nationaldataplatform.org

Category I: National Data Platform

A Federated National AI-Ready Data Ecosystem for Discovery, Innovation, Education, and Workforce Development

NSF IDSS · #2609447

Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

Introductory Presentation for IDSS PIs

NATIONAL DATA PLATFORM

What We'll Cover

1. The national challenge & NDP vision
2. Our starting points
3. Architecture: connector nodes, platform, portal
4. Five AI-driven workflow patterns
5. Eight co-design application domains
6. Production operations & user access
7. Team, timeline & collaboration opportunities

PI: Ilkay Altintas, UC San Diego

Category I: National Data Platform

The Challenge: Fragmented U.S. Data & AI Ecosystem

The nation's distributed CI needs to keep pace with AI-driven science

NATIONAL DATA PLATFORM

Data is Siloed

Repositories at NSF facilities, federal agencies, and universities exist in isolation — incompatible formats, access barriers, no common discovery layer.

Compute is Fragmented

HPC centers, edge instruments, and commercial clouds are powerful but disconnected. Bridging them requires manual, error-prone effort.

AI is Hard to Integrate

Modern AI workflows demand seamless movement from data ingestion → model training → inference. Disconnected systems break this loop.

Workforce is not Ready

Students and researchers lack hands-on access to integrated CI environments. Education lags behind the tools science actually uses.

Result: The rate of AI progress is outpacing the nation's CI and workforce capacity.

How can we bridge the nation's distributed data repositories and facilities with its computing capabilities to accelerate the translation of data and AI into progress?

NDP Vision & Goals

Transforming fragmented CI into a unified, AI-ready national ecosystem

NATIONAL DATA PLATFORM

Vision: Enable NDP users to seamlessly connect to and harness U.S. CI — transforming data-to-knowledge from being data-limited to knowledge-driven through composable, discoverable, interoperable and trustworthy AI services at the fingertips of researchers, educators, and learners nationwide.

Unify

Connect distributed national data and compute into an AI-ready ecosystem where multimodal data, models, and services can be composed and reused.

Accelerate

Enable seamless movement from data discovery and AI training to inference and simulation through interoperable, extensible and trustworthy services.

Democratize

Reduce barriers via a user-centered portal, ready-to-use workflows, and broad support across institutions and experience levels.

Innovate

Accelerate a growing ecosystem of AI and data services through standardized connector nodes and a robust build-test-deploy capability.

NATIONAL DATA PLATFORM

Target Capabilities

Multimodal Governance and Interoperability Fabric

Who is NDP for?

A broad user community from individual researchers to entire institutions and the nation

Researchers & Scientists

Seamlessly access distributed datasets, models, and AI tools. Build and execute scalable workflows without managing CI.

Data & Software Engineers

Extend the ecosystem via open APIs and SDKs. Register, deploy, and maintain composable services for reuse.

AI Developers

Use the Innovation Factory to transform research code into reusable community assets — prototype to production.

Program Managers

Leverage observability dashboards and FAIR metrics to monitor use, impact, and resource utilization.

Educators & Students

Access the Education Hub for hands-on training, data challenges, and AI-in-science learning.

Institutions & Communities

Connect local data and compute via Connector Nodes. Maintain autonomy while joining the national federation.

And all IDSS projects...

NATIONAL DATA PLATFORM

Our Starting Points

Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

UC San Diego
The University of Utah
Texas — The University of Texas at Austin
University of Washington
UMass
Princeton University
Duke University
Loyola Marymount University
University of Oregon
Purdue University
Iowa State University
NCSA
SCI — www.sci.utah.edu
eScience Institute
Pittsburgh Supercomputing Center
MGHPCC — Massachusetts Green High Performance Computing Center
TACC — Texas Advanced Computing Center

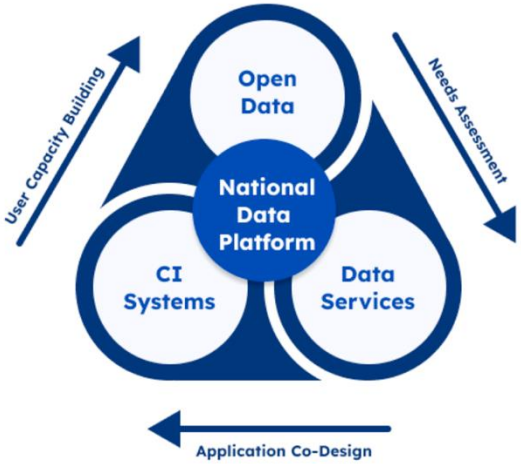
PI: Ilkay Altintas, UC San Diego

Last three years…

nationaldataplatform.org

National Data Platform Pilot

NATIONAL DATA PLATFORM



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

A scalable **platform** for using, developing and deploying composable services and application workflows at **distributed connector nodes**

Central Deployment of NDP

- **Extensive** catalog of data & services
- Advanced **data discovery** tools
- **Collaborative** workspaces
- **Interactive** classrooms & data challenges
- Open knowledge networks & **conceptual search**

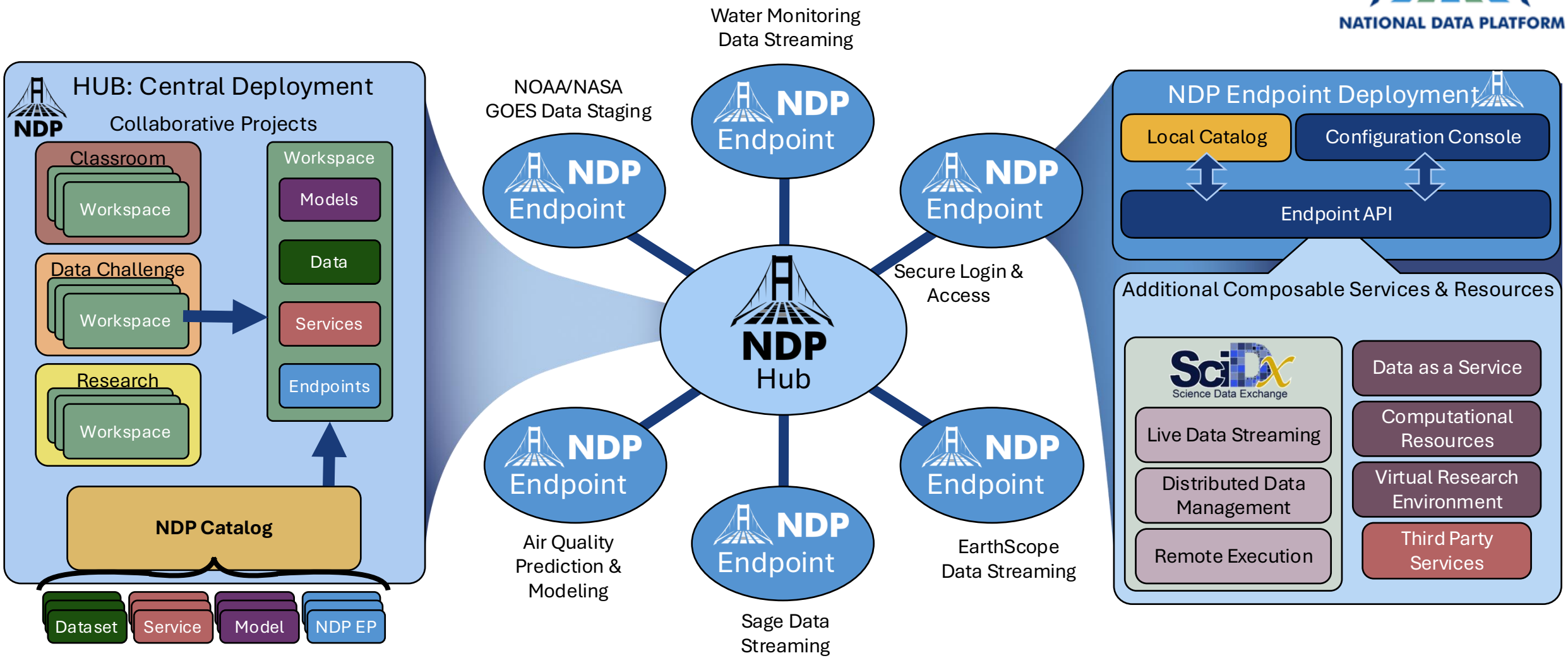
Lightweight software stack that enables a **custom data experience**

- Deployed virtually **anywhere**
  - **Near data** or compute
- Connects to **NDP core services**
- **Customized** services and data access
- **Bring-your-own** resources
- **Community** controlled

PI: Ilkay Altintas, UC San Diego

Category I: National Data Platform

NATIONAL DATA PLATFORM



https://www.nationaldataplatform.org/

PI: Ilkay Altintas, UC San Diego

Tapis

About Tapis

- Cloud-hosted framework of secure REST APIs for research computing
- Supports data management and code execution across HPC, HTC, and cloud systems
- Includes SDKs, CLI, and ecosystem tooling for integration into workflows and gateways
- Helps investigators use computing resources across institutions and scientific domains

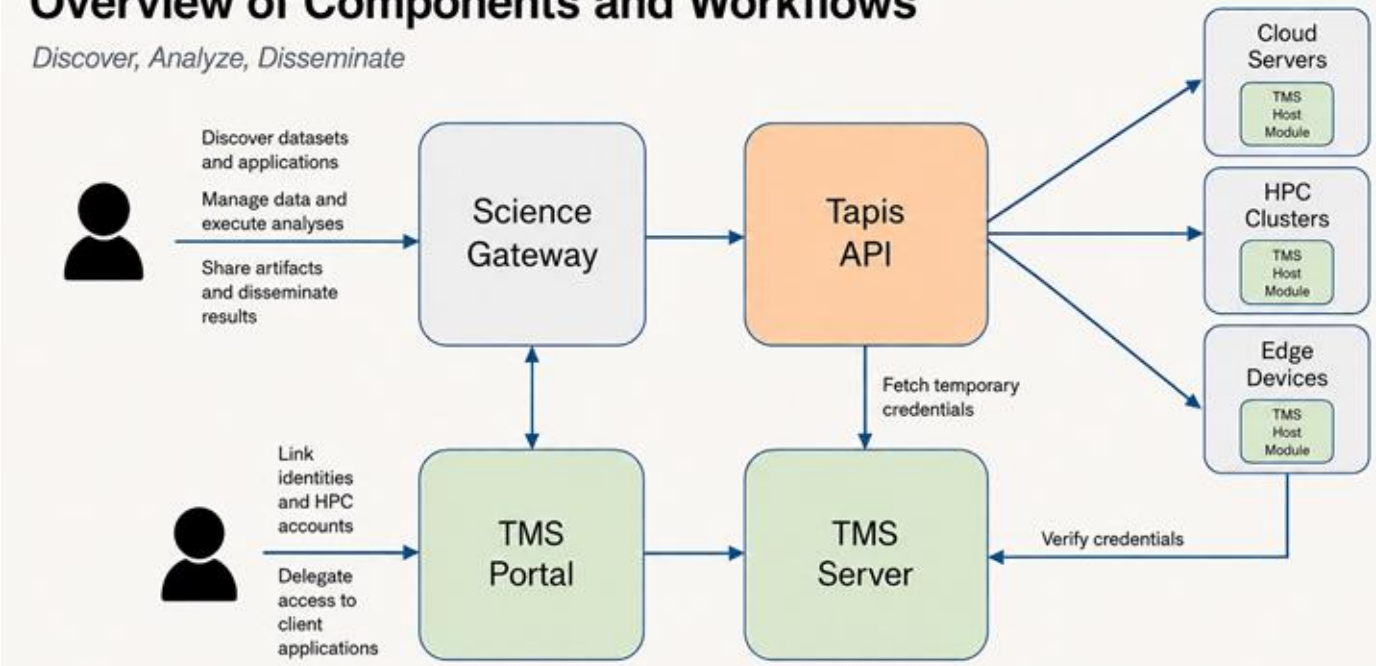
High-Level Objectives

1. **Programmable Infrastructure**
   Expose programmable interfaces to complex cyberinfrastructure so workflows can be automated
2. **Repeatability**
   Record inputs, outputs, software versions, and execution environments to support repeatability and reproducibility
3. **Secure Collaboration**
   Provide access controls on resources so they can be shared privately with individuals or research communities

NATIONAL DATA PLATFORM

Overview of Components and Workflows

Discover, Analyze, Disseminate



TMS: Automated Credential Management and Remote Command Execution

Core Features

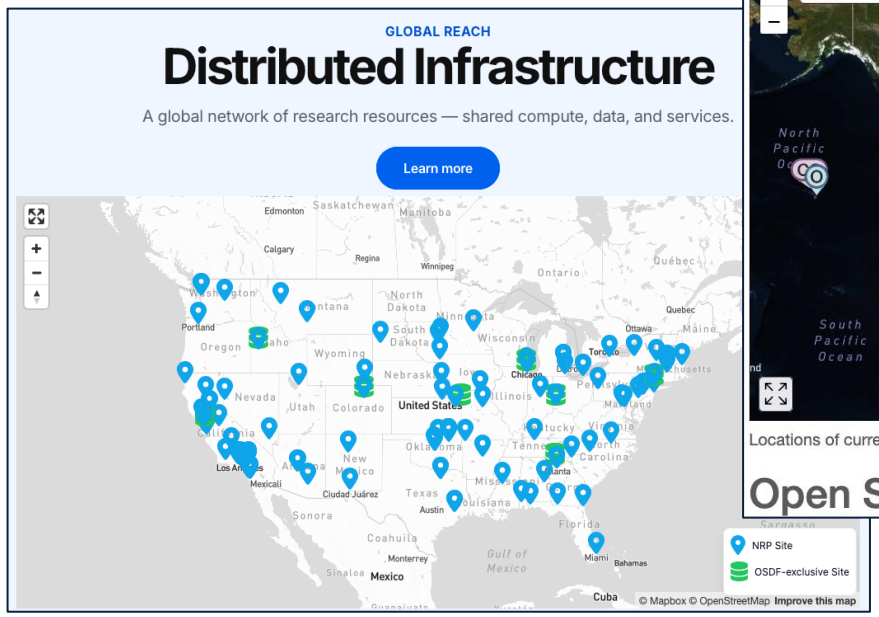
- No manual key distribution
- No user secrets shared with applications
- Account access using federated identities
- Secure execution
- Easy deployment and administration

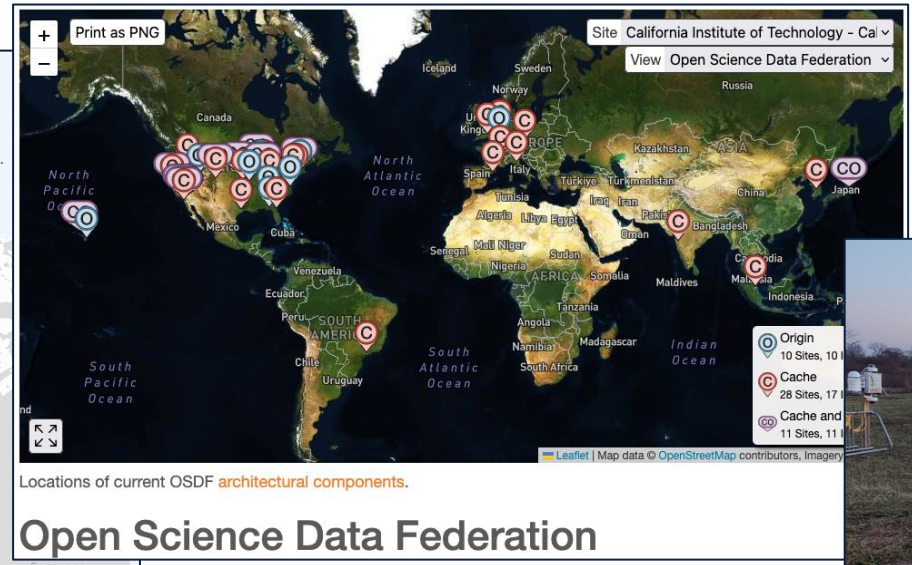
In production on all major TACC HPC systems including Stampede3, Frontera, Vista, Lonestar6 and Corral.

NRP, Cloudbank, Sage, Expanse and Pelican/OSDF

NATIONAL DATA PLATFORM

NATIONAL RESEARCH PLATFORM





Locations of current OSDF architectural components.

Open Science Data Federation

Sage

Sage Grande Testbed

CloudBank2 ACCELERATING SCIENCE and ENGINEERING in the COMMERCIAL CLOUD

Managed Services
User Portal
Education Hub
Financial Monitoring
*Data Transfer
*Community & Gateway Tools
*Regulatory Data Management

Support
Training
Help Desk
Documentation
Curated Research Workflows
*Cloud Clinics
*Application Engineering

SERVICE-BASED RESOURCE
Community-Piloted | Launch-Ready

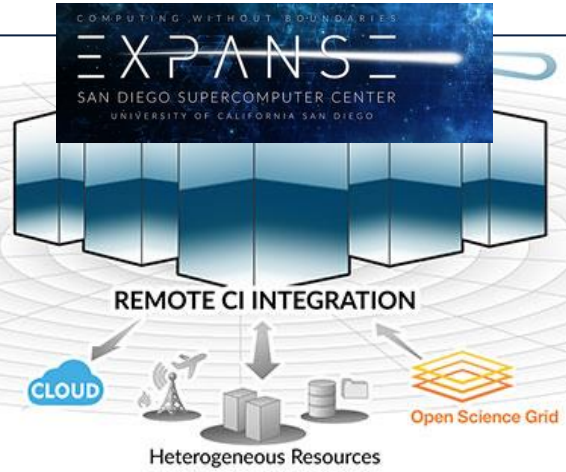
Explore | Discover | Accelerate

Multi-Cloud
Amazon Web Services
Google Cloud
IBM Cloud
Microsoft Azure
Future Commercial Clouds*

Example Cloud Applications
Artificial Intelligence
High-Performance Computing
Network Research
Edge Computing
Elastic Computing
Quantum Computing
Secure Computing*
Large-scale Collaborations*

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

NATIONAL DATA PLATFORM

Architecture and Cyberinfrastructure

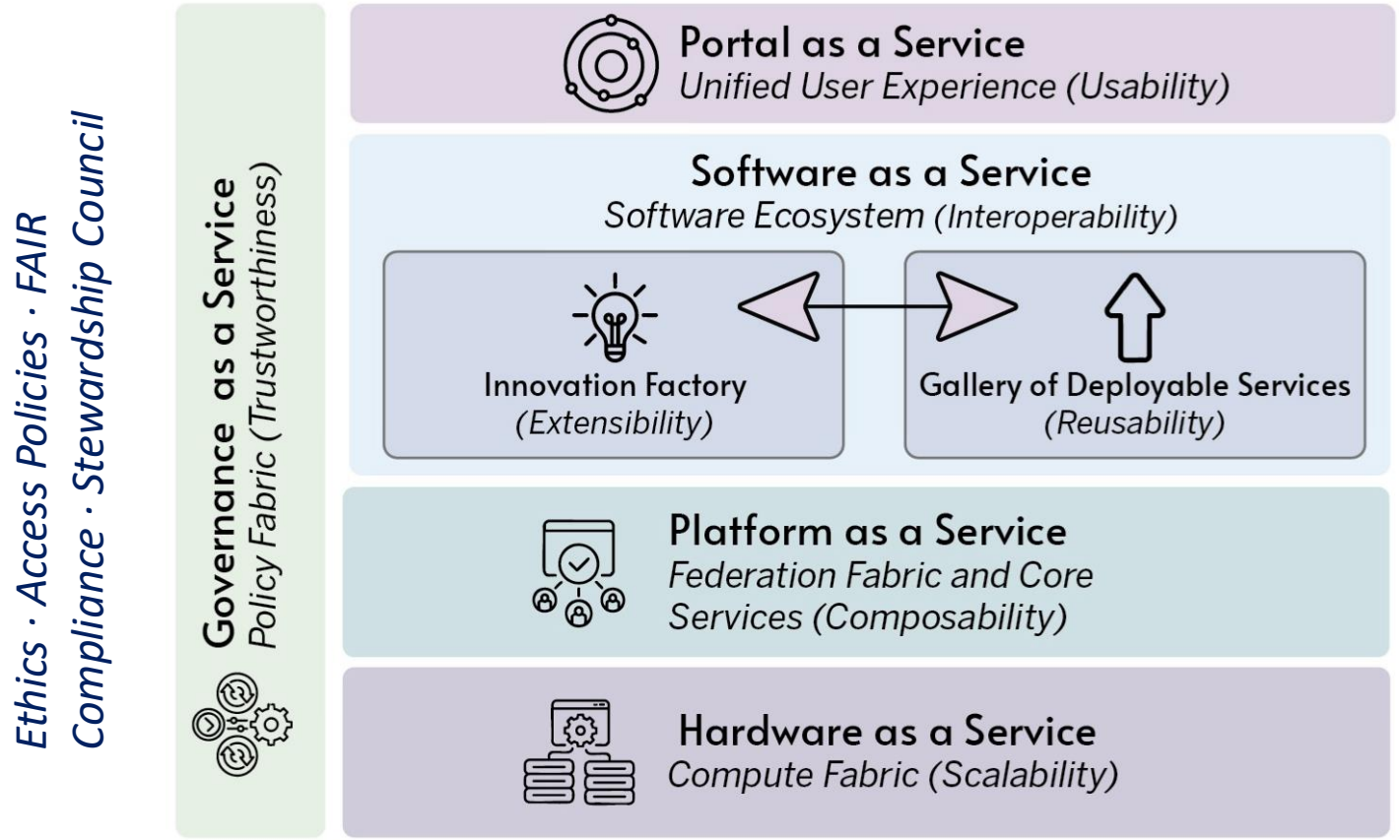
Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

UC San Diego
The University of Utah
TEXAS The University of Texas at Austin
University of Washington
UMass
Princeton University
Duke University
Loyola Marymount University
University of Oregon
Purdue University
Iowa State University
NCSA
SCI www.sci.utah.edu
eScience Institute
Pittsburgh Supercomputing Center
MGHPCC Massachusetts Green High Performance Computing Center
TACC Texas Advanced Computing Center

NDP Architecture:
A Layered Microservices Ecosystem with 7 principles

NATIONAL DATA PLATFORM

Five integrated layers from hardware to user experience



Unified User Experience — Discovery · Workspaces · Dashboards · Collaboration

Gallery of Deployable Services · Innovation Factory (Build → Test → Validate → Deploy)

Auth & Identity · Catalog · Discovery · HPC/HTC Integration (Tapis) · Kubernetes Orchestration

Head Nodes + Connector Nodes at HPC centers · Cloud · Host-owned

A user-centered architecture, backed by scalable hardware, a composable federation platform, and an interoperable, extensible software ecosystem.

PI: Ilkay Altintas, UC San Diego

Category I: National Data Platform

Hardware: Compute Fabric

A federation of Head Nodes and Connector Nodes bridging data and compute

NATIONAL DATA PLATFORM

Head Nodes (Control Plane)

- Host core infrastructure & portal in highly available, load-balanced configuration
- Coordinate federation-wide policies & maintain consistent state
- 3 deployments: SDSC (San Diego) · CHPC (Utah) · AWS (cloud backup)
- Ensure continuity of operations and production-grade redundancy

Connector Nodes (Execution Layer)

National HPC Nodes (9)

NDP-owned — proximal to national HPC CI
Y1: SDSC · TACC · NCSA · PSC · MGHPCC · CHPC — Y2: Duke · Iowa State · Princeton

Cloud-hosted Nodes

Elastic on-demand resources near large or sensitive data; minimize data-movement — target 10 by end of Y2

Host-owned Nodes

Institutions bring their own resources via NRP/Nautilus — local control + federation — target 9 by end of Y2

All nodes run the NDP Federation Platform Stack · Orchestrated by Kubernetes via NRP or similar CI

Builds on the successful BYOR model developed by NRP (funded 44 GPU nodes in 2021 - at present, the NRP comprises 490 nodes) and OSDF (first award funded 3 caches – at present, 21 origins and 42 caches deployed globally)

PI: Ilkay Altintas, UC San Diego

Category I: National Data Platform

Platform: Federation Fabric & Core Services

NATIONAL DATA PLATFORM

The software stack that unifies distributed nodes into a composable operational platform

Auth & Identity

InCommon · KeyCloak · TMS

Fine-grained access control for data, models, services, and workflows across the federation

Catalog Service

Federated registry of multimodal data, AI/ML models, services, and workflows.

Heterogeneous metadata (DCAT, schema.org, Croissant, EML) and knowledge graphs

Provenance, insights & FAIR tracking.

Registration

AI-assisted intake for all digital assets

LLM-driven workflows (LangGraph, CrewAI) for metadata mapping, validation, enrichment, and external catalog stewardship.

Search & Discovery

Hybrid retrieval: vector search + graph reasoning + semantic query translation.

Agentic discovery, data–model matching, privacy-preserving queries.

HPC/HTC Integration

Tapis API layer — abstraction of HPC/HTC complexity.

Job submission, monitoring, containerization, data staging.

Workflows for simulation and AI.

Trust & Policy

Governance guidelines, ethical AI principles, data-sharing agreements, compliance with NSF open-science standards.

Policy-as-code agentic governance engine across the ecosystem.

Innovation Factory: Build → Test → Validate → Deploy

NATIONAL DATA PLATFORM

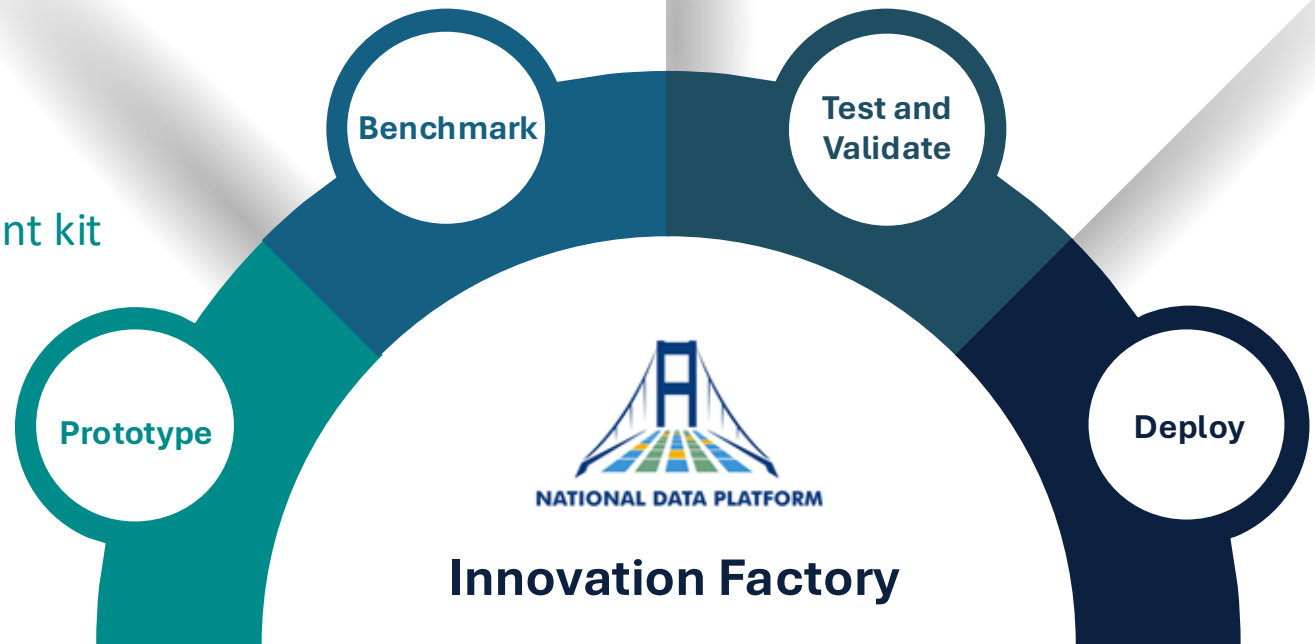
A structured end-to-end pipeline for turning ideas into production-ready digital assets

Only vetted, benchmarked, and trusted assets enter the Gallery of Deployable Services to ensure reliability, reproducibility, and operational rigor.

- Software development kit
- Container widget
- Catalog integration
- Workflow sandbox

- Data & AI benchmarks
- Telemetry
- Workflow portability

- Performance
- Security and FAIR compliance
- Reproducibility



- Production deployment
- Gallery and catalog integration
- Usage tracking
- Version control

Portal & Governance

NATIONAL DATA PLATFORM

A centralized and customizable web-based entry point backed by responsible, community-driven governance

NDP Portal as a Service

- Single sign-on & role-based user journeys (researcher · educator · developer · student)
- Dataset and digital asset discovery across the federated catalog
- Deploy services, visualize & monitor workflows
- Collaborative workspaces (CollabStudios) and shared outcomes
- Integrated training modules and AI services
- Programmatic access via APIs in addition to GUI
- "Bring Your Own Resources" to link existing allocations & compute
- Built on NDP Workspaces; Year-1 usability assessment with SGX3
- Customizable by domain-specific communities

Governance Fabric

Stewardship Council

Oversees lifecycle and stewardship of digital assets; quarterly reviews

Interoperability Advisory Board

National experts in CI, domain science, data, education — annual strategy review

Policy-as-Code Engine

Ethical AI principles, data-sharing agreements, FAIR compliance, provenance tracking, owner-controlled governance

External Evaluator

Independent evaluation from UT Austin on scientific, operational, and educational outcomes

PI: Ilkay Altintas, UC San Diego

What makes NDP unique?

NATIONAL DATA PLATFORM

Filling the gaps existing CI cannot — the missing AI-integration layer for multimodal data & AI services.

Dimension	Existing CI Gap	NDP Novel Capability
Architecture	Siloed HPC, instruments, and data repositories	Federated composable edge-HPC CI fabric for dynamic data and AI services
Interoperability	Manual interoperability at institutional boundaries	Seamless automated federation through standardized services and APIs
Data Access	Static archives and limited access services	Dynamic AI-ready multimodal data and services on demand
Compute for AI	Centralized HPC; AI as a batch component	Elastic edge-HPC continuum; full multimodal AI services instantiated on demand
Science Workflows	Difficult data/AI integration; manual hand-offs	Accelerated AI-integrated workflows from ingestion to inference
Education	Blackboard as primary classroom medium	Reusable, compute-ready, reproducible digital assets and lab environments
Partnerships	Separately managed research and private CI	Composable data and AI services across public and private resources

NATIONAL DATA PLATFORM

Design and Innovation Process

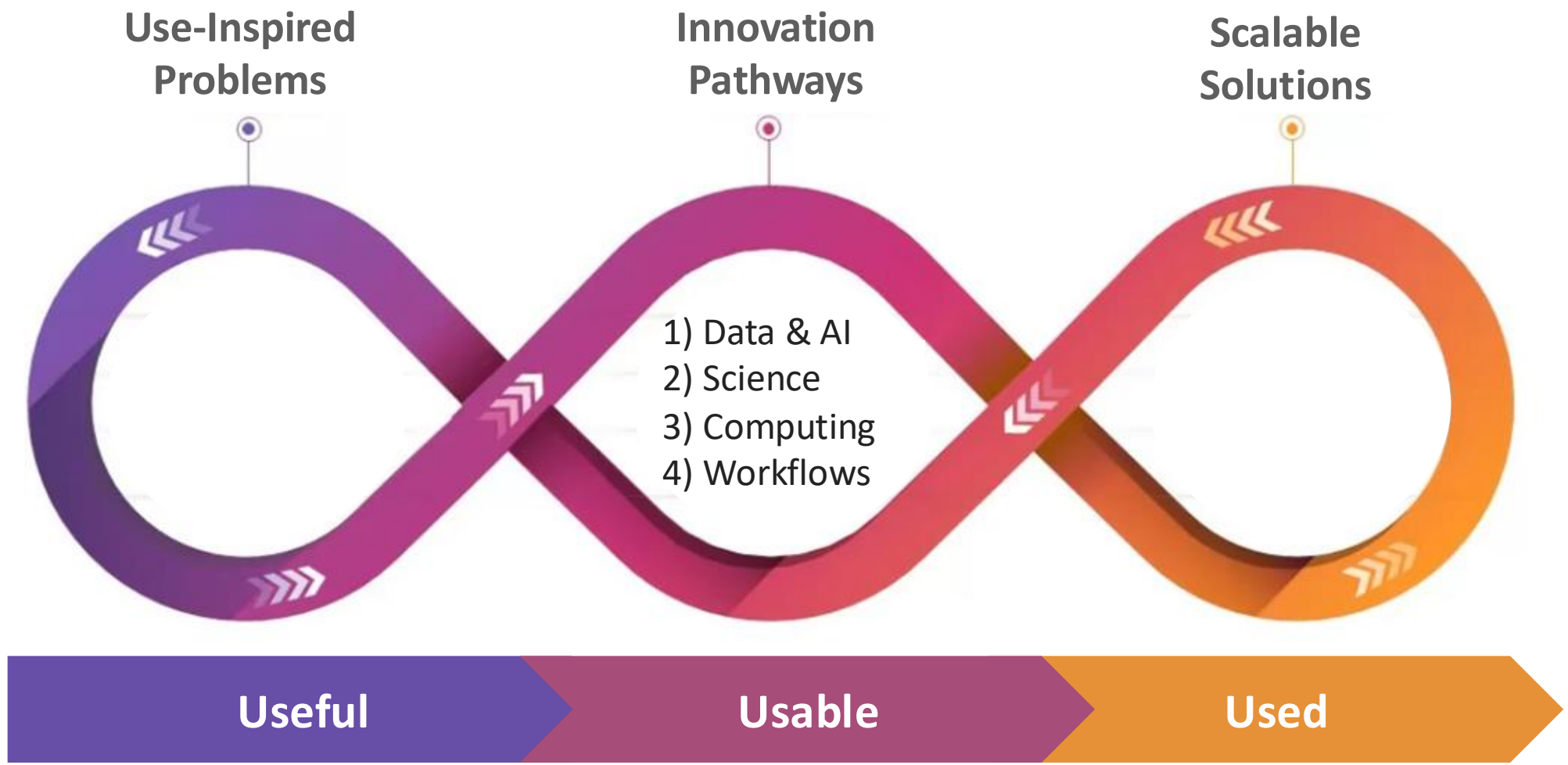
Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

UC San Diego
The University of Utah
The University of Texas at Austin
University of Washington
UMass
Princeton University
Duke University
Loyola Marymount University
University of Oregon
Purdue University
Iowa State University
NCSA
www.sci.utah.edu
eScience Institute
Pittsburgh Supercomputing Center
MGHPCC Massachusetts Green High Performance Computing Center
TACC Texas Advanced Computing Center

Societal Computing and Innovation Lab

Innovation Approach

NATIONAL DATA PLATFORM



8 Co-Design Application Drivers

Domain-agnostic design driven by real scientific needs across disciplines

NATIONAL DATA PLATFORM

Domain	Partner / Institution	Research Focus	NSF Program/Area
Agriculture	B. Ganapathysubramanian, Iowa State U.	Drone sensing, edge AI, crop digital twins	AI Institute
Air Quality	D. Mendoza, University of Utah	Anomaly detection, forecast & messaging agents	NSF OAC
Biomedical	A. Randles, Duke University	Cardiovascular simulation & AI digital twin	NSF DMS FDT-BioTech
Earthquake	D. Moctezuma, University of Oregon	Edge sensor ingestion, simulation on HPC	NSF EAR CST
Linguistics	J. Coleman, Loyola Marymount U.	Indigenous language revitalization via LLM agents	NSF BCS
Hydrology	R. Maxwell, Princeton University	Large-scale hydrologic models + ML surrogates	NSF CSSI / NAIRR Pilot
Material Sci.	T. Sparks, University of Utah	Autonomous materials discovery, multimodal data	NSF DMR
Particle Physics	M. Liu, Purdue University	Edge-HTC AI/ML for extreme-volume collision data	NSF PHY HDR

Design & Production: The Dual-Track Model

NATIONAL DATA PLATFORM

Continuous co-design feeds into stable production services — both tracks run in parallel throughout the 5-year project

DESIGN TRACK

Research Application Co-Design

8 domain partners prototype workflow patterns around real scientific projects. Each yields a generalizable workflow template contributed to the Innovation Factory.

Workforce Development Co-Design

Curriculum and training co-designed with workforce programs, community colleges, and public agencies. Annual hackweeks, data challenges, and credential modules.

Digital Assets Lifecycle Co-Design

Shared practices for ingestion → curation → provenance → validation → deployment → stewardship — developed with NSF facilities and community repositories.

Infrastructure & Services Co-Design

Bidirectional collaboration with CI partners (ACCESS, OSDF, Sage, CloudBank, NAIRR Pilot): NDP integrates existing capabilities and feeds usage patterns and performance data back.

PRODUCTION TRACK

Research Hub

Production-grade, stable access to integrated AI-ready data, compute, and services for researchers.

Education & Training Hub

Curated learning environments with datasets, lab environments, and credential pathways for educators and students.

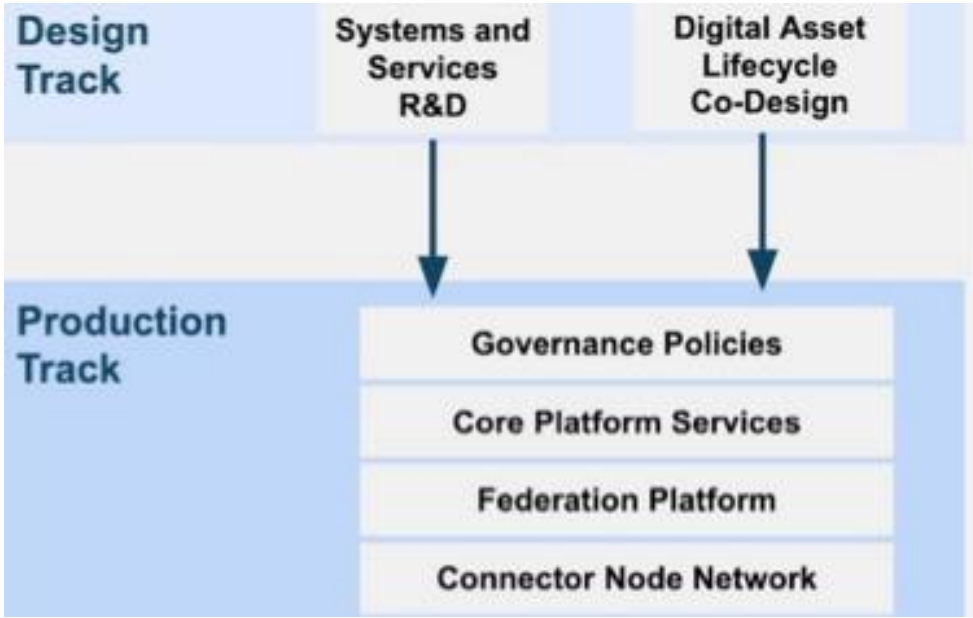
Gallery of Services

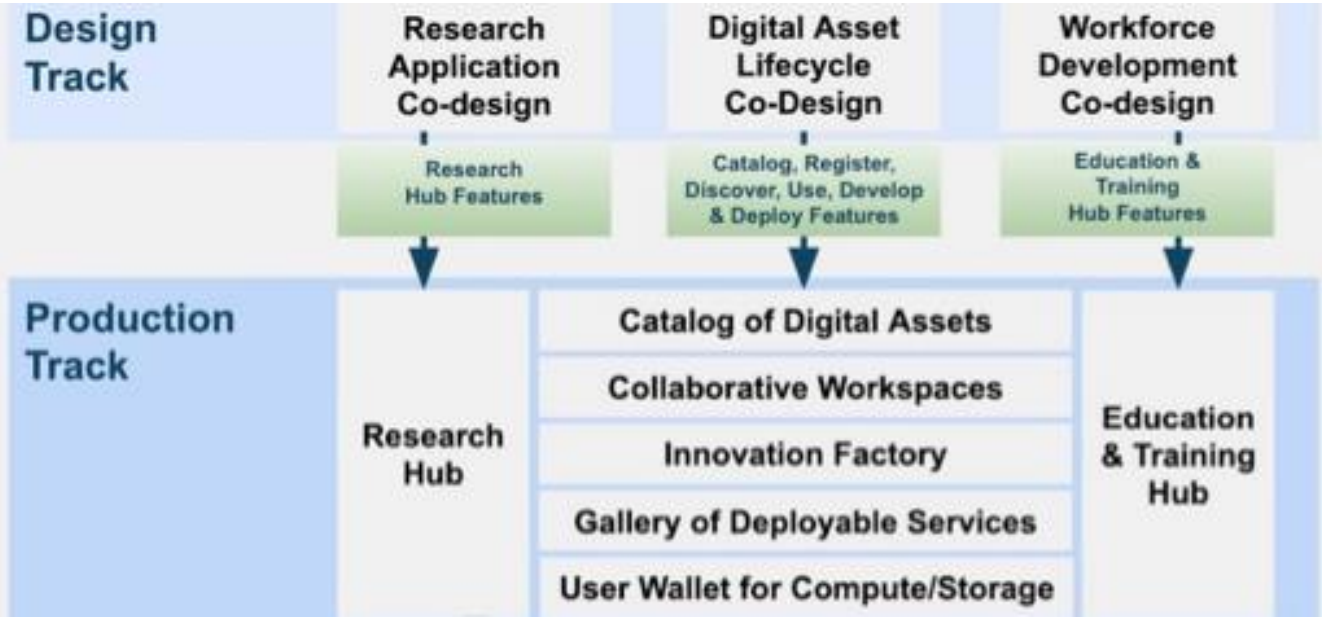
Vetted, benchmarked, production-ready workflows and data/AI services published for community reuse.

Operations & Observability

≥99% uptime · Kubernetes orchestration via NRP · OpenTelemetry monitoring · CI/CD pipelines

NATIONAL DATA PLATFORM





Continuous Design Track Influence on the Production Track

PI: Ilkay Altintas, UC San Diego

Category I: National Data Platform

From Co-Design to Community-Wide Services

How NDP turns domain-specific research into reusable, generalizable workflow patterns

NATIONAL DATA PLATFORM

Domain Prototype

One research application per pattern — prototype co-developed with a single domain partner via direct engagement model.

Working Group

Annual facilitated convening (e.g. Digital Twins WG, Agentic AI WG) with project partners + broader community members with similar applications.

Generalized Template

Reusable, repurposable workflow templates published as Innovation Patterns. Available to all users across disciplines.

Education Integration: A technical education specialist joins each working group to co-develop classroom modules, data challenges, and hackweek curricula alongside the generalized workflow templates.

Guiding Co-Design Principles

Domain-Agnostic

Each pattern designed to generalize beyond its seed application to any domain with similar workflow needs.

Community-Driven

Working groups include non-partner community members — broadening input and accelerating cross-disciplinary adoption.

Reusable by Design

Templates enter the Innovation Factory with standardized SDKs, CI/CD pipelines, and FAIR metadata from day one.

Education-First

Every workflow pattern produces educational materials: classroom modules, workforce training, and open data challenges.

5 AI-Driven Workflow Patterns

NATIONAL DATA PLATFORM

Co-designed application drivers expose shared patterns generalized into reusable NDP services

Agentic AI & Event-Driven Decision Support

Science at the Edge-to-HPC Continuum

Simulation-in-the-Loop Learning & Digital Twins

Autonomous Experiments & Laboratories

Multimodal Data Fusion & Knowledge Graphs

Each pattern is co-developed with 1–2 domain applications, generalized into reusable workflow templates, and published via the Innovation Factory.

NATIONAL DATA PLATFORM

Pattern 01: Agentic AI & Event-Driven Decision Support

A distributed network of autonomous AI agents that monitor, detect, decide, and act based on streaming data — enabling rapid, dynamic, context-aware responses.

Air Quality

D. Mendoza, U. of Utah

Anomaly detection, threshold alerts, forecast generation, and public messaging agents. Sensors on transit buses feed real-time data into agent pipelines.

Linguistics

J. Coleman, Loyola Marymount U.

Indigenous language revitalization via LLM-based translation, data augmentation, and access control agents. Preserves endangered languages at scale.

NDP Impact

- Scalable & reusable for any domain requiring real-time monitoring and human-interpretable AI outputs
- Elevates governance and community-defined controls as first-class capabilities for responsible AI
- Workflow templates for agentic pipelines contributed to the NDP Innovation Factory

Pattern 02: Science at the Edge-to-HPC Continuum

Multi-tier edge–cloud–HPC workflows orchestrated by policy for latency, accuracy, and resource optimization. Data and models flow seamlessly across tiers via federated Kubernetes clusters.

Agriculture

B. Ganapathysubramanian, Iowa State U.

Drone-based sensing and edge AI for crop monitoring. Training on HPC. Demonstrates 'build once, run anywhere' containerized connector node paradigm.

Earthquake

D. Moctezuma, U. of Oregon

Rapid ingestion of seismic sensor data at the edge, simulation on HPC for Cascadia subduction zone response workflows.

NDP Impact

- Composable, location-aware workflow templates usable across scientific domains
- Specialized NDP Connector Node containers: 'build once, run anywhere' paradigm
- Reduces data movement and latency for time-critical science applications

Pattern 03: Simulation-in-the-Loop Learning & Digital Twins

Hybrid workflows integrating physics-based models, simulation outputs, and ML surrogates to support scientific discovery and decision-making. AI learns from simulation; simulations adapt to AI predictions.

Hydrology

R. Maxwell, Princeton U.

Large-scale hydrologic models (ParFlow) augmented by ML surrogates. Rapid subsurface water flow prediction at continental scales.

Wildfire

PI Altintas, UC San Diego

Fire response digital twin integrating real-time sensor feeds, physics fire models, and ML prediction for emergency decision support.

Biomedical

A. Randles, Duke U.

Cardiovascular simulation coupled with an AI digital twin for patient-specific hemodynamics.

NDP Impact

- Workflow patterns for large-scale simulation enabling AI and digital twins
- Educational materials for teaching physics-informed ML and AI for physical sciences
- Demonstrated at national scale via NDP Pilot, now productionized

Pattern 04: Autonomous Experiments & Laboratories

NATIONAL DATA PLATFORM

Continuous experimental cycles integrating streaming data from instruments, AI-guided parameterization and hypothesis generation, and human-in-the-loop autonomy.

Material Science

T. Sparks, U. of Utah

Autonomous materials discovery using multimodal data and Bayesian optimization. AI suggests next experiment; instruments execute; results feed back.

Particle Physics

M. Liu, Purdue U.

Edge-HTC AI/ML for particle collision data of extreme volume. Rapid filtering and classification at the edge before archiving.

NDP Impact

- Domain-specific AIs that accelerate clinically relevant materials innovation
- Patterns for labs seeking to connect physical instruments to national CI
- Benchmarking and educational materials for AI in materials and physics domains

Pattern 05: Multimodal Data Fusion & Knowledge Graphs (carried over from Pilot)

Harmonization of diverse data types (spatial, temporal, sensor, and text) through AI-based fusion, semantic enrichment, and graph-based discovery. Foundation for all NDP data and AI services.

Agriculture

PI Altintas Lab, UCSD

Multimodal data fusion across drone imagery, soil sensors, weather, and satellite data for precision agriculture and crop stress detection.

Wildfire

PI Altintas Lab, UCSD

Edge sensor-driven experiments fusing fire weather, fuel maps, topography, and real-time video for comprehensive situational awareness.

NATIONAL DATA PLATFORM

Collaborations and Partnerships

Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

Connecting & Amplifying Existing National CI

NATIONAL DATA PLATFORM

NDP as the coordination layer — not duplicating, but federating

NSF Computing CI

- ACCESS · LCCF · NAIRR Pilot
- OSDF (Open Storage Data Fabric)
- CC* (Campus Cyberinfrastructure)
- Sage (CI for AI at the Edge)

Foundation Platforms

- NDP base platform
- Tapis — HPC/HTC API layer
- National Research Platform (NRP)
- CloudBank

Science Programs

- NSF AI Institutes
- CSSI, SCIPE, CICI projects
- AI & Cloud Testbeds
- Domain workflows from NSF instruments

Data Repositories

- NASA & NOAA data repositories
- Domain data commons
- Community data repositories
- External agency open datasets

Education Networks

- CENIC, Edge regional networks
- ADSA workforce training
- Community colleges & public agencies
- Regional workforce development offices

NDP standardizes APIs, metadata schemas and federated identity across all CI toward an AI-ready national fabric.

NATIONAL DATA PLATFORM

Production Offerings

Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

UC San Diego
THE UNIVERSITY of UTAH
TEXAS The University of Texas at Austin
UNIVERSITY of WASHINGTON
UMASS
PRINCETON UNIVERSITY
Duke UNIVERSITY
Loyola Marymount University
UNIVERSITY OF OREGON
PURDUE UNIVERSITY
IOWA STATE UNIVERSITY
NCSA
SCI www.sci.utah.edu
eScience Institute
PITTSBURGH SUPERCOMPUTING CENTER
MGHPCC MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER
TACC TEXAS ADVANCED COMPUTING CENTER

Production Offerings: Research Hub & Education Hub

NATIONAL DATA PLATFORM

Production-grade, fully supported platforms for science and workforce development

Research Hub

- Stable, supported access to integrated AI-ready data, compute, and services
- Unified interfaces for discovery, access, and execution of multimodal workflows
- Run end-to-end: data ingestion → preprocessing → model training → simulation → analysis
- No need to manage underlying HPC, cloud, or edge complexity
- Built-in lifecycle tools, shared workspaces, collaboration studios, gallery of services, innovation factory
- ≥ 99% uptime · < 10s end-to-end response · 1000 concurrent sessions target

Education & Training Hub

- Production-grade platform for the next-generation AI- and data-enabled workforce
- Curated datasets, sample workflows, hands-on lab environments
- Course modules, tutorials, asynchronous training, and hackathon infrastructure
- Annual hackweeks — active learning, peer mentoring, community development
- Co-designed with NSF workforce programs, community colleges, public agencies
- Year 5 target: 2,000 training modules · 4,000 completed certifications

Performance Objectives & Year-5 Targets

Measurable goals that ensure NDP delivers at national scale

NATIONAL DATA PLATFORM

Asset Registration
10,000 indexed assets · 500 cross-repo users · 50 knowledge graphs · 3s query response

Workflow Reuse
20 reused patterns · 300 new workflows · 1-day prototype to production transition

Federation Reach
60 federated connector nodes · 10 interoperable institutional partners · All U.S. states

User Impact
90% satisfaction · 1-day data discovery to workflow completion · 6,000 unique portal users

Service Utilization
600 deployable AI & data services · 20 deployments/service · 5,000 workflow executions · 95% uptime

Education & Training
2,000 training modules · 4,000 completed trainings/certificates · 20-30% annual contributor growth

Innovation Factory
1,000 services/workflows transitioned to production · 50% adoption rate of new NDP services

User Support
≤24h ticket response · 90% resolved in 2 business days · 90% support satisfaction

User Access Model & Resource Policy

NATIONAL DATA PLATFORM

Open discovery to full allocation — a tiered, transparent access framework

Open Discovery	Anyone can freely explore public catalog metadata, documentation, and training resources. No account needed.
Registered Access	Researchers, educators, and students obtain access via federated identity. Create workspaces, collaborate, integrate activities.
Transient Resource Access	Deep, sustained use of compute, storage, or connector nodes. Lightweight allocation via quarterly Governance & Stewardship Council review.
Bring Your Own Resources	Link external storage, compute, or cloud allocations via NRP/Nautilus. Local data stays local; federation adds national reach.

All access tiers are governed by the NDP Policy Fabric · Open science principles · NSF cooperative agreement requirements

NATIONAL DATA PLATFORM

Project Timeline and Work Packages

Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

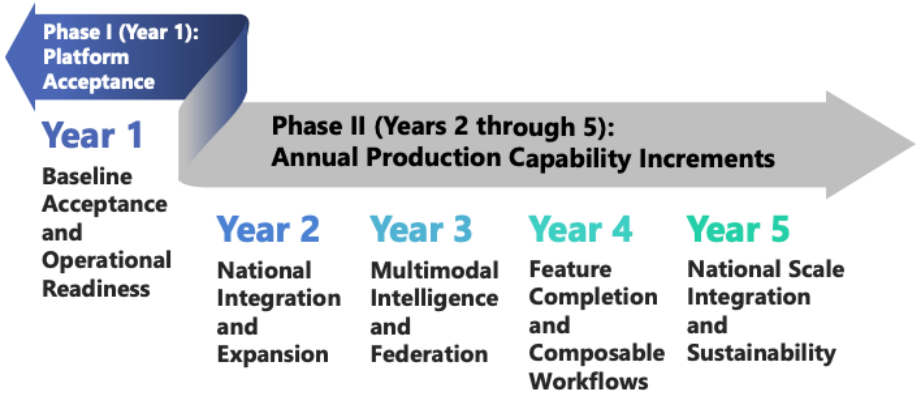
PI: Ilkay Altintas, UC San Diego

Category I: National Data Platform

Project Timeline: 5-Year Execution Plan

NATIONAL DATA PLATFORM

Phase I establishes foundations; Phase II delivers production capability increments annually



Annual Activities

- Co-Design Workshops
- Hackweeks
- Interoperability Advisory Board Meetings
- External Evaluation

Monthly Activities

- Community Webinars
- Online Tutorial Releases
- Science and Education Advisor Touch Bases
- Monthly User Stories

PHASE I (Year 1)

PHASE II (Years 2–5) — Annual Production Capability Increments

Year 1
Foundation

- Deploy federation stack and core platform services: Catalog, auth, workspace services
- Head nodes + 15 (6 NDP-owned) connector nodes
- Co-design partnerships launched
- MVP acceptance testing

Year 2
Expansion

- 28 (3 NDP-owned) connector nodes
- Research Hub enters production
- Education Hub v2 launched
- First workflow patterns published
- External user onboarding updated process

Year 3
Integration

- Annual technology refresh cycle defined
- Agentic data services deployed
- 10+ institutional integrations
- Hackweek programs expanded
- Performance benchmarking enters production

Year 4
Scale

- 10+ connector nodes
- Innovation Factory at capacity
- Education Hub curriculum growth
- Governance engine policy-as-code enters production
- National open data challenges launched

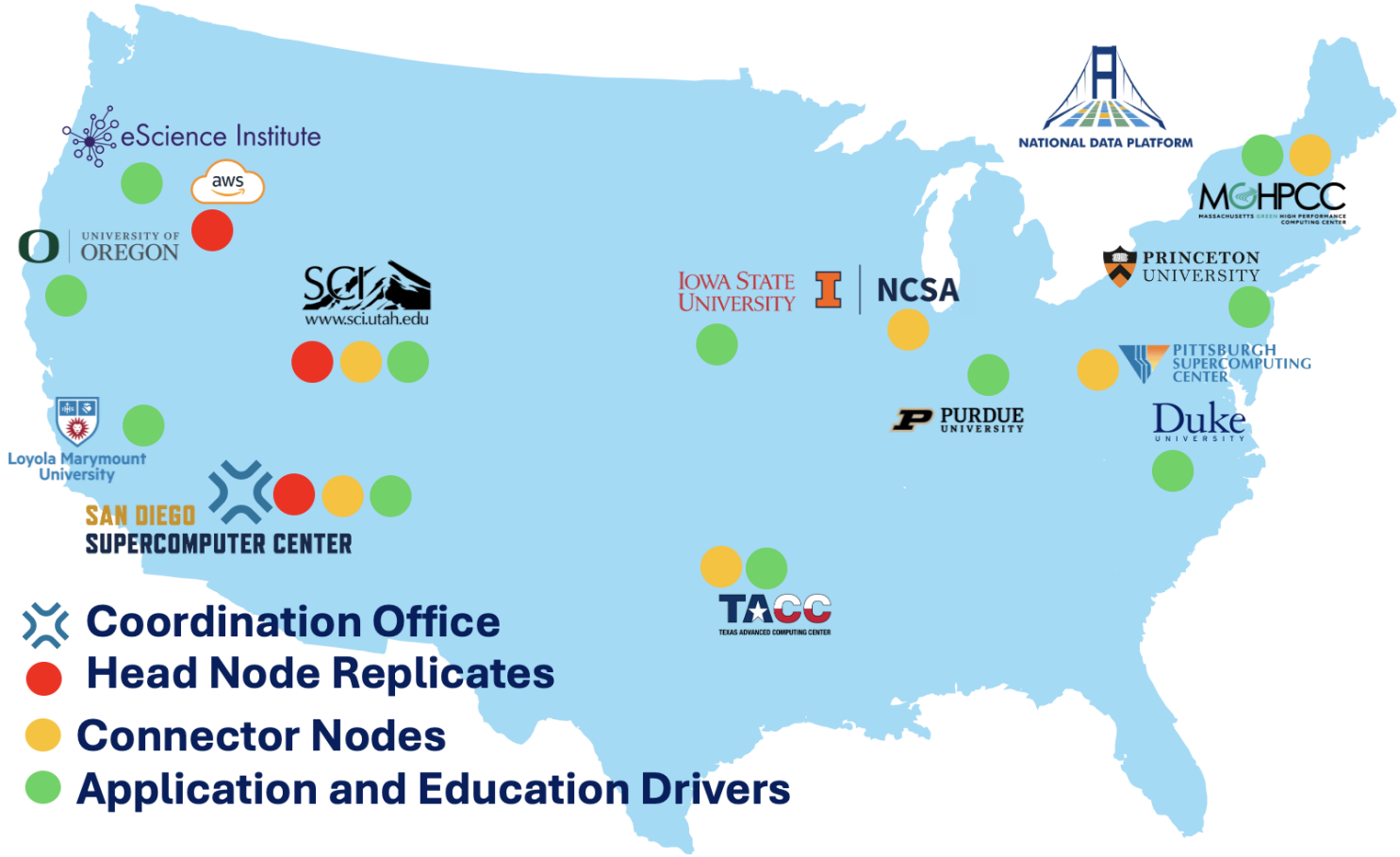
Year 5
Sustainability

- All Year-5 targets achieved
- Sustainability roadmap finalized
- 60+ connector nodes federated
- External evaluator final report
- Next-generation NDP designed

Partner Institutions

A multidisciplinary team with decades of experience operating national-scale CI

NDP Collaborators and Nodes



Ilkay Altintas (PI)	SDSC, UC San Diego
Anthony Arendt	University of Washington
Ashley Atkins	SDSC, UC San Diego
Penny Atkins	SCI, University of Utah
Rohan Basuroy	University of Utah
Paola Buitrago	PSC, Carnegie Mellon University
Jared Coleman	Loyola Marymount University
Lester Crawl	UC San Diego
Maytal Dahan (Co-PI)	TACC, The University of Texas at Austin
Melissa Floca	SDSC, UC San Diego
Baskar Ganapathysubramanian	Iowa State University
Amarnath Gupta	SDSC, UC San Diego
Rafael Coelho Lopes de Sá (Co-PI)	MGHPCC, University of Massachusetts Amherst
Mia Liu	Purdue University
John Daniel Maloney	NCSA, University of Illinois Urbana-Champaign
Reed M. Maxwell	Princeton University
Diego Melgar Moctezuma	University of Oregon
William L. (Bill) Miller	CHPC, University of Utah
Daniel Mendoza	University of Utah
Viswanath Nandigam	SDSC, UC San Diego
Manish Parashar (Co-PI)	SCI, University of Utah
Amanda Randles	Duke University
Scott Sparks	University of Utah
Sarah Stone (Co-PI)	University of Washington
Frank K. Wuerthwein	SDSC, UC San Diego

Summary: What NDP Delivers

NATIONAL DATA PLATFORM

A national-scale AI-ready ecosystem ready for the next decade of science

1	**A Complete Ecosystem**	Not just another HPC center or data repository — NDP is the integration layer that connects them all into a composable, AI-ready whole.
2	**5 Generalizable Workflow Patterns**	From agentic AI to digital twins to autonomous labs — patterns extracted from 8 real co-design domains, made reusable for any discipline.
3	**Production-Grade from Day 1**	Research Hub and Education Hub with ≥99% uptime targets, built on proven NDP, Tapis, NRP, and CloudBank platforms with 3+ years of operational track record.
4	**Workforce Ready**	Education Hub, annual hackweeks, data challenges, and curriculum modules ensure the next generation is AI- and data-fluent.
5	**Open, FAIR, and Governable**	Open source, FAIR-by-design, community governance, ethical AI principles for public-interest science.

nationaldataplatform.org

Category I: National Data Platform

Thank you!

Presenter: Ilkay Altintas, ialtintas@ucsd.edu

NSF IDSS · #2609447

NATIONAL DATA PLATFORM

UC San Diego · The University of Utah · The University of Texas at Austin

University of Washington · UMass · Princeton University · Duke University

Loyola Marymount University · University of Oregon · Purdue University · Iowa State University

San Diego Supercomputer Center · SCI www.sci.utah.edu · eScience Institute

MGHPCC Massachusetts Green High Performance Computing Center · TACC Texas Advanced Computing Center

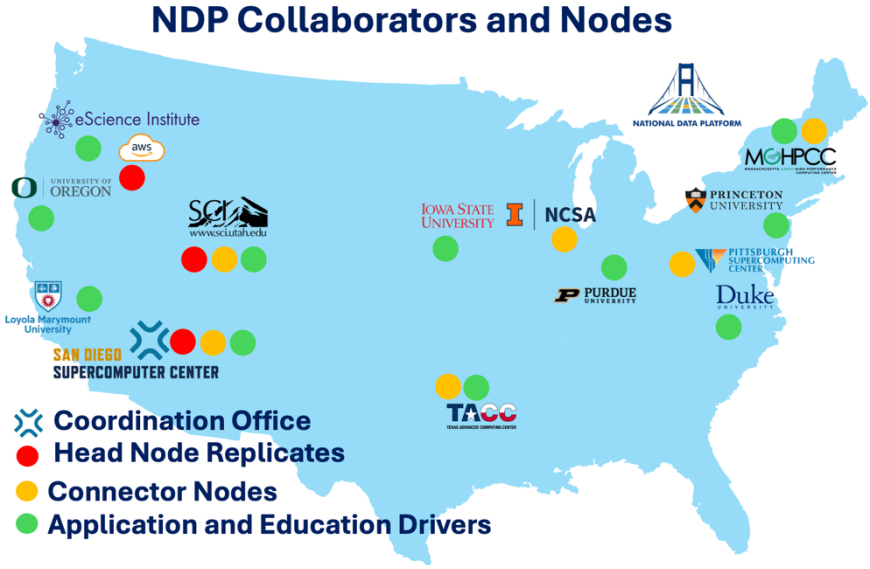
Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

Pittsburgh Supercomputing Center · NCSA

NATIONAL DATA PLATFORM

Category I: National Data Platform

A Federated National AI-ready Data Ecosystem for Discovery, Innovation, Education, and Workforce Development



Federated Data Access

Seamlessly bridge distributed national data repositories and HPC and cloud resources.

AI-Ready Infrastructure

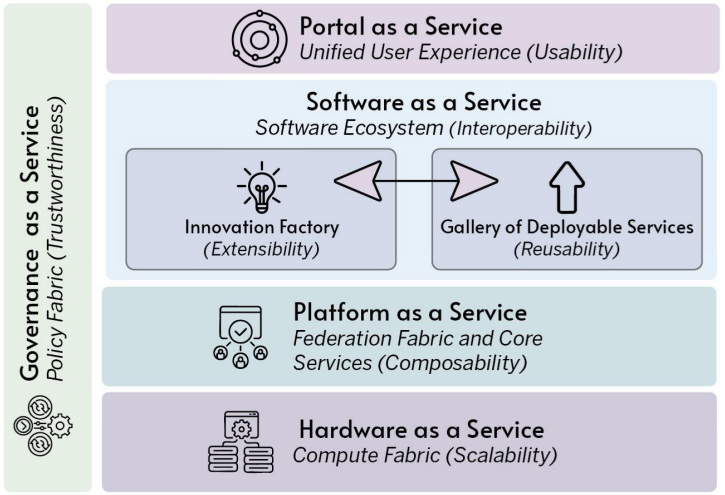
Deploy multimodal AI services close to data across edge, cloud, and HPC environments.

Innovation Factory

Build, test, validate, and deploy composable scientific workflows and AI tools.

Workforce Development

Hands-on learning, curriculum, and training for researchers and students nationwide.



A user-centered architecture, backed by scalable hardware, a composable federation platform, and an interoperable, extensible software ecosystem.

Agentic AI & Event-Driven Decision Support

Air quality agents · Language revitalization · Fusion energy

Edge-to-HPC Continuum Science at the Edge

Agriculture drones · Earthquake rapid response · Edge AI

Simulation-in-the-Loop Learning & Digital Twins

Hydrology · Cardiovascular health · Wildfire

Autonomous Experiments & Laboratories

Materials discovery · Particle physics · Drug design

Multimodal Data Fusion & Knowledge Graphs

Spatial · temporal · sensor · image · text integration

UC San Diego

THE UNIVERSITY of UTAH

TEXAS The University of Texas at Austin

UNIVERSITY of WASHINGTON

UMass

NSF # 2609447

Ilkay Altintas (PI), UC San Diego
Manish Parashar, University of Utah
Maytal Dahan, University of Texas
Sarah Stone, University of Washington
Rafael Coelho Lopes de Sa, UMass

https://www.nationaldataplatform.org