University of Pittsburgh

Report of the Data Science Task Force

November 30, 2020

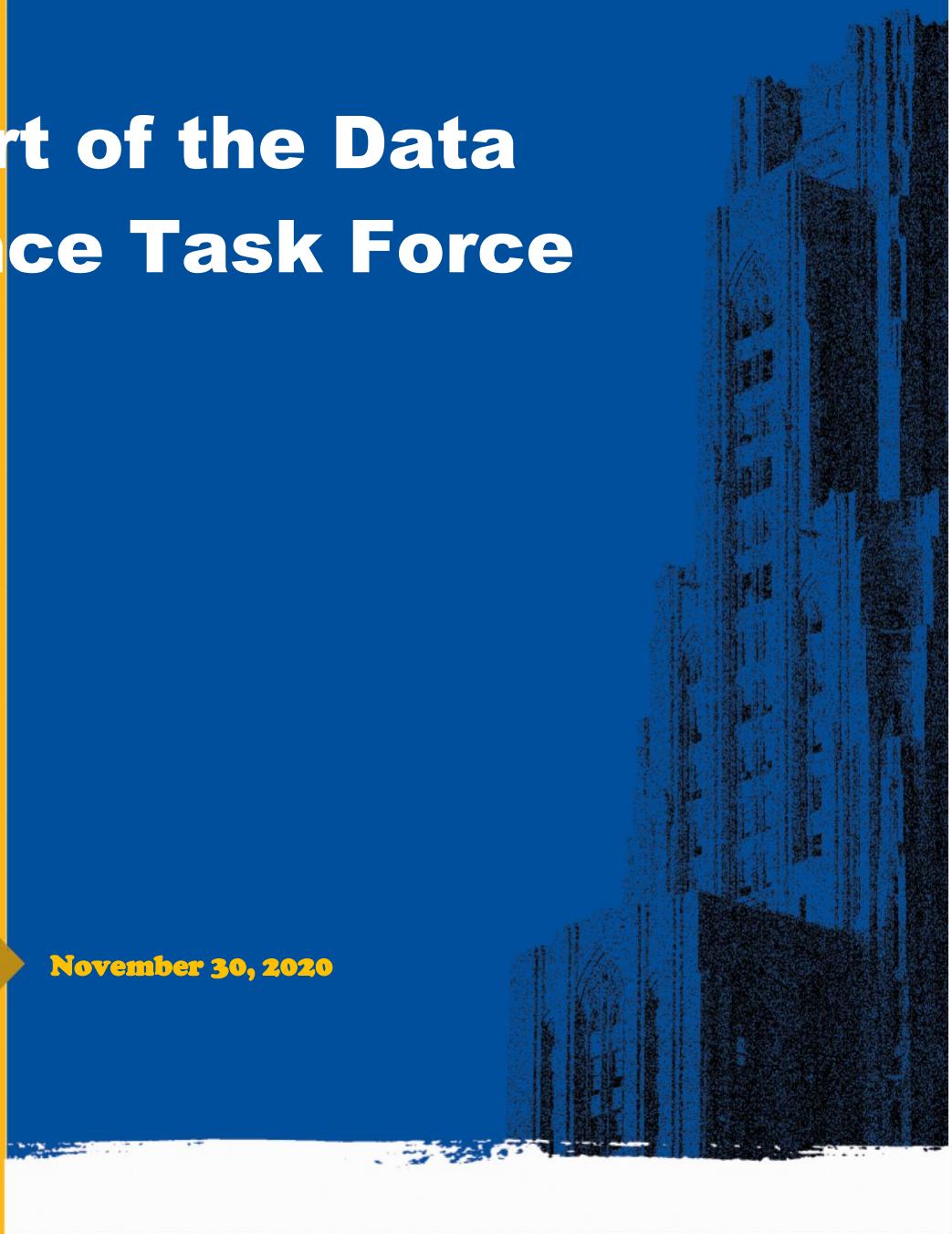

Table of Contents

Executive Summary

The Data Science Task Force considered how to advance data science at Pitt. One opportunity is to think very big and build a large presence and a strong brand in data science. A more modest opportunity is to create structures that build on the strengths that are here and provide synergy across the structures, while gradually expanding our capabilities, especially to improve sharing and accessibility to data, tools, curriculum, and expertise. Improved visibility of existing effort is not necessarily a big-ticket item and can be done in either context. While data science initiatives are already well established at many institutions, it is not too late to be a leader in this area. The task force identified several areas in which Pitt has the opportunity to step out of the pack and take a national leadership position:

1) Equal and integrated emphasis of both data science methods development (statistics, information science, etc.) and innovative discipline-focused uses of data.
2) Emphasis on ethics, social responsibility, and social impact of data use.
3) Methods and use of non-sampled or non-scientific data, e.g., full-capture of consumer data such as social media data, consumer purchasing data, and geo-tracking data.
4) Capitalize on Pitt's health sciences strength in building a brand in data science.

To work toward a leadership position at the confluence of these areas of *responsible data science*, the DSTF recommends four overarching goals and a set of actions for each goal. The actions are listed from relatively small and immediate ones to larger and longer-term transformation, and are described in greater detail in the subsequent sections of this report.

Goal 1: Create shared understanding.
Increase the reputation, visibility, and awareness of responsible data science within and outside the Pitt community. Create a shared and unified understanding of data science and of its importance across disciplines.

> Action 1: Establish a group of "data science liaisons" to ensure diverse representation of faculty, staff, students, and alumni to form an initial community to seed, welcome, nourish, and mentor the growth of a larger, inclusive community of individuals using, critiquing, governing, and regulating data science, as well as individuals who have curiosity, new interest, or need for expertise with data, but are finding impediments to doing so. (short term)
>
> Action 2: Establish a regular University-wide/Provost "Distinguished Data Scientist Lecture Series" to invite distinguished visitors to advise and speak on uses and methods, ethics, laws, and critique of data and data methods. (short term)
>
> Action 3: Unify, market, and communicate "Data Science Success and Opportunity" that reinforces a message of responsible, use-driven data science. (short to medium term)

Action 4: Create and continuously update a "Data@Pitt" online web hub to aggregate and disseminate opportunities, success stories, events, activities, education pathways, and initiatives related to data science. (short to medium term)

Goal 2: Require fluency and knowledge.
Require every undergraduate student to acquire a basic understanding of data and data methods, including considerations of responsibility, as part of their learning at Pitt.

Action 5: Mandate that every school with an undergraduate program develop inclusive curriculum, coupled to practical experience with actual datasets, questions, methods and tools, that provides all undergraduates with preparation in data concepts and skills. (medium term)

Goal 3: Catalyze skill acquisition.
Create, support, and incentivize inclusive, flexible undergraduate and graduate educational programs and shared educational resources to offer training in data science – in context of a broad variety of domains – to students, postdocs, staff, and faculty.

Action 6: Identify gaps in existing curriculum, develop a set of shared educational resources for these gaps, and provide a central repository of curricular materials at both the undergraduate and graduate levels. (short to medium term)

Action 7: Establish and/or charge an organizational entity to coordinate training and education, assessment, development of curriculum, collecting and disseminating project opportunities, courses and course content materials. (medium to long term)

Goal 4: Coordinate strategy and action.
Implement a structure that (i) knits together, in a visible, accessible, and central place, people and practices in data science; (ii) serves as an evolving source of knowledge in developing and applying responsible data science to overcome diverse, challenging problems, including ethics, policy, and legal aspects; and (iii) animates extraordinary ambitions and success in collaborations transcending disciplinary and community limitations.

Action 8: Establish a dedicated, full-time position and charge a leader with a mandate to advance and coordinate data science for Pitt. (short to medium term)

Action 9: Use the Pitt Momentum or other funding mechanism to encourage, initiate, and support action on the highest impact actions in this report and work toward the development of a polished concept, with pilot implementation, that can be a springboard for a major gift. (short term)

Action 10: Create and support an institutional structure – a "coordination tower" – to coordinate existing and emerging elements (layers) of data science. (long term)

1 Motivation and Goals for the Task Force Work

The data revolution has brought about powerful capabilities to gather, retrieve, analyze, visualize, and communicate about data at unprecedent scales in near real time, which has led to new insights, discoveries, and ways of working and living. This revolution has been far-reaching to epistemological and pragmatic change in nearly every field and facet of life, whether in education, government, the natural sciences, medicine, engineering, entertainment, social science, the arts, humanities, or business. While applications of data science have many benefits, they have also brought new actual and potential harms and digital divides, including but not limited to emerging challenges related to disinformation, privacy, algorithmic bias, and the future of work. Our ability as a community and a society to fulfill the promise of data science, while avoiding the emerging pitfalls, will be navigated by our students, faculty, and staff. Empowering students with data skills, knowledge, and ethics, and faculty and staff with the necessary infrastructure to collect, steward, retrieve, and analyze data is vital to the educational, research, and outreach missions of Pitt.

The Data Science Task Force (DSTF) was formed to answer the question of "how should the Pitt academic community collectively act on the urgent need, given our context, strengths, and individual efforts in data-related areas?" The specific charge (cf. Appendix) was to recommend a coordinated strategy to catalyze, nourish, and sustain educational programs and research initiatives that (1) equip undergraduate and graduate students with the knowledge and skills necessary for the increasingly data-oriented world; (2) develop and use data science methods in research; and (3) attract and retain faculty using data and associated methods in their disciplines.

The DSTF sees an opportunity to "think big" to build a large presence and a strong brand in data science. Most of the ingredients are in place to unleash a transformation that synergistically attends to and accelerates research and education using, developing, or examining data and data-oriented methods throughout the University of Pittsburgh. What is required is to forge an *institutional focal point* of data science that would bring together, catalyze, and support a community of researchers, educators, and practitioners who see opportunity and want to use data and data methods in their work, as well as those individuals who are already developing or applying data techniques.

1.1 A Conceptual Structure for Data Science at the University of Pittsburgh

Figure 1 expresses that Pitt has a diverse community of expertise across many domains over the breadth of the campus community (bottom layer) in data and data methods. However, too few teams have put together, or know where to find within the university, the comprehensive data science pipelines of identifying societal challenges and collecting, organizing, analyzing, interpreting, and communicating necessary information to offer solutions to those challenges (upper layers). A focal point is necessary to coordinate and communicate practices and between people across campus. While every scholarship project (vertical arrows) is different, there are

common data science lessons and innovations that can accelerate breakthroughs in a range of fields.

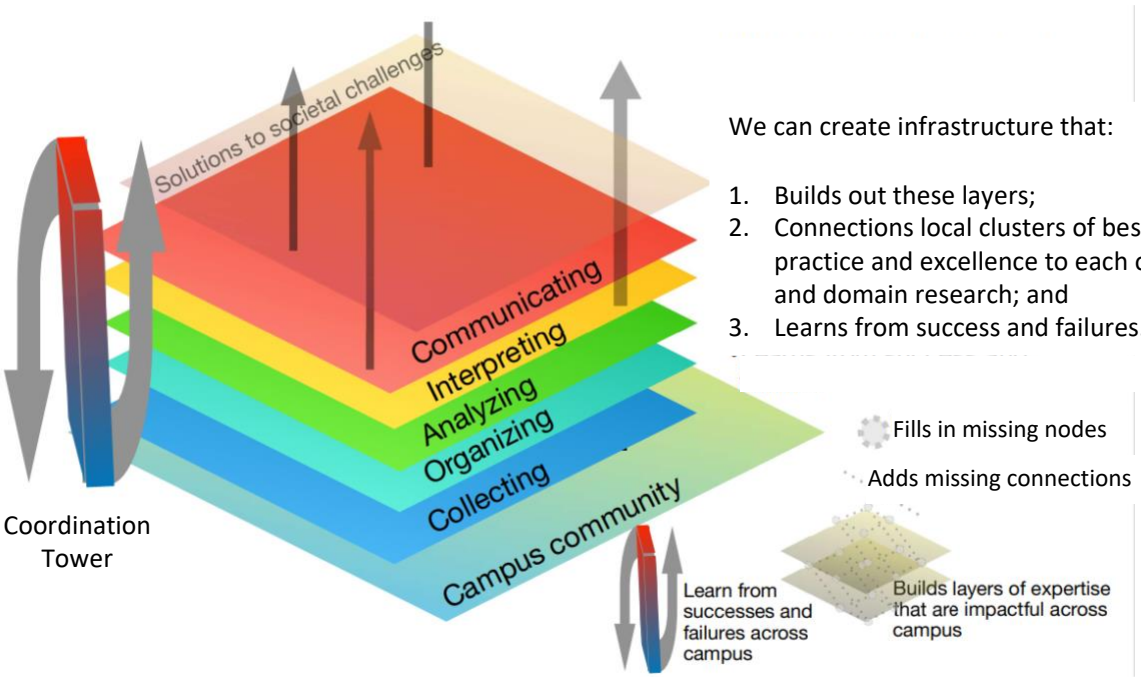

Figure 1: Institutional focal point for data science at Pitt

The broad area of each layer represents that innovative methods for data science will have the greatest impact when shared across the university, not simply existing in one domain silo or a silo of their own. The suggested transformation aims to build out these layers (multi-colored for each suggested step), connecting local clusters of excellences to each other and domain research, evolving over time. Here, research laboratories and additional university resources have multi-dimensional expertise. Some specialize in both collecting and organizing specific types of information (e.g., data in astronomy or public health), while others offer solutions to the analysis and building of models as well as interpretation.

An institutional focal point of data science could explicitly enable these layers of expertise across campus, with the aim of facilitating researchers, teachers, and other data science experts in connecting within each layer and across layers to accelerate and increase the impact of their work. Teams within each layer will learn from the successes and failures of specific data science applications to continuously adjust and evolve the infrastructure to new technological developments, inquiries of discovery, community needs, and unforeseen obstacles. Teams across layers will build out specific research projects or "thrusts."

The vertical block to the left of the stack of layers represents an essential "coordination tower" function (akin to an airport's control tower) and organization to create the focal point. The purpose of the "coordination tower" is to document, coordinate, and energize data science

research, teaching, and engagement within layers and across them. Pitt's data science focal point should embody the ambition to build an evolving and resilient fueling and talent center of data initiatives. The DSTF conducted an informal benchmark analysis of 41 data science initiatives at other institutions (cf. Appendix). Among these efforts, only 7 appeared to take a similarly comprehensive and holistic approach in structure as suggested by the DSTF, incorporating doctoral, masters and undergraduate education and research activities[1]. This may represent an opportunity for Pitt to step out front and to act big across the institution. The time is short, though, as many of the initiatives beyond the 7 identified as already comprehensive are moving in this direction. At the very least, the analysis suggests that Pitt should consider quick action.

1.2 Human and Institutional Resources

The creation of a focal point in data science will require an investment of resources. The DSTF recognizes the challenges faced by the University at the present time due to the COVID-19 pandemic response. A more modest opportunity also exists to create initiatives that provide awareness and coordination of current programs, while gradually expanding our capabilities, especially to improve sharing and accessibility to data, tools, curriculum, and expertise. Improved visibility of existing effort is not necessarily a big-ticket item and can be done in either context.

Over the course of several meetings, individual discussions, and group activities, the DSTF developed a provisional *framework of goals and actions* that could bring about the transformation to create the institutional focal point. The framework is focused on *human resources and nourishing a community* of data science education and research rather than computational and storage infrastructure, institutional data use, specific research topics, or curricular choices. While instantiations of these elements will be needed, **the DSTF believes that "people" (including students, postdoctoral fellows, staff, faculty, and alumni) are *essential* to coordinating and growing Pitt's opportunities, visibility, and capabilities in data science.**

Effort underway by Pitt IT and Pitt Research will support researchers and educators using data science methods. Pitt IT is facilitating governance, sharing and use of institutional data that could open up new research, educational, and operational avenues. The effort by Pitt Research to coordinate and integrate storage, computation, and application services could dramatically ease and reduce cost of access to resources in support of a data science community. Acting on the DSTF's recommendations in concert with the initiatives by Pitt IT and Pitt Research, the University would have a comprehensive approach to transformation that includes all of the necessary ingredients – the people, data, and infrastructure – for the focal point recommended by the DSTF. Alignment to the Plan for Pitt would further ensure strategic action in support of the University.

[1] These institutions were identified by examining their structure. They have undergraduate, masters, PhD curriculum, as well as perhaps a dedicated center and designated leadership. Independent of their structure and programmatic activities, the issues addressed by data science institutes are often interdisciplinary in their nature.

2 Task Force Consensus on Guiding Principles

2.1 The Need for Immediate Action

The Data Science Task Force made several observations that motivate the need and the recommended goals and actions for transformation. First and foremost, an immediate response is needed. Pitt must act quickly to ensure that societal needs do not overtake our actions to coordinate data activities even before we start. It is not too late but if we want students to continue to come to Pitt, we need to quickly act on this proposed framework and send signals now that we are investing and innovating in data science skills, knowledge, ethics, and applications. Workforce development in the Pittsburgh region and nationally is a driver: numerous reports indicate that data is affecting every aspect of the economy with up to 2.7 million job postings related to data science and analytics in 2020[2]. Another driver is civic responsibility: false and misleading information (e.g., deepfake videos) can be easily created and disseminated on the Internet, and an understanding of data methods is increasingly necessary to sort through fact, fiction, and the "in between" to make informed decisions. Citizens are called upon to make decisions with cascades of complex and sometime contradictory data every day. Pitt must be a place where our students and community helps to arrange the pixels of the future in productive and positive ways, as opposed to being trapped in front of screens, struggling to understand and unable to turn data into insights for themselves and others. The Pitt community needs to increase support of how data and associated methods have become critical to the creation of new knowledge and discovery, which are fundamental to any research university.

A survey instrument was developed by the Data Science Task Force to learn the current prevalence and trends in data science for educational programs and research at Pitt[3]. The survey was sent to department chairs in all schools, including the health sciences. Supporting employment trends, 89% of the department chair respondents indicated that skills and knowledge in data science are *significantly increasing* (28 of 64 valid responses) or *increasing* in importance (29) for employers. Only 11% indicated a *slight increase* (4) or *no change* (3)[4].

[2] Investing in America's Data Science and Analytics Talent - The Case for Action, Business Higher Education Forum, April 2017.

[3] Individuals responding include Department Chairs and Associate Deans: 19 from the Dietrich School of Arts and Sciences; 5 from the Graduate School of Public Health; 1 from the Graduate School of Public and International Affairs; 1 from the Katz Graduate School of Business; 2 from the School of Computing and Information; 1 from the School of Dental Medicine; 2 from the Swanson School of Engineering; 2 from the School of Education; 5 from the School of Health and Rehabilitation Sciences; 1 from the School of Law; 17 from the School of Medicine; 1 from the School of Nursing; 2 from the School of Pharmacy; 1 from the School of Social Work; and 4 from University Centers.

[4] Selection bias is probable – i.e., it is expected that data-oriented programs were more likely to respond. Even with this bias, 57 departments that have significantly increasing or increasing importance is quite compelling.

2.2 Focus on Responsible Data Science

The DSTF recommends connecting "use", "responsibility", and "data", into a framing of *responsible, use-driven data science*[5] that provides societal benefit. In particular, we noted that data methods at Pitt are widely adopted for a specific purpose, or *use.* COVID-19 is a compelling example: data and modeling have been used to inform policy decisions and resource planning, make discoveries about the virus and its behavior, categorize the credibility of information about the pandemic, automatically remind about social distancing, counteract the spread of misinformation about the virus, and enable contact tracing. Typically, "use comes first" to drive data collection and inquiry, rather than the development of tools and the creation of new methods (e.g., types of machine learning). Tools and methods are commonly created for an intended purpose, or a family of purposes, by researchers at Pitt, although there is important theoretical and tool development underway as well. Indeed, the coupling of method and purpose is often at play no matter the form of scholarship, whether discovery, critique (e.g., the Mellon Sawyer Seminar[6] has critically examined the intersection of data and society), learning, or translation.

Many uses of data science have ethical, legal, and policy considerations, or dimensions of *responsibility*, e.g., consider the implications of privacy, confidentiality, and accessibility of mobility and health information for contact tracing during the pandemic[7]. Ethical and legal issues about how to ensure appropriate deployment and protections abound, including what we *should* and *should not* do as individuals and a society (e.g., IBM deciding not to pursue facial recognition as a business due to potential human rights and privacy abuses[8], and whether/how government regulates such technologies, like San Francisco's ban[9]). There are also considerations of transparency, interpretability, bias, and accountability that interact with methods, datasets, and model use. For example, ProPublica's well-known reporting in 2016 about data science tools used to predict risk scores for recidivism revealed racial bias in machine learning models[10]. The models examined in the ProPublica study were hidden behind a cloak of intellectual property. The survey of department chairs also supports this observation: at both the undergraduate and graduate levels, respondents indicated that *responsibility and ethics of data and data methods* were

[5] Other institutions also recognize the need to integrate responsibility and ethics with method and application in their data science initiatives, e.g., University of Virginia's School of Data Science.

[6] Information Ecosystems: A Mellon Sawyer Seminar, Digital / Critical Interdisciplinary Methods, University of Pittsburgh, 2019-2020. Web site: https://sites.haa.pitt.edu/digitalcriticalmethods/sawyer-seminar-ay-2019-2020/

[7] Privacy and Ethics Recommendations for Computing Applications Developed to Mitigate COVID-19, NSCAI Commissioners Eric Horvitz, Mignon Clyburn, Jose-Marie Griffiths, and Jason Matheny; National Security Commission on Artificial Intelligence, May 6, 2020.

[8] Arvind Krishna, IBM Chief Executive Officer. IBM CEO's Letter to Congress on Racial Justice Reform. June 8, 2020.

[9] Kate Conger, Richard Fausset and Serge F. Kovaleski. San Francisco Bans Facial Recognition Technology. New York Times. May 14, 2019.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. May 23, 2016.

among the top two areas of increasing importance to their educational programs (most important for undergraduates and second most important for graduate). It is also an area where the programs at Pitt have, on average, less educational content.

Using algorithmically-derived decisions and expecting *understandable* (interpretable, transparent) explanations of outcomes requires the support of methods, legal structures, and the models themselves. Students will need to learn and value these issues to be well-informed citizens, to follow through on their civic responsibility, and increasingly in their own work. For example, big data models (e.g., deep neural networks) are widely deployed to guide or make decisions, often autonomously. These models may be "roll your own", or possibly procured from an external source. As a result, there can be a tendency to focus on "the specs" of particular algorithms or software applications rather than the *outcomes* from those algorithms or applications, i.e., validation and audit. This can lead to an underappreciation, or outright ignoring, important issues, like bias, during procurement, and scapegoating of outcomes during deployment ("the algorithm said it"). It can even lead to subtly and inherently "handing off" policy decisions to external, unregulated entities by procuring opaque algorithms. With the wide use of data science – in education, government and commercial sectors – the *need to critically understand, value, and act with responsibility* is of the utmost importance, much less the need to design and develop such datasets, tools, and methods with responsibility at the forefront.

2.3 Build on the Breadth of Current Data Science Activities at Pitt

A third motivation for the DSTF's recommendations concerns how much data science is present at Pitt. Through an environmental scan of degree programs conducted by the DSTF and the survey of department chairs, it is apparent that a wealth of expertise exists throughout campus, including curriculum, research, and infrastructure. Data science has percolated into many areas; for example, the DSTF scanned undergraduate and Master programs, and identified 40 that are "data oriented"[11]. Figure 2 shows an informal characterization of the programs and the nature of their data methods. This figure is based on qualitative examination of degree requirements using the "data acumen areas" from the NASEM 2018 report on data science education[12]. The bars in this figure show the amount of emphasis in each area of the NASEM report on a scale of 1 (minimum) to 4 (high). The height of each bar is the total emphasis among the 10 areas of data acumen contained in the NASEM report. The graph considers only *required* courses for a degree program; there are also numerous elective paths in many programs that also provide thorough training in data science. As the figure shows, curriculum expertise is broad (i.e., the range of schools/degrees) and deep within specific areas (i.e., the range of the aggregated characterization).

[11] Due to time constraints, the DSTF focused on undergraduate and Master programs and partially relied on the 2019-2020 Pitt Undergraduate and Graduate Course Catalogs. The group focused on programs that had at least four required courses on topics drawn from the categories of the NASEM 2018 report. Further work is needed to consider programs with many electives, PhD and joint degrees, and to validate these results with department chairs.

[12] National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104

The survey of department chairs also supports this observation: 28 undergraduate, 48 graduate, and 9 non-credit educational programs expect to teach students about data and data methods in the next five years. Among the 52 responses about research, 48 respondents indicated that data and data methods are *important* or *significantly important* to their research programs today, and 44 indicated that data and data methods will further *increase* or *significantly increase* in importance to research over the next 5 years. Within this group of responses, 23 respondents use data methods, 3 develop data methods, and 26 both use and develop data methods. Interestingly, the results reveal that many departments – beyond the ones most closely associated with data science, e.g., statistics, informatics, biostatistics, mathematics, and computer science – are not only applying but also *innovating* data science methods.

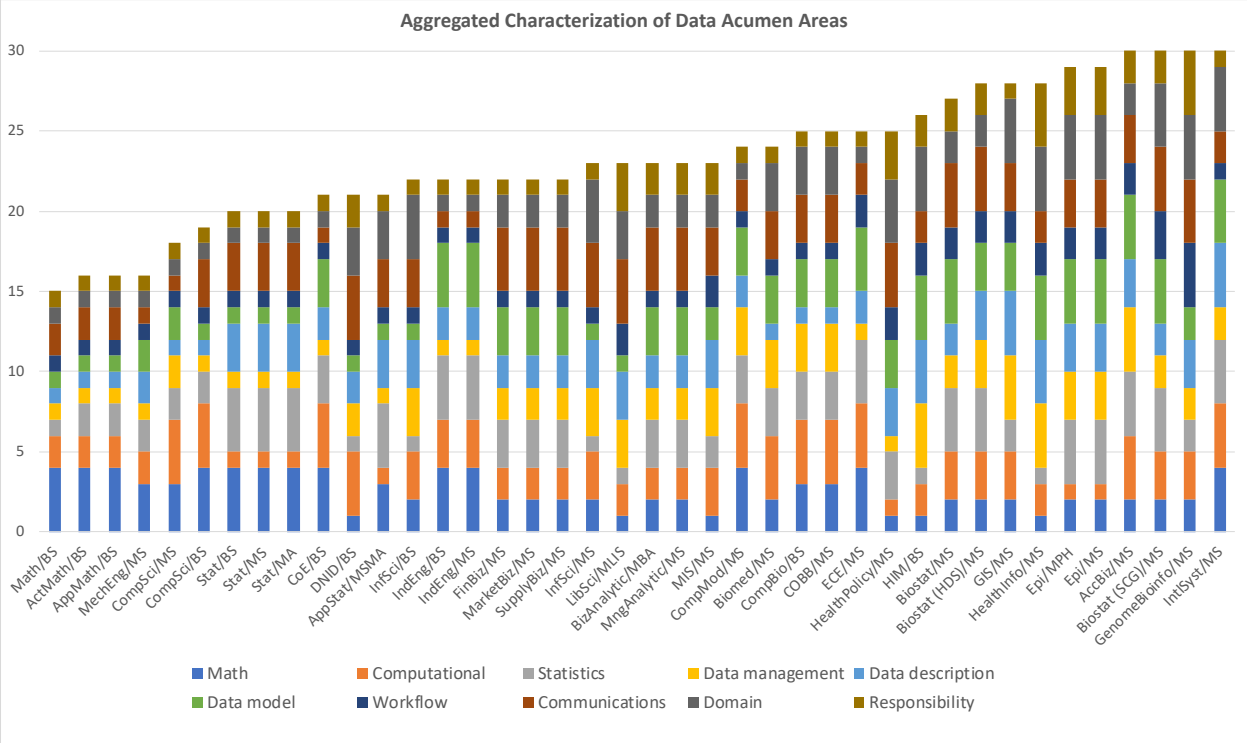

Figure 2: Data Oriented Undergraduate and Graduate Programs at Pitt

Taken together, the data from Figure 2 and the survey data show that a large portion of responding departments have a data orientation in education and research, which is expected to increase in the future. Although the DSTF did not specifically align the survey responses to the environmental scan, it is also worth noting that the aggregate number of programs from the survey that expect to teach data or data methods is 72, which is more than the number of programs identified as "data oriented" by the scan. This further hints at considerable expertise, need and interest in data science throughout departments of different schools. There are also related programs and initiatives by the University Library System (ULS), the Health Sciences Library System (HSLS), the Center for Research Computing (CRC), the University Center for Social

and Urban Research (UCSUR), and the Learning and Research Development Center (LRDC) which were not part of the department chair survey.

Despite the widespread nature of data science at the University, the DSTF noted that many were not aware of the activities. More challenging, barriers to entry are impeding some disciplines from accessing even adjacent data science expertise: we anecdotally discussed and heard about significant interest but also sensed that many researchers and educators are unsure of how to engage with data and data methods, even though they want to do so. Faculty have ideas for ground-breaking uses and innovations with data and associated techniques, but the existence of domain silos and the absence of visible and legible scaffolding has been an obstacle to many of these reputation-enhancing projects. There is opportunity to leverage and share knowledge given the range of expertise and interest. In essence, many puzzle pieces exist throughout campus, and they should be put together into a comprehensive framework to bring about a focal point of data science. Subgraphs of related expertise need to be intentionally connected between disciplines, illuminating new pathways of data science uses and breakthroughs. This observation reinforces the charge put forth by the Provost: we found that several capabilities exist that could be put together to make the sum of the parts greater than the whole.

2.4 Human Resources will be Key to Success

Finally, not only do data capabilities exist on campus, but the interest of researchers, educators, and practitioners to forge a community, place, and organization to act with intention and coordination also exists to leverage and amplify data science. This level of interest could be powerful for recruitment, retention, and development of faculty, students, and staff. The framework recommended by the DSTF reflects this observation: much of what we propose involves supporting and developing human resources, and gathering them into a "community of the eager". One respondent to the survey of department chairs wrote, "within one school our bench is small and we would love opportunities to collaborate with other departments who have other skills." From the DSTF's own discussions and social networks, this seems to be a commonly held view. In fact, within the Data Science Task Force itself, several members expressed interest to start working on the recommendations right away, such as developing shared curriculum materials or putting together a campus-wide lecture series. While there is grassroots interest, a mandate and support from Pitt's senior leadership, an organizational structure, and a strong leader will further champion and connect the community for the most impact.

3 Detailed Recommendations

To act on the opportunity in front of Pitt and to undertake the necessary transformation, the DSTF recommends a framework of goals and actions to coordinate and grow data science across campus into an institutional focal point. The framework is based on the DSTF's observations, background study, and the committee's expertise, knowledge, and experiences. The goals provide direction, and the actions lay out tangible and specific steps that can be undertaken, building toward the opportunity. Although the actions are organized to incrementally create a focal point of data science at the University, an alternative strategy could start first with hiring

and charging a leader to nourish and direct effort for the actions. Several of the actions can be carried out simultaneously. Others can be done right away to address the immediacy of need.

During its work, the DSTF encountered numerous definitions of data science – some put emphasis on "data" and its transformation into knowledge; others draw on the intersection of disciplines; and still others consider the "data pipeline"[13]. With such ambiguity, the DSTF suggests a definition to guide the implementation of the recommendations. This definition emphasizes discovery, exploration, and knowledge creation, i.e., the *uses* of data and data methods:

> *Data Science draws upon multiple disciplines to develop or apply computational and analytic methods to responsibly explore data and discover actionable insights and accelerate translation into implementation with real-world evidence.*

3.1 Goal 1: Create shared understanding.

> **Increase the reputation, visibility, and awareness of responsible data science within and outside the Pitt community, create a shared and unified understanding of data science, and for its importance across disciplines.**

Following from the DSTF's definition of data science, this goal aims to create shared understanding of data science to increase reputation, visibility, and awareness with internal and external stakeholders. Acting on the goal will lead to a better appreciation and knowledge of how data science presents itself at Pitt in research and educational programs. It will crisply define to internal and external stakeholders "responsible use-driven data science" and how it manifests at Pitt in our scholarship of all forms. The goal will help students identify programs that best align with their aspirations, and it will clarify differences among choices of curricula. It will also create awareness of opportunities and resources at Pitt, which will be beneficial to recruit and retain faculty whose scholarship involves data. It will also help connect data-oriented education and research to industry and corporate opportunities, especially data-driven startups.

Action #1 (short term): Establish a group of "data science liaisons" to ensure diverse representation of faculty, staff, students, and alumni to form an initial community to seed, welcome, nourish, and mentor the growth of a larger, inclusive community of individuals using, critiquing, governing, and regulating data science, as well as individuals who have curiosity, new interest, or need for expertise with data, but are finding impediments to doing so.

We propose forming a group of stakeholders from the university community to actively raise the visibility of responsible, use-driven data science inside and outside of Pitt by communicating our data science success stories, the challenges with data and how they were overcome, and explaining the importance of data and data methods in education and research. The liaisons should be formed to ensure diverse and equitable representation and to be purposefully drawn

[13] Victoria Stodden. The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science. Communications of the ACM, July 2020, Vol. 63 No. 7, pages 58-66. DOI: 10.1145/3360646

from many varied parts of the university and include student, staff, faculty and alumni members. The liaisons would call on their professional networks to raise awareness, and support and mentor others in doing the same (e.g., through Pitt Commons). Importantly, they would lead "by doing" in their own data-fueled initiatives to point out the available resources, challenges, open questions, and opportunities from which many faculty and students can benefit, learn, and contribute.

The liaisons would form an initial "community of the eager" to seed and nourish the growth of a community of individuals using, critiquing, governing, and regulating data science, as well as individuals who have curiosity, new interest, or need for expertise with data, but are finding impediments to doing so. This group could also serve to advocate for ongoing enhancement of inclusive and equitable teaching practices of data and data methods and support evidence-informed assessment of such practices.

Some incentives and recognition structures could help the liaisons group to develop and maintain cohesion, including informal ones like helping to select and host distinguished speakers for a data science lecture series (cf. next action). Raising the profile of the liaisons' work will help grow data science, and bring others into data science. It might be useful to provide investments in liaisons' labs in return for a commitment to run outward facing hands-on workshops on relevant data science tools and infrastructure as well as how they overcame previous obstacles. Artifacts such as Gists, Jupyter notebooks, code, best practices, datasets, challenge issues, and policies can be hosted and collected for future asynchronous use. Over time, the liaisons could evolve into a formal structure to support the institutional framework envisioned for Goal 4.

Action #2: Establish a regular University-wide/Provost "Distinguished Data Scientist Lecture Series" to invite distinguished visitors to advise and speak on uses and methods, ethics, laws, and critique of data and data methods.

The DSTF proposes a speaker series to *inform* and *form* a community of data science. The series could bring distinguished educators, researchers, and practitioners to the Pitt campus to advise and speak on the uses, ethics, laws, and critique of data and methods (e.g., Kate Crawford from the AI Now Institute; Cathy O'Neil, author of "Weapons of Math Destruction"), including potential harms and dangers of marginalization from data. We suggest a focus on general and timely topics to specially attract attendees who are new to responsible, use-driven data science. With appropriate advertising, hot topics, and speakers, the series could become "*the place and the time*" to learn and network on data for all disciplines. In addition to members of the Pitt community, we recommend involving alumni, local businesses, startups, and user groups (e.g., the Pittsburgh Tableau user group and tech meetup groups).

Initially, the data science liaisons could be the group to organize and host the series. They could seek speaker nominations from campus to foster diversity, engagement and awareness. To build relationships and knowledge, the series could be organized around a series of activities with each visitor, including a talk, discussion with the speaker, and small group learning roundtables on specific topics. Lastly, while an on-campus visit would be much preferred, we've learned from

the pandemic that virtual talks can lead to extremely high participation rates. The series could be virtual, with a hybrid approach to encourage participation and to benefit from in-person interaction. Using Zoom during this academic year to start the series could quickly establish an exciting and unique series to bring in a large general audience on campus and beyond in the Pittsburgh area.

Action #3 (short to medium term): Unify, market, and communicate "Data Science Success and Opportunity" that reinforces a message of responsible, use-driven data science.

This action will bring together data science activities into a consistent message that can be broadcast through "one megaphone". The message would strengthen a unifying theme and message of responsible, use-driven data science. The megaphone would be a communication and marketing plan to explain and inform students, faculty, and others about the many data science activities at Pitt. For example, training and degree programs, activities (e.g., the Distinguished Data Science Lecture series), and research outcomes could be advertised to recruit students, faculty, researchers, and others with interest and curiosity about data methods.

Specific messages could include guidance to potential and current students about Pitt's many education programs, and how those programs align to students' career aspirations and goals. A high school student that wants to create new data methods, for example, might be informed about the nascent data science undergraduate major from Statistics, Mathematics, and Computer Science. A student interested in astronomical survey and data analysis could be directed to the Department of Physics and Astronomy, or a student interested in data methods and economics might be pointed to the Master of Science in Quantitative Economics. A graduate student who is interested in biostatistics and social determinants of health could be informed about the Graduate School of Public Health's Biostatistics MS degree with concentration in Health Data Science. Another student interested in nursing and data could be directed to Nursing's informatics program. Further students could be informed about the online Master of Science in Health Informatics Data Science track from the School of Health and Rehabilitation Sciences, Katz's and CBA's offerings on Business Analytics (certificate), the Intelligent Systems Program, Computational Biomedicine and Biotechnology program, and Biomedical Informatics. Students with career goals at the intersection of law, public policy, and international governance of data might be direct to the School of Law and/or the Graduate School of Public and International Affairs. Students that want to pursue research and education in data-driven education could be provided recommendations about opportunities in the School of Education and the Learning Research and Development Center. Indeed, as this long list of examples illustrates, *every* school at Pitt has programs that involve data science, and a unified message about what's available could help students *at all levels* – high school, undergraduates, and graduate – to explore new educational and scholarship opportunities, and imagine their career possibilities.

Other training opportunities and workshops, such as R programming from the University Library System, data programming fundamentals from the School of Computing and Information, FAIR data sharing from the Health Sciences Library System, or getting started with Pandas and Python

from the Center for Research Computing, could be described and advertised in one place as part of the unified message. 9 respondents to the department chair survey indicated their departments have non-credit bearing educational programs, which could also be brought together with these training opportunities. Such a central resource would be particularly supportive and useful to staff, graduate students, postdoctoral fellows, and faculty that want to acquire new skills in data. This approach may also bring awareness and a chance for academic units already involved in, or desiring to develop, degree programs and training to work together, share materials, and leverage their programs, rather than duplicate disparate effort. It may even lead to organizing and coordinating these programs into non-credit "badges" and bridge programs[14], similar to Google's and Cisco's certificates[15,16], except offered by and for the Pitt community.

"Data science success stories" could be sought and developed, possibly by the Office of University Communications and Marketing. These stories should highlight the full spectrum of uses of data science by everyone, no matter expertise or domain area, such as the recent Grief 2 Action effort by the Center for Analytical Approaches to Social Innovation[17]. Stories that describe data science in unexpected ways that touch upon daily life (e.g., planning of bicycle routes and parking meters) by individuals or groups new to data methods could be especially powerful. The data science liaisons could help provide assistance to University Communications and Marketing to identify and develop content. Such stories could serve to highlight compelling and powerful exemplars of how responsibility and use manifest themselves in data-driven research and education. The COVID-19 pandemic, for example, has numerous stories about data, e.g., how to automatically sort through information and determinate its reliability, how to model the social determinants of health to guide social distancing intervention, or how the unprecedented public sharing of COVID-19 public health data has intersected with data journalism and communication. These stories could be the foundation for a recurring feature about "Data@Pitt" in PittWire. They may also be used to create feature stories, or even an annual special issue, in Pitt Magazine – whether about outcomes, critique, tools, datasets, or methods. Using this content, a separate marketing campaign on data science could be used to disseminate the success stories, education opportunities, and research outcomes to potential student, faculty, and staff recruits.

This action will require a thorough and ongoing inventorying of data science – education programs, datasets, methods, tools, and research. Fortunately, some of this work has been done by the DSTF as part of our environmental scan. The DSTF identified, characterized, and cataloged 40 data-oriented undergraduate and Master programs (cf. Figure 2), and sought qualitative

[14] A type of bridge program could prepare students without backgrounds in data and data methods to undertake graduate study in a data area, similar to programs like the Swanson School of Engineering's Compass Program, https://www.engineering.pitt.edu/graduatecompass/

[15] Google Career Certificates. https://grow.google/certificates/

[16] Cisco Data and Analytics. https://www.cisco.com/c/en/us/training-events/training-certifications/training.html

[17] Graduate School of Public and International Affairs, Center for Analytical Approaches to Social Innovation. https://www.gspia.pitt.edu/CentersandInstitutes/Centers/CenterforAnalyticalApproachesToSocialInnovation

survey feedback from department chairs about their education and research programs, as well as the trends in their disciplines. These are useful resources; as far as the DSTF is aware, this is the first attempt to broadly catalog and characterize data science education at the University. In the short term, the existing scan could be augmented with information about training programs from the Center for Research Computing, the University Library System, the Pittsburgh Supercomputing Center, the Health Sciences Library System, the School of Computing and Information, and many others. Although the DSTF's scan is only a starting point, it could be used to immediately create content for a campaign to inform current and potential students, postdocs, faculty, and staff where to look for data science. The data science liaisons could again be helpful in seeking leads and advising on the full cataloging of activity.

Action 4 (short to medium term): Create and continuously update a "Data@Pitt" online web hub to aggregate and disseminate opportunities, success stories, events, activities, education pathways, and initiatives related to data science.

In addition to existing communication channels, a dedicated web presence – "Data@Pitt" – should be developed and stood-up to aggregate and share information about activities, stories, education, and research associated with data and data methods. The web presence could be crafted to be the "one-stop shop for all things related to data". Many of the recommended actions of this report could feed directly into this portal. For example, the web portal could provide a tool to ask students about their background, interests, and career aspirations, and then "match" them to a suggested set of data-oriented activities and degree programs. It would be a powerful tool to assist students in learning about new possibilities and exploring their interests.

The portal may also link to resources from Pitt Research (e.g., policies, grant opportunities in data), the University Library System (e.g., dissertations related to data science), the Health Sciences Library System (e.g., datasets, tools, best practices), the CRC and PSC (e.g., computational and storage services), Pitt IT (e.g., institutional data and infrastructure), and others. An "ask the experts" feature could be available to provide a search capability to find Pitt educators, researchers, mentors, and alumni who are expect on particular data science topics, tools, and datasets available. This feature may be extended to include a natural language interface and a recommender to sort through information about Pitt's data science activities to suggest educational programs/pathways, people, opportunities, and other information relevant to a particular interest or direction of inquiry. Although a fully developed web presence will take resources, some small steps in the near term, such as advertising data-oriented education programs (workshops, degrees, etc.), centrally promoting talks on data science (including the Distinguished Data Science Speaker series), and distributing success stories on a single aggregated web site would go a long way to presenting a unified view of data at Pitt.

3.2 Goal 2: Require fluency and knowledge.

Require every undergraduate student to acquire a basic understanding of data and data methods, including consideration of responsibility, as part of their learning at Pitt.

This goal aims to encourage developing inclusive curriculum that provides *every* undergraduate with opportunities for strong preparation in responsible, use-driven data science. We also suggest the development of an "ecosystem" that will promote opportunities and leverage curriculum for students to participate in data science. By collaborating on the ecosystem, there will be more cross-pollination of education and research ideas across domains, and new professional connections created, leading to more interdisciplinary projects, including ones of larger scale.

We have many undergraduate programs and likely will acquire more in coming years. Having the same specific data science requirements for students in widely different programs seems unwise. For example, students in the School of Computing and Information (SCI) should become able to produce routines and systems that embody data science. Hence, current guidelines for such programs often require programming language learning early in the curriculum so that later courses can include exercises using Python, R, TensorFlow and perhaps other languages. On the other hand, a student in a humanities major may need a very different grounding in order to leave Pitt as a well-educated person able to succeed in career choices and able to participate strongly as a well-informed citizen. Solid preparation may include engagement using data science in projects within their studies or in civic engagement as well as an understanding of how data science works to produce new knowledge in the disciplines of their collaborators.

There is also need to educate across boundaries, and to teach about the intersections of data science and domains, and vice versa; e.g., SCI students need to understand responsible uses of data, and the role of social context in their future contributions. Humanities students need to understand the broad contours of data science to contribute to our understanding of data science influence on society and the human experience. Law students need to understand the potential technical capabilities of new data science methods to appropriately govern and regulate them without prematurely limiting innovation.

Action #5 (medium term): Mandate that every school with an undergraduate program develop inclusive curriculum, coupled to practical experience with actual datasets, questions, methods and tools, that provides all undergraduates with preparation in data concepts and skills.

We recommend that each school with an undergraduate program should develop opportunities that equip students with knowledge and experience in collecting, storing, analyzing, interpreting, critiquing, and communicating insights from data. Undergraduates should be exposed to the concepts of data – the use and misuse of data in science and/or society – through at least one course. These concepts should be taught using experiences and content relevant to students from a range of backgrounds. Learning outcomes related to data science and their manifestation will naturally differ between schools, departments, and degree programs, as content is tailored to specific context. Course and program assessment plans should incorporate evaluation of these tailored experiences, e.g., critical thinking about data. Although there is no "one size fits all" for such a course(s) across disciplines, particularly when tied closely to use, there may be opportunities to share curriculum components and best practices.

In the medium term, the DSTF recommends that data science evolve into a general education requirement in each school. A mandate from the Provost to examine how best to incorporate a general education requirement (e.g., should it be treated as a science or quantitative requirement, some sort of new requirement, embedded in courses, what might be replaced?) will likely be needed to act quickly with purpose. Although broad curricular goals and guidance might be set by the University, any new requirement comes at an inevitable cost. Credits allocated to data science, i.e., a general education requirement, will come from a finite pool, and the number of credits required for a degree is unlikely to change. Accordingly, existing courses may need to be adapted to incorporate data methods, while courses that serve multi-purposes may need to be created afresh. Such an approach would likely require a somewhat prescriptive approach to "push" units to create a plan on how they would incorporate and assess aspects of data science in their curriculum, or how they will work to add a general education requirement.

Table I: Data acumen areas from NASEM 2018

Mathematical Foundations	Data Modeling and Assessment
Computational Foundations	Workflow and Reproducibility
Statistical Foundations	Communication and Teamwork
Data Management	Domain-specific Considerations
Data Description and Visualization	Ethical Problem Solving

A 2018 report from the National Academies of Sciences, Engineering, and Medicine provides thorough analysis and recommendation of data science competencies that should be taught[18]. This report offers a roadmap of 10 data acumen areas (Table I) to develop curriculum that spans the full "data science life cycle" – i.e., acquisition, cleaning, using/reusing, and preserving/destroying data with consideration to ethics, policy regulations, stewardship, platform, and disciplines[19].

Figure 3 shows the amount of emphasis given on average by the 40 programs scanned by the DSTF (blue bar, "Emphasis"). The emphasis ranges from 1 (minimum/unknown) to 4 (high emphasis). The figure also plots responses from the survey of department chairs about their programs. In particular, the survey asked "Looking forward by 5 years, how would you characterize the importance of the following data acumen areas to your undergraduate students?" The same question was asked about graduate students. The range is 1 (no increase/not applicable) to 4 (significantly increased). Considering current programs, there is a gap between the average from the scan and the survey for Workflow and Reproducibility, and

[18] Other reports arrive at similar conclusions as NASEM, although they may be tailored to discipline. For example, the Association for Computing Machinery (ACM) – the primary professional organization for computer science – has an ongoing effort to recommend data science curricula. ACM Data Science Task Force, Computing Competencies for Undergraduate Data Science Curricula, Draft 2, December 2019. http://dstf.acm.org/DSReportDraft2Full.pdf

[19] The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science. Victoria Stodden. *Communications of the ACM*, July 2020, Vol. 63, No. 7, pages 58-66. DOI: 10.1145/3360646

Ethical Problem Solving (Responsibility). Both areas have gained increased interest in recent years, and this gap is likely due to the evolving emphasis between current and future trends. Data Management, Data Visualization, and Domain-specific Considerations also have gaps.

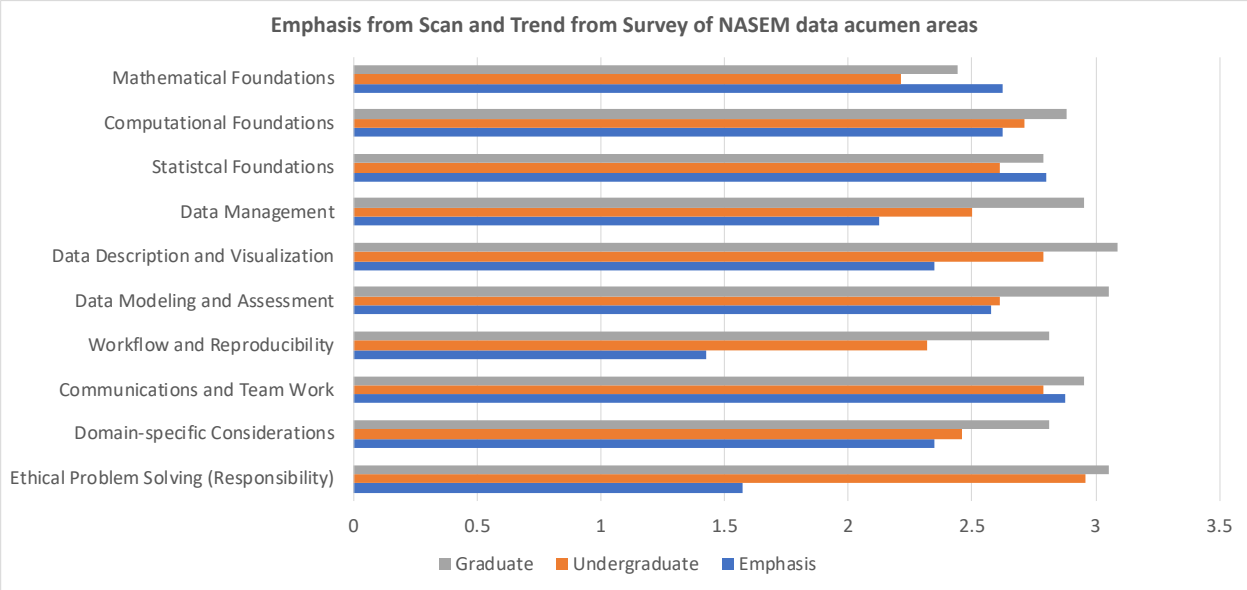

Figure 3: Amount of emphasis from scan and importance in trend from survey

In addition to the areas of the NASEM's roadmap, the DSTF strongly recommends that responsibility, ethics, and evaluation of *complex evidence* should be emphasized in curriculum. For example, some course topics should consider misuses of data and evidence, critique of methods (especially bias), accessibility, and marginalization and disadvantage through "data haves" and "data have nots". Critical thinking should be emphasized as well. Specifically, many challenges that we see in society come from not appreciating that we that we live in complex systems, and thinking there are simple solutions, particularly ones that might be driven by data. Consideration of these factors are all important elements to inclusiveness and good citizenship.

The DSTF further recommends that skills to effectively participate and translate among domains through interdisciplinary and diverse teams must be part of any data science curriculum. Data science draws on fields and subfields with different ontologies and epistemologies. People often must work in teams whose members have different backgrounds, experiences, and expertise, grounded in different basic understanding. Excellent training would give students a sense of the disciplinary ontologies relevant to their own field(s) and fields nearby, and the ability to collaboratively work, communicate across fields, and work as part of diverse teams.

Real world data is inherently messy and data analysis is filled with ambiguity and loosely defined questions and inquiries. Consequently, the DSTF recommends that the curriculum should be placed into context through actual cases of responsible, use-driven data science, i.e., practical experience with datasets, questions, methods, and tools. This experience could be offered through opportunities in community-engaged or societally-engaged data science, ideally with Pitt's many community partners, UCSUR, the Digital Inclusion Center, and others, to incorporate

projects, internships, and other 'learn-by-doing'. It will be crucial to the success of data science at Pitt as well as for the future of Pitt's competitiveness in this data-driven world that this expertise is available as a layer across the university, and not simply as a place that you have to know about or be adjacent to in order to find and access the necessary knowledge infrastructure.

An important consideration for learn-by-doing is the effort required to adopt examples in curriculum relevant to students from many backgrounds to create belongingness in the classroom. Fortunately, resources exist to reduce the barrier of entry by faculty. For example, Xprize created an alliance of organizations to apply the power of data to stopping the COVID-19 pandemic. As part of their effort, they built a catalog of open source projects, tasks, and datasets that could be used to provide societally relevant and compelling experiences to students. Kaggle also supports and guides projects for analysis, software code, and datasets. In addition, they conduct open "competitions" that could be a source of projects and experiences. A recent example is Cornell Lab of Ornithology's project at Kaggle to develop a tool for birdcall identification to improve the accuracy of population censuses of migratory birds. Kaggle's crowd-source model could be applied at Pitt to marshal datasets, tools, and analysis questions from researchers into ready-made projects that would ease adoption of meaningful experiences, benefit scholarship, and excite students.

Lastly, advisors should be trained to effectively and proactively assist students in selecting courses and seeking experiences about responsible, use-driven data science.

3.3 Goal 3: Catalyze skill acquisition.

> **Create, support, and incentivize inclusive, flexible undergraduate and graduate educational programs and shared educational resources to offer training in data science – in context of a broad variety of domains – to students, postdocs, staff, and faculty.**

While goal 2 propels data science out to diverse parts of the University for undergraduate education, goal 3 pulls people into coherent networks of data science training and applications who might not be there otherwise. This goal is intended to scaffold individuals that seek knowledge and training to use data and data methods in their work, especially in their careers and at the graduate level and including faculty and staff. Goal 2 is principally focused on *literacy and fluency* for all undergraduate students. On the other hand, goal 3 is focused on *competency* to conduct research in data science or careers that apply data science to transform other career paths (e.g., campaigns, pharma, libraries, etc.).

We suggest that a range of advanced educational opportunities need to be offered that are coordinated but tailored to uses within different disciplines to provide *specific* advanced training and education programs. Our recommendation is not one about centralizing the educational programs themselves, but rather a recommendation to develop and provide services and resources to enable individual programs throughout campus to leverage one another's curricular work and knowledge to mutually benefit all programs.

Action #6 (short to medium term): Identify gaps in existing curriculum, develop a set of shared educational resources for these gaps, and provide a central repository of curricular materials at both the undergraduate and graduate levels.

There has been significant interest for curriculum development in data science by many programs, especially at the graduate level, e.g., Nursing, Pharmacy, Health and Rehabilitation Sciences, Statistics, Economics, Public Health, as well as Computational Social Science. While curricular needs are often specific to a discipline, usually oriented around particular inquiries, tools, and datasets, the *underlying* skills and competencies that need to be taught are actually similar. For instance, learning to program in Python is a current commonly needed skill, regardless of the discipline. Yet, the datasets and analyses that may be explored in a programming course on Python would probably need to differ from one domain to the next (e.g., consider the differences among infectious disease data, electronic health record data, and digitized historical texts and images). Thus, we recommend supporting multiple, diverse efforts in a broadly defined data science space, to allow innovation to flourish, and not to prematurely seek to reduce and consolidate offerings in the name of efficiency. Complement efficiency with incentives, such as new professional opportunities. At the end of this process, we aim to see a more effective, collaborative delivery of existing degree packages, as well as the identification of any educational gaps, and mechanisms to fill said gaps.

A set of flexible and customizable resources could be developed, supported, and disseminated to assist individual units in developing their own advanced data science training. These resources could take many forms and serve multiple purposes, e.g., assessment approaches of data-oriented teaching, curricular materials and courses; course syllabi; learning outcomes with data; best practices; lecture slides; projects and assignments; or other "modules" that can be adopted or adapted to a particular context. For example, an introductory learning module on data mining might be designed to be customizable with domain datasets and exploration problems, even though the *same method* is taught regardless of the area. The availability of these resources would remove impediments faced by many disciplines in developing and evaluating new courses. It would also help to continue to keep curriculum up to date to maintain pace with the rapid advances in data science. Such resources would provide a path to accelerating curriculum development and offerings.

This action requires defining the marketplace using our definition of responsible, use-driven data science – including those who *use* data and methods, and those who gather data and *develop* the methods; identifying gaps and opportunities to share materials and expertise; seeking buy-in and participation of programs that want/need to offer advanced training; developing incentivization structures; and organizing a plan to assess, continuously improve, share, and sustain the materials.

Action #7 (medium to long term): Establish and/or charge an organizational entity to coordinate training and education, development of curriculum, collecting and disseminating project opportunities, courses and course content materials.

Progress on both goals 2 and 3 will lead to a rich ecosystem for data science education, including short online modules, "pop-up" training, non-credit workshops, micro-credentials, and degree programs[20]. Thus, the DSTF recommends that an organizational entity is charged, or a new one created, to shepherd and maintain the curricular ecosystem. This might be an entity, like the University Center for Teaching and Learning, or part of the organization suggested in Goal 4.

Specifically, the organization could provide a number of valuable capabilities:

- Advise schools/units on specifics of data science curriculum (uses, methods, tools, and datasets), including ethics, citizenship, and responsibility;
- Prepare, deliver, and continuously create content and maintain learning modules that support multiple purposes (courses, workshops), including gen eds;
- Curate and steward a range of exemplar datasets that can be shared for educational purposes and used in a variety of curricular activities;
- Develop and assist in contextualizing modular content to disciplinary needs, datasets, questions, and tools;
- Create assessment tools and validation strategies for inclusive teaching and learning along the full spectrum of basic data literacy to advanced methods and uses;
- Connect undergraduate students to learn-by-doing experiential opportunities that ask 'real-world questions' using production datasets and tools;
- Seek and curate ideas for data-oriented student projects from researchers and scholars throughout campus, packaging the ideas into well documented and ready-made experiences (e.g., similar to Kaggle competitions and hackathons) that could be offered to students and/or easily adopted by faculty in their courses;
- Recruit and attract top undergraduate and graduate students who seek to focus on data sciences as part of their learning, possibly with scholarships;
- Develop and offer incentives, recognition, and awards to encourage participation and contribution in an ecosystem of data science curricular materials; and,
- Provide advanced training to faculty, graduate students, postdocs and staff, such as summer bootcamps, workshops, and seminars focused on particular skill needs.

3.4 Goal 4: Coordinate strategy and action.

Implement a structure that (i) knits together, in a visible, accessible, and central place, people and practices in data science; (ii) serves as an evolving source of knowledge in developing and applying responsible data science to overcome diverse, challenging problems, including ethics, policy, and legal aspects; and (iii) animates extraordinary ambitions and success in collaborations transcending disciplinary and community limitations.

[20] National Academies of Sciences, Engineering, and Medicine. 2018. *Graduate STEM Education for the 21st Century*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25038

The DSTF urges Pitt to create an institutional focal point – in actuality, a formal *structure* – to embody the ambition to build an evolving and resilient organization that coordinates and fuels data science innovation[21]. This structure should connect and involve academic units, research programs, infrastructure, and operations, including Pitt IT and Pitt Research.

To build capabilities in data science to a point of excellence and impact across the full range of research, teaching, and community engagement ambitions, Pitt will need to prioritize: Pathways, practices, and policies for connecting data science capabilities around the university, and between the university and organizational partners, in "layers" of data science activity (cf. Figure 1), such as data identification and collection (one layer); data curation and storage (another layer); data modeling and analysis (a third layer); data interpretation and visualization (a fourth layer); and data sharing and communication (a fifth layer); as well as other emerging layers. These capabilities currently exist in unconnected silos defined by discipline and department.

This goal aims to connect the silos into learning-based systems or communities, which leverage their complementary strengths but do not undercut their independence to maximize the power to tackle large-scale, complex problems; scholarly impact; and social benefit.

Action #8 (short to medium term): Establish a dedicated, full-time position and charge a leader with a mandate to advance and coordinate data science for Pitt.

This action recognizes the importance of and the need for commitment by a leader to rally and bring together people to harness their collective networks and social credit to communicate, share, and coordinate data science in all of its forms. A dedicated position is urged to be created, possibly within the Provost's office or some other academic unit that has experience in relevant areas and responsibilities related to data science. This position would connect and orchestrate points of data science for an institutional focal point, drawing *together* resources from throughout campus. A leader in this position would ensure accountability in moving the proposed actions forward, and championing key activities, such as organizing the data science liaisons, developing the distinguished speaker series, rallying support and buy-in for curriculum and general education requirements in data science, and curating content for Data@Pitt. It is worth noting that 26 of the 41 benchmarked efforts at other institutions have a formal leadership position for data science. All 7 of the most comprehensive efforts have a leader for data science.

The leader should be charged to put the recommended actions, and others that may emerge, into play. A strong leader equipped with a clear mandate will give a basis for Deans and unit heads to prioritize coordination. To make this basis stronger, data science initiatives could be connected to unit strategic plans and Resource Proposals, and tied into the Plan for Pitt 2025.

[21] Micaela S. Parker, Arlyn E. Burgess and Philip E. Bourne. Ten Simple Rules for Starting (and Sustaining) an Academic Data Science Initiative. June 2, 2020. https://doi.org/10.31219/osf.io/wu4fv

Action #9 (short term): Use Pitt Momentum or other funding mechanism to encourage, initiate, and support action on the highest impact actions in this report and work toward the development of a polished concept, with pilot implementation, that can be a springboard for a major gift.

Many of the DSTF's recommendations require resources, and the current economic environment will make this challenging. Some initial steps could be taken, however, for quick effect and impact with small funding (e.g., the Data Science Success Stories). Others could be turned into a seed project for a Pitt Momentum or other proposal opportunity to initiate action. In particular, several ideas formulated by the DSTF for undergraduate and graduate training and novel teaming of data scientists could be well suited for a pilot that more fully seeks buy-in, develops ideas, determines needs, and implements assessment. This could benefit students and postdocs right away through a pilot education program. Such a project will be important to establish expertise, credibility, and preliminary evidence for large external sponsorship of a full effort that spans disciplines, e.g., an NSF Research Traineeship or a NIH/NLM training program.

Learning how to coordinate and amplify data science by *actually* doing it, will help refine ideas and approaches and crystalize them into a polished concept that could be a springboard for a major gift. With the support of University senior leadership, a concept sheet for a well-defined data science effort could be developed. From anecdotal information uncovered by the DSTF's benchmark effort, this mirrors approaches taken at other universities, like the University of Virginia and Princeton, both of which received major gifts for data science. Given the current financial climate, this "go slow, show what we can do" approach would move the DSTF actions forward, in preparation for an improved economic outlook in a few years.

Action #10 (long term): Create and support an institutional structure – a "coordination tower" – to coordinate existing and emerging elements (layers) of data science.

An institutional home by which our recommended pathways, practices, and actions are documented, coordinated, and energized, with physical, organizational, and conceptual embodiments is needed. This home would manifest itself in different ways as a physical place, a group of dedicated people, and a mission, supported effectively by organizational and financial capital. In essence, this would be a "coordination tower" that orchestrates data science across units at the University. The "coordination tower" metaphor should evoke not simply a complex airport (or railyard); instead, it should invoke a "JPL" metaphor, as in "Jet Propulsion Laboratory" and its role in advancing rocket science, not simply building a single rocket but rather supporting and catalyzing an *entire* science. In this case, the coordination tower of this action is akin to a "*Data Propulsion Laboratory"*. It captures the layers of the stack shown in Figure 1.

That "JPL-style" coordination function would enable Pitt to not only empower teams that propel research upwards by solving complex social problems and making new scientific discoveries, but also keep, propagate, and communally learn from the lessons of data science innovation attempts (both successful and less so) from across the university.

A crucial opportunity is for the coordination tower to develop, learn about and share solutions for increasing diversity and equity, both in data science as research and teaching fields and in identifying and addressing societal problems via responsible data science-supported practice. Another opportunity is to evolve data practices and to promulgate practice throughout University research, education, and operations. It would also be a chance to bring together core competencies of the different data layers and grow the layers together and learn across and within the layers.

United with the data science liaisons recommended as part of Goal 1, the coordination tower could facilitate a leadership cohort that can help evolve responsible uses of data science, as well as the methods themselves, and recruit resources and formulate new initiatives under the direction of the leader recommended in Action 8. Such champions and resources might include a cadre of experts, from across campus, and dedicated professional staff who serve as consultants, investigators, and educators to support and develop the use of data science in research, education, and operations. Another important opportunity is to evolve data practices, learning what is good, what we have seen to work, what is working, and what is not working. Lastly, and certainly not the least, the coordination tower could serve an opportunity to facilitate, connect, and grow communities that are today using data methods, as well as communities that may want to start using these methods but lack critical mass or expertise to do so.

4 Next Steps and Evolution of the Task Force

The DSTF recommends that further consultation is undertaken to interview stakeholders about the envisioned institutional focal point and the formal structure needed to bring it to fruition. The stakeholders should include department chairs (particularly the respondents to the DSTF's survey and the contacts that they identified), student groups and postdoctoral researchers, the Council of Deans, the Provost's Cabinet, Pitt Research, Pitt IT, UCTL, UCSUR, and the Pittsburgh Supercomputing Center. With input from these stakeholders, the goals and actions should be further refined, and key performance indicators and associated targets should be developed for each of the actions to be implemented.

Additionally, the DSTF recommends that work starts right away on Action 1 (data science liaisons) and Action 2 (distinguished speaker series). Several members of the DSTF are eager to work together and to invite members of the University community to join in establishing both the data science liaisons' group and the speaker series. This would serve to attract and involve interested parties in providing feedback and consultation for further development and eventual implementation of the recommendations of this report. Looking further down the road, some members of the DSTF are also interested in beginning work on involving relevant constituents in developing shared curricular materials, integrating, piloting, and assessing these materials in their degree programs. Lastly, there is interest in the DSTF to develop a proposal for a Pitt Momentum Scaling Grant or other funding vehicle to seek seed funding for these first steps, especially related to data science for social justice.

5 Appendix Task Force Members

Colin Allen	Center for Philosophy of Science
Michael Becich	School of Medicine/DBMI
Michael Blackhurst	University Center for Social and Urban Research
Aaron Brenner	University Library System
Bruce Childers (chair)	School of Computing and Information
Michael Colaresi	Dietrich School of Arts & Sciences – Social Science/Political Science
Kevin Crowley	School of Education
Eleanor Feingold	Graduate School of Public Health
Nathan Glasgow	Postdoctoral Fellow
Catherine Greeno	School of Social Work
Mike Holland	Pitt Research
Heng Huang	Swanson School of Engineering
Satish Iyengar	Dietrich School of Arts & Sciences – Natural Sciences/Statistics
Brett Jones	Undergraduate Student Representative
Chris Kemerer	Katz Graduate School of Business / College of Business Administration
Alison Langmead	Dietrich School of Arts & Sciences – Humanities/HAA
Young Ji Lee	School of Nursing
Alan Lesgold	Learning Research and Development Center
Sera Linardi	Graduate School of Public and International Affairs
Michael Madison	School of Law
Mary Marazita	School of Dental Medicine
Bambang Parmano	School of Health and Rehabilitation Sciences
Koastas Pelechrinis	School of Computing and Information
Melissa Ratajeski	Health Sciences Library System
Ralph Roskies	Research Computing
Brett Say	University Honors College
Sera Thornton	University Center for Teaching and Learning
Steve Wisniewski	Office of the Provost
Xiang-Qun (Sean) Xie	School of Pharmacy

6 Appendix Charge to the Task Force

From scientific and mathematical discoveries through computation and pattern recognition, to the visualization and analysis of demographic and linguistic data, and the analysis of texts and images, the continuing expansion of digital data and our ability to store, retrieve, and analyze it is bringing about an epistemological revolution. We are producing knowledge in new ways an at an ever-increasing pace. Empirical investigation has replaced theoretical speculation in areas as diverse as systems biology and the micro-foundations of macroeconomic phenomena. It is also a pragmatic revolution. Data science now permeates education, government, medicine, engineering, entertainment, science, the arts, humanities, and business, touching nearly every facet of life. In the university, we are enabling education and investigation to dig deeper and scan a broader phenomenal horizon. This infusion of data has led to a pressing need to teach students data-related skills, knowledge, ethics, and literacy, and to support faculty and staff in the collection, stewarding, retrieval, and analysis of data, whether in educational programs, research initiatives, or operations.

Many academic units at the University of Pittsburgh recognize this need and are responding individually with initiatives around data science. We have a burgeoning variety of “analytics,” “digital,” “computational,” “informatics,” and “-omics,” in fields as far flung as business, health sciences, social sciences, law, and humanities, as well as adaptive learning and student advising platforms and practices. Coordinating and collaborating on these activities could leverage our collective ability, jointly build capacity to meet the demand, and foster opportunities for new interdisciplinary educational and research programs. Other universities are undertaking similar coordination through a range of approaches, such as independent institutes; schools of data science; virtual divisions of data-oriented academic units; research and learning hubs in data science; data spaces and services in libraries; computing for machine learning, visualization, and data management; and interdisciplinary data science degrees. Pitt has enormous opportunities to create an exciting new initiative to bring all of this together with the recently launched the School of Computing and Information, and the plans for a new building to house the School and data science-related research and educational activities.

How should the broader Pitt academic community collectively act on the urgent need, given our context, strengths, and individual efforts in data-related areas? To this end, the Provost charges the Ad Hoc Committee on Data Science with recommending a coordinated strategy to catalyze, nourish, and sustain educational programs and research initiatives that (1) equip undergraduate and graduate students with the knowledge and skills necessary for the increasingly data-oriented world; (2) develop and use data science methods in research; and (3) attract and retain faculty using data and associated methods in their disciplines. In doing so, please answer these questions and any others that are deemed relevant by the committee:

- What current and future data-related educational and research programs exist at Pitt, and which would benefit from a coordinated strategy? How could new opportunities be enabled?

- What approaches would facilitate leveraging and interfacing current and future initiatives to ensure proficiency in data-related areas by our students, and to support and accelerate research relying on data science?
- What nomenclature and standards should be adopted to enable a shared foundation of education and research capacity in data science by and for the Pitt community?
- How can the identified approaches be implemented? How should the approaches be prioritized, and what should be done in the short, medium, and long terms?

In answering the questions, please consult the Pitt community broadly to seek their input. Please establish an inclusive process for this engagement at the start of the committee's work.

By January 30, 2020, please submit an interim report that summarizes the committee's findings to that date, with the complete report due by June 30, 2020.

7 Appendix Task Force Process and Activities

The Data Science Task Force was formed by the Provost in Fall 2019 and charged with producing a set of recommendations to coordinate and amplify data and data methods at Pitt. The 29-member DSTF has members from all schools, University centers, University Library System, Health Sciences Library System, Pitt Research, undergraduate and graduate students, and the Office of the Provost.

The DSTF first convened in November 2019 and held several full task force and individual group meetings throughout the Spring term 2020, including a "mini retreat" in April 2020. Initial feedback was sought through a survey of department chairs and conversations by members of the DSTF. The committee's work has progressed in phases:

> Phase 1: Determine definition of "data science"; conduct an environmental scan of data-oriented programs; and, benchmark external institutions to learn their experiences in coordinating data science activities and effort. The DSTF organized itself into three groups to tackle each of these items, and then came back together to collectively synthesize and combine information.
>
> Phase 2: An online retreat was held to determine the provisional goals based on the background work. The retreat resulted in five distinct goals, which were subsequently combined through asynchronous discussion and a survey into the four goals of this report.
>
> To develop actions for each goal, the DSTF divided itself into four separate groups, which held a series of individual ideation meetings to arrive at recommended actions. Each group produced a report, which are available separately. The DSTF came back together to review, combine, and synthesize the full collection of actions from the groups into the set of 10 actions in this report.
>
> Phase 3: Share provisional recommendations for a definition, goals, and actions for feedback and refinement.

The first two phases were completed in Spring 2020, and this draft report was developed over the Summer 2020. The third phase was delayed to late Fall 2020 due to the Covid-19 pandemic. In the Fall, the DSTF's worked to refine recommendations, seek feedback, and produced the set of recommendations.

Upon completion of the third phase, the DSTF could evolve into a group that undertakes some of the actions outlined in the report. Several members of the DSTF expressed a commitment to be part of a group to initiate action.

8 Appendix Department Chair Survey Data

The information below summarizes the responses from the survey to department chairs.

To what extent is data important to the learning outcomes of your educational programs?

Not applicable	2
Slightly important	9
Important	18
Significantly important	35

What is the trend of importance for skills and knowledge in data science by employers of your students?

No change in importance	3
Slightly increasing importance	4
Increasing importance	29
Significantly increasing importance	28

Do your current undergraduate degree programs teach students how to develop data science methods, how to apply data science methods, how to address fundamental or critical questions about the nature and uses of data, or some combination of these? Select all that apply.

Not applicable	35
Teaches how to apply data science methods	19
Teaches how to develop data science methods	13
Teaches how to address fundamental or critical questions about the nature of uses of data	16

Considering current undergraduate degree programs offered by your department, how would you characterize the methods that your undergraduate students are required to learn in the following data acumen areas? If you have multiple programs, please select all choices that apply.

	Not used	Basic	Comprehensive	Emerging	Avg.
Mathematical foundations	5	11	9	2	2.30
Computational foundations	5	12	6	4	2.33
Statistical foundations	4	14	8	1	2.22
Data management and curation	7	14	5	1	2.00
Data description and visualization	5	12	7	3	2.30
Data modeling and assessment	9	10	8	0	1.96
Workflow and reproducibility	13	7	6	1	1.81
Communication and teamwork	2	11	12	2	2.52
Domain-specific data and data methods	9	8	8	2	2.11
Responsibility and ethics of data and data methods	3	16	5	3	2.30

Do your current graduate degree programs teach students how to develop data science methods, how to apply data science methods, how to address fundamental or critical questions about the nature and uses of data, or some combination of these? Select all that apply.

Not applicable	13
Teaches how to apply data science methods	31
Teaches how to develop data science methods	21
Teaches how to address fundamental or critical questions about the nature of uses of data	35

Considering current graduate degree programs offered by your department, how would you characterize the methods that your undergraduate students are required to learn in the following data acumen areas? If you have multiple programs, please select all choices that apply.

	Not used	Basic	Comprehensive	Emerging	Avg.
Mathematical foundations	9	19	9	10	2.43
Computational foundations	8	18	12	9	2.47
Statistical foundations	4	16	21	6	2.62
Data management and curation	4	19	20	4	2.51
Data description and visualization	4	14	18	11	2.77
Data modeling and assessment	5	19	12	11	2.62
Workflow and reproducibility	10	16	18	3	2.30
Communication and teamwork	3	13	24	7	2.74
Domain-specific data and data methods	5	14	17	11	2.72
Responsibility and ethics of data and data methods	3	14	24	6	2.70

In the next 5 years, which of your educational programs will teach students about data or data methods?

Undergraduate	28
Graduate	44
Noncredit	9
None	3

Looking forward by 5 years, how would you characterize the importance of the following data acumen areas to your undergraduate students?

	No change	Slightly increased	Increased	Significantly increased	Avg.
Mathematical foundations	6	13	6	3	2.21
Computational foundations	2	11	8	7	2.71
Statistical foundations	2	11	11	4	2.61
Data management and curation	3	11	11	3	2.50
Data description and visualization	1	10	11	6	2.79
Data modeling and assessment	3	12	6	7	2.61
Workflow and reproducibility	2	18	5	3	2.32
Communication and teamwork	1	9	13	5	2.79
Domain-specific data and data methods	2	14	9	3	2.46
Responsibility and ethics of data and data methods	1	7	12	8	2.96

Looking forward by 5 years, how would you characterize the importance of the following data acumen areas to your graduate students?

	No change	Slightly increased	Increased	Significantly increased	Avg.
Mathematical foundations	5	21	10	7	2.44
Computational foundations	2	14	14	13	2.88
Statistical foundations	0	16	20	7	2.79
Data management and curation	1	12	18	12	2.95
Data description and visualization	0	11	17	15	3.09
Data modeling and assessment	2	8	19	14	3.05
Workflow and reproducibility	2	14	17	10	2.81
Communication and teamwork	2	9	21	11	2.95
Domain-specific data and data methods	1	15	18	9	2.81
Responsibility and ethics of data and data methods	1	9	20	13	3.05

To what extent is data important to research in your department today?

Not applicable	0
Slightly important	7
Important	18
Significantly important	30
Average importance	3.42

Over the next 5 years, what is the trend in importance of data science methods to research in your department?

No change in importance	3
Slightly increasing importance	8
Increasing importance	17
Significantly increasing importance	27
Average importance	3.24

Do researchers in your department use data science methods, develop data science methods, or a combination of both? Select all that apply.

Not applicable	0
Researchers use data science methods	23
Reserchers develop data science methods	3
Researchers use and develop data science meethods	26

How would you characterize the methods used in research in your department for the following data acumen areas? Please select all choices that apply.

	Not used	Basic	Comprehensive	Emerging	Avg.
Mathematical foundations	6	19	13	12	2.62
Computational foundations	2	18	18	13	2.82
Statistical foundations	3	9	29	10	2.90
Data management and curation	1	21	23	5	2.64
Data description and visualization	0	13	27	10	2.94
Data modeling and assessment	3	13	16	19	3.00
Workflow and reproducibility	6	21	20	3	2.40
Communication and teamwork	4	12	30	4	2.68
Domain-specific data and data methods	3	15	21	11	2.80
Responsibility and ethics of data and data methods	3	16	22	9	2.74

9 Appendix Benchmark Scan

To identify benchmark institutions, the group started with a list from Evaluation of the Moore-Sloan Data Science Environments[22] and those used for benchmarking by the Pitt Data Management Committee Policy Working Group in 2017. This list was augmented with institutions that have unique or university-wide approaches, or that the group had personal knowledge about. In early 2020, working from the list of 41 institutions, the group gathered information from public sources about key attributes and captured these in a table for comparison.

Institution	Clear definition	Dedicated space	Formal Leadership	Fellowships	Professional Cert.	PhD	Masters	Undergrad	HS/pipeline	Othr training & Support	Center	Institute	Initative	School/College	Dept/Division	No structure
University of California Berkeley	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓			✓	
University of Washington (Seattle)	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓					
UVA	✓		✓	✓	✓	✓	✓	✓	✓		✓			✓		
NYU	✓		✓	✓	✓	✓	✓	✓		✓	✓					
University of Illinois at Urbana	✓		✓		✓	✓	✓	✓				✓	✓			
University of Michigan			✓	✓			✓	✓	✓	✓	✓	✓				
Penn State			✓		✓	✓	✓	✓		✓	✓	✓				
Duke			✓	✓	✓		✓	✓			✓		✓			
University of Rochester	✓	✓	✓				✓	✓		✓		✓				
MIT			✓		✓	✓	✓	✓			✓			✓		
University of Wisconsin Madison		✓	✓			✓	✓	✓				✓		✓		
University of North Carolina-Chapel Hill	✓		✓		✓	✓	✓	✓				✓				
CalTech and JPL	✓		✓		✓	✓		✓		✓	✓					
Harvard	✓		✓			✓	✓	✓		✓			✓			
Johns Hopkins	✓		✓		✓		✓	✓		✓		✓				
Columbia's Data Science Institute	✓		✓		✓	✓	✓					✓				
Ohio State		✓			✓		✓	✓		✓		✓				
Purdue University	✓	✓	✓				✓	✓					✓			
University of Chicago			✓		✓		✓	✓	✓		✓					
Georgetown University			✓		✓		✓				✓	✓	✓			
Princeton University	✓						✓	✓		✓	✓					
University of California San Diego			✓			✓	✓	✓				✓				
University of California San Francisco			✓	✓					✓	✓		✓				
University of Texas San Antonio		✓	✓				✓	✓						✓		
Case Western Reserve University	✓					✓	✓	✓							✓	
Georgia Tech			✓		✓		✓	✓				✓				
University of Florida	✓	✓	✓	✓								✓				
Boston University		✓	✓								✓		✓			
University of Massachusetts Amherst	✓						✓	✓			✓					
Carnegie Mellon University							✓	✓	✓						✓	
Northwestern	✓						✓			✓	✓					
University of Pennsylvania	✓						✓	✓				✓				
University of Texas Main Campus					✓		✓	✓				✓				
Yale						✓	✓	✓							✓	
Texas A&M			✓				✓					✓				
University of Toronto					✓		✓	✓								
Arizona State	✓						✓	✓								
Michigan State	✓							✓							✓	
Northeastern							✓	✓								✓
Washington University (St. Louis)						✓	✓					✓				
Stanford															✓	
	20	9	26	8	18	16	35	32	5	13	14	19	6	4	6	1

[22] Luba Katz, Evaluation of the Moore-Sloan Data Science Environments, Abt Associates, February 2019