

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







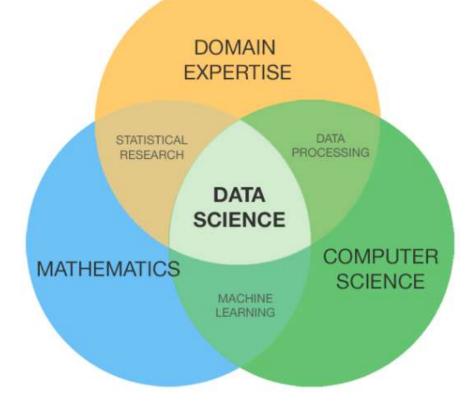
Computational and	Information Sciences and
Technology Office	
Code 606.0	
Chief:	Dr. Dan Duffy
Associate Chief:	Bob Peirce

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



What is Data Science and AI/ML?





ARTIFICIAL INTELLIGENCE

Programs with the ability to learn and reason like humans

MACHINE LEARNING

Algorithms with the ability to learn without being explicitly programmed

DEEP LEARNING

Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data



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

GSFC 6 Month Al Roadmap



Omar Hatamleh



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 Al/ML currently around 300 members with representation from codes 100 - 700

GSFC GenAl 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 publiccloud capabilities can get access to the SMCE.





Discover Supercomputer

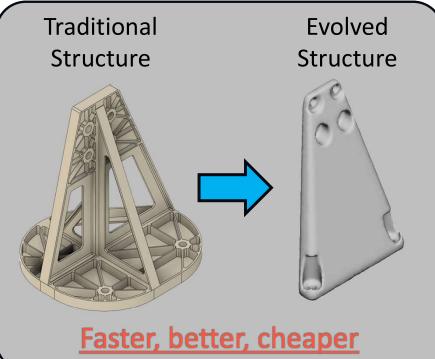
The centerpiece of the NCCS is the over 129,000-core "Discover" supercomputing cluster, an assembly of Linux scalable units capable of over 6.8 petaflops, or 6,800 trillion floating-point operations per second.

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





Evolved Design:



Code Assistant:

File Edit Selection View Go Run Terminal Help		Ctul+1
	GenericVerticalAxis.cs X	change this to use any
label.transform.localPosition = labelPos;	Unity > Assets > MRET > Cor	(bedrock-claude-3.5-sonnet) (Press 'Enter' to confirm or 'Escape' to cancel)
Tabel (Label) of a local of a label of a		
label.GetComponent <tmp_text>().text = string.F</tmp_text>		
		(needsRedraw (tickSpacing != lastTickSpacing)) (//Tick spacing needs to be adjusted because user has gotten
<pre>// Helper method to ensure we have enough child object</pre>		incontrol on theory
<pre>void EnsureChildObjects(Transform parent, int required {</pre>		
while (parent.childCount < requiredCount)	87 Add	to chat (Ctrl+L) Edit highlighted code (Ctrl+I).
{	88 void se	tTickHeights(int tickSpacing)(
<pre>GameObject newChild = new GameObject("TickObje newChild.transform.SetParent(parent, false);</pre>		vertical = this.transform.GetChild(0);
// Add necessary components (e.g., THP_Text fo		tical.transform.localScale = new Vector3(50000, 50000, 1000 * 5 * tickSpacing * (float)LayerManager.VerticalScal
		tical.transform.localPosition = new Vector3(0,0,-(tickSpacing * 5 * 1000) / 2 * (float)LayerManager.VerticalScal
		ticks - this.transform.GetChild(1);
		labels = this.transform.GetChild(2);
n this modified version:		(int i = 0; i < 5; i ++)
		var alt = (i + 1) * tickSpacing;
1. We've added a numberofflicks parameter to control		<pre>var latTick = ticks.transform.GetChild(1*2); var lonTick = ticks.transform.GetChild(1*2 + 1);</pre>
how many ticks to create.		<pre>var label = labels.transform.GetChild(1'2 + 1);</pre>
The vertical scale and position are now based on		Var label - labels. transform. GetChild(1/)
numberOfTicks.		var tickPos - latTick.transform.localPosition:
3. We've added an EnsureChildObjects helper method		tickPos.z = -alt * 1000 * (float)LaverManager.VerticalScaleModifier; //new scaling
to dynamically create child objects if needed.		latTick.transform.localPosition = tickPos:
The main loop now iterates based on numberOfTicks.		IonTick-transform.localPosition = tickPos;
		var labelPos = label.transform.localPosition;
You would call this function like:		labelPos.z = -alt * 1000 * (float)LayerManager.VerticalScaleModifier; //new scaling
		label.transform.localPosition = labelPos;
<pre>setTickHeights(5, 10); // 10 ticks, each representing</pre>		
		label.GetComponent <tmp_text>().text = String.Format("(0) km",alt);</tmp_text>
Remember to implement the EnsureChildObjects method		
to create new child objects with the necessary components		
(like TMP_Text for labels) when needed. This approach		tTickHeightsPressure()(
allows you to change the number of tickmarks dynamically	114 Var	altitudeLabelPairs = new List<(int altitude, int pressure)>{
while ensuring that you have enough child objects to	PROBLEMS OUTPUT DEB	
represent them.		
		Desktop\MRET_Kore>
	* History restored	
	In the base of the state of the	
	On branch sdangelo/cr	oss-section jate with 'origin/sdangelo/cross-section'.
	rour branch 15 up to i	and with or ignify sumgers/cross-section .



AI Use Cases

Proposal Assistant

HASA GSPC

	Goddard Information A	ssistant	Chat I	Manage Content		
Writing Proposals: Guide first time submitters, early career PIs				🔗 Ölesi chin	≢ Adju	
Evaluating proposals: Red team reviews on demand						
	Chat with your work data					
\$13M-30M Boost in ROSES Proposal Win Rate and Labor Saved		Idard Information Assistant uses AI. Check for mistakes. <u>Transpared</u>				
Useful for IRAD, ROSES, and competed mission proposals	What are the phases of the mission lifecycle?	What is a Systems Engineering Management Plan?	Summarize the Risk Management process at Goddard.			
 Ongoing work: Surfacing related lessons learned and NTRS documents 	Type a new question (e.g. What are the elements in a risk statement?) I			>		
 FY25 IRAD Step 2 selected "Goddard Proposal Assistant for Missions" 		- refron C recificantese				

A B C D E F

Michael Biskach



Generative AI (GPT) for data mining PI: Armin Mehrabian (619)



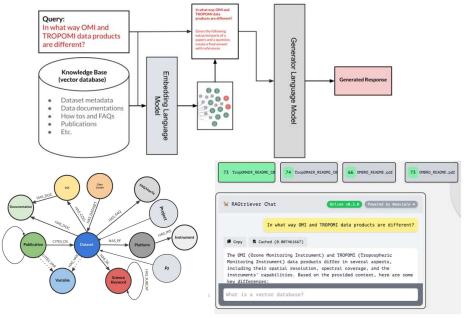
Description

- 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.

Science Impact

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

Graph-based Retrieval Augmented Generation

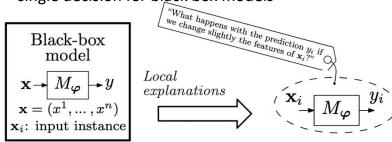


eXplainable Artificial Intelligence (XAI) – Tools to explain ML models and predictions

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

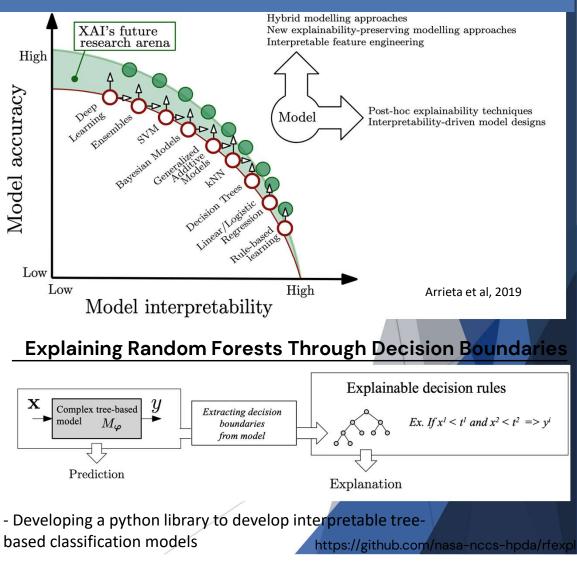
Using SHAP Values for Local Explanations

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





GSFC Foundation Model: SatVision PI: Mark Carroll (606.3) and Jie Gong (613)

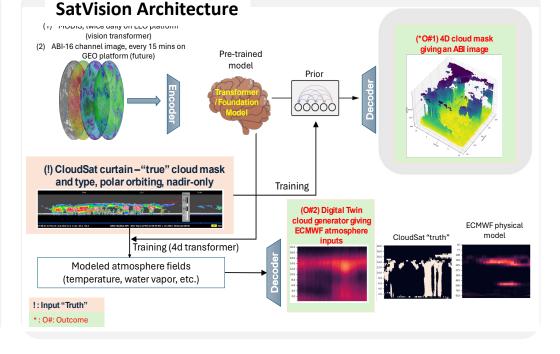


Description

- 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

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.





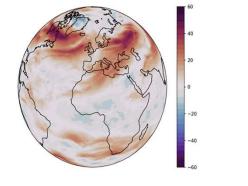


Goals

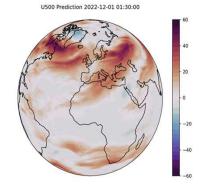
- AI FM for Weather and Climate not just on Forecasting/Prediction but for different categories of downstream applications
- Model will multiresolution both spatial and temporal to be able to use different types of data such as MERRA, ERA and HRR

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



U500 Target 2022-12-01 01:30:00



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]



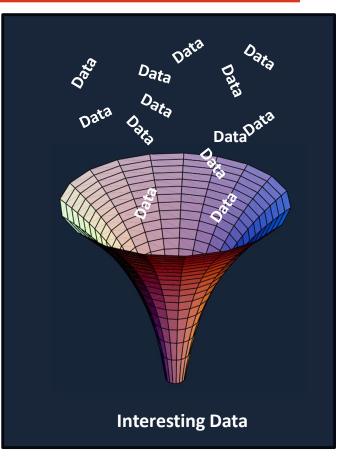


- Currently there are at least a dozen Foundation Models trained on climate data (MERRA-2, ERA-5, CMIP-6, etc.)
- There is limited information on how well these models can perform science relevant tasks
- We proposed a STG to develop a quantitative assessment of these models using scientists suggested case studies
 - o 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













- Check out our website <u>https://science.gsfc.nasa.gov/sed/index.cfm?fuseAction=home.main&&navOrgCode</u> <u>=606.3</u>
- AI Center of Excellence <u>https://ai.gsfc.nasa.gov</u>
 - Coming soon AI CoE GenAI working group to advance LLM work on center
- Al Inventory for 2024 <u>Al_projects_update_2024.xlsx</u>
- Introduction to Machine Learning
 - <u>https://appliedsciences.nasa.gov/join-mission/training/english/arset-fundamentals-machine-learning-earth-science</u>
- Python training
 - <u>https://www.nccs.nasa.gov/nccs-users/user-events/python-classes</u>
- Reach out to me directly if you want to discuss a potential project <u>mark.carroll@nasa.gov</u>
- Questions?