National Data Platform (NDP) as an AI Research Resource for All

Presentation and LLM as a Service Tutorial at the 5NRP Meeting
March 19, 2024 - San Diego

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**

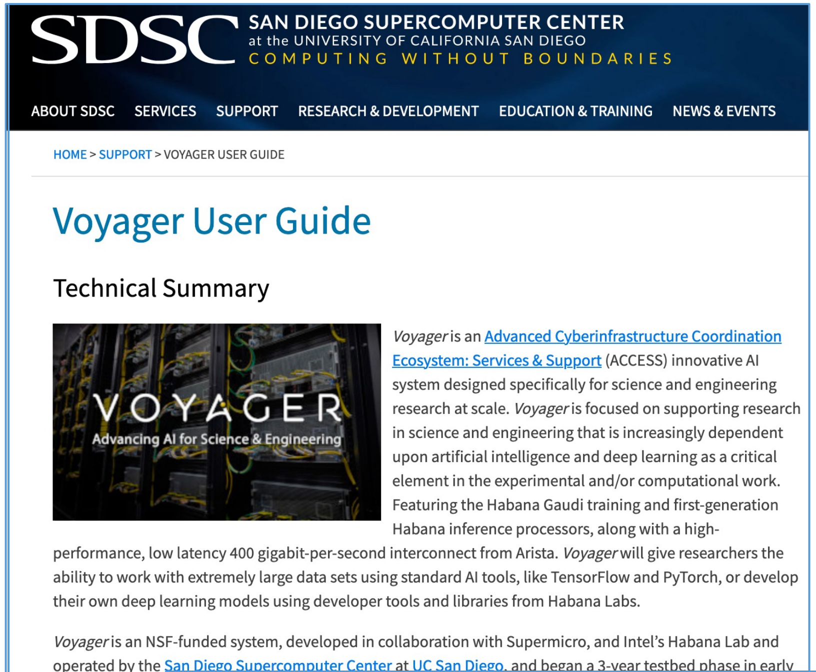
Joint Faculty Appointee, **Los Alamos National Laboratory**

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Some Examples of AI at SDSC

- ○ Exploring new architectures in support of AI in research and engineering
- ○ Teaching best practices for machine learning and data science applications
- Building methods for AI-integrated societal impat
- NAIRR Pilot Activities



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

Voyager User Guide

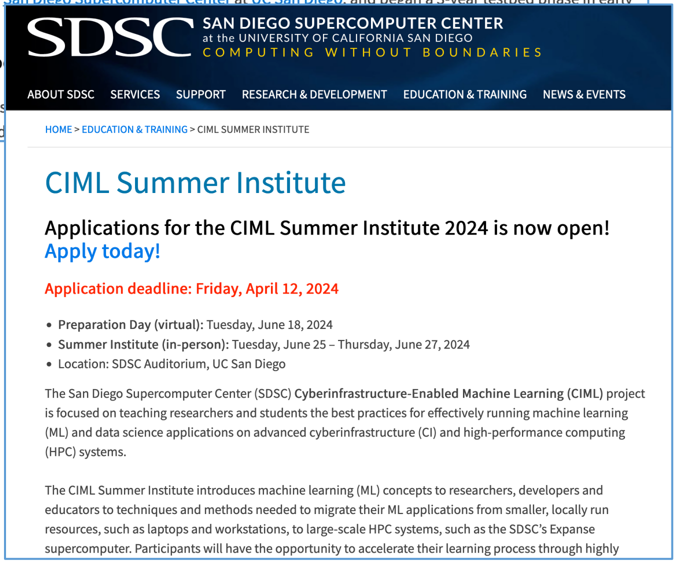
Technical Summary

Voyager is an Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) innovative AI system designed specifically for science and engineering research at scale. *Voyager* is focused on supporting research in science and engineering that is increasingly dependent upon artificial intelligence and deep learning as a critical element in the experimental and/or computational work. Featuring the Habana Gaudi training and first-generation Habana inference processors, along with a high-performance, low latency 400 gigabit-per-second interconnect from Arista. *Voyager* will give researchers the ability to work with extremely large data sets using standard AI tools, like TensorFlow and PyTorch, or develop their own deep learning models using developer tools and libraries from Habana Labs.

Voyager is an NSF-funded system, developed in collaboration with Supermicro, and Intel's Habana Lab and operated by the San Diego Supercomputer Center at UC San Diego, and began a 3-year testbed phase in early 2022.

Resource Allo

- Current Status
- 3-year testbed



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

CIML Summer Institute

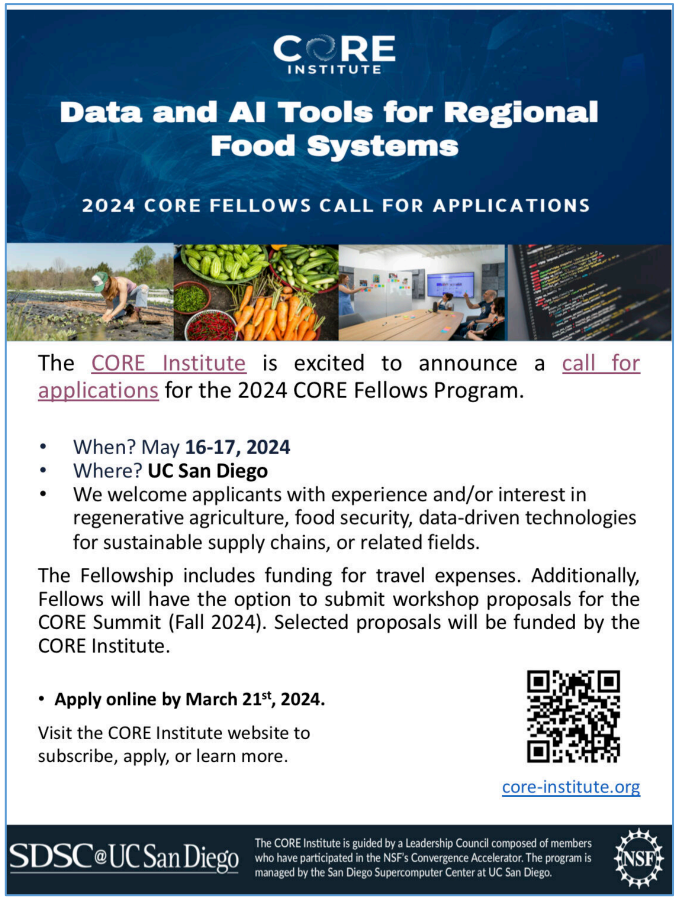
Applications for the CIML Summer Institute 2024 is now open! Apply today!

Application deadline: Friday, April 12, 2024

- **Preparation Day (virtual):** Tuesday, June 18, 2024
- **Summer Institute (in-person):** Tuesday, June 25 – Thursday, June 27, 2024
- Location: SDSC Auditorium, UC San Diego

The San Diego Supercomputer Center (SDSC) **Cyberinfrastructure-Enabled Machine Learning (CIML)** project is focused on teaching researchers and students the best practices for effectively running machine learning (ML) and data science applications on advanced cyberinfrastructure (CI) and high-performance computing (HPC) systems.

The CIML Summer Institute introduces machine learning (ML) concepts to researchers, developers and educators to techniques and methods needed to migrate their ML applications from smaller, locally run resources, such as laptops and workstations, to large-scale HPC systems, such as the SDSC's Expanse supercomputer. Participants will have the opportunity to accelerate their learning process through highly



CORE INSTITUTE

Data and AI Tools for Regional Food Systems

2024 CORE FELLOWS CALL FOR APPLICATIONS

The CORE Institute is excited to announce a call for applications for the 2024 CORE Fellows Program.

- When? May **16-17, 2024**
- Where? **UC San Diego**
- We welcome applicants with experience and/or interest in regenerative agriculture, food security, data-driven technologies for sustainable supply chains, or related fields.

The Fellowship includes funding for travel expenses. Additionally, Fellows will have the option to submit workshop proposals for the CORE Summit (Fall 2024). Selected proposals will be funded by the CORE Institute.

- **Apply online by March 21st, 2024.**

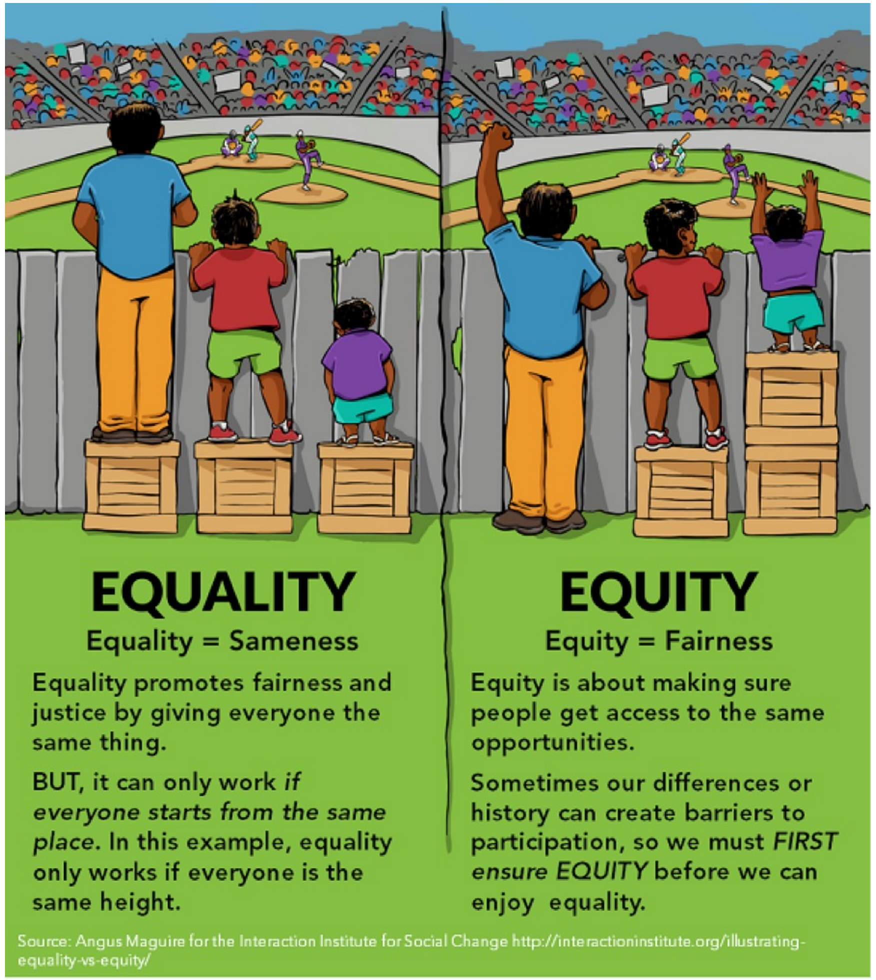
Visit the CORE Institute website to subscribe, apply, or learn more.

core-institute.org

SDSC@UC San Diego

The CORE Institute is guided by a Leadership Council composed of members who have participated in the NSF's Convergence Accelerator. The program is managed by the San Diego Supercomputer Center at UC San Diego.

Equity of AI in Research



EQUALITY

Equality = Sameness

Equality promotes fairness and justice by giving everyone the same thing.

BUT, it can only work *if everyone starts from the same place*. In this example, equality only works if everyone is the same height.

EQUITY

Equity = Fairness

Equity is about making sure people get access to the same opportunities.

Sometimes our differences or history can create barriers to participation, so we must *FIRST ensure EQUITY* before we can enjoy equality.

Source: Angus Maguire for the Interaction Institute for Social Change http://interactioninstitute.org/illustrating-equality-vs-equity/

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

2023

Artificial Intelligence in Science

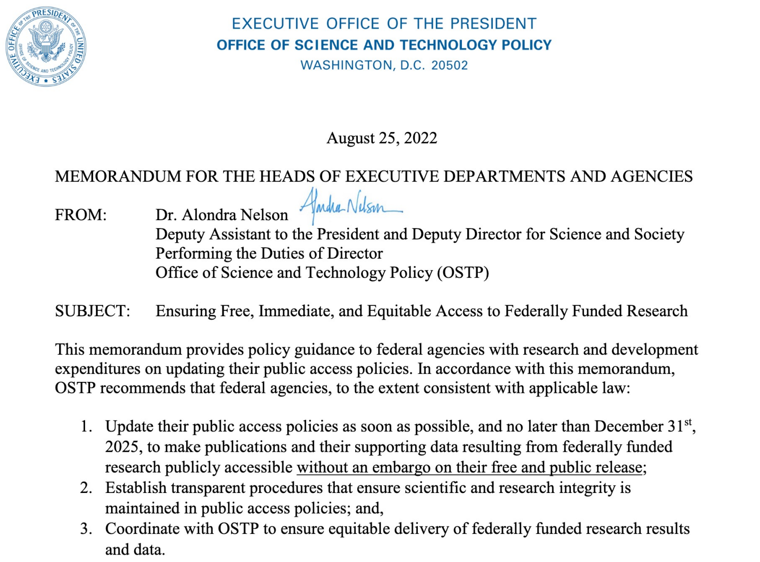
CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH

OECD

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

An Implementation Plan for a National Artificial Intelligence Research Resource

January 2023



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

August 25, 2022

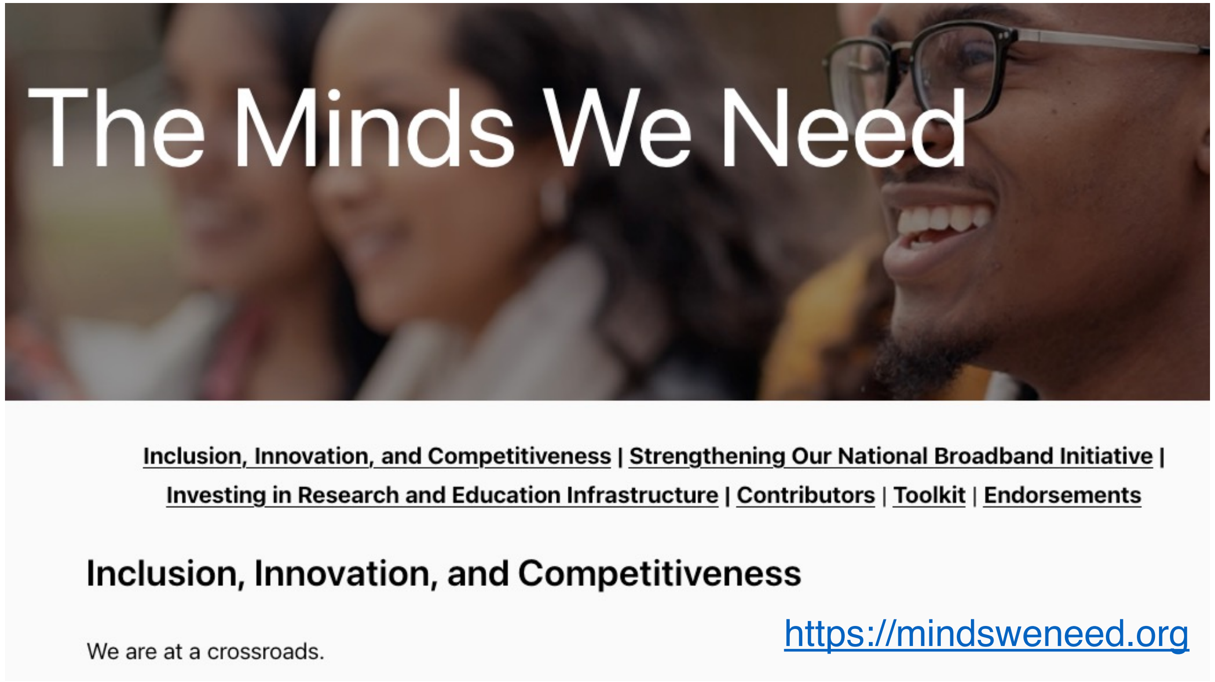
MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

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.



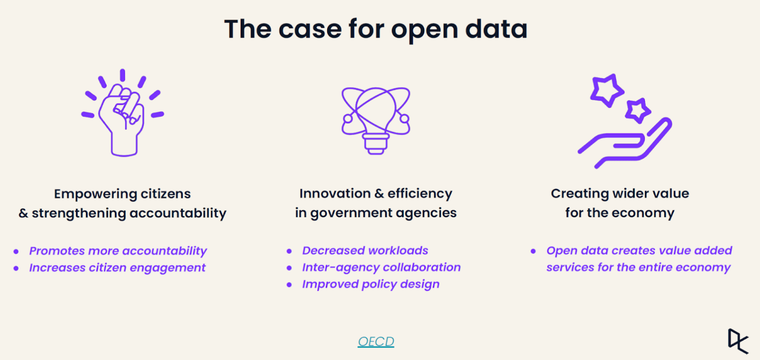
The Minds We Need

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



The case for open data

Empowering citizens & strengthening accountability

- *Promotes more accountability*
- *Increases citizen engagement*

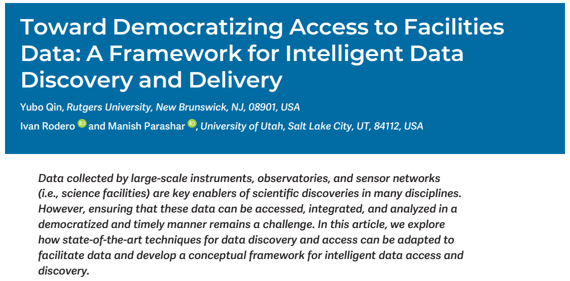
Innovation & efficiency in government agencies

- *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



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

Even access is solved

Open Questions for Equitable Open Research

What are the foundational data abstractions, catalogs, multipurpose services and expandable workflows for data-driven and AI-integrated application patterns?

How can everyone effectively access and utilize these abstractions and services?

FOUNDATIONAL ABSTRACTIONS, CATALOGS, AND SERVICES

How can services and workflows be developed and deployed on top of production-ready CI?

How can equity be ensured for all to access & use CI from storage to the edge-to-HPC computing continuum?

EQUITABLE OPEN CI USE

What are the governance and open science, open data and open CI requirements and challenges?

What are the required guardrails for protecting privacy, civil rights and civil liberties that will **ensure a more equitable use of such data systems and services for everything from education to new AI training and application development**?

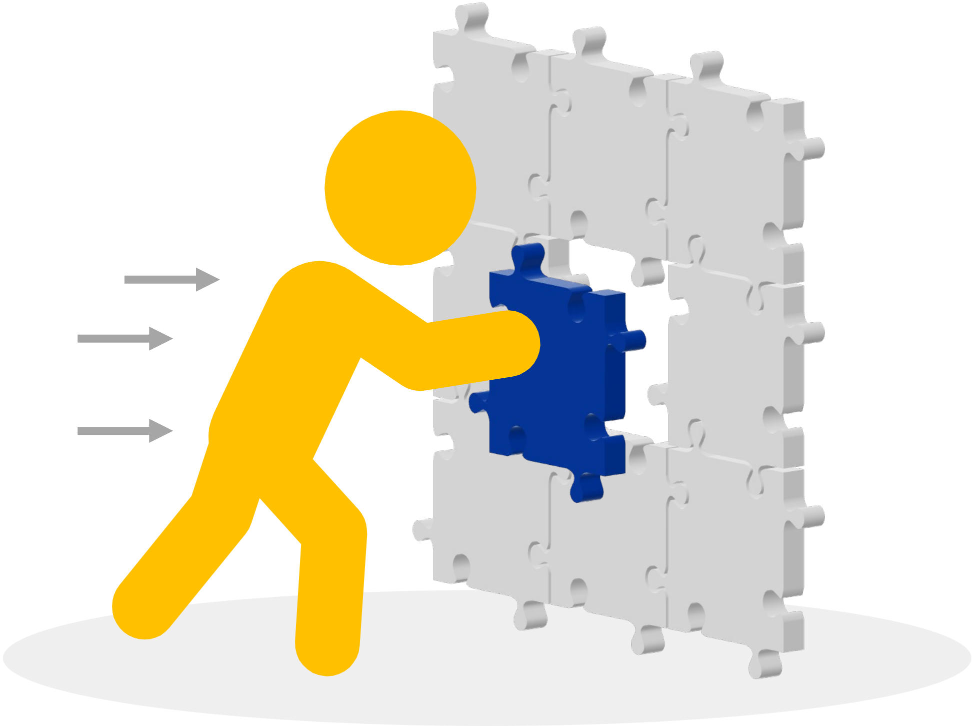
NEEDS, REQUIREMENTS AND CHALLENGES

Architecting for Equity of Research Workflows for All

- Involve diverse users in architecting around access, use, expertise and education gaps
- Improve the experience of working with data
  - e.g., serve data and knowledge systems around it
- Create an ecosystem approach to capacity building
  - e.g., through services, platforms, education of many types
- Incubate use-inspired solutions to scale
- Explore new models of allocation
  - e.g., service unites, credits, tokens, aggregated workflow coops
- Develop models of sustainability and scale
  - e.g., public/private partnerships, NGOs, consortiums, cooperatives

NATIONAL DATA PLATFORM

Addressing the Missing Middle for AI-enabled Data-driven Research and Education Workflows

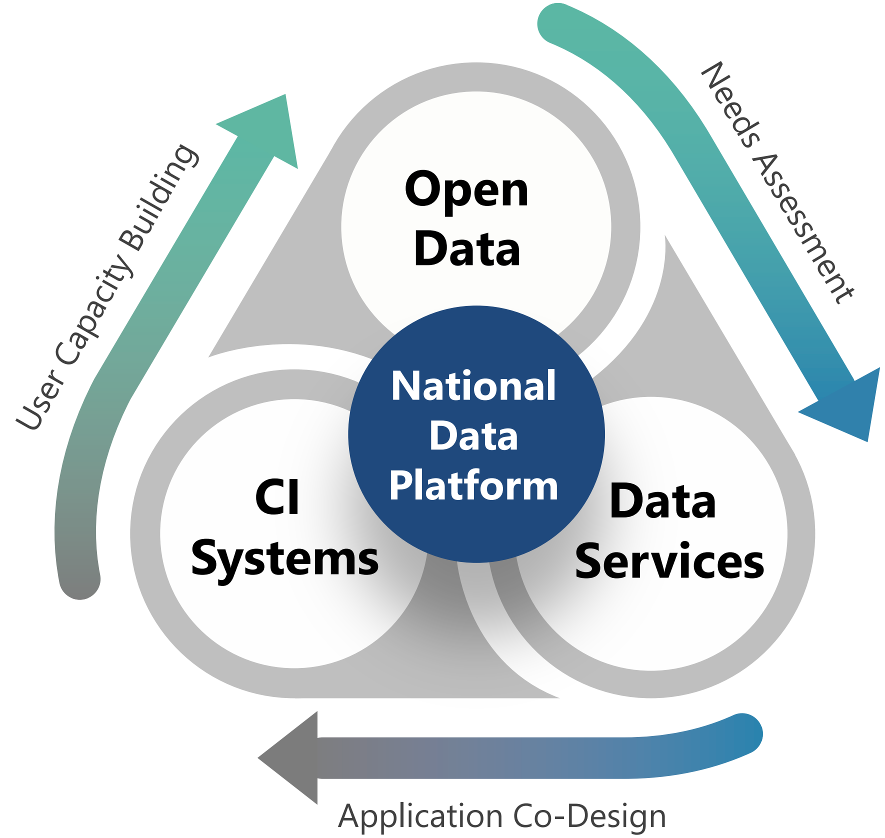


http://www.nationaldataplatform.org

NATIONAL DATA PLATFORM

National Data Platform Pilot: Services for Equitable Open Access to Data

nationaldataplatform.org



National Data Platform is a federated and extensible data and service ecosystem to promote collaboration, innovation and equitable use of data on top of existing cyberinfrastructure capabilities.

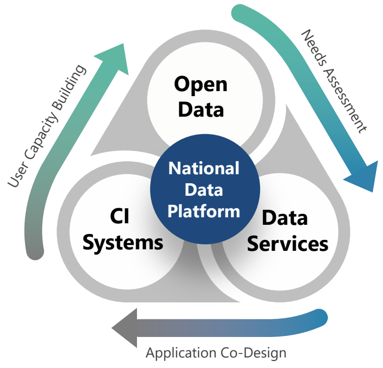
NDP enables AI-integrated science workflows that foster discovery, decision-making, policy formation and societal impact related to wildfire, climate, earthquake and food security among others.

Link to the award abstract: https://www.nsf.gov/awardsearch/showAward?AWD_ID=2333609

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

NATIONAL DATA PLATFORM

Our Use-Inspired Approach



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

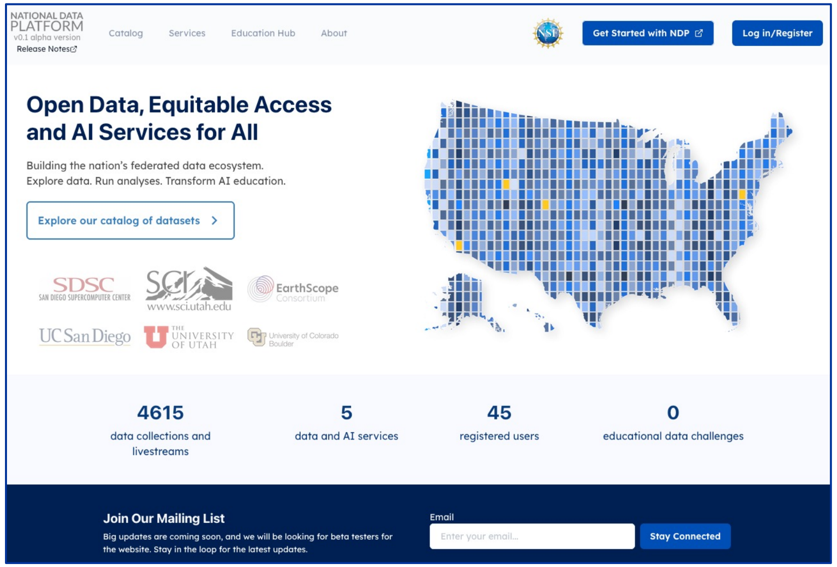
Composable Services

Composable Systems and Platforms

Sustainable and Scalable Use

- Distributed in nature
- Composition as a principle
- Sustain the backbone and the interface
- Integrate in education systems

Collaboration, Incubation, Allocation and Partnership Models


NATIONAL DATA PLATFORM
v0.1 alpha version
Release Notes

Open Data, Equitable Access and AI Services for All

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

4615
data collections and livestreams

5
data and AI services

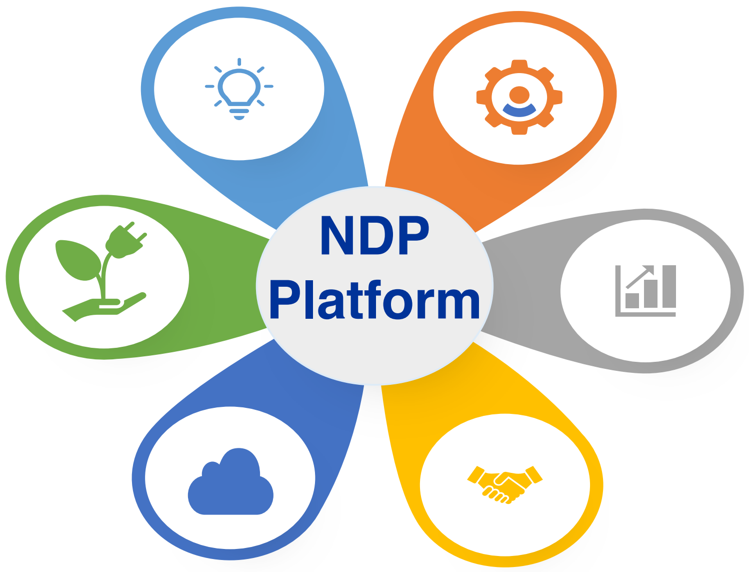
45
registered users

0
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.

Email
Enter your email...
Stay Connected



NATIONAL DATA
PLATFORM

Discover and use interconnected data hubs and user-facing services deployed on CI

Develop and deploy services, application workflows and educational challenges

Fostering scientific understanding, decision-making, policy formation and societal impact

Focus: use of existing data repositories to scientists and nonexpert users, making technology accessible to those without access to data and AI expertise

Objectives:

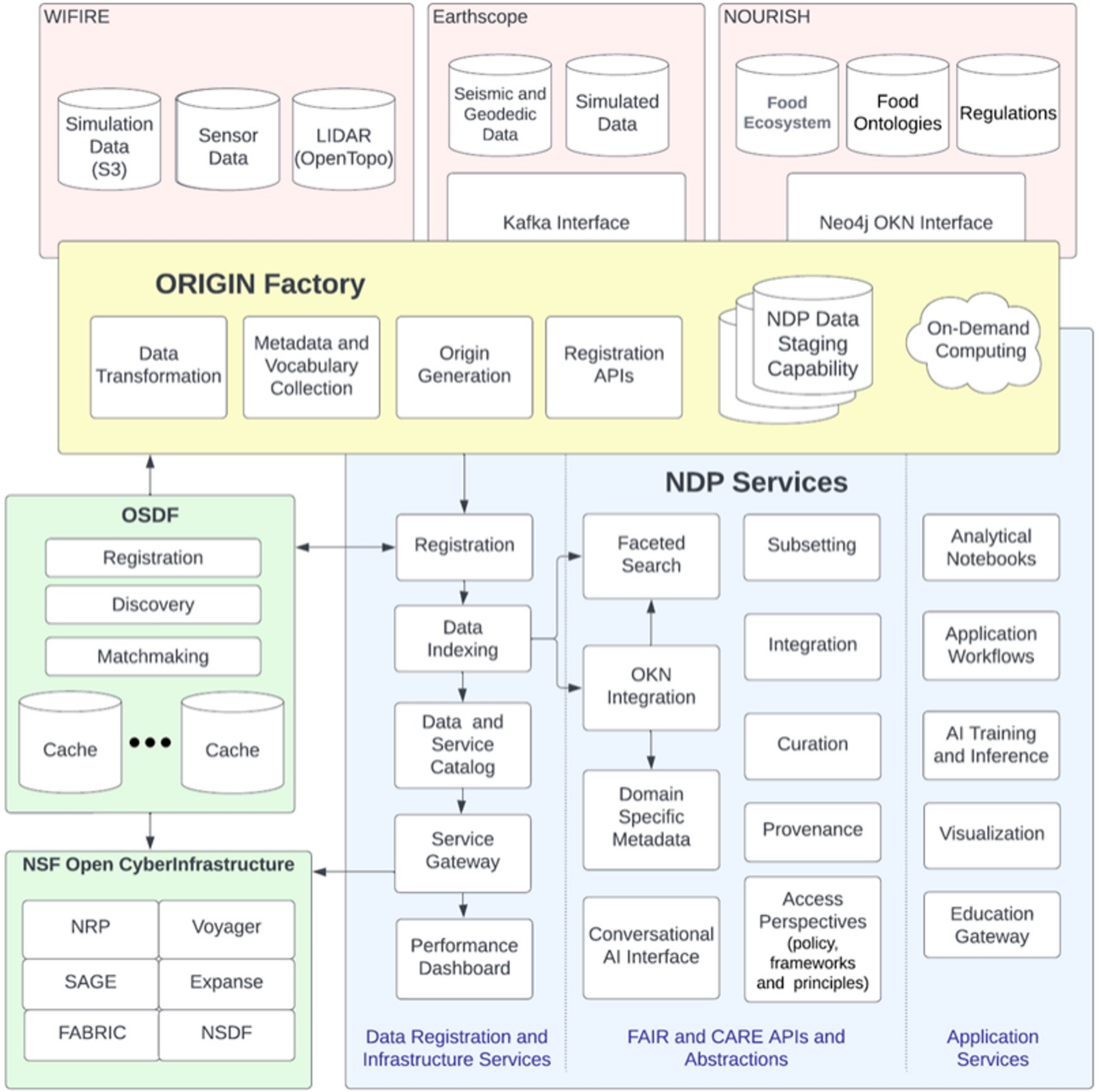
- contribute to a more equitable data and AI research
- build a broadly accessible data ecosystem
- enable diversity in
  - data sources
  - perspectives and experiences of students and researchers
  - research practices and governance processes
- equitably manage both the benefits and risks related to AI

Case studies initially focused on earth sciences and food security but designed to be generalizable.

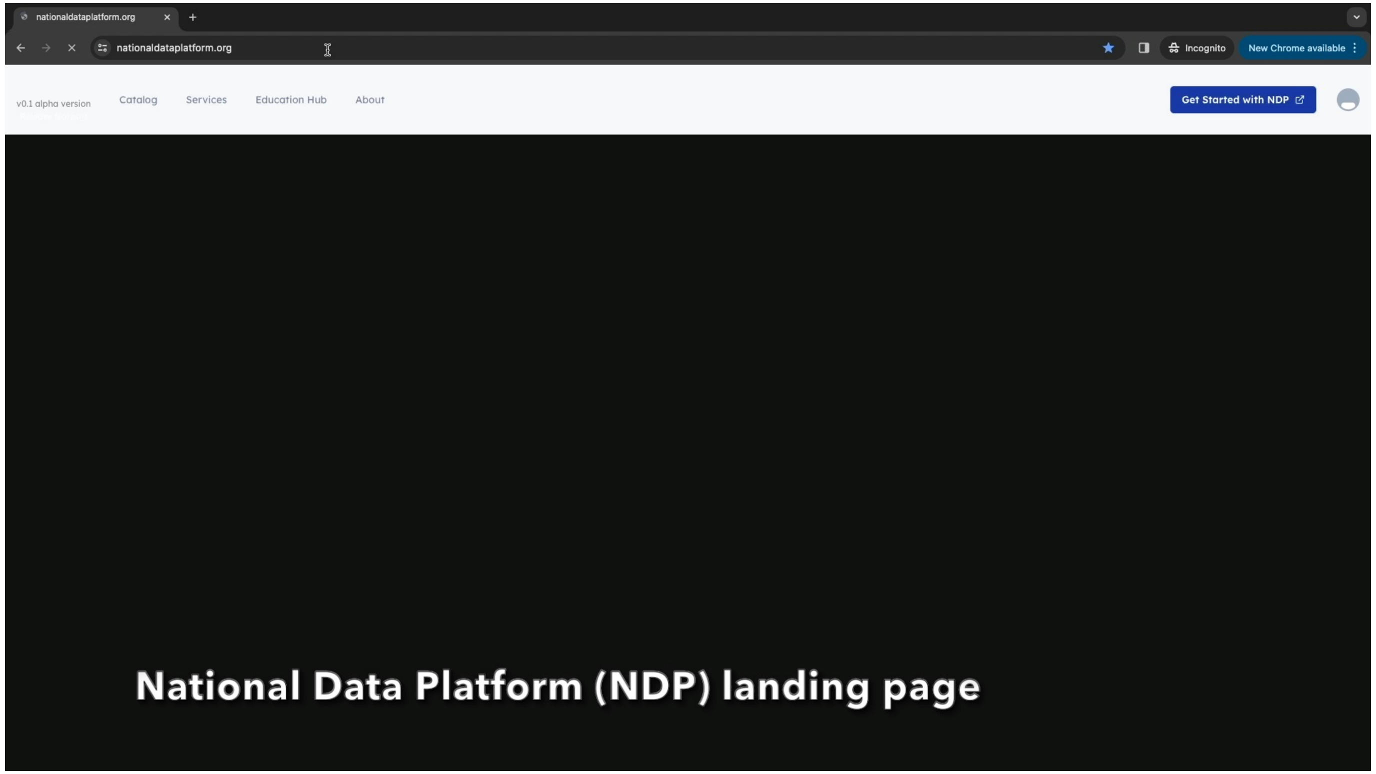
NATIONAL DATA PLATFORM

Create reusable capabilities and amplify the value of existing data repositories to benefit science, society and education.

Reference Architecture

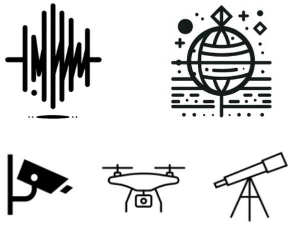


Links data and cyberinfrastructure with domain specific platforms to enable value add services and open educational capabilities.

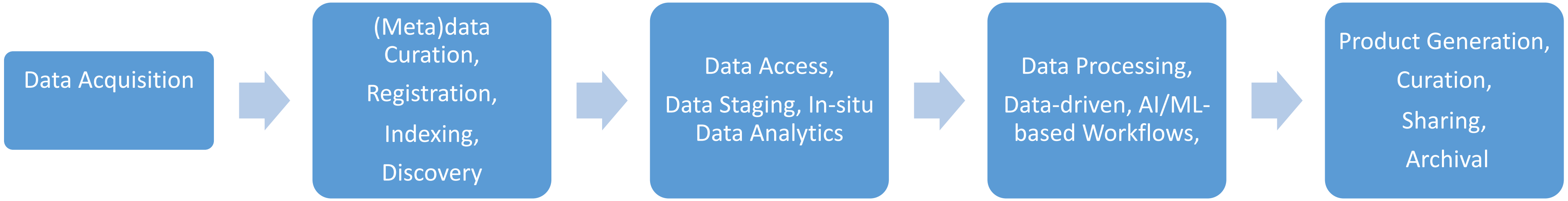


National Data Platform (NDP) landing page

Example NDP AI in Science Workflow



Data sources (instruments, sensors, facilities, repositories, etc.)


Data Acquisition
(Meta)data Curation, Registration, Indexing, Discovery
Data Access, Data Staging, In-situ Data Analytics
Data Processing, Data-driven, AI/ML-based Workflows,
Product Generation, Curation, Sharing, Archival

Data Acquisition

- Data stream adaptor
- Object-based data adaptor (e.g., AWS S3)
- API-based data adaptor

(Meta)data Curation, Registration, Indexing, Discovery

- Metadata catalog (CKAN)
- Search engine

Data Access, Data Staging, In-situ Data Analytics

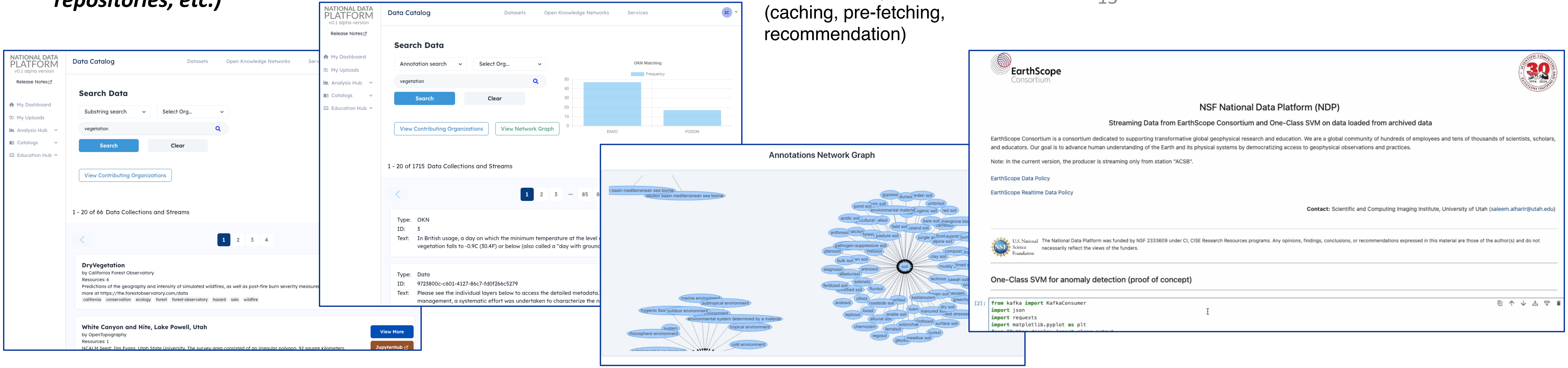
- Pelican (OSDF) data origins; data federation
- High-performance data staging/in-situ processing
- Data access optimization (caching, pre-fetching, recommendation)

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

- AI Gateway (Jupyter)
- Resources: PATh/OSG, CloudBank, ACCESS
- Data-driven stream processing

Product Generation, Curation, Sharing, Archival

- Data sink adaptor
- Data re-streaming
- Archival at Pelican (OSDF)



NATIONAL DATA PLATFORM v0.1 alpha version

Data Catalog — Search Data (Substring search, vegetation)

1 - 20 of 66 Data Collections and Streams

DryVegetation
by California Forest Observatory
Resources: 6
Predictions of the geography and intensity of simulated wildfires, as well as post-fire burn severity measure
more at https://the.forestobservatory.com/data
california conservation ecology forest forest-observatory hazard salo wildfire

White Canyon and Hite, Lake Powell, Utah
by OpenTopography
Resources: 1
NCALM Seed: Jim Evans, Utah State University. The survey area consisted of an irregular polygon, 92 square kilometers.

Data Catalog — Search Data (Annotation search, vegetation)

1 - 20 of 1715 Data Collections and Streams

Type: OKN
ID: 3
Text: In British usage, a day on which the minimum temperature at the level of vegetation falls to -0.9C (30.4F) or below (also called a "day with ground

Type: Data
ID: 9723800c-c601-4127-86c7-fd0f266c5279
Text: Please see the individual layers below to access the detailed metadata. management, a systematic effort was undertaken to characterize the n

Annotations Network Graph

NSF National Data Platform (NDP)

Streaming Data from EarthScope Consortium and One-Class SVM on data loaded from archived data

EarthScope Consortium is a consortium dedicated to supporting transformative global geophysical research and education. We are a global community of hundreds of employees and tens of thousands of scientists, scholars, and educators. Our goal is to advance human understanding of the Earth and its physical systems by democratizing access to geophysical observations and practices.

Note: In the current version, the producer is streaming only from station "ACSB".

EarthScope Data Policy

EarthScope Realtime Data Policy

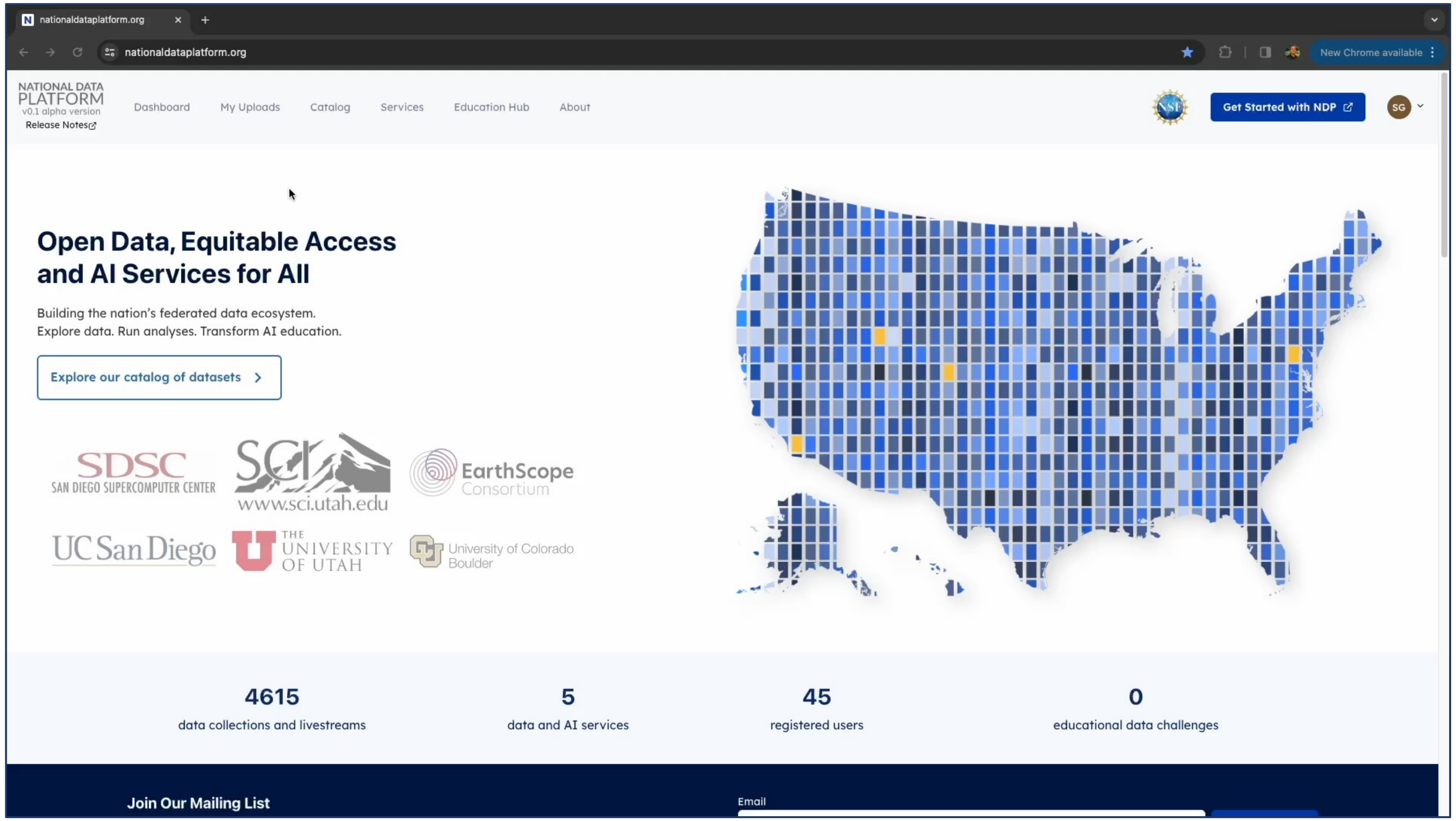
Contact: Scientific and Computing Imaging Institute, University of Utah (saleem.alharir@utah.edu)

The National Data Platform was funded by NSF 2333609 under CI, CISE Research Resources programs. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

One-Class SVM for anomaly detection (proof of concept)

```
[2]: from kafka import KafkaConsumer
     import json
     import requests
     import matplotlib.pyplot as plt
```

Example NDP NAIRR Integration



nationaldataplatform.org

NATIONAL DATA PLATFORM v0.1 alpha version Release Notes

Dashboard | My Uploads | Catalog | Services | Education Hub | About

Get Started with NDP

Open Data, Equitable Access and AI Services for All

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

Explore our catalog of datasets >

SDSC San Diego Supercomputer Center · SCI www.sci.utah.edu · EarthScope Consortium · UC San Diego · The University of Utah · University of Colorado Boulder

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

Join Our Mailing List

Email

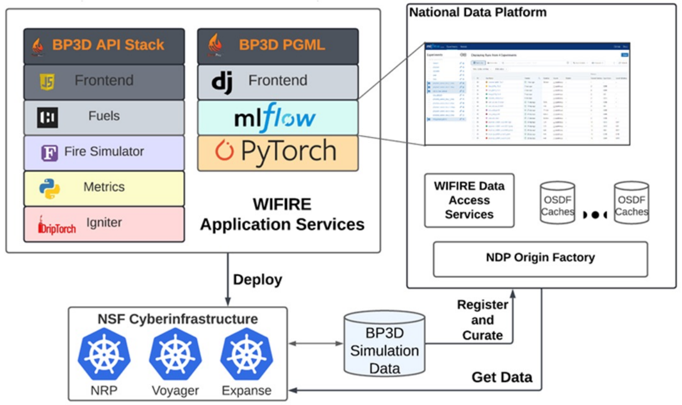
- Find NAIRR resources
- Learn from notebook examples
- Use NAIRR allocations

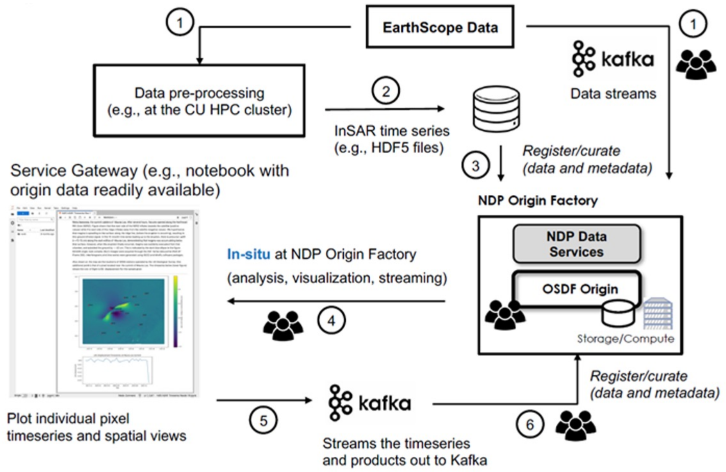
- Build new models and data
- Publish via GitHub and HuggingFace
- Create narrative stories and educational resources for others

NATIONAL DATA PLATFORM

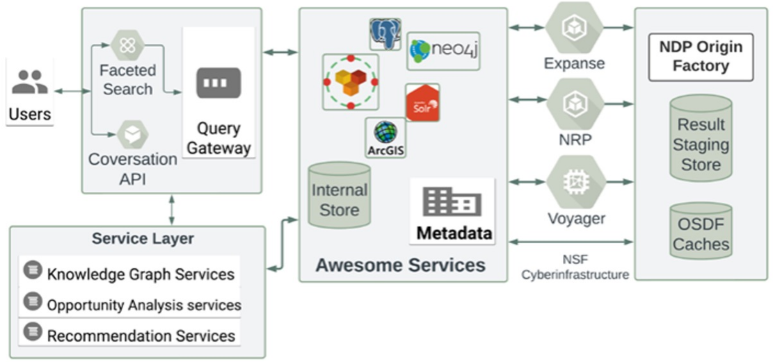
Case Studies for Generalizable Workflows

- **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.



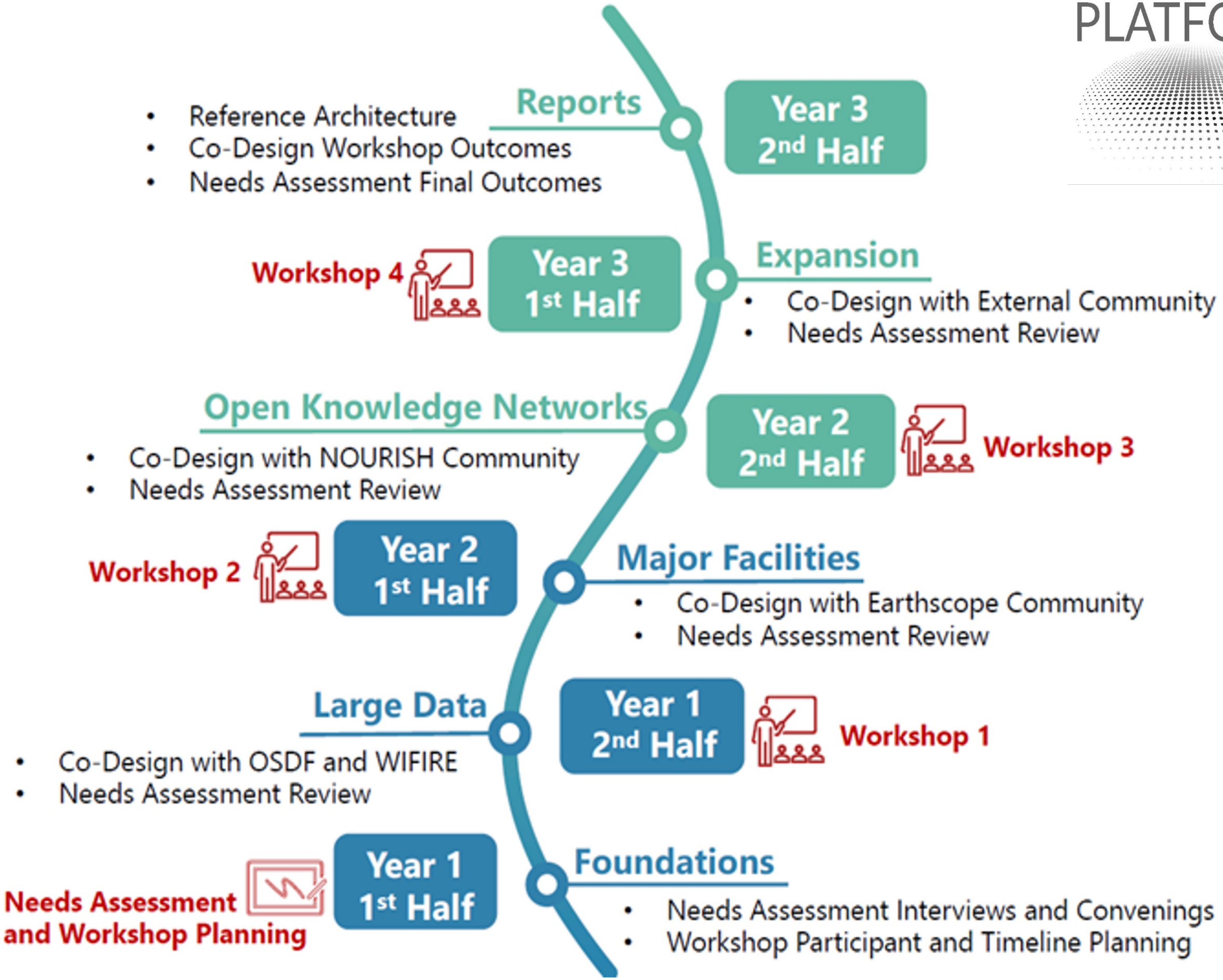


1
EarthScope Data
1
Data pre-processing (e.g., at the CU HPC cluster)
2
InSAR time series (e.g., HDF5 files)
kafka
Data streams
3
Register/curate (data and metadata)
Service Gateway (e.g., notebook with origin data readily available)
NDP Origin Factory
NDP Data Services
OSDF Origin
Storage/Compute
In-situ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



Community expansion and stakeholder engagement

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



Year 1 1st Half

Needs Assessment and Workshop Planning

Foundations

- Needs Assessment Interviews and Convenings
- Workshop Participant and Timeline Planning

Year 1 2nd Half

Workshop 1

Large Data

- Co-Design with OSDF and WIFIRE
- Needs Assessment Review

Year 2 1st Half

Workshop 2

Major Facilities

- Co-Design with Earthscope Community
- Needs Assessment Review

Year 2 2nd Half

Workshop 3

Open Knowledge Networks

- Co-Design with NOURISH Community
- Needs Assessment Review

Year 3 1st Half

Workshop 4

Expansion

- Co-Design with External Community
- Needs Assessment Review

Year 3 2nd Half

Reports

- Reference Architecture
- Co-Design Workshop Outcomes
- Needs Assessment Final Outcomes

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.

Three Co-Design Workshops

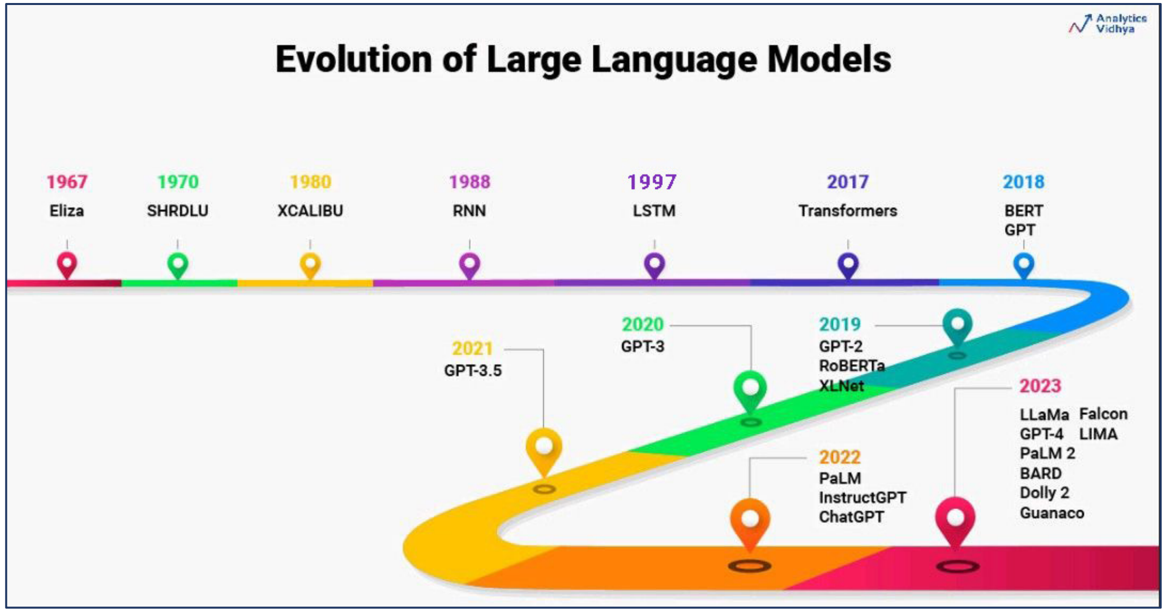
Each will include a breakout session to develop a data challenge question specific to large data (W1); streaming data (W2); and graph data (W3).

NATIONAL DATA PLATFORM

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

An NDP Service Example:
Generative AI and Large Language Models (LLM)

- Huge generative potential
- Ability to create human-like outputs
- Integration with complex models
- Libraries advanced technologies
  - e.g., GPT, Prompt Engineering, and vector storage
- Shortcomings on domain expertise
- Need domain-specific LLMs
  - with human-curated data and controlled knowledge



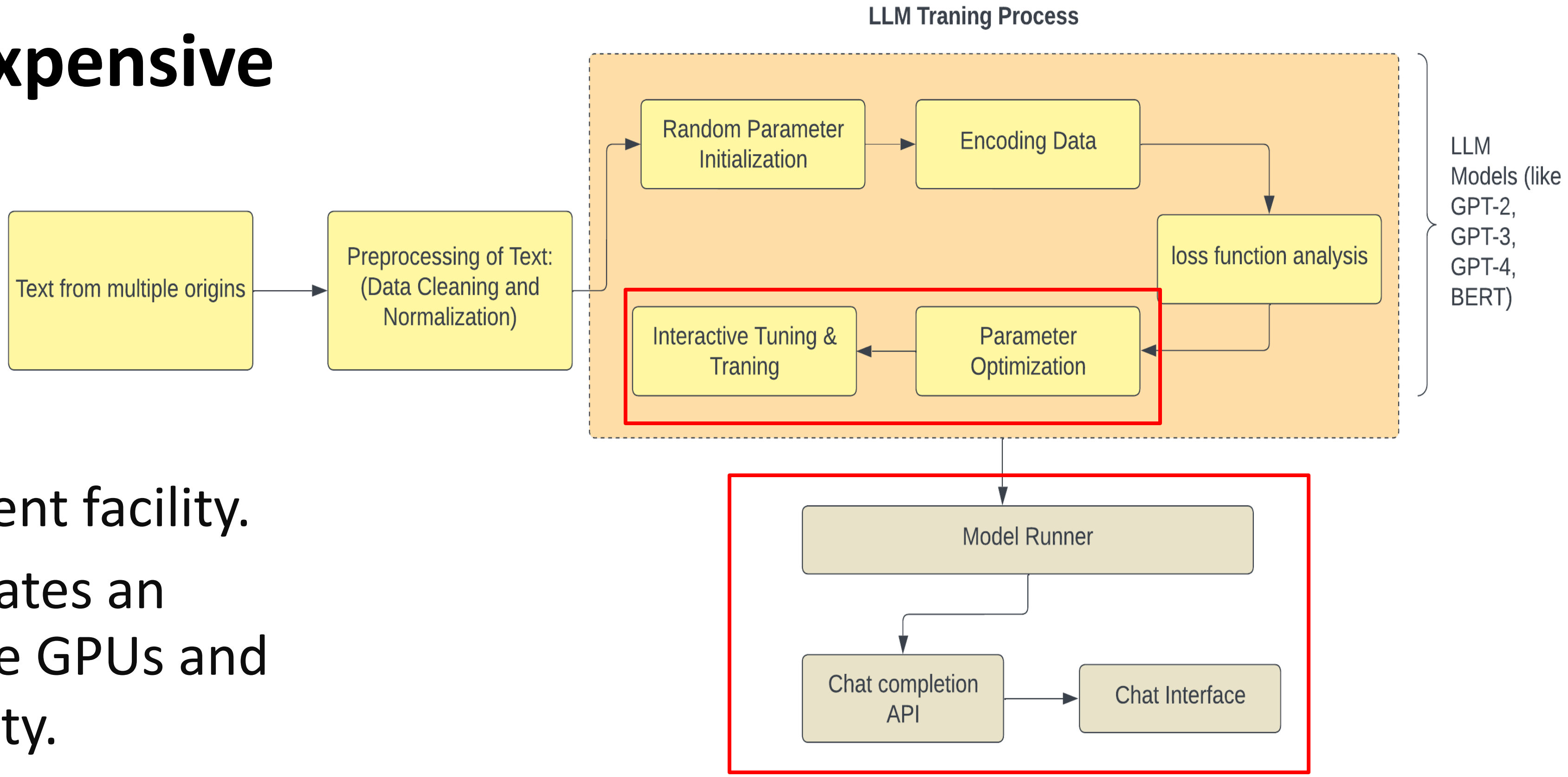
Source: https://www.analyticsvidhya.com/blog/2023/07/beginners-guide-to-build-large-language-models-from-scratch

> “Generative AI helped workers avoid awful ideas, but it also led to more average ideas”
> - **Harvard Business Report (March – April 2024)**

Accessing and Using LLMs is an Equity Issue

LLM Deployment is Expensive

- Even tuning an LLM can incur substantial costs, necessitating 4-5 AT100 GPUs, expansive nodes, and an equipped deployment facility.
- Operating an LLM necessitates an infrastructure with multiple GPUs and substantial memory capacity.


LLM Traning Process
Text from multiple origins
Preprocessing of Text: (Data Cleaning and Normalization)
Random Parameter Initialization
Encoding Data
loss function analysis
Parameter Optimization
Interactive Tuning & Traning
LLM Models (like GPT-2, GPT-3, GPT-4, BERT)
Model Runner
Chat completion API
Chat Interface

LLM Deployments

Commercial LLM

Pros:

- Ready to use
- Large knowledge base
- Low latency and distributed deployment
- Robust API with security

Cons:

- Lack of domain-specific knowledge
- High costs
- Information security and privacy concerns

Community LLM

Pros:

- Domain-specific knowledge
- Average latency
- Low costs
- Community-owned

Cons:

- Need infrastructure
- Dedicated tech and knowledge team
- Community volunteers to manage everything
- Privacy concerns

Private LLM

Pros:

- Usage mode is private access
- No knowledge control and security issues

Cons:

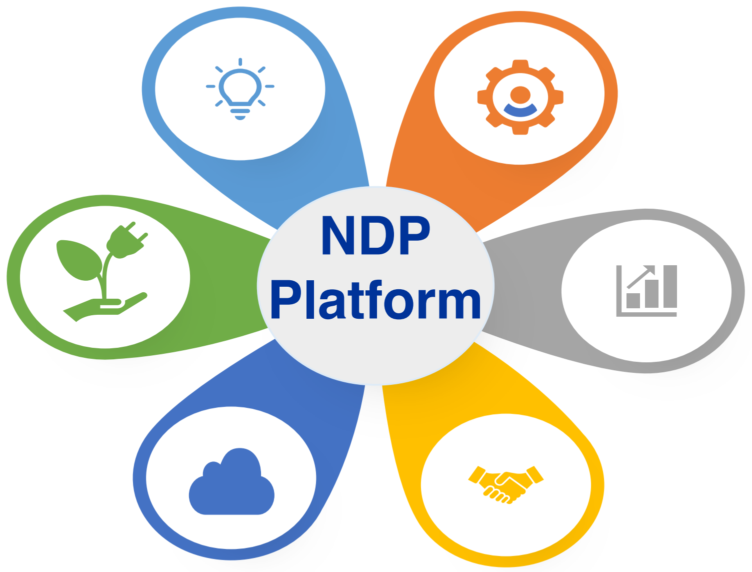
- Required hardware and software run models

NDP
LLM as a Service

- Tailored Model Selection
- Enhanced Data Control
- Privacy and Security
- Cost Efficiency
- OpenAI API and LangChain Support

LLM Client Service

- Use an existing model
- Add context with domain-specific documents



LLM Training Service

- Fine-tune an existing model to create a new model
- Use a larger corpus for training
- Deploy as a service

NDP LLM as a Service

Alignment with NAIRR Objectives

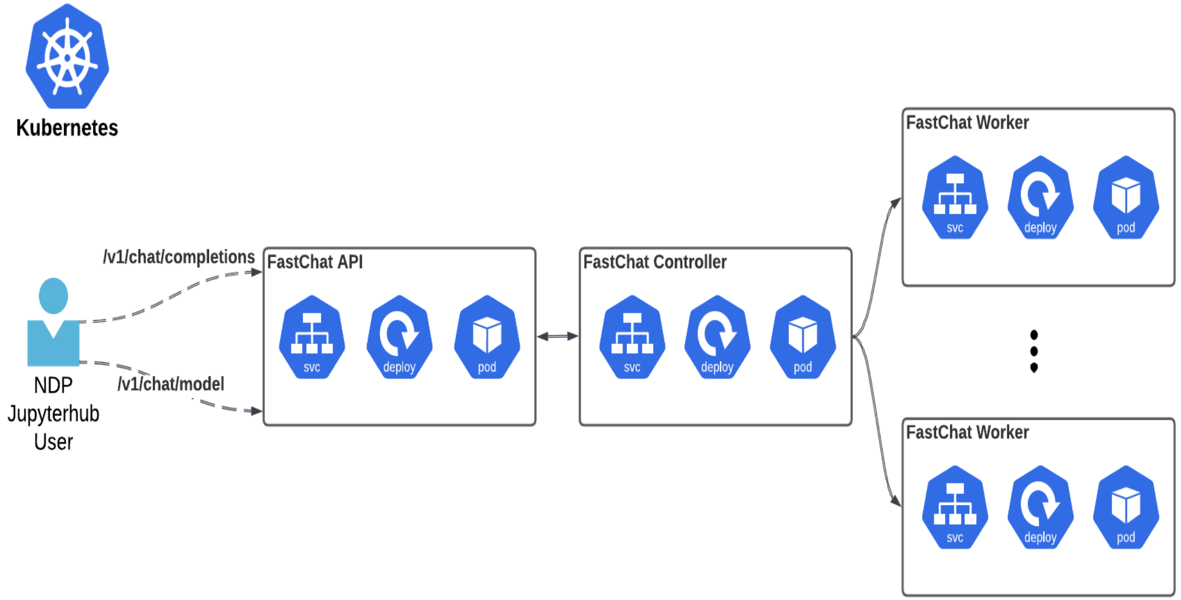
- **Capacity** to support many users with a spectrum of backgrounds
- **Capabilities**
  - Ability to train *(and use)* resource-intensive AI models on CI resources
  - Ability to make use of a mix of computational resources
  - Option to select which resources to use through a range of mechanisms, including ... optionally interactive “notebook”-like environments
  - A NAIRR system should include at least one large-scale machine-learning supercomputer capable of training 1 trillion-parameter models

↑

In today’s tutorial we are using a model with
***7B** parameters running on NRP*

NDP LLM Deployment Architecture

- FastChat
  - Open source LLM execution library
  - Deployed on Nautilus
    - API Server
    - Controller
    - Worker (serves different or the same LLMs)
- Currently all workers are serving the following LLMs
  - eci-io/climategpt-7b,
  - ECarbenia/grimoiresigils
  - text-embedding-ada-002



Part 2 Comin Up at 4:30pm:

NDP LLM-as-a-Service on NRP Tutorial

How can we work with you?

UC San Diego · UTAH · University of Colorado Boulder · SDSC SAN DIEGO SUPERCOMPUTER CENTER · SCI www.sci.utah.edu · EarthScope Consortium

Contact: Ilkay Altintas, Ph.D.

Email: ialtintas@ucsd.edu