SPECIAL
COMPETITIVE
STUDIES
PROJECT

Data's Role in Unlocking Scientific Potential

OCTOBER 2024

Rama Elluru and Karina Barao

Contributors: Dr. Tom Mitchell and Dr. David Danks

The United States' innovation power depends on accelerating its scientific research. In turn, scientific research depends on developing theories and methodological tools. Data is a critical component needed to test and transform these theories and methodologies into practical applications for societal prosperity. Without a strong commitment to leveraging its data assets, the United States risks falling behind in the race for technological and scientific leadership.

The United States has many advantages, not least of which is its potential access to massive amounts of data across its public and private sectors.[1] The country is home to global technology companies, world-renowned universities, and it dominates the big data and data analytics markets.[2] The United States is also home to the most data centers, and is the world's largest data producer.[3] With these valuable assets, we should be doing much more to leverage data for societal good, particularly in science.

China understands the value of data and is developing strategic policies to strengthen its competitive technological advantage. A crucial part of Beijing's approach is collecting as

[1] See Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data, Epoch AI (2024).

[2] Bhaskar Chakravorti, et al., Which Countries Are Leading the Data Economy?, Harvard Business Review (2019)

[3] Top 10: Countries with the Most Data Centres, Data Centre Magazine (2024); 11 Insightful Statistics on Data Market Size and Forecast, Edge Delta (2024); See National Data Action Plan, Special Competitive Studies Project at 2-3 (2022).

much foreign data as it can while hoarding its own.[4] Domestically, the People's Republic of China (PRC) is continuing to adopt comprehensive data regulations,[5] incentivizing data sharing, and its leaders are currently considering recognizing data as a commodity.[6] The end result could be China leveraging data to leapfrog the United States in scientific prowess. The United States needs to match this strategic data effort with a comprehensive data policy to remain competitive and strengthen its global leadership.[7]

This paper builds on "How Can AI Accelerate Science, and How Can Our Government Help?" by Dr. Tom Mitchell. Dr. Mitchell lays out how Artificial Intelligence (AI) can help create a paradigm of "community scientific discovery."[8] Here, we delve into the critical role of data in creating a collaborative scientific community and provide concrete recommendations to optimize U.S. data assets for scientific discovery.

Introduction

Scientific progress is being hindered by the lack of a truly collaborative scientific community. The vast majority of today's scientific research, across nearly every field, is siloed. Dr. Mitchell states: "Scientists and their research teams come up with a hypothesis, conduct an experiment to test the hypothesis, publish the results, perhaps share some experimental data, and then repeat this process. Other scientists can build on these results by reading the published paper, but this process is error-prone and highly inefficient for several reasons: (1) individual scientists have no hope of reading all of the published articles in their field, and, therefore, operate in partial blindness to other relevant research, (2) the full details of the experiments described in the journal publications necessarily omit certain experimental data and many details, making it difficult or impossible for others to replicate or build on their results, and (3) the analysis of a single

[4] See National Data Action Plan, Special Competitive Studies Project at 3 (2022).

[5] China's regulations cover data privacy, collection and reporting practices, data storage and protection requirements. See Eva Xiao, China Passes One of the World's Strictest Privacy Laws, Wall Street Journal (2021).

[6] In Depth: China's Efforts to Unlock the Value of Data as an Asset, Caixin Global (2024); National Data Action Plan, Special Competitive Studies Project at 12 (2022); CNIPA Announces Pilot Areas for the Work on Data Intellectual Property, LexisNexis (2022).

[7] SCSP published a national data action plan to address this void. The plan provides recommendations to the United States as regulator, holder of valuable public data, and convener of public private partnerships (PPPs) — collaborations between relevant stakeholders that includes government, industry, and academia). See National Data Action Plan, Special Competitive Studies Project at 7 (2022); Vision for Competitiveness: Mid-Decade Opportunities for Strategic Victory, Special Competitive Studies Project at 37 (2024). In this paper, we apply the PPP recommendations to a critical domain in the global competition, scientific research.

[8] Tom Mitchell, How Can AI Accelerate Science, and How Can Our Government Help?, Block Center for Technology and Society (2024).

experimental dataset is typically done in isolation, failing to incorporate data (and hence valuable information) from other relevant research conducted by other scientists."[9]

Dr. Mitchell provides recommendations for more collaborative scientific discovery. He suggests the U.S. Government support the components necessary to establish novel computerized research assistants that help the scientific community overcome the current fragmentation in scientific research. This includes the ability to share and jointly analyze diverse datasets contributed by the whole scientific community.[10]

Increased access to experimental data will improve scientific hypothesis generation, accelerate the hypothesis-to-experimentation loop, and enable faster deployment of scientific achievements. Robust data sharing mechanisms enable researchers to leverage a global network of insights, paving the way for unprecedented advancements and unthinkable scientific progress.

However, establishing data sharing infrastructure requires addressing the lack of incentives for researchers to share data, which principally results from privacy, security, legal (e.g., liability), and intellectual property (IP) concerns.[11] If a researcher holds a wealth of valuable data, there is little motivation to share it when there is no reliable infrastructure in place to address these legitimate concerns. Without adequate infrastructure, the risks of sharing data often outweigh the benefits. Additionally, establishing and sustaining adequate data sharing infrastructure requires both time and resources (e.g., ongoing maintenance of data). Funding agencies, however, typically do not provide financial support for these efforts.

In addition, not all data that can be used for training AI for science is made available to the scientific community, easily accessible, and in formats that allow for interoperability and AI training. Even when the desire to share exists, the lack of incentives for data sharing, such as monetary or recognition, hinders collaboration.

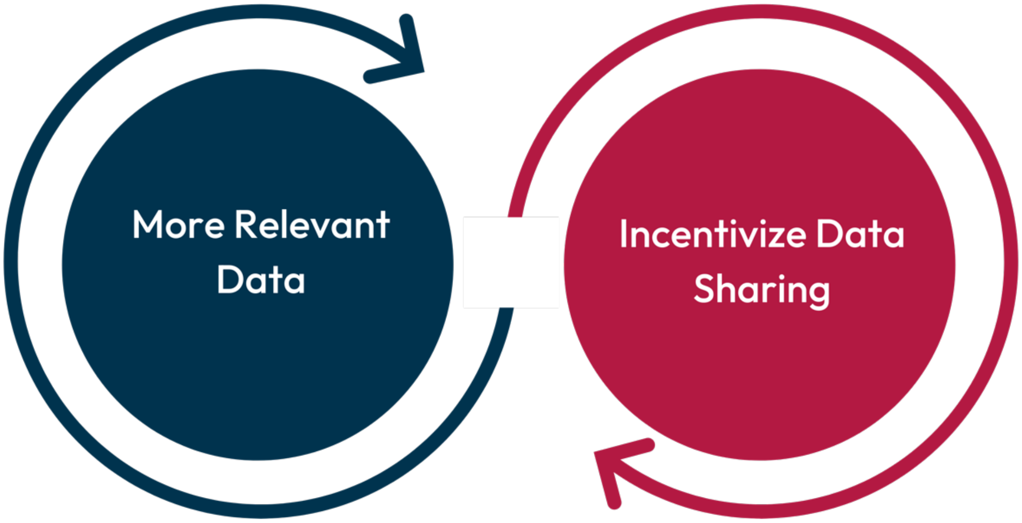
[9] Tom Mitchell, How Can AI Accelerate Science, and How Can Our Government Help?, Block Center for Technology and Society (2024).

[10] Tom Mitchell, How Can AI Accelerate Science, and How Can Our Government Help?, Block Center for Technology and Society (2024).

[11] For example, researchers are unlikely to share their data for another's gain without some sort of compensation, such as monetary or recognition.

Data is the Connective Tissue of the Scientific Community

To make full use of AI and data, the U.S. Government must implement policies that unlock the value of both governmental[12] and non-governmental data assets. This should be done in partnership with industry and academia.[13]



The Need for More Relevant Data

The entire scientific research community – industry, academia, and government (including federally funded R&D centers) – should commit to making more high-quality data available, usable, and accessible. This commitment will further "community scientific discovery" and enable the training and utilization of AI models for scientific endeavors.

To turn the opportunity of data into reality requires:

- **Access to significant experimental data.** Dr. Mitchell writes that, "One lesson from text-based foundational models is that the more data they are trained on, the more capable

[12] Examples of successful U.S. Government dataset programs include: (1) the Materials Genome Initiative, a multi-agency partnership, that aims to accelerate new materials discovery and development by integrating data and computational tools to predict new materials; (2) the NIH Human Connectome Project (HCP), a consortium project to map the neural pathways of the brain, that houses neuroimaging data. (Note: There are regulatory hurdles to sharing MRI image data (e.g., HIPAA)), and; 3) a DOE tool for accessing its scientific datasets. See Materials Genome Initiative Strategic Plan, NIST (2021); CCF Data from the Human Connectome Projects, National Institutes of Health (last accessed 2024); DOE Data Explorer, U.S. Department of Energy (last accessed 2024).

[13] See National Data Action Plan, Special Competitive Studies Project at 13 (2022).

they become. Empirical scientists are also well aware of the value of more diverse experimental data. To achieve multiple order-of-magnitude advances in science, and to train the types of foundational models we desire, will require a very significant advance in our ability to share and jointly analyze diverse datasets contributed across the entire scientific community."[14] This requires access to more relevant data and incentives to share.

- **Developing diverse datasets.** Training AI models for science will, in many cases, require diverse types of experimental data collected across multiple laboratories and scientists. For example, training an AI foundational model on how cells behave, which may lead to advances in understanding diseases such as cancer will require integrating diverse and multimodal datasets.[15] The datasets will include data describing the temporal and spatial distribution of organelles within the cells and how they interact, the expression of various genes across the cell cycle, the impact of disease agents on the cell, how the cell sends and receives messages from neighboring cells, and more. Scientists do not necessarily require massive data quantities, compared to the immense resources required to train today's large, data-intensive language models. Instead, scientists can conduct experiments with significantly less data, so long as it is relevant, high-quality data.

- **Unlocking relevant hidden data.** Dedicated repositories identifying available raw experimental data are necessary to ensure that the research community is aware of data that is not included in publications. In addition, the data that is shared must come with adequate metadata information. For example, scientists need to understand how data was generated, what measurement methods were used, how the data was structured, and other critical details to fully interpret the results. These types of practices would not only preserve the integrity of the scientific record, but also would provide researchers with the available knowledge to facilitate replication and validation of experiments.

Data Sharing

Necessary infrastructure for data sharing is lacking.[16] The United States does not have the resources, talent integration capacity, data management tools, standardization frameworks,

[14] Tom Mitchell, How Can AI Accelerate Science, and How Can Our Government Help?, Block Center for Technology and Society (2024).

[15] See Virtual Cells, Chan Zuckerberg Initiative (last accessed 2024) ("Building virtual cells will require vast amounts of diverse and multimodal biological data.").

[16] SCSP has recommended scaling cloud-based laboratory models, such as Carnegie Mellon University's distributed Cloud Lab, to develop infrastructure capable of unlocking the full potential of biotechnology.

See National Action Plan for U.S. Leadership in Biotechnology, Special Competitive Studies Project (2023).

and data sharing policies, which are essential for the uptake, integration, and analysis of data.[17] If data sharing partnerships are to be established, it will be necessary to rectify these shortcomings.

U.S. Government efforts to enhance data accessibility and quality have increased in recent years, focusing on making datasets more open and suitable for AI applications, which can increase the efficacy of a data sharing partnership. For example, the Department of Commerce (DOC) has established a dedicated working group to develop comprehensive guidelines for publishing AI-ready, machine-readable data.[18] Washington must bolster its departments and agencies efforts through funding requirements or grants.[19] Past and present public private partnerships illustrate how these types of collaborations can be beneficial to science, and advance the development and deployment of AI-ready, open data for the benefit of both public and private sectors.[20]

Recommendations

The Special Competitive Studies Project's (SCSP) National Data Action Plan offers a viable mechanism for implementing the recommendations made by Dr. Mitchell, for the United States to "advance in our ability to share and jointly analyze diverse datasets contributed across the entire scientific community."[21]

SCSP recommends two key actions for the U.S. Government to effectively foster a collaborative data ecosystem. First, the U.S. Government should lead efforts to increase awareness among government, academia, civil society, and industry stakeholders about existing and available data assets. While some data repositories already exist across sectors, a centralized repository that catalogs available datasets would lead to more transparency and efficiency. Second, leveraging its unique convening power, the U.S. Government should spearhead the establishment of

[17] See National Data Action Plan, Special Competitive Studies Project at 31 (2022).

[18] To ensure a collaborative approach, the DOC has issued a request for information (RFI) to gather insights and feedback from stakeholders on the creation, curation, and distribution of its data. In July 2024, the Office of the Chief Data Officer furthered this initiative by hosting a workshop on the Future of AI and Open Government Data, bringing together multidisciplinary experts to discuss the critical elements outlined in the RFI. Preparing Open Data for the Age of AI, U.S. Department of Commerce (2024); 89 Fed. Reg. 27411, AI and Open Government Data Assets Request for Information, U.S. Department of Commerce (2024).

[19] See Nyah Stewart, Fueling Innovation: Insights into Federal AI R&D Investment, Special Competitive Studies Project at 17 (2024).

[20] The COVID-19 High-Performance Computing Consortium, formed in March 2020, united tech companies, academic institutions, and federal agencies to provide researchers with advanced computing resources including datasets to accelerate COVID-19 research, see The COVID-19 High-Performance Computing Consortium, National Library of Medicine (2022); The National Cancer Institute's Cancer Research Data Commons (CRDC) integrates datasets from NIH-funded research and the private sector, enables data contributions and search, and provides computational and analytical resources for the research community. See Cancer Research Data Commons, U.S. National Institutes of Health (2024).

[21] Tom Mitchell, How Can AI Accelerate Science, and How Can Our Government Help?, Block Center for Technology and Society (2024).

formalized data sharing public-private partnerships (PPP) to ensure cross-sector collaboration.

I. Create comprehensive data inventories across scientific domains.

The Secretary of Commerce, acting through the Department of Commerce Chief Data Officer and Director of the National Institute of Standards and Technology (NIST), and in coordination with the Federal Chief Data Officer Council (CDO Council) should establish a government-managed inventory where organizations (e.g., industry, university, research institutes) catalog information about their datasets, including key metadata such as purpose, description, domain, measurements and accreditation. Federal oversight of this repository, similar to existing practices on platforms like data.gov,[22] would promote scientific participation by making high-quality datasets more visible and accessible. To encourage participation, the government could offer incentives such as academic recognition and citation credit for researchers whose datasets are used by others.[23] Additionally, organizations would be responsible for regularly updating their entries as datasets change to ensure the inventory remains current and relevant. The utility, including searchability, of data sharing websites depends on the adoption of data and metadata sharing standards.[24]

Examples of data inventories successfully leading to scientific discoveries include:

- The Broad Institute's Center for the Development of Therapeutics (CDoT),[25] which maintains an extensive inventory of one million molecular compounds. These compounds have known attributes that are effective in treating specific diseases, and AI models have been employed to explore their potential in treating other conditions.[26] Using AI and CDoT's Drug Repurposing Hub, researchers discovered a new antibiotic, Halicin, which can effectively combat certain antibiotic-resistant bacteria, a long-standing scientific problem.[27] Such breakthroughs in antibiotic research would not have been possible without a robust and diverse catalog of molecules.

- The National Institutes of Health (NIH) houses the National Library of Medicine (NLM), the world's largest biomedical library, which contains extensive collections of clinical trial

[22] The Government Services Administration (GSA) keeps a catalog of federal agency data. See Data.gov (last accessed 2024).

[23] For example, Google Scholar primarily indexes its results on published papers. If researchers were able to receive academic recognition for their datasets, by making them accessible to others, it could provide greater incentive for researchers to share their data.

[24] Several scientific data sharing platforms already exist, like the Open Science Framework (OSF) and ResearchBox, among others. However, the abundance of these platforms underscores a challenge: the absence of a single, centralized repository for finding usable scientific data.

[25] The Center for the Development of Therapeutics (CDoT), Broad Institute (2024).

[26] Dhruv Khullar, How Machines Learned to Discover Drugs, The New Yorker (2024).

[27] Nancy S. Loving, Researchers Use Artificial Intelligence to Identify New Antibiotic, Equimanagement (2024).

data, biomedical literature, and health informatics. Its resources are essential for developing new therapies, improving healthcare practices, and advancing public health outcomes. By collaborating with institutions globally, the NLM promotes innovation and ensures scientific information is accessible for solving critical medical challenges.[28]

- NASA's Exoplanet Catalog is a continuously updated database with detailed information on over 5,600 confirmed exoplanets. It provides 3D models and statistics for each, enabling scientists to explore distant worlds and allowing researchers to compare planetary characteristics, study systems, and track discoveries, all of which enhance the understanding of planetary formation beyond the earth's solar system.[29]

II. Create scientific data sharing public-private partnerships.

The above U.S. Government entities should establish dedicated repositories for scientific and experimental data across various scientific disciplines through public-private partnerships.[30] Identifying a department to take the lead on this action will prevent valuable data from being siloed across various departments and agencies.

These U.S. Government entities can authorize appropriate senior leaders to oversee stakeholder engagement to create public-private partnerships for sharing data in priority[31] scientific domains (e.g., materials science, biology, chemistry, computer science). Each should be assigned specific scientific challenges to solve.[32] The PPPs should be equipped with agreements and controls surrounding privacy, data security, and democratizing access for SMEs and researchers. They must also create incentives,[33] address anti-competitive concerns, and craft proprietary IP or IP-type protections tailored to the relevant PPP.

Different scientific fields face unique real-world constraints when it comes to establishing effective incentive structures for data sharing. For example, the considerations for drug design, materials science, and neuroscience can vary significantly, requiring tailored approaches to address each field's specific challenges. Targeted and narrowly scoped PPPs can help mitigate these constraints by aligning incentives with the distinct needs of each research domain. For

[28] Accelerating Biomedical Discovery and Data-Powered Health, National Institute of Health (2024).

[29] Exoplanet Catalog, National Aeronautics and Space Administration (2024).

[30] See National Data Action Plan, Special Competitive Studies Project at 30 (2022).

[31] DOC can manage domains outside of its mission space with guidance from relevant agencies such as the Department of Health & Human Services, Department of Energy, and National Institutes of Health.

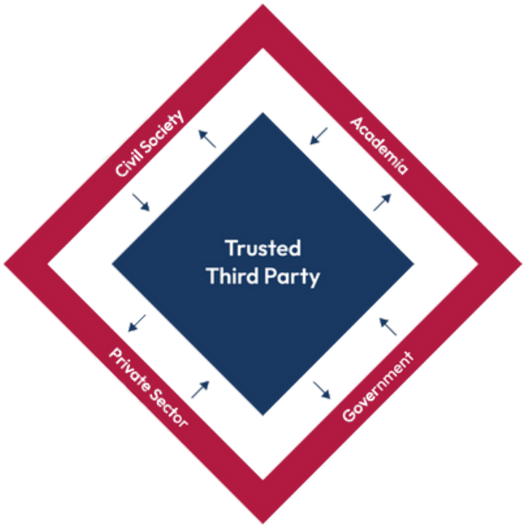
[32] When the processes for building and maintaining successful data sharing PPPs have been established, the processes can be scaled to address larger and larger scientific challenges. PPPs can further be combined where appropriate.

[33] Incentives can be in the form of acknowledgment through citations of data sources, monetary compensation, enlarging and sustaining free market competition, or shared commitments for societal prosperity.

instance, scientists working with personally identifiable information (PII) have already developed best practices for accessing and using data ethically.

The United States, either by acting as a broker or by enlisting a trusted third party, should require that all data shared adheres to standards that ensure quality and interoperability.[34] The PPP participants should agree upon data sharing standards, potentially lifted from relevant industry, academic and international data standards. These standards can facilitate functional data sharing, integration, and collaboration among researchers, enhancing the overall reach and impact of the datasets. The PPP itself should provide any support necessary for satisfying these requirements.

Data Sharing Public-Private Partnerships



Various types of existing data sharing PPPs can serve as a model for the PPPs proposed here.[35] These PPPs often use trusted third parties for data access and sharing to unlock the opportunities of aggregating private and public sector data in a controlled and trusted manner for the collective benefit of all participants. Each PPP is shaped by its mission and participants, but there are common elements across successful data sharing PPPs:

- They are formed to address a discrete, critical and urgent problem or opportunity (e.g., open ended research), which provides a justification for participants to engage.[36]

- All participants view the PPP as independent and trustworthy, with clear guardrails on data accessibility, sharing, and use.[37]

[34] This would be similar to how the countries and scientists involved in the Human Genome Project enforced data standards they had agreed upon. See Summary of the Report of the Second International Strategy Meeting on Human Genome Sequencing, Human Genome Project (1997).

[35] Ted Senkrecht, Tales & Tips from the Trenches: Extend the Impact of Enterprise Data through Partnerships, MITRE (2021) ("Data-sharing PPPs involve multi-party collaboration around information sharing and analysis to take action on complex problems without boundaries. These PPPs are predicated on shared decision-making, shared resourcing, and shared benefit to the partners and the public.").

[36] Ted Senkrecht, Tales & Tips from the Trenches: Extend the Impact of Enterprise Data through Partnerships, MITRE (2021) (Data sharing PPPs have a "common mission - Partners are driven by a sense of urgency and the realization that their own interests are served by working together on a shared goal.").

[37] Robert Groves & Adam Neufeld, Accelerating the Sharing of Data Across Sectors to Advance the Common Good, Georgetown University at 19 (2017) ("[A] key barrier to sharing data across sectors is trust. Companies want to make sure the government does not use the data in ways that hurt their business interests, and the government taking control of private sector data imposes security and legal risks. The public's privacy concerns also seem to diminish when the government is not the one combining data.").

- There is a strong focus on ensuring privacy and security in the PPP’s operations, to reassure the public that there are adequate protections against unintended use of the shared data.[38]

- The partners establish clarity about contractual and legal issues. These can include potential regulatory violations, the treatment of IP or IP-type protection and data rights, policies for avoiding conflicts of interest, the role of an independent, trusted third party if one is established, avoiding antitrust/anti-competition issues, funding and authorization of the PPP.[39] The potential legal hurdles also depend on the composition of stakeholders (e.g., domestic versus international).

Trusted intermediaries can responsibly accumulate data from a range of private and public entities, link and analyze the data into actionable information, and share both the insights and the underlying data with all parties. This creates a repository of experimental data that the scientific community can leverage for research.

Infrastructure Needed for Successful Data Sharing Public-Private Partnership

Addressing the challenges of data sharing within PPPs requires infrastructure that not only supports collaboration but also mitigates concerns around standardization, access to high-quality and relevant data, privacy and security, IP, and data scarcity. These concerns must be addressed, as they can damage stakeholder participants' trust and stem the flow of data necessary to the PPP. Strengthening the PPP's infrastructure can significantly reduce barriers to data sharing. Below, we outline a series of solutions to these challenges, which can create a secure and efficient data sharing ecosystem.

[38] Robert Groves & Adam Neufeld, Accelerating the Sharing of Data Across Sectors to Advance the Common Good, Georgetown University at 20-21 (2017) (“The technology for protecting privacy has evolved substantially since organizations simply deleted names, addresses, and Social Security numbers from spreadsheets. Technical approaches (such as query tools, synthetic data, and multiparty shared computing) and mathematical methods (such as differential privacy) now allow for far more sophisticated ways to reduce the risk of re-identifying people. … Decisions on technical infrastructure are intertwined with privacy, cybersecurity, and confidentiality concerns.”).

[39] Ted Senkrecht, Tales & Tips from the Trenches: Extend the Impact of Enterprise Data through Partnerships, MITRE (2021).

Infrastructure Needed for Successful Data Sharing Public-Private Partnerships

U.S. Government Data
Data IP Rights
Data Standards
PETs
Synthetic Data

1. Develop clear data standards for interoperability.

The PPP's stakeholders, particularly the scientists and individuals responsible for generating and utilizing the data, should establish clear data standards tailored to the partnership's specific use cases. For example, the data shared should be of high quality, and include documentation of data collection methods, validation, and error-checking processes. The datasets should meet agreed-upon thresholds for accuracy, completeness, and reliability to be shared within the partnership. These standards should also promote rapid and open data exchange, ensuring that datasets are accessible, usable, and interoperable across different domains and institutions within the partnership.[40] Key aspects could include requirements for timely sharing of data, standard formats for documentation and metadata, and availability of data without restrictions to the stakeholders in the PPP. Standards Development Organizations can draw from the data standards and principles established by the PPP, if public, to guide their own standards development efforts. These criteria can help stakeholders create a culture of openness that accelerates scientific discovery and maximizes the value of shared data for the broader research community. Consensus on data sharing standards is more achievable when the ultimate goal is narrowly defined. For example, the success of the Human Genome Project's data sharing can be credited to the Bermuda Principles, which established agreed-upon guidelines by the data contributors themselves for making all human genome sequences publicly accessible.[41] These principles also outlined specific

[40] By adopting a framework similar to the Bermuda Principles of the Human Genome Project (a set of agreed upon principles for the release and public access of genomic data), data sharing PPPs can promote interoperability data exchange that benefits all stakeholders involved in the collaboration. See Summary of the Report of the Second International Strategy Meeting on Human Genome Sequencing, Human Genome Project (1997).

[41] See How a Field Built in Data Sharing Became a Tower of Babel, Nature (2021); Previous U.S. Government efforts, such as brain imagery data centers through the National Science Foundation (NSF), have faced challenges in implementation of standards. It is important to review and understand these lessons learned to inform future initiatives aimed at standardizing domain-specific data.

standards for data quality and sequencing annotations, ensuring that shared data met relevant criteria.

2. Increase access to U.S. Government and national laboratories' data.

Increasing access to more relevant data is a critical component of the infrastructure needed for effective data sharing. The U.S. Government and national laboratories hold invaluable data that could form the backbone of a national research infrastructure, albeit with careful consideration of national security interests when sharing and collating sensitive data.[42] Public-private partnerships, like the National Artificial Intelligence Research Resource Pilot (NAIRR) are already working to facilitate access to both U.S. Government and non-U.S. Government datasets. Increasing scientists' access to existing and proposed government programs, principally the National Science Foundation's Directorate for Technology, Innovation and Partnerships (TIP)[43] and the Department of Energy's Frontiers in Artificial Intelligence for Science, Security and Technology (FASST) program also will be necessary to make these high-quality resources more widely available.[44]

3. Incentivize the creation and use of privacy enhancing technologies.

Protecting sensitive information is critical to maintaining public trust and compliance with standards. Privacy-enhancing technologies (PETs) play a crucial role in addressing these concerns, and they offer sophisticated tools and techniques for doing so.[45] PETs include anonymizing data, differential privacy, homomorphic encryption, secure multi-party computation, and federated learning.[46] By integrating PETs into the infrastructure of a data sharing partnership, organizations can make their data available in a secure environment without compromising privacy.[47] This strengthens security and promotes trust among stakeholders, encouraging further data sharing. In an era where data is increasingly seen as a critical asset, the adoption of PETs is essential for balancing the

To mitigate the risk of non-standardized data sharing, it is crucial to establish standards that are collaboratively developed and agreed upon by the PPP stakeholders.

[42] See The National Artificial Intelligence Research Resource (NAIRR) Pilot, U.S. National Science Foundation (2024); Frontiers in Artificial Intelligence for Science, Security and Technology (FASST), U.S. Department of Energy (2024).

[43] Technology, Innovation and Partnerships, U.S. National Science Foundation (2024).

[44] One pillar of the proposed FASST program is to make repositories of DOE data accessible to partners in the scientific community. See Frontiers in Artificial Intelligence for Science, Security and Technology (FASST), U.S. Department of Energy (2024).

[45] The AI Executive Order (EO) defines privacy-enhancing technologies as ("any software or hardware solution, technical process, technique, or other technological means of mitigating privacy risks arising from data processing, including by enhancing predictability, manageability, dissociability, storage, security, and confidentiality.") Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, White House (2023).

[46] Emerging Privacy-Enhancing Technologies, Organisation for Economic Co-operation and Development (2023).

[47] See National Data Action Plan, Special Competitive Studies Project at 19-21 (2022).

benefits of data-driven innovation with the necessity of protecting privacy, and federal efforts are ongoing to explore the utility of these technologies.[48]

4. Explore data IP rights protections.

Current IP laws have gaps that fail to protect valuable datasets. Inadequate protections could discourage the sharing of critical datasets within the scientific community.[49] For example, under current IP laws, data has limited protections under copyright laws. Thus, companies typically protect their data as a trade secret which prohibits them from making their data publicly available. Assessing the utility of IP rights for data can help determine if this is a viable way for the United States to protect its data assets while promoting a culture of data sharing and collaboration.

The National Security Commission on AI proposed that the Secretary of Commerce conduct a thorough study examining the need for intellectual property and IP-like protections to incentivize the creation, and sharing of new datasets,[50] which can be specifically tailored for datasets in scientific domains to foster new discoveries. Such a study has yet to be conducted.

China has already experimented with applying intellectual property protections to data. In 2022, the Office of the China National Intellectual Property Administration (CNIPA) announced it was beginning pilot projects[51] to apply intellectual property rights to data. In 2023, the pilots were reported to be making "positive progress" in eight provinces.[52] Earlier this year, China unveiled a strategy to strengthen its IP system, with an emphasis on "deeply promot[ing] the establishment of data intellectual property protection rules."[53]

[48] The AI EO specifically directs federal agencies to mitigate privacy risks through the use of PETs by issuing "guidelines to evaluate the efficacy of differential-privacy-guarantee protections, including for AI," and to lead ongoing research and development in this vital area by directing NSF to, "fund the creation of a Research Coordination Network (RCN) dedicated to advancing privacy research and, in particular, the development, deployment, and scaling of PETs," which has already met EO deadlines. See Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, White House (2023); NSF and DOE Establish a Research Coordination Network Dedicated to Enhancing Privacy Research, National Science Foundation (2024).

[49] See Mid-Decade Challenges to National Competitiveness, Special Competitive Studies Project at 74-75 (2022).

[50] The study should include a full market and policy impact analysis of IP protections for data and assess which policies and/or legislation should be proposed if protection is deemed necessary. See Final Report, National Security Commission on Artificial Intelligence at 471 (2021) ("While protections for data might be a future need, the U.S. should be proactive in assessing and addressing the necessity of such protections. The study can explore ways to protect and incentivize creation of datasets while allowing the data to be shared at some point, particularly with smaller entities that might not otherwise be able to enter the market. An analysis of the strengths and weaknesses of the European sui generis database protections should inform the study.").

[51] CNIPA Announces Pilot Areas for the Work on Data Intellectual Property, LexisNexis (2022).

[52] Positive Progress made in the Pilot Work of IP protection of Data, Lexology (2023).

[53] China Releases the Plan for Promoting the Construction of a Powerful Intellectual Property Country in 2024, China IP Law Update (2024).

The United States should proactively explore the utility of tailored IP rights for data.By providing IP or IP-type protections for data, the United States can not only safeguard its high-quality datasets, but also remain competitive on the global stage, ensuring that it continues to lead in science and technology.

5. Utilize a vital research tool - synthetic data.

Each PPP should consider incorporating synthetic data in their data sharing initiatives where appropriate. Synthetic data can increase the amount and quality of relevant data available to the scientific research community, albeit in limited circumstances.[54] AI models that have been validated and tested in specific domains and for specific purposes (i.e., they are trustworthy) can be prompted to output "simulated" experimental data. Synthetic data essentially extrapolates existing knowledge to new contexts. For example, a validated AI model that has been trained on real patient genomic data can be prompted to generate synthetic genomic data for hypothetical patients with certain characteristics. Similarly, data generated from digital twins (i.e., a virtual replica of a physical system, process, system) also can be used to accelerate the training of AI models. For example, data generated from a virtual replica of a machine (e.g., a microreactor) can be used to predict machine failures, enabling predictive maintenance and increased productivity.[55]

Synthetic data can act as a proxy for instances in which real world data is incomplete, biased, or difficult to obtain. Leveraging synthetic data for scientific research not only enlarges the corpus of relevant data available to scientists, but also has the potential to mitigate privacy concerns that are raised by using real patient genomic data. Of course, synthetic data will have limited use in circumstances when the research involves questions on whether current knowledge works in a new context, or when there is no pre-existing knowledge base from which to extrapolate.

Conclusion

The United States holds valuable datasets spanning government, industry, and academia, which should be leveraged as a competitive advantage in scientific research. To fully harness this potential, the United States must establish infrastructure for data that is currently siloed across the data generation ecosystem. This infrastructure should take the form of public-private

[54] Synthetic Data Generation: Definition, Types, Techniques, & Tools, Turing (2023).

[55] Idaho National Laboratory Demonstrates First Digital Twin of a Simulated Microreactor, U.S. Department Of Energy (2022); Kosmas Alexopoulos, et al., Digital Twin-Driven Supervised Machine Learning for the Development of Artificial Intelligence Applications in Manufacturing, International Journal of Computer Integrated Manufacturing (2020).

partnerships, facilitating collaboration and data sharing. The United States can lead in scientific discovery and innovation by implementing these data sharing frameworks, ensuring these valuable resources are maximized. The U.S. Government should ensure the infrastructure includes the following: clear data standards, access to rich U.S. Government data, comprehensive data inventories, use of PETs, exploration of data IP rights, and utilization of synthetic data. By doing so, the United States can continue to be at the forefront of global scientific progress, transforming these datasets into powerful tools for groundbreaking advancements.