Multidomain Reusable Artificial Intelligence Tools
Abstracts of selected proposals
(NNH22ZDA001N-MDRAIT)

Below are the abstracts of proposals selected for funding for the Multidomain Reusable Artificial Intelligence Tools program. Principal Investigator (PI) name, institution, and proposal title are also included. 18 proposals were received in response to this opportunity. On April 28, 2023, 8 proposals were selected for funding.

Srija Chakraborty/Universities Space Research Association, Columbia
Domain Adaptive Anomaly Detectors for Extracting and Identifying Rare Class Signals

Recent years have witnessed a deluge of data in Earth and Space Sciences providing a unique opportunity to offer insights into different physical events of interest. These events are often rare and embedded in large datasets and can be challenging to discover due to the dominant uninteresting data" and worsened by the higher dimensions of multispectral observations. Anomaly detectors effectively leverage large datasets by modeling the dominant class and are ideally suited to then detect the anomalous, rare class from its deviation from the model. Moreover, these detectors are unsupervised, minimizing the labeling effort from subject matter experts (SME). This study proposes an anomaly detection approach that adapts to large, multispectral datasets to extract interesting, rare class signals followed by its categorization using domain knowledge from SME to maximize the scientific return from the instruments.

The proposed detectors have been applied on daily multispectral nighttime data from NASA's Black Marble Product acquired by the Visible Infrared Imaging Radiometer Suite (VIIRS) to detect thermal anomalies caused by combustion from fires and gas flaring. Accurate accounting of these events is crucial due to associated greenhouse gas (GHG) emissions and satellite-based detections are used to derive bottom-up emission estimates. We have used Reed Xiaoli Detectors, Autoencoders, and Variational Autoencoders to detect combustion in six test sites over diverse geographic areas during peak occurrence months from 2018-2020. Our approach shows high overlap with current experimental VIIRS datasets and visible flaring sites in Landsat, and we also find likely combustion missing in these datasets. A containerized workflow has been implemented on the Nautilus Hypercluster by interfacing with LAADS for scalability and can thus be viewed at TRL 5. We will expand our approach to detect offshore flares and also expect to detect volcanoes and shipping vessels. We will further mature our approach by using multispectral context to potentially cluster and map the detections. The detections will be evaluated with existing VIIRS datasets and visible signatures of fires and flaring sites in Landsat. The two-step process of deriving VIIRS-based accounts of GHG emitters followed by identification of likely event class should improve satellite-based monitoring of emission events and support further decision-making.

We will also apply the anomaly detectors on multispectral data from the Atmospheric Imaging Assembly (AIA) onboard the Solar Dynamics Observatory (SDO) to automate the detection of a new class of solar energetic events at the limit of instrument detectability and often confused as cosmic rays. We will first apply minimal preprocessing to eliminate saturated pixels. The detectors will then model the dominant class in the dataset and extract anomalous signals relating to solar energetic events from their deviations. We will examine the multispectral signature and use domain knowledge from SME to group potential differences in solar energetic events. Our detections will be compared with anomalous events already curated by SME for validation. By modeling a new class of solar emission and detecting cosmic rays penetrating into the solar system, the detectors are relevant to determine the origins of the Sun's activity, its interaction with the solar system and interstellar medium, and predicting space environment variability.

This effort will create an open-source anomaly detection methodology to extract and identify interesting signals" from multispectral Earth and Space Science datasets and accelerate SME analysis. Post-project, we will periodically evaluate the detectors on new events to identify scenarios for model adjustment. We expect our approach to be applicable to a wider range of instruments and create data-driven, knowledge-guided rare class signal catalogs.

Alexander Engell/NextGen Federal Systems, LLC
Multidomain-Multimodal Data Dashboard for Big Data Analysis and AI/ML Applications

Both heliophysics and Earth science disciplines are continually growing the volume and complexity of data amassed due to new missions and more sophisticated models. Meeting science objectives to perform efficient data discovery and apply artificial intelligence (AI) and machine learning (ML) applications at scale is challenging. Innovative and open-source software (OSS) is a cornerstone and pathway to improving these AI/ML abilities.
To advance Heliophysics and Earth Science research through AI/ML techniques, reusable and easy to use customizable tools for data visualization, annotation, and exploration are required. NextGen Federal Systems (NextGen) is excited to propose the Multidomain-Multimodal Data Dashboard (M2D2) tool developed under previous research efforts to address this need. M2D2 will be generalized with new OSS code to support historical and real-time data visualization and analysis at scale for a variety of SMD data types, formats, and coordinate systems.

M2D2 is an end-to-end software system for web-based data visualization and is currently at a TRL 5. It is used to display historical and real-time time-series and imagery data in AWS GovCloud, which is a relevant environment for cloud-based analysis.
M2D2 was first developed as part of NextGen's Soil Moisture Advancements from Research to Transition (SMART) project. SMART applied ML to generate synthetic soil moisture products using alternate remote sensing datasets. The M2D2 tool was further

enhanced through development of our Space Radiation Intelligence System (SPRINTS) capability, which is a cloud-hosted collaborative ecosystem for repeatable heliophysics data processes, science, and ML forecasting. The SPRINTS workflow currently uses M2D2 dashboards to annotate (define and associate) solar X-ray flares and solar proton events for ML modeling to predict solar radiation events.

Supported M2D2 modalities include time-series measurements, interval-based events, time animated imagery, 4D in situ measurement representations, and forecast displays. Imagery and time-series annotations, and linked time-series and remote-sensed imagery scrubbing are available to use across a variety of widget panels that can make up a custom dashboard display. The multimodal and annotation capabilities are key to ML efforts with large volumes of diverse data.

For the MDRAIT program, M2D2 will be enhanced for greater interoperability across data enterprises, scientific workflows, and ML applications. Regarding MDRAIT's notional areas of interest, M2D2 applies to the data visualization and AI-ready data preparation categories.

Annotation tools implemented in M2D2 dashboards will support time-series analysis of a variety of time-based data types (EUV post-flare loop plasma cooling, atmospheric dynamics, soil properties, radio occultation profiles, etc.). These annotation tools will support and enable the creation of ML-ready datasets (for example, ground-truth labels of events of interest).

In this project, a generalized version of the M2D2 tool will be customized to show how the baseline annotation feature can be applied to different domains, demonstrating its reusability. The customized dashboards will provide easy to understand blueprints for how a user can take the generalized M2D2 codebase and customize it to perform complex data interactions and analysis for their science and ML objectives.

Targeted product owners include but are not limited to the Python Heliophysics Community (PyHC), Helioanalytics, NASA HelioCloud, NASA Science Managed Cloud Environment (SMCE), the Pangeo community, and NOAA NESDIS and SWPC scientists and engineers. Focus on M2D2 development will be on how it can be matured to streamline AI/ML processes in the heliophsyics and Earth science domains.

Matthew Finley/University of Iowa, Iowa City
MAGSTAR: Multi-Mission MAGnetometer Denoising and Sensor Resiliency through STatistical Decomposition and ARtificial Intelligence

This 12-month proposal will support the development, testing, and generalization of a robust denoising technique for magnetic field sensors. The resulting toolset will utilize statistical techniques to enable noise mitigation for several missions in multiple disciples when a pair of magnetometers is available, with machine learning providing resilience when only a single magnetometer is present.

Magnetometer measurements are necessary to understand the mechanisms that couple mass and energy throughout our solar system. However, magnetic field measurements are often contaminated by interference from other subsystems on the host spacecraft. One common technique for the mitigation of this interference is by deploying a pair of magnetometers (i.e., a gradiometer) along a common boom away from the main body of the host spacecraft. The magnetic gradient between the two sensors can be fit to an ideal dipole field which is representative of the spacecraft's magnetic field and subsequently removed, eliminating the local interference. However, many missions not designed for magnetic cleanliness utilize shorter booms to reduce technical complexity and implementation costs. This reduces the effectiveness of standard gradiometric denoising techniques, since field measurements made closer to the host spacecraft will experience multi-pole moments that cannot be removed without careful pre-flight calibration and characterization of all possible interference sources, which is often impossible due to the limited magnetic cleanliness provided by most vehicle assembly facilities.

To overcome this obstacle, an end-to-end software system has been previously developed by the proposer to simultaneously decompose inboard and outboard gradiometric magnetic field measurements into physically meaningful sub-signals using a statistical technique called Singular Spectrum Analysis. These decomposed signals are then labeled as local spacecraft interference or residual geophysical signal through statistical analysis of the magnetic gradient between the two sensors. The labeled interference terms are subsequently removed, improving data product fidelity without the need for careful pre-flight calibration or characterization of other spacecraft subsystems. Machine learning is then used to enable robust alternative classification of the decomposed signals in the case where only a single magnetic field measurement is available. This software has been previously prototyped for the CASSIOPE Magnetic Field Instrument.

This proposal will develop this existing Technology Readiness Level (TRL) 5 software suite (e.g., end-to-end software system, tested in relevant environment") to TRL 6 (e.g., prototype implementations of the software demonstrated on full-scale realistic problems") by improving and generalizing the software such that it can be validated against large datasets from several different missions. Additionally, post-project support will be provided by long-term storage and availability of the developed codebase in a public repository. The missions that this proposal targets for interference mitigation are Parker Solar Probe (Heliophysics), GRACE-FO (Earth Science), and MAVEN (Planetary Science). This robust and generalizable interference mitigation toolset will enable the following goals and objectives from various NASA divisions:

- Science Mission Directorate's Strategy for Data Management and Computing for Groundbreaking Science 2019-2024. (Goal 2: Continuous Evolution of Data and Computing Systems" via Strategy 2.4: Invest in the tools and training necessary to enable breakthrough science through application of AI/ML.")
- Earth Science Division's objective to understand the dynamics of Earth's magnetic field and its interactions with the rest of Earth's systems."
- Heliophysics Divisions's primary objective to understand the Sun and its interactions with the Earth and the Solar System."

Dustin Kempton/Georgia State University Research Foundation, Inc.
Tools for Preparing Machine Learning Ready Multivariate Time Series Datasets

The goal of the proposed works are to increase accessibility and scientific readiness of a set of software tools. These tools are a core functionality for constructing and processing datasets that can be used as testbeds for heliophysicists, earth scientists, and machine learning (ML) practitioners. These tools are applicable to any domain in which multi-feature time series are utilized for a supervised machine learning task. The main focus of our development interest revolves around the software elements created along with the benchmark dataset named Space Weather ANalytics for Solar Flares (SWAN-SF). This dataset was produced for testing solar flare prediction algorithms, and already has over 2,400 downloads and supported a worldwide best model competition within the 2019 IEEE BigData conference (86 teams participated). We plan to build on its gaining popularity and expand the capabilities of the software infrastructure used to reprocess the dataset for different task definitions, sampling strategies, feature selection methods, and other pre-processing steps which may impact the reliability of any models trained using the multivariate time series dataset.

The project will build upon the cyber-infrastructure we have been developing for data-enabled scientific research on time series metadata derived from many years worth of observations of solar magnetic active regions. Beyond producing several methods and pieces of code to systematically generate, clean, and label a multivariate time series dataset, additional methods and code were devised to encourage the proper utilization of the data. For instance, to provide a machine learning-ready dataset with consistent descriptive feature observation periods and target feature prediction periods, the larger time series of parameters for any instance (e.g. an active region) required a sampling mechanism to be applied to it. Furthermore, if improper methods are used when constructing training and testing subsets of the larger dataset, this can lead to overly optimistic results being reported that may lead to unexpected degraded performance when unseen data is encountered during deployment. So, in order to mitigate this possibility, we have also developed a partitioning strategy to break such datasets into non-overlapping subsets seeking to have similar numbers of positive instance labels in each partition.

The software developed throughout this project shall be applicable to time series data analysis and management in particular, and data partitioning and balancing in general, in the domains of Heliophysics, Earth Sciences, and beyond. The developed toolkit will be useful for STEM education since it provides organized data and tools with the proper balance of ease of use while also allowing for the use of complex data and problems in a teaching environment. The collection of projects and code developed to work with the MVTS data discussed above can be seen as a cohesive ecosystem made possible by a standardized methodology and format of producing data and models trained on those data products. However, all of the functionality is dispersed through several different repositories, designed and implemented by different individual researchers within our lab,

over years of preparation. These different software products are generally in a state of TRL 5 to 6 where code has been developed to produce and process data, and has been validated through several project integrations. However, they are generally not in a well-documented and easily usable state for wider user adoption. In order to increase the TRLs of these software products, and allow for wider adoption, these repositories and functionalities need to be condensed into a cohesive, well-documented, open-sourced, software library, with both a well-written set of demonstrations and community outreach to inform of its existence.

James Parr/Trillium Technologies, Inc.
ITI (Instrument to Instrument) Tool - Unlocking Collaborative Sensor Webs

Science Goals
The Instrument-To-Instrument (ITI) tool is an integrated software package that translates between images acquired by different observing sensors. We believe ITI is ideal for MDRAIT as it enables machine learning (ML) methods to up-scale resolution (super-resolution), estimate realistic images in unseen colour bands and cross-calibrate brightness scales between disparate instruments. The augmented ITI data product is proven to be statistically similar to high-quality modern data, meaning that it can readily be used in combination with new instruments to create observations as a coherent system - i.e collaborative sensor webs. The goals of this proposal are 1) to enhance and expand ITI to work with Earth observation (EO) data and Earth-science use cases; 2) make the package easier to use and thus entry barriers and expand the user-base.

Methods (current -> changes -> final state)
ITI has previously been applied to five different Heliophysics use cases and implements a framework for loading different solar observatory data, including HMI, HiNode, SoHo, EMI, STEREO, KSO and SDO. We see ITI as a true general AI tool, similar to 'CURATOR' - a self-supervised learner developed by our team: https://github.com/spaceml-org/Self-Supervised-Learner which has been adapted to multiple NASA use cases, from Worldview to Hubble.

ITI is implemented as a Python module with bundled examples, and is classified at TRL 5. We propose to build on the current ITI tool by 1) building new ML models to enhance EO data and translate between all EO instruments run by NASA, ESA and significant commercial partners; 2) building data-loaders and handlers for managing data at scale; and 3) creating excellent documentation and tutorials to lower the usage barriers. The enhanced ITI tool would be classified as TRL 6+, with new capabilities and a more accessible and discoverable interface.

Validation Plan
The validity of this tool in both science domains will be demonstrated through building a benchmarking framework, where users can rapidly determine the statistical similarity of their own ITI pipelines against empirical models using ITI standardized datasets. This can be done for a variety of scenarios allowing for a thorough analysis of the results and

confidence of integration into new ML pipelines. We will build open-source analysis ready datasets that leverage well known assets in NASA's observation portfolio. The framework will enable trained ML models to be compared to empirical models, measuring reduced error and highlighting their superiority at all conditions, altitude, incidence angle and other variables. This benchmarking framework will be designed to be easily operated by external users creating a lively peer-review community to further enable validation.

Maintenance
Post-project support will maintain the capabilities of this tool through expanding the user base through the provision of explanatory films, data snippets and annotated notebooks for easy onboarding on SpaceML.org - an open science portal for AI tools and data. The user community will be supported and expanded with hosting annual meetings and forums to discuss derivatives, improvements and opportunities, and onboarding new researchers. Together, these efforts will deliver a user community aware of the tool, and upskilled in its use and maintenance and able to continue to iterate and build effectiveness

---

Umaa Rebbapragada/Jet Propulsion Laboratory
Anomaly Visualization for Earth and Heliophysics GNSS Data using DORA

Our objective is to enable easy discovery, visualization and analysis of dynamic or transient time variable phenomena from GNSS and other multivariate time series datasets that are ubiquitous in Earth science and heliophysics. We meet our objective by proposing updates to an easy-to-use open-source cross-domain toolkit for anomaly visualization called Domain-agnostic Outlier Ranking Algorithms (DORA) using a heliophysics dataset that is also highly relevant to Earth science. Our proposed use case is the detection of high-latitude ionospheric scintillation events in Global Navigation Satellite Systems (GNSS) data from stations at the Canadian High Arctic Ionospheric Network (CHAIN) that have validation data, as well as solar wind, magnetospheric, and ionospheric data from OMNI Web, hosted by NASA's Space Physics Data Facility. The proposed work builds upon work in progress that adds anomaly detection capabilities for the discovery of seismic and tropospheric events of arbitrary length from Earth Science GNSS data and brings DORA to TRL 5 by the start of this award.

The goals of our proposed work are to bring DORA to TRL 6 by 1) performing exhaustive experiments on CHAIN and OMNI Web data related to our use case to ensure high quality results and cross-domain compatibility on GNSS data, and 2) build visualization capability for multivariate time series into DORA to allow for science analysis and interpretation of results, and 3) submit a publication detailing DORA's detected scintillation events from CHAIN, and a comparison to the current state-of-the-art (McGranaghan et al., 2018). We will push our new capabilities to DORA's open-source repository, including new data readers that will be compatible with CHAIN and OMNI Web data.

This is truly cross-domain work that advances the state-of-the-art in heliophysics and lowers barriers to usage of the latest machine learning software through a single, easy-to-use interface.

Improved, high quality catalogs of scintillation events improve study of how space weather and solar effects impact GNSS signal quality and lead to advances in the science of space weather. The improved understanding of the propagation environment for satellite signals is central to NASA's Space Communications and Navigation program (SCaN) and a key science objective of NASA's Heliophysics Science Division. Given the global distribution of GNSS receivers, diversity of transient/dynamic phenomena in both Earth science and heliophysics, and lack of receiving stations that directly measure ground truth, an unsupervised learning toolkit that lowers barriers to catalog generation will increase the science yield of many NASA Earth and space science data archives.

DORA already meets many of the objectives of this call. It is an easy to use, cross-domain software suite for anomaly visualization that scientists can leverage without writing much code or knowing the state-of-the-art in anomaly detection. DORA has already been prototyped and evaluated with use cases in Earth science, astrophysics, and planetary science, and will be finished prototyping an Earth science GNSS use case by the start of this award. DORA is capable of serving science across the entirety of the NASA Science Mission Directorate.

This is a true collaboration between science and data science. The DORA team is an experienced group of research data scientists with an extensive list of collaborators including members of the Dark Energy Survey, NASA Harvest, Planetary Data Service Imaging Node, and Enhanced Solid Earth Science Earth Science Data Record System. We are bringing our collaborative framework to heliophysics to take an important step toward unifying current work in transient event detection in both Earth-science and heliophysics-focused GNSS systems.

Benoit Tremblay/University Corporation For Atmospheric Research (UCAR)
Leveraging Satellite Observations and Deep Learning to Generate 3D Representations of the Solar Atmosphere and Earth’s Atmosphere

Reconstructing the 3D geometry of objects from 2D images is very challenging. While we have access to a wealth of satellite observations of the Sun and Earth, the number of simultaneous viewpoints and the spatial coverage tend to be limited and the rendering of 3D objects as 2D images through line-of-sight integration results in projection effects and the loss of valuable 3D geometrical information. We propose deep learning to address the need for 3D representations of the Sun and Earth from satellite observations that account for how light travels in the corresponding medium.

We build upon previous work performed for the Sun, motivated by the following. Currently, EUV-observing instruments are limited in their numbers and are constrained to viewing the Sun from its equator (i.e., the ecliptic). For example, the Solar Dynamics

Observatory (SDO; 2010-present) provides images of the Sun in EUV from the perspective of the Earth-Sun line. Two additional viewpoints are provided by the STEREO twin satellites pulling Ahead (STEREO-A; 2006-present) and falling Behind (STEREO-B; 2006-2014) of Earth's orbit. While Solar Orbiter while get close (2020-present), no satellites observe the solar poles directly. However, a complete image of the 3D Sun is required fully understand the dynamics of the Sun (from eruptive events to the resultant space weather in the solar system), to forecast EUV radiation to protect our assets in space, to relate the Sun to other stars in the universe, and to generalize our knowledge of the Sun-Earth system to other host stars.

To maximize the science return of multiple viewpoints, we developed a novel approach that unifies and smoothly integrates data from multiple perspectives into a consistent 3D representation of the solar corona. We leveraged Neural Radiance Fields (NeRFs) which are neural networks that achieve state-of-the-art 3D scene representation and generate novel views from a limited number of input images. We adapted a Sun NeRF (SuNeRF) to generate a physically-consistent representation of the 3D Sun, with the inclusion of radiative transfer and geometric ray sampling that matches the physical reality of optically thin plasma in the solar atmosphere. We used simulation data to validate the SuNeRF's ability to provide new state-of-the-art results in 3D representations of the full Sun, including viewing the poles, despite learning only from ecliptic viewpoints. Using SuNeRFs, the heliophysics community will be able to study the Sun from completely unprecedented perspectives.

With this call, we propose to train SuNeRF models using EUV observations captured from multiple viewpoints (SDO, STEREO-A, STEREO-B, and Solar Orbiter) in order to then generate novel views beyond what existing satellites can achieve. The proposed work would build and expand upon existing prototypes that the team has developed during the Frontier Development Lab. Additionally, we propose to streamline our framework to expand applications to new physical domains. As a proof of concept, we focus on the treatment of cloud swaths captured by the MODIS instrument and developing SuNeRFs modules to address the physical reality of the problem. In the end, we will provide an easily-customizable and reproducible pipeline for the heliophysics and geophysics communities.

SuNeRFs leverage existing multi-viewpoint observations (e.g., constellation missions) and act as virtual instruments that can fly out and be placed anywhere at no additional cost. Our method is also an example of how novel deep learning techniques can be used to significantly enhance observational capabilities by the creation of virtual instruments. These virtual instruments need not be limited to the domain of a single NASA SMD division, but could potentially combine stellar, planetary, and heliophysics observational platforms.

Chaowei Yang/George Mason University

Transitioning a Training Dataset Labeling Tool (TDLT) to Support Discoveries in Earth Science and Heliophysics

Abstract: We propose to create a generalizable training dataset labeling tool for both Earth and Heliophysics by improving upon a prototype tool designed for Earth science. Developed by the NSF Spatiotemporal I/UCRC at George Mason University (GMU) in collaboration with NASA Goddard CISTO and partially funded by AIST 2021, an automatic Sea Ice labeling tool will be extended to label to identify coronal holes using Extreme Ultraviolet (EUV) solar imagery. The project will be implemented by an interdisciplinary team of experts in AI/ML data labeling for Earth science (PI Yang), Heliophysics image processing and discovery (CoIs Zhang and Kirk) and cross-domain computing operation (CoI Duffy). The end result will be the basis of a foundational training dataset labeling tool that can be readily deployed and rapidly trained across multiple science domains via the following steps:

- Analyzing polar region sea ice images for climate change and detecting coronal holes in EUV images obtained by Atmospheric Imaging Assembly (AIA) instrument onboard SDO (Solar Dynamic Observatory) for heliophysics science.
- Based on previous NASA and NSF project results, integrating, and interfacing the data labeling system to include the functionality of the open source including a) image loading, b) automatic initial labeling, and c) manual adjustment of the labeling.
- Enhancing the integrated system by a) further improving the accuracy and data management of the training datasets labeling process, b) adjusting the interface to fit in heliophysics user requirements, c) integrating the training labeling process with CISTO computing infrastructure in a cloud computing fashion, d) minimizing the system configuration needed for the Earth and heliophysics domains, e) creating Docker/container solution to enable rapid deployment.
- Collaborating with NASA Goddard CISTO user communities to a) capture the user requirements, b) analyze and recommend user interface adjustment, c) integrate the training dataset labeling into CISTO tool offerings to Goddard earth and heliophysics scientists, and d) adjust the tool according to user feedback for long term operation support.
- Interfacing the data labeling system with the open-source AI/ML software suite at CISTO to support both domains and beyond.
- Providing the system as an open-source solution (with Apache 2.0 license) to the NASA Earth and heliophysics communities as an advanced training dataset labeling capability solution.

The developed capabilities will provide a translational capability to the training dataset labeling for Earth and heliophysics domains, increasing the efficiency of data exploration and decreasing the time that scientists must spend on searching, creating, and obtaining the training datasets most applicable to their AI/ML-based scientific research. The system will also be maintained at NASA Goddard CISTO to create an environment using JupyterHub in the cloud where anyone can register to get access to the data, training sets, and sample notebooks for open science and support multiple domains adopting AI/ML with readily available training datasets and management system.

Keywords: Machine Learning, Heliophysics, Coronal Hole, Sea Ice, Climate Change, Training Dataset

TRL: The current tool has been prototyped and initially tested by Earth scientists at a TRL 5 and will be expanded to support Heliophysics and offered to both domain scientists at a TRL 6.