National Aeronautics and
Space Administration

AI Foundation Models for Science

Dr. Rahul Ramachandran
NASA Marshall Space Flight Center
Office of Data Science and Informatics

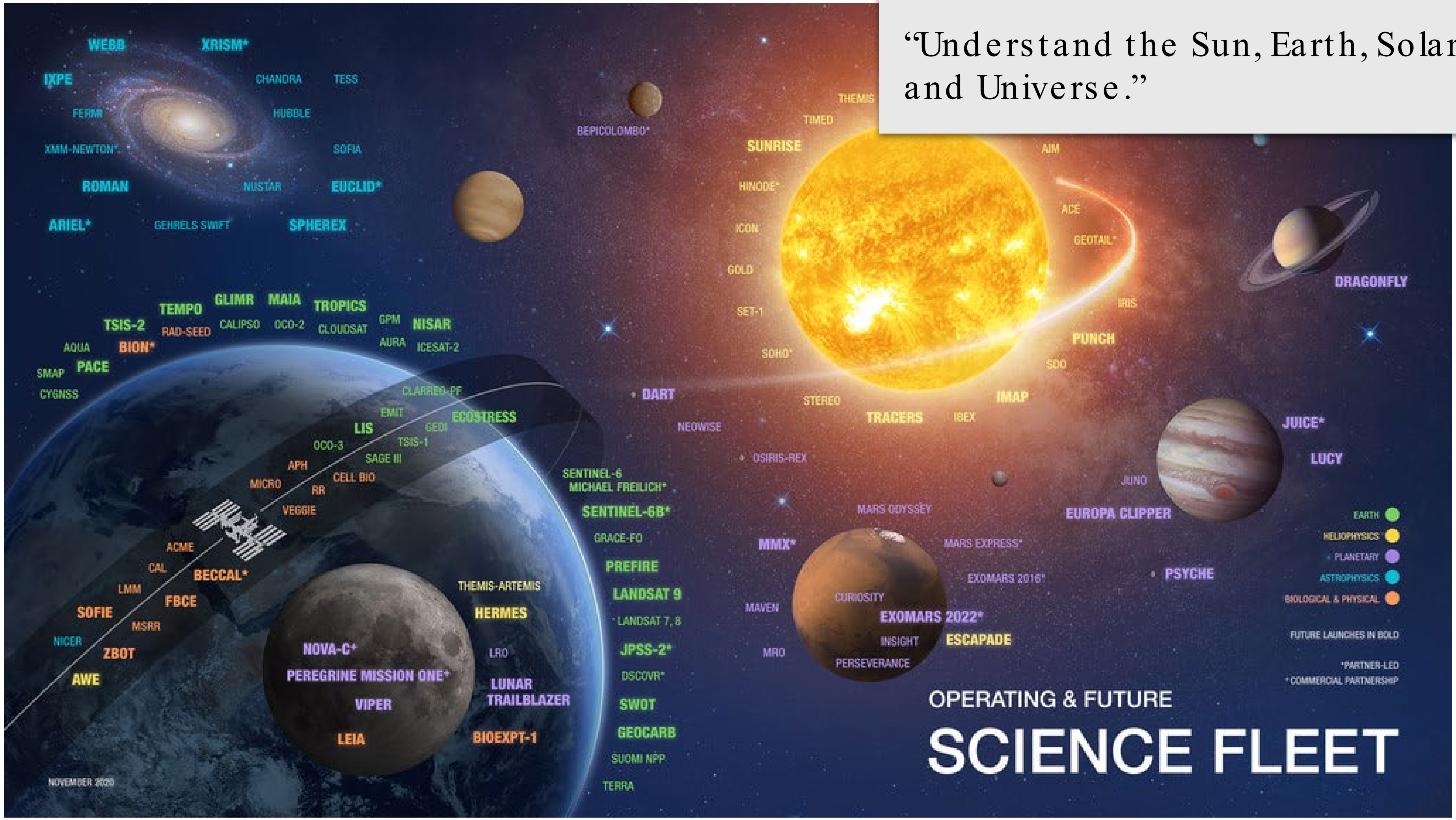
rahul.ramachandran@nasa.gov

National Aeronautics and
Space Administration

NASA (Space) Science

NASA's Strategic Goal 1.1

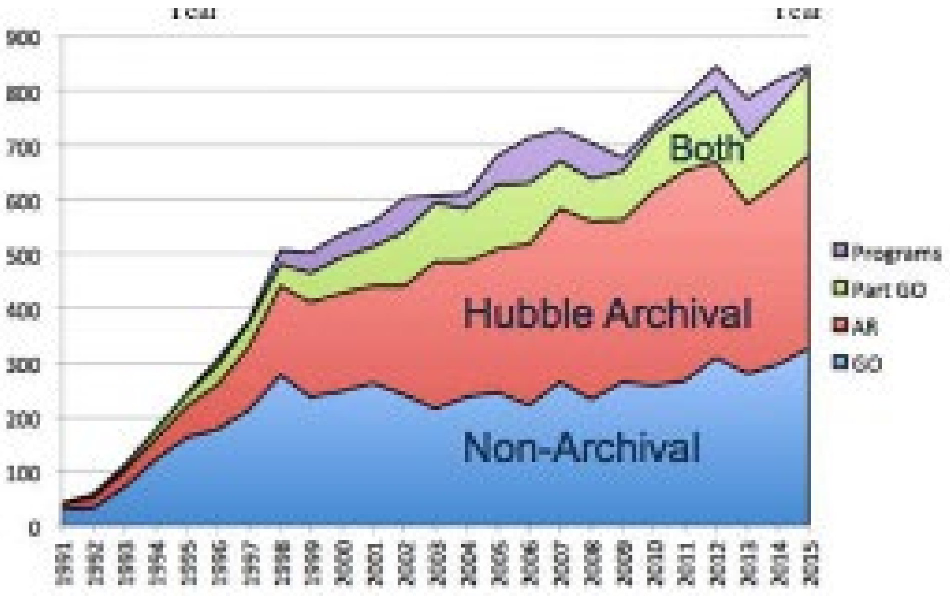
"Understand the Sun, Earth, Solar System, and Universe."



Value of NASA's 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

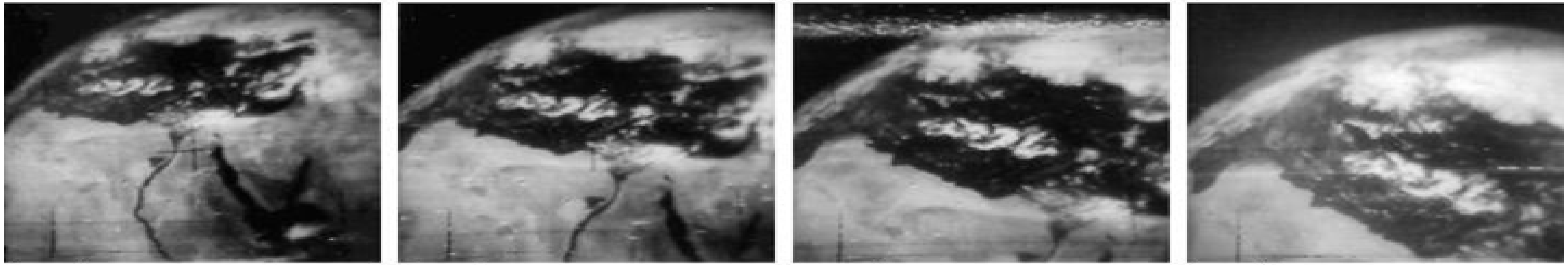


Number of Hubble papers as a function of time. Archival research dominates paper output after the first few years.

Problem of Scale:
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 themehow to obtain information and understanding from these data.”

– National Research Council, 1982



frame 6 frame 7 frame 8 frame 9

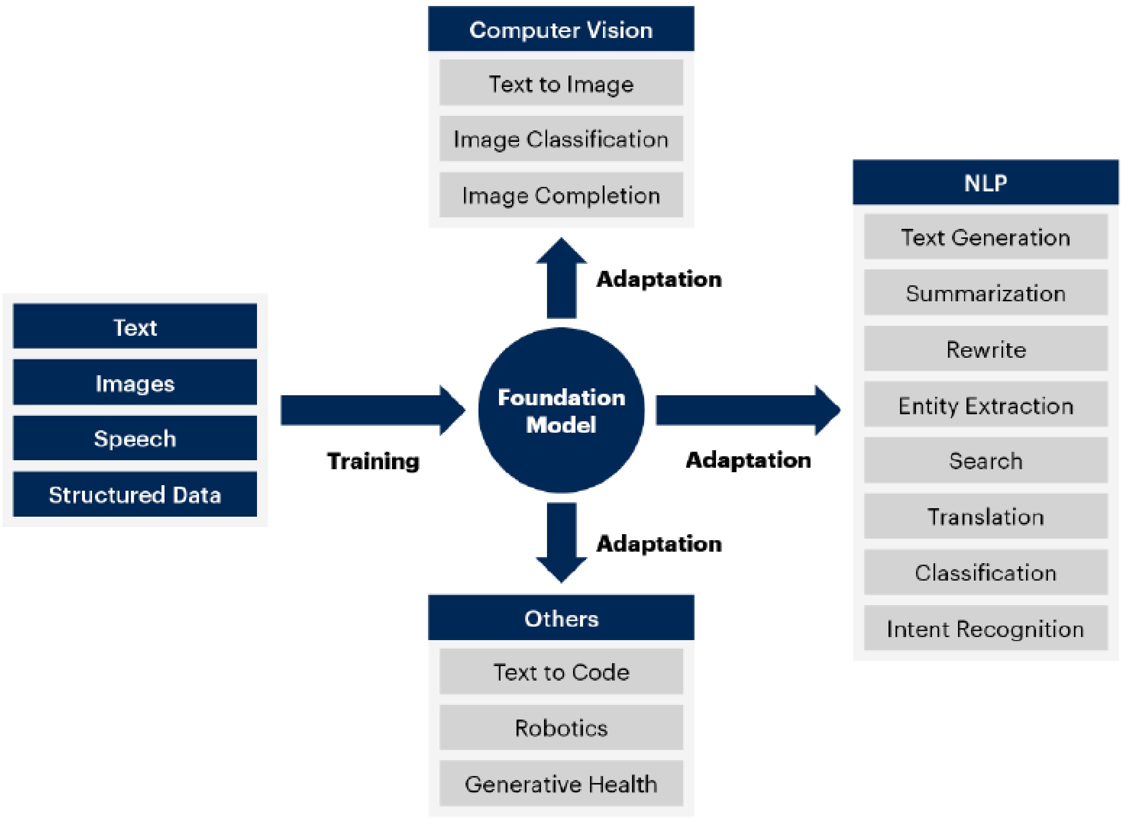
Photography from TIROS 1, Launched April 1, 1960
NASA Goddard Space Flight Center

The challenge: manage overwhelming data and distill it into meaningful knowledge.

What Are Foundation Models?

- Foundation models (FM) are AI models that are *designed to replace task - specific models and be applied to many different downstream applications* .
- FM are trained using *self-supervised techniques* and can be built on any type of *sequence data* .
  - Self supervised learning removes the existing roadblock for developing a large annotated dataset for training.
- FM models have the following characteristics:
  - Mostly utilize transformer architectures that utilize the notion of attention or self-attention, allowing the network to model the influence of distant data elements on each other both in time and space
  - Exhibit *emergent properties that are induced from the data* , rather than explicitly constructed
  - Can be applied to downstream tasks by *using few shot learning and fine tuning*
  - Have to be *trained at scales* that limits the ability to a handful organizations

Foundation Models - Characteristics and Applications



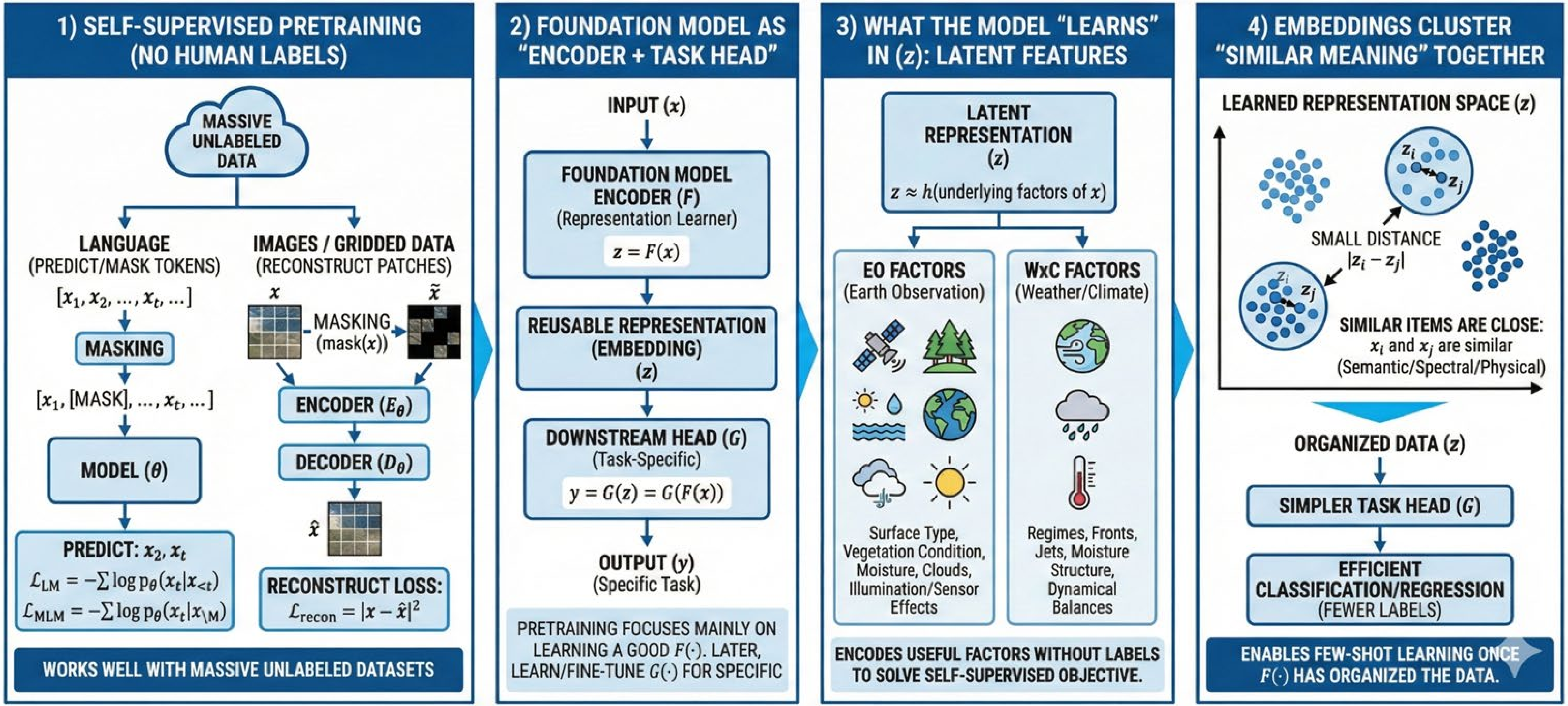
Source: Gartner
769102_C

Image source: Gartner Report - Innovation Insight for Artificial Intelligence Foundation Models, Published 27 Oct 2022

Gartner

Self-Supervised Learning: FMs from Unlabeled Data

Core Concept: model learns hidden patterns in the data, leveraging massive unlabeled datasets



Why are FMs Important for Science?

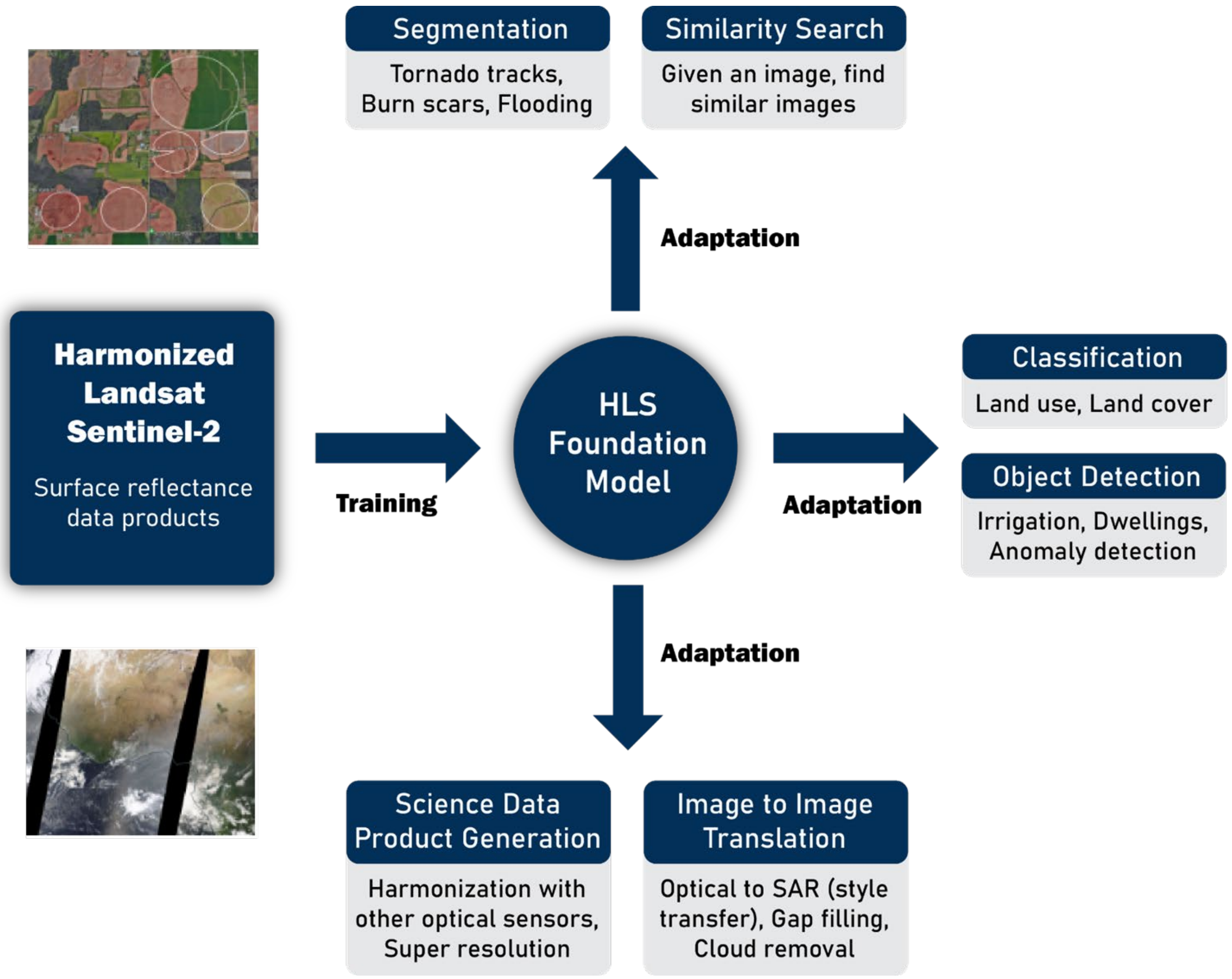
Obvious Practical Benefits

- Accelerate Downstream Applications:
- Lower Barriers to AI Integration:- simplify the process of incorporating AI into scientific research and analysis.
- Remove Training Data Bottlenecks - reduce the dependency on large, labeled training datasets for every new task.
- Increase Data Utilization- drive greater use and exploration of NASA's vast data archives.
- Reduce Cost of Entry- drastically lower the initial effort and resources needed to build new downstream applications.

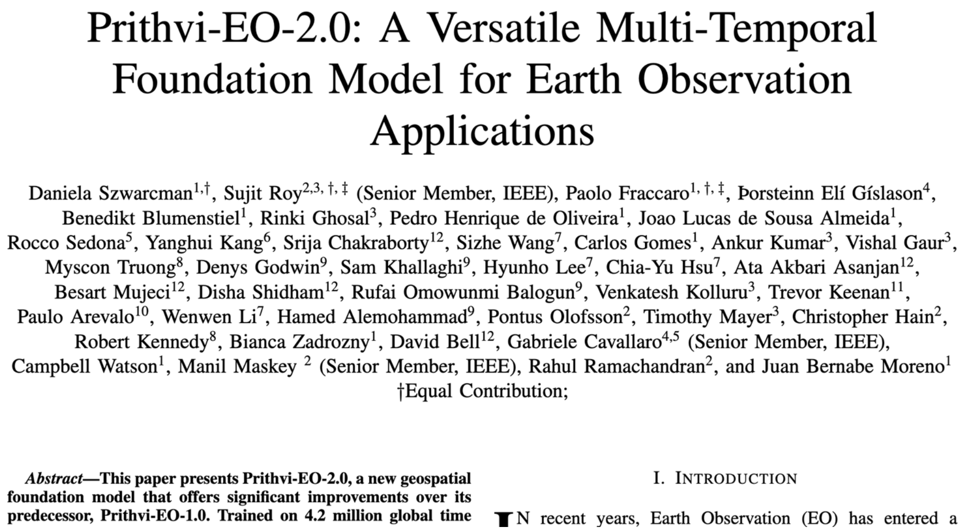
Strategic Benefits for NASA Science

- Enable Multi-Instrument, Multi-Mission Synthesis - new methods for combining and analyzing data from diverse instruments and missions.
- Accelerate Data-Driven Model Development - the creation of sophisticated models, like digital twins, for complex Earth and space systems.
- Explore New Scientific Spaces - utilize data embeddings to uncover novel relationships, hidden processes, and unexpected patterns in data.
- Standardize Scientific Data Representations?

Prithvi-EO: Generalist Geospatial FM



- Built with collaboration with IBM Research, UAH, Clark University, ASU and others
- Initially released (2023) 100M parameter model pretrained on HLS CONUS data
- Evaluated for adaptation for different categories of downstream tasks



Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications

Daniela Szwarcman[1,†], Sujit Roy[2,3,†,‡] (Senior Member, IEEE), Paolo Fraccaro[1,†,‡], Þorsteinn Elí Gíslason[4], Benedikt Blumenstiel[1], Rinki Ghosal[3], Pedro Henrique de Oliveira[1], Joao Lucas de Sousa Almeida[1], Rocco Sedona[5], Yanghui Kang[6], Srija Chakraborty[12], Sizhe Wang[7], Carlos Gomes[1], Ankur Kumar[3], Vishal Gaur[3], Myscon Truong[8], Denys Godwin[9], Sam Khallaghi[9], Hyunho Lee[7], Chia-Yu Hsu[7], Ata Akbari Asanjan[12], Besart Mujeci[12], Disha Shidham[12], Rufai Omowunmi Balogun[9], Venkatesh Kolluru[3], Trevor Keenan[11], Paulo Arevalo[10], Wenwen Li[7], Hamed Alemohammad[9], Pontus Olofsson[2], Timothy Mayer[3], Christopher Hain[2], Robert Kennedy[8], Bianca Zadrozny[1], David Bell[12], Gabriele Cavallaro[4,5] (Senior Member, IEEE), Campbell Watson[1], Manil Maskey [2] (Senior Member, IEEE), Rahul Ramachandran[2], and Juan Bernabe Moreno[1]
†Equal Contribution;

Abstract—**This paper presents Prithvi-EO-2.0, a new geospatial foundation model that offers significant improvements over its predecessor, Prithvi-EO-1.0. Trained on 4.2 million global time**

I. INTRODUCTION

IN recent years, Earth Observation (EO) has entered a

Prithvi WxC: Weather and Climate Foundation Model

Motivation

Enhance atmospheric analysis to address applications that aren't focused solely on forecasting such as parameterization and downscaling.

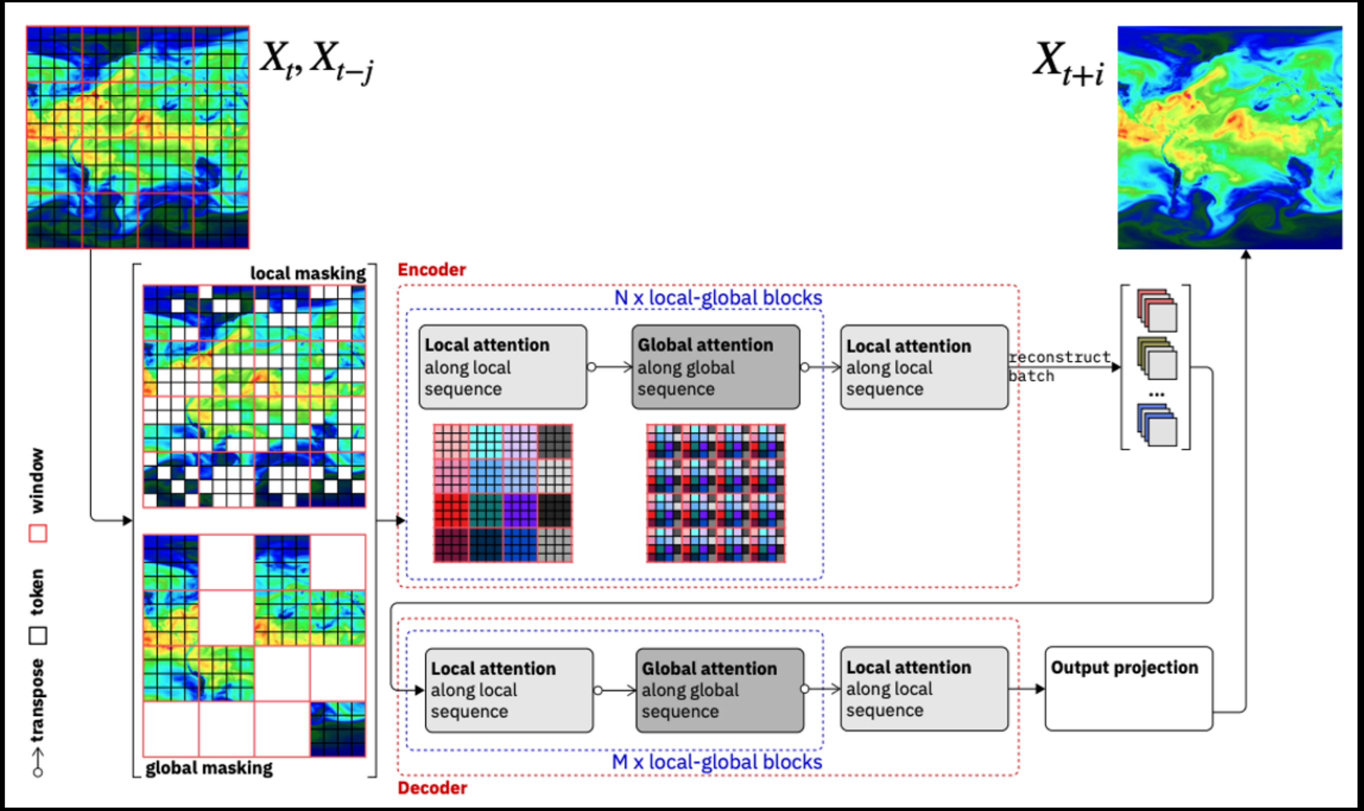
Science Use Cases

1. **Hurricane forecasting:** Outperformed FourCastNet models, with a mean track error of 63.9 km for Hurricane Ida
2. **Downscaling:** Achieved significant improvements in spatial and temporal RMSE (4x better than interpolation baselines)
3. **Climate model parameterization:** Successfully fine-tuned for gravity wave flux prediction, improving sub-grid atmospheric process representation

Prithvi WxC: Weather and Climate Foundation Model

Architecture

- 2.3 billion parameter foundation model for weather and climate applications
- Developed using 160 atmospheric variables from NASA's MERRA-2 dataset (1980–2019)
- Transformer-based encoder-decoder architecture capturing regional and global dependencies
- Combines masked reconstruction and forecasting objectives for generalizability
- Open-source release on Hugging Face for public use



$$\frac{\hat{X}_{t+\delta t} - C_{t+\delta t}}{\sigma_C} = f_\theta\left[M_{0.5}\left(\frac{X_t - \mu}{\sigma}, \frac{X_{t-\delta\tau} - \mu}{\sigma}\right); \frac{C_{t+\delta t} - \mu}{\sigma}, S, \delta t, \delta\tau\right]$$

SURYA: Heliophysics Foundation Model

Motivation

Boost the accuracy and efficiency of space weather forecasting including the degradation/disruption of satellite navigation (GPS) and radio communications and radiatio

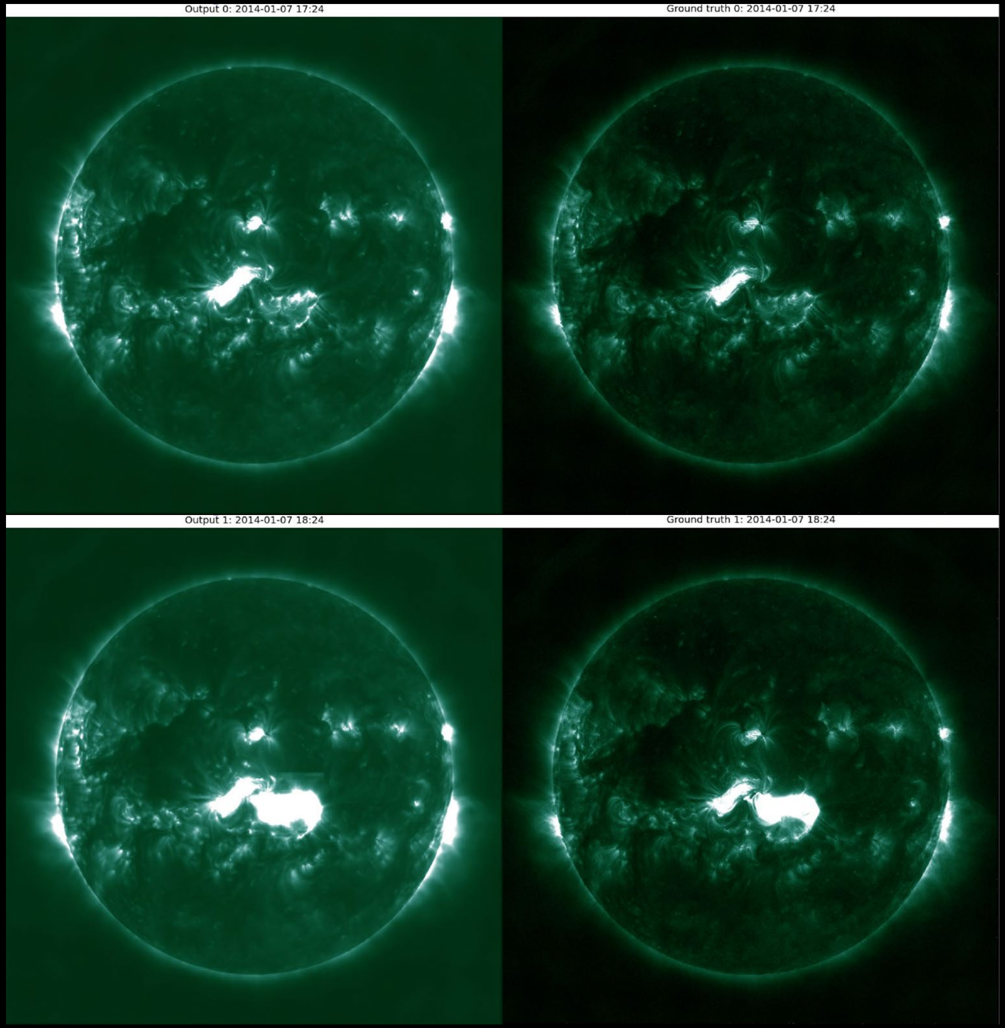
Science

1. **Solar Flare Forecasting:** Predict M- or X-class flares within a 24-hour window (binary classification)
2. **Active Region Emergence Forecasting:** 24-hour generative prediction of continuum intensity to predict precursors to flare activity and magnetic storms
3. **Solar Wind Forecasting:** Regression task predicting wind speed up to 4 days ahead
4. **Solar EUV spectra prediction:** Forecast EUV irradiance across 1343 channels, potentially creating a "virtual EVE" capability

Surya: Heliophysics Foundation Model

Data Source

NASA's Solar Dynamics Observatory (SDO), specifically AIA and HMI instruments, uses 13 channels (8 AIA, 5 HMI).



Zeroshot rollout of a flare event

Open Science as an Ethos

Essential for AI for Science

Challenges with Commercial AI Models

- Proliferation of closed, proprietary models developed by commercial vendors
- Limited visibility into pretraining data, model architecture, compute resources, and methodological choices
- Models function as black boxes, constraining scientific interpretability
- Release cycles driven by hype rather than verifiable scientific rigor

What the Scientific Community Must Ensure

- Collaborative model development grounded in sound scientific principles
- Open and accessible design artifacts (data, code, documentation, evaluation protocols)
- Independent evaluation by diverse science teams, not just model developers
- Transparent communication of strengths, limitations, and appropriate use cases
- Explainability and interpretability (based on first principles) to understand what the model has learned, where it performs reliably, and where it systematically fails

Thank You

Dr. Rahul Ramachandran

NASA Marshall Space Flight Center
Office of Data Science and Informatics

rahul.ramachandran@nasa.gov