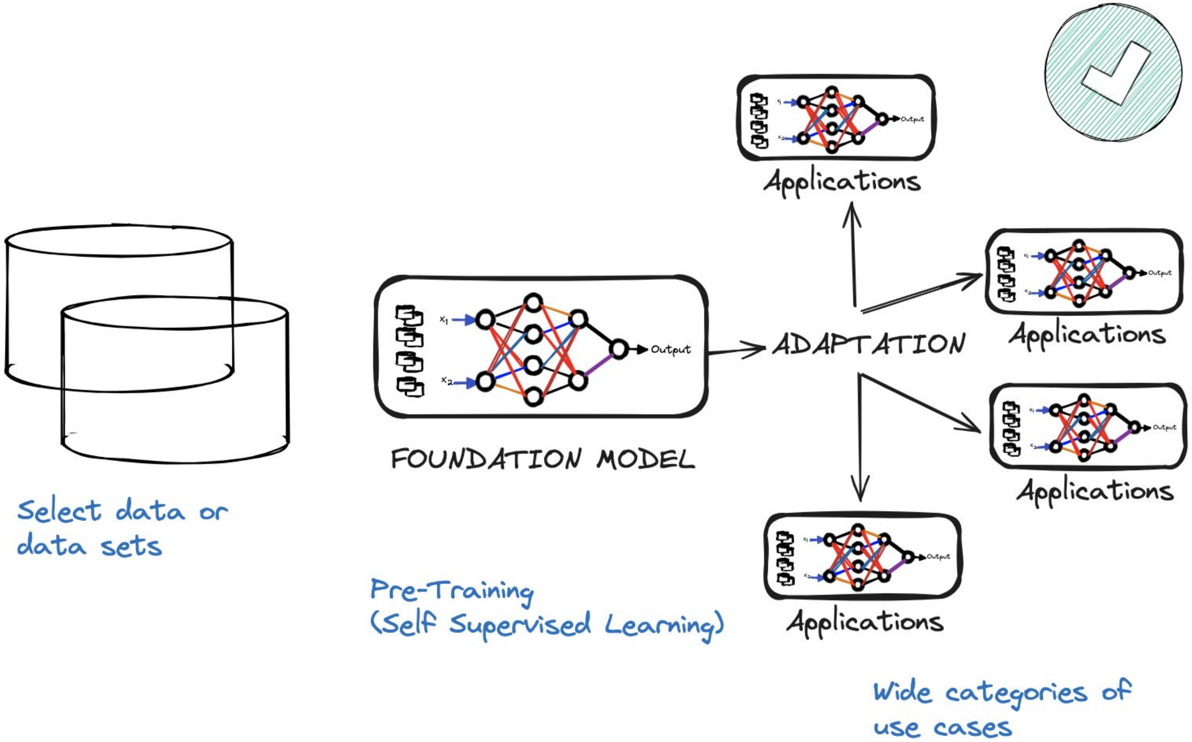
On the Opportunities and Risks of Foundation Models

arXiv:2108.07258v2 [cs.LG] 18 Aug 2021

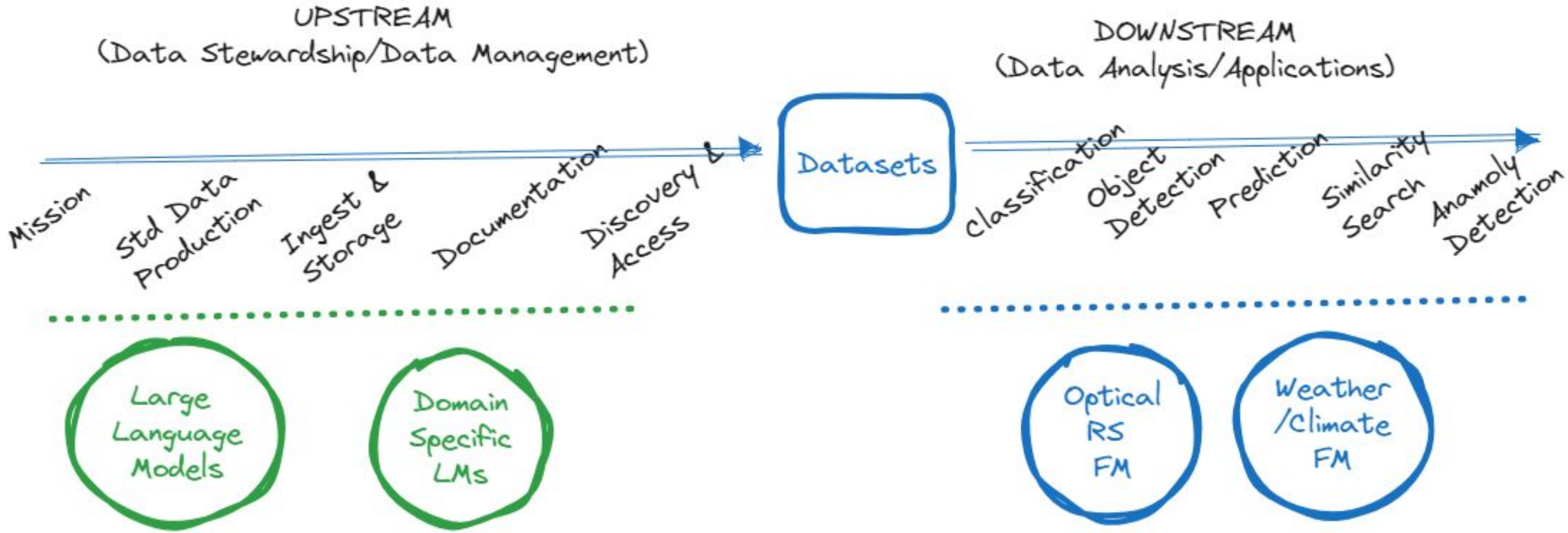
Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, reasoning, human interaction) and technical principles (e.g., model architectures, training procedures, data, systems, security, evaluation, theory) to their applications (e.g., law, healthcare, education) and societal impact (e.g., inequity, misuse, economic and environmental impact, legal and ethical considerations). Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities, and their effectiveness across so many tasks incentivizes homogenization. Homogenization provides powerful leverage but demands caution, as the defects of the foundation model are inherited by all the adapted models downstream. Despite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties. To tackle these questions, we believe much of the critical research on foundation models will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature.



AI FM Adoption Both Upstream and Downstream



Build FM for high value datasets/thematic areas

Foundation Models for Science - Our List

- **LLM for NASA's Science Mission Directorate**
  - Base Encoder Model - encoder-only transformer model, tailored for SMD applications
  - Sentence Transformer - Generates embeddings for queries and sentences, enhancing information retrieval
  - Passage Reranker - fine-tuned model that takes a search query, and a passage, and calculates the relevancy score of the passage w.r.t the query
- **Prithvi** - Geospatial FM based on HLS data to support Land Surface Processes and Application
  - Initial version
  - *Working on updated global version*
- **Weather and Climate (WxC) FM** - the focus is not just on Forecasting/Prediction but also on different categories of downstream applications
  - *Release May 2024*
- **Space Weather FM** - based on SDO for space weather applications (just started)

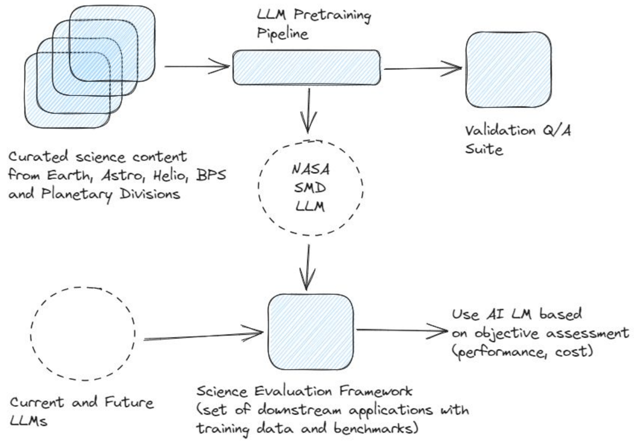
Indus - Language Model for NASA Science Mission Directorate

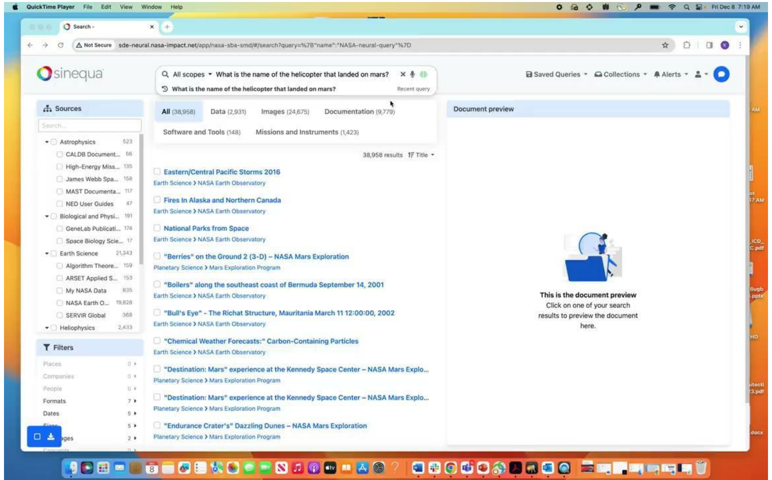
Goals

- Unite SMD's NLT experts to develop a NASA-specific LLM.
- Design and evaluate an LLM tailored to SMD's unique needs.
- Highlight potential LLM applications to encourage SMD collaboration.
- Establish a unified framework for LLM and strategies for sharing resources.

Approach

- Compile sources from ESD, Astro, Helio, BPS, and Planetary
- Utilize IBM's training pipeline
- Implement a Q/A Science Validation Suite for model training.
- Establish a Baseline Encoder Model for ongoing refinement.
- Create a Science Evaluation Framework for objective assessment of current and future models.



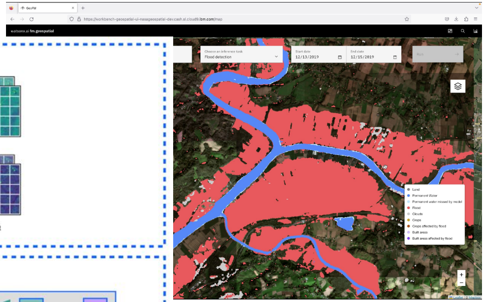


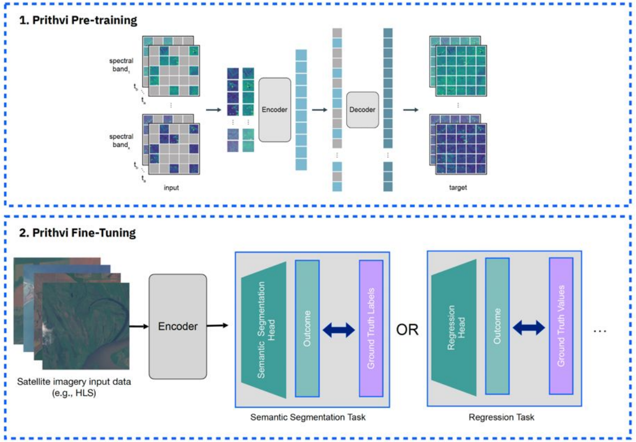
The RAG-Enabled Science Discovery Engine

NASA

IMPACT
Interagency Implementation and Advanced Concepts Team

Prithvi - Geospatial FM for HLS Data





Goals

- Proof of concept investigating the value of FM for science
- Understand the preprocessing required
- Understand the design considerations

Approach

- Masked Autoencoder (MAE) architecture - self-supervised learning model that learns to reconstruct the original image from a version where some parts have been masked out.
- Modifications for Satellite Imagery: 3D Positional Embedding, 3D Patch Embedding to handle the spatiotemporal nature of the data

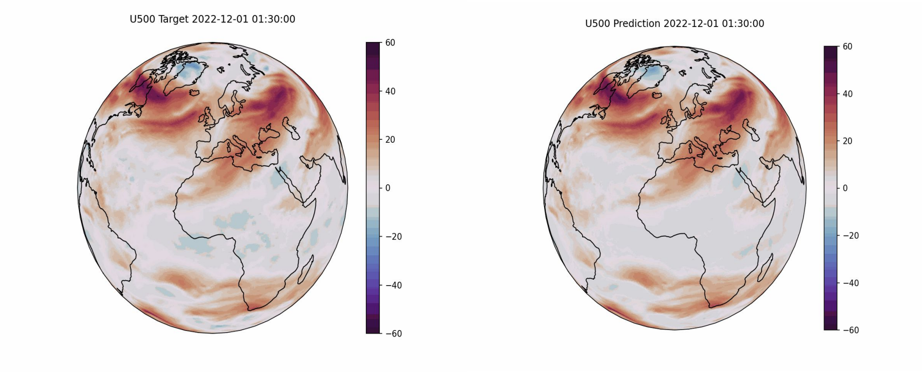
Key Takeaways

- Not expected to outperform all state-of-the-art models
- Advantages of Foundation Models:
  - Data Efficiency
  - Less Compute
  - Ability to generalize

WxC (Weather & Climate) Foundation Model

Goals

- AI FM for Weather and Climate not just on Forecasting/Prediction but for different categories of downstream applications
- Model will multiresolution both spatial and temporal to be able to use different types of data such as MEERA, ERA and HRR



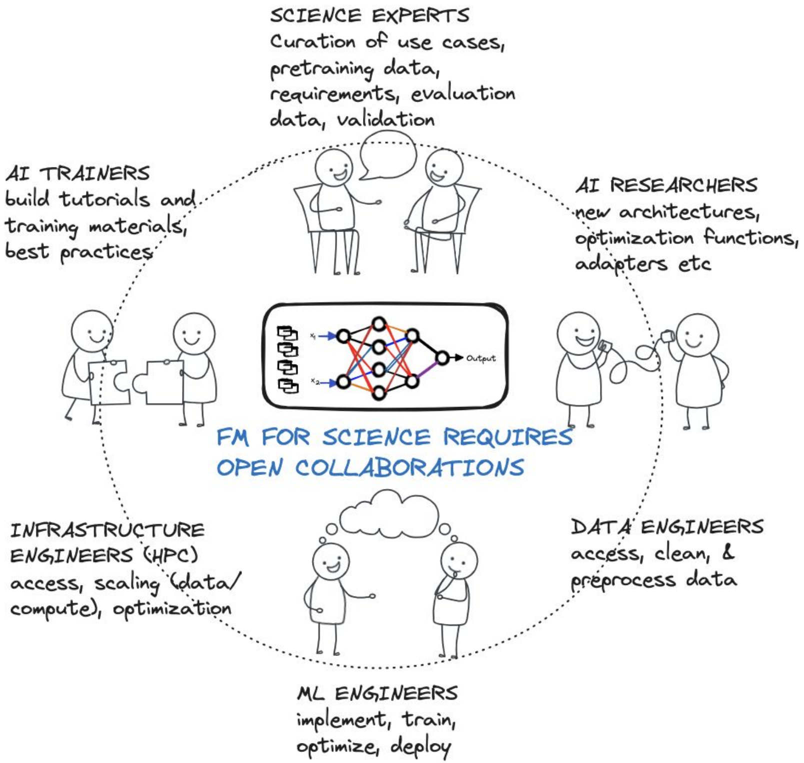
Approach

- Core architectures under consideration: SWIN, Hiera
- Extensions/modifications include:
  - Multi-level and multi-resolution approaches to accommodate data at different spatial and temporal scales.
  - Diffusion-based architectures to incorporate additional information and enhance model predictions.
- Evaluation using seven different types of use cases

Team

Broader participation for Science experts to ensure right direction, evaluation and future adoption of the model in their workflows [NASA, DOE ORNL, IBM Research, NVIDIA, Academia - University of Colorado, University of Alabama in Huntsville, Stanford]

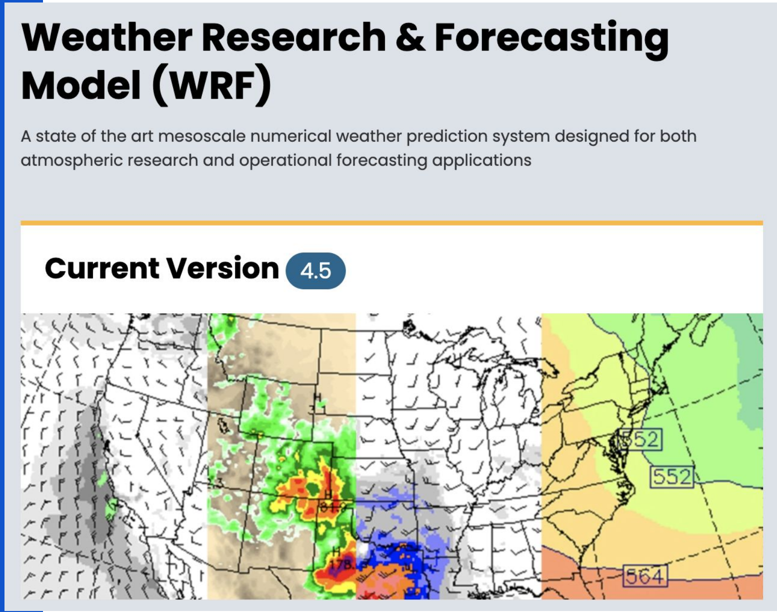
Building these FM requires OPEN Collaborations



- Diverse skill sets are required to build a useful AI FM (beyond just writing a paper)
- Compute resources
- Access to data

Difficult for a single team/organization to do this well for science

Collaboration Exemplar:
Weather and Research Forecasting (WRF) Model

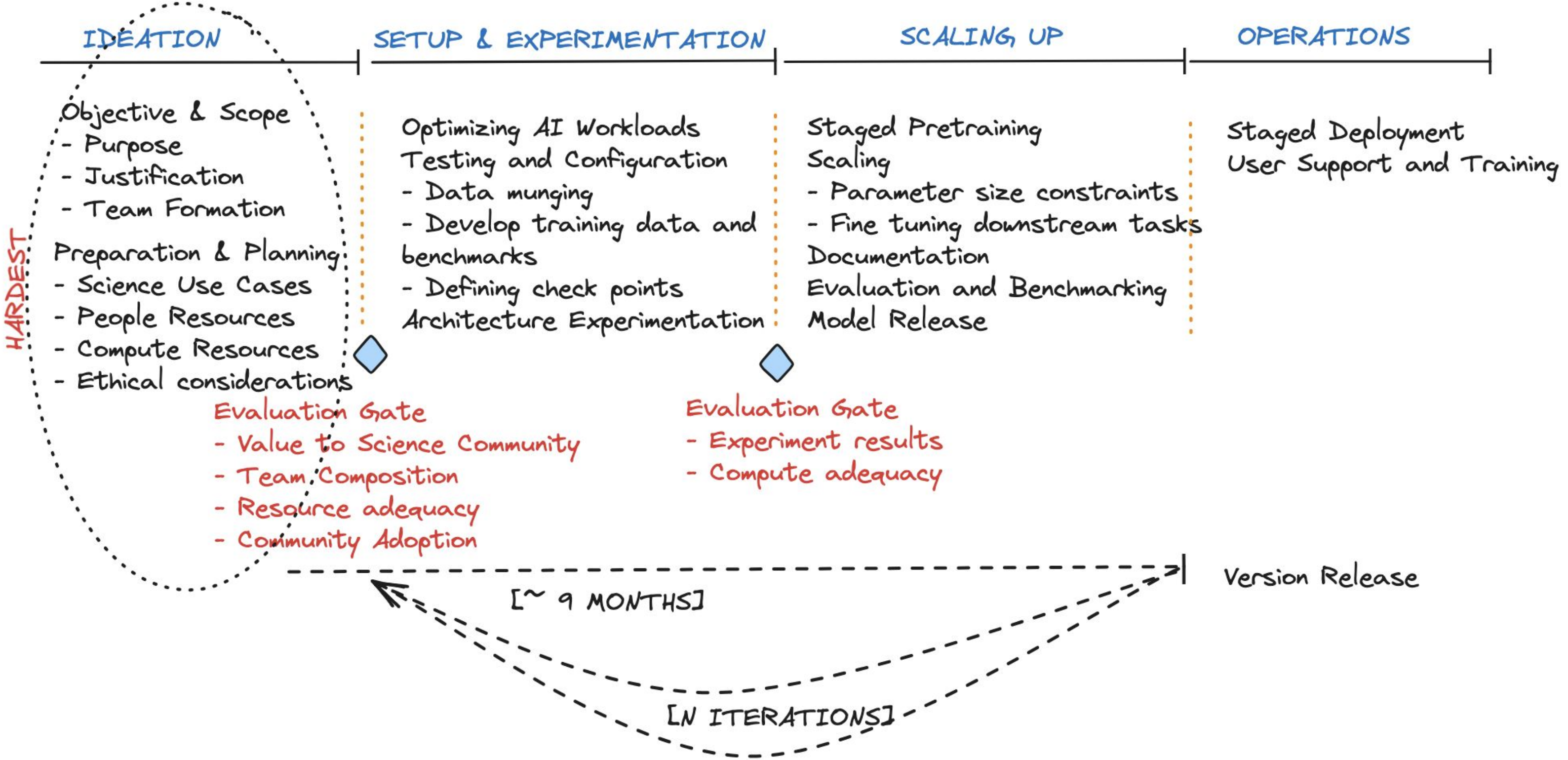


- Collaborative effort to create a next-generation model for weather prediction that would be useful for both research and real-time forecasting
    - NCAR, NOAA, U.S. Air Force, Naval Research Laboratory, University of Oklahoma, Federal Aviation Administration (FAA)
- Development of WRF began in the late 1990s, with the first official release occurring in 2000
- Massive user community exceeding 30,000 individuals across 150 countries

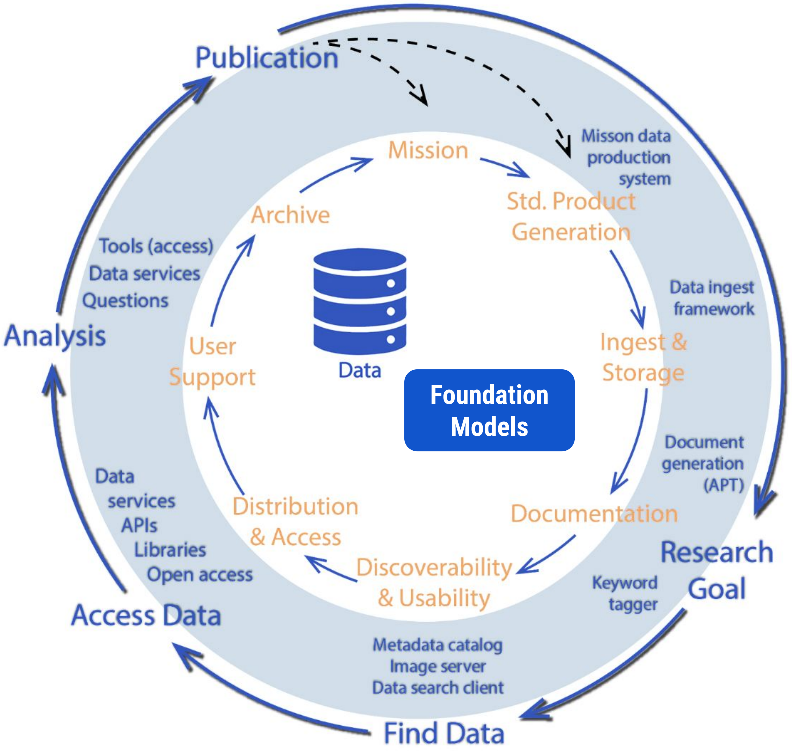
Key Features of the approach

- Broad collaboration including academic institutions, federal agencies, and other research organizations
- Community-driven approach ensured model was useful to wide range of users enabling adoption
- Education and training were an integral part of the WRF development effort

“Playbook” for building FM for Science



Building Blocks: Future State

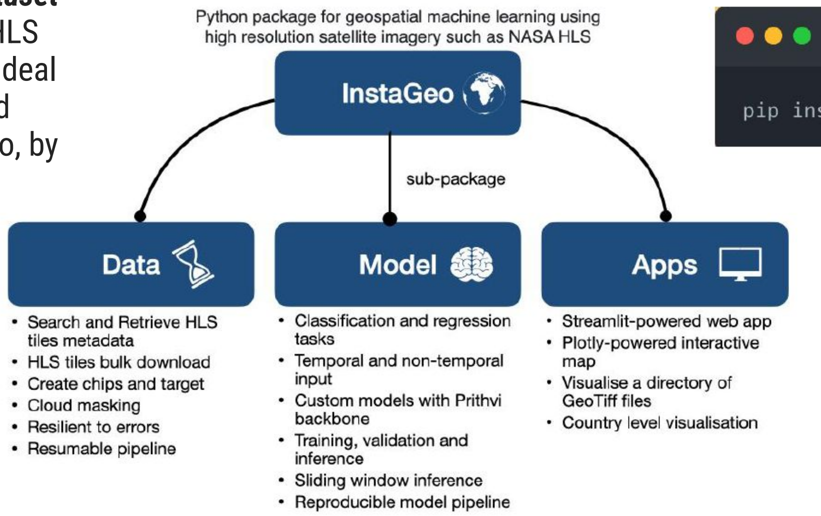


Data + Tools/Infrastructure + AI Foundational Models -> *Accelerate Research and Applications*

Prithvi Success Story: InstaDeep

"To provide more context, we started by using NASA's GLDAS dataset; however, its lower spatial resolution (0.25 degrees) ….. many challenges such as discrepancies in their update frequency and granularity, missing values, mismatches in data quality and sustained collection protocols. After all this pain, with suboptimal results, ***discovering your NASA Harmonized Landsat and Sentinel-2 (HLS) dataset and the Prithvi model in your paper was a breakthrough moment for us***. The HLS dataset's 30m spatial resolution and 3-5 day temporal resolution offered an ideal solution. We achieved state-of-the-art performance in locust breeding ground prediction by fine-tuning the Prithvi model with HLS multispectral bands. Also, by only relying on the HLS, it immediately enables us to potentially deploy the resulting model for operational use"
- Arnu Pretorius

- The HLS Data and Prithvi Foundation Model were instrumental in developing a cutting-edge model for locust breeding ground prediction.
- The creation of InstaGeo, an open source data pipeline designed to search, retrieve, preprocess HLS data from Earth Data, and finetune/adapt Prithvi for specific applications.



Content attribution: Arnu Pretorius, Yusuf Ibrahim InstaDeep

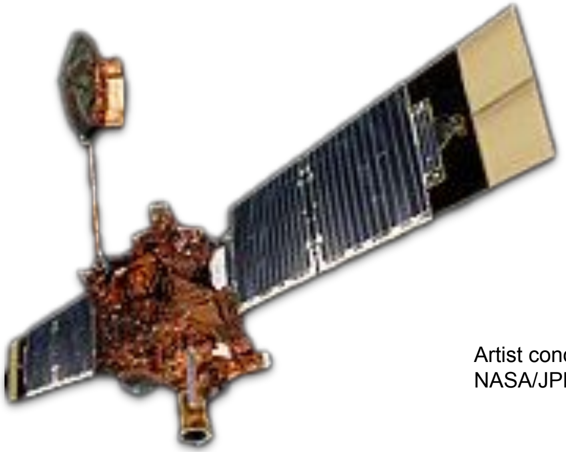
Contact or Connect with Me

rahul.ramachandran@nasa.gov

The Enduring Legacy of NASA Science Data

"The National Aeronautics and Space Administration (NASA) has become a knowledge agency. Long after the Mars Surveyor has gone silent, Hubble has met the same fate as Mir, and the Moderate Resolution Imaging Spectroradiometer has produced its final set of images, what will endure are the volumes of valuable data that these instruments and many others have collected over their lifetimes…."

– National Research Council, 2002



Artist concept of the Mars Global Surveyor
NASA/JPL-Caltech/Corby Waste