Accelerating Scientific Discoveries: NASA's Use of Large Language Models for Unlocking the Value of Science Data and Information

Kaylin Bugbee, Rahul Ramachandran

NASA Marshall Space Flight Center

March 14, 2024

“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. Assessment of the Usefulness and Availability of NASA's Earth and Space Science Mission Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/10363.

Science at NASA

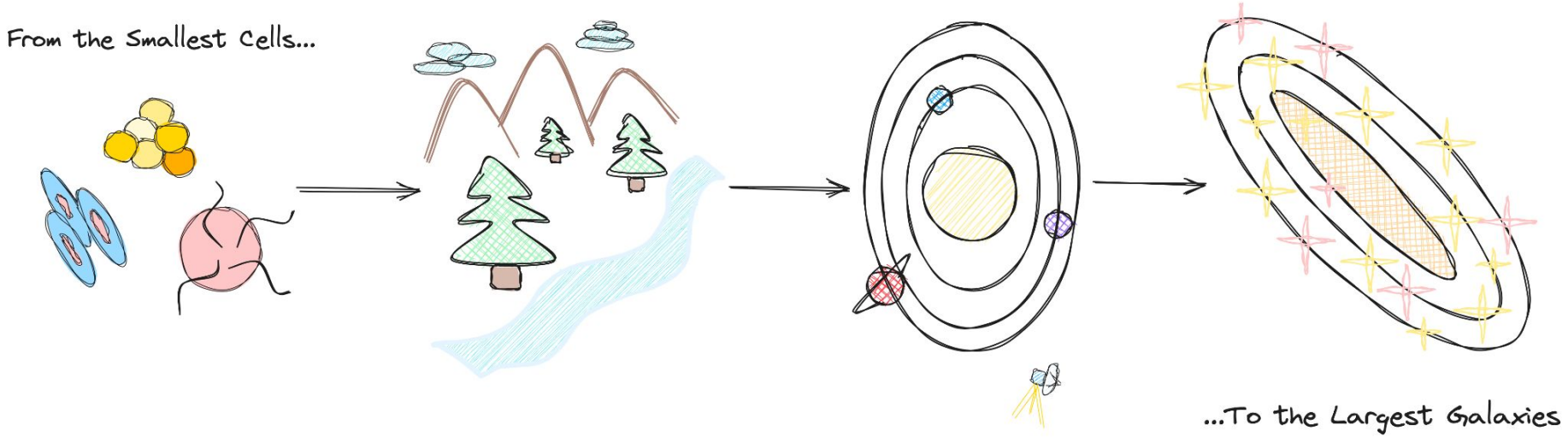
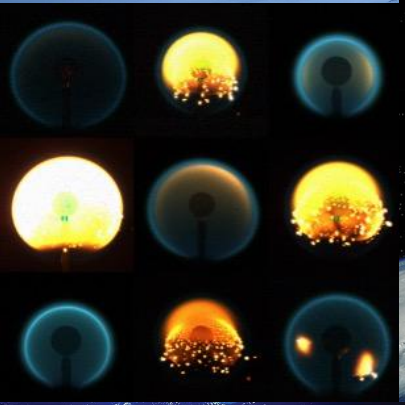
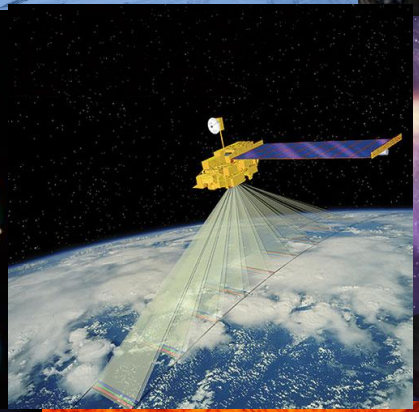
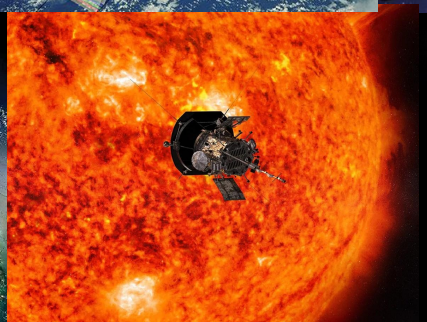


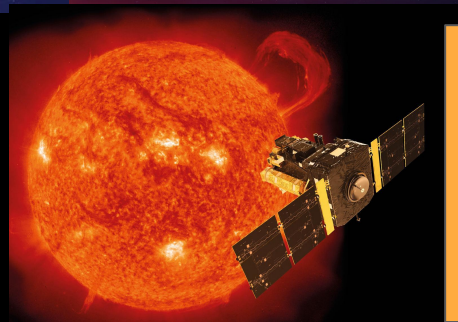
Image Credit: SDE team









Images Credit: NASA

SMD by the Numbers

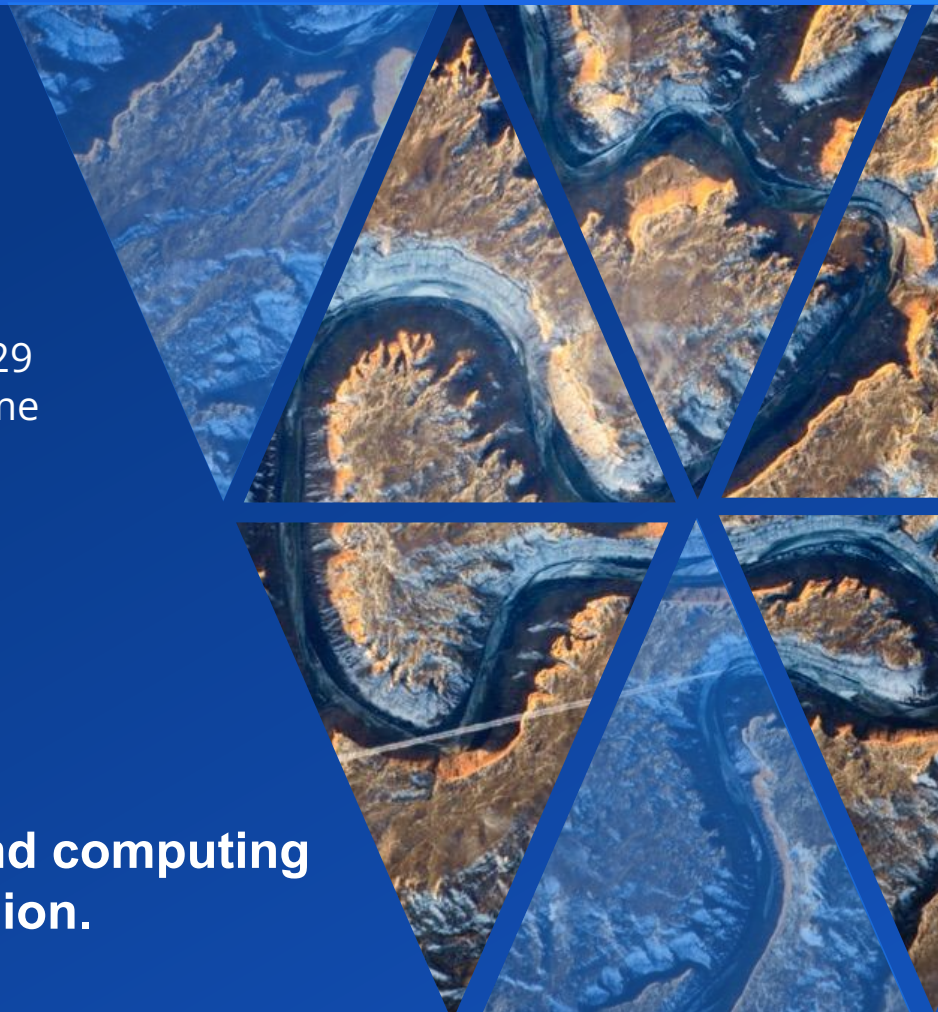
10 M
FY23 unique users

>500 PB
Projected FY29 archive volume

140 PB
FY23 Total volume

1.5 B
Total files

50 PB/year
Ingest Rate



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.

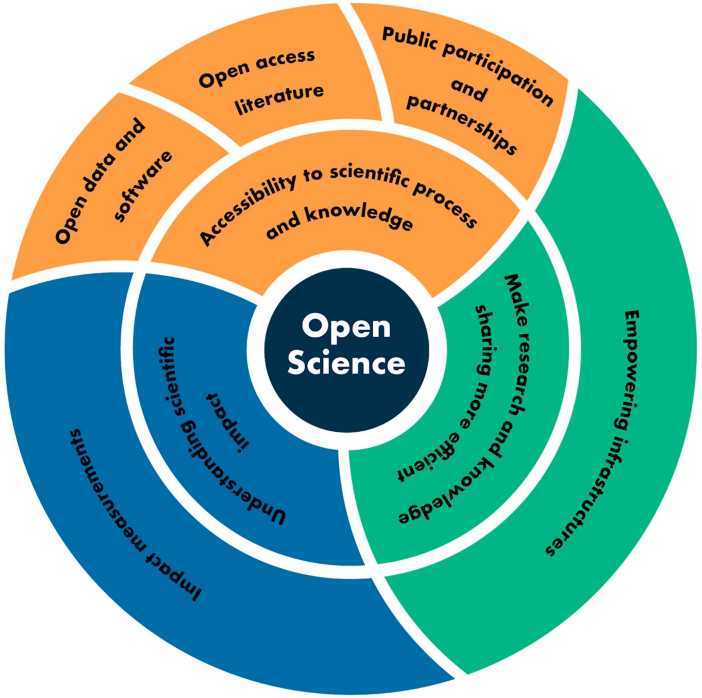
LLMs offer ***transformative benefits*** for NASA science, streamlining both ***science and data system operations***, thereby accelerating the pace of discovery and enhancing data utility.

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

Foundational Principle: Open Science



Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From open data to open science. *Earth and Space Science*, 8, e2020EA001562. https://doi.org/10.1029/2020EA001562

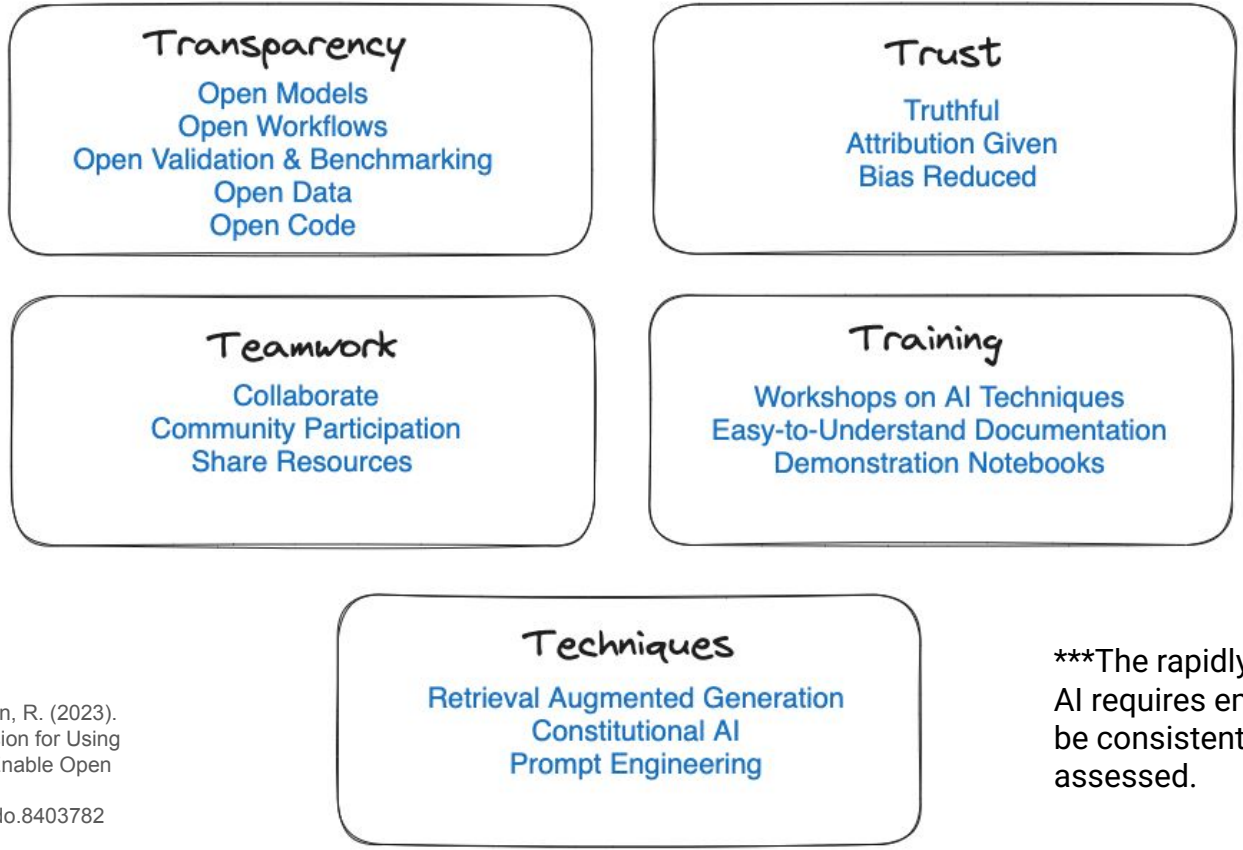
"Open science is a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding."

For data systems, open science has 3 core focus areas:

- Increasing accessibility to the scientific process & knowledge;
- Making research & knowledge sharing more efficient;
- Understanding scientific impact.

AI and Large Language Models will transform these 3 focus areas.

Open Science Principles for the AI Lifecycle



Bugbee, K., & Ramachandran, R. (2023). Architecting the Future: A Vision for Using Large Language Models to Enable Open Science. Zenodo. https://doi.org/10.5281/zenodo.8403782

***The rapidly changing nature of AI requires emerging techniques be consistently monitored and assessed.

Image Credit: NASA IMPACT team

Foundational Principle: Scientific Curation

- Scientific content curation involves authoritative experts who identify, gather, validate, synthesize, organize, and present contextual details necessary to discover, understand, and use scientific data and knowledge effectively.
- Significantly enhances **Retrieval-Augmented Generation (RAG)** by helping to ensure quality, relevance, and accuracy.
- At NASA, we've identified 600+ NASA-funded science sources:
  - Unstructured (e.g. web content) and structured information (e.g. data repository APIs) included;
  - This includes metadata about data, documents, images, code repositories, and tools.
- These sources are indexed in the **Science Discovery Engine** (SDE) which serves as an authoritative source for RAG applications.

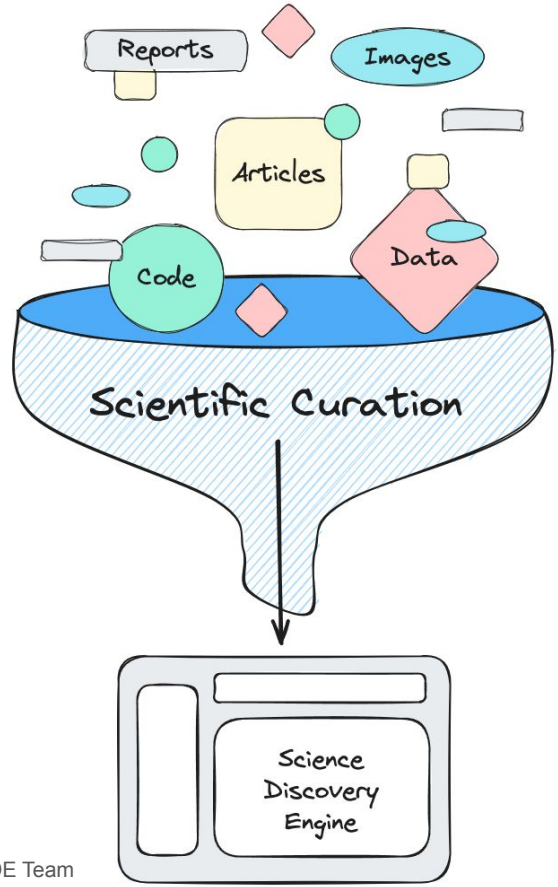


Image Credit: SDE Team

Citation: Bugbee, K., D. Smith, S. Wingo, and E. Foshee (2023), The art of scientific curation, Eos, 104, https://doi.org/10.1029/2023EO230201. Published on 19 May 2023.

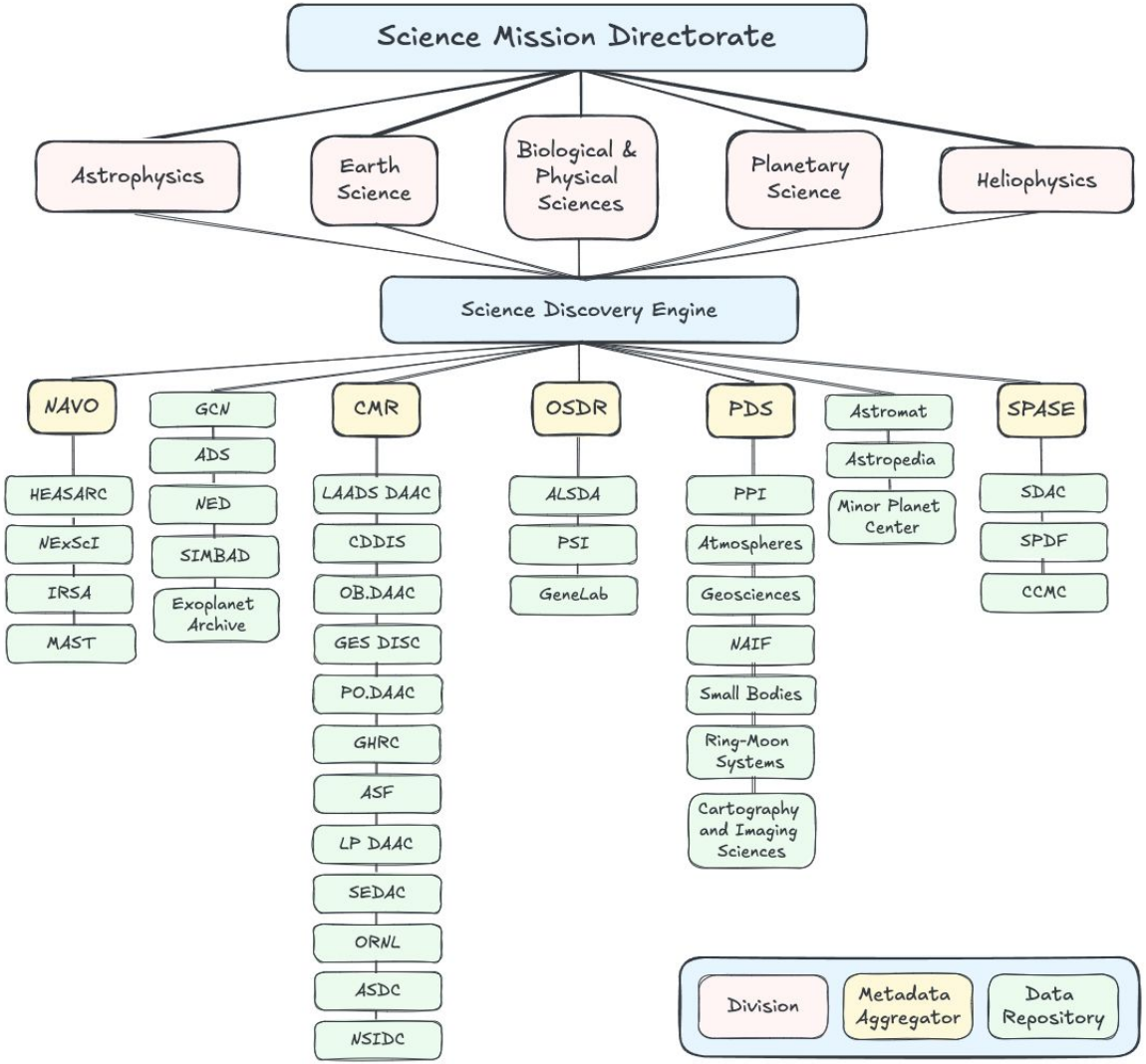


Image Credit: SDE Team

Foundational Principle: Scientific Curation

- The SDE is a source of trusted and curated open science data and information.
- The SDE includes:
  - Metadata about science data
  - Code
  - Documentation
  - Images
  - Tutorials
  - Mission and instrument information
- Includes over 84,000 metadata records and thousands of documents from hundreds of sources across NASA science.

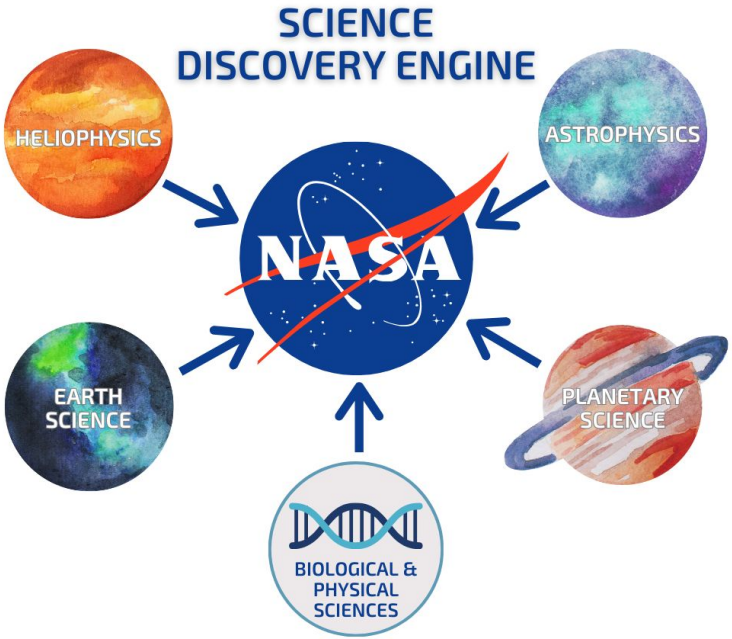


Image Credit: SDE Team

Access at:
https://sciencediscoveryengine.nasa.gov/

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

Information Management: Both Upstream and Downstream

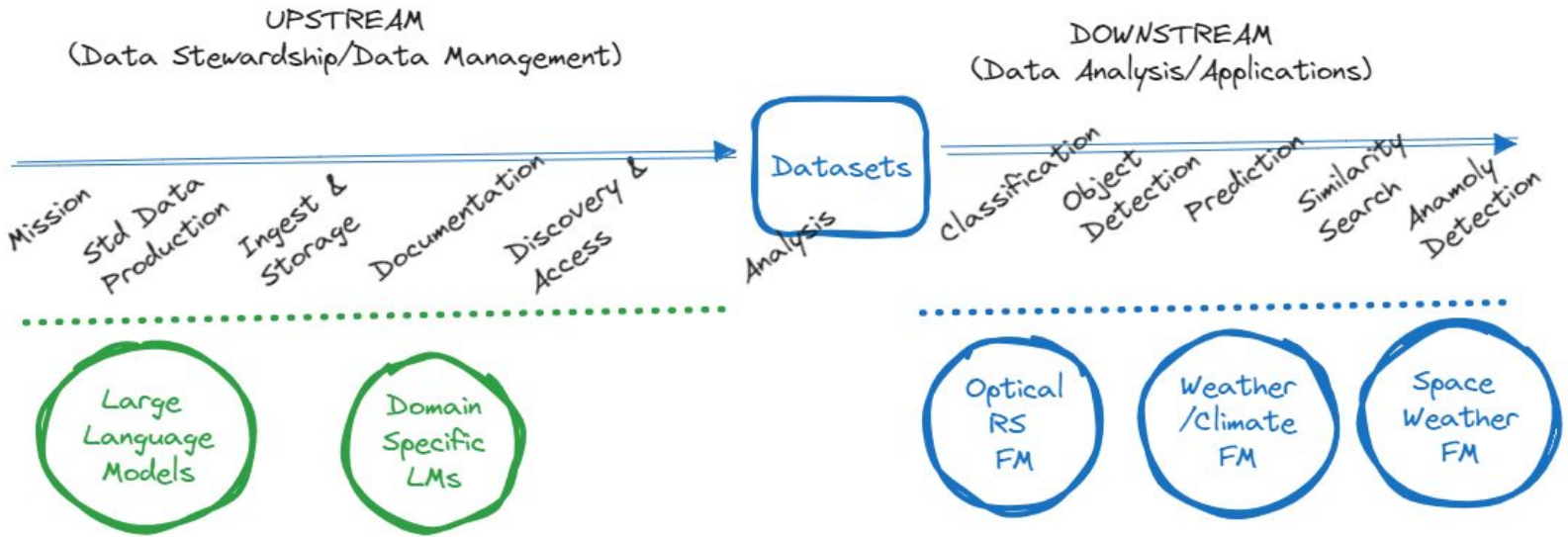


Image Credit: SDE Team

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

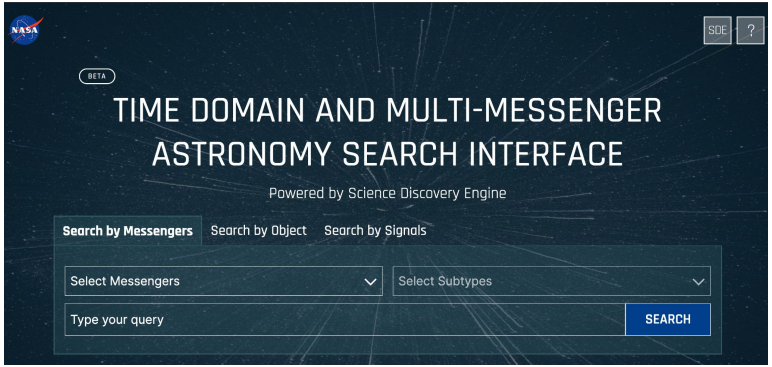
Key Takeaways from the Study

- **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

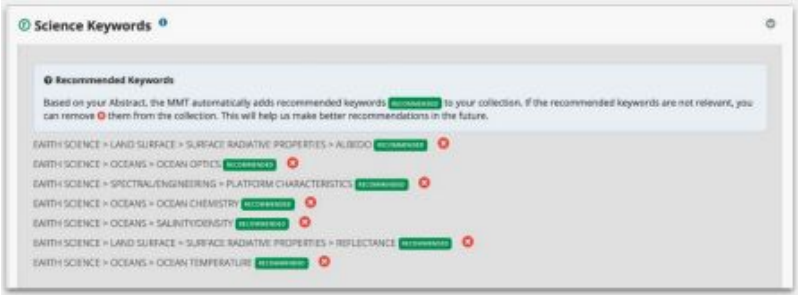
- Automated content curation.
  - Using LLMs to automate document classification for curated search and discovery experiences.
  - TDAMM, division classifier, GCMD keyword recommender, division classifier.



- 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.



When publishing a collection record through the MMT, recommendations generated by GKR automatically show up in the Science Keywords section on the Descriptive Keywords tab.

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.

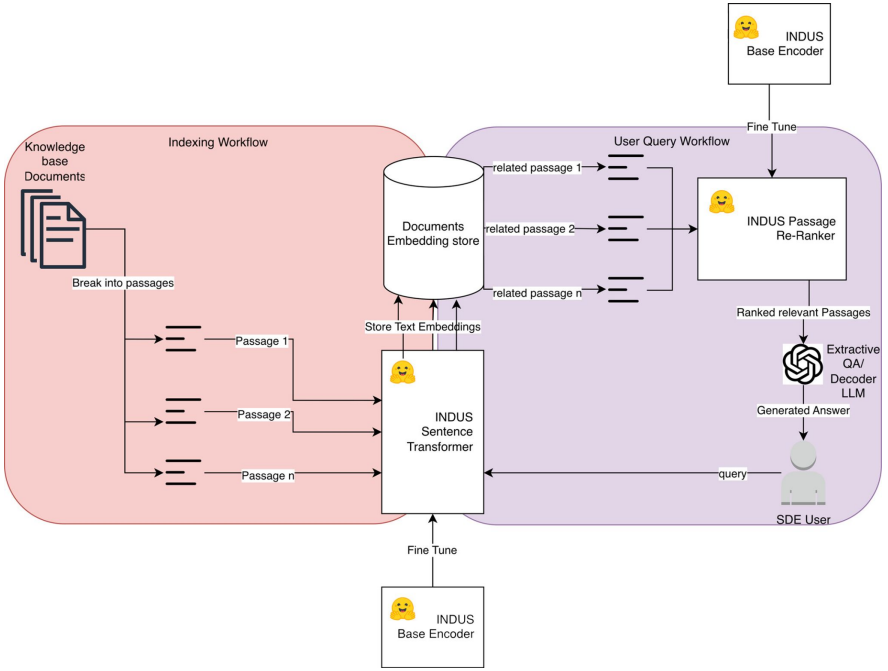


Image Credit: Ramachandran, R., & Bugbee, K. (2025). Balancing practical uses and ethical concerns:The role of large language models in scientific research. *Perspectives of Earth and Space Scientists*, 6, e2024CN000258. https://doi.org/10.1029/2024CN000258

LLM Applications: Enhanced Search

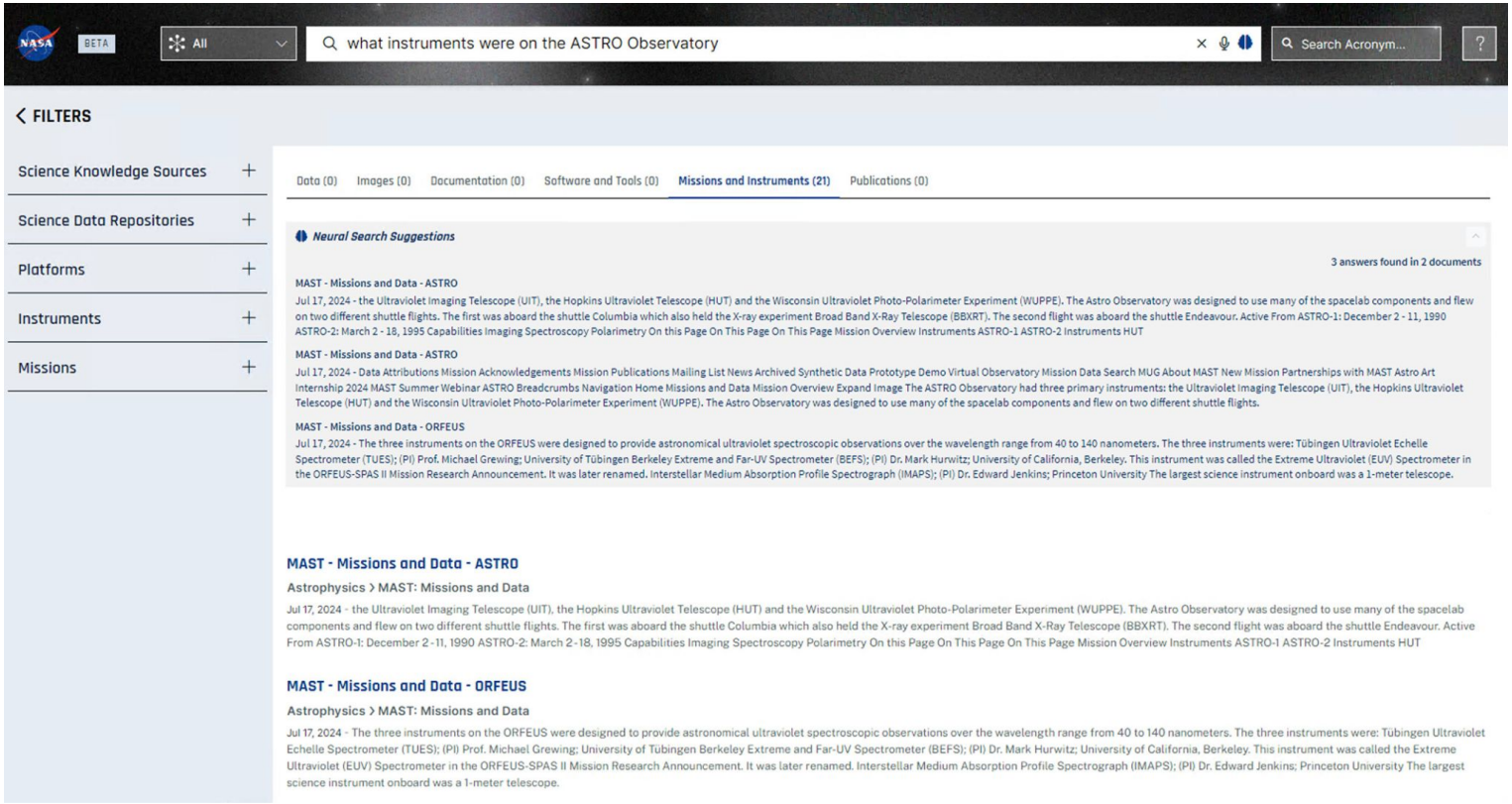


Image Credit: SDE Team

Future Direction: Accelerated Discovery

Research Workflows

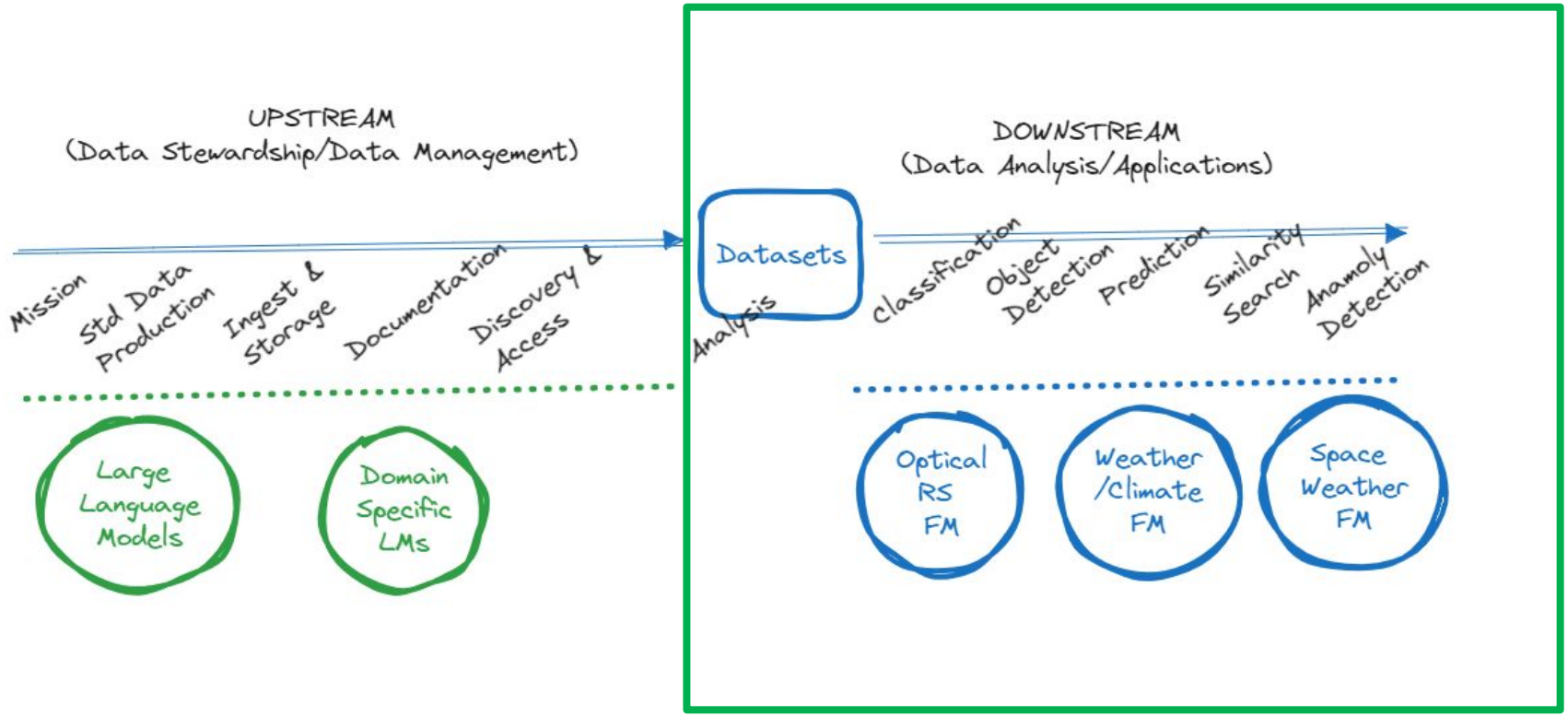
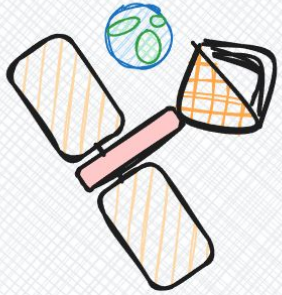
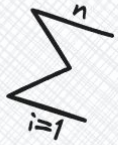


Image Credit: NASA IMPACT Team





Research Process

Algorithm Theoretical Basis Document V2.1

Scientists et al

Formulate Research Question or Problem

Could be related to climate change, disaster response, forecasting capabilities, etc

Conduct literature review of papers related to the topic

Need to understand what's been previously done and where gaps in our knowledge are

Choose which dataset(s) you'd like to use

Incorporating User Guides, metadata, validation studies, ATBDS, calibration reports, etc help users understand the benefits and limitations of the data

Apply Data Preprocessing

Depending on variables such as errors in calibration or atmospheric correction, some additional processing may need to be applied

Conduct Modeling and/or Data Analysis

Analyze the data with statistical or numerical methods, conduct data validation, and interpret results

Publish Findings

Submit work to peer-reviewed journals and share code/data in open repositories. Include data and data documentation citations (e.g. ATBDs)

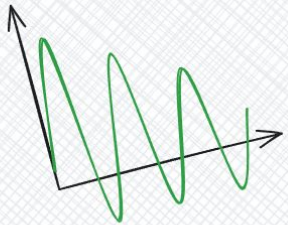


Image Credit: NASA IMPACT Team

Accelerated Discovery

We envision Accelerated Discovery as:

- An extensible Agentic Framework for Augmenting and Accelerating process of scientific research.
- Utilize agents to orchestrate various components of the research process:
  - Search Query and Intent understanding
  - Deep, Agentic Literature Search
    - Powered by SDE (for Data, Code and other NASA related documents) and SciX (for Literature Search)
  - Data Search and Discovery
  - Data Analysis
  - Report (and Insight) Generation

Future Research Workflow

Singular Scientific Research Work Area

Idea Generation

Hypothesis Generation or Context Setting

Actions:

1. *Generate hypotheses*
2. *Validate hypotheses*
3. *Understand the latest paradigms*

Outputs:

1. *Hypothesis*
2. *Curated literature list with references file (bibtex)*
3. *Summary of each document*
4. *Comparison of semantically similar papers*
5. *Recommended relevant data*

Research Iteration

Access & Analyze Data, Create Visualizations, Validation

Actions:

1. *Analyze data*
2. *Create visualizations*
3. *Validate data*
4. *Fine tune models*

Outputs:

1. *Guided ‘templates’ for common analysis types with data outputs*
2. *Visualizations and plots for inclusion in paper*
3. *Validation statistics*
4. *Updated model*

Publish Findings

Write Papers, Review Papers, Share Data & Code

Actions:

1. *Write paper*
2. *Review paper*
3. *Share supporting objects*

Outputs:

1. *Paper written per topical area template, built upon content created in ‘research’ phase*
2. *Review paper for novelty, proper citations, grammar*
3. *Data, code, images shared with proper metadata and a DOI*

Conclusions

- LLMs are transforming the scientific research and data management paradigms.
- We envision a number of issues with building and using AI/LLMs including:
  - Content overload and content pollution of dis/misinformation;
  - Potential to favor known facts over emerging ideas;
  - Lack of trust in research outputs;
  - Irresponsible use of AI;
  - Potential to limit both creativity and depth of questions asked.

Conclusions

- It will be important for NASA to consider ways to mitigate these issues such as:
  - Providing curation services for data and information to:
    - Ensure the correct use of complex data;
    - Expose topics or ideas on the margins.
  - Ensuring trust by providing traceability, credit and use of authenticated sources.
  - Providing transparency in tools and processes to ensure details and assumptions are accessible.

THANK YOU!

Questions?

Email me at kaylin.m.bugbee@nasa.gov