

NIH Strategic Plan For Data Science

2025–2030



Introductory Letter



Biomedical research relies on an interplay between scientific observation, hypothesis development, experimental design, data analysis, and interpretation. The increasingly powerful data resources produced by NIH-funded research together with a rapid proliferation of data science technologies available to modern biomedical and behavioral research have advanced our understanding of living systems and has led to medical breakthroughs. But we can do so much more. Currently, optimal use of data for research remains limited by challenges with managing, sharing, and integrating data and technology in a usable, secure, and equitable way. But more inspiring than daunting, is the opportunity that data science offers to dramatically accelerate improved health outcomes for all.

With its mandate to galvanize the energy, talent and creativity of the broadest possible research community, the National Institutes of Health (NIH) must play a critical role in establishing and maintaining a Biomedical Data Ecosystem that moves beyond data collection and sharing to actively fostering effective and ethical data use. The 2025 - 2030 NIH Strategic Plan for Data Science provides a bold vision for the future of data science at the NIH. Five goals — built on the solid foundation laid by the initial data science strategy and rooted in NIH's mission to seek fundamental knowledge to protect and enhance health — prepare NIH to adapt and advance in a changing data science ecosystem.

This strategic plan supports the vision of NIH to enhance data-driven discovery to advance biological knowledge, better characterize the health and health outcomes of all people and foster the development of new technologies and care delivery approaches. Moreover, this strategy encourages the scientific community to share data and to integrate data science practices into their research. But this work must extend beyond the mainstream scientific community. For example, each person has something important to contribute to advancing health by allowing their personal health data to be used for research. NIH must therefore earn and maintain trust from communities with different backgrounds and experiences through engagement and empowerment of their contributions to data and data science. Importantly, the plan addresses this imperative by emphasizing cross-disciplinary collaborations, including public-private partnerships and researcher-clinician-community partnerships to find opportunities and solve challenges together.

I'm thrilled to introduce this new strategic plan that will carry us forward through the next five years. This plan embraces unprecedented technological changes like the development and availability of generative AI and growing interest in quantum information sciences. Growth in the diversity and extent of biomedical and behavioral data is equally rapid, with the research community gaining access to powerful new NIH-supported assets like the *All of Us* research program, Human Pangenome, Human Connectome Project and so much more. The NIH Biomedical Data Ecosystem, with its development advanced by this new strategic plan, will bring increasingly effective data and tools that enable the broadest possible research community to contribute to our mission to bring health to all people. It's an incredible time to engage in biomedical and behavioral research and I can't wait to see what the future brings!

Monica M. Bertagnolli, M.D.
Director, National Institutes of Health

Table of Contents

Introduction.....	2
Emerging Opportunities for Biomedical and Behavioral Data Science.....	4
Plan Content and Implementation	8
GOAL 1	
Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing.....	10
Objective 1-1: Support the Biomedical Community to Manage, Share, and Sustain Data.....	12
Objective 1-2: Enhance FAIR Data and Greater Data Harmonization	14
Objective 1-3: Strengthen NIH's Data Repository and Knowledgebase Ecosystem.....	16
Partnerships and Measuring Progress.....	19
GOAL 2	
Develop Programs to Enhance Human-Derived Data for Research.....	20
Objective 2-1: Improve Access to and Use of Clinical and Real-World Data.....	22
Objective 2-2: Adopt Health IT Standards for Research.....	24
Objective 2-3: Enhance the Adoption of Social and Environmental Determinants of Health for Health Equity.....	25
Objective 2-4: Support Cross-disciplinary Training to Empower Clinical Data Science.....	26
Partnerships and Measuring Progress.....	27
GOAL 3	
Provide New Opportunities in Software, Computational Methods, and Artificial Intelligence.....	28
Objective 3-1: Support New Opportunities to Enhance Artificial Intelligence, Including Ethical AI for Biomedicine	30
Objective 3-2: Develop Cutting Edge Software Technologies	32
Objective 3-3: Support FAIR Software Sustainability	33
Partnerships and Measuring Progress	35
GOAL 4	
Support for a Federated Biomedical Research Data Infrastructure	36
Objective 4-1: Develop, Test, Validate, and Implement Ways to Federate NIH Data and Infrastructure	38
Partnerships and Measuring Progress.....	41
GOAL 5	
Strengthen a Broad Community in Data Science.....	42
Objective 5-1: Increase Training Opportunities in Data Science	44
Objective 5-2: Develop and Advance Initiatives to Expand the Data Science Workforce	45
Objective 5-3: Enhance Data Science Collaboration within the NIH IRP	46
Objective 5-4: Broaden and Champion Capacity Building and Community Engagement Efforts.....	47
Partnerships and Measuring Progress.....	49
Endnotes	50
Appendix 1: Accomplishments from the First NIH Strategic Plan for Data Science	52
Appendix 2: Acronyms	54

Introduction

Modern biomedical and behavioral science benefits from the fundamental transformation of basic biological and biomedical experiments and data science-enabled clinical studies that drive new discoveries. Data enable new opportunities for scientific inquiry; hence, this updated National Institutes of Health (NIH) Strategic Plan for Data Science sets a bold vision for the future, one in which data generated during care of individuals and data generated from biomedical and basic research become powerful inputs that enhance our understanding of fundamental biology and enable the development of new clinical treatments and diagnostic technologies.



Data science is a concept, a complex and iterative process that involves working with larger, more complex datasets that often require advanced computational and statistical methods to analyze. These complex datasets include genomics, transcriptomics, proteomics, metabolomics, imaging, and other data that underly basic biological experimentation.

Data science also consists of clinical trial data; real-world data (RWD) including electronic health data, wearable data and geospatial data, and health-derived data; survey data; data from social and observational studies; and data on social and environmental determinants of health. The vision, articulated in this Strategic Plan, supports the NIH Policy for Data Management and Sharing (DMS Policy)¹ and embraces data-driven discovery as a powerful tool to grow biological knowledge, better characterize the health and health outcomes of all people, and fosters ethical use of new methodologies arising from artificial intelligence (AI) and machine learning (ML). Progress

towards the promise of data-driven discovery requires innovation and creativity across the NIH Institutes, Centers, and Offices (ICOs) that is coordinated by and stimulated with the resources of the Office of Data Science Strategy (ODSS). To accomplish the goals set forth with this vision, NIH will address key challenges and outline opportunities relevant to:

- **Generation and dissemination** of Findable, Accessible, Interoperable, and Reusable (FAIR) data in a manner that will foster greater data sharing and add value to NIH research investments.
- **Enacting** cost-effective strategies for sustainable, secure, and accessible biomedical data repositories and knowledgebases.
- **Acquiring and protecting** data obtained from electronic health records (EHRs) and other RWD, including data captured from outside of traditional health care settings, in a way that preserves privacy and promotes participant permission.
- **Promoting** innovation and use of trustable AI approaches that reduce bias and risks, and ensure approaches are FAIR, validated, and explainable.
- **Creating** opportunities for exploration of new technologies and computing paradigms for biomedical research.
- **Decreasing** health and training disparities across institutions, regions, and global partners in data science.

In support of the NIH mission and the goals of the Department of Health and Human Services (HHS) to increase data sharing, modernize data infrastructure, and develop AI capacity, the 2025-2030 NIH Strategic Plan for Data Science articulates the NIH's strategic views, goals, and objectives to advance data science in the next five years. By addressing these challenges, NIH will pioneer robust data governance frameworks, ensuring data integrity, security, interoperability, and accessibility, while promoting cross-disciplinary collaborations that accelerate scientific discovery. The plan also acknowledges the uniqueness of Tribal data in that Tribal nations are sovereign nations, and Tribal data sovereignty will be respected.

The 2025-2030 NIH Strategic Plan for Data Science builds on accomplishments from significant collaborations of NIH ICOs under the initial NIH Strategic Plan for Data Science.² Experiences with public-private partnerships and alignment with activities across the Federal sector demonstrate that the NIH need not solve all data science challenges alone, but rather that NIH can and should collaborate

» Goals

This **Strategic Plan** will prepare NIH to face the acceleration of sophisticated new technologies, address the rapid rise in the quantity and diversity of data and ensure that new data are connected to existing knowledge by accomplishing five overarching goals:

- Goal 1:** Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing 
- Goal 2:** Develop Programs to Enhance Human-Derived Data for Research 
- Goal 3:** Provide New Opportunities in Software, Computational Methods, and Artificial Intelligence 
- Goal 4:** Support for a Federated Biomedical Research Data Infrastructure 
- Goal 5:** Strengthen a Broad Community in Data Science 

with partners to advance solutions that are sufficiently robust to be applied within the biomedical research enterprise. Advancing data science requires new partnerships with groups including, but not limited to, health care delivery systems, private sector industries in technology and pharmaceuticals, Tribes, non-profit patient representative groups and community partners, and other government agencies.

This document includes a summary of emerging opportunities and challenges facing NIH and delineates strategic objectives for each of the five goals. Associated with each strategic objective are suggested implementation tactics and evaluation schemes. Achievement of the vision set forth in this Plan will position the NIH for accelerating discovery in biomedicine and health; improving and growing existing knowledge; mitigating health disparities; improving health equity through more relevant, comprehensive scientific findings; and achieving a workforce of researchers and clinicians sophisticated in the use of data science methods for discovery and care.



Emerging Opportunities for Biomedical and Behavioral Data Science

Significant advances in data science have been made since the initial NIH Strategic Plan for Data Science (Appendix I). For example, NIH has maintained data sharing policies for several decades and has taken a bold step forward with the final DMS Policy,³ which articulates the need to prospectively plan for how scientific data and accompanying metadata will be managed and shared. NIH defines metadata as information intended to make scientific data citable, interpretable, and reusable.⁴ NIH will continue to support data management and sharing capabilities that enable researchers to appropriately share data in ways that reduce barriers and overall cost, and honor Tribal data sovereignty.

New capabilities and resources are needed that enable researchers to improve the automated collection of valuable metadata during the research process. These capabilities and resources should be consistent with community expectations and standards and should enable easier sharing of these data in appropriate repositories. Moreover, new opportunities and guidelines are needed to enhance trustworthy data repositories in a manner that aligns with global community expectations and contains open metrics that illustrate the impact of data sharing. Finally, developing new methodologies to allow for computational interoperability across data repositories and knowledgebases and mechanisms facilitating the connection of new data to existing knowledge will enable greater and more meaningful research to emerge from the underlying data. Acknowledging Tribal sovereignty, we know that advancing these new capabilities and resources can be done in a way that honors Tribal data sovereignty.

Today, there is potential to create federated networks that connect the billions of data points stored in EHRs, other RWD such as wearable data, and clinical trial data obtained from healthcare providers and medical research institutions across the country. However, to maximize the potential of these data to discover new treatments and cures, there needs to be broad adoption of standardized data exchanges and integration. Through the Health Level Seven International (HL7[®])⁵ Fast Healthcare Interoperability Resources (FHIR[®])⁶ specification, certified health information technology (IT) products will have standardized application programming interface (API) capabilities to facilitate health data sharing. Leveraging and building on the FHIR[®] standard to exchange and share not only EHR data, but also phenotypic data obtained from clinical and genomics studies, clinical records and related social determinants of health (SDoH) data, and eventually other data from medical devices and wearable sensors, provides promising new avenues for clinical research. The ability to gather individual health data over time offers tremendous opportunities to accelerate research and medical breakthroughs and enable individualized preventions and treatments and is the vision of the NIH's All of Us⁷ program. Recently, the National Clinical Cohort Collaborative (N3C)⁸ illustrated the power of a collective data initiative. The N3C extracts data in four common models from 77 health systems, which represents more than 230 organizations. Data are harmonized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model on a weekly basis. N3C represents the largest de-

» **Bridge to Artificial Intelligence (Bridge2AI)** will propel biomedical research forward by supporting widespread adoption of AI that tackles complex biomedical challenges beyond human intuition.

The program will help:

- 1 **Generate**
New flagship biomedical and behavioral datasets
- 2 **Develop**
Software and standards to unify data attributes
- 3 **Create**
Automated tools to accelerate the creation of FAIR and ethically sourced datasets
- 4 **Provide**
Resources to disseminate data, ethical principles, tools, and best practices
- 5 **Create**
Training that bridges the AI, biomedical, and behavioral research communities

identified limited datasets for COVID-19 research and uses privacy-preserving linkages to other RWD, such as Centers for Medicare & Medicaid Services (CMS) and mortality data. Advances such as those seen in the All of Us and N3C programs require standardized vocabularies and ontologies that include communities from different areas of biomedical science and medicine. Experiences during COVID-19 also emphasized the importance of using and promoting Common Data Elements (CDEs), as was illustrated in the Rapid Acceleration of Diagnostics (RADx)⁹ initiative which developed a core set of CDEs used across all RADx-funded projects. In addition, RADx's Mobile At-Home Reporting through Standards (MARS)¹⁰ program established a core set of CDEs and a common HL7[®] specification to facilitate standardized public health reporting of at-home COVID-19 test results. CDEs continue to lack standardized semantics and ontologies. As a goal, this updated Strategic Plan for Data Science advocates for the creation of minimal sets of core CDEs, enabled by

“New paradigms in data discovery and knowledge generation that utilize the integrative power of foundational models would enable researchers and citizen scientists to explore and use data to address complex questions involving diverse and heterogeneous datasets.**”**

creating standardized concepts with allowable responses and data representations, that would enable and broaden the use of clinical and health data.

Another challenge for the data science research community is leveraging the power of data linkages for the same individual across multiple data resources in a way that preserves participant privacy, confidentiality, and their preferences for sharing. The situation is further complicated by inclusion of time-dependent participant data collection methods and requires data linkages. Linkage is more easily accomplished if data have compatible standards and data models. Addressing these challenges requires new governance policies for data linkage and approaches to ensure participants' autonomy is respected. Technical capabilities are needed as well to support data harmonization and aggregation across different sources, including new methodology for collecting, integrating, and sharing social and environmental determinants of health data. These challenges open the door for new algorithms that incorporate secure data governance and participant consent, including privacy-preserving computing, generative AI, foundation models, and blockchain methods.

Machine learning, deep learning, and AI technologies hold significant opportunities to advance basic and clinical research and to improve health and healthcare at individual and community levels. Recognizing these opportunities, NIH launched the Bridge to Artificial Intelligence (Bridge2AI¹¹) program in 2022 to produce new flagship biomedical and behavioral datasets that adhere to FAIR principles and integrate ethical considerations (see textbox **Bridge2AI** on page 5). Creating AI-ready data requires the necessary tools to collect FAIR data at the beginning of the research process (FAIR by design-intentional integration of FAIR principles from the beginning of the data lifecycle) and methods that are complementary to the Collective benefit Authority to control, Responsibility, and Ethics (CARE) principles¹² for Indigenous data sovereignty and data

governance. New capabilities should respect Tribal data sovereignty, including Tribal laws, research codes, ordinances, and cultural protocols.

Efforts are also ongoing to utilize cloud service providers for data storage and management and to create interoperable data systems, to enhance biomedical-AI for ethical and unbiased data and algorithms^{13 14} and to create tools that enable researchers to collect, find, and utilize FAIR data and software. Through the Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES)¹⁵ initiative, the NIH partnership with cloud service providers Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure has resulted in significant increases in data storage, data access, and the use of computational data platforms. Many NIH ICOS have leveraged STRIDES and have created cloud-based data repositories. However, much of these data remain siloed and are not utilized to their fullest potential. Addressing this challenge will require NIH to leverage a modern, federated data architecture approach. This approach will enable NIH to create cost-effective and sustainable practices that are tailored to the needs of individual ICOS and will allow researchers to take full advantage of biomedical data in the cloud with shared scientific analysis capabilities at unprecedented scales.

Today, technological innovations in AI and new capabilities to optimize large language models (LLMs) have generated considerable interest in the possibility of AI to recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets. Yet, challenges remain in creating transparent and explainable datasets and models. The need for ethical principles and frameworks, and connecting those principles/frameworks to practice, for developing and using AI in biomedical research remains an important priority and an unmet need. New paradigms in data discovery and knowledge generation¹⁶ that utilize the integrative

» **Digital NIH** identifies new governance and funding approaches as well as capabilities organized by four functional areas: Extramural Research Management, Intramural Basic Research, Intramural Clinical Research, and Administration and Management. Digital NIH is supported by efforts in the five cross-cutting themes:

- » A common architecture with well-defined standards to enable integration
- » Innovative, cutting-edge storage, analytics, and computational infrastructure
- » Increased technical competency of the workforce at all levels
- » Technology to support an anywhere, anytime workplace of the future
- » Risk-based, embedded cybersecurity protections

power of foundational models would enable researchers and citizen scientists to explore and use data to address complex questions involving diverse and heterogeneous datasets.

Beyond AI, other technological breakthroughs are emerging including advances in quantum computing, quantum sensing, privacy-enhancing computing such as federated learning, and privacy-preserving data sharing such as blockchain. Other applications are emerging in scientific fields such as physics engineering, and computer science, as well as in other industries from manufacturing to finance. The challenges of building teams with the right mix of expertise to advance technology and research at the intersection of these fields makes them particularly opportunity rich. For example, leveraging the gains in instrumentation and computation from quantum information systems (e.g., quantum sensors and computers) is providing insights into new molecular properties, improvements in RWD collection, and new instrumentation for exploring biomedical and behavioral phenomena. Expanding NIH investments

in these emerging technologies will better position the biomedical research community and the agency to take full advantage of these and other new capabilities.

To make progress in the next five years, NIH must leverage the exponential growth in the amount and variety of data by developing and using sophisticated approaches to data management and embracing new technologies, including new methods in data reduction and compression for downstream analysis. Cloud computing will continue to play a significant role in the ability to share and utilize scientific workflows at an unprecedented scale. For example, the National Library of Medicine (NLM)'s entire compendium of the Sequence Read Archive database is available on AWS through the Open Data Sponsorship Program. As a result, researchers can search and align over 19 million diverse samples (16.8 petabytes) to answer questions about evolutionary and comparative biology. By taking advantage of current and emerging data and computing technologies from the commercial and public sectors, NIH could realize exabyte-scale data science in the next decade.



Plan Content and Implementation

This updated Strategic Plan for Data Science is organized into five overarching goals, corresponding to strategic objectives and implementation tactics. These goals will ensure that NIH's strategic approach will address concerns to reduce unintended biases, protect participant privacy, honor Tribal sovereignty and increase transparency in AI while collectively improving data and tools for research. This Strategic Plan also supports the DMS Policy by developing new capabilities to streamline data access with renewed emphases on metadata consistency and accuracy, use of CDEs, and support for utilizing community driven schemas and ontologies to enable data discovery across collections and repositories.



These efforts will increase the scientific community's ability to address new challenges in accessing, managing, analyzing, integrating, and making reusable the huge amounts of data being generated by the biomedical research enterprise.

This Strategic Plan also addresses data science gaps in the NIH Intramural Research Program (IRP) and encourages increased and targeted collaborations to realize sharable opportunities for data and software. New to this Strategic Plan are objectives to enhance NIH's ability to leverage AI/ML technologies for biomedical and behavioral research, enhance clinical and health care data for research, and integrate policy, ethics, and health equity into its visions and objectives.

The implementation tactics are a roadmap for how the overarching goals and strategic objectives will be achieved. Details of these implementation tactics will be determined by the NIH Associate Director for Data Science in collaboration with working groups established by the NIH Scientific Data Council (SDC) and NIH Data Science Policy Council (DSPC), in consultation with the NIH ICOS, other federal and international agencies, the research community, the private sector, and other key stakeholder groups. NIH will continually assess and adjust these priorities based on the needs of NIH and its stakeholders and new opportunities in response to new technologies and capabilities.

Through the implementation of this Strategic Plan during the next five years, NIH will:

- **Develop** new programs to support innovative approaches to data curation, harmonization, interoperability, and validation and increase support for communities to develop and implement new CDEs and standards in priority disease areas.
- **Increase** support for research on clinical and healthcare data science, including new methods for privacy protection, participant informed consent, and data governance.

- **Increase** support for developing tools to collect and analyze data from wearable devices and other new RWD technologies.
- **Develop** new research, training programs, and collaborations in AI and bioethics.
- **Provide** new ways for researchers to search, discover, access, and analyze data across resources and enhance the accuracy, validity, transparency, and reproducibility of these capabilities.
- **Engage** researchers and communities in data science training across biomedical, social, environmental, and behavioral disciplines.
- **Provide** researchers with incentives to adopt data science practices and tools.

This Strategic Plan aligns with the recently released document *Digital NIH: Innovation, Technology, and Computation for the Future*.¹⁷ The Digital NIH proposes new approaches to manage and govern NIH technology investments; describes a framework to guide implementation of high-priority, high-value capabilities; and identifies cross-cutting capabilities that will support data science within NIH (see textbox **Digital NIH** on page 7).

This Strategic Plan will provide direction to encourage greater integration of data science to improve access to and use of biomedical and behavioral data that supports strong ethical, transparent, and anti-bias frameworks and respects Tribal data sovereignty. These efforts will increase the scientific community's ability to address new challenges in accessing, managing, analyzing, integrating, and making reusable the huge amounts of data being generated by the biomedical research enterprise.



GOAL 1

Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing

NIH has a long-standing commitment to data management and sharing across two decades of policies with the goal to create and support a data sharing culture. For example, the final DMS Policy emphasizes the importance of good data management practices and encourages data management and data sharing that reflect practices within research communities. NIH also recognizes Tribal sovereignty and the rights associated with Tribal sovereignty around data collection, data management, and data sharing (reflected in the supplemental notice to the NIH DMS Policy, NOT-OD-22-214¹⁸), and the need to provide options for Tribal ownership of Tribal members' data and control of data sharing. Data management and sharing should reflect practices consistent with FAIR principles to be most beneficial to the research community. NIH-supported and NIH-managed repositories are the building blocks of the NIH data ecosystem and one of the primary mechanisms by which NIH makes the results of federally funded data available to the research community and the public. Federally funded data repositories should adopt the Office of Science, Technology, and Policy (OSTP) Desirable Characteristics of Data Repositories¹⁹ and should align with community standards such as the Transparency, Responsibility, User focus, Sustainability, and Technology (TRUST) principles.²⁰

The TRUST principles provide a framework for formalizing the capabilities of a repository to efficiently serve its intended scientific community. Together with the FAIR, CARE, and TRUST principles, the National Science and Technology Council (NSTC) Guidance provides a framework for formalizing the desired capabilities of a repository (see textbox **NSTC Desired Characteristics of Data Repositories** on page 11). NSTC ensures that federal investment in research that results in scientific data is accessible to accelerate biomedical discoveries, advance human health, and maximize America's return in dollars invested in scientific research.

NIH will continue to promote and support researchers' ability to comply with the DMS Policy expectations by providing resources and guidance to researchers.

NIH will also develop new frameworks to handle the needs of modern data science challenges. For example, National Center for Advancing Translational Sciences (NCATS) is developing the Maintainable, Observing, Securing, and Timing (MOST) framework to augment the FAIR and CARE principles. The MOST framework is a new paradigm for management of data "in use" that emphasizes the importance of

maintainable data infrastructure and policies, observing and understanding data as it is generated. This ensures data security compliance during data ingestion, validation, and utilization. This framework has been used to guide several of the largest initiatives such as the Rare Diseases Clinical Research Network (RDCRN)²¹, N3C, and A Specialized Platform for Innovative Research Exploration (ASPIRE)²². In addition, NIH will promote data repository interoperability. NIH seeks to create a FAIR-enabled data ecosystem that will break down data silos and promote greater findability and accessibility of data, thereby preventing unnecessary duplication of efforts and maximizing NIH investments.

» **National Science and Technology Council (NSTC) Desired Characteristics of Data Repositories** are designed to be relevant to all repositories that manage and share data resulting from Federally funded research. The characteristics are organized across three themes:

- **Organizational Infrastructure**
- **Digital Object Management**
- **Technology**

In addition, additional characteristics for repositories storing human data must be able to address privacy protections, confidentiality, and security.



Support the Biomedical Community to Manage, Share, and Sustain Data

The DMS Policy established requirements that emphasize the importance of good data management practices. It also established the expectations for maximizing the appropriate sharing of scientific data generated from NIH-funded or -conducted research, with justified limitations or exceptions. This policy applies to research funded or conducted by NIH that results in the generation of scientific data. The DMS supplemental notice²³ for American Indian/Alaska Native (AI/AN) populations acknowledges Tribal sovereignty and Tribal rights associated with data management and sharing of data collected from AI/AN populations. By requiring researchers to anticipate their needs for managing and sharing scientific data, NIH will ensure that researchers develop data management and sharing plans that include where and how their scientific data will be shared and any anticipated limitations. This forward-thinking policy integrates data management and sharing into the routine conduct of a scientific project, and in the process, NIH aims to shift the biomedical research culture into one in which data sharing and data reuse are the rule rather than the exception.²⁴ NIH is developing resources to support the DMS Policy compliance activities of their funded investigators, including the NIH Office of Extramural Research (OER) sharing website,²⁵ the NIH Biomedical Informatics Coordinating Committee's portal to key data repositories,²⁶ the Network of the National Library of Medicine (NNLM) Toolkit for the DMS Policy,²⁷ and the National Institute of Child Health and Human Development's (NICHD) Data Repository Finder²⁸ to support the development of data management and sharing plans. These efforts should facilitate the coordinated implementation of the policy across NIH to minimize compliance burden on researchers. Supporting the DMS Policy requires a coordinated effort between the Office of Science Policy, OER, Office of Intramural Research, ODSS, and other NIH ICOS. As such, NIH will establish and enhance harmonized guidelines, processes, data sharing tools, and training in data management and sharing and will explore funding and governance models to ensure a sustainable NIH data sharing infrastructure for researchers, NIH staff, and for data stewards and librarians including those at low resourced institutions.

Implementation Tactics

- **Strengthen** the core data management competencies of researchers, data stewards, data librarians, and NIH program and grants management officers with tools and training:
 - For researchers: core data management and sharing competencies.
 - For data stewards²⁹ and data librarians: promote and enhance FAIR data sharing at their institutions.
 - For NIH staff: evaluate and improve data management and sharing practices and plans.
- **Enhance** programs that provide for credit and incentives for sharing data, including working with publishers, academic institutions, and other funding organizations and agencies.
- **Develop** metrics to measure data sharing, reuse, and impact.
- **Establish** a data steward program to guide data sharing and leverage existing activities at the NNLM National Center for Data Services³⁰ and support additional partnerships, including with societies and associations, for training.
- **Support** tools that will assist researchers with the process of preparing, annotating, and sharing their data.
- **Coordinate** implementation processes across NIH to simplify and streamline compliance with data management and sharing policies.
- **Continue** to provide guidance to researchers on cost management and estimating costs for data management and sharing.

“ NIH will establish and enhance harmonized guidelines, processes, data sharing tools, and training in data management and sharing and will explore funding and governance models to ensure a sustainable NIH data sharing infrastructure for researchers, NIH staff, and for data stewards and librarians, including those at low resourced institutions.



Enhance FAIR Data and Greater Data Harmonization

Through enhanced data sharing efforts and the NIH STRIDES Initiative, NIH-supported investigators have generated and made available over 200 petabytes of data in the cloud. This represents a significant amount of biological data including genomics data, clinical study data, phenotypic and ‘omics data, fitness measures and survey data, and data derived from electronic health care systems and SDoH, taking into account protections for AI/AN Tribal data. These data are most valuable when researchers can combine datasets to answer challenging questions such as “What are the physiological changes evident in Long COVID?” or “What is the relationship between obesity and diabetes in populations with inadequate health care and how does this differ across geographic areas?” To address questions such as these requires interoperability or standards for exchanging datasets with supported efforts in common data models, CDEs and adoption of the USCDI+ (United States Core Data for Interoperability) for research. The NIH CDE Task Force includes a governance committee that reviews and provides NIH endorsements to submitted CDEs, which are then deposited into the NIH CDE Repository,³¹ hosted by the NLM, for use by researchers to help share and combine datasets. Other resources for CDEs include the National Cancer Institute’s (NCI) Enterprise Vocabulary Services (EVS)³² and the cancer Data Standards Registry and Repository (caDSR).³³ Operationalizing these harmonization efforts requires both hybrid and automated curation methods, which are critical to ensuring data can be used with powerful data science tools, like AI.

During the past five years, a number of large NIH programs have undertaken an effort to enhance data harmonization, including Helping End Addiction Long-Term (HEAL) Initiative³⁴ (see textbox **HEAL CDE program** on page 15) that is standardizing metadata; NCI’s Cancer Data Aggregator³⁵ that is mapping harmonized data elements to FHIR®, OMOP,³⁶ and other data models; National Institute of Dental and Craniofacial Research (NIDCR)’s FaceBase³⁷ that is providing guides to scientists to produce harmonizable metadata/data; and other large

programs such as *All of Us*, which is harmonizing clinical data to the OMOP Common Data Model. In addition, community efforts such as the Minimal Common Oncology Data Elements (mCODE)³⁸ provides an agreed upon data standard that can be widely adopted and can increase high-quality data for all cancer types. International funders, such as the International Alliance for Mental Health Research Funders, have established a community of funders, medical journals and data measurement experts that are committed to adopting an agreed upon set of common measures for mental health science.³⁹ The National Institute of Mental Health (NIMH) has been central to the implementation of this effort by using its centralized data archive to monitor compliance with the use of CDEs.⁴⁰ NIH applauds these efforts to support the development of standardized outcome measures in basic and clinical research. Creating standardized outcome measures, when appropriate, will allow for unbiased analysis, interpretation, and reporting of results. These standardized measures of behavioral and health outcomes facilitate cross-study comparisons and improve the interpretability and reporting of research findings and translation into evidence-based clinical practice but will need to be balanced with the flexibility that enables innovative clinical research.

To further enhance the data ecosystem, NIH will encourage the use of community agreed upon standard schemas and metadata, enhance automated ontologies and automated curation processes, and create capabilities for greater data discovery and interoperability across multiple data repositories and knowledgebases. When possible, NIH will seek to enhance and build on existing versions of these tools and standards to avoid unnecessary proliferation of standards, which would make it harder for researchers to find and choose the right tools for their work. In addition, NIH will promote the accessibility of data and research tools to a broader range of the scientific community. This objective will inform future activities under the NIH Plan to Enhance Public Access to the Results of NIH-Supported Research.⁴¹

“NIH will promote the accessibility of data and research tools to a broader range of the scientific community.**”**



Implementation Tactics

- **Enhance** abilities to improve data and metadata quality, including quality assurance, quality control, and accessibility (e.g., guidance on data accessibility practices and policies on accessibility of Tribal data).
- **Encourage** usage of open and standardized schemas, ontologies, and data formats.
- **Enhance** processes for ontology use and for enhanced data curation including enrichment of metadata.
- **Create** a minimal set of consistent and computable common data elements, or concepts, with consistent definitions and terminology.

Helping End Addiction Long-Term (HEAL) Common Data Element (CDE) Program

The NIH HEAL Initiative research portfolio spans a broad array of data types that are a rich resource for future studies. The NIH HEAL Initiative’s CDE Program supports the initiative’s [Public Access and Data Sharing policy](#), which requires researchers to develop plans to share their project’s underlying primary data through a repository that is appropriate for the data type and research discipline, and will connect and expose data via the [HEAL Platform](#) (HEAL acknowledges Tribal sovereignty and the rights associated with data management and sharing of American Indian/Alaska Native data expressed in the DMS supplemental Notice⁴²).

To facilitate cross-study comparisons and improve the interpretability of findings, clinical pain research grantees collaborate and agree to use CDEs across nine core pain domains for patient-reported outcomes (PROs).

Strengthen NIH's Data Repository and Knowledgebase Ecosystem

Data repositories and knowledgebases are essential to increasing the information value of the scientific research enterprise and serve as important components of the implementation of the DMS Policy for preserving, archiving, and sharing scientific data, along with the supplemental Notice⁴³, which allows Tribal control and oversight of data management and sharing for their American Indian and Alaska Native populations. As the size and diversity of data collected and stored from biomedical research continues to increase and we transition to a modernized data ecosystem, the need for making scientific research data and information FAIR and the important role of repositories and knowledgebase in bringing this to fruition is even more evident. There is a current need for an organized system linking new data to established knowledge such as found in textbooks. As articulated in the first Strategic Plan for Data Science, NIH makes a distinction between data repositories and knowledgebases as follows:

› **Biomedical data repositories** accept the submission of relevant data from the research community and store, organize, validate, archive, preserve, and distribute data in compliance with the FAIR data principles. Curation focuses on quality assurance and quality control.

› **Biomedical knowledgebases** extract, accumulate, organize, annotate, and link the growing body of information that is related to and relies on core datasets. Curation of information is often required in knowledgebases.

NIH has separately supported data repositories and knowledgebases as valuable assets and recognizes that these unique resources require funding mechanisms and review panels tuned to the needs of data science resources. Data resources and good data management practices are the key to data and knowledge discovery, data integration, and data reuse. To sustain a healthy and productive data resource ecosystem, it is critical that data repositories and knowledgebases:

› **Deliver** scientific impact to the communities that they serve.

› **Employ and promote** good data management practices and efficient operation for quality and services.

› **Engage** with the user community and continuously address their needs.

› **Implement, adopt, or contribute** to openly shared metrics (e.g., usage, utility, and scientific impact).

› **Provide** sufficient metadata and semantic annotation.

› **Support** a process for data life-cycle analysis, long-term preservation, and trustworthy governance.

NIH supports both unrestricted access and controlled access data repositories. Controlled access data repositories manage and share research participant data, at the individual or aggregate/cohort level, in order to respect research participants' privacy and autonomy. Controlled access datasets often have data use limitations requiring NIH authorization for data access, which, although necessary, can pose challenges to accessibility by the research community and access to controlled data in a timely fashion. Controlled access processes are currently labor- and resource-intensive, which limits their scalability. To accelerate research, and to maintain participants' data protections, NIH seeks to streamline and semi-automate controlled access processes by developing, testing, and deploying the use of emerging technological advancements where feasible and appropriate. In addition, NIH seeks to develop common approaches and infrastructure for addressing data management incidents across NIH-supported repositories.

In recent years, NIH has developed a number of capabilities to promote a data ecosystem, including a collaborative approach for data management and sharing with seven generalist repositories (see textbox **Generalist Repository Ecosystem Initiative (GREI)** on page 17) and support for the use of persistent unique data identifiers through a consortium membership with DataCite.⁴⁴ By partnering with DataCite, NIH data resources will be able to enhance data sharing and enable researchers to cite and reuse research outputs. These efforts aim to strengthen data management and sharing by enhancing data visibility, data citation in scholarly publications, data preservation, future data reuse, and data access.



As the size and diversity of data collected and stored from biomedical and behavioral research continue to grow and NIH enhances its modernized data ecosystem, the need for making these research data and information FAIR underscores the important role of data repositories and knowledgebases. Institutional repositories play an important role in increasing the richness of this ecosystem, which NIH is fostering through activities like the Data Management Center of Excellence-led trainings and community building efforts with the Data Curation Network.⁴⁵ Moreover, to maintain the scientific value of data, repositories and knowledgebases are increasingly required to embrace trustworthy principles. The recently formulated TRUST principles provide a framework for formalizing the capabilities of a repository to efficiently serve the intended scientific community. Developing sustainable data resources requires an understanding and use of metrics for evaluating the usage, utility, and

Generalist Repository Ecosystem Initiative (GREI) is a collaborative effort with Dataverse, Dryad, Figshare, OSF, Mendeley Data, Vivli, and Zenodo to:

› Establish a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure

› Raise general awareness and help researchers to adopt FAIR principles to better share and reuse data

The aim of the GREI is to establish consistent metadata, develop use cases for data sharing, train and educate researchers on FAIR data and the importance of sharing.

impact of a given repository. Moreover, promoting equitable access to research products with appropriate security controls, privacy protections, including human subjects' protections, as outlined in the "Desirable Characteristics of Data Repositories for Federally Funded Research" will continue to be central to the NIH goals.

As important in adopting FAIR principles are the principles for the governance of data generated by or specific to American Indians and Alaska Natives (Indigenous data). Indigenous data are intrinsic to Indigenous Peoples' capacity and capability to realize their human rights and reflect the crucial role of data in advancing Indigenous innovation and self-determination. The CARE principles outline goals for Indigenous data governance that reaffirm the principles of Indigenous self-governance and self-determination. As a first step, NIH developed supplemental information to the DMS Policy on "Responsible Management and Sharing of American Indian/Alaska Native Participant Data"⁴⁶ as a result of Tribal consultation. Similarly, international data sharing, especially involving data generated in low- and middle-income countries, should be respectful of regional and population-specific data governance considerations.

The ubiquitous use of data resources in biomedical research, coupled with a greater emphasis on data management and sharing, has greatly amplified the need for NIH to ensure the stability and robustness of widely used data resources. A particular challenge is long-term support for data resources – especially in light of their growing size, complexity, and demands from the research community. With growing concerns about the sustainability of data resources, NIH aims to articulate a coherent framework for their long-term support.

The challenges that NIH faces with respect to the support of widely used data resources are mirrored at the federal and international level. NIH provides the largest amount of support for the most widely used biomedical data resources, with resources managed by NLM serving as an important node in the international biomedical data ecosystem. For this reason, NIH has been involved in a number of efforts including CoreTrustSeal,⁴⁷ Research Data Alliance (RDA),⁴⁸ DataOne,⁴⁹ Open Science Framework,⁵⁰ DataCite,⁵¹ the Wellcome Trust,⁵² and the Global Biodata Coalition.⁵³ In addition, for more than 30 years, NLM has worked globally to preserve data and enable broad data sharing by coordinating with critical resources such as those comprising the International Nucleotide Sequence Database Collaboration,⁵⁴ and continuing to develop relationships with important global actors, such as

the World Health Organization. These organizations provide a platform for the international community to work together to better coordinate the management and sharing of scientific data. Over the next five years, NIH will work closely with these organizations and new efforts to help ensure the long-term sustainability of the global biodata ecosystem that is relied upon by NIH-funded and all other biomedical researchers worldwide. NIH will continue to encourage data resources to partner with service providers to provide essential data science services and capabilities.

Implementation Tactics⁵⁵

- › **Enhance** data repositories and knowledgebases that promote equitable access to all in alignment with the OSTP memo about Desirable Characteristics of Data Repositories for Federally Funded Research.
- › **Enhance** FAIR, CARE, and TRUST capabilities that ensure secure and effective data management and promote data governance and data sovereignty.
- › **Support** methods and programs with Tribal communities to develop Tribal data governance and sharing that recognize Tribal rights in data.
- › **Promote** shared data management practices, utilize open metrics for impact including enhancing data citation practices, and provide guidance on data preservation and long-term data archiving.
- › **Develop** a comprehensive, coherent, and acceptable sustainability framework for identifying and supporting the portfolio of the most widely used and impactful NIH data resources.
- › **Develop** a single policy framework that governs controlled data access repositories and standardized language for institutions and researchers.
- › **Streamline** controlled data access processes across NIH repositories, including greater use of automation.
- › **Develop** a common approach and infrastructure for addressing data management incidents across controlled access data repositories.
- › **Develop** a single approach to help investigators find and appraise the relevance of controlled access data in NIH repositories, which enable metadata sharing.
- › **Enhance** the visibility and use of NIH intramural research datasets and data resources.
- › **Develop** methods to promote computational interoperability and, when possible, harmonization across data repositories and knowledgebases.
- › **Connect** new data-derived knowledge to existing knowledge.

GOAL 1

Partnerships and Measuring Progress

Potential measures of progress for this goal include data resource key performance indicators for both data resources and for individual datasets, quantity and interoperability of databases and knowledgebases, quantity and citations of datasets deposited (over baseline), ability to find datasets across multiple resources, and data lifecycle frequently asked questions (FAQs). NIH will support and engage in partnership and collaboration across multiple stakeholders including RDA, GO-FAIR, biomedical societies, and international partnerships such as with Global Alliance for Genomics and Health (GA4GH), ELIXIR, and Global Biodata Coalition.



GOAL 2

Develop Programs to Enhance Human-Derived Data for Research

Data discoveries that aim to improve human health and underpin new treatments require a wide range of participant data including clinical data, gathered for the broad purpose of clinical research: health care data including medical history, records, and information that is necessary for care and treatment of patients, enhanced through linkages to social determinants of health (SDoH) and environmental determinants of health (EDoH) data. During the last decade, the United States has seen an increase in the generation and usage of these data in research including through efforts such as All of Us. These efforts are enhanced by large scale data collection and curation that utilize agreed upon common standards and data models. While progress has been made, integration of multiple types of RWD with other data sources remains a challenge because the interpretation of health care-derived data for research purposes is highly dependent on the context of the interactions between patients, their healthcare providers, and their health environment.

Enabling the biomedical and behavioral research community to take full advantage of the multitude of health-derived data requires the adoption and integration of health care data standards with research data standards. NIH will continue and expand work with current and new partners from federal agencies, medical institutions, and health IT developers and vendors, where appropriate, to bridge the technology or data gaps between health care settings and clinical research. To enable researchers to gather and integrate data of interest to address health related questions, NIH will improve access to data repositories that hold participant-derived data and will enhance abilities to link RWD from multiple sources, with appropriate informed consents from the participants, and agreements with Tribal Nations where applicable. NIH will support approaches to leverage or build on existing programs, bring new partnerships together to enhance clinical data science, and support cross-training between clinician researchers, data scientists, and other technical experts/stakeholders. A major goal is to increase the use and utility of health care-derived data for research, with proper security and privacy safeguards. To achieve this goal, activities that integrate clinical data and RWD including data from wearables and data originating from health care settings such as mental health, dental, pathology, and ophthalmology settings should be developed.



Improve Access to and Use of Clinical and Real-World Data

The health care enterprise is a rich source of data for biomedical and behavioral researchers. However, methods of and policies for sharing these data with the wider research community differ in complexity from more traditional research settings and from data sharing expectations and approaches. Unique challenges in data quality, privacy and confidentiality, policy, regulatory, and ethical issues associated with health care and administrative data will require considerations for data sharing and its uses. Consent for collecting, using, and sharing these data is essential for respecting participant rights and contributes to public trust in the research enterprise. NIH will increase capabilities for consent processes and transparency in how participant data is used in research, which may include participant preferences for sharing data, along with Tribal consent where applicable. This is particularly pertinent to the specific challenges for data science in clinical use cases to build trust, explainability, and transparency into the systems and processes leveraging participant data. These activities are consistent with and building on recent NIH guidance and templated informed consent language⁵⁶ for secondary research use of data and specimens beyond addressing transparency and consent, it is equally important to ensure participants see tangible returns from their involvement in research. This includes offering participants access to individualized or aggregated study results, health insights, or tools that allow them to engage with their own data. Moreover, engaging participants in data governance processes, including their preferences for sharing data, ensures that their voices help shape how their data is used, shared, and protected. These steps not only foster trust but also empower participants as co-creators in the research process, aligning with NIH's mission to enhance health equity and promote inclusive research practices.

In addition, wearable device data require substantial efforts to extract, transform, and structure the data in order to reduce the risk of exposing personal information and improve its suitability for research.

There are several existing models for sharing health data with researchers: independent hospitals forming networks or consortia to exchange data with each other and with select external researchers; professional societies engaging with their member institutions and membership to establish data sharing agreements and new channels for data sharing (e.g., National Institute of Biomedical Imaging and Bioengineering (NIBIB)'s Medical Imaging and Data Resource Center (MIDRC));⁵⁷ data enclaves or secure networks that support federated learning where computational tools can be sent and data can be stored or disseminated without the need for data exchange; and NIH-supported enclaves (e.g., NCATS N3C and *All of Us*). Although each approach comes with benefits and challenges, all are important components of the NIH data ecosystem. NIH is committed to improving data FAIRness, transparency of data governance and stewardship expectations, and requirements for accessing and using data derived from care.

In understanding the relationship between health, environment, culture, and lifestyle, researchers are finding that linking and combining individual-level health data with other RWD and digital sources improves our understanding. Integrating multi-modal data—such as genomic, clinical, wearable, and SDoH data—into AI systems provides a more comprehensive and diverse dataset, which can help identify and mitigate biases that arise from reliance on any single data source, ensuring more equitable and accurate insights across populations. However, challenges remain in developing multi-modal data from richly characterized research participants. In addition, linked data provides greater opportunities for researchers to study epidemiological factors. For example, the National Eye Institute (NEI) recently articulated the need to include vision-specific data missing from large-scale research efforts, such as the NIH *All of Us* Research Program and the Genotype-Tissue Expression Project (**NEI Strategic Plan**).⁵⁸ Similarly, new and improved sources of environmental data continue to emerge from industry,



Implementation Tactics

- **Enhance** methods for consent, including consent standards that facilitate data sharing, in cases where data are combined from multiple sources and/or combined over longitudinal studies, with additional considerations for populations with health disparities.
- **Enhance** methods for obtaining and giving research participants power to manage permission for use of their data, including the ability to opt in or out of the use of all or some of their data for research.
- **Create, test, validate, and adopt** accompanying security and privacy-preserving methods and tools to protect research participants while enabling researchers to use multi-modal and digital data combined from multiple sources including through partnerships with other agencies, where appropriate.
- **Establish and promote** standards for new types of health data, such as data captured from home health care devices.
- **Support** AI tools to extract data from electronic medical records and integrate it with real-world and clinical trial data, with proper safeguards for patients' privacy, security, and confidentiality.
- **Support** pilot projects that demonstrate how to combine sparse, multi-modal data into predictive models, potentially using transformer models.
- **Enable** federated frameworks that will allow sensitive data to be utilized in clinical research, including fostering data linkages and interoperability across existing NIH-supported RWD platforms.
- **Develop** ethical, governance, and policy frameworks to guide data linkages in different use case scenarios.
- **Leverage** existing agreements and promote infrastructure to create avenues for researchers to use and access health care and administrative datasets, enhancing participant awareness and consent of data use, especially for vulnerable under-resourced populations.

» OBJECTIVE 2-2

Adopt Health IT Standards for Research

Data sharing is essential to expedite the translation of research into knowledge, products, and procedures that will improve human health and accelerate the development and improvement of treatments for diseases. While there may be benefit to biomedical and behavioral research in connecting and sharing the billions of data points stored in EHRs and clinical trial records across thousands of healthcare providers, there are significant challenges in making use of these data for research. For example, these data lack consistency in standardization. NLM maintains the Unified Medical Language System **UMLS**⁶¹ to distribute key terminologies, classification, and coding standards, supports for coding systems that are now required for use in certified EHRs (e.g., Systematized Nomenclature of Medicine Clinical Terms (**SNOMED CT**), Logical Observation Identifiers Names and Code (**LOINC**), **RxNorm**, among others), and associated resources to promote more effective and interoperable biomedical information systems and services.

Data sharing has made significant progress in the health care community, in part due to the development and adoption of terminology and exchange standards. In 2020, NIH held a virtual workshop entitled *Advancing the Use of Fast Healthcare Interoperability Resources (FHIR®) in Research*.⁶² This workshop brought together leaders in data science and research from across federal agencies to develop a framework for increasing the use of FHIR® for research (see textbox **FHIR®** on page 24). The workshop discussed the interplay needed between policy and technical advances, the opportunities for FHIR® to expand the sources of data that can be integrated into the larger ‘system of care’ to support both clinical care and clinical research, the opportunity that FHIR® presents to increase data reuse across both clinical care and research settings and is enabling patients to access their own clinical data. In addition, FHIR® or other such systems should facilitate population science and SDoH standards to foster the integration of applied research. FHIR® can also be used in conjunction with mature standards, such as those developed by the Clinical Data Interchange Standards Consortium, that have additional advantages, such as harmonized data elements that do not vary across implementations.

To further advance NIH’s goal to bridge the gap between health care settings and applied and clinical research, NIH will strengthen the use of

ontologies with standard biomedical vocabularies and terminologies (e.g., SNOMED CT, LOINC) and exchange standards such as FHIR®. NIH will partner with health data standards bodies and organizations and other federal agencies that work with health data standards. A successful example of this is RADx’s MARS program, which coordinated with federal agencies (Assistant Secretary for Technology Policy/Office of the National Coordinator for Health Information Technology (ASTP/ONC), Food and Drug Administration (FDA), and Centers for Disease Control and Prevention (CDC)) and test manufacturers to establish HL7® v2 and FHIR® standards for capturing data from at-home COVID-19 tests.

Fast Healthcare Interoperability Resources (FHIR®) standard enables electronic healthcare data exchange through an API. An API is a specified set of protocols and data standards that establish the ground rules by which one information system directly communicates with another. Software developers can seamlessly connect their system to another through a FHIR® API to transmit electronic health data.

Implementation Tactics

- **Implement** agile programs that convene researchers and developers to develop, test, validate and adopt health IT technologies and standards based on scientific use cases and provide feedback based on lessons learned.
- **Promote** development, training, and adoption of FHIR®, USCDI+, and other health IT standards to enable further tools for clinical research and for data exchange in research infrastructure, cohort discovery, and applied real-world research.
- **Partner** with other agencies such as the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health Information Technology (ASTP), large health care systems, health technology groups, and researchers to develop use cases outlining how health data standards can benefit and enhance scientific data analysis.

» OBJECTIVE 2-3

Enhance the Adoption of Social and Environmental Determinants of Health for Health Equity

Technological advances have made a significant impact on positive health outcomes; however, advances have not benefited all Americans equally. Health disparities persist, disproportionately affecting racial and ethnic minority populations, Tribal Nations, individuals of less privileged socioeconomic status (SES), underserved rural residents, sexual and gender minorities (SGMs), individuals with disabilities, and any subpopulations that can be characterized by two or more of these descriptions. It has long been recognized that health and treatment outcomes are not solely determined by clinical procedures but that environmental, behavioral, and social factors also play a crucial role. Environmental risk factors such as exposure to pollutants, air and water contamination, and health impacts from climate change are entangled with SES and may increase health disparities. These factors affect health at both the individual and community levels, as evidenced by the health disparity in many communities when compared against the national health indices for a spectrum of diseases and conditions. Advances in data science can help researchers better understand the social and environmental factors associated with racial or ethnic minority group health outcomes and can lead to more effective interventions. This can be accomplished by adopting standardization, collection, reporting, and leveraging of measures of health determinants in both existing and emerging data sources and fostering appropriate data linkages between clinical research data, SDoH data^{63 64} and EDoH data.⁶⁵ In addition, the National Academies of Sciences, Engineering, and Medicine is developing a vision for data infrastructure for federal statistics and social and economic research in the 21st century.⁶⁶ In this new Strategic Plan for Data Science, there is an opportunity to broaden and enhance consensus-driven SDoH⁶⁷ and EDoH standards for data capture and integration across a variety of systems. NIH will engage communities and stakeholders to develop demonstration projects, real-world pilots and use cases to identify and implement SDoH and EDoH data and CDEs for specific diseases/conditions. This objective is aligned with the NIH UNITE⁶⁸ initiative to facilitate new research in health disparities and minority health research (HD/MH).

“ Advances in data science can help researchers better understand the social and environmental factors associated with racial or ethnic minority group health outcomes and can lead to more effective interventions. ”

Implementation Tactics

- **Identify** individual and contextual-level SDoH/EDoH of health data and their associated value set.
- **Support** demonstration projects to test how best to capture SDoH/EDoH of health data for interoperable electronic data exchange.
- **Develop** infrastructure and tools for extracting structured and unstructured SDoH/EDoH from multiple sources and enable iterative models to include SDoH /EDoH in training.
- **Enable** linkage of SDoH/EDoH with other data such as clinical, RWD, wearable sensor, ‘omics data, and administrative data and develop demonstration projects to show technical feasibility of such linkages when appropriate and when there are no increase risks of reidentification for small communities.
- **Support** real-world pilots to integrate social and environmental determinants with clinical CDEs.
- **Support** the collection and use of repeated measures of individual and community-level SDoH/EDoH over time, across the lifespan, or during each care event linked to EHR and health records data to improve disease surveillance, prediction, and population health interventions.
- **Support** training programs and activities for underrepresented groups to expand use of SDoH/ Behavioral/EDoH data models and data collections.



» OBJECTIVE 2-4

Support Cross-disciplinary Training to Empower Clinical Data Science

NIH recognizes that to maintain and enhance clinical research informatics as a career path requires not only clinical training but also training in informatics, analytics, ethics, and standards. This training will focus on appropriate use of data generated from clinical, health care, and real-world settings to better understand the regulatory and policy standards in the generation and use of these data. Equally important is the need to provide health science training to individuals with strong backgrounds in data science. Cross-training between data scientists and clinical researchers would pave the way for interdisciplinary research and could help to reach across new research areas (**National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Strategic Plan⁶⁹**, **National Institute of Neurological Disorders and Stroke (NINDS) Strategic Plan⁷⁰**). Data management and data linking requires partnerships across ethics, social, technological and data science fields. Research involving linking multiple data types, and exposure to new opportunities in technologies, will require a diverse cadre of colleagues for future collaborations. In addition, other health related research fields, including dental and ophthalmology, can benefit from enhanced data science training, with a goal to integrate clinical data, imaging data, and ‘omics data with diverse data types from other health-related fields including the SDoH.

Implementation Tactics

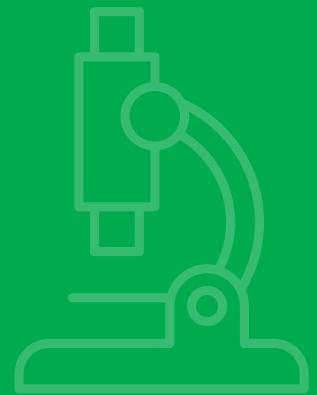
- **Support** cross-training between data scientists, clinical researchers, and nurses engaged in research at various stages of the academic tracks.
- **Develop** training on consent practices and ethical use of data that go beyond legal and regulatory requirements with special considerations for linked/merged data and data from underrepresented communities or Tribal populations.
- **Develop** trainings on data sharing, management, transparency, provenance, and data quality for clinical research.
- **Create** networking opportunities for clinical and data science researchers to develop collaborations, build teams, and learn from experts on these topics.

GOAL 2

Partnerships and Measuring Progress

NIH understands the strength of partnerships and collaborations for innovation in biomedical and behavioral research. NIH will seek partnership and collaboration across multiple stakeholders including the ASTP in implementing relevant data standards, large health care systems and health technology groups, relevant standards development organizations, and to develop collaborations with bioethics organizations. Importantly, NIH will increase and improve opportunities for community engagement and partnership with underserved communities to collectively build tools and frameworks for biomedical and clinical data science uses.

For this goal, potential measures of progress need to address how these activities are advancing biomedical research and include greater use of RWD and SDoH/EDoH data, increased utilization of FHIR® for data exchange including creating or fine-tuning implementations to address research needs, new examples of discovery and harmonization, new or enhanced CDEs for interoperability, and increased use of existing and new standards in clinical and research applications and increasing the number of and reducing processing time for data access requests.





GOAL 3

Provide New Opportunities in Software, Computational Methods, and Artificial Intelligence

Immense amounts of data are generated throughout the biomedical research enterprise from fundamental experiments using cells and research organisms to clinical studies and community-level epidemiological research. These data have value not only for the original research question, but also for secondary data analyses for study replication or for other researchers asking different questions. Harnessing research data for data-driven discovery remains a major challenge that requires attention to data quality, quantity, computability, and standards as well as new methods in computational and AI modeling.

AI/ML are a collection of data-driven technologies with the potential to significantly advance biomedical research. Advances in the field of AI have led to exciting opportunities, including improvements in protein structure prediction and protein design, computer-aided diagnosis on medical images, better understanding of Long COVID phenotypes, and LLMs to interpret clinical, electronic health care records and reports to aid in clinical decision support. AI algorithms can analyze vast amounts of data, identify complex patterns, and gain deeper insights into fundamental scientific phenomena. This approach allows researchers to unlock new avenues for exploration, drive scientific discoveries, and further our understanding of the underlying principles in various fields of study. With the ability to process billions of parameters, AI could significantly improve future health research in recommender systems, rapid annotations including medical and tissue image processing, and the daunting task of organizing large bodies of medical information. However, utilizing AI for biomedical research and health

care practices is still hampered by inconsistent, incomplete, biased, and low-quality data. The task of making data FAIR and AI/ML-ready is not only algorithmic. It requires multi-disciplinary expertise, experimentation and, often, iterative feedback from AI/ML applications and experts. In particular, ground-truth standardization and validation of training datasets is particularly important in biomedical applications where bias and inaccuracies could have misleading and inaccurate results.

Across the federal landscape, AI is seen as a priority and as such the National AI Initiative Act of 2020⁷¹ called on the National Science Foundation (NSF), in coordination with OSTP, to form a National AI Research Resource (NAIRR) Task Force. This Task Force laid out a plan to establish the NAIRR with four measurable goals in mind, namely to 1) spur innovation, 2) increase diversity of talent, 3) improve capacity, and 4) advance trustworthy AI. The roadmap to implementing the NAIRR⁷² calls for an all-of-government approach to leveraging resources,

such as massive compute infrastructures, large data, and a growing talent pool of researchers, to realize this vision.

To enhance the robustness and utility of data analysis and processing methods, NIH will take advantage of new innovations in open and FAIR software and algorithms. NIH will support partnerships to co-design emerging capabilities including new methods in AI including generative AI, computational image analysis, and machine vision; new infrastructures such as quantum information sciences; automated workflows and new tools for researchers to leverage data in a transparent, explainable, FAIR, and ethical manner; and new ways of enabling communities to develop software through collaborative projects. In a world that could unintentionally create a technological divide, NIH must strengthen and diversify its data and software expertise and the technological workforce and make these resources accessible to underrepresented populations in data science and, more broadly, biomedical and behavioral research.

Support New Opportunities to Enhance Artificial Intelligence, Including Ethical AI for Biomedicine

AI (which includes knowledge representation, ML, natural language processing, computer vision and perception, deep learning, and language models) has shown promise in assisting with medical diagnoses and in better understanding of underlying biological processes. Important considerations when using AI methods include transparency; data and algorithm biases; and ethical, legal, and social implications. Making data FAIR and AI/ML-ready also requires interdisciplinary skills. Particularly for biomedical and behavioral research, AI/ML-readiness should be guided by attention to individual and societal impacts of datasets used to train the AI/ML models. While different classes of AI may have unique data requirements, in general AI requires machine readable data that are well described with ontologies and schema so that data can be parsed by the algorithm. Including data quality, such as accuracy, completeness, consistency, and reliability, in AI metadata standards will help address the trustworthiness of the information provided as a result of the AI algorithms. Biases in datasets, algorithms, and applications raise risks and increase potential harms related to privacy, confidentiality, health, and adverse cultural and social impacts with consequences to people, organizations, and communities, particularly for disadvantaged, disempowered, or marginalized under-resourced groups. The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)⁷³ identifies risks and/or potential harms and discusses how managing these risks will lead to more trustworthy AI systems and enable AI developers and users to better understand and take responsibility for the potential limitations and uncertainties in their models and systems.

There are many challenges that hinder the widespread use and deployment of AI/ML capabilities. Training AI/ML algorithms often requires big and diverse datasets; yet, many underrepresented communities have a long history of being absent or misrepresented in existing biomedical and behavioral datasets including clinical, observational, and data

generated in the course of care. Additionally, there is a lack of diversity among researchers, which may unintentionally lead to bias in the design, use, and deployment of AI/ML models in health care and research. Further, non-traditional measures, including SDoH, are important for disease outcomes and health care delivery and, when missing, may adversely affect predictions. SDoH, including lived experiences, social networks, poverty, education level, stress level, access to healthy foods and health care, and exposure to hazards, may play an important role in the diagnoses and treatments of patients from underrepresented communities. Their omission from consideration may lead to fatal outcomes, misdiagnosis, and lack of generalization. Underrepresented communities, which are often disproportionately affected by diseases and health conditions, have the potential to contribute expertise, data, diverse recruitment strategies, and cutting-edge science; and to inform the field on the most urgent research questions; but may lack financial infrastructure, and data science training capacity to apply AI/ML approaches to research questions of interest to them. Recognizing these challenges, NIH launched the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) in 2021 to address health disparities and health inequities using AI/ML (see **textbox AIM-AHEAD** on page 31) and Science Collaborative for Health Disparities and Artificial Intelligence Bias Reduction (ScHARE). Ethical and unbiased data and algorithms are necessary to create safe, secure, and trustworthy AI⁷⁴ for people and for a civil society. According to NIST, trustworthy and responsible AI includes essential building blocks of accuracy, explainability and interpretability, privacy, reliability, robustness, safety, security, and mitigation of harmful bias. This requires the development of assessment frameworks for measuring bias attributes in existing datasets and algorithms across the continuum of AI development and use. Connecting these principles to frameworks that can be used in practice by researchers is essential for continuing to build trust and transparency.

Important goals for NIH are to enhance AI methodology and technologies that expand on the unique opportunities for biomedical and health research; and to ensure that AI/ML capabilities are equitably beneficial across populations in the United States and globally. NIH activities align with and support the White House's Blueprint for an AI Bill of Rights⁷⁵ to ensure that AI algorithms and systems are used and designed in an equitable way. In partnership with the NIH ICOs, the agency will support emerging technologies and AI to integrate multiple streams of data including genomic, nutritional, sensor-based, social and behavioral, exposure, and community-level data to develop explanatory theoretical models, to inform prevention efforts, and to address health disparities (**NICHD Strategic Plan**,⁷⁶ **NIMH Strategic Plan**,⁷⁷ **National Heart, Lung, and Blood Institute (NHLBI) Strategic Vision**⁷⁸). Support to increase opportunities for research communities to include concepts of diversity, equity, inclusion, and accessibility in the development of trustworthy AI-enabled infrastructures and training is a priority.

Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) fosters and supports mutually beneficial partnerships to increase the participation of underrepresented researchers and communities, and build capacity and capabilities of AI/ML in these communities through:

- › Access to high-quality AI/ML-ready data from diverse populations
- › Coordinate federated data approaches and computing infrastructure
- › Train diverse data science workforce
- › Support research questions that connect EHRs, SDoH/EDoH and other related datasets to detect and mitigate biases, develop predictive models, and incorporate community-engaged research

Implementation Tactics

- › **Develop** socio-technical solutions, including guidelines and principles, for ethical AI and to redress biases in training sets (containing ground truth), and algorithms, and support their effective assessment, validation, and adoption.
- › **Establish and operationalize** community engagement for diverse, equitable and inclusive data, methods, and sources for AI.
- › **Support** research in the development, validation, and use of synthetic clinical datasets for AI training and applications, when appropriate.
- › **Facilitate** the development and use of safe, trustworthy, and responsible AI models (including generative AI models) for biomedical and behavioral research and promote validation and benchmarking of AI models.
- › **Develop** tools and training opportunities to help researchers create and prepare data that are FAIR and AI-Ready, including ontologies, schema, and data quality measures.
- › **Support** the development of AI models, with appropriate metadata (model cards) that are explainable, transparent, and FAIR.
- › **Leverage** new technologies and methods for foundational models to accelerate biomedical and behavioral research.
- › **Support** opportunities to develop new AI technologies that will enable the translation of data to knowledge, including AI tools to enable data cleaning, harmonization, integration, and metadata collection.
- › **Enhance** NIH capabilities in AI and responsible use of AI through partnerships across federal agencies and communities to develop new methods, standards, and policies in AI.
- › **Mitigate** the costs to researchers of adopting and responsibly using novel AI technologies.



» OBJECTIVE 3-2

Develop Cutting Edge Software Technologies

NIH is poised to take advantage of the integration of real-world devices, the increased scale of computational resources and significant automation in software and algorithms to advance biomedical discoveries and innovation. For example, new methods that can integrate multidimensional data from a variety of sources including molecular, wearable sensors, environmental, and survey data are needed to develop predictive and actionable models of weight gain, weight loss, and weight loss maintenance and to clarify the role of obesity in the risk, prevention, and treatment of cardiopulmonary and sleep disorders (**NHLBI Strategic Vision**⁹). Multi-dimensional data integration remains a significant challenge for biomedical and behavioral research.

Additionally, low code/no code technologies provide a growing opportunity for trainees and citizen scientists to develop functional applications via ‘drag-and-drop’ software platforms or on the web, with appropriate training. New opportunities to enhance biomedical and behavioral research through the support of digital twinning approaches to model organs, systems, individuals, and populations; new capabilities for privacy-preserving computing and privacy-preserving technologies; and quantum computing should be explored. Finally, considerations for transparency in software and algorithms should be supported.

Implementation Tactics

- › **Adopt and adapt** emerging and specialized methods, algorithms, tools, software, and workflows for biomedical and behavioral scientific discovery, while mitigating the cost of adoption.
- › **Enhance** tools and workflows for greater automation of research and analytical tasks, while maintaining robust ethical standards.
- › **Leverage** new passive and mobile devices and technologies for data collection and analysis with improved practices for informed consent.
- › **Facilitate** FAIR software, with sufficient documentation and metadata, and enhance ethical frameworks.
- › **Leverage** advances in computational methodology to create new opportunities for ethical and social science research.
- › **Investigate** the potential of digital twinning approaches to organs, systems, individuals, and populations.
- › **Explore** opportunities to combine theory-based modeling and simulations with data-driven capabilities.
- › **Promote** opportunities to engage new communities in software development and make these resources accessible, such as through low-code development, to underrepresented communities interested in data science.

» OBJECTIVE 3-3

Support FAIR Software Sustainability

Software is an integral component of biomedical behavioral research due in part to the speed and growth of new technology innovations in the software and computing fields including AI, computer transistors, and microchips. NIH collaborates across 19 ICOS to support the development and enhancement of software tools for open science⁸⁰ by fostering new collaborations between biomedical and clinical scientists and software engineers. For example, significant progress has been made in developing computing models for client-server architectures for data acquisition and progress in developing cloud-based data management and data analytics. Through partnerships with Cloud Service Providers GCP, AWS and Microsoft Azure, NIH has realized over 275 million compute hours for data analysis in the cloud. Yet challenges remain in creating FAIR for Research Software (FAIR4RS).⁸¹ The FAIR4RS software principles, similar to the FAIR Data principles, ensure that software will be usable beyond a single laboratory or investigator. FAIR software principles foster practices to ensure that research software is sustained by larger biomedical research communities over time. NIH recently issued best practices for software sharing⁸² that align with the FAIR software principles.

To develop FAIR and sustainable software⁸³ at a scale beyond single academic laboratories requires multi-disciplinary collaborations from biomedical, computer science, and related fields. Today NIH and other federal agencies and nonprofits are tackling software sustainability head-on, including the NSF program on Cyberinfrastructure for Sustained Scientific Innovation⁸⁴ the recent NIH supplements to support enhancement of software tools for open science, the Schmidt Futures Virtual Institutes for Scientific Software⁸⁵ and the Chan Zuckerberg Initiative’s program for Essential Open-Source Software for Science.⁸⁶ A long-standing program at NIH is NCI’s Information Technology for Cancer Research (ITCR)⁸⁷ program. The ITCR program serves the informatics needs of cancer research continuum and provides support for informatics resources across the development lifecycle (see textbox **ITCR** on page 33). These efforts are critical to building a community around open-source software and NIH intends to

foster such a community to ensure it can achieve the goals laid out in this strategy for sustainable biomedical and behavioral research software.

These programs have a common theme: to enable investigators to adapt and enhance software and tools to take advantage of new technologies and computing paradigms and to optimize research software for robustness and ultimately to increase software sustainability.

Information Technology for Cancer Research (ITCR) supports investigator-initiated, research-driven informatics technology development spanning all aspects of cancer research. The ITCR lifecycle approach includes separate funding in the following areas:

- › **Algorithm Development**
- › **Prototype and Hardening of Software**
- › **Enhancement and Dissemination of Software**
- › **Software Sustainability**

Implementation Tactics

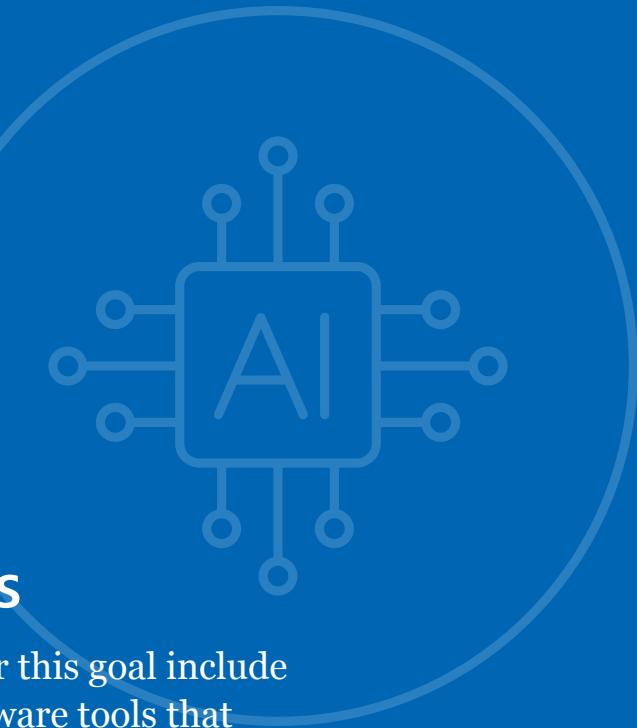
- › **Foster** communities that enhance software development and dissemination.
- › **Improve** visualization tools to support the scale and variety of modern biomedical data.
- › **Establish** metrics and best practices for software sustainability and integrate these into software development lifecycle.
- › **Facilitate** research activities for software engineers and biomedical and computational researchers to collaborate.
- › **Develop** mentorship programs that pair experienced software engineers with early-career researchers and software developers.
- › **Explore** innovative models for public-private partnerships to support software and data innovation and sustainability.



GOAL 3

Partnerships and Measuring Progress

Potential measures of progress for this goal include an increase in the number of software tools that align with the FAIR principles and have a measurably enhanced user experience, increased citation of NIH software across broader communities, software tools that support an increase of use cases across various scientific domains, and integration of tools from other domains into biomedical research. Additional measures of progress include increasing the broader populations' representation among AI developers and participating in auditing AI models to ensure their needs are being met, development of ethical frameworks and tools for software and algorithms, and greater transparency in the development and processing of data and models. NIH will seek partnerships and collaborations with other federal agencies such as NSF and Department of Energy (DOE), for/non-profit organizations, and societies and communities such as the Research Software Alliance.⁸⁸





GOAL 4

Support for a Federated Biomedical Research Data Infrastructure

Throughout the last five years, NIH has seen a remarkable growth in the support for and use of biomedical data repositories and platforms for biomedical research. Today, more and more of these data infrastructures are now moving entirely to the cloud. By moving data infrastructures to the cloud, NIH utilizes advanced cybersecurity controls and scales data management and computation that can take advantage of new technologies while simultaneously creating cost efficiencies and enhancing a positive user experience. The challenge now is to provide greater connections across NIH cloud-based data platforms for easier access to multiple datasets, streamline of similar functions, and enabling more facile analytics.⁸⁹ These resources are cloud-based data infrastructures that provide the research community with data and analytical tools, applications, and workflows in secure environments.

NIH ICOS are also developing data ecosystems including the **NIH Cloud Platform**

Interoperability initiative, which is partnering with the Advanced Research Projects Agency for Health (ARPA-H) Biomedical Data Fabric (BDF) Toolbox program to provide data services to break down data research silos;⁹⁰ the **NIBIB's MIDRC**,⁹¹ which provides open access to 300k+ curated, AI-ready COVID-19 imaging studies and has demonstrated interoperability with BioData Catalyst and the N3C; **National Institute of Allergy and Infectious Diseases (NIAID)** (see textbox **NIAID Data Ecosystem** on page 37); the National Institute on Minority Health and Health Disparities (NIMHD) and the National Institute of Nursing Research (NINR) **ScHARe**, which hosts population science, SDoH, behavioral and environmental datasets to advance health disparity, health care delivery and health outcomes research and foster strategies to mitigate AI biases; the **Common Fund Data Ecosystem**;⁹² and NCBI's **NIH Comparative Genomics Resource**⁹³ which aims to integrate genomic data across all eukaryotic species.

With recent and significant migrations of data resources to the cloud, and the ability to enable petabyte scale data analytics, NIH has the

responsibility to integrate these resources into a federated data infrastructure that leverages ideas from industry and cutting-edge research. The benefit of federating NIH data resources includes: 1) easier access to and use of data across multiple Institutes supported data platforms, 2) economies of scale for NIH to support and maintain shared tools and capabilities, 3) opportunities for communities to collaboratively develop and share new methods and workflows and 4) oversight by the community for greater transparency and autonomy of data use. In collaboration with the NIH ICOS, the agency will support development of innovative data sharing platforms, data analytics, and their integration. This is integral to the missions of each NIH ICO (specific examples found in the **National Institute of Environmental Health Sciences (NIEHS) Strategic Plan**)⁹⁴ and to the overall mission of NIH. The broad use of big data frameworks and FAIR principles, with continued emphasis on partnerships within and outside NIH, will result in new discoveries. NIH will also incorporate the importance of Tribal sovereignty regarding American Indian and Alaska Native data, through Tribal consultation regarding best practices for data management and access processes.



» **National Institute of Allergy and Infectious Diseases (NIAID) Data Ecosystem** enables simultaneous search across 15 infectious and immune-mediated disease and general data repositories based on metadata. The NIAID Data landscape is highly distributed and requires an ecosystem approach that allows for freedom to operate regarding system, syntactic, and semantic interoperability while requiring a minimal set of FAIR-compliant metadata about existing data access protocols used by the repositories. The NIAID approach to the ecosystem is to leverage FAIR metadata to describe data as well as API's and other data access protocols to create a FAIR compliant, interoperability layer on top of a diverse landscape of data, software, and services.

Develop, Test, Validate, and Implement Ways to Federate NIH Data and Infrastructure

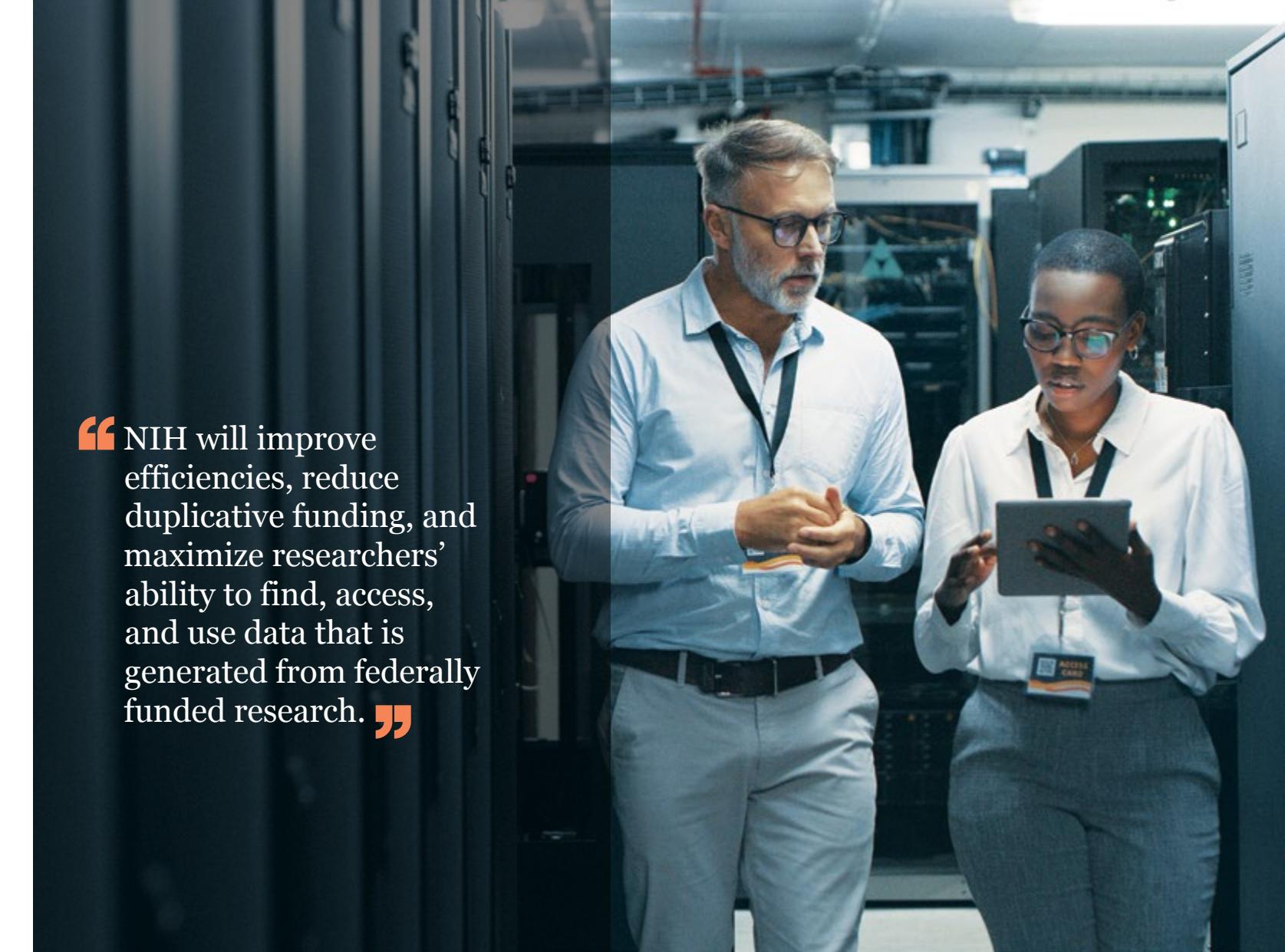
As articulated in the **National Institute on Aging (NIA) Strategic Plan**,⁹⁵ NIH needs to develop comparable databases on health outcomes, risk factors, and determinants of health disparities. An emerging paradigm in data science is data-oriented infrastructure that moves away from a traditionally centralized data infrastructure. A data-oriented infrastructure is distributed between data repositories, or nodes, and has shared capabilities and services to allow for maximum interoperability and economies of scale. This requires a common and coordinated data access process with shared policy and governance. The challenge of a distributed data infrastructure is to create a fabric of harmonized services (e.g., identity and data access management (Authentication [AuthN]/Authorization [AuthZ]), data catalogues, search capabilities, and APIs that are commonly shared, or federated, across the data repositories. In a federated paradigm, NIH ICOs, and organizations supporting biomedical and behavioral research data infrastructures, will control and manage their own data, adopt common processes and interfaces, analysis tools, and services that can be used broadly for biomedicine and behavioral research. In line with this vision, the goal of this objective is to improve efficiencies and maximize researchers' ability to find, access, and use data that are generated from federally funded research and stored in cloud-based data repositories. NIH's vision is to build a connected and federated data ecosystem to ensure that data repositories can be used together rather than in isolation, while honoring Tribal sovereignty when American Indian and Alaska Native data is involved. Several NIH ICOs have collaborated and developed early capabilities including a common approach for researcher's access to controlled access data across a set of data repositories (see textbox **RAS** on page 38), and an approach to implement guidelines and technical standards to empower end-user analyses across participating cloud platforms. In 2020, these interoperability standards were piloted and as a result researchers were able to demonstrate data access

across multiple cloud-based NIH data repositories and perform combined analysis with meaningful results.⁹⁶ These initial piloted efforts are the genesis of a NIH wide federated data ecosystem and are articulated as priorities in the Future Advanced Computing Ecosystem Strategic Plan FY2022 Implementation Roadmap⁹⁷ priorities.

To capitalize on these early successes, NIH will support and enhance a federated biomedical data research infrastructure that will create, test, validate, and implement a set of sharable services (e.g., common search capabilities, APIs, identify and access management (IAM) services, and shared workspaces/sandboxes). NIH will also explore flexible computing infrastructures that allow for greater federation with non-NIH-supported data resources. By doing so NIH will improve efficiencies, reduce duplicative funding, and maximize researchers' ability to find, access, and use data that is generated from federally funded research.

Researcher Auth Service (RAS) Initiative is advancing NIH's data infrastructure and ecosystem. RAS is an identity and data access and management service provided by NIH's Center for Information Technology to facilitate consistent and user-friendly researcher access to NIH's controlled-access data. RAS has adopted the [GA4GH standards](#) for integrating researcher-focused applications and data repositories over the OpenID Connect (OIDC) platform. RAS supports the [FY 2023 Federal Cybersecurity R&D Strategic Plan Implementation Roadmap](#) to protect systems and ensure confidentiality, integrity, availability and privacy of data and [Executive Order on Improving the Nation's Cybersecurity](#). The RAS initiative is advancing data infrastructure and ecosystem goals as defined in the 2018 NIH Strategic Plan for Data Science.

“ NIH will improve efficiencies, reduce duplicative funding, and maximize researchers' ability to find, access, and use data that is generated from federally funded research. ”



Implementation Tactics

- **Enhance** utilization of cloud and hybrid computing architectures and provide opportunities for low-resource institutions to access and utilize NIH-supported cloud capabilities.
- **Support** efficiencies and sharable technologies across NIH data platforms, including increasing new and existing technology industry partnerships.
- **Harmonize** policies for data federation for data use across resources, including expanding RAS to increase researchers' ability to access data and ensure accountability for privacy protection, Tribal data sovereignty, and cybersecurity of systems.
- **Ensure** a robust and connected data resource ecosystem that includes supporting linkages, shared workspaces, and interoperability across NIH-supported cloud platforms for curation, analysis, and sharing of data and metadata.
- **Develop** new capabilities for data search and discovery by enhancing metadata standards, indexing techniques, and improving data interoperability and harmonization.
- **Explore** new paradigms in computing for biomedical and behavioral applications.
- **Develop and implement** guidance for systems that enable sharing and analysis of sensitive human data.



GOAL 4

Partnerships and Measuring Progress

For this goal, the potential measures of progress should advance biomedical research through the use of an integrated infrastructure, include ease of findability of datasets and the ability of researchers to integrate NIH data, and increase ability to use data across the NIH data ecosystem as measured by publications. Additional metrics include guidance and best practices on interoperability so that data, analysis tools and models of biological or population systems can be shared more easily. NIH will require partnership and collaboration across multiple NIH ICOS and scientific organizations such as GA4GH, RDA, and Research Software Alliance, and with technology industries such as Cloud Service Providers.





GOAL 5

Strengthen a Broad Community in Data Science

Advances in data science have the potential to advance our understanding of biomedical and behavioral phenomena but may require the research workforce to adopt new and sometimes unfamiliar technologies and practices. Adoption of new technologies and practices can be burdensome by, for example, requiring substantial training, incurring new costs, or communicating with new partners. Those barriers may fall unequally on the research workforce, exacerbating disparities in access to some of the most promising technologies and stymying NIH's efforts to reduce these disparities. Therefore, NIH must carefully consider how to facilitate adoption of the technologies and practices discussed here and that will be developed in the future. By carefully considering the adoption incentives faced by the research workforce, NIH can maximize progress across fields while continuing to reduce disparities among biomedical and behavioral researchers.

As data science is a necessity in most biomedical and behavioral research, there is a need to develop and nurture data science talent from a diverse array of scientific interests. NIH is committed to growing a stronger and broader community of data scientists including:

- › Data science literate researchers who feel comfortable reading and understanding reported outcomes resulting from data science approaches.
- › Data science savvy researchers who are data science literate and can actively use data science approaches in designing research projects and initiating and/or participating in collaborations with data scientists.
- › Data scientists who are skilled in areas that include bioinformatics, AI/ML, clinical informatics, cloud computing, statistics, computational science, software design and programming, bioinformatics, foundational models, visualization, predictive analytics, modeling and simulation, and ethical data management and sharing.

Following the first Strategic Plan for Data Science, NIH, through ODSS, has worked collaboratively with ICOS on programs that train and educate researchers in data science. These collective efforts will continue to grow, with a particular focus in enhancing the diversity of the data science community so that it better reflects the diversity of the United States. Diversity of backgrounds and scientific area expands the range of research questions, facilitates the translation of scientific data and findings to different communities and helps to build trust in all communities. The time has never been better for all biomedical and behavioral researchers to take full advantage of data science including new innovations from cloud computing, utilizing the availability of significant amounts of biomedical and behavioral data, and new advances in AI/ML. In working to ensure that data science advances in biomedical research can benefit all populations, NIH will help to create a vibrant, innovative, and inclusive data science community.



“Promoting focused data science training for graduate students and postdoctoral fellows will help these early researchers develop into independent investigators with data science acumen.”

» OBJECTIVE 5-1

Increase Training Opportunities in Data Science

Drawing on the foundation of the network of existing extramural training programs, NIH will coordinate across the ICOS to promote data science training and education. In alignment with the Notice of NIH’s Interest in Diversity,⁹⁸ NIH will boost investment in programs that increase the number of underrepresented individuals in data science, including but not limited to racial/ethnic minorities (Blacks or African Americans, Hispanics or Latinos, American Indians or Alaska Natives, Native Hawaiians and other Pacific Islanders), individuals with disabilities, individuals from disadvantaged backgrounds, and women. The aim is to strengthen the support for students and scientists from pre-college through early investigator levels and provide them with a continuum of competitive funding opportunities in data science. Studies show that bright minds may be lost to science long before reaching the college years. Early intervention strategies in research education are therefore necessary to provide a foundation on which essential data skills and visions can develop. Early intervention strategies in research education, such as those supported by the Science Education Partnership Award (SEPA)⁹⁹ at National Institute of General Medical Sciences (NIGMS) and the Youth Enjoy Science program¹⁰⁰ at NCI, are therefore necessary to provide a foundation on which essential data skills and visions can develop. Promoting focused data science training for graduate students and postdoctoral fellows will help these early researchers develop into

independent investigators with data science acumen. In addition, professional and career development support such as access to mentors, soft skills training, and resilience and wellness support are critical to retain the data science trainees in biomedical and behavioral research.

Implementation Tactics

- › **Support** data science training for students and scientists at all academic and career levels from pre-college through early investigators.
- › **Match** training mechanisms, including novel approaches and flexible content, to meet the evolving needs of the biomedical and behavioral workforce.
- › **Enhance** diversity among data science trainees by promoting diversity-focused training and education initiatives.
- › **Increase** pairing of technical data science training with domain-specific knowledge training in NIH training programs.
- › **Increase** the use of hands-on training in new areas such as AI/ML.
- › **Develop** requirements of foundational elements in data science training such as data ethics and cybersecurity.



» OBJECTIVE 5-2

Develop and Advance Initiatives to Expand the Data Science Workforce

Since the first publication of the Strategic Plan for Data Science, significant progress has been made within NIH to enhance its administrative and programmatic data science workforce. For example, in 2020 NIH launched the Data and Technology Advancement National Service Scholars (DATA Scholars) Program.¹⁰¹ These scholars spend one to two years transforming NIH programs by applying cutting-edge methods to health-related challenges. NIH has also implemented the Civic Digital Fellowship¹⁰² program in collaboration with the non-profit organization Coding it Forward to bring to the NIH early-career technologists to spend a summer in data-related projects in NIH program offices. This program successfully supported 80 fellows over four years and provides a solid foundation for NIH to expand to a longer-term program. In addition to these programs, some NIH ICOS have initiated new Offices of Data Science to oversee data management and sharing, the responsible use of data, data science training to staff, and new funding programs in data science. These efforts strengthen the data science workforce within NIH and provide a strong foundation for continued

growth. In the extramural community, NIH will focus on promoting the use of data science approaches for established investigators, enhancing the diversity of the data scientists, and supporting the growth of data science skills among clinician scientists.

Implementation Tactics

- › **Enhance** the diversity of data science investigators and broaden the reach of data science in the biomedical and behavioral research community.
- › **Facilitate** cross-disciplinary trainee programs in data and biomedical sciences.
- › **Enhance** the data science knowledge and skill building for biomedical and clinician scientists including cross-disciplinary skillsets.
- › **Facilitate** recruitment and retention of diverse data science talents at the NIH.
- › **Develop** a pathway for early-career data scientists to join the NIH.



» OBJECTIVE 5-3

Enhance Data Science Collaboration within the NIH IRP

In addition to promoting data science training in the extramural community, NIH will also work to enhance the recruitment of data science trainees from diverse backgrounds in the IRP.¹⁰³ With approximately 1,150 Principal Investigators, more than 2,600 Non-Principal Investigators and more than 5,000 trainees conducting basic, translational, and clinical research, NIH IRP is the largest biomedical research institution and conducts long-term and high-impact science that would otherwise be difficult to undertake. Moreover, NIH supports Biowulf,¹⁰⁴ a high-performance computing system specifically for use by the intramural NIH community. Biowulf is consistently ranked in the top 100-200 of the Top 500 computing infrastructures worldwide and provides access to a wide range of computational applications for genomics, molecular and structural biology, mathematical and graphical analysis, image analysis, and other scientific fields. NIH will build a strong and diverse cohort of intramural data science students and researchers, develop a supportive network for the data science trainees in the IRP and enhance the intramural computational capabilities to realize new opportunities and partnerships not only across NIH, but also with industries and other organizations.

Implementation Tactics

- › **Coordinate** with the NIH Office of Intramural Training and Education to develop a data science-focused intramural cross-disciplinary training program that supports mentored research experiences for postbaccalaureate, post-master's and postdoctoral fellows from diverse backgrounds.
- › **Support** cross-Institute intramural data science projects and enhance interconnectivity of data scientists of all levels.
- › **Enable** federated capabilities, for data and software, within the NIH IRP.
- › **Facilitate** opportunities for intramural researchers to partner with the private sector.
- › **Enhance** NIH's intramural computing environment to utilize new opportunities in cloud computing, AI/ML, and other data science and computing initiatives.

» OBJECTIVE 5-4

Broaden and Champion Capacity Building and Community Engagement Efforts

Developing and sustaining a biomedical and behavioral research workforce that is reflective of the communities being served and supported in an environment that nurtures their success is essential to truly advancing health equity. However, for some investigators and institutions, including Minority Serving Institutions (MSIs) and low-resource institutions, data science challenges remain, including easy access to cloud computing environments, sufficient training and mentoring in data science, and opportunities to apply unique expertise to conduct data science focused health disparities research. NIH is committed to broaden the participation of MSIs and low-resource institutions in the data science community and support efforts to increase human capacity, build partnerships and strengthen research infrastructure.

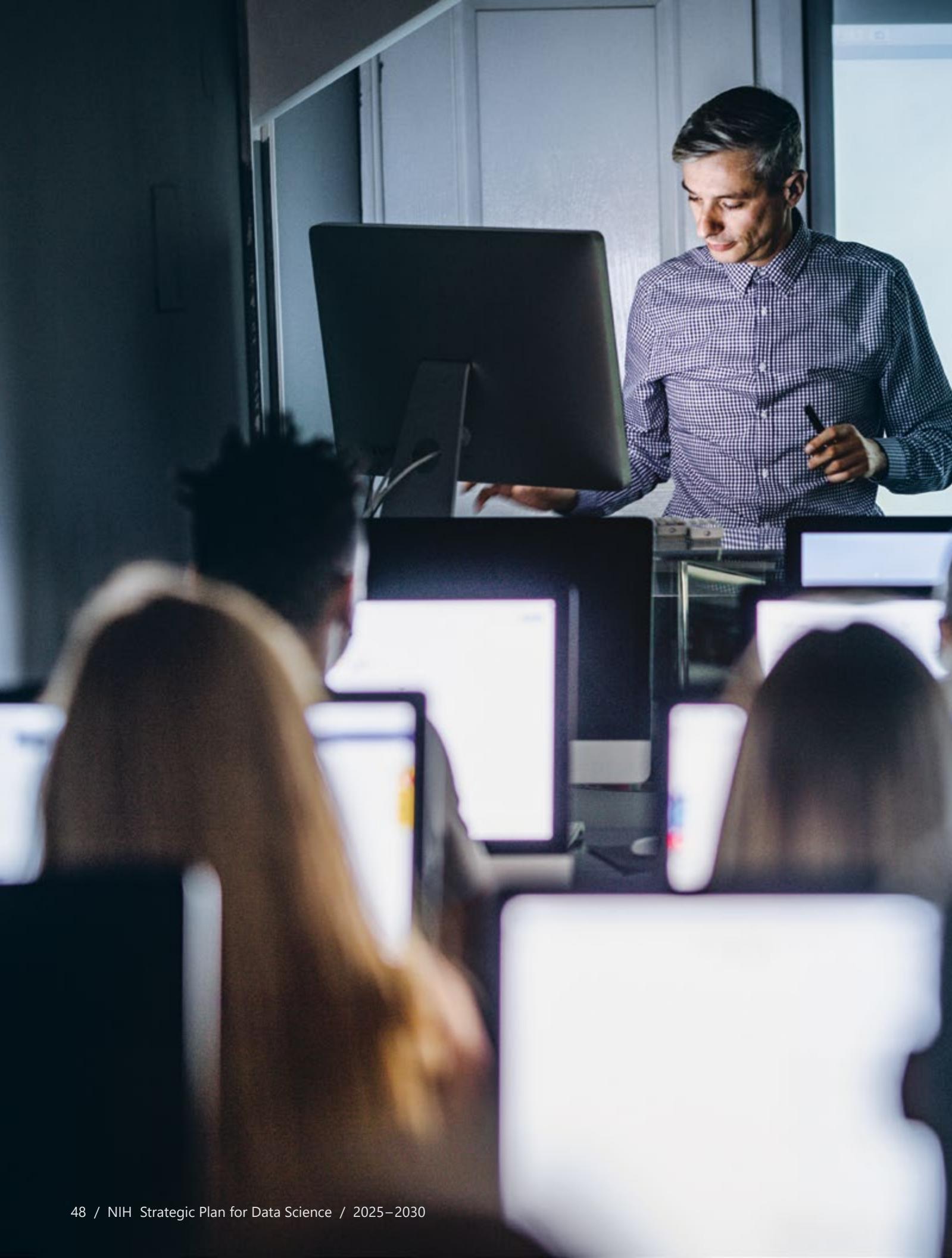
Following the first Strategic Plan for Data Science, new programs have resulted from data science partnerships with NIGMS, including enhancement to the IDeA Networks of Biomedical Research Excellence (INBRE) program (see textbox INBRE on page 47) to support new data science cores and the development of cloud-based learning modules for the NIH CloudLab.¹⁰⁵ Support for National Human Genome Research Institute (NHGRI)'s **Educational Hub for Enhancing Diversity in Computational Genomics and Data Science**, partnership with NIMHD to **Enhance Data Science Capacity Research Centers in Minority Institutions (RCMIs)**, and new data science training efforts in the Ruth L. Kirschstein National Research Service Award (NRSA) Institutional Research Training Grant. Partnership with the NIH Common Fund and Fogarty International Center has enhanced data science capacity in low- and middle-income countries through the Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa¹⁰⁶) program. These efforts offer a platform for research and collaboration as well as a way to inspire interest in aspiring new data scientists. NIH will continue to develop and expand activities and events to attract a wider community.

IDeA Networks of Biomedical Research Excellence (INBRE) fosters the development, coordination and sharing of research resources, and expertise that will expand research opportunities and increase the number of competitive investigators in IDeA-eligible states.

Recently the INBRE program has required a Data Science Core for Biomedical Research. The Data Science Core will provide resources for research, education, and training to expose undergraduate students to data science research and engage a broader community with expertise in biomedical data sciences and related disciplines such as ML, AI, and virtual reality technologies.

Implementation Tactics

- › **Collaborate** with existing NIH programs, such as the Institutional Development Award (IDeA) at NIGMS, the RCMI program at NIMHD and the Partnerships to Advance Cancer Health Equity (PACHE) program at NCI, to develop and expand programs to enhance data science capacity, particularly in MSIs and low-resource institutions.
- › **Leverage** datasets in NIH-supported data repositories and data platforms as training resources.
- › **Build** synergies across government, academic, nonprofit, international, and industry stakeholders focused on data science workforce development and training.



GOAL 5

Partnerships and Measuring Progress

For this goal, potential measures of progress include an increase in the number and diversity of data science trainees from geographically diverse institutions across the United States and the number of trainees leveraging NIH-supported data platforms, increased numbers of trainees who matriculate to data science careers, increased number of data scientists recruited to the NIH and increased numbers of intramural scientists developing and utilizing NIH-supported software. Additional measures may include the products of the trainees and scientists, including publications, patents, models and technologies. NIH will seek partnerships and collaborations with other federal agencies, non-profit organizations, and private sector industries.



Endnotes

- 1 <sharing.nih.gov/data-management-and-sharing-policy>
- 2 <datascience.nih.gov/strategicplan>
- 3 <sharing.nih.gov/data-management-and-sharing-policy>
- 4 <sharing.nih.gov/data-management-and-sharing-policy/>
[data-management](#)
- 5 [hl7.org/](#)
- 6 <datascience.nih.gov/clinical-informatics>
- 7 [allofus.nih.gov/](#)
- 8 [ncats.nih.gov/n3c](#)
- 9 <www.nih.gov/research-training/medical-research-initiatives/radx>
- 10 [nibib.nih.gov/covid-19/radx-tech-program/mars](#)
- 11 [commonfund.nih.gov/bridge2ai](#)
- 12 [www.gida-global.org/care](#)
- 13 [www.nimhd.nih.gov/resources/schare/](#)
- 14 [ncats.nih.gov/funding/challenges/bias-detection-tools-in-health-care](#)
- 15 [datascience.nih.gov/strides](#)
- 16 [datascience.nih.gov/artificial-intelligence/initiatives/ethicalframework2024](#)
- 17 [ocio.nih.gov/digital-nih#:~:text=Digital%20NIH%20represents%20a%20new,over%20the%20next%20five%20years.](#)
- 18 NIH DMS Policy, NOT-OD-22-214 <grants.nih.gov/grants/guide/notice-files/not-od-22-214.html>
- 19 <whitehouse.gov/ostp/news-updates/2022/05/30/nstc-desirable-characteristics-of-data-repositories-for-federally-funded-research/>
- 20 [pubmed.ncbi.nlm.nih.gov/32409645/](#)
- 21 [rarediseasesnetwork.org](#)
- 22 [ncats.nih.gov/research/research-activities/aspire](#)
- 23 NOT-OD-22-214
- 24 [pubmed.ncbi.nlm.nih.gov/34734966/](#)
- 25 [sharing.nih.gov/data-management-and-sharing-policy](#)
- 26 [www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html](#)
- 27 [nnlm.gov/guides/nnlm-toolkit-nih-data-management-and-sharing-policy](#)
- 28 [dash.nichd.nih.gov/](#)
- 29 For this document, data stewards provide guidance on data quality, inclusion of appropriate metadata, appropriate data standard usage, data governance and limitations of data use.
- 30 [nnlm.gov/about/centers/ncds](#)
- 31 [cde.nlm.nih.gov/home](#)
- 32 [evs.nci.nih.gov/](#)
- 33 [datascience.cancer.gov/resources/metadata](#)
- 34 [heal.nih.gov/](#)
- 35 [datacommons.cancer.gov/explore/core-standards-services#cancer-data-aggregator](#)
- 36 [ohdsi.org/data-standardization/](#)
- 37 [facebook.org/](#)
- 38 [health.mitre.org/mcode](#)
- 39 [iamhrf.org/projects/driving-adoption-common-measures](#)
- 40 [grants.nih.gov/grants/guide/notice-files/NOT-MH-20-067.html](#)
- 41 [nda.nih.gov/nda/nimh-common-data-elements](#)
- 42 NOT-OD-22-214
- 43 NOT-OD-22-214
- 44 [datascience.nih.gov/news/nih-joins-datacite-consortium](#)
- 45 [datascience.nih.gov/data-ecosystem/data-management-center-of-excellence](#)
- 46 [datacurationnetwork.org/](#)
- 47 [grants.nih.gov/grants/guide/notice-files/not-od-22-214.html](#)
- 48 [coretrustseal.org](#)
- 49 [rd-alliance.org](#)
- 50 [dataone.org](#)
- 51 [osf.io](#)
- 52 [datacite.org](#)
- 53 [wellcome.org](#)
- 54 [globalbiodata.org/](#)
- 55 Implementation tactics shall not supersede Tribal data sovereignty or Tribal research laws and ordinances.
- 56 [sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/considerations-for-obtaining-informed-consent](#)
- 57 [midrc.org](#)
- 58 [nei.nih.gov/about/strategic-planning](#)
- 59 [osp.od.nih.gov/policies/artificial-intelligence/](#)
- 60 [www.nichd.nih.gov/about/org/od/odss#projects](#)
- 61 [nlm.nih.gov/research/umls/index.html](#)
- 62 [datascience.nih.gov/stoking-flames-virtual-workshop-advancing-use-fhir-research-summary](#)
- 63 Science Collaborative for Health disparities and Artificial intelligence bias Reduction (ScHARe): [nimhd.nih.gov/resources/schare/](#)
- 64 [nimhd.nih.gov/about/strategic-plan](#)
- 65 [niehs.nih.gov/about/strategicplan](#)
- 66 [nationalacademies.org/our-work/toward-a-vision-for-a-new-data-infrastructure-for-federal-statistics-and-social-and-economic-research-in-the-21st-century](#)
- 67 [ahrq.gov/sdoh/data-analytics/sdoh-data.html](#)
- 68 [nih.gov/ending-structural-racism/unite](#)
- 69 [niddk.nih.gov/about-niddk/strategic-plans-reports/niddk-strategic-plan-for-research](#)
- 70 [ninds.nih.gov/about-ninds/what-we-do/strategic-plans-evaluations/strategic-plans](#)
- 71 [congress.gov/116/crp/116hrpt617/CRPT-116hrpt617.pdf#page=1218](#)
- 72 [whitehouse.gov/ostp/news-updates/2023/01/24/national-artificial-intelligence-research-resource-task-force-releases-final-report/](#)
- 73 [doi.org/10.6028/NIST.AI.100-1](#)
- 74 [federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government](#)
- 75 [whitehouse.gov/ostp/ai-bill-of-rights/](#)
- 76 [nichd.nih.gov/about/org/nichd-strategic-plan-2025](#)
- 77 [nimh.nih.gov/about/strategic-planning-reports](#)
- 78 [nhlbi.nih.gov/about/strategic-vision](#)
- 79 [nhlbi.nih.gov/about/strategic-vision](#)
- 80 [datascience.nih.gov/tools-and-analytics/administrative-supplements-to-support-enhancement-of-software-tools-for-open-science](#)
- 81 [pubmed.ncbi.nlm.nih.gov/36241754/](#)
- 82 [datascience.nih.gov/tools-and-analytics/best-practices-for-sharing-research-software-faq](#)
- 83 ADORE software Toolkit - Google Docs
- 84 [new.nsf.gov/funding/opportunities/cssi-cyberinfrastructure-sustained-scientific-innovation](#)
- 85 [schmidtfuture.com/our-work/virtual-institute-for-scientific-software](#)
- 86 [chanzuckerberg.com/eoss](#)
- 87 [cancer.gov/about-nci/organization/cssi/research/itcr](#)
- 88 [researchsoft.org](#)
- 89 Current cloud-based data platforms include the NHLBI's BioData Catalyst, an ecosystem that offers data, analytic tools, applications, and workflows in secure workspaces to accelerate reproducible biomedical research to drive scientific advances that can help prevent, diagnose, and treat heart, lung, blood, and sleep disorders on over 400,000 individuals; NCI's Cancer Research Data Commons (CRDC) which provides secure access to a large, comprehensive, and expanding collection of cancer research data; the Kids First Data Resource which houses data on 44 childhood cancer and structural birth defects cohorts; the National Human Genome Research Institute's (NHGRI) Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) genomic data sharing and analysis platform; All of Us, which includes collected data from over 500,000 participants; the NIH database of Genotypes and Phenotypes (dbGaP) which has controlled access data including with sequence, genotype, and/or phenotype data from over 3.2 million research participants; the NIMH Data Archive (NDA) which provides multiple NIMH Institutes the infrastructure to hold data measured from human subjects and which has data from over 640,000 research participants; and other NIH-supported data platforms.
- 90 [datascience.nih.gov/nih-cloud-platform-interoperability-effort](#)
- 91 [midrc.org](#)
- 92 [commonfund.nih.gov/dataecosystem](#)
- 93 [ncbi.nlm.nih.gov/comparative-genomics-resource](#)
- 94 [niehs.nih.gov/about/strategicplan](#)
- 95 [nia.nih.gov/about/aging-strategic-directions-research](#)
- 96 [pubmed.ncbi.nlm.nih.gov/35199087/aacrjournals.org/cancerres/article/84/9/1396/745074/aacrjournals.org/cancerres/article/84/9/1388/745077](#)
- 97 [nitrd.gov/pubs/FACE-SP-FY22-Implementation-Roadmap.pdf](#)
- 98 [grants.nih.gov/grants/guide/notice-files/NOT-OD-20-031.html](#)
- 99 [nigms.nih.gov/capacity-building/division-for-research-capacity-building/science-education-partnership-awards-sepa](#)
- 100 [cancer.gov/about-nci/organization/crched/about-health-disparities/resources/yes-r25-fact-sheet.pdf](#)
- 101 [datascience.nih.gov/data-scholars-2023](#)
- 102 [codingitforward.com/fellowships](#)
- 103 [irp.nih.gov/](#)
- 104 [hpc.nih.gov/](#)
- 105 [cloud.nih.gov/resources/cloudlab/](#)
- 106 [commonfund.nih.gov/africadata](#)

APPENDIX 1

Accomplishments from the First NIH Strategic Plan for Data Science

GOAL 1

Support a Highly Efficient and Effective Biomedical Research Data Infrastructure

- › NIH Partnership with GCP, AWS, and Microsoft Azure through the STRIDES program has resulted in over 200PB of biomedical data on the cloud, 320 compute hours, 5,000 researchers trained, 1,300 programs working in the cloud and the development of the NIH CloudLab.
- › Development of the RAS for single-sign on and efficient data access across NIH data platforms includes integrating over 30 data programs into RAS, and partnership with Internet2.
- › NIH has made further efforts to connect NIH data platforms through the NIH Cloud Platform Interoperability program with a partnership between NLM, NHGRI, NHLBI, NCI, and the Common Fund, with a result of single sign-on and cross-platform analysis of data.

GOAL 2

Promote Modernization of the Data-Resources Ecosystem

- › NIH has supported funding opportunities for data resources (databases and knowledgebases) that has resulted in 17 new awards across 7 NIH Institutes and Centers (ICs).
- › NIH has supported supplemental funding to existing databases to align with the OSTP characteristics for FAIR data repositories, resulting in 21 awards across 12 NIH ICs.
- › NIH has also launched the GREI, partnering with 7 generalists repositories to establish a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure across these repositories. This initiative conducted a number of webinars with over 1,100 attendees, enabled open metrics in the MakeDataCount Project and created "Search by Funder and Grant ID" metadata fields in participating repositories.
- › NIH has also partnered with DataCite to support the ability to find and cite NIH-funded data, via the use of persistent unique identifiers.
- › NIH has partnered with NLM and the Data Curation Network to provide ongoing training in data management and sharing for researchers, data resource staff, and NIH program staff.
- › NIH has partnered with FASEB to offer the first ever Data Sharing and Data Reuse prize, resulting in over 100 applicants and 12 finalists, with two grand prize winners.
- › NIH has also partnered with HL7® to support training in FHIR® and supported NIH Institutes to leverage FHIR® for clinical data platforms. NIH partnered with the RDA and HL7® to develop and publish a FHIR® implementation guide with 6 real-world use cases by assessing the impact of FHIR® implementation using FAIR data metrics.

GOAL 3

Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools

- › NIH has supported supplemental funding for NIH-funded software, tools, and workflows to develop robust, sustainable, 'cloud-ready', capabilities, resulting in 94 awards across 19 NIH ICs.
- › NIH partnered with NSF on the Smart and Connected Health for AI and data science, resulting in 17 awards across 11 NIH ICs.
- › NIH developed a Software Best Practices document for sharing research software and source code, developed under research grants in any stage of development, in a free and open format.

GOAL 4

Enhance Workforce Development for Biomedical Data Science

- › NIH launched the Data and Technology Advancement (DATA) National Service Scholar Program to bring experts in data and computer scientists and engineers to tackle challenging biomedical data problems with the potential for substantial public health impact, resulting in 17 DATA Scholars across 13 NIH ICs.
- › NIH partnered with the Civic Digital Fellows program to bring 80 Coding-it-Forward fellows to NIH for four consecutive summers.
- › NIH has supported code-a-thons to engage underrepresented communities and increase their participation in data science, including coding partnerships with the African society for Bioinformatics and Computational Biology.
- › NIH has expanded the SEPA, NARCH, INBRE programs to include new initiatives in data science, resulting in 9 new awards.
- › NIH has also expanded diversity supplements in data science to existing grants, resulting in 15 new awards.

GOAL 5

Enact Appropriate Policies to Promote Stewardship and Sustainability

- › NIH published the 2023 NIH DMS Policy, with new training and infrastructure support for the implementation of this policy.

APPENDIX 2

Acronyms

AI Artificial Intelligence	FHIR® Fast Healthcare Interoperability Resources	NAIRR National AI Research Resource	OER Office of Extramural Research
AIM-AHEAD Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity	GA4GH Global Alliance for Genomics and Health	NCATS National Center for Advancing Translational Sciences	OIDC OpenID Connect
API Application Programming Interface	GCP Google Cloud Services	NCI National Cancer Institute	OMOP Observational Medical Outcomes Partnership
ARPA-H Advanced Research Projects Agency for Health	GREI Generalist Repository Ecosystem Initiative	NEI National Eye Institute	ONC Office of the National Coordinator for Health Information Technology
ASPIRE A Specialized Platform for Innovative Research Exploration	HD/MH Health Disparities and Minority Health	NHGRI National Human Genome Research Institute	OSP Office of Science Policy
ASTP Assistant Secretary for Technology Policy	HEAL Helping End Addiction Long-Term	NHLBI National Heart, Lung, and Blood Institute	OSTP Office of Science, Technology, and Policy
AWS Amazon Web Services	HHS Department of Health and Human Services	NIA National Institute on Aging	PACHE Partnerships to Advance Cancer Health Equity
BDF Biomedical Data Fabric	HL7® Health Level Seven International	NIAID National Institute of Allergy and Infectious Diseases	PROs Patient-Reported Outcomes
Bridge2AI Bridge to Artificial Intelligence	IAM Identify and Access Management	NIBIB National Institute of Biomedical Imaging and Bioengineering	RAS Researcher Auth Service
caDSR Cancer Data Standards Registry and Repository	ICs Institutes and Centers	NICHD National Institute of Child Health and Human Development	RCMI Research Centers in Minority Institution
CARE Collective benefit, Authority to control, Responsibility, and Ethics	ICOs Institutes, Centers, and Offices	NIDCR National Institute of Dental and Craniofacial Research	RDA Research Data Alliance
CDC Centers for Disease Control and Prevention	IDeA Institutional Development Award	NIDDK National Institute of Diabetes and Digestive and Kidney Diseases	RDCRN Rare Diseases Clinical Research Network
CDE Common Data Element	INBRE IDeA Networks of Biomedical Research Excellence	NIEHS National Institute of Environmental Health Sciences	RDW Real-World Data
CMS Centers for Medicare & Medicaid Services	IRP Intramural Research Program	NIGMS National Institute of General Medical Sciences	ScHARe Science Collaborative for Health disparities and Artificial intelligence bias Reduction
DMS Policy NIH Policy for Data Management and Sharing	IT Information Technology	NIH National Institutes of Health	SDC Scientific Data Council
DOE Department of Energy	ITCR Information Technology for Cancer Research	NIMH National Institute of Mental Health	SDoH Social Determinants of Health
DSPC Data Science Policy Council	LLM Large Language Model	NIMHD National Institute on Minority Health and Health Disparities	SEPA Science Education Partnership Award
EDoH Environmental Determinants of Health	LOINC Logical Observation Identifiers Names and Codes	NINDS National Institute of Neurological Disorders and Stroke	SES Socioeconomic Status
EHRs Electronic Health Records	MARS Mobile At-Home Reporting through Standards	NINR National Institute of Nursing Research	SGMs Sexual and Gender Minorities
EVS Enterprise Vocabulary Services	mCODE Minimal Common Oncology Data Elements	NLM National Library of Medicine	SNOMED CT Systematized Nomenclature of Medicine Clinical Terms
FAIR Findable, Accessible, Interoperable, and Reusable	MIDRC Medical Imaging and Data Resource Center	NNLM Network of the National Library of Medicine	STRIDES Science and Technology Research Infrastructure for Discovery, Experimentation and Sustainability
FAIR4RS FAIR for Research Software	ML Machine Learning	NRSA National Research Service Award	TRUST Transparency, Responsibility, User focus, Sustainability, and Technology
FAQs Frequently Asked Questions	MOST Maintainable, Observing, Securing, and Timing	NSF National Science Foundation	UMLS Unified Medical Language System
FDA Food and Drug Administration	MSI Minority Serving Institution	NSTC National Science and Technology Council	USCDI United States Core Data for Interoperability
	N3C National Clinical Cohort Collaborative	ODSS Office of Data Science Strategy	



“ This new strategic plan is built on more than five years of investment into the creation of a strong foundation for a modern biomedical data ecosystem. As we look boldly to the future, we will build on that foundation to advance new, sustainable, trustworthy capabilities and enable a strong workforce community that leverages data science to inform discovery and care for all. ”

› **Dr. Susan Gregurick**
*Associate Director for Data Science
and the Director of ODSS*





National Institutes of Health
Office of Data Science Strategy



Visit us online:
bit.ly/4hfXTmz