National Institutes of Health
Office of Data Science Strategy

ODSS Data Infrastructure and Cloud Programs Overview

Laura Biven, Ph.D.
Lead, Integrated Infrastructure and Emerging Technologies
Office of Data Science Strategy
National Institutes of Health

https://datascience.nih.gov/

1/17/2024

Agenda

- **ODSS Data Infrastructure & Cloud Programs**
  - STRIDES - *NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative*
  - Other NIH cloud-based Workspaces
  - Cloud Lab
  - RAS - *Researcher Auth Service Initiative*
  - NCPI – *NIH Cloud Platform Interoperability Program*
- **How to participate in ODSS data infrastructure and cloud programs?**
- **Cloud Supplement Programs**
  - HVD – *High-Value Datasets Program*
  - Cloud supplement NOSI

The NIH STRIDES Initiative

STRIDES: Science & Technology Research Infrastructure for Discovery, Experimentation, & Sustainability

Overview

Serving **both the NIH intramural and extramural research communities**, the STRIDES Initiative accelerates biomedical research in the cloud by:

- Simplifying access
- Reducing costs
- Lowering technological barriers
- Standardizing administrative & financial processes

Core Motivations

1. **Democratization of computational research & data science**
   Leveling the playing field for those traditionally underrepresented in biomedical research
2. **Cost savings & efficiencies for the research community**
   More usage begets more savings and greater overall discounts for all
3. **Strong partnerships with cloud providers**
   Resulting in collaborative R&D engagements and more direct focus and support on research

Partnerships with: aws Google Cloud Microsoft Azure

STRIDES OVERVIEW

Value to Participants

STRIDES participants benefit from a variety of exclusive features, from competitive pricing to training expertise.

- **Competitive** pricing & financial benefits
- **Professional** service consultations
- **Flexible** business model
- **Expanded** communication reach
- **Expert** support from cloud providers
- **Reach-through** to additional partners
- **Training** expertise and scaling capacity

Impact to Date*

247+	PETABYTES OF DATA
491M+	COMPUTE HOURS
1,650+	RESEARCH PROGRAMS
$72M+	COST SAVINGS
5350+	PEOPLE TRAINED

*as of August 31, 2023

Major NIH & NIH-Funded Research Programs Supported

All of Us
RESEARCH PROGRAM

FRAMINGHAM HEART STUDY
Three Generations of Dedication
1948 1971 2002
A Project of the National Heart, Lung, and Blood Institute and Boston University

National Ophthalmic Disease Genotyping Network
eyeGENE®
A program of the National Institutes of Health

GTEx Portal

PRIMED
consortium

HuBMAP
Human BioMolecular Atlas Program

AMP PD

HEAL
NIH · Helping to End Addiction Long-term

NHLBI
TOPMed

Metabolomics Workbench

Common Fund Data Ecosystem

UniProt

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

SPARC

AnVIL

MoTrPAC
The Molecular Transducers of Physical Activity Consortium

HMP

DK net

eRA

The Down Syndrome Consortium

Adolescent Brain Cognitive Development

Glycoscience

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM

NIMH Data Archive
NDA | ABCD | CCF | NIAAA$_{DA}$ | OAI

NATIONAL CANCER INSTITUTE
GENOMIC DATA COMMONS
Access the Data
#NCIGDC

NIA IMPACT
COLLABORATORY
TRANSFORMING DEMENTIA CARE

IDG
ILLUMINATING the DRUGGABLE GENOME

SenNet

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®,

NIH Center for Information Technology

Galaxy
PROJECT

Example NIH cloud-based Workspaces

- CRDC - Cancer Research Data Commons
- AnVIL - NHGRI Analysis Visualization and Informatics Lab-space
- BioData Catalyst
- CFDE – Common Fund Data Ecosystem
- HEAL Initiative
- AoU Workbench
- ScHARe - Science Collaborative for Health disparities and Artificial intelligence bias REduction

CLOUD TRAINING AT NIH

NIH Cloud Lab: Experiment in the Cloud

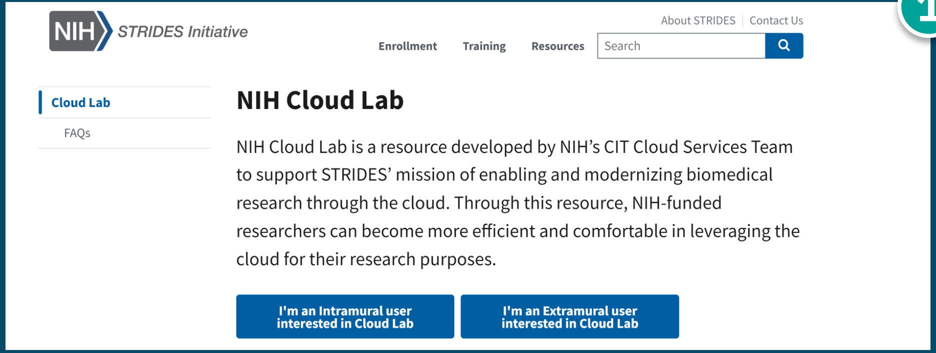
NIH Cloud Lab is a no-cost, 90-day program for NIH intra- and extramural researchers to try commercial cloud services in an NIH-approved environment. Cloud Lab provides training and guardrails to protect against financial and security risks.

How It Works

1. **Fill out** interest form
2. **Get** account and $500 of credits
3. **Access** tailored cloud trainings
4. **Practice and learn** for 90 days

NIH Cloud Lab Sign Up Page

1


NIH STRIDES Initiative
About STRIDES | Contact Us
Enrollment
Training
Resources
Search
Cloud Lab
FAQs

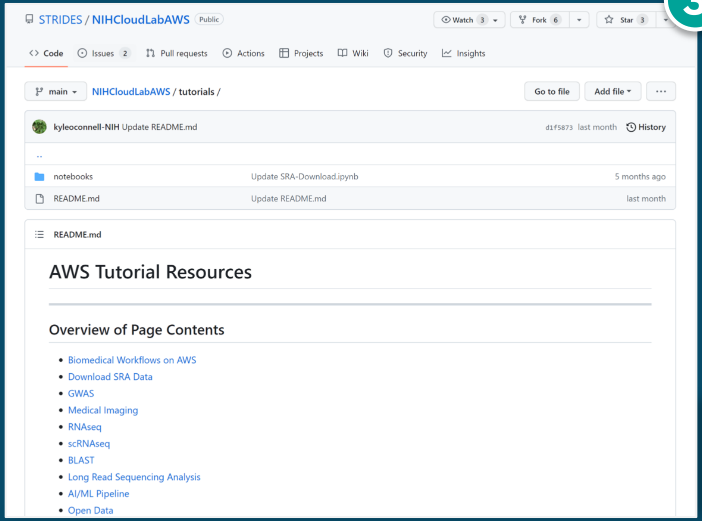
NIH Cloud Lab

NIH Cloud Lab is a resource developed by NIH's CIT Cloud Services Team to support STRIDES' mission of enabling and modernizing biomedical research through the cloud. Through this resource, NIH-funded researchers can become more efficient and comfortable in leveraging the cloud for their research purposes.

I'm an Intramural user interested in Cloud Lab
I'm an Extramural user interested in Cloud Lab

NIH Cloud Lab AWS Tutorial Repository

3



NIHCloudLabAWS / tutorials /

kyleoconnell-NIH Update README.md d1f5873 last month

..

notebooks	Update SRA-Download.ipynb	5 months ago
README.md	Update README.md	last month

README.md

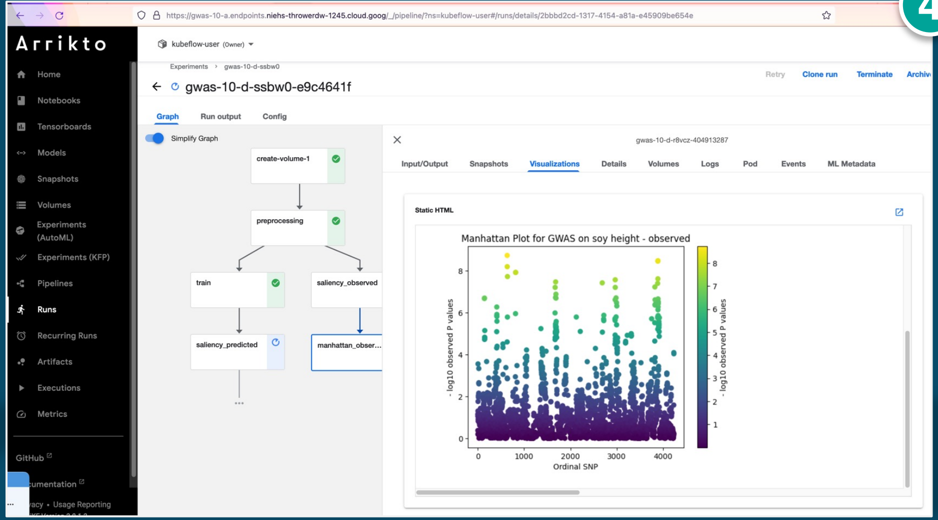
AWS Tutorial Resources

Overview of Page Contents

- Biomedical Workflows on AWS
- Download SRA Data
- GWAS
- Medical Imaging
- RNAseq
- scRNAseq
- BLAST
- Long Read Sequencing Analysis
- AI/ML Pipeline
- Open Data

Example of NIH Cloud Lab Use Case

4



https://gwas-10-a.endpoints.niehs-throwerdw-1245.cloud.goog/_/pipeline/?ns=kubeflow-user#/runs/details/2bbbd2cd-1317-4154-a81a-e45909be654e

gwas-10-d-ssbw0-e9c4641f

NIH Use Cases

Evaluate Utility & Cost

Provides an easy route to evaluate the cloud's utility/cost for a project without major time or financial commitments

Develop New Tools

Allows experienced teams to prototype new architectures and evaluate software and hardware combinations

Share Ideas

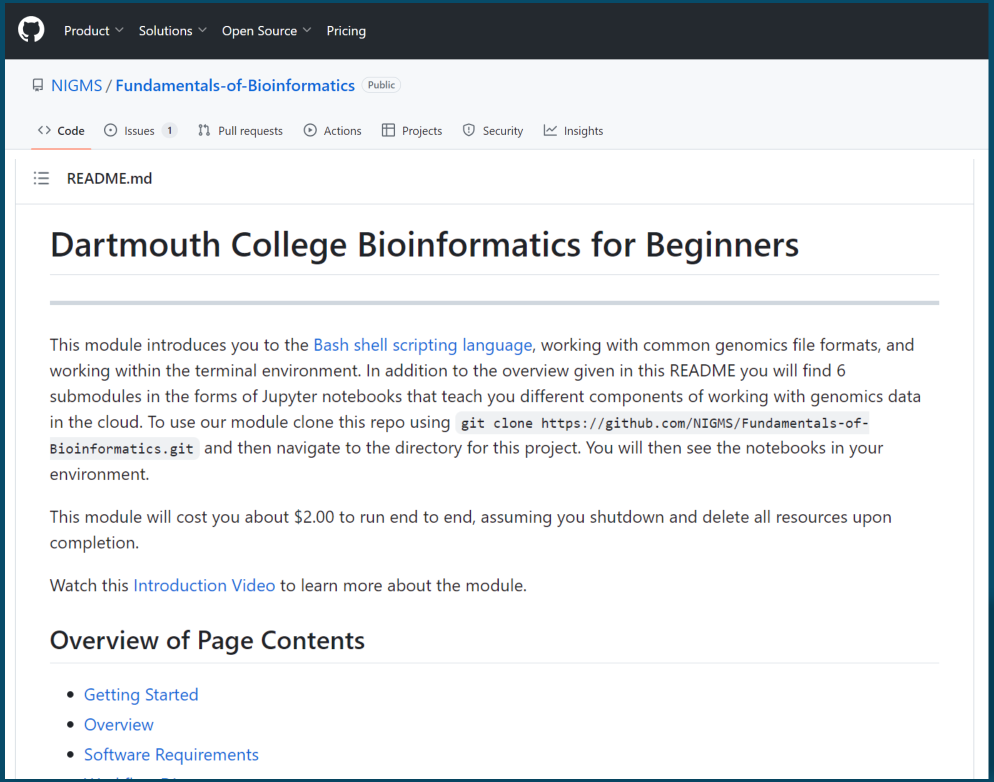
Connects NIH'ers from across ICs to share ideas on how to conduct biomedical research in the cloud

Learn New Skills

Simplifies access to tools and cloud environments that participants can use for training purposes

NIH Cloud Lab Tutorials

NIH Cloud Lab provides GitHub repositories with general resources on Amazon Web Services , Google Cloud, and Microsoft Azure. The program also provides a GitHub repository with twelve interactive, cloud-based learning modules created with funding from the National Institute of General Medical Sciences. Accompanying videos are available on YouTube.



Dartmouth College Bioinformatics for Beginners

This module introduces you to the Bash shell scripting language, working with common genomics file formats, and working within the terminal environment. In addition to the overview given in this README you will find 6 submodules in the forms of Jupyter notebooks that teach you different components of working with genomics data in the cloud. To use our module clone this repo using `git clone https://github.com/NIGMS/Fundamentals-of-Bioinformatics.git` and then navigate to the directory for this project. You will then see the notebooks in your environment.

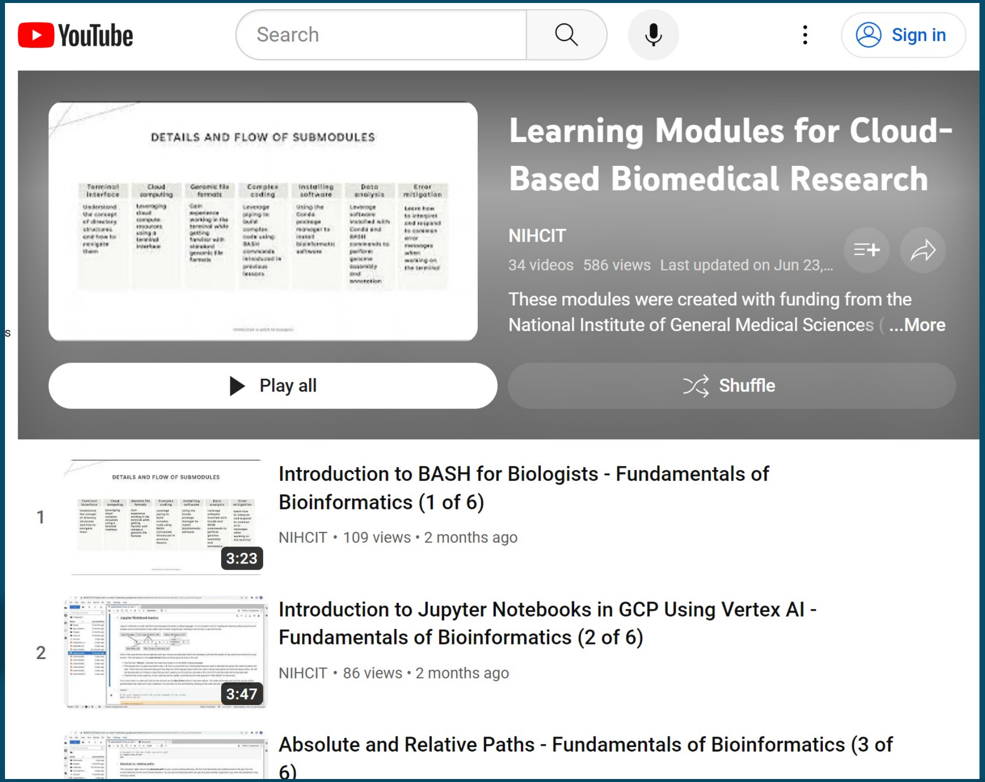
This module will cost you about $2.00 to run end to end, assuming you shutdown and delete all resources upon completion.

Watch this Introduction Video to learn more about the module.

Overview of Page Contents

- Getting Started
- Overview
- Software Requirements

Interactive Cloud-Based Learning Module in GitHub



Learning Modules for Cloud-Based Biomedical Research

NIHCIT

34 videos 586 views Last updated on Jun 23,...

These modules were created with funding from the National Institute of General Medical Sciences (

1 Introduction to BASH for Biologists - Fundamentals of Bioinformatics (1 of 6)
3:23
NIHCIT • 109 views • 2 months ago

2 Introduction to Jupyter Notebooks in GCP Using Vertex AI - Fundamentals of Bioinformatics (2 of 6)
3:47
NIHCIT • 86 views • 2 months ago

Absolute and Relative Paths - Fundamentals of Bioinformatics (3 of 6)

Playlist of Module Videos on YouTube

NIH Researcher Auth Service (RAS)

CHALLENGES

Internal NIH and external researchers must **maintain separate accounts to access the same dataset across different platforms** resources and are required to **sign in multiple times**

Complex data ecosystems: Search portals, data repositories, and platforms manage their own identity and access management (“auth”) software.
Authentication info does not travel with researchers moving between platforms

No standard protocol for describing authorization info; authorizations from NIH dbGaP Data Access Committee (DAC) decisions replicated in multiple disconnected data repositories

Access to controlled-access data via **username and password represents security risk**

Tracking efforts are disjointed as auditing and logging is not standardized across data repositories

SOLUTIONS TO SUPPORT SCIENCE

Simplified process enables a researcher to **log in once** or link accounts securely using preferred credentials from multiple identity providers

Delegated responsibility to NIH: Only NIH RAS tokens can be used check the user’s identity before a system or data access event to NIH-funded/controlled data and tools (NIH makes decision). **Multi-Factor authentication (MFA) is standard**

Authorizations from NIH dbGaP Data Access Committee (DAC) decisions are centralized and provisioned only upon login

A Zero Trust security approach among partners ensures that repositories adhere to important security controls in the Interconnection Security Agreement (ISA). **Multifactor authentication (MFA)** for higher level of access security

Consistent auditing and logging of events throughout the process

NIH Cloud Platform Interoperability Program

- The NCPI program is a **partnership** between multiple NIH-supported participating systems (currently AnVIL, BioData Catalyst, CRDC, dbGaP, and Kids First) developing and implementing technical standards to enable interoperability and facilitate a federated data ecosystem.

- The goal of NCPI is to enable a **federated data ecosystem** that will facilitate **researcher-driven analyses** of datasets across multiple NIH cloud-based platforms and repositories.

- This will be accomplished through testing and implementing **standards and approaches for systems interoperability** and universal authentication & authorization

Currently Participating Platforms

AnVIL

The Analysis, Visualization, and Informatics Lab-space (AnVIL) is the National Human Genome Research Institute’s genomic data resource that leverages a cloud-based infrastructure for democratizing genomic data access, sharing and computing across large genomic, and genomic-related data sets.

BioData Catalyst

NHLBI BioData Catalyst, supported by the National Heart, Lung, and Blood Institute (NHLBI), is a cloud-based platform providing tools, applications, and workflows in secure workspaces. By increasing access to NHLBI datasets and innovative data analysis capabilities, BioData Catalyst accelerates efficient biomedical research that drives discovery and scientific advancement, leading to novel diagnostic tools, therapeutics, and prevention strategies for heart, lung, blood, and sleep disorders.

Cancer Research Data Commons

The goal of the National Cancer Institute's Cancer Research Data Commons (CRDC) is to empower researchers to accelerate data-driven scientific discovery by connecting diverse datasets with analytical tools in the cloud. The CRDC is built upon an expandable data science infrastructure that provides secure access to many different data across scientific domains via Data Commons Framework

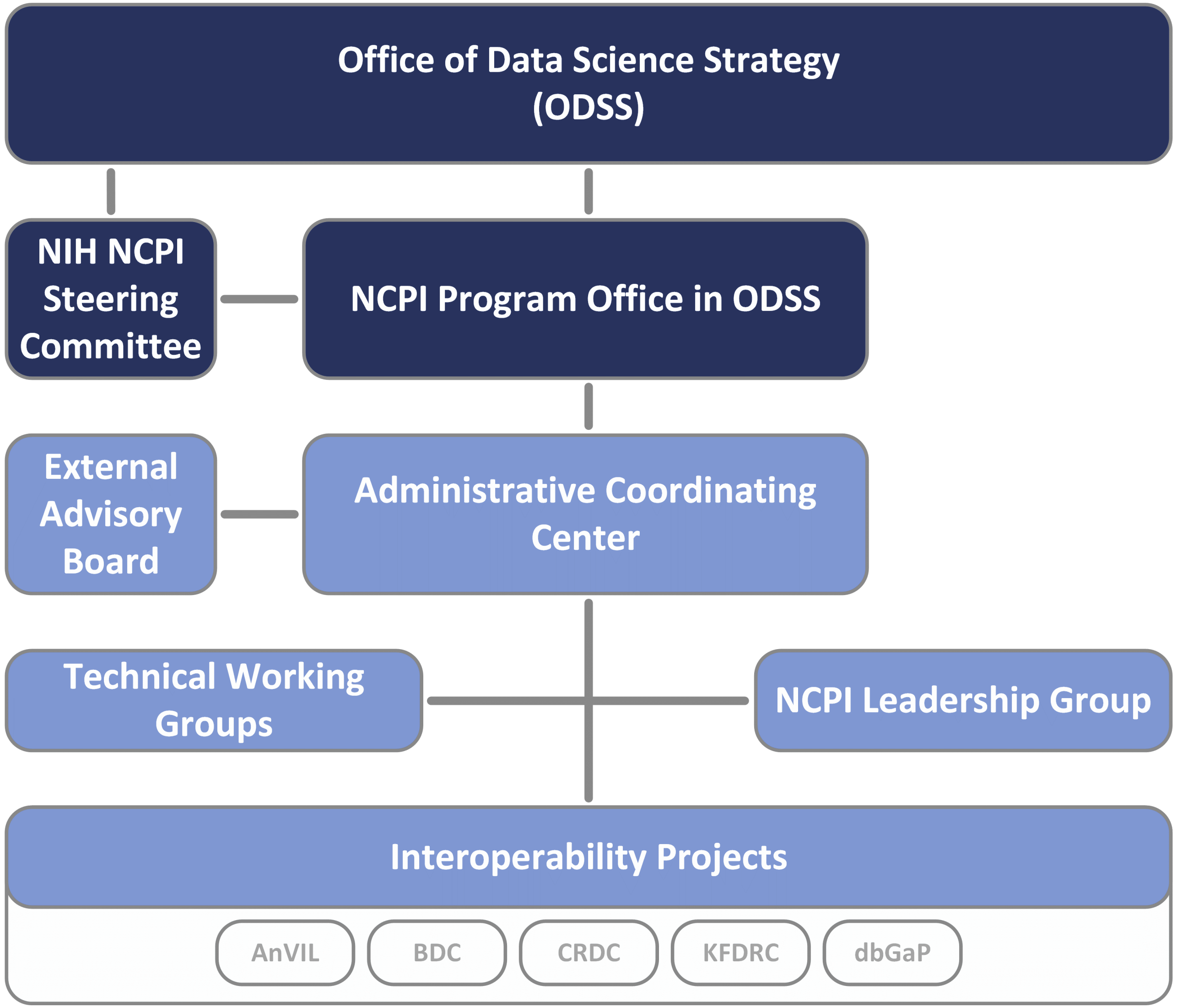
Kids First Data Resource Center

The NIH Common Fund’s Gabriella Miller Kids First Pediatric Research Program’s (“Kids First”) vision is to “alleviate suffering from childhood cancer and structural birth defects by fostering collaborative research to uncover the etiology of these diseases and by supporting data sharing within the pediatric research community.”

National Center for Biotechnology Information

The National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) hosts and manages the Database of Genotypes and Phenotypes (dbGaP) and NIH’s Sequence Read Archive (SRA). dbGap provides and manages access to protected data related to human studies that have investigated the interaction of genotype and phenotype. SRA is the largest archive for public controlled-access next-generation sequencing data.

The NCPI Community



Office of Data Science Strategy (ODSS)

NIH NCPI Steering Committee

NCPI Program Office in ODSS

External Advisory Board

Administrative Coordinating Center

Technical Working Groups

NCPI Leadership Group

Interoperability Projects

NCPI Administrative Coordinating Center (ACC)

The NCPI is supported by Administrative Coordinating Center (ACC). The ACC provides technical, administrative, coordination, and project management support for the five primary task areas outlined below:

- Establish, facilitate, and monitor the technical implementation of interoperability projects
- Provide project management and coordination for NCPI partners and collaborators
- Support training, outreach, and other community building activities
- Support NCPI's governance structure and coordinate working groups
- Support, facilitate, and monitor adoption of RAS by NCPI

Interoperability Technologies

- **Researcher Auth Service (RAS)** is an effort by the NIH’s Center for Information Technology (CIT) to provide a common mechanism by which researchers can establish their identity and access data they are authorized to use across NCPI systems. The RAS API allows seamless access to researchers for integrated data repositories.

- **The Global Alliance for Genomics and Health (GA4GH) Data Repository Service (DRS)** provides generic interface to data repositories so data consumers, including workflow systems, can access data objects in a single, standard way regardless of where they are stored and how they are managed.

- **Fast Healthcare Interoperability Resources (FHIR)** is a standard describing data formats and elements (known as “resources”) and an API for exchanging electronic health records (EHR). One of its goals is to facilitate interoperation between legacy health care systems, to make it easy to provide health care information to health care providers and individuals on a wide variety of devices.

- **The Portable Format in Bioinformatics (PFB)** is an Avro-based file format that bundles schema, data, ontologies/controlled variables, and pointers to data files in a single, serializable format that can be sent easily across systems and has the flexibility for different data models.

- **The Workflow Execution Service (WES)** is an API developed by the GA4GH Cloud Work Stream that describes a standard protocol for running the same genomic data analysis in different environments and still obtaining the same results. The WES API is part of a larger framework to seamlessly bring algorithms to genomic data.

How to participate in ODSS data infrastructure and cloud programs?

- STRIDES
  - Please send an email to STRIDES@nih.gov to request a consultation or new cloud account
  - Intramural users can also request support via: ServiceNow - Cloud Services - Enterprise Cloud Platforms
- Cloud Lab
  - Intramural users – enroll in Cloud Lab intramural registration page.
  - Extramural users – enroll in Cloud Lab extramural registration page.
- RAS - visit the NIH RAS Service Offerings website for more information and contact information.
- NCPI – visit NCPI Administrative Coordination Center (ACC) website.
- Cloud supplement programs
  - High-Value Datasets Program – watch out the email forwarded from you Scientific Director in Oct. ~ Dec., or send an email to Dr. Fenglou Mao in November, or attend ODSS monthly meetings such as TIWG/FAIR/Data Share and Reuse/Townhall.
  - Cloud supplement NOSI – sign on ODSS newsletter.

Cloud Supplement Programs

- Supplement to intramural projects and contracts
  - High-Value Datasets Program (HVD)
  - Past HVD programs – HVD 20 ~ 23
  - Active HVD Program – HVD 24
- Supplement to extramural awards
  - NOT-OD-23-070 - Notice of Special Interest (NOSI): Administrative Supplements to Support the Exploration of Cloud in NIH-supported Research

Acknowledgement

STRIDES and Cloud Lab

- Nick Weber
- Joshua Stultz
- Dana Gaffney
- Rachel Malashock
- Wayne Chen
- Vishal Thovarai
- Henrique Ludwig
- Gavin Brennan
- Jonny Coleman
- Joshua Jaggat
- Yugandhar Guntaka
- Thad Carlson
- Kyle O’Connell
- Antej Nuhanovic
- Mohan Muthukumarasamy
- Warren Mattocks

ODSS

- Susan Gregurick
- Christopher Siwy
- Fenglou Mao
- Carol Loose
- Gorge Coy
- Ashley Hackket

NCPI

- NCPI Administrative Coordinating Center (RTI and Deloitte)
- NIH NCPI Steering Committee
- Christopher Siwy

RAS

- Jeff Erickson
- OCIO/CIT Leadership
- ODSS Technical Working Group
- RAS Development Team
- NIH IC System Owners and PIs/PMs/Development Teams
- NCBI and dbGaP Team
- RAS Governance Team and Security Advisory Group