

Will AI Foundation Models Transform Scientific Research?

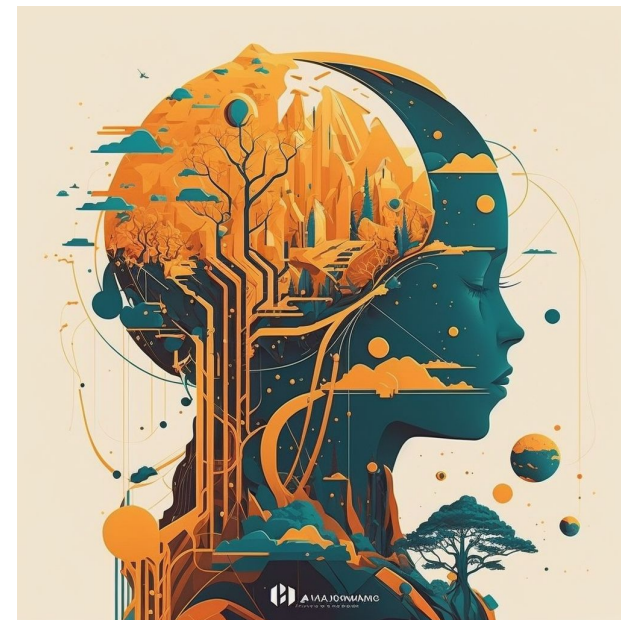
Dr. Rahul Ramachandran*
Senior Research Scientist, NASA/MSFC
Project Manager, IMPACT

Contact:
rahul.ramachandran@nasa.gov



Presentation Outline

- Programmatic Overview (drivers)
- Inflection Point in AI
 - Foundation Models
- Large Language Model - Foundation Model Exemplar
- Geospatial Foundation Models
 - HLS - Optical Remote Sensing Data
 - Weather/Climate (initial results)
- Summary
 - What will the future look like?
 - Takeaways/Challenges



The background of the slide features a light blue architectural blueprint with various lines and structures. Overlaid on the right side is a large, tilted sequence of numbers in a darker blue font. The numbers are arranged in a grid-like pattern, with some numbers appearing larger than others, creating a sense of depth and complexity. The overall aesthetic is technical and modern.

Programmatic Overview



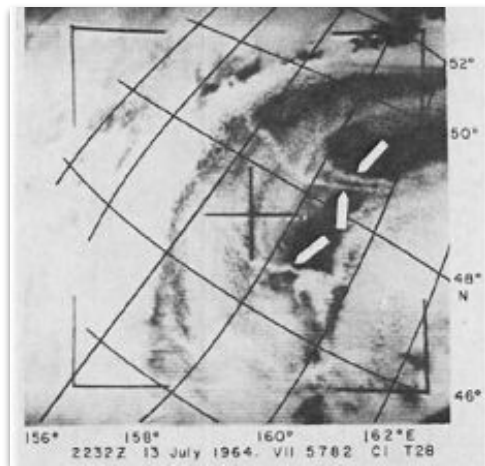
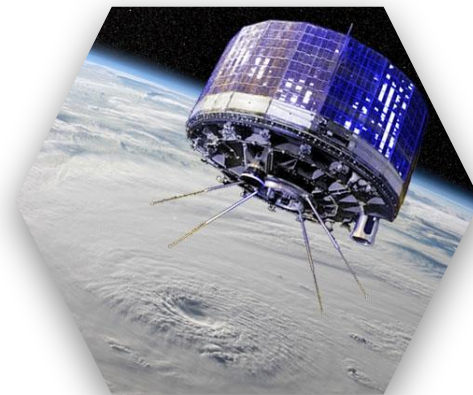
Value of Our Science Data

Established by NASA in 1958

Experimental satellites developed with the Weather Bureau

First launched in April 1960

Series continued into the 1980s



TIROS-1 (or TIROS-A) was the first successful weather satellite launched by NASA

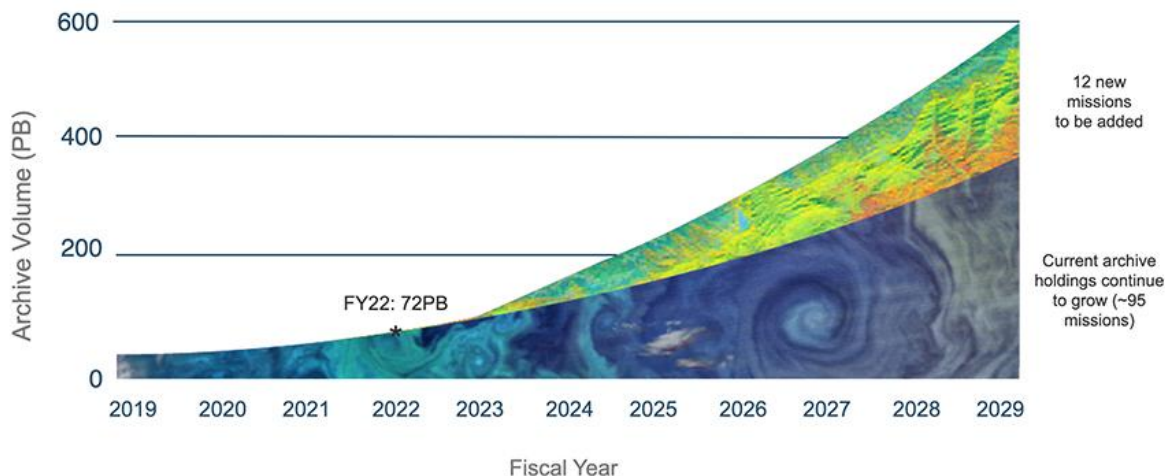
Observation of anomalous cloud lines in the TIROS satellite imagery by Conover (1966), which he attributed to the impact of smoke emissions from ships on marine stratus clouds (Noone, 2000). Twomey et al. (1968) later showed that this phenomenon is caused by aerosols in ship exhaust functioning as cloud condensation nuclei increases the albedo of clouds. Twomey's work eventually motivated the research on the indirect radiative effect of aerosols, a [field of research active to the present day](#).

NASA data changes the direction of science.



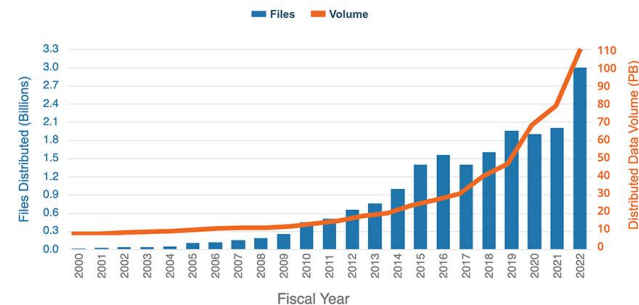
Metrics from 2022

Earth Science Data Archive Growth Projection

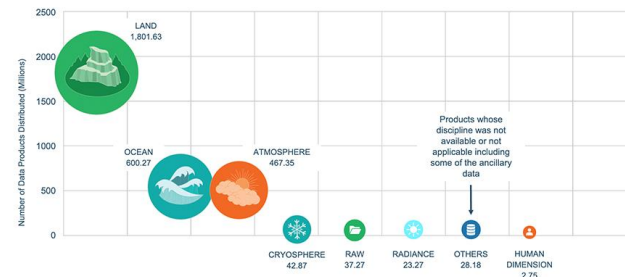


ESDS faces a scaling problem!

Total Data Volume and Data Files Distributed by Year

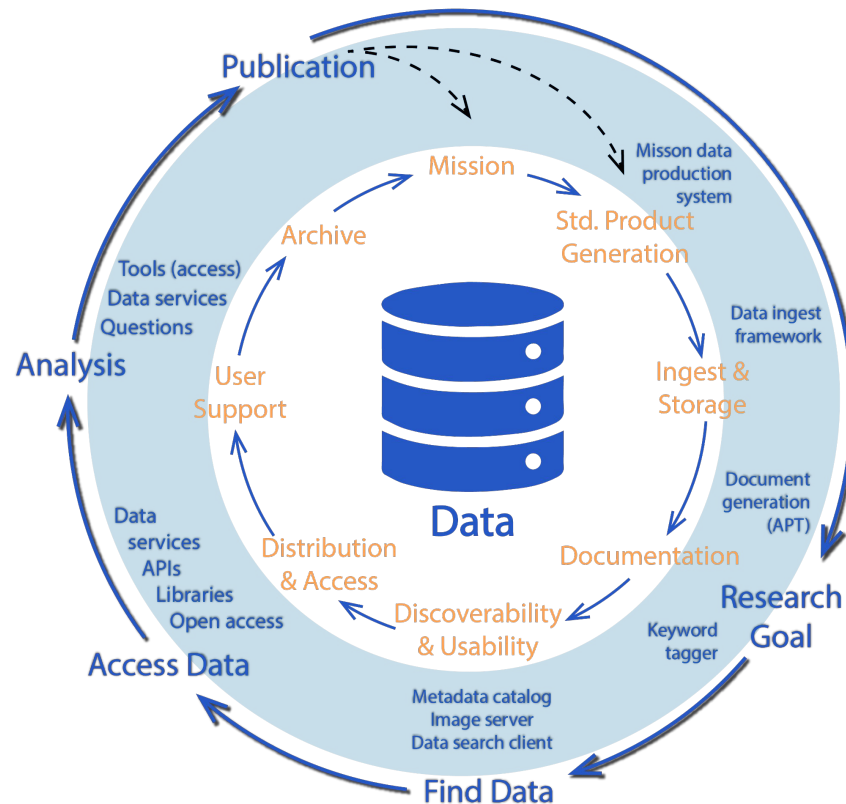


FY2022 Number of Data Products Distributed by Discipline

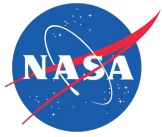




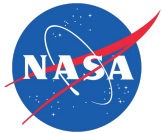
Building Blocks



Data + Tools/Infrastructure -> Enable Research Lifecycle



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



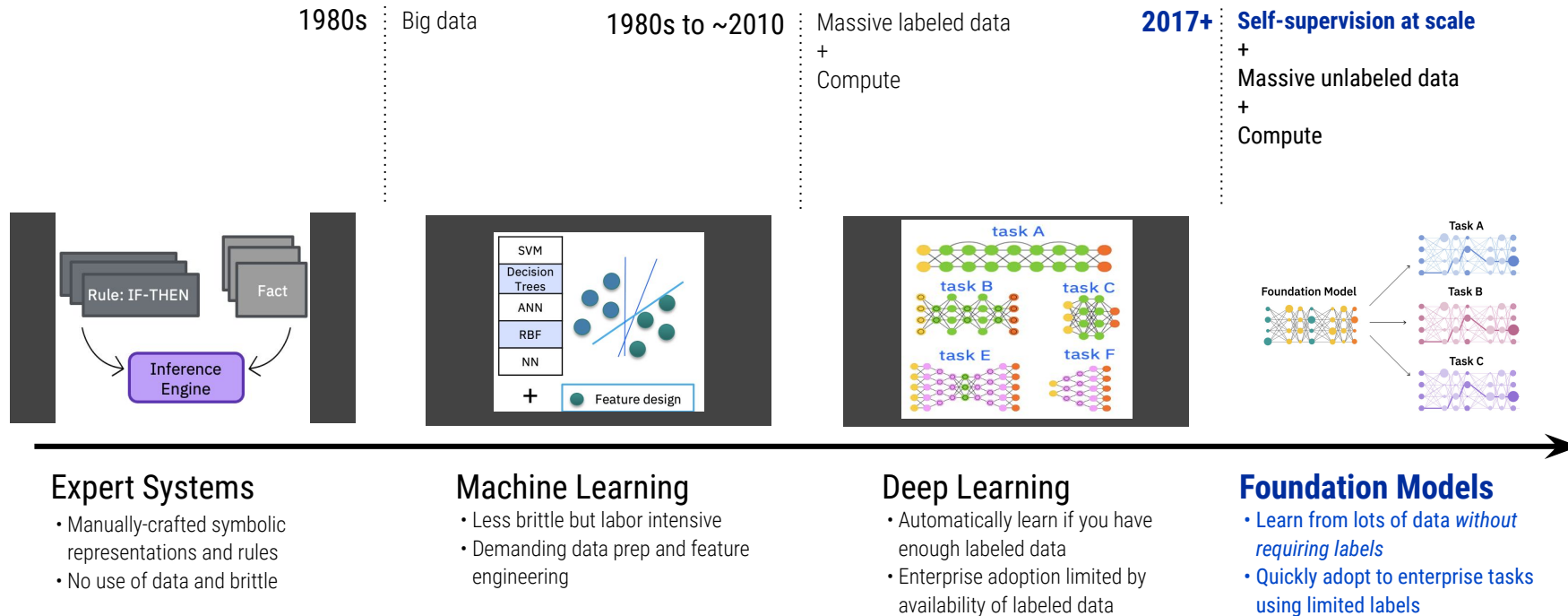
Augmentation not Automation

The background is a light blue abstract composition. It features a dense network of thin, white, perspective-drawn lines that create a sense of depth and structure, reminiscent of a complex architectural plan or a data visualization. Overlaid on these lines are several large, semi-transparent numbers in a dark blue font. These numbers are arranged in a way that suggests a sequence or a path, with some numbers appearing to be part of a larger, partially visible sequence like '173354230931' and '1864'. The overall effect is one of technological complexity and forward momentum.

Inflection Point in AI



Poised to Dramatically Accelerate Enterprise AI Adoption





Foundation AI Models



- Pre-trained on a comprehensive data set and used for various downstream tasks
- Substantially reduce the downstream effort for building AI applications, including the need for large labeled training datasets
- Model captures **emergent behavior** within the modeled system from the data.

On the Opportunities and Risks of Foundation Models

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 Dorothea Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kavin 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 Faissal 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 Avaniika 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 Yussuf 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[†]

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.

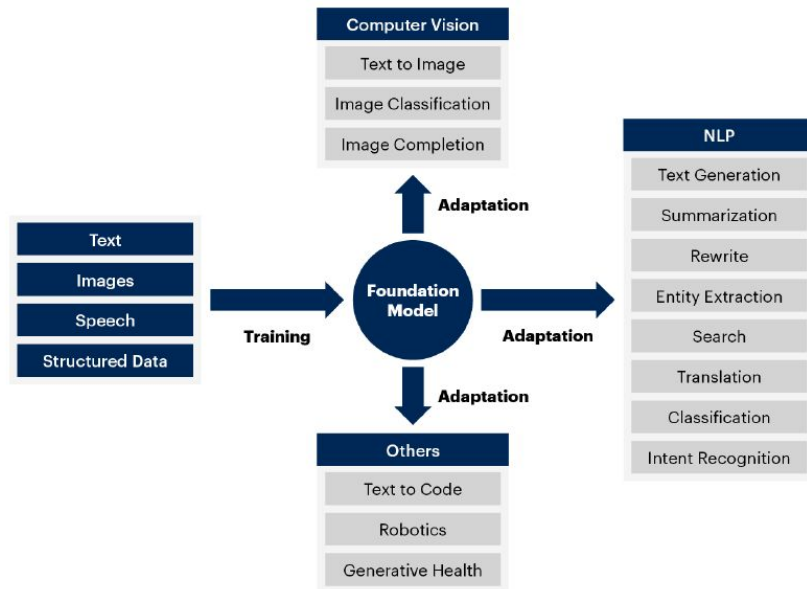
arXiv:2108.07258v2 [cs.LG] 18 Aug 2021



What Are Foundation Models?

- Foundation models (FM) are AI models that are **designed to replace task-specific models and be applied to different downstream applications**.
- FMs 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:
 - Leverage 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
 - Can be applied to downstream tasks by **using few-shot learning and fine tuning**
 - Some have to be **trained at scales** that limits the ability to a handful organizations

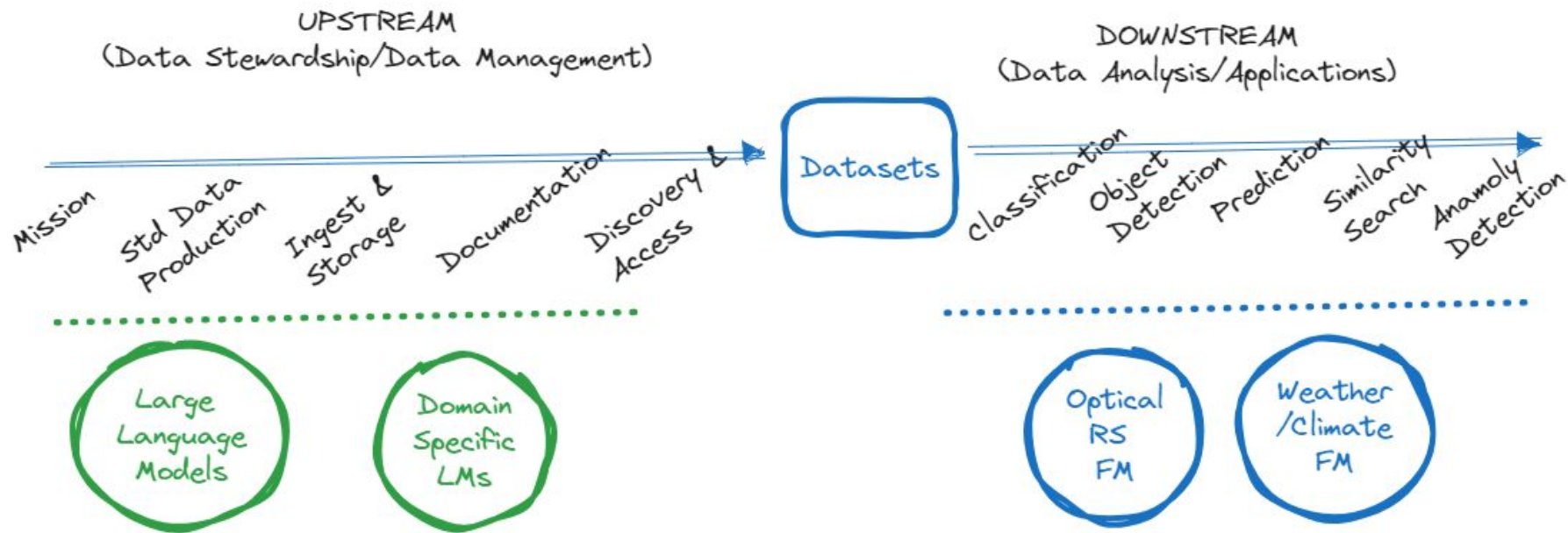
Foundation Models - Characteristics and Applications



Source: Gartner
769102_C



AI FM Adoption Both Upstream and Downstream

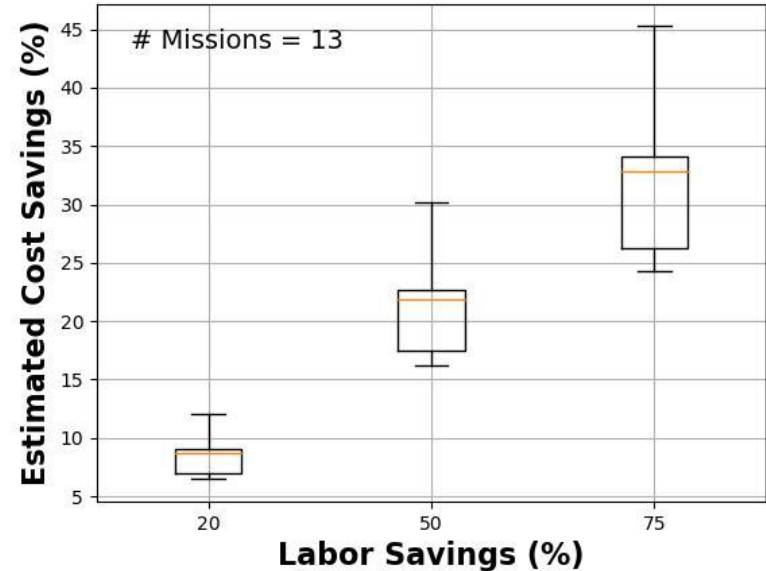




Impact of ML and AI on Mission Cost: Ingest and Archive



- Data: 13 actual budget estimates from MCE team
- Of 66 defined tasks for data ingesting and archiving, **22 can be augmented by AI**.
- Greatest number of tasks (10) are in “Science Data and Engineering”.
- Greatest density of tasks are in “Science Data and Engineering” (10/19) followed by “Information Support” (2/4).
- Examples of potential labor savings include:
 - Streamlining user support with a chatbot
 - Write dataset documentation
 - Validation of delivered metadata (i.e., abstract quality control)



Box and whisker plot showing estimated savings by using NLP in data ingestion and archiving. Yellow line is mean, boxes are the 1st and 3rd quartiles, and whiskers show the range.



Downstream: 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.

Potential innovative solutions in which to invest include techniques that require less data, such as transfer learning, representation learning, semi-supervised learning, and unsupervised learning.



Why Are FMs Important for Science?

Scale of these models results in better performance for downstream tasks.

- FM-based applications show better accuracy over models built from scratch.

Drastically reduce the cost of entry to build downstream applications

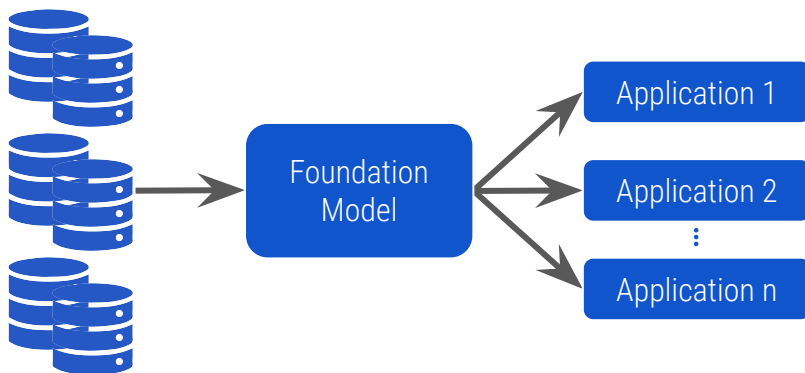
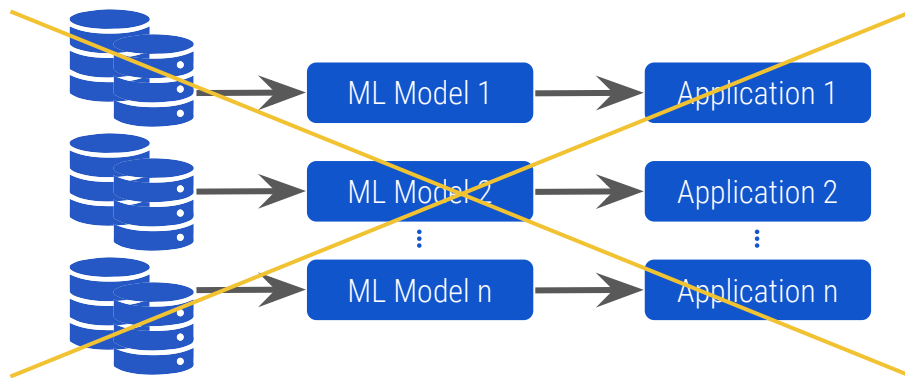
- Time/effort required to build and run an application is low compared to building from scratch.

FM models for select datasets can accelerate downstream applications (data quality monitoring, feature detection, monitoring, anomaly detection, prediction)

- Make it easier to infuse AI into scientific research and analysis
- Remove the training data bottleneck
- Increase the use of NASA data



Foundation AI Models for Earth Science Data

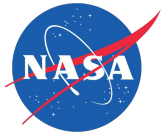


- Can we build foundation AI models for domains or representative types of data?
 - Optical remote sensing data, SAR, climate simulations
 - Can these models capture underlying physical processes?

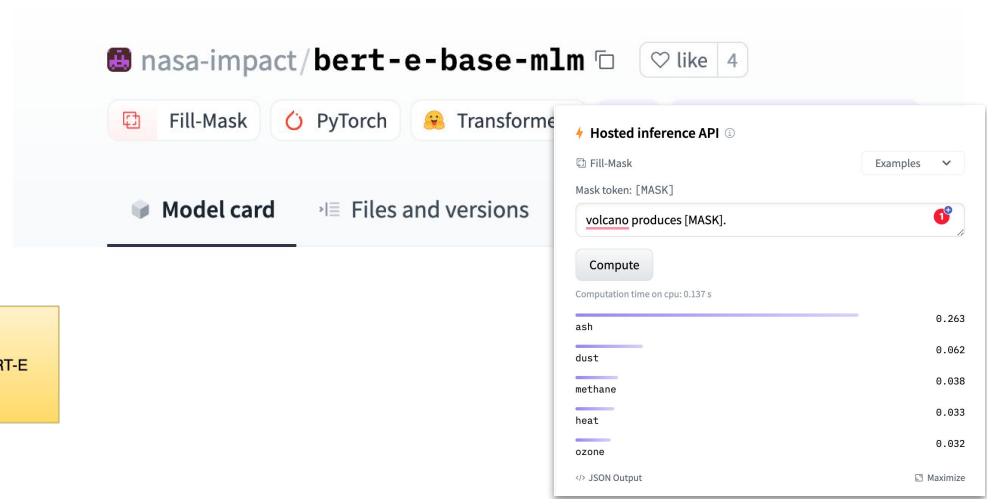
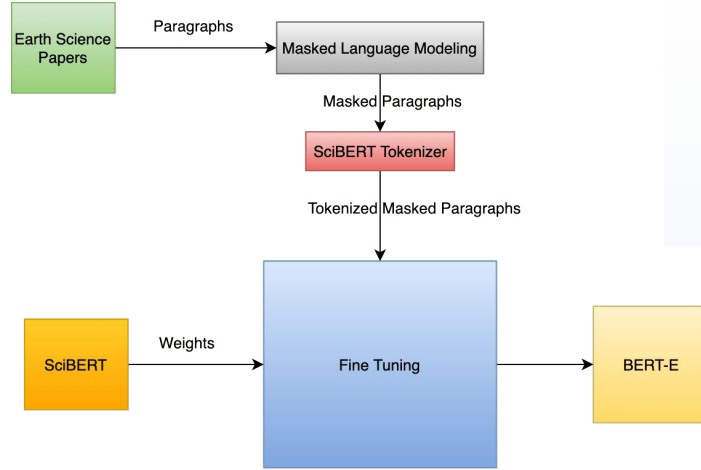
The background features a light blue architectural blueprint with various lines and structures. Overlaid on the right side is a large, tilted sequence of numbers: 173354230931, 5039756851864, 1923427064, 4249064, 196502944249064, 12000625598064.

Upstream: Large Language Models For Science

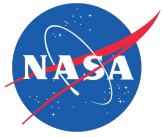
Collaboration with IBM Research



BERT-E



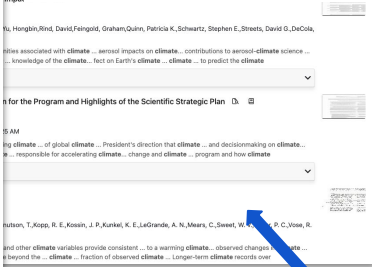
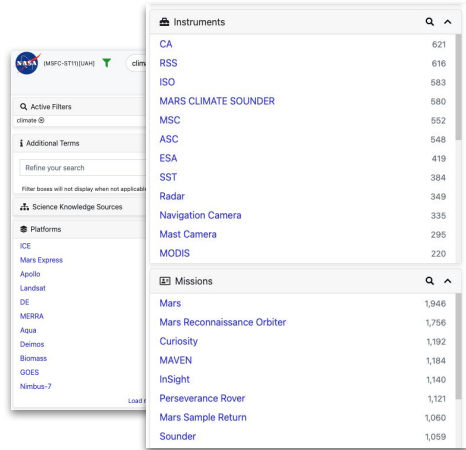
- Collected Earth science articles with a total of almost 6 million paragraphs
- Masked language modeling (MLM) is used to mask random words in the paragraph. The model trains by predicting the word and validating against the actual word.
- Initial model is trained using the abstracts from the collected corpus.
- Model performance on downstream tasks is limited due to training limited to just abstracts and the use of general science vocabulary.



AI Language Model to Improve Operational Processes



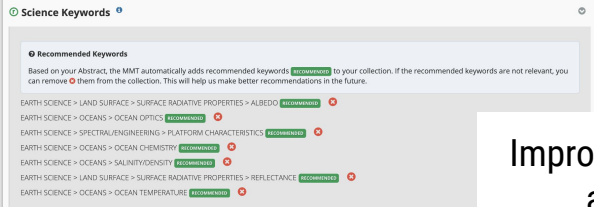
Augmenting facet design to
improve search and discovery



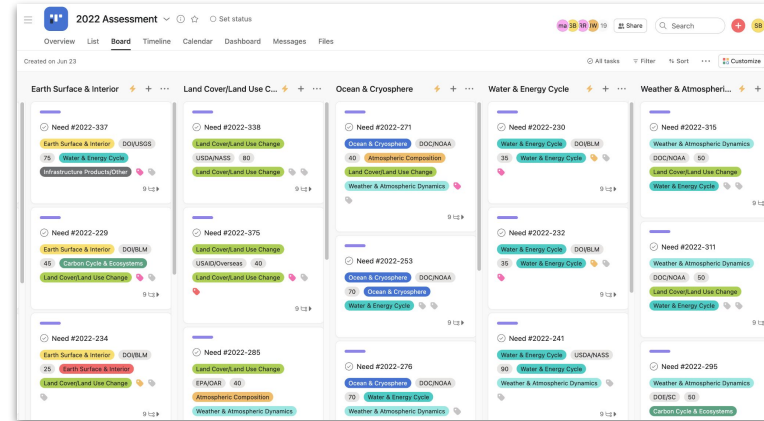
AI Language
Model

Using GKR

- GKR is currently only available through MMT (no direct API access)
- Available in UAT (<https://mmt.uat.earthdata.nasa.gov>) on Monday, September 13th
- When publishing a collection record through MMT in UAT, GKR recommendations will automatically show up in Science Keywords on the Descriptive Keywords tab:



Improving data stewardship by
automating keyword

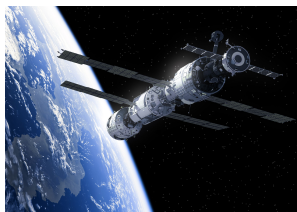


Automated binning of SNWG surveys into
NASA ES focus areas



Language Model for Earth Science

Train a new domain specific foundation model on Earth science literature that improves downstream tasks and enables information retrieval



- Models built using FM stack
- Standard RoBERTa base architecture (125M)
- 120k papers (after data preprocessing and duplicate removals, reduced from 275k papers)
- 64 NVIDIA A100 GPUs
- 10 hrs of training
- 67k words/GPU/sec throughput

NASA RoBERTa model

CAPE increase due to [*anthropogenic greenhouse warming*].

The second most abundant element in the atmosphere is [*oxygen*].

[*Oxygen*] atoms combine to form dioxygen.

An igneous rock is a rock that crystallizes from [*magma*].

An igneous rock is a rock that crystallizes from [*igneous melt*].

Standard English RoBERTa

CAPE increase due to [*weather*].

The second most abundant element in the atmosphere is [*water*].

[*The hydrogen*] atoms combine to form dioxygen.

An igneous rock is a rock that crystallizes from [*ashes*].

An igneous rock is a rock that crystallizes from [*the ground*].



Using the Language Model for Question Answering



Question Answering

What would you like to know?

Is there a dust index that can be used to detect airborne dust over china

Ask

Question

Is there a dust index that can be used to detect airborne dust over china

Found 15 answers matching your question.

Answer

A dust aerosol index (DAI) algorithm based on measurements in deep blue (412 nm), blue (440 nm), and shortwave IR (2130 nm) wavelengths using Moderate Resolution Imaging Spectroradiometer (MODIS) observations has been developed

Confidence: 52%

Was this answer useful?

Answer

dust aerosol index (DAI) algorithm based on measurements in deep blue (412 nm), blue (440 nm), and shortwave IR (2130 nm) wavelengths using Moderate Resolution Imaging Spectroradiometer (MODIS) observations has been developed

Confidence: 31%

Was this answer useful?

Answer

Infrared Difference Dust Index (IDDI).

Confidence: 39%

Was this answer useful?

Answer

Daily aerosol optical thickness (AOT) at 0.55 μm over the desert regions is needed as a source of validation for numerical models such as the United Kingdoms Numerical Weather Prediction Unified Model. We examined the relationship between monthly mean ultraviolet (UV) absorbing aerosol index (AI) from the Ozone Monitoring Instrument (OMI) that is available on a daily basis with the Multiangle Imaging Spectroradiometer (MISR) AOT that is available once every nine days over North Africa

Confidence: 43%

Was this answer useful?

Retriever

ColBERTRetriever

Select a retriever

Retriever Settings

Checkpoint

Maximum number of retrieved documents

1 100 5

Corpus

nasa-corpus

Select a corpus

Reader

ExtractiveReader

Select a reader

Reader Settings

Model

PrimeQA/nq_tydi_sq1-reader-xlmr_large-202211

Maximum number of answers

1 5 3

Maximum answer length

2 2000 100

Proof of concept demo (IBM Research collaboration)



Enhancing Scientific Efficiency through AI (FIRMS Q&A)

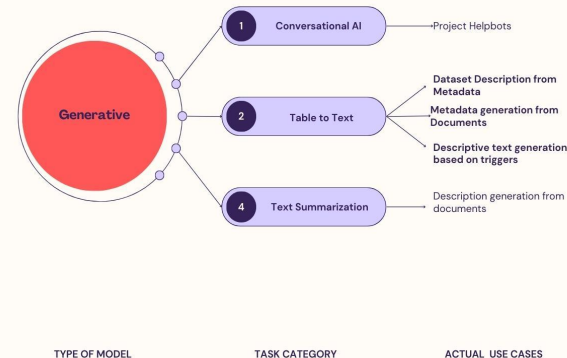
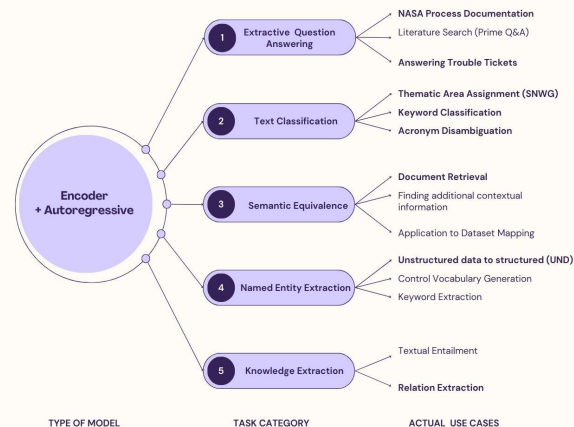
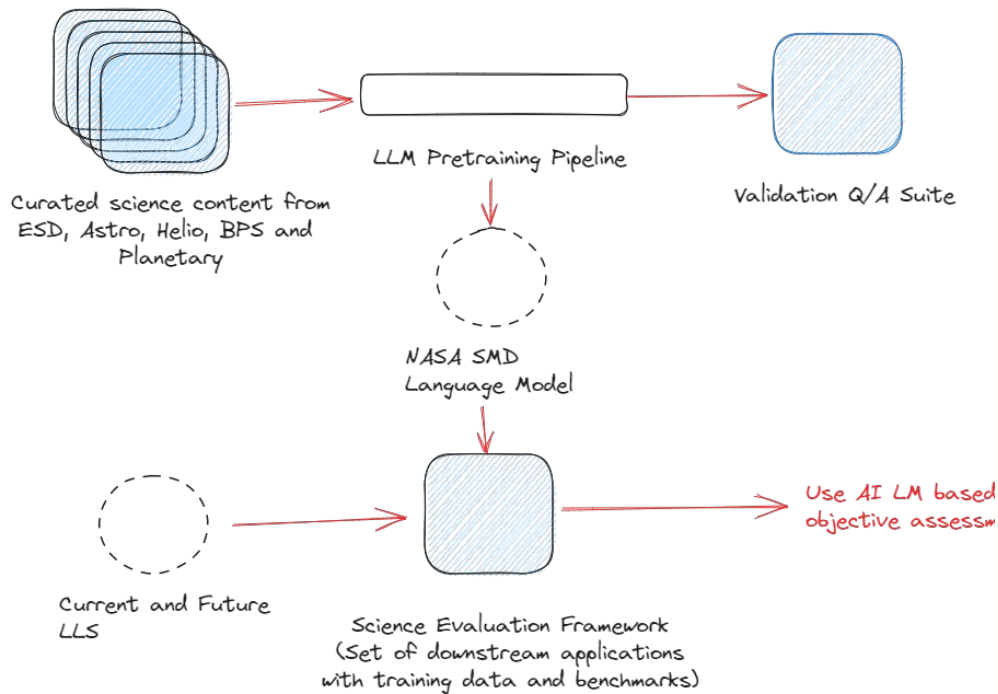


The screenshot displays a Jupyter Notebook environment with the following elements:

- Header:** The notebook is titled "NLP-API-trend-analysis". The top bar includes a menu (File, Edit, View, Insert, Runtime, Tools, Help), a status message "All changes saved", and action buttons for Comment, Share, and settings.
- Left Sidebar:** Contains icons for file management, search, and a variable explorer showing a variable named "result".
- Code Cell:** The first cell contains the code: `[139] answer, result = query("how many active fires were there in canada yesterday?")`. It has a green checkmark icon on the left.
- Text Cell:** The second cell contains the text: `[140] answer`. It also has a green checkmark icon.
- Output Cell:** The third cell shows the output of the query, labeled "result", with a play button icon. It includes a toolbar with navigation and action icons.
- Bottom Status Bar:** Shows a green checkmark, "0s", and the text "completed at 10:14AM".



Foundation Language Model for Science

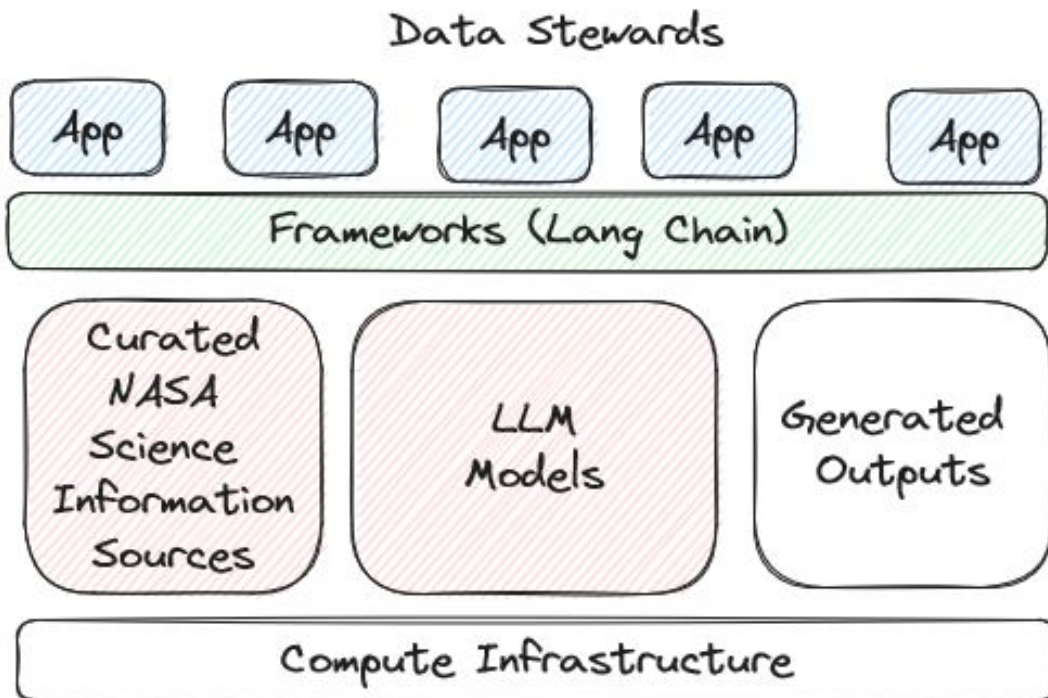


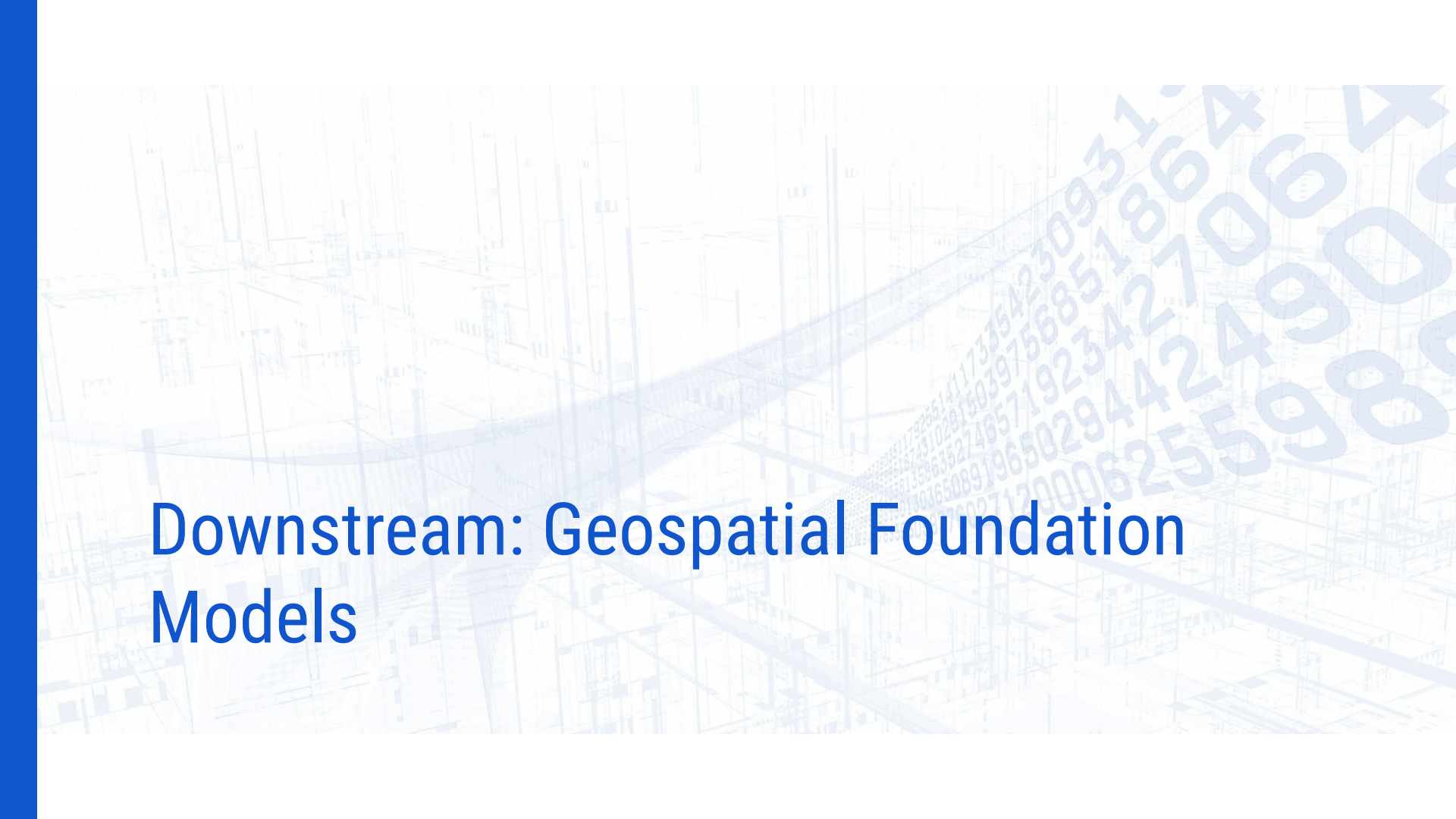


Platform for Building LLM-Based Data Stewardship Applications



LLM Applications for Data Stewardship/Management



The background features a light blue architectural blueprint with various lines and structures. Overlaid on the right side is a large, tilted sequence of numbers in a darker blue font. The numbers are arranged in a grid-like pattern, with some numbers appearing to be part of a larger sequence or code.

Downstream: Geospatial Foundation Models

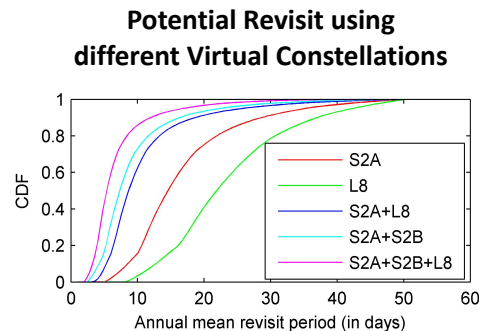
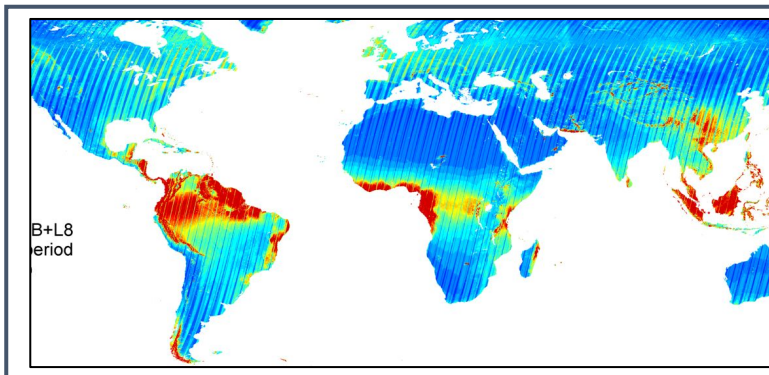
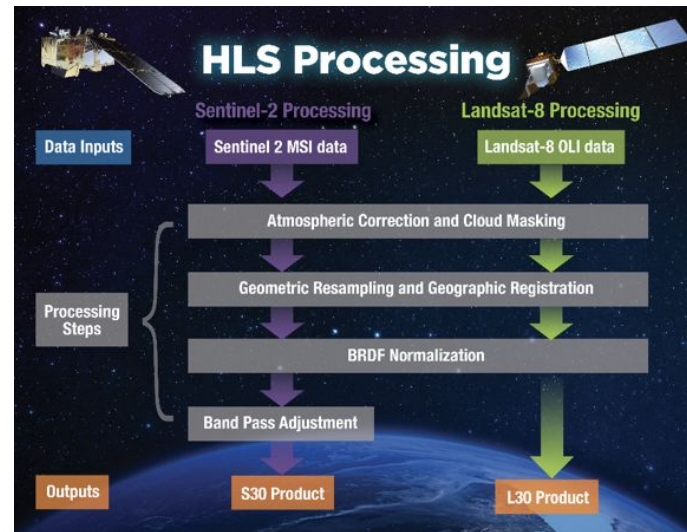


Foundation AI Model for Optical Remote Sensing Data



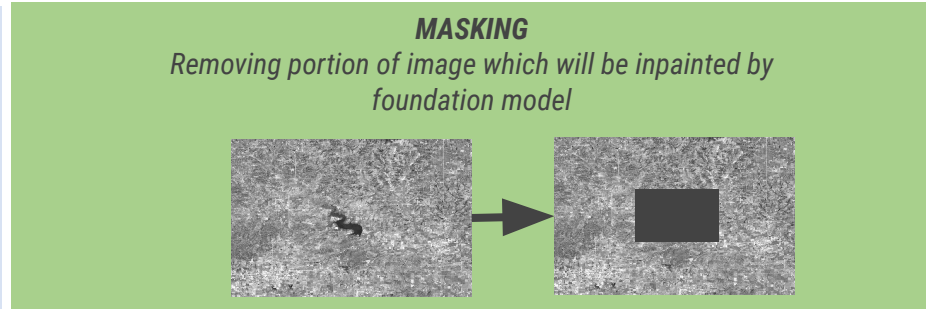
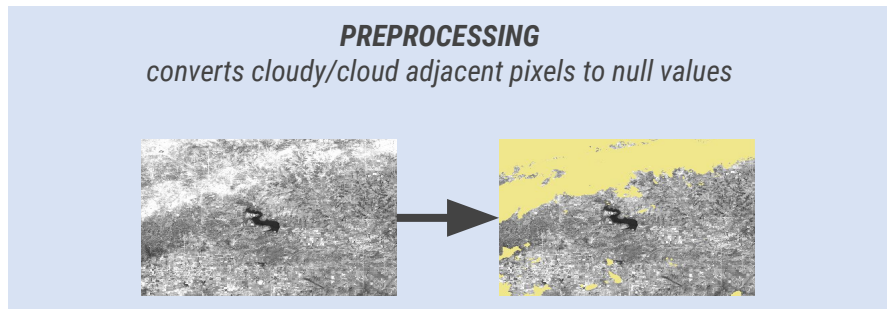
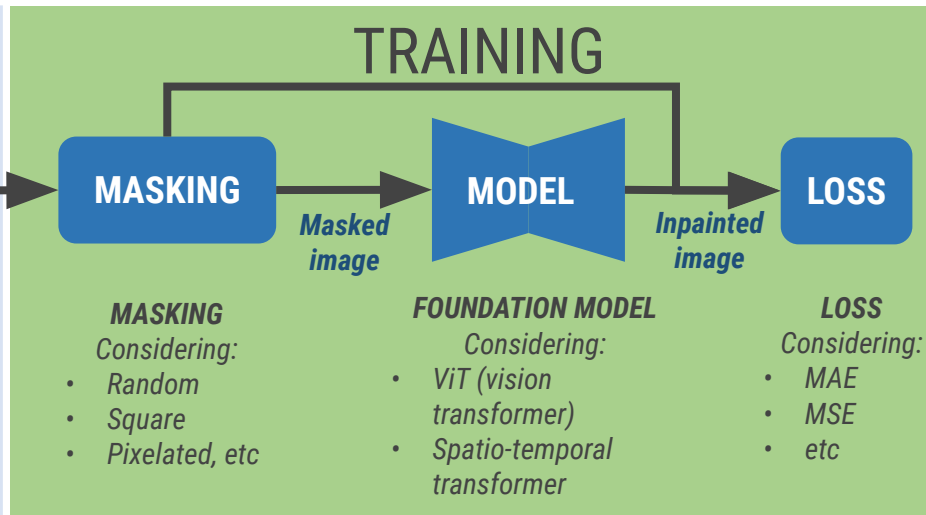
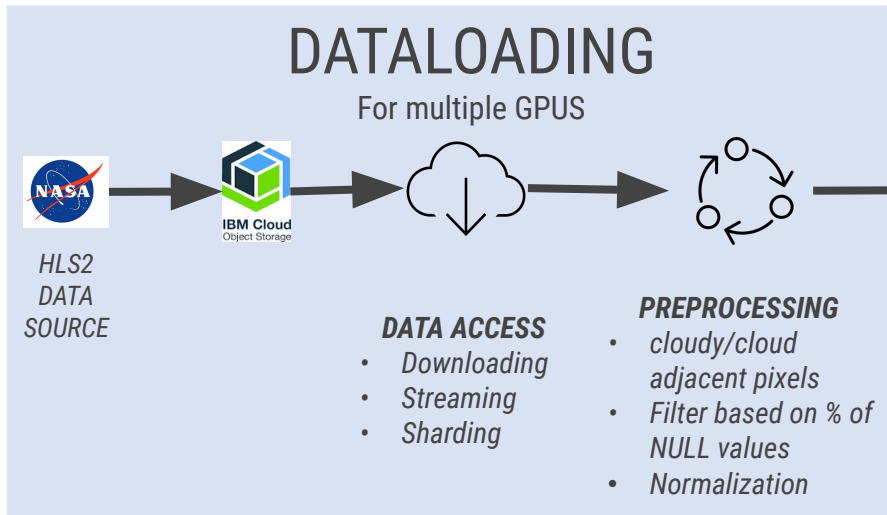
Harmonized Landsat Sentinel (HLS)

- “Seamless” near-daily 30m surface reflectance record including atmospheric corrections, spectral and BRDF adjustments, regridding
- Merges Sentinel-2 and Landsat data streams and can provide **2-4 day global coverage**



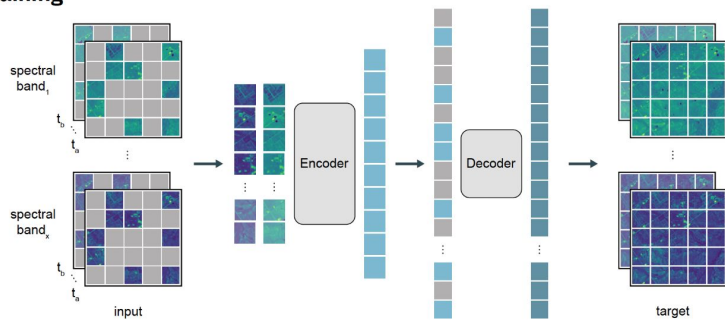


HLS FM Training Pipeline

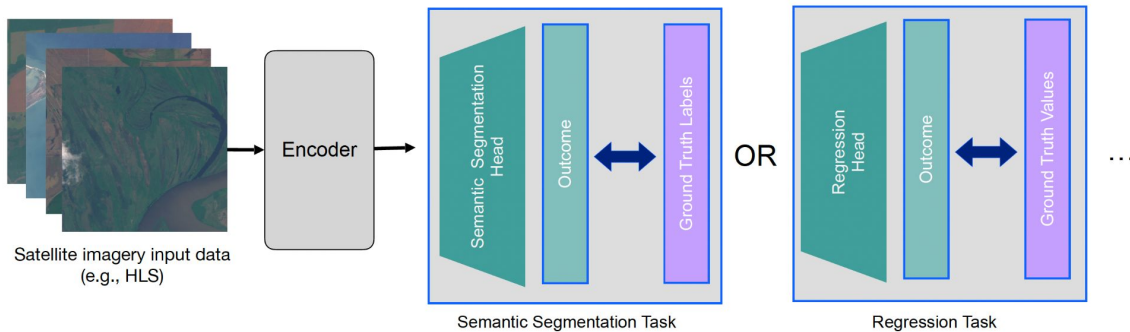


HLS FM Training Pipeline

1. Prithvi Pre-training

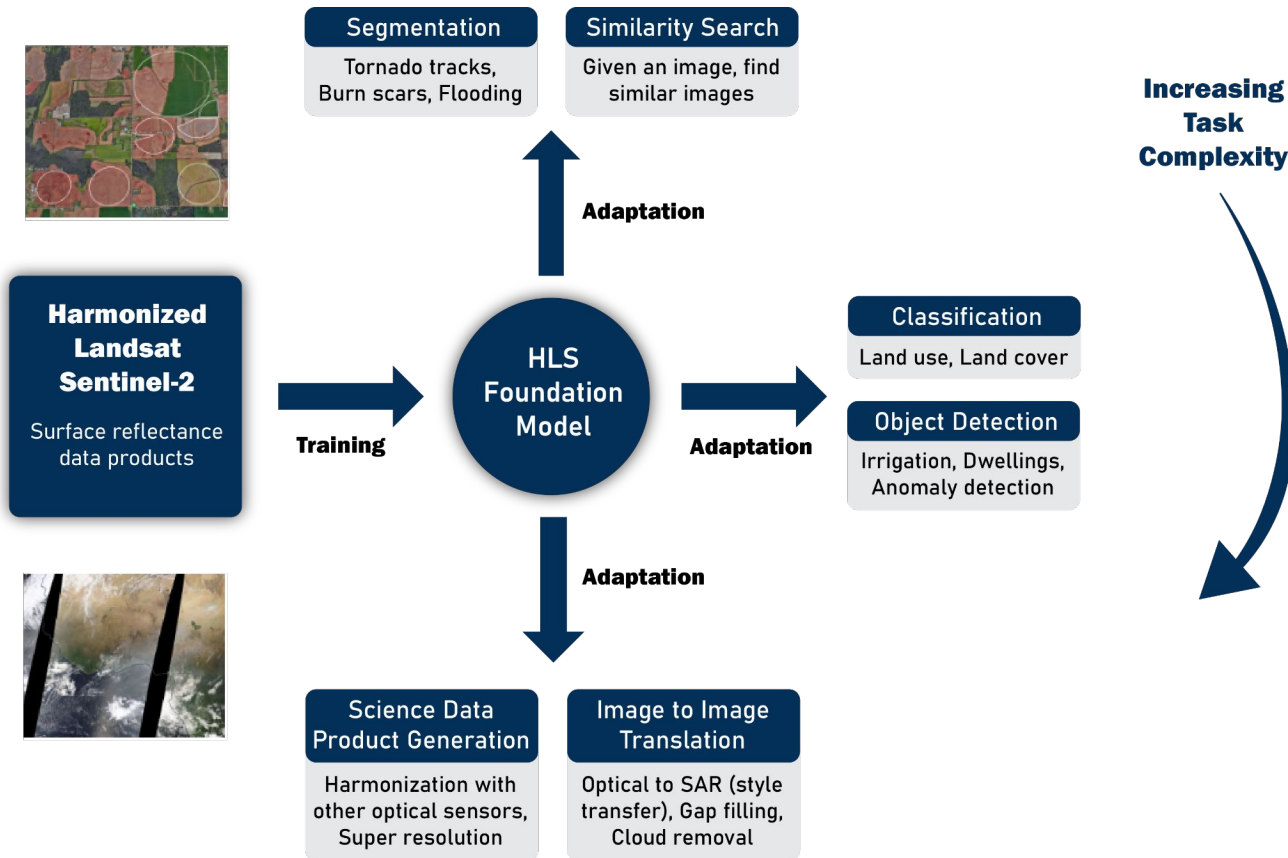


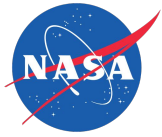
2. Prithvi Fine-Tuning



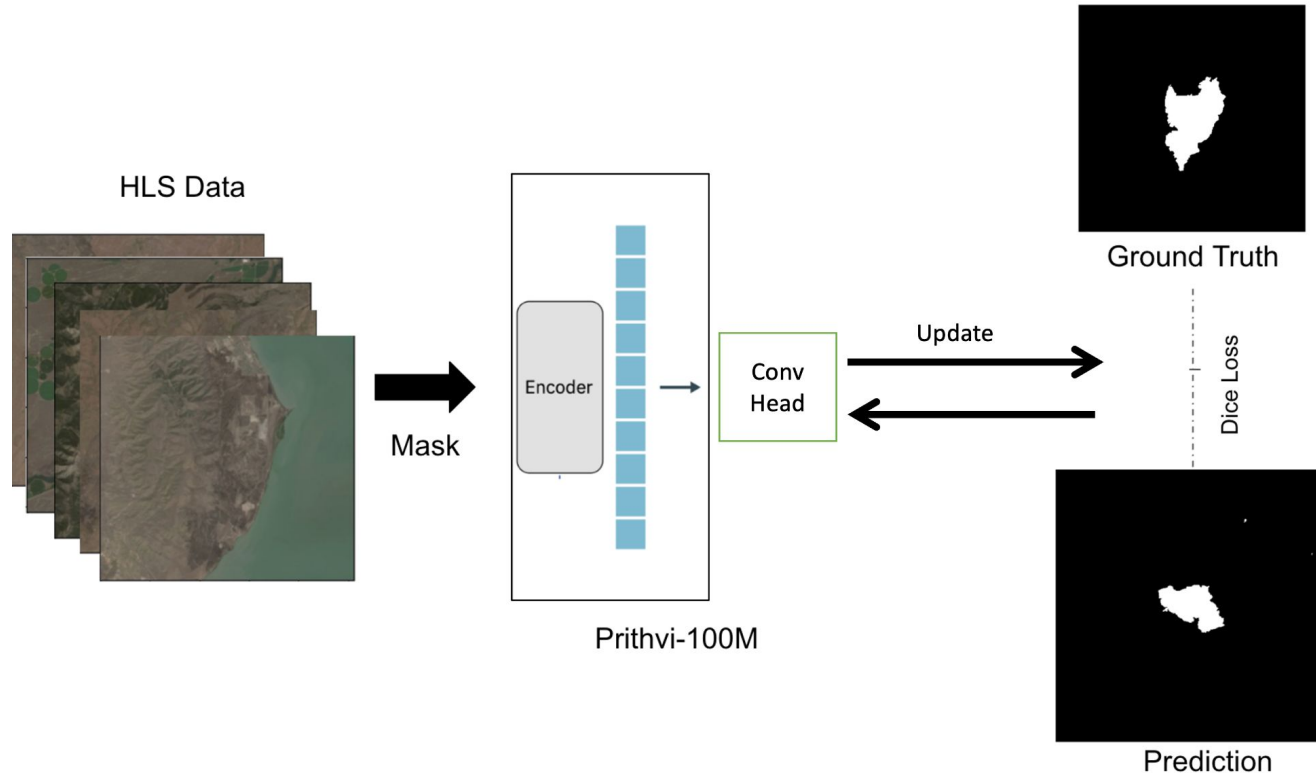


HLS FM Downstream Adaptation



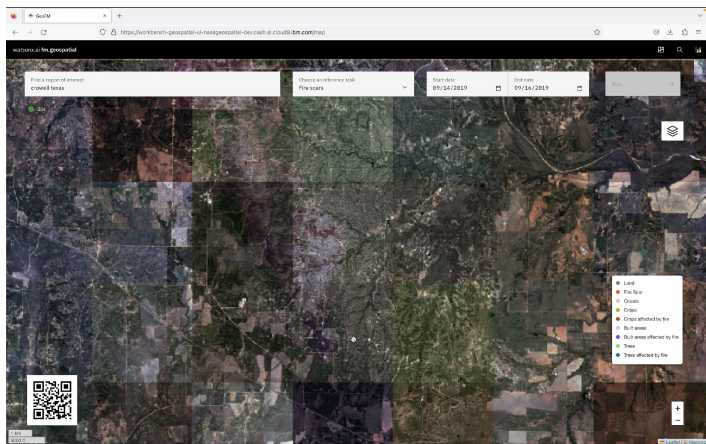


Prithvi (HLS FM) - Fine Tuning Pipeline for Burn Scar



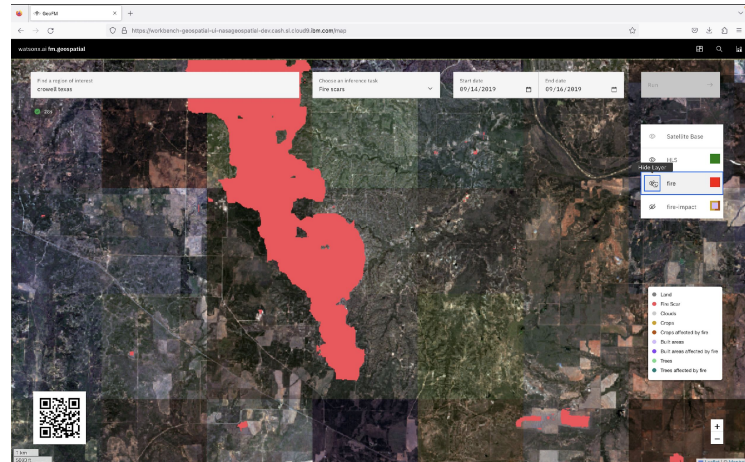


Fire Scar Mapping

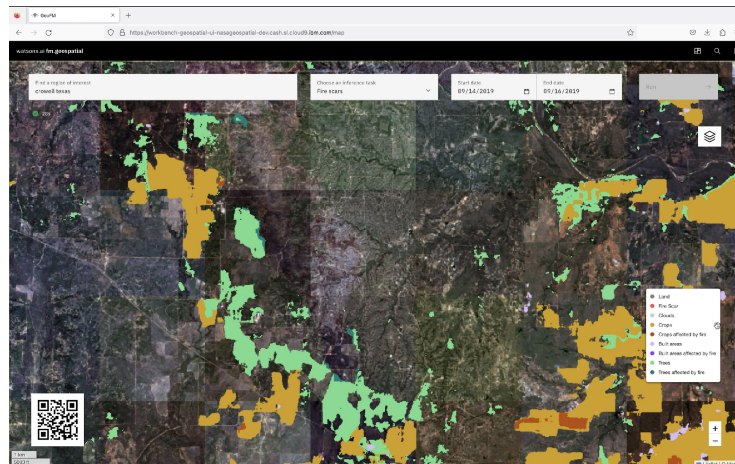


Insights: fire scar
detection

<< Inference >>



Insights: Fire impact

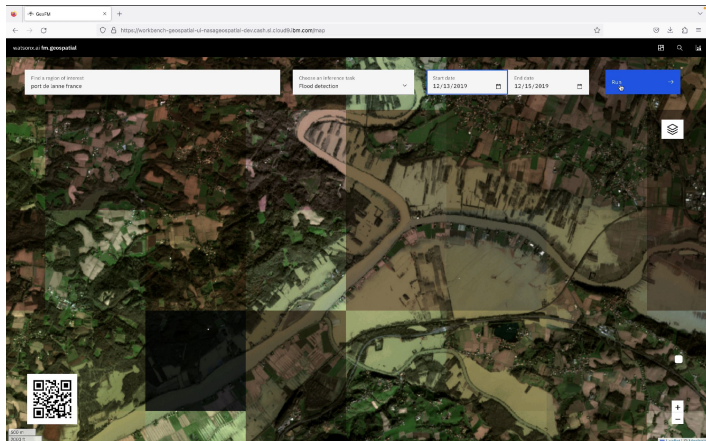


"Prompt": Image(s) (spatial + temporal domains)

	IoU (fire scar class)	F1 (fire scar class)	IoU	F1 score	Accuracy
U-Net (DeepLabV3) [48]	45.96	62.98	93.38	96.11	96.22
ViT-base [19]	51.87	68.31	94.01	96.56	96.56
Swin [46]	56.04	71.83	94.16	96.71	96.50
Prithvi (not pretrained)	59.31	74.46	94.62	97.0	96.80
Prithvi (pretrained)	61.55	76.20	95.04	97.25	97.10



Flood Mapping

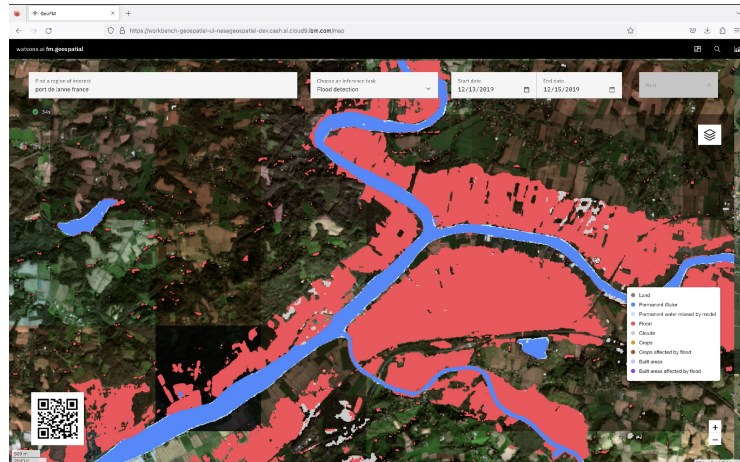


Insights: Flood detection

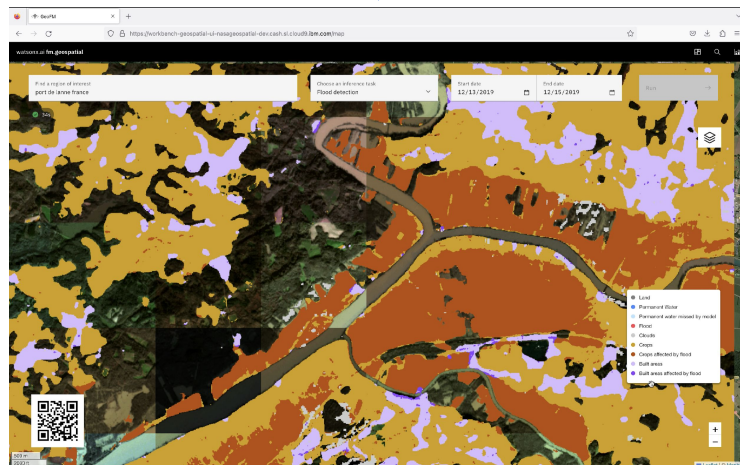
<< Inference >>
(e.g., flood task)

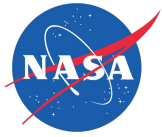
“Prompt”: Image(s) (spatial + temporal domains)

	IoU (water class)	F1 (water class)	IoU	F1 score	Accuracy
Baseline [44]	24.21	—	—	—	—
U-Net-based SOTA [45]	69.12	81.74	93.85	96.65	96.44
ViT-base [19]	66.52	79.89	90.92	94.97	94.97
Swin [46]	74.75	85.55	92.38	95.90	94.73
Prithvi (not pretrained)	79.23	88.41	94.52	97.09	97.07
Prithvi (pretrained)	80.10	88.95	94.78	97.23	97.23



Insights: Flood impact

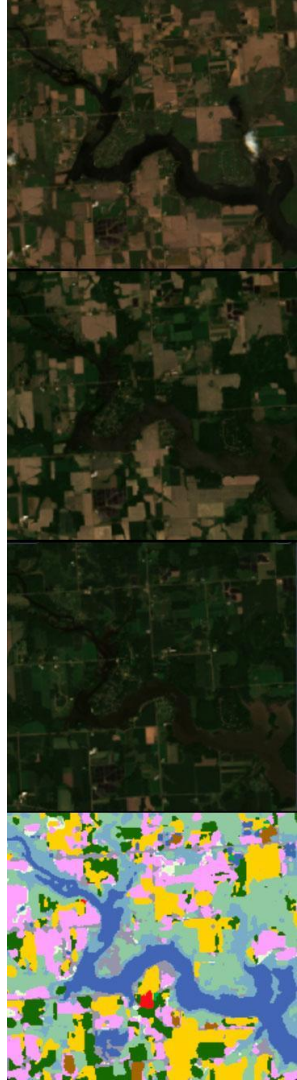




Multi-Temporal Crop/LC Mapping

- 3 cloud-free scenes of HLS predicting USDA CDL
- Overall comparable performance between Prithvi and U-Net baseline
- Prithvi has better performance with fewer samples for fine-tuning

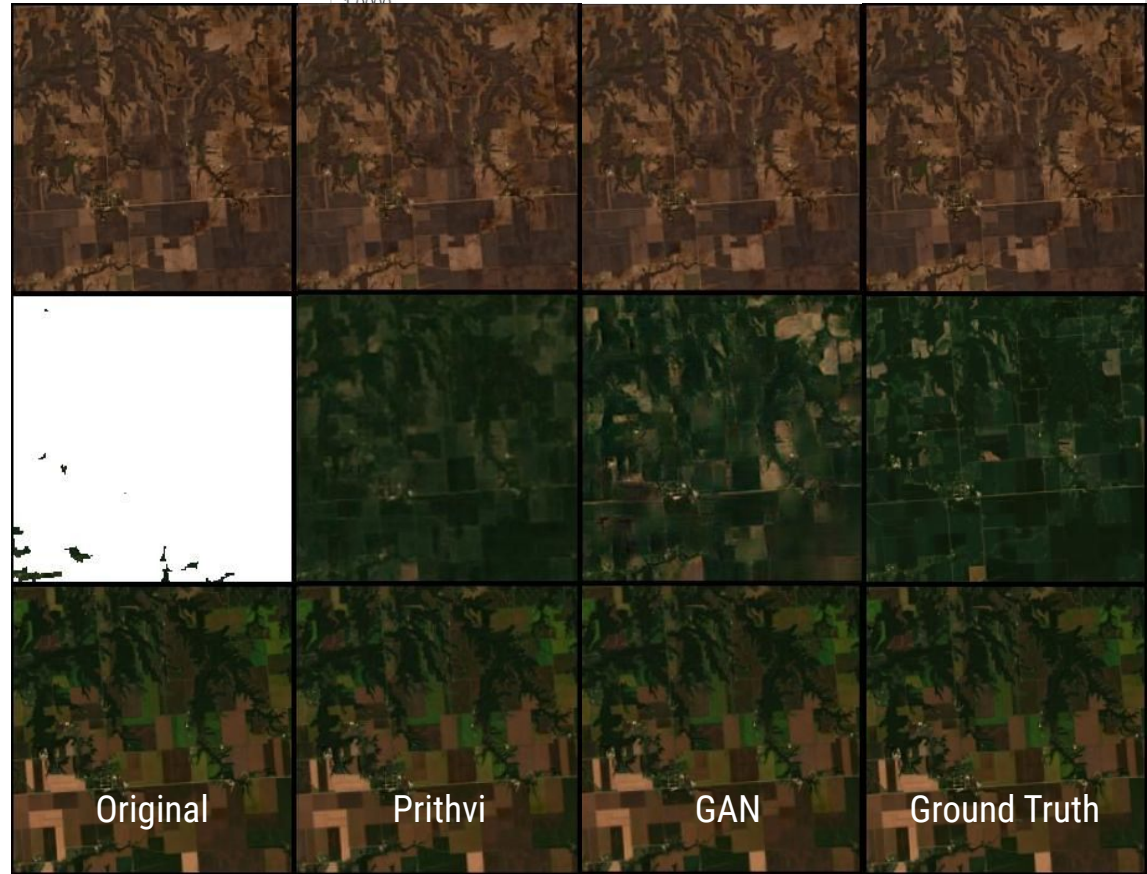
Metric	Prithvi	U-Net
aAcc	0.6064	0.6305
mIoU	0.4269	0.4208
mAcc	0.6406	0.6191





Gap Filling of Cloudy Pixels

- Filling gaps in the middle scene of a 3-scene time series.
- Scenes are selected within a 6-month acquisition window
- Prithvi performs better across all sample sizes and converges faster.



Attribution: Dr. Alemohammad
Clark University



Weather/Climate FM: Clifford Fourier Neural Operator

Dataset Trained on Merra2 dataset taking U500, V500 and Z500 to see if model is able to learn dynamics and solve pde like shallow water equation.

Trained over 7 months, predicted over 1 month

Implementation

1. Defined 2 basis vector e_1, e_2 and 1 multi-vector $e_1 e_2$
2. Take input multivector and create a dual function given by:

$$a = a_0 + a_1 e_1 + a_2 e_2 + a_{12} e_1 e_2$$

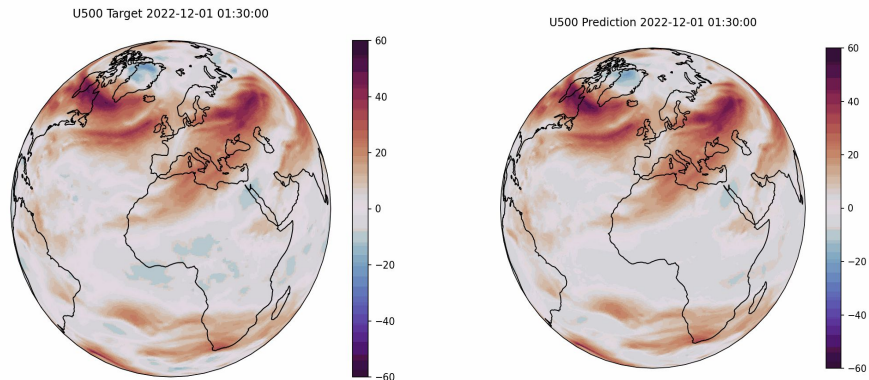
$$(a_0 + a_{12} i_2) + e_1 (a_1 + a_2 i_2)$$

Spinor Vector

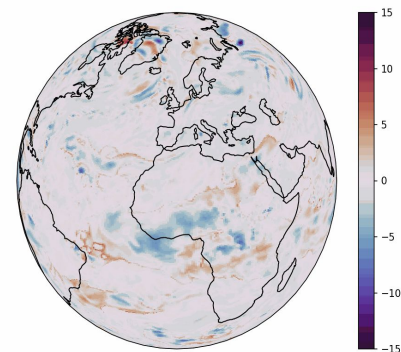
3. Perform Real Fourier transform on each part and revert back and concatenate both.

Key Takeaway

1. Model is able to learn the dynamics related to geometry
2. Can be used to train the model to make it resolution invariant by mapping $u(x) \rightarrow v(x)$
3. Model computation may grow as we will need to perform FFT on vector and spinor part
4. Implementation with transformer can help in efficiently mixing tokens



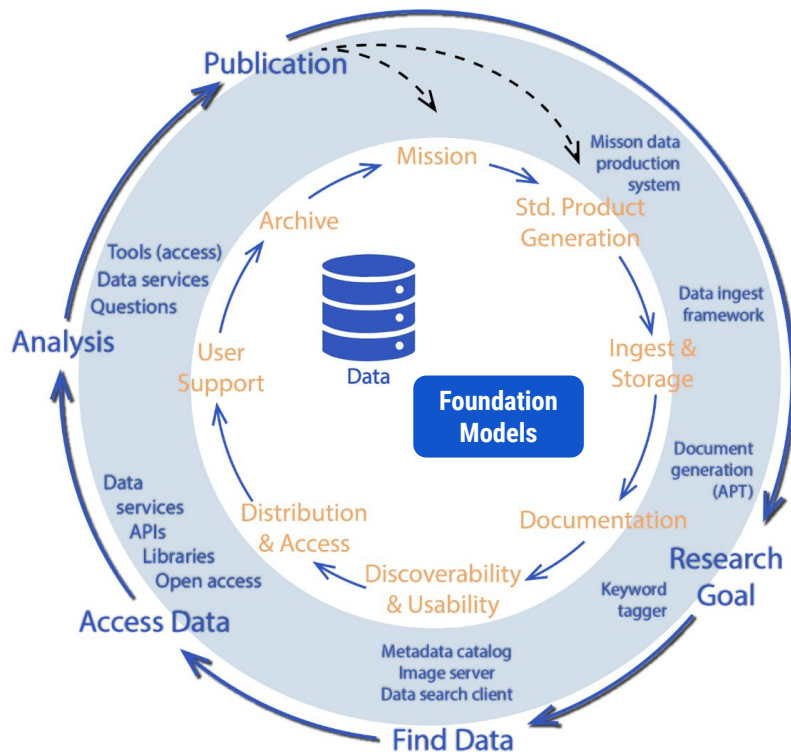
U500 Diff (Target - Prediction) 2022-12-01 01:30:00



Attribution: Dr. Roy NASA/IMPACT UAH



Building Blocks: Future State



“most significant innovations end up generating second- or third-order effects that are difficult to predict in advance...”

– Steven Johnson

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



FM Adaptation: Shift in the way we will use AI Models

Foundation Model
(Pretrained Model)

1. Linear Probing - retrain only a few top layers in the pretrained model
2. Fine Tuning - retrain the entire network, expensive
3. Prefix Tuning - add additional heads to the pretrained model

AI/ML
Expertise

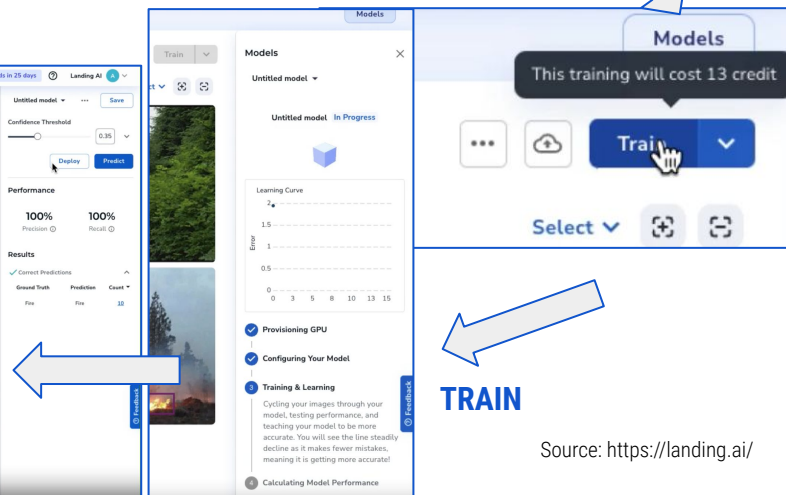
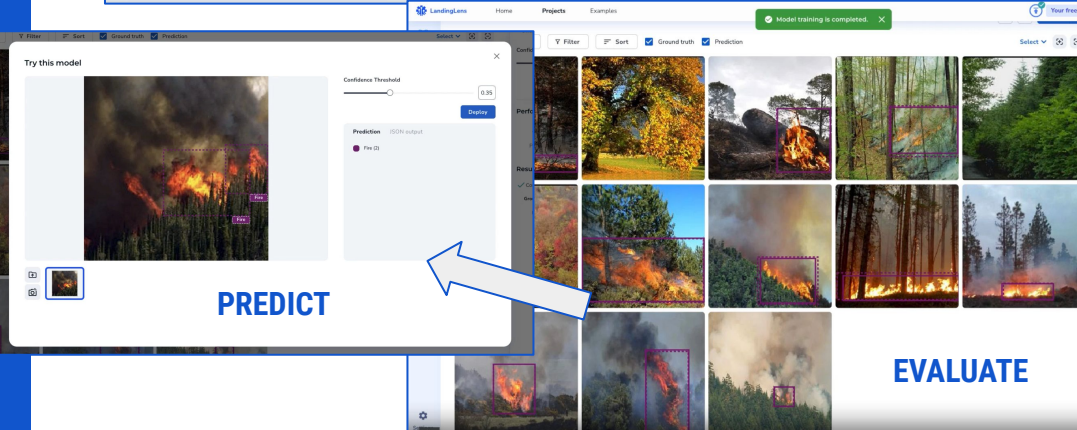
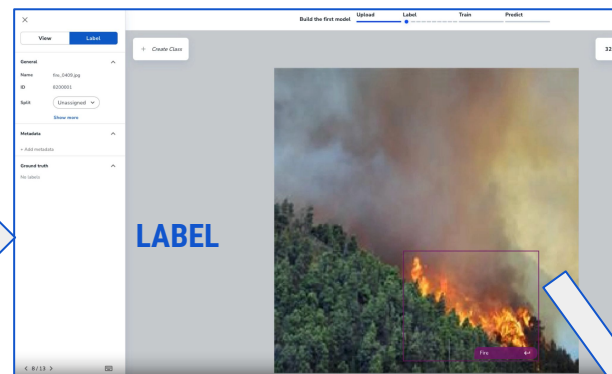
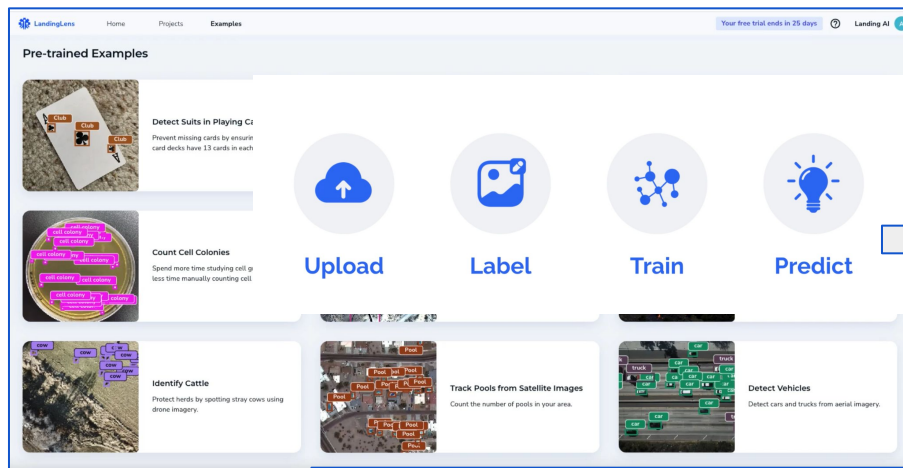
-
4. Prompt Tuning - Zero shot learning
 5. In Context Learning
 - Few shot learning, Chain of Thought

Domain
Expertise

*Iterative, human in the loop and requires
informative set of attributes for the task and well
crafted prompts*



What Does the Future Look Like? (1)



TRAIN

Source: <https://landing.ai/>

What Does the Future Look Like? (2)

	<i>Land cover classes present in the image</i> discontinuous urban fabric, inland waters, forests, water courses, urban fabric, agricultural areas, water bodies, pastures, forest and seminatural areas, broad-leaved forest, artificial areas
	<i>Land cover question</i> Which L2 land cover classes are in the scene? Forests, inland waters, pastures and urban fabric
	<i>Land cover classes present in the image</i> discontinuous urban fabric, open spaces with little or no vegetation, urban fabric, transitional woodland/shrub, agricultural areas, heterogeneous agricultural areas, scrub and/or herbaceous vegetation associations, burnt areas, land principally occupied by agriculture, with significant areas of natural vegetation, forest and seminatural areas, artificial areas
	<i>Yes/No question</i> Are some urban fabric present? Yes

Chappuis et al, 2022, Prompt-RSVQA: Prompting visual context to a language model for Remote Sensing Visual Question Answering, Computer Vision Foundation

Remote Sensing Visual Question Answering

Posted on 14.06.2022

Open Data

Open Science

The interaction on remote sensing images in natural language is a new task for the Earth observation community. Prof Devis Tuia's team share openly a training dataset to explore this domain.

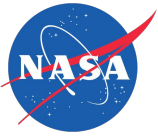
A training dataset designed to for remote sensing visual question answering, hosted on its dedicated website with accompanying scripts, unlicensed.

Why Open?

RSVQA is a new task for the Earth observation community. Prof. Devis Tuia and his team wanted to explore something new (the interaction on remote sensing images in natural language) and they felt that other people would like to join them and develop new ways to tackle this new problem. They also spent months designing a dataset to get started, which they think could be a barrier for many people to get started and focus here rather than on existing super-already-studied other datasets. As such, making the dataset openly available seemed a natural way to enable new progresses.

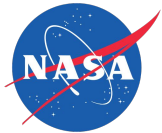
Who benefits?

It is a dataset for anyone in the remote sensing community, so scientists working with satellites and developing algorithms to extract information from them, but also, more at large, the whole computer science and computer vision communities. The original VQA task was first proposed in computer vision to help visually



"Discovery consists of seeing what everybody has seen and thinking what nobody has thought."

– Nobel Prize winner Albert von Szent-Györgyi



FM for Science - Challenges

Pretraining FM is not a trivial task.

Broader Goals

- Understand the resources/infrastructure needed to build these models
 - Inform the teams responsible for infrastructure so that they can develop their roadmaps to align
- Develop a “play book” to build FM for science
 - Systematic step-by-step approach from data set selection to release
- Develop community evaluation suite and benchmarks
- Foster collaborations in a sustainable manner
 - Requires large teams with diverse expertise



Attributions

NASA MSFC/IMPACT

Sujit Roy, Ankur Kumar, Chris Phillips, Iksha Gurung, Kumar,
Rahul Ramachandran, **Manil Maskey**

IBM Research

Johannes Jakubik, Linsong Chu, Paolo Fraccaro, Ranjini, Kamal Das, Daiki Kimura, Naomi Simumba,
Daniela Szwarcman, Michal Muszynski, Carlos Gomes, **Kommy Weldemariam**, Bianca Zadrozny, **Raghu
Ganti**,...

Clark University

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

NASA HQ

Tsengdar Lee, **Kevin Murphy**

NASA GSFC

Dan Duffy, Mike Little



Contact or Connect with Me



rahul.ramachandran@nasa.gov





Backup

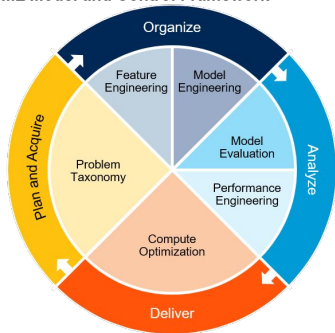
Organization to support different disciplines

Domain experts + data science team
solve problems, scale solutions to production

Data Science team provides competency in

- Big data
- Management, stewardship, system
- Analytics, machine learning, and AI
- Leveraging HPC and cloud platforms
- Development and MLOps

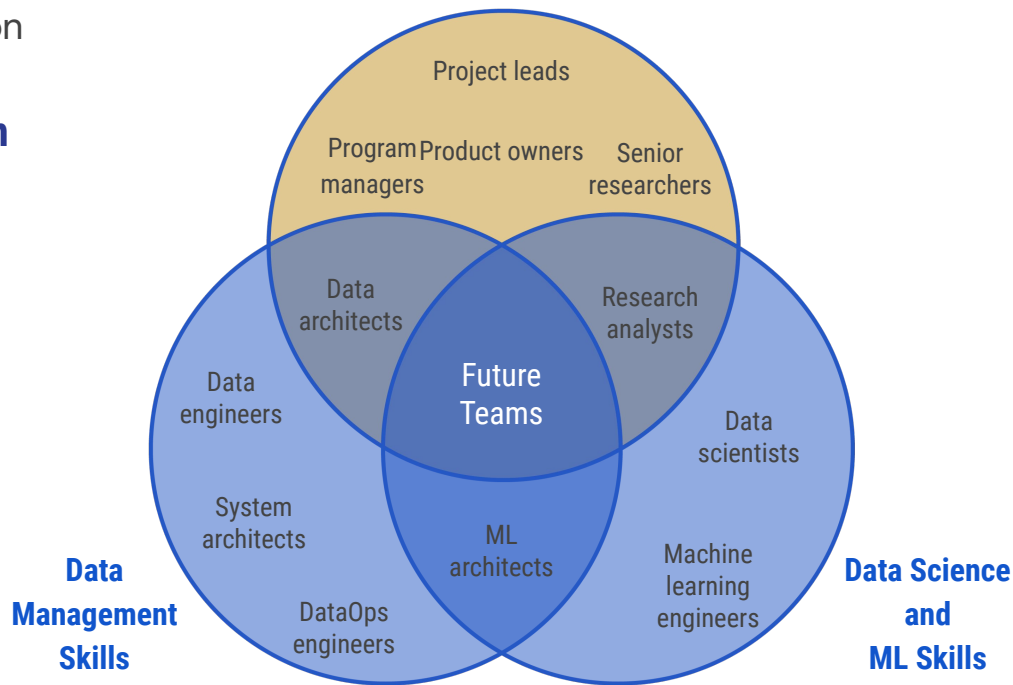
Continuous ML Model and Control Framework



ID: 365935

© 2018 Gartner, Inc.

Earth/Helio/Planetary/Astro/BPS and Other Domains



Data Science teams can support multiple projects from different disciplines