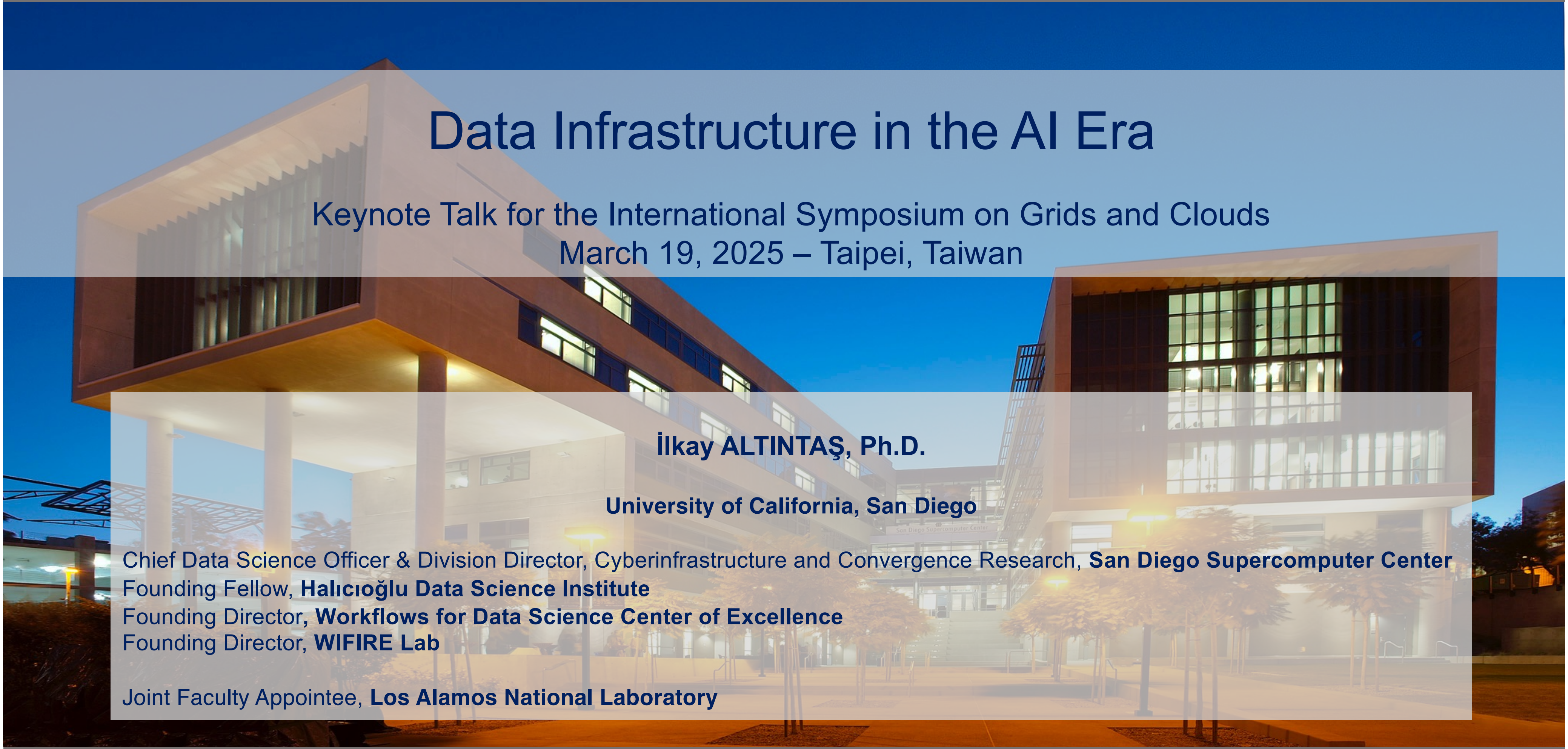


Data Infrastructure in the AI Era

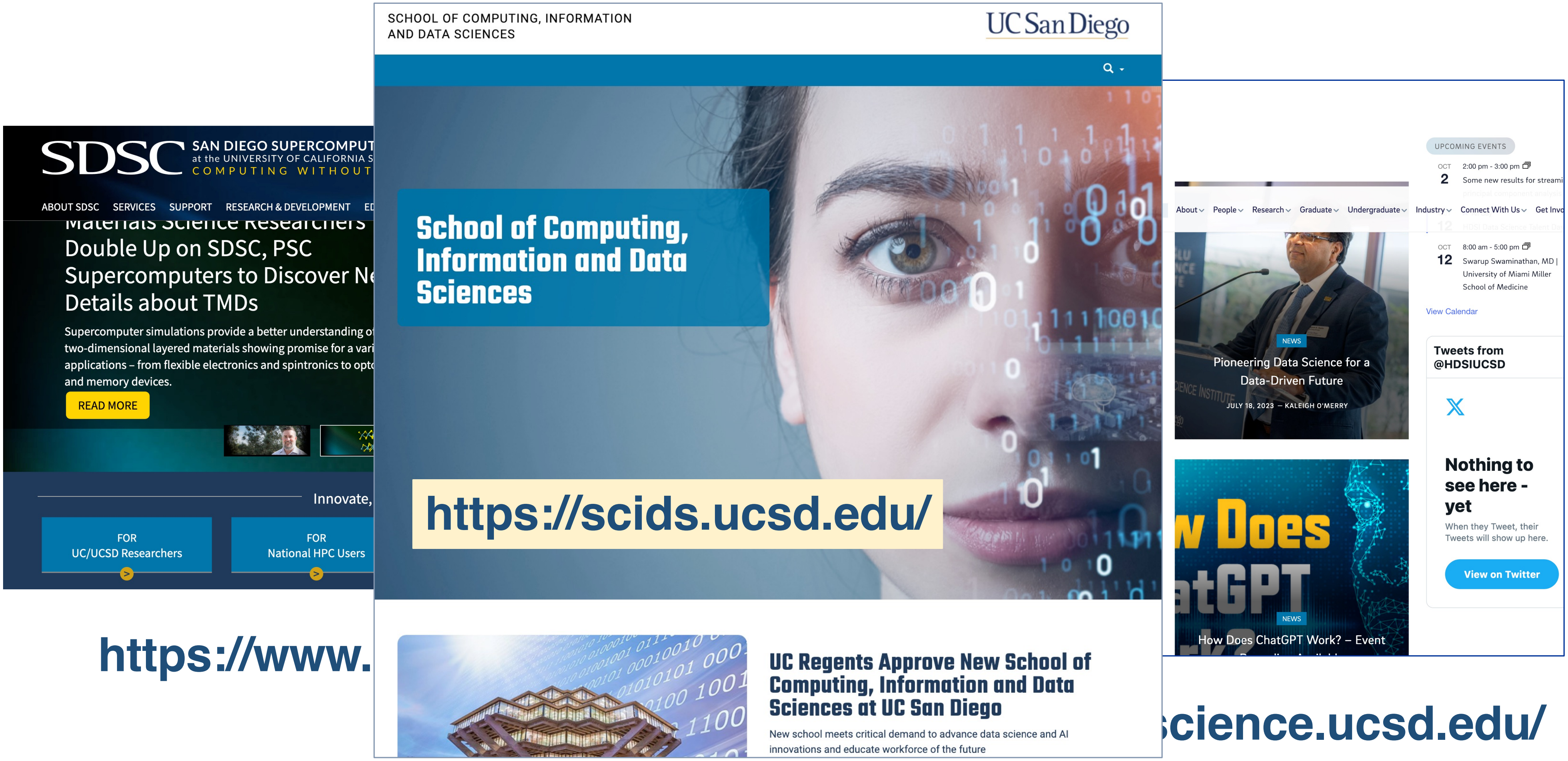
Keynote Talk for the International Symposium on Grids and Clouds
March 19, 2025 – Taipei, Taiwan

İlkay ALTINTAŞ, Ph.D.

University of California, San Diego

Chief Data Science Officer & Division Director, Cyberinfrastructure and Convergence Research, **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**



SDSC SAN DIEGO SUPERCOMPUT
at the UNIVERSITY OF CALIFORNIA S
COMPUTING WITHOUT

ABOUT SDSC SERVICES SUPPORT RESEARCH & DEVELOPMENT ED

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

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

READ MORE

Innovate,

FOR UC/UCSD Researchers

FOR National HPC Users

SCHOOL OF COMPUTING, INFORMATION AND DATA SCIENCES

UC San Diego

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

About People Research Graduate Undergraduate Industry Connect With Us Get Invo

NEWS

Pioneering Data Science for a Data-Driven Future

JULY 18, 2023 — KALEIGH O'MERRY

NEWS

How Does ChatGPT Work? – Event

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 @HDSIUCSD

Nothing to see here - yet

When they Tweet, their Tweets will show up here.

View on Twitter

https://www.

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

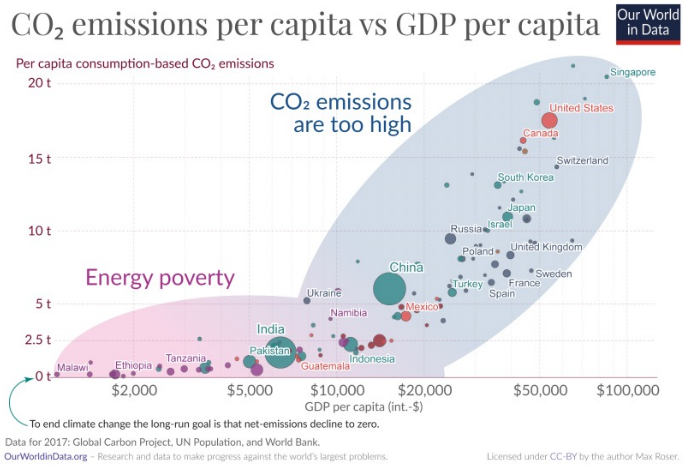
CICORE

Cyberinfrastructure | Convergence Research | Education

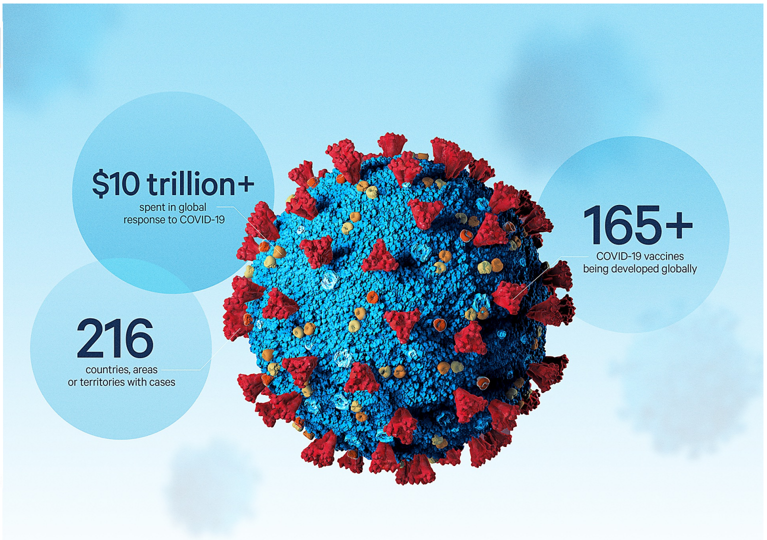
NSF
U.S. DEPARTMENT OF ENERGY | Office of Science
NIH National Institutes of Health
NOAA
U.S. Department of Homeland Security
USDA
Forest Service U.S. Department of Agriculture
NASA
University of California
NIST
Joint Fire Science Program
USGS science for a changing world
Department of Defense United States of America
Cal Fire
SDGE
Cal OES Governor's Office of Emergency Services
Intterra

WCRP
GRAND CHALLENGES
MELTING ICE AND GLOBAL CONSEQUENCES
CLOUDS, CIRCULATION AND CLIMATE SENSITIVITY
REGIONAL SEA LEVEL CHANGE AND COASTAL IMPACTS
WATER FOR THE FOOD BASKETS OF THE WORLD
WEATHER AND CLIMATE EXTREMES
CARBON FEEDBACKS IN THE CLIMATE SYSTEM
NEAR-TERM CLIMATE PREDICTION

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



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

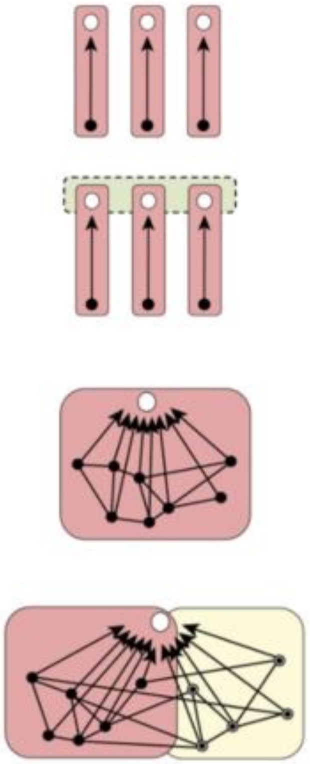
The biggest challenges of our time are too difficult to solve alone!

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 gal 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

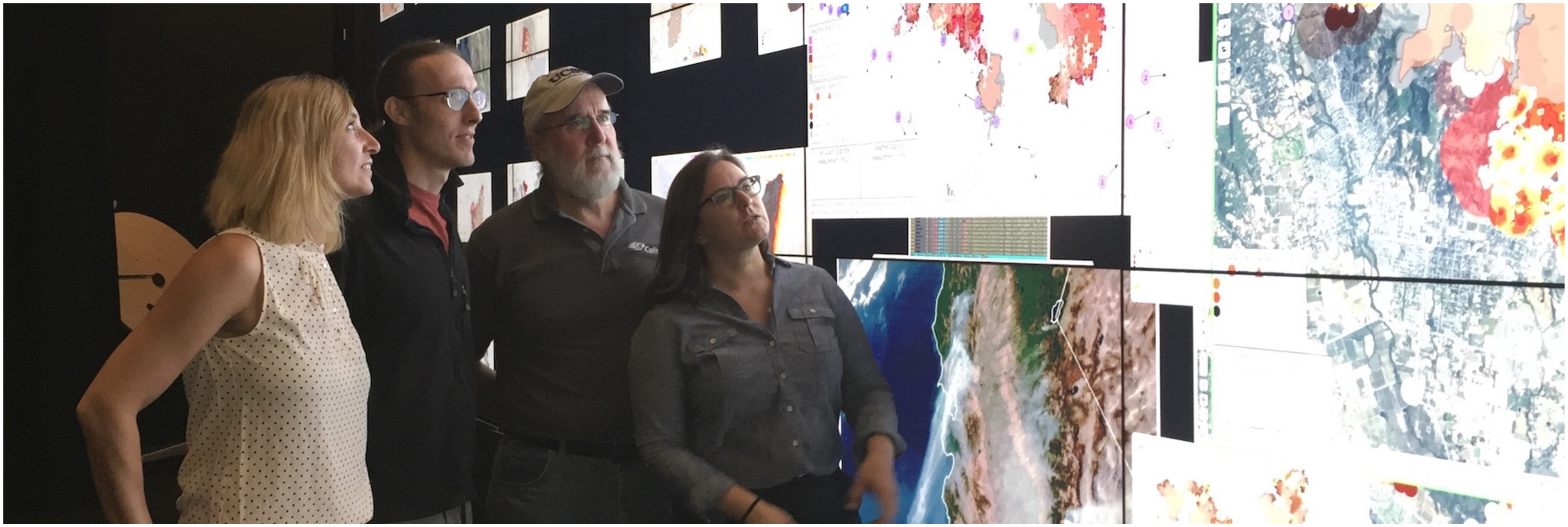
Convergence
- Crosses disciplinary and sectorial boundaries
- Common goal setting
- Develops integrated knowledge for science and society
- Creates new paradigms

◉ 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



SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

Three Main Components

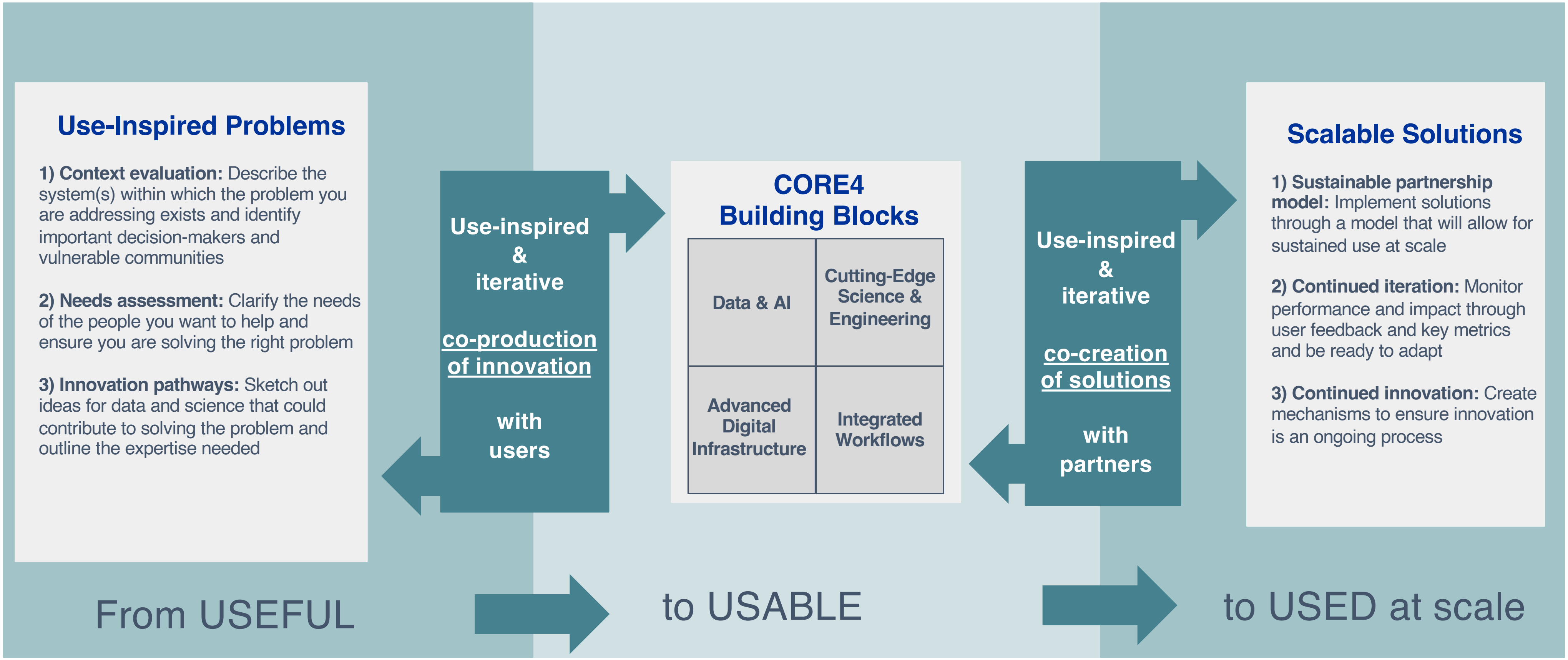
Composable Workflows + Collaborative Innovation + Impact Network

CORE INSTITUTE

https://www.core-institute.org/

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

Translating Fire Research into Impact

WIFIRE

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

Actionable Open Fire Science and AI:

Right Model and Right Data
for the Right Decision Support Workflow
at the Right Time
with the Right Communication

…before, during, and after a fire.

Where are we headed at WIFIRE?

- **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
- **Community Data Platforms:** WIFIRE's ***Wildfire Science & Technology Commons*** and ***Wildfire and Landscape Resilience Data Hub*** to develop standards, tools and techniques to share data and data-driven models with partners including NIST, CAL FIRE, and SDGE
- **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
- **Immersive Fire Environment:** WIFIRE's ***Immersive Forest Project*** leverages the AI-readiness of scientific data for new modes of teaching, training, decision-making, and public communication,
- Our platforms and products are fueled by over a dozen research projects and partnerships focused on ***moving science to practice***

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.

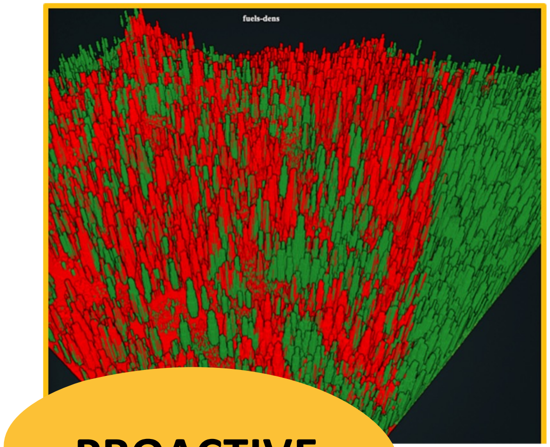
REACTIVE

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 U S DEPARTMENT OF AGRICULTURE

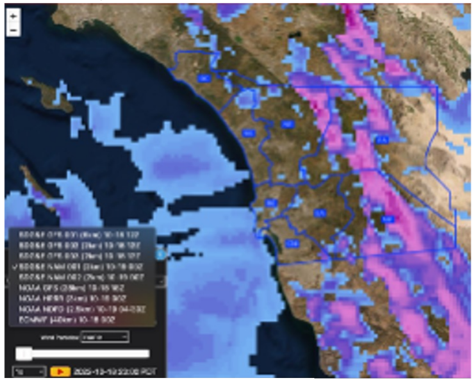
NSF

Data and Computing Platforms

WIFIRE LAB

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.



NSF NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE SDGE

Wildfire and Landscape Resilience Data Hub

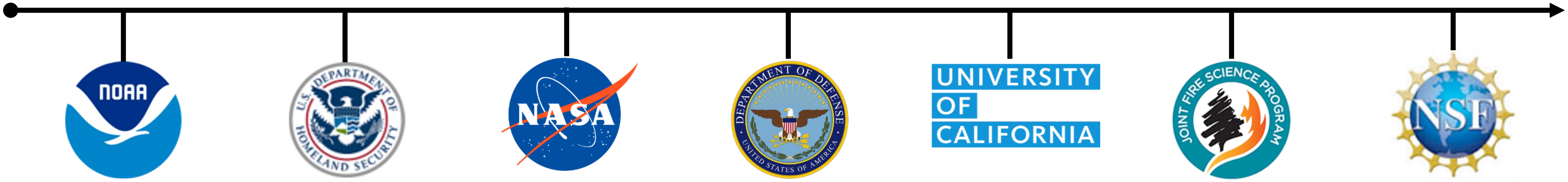
The Data Hub is a federated data ecosystem for California's Wildfire and Forest Resilience Task Force, providing a "single view" over existing data to fulfill the reporting requirements for California's Million Acre Strategy to treat 1 million forested acres per year to reduce wildfire risk. It will 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.

THE FRONT LINES OF LANDSCAPE HEALTH

In a response commensurate to the challenge, Governor Gavin Newsom created the California Wildfire & Forest Resilience Task Force, bringing together an unprecedented coalition of the best available resources for preventing catastrophic wildfires by creating healthier, more sustainable natural environments. The Governor's Task Force is a proactive effort that is already showing progress towards long term forest health and safe, sustainable coexistence with fire.

CAL FIRE

Additional Grants Fueling R&D


NOAA
U.S. DEPARTMENT OF HOMELAND SECURITY
NASA
DEPARTMENT OF DEFENSE UNITED STATES OF AMERICA
UNIVERSITY OF CALIFORNIA
JOINT FIRE SCIENCE PROGRAM
NSF

- 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

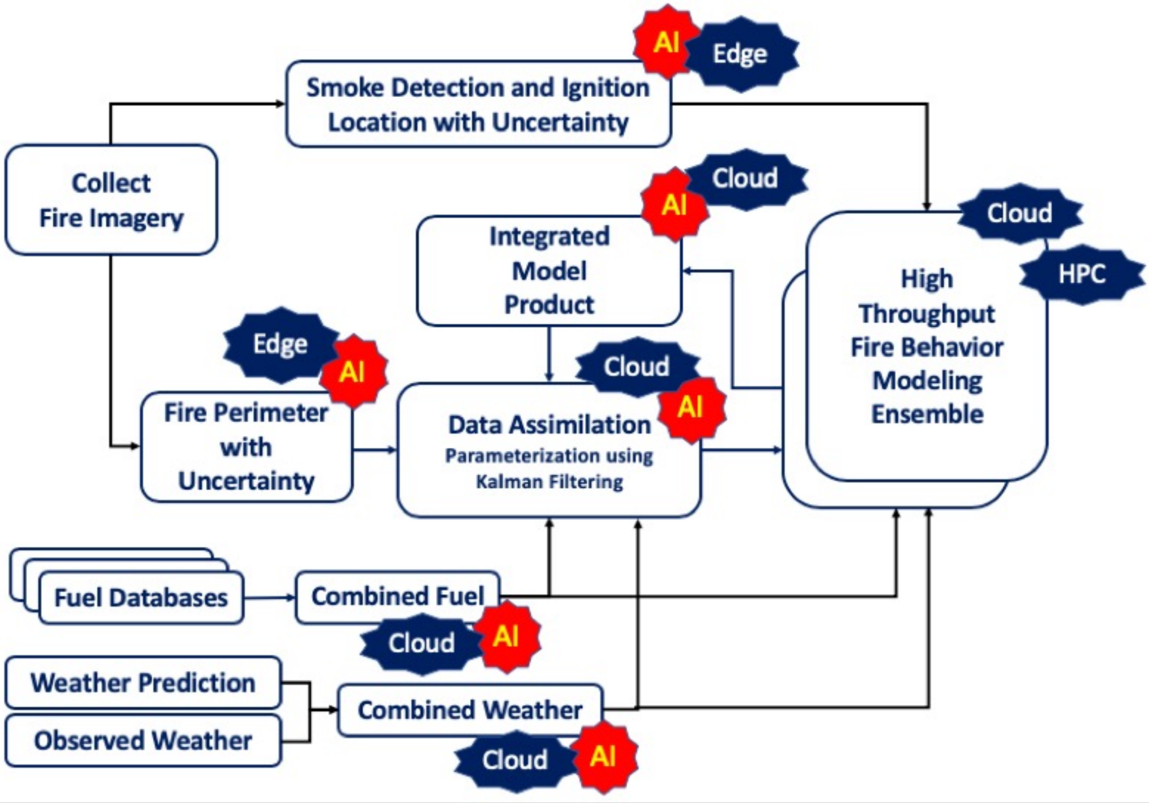
AI and Computing Needs for Dynamic Data-Driven Fire Modeling

-- Characterizing the dynamic fire environment : Variation of wind, smoke, moisture, fuels, fire perimeter, …

-- Detection of fire ignitions

-- Decision support for fire management

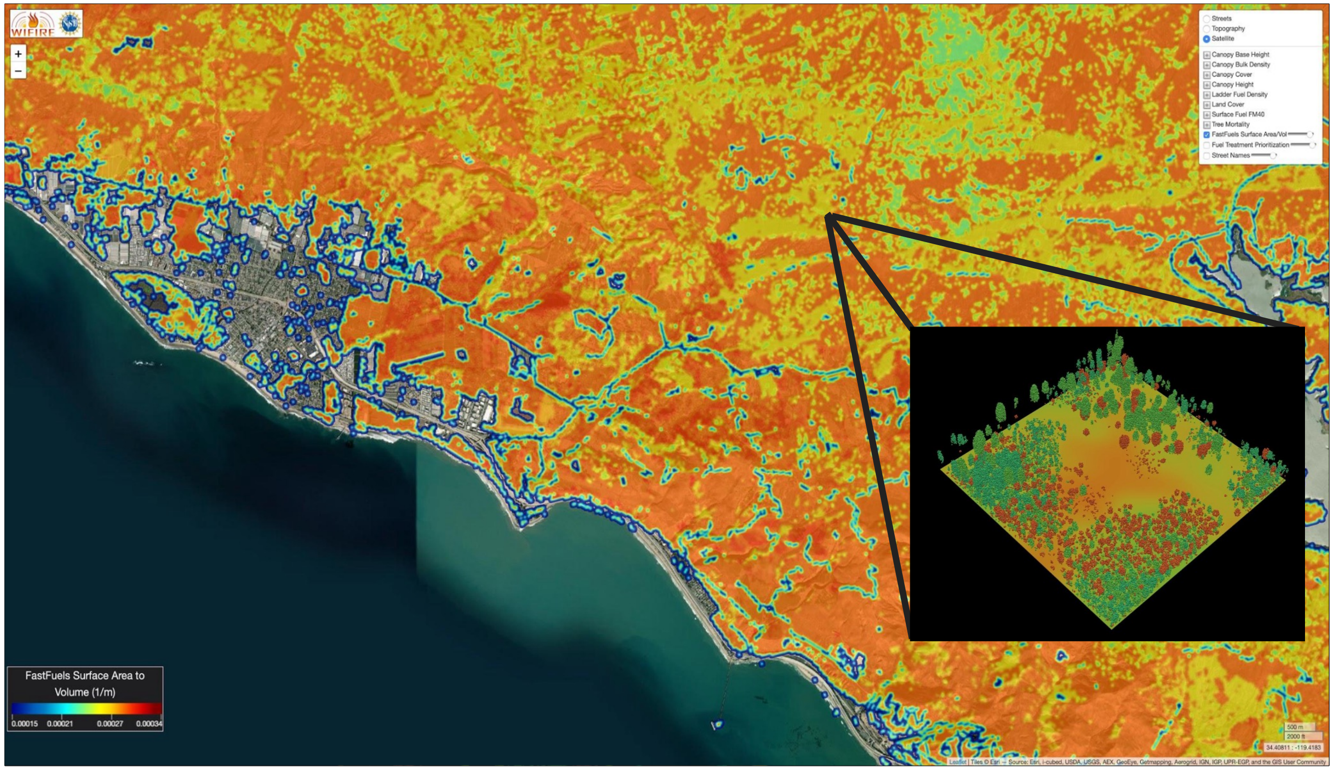
-- Prediction of potential fire ignitions



AI Techniques to Condition Data and Improve Model Accuracy

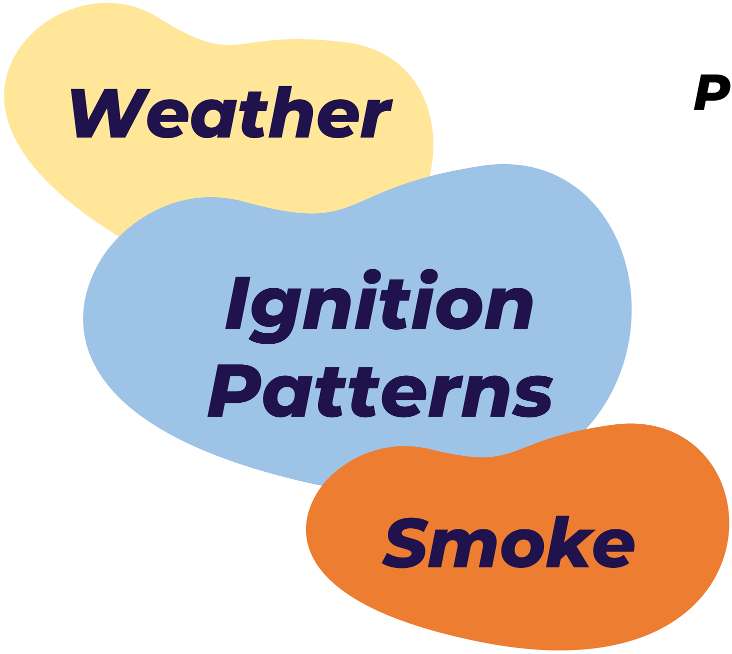
3D vs 2D

900x more detailed



Collaboration with Rod Linn (LANL), Kevin Hiers (TTRS) and Russ Parsons (USFS)

AI Techniques to Improve Decision Making



Weather

Ignition Patterns

Smoke

PHYSICS-GUIDED MACHINE LEARNING

To improve predictive fire behavior models

OPTIMIZATION

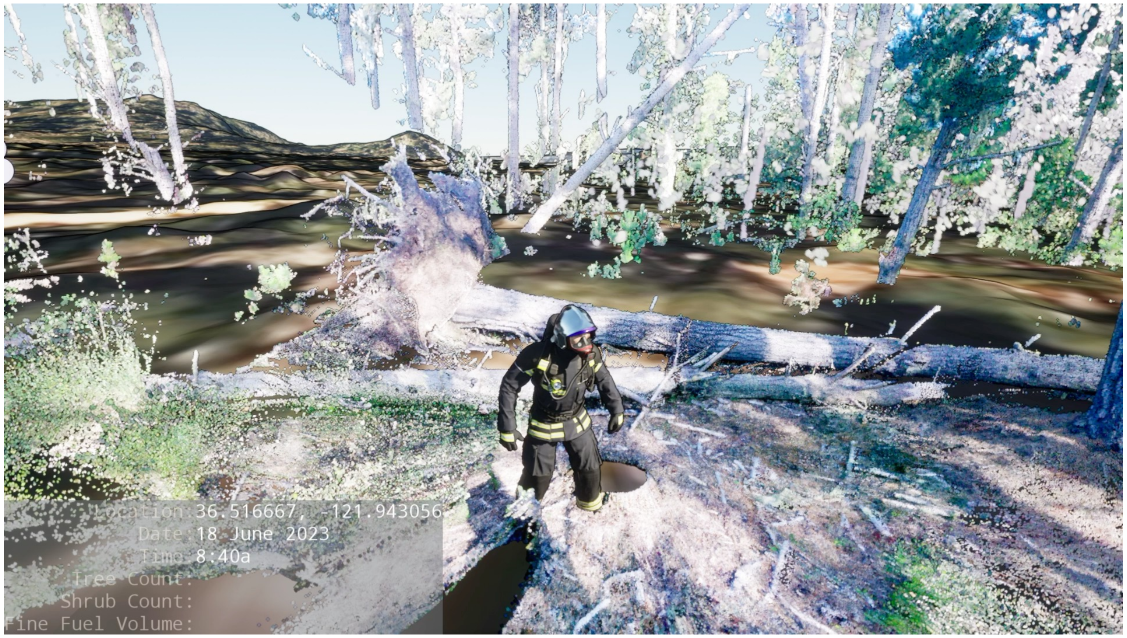
To address complex tradeoffs and prioritization

EXPLAINABLE AI

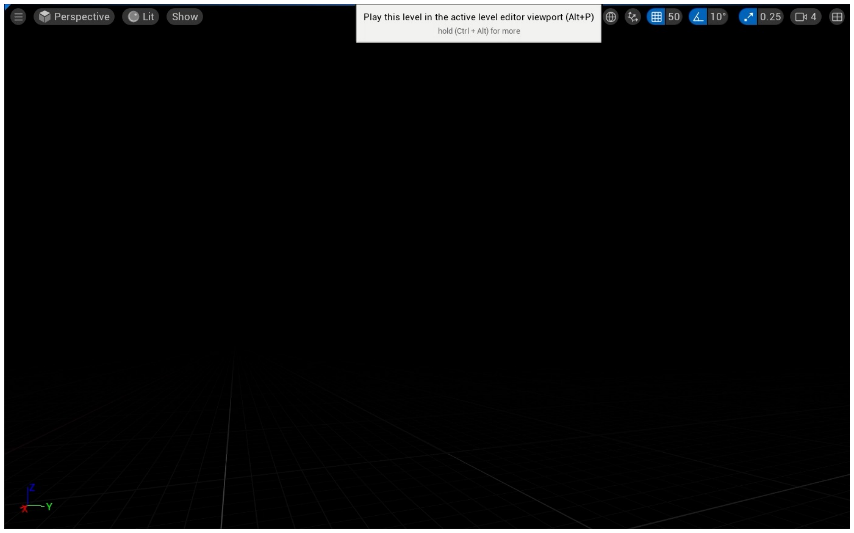
To increase scientific understanding and interpretability all along the decision-making chain

AI in Science Communication

Visualization of multiple terrestrial LiDAR scans in the Immersive Forest prototype



Location:36.516667, -121.943056
Date: 18 June 2023
Time: 8:40a
Tree Count:
Shrub Count:
Fine Fuel Volume:


Perspective
Lit
Show
Play this level in the active level editor viewport (Alt+P)
hold (Ctrl + Alt) for more
50
10°
0.25
4
Y

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)

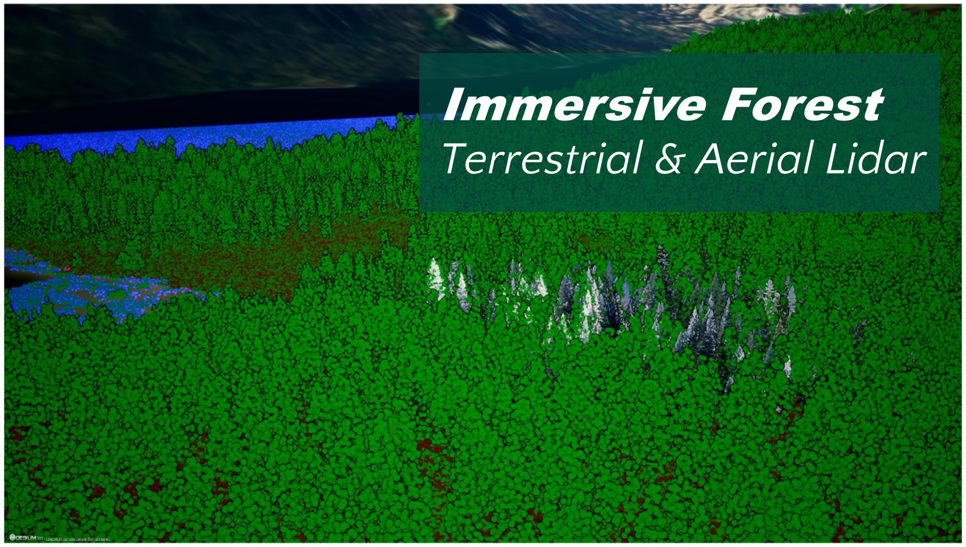
SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

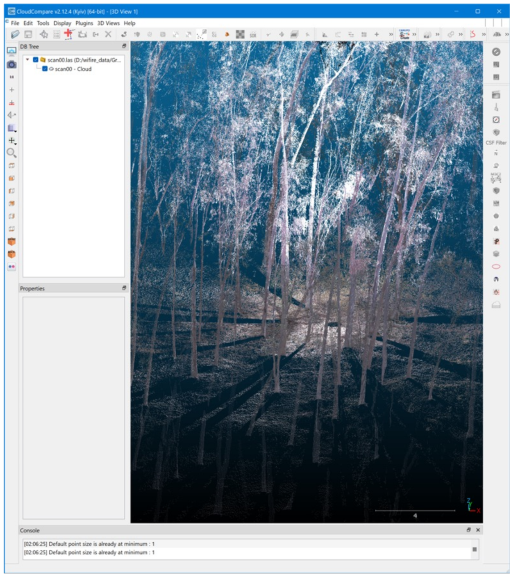
Immersive Fire Digital Twin

WIFIRE's **Immersive Forest** leverages the 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.

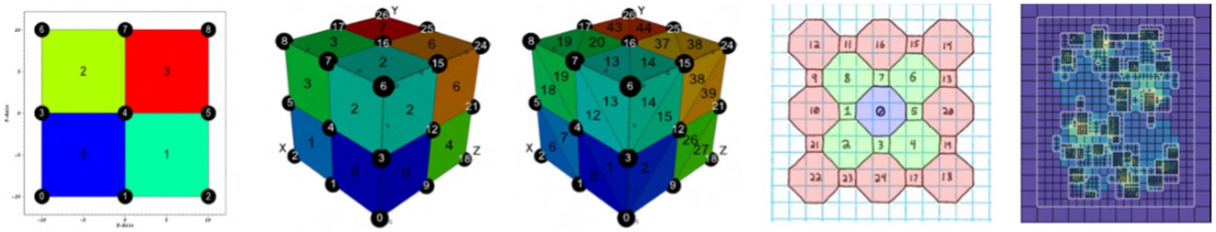


Immersive Forest
Terrestrial Lidar

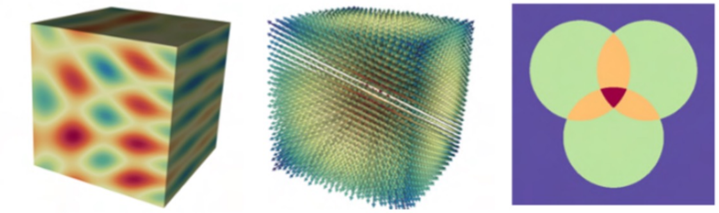
Many Scientific Data Types in Digital Twin


CloudCompare v2.12.4 (Kyiv) [64-bit] - [3D View 1]
File Edit Tools Display Plugins 3D Views Help
DB Tree
scan00.las (D:/wifire_data/Gr...
scan00 - Cloud
Properties
CSF Filter
4
Console

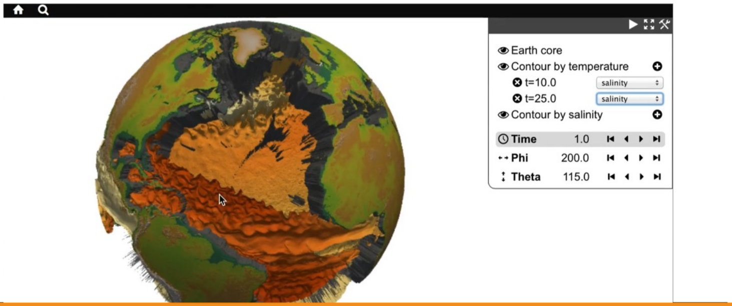
```
[02:06:25] Default point size is already at minimum : 1
[02:06:25] Default point size is already at minimum : 1
```



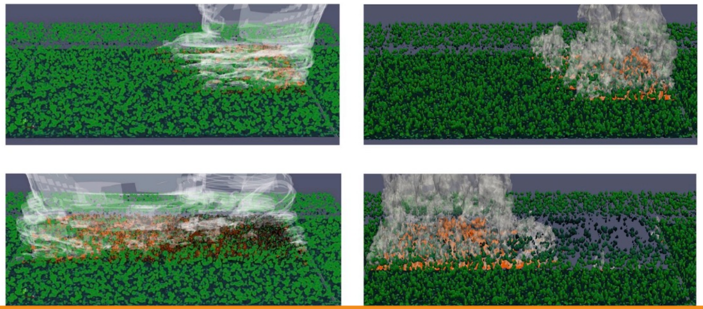
Topologies: 1D/2D/3D - Uniform, Rectilinear, Structured, Unstructured, Polygonal, Polyhedral, AMR

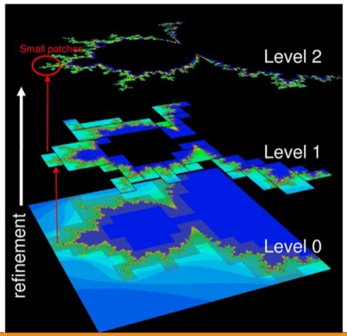


Fields: Scalar, Vector, Multi-material

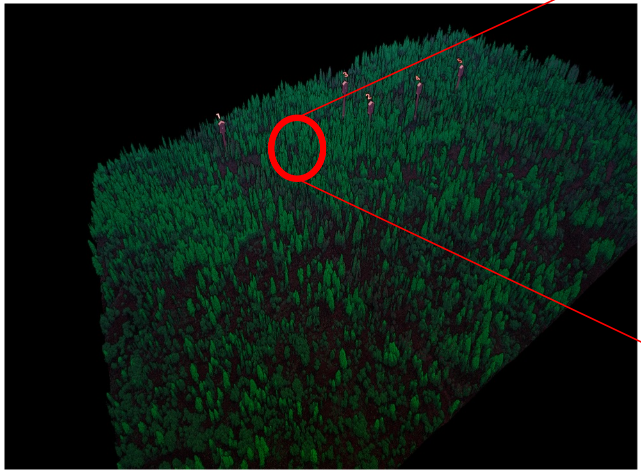


Earth core
Contour by temperature
t=10.0 salinity
t=25.0 salinity
Contour by salinity
Time 1.0
Phi 200.0
Theta 115.0

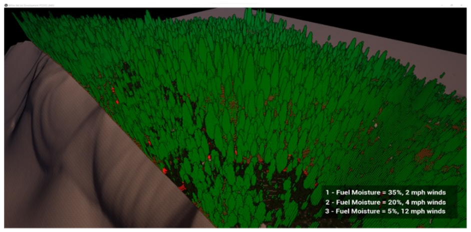


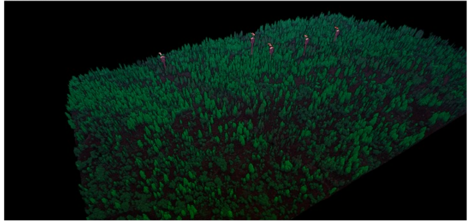


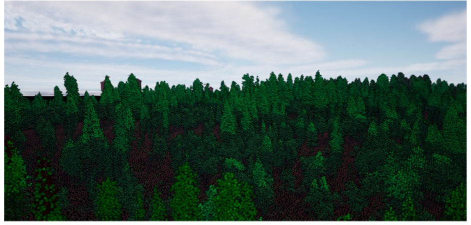
Immersive Forest for Multimodal Communication



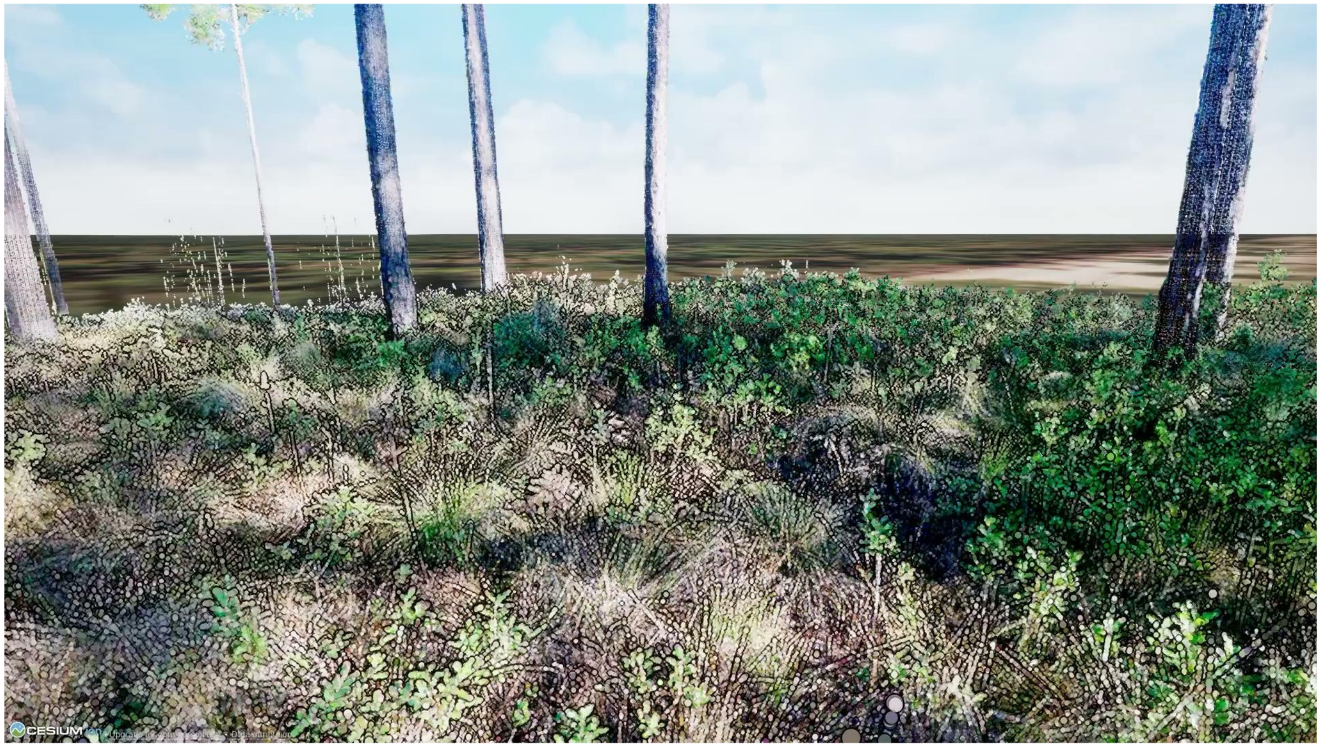
Terrestrial LiDAR contextualized within Aerial scan


1 - Fuel Moisture = 35%, 2 mph winds
2 - Fuel Moisture = 20%, 4 mph winds
3 - Fuel Moisture = 5%, 12 mph winds



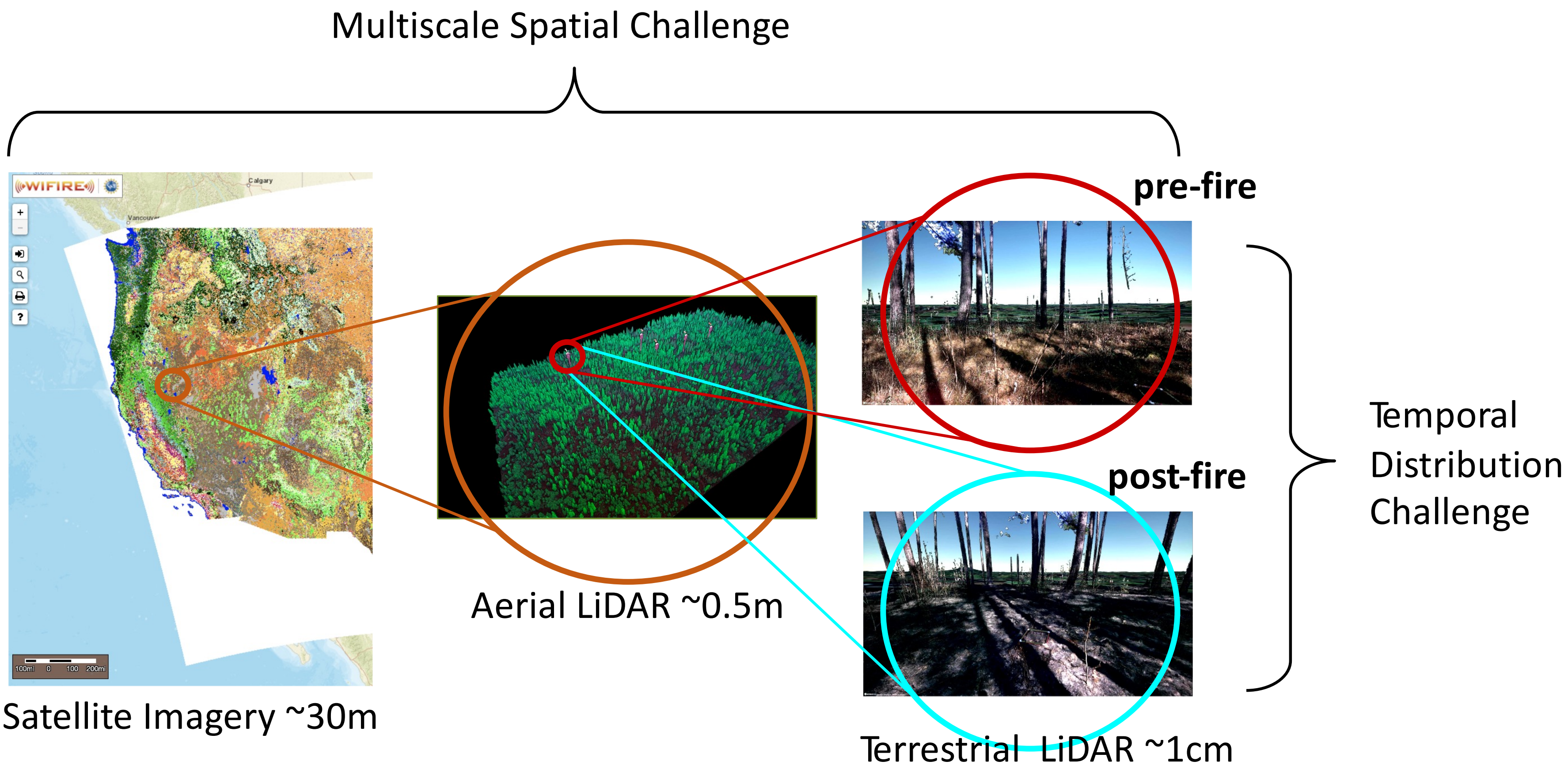


LiDAR Processing & Visualization



I. Moreno, I. Nealey, D. Roten, M. Nguyen, D. Crawl, K. O'Laughlin, M. Floca, S. Pokswinski, and İ. Altıntaş, "Visualization and Labeling of Terrestrial LiDAR Data for Three-Dimensional Fuel Classification," in Proceedings of the IEEE eScience 2023 Conference, 2023, pp. 1-2. doi: 10.1109/e-Science.58273.2023.10254841.

Knowledge Representation: Spatiotemporal Data and its Challenges



Multiscale Spatial Challenge

pre-fire

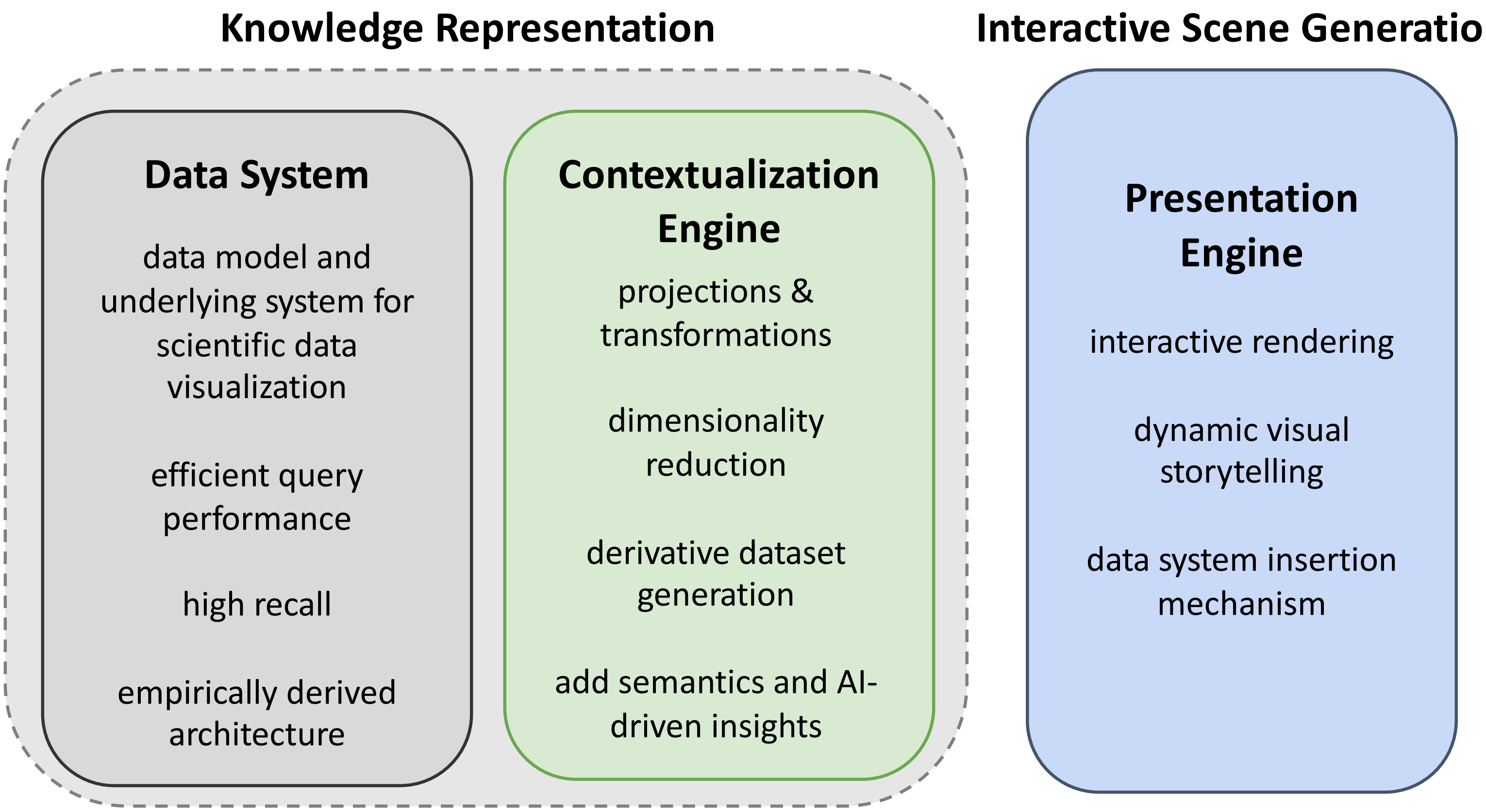
post-fire

Temporal Distribution Challenge

Satellite Imagery ~30m

Aerial LiDAR ~0.5m

Terrestrial LiDAR ~1cm



Knowledge Representation

Interactive Scene Generation

Data System

data model and underlying system for scientific data visualization

efficient query performance

high recall

empirically derived architecture

Contextualization Engine

projections & transformations

dimensionality reduction

derivative dataset generation

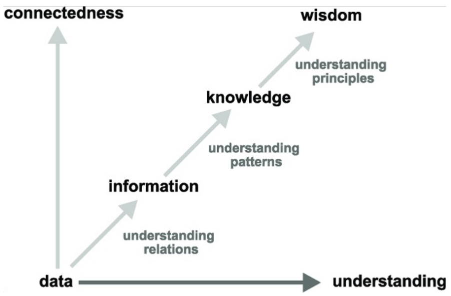
add semantics and AI-driven insights

Presentation Engine

interactive rendering

dynamic visual storytelling

data system insertion mechanism



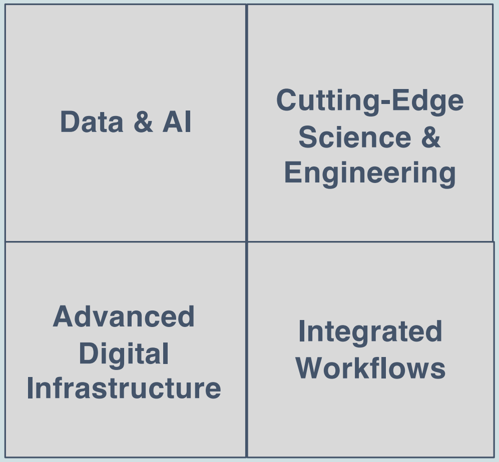
Data → Information → Knowledge → Interpretation and Stories

This type of work needs the CORE4 building blocks.

CICORE

Cyberinfrastructure | Convergence Research | Education

CORE4 Building Blocks



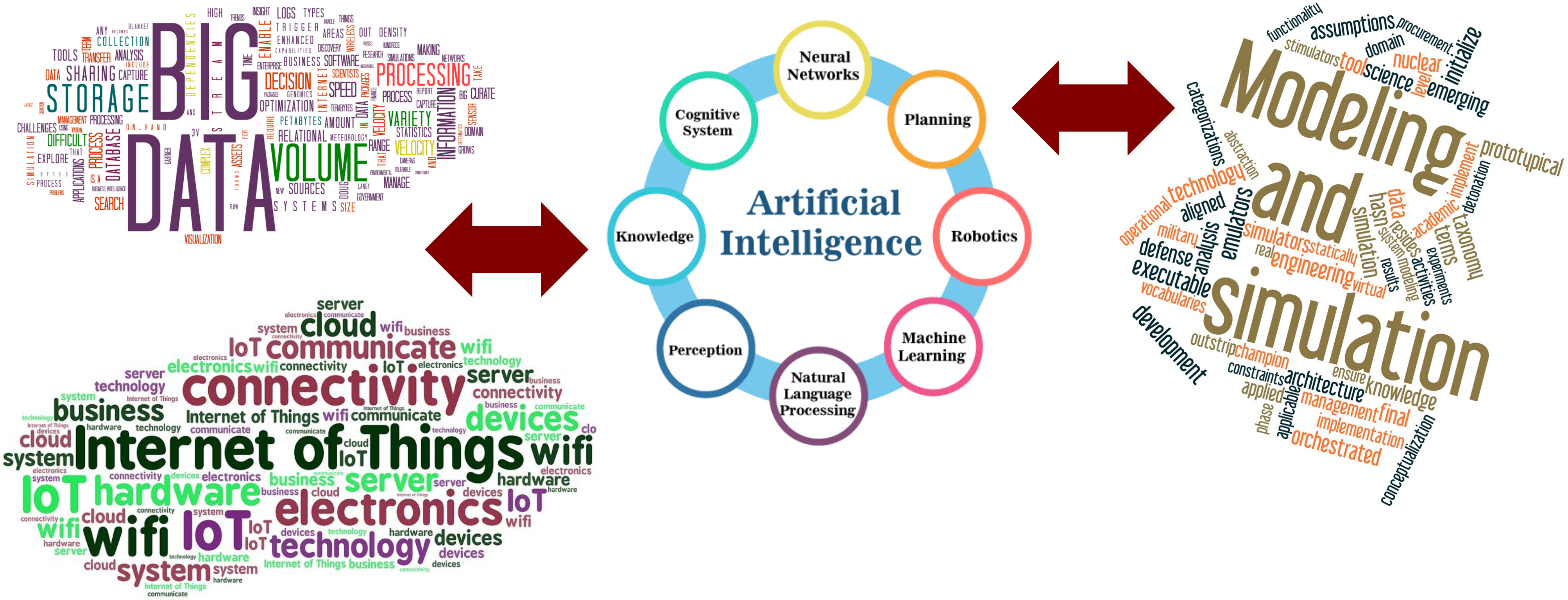
Data & AI

Cutting-Edge Science & Engineering

Advanced Digital Infrastructure

Integrated Workflows

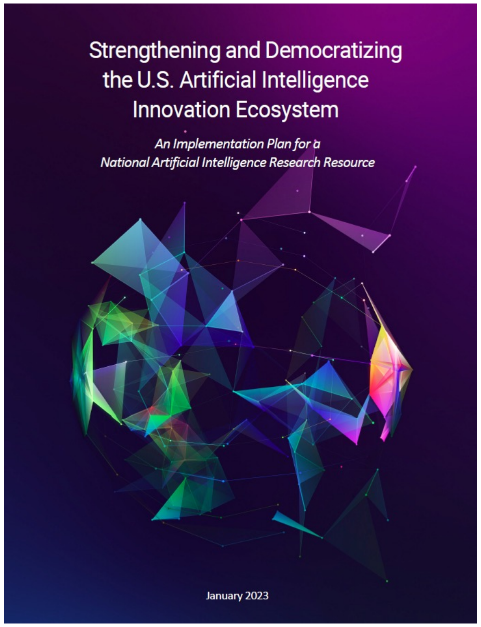
AI-Integrated Applications at the Digital Continuum



Artificial Intelligence in Science

CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH

OECD



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

An Implementation Plan for a National Artificial Intelligence Research Resource

January 2023

CSIRO

Australia's National Science Agency

Artificial intelligence for science

Adoption trends and future development pathways

November 2022

ANL-22/91

ADVANCED RESEARCH DIRECTIONS ON

AI FOR SCIENCE, ENERGY, AND SECURITY

Report on Summer 2022 Workshops

Jonathan Carter
Lawrence Berkeley National Laboratory

John Feddema
Sandia National Laboratories

Doug Kothe
Oak Ridge National Laboratory

Rob Neely
Lawrence Livermore National Laboratory

Jason Pruet
Los Alamos National Laboratory

Rick Stevens
Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY · U.S. DEPARTMENT OF ENERGY Office of Science · NNSA

May 2023

AI in Science and Research

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

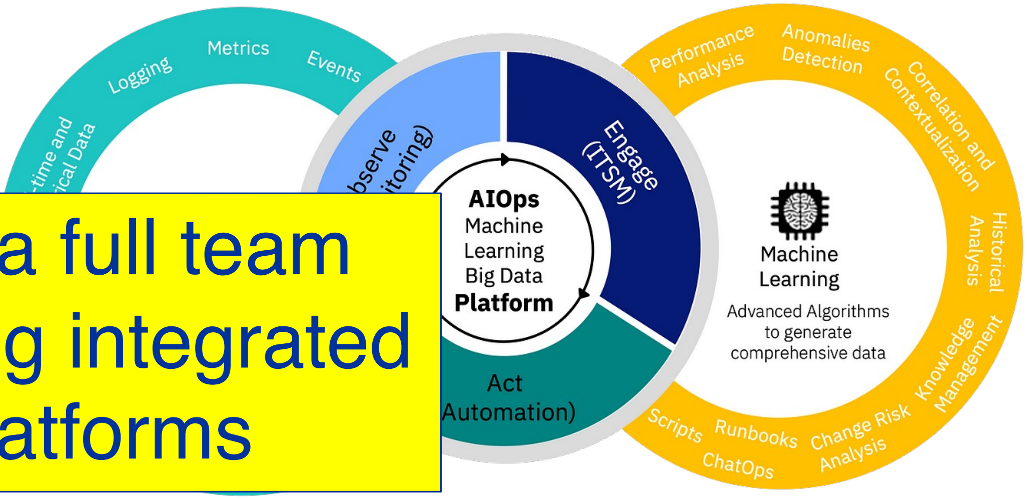
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

AI in Science Readiness

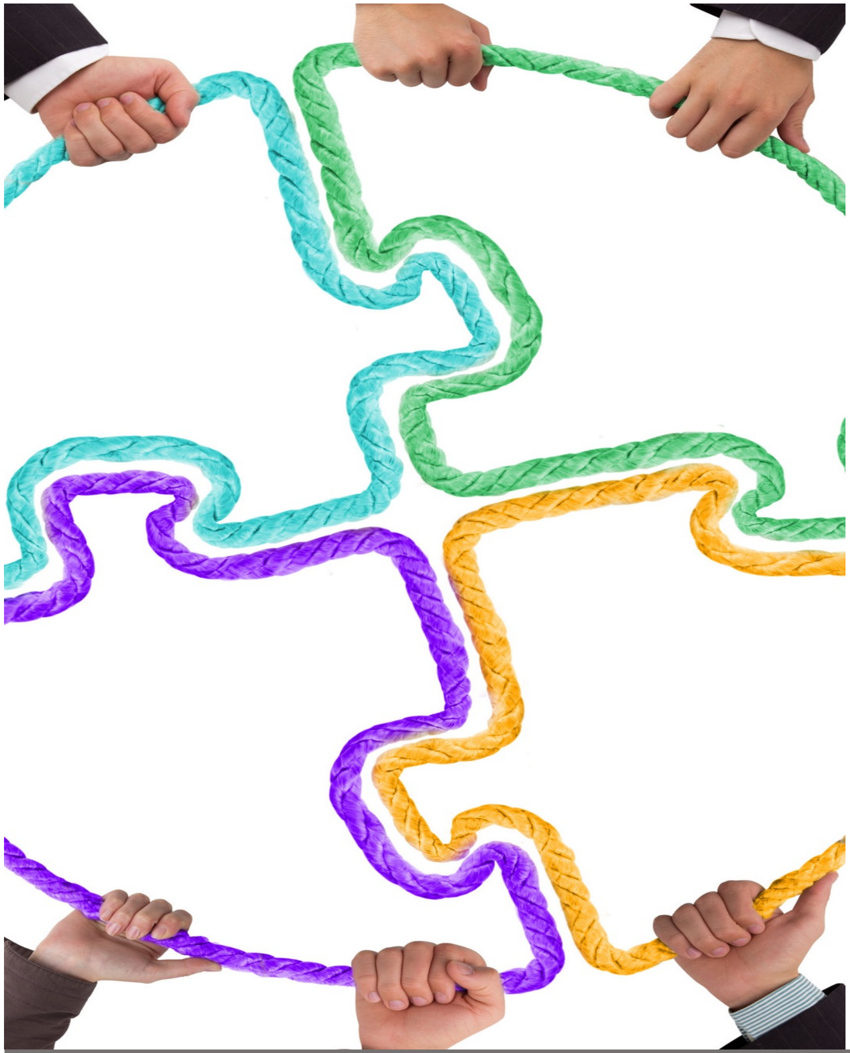
“not just science + AI methods”

- Data federation and hubs
- Data quality and volume
- Knowledge managem
- 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



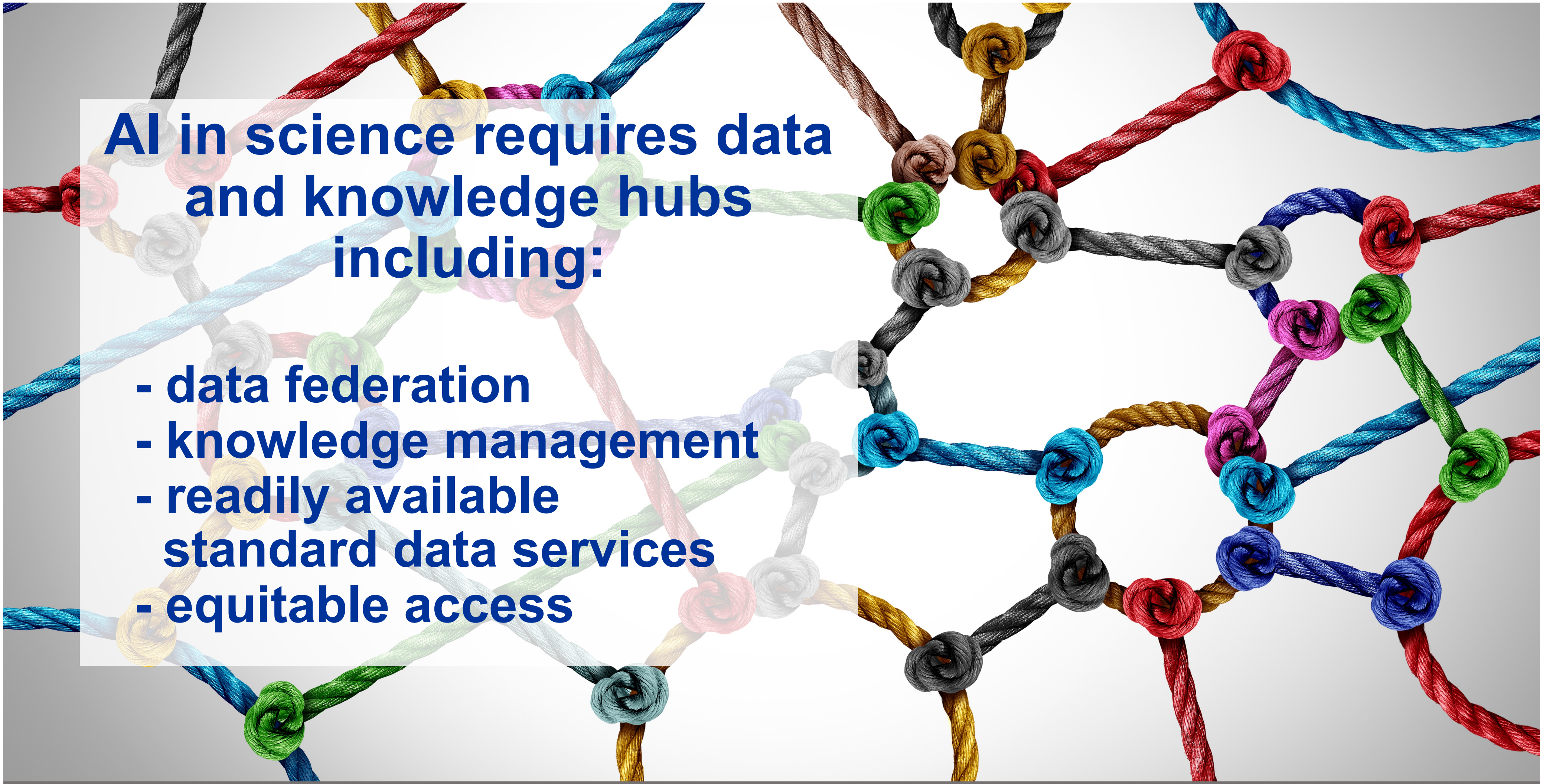
Requires a full team and enabling integrated data platforms

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.

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

Dynamic composability matters.

COMPOSABLE SERVICES

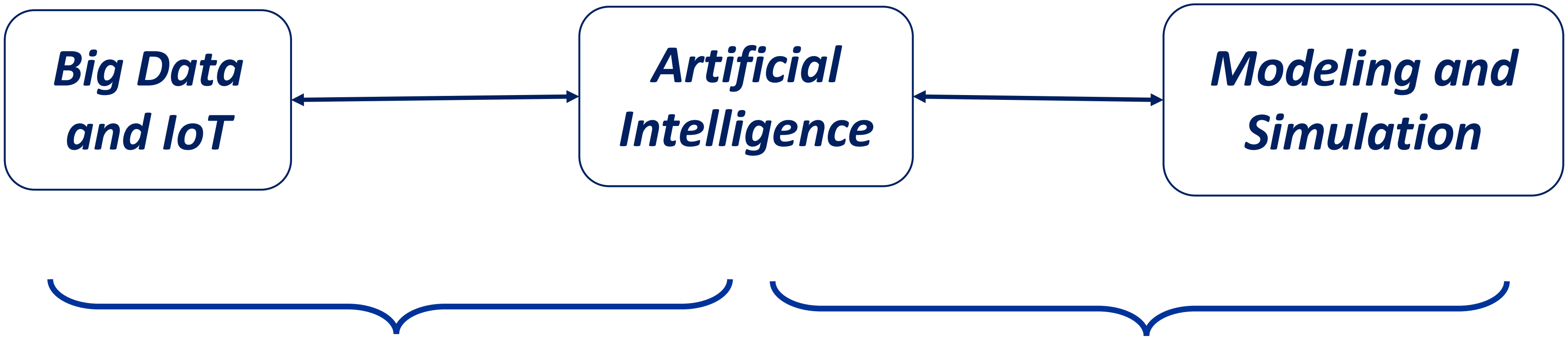
e.g., model and data archives, learning and analytics, simulation, training

RESOURCE MANAGEMENT

e.g., container orchestration, optimization

COMPOSABLE SYSTEMS

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



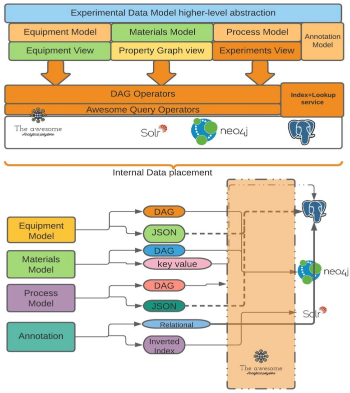
Big Data and IoT

Artificial Intelligence

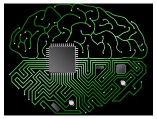
Modeling and Simulation

Capability

Capacity

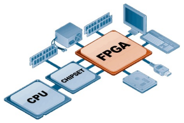


Big Data



xPU → GPU, CPU, TPU, IPU, QPU,…

Edge



FPGA

Cloud, HPC, Storage

Some Composable Systems

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

EXPANSE

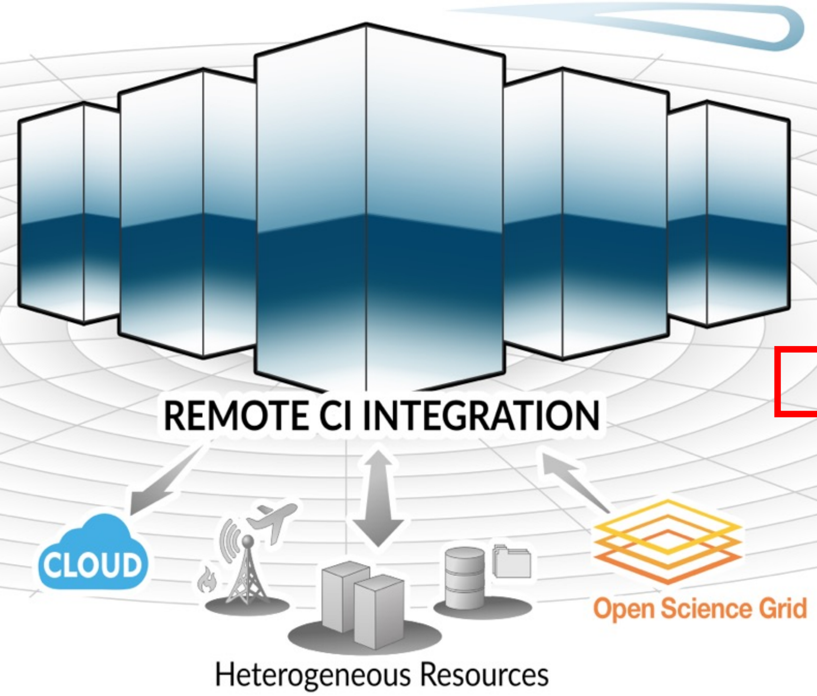
COMPUTING WITHOUT BOUNDARIES
5 PETAFLOP/S HPC and DATA RESOURCE

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

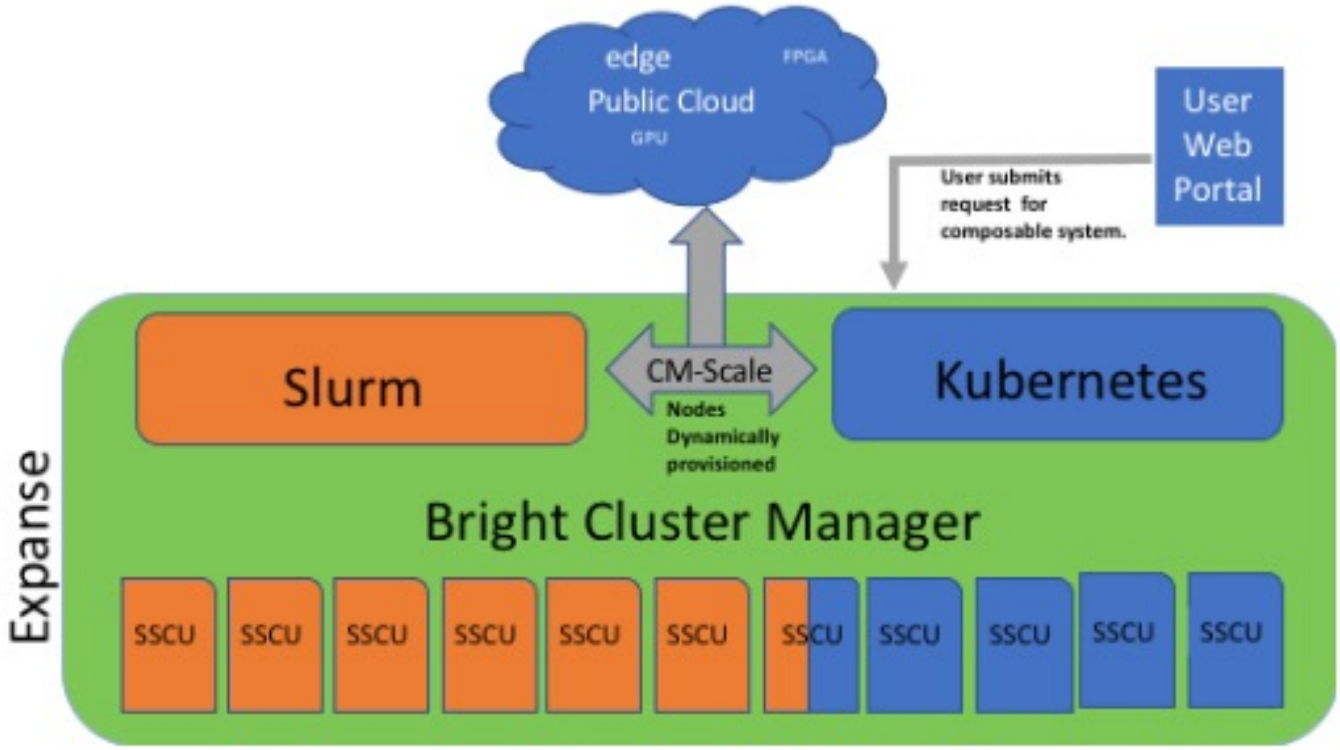


LONG-TAIL SCIENCE

Multi-Messenger Astronomy
Genomics
Earth Science
Social Science

INNOVATIVE OPERATIONS

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

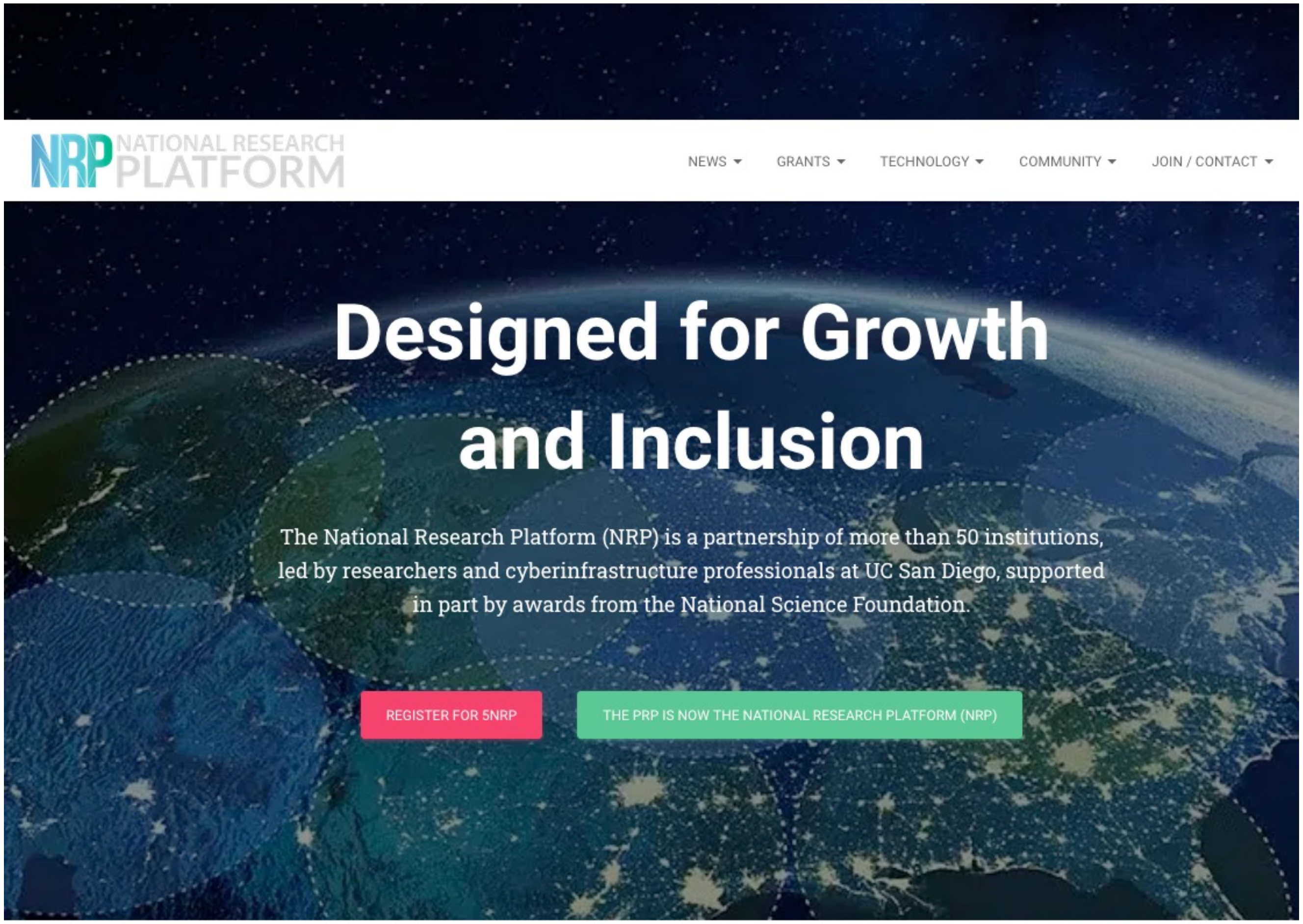


Expanse Composable Systems Framework

National Research Platform

Nautilus

https://nationalresearchplatform.org/


NRP NATIONAL RESEARCH PLATFORM
NEWS ▾
GRANTS ▾
TECHNOLOGY ▾
COMMUNITY ▾
JOIN / CONTACT ▾
REGISTER FOR 5NRP
THE PRP IS NOW THE NATIONAL RESEARCH PLATFORM (NRP)

Designed for Growth and Inclusion

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.

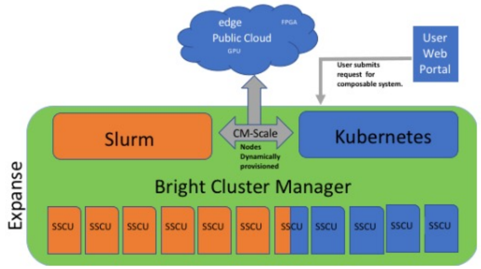
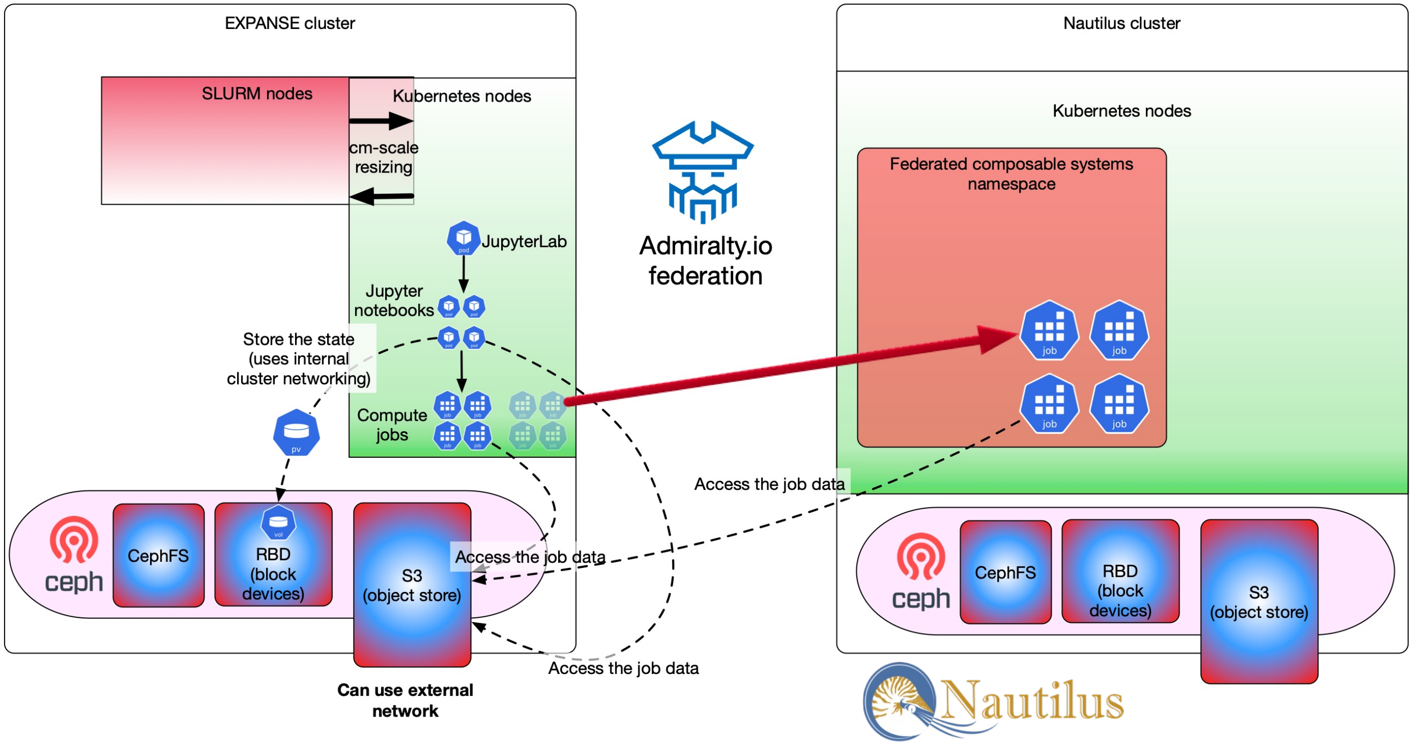

edge
Public Cloud
FPGA
GPU
User Web Portal
User submits request for composable system.
Slurm
CM-Scale
Nodes Dynamically provisioned
Kubernetes
Bright Cluster Manager
Expanse
SSCU SSCU SSCU SSCU SSCU SSCU SSCU SSCU SSCU SSCU SSCU

Figure 5.1 Expanse Composable Systems Framework

First composable cluster is federated!

EXPANSE (Enthalpy) + CHASE-CI (Nautilus)

EXPANSE (Enthalpy)



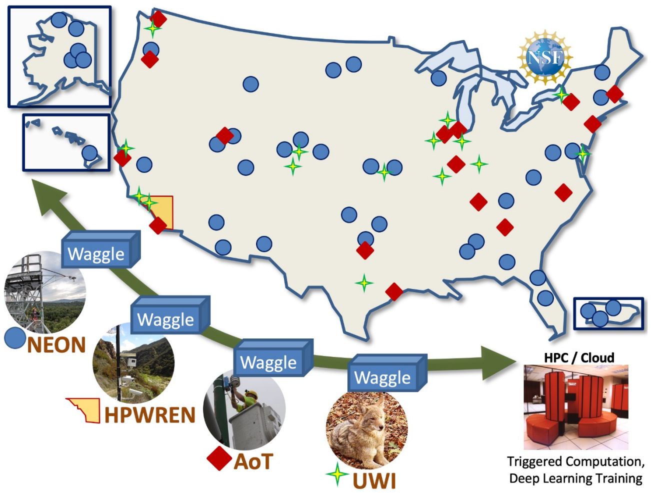
Sensor/Instrument ***HPC/Cloud***

AI@Edge and the Digital Continuum

Slide Source: *Pete Beckman, ANL*

SAGE

Cyberinfrastructure for
AI at the Edge
sagecontinuum.org

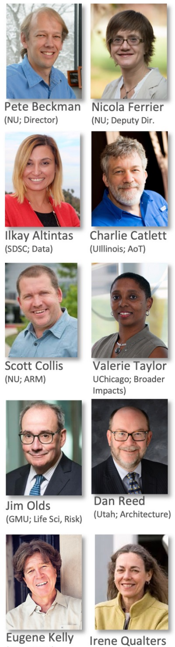


Triggered Computation,
Deep Learning Training

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.

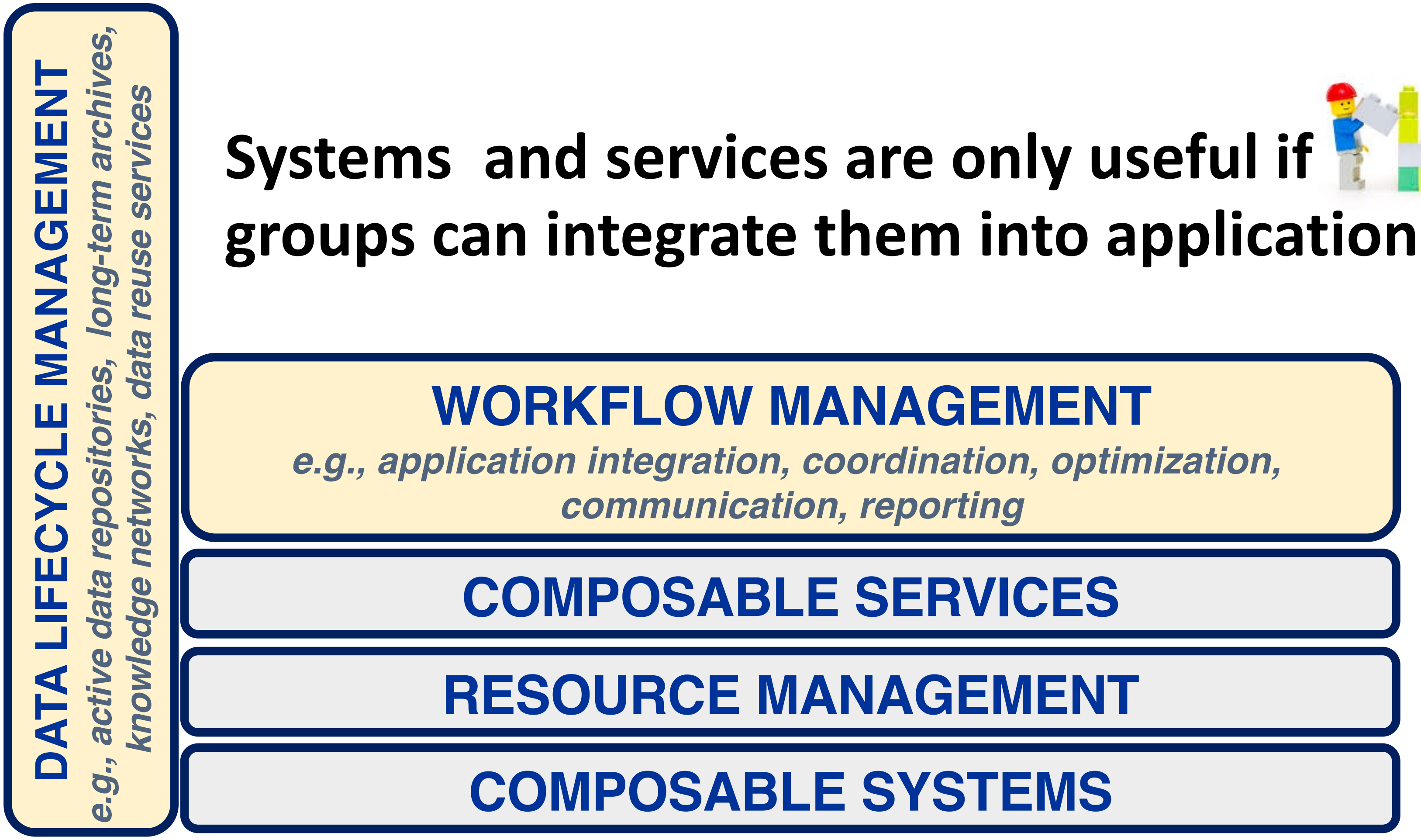
Education & Training

Leadership Team



Pete Beckman (NU; Director)
Nicola Ferrier (NU; Deputy Dir.
Ilkay Altintas (SDSC; Data)
Charlie Catlett (UIllinois; 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)

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



DATA LIFECYCLE MANAGEMENT

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

WORKFLOW MANAGEMENT

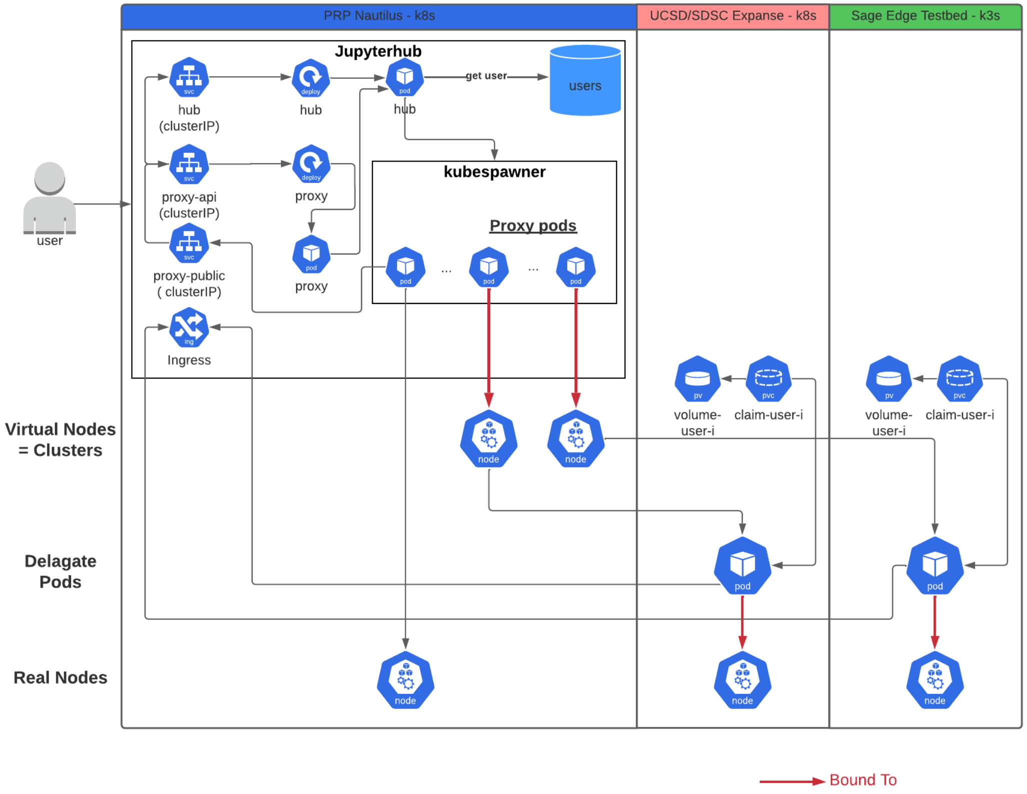
e.g., application integration, coordination, optimization, communication, reporting

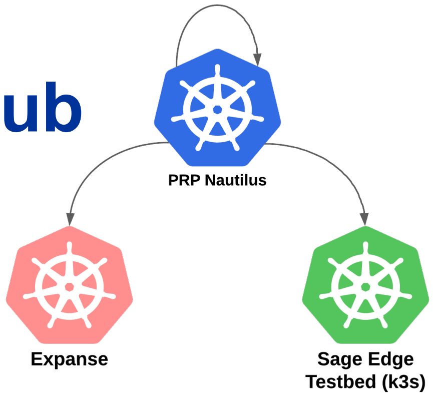
COMPOSABLE SERVICES

RESOURCE MANAGEMENT

COMPOSABLE SYSTEMS

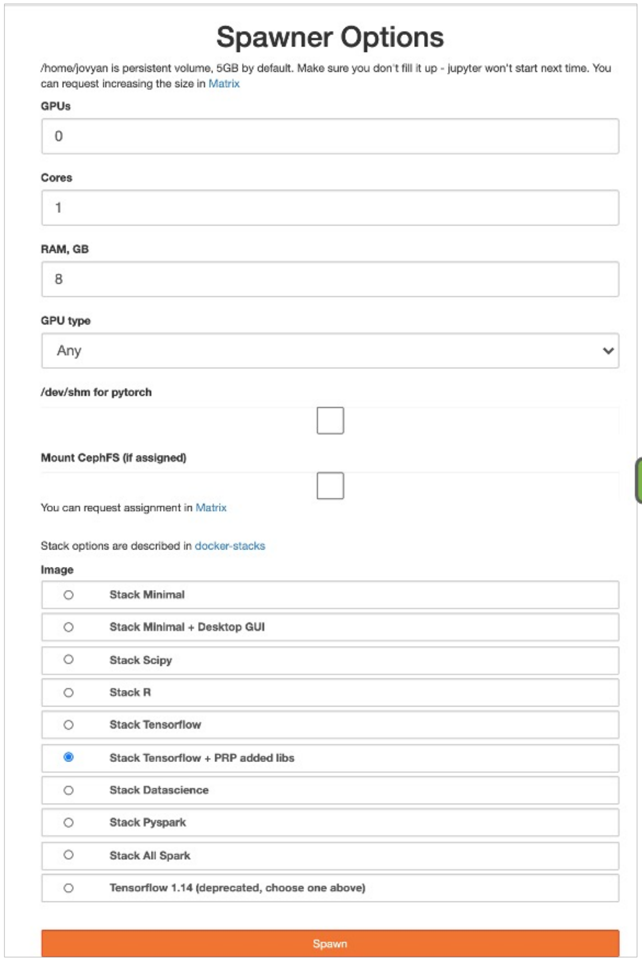
Integration of NSF EXPANSE, NRP and Sage
A Composable System Deployment of JupyterHub





- 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

0

Cores

1

RAM, GB

8

GPU type

Any

/dev/shm for pytorch

□

Mount CephFS (if assigned)

□

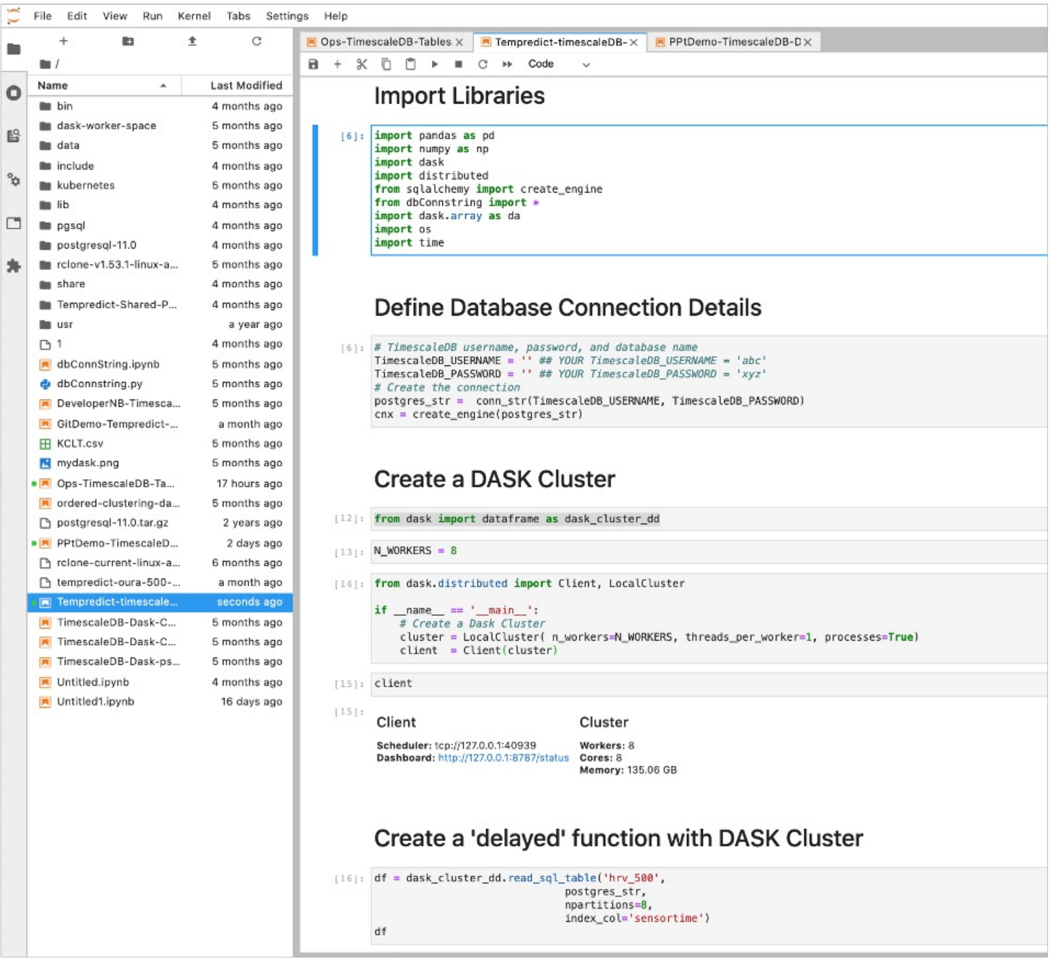
You can request assignment in Matrix

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)

Kubernetes Pod Spawned for Exploration



/

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

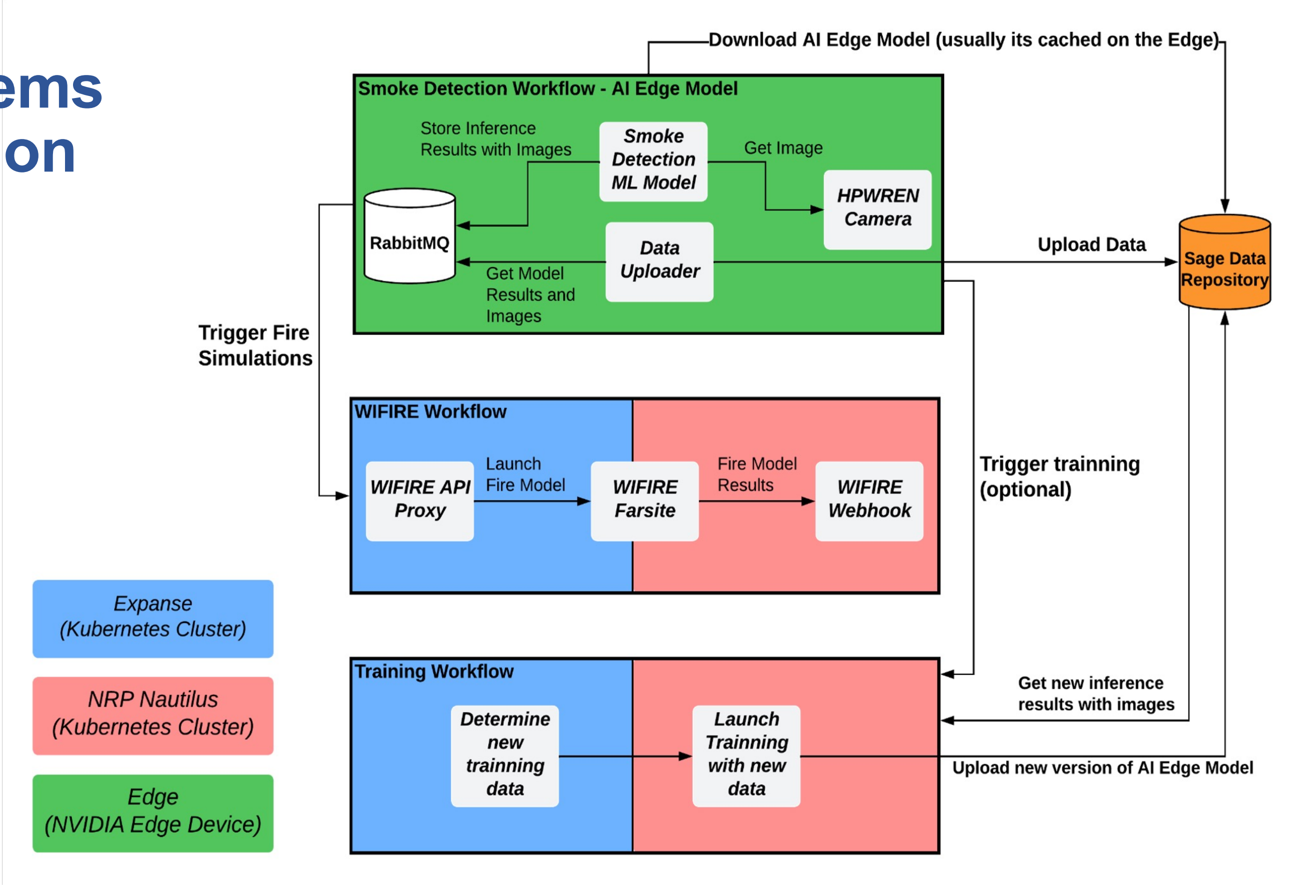
```
[12]: from dask import dataframe as dask_cluster_dd

[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)

[15]: client

[15]:
      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



Download AI Edge Model (usually its cached on the Edge)

Smoke Detection Workflow - AI Edge Model

Store Inference Results with Images

Smoke Detection ML Model

Get Image

HPWREN Camera

RabbitMQ

Data Uploader

Get Model Results and Images

Upload Data

Sage Data Repository

Trigger Fire Simulations

WIFIRE Workflow

WIFIRE API Proxy

Launch Fire Model

WIFIRE Farsite

Fire Model Results

WIFIRE Webhook

Trigger trainning (optional)

Expanse (Kubernetes Cluster)

NRP Nautilus (Kubernetes Cluster)

Edge (NVIDIA Edge Device)

Training Workflow

Determine new trainning data

Launch Trainning with new data

Get new inference results with images

Upload new version of AI Edge Model

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

RESPONSIBILITY

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

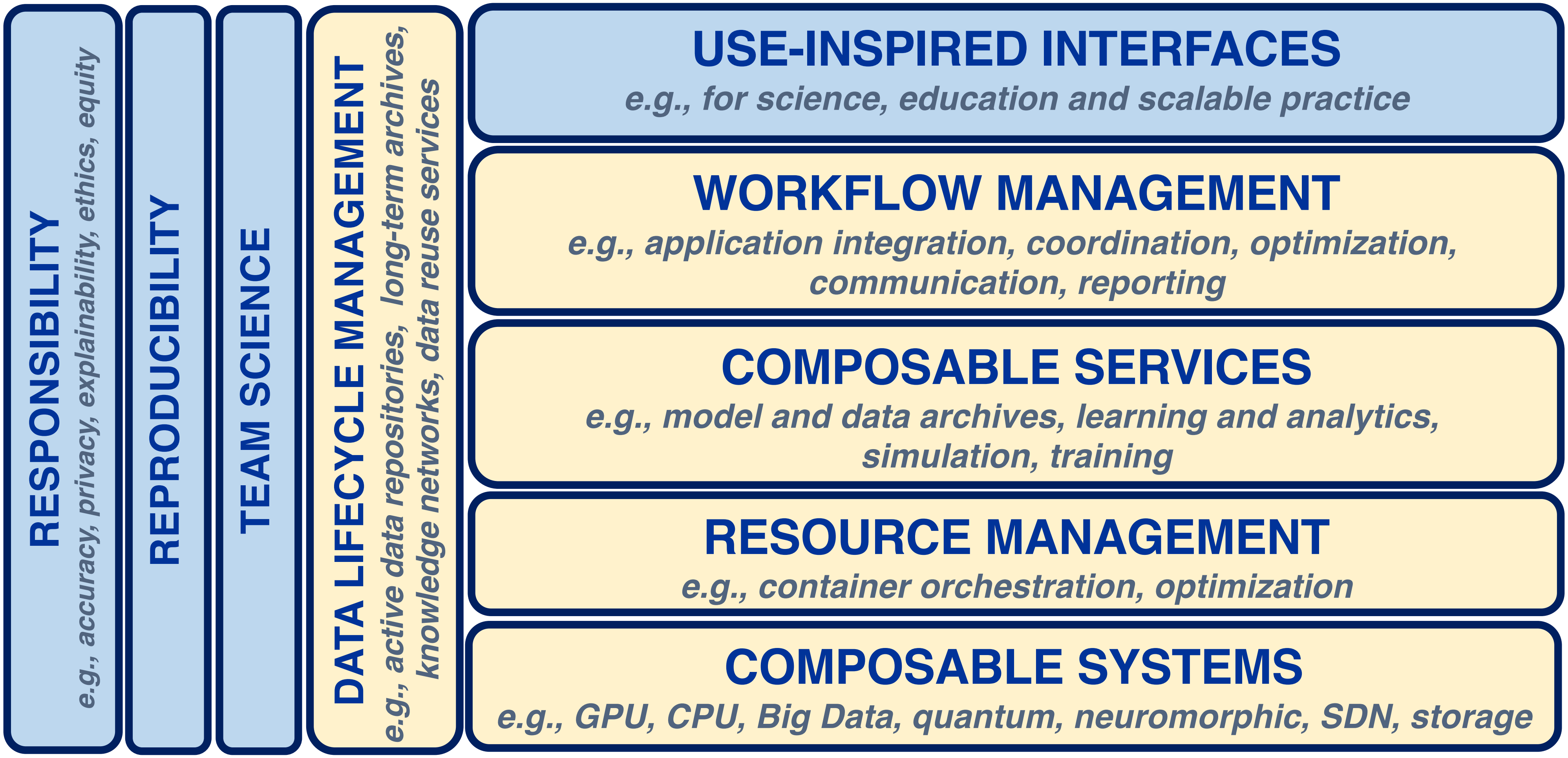
REPRODUCIBILITY

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.



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

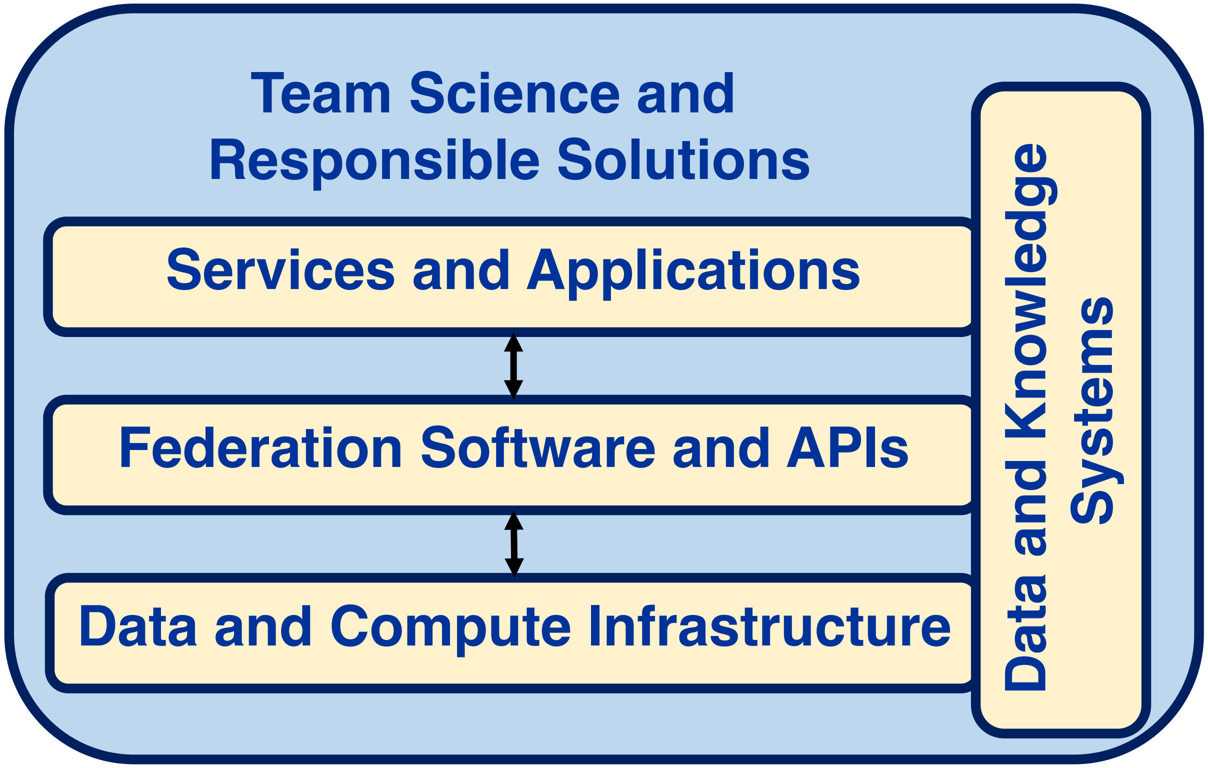
RESOURCE MANAGEMENT

e.g., container orchestration, optimization

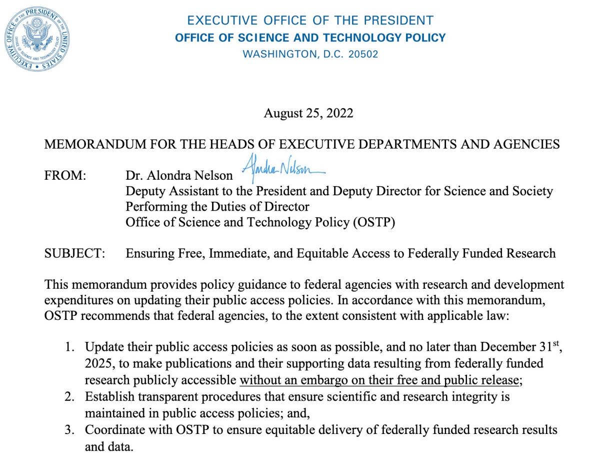
COMPOSABLE SYSTEMS

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

Use-Inspired Composability from Systems to Services


Team Science and Responsible Solutions
Services and Applications
Federation Software and APIs
Data and Compute Infrastructure
Data and Knowledge Systems

- 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



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

August 25, 2022

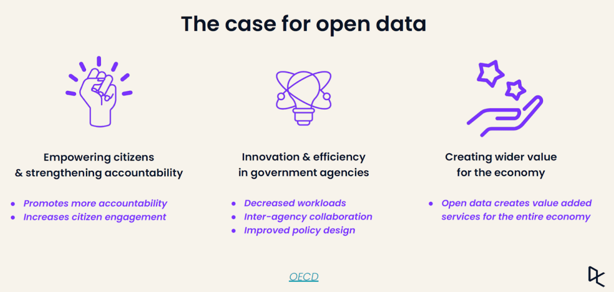
MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson 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*

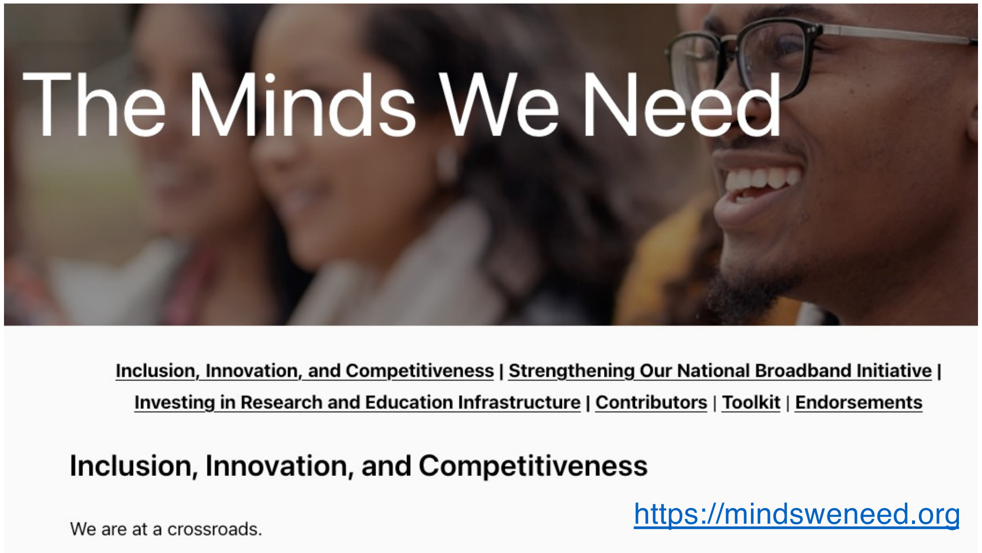
Innovation & efficiency in government agencies

- *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



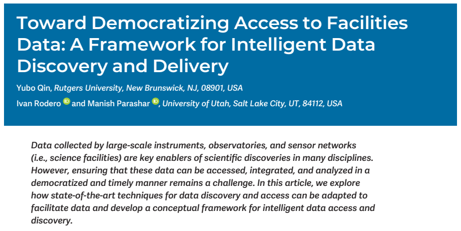
The Minds We Need

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.

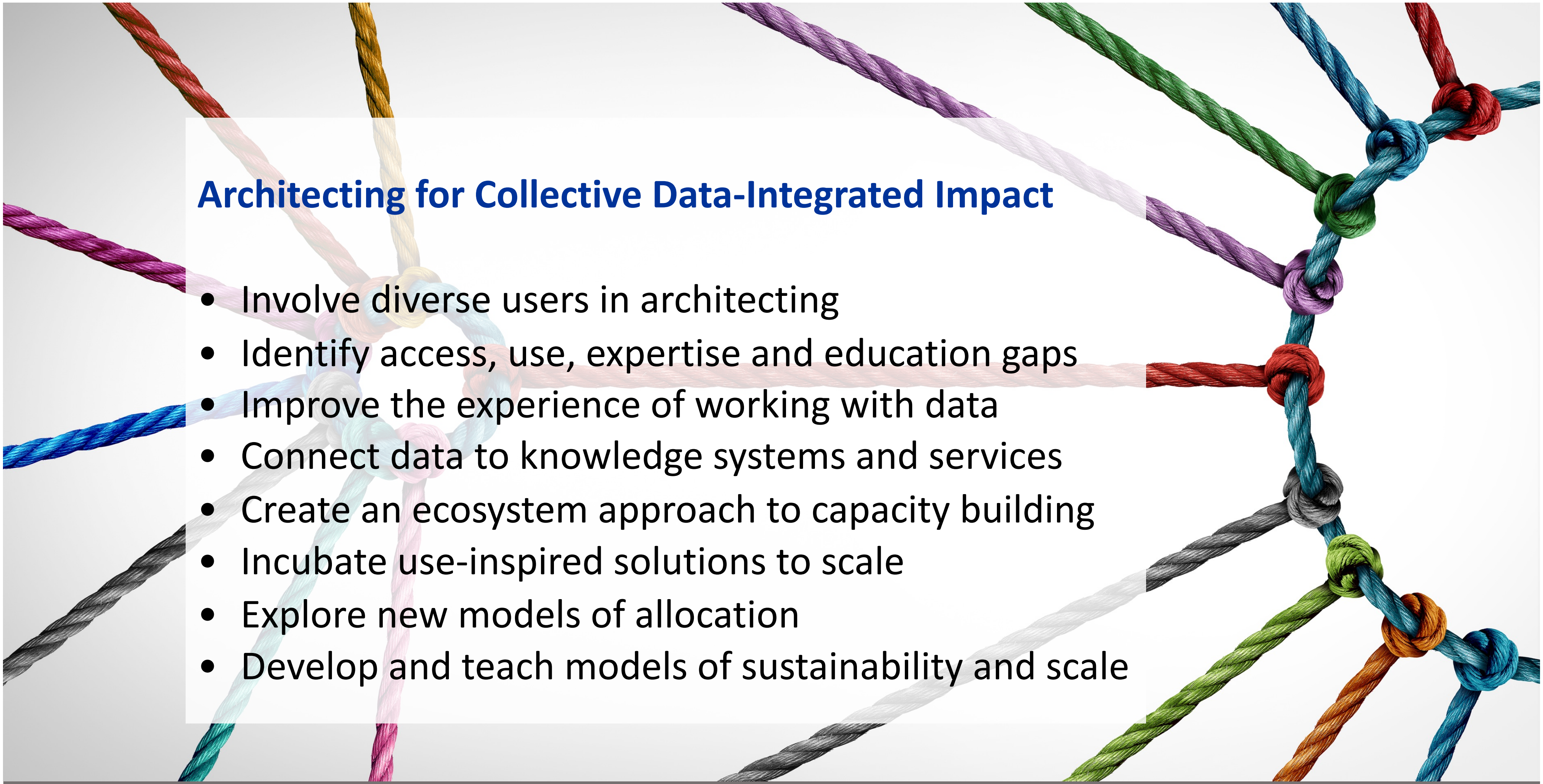
The Missing *Millions*

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

October 3, 2021

NSF OAC 2127459

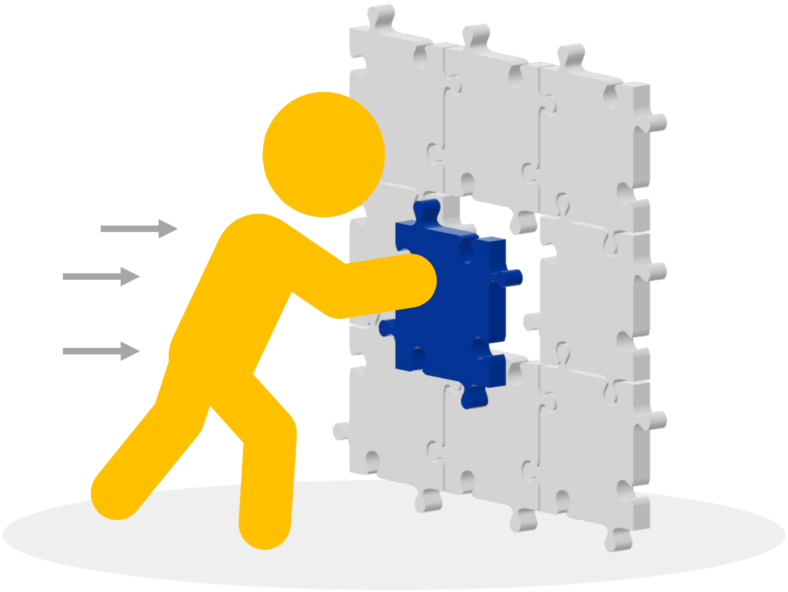
Democratization of CI and Data Access



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

How do we bridge the data gaps?



http://www.nationaldataplatform.org

NATIONAL DATA PLATFORM

Bridging the Data Gaps for AI

UC San Diego

UTAH

University of Colorado Boulder

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

SCI
www.sci.utah.edu

EarthScope Consortium

http://www.nationaldataplatform.org

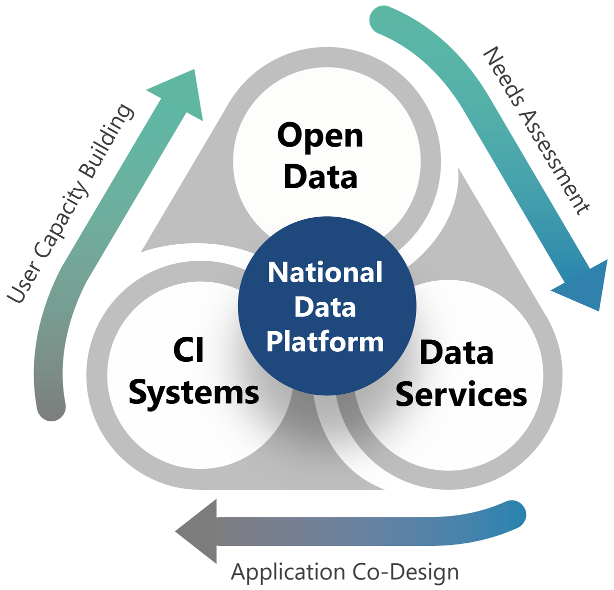
Award abstract: https://www.nsf.gov/awardsearch/showAward?AWD_ID=2333609

What is the National Data Platform?

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

FOCUS AREAS:

- 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



http://www.nationaldataplatform.org

SDSC SAN DIEGO SUPERCOMPUTER CENTER

www.sci.utah.edu

UTAH

University of Colorado Boulder

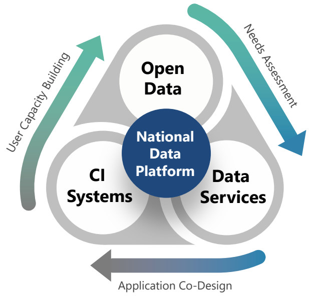
EarthScope Consortium

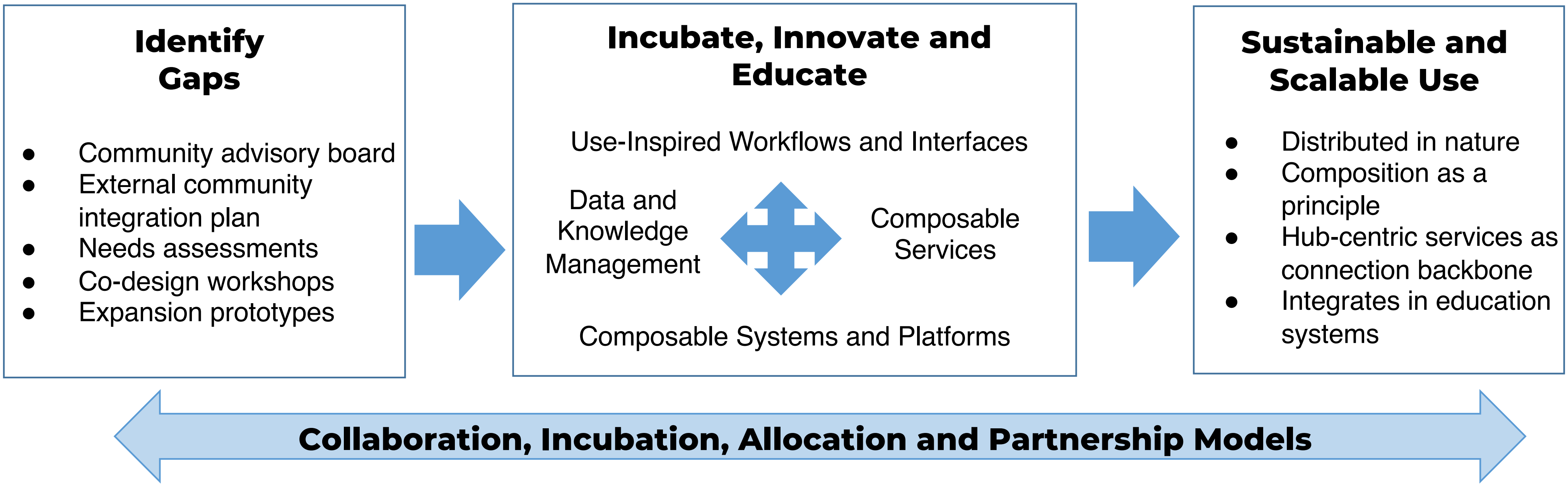
http://www.nationaldataplatform.org

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

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

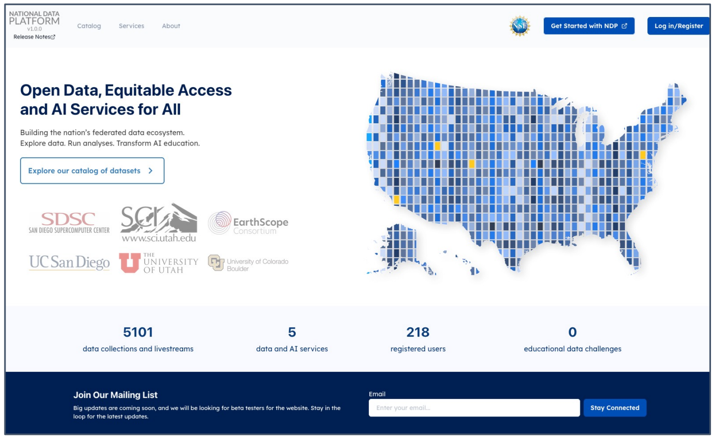
- Distributed in nature
- Composition as a principle
- Hub-centric services as connection backbone
- Integrates in education systems

Collaboration, Incubation, Allocation and Partnership Models

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE



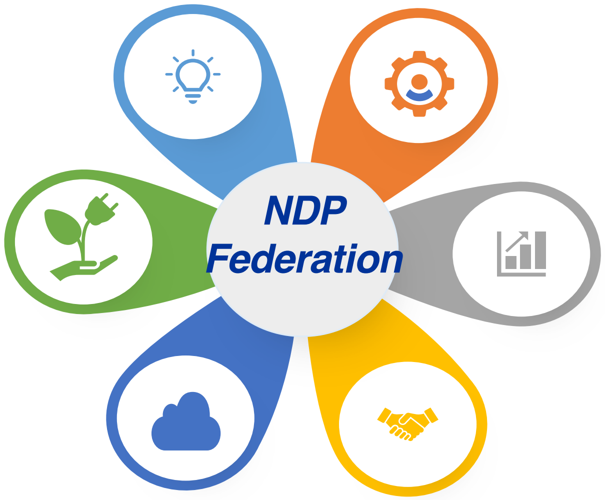
Open Data, Equitable Access and AI Services for All

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

5101	5	218	0
data collections and livestreams	data and AI services	registered users	educational data challenges

Join Our Mailing List

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



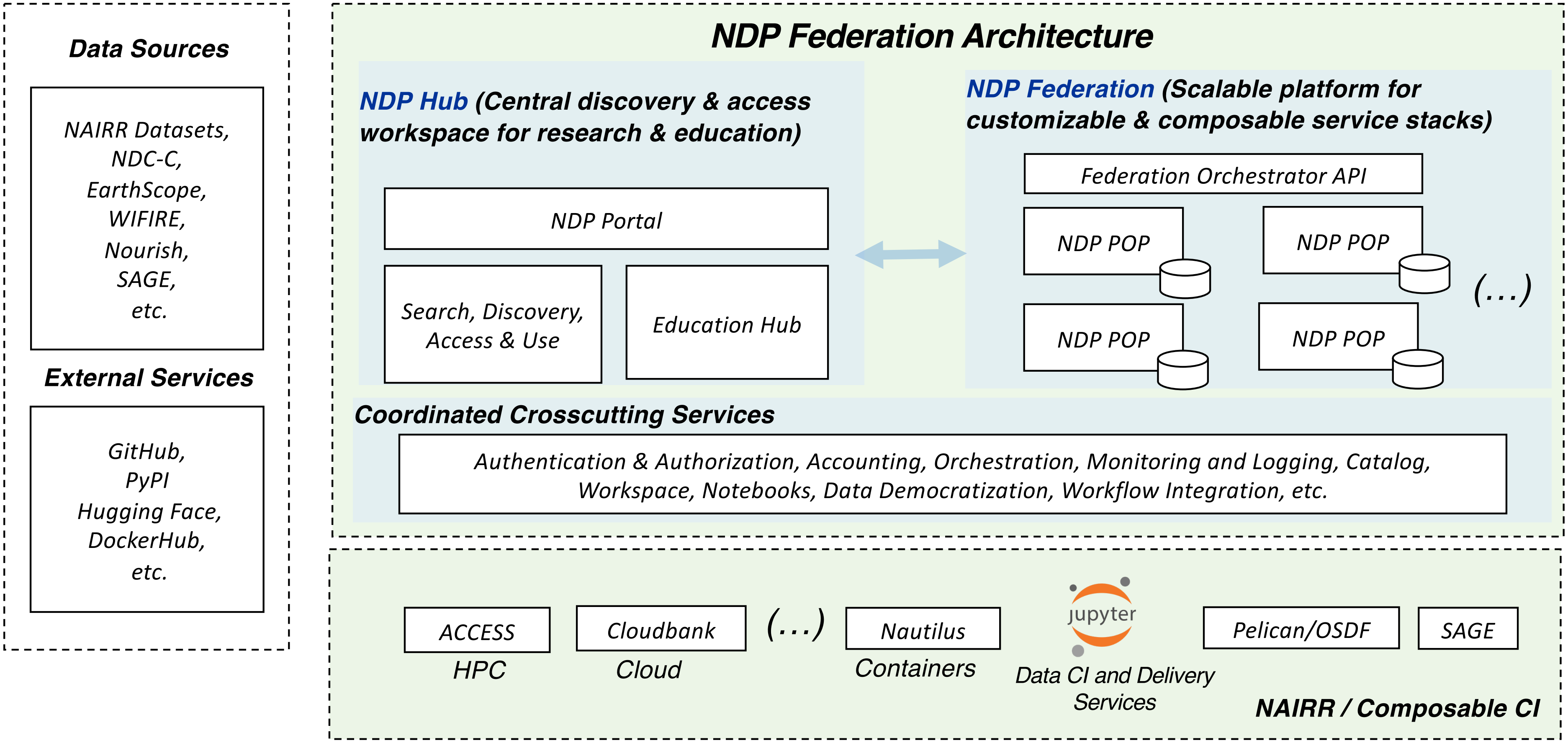
NATIONAL DATA PLATFORM

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***

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

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)

Federation Orchestrator API

NDP POP

NDP POP

NDP POP

NDP POP

(…)

Coordinated Crosscutting Services

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

ACCESS
HPC

Cloudbank
Cloud

(…)

Nautilus
Containers

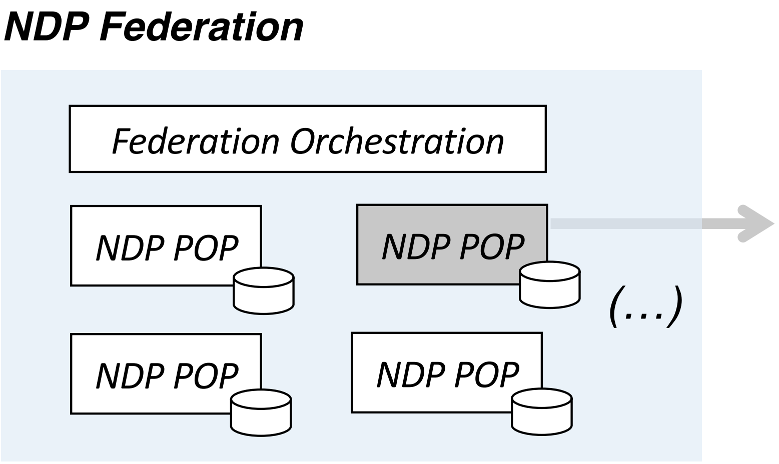
jupyter
Data CI and Delivery Services

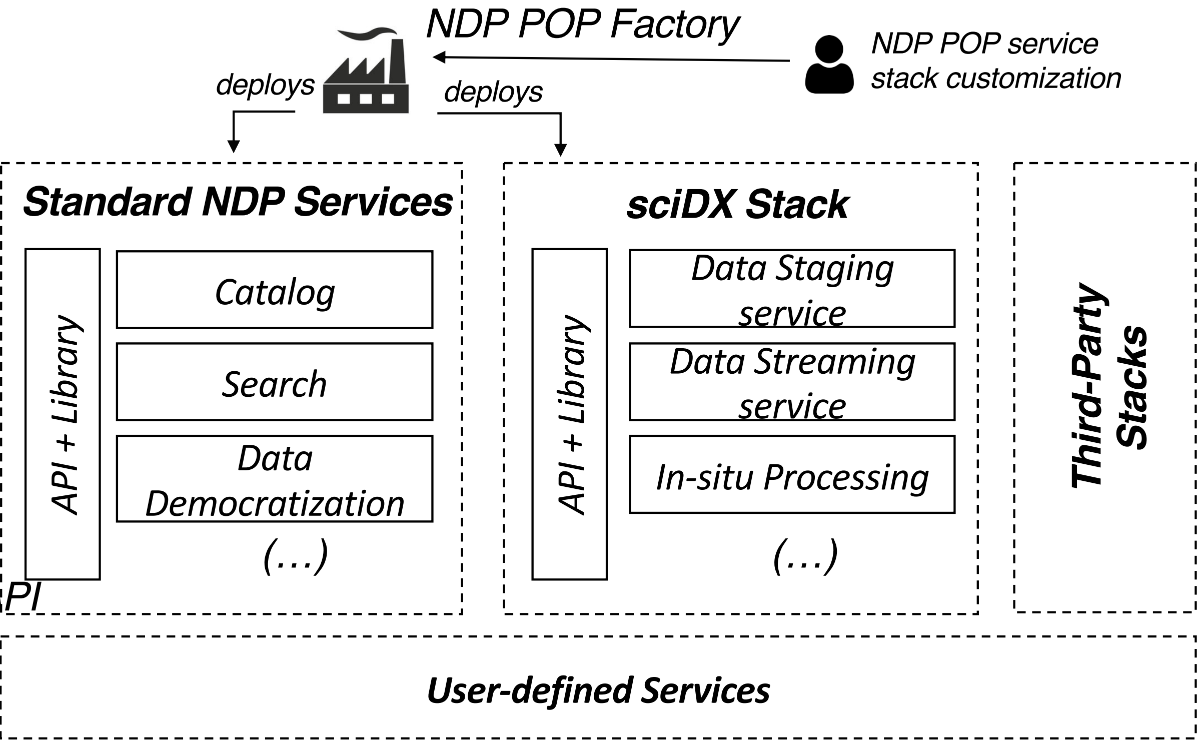
Pelican/OSDF

SAGE

NAIRR / Composable CI

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





Workflow composition currently via API and/or Python client library

pypi

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

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

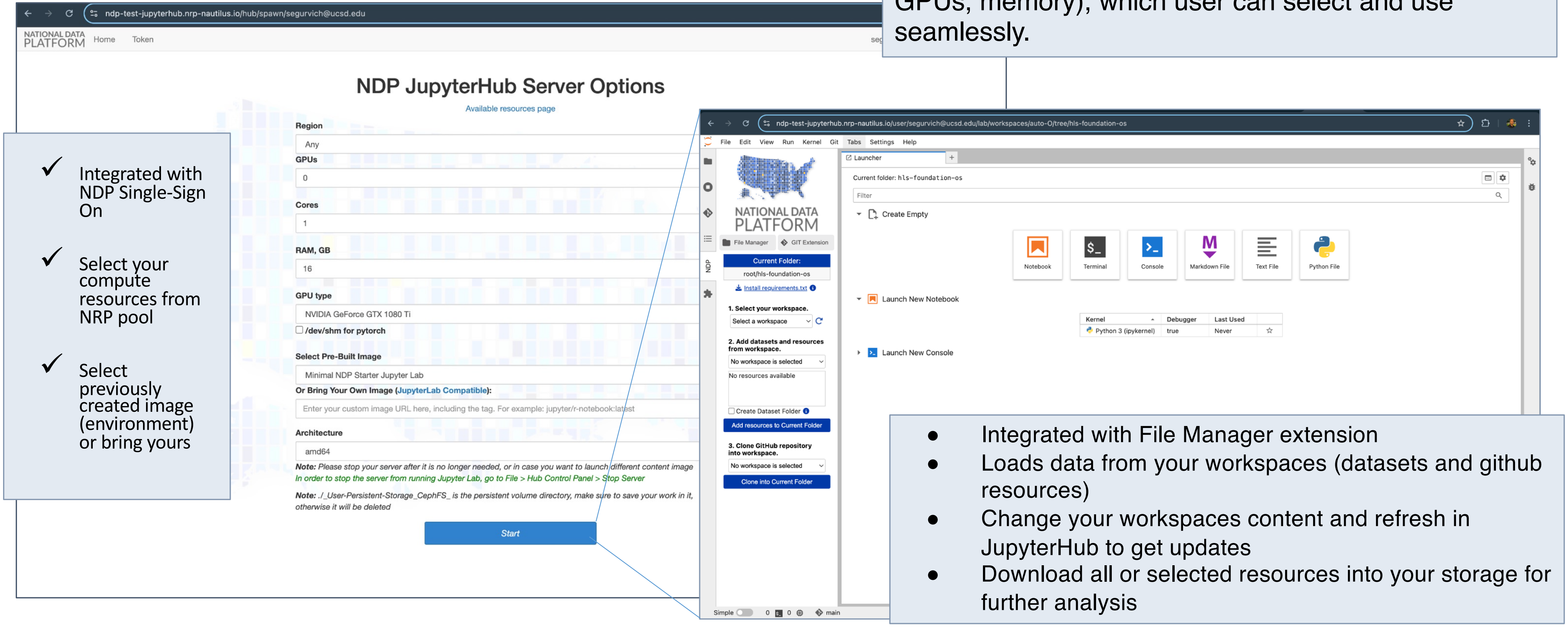
Deployment models: docker kubernetes Cloud HPC

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

NDP JupyterHub (Sandbox)

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

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



ndp-test-jupyterhub.nrp-nautilus.io/hub/spawn/segurvich@ucsd.edu

NATIONAL DATA PLATFORM Home Token

NDP JupyterHub Server Options

Available resources page

Region: Any

GPUs: 0

Cores: 1

RAM, GB: 16

GPU type: NVIDIA GeForce GTX 1080 Ti

□ /dev/shm for pytorch

Select Pre-Built Image: Minimal NDP Starter Jupyter Lab

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

Architecture: amd64

Note: *Please stop your server after it is no longer needed, or in case you want to launch different content image*
In order to stop the server from running Jupyter Lab, go to File > Hub Control Panel > Stop Server

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

Start

ndp-test-jupyterhub.nrp-nautilus.io/user/segurvich@ucsd.edu/lab/workspaces/auto-O/tree/hls-foundation-os

File Edit View Run Kernel Git Tabs Settings Help

NATIONAL DATA PLATFORM

File Manager | GIT Extension

Current Folder: root/hls-foundation-os

Install requirements.txt

1. Select your workspace.
Select a workspace

2. Add datasets and resources from workspace.
No workspace is selected
No resources available
□ Create Dataset Folder
Add resources to Current Folder

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

Launcher

Current folder: hls-foundation-os

Filter

Create Empty: Notebook, Terminal, Console, Markdown File, Text File, Python File

Launch New Notebook

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

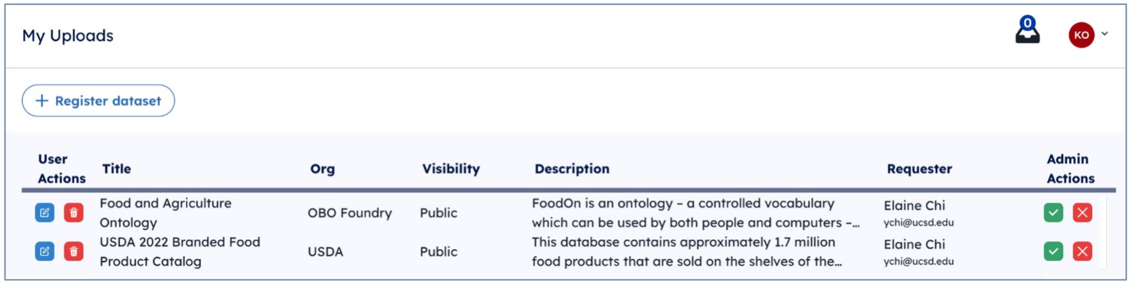
Launch New Console

Simple 0 0 main

- 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

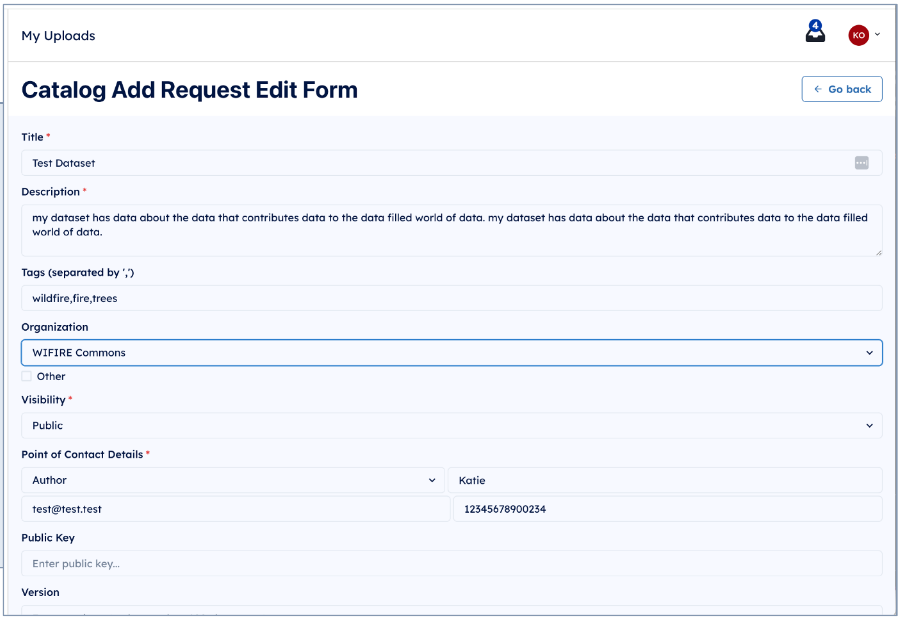
Goal: Users can add dataset references to either NDP centralized catalog or POP-specific catalog



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	✓ ✕

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



Catalog Add Request Edit Form

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

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

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

Slide Source:
Manish Parashar and Ivan Rodero

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

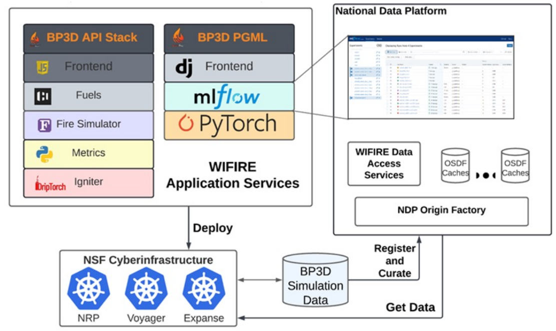
Data Streaming Service

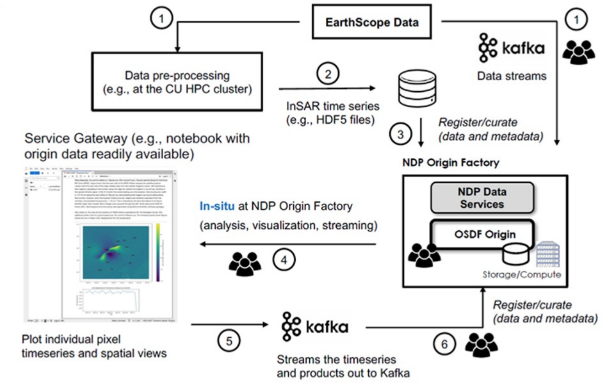
- Streams registration, curation/archival for discovery and access
- User-defined operations on streaming data (semantically specialized abstractions)
- Combine streaming data with archived/playback data
- Mechanism for online data product generation (i.e., new data streams

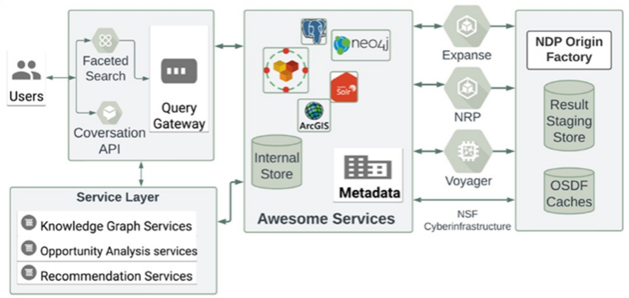
In-situ AI workflow execution runtime (on staged and streaming data)

Case Studies for Generalizable Workflows

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




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



INTRODUCING THE

WILDFIRE TECHNOLOGY COMMONS

We believe that avoiding devastating wildfires requires urgent, innovative, and collaborative solutions. The Wildfire Technology Commons is a bold new initiative designed to accelerate technological innovations for wildfire management and mitigation. We are building a community platform around open data, cutting-edge science, AI, and shared knowledge.

https://www.wildfirecommons.org

JOIN THE NETWORK

CONTRIBUTE DATA & MODELS

BECOME A PATHFINDER

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

NDP Data Challenges for students and researchers

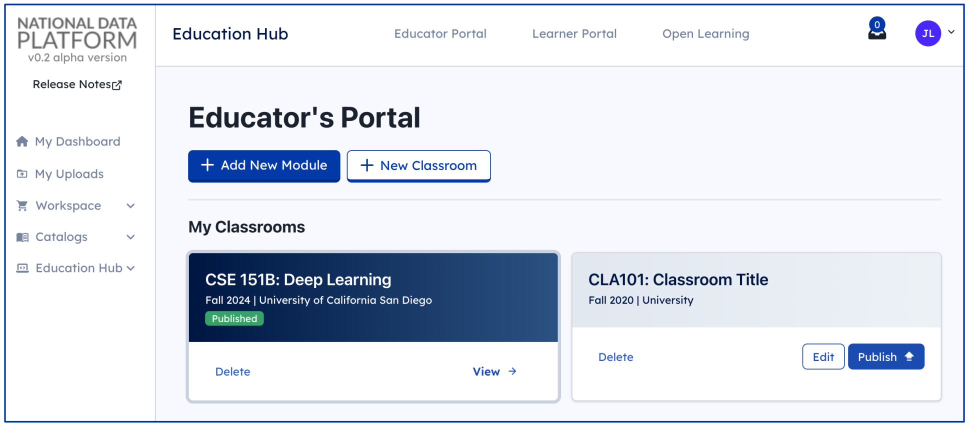
NDP Education Hub to provide participants access to the NDP data ecosystem

NATIONAL DATA PLATFORM

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

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

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.



Educator's Portal

My Classrooms

CSE 151B: Deep Learning
Fall 2024 | University of California San Diego
Published

CLA101: Classroom Title
Fall 2020 | University

Education and capacity building through data challenges

DATA CHALLENGE

FIRE-READY FORESTS

Next Generation Science for Wildfire Resilience

Hosted by the Prowess Center on the National Data Platform in partnership with the Wildfire Commons
More information at https://prowesscenter.org/

NATIONAL DATA PLATFORM

WILDFIRE
SCIENCE & TECHNOLOGY
COMMONS

PROWESS

UC San Diego

Los Alamos
NATIONAL LABORATORY

To sum up…

Emerging new applications require integrated AI in dynamically composed workflows, but there are significant data gaps to be addressed.

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

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?

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.