An experiential session on Large Geospatial Foundation & Language Models

Manil Maskey, PhD

Senior Research Scientist/Data Science & Innovation Lead
NASA MSFC & NASA SMD/HQ
IEEE GRSS Earth Science Informatics Technical Committee Chair

IEEE GRSS High Performance and Disruptive Computing in Remote Sensing Summer school
June 5-6, 2024
Santiago de Compostela, Spain

Goals

Continue the summer school series: building capacity around data science with focus on computing

Explore research and applications of Geospatial Foundation Models and Large Language Models

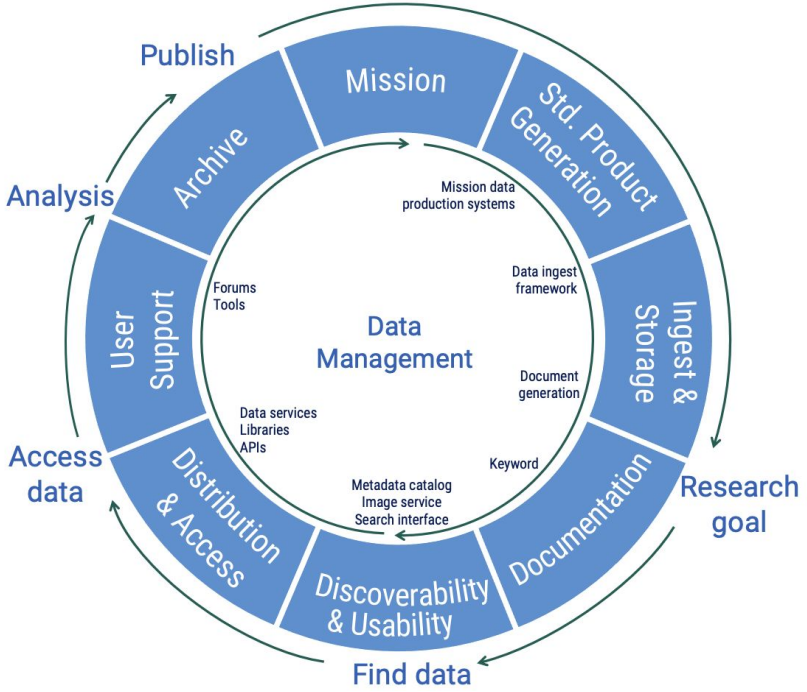
Provide hands on experience to support science research lifecycle:

- Use of LLMs for efficiency
- Use of different computing platforms to
  - Fine-tune geospatial foundational models
  - Build application and inference

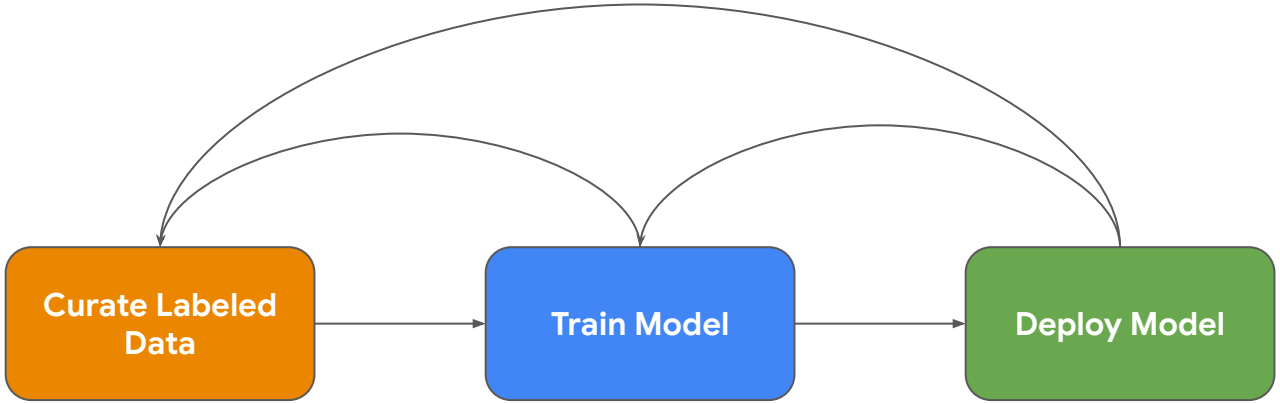
Provide forum to exchange Ideas

Foster collaboration

AI to support science research lifecycle



Supervised learning



AI challenges in Earth science

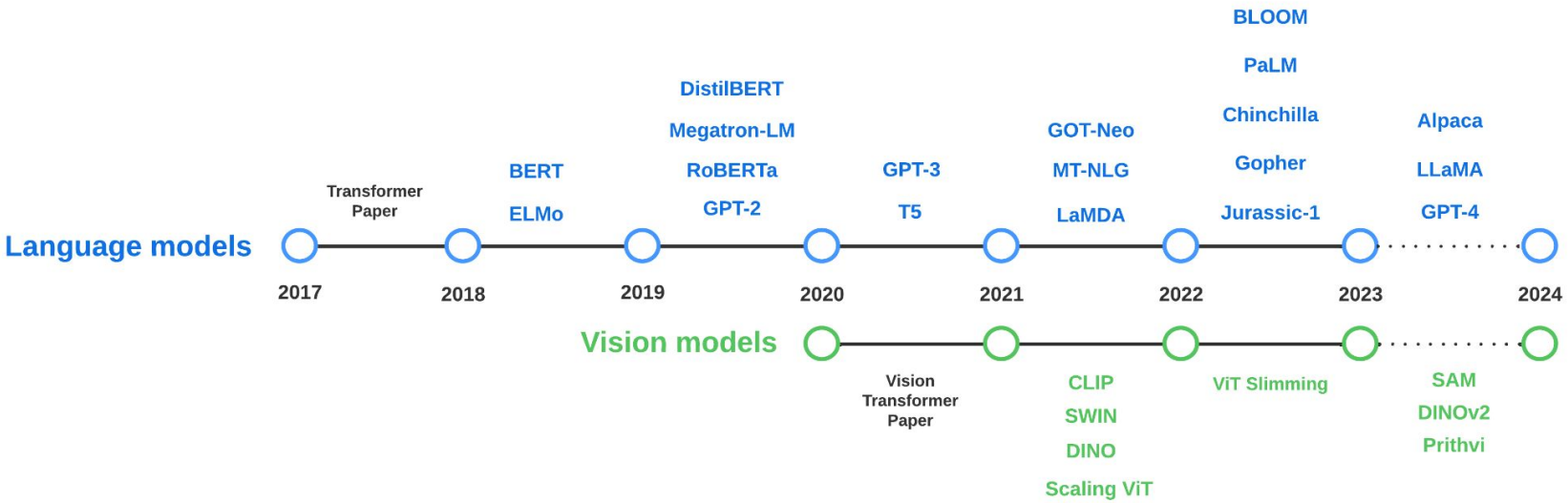
Advancing Application of Machine Learning Tools for NASA's Earth Observation Data

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

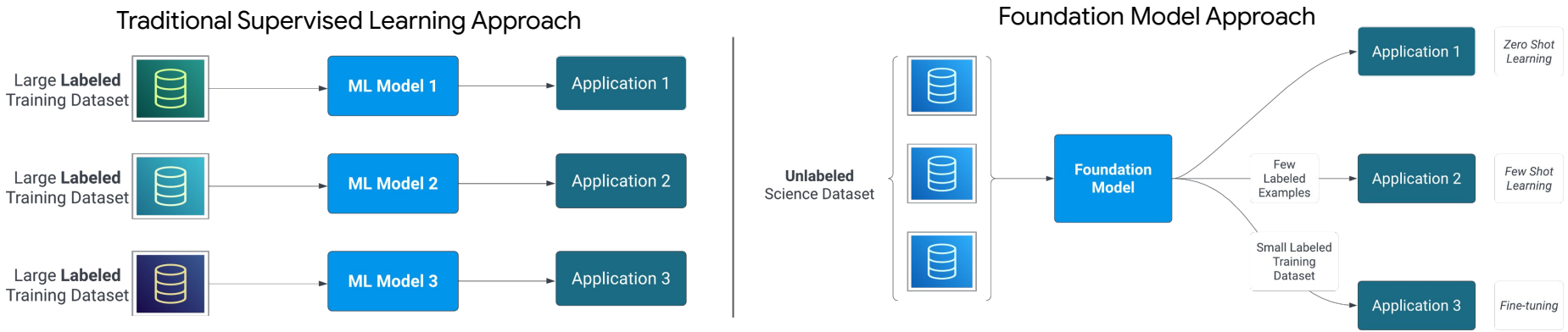
- **Training data** is the main component of supervised machine learning techniques and is increasingly becoming the **main bottleneck to advance applications of machine learning** techniques in Earth science.
- Geoscience models must **generalize across space and time**; however, for supervised learning one needs large training datasets to build generalizable models.

Maskey et al. "Advancing AI for Earth Science: A Data Systems Perspective," AGU Eos 2020

Foundation Models



AI Foundation Models

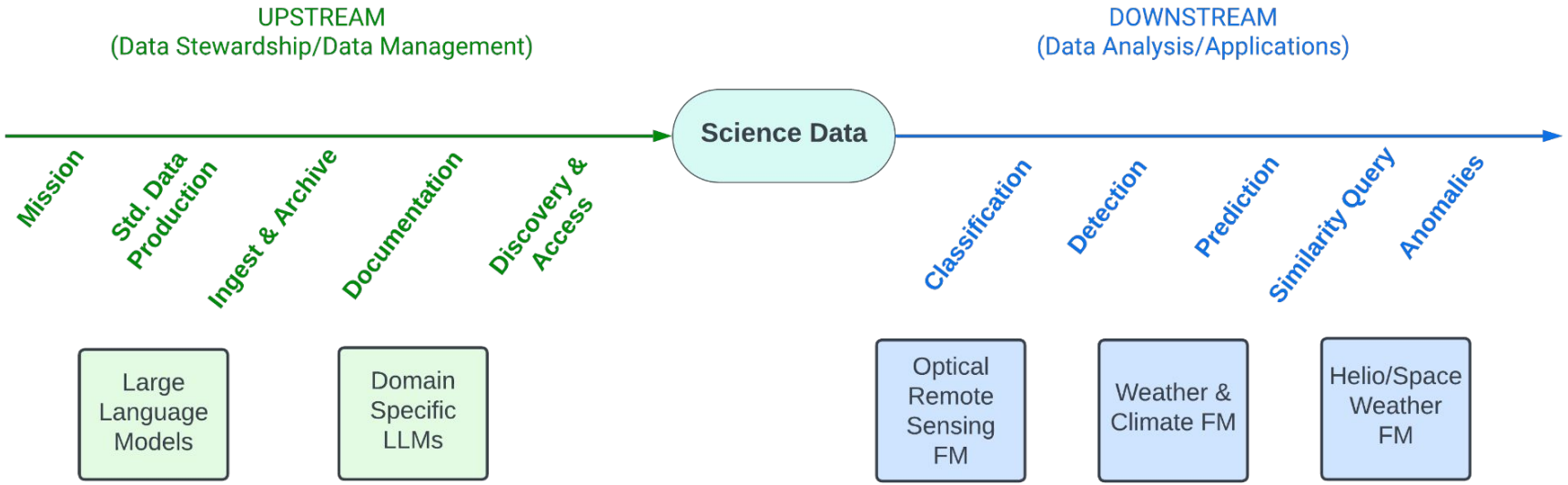


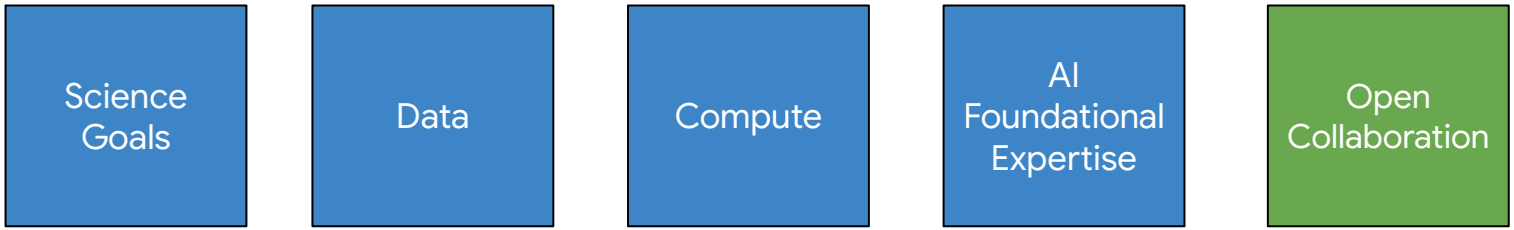
An AI Foundation Model is a large machine learning model pre-trained on a vast amount of data using self supervision, enabling it to perform a wide range of tasks.

Advantages

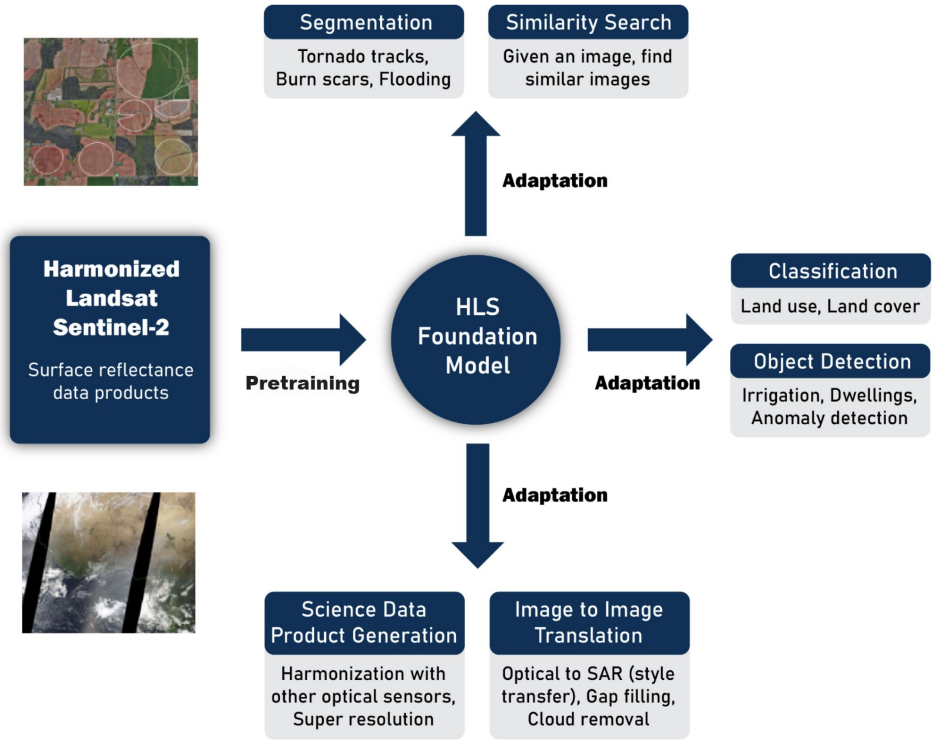
- Maximizes the potential of archives of time series datasets
- Reduces effort for building AI applications
- A single foundation model can be fine-tuned for a wide range of applications
- Foundation models often achieve state-of-the-art performance on various tasks, even with limited labeled data

AI Adoption both Upstream and Downstream




Science Goals
Data
Compute
AI Foundational Expertise
Open Collaboration

HLS Foundation Model (Prithvi)



Collaboration with IBM Research

Initial version released are 100M and 300M parameter models

Masked Autoencoder where attention mechanism is extended in space and time

Being evaluated for adaptation for different categories of downstream tasks

Collaborators

- IBM, UAH, Clark University, ORNL, Hugging Face

Why HLS?

Merging Sentinel-2 and Landsat data

2-4 day global coverage

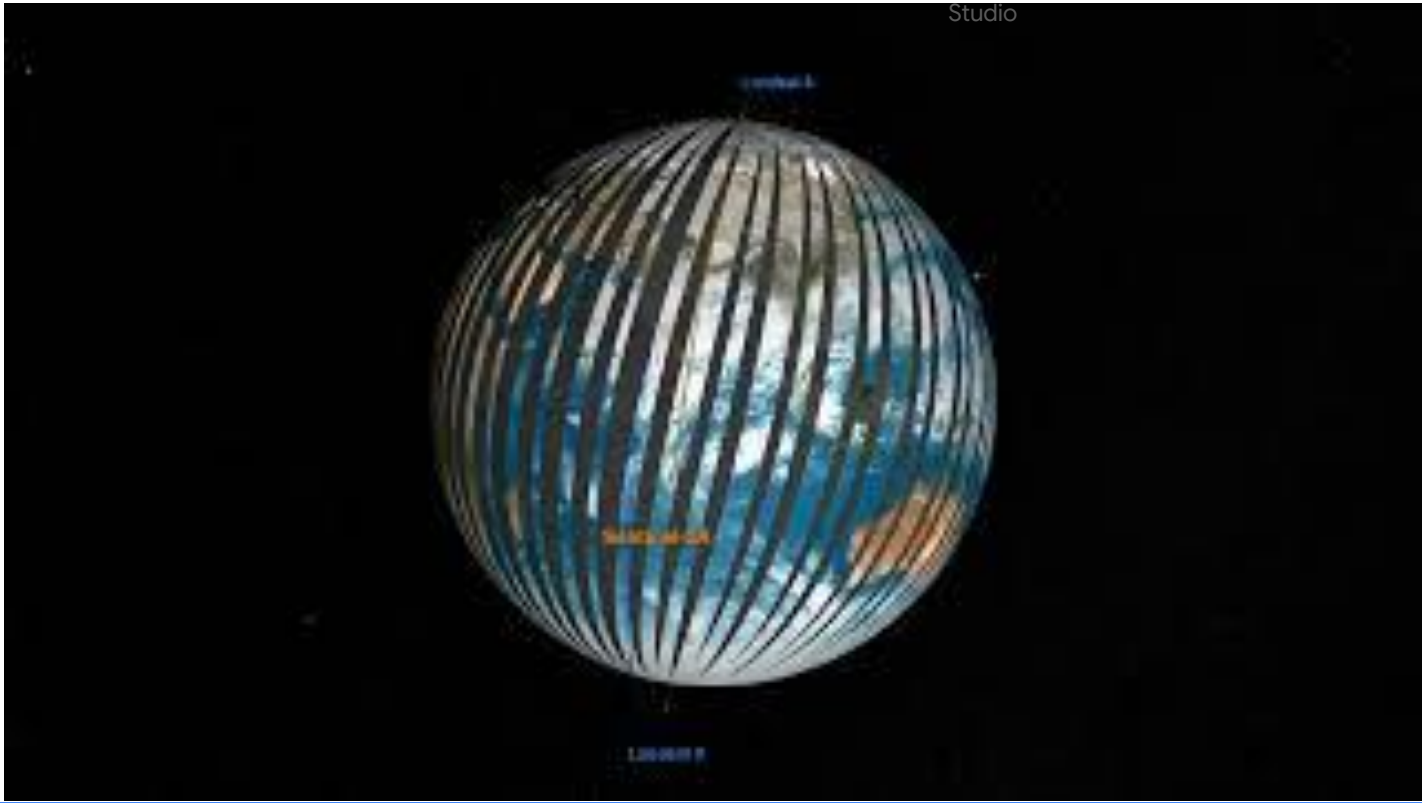
“Seamless” near-daily 30m surface reflectance

~8000 citations

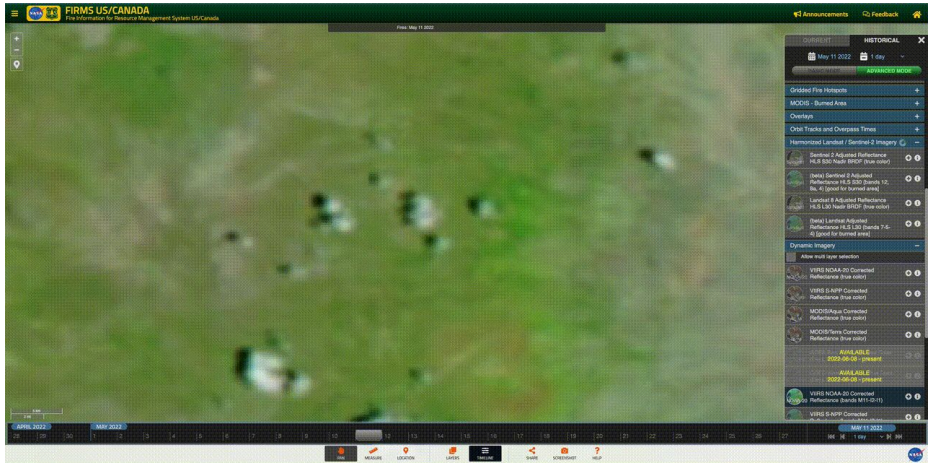
2 PB+ data transferred to users

2nd most downloaded data from NASA Earth observation data archive

Video Credit: Scientific Visualization Studio

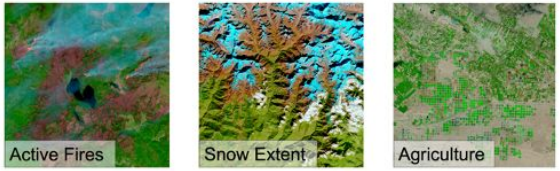


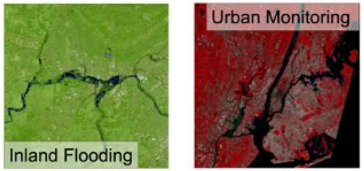
HLS applications

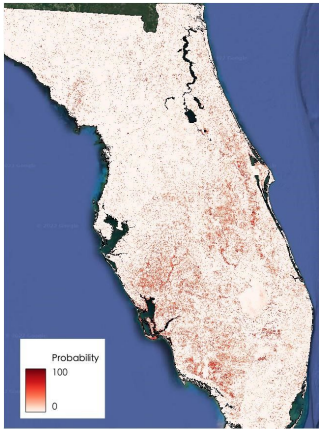


Wildfire monitoring in FIRMS.

Reservoir water level monitoring.





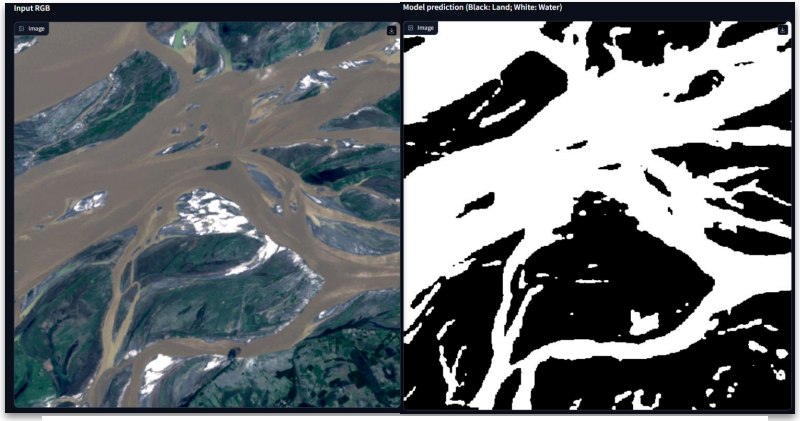


Assessing and reporting damage during Hurricane Ian.

Few highlights from Prithvi

Significant Achievements:

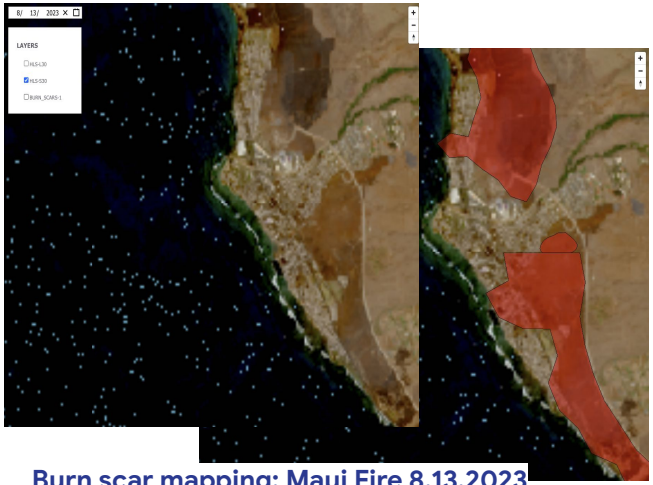
- Successfully evaluated the model across four distinct downstream tasks.
- Fewer training data (up to 87% less) to meet established benchmarks.



Model fine-tuned to segment the extent of floods on Sentinel-2

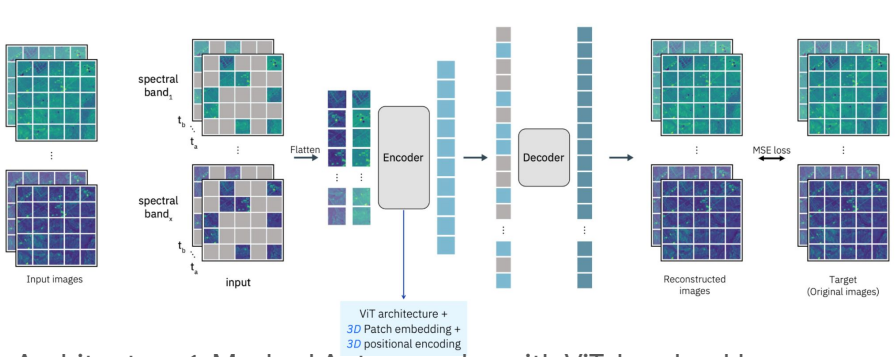
Community adoption:

- Demoed at the White House Demo Day
- 80+ downloads from Hugging Face, and used during Space Apps Challenge (42 submissions)
- Part of ESRI Living Atlas
- Used in teaching remote sensing
- Scaling at Julich Supercomputing Centre (JSC)

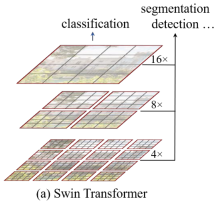


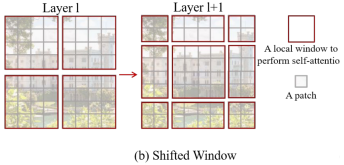
Burn scar mapping: Maui Fire 8.13.2023

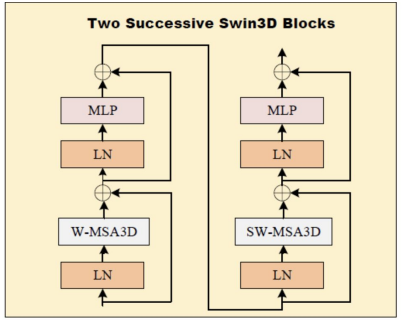
Geospatial Foundation Model: Scaling Prithvi



Architecture 1: Masked Autoencoder with ViT-L as backbone







Architecture 2: Swin 3d Architecture*

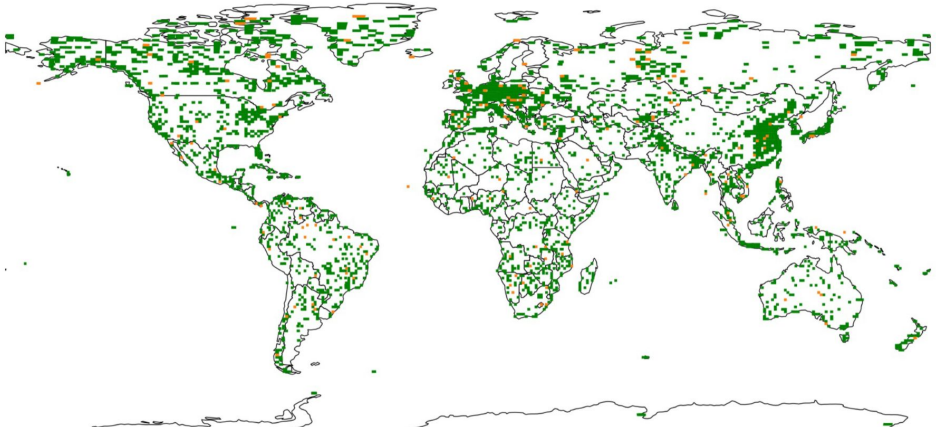
Architecture:

- Training model on Swin 3D.
- 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

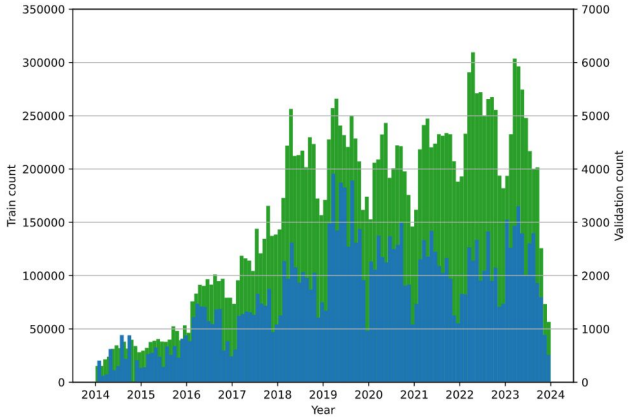
3156 train tiles (green)
168 validation tiles (orange)

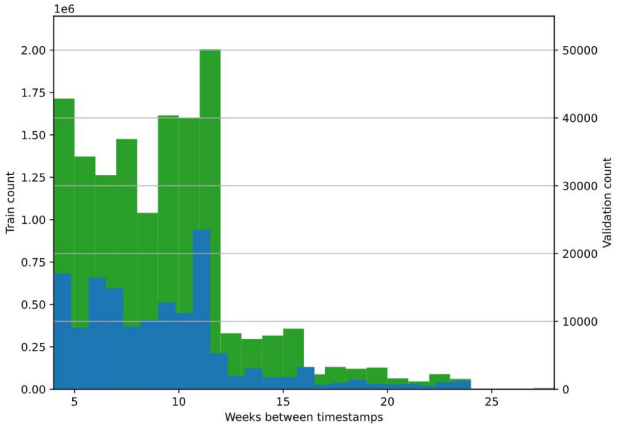


Data 2015 - 2024 HLS

4.8 mil samples

Temporal distribution



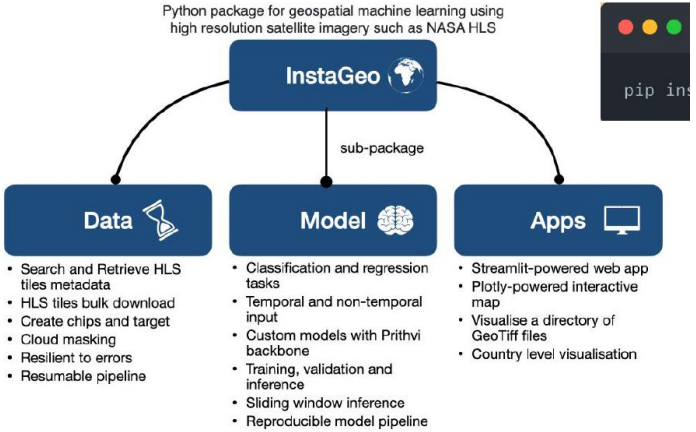


Credit: IBM Research & NASA IMPACT

Downstream applications

- Burn scar detection
- Flood water delineation
- Crop classification
- Cloud gap filling
- LULC (HK region)
- Eddy covariance
- Insect damage estimation
- *Locust breeding ground prediction**
- *Semantic segmentation of mangrove forest**

Locust breeding ground prediction.

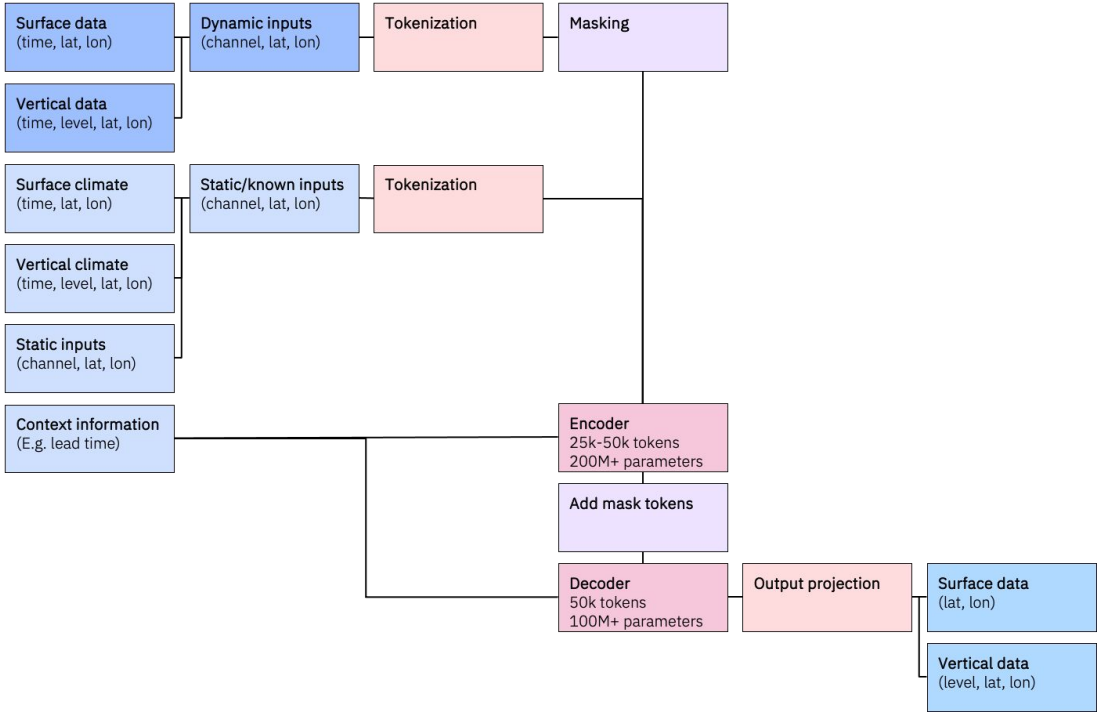


Source: Arnu Pretorius, Yusuf Ibrahim InstaDeep

Weather and Climate

Weather Foundation Model

Pretraining



Credit: IBM Research

Key features

- Grid free model. Can be used for both Euclidean and spherical topologies.
- Encoder/decoder are fully attention based.
- All auxiliary information (e.g. lead time) injected via context tokens.
- Tokenization & output projection can use convolutions.
- No hard-coded token or window sizes.

Input data

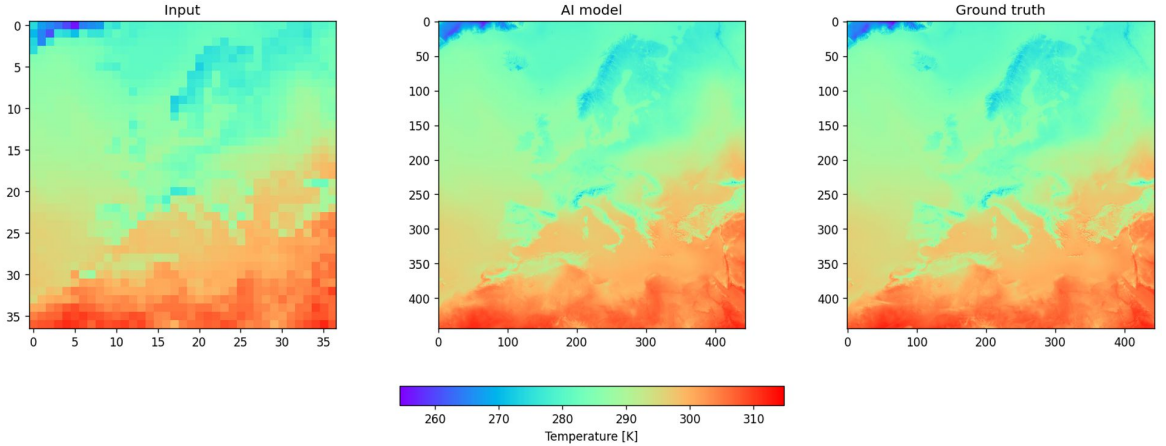
- 40 - yrs of Merra-2 for pre-training
- 163 variables - > 150TB
- Time-dependent inputs use 2 timestamps. Vertical, temporal and parameter dimensions are all stacked.

Encoder - 200 M Parameter
Decoder - 100 M parameter

Weather & Climate Foundation Model: Fine-tuning for downscaling application

Strategy for downscaling

Daily average temperature (EURO-CORDEX)



Initial Release Downstream Applications

1. downscaling
2. wind prediction at wind farms
3. gravity wave flux parameterization
4. hurricane intensity and track estimation

Credit: IBM Research

Science 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

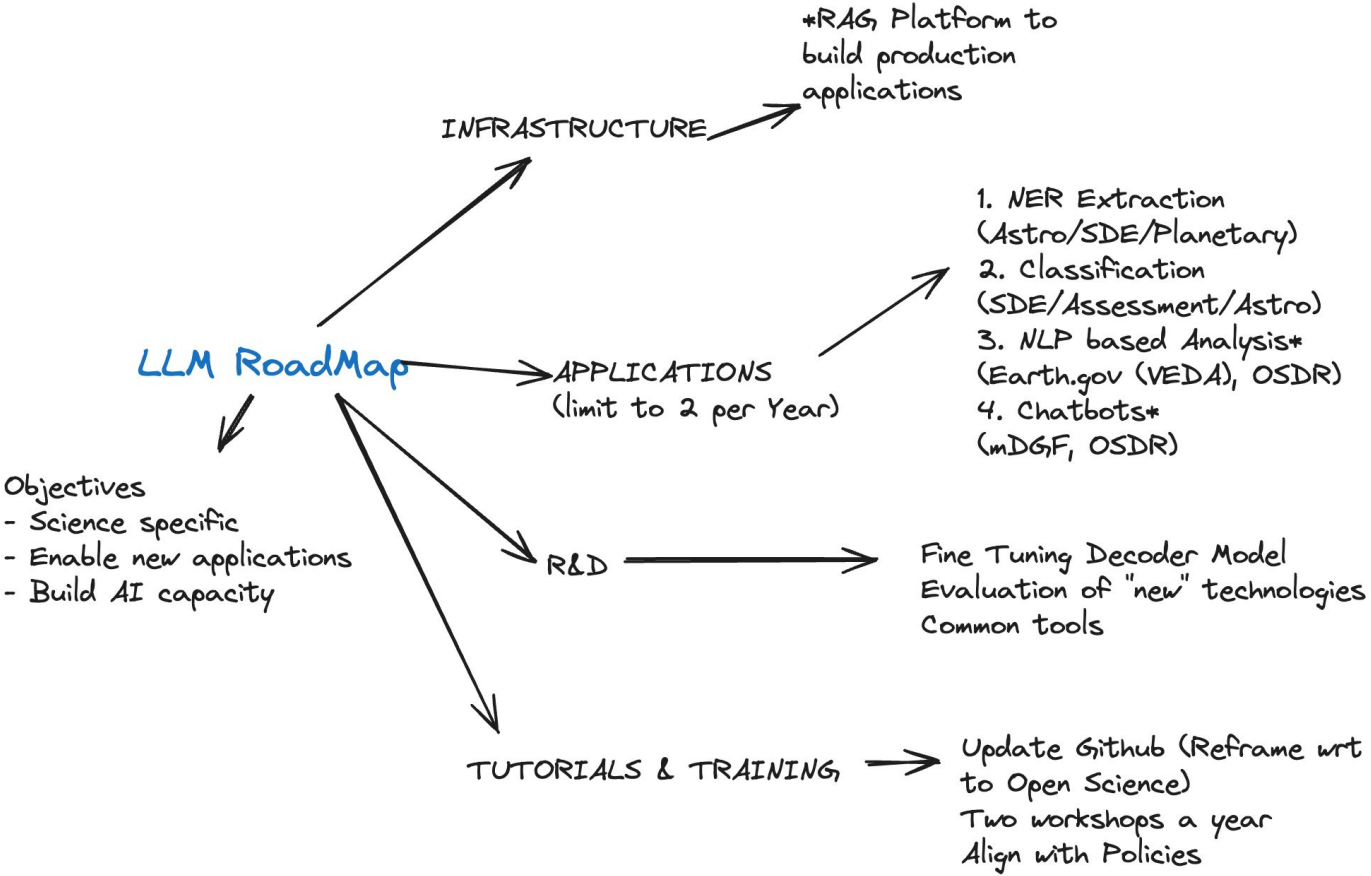
Useful in question-answering systems, document retrieval, and chatbots

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%

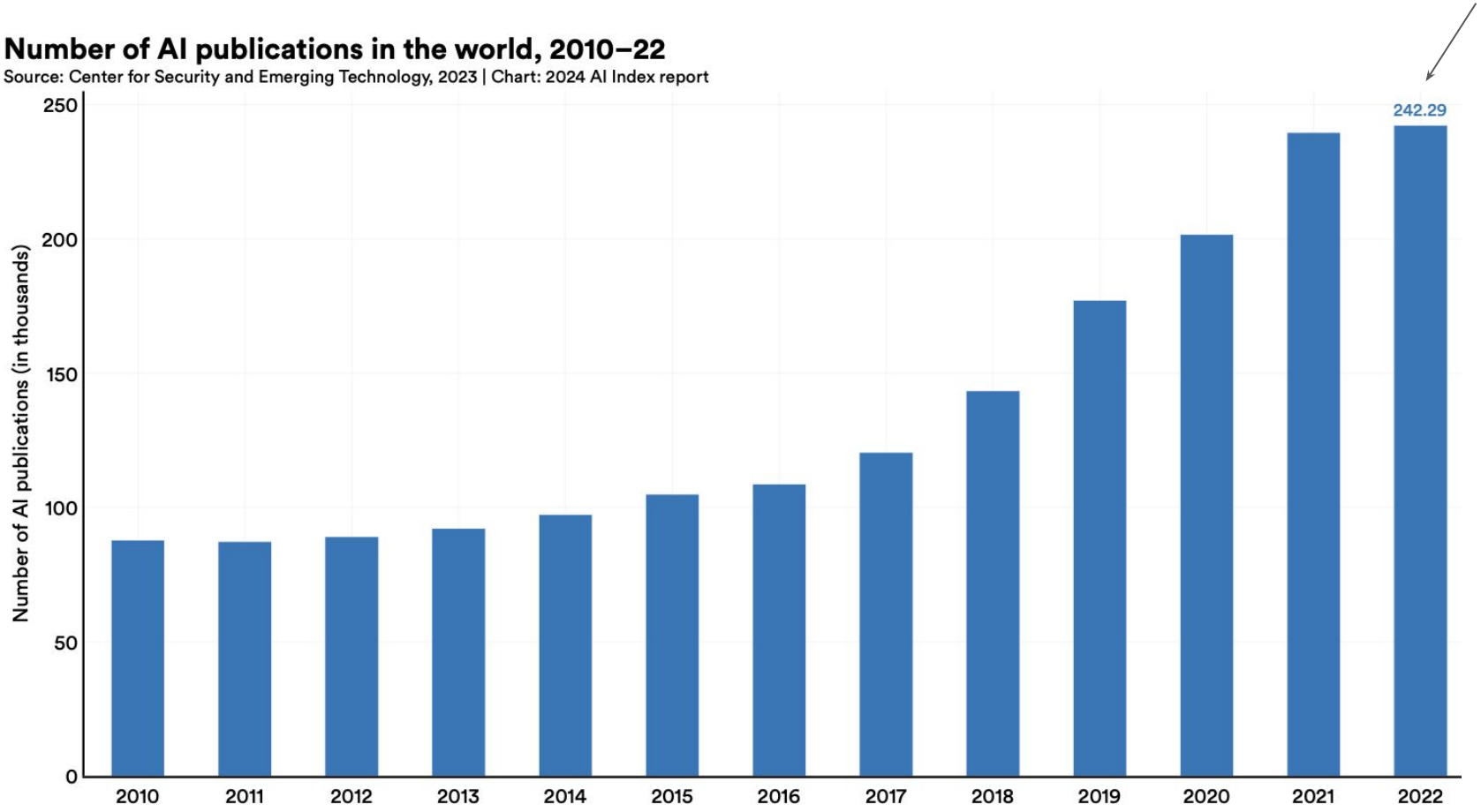
Curated resources for training

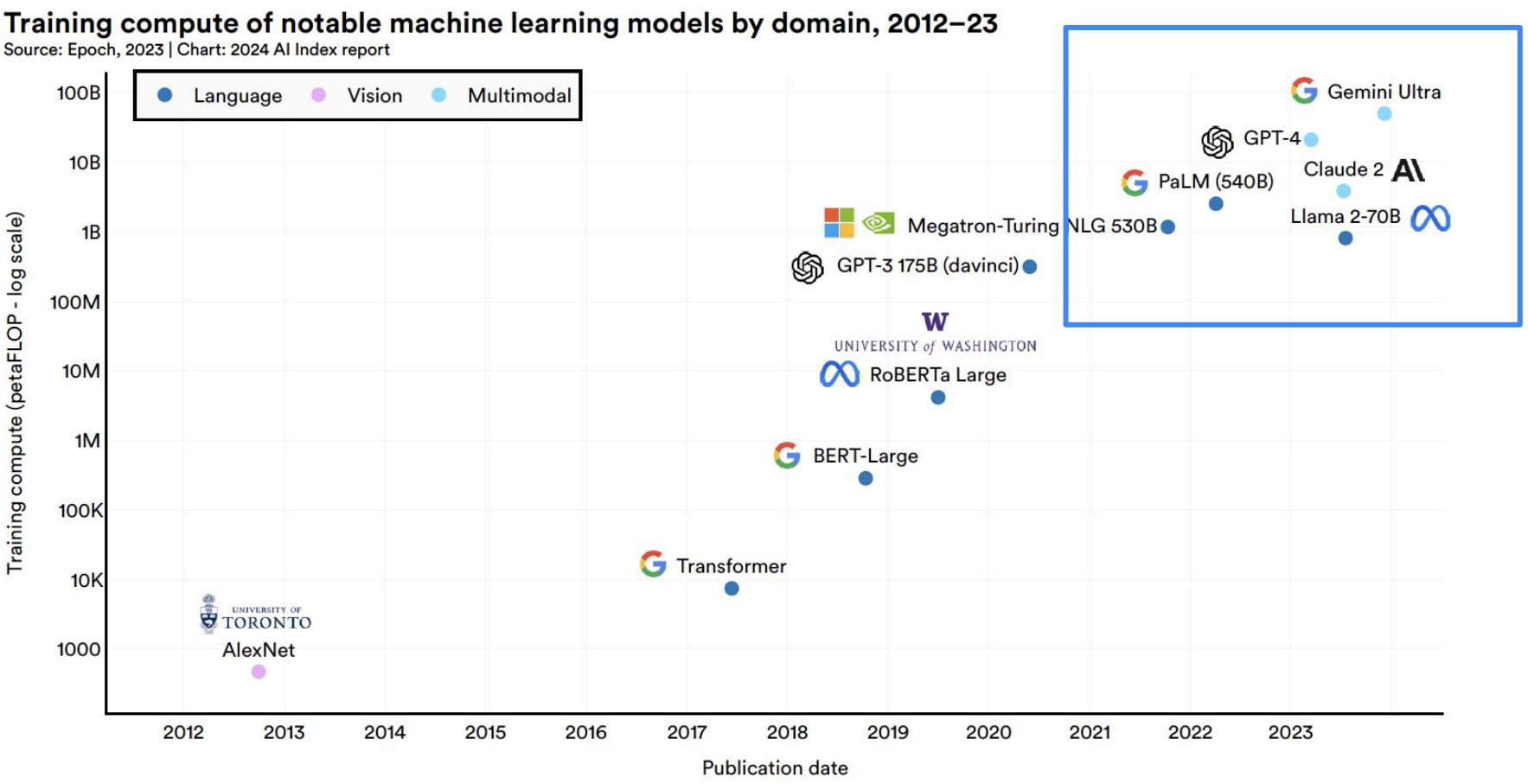
Path forward

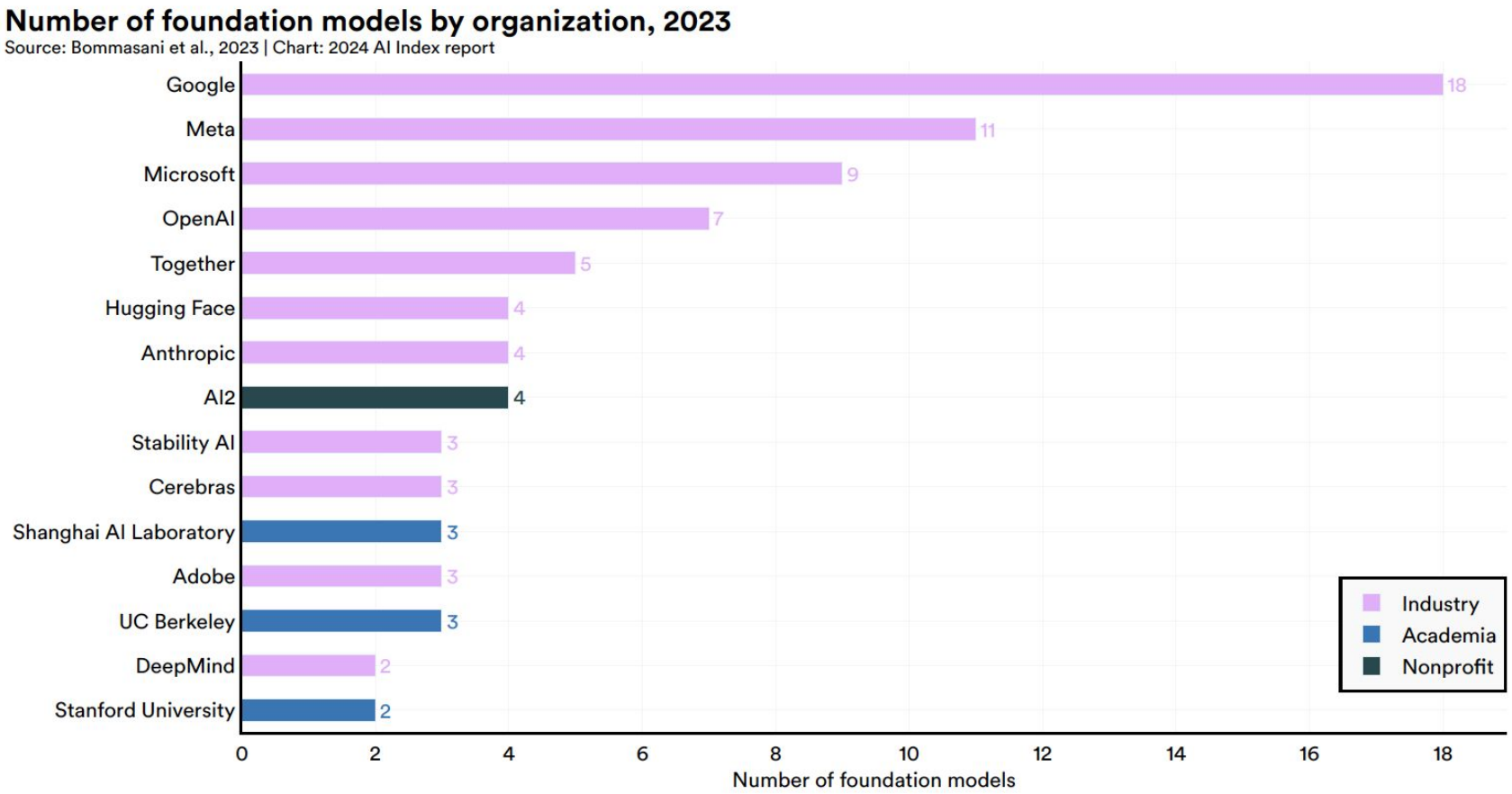


Our principles

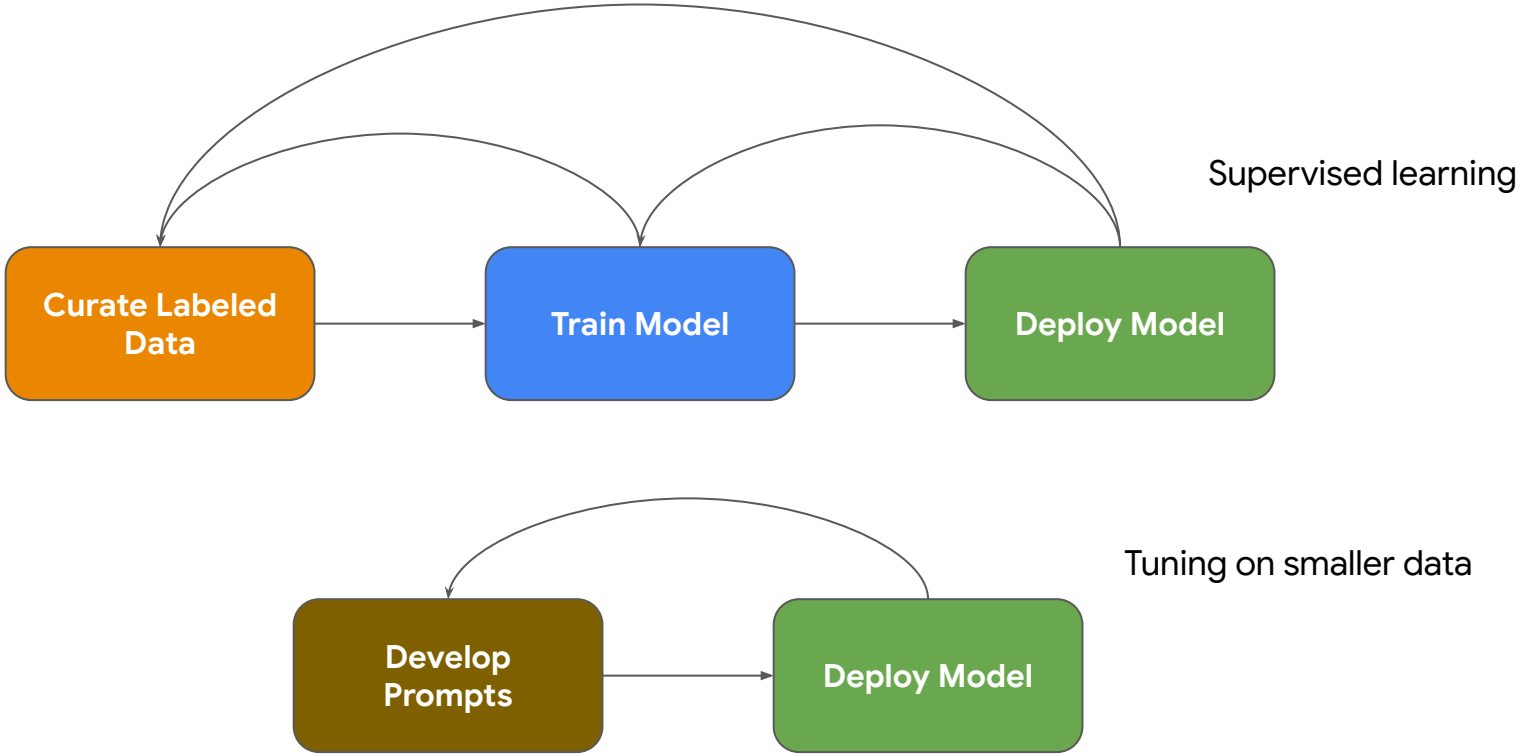
- Open science
- Core focus on science and applications
- Early involvement of domain experts
- Continuous evaluation of impact/values - benefits are multifaceted
- Inclusion of multi-party stakeholders
- Aim for extreme transparency - encourage participation and trust


Number of AI publications in the world, 2010–22
Source: Center for Security and Emerging Technology, 2023 | Chart: 2024 AI Index report
Number of AI publications (in thousands)
250
200
150
100
50
0
242.29
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

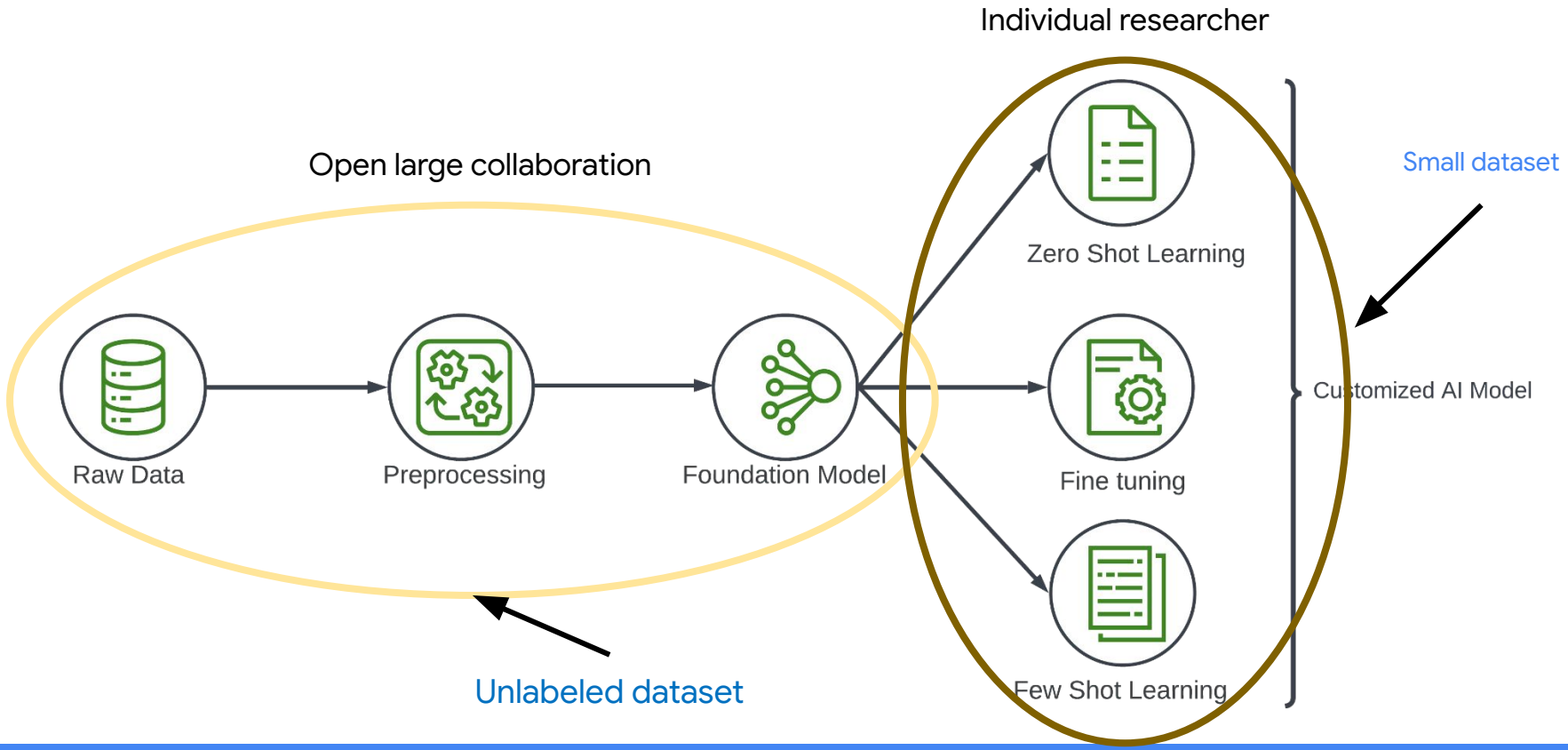

Training compute of notable machine learning models by domain, 2012–23
Source: Epoch, 2023 | Chart: 2024 AI Index report
Language
Vision
Multimodal
Training compute (petaFLOP - log scale)
100B
10B
1B
100M
10M
1M
100K
10K
1000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Publication date
Gemini Ultra
GPT-4
Claude 2
PaLM (540B)
Llama 2-70B
Megatron-Turing NLG 530B
GPT-3 175B (davinci)
UNIVERSITY of WASHINGTON
RoBERTa Large
BERT-Large
Transformer
UNIVERSITY OF TORONTO
AlexNet


Number of foundation models by organization, 2023
Source: Bommasani et al., 2023 | Chart: 2024 AI Index report
Google
Meta
Microsoft
OpenAI
Together
Hugging Face
Anthropic
AI2
Stability AI
Cerebras
Shanghai AI Laboratory
Adobe
UC Berkeley
DeepMind
Stanford University
18
11
9
7
5
4
4
4
3
3
3
3
3
2
2
0
2
4
6
8
10
12
14
16
18
Number of foundation models
Industry
Academia
Nonprofit

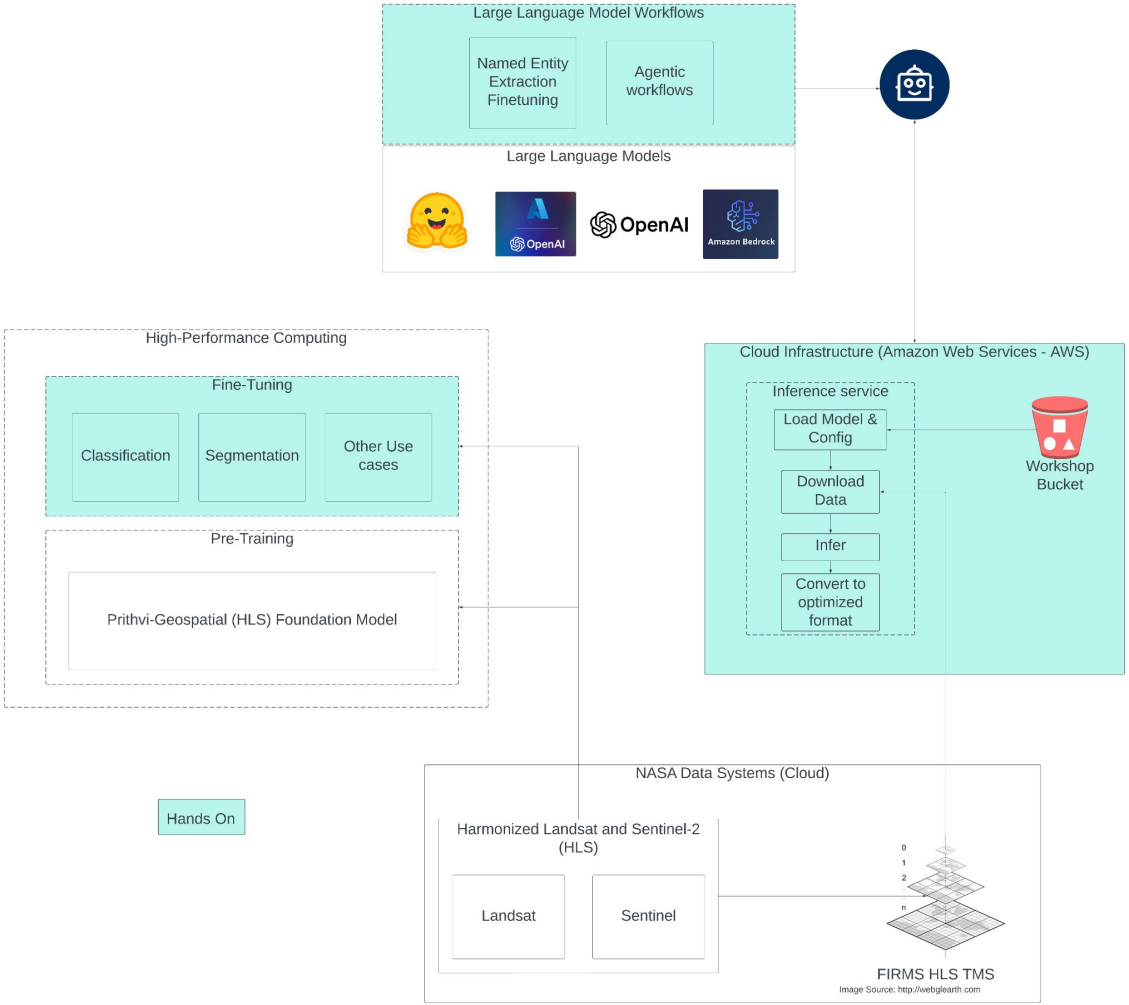
Changing paradigm



Reduced data and resource needs for custom models



Plan for today and tomorrow



Takeaway: Expanding proficiency in Large vision and language models

1. Using LLM and FM to enhance research processes and scale analysis.
2. Establish a data science environment and conduct interactive analysis.
3. Develop a strong understanding in geospatial foundation models.
4. Apply a finely-tuned geospatial foundation model effectively.
5. Learn best practices to maximize potential of LLMs.

We hope.....

You learn something new. Something you didn't know before you came here and will challenge you to go back and learn more

You meet a potential collaborator

You feel listened to

You have a good time

Thank you

manil.maskey@nasa.gov

- IEEE GRSS
- Earth Science Informatics TC
- HDCRS WG
- AIFMDT WG
- University of Santiago de Compostela
- IBM Research
- All the participants
- Iksha & Kumar - NASA IMPACT