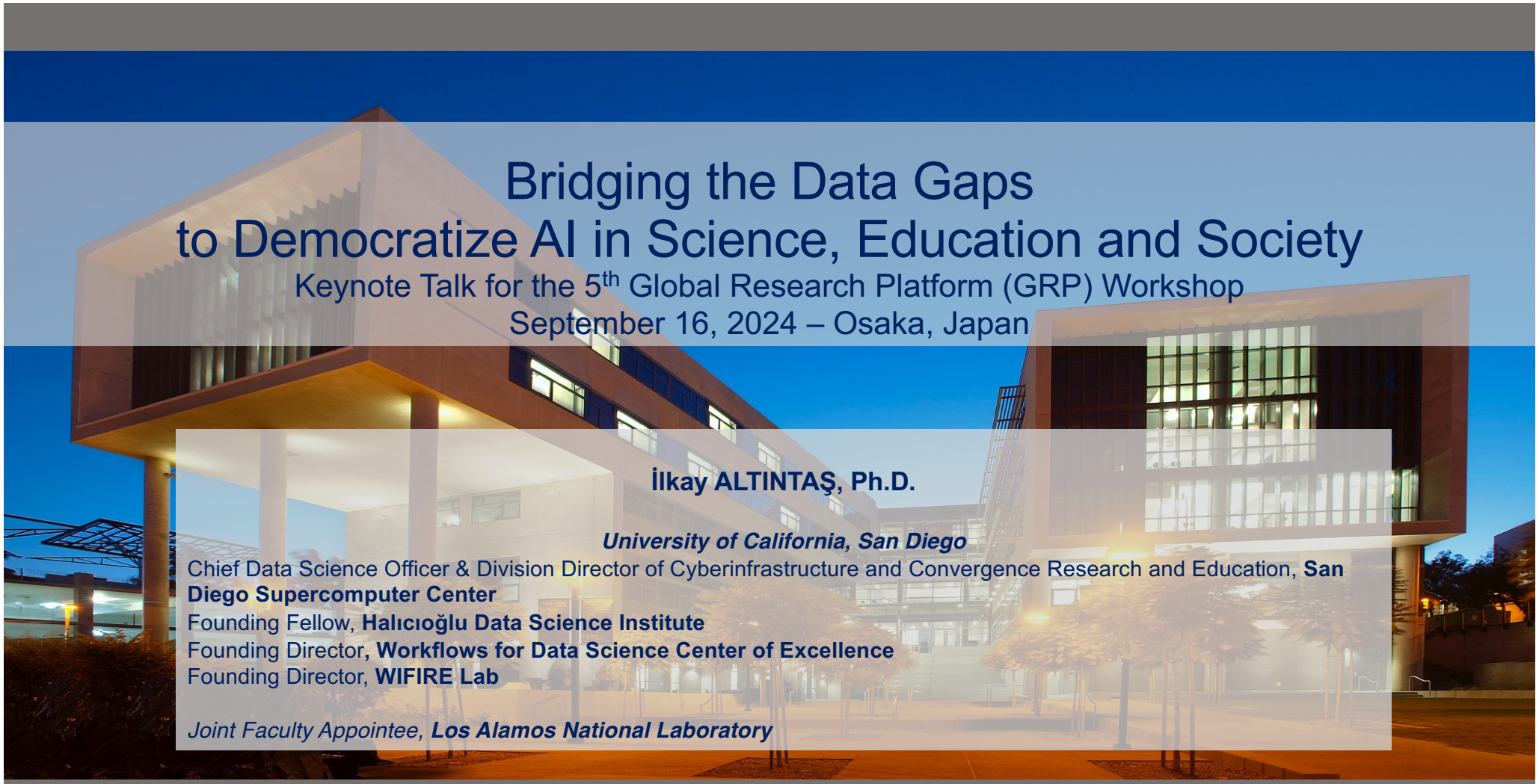


Bridging the Data Gaps to Democratize AI in Science, Education and Society

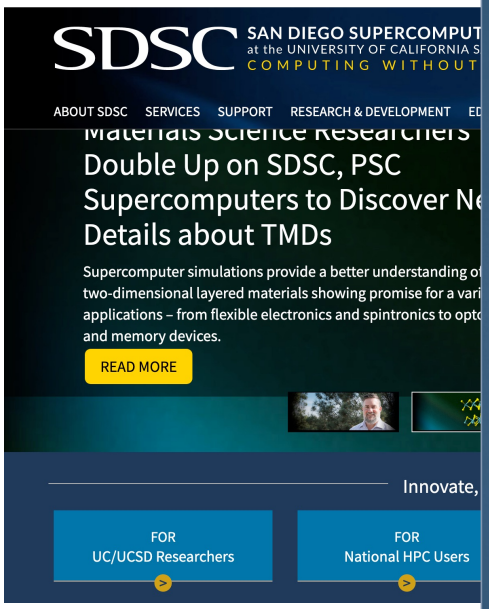
Keynote Talk for the 5th Global Research Platform (GRP) Workshop
September 16, 2024 – Osaka, Japan

İlkay ALTINTAŞ, Ph.D.

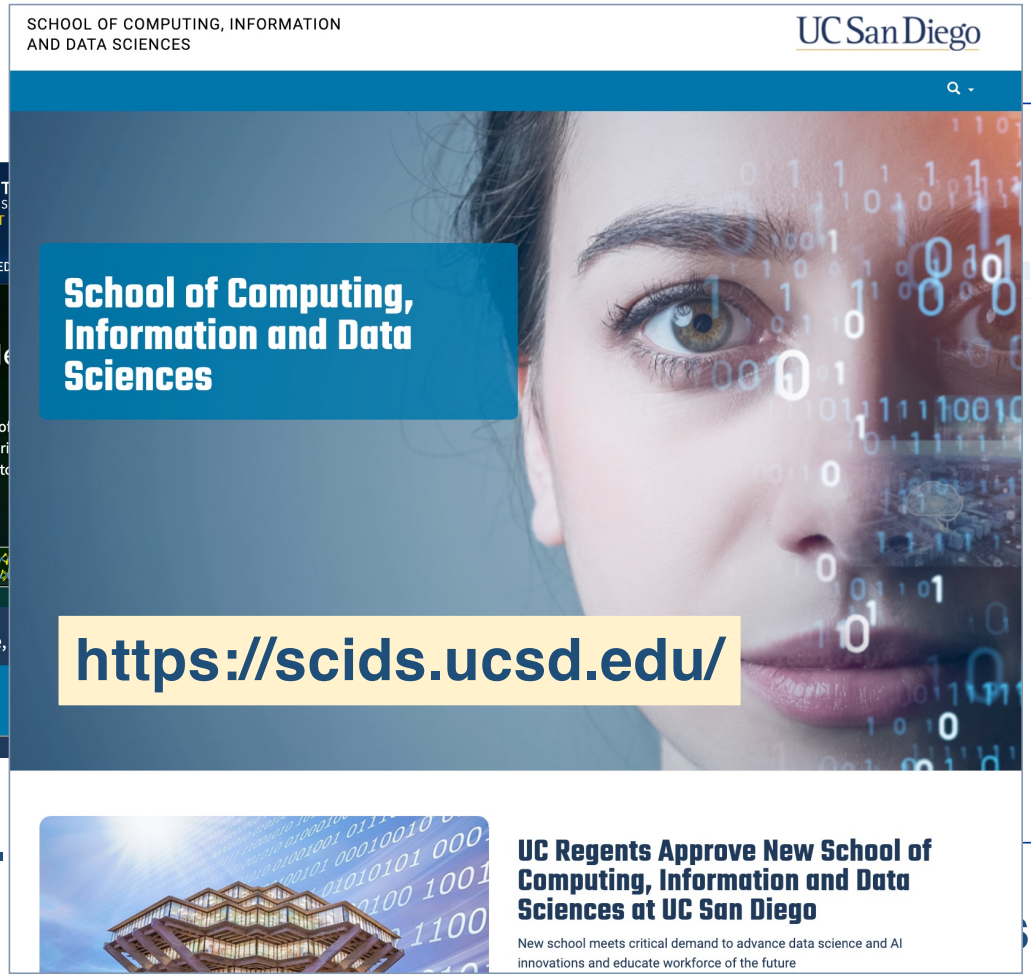
University of California, San Diego

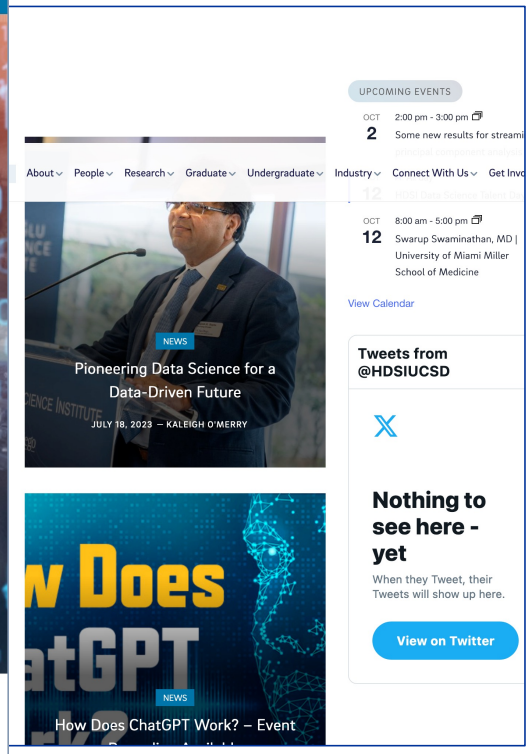
Chief Data Science Officer & Division Director of Cyberinfrastructure and Convergence Research and Education, **San Diego Supercomputer Center**
Founding Fellow, **Halıcıoğlu Data Science Institute**
Founding Director, **Workflows for Data Science Center of Excellence**
Founding Director, **WIFIRE Lab**

Joint Faculty Appointee, ***Los Alamos National Laboratory***



https://www.





science.ucsd.edu/

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Cyberinfrastructure and Convergence Research Division @SDSC

Translating cyberinfrastructure research for impact at scale

- "Big" Data
- Computational Science
- Data Science
- Cyberinfrastructure
- Collaborative Problem Solving
- Convergence Research
- Experiential Education

CICORE

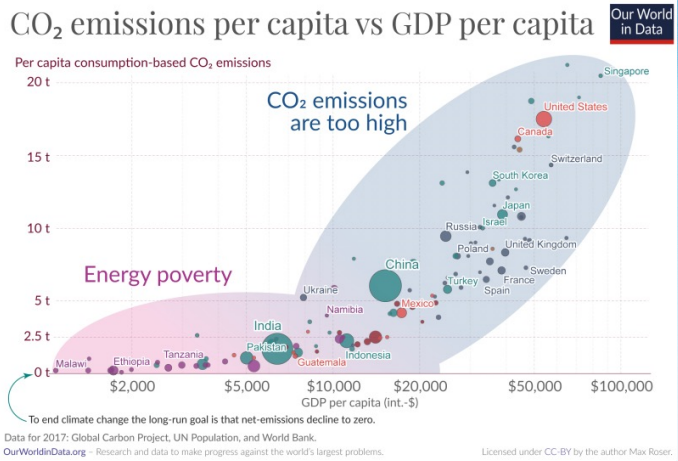
Cyberinfrastructure | Convergence Research | Education

SDSC SAN DIEGO SUPERCOMPUTER CENTER

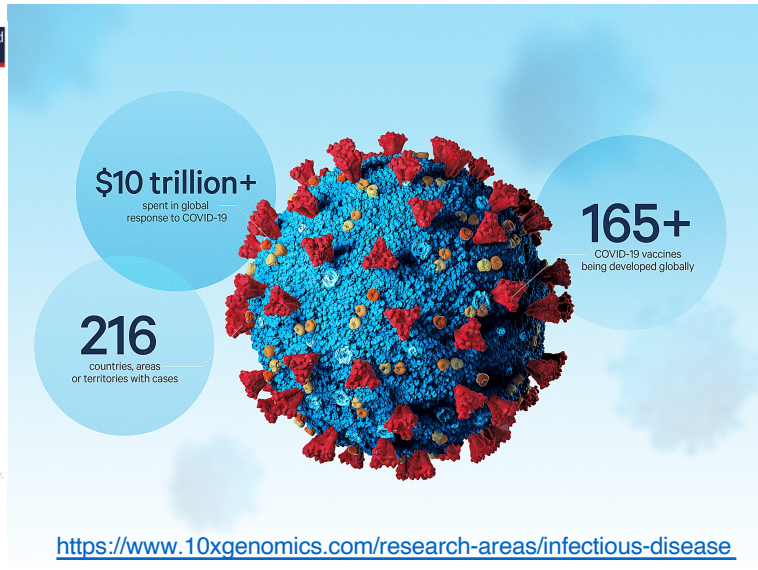
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

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



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



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

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

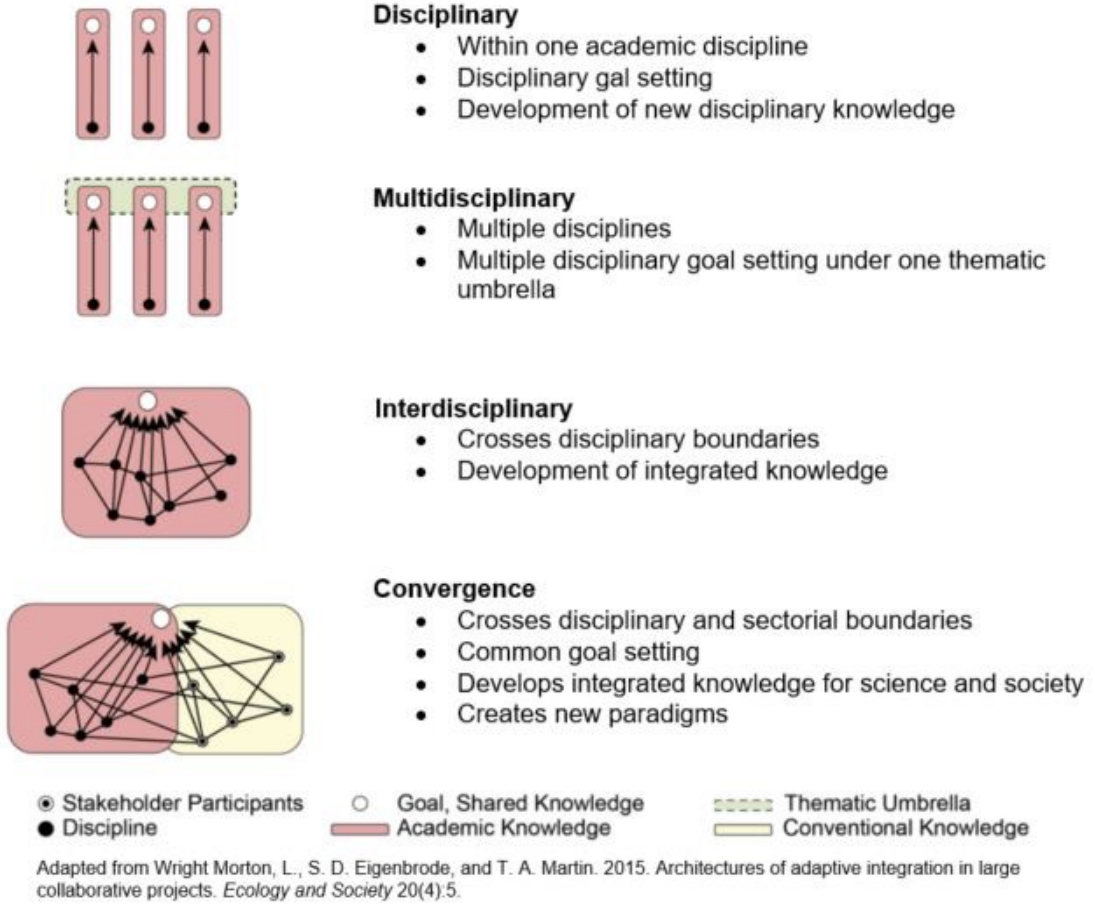
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Convergence research is:

driven by a specific and compelling societal problem

and

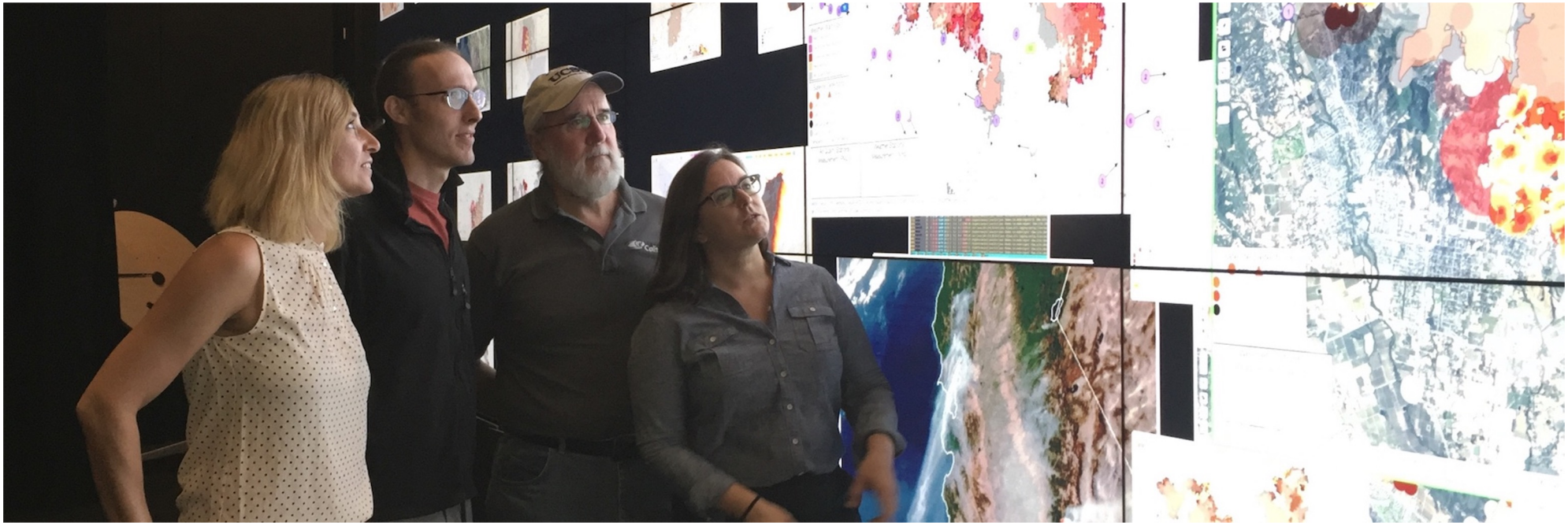
works towards integrating innovative and sustainable solutions into society



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



SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Three Main Components

Composable Workflows + Collaborative Innovation + Impact Network

- 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**

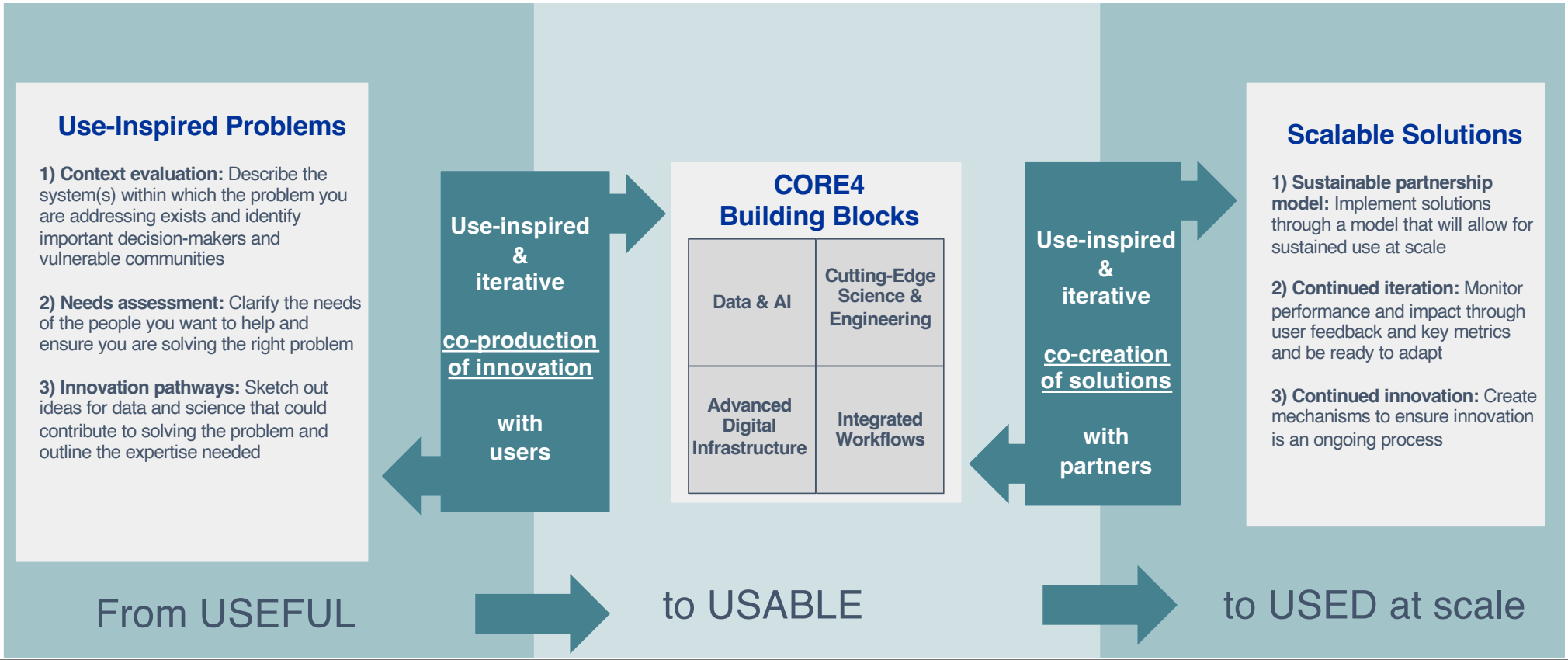
- Create **collaborative pathways** between UC San Diego and Los Alamos National Lab
- Accelerate the advancement of **science and technology as a basis for responsible and proactive approaches** to environmental challenges
- Leverage cutting-edge capabilities in scalable computing and diverse scientific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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

CORE Institute Innovation Approach

Creating Breakthrough Technological Innovations for Complex Societal Challenges



SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

Translating Fire Research into Impact

WIFIRE LAB

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

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

WIFIRE LAB

FIREMAP

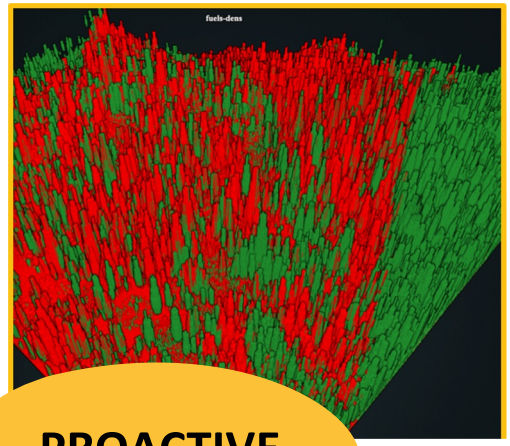
Firemap is currently being used by firefighters in Colorado, in collaboration with Intterra, and firefighters in California through the FIRIS program under the California Governor's Office of Emergency Services and CALFIRE. FIRIS uses Firemap to provide real-time information on weather conditions and 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.

REACTIVE

CAL FIRE · Cal OES GOVERNOR'S OFFICE OF EMERGENCY SERVICES

BurnPro3D

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 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.



PROACTIVE

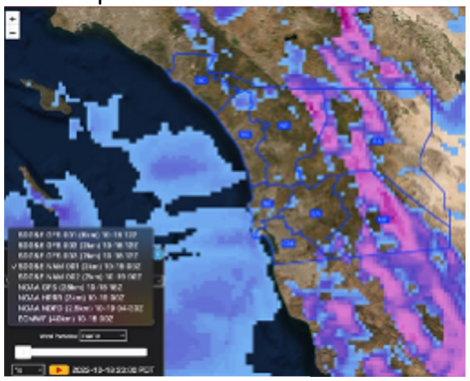
Los Alamos NATIONAL LABORATORY · Forest Service US Department of Agriculture · NSF

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Data and Computing Platforms

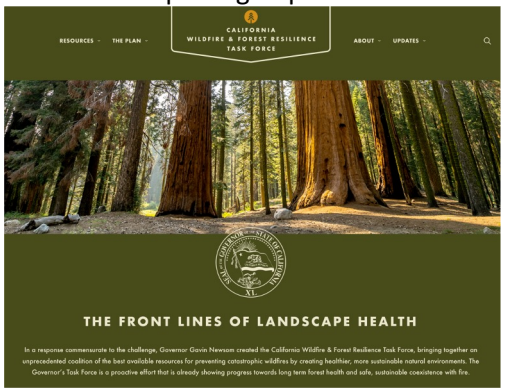
Wildfire Science and Technology Commons

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.



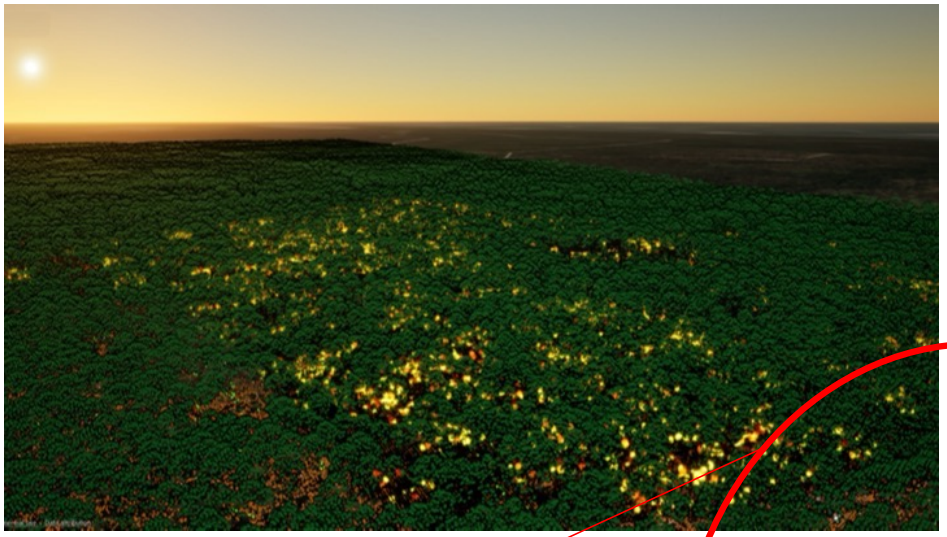
SDSC SAN DIEGO SUPERCOMPUTER CENTER

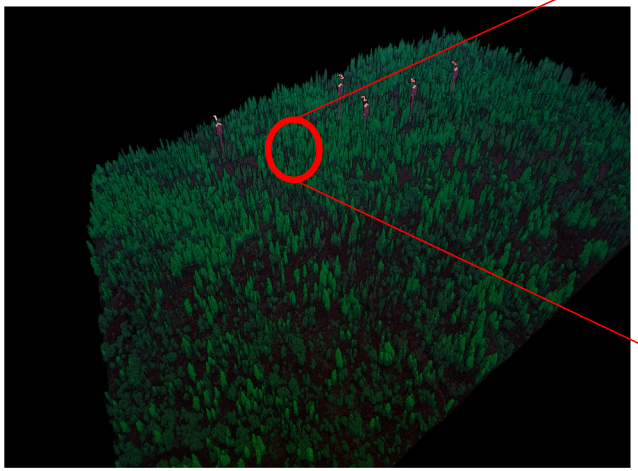
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

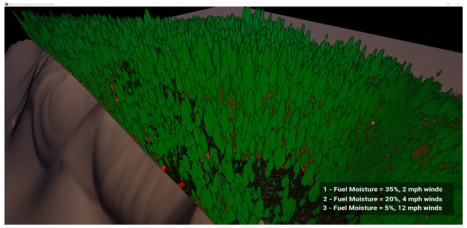

UC San Diego
Immersive Forest
wifire.ucsd.edu/

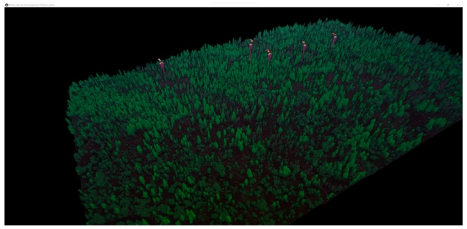
Immersive Forest
for
Multimodal Communication

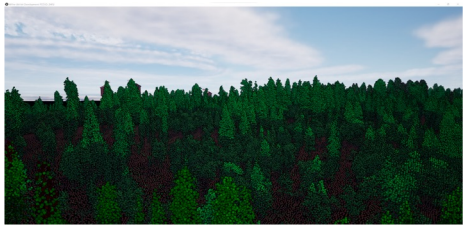




Terrestrial LiDAR contextualized within Aerial scan

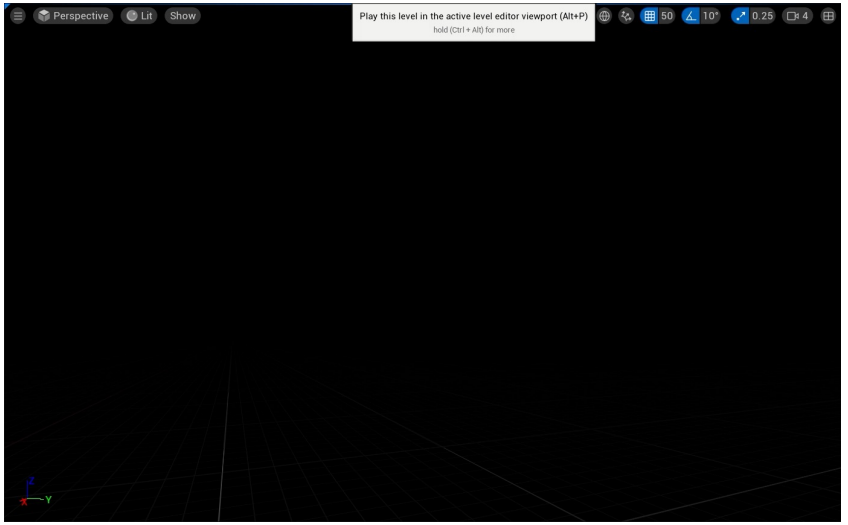






AI in Science Communication

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.

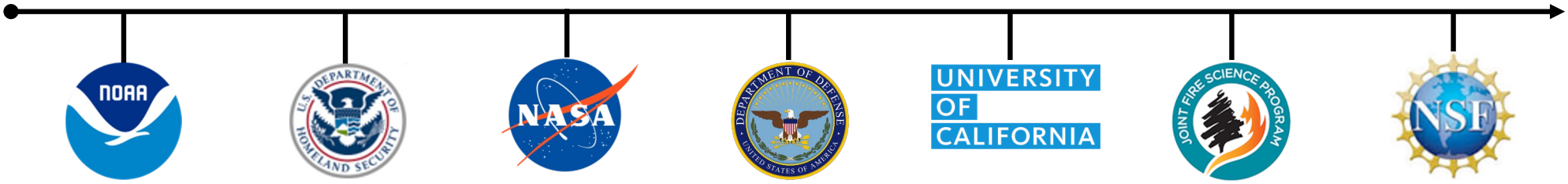
Animations by: Isaac Nealey (left, bottom), Ivannia Gomez (top)

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

This type of work needs the CORE4 building blocks.

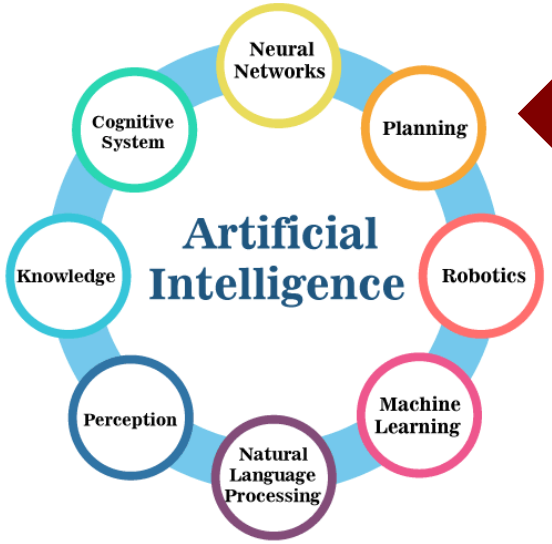
CICORE
Cyberinfrastructure | Convergence Research | Education

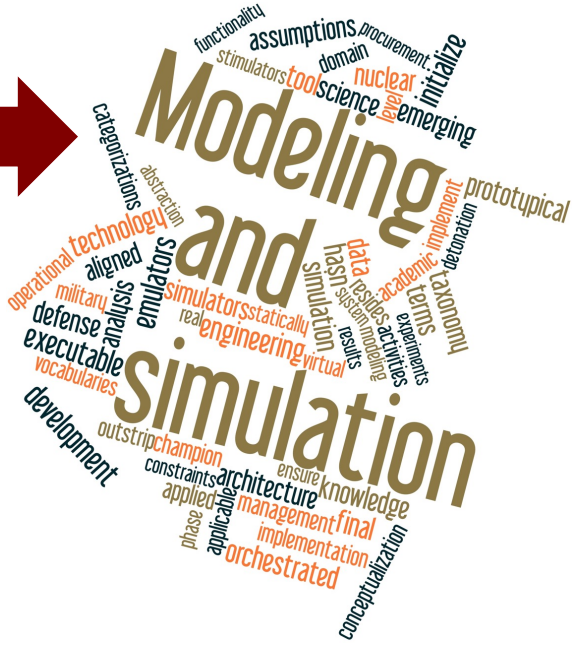
CORE4 Building Blocks

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

AI-Integrated Applications at the Digital Continuum





İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

AI in Science and Research 2023

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

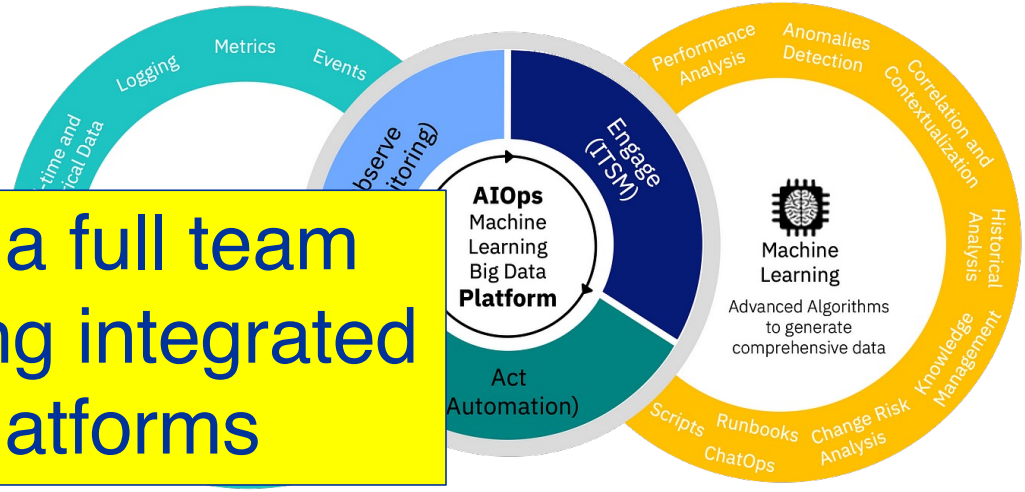
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

AI in Science Readiness

"not just science + AI methods"

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

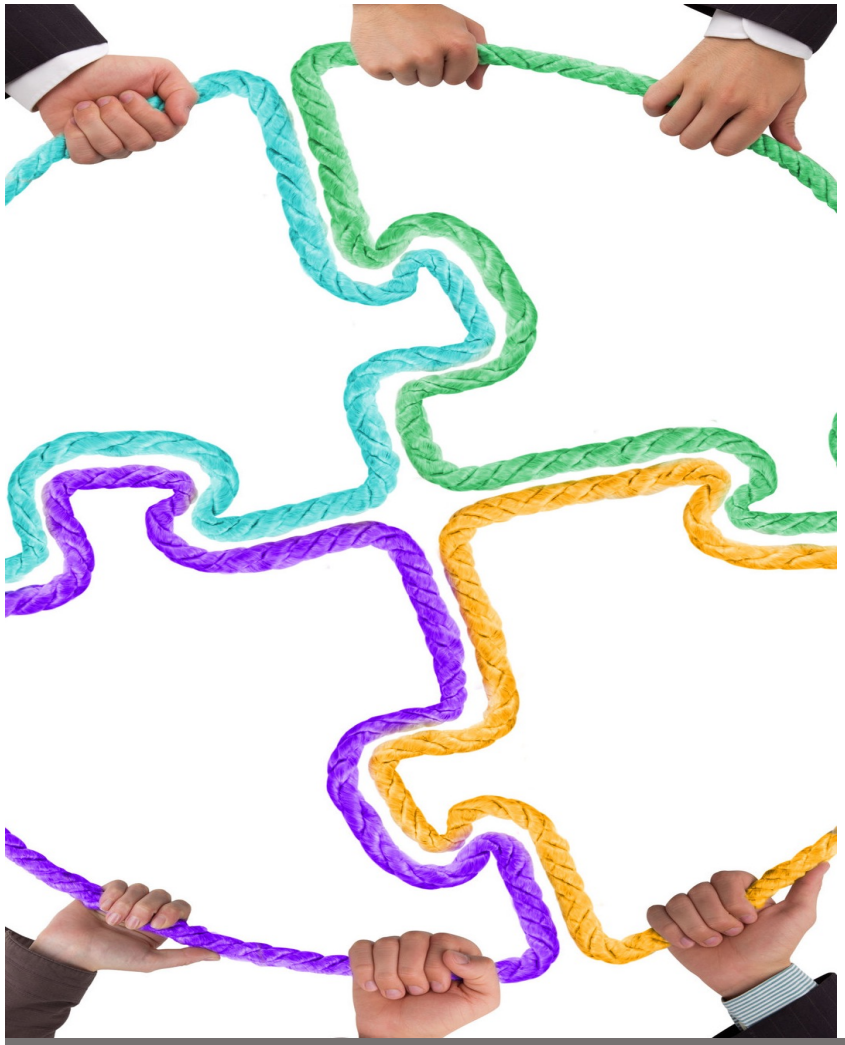
Requires a full team and enabling integrated data platforms



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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

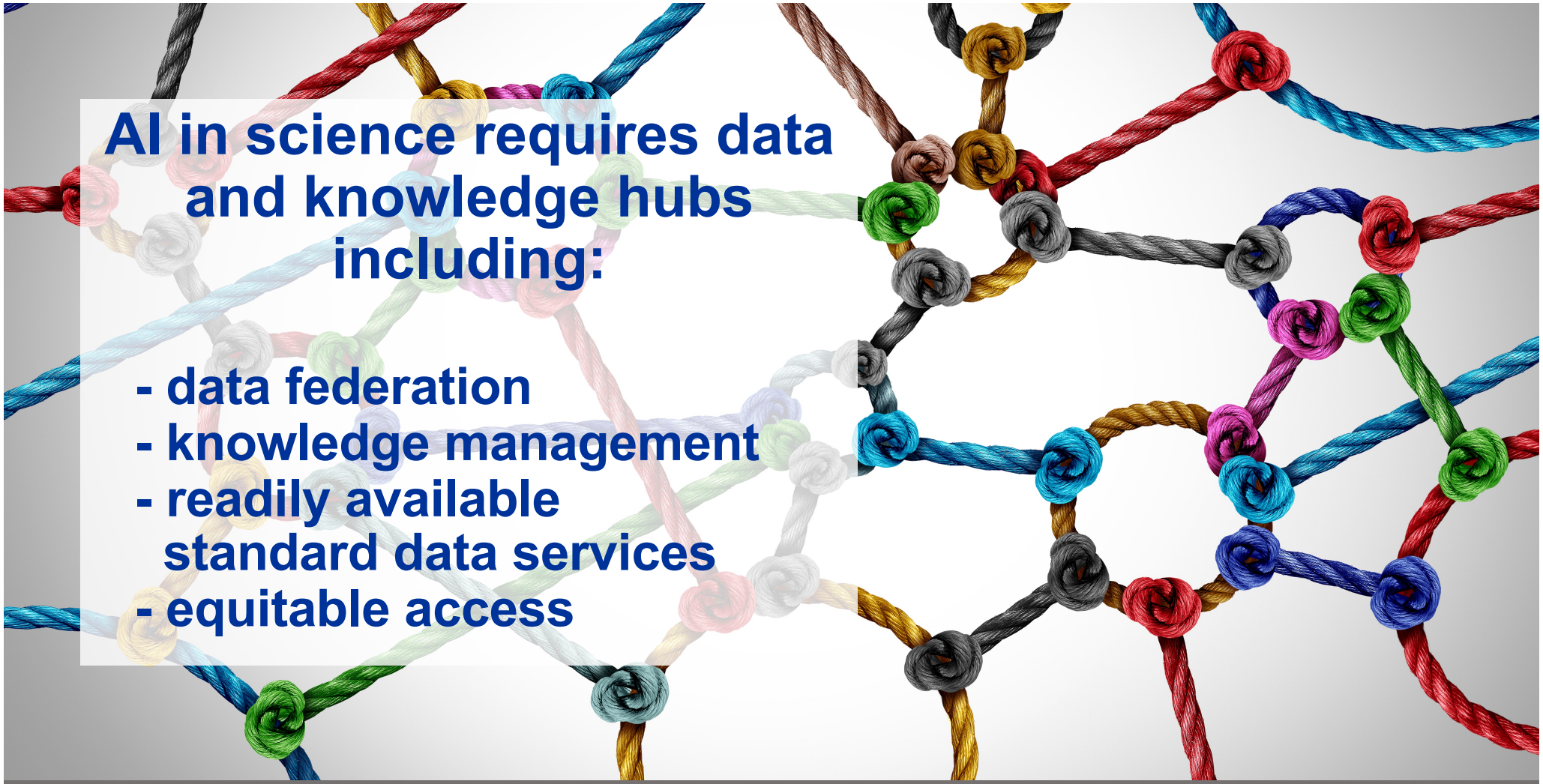
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



Workflow integration requires a digital continuum composed through:

- system federation
- reusable capability services
- solutions integrating services

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Integration requirements…

Dynamic composability matters.

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

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Dynamic composability matters.

COMPOSABLE SERVICES

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

RESOURCE MANAGEMENT

e.g., container orchestration, optimization

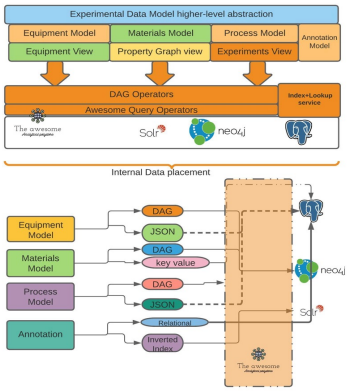
COMPOSABLE SYSTEMS

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

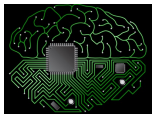
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Big Data and IoT ⟷ **Artificial Intelligence** ⟷ **Modeling and Simulation**

Capability ⟷ Capacity

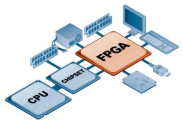


Big Data



xPU → GPU, CPU, TPU, IPU, QPU,…

Edge



FPGA

Cloud, HPC, Storage

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Some Composable Systems

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

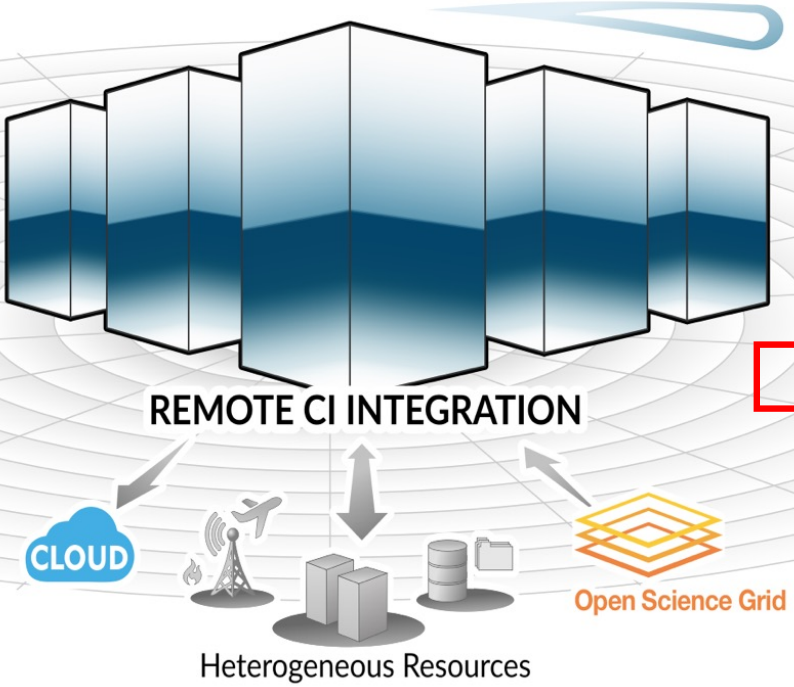
SDSC
SAN DIEGO SUPERCOMPUTER CENTER

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

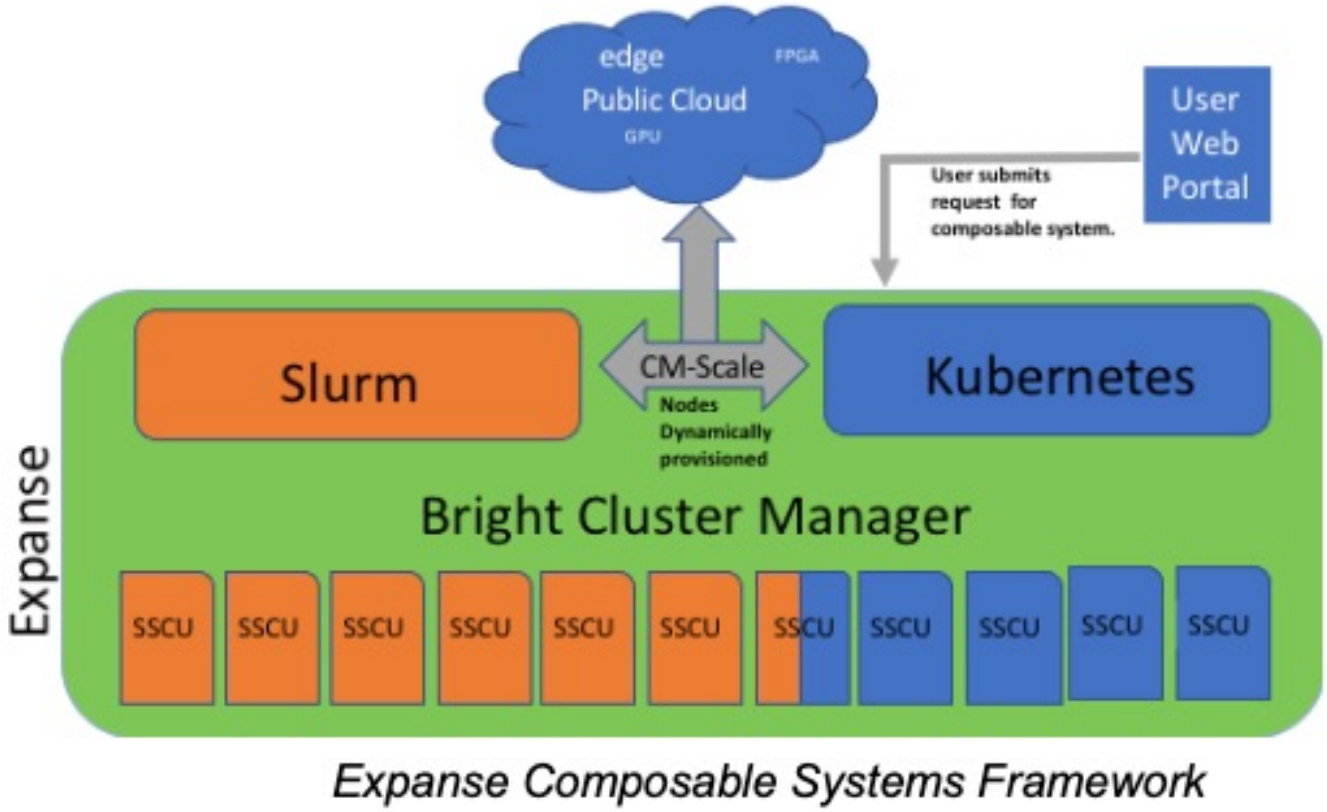
DATA CENTRIC ARCHITECTURE
12PB Perf. Storage: 140GB/s, 200k IOPS
Fast I/O Node-Local NVMe Storage
7PB Ceph Object Storage
High-Performance R&E Networking



LONG-TAIL SCIENCE
Multi-Messenger Astronomy
Genomics
Earth Science
Social Science

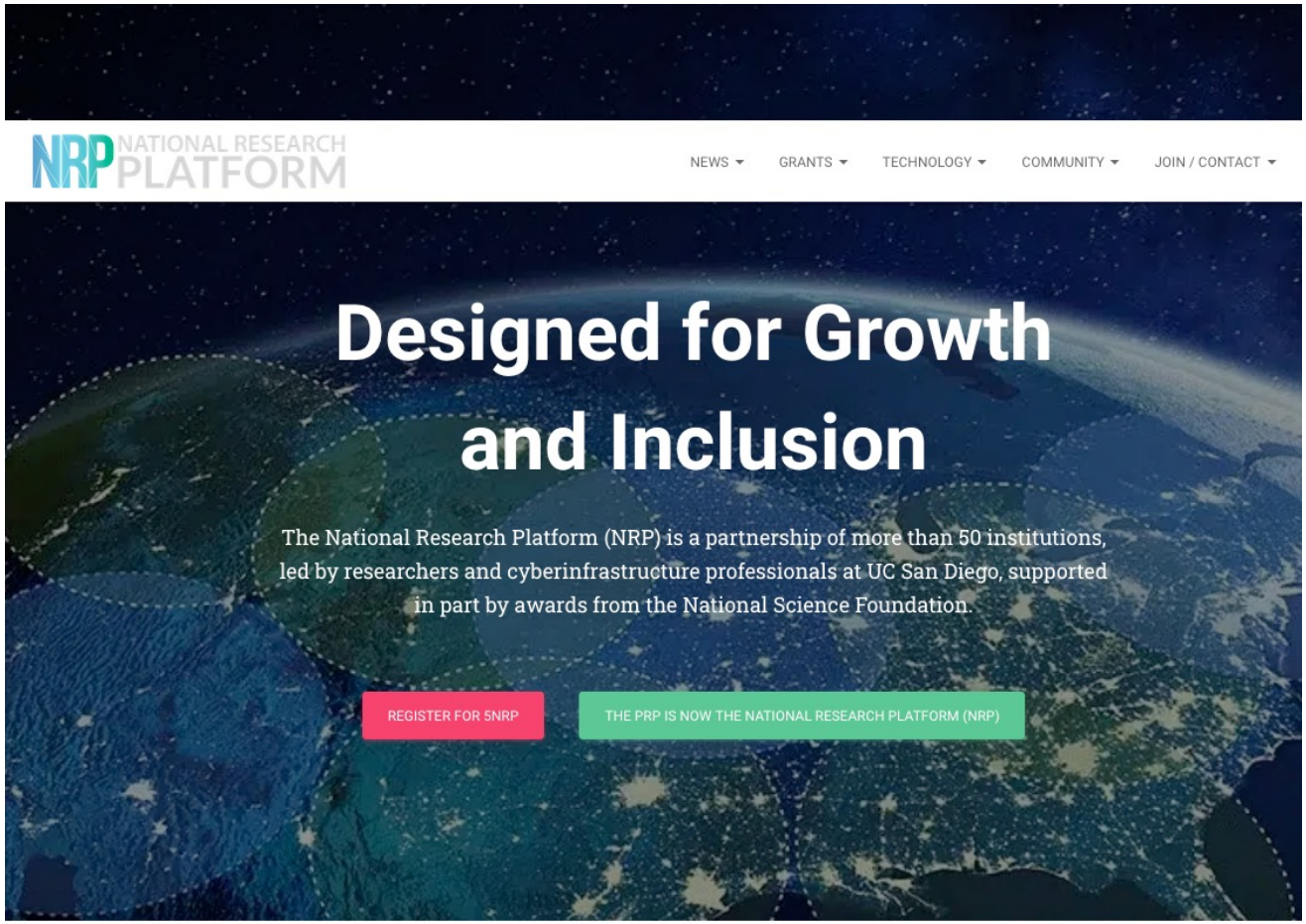
INNOVATIVE OPERATIONS
Composable Systems
High-Throughput Computing
Science Gateways
Interactive Computing
Containerized Computing
Cloud Bursting

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



Expanse Composable Systems Framework

National Research Platform



Nautilus

https://nationalresearchplatform.org/

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

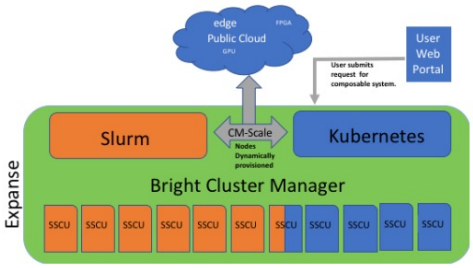
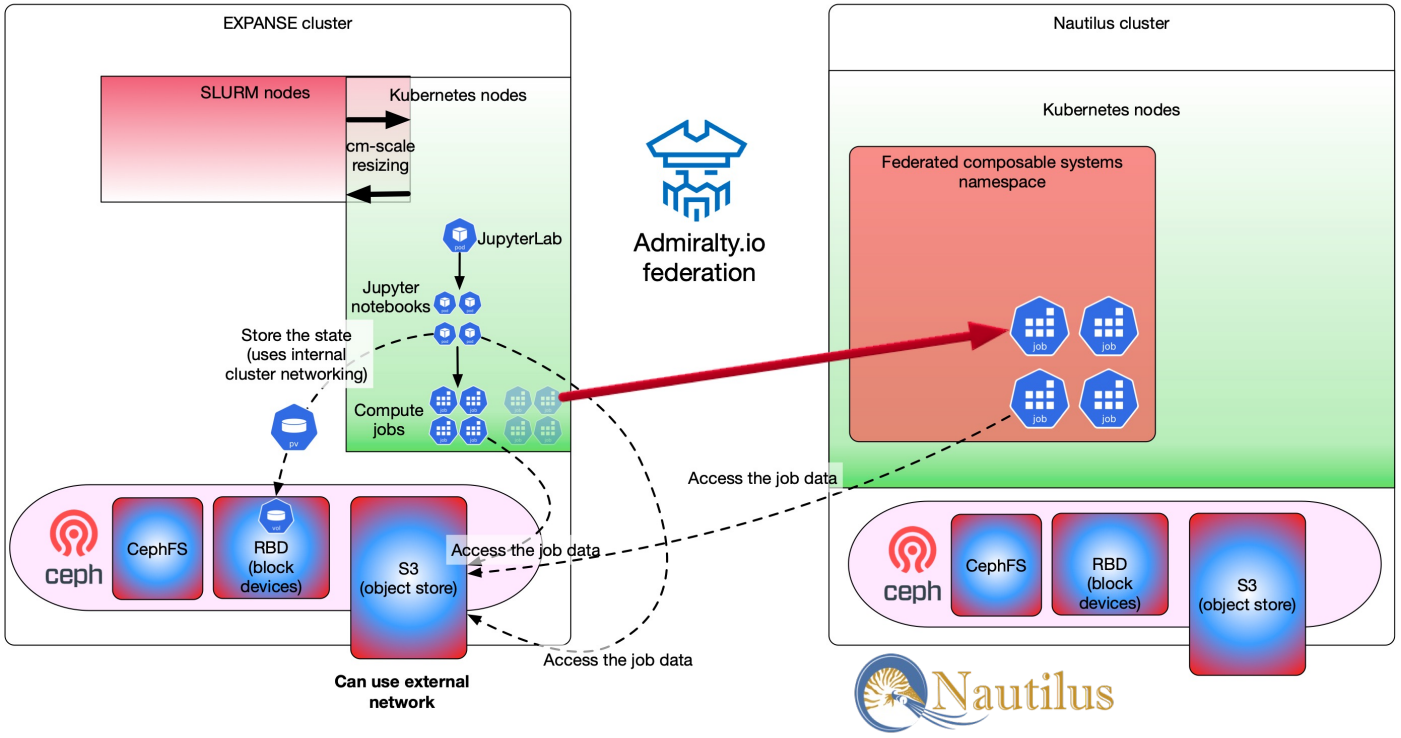


Figure 5.1 Expanse Composable Systems Framework

First composable cluster is federated!

EXPANSE (Enthalpy) + CHASE-CI (Nautilus)

EXPANSE (Enthalpy)



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Sensor/Instrument — HPC/Cloud

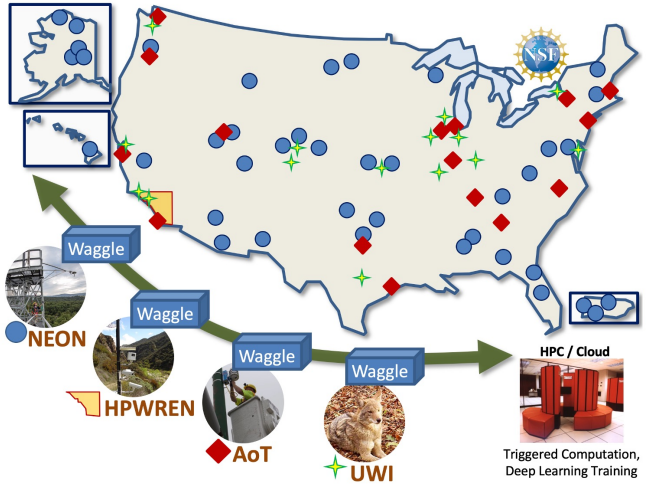
AI@Edge and the Digital Continuum

Slide Source: *Pete Beckman, ANL*

SAGE

Cyberinfrastructure for
AI at the Edge
sagecontinuum.org

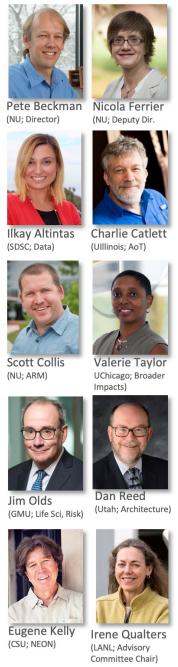
NSF



Northwestern University · The University of Chicago · Argonne National Laboratory · Colorado State University · UC San Diego · George Mason University · Array of Things · ARM · neon National Ecological Observatory Network · HPWREN · SDSC San Diego Supercomputer Center · Lincoln Park Zoo For Wildlife. For All.

Education & Training

Leadership Team



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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

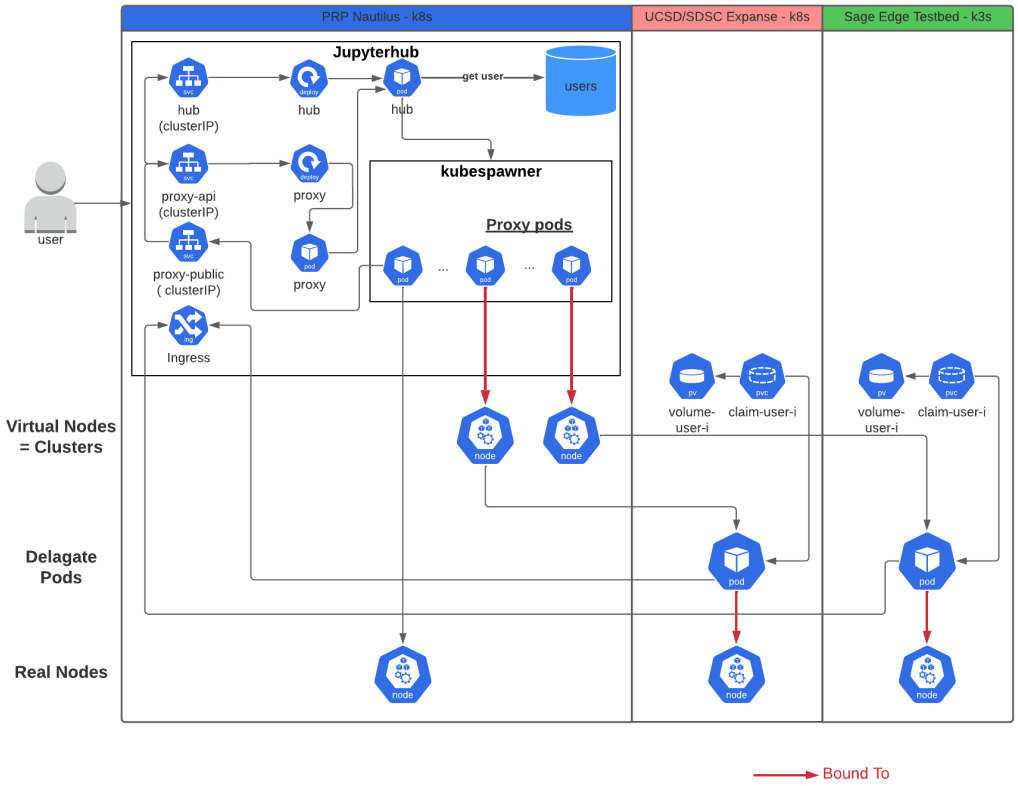
COMPOSABLE SERVICES

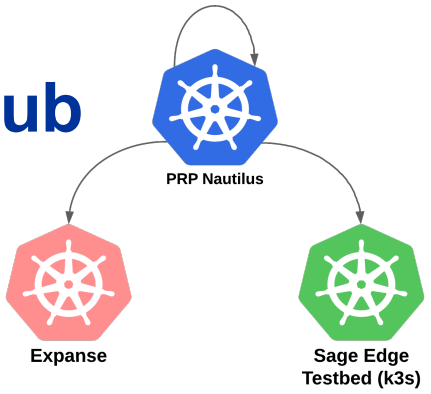
RESOURCE MANAGEMENT

COMPOSABLE SYSTEMS

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Integration of NSF EXPANSE, NRP and Sage
A Composable System Deployment of JupyterHub





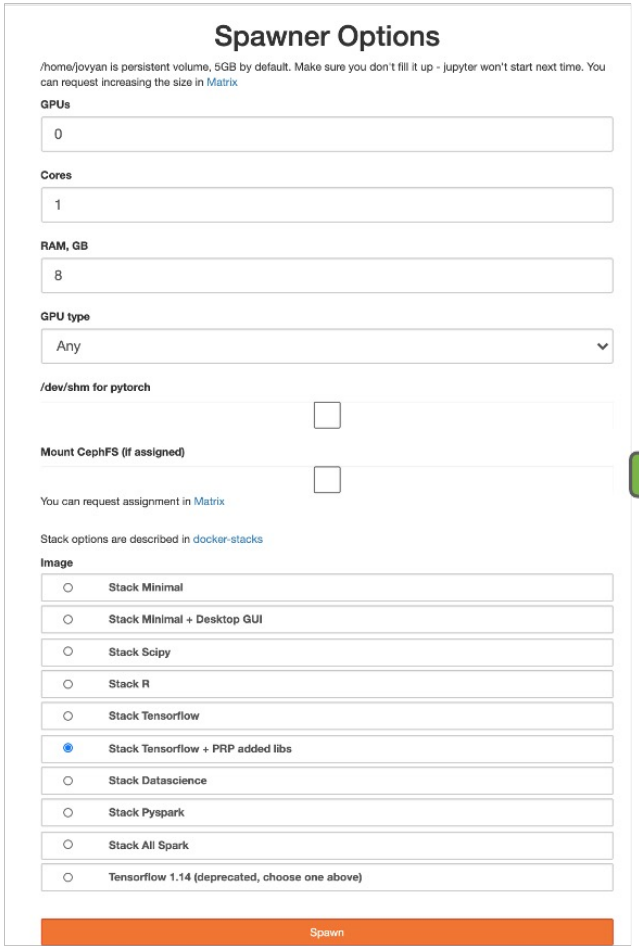
- Edge-Cloud Unified Environment for prototyping AI models to deploy on the Edge
- A user can easily be provided the right environment for developing their AI Edge Application

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

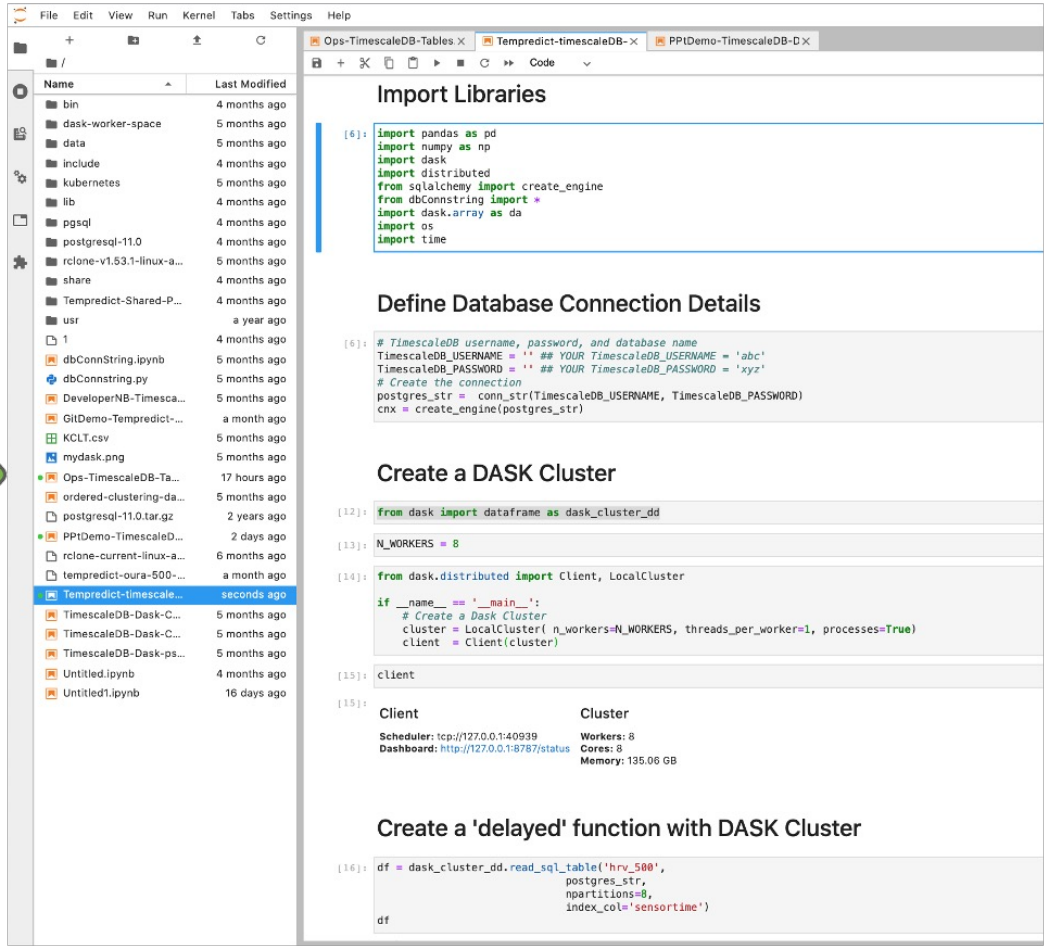
SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE



Kubernetes Pod Spawned for Exploration



SDSC SAN DIEGO SUPERCOMPUTER CENTER

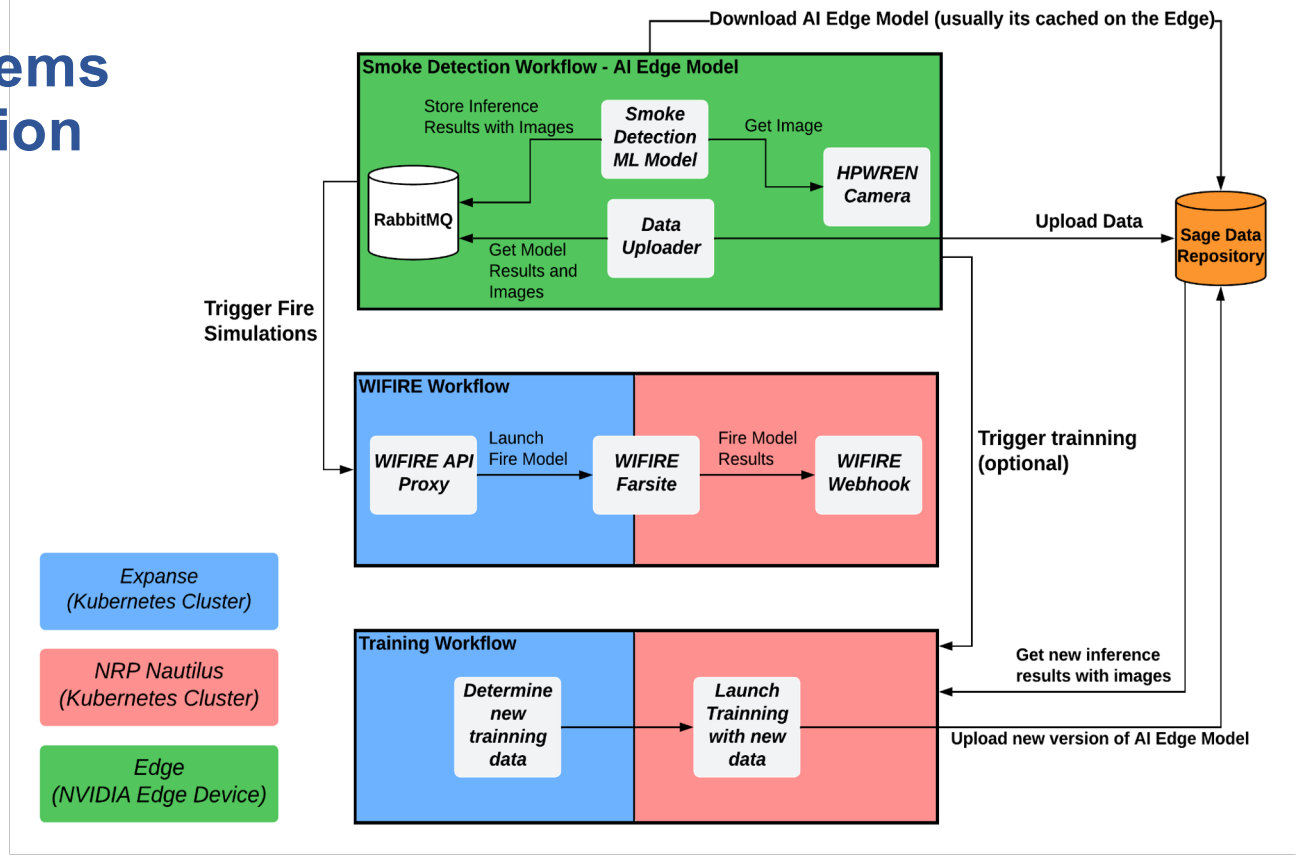
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

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

We need to stay use-inspired!

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

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

REPRODUCIBILITY

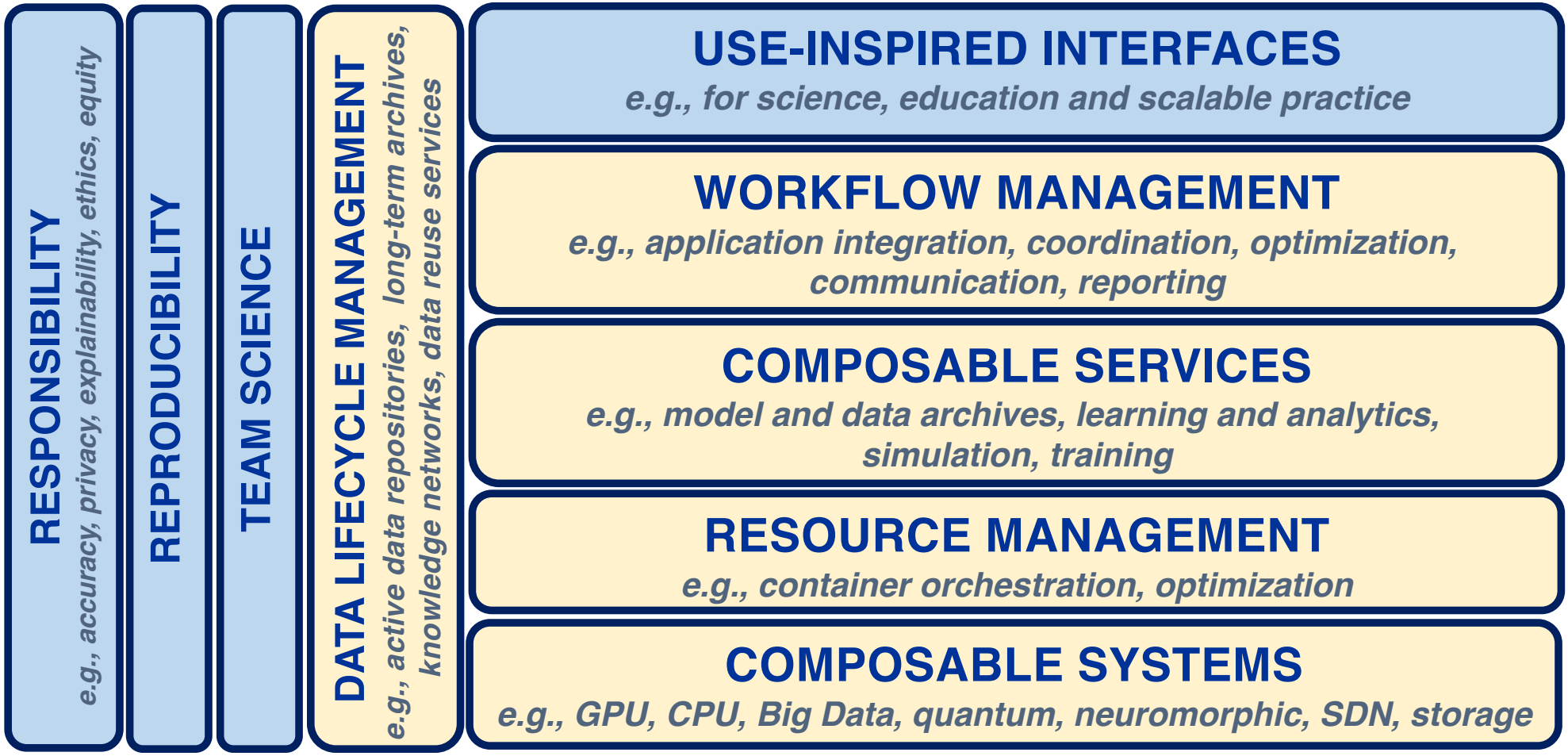
TEAM SCIENCE

USE-INSPIRED INTERFACES

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

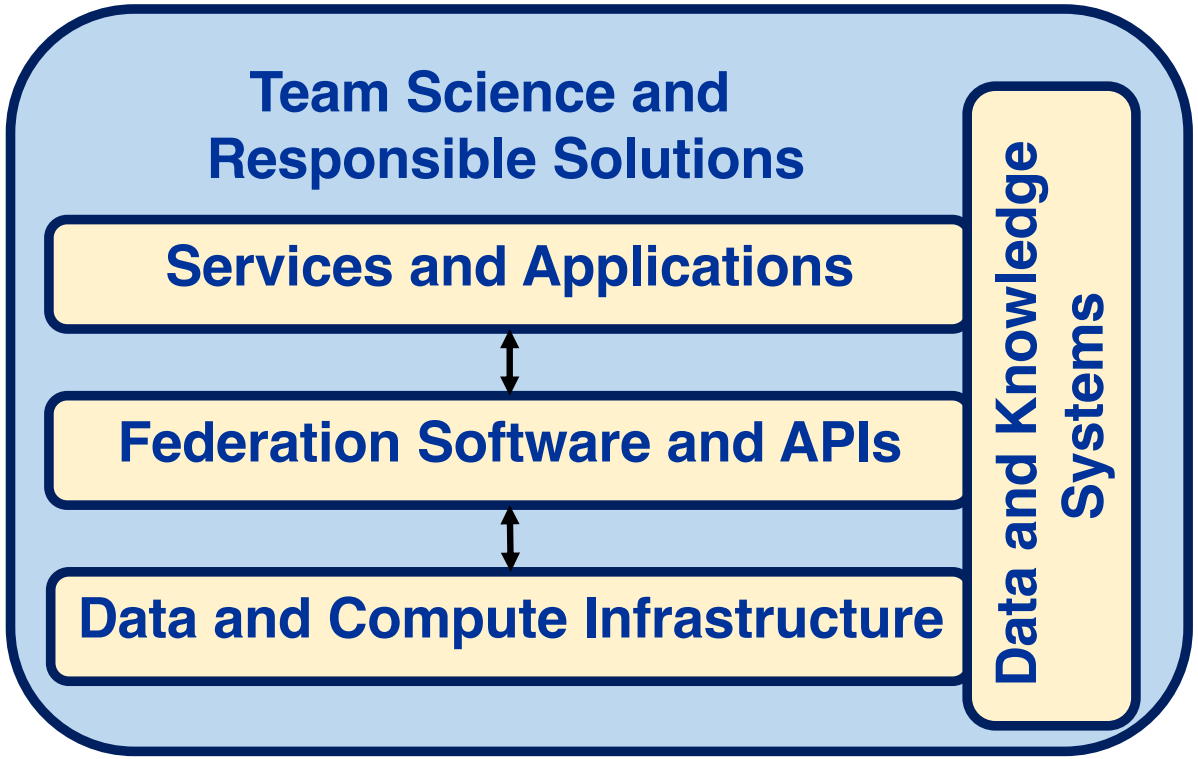
Tools that enhance teamwork and use need to be coupled with responsible AI systems.

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Schmidt AI in Science Postdoc Research

Schmidt Sciences

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

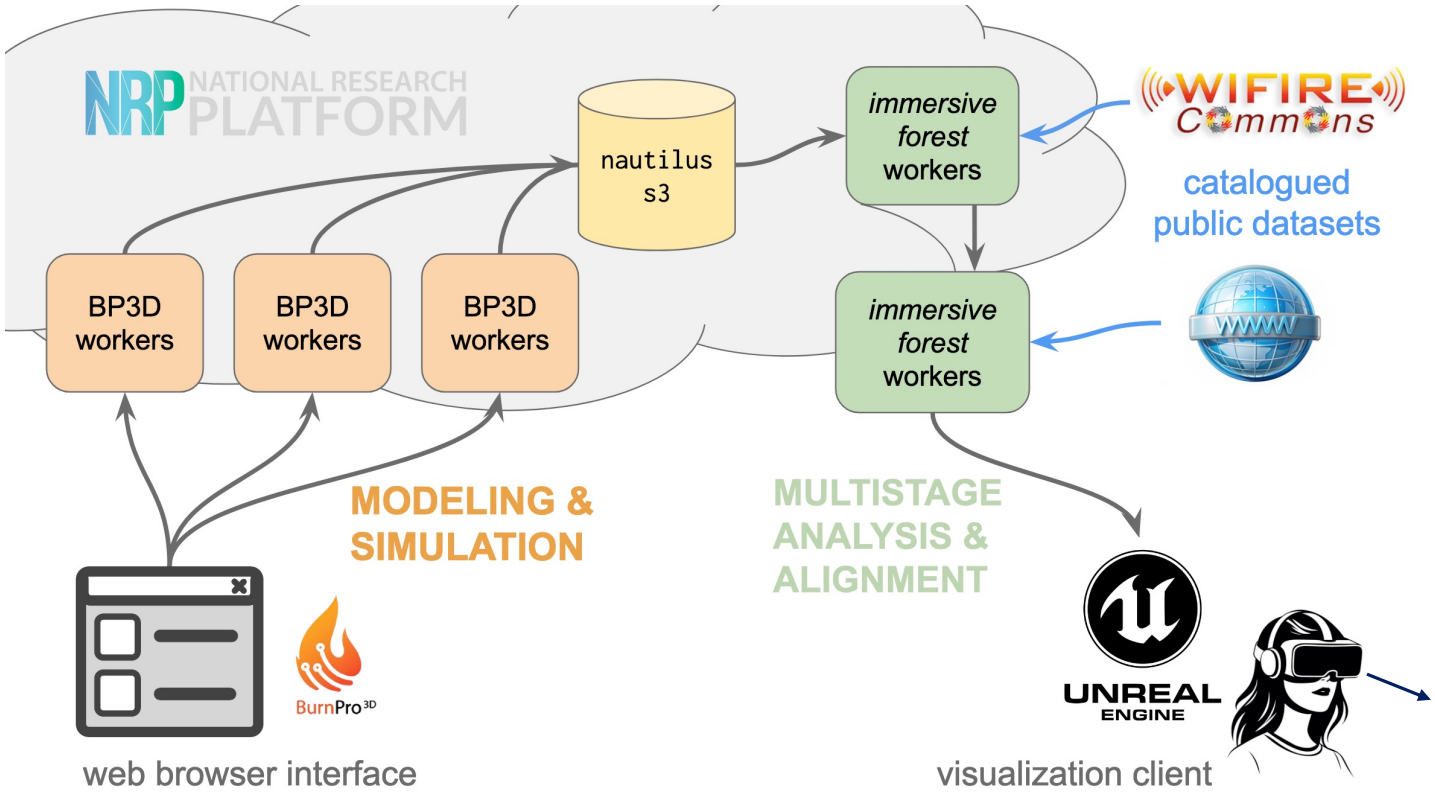
Bathymetry from space
Machine learning model to understand small-scale ocean dynamics

The effect mutations implicated in autism can have in protein oscillation
Deep learning model to predict the oscillation of protein in cell-cell communication

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE



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

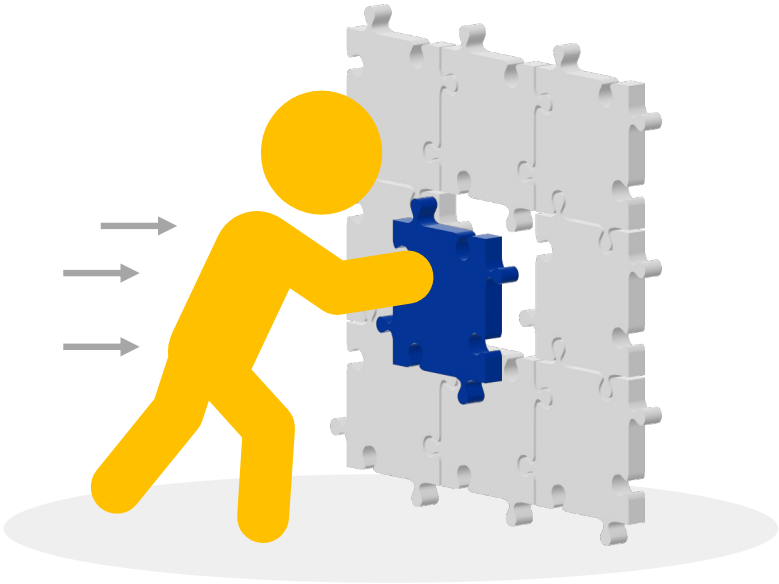
WIFIRE LAB

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

What do we do about the data gaps?



http://www.nationaldataplatform.org

UC San Diego

UTAH

University of Colorado Boulder

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

SCI
www.sci.utah.edu

EarthScope Consortium

http://www.nationaldataplatform.org

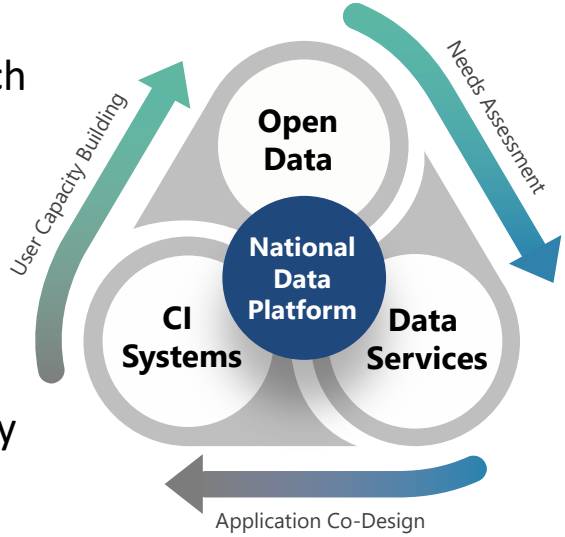
Award abstract: https://www.nsf.gov/awardsearch/showAward?AWD_ID=2333609

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

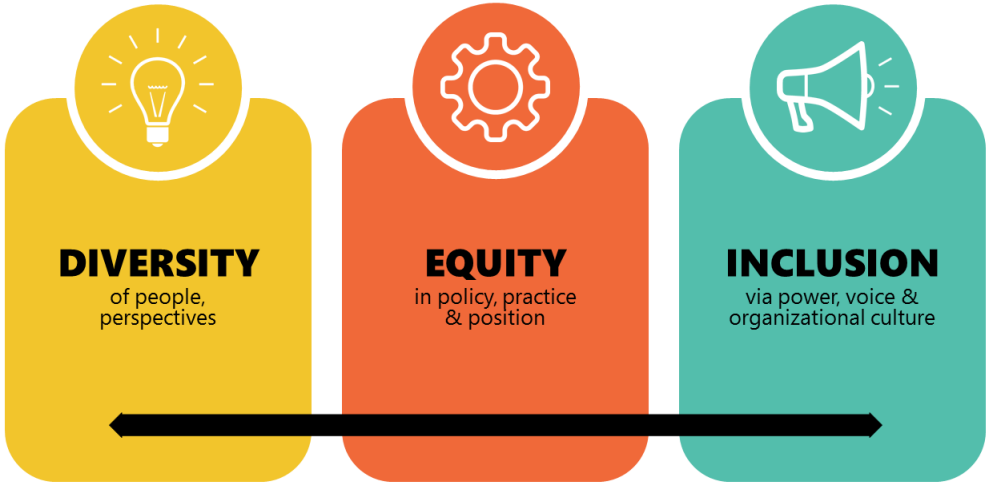
Addressing Open Questions for Equitable Open Access

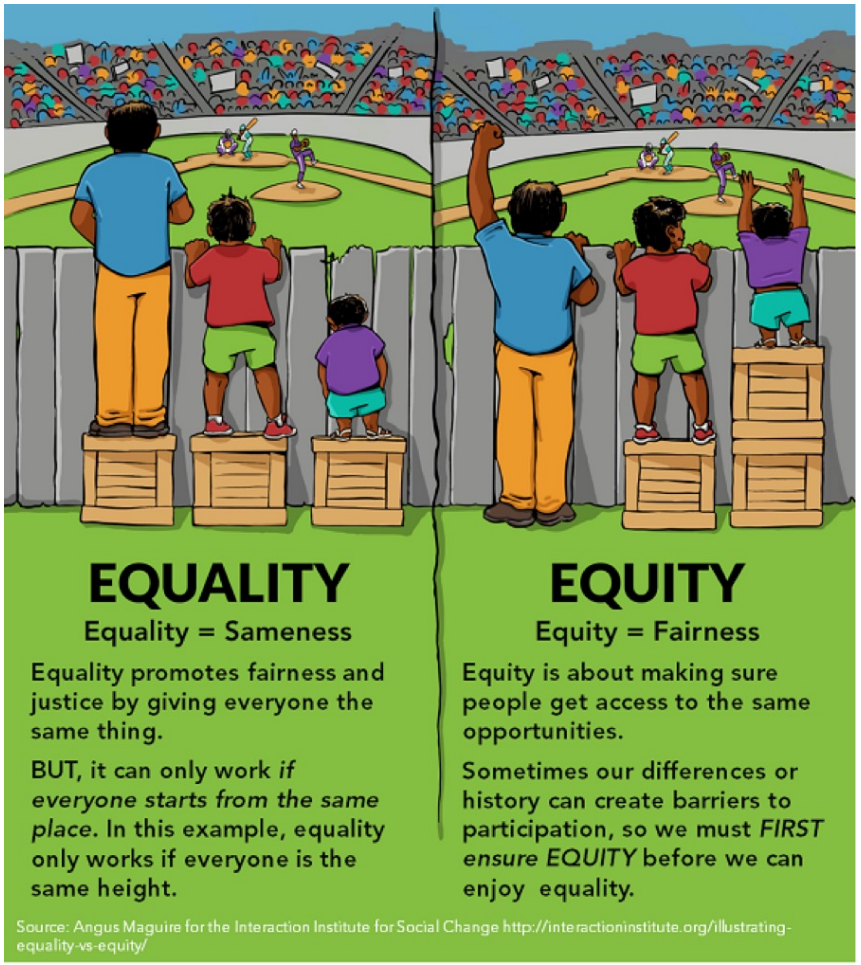
Foundational Abstractions and Services	• What are the foundational data abstractions and services that can serve as multipurpose and expandable building blocks for data-driven and AI-integrated application patterns? • How can everyone effectively access and utilize these abstractions and services?
Equitable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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

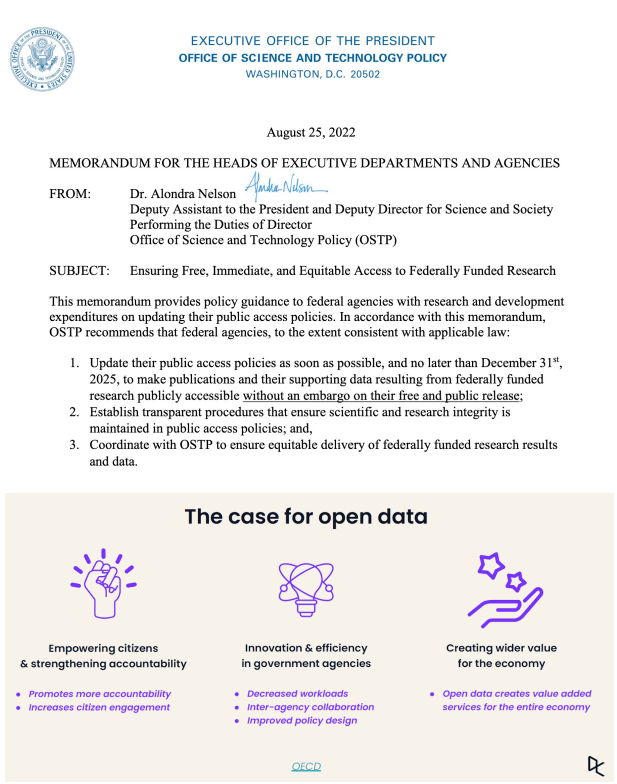
Diversity is a fact.
Equity is a choice.
Inclusion is an action.
Belonging is an outcome.

- Arthur Chan





İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



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

August 25, 2022

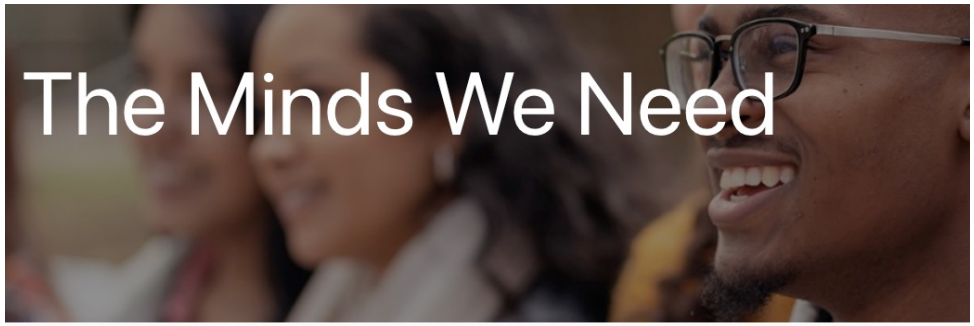
MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

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.

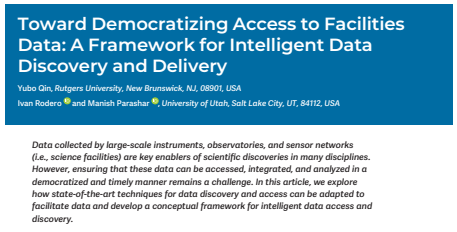


Inclusion, Innovation, and Competitiveness | Strengthening Our National Broadband Initiative | Investing in Research and Education Infrastructure | Contributors | Toolkit | Endorsements

Inclusion, Innovation, and Competitiveness

We are at a crossroads.

https://mindsweneed.org

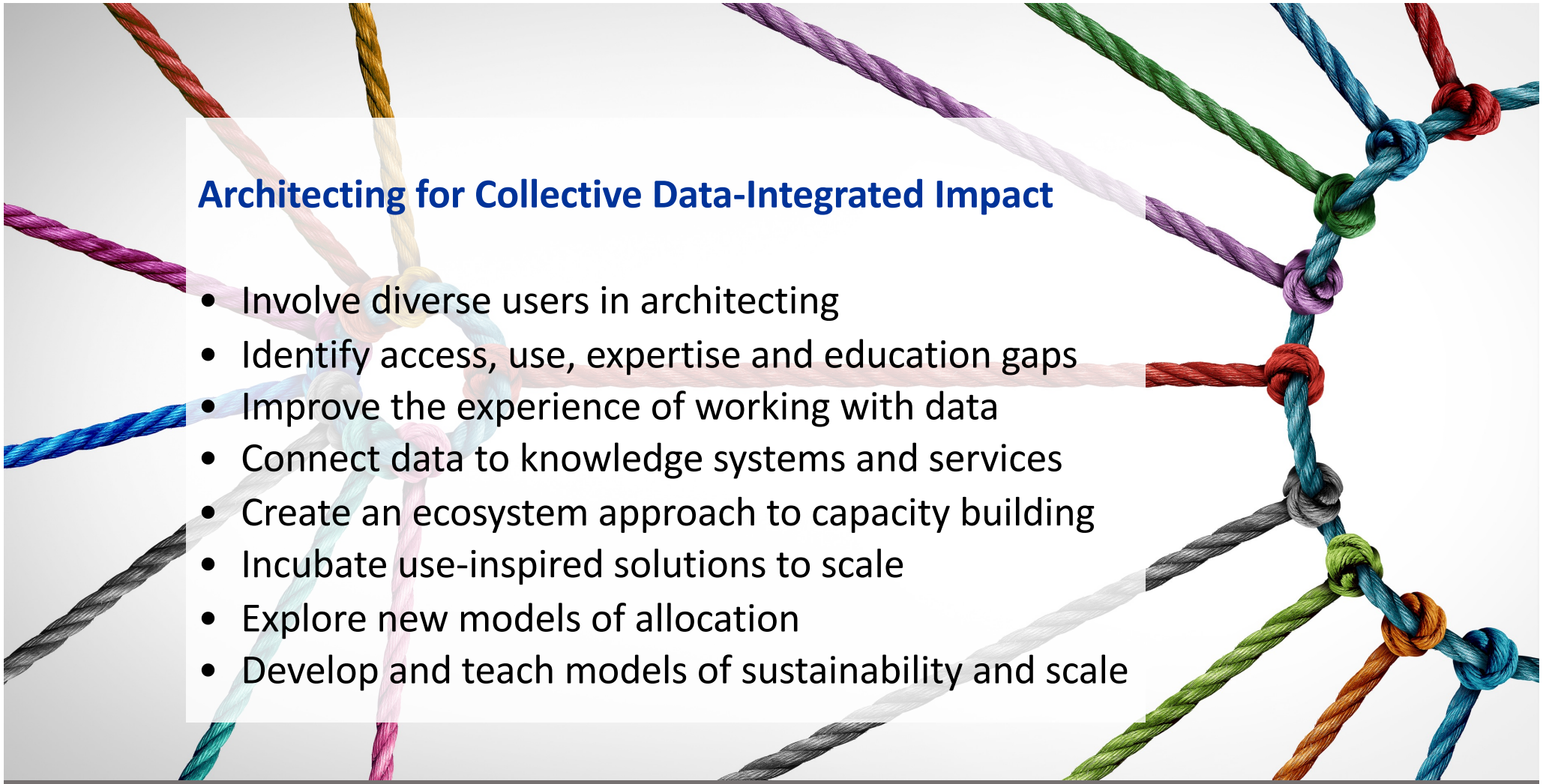


Democratization of CI and Data Access

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE



Architecting for Collective Data-Integrated Impact

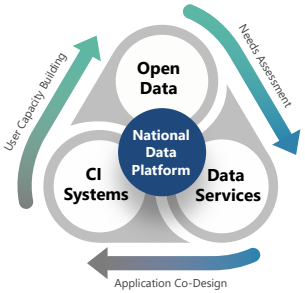
- Involve diverse users in architecting
- Identify access, use, expertise and education gaps
- Improve the experience of working with data
- Connect data to knowledge systems and services
- Create an ecosystem approach to capacity building
- Incubate use-inspired solutions to scale
- Explore new models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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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

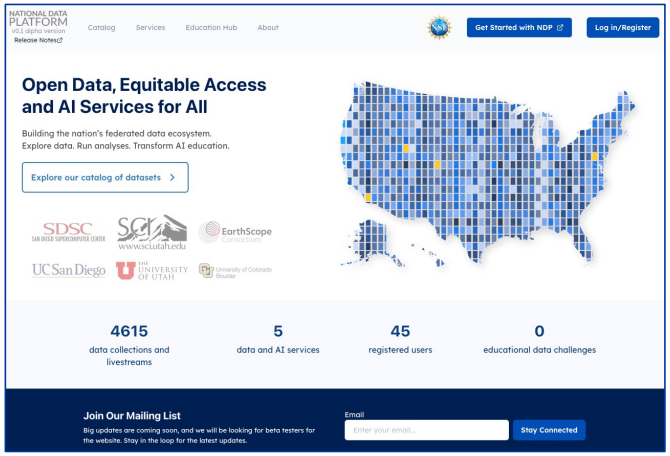
- Distributed in nature
- Composition as a principle
- Hub-centric services as connection backbone
- Integrates in education systems

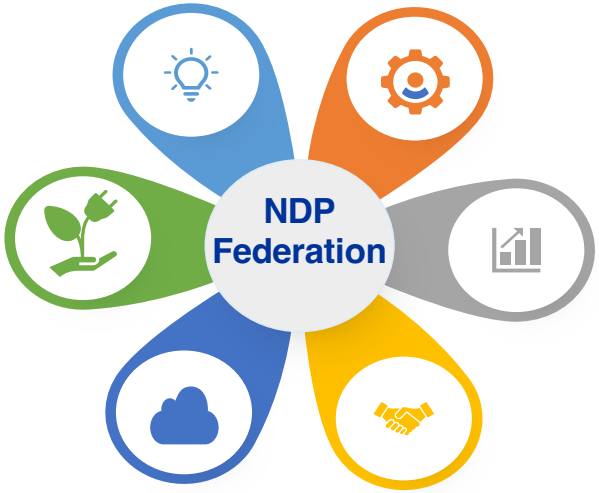
Collaboration, Incubation, Allocation and Partnership Models

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE





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

NATIONAL DATA PLATFORM

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Current 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

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

Cloudbank

(…)

Nautilus

jupyter

Pelican/OSDF

SAGE

HPC

Cloud

Containers

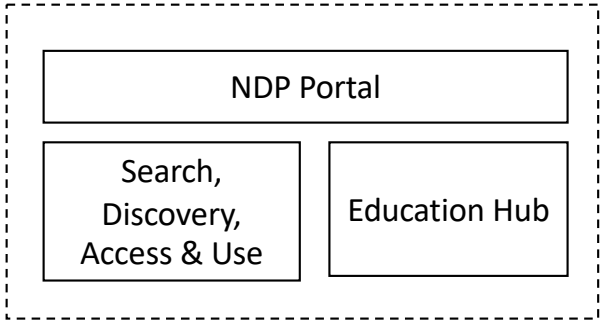
Data CI and Delivery Services

Composable CI

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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

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

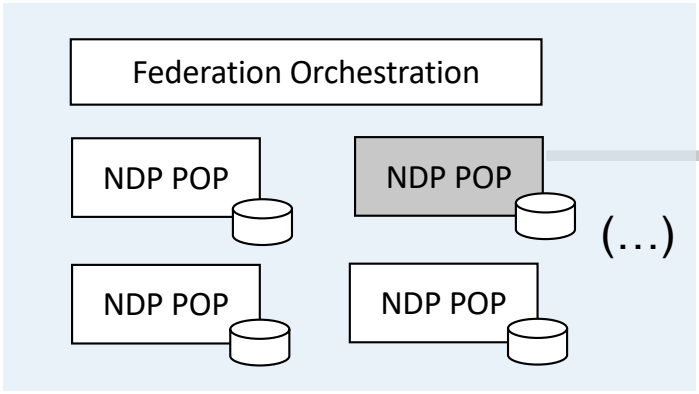
Planned Future Work

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

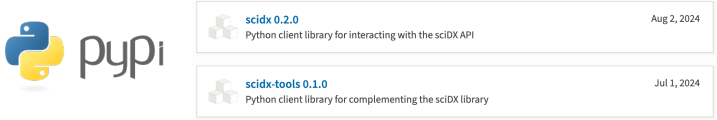
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

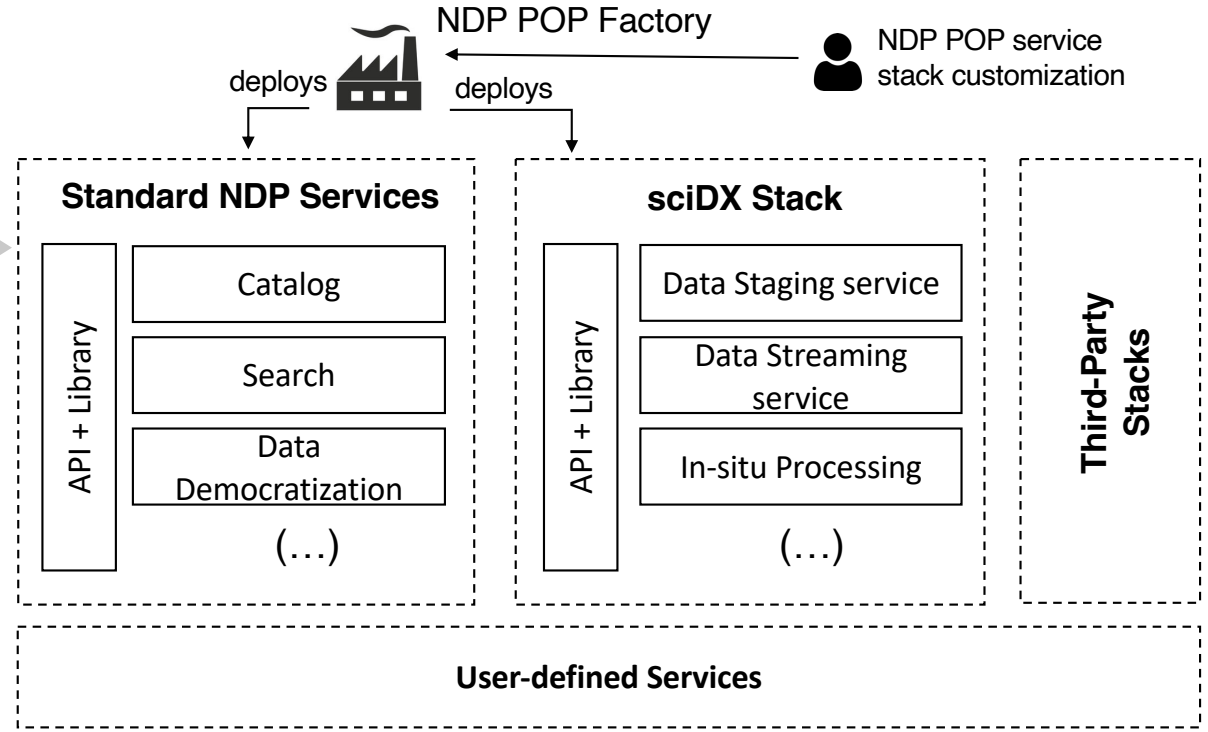
NDP POP: Distributed Points of Presence with Customizable, Composable Service Stacks

NDP Federation



Workflow composition currently via API and/or Python client library



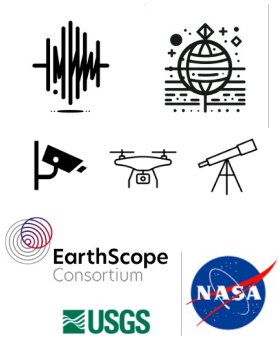


Deployment models: docker kubernetes Cloud HPC

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Typical NDP Workflow with Composable Capabilities



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

Data Acquisition

- Data adaptors: Customizable metadata/data ingestion; heterogeneous data sources & types

(Meta) data Curation, Registration, Indexing, Discovery

- Metadata catalog
- Extensible search engine
- Recommendation services

Data Access, Data Staging, In-situ Data Analytics

- High-performance data staging & in-situ processing
- Data access optimization (caching, pre-fetching, recommendation)
- Leverage Pelican data origins; data federation

Data Processing, Data-driven, AI/ML-based Workflows,

- Services Gateway (support notebooks)
- Leverage NAIRR/NSF ACCESS/Cloud resources.
- Data-driven stream processing

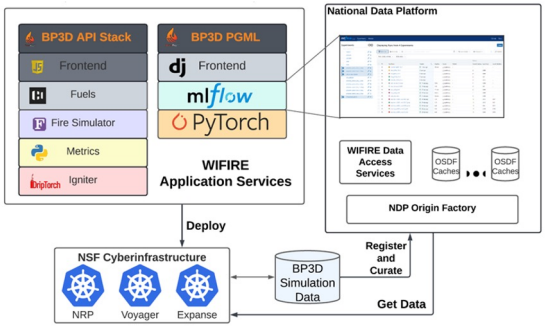
Product Generation, Curation, Sharing, Archival

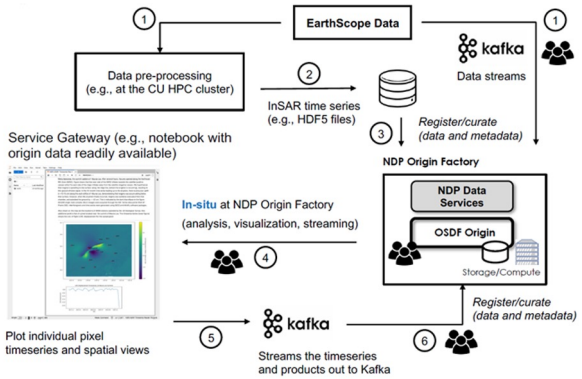
- Data product & re-streaming
- Archival support (including using Pelican)

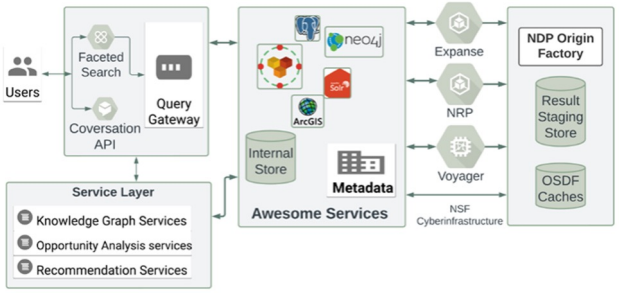
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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.

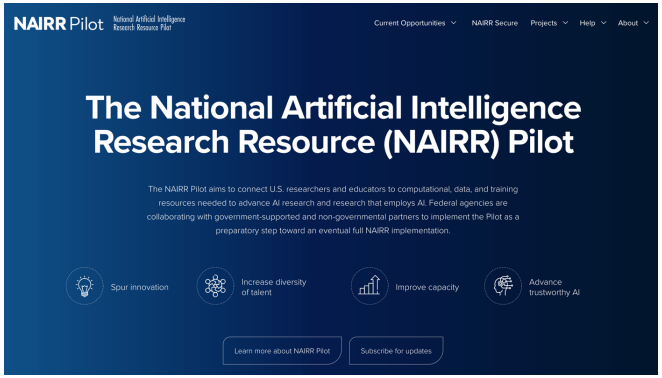






İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Example NDP-NAIRR AI in Science Workflow

Data Acquisition

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

Registration, Indexing, Discovery

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

Workspace Definition

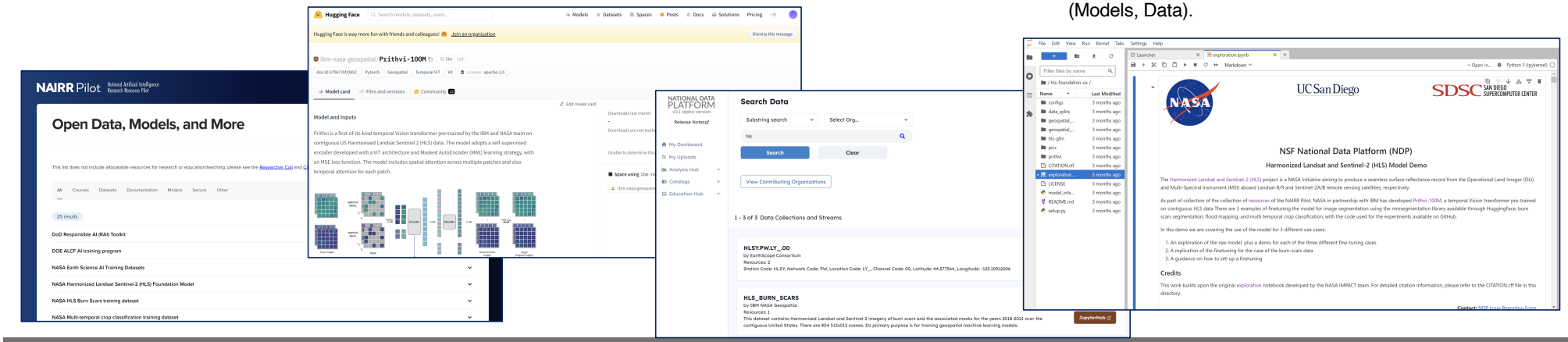
- Data and Models are included into user's workspace, along with the necessary libraries, services and files to work on a new project.

Data-driven, AI/ML-based workflows

- 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).

Product Generation, Curation, Sharing, Archival

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



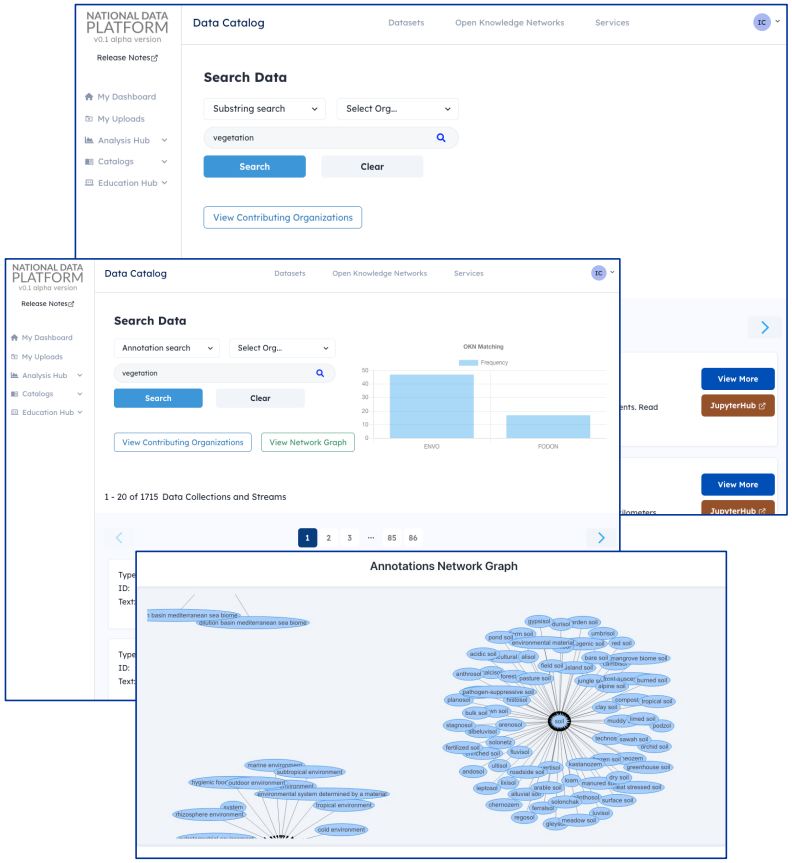
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

NDP Hub Functionality

September 2024 Release

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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:

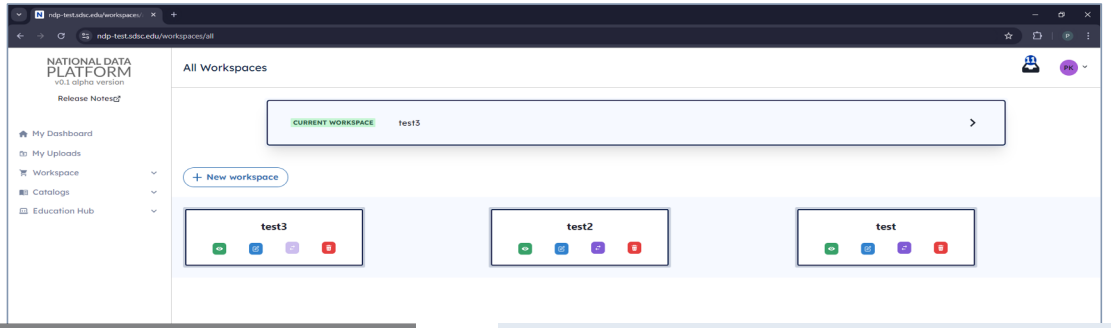
- Extract entity annotations from the metadata text and integrate them with the ontology to enhance search functionality.
- Create a vector store and develop a search pipeline that handles queries in natural language.
- Optimize the system's performance to ensure fast and accurate retrieval of relevant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

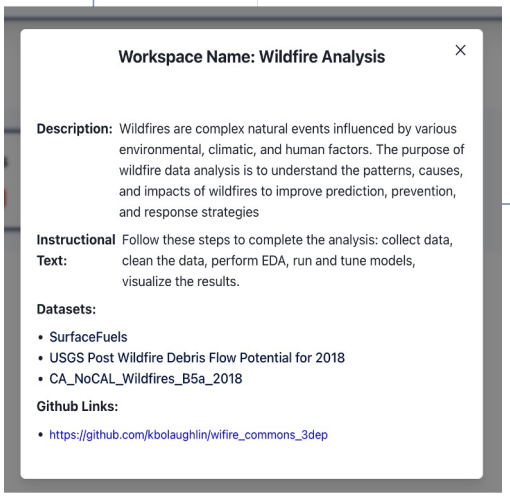
SDSC SAN DIEGO SUPERCOMPUTER CENTER
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

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



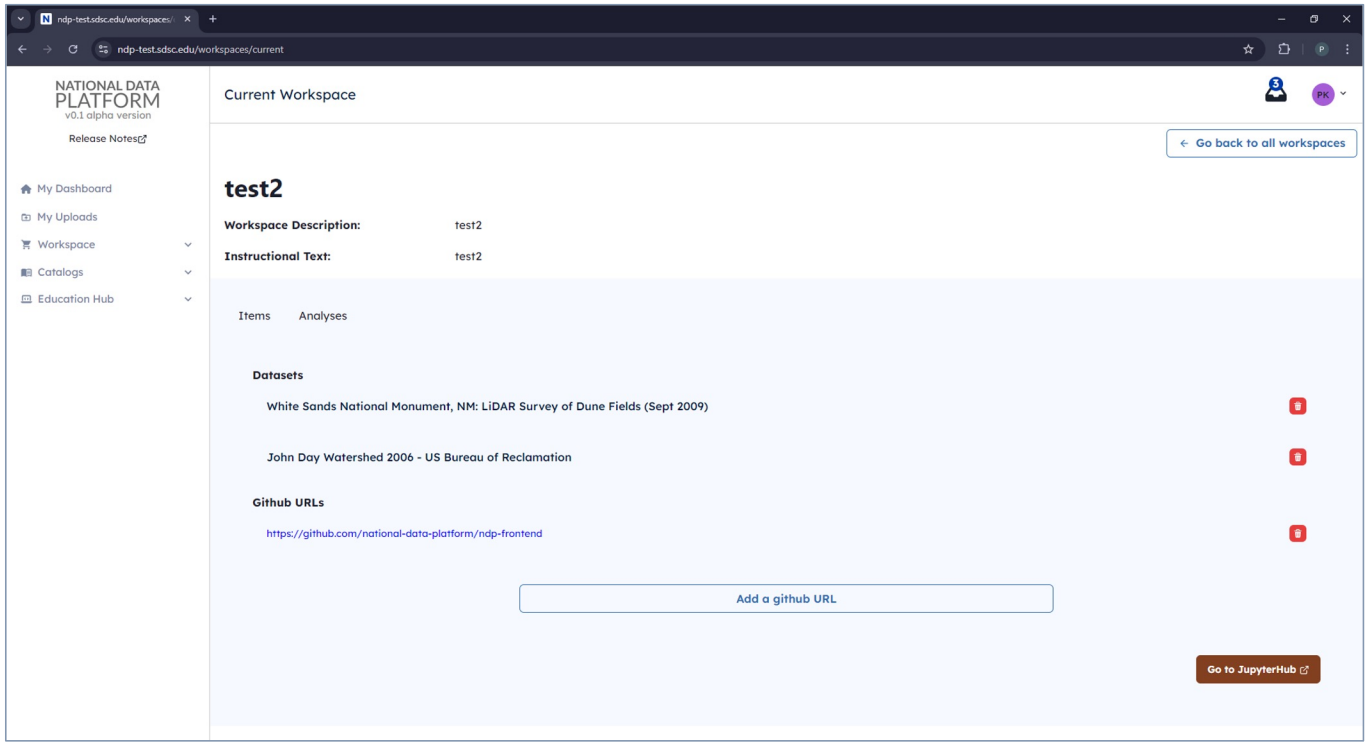


Users can:

- *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.*

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Current Workspaces



Users can:

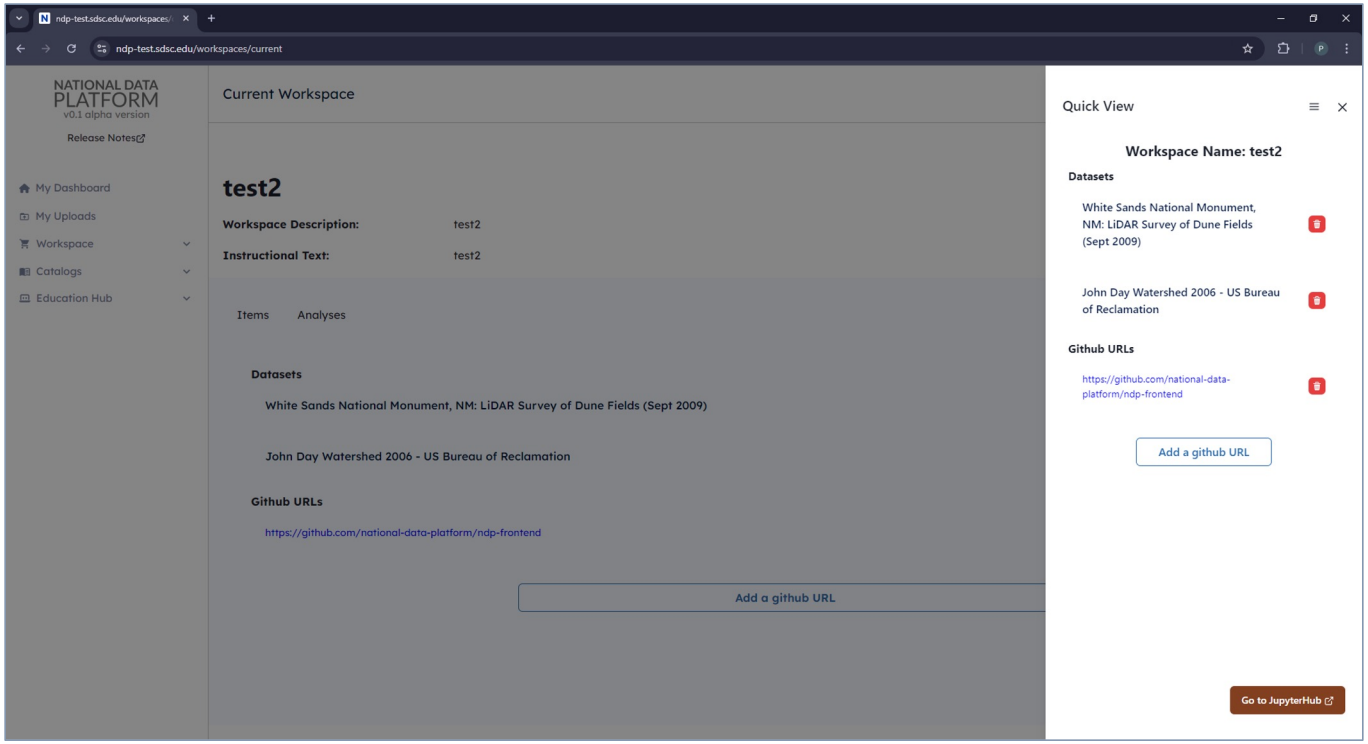
- view their current workspace along with all the details and resources
- delete datasets and add/delete github links from their current workspace
- navigate to JupyterHub using the button at the bottom

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Current Workspace Quick View Drawer

Users can:

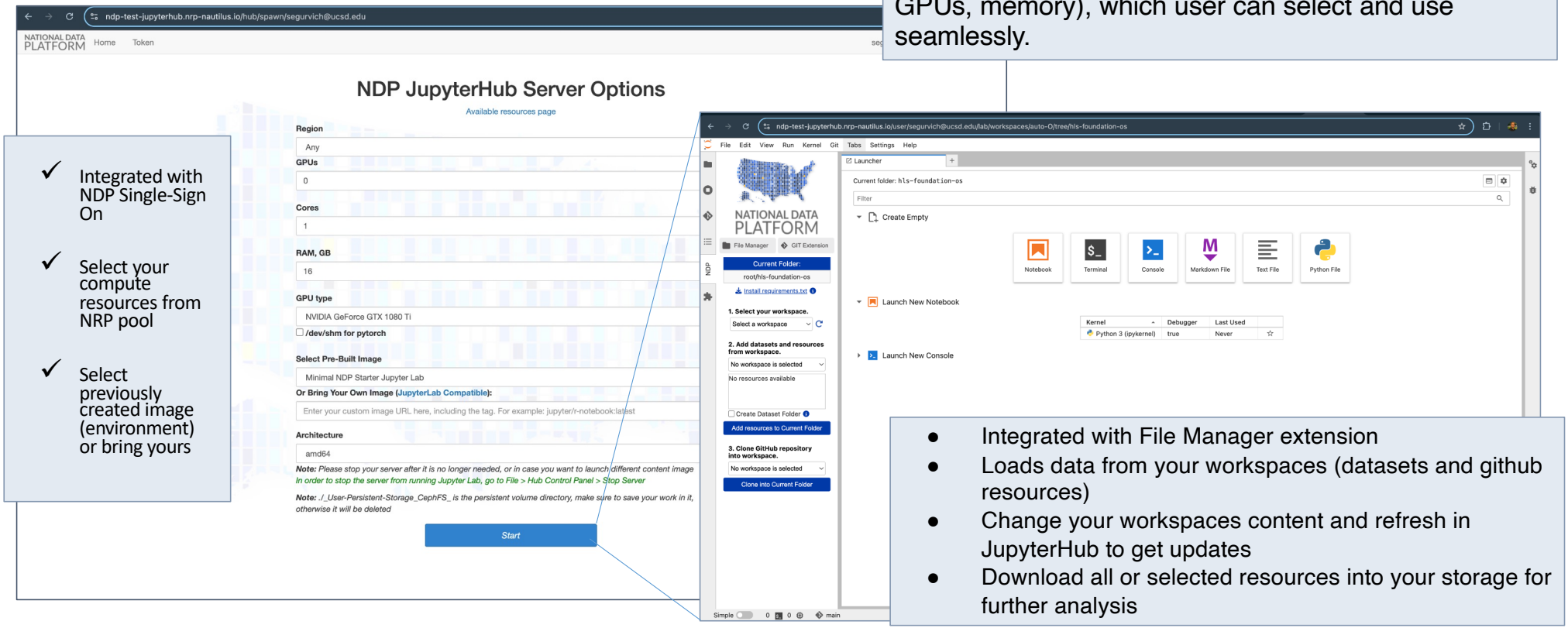
- see a quick view of their current workspace along with all the details and resources
- enter this view by clicking the drawer button on the top right on the navbar next to their avatar
- delete datasets and add/delete github links from their current workspace
- navigate to JupyterHub using the button at the bottom



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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.

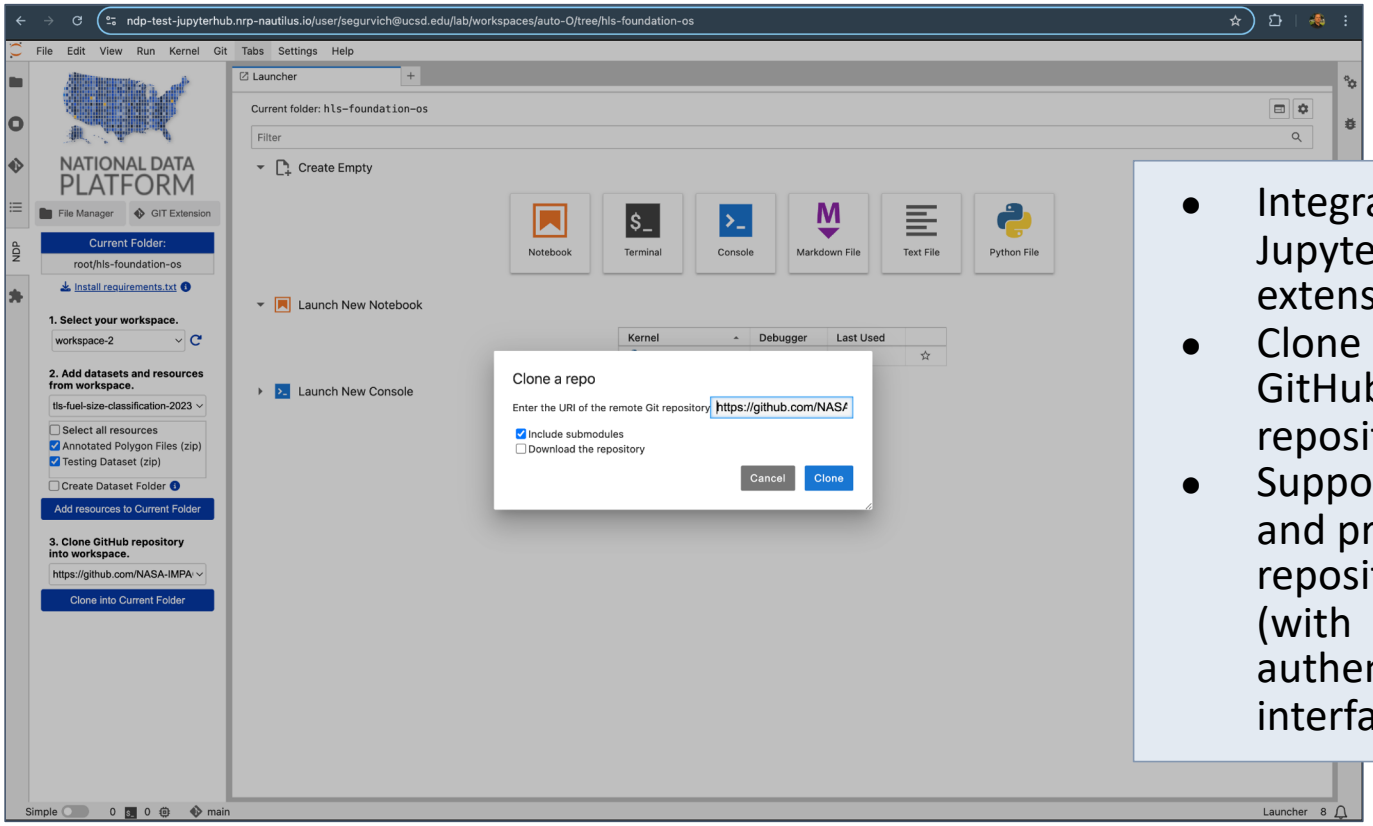


- ✓ Integrated with NDP Single-Sign On
- ✓ Select your compute resources from NRP pool
- ✓ Select previously created image (environment) or bring yours

- 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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

NDP JupyterLab Extension



- Integrated with Jupyter GIT extension
- Clone attached GitHub repositories
- Supports public and private repositories (with authentication interface)

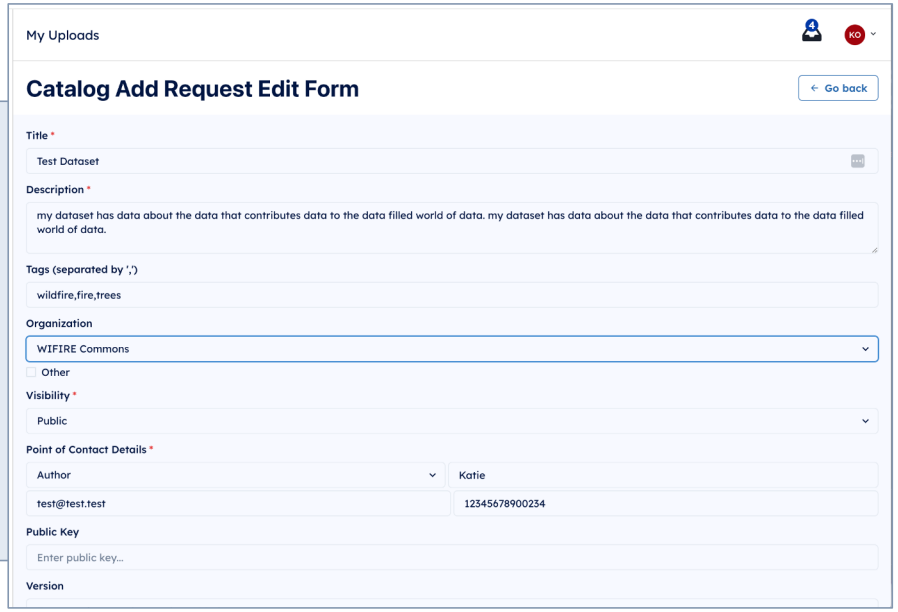
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Education and capacity building through data challenges

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.

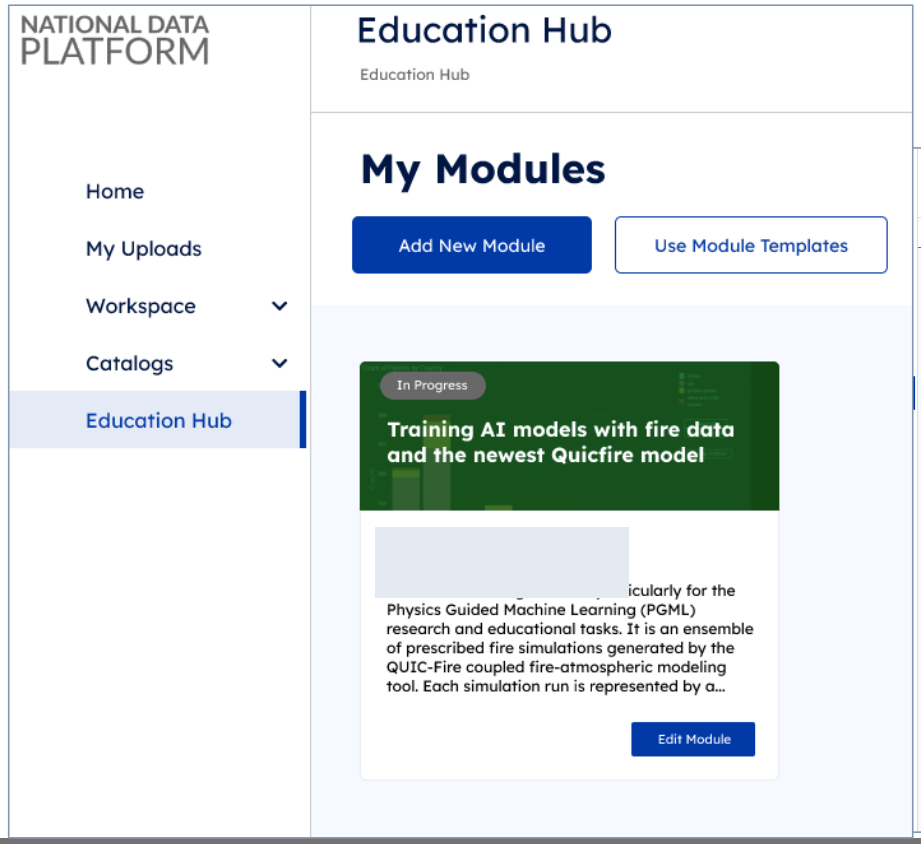
NATIONAL DATA PLATFORM

Three Co-Design Workshops

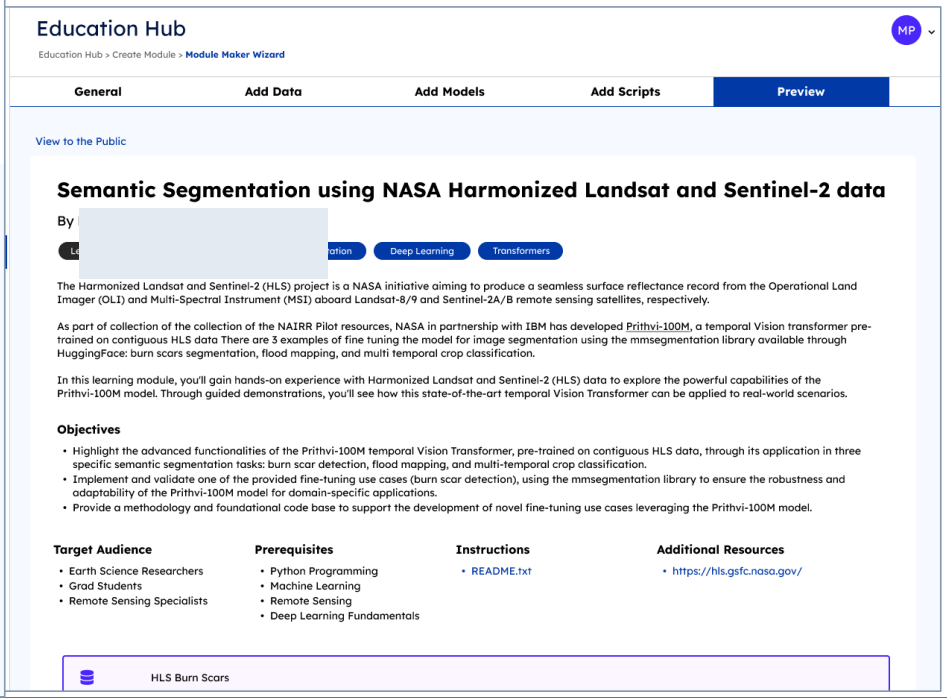
Each will include a breakout session to develop a data challenge question specific to large data (W1); streaming data (W2); and graph data (W3).

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)



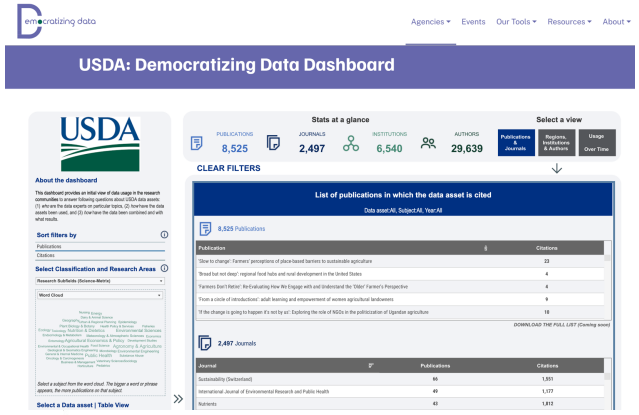
New Learning Module Development Wireframes

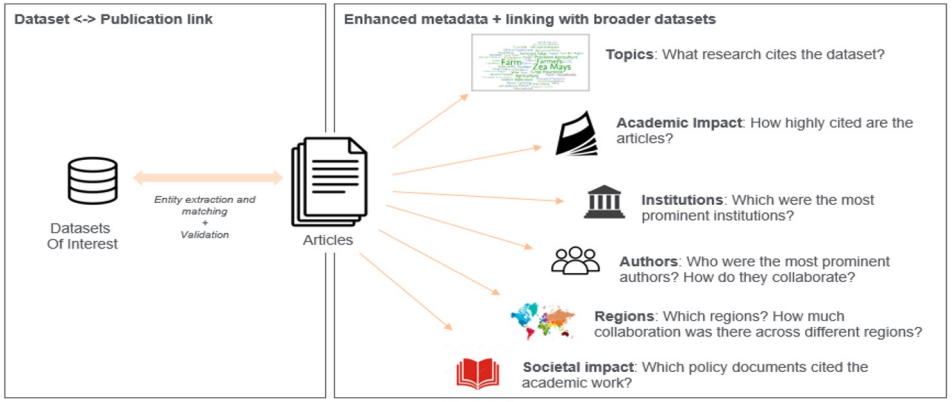


İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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/





İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

NDP PoP Examples & Documentation

September 2024 Release

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

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

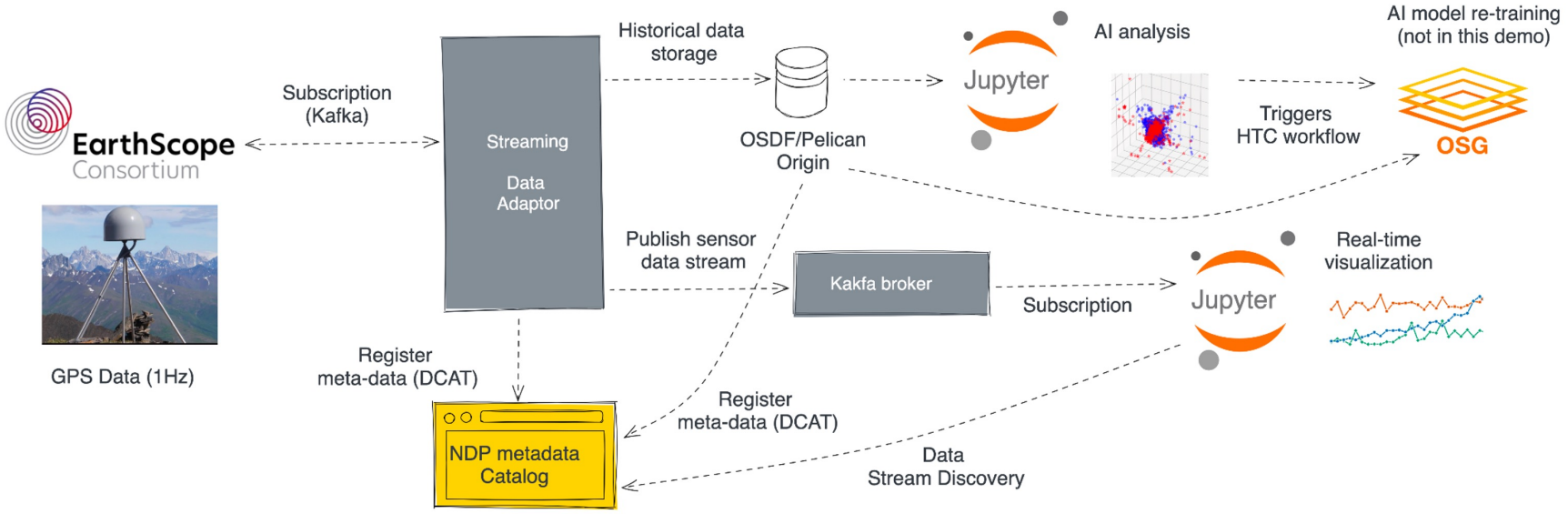
- Streams registration, curation/archival for discovery and access
- User-defined operations on streaming data (semantically specialized abstractions)
- Combine streaming data with archived/playback data
- Mechanism for online data product generation (i.e., new data streams

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

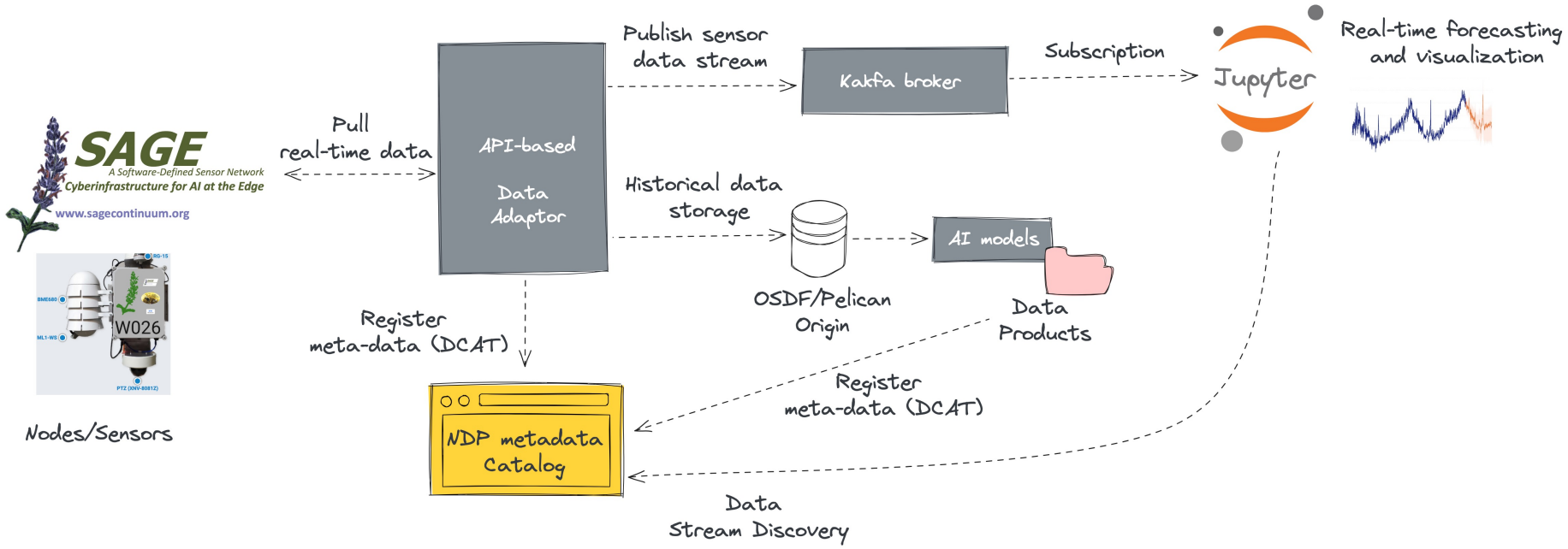


İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Example 2: SAGE data streaming/analysis enabled by NDP

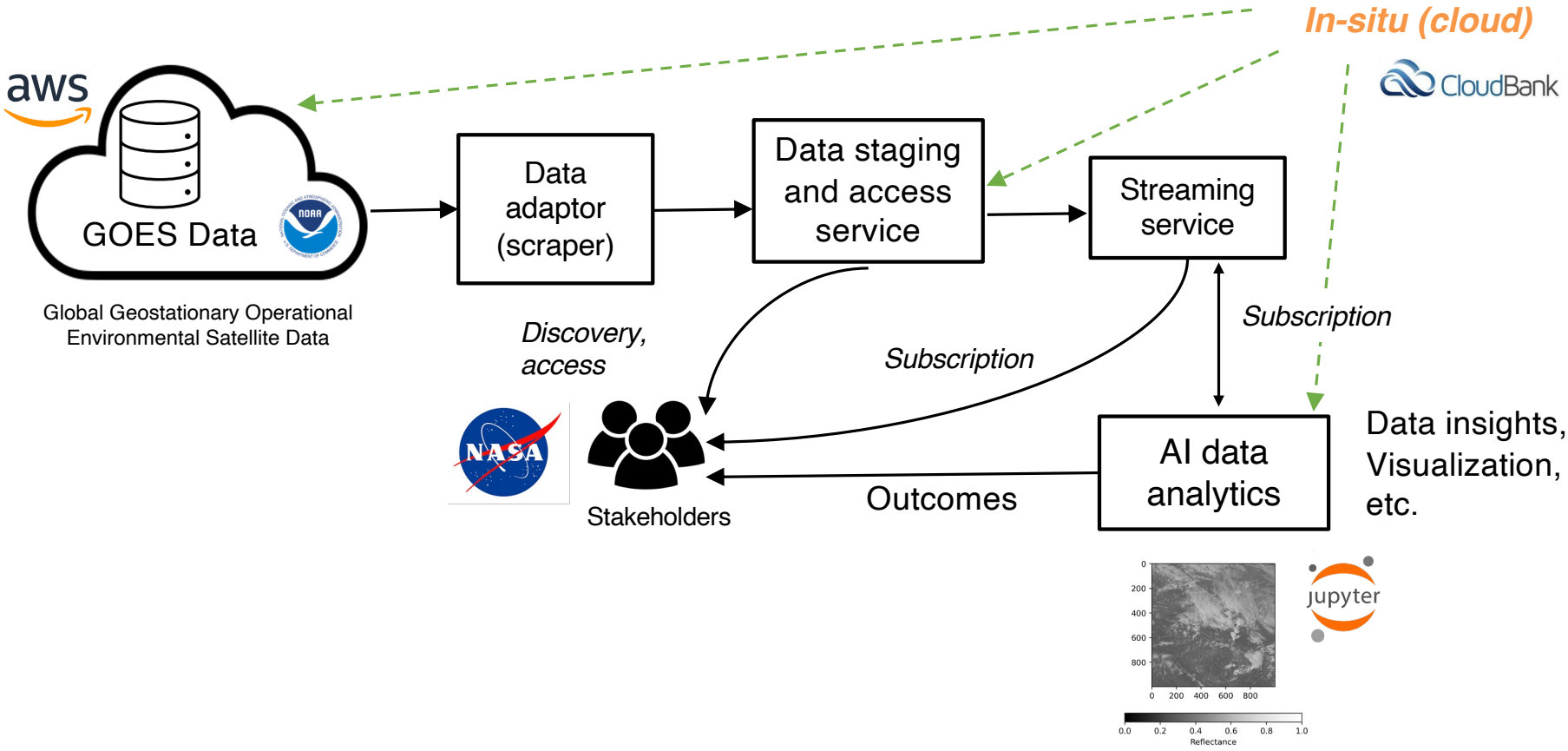
SAGE data streams

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



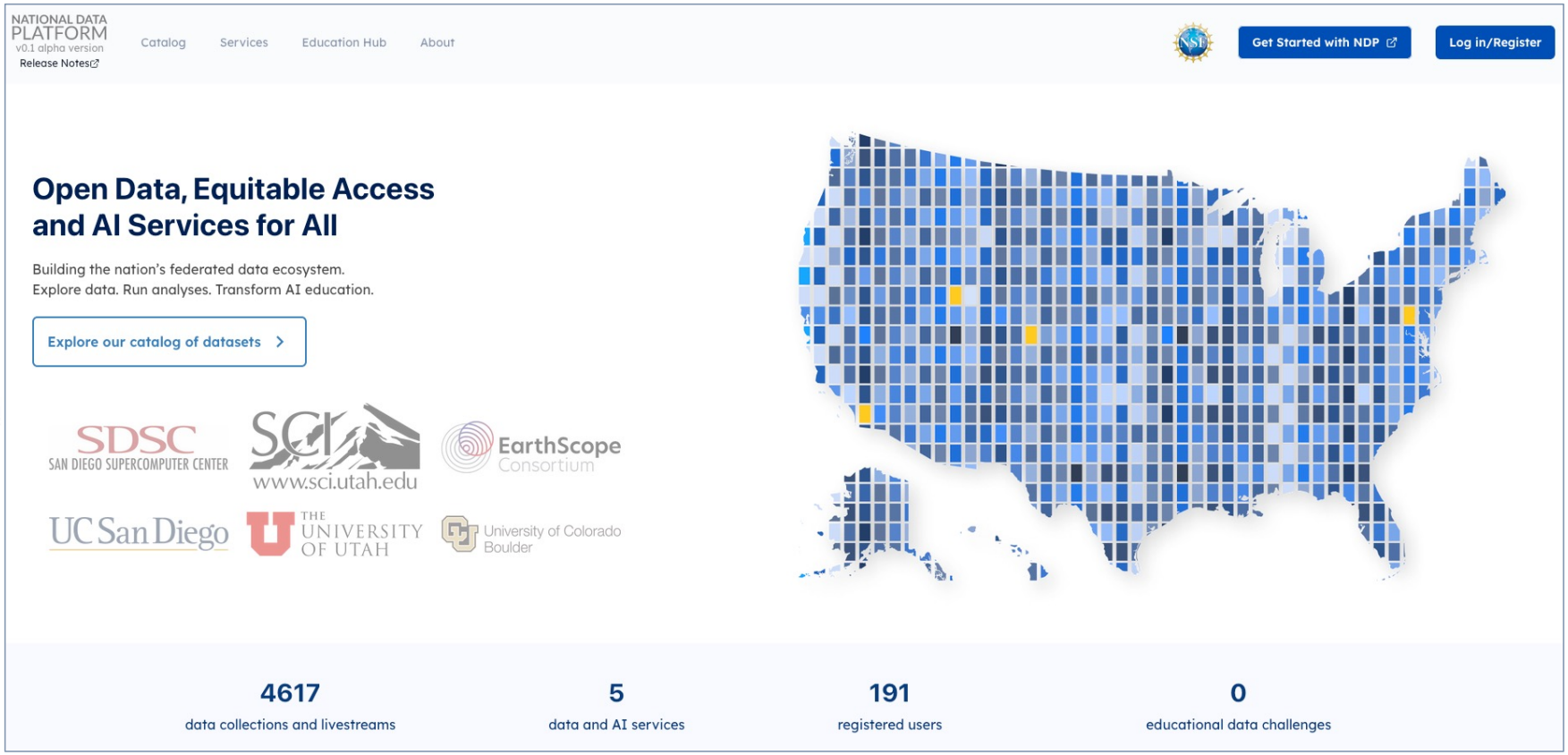
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Example 3: Fire Detection using NASA GOES Satellite Data
(under development)



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Any questions? Contact ndp@sdsc.edu



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

To sum up…

Emerging new applications require integrated AI in dynamically composed workflows.

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

Embrace Complexity!

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)