The National Data Platform (NDP): Democratizing Data and Responsible Artificial Intelligence

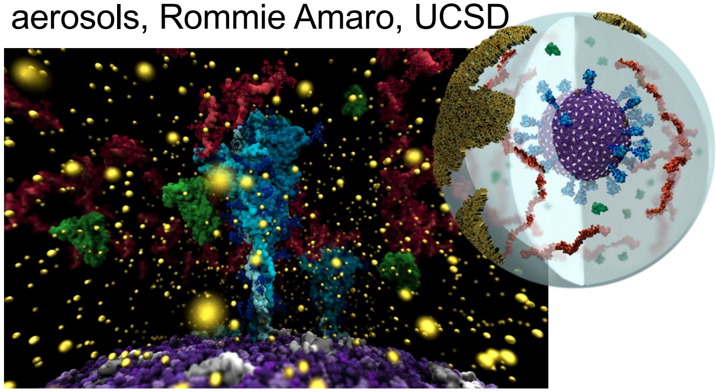
Manish Parashar

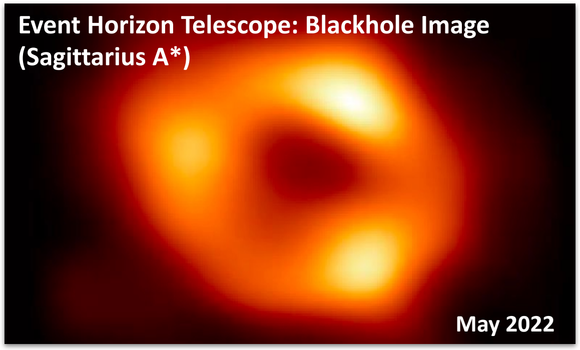
Director, Scientific Computing & Imaging (SCI) Institute
Chair in Computational Science and Engineering
Presidential Professor, Kahlert School of Computing

6NRP
San Diego, CA
January 30, 2025

Science / Society Transformed by Data, Cyberinfrastructure

Modeling of the delta virus inside respiratory aerosols, Rommie Amaro, UCSD





Event Horizon Telescope: Blackhole Image (Sagittarius A*)

May 2022

Cyberinfrastructure is a key enabler of discoveries & innovations

Data-driven Urgent Science

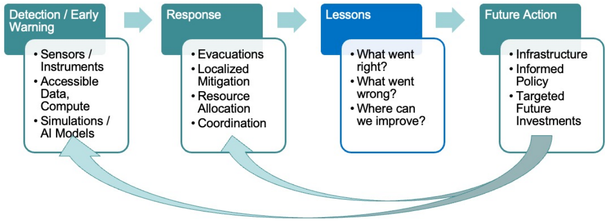
Earthquakes & Tsunamis

Extreme Weather

Cyber-Attacks

Pandemics

Industrial Disasters



Detection / Early Warning
- Sensors / Instruments
- Accessible Data, Compute
- Simulations / AI Models

Response
- Evacuations
- Localized Mitigation
- Resource Allocation
- Coordination

Lessons
- What went right?
- What went wrong?
- Where can we improve?

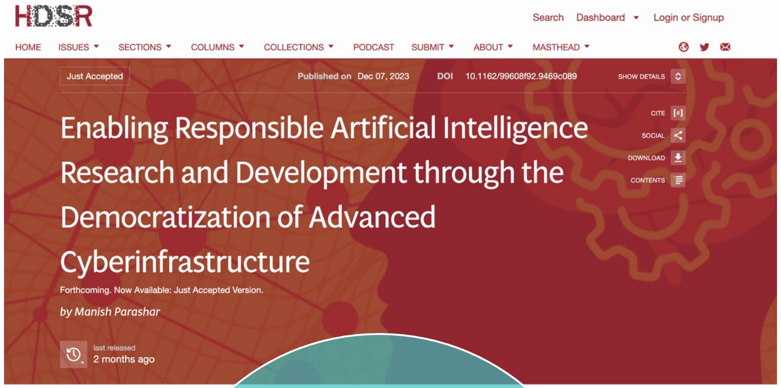
Future Action
- Infrastructure
- Informed Policy
- Targeted Future Investments

The Transcendence of AI

THE LEARNING HEALTH SYSTEM SERIES
Artificial Intelligence in Health Care
The Hope, the Hype, the Promise, the Peril
Michael Matheny, Sonoo Thadaney Israni, Mahnoor Ahmed, and Danielle Whicher, Editors
NATIONAL ACADE
Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem
An Implementation Plan for a National Artificial Intelligence Research Resource
January 2023
CSIRO
Australia's National Science Agency
Artificial intelligence for science
Chemical shifts from tiny NMR samples pp. 38 & 67
Regulating products that target gut microbiomes p. 39
Preschool games promote math skills in India p. 47
Science
$15 7 JULY 2017 sciencemag.org
AAAS
AI
SFORMS
IENCE
p. 16
Artificial Intelligence in Science
CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH
OECD
ANL-22/91
ADVANCED RESEARCH DIRECTIONS ON
AI FOR SCIENCE, ENERGY, AND SECURITY
Report on Summer 2022 Workshops
Jonathan Carter
Lawrence Berkeley National Laboratory
John Feddema
Sandia National Laboratories
Doug Kothe
Oak Ridge National Laboratory
Rob Neely
Lawrence Livermore National Laboratory
Jason Pruet
Los Alamos National Laboratory
Rick Stevens
Argonne National Laboratory
ENERGY
ENERGY
Office of Science
NNSA
May 2023
CRC Press
Integration of AI-Based Manufacturing and Industrial Engineering Systems with the Internet of Things
Edited by
Pankaj Bhambri, Sita Rani, Valentina E. Balas, and Ahmed A. Elngar

Democratizing Responsible Data, AI is important

- Key characteristics for responsible AI (NIST) *validity, safety security, accountability, privacy enhancement, fairness, and explainability.*
- The quality and impact of research and the pace of innovation are linked to the diversity of the contributions.
- Especially true for AI-enabled research
  - In case of AI, quality depends on who is developing and use AI, and where the data coming from
  - Greater inclusivity in contribution to research and development increases the diversity of approaches and the fairness of the results.
- Many barriers: awareness, ability, access, association, …



HDSR

Just Accepted

Published on Dec 07, 2023

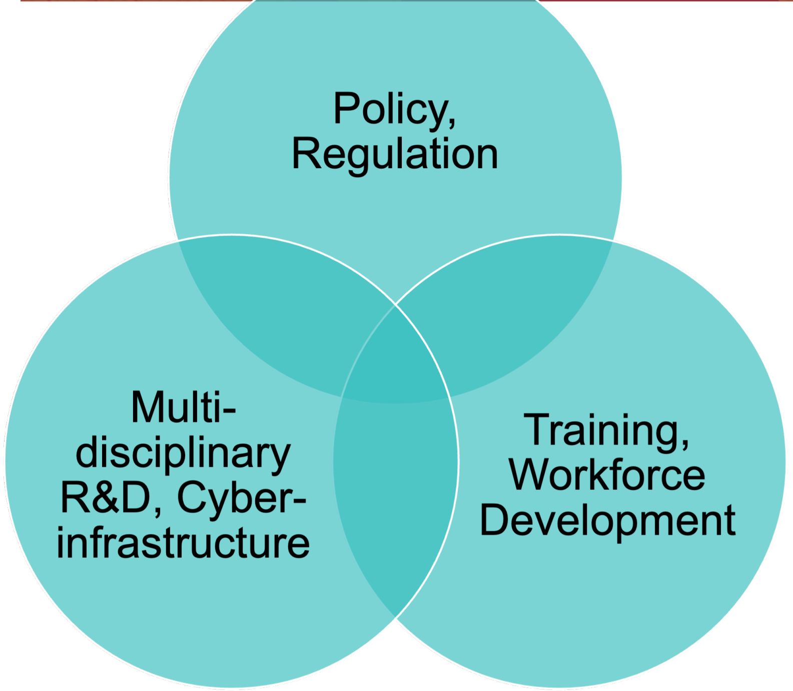
DOI 10.1162/99608f92.9469c089

Enabling Responsible Artificial Intelligence Research and Development through the Democratization of Advanced Cyberinfrastructure

Forthcoming. Now Available: Just Accepted Version.

by Manish Parashar

last released
2 months ago



Policy, Regulation
Multi-disciplinary R&D, Cyber-infrastructure
Training, Workforce Development

NAIRR: Democratizing the AI R&D Ecosystem

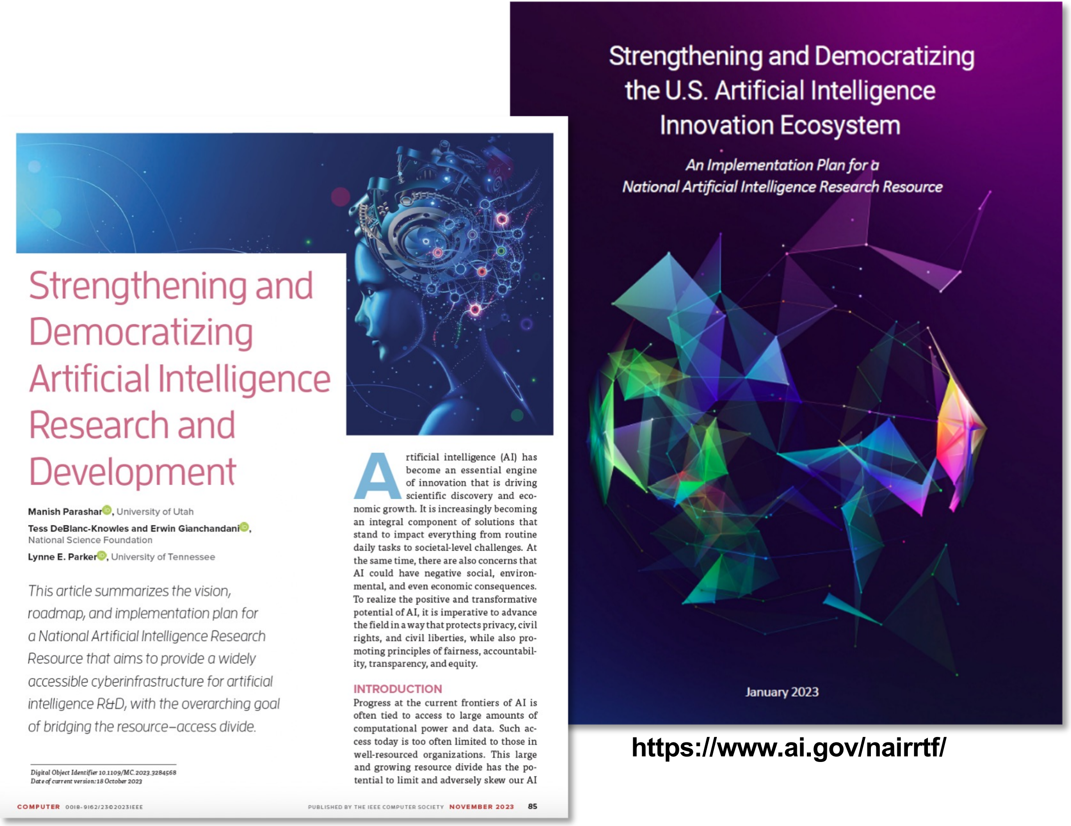
Goals: Strengthen and democratize the U.S. AI Innovation ecosystem in a way that protects privacy, civil rights, and civil liberties.

Spur **innovation**

Increase the **diversity** of talent in AI

Improve U.S. **capacity** for AI R&D

Advance **trustworthy AI**



Strengthening and Democratizing Artificial Intelligence Research and Development

Manish Parashar, University of Utah
Tess DeBlanc-Knowles and Erwin Gianchandani, National Science Foundation
Lynne E. Parker, University of Tennessee

This article summarizes the vision, roadmap, and implementation plan for a National Artificial Intelligence Research Resource that aims to provide a widely accessible cyberinfrastructure for artificial intelligence R&D, with the overarching goal of bridging the resource–access divide.

Artificial intelligence (AI) has become an essential engine of innovation that is driving scientific discovery and economic growth. It is increasingly becoming an integral component of solutions that stand to impact everything from routine daily tasks to societal-level challenges. At the same time, there are also concerns that AI could have negative social, environmental, and even economic consequences. To realize the positive and transformative potential of AI, it is imperative to advance the field in a way that protects privacy, civil rights, and civil liberties, while also promoting principles of fairness, accountability, transparency, and equity.

INTRODUCTION

Progress at the current frontiers of AI is often tied to access to large amounts of computational power and data. Such access today is too often limited to those in well-resourced organizations. This large and growing resource divide has the potential to limit and adversely skew our AI

Digital Object Identifier 10.1109/MC.2023.3284568
Date of current version: 18 October 2023

COMPUTER 0018-9162/23©2023IEEE PUBLISHED BY THE IEEE COMPUTER SOCIETY NOVEMBER 2023 85

Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem

An Implementation Plan for a National Artificial Intelligence Research Resource

January 2023

Computer, vol. 56, no. 11, pp. 85-90, Nov. 2023, doi: 10.1109/MC.2023.3284568.

https://www.ai.gov/nairrtf/

NATIONAL DATA PLATFORM

Bridging the Data Gaps for AI

National Data Platform - NDP
Services for Equitable Open Access to Data

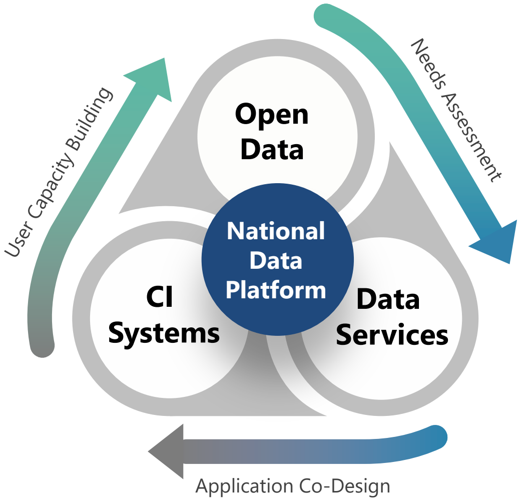
#2333609

A **federated** and **extensible** data ecosystem to promote innovation and collaboration through the equitable use of data leveraging existing and future national cyberinfrastructure capabilities.

https://www.nationaldataplatform.org/

FOCUS AREAS:

- **Platform** for data-enabled and AI-integrated workflows
  - Facilitates data registration and discovery via a **centralized hub**
  - Democratizes data access and use via **distributed points of presence**
  - Cultivates resources for **classroom education** and **data challenges**
  - Assists research and learning through **personalized workspaces**

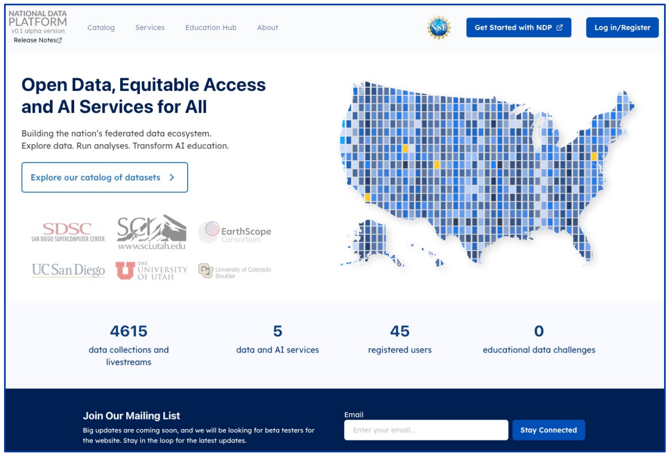


- **Applications** in climate and AI with data diverse scientific data repositories including NSF facilities, NAIRR, NASA, USGS, NOAA and USDA

- **Partnerships** to foster scientific discovery, decision-making, policy formation and societal impact

UC San Diego | THE UNIVERSITY OF UTAH | University of Colorado Boulder | SAN DIEGO SUPERCOMPUTER CENTER | SCI www.sci.utah.edu | EarthScope Consortium

Centralized portal for discovery through collaborative workspaces for research and education



Open Data, Equitable Access and AI Services for All

Building the nation's federated data ecosystem.
Explore data. Run analyses. Transform AI education.

4615	5	45	0
data collections and livestreams	data and AI services	registered users	educational data challenges

Join Our Mailing List

Big updates are coming soon, and we will be looking for beta testers for the website. Stay in the loop for the latest updates.

NATIONAL DATA PLATFORM

A scalable **platform** for developing and deploying services at **distributed points of presence**

SDSC SAN DIEGO SUPERCOMPUTER CENTER

SCI www.sci.utah.edu

UTAH U

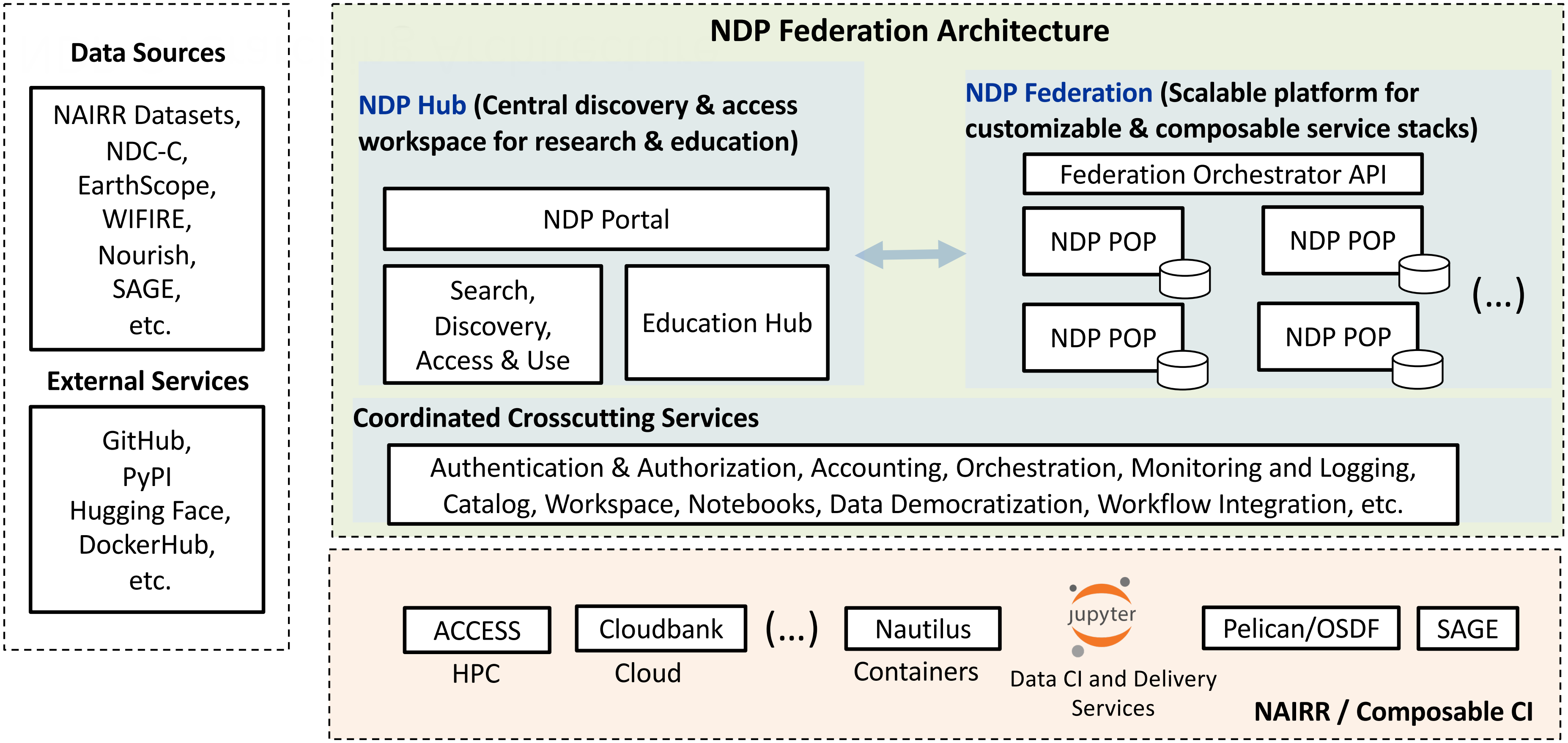
University of Colorado Boulder

EarthScope Consortium

http://www.nationaldataplatform.org

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

NDP Overarching Architecture



Data Sources

NAIRR Datasets, NDC-C, EarthScope, WIFIRE, Nourish, SAGE, etc.

External Services

GitHub, PyPI Hugging Face, DockerHub, etc.

NDP Federation Architecture

NDP Hub (Central discovery & access workspace for research & education)

NDP Portal

Search, Discovery, Access & Use

Education Hub

NDP Federation (Scalable platform for customizable & composable service stacks)

Federation Orchestrator API

NDP POP

NDP POP

NDP POP

NDP POP

(…)

Coordinated Crosscutting Services

Authentication & Authorization, Accounting, Orchestration, Monitoring and Logging, Catalog, Workspace, Notebooks, Data Democratization, Workflow Integration, etc.

ACCESS

HPC

Cloudbank

Cloud

(…)

Nautilus

Containers

jupyter

Data CI and Delivery Services

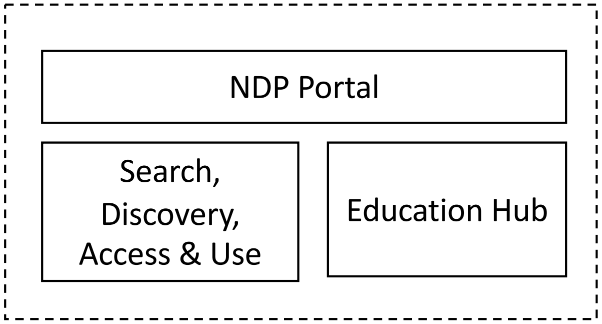
Pelican/OSDF

SAGE

NAIRR / Composable CI

NDP Hub: Central discovery & access workspace for research & education

NDP Hub



- NDP Portal (point of access)
  https://nationaldataplatform.org
- Metadata registration and indexing
  - Contributing organizations
  - Harvested metadata from NDP POPs
- Data search
  - String and conceptual search
  - Open Knowledge graphs / via LLMs

NDP Standard Services

<u>Public</u>:

- Extensible Data Catalog and Search Services
- Education Hub Informal Learning Modules

<u>Login-enabled</u>:

- Keycloak Role-Based Access Service
- User Workspaces
- AI Gateway with Custom JupyterHub Service
- Data Catalog and OKN Ingestion
- External Model Ingestion
- Data Exploration Services
- MLFlow Dashboard Service
- Education Hub Classroom
- Education Hub Challenge
- Democratizing Data Dashboard

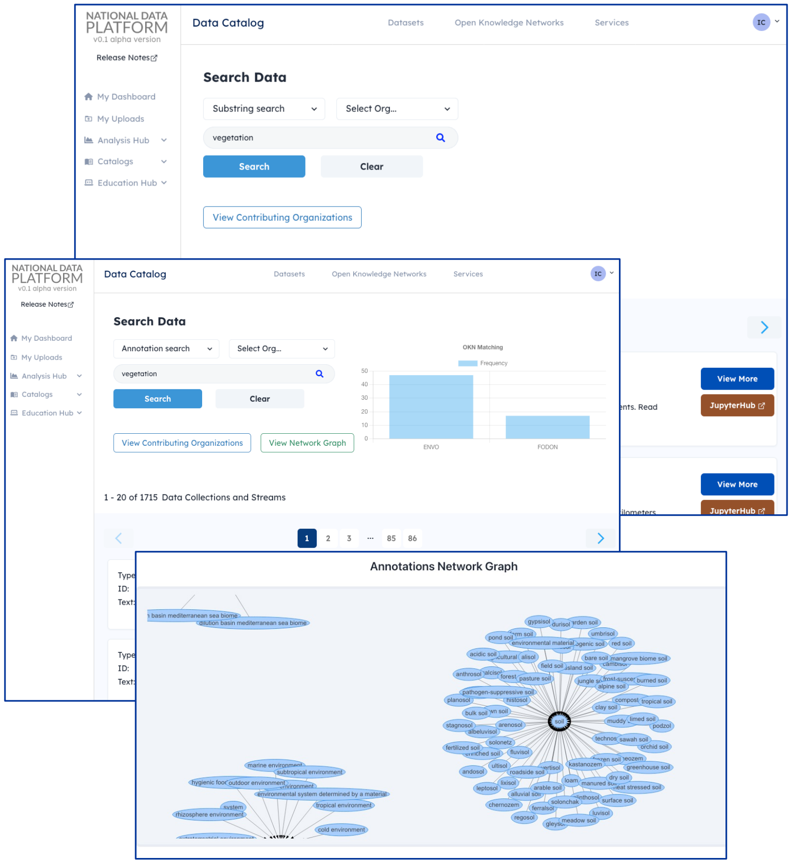
Hub Capabilities Under Development

- Sage Data and Edge Code Integration Service
- Service Catalog and Discovery Service
- Educational Hub Expansion
- Streaming Data Services
- Pelican Registration Service
- Integrated Workflows

Planned Future Work

- OKN Integration
- Data Curation
- Data Subsetting
- Data Provenance
- Educational Toolkits
- Open Science Chain Provenance Service
- Gateway Services

NDP Hub: Data Search and Discovery



Substring search

vegetation

Annotation search

vegetation

1 - 20 of 1715 Data Collections and Streams

Current Capabilities:

- Search capabilities to include not just text in metadata and ontology concepts but also time and location data.
- Ability to search time and time ranges within the data, such as from "27 September 2020" to "24 January 2021."
- Location-based searches can now be combined using specific location names (e.g., "San Luis Obispo") or boundary polygons.
- Support free-text search across "all metadata" without specifying particular fields.
- Utilize Lucene, a popular search syntax, to improve search functionality.

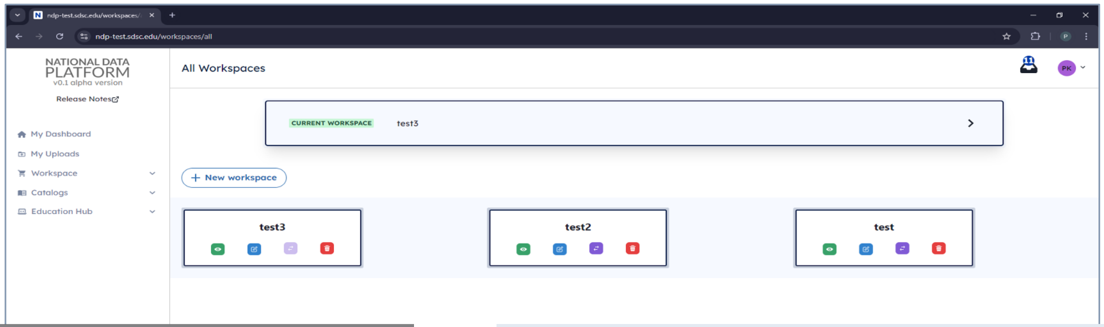
Ongoing Work :

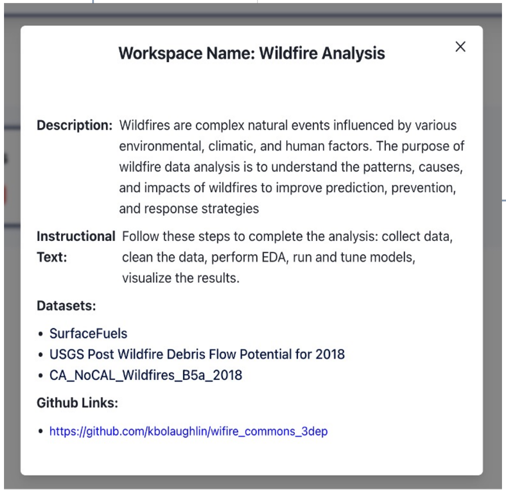
- Extract entity annotations from the metadata text and integrate them with the ontology to enhance search functionality.
- Create a vector store and develop a search pipeline that handles queries in natural language.
- Optimize the system's performance to ensure fast and accurate retrieval of relevant information.

NDP Workspaces (Version 1 – September 2024)

Goal : Craft persistent and customizable workspaces with datasets and services to launch into a sandbox

- Create customized workspaces for varied use cases
- Search and add datasets to use in sandbox (HPC Env)
- Add github links for file access
- Launch packaged workspace into sandbox





Workspace Name: Wildfire Analysis

Description: Wildfires are complex natural events influenced by various environmental, climatic, and human factors. The purpose of wildfire data analysis is to understand the patterns, causes, and impacts of wildfires to improve prediction, prevention, and response strategies

Instructional Text: Follow these steps to complete the analysis: collect data, clean the data, perform EDA, run and tune models, visualize the results.

Datasets:

- SurfaceFuels
- USGS Post Wildfire Debris Flow Potential for 2018
- CA_NoCAL_Wildfires_B5a_2018

Github Links:

- https://github.com/kbolaughlin/wifire_commons_3dep

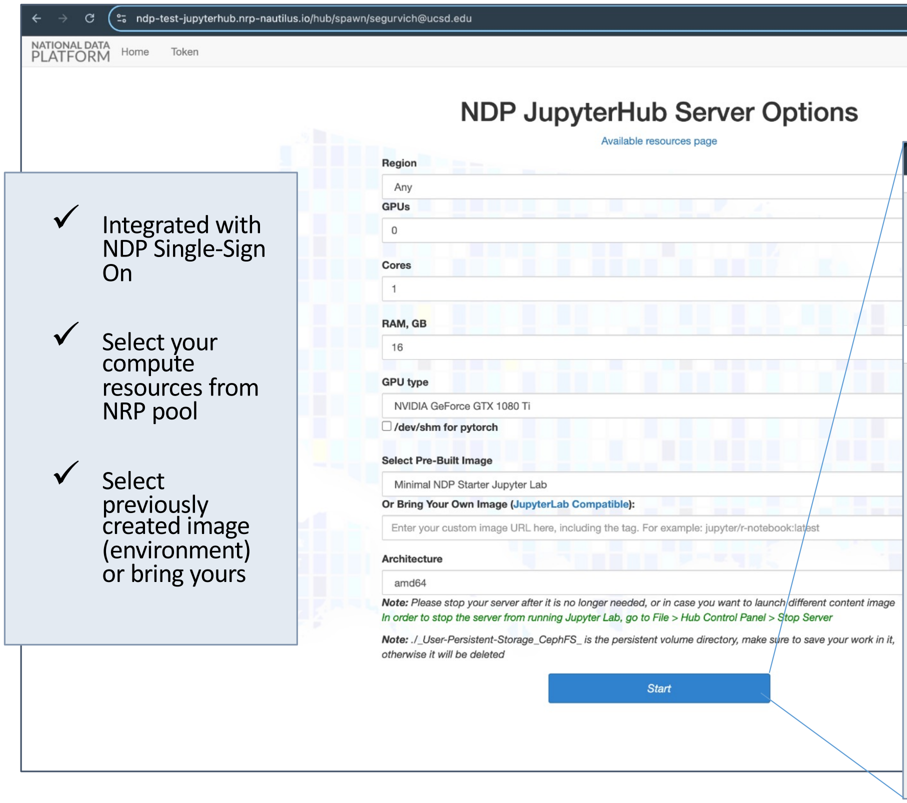
Users can:

- view all their workspaces
- create new workspaces by clicking on the “New Workspace” button
- use workspace action buttons to preview, edit, switch and delete
- add datasets to their current workspace from the catalog page.

NDP JupyterHub (Sandbox)

A compute environment for data analysis, machine learning training or any other computational tasks, built on top of NRP (Nautilus) cluster. Different datasets and tasks will require powerful compute resources (CPUs, GPUs, memory), which user can select and use seamlessly.

- ✓ Integrated with NDP Single-Sign On
- ✓ Select your compute resources from NRP pool
- ✓ Select previously created image (environment) or bring yours


ndp-test-jupyterhub.nrp-nautilus.io/hub/spawn/segurvich@ucsd.edu
NATIONAL DATA PLATFORM
Home
Token

NDP JupyterHub Server Options

Available resources page

Region: Any

GPUs: 0

Cores: 1

RAM, GB: 16

GPU type: NVIDIA GeForce GTX 1080 Ti

☐ /dev/shm for pytorch

Select Pre-Built Image: Minimal NDP Starter Jupyter Lab

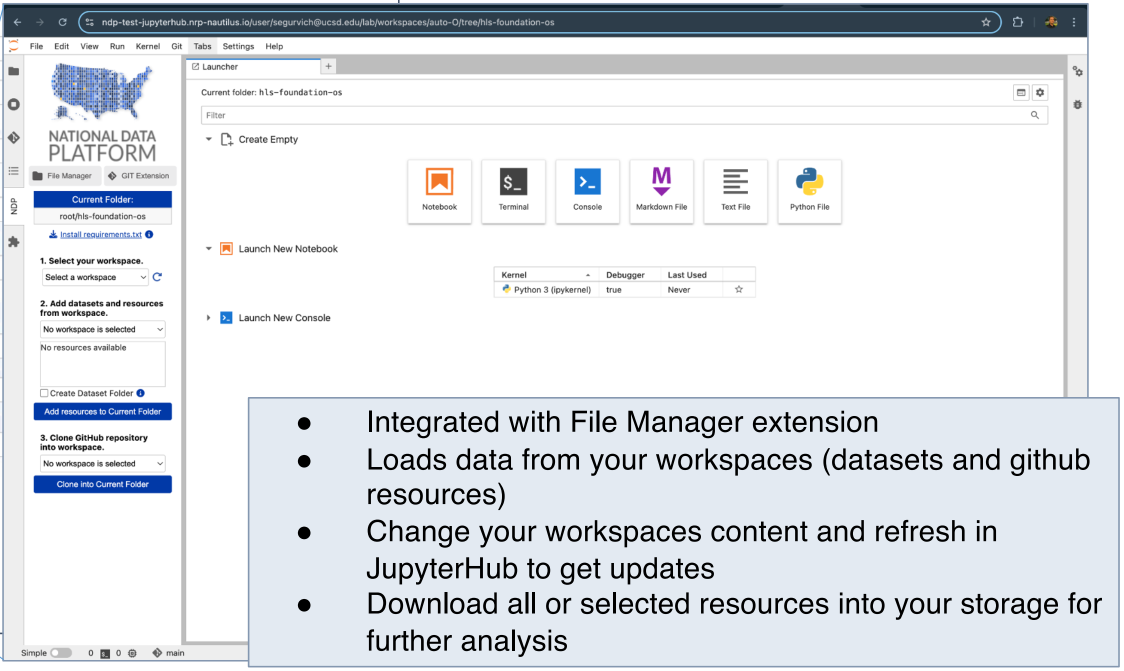
Or Bring Your Own Image (JupyterLab Compatible): Enter your custom image URL here, including the tag. For example: jupyter/r-notebook:latest

Architecture: amd64

Note: Please stop your server after it is no longer needed, or in case you want to launch different content image

In order to stop the server from running Jupyter Lab, go to File > Hub Control Panel > Stop Server

Note: ./_User-Persistent-Storage_CephFS_ is the persistent volume directory, make sure to save your work in it, otherwise it will be deleted



Current Folder: root/hls-foundation-os

Install requirements.txt

1. Select your workspace. Select a workspace

2. Add datasets and resources from workspace. No workspace is selected

No resources available

☐ Create Dataset Folder

Add resources to Current Folder

3. Clone GitHub repository into workspace. No workspace is selected

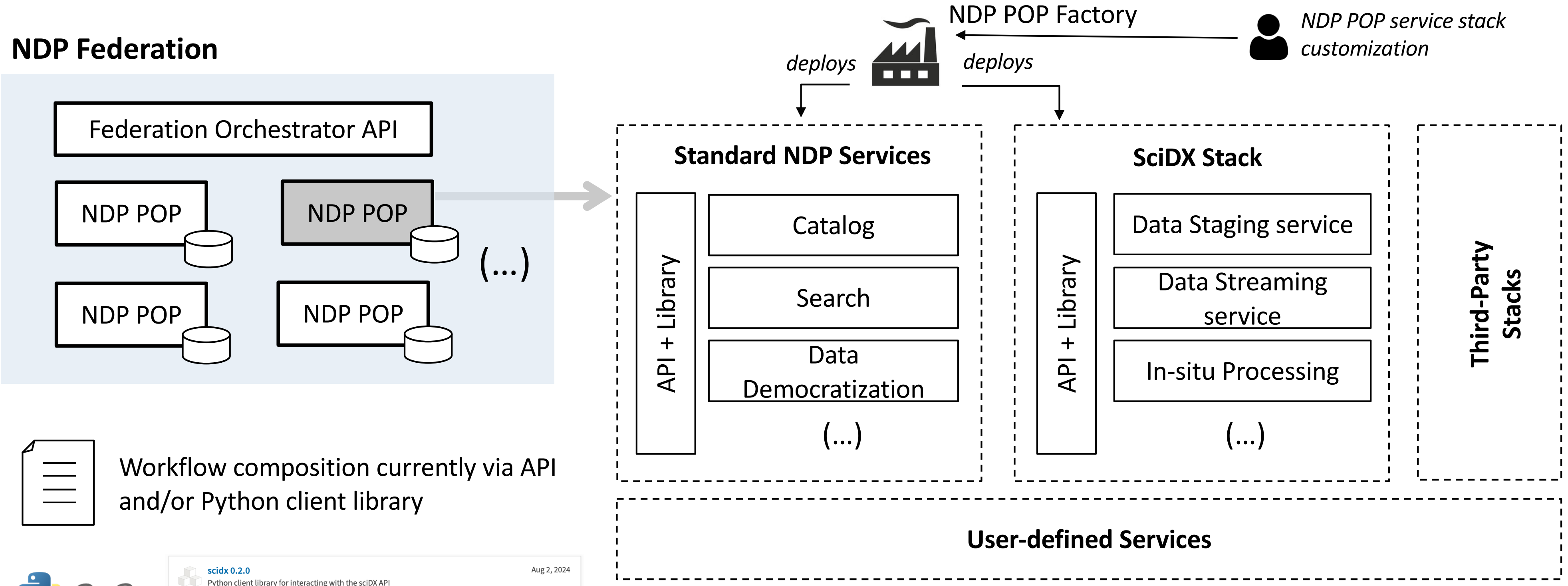
Clone into Current Folder

Current folder: hls-foundation-os

Kernel	Debugger	Last Used	
Python 3 (ipykernel)	true	Never	☆

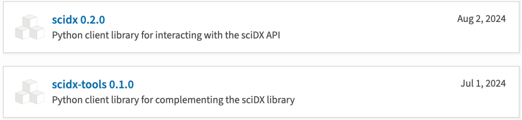
- Integrated with File Manager extension
- Loads data from your workspaces (datasets and github resources)
- Change your workspaces content and refresh in JupyterHub to get updates
- Download all or selected resources into your storage for further analysis

NDP Data POP: Distributed Points of Presence with Customizable, Composable Service Stacks



Workflow composition currently via API and/or Python client library

pypi


scidx 0.2.0 Aug 2, 2024
Python client library for interacting with the sciDX API

scidx-tools 0.1.0 Jul 1, 2024
Python client library for complementing the sciDX library

Deployment models: docker kubernetes Cloud HPC

-- Standalone -- *-- Scalable (cluster) --*

Science Data Exchanges (SciDx) Services

A customizable Data-Pop software stack for in-situ data access & processing

SciDx Science Data Exchange

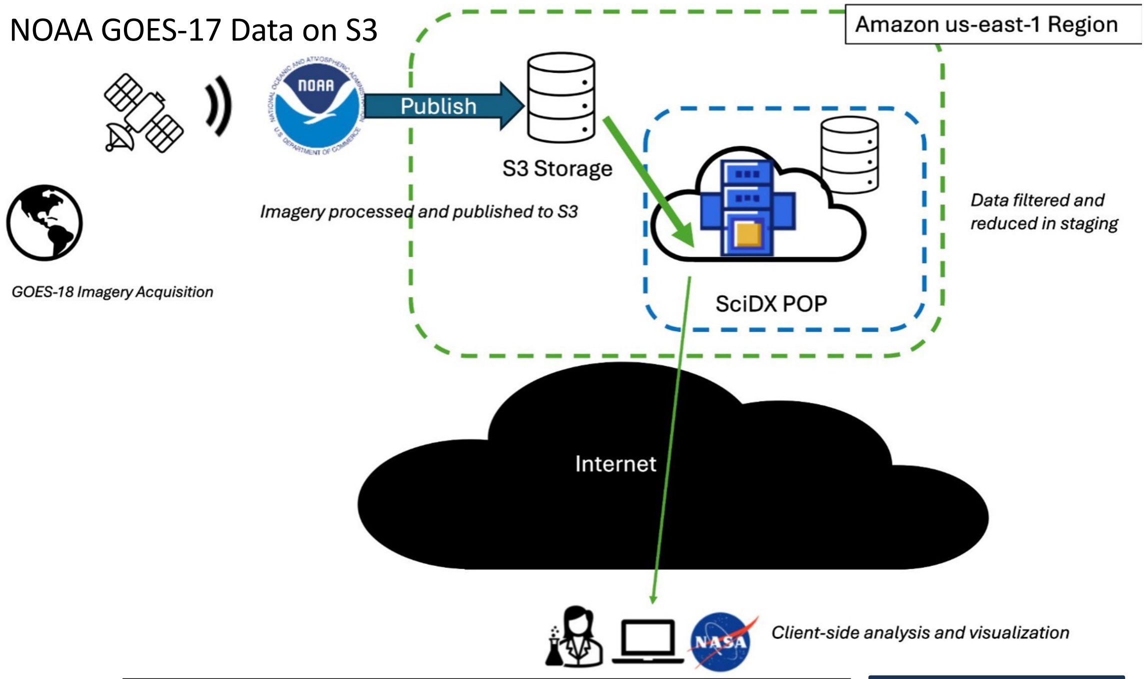
SciDx Staging Services

- Transient resources for in-situ (close to the data) data processing and access
  - High-performance in-memory processing
  - Server-side data transformations (e.g., sub-setting, reduction, user-defined analysis, etc.)
  - Caching/sharing of data, results, and data-products
  - Registration of data-triggers
- Efficient management of data in-motion
  - Streamline workflows; minimize data transfers
  - Perform ETL operations at data source

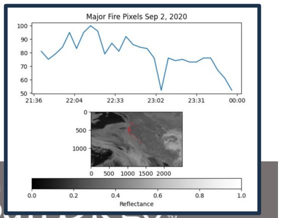
SciDx Staging Service: Wildfire Monitoring Usecase

SciDx Science Data Exchange

- Monitor fire hotspots based on satellite data that updates every 5 minutes
- Not interested in the entire data product, just pixels that reach severity threshold
- Per-pixel evaluation as a user-defined transformation is performed on each new data update
- The user subscribes to the results of the transformation
- Reduction in data cost, latency, time to solution


NOAA GOES-17 Data on S3
Amazon us-east-1 Region
NOAA
Publish
S3 Storage
Imagery processed and published to S3
GOES-18 Imagery Acquisition
Data filtered and reduced in staging
SciDX POP
Internet
NASA
Client-side analysis and visualization

```
result =
client.query_array(source='goes18-radc',
                   var_name='Rad',
                   lb=(0,2500),
                   ub=(2499,4999),
                   timestamp='2024-08-02T00:35:00',
                   time_direction=PAST)
```


Major Fire Pixels Sep 2, 2020

Science Data Exchanges (SciDx) Services

A customizable Data-Pop software stack for in-situ data access & processing

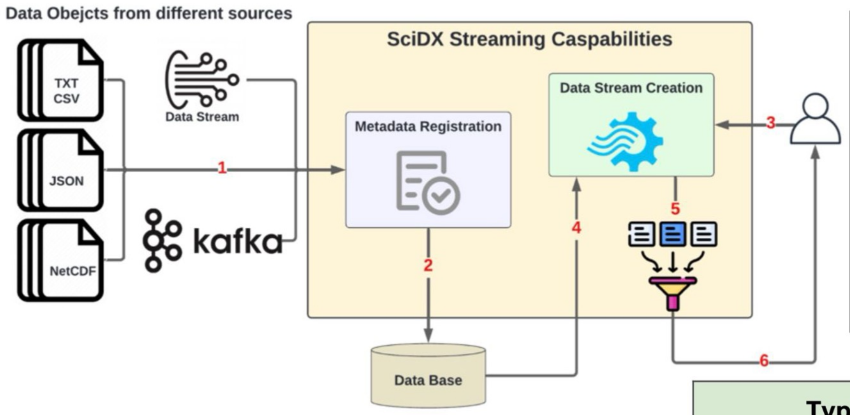
SciDx Staging Services

- Transient resources for in-situ (close to the data) data processing and access
  - High-performance in-memory processing
  - Server-side data transformations (e.g., sub-setting, reduction, user-defined analysis, etc.)
  - Caching/sharing of data, results, and data-products
  - Registration of data-triggers
- Efficient management of data in-motion
  - Streamline workflows; minimize data transfers
  - Perform ETL operations at data source

SciDx Streaming Service

- Streams registration, curation/archival for discovery and access
- User-defined operations/filters on streaming; containerized execution
- Combine streaming data with archived/playback data
- Mechanism for online data product generation (i.e., new data streams)

SciDx Streaming Service



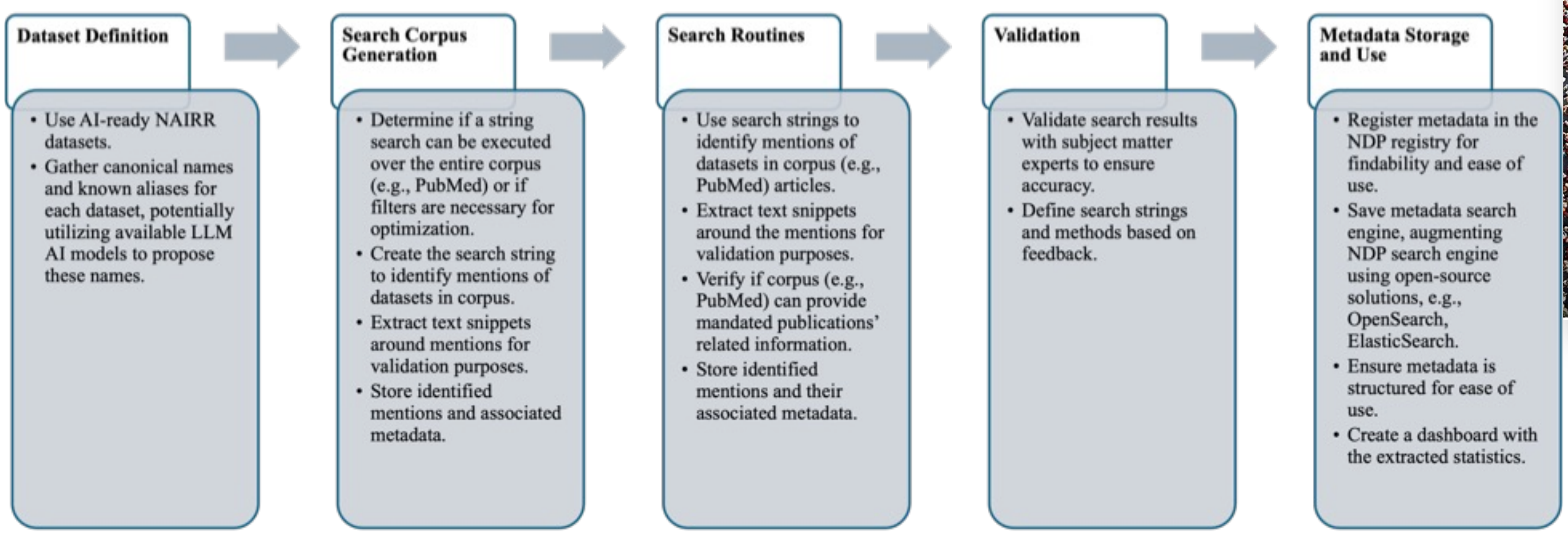
1. Register metadata for various Data Objects
2. Store metadata
3. User queries specific data (e.g., ‘Temp’)
4. Retrieve data objects matching user criteria
5. Apply user-defined filters (e.g., ‘a>10’)
6. Deliver unified, filtered data as a stream

Type	Explanation	Example
Column Comparisons	Column-to-column comparisons	`x > y`
Mathematical Operations	Addition, subtraction, multiplication and division	`x > 10*y`
IN Operator	Check if values are in a list	`station IN ['A', 'B']`
Conditional Logic (IF-THEN-ELSE)	Apply rules based on conditional statements	`IF x > 20 THEN alert = High ELSE y = 10`
Logical Operators (AND, OR)	Combine multiple conditions using AND and OR operators	`IF x > 10 OR z = 20 THEN alert = High ELSE alert = Low`
Window-Based Filtering	Calculate aggregates (mean, sum, max, min) over sliding windows	`IF window_filter(9, sum, x > 20) THEN alert = High`

SciDx: Advanced Search & Discovery

democratizing data

https://democratizingdata.ai/



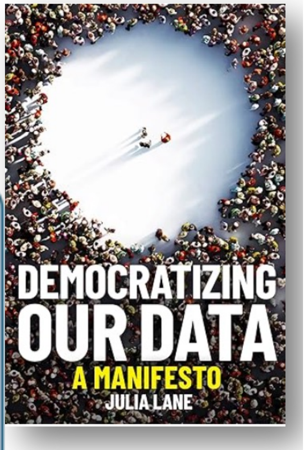
Dataset Definition
- Use AI-ready NAIRR datasets.
- Gather canonical names and known aliases for each dataset, potentially utilizing available LLM AI models to propose these names.

Search Corpus Generation
- Determine if a string search can be executed over the entire corpus (e.g., PubMed) or if filters are necessary for optimization.
- Create the search string to identify mentions of datasets in corpus.
- Extract text snippets around mentions for validation purposes.
- Store identified mentions and associated metadata.

Search Routines
- Use search strings to identify mentions of datasets in corpus (e.g., PubMed) articles.
- Extract text snippets around the mentions for validation purposes.
- Verify if corpus (e.g., PubMed) can provide mandated publications' related information.
- Store identified mentions and their associated metadata.

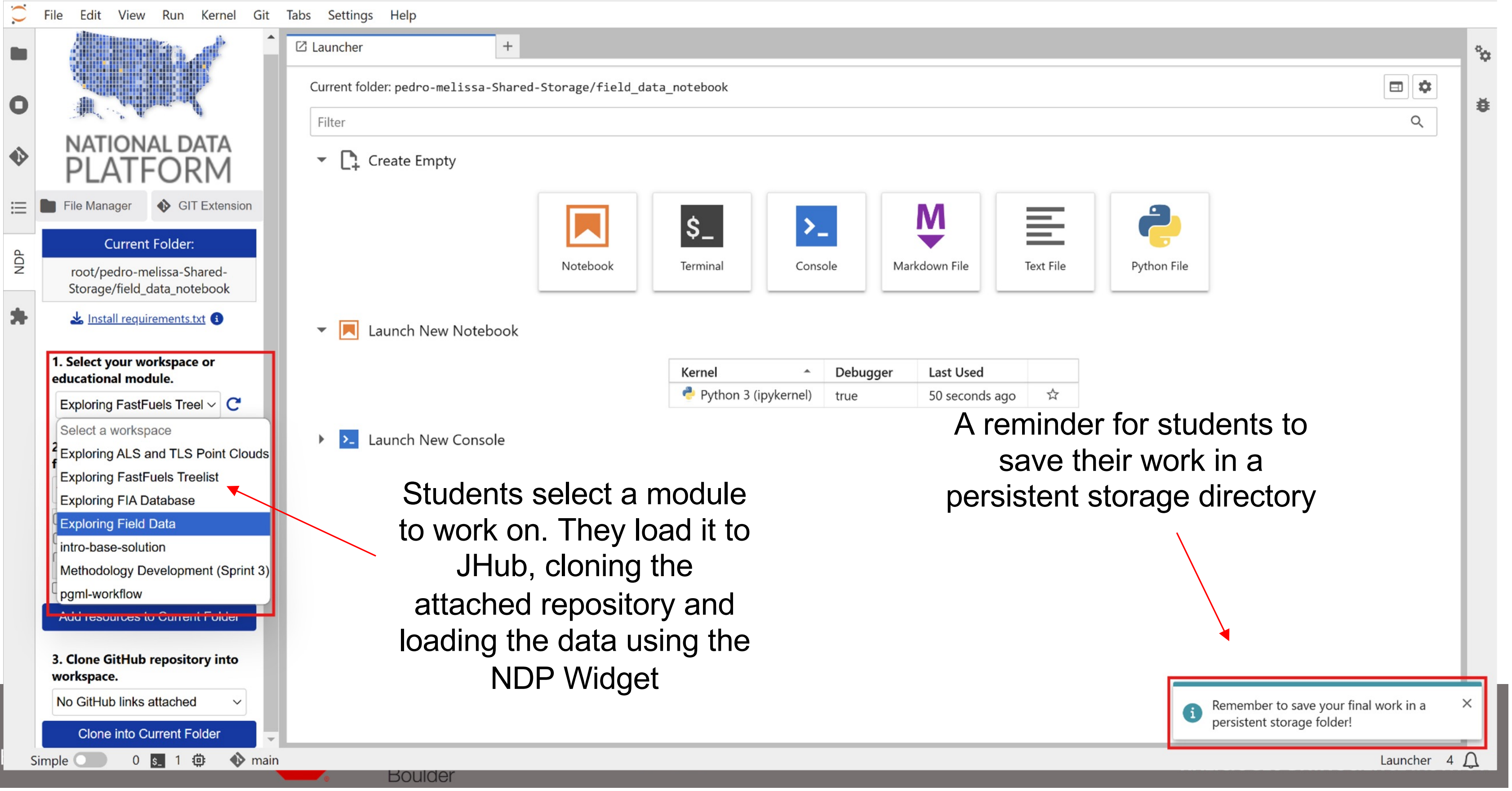
Validation
- Validate search results with subject matter experts to ensure accuracy.
- Define search strings and methods based on feedback.

Metadata Storage and Use
- Register metadata in the NDP registry for findability and ease of use.
- Save metadata search engine, augmenting NDP search engine using open-source solutions, e.g., OpenSearch, ElasticSearch.
- Ensure metadata is structured for ease of use.
- Create a dashboard with the extracted statistics.



SDSC SAN DIEGO SUPERCOMPUTER CENTER · SCI www.sci.utah.edu · UTAH · University of Colorado Boulder · EarthScope Consortium · http://www.nationaldataplatform.org · UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

NDP+NRP: Use the NDP widget to import datasets and conduct analysis within NRP.



Current Folder:

root/pedro-melissa-Shared-Storage/field_data_notebook

Install requirements.txt

1. Select your workspace or educational module.

Exploring FastFuels Treel

- Select a workspace
- Exploring ALS and TLS Point Clouds
- Exploring FastFuels Treelist
- Exploring FIA Database
- Exploring Field Data
- intro-base-solution
- Methodology Development (Sprint 3)
- pgml-workflow

3. Clone GitHub repository into workspace.

No GitHub links attached

Current folder: pedro-melissa-Shared-Storage/field_data_notebook

Filter

Kernel	Debugger	Last Used
Python 3 (ipykernel)	true	50 seconds ago

Remember to save your final work in a persistent storage folder!

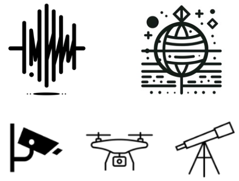
Students select a module to work on. They load it to JHub, cloning the attached repository and loading the data using the NDP Widget

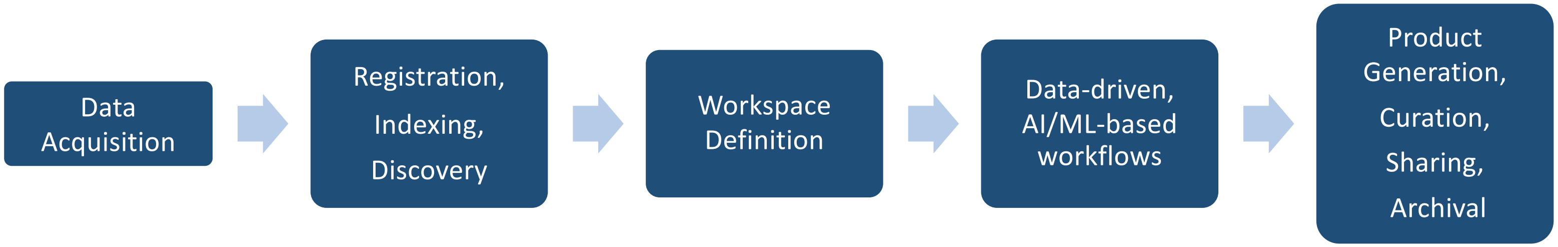
A reminder for students to save their work in a persistent storage directory

SDSC SAN DIEGO SUPERCOMPUTER C

Boulder

Example NDP-NAIRR AI in Science Workflow




Data Acquisition
Registration, Indexing, Discovery
Workspace Definition
Data-driven, AI/ML-based workflows
Product Generation, Curation, Sharing, Archival

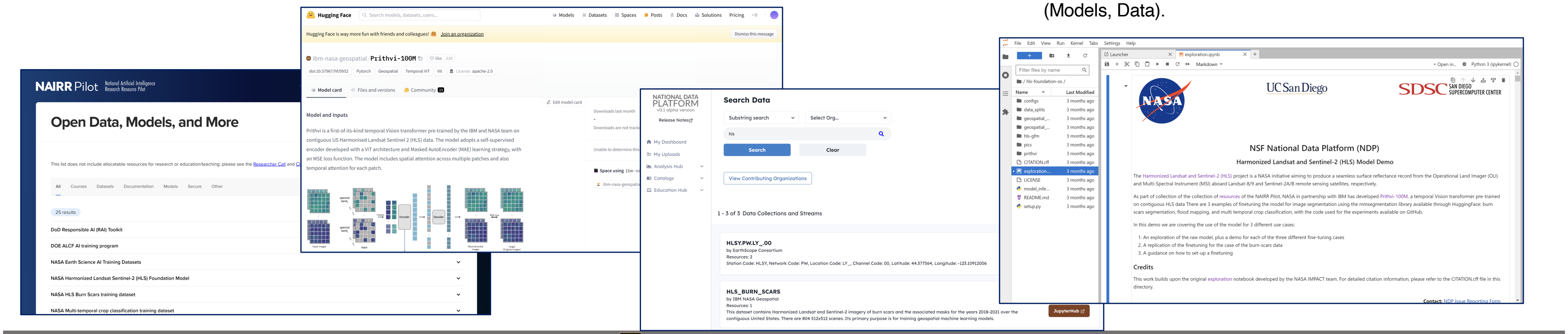
- Data and Models are identified as part of the Open NAIRR Resources.
- Resources are collected from HuggingFace

- Data and Models are registered into NDP catalog (CKAN)
- Data origin is created in OSDF to optimize data transfer

- Data and Models are included into user's workspace, along with the necessary libraries, services and files to work on a new project.

- Analysis and AI/ML workflow is supported by AI Gateway (JupyterHub), using NRP's Nautilus.
- High Performance processing for new resource(s) development (Models, Data).

- Final products pushed to OSDF/HuggingFace/GitHub and registered into NDP's catalog .



Open Data, Models, and More

This list does not include allocatable resources for research or education/teaching; please see the Researcher Call and Cl

25 results

DoD Responsible AI (RAI) Toolkit

DOE ALCF AI training program

NASA Earth Science AI Training Datasets

NASA Harmonized Landsat Sentinel-2 (HLS) Foundation Model

NASA HLS Burn Scars training dataset

NASA Multi-temporal crop classification training dataset

Hugging Face is way more fun with friends and colleagues! 🤗 Join an organization

ibm-nasa-geospatial/**Prithvi-100M** like 218

doi:10.57967/hf/0952 Pytorch Geospatial Temporal ViT Vit License: apache-2.0

Model and Inputs

Prithvi is a first-of-its-kind temporal Vision transformer pre-trained by the IBM and NASA team on contiguous US Harmonised Landsat Sentinel 2 (HLS) data. The model adopts a self-supervised encoder developed with a ViT architecture and Masked AutoEncoder (MAE) learning strategy, with an MSE loss function. The model includes spatial attention across multiple patches and also temporal attention for each patch.

Downloads last month
-
Downloads are not tracke

Unable to determine this

Space using ibm-na

ibm-nasa-geospatia

Search Data

hls

1 - 3 of 3 Data Collections and Streams

HLSY.PW.LY_.00
by EarthScope Consortium
Resources: 2
Station Code: HLSY, Network Code: PW, Location Code: LY_, Channel Code: 00, Latitude: 44.377564, Longitude: -123.10912006

HLS_BURN_SCARS
by IBM NASA Geospatial
Resources: 1
This dataset contains Harmonized Landsat and Sentinel-2 imagery of burn scars and the associated masks for the years 2018-2021 over the contiguous United States. There are 804 512x512 scenes. Its primary purpose is for training geospatial machine learning models.

/ hls-foundation-os /

Name	Last Modified
configs	3 months ago
data_splits	3 months ago
geospatial_...	3 months ago
geospatial_...	3 months ago
hls-gfm	3 months ago
pics	3 months ago
prithvi	3 months ago
CITATION.cff	3 months ago
exploration....	3 months ago
LICENSE	3 months ago
model_infe...	3 months ago
README.md	3 months ago
setup.py	3 months ago

NSF National Data Platform (NDP)

Harmonized Landsat and Sentinel-2 (HLS) Model Demo

The Harmonized Landsat and Sentinel-2 (HLS) project is a NASA initiative aiming to produce a seamless surface reflectance record from the Operational Land Imager (OLI) and Multi-Spectral Instrument (MSI) aboard Landsat-8/9 and Sentinel-2A/B remote sensing satellites, respectively.

As part of collection of the collection of resources of the NAIRR Pilot, NASA in partnership with IBM has developed Prithvi-100M, a temporal Vision transformer pre-trained on contiguous HLS data There are 3 examples of finetuning the model for image segmentation using the mmsegmentation library available through HuggingFace: burn scars segmentation, flood mapping, and multi temporal crop classification, with the code used for the experiments available on GitHub.

In this demo we are covering the use of the model for 3 different use cases:

1. An exploration of the raw model, plus a demo for each of the three different fine-tuning cases
2. A replication of the finetuning for the case of the burn-scars data
3. A guidance on how to set-up a finetuning

Credits

This work builds upon the original exploration notebook developed by the NASA IMPACT team. For detailed citation information, please refer to the CITATION.cff file in this directory.

Contact: NDP Issue Reporting Form

SDSC SAN DIEGO SUPERCOMPUTER CENTER

www.sci.utah.edu

UTAH

University of Colorado Boulder

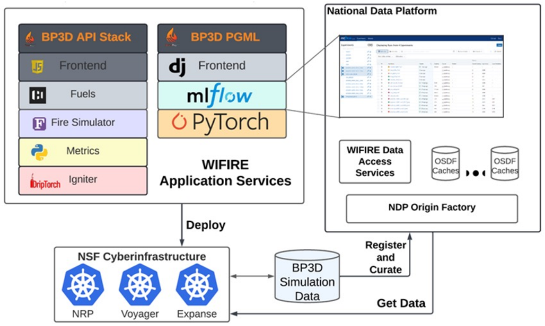
EarthScope Consortium

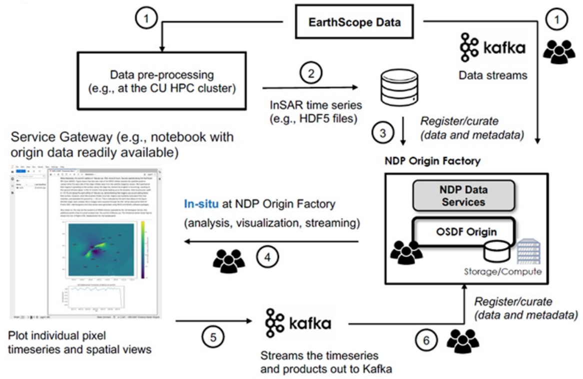
http://www.nationaldataplatform.org

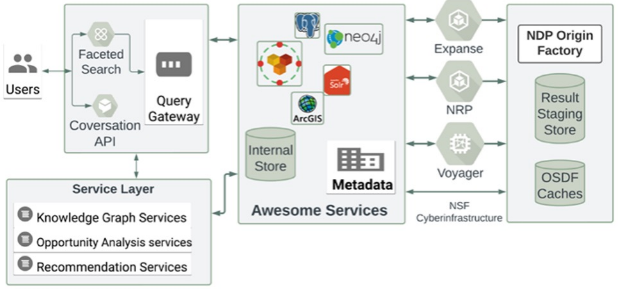
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Case Studies for Generalizable Workflows

- **Representative examples** of important patterns that exist in science today for working with
  - large datasets
  - streaming data from facilities
  - graph data from open knowledge networks
- Implemented as production-quality specialized value-added services
- Domains of wildland fire, earthquakes, and food security
- Will be generalized for replication by external communities.

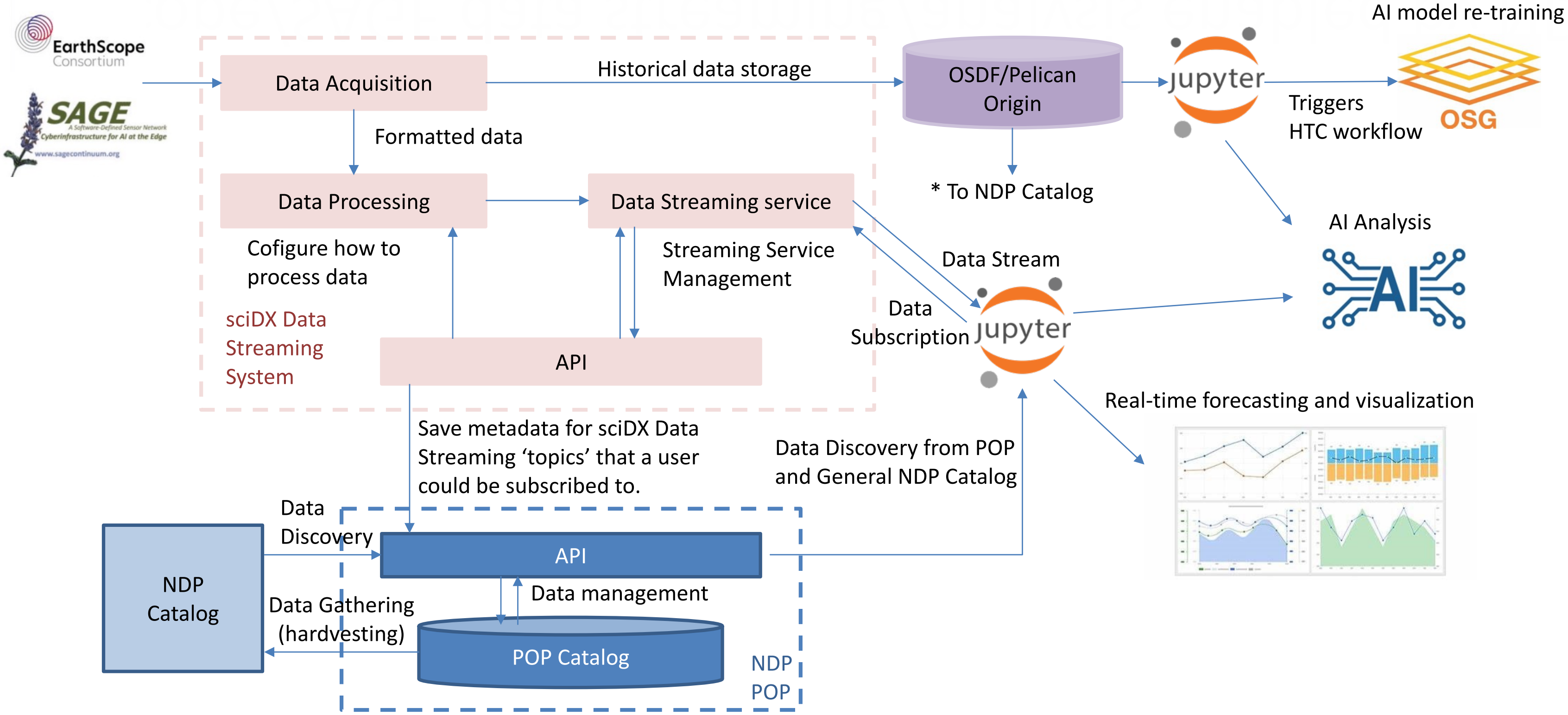



1
EarthScope Data
1
kafka
Data streams
Data pre-processing
(e.g., at the CU HPC cluster)
2
InSAR time series
(e.g., HDF5 files)
3
Register/curate
(data and metadata)
Service Gateway (e.g., notebook with origin data readily available)
In-situ at NDP Origin Factory
(analysis, visualization, streaming)
4
NDP Origin Factory
NDP Data Services
OSDF Origin
Storage/Compute
Plot individual pixel timeseries and spatial views
5
kafka
Streams the timeseries and products out to Kafka
6
Register/curate
(data and metadata)



EarthScope/SAGE data streaming/analysis enabled by NDP POP

- Real-time high-precision GNSS stations and SAGE data streams



AI model re-training

Data Acquisition

Historical data storage

OSDF/Pelican Origin

Triggers HTC workflow

OSG

Formatted data

* To NDP Catalog

Data Processing

Data Streaming service

AI Analysis

Cofigure how to process data

Streaming Service Management

Data Stream

Data Subscription

sciDX Data Streaming System

API

Real-time forecasting and visualization

Save metadata for sciDX Data Streaming ‘topics’ that a user could be subscribed to.

Data Discovery from POP and General NDP Catalog

Data Discovery

NDP Catalog

API

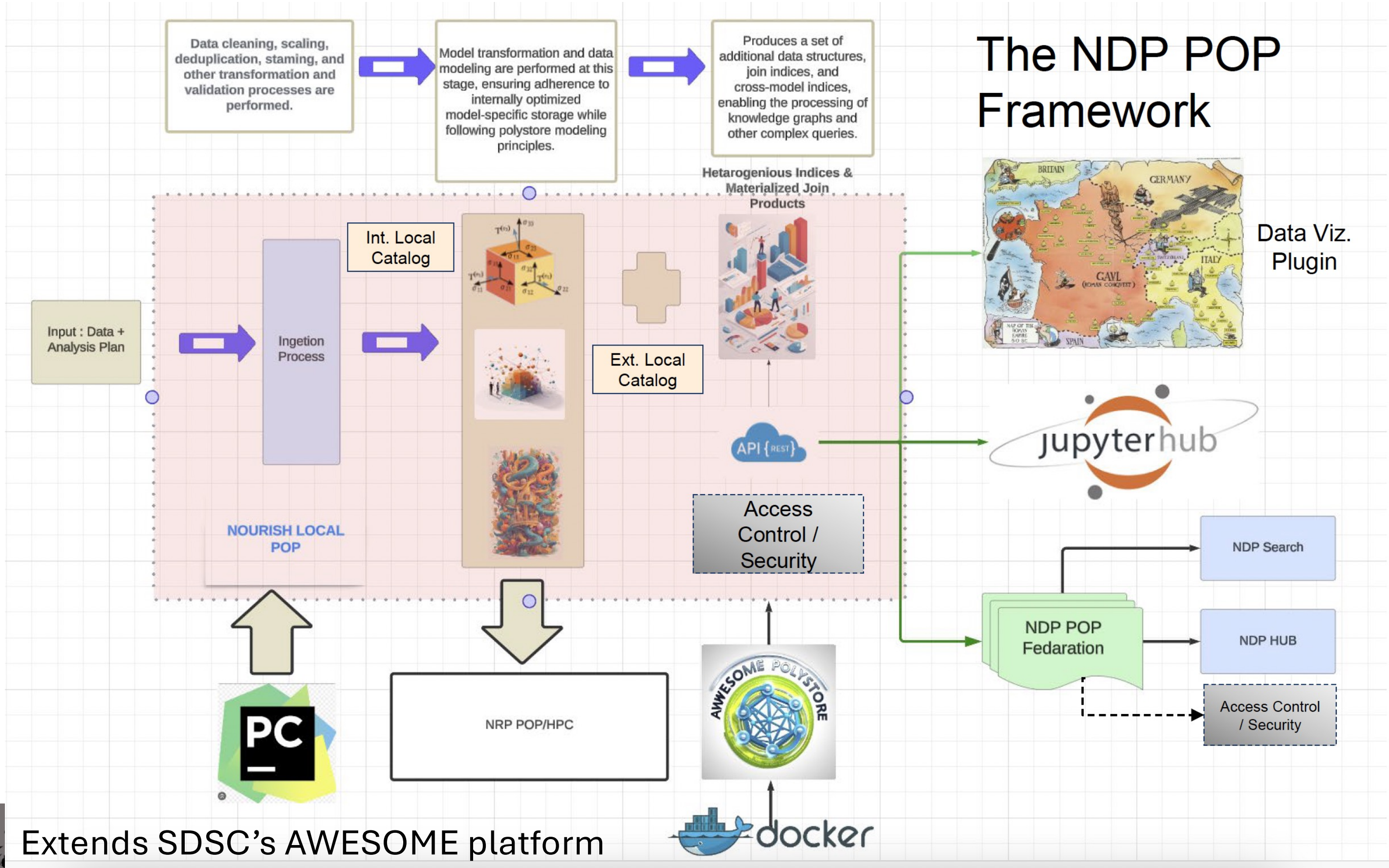
Data management

Data Gathering (hardvesting)

POP Catalog

NDP POP

NDP + NDR: Nourish NDF POP



Data cleaning, scaling, deduplication, staming, and other transformation and validation processes are performed.

Model transformation and data modeling are performed at this stage, ensuring adherence to internally optimized model-specific storage while following polystore modeling principles.

Produces a set of additional data structures, join indices, and cross-model indices, enabling the processing of knowledge graphs and other complex queries.

Hetarogenious Indices & Materialized Join Products

The NDP POP Framework

Data Viz. Plugin

Input : Data + Analysis Plan

Ingetion Process

Int. Local Catalog

Ext. Local Catalog

NOURISH LOCAL POP

Access Control / Security

NRP POP/HPC

NDP POP Fedaration

NDP Search

NDP HUB

Access Control / Security

Extends SDSC’s AWESOME platform

NATIONAL DATA PLATFORM

Bridging the Data Gaps for AI

UC San Diego

UTAH

University of Colorado Boulder

SDSC SAN DIEGO SUPERCOMPUTER CENTER

SCI www.sci.utah.edu

EarthScope Consortium

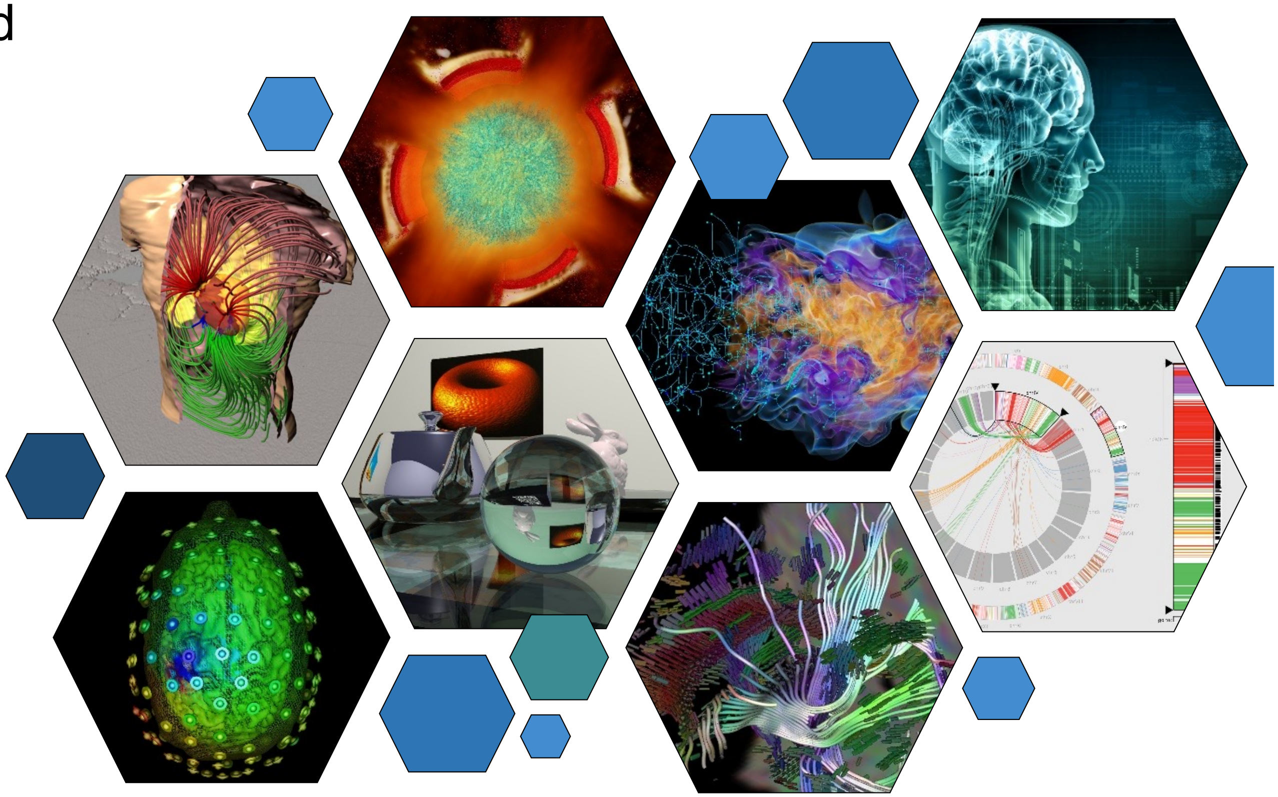
http://www.nationaldataplatform.org

Award abstract: https://www.nsf.gov/awardsearch/showAward?AWD_ID=2333609

SCI Institute

Transformation of science and society through ***translational research and innovation***

- Inter/transdisciplinary, collaborative, convergent
- Core strengths in: Visualization & imaging; Scalable analytics; Advanced computing & data
- Software/system development and distribution integral to our research processes



One-U Responsible AI Initiative

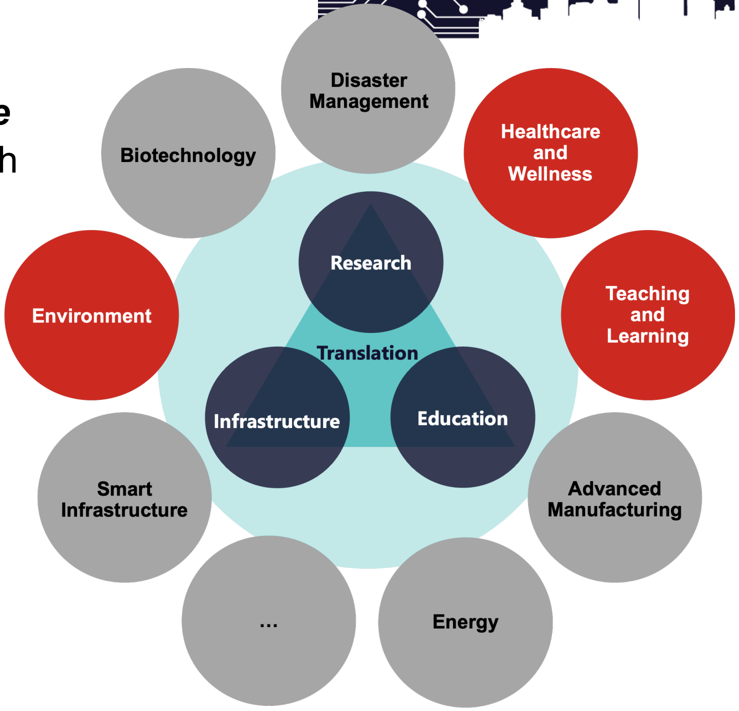
University of Utah’s $100M AI Research Initiative Led by SCI

Responsibly advance ***translational AI*** to achieve societal good

- Catalyze ***transdisciplinary excellence in responsible AI*** to bring together the use-inspired/applied AI research and technological expertise, advanced cyberinfrastructure, and translational workforce

Initial research focus on regionally important applications

1. Environment
2. Healthcare, societal wellness, and public services
3. Future of teaching and learning



Disaster Management

Healthcare and Wellness

Teaching and Learning

Advanced Manufacturing

Energy

...

Smart Infrastructure

Environment

Biotechnology

Research

Translation

Infrastructure

Education

Thank you!

One-U RAI Opportunities

- **Distinguished Visitors Program:** Supports visits from a few days to a full year for faculty, up to two years for postdoctoral fellows.
- **Postdoctoral Fellows Program:** Supports postdocs in areas related to responsible AI for up to 2 years.

Manish Parashar

Email: manish.parashar@utah.edu

WWW: manishparashar.org / sci.utah.edu / rai.utah.edu

one-u-responsible-ai-initiative

@OneU_RAI