NIH STRATEGIC PLAN FOR DATA SCIENCE

Introduction

As articulated in the National Institutes of Health (NIH)-Wide Strategic Plan[1], our nation and the world stand at a unique moment of opportunity in biomedical research, and data science is an integral contributor. Understanding basic biological mechanisms through NIH-funded research depends upon vast amounts of data and has propelled biomedicine into the sphere of "Big Data" along with other sectors of the national and global economies. Reflecting today's highly integrated biomedical research landscape, NIH defines data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."

NIH supports the generation and analysis of substantial quantities of biomedical research data (see, for example, text box "Big Data from the Resolution Revolution[2]"), including numerous data sets emanating from fundamental research using model organisms (such as mice, fruit flies, and zebrafish), clinical studies (including medical images), and observational and epidemiological studies (including data from electronic medical records). By 2025, the total amount of genomics data alone is expected to equal or exceed totals from the three other major producers of large amounts of data: astronomy, YouTube, and Twitter.[3] Indeed, next-generation sequencing data, stored at the National Institutes of Health (NIH's) National Center for Biotechnology Information (NCBI), has been growing exponentially for many years and shows no signs of slowing (see Fig. 1, below).

Big Data from the Resolution Revolution

One of the revolutionary advances in microscope, detectors, and algorithms, cryogenic electron microscopy (cryoEM) has become one of the areas of science (along with astronomy, collider data, and genomics) that have entered the Big Data arena, pushing hardware and software requirements to unprecedented levels. Current cryoEM detector systems are fast enough to collect movies instead of single integrated images, and users now typically acquire up to 2,000 movies in a single day. As is the case with astronomy, collider physics, and genomics, scientists using cryoEM generate several terabytes of data per day.

The generation of most biomedical data is highly distributed and is accomplished mainly by individual scientists or relatively small groups of researchers. Moreover, data also exist in a wide variety of formats, which complicates the ability of researchers to find and use biomedical research data generated by others and creates the need for extensive data "cleaning." According to a 2016 survey, data

[1] NIH-Wide Strategic Plan Fiscal Years 2016-2020: Available at: https://www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2016-2020-508.pdf

[2] Baldwin PR, Tan YZ, Eng ET, Rice WJ, et al. Big data in cryoEM: automated collection, processing and accessibility of EM data. Curr Open Microbiology 2018;43:1–8.

[3] Stephens, et al., Big Data: Astronomical or Genomical? PLOS Biology 2015 (July 7, 2015) Available at: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

scientists across a wide array of fields said they spend most of their work time (about 80 percent) doing what they least like to do: collecting existing data sets and organizing data.[4] That leaves less than 20 percent of their time for creative tasks like mining data for patterns that lead to new research discoveries.

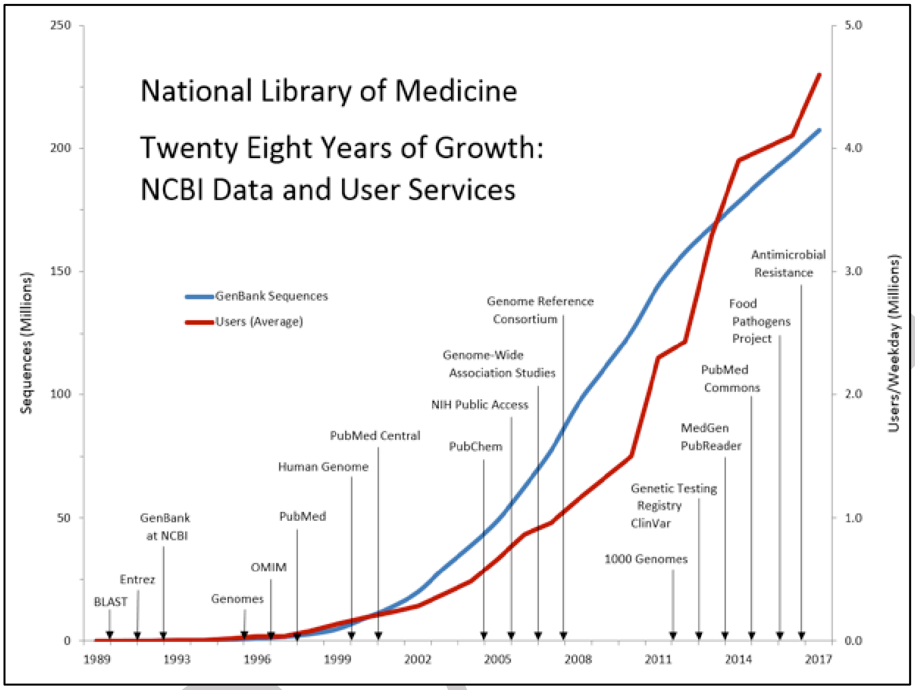


Figure 1. Growth of NCBI Data and Services, 1989-2017 Credit: NCBI

A New Era for Biomedical Research

Advances in storage, communications, and processing have led to new research methods and tools that were simply not possible just a decade ago. Machine learning, deep learning, artificial intelligence, and virtual-reality technologies are examples of data-related innovations that may yield transformative changes for biomedical research over the coming decade. The ability to experiment with new ways to optimize technology-intensive research will inform decisions regarding future policies, approaches, and business practices, and will allow NIH to adopt more cost-effective ways to capture, access, sustain, and reuse high-value biomedical data resources in the future. To this end, NIH must weave its existing data-science efforts into the larger data ecosystem and fully intends to take advantage of current and emerging data-management and technological expertise, computational platforms, and tools available from the commercial sector through a variety of innovative public-private partnerships.

Note: Please see ***Glossary*** *for definitions of terms related to data science.*

[4] CrowdFlower 2016 Data Science Report. Available at: http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

The fastest supercomputers in the world today perform a quadrillion (10^{15}) calculations each second: known as the petascale level. The next frontier is exascale computing (which is 1,000 times faster than petascale, or a quintillion (10^{18}) calculations each second). Reaching exascale-level computing is a technical milestone that is expected to have profound impacts on everyday life. At the exascale level of computing speed, supercomputers will be able to more realistically mimic the speed of life operating inside the human body, enabling promising new avenues of pursuit for biomedical research that involves clinical data. These data-intensive programs may well be among the earliest adopters and drivers of exascale computing: They include the All of Us Research Program and the Cancer Moonshot℠ components of the Precision Medicine Initiative, the Human Connectome project, the Brain Research through Advancing Innovative Neurotechnologies (BRAIN®) initiative, and many others.

Clinical Data and Information Security

Throughout the research enterprise, NIH must continue to balance the need for maximizing opportunities to advance biomedical research with responsible strategies for sustaining public trust, participant safety, and data security. Proper handling of the vast domain of clinical data that is being continually generated from a range of data producers is a challenge for NIH and the biomedical research community, including the private sector. Patient-related data can arise from a wide array of sources, including specialized research projects and trials; epidemiology; genomic analyses; clinical-care processes; imaging assessments; environmental-exposure records; and a host of social indicators now linked to health such as educational records, employment history, and genealogical records. NIH must develop, promote, and practice robust and proactive information-security approaches to ensure appropriate stewardship of patient data and to enable scientific advances stemming from authentic, trusted data sources.

Data quality and integrity must be maintained at all stages of the research life cycle—from collection through curation, use, dissemination, and retirement. It is essential that NIH implement comprehensive security controls consistent with the risk of harm if data are breached or corrupted. NIH must also continually revisit its approaches to keep pace with ever-increasing information security threats that arise in the global information technology environment. This work must be done in close partnership with private, public, and academic entities that have expertise in information security and related areas.

Current Data Science Challenges for NIH

As an initial step to strengthen the NIH approach to data science, in 2014, the NIH Director created a unique position, the Associate Director for Data Science, to lead NIH in advancing data science across the Agency, and established the Big Data to Knowledge (BD2K) program. NIH's past investment in the BD2K software-development initiative launched in 2014 produced a number of tools and methods that can now be refined and made available to help tackle a variety of challenges. These include data-compression formats, suites of algorithms, web-based software, application-programming interfaces (APIs), public databases, computational approaches, among others.

In subsequent years, NIH's needs have evolved, and as such the agency has established a new position to advance NIH data science across the extramural and intramural research communities. The inaugural NIH Chief Data Strategist, in close collaboration with the NIH Scientific Data Council and NIH Data Science Policy Council, will guide the development and implementation of NIH's data-science activities and provide leadership within the broader biomedical research data ecosystem. This new leadership position will also forge partnerships outside NIH's boundaries, including with other federal and international funding agencies and with the private sector to ensure synergy and efficiency, and prevent unnecessary duplication of efforts.

As a result of the rapid pace of change in biomedical research and information technology, several pressing issues related to the data-resource ecosystem confront NIH and other components of the biomedical research community, including:

- The growing costs of *managing* data could diminish NIH's ability to enable scientists to *generate* data for understanding biology and improving health.
- The current data-resource ecosystem tends to be "siloed" and is not optimally integrated or interconnected.
- Important datasets exist in many different formats and are often not easily shared, findable, or interoperable.
- Historically, NIH has often supported data *resources* using funding approaches designed for *research* projects, which has led to a misalignment of objectives and review expectations.
- Funding for tool development and data resources has become entangled, making it difficult to assess the utility of each independently and to optimize value and efficiency.
- There is currently no general system to transform, or harden, innovative algorithms and tools created by academic scientists into enterprise-ready resources that meet industry standards of ease of use and efficiency of operation.

As a public steward of taxpayer funds, NIH must think and plan carefully to ensure that its resources are spent efficiently toward extracting the most benefit from its investments. Because of these issues, NIH has adopted a unified vision, and a corporate strategy for attaining that vision, that will best serve the biomedical research enterprise in the coming decades.

Plan Content and Implementation

This document, the *NIH Strategic Plan for Data Science* describes NIH's Overarching Goals, Strategic Objectives, and Implementation Tactics for modernizing the NIH-funded biomedical data-resource ecosystem. In establishing this plan, NIH addresses storing data efficiently and securely; making data usable to as many people as possible (including researchers, institutions, and the public); developing a research workforce poised to capitalize on advances in data science and information technology; and setting policies for productive, efficient, secure, and ethical data use. As articulated herein, this strategic plan commits to ensuring that all data-science activities and products supported by the agency adhere to the FAIR principles, meaning that data be Findable, Accessible, Interoperable, and Reusable (see text box “What is FAIR?”).[5]

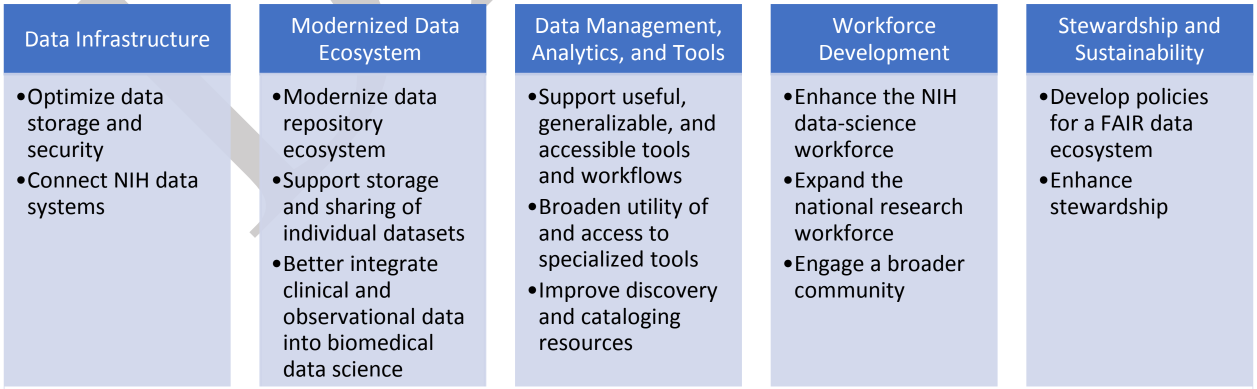
Recognizing the rapid course of evolution of data science and technology, this plan maps a general path for the next five years but is intended to be as nimble as possible to adjust to undiscovered concepts and products derived from current investments from NIH and elsewhere in the public and private sectors. Frequent course corrections are likely based upon the needs of NIH and its stakeholders and on new opportunities that arise because of the development of new technologies and platforms.

What is FAIR?

Biomedical research data should adhere to FAIR principles, meaning that it should be Findable, Accessible, Interoperable, and Reusable.

- To be **Findable**, data must have unique identifiers, effectively labeling it within searchable resources.
- To be **Accessible**, data must be easily retrievable via open systems and effective and secure authentication and authorization procedures.
- To be **Interoperable**, data should “use and speak the same language” via use of standardized vocabularies.
- To be **Reusable**, data must be adequately described to a new user, have clear information about data-usage licenses, and have a traceable “owner’s manual,” or provenance.

As outlined in the five Overarching Goals and correspondent Strategic Objectives (Fig. 2), NIH’s strategic approach will move toward a common architecture, infrastructure, and set of tools upon which individual Institutes and Centers (ICs) and scientific communities will build and tailor for specific needs. A Software as a Service (SaaS) framework, in which software licensing and delivery are provided and hosted by centralized resources, will greatly facilitate access to, analysis and curation of, and sharing of all NIH-funded data. Adhering to NIH’s data-science vision outlined herein, and compatible with the NIH mission, the NIH Chief Data Strategist, in conjunction with the NIH Scientific Data Council and NIH Data Science Policy Council, will serve as leads for implementing this strategic plan. Evaluation is a critical component of stewardship of federal resources, and over the course of 2018, NIH will develop performance measures and specific milestones that will be used to gauge the progress of this strategic



Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
•Optimize data storage and security •Connect NIH data systems	•Modernize data repository ecosystem •Support storage and sharing of individual datasets •Better integrate clinical and observational data into biomedical data science	•Support useful, generalizable, and accessible tools and workflows •Broaden utility of and access to specialized tools •Improve discovery and cataloging resources	•Enhance the NIH data-science workforce •Expand the national research workforce •Engage a broader community	•Develop policies for a FAIR data ecosystem •Enhance stewardship

Figure 2. NIH Strategic Plan for Data Science: Overview of Goals and Objectives

[5] *Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data. *2016;3:160018.*

plan and guide any necessary course corrections. Example metrics and milestones appear at the end of each Goal section. These will be established in concert with the NIH Chief Data Strategist and NIH IC leadership.

Cross-Cutting Themes

The Overarching Goals, Strategic Objectives, and Implementation Tactics outlined in this plan are highly integrated. The central aim is to modernize the data-resource ecosystem to increase its utility for researchers and other stakeholders, as well as to optimize its operational efficiency. The many connections between infrastructure, resources, tools, workforce, and policies call us to articulate a number of cross-cutting themes, presented below, that layer across the intentions and actions outlined in this document.

- Support common infrastructure and architecture on which more specialized platforms can be built and interconnected.
- Leverage commercial tools, technologies, services, and expertise; and adopt and adapt tools and technologies from other fields for use in biomedical research.
- Enhance the nation's biomedical data science research workforce through improved training programs and novel partnerships.
- Enhance data sharing, access, and interoperability such that NIH-supported data resources are FAIR.
- Ensure information security of patient and participant data.
- Improve the ability to capture, curate, store, and analyze clinical data for biomedical research.
- Coordinate and collaborate with other federal, private and international funding agencies and organizations to promote economies of scale and synergies and prevent unnecessary duplication.

Data science holds significant potential for accelerating the pace of biomedical research. To this end, NIH will continue to leverage its roles as an influential convener and major funding agency to encourage rapid, open sharing of data and greater harmonization of scientific efforts. Through implementing this strategic plan, NIH will enhance 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 ecosystem.

Overarching Goals, Strategic Objectives, and Implementation Tactics

Ensuring that the biomedical research data-resource ecosystem is FAIR (see text box) is a laudable but complex goal to achieve on a large scale, especially given the international expanse of biomedical research data resources and their use. NIH as the world's largest funder of biomedical research can play a leadership role by developing practical and effective policies and principles related to the storage, use, and security of biomedical data. In this strategic plan, NIH articulates specific priorities that address developing reliable, accessible, and appropriately secured modes of storage; transforming a fragmented set of individual components into a coordinated, efficient, and optimally useful ecosystem; reducing unnecessary redundancies and increasing synergies and economies of scale; and strengthening coordination and interactions—both within NIH and between NIH and its stakeholder communities. Paramount is the need to establish organizing principles and policies for an efficient yet nimble funding model for data science infrastructure that serves the needs of NIH and its stakeholders.

GOAL 1
Support a Highly Efficient and Effective Biomedical Research Data Infrastructure

NIH ICs routinely support intramural and extramural research projects that generate tremendous amounts of biomedical data. Regardless of format, all types of data require hardware, architecture, and platforms to capture, organize, store, allow access to, and perform computations with it. As projects mature, data have traditionally been stored and made available to the broader community via public repositories or at data generators' or data aggregators' local institutions. This model has become strained as the number of data-intensive projects—and the amount of data generated for each project—continues to grow rapidly.

Objective 1-1 | Optimize Data Storage and Security

Large-scale cloud-computing platforms are shared environments for data storage, access, and computing. They rely on using distributed data-storage resources for accessibility and economy of scale—similar conceptually to storage and distribution of utilities like electricity and water. Cloud environments thus have the potential to streamline NIH data use by allowing rapid and seamless access, as well as to improve efficiencies by minimizing infrastructure and maintenance costs. NIH will leverage what is available in the private sector, either through strategic partnerships or procurement, to create a workable Platform as a Service (PaaS) environment. NIH will partner with cloud service providers for cloud storage, computational, and related infrastructure services needed to facilitate the deposit, storage, and access to large, high-value NIH data sets (see text box "Science in the Cloud: The NIH Data Commons"). These negotiations may result in partnership agreements with top infrastructure providers from U.S.-based companies whose focus includes support for research. Suitable cloud environments will house diverse data types and high-value data sets created with public funds and ensure that they are stable and secure, to protect against data compromise or loss and available for research use and reuse.

NIH's cloud marketplace initiative will be the first step in a phased operational framework that establishes a SaaS paradigm for NIH and its stakeholders.

Science in the Cloud: The NIH Data Commons

One of the first steps NIH is taking to modernize the biomedical research data ecosystem is funding the NIH Data Commons pilot: Its main objective is to develop the ability to make data FAIR through use of a shared virtual space to store and work with biomedical research data and analytical tools. The NIH Data Commons will leverage currently available cloud-computing environments in a flexible and scalable way, aiming to increase the value of NIH-supported data by democratizing access and use of data and analytical tools and allowing multiple datasets to be queried together. To begin, the NIH Data Commons will enable researchers to work with three test data sets: the National Heart, Lung, and Blood Institute's Trans-Omics for Precision Medicine (TOPMed) program, the NIH Common Fund's Genotype-Tissue Expression (GTEx) program, and various model-organism data repositories.

Implementation Tactics:

- Leverage existing federal, academic, and commercial computer systems for data storage and analysis.
- Adopt and adapt emerging and specialized technologies (see text box "Graphical Processing Units").
- Support technical and infrastructure needs for data security, authorization of use, and unique identifiers to index and locate data.

Objective 1-2 | Connect NIH Data Systems

More than 3,000 different groups and individuals submit data via NCBI systems daily. Among these are genome sequences from humans and research organisms; gene-expression data; chemical structures and properties, including safety and toxicity data; information about clinical trials and their results; data on genotype-phenotype correlations; and others. Beyond NIH-funded scientists and research centers, many other individuals and groups contribute data to the biomedical research data ecosystem, including other federal agencies, publishers, state public-health laboratories, genetic-testing laboratories, and biotech and pharmaceutical companies. NIH will develop strategies to link high-value NIH data systems, building a framework to ensure they can be used together rather than existing as isolated data silos (see text box, below, "Biomedical Data Translator"). A key goal is to promote expanded data sharing to benefit not only biomedical researchers but also policymakers, funding agencies, professional organizations, and the public.

Graphical Processing Units

The workhorses of most computers are central-processing units, or CPUs, which perform computing functions as specified by computer programs. Specialized versions of these, called graphical processing units, or GPUs, are dedicated exclusively to imagery, or graphics. These elements have driven the motion-picture and video-game industries, and, more recently, have been adapted for use in biomedical research with very large and complex data sets such as molecular, cellular, radiological, or clinical images.

Implementation Tactics:

- Link the NIH Data Commons (see text box, above) and existing, widely-used NIH databases/data repositories using NCBI as a coordinating hub.
- Ensure that new NIH data resources are connected to other NIH systems upon implementation.
- When appropriate, develop connections to non-NIH data resources.

Biomedical Data Translator

Through its Biomedical Data Translator program, the National Center for Advancing Translational Sciences (NCATS) is supporting research to develop ways to connect conventionally separated data types to one another to make them more useful for researchers and the public. The Translator aims to bring data types together in ways that will integrate multiple types of existing data sources, including objective signs and symptoms of disease, drug effects, and other types of biological data relevant to understanding the development of disease and how it progresses in patients.

Goal 1: Evaluation

For this Goal, “Support a Highly Efficient and Effective Biomedical Research Data Infrastructure,” potential measures of progress include: quantity of cloud storage and computing used by NIH and by NIH-funded researchers; unit costs for cloud storage and computing; number of technologies adapted for use by NIH-funded resources; quantity of NIH data resources incorporated into the NIH Data Commons; and quantity of NIH data resources linked together.

GOAL 2
Promote Modernization of the Data-Resources Ecosystem

The current biomedical data-resource ecosystem is challenged by a number of organizational problems that create significant inefficiencies for researchers, their institutions, funders, and the public. For example, from 2007 to 2016, NIH ICs used dozens of different funding strategies to support data resources, most of them linked to research grant mechanisms that prioritized innovation and hypothesis testing over user service, utility, access, or efficiency. In addition, although the need for open and efficient data sharing is clear, where to store and access datasets generated by individual laboratories—and how to make them compliant with the FAIR principles—is not yet straightforward. Overall, it is critical that the data-resource ecosystem become seamlessly integrated such that different data types and information about different organisms or diseases can be easily used together rather than existing in separate data "silos" with only local utility. Wherever possible, NIH will coordinate and collaborate with other federal, private, and international funding agencies and organizations to promote economies of scale and synergies and prevent unnecessary duplication.

Objective 2-1 | Modernize the Data Repository Ecosystem

To promote modernization of the data-repository ecosystem, NIH will refocus its funding priorities on the utility, user service, accessibility, and efficiency of operation of repositories (see Current Data Science Challenges for NIH). Wherever possible, data repositories should be integrated and contain harmonized data for all related organisms, systems, or conditions, allowing for seamless comparison. To improve evaluation of data-repository utility, and allow those who run them to focus on the particular goals they need to achieve to best support the research community and operate as efficiently as possible, NIH will distinguish between databases and knowledgebases (see text box "Databases and Knowledgebases: What's the Difference?") and will support each separately from one another as well as from the development and dissemination of tools used to analyze data (see Goal 3 for NIH's proposed new strategies for tool development). Although a grey area does exist between databases and knowledgebases, and some data types currently appropriate for a knowledgebase may eventually harden and become core data more appropriate for a database, this distinction will allow improved focus and coherence in the support and operation of modern data resources. Funding approaches used for databases and knowledgebases will be appropriate for resources and focus on user service, utility,

Databases and Knowledgebases: What's the Difference?

Databases are data repositories that store, organize, validate, and make accessible the core data related to a particular system or systems. For example, the core data for a model organism database might include genome, transcriptome, and protein sequences and functional annotations of gene products.

Knowledgebases accumulate, organize, and link growing bodies of information related to core datasets. A knowledgebase may contain information about expression patterns, splicing variants, localization, and protein-protein interaction and pathway networks related to an organism or set of organisms. Knowledgebases typically require significant curation beyond the quality assurance/quality control and annotation needed for databases.

and operational efficiency rather than on research project goals. As such, NIH will establish procedures and metrics to monitor data usage and impact—including usage patterns within individual datasets.

Implementation Tactics:

- Separate the support of databases and knowledgebases (see text box, above).
- Use appropriate and separate funding strategies, review criteria, and management for each repository type.
- Dynamically measure data use, utility, and modification.
- Ensure privacy and security.
- Create unified, efficient, and secure authorization of access to sensitive data.
- Employ explicit evaluation, lifecycle, sustainability, and sunsetting expectations for data resources.

Objective 2-2 | Support the Storage and Sharing of Individual Datasets

Currently, most data generated by biomedical researchers are small-scale datasets produced by individual laboratories. In contrast, large, organized consortia, including various programs specific to NIH ICs and the NIH Common Fund, generate large, high-value datasets that are relatively small in numbers but used by thousands of researchers. Whereas these large datasets generally reside in dedicated data resources, a current dilemma for NIH is determining how to store and make accessible all of the smaller datasets from individual laboratories. NIH will create an environment in which individual laboratories can link datasets to publications in the NCBI's PubMed Central publication database. Part of that effort includes development of the NIH Data Commons Pilot, the first trans-NIH effort to create a shared, cloud-based environment for data storage, access and computation.

Implementation Tactics:

- Link datasets to publications via PubMed Central and NCBI.
- Longer-term: Expand NIH Data Commons to allow submission, open sharing, and indexing of individual, FAIR datasets.

Objective 2-3 | Leverage Ongoing Initiatives to Better Integrate Clinical and Observational Data into Biomedical Data Science

NIH has several large-scale, ongoing efforts that are building high-value resources that include clinical and observational data from individual volunteers and patients (for one example, see text box, below, "TB Portals"). These efforts also include the All of Us Research Program and Cancer Moonshot^{SM}, the National Heart, Lung and Blood Institute's Trans-Omics for Precision Medicine (TOPMed) program, the Environmental Influences on Child Health Outcomes (ECHO) program, and various datasets generated by scientists conducting research in the NIH Clinical Center (see text box, below, "Big Data for Health"). NIH will leverage these and other related initiatives to integrate the patient health data they contain into the biomedical data-science ecosystem in ways that maintain security and confidentiality and are consistent with informed consent, applicable laws, and high standards for ethical conduct of research (see Clinical Data and Information Security).

TB Portals

The National Institute of Allergy and Infectious Diseases Tuberculosis (TB) Portals Program is a multi-national collaboration for data sharing and analysis to advance TB research. A consortium of clinicians and scientists from countries with a heavy burden of TB, especially drug-resistant TB, work together with data scientists and information technology professionals to collect multi-domain TB data and make it available to the clinical and research communities.

The Office of the National Coordinator for Health Information Technology (ONC), within the Department of Health and Human Services, leads national health information technology efforts. ONC coordinates nationwide efforts to implement and use the most advanced health information technology and the electronic exchange of health information, enhancing the ability to extract and share data between federal health agencies such as the Food and Drug Administration, the Centers for Disease Control, and the Centers for Medicare and Medicaid Services. ONC's interoperability initiatives are those that make electronic medical records more accessible for research, and NIH continues to adopt these practices such as standardizing terminology, or vocabularies, for clinical care. The NIH Clinical Center has in place the Biomedical Translational Research Information System (BTRIS), which is a resource available to the NIH intramural community that brings together clinical research data from the Clinical Center and other NIH ICs. BTRIS provides clinical investigators with access to identifiable data for subjects on their own active protocols, while providing all NIH investigators with access to data without personal identifiers across all protocols. Additionally, NIH encourages researchers to use common data elements, or CDEs, which helps improve accuracy, consistency, and interoperability among data sets within various areas of health and disease research.

Big Data for Health

Two signature NIH projects that aim to garner health insights from human data are the All of Us Research Program and the Cancer Moonshot. The All of Us Research Program aims to gather data over time from 1 million or more people living in the United States, with the ultimate goal of accelerating research and improving health. Scientists plan to use All of Us Research Program data to learn more about how individual differences in lifestyle, environment, and biological makeup can influence health and disease. Participants in the All of Us Research Program may be invited to use wearable sensors that will provide real-time measurements of their health and environmental exposures, significantly expanding this type of research. The Cancer Moonshot aims to accelerate cancer research to make more therapies available to more patients, while also improving our ability to prevent cancer and detect it at an early stage. Data-intensive strategies include mining past patient data to predict responses to standard treatments and future patient outcomes, developing a three-dimensional cancer atlas to view how human tumors change over time, and a Cancer Research Data Commons.

Implementation Tactics:

- Create efficient linkages among NIH data resources that contain clinical and observational information.

- Develop and implement universal credentialing protocols and user authorization systems that work across NIH data resources and platforms.
- Promote use of the NIH Common Data Elements Repository.

Goal 2: Evaluation

For this Goal, "Promote Modernization of the Data-Resources Ecosystem," potential measures of progress include: data resource use metrics (overall and individual datasets); quantity of databases and knowledgebases supported using resource-based funding mechanisms; cost-efficiency improvements over baseline (e.g., cost per unit of use by researchers, cost per amount of data stored, average cost per resource) of NIH-supported data resources; quantity of datasets deposited (over baseline) and linked to publications in PubMed Central; linkages between NIH clinical and observational data resources; and usage of the NIH Common Data Elements Repository.

GOAL 3

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

Extracting understanding from large-scale or complex biomedical research data requires algorithms, software, models, statistics, visualization tools, and other advanced approaches such as machine learning, deep learning, and artificial intelligence (see text box “Thinking Machines[6]”). Accomplishing NIH’s goal of optimizing the biomedical data-science ecosystem requires prioritizing development and dissemination of accessible and efficient methods for advanced data management, analysis, and visualization. To make the best tools available to the research community, NIH will leverage existing vibrant tool-sharing systems to help establish a more competitive “marketplace” for tool developers and providers than currently exists. By separating the evaluation and funding for tool development and dissemination from support for databases and knowledgebases, innovative new tools and methods should rapidly overtake and supplant older, obsolete ones. The goal of creating a more competitive marketplace, in which open-source programs, workflows, and other applications can be provided directly to users, could also allow direct linkages to key data resources for real-time data analysis.

Thinking Machines

Artificial intelligence, machine learning, and deep learning are computer algorithms that change as they are exposed to more data. Many are adept at recognizing objects in images, thus enabling machines to find patterns in data in ways that are not only cost-effective but also potentially beyond human abilities. “Centaur radiologists,” hybrids of human and computer expertise, will likely offer the opportunity to interpret data from a range of sources, including electronic medical records. Such hybrid person-machine interpretations will likely have a formative role in customizing care to individual patients.

Objective 3-1 | Support Useful, Generalizable, and Accessible Tools and Workflows

Historically, because data resources have generally been funded through NIH research grants, applicants have emphasized development of new tools in order to meet innovation expectations associated with conducting research. This strategy can shift the focus of data resources away from their core function of providing reliable and efficient access to high-quality data. In addition, coupling review and funding of data resources to tool development can inhibit the type of open competition among developers that allows support of the most innovative and useful tools (see Current Data Science Challenges for NIH). To address these concerns, NIH will evaluate and fund tool development separately from support of databases and knowledgebases. NIH will also promote the establishment of environments in which high-quality, open-source data management, analytics, and visualization tools can be obtained and used directly with data in the NIH Data Commons and/or other cloud environments. A key step will be leveraging through partnerships or procurement expertise in systems integration/engineering to refine and harden tools from academia to improve software design, usability, performance, and efficiency.

[6] Dreyer KJ, Geis JR. When Machines Think: Radiology's Next Frontier. Radiology. 2017 Dec;285(3):713-718. doi: 10.1148/radiol.2017171183.

Implementation Tactics:

- Separate support for tools development from support for databases and knowledgebases.
- Use appropriate funding mechanism, scientific review, and management for tool development.
- Establish partnerships to allow systems integrators/engineers from the private sector to refine and optimize prototype tools developed in academia to make them efficient, cost-effective, and widely useful for biomedical research.
- Employ a range of incentives to promote data-science and tool innovation including “hackathons,” prizes, public-private partnerships, and other approaches.

Objective 3-2 | Broaden Utility, Usability, and Accessibility of Specialized Tools

An important opportunity and challenge is to adopt for use in biomedical research tools that have been developed by fields outside of the biomedical sciences. For example, the same software used by NASA scientists to determine the depths of lakes from space is being tested for use in medical-image analysis for mammography, X Rays, computerized tomography (CT) and magnetic resonance imaging (MRI) scans, as well as for ultrasound measurements.[7] Specialized tools developed for one subfield of biomedical research might also be adopted for different purposes by researchers in other areas.

It will also be important to develop and adopt better tools for collecting and efficiently assimilating data from disparate and dynamic sources that combine to inform us about the health of individuals and populations. Novel data-science algorithms will likely create new knowledge and innovative solutions relevant to health disparities and disease prevention. New approaches and tools have the potential to transform data combined from various structured and unstructured sources into actionable information that can be used to identify needs, provide services, and predict and prevent poor outcomes in vulnerable populations. Toward identifying and stratifying risk, employing standardized data formats and vocabularies should advance our understanding of relationships between demographic information, social determinants of health, and health outcomes. One especially ripe opportunity for use in community-based research is broader use of rapidly evolving mobile-device technologies and information-sharing platforms. These resources, such as wearable devices, can capture a wide variety of health and lifestyle-related data from individual volunteers that could help transform our understanding of both normal human biology and disease states.

[7] NASA Technology Transfer Program: Hierarchical Image Segmentation: Available at https://technology.nasa.gov//t2media/tops/pdf/GSC-TOPS-14.pdf

Finally, there is a critical need for better methods to mine the wealth of data available in electronic medical records (see text box “Linking Genomic Data to Health: eMERGE”). These records present great opportunities for advancing medical research and improving human health—particularly in the area of precision medicine—but they also pose tremendous challenges (see “Clinical Data and Information Security”). For example, patient confidentiality must be assured, and the level of access granted by each individual to researchers has to be obtained, recorded, and obeyed. Equally challenging is the fact that electronic medical records are controlled by thousands of different hospitals and other organizations using dozens of different commercial computer platforms that do not currently share a uniform language or data standards. Because of these challenges, NIH will support additional research to find better ways to allow clinical data to be used securely, ethically, and legally, to advance medicine. NIH will also work with other federal and state agencies, private healthcare and insurance providers, and patient advocacy groups to find more efficient paths to realize the promise of electronic medical records and other clinical data for medical research.

Linking Genomic Data to Health: eMERGE

eMERGE is an NIH-funded national network organized and funded by the National Human Genome Research Institute that combines DNA biorepositories with electronic medical record systems for large scale, high-throughput genetic research in support of implementing genomic medicine. Begun in 2007, the project merges genetic data from patients with their electronic medical records and facilitates sharing of the data for biomedical research. eMERGE currently involves 14 sites across the country that collect, store, and share patient data for biomedical discovery aiming to improve patient care.

Implementation Tactics:

- Adopt and adapt emerging and specialized methods, tools, software, and workflows.
- Promote development and adoption of better mobile-device and data-interface tools.
- Support research to develop improved methods for using electronic medical records and other clinical data securely and ethically for medical research.

Objective 3-3 | Improve Discovery and Cataloging Resources

Data that is not easily located is likely to be underused and of little value to the broader research community. NIH has invested in developing resources, such as the Data Discovery Index that is part of the NIH Data Commons pilot, to enable data reuse. Such a resource will exceed a mere cataloging function and will also contain platforms and tools by which biomedical researchers can find and reuse data and that will support data-citation metrics. NIH will continue to invest in development of improved approaches for making data findable and accessible. For example, the NIH Data Commons pilot is creating search and analysis workspaces that support a broad range of authenticated users, and where users with all levels of expertise can access and interact with data and tools. Collaboration will be integral to this approach, which, in addition to continued research and development, must involve a community-driven process for identifying and implementing optimal standards to improve indexing, understandability, reuse, and citation of datasets.

Implementation Tactics:

- Promote community development and adoption of uniform standards for data indexing, citation, and modification-tracking (provenance).

Goal 3: Evaluation

For this Goal, "Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools," potential measures of progress include: quantity of new software tools developed; citations of tool use; downloads of tools from tool repositories; and quantity of tools adapted from other fields for use in biomedical research.

GOAL 4
Enhance Workforce Development for Biomedical Data Science

Innovative contributions to biology from computer science, mathematics, statistics, and other quantitative fields have facilitated the shift in biomedicine described throughout this document. NIH recognizes that data scientists perform far more than a support function, as data science has evolved to be an investigative domain in its own right. While NIH has supported quantitative training at various levels along the biomedical career path, more needs to be done to facilitate familiarity and expertise with data-science approaches and effective and secure use of various types of biomedical research data. There is also a need to grow and diversify the pipeline of researchers developing new tools and analytic methods for broad use by the biomedical research community. Finally, data-science approaches will be essential for NIH to achieve the stewardship goals outlined in the NIH-wide strategic plan and are likely to facilitate the agency's ability to monitor demographic trends among its workforce and thus address diversity gaps (see text box, below, "Data Science and Diversity").

Objective 4-1 | Enhance the NIH Data-Science Workforce

Given the importance of data science for biomedical research, NIH needs an internal workforce that is increasingly skilled in this area. This includes ensuring that NIH program and review staff who administer and manage grants and coordinate the evaluation of applications have sufficient experience with and knowledge of data science. To begin to address this need, NIH will develop training programs for its staff to improve their knowledge and skills in areas related to data science. In addition, NIH will recruit a cohort of data scientists and others with expertise in areas such as project management, systems engineering, and computer science from the private sector and academia for short-term (1- to 3-year) national service sabbaticals. These "NIH Data Fellows" will be embedded within a range of high-profile, transformative NIH projects such as All of Us, the Cancer Moonshot^{SM} and the BRAIN initiative and will serve to provide innovation and expertise not readily available within the federal government.

Data Science and Diversity

Data-driven approaches offer opportunities to address gaps in scientific workforce diversity, by observing and influencing system-wide patterns. NIH has applied this strategy to understand factors driving recruitment and retention of undergraduate students from underrepresented groups as well as to study various approaches to mentoring. Additional systems-based approaches to workforce modeling, including at the faculty level at NIH-funded institutions, offer promise for understanding effects on diversity in various contexts.

Implementation Tactics:

- Develop data-science training programs for NIH staff.
- Launch the NIH Data Fellows program.

Objective 4-2 | Expand the National Research Workforce

Modern biomedical research is becoming increasingly quantitative and it is essential that the next generation of researchers be equipped with the skills needed to take advantage of the growing promise of data science for advancing human health. NIH will work to ensure that NIH-funded training and fellowship programs emphasize teaching of quantitative and computational skills and integrate training in data-science approaches throughout their curricula and during mentored research. In addition, NIH will partner with institutions to engage librarians and information specialists in finding new paths in areas such as library science that have the potential to enrich the data-science ecosystem for biomedical research.

Implementation Tactics:

- Enhance quantitative and computational training for graduate students and postdoctoral fellows.
- Build on diversity-enhancing efforts in data science, such as the NIH BD2K Diversity Initiative.[8]
- Engage librarians and information specialists in developing data-science solutions and programs.
- Employ data-driven methods to monitor workforce diversity.

Objective 4-3 | Engage a Broader Community

As a field, data science crosses boundaries between research and practice, as well as between science and policy. NIH will promote knowledge exchange and development of best practices for the collection, organization, preservation, and dissemination of information resources across communities. Part of this effort is nurturing cultural change, emphasizing the role of data science in discovery and health, and enabling citizen scientists access to data without compromising its privacy or security. NIH recognizes its role in the larger data-science ecosystem and that NIH-generated biomedical data are used often by the private sector, clinicians, and other public groups. As part of the BD2K effort, NIH encouraged development of new or significantly adapted interactive digital media that engages the public, experts or non-experts, in performing some aspect of biomedical research via crowdsourcing. NIH will work to find additional ways to engage the public and healthcare providers in making use of biomedical data and data-science tools. Doing so will help to expand the biomedical “sandbox” to researchers without access to large-scale computational resources, such as non-research academic organizations, community colleges, and citizen scientists.

Implementation Tactics:

- Give citizen scientists access to appropriate data, tools, and educational resources (see text box “Citizen Science”).
- Develop materials to train healthcare providers in data science-related clinical applications.

[8] Canner JE, McEligot AJ, Pérez ME, Qian L, Zhang X. Enhancing Diversity in Biomedical Data Science. Ethn Dis. 2017;27(2):107-116.

Citizen Science

Citizen science aims to bring science and science policy processes to the public, to ensure that science is responsive to public concerns and needs and to harness an additional resource to capture data and innovation in a wide variety of fields. One example is EyeWire, an online resource launched in 2012 that challenges players to map the three-dimensional structure of neurons in the retina of a mouse using real electron-microscopy images. About 70,000 players from more than 130 countries play this scientific "game," which is helping researchers discover how neurons connect to process visual information.

Goal 4: Evaluation

For this Goal, "Enhance Workforce Development for Biomedical Data Science," potential measures of progress include: quantity of new data science-related training programs for NIH staff and participation in these programs; number of NIH Data Fellows recruited; number of NIH-funded training programs that increase requirements for quantitative and computational skills development; and demographic, geographic, and disciplinary diversity of NIH-funded quantitative scientists.

GOAL 5
Enact Appropriate Policies to Promote Stewardship and Sustainability

Creating and maintaining an efficient and effective biomedical data-science ecosystem requires policies and practices appropriate for optimal governance, financial management, evaluation, and sustainable stewardship of resources. Because cultural issues are central to implementing policies, appropriate reward, review, and expectation systems are central to making data FAIR and for incentivizing researchers to share their data and analysis tools widely for reuse by others. To ensure researchers collecting data understand and comply with data-security and confidentiality standards, it will be important for NIH to collaborate with the research community on strategies to guide general practice in data security and privacy matters and to collaborate with industry leaders who set standards in the information-security arena.

Objective 5-1 | Develop Policies for a FAIR Data Ecosystem

Currently, most biomedical data do not adhere to FAIR principles and thus are difficult to find and access. Moreover, complex or integrated analysis requires that data are interoperable and reusable across multiple domains with high fidelity. Thus, through appropriate policies and practices and as a core data-management activity, NIH will strive to ensure that all data in NIH-supported data resources are FAIR. The NIH Data Commons Pilot will be a starting point toward accomplishing this objective. While freely sharing high-value data is a critical goal for advancing research, NIH must ensure that its policies are achievable and sustainable and do not impose unnecessary burdens or untenable expectations on grantee institutions. Therefore, policies must reflect the data use and evaluation metrics and methods that will be established in Objective 2-1 to guide what data need to be made accessible and when they should be moved to less-accessible but less-expensive archive storage or retired altogether. NIH will also promote community-guided development of model open data-use licenses that will facilitate data sharing while simultaneously allowing protection of confidentiality and intellectual property. In addition, the NIH Data Commons pilot is establishing ways to use controlled-access data through appropriate authentication and protocols.

Implementation Tactics:

- Create rational and supportable data-sharing and data-management policies.
- Promote development of community standards that support FAIR principles for data storage.
- Develop model open-data use licenses to enable broad access to datasets.
- Optimize security management and access policies.

Objective 5-2 | Enhance Stewardship

The rapidly growing amount of data generated by the biomedical research enterprise creates an urgent need for developing clear guidelines for what data must be stored and shared, where and in what form it must be stored, as well as practical solutions for sustaining valuable data resources and determining priorities for data-resource funding. In addition, to produce the most scientific value for taxpayers' investments and to provide researchers with the best access to data resources possible, NIH must work with the community to improve the efficiency of operation of these resources and, wherever possible, create synergies and economies of scale. Toward achieving these goals, NIH will collaborate with its stakeholders—including academia, other U.S. and international funding agencies, journals, and the private sector—to establish a wide range of metrics to dynamically measure data use, utility, and modification, as well as measures of the operational efficiency of the resources themselves. Creating incentives and expectations for depositing FAIR-compliant data in NIH-funded repositories, data commons, or other NIH data systems, will enhance data sharing and reuse and allow NIH to accurately assess data usage and lifecycles. This information will be essential for making informed decisions about priorities for data-resource support. In addition, NIH will engage the broader data-science community in testing the utility of the NIH Data Commons as it is developed. As it refines these systems, NIH will seek input from entities with expertise in research ethics, privacy regulations and statutes, and data security to ensure NIH-supported data resources maintain research-participant confidentiality.

Implementation Tactics:

- Develop standard use, utility, and efficiency metrics and review expectations for data resources and tools.
- Establish sustainability models for data resources.
- Develop a reward and expectation system for investigators to make data FAIR and for ensuring open-source data-analysis tools are available.

Goal 5: Evaluation

For this Goal, "Enact Appropriate Policies to Promote Stewardship and Sustainability," potential measures of progress include: establishment and use of open-data licenses; development and performance of metrics to assess data resource use, utility and efficiency of operation; and usage of data-security protection guidelines by NIH-funded researchers using clinical data.

Conclusion

Accessible, well-organized, secure, and efficiently operated data resources are critical enablers of modern scientific inquiry. With publication of the *NIH Strategic Plan for Data Science*, NIH aims to maximize the value of data generated through NIH-funded efforts to enable biomedical discovery and innovation. Doing so is critical for keeping the United States at the forefront of biomedical research, ensuring continued advances toward improving the nation's health.

This strategic plan is highly interconnected but rests upon five pillars, its Overarching Goals:

- Support a Highly Efficient and Effective Biomedical Research Data Infrastructure
- Promote Modernization of the Data-Resources Ecosystem
- Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools
- Enhance Workforce Development for Biomedical Data Science
- Enact Appropriate Policies to Promote Stewardship and Sustainability

As articulated in this plan, NIH considers essential the need to coordinate and collaborate with other federal, private and international agencies and organizations in the data-resource ecosystem to promote economies of scale and synergies and prevent unnecessary duplication. NIH's goal is to maximize the utility of the data resource ecosystem for researchers and to ensure optimal scientific return on investment for taxpayers. A central facet of the plan is engagement of industry partners who have expertise in key areas related to information technology, complementing the research strengths of NIH and the academic community, and helping ensure that we achieve the plan's Goals and Objectives in an efficient manner. As outlined above, NIH's vision is to move toward a common framework upon which individual ICs and scientific fields will build and adapt. The inaugural NIH Chief Data Strategist, in conjunction with the NIH Scientific Data Council and NIH Data Science Policy Council, will lead efforts to manage this new system and will take the lead on implementing this strategic plan, in close collaboration with NIH IC leadership and staff.

NIH recognizes fully that technological advancement and unprecedented growth in biomedical data have created great opportunities, but they have also introduced great challenges for protecting the privacy and security of patient and other research data. NIH must work with its stakeholders and experts in the private sector and other federal agencies to promote and practice robust and proactive information-security procedures to ensure appropriate stewardship of patient and research-participant data while at the same time enabling scientific and medical advances. Because the data-science ecosystem is evolving rapidly, this strategic plan is intended to be nimble enough to make necessary course corrections through an ongoing process of performance evaluation, needs assessment, and research and technology landscape surveys. NIH's corporate strategy for positioning itself within a highly dynamic biomedical data-science enterprise will poise the agency to maximize the utility of data science for discovery while minimizing duplication of effort and ensuring that its research investments are cost-effective for American taxpayers.

Glossary

Algorithm—a process or set of rules to be followed in calculations or other problem-solving operations

Application-programming interface (API)—clearly defined method of communication between different software components

Artificial intelligence—the power of a machine to copy intelligent human behavior

Cloud computing—an internet-enabled shared system of resources usable by many individuals and groups

Common data element (CDE)—piece of data common to multiple datasets across different studies (may be universal or domain-specific)

Crowdsourcing—a distributed model in which individuals or groups obtain services, ideas, or content from a large, relatively open and often rapidly-evolving group of internet users

Database/Data repository—virtual data storage that stores, organizes, validates, and makes accessible core data related to a particular system or systems

Data commons—a shared virtual space in which scientists can work with the digital objects of biomedical research such as data and analytical tools

Data ecosystem—a distributed, adaptive, open system with properties of self-organization, scalability, and sustainability inspired by natural ecosystems.

Data integrity—the accuracy and consistency of data stored in a database/data repository, data warehouse, data mart or other construct

Data science—interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data

Dataset—collection of related sets of information composed of separate elements that can be manipulated computationally as a unit

Data visualization—effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected can be exposed and recognized more easily with data-visualization techniques

Deep learning—type of machine learning in which each successive layer uses output from the previous layer as input; similar to communication patterns in a biological nervous system

Domain-specific—for biomedical data, designed and intended for use in studies of a particular topic, disease or condition, or body system (compare to universal)

Electronic medical record—digital version of a patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users

Enterprise system—computer hardware and software used to satisfy the needs of an organization rather than individual users

Exascale computing—a computer system capable performing one quintillion (10^{18}) calculations per second.

Extramural—research or other activities supported by NIH and conducted by external organizations, and funded by grants, contracts, or cooperative agreements from NIH.

Genotype—the genetic make-up of an individual organism

Hardening—process of optimizing a tool or algorithm to industry standards to ensure efficiency, ease of use, security, and utility.

Hardware—collection of physical parts of a computer system

Indexing—methods to allow data finding and retrieving

Interoperability—in computer systems, the ability to exchange and make use of information from various sources and of different types

Intramural—research or other activities conducted by, or in support of, NIH employees on its Bethesda, Maryland, campus or at one of the other NIH satellite campuses across the country.

Knowledgebase—virtual resource that accumulates, organizes, and links growing bodies of information related to core datasets

Machine learning—a field of computer science that gives computers the ability to learn without being explicitly programmed by humans

Metadata—data that describe other data. Examples include title, abstract, author, and keywords (publications); organization and relationships of digital materials; and file types or modification dates.

NIH IC—NIH Institute or Center

'Omics —collective characterization and measurement of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms. Examples include genomics, proteomics, metabolomics, and others.

Petascale computing—a computer system capable performing one quadrillion (10^{15}) calculations per second. It is currently used in weather and climate simulation, nuclear simulations, cosmology, quantum chemistry, lower-level organism brain simulation, and fusion science.

Phenotype—the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment

Platform—group of technologies (software and hardware) upon which other applications, processes, or technologies are developed

Platform as a Service (PaaS)—a type of cloud computing that allows users to develop, run, and manage applications without the complexity of building and maintaining an overarching infrastructure

Provenance—timeline of ownership, location, and modification

Retirement (of data)—the practice of shutting down redundant or obsolete business applications while retaining access to the historical data

Software—programs and other operating information used by a computer

Software as a Service (SaaS)—software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted

System integrator/system engineer—individual who refines and hardens tools from academia to improve user design, authentication and testing, and optimize productivity, efficiency, and outcomes/performance

Unique identifiers—an alphanumeric string (such as 1a2b3c) used to uniquely identify an object or entity on the internet

Universal—for biomedical data, usable in research studies regardless of the specific disease or condition of interest (compare to domain-specific)

Wearables—devices that can be worn by a consumer that collect data to track health

Workflow—defined series of tasks for processing data