INNOVATING THE DATA ECOSYSTEM: AN UPDATE OF THE FEDERAL BIG DATA RESEARCH AND DEVELOPMENT STRATEGIC PLAN

A Report by the

BIG DATA INTERAGENCY WORKING GROUP

NETWORKING AND INFORMATION TECHNOLOGY

RESEARCH AND DEVELOPMENT SUBCOMMITTEE

of the

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

November 2024

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, health, foreign relations, and the environment, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of federal research and development (R&D) in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the federal government. More information is available at https://www.whitehouse.gov/ostp.

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the executive branch coordinates science and technology policy across the diverse entities that make up the federal research and development enterprise. A primary objective of the NSTC is to ensure science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at https://www.whitehouse.gov/ostp/nstc.

About the Subcommittee on Networking & Information Technology Research & Development

The Networking and Information Technology Research and Development (NITRD) Program has been the nation's primary source of federally funded work on pioneering information technologies (IT) in computing, networking, and software since it was first established as the High-Performance Computing and Communications program following passage of the High-Performance Computing Act of 1991. The NITRD Subcommittee of the NSTC guides the multiagency NITRD Program in its work to provide the R&D foundations for ensuring continued U.S. technological leadership and for meeting the nation's needs for advanced IT. The National Coordination Office (NCO) supports the NITRD Subcommittee and its Interagency Working Groups (https://www.nitrd.gov/about/).

About the Big Data Interagency Working Group

The Big Data Interagency Working Group (BD IWG) coordinates federal R&D to enable timely and effective analysis, decision-making, and discovery based on large, diverse data. BD R&D expands big data and data science capabilities, providing the foundation for algorithm-driven businesses and catalyzing innovations critical to the nation.

About This Document

This document, Innovating the Data Ecosystem: An Update of The Federal Big Data Research and Development Strategic Plan, updates the 2016 Federal Big Data Research and Development Strategic Plan. This plan updates the vision and strategies on the R&D needs for big data laid out in the 2016 Strategic Plan through six strategic areas (enhance the reusability and integrity of data; enable innovative, user-driven data science; develop and enhance the robustness of the federated ecosystem; prioritize privacy, confidentiality, ethics, and security; develop necessary expertise and diverse talent; and enhance U.S. global leadership) to enhance data value and reusability and responsiveness to federal policies on data sharing and management.

Copyright Information

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

Chair
Arati Prabhakar, Assistant to the President for Science and Technology; Director, Office of Science and Technology Policy

Executive Director (Acting)
Lisa E. Friedersdorf, Office of Science and Technology Policy

SUBCOMMITTEE ON NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT (NITRD)

Co-Chair
Joydip Kundu, Deputy Assistant Director for Computer and Information Science and Engineering, National Science Foundation (NSF)

Co-Chair
Craig Schlenoff, Director, NITRD National Coordination Office (NCO)

BIG DATA INTERAGENCY WORKING GROUP

Co-Chair
Laura Biven, Integrated Infrastructure and Emerging Technologies Lead, Office of Data Science Strategy, National Institutes of Health (NIH) (through July 2024)
Manil Maskey, Data Science and Innovation Lead, Office of Chief Science Data Officer, National Aeronautics and Space Administration (NASA)

Co-Chair
Amy Walton, Deputy Director, Office of Advanced Cyberinfrastructure, Computer and Information Science and Engineering Directorate, National Science Foundation (NSF)

Technical Coordinator

Ji Hyun Lee, NITRD National Coordination Office

Writing Team Members

Rajeev Agrawal, Army, Department of Defense (DOD)
Laura Biven, National Institutes of Health (NIH)
Ishwar Chandramouliswaran, National Institutes of Health (NIH)
Wo Chang,[1] National Institute of Standards and Technology (NIST)
Allison Dennis, National Institutes of Health (NIH)
Hal Finkel, Office of Science, Department of Energy (DOE)
Daryl Hess, National Science Foundation (NSF)
Margaret Lentz, Office of Science, Department of Energy (DOE)
Heather Masson-Forsythe, National Science Foundation (NSF)
Aaron Mannes,[1] Department of Homeland Security (DHS)
William Miller, National Science Foundation (NSF)
Maria Seale, Army, Department of Defense (DOD)
Scott Sellars, Department of State (DOS)
Sylvia Spengler, National Science Foundation (NSF)
Lisa Ulmer,[1] National Science Foundation (NSF)
Amy Walton, National Science Foundation (NSF)

[1] Contributed to this document while affiliated with the listed organizations.

Table of Contents

Acronyms

AI	Artificial Intelligence
APIs	Application Programming Interfaces
BD IWG	Big Data Interagency Working Group
CARE	Collective Benefit, Authority to Control, Responsibility, and Ethics
DHS	Department of Homeland Security
DOD	Department of Defense
DOE	Department of Energy
FAIR	Findability, Accessibility, Interoperability, and Reusability
FAIR4ML	FAIR for Machine Learning
FAIR4RS	FAIR for Research Software
FSDRs	Federally Supported Data Repositories
HPC	High Performance Computing
IARPA	Intelligence Advanced Research Projects Activity
LLMs	Large Language Models
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology
MRI	Magnetic Resonance Imaging
NCO	National Coordination Office
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NITRD	Networking and Information Technology Research and Development
NSF	U.S. National Science Foundation
NSTC	National Science and Technology Council
OSTP	Office of Science and Technology Policy
PPDSA	Privacy Preserving Data Sharing and Analytics
PIDs	Persistent Identifiers
R&D	Research and Development
SOPs	Standard Operating Procedures
TRUST	Transparency, Responsibility, User Focus, Sustainability, and Technology

Executive Summary

Data has become a critical national asset, driving advances in research, industry, commerce, and education. Federally supported research is a valuable source of large volumes of information-rich data that, when appropriately harnessed, shared, and managed, have the potential to enable transformative insights. This, in turn, can significantly improve how data are collected, analyzed, and used, benefiting both the nation and society. Computational and data-driven capabilities underpin key areas of national interest, such as disaster preparedness and response, precision medicine, integration of renewable energy sources into an advanced electric grid, and competitiveness in artificial intelligence (AI), quantum computing, and digital twins.

The federal government recognizes the critical role data plays in driving innovation and has launched several initiatives to support its availability and responsible use. Initiatives such as the Year of Open Science by the White House OSTP promote open data sharing and responsible data management. The National Artificial Intelligence Resource Task highlights the importance of robust and accessible data as a foundation for U.S. leadership in AI. Additionally, the Executive Order on Promoting the Use of Trustworthy AI in the Federal Government underscores a strong commitment to ensuring that data is unbiased, ethical, and reliable, reinforcing the critical role of data integrity.

Building on these efforts, this document outlines a new vision and strategies to address the evolving big data research and development needs. It updates *The Federal Big Data Research and Development Strategic Plan* to address the notable developments and substantial changes in technologies and data management over the past several years. Additionally, this update considers the ethical and workforce implications of big data capabilities, and incorporates insights from federal agencies, findings from a community workshop, and input from the public. This update is essential for ensuring that the United States remains equipped to tackle the most pressing challenges and to empower the nation to make informed decisions that directly impact every sector.

Key advances in this updated strategic plan include the adoption of an adaptive data ecosystem view first envisioned in the 2016 plan. There is also a greater emphasis on the importance of all stages of the data lifecycle to enhance the value and reusability of data. Furthermore, this strategy responds to federal policies on data sharing and data management deployed since 2016 and recognizes that AI is rapidly changing the landscape for demand of data.

This plan is composed of a vision statement and six strategic areas that offer guidance on federal investments to support and propel the value of big data for science and innovation.

> Vision: An adaptive data ecosystem built on large, diverse datasets that can effectively and efficiently support the extraction and analysis of information. This ecosystem spans the entire data lifecycle, enabling new capabilities for federal agencies and the nation.

The following six strategic areas support this Vision.

- Strategy 1: Enhance the Reusability and Integrity of Data.
- Strategy 2: Enable Innovative, User-driven Data Science.
- Strategy 3: Develop and Enhance the Robustness of the Federated Ecosystem.
- Strategy 4: Prioritize Privacy, Confidentiality, Ethics, and Security.
- Strategy 5: Develop Necessary Expertise and Diverse Talent.
- Strategy 6: Enhance U.S. Global Leadership.

The adaptive data ecosystem provides a critical foundation to accelerate the process of scientific discovery and innovation, lead to new fields of research and new areas of inquiry that would otherwise be impossible, promote new economic growth, and educate the next generation of 21st-century scientists and engineers. This big data vision and six strategies will inform and guide future big data research and development (R&D) opportunities across U.S. federal agencies.

Introduction

Vision: An adaptive data ecosystem built on large, diverse datasets that can effectively and efficiently support the extraction and analysis of information. This ecosystem spans the entire data lifecycle, enabling new capabilities for federal agencies and the nation.

In an era of digital transformation, the internet of things, and increasing adoption of emerging technologies, data is increasingly playing a pivotal role. The nation faces a myriad of challenges and opportunities of national significance that require combined computational and data-driven capabilities. These include preparing for and responding to natural disasters, expanding personalized medicine, integrating renewable energy sources into an advanced power grid, and sustaining a competitive edge in key science and technology areas, such as machine learning (ML), AI, and quantum computing.

Significant progress has been made since the release of *The Federal Big Data Research and Development Strategic Plan*[2] in 2016. For example, there have been substantial successes in data-intensive computing at scale, both for high performance computing and in the cloud, including advances in development and implementation of large language models (LLMs). Instrumented systems and environments now deploy large numbers of heterogeneous sensors and sensor networks to collect vast amounts of data. Best practices in data management have matured substantially, including the widely accepted Findability, Accessibility, Interoperability, and Reusability (FAIR) Principles[3] and the awareness of Collective Benefit, Authority to Control, Responsibility, and Ethics (CARE) Principles for Indigenous Data Governance.[4] Incentives for data sharing have grown, including data management and sharing policies from U.S. federal research funding agencies. In addition, there are a growing number of management frameworks and standards for other components of the data ecosystem, such as repositories, software, and workflows.

There is also an increasing understanding in the big data community of how different organizations, stakeholders, and components of a big data ecosystem need to interact. Direct connections between sensors and instruments, edge and centralized computing are now being leveraged for more real-time and distributed analyses. Data repositories are key building blocks of the ecosystem that have a critical role in ensuring interconnectedness and interoperability.[5] The intervening years since the publication of the 2016 Strategic Plan have seen heightened expectations and serious concerns for the ethical development and use of data, such as the implications of bias within deployed AI models.

These advancements and emerging trends have shifted the bottlenecks for achieving the vision of a big data ecosystem. Current challenges focus on the problems that are more apparent when data and systems are shared or integrated. These challenges include data and system interoperability, trustworthiness of data and the insights derived from data, data management and analysis tools,

[2] https://www.nitrd.gov/pubs/bigdatardstrategicplan.pdf
[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[4] https://www.gida-global.org/care
[5] https://www.nitrd.gov/pubs/NITRD-BDIWG-Workshop-FSDRs-2022.pdf

practices for ensuring ethical development and use of data, and ensuring broad reusability and AI-readiness of shared data.

Additionally, technological advancements and changing user needs require that the envisioned adaptive data ecosystem be able to support a variety of use cases. The NITRD BD IWG identified the following use cases:

- Time-Sensitive Patterns: Science cases that require end-to-end urgency, such as streaming data for real-time analysis, real-time experiment steering, real-time event detection, AI-model inference, and deadline scheduling to avoid falling behind.[6, 7]
- Data-Integration-Intensive Patterns: Science cases that demand combining and analyzing data from multiple sources, such as AI federated learning, combined analysis of data from multiple sites/locations, experiments, and/or simulations.[6, 7]
- Long-Term Campaign Patterns: Science cases that require sustained access to resources over a long time to accomplish a well-defined objective, such as sustained simulation production, large data processing, and archiving for collaborative use and reuse.[6, 7]
- Internet of Things Patterns: Science cases that involve interconnecting computing devices embedded in instruments, sensors, or other systems of interest.
- Foundation Models: Science cases where self-supervision at scale enables models to learn from vast and well-curated datasets, allowing them to generalize effectively across applications while also minimizing biases and addressing ethical concerns.
- Digital Twins: Science cases where a computational model is twinned "…with a physical counterpart to create a system that is dynamically updated through bidirectional data flows as conditions change."[8]
- AI Agents: Science cases that deploy software to interact with and take actions autonomously based on the perceived environment.

Based on these challenges and inputs from the NITRD BD IWG member agencies, a public workshop,[9] and responses to a request for information,[10] six strategic areas were identified for future federal big data R&D. The following sections provide further information on these strategic areas.

Strategy 1: Enhance the Reusability and Integrity of Data

Many of the current challenges in data science rely on reusing data in ways different from the original purpose for which they were created or generated, and on bringing together data from multiple sources in unique ways. Principled approaches to data management, such as the FAIR and CARE principles, enhance the broad reusability of data. Community standards such as Common Data Elements,[11] curated ontologies, and data models all facilitate interoperability across data resources. The ability to

[6] https://www.osti.gov/biblio/1984466
[7] https://doi.org/10.2172/1984466
[8] https://climatemodeling.science.energy.gov/news/report-digital-twins
[9] https://www.nitrd.gov/pubs/NITRD-BDIWG-Workshop-FSDRs-2022.pdf
[10] https://www.federalregister.gov/documents/2022/07/01/2022-14084/request-for-information-on-the-federal-big-data-research-and-development-strategic-plan-update
[11] https://cde.nlm.nih.gov/guides

semantically link different modalities of data — for example, images and text, clinical health data and genomic information, simulation results and real-world observations — is also increasingly important for improving our understanding of complex systems like the human body and disease treatment, climate, environment, and engineered systems. Interoperability across data sources and among data modalities remains the most challenging of the FAIR Principles.

This strategy encompasses five key areas: Enabling and Incentivizing Open Science; Pursuing Standards and Provenance Development for FAIR Data; Leveraging AI Across the Data Life Cycle; Stewarding Data Assets as Long-term Infrastructure; and Developing and Stewarding New, Strategic Data Assets.

Enabling and Incentivizing Open Science

The U.S. government has advanced a culture of open science — "the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity".[12] Key initiatives supporting this include the Federal Data Strategy,[13] the OSTP Memo on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research,[14] and the data sharing policies of research funding agencies,[15] particularly for scholarly publications and data. The publication and dissemination of research results in the form of peer-reviewed articles and conference presentations and proceedings have provided recognizable ways to credit scientific accomplishments. Although there are community efforts to develop shared approaches to measuring the impact of data resources, there is currently no widely used benchmark for the value of sharing data and related digital products. The enormous positive impact that data sharing can have on scientific progress, data consistency, collaboration, cost savings, innovation, transparency, and societal issues can motivate the scientific research community to develop new ways of encouraging researchers to release data sets and digital research products by rewarding those who promote Open Science Principles. For example, it is increasingly common to see published research findings based on reused data.

Advancing the goals of open science will require both greater incentives and lower barriers or burdens for researchers to align with open science practices. This could include the refinement of policies from research organizations, federal funding agencies, publishers, and other institutions related to sharing the results of research, such as greater acknowledgement for contributions to open science and recognition during promotion and hiring decisions. Incentives are also needed to encourage community experts to take ownership of data and to develop mechanisms for automated data quality checks. Additionally, reducing the barriers for researchers should include a suite of tools at every stage of the data lifecycle for aligning data with existing standards and ontologies; improved practices for citing data, particularly ad hoc composite data sets; and tools for automated provenance capture to enhance the reusability of shared data.

[12] https://open.science.gov/
[13] https://strategy.data.gov/
[14] https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf
[15] https://smd-cms.nasa.gov/wp-content/uploads/2023/08/smd-information-policy-spd-41a.pdf

Open science can play an important role in realizing convergent science (integrating expertise from multiple scientific and engineering fields to address specific challenges) and realizing a scientific incentive for sharing data and maintaining high-quality, well-curated data collections across disciplines. There are examples[16, 17] of data-intensive approaches enabling the solution or significant progress over traditional science and engineering methods. The more fields of science and engineering that participate in this way, the more problems that require interdisciplinary or cross-disciplinary data-intensive approaches come in range. Solving problems on this scale brings significant rewards, such as career opportunities, which help to incentivize the creation of high-quality data collections across communities. Mechanisms should be developed to provide additional incentives for communities to develop and share their data.

To further the goals of open science, it would be beneficial to expand the definition of research products to include software, AI/ML models, data services, and standards. Indeed, much of this is already in progress as made evident by the many grassroots organizations adapting FAIR Principles for various research products, such as FAIR for Research Software (FAIR4RS)[18] and FAIR for Machine Learning (FAIR4ML).[19] Sharing all the research products relevant to research findings provides a rich context and meaning to shared data.

Finally, open science efforts need to improve the quality and reusability of data — a process that begins early in the data lifecycle, starting at data creation or collection and possibly even earlier during research design. Tools that automate the capture and organization of standardized data and metadata in sharable and reusable ways have the potential to significantly reduce the economic and time burdens researchers face, which consequently could improve data and metadata quality and reduce the overall burden on researchers for making data FAIR. Metadata and data models are inherently linked to the conceptual modeling and theories of the underlying system. Therefore, the development of metadata and the degree to which it can be standardized is, to some extent, a reflection of the consensus understanding in that field of research. Efforts for data automation and metadata capture need to take this into account. Approaches to data and metadata will need to evolve with changing theories and understanding of the underlying system without requiring complete reprocessing of existing data.

Pursuing Standards and Provenance Development for FAIR Data

Data standards are an essential means to achieving interoperability but require intentional and purposeful efforts to harmonize data and use common tools. Standard formats enable syntactic interoperability, and common data elements and standard terms and ontologies enable semantic interoperability. For example, standards for Application Programming Interfaces (APIs) assist with data access, movement, and the ability to bring analysis to data. However, implementing standards is currently time- and labor-intensive and is often the rate-limiting step affecting interoperability. It is essential to explore less burdensome, automated ways of implementing and mapping across data standards, as well as providing testbeds to assess the impact of standards on interoperability goals. In

[16] https://science.nasa.gov/open-science/mars-mapping-open-source/
[17] https://nationaldataplatform.org/
[18] https://www.rd-alliance.org/groups/fair-research-software-fair4rs-wg/
[19] https://www.rd-alliance.org/groups/fair-machine-learning-fair4ml-ig/

addition, community involvement and research are necessary to ensure that standards meet the specialized needs of each community, thereby increasing the odds of these communities adopting them.

Other essential enablers of data reuse and interoperability are data models and the ability to map data from one model to another. More robust and agile approaches are needed for extending, expanding, and enhancing data models, including standardized formats and vocabularies. Efforts to develop standards require continuous community-level maintenance and enhancements, often involving time-consuming, labor-intensive efforts that are often un- or under-rewarded in traditional career paths. R&D of AI tools can help to create more streamlined and less labor-intensive processes, resulting in enhanced data models. Additional effort focused on simplifying integration and mappings is required to develop and maintain data models.

Standardizing contextual information — about the motivation for creating a particular dataset, the provenance of actions and processing, statistical properties, and other attributes — enables informed, responsible, and ethical use and reuse of data. There is a need to understand which contextual elements best enhance ethical reuse, reduce the burden of creating such information, and make this information machine-readable and actionable. This need, and its importance in enabling informed, responsible, and ethical reuse of data, will be further elaborated in Strategy 4: Prioritize Privacy, Confidentiality, Ethics, and Security. The development of data sheets and other similar documentation is a promising first step, but further investigation is needed to assess the impact of such metadata and develop a more holistic approach to provenance across the data lifecycle.

Persistent Identifiers (PIDs) are essential for reproducibility of research, long-term stewardship of digital objects, and cultivation of a culture that acknowledges value in producing a variety of research products. Further exploration is essential to ensure the metadata associated with PIDs is actionable, and to link PIDs (for example, for papers and the associated data or specimen IDs) in agile and robust ways. A thriving data ecosystem requires a network of linked PIDs for a variety of ecosystem components, such as data, software, instruments, people, publications, and institutions. For example, PIDs associated with data that are compliant with standards may enable effective and accessible ways to manage and search community-level data collections. More research and evaluations are needed to develop PIDs metadata to support a variety of science use cases across the current and evolving future ecosystem. Additional emphasis is needed on the use of PIDs reproducibly for versioning, data subsets, and data collections.

Leveraging AI Across the Data Life Cycle

AI and automation can make it easier for researchers to create and steward FAIR data throughout the data lifecycle. This life cycle typically starts with data creation or acquisition and proceeds through stages such as processing, analysis, curation, delivery, and use/reuse.[20] Augmenting tasks such as the capture of information directly from workflows, performing quality control and quality assessment, data reduction, and any number of automatable data management and processing steps can significantly enhance data search, discoverability, and reuse.

[20] Hanisch, R., et al., NIST Research Data Framework (RDaF), Version 2.0, National Institute of Standards and Technology, 2024, https://doi.org/10.6028/NIST.SP.1500-18r2.

AI is already being used in some repositories for quality control, anomaly detection, and other curation tasks. With the rise of generative AI, there are also opportunities to use these capabilities in areas such as data standardization or data mapping to specific ontologies (and between ontologies). Further research could explore using generative AI to automate the development and use of ontologies, taxonomies, controlled vocabularies, and common data elements. This could help to reduce the workload while increasing adoption and collaboration.

Technologies such as generative AI can be used to input missing data, including image data, thereby creating more dense datasets for AI training or other analysis capabilities. Generative AI can also assist in the creation of synthetic datasets which are in demand, for example, as replacements for data that would otherwise be sensitive and require limited access. For example, synthetic data (together with validation) has significant potential in biomedical and health research. The addition of synthetic data can also support more robust ML model development and testing, often reducing the temporal and economic toll of gathering copious real-world data. Advancements in AI will continue to advance data itself and the tools and findings that stem from it. Continuous learning AI paradigms can support the analysis and curation of time-series data, including real-time data.

Stewarding Data Assets as Long-term Infrastructure

Data repositories and the communities that govern and maintain them are critical building blocks of the data infrastructure. They provide data, tools, and related resources for reuse and are a primary means by which the federal government provides access to the data resulting from federally funded research. The vision for an adaptive data ecosystem relies on strengthening the reliability and capability of repositories and enabling user-driven analyses of data from multiple sources. Research repositories should be considered as long-term infrastructure and be supported as such to provide continuity of service to the research community and a fertile environment for training future generations of scientists.

Since 2016, the number of data repositories has increased and data reusability has been enhanced through development of repository standards and guidelines. Such guidelines included the TRUST (Transparency, Responsibility, User Focus, Sustainability, and Technology) Principles[21] and the *Desirable Characteristics of Data Repositories for Federally Funded Research* guidelines,[22] which provided measurements of what would be trustworthy, reliable, and persistent data repositories. External credentials such as CoreTrustSeal[23] have also supplied access to both open and restricted research data. Continued focus is needed on evolving the quality standards and ensuring interoperability among ecosystem components. Aligning with such guidelines requires extensive data curation, typically performed by a repository in collaboration with data submitters. Data curation continues to be a rate-limiting process that delays sharing. It is also a process prone to human error.

[21] Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

[22] https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf

[23] https://www.coretrustseal.org/

Robust and effective data curation tools, processes, and skilled workforce are needed, including automated or intelligent mechanisms, to perform as much data curation as possible.

Developing and Stewarding New, Strategic Data Assets

Shared data, much like new scientific instruments, have spurred unique advances in research and innovation. For example, ImageNet[24] and MNIST (Modified National Institute of Standards and Technology)[25] databases propelled the development of image-based deep learning. The Protein Data Bank[26] facilitated new AI advances in protein folding. Additionally, access to geospatial data has led to map products that have transformed our personal navigation capabilities. However, the increasing usage of technologies like cloud computing has led to the creation of multiple copies of data, making it difficult to maintain data authoritativeness. Data authoritativeness signifies a "single source of truth" for a specific domain and is often used as the reference point for decision-making and scientific research. It is becoming increasingly important to develop robust methods for validating authoritative data to ensure that innovations are built on reliable and trustworthy information.

Metrics and other mechanisms to prioritize data and information assets to propel innovation and provide support during times of crisis should also yield suggestions for new data assets that should be developed from the perspective of cross-disciplinary research or national security. Traditional prioritization mechanisms within any single agency might miss opportunities and needs. Cross-agency efforts are needed to identify data and information gaps and explore how to develop new data assets with broad applicability and benefit. Leveraging the adaptative data ecosystem by supporting a broad spectrum of applications improves the likelihood of identifying such gaps. This approach suggests the need for an "adaptive data ecosystem stress test" where problems that cannot be addressed for want of data or tools are identified.

Some applications will require a co-design approach whereby data generation and applications are developed in tandem. This has proved to be a promising approach for AI/ML applications where performance characteristics of the AI model can be tuned with adjustments to the underlying training, inference, or prompting data. Current efforts are largely based on educated trial and error. Understanding the theoretical link between data and AI model performance — and being able to better articulate what it means for data to be AI-ready — requires additional investigation. Additionally, it is critical to address the issue of authoritativeness, ensuring that shared data used in these AI/ML applications is validated.

Given the growing importance of AI in almost every aspect of science and society, it is increasingly important to understand what is necessary to ensure that shared data is useful for the development of new AI technologies and applications. Federally funded research activities generate a wealth of increasingly available and reusable data for research communities; however, only some of these data can be used efficiently and effectively by AI/ML applications. This emphasizes the need for a data-

[24] https://www.image-net.org/

[25] https://www.nist.gov/itl/products-and-services/emnist-dataset

[26] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.

centric AI approach where the focus is not just on developing advanced AI models but on systematically improving data throughout the AI lifecycle.

Preparing data for AI/ML applications (making data "AI-ready") is more than formulaic and requires engagement with and feedback from AI/ML model developers. Some aspects of AI-readiness are better understood than others. For example, data must conform to specific data formats to be analyzed by AI/ML tools that build and deploy AI/ML applications, such as PyTorch[27] and TensorFlow.[28] Through data and metadata standards such as ontologies, taxonomies, and controlled vocabularies, the FAIR Principles facilitate combining data from different sources to collect the data necessary for AI/ML. This data-centric approach ensures that the underlying data foundation of AI is robust and reliable.

Other aspects of what is needed to make data AI/ML ready must be discovered through iterative and exploratory testing. The challenges include how to best represent data and information for a particular AI/ML use case, how to correct for noise, understanding and managing correlations among the data, and what level of specificity or uncertainty of labels is tolerable for a desired AI/ML application. Some initial progress has been made by community-driven initiatives like Croissant,[29] a high-level format for AI-ready datasets that seeks to enhance the accessibility, usability, and tool support for existing datasets.

Decentralized ML, also referred to as federated or distributed learning, is a paradigm of ML where a model is trained iteratively on data in multiple locations. This paradigm can facilitate the use of data that, for privacy or other reasons, cannot be aggregated or moved. Preparing data for decentralized ML requires harmonization and testing as well as capabilities for standardized access and, possibly, enhanced data and model governance to protect privacy. Decentralized ML represents an area in which FAIR Principles could significantly improve the research process. It is also crucial to ensure the authoritativeness of decentralized data sources.

Repositories are increasingly using mechanisms such as hackathons and user engagements to help prepare and process their data to be used in AI applications. The results from these assessments can be shared with researchers submitting data to the repositories so that data could be more AI-ready from the outset. Hackathons, particularly if tied to evaluations, can help identify issues with the data repository. Multi-repository hackathons could also be harnessed to look at integration challenges. Additionally, hackathons have significant potential to scale up when virtual.

Beyond AI-readiness, new computing paradigms such as quantum computing, neuromorphic computing, and DNA computing, will likely require new representations of information and data, including data models, metrics, and concepts of provenance. A deeper examination is needed to understand what types and formats of data will advance emerging computational paradigms.

[27] https://pytorch.org/
[28] https://www.tensorflow.org/
[29] https://github.com/mlcommons/croissant

Strategy 2: Enable Innovative, User-driven Data Science

Data enables human insights and decision-making. The envisioned adaptive data ecosystem presented in this document will allow researchers to explore connections and hypotheses in flexible ways. Researchers ask questions of data through a combination of reusable, general purpose tools and workflows, and cutting-edge capabilities that may open new avenues of research. Tracking uncertainties and potential biases in data (data quality) and data analysis (model approximation) enables appropriate and responsible interpretation of analysis results.

This strategy encompasses five key areas: Advancing Data Services; Enhancing and Developing Multipurpose Tools for Data Analysis and Use; Advancing New Capabilities for Deriving Insights from Data; Enhancing the Connection between Data Repositories and Users; and Developing Metrics, Validation, and the Explanation of Uncertainties.

Advancing Data Services

Search engines, data libraries, access tools, networking and data transfer tools, and ontologies and knowledge graphs are data services that enable users to find, access, integrate, and reuse data. Yet most data services exist to serve specific projects or communities. Universal data services remain out of reach and require technologies that are extensible and composable. Efforts to build knowledge graphs and universal metadata schema, such as FAIR Digital Objects or Schema.org, along with the upcoming updates to the federal data catalog metadata standards[30] could prove useful. More efforts are also needed to ensure that data services are FAIR and stewarded as long-term infrastructure.

Enhancing and Developing Multipurpose Tools for Data Analysis and Use

Enhanced visualization tools will enable diverse audiences to derive value from massive data sets across many use cases and data types. Moreover, new technologies such as those with virtual- and augmented-reality capabilities are promising, not only for enhancing the productivity of experts, but also for making data-driven information accessible to a broader audience. Increased R&D efforts are required to effectively visualize complex data and information which encompasses unstructured and multimodal data. These visualizations are crucial for enabling more intuitive, real-time, high-risk, and high-reward decision-making.

New tools and capabilities are also needed to enable interactive and scalable data exploration that enhances or replaces human sight, empowers more diverse data consumers, and stimulates new approaches for data interpretation and analysis. The ability of LLMs and other generative AI capabilities to summarize large amounts of data and converse with humans could prove to be a useful alternative or complement to visual examination of data.

Search and discovery tools for data are poised for significant advancements in terms of performance, scalability, and transparency where biases in search are clearly conveyed and understood. Various tools are needed to serve the plethora of search and discovery use cases, ensuring that search results become more intuitive through visualizations and interactive features.

[30] https://doi-do.github.io/dcat-us/

Advancing New Capabilities for Deriving Insights from Data

AI/ML advances show the power of data-driven techniques to make enormous strides in decision-making and data discovery. For example, models can help researchers determine which data are informative and worth retaining for analysis while removing excess information. They can perform self-consistency and other checks of the data, providing a measure of confidence in analysis results. These data-driven technologies can become more capable, efficient, and explainable when combined with scientific theory and techniques such as modeling and simulation.

Digital twins are virtual models that incorporate both simulation and real-world data for a more accurate, longitudinal representation of a complex system. Digital twins have the potential to fundamentally change how big data, physical simulation, and precise sensing combine to drive next-generation technologies. Real-time multimodal data acquisition and analysis methods, integrated with AI-enhanced simulations, have the potential to generate digital twins of highly complex systems ranging from personalized medicine to aircraft maintenance. The tools and techniques necessary to process, compress, move, and integrate the massive amounts of necessary data require further research. Similarly, new foundational mathematics is needed to ensure the accuracy and robustness of digital twins, including validation, verification, and uncertainty quantification techniques. The massive scale of data that modern sensors and active measurement devices can generate precludes storing, moving, and analyzing data in its raw form. This means digital twins and other workflows require multi-stage processing pipelines which analyze some data at the edge before merging that data with state-of-the-art systems simulations farther upstream.

Enhancing the Connection between Data Repositories and User Communities

Currently, researchers spend considerable time and effort understanding aspects of data and overcoming barriers that have substantial impact on their analyses, such as identifying sources of noise or uncertainties. Enhancing the relationship between data repositories and their user communities could help researchers more readily assess, for example, when two datasets are compatible; how to identify noise; and whether the dataset is balanced or otherwise fit for purpose. Drawing on the expertise of data repository owners and managers will enable researchers to reuse data more productively and easily.

Developing Metrics, Validation, and the Explanation of Uncertainties

Metrics are currently being developed and deployed to help illustrate the impact of data and data sharing. Open Metrics[31] and Make Data Count[32] are examples of such efforts. The workshop report for *Pioneering the Future of Federally Supported Data Repositories*[33] discussed invoking the FAIR principles for "FAIRed" impact measures on how data have been found, accessed, interoperated, and reused. Further development of metrics will help researchers identify and prioritize data that meet their research needs. Community involvement in this process can ensure effective metrics that meet their needs.

[31] https://openmetrics.io/
[32] https://makedatacount.org/
[33] https://www.nitrd.gov/pubs/NITRD-BDIWG-Workshop-FSDRs-2022.pdf

As the agile data ecosystem develops, improved metrics will also be needed to assess the availability and robustness of various elements of the ecosystem (the degree to which data are current, the availability of research cyberinfrastructure, the performance of scientific instruments), perhaps in real time. Such metrics could support researchers in determining when to run a particular workflow, or which datasets to use.

Metrics and measures of uncertainty, both quantitative and qualitative, are needed for workflows. Researchers need interoperable tools with computational provenance information for assessing how data and processing steps affect the reliability and interpretation of analysis results. Providing this information in near real time will be essential for responsible and accurate decision making for some use cases.

Strategy 3: Develop and Enhance the Robustness of the Federated Ecosystem

The ability to combine data in new and innovative ways to derive insights requires a federated system of shared infrastructure that prioritizes interoperability. This ecosystem's anticipated complexity and dynamism requires new understanding and services to assess and manage the infrastructure in real time. Shared computing, storage, networking, and data infrastructure will require careful management to ensure workflows do not interfere with one another and to ensure information is not inadvertently leaked. There are still open questions regarding the mathematical and technical foundations for designing, developing, and deploying federated, shared infrastructure.

This strategy encompasses seven key areas: Developing National-Scale Infrastructure; Creating Workflow Capabilities that Traverse the Ecosystem and Data Life Cycle; Developing Data Services for Federated, Shared Infrastructure; Enhancing the Energy Efficiency of Data Management; Managing Data Volumes; Managing Rapid Technology Changes; and Pursuing Governance Paradigms for Distributed Accountability.

Developing National-Scale Infrastructure

For many scientific areas, the scale of infrastructure needed to handle current and predicted data rates and volumes is beyond the scope of any single research organization or funding agency. To address this issue, a national-scale data and computing infrastructure are needed. Some prototypes, such as the National AI Research Resource Pilot,[34] are steps in this direction. The benefits of such a national-scale infrastructure include the promise to broaden participation in data science; enable faster, easier, and more equitable access to data-driven science; and enable science at scale. For an adaptive and interoperable data ecosystem at the national scale, it is critical to develop reference architectures that allow designers to understand and define requirements for all ecosystem components: Experimental and Observational Infrastructure, Knowledge Infrastructure, and Research Cyberinfrastructure,[35] and to address interoperability among those components. Such reference architectures will also help funders and operators distinguish components of the ecosystem that require long-term sustainability

[34] https://new.nsf.gov/focus-areas/artificial-intelligence/nairr

[35] https://www.whitehouse.gov/wp-content/uploads/2021/10/NSTC-NSO-RDI-_REV_FINAL-10-2021.pdf

versus ones where agility is needed. A variety of reference architectures need to be tested for their impact on interoperability and other ecosystem goals.

Creating Workflow Capabilities that Traverse the Ecosystem and Data Life Cycle

The envisioned agile data ecosystem is characterized by federated systems of resources that are shared by multiple users and applications and, in some cases, dynamically adjusted to meet user needs. Ensuring reproducible science and repeatable workflows will require new research on software containerization, software-hardware compatibility, and standard APIs. Additionally, there needs to be a focus on managing dynamic, complex systems of shared software and hardware resources.

Increased effort will be needed for the development of composable workflow tools. Workflows that are local to a scientific instrument or laboratory, perhaps using a notebook or local data acquisition system, will be seamlessly connected to other workflows that manage allocations on computing systems or that triage data before submitting to a repository. Scientific workflows, digital laboratory notebooks, and other tools to capture and share processes are relatively new compared to other elements of the data ecosystem. Interoperability and composability of these tools remain a significant challenge.[36]

Developing Data Services for Federated, Shared infrastructure

Data services provide researchers with access to ecosystem-level capabilities and information that are beyond the capability of any single node or component in the federated system. Examples of this include data search engines; scheduling and allocation tools for computing resources; data access mechanisms; data transport tools; and the virtualization of storage infrastructure. Data services become increasingly complex as the ecosystem grows and becomes more federated, incorporating more shared infrastructure. Hence, as the ecosystem develops, effective and robust data services will require continual maintenance. Inexpensive, efficient, and effective data services that meet the needs of current and future use case patterns need to be developed. For example, new allocation tools will be needed for incorporating a wide variety of computing hardware, including quantum computing, and search capabilities will need to significantly advance to include uncertainties, bias metrics, and other information for responsible reuse. Furthermore, increased desire to automate aspects of data management and analysis will require data services to interface with machines.

Enhancing the Energy Efficiency of Data Management

Over the past decade, there has been an enormous increase in the volume, variety, and quality of shared data, along with advances in methods of analyzing previously collected data. Current standard technologies used for data storage and computation are environmentally expensive. Many of these technologies rely on scarce resources like rare earth minerals and use large amounts of water and energy dedicated to data center operation and cooling. Therefore, embodied emissions and electronic waste is a byproduct of data storage and computing infrastructure. The current rate of data generation is outpacing space and resources available. However, data reuse and curation can help mitigate these issues, along with continued federal focus on research to transform the sustainability of computing and

[36] https://nvlpubs.nist.gov/nistpubs/ir/2024/NIST.IR.8520.pdf

data storage. Some fruitful avenues include DNA-based storage and data reduction across the data lifecycle. Other examples include the IARPA (Intelligence Advanced Research Projects Activity) Molecular Information Storage,[37] NSF's Design for Environmental Sustainability in Computing program,[38] and the Department of Energy (DOE's) investments in data reduction.[39]

In a 2020 report,[40] the NSTC Subcommittee on Future Advanced Computing Ecosystem called for a computing ecosystem with "reconfigurability, programmability, reliability, and energy-efficiency" as its key attributes. The need for efficiencies in terms of energy and raw materials is particularly acute in big data applications. For example, training using GPT-3 can consume as much energy as 1,450 U.S. households per month.[41] Furthermore, the uneven distribution of associated environmental costs from this scale of computing has raised ethical concerns.[42] Sustainable growth in data innovation will require new energy and resource-efficient advances in computing and storage architecture, along with continued federal investment in lower energy technologies that are on the horizon.

Managing Data Volumes

The widespread adoption of digital technologies across industries, the increasing number of instruments and smart sensors, the development of an internet of things, and other technological advancements are creating data at unprecedented rates. Today's large-scale simulations are producing vast amounts of data that are revolutionizing scientific thinking and practices. For example, aggregate data generation across DOE's Basic Energy Sciences light sources will approach the exabyte range per year by 2028;[43] and the international fusion energy experiment, ITER, is expected to need more than an exabyte of storage by mid-2030s,[44] which is the equivalent to over 100 billion digitized songs. Additionally, the current and upcoming exascale supercomputers, such as Frontier[45] and Aurora,[46] are capable of computing at 10^{18} operations per second, with storage systems expected to write data at 10^{12} bytes per second. Furthermore, the National Institutes of Health (NIH's) Sequence Read Archive is now 12 petabytes, and recent upgrades to the Stanford Linear Accelerator Center National Laboratory's Linac Coherent Light Source are expected to produce up to one petabyte of data per day. The result is an ever-increasing size of data repositories, leading to increased pressure on data storage platforms and data transport, as well as expanding cybersecurity concerns and associated

[37] https://www.iarpa.gov/research-programs/mist
[38] https://new.nsf.gov/funding/opportunities/design-environmental-sustainability-computing-desc
[39] https://news.fnal.gov/2021/09/doe-invests-13-7-million-for-research-in-data-reduction-for-science/
[40] https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf
[41] https://www.ll.mit.edu/news/ai-models-are-devouring-energy-tools-reduce-consumption-are-here-if-data-centers-will-adopt
[42] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922.
[43] https://www.es.net/assets/Uploads/22-SN-35038-BES-Report-_Full.pdf
[44] https://www.es.net/assets/Uploads/20220613-FES-Final.pdf
[45] https://www.olcf.ornl.gov/frontier/
[46] https://www.anl.gov/aurora

costs. Additionally, researchers are finding it increasingly difficult to find the data they need or to carry out analyses in reasonable time and cost parameters.

Substantial efforts are underway to create the capabilities needed to manage, organize, and analyze these data, yet the size of data continues to be a major cost driver for data-driven science, and increasing data volumes continually stretch the scalability of data management and analysis capabilities. The tension between the vast amounts of data generated for scientific analysis and the ability to manage, store, and analyze those data is expected to continue. A multipronged strategy needs to be developed and employed, both to scale up capabilities and tools and to strategically reduce the growth of data without compromising scientific outputs.

When determining what kinds of data need to be retained and curated and for how long, versus what data does not hold long-term scientific value, decision should be informed by the community, specific use cases, and/or potential impact, including peer review should considered as part of the process. This decision-making process extends from single principal investigator science (for example, determining if it is scientifically necessary to keep and share all generated raw data) to large scale science (such as implementing Large Hadron Collider upgrades that will eventually filter out most of the data generated at the source before sending it through the data system). Implementing multiple approaches at various points in the data lifecycle are needed. Potential strategies could involve:

- Employing better workflow solutions that remove the need for duplicated data.
- Enabling federated or distributed analyses to reduce the need for multiple copies of data.
- Deploying data compression and reduction techniques across the collection and processing pipelines to prioritize information capture.
- Developing approaches to data curation and cost-benefit approaches to data retention.
- Exploring novel data storage technologies for more cost-effective and dense data storage technologies such as storing data in synthetic DNA sequences.

New data models are also needed to deal with vast amounts of metadata, as seen in examples such as genome annotations. Machine-readable metadata can facilitate new capabilities in search, discovery, and interoperability. As metadata becomes more complex and requires comparable or greater storage than the associated data, fundamentally new approaches to data management, storage, and retrieval are needed.

On a more macro-scale, the number of repositories and the global volume of data also needs to be managed strategically. As highlighted in the *Big Data: Pioneering the Future of Federally Supported Data Repositories Workshop Report,*[47] there is a need to foster a culture of continuous evaluation and improvement of Federally Supported Data Repositories (FSDRs) and other infrastructure investments to maximize scientific impact from data. Thinking of large-scale data resources as infrastructure with a lifecycle consisting of design, build, operate, and decommissioning could also be helpful.

Managing Rapid Technology Changes

One way in which the federal government has enhanced the capabilities of data repositories is through the connection to large-scale computing resources, such as commercial cloud platforms or high-

[47] https://www.nitrd.gov/pubs/NITRD-BDIWG-Workshop-FSDRs-2022.pdf

performance computing resources. There has been a thousand-fold increase in computational capabilities across the computing landscape, making computing on data more streamlined and elastic.

The multiplicity of computing hardware, providers and business models, and analysis tools and platforms are an advantage in meeting the wide variety of analysis needs but is also a challenge in developing interoperability and enabling federated analyses. Rapid changes in the underlying hardware and software pose risks to computing workflows and the ability to store, move, and access data. Avoiding technology debt across the ecosystem is a significant concern that requires attention to identify new technology opportunities and risks and to develop software and data management techniques that can quickly adapt. For example, industry partnerships with cloud service providers that manage the transfer of data to new storage hardware on behalf of their customers can be of significant benefit.

As already apparent, the next decade promises to create multiple novel computing architectures for specialized applications such as AI, and many of these architectures are being developed by and for industry. Similar innovation and services are needed for data storage, management, and transport. There is a need to develop innovative cloud solutions for data management to optimize cloud workloads, and to develop standards and APIs for interoperability. Furthermore, the promise of cloud computing to democratize access to data and greatly broaden participation in data-driven research has not been fully realized. More effort is needed to resolve "last mile" challenges in cloud access and training.

Pursuing Governance Paradigms for Distributed Accountability

Data governance is a combination of technology, policy, and data management that engenders trusted relationships among data stewards and users. Appropriate data governance is essential to realizing the promise of an agile data ecosystem. Data governance is an explicit focus of stewards and users of controlled access or sensitive data where bidirectional trust is a prerequisite for data access. Data governance and trust in data integrity is important for open data as well, but it is not as explicitly prioritized. Further exploration is essential to develop technologies that support trusted data sharing and to assess current and future data governance solutions against the goals of the data ecosystem.

Strategy 4: Prioritize Privacy, Confidentiality, Ethics, and Security

Big data and its capabilities have contributed to incredible progress across a variety of fields, but these capabilities and data have also brought about situations where it is causing or contributing to harm. Increases in computing power, coupled with the increasing availability of data, are leading to amazing advances in science, but also result in new risks. Examples include the increasing capacity for computational systems to identify individuals by their voice,[48] and the ability to reconstruct an individual's face from MRI scans.[49] In both cases, risks to privacy and confidentiality arose because of technologies that post-date the collection and sharing of data. It is vital to build trust and mitigate the risks brought on by the use and reuse of data.

[48] https://www.nist.gov/publications/voice-biometrics-future-trends-and-challengesahead

[49] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8154695/

This strategy outlines a four-pronged approach to R&D for new tools, methods, and governance that ensure the ethics, privacy, and security of data. Each of these elements contributes to building trust:

- Building Transparency: Ensuring every part of the data lifecycle — from collection and curation to processing and analysis — is transparent so other users can understand any limitations of the data, such as, for example, patient permissions and how to use data correctly and most effectively.
- Understanding Bias and Improving Equity: Ensuring the inclusion of diverse voices and potentially impacted groups throughout the data lifecycle to better understand potential misuse or bias.
- Strengthening Privacy, Confidentiality, and Data Security: Ensuring that data is secure, while protecting privacy and confidentiality.
- Pursuing Risk Management, Mitigation, and Vigilance: Ensuring continuous assessment, communication, and management of risks to and from data throughout the data lifecycle.

Building Transparency

Transparency is the foundation of data ethics, privacy, and security. Transparent data refers to data in which the collection, permissions, classification, curation, processing methods, and usage information are openly communicated throughout the data life cycle. This transparency allows analysts, affected groups, and decision-makers to have a clear understanding of the data's features and enables them to make informed and ethical decisions about its use. In addition to measurable statistical and structural properties of data, ethical reuse is enhanced when users understand why the data was collected, its intended use, and what study was being undertaken. This contextual information can give clues about where potential imbalances or biases may arise.

Enhanced understanding of the specific types of contextual information is necessary for ethical reuse or trusted decision-making by the multiple stakeholders in data-driven research and applications. Research is also needed to understand how this information can be organized through documentation and metadata, and, eventually, how it can be machine readable, actionable, and updatable with opportunities for contributions from various data users and stakeholders. New derivative data products should contain metadata with provenance and data versioning approaches. More effort is needed to increase the adoption of practices that disseminate information about incentives and provenance, thereby increasing their utility.

Understanding Bias and Improving Equity

For complex systems and applications of national scale and importance, biases in data are inevitable and can propagate through the data and analysis life cycle. These biases have the potential to result in undesirable outcomes such as inequitable benefit across population groups. Understanding the impacts of biases on scientifically and socially important outcomes is an area where further investigation is needed. For example, AI models trained on homogeneous populations often perform poorly against populations outside that training group. Biases in real-world data are often artifacts of known and unknown behavioral biases. Further studies are necessary to leverage data biases to gain a deeper understanding of the underlying social and real-world factors.

Currently, the most robust way to predict unintended consequences such as bias and inequity from data is through iterative and periodic participation and feedback from relevant stakeholders. To highlight their varying perspectives and scope, these stakeholders include the public, research subjects, the institutions that support data collections and sharing, individuals who engage in data curation and labeling, federal agencies, and foreign governments. Each of these stakeholders will have different perspectives and may weigh risks differently. An examination of ethics is needed to guide the incorporation of varying perspectives and best practices for building consensus around acceptable risk. Scalable approaches to community engagement and approaches that incorporate individual-level input need to be developed. Improving widespread data literacy will also help enable more productive and efficient stakeholder interactions.

An expanded application of sociology is needed to understand a) the most effective ways to communicate risks with the public and between parties exchanging data and b) known research on best practices and how it can be best implemented. Since one of the goals of making government data more open is to encourage data reuse, best communication practices are especially needed to improve the transfer of knowledge about the nuances of a dataset's potential risk through its metadata, documentation, and provenance.

Although some research communities such as human subject research may be well versed in proper modes of communication, existing governing bodies approving data collection activities may not be able to accurately predict risk in this rapidly changing landscape. In turn, they may not be prepared to communicate effectively in the wake of sweeping consequences of data disclosures. It is, therefore, important that governing bodies are aware of and maintain clear channels of communication with experts in data use, risk, and management.

Strengthening Privacy, Confidentiality, and Data Security

Ethical data must protect the confidentiality and privacy of personal or sensitive data. Quickly evolving big data technologies, tools, environments, and practices all introduce new risks to security, confidentiality, and privacy and require new technologies, practices, and policies to keep pace.

When planning for future technological developments, it's important to take into consideration data privacy, confidentiality, and security concerns. For instance, malicious actors are currently able to capture existing encrypted big data in bulk. While this data may not be vulnerable now, it could become vulnerable as advanced computing technologies continue to develop. Data security encryption standards are already being devised. The National Institute of Standards and Technology (NIST), for example, is developing quantum-resistant cryptography to help address this concern.[50] Another example is malicious actors planting backdoors into AI models, representing a functional encoding of large training datasets. Detection of AI models with planted backdoors is an active area of research and is funded by several federal agencies.[51, 52]

[50] https://www.nist.gov/news-events/news/2024/08/nist-releases-first-3-finalized-post-quantum-encryption-standards

[51] https://www.iarpa.gov/research-programs/trojai

[52] https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception

The first requirement for big data security is that cybersecurity cannot be an afterthought; it must be a consideration as environments are established and projects are undertaken. Establishing cybersecurity includes policies for encryption, access control, monitoring, security training, backup and recovery, and response and disclosure should a security incident occur.[53] Identifying best practices for integrating cybersecurity into big data is also essential.[54]

Privacy and confidentiality present another challenge for data-driven research and decision-making. As data analysis tools become more sophisticated, even anonymized data carries a higher risk of re-identification, potentially leading to harm if sensitive information is exposed. Big data increases the risk of re-identification through blending or linking datasets, where combining diverse sources allows patterns to emerge that can reveal individual identifiers. This highlights the need for stronger privacy protections and more robust techniques to mitigate both disclosure and harm. Privacy Preserving Data Sharing and Analytics (PPDSA) offers significant potential to enable data-sharing while mitigating privacy concerns by obscuring details within the data or introducing noise. These methods allow for trend analysis without identifying individuals or revealing underlying data. However, it can involve trade-offs, as the added noise may result in less accurate or incomplete data. There are various technical approaches including differential privacy, secure multi-party computation, federated learning, homomorphic encryption, and synthetic data generation. However, these methods can be extremely resource-intensive and costly, making them impractical as near-term solutions for scalable, big data applications. Recent government efforts have highlighted the potential value of PPDSA technologies and sought to advance their development,[55] but widespread implementation remains as the long-term goal due to significant challenges involved.

Privacy issues require blending technical capabilities with ethical design and strong data communication. There are inherent tradeoffs between protecting privacy, which can limit access to data, and trust in a model's output. Without guarantees of privacy, people and organizations will be unwilling to share their data.

Pursuing Risk Management, Mitigation, and Vigilance

A fundamental challenge to responsible and trustworthy big data capabilities is the potential for harm from data changes based on the technologies and data available, as well as changing societal norms. Initial decisions about data collection or structure can have unpredictable downstream effects with significant consequences. Long-term strategies are needed to ensure the risks from data are continually managed through adaptation.

Risk can be assessed by factoring in the likelihood that an adverse event might occur with the potential impact or harm of that event occurring. Risk assessment can be precise and still account for substantial uncertainty. Reliable risk assessment depends on a deep and broad understanding of the possible outcomes. Existing resources such as the Federal Committee on Statistical Methodology's Data

[53] https://www.ftc.gov/business-guidance/resources/data-breach-response-guide-business

[54] https://www.cisa.gov/sites/default/files/2023-04/principles_approaches_for_security-by-design-default_508_0.pdf

[55] https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf

Protection Toolkit[56] can be used to help with risk assessment. However, deliberate efforts are needed to identify changes to and emergence of risks stemming from the advancement of technologies and the expansion of the federally supported data ecosystem.

The likelihood of disclosing data containing private or controlled information can be determined in part from predictable and defined factors such as who has access to a given dataset. However, many other factors contributing to this likelihood are much more difficult to define. For example, the likelihood that personal information can be derived from a dataset can be elevated by the so-called mosaic effect, which occurs when seemingly uninformative data can be combined to identify what should be unavailable or unidentifiable attributes. In addition to data disclosure, other types of information and deductions from multiple datasets may be private, controlled, or even classified. The ability to make novel observations from many-to-many data set linkages is precisely what makes these efforts valuable for scientific discovery. Indeed, the goal of open data infrastructure is increasingly to facilitate the combination of useful data.

Federal agencies not only need to be reactive to developments coming from the private sector but should continue supporting novel research to find and address risks proactively. Conclusions about risk and mitigation drawn from federally funded work can oftentimes be adopted widely without the added proprietary complications of the private sector. Research advancing the theoretical prediction of the risks arising from many-to-many linkages and application of novel methodologies, such as deep learning within AI, is needed to ensure these risks can be accurately assessed. Practically, research to promote the development of risk assessment and uncertainty estimation methodology is needed to provide an evidence-based foundation to guide risk analysis. Ethical data demands a culture of vigilance. Given the ever-changing risk landscape, technologies and approaches are needed to reassess risk routinely and consistently, and enable unacceptable changes to be flagged and appropriately addressed. Risk assessment should be practiced throughout the entire project lifecycle — from research funding and planning, through design, execution, dataset publication, and post-publication.

Support is needed to provide centralized approaches such as checklists and shared methodology paired with policies to require their use. Review boards should include experts up to date in their knowledge of threats. Transparent data incorporated with diverse perspectives will help make potential pitfalls more visible. However, adapted and new training and education programs will be needed to build this expertise.

Data vigilance demands more than research and technical tools; it requires fostering a culture of continued risk management. This process includes assessing a data project for potential bias and other ethical issues, managing unforeseen consequences, and implementing necessary changes and protections.

It is also important that at least some aspects of this risk management work be accessible to non-specialists. Those involved at all levels of data collection and analysis should be aware of processes and considerations where relevant. Risk management efforts should seek diverse voices from multiple communities to be included in the discussion. Ethically addressing risk will rely on frameworks, which may change depending on situation, domain, and time. Priority should be given to understanding how

[56] https://nces.ed.gov/fcsm/dpt

to best develop and update these frameworks (in coordination with affected communities) and to ensuring they are put into use effectively.

New tools and testbeds are also needed to enable data vigilance and to assess risks and benefits across the data lifecycle. These technical capabilities would enable the systematic examination of data and the quantification of risk, including the analysis of how multiple disparate datasets might interact. Finally, technical capabilities are needed to offer options to mitigate biases and other issues. These capabilities will rest on machine-readable metadata about dataset provenance, structure, collection, and curation. It is also crucial that at least some aspects of this risk management work be accessible to non-specialists so that diverse voices from multiple communities can be included in the discussion.

Strategy 5: Develop Necessary Expertise and Diverse Talent

The adaptive data ecosystem is a critical foundation for decision making, and fully unleashing its potential requires both technology and a workforce equipped with evolving skill sets. This includes not just professionals in established fields, but also the creation of entirely new professions with a variety of skill sets, many of which are just emerging as distinct areas of expertise. Additionally, a data literate community is essential for maximizing the benefits of big data and ensuring widespread engagements across sector. For example, the skills needed to manage a repository are different from those needed to analyze data or develop and use ontologies. Drawing from new, diverse talent pools can help recruit for these specialized positions. Access to continued education and training programs for the existing workforce can also support growth of multidisciplinary knowledge and skill sets in the field.

This strategy encompasses four key areas: Developing Specialized Expertise throughout the Data Ecosystem; Broadening Participation of Diverse Talent; Developing Data Expertise within Government; and Improving Data Literacy Writ Large.

Developing Specialized Expertise throughout the Data Ecosystem

A data ecosystem requires a broad and deep array of skills that can maintain operational stability and robustness as well as innovate and provide new capabilities. Developing and maintaining a data ecosystem requires a variety of experts, some of whom are not currently recognized by traditional forms of attribution, job opportunities, or career paths. These areas of expertise are specific to a given skillset that can be harnessed and expanded for the big data ecosystem. Examples include software engineers who write robust applications; data engineers who design, organize and maintain capabilities for data collection, access, and use; cybersecurity specialists who help protect the infrastructure and data assets; and data scientists who use analytical tools and techniques to extract or deduce meaningful insights. The ecosystem also includes data stewards, repository owners, and information and library scientists to ensure the quality and availability of data and information. Cyberinfrastructure professionals with expertise in computing and networking (to maintain everything from cooling systems to connectivity among datasets and models), along with cybersecurity experts, are also needed.

It is important that this growing workforce is multidisciplinary and includes skill sets in emerging technologies to rapidly address the needs of the adaptive data ecosystem, including increased data

volumes and high performance with such volumes. Additionally, ethicists, lawyers, and social scientists with expertise in data must play a crucial role in shaping and refining capabilities from the initial stages to preparing data for practical applications. They can also contribute to the development of social and technical frameworks for the ethical handling and analysis of data. An adaptive data ecosystem both requires and nurtures this diversity of expertise.

The NIST Research Data Framework,[57] a consensus document based on inputs from a broad range of research data stakeholders, provides guidance that supports the development of expertise throughout the data ecosystem. It covers: a map of the research data space; a guide for stakeholders in research data to understand best practices for research data management and dissemination; and tips to understand costs, benefits, and risks associated with research data management.

Recent reports have identified the needs and opportunities for incorporating new "cyberinfrastructure professionals" into the workforce and for broadening participation in these professions;[58, 59, 60, 61] these reports have also identified the need for appropriate development programs. More work is required to develop consensus around the roles and expertise needed across the data ecosystem. Such an agreement will support robust recruitment and professional development and assist with developing this talent pool.

Broadening Participation of Diverse Talent

Since 2013, data science has become a profession, with data science schools of higher education and data science curricula providing a variety of preparatory skills and experiences. Reducing barriers to entry into data science professions is important for maintaining global competitiveness and broadening participation. NIH noted that individuals interested in studying health disparities with data science often enter the field through the social sciences rather than through a laboratory science background, highlighting possible expansion pathways into data science careers. The broad participation of diverse talent can be nurtured through accessible infrastructure, training, user-friendly tools, visualization, and general improvements to data literacy.

Developing Data Expertise within Government

Developing talent within federal agencies is important for the oversight of data science activities and research. Government agencies will need to improve data literacy within agencies, nurture a culture of data professionals, and pursue hiring and retention mechanisms that respond to market forces.

The government has recognized the need for competitive hiring and recruitment through, for example, the Office of Personnel Management's Data Science series created in 2021 and exceptional service hiring mechanisms. NSF has also launched training programs such as Training-based Workforce

[57] https://doi.org/10.6028/NIST.SP.1500-18r2
[58] https://www.nitrd.gov/pubs/National-Strategic-Computing-Initiative-Update-2019.pdf
[59] https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf
[60] https://par.nsf.gov/servlets/purl/10354843
[61] https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

Development for Advanced Cyberinfrastructure[62] and Strengthening the Cyberinfrastructure Professionals Ecosystem.[63]

It is important to acknowledge that the need for skills in data handling and analysis is cross-agency and spans a range of occupations. Training and up-skilling across agencies and occupations is also important as data science is relevant to many positions and government functions.

Improving Data Literacy Writ Large

Across all populations and sectors, individuals face increasing opportunities and requirements to be both consumers and producers of data. Just as financial literacy is essential to individual productivity and security in today's world, data literacy is also becoming an essential skill. Individuals should understand basic principles about how data can accurately (or with bias) represent the real world and how meaning can be derived from data. They should understand how their own data is used to improve products and services, as well as to generate wealth, and the potential risks to privacy.

In addition to serving the individual, people who are data literate are better positioned to provide input on the development of data-driven technologies and potentially influence the trajectory of these technologies, their regulation, and their prudent use.

More training materials and sandbox-type educational tools are needed to develop an expanded expert workforce in data management and data engineering. Addressing this need can be facilitated in part through long-term support for FSDRs. This support encourages the development of a trained workforce that plays a crucial role in advancing, maintaining, and operating these resources by providing job opportunities and important career paths in FSDRs management. Additionally, long-term support will encourage expert repository managers and stewards not only to address operational needs, but also to conduct research in advanced repository management areas. For example, innovative services and technology components, such as AI-based tools, could enable advanced dynamic infrastructure reconfiguration and repository management through automated capture and tracking of digital asset provenance.

Strategy 6: Enhance U.S. Global Leadership

Data fuel research and innovation and are critical assets for addressing national challenges and opportunities. It is valuable for the U.S. to serve in leadership roles in sciences that produce strategic data assets, develop interoperability and critical tools, and advance our standing as sought-after partners in open science and research. Collaborations with a wide spectrum of stakeholders — such as international partners, private sector, and non-profit entities — are essential for the U.S. to grow its leadership position and ensure data and data science align with democratic values.

Federation among many distinct elements is a hallmark of the adaptive data ecosystem envisioned in this strategic plan. The adaptive data ecosystem is composed of many data resources, computational and network systems, standards, services, individuals, and stakeholder communities. The expertise to

[62] https://www.nsf.gov/pubs/2022/nsf22574/nsf22574.htm
[63] https://www.nsf.gov/pubs/2023/nsf23521/nsf23521.pdf

operate and innovate for the ecosystem resides in U.S. academic, private, and government sectors as well as internationally. Achieving interoperability requires intentional, purposeful efforts to make these elements cohesive and reflective of the needs of multiple stakeholders. Engaging with communities of stakeholders is key to ensuring informed and ethical data practices; and collaborations and data governance are key to achieving the adaptive data ecosystem. Perhaps more than in any other field of research right now, ensuring synergies between private sector and government-supported efforts in big data will be essential to all the above strategies.

This strategy encompasses two key areas: Enhancing Cross-sector and International Partnerships and Pursuing Sustainable Stakeholder Engagement.

Enhancing Cross-sector and International Partnerships

Research is inherently international, and a strong U.S. position benefits U.S. researchers in global collaborations and competition. Critical research data may be collected at international locations, such as earth science data, and by foreign organizations, such as public health research data. Furthermore, the diversity of talent and perspectives that advance science come from across the globe. The U.S. has a robust, globally respected, open, and merit-based research system that is well-supported and fuels global research collaboration. But international research collaborations, data sharing, and other partnerships, such as the Data Privacy Framework[64] and the Joint Statement of Intent between the U.S. and the European Organization for Nuclear Research (CERN),[65] make the U.S. science enterprise even more robust.

There is a need for a strategy that leverages U.S. competitive strengths and principles while simultaneously reemphasizing strong international engagement across the science and technology research spectrum. International efforts are underway for strategies to address their own needs, for example, the European Open Science Cloud, the African Open Science Platform, and the Australian Research Data Commons, which will develop regional IT infrastructure. There is a need to energize cross-agency efforts to develop strategic connections with international counterparts on data and data systems.

As articulated in Strategy 3: Develop and Enhance the Robustness of the Federated Ecosystem, there is also a need to develop national-scale federated data infrastructure, which balances needs for data sharing and interoperability across disciplines with the unique needs of the various constituencies.

Big data is a hallmark of most Fortune 500 companies and major industries today. The private sector has advanced large-scale data collection, management, and analysis capabilities with enormous impact on almost all aspects of life. Federally supported R&D has seeded many of these advances, and federal agencies are now partners with the private sector as consumers and co-creators of technologies aimed at areas of national need. Many experts in big data fields pursue careers that encompass academia, government, and the private sector. However, sustaining productive engagement across sectors necessitates the continuation and enhancement of federal agencies' roles. These roles include

[64] https://www.dataprivacyframework.gov/

[65] https://www.state.gov/joint-statement-of-intent-between-the-united-states-of-america-and-the-european-organization-for-nuclear-research-concerning-future-planning-for-large-research-infrastructure-facilities-advanced-scie/

serving as supporters of foundational and translational research, consumers of private sector technologies and offerings, and co-developers of technologies aimed at areas of national significance.

International partners also play important roles in the adaptive data ecosystem in its broadest sense. There are examples of global standards for data sharing already in place, such as those set by the Global Alliance for Genomics and Health for the responsible use of genomic data. An example of a critical priority area for data sharing is human health, as results have clear and substantial positive global impacts on human well-being. Collaborations with international partners are also important, as the U.S. can benefit from pre-competitive R&D and data sharing, and strategic partnerships can provide irreplaceable and expedited knowledge and resources required to stay competitive in areas such as AI.

Pursuing Sustainable Stakeholder Engagement

As noted in Strategy 4: Prioritize Privacy, Confidentiality, Ethics, and Security, stakeholder engagement is critical to the responsible development of trustworthy and ethical big data capabilities and therefore essential for maintaining open and trusted international collaborations. Productive stakeholder engagement requires understanding and respecting an individual stakeholder's goals, requirements, limitations, incentives, and success metrics. Stakeholders vary widely and are not only the data generators and data users, but also the public that benefits from data results. Dissemination of data results back to the public is therefore an important component of stakeholder engagement. Other examples of stakeholders include industry, academia, government agencies, non-profit organizations, Indigenous communities, research participants, students, funders, end-users, and the Global South as both data partners and stakeholders. The sustainability of a data ecosystem requires active, continued participation from multiple stakeholders, including those beyond our borders. New practices and incentives are needed for longitudinal stakeholder engagement and input.

Sustainability is a requirement for a long-lasting, high-value ecosystem. Respecting differences between nations and cultures, such as providing multi-language resources for training and education, will increase the uptake and potential user base of the system. The sustainability of the system can be maintained through clearly defined policies, regulations, and leadership practices. A transparent governance mechanism decreases the barriers for participation by building trust between parties, limiting a "winner-takes-all" environment, and increasing the incentive to participate in the shared resource. Metrics will also need to be created to measure the value of data, resources, and resulting output to monitor the return on investment and to also know when to sunset programs that are duplicative or no longer valuable.

This strategy will play a key role in shaping a data ecosystem that can meet the challenges of rapidly growing data needs. By continuing to refine tools, infrastructure, and collaborative efforts, this adaptive system will support the efficient use of data across a range of fields. Strengthening partnerships and enhancing automation capabilities will enable federal agencies and the scientific community to maximize the potential of the data. This proactive approach will help sustain the U.S. leadership in data-driven research and innovation while ensuring responsible data management.

Appendix: Synergistic Efforts and Initiatives Related to Big Data Vision

This strategic plan update is one of several initiatives across the federal government seeking to overcome technical and social challenges through national-scale approaches. Some examples of other initiatives and actions include:

- The *National Artificial Intelligence Research and Development Strategic Plan 2023 Update*.[66] This document is synergistic with the AI R&D Strategic Plan by articulating needs in data infrastructure and tools that are the basis for new AI capabilities.
- The *National Artificial Intelligence Research Resource Task Force Report*,[67] which provides a blueprint for an exemplar data ecosystem focused on advancing and democratizing AI.
- The Future of Advanced Computing Ecosystem Report, *Pioneering the Future Advanced Computing Ecosystem: A Strategic Plan,*[68] dovetails with this big data plan with a focus on integrating computing infrastructure to support critical applications and foster strategic partnerships across government, academia, and industry.
- The *Federal Cybersecurity Research and Development Strategic Plan 2023*[69] identifies advances in cybersecurity that are needed to prevent malicious activities and strengthen public trust in the digital ecosystem.
- The Fifth U.S. Open Government National Action Plan[70] provides a framework aimed at enhancing public access to government data to promote transparency and accountability.
- The OSTP Memo *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*[71] articulates U.S. priorities in open science and provides policy guidance to promote improved public access to federally funded research results. This plan further details the R&D needs to advance these priorities.

[66] https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf

[67] https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

[68] https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf

[69] https://www.nitrd.gov/pubs/Federal-Cybersecurity-RD-Strategic-Plan-2023.pdf

[70] https://www.gsa.gov/governmentwide-initiatives/us-open-government/current-action-plan/summary-of-progress

[71] https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf