

Advancing Data Science Among the Federal Public Health Workforce: The Data Science Upskilling Program, Centers for Disease Control and Prevention

Mary Catherine P. Bertulfo, MPH; Robert D. Kirkcaldy, MD, MPH; Laura H. Franzke, PhD; Sridhar R. Papagari Sangareddy, PhD; Faisal Reza, PhD

ABSTRACT

Context: Data can guide decision-making to improve the health of communities, but potential for use can only be realized if public health professionals have data science skills. However, not enough public health professionals possess the quantitative data skills to meet growing data science needs, including at the Centers for Disease Control and Prevention (CDC).

Program: The Data Science Upskilling (DSU) program increases data science literacy among staff and fellows working and training at CDC. The DSU program was established in 2019 as a team-based, project-driven, on-the-job applied upskilling program. Learners, within interdisciplinary teams, use curated learning resources to advance their CDC projects. The program has rapidly expanded from upskilling 13 teams of 31 learners during 2019-2020 to upskilling 36 teams of 143 learners during 2022-2023.

Evaluation: All 2022-2023 cohort respondents to the end-of-project survey reported the program increased their data science knowledge. In addition, 90% agreed DSU improved their data science skills, 93% agreed it improved their confidence making data science decisions, and 96% agreed it improved their ability to perform data science work that benefits CDC.

Discussion: DSU is an innovative, inclusive, and successful approach to improving data science literacy at CDC. DSU may serve as an upskilling model for other organizations.

KEY WORDS: applied learning program, data science, inclusivity, interdisciplinary teams, workforce development

Author Affiliations: Public Health Workforce Branch, Division of Workforce Development, National Center for State, Territorial, Local, and Tribal Public Health Infrastructure and Workforce (Ms Bertulfo and Drs Kirkcaldy and Franzke); Office of Science (Dr Papagari Sangareddy); and Informatics and Data Analytics Branch, Immunization Services Division, National Center for Immunization and Respiratory Diseases (Dr Reza), Centers for Disease Control and Prevention, Atlanta, Georgia.

This work was supported by the Center for Disease Control and Prevention's Data Modernization Initiative and the American Rescue Plan.

The authors have indicated they have no potential conflicts of interest to disclose.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (<http://www.JPHMP.com>).

Correspondence: Mary Catherine P. Bertulfo, MPH, National Center for State, Territorial, Local, and Tribal Public Health Infrastructure and Workforce, Centers for Disease Control and Prevention, 2400 Century Center, Atlanta, GA 30341 (ktt8@cdc.gov).

Copyright © 2024 Wolters Kluwer Health, Inc. All rights reserved.

DOI: 10.1097/PHH.0000000000001865

The National Institutes of Health defines data science as “the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex set of data.”¹ Data science requires collaboration across disciplines and sectors, including health care and government, to translate data into information.² The substantial volume, variety, and variability of health data collected by governments create the urgent need for governmental public health to modernize systems and upskill a multidisciplinary workforce in data science.³

In 2019, the US government released the *Federal Data Strategy* and *National Artificial Intelligence Research and Development Strategic Plan* that acknowledged the urgent need for governmental data modernization and outlined a framework to advance public sector data science.^{4,5} US government agencies collect substantial amounts of data from

different sources, and high-quality data are required to guide sound decisions.⁶ However, governments often lag industry in maximizing big data use, with industry reports and journal articles proposing government agencies make strides by prioritizing data modernization to improve decision-making for national challenges.⁷ Big data are anticipated to advance health and health care into the future but require a competent workforce to analyze and process these data.²

The Council for State and Territorial Epidemiologists (CSTE) has recognized the need for public health professionals to build data science skills and recommends the public sector fund data science training for their workforce to mitigate the challenges related to recruiting data scientists, such as higher salaries available in the private sector.⁸ One strategic approach to upskilling public health professionals in data science is to leverage the workplace environment where interdisciplinary teams work together to solve complex public health problems.

Recognizing data science needs and opportunities, the Centers for Disease Control and Prevention (CDC) started the Data Modernization Initiative (DMI) in 2020 to strategically address and fund infrastructure to integrate data science capacity and capability into the agency.⁹ Among CDC's core capabilities include maintaining a diverse workforce and deploying world-class data and analytics.¹⁰

Approach

To meet emerging and complex public health challenges, CDC's workforce must strengthen data science skills, specifically to work with new types of data and sources.¹¹ The workforce component of DMI has a multifaceted approach to address workforce challenges and priorities, including recruitment of data scientists and upskilling the current workforce.⁹

CDC started the Data Science Upskilling (DSU) program in 2019 to address agency data science knowledge gaps. Through DSU, CDC seeks to increase data science capacity and capabilities, sustain the data science workforce, and promote a learning culture. DSU is a team-based, project-driven, applied upskilling program that provides CDC staff with robust foundations in data science. DSU is offered as an online program, where DSU staff facilitate applied learning experiences and provide just-in-time technical advising (TA).

Previous data science training programs at the agency limited participation to certain job series and grades and used in-person training models, which posed barriers for staff who lacked travel funds to participate. DSU's online availability and broad

eligibility reflect the vision of inclusivity from DSU founders (S. Papagari Sangareddy and F. Reza) and their belief that opportunities to learn and apply data science knowledge should be widely accessible to the workforce, regardless of education level, position, job series, or learning style. Because the program is offered online, learners and alumni participate from across the agency's Centers, Institutes, and Offices (CIOs), including international and regional offices.

This report describes DSU's history and current state and includes implications of a curated data science upskilling program and learning community for other federal agencies and state, territorial, local, and tribal (STLT) public health agencies.

Planning for a structured data science program at CDC

In 2018, S. Papagari Sangareddy and F. Reza conducted an internal landscape analysis and found limited public health data science learning programs but strong interest among staff to learn data science. They recruited CDC data science experts to form a Scientific Advisory Board (SAB) that assessed the need for a training program and catalogued internal data science trainings. DSU staff, SAB, and DMI staff developed DSU's mission (ie, *to improve the data science capacity and capabilities of the CDC workforce*) and objective (ie, *to provide an applied learning environment in data science to teams that support high-priority data modernization projects at the agency*).

DSU opted for a team-based approach that allows self-assembled teams of public health professionals with different technical experiences to upskill collaboratively in data science and apply their knowledge to advance their existing or proposed CDC projects. This approach is based on interdisciplinary team-based solutioning, which catalyzes learning and innovation, and postulates that no single professional can solve the complex public health problems that require data science skills and knowledge.^{12,13}

CDC Data Science Upskilling Program

DSU is managed within the Public Health Workforce Branch, Division of Workforce Development (DWD), in the National Center for State, Territorial, Local, and Tribal Public Health Infrastructure and Workforce at CDC. Teams participate part-time during a 10-month period while maintaining their normal work schedules. The applied learning environment allows teams to synthesize new information, learn from failure, and apply new knowledge directly to their daily work. This learning model can be attractive for

supervisors because their staff will directly apply what they learn to an agency priority data science project.

DSU staff recruit applicants using internal agency-wide communication platforms. Interdisciplinary teams of 2 to 5 members apply to DSU with a proposed data science project. Examples of projects include development of a data visualization tool to support global partners' monitoring of public health programs, use of natural language processing to detect health misinformation on social media, and analysis of satellite imagery to improve pedestrian safety. Teams can be newly formed of CDC staff who want to advance a new project or can be teams of CDC staff who are already collaborating on an existing project. Team members can belong to the same organizational unit or belong to different divisions, branches, and centers. An applying team must appoint a team captain, who is responsible for submitting one team application (and if the team is accepted into DSU is responsible for overall project management). Multiple DSU reviewers score applications based on the proposed project's alignment with DSU learning resources and agency priorities and whether the team clearly articulated its learning objectives.

A cohort is the group of teams that begin and advance through the program together. Subject matter experts (SMEs) are CDC staff, contractors, and academic partners who mentor teams regarding their data science learning and projects. SMEs have mentoring experience and technical expertise in a range of data science and public health topics. DSU program staff include a program lead, instructional designers, an evaluator, and program support specialists.

Competency Identification and Program Components

DSU definitions of data science literacy, competencies, and framework were based on the work of Dichev and Dicheva.¹⁴ They defined data science literacy as the ability to “collect, evaluate, analyze and interpret data, present derived results, and take ethically sound action based on them” with a framework comprising 5 data science literacy domains: visualization, statistical, machine learning, computational, and ethical. An internal 2020 gap analysis performed by DSU staff confirmed these domains were crucial to CDC and were adapted into DSU competencies. DSU staff identified additional competencies that include problem-solving, project management, and research methods to collect data and minimize bias.

DSU staff created a flexible and inclusive learning model that includes multiple complementary learning resources to accommodate different learning styles and study habits. The learning resources include

access to massive open online courses (MOOCs), TA, and experiential approaches.

Skills-based learning

Skills-based learning is experiential and provides learners with opportunities to apply new knowledge in a safe environment.¹⁵ Each DSU team develops its own learning and project objectives. Although each team has different objectives, all teams use an open-source programming language (ie, Python or R), and all projects include data cleaning and data wrangling. DSU staff use team objectives to tailor boot camps, MOOCs, and biweekly 2-hour workshops (“DSU Fridays”).

Boot camps are intensive, weeklong, online, live, instructor-led learning opportunities. Boot camps build foundations in open-source programming language skills. MOOCs are self-paced courses accessible at any time. Immediate access to open-source tools and MOOCs provides just-in-time learning while teams work on projects in a fast-paced environment.

Teams are encouraged to discuss their new knowledge, challenges, and solutions with peers, DSU staff, and SMEs during DSU Fridays to promote knowledge sharing. Problem-solving and data visualization are topics explored in every cohort; additional topics are recommended by the learners themselves. DSU Fridays further serve as a learning community for participants to remain connected, peer-motivated, and accountable to their team and projects goals. Learners are exposed to a range of data science topics they can further explore during TA sessions.

Performance support

Performance support includes processes by which SMEs facilitate learning within teams through direct recurring TA. Each team is supported by 2 to 3 SMEs who coach teams through project design and execution. SMEs identify areas where teams require support, including coding and communicating analytic information. SMEs troubleshoot problems in projects and facilitate sustained team progress.

Engaging as a learning community, learners help each other overcome project hurdles and achieve milestones. The learning community, in turn, provides a consistent and supportive environment that draws from the cohort's knowledge and experiences, while building a network within CDC of DSU learners and alumni who are conversant in data science. During the DSU symposium, the culminating end-of-year project-sharing event, teams describe their project, results, data science learning journey, and effects upskilling had on their project. Symposium presentations are the

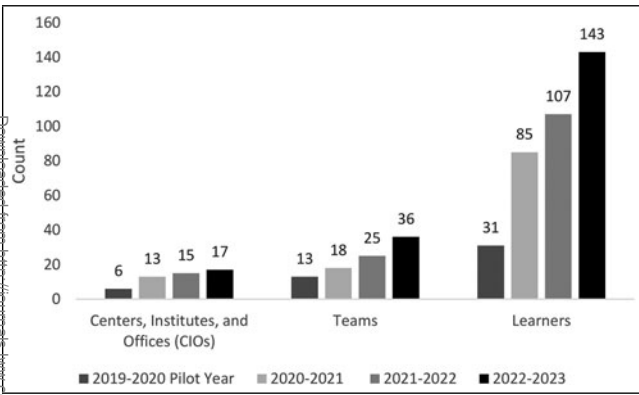


FIGURE Number of Centers, Institutes, and Offices, Teams, and Learners Participating in the Data Science Upskilling Program by Year, Centers for Disease Control and Prevention, 2019-2023

final learning activity of the program; all participants receive a certificate of program completion.

DSU Participation and Evaluation Results

In 2019, DSU piloted the program with 13 teams comprising 49 individual learners; the number of participating CIOs, teams, and learners increased each year (Figure). DSU has increased its acceptance rate from 33% in 2020 to 76% in 2022 to address steady demand. In 2022, DSU enrolled 36 teams and 143 learners, the largest cohort to date (Figure). Overall, DSU has trained 366 learners in 17 CIOs and advanced 92 data science projects.

DSU learners and graduates have a range of professional backgrounds, including economists, health scientists, and geneticists (Table 1). They began the program with a spectrum of data science knowledge (from beginner to advanced), yet none were designated as data scientists. In part, this is attributable to the lack of a federal data science job series, which was not available until 2022. In our experience, the positions of DSU participants have ranged from GS-9 to 15.

Based on end-of-program surveys, substantial proportions of learners reported that DSU increased their data science knowledge, improved their data science knowledge, skills, and confidence in data science decision-making, and improved their ability to do data science work at CDC, which resulted in improved data science literacy outcomes (Table 2; see Supplemental Digital Content Table, available at <http://links.lww.com/JPHMP/B279>).

Lessons Learned

DSU staff identified lessons learned and made programmatic changes based on challenges that they or

TABLE 1
Job Series of Learners in the Data Science Upskilling Program by Cohort, Centers for Disease Control and Prevention, 2021-2023

	2021-2022 (N = 71) ^a	2022-2023 (N = 101) ^a	Total
Biological scientist (0401)			
Computer scientist (1550)	1	3	4
Data scientist (1560)	0	4	4
Economist (0110)	1	0	1
Emergency management (0089)	0	3	3
Geneticist (0440)	0	1	1
General engineer (0801)	0	1	1
General health scientist (0601)	33	37	70
Information technology management (2210)	0	2	2
Industrial hygienist (0690)	1	0	1
Medical officer (0602)	3	1	4
Mathematical statistician (1529)	2	3	5
Microbiologist (0403)	3	5	8
Mining engineer (0880)	0	2	2
Public health program specialist (0685)	2	7	9
Social/Behavioral scientist (0100)	0	2	2
Statistician (1530)	4	2	6
Fellows without an assigned job series	21	25	46

^aThirty-six missing responses for 2021-2022 cohort and 42 missing responses for 2022-2023 cohort.

teams experienced. These changes might have contributed to improved learner satisfaction over time. In 2020, supervisors were allowed to nominate teams and propose projects for DSU. However, the nominated teams might not have wanted to participate in DSU and may have lacked team cohesion, been unclear about their role on the team, or been unfamiliar with the proposed project. Thus, multiple teams dropped out of the cohort. Remaining teams were able to overcome such challenges. But this experience highlighted the wider need for improved project management and leadership skills within teams.

The following year, DSU staff required each team to apply with a designated team captain. Rather than supervisors nominating a team and project, the team captains were required to work with their teams to complete and submit the application, including gathering letters of support from supervisors. The team captains' role has been to manage the project and to lead, motivate, and manage their team during the program. Team captains are mentored by DSU staff to improve these skills, and the DSU learning

TABLE 2
Results of End-of-Year Satisfaction Surveys From Participants in the DSU Program by Cohort, Centers for Disease Control and Prevention, 2019-2023

	2019-2020 (N = 31) ^a	2020-2021 (N = 37) ^a	2021-2022 (N = 71) ^a	2022-2023 (N = 97) ^a
Agreed that DSU increased their data science knowledge ^b	27 (87%)	31 (84%)	71 (100%)	97 (100%)
Agreed that DSU improved data science skills ^b	25 (81%)	30 (81%)	63 (89%)	87 (90%)
Agreed that DSU improved confidence in making data science decisions ^b	24 (77%)	30 (81%)	65 (92%)	90 (93%)
Agreed that DSU improved ability to do data science work at CDC ^b	26 (84%)	31 (84%)	67 (94%)	93 (96%)

Abbreviations: CDC, Centers for Disease Control and Prevention; DSU, Data Science Upskilling.

^aEighteen missing responses for 2019-2020 cohort, 47 missing responses for 2020-2021 cohort, 28 missing responses for 2021-2022 cohort, and 44 missing responses for 2022-2023 cohort.

^bSee Supplemental Digital Content Table (available at <http://links.lww.com/JPHMP/B279>) for wording of questions each year.

environment provides team captains with a safe space to practice leadership and project management skills. Since these changes were implemented, every team accepted into DSU has successfully completed the program.

Discussion

Harnessing evolving data science approaches and infrastructure to address public health threats will require a data-literate public health workforce.⁹ DSU contributes to development of a skilled and diverse data science workforce through CDC staff upskilling, by using a team-based, experiential, and project-driven learning approach that tailors to learner needs. DSU’s rapid expansion is highlighted by the increasing number of learners and participating CIOs. The growing number of learners and graduates contributes to an expanding CDC data science learning community. DSU’s approach advances inclusivity by welcoming learners with varied data science literacy levels, professional backgrounds, and educational degrees.

Program graduates reported increased data science literacy and competency, suggesting that DSU is a successful model for data science upskilling of the federal public health workforce. DSU’s adaptability and emphasis on centering learners’ needs might account for the positive perceptions of the program. Increases in applications highlight the strong demand. Yet, even with growing demand, applicants might not be accepted because of limited program resources; an unmet and highly sought-out need remains.

The skills-based learning and the learning community are major cornerstones of DSU.¹⁶ Knowledge sharing among teams regarding how they overcame problems and team-based learning in a supportive

environment reinforce learning and collaboration.¹² The growing learning community creates a network of data science peers and mentors that can foster continuous learning within the community and holds promise for contributing to a culture of data modernization and learning at CDC.

Online training allows those based in CDC offices outside of Atlanta, Georgia (eg, other US locations or internationally), to access DSU and advance agency priority projects. In-person training can be a barrier for many CIOs because of budgetary, location, and time constraints. The online approach allowed DSU to continue uninterrupted during the COVID-19 pandemic, when many staff worked remotely. The accessibility makes it feasible for DSU to meet the mission of increasing data science capabilities and capacities within CDC and aligns with the agency’s core capability to develop a diverse, skilled workforce and treating data analytics as an asset.¹⁰

CDC is committed to increasing diversity and inclusion in the public health workforce, particularly to expand the talent pool in scientific, technology, engineering, and math (STEM) roles.¹⁷ DSU’s broad eligibility, no-cost participation, and curated learning approach support CDC’s and DWD’s commitment to increasing inclusivity in STEM and workforce development.^{17,18} DSU online availability is accommodating and accessible for many different types of learners working across the globe. DSU learners represent a range of job series and successfully complete the program, demonstrating that data science skills and knowledge are attainable and appropriate for people in many positions at CDC.

DSU is intended to introduce and increase data science literacy. Ideally, the program would like to expand the size of cohorts. However, the number of

Downloaded from <http://journals.lww.com/jphmp> by BhdMf5epHKav1zEoum1tQINa+KJLNEZgpsiH04XMf0hCwvCX1A WnY9pI/QHD3I3D00dRy7TvxSH4C3V/C4/OA/vpDda8KKGKv0Ymy+78= on 01/26/2024

Implications for Policy & Practice

- DSU program contributed substantially to increasing learners' data science awareness, skills, and knowledge, and advancing many crucial public health data science projects.
- DSU and its inclusive, flexible, team-based, project-based, and applied learning-in-place approach can serve as a useful model for other public health entities that are interested in upskilling their workforce in data science or other topic areas.

participants and the extent of training content are currently capped by available resources. Expansion would require additional SMEs with sufficient time and experience to support diverse projects, alumni, and individual learning experiences. Additional workforce development efforts, such as recruiting and retaining of data scientists and providing additional data science training opportunities, might be needed to fully realize the potential for data science at CDC.⁹ A near-term DSU goal is to leverage alumni to mentor current learners and build an agency-wide data science mentoring network.

In 2021, the US General Services Administration featured DSU as a case study to aid other federal agencies in closing the data science skill gaps.¹⁹ Numerous federal agencies subsequently expressed interest in developing a similar program. DSU staff hosted a series of meetings for other federal agencies to learn about DSU competencies, evaluation, TA, and lessons learned. In addition, CSTE piloted the Data Science Team Training (DSTT) program, which is modeled after DSU, and focuses on upskilling professionals working on data science projects at STLT public health departments.^{16,20} DSU and DSTT bidirectionally share lessons learned, including curating high-quality data science training content and conducting the symposium.

References

1. Payne P, Bernstam E, Starren J. Biomedical informatics meets data science: current state and future directions for interaction. *JAMIA Open*. 2018;1(2):136-141.
2. Zhang X, Pérez-Stable E, Bourne P, et al. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn Dis*. 2017;27(2):95-106.

3. Bunnell R, Ryan J, Kent C. CDC Office of Science and CDC Excellence in Science Committee. *AJPH*. 2021;111(8):1489-1496.
4. Office of Management and Budget, the CDO Council, and the General Services Administration. Federal Data Strategy. <https://strategy.data.gov>. Accessed March 3, 2023.
5. Select Committee on Artificial Intelligence of the National Science & Technology Council. The National Artificial Intelligence Research and Development Strategic Plan: 2019 update. <https://catalog.data.gov/dataset/the-national-artificial-intelligence-research-and-development-strategic-plan-2019-update>. Published June 2019. Accessed March 3, 2023.
6. Matheus R, Janssen M, Maheshwari D. Data science empowering the public: data-driven dashboards for transparent and accountable decision-making in smart cities. *Gov Inf Q*. 2020;37(3):101284.
7. Kim GH, Trimi S, Chung JH. Big-data applications in the government sector. *Commun ACM*. 2014;57(3):78-85
8. Hagan C, Holubowich E, Criss T; Council of State and Territorial Epidemiologists. Driving public health in the fast lane: the urgent need for a 21st century data superhighway. <https://resources.cste.org/data-superhighway/mobile/index.html>. Published 2019. Accessed October 13, 2023.
9. Deputy Director for Public Health, Centers for Disease Control and Prevention. Data Modernization Initiative. www.cdc.gov/surveillance/data-modernization/index.html. Published December 15, 2022. Accessed January 9, 2023.
10. Centers for Disease Control and Prevention. Core capabilities. <https://www.cdc.gov/about/strategic-plan/capacity-priority.html>. Published April 29, 2022. Accessed March 3, 2023.
11. DeSalvo K, Wang C, Harris A, Auerbach J, Koo D, O'Carroll P. Public Health 3.0: a call to action for public health to meet the challenges of the 21st century. *Prev Chronic Dis*. 2017;14:E78.
12. Olenick M, Allen L, Smego RA Jr. Interprofessional education: a concept analysis. *Adv Med Educ Pract*. 2010;1:75-84.
13. Goldsmith J, Sun Y, Fried L, Wing J, Miller G, Berhane K. The emergence and future of public health data science. *PHR*. 2021; 42:1604023.
14. Dichev C, Dicheva D. Towards data science literacy. *Procedia Comput Sci*. 2017;108:2151-2160.
15. Dula C, Porter A. Addressing challenges in skills-based education through innovation and collaboration. *Am J Pharm Educ*. 2021; 85(7):8788.
16. Centers for Disease Control and Prevention. Developing state-of-the-art skills. <https://www.cdc.gov/surveillance/data-modernization/snapshot/2022-snapshot/stories/state-of-the-art-skills.html>. Published April 11, 2023. Accessed May 19, 2023.
17. Centers for Disease Control and Prevention. Building an inclusive STEM workforce. <https://www.cdc.gov/stem/workforce/building-stem-workforce.html>. Published July 27, 2021. Accessed March 19, 2023.
18. Centers for Disease Control and Prevention. Division of Workforce Development. <https://www.cdc.gov/csels/dsepd/index.html>. Published July 27, 2021. Accessed March 19, 2023.
19. Office of Management and Budget, the CDO Council, and the General Services Administration. Federal Data Strategy. <https://resources.data.gov/assets/documents/CDOC%20Data%20Skills%20Case%20Studies%20v6.pdf>. Accessed March 19, 2023.
20. Centers for Disease Control and Prevention. Informatics and Data Science Workforce Programs. <https://www.cdc.gov/idswd/index.html#:~:text=Data%20Science%20Team%20Training%20%28DSTT%29%20Program%20CDC%2C%20in,state%2C%20territorial%2C%20local%2C%20and%20tribal%20public%20health%20agencies>. Published February 2, 2023. Accessed March 19, 2023.