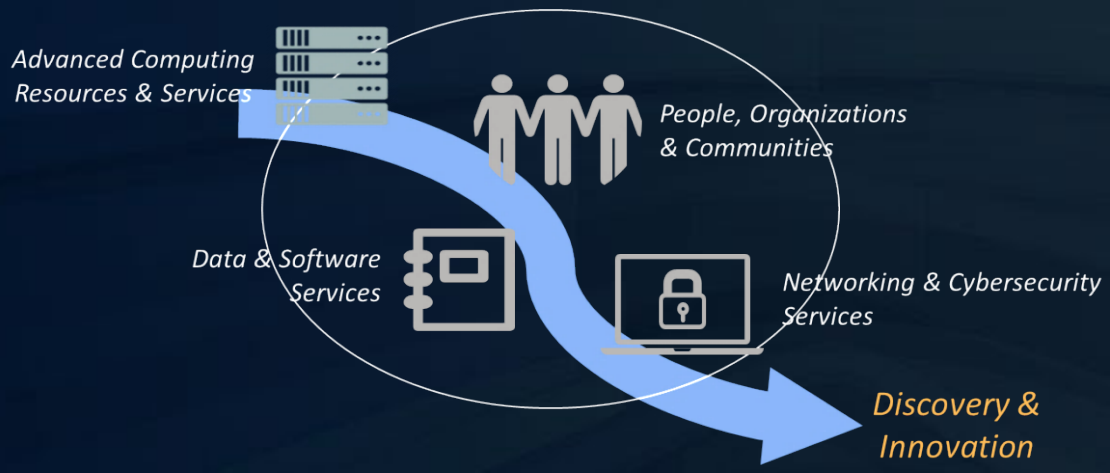
Transforming Science Through Cyberinfrastructure

NSF's Blueprint for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century



Blueprint for a National Data and Software Cyberinfrastructure

Office of Advanced Cyberinfrastructure
Directorate for Computer & Information Science & Engineering
National Science Foundation

December 2020

Transforming Science Through Cyberinfrastructure: National Data and Software Cyberinfrastructure

NSF's Blueprint for a National Data and Software Cyberinfrastructure for Accelerating Science and Engineering in the 21st Century

Executive Summary

The national research cyberinfrastructure (CI) has become critical to computational and data-intensive research across all of science and engineering (S&E) in the 21st century. The National Science Foundation (NSF) recently shared a vision, developed by its Office of Advanced Cyberinfrastructure (OAC), for ***an agile, integrated, robust, trustworthy, and sustainable CI ecosystem that drives new thinking and transformative discoveries in all areas of S&E research and education***. The envisioned CI ecosystem integrates advanced CI resources, services, and expertise towards collectively enabling new, transformative discoveries across all of S&E.

This document is the fourth in a series of blueprint documents that outline NSF's plan for realizing this vision. It presents a forward-looking blueprint for a national data and software cyberinfrastructure (Data and Software CI) ecosystem to enable and accelerate S&E research and outlines a plan for translating this blueprint into action. Aligned with NSF's mission, this plan is built on community input and driven by the current and future S&E needs of society. NSF, through OAC, is aggressively moving ahead to put this blueprint into action. This action plan includes enhancing NSF's current programs in this landscape as well as developing new, complementary activities.

Table of Contents

1 Introduction

The national research cyberinfrastructure (CI) has become critical to computational and data-intensive research across all of science and engineering (S&E) in the 21st century. It is a key catalyst for discovery and innovation and plays a critical role in ensuring US leadership in S&E, economic competitiveness and national security, consistent with the National Science Foundation's (NSF's) mission. NSF, through the Office of Advanced Cyberinfrastructure (OAC), has shared a vision[1] that calls for the broad availability and innovative use of ***an agile, integrated, robust, trustworthy and sustainable CI ecosystem*** that can ***drive new thinking and transformative discoveries in all areas of S&E research and education***.

NSF's vision for an agile, integrated, robust, trustworthy and sustainable CI ecosystem that drives new thinking and transformative discoveries in all areas of S&E research and education

- View CI more holistically: CI continuum seamlessly integrating a spectrum of resources, tools, services, and expertise to enable transformative discoveries.
- Support translational research: Core innovations → development of community tools and frameworks → deployment and operation of sustainable production CI.
- Balance innovation with stability: Ensure continuity in production computational capacity while fostering innovation and transition to production.
- Couple discovery and CI innovation cycles: Rapidly address new challenges and opportunities in an era of disruptive technologies and evolving science needs.
- Improve usability: Ease pathways for discovering, accessing, understanding and using powerful CI capabilities and services to enhance researcher productivity and scientific impact.

This document is the fourth in a series of blueprint documents that outline NSF's plan for realizing this vision. It presents a forward-looking blueprint for a robust, secure, trusted, performant, scalable, and sustainable data and software cyberinfrastructure (Data and Software CI) ecosystem to enable and accelerate S&E research and outlines a plan for translating this blueprint into action. This blueprint is informed by community input via advisory bodies, requests for information (RFIs, including the recent *RFIs on "Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research[2]"* and *"Future Needs for Advanced Cyberinfrastructure to Support Science and Engineering Research[3]"*), surveys (including the recent *NSF CSSI Community Survey*[4]), and workshops and conferences, as well as by national initiatives (including those listed in the vision document). NSF, through OAC, is aggressively moving ahead to put the blueprint outlined in this document into action.

[1] "Transforming Science Through Cyberinfrastructure: NSF's Blueprint for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century," https://www.nsf.gov/cise/oac/vision/blueprint-2019/.
[2] "Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research," https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp.
[3] "Dear Colleague Letter: Request for Information on Future Needs for Advanced Cyberinfrastructure to Support Science and Engineering Research (NSF CI 2030)," https://www.nsf.gov/pubs/2017/nsf17031/nsf17031.jsp.
[4] "CSSI Community Survey," https://cssi-pi-community.github.io/2020-meeting/slides/FredHansen_slides.pptx.

2 NSF's Current Data and Software CI Landscape

NSF's Data and Software CI programs have been long-term investments focused on catalyzing new thinking, paradigms, and practices in developing and using Data and Software CI tools, capabilities, and services to understand natural, human, and engineered systems. S&E challenges and use cases drive CI development, and successful CI systems strike a balance that reflects both the underlying technology and disciplinary research needs. NSF's Data and Software CI programs have continued to evolve in response to increasing complexity in the application requirements, rapid changes and diversity in the underlying hardware and networking components, the accelerated use of new data representations and processing paradigms and the convergence of data, software and services into unified instruments essential to all S&E domains.

NSF's current Data and Software CI portfolio includes the following programs:

- Cyberinfrastructure for Sustained Scientific Innovation (CSSI)
- Harnessing the Data Revolution (HDR)
- Computational and Data Enabled Science and Engineering (CDS&E)
- Cyberinfrastructure for Emerging Science and Engineering Research (CESER)
- Big Data Regional Innovation Hubs (BD Hubs)
- Discipline-Specific Programs on Data and Software CI Development
- Other Complementary CI-Focused Programs at NSF

We describe each of these programs in detail in the following subsections.

2.1 Cyberinfrastructure for Sustained Scientific Innovation (CSSI)

The Cyberinfrastructure for Sustained Scientific Innovation (CSSI)[5] umbrella program aims to create a Data and Software CI ecosystem that scales from individuals or small groups of researchers/innovators to large communities. Recognizing the need to rapidly respond to evolving research community priorities, NSF envisions support for the creation of such an ecosystem to be complemented by investments in foundational CI community services.

The CSSI program builds on the long-running Data Infrastructure Building Blocks (DIBBs)[6] and Software Infrastructure for Sustained Innovation (SI^2)[7] programs. The DIBBs program encouraged the development of robust and shared Data CI capabilities to accelerate interdisciplinary and collaborative research in areas of inquiry stimulated by data. The SI^2 program recognized that software permeates all aspects and layers of CI (from application codes and frameworks, programming systems, libraries, and system software, to middleware, operating systems, networking, and the low-level drivers), and aimed to catalyze new software thinking, paradigms, and practices in S&E.

The newer CSSI umbrella program targets services that address all aspects of CI, from embedded sensor systems and instruments, to desktops and high-end data and computing systems, to major instruments and facilities. *The program nurtures the interdisciplinary processes required to*

[5] "Cyberinfrastructure for Sustained Scientific Innovation (CSSI): Elements and Framework Implementations," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505505.
[6] "Data Infrastructure Building Blocks (DIBBs)," https://www.nsf.gov/pubs/2017/nsf17500/nsf17500.htm.
[7] "Software Infrastructure for Sustained Innovation (SI^2)," https://www.nsf.gov/pubs/2016/nsf16532/nsf16532.htm.

support the entire data and software lifecycle *and the integration of development, deployment, and support of CI services with innovation and research.* Furthermore, the program aims to catalyze the development of sustainable CI communities that transcend scientific and geographical boundaries. The program envisions vibrant partnerships among academia, government laboratories and industry, including international entities, for the development and stewardship of ***sustainable CI services that can enhance productivity*** and accelerate innovation in S&E. Integrated education activities play a key role in developing and sustaining the Data and Software CI over time and in creating a workforce capable of fully realizing its potential to transform S&E.

2.2 Harnessing the Data Revolution (HDR)

In 2016, NSF unveiled a set of "Big Ideas[8]," ten bold, long-term research and process ideas that identify areas for future investment at the frontiers of S&E. The Big Ideas represent unique opportunities to position our Nation at the cutting edge of global S&E leadership by bringing together diverse disciplinary perspectives to support convergence research. NSF is implementing the Big Ideas vision through focused NSF-wide programs and solicitations that involve all NSF directorates and offices. CI challenges and opportunities cut across, and OAC is working with the other directorates and offices to provide CI leadership.

Among the NSF Big Ideas, Harnessing the Data Revolution (HDR)[9] explicitly addresses Data CI as one of its goals. HDR is a national-scale activity to enable new modes of ***data-driven discovery*** that will allow fundamental questions to be asked and answered at the frontiers of S&E. The HDR vision is realized through a set of interrelated efforts seeking to establish theoretical, technical, and ethical frameworks that will be applied to tackle data-intensive problems in S&E, contributing to data-driven decision-making that impacts society. The HDR Institutes program seeks to create an integrated fabric of interrelated institutes that can accelerate discovery and innovation in multiple data-intensive S&E areas by ***harnessing diverse data sources*** and developing and applying new methodologies, technologies, and CI for data management and analysis. The HDR Institutes will support convergence between S&E research communities as well as expertise in data science foundations, systems, applications, and CI.

2.3 Computational and Data-Enabled Science and Engineering (CDS&E)

The goal of the Computational and Data-Enabled Science and Engineering (CDS&E)[10] program is to identify and capitalize on opportunities for major scientific and engineering breakthroughs through new computational and data analysis approaches. The intellectual drivers may be in an individual discipline, or they may cut across more than one discipline in various NSF directorates. The key identifying factor is that the outcome relies on the development, adaptation, and utilization of one or more of the capabilities offered by the advancement of both research and infrastructure in computation and data, either through cross-cutting or disciplinary programs.

[8] "NSF10 Big Ideas," https://www.nsf.gov/news/special_reports/big_ideas/.
[9] "Harnessing the Data Revolution," https://www.nsf.gov/cise/harnessingdata/.
[10] "Computational and Data-Enabled Science and Engineering (CDS&E)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504813.

The CDS&E program in OAC specifically addresses ***research in CI*** *itself with the clear potential to impact multiple research disciplines through the development of the paradigms, algorithms and processes needed to provide general CDS&E solutions as part of comprehensive, integrated, sustainable and secure Data and Software CI.* The CDS&E program encourages adventurous ideas that generate new paradigms and that create and apply novel techniques, generating and utilizing digital data in innovative ways to complement or dramatically enhance traditional computational, experimental, observational, and theoretical tools for scientific discovery and application.

2.4 Cyberinfrastructure for Emerging Science and Engineering Research (CESER)

The Cyberinfrastructure for Emerging Science and Engineering Research (CESER)[11] program aims to catalyze new S&E discovery pathways through early-stage collaborative activities between disciplinary scientists and engineers as well as developers/implementers of innovative CI capabilities, services, and approaches.

Through a recent Dear Colleague Letter (DCL) on *Pilot Projects to* ***Integrate*** *Existing Data and Data-Focused Cyberinfrastructure to Enable Community-level Discovery Pathways*[12], the CESER program encourages pilot projects that bring together researchers and CI experts to develop the means of combining existing community data resources and shared Data CI into new integrative and highly performing ***data-intensive discovery workflows*** that empower new scientific pathways. Aims of such pilot projects can include, but are not limited to: (1) improving the end-to-end process of accessing, integrating and transforming research and education data to knowledge and discovery for one or more communities; (2) creating new workflows and new usage modes to address multi-disciplinary and cross-domain scientific objectives; (3) addressing emerging community-scale scientific data challenges such as real-time, streaming and on-demand data access, data discovery through knowledge networks and intelligent data delivery, access to data with privacy concerns, and data fusion, integration and interoperability; (4) enhancing the performance and robustness of community-scale data integration and discovery workflows such as through automated curation, end-to-end performance monitoring, provenance tracking, and means of assuring data trustworthiness; and (5) federating learner data to empower innovative assessment tools for large-scale modeling of learning gains.

2.5 Big Data Regional Innovation Hubs (BD Hubs)

NSF's data portfolio also includes the Big Data Regional Innovation Hubs (BD Hubs)[13], launched in 2015, that aim to ***nucleate regional collaborations*** and multi-sector projects, and foster innovation in data science. Four BD Hubs were funded as part of this program within each of the Census Regions of the country—Midwest[14], Northeast[15], South[16], and West[17]. The BD Hubs serve

[11] "Cyberinfrastructure for Emerging Science and Engineering Research (CESER)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505385.
[12] "Dear Colleague Letter (DCL) on Pilot Projects to Integrate Existing Data and Data-Focused Cyberinfrastructure to Enable Community-level Discovery Pathways," https://www.nsf.gov/pubs/2020/nsf20085/nsf20085.jsp.
[13] "Big Data Regional Innovation Hubs," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505185
[14] "Midwest Big Data Innovation Hub," https://midwestbigdatahub.org.
[15] "Northeast Big Data Innovation Hub," https://nebigdatahub.org.
[16] "South Big Data Innovation Hub," https://southbigdatahub.org.
[17] "West Big data Innovation Hub," https://westbigdatahub.org.

as a venue for building and fostering local and regional data-related activity in city, county, and state governments, in local industry and non-profits, and in regional academic institutions. Collaborative activities and partnerships emerging from a regional focus contribute to building and sustaining a successful national big data innovation ecosystem. The last *Big Data Regional Innovation Hubs (BD Hubs) - Accelerating the Big Data Innovation Ecosystem* solicitation (NSF 18-598)[18] solicitation aimed to continue the operation of a national network of BD Hubs. It built upon demonstrated strengths of the program, which is responsive to the recent developments in data-intensive S&E domains.

2.6 NSF's Discipline-Specific Data and Software CI Development Programs

Several NSF directorates administer the development of discipline-specific Data and Software CI to support research and innovation in their own fields. Some of these efforts are listed below.

EarthCube[19] is a community-driven activity sponsored through a partnership between NSF's Directorate for Geosciences (GEO) and OAC to transform research in the academic geosciences community. EarthCube aims to create a well-connected and facile environment to share data and knowledge in an open, transparent, and inclusive manner, thus accelerating our ability to understand and predict the Earth system. EarthCube supports projects which build capabilities to improve geosciences data use and reuse for observational, experimental, and computational research that is interoperable with emerging standards and resources. EarthCube also supports data facilities and data resources to adopt robust standards and/or implementation of pilot tools/activities to improve discovery, interoperability, and access to data and CI services. In conjunction with EarthCube's Council of Data Facilities developments, these projects are facilitating the adoption of new semantic web standards and machine-readable publishing patterns, such as for the EarthCube data repository and resource registries.

The **Infrastructure Capacity for Biological Research: Cyberinfrastructure**[20] program supports the implementation or major improvement of CI that specifically advances or transforms contemporary biology, and that is broadly applicable to a wide range of researchers. Proposed projects may include CI related to any level of biological phenomena (e.g., molecular, cellular organismal, ecosystem, biome), but there needs to be a clear articulation of how the proposed capacity will lead to scientific understanding in biology. The scope of the proposed infrastructure can include but is not limited to: design and construction of databases from new or existing data sources; software and methods for making use of new technologies for the acquisition, communication or visualization of biological data; software or ontologies related to data discovery, data mining, data integration or data visualization; tools that facilitate biological research workflows, analytical pathways, or integration between the field and laboratory or between observations, experiments, and models; or scientific gateways. It is expected that

[18] "Big Data Regional Innovation Hubs (BD Hubs) – Accelerating the Big Data Innovation Ecosystem," https://www.nsf.gov/pubs/2018/nsf18598/nsf18598.htm.
[19] "EarthCube: Developing a Community-Driven Data and Knowledge Environment for the Geosciences," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504780.
[20] "Infrastructure Capacity for Biological Research: Cyberinfrastructure," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505538.

projects will produce finished products that are publicly accessible and useable for the biological research community.

The **Designing Materials to Revolutionize and Engineer our Future (DMREF)**[21] program is the primary program by which NSF participates in the Materials Genome Initiative (MGI) for Global Competitiveness[22]. MGI recognizes the importance of materials S&E to the well-being and advancement of society and aims to "deploy advanced materials at least twice as fast as possible today, at a fraction of the cost." MGI integrates materials discovery, development, property optimization, and systems design with a shared computational framework. DMREF aims to drive the development of new tools, new infrastructure, and the integration of computation, data analytics, AI, experiment and theory. These include: new data analytics tools and statistical algorithms; advanced simulations of material properties in conjunction with new device functionality; advances in predictive modeling that leverage machine learning (ML), AI, data mining, and sparse approximation; data infrastructure that is accessible, extensible, scalable, and sustainable; the development, maintenance, and deployment of reliable, interoperable, and reusable software for the next-generation design of materials; and new collaborative capabilities for managing large, complex, heterogeneous, distributed data supporting materials design, synthesis, and longitudinal study.

Some of the other discipline-specific Data and Software CI development programs at NSF include CISE Community Research Infrastructure (CCRI)[23], Human Networks and Data Science (HNDS)[24], Sustaining Infrastructure for Biological Research (Sustaining)[25], and Earth Sciences: Instrumentation and Facilities (EAR/IF)[26].

2.7 Other Complementary CI-Focused Programs at NSF

NSF offers a set of other relevant programs which complement the above-mentioned Data and Software CI programs in terms of CI research, trustworthiness, instrumentation, and networking support. Some of these programs are listed below.

The **OAC Core Research**[27] program aims to address the CI research challenges that significantly impact the future capabilities of advanced research CI by engaging a diverse community of computer and computational S&E researcher and students. The context is emerging translational research challenges in the design, development, deployment, experimentation, and application of advanced research CI. The OAC Core Research investments are expected to be multi-disciplinary, extreme-scale, driven by S&E research, end-to-end solutions, and deployable as robust research CI.

[21] "Designing Materials to Revolutionize and Engineer our Future (DMREF),"
https://www.nsf.gov/pubs/2021/nsf21522/nsf21522.htm.
[22] "Materials Genome Initiative (MGI) for Global Competitiveness,"
https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf.
[23] "CISE Community Research Infrastructure (CCRI)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12810.
[24] "Human Networks and Data Science (HNDS)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505702.
[25] "Sustaining Infrastructure for Biological Research (Sustaining)," https://www.nsf.gov/pubs/2021/nsf21503/nsf21503.htm.
[26] "Earth Sciences: Instrumentation and Facilities (EAR/IF)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=6186.
[27] "OAC Core Research," https://nsf.gov/funding/pgm_summ.jsp?pims_id=505670.

The **Principles and Practice of Scalable Systems (PPoSS)**[28] program aims to support a community of researchers who will work symbiotically across multiple disciplines to perform basic research on scalability of modern applications, systems, and toolchains. The intent is that these efforts will foster the development of principles that lead to rigorous and reproducible artifacts for the design and implementation of large-scale systems and applications across the full hardware/software stack. These principles and methodologies should simultaneously provide guarantees on correctness and accuracy, robustness, and security and privacy of systems, applications, and toolchains. Importantly, PPoSS specifically seeks to fund projects that span the entire hardware/software stack and will lay the groundwork for sustainable approaches for engineering highly performant, scalable, and robust computing applications.

The **Cybersecurity Innovation for Cyberinfrastructure (CICI)**[29] program aims to develop, deploy and integrate solutions that benefit the broader scientific community by securing science data, workflows, and infrastructure. CICI recognizes the unique nature of modern, rapid collaborative science and the breadth of security expertise, infrastructure and requirements among different practitioners, researchers, and scientific projects. CICI seeks projects in three program areas: (1) usable and collaborative security for science, (2) reference scientific security datasets, and (3) scientific infrastructure vulnerability discovery.

The **Campus Cyberinfrastructure (CC*)**[30] program invests in coordinated campus-level CI (including networking, computing, and storage) improvements, innovation, integration, and engineering for science applications and distributed research projects. Learning and workforce development (LWD) in CI is explicitly addressed in the program. A common theme across all aspects of the CC* program is the critical importance of the partnership among campus-level CI experts, including the campus information technology (IT)/networking/data organization, contributing domain scientists, other research groups, and educators necessary to engage in and drive new CI capabilities and approaches in support of scientific discovery.

The **Major Research Instrumentation (MRI)**[31] program serves to increase access to multi-user scientific and engineering instrumentation for research and research training in our Nation's institutions of higher education and not-for-profit scientific/engineering research organizations. A MRI award supports the acquisition or development of a multi-user research instrument that is, in general, not appropriate for support through other NSF programs. MRI provides support to acquire critical research instrumentation without which advances in fundamental S&E research may not otherwise occur. MRI also provides support to develop next-generation research instruments that open new opportunities to advance the frontiers in S&E research.

The **Mid-scale Research Infrastructure** (**Mid-scale RI-1**[32] and **Mid-scale RI-2**[33]) programs intend to provide NSF with an agile, Foundation-wide process to fund research infrastructure capabilities in the mid-scale range between the MRI and the higher end Major Research

[28] "Principles and Practice of Scalable Systems (PPoSS)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505751.
[29] "Cybersecurity Innovation for Cyberinfrastructure (CICI)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505159.
[30] "Campus Cyberinfrastructure (CC*)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748.
[31] "Major Research Instrumentation (MRI)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5260.
[32] "Mid-scale Research Infrastructure-1 (Mid-scale RI-1)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505602.
[33] "Mid-scale Research Infrastructure-2 (Mid-scale RI-2)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505550.

Equipment and Facilities Construction (MREFC[34]) project thresholds. Mid-scale Research Infrastructure projects directly enable advances in fundamental science, technology, engineering, and mathematics (STEM) in one or more of the research and education domains supported by NSF. Mid-scale RI projects emphasize strong scientific merit with an identified research infrastructure need of the research community at a national needs level.

For other relevant programs in the NSF CI landscape including leadership-class computing systems (e.g., *LCCF*[35]), innovative advanced computing systems and services (e.g., *ACCS*[36]), cybersecurity (e.g., *SaTC*[37]), and learning and workforce development (e.g., *CyberTraining*[38]), please refer to the first blueprint, *"Transforming Science Through Cyberinfrastructure: NSF's Blueprint for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century."*[1]

3 A Blueprint for National Data and Software CI Ecosystem for Enabling and Accelerating Science and Engineering in the 21st Century

This section presents NSF's forward-looking blueprint for a national Data and Software CI ecosystem and outlines a plan for translating this blueprint into action. Aligned with NSF's mission, this plan is built on community input and driven by the current and future S&E needs of society.

3.1 Building on Community Inputs

This blueprint is informed by community input including through advisory bodies, requests for information (RFIs), and workshops and conferences, as well as by national initiatives. Below we describe a few key examples.

1. *Responses to the NSF Request for Information (RFI) on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research (NSF 20015).*[2]

This RFI invited the community to provide input to NSF on specific data-intensive S&E research questions and challenges and the essential data-related CI services and capabilities needed to publish, discover, transport, manage and process data in secure, performant and scalable ways to enable data-intensive research. NSF received more than 100 responses to this RFI, comprising contributions from over 340 authors from over 150 research institutions and other organizations across a wide range of S&E research domains supported by NSF. Ninety-eight of the responding primary authors consented to allow NSF publication of their response per a Creative Commons license, and these responses are available online[39] through NSF OAC webpage. The RFI responses

[34] "NSF Large Facilities Manual," (NSF 17-066), https://www.nsf.gov/pubs/2017/nsf17066/nsf17066.pdf.
[35] "Towards a Leadership-Class Computing Facility - Phase 1," https://www.nsf.gov/pubs/2017/nsf17558/nsf17558.htm.
[36] "Advanced Computing Systems & Services: Adapting to the Rapid Evolution of Science and Engineering Research," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503148.
[37] "Secure and Trustworthy Cyberspace (SaTC)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504709.
[38] "Training-based Workforce Development for Advanced Cyberinfrastructure (CyberTraining)," https://www.nsf.gov/pubs/2019/nsf19524/nsf19524.htm.
[39] "Responses to NSF 20-015," https://www.nsf.gov/cise/oac/datacirfi/rfi_responses.jsp.

collectively serve as important input towards NSF's further development and implementation of its Data CI investment strategy and programs going forward.

2. Analysis of Community Response to NSF CSSI Program and Solicitation, 2020.[4]

NSF funded a survey[40] of 251 Principal Investigators (PIs), co-PIs, and others in the CI research community. The survey was designed to gather input that would inform decisions about the CSSI solicitation and decisions about the program's direction and focus. The respondents were primarily PIs or co-PIs. More than 50% had served as reviewers in the past year. Nearly all of the respondents had previously submitted proposals to NSF, 73% of whom received funding. More than 90% listed universities as their primary organization. Survey responses were received from 42 U.S. states.

3. Responses to the NSF Request for Information (RFI) on Future Needs for Advanced Cyberinfrastructure to Support Science and Engineering Research (NSF CI 2030).[3]

This RFI, initiated on behalf of the NSF Advisory Committee for Cyberinfrastructure (ACCI), explored the advanced CI needs of the S&E community over the next decade. NSF received more than 130 responses to this Request for Information (RFI), comprising contributions from over 360 authors across the full spectrum of S&E research domains. RFI response authors agreed to allow NSF to make their responses public per a Creative Commons license, and these responses are available online[41] through NSF OAC webpage. NSF is using these contributions to assist in formulating plans for supporting the NSF community in the exploration, development, and deployment of an advanced CI for the next decade.

4. Other community input based on workshop reports, advisory committee reports, and strategic planning documents on future cyberinfrastructure, data, and computing initiatives, including but not limited to the following:

- The 2020 NSF CSSI Principal Investigator (PI) Workshop Report[42];
- The 2019 Community Visioning Workshop on the Future Direction of the CSSI Program[43];
- The 2018 NSF SI2 Principal Investigator (PI) Workshop Report[44];
- The 2018 NSF Advisory Committee for Cyberinfrastructure (ACCI) Report on CI2030: Future Advanced Cyberinfrastructure[45];
- The 2017 NSF Data Infrastructure Building Blocks (DIBBs) PI Workshop Report[46];

[40] "Assessment and Evaluations of NSF OAC-Funded Program Impact on the Scientific Community," https://www.nsf.gov/awardsearch/showAward?AWD_ID=1930025.
[41] "Responses to the NSF CI 2030 RFI," https://www.nsf.gov/cise/oac/ci2030/rfi_responses.jsp.
[42] "CSSI PI Meeting Report", https://cssi-pi-community.github.io/2020-meeting/CSSI-2020-FinalReport-Public.pdf.
[43] "Future Directions of the Cyberinfrastructure for Sustained Scientific Innovation (CSSI) Program Workshop Report", https://drive.google.com/file/d/104GRCzv5fX0gUvdPNaWgxL6qFI3VCQIn.
[44] "The 2018 NSF Software Infrastructure for Sustained Innovation (SI2) Principal Investigator (PI) Workshop Report," https://si2-pi-community.github.io/2018-meeting/.
[45] "CI2030: Future Advanced Cyberinfrastructure - A report of the NSF Advisory Committee for Cyberinfrastructure," https://www.nsf.gov/cise/oac/ci2030/ACCI_CI2030Report_Approved_Pub.pdf.
[46] "The 2017 NSF Data Infrastructure Building Blocks (DIBBs) PI Workshop Report," https://dibbs17.org/report/DIBBs17FinalReport.pdf.

- NSF Cyberinfrastructure Framework for Twenty-First Century Science and Engineering (CIF21) Initiative and Vision Document[47];
- Realizing the Potential of Data Science - Final Report from the NSF CISE Advisory Committee Data Science Working Group[48] ;
- National Strategic Computing Initiative Update: Pioneering the Future of Computing - A Report by the Fast-Track Action Committee on Strategic Computing[49]; and
- The Federal Big Data Research and Development Strategic plan - A Report by the Networking and Information Technology Research and Development (NITRD)[50].

3.2 The Road Ahead – Planning for the Future

Digital data and its analysis, as well as the underlying software applications, systems and services, continue to play an increasingly central role in all areas of S&E research. Many research communities supported by NSF are still challenged by the need to access, manage, integrate, and use more massive and diverse scientific datasets than ever before to conduct research. For this reason, fostering S&E-driven, robust, secure, trusted, performant, scalable, and sustainable Data and Software CI is a key component of NSF's vision[1] for a national CI ecosystem. Such CI must be designed to flexibly accommodate both existing and future disciplinary and multi-disciplinary application challenges and to provide essential capabilities and services that enable new and evolving integrative and cross-disciplinary S&E efforts to translate data from diverse and distributed data sources to knowledge and discovery. NSF accordingly continues to invest in the creation of a wide range of Data and Software CI tools, services, resources, and solutions for use by the various disciplinary communities that it supports in order to enable this transformation.

Scientific endeavors are getting increasingly collaborative, cross-disciplinary, and convergent. NSF thus recognizes the importance of promoting holistic CI approaches to address the growing and evolving end-to-end data lifecycle and analytics workflow challenges both within and across research fields. This holistic view is predicated on harmonization, integration and interoperability among Data and Software CI resources, tools, services and expertise to achieve accessible, seamless and flexible end-to-end discovery pathways that drive new thinking and enable transformative discoveries.

[47] ''NSF Cyberinfrastructure Framework for Twenty-First Century Science and Engineering (CIF21)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504730.
[48] "Realizing the Potential of Data Science - Final Report from the NSF CISE Advisory Committee Data Science Working Group," https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf.
[49] "National Strategic Computing Initiative Update: Pioneering the Future of Computing," https://www.nitrd.gov/pubs/National-Strategic-Computing-Initiative-Update-2019.pdf.
[50] "The Federal Big Data Research and Development Strategic plan - A Report by the Networking and Information Technology Research and Development (NITRD)," https://www.nitrd.gov/pubs/bigdatardstrategicplan.pdf.

NSF's overarching strategies in building the national Data and Software CI ecosystem:

1. Support domain-specific and customized Data and Software CI
2. Prioritize and invest in transdisciplinary community Data and Software CI
3. Close the gap between research, development and sustained production of CI
4. Complement Software and Data CI with other NSF CI efforts
5. Promote Data and Software CI community building
6. Invest in Data and Software CI learning and workforce development

NSF will continue to encourage the development of innovative, robust and trustworthy Data and Software CI capabilities to enable and accelerate interdisciplinary and collaborative research in areas of inquiry stimulated by dynamic and diverse data sources. NSF investments in Data and Software CI ecosystem will balance innovation with stability to enable new services, capabilities, and resources to advance scientific discoveries and collaborations, while continuing to address emerging requirements, novel technologies and concerns such as reproducibility, privacy and trust. These investments are expected to build upon, integrate with, and contribute to existing CI, serving as evaluative resources while developments in national-scale access, policy, interoperability and sustainability continue to evolve.

NSF's ***overarching strategies*** in building the national Data and Software CI ecosystem include the following:

Support Domain-specific and Customized Data and Software CI: Some S&E domains have very specific data and software requirements unique to their disciplines and they need to be well supported with Data and Software CI resources tailored and optimized for those applications. NSF will continue to support domain-specific and customized Data and Software CI solutions for such domains. NSF's existing programs such as *CSSI*[5], *HDR*[9], *CDS&E*[10], and *CESER*[11], together with discipline-specific programs listed in Section 2.6 are well-suited for such efforts.

Prioritize and Invest in Transdisciplinary Community Data and Software CI: NSF will prioritize and invest in the key elements of a broadly accessible, interoperable, and reusable transdisciplinary community Data and Software CI. The lack of easily reusable capabilities can sometimes result in an ecosystem of duplicated functionality. For this reason, ***not only the data, but also the CI tools and services should be FAIR***[51] *– findable, accessible, interoperable, and reusable.* NSF is invested in understanding how broader cross-disciplinary and domain-agnostic solutions can be devised and implemented, along with the structural, functional and performance characteristics such cross-disciplinary solutions must possess. Such new CI services and capabilities should allow for

[51] NSF acknowledges that practices concerning making the data FAIR may vary and have some restrictions determined by applicable laws, university and research institution policies, funder terms, privacy, intellectual property and licensing agreements, and the ethical context of research. NSF's policy on public access was originally presented in NSF 15-052: *NSF Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation, March 15, 2015* (https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf). The policy has been further refined in two Dear Colleague Letters: NSF 19-026: *Dear Colleague Letter: Effective Practices for Data, May 20, 2019* (https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.pdf) and NSF 20-068: *Dear Colleague Letter: Open Science for Research Data, March 27, 2020* (https://nsf.gov/pubs/2020/nsf20068/nsf20068.jsp).

seamless integration and interoperability with existing CI; support a wide variety of S&E drivers, users and usage modes; and foster the initiation of future modes of discovery. *NSF's plans to realize this goal are detailed in* ***Section 3.3.***

Close the Gap Between Research, Development and Sustained Production of CI: NSF is developing a strategy that will close the gap between research, development and sustained production of Data and Software CI. The envisioned *Pathways to Production* will balance innovations with stability and continuity in production-quality Data and Software CI while ensuring that there are opportunities to explore innovations that address emerging requirements, novel technologies and concerns such as reproducibility, privacy and trust, and to transition these innovations to production when appropriate. It is essential to have a clear plan for scaling Data and Software CI research prototypes and early implementations developed through other NSF programs mentioned above and ***transition them to production*** in order to increase productivity and ensure sustained scientific innovation across S&E domains. *NSF's plans to realize this goal are detailed in* ***Section 3.4****.*

Complement Data and Software CI with Other NSF CI Efforts: Data and Software CI needs to be complemented with shared-use computing, networking and data CI, sophisticated research instruments and platforms, robust and trustworthy services and data products that are openly, reliably and pervasively accessible by a broad community of researchers and/or educators. NSF recognizes that such research instrumentation is critical for advances in fundamental S&E, and it will continue investing in the creation of this infrastructure through its existing programs such as *CICI*[29]*, CC**[30]*, MRI*[31]*, Mid-scale RI-1*[32] *and RI-2*[33]*, LCCF*[35]*,* and *ACCS*[36]. The size of these investments can change from small sensor systems for data acquisition to large repositories for storage and high-end computing facilities for the analysis of the massive data sets.

Promote Data and Software CI Community Building: NSF will continue to invest in developing a broad and diverse Data and Software CI community, promoting coordination and exchange between the CI and research communities and facilitating the dissemination of best practices for design, development, and operation of CI resources and capabilities. NSF's recent *BD Hubs*[13] program is a major step towards achieving this goal by serving as a venue for building and fostering local and regional data-related activities in city, county, and state governments, in local industry and non-profits, and in regional academic institutions. NSF's new investments in *CI Centers of Excellence (CoE)*[52] takes this a step further and aims to facilitate community building and sharing by supporting hubs of expertise and innovation targeting specific areas, aspects, or stakeholder communities of the research CI ecosystem. Supported CI CoEs provide expertise and services related to CI technologies and solutions; gather, develop, and communicate community best practices; and serve as readily-available resources for both the research community and the CI community.

Invest in Data and Software CI Learning and Workforce Development: All Data and Software CI programs at NSF will continue to include an integrated component focused on the training and professional development of a skilled workforce with expertise ranging from CI research and development to CI deployment and its application to different domains. This is critical in

[52] "Cyberinfrastructure Centers of Excellence (CoE)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505744.

preparing, nurturing, and growing the national scientific research workforce for creating, utilizing, and supporting advanced CI to enable and potentially transform fundamental S&E research and contribute to the Nation's overall economic competitiveness and security. In addition to this integrated learning and workforce development (LWD) component across Data and Software CI programs, NSF's dedicated LWD program in CI, *CyberTraining*[38], will continue to invest in innovative and scalable training, education, and professional development activities which will lead to transformative changes in the state of research workforce preparedness for advanced CI-enabled research in the short and long terms.

3.3 Key Elements of the Envisioned National Data and Software CI Ecosystem

As mentioned in the previous section, NSF acknowledges that some S&E domains have very specific data and software requirements unique to their disciplines and they need to be well supported with Data and Software CI solutions tailored and optimized for those applications. NSF will continue to support domain-specific and customized Data and Software CI tools and services for such domains. On the other hand, NSF will also prioritize and invest in the key elements of a broadly accessible, interoperable, and reusable strategic community Data and Software CI. For this purpose, NSF identified the key elements of the envisioned national Data and Software CI ecosystem, which will promote easy discovery of relevant data, metadata, and CI components; interoperability between different data types, distributed data sources, and software developed and managed by different organizations; deliver easy and fast access to data, computation, and other CI resources; protection and reliability of data and software; and high-performance processing, analysis, and interpretation of the data through shared tools, technologies and services.

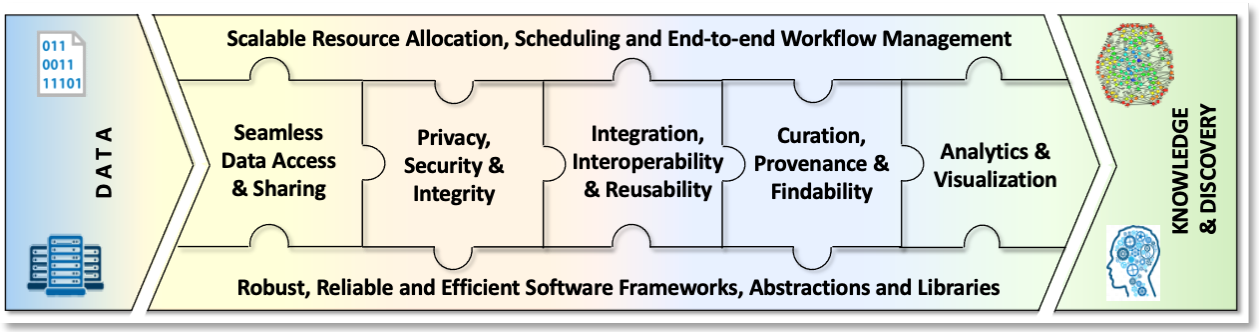


Figure 1: Key elements of the envisioned national Data and Software CI ecosystem to transform data into knowledge and discovery.

The ***key elements*** of the envisioned national Data and Software CI ecosystem (as outlined in Figure 1) include but are not limited to the following:

Seamless Data Access and Sharing. Digital data is being generated at increasing rates and densities from a growing array of sources, including major surveys, mobile and embedded systems, sensors, connected Internet-of-Things (IoT) devices, observatories, scientific instruments, publications, experiments, simulations, and analyses. Smart data placement and seamless data access, transfer, streaming and sharing services that would support computing continuum from the micro level (e.g., edge devices, sensors, IoT) to the macro level (e.g., data

centers, clouds, supercomputers) and enable access to data anywhere, anytime, from any device is a pressing necessity. This could be done either by taking the generated data, filtering/transforming them, and proactively moving them to locations where it will be processed (in-situ or offline) or by moving the computation to where the relevant data are located. Since data are the core for research and insight for a broad set of academic disciplines, fast and seamless access to data in a usable form becomes critical for innovative research and educational programs across S&E domains.

Privacy, Security and Integrity. While sharing data and code within and across public and private sectors is a critical aspect of collaborative scientific discovery, issues of data and code integrity, privacy and security are paramount in this process. Emerging domains like edge computing, confidential cloud computing, and secure distributed computation introduce new security vulnerabilities and privacy concerns especially when designed explicitly for, and operated at, extreme scale. These issues must be addressed to ensure controlled and proper dissemination of data and code in order to ensure trust among the various stakeholders. Defining privacy, security and integrity measures, policies, and regulations of both data and software will be critical elements in the advancement of collaborative science. Trust in the integrated privacy and security measures for sensitive datasets, end-to-end scientific workflows, and software artifacts will be integral elements of the national Data and Software CI ecosystem.

Integration, Interoperability and Reusability. A broad class of new S&E applications must deal with data and software from multiple sources that may be heterogeneous in a variety of ways, such as the type, syntax and semantics of the data, the quality of the data, the platform and interfaces of the software, and the policy regime under which the data and software were produced and by which they can be used. Developing robust, scalable and flexible solutions that would provide interoperability between these intrinsically diverse and disparate data and software components is a key requirement for the S&E applications which depend on them. Integrated data, software, and CI components with seamless interoperability will serve as an impetus for novel scientific discoveries by enabling more complex end-to-end scientific workflows.

Curation, Provenance and Findability. Significant efforts are needed to curate datasets – to clean, enrich, and standardize the data, record the context as well as semantics associated with the data, and log the analyses performed on the data as well as the code used for those analyses – in order to make them more useful for scientists involved in data discovery and analysis. Effective and proper reuse of data and code demands that the data and code context be appropriately registered and that their semantics be extracted and represented. Research in automated data tagging, metadata generation and registration, semantic representation methods, ontologies, and provenance of data and software will be essential for the discovery and collaborative exploration of all relevant data by researchers across all S&E domains.

Analytics and Visualization. Synthesis of the information content and deriving insight from massive, dynamic, ambiguous, and even conflicting data can be achieved through advanced data analytics and visualization techniques. Transforming data into new knowledge and understanding is a crucial step for advancement in S&E, and this can be done using advanced data analysis tools and collaborative visual interfaces. A new generation of data analytics and visualization systems and services will help absorb vast amounts of data and enhance

researchers' ability to interpret and analyze otherwise overwhelming data. In this way, researchers will be able to detect the expected and discover the unexpected, uncovering hidden associations within vast data sets and making new scientific breakthroughs.

Software Frameworks, Abstractions and Libraries. New abstractions and programming frameworks will be necessary to simplify the challenges of programming scalable and parallel systems, while achieving maximal performance through exploitation of parallelism for scheduling computation, communication, and output for interactive as well as batch-oriented S&E applications. The proper set of abstractions must be provided to enable applications to specify their resource requirements and execute efficiently in an environment with shared resources. The development of domain-specific as well as cross-domain robust, reliable and efficient software frameworks, abstractions, and libraries should continue since they are critical for the rapid advancement of S&E.

Resource Allocation, Scheduling and End-to-End Workflow Management. End-to-end data processing and analysis is generally performed via data analysis pipelines and workflows. For this reason, many S&E communities depend on access to services that enable the creation of robust, reliable, efficient and scalable scientific workflows, and integration of the diverse data, computing, analysis, and monitoring capabilities. Comprehensive tools and best practices are needed to ensure that existing analysis pipelines are efficient, reliable, and scalable and that the results can be replicated at some future point in time if needed. A next generation of resource allocation, task scheduling, and end-to-end workflow management solutions is needed to allow for the efficient and scalable processing, analysis, visualization, and sharing of large datasets generated among highly diverse and interdisciplinary groups.

While no one technical solution will likely be able address the expansive S&E research enterprise that NSF supports, NSF is interested in understanding how different data-related CI solutions might support heterogenous ensembles of data-intensive disciplines – owing, for instance, to common requirements due to similarities in data set sizes, types and utilization workflows, or to novel shared goals for cross-disciplinary data integration and discovery. NSF is especially interested in Data and Software CI systems and services that build on existing and future data sources (including repositories) and address the current bottlenecks involved in publishing, discovery, access, management and processing of the data. NSF's future investments in Data and Software CI will continue to be guided by the research needs and priorities of the science, engineering, and education communities.

3.4 Data and Software CI Pathways to Production

NSF will continue to catalyze the development, implementation and evolution of a functionally complete national Data and Software CI that integrates data, computing, and networking CI assets and services to support S&E research and education. NSF will invest in the establishment of world-class Data and Software CI systems and services that are secure, efficient, reliable, accessible, usable, pervasive, persistent and interoperable, and that are able to exploit the full range of research and education tools available at any given time. NSF will promote the development of partnerships to facilitate the sharing and integration of distributed Data and Software CI components deployed and supported at the local, regional, national, and

international levels. Significant resources already exist at different levels, and it is essential to integrate such resources into the national Data and Software CI fabric.

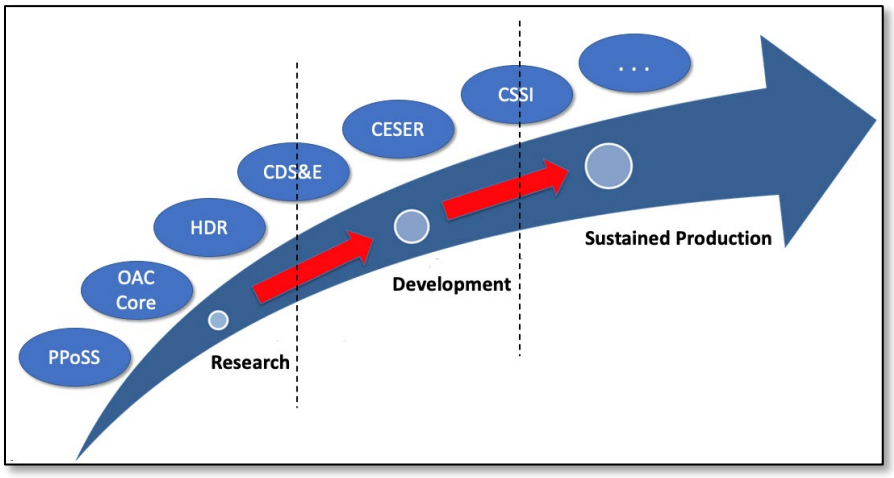


Figure 2: Data and Software CI pathways to production.

NSF investments in Data and Software CI ***pathways to production*** will be under three broad categories (as outlined in Figure 2) which are described below.

Data and Software CI Research. NSF recognizes and supports foundational and translational research to catalyze core Data and Software CI innovations essential to address disruptive changes in applications and technologies as well as the emergence of new concerns (e.g., energy efficiency, privacy, trust, transparency). There are multiple open research issues leading to advances in technologies for storing, accessing, sharing, integrating, and analyzing data, as well as for developing, managing and sustaining complex software. Fundamental understanding is needed not only in modeling and theory but also in designing new architectures, novel visualizations, and the effective utilization and optimization of data, software, computing, and communications resources. Insertion of these advances into the next generation of Data and Software CI needs to occur through close collaboration between the researchers, user communities, developers, and the providers of these systems, tools, and resources. To address new challenges in data-to-knowledge pipeline, there should be continuous and increasing investments in research on Data and Software CI technologies for large-scale data collection, management, analysis, interpretation, preservation, and security. Machine-learning approaches, including deep learning systems, are also needed to build better data-driven models that can be used to augment human decision making reliably. NSF's current programs such as *PPoSS*[28], *OAC Core*[27], *HDR*[9], and *CDS&E*[10], and new initiatives such as *AI Research Institutes*[58], will continue to support research in novel and innovative techniques in Data and Software CI.

Data and Software CI Development. NSF is heavily invested in supporting the development of a robust, secure, trusted, performant, scalable, and sustainable Data and Software CI ecosystem to enable and accelerate S&E research. Consistent with the *Federal Big Data Research and Development Strategic Plan*[50], NSF recognizes that there are significant data handling challenges common across disciplines, while some challenges are specific to particular disciplines. Some aspects of Data and Software CI may focus on specific application domains, while others are

common and shared across multiple research domains. Investments in both categories are critical for creating the envisioned Data and Software CI ecosystem that drives new thinking and transformative discoveries in all areas of S&E research and education. The former is important so that domains with specific and complex data and software challenges can be well supported with resources optimized for those applications; and the latter so that a shared infrastructure can offer access to resources that an individual community alone would not be able to build and sustain. NSF's current programs such as *CSSI*[5] and *CESER*[11] will continue to support the development of new Data and Software CI systems and services that are findable, accessible, interoperable, reusable, provenance traceable, and sustainable.

Data and Software CI Sustained Production. NSF aims to enable the deployment and operation of sustained production-quality Data and Software CI systems, tools and services. For this reason, NSF is developing a strategy that balances innovations with stability and continuity in production-quality Data and Software CI while ensuring that there are opportunities to explore innovations and to transition these innovations to production when appropriate. It is essential to have a clear plan for scaling Data and Software CI research prototypes and early implementations developed through other NSF programs mentioned above and transition them to production in order to increase productivity and ensure sustained scientific innovation across S&E. NSF is planning new initiatives complementing its existing programs (such as *CSSI*[5]) in this area, closing the gap between research, development and sustained production of Data and Software CI. The envisioned *Pathways to Production* will balance innovations with stability and continuity in production-quality Data and Software CI while ensuring that there are opportunities to explore innovations and to transition these innovations to production when appropriate in order to increase productivity and ensure sustained scientific innovation across S&E domains.

3.5 Putting the Plan into Action

NSF, through OAC, is aggressively moving ahead to put the blueprint for the national Data and Software CI outlined in the previous sections into action through programs and projects in FY 2021 and beyond. This action plan includes enhancing NSF's current programs in this landscape as well as development of new initiatives complementing the existing programs. A notional execution timeline for this plan is presented in Figure 3.

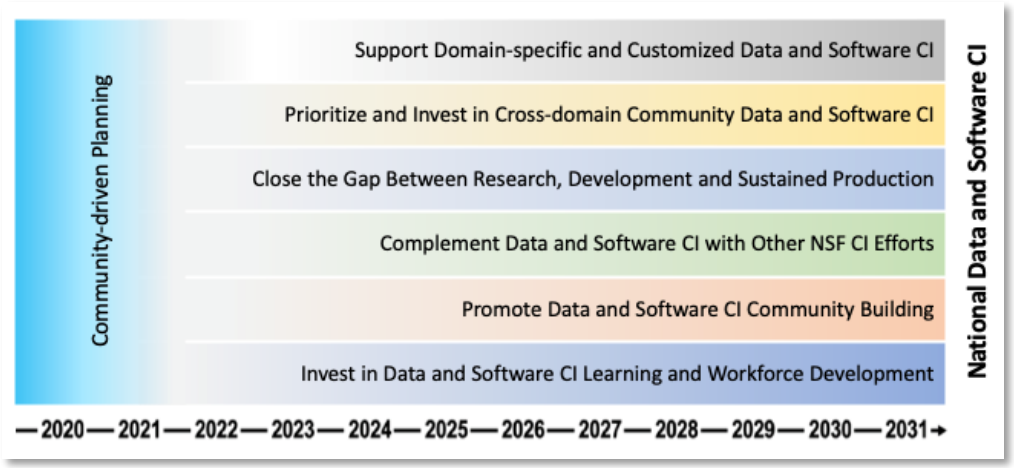


Figure 3: Notional execution timeline for the national Data and Software CI blueprint.

4 Ongoing Strategic Planning and Community Engagement

NSF will continue its strategic planning and community engagement activities as it explores other elements of the national CI ecosystem, including networking, cybersecurity, and learning and workforce development. NSF notes that this blueprint is the fourth in a series focused on different elements of the national CI ecosystem being developed in partnership with the community.

Looking to the future, the road ahead for research CI promises to be exciting with many new opportunities. In the near term, NSF's Big Ideas are moving into full gear with multiple new solicitations. These investments are now complemented by recent relevant national initiatives in areas such as Quantum Information Science (including the National Quantum Initiative Act[53]) and Artificial Intelligence (including an Executive Order on Maintaining American Leadership in Artificial Intelligence[54]). For example, NSF's investments in the Quantum Leap Challenge Institutes[55] and the Quantum Computing & Information Science Faculty Fellows (QCIS-FF; NSF 19-507)[56], as well as its investments in foundational and use-inspired artificial intelligence[57] research [see the recently-launched National Artificial Intelligence (AI) Research Institutes program (NSF 20-604)[58] which includes a specific theme on *"AI and Advanced Cyberinfrastructure"*], will help define the nature and structure of the CI ecosystem over the longer term. NSF looks forward to continuing to work with the community to define the future of CI research and research CI, with the overarching goal of realizing an integrated national CI ecosystem that transforms all S&E research and education.

5 Conclusion

The NSF-funded CI ecosystem is playing an increasingly critical role across all of S&E research and education, enabling discoveries and driving innovation. It is an essential part of the national CI ecosystem critical for ensuring US leadership in S&E, economic competitiveness, and national security. As a result, it is essential that NSF strategically evolves this CI ecosystem in response to the changing nature of S&E needs and the changing technology landscape, all the while being informed by community inputs. Building upon NSF's recently-articulated vision for a national CI ecosystem that integrates computational, data, software, networking, and security resources, tool and services, and computational and data skills and expertise, this document presented NSF's blueprint for a national Data and Software CI to support S&E research and education in the 21st century. It also outlined a plan to implement this blueprint. The vision and blueprint presented in this document have been informed by the community through advisory bodies, requests for information (RFIs), and workshops and conferences, as well as by national initiatives.

[53] "National Quantum Initiative," https://www.congress.gov/115/bills/hr6227/BILLS-115hr6227enr.pdf.
[54] "Executive Order on Maintaining American Leadership in Artificial Intelligence," https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.
[55] "NSF Quantum Leap Big Idea," https://www.nsf.gov/news/special_reports/big_ideas/quantum.jsp.
[56] "Quantum Computing & Information Science Faculty Fellows (QCIS-FF)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505535.
[57] "Artificial Intelligence (AI) at NSF," https://nsf.gov/cise/ai.jsp.
[58] "National Artificial Intelligence (AI) Research Institutes," https://www.nsf.gov/pubs/2020/nsf20604/nsf20604.htm.

NSF intends to continue to work with the community to evolve and implement the vision and blueprint presented in this document and develop complementary blueprints for other key elements of the national CI ecosystem.