Data Science Group (606.3)

Mark Carroll – Lead

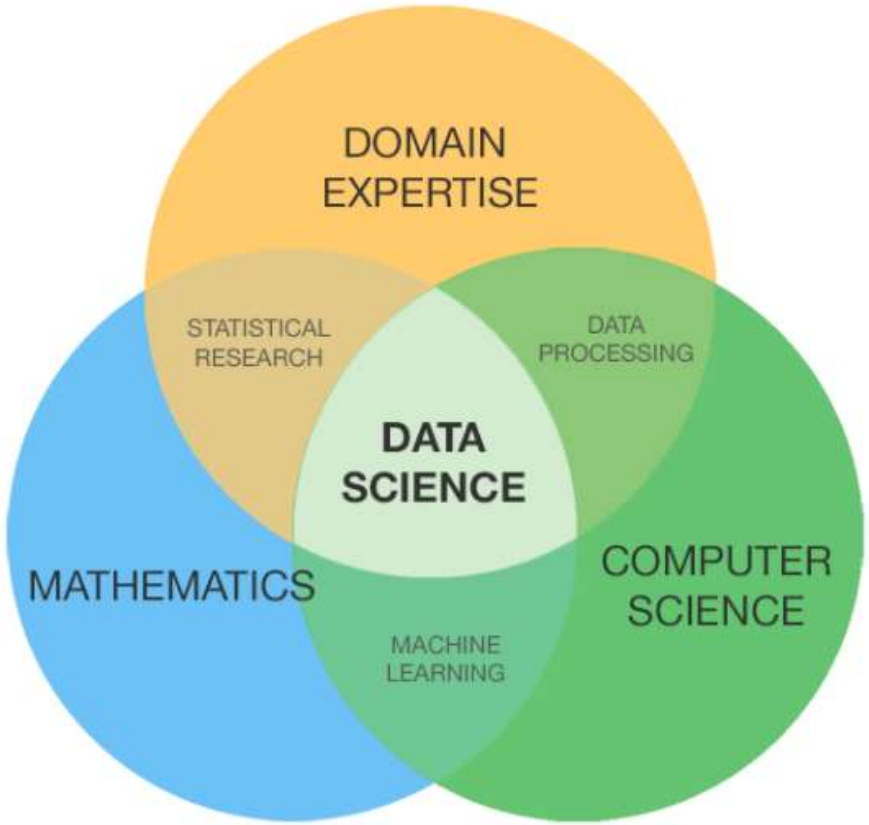
John Schnase, Paul Montesano, Roger Gill, Glenn Tamkin, Savannah Strong, Tom Maxwell, Jian Li, Jordan Caraballo-Vega, Caleb Spradlin, Mariana Blanco-Rojas, Melanie Frost

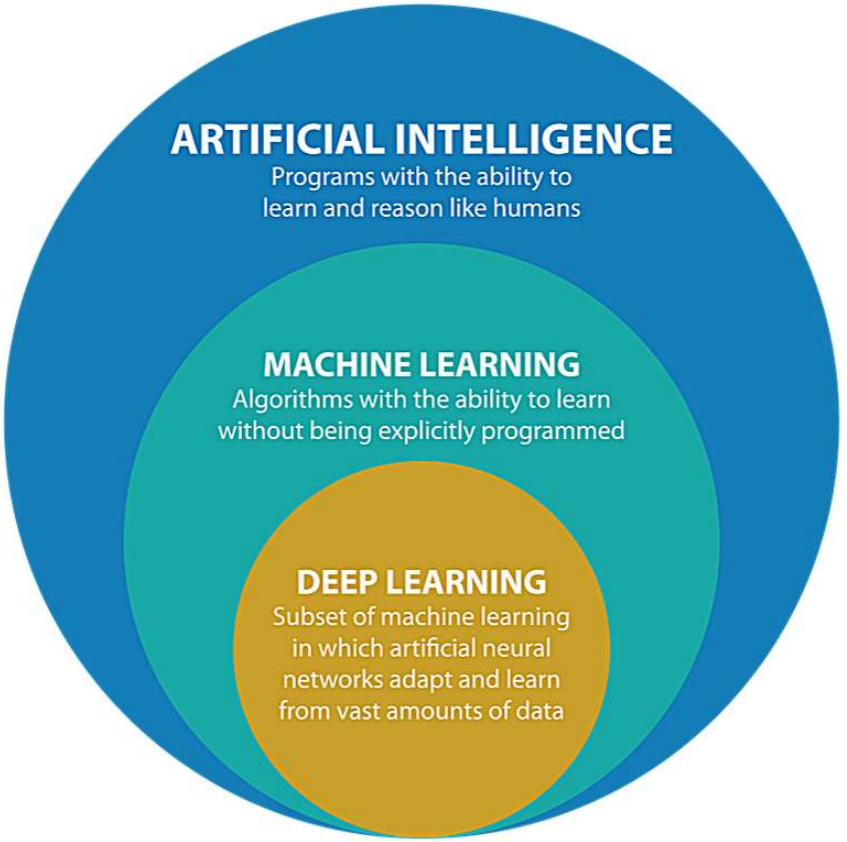
606 organization structure

Computational and Information Sciences and Technology Office Code 606.0	
Chief:	Dr. Dan Duffy
Associate Chief:	Bob Peirce

Networks and IT Security Code 606.1	High End Computing Code 606.2	Data Science Group Code 606.3	Science Visualization Studio Code 606.4
Lead: Bill Fink	Lead: Laura Carriere	Lead: Dr. Mark Carroll	Lead: Dr. Mark Subbarao

What is Data Science and AI/ML?





BLUF: GSFC is leading in AI and pathfinding for the agency

Established GSFC CAIO team and charter

Defined Goals, Objectives and Key Results

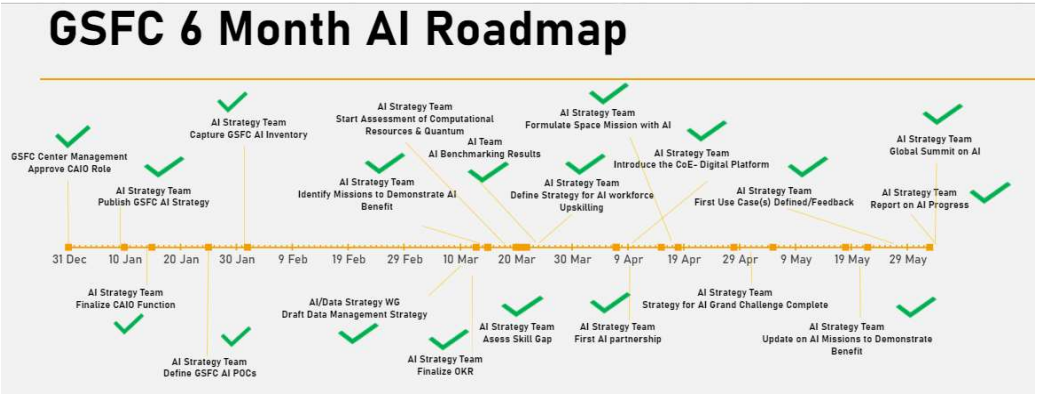
Established and enabled AI communities

Provided workforce training and upskilling

Established a digital platform to coordinate all activities

Established strategic partnerships (RAISR)

Delivered prototypes that demonstrate mission value



BLUF GSFC AI Strategy

Building and Enabling Communities

GSFC AI Lead Community

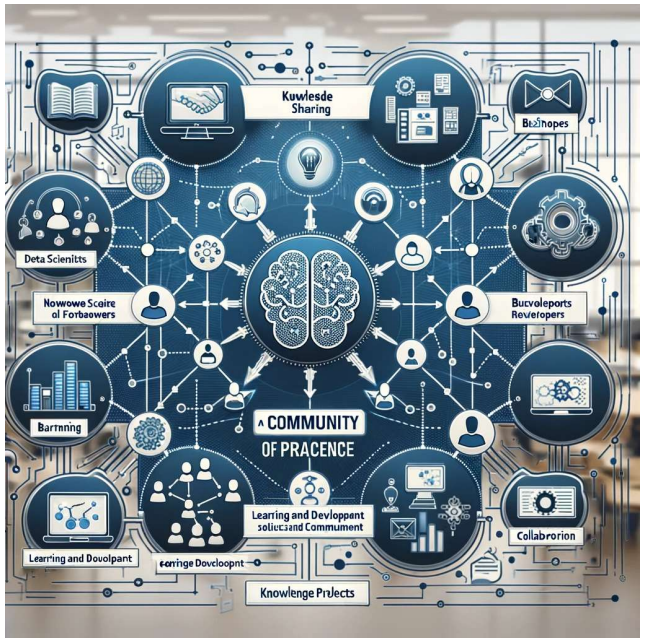
Empowered directorate leads that provide org needs (training, tools, etc.), use cases, projects, and communications. Ensures GSFC AI strategy is inclusive of all organizations.

GSFC AI Center of Excellence (CoE)

Exemplar leaders that enable scientific research, missions, and projects with the latest advances in AI/ML currently around 300 members with representation from codes 100 - 700

GSFC GenAI Community of Practice (CoP)

Leaders in Generative AI that showcase projects, collaborate on use cases, and bring the latest in technology, tools, and methodologies. Active and fast-growing ChatGSFC MS Team with over 100 members from diverse disciplines.



AI Communities are key to driving culture change

Compute systems

Explore/ADAPT Science Cloud

Explore combines high-performance computing and virtualization technologies to create an on-site private cloud. This managed virtual machine (VM) environment is specifically designed for large-scale data analytics.

The system allows researchers to bring their applications to the data and define the environment in which those applications run. The science results can then be stored for future analysis or shared with other users.

Science Managed Cloud Environment (SMCE)

The Science Managed Cloud Environment (SMCE) is a managed Amazon Web Service (AWS) based infrastructure for NASA funded projects that can leverage cloud computing capabilities.

While the SMCE was started to meet the needs of AIST projects, any NASA project that can leverage AWS public-cloud capabilities can get access to the SMCE.

Discover Supercomputer

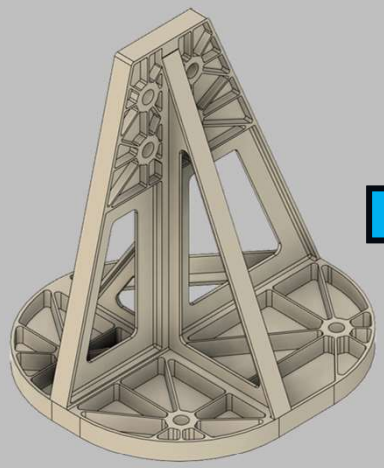
The centerpiece of the NCCS is the over 129,000-core "Discover" supercomputing cluster, an assembly of Linux scalable units capable of over 6.8 petaflops, or 6,800 trillion floating-point operations per second.

Discover is particularly suited for large, complex, communications-intensive problems employing large matrices and science applications, which benefit from its ecosystem of software ecosystem.

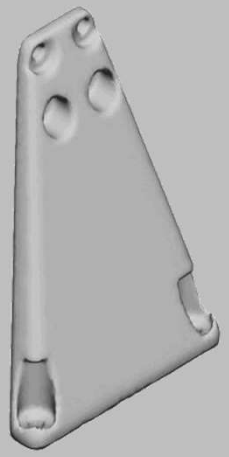
Generative AI Development at Goddard

Evolved Design:

Traditional Structure

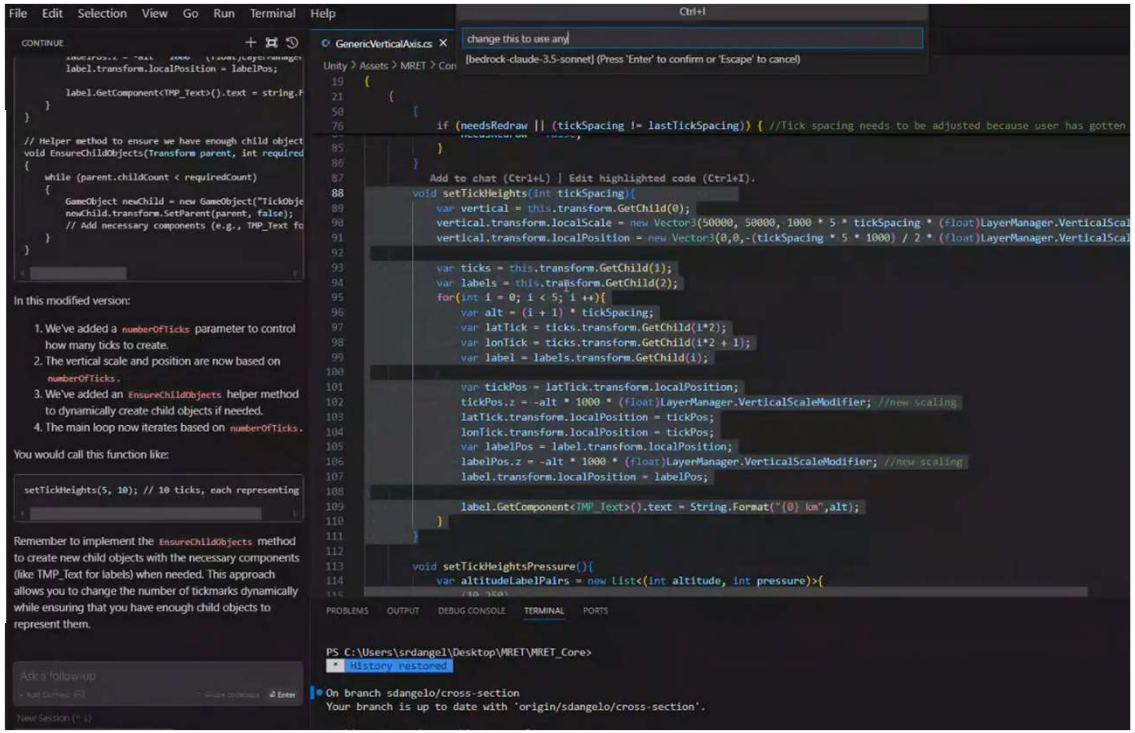


Evolved Structure



Faster, better, cheaper

Code Assistant:



AI Use Cases

Proposal Assistant

Writing Proposals: Guide first time submitters, early career PIs

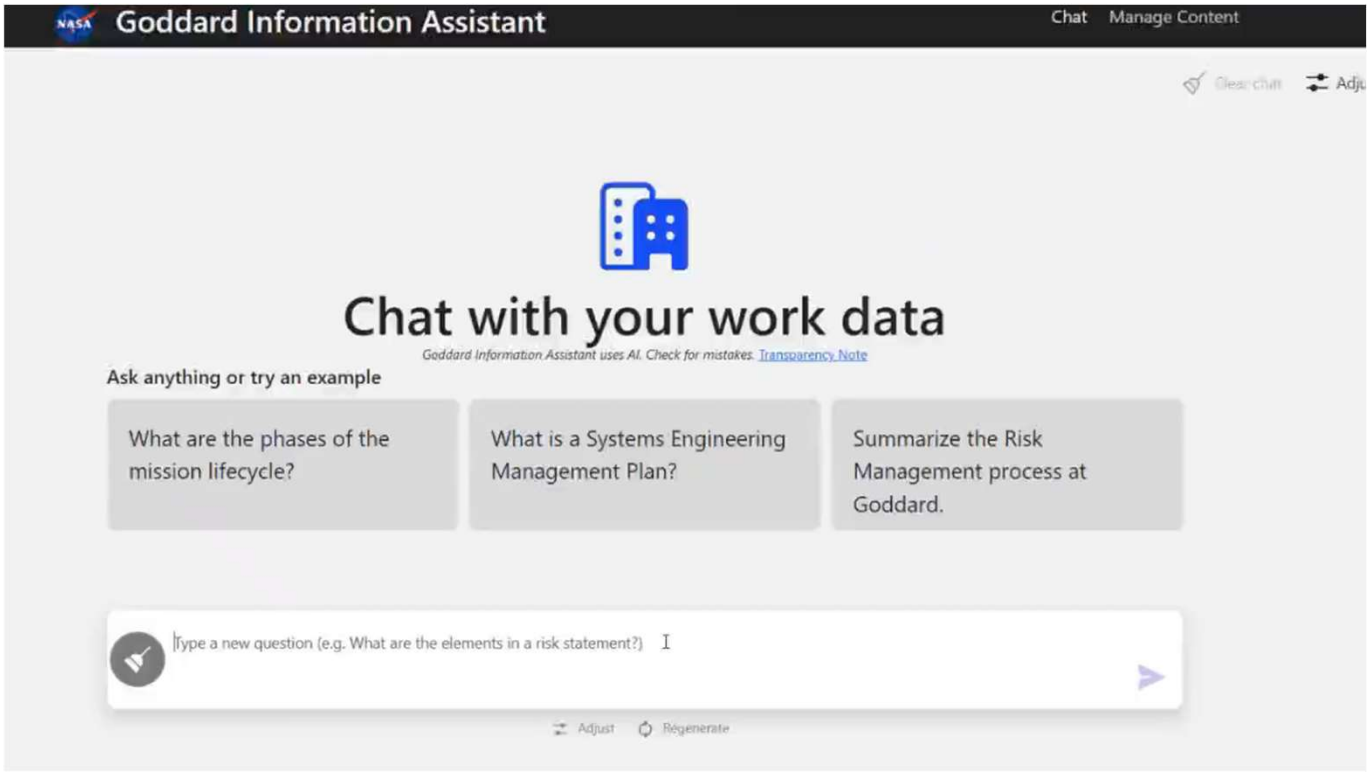
Evaluating proposals: Red team reviews on demand

$13M-30M Boost in ROSES Proposal Win Rate and Labor Saved

Useful for IRAD, ROSES, and competed mission proposals

Ongoing work:

- Surfacing related lessons learned and NTRS documents
- FY25 IRAD Step 2 selected “Goddard Proposal Assistant for Missions”



Michael Biskach

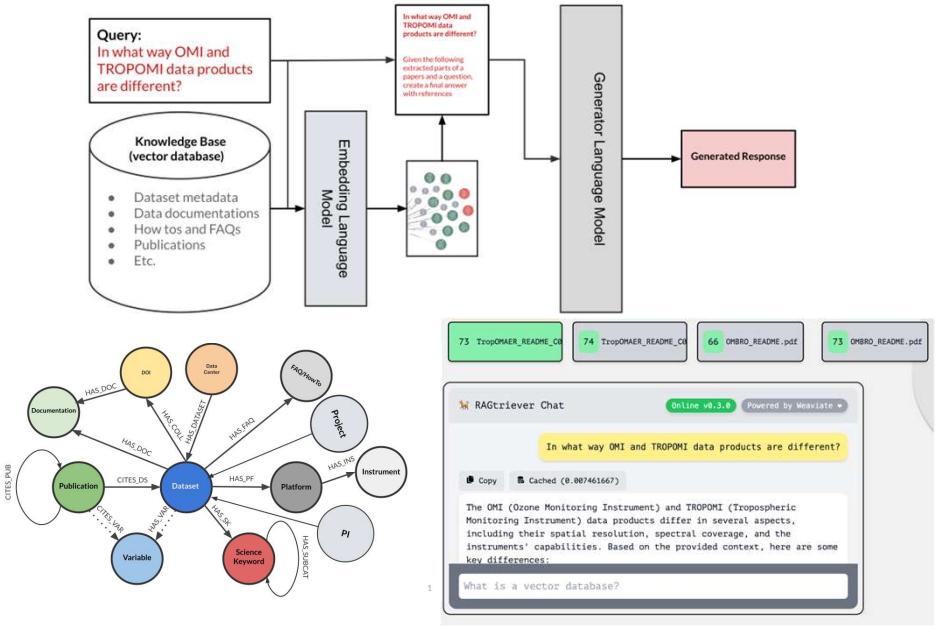
Generative AI (GPT) for data mining

PI: Armin Mehrabian (619)

Description

- NASA Goddard Earth Sciences Data and Information Services Center (GES-DISC) is one of 12 Distributed Active Archive Centers (DAACs).
- GES-DISC offers over 1,500 satellite datasets.
- Traditional search and data discovery methods rely mainly on curated metadata, often leaving other knowledge sources isolated.
- Our project creates a comprehensive knowledge graph that integrates dataset metadata and related publications.
- Leveraging this knowledge graph with graph machine learning techniques, we enhance user search and discovery. For instance, link prediction can uncover and add missing connections.
- By developing a GraphRAG, we can effectively address scientists' research questions and recommend relevant datasets and methodologies.

Graph-based Retrieval Augmented Generation

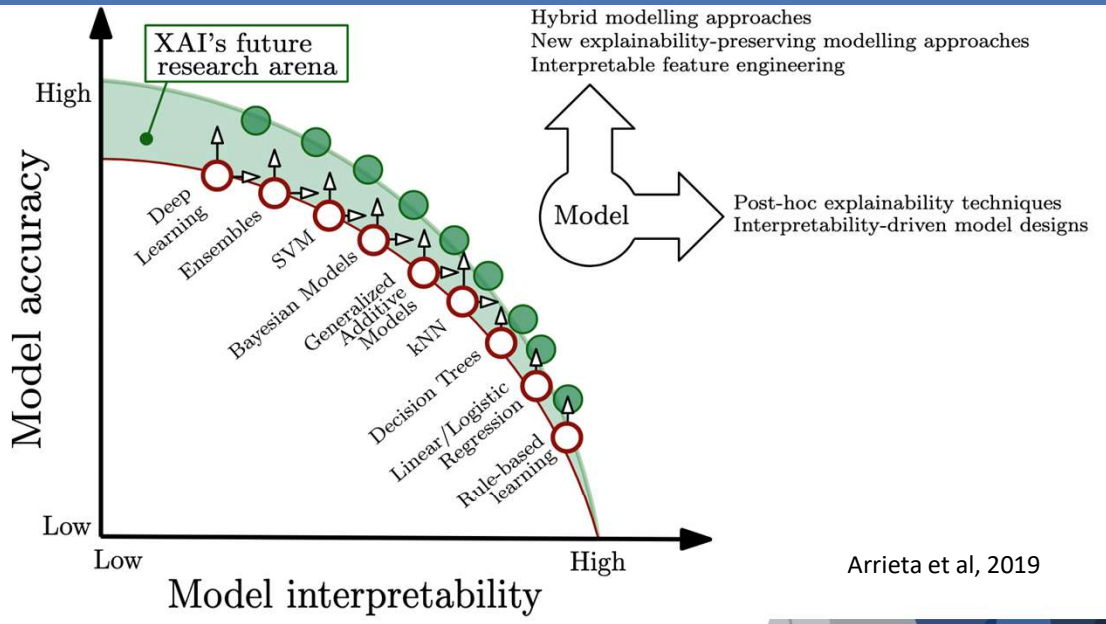


Science Impact

The GES-DISC knowledge graph enhances data discovery, revealing hidden connections and accelerating Earth science research.

eXplainable Artificial Intelligence (XAI) – Tools to explain ML models and predictions

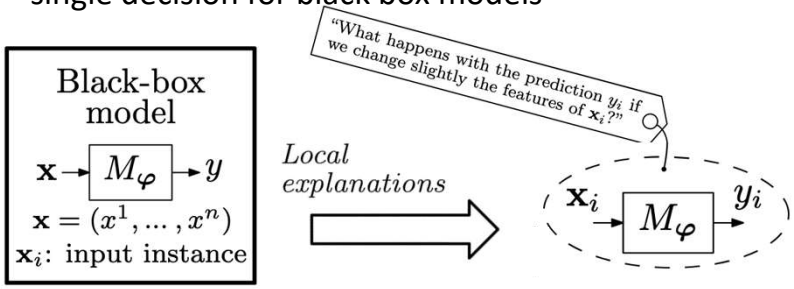
As we come to rely on inferences given by machine learning models, it is important that these models be accurate and interpretable



Arrieta et al, 2019

Using SHAP Values for Local Explanations

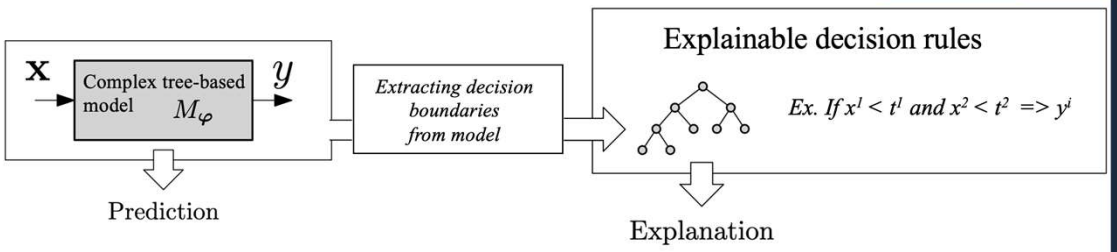
- Using Shapely values to provide explanations of single decision for black box models



Arrieta et al, 2019

https://shap.readthedocs.io/en/latest/index.html

Explaining Random Forests Through Decision Boundaries



- Developing a python library to develop interpretable tree-based classification models

https://github.com/nasa-nccs-hpda/rfexpl

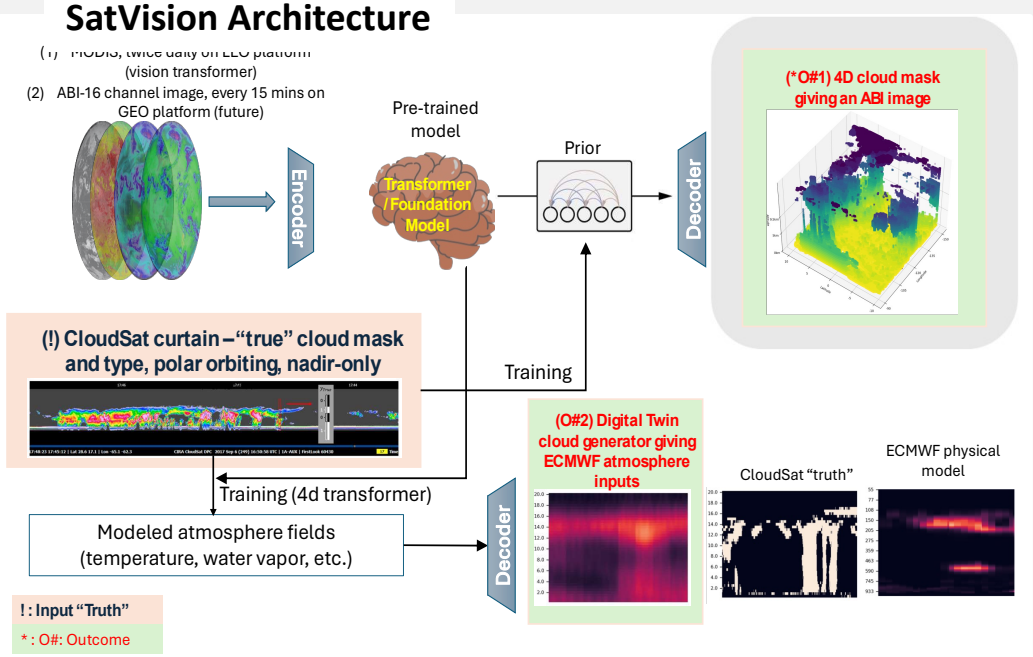
GSFC Foundation Model: SatVision

PI: Mark Carroll (606.3) and Jie Gong (613)

Description

- **A Foundation model is a class of ML models often built on "transformer" architecture and trained on huge volumes of training data**
- **SatVision is a Foundation Model designed and built at Goddard to address a science need for interrogating satellite data for atmospheric modeling**
- **This 3 Billion parameter model is being trained on 100 Million images from MODIS**
- **Initial development was performed on High Performance Compute at Goddard, final model was trained on the Frontier Supercomputer at Oak Ridge National Lab**
- **Once trained this model will be applicable to other similar spectrometer data such as GOES – ABI**
- **Release expected in Summer 2024**

SatVision Architecture



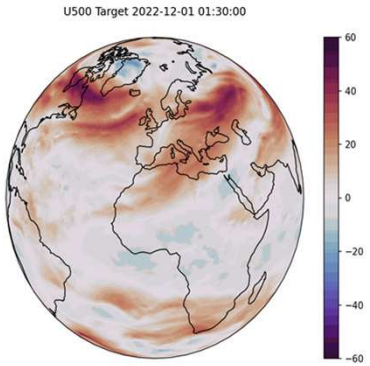
Science Impact

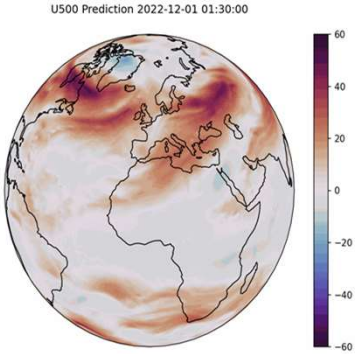
The SatVision all-sky foundation model will generate 3D cloud structures from 2D spectrometer images using sparse training. These AI/ML results provide first of their kind cloud 3D structure representation for physics-based models such as GEOS-5 to improve the cloud radiative effect and hydrological effect and ultimately improve weather forecast and climate projection.

WxC (Weather & Climate) Foundation Model

Goals

- AI FM for Weather and Climate not just on Forecasting/Prediction but for different categories of downstream applications
- Model will multiresolution both spatial and temporal to be able to use different types of data such as MERRA, ERA and HRR





Approach

- Core architectures under consideration: SWIN, Hiera
- Extensions/modifications include:
  - Multi-level and multi-resolution approaches to accommodate data at different spatial and temporal scales.
  - Diffusion-based architectures to incorporate additional information and enhance model predictions.
- Evaluation using seven different types of use cases

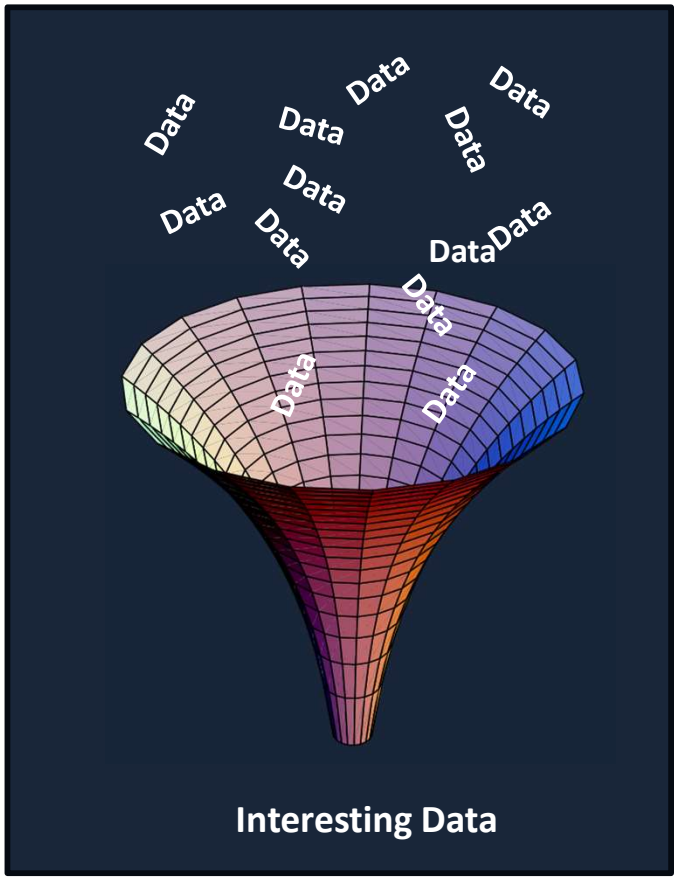
Team

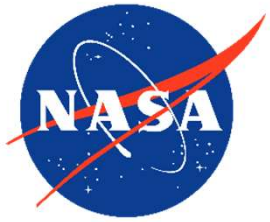
Broader participation for Science experts to ensure right direction, evaluation and future adoption of the model in their workflows
[NASA, DOE ORNL, IBM Research, NVIDIA, Academia - University of Colorado, University of Alabama in Huntsville, Stanford]

Quantitative Evaluation of Foundation Models

- Currently there are at least a dozen Foundation Models trained on climate data (MERRA-2, ERA-5, CMIP-6, etc.)
- There is limited information on how well these models can perform science relevant tasks
- We proposed a STG to develop a quantitative assessment of these models using scientists suggested case studies
  - All work to be done on Discover
  - Resulting models, scripts and data will be open to all of Goddard
  - Will represent a true “apples to apples” comparison of performance of the various FM on real world problems

Why use Data Science?





More information

- Check out our website https://science.gsfc.nasa.gov/sed/index.cfm?fuseAction=home.main&&navOrgCode=606.3
- AI Center of Excellence https://ai.gsfc.nasa.gov
  - Coming soon AI CoE GenAI working group to advance LLM work on center
- AI Inventory for 2024 AI_projects_update_2024.xlsx
- Introduction to Machine Learning
  - https://appliedsciences.nasa.gov/join-mission/training/english/arset-fundamentals-machine-learning-earth-science
- Python training
  - https://www.nccs.nasa.gov/nccs-users/user-events/python-classes
- Reach out to me directly if you want to discuss a potential project mark.carroll@nasa.gov
- Questions?