

# AI-Driven Accelerated Discovery for NASA's Science Mission Directorate

**Kaylin Bugbee<sup>1</sup>, Rahul Ramachandran<sup>1</sup>, Juan Bernabe  
Moreno<sup>2</sup>**

\*Presented by Muthukumaran Ramasubramanian<sup>3</sup>

ESA-NASA International Workshop on AI Foundation Model for EO  
May 6, 2025

[1] Marshall Space Flight Center; [2] IBM Research; [3] University of Alabama in Huntsville



# Science at NASA

From the Smallest Cells...

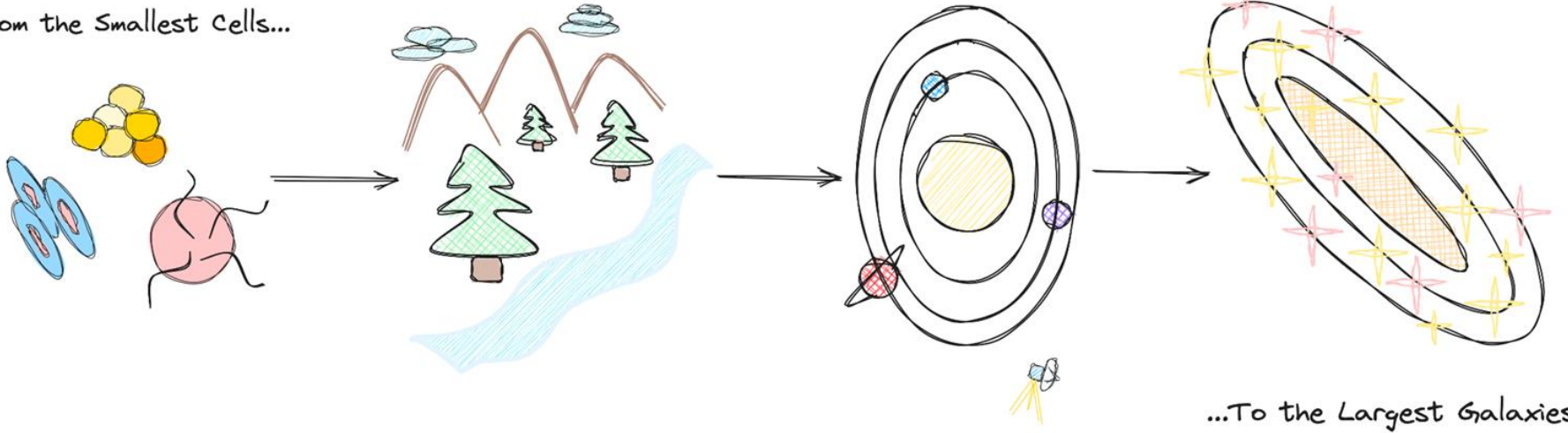


Image Credit: SDE team

# SMD by the Numbers

 **10 M**  
FY23 unique users

 **140 PB**  
FY23 Total volume

 **50 PB/year**  
Ingest Rate

 **>500 PB**  
Projected FY29  
archive volume

 **1.5 B**  
Total files

 **Efficient data management and computing  
are essential for NASA's mission.**



# Motivation and Foundational Principles



# Motivation to Use LLMs for Science

- Large Language Models (LLMs) increase research **efficiency**, saving scientists significant time.
  - **Streamline workflows**: data discovery, access, literature review, and coding.
  - **Reduce "data-wrangling"** time from 80% to significantly less, accelerating scientific discovery.
- LLMs **enhance data system value**: improving data visibility, access, usability, and value.
  - Facilitate new discovery pathways and applications, increasing data system use.
  - Enable easier and more **contextual information retrieval**, aligning with user expectations.

LLMs offer **transformative benefits** for NASA science, streamlining both **science and data system operations**, thereby accelerating the pace of discovery and enhancing data utility.



# Foundational Principle: AI Ethics for Science

- Prioritize ***open models, data, workflows, and code*** for transparency and collaboration.
- Devote significant effort to ***formulating clear, concise, and correct questions***.
  - Embrace Albert Einstein's approach: Spend most of the time defining the problem accurately to enable rapid solutions.
  - Ensure questions are precisely worded to retrieve relevant answers.
- Apply Carl Sagan's "***baloney detection kit***" for critical analysis and fact verification.
  - Adhere to the principle of "***trust but verify***" to maintain rigorous scrutiny.
  - ***Avoid unquestioning acceptance*** of AI-generated outputs.
- ***Approach AI as a collaborative partner***, not a sole decision-maker.
  - Employ the ***co-pilot analogy***, emphasizing shared responsibility in AI interaction.
  - Remain accountable for the integrity and accuracy of AI-assisted outcomes.

Ethical AI use demands ***openness, critical questioning, collaborative partnership, and vigilant verification*** to ensure responsible and effective outcomes



# Large Language Models for Science



# Large Language Models for Science

- At NASA's Science Mission Directorate (SMD), we aim to understand the appropriate use of LLMs within the scientific enterprise by both researchers and developers of scientific applications.
- This involves investigating:
  - Whether NASA should build its own LLM for science;
  - Determining the type of model needed (encoder vs. decoder/generative);
  - Deciding between building from scratch or fine-tuning an existing open model like Meta's Llama;
  - Curating pre-training materials such as journal papers and technical reports.

Ramachandran, R., & Bugbee, K. (2025). Balancing practical uses and ethical concerns: The role of large language models in scientific research. *Perspectives of Earth and Space Scientists*, 6, e2024CN000258. <https://doi.org/10.1029/2024CN000258>





## Large Language Models Study

- Leveraged our partnership with IBM Research to develop a LLM framework for NASA Science Mission Directorate (SMD).
- Curated sources from the five SMD topical areas for model pre-training.
  - Collaboration with subject matter experts (SMEs) from across the NASA science topical areas was essential
  - Resulted in a diverse data set totaling 66.2 billion tokens for model pre-training
- Developed an encoder-only transformer model, named INDUS, tailored for SMD applications.
  - Useful for various tasks such as named entity recognition, extractive question answering, text classification, semantic equivalence for document retrieval, and knowledge extraction for relationships
  - Also developed a distilled version of the model, five times smaller (30M parameters) than the original 125M model, with only marginally reduced performance.



# Large Language Models Study

- Teams from the different science areas created a targeted question and answer (Q/A) suite to support the encoder model's training, focusing on precision in scientific inquiry and retrieval.
- Also created a sentence transformer:
  - Assists in information retrieval by efficiently understanding text semantics.
  - Generates embeddings for queries and sentences, enhancing information retrieval by converting them into high-dimensional vectors that capture deep semantic meanings.

## **Learn more about the technical details of INDUS here:**

Bhattacharjee, B., Trivedi, A., Muraoka, M., Ramasubramanian, M., Udagawa, T., Gurung, I., et al. (2024). INDUS: Effective and efficient language models for scientific applications. arXiv, 98–112. <https://doi.org/10.18653/v1/2024.emnlp-industry.9>

## **Find INDUS on Hugging Face here:**

Maraoka, M., Bhattacharjee, B., Ramasubramanian, M., Gurung, I., Ramachandran, R., Maskey, M., et al. (2023). Nasa-impact/nasa-smd-ibm-v0.1 (Version 0.1) [Software]. Hugging Face. <https://doi.org/10.57967/hf/1429>



## Key Takeaways from the Study

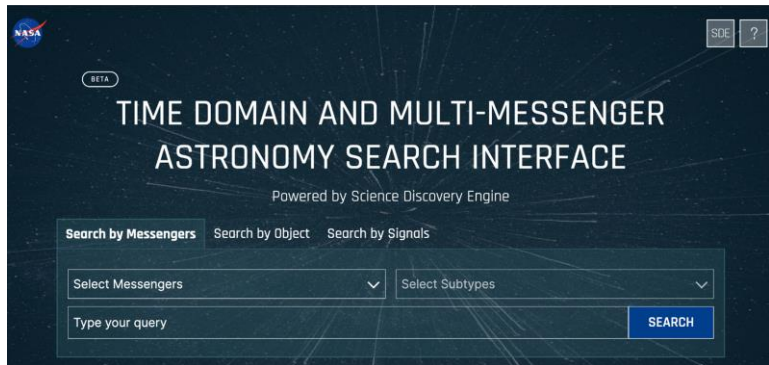
- **Collaboration and partnerships were critical to this study.**
  - IBM Research for LLM expertise and needed compute resources.
  - SMD subject matter experts for domain expertise.
- **A single type of model is not the right answer for all solutions.**
  - A suite of models, including both encoders and decoders, is essential for supporting diverse downstream applications.
- **Retrieval augmented generation (RAG) is a cost-effective, low-risk strategy for more responsibly using LLMs for science.**
  - RAG involves using domain-specific encoder models, such as INDUS, within a RAG framework for information retrieval, which can be integrated with existing off-the-shelf generative models like Llama or GPT.
  - RAG combines document retrieval (via an encoder) and generative modeling to enhance answer accuracy and relevance by providing contextual grounding.



# LLM Applications: Stewardship Workflows

Image Credit: SDE Team

- Automated content curation.
  - Using LLMs to automate document classification for curated search and discovery experiences.
  - TDAMM, division classifier, GCMD keyword recommender, division classifier.
- Example: Time Domain and Multi-Messenger Astronomy Search Interface.
  - Relatively new field in astrophysics - observations cover a wide range of time-varying and types of phenomena/messengers where messengers can be cosmic rays, electromagnetic radiation, gravitational waves and neutrinos.
  - Relevant data and information is dispersed across a number of archives and repositories—there is a need to make search and discovery easier for the TDAMM community.
  - Used astroBERT to build a classifier to streamline content curation.





# LLM Applications: Stewardship Workflows

- Example: Earth Science Keyword Recommender.
  - Global Change Master Directory (GCMD) Keywords are a set of controlled vocabularies for the Earth sciences. Used to label metadata and other documentation.
  - GCMD Keyword recommender helps human curators accurately tag their data with relevant GCMD keywords.
  - We enhanced the keyword recommender by fine-tuning INDUS for the classification task.
  - New fine-tuned model outperforms other models including RoBERTa and the existing model.

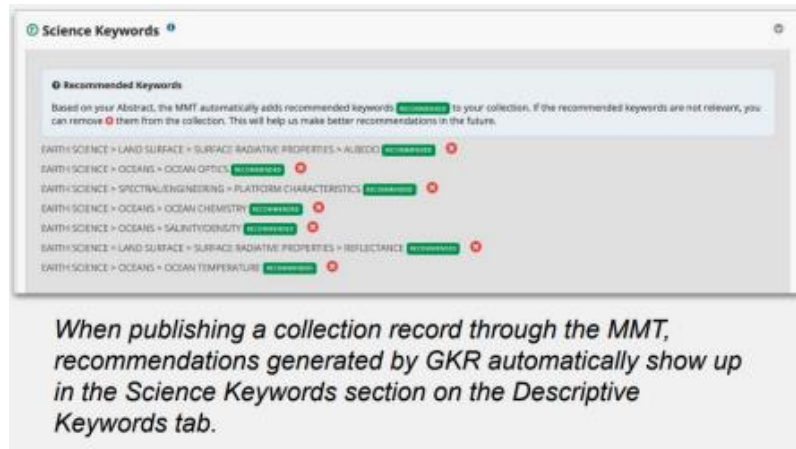


Image Credit: Bhattacharjee et al.  
<https://ntrs.nasa.gov/api/citations/20240015188/downloads/AGU%202024%20INDUS%20Applications.pdf>



# LLM Applications: Enhanced Search

- Integrated the sentence transformer and passage re-ranker adapted from the base INDUS encoder model with a general model ChatGPT within a RAG framework into a prototype Science Discovery Engine (SDE) environment.
- Qualitative and quantitative assessments indicate improved relevancy of search results.

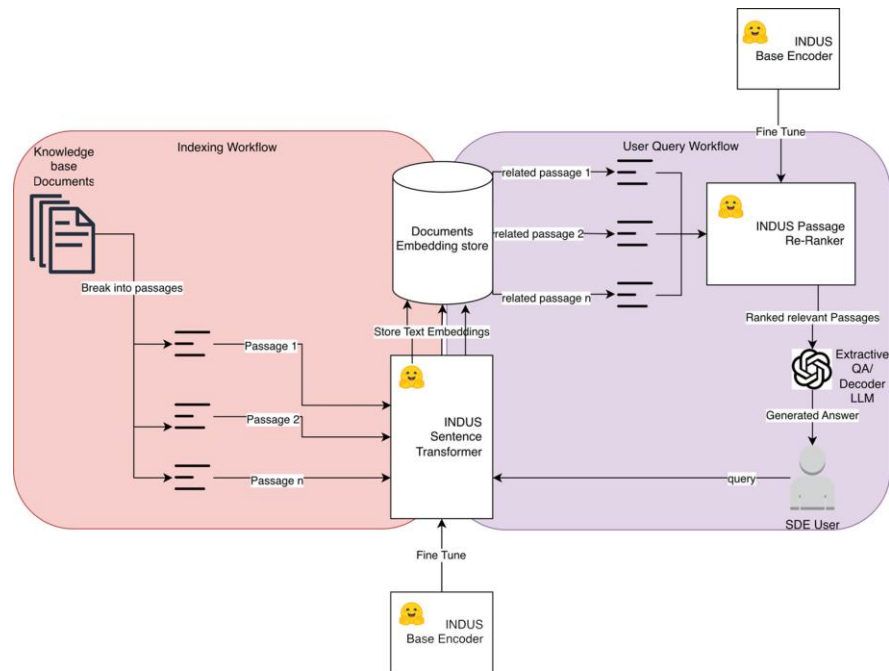


Image Credit: Ramachandran, R., & Bugbee, K. (2025). Balancing practical uses and ethical concerns: The role of large language models in scientific research. *Perspectives of Earth and Space Scientists*, 6, e2024CN000258. <https://doi.org/10.1029/2024CN000258>



# LLM Applications: Enhanced Search

BETA All

FILTERS

Science Knowledge Sources +

Science Data Repositories +

Platforms +

Instruments +

Missions +

Neural Search Suggestions

3 answers found in 2 documents

MAST - Missions and Data - ASTRO

Jul 17, 2024 - the Ultraviolet Imaging Telescope (UIT), the Hopkins Ultraviolet Telescope (HUT) and the Wisconsin Ultraviolet Photo-Polarimeter Experiment (WUPPE). The Astro Observatory was designed to use many of the spacelab components and flew on two different shuttle flights. The first was aboard the shuttle Columbia which also held the X-ray experiment Broad Band X-Ray Telescope (BBXRT). The second flight was aboard the shuttle Endeavour. Active From ASTRO-1: December 2 - 11, 1990 ASTRO-2: March 2 - 18, 1995 Capabilities Imaging Spectroscopy Polarimetry On this Page On this Page On this Page Mission Overview Instruments ASTRO-1 ASTRO-2 Instruments HUT

MAST - Missions and Data - ASTRO

Jul 17, 2024 - Data Attributions Mission Acknowledgements Mission Publications Mailing List News Archived Synthetic Data Prototype Demo Virtual Observatory Mission Data Search MUG About MAST New Mission Partnerships with MAST Astro Art Internship 2024 MAST Summer Webinar ASTRO Breadcrumbs Navigation Home Missions and Data Mission Overview Expand Image The ASTRO Observatory had three primary instruments: the Ultraviolet Imaging Telescope (UIT), the Hopkins Ultraviolet Telescope (HUT) and the Wisconsin Ultraviolet Photo-Polarimeter Experiment (WUPPE). The Astro Observatory was designed to use many of the spacelab components and flew on two different shuttle flights.

MAST - Missions and Data - ORFEUS

Jul 17, 2024 - The three instruments on the ORFEUS were designed to provide astronomical ultraviolet spectroscopic observations over the wavelength range from 40 to 140 nanometers. The three instruments were: Tübingen Ultraviolet Echelle Spectrometer (TUES); (PI) Prof. Michael Grewing; University of Tübingen Berkeley Extreme and Far-UV Spectrometer (BEFS); (PI) Dr. Mark Hurwitz; University of California, Berkeley. This instrument was called the Extreme Ultraviolet (EUV) Spectrometer in the ORFEUS-SPAS II Mission Research Announcement. It was later renamed. Interstellar Medium Absorption Profile Spectrograph (IMAPS); (PI) Dr. Edward Jenkins; Princeton University The largest science instrument onboard was a 1-meter telescope.

MAST - Missions and Data - ASTRO

Astrophysics > MAST: Missions and Data

Jul 17, 2024 - the Ultraviolet Imaging Telescope (UIT), the Hopkins Ultraviolet Telescope (HUT) and the Wisconsin Ultraviolet Photo-Polarimeter Experiment (WUPPE). The Astro Observatory was designed to use many of the spacelab components and flew on two different shuttle flights. The first was aboard the shuttle Columbia which also held the X-ray experiment Broad Band X-Ray Telescope (BBXRT). The second flight was aboard the shuttle Endeavour. Active From ASTRO-1: December 2-11, 1990 ASTRO-2: March 2-18, 1995 Capabilities Imaging Spectroscopy Polarimetry On this Page On this Page On this Page Mission Overview Instruments ASTRO-1 ASTRO-2 Instruments HUT

MAST - Missions and Data - ORFEUS

Astrophysics > MAST: Missions and Data

Jul 17, 2024 - The three instruments on the ORFEUS were designed to provide astronomical ultraviolet spectroscopic observations over the wavelength range from 40 to 140 nanometers. The three instruments were: Tübingen Ultraviolet Echelle Spectrometer (TUES); (PI) Prof. Michael Grewing; University of Tübingen Berkeley Extreme and Far-UV Spectrometer (BEFS); (PI) Dr. Mark Hurwitz; University of California, Berkeley. This instrument was called the Extreme Ultraviolet (EUV) Spectrometer in the ORFEUS-SPAS II Mission Research Announcement. It was later renamed. Interstellar Medium Absorption Profile Spectrograph (IMAPS); (PI) Dr. Edward Jenkins; Princeton University The largest science instrument onboard was a 1-meter telescope.

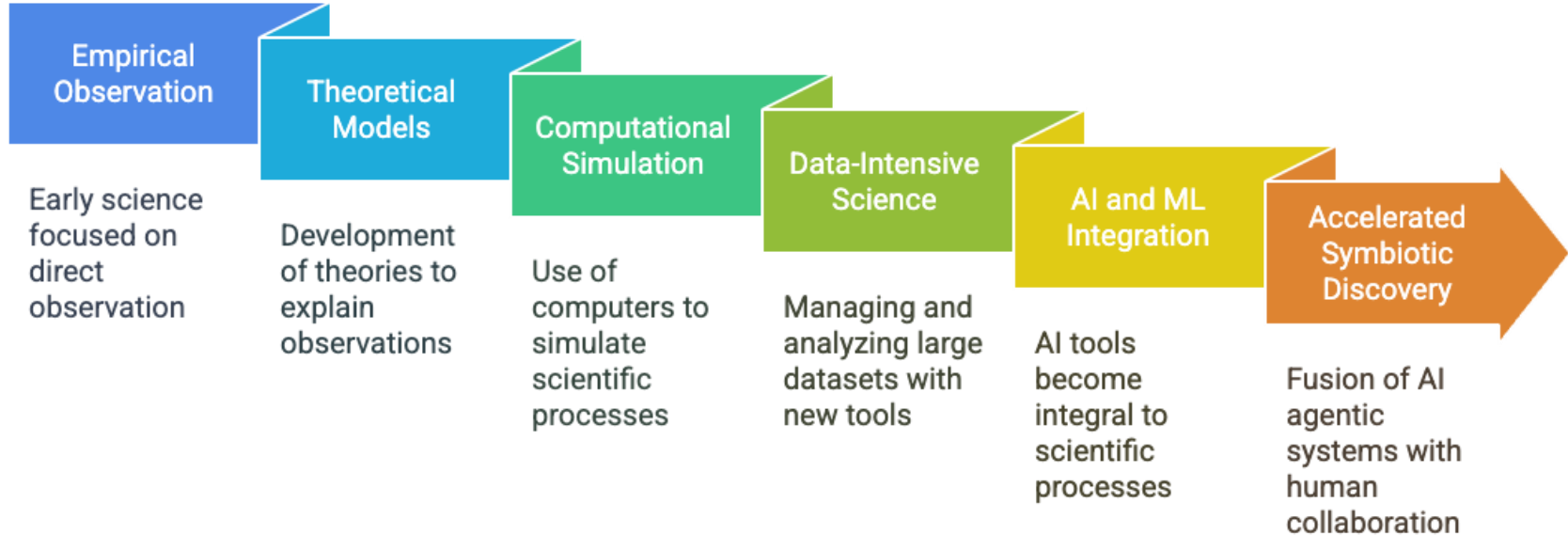
Image Credit: SDE Team



# Future Direction: Accelerated Discovery



## Evolution of Scientific Paradigms



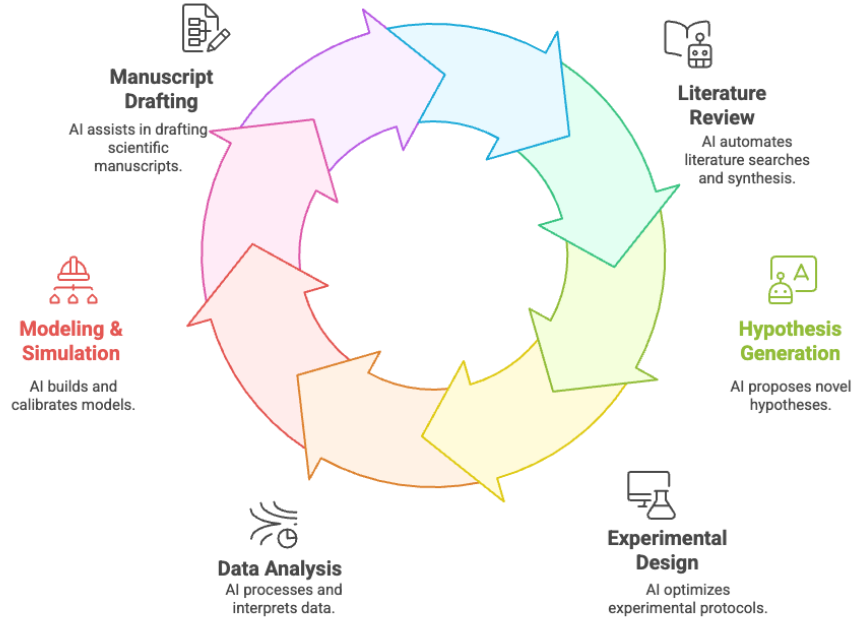
Made with  Napkin

**Repositions AI from a tool to collaborator**

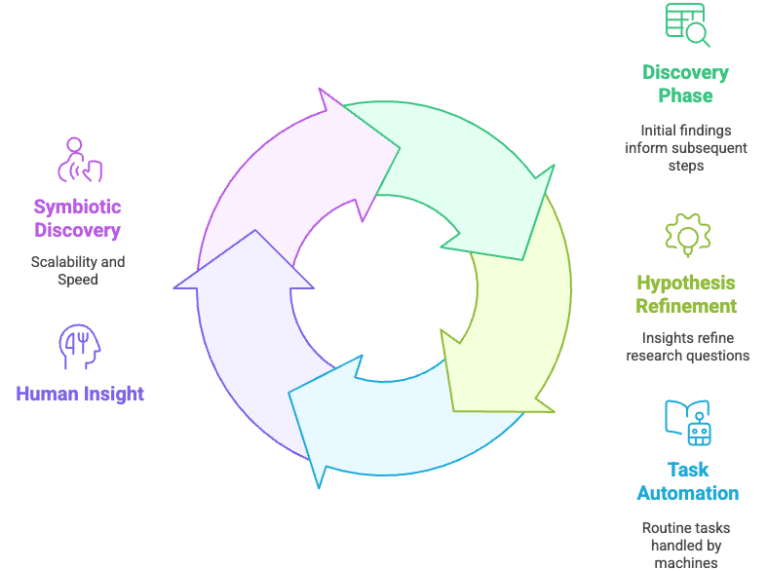
*Inspired by and building on Jim Gray's Science Paradigms*

Image Credit: Rahul Ramachandran

## Symbiotic Scientific Discovery Cycle



## Closed-Loop Research Cycle



Made with Napkin

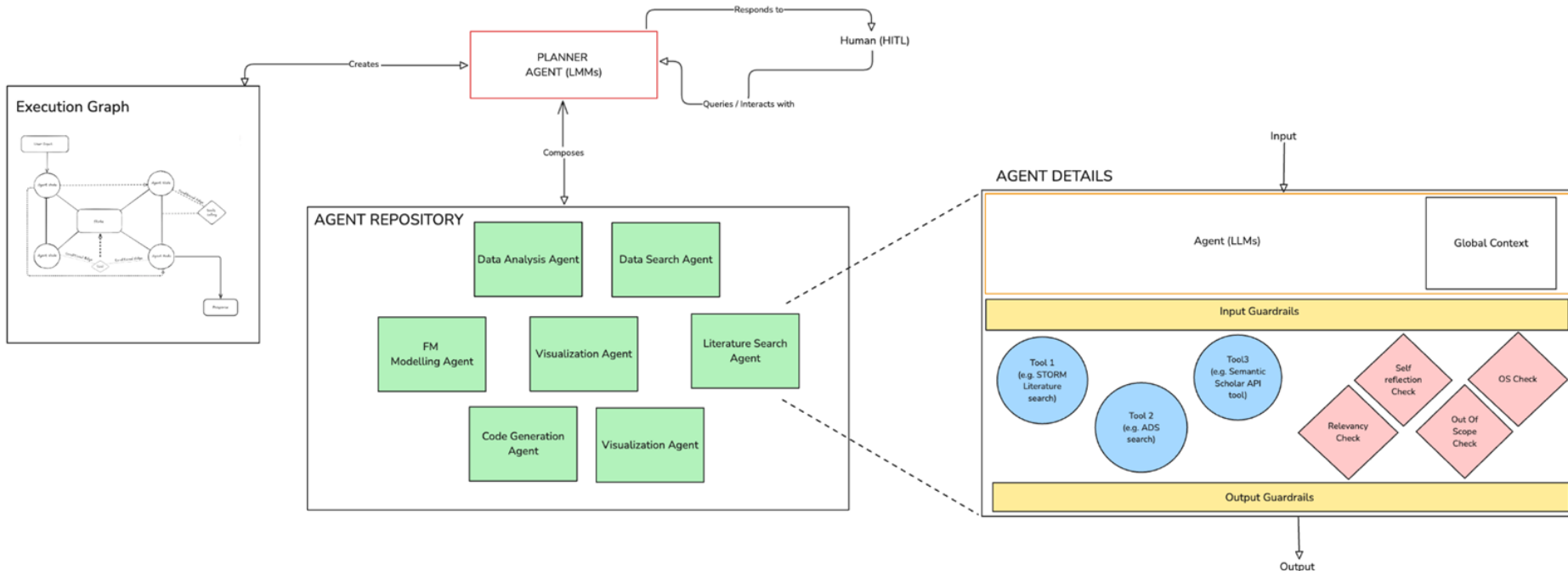
*The scientific research process augmented by AI.*



*AI can help a scientist iterate and refine each step of the research process.*

Made with Napkin

# Conceptual Multi-Agent Architecture





AD

Do literature review on oil palm classification using remote sensing and ML from 2020-2025, identify gaps, suitable datasets, models, and suggest future research avenues.



earthdata-search [Codespaces: silver train]

EXPLORER

▼ EARTHDATA-SEARCH [CODESPACES: SILVER TRAIN]

- > .github
- > bin
- > cdk
- > layers
- ▼ migrations
  - JS 1556745445952\_colormaps.js
  - JS 1559965681107\_users.js
  - JS 1559966487087\_projects.js
  - JS 1559968798898\_retrievals.js
  - JS 1559970480679\_retrieval-collections.js
  - JS 1559971914388\_retrieval-orders.js
  - JS 1562705069972\_rename-order-search-par...
  - JS 1563477325342\_access-configurations.js
  - JS 1563913294480\_shapefiles.js
  - JS 1563990176916\_alter-retrieval-order-num...
  - JS 1564488272729\_add-environment-to-user...
  - JS 1564489093225\_remove-environment-def...
  - JS 1565375842170\_add-order-information-to...
  - JS 1565744336991\_add-default-to-order-stat...
  - JS 1567513762242\_user-tokens.js
  - JS 1567802528030\_add-on-delete-cascades.js
  - JS 1568146028532\_change-users-not-null.js
  - JS 1585836600012\_clear-site-preferences.js
  - JS 1598036919806\_add-parent-shapefile-id-s...

▼ OUTLINE

▼ TIMELINE

[Preview] README.md X

## Earthdata Search

no status codecov 93% Snyk security monitored

### About

Earthdata Search is a web application developed by NASA EOSDIS to enable data discovery, search, comparison, visualization, and access across EOSDIS' Earth Science data holdings. It builds upon several public-facing services provided by EOSDIS, including the Common Metadata Repository (CMR) for data discovery and access, EOSDIS Earthdata Login (EDL) authentication, the Global Imagery Browse Services (GIBS) for visualization, and a number of OPeNDAP services hosted by data providers.

### License

Use Cmd/Ctrl + Shift + P -> View Creation Log to see full logs

- ✓ Finishing up...
- ✗ Detected Node.js 16.20.2
- ✓ Running postCreateCommand...
- ✓ Done

Finished configuring codespace. Press any key to exit.

bash

GitHub ...

Ask Accelerated Discovery Planner

Accelerated Discovery Planner is powered by AI, so mistakes are possible. Review output carefully before use.

Type your query

EXPLORER

▼ EARTHDATA-SEARCH [CODESPACES: SILVER TRAIN]

> .github

> bin

> cdk

> layers

▼ migrations

J\$ 1556745445952\_colormaps.js

J\$ 1559965681107\_users.js

J\$ 1559966487087\_projects.js

J\$ 1559968798898\_retrievals.js

J\$ 1559970480679\_retrieval-collections.js

J\$ 1559971914388\_retrieval-orders.js

J\$ 156270569972\_rename-order-search-par...

J\$ 1563477325342\_access-configurations.js

J\$ 1563913294480\_shapefiles.js

J\$ 1563990176916\_alter-retrieval-order-num...

J\$ 1564488272729\_add-environment-to-user...

J\$ 1564489093225\_remove-environment-def...

J\$ 1565375842170\_add-order-information-to...

J\$ 1565744336991\_add-default-to-order-stat...

J\$ 1567513762242\_user-tokens.js

J\$ 1567802528030\_add-on-delete-cascades.js

J\$ 1568146028532\_change-users-not-null.js

J\$ 1585836600012\_clear-site-preferences.js

J\$ 1598036919806\_add-parent-shapefile-id-s...

> OUTLINE

> TIMELINE

[Preview] README.md ✕

# Earthdata Search

no status 93% monitored

## About

Earthdata Search is a web application developed by NASA EOSDIS to enable data discovery, search, comparison, visualization, and access across EOSDIS' Earth Science data holdings. It builds upon several public-facing services provided by EOSDIS, including the Common Metadata Repository (CMR) for data discovery and access, EOSDIS Earthdata Login (EDL) authentication, the Global Imagery Browse Services (GIBS) for visualization, and a number of OPeNDAP services hosted by data providers.

## License

Use Cmd/Ctrl + Shift + P -> View Creation Log to see full logs

- ✓ Finishing up...
- ✗ Detected Node.js 16.20.2
- ✓ Running postCreateCommand...
- ✓ Done

Finished configuring codespace. Press any key to exit.

PROBLEMS

OUTPUT

DEBUG CONSOLE

TERMINAL

PORTS

bash

GitHub ...

CHAT

You 15 January 2024, 02:22 AM

Can you build a jupyter notebook to simulate earthdata search

ADP 15 January 2024, 02:22 AM

Sure! Below is a basic outline of a Jupyter Notebook to simulate an Earthdata Search experience. This notebook will include:

1. A simple search interface using input widgets.
2. A simulated dataset to search from.
3. Basic filters (e.g., date range, region).
4. A results display with mock data.

✓ Earthdata Search Simulation Notebook ( earthdata\_search\_sim.ipynb )

```
python

# Earthdata Search Simulation - Jupyter Notebook
# Author: [Your Name]
# Description: Simulates Earthdata-like search on mock data

import pandas as pd
import numpy as np
from IPython.display import display, Markdown
import ipywidgets as widgets
from datetime import datetime

# -----
# 1. Simulated Earthdata Catalog
```

Type your query

+

→



## Conclusions

- NASA is positioned to enable a transformative shift in scientific discovery.
- Accelerated Symbiotic Discovery will
  - Integrate agentic AI, multimodal foundation models, scalable cloud and high-performance computing (HPC) infrastructure, and existing knowledge-representation frameworks
  - Accelerate scientific discoveries from the smallest of cells to the largest of galaxies.



## Call to Action

- We envision this agentic infrastructure to be open, transparent, and reproducible
- **We welcome collaboration, thoughts and feedback on this effort**



# THANK YOU!

## Questions?

Email:

[kaylin.m.bugbee@nasa.gov](mailto:kaylin.m.bugbee@nasa.gov)

[rahul.ramachandran@nasa.gov](mailto:rahul.ramachandran@nasa.gov)