National Institutes of Health
Office of Data Science Strategy

Creating a FAIR and Equitable Data Ecosystem

Susan K. Gregurick, Ph.D.
Associate Director for Data Science
National Institutes of Health

Topics for Today

- Creating Data Science Infrastructure
- Enhancing Data Science Capacity
- Expanding Use of Generalist Repositories
- Delivering Data Science Outcomes

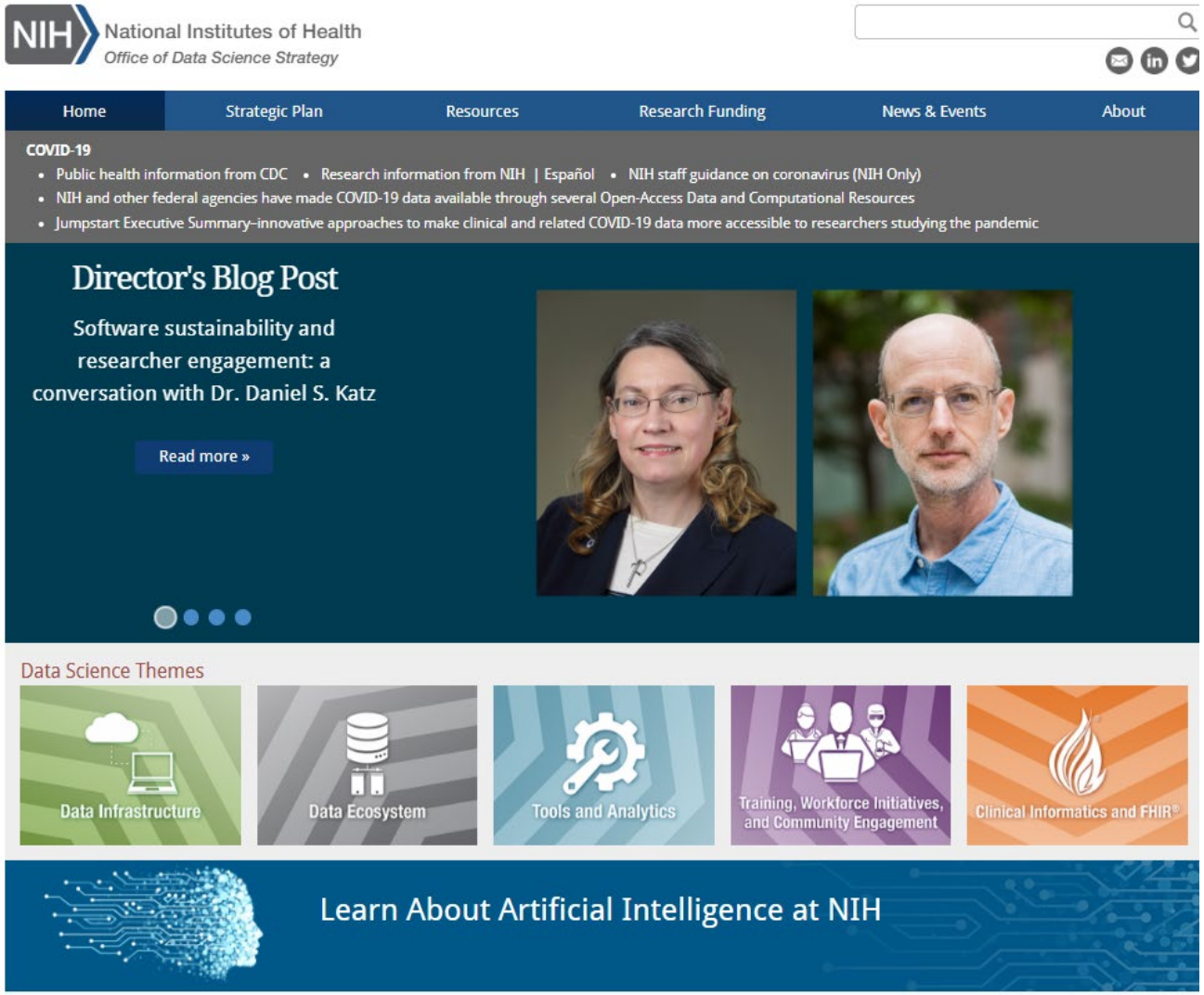
Our vision is built on the Strategic Plan for Data Science

- Support data infrastructure and architecture
- Leverage commercial tools, technologies, services, and expertise
- Enhance the nation’s biomedical data-science research workforce
- Enhance data sharing, access, and interoperability
- Ensure the security and confidentiality of participant data
- Develop & promote data standards, vocabularies and ontologies

Office of Data Science Strategy

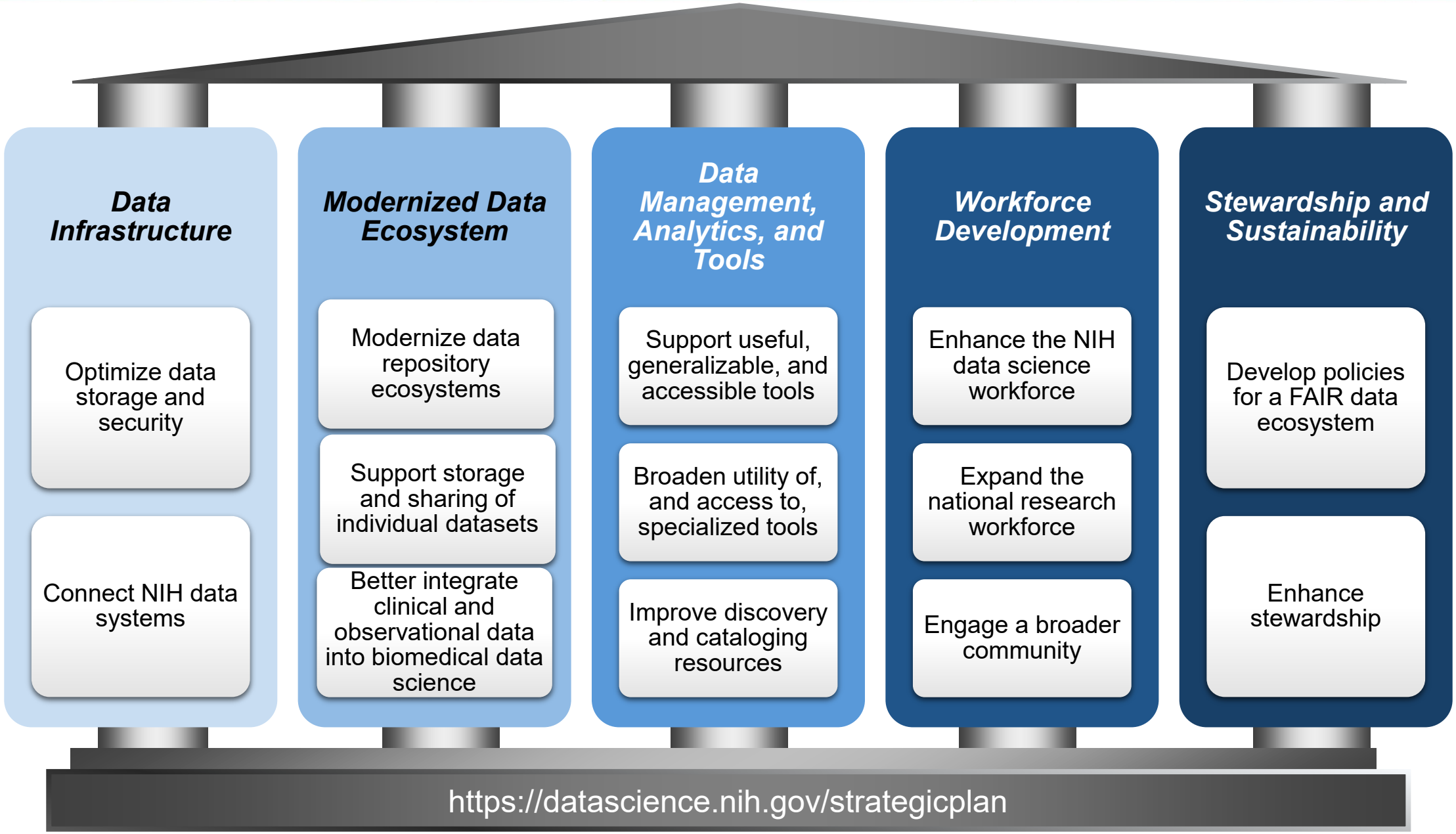
The NIH **Office of Data Science Strategy (ODSS)**, in the Office of the Director:

- Provides **leadership and coordination** on the strategic plan for data science
- Develops and implement NIH's vision for a **modernized** and **integrated** biomedical data ecosystem
- Enhances a **diverse and talented** data science workforce
- **Builds strategic partnerships** to develop and disseminate advanced technologies and methods



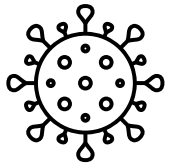
datascience.nih.gov

NIH Strategic Plan for Data Science – Goals & Objectives



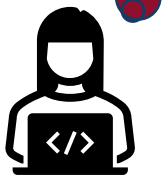
Accomplishments

200 PB data on 3 Clouds



Identify and share coronavirus sequences collected by the global research community

9+ NIH IC data platforms allow for single sign-on of researchers

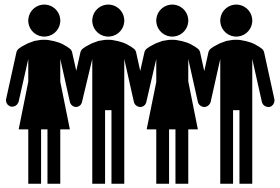


Researchers can data from the GTEx AnVIL platform & the Cancer Data Commons for combined analysis of LINE-1 expression

Programs to support databases and knowledgebases, includes DataCite

Aligning data resources to support NIH Data Management and Sharing Policy

Programs to enhance data workforce



NIH 1–2-year sabbaticals for fellows to work at NIH on challenging data science problems

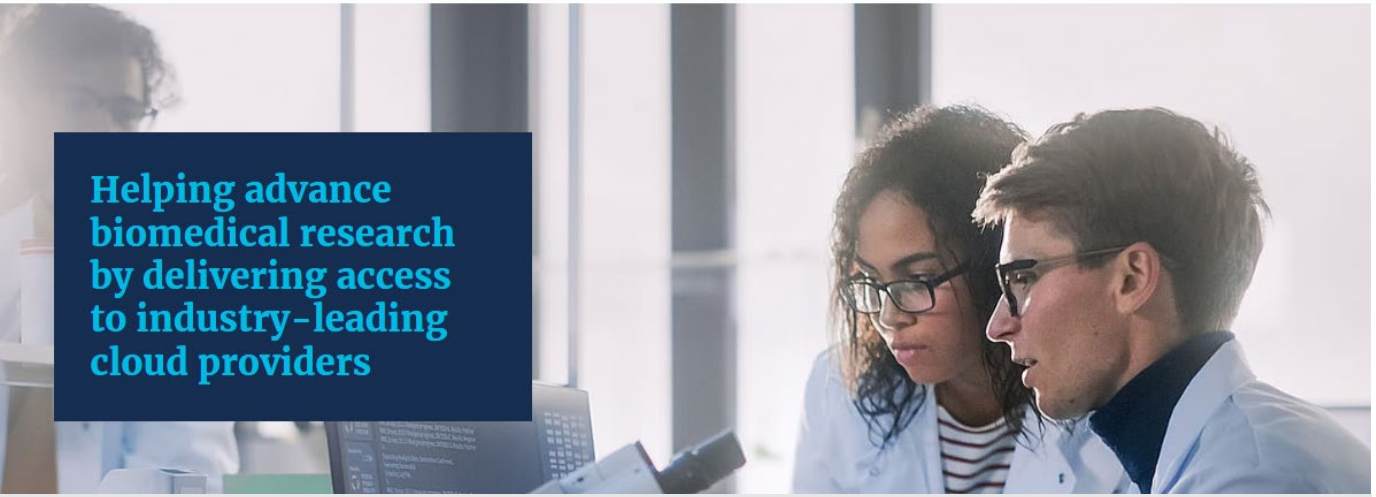
Creating Data Science Infrastructure

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



The STRIDES Initiative aims to help NIH and its institutions accelerate biomedical research by reducing barriers in utilizing commercial cloud services. This initiative aims to harness the power of the cloud to accelerate biomedical discovery. NIH and NIH-funded researchers can take advantage of STRIDES benefits.

Benefits:

- Discounts on partner services
- Professional services consultations
- Access to training
- Potential collaborative engagements

>200
Petabytes of Data

274M
Compute Hours

>995
NIH & NIH-funded Research Programs/ Projects

$41M
Cost Savings

>4700
People Trained

https://datascience.nih.gov/strides

Data is the new oil!

Genetic Expression and Variation Analysis

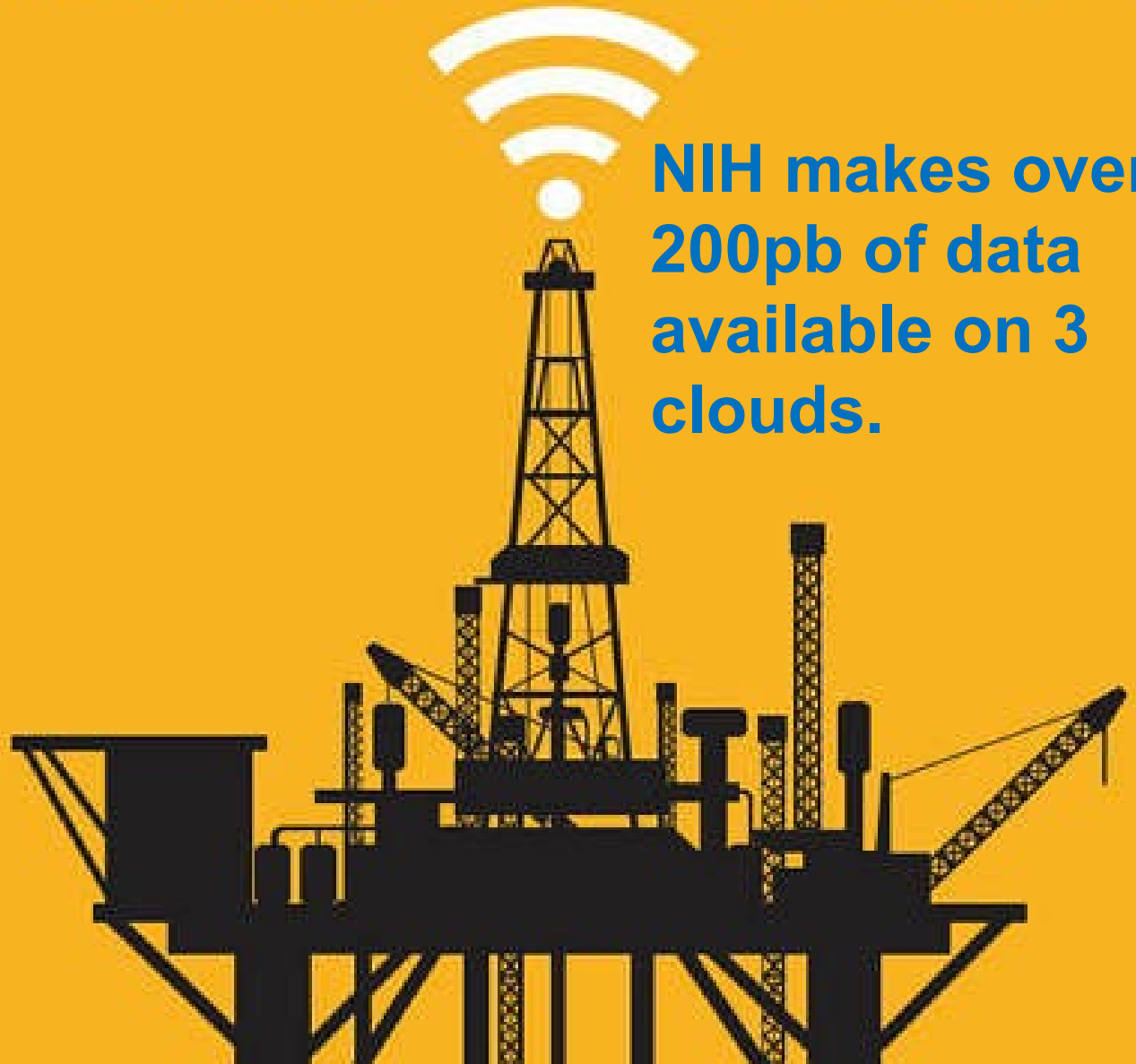
Microbiome Analysis

Cellular Structure and Functional Analysis

Neuroscience Analysis

Genomic and Phenotypic Analysis

Neuronal Image Analysis



NIH makes over 200pb of data available on 3 clouds.

Metabolomics Analysis

Whole Genome Sequence Analysis

Single-Cell ‘Omics Analysis

Microscopy Image Analysis

Cryo-Electron Microscopy Analysis

Clinical Analytics, new applications of FHIR

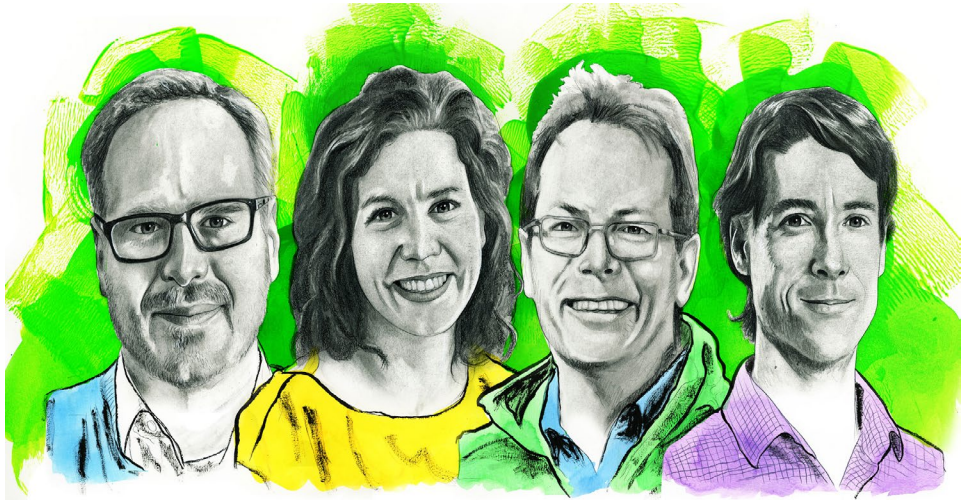
Advancement in cloud means more is data available!

During the last five years NIH has seen a remarkable growth in the support for and use of biomedical controlled access data repositories and platforms for analytical research

- **NHLBI's BioData Catalyst** which provides access to NHLBI clinical study data
- The **NCI Cancer Research Data Commons** (CRDC) which provides access to a large collection of cancer research data
- The **Kids First Data Resource** which houses data on 44 childhood cancer and structural birth defects cohorts
- The **NHGRI Genomic Data Science** Analysis, Visualization, and Informatics Lab-space (AnVIL) genomic data sharing and analysis platform
- **AllofUs** which contains data from over 500,000 participants
- **NIH database of Genotypes and Phenotypes** (dbGaP) which controls access to sequence data and genotype and/or phenotype data from over 3.2 million participants

These resources are **cloud-based data infrastructures** that provide the research community with data and analytical tools, applications, and workflows in secure environments

STRIDES Enabled Research of 2022 'TIME 100’ Most Influential People in the World



Michael Schatz, Karen Miga, Evan Eichler, Adam Phillippy

STRIDES, an ODSS cloud storage and compute investment, facilitated the Telomere to Telomere (T2T) Consortium’s move of computational work to the cloud.

Due to the Telomere-to-Telomere Consortium (T2T) of scientists, we can see the full map of the human genetic landscape and the elements that uniquely fingerprint an individual’s genome.

Hoyt, S. et al. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science, 376*(6588). DOI: 10.1126/science.abk3112

Addressing Barriers to Cloud

Reducing barriers to entry

Provide NIH and NIH-funded groups an easy route to access the cloud so they can quickly evaluate its utility for their project without having to make major time or financial commitments

Technical development

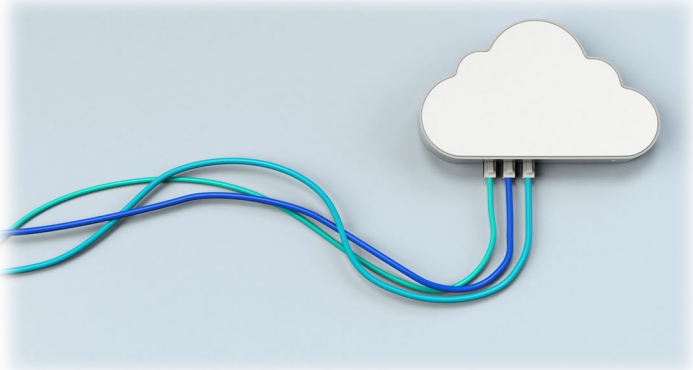
Allow experienced teams access to the cloud environment(s) so they can prototype new architectures, and/or evaluate new software/hardware combinations in a cloud environment

Training

Provide access to the cloud, simplifying access to tools and environments that can be used for training purposes

NIH Cloud Lab Overview

A cloud testbed allowing researchers to
“try before they buy”



Primary Cloud Lab Use Cases

Exploring the Cloud Consoles
Researchers can gain an understanding of the look and feel of cloud environments before they jump into a full STRIDES account for research

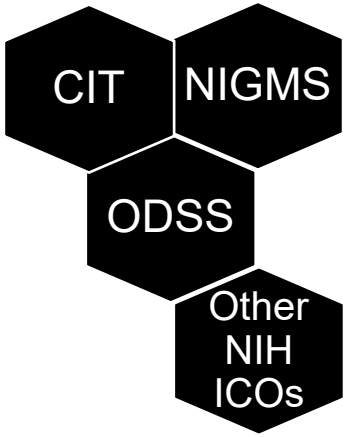
Supplementing Cloud Training
Researchers can use the sandbox to strengthen their understanding of cloud training or follow along with training content in a separate environment.

Experimenting with Simple Cloud Solutions
Researchers interested in solutions for specific scientific tasks can use the sandbox to build proof of concept or other simple solutions to understand LOE and other details for production.

Benchmarking Costs
Testing out different tools and configurations (instance types, sizes, etc.) to optimize research analyses

Integrating NIGMS IDeA Modules into the NIH Cloud Lab

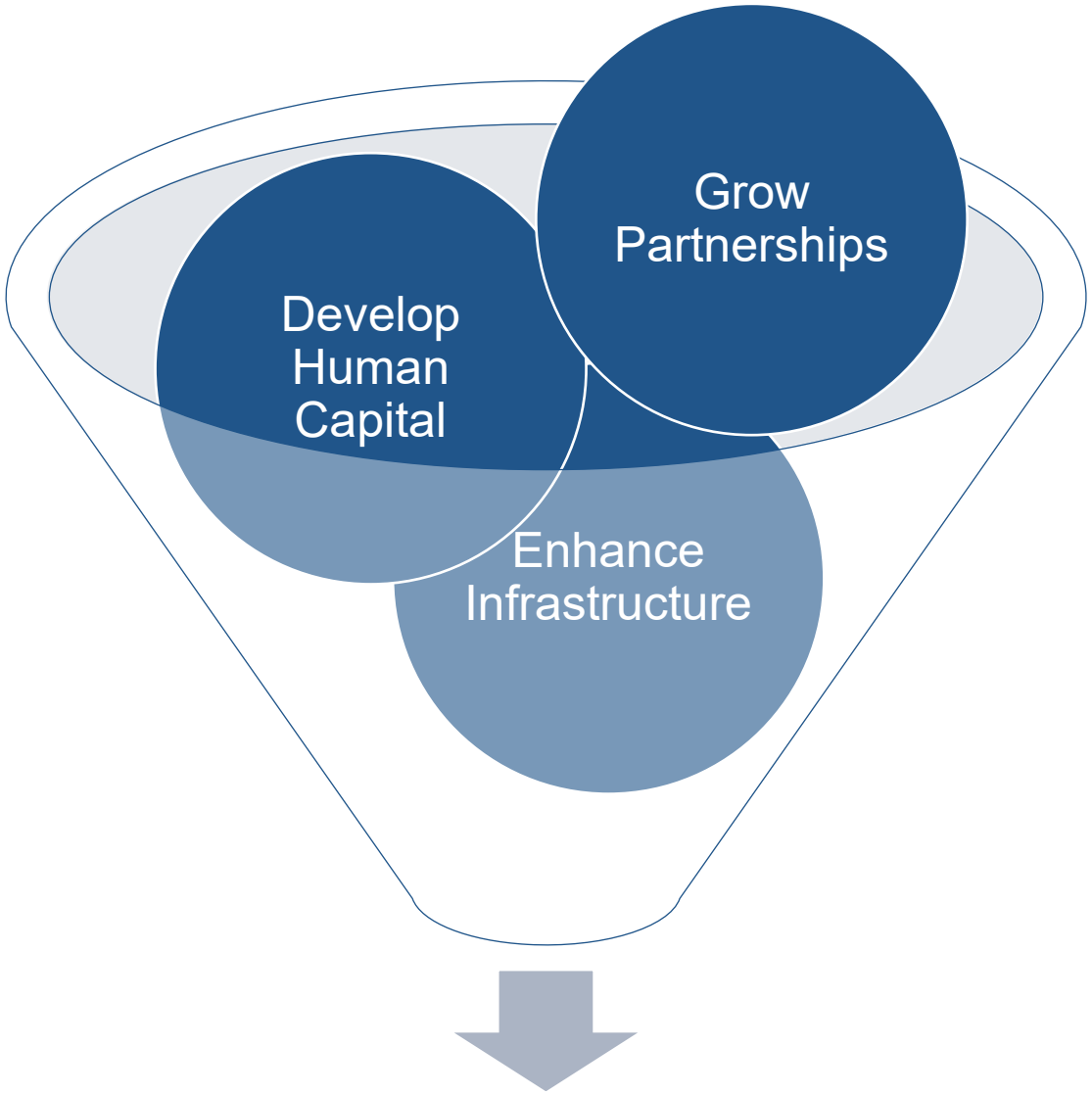
Both NIGMS Sandbox and NIH Cloud Lab enable hands-on data analysis in the cloud



Program	NIGMS Sandbox Support for NOSI supplements and Google Professional Services to develop cloud learning modules for 12 extramural institutions	NIH Cloud Lab STRIDES product providing up to 90 days of “credits” for intra- and (future) extramural researchers to try using the cloud
Core Service	Practical research modules/tutorials	Simple, streamlined access to cloud
Partners	GCP, (AWS coming)	GCP, AWS
Customers	Extramurally focused	Intramural, with extramural in progress
Participants	12 research programs	31 accounts, 12 ICs
Duration	2021 to 2023	2022 to ongoing
Funders	NIGMS, ODSS	ODSS, CIT

Enhancing Data Science Capacity

Promote Capacity Building in Low-Resource Institutions

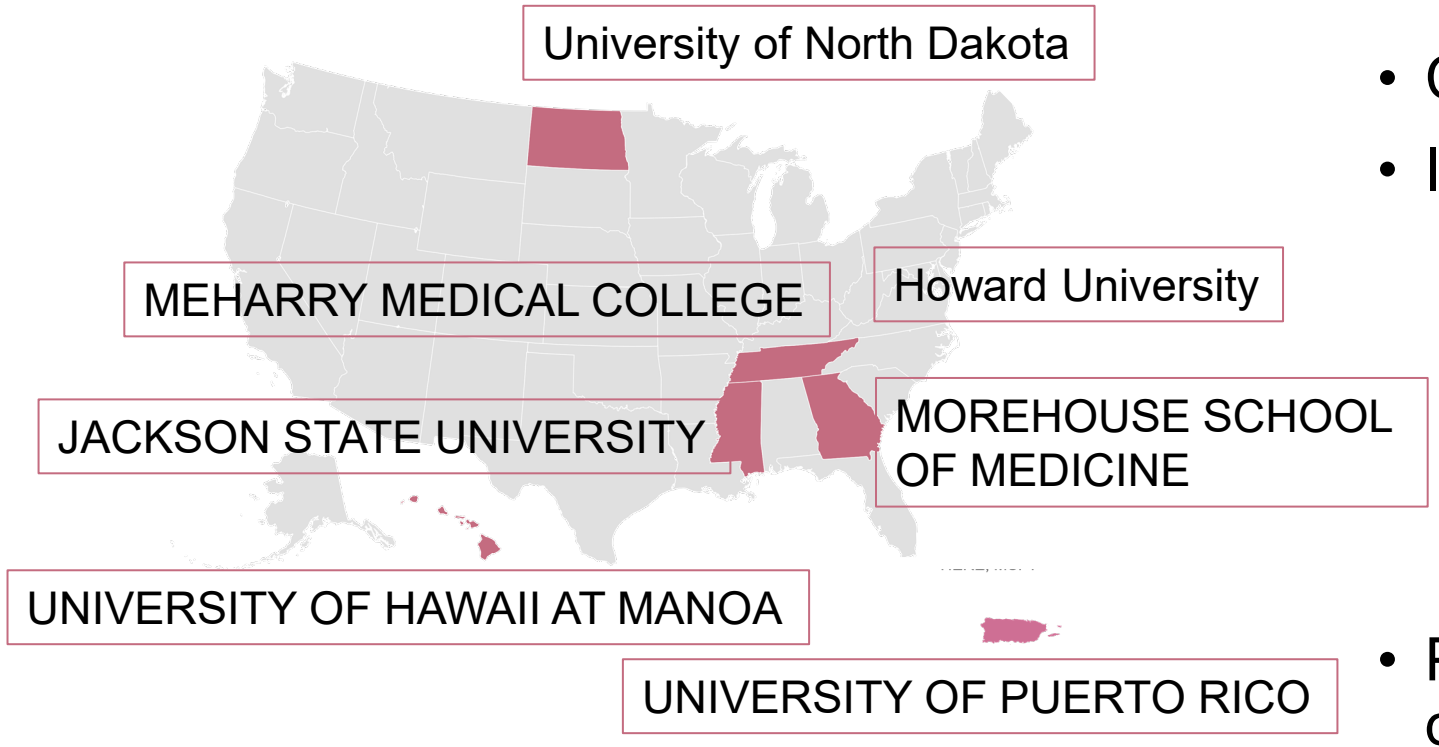


- Administrative Supplements to Enhance Data Science Capacity
- Co-funding of Native American Research Centers for Health (NARCH)
- Administrative Supplements to Develop Cloud-Based Learning Modules

Administrative Supplements to Enhance Data Science Capacity

Leverage existing NIH infrastructure-building programs to enhance capacity building in data science at **low-resource** institutions

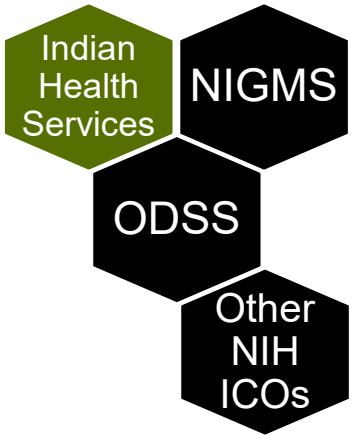
NIMHD NIGMS ODSS



- Currently in Year 2
- Implemented activities include:
  - Development and improvement of courses and curriculum
  - Providing training events
  - Building collaborations
  - Evaluation of impact of the activities and effectiveness of the overall projects
- Participants include undergraduate and graduate students as well as faculty investigators and community leaders

Co-funding Support for Native American Research Centers for Health

The NARCH program supports opportunities for conducting research and career enrichment to meet health needs prioritized by American Indian/Alaska Native (AI/AN) tribes or tribally based organizations



CHEROKEE NATION	Building Tribal Capacity through Informing the Development of Tribal Research Codes to Govern Genomic Research: A Collaboration between Cherokee Nation and the University of Oklahoma
INTER TRIBAL COUNCIL OF ARIZONA, INC	Enhancing Early Career Research Ethics to Support Indigenous Research Governance (faculty development around data governance)
FOND DU LAC RESERVATION	Capacity Building (Develop data repository, data training, trainings on research ethics)

Co-funding Support for Science Education Partnership Awards (SEPA)

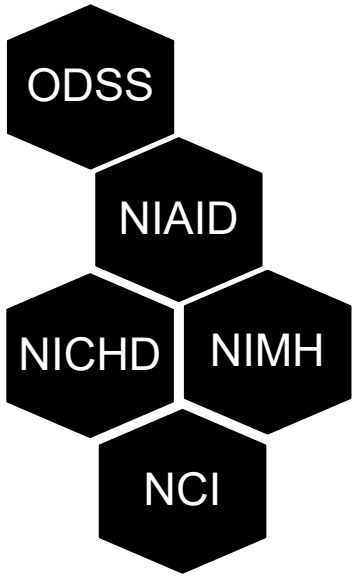
The NIH SEPA program supports STEM and Informal Science Education activities for pre-kindergarten to grade 12 (P–12) students from diverse backgrounds. ODSS currently co-funds two SEPA programs to help grow a pool of well-prepared young students in data science.

NIGMS

ODSS

- The Knox Scholars Data Science Research Program from ***Health Resources in Action, Inc.*** supports high school students from the Boston area who are underrepresented in STEM fields (Black and Latinx youth, first-generation college students, low-income students).
- The Data Detectives: Using Real Data to Solve Real Community Health Problems from ***Emory University*** provides underrepresented middle school students with a curriculum focused on using population-level Big Data for community health needs assessment, planning, analysis, evaluation and application.

New Notice of Special Interest: NOT-AI-23-010



Administrative Supplements for R25 Data Science Training for Infectious and Immune-mediated Disease Research

- Support applications that propose to enhance existing NIH Research Education awards (R25) with data science training relevant to infectious- and immune-mediated disease research
- Encourage applications that improve training programs and foster transdisciplinary collaboration between biomedical, behavioral, clinical, and computational scientists for research on infectious- and immune-mediated diseases, consistent with the NIH Strategic Plan for Data Science

Upcoming submission date: February 14, 2023

https://grants.nih.gov/grants/guide/notice-files/NOT-AI-23-010.html

NIH Data Management and Sharing Policy

Final NIH Data Management and Sharing (DMS) Policy

Effective January 25, 2023, researchers must submit a Data Management and Sharing Plan detailing how data and metadata will be preserved, managed, and shared, including restrictions or limitations.

The DMS Policy (NOT-OD-21-013) applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. This includes research funded or conducted by extramural grants, contracts, Intramural Research Projects, or other funding agreements regardless of NIH funding level or funding mechanism.

The DMS Policy does not apply to research and other activities that do not generate scientific data, including training, infrastructure development, and non-research activities.

Flexible Policy

- **All data should be managed but not all data needs to be shared**
  - **What's in:** All NIH-supported research generating *scientific data* "Recorded factual material… of sufficient quality to validate and replicate research findings" – published or unpublished
  - **What's out:** lab notebooks, preliminary analyses, case report forms, physical objects
- **Data should be accessible as soon as possible**
  - No later than publication or end of award
  - Considerations regarding how long data should be shared (e.g., journal policies, repository policies)

'Which brings us to my next point.'

Elements of a Data Management and Sharing Plan

- **Data type -** Data to be preserved and shared
- **Related tools, software, code -** Tools and software needed to access/manipulate data
- **Standards -** Standards to be applied to scientific data/metadata
- **Data preservation, access, timelines -** Repository to be used, persistent unique identifier, and when/how long data will be available
- **Access, distribution, reuse considerations -** Factors for data access, distribution, or reuse
- **Oversight of data management –** How Plan compliance will be monitored/ managed and by whom

NOT-OD-21-014 – Supplemental Information to the NIH Policy for Data Management and Sharing: Elements of an NIH Data Management and Sharing Plan

Allowable Costs

- **Reasonable costs allowed in budget requests**
  - Curating data/developing supporting documentation
  - Preserving/sharing data through repositories
  - Local data management considerations

- **NOT considered data sharing costs**
  - Infrastructure costs typically included in indirect costs
  - Costs associated with the routine conduct of research (e.g., costs of gaining access to research data)

NOT-OD-21-015 – Supplemental Information to the NIH Policy for Data Management and Sharing: Allowable Costs for Data Management and Sharing

Sharing.nih.gov

DATA MANAGEMENT AND SHARING POLICY | GENOMIC DATA SHARING POLICY | OTHER SHARING POLICIES | ACCESSING DATA | ABOUT

Expediting the Translation of Research Results to Improve Human Health.

Featured News & Events

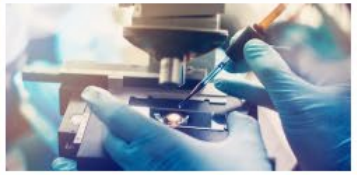
Questions? Answers! Feb 1-2 NIH Virtual Grants Conference sessions and data sharing booth.

VIEW MORE →

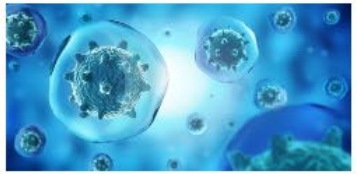
Explore the areas in which NIH has sharing policies.

Scientific Data

Genomic Data



Research Tools



Model Organisms

Clinical Trials

Research Publications

Creating Data Sharing Capabilities

The NIH Data Sharing Landscape

NIH strongly encourages
open access data sharing repositories
as a first choice.

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Datasets up to **2 gigabytes**

PubMed Central

Stores publication-related supplemental materials and datasets directly associated publications.

Datasets up to **20 gigabytes**

Generalist Repositories

Datasets associated with publications or otherwise and links to PubMed.

High priority datasets **petabytes**

Cloud Partners
(STRIDES Program)

Store and manage large scale, high priority NIH datasets.

Positioning repositories for sharing (NOT-OD-23-044)

monarch INITIATIVE

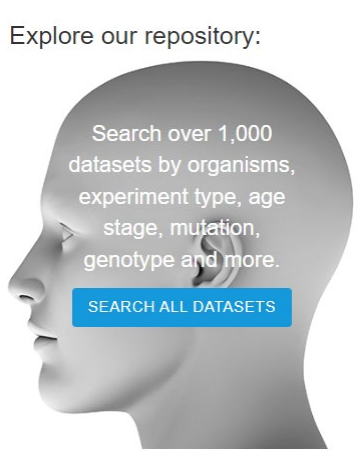
Mouse Phenome Database at The Jackson Laboratory

ALLIANCE of GENOME RESOURCES

UniProt

DBAASP v3.0

Database of antimicrobial activity and structure of peptides



National Toxicology Program
U.S. Department of Health and Human Services

Chemical Effects in Biological Systems

FY21-FY22: 21 Awards, $4.4M

Biomedical focus areas: Alzheimer's, traumatic brain injuries, obesity nutrition, mental health, immune response, environmental data, vision, ontology

Data types: imaging, behavioral measures, clinical, claims data, EHRs, - omics, environmental health data, brain data, speech and language

IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

WORLDWIDE PDB
PROTEIN DATA BANK

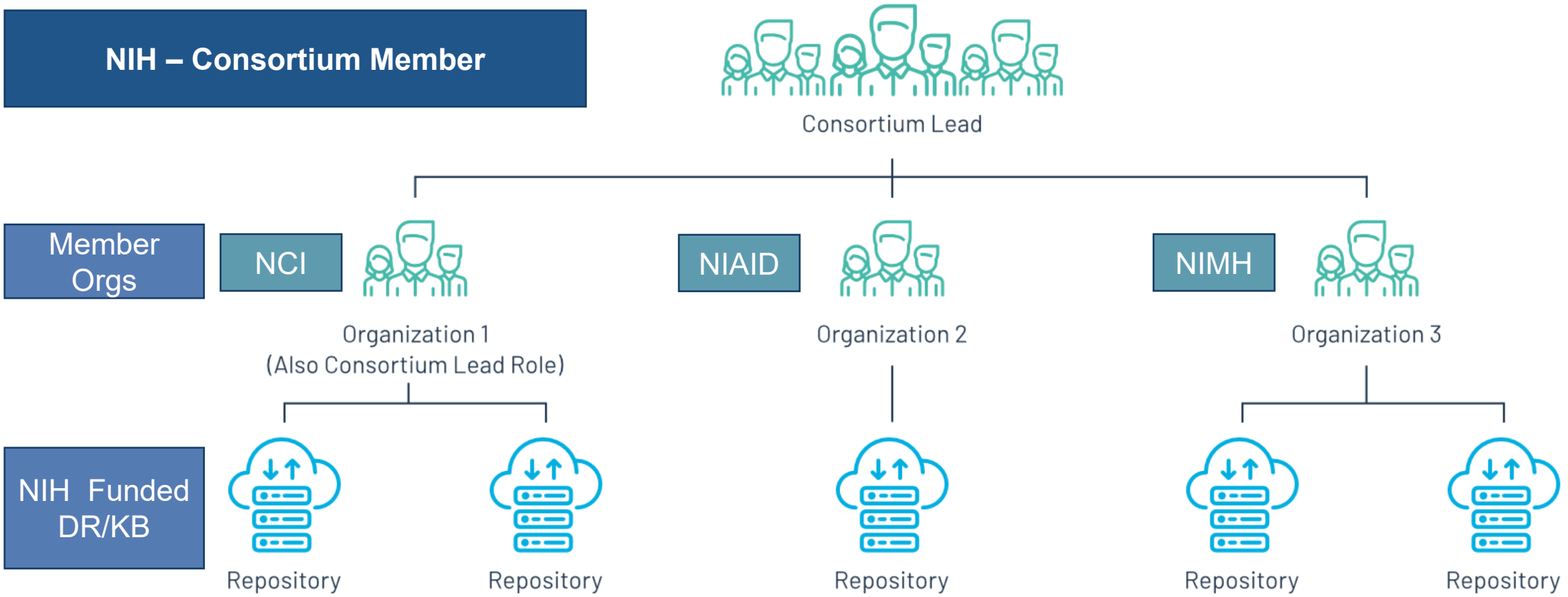
Findable Accessible Interoperable Reusable

TRUST
Transparency Responsibility User focus Sustainability Technology

NHGRI | NIA | NIAID | NICHD | NIDA | NIDCD | NIDCR | NIDDK | NIEHS | NIMH | NINDS | NIGMS

DataCite

NIH became a DataCite consortium member to meet a critical need to mint digital object identifiers, thereby supporting the implementation of FAIR principles for data generated from NIH funded and conducted research.



Generalist Repository Ecosystem Initiative (GREI)

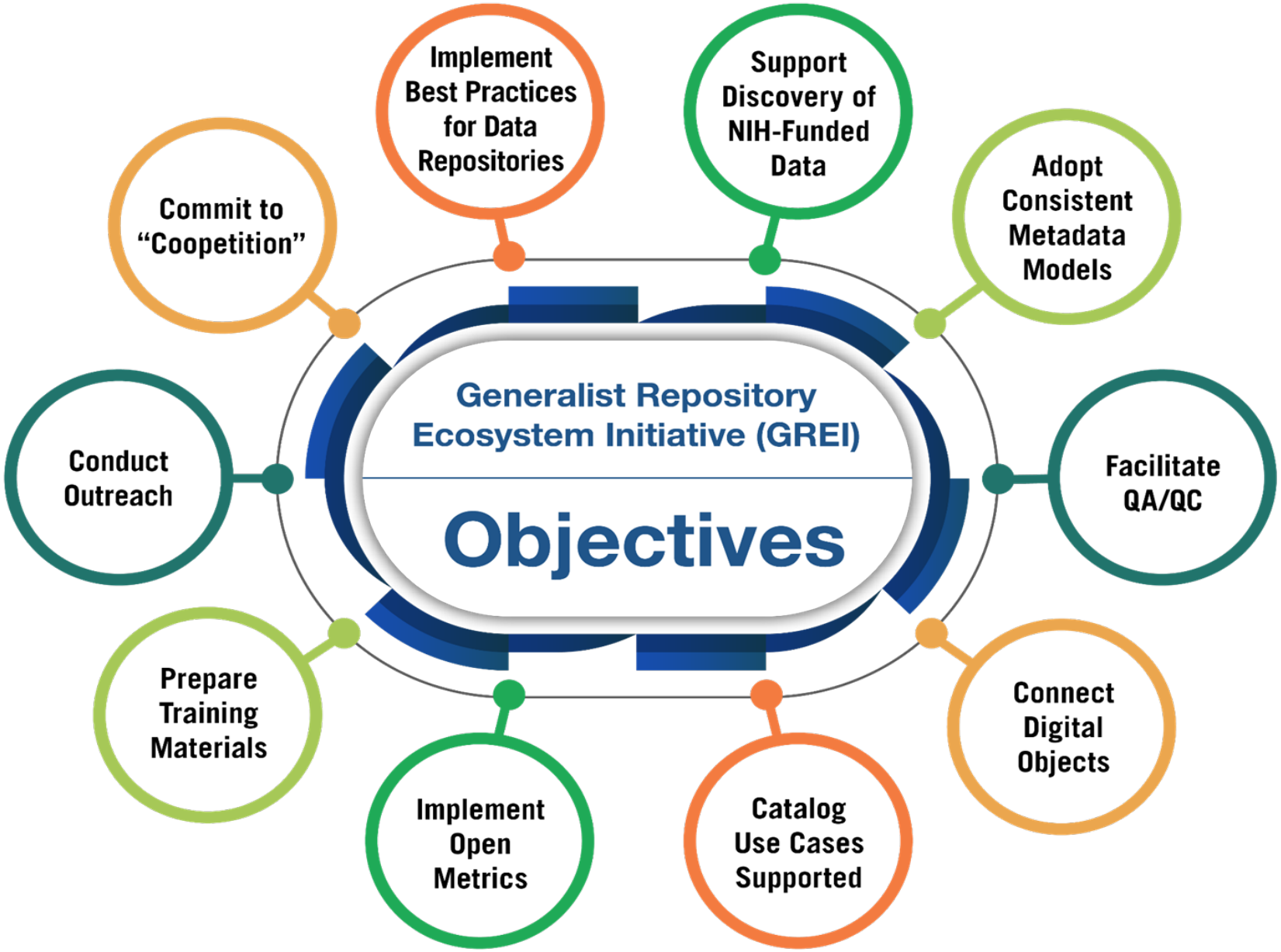
GREI Mission and Goals

NIH Generalist Repository Ecosystem Initiative

The mission of GREI is to establish a common set of capabilities, services, metrics, and social infrastructure; raise general awareness and facilitate researchers to adopt FAIR principles to better share and reuse data.

This initiative will further enhance the biomedical data ecosystem and help researchers find and share data from NIH-funded studies in generalist repositories.

Goals of the Generalist Repository Ecosystem Inititative

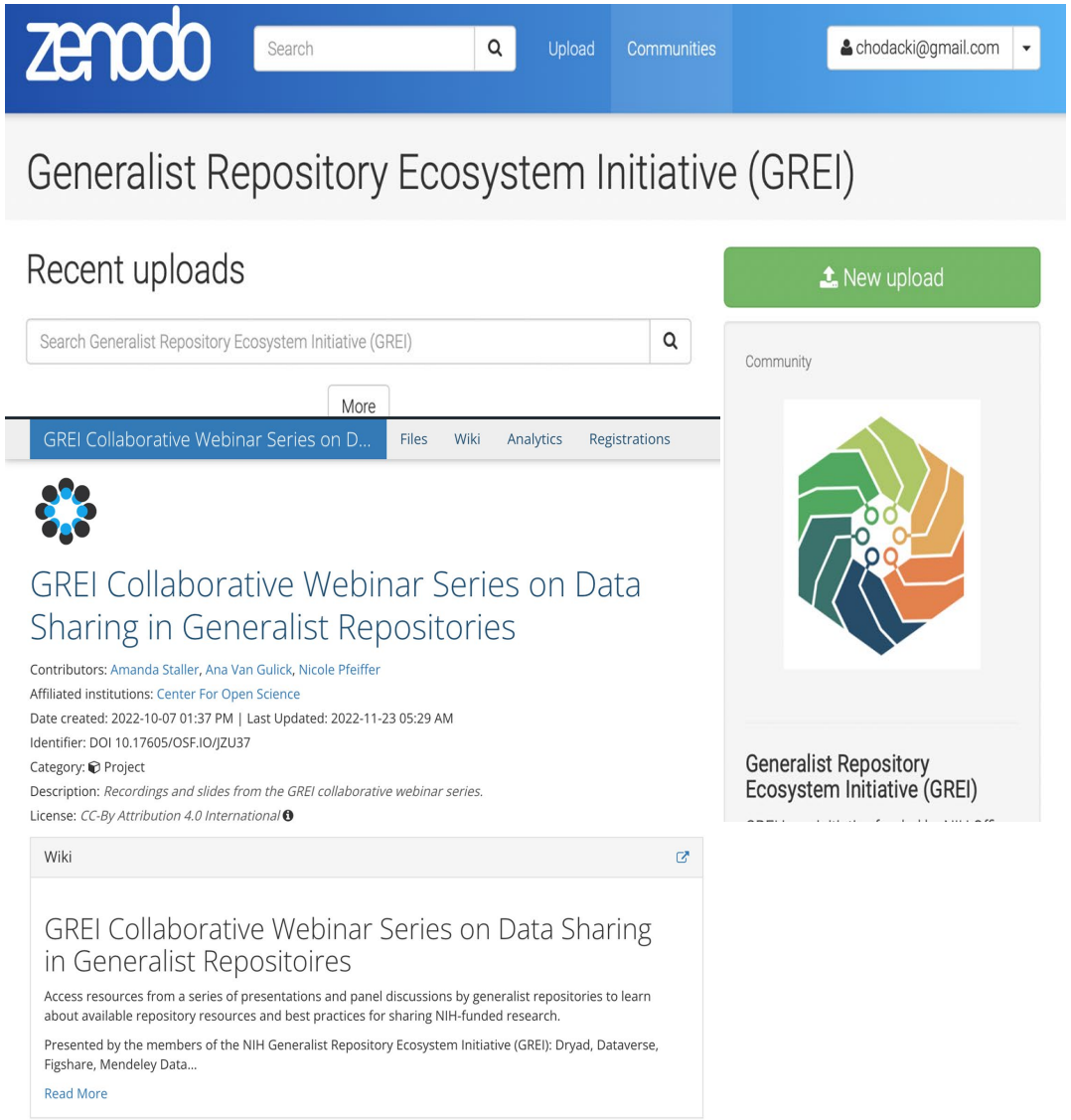


https://datascience.nih.gov/data-ecosystem/exploring-a-generalist-repository-for-nih-funded-data

GREI Accomplishments

- Hosted four trainings with over 1,100 participants
  - Introduction to Generalist Repositories for NIH Data Sharing
  - Meet the GREI Generalist Repositories
  - How to Include Generalist Repositories in Your NIH Data Management and Sharing Plans
  - Best Practices for Sharing Data in a Generalist Repository
- Launched Make Data Count, a global, community-led initiative focused on the development of open research data assessment metrics.
- Repositories created search by funder and grant ID metadata fields and established a repository community.

Connecting with GREI



- GREI Collaborative Webinar Series: Recordings and Slides available: https://doi.org/10.17605/OSF.IO/JZU37
- GREI Community on Zenodo. New location for slides, whitepapers, etc.: https://zenodo.org/communities/grei/
- NIH Program: grei@nih.gov

…more throughout 2023!

In-Person Data Curation Workshop

DATA CURATION NETWORK

The Data Curation Network Collaborative Learning series will offer an in-person data curation workshop **April 12-13, 2023**, at the NIH Campus in Bethesda, Maryland.

- Free to all selected applicants
- Uses CURATE(D) workflow as a training foundation
- Brings together library data specialists and discipline and functional experts in a peer-to-peer learning environment
- **Specifically geared towards institutions that have limited curation support**

Apply by Feb. 17:
https://bit.ly/DCNworkshop

ODSS Data Sharing & Reuse Seminar Series

Highlighting exemplars of data sharing/reuse monthly on 2nd Friday

Past Speakers:

Karen E. Adolph, PhD
Databrary: Secure and Ethical Sharing of Research Video as Data and Documentation

Purvesh Khatri, PhD
Adventures of a Data Parasite: Accelerating Clinical Translation Using Heterogeneity in Public Data

Alexander Ropelewski
The Brain Image Library: A Resource for Sharing Microscopy Data

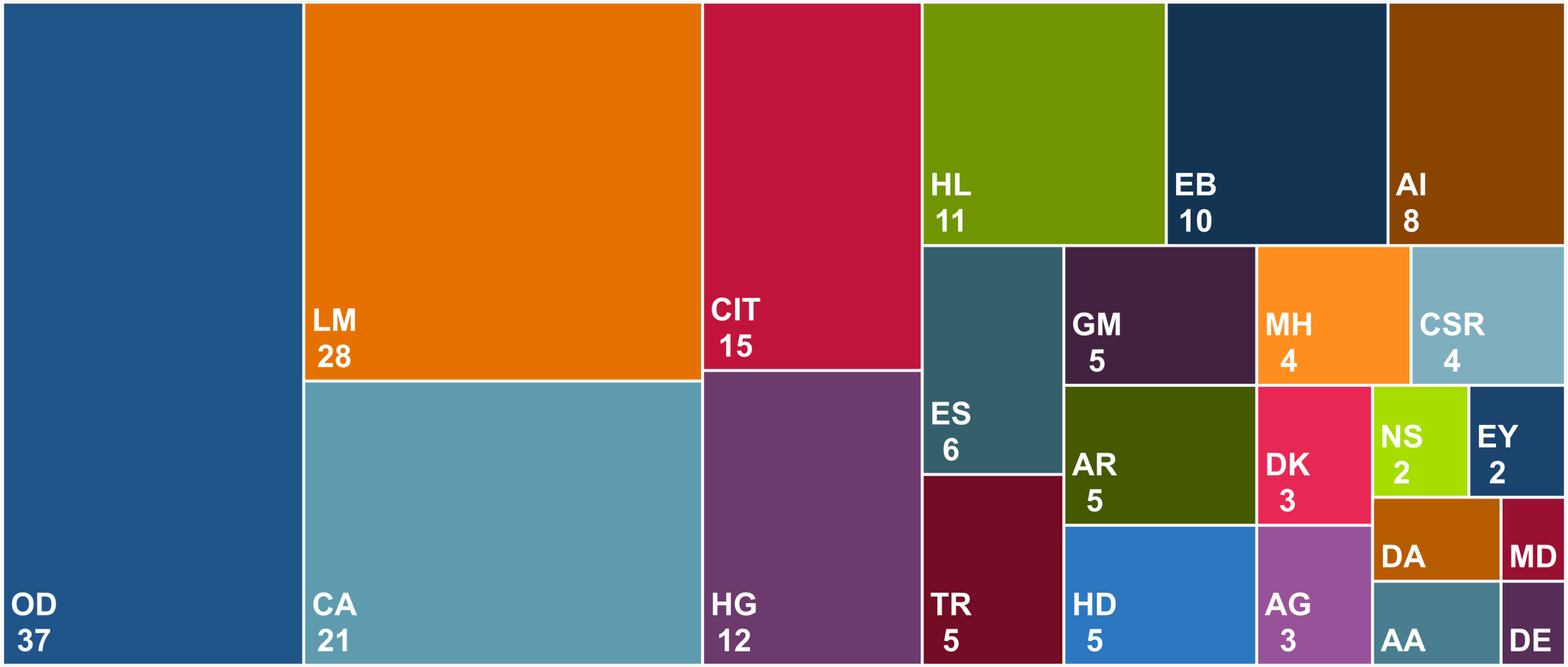
https://datascience.nih.gov/nih-data-sharing-and-reuse-seminar-series

Future Opportunities

- Support capabilities that will develop and adopt common services, tools, and standards, for more connected NIH data systems
- Integrate ethics, transparency, and bias in the development of data science methods, and tools, including AI/ML
- Support community driven and open standardization methods, tools, workflows to ensure interoperable FAIR data and software
- Enhance participants ability to share their data, while preserving anonymity, and streamline researcher access to de-identified health and clinical data

Catalyzing Data Science Across NIH

More than 190 NIH staff from 23 ICOs contributed to these activities



Thank you for your time and attention