National Institutes of Health
Office of Data Science Strategy

Catalyzing data science in research through collaborations

Susan Gregurick, Ph.D.

Associate Director for Data Science, NIH Office of Data Science Strategy

September 18, 2023

Data Science in the next 5 years

- *Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing*
- *Develop Programs to Enhance Human Derived Data for Research*
- *Provide New Opportunities in Software, Computational Methods, and Artificial Intelligence*
- *Support for a Federated Biomedical Research Data Infrastructure*
- *Strengthen a Broad Community in Data Science*

https://dpcpsi.nih.gov/council/september-7-2023-agenda

Agenda

1. Creating and Sharing FAIR Data
2. Developing and Sustaining Software
3. Utilizing the Cloud
4. Federating NIH Data Platforms
5. AI

Office of Data Science Strategy | OD/DPCPSI

Provides NIH-wide leadership and coordination for a modernized NIH data resource ecosystem*

What we do

Provides **leadership and coordination** on the strategic plan for data science

Develops NIH's vision for a **modernized** and **integrated** biomedical data ecosystem

Enhances a **diverse and talented** data science workforce

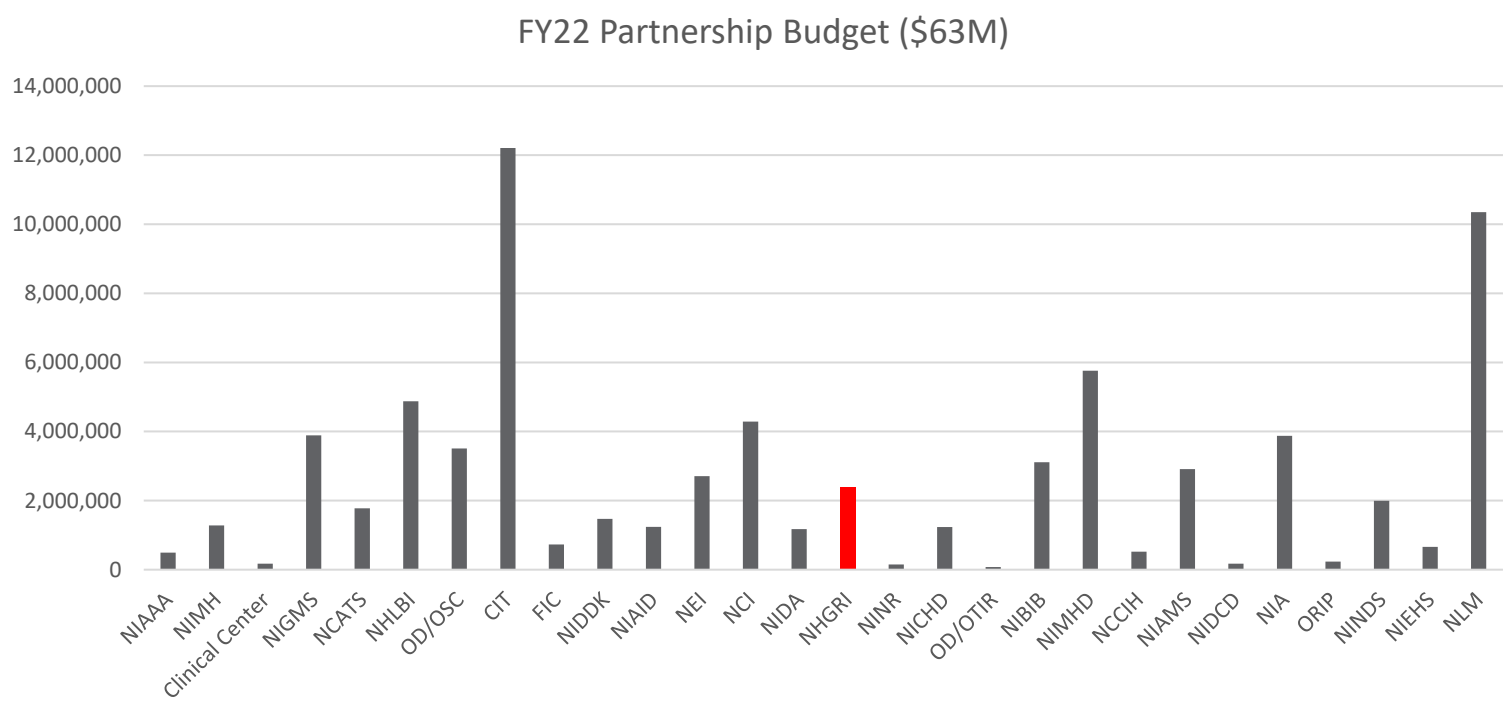
Builds strategic partnerships to advanced technologies and methods

*Over 140 NIH staff from 26 ICOs work with ODSS on working groups and projects

NHGRI Council September 18 2023

Data Science Partnerships | 2022

Support from ODSS to NIH Institutes, Centers, and Offices



Data Science Focus Areas

- Artificial Intelligence
- Enhancing Data Infrastructures
- Supporting Data Sharing
- Developing Tools and Analytics
- Training and Data Science Workforce Development

Creating and Sharing FAIR Data

NIH POLICY FOR DATA MANAGEMENT & SHARING

- TWO BASIC REQUIREMENTS
  - Submission of a Data Management & Sharing “Plan” for all NIH-funded research
  - Compliance with the ICO-approved Plan
- Effective January 25, 2023 *(replaces 2003 Data Sharing Policy)*

Improving Discoverability of Existing Resources and Improving Data Within These Resources

https://findwise.com/blog/data-that-really-saves-lives-and-possibly-your-organisation/

(*used with permission from Ingrid Dillo*)

TRUST

Transparency Responsibility User focus Sustainability Technology

The TRUST Principles

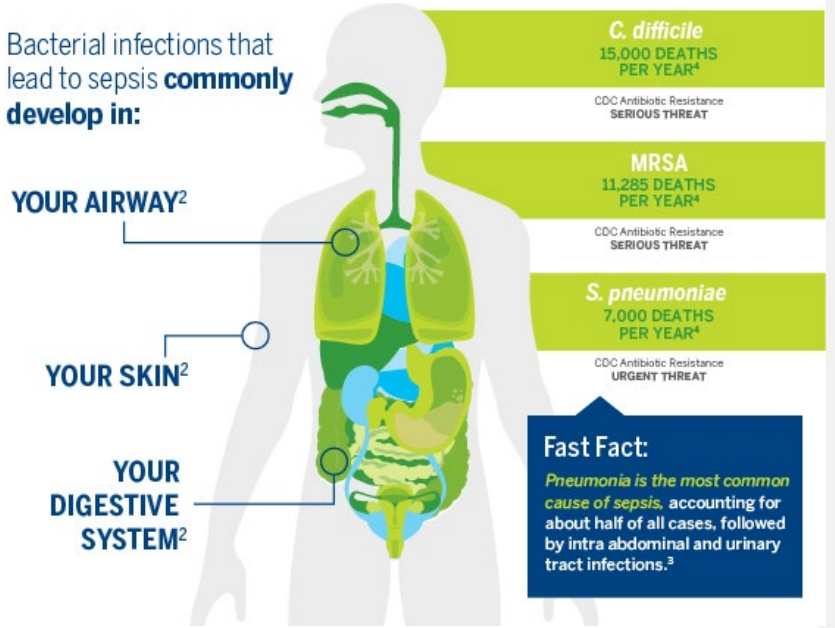
Principle	Guidance for Repositories
Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

Source: Lin et al., 2020. The TRUST Principles for Digital Repositories. Scientific Data https://doi.org/10.1038/s41597-020-0486-7

NIH's Support for FAIR Data

Maria Vazquez Gillamet, Associate Professor of Medicine, Washington University in St. Louis

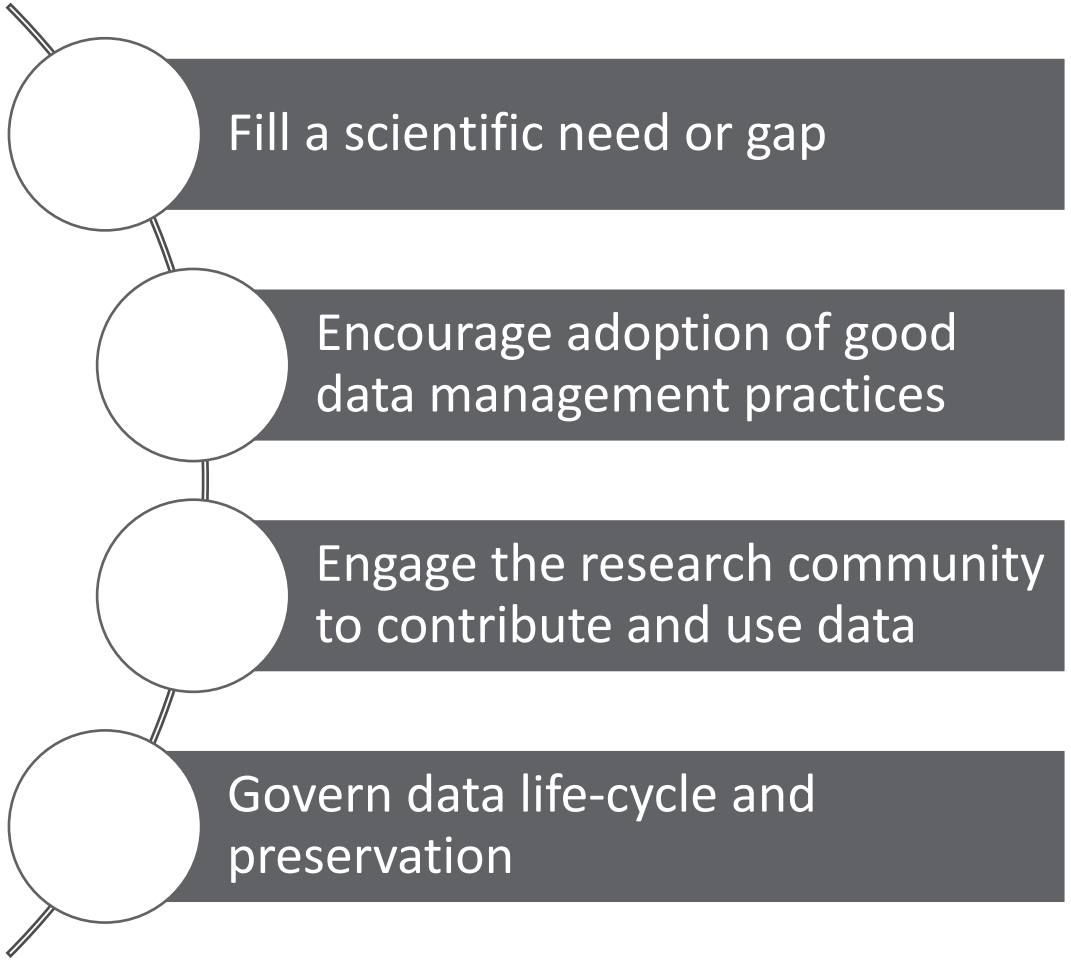
Developing data enrichment tools that will advance the use of electronic health records data to stratify septic patients at risk for infections caused by multidrug resistant microbes



Impact of 3 NOFOs in 2021 to 2023

117	awards
11	IDeA States
18	NIH ICOs
Alzheimer's, Parkinson's, Cardiovascular, Cancer, Aging	Disease focus
Imaging, EHRs, -omics, Speech, environmental, drug, survey	Data Focus

NIH's Support for FAIR & TRUST Repositories



Impact of 2 NOFOs in 2020 to 2023

21	awards
2	IDeA States
7	NIH ICOs
Alcohol Research, Virus Taxonomy, Vaccine Information, Chemotherapy Drugs, Drosophila, Human Pathways, GWAS, Neurotrauma	Science focus

NHGRI Council September 18 2023

NIH's Support for Data Resources

Alex Bateman, Senior Team Leader - Protein Sequence Resources, European Molecular Biology Laboratory

UniProt: A Protein Sequence and Function Resource for Biomedical Science

- **NIH-wide program** to support data management and data sharing through targeted Data Resource Funding
- **Lead Funding IC for UniProt**: NIGRI
- **Partner ICs** include NIA, NIAID, NCI, NIDDK, NEI, NIGMS, NHLBI, OD/ODSS

NEW Funding Announcement PAR-23-236 & PAR-23-237

Support Early Stage and Established Data Resources

Standard NIH Due Dates

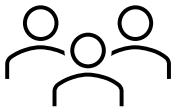
LOI Strongly Encouraged

Special Emphasis Panel Review

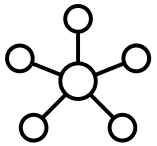
ODSS facilitates NIH-wide program, with co-funding

Supporting Data Management and Sharing | Next 5 Years

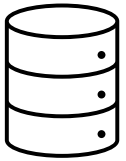
The activities below will establish guidelines, processes, data sharing tools and training. NIH will increase community development and use of community agreed upon standards, enhance AI capabilities in data curation and harmonization. Ultimately the goal is to ensure the long-term sustainability of NIH funded data assets.



Support the biomedical community to manage, share, and sustain data

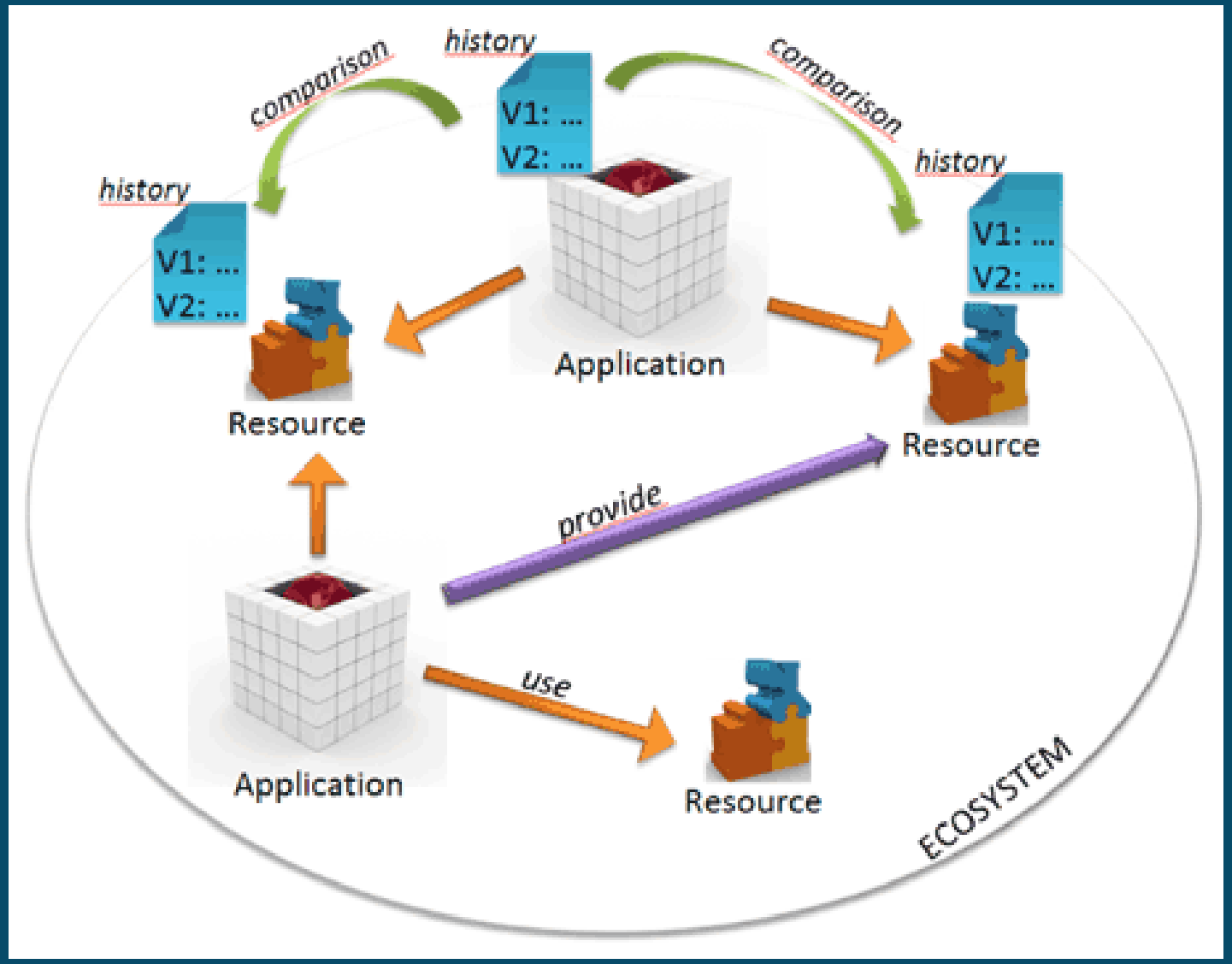


Enhance FAIR data and greater data harmonization



Strengthen NIH’s data repository and knowledgebase ecosystem

Software is an Important Part of a Data Ecosystem



NHGRI Council September 18 2023

NIH's Support for Software to Enhance Open Science

- Enhance software engineering of valuable scientific tools
- Encourage new collaborations between biomedical and clinical scientists and software engineers
- Make research tools reliable and sustainable across multiple computing environments
- Improving reuse and effectiveness of NIH-developed software for open science

Impact of 4 NOFOs in 2020 to 2023

$22.2M	ODSS funds
126	awards
10	IDeA States
19	NIH ICOs

Examples of Software to Enhance Open Science

Software Engineering for Cloud-Native Toolkits

Gabor Marth, RUFUS genomic structural variation (NHGRI)

Extracting Data for Sharing on Cloud

Melanie Fried-Oken, Brain-Computer Interface software to collect & share severe speech defect data using cloud (NIDCD)

Extracting Data for Sharing on Cloud

Melanie Fried-Oken, Brain-Computer Interface software to collect & share severe speech defect data using cloud (NIDCD)

Community Outreach

Gerardo Andres Cisneros, multi-scale modeling & dynamic “graphic novel” on Twitter for LatinXChem (NIGMS)

https://dpcpsi.nih.gov/council/january-19-20-2023-agenda

NEW Funding Announcement in FY24!

Research Software Engineering Program (R50)

Pilot a new model to support research software engineers in biomedical and behavioral research

Software Sustainability Program (R03)

Foster software foundations to increase robustness, reproducibility, and reusability of NIH supported open software

Partnership with NSF | Smart and Connected Health

Accelerate innovations in computer and information science and engineering to support the transformation of health and medicine

Areas of Interest:

- FAIR and Trustworthy AI
- Transformative Analytics
- Next Generation Multimodal and Reconfigurable Sensing Systems
- Cyber-Physical Systems
- Robotics
- Biomedical Image interpretation
- Unpacking Health Disparities and Health Equity

Register for the webinar and apply: https://bit.ly/45GoRP5

NEW Funding Announcement!
NOT-OD-21-011

NSF application criteria and review

Webinar on Sept. 25

Full Proposals due
November 9, 2023

10 to 16 projects will be funded, ODSS co-funds Data Science projects

23 NIH ICOs participating

Cloud Computing & Biomedical Research

"In a cloud environment, you don't need to own a data center to do research at a global scale. Today, the use of cloud-based tools enables analyses across **petabytes of biomedical data** to identify patterns and markers for disease predisposition, prediction, and causality."

David Glazer
Terra CTO, Verily

Alexander Titus, PhD
Strategic Business Executive, Global Public Sector, Google Cloud

Google Blog, September 24, 2021

STRIDES Initiative | Value to Participants

Participants in the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative benefit from:

- **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

Google AWS Microsoft Azure

Impact *as of June 30, 2023*

248+	PETABYTES OF DATA
440M+	COMPUTE HOURS
1500+	RESEARCH PROGRAMS
$65M+	COST SAVINGS
5300+	PEOPLE TRAINED

Supporting Researchers to use STRIDES

Encourage and enable researchers to explore and test opportunities by incorporating cloud capabilities

- *Paul Sternberg* (California Institute of Technology)- **Text mining in the Cloud**
- *Michael Schatz* (Johns Hopkins University)- **Long Read Variant Frequency Database on AnVIL**
- *Ben Heavner* (University of Washington)- **Building a cross-study data set for the PRIMED consortium**

Impact from 2022 (NOT-OD-23-070)

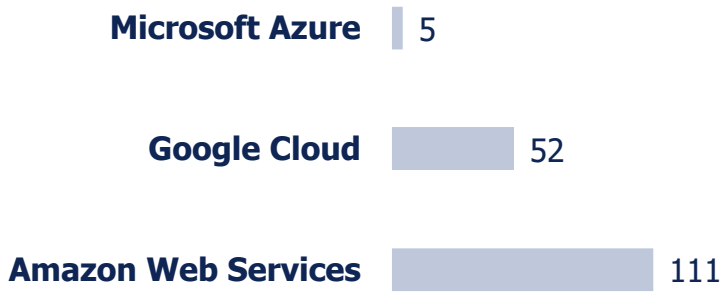
$1.8M	ODSS in funding
40%	Intramural Researchers
41%	For new Staffing costs
7	NIH ICOs

NHGRI Council September 18 2023

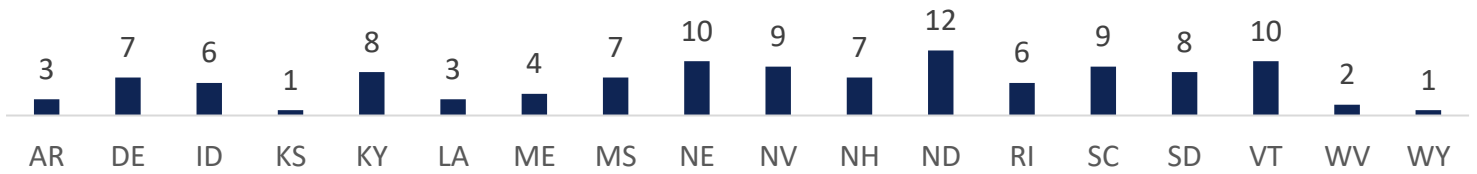
NIH Cloud Lab | Experiment in the Cloud

Through this resource, NIH-funded researchers will become more efficient and comfortable in leveraging the cloud for their research purposes.

Accounts Issued as of August 2023



Accounts Issued in Institutional Developmental Award States



Use Cases

Evaluate Utility & Cost

Reduces the financial, labor, and time commitments required to evaluate the cloud's utility/cost for a project

Develop New Tools

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

Share Ideas

Enables researchers across the world 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

Partnering to Improve Data Science Education

NIH awards $5.8M to North Carolina Agricultural and Technical State University **to create genomic data science educational hub for early career researchers**

Funding to enhance diversity in genomic data science through cloud computing

Goals

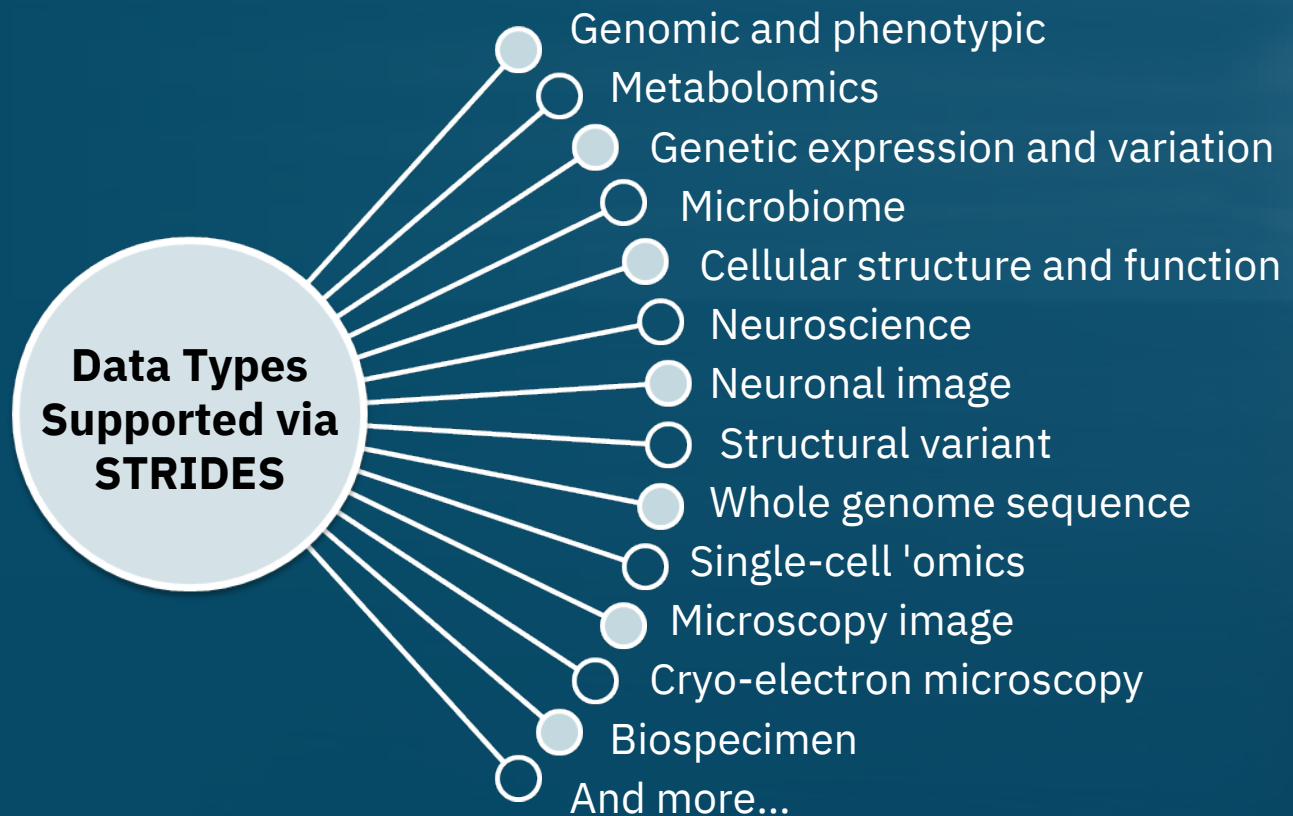
- Proved an anchor for institutions that are serving students from diverse backgrounds
- Host workshops and hands-on learning sessions
- facilitate educational and research initiatives
- Help students build a solid foundation

Partnership with NHGRI (Lead) and support from All of Us and ODSS

Bringing the Data (platforms) and Analysis together

Data IS a Resource

As the scientific importance of data continues to grow, researchers need to be equipped to collect, analyze, and manage a growing list of data types.



Support for FAIR Data is critical for biomedical research

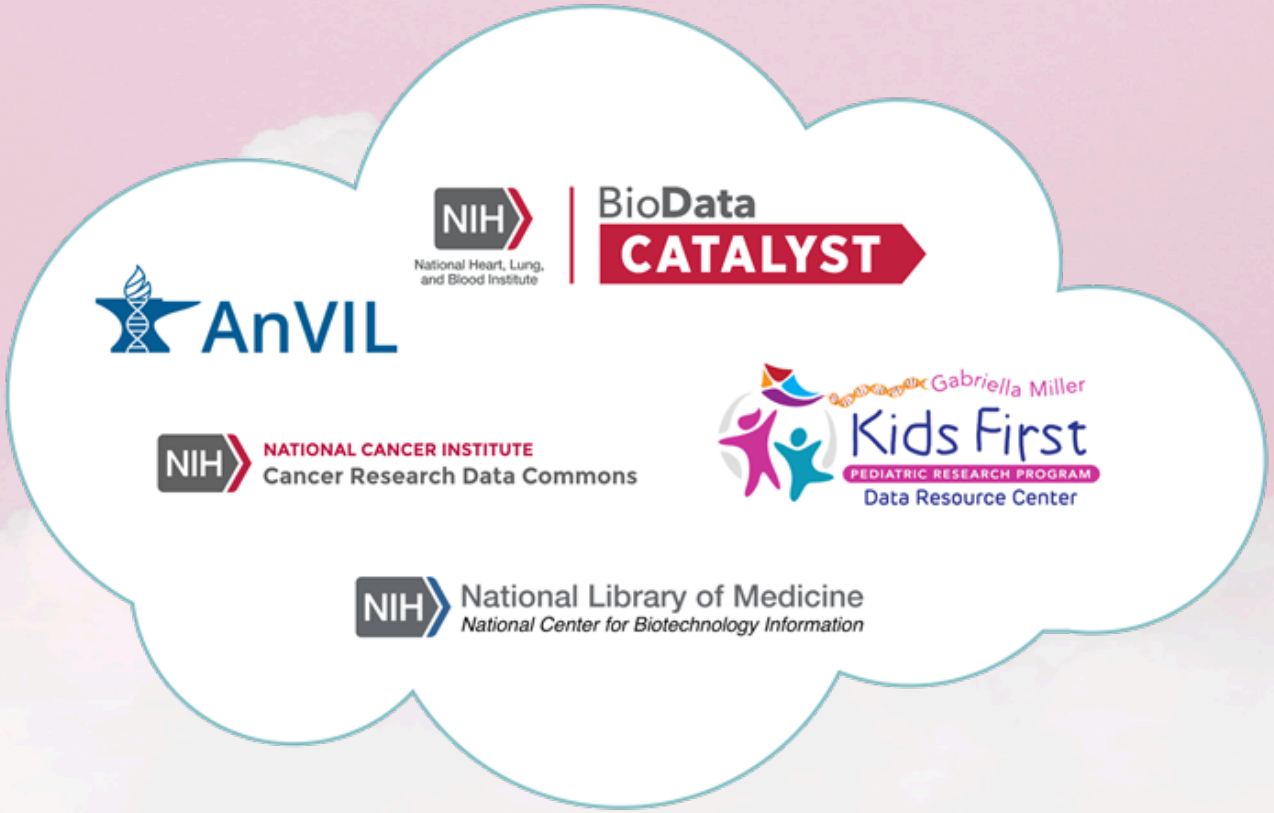
Connecting NIH's many data systems is a critical step toward improving the fairness and value of data in the ecosystem

- NIH has invested significant resources into the generation of large-scale datasets, such as human genomic, clinical, and imaging data. Many of these data are stored in different cloud-based repositories that are stewarded by many ICs

- This can cause researchers to struggle to find, access, aggregate, and co-analyze datasets across different data repositories

- Integrating data from multiple NIH systems in a bespoke manner can be costly, time-consuming, and requires expertise in cloud computing and computer programming.

Creating an Integrated Cloud-based Ecosystem



The **NIH Cloud Platform Interoperability (NCPI) effort** aims to establish and implement guidelines and technical standards to empower greater data sharing and end-user analyses across participating NIH cloud platforms

https://anvilproject.org/ncpi

About NCPI

Goal

The goal of the NIH Cloud Platform Interoperability (NCPI) program is to enable a federated data ecosystem that will facilitate researcher-driven analyses of datasets across multiple cloud-based platforms and repositories

Current Participating Platforms and Systems

- AnVil
- BioData Catalyst
- Cancer Research Data Commons
- Gabriella Miller Kids First Data Resource
- dbGaP/SRA

Guiding Principles

- Form trust relationships and interoperate with other platforms
- Follow the golden rule of data resources
- Support the principle of least restrictive access
- Agree on standards, compete on implementations
- Plan to support patient partnered research

Example Interoperability Technologies

- Researcher Auth Service (RAS)
- Global Alliance for Genomics and Health (GA4GH) Data Repository Service (DRS)
- Fast Healthcare Interoperability Resources (FHIR)
- Portable Format in Bioinformatics (PFB)

Interoperability Projects | Create, Test, Improve

Potential areas of interest underlying interoperability projects may include but are not limited to:

- The ability to search for data across the NCPI partner systems
- Executing the same analysis workflows on different platforms and validating the workflow equivalency
- Testing new IT standards to foster cross-platform analysis
- Creating resources for estimating cloud computing costs for popular workflows in various platforms
- Improving semantic and syntactic interoperability to support a specific combined analysis

NIH Cloud Platform Interoperability (NCPI) effort

$16.8M Total Funding in FY22 & FY23 including $500K from NHGRI to support:

- **Administrative and Coordination Center** (Research Triangle Institute)
- **Development of technology based on Science Use Cases across multiple NIH Data platforms.** Use cases include studies of rare genetic variants, utilization of FHIR, sex-biased chromosome association in targeted diseases, and Hispanic colorectal cancer health disparities
- **Access to multiple datasets** across NIH data Systems, using Research Auth Services, with appropriate data access processes and security assurances
- **Support** for new researchers, training, outreach
- **Partnership with GA4GH**

Artificial Intelligence

The Promise of AI

Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation

MICHELLE M. CLARK, AMBER HILDRETH, SERGEY BATALOV, YAN DING, SHIMUL CHOWDHURY, KELLY WATKINS, KATARZYNA ELLSWORTH, BRANDON CAMP, CYRIELLE I. KINT, [...] STEPHEN F. KINGSMORE +52 authors Authors Info & Affiliations

SCIENCE TRANSLATIONAL MEDICINE · 24 Apr 2019 · Vol 11, Issue 489 · DOI: 10.1126/scitranslmed.aat6177

The Challenges in AI

nature

Explore content · About the journal · Publish with us · Subscribe

nature › news › article

NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

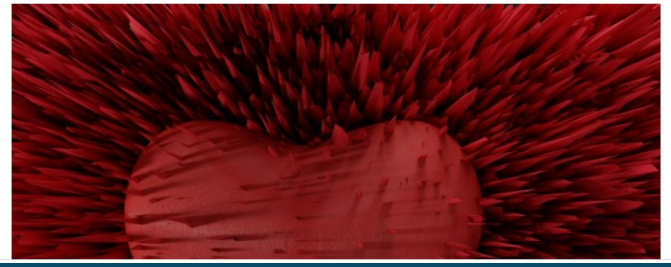
Heidi Ledford

Opportunities to Improve AI

POPULATION HEALTH NEWS

SDOH Improves Performance of Heart Failure Mortality Predictive Model

Researchers have found that machine-learning models that incorporate social determinants of health data perform better than traditional methods of predicting heart failure deaths among Black patients.

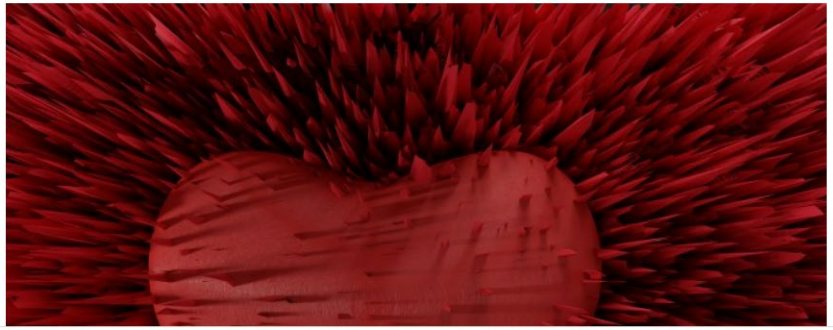


- Unrepresentative data
- Bias within training data
- Bad design/asking the wrong question
- Bias in algorithm development and implementation
- Lack of diversity of researchers
- Lack of data in lived experiences, historical/cultural contexts such as social determinants of health (SODH)

POPULATION HEALTH NEWS

SDOH Improves Performance of Heart Failure Mortality Predictive Model

Researchers have found that machine-learning models that incorporate social determinants of health data perform better than traditional methods of predicting heart failure deaths among Black patients.



JAMA Cardiol. 2022;7(8):844-854.
doi:10.1001/jamacardio.2022.1900

NIH and AI | Programs, Workshops, Meetings

- **Bridge2AI** to generate new “flagship” data sets and best practices for machine learning analysis.
- **AIM-AHEAD** to enhance the participation and representation of researchers and communities currently underrepresented in the development and use of AI/ML
- **AI-Ethics** supplements (FY22)
- **AI-Workforce** supplements (FY21)
- **AI-Readiness** supplements (FY21,22,23)
- **2021 AI-Ethics Micro Labs**, AI-Ethics IdeasLab
- **NSF/NIH Workshop** on AI for Biology
- ***NASEM*** AI Code of Conduct Study

NHGRI Council September 18 2023

Support for Collaborations to Improve the AI/ML-Readiness of NIH-Supported Data

UniProt - Protein sequence and function embeddings for AI/Machine Learning readiness

Alex Bateman, European Molecular Biology Laboratory

Retinal Circuitry

Bryan Jones, University Of Utah

Development and validation of a computable knowledge framework for genomic medicine

Alex Wagner, Research Institute Nationwide Children's Hospital

Impact of 3 NOFOs in 2021 to 2023

$29.7M	ODSS funds
107	awards
3	IDeA States
20	NIH ICOs

New Opportunities in Artificial Intelligence

The activities below will introduce new opportunities to support collaborations in developing socio-technical solutions, including guidelines and principles, for ethical AI, including new technologies and methods for foundational models.

- **Develop** social and technical solutions for ethical AI
- **Create** and validate an approach for using synthetic clinical datasets for AI
- **Leverage** new technologies and methods for AI and foundational models to accelerate biomedical and behavioral research
- **Develop** new AI technologies that will enable the translation of data to knowledge
- **Enhance** NIH capabilities in AI through partnerships across federal agencies and communities

Thank you to:

- *Heidi Sofia* is a co-chair of the Software Working Group and participates in the AI-Ethics Working Group.
- *Ajay Pillai* is a co-chair in the Database/Knowledgebase Working Group.
- *Valentina Di Francesco* is a lead in the NCPI program and collaborated on the DATA Scholar Program. Valentina also participates in the STRIDES Extramural Working Group.
- *Nicole Lockhart* participates in the AI-Ethics Working Group.
- *Ken Wiley* participates in the NCPI Steering Committee.
- *Shurjo Sen* serves on the AIM-AHEAD NIH Advisory Group
- *Christopher Wellington* participates in the Data Repository and Knowledgebase Working Group.

Additional Slides as an FYI

Additional ODSS Activities to support Data Sharing

- **Data Management Center of Excellence (MITRE)** to support development of data management and sharing best practices and materials, including trainings, to NIH staff
- **Generalist Repositories Ecosystem Initiative** to provide a home for data that don't have a home today & enable the concept of "**Co-opetition**" for competitors to work together to develop a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure
- Established NIH as a **DataCite member** to meet the critical need to mint PID's - digital object identifiers (DOIs) for data generated from NIH funded and conducted research
- Partnered with FASEB to organize the **DataWorks! Prize** (NIH Challenge) to incentivize researchers on better data management practices for data sharing and reuse data - www.herox.com/dataworks