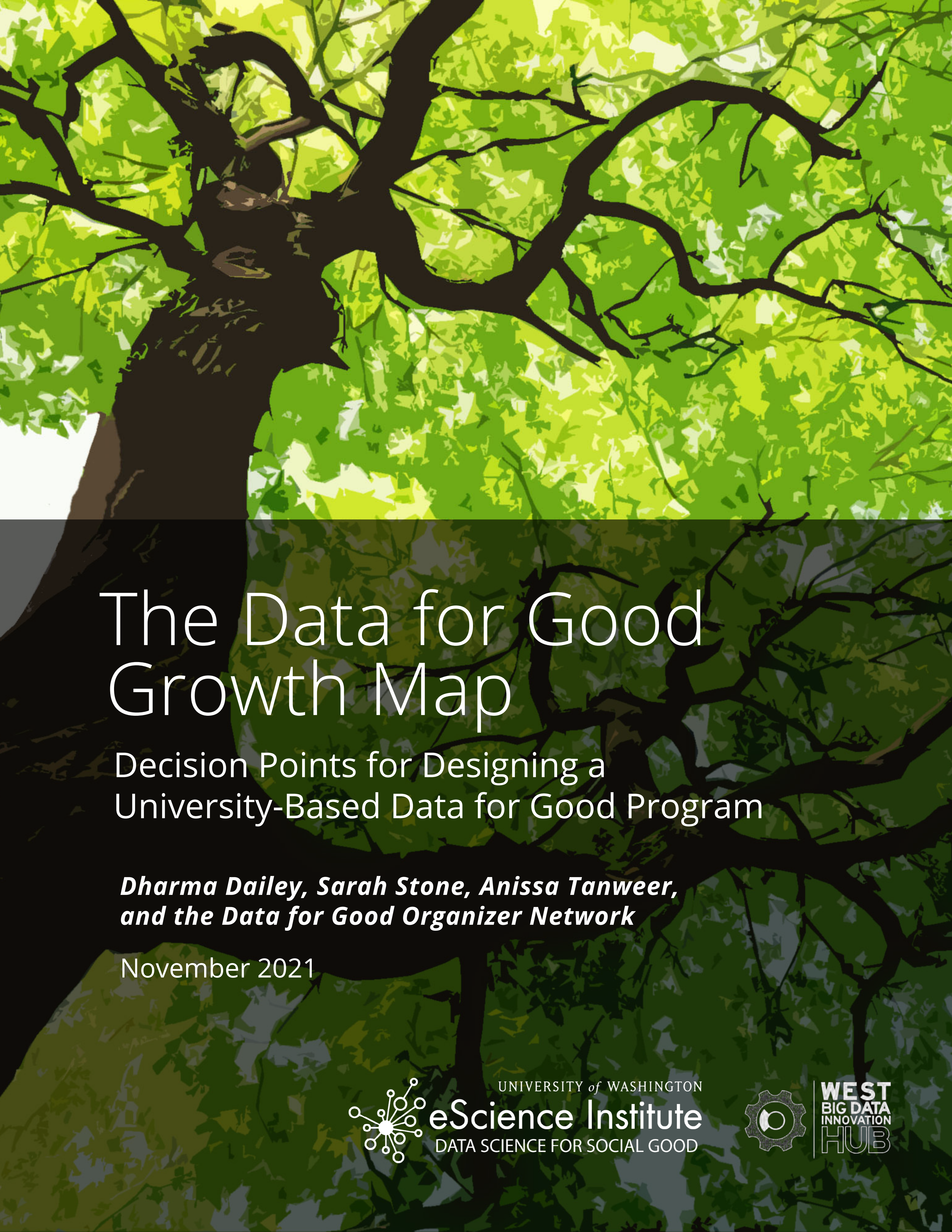


The Data for Good Growth Map

Decision Points for Designing a University-Based Data for Good Program

Dharma Dailey, Sarah Stone, Anissa Tanweer, and the Data for Good Organizer Network

November 2021

UNIVERSITY *of* WASHINGTON
eScience Institute
DATA SCIENCE FOR SOCIAL GOOD

WEST
BIG DATA
INNOVATION
HUB

The Data for Good Growth Map

Decision Points for Designing a University-Based Data for Good Program

Figures

INTRODUCTION

Advancing socially beneficial data science

Increases in the availability of data and computational resources in the 21st century have spawned much innovation in harnessing large, complex, and noisy data to advance knowledge. One important area of innovation in this “data revolution” is the application of data science techniques to questions that have profound implications for both public policy and social practices. Organizations across a range of sectors—including governments, nonprofits, universities, and private companies—are striving to improve their social impact by leveraging data in new ways. Artfully applying data science to social problems entails carefully executing rigorously developed data science methods while attending with care to how data science is integrated into social interventions. Making a positive social impact with data-intensive technologies, therefore, requires partnerships across a range of stakeholders that can provide both technical expertise and nuanced understanding of the social issues and contexts in question. In recent years, a number of such partnerships have emerged through university-based initiatives that we characterize here as Data for Good Programs (D4G). In many cases, these programs were inspired by the first Data Science for Social Good Program, which started at the University of Chicago in 2013. Data for Good programs share a common mandate to educate students through real-world, team-based data science projects intended to positively impact society. Beyond this common premise, however, Data for Good programs vary widely by the types of students served, projects selected, learning experiences provided, program structure, and resources required. In this paper, we explore some of these commonalities and variations, highlighting key decision points that an institution should consider when launching their own Data for Good program.

Common Characteristics of Data for Good Programs

- University-Hosted
- Project & Team Based
- Educate Students
- Intend a Social Impact
- Integrate Stakeholders
- Use Data Science Techniques

Figure 1.

Growing the Data for Good Organizer Network

From July 2020–April 2021, a network of Data for Good program organizers and those doing related work from 17 universities, including nine active Data for Good programs and four in development, met regularly to share their experiences and discuss practices. These Data for Good organizers also participated in a survey that collected detailed information about their programs. For more information about contributors, see the various Appendices and the List of Contributors (p. 62). Aware that many university scholars are considering starting a Data for Good Program to meet the high demand for applied data science education in their own communities, we decided to share what we had learned together. With support from the West Big Data Innovation Hub, a team from the University of Washington's eScience Institute distilled the insights generated through group discussions and survey results to produce a series of "growth maps." Each growth map highlights key decision points to consider when designing a Data for Good program. By elaborating on these high-level decision points, we hope to assist "seedling" programs interested in charting their own plan for growth.

What is "Data for Good"?

In this document, we use the terms "good," "social good," and "social benefit" in an aspirational way. We recognize that social good is not a monolithic or self-evident idea, but rather, it is relational, dynamic, and contested. What is good for some may not be good for all, and who gets to define what counts as good is an inherently political question with implicit power asymmetries. As scientific and cultural norms evolve around what good means, so do our notions of "data for good."

The Data for Good Organizer Network has not adopted a single definition of social good, meaning that the programs represented here emphasize different kinds of social impact. Projects that may be viewed as serving a social good in one program may not be defined that way in another program. We view these differences as a reflection of the lively dialogue currently taking place among researchers and within society about the appropriate application of data-intensive techniques to social concerns. We conduct this work against a backdrop of increasing scrutiny of the role that data-intensive systems play in society. These concerns of researchers and the public include how such systems perpetuate social inequities of race, class, gender, and other faultlines (Eubanks, 2018; Noble, 2018; O'Neil, 2016); the growing discomfort of the public in providing data to these systems (Perrin, 2020); the widening critique of business models that rely upon them (Zuboff, 2019); and ever-louder demands for improved public oversight of them (Engler, 2020; Rainie, Anderson, and Page, 2017).

> ***Our purpose is not to naively apply data-intensive methods to questions of profound social consequence, but to reflexively and systematically explore what it takes to do "good" with data.***

D4G programs that convene interdisciplinary teams of students and faculty alongside government, nonprofit, and industry partners are valuable venues for fostering dialogue and envisioning better practices. Both individually and collectively, these programs are helping to explore and envision data-intensive practices that are scientifically rigorous and socially responsible. We acknowledge that there is much work to do to achieve these admirable goals. Our purpose is not to naively apply data-intensive methods to questions of profound social consequence, but to reflexively and systematically explore what it takes to do “good” with data.

Sectors Leveraging Data in New Ways

Governments

Nonprofits

Universities

Businesses

The wider ecosystem of Data for Good

These university-led, project-based, student-oriented D4G programs are part of a wider ecosystem of interventions helping to fill the high demand for applying data science research and education to social concerns. D4G programs stand among (and complement) other offerings currently seeing an expansion across universities. For example, the curricula of formal data science degrees often feature capstone experiences that give students hands-on data science experience prior to graduation, frequently in partnership with social sector organizations. Another option gaining traction is to integrate data science training into curricula in the social sciences, public policy, and related domains. For example Georgetown University offers a Master of Science in Data Science for Public Policy (MS-DSPP) that prepares students for data-intensive work in the public sector.

In addition to new courses and curricular offerings, other complementary approaches are tailored to supporting applied research in the public sector. For example, the academic-led nonprofit Research4Impact matches researchers with social sector organizations and gives workshops on developing strong research-practitioner partnerships. Such academic-led initiatives are, in turn, part of an even broader ecosystem supporting the integration of data-intensive work into nonprofit and public sector organizations. For example, DataKind partners pro-bono professional data scientists with nonprofit and government organizations, while Data Analysts for Good offers data skills training to social sector professionals.

Among these valuable and varied initiatives, Data for Good programs are distinguished by their trifecta commitment to education, service, and research. As we describe, balancing these missions within a single program takes thoughtful consideration, yet the rewards of this approach are manifold.

How to read this document

We anticipate that the Growth Map in this document will be a reference for programs at different stages of planning. Decision points are illustrated in a series of “Growth Map” diagrams, which provide a visual overview of topics that are further addressed in each section. Note that a number of these decision points are related and mutually reinforcing, and therefore some topics are touched on in multiple Growth Maps. For example, decisions made about “Learning Support” and “Curriculum” should be closely coupled to and informed by a program’s learning goals. For this reason, “Learning Goals” are briefly introduced in the “Learning Support” map and then elaborated on in the “Curriculum” map. Therefore, it may be helpful to first skim all figures before reading the text.

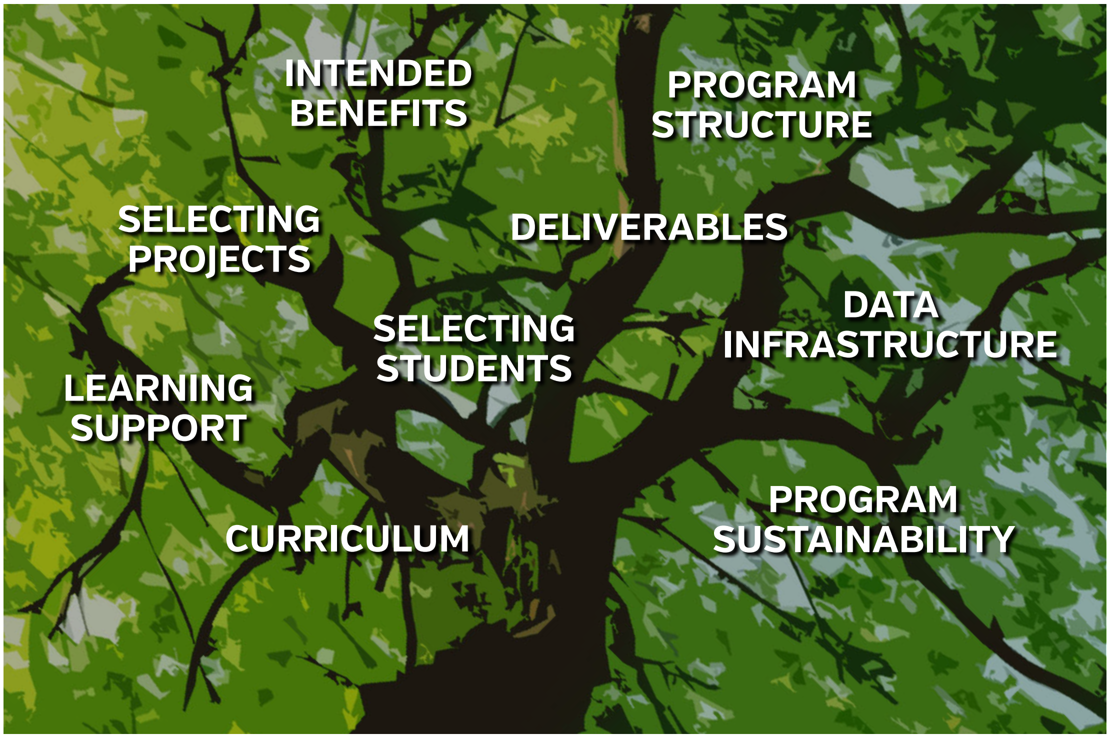
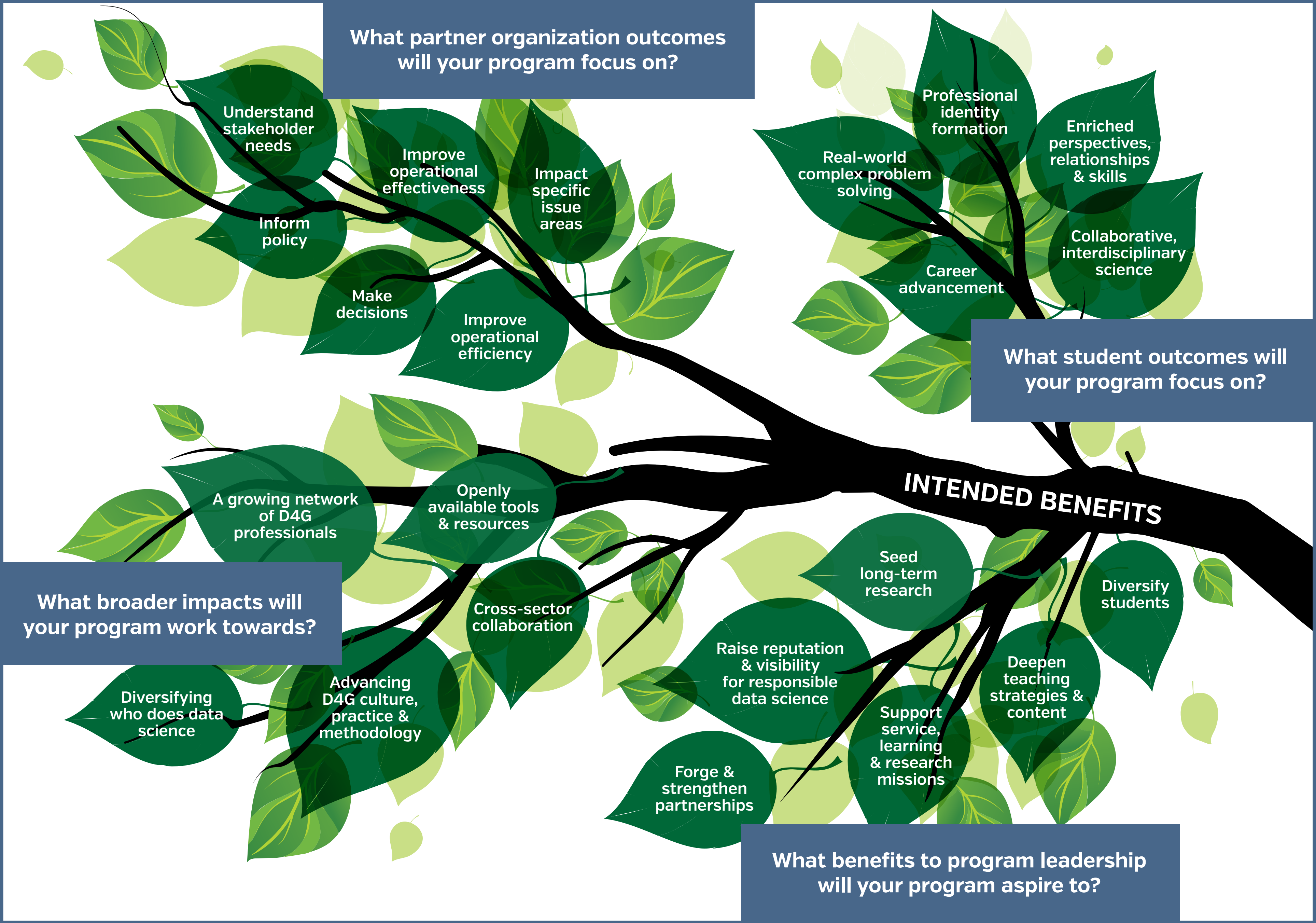


Figure 2. Growth Map Topics (clickable when viewed as PDF)



What partner organization outcomes will your program focus on?

Understand stakeholder needs

Inform policy

Make decisions

Improve operational effectiveness

Impact specific issue areas

Improve operational efficiency

What student outcomes will your program focus on?

Professional identity formation

Real-world complex problem solving

Enriched perspectives, relationships & skills

Career advancement

Collaborative, interdisciplinary science

INTENDED BENEFITS

What broader impacts will your program work towards?

A growing network of D4G professionals

Openly available tools & resources

Cross-sector collaboration

Diversifying who does data science

Advancing D4G culture, practice & methodology

What benefits to program leadership will your program aspire to?

Seed long-term research

Diversify students

Raise reputation & visibility for responsible data science

Deepen teaching strategies & content

Support service, learning & research missions

Forge & strengthen partnerships

Figure 3. Intended Benefits Growth Map

THE INTENDED BENEFITS OF DATA FOR GOOD PROGRAMS

Ideally, Data for Good programs accrue benefits for multiple parties, including students, partner organizations, program mentors, leaders, and broader swaths of society. Here, we position the benefits of D4G programs as “intended” for two reasons: First, we recognize that aligning benefits across multiple parties is always a challenge, particularly when the parties are convening to address a fraught or contested social concern. Second, the benefits discussed below reflect the intentions and rationales for running D4G programs that organizers expressed in our discussions. Though positioned in an aspirational manner, many of the intended benefits listed below are observed outcomes of one or more programs.

What student outcomes will your program focus on?

Data for Good programs train students to use data science approaches on real-world data in close collaboration with partner organizations and stakeholders as part of an interdisciplinary team. This training setting advances multiple learning goals, which are addressed in detail in the “Curriculum” section on p. 35. Here, we provide a higher-level overview of the benefits that students in D4G programs accrue.

The **cross-cutting problems** encountered in D4G programs require students to work at the boundary between academy and community to understand the problem at hand. To achieve satisfactory results, students must collaboratively frame the problem with project partners and negotiate deliverables. This stands in contrast to the more clear-cut kinds of problems students encounter in data science classrooms. Working from beginning to end with an actual client or partner while using real-world data **enriches students’ perspectives, relationships, and skills**. Compared to a more traditional classroom setting, students in a D4G program engage more deeply with data science skills that must be practiced to be learned, such as good documentation, version control, and data cleaning. Likewise, applying data-intensive methods to social concerns affords deeper thinking about the ethical dimensions of data science practice.

The cross-cutting nature of D4G problems requires intensive collaboration among people with different kinds of expertise. Students gain valuable experience doing **collaborative interdisciplinary science** as they integrate different disciplinary approaches into problem solving, and navigate the confusion and disagreements that inevitably arise in cross-disciplinary work. Being exposed to a range of scientific ways of knowing and doing beyond those of their home discipline and the academy helps students cultivate an interdisciplinary perspective on what counts as research. Such interdisciplinary education can shift perceptions of STEM research among students. Students who already view themselves as STEM-capable come to appreciate the value of integrating knowledge and practices from the arts, humanities, and professional fields into applied scientific research. Students who have not considered STEM careers may do so. Likewise, working closely with community partners to address a real-world problem can broaden students' perspectives on what data science can do and provide them with a greater sense of agency. In these ways, D4G programs can be important avenues for **forming professional identities**.

D4G programs are designed to help students grow as professionals. More advanced students step into leadership roles, serving as team leaders or mentors. Students also receive ample opportunities to improve their communication and other professional skills in the course of D4G work. They gain public scholarship experience through program presentation opportunities, experiences that can enable career advancement. In other words, the addition of a D4G project in a student's data science portfolio helps them **launch their career** and get a job. For example, the University of Virginia reported that their students routinely were hired by nonprofit and government project partners. More generally, alumni from the University of Washington program said that D4G projects often became a focal point of job interviews. For other students, D4G programs lead them to extend their academic careers. Many pursue PhDs and other research opportunities after deep exposure to applied research encountered in a D4G program. Additionally, students build professional relationships and networks that extend beyond the program's duration.

Students gain valuable experience doing collaborative interdisciplinary science.

What partner organization outcomes will your program focus on?

Nonprofit and government organizations have a variety of motivations for partnering with universities in D4G programs. D4G projects typically aspire to **impact a specific issue area** within the purview or jurisdiction of their project partners. Analytic tools and protocols vetted by scientists may be integrated into organizational processes for addressing that issue, and new operational processes can be devised. Research results can directly **inform decision making** or assessments related to that particular issue. Thus, successful D4G projects often deepen insights into a particular problem, expand a partner organization's capacity to tackle a problem, or **improve operational efficiency**. For example, the DSSG program at the University of Chicago (now based at Carnegie Mellon University) once worked with the City of Cincinnati to optimize responses by Emergency Medical Services.

Another viable D4G project outcome is to help partner organizations **better understand how they are serving stakeholders** or constituents in order to inform policy. For example, a 2019 project at the University of Washington used Washington State Department of Transportation traffic data, combined with other data sources, to address equity concerns. The goal of this project was to identify whether constituents would be unduly impacted by highway tolls and other transportation management policies. Such a project, if further developed, could

improve operational effectiveness by helping WSDOT better serve their constituents. The project helped WSDOT explore a knowledge gap in public policy that the state legislature had specifically requested they investigate, thereby **informing policy**.

However, program organizers emphasize that some of the most important outcomes for project partners are less immediate or more diffuse. For example, engagement on D4G projects can increase the general capacity of partner organizations to design and implement data-intensive projects. Additionally, Data for Good projects can serve as a vehicle for “project discovery”. One common type of D4G project is a proof-of-concept project that discovers new ways to glean insights from a project partner’s data sources. Such pilot projects generate interest and momentum for data-intensive knowledge production and decision-making. They can help to make arguments for further investment within a partner organization and can lead to larger projects later on. It is important to note that project outcomes can fall along a continuum from project discovery—whereby a D4G program helps an organization to answer the question, “What are ways we can leverage our data to be more efficient or effective?” in an open-ended manner—to those that help an organization to implement or refine particular data-intensive technologies or processes. For example, the University of Warwick program has helped the government of Chile devise computationally-assisted processes to better prioritize environmental complaints.

The close collaborations that take place through D4G programs fostered knowledge exchange between researchers and practitioners that generalized into how partner organizations did data intensive work. In some cases, the sphere of influence spread beyond partner organizations to increase the capacity of a broader group of downstream stakeholders and users. For example, a D4G project focused on making the best use of available data on homelessness for one partner organization in the area around the University of North Florida was the impetus for conversations about doing data analysis, filling “data holes,” and sharing data among several agencies. Thus the project indirectly benefited additional organizations, improving homelessness services in the region.

What broader impacts will your program work towards?

The nature of complex social problems is such that they cannot be solved by a single intervention, much less one that takes place in the 10–14 week timeframe of a typical Data for Good program. Even a highly successful Data for Good project will inevitably offer, at most, a partial solution to a social challenge. D4G programs, therefore, tend to partner with organizations that have deep and sustained engagement with the specific problem space. Sustained partnerships help the short-term, intensive work done in D4G programs be incorporated into larger, long-term efforts. As such, program organizers do not have grandiose expectations that their interventions will change the world overnight. Rather, they recognize that their work is contributing one small piece to a complicated puzzle.

The true impact of partial solutions, as applied to the types of social concerns addressed in any given D4G project, is difficult to assess. Yet, a broader set of potential impacts become apparent when we consider the benefits of running D4G programs for multiple years through many universities.

An aggregate benefit of enabling students to tackle social concerns with data science has been **a growing network of Data for Good professionals.** Providing a pipeline of students capable of supporting social sector data science is one way that D4G programs have contributed to raising the data science capacity of public and nonprofit organizations. Since the first D4G programs initially launched, in-house data science positions at government organizations and NGOs have become commonplace. D4G alumni are frequently hired by government agencies and NGOs, and launch data science consultancies tailored to addressing social concerns. Thus, just as D4G programs help individuals launch their careers, they are playing a role in establishing D4G as a professional field. One result of this professionalization of D4G is the active community building that takes place among program alumni. For example, the Data Science for Social Good program at Carnegie Mellon (formerly at the University of Chicago) maintains an active alumni network of data science professionals across the globe. The ongoing engagement of alumni has encouraged CMU to establish a volunteer network (Solve for Good) to address the overflow of D4G projects proposed through their network.

D4G programs have become sites for intensive **cross-sector collaboration** between partner organizations with deep expertise on a pressing social concern and researchers versed in sound data science methods. Such cross-sector co-investigation is essential to ensuring that interventions do more good than harm. In the aggregate, these co-investigations help to clarify and articulate the techniques, processes, procedures and concepts of most value to data science for social good.

Incrementally, each D4G intervention and its associated public and academic scholarship contribute to greater awareness and dialogue among researchers, students, partners, and the public. Each helps us understand how to navigate the potential benefits of data science applied to social concerns, in light of its great potential harms. In this way, Data for Good programs are an important avenue for both **advancing the culture and practice of doing data science for social good** and **advancing the methodology of Data for Good**. For example, grounded by experience in real-world social and public sector projects, D4G leaders and alumni may bring a different lens to policy and practitioner debates about the use of artificial intelligence. For instance, the Translational Data Analytics Institute at the Ohio State University is encouraging affiliated researchers and students to contribute their data science expertise to policy papers and briefs on AI, thereby helping define the responsible application of data science techniques.

Another important societal benefit of D4G programs resides in their capacity to help **diversify who does data science**. D4G organizers have observed in their programs that the transdisciplinary nature of D4G work, along with its social mission, appeal to groups historically underrepresented in STEM, such as women and racial/ethnic minorities (for more on this phenomenon see "Selecting and Advancing D4G projects" p. 15). For example, the University of Massachusetts Amherst program consistently attracts more women than comparable computer science offerings on their campus. Likewise, D4G programs appeal to a wide range of disciplinary majors with diverse skills sets. Therefore, D4G programs can be platforms for diversifying data science and related fields. By helping to attract and retain students who may not have otherwise considered data science as a career path, D4G programs can help universities meet the demand for data science professionals. At the same time, attracting students with a broader range of lived experiences and intellectual training enriches the field of applied data science (National Academies Press, 2020a; Rawlings-Goss et al., 2018).

> ***Data for Good programs reflect the vision and values of universities and can support a university's service, learning, and research missions.***

Many D4G projects have developed **openly available tools and resources** that enable others to address analogous problems. Open-source code libraries, data repositories, and well-documented data science tools that are advanced in D4G projects can be used by researchers, practitioners, and educators working on the same or similar problems. For example, the Algorithmic Equity Toolkit developed with the ACLU Washington and community groups during the 2019 University of Washington DSSG helps citizens to understand and engage with public policy, thus informing the use of algorithmic technologies by governments (**https://www.aclu-wa.org/AEKit**).

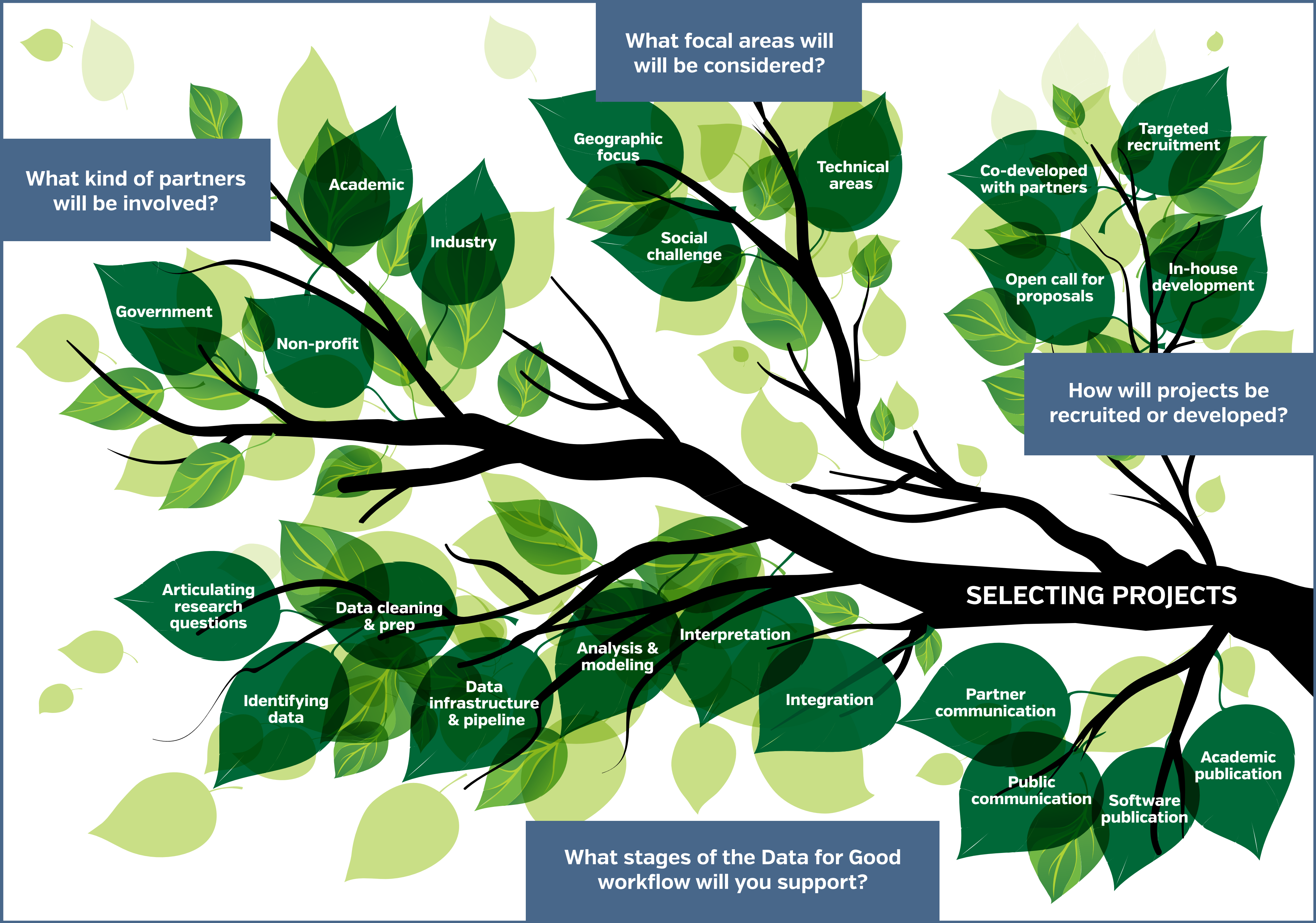
What benefits to program leadership will your program aspire to deliver?

The benefits of running a D4G program may also accrue in the program's leadership and home institution. Data for Good programs reflect the vision and values of universities and can **support a university's service, learning, and research missions.** They can also demonstrate a commitment to civic engagement and responsibly doing data science. For example, the Data Science for Social Good program at the University of Washington was part of the university's application to be recognized as a "Community Engaged Campus," according to the Carnegie Foundation's classification. Data for Good programs can **raise the reputation and visibility of responsible data science** for an institution, highlighting a university's strengths in data science. For example, the University of Warwick found their Data Science for Social Good program to be a valuable recruitment mechanism to attract highly talented doctoral students.

Data for Good programs enhance educational offerings by **deepening teaching strategies and content**. For example, educators at the University of Massachusetts Amherst Data Science for the Common Good program found that D4G projects provide rich, engaging course material to later bring into data science classrooms. That is, the substantive preparatory work required to scaffold specific learning outcomes based on real-world examples within a data science classroom can be fulfilled to some degree by repurposing materials generated in D4G programs. Amherst found course materials created in this manner strongly appealed to students. The University of North Florida's FL-DSSG program was viewed as an innovative service learning model, supporting and inspiring other campus initiatives. D4G programs can also attract and place public sector projects that feed into other campus-based programs. Many D4G programs reported connecting government and nonprofit organizations to other entities on campus. Their solicitations for projects often attracted ideas better suited to project-based courses such as data science masters capstones or classes focused on building data infrastructures. This enriched the types of projects available to students throughout the university.

These benefits similarly enrich the academic unit that leads a D4G program. Because D4G projects appeal to a wide range of students, a D4G program can fulfill a unit's educational mission, helping to **diversify students** that an organization reaches. D4G projects are frequently of interest to the public, university leadership, potential collaborators, potential donors, and the media. Therefore, Data for Good programs increase the visibility of their lead organizations. Service and research missions are advanced by the intensive collaborations that take place through D4G programs. These collaborations can **forge or strengthen partnerships** within and beyond the university and **seed long-term research**. Collaborations may continue long after the program has ended, enabling ongoing service research and the pursuit of long-term funding opportunities.

D4G programs provide a structure for applied research that can elevate the capacity of individual academic project leads and mentors to contribute to social interventions. The intensive cross-disciplinary work that takes place in D4G programs exposes project leads and mentors to new ways of doing data science. Mentors in D4G programs have commented that methods and concepts they encountered when mentoring D4G teams inform their own future work. Likewise, the multidisciplinary nature of the program can bring about alliances that would not have occurred otherwise. Although academic publications are often not the immediate or primary deliverable of a D4G project, they are commonly produced as secondary outputs further downstream; these visible and recognizable outputs are typically viewed favorably within academic communities and can help support the career advancement of academic researchers involved in D4G projects. Additionally, D4G projects sometimes lead to longer-term funding opportunities that can advance and sustain researchers' careers. As sites of experiential learning, D4G programs can deepen mentors' teaching strategies and D4G projects provide engaging examples to incorporate into classroom learning. Finally, project leads and mentors often report personal enjoyment, satisfaction, and a sense of increased perceived prestige from participating in D4G programs.



SELECTING PROJECTS

What kind of partners will be involved?

- Academic
- Industry
- Government
- Non-profit

What focal areas will will be considered?

- Geographic focus
- Technical areas
- Social challenge

How will projects be recruited or developed?

- Targeted recruitment
- Co-developed with partners
- Open call for proposals
- In-house development

What stages of the Data for Good workflow will you support?

- Articulating research questions
- Data cleaning & prep
- Identifying data
- Data infrastructure & pipeline
- Analysis & modeling
- Interpretation
- Integration
- Partner communication
- Public communication
- Software publication
- Academic publication

Figure 4. Selecting Projects Growth Map

SELECTING AND ADVANCING DATA FOR GOOD PROJECTS

The project-based model Data for Good programs are built around providing rich learning, service, and research opportunities. This model incorporates evidence-based best practices in STEM education that are proven to improve student outcomes and broaden STEM participation. The deep integration of knowledge and skills afforded by small-team projects is highly valued by education leaders and employers (National Academies Press, 2020a). Through D4G projects, students acquire the four types of skills identified by data science educators and practitioners as essential to the field: foundational skills, translational skills, ethical skills, and professional skills (National Academies Press, 2018). Through facilitated project work, students apply each of these skill areas to a data science project of social consequence. This prepares them to be capable and responsible professional data science practitioners.

The kind of "experiential learning" that takes place through team-based project work is known to benefit all students, but especially benefits students underrepresented in STEM fields (Theobald et al., 2020). D4G projects enable students to make strong conceptual connections between the practices and procedures of STEM and the goals of applied STEM. Making connections between the "how" and the "why" of STEM has been shown to increase student interest in STEM fields, appealing to those who might otherwise overlook STEM-related careers (Stienberg & Diekman, 2018). This kind of situated learning that addresses real problems has been shown to improve students' confidence in problem solving (Vaz & Quinn, 2014) and learning outcomes (Pomalaza-Ráez & Groff, 2003). Experiences that integrate consideration of the broader context of a STEM problem are especially beneficial to attracting and retaining women and students of color (National Academies Press, 2020b).

"Social good" projects may be especially effective for broadening and deepening the pool of STEM talent because they enable students to exercise "communal values." That is, students can apply STEM concepts and skills to a cause that foregrounds the values of collaboration and helping others. Tying STEM work to communal values generally increases student satisfaction, but is a particularly important strategy for attracting and retaining women in STEM (Belanger et al., 2020; Boucher, Fuesting, Diekman, & Murphy, 2017; National Academy of Sciences, 2020b).

At the same time, D4G projects are service opportunities structured around pressing social problems. They simultaneously support service research at two levels. Individually, each project advances a research question of importance to partners and stakeholders with intimate connections to the issue at hand. Collectively, D4G projects help advance our understanding of the artful and ethical application of data science to the complex interventions that characterize public sector and nonprofit knowledge work. Because the integration of data science into social interventions is still a nascent activity, D4G programs can be viewed as sites of research that refine and advance state-of-the-art for applied data science. D4G programs are less sites where scientists work on behalf of a social benefit organization. Rather, they are where scientists and expert practitioners are co-investigators.

To tap into the rich set of opportunities afforded by a D4G project, it is beneficial to have a clear idea of the type of projects a program will support: What focal areas will be considered? What kind of partners will be involved? What stages of the Data for Good workflow will be supported? How will projects be recruited or developed?

What focal areas will be considered?

The kind of projects a D4G program takes on vary in terms of their **social challenges** addressed, **technical areas** of interest, and **geographic focus**. Some programs chose to narrow their focus to one or more of these areas. For example, the University of British Columbia Data Science for Social Good program has largely focused on “smart city” projects in their region, often repeatedly working with the same project partners (like the City of Surrey, BC) and sometimes working on the same project over multiple years. There are several advantages to focusing a program in this way. Ongoing relationships can reduce program overhead, in terms of vetting and onboarding partners and mentors. The continuity of work and relations made possible by a committed focus enables the kind of buy-in and incremental progress that increases the likelihood that projects will continue and be sustained past an initial exploration or proof-of-concept. Finally, by longitudinally working on local social challenges, a program can become a visible, regional resource for deepening the capacity of local governments and nonprofits to make use of data science.

Projects of Contributing Programs by Thematic Area

Public Health	34
Environment & Natural Resources	18
Transportation	18
Planning & Development	17
Education	13
Governance	10
Employment & Workforce	8
Human Services	8
Public Safety	8
Socio-economic Disparities	7
Homelessness & Housing	6
Incarceration & Criminal Justice	4
Disaster Response	3
Energy	3
Infrastructure	3
Innovation	2
Public Information	2
Philanthropy	1
Other	10
Total	**175**

Figure 5.

Contrasting the local-intensive approach is a more distributed approach. The Data Science for Social Good program at Carnegie Mellon University (formerly at the University of Chicago) has tackled a wide variety of social challenges in locations across the globe. The program’s broad geographic focus and embrace of a wide spectrum of social challenges attracts a wide pool of potential projects, collaborators, and funders, while building D4G expertise and relationships across borders. At the same time, projects have tended to focus on the technical area of predictive analytics, ranging from detecting government fraud, to identifying wild animals from satellite imagery, to predicting educational outcomes for at-risk students.

As shown in the table to the left, contributing Data for Good programs have tackled a wide range of social challenges. Among 175 projects that took place through 2020, more than 19 distinct areas of social concern were addressed. (See Appendix 2 for a list of projects from contributors.)

What kind of partners will be involved?

D4G programs also must decide what kinds of partners they will work with. By partners, we mean persons or entities who are directly involved in defining the scope of the problem to be addressed, formulating a question to be answered, proposing an intervention, contributing data and expertise, or providing material support.

D4G programs frequently team up with **government** and **nonprofit** organizations that have pressing data science challenges. **Industry** and **academic** partners are also sometimes involved. Given that the risks and benefits of partnering on a D4G project vary by sector, our contributors recommend being clear-eyed about how the different incentive structures for each may affect project work and impact. For example, if partnering with an academic researcher, there will likely be pressure to produce peer-reviewed publications. Projects with government partners may be subject to a high degree of public scrutiny, and come with a risk that the project could be scrapped when a new administration is elected. Nonprofit organizations experience high rates of turnover. Industry partners likely have a profit motive, even if it is indirect, and are often in need of burnishing their public image.

Contributing programs were divided on whether they would consider working with industry partners. The University of Warwick viewed partnering with industry as incommensurate with the mission of their program, while the Stanford Data Science for Social Good program saw industry allies to be better positioned as program sponsors than project partners. Other programs report they would consider an industry partner if the project aligned with their internal definition of a social good project. For example, the University of Massachusetts Amherst partnered with for-profit social entrepreneurs to focus on detecting COVID misinformation. The University of British Columbia considered working with industry partners when they already had existing relationships with a government or nonprofit project, as long as the public partner is the main applicant and benefactor of project outcomes.

Projects may occasionally have multiple partners contributing data, expertise, or material support. For example, the Stanford Data Science for Social Good program collaborated with academic experts in the Human Trafficking Data Lab at Stanford and the Brazilian Federal Labor Prosecution Office. This government partner provided data, insights, and context about investigating and prosecuting human traffickers, to identify characteristics of businesses that are more likely to be involved in human trafficking. D4G programs occasionally have even considered funders to be substantively involved partners and not just sources of monetary support. For example, the Data Science for Public Good Young Scholars programs at Virginia Tech, Iowa State University, Oregon State University, and the University of Virginia worked with the U.S. Cooperative Extension Services and the Bill & Melinda Gates Foundation to develop an analytical approach that utilized evidence-based policies and programs that reduce poverty and improve economic mobility by identifying barriers in rural communities. The process was based on the Community Capitals Framework (CCF) and quantifies the assets of rural and urban communities using seven community capitals: natural, cultural, human, social, political, financial and built.

D4G program organizers must be cognizant that this work requires significant investment in time, energy, and resources from their project partners. Project partners must be capable and motivated to do the work that complements the work of D4G students and mentors, including offering expertise and guidance during the program and maintaining the work after the program ends. To assess a potential D4G partner, programs considered whether the organization had the capability to turn the project results into an impact and maintain it over time. When assessing the technical capacity of potential project partners, the University of Massachusetts Amherst program used how much common language was shared between the project partners and program leadership as one gauge for the kind of engagement strategy that would be needed to make the project a success.

Importantly, such demands cannot be made of a partnering organization if the project is not an authentic fit to their needs, resources, and mission. That is, D4G organizers can only expect significant investments in time, energy, and resources from project partners if and when the project is making significant and substantive contributions to their partners' goals.

> *Given the compressed time frame, it is important to consider what aspects of a Data for Good workflow will be supported with the program.*

What stages of the Data for Good workflow will you support?

Given the compressed time frame of a Data for Good program, it is important to consider what aspects of a Data for Good workflow will be supported within the program. We acknowledge that there are many ways of conceiving of an applied data science workflow, and draw upon our experience running Data for Good programs to call attention to several common steps associated with Data for Good projects.

In a **Data for Good Workflow**, we include these steps: articulating a research question; identifying or generating data; data cleaning and preparation; developing a data infrastructure and pipeline; conducting analysis and modeling; interpretation of results; communicating with partners; integrating deliverables into an intervention of some kind; communicating to the public about the project; publishing tools and code repositories; and academic publishing. While every project follows some chronological evolution, these steps tend not to follow in a uniform sequence but instead overlap and repeat. For example, partner communication is often not a single "step" but a process that takes place throughout, and exploratory analysis may reveal that data needs more cleaning than was initially thought. For this reason, we have chosen not to impose a particular order or cycle for the various components of a Data for Good Workflow. However, we recognize that there can be value in further explaining and understanding the rhythm and patterns of a data science practice, and point the reader to other resources that can help make sense of this work. (Alspaugh et al., 2018; Crisan, 2020; Grolemund & Wickham, 2014; Yang et al., 2020; Moreno, González, & Viedma, 2019; Zhang et al., 2020).

Given that most D4G programs correspond to the academic calendar and are only 10-14 weeks in length, it is rarely possible to complete all aspects of a data science workflow during the compressed timeline of a D4G program. Program organizers must therefore decide what work can be tackled by their students and what needs to be accomplished by other means. Students who apply to D4G programs typically are eager to learn canonical data science techniques and apply them to real-world data. Likewise, project partners are eager to see results of these techniques. Therefore, D4G programs tend to give students substantive hands-on experience with data analysis and modeling. Yet a complete Data for Good workflow has many essential steps both preceding and following the analysis and modeling phase of a project. Programs need to consider how each of the essential stages of the work will be achieved. Which will take place within the program? Which will occur prior to or following the program? For those stages of work tackled during the program, what level of support will be offered?

Most program organizers agreed that a research question already needs to be articulated and data needs to be in hand before students begin working on D4G projects. They differ, however, on how they ensure that this is the case. Some collaborate with project partners on these early stages of work, while others make their completion a prerequisite for establishing a partnership. For example, programs at Georgetown University, Iowa State University, the University of Warwick, the University of Massachusetts Amherst, the University of North Florida, the University of Oregon, and the University of Virginia work with potential project partners to articulate research questions and identify data. Working with partners during these early stages has enabled these D4G programs to collaborate with nonprofit and public organizations that may not otherwise be experienced at framing a problem in data science terms or designing a research project. In this way, they successfully expand the pool of people and organizations that can participate in D4G efforts. Yet this commitment to co-design with project partners can be resource intensive, demanding prolonged engagement. Some universities put fewer resources into early stage work, supporting identification of data sources only on occasion (as was the case at the University of British Columbia) or posing that aspect of work as beyond the scope of the program. For example, the University of Washington and Stanford University make clear to potential project partners that having a primary dataset in hand is a prerequisite to being considered for participation in their programs.

Though most D4G projects are designed so that students get ample time doing analysis, modeling, and interpretation, a successful D4G project can also be designed primarily around other phases of the work. For example, one team of students in the University of Washington Data Science for Social Good program spent a significant portion of their project time improving an R package for identifying vote dilution in the U.S. The package was already being used to litigate challenges to the Voting Rights Act, and the team added important functionality, refined the code base, and created tutorials—relatively late-stage activities in the D4G workflow that made the package more accessible to a wider user base.

In theory, any aspect of data science work that fits into the program time frame can be considered. Indeed, an important difference between D4G programs and learning opportunities such as classroom exercises is that students get the opportunity to work with data in a way that is closer to “the real world.” This means that students are exposed to a more complete range of the D4G workflow, and spend time on the crucial steps of cleaning and transforming data in preparation for analysis. Yet students generally most enjoy learning and performing analytic tasks using data science techniques that lead to a result they can share. Thus it is important to anticipate how much prep and transformation will be needed to conduct their analyses, and to take measures to ensure that students won't spend their entire experience in the program cleaning data. In all cases, it is helpful to clearly communicate what aspects of the work will be taken on by students, ideally prior to their commitment to the program.

Consensus considerations for selecting projects

Though D4G organizers have different ways of demarcating what aspects of the data science workflow they consider to be in-scope versus out-of-scope for their programs, we have generally found consensus on several non-negotiable features that render a project tractable. First, due to the academic mission of D4G programs, projects need to be structured around a research question and involve substantive analytical tasks rather than exclusively requiring data engineering. Second, projects that cannot deliver data in hand prior to the beginning of a program are usually non-starters because it is not feasible to both generate data and conduct robust analysis in the compressed timeline of an intensive internship or fellowship program. The exception to this rule are projects that center on project discovery rather than analysis (see "What partner organization outcomes will your program focus on?" p. 10). Third, to increase the likelihood that projects will have a positive social impact, they must demonstrate the **potential for longevity and sustainability**. Perhaps most importantly, it is essential that **ethical concerns have been carefully considered**, and that a **strong rationale for why and how the project will make a positive social impact** is clearly articulated.

How will projects be recruited or developed?

D4G projects come about in different ways. A program may **co-develop** proposals with partners or develop them **in-house**. Others may come about through **open calls** for project proposals. These strategies can be augmented by **targeted promotion** of the program.

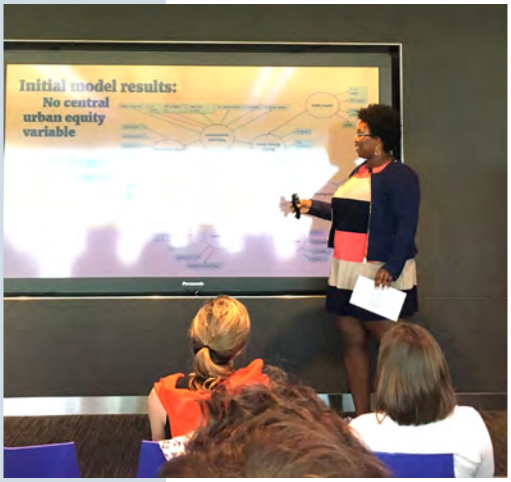
Given the intensive nature of collaboration between D4G programs and their partners, among the Data for Good Organizer Network that contributed to this document, project recruitment strategies tend to be targeted and hands-on. As noted above, (see "What stages of the Data for Good Workflow Will You Support?," p. 19) some programs **co-develop projects** with selected partners, working collaboratively from early stages of ideation and planning. A seemingly less intensive approach is to solicit projects through an **open call for proposals**. The University of Washington recruits projects through this open call approach. However, proposals solicited through an open call still require program leadership and project partners to work together in advance of the program to refine the project plan. It is beneficial to plan time after an open call for possible iteration with candidate project partners, as it will be necessary to review the research question, project design, data permissions and stakeholder engagement together prior to making a final decision on the project's viability.

Situated somewhere between the extremes of co-development and circulating open calls for proposals, several programs reported that they employ **targeted recruitment** to some extent. For example, the University of Warwick has primarily identified projects through an open call for proposals but has also drawn on existing relationships with nonprofit and government organizations by actively reaching out and helping them develop a project idea and scope. Regardless of the manner projects are recruited, however, Warwick's project partners retain ownership of the problem being addressed and drive the project's direction and deliverables. Organizers of the Stanford University program also use targeted recruitment, identifying potential projects by scanning public seminars and publications by Stanford faculty throughout the year with the aim of identifying partners and projects for their summer program. Projects are then further developed by program organizers, a Stanford faculty member, and an external organization. Because faculty and staff are an integral part of project development, program leadership and project leadership sometimes overlap in what may be considered **in-house development**.

COMMON PROGRAM DELIVERABLES

What will you emphasize in terms of deliverables?

Data for Good programs produce a diverse set of deliverables that cater to different aspects of the D4G mission. Common program deliverables serve to advance an intervention, public scholarship, and academic research. Some common deliverables best align with a particular aspect of the D4G mission, such as an academic publication supporting academic research. However, many common D4G deliverables can be tailored to serve project stakeholders, the public, and researchers, depending on the program and project in question. To run projects as smoothly as possible and foster a positive experience for participants, it is important for programs to consider what deliverables to prioritize by program and project.



Common analytic tools and products developed to support a specific intervention may include: the design and implementation of a data infrastructure; visualizations that summarize analyses; an interactive tool such as an interactive dashboard that enables further analysis; and the code repository that is produced in the course of data analysis. Other frequent D4G deliverables are work processes and protocols that guide practitioners through appropriate data science techniques. Equally important to the creation of such analytic tools and work processes is documentation that explains and contextualizes their creation and development. Teams frequently hand off deliverables to project partners in tandem with a written and/or oral report, which generally includes relevant recommendations for advancing the work.

Any given project will include several of these deliverables, tailored as appropriate. For example, when the primary goal of a project is to provide a one-off analysis for a policy question a project partner is debating, project deliverables may look like this: At the conclusion of the program, a team may meet with the project partner to present and review their analysis. The project partners may have methodological questions, and request a link to a code repository and project documentation. The presentation may be accompanied by a report. When the project deliverables include the development of an interactive decision-support tool, this will likely require additional meetings with the IT department, other technical support, and the analysts who will be using the tool. Planning for and prioritizing the right combination of deliverables and hand-off events in this way smooths the path for the partner organization to take full advantage of D4G project work.

In addition to what is shared with project partners, all D4G programs require public-facing outputs from project teams. Programs conclude with a public seminar in which students typically share their work through a brief public talk or a poster session. Students are generally highly motivated by the opportunity to present their work publicly. Because the final presentation raises students' awareness of what they have accomplished, it is often an intellectual and emotional highlight for them. At the same time, D4G presentations can be valuable venues for public science communication, often generating enthusiastic interest from project stakeholders, the media, and members of the public. Several programs have found that videos of final presentations are particularly valuable for demonstrating what has been achieved through the program to funders, sponsors, potential project partners and future students.

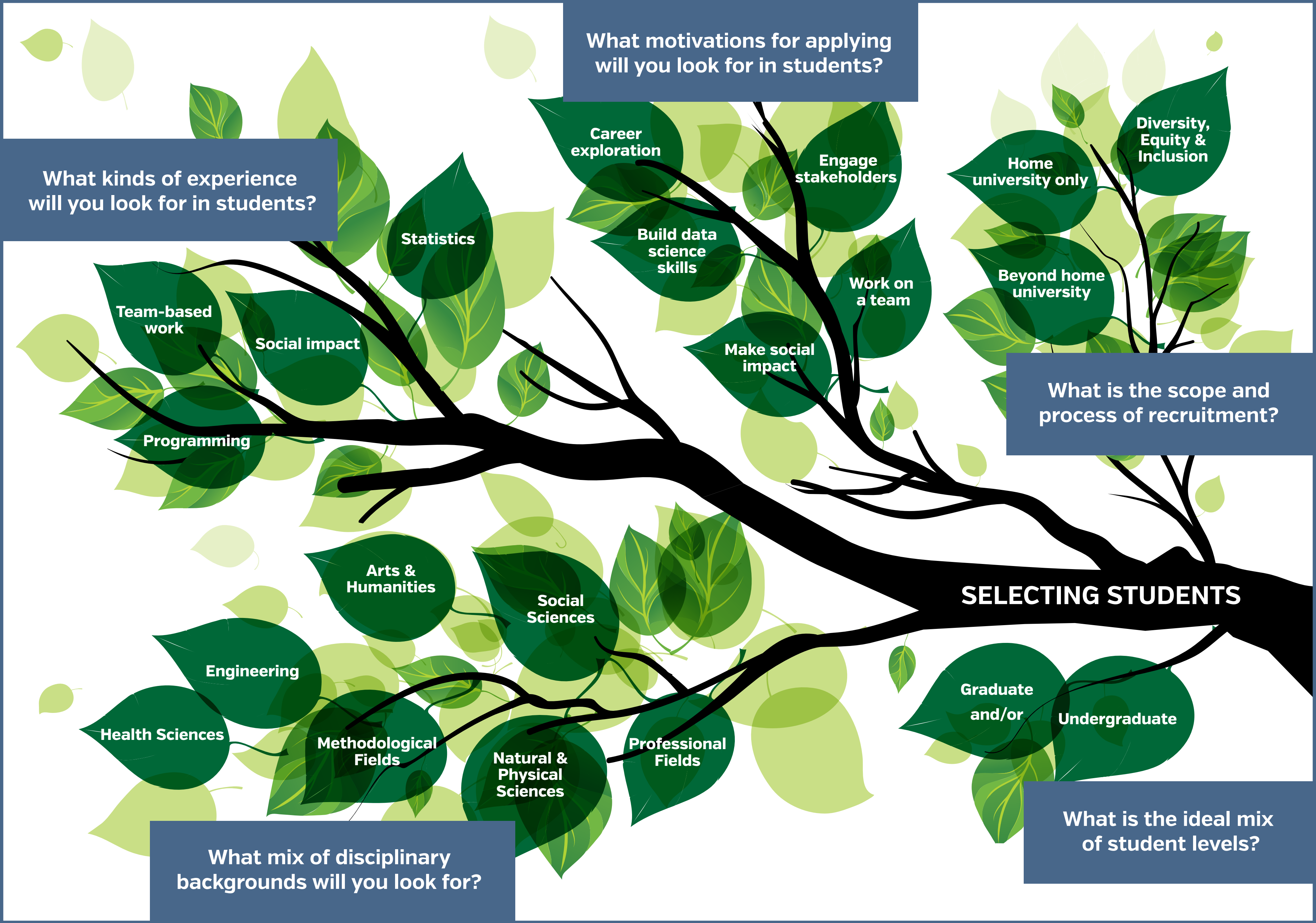
Common Program Deliverables

- Interactive tool
- Visualizations
- Code repository
- Software
- Work processes & protocols
- Documentation
- Report
- Recommendations
- Public seminar
- Project website
- Poster
- Publication

Figure 6.

A D4G project's web presence varies by program and project. Among the programs contributing here, the minimum web presence for a project was a brief project description on a program's webpage, though some programs require a project website and public code repository. In most cases a web presence will include or point to a summary of the project, highlights of a team's analysis; methodology; interactive dashboards; code repositories; posters; recording of public talks; and/or publications. Making these aforementioned materials publically available furthers reproducible science, public service, and student learning.

Inevitably, the compressed timeframe of the program makes it difficult to achieve all of these goals with equal rigor. One common tradeoff in balancing priorities for what the program will achieve is to defer work on academic publications until after the program ends. Programs can support academic publications and presentations after the program concludes. For example, they can provide funds for students to attend conferences and create posters. It is strongly advisable to have conversations about authorship at the outset of a project, even if publications are unlikely to be produced or likely to be produced at a later date. For example, the University of Washington DSSG program explicitly states the expectation that all team members should be included as authors on any future publication based on work they completed during the summer program, even if they were not available to contribute to the activity of writing after the program ended. Such expectations may vary by project and program, but all team members should be aware of what constitutes authorship and how author order will be determined.



SELECTING STUDENTS

What kinds of experience will you look for in students?

Statistics

Team-based work

Social impact

Programming

What motivations for applying will you look for in students?

Career exploration

Engage stakeholders

Build data science skills

Work on a team

Make social impact

What is the scope and process of recruitment?

Home university only

Diversity, Equity & Inclusion

Beyond home university

What mix of disciplinary backgrounds will you look for?

Arts & Humanities

Social Sciences

Engineering

Health Sciences

Methodological Fields

Natural & Physical Sciences

Professional Fields

What is the ideal mix of student levels?

Graduate

and/or

Undergraduate

Figure 7. Selecting Students Growth Map

SELECTING AND SERVING DATA FOR GOOD STUDENTS

An important consideration for each D4G program is determining the kind of students the program will serve. Fortunately, D4G programs have shown that they appeal to a wide range of graduate and undergraduate students with diverse interests and backgrounds. This means that the D4G model can be tailored to meet different educational goals, and can also serve as an effective mechanism for broadening participation in data science (see also, “Selecting and Advancing Data for Good Projects,” p. 15).

D4G program organizers may want to ask themselves: What is the ideal mix of student levels? What mix of disciplinary backgrounds will you look for? What kinds of experience do you look for in students? What motivations for applying will you look for in students? What will be the scope and process of recruitment?

What is the ideal mix of student levels?

D4G programs may choose to serve **graduate students,** undergraduate students or both. Programs contributing to this document selected mainly graduate students, with some **undergraduates**. However, several programs reverse that trend, with teams of undergraduates supported by graduate student mentors. Less common are programs that exclusively serve graduates or undergraduates. Programs may lean toward more strict or flexible guidelines when conceptualizing their ideal mix of student levels. For example, the University of Warwick accepts recent baccalaureate graduates, and occasionally a postdoctoral student. They give more consideration to whether an applicant is likely to fit within a group of students than their precise status designation. Several examples of the mix of student levels at different programs are given in “What roles will support the program” (p. 45).

What mix of disciplinary backgrounds will you look for?

According to the National Academies of Sciences, data science demands a mix of “foundational, translational, ethical, and professional skills” (National Academies Press, 2018). To achieve such a mix for D4G projects, it helps to attract students who represent a breadth of interests and experiences. Thus it is beneficial for programs to consider in advance the level of experience and range of disciplines they want to attract to the program and have work on individual projects.

The Data for Good programs contributing to this document have drawn students across a wide swath of disciplines: **arts and humanities**; **engineering**; **health sciences**; **methodological fields** (e.g., computer science, math); **natural and physical sciences**; professional fields (e.g., business, public administration); and **social sciences**. The University of North Florida has achieved a **mix of disciplinary backgrounds** by pairing students from non-STEM majors with those from STEM majors, thereby providing a high-quality, multi-disciplinary educational experience to both groups. Non-STEM students benefit from working with teammates who have a strong background in the mathematical and computational skills that are foundational to data science. Students already steeped in data science methodologies benefit from working with peers versed in approaches from the social sciences, arts and humanities that are crucial for understanding social contexts, social theories, and social dynamics.

The D4G model can serve as an effective mechanism for broadening participation in data science.

What kinds of experience do you look for in students?

Assembling a cohort of students and mentors with the skills and perspectives needed to successfully complete a D4G project while simultaneously providing an educational opportunity requires careful consideration. In order to successfully complete a D4G project in the typical 10–14 week timeframe, most programs require students to arrive with a grasp of foundational data science skills, such as **statistics** and **programming**, along with experience in other requisite skill areas, such as **team-based work** or **social impact work**. Requiring previous experience may be why graduate students made up the majority of accepted fellows in the programs we surveyed. However, because D4G programs have an educational mandate, it is important that student participants do not exclusively work in areas they have already mastered, and that they are given opportunities to learn new skills and knowledge. Of course, students who already have abundant experience in the skill areas developed through the program can serve as mentors to D4G teams; indeed, several of the programs brought students on as both learners and mentors.

Although a mix of pre-existing skill sets and knowledge levels is desirable overall, D4G organizers have learned to be cautious about creating teams that have wide technical disparities between individuals. Data science experts recognize that technical skills are only one component of a rigorous and ethical data science practice, and that qualitative and critical ways of thinking are also essential (National Academy of Science, 2018). This is true in data science broadly, even more so when data-intensive methods and technologies are deployed in the pursuit of "social good." Nonetheless, D4G organizers note that students enter D4G programs in no small part because they want to advance their technical competencies. Small disparities in technical skills can be bridged through pedagogical interventions such as pair coding, peer-to-peer learning, one-on-one mentoring, and hands-on tutorials. Yet, when students in a team have large disparities in pre-existing technical skills, there is a tendency for less technically-advanced students to suffer from imposter syndrome or default to performing non-technical work, depriving them of the very learning opportunities they seek. A skills imbalance can impact the team's dynamic, resulting in lower performance and satisfaction with the program. Therefore, it is ideal to design a D4G project that provides an opportunity for all students to advance their technical competency. This task is made easier when students enter the program feeling they have roughly equivalent technical acumen as their peers, even when their specific skill sets may vary.

What motivations for applying will you look for in students?

Another important aspect of student selection is the students' motivation to use their data science skills for social good and to work on projects that are aligned with socially beneficial values. In fact, student motivation was the only criterion unanimously ranked a "high" priority for selecting students among the D4G organizers surveyed. Though students are motivated to apply to D4G programs in part to **advance their technical competency**, this motivation alone is not enough to create a satisfactory match. Students only interested in manipulating data without understanding social contexts and implications tend to be frustrated by the other essential aspects of socially beneficial data science. They may be more interested in gaining experience using cutting-edge tools rather than using the appropriate tool for the problem at hand. Students who thrive in D4G programs express a genuine **desire to achieve social impact** and have demonstrated prior experience in fulfilling that goal. These students are eager not only to improve their technical skills, but also to engage with other essential aspects of applied social research, such as **working with diverse colleagues and stakeholders**, and critically assessing ethical concerns. Students motivated by a desire to do data science in a way that is reproducible and human-centered also tend to thrive in a D4G program, as do students who come to D4G programs with the intention of **exploring career options** pertaining to socially beneficial data science.

What scope and process of student recruitment will you use?

Having a recruitment plan that approaches **Diversity, Equity, and Inclusion** with intention will help assure that a program attracts students who can contribute to and benefit from the D4G learning environment. To encourage disciplinary diversity, it can be helpful to ensure that the program is promoted through channels that cater to non-STEM majors, such as those in the arts and humanities, business, and public administration. Beyond skill level, student level, and disciplinary backgrounds, programs will want to consider how they might attract talented students who are traditionally underrepresented in STEM fields, including people of color; women and gender minorities; LGBTQIA people; those with disabilities; first-generation college students; and those who are socioeconomically challenged. To address the persistent underrepresentation of women, minorities, and people with disabilities in STEM, it is helpful to have a framework for evaluating candidates that **assesses prior achievements relative to opportunity**. The University of California Berkeley attempts to fairly assess career progression and achievement by considering achievements in light of the opportunities available to applicants. The institute weighs "quality of experience over quantity." The University of British Columbia aspires to include 50% or more of the accepted students from underrepresented groups. To assist these admission statistics, the University of British Columbia asks applicants to identify their gender and whether they are from an underrepresented group.

To reach underrepresented racial minorities, Vanderbilt University's Data Science Institute shares opportunities for students with the university's Associate Dean of Diversity Recruitment. In turn, the dean promotes these opportunities in multiple venues catering to minority students, including historically Black colleges and universities. The University of Ohio's Translational Data Analytics Institute has developed relationships with educators at institutions that disproportionately serve students who are racial minorities and from socioeconomically disadvantaged backgrounds (both of which are underrepresented in STEM fields and careers). When circulating recruitment opportunities with other data science educators, the Michigan Institute for Data Science at the University of Michigan includes a sentence remarking that they encourage underrepresented students to apply. The University of Washington offers a privately funded "Opportunity Scholarship" for students facing adverse circumstances. Students can apply this funding to expenses such as housing, childcare, telecommunication, and technology. All DSSG candidates are eligible to apply for this scholarship in addition to the regular program stipend, but if requests for funding exceed the scholarship funds available, priority is given to students who identify as belonging to underrepresented or disadvantaged groups.

A decision that impacts several aspects of a D4G program is choosing whether students will be drawn exclusively from the **home university** or **beyond** it. Recruiting only from the home university (or only those in close proximity) can reduce the need for housing, travel, and relocation expenses. Local recruitment can also lessen the logistical hurdles associated with paying students who are not in your university system. A program catering to local students may appeal to students who would not be able to afford travel and relocation costs. Local students may also stay engaged with projects after the program ends, benefiting both students and projects alike. At the same time, recruiting locally limits the pool of applicants and may reduce opportunities for students to broaden their professional and peer networks beyond their home university.

Most D4G programs accept **international students**. D4G programs offering stipends will require that international students are authorized to work in their host country. Programs considering accepting international students will need to consider visa requirements. If offering visa support, account for the lead time of successfully completing a visa application or work authorization addendum. In the United States, obtaining summer work authorizations is typically pro forma with the most common types of student visas, but requires that international students coordinate with the appropriate office at their home institution well in advance of applying to the D4G program. Payment to undocumented student participants can pose a challenge depending upon the internship structure and stipend/award distribution mechanism on your campus. Program organizers may want to consult the organization that addresses concerns of undocumented students on their campus, such as an Office of Minority Affairs.

Data for Good programs tend to be popular, and many programs reported receiving hundreds of applications per year. To manage the volume of applications, a number of programs recruit volunteers from outside their core program staff to help in the review process. For example, some enlist program alumni to review applications while others rely on data scientist volunteers from industry. Most programs also opt for a layered or staged selection process. For example, many D4G programs follow some combination of the following steps:

1. **Application screening.** At this stage, a reviewer quickly identifies applications from individuals who do not meet basic eligibility criteria, have not completed all parts of the application, or do not demonstrate proficiencies deemed necessary for success in the program. It should be noted that these baseline criteria vary considerably from program to program. For example, some programs may require proficiency in a particular programming language, while others may not.

2. **Application assessment.** At this stage, applicants who pass the initial screening are evaluated based on the merits of their application materials. It is advisable to have a standardized process and rubric for reviewers to follow, and for each application to receive multiple reviews.

3. **Interview assessment**. A subset of applicants with highly rated application materials are often invited to participate in at least one interview. Some programs conduct separate technical and nontechnical interviews, others ask a combination of technical and nontechnical questions in a single interview, and others ask only nontechnical questions. Regardless of which interview style is adopted, interviewees should be told what to expect.

4. **Final selection and placement**. The final selection of applicants is often made with consideration for the combination of interests, skills, and backgrounds that would make for an ideal cohort. At this stage, organizers synthesize everything they know about candidates from their written application materials and interviews to find the best possible match for the projects they will be developing during the program.

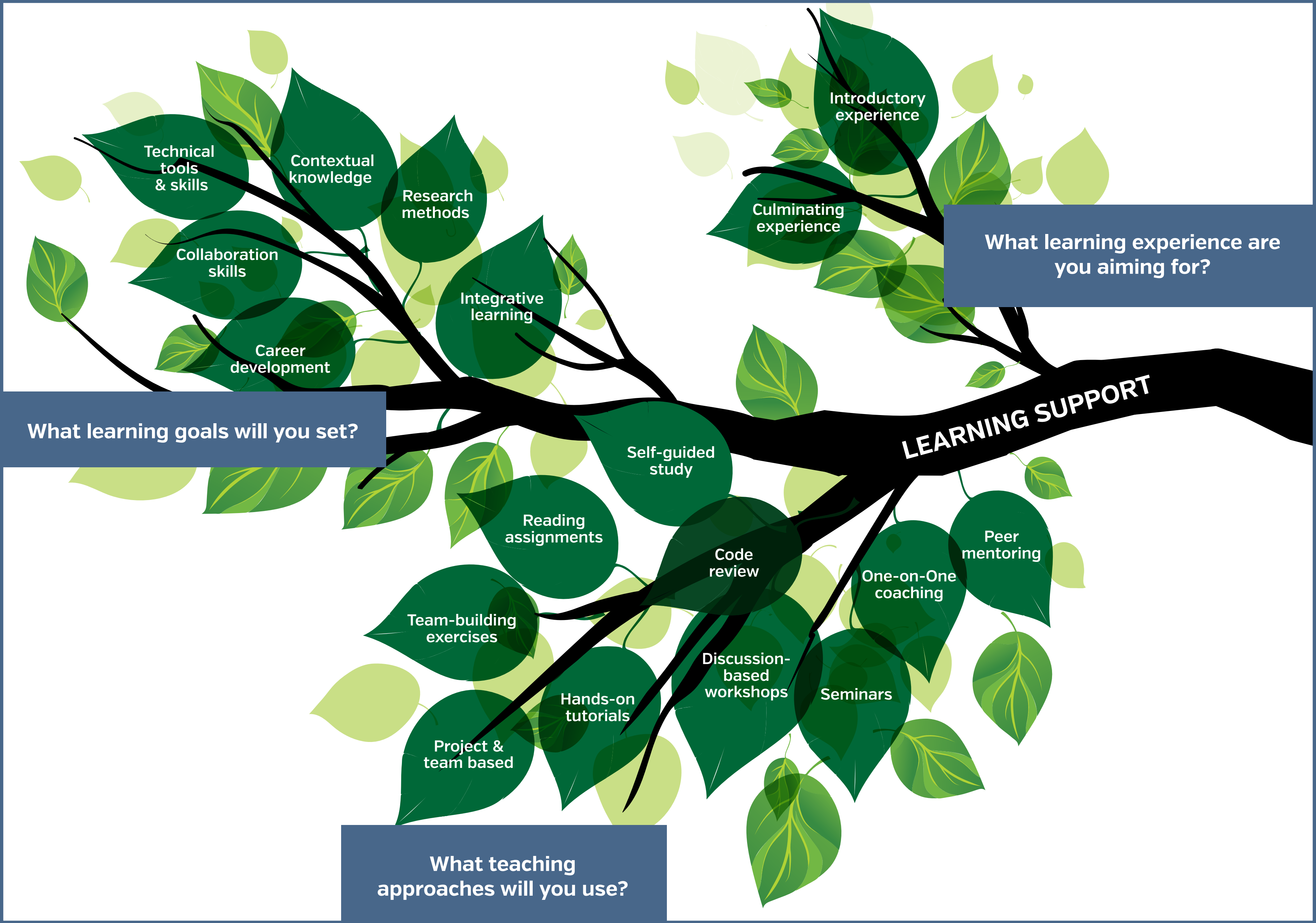


Figure 8. Learning Support Growth Map

LEARNING SUPPORT

Through project-based learning augmented with other modes of instruction, D4G programs enable students to engage with many aspects of socially beneficial data science. When designing a D4G program, consider the kind of learning experiences you wish to create, the learning goals you hope to achieve, and teaching approaches you will adopt. These considerations are interrelated and mutually reinforcing, and should ultimately inform the program's curriculum (which will be addressed in the following section).

What learning goals will you set?

We identified several common learning goals across D4G programs. Programs introduced several kinds of **contextual knowledge** required for responsible applied research. Because of the highly collaborative nature of applied data science, programs also helped students develop and reflect upon their **collaboration skills**. As may be expected, requisite **research methods** and **technical tools and skills** were also addressed in learning goals. Because D4G programs frequently are a platform for career exploration, many programs incorporated **career development** activities and coaching. To understand how these learning goals inform curriculum development, see the next section, "Curriculum," p. 35.

> ***Many programs use a mix of teaching approaches to complement project-based teamwork.***

What learning experience are you aiming for?

D4G programs can be tailored to a variety of learning experiences. For example, the University of North Florida's program was conceived as a way to **introduce applied data-intensive work** to relative newcomers, whereas the University of Massachusetts Amherst's program provided a **culminating capstone-like experience** to more advanced students. Several programs offered both of these experiences in one program by assembling teams composed of students with different experience levels. For example, the University of Virginia had teams of undergraduates mentored by graduate students.

The learning experience a program strives for will inform their choices of curriculum and teaching approaches. For example, an introductory experience probably requires more formal modes of instruction, while a capstone experience allows for more time to be spent on independent project work.

What teaching approaches will you use?

In tandem with planning an overall learning experience, it is helpful to consider what teaching approach will support each learning goal. For example, when students need support for a particular technical skill, will they be informally mentored one-on-one, participate in a group tutorial, or be directed toward a self-study resource? In practice, it can be challenging to anticipate which students will need the most assistance, and the most effective mode of support in advance of the program. One way programs address this challenge is by building flexibility and redundancy into their education plan. This means employing multiple modes of instruction and having mentors on hand who can adapt to students' needs.

As mentioned in the Selecting Projects section above, educational research shows **project and team based work** motivates students and can foster deep intellectual engagement (National Academies Press, 2020b). However, many programs have found it helpful to complement project- and team-based work with a mix of teaching approaches such as **seminars, discussion-based workshops, hands-on tutorials, reading assignments, team-based exercises, code review, self-guided study, one-on-one coaching, and peer mentoring**. In these ways, D4G programs are rich, multimodal learning environments that provide multiple opportunities to enhance students' understanding of data science for social good.

Photo Credit: Kevin Lin

Each mode of instruction lends itself to supporting students' awareness of a different dimension of D4G work. At the University of Warwick, general lectures and tutorials gave teams baseline data science technical skills, such as version control and proper coding practices. Students enjoyed talks by invited outside experts, who helped students understand the broader context of applied data science and share career advice. These planned activities were augmented by ample one-on-one coaching—a strategy adaptable to the needs of each student and each project. Through one-on-one coaching, emergent needs for training were identified for a larger group of students and then group trainings arranged. Reading assignments or self-guided study helped students gain contextual knowledge about their projects and locate information on data science methods and principles. For example, the Data Science for Social Good program at Carnegie Mellon University (formerly at the University of Chicago) created a reference on techniques used in the program for students and mentors, called *The Hitchhiker's Guide to Data Science for Social Good* (DSSG Fellowship, 2020). The University of Warwick suggests *The Turing Way: A Handbook for Reproducible Data Science* (Arnold et al. 2019) as a reference to students.

Tailoring learning priorities

Research methods, technical tools and skills, and **contextual knowledge** are contingent upon the projects that a program selects. Programs adapt their curricula accordingly (see "Curriculum" p. 35). Contextual knowledge is unique to each project and requires tailored investigation. Technical tools and skills are also often project-specific: Natural Language Processing tutorials and coaching may be offered to students taking on a project employing NLP; Shiny tutorials may be given to those building an interactive dashboard; and GIS tutorials provided to those taking on a mapping project. Preparing for these more contingent aspects of learning generally means lining up mentors with requisite knowledge in advance of the program, and sometimes recruiting specialists beyond the core organizing team. The alternative is to ensure that a given project can be accomplished with the skills students and mentors already have or can readily attain. Therefore, it is helpful to consider learning goals and curriculum in relation to program priorities and individual projects.

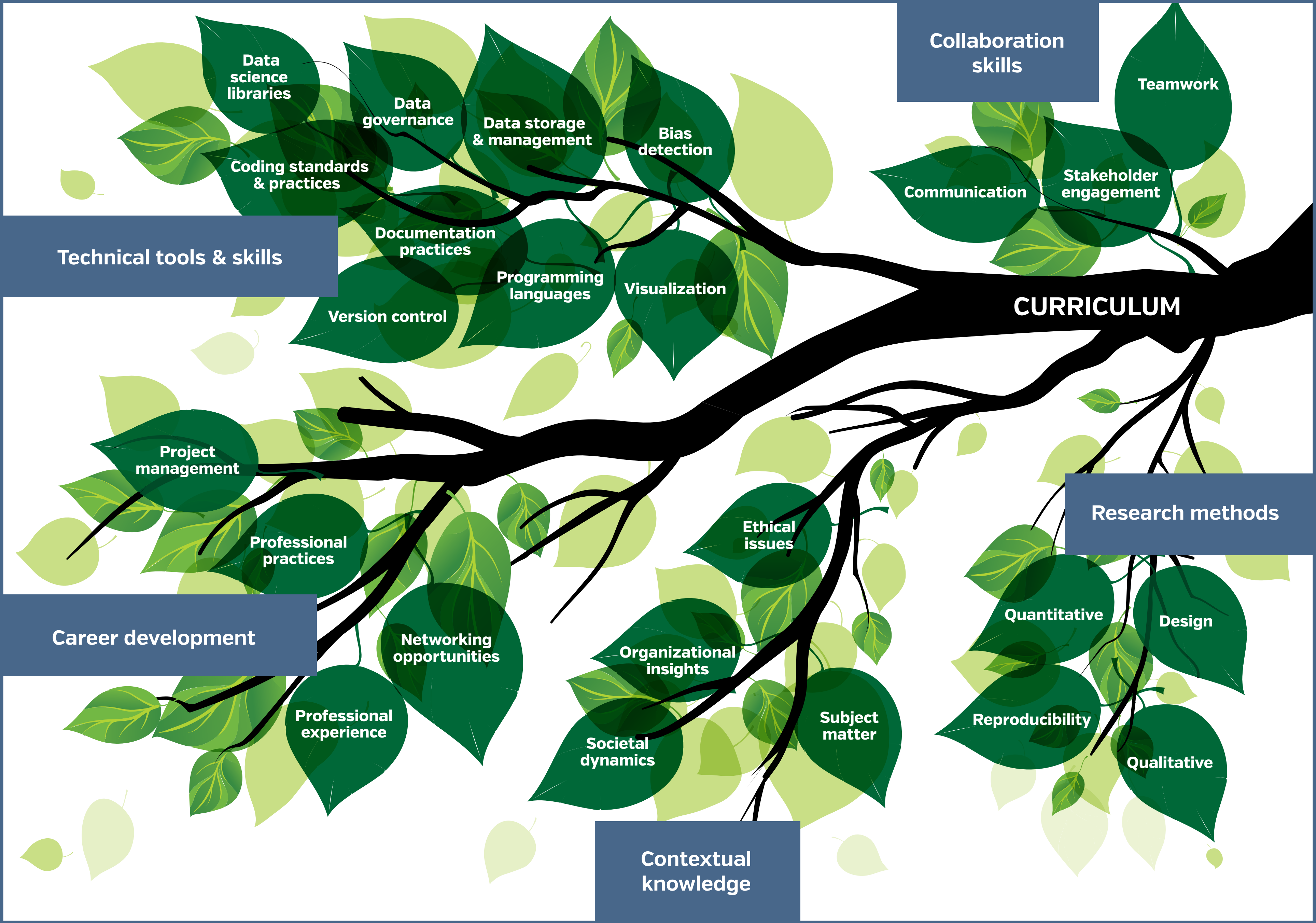


Figure 9. Curriculum Growth Map

CURRICULUM

Integrative learning through projects

As mentioned in the “Selecting Projects” section (p. 15), Data for Good programs foster integrative learning across the National Academies of Science’s four core areas of data science education (2018): foundational skills, translational skills, ethical skills, and professional skills. By creating a venue where each of these skills is advanced in relation to the others, students can deepen their understanding of both theory and practice of doing data science for a social good. They emerge better prepared to apply data science to matters of social concern in the future, whether through civic engagement activities or career activities. These core areas are attended to in Data for Good programs through the Learning Goals named in the previous section: Contextual Knowledge, Collaboration Skills, Research Methods, Technical Tools and Skills, and Career Development. In this section, we elaborate on how each of these goals translate into curricula..

Contextual knowledge

Data for Good programs are designed to help students grasp several kinds of **contextual knowledge** important to the effective and ethical application of data science to social issues. **Organizational insights** help students to understand how to best tailor their project deliverables so that their project partners can use them. Most projects also benefit when students are exposed to background **subject matter** relevant to the social challenge being addressed, such as knowledge, frameworks, and methods established by prior work. Because social interventions are inherently complicated and frequently contested, D4G students must be familiar with the core **societal dynamics** around a project, including the historical and structural dimensions of the problem they seek to address. Relatedly, D4G students must explore **ethical issues** that are inevitably related to their project and pertain to the application of data science to social issues. When attention is given to these themes within the curriculum, students are able to understand the nuances of their intervention and how it relates to the broader social sphere.

When contextual knowledge is sufficiently addressed, students often feel more comfortable and satisfied with what they accomplish in the program. The intentional introduction of contextual knowledge prepares students for a deeper integration of the “hows” and “whys” that pertain to their project and sets the stage for ethical decision-making. For example, contextual knowledge is important for determining whether feature engineering choices are fair and accountable to those represented in the data they are analyzing.

All programs in our network strive to give students the opportunity to work directly with project partners throughout the program.

Orienting students to the broader societal and organizational dynamics around their projects is particularly helpful when introduced at the beginning of the program. Background readings followed by facilitated team discussions are an efficient first step. Additionally, subject matter experts may be invited to engage in informal discussions with students, or students may conduct interviews with them. Organizational dynamics may be introduced by project partners in early meetings. Awareness of organizational context and subject matter may be further enhanced by site visits. All of these approaches help students understand the "why" of their project. As students dive deeper into the "how" of their projects, discussions with mentors can sensitize them to the ethical dilemmas associated with data science methods. Program organizers have incorporated a range of processes and activities for integrating ethical thinking and decision-making into the daily work of D4G projects, including conducting power analyses on their projects, facilitating interactive exercises that surface differences in privilege among participants, brown bag lunches with data science ethicists, readings from literature on "critical data studies," written reflections on the ethical issues implicated in projects, and engaging with tools designed to prompt discussions about ethics in technology development (Ballard et al., 2019; Calderon et al., 2021.; Kong & Hseih, n.d.). Importantly, opportunities for exploring and developing these forms of contextual knowledge should be ongoing and woven throughout the program.

Activities focused on direct **stakeholder engagement** are crucial to fostering the acquisition of contextual knowledge. All programs in our network strove to give students the opportunity to work directly with project partners throughout the program. For example, the University of Massachusetts Amherst has made students responsible for communication with their partnering organization (which they referred to as "clients") by the midpoint of their program. Students often also engaged with stakeholders who were not as directly involved in projects as partners, such as downstream users and affected communities. By interviewing a project's intended users, intended beneficiaries, and subject matter experts, students gain insights that help them refine their project's deliverables, improving the likelihood the intervention they devise will be effective. To facilitate interactions with these broader sets of stakeholders, program organizers have had success in relying on project partners to broker relationships (for example, a social service agency may arrange a meeting with agency clients). Making stakeholder engagement activities a formally recognized part of a D4G curriculum can help students and mentors gain experience and confidence in this essential aspect of applied research. For example, the University of Washington DSSG program has a designated "Human-Centered Design Mentor" to support teams planning and executing stakeholder engagements and integrating learnings derived from stakeholders into project work.

Collaboration skills

Because data science that aims to be socially beneficial is inherently an interdisciplinary activity, collaboration skills are another essential component of D4G curricula. Many academic leaders view the ability to effectively collaborate across disciplines as an essential 21st century skill (National Academies Press, 2020a), and collaboration and communication skills are consistently ranked by employers as among the most important.

To help students successfully collaborate across disciplines and within teams, programs give attention to the mechanics of effective **communication**, including public speaking, science communication, internal communication within teams, and communication with stakeholders. In terms of internal communication, attention is given to team roles and team formation. In terms of cross-sector and public communication, all programs end with public talks or poster sessions. Most augment work on projects with activities that help students reflect upon or improve their **teamwork** skills by co-developing documents that articulate norms of interaction within their teams, participating in team-building exercises, socializing, conducting self-assessments or team-assessments, and checking in with mentors.

Research methods

As may be expected, students in a Data for Good Program must engage with research design, quantitative methods, and qualitative methods. D4G programs are also a good environment for learning the concepts and mechanics of reproducibility.

A D4G curriculum is likely to address **quantitative research** methods closely associated with applied data science, such as machine learning, geospatial analysis, exploratory data analysis, predictive analytics, and causal analysis. Each of these methodologies are vast, and can only be selectively addressed in a D4G program. Particular quantitative methods are chosen with consideration for project work and student interest. Students generally appreciate tutorials that advance their quantitative skills. Therefore, program organizers report sometimes offering tutorials in quantitative methods even if those skills will not directly or immediately be implemented in project work.

We note that many of the activities that help students attain contextual knowledge about their projects (see "Contextual knowledge" p. 36) can be considered **qualitative research**—a fact students can miss if it is not pointed out to them. Activities such as stakeholder engagements are fruitful opportunities to lightly introduce qualitative research practices. For example, the University of Washington required all teams to do a virtual or in-person field site visit to a stakeholder venue in combination with an interview or group discussion within the first few weeks of the program. Teams prepared for the visit according to standard field research practices by writing an interview protocol together and discussing what they wanted to learn from the experience. Shortly after the visit, they reflected on new insights and how they might be incorporated into ongoing work. Once introduced, the collaborative process of “plan-do-reflect” is often adopted by teams for the duration of the program. Further, teams typically elect to do many more stakeholder engagements beyond the required one. These activities are supported by the Human-Centered Design Mentor (see “Contextual knowledge” p. 36 and “What roles will support the program?” p. 45).

Data science educators in our network have observed that it is difficult to conceive of the conceptual and practical aspects of **reproducibility** until one has a project that demands it. Therefore, D4G programs are exceptional learning environments for teaching this essential aspect of data science. Many programs incorporate tutorials on documentation practices and version control that help to scaffold project work by introducing tools such as Jupyter Notebooks, Docker, and GitLab. Concepts such as continuous integration, unit testing, and binders are introduced just as students are ready to consider them in relation to a hands-on implementation. This sets the stage for helping students to consider reproducibility in relation to **research design**. For example, will the team report only the final outcome of their analyses or all of the interim analyses including those that did not work? Will they strive to produce code that is reproducible only in relation to project data or that will also work with other data sets? In these ways, students gain experience with the technical, conceptual, and ethical dimensions of reproducibility.

Technical tools & skills

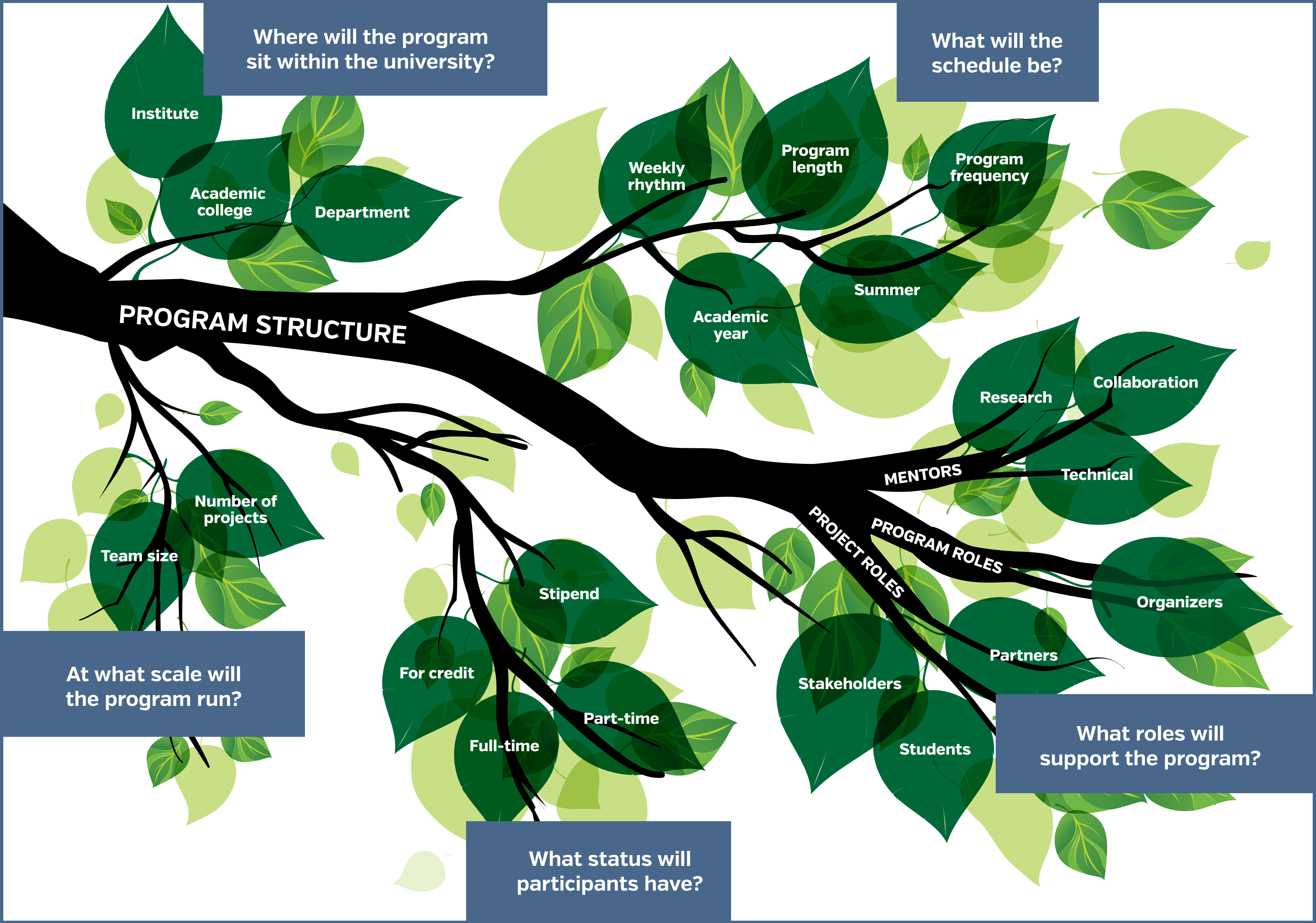
All D4G programs teach a suite of technical tools and skills, but the specifics vary widely both across programs and from year-to-year within programs. Regardless, most programs teach some version of tools and techniques in the following technical areas:

- Bias detection
- Coding standards and practices
- Data science libraries
- Data governance
- Data management and storage
- Documentation practices
- Programming languages
- Version control
- Visualization

Career development

Many students are motivated to participate in Data for Good programs in part because they are venues for career exploration and development. Students build their data science portfolio through the **professional experience** they get working on a D4G project, and relationships they develop through the program provide important **networking opportunities**. For example, the University of British Columbia and the University of Virginia have had a number of students approached by program partners to be hired. Brown bag lunches or career panels with data science professionals further expand students' understanding of potential careers. The University of Washington notes that several of their program alumni have switched career paths to work in social sector data science.

Many programs augment the **project management** experience gained through project work by offering training and coaching on topics such as time management and agile development. In some programs, a student project manager is assigned to a project. Project management is but one example of a number of **professional practices** that mentors help students hone in Data for Good programs. At the University of Virginia students are coached on how to talk to reporters about their projects. The University of Massachusetts Amherst has found that master's students who arrive in their program with high technical proficiency can benefit from coaching in professional skills such as writing an engaging LinkedIn profile and understanding how to introduce themselves to project partners.



Where will the program sit within the university?

Institute

Academic college

Department

What will the schedule be?

Weekly rhythm

Program length

Program frequency

Academic year

Summer

PROGRAM STRUCTURE

At what scale will the program run?

Number of projects

Team size

What status will participants have?

Stipend

For credit

Part-time

Full-time

What roles will support the program?

MENTORS

Research

Collaboration

Technical

PROGRAM ROLES

Organizers

Partners

PROJECT ROLES

Stakeholders

Students

Figure 10. Program Structure Growth Map

STRUCTURING A DATA FOR GOOD PROGRAM

High-level decisions concerning program structure include: Where will the program sit within the university system? At what scale will the program run? What will the program schedule be? Will students receive course credit? If they get paid, will they work full- or part-time? And what are the necessary roles to support the program?

Where will the program sit within the university?

An important decision that will impact the resources, relationships, constraints and opportunities of a D4G program is where the program will sit within the university system. Programs may be situated within traditional academic units on campus—such as in an **academic college** or a **single department**—or across multiple departments or schools. For example, the University of Warwick DSSG program is a joint initiative of Business, Maths, and Computer Science, with additional funding provided by the University. While this widens the pool of people involved and distributes the cost, having multiple departments involved creates additional administrative challenges. On the other hand, having university support has made many things possible that might have been difficult for a single department. Several D4G programs are affiliated with university-based data science **institutes** or centers, including the programs at the University of British Columbia, the University of Massachusetts Amherst, and the University of Washington. Running a program within such interdisciplinary centers allows programs to tap into an academically diverse campus network. The University of Virginia's program is housed within the Biocomplexity Institute's division of Social and Decision Analytics, enabling D4G projects to be carved out of the Institute's long-term research agenda.

At what scale will the program run?

It is possible to run a successful D4G program at different scales, in terms of the **number of projects** and **team size**. All programs have kept the number of students on a team relatively small—between two and five students. There is far more variety in how many projects a program will support in a given year. Some programs have focused on running just one or two projects, while others have taken on up to 14 projects. These differences mean that operating budgets also vary widely across programs. Among those that shared cost figures with us, program budgets ranged between $40,000 and $350,000 per year. The former was a program serving up to 12 paid interns, each working part-time on one of two projects over a summer. The low financial cost reflects not only the relatively small size of the program, but also the fact that several aspects of the program—including mentoring time—were supported by in-kind resources from university, faculty, and project partners. The most expensive program among those that shared their operating budgets served 24 students, each working full-time on one of six projects. It offered larger stipends than most of the other programs in part because they exclusively hired graduate students. The modal program among those that shared an operating budget served about 16 students and had an operating budget between $200,000 and $300,000, depending on the amount of in-kind support received for faculty, staff, and external mentors. The largest program costs were personnel salaries and student stipends.

What will the schedule be?

Programs confront a number of scheduling considerations including whether the program should run during the summer or academic year, the program length, the program frequency, and the weekly rhythm.

Summer or academic year. Most D4G programs run during the **summer** outside of normal course offerings. During summer, program leadership, mentors, and students generally have more time to devote to an intensive project. Summer is also ideal for programs that want to include students or faculty beyond their home university. Conversely, as discussed below, running projects during the academic year can have advantages when serving students exclusively from the home university.

Program length is largely determined by the home university schedule. However, if a program accepts students from other universities, this may impact start and end dates and program length because it will need to accommodate both semester-based and quarter-based academic calendars. For these reasons, program lengths ranged between 10–14 weeks.

However, university programs that build coursework around D4G projects during the school year, like Georgetown University and the University of Massachusetts Amherst, are able to run D4G projects more frequently. Running programs every quarter or semester can enable more students to engage in D4G project work. In contrast, the **program frequency** of most D4G programs is only once a year. Thus, they have to be selective about which stages of the data science lifecycle they target within their limited time frame [see "What stages of the Data for Good workflow will you support?" p. 19].

Weekly rhythm. The weekly schedule sets the balance each program chooses to strike between uninterrupted project work and activities that round out project work. Each program must decide how much time to allot to tutorials, career development conversations, one-on-one check-ins, stakeholder engagements, team building exercises, and reading assignments. At the University of Washington, the initial weeks of a program tend to be frontloaded with methods and technical tutorials, team-building exercises, and activities that establish contextual knowledge about the project. For a smoother start, it's helpful to have a detailed schedule planned out for the first two weeks prior to the first day.

What status will student participants have?

Programs may offer students positions as interns or fellows who earn a **stipend**. They may or may not offer them **course credit**. Work may either be **full-time or part-time**.

In theory, stipends and credit can be offered by the same program. In practice, among the contributors to our discussions, there was a split between programs that offered stipends but not course credit and programs that offered credit and not stipends. Some program organizers who considered offering credit for participation in their programs (in addition to providing monetary compensation) noted bureaucratic hurdles within their universities.

Most D4G programs chose to offer **stipends**, granting students the status of either "fellows" or "interns." Though we do not have direct examples of programs that do so, one can imagine a D4G program that incorporates students who work as volunteers, and programs frequently vet questions about volunteer participation. In our discussions, though, organizers from multiple programs said they had disappointing results trying to incorporate volunteers into their programs. Moreover, most D4G program organizers strongly felt that it was important to compensate students monetarily when project work was not offered for course credit or tied directly to a student's individual research. Several programs reported that positioning the program as a paid internship opportunity improved students' motivation for participation and their level of commitment. Another benefit of offering stipends is that they may enable students to attend who otherwise would not be able to do so, thereby contributing to student diversity.

Alternatively, running a D4G program as a for-credit option within a degree-granting program can enable more students to take advantage of the team-based, project-based experiential learning that the D4G model provides, especially if participation counts toward a degree requirement such as a capstone course. Running D4G projects during the academic year may also enable more prolonged student engagement and exposure to more aspects of the data science life cycle. For these reasons, Georgetown only offers for-credit opportunities, running D4G projects within courses. The University of Massachusetts Amherst has run a not-for-credit, paid stipend program for undergraduates during the school year (funded by an NSF grant) and offered a not-for-credit, paid stipend for graduate students during the summer. However, they are considering shifting to a for-credit program model.

Another dimension of participant status that is particularly relevant to summer-based extracurricular programs is whether students will work full-time or part-time on their D4G projects. A part-time program is less resource intensive, so enables participation by students who are not in a position to set aside other obligations, such as dissertation research, teaching assistantships, or caretaker responsibilities. Regardless, most program organizers felt that full-time work was necessary to meet the challenge of completing a D4G project within the short durations of their program. Therefore, most programs operated on a full-time work week during the summer months.

What roles will support the program?

The technologically-complex cross-disciplinary research and learning that takes place in D4G programs requires a clear articulation of roles and responsibilities. Programs varied by how they assigned responsibilities to the Program Roles among Organizers, and Project Roles to Students, Mentors, Partners, and Stakeholders.

Program Roles

D4G programs require a fair degree of administrative support, and program **organizers** report being engaged throughout the year. For programs running once a year, preparation becomes more intensive about six months out. For summer programs, the least amount of activity is likely directly after the program ends in the fall. Even in this relatively calm time, organizers reported having debriefing sessions, reviewing exit surveys of outgoing students, and beginning early conversations with prospective project partners. Therefore, D4G program leaders should consider how organizational responsibilities will be distributed. Common responsibilities included: project selection, development, and management; recruitment and selection of students, mentors, partners, and support staff; assembling project teams of individuals with complementary skills, experience, and work styles; partner relations; communications; curriculum development; scheduling; and fundraising. Technical infrastructure and management of facilities and finances must be designed, resourced, and managed. A number of programs rely on university staff to help with program administration, while others recruit students for administrative support.

Project Roles

Students (Learners). As discussed above, students are frequently interns, fellows, or credit-earning students engaged in hands-on project work. However, many programs blend learning and project leadership roles based on a student's experience level. At the University of Virginia, PhD students manage and support a team of undergraduates with oversight and guidance from a postdoc or faculty member. At the University of North Florida, more advanced undergraduates or graduate students are appointed leaders for their project teams. In most D4G programs students participate in a single project. However, at Iowa State University, each student participates on two teams, and project teams generally are composed of different students for each project, thus increasing collaboration opportunities among fellow students.

Mentors. There are three distinct areas of mentoring common in Data for Good programs. **Research mentoring** entails helping students answer a research question using sound methods. **Collaboration mentoring** can be defined as ensuring students work effectively with their teams and project partners. **Technical mentoring** involves helping students acquire the tools and techniques for managing and manipulating data.

Programs have a variety of ways to ensure these different kinds of mentoring are available to students. Several programs employ senior students (usually PhD students) as technical or research mentors. For example, at Stanford, faculty serve as research mentors and PhD students are technical mentors. At the University of Virginia, postdocs or senior PhD students provide research mentoring to undergraduate interns.

At the University of Washington, staff research scientists typically provide technical and research mentoring. Their role is designed to complement the strengths of the project leads who "own" the project and provide crucial mentoring in one or both of those areas. Staff research scientists and project leads also share responsibility for collaboration mentoring by providing project management support when helpful or necessary. Likewise, at the University of British Columbia, staff research scientists provide technical and research mentoring through structured periodic meetings and ad-hoc office hours. General research supervision is also provided by faculty at the University of British Columbia.

> ***It is possible to run a successful program at different scales, in terms of number of projects and team size.***

Support for collaboration and project management is handled differently, depending on the program. It is sometimes handled by the same individuals responsible for technical and research mentoring; other times it is a distinct role. For example, the University of Warwick recruits project managers who help teams collaborate, organize their time, and communicate with their project partners.

While many programs draw on mentoring expertise from within their own organizations, some programs choose to recruit external mentors. Programs at the University of Warwick and Carnegie Mellon University (formerly at the University of Chicago) have relied on professional data scientists on sabbatical from industry to mentor teams of graduate students. In other cases, external mentors have been brought in for short-term engagements such as tutorials, workshops, and career conversations. The University of North Florida invites industry data scientists and academic faculty to bi-weekly meetings with students, in which students discuss issues faced by project teams, and mentors share best practices and suggested solutions. Industry data scientists were also recruited to mentor students at the University of British Columbia during the first three years of the the program—a model put on hold due to the coronavirus pandemic.

Partners. Research intended to address a social concern generally occurs in an ecosystem of stakeholders. Navigating multiple relationships is one of the most challenging, but essential, aspects of applied research. All D4G programs contributing to this document emphasized helping students understand the organizational and societal relations shaping their projects (see also, "Contextual knowledge" p. 36). The primary means of doing so was by giving students the opportunity to work directly with partnering organizations. Programs viewed these relationships as having a high educational value for students and potential benefit to society. Therefore, substantive direct interaction between students and project partners was one of the signature aspects of the contributing Data for Good programs (see also, "What kind of partners will be involved?" on p. 18).

Programs variously position project partners as clients, sponsors, project leads, advisors, trainers, or mentors. This suggests differences in the nature of their partnerships. Yet, on a practical level, each program sets the expectation of substantive interaction among project partners, teams, and mentors. Frequent contact between students and partners throughout a program is seen as mutually beneficial and is often crucial to project success. Meetings between project partners and students have been an integral part of the program at the University of Virginia. At the University of Warwick, project teams and partners have two longer interactions at the start and end of the project with shorter weekly calls in-between. Likewise, the University of British Columbia requires weekly meetings between students and project "sponsors," and even more frequently when a sponsor is available. At the University of North Florida, students meet every other week with their "client" and all meetings are scheduled in the first week of the program. At a minimum, when access to a client is limited, Iowa State University schedules an initial meeting, a midpoint meeting, and final meeting. At the University of Washington, the relationship between project teams and project partners is even more intensive. Project partners are considered to be "Project Leads" rather than clients, and directly guide the intellectual work of the project. They are expected to co-work at least 16 hours each week alongside students and staff data scientists, regardless of whether they hail from the academic, nonprofit, or government sectors. This relatively high level of commitment assures that students and projects are well supported, and increases the likelihood that the project will be sustained beyond the duration of the summer program. However, this high level of commitment also limits the pool of projects to those with Project Leads willing and able to devote such a large amount of time to a single project.

Photo Credit: Mina Park

Stakeholders. Whereas partners are most directly involved with a D4G project, a wider range of stakeholders may indirectly influence it or be affected by it. Additionally, in many—if not most—instances of data science applied to social concerns, those who devise and perform the analyses and those who devise and implement interventions based upon those analyses are not the people who are most directly impacted by the analyses. For example, human service agencies may use data science to inform programs for people with drug dependencies or children at risk of abuse. In these cases, minor differences in analysis can be very consequential for the people involved in those programs. For these reasons, the contributing programs in our network believed it was essential for students to become sensitized to the broader social context of the research they engaged in beyond their point of contact at a partnering organization (see also, "Contextual knowledge" p. 36). Project teams often benefited from engaging with stakeholders who were unaffiliated with the partnering organization but nonetheless were impacted by their actions. For example, a team working to improve pedestrian routing at the University of Washington partnered with the Taskar Center for Accessible Technology. The team conducted multiple interviews and site visits with stakeholders beyond that organization, including contributors to OpenStreetMaps, individuals who used various styles of wheelchairs, and contractors who conducted accessibility audits for local transit agencies.

In the limited timespan of a D4G project, seeking input from a comprehensive set of stakeholders is not always possible. However, all teams benefit from having a conceptual understanding of the stakeholders who are more broadly attached to their project, especially the ultimate intended beneficiaries of their work. Teams tend to flounder without a shared understanding of the social world around their project, their intended impact, and how they relate to what they are doing. This conceptual understanding usually starts to take shape with an introduction to the social challenge provided by the subject matter experts associated with the project, and can be further bolstered through a systematic stakeholder analysis. However, students' conceptual understanding of the social world, relations, and impact of their projects tends to evolve throughout the program. Mentors can accelerate these understandings by frequently revisiting these concepts.

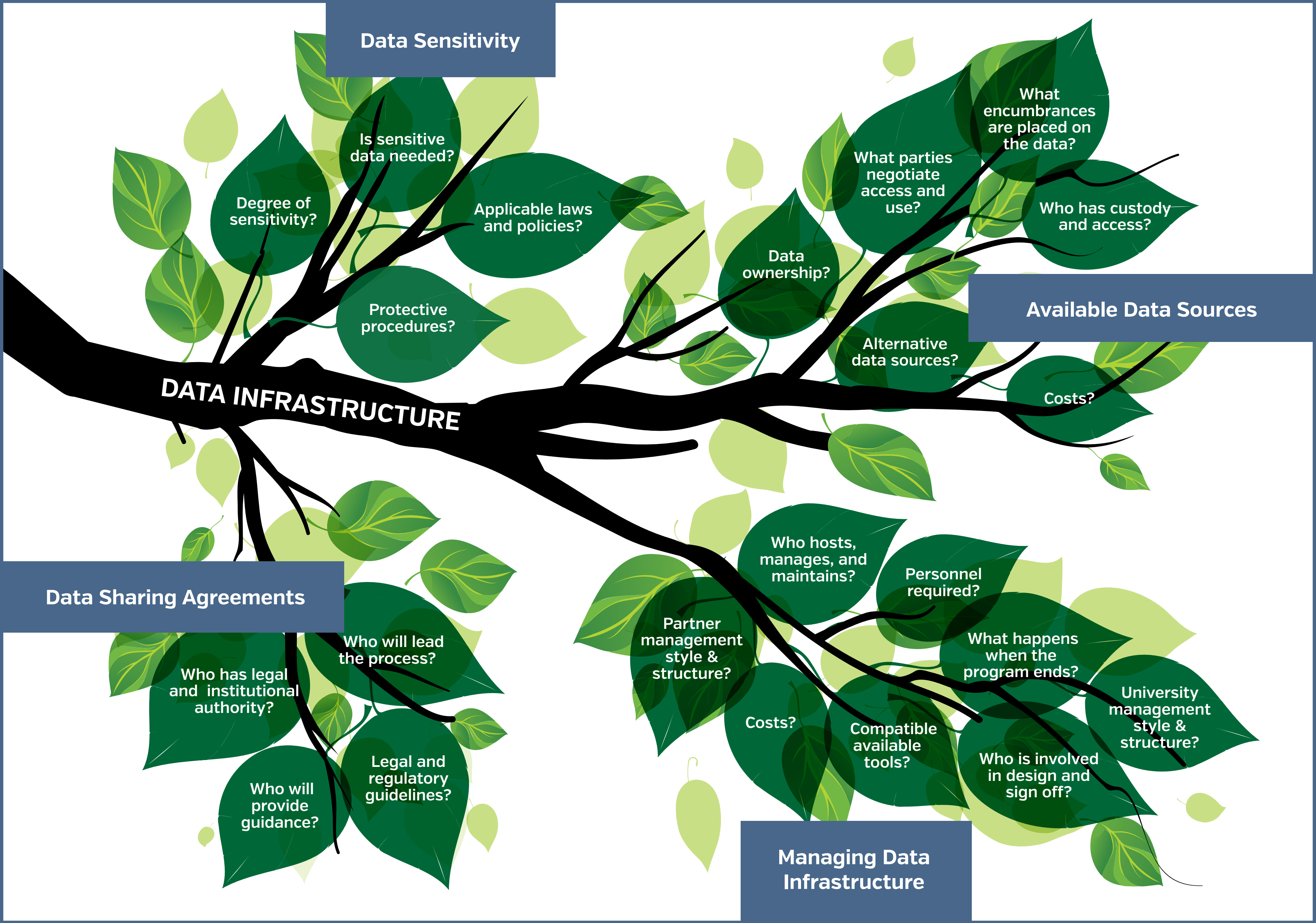

Data Sensitivity
Is sensitive data needed?
Degree of sensitivity?
Applicable laws and policies?
Protective procedures?
DATA INFRASTRUCTURE
Data ownership?
What parties negotiate access and use?
What encumbrances are placed on the data?
Who has custody and access?
Available Data Sources
Alternative data sources?
Costs?
Data Sharing Agreements
Who will lead the process?
Who has legal and institutional authority?
Legal and regulatory guidelines?
Who will provide guidance?
Who hosts, manages, and maintains?
Personnel required?
Partner management style & structure?
What happens when the program ends?
University management style & structure?
Costs?
Compatible available tools?
Who is involved in design and sign off?
Managing Data Infrastructure

Figure 11. Data Infrastructure Growth Map

DATA INFRASTRUCTURE

Foundational to the success of a Data for Good program is having a data infrastructure in place that enables the work of applied research while adequately addressing the inherent ethical, legal, and technical considerations of applying data science to social concerns. Many of the decisions that D4G programs make while developing data infrastructure can apply more broadly to any data science project. We summarize these general considerations below as a list of questions. We then touch on some of the ways these general considerations intersect with D4G programs more specifically. Finally, we offer strategies for developing data infrastructures employed by D4G programs.

General considerations for developing data infrastructure

The following considerations were derived from a D4G Organizer Network discussion in which program leaders reflected on their experience developing data infrastructure for their programs. Questions marked by an asterisk (*) are revisited in the following section on considerations particularly pertinent to Data for Good programs.

Evaluating the Availability of Data Sources

*Who owns the data?**

Who has custody and access to the data?

*What parties need to be part of negotiations around access and use?**

What encumbrances are placed on the data?

Are there open source or other alternative options for data?

What costs are entailed?

Evaluating the Sensitivity of the Data

*Is sensitive data needed?**

How sensitive is the data?

How will those represented in the data be protected? [E.g. de-identification, restricted access]

What laws, procedures, and policies apply? [E.g. GDPR, HIPAA]

Establishing Data Sharing Agreements

*Who has the legal and institutional authority to enter into a data sharing agreement?**

Who will lead the process of creating a data sharing agreement?

What legal and regulatory guidance and constraints apply?

Who will provide legal, regulatory and technical guidance on data sharing agreements?

Managing Data Infrastructure

*Who will host, manage, and maintain the technical infrastructure?**

What tools will be available with a given infrastructure?

What personnel will be needed?

What costs are entailed?

What is your partner organization's management style and structure for overseeing data infrastructure?

*What is your university's management style and structure for overseeing data infrastructure?**

Who will be involved in developing and signing off on data infrastructure at your university?

*What infrastructural resources will support the work after your program ends and how should this be taken into account during your program? **

Data infrastructure considerations specific to Data for Good programs

We now touch on a few ways that the general considerations in developing a technical infrastructure intersect with characteristics of Data for Good programs. We defined these characteristics earlier as programs that are hosted at a university, project- and team-based, educate students, seek social impact, integrate stakeholders, and use data science techniques.

University-hosted

Because Data for Good programs are university-hosted, a minimum of three entities will be involved in establishing data policies, as well as technical and legal arrangements for a D4G project: the partner organization, D4G program leadership, and university administration. It is important for program leaders to understand their home university's management approach to projects' data infrastructure.

Our contributors reported a range of experiences regarding data infrastructure at hosting universities. Some universities make it relatively easy for programs to create bespoke data sharing agreements tailored to a D4G project. Other universities have steered programs toward templated data sharing agreements and uniform data sharing policies. Some programs reported relatively easy access to legal advice for establishing data sharing agreements. Others have encountered challenges, such as not fitting into existing mechanisms to receive legal guidance from their university or receiving guidance that did not adequately consider the needs of all parties involved. For example, a university may have legal expertise available to support the development of data sharing agreements for project work specifically funded by outside sponsors, but lack a framework for the non-monetary-based data sharing typical of D4G programs. Program organizers may also receive confusing or conflicting guidance depending on which office they interact with on campus. For example, legal advice on data sharing agreements may conflict with records management policies that are designed to comply with public records laws.

Likewise, each university's management style for technical resources should be considered by programs. Some university managers may steer programs toward specific technical resources such as a preferred trusted data collaborative. Others require a technical review by IT to assure that procedures are in place so that data will not be accidentally shared.

In summary, the kind of support, degree of support, and mechanisms of support for data infrastructure that a program will receive reflects the particulars of each institution. We encourage those starting programs to become familiar with the relevant data policies, technical and legal arrangements, and administrative philosophy that prevail at their university.

Project-based

The data infrastructure of a D4G project is not one-size-fits-all, but varies depending on the nature of the project. This places a burden on programs to customize data infrastructure for every project. A data infrastructure's management strategy, associated costs, and ease of use will differ by project. Commonly, a D4G project will rely on some combination of the following: a university's computing power and networks, commercial cloud services (e.g., Azure, AWS, GCP), partner organizations' data infrastructure, and trusted data collaboratives (university-based or commercial).

D4G projects are typically short-term collaborations that take place over an academic term or year, and address one aspect of project partners' ongoing efforts. As such, D4G organizers must think carefully about how the data infrastructure of a given project will transition after the program has ended. Infrastructures used during the program may need to be actively dismantled (e.g., cloud services wiped of data and permissions revoked). Plans may need to be made for infrastructure components to be taken over or replicated by a project partner.

Team-based student education

Decisions made about data infrastructure can profoundly impact students' learning and work experience within a D4G program. Therefore, the needs of students must be considered in tandem with other data infrastructure considerations.

Students working in teams need a data infrastructure that enables them to work with data in a collaborative fashion. However, enabling multiple individuals to simultaneously work on the same data set requires specific technical provisions that may differ from those employed in other research settings. The difference in technical provisions is greatest when working with sensitive data. Though D4G programs take on projects involving data that ranges from less sensitive to more sensitive, the social and technical arrangements required to work with highly sensitive data may be better suited to longer-term collaborations. For this reason, some D4G programs steer away from projects that require highly sensitive data. For example, the University of British Columbia seeks out projects that can be developed around open data sources. Alternatively, data can be pre-processed to remove sensitive information. For example, project leads at the University of Washington are asked to anonymize data prior to making it available to students or placing it within a secure data infrastructure. If students have to handle personally identifiable data (PII) or other sensitive data, programs must plan to provide additional support to acquire more exacting proficiencies and specialized certifications.

It is important for program leaders to understand their university's approach to data infrastructure management.

Another aspect of data infrastructure that can profoundly impact students' experience in a D4G program concerns intellectual property. Non-disclosure agreements (NDAs) can prevent students from being able to discuss their work or build upon it. These limitations can diminish the value of participating in a program, as students often include program work in their resumes, data science portfolios, and public papers. Intellectual property constraints can prevent students from owning the code they create or from making it publicly available. An ideal intellectual property arrangement for a D4G program is one where students are able to discuss projects and own the copyright for the code they create, but partners have permission to reuse the code.

Integrate stakeholders immersed in the issue of concern

The social good mission of D4G programs means working with stakeholders immersed in addressing social concerns. Typically, project partners are nonprofits with a social mission, or government organizations. There are a range of data infrastructure considerations related to working with these different types of partners.

D4G organizers have noted a number of data infrastructure challenges clustered around smaller, less-resourced nonprofit and government organizations. Such organizations can be more reliant on external vendors to provide data infrastructure and services. In some cases, less-resourced organizations had less expertise in negotiating contracts with vendors and signed away ownership or rights of use, hindering academic collaborations. For example, in reviewing a contract between a database management vendor and a small government organization that was a prospective partner, researchers at Iowa State University discovered that the vendor demanded $500 each time data was shared with someone outside of the partner organization. The small government partner did not understand that when they signed a contract establishing a management service for their data, they had legally given away ownership of their data. Nor did they understand that they were constrained in what they could do with their data. Many D4G projects are designed around such obstacles, which ultimately are the result of systemic power imbalances between data infrastructure providers and the small organizations they serve.

Troublingly, D4G programs that partnered with less-resourced organizations have repeatedly encountered such issues. Multiple D4G programs reported that potential partners were unaware they did not own their data. A due diligence determination of ownership is critical for many reasons. If data are used without proper permissions, organizations may lose access to data they need for operations and funding. Leadership of several D4G programs saw it as their responsibility to ensure due diligence on such issues, viewing the constraints on what small government and nonprofits can do with their data as a social equity concern. They have taken an active stance by working with partners to overcome such obstacles.

However, our discussions also revealed the same kind of power imbalances can be perpetuated between universities and small government or nonprofit organizations. In such cases, D4G program leads found themselves in a dual role of trying to advocate for the interests of project partners while abiding by the policies and procedures of their home university.

The structural inequities touched on here suggest a broader set of issues concerning public data infrastructures beyond what D4G programs can take on. One practical stop-gap measure recommended by the University of Warwick was for program leads to initiate the first drafts of data sharing agreements rather than leaving them in the hands of university or project partners. Likewise, the University of North Florida drafted a data sharing agreement template in the form of an MOU tailored to the kinds of project partners they worked with. They worked with their University administration to attain approval of the template, thereby simplifying the negotiation process between project partners and the university for each D4G project.

Strategies for developing data infrastructures employed by D4G programs

D4G programs have found the following practices helpful for designing a D4G data infrastructure:

- **Prepare a list of questions to structure conversations around data sharing.** Having a prepared set of questions for project partners to help vet, scope, and plan data use and infrastructure development can streamline and accelerate the processes of data discovery and vetting. It is helpful if questions are phrased such that they can be answered by individuals who are not legal or technical experts on data sharing. Using more accessible language can expand the organizations that a D4G program can serve to include organizations that lack the expertise to put together a data sharing plan on their own. Questions should include who owns the data, how it is accessed, how it is downloaded, and what contracts it is subject to. The questions laid out at the beginning of this section on p. 50 provide a good starting place for generating such a list of questions.

- **Identify alternative sources of data.** Several programs reported needing alternative data sources when an initial data source became unavailable. Ideally, alternative data sources will be identified at the earliest stages of data discovery and project development.
- **Give plenty of lead time to put data arrangements in place.** It is prudent to anticipate that putting data sharing agreements in place, adding project participants to existing data sharing agreements, gaining custody of data, and assembling the technical infrastructure can take at least several months. While the lead time can be shorter in some cases, it is possible for hurdles to arise at any point. For example, partner agencies often have backlogs in either the technical or legal procedures that clear the path to data availability, and data agreements likely need to be cleared by university legal council prior to starting project work. For such reasons, the University of British Columbia begins negotiations four to five months before the start of the program.
- **Set a deadline for when to postpone a project for another time.** Occasionally, aligning data arrangements slows to the point when the best option is to forgo moving forward on the project. The University of North Florida has a deadline for having arrangements settled and data in hand two weeks prior to their program's start. During the project selection interview, UNF reviews data available from the project partners to ensure it can answer the proposed questions. If project partners do not provide access to relevant datasets by the deadline, the project scope is adjusted to address the problem using publicly available data. The University of Washington's deadline is two months prior to the program's start. This gives time to select an alternate project or to integrate students and other resources reserved for a project into those projects that are moving forward.

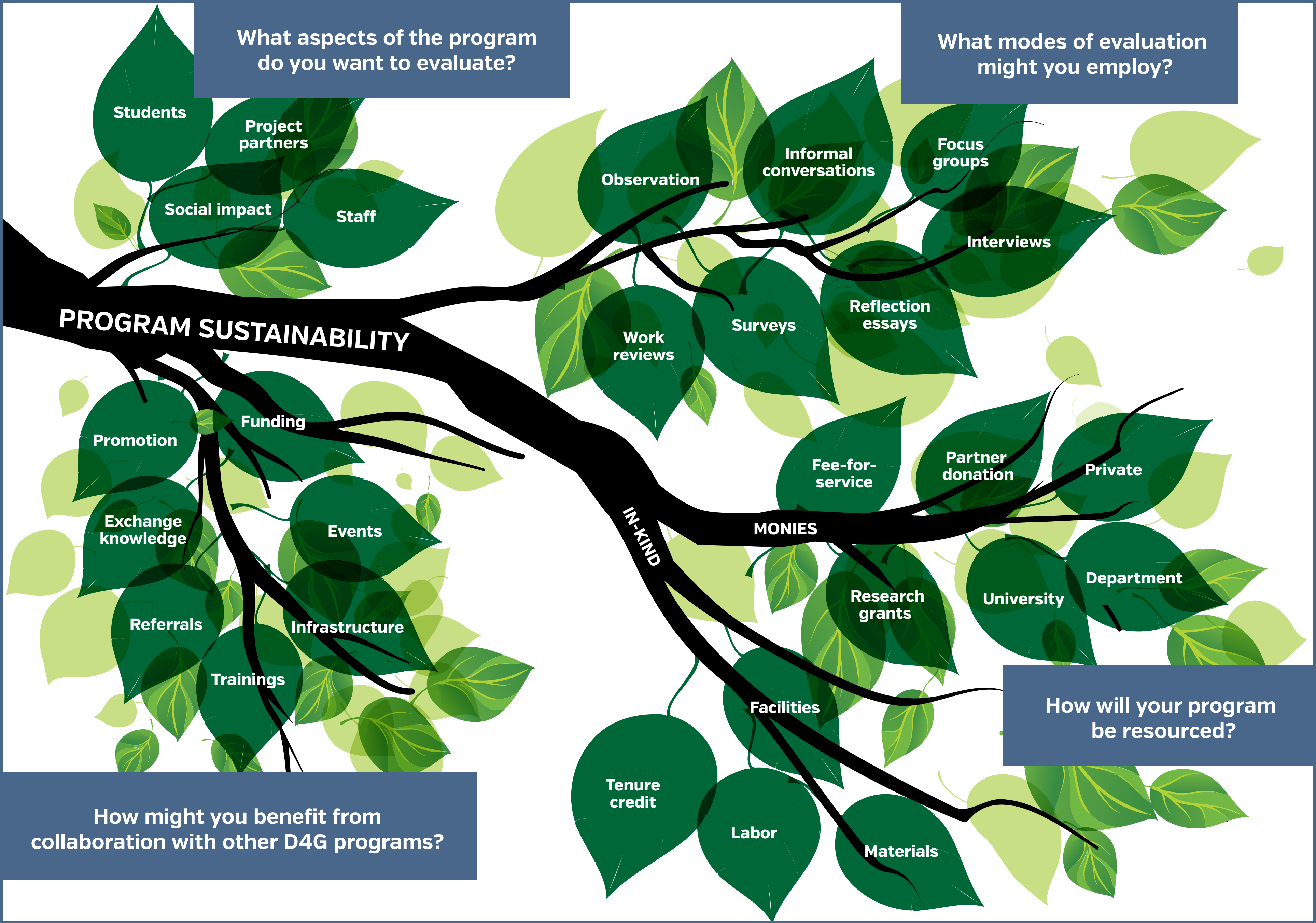


Figure 12. Program Sustainability Growth Map

SUSTAINING A DATA FOR GOOD PROGRAM

Questions of sustainability that apply to any new program are: How will it be resourced? What will be the measures of success and how will one know when it has been achieved? In this section, we highlight some of the ways D4G programs in our network have approached these questions. In particular, we touch on common ways programs are resourced, the impacts that are common priorities among programs and the modes of evaluation that are commonly employed. Additionally, given the trend of collaboration across programs among our contributors, we note that many D4G programs benefit from cross-program collaborations.

How will your program be resourced?

Programs frequently rely upon diversified funding sources from **private donors, fee-for-service contracts, research grants, university funds** or **department funds**, and **partner donations.** As noted above (see “At what scale will the program run?” p. 42), the total cost of programs varies broadly depending upon the number of students, projects, and the level of in-kind support.

Private donors. Generally, the most flexible sources of funding for D4G programs are private donors such as foundations or industry partners who are interested in supporting socially beneficial data science. Private donations in the form of gifts can be used for expenses not covered by other funding sources. A challenge of private donor funding is that it can be difficult to achieve a long-term commitment to funding, as private donors often want to diversify what they fund and tend to be interested in starting new efforts rather than providing sustaining support. Nevertheless, a number of programs are sustained to some degree through private sponsorship.

Research grants. Another approach that D4G programs take is line-item funding for a D4G project within a research grant. This may be beneficial for grants that require or otherwise privilege a research proposal that incorporates a strong student training, community impact, or service component. In such cases, a grant proposal can be strengthened by adding a D4G project as a line item. A drawback to this approach is that it typically requires multi-year planning to execute, with the research being proposed and funds granted well in advance of the D4G program. However, this approach has worked well for the University of Virginia.

University and department funds. Several programs receive some degree of financial support from their home university. Others self-fund from the general operating budget of their unit. The rationale for obtaining such funding is that D4G programs can support the educational, service, and research missions of the university. Additionally, D4G programs often execute projects that draw interest and attention from prospective students, media, donors, and the public.

Fee-for-service is another revenue source that can work well for small projects. This funding model can work well when the contracting agency—a nonprofit organization or a local government—is accustomed to contracting with academic units. Fee-for-service projects offer the opportunity for a mutually beneficial partnership to be formed. The fee typically covers the cost to the academic unit for student participation. For example, at Iowa State University, a fee-for-service model has allowed for funds to be used to cover student transportation, cost of materials (i.e., printed reports) and a small stipend to support a brief period of pre- or post-course project management work for a student intern or faculty member. A formal fee-for-service agreement should be well structured to set clear expectations for both parties, including deliverables and rights of ownership related to the data and intellectual property. Alternatively, some programs offset costs through **partner donations**, posed as optional and voluntary contributions.

Many D4G programs benefit from cross-program collaborations.

In-kind support is as vital to many D4G programs as monetary funds, and frequently includes **facilities, materials, labor,** or **tenure credit.** For example, facilities may be provided at little or no cost by a university or department. A collaborating academic unit or industry partner can donate material resources such as computing power or data hosting. Social event spaces and catering could be provided by a collaborating organization or private donor. Project partners or industry mentors frequently provide gratis labor. Program organizers also sometimes contribute portions of their labor gratis. For example, faculty participating in the D4G program at the University of North Florida can count portions of their participation toward tenure for teaching, research, and service requirements. This makes it more feasible for faculty to participate in the program.

How will your program approach evaluation?

What aspects of the program do you want to evaluate? Integral to questions of ongoing program sustainability is how organizers understand and communicate their programs' impact. It can be beneficial for program leadership to consider what success looks like in terms of **social impact gains** as well as **gains made by project partners, students, faculty, and staff**, as all of these affect ongoing program sustainability. It is ideal for program leadership to discuss and set goals, definitions of success, and program priorities, giving consideration to how success may be evaluated for each. These goals may be tailored to leadership's interest and focus. For example, the University of North Florida is keenly interested in how their program impacts students' professional identity formation and perceptions of STEM fields, so they orient their evaluation around those goals.

What modes of evaluation might you employ? Contributing programs employ several modes of evaluation to understand their programs. **Informal conversations, focus groups, interviews, observations, reflection essays, surveys**, and **reviews of project deliverables** have each proven helpful for understanding program strengths and opportunities for improvement.

D4G organizers have adopted a number of qualitative approaches for appraising their program's impact and effectiveness. During the program, informal conversation and observation are frequently used to solicit verbal feedback from students, mentors, and project partners. Some programs have also incorporated brief weekly surveys or verbal check-ins with students to understand their experiences as the program progresses. Such weekly assessments help organizers respond to issues as they arise. Other options include a survey distributed directly after tutorials and workshops, which can help a program steer mentoring resources to content that most benefits students. Exit surveys, interviews, and focus groups can give organizers a sense of the overall perception of a program from the perspective of students, mentors, and project partners. A debrief among the program's leadership shortly after the program concludes is another relatively easy way to capture important insights on program successes and opportunities for change.

Objective measurement of a program's impact can be more difficult to achieve than the qualitative appraisals mentioned above. Our conversations surfaced the challenges to assessing lasting social impact (including a lack of resources for longitudinal evaluation) and to objectively measuring gains made by students, partners, or staff in the program. One program in the D4G network attempted pre- and post-skill evaluations for student participants, but found that the assessment tool did not match what students actually learned in the program. Another program experimented with having external industry experts review students' projects. When expert reviewers are not educators, some preparation may be needed to orient them to the dual mission of the programs (as a vehicle for both service and education) to calibrate their expectations of student work.

Whatever the means of gaining insights about the program's impact, the most important element of an evaluation plan is ensuring that program leadership makes time for informed reflection on how the program is performing and that there is room to incorporate insights into future iterations of the program.

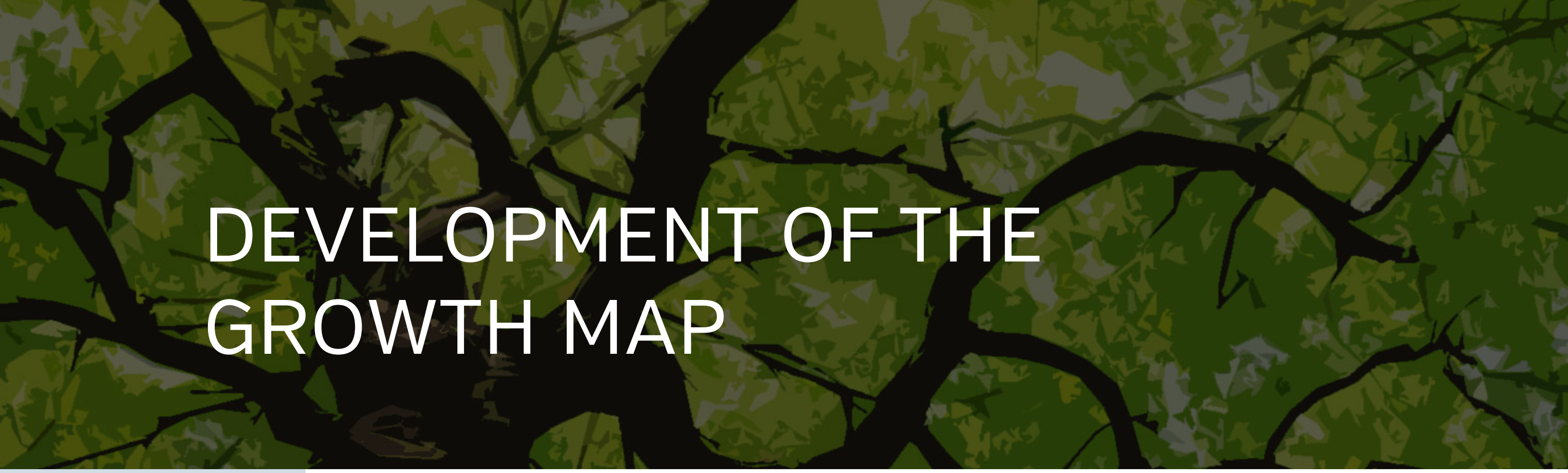
How might you benefit from collaboration with other D4G programs?

Finally, in reviewing resources that D4G programs have martialed for their programs, we observed that quite a few programs collaborated with their peers at other universities to: pursue **joint funding; run shared events** and **trainings;** reduce **infrastructure** costs; **cross-promote** their programs; tap into each other's networks for resources and **referrals;** and **exchange knowledge.** Therefore, it may be fruitful for new programs to consider how they might collaborate with programs at other universities.

D4G program organizers benefited from the discussions that informed this document. This cross-university collaboration is but one among many in our network. When starting, the University of British Columbia's program took inspiration from the University of Washington's Data Science for Social Good program. During the University of British Columbia's first year, both programs were funded through the Cascadia Urban Analytics Cooperative and Microsoft. The two programs then collaborated on joint program activities. After pursuing joint funding, the University of Virginia, Virginia Tech, Oregon State University, and Iowa State University also collaborated on student training. Both students and projects gained greater visibility through a shared website and a joint symposium including a student poster session at the conclusion of the summer programs. The University of Warwick and the Turing Institute initially affiliated with the "DSSG Europe" initiative started by the DSSG program at Carnegie Mellon University (formerly at the University of Chicago). This affiliation enabled Warwick and Turing to recruit top students from an extensive worldwide network and reduce their in-house infrastructure resources through a shared website. Through the generation and sharing of this document, we hope to inspire and support new collaborations as new D4G programs develop.

CONCLUSION

In D4G programs, university-based researchers work closely with partners and stakeholders to address pressing social concerns while training the next generation of Data for Good professionals in the art of interdisciplinary, applied data science research. As this document has made clear, organizers of D4G programs must balance multiple priorities and commitments. They must select and support projects that are not only feasible to complete in a compressed time frame, but also provide sufficient learning challenges for students, and meaningful impact for project partners and stakeholders. They must equally support the development of technical skills, research methods, interdisciplinary collaboration, and contextual expertise such as ethics and subject matter knowledge. They must structure and staff their programs to advance these multiple aims, identify and develop the infrastructures to enable data for good work, and find ways to justify and sustainably support a resource-intensive program. All of this makes running a D4G program complex and time-intensive, and institutions considering launching such an initiative face a number of important decision-points in designing their programs. But as our network of organizers has found, the consideration and care that goes into developing D4G programs produces enormous opportunities. These programs are important sites for exploring, advancing, and articulating the theory and practice of data science applied ethically to social concerns. They are platforms for cultivating cross-sector collaborations to tackle some of the most challenging issues facing society today. And D4G programs have proven to be desirable experiences for a wide range of students, including those underrepresented in STEM—expanding career opportunities of participating students and fostering a new generation of critical and capable data-intensive researchers. By elaborating on the high-level decision points that shape D4G programs, we hope to assist "seedling" programs in charting their own plans for growth and contribution to this growing field.



DEVELOPMENT OF THE GROWTH MAP

Method summary

The Growth Map presented here originated from a survey of D4G organizers and subsequent discussions facilitated by the University of Washington eScience Institute in collaboration and coordination with many partners. The survey was initially devised as a prelude to an in-person workshop of D4G leaders planned for March 2020. Instead, due to the COVID-19 pandemic, online discussions were held between July 2020 and April 2021. A benefit of this shift was that more programs at all stages of development were able to participate. In total, Data for Good program organizers representing nine active Data for Good programs and four in development participated in these discussions. Additionally, this work benefited from contributions by those running programs and organizations doing adjacent and related work. In all, contributors hail from 17 universities. By looking at the patterns of variation and convergence that emerged from the survey and discussions, we were able to identify key decision points for Data for Good programs.

Authors

Dharma Dailey, University of Washington, Research Associate, eScience Institute & Human-Centered Design Mentor, UW Data Science for Social Good

Sarah Stone, University of Washington, Executive Director, eScience Institute & Director, UW Data Science for Social Good

Anissa Tanweer, University of Washington, Research Scientist, eScience Institute & Program Chair, UW Data Science for Social Good

Contributors

The following individuals contributed via survey, monthly discussions, or both:

Data Science for Public Policy, Georgetown University, Massive Data Institute
Michael Bailey, Director Massive Data Institute, MDI Scholars Program, Georgetown University

Data Science for Social Good, Carnegie Mellon University (formerly University of Chicago)
Rayid Ghani, Distinguished Career Professor, DSSG Program Director and Founder
Kit Rodolfa, Senior Research Scientist

Data Science for Social Good, University of British Columbia, Data Science Institute
Kevin Lin, Research Administrator, Data Science Institute
Raymond Ng, Professor, Computer Science, and Director, Data Science Institute

Data Science for Social Good, University of Warwick, Turing Institute
Juergen Branke, Director of DSSGx UK (Warwick); Professor of Operational Research & Systems at the University of Warwick
Sebastian Vollmer, Founder of DSSGx now Professor of Computer Science at University of Kaiserslautern and German Center for Artificial Intelligence

Data Science for the Common Good, University of Massachusetts Amherst, Center for Data Science
Matthew J. Hale Rattigan, Director of Community Initiatives

Data Science for the Public Good, Iowa State University
Cassandra Dorius, Associate Professor, DSPG Program Co-Principal Investigator
Shawn Dorius, Associate Professor, DSPG Program Co-Principal Investigator
James Reecy, Associate Vice President for Research, DSPG Program Co-Principal Investigator
Christopher J. Seeger, Professor and Extension Specialist, DSPG Program Co-Principal Investigator

Data Science for the Public Good, University of Virginia, Biocomplexity Institute
Sallie Keller, Distinguished Professor in Biocomplexity, Director of the Social and Decision Analytics Division within the Biocomplexity Institute and Professor of Public Health Sciences
Gizem Korkmaz, Research Associate Professor, University of Virginia Data Science for the Public Good Young Scholars Program
Aaron Schroeder, Research Associate Professor

Data Science for the Public Good, Virginia Tech, Virginia Cooperative Extension
Michael Lambur, Associate Director, Program Development

Florida Data Science for Social Good (FL-DSSG), University of North Florida
Dan Richard, Associate Professor, Psychology, Co-Director, Florida Data Science for Social Good
Karthikeyan Umapathy, Associate Professor and FIS Distinguished Professor in Computing, Co-Director of Florida Data Science for Social Good

Research4Impact, Johns Hopkins University
Adam Levine, SNF Agora Associate Professor of Health Policy & Management Bloomberg School of Public Health and President of Research4Impact

Stanford Data Science for Social Good, Stanford University
Balasubramanian Narasimhan, Senior Research Scientist in the Department of Statistics and the Department of Biomedical Data Sciences; Director of the Data Coordinating Center in the Department of Biomedical Data Sciences
Chiara Sabatti, Professor of Biomedical Data Science and Statistics, DSSG faculty coordinator
Ben Stenhaug, PhD Student and DSSG Mentor

The Ohio State University, Translational Data Analytics Institute

Tanya Berger-Wolf, Director TDAI, Professor, Computer Science and Engineering; Evolution, Ecology, and Organismal Biology; and Electrical and Computer Engineering
Steve Chang, TDAI Technology Solutions Engineer
Jenna McGuire, TDAI Associate Director
Joshua Roush, TDAI Internship and Student Programs Coordinator
Cathie Smith, TDAI Managing Director

University of California Berkeley, Berkeley Institute for Data Science

Maryam Vareth, Health and Life Sciences Lead, Co-Director Innovate for Health Initiative

University of Michigan, Michigan Institute for Data Science

Jing Liu, Managing Director Michigan Institute for Data Science

University of St. Andrews, School of Computer Science

Diego Arenas, Research Engineer, EngD Student

University of Washington Data Science for Social Good

Dharma Dailey, Research Associate, eScience Institute & Human-Centered Design Mentor, UW Data Science for Social Good
Emily Keller, Program Coordinator, UW Data Science for Social Good
Sarah Stone, Executive Director, eScience Institute & Director, UW Data Science for Social Good
Anissa Tanweer, Research Scientist, eScience Institute & Program Chair, UW Data Science for Social Good

Vanderbilt University, Data Science Institute

Andreas Berlind, Associate Professor of Physics and Astronomy, Co-Director of the Data Science Institute
Amanda Harding, Assistant Director, Data Science Institute

Support

We thank the eScience Institute (University of Washington) and the NSF West Big Data Innovation Hub for supporting this project.

Attribution

The Data for Good Growth Map: Decision Points for Designing A University-Based Data for Good Program. Dharma Dailey, Sarah Stone, Anissa Tanweer and the Data for Good Organizer Network. University of Washington, 2021.

Layout, Illustration and Graphic design by Monique Heileson.

Photos provided by the Data Science for Social Good programs at the University of British Columbia, the University of North Florida, and the University of Washington.

Use

REFERENCES

Alspaugh, S., Zokaei, N., Liu, A., Jin, C. & Hearst, M. A. (2018). Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics,* 25(1), 22-31.

Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., & Whitaker, K. (2019). The Turing Way: A Handbook for Reproducible Data Science. *Zenodo.* http://doi.org/10.5281/zenodo.3233986

Ballard, S., Chappell, K. M., & Kennedy, K. (2019). Judgment Call The Game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 421-433). https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall

Belanger, A. L., Joshi, M. P., Fuesting, M. A., Weisgram, E. S., Claypool, H. M., & Diekman, A. B. (2020). Putting belonging in context: Communal affordances signal belonging in STEM. *Personality and Social Psychology Bulletin, 46*(8), 1186-1204.

Boucher, K. L., Fuesting, M. A., Diekman, A. B., & Murphy, M. C. (2017). Can I work with and help others in this field? How communal goals influence interest and participation in STEM fields. *Frontiers in Psychology, 8,* 901.

Calderon, A., Taber, D., Qu, H., & Wen, J. (2021). *AI Blindspot: A discovery process for spotting unconscious biases and structural inequalities in AI systems.* MIT Media Lab/Berkman Klein Center. https://aiblindspot.media.mit.edu/

Crisan, A., Fiore-Gartland, B., & Tory, M. (2020). Passing the Data Baton: A Retrospective Analysis on Data Science Work and Workers. *IEEE Transactions on Visualization and Computer Graphics, 27(2)*, 1860-1870.

Data Science for Social Good Fellowship. (2020). The *Hitchhiker's Guide to Data Science for Social Good.* Carnegie Mellon/University of Chicago. https://github.com/dssg/hitchhikers-guide

Engler, A. (2020). *Tech cannot be governed without access to its data.* Brookings Institution Tech Tank. https://www.brookings.edu/blog/techtank/2020/09/10/tech-cannot-be-governed-without-access-to-its-data/

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

Grolemund, G., & Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review, 82(2),* 184-204.

Ko, P., & Hsieh, M. N.d. *The Tarot Cards of Tech. Artefact Group.* http://tarotcardsoftech.artefactgroup.com/

Liu, Y., Althoff, T., & Heer, J. (2020). Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Moreno, C., González, R. A. C., & Viedma, E. H. (2019). Data and artificial intelligence strategy: A conceptual enterprise big data cloud architecture to enable market-oriented organisations. *The International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI), 5(6)*, 7-14.

National Academies of Sciences, Engineering, and Medicine. (2018). *Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report.* National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2020). *A. The Integration of the Humanities and Arts with Sciences, Engineering, and Medicine in Higher Education: Branches from the Same Tree.* National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2020). B. *Promising Practices for Addressing the Underrepresentation of Women in Science, Engineering, and Medicine: Opening Doors.* National Academies Press.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

O'Neil, A. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

Pomalaza-Ráez, C., & Groff, B. H. (2003). Retention 101: Where robots go… students follow. *Journal of Engineering Education,* 92(1), 85-90.

Perrin, A. (2020). *Half of Americans have decided not to use a product or service because of privacy concerns.* Pew Research Center. https://www.pewresearch.org/fact-tank/2020/04/14/half-of-americans-have-decided-not-to-use-a-product-or-service-because-of-privacy-concerns/

Rainie, L., Anderson, J., & Page, D. (2017). *Code-dependent: Pros and cons of the algorithm age.* Pew Research Center. https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/

Rawlings-Goss, R., Cassel, L., Cragin, M., Cramer, C., Dingle, A., Friday-Stroud, S., Herron, A., , Horton, N., Inniss, T., Jordan, K., Ordóñez, P., Rudis, M., Rwebangira, R., Schmitt, K., Smith, D., & Stephens, S. (2018). *Keeping Data Science Broad: Negotiating the Digital and Data Divide Among Higher-Education Institutions.* South Big Data Innovation Hub.

Steinberg, M., & Diekman, A. B. (2018). Considering "why" to engage in STEM activities elevates communal content of STEM affordances. *Journal of Experimental Social Psychology,* 75, 107-114.

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., 2nd, Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., Lowe, A., Newman, S., Okolo, V., Olroyd, S., Peecook, B. R., Pickett, S. B., Slager, D. L., Caviedes-Solis, I. W., Stanchak, K. E., Sundaravardan, V., Valdebenito, C., Williams, C. R., Zinsli, K., & Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences of the United States of America,* 117(12), 6476-6483.

Vaz, R., & Quinn, P. (2014). Long term impacts of off-campus project work on student learning and development. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (pp. 1-5).

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Profile Books.

Zhang, A. X., Muller, M., & Wang, D. (2020). How do data science workers collaborate? Roles, workflows, and tools. *In Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), 1-23.

APPENDIX 1

Contributing Programs at a Glance

Data for Good Programs

Home Institution	Program Name	Start	Cycles thru 2020	Students thru 2020*	Projects thru 2020*	Website
Carnegie Mellon University (formerly University of Chicago)	Data Science for Social Good	2013	9	300	100	https://www.dssgfellowship.org
University of Virginia, Biocomplexity Institute	Data Science for the Public Good Young Scholars Program	2014	7	98	69	https://biocomplexity.virginia.edu/institute/divisions/social-and-decision-analytics/dspg
University of Washington, eScience Institute	UW Data Science for Social Good	2015	6	86	21	https://escience.washington.edu/dssg/
University of British Columbia, Data Science Institute	Data Science for Social Good	2016	4	56	15	https://dsi.ubc.ca/data-science-social-good
University of North Florida	Florida Data Science for Social Good (FL-DSSG)	2017	4	31	15	https://dssg.unf.edu
University of Warwick, The Alan Turing Institute	Data Science for Social Good Fellowship at Warwick/Turing (UK)	2019	2	33	8	https://warwick.ac.uk/research/data-science/warwick-data/dssgx/
University of Massachusetts Amherst, Center for Data Science	Data Science for the Common Good	2019	2	25	10	https://ds.cs.umass.edu/industry/data-science-common-good
Stanford University	Stanford Data Science for Social Good	2019	2	17	5	https://datascience.stanford.edu/programs/data-science-social-good-summer-program
Iowa State University	Data Science for the Public Good	2020	1	12	6	https://dspg.iastate.edu

*Numbers are self-reported per university. Where collaborations have occurred, students and projects may be double counted.

Additional Data Science Organizations Consulted	Website
Georgetown University, McCourt School of Public Policy, Massive Data Institute/ Data Science for Public Policy	https://mccourt.georgetown.edu/research/the-massive-data-institute
research4impact	https://r4impact.org
The Ohio State University, Translational Data Analytics Institute	https://tdai.osu.edu
University of California Berkeley, Berkeley Institute for Data Science	https://bids.berkeley.edu
University of Michigan, Michigan Institute for Data Science	https://midas.umich.edu
Vanderbilt University, Vanderbilt Data Science Institute	https://www.vanderbilt.edu/datascience

APPENDIX 2

D4G Projects of Contributing Programs

This list was compiled by reviewing program websites as they appeared early 2021. One overarching category was assigned to each project. However, as suggested by the project titles, it is common for D4G projects to traverse several thematic areas.

Project	Year	University
Disaster Response		
Automatic damage annotation on post-hurricane satellite imagery	2018	University of Washington
Measuring disaster damage with tweets	2013	CMU (formerly at U. of Chicago)
Smarter crowdsourcing for crisis maps	2013	CMU (formerly at U. of Chicago)
Education		
Avenues of change - Early child education project	2017	University of British Columbia
Identifying factors that contribute to post-secondary school success	2019	University of Massachusetts Amherst
Access to out-of-school opportunities and student outcomes	2018	University of Washington
Analyzing impacts of the arts education program	2019	University of North Florida
Getting students into college	2013	CMU (formerly at U. of Chicago)
Identifying and influencing students at risk of not finishing high school	2016	CMU (formerly at U. of Chicago)
Identifying factors driving school dropout and improving the impact of social programs in El Salvador	2018	CMU (formerly at U. of Chicago)
Identifying high school students who may not graduate on time	2015	CMU (formerly at U. of Chicago)
Increasing graduation rates and improving college readiness for high school students	2014	CMU (formerly at U. of Chicago)
Mining insights from the adult learners educational data	2020	University of North Florida
Predicting college persistence among high school students	2015	CMU (formerly at U. of Chicago)
Predicting students that will struggle academically by third grade	2016	CMU (formerly at U. of Chicago)
Student enrollment prediction for budget allocation	2014	CMU (formerly at U. of Chicago)

Project	Year	University
Employment & Workforce		
Fairfax County labor markets: Characterizing local workforce and employment networks	2020	University of Virginia
Improving local labor market matching using high frequency resume and jobs data	2015	CMU (formerly at U. of Chicago)
Matching jobseekers with interventions to improve employment outcomes in Portugal	2019	CMU (formerly at U. of Chicago)
Modeling career pathways of veterans in the DC Metro area	2020	University of Virginia
Predicting long-term unemployment in Portugal	2018	CMU (formerly at U. of Chicago)
Predicting risk of long-term unemployment	2017	CMU (formerly at U. of Chicago)
Skilled Technical Workforce (STW) estimates for states x years (2010 to 2019) and benchmarking	2020	University of Virginia
Skilled technical workforce: Demand, supply, and pathways	2019	University of Virginia
Energy		
Building open source tools to analyze smart meter data	2014	CMU (formerly at U. of Chicago)
Developing NLP Tools for sharing of indigenous and community knowledge	2019	University of British Columbia
Natural Language Processing of letters of comment for pipeline applications	2020	University of British Columbia
Predicting building energy savings	2013	CMU (formerly at U. of Chicago)
Environment & Natural Resources		
Analysis of 25 years' worth of water-quality data collected by citizen scientists	2019	University of Massachusetts Amherst
Assessing the precision and accuracy of data collected by students	2019	University of North Florida
Building a network of land ownership in Kenya	2020	Stanford University
Data-driven digital engagement for environmental causes	2015	CMU (formerly at U. of Chicago)
Detecting animals in photographs	2019	University of Massachusetts Amherst
Developing a fishing risk framework from satellites and ocean data	2017	CMU (formerly at U. of Chicago)
Identifying CAFO characteristics using satellite imagery	2020	Stanford University
Increasing accessibility of biodiversity data in Metro Vancouver	2019	University of British Columbia
Measuring and predicting carbon emissions in Appalachian Mountain Club operations and facilities	2020	University of Massachusetts Amherst

Project	Year	University
Environment & Natural Resources		
Predictive enforcement of pollution and hazardous waste violations	2015	CMU (formerly at U. of Chicago)
Predictive enforcement of pollution and hazardous waste violations in New York State	2016	CMU (formerly at U. of Chicago)
Quantifying traffic dynamics to better estimate and reduce air pollution exposure in London	2019	CMU (formerly at U. of Chicago)
Strengthening capacities, knowledge and data sharing platforms for sustainable development	2017	University of Washington
Using sensor data to inform and evaluate environmental initiatives	2014	CMU (formerly at U. of Chicago)
Governance		
Anomalies' detection in public procurement processes	2019	University of Warwick
eiCompare: Making every vote count	2020	University of Washington
Identifying and analyzing corruption risks in public administration	2020	University of Warwick
Identifying fraud & collusion in international development projects	2015	CMU (formerly at U. of Chicago)
Improving government response to citizen requests online	2016	CMU (formerly at U. of Chicago)
Prediction & identification of collusion in international development projects	2014	CMU (formerly at U. of Chicago)
Reducing corruption in public procurement processes	2019	CMU (formerly at U. of Chicago)
Text analysis of government spending bills	2014	CMU (formerly at U. of Chicago)
Tracing policy ideas from lobbyists through state legislatures	2015	CMU (formerly at U. of Chicago)
Homelessness & Housing		
ADUniverse: Evaluating the feasibility of (affordable) accessory dwelling units in Seattle	2019	University of Washington
Changing homelessness - Creating profile of homelessness and shelter services	2017	University of North Florida
Improving outcomes for rough sleepers in the UK	2019	CMU (formerly at U. of Chicago)
Improving outcomes for rough sleepers through public reporting	2019	University of Warwick
Measuring the effectiveness of interventions on improving outcomes for homeless individuals	2014	CMU (formerly at U. of Chicago)
Proactive outreach to reduce harassment of NYC rental housing tenants	2018	CMU (formerly at U. of Chicago)

Project	Year	University
Human Services		
Data-driven prioritisation of independent fostering agency inspections	2019	CMU (formerly at U. of Chicago)
Data-driven prioritisation of independent fostering agency inspections	2019	University of Warwick
Enhancing the distribution of social services in Mexico	2016	CMU (formerly at U. of Chicago)
Family support services of North Florida – Patterns and trends in child welfare resource systems	2018	University of North Florida
Identifying frequent users of multiple public systems for more effective assistance	2016	CMU (formerly at U. of Chicago)
Improving early and middle childhood outcomes	2018	University of British Columbia
Pilot a 'Systems of Care' data infrastructure to support state prevention, treatment and safety response efforts	2020	Iowa State University
Risk assessing early years providers	2020	University of Warwick
Infrastructure		
Broadband data validation: Comparing U.S. broadband coverage	2019	University of Virginia
Early warning system for water infrastructure problems	2016	CMU (formerly at U. of Chicago)
Optimizing waste collection from portable sanitation in Kenya	2016	CMU (formerly at U. of Chicago)
Innovation		
Measuring the public funding of R&D: A feasibility study	2019	University of Virginia
RnD abstracts: Emerging topic identification	2020	University of Virginia
Incarceration & Criminal Justice		
Analyzing impacts of the school to prison pipeline program	2020	University of North Florida
Halifax County: Factors of incarceration and recidivism	2020	University of Virginia
Preventing juvenile interactions with the criminal justice system	2016	CMU (formerly at U. of Chicago)
Returning citizens re-entry program	2019	University of Virginia
Other		
Army performance measurement: Content and themes	2020	University of Virginia
Enlarge the ISU extension community helpline services	2020	Iowa State University

Project	Year	University
Other		
Identifying unlabeled objects in images of Pompeii frescoes	2020	University of Massachusetts Amherst
Improving predictions for targeted human trafficking investigations in Brazil	2020	Stanford University
Measuring community embeddedness near army installations: A feasibility study	2019	University of Virginia
Measuring the universe of Open Source Software (OSS)	2019	University of Virginia
Predicting YMCA membership churn	2019	University of Massachusetts Amherst
Reducing response times to citizen legal questions across Africa	2019	CMU (formerly at U. of Chicago)
Sectoring Open Source Software: Where do GitHub contributions come from?	2020	University of Virginia
The American soldier in World War II: Extracting insights from historical textual data	2020	University of Virginia
Philanthropy		
The giving graph- Grassroots philanthropy meets social networks	2013	CMU (formerly at U. of Chicago)
Planning & Development		
Assessing community well-being through open data and social media	2015	University of Washington
Assessing factors of economic mobility through a political capital lens	2020	University of Virginia
BC tourism resources project	2017	University of British Columbia
Develop a community capitals data infrastructure to support community economic mobility	2020	Iowa State University
Economic mobility baseline and comparative analysis for the South Wasco County School District Area, Oregon	2020	University of Virginia
Enhancing municipal planning forecasting	2019	University of Massachusetts Amherst
Evaluating residential property data quality	2020	University of Virginia
Identifying rooftop usage in Rotterdam	2017	CMU (formerly at U. of Chicago)
Improving long-term financial soundness by identifying causes of home abandonment in Mexico	2015	CMU (formerly at U. of Chicago)
Investment and intergovernmental project	2017	University of British Columbia
Predictive analytics for smarter city services	2013	CMU (formerly at U. of Chicago)

Project	Year	University
Planning & Development		
Proactive blight reduction and neighborhood revitalization	2015	CMU (formerly at U. of Chicago)
Sustainable tourism in Tuscany	2017	CMU (formerly at U. of Chicago)
Targeted approach to returning vacant land to productive use	2013	CMU (formerly at U. of Chicago)
Targeted urban investments to improve future economic outcomes	2014	CMU (formerly at U. of Chicago)
Uncovering the hidden universe of rental units in Surrey	2018	University of British Columbia
Public Health		
Baptist Health Y Healthy Living Centers – Addressing metabolic syndrome	2018	University of North Florida
Barriers to health care access and use in Patrick County, VA	2020	University of Virginia
Contributions of service combinations on healthy child outcomes	2019	University of North Florida
Detecting and linking pharmaceutical innovators in news articles	2020	University of Virginia
Detecting pharmaceutical innovations in news articles using machine learning	2019	University of Virginia
Developing health assessment scores at the city/town level	2019	University of Massachusetts Amherst
Examining opiate adverse events in minority populations	2019	Stanford University
Fairfax County CommunityScapes	2019	University of Virginia
Generating insights into patient care and treatment	2020	University of Massachusetts Amherst
Helping decision-makers to keep up to date with new research	2019	University of Warwick
Identify communities in greatest need of excessive alcohol prevention efforts	2020	Iowa State University
Identify communities ready and able to support substance use recovery centers	2020	Iowa State University
Improving social services interactions	2014	CMU (formerly at U. of Chicago)
Increasing the efficiency of creating meta-reviews in biomedical research	2019	CMU (formerly at U. of Chicago)
Increasing the efficiency of heart function assessment and diagnosis through echocardiography	2019	CMU (formerly at U. of Chicago)
Matchmaking between patients and doctors in a large healthcare network	2017	CMU (formerly at U. of Chicago)

Project	Year	University
Public Health		
Mayo Clinic Wellness Rx - Helping community to address health disparities	2017	University of North Florida
Mining online data for early identification of unsafe food products	2016	University of Washington
Natural Language Processing for peer support in online mental health communities	2019	University of Washington
Predicting and reducing adverse birth outcomes	2015	CMU (formerly at U. of Chicago)
Predicting platelet usage	2019	Stanford University
Predicting success in mother-child interventions	2014	CMU (formerly at U. of Chicago)
Predictive analytics to prevent lead poisoning in children	2014	CMU (formerly at U. of Chicago)
Reducing maternal mortality rates in Mexico	2014	CMU (formerly at U. of Chicago)
Reducing recidivism and improving outcomes for people with complex health needs	2018	CMU (formerly at U. of Chicago)
Supporting proactive diabetes screenings to improve health outcomes	2018	CMU (formerly at U. of Chicago)
Targeting the uninsured for health insurance enrollment	2014	CMU (formerly at U. of Chicago)
Tracking the impact of early childhood health programs	2013	CMU (formerly at U. of Chicago)
Understanding the impact of COVID-19 on the delivery of emergency medical services	2020	University of Virginia
Understanding the patterns of recidivism in mental health	2019	University of North Florida
Use of machine learning techniques to classify laboratory test results	2018	University of British Columbia
Use of machine learning techniques to classify laboratory test results (Phase 2)	2019	University of British Columbia
Using electronic medical records data to prevent cardiac arrests (Code Blue)	2013	CMU (formerly at U. of Chicago)
Yoga 4 Change - Analyzing impacts of yoga curriculum on stress and mood levels	2017	University of North Florida
Public Information		
Detecting and tracking online Covid misinformation	2020	University of Massachusetts Amherst
Identifying coronavirus disinformation risk on news websites	2020	University of Washington

Project	Year	University
Public Safety		
Building a deeper police early intervention system	2016	CMU (formerly at U. of Chicago)
Early intervention system for adverse police interactions	2015	CMU (formerly at U. of Chicago)
Economic and social impact of Arlington restaurant initiative	2019	University of Virginia
Ensemble forecasts for wildfire smoke	2020	University of British Columbia
Expanding our early intervention system for adverse police interactions	2016	CMU (formerly at U. of Chicago)
Improving workplace safety through proactive inspections	2018	CMU (formerly at U. of Chicago)
Improving outcomes for repeat/frequent 911 callers to emergency services	2019	CMU (formerly at U. of Chicago)
Optimizing the quality and delivery of city emergency medical services	2016	CMU (formerly at U. of Chicago)
Socio-economic Disparities		
Algorithmic Equity Toolkit	2019	University of Washington
Finding data-driven insights in the fight against hunger	2019	University of North Florida
Girls Incorporated of Jacksonville – Breaking the cycle of poverty	2018	University of North Florida
Identifying children and families with low-incomes and early learning needs	2020	University of North Florida
Identifying new opportunities for food bank donation from food service retail	2015	CMU (formerly at U. of Chicago)
Identifying skills gaps to reduce unemployment	2014	CMU (formerly at U. of Chicago)
The Performers Academy – Empowering at-risk youths through the arts	2018	University of North Florida
Transportation		
Can traffic sensor data detect vehicle cruising?	2017	University of Washington
Exploratory data analysis and visualization for Surrey's electric vehicle strategy and heavy-duty vehicle approach	2019	University of British Columbia
Global Open Sidewalks: Creating a shared open data layer and an OpenStreetMap data standard for sidewalks	2016	University of Washington
Improving incident response in the Netherlands	2017	CMU (formerly at U. of Chicago)
Improving traffic safety through video analysis	2018	CMU (formerly at U. of Chicago)
Improving transit services using ORCA data	2017	University of Washington
Open sidewalk graph for accessible trip planning	2015	University of Washington

Project	Year	University
Transportation		
Predicting crosswalk locations to enhance road safety and create equity	2020	University of British Columbia
Predicting when Divvy bike share stations will be empty or full	2013	CMU (formerly at U. of Chicago)
Rerouting solutions and expensive ride analysis for King County Paratransit	2015	University of Washington
Seattle Mobility Index Project	2018	University of Washington
Simulating better bus service	2013	CMU (formerly at U. of Chicago)
Surrey transportation project	2017	University of British Columbia
Transportation energy and emissions baseline and forecast for ongoing modelling and policy analysis	2018	University of British Columbia
Understanding and reducing inequities in transportation in the West Midlands	2019	University of Warwick & CMU (formerly at U. of Chicago)
Understanding congestion pricing, travel behavior, and price sensitivity	2019	University of Washington
Use of ORCA data for improved transit system planning and operation	2016	University of Washington

APPENDIX 3

2020 Survey of 8 D4G Programs Highlights

In early 2020 the University of Washington conducted a survey of eight Data for Good programs that were in operation as of 2019. The survey helped to shape the discussions among a broader set of contributors that led to the Data for Good Growth Map paper. This appendix provides survey highlights of interest to a broader audience. Programs represented in the survey include those hosted at the following universities: Carnegie Mellon University (formerly at the University of Chicago) (CMU), Stanford University, University of British Columbia (UBC), University of Massachusetts Amherst, University of North Florida (UNF), University of Virginia (UVA), University of Warwick, and University of Washington.

Student Participation

Student are paid
Yes (8 of 8)

Monthly pay
$1200 to $4200

Outside formal course offerings
Always (8 of 8)

Course credit
Yes (2) No (6)

Graduate student involvement
All (1) Most (5) Some (2)

Undergraduate involvement
Most (2) Some (5) None (1)

Students accepted from outside home university
Yes (5) No (3)

International students accepted
Yes (8 of 8)

Disciplinary Backgrounds	Many or Most	Some	None
Methodological Fields	7	1	
Social Sciences	4	3	1
Health Sciences	2	3	3
Humanities	2	3	3
Engineering	1	6	1
Natural & Physical Sciences	1	4	3
Other	1		4
Professional		5	3

Commonly, D4Gs were run as internships or fellowships. All programs paid students though pay range varied. The lowest at 1200 per month was for a half-time program. The highest paying program only accepted graduate students. Most D4G students across programs were at the graduate level, though only one program excluded undergraduates (UMASS Amherst). At two programs (UBC, UVA), most students were undergraduates. Programs attracted and selected students across the spectrum of disciplinary backgrounds. Seven of eight programs reported many or most students hailed from methodological fields such as computer sciences, stats, math, and data science. Social science students were the next most represented. The remaining disciplines of engineering, health sciences, humanities, natural and physical sciences and professional fields provided “some” students. Reflecting differences between programs some reported “many” students from health sciences and humanities, while others had no students from these backgrounds.

Priority Criteria for Selecting Students

	High or Medium	Low or N/A
Ability to work w/ diverse stakeholders	8	
Conduct during an interview	8	
Contributions to diversity	8	
Evidence of motivation	8	
Programming experience	8	
Career plan/trajectory	6	2
Disciplinary background	6	2
Experience w/ team-based project work	6	2
Math/statistical knowledge	6	2
Previous work on social good projects	6	2
Research experience	6	2
Student interests	6	2
GPA	4	4
Outside references	4	4
Skill set that matches project needs	4	4
Design experience	3	5
Other	2	6
Testing/skill evaluation	1	7

Top selection criteria across programs included: ability to work with diverse stakeholders, conduct during an interview, contributions to diversity, evidence of motivation, and programming experience.

Most programs also considered a student's career plan/trajectory, disciplinary background, experience with team-based project work, knowledge of math and statistics, previous work on social good projects, research experience, and student interests.

It was less common for programs to give medium-to-high consideration to a prospective students' design experience, outside references, skill evaluations, project-specific skills, GPA, or other criteria (elaborated on in free text response as "critical thinking ability", "public presentation skills, ability to handle team conflicts, and knowing how to ask for help to resolve problems"). Though programs differed in the criteria they prioritized, all 18 criteria were rated medium-to-high priority by at least one program.

Program Curricula

Program Curricula	n of 8
Domain area knowledge	8
Ethics	8
Professional/career development	8
Project management	8
Science communication/presentation	8
Team development/collaboration	8
Data science tools	7
Quantitative methods training	7
Version control/reproducibility	7
Computer programming	6
Qualitative methods training	6
Stakeholder engagement	6
Design	5
Other topics	2
Other research skills	

D4G curriculum topics touched upon a range of D4G skills and concepts. The high concurrence of 13 topics named in the survey in combination with the low number of "other" responses, suggests that the 13 topics named in the survey comprised the core curriculum topics across programs. Domain area knowledge, ethics, professional/career development, project management, science communication, and team development were addressed by all reporting programs. The least frequently covered topic (design) was covered by five of eight programs.

Technical Topics

Technical skills that were most commonly taught across the programs were: data visualization, data science libraries, exploratory data analysis, Git & GitHub, and machine learning.

Technical skills reported as least commonly currently taught were object-oriented programming, web scraping, web design/apps, unit testing, high performance computing, and database design.

All 23 technical topics named in the survey were covered by at least one program. Yet, no single technical topic was planned to be covered in the next session by all eight programs. Likewise, 13 skills taught in previous sessions of a program would not be taught by the same program in next session.

This variability among the technical topics addressed within and across programs likely reflects differences in the work demands between different projects as well as differences in level of experience among students accepted into different programs.

Technical Topics	n of 8
Data visualization	7
Data science libraries	6
Exploratory data analysis	6
Git & GitHub	6
Machine learning	6
Cloud computing	5
Coding standards & best practices	5
Data privacy & security	5
Documentation practices	5
GIS/geospatial tools	5
Pipelines & computational workflows	5
Other programming languages	4
Pair programming	4
SQL	4
Python	3
R	3
Software design	3
Database design	2
High performance computing	2
Unit testing	2
Web design/apps	2
Web scraping	2
Object-oriented programming	1

Project Selection and Support

Project recruitment	Frequent	Sometimes	Never
Targeted solicitation	6	2	
Co-develop with partners	5	2	1
Open call	4	1	3
Develop in-house	3		5
Other		1	3

Programs used markedly different strategies to recruit projects. For example four programs frequently held an open call while three said they never did so. Likewise, five of eight programs reported never developing projects in house, while three reported frequently doing so.

Who drives project ideas?	Frequent	Sometimes	Never
Government agencies	7	1	
Non-profits	5	3	
Academics	4	3	1
Private companies	1	2	5
Other			2
Philanthropists		4	2

Government agencies, nonprofits, and academics where reported as the most common drivers or contributors to project ideas. Though some programs had explicit commitments to work with government agencies and others did not, government agencies stood out as common drivers for project ideas across programs.

Types of support provided to partners	Frequent	Sometimes	Never
Building an infrastructure/pipeline	8		
Communicating work	8		
Prepping data	8		
Data analysis	7		
Articulate research questions	5	3	
IP data sources	4	2	2
Project design	3	5	
Integrating outputs	2	5	1
Academic publications/presentations	1	7	

Programs most commonly supported project partners by: helping to communicate the work; data analysis/modeling/interpretation; prepping data; and building a data infrastructure/pipeline. Less frequently, D4G programs helped partners identify data sources, articulate research questions, integrate outputs, design a project, and produce academic publications and presentations.