

Bridging the Data Gaps to Democratize AI in Science, Education and Society

Keynote Talk for the 7th Rucio Community Workshop
September 16, 2024 – Osaka, Japan

İlkay ALTINTAŞ, Ph.D.

University of California, San Diego

Chief Data Science Officer & Division Director of Cyberinfrastructure and Convergence Research and Education, **San Diego Supercomputer Center**

Founding Fellow, **Halıcıoğlu Data Science Institute**

Founding Director, **Workflows for Data Science Center of Excellence**

Founding Director, **WIFIRE Lab**

Joint Faculty Appointee, Los Alamos National Laboratory



School of Computing, Information and Data Sciences

<https://scids.ucsd.edu/>



UC Regents Approve New School of Computing, Information and Data Sciences at UC San Diego

New school meets critical demand to advance data science and AI innovations and educate workforce of the future

SDSC SAN DIEGO SUPERCOMPUTER CENTER at the UNIVERSITY OF CALIFORNIA SAN DIEGO COMPUTING WITHOUT BOUNDARIES

ABOUT SDSC SERVICES SUPPORT RESEARCH & DEVELOPMENT EDUCATION

Materials Science Researchers Double Up on SDSC, PSC Supercomputers to Discover New Details about TMDs

Supercomputer simulations provide a better understanding of two-dimensional layered materials showing promise for a variety of applications – from flexible electronics and spintronics to optical and memory devices.

READ MORE



Innovate,

FOR UC/UCSD Researchers

FOR National HPC Users

<https://www.sdsc.ucsd.edu/>

UPCOMING EVENTS

- OCT 2 2:00 pm - 3:00 pm Some new results for streami
- OCT 12 8:00 am - 5:00 pm Swarup Swaminathan, MD | University of Miami Miller School of Medicine

[View Calendar](#)

Tweets from @HDSIUUSD

Nothing to see here - yet

When they Tweet, their Tweets will show up here.

[View on Twitter](#)

NEWS

Pioneering Data Science for a Data-Driven Future

JULY 18, 2023 - KALEIGH O'MERRY

NEWS

How Does ChatGPT Work? - Event

science.ucsd.edu/

Cyberinfrastructure and Convergence Research Division @SDSC

Translating cyberinfrastructure research for impact at scale

CI Methods and Systems

- “Big” Data and Knowledge Systems
- Computational Data Science
- Machine Learning and AI
- Advanced Computing

Convergence Research

- Collaborative Problem Solving
- Use-inspired Design
- Sustainable and Scalable Solutions

Experiential and Classroom Education

CICORE

Cyberinfrastructure | Convergence Research | Education



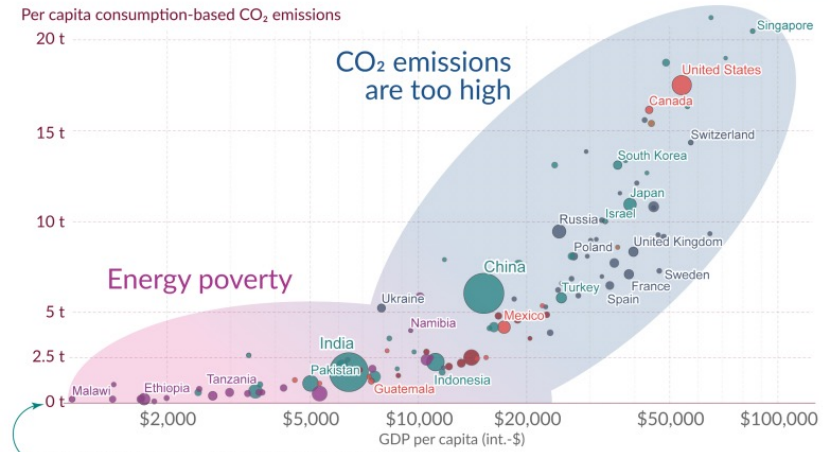
WCRP

GRAND CHALLENGES



CO₂ emissions per capita vs GDP per capita

Our World in Data



To end climate change the long-run goal is that net-emissions decline to zero.
 Data for 2017: Global Carbon Project, UN Population, and World Bank.
 OurWorldInData.org - Research and data to make progress against the world's largest problems. Licensed under CC-BY by the author Max Roser.

<https://ourworldindata.org/worlds-energy-problem>

\$10 trillion+ spent in global response to COVID-19

165+ COVID-19 vaccines being developed globally

216 countries, areas or territories with cases

<https://www.10xgenomics.com/research-areas/infectious-disease>

The biggest challenges of our time are too difficult to solve alone!

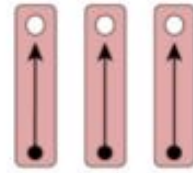
<https://www.wcrp-climate.org/learn-grand-challenges>

Convergence research is:

driven by a specific and compelling societal problem

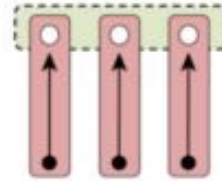
and

works towards integrating innovative and sustainable solutions into society



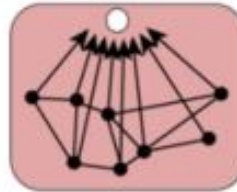
Disciplinary

- Within one academic discipline
- Disciplinary goal setting
- Development of new disciplinary knowledge



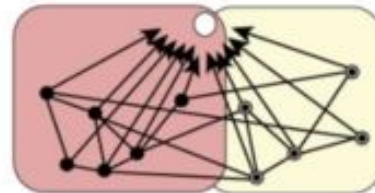
Multidisciplinary

- Multiple disciplines
- Multiple disciplinary goal setting under one thematic umbrella



Interdisciplinary

- Crosses disciplinary boundaries
- Development of integrated knowledge



Convergence

- Crosses disciplinary and sectorial boundaries
- Common goal setting
- Develops integrated knowledge for science and society
- Creates new paradigms

● Stakeholder Participants
● Discipline

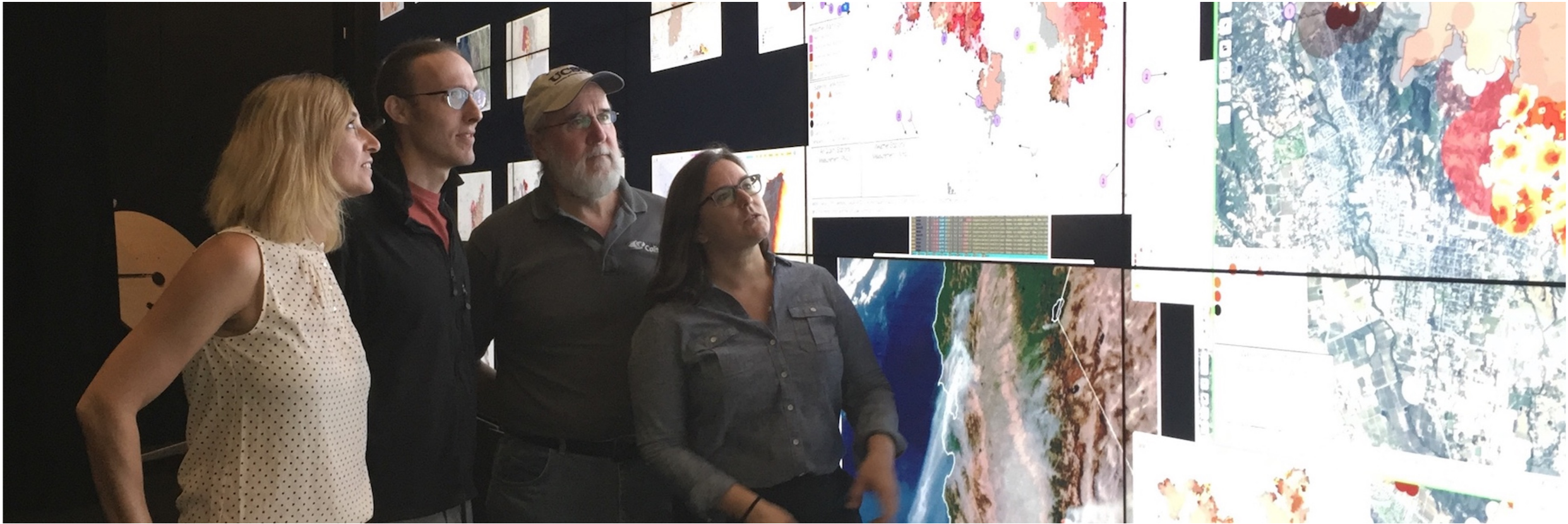
○ Goal, Shared Knowledge
■ Academic Knowledge

▭ Thematic Umbrella
▭ Conventional Knowledge

Adapted from Wright Morton, L., S. D. Eigenbrode, and T. A. Martin. 2015. Architectures of adaptive integration in large collaborative projects. *Ecology and Society* 20(4):5.

Translating Research into Impact

through Democratizing Access to Cyberinfrastructure



Three Main Components

Composable Workflows + Collaborative Innovation + Impact Network



Workflows for Data Science
Center of Excellence

- Develop methodologies and tools to enable **collaborative workflow-driven data science**
- Create solution architectures on top of **big data and advanced computing platforms**
- Push the boundaries of the **computing continuum through composable systems and services**



Proactive Wildfire & Environmental Sustainability Solutions

- Create **collaborative pathways** between UC San Diego and Los Alamos National Lab
- Accelerate the advancement of **science and technology as a basis for responsible and proactive approaches** to environmental challenges
- Leverage cutting-edge capabilities in scalable computing and diverse scientific expertise to foster **solution-focused, community-facing innovation**



- Catalyze an **impact network** of students, researchers, practitioners, industry leaders, and public policy professionals committed to **collaboratively engaging in research** that is **driven by specific and compelling societal problems** and requires deep integration across disciplines and sectors to create solutions
- Provide participants with a **foundational experience** to position them **for impact** throughout their careers on the most challenging issues of our time.

CORE Institute Innovation Approach

Creating Breakthrough Technological Innovations for Complex Societal Challenges

Use-Inspired Problems

- 1) Context evaluation:** Describe the system(s) within which the problem you are addressing exists and identify important decision-makers and vulnerable communities
- 2) Needs assessment:** Clarify the needs of the people you want to help and ensure you are solving the right problem
- 3) Innovation pathways:** Sketch out ideas for data and science that could contribute to solving the problem and outline the expertise needed

Use-inspired & iterative
co-production of innovation
with users

CORE4 Building Blocks

Data & AI	Cutting-Edge Science & Engineering
Advanced Digital Infrastructure	Integrated Workflows

Use-inspired & iterative
co-creation of solutions
with partners

Scalable Solutions

- 1) Sustainable partnership model:** Implement solutions through a model that will allow for sustained use at scale
- 2) Continued iteration:** Monitor performance and impact through user feedback and key metrics and be ready to adapt
- 3) Continued innovation:** Create mechanisms to ensure innovation is an ongoing process

From USEFUL

to USABLE

to USED at scale

Translating Fire Research into Impact



Mission: Develop technologies with the fire management community driven by cutting-edge science and data

Vision: Enable tools that can have an impact at the scale of the environmental challenges we face today



wifire.ucsd.edu

Where are we headed at WIFIRE Lab?

- **Wildfire Response:** WIFIRE's Firemap platform in collaboration with CALOES and CAL FIRE through California's Fire Integrated Real-Time Intelligence System (FIRIS) and with partners in Colorado
- **Beneficial Fire:** WIFIRE's BurnPro3D platform for prescribed burn planning and implementation in collaboration with 3D fuel and fire modeling efforts at USGS, DOD, USFS, and LANL
- **Data and Model Sharing:** WIFIRE's Wildfire Technology Commons to develop standards, tools and techniques to share data and data-driven models to enable scientific workflows and AI innovation in collaboration with partners including NIST, CAL FIRE, and SDGE
- **Immersive Visualization:** AI-readiness of scientific data for new modes of teaching, training, decision-making, and public communication, including 3D outputs from vegetation modeling and fire science simulations and real-world information collected with cameras and sensors

Operational Products

FIREMAP

Firemap is currently being used by firefighters in Colorado, in collaboration with Intterra, and firefighters in California through the FIRIS program under the California Governor's Office of Emergency



REACTIVE

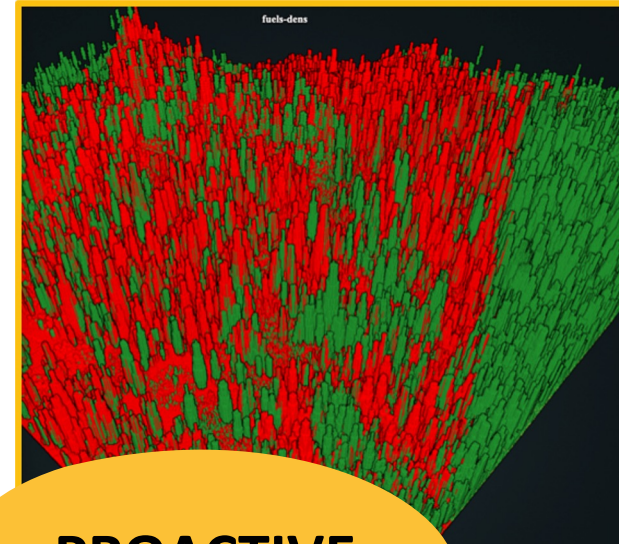
Services and CALFIRE. FIRIS uses Firemap to provide real-time information on weather conditions and fire ignitions and to monitor and predict direction and speed of fire spread, as well as communities at risk. It has revolutionized initial attack response for the most dangerous fires across California.



Cal OES
GOVERNOR'S OFFICE
OF EMERGENCY SERVICES

BurnPro^{3D}

In alignment with the nation's goal to increase fuel treatments to reduce wildfire risk, BurnPro3D is designed to support the preparation of burn plans as well as the implementation of prescribed



PROACTIVE

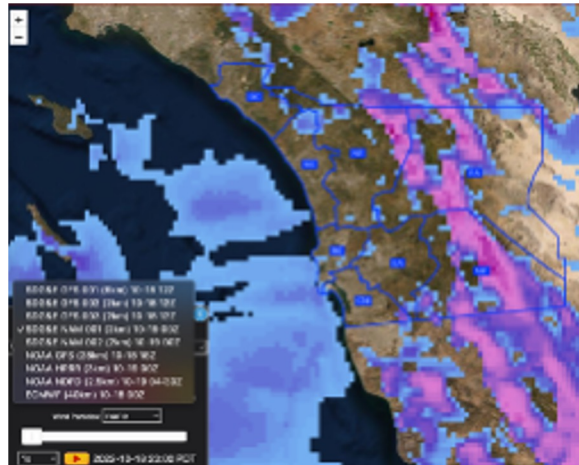
burns. The interface allows burn bosses to create and visualize high-resolution 3D fire simulations and compare fuel consumption and risk under different weather and ignition scenarios. It uses 3D FastFuels data developed by the US Forest Service and the QUIC-Fire coupled fire/atmosphere model developed at Los Alamos National Lab.



Data and Computing Platforms

Wildfire Science and Technology Commons

The Commons enables the development of foundational AI techniques to fuse and learn from data and to make scientific models interpretable and complex decisions easier. It connects next-generation data and



models for anyone interested in developing solutions. For example, it enables an integrated fire weather intelligence platform focused on reducing risk related to power lines for Southern California. A new phase of development was recently supported through congressionally directed spending proposed by California Sen. Padilla, Rep. Vargas, and Rep. Jacobs.



Wildfire and Landscape Resilience Data Hub

The Data Hub is a federated data ecosystem for California's Wildfire and Forest Resilience Task Force, providing a "single view" over existing data to fulfill the reporting requirements for California's

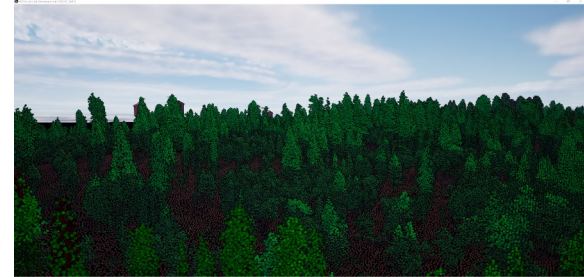
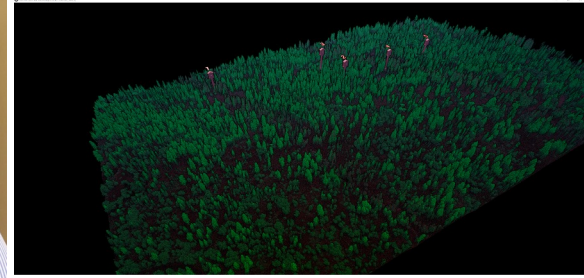
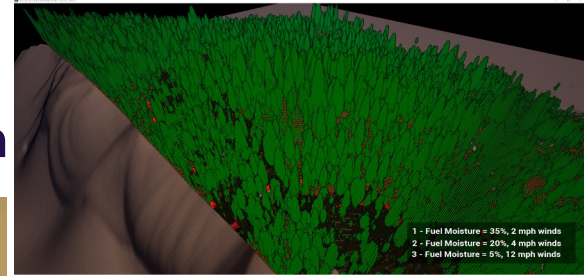


Million Acre Strategy to treat 1 million forested acres per year to reduce wildfire risk. It will provide public, open, and fair access to data, analytic tools, and customizable reports via the Data Hub explorer web viewer, as well as access to data through APIs.



Immersive Forest

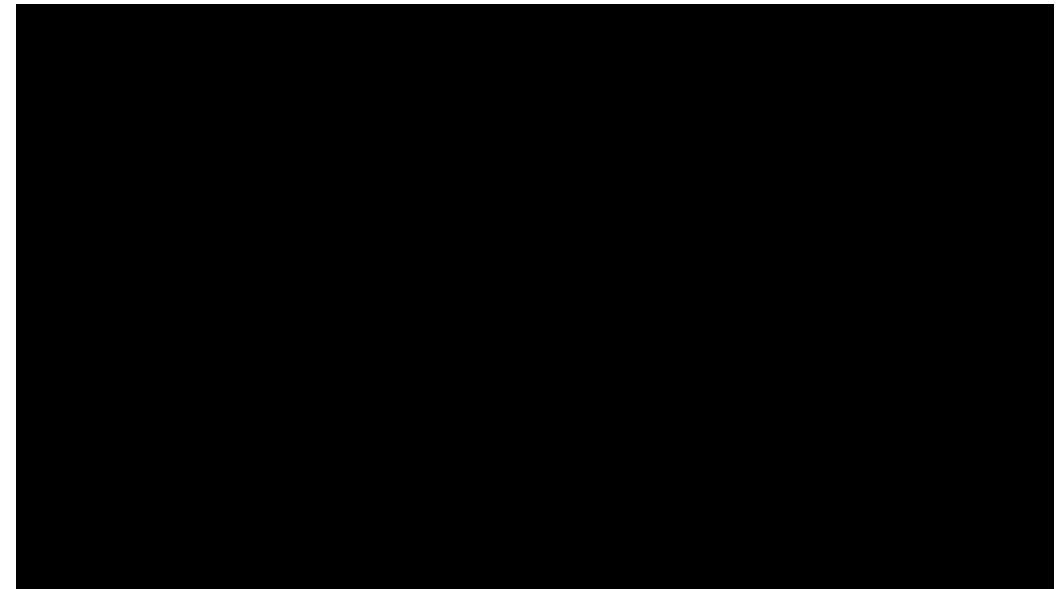
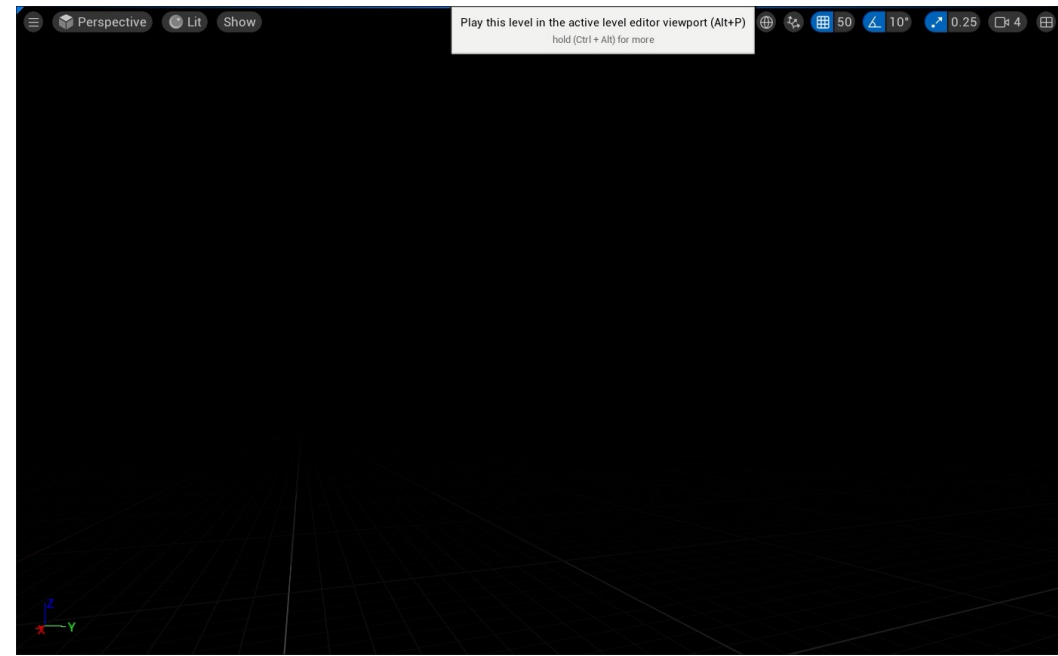
Immersive Forest for Multimodal Communication



*Terrestrial LiDAR contextualized
within Aerial scan*

AI in Science Communication

Visualization of multiple terrestrial LiDAR scans in the Immersive Forest prototype



Immersive AI-integrated visualization of scientific data and simulations for training, decision making, and public communication.

Animations by: Isaac Nealey (left, bottom), Ivannia Gomez (top)

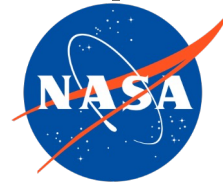
Additional Grants Fueling R&D



Evaluation of satellite-based fire detection and fire radiative power applications



Ground sensing and in-situ edge computing for monitoring and decision-making



Open fire models to predict wildfire spread over 3-5 days



Workflows for DOD prescribed fire managers participating in the National Innovation Landscapes Network



Prescribed fire planning and monitoring tools and workforce training for California agencies



Multi-modal data to improve characterization of fuels at large spatial extents and fine spatial scales



Immersive visualization of scientific data for new modes of training, decision-making and communication

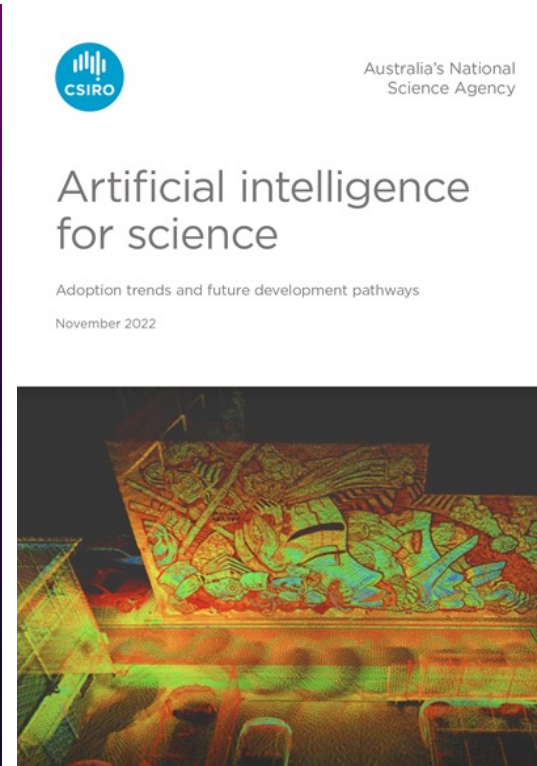
**This type of work
needs the CORE4
building blocks.**

CICORE

Cyberinfrastructure | Convergence Research | Education

CORE4 Building Blocks

Data & AI	Cutting-Edge Science & Engineering
Advanced Digital Infrastructure	Integrated Workflows



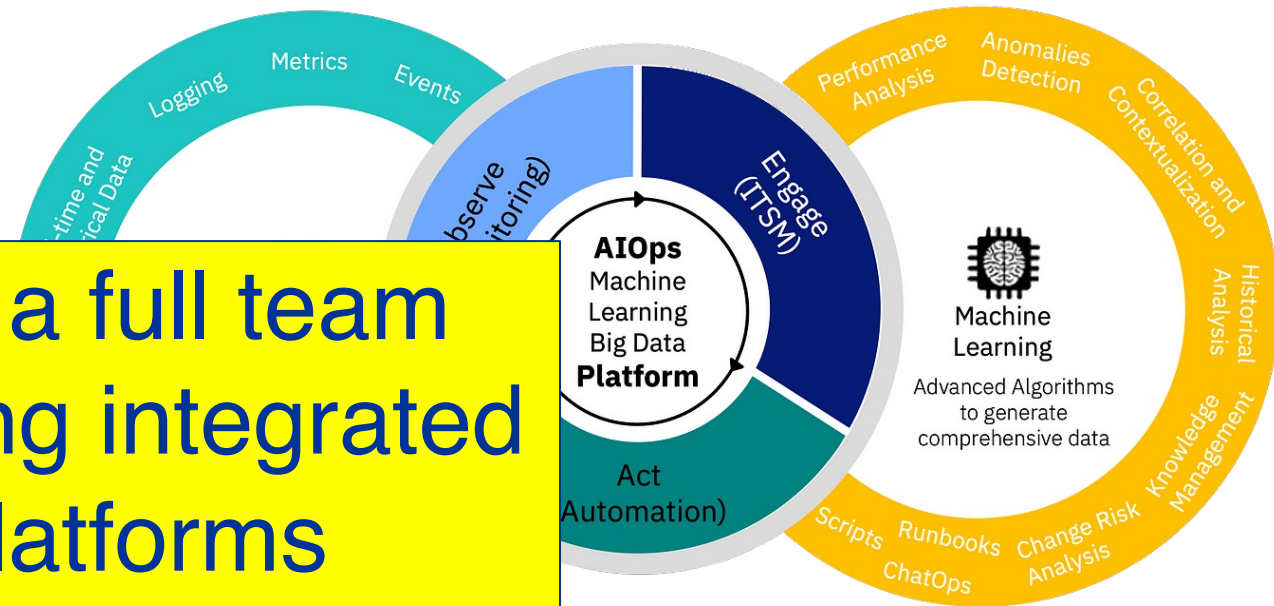
AI in Science and Research 2023

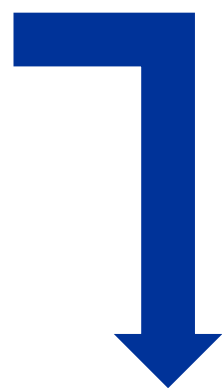
AI in Science Readiness

“not just science + AI methods”

- Data federation and hubs
- Data quality and volume
- Knowledge management
- Benchmarks
- Scalability up and down
- Workflow management
- Software integration and engineering
- Dev ops (also called AI ops and data ops)
- Interpretability and explainability
- Workforce training and culture/incentive building

Requires a full team
and enabling integrated
data platforms





Systems should enable seamless integration of AI-integrated application workflows by teams!





Workflow integration requires a digital continuum composed through:

- system federation**
- reusable capability services**
- solutions integrating services**



AI in science requires data and knowledge hubs including:

- data federation
- knowledge management
- readily available standard data services
- equitable access

Integration requirements...



Dynamic composability matters.

Systems and services are useful if groups can integrate them into applications.



TEAMWORK

Tools that enhance teamwork and use need to be coupled with responsible AI systems.



Dynamic composability matters.

COMPOSABLE SERVICES

e.g., model and data archives, learning and analytics, simulation, training

RESOURCE MANAGEMENT

e.g., container orchestration, optimization

COMPOSABLE SYSTEMS

e.g., GPU, CPU, Big Data, quantum, neuromorphic, SDN, storage

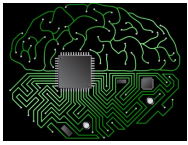
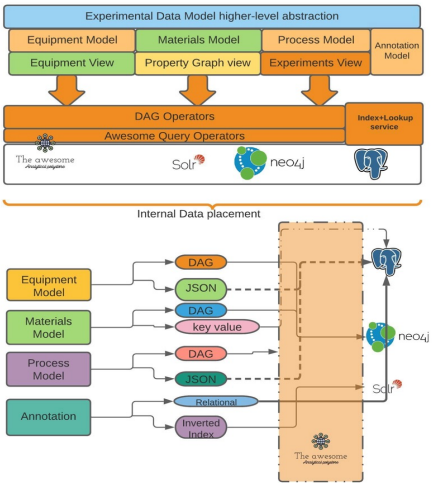
Big Data and IoT

Artificial Intelligence

Modeling and Simulation

Capability

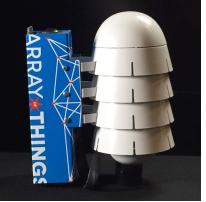
Capacity



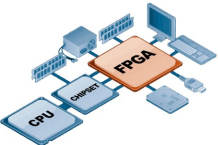
xPU → GPU, CPU, TPU, IPU, QPU, ...



Big Data



Edge



FPGA

Cloud, HPC, Storage

Some Composable Systems

EXPANSE

COMPUTING WITHOUT BOUNDARIES
5 PETAFLOP/S HPC and DATA RESOURCE

HPC RESOURCE

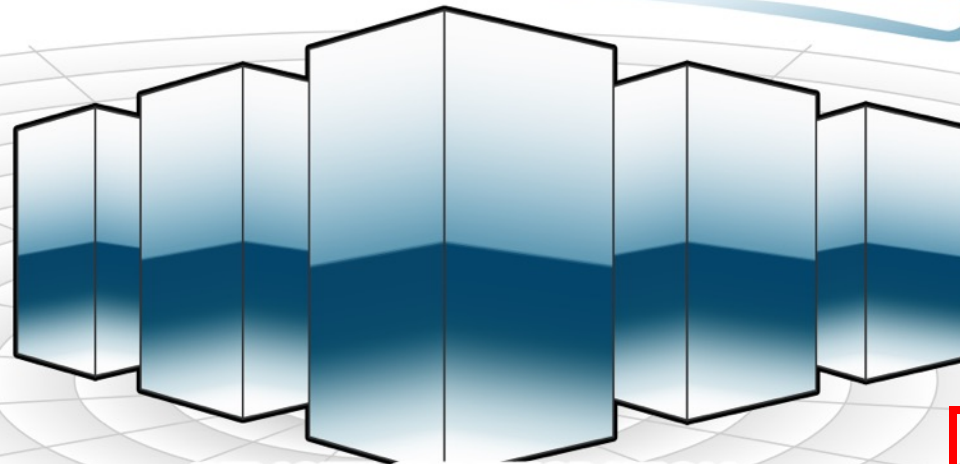
- 13 Scalable Compute Units
- 728 Standard Compute Nodes
- 52 GPU Nodes: 208 GPUs
- 4 Large Memory Nodes

LONG-TAIL SCIENCE

- Multi-Messenger Astronomy
- Genomics
- Earth Science
- Social Science

INNOVATIVE OPERATIONS

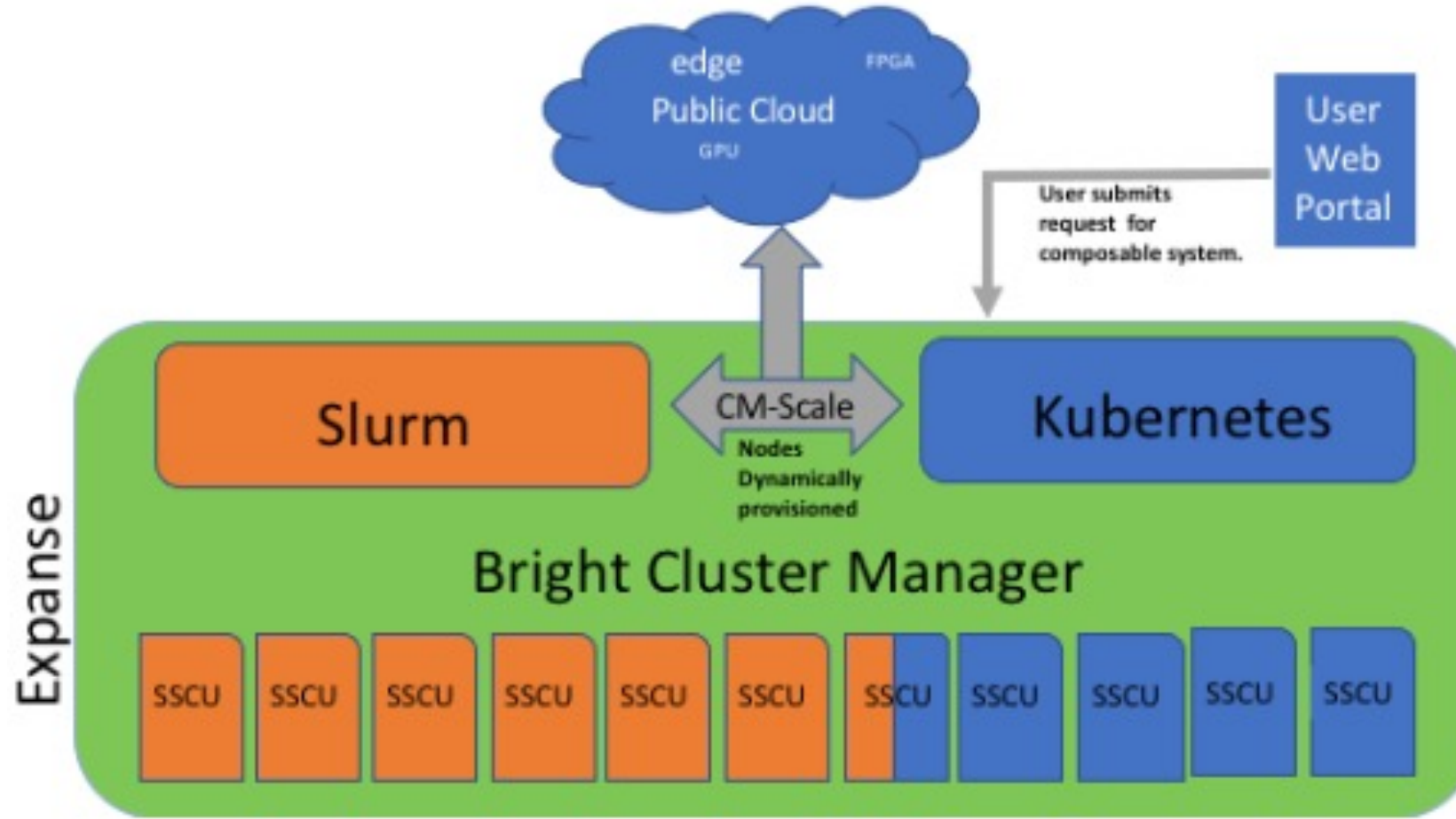
- Composable Systems
- High-Throughput Computing
- Science Gateways
- Interactive Computing
- Containerized Computing
- Cloud Bursting



REMOTE CI INTEGRATION

- ### DATA CENTRIC ARCHITECTURE
- 12PB Perf. Storage: 140GB/s, 200k IOPS
 - Fast I/O Node-Local NVMe Storage
 - 7PB Ceph Object Storage
 - High-Performance R&E Networking





Expanse Composable Systems Framework

National Research Platform



<https://nationalresearchplatform.org/>

A screenshot of the National Research Platform website. The header features the "NRP NATIONAL RESEARCH PLATFORM" logo on the left and navigation links for "NEWS", "GRANTS", "TECHNOLOGY", "COMMUNITY", and "JOIN / CONTACT" on the right. The main content area has a dark blue background with a satellite-style map of the world. The headline "Designed for Growth and Inclusion" is prominently displayed in white. Below it, a paragraph states: "The National Research Platform (NRP) is a partnership of more than 50 institutions, led by researchers and cyberinfrastructure professionals at UC San Diego, supported in part by awards from the National Science Foundation." Two call-to-action buttons are visible: a pink one labeled "REGISTER FOR 5NRP" and a green one labeled "THE PRP IS NOW THE NATIONAL RESEARCH PLATFORM (NRP)".

First composable cluster is federated!

EXPANSE (Enthalpy) + CHASE-CI (Nautilus)

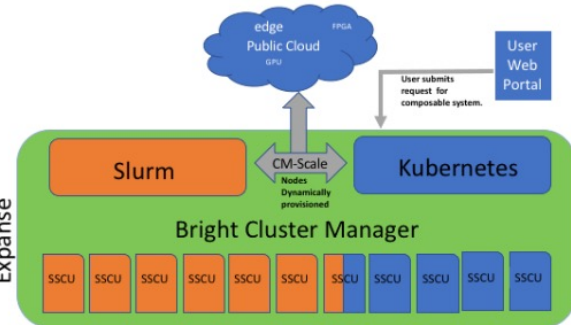
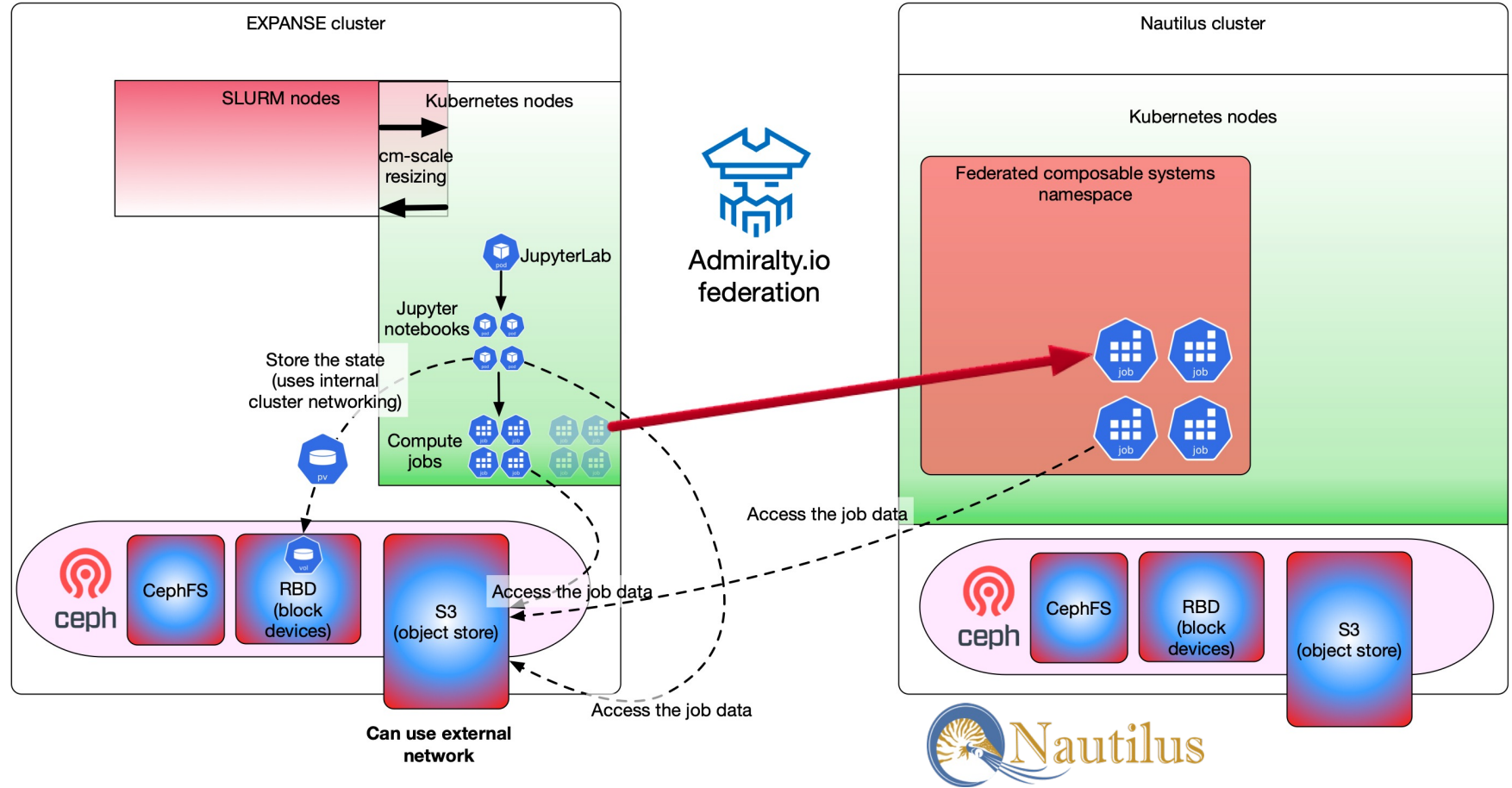


Figure 5.1 Expance Composable Systems Framework

**EXPANSE
(Enthalpy)**



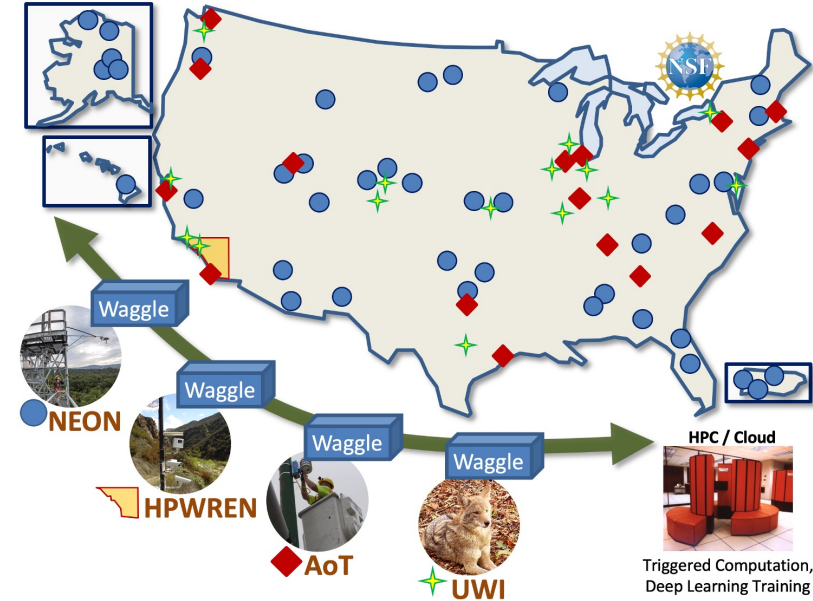
AI@Edge and the Digital Continuum

Slide Source: Pete Beckman, ANL



SAGE
Cyberinfrastructure for
AI at the Edge
sagecontinuum.org

NSF



Leadership Team

 Pete Beckman (NU; Director)	 Nicola Ferrier (NU; Deputy Dir.)
 Ilkay Altintas (SDSC; Data)	 Charlie Catlett (Illinois; AoT)
 Scott Collis (NU; ARM)	 Valerie Taylor UChicago; Broader Impacts)
 Jim Olds (GMU; Life Sci, Risk)	 Dan Reed (Utah; Architecture)
 Eugene Kelly (CSU; NEON)	 Irene Qualters (LANL; Advisory Committee Chair)

Education & Training



Northwestern University
Argonne NATIONAL LABORATORY
Colorado State University
UC San Diego
GEORGE MASON UNIVERSITY
ARM
neon National Ecological Observatory Network
HPWREN SDSC SAN DIEGO SUPERCOMPUTER CENTER
LINCOLN PARK ZOO FOR WILDLIFE. FOR ALL.

DATA LIFECYCLE MANAGEMENT

*e.g., active data repositories, long-term archives,
knowledge networks, data reuse services*

Systems and services are only useful if groups can integrate them into applications.



WORKFLOW MANAGEMENT

*e.g., application integration, coordination, optimization,
communication, reporting*

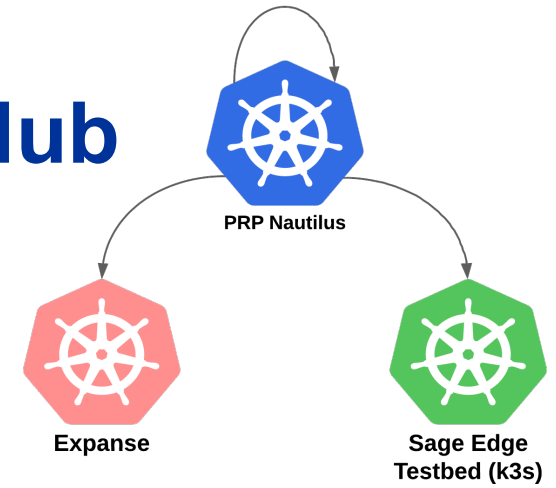
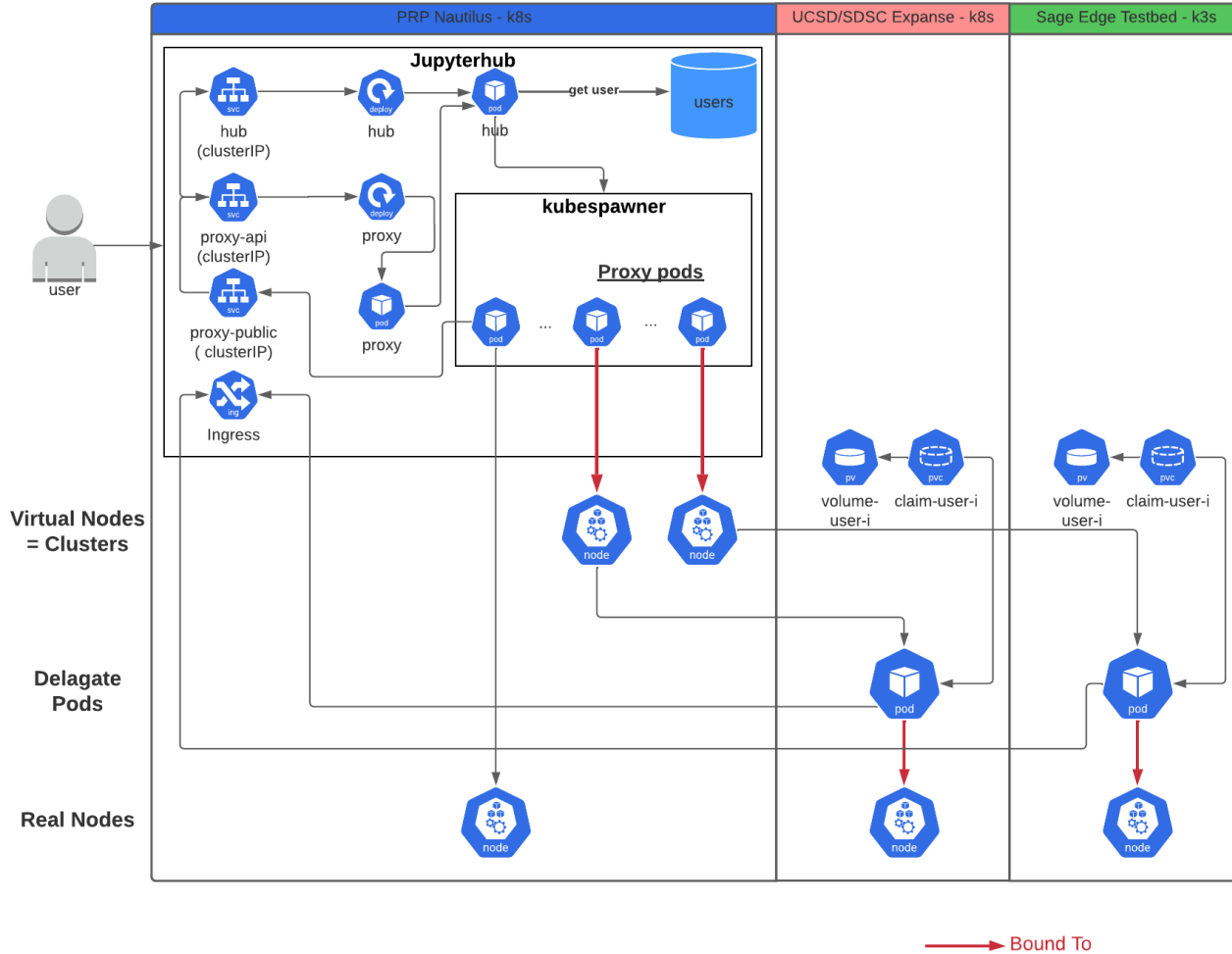
COMPOSABLE SERVICES

RESOURCE MANAGEMENT

COMPOSABLE SYSTEMS

Integration of NSF EXPANSE, NRP and Sage

A Composable System Deployment of JupyterHub



- Edge-Cloud Unified Environment for prototyping AI models to deploy on the Edge
- A user can easily be provided the right environment for developing their AI Edge Application

I. Altintas et al., "Towards a Dynamic Composability Approach for using Heterogeneous Systems in Remote Sensing," 2022 IEEE e-Science doi: 10.1109/eScience55777.2022.00047

Spawner Options

/home/jovyan is persistent volume, 5GB by default. Make sure you don't fill it up - jupyter won't start next time. You can request increasing the size in [Matrix](#)

GPUs

Cores

RAM, GB

GPU type

/dev/shm for pytorch

Mount CephFS (if assigned)

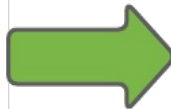
You can request assignment in [Matrix](#)

Stack options are described in [docker-stacks](#)

Image

- Stack Minimal
- Stack Minimal + Desktop GUI
- Stack Scipy
- Stack R
- Stack Tensorflow
- Stack Tensorflow + PRP added libs
- Stack Datascience
- Stack Pyspark
- Stack All Spark
- Tensorflow 1.14 (deprecated, choose one above)

Spawn



Kubernetes
Pod
Spawned
for
Exploration

File Edit View Run Kernel Tabs Settings Help

Name	Last Modified
bin	4 months ago
dask-worker-space	5 months ago
data	5 months ago
include	4 months ago
kubernetes	5 months ago
lib	4 months ago
pgsql	4 months ago
postgresql-11.0	4 months ago
rclone-v1.53.1-linux-a...	5 months ago
share	4 months ago
Tempredict-Shared-P...	4 months ago
usr	a year ago
1	4 months ago
dbConnString.ipynb	5 months ago
dbConnstring.py	5 months ago
DeveloperNB-Timesca...	5 months ago
GitDemo-Tempredict-...	a month ago
KCLT.csv	5 months ago
mydask.png	5 months ago
Ops-TimescaleDB-Ta...	17 hours ago
ordered-clustering-da...	5 months ago
postgresql-11.0.tar.gz	2 years ago
PPTDemo-TimescaleD...	2 days ago
rclone-current-linux-a...	6 months ago
tempredict-oura-500-...	a month ago
Tempredict-timescale...	seconds ago
TimescaleDB-Dask-C...	5 months ago
TimescaleDB-Dask-C...	5 months ago
TimescaleDB-Dask-ps...	5 months ago
Untitled.ipynb	4 months ago
Untitled1.ipynb	16 days ago

Import Libraries

```
[6]: import pandas as pd
import numpy as np
import dask
import distributed
from sqlalchemy import create_engine
from dbConnstring import *
import dask.array as da
import os
import time
```

Define Database Connection Details

```
[6]: # TimescaleDB username, password, and database name
TimescaleDB_USERNAME = '' ## YOUR TimescaleDB_USERNAME = 'abc'
TimescaleDB_PASSWORD = '' ## YOUR TimescaleDB_PASSWORD = 'xyz'
# Create the connection
postgres_str = conn_str(TimescaleDB_USERNAME, TimescaleDB_PASSWORD)
cnx = create_engine(postgres_str)
```

Create a DASK Cluster

```
[12]: from dask import dataframe as dask_cluster_dd
```

```
[13]: N_WORKERS = 8
```

```
[14]: from dask.distributed import Client, LocalCluster
if __name__ == '__main__':
    # Create a Dask Cluster
    cluster = LocalCluster(n_workers=N_WORKERS, threads_per_worker=1, processes=True)
    client = Client(cluster)
```

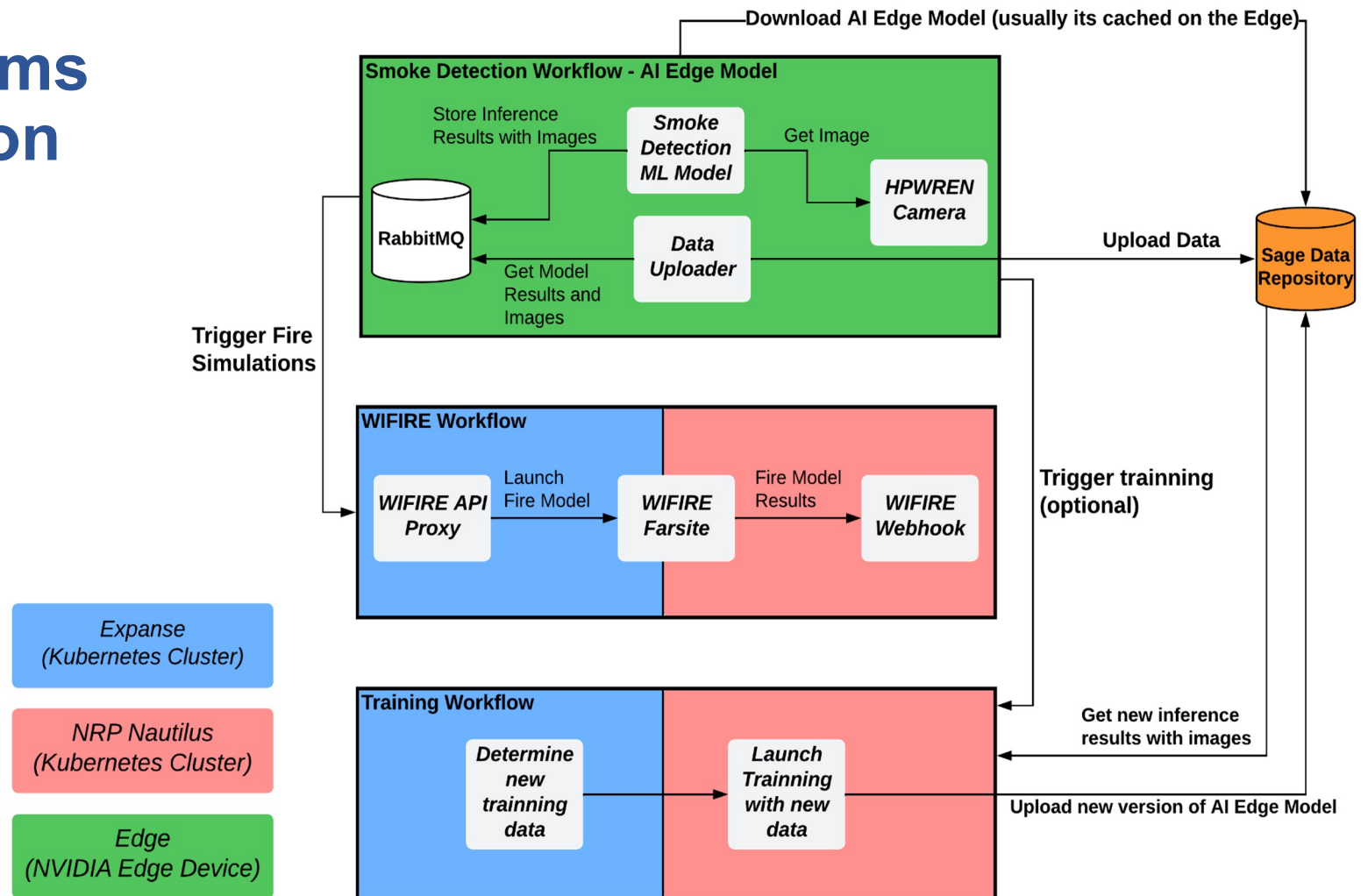
Client	Cluster
Scheduler: tcp://127.0.0.1:40939	Workers: 8
Dashboard: http://127.0.0.1:8787/status	Cores: 8
	Memory: 135.06 GB

Create a 'delayed' function with DASK Cluster

```
[16]: df = dask_cluster_dd.read_sql_table('hrv_500',
postgres_str,
npartitions=8,
index_col='sensortime')
df
```


Fire Simulations using Composable Systems and Edge Smoke Detection

- Three workflows
 - Smoke – Sage Edge App
 - Fire simulator
 - AI Training
- Both the fire simulator and training workflows are can be run on Expanse or Nautilus through the federation layer



I. Altintas et al., "Towards a Dynamic Composability Approach for using Heterogeneous Systems in Remote Sensing," 2022 IEEE e-Science
doi: 10.1109/eScience55777.2022.00047



Humpty Dumpty sat on the AI wall,
Assessing systems, both large and small.

**We need to stay
use-inspired!**

—but are they prepared,
blems for which they're declared?"



The king's horses ran simulations fast,
While the king's men studied data amassed.
Yet despite their efforts, and all they'd apply,
True AI readiness still passed them by.

RESPONSIBILITY

e.g., accuracy, privacy, explainability, ethics

REPRODUCIBILITY

TEAM SCIENCE

USE-INSPIRED INTERFACES

e.g., for science, education and scalable practice

Tools that enhance teamwork and use need to be coupled with responsible AI systems.

TEAMWORK

RESPONSIBILITY

e.g., accuracy, privacy, explainability, ethics, equity

REPRODUCIBILITY

TEAM SCIENCE

DATA LIFECYCLE MANAGEMENT

e.g., active data repositories, long-term archives, knowledge networks, data reuse services

USE-INSPIRED INTERFACES

e.g., for science, education and scalable practice

WORKFLOW MANAGEMENT

e.g., application integration, coordination, optimization, communication, reporting

COMPOSABLE SERVICES

e.g., model and data archives, learning and analytics, simulation, training

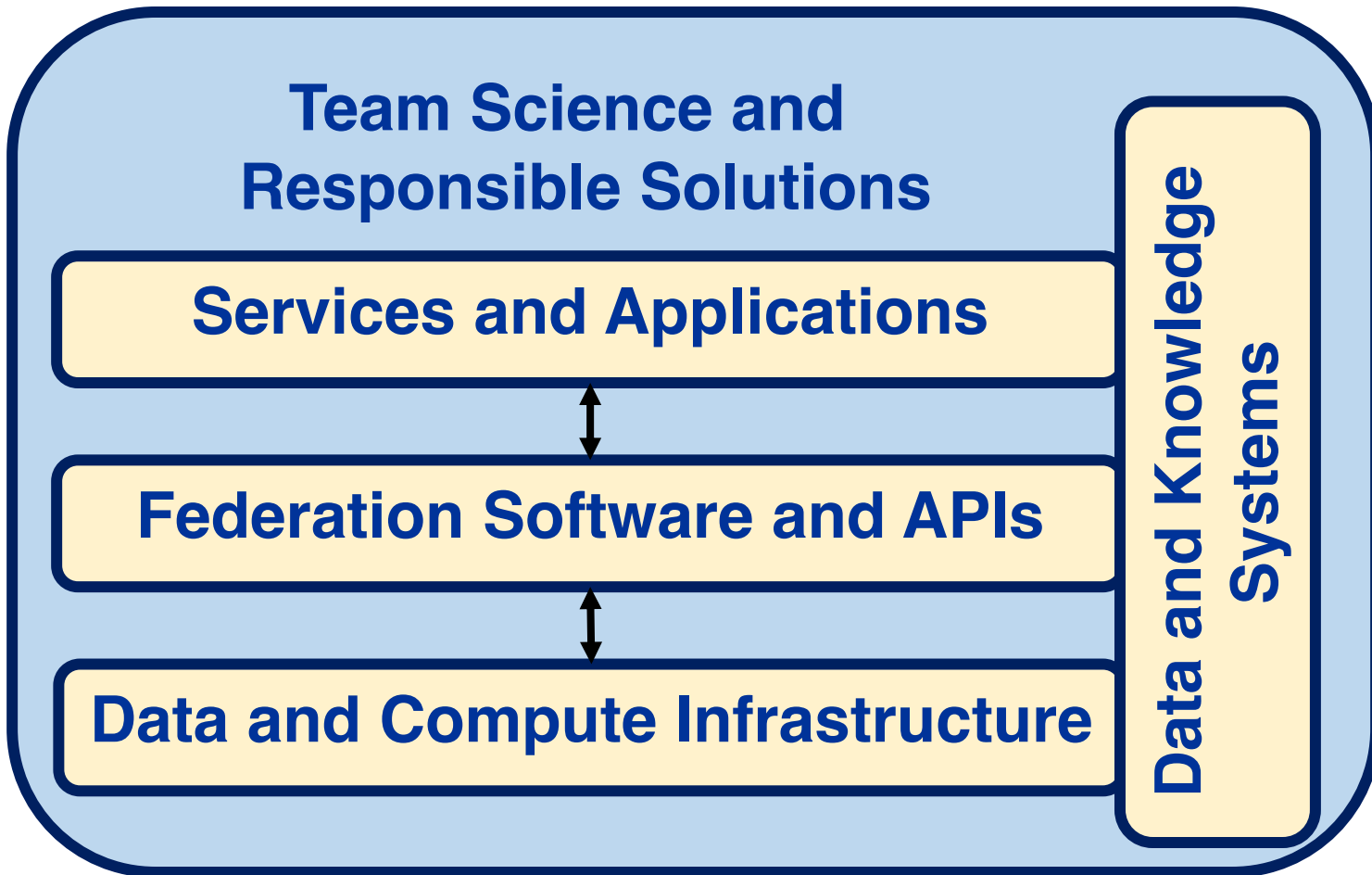
RESOURCE MANAGEMENT

e.g., container orchestration, optimization

COMPOSABLE SYSTEMS

e.g., GPU, CPU, Big Data, quantum, neuromorphic, SDN, storage

Use-Inspired Composability from Systems to Services



- User-centered design and experience
- Improved FAIR data capacity
- Capability-based integration
- Create plug and play microservices
- Run across many systems
- Dynamically measure, manage and provision resources

Schmidt AI in Science Postdoc Research



Computational microscopy of respiratory viruses in aerosols

Exploring different models to simulate and visualize the behavior of viruses in the respiratory tract

AI-Powered analysis of molecular simulations

High-affinity generative model for target proteins

Data-driven development of neural-network potentials from quantum chemistry data

ML model to be used as a surrogate for expensive high-level chemistry calculations

Drug resistance evolution in HIV patients

Leverages machine learning system for heterogeneous cryo-EM reconstruction of proteins and protein complexes from single-particle cryo-EM data

The relationship between life span of the plant roots microscopy data and wildfire

Deep learning model to estimate life span

Small coronary artery calcium detectability

Deep learning model to segment and visualize chambers of the heart

Earth system modelling

Deep learning model to use data extracted from ECMWF to calibrate earth systems simulation

Brain activity of diving seals reveals short sleep cycles at depth

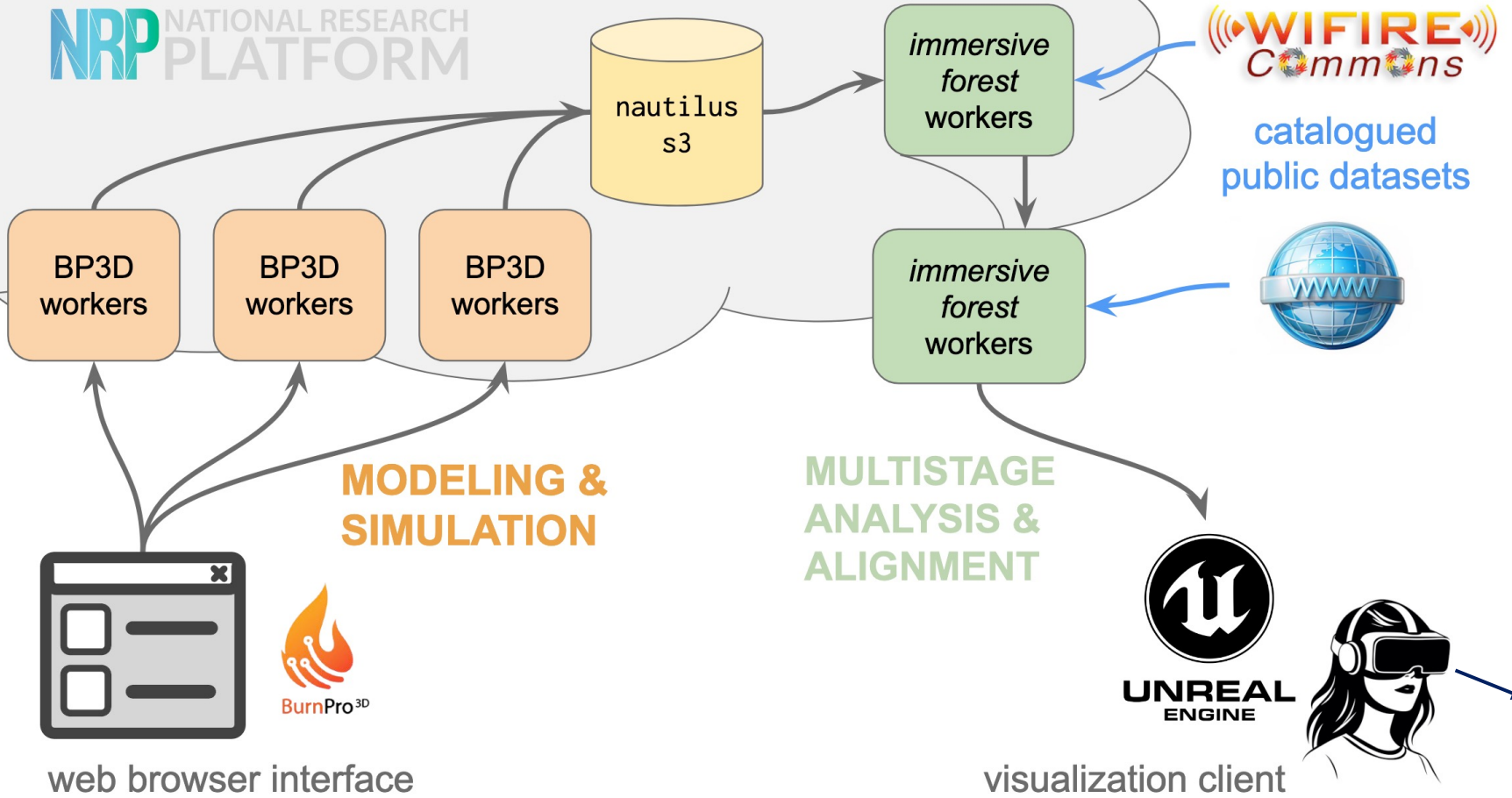
Linear regression models to assess the impact of age, recording location and design iteration

Bathymetry from space

Machine learning model to understand small-scale ocean dynamics

The effect mutations implicated in autism can have in protein oscillation

Deep learning model to predict the oscillation of protein in cell-cell communication



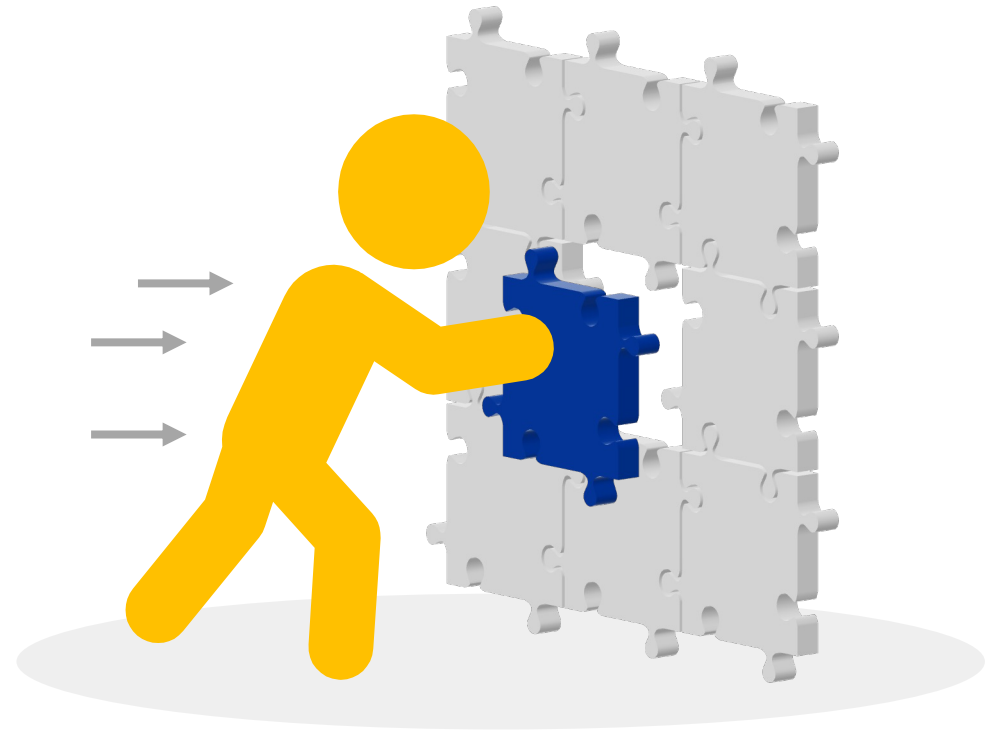
Current prototype capabilities

- Terrestrial LiDAR contextualized with aerial LiDAR for VR
- Georeferenced panoramic projection of terrestrial LiDAR for mobile
- Watch and interact with fire simulations in 3D under a variety of weather conditions
- Move through multiple LiDAR scans across the landscape to compare pre- and post-burn vegetation in 3D



3D Immersive Forest using NRP

What do we do about the data gaps?



<http://www.nationaldataplatform.org>



NATIONAL DATA PLATFORM

Bridging the Data Gaps for AI

UC San Diego



University of Colorado
Boulder

SDSC
SAN DIEGO SUPERCOMPUTER CENTER



EarthScope
Consortium

<http://www.nationaldataplatform.org>



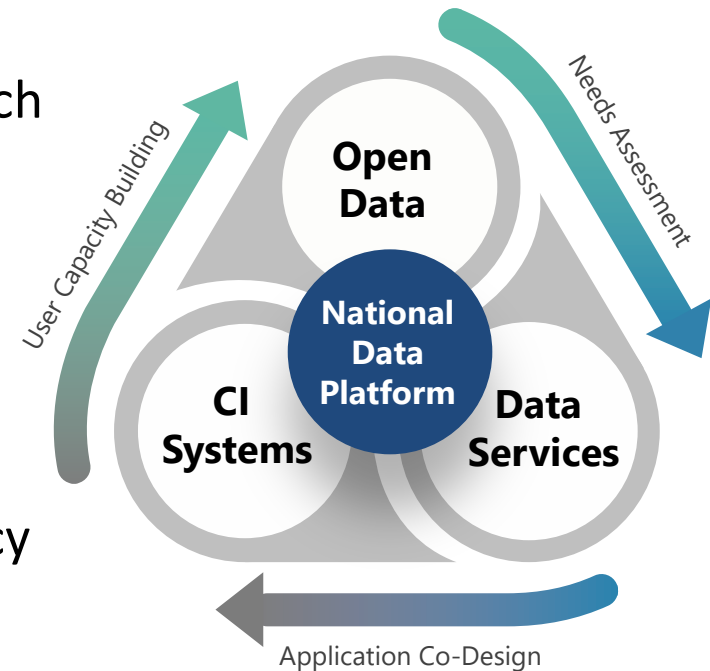
Award abstract: https://www.nsf.gov/awardsearch/showAward?AWD_ID=2333609

National Data Platform Pilot (NDP): Services for Equitable Open Access to Data

A **federated** and **extensible** data ecosystem to promote collaboration, innovation and equitable use of data using existing and future national cyberinfrastructure (CI) capabilities.

- A broad data ecosystem to enable data-enabled and AI-integrated research and education workflows
 - Facilitates data registration, discovery and usage through a centralized hub
 - Enhances distributed CI capabilities through distributed points of presence
 - Cultivates resources for classroom education and data challenges
 - Assists research and learning through personalized workspaces
- Partnership pathways to foster scientific discovery, decision-making, policy formation and societal impact
 - Focus areas: Wildfire, climate, earthquake and food security, among others

46



Addressing Open Questions for Equitable Open Access

Foundational Abstractions and Services

- What are the foundational data abstractions and services that can serve as multipurpose and expandable building blocks for data-driven and AI-integrated application patterns?
- How can everyone effectively access and utilize these abstractions and services?

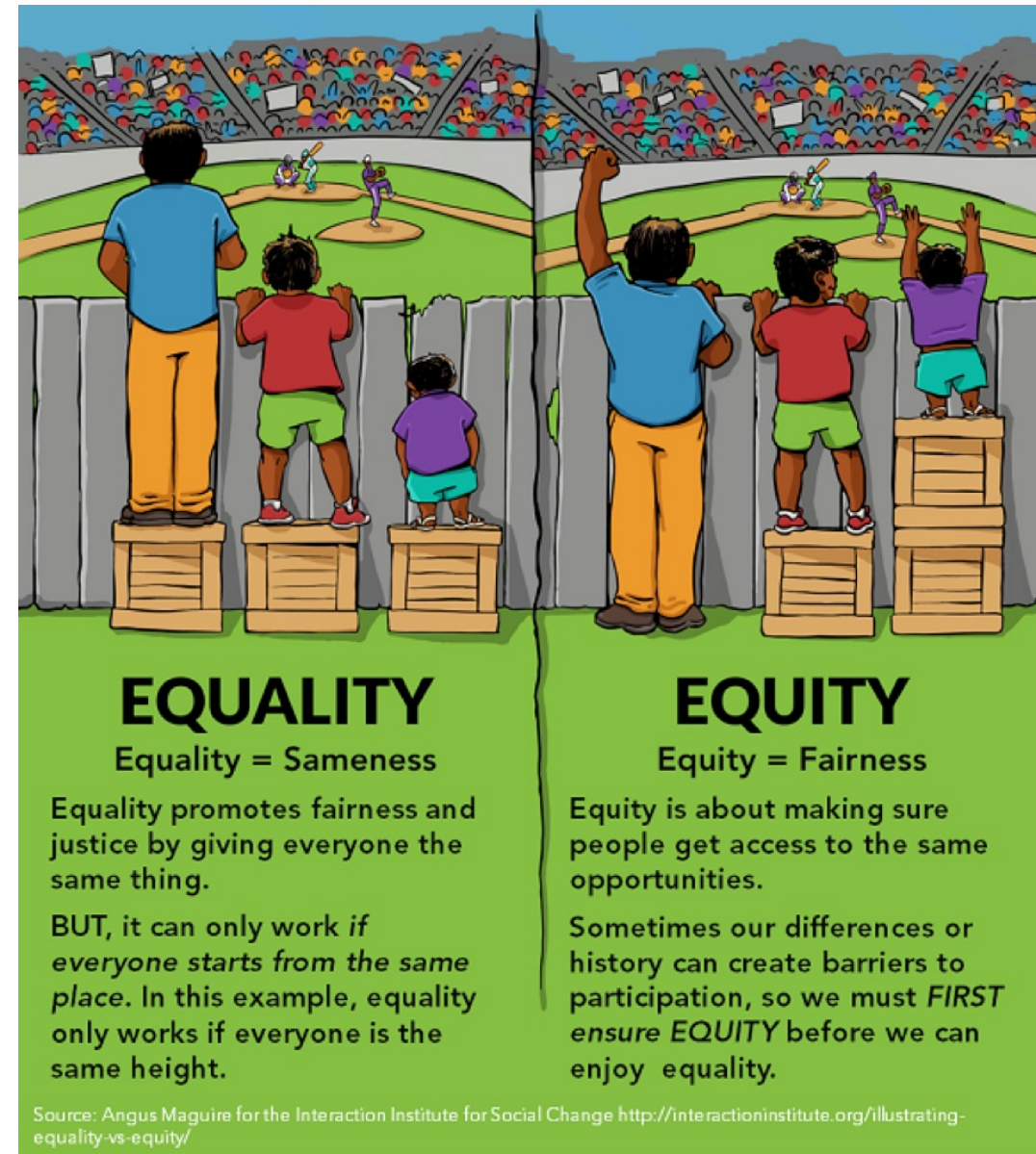
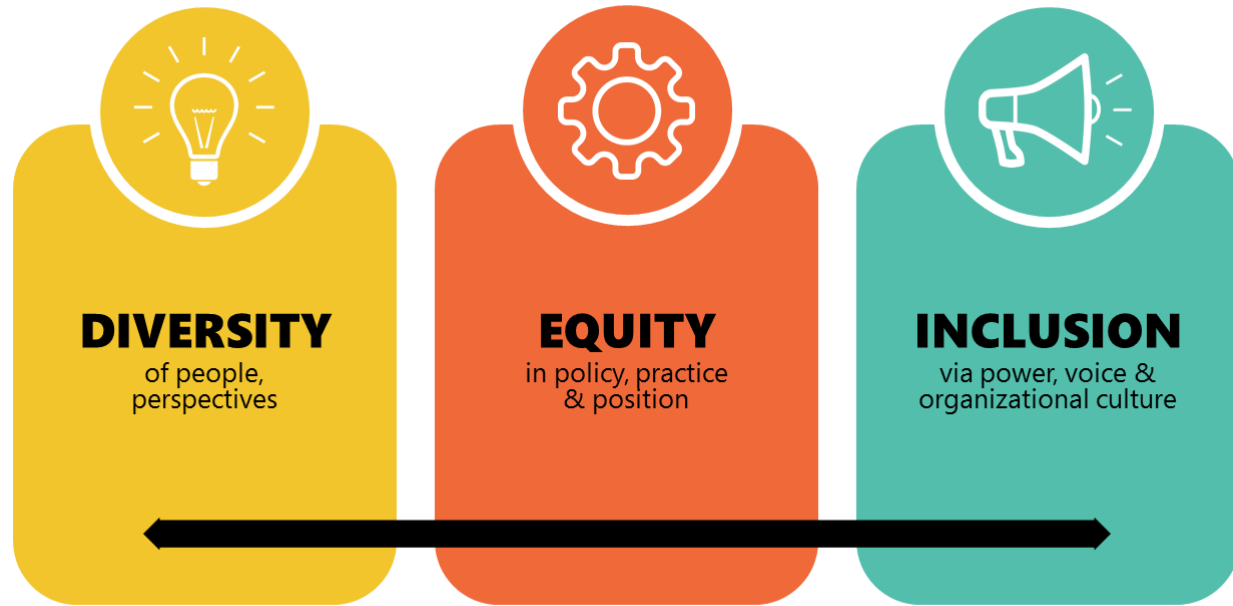
Equitable and Open CI Use

- How can such foundational data abstractions and services be developed and deployed on top of existing production-ready CI, including storage and the edge-to-HPC continuum?
- How can we ensure equity of data access and use across distributed CI?

Needs, Requirements and Challenges

- What are the requirements and challenges for governance of open science, open data and open CI?
- What are the required guardrails for protecting privacy, civil rights and civil liberties that will ensure a more equitable use of data systems and services?

Diversity is a fact.
Equity is a choice.
Inclusion is an action.
Belonging is an outcome.
- Arthur Chan





EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES


FROM: Dr. Alondra Nelson 
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:


1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

The case for open data




Empowering citizens & strengthening accountability

- Promotes more accountability
- Increases citizen engagement



Innovation & efficiency in government agencies

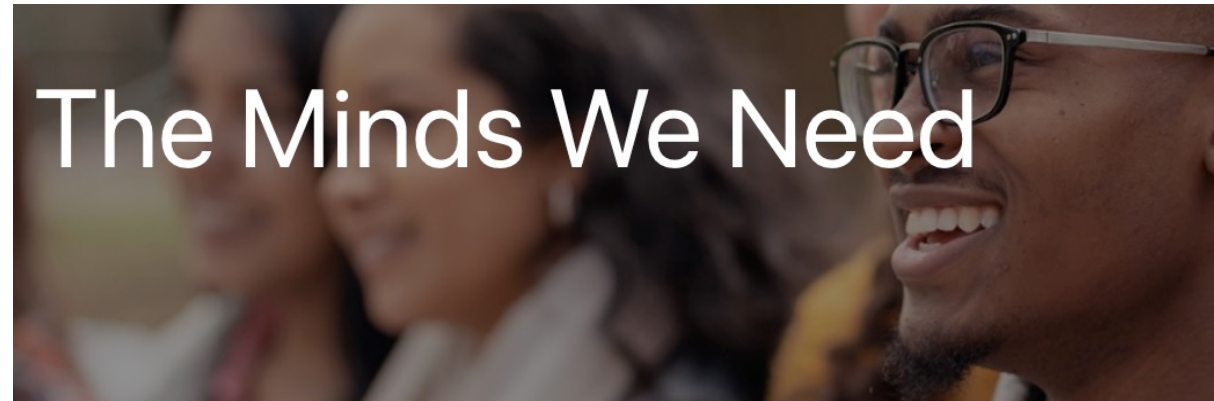
- Decreased workloads
- Inter-agency collaboration
- Improved policy design



Creating wider value for the economy

- Open data creates value added services for the entire economy

OECD



The Minds We Need

Inclusion, Innovation, and Competitiveness | Strengthening Our National Broadband Initiative | Investing in Research and Education Infrastructure | Contributors | Toolkit | Endorsements

Inclusion, Innovation, and Competitiveness

We are at a crossroads.

<https://mindsweneed.org>

Toward Democratizing Access to Facilities Data: A Framework for Intelligent Data Discovery and Delivery

Yubo Qin, Rutgers University, New Brunswick, NJ, 08901, USA
Ivan Rodero and Manish Parashar, University of Utah, Salt Lake City, UT, 84112, USA

Data collected by large-scale instruments, observatories, and sensor networks (i.e., science facilities) are key enablers of scientific discoveries in many disciplines. However, ensuring that these data can be accessed, integrated, and analyzed in a democratized and timely manner remains a challenge. In this article, we explore how state-of-the-art techniques for data discovery and access can be adapted to facilitate data and develop a conceptual framework for intelligent data access and discovery.

The Missing Millions

Democratizing Computation and Data to Bridge Digital Divides and Increase Access to Science for Underrepresented Communities

October 3, 2021
NSF OAC 2127459

Democratization of CI and Data Access

The background of the slide features a network of colorful ropes (red, yellow, purple, green, blue, pink, grey, orange) that are knotted together in various ways, symbolizing interconnectedness and collaboration. The ropes are set against a light grey background.

Architecting for Collective Data-Integrated Impact

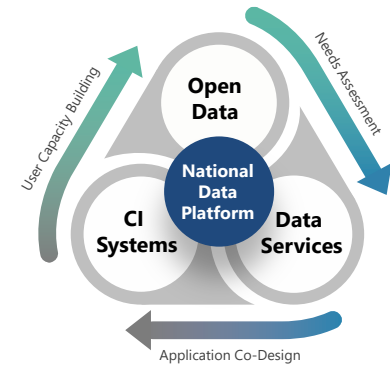
- Involve diverse users in architecting
- Identify access, use, expertise and education gaps
- Improve the experience of working with data
- Connect data to knowledge systems and services
- Create an ecosystem approach to capacity building
- Incubate use-inspired solutions to scale
- Explore new models of allocation
- Develop and teach models of sustainability and scale



NATIONAL DATA PLATFORM

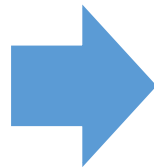
Our Use-Inspired Approach

Solving data gaps one workflow template at a time...



Identify Gaps

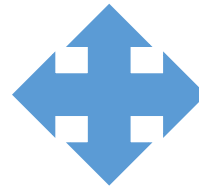
- Community advisory board
- External community integration plan
- Needs assessments
- Co-design workshops
- Expansion prototypes



Incubate, Innovate and Educate

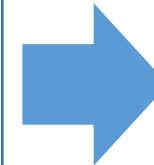
Use-Inspired Workflows and Interfaces

Data and Knowledge Management



Composable Services

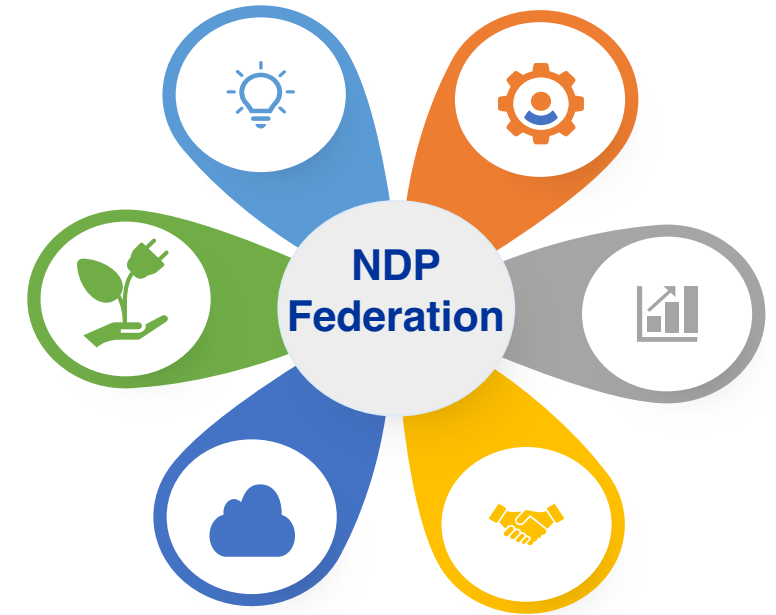
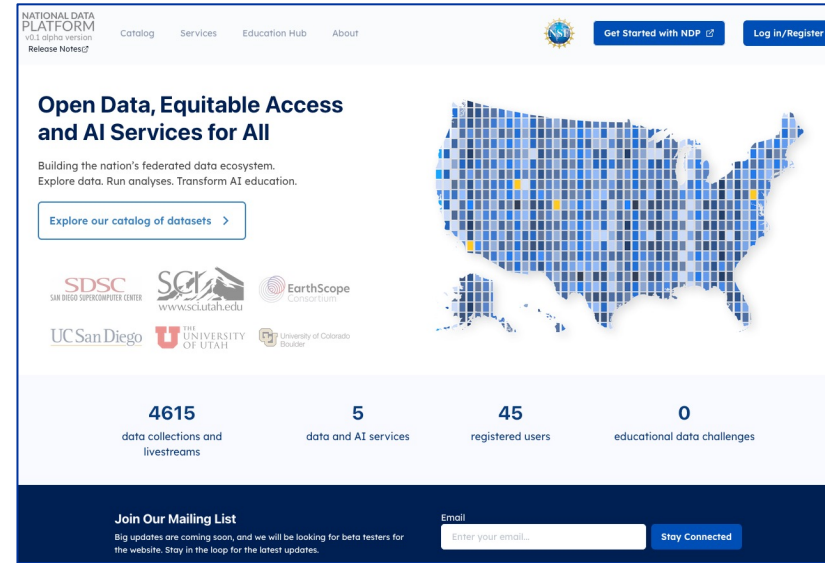
Composable Systems and Platforms



Sustainable and Scalable Use

- Distributed in nature
- Composition as a principle
- Hub-centric services as connection backbone
- Integrates in education systems



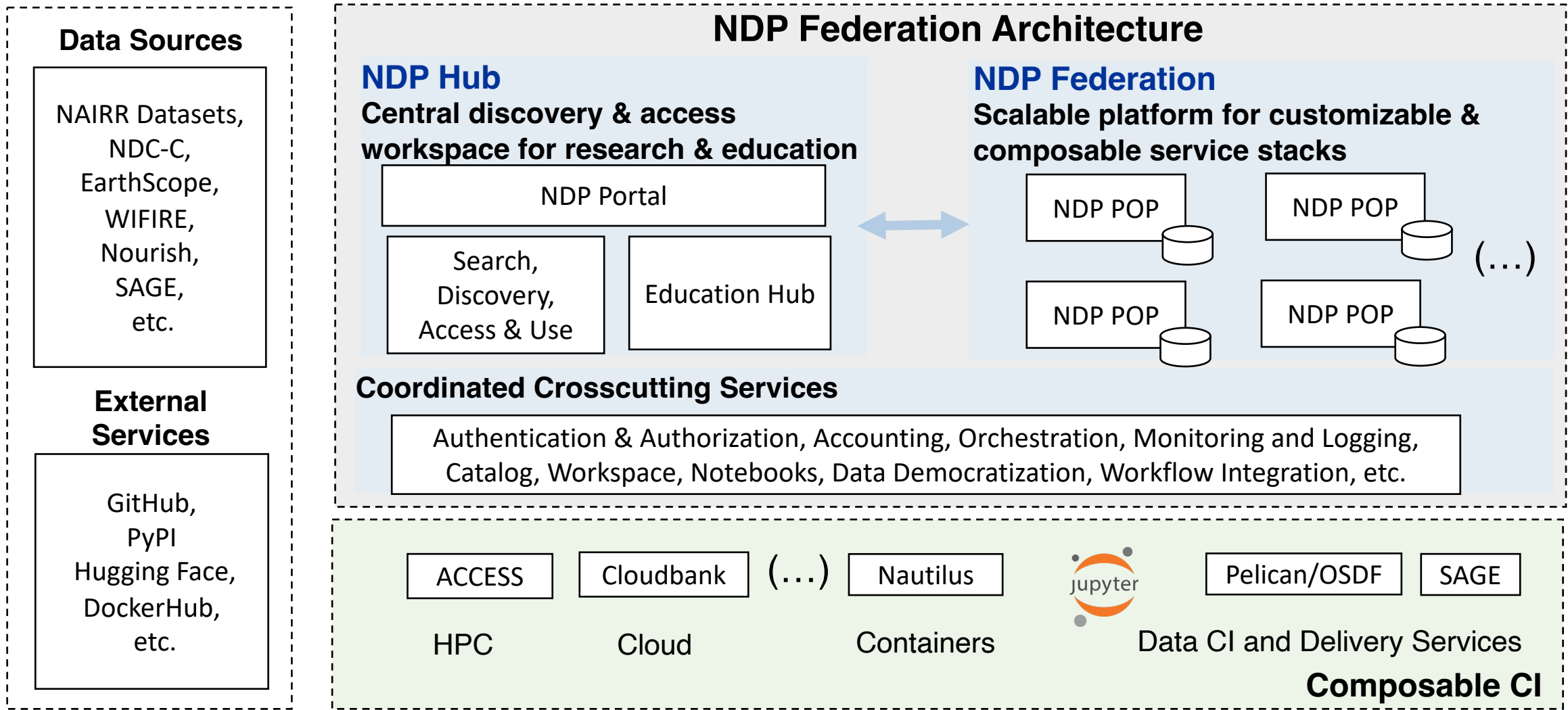


Centralized portal for discovery, access and use workspaces for research and education



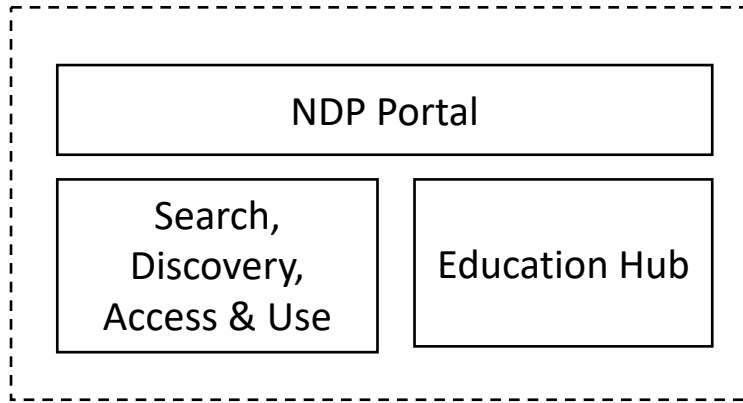
A scalable **platform** for using, developing and deploying services and application workflows at **distributed points of presence**

Current NDP Overarching Architecture



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

NDP Hub



- NDP Portal (point of access)
<https://nationaldataplatfrom.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

Public:

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

Login-enabled:

- 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

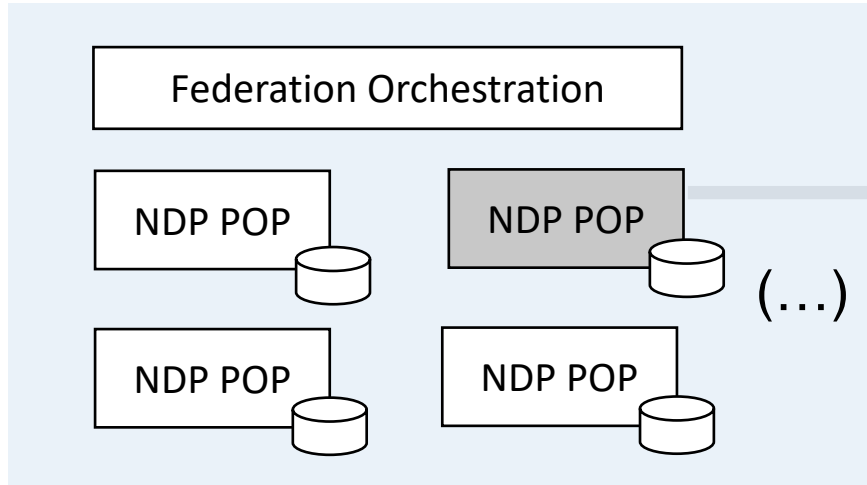
- Sage Data and Edge Code Integration Service
- Service Catalog and Discovery Service
- Educational Hub Expansion
- Streaming Data Services
- Pelican Registration Service
- Integrated Workflows

Planned Future Work

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



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

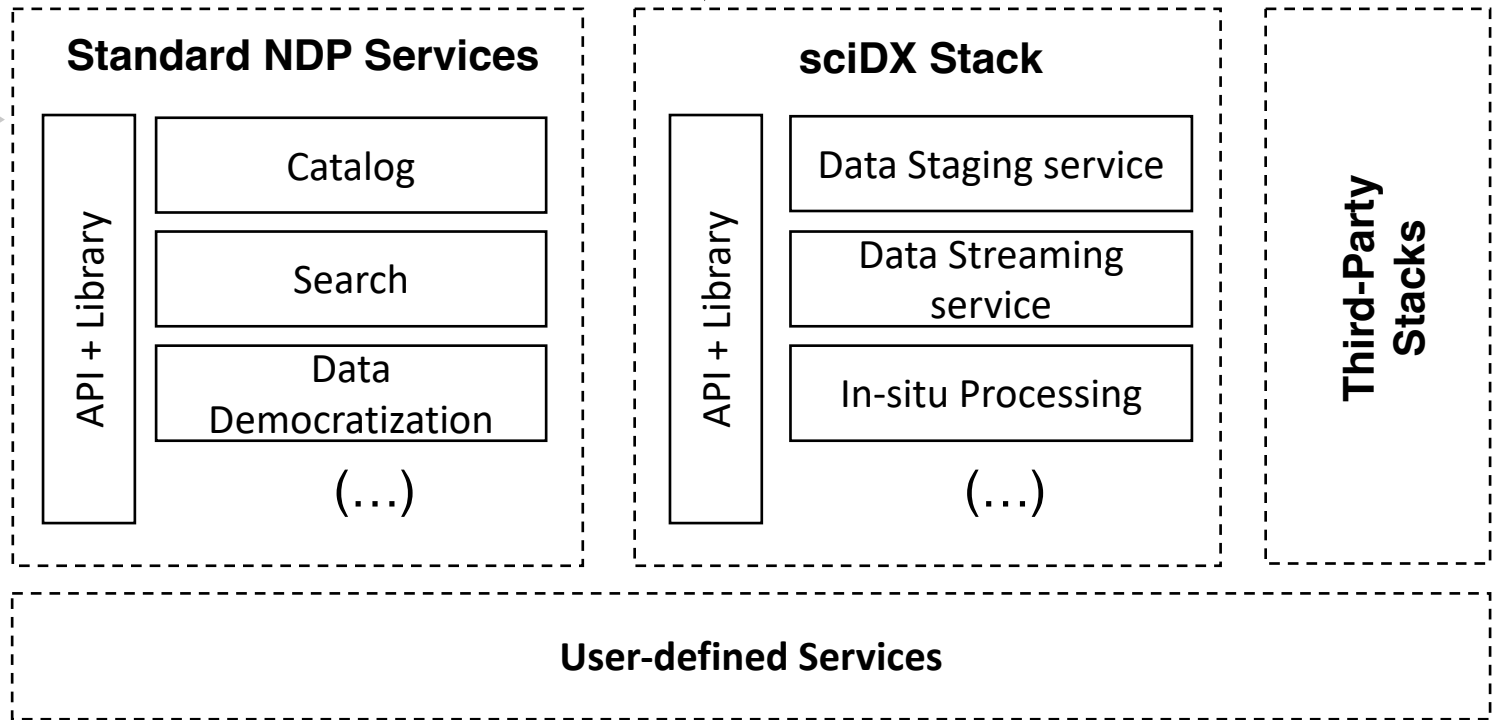
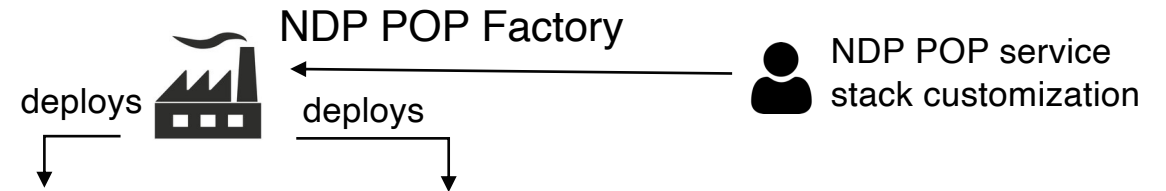
NDP Federation



Workflow composition currently via API and/or Python client library



 scidx 0.2.0 Python client library for interacting with the sciDX API	Aug 2, 2024
 scidx-tools 0.1.0 Python client library for complementing the sciDX library	Jul 1, 2024



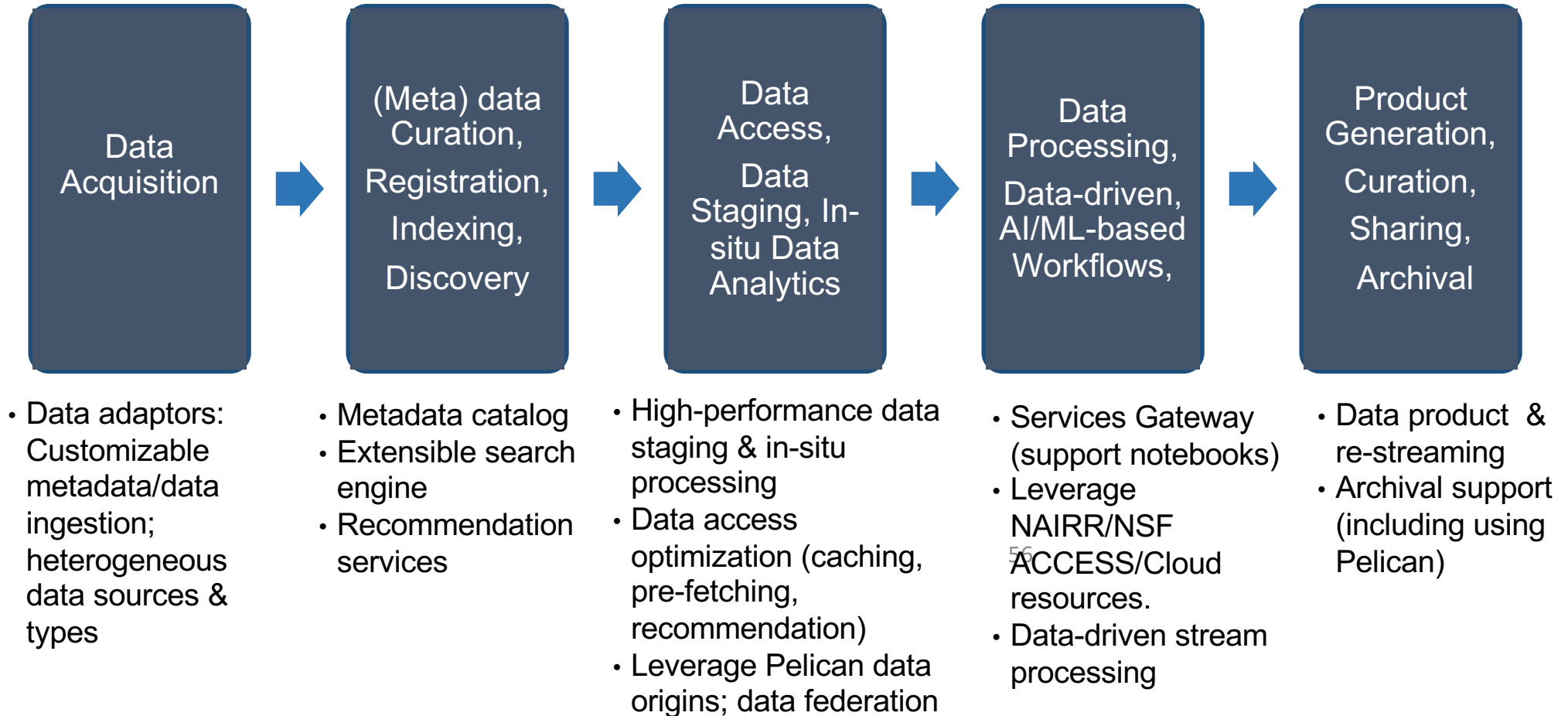
Deployment models:  **docker**  **kubernetes** **Cloud** **HPC**

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

Typical NDP Workflow with Composable Capabilities



Data sources:
NAIRR datasets, repositories, instruments, sensors, facilities, etc.

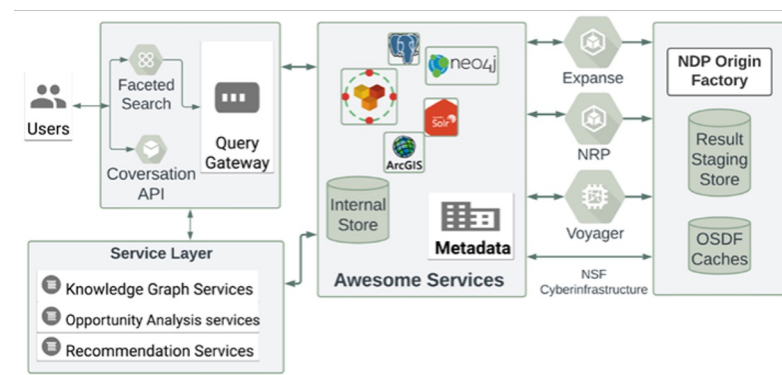
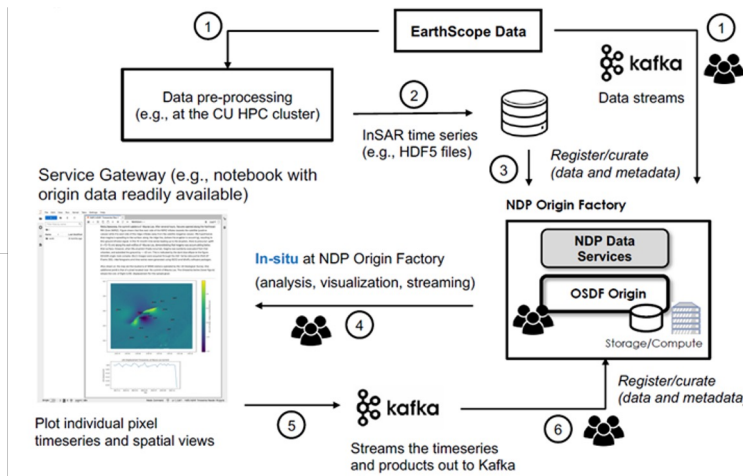
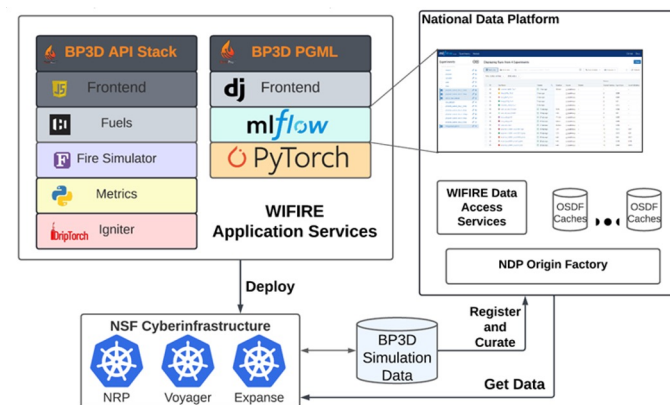




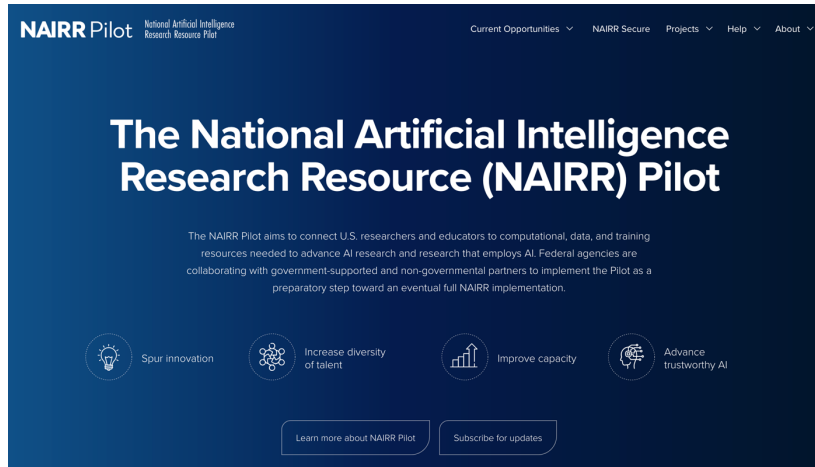
Case Studies for Generalizable Workflows

NATIONAL DATA PLATFORM

- **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.



Planned Extensions for NAIRR (September 2024 – August 2025)



NAIRR Data Resource Catalog

- Ingestion Process for NAIRR Data
- FAIR NAIRR Catalog
- Conversational Search Interfaces

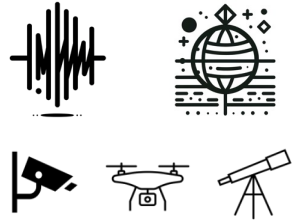
NAIRR CloudBank Research Workflows

- Provisioning and Accounting
- CloudBank Workflow Deployment
- Collaborate with NAIRR science pilots

NAIRR Classroom Workflows

- NAIRR Educator Workflows
- NAIRR Student Workflows
- Community Engagement

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 .

NAIRR Pilot National Artificial Intelligence Research Resource Pilot

Open Data, Models, and More

This list does not include allocatable resources for research or education/teaching; please see the [Researcher Call](#) and [FAQ](#).

AI Courses Datasets Documentation Models Secure Other

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 Search models, datasets, users...

ibm-nasa-geospatial | **Prithvi-100M** | 235 likes

Model card | Files and versions | Community

Model and Inputs

Prithvi is a first-of-its-kind temporal Vision transformer pre-trained by the IBM and NASA team on contiguous US Harmonized 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.

NATIONAL DATA PLATFORM v0.5 alpha version

Search Data

Substring search: Select Org.:

1 - 5 of 3 Data Collections and Streams

- HLSYPWLY_00** by EarthScope Consortium
Resources: 2
Station Code: HLSY, Network Code: PW, Location Code: LY, Channel Code: 00, Latitude: 44.37754, Longitude: -123.1912006
- 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.

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name	Last Modified
configs	3 months ago
data_splits	3 months ago
geospatial_...	3 months ago
geospatial_...	3 months ago
hls_gfn	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_info...	3 months ago
README.md	3 months ago
setup.py	3 months ago

UC San Diego **SDSC** SAN DIEGO SUPERCOMPUTER CENTER

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 mmssegmentation 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

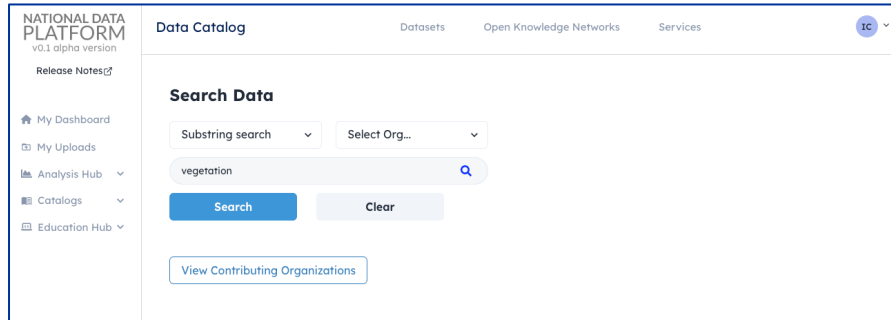
This work builds upon the original exploration notebook developed by the NASA IMPACT team. For detailed citation information, please refer to the CITATION.cff file in this directory.

NDP Hub Functionality

September 2024 Release

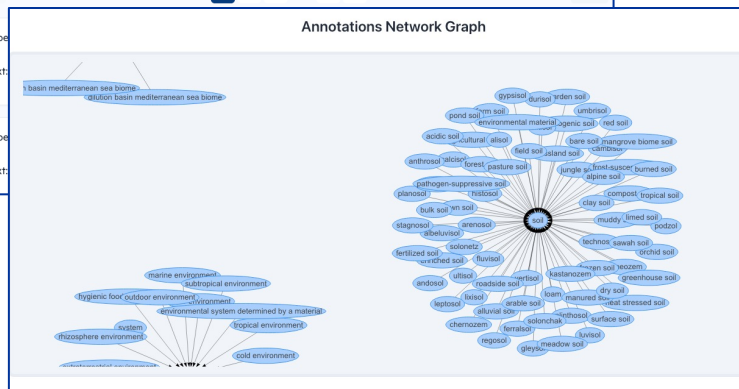
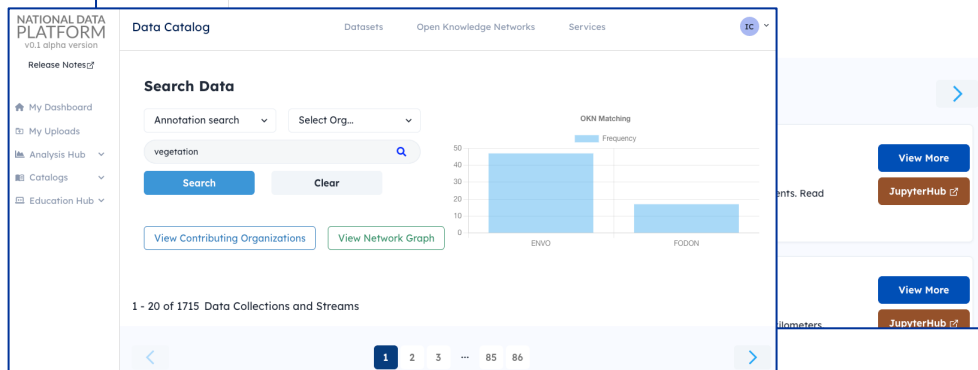
60

NDP Hub: Data Search and Discovery



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.



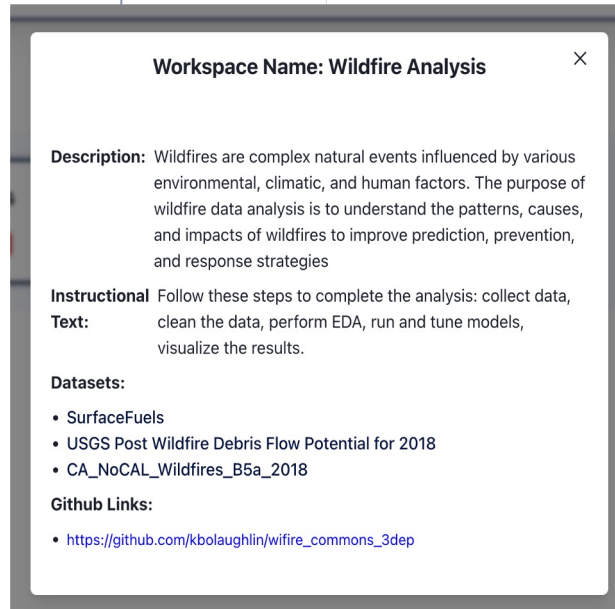
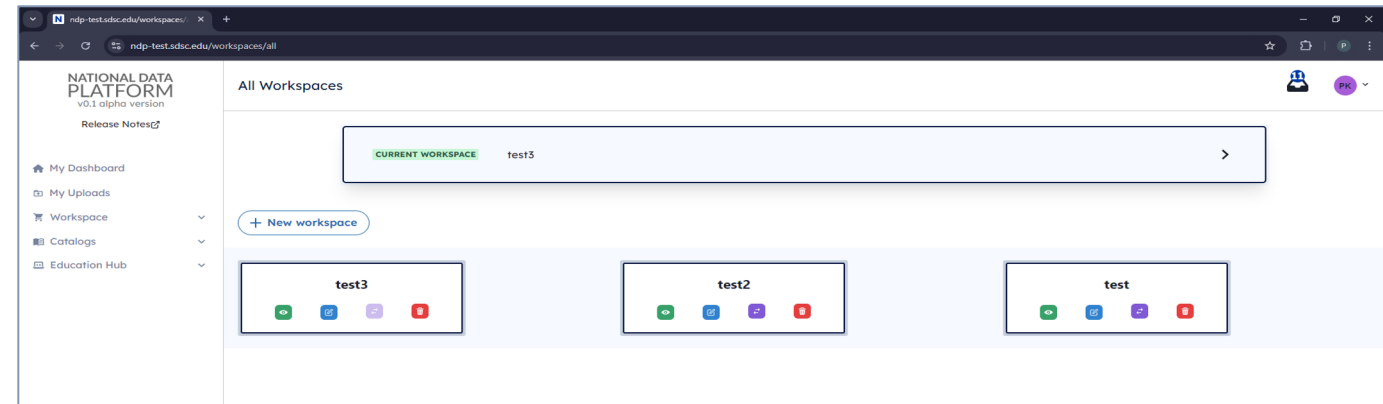
Future Work Post-September 2024 to include NAIRR Data Resources:

- 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



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.

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

Start

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

NDP JupyterHub Interface

Launcher

Current folder: hls-foundat ion-os

Filter

Create Empty

Launch New Notebook

Launch New Console

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 Catalog Addition

Goal: Users can add dataset references to either NDP centralized catalog or POP-specific catalog

Curated Public Catalog Add Request:

- Provide all metadata and data access information
- Designated data approvers evaluate dataset quality
- Add or reject datasets for access to community

My Uploads

[+ Register dataset](#)

User Actions	Title	Org	Visibility	Description	Requester	Admin Actions
 	Food and Agriculture Ontology	OBO Foundry	Public	FoodOn is an ontology – a controlled vocabulary which can be used by both people and computers –...	Elaine Chi ychi@ucsd.edu	 
 	USDA 2022 Branded Food Product Catalog	USDA	Public	This database contains approximately 1.7 million food products that are sold on the shelves of the...	Elaine Chi ychi@ucsd.edu	 

My Uploads

Catalog Add Request Edit Form

[← Go back](#)

Title *
Test Dataset

Description *
my dataset has data about the data that contributes data to the data filled world of data. my dataset has data about the data that contributes data to the data filled world of data.

Tags (separated by ',')
wildfire,fire,trees

Organization
WIFIRE Commons
 Other

Visibility *
Public

Point of Contact Details *
Author: Katie
test@test.test 12345678900234

Public Key
Enter public key...

Version

NDP Data Challenges for students and researchers

Designed to ensure that we are developing broadly accessible services for equitable education and community building.

NDP Education Gateway to provide participants access to the NDP data ecosystem

The challenge questions will require using data and models in an environment that requires computing and huge data stores, which would typically be unavailable to a student or researcher without the NDP Education Gateway.

Three Co-Design Workshops

Each will include a breakout session to develop a data challenge question specific to large data (W1); streaming data (W2); and graph data (W3).

Education and capacity building through data challenges



NATIONAL DATA PLATFORM

Data challenge toolkits will be developed after each data challenge so that other institutions can easily design their own data challenges to be run through the NDP Education Gateway.

NDP Education Hub (Version 1 – September 2024)

The screenshot shows the 'Education Hub' section of the National Data Platform. On the left is a navigation sidebar with 'Education Hub' selected. The main content area is titled 'Education Hub' and 'My Modules'. It features two buttons: 'Add New Module' and 'Use Module Templates'. Below these is a module card titled 'Training AI models with fire data and the newest Quicfire model' with an 'In Progress' status. The card includes a description: '... particularly for the Physics Guided Machine Learning (PGML) research and educational tasks. It is an ensemble of prescribed fire simulations generated by the QUIC-Fire coupled fire-atmospheric modeling tool. Each simulation run is represented by a...' and an 'Edit Module' button.

New Learning Module Development Wireframes

The wireframe shows a page for a learning module. The title is 'Semantic Segmentation using NASA Harmonized Landsat and Sentinel-2 data'. It includes a 'View to the Public' link, a 'By' field, and tags for 'Deep Learning' and 'Transformers'. The main text describes the 'Harmonized Landsat and Sentinel-2 (HLS) project' and the 'Prithvi-100M' model. It lists 'Objectives', 'Target Audience', 'Prerequisites', 'Instructions', and 'Additional Resources'. At the bottom, there is a section for 'HLS Burn Scars' with a data icon.

Education Hub MP

Education Hub > Create Module > **Module Maker Wizard**

General Add Data Add Models Add Scripts **Preview**

View to the Public

Semantic Segmentation using NASA Harmonized Landsat and Sentinel-2 data

By [Redacted]

Deep Learning Transformers

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

In this learning module, you'll gain hands-on experience with Harmonized Landsat and Sentinel-2 (HLS) data to explore the powerful capabilities of the Prithvi-100M model. Through guided demonstrations, you'll see how this state-of-the-art temporal Vision Transformer can be applied to real-world scenarios.

Objectives

- Highlight the advanced functionalities of the Prithvi-100M temporal Vision Transformer, pre-trained on contiguous HLS data, through its application in three specific semantic segmentation tasks: burn scar detection, flood mapping, and multi-temporal crop classification.
- Implement and validate one of the provided fine-tuning use cases (burn scar detection), using the `mmsegmentation` library to ensure the robustness and adaptability of the Prithvi-100M model for domain-specific applications.
- Provide a methodology and foundational code base to support the development of novel fine-tuning use cases leveraging the Prithvi-100M model.

Target Audience

- Earth Science Researchers
- Grad Students
- Remote Sensing Specialists

Prerequisites

- Python Programming
- Machine Learning
- Remote Sensing
- Deep Learning Fundamentals

Instructions

- [README.txt](#)

Additional Resources

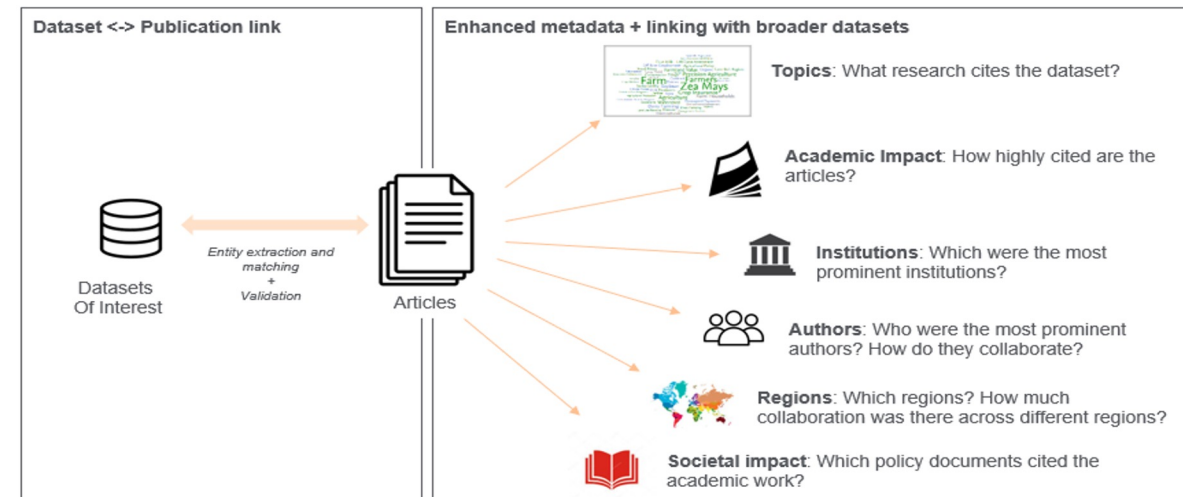
- <https://hls.gsfc.nasa.gov/>

HLS Burn Scars

Democratizing Data (DD) Service

- Composable DD service for search and discovery within NDP requires a detailed and structured approach
 - Extracting and utilizing publication metadata using multiple corpora (e.g., Scopus, OpenAlex, PubMed)
 - Integrating NAIRR datasets starting with USDA, NIH, NASA, and NOAA.
 - Exploring a generalized approach to support the integration to other corpora and AI-ready NAIRR datasets
- Leveraging <https://democratizingdata.ai/>

The screenshot shows the USDA: Democratizing Data Dashboard. At the top, there's a navigation bar with 'Agencies', 'Events', 'Our Tools', 'Resources', and 'About'. Below that is a purple header with the title 'USDA: Democratizing Data Dashboard'. The main content area includes a 'Stats at a glance' section with icons for Publications (8,525), Journals (2,497), Institutions (6,540), and Authors (29,639). There are also buttons for 'Publications & Journals' and 'Usage Over Time'. Below this is a 'CLEAR FILTERS' button and a 'Select a view' dropdown. The central part of the dashboard is a table titled 'List of publications in which the data asset is cited'. The table has columns for 'Publication', 'Citations', and a 'Downloads' link. The table lists several publications with their respective citation counts. On the left side, there's a sidebar with 'About the dashboard', 'Sort filters by', 'Select Classification and Research Areas', and a 'Word Cloud'.



NDP PoP Examples & Documentation

September 2024 Release

68

Science Data Exchanges (sciDX) Services: Data Staging and Streaming Services

Science Data Exchange (sciDX): Customizable software stack for in-situ data access & processing

Data Staging Service

- In-situ (close to the data) data processing and access
- High-performance in-memory processing
- Server-side data transformations (e.g., subsetting, reduction, user-defined analysis, etc.)
- Caching/sharing of data, query results, and data products with user and group isolation

Data Streaming Service

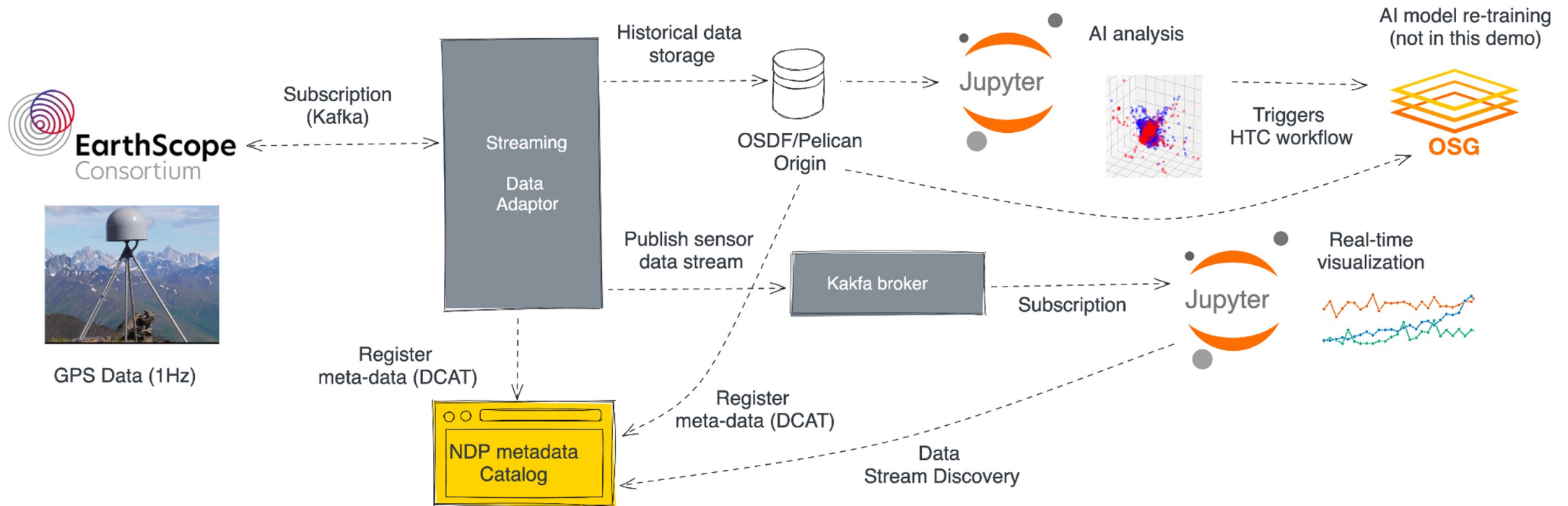
- Streams registration, curation/archival for discovery and access
- User-defined operations on streaming data (semantically specialized abstractions)
- Combine streaming data with archived/playback data
- Mechanism for online data product generation (i.e., new data streams)

In-situ AI workflow execution runtime (on staged and streaming data)

Example 1: EarthScope data streaming/analysis enabled by NDP POP

Real-time high-precision GNSS stations

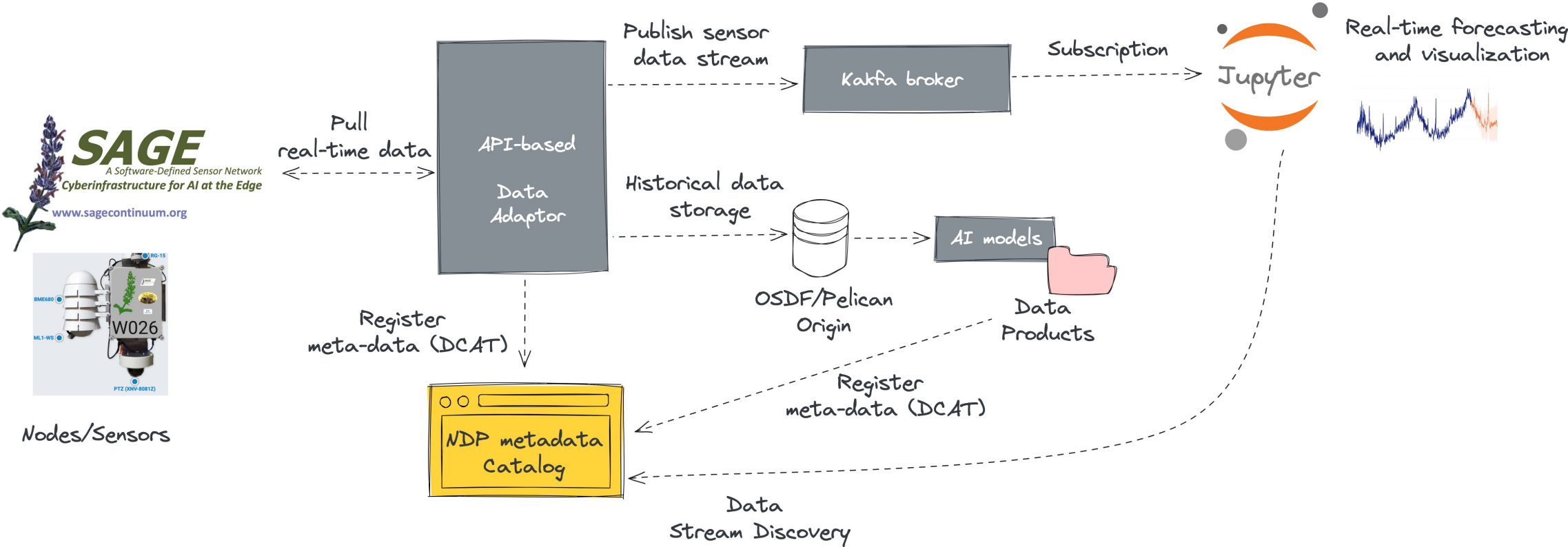
- AI analysis: anomaly detection from archived data from OSDF/Pelican
- Real-time data visualization



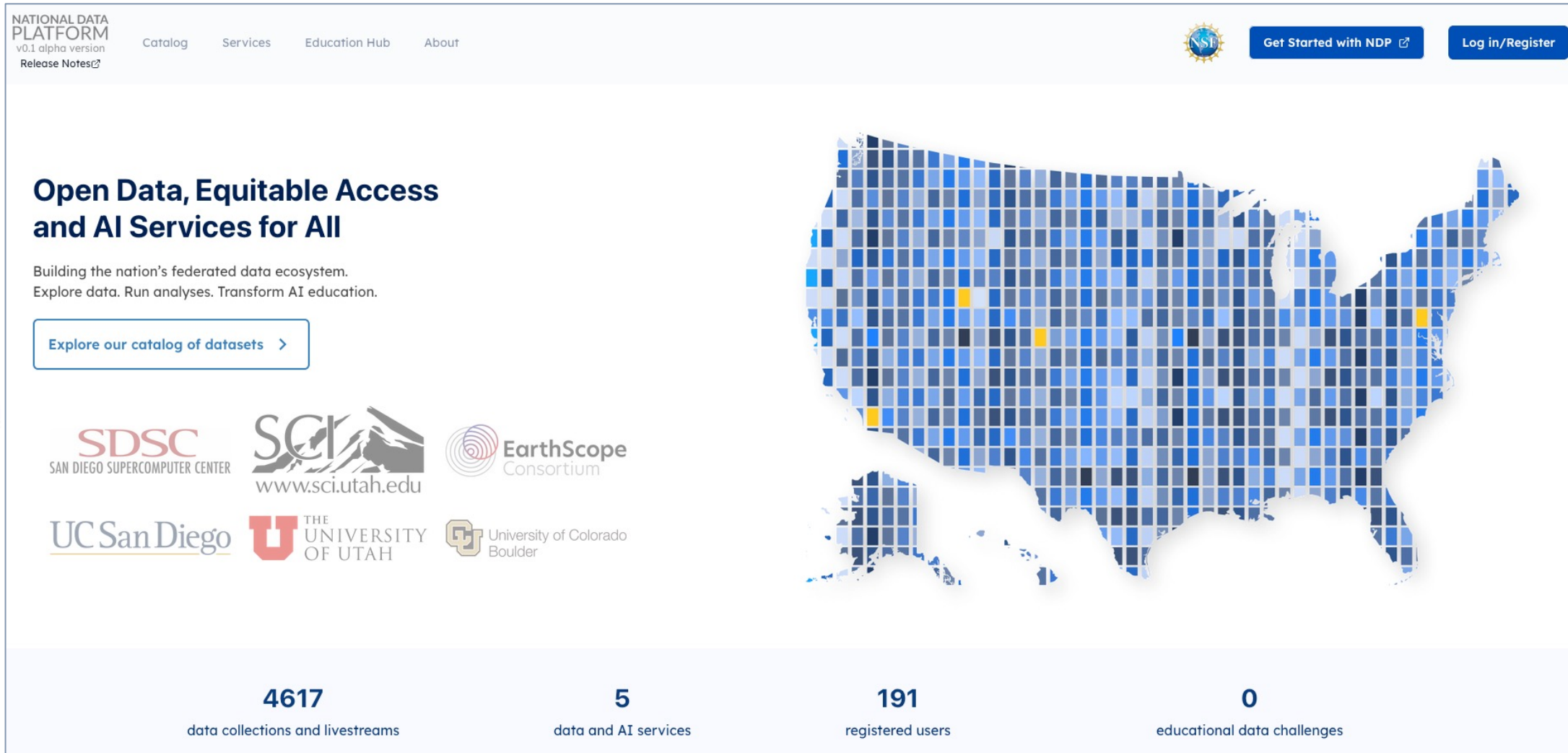
Example 2: SAGE data streaming/analysis enabled by NDP

SAGE data streams

- Real-time data visualization (temperature)
- Time series forecasting (proof-of-concept)



Any questions? Contact ndp@sdsc.edu



The screenshot shows the homepage of the National Data Platform (NDP). At the top left, it says "NATIONAL DATA PLATFORM v0.1 alpha version" with a "Release Notes" link. Navigation links include "Catalog", "Services", "Education Hub", and "About". On the top right, there is an NSF logo, a "Get Started with NDP" button, and a "Log in/Register" button. The main heading is "Open Data, Equitable Access and AI Services for All". Below this, it states "Building the nation's federated data ecosystem. Explore data. Run analyses. Transform AI education." and includes a button "Explore our catalog of datasets". A large map of the United States is shown on the right, composed of a grid of blue and yellow squares. At the bottom, four statistics are listed: 4617 data collections and livestreams, 5 data and AI services, 191 registered users, and 0 educational data challenges. Logos for SDSC, SCI, EarthScope Consortium, UC San Diego, The University of Utah, and University of Colorado Boulder are displayed in the middle section.



Artwork: **Jen Stark, Cosmographic, 2014**, acid-free paper, holographic paper, glue, wood, acrylic paint, 34 x 37 x 4 in.

To sum up...

Emerging new applications require integrated AI in dynamically composed workflows.



Complexity comes at a cost

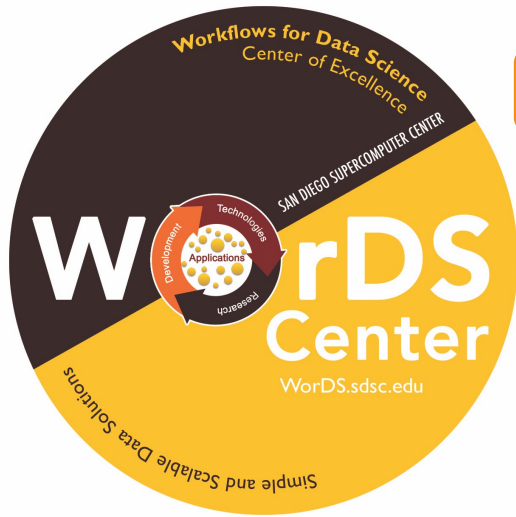
- Composable systems is not a turnkey functionality
- Requires collaboration with and between infrastructure providers

Convergence research helps

- End-to-end data pipelines need to be defined for each application along with microservice execution
- Use-inspired design and translational CS helps to focus the effort

Contact: Ilkay Altintas, Ph.D.

Email: ialtintas@ucsd.edu



<https://words.sdsc.edu/>

<https://wifire.ucsd.edu/>



We are hiring!
https://www.sdsc.edu/about_sdsc/careers.html

Questions?



Office of Science



The presented work is collaborative work with many wonderful individuals, and parts of it are funded by various government agencies, UC San Diego and various industry, government and foundation partners.