National Science Foundation

CONVERGENCE ACCELERATOR SUMMARY REPORT: Descriptive Characteristics from 2019 to 2021

April 2023

A report from the Evaluation and Assessment Capability Section of the National Science Foundation.

About the Evaluation and Assessment Capability Section

The Evaluation and Assessment Capability (EAC) Section bolsters NSF efforts to make informed decisions and promote a culture of evidence. Located in the Office of Integrative Activities of the Office of the Director, EAC provides centralized technical support, tools, and resources to conduct evidence-building activities and to build capacity for evidence generation and use across the agency. EAC is led by NSF's Chief Evaluation Officer.

About this report

This report was prepared for EAC under contract number 49100421D0014 and order number 49100421F0211. The views expressed are those of the authors and should not be attributed to NSF, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Preferred citation

Wyhof, Annie, Michael Lee, Claire Lecornu, Mia Thomas, Shay Moore, Christa Reid, Dylan Solden, and Roland Stephen. 2023. *Convergence Accelerator Summary Report: Descriptive characteristics from 2019 to 2021*. Alexandria, VA: National Science Foundation.

NSF Quality Certification: Level 1

Quality Certifications

Level 1 — The author(s)/contractor(s) are responsible for the quality and conclusions presented in this report.

Level 2 — NSF verified that this report underwent quality assurance procedures and contributed to assessing its content.

Level 3 — NSF independently reproduced the analysis presented in this report.

Abstract

The Convergence Accelerator (CA) program at the National Science Foundation pursues national-scale, societal challenges through use-inspired convergence research. Its team-based approach is designed to have impact far beyond the laboratory by reshaping the way participating researchers approach their work. SRI International (SRI) has collected administrative data from 2019 to 2021 for Tracks A through F of the CA program. From these data, SRI generated descriptive statistics about CA applicants, grantees, and teams. This report presents SRI's findings about proposals submitted, partnerships, team size, requested personnel funding, proposed time commitments to projects, and the educational backgrounds and professional experiences of senior personnel, drawn from proposal metadata, budgets, and biographical sketches (biosketches). The outputs of this report seek to support NSF's ability to assess, evaluate, and communicate the impact of the CA.

Table of Contents

List of Tables

List of Figures

1. Executive Summary

Launched in 2019, the Convergence Accelerator (CA) is a National Science Foundation (NSF) program designed to move scientific discovery out of the lab and into practice to achieve long-lasting societal impact. The CA program provides "researchers and innovators the opportunity to accelerate their research toward tangible solutions that make a difference" (NSF CA Overview n.d.) by employing a novel combination of convergent research (i.e., a team-based, interdisciplinary approach to problem solving) and research acceleration through a mixture of collaboration and competition between teams.

Each year since 2019, the CA program has supported a cohort of teams to address the research topics chosen by CA program staff for that year's cohort. Topics are selected after an extensive process of soliciting and reviewing topic ideas from the public. Teams submit proposals for the specific selected research topic they intend to address. For awarded teams, there are two phases to conducting the research. During Phase 1, CA teams spend up to one year attending training and developing a proof of concept. After Phase 1, teams apply to participate in Phase 2 for which a smaller number of teams is selected. In Phase 2, CA teams are expected to provide deliverables that address societal needs at scale.

In 2021, in order to better understand the key characteristics of the researchers and teams that have participated in the CA program, NSF contracted with SRI International (SRI) to collect and aggregate data drawn from proposals to the program. The results of this work are contained in the report that follows. In it are descriptive statistics about CA applicants, grantees, proposed project personnel and team characteristics for the 2019, 2020, and 2021 cohorts. Specifically, SRI examined the biographical sketches (biosketches) and budget tables found in competitively reviewed proposals for Phase 1 and Phase 2 of the CA program for the 2019 and 2020 cohorts and Phase 1 of the CA program for the 2021 cohort.

Our approach focused on biosketches and budget tables because these sections contain the most complete and consistently formatted data for institutions, teams, and proposed project personnel. Biosketches provide information on the professional background for some, but not all, proposed project personnel. Budget tables provide information about institutions, proposed funding levels, and the roles assigned to proposed project personnel for each CA team.

There are notable limitations to our approach. Our sample of competitively reviewed proposals undercounts the actual number of such proposals for the CA program from 2019 to 2021. Next, while the biosketch and budget table sections were the most complete and consistently formatted proposal sections with information about team composition and personnel, they required parsing of unstructured data from NSF's Solr database using natural language processing and manual labeling. While mitigation steps were taken, miscategorized and dropped data are possible. Furthermore,

because SRI analyzed project proposals rather than reports from projects as they were implemented, the statistics presented on team composition, background, and budgets are based on data as proposed at the time of submission.

Given these limitations, the findings of this report are descriptive and should not be interpreted as establishing causal relationships. However, these descriptive findings will allow NSF and other stakeholders to better understand the characteristics of the CA program.

Key Findings

- **Geographic participation increased:** From 2019 to 2021, the geographic dispersion of Phase 1 applicants and grantees increased.
    - For the 2019 cohort, Phase 1 proposals were submitted by and awarded to lead institutions in 53% and 41% of U.S. states (inclusive of Washington D.C.), respectively.
    - Examining the 2019, 2020, and 2021 cohorts together, Phase 1 proposals were submitted by and awarded to lead institutions in 78% and 53% of U.S. states, respectively.
    - Examining the 2019, 2020, and 2021 cohorts together, lead institutions in 11 states have never submitted a Phase 1 proposal, and lead institutions in an additional 13 states have submitted a Phase 1 proposal but have not received an award.

- **The number and types of institutions on a proposal varied between phases:** The average number of institutions on proposals varies between phases and across cohorts. Also, the percent of proposals that are grantee partnership projects—those that are multi-organizational and have at least one non-academic partner—increased between Phase 1 and Phase 2.
    - For most cohorts, the average number of institutions on a proposal increased between Phase 1 and Phase 2.
    - Phase 1 proposals commonly featured a multi-institutional partnership with at least one non-academic partner with an average of 48% (across Tracks A to F).
    - Partnership formation increased in Phase 2 to about 76% of proposals (across Tracks A to D). Further, the private and nonprofit sectors increased their representation in proposals in Phase 2. Specifically, 57% of Phase 2 proposals included at least one private sector institution, 38% included at least one nonprofit institution, and 37% had *both* a private sector and nonprofit partner.

- **Teams grew between phases:** Concordant with the different activities associated with each phase, team size and variation in team composition increased from Phase 1 to Phase 2.

- For 2019 to 2021 cohorts, the median number of people on a team increased from 14 in Phase 1 proposals to 49 in Phase 2 proposals.
- When examining the team roles across all Phase 1 proposals, the most common role listed is senior personnel, with an average of six senior personnel per proposal team.
- When examining the team roles across all Phase 2 proposals, the most common team roles are graduate students (an average of 12 per proposal team) and other professionals (also with an average of 12 per proposal team).

- **Higher education was the largest sector in the CA program:** When examining the sectors associated with both the institutions included in proposals and the proposed project personnel who submit biosketches, the most frequent was the higher education sector.
  - In Phase 1 proposals, 91% of lead institutions are higher education institutions.
  - In Phase 2 proposals, while almost all of which are multi-organizational, 72% of all institutions on proposals are in the higher education sector.
  - Among the project personnel with data in Phase 1 and Phase 2 proposals, 89% worked in higher education at the time of proposal submission.

- **Project personnel are highly qualified:** Most project personnel that submitted a biosketch hold a PhD.
  - Of the proposed project personnel with available data, 86% have obtained a PhD.
  - Most proposed project personnel with a PhD have at least a decade of expertise at the time of proposal (that is to say, a decade following award of their PhD).
  - Examining all of the degrees earned by project personnel with available data, almost all of the degrees were earned in a science or engineering field.

2. Introduction

In 2016, the National Science Foundation (NSF) unveiled a set of 10 "Big Ideas," cutting-edge research agendas and processes uniquely suited to NSF's broad portfolio of investments that require collaboration with industry, private foundations, other agencies, and universities. NSF identified Growing Convergence Research (GCR) as one of the 10 Big Ideas. The NSF's Convergence Accelerator (CA) program specifically addresses policymakers' increasing interest in accelerating use-inspired convergence research to address national-scale, societal challenges.

Each year since 2019, the CA program has admitted a cohort of teams that conduct convergence research around a topic chosen by CA program staff. Topics are chosen each year after an extensive process of soliciting ideas from the public. Each of the 2019, 2020, and 2021 cohorts pursued one of two research tracks. In the CA program, there are two phases to conducting convergence research. During Phase 1, CA teams spend up to one year attending training and developing a proof of concept. After Phase 1, teams apply to participate in Phase 2, for which a smaller number of teams are selected. In Phase 2, CA teams spend up to two years refining their product and developing a business model to complete the translation of research into use.

The purpose of this report is to provide descriptive statistics on applicants, grantees, and projects for the 2019, 2020, and 2021 CA cohorts. The approach SRI International (SRI) adopted focused on analyzing CA proposals using proposal data from the Solr repository of NSF materials. Our approach relied on text analytics in combination with manual data cleaning.

Specifically, we examined data on competitively reviewed proposals for Phase 1 and Phase 2 of the CA program for the 2019 and 2020 cohorts and Phase 1 of the CA program for the 2021 cohort (at the time of our analysis, the CA program had not yet solicited proposals for Phase 2 for the 2021 cohort). This includes Tracks A through F, which cover the following research topics: Open Knowledge Networks; Artificial Intelligence (AI) and the Future of Work; Quantum Technology; AI-Driven Innovation via Data and Model Sharing; Networked Blue Economy; and Trust and Authenticity in Communication Systems.

We chose proposals as the principal data source in to get the most complete understanding of CA teams and their key characteristics. Specifically, we focused on proposals' biographical sketches (biosketches) and budget tables, as these sections contain the most complete and consistently formatted proposed project personnel and team data of all the sections in the proposals, making them more ready for analysis.

There are notable limitations to our approach. Our sample of competitively reviewed proposals under counts the actual number of such proposals for the CA program from 2019 to 2021. Next, while the biosketch and budget table sections contain information about team composition and personnel, they required parsing of unstructured data

using natural language processing and manual labeling. While mitigation steps were taken, miscategorized and dropped data are possible. Furthermore, because SRI analyzed project proposals rather than reports from projects as they were implemented, the statistics presented on team composition, background, and budgets are based on data as proposed at the time of submission.

Given these limitations, the findings of this report are descriptive and should not be interpreted as establishing causal relationships. Nonetheless, our descriptive findings are detailed and allow NSF and other stakeholders to understand the particular characteristics of the institutions, teams, and personnel involved in the CA program.

In the sections that follow, we begin by reviewing the history of the CA program and provide an overview of program cohorts and tracks from 2019 to 2021. Next, we describe the data sources used, as well as the methods and limitations of our descriptive analysis. We then present descriptive statistics on proposing institutions, team size, proposed personnel budgets and time commitments, institutional partnerships, and the academic background of project personnel. Finally, we describe the data challenges we encountered.

Throughout the presentation of our data methods and findings, there are a few terminology points to note. Unless otherwise specified, “institutions” refers to both lead institutions and non-lead institutions on a proposal. Unless otherwise specified, “proposals” refers to all CA proposals in our analytical sample, both Phase 1 and Phase 2 proposals. Similarly, “awards” refers to all Phase 1 and Phase 2 awards except where noted.

3. Program History

The primary goal of the CA program is to accelerate the transition of basic research and discovery into use in order to deliver solutions with societal impact. In pursuit of this goal, the program is designed to foster multi-disciplinary research, cross-cutting partnerships (across different institutional sectors), and coopetition (a mix of competition and cooperation among teams). This distinctive approach allows researchers and innovators to accelerate their research and develop tangible, sustainable solutions with a real-world application.

The CA program's unique model operates as follows: First, program staff solicit ideas that would be a suitable focus for future research through a Dear Colleague Letter (DCL), a mechanism NSF uses to inform the research community about upcoming opportunities. Anyone, regardless of affiliation, is allowed to submit a topic idea. Based on the topic ideas received, proposers of the ideas are selected to organize a workshop that brings together subject matter experts, stakeholders, and other interested parties to further refine the topic. These workshops then guide CA program staff in finalizing topics for the next year's CA cohort. Program staff choose two or more topics, called tracks.[1] Project teams then respond to an NSF solicitation with their proposals to conduct research for their chosen track. Project teams that receive awards for any of the tracks in a given year comprise a cohort, integrating their work through facilitated collaboration.

Once teams for a cohort have been selected, Phase 1 of convergence research begins. Phase 1 awards are for up to $1 million for a nine-month grant.[2] During this period, grantees participate in training on human-centered design, team science, and storytelling. They also receive individualized coaching, a distinctive feature of the program. In addition, teams are encouraged to collaborate with other teams as they execute their research. At the end of Phase 1, teams are expected to have developed a conceptual prototype and participate in a formal NSF pitch that describes their research and proposed solution to the societal issues associated with their track.

In Phase 2 of convergence research, the final phase in the CA program, teams are selected to participate based on their pitch and formal Phase 2 proposal. Phase 2 awards are for up to $5 million over two years. During Phase 2, grantees refine their product, build partnerships, and develop a plan to sustain their idea beyond the CA program in order to complete the translation of research into use. At the end of Phase 2, teams are expected to provide deliverables that can address societal needs at scale without further support and funding from NSF. At the time of our analysis, described in this report, no team had yet completed Phase 2.

---

[1] The number of tracks is subject to the availability of funds.
[2] The maximum Phase 1 award for the 2019 and 2020 cohorts was $1,000,000; all future cohorts have had a maximum award amount of $750,000.

Table 3.1 provides a timeline of major milestones for the CA program. It lists the dates of many of the first steps in the program, but it does not include all program activities since its founding. The DCL for the first (pilot) cohort occurred without the use of a workshop to refine topics. Awards for each phase for each cohort have occurred in September. In March 2022, the CA program was moved to the new Directorate for Technology, Innovation, and Partnerships (TIP) within NSF. TIP focuses on creating technological breakthroughs, scaling use-inspired and translational research to meet societal and economic needs, making it an ideal home for the CA program.

Table 3.1: Select History of the CA Program from 2018 to 2022

Date	Key Program Milestones
06/2018	Program launches within the Office of Integrative Activities
03/2019	First DCL published for the first cohort
05/2019	First DCL published for future track topic workshops
06/2019	Research project proposals submitted for the first cohort
Fall 2019	First track topic workshops held
09/2019	First Phase 1 awards made for 2019 cohort
09/2020	First Phase 2 awards made for 2019 cohort
09/2020	Phase 1 awards made for 2020 cohort
09/2021	Phase 2 awards made for 2020 cohort
09/2021	Phase 1 awards made for 2021 cohort
03/2022	Program moved to Directorate for Technology, Innovation, and Partnerships

Source: NSF Convergence Accelerator website: https://beta.nsf.gov/funding/initiatives/convergence-accelerator.
Note: This table presents dates of CA program milestones.

Table 3.2 presents details of each cohort and research track, as of the 2021 cohort. Each track-specific summary contains an overview of the problem space for each track, solution possibilities, and a brief review of related literature to describe the state of each field that the CA program seeks to accelerate.[3]

[3] The process for gathering the information about each track shared in the remainder of this section began by identifying and obtaining any workshop materials that informed the creation of each track. Our track summaries were also informed by materials about tracks found in public NSF materials and a brief literature review of publications related to topics within the CA tracks.

Table 3.2: CA Program Cohorts, Tracks, and Topics from 2019 to 2021

Cohort	Track	Track Topic
2019	A	Open Knowledge Networks
2019	B	AI and the Future of Work
2020	C	Quantum Technology
2020	D	AI-Driven Innovation via Data and Model Sharing
2021	E	Networked Blue Economy
2021	F	Trust and Authenticity in Communication Systems

Source: NSF Convergence Accelerator website: https://beta.nsf.gov/funding/initiatives/convergence-accelerator.
Note: This table presents the CA program cohorts, tracks, and track topics through 2021.

Track A is about Open Knowledge Networks (OKNs). It funds researchers from disciplines such as computer science, social science, communications, and psychology to create tools and a non-proprietary infrastructure for building OKNs. Track A is currently in Phase 2, in which NSF has funded six teams. The anticipated completion date for Track A is September 2022.[4]

Track B is about the Future of Work, with sub-tracks focusing on artificial intelligence (AI), Future Jobs, and the National Talent Ecosystem. It funds researchers from disciplines such as public policy, labor economics, computer science, business administration and management science, education, and real estate development to prepare workers for the jobs of the future. The track is currently in Phase 2, in which four teams were funded. The anticipated completion date for Track B is September 2022.

Track C, Quantum Technology, funds researchers from disciplines such as quantum theory, physics, physical sciences, computer science, engineering, economics, psychology, and sociology to develop quantum technologies toward a universal quantum computer. Track C is currently in Phase 2, in which NSF has funded four teams. The anticipated completion date for Track C is September 2023.

Track D, AI-Driven Innovation, funds researchers from disciplines such as data science, engineering, computer science, mathematics, and social science to explore the gap between fundamental research and the implementation of AI and data-driven tools in society. Track D supports research that focuses on (a) dataset(s) of interest and importance to one or more communities and (b) the tools, platforms, and protocols that make the data FAIR (Findable, Accessible, Interoperable, Reusable). Track D is currently in Phase 2, in which NSF has funded six teams. The anticipated completion date for Track D is September 2023.

[4] Phase 2 completion dates are subject to change as teams can file for a no-cost extension.

Track E, Networked Blue Economy, funds researchers from disciplines such as biology, oceanography, computer science, engineering, social science, art, and communications to conduct translational research toward an integrated ecosystem for ocean innovation, exploration, and sustainable utilization. Track E is currently in Phase 1, in which NSF funded 14 teams. Phase 1 for Track E is expected to be completed in September 2022, with Phase 2 running September 2022–2024.

Track F, Trust and Authenticity in Communication Systems, funds researchers from disciplines such as computer science, sociology, psychology, political science, and law to address inauthentic behavior in online and digital systems. Track F is currently in Phase 1 in which NSF has funded nine teams. Phase 1 for Track F is expected to be completed in September 2022, with Phase 2 occurring September 2022–2024.

4. Data, Methods, and Limitations

4.1 Data Sources and Sample Coverage

We examined CA proposal data, such as budget tables and biosketches, and proposal metadata, such as award status and submission date, to produce descriptive statistics about CA applicants, grantees, project personnel, and teams to provide deeper insights into characteristics and trends within the CA program. We used two NSF data systems: Solr and Report Server SQL.

Solr is a database containing the machine-readable full text of NSF proposals. We extracted data containing the proposal data (budget tables and biosketches) and proposal metadata from NSF's Solr server via an Application Programming Interface (API) request on December 1, 2021 using NSF's virtual desktop environment. The metadata fields used in analyses for this report include NSF's ID number for the proposal, the proposal's title, the name of the lead institution, the state in which the lead institution for that proposal was located, proposal award status, the date NSF received the proposal, the date NSF made a recommendation or decision on the proposal, a flag variable indicating whether the proposal is a collaborative submission, the program announcement to which a team submitted the proposal, and the managing program code for the proposal. In total, we used the Solr data system to obtain 815 proposal PDFs and tabular proposal data containing 815 rows and 89 fields.

We chose Solr as the primary data source for this report because only Solr contains complete machine-readable proposal data. The use of proposal data is key to our approach because only proposal content can show the full extent of team composition, which can be used to calculate trends. The CA focuses heavily on collaboration across institutions, which cannot be understood with metadata available in the NSF Report Server SQL database alone. For example, there is a proposal metadata field in the SQL tables that can be used to ascertain the number of Principal Investigators (PIs) and Co-Principal Investigators (Co-PIs) per team; however, there is no metadata field that contains the total size of the team.

Report Server SQL database provided additional administrative data. This database has many proposal metadata fields in common with Solr, such as when a team submitted a proposal and when NSF made an award decision for a proposal. However, the Report SQL database does not contain proposal data (e.g., biosketches, budget tables).[5] Data parsed from proposal components, such as the count of senior and other personnel from proposal budget data, are also not available in Report SQL.

We obtained performing organization classification for each institution from Report Server SQL since it is not included in the Solr database. This variable contains the self-

[5] We also consulted NSF's Oracle database to confirm the number of competitively reviewed proposals for each fiscal year, shown in Table 4.2. We did not use additional data from this source.

reported classification of the lead institution for each proposal.[6] The lead institution is the institution responsible for submitting the proposal to NSF and that will directly receive the NSF grant, if awarded.[7] We did not use any other data from Report SQL.

4.1.1 Analytical Sample of Interest

The data sample that we collected from Solr included competitively reviewed proposals submitted to the 2019, 2020, and 2021 cohorts of the CA program, which we obtained with a query based on specifying the program announcements for the CA program.[8] The proposal statuses discovered in the data are shown in Table 4.1.

Our analytical sample consists of only competitively reviewed proposals—meaning with set of proposals with statuses of "Proposal has been awarded" (124) or "Decline, DDConcurred" (273). Of the 124 proposals with awarded status, we did not include six proposals because they were not submitted for consideration in the 2019, 2020, or 2021 cohorts.[9] Thus, our analytical sample includes 118 awarded proposals and 273 declined proposals for a total of 391 proposals.

[6] Specifically, the performing organization field is available in Report SQL but not in Solr.
[7] As a hypothetical example, if College University planned to work with researchers from Nonprofit International and Private Company Inc. through subcontracts to each institution, the lead institution would be College University. The classification describes what type of institution the lead institution is, for example, "PhD and Equivalent Degree (Public, State)" or "Large Corporation – Business."
[8] We used the following CA program announcements to obtain the data: PD 19-095Y, NSF 20-555, NSF 20-565, and NSF 21-572 (NSF 2019, NSF 2020, NSF 2021 [May]).
[9] These six proposals instead concerned workshops for CA track topic development.

Table 4.1: Sample of Proposal Status for CA Proposals from 2019 to 2021

Proposal Status	Observed Entries in Solr	Count
Awarded*	Proposal has been awarded	124
Declined*	Decline, DDConcurred	273
Not competitively reviewed	Ret w\o Rev-Prel not invited, DDConcurrd	1
Not competitively reviewed	FL Withdrawn, Other	11
Not competitively reviewed	FL Withdrawn Preproposal	9
Not competitively reviewed	Prel Ret w\o Rev - Recd past deadln	1
Not competitively reviewed	Ret w\o Rev-Recd past deadln, DDConcurrd	1
Not competitively reviewed	Ret w\o Rev-Not resp NSF FundOp, DDConcrd	1
Not competitively reviewed	Ret w\o Rev - Dup in review, DDConcurred	1
Pending	Pending, Review Package Produced	32
Pending	Pending, PM recommends decline	1
White paper – not invited	Full Proposal not invited	263
White paper – invited	Full Proposal invited	97
		Total: 815

Source: NSF administrative data from Solr (*status* field); accessed 12/1/2021.
Note: This table contains the status values in Solr for all available CA proposals using the CA program announcements. The count for each status follows, and the statuses of interest for the analytical sample are marked with an asterisk (*), indicating their inclusion. Proposals that NSF did not competitively review were either returned or withdrawn. NSF made 118 awards to CA projects; NSF made six additional awards for track topic workshops. These six awards are not used in the analytical sample, as described in the introduction of Table 4.2.

Table 4.2 describes the data coverage of our Solr analytical sample. Population count per fiscal year is based on the number of competitive proposals listed in the Program Portfolio dashboard of NSF's Oracle Enterprise Reporting database.

Table 4.2: Sample Coverage for Competitively Reviewed CA Proposals by Fiscal Year and select Track

Time or Track	Population Count	Sample Count
By FY		
FY 2019	50	44
FY 2020	153	134
FY 2021	195	191
FY 2022	52	21
FY Unknown	0	1
By Recent Track		
Track E	137	99
Track F	73	41

Source: NSF administrative data from Solr (*dd_rcom_date* field – accessed 12/1/2021), NSF administrative data from Oracle (accessed 6/15/2022), and CA program information shared with SRI via email (received 3/18/2022).
Note: This table shows the data coverage of SRI's analytical sample.

For all fiscal years, the population count is higher than our sample count.[10] In both Track E and Track F, the population counts are higher than the Solr sample count.[11] Overall, by fiscal year, and by track, our sample under counts the population and the magnitude of this under count varies by track and fiscal year.

The descriptive statistics provided in this report use several units of analysis. Each figure and table in this report is presented with the sample size and unit of analysis for those data. The unit of analysis for statistics based on proposal metadata, such as the geographic location of the lead institution, is a competitively reviewed proposal observed in Solr. For statistics based on biosketch information, the unit of analysis is individual biosketches submitted on competitively reviewed proposals available in Solr with available biosketch data. For example, the unit of analysis for a table showing the number of individuals working at universities is people, parsed from biosketches submitted on competitively reviewed proposals. For statistics based on data from proposed budget tables, the unit of analysis is proposals or institutions. For example, when describing the team size, a team is considered to be the people proposed in the budget tables of one proposal.

[10] There is one proposal in our analytical sample that was competitively reviewed (and subsequently awarded) and whose value for the division director recommend for award date was listed in Solr as "NA"; this is listed in the table as "FY Unknown".
[11] As of March 18, 2022, 51 additional proposals were submitted through the Broad Agency Announcement (BAA) mechanism and were not available via Solr at the time of our API request, so they are not included in the report. However, they are included in the population counts for Track E and Track F.

Table 4.3 shows the sample sizes for all three data sources. Relative to our analytical sample of 391 competitively reviewed proposals, the biosketch data are a sample of 389. One of the competitively reviewed proposals was excluded because it had irregularly formatted data and an unreadable PDF, so we could not manually reference the biosketch data for input. We excluded the other competitively reviewed proposal from analysis because the biosketch contained a paragraph “bio blurb” that did not follow the NSF biosketch structure or content requirements.

Table 4.3: Sample Size of Data Sources for Competitively Reviewed CA Proposals from 2019 to 2021

Data Source	Sample Size	Tracks Affected (Count)
Proposal Metadata	391	N/A
Biosketches	389	A (1), B (1)
Budget Tables	386	A (3), B (1), E (1)

Source: NSF administrative data from Solr (*budget* and *bio* fields); accessed 12/1/2021.
Note: This table contains the sample sizes of the three data sources used in analysis of CA proposals. Analyses conducted with proposal metadata have a sample size of 391. Analyses conducted with budget table data have a sample size of 386. Analyses conducted with biosketch data have a sample size of 389. The unit of analysis is competitively reviewed proposals.

Relative to our analytical sample of 391 competitively reviewed proposals, 5 proposals are not included in the sample on budget tables. For 4 of these 5 proposals, the data from the budget tables were irregularly formatted and could not be accurately reconciled, either due to underlying data difficulties or an unreadable PDF that could not be referenced manually. The remaining proposal had budget tables attached but no personnel listed on the tables. Proposed project personnel either left all lines blank or only filled the line in with a zero.

Next, we compared the senior personnel proposed in budget tables and the senior personnel whose biosketches are included in the proposal. Table 4.4 shows the total number of people for whom we have data from the biosketches and proposed in budget tables.

Table 4.4: Count of Biosketches and Budgeted Personnel for Competitively Reviewed CA Proposals from 2019 to 2021

Category	Count – All Proposals	Count – Awarded Only
Submitted Biosketches	3,467	1,137
Budgeted Senior Personnel	2,842	886
Budgeted Total Personnel	7,460	2,365

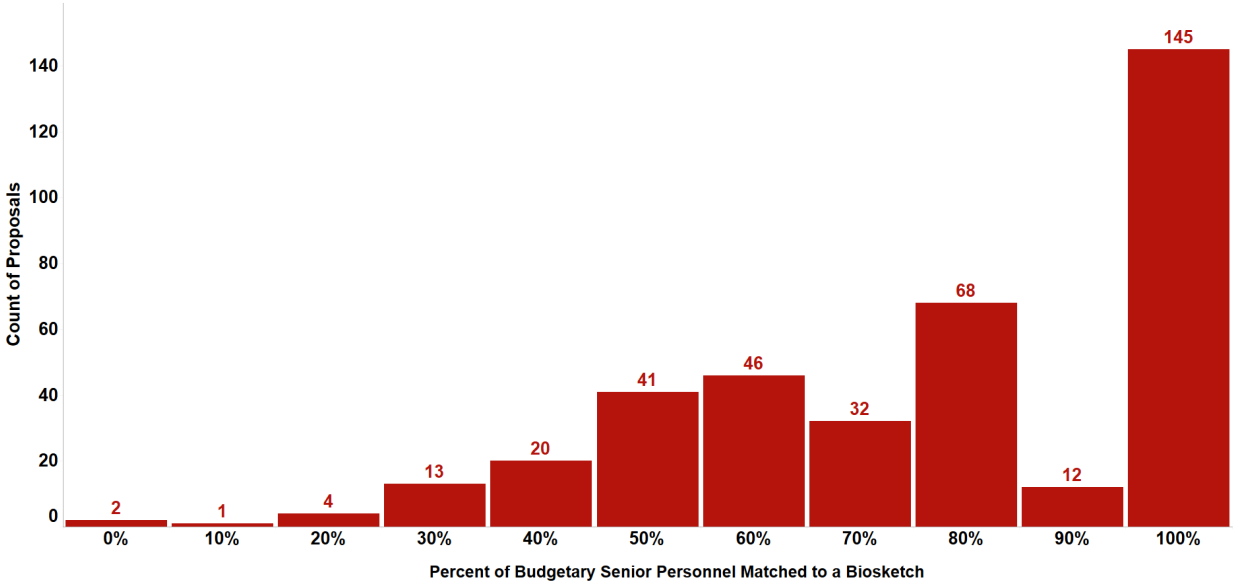
Source: NSF administrative data from Solr (*budget* and *bio* fields); accessed 12/1/2021
Note: This is a table of the count of proposed personnel on budget tables for all submitted proposals and the count of proposed personnel on budget tables for all awarded proposals. The sample size is 391 proposals with budget table or biosketch data, and the unit of analysis is people, as observed in submitted biosketches or submitted budget tables. There are 7,460 proposed budgeted personnel for all submitted proposals and 3,467 biosketches in all submitted proposals.

Examining proposals using biosketches and budget tables can seemingly yield divergent descriptions of team composition, discipline, and expertise. We view these data sources as complementary and find value in looking at CA teams from the point of view of both the budget personnel and biosketch personnel. Specifically, the budget tables permit describing the whole team whereas the biosketch data permit describing the backgrounds of specific team members. Hence, we linked the biosketch and budget table data by comparing the extracted names of individuals from the budget table and biosketch sections within a proposal. The resulting sample size is 384 proposals with both budget table and biosketch data.[12]

Figure 4.1 shows a histogram representing the percentage of named senior personnel in a budget table of a submitted proposal that we mapped to a submitted biosketch. Approximately 34% of proposals (134 proposals out of 384 total proposals) have biosketches for between 90% and 100% of all named senior personnel proposed in a budget table. Around 16% of proposals (64 out of 384 total proposals) have biosketches for between 70% and 80% of named senior personnel proposed in a budget table.

[12] A decline relative to our analytical sample of 391 due to the exclusion of five proposals that do not have budget table data and two different proposals that do not have biosketch data.

Figure 4.1: Percent of Named Senior Personnel Matched to a Biosketch for Competitively Reviewed CA Proposals from 2019 to 2021



Source: NSF administrative data from Solr (*bio* and *budget* fields); accessed 12/1/2021
Note: This histogram illustrates how many named senior personnel proposed in a budget table of a proposal have a biosketch included in that proposal. Thirty-four percent of proposals (134 proposals out of 38 total proposals with biosketch and budget table data) have biosketches for between 90% and 100% of all named senior personnel proposed in a budget table. Proposals that included no biosketches or no budget table data are excluded from this figure. The sample size is 384 proposals, and the unit of analysis is competitively reviewed proposals.

The proposals that fall in the lower bins (indicating a failed link between biosketches and persons in a budget table) do so not because the data are missing, but because of large discrepancies between the personnel listed in the budget tables and the personnel whose biosketches are included with the proposal. Some of these proposals in the lower portion of the chart had a relatively small number of biosketches compared to the number of named senior personnel occurring in the budget table data for that proposal. Others listed many people in the budget table and included many biosketches, but there was little crossover between the two groups. One proposal did not name any senior personnel in their budget table, even though the rest of the budget table form was filled out.

4.1.2 Supplemental Data Collection

To describe the nature and frequency of partnerships, we defined four institutional sectors: higher education, the private sector, the public sector, and nonprofit organizations. To classify the institutional sector of each lead institution, we mapped the performing organization field values from Report Server SQL to one of the four

sectors.[13] To supplement the sector classification for lead institutions, we used manual internet searches to collect data about the institutional sectors of non-lead institutions listed in the budget tables of CA proposals. We used institution names to conduct an internet search for the institution to determine its sector category. For example, we have classified an institution whose website indicated it is organized as a 501(c)(3) as belonging to the nonprofit sector.

To describe the higher education institutional sector, we used the Carnegie Classification of Institutions of Higher Education to categorize higher education institutions according to their research output (Carnegie Classifications 2021). Specifically, we matched institutions by name to the Carnegie Classification data. We did not link institutions for which we could not determine an equivalent institutional match to a Carnegie Classification.

To describe academic degree fields and the interdisciplinarity of proposed CA project personnel, we relied on the Taxonomy of Disciplines (TOD), a hierarchy of academic disciplines maintained by NSF's National Center for Science and Engineering Statistics (NCSES) (Taxonomy of Disciplines 2021). In the current version of the TOD, each degree discipline is assigned a taxonomy code based on six levels of classification, which help organize the distinct degree fields of proposed project personnel as reported on their biosketches. Specifically, we identified disciplines from the text of each biosketch and then matched these identified disciplines to discipline names from the TOD.[14] Merging academic degree disciplines with the classification levels from TOD allowed us to understand the interdisciplinarity of proposed CA project personnel and participants.

4.2 Methods and Variables of Interest

After accessing the CA analytical sample and assessing its coverage, we transformed the raw text data from the budget table and biosketch sections into the variables of interest necessary to create this report. We also added new CA-specific metadata to enable analysis by track and cohort.

A hypothetical example of the personnel section of a budget table is shown in Figure 4.2. The variables of interest extracted from each budget table are as follows: the number of senior personnel and their names and leadership roles, the number of non-senior personnel by team role, and the proposed time commitments and proposed budget amounts for each line in the personnel section of the budget table.[15]

---

[13] We did not collect information about non-lead institutions included on a proposal from Report Server SQL.
[14] We manually reviewed biosketch lines from the professional preparation section that our automated extraction method did not flag as containing a discipline. We manually matched biosketch text for which we could not identify an exact match in the TOD (e.g., we would match the discipline "Limnology" to "Aquatic Biology, Limnology" manually).
[15] Please refer to Appendix A for more details about how we extracted all variables of interest.

Figure 4.2: Hypothetical Submission by a Proposed Institution of a Proposal Budget Table

SUMMARY PROPOSAL BUDGET YEAR 1

ORGANIZATION **College University**				FOR NSF USE ONLY PROPOSAL NO.	DURATION (months) Proposed / Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR **Martina Gavri'el**				AWARD NO.	
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Martina Gavri'el - PI**	1.1	0.00	0.00	**28,470**	
2. **Kinich Bharata - Co-PI**	1.8	0.00	0.00	**29,946**	
3. **Giampiero Antonin - Co-PI**	2.3	0.00	0.00	**20,953**	
4. **Adrianna Flint - Senior Scientist**	2.5	0.00	0.00	**18,529**	
5. **Korbin Brylee - Faculty**	1.7	0.00	0.00	**20,612**	
6. (**2**) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	5	0.00	0.00	**37,076**	
7. (**7**) TOTAL SENIOR PERSONNEL (1 - 6)	14.4	0.00	0.00	**155,586**	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (**1**) POST DOCTORAL SCHOLARS		0.00	0.00	**40,544**	
2. (**1**) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)		0.00	0.00	**31,381**	
3. (**2**) GRADUATE STUDENTS				**16,687**	
4. (**2**) UNDERGRADUATE STUDENTS				**13,458**	
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**	
6. (**0**) OTHER				**0**	

Source: Hypothetical data filled in on the NSF budget template included with proposals.
Note: The names and data presented here are randomly generated and not drawn from actual CA proposal data. This figure is an example of a proposed budget table submitted for the CA program.

The number of senior personnel, their names, and leadership roles are shown in Section A of Figure 4.2. Senior personnel are those who have a leadership role, which includes PIs, Co-PIs, faculty, and other senior associates. For the first five senior personnel of a proposal, the name and leadership role of each person is listed on a separate line. If there are more than five senior personnel, the number of additional senior personnel is listed in line 6 of Section A. Line 7 of Section A shows the total number of senior personnel. Thus, as shown by the hypothetical example in Figure 4.2, there are seven senior personnel. The first five senior personnel have the following names and leadership roles: Martina Gavri'el - PI, Kinich Bharata - Co-PI, Giampiero Antonin - Co-PI, Adrianna Flint - Senior Scientist, and Korbin Brylee - Faculty. We extracted the variables associated with senior personnel using a combination of automated and manual approaches.

The number of non-senior personnel is shown in Section B of Figure 4.2. This represents the number of personnel in non-leadership roles including postdoctoral scholars, other professionals, graduate students, undergraduate students, and secretarial - clerical. As shown by the hypothetical example in Figure 4.2, there is one postdoctoral scholar, one other professional, two graduate students, two undergraduate students, and no secretarial - clerical or other non-senior personnel. We extracted the number of non-senior personnel using an automated approach.

The proposed time commitments are shown for the senior personnel in Section A and for postdoctoral scholars and other professionals in Section B of Figure 4.2. This represents the proposed amount of time that will be devoted to the project. There are three possible measures of time commitment: calendar year person-months, academic year person-months, and summer person-months. As an example, in Figure 4.2, Martina Gavri'el has 1.1 calendar year person-months committed to the project. We were able to develop an automated approach to extract these variables since the text adjacent to these values was consistent.

The proposed budget for each line in Section A and Section B are shown in the "Funds Requested by proposer" field in Figure 4.2. This represents the amount of funds that are requested in the proposal for each role. For example, in Figure 4.2, the proposed budget for total senior personnel is $155,586. We extracted this value using the same automated approach that was used to extract the proposed time commitments.

We also extracted information from particular sections of a biosketch, which contain the following sections: personnel, professional preparation, appointments, products, and synergistic activities. A hypothetical example of the first page of a biosketch is shown in Figure 4.3, illustrating the personnel, professional preparation, and appointment sections of a biosketch.

The personnel information of a biosketch is shown in the top section of a biosketch as shown in Figure 4.3. This includes the name of the person submitting the biosketch, the person's title, and the institution where the person works. In Figure 4.3, for example, the name of the person submitting the biosketch is Martina Gavri'el, her title is Associate Professor of Discipline, and she works at College University. We tried a combination of automated and manual approaches to extract the personnel information from a biosketch, but not all biosketches had a standardized personnel information section.[16] Consequently, the personnel section was not used in the biosketch analysis.

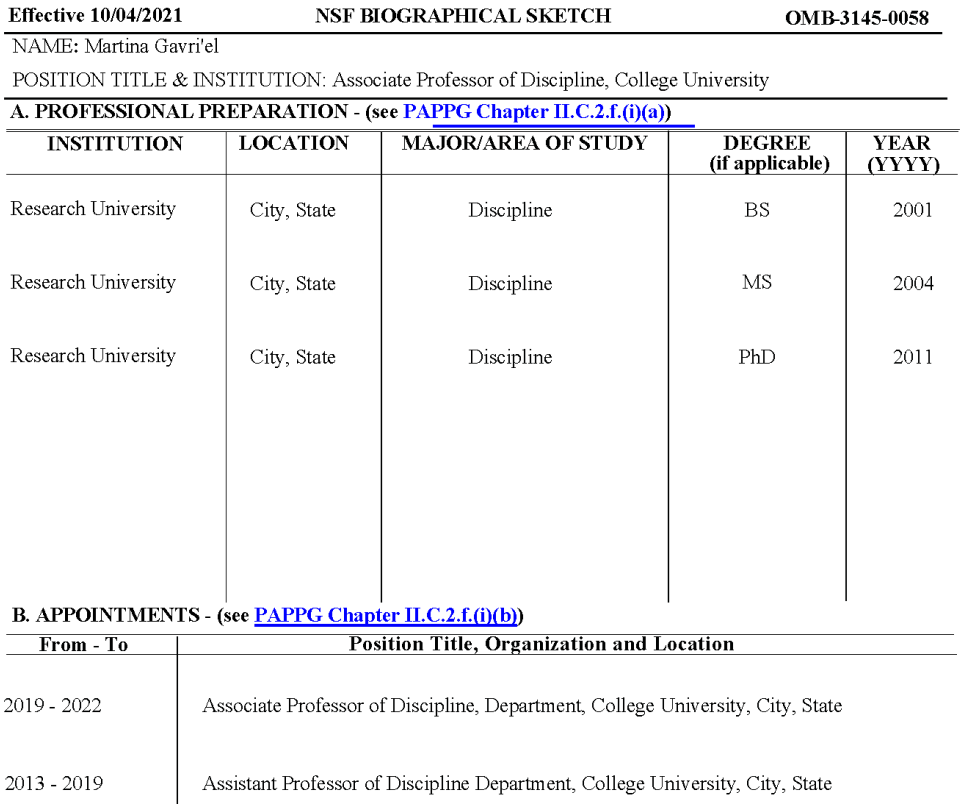
The professional preparation information of a biosketch is shown in Section A of Figure 4.3. This includes the institution, location, the area of study (discipline), degree attained, and year of degree completion. An example of professional preparation information is shown in Figure 4.3, where the institution is Research University, the location is City, State, the area of study is labeled as Discipline, the degree attained is PhD, and the year of degree completion is 2011. We extracted this information using automated processes involving regular expression queries.

The appointments information of a biosketch is shown in Section B of Figure 4.3. The information for each appointment includes the start and end date, position title, organization, and location. An example of appointment information is shown in Figure 4.3, where the start date of the appointment is 2019, the end date of the appointment is

[16] For example, some biosketches only included the name of the person submitting the biosketch, while others included contact information.

2022, the title is Associate Professor of Discipline, the organization is Department, College University, and the location is City, State. We extracted this information using automated processes involving regular expression queries.

Figure 4.3: Hypothetical Submission by a Proposed Senior Personnel of a Biosketch



Effective 10/04/2021 **NSF BIOGRAPHICAL SKETCH** **OMB-3145-0058**

NAME: Martina Gavri'el

POSITION TITLE & INSTITUTION: Associate Professor of Discipline, College University

A. PROFESSIONAL PREPARATION - (see PAPPG Chapter II.C.2.f.(i)(a))

INSTITUTION	LOCATION	MAJOR/AREA OF STUDY	DEGREE (if applicable)	YEAR (YYYY)
Research University	City, State	Discipline	BS	2001
Research University	City, State	Discipline	MS	2004
Research University	City, State	Discipline	PhD	2011

B. APPOINTMENTS - (see PAPPG Chapter II.C.2.f.(i)(b))

From - To	Position Title, Organization and Location
2019 - 2022	Associate Professor of Discipline, Department, College University, City, State
2013 - 2019	Assistant Professor of Discipline Department, College University, City, State

Source: Hypothetical data filled in on the NSF fillable biosketch template included with proposals.
Note: The names and data presented here are randomly generated and not drawn from actual CA proposal data. This figure is an example of the first page of a biosketch submitted for the CA program.

The products information of a biosketch contains up to five publications or products that are closely associated with the proposed project (e.g., a relevant journal article). We did not extract any publications or productions from the products section of the biosketch.

The synergistic activities information of a biosketch contains up to five examples that demonstrate the broader impact of the person's professional and scholarly activities (e.g., a service committee on which a faculty member served). We did not extract any synergistic activities.

4.3 Limitations

There are notable limitations to our approach. For one, this report contains descriptive statistics using data from proposals submitted to the CA program, rather than

descriptive statistics generated from reports of awards after projects have been funded.[17] The trends described in the descriptive statistics tables and figures based on proposal data may differ from the patterns of funded teams. The proposal data are early formation of ideas and teams, both of which are subject to change. Thus, examining proposal data may not accurately represent the true nature of funded proposals when describing characteristics like team size and project area of focus.

Next, while nearly all proposals have both biosketch and budget table data, two proposals did not have biosketches five proposals did not have budget tables. That is, the sample size for proposals with biosketches is 389 out of a possible 391 proposals for all analyses. The sample size for proposals with budget tables is 386 out of a possible 391 proposals for all analyses. This difference in data availability introduced a small limitation to our analysis.[18]

Further, many person and team characteristics of interest residing within submitted biosketches are not structured, readily accessible data within Solr and proposed project personnel from 2019 to 2021 did not always fill out forms completely or consistently. This complicated data extraction efforts and affected the data quality for the budget table and biosketch descriptive statistics.

For biosketches, proposed project personnel used a variety of ways to number and name sections. Additionally, project personnel did not use a consistent format when supplying responses to the personnel section.[19] For budget tables, some project personnel did not fill out all lines of the budget tables completely, and some left empty boxes instead of inputting the value zero for items with no value. Thus, many rows contained fewer than the four values expected by our automated extraction tool. To capture the full information about lines missing time commitments and proposed personnel amounts, we manually reviewed each nonstandard row.[20]

Also, the roles into which team members are grouped in budget tables are limited to the degree to which the available role options align with each team member's planned team activities. Some roles, such as PI and Co-PI, are formally defined. Others, such as faculty, postdoctoral scholar, and graduate student, map to roles common across academia. The other senior associate and other professional roles are not as clearly defined and thus may introduce variability regarding who is included in those roles.

---

[17] We selected this approach because it allows for comparison of awarded versus declined proposals and because there are more detailed raw data available for proposals than for project reports.
[18] Another data issue worth noting that may have introduced a small limitation to our analysis includes having to manually collected and appended CA-specific data necessary for analysis, such as the CA track and cohort to which teams submitted their proposals to our datasets. Although we used a rigorous process for manually identifying these metadata, it is possible that we miscategorized a proposal that did not state the desired track in its title and that we therefore used proposal content data to classify.
[19] For example, some project personnel listed their current occupation as well as their employing institution. Others only included their name and contact information.
[20] This process included checking the PDF version of the proposal, finding that budget table within the PDF, identifying from where the nonstandard value(s) came, and inputting zeroes for boxes left blank on the form.

Finally, it is clear that our sample of competitively reviewed proposals under counts the true population of proposals. Consequently, our geographic description of submissions and awards may under state the geographic growth and dispersion of the CA program from 2019 to 2021.[21] For our description of CA teams, team members, and collaborations among institutions, our under-counted sample could potentially yield a non-representative sample of the CA program from 2019 to 2021.[22]

Given all data limitations, our findings are descriptive and should not be interpreted as establishing causal relationships. They nonetheless seek to bolster NSF's, and other stakeholders', understanding of particular characteristics of the institutions, teams, and personnel involved in the CA program. Further, our data lessons learned may prove helpful towards future work on describing, assessing, and evaluating the CA program.

[21] For example, our under count of competitively reviewed proposals of approximately 130 proposals may result in an under count of the number of lead institutions within various states that have submitted a Phase 1 proposal to the CA program or submitted and received a Phase 1 award from 2019 to 2021. Further, our findings could overestimate the number of states whose lead institutions did not submit a Phase 1 from 2019 to 2021. Similarly, our findings could overestimate the number of states whose lead institutions submitted a Phase 1 proposal but did not receive an award from 2019 to 2021.

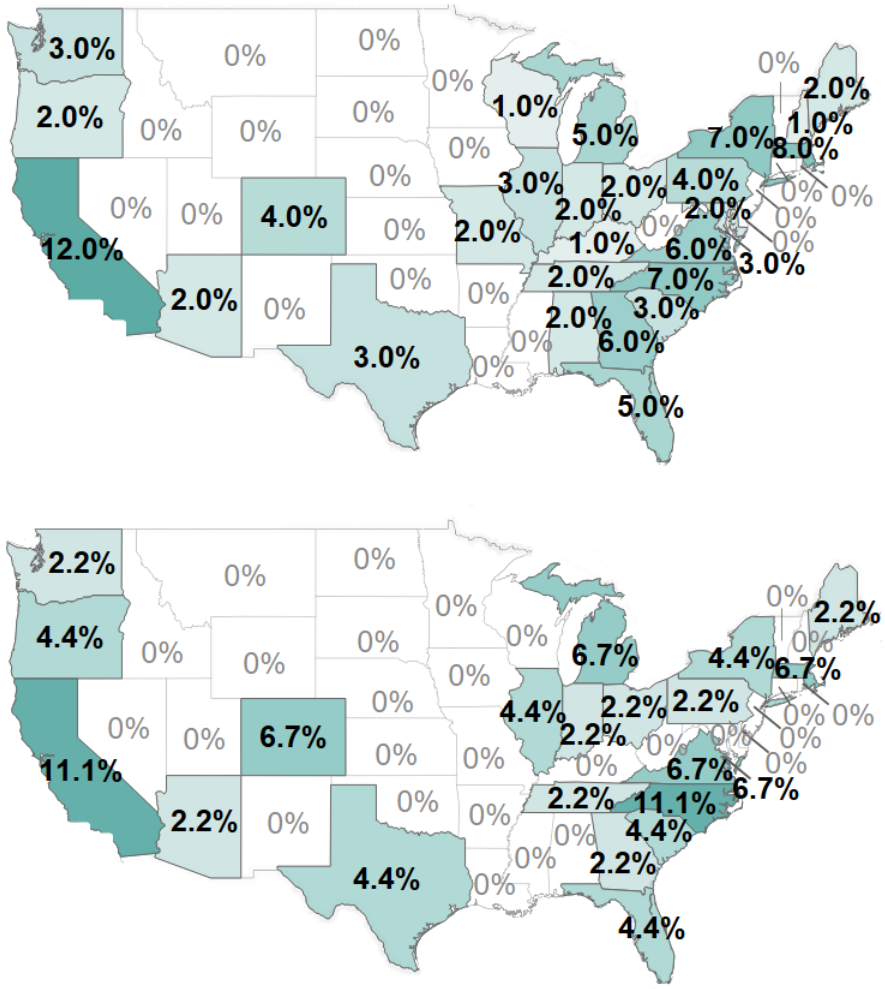
[22] That is to say, it is possible there would exist statistically significant discrepancies (in either direction) between our reported descriptive results and the results based on an underlying sample of all competitively reviewed proposals (and subsequent related data, such as biosketches and budget tables) rather than our under-counted sample.

5. Descriptive Findings

5.1 Program Growth

To describe program growth, we examined which states accounted for the highest shares of CA proposals from lead institutions. Figure 5.1 is a map of the distribution of lead institutions on submitted (top) and awarded (bottom) Phase 1 proposals for the 2019 cohort. Lead institutions in California make up 12% of submitted Phase 1 proposals for the 2019 cohort (12 out of 100 total 2019 Phase 1 proposals), the largest share of any state. Proposals submitted by lead institutions in Massachusetts make up 8% of submitted Phase 1 proposals for the 2019 cohort (8 out of 100 total 2019 Phase 1 proposals).

Figure 5.1: Percent of Submitted and Awarded 2019 Competitively Reviewed Phase 1 Proposals by State



Source: NSF administrative data from Solr (*inst_state* and *status fields*); accessed 12/1/2021. *Note:* This figure maps the distribution of lead institutions on submitted (top) and awarded (bottom) Phase 1 proposals for the 2019 cohort. The sample size is 100 submitted (top) and 45 awarded (bottom) proposals from Phase 1 of the 2019 cohort, and the unit of analysis is competitively reviewed Phase 1 proposals. Proposals from lead institutions in California make up 12.0% of submitted Phase 1 proposals for the 2019 cohort (12 out of 100 total 2019 Phase 1 proposals). Proposals with PIs at lead institutions in California and North Carolina each make up slightly more than 11% of Phase 1 awards for the 2019 cohort (5 out of 45 awarded 2019 Phase 1 proposals). Hawaii and Alaska are not shown in the figure; both make up 0% of lead institutions in 2019 submitted and awarded proposals.

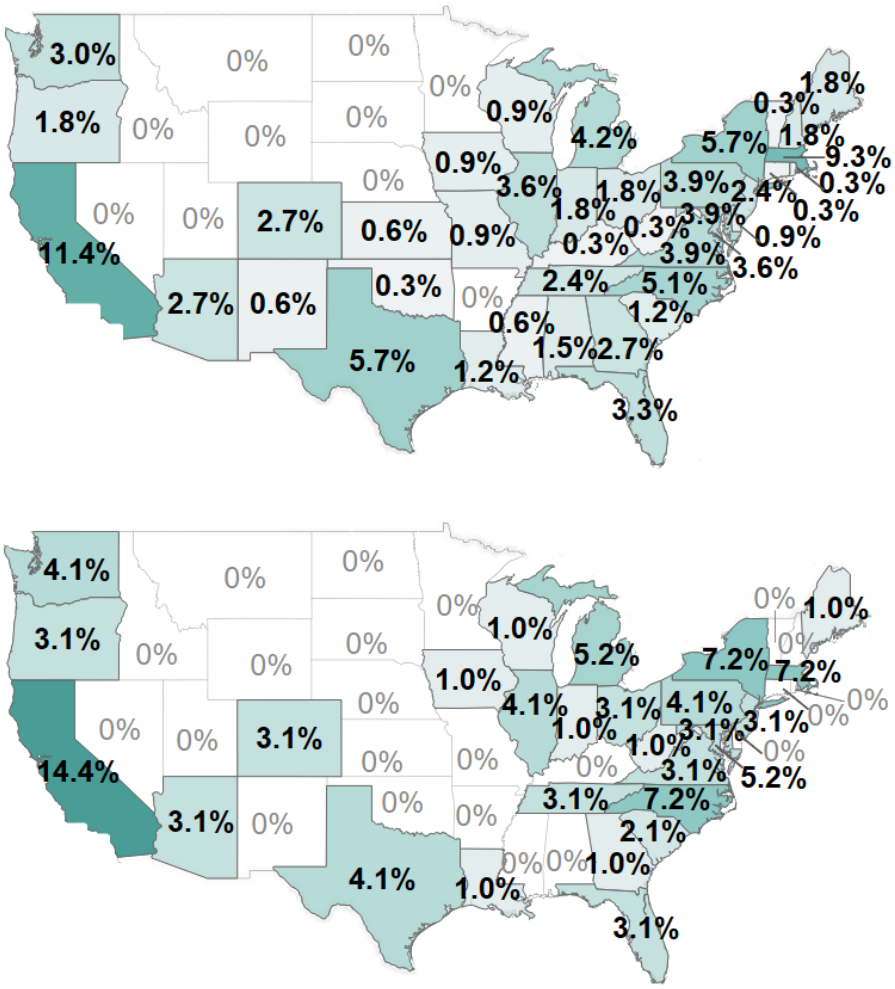
Among awarded proposals, lead institutions in California and North Carolina each make up slightly more than 11% of 2019 Phase 1 awards (5 out of 45 awarded Phase 1 proposals for the 2019 cohort), making California and North Carolina tied for the states with the most lead institutions with awarded proposals. Proposals submitted by lead institutions in Colorado, Washington, D.C., Massachusetts, Michigan, and Virginia each

make up approximately 7% of 2019 Phase 1 awards (3 out of 45 awarded Phase 1 proposals for the 2019 cohort). Fifty-three percent of U.S. states (27 out of 51 states, inclusive of Washington, D.C.) contain a lead institution that submitted a Phase 1 proposal in 2019, and 41% (21 out of 51 states) contain a lead institution that was awarded a Phase 1 proposal in 2019.

The data in Figure 5.2 below show increased geographic dispersion for lead institutions with submitted and awarded Phase 1 proposals over the first three years of the CA program combined. When looking at the first three program cohorts collectively, proposals from lead institutions in California make up the largest share with about 11% of submitted Phase 1 proposals for the 2019 to 2021 cohorts (38 out of 334 total 2019 to 2021 Phase 1 proposals). Proposals from lead institutions in Massachusetts make up around 9% of submitted Phase 1 proposals for the 2019 to 2021 cohorts (31 out of 334 total 2019 to 2021 Phase 1 proposals). Proposals from lead institutions in Texas and New York each make up slightly less than 6% of submitted Phase 1 proposals for the 2019 to 2021 cohorts (19 out of 334 total 2019 to 2021 Phase 1 proposals).

Among awarded proposals, lead institutions in California make up slightly more than 14% of awarded Phase 1 proposals for the 2019 to 2021 cohorts (14 out of 97 awarded 2019 to 2021 Phase 1 proposals), indicating that proposals from California institutions are awarded at a rate disproportionate to their submission count. Lead institutions in Massachusetts make up about 7% of awarded Phase 1 proposals for the 2019 to 2021 cohorts (7 out of 97 awarded 2019 to 2021 Phase 1 proposals). This pattern of program funding is consistent with the overall pattern of NSF research and development (R&D) funding received by states. In FY2020, California received the highest share of NSF's federal obligations for research and experimental development with almost 13%, followed by Massachusetts and New York, each with about 7% (NCSES 2021).

Figure 5.2: Percent of Submitted and Awarded Competitively Reviewed Phase 1 Proposals from 2019 to 2021 by State



Source: NSF administrative data from Solr (*inst_state* and *status fields*); accessed 12/1/2021.
Note: This is a map of the distribution of lead institutions on submitted (top) and awarded (bottom) Phase 1 proposals for all cohorts in the program to date as of this analysis (i.e., the 2019, 2020, and 2021 cohorts). The sample size is 334 submitted (top) and 97 awarded (bottom) Phase 1 proposals, and the unit of analysis is competitively reviewed Phase 1 proposals. Hawaii and Alaska are not shown in the figure. Of lead institutions on all submitted proposals, Hawaii and Alaska make up 0.6% and 0%, respectively. Both make up 0% of lead institutions on 2019 to 2021 awarded proposals.

Table 5.1 shows the count of U.S. states (including Washington, D.C.) with a Phase 1 submission from a lead institution by cohort grouping. The first cohort (2019) is compared to an aggregation of all cohorts to date (2019, 2020, and 2021). For the 2019 cohort, about 53% of states (27 out of 51 states) had a lead institution submit a Phase 1 proposal. The participation rate for all cohorts to date is higher at over 78% (40 out of 51 states).

Similarly, the number of states receiving a Phase 1 award has increased over time. For the 2019, over 41% of states (21 out of 51 states) had an institution that received a Phase 1 award. For all cohorts to date, about 53% (27 out of 51 states) had an institution that received a Phase 1 award.

Table 5.1: Count of States with Any Lead Institutions for Competitively Reviewed CA Proposals from 2019 to 2021

Counts	First Year (2019 Cohort)	Program to Date (2019–2021 Cohorts)
States with Submissions	27	40
States with Awards	21	27

Source: NSF administrative data from Solr (*inst_state* and *status* fields); accessed 12/1/2021.
Note: This table contains the counts of U.S. states (including Washington, D.C.) that dad any lead institutions that submitted a proposal for the 2019 cohort and for all cohorts in the program to date as of this analysis (i.e., the 2019, 2020, and 2021 cohorts). The sample size is 391 proposals, and the unit of analysis is U.S. states. For the 2019, 2020, and 2021 cohorts combined, about 78% of states (40 out of 51 states) had lead institutions that submitted a Phase 1 CA proposal and nearly 53% of states (27 out of 51 states) had an institution that received Phase 1awards.

Of the states that did not have a lead institution that received a Phase 1 award, nearly 80% (24 out of 30 states) did not have a lead institution submit a proposal for the 2019 cohort.[23] That is, lead institutions in 24 states did not submit a proposal, and thus did not receive an award, for the 2019 cohort. This is compared to about 46% (11 out of 24 states) for the program to date. That is, lead institutions in 11 states did not submit a proposal, and thus did not receive an award, when examining all submissions and awards from 2019 to 2021.

In summary, the geographic reach of the CA program expanded from 2019 to 2021. The number of states with any lead institution submitting a Phase 1 proposal has increased from 53% to 78%. The number of states with a lead institution receiving a Phase 1 award has increased from 41% to 53%. The number of states with lead institutions that had not yet submitted a Phase 1 proposal has declined from 24 to 11 states.

5.2 Proposed Team Size

To describe proposed team size, we examined the size and composition of proposed CA teams for Phase 1 and Phase 2 proposals for all cohorts in the program subject to this analysis (i.e., 2019, 2020, and 2021). We used proposal budget tables to obtain data on team size, overall, and by CA team role. Lastly, we assumed a team is defined as the people proposed in the budget tables for a given proposal.

[23] See Appendix B for more details.

Table 5.2 below compares all proposed Phase 1 and Phase 2 team sizes. For Phase 1 submissions, the mean proposed team size is about 15 people, and the median proposed team size is 14 people. The largest proposed team size is 47 people, and the smallest is one person. In contrast, proposed Phase 2 teams are on average much larger, which is expected since Phase 2 awards are for a longer duration, have higher award amounts, and have a broader scope. The mean proposed Phase 2 team size is about 50 people, and the median proposed team size is 49 people.

Table 5.2: Team Size Descriptive Statistics for Competitively Reviewed CA Proposals from 2019 to 2021 by Phase

Team Size Metric	Phase 1	Phase 2
Mean	14.5	49.8
Median	14	49
Range	46	89
Minimum	1	3
Maximum	47	92

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table compares proposed Phase 1 and Phase 2 team sizes. Phase 1 data are for the 2019 to 2021 cohorts; Phase 2 data are for the 2019 and 2020 cohorts. The sample size is 386 proposals, and the unit of analysis is people, as identified by budget tables on competitively reviewed proposals. The mean proposed team size is about 15 people for Phase 1 proposals and about 50 people for Phase 2 proposals.

Table 5.3 compares proposed Phase 1 and Phase 2 team sizes by track. Although the research domains are very different across tracks, there is little variation in typical team size by track for each phase. The median proposed team size for Phase 1 varied by track between a low of 13 and a high of 16 people. The median proposed team size for Phase 2 varied between a low around 42 and a high around 60 people.

Table 5.3: Team Size Descriptive Statistics for Competitively Reviewed CA Proposals from 2019 to 2021 by Phase and Track

Team Size	Phase 1						Phase 2			
Metric	A	B	C	D	E	F	A	B	C	D
Mean	13.8	13.9	16.3	17.5	13.5	12.6	49.6	46.8	57.9	46.5
Median	13.0	15.0	15.5	16.0	13.0	13.0	49.0	49.0	59.5	41.5
Range	33	25	39	41	26	21	89	64	61	18
Minimum	2	3	5	6	1	3	3	18	30	40
Maximum	35	28	44	47	27	24	92	82	91	58

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table compares proposed team size by track and phase. The sample size is 386 proposals, and the unit of analysis is people, as identified by budget tables on competitively reviewed proposals. The mean proposed team size for Phase 1 varied between a low around 13 and a high around 18 people. The mean proposed team size for Phase 2 varied between a low around 47 and a high around 58 people.

Table 5.4 shows the average number of proposed personnel on a CA research team in each CA team role for all Phase 1 and Phase 2 proposal submissions. There is a total of 4,820 people proposed on Phase 1 budget tables from the 2019 to 2021 cohorts.

Table 5.4: Average Team Composition for Competitively Reviewed CA Proposals from 2019 to 2021 by Role and Phase

CA Team Role	Phase 1	Phase 2
Senior Personnel	6.0	10.4
Other Senior	0.6	1.5
Other Professionals	2.2	12.0
Postdocs	0.8	4.5
Graduate Students	2.9	12.8
Undergraduate Students	0.9	3.4
Other	1.1	5.2
TOTAL	**14.5**	**49.8**

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table shows the average number of proposed personnel on a CA research team by CA team role for all Phase 1 (2019 to 2021 cohorts) and Phase 2 (2019 to 2020 cohorts) proposal submissions. The sample size is 386 proposals, and the unit of analysis is people, as identified by budget tables on competitively reviewed proposals. There is a total of 4,820 people proposed on Phase 1 budget tables and 2,640 people proposed in Phase 2 budget tables in CA proposals.

For Phase 1 proposals from the 2019 to 2021 cohorts, named senior personnel make up slightly less than 42% of people listed in CA proposal budget tables (2,008 out of 4,820 total individuals on Phase 1 proposals). The average proposed Phase 1 team has six people in the senior personnel role. Graduate student is the next most common role, with about three people in that role proposed on the average Phase 1 budget table (see the discussion in the previous section on the details of the budget table). Graduate student is the next most common role, with about three people in that role proposed on the average Phase 1 budget table (see the discussion in the previous section on the details of the budget table).

Because Phase 2 teams are much larger than Phase 1 teams, the average number of proposed project personnel increased between the phases for each team role. Graduate students and other professionals make up the greatest shares of proposed Phase 2 teams with 12 to 13 people in each role on the average team. The senior personnel role is close behind with about 10 people on the average Phase 2 team. Overall, and in ways consistent with the design of the CA program, teams in Phase 2 increase in size across project roles.

5.3 Proposed Personnel Amounts

To describe proposed personnel budgets and proposed time commitments of CA team members, we examined data the proposed budget tables. Budget tables contain proposed budget amounts for all team roles, including graduate students and undergraduates, and proposed time commitments for the non-student team roles.

Table 5.5 shows the average budget per person per month by team role in Phase 1 proposals for the 2019 to 2021 cohorts. There is a wide range of mean budget amounts per person per month across the tracks and team roles. For example, the senior personnel team role ranges from a low of $13,600 per person per month in Track E to a high of $18,690 in Track A. This is consistent with the slightly longer median period since the award of PhD observed for tracks A and C as noted in Table 5.23 below. Across all but Track F, the senior personnel role has the highest budgeted amount per person per month. For every track, the postdocs role has the lowest amount budgeted.[24]

[24] See Section 4 for definitions of personnel roles.

Table 5.5: Average Budget Amounts per Person per Month for Competitively Reviewed Phase 1 Proposals from 2019 to 2021 by Role and Track

CA Team Role	Track A	Track B	Track C	Track D	Track E	Track F
Senior Personnel	$18,690	$15,114	$17,252	$16,210	$13,600	$14,764
Other Senior	$13,487	$13,261	$11,982	$14,526	$12,861	$12,834
Other Professionals	$10,314	$9,978	$13,687	$11,998	$8,749	$15,022
Postdocs	$6,432	$6,979	$5,441	$5,455	$5,091	$4,740

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table contains the mean proposed budget amount per person per month in each team role on Phase 1 proposals from 2019 to 2021 by track. The sample size is 333 Phase 1 proposals with budget tables, and the unit of analysis is U.S. nominal dollars. Team role averages are only presented for proposals that had at least one person in that role. The mean proposed personnel budget amount per person per month for senior personnel for Phase 1 varies from $13,600 (Track E) to $18,690 (Track A) across Tracks A to F.

Table 5.6 shows the average budget per person per month by team role in Phase 2 proposals for the 2019 to 2020 cohorts by track. Though Phase 2 proposals have much larger budgets than Phase 1 proposals, after normalizing by per person per month, the budgeted amounts for each team role by track in Phase 2 are very similar to those in Phase 1. As in Phase 1, across Tracks A to D, the senior personnel team role has the highest budgeted amount per person per month of the four team roles with available time commitments data.

Table 5.6: Average Budget Amounts per Person per Month for Competitively Reviewed Phase 2 Proposals for the 2019 and 2020 Cohorts by Role and Track

CA Team Role	Track A	Track B	Track C	Track D
Senior Personnel	$16,650	$15,348	$17,385	$17,815
Other Senior	$14,727	$15,170	$15,375	$11,540
Other Professionals	$11,307	$8,614	$13,867	$8,588
Postdocs	$5,737	$5,451	$5,223	$5,146

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table contains the mean proposed budget amount per person per month in each team role by track on competitively reviewed Phase 2 proposals for the 2019 and 2020 cohorts. The sample size is 53 competitively reviewed Phase 2 proposals with budget tables, and the unit of analysis is U.S. nominal dollars. Team role averages are only presented for proposals that had at least one person in that role. The mean proposed personnel budget per person per month for senior personnel for Phase 2 ranges from $15,348 (Track B) to $17,815 (Track D) across Tracks A to D.

Differences in the mean budget amount per person per month can depend on a combination of factors, such as the mean budgeted amount per role in a proposal, the total personnel count per role in a proposal, and the total months committed for each team role in a proposal. Table 5.7 shows the data for the average proposed time commitments per person for selected CA team roles. Phase 2 data are for 2019 to 2021

cohorts; Phase 2 data are for 2019 and 2020 cohorts. When looking at the time commitment by phase, there is little variability by team role across tracks. The Phase 1 average proposed time commitment per person for senior personnel ranged from a low of 1.4 months (Track A) to a high of 1.7 months (Track E). The Phase 2 average proposed time commitment per person for senior personnel ranged from a low of 2.3 months (Track C) to a high of 3.8 months (Track B).

The Phase 2 proposed time commitment per person for postdocs ranged from a low of 10.1 months (Track D) to a high of 11.0 months (Track B). In general, Phase 2 proposals have a longer proposed time commitment per role compared to Phase 1 proposals. This is consistent with the longer duration of Phase 2 projects (24 months for Phase 2 compared to 9 months for Phase 1). Across both phases and all tracks, postdocs have the highest averages for months committed.

Table 5.7: Average Months Committed per Person for Competitively Reviewed CA Proposals from 2019 to 2021 by Phase and Track

CA	Phase 1						Phase 2			
Team Role	A	B	C	D	E	F	A	B	C	D
Senior Personnel	1.4	1.5	1.6	1.5	1.7	1.6	3.3	3.8	2.3	3.3
Other Senior	1.0	1.0	1.0	1.1	1.3	0.7	2.6	4.4	2.1	3.1
Other Professionals	3.8	3.8	2.7	4.4	2.3	3.9	4.8	5.9	2.9	6.5
Postdocs	6.5	11.5	6.9	6.9	7.3	7.6	10.5	11.0	10.3	10.1

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table contains the proposed time commitments per person for CA project personnel for the team roles for which time commitment data were available. The sample size is 386 proposals, and the unit of analysis is proposed months of work for a CA project, as identified by budget tables on competitively reviewed proposals. Team role averages are only presented for proposals that had at least one person in that role. The Phase 1 proposed time commitment per person for senior personnel ranged from a low of 1.4 months (Track A) to a high of 1.7 months (Track E). The Phase 2 proposed time commitment per person for senior personnel ranged from a low of 2.3 months (Track C) to a high of 3.8 months (Track B).

Table 5.8 shows the average total proposed budgets for each team role in Phase 1 proposals for the 2019 to 2021 cohorts. Across all tracks, undergraduates have the lowest average total budget amounts ($11,792 to $26,306) compared to all other roles, while senior personnel generally have the highest average total budget amounts ($98,419 to $122,097).

Table 5.8: Total Average Personnel Budget Amounts for Competitively Reviewed Phase 1 Proposals from 2019 to 2021 by Role and Track

CA Team Role	Track A	Track B	Track C	Track D	Track E	Track F
Senior Personnel	$99,037	$98,419	$143,617	$135,770	$117,479	$122,097
Other Senior	$41,621	$32,145	$24,073	$59,249	$29,353	$23,195
Other Professionals	$100,288	$91,841	$68,724	$75,610	$60,218	$73,459
Postdocs	$54,334	$74,148	$80,087	$82,922	$54,994	$55,979
Grad Students	$77,363	$83,353	$70,258	$92,536	$51,108	$65,511
Undergraduates	$20,462	$22,538	$26,306	$20,877	$11,792	$13,214

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table contains the mean proposed total budget amount in each team role by track on competitively reviewed Phase 1 proposals from 2019 to 2021. The sample size is 333 Phase 1 proposals with budget tables, and the unit of analysis is U.S. nominal dollars. Team role averages are only presented for proposals that had at least one person in that role. The mean proposed budget for senior personnel for Phase 1 varies from $98,419 (Track B) to $143,617 (Track C) across Tracks A to F.

Table 5.9 shows the average total proposed budgets for each team role in Phase 2 proposals for the 2019 and 2020 cohorts. The proposed budget amounts for each team role on Phase 2 proposals are generally higher than on Phase 1 proposals, and this is consistent with the larger scope and objectives of Phase 2 projects. The team roles with the highest average total personnel budgets are other professionals, especially Track B ($772,506) and Track D ($742,212).

Table 5.9: Total Average Personnel Budget Amounts for Competitively Reviewed Phase 2 Proposals from 2019 to 2020 by Role and Track

CA Team Role	Track A	Track B	Track C	Track D
Senior Personnel	$535,974	$436,792	$408,714	$499,536
Other Senior	$128,606	$204,922	$162,086	$127,392
Other Professionals	$547,174	$772,506	$399,466	$742,212
Postdocs	$290,396	$260,160	$439,731	$265,004
Grad Students	$381,803	$420,947	$336,636	$312,757
Undergraduates	$65,170	$56,148	$21,340	$60,725

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table contains the mean total proposed budget amount in each team role by track on competitively reviewed Phase 2 proposals for the 2019 and 2020 cohorts. The sample size is 53 competitively reviewed Phase 2 proposals with budget tables, and the unit of analysis is U.S. nominal dollars. Team role averages are only presented for proposals that had at least one person in that role. The mean proposed personnel budget for other professionals for Phase 2 varies from $399,466 (Track C) to $772,506 (Track B) across Tracks A to D.

Overall, Phase 2 proposals generally have a higher proposed budget and personnel time commitment as compared to Phase 1 proposals across all tracks. However, when

standardizing to a common person per month basis, Phase 2 proposals have similar budget amounts per person per month compared to Phase 1 proposals. Even so, the results may indicate Phase 2 provides more professional opportunities for postdocs, graduate students, and other professionals (as indicated by the average months committed per person by project role in Table 5.7).

5.4 Measures of Partnerships

To describe collaboration among institutions, we examined multi-institution awards, the distribution of institutions on CA proposals across different institutional sectors, and the characteristics of grantee partnership projects.

To describe multi-institutional awards, Table 5.10 compares the mean number of institutions on competitively reviewed proposals for the 2019 to 2021 cohorts by phase.[25] The mean number of institutions on a Phase 1 proposal varied from a low of 2.5 for the 2019 cohort to a high of 3.7 institutions for the 2020 cohort. The mean number of institutions on Phase 2 proposals increased for both the 2019 and 2020 cohorts, to 4.9 and 6.6 institutions respectively. This is expected as teams develop new partnerships during Phase 1 and collaborate toward developing a sustainability plan and a prototype with practical use.

Table 5.10: Average Number of Institutions in Competitively Reviewed Proposals from 2019 to 2021 by Cohort and Phase

Phase	2019 Cohort	2020 Cohort	2021 Cohort
Phase 1	2.54	3.67	3.17
Phase 2	4.89	6.56	N/A

Source: NSF administrative data from Solr (*budget* field); accessed 12/1/2021.
Note: This table compares the mean number of institutions in Phase 1 and Phase 2 proposals by cohort. Phase 1 data are for 2019 to 2021 cohorts; Phase 2 data are for 2019 and 2020 cohorts. The sample size is 386 proposals, and the unit of analysis is institutions, as observed on budget tables on competitively reviewed proposals. The mean number of institutions on Phase 1 proposals varied from a low of 2.5 institutions in the 2019 cohort's proposals to a high of 3.7 institutions in the 2020 cohort's proposals. The mean number of institutions on Phase 2 proposals increased from 4.9 institutions for the 2019 cohort to 6.6 institutions for the 2020 cohort.

Table 5.11 compares the mean number of institutions on competitively reviewed Phase 1 proposals by award decision by track across the 2019 to 2021 cohorts. Across all tracks, the mean number of institutions on awarded and declined proposals are all around three. Given a sample size of 333, this suggests that there is unlikely to be a relationship between team size and proposal success. The mean number of institutions on both awarded and declined proposals by track varied from a low around 2 institutions

[25] Because proposals had not yet been solicited for Phase 2 of the 2021 cohort at the time of this analysis, data on the average number of institutions for 2021 cohort Phase 2 proposals are not available.

(Track B) to a high around 4 institutions (Track C). This shows that most Phase 1 proposals, whether awarded or declined, were a partnership between at least two institutions.

Table 5.11: Average Number of Institutions in Competitively Reviewed Phase 1 Proposals from 2019 to 2021 by Award Decision and Track

Track	Awarded	Declined
All Tracks	3.04	3.16
Track A	3.13	2.50
Track B	2.41	2.21
Track C	3.64	4.00
Track D	3.00	3.78
Track E	3.29	3.21
Track F	3.33	2.97

Source: NSF administrative data from Solr (*status* and *budget* fields); accessed 12/1/2021.
Note: This table compares the mean number of institutions on Phase 1 proposals by award decision across the 2019 to 2021 cohorts. The sample size is 333 proposals, and the unit of analysis is institutions, as observed on budget tables on competitively reviewed Phase 1 proposals. The mean number of institutions on awarded Phase 1 proposals varied from a low around two institutions on Track B proposals to a high around four institutions on Track C proposals.

Table 5.12 shows the number of institutions by cohort and award decision for Phase 1 proposals for the 2019, 2020, and 2021 cohorts. Awards represented 45% of the total number of proposals (100) in 2019; 31% of the total number of proposals (94) in 2020; and, 16% of the total number of proposals (140) in 2021. Approximately 47% of awarded proposals had three or more institutions in 2019 (21 out of 45 awarded).[26] In the 2020 cohort, about 76% of awarded proposals had three or more institutions (22 out of 29 awarded). In the 2021 cohort, 78% of awarded proposals had three or more institutions (18 out of 23 awarded). These results show that between 2019 and 2021 an increasing share of awarded proposals had three or more institutions on their proposed budget tables—that is, a growth in institutional collaboration consistent with the intensions of the program.[27]

[26] For the 2019 cohort, 20 awards went to institutions that had no partners; however, in 2020 and 2021 this occurrence declined to 3 and 2, respectively.
[27] We also note a substantial increase in the overall number of submitted proposals in 2021, indicating increased interest in program participation.

Table 5.12: Number of Institutions in Competitively Reviewed Phase 1 Proposals by Cohort and Award Decision

No. of Institutions	2019 Cohort		2020 Cohort		2021 Cohort	
	Proposals Awarded	Proposals Declined	Proposals Awarded	Proposals Declined	Proposals Awarded	Proposals Declined
1	20	18	3	7	2	18
2	4	15	4	7	3	25
3	10	13	13	18	9	32
4	3	4	4	12	5	19
5	4	4	3	9	3	14
6	0	1	1	8	1	5
7	1	0	1	0	0	1
8	0	0	0	3	0	0
9	2	0	0	0	0	2
10	1	0	0	0	0	0
11	0	0	0	1	0	0
TOTAL	**45**	**55**	**29**	**65**	**23**	**116**

Source: NSF administrative data from Solr (*budget* and *status* fields); accessed 12/1/2021.
Note: This table shows the number of institutions on Phase 1 proposals by award decision and cohort. The sample size is 333 Phase 1 proposals, and the unit of analysis is competitively reviewed Phase 1 proposals. In the 2019 cohort, approximately 76% of awarded proposals had one to three institutions (34 out of 45 total awarded Phase 1 proposals for the 2019 cohort). In the 2020 cohort, about 69% of awarded proposals had one to three institutions (20 out of 29 total awarded Phase 1 proposals for the 2020 cohort). In the 2021 cohort, slightly less than 61% of awarded proposals had one to three institutions (14 out of 23 total awarded Phase 1 proposals for the 2021 cohort).

To describe proposals across institutional sector, we categorized institutions on proposal budget tables into one of four distinct institutional sectors: higher education, private, nonprofit, and public. Table 5.13 shows the percent and count of all institutions and lead institutions on CA proposal budget tables for each institutional sector, by phase. Phase 1 data cover the 2019 to 2021 cohorts; Phase 2 data cover the 2019 and 2020 cohorts.

For all institutions on Phase 1 proposals, approximately 75% (778 out of 1,040 institutions on Phase 1 proposals) are in the higher education sector. Approximately 13.5% of institutions in Phase 1 CA proposals (140 out of 1,040 institutions on Phase 1 proposals) are in the private sector. Almost 9% come from the nonprofit sector (93 out of 1,040 institutions). The higher education sector accounts for around 91% of lead institutions on Phase 1 proposals (304 out of 333 lead institutions on Phase 1 proposals). Less than 2% of Phase 1 proposals (5 out of 333 lead institutions on Phase 1 proposals) are led by a nonprofit or public sector institution.

The sector distribution of institutions does not vary much between Phase 1 and Phase 2. In Phase 1, higher education constitutes 75% of the participants and 91% of the lead institutions, while the private sector represents 13.5% of all participants, leading about 7% of the time. For all institutions on Phase 2 proposals, the higher education sector has the highest share at about 72% (207 out of 286 institutions on Phase 2 proposals). The private sector accounts for about 16% of institutions on Phase 2 proposals (46 out of 286 institutions on Phase 2 proposals). For all lead institutions on Phase 2 proposals, the higher education sector accounts for 93% (49 out of 53 lead institutions on Phase 2 proposals). No Phase 2 proposals are led by a public sector institution, and one is led by a nonprofit.

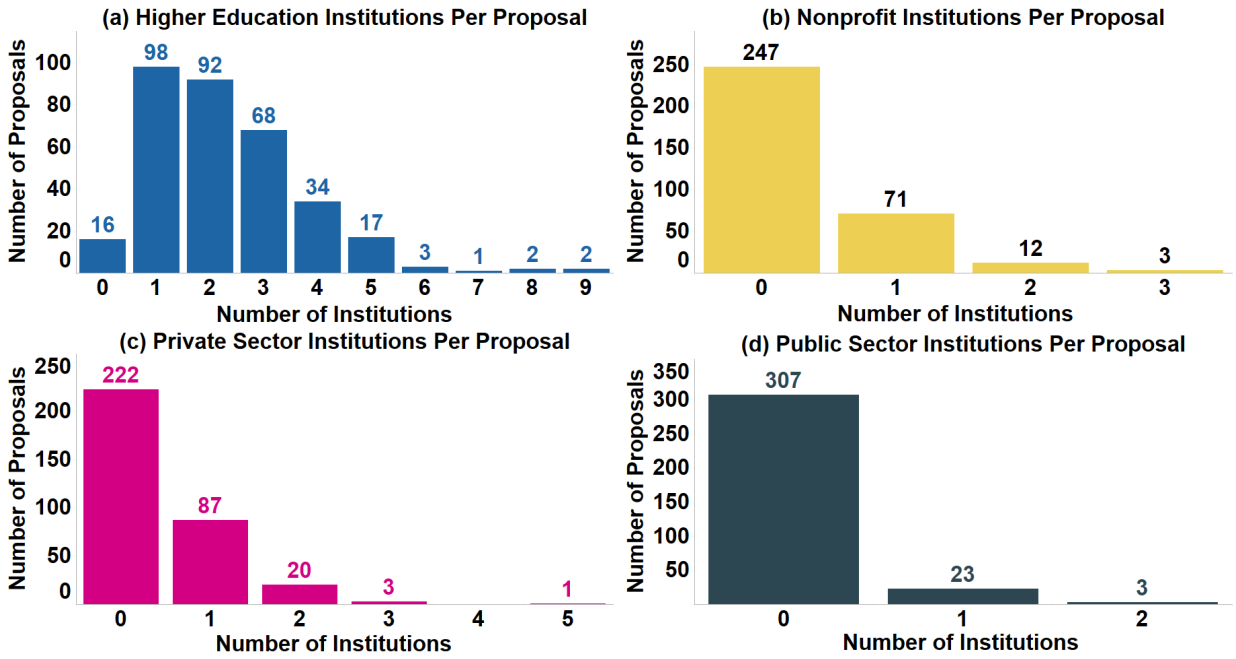
Table 5.13: Percent of Total and Lead Institutions for Competitively Reviewed CA Proposals from 2019 to 2021 by Phase and Sector

Phase	Inst.	Higher Ed.	Nonprofit	Private Sector	Public Sector
Phase 1	Total	74.8% (778)	8.9% (93)	13.5% (140)	2.8% (29)
Phase 1	Lead	91.3% (304)	1.2% (4)	7.2% (24)	0.3% (1)
Phase 2	Total	72.4% (207)	9.4% (27)	16.1% (46)	2.1% (6)
Phase 2	Lead	92.5% (49)	1.9% (1)	5.7% (3)	0.0% (0)

Source: NSF administrative data from Solr (*budget* and *inst* fields), accessed 12/1/2021; and Report SQL (*perf_org_txt* field), accessed 5/18/2022.
Note: This table contains the counts and percentages of all proposed institutions and lead institutions across institutional sectors by phase. A lead institution is the institution responsible for submitting a proposal and that ultimately is awarded an NSF grant; total institutions include lead institutions and any other institutions that submitted a budget table on the proposal. Phase 1 data are for 2019 to 2021 cohorts; Phase 2 data are for 2019 and 2020 cohorts. The sample size is 386 proposals, and the unit of analysis is institutions, as observed on budget tables on competitively reviewed proposals.

Figure 5.3 shows the frequency of higher education, nonprofit, private, and public institutions proposed on Phase 1 proposals from the 2019, 2020, and 2021 cohorts, either as a lead or non-lead institution. Higher education participated in 95% (317 of 333) of all Phase 1 proposals. Around 33% of proposals (111 out of 333) included at least one private sector institution. About 8% of proposals (26 out of 333) included a public sector institution.

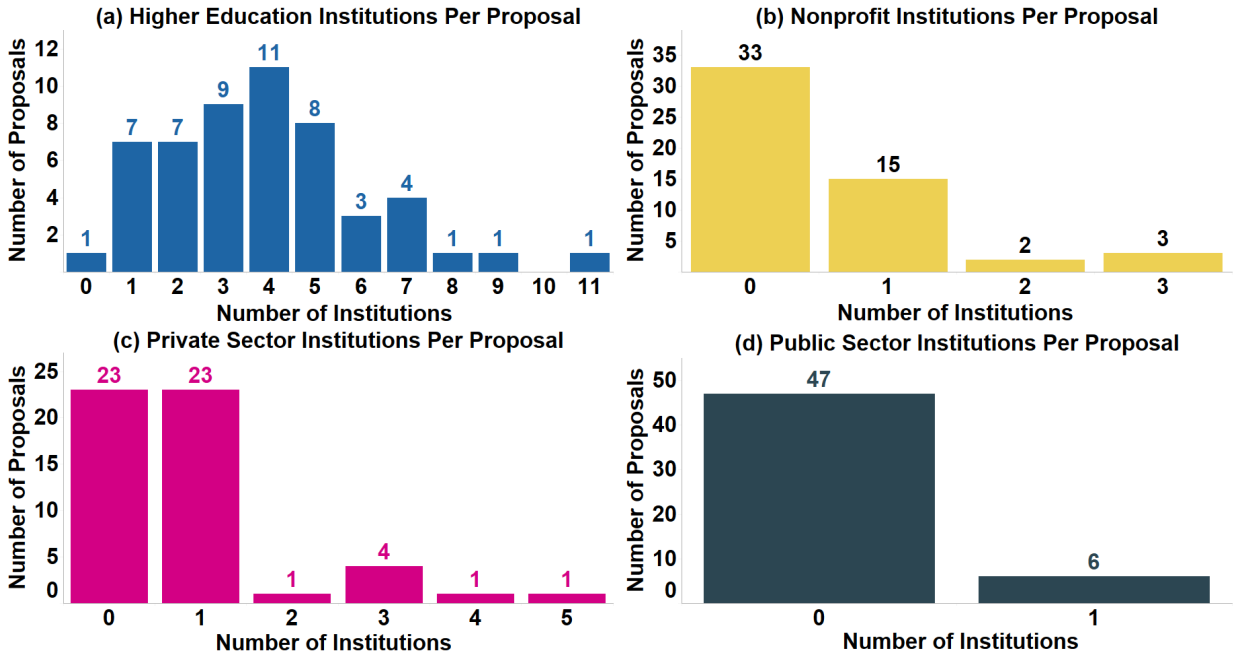
Figure 5.3: Number of Institutions per Competitively Reviewed Phase 1 Proposal from 2019 to 2021 by Institutional Sector



Source: NSF administrative data from Solr (*budget* field), accessed 12/1/2021; and Report SQL (*perf_org_txt* field), accessed 5/18/2022.
Note: This figure illustrates the frequency of higher education, nonprofit, private, and public institutions included in Phase 1 proposals for the 2019 to 2021 cohorts. The sample size is 333 Phase 1 proposals with budget tables, and the unit of analysis is competitively reviewed Phase 1 proposals. Higher education participated in 95% (317 of 333) of all Phase 1 proposals with budget tables.

Figure 5.4 shows the frequency of higher education, nonprofit, private, and public institutions included in Phase 2 proposals from the 2019 and 2020 cohorts. About 98% (52 out of 53) of Phase 2 proposals included a higher education institution and 11% (647 out of 53 proposals) included a public sector institution. Around 57% of Phase 2 proposals (30 out of 53) included at least one private sector institution, while about 38% (20 out of 53 proposals) included at least one nonprofit institution. Thirty-seven percent of Phase 2 proposals (20 out of 53) have *both* a nonprofit and private sector partner.

Figure 5.4: Number of Institutions per Competitively Reviewed Phase 2 Proposal from 2019 to 2020 by Institutional Sector



Source: NSF administrative data from Solr (*budget* field), accessed 12/1/2021; and Report SQL (*perf_org_txt* field), accessed 5/18/2022.
Note: This figure illustrates the frequency of higher education, nonprofit, private, and public institutions included in Phase 2 proposals for the 2019 and 2020 cohorts. The sample size is 53 competitively reviewed Phase 2 proposals with budget tables, and the unit of analysis is competitively reviewed Phase 2 proposals. Of Phase 2 proposals with budget tables, about 98% (52 out of 53) of Phase 2 proposals included a higher education institution.

One definition of a grantee partnership is a multi-organizational project that has at least one non-academic partner. Table 5.14 shows the percentage of all submitted proposals, by phase and track, that are grantee partnership projects. Across Tracks A to F, 48% of Phase 1 proposals are grantee partnership projects. For Phase 1, the percentage of proposals classified as grantee partnerships ranges from Track A's low around 33% of proposals (16 out of 49 Phase 1 proposals to Track A) to Track C's high around 64% of proposals (23 out of 36 Phase 1 proposals to Track C).

For Phase 2, teams are encouraged to partner with other institutions and are required to involve stakeholders beyond the lead institution. As a result, the expectation is an increase in grantee partnerships. From Table 5.14, the percentage of proposals classified as grantee partnerships increases from Phase 1 to Phase 2, within-each track and overall. In Phase 2, proposals classified as grantee partnerships ranges from Track A's low around 67% of proposals (12 out of 21 Phase 2 proposals to Track A) to Track C's high of 90% of proposals (9 out of 10 Phase 2 proposals to Track C). Across Tracks A to D, nearly 76% of Phase 2 proposals are grantee partnership projects.

Table 5.14: Percent of Grantee Partnerships in Competitively Reviewed CA Proposals from 2019 to 2021 by Phase and Track

Track	Phase 1	Phase 2
All Tracks	47.7% (159)	75.5% (40)
Track A	32.7% (16)	66.7% (12)
Track B	39.2% (20)	73.7% (14)
Track C	63.9% (23)	90.0% (9)
Track D	44.8% (26)	83.3% (5)
Track E	59.2% (58)	N/A
Track F	39.0% (16)	N/A

Source: NSF administrative data from Solr (*budget* field), accessed 12/1/2021; and Report SQL (*perf_org_txt* field), accessed 5/18/2022.
Note: This table shows the percentage of all submitted proposals, by phase and track, that we classified as multi-institutional with at least one non-academic partner based on the institutions proposed in that proposal's budget tables. Phase 1 data are for 2019 to 2021 cohorts; Phase 2 data are for 2019 and 2020 cohorts. The sample size is 386 proposals, and the unit of analysis is a competitively reviewed proposal. For Phase 1, the percentage of proposals classified as grantee partnerships ranges from Track A's low around 33% of proposals (16 out of 49 Phase 1 proposals to Track A) to Track C's high around 64% of proposals (23 out of 36 Phase 1 proposals to Track C). Within each track the percentage of proposals classified as grantee partnerships increases from Phase 1 to Phase 2.

Next, we examined the participation of project personnel across different kinds of institutions using budget table data. Table 5.15 shows the percentage and count of named senior personnel within each leadership role that are proposed on the budget tables of institutions in the higher education, private, nonprofit, and public sectors for proposals from 2019 to 2021.[28]

[28] In cases in which a partner institution plans to include personnel from a subcontracted institution on their team, the personnel on each budget table can include both the direct employees of that institution and any employees whose institution would be subcontracted to the institution submitting the budget table. Therefore, an individual may be proposed on the budget table of one institution but list a different institution as their current place of employment on their biosketch if one was included in the proposal. For example, if Nonprofit International, which is not the lead institution on its proposal, proposed to work with an individual from Private Company, Inc. by providing this institution a portion of the funds for Nonprofit International, the individual from Private Company, Inc. would appear on the budget table for Nonprofit International. Since direct employees and subcontracted employees are not differentiated in the budget tables, the individual from Private Company, Inc. would be counted in the nonprofit sector, rather than the private sector, as they are on the budget table for a nonprofit institution. We did not conduct an analysis comparing the institutional affiliations for individuals whose biosketch was linked to a named senior personnel from the budget tables, so the prevalence of these cases is not reported.

Table 5.15: Leadership Roles for Competitively Reviewed CA Proposals from 2019 to 2021 by Institutional Sector

Leadership Role	Higher Ed.	Nonprofit	Private Sector	Public Sector
PI	82.9% (383)	6.3% (29)	9.3% (43)	1.5% (7)
Co-PI	82.9% (704)	6.4% (54)	9.9% (84)	0.8% (7)
Faculty	98.0% (300)	2.0% (6)	0.0% (0)	0.0% (0)
Other Senior Assoc.	69.2% (457)	7.6% (50)	19.7% (130)	3.5% (23)

Source: NSF administrative data from Solr (*budget* field), accessed 12/1/2021; and Report SQL (*perf_org_txt* field), accessed 5/18/2022.
Note: This table shows the percentage and count of CA senior personnel within each leadership role that are proposed on the budget tables of institutions in the higher education, private, nonprofit, and public sectors for proposals from 2019 to 2021. The sample size is 368 proposals with budget tables that list leadership roles for senior personnel, and the unit of analysis is people, as observed on budget tables on competitively reviewed proposals. Of senior personnel in the PI leadership role, slightly less than 83% (383 out of 462 PIs) were proposed on budget tables for higher education institutions.

Slightly fewer than 83% (383 out of 462 PIs) of senior personnel were proposed on budget tables from higher education institutions. Compared to the different leadership roles examined, the “other senior associate” leadership role was the least likely to be on the budget tables from higher education institutions (about 70%, or 457 out of 660 senior personnel).

Table 5.16 shows the count of people proposed on budget tables for higher education institutions by the institution’s Carnegie classification.[29] Of CA project personnel proposed on budget tables from higher education institutions, 82% (4,758 out of 5,803 people proposed on budget tables for higher education institutions) were proposed on budget tables for institutions classified as Very High Research Activity institutions.[30] Around 13% (779 out of 5,803 people proposed on budget tables for higher education institutions) of project personnel were proposed on budget tables for institutions classified as High Research Activity institutions. The remaining 5% (266 out of 5,803 people proposed on budget tables for higher education institutions) were proposed on budget tables for higher education institutions that fall into other Carnegie classifications.

[29] As with Table 5.16, analyzing the sector of institutions that submitted budget tables, the same limitation of direct and subcontracted employees appearing on the same budget table applies to this analysis.
[30] As of September 2022, there were 146 institutions in this category.

Table 5.16: Personnel Proposed to Higher Education Institutions by Carnegie Classifications from 2019 to 2021

Carnegie Classification	Count	Percent
Doctoral Universities - Very high research activity	4758	82.0%
Doctoral Universities - High research activity	779	13.4%
Doctoral/Professional Universities	58	1.0%
Master's Colleges and Universities - Larger programs	72	1.2%
Master's Colleges and Universities - Medium programs	26	0.4%
Special Focus Four Year: Research Institution	41	0.7%
Other Classifications	69	1.2%

Source: NSF administrative data from Solr (*budget* field), accessed 12/1/2021; and Report SQL (*perf_org_txt* field), accessed 5/18/2022.
Note: This table shows the count of people proposed on budget tables for higher education institutions by the Carnegie classification of that institution. The sample size is 365 proposals from 2019 to 2021 with budget tables for institutions that could be matched to entries in the Carnegie Classification, and the unit of analysis is people, as observed on budget tables. Of CA project personnel proposed on budget tables for higher education institutions, around 82% (4,758 out of 5,803 people proposed on budget tables for higher education institutions) were proposed on budget tables for institutions classified as Very High Research Activity institutions.

Overall, Phase 1 proposals commonly featured a multi-institutional partnership with at least one non-academic partner with an average of 48% (across tracks A to F) and this formation of partnerships increased in Phase 2 to about 76% (across tracks A to D). Additionally, senior personnel were mostly from the higher education sector. For those from the higher education sector, most were from research-intensive higher education institutions. Additionally, the private sector did participate in 33% of Phase 1 proposals. Further, the private and nonprofit sectors increased their representation in proposals in Phase 2. Specifically, 57% of Phase 2 proposals included at least one private sector institution and 38% included at least one nonprofit institution.

5.5 Biosketch Analyses

We examined the individual personnel included in NSF proposals using data obtained from proposal biosketches, which provide details about the educational and professional backgrounds of many, but not all, personnel proposed to CA projects.[31] These details on personnel backgrounds provide insight into the interdisciplinary composition of personnel by tracks and their institutional background.

Table 5.17 presents the principal disciplines as reported by proposed project personnel on their biosketches. Across all tracks, between about 72% and 92% of degrees earned by proposed project personnel on awarded proposals were in science, technology, engineering, and mathematics (STEM) disciplines. Across all tracks between 70% and

[31] See the discussion on the coverage of biosketch data in Section 4.

91% of degrees earned on declined proposals were in STEM disciplines, suggesting no distinction in the prevalence of STEM disciplines between awarded or declined status.

Additionally, the results indicate the prominence of some disciplines due to the domain addressed by the track. For example, proposals to Track C, Quantum Technology, are heavily populated by personnel with the Physical Sciences as their background and proposals to Track E, Networked Blue Economy, have Life Sciences as their background.

Table 5.17: Discipline of STEM Degrees Earned by Proposed Project Personnel in Submitted Biosketches from 2019 to 2021 by Track and Award Status

Degree Discipline	A	B	C	D	E	F
Awarded All STEM	87.8%	76.4%	90.6%	92.0%	91.6%	71.8%
Engineering	25.5%	25.4%	20.5%	32.4%	19.0%	9.8%
Computer Sciences	16.2%	21.4%	12.2%	21.7%	6.8%	26.3%
Physical Sciences	4.5%	2.8%	41.4%	2.6%	2.8%	1.6%
Life Sciences	13.8%	5.5%	3.2%	22.0%	33.7%	5.1%
Other STEM	27.7%	21.2%	13.3%	13.3%	29.3%	29.0%
Declined All STEM	90.1%	80.4%	91.0%	89.9%	88.5%	69.8%
Engineering	21.4%	23.2%	18.4%	24.4%	28.1%	14.6%
Computer Sciences	25.0%	21.1%	11.0%	23.2%	6.8%	21.2%
Physical Sciences	7.5%	3.6%	49.4%	6.4%	5.1%	1.8%
Life Sciences	10.7%	7.3%	3.9%	16.1%	21.2%	8.9%
Other STEM	25.5%	25.2%	8.3%	19.0%	27.3%	23.2%

Source: NSF administrative data from Solr (*bio* and *status* fields); accessed 12/1/2021.
Note: This table shows the percent of all degrees earned by proposed project personnel in select STEM disciplines by track and award status using the NCSES Taxonomy of Disciplines hierarchy Level Two. The sample size is 389 proposals from 2019 to 2021 with biosketches featuring degree(s) earned, and the unit of analysis is degrees earned, as observed on biosketches submitted with competitively reviewed proposals.

We also examined where personnel are employed to understand the mix of institutions participating in the program. Table 5.18 shows the employment background of proposed project personnel in Phase 1 (2019 to 2021 cohorts) and Phase 2 (2019 and 2020 cohorts) by institutional sector based on their job at the time of proposal. From the biosketch analysis, we identified 2,139 current appointments of project personnel in Phase 1 proposals and 588 current appointments of project personnel in Phase 2 proposals.

Table 5.18: Current Institutional Sector of Proposed Project Personnel in Submitted Biosketches by Phase

Sector	Phase 1	Phase 2
Higher Ed.	89.3% (1910)	88.9% (523)
Nonprofit	2.4% (52)	2.6% (15)
Private	6.1% (130)	7.0% (41)
Public	2.2% (47)	1.5% (9)

Source: NSF administrative data from Solr (*bio* field); accessed 12/1/2021.
Note: This table shows the count of proposed project personnel in each institutional sector based on their job at the time of proposal submission by phase. Phase 1 data are for 2019 to 2021 cohorts; Phase 2 data are for 2019 and 2020 cohorts. The sample size is 384 proposals with biosketches featuring current roles, and the unit of analysis is people, as observed on biosketches submitted with competitively reviewed proposals. Of the project personnel's appointments at the time of Phase 1 proposal submission, slightly more than 89% (1,910 out of 2,139 total Phase 1 current appointments) were held in the higher education sector.

In Phase 1 proposals, slightly more than 89% of the current appointments identified in biosketches (1,910 out of 2,139 total Phase 1 current appointments) are in the higher education sector. About 6% (130 out of 2,139 total Phase 1 current appointments) are in the private sector. The sector distribution for current appointments is very similar between the phases. For Phase 2, slightly less than 89% of the current appointments (523 out of 589 total Phase 2 current appointments) are held in the higher education sector, while less than 2% (9 out of 589 total Phase 2 current appointments) are in the public sector.[32]

Table 5.19 shows the distribution of institutional sectors for project personnel's current employing institutions by track. The percentage of project personnel with employing institutions in the higher education sector varies from a low around 83% of project personnel for Track E (544 out of 652 total Track E project personnel) to a high around 93% of project personnel for Track F (217 out of 233 total Track F project personnel). In Track E, more than 5% of project personnel (35 out of 652 total Track E project personnel with current appointments) have current appointments at institutions in the nonprofit sector, and slightly less than 9% have current appointments at private sector institutions (57 out of 652 total Track E project personnel with current appointments).

[32] This further demonstrates the participation of the higher education sector in CA program proposals across every metric: with higher education personnel participating at a higher rate, 89.3%, than higher education institutions at 74.8% (as noted in Table 5.13) in Phase 1 proposals.

Table 5.19: Current Institutional Sector of Proposed Project Personnel in Submitted Biosketches from 2019 to 2021 by Track

Sector	Track A	Track B	Track C	Track D	Track E	Track F
Higher Ed.	91.1% (410)	91.2% (414)	91.6% (392)	89.6% (466)	83.4% (544)	93.1% (217)
Nonprofit	0.7% (3)	2.6% (12)	0.7% (3)	1.3% (7)	5.4% (35)	3.0% (7)
Private	4.2% (19)	5.9% (27)	5.8% (25)	6.9% (36)	8.7% (57)	3.0% (7)
Public	4.0% (18)	0.2% (1)	1.9% (8)	2.1% (11)	2.5% (16)	0.9% (2)

Source: NSF administrative data from Solr (*bio* field); accessed 12/1/2021.
Note: This table shows a comparison of the sector of project personnel with biosketches' current employing institution by track. The sample size is 384 Phase 1 and Phase 2 proposals with biosketches listing current appointment, and the unit of analysis is people, as observed on biosketches submitted with competitively reviewed proposals. The percentage of project personnel with employing institutions in the higher education sector varies from a low around 83% of project personnel for Track E (544 out of 652 total Track E project personnel with current appointments listed in biosketches) to a high around 93% of project personnel for Track F (217 out of 233 total Track F project personnel with current appointments listed in biosketches).

Additionally, we analyzed the occupation titles of each current higher education appointment into standardized university occupations to describe the academic personnel involved in CA proposals. Table 5.20 presents the counts of current university occupations for proposed project personnel working in higher education. Of project personnel with current occupations at higher education institutions, nearly half (about 49%, or 1,162 out of 2,388 total) were full, associate, or assistant professors. Of project personnel with current occupations at higher education institutions, about 20% (471 out of 2,388 total people) were researchers.

Table 5.20: University Occupations at the Time of Competitively Reviewed CA Proposal Submission from 2019 to 2021

University Occupation	Count	Percent
Full Professor	447	18.7%
Associate Professor	322	13.5%
Assistant Professor	393	16.5%
Researcher	471	19.7%
Administrator	476	19.9%
Other	279	11.7%

Source: NSF administrative data from Solr (*bio* field); accessed 12/1/2021.
Note: This table presents the counts of current university occupation classifications for proposed project personnel working in higher education. The sample size is 381 proposals with biosketches with current appointments in the higher education sector, and the unit of analysis is people, as observed on biosketches submitted with competitively reviewed proposals. Of project personnel with current occupations at higher education institutions, around 49% (1,162 out of 2,388 total people) were full, associate, or assistant professors.

Table 5.21 shows the highest degree attained by proposed project personnel who submitted biosketches, for personnel employed in all sectors. Of project personnel whose biosketch lists a degree, slightly less than 86% (2,967 out of 3,467 people) earned a PhD or other terminal degree as their highest degree. Of project personnel whose biosketch lists a degree, slightly more than 10% (356 out of 3,467 people) earned a master's degree as their highest degree. This means that approximately 96% of project personnel whose biosketch lists a degree (3,323 out of 3,467 people) earned a graduate degree.

Table 5.21: Highest Degree Attained by Proposed Project Personnel for Competitively Reviewed CA Proposals from 2019 to 2021

Degree	Count	Percent
PhD	2,967	85.6%
Master's	356	10.3%
Bachelor's	129	3.7%
No Degree or Other (e.g., Associate's)	15	0.4%

Source: NSF administrative data from Solr (*bio* field); accessed 12/1/2021.
Note: This table shows the highest degree attained by proposed project personnel who submitted biosketches. The sample size is 389 proposals, and the unit of analysis is people, as observed on biosketches submitted with competitively reviewed proposals. Of project personnel whose biosketch lists a degree, slightly more than 86% (2,967 out of 3,467 people) earned a PhD or other terminal degree as their highest degree. Of project personnel whose biosketch lists a degree, approximately 96% (3,323 out of 3,467 people) earned a graduate degree.

Table 5.22 shows the median number of years since Phase 1 project personnel with PhDs completed their PhD. As indicated in Table 5.21, about 86% of project personnel (2,967 out of 3,467 individuals with biosketches showing a degree) earned a PhD as their highest degree. Attainment of a PhD typically marks the beginning of an individual's professional career in research, so we used the year a person's PhD was attained as a proxy measure for estimating the length of time a person has spent in the workforce as an expert in their field.

Tracks A, C, and E each have a median of 15 years since PhD completion. Track F has the lowest median of years since PhD completion with 12. The median for PhD holders across all tracks is 14. As a point of reference, the median age at award of doctorate for graduates in the four broad STEM fields of study ranges from 29.6 to 31.1 years of age

(NCSES 2020).[33] This could indicate that Phase 1 project personnel with a PhD have a median age of about 44 years and 14 years of experience in the scientific workforce.[34]

Table 5.22: Median Years since PhD Completion in Competitively Reviewed Phase 1 Proposals from 2019 to 2021 by Track

Track	Median Years since PhD Completion
All Tracks	14
A	15
B	13
C	15
D	13
E	15
F	12

Source: NSF administrative data from Solr (*bio* field); accessed 12/1/2021.
Note: This table shows the median number of years since Phase 1 project personnel with biosketches showing a PhD degree completed their PhD by track for the 2019 to 2021 cohorts. The sample size is 384 proposals with biosketches in which project personnel earned a PhD, and the unit of analysis is years, as observed on biosketches submitted with competitively reviewed proposals. Tracks A, C, and E each have a median of 15 years since PhD completion.

Overall, the biosketches indicate that the personnel proposed for CA projects are mostly from academic institutions with PhDs, typically have at least 10 years of experience since receiving their PhDs, and most specialize in STEM disciplines.

[33] The four broad STEM fields of study are: life sciences, physical sciences and earth sciences, mathematics and computer sciences, engineering.
[34] As a benchmark, the median age at receiving support on a first NIH R01 award for PIs with a PhD degree in 2020 is 41 (NIH 2021, Table 5).

6. Data Lessons Learned

The descriptive statistics in this report were primarily from Solr.[35] We chose Solr because it is a platform that contains over 10 years of NSF proposals in a machine-readable format. We used this platform to extract 391 competitively reviewed CA project proposals submitted to Tracks A through F for both Phase 1 and Phase 2, where applicable. These proposals were extracted as a single JSON file with all of the data from each proposal compressed into a single row.

When examining Solr, we learned five data lessons. First, our analytical sample of competitively reviewed proposals based on data pulls of select program announcement codes under counted the actual number of such proposals for the CA program from 2019 to 2021. Using program element codes or other queries may have improved our sample counts. However, an additional 51 proposals were submitted through a broad agency announcement, which at the time of our analysis were not available on the Solr platform and thus were not included in our analysis. Thus, exact parity was not possible.

Second, we learned linking biosketches to budget table data proved difficult. Even after matching names on biosketches to names in budget tables, we could not discover a way to reconcile the variation in proposed project personnel for whom we had a biosketch. Thus, we decided to present statistics based on biosketches separately from those based on budget tables.

Third, we learned the proposal data of budget tables and biosketches required extensive cleaning. For example, we were interested in analyzing proposal data by CA track and phase, but because these data were not included in the data as a standalone field, we manually collected the data for each proposal. Further, inconsistencies in completeness and terminology used in the proposals extracted from Solr (e.g., leadership roles, section titles) made it difficult to employ automated text analytics techniques, so we conducted significant manual work to parse information in budget tables and biosketches to standardize their content for analysis.

Fourth, we learned that institutional classifications (such as the Carnegie Classification, for example) are not readily available in Solr. To fill this gap, we decided to match the names of higher education institutions on budget tables to those in the Carnegie Classification of Institutions of Higher Education. However, these data additions required significant manual intervention.

Fifth, we learned that disciplinary classifications of proposed team members are not readily available in Solr. To fill this gap, we used NSF's Taxonomy of Disciplines to

[35] Notably, we did use NSF's Report Server SQL to describe the different types of institutions that apply to the CA program and those that received awards. When examining NSF's Report Server SQL, we learned two lessons. First, we learned that self-reported data was readily available for the lead institution of each submitted CA proposal (specifically if the lead institution is a nonprofit, public, private, or higher education institution). Second, we learned manual data collection efforts were required for similarly describing all institutions on a proposal.

generate the descriptive statistics regarding the disciplinary mix across team members, but these data additions required significant manual intervention.

Appendix A: Methodological Details

This appendix describes the procedures SRI used to analyze the data in this report.

Data Preparation

Before we could analyze the selected proposal data collected from NSF's Solr platform, we split the text for both the budget table and biosketch sections using the new line characters included in the optical character recognition (OCR) output of each field such that one row in a table would represent one row of tabular data. We extracted the data for the summary report through this iterative build-up of fields at the line level. We further prepared the data using an initial cleaning process.

This cleaning process included forcing the encoding to UTF-8, excluding non-ASCII characters, forcing all characters to be lowercase, and trimming white space. We replaced rows that did not contain any alphabetical characters (such as those that only contained numerical or punctuation information) with an empty string.

Proposal Metadata

We needed to assign the track and phase associated with each proposal before we could analyze the proposals in the desired manner. After attempting several automated methods for parsing the track from proposal titles and content, we decided to use a fully manual approach to classifying proposals. Because of the large differences across CA tracks and cohorts, it was important to ensure full reliability of the data through a manual method. We conducted a manual review of each proposal to determine the track to which it was submitted. We reviewed the PDF version of each proposal and used the text content of the proposal, submission date, funding code, and other context in the PDF to make track determinations.

We then used the track data to create a cohort field and compared the cohort year against the *Division Director recommend for award date* for each proposal to determine the phase to which each proposal was submitted.

Budget Summaries

Our analysis relies on the senior personnel and other personnel sections and does not use other budget table sections. We began budget table work by splitting the data into individual budget sheets and adding a label for the section to which each line belonged. We used the text preceding the first space of each line to place line numbers within a section. Line numbers in the budget sheets start with the line number of that section, a period, and a space.

Section A of the budget summary covers named and unnamed senior personnel, showing the count, proposed time commitments, and requested funding for these senior personnel. As discussed in Section 4, the proposed senior personnel listed in the budget table are a smaller set of individuals than the group of proposed personnel

considered senior enough to warrant a biosketch. NSF provided instructions on who should be included as proposed senior personnel on budget tables, but those instructions varied from year to year. Section B of the budget summary covers the count, proposed time commitments, and requested funding for non-senior personnel by project role, such as postdoctoral researchers and graduate students.

Proposed project personnel did not always fill the lines of the budget summaries completely or consistently, and sometimes left boxes blank instead of inputting the value 0 for items with no value. This scenario occurred in approximately 7% of lines in the budget summaries (2,119 out of 30,502 total lines). In these cases, the lines of budget summaries contained fewer than four values (for example, the project personnel may have filled in the funds requested values and not the time commitment values). To capture the full information about each line, we manually reviewed each nonstandard row. This process included checking the PDF version of the proposal, finding that budget summary within the PDF, identifying from where the nonstandard value(s) came, and inputting the correct field values, with 0's for spots left blank on the form.

Finally, we identified the line containing the institution described by each budget summary. These institution names were separated by code from the instructional text of the line, using the clean text field.

Biosketches

To begin the biosketch data extraction process, we first split biosketch data for each proposal into individual biosketches and identified the name of the individual whom each biosketch describes. This process involved carefully checking each row for whether it was the beginning of a new biosketch and labeling it as such with the name associated with the biosketch. To further ensure accuracy of labeled data, we conducted a procedure to determine inter-rater reliability. This process is described in Appendix D.

After splitting the biosketches, we identified the sections of each biosketch. As the first step in the process to identify the sections of the biosketch, the list of the most frequent lines of text was exported for labeling to identify those lines that refer to the professional preparation section, appointments section, and products section. Ultimately, we successfully marked with flags 3,467 individual biosketches, indicating the start of the personnel, professional preparation, appointments, and product sections.

We parsed the professional preparation section of the biosketch to extract—for each educational experience—the institution at which the degree was obtained or attempted, the discipline or major of the degree, the degree obtained (where applicable), and the years that educational experience started and ended. We used the extracted information to generate the highest degree listed and the time since their most recent degree.

To identify the type of degree obtained, we prepared a series of keywords for each level of degree—bachelor's, master's, doctoral, and postdoctoral. We prepared a separate list of keywords to capture professional degrees and other degree types that did not fall under one of the four primary levels of degrees. Through this process, we identified a mention of a degree in 11,033 lines of text and could not identify a mention of a degree in 6,541 lines of text.

To identify the discipline from the professional preparation section, we identified the presence of the disciplines by comparing the biosketch text to a list of disciplines from the NCSES Taxonomy of Disciplines (TOD) using a regular expression query. The TOD is based on the 2010 National Center for Education Statistics Classification of Instructional Programs (CIP) (Taxonomy of Disciplines 2021). We identified 9,176 records of disciplines and their associated degrees using regular expression queries; we could not identify 8,398 records of disciplines using regular expression queries, and we used manual labeling to identify the discipline.

We also parsed the appointments section of the biosketches. We obtained the start and end years of each appointment and the institution at which the individual held the job. For appointments listed as current, we calculated the end year to be the year in which the proposal was submitted. To determine the institution, we used an automated extraction method to identify if the appointment text contained the name of a college or university whose name is present in the Carnegie Classification of Higher Education.

Using this approach, we identified the name of the higher education institution in 9,589 lines from the appointment sections of biosketches. We could not match 16,437 lines of appointment text. In 2,369 of these unmatched lines, the automated extraction method detected the word "university" or "college," but the extraction method did not identify a specific institution name. Lines of appointment text that we associated with the name of a higher education institution or in which we detected the word "university" or "college" were classified as being part of the higher education sector. For the remaining lines of appointment text, the institutions were identified using a Named Entity Recognition (NER) model, and the sector to which the institutions belonged was manually classified.

We used a keyword classification method to identify the presence of university roles for lines in the appointment section. Our biosketch appointments analysis focuses on academic appointments at higher education institutions. The university roles we classified for the appointments were: administrators, researchers, full professors, associate professors, assistant professors, adjunct professors, and other university roles.

Once we manually labeled the institution sector classifications for non-higher education institutions, we combined these appointments with the university appointments. From the combined dataset, we set up a process to determine whether each individual biosketch had listed a current appointment. Of the 3,447 individual biosketches, 2,753 had a current role identified, and 694 did not.

Appendix B: Supplemental Details to 5.1 Program Growth

The following tables accompany the descriptive statistics discussion on program growth in Section 5, particularly Table 5.1. Table B.1 provides information on states with lead institutions that did not submit a Phase 1 proposal for the 2019 cohort and the sum of the 2019 to 2021 cohorts. Table B.2 provides information on states with lead institutions that submitted a Phase 1 proposal but did not receive a Phase 1 award for the 2019 cohort and the sum of the 2019 to 2021 cohorts.

Table B.1: States with a Lead Institution that did not Submit a Phase 1 Proposal

No Submission in 2019	No Submission in 2019–2021
Alaska	Alaska
Arkansas	Arkansas
Connecticut	Idaho
Delaware	Minnesota
Hawaii	Montana
Idaho	Nebraska
Iowa	Nevada
Kansas	North Dakota
Louisiana	South Dakota
Minnesota	Utah
Mississippi	Wyoming
Montana	
Nebraska	
Nevada	
New Jersey	
New Mexico	
North Dakota	
Oklahoma	
Rhode Island	
South Dakota	
Utah	
Vermont	
West Virginia	
Wyoming	

Source: NSF administrative data from Solr (*inst_state* and *status* fields); accessed 12/1/2021.
Note: This table lists the states in which no lead institution submitted a Phase 1 proposal to the CA program for the 2019 cohort and the sum of the 2019 to 2021 cohorts. The analysis is based on a sample of 334 Phase 1 proposals and 97 awarded Phase 1 proposals. In the 2019 cohort, about 47% (24 out of 51 total states, which includes Washington D.C.) of states had a lead institution that did not submit a proposal to the CA program. This declined to around 22% (11 out of 51 total states) when examining the 2019 to 2021 cohorts cumulatively. That is, nearly 78% of states had a lead institution that submitted a proposal to the CA program when examining the 2019 to 2021 cohorts cumulatively.

Table B.2: States with a Lead Institution that Submitted a Phase 1 Proposal but did not Receive a Phase 1 Award

Submitted but No Award in 2019	Submitted but No Award in 2019–2021
Alabama	Alabama
Kentucky	Connecticut
Maryland	Delaware
Missouri	Hawaii
New Hampshire	Kansas
Wisconsin	Kentucky
	Mississippi
	Missouri
	New Hampshire
	New Mexico
	Oklahoma
	Rhode Island
	Vermont

Source: NSF administrative data from Solr (*inst_state* and *status* fields); accessed 12/1/2021.
Note: This table lists the states in which a lead institution submitted a Phase 1 proposal to the CA program but did not receive an award for the 2019 cohort and the sum of the 2019 to 2021 cohorts. The analysis is based on a sample of 334 Phase 1 proposals and 97 awarded Phase 1 proposals. For the 2019 cohort, lead institutions in 27 states submitted a Phase 1 proposal. Of these 27 states, lead institutions in 21 states received awards and lead institutions in 6 states (shown above) did not receive an award. Examining the 2019 to 2021 cohorts cumulatively, lead institutions in 40 states submitted a Phase 1 proposal. Of these 40 states, lead institutions in 27 states received awards and lead institutions in 13 states (shown above) did not receive an award. That is, states with proposal submissions increased from about 53% (21 out of 51 total states) to 78% (40 out of 51 total states). Further, states receiving awards increased from 41% (21 out of 51 total states) to almost 53% (27 out of 51 total states).

Appendix C: Quality Assurance

This section outlines the steps taken by the team to read, replicate, and trust the code and data in the summary report clearly and easily. The quality assurance (QA) steps involved in drafting the content of this report include checking the analyses in this report, ensuring that manual processes were standardized and accurate, and checking the code for bugs, issues, and opportunities for improvement.

For this report, the following quality assurance steps were taken: (1) at least a two-person review of every figure included in this document; (2) a thorough copy edit; and (3) a high-level content review by senior staff.

The QA activities related to producing the data used in this report are further detailed below.

Replication

We listed the fields acquired from Solr and the fields we created for analysis in a data dictionary. The following elements were included for each field: field name, definition, and if the field was obtained via Solr or data science. The data dictionary describes only those fields that were obtainable, either because they came from Solr directly or because they were obtained through transformation of data from Solr via data science methods.

All folders in the code repository are numbered in ascending order in which their code should be run. Within each one of these folders, the same numbering process has been followed. As programmers and reviewers named and verified the code in each folder, it was checked to ensure that the numbering matches all other files in the code repository.

Inter-coder Reliability

For the manual labeling tasks associated with this report, the labelers on a given task manually completed independent work that was not influenced by the efforts of the other labelers, if other labelers were involved. Once completed, the work of the labelers for that task was compared to each other. Once a standardized process was determined for the discrepancies, the labelers completed the labeling process again with the new criteria. This process was repeated until all labelers for that task came to the same conclusion on manual elements of the report. This step was very important in ensuring that a manual process can be trusted and repeated at a later date. The use of multiple labelers allowed for the range of potential labeling errors to be identified through this process.

Code Checks

We implemented a standardized system to review all code used to create the deliverables and data products for this report. First, the QA process was standardized across the different R programming files and implemented on the code repository. This

process (1) ensures that all code correctly runs as intended; (2) ensures that all fields used or created in the file match the naming conventions of other programming files; and (3) identifies opportunities to improve the code.[36]

Second, the reviewer ensures that fields created in the new code file match the naming conventions of other fields in this analysis. The lack of consistent naming conventions can also increase the risk of introducing errors in future analysis. This step is completed by comparing all the field names of the exported items against the preexisting names in the data dictionary. If a new field is created and subsequently used in an export for analysis or visualization, that item is added to the data dictionary and given a definition. If the field is new but does not follow a regular naming convention, this is addressed.

Third, the process includes a step about improvement to the code. This step does not find any errors with the code, but rather reviews for accessibility, speed, and reliability. The reviewer of this section uses their expert knowledge to see if the run time of the code, which is always included in each code file, can be improved upon. Additional steps depending on the code can be advised, such as improving loops, making documentation clearer, or identifying ways to split up large processing steps that might be burdensome on certain users' computers with less computing power.

After completing the checklist, a reviewer then identifies any data products or deliverables that use the outputs of the code. This typically takes the form of data tables, figures, and report numbers. The reviewer ensures that all references to this output are accurate. This may include simple calculations and one-to-one comparisons. This step is important to ensure that the data are correctly interpreted and consistently applied throughout all deliverables.

Lastly, the programmer and the reviewer(s) schedule a meeting to walk through the findings. This meeting covers the process, the bugs and issues identified, and the approaches for correcting or improving the code. If no issues were found in the code and no improvements are necessary, this step can take the form of an email verifying that the reviewer has completed the process.

[36]All folders within the code repository were delivered to NSF with an instructional documentation file that includes the purpose of the code and the expected outcome. The reviewer reads these instructions and then executes the code. The first ask is to verify that the code is executing its intended purpose. If it is not, this is addressed in the code.

Appendix D: Key to Acronyms

AI	Artificial Intelligence
BAA	Broad Agency Agreement
CA	Convergence Accelerator
Co-PI	Co-Principal Investigator
DCL	Dear Colleague Letter
FAIR	Findable, Accessible, Interoperable, Reusable
FAQ	Frequently Asked Questions
GCR	Growing Convergence Research
LOI	Letter of Intent
NA	Not Applicable
NCSES	National Center for Science and Engineering Statistics
NER	Named Entity Recognition
NSF	National Science Foundation
OKNs	Open Knowledge Networks
PI	Principal Investigator
QA	Quality Assurance
SQL	Structured Query Language
STEM	Science, Technology, Engineering, and Mathematics
TOD	Taxonomy of Disciplines

References

Carnegie Classifications. N.d. *The Carnegie Classification of Institutions of Higher Education*. Retrieved December 1, 2021 (https://carnegieclassifications.acenet.edu/).

National Center for Science and Engineering Statistics (NCSES). 2021. "Table 27: Median age and age distribution of doctorate recipients, by broad field of study, sex, citizenship status, ethnicity, and race: 2020." *Survey of Earned Doctorates*. Retrieved October 28, 2022 (https://ncses.nsf.gov/pubs/nsf22300/data-tables#group3).

National Center for Science and Engineering Statistics (NCSES). 2022. *Federal Funds for Research and Development: Fiscal Years 2020–21*. NSF 22-323. Retrieved October 24, 2022 (https://ncses.nsf.gov/pubs/nsf22323/).

National Institutes of Health. Office of Extramural Research. 2021. "Table 5: Age at receiving support on a first NIH R01 award for Principal Investigators with a PhD degree according to fiscal year." *Long-Term Trends in the Age of Principal Investigators Supported for the First Time on NIH R01-Equivalent Awards*. Retrieved October 28, 2022 (https://nexus.od.nih.gov/all/2021/11/18/long-term-trends-in-the-age-of-principal-investigators-supported-for-the-first-time-on-nih-r01-awards/).

National Science Foundation (NSF). N.d. *2020 Workshops Reports for the 2021 Topics*. Retrieved December 1, 2021 (https://beta.nsf.gov/funding/initiatives/convergence-accelerator/resources/2020-workshop-reports-2021-topics).

National Science Foundation (NSF). N.d. "Convergence Accelerator Overview." Retrieved September 20, 2022 (https://beta.nsf.gov/funding/initiatives/convergence-accelerator).

National Science Foundation (NSF). N.d. "Frequently Asked Questions (FAQ) for the Convergence Accelerator." Retrieved December 1, 2021 (https://www.nsf.gov/pubs/2019/nsf19056/nsf19056.jsp)

National Science Foundation (NSF). March 15, 2019. "Dear Colleague Letter: NSF Convergence Accelerator Pilot (NSF C-Accel)." Retrieved December 1, 2021 (https://www.nsf.gov/pubs/2019/nsf19050/nsf19050.pdf).

National Science Foundation (NSF). May 11, 2020. *NSF Convergence Accelerator Phase 1 and Phase 2*. Retrieved December 1, 2021 (https://www.nsf.gov/pubs/2020/nsf20565/nsf20565.pdf).

National Science Foundation (NSF). May 13, 2020. *NSF Convergence Accelerator Pilot Phase 2*. Retrieved December 1, 2021 (https://www.nsf.gov/pubs/2020/nsf20555/nsf20555.htm).

National Science Foundation (NSF). May 5, 2021. *NSF Convergence Accelerator Phase 1 and II for the 2021 Cohort*. Retrieved December 1, 2021 (https://www.nsf.gov/pubs/2021/nsf21572/nsf21572.pdf).

National Science Foundation (NSF). October 4, 2021. *Proposal and Award Policies and Procedures Guide*. Retrieved December 1, 2021 (https://www.nsf.gov/pubs/policydocs/pappg22_1/nsf22_1.pdf).

Acknowledgments

This report was prepared for EAC under contract number 49100421D0014 and order number 49100421F0211. The views expressed are those of the authors and should not be attributed to NSF, nor does mention of trade names, commercial products, or organizations imply endorsement of the same by the U.S. Government.

This report benefited from extensive contributions from NSF staff. Christina Freyman, Alex Loewi, Doug Maughan, Robyne McRey, and Taylor Rhodes provided guidance throughout the report preparation process. Staff at SRI International, Center for Innovation Strategy and Policy, not acknowledged elsewhere also assisted with report composition, including Krystal Bouverot, Paul Liu, Kelsey Cooper, Will Ebert, Rayelene Covelle, and Brandon Van Hoff. Michelle Goryn at SRI International provided editorial assistance.

Disclosures

This report was prepared for EAC by SRI International under contract number 49100421D0014 and order number 49100421F0211. The views expressed are those of the authors and should not be attributed to NSF, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Citation

Wyhof, Annie, Michael Lee, Claire Lecornu, Mia Thomas, Shay Moore, Christa Reid, Dylan Solden, and Roland Stephen. 2023. *Convergence Accelerator Summary Report: Descriptive characteristics from 2019 to 2021*. Alexandria, VA: National Science Foundation.