Big Data Innovation Hubs National Coordination Committee:

Advancing Data Science Research: An Integrated Analysis

July 31, 2020

Rebecca Joy Norlander, John Voiklis, Elizabeth Attaway, Rupu Gupta, Joanna Laursen Brucker, John Fraser, & Uduak Grace Thomas

Recommended Citation

Norlander, R. J., Voiklis, J., Attaway, E., Gupta, R., Brucker, J. L., Fraser J. & Thomas U. G. (2020). Advancing Data Science Research: An Integrated Analysis. Knology Publication #NSF.159.648.03. New York: Knology.

Date of Publication

July 31, 2020

Prepared for

Florence Hudson
Executive Director
Northeast Big Data
Innovation Hub

This material is based upon work supported by the National Science Foundation under Grant No. (NSF #1946615, #1946558, and #1947000). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Knology produces practical social science for a better world.

tel: (347) 766-3399
40 Exchange Pl. Suite 1403
New York, NY 10005

tel: (442) 222-8814
3630 Ocean Ranch Blvd.
Oceanside, CA 92056

Knology Publication #

NSF.159.648.03

Executive Summary

The National Science Foundation (NSF)'s **Harnessing the Data Revolution** (HDR) Big Idea is a visionary, national-scale activity to enable new modes of data-driven discovery, allowing fundamentally new questions to be asked and answered in science and engineering frontiers, generating new knowledge and understanding, and accelerating discovery and innovation.

As part of this initiative, NSF seeks to identify what is needed to advance a robust ecosystem of data science research. Currently, the regional Big Data Hubs work to identify areas of collaboration and opportunities for supporting data science research, gathering input from the broader data science community. Knology, a nonprofit research organization, was selected by the Northeast Hub to handle data collection and analysis across a survey, online discussion, and conference with current HDR PIs and other NSF-identified stakeholders, and then synthesize findings in a public report.

This report highlights avenues for continued growth and concrete suggestions for possible next steps in five key areas:

- Collaboration between data scientists & subject matter experts
- Framing education and training opportunities
- Re-thinking the data
- Identifying best practices and creating repositories
- Broader cross-sector engagement

Our research demonstrated that participants overwhelmingly saw the need for effective **collaboration** within and beyond their current network, and desired increased teamwork. The data revealed both the importance and challenges of collaborations between subject matter experts and data science researchers. Supporting these kinds of partnerships requires the building of a culture of collaboration, and tools to enable clear communication across disciplines. Participants offered various suggestions for improving communication including developing shared vocabulary and language for facilitating interactions between stakeholders. The idea of an "HDR Dictionary", or data and domain science lexicon could be helpful and perhaps transformational.

Also critical to growing the field are efforts to **educate** and **train** the next-generation of data science researchers. Participants highlighted the importance of providing viable pathways for students, trainees, early career researchers, and academic scholars to advance. Recommendations from the field included providing interdisciplinary training modules for learning and opportunities for faculty to serve as advisors or mentors to students and post-docs. Respondents also thought that enabling students to get involved in real world projects would be of benefit to the field as a whole. These projects would also expose them to new types of data and, perhaps, motivate them to continue learning. Other suggestions included providing more funded opportunities for students and trainees, as well as rotations through labs in different disciplines.

The need for effective methods of collecting, storing, and sharing data surfaced frequently in data across all three events. Participants had suggestions for ways to **re-think data** practices to increase interoperability, thereby furthering collaborative efforts. Concrete recommendations here included standardizing ways of storing and sharing data; creating measures for assessing data completeness and quality, a discussion forum for data quality and improvement; and developing incentives and assessment strategies to encourage and reward open access of datasets. These ideas fall under the broader theme of **identifying best practices** for data collection and use. Participants also brought up the topic of best practices in the context of preparing the research workforce for careers in data science. One specific area that participants highlighted was a need for best practices for assessment or measurement to provide a mechanism for the field to measure its effectiveness against agreed upon criteria. In practice, the could include creating an evaluation guide and pairing it with a repository of validated instruments. Participants also highlighted the importance of recognizing interdisciplinary work and considering non-traditional metrics that prioritize real-world impact as markers of success.

Lastly, participants explored ways of accomplishing **broader cross-sector engagement,** especially with communities underrepresented in data science. There exists a wide array of potential stakeholders to be involved in and benefit from data science research. Participants proposed systematically enabling and training PI teams to do stakeholder mapping as a way of advancing HDR as well as supporting equity, diversity, and inclusion. Creating various research products – other than journal articles – was considered fundamental to the pursuit of broader cross-sector engagement. Recommendations included developing identifiers and standard citation practices for different types of research products.

Taken together, these interrelated findings help point the way toward what a robust data science research ecosystem would look like. This report also provides concrete steps that will help the community attain and sustain that ecosystem once it is in place.

Table of Contents

List of Tables

List of Figures

Introduction

The National Science Foundation (NSF)'s **Harnessing the Data Revolution** (HDR) Big Idea is a visionary, national-scale activity to enable new modes of data-driven discovery, allowing fundamentally new questions to be asked and answered in science and engineering frontiers, generating new knowledge and understanding, and accelerating discovery and innovation. The HDR vision is realized via a coordinated set of program solicitations (in TRIPODS, Data Science Corps, Ideas Lab, Institutes, and Big Data Innovation Hubs) resulting in an ecosystem of interrelated activities. Over the last two years, NSF has made awards that enable (i) research in the foundations of data science; frameworks, algorithms, and systems for data science; and data-driven research in science and engineering; (ii) advanced cyberinfrastructure; and (iii) education and workforce development.

In support of the funded activity of HDR, NSF is interested in identifying what is needed to advance a robust ecosystem of data science research. The Big Data Hub Leadership was selected by NSF to lead the process of determining potential areas of collaboration and opportunities for supporting data science research. Knology, a nonprofit research organization, has been selected to support the process through data collection and analysis that will eventually result in a public report that reflects a breadth of input from the data science community of researchers and practitioners. This work has been supported by NSF awards 1946615, 1946558, and 1947000.

This Report

This is the comprehensive report of all HDR Coordination activities supported by Knology researchers. The report has three chapters that each correspond to field-wide HDR activities, and a fourth chapter that integrates findings across those three activities.

Chapter 1 – HDR Stakeholder Survey

In winter 2019-20, Knology designed a survey that gathered perspectives regarding the possible support for and expectations of an HDR coordinating entity. In January 2020, the survey was shared with stakeholders that included HDR grantees and the Big Data Hub community. Analysis was conducted in February, 2020. Chapter 1 presents topline findings from the stakeholder survey.

Chapter 2 – Microlab Virtual Discussion

A virtual "Microlab" (and subsequent HDR PI Conference) were organized by Knowinnovation, a partner institution working with the South Big Data Hub. Knology attended planning meetings with Knowinnovation to provide insights from the Stakeholder Survey and advise on the Microlab and PI Workshop design. An appendix to Chapter 2 presents the outputs generated by the participants of the Microlab Virtual Discussion, held on March 16, 2020.

Chapter 3 – Conference Summary Themes

An in-person HDR PI Conference for HDR grantees had been planned for April 28-30, 2020. Due to Covid-19, the meeting shifted to a virtual format, facilitated by Knowinnovation. A researcher from Knology attended HDR Workshop and wrote short summary themes of all workshop sessions. Those summaries are presented in Chapter 3.

Chapter 4 – Synthesis and Key Findings

Knology researchers looked across all data and analysis presented in the first three chapters, representing the three major activities associated with HDR coordination, synthesized the information and highlighted key findings. This integrated analysis is presented in Chapter 4.

What Happens Next

A draft version of this report will be disseminated to interested stakeholders by NSF program officers and Big Data Hub staff in early August 2020. The public comment period will last until August 31, 2020. Knology researchers will use the feedback provided to revise this report. The final version will be submitted to NSF by September 30, 2020 and made widely available to the field.

HDR Stakeholder Survey

Knology, in partnership with the regional Big Data Hubs, developed a survey that was distributed to current HDR awardees and other stakeholders. The survey was aimed at better understanding the potential value of a coordinating entity. The results were intended to inform the process of developing a solicitation for a future coordinating entity, and to support the Microlab Virtual Discussion event and the HDR PI Conference facilitated by project partners, Knowinnovation.

Participants

The survey was deployed to all current HDR PIs and other NSF-identified stakeholders, and via Twitter and newsletter by each of the four regional Big Data Hubs. It was also delivered through the eScience Bulletin, which is the weekly newsletter for the University of Washington eScience Institute, a PI Institute for the West Hub. We received 162 responses to the survey, 92 of whom identified as currently working on a project being funded by the HDR grant mechanism. Some respondents also indicated that they have experience with other coordinating entities. We organized the survey responses into four groups based on the intersection of these two variables: working on an HDR project (yes or no), and involvement with another professional network or entity (yes or no). In this analysis, we refer to the groups currently working on HDR projects as "HDR grantees," while recognizing that not everyone working on an HDR grant is necessarily a PI.

Table 1: Survey response groupings based on two variables

Survey Responses	*N*
Yes HDR and Yes Other Entity	34
Yes HDR and No Other Entity	58
No HDR and Yes Other Entity	40
No HDR and No Other Entity	30

Choices & Ratings

Methods

Analyses of survey responses are based on the four participant groups (Table 1). The first analysis we did for each question looked at which items exceeded or fell short of the global mean. In determining the important functions of a coordinating entity, it seemed less important to tease apart fine distinctions between the four groups in Table 1. Instead, the team assessed gross trends in the data.

The second analysis focused on the HDR grantees and used logistic regression to test whether their responses differed from those running other data science projects. We also verified whether the two groups of HDR grantees (those previously involved or not previously involved with a coordinating entity) differ from each other.

We organized the results of both sets of analysis by survey topic. We used figures to summarize responses, and did a gross analysis of patterns and the finer distinctions between HDR grantees.

Results

Knology asked current HDR grantees and other stakeholders about several topics including:

- What tools or resources they plan to create as part of an active data science project;
- How they plan to share those tools or resources;
- Where they generally look for resources that they need;
- How much each of 12 activities would help HDR and non-HDR data science projects;
- Which of the nine functions of a coordinating entity do they value most?

Creating Tools & Resources

In general, respondents in the four groups said that they planned to produce some combination of the four specified tools or resources as part of projects that they were working on. They responded at rates that exceeded the global average. A negligible number of respondents said that their projects would yield no shareable products (Figure 1).

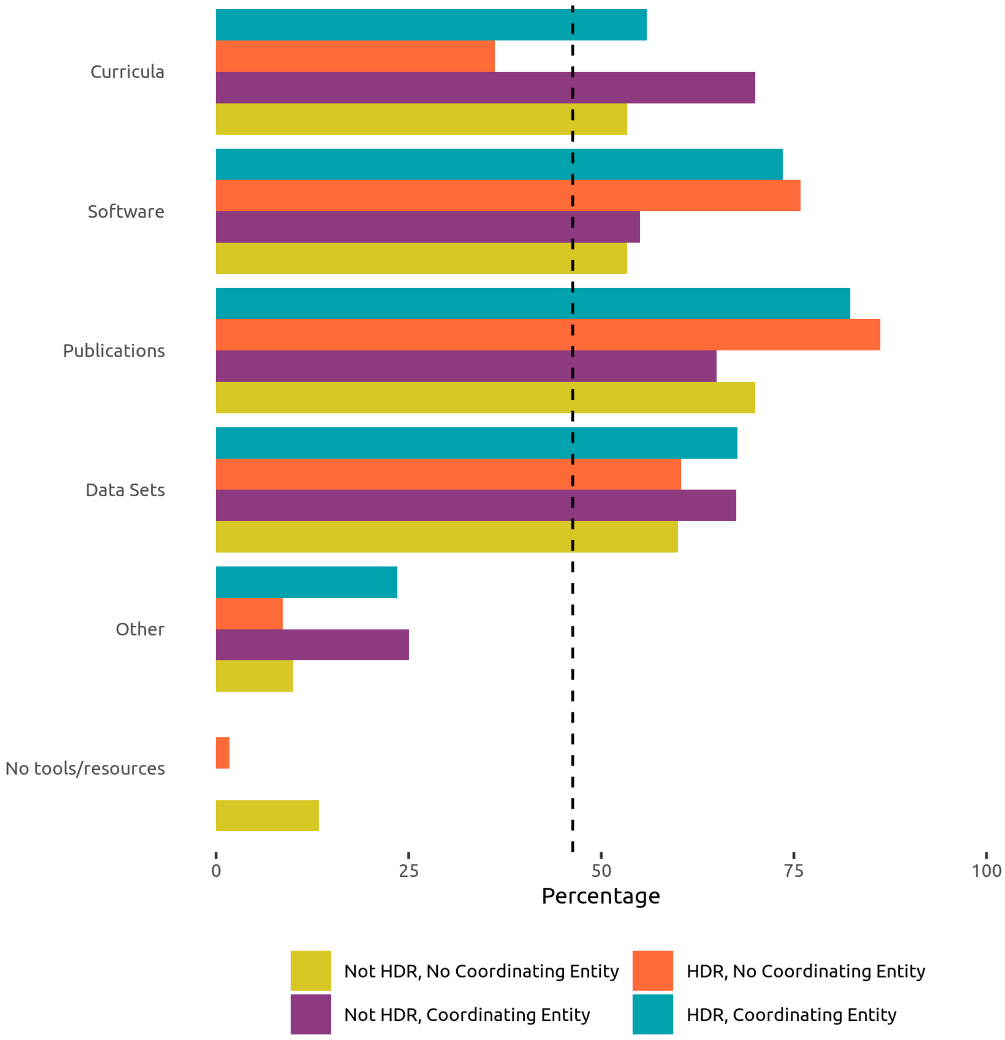

Figure 1 Distribution of responses to “What tools/resources will you be creating as part of a data science project that you may actively be working on?”

When we contrasted HDR grantees to each other and to other groups of respondents, we found no differences in the distribution of responses that exceeded chance occurrence (all main effects *p* > .1). As apparent in Figure 1 above, a notable, but likely chance, difference between the HDR grantees is that those who were not previously involved with a coordinating entity were much less likely to report planning to produce curricular materials ($\chi^2 = 24.15$, *p* < .001). Table 2 details the other tools and resources that respondents said they would create in addition to those specified in the question.

Table 2. Other tools and resources respondents said they would create as part of a data science project.

What tools / resources will you be creating as part of a data science project that you may actively be working on? Please select all that apply. - Other - Text
A community of practice
A working educational service
Advising students to join summer programs, or to learn material before going on the job market
Algorithms
Best practices for team science
Community
Convergence processes
Cultural Map-   interactive internet accessible map for a Native American rural community
Documentation
Experimental data
Firmware
I provide support to enable these programs
Insights/models
Jupyter notebooks
Learning resources
Modeling framework
My grant is to build a community organization for the data community
Predictive model and validation layer using NLP
Products and Services
Science Gateway
Science
User guidance

Note. Respondents could include multiple answers in an open-ended question format which were then listed separately. As a result, the total number of rows in the table may be more than the number of respondents.

Sharing Tool & Resources

Overall, a clear majority of respondents in all groups said that they would use Web portals to share the tools and resources that their projects produce. Respondents in the two groups that were previously involved with coordinating entities said they planned to use *Mailing Lists* and *Live Events* at rates exceeding the global average rate (see Figure 2).

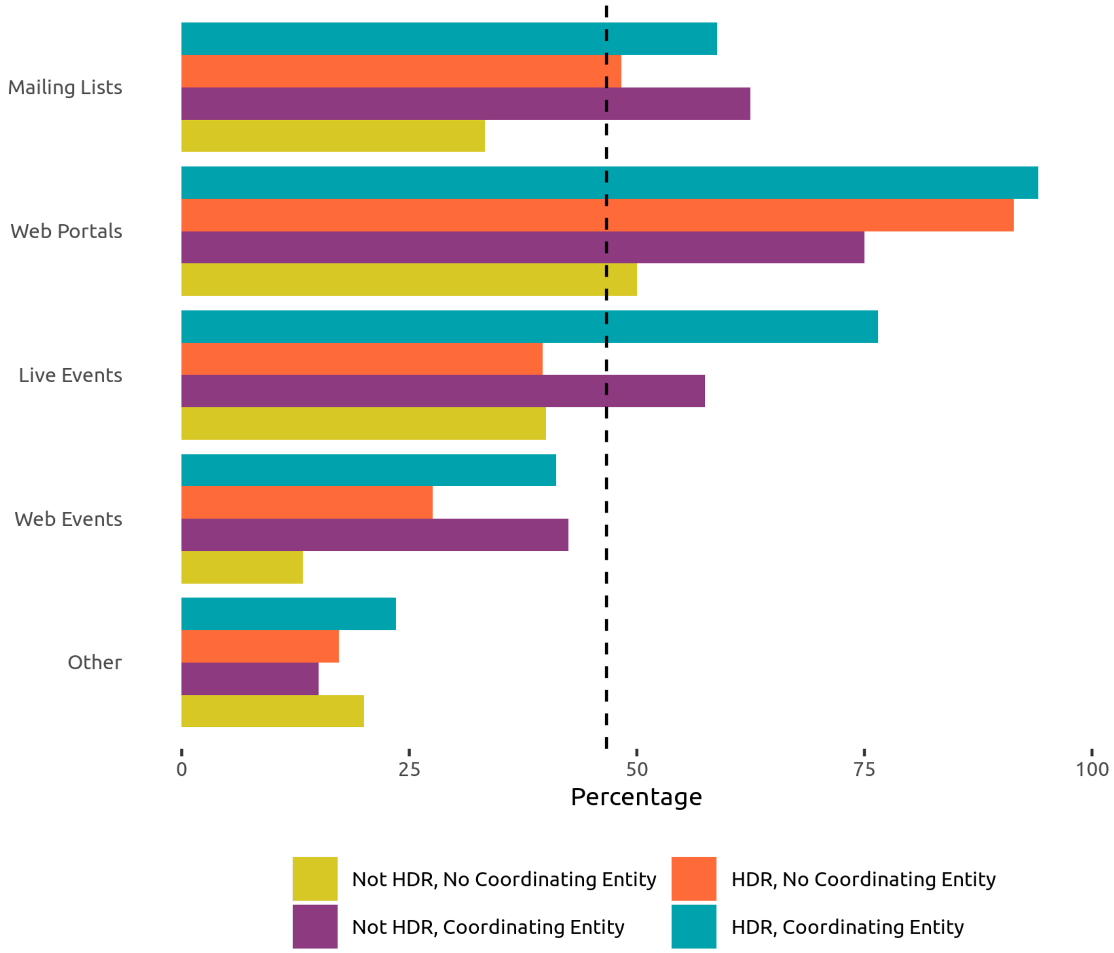

Figure 2 Distribution of responses to the survey question “What plans does your project have for sharing those tools/resources with others in the data science field?”

When we compared HDR grantees to each other and to other groups of respondents, we found that HDR grantees are more likely to report sharing tools and resources across a larger number of outlets ($\chi^2 = 5.79$, $p = .02$). Most of the difference can be attributed to using Web Portals ($\chi^2 = 15.86$, $p < .01$) and to HDR grantees who were involved with a coordinating entity ($\chi^2 = 10.25$, $p < .01$). Table 3 details other plans for sharing tools and resources that respondents reported.

Table 3. Other plans for sharing tools and resources with others in the data science field.

What plans does your project have for sharing those tools / resources with others in the data science field? Please select all that apply. - Other - Text
APIs
Blogs/Medium
Cloud resource collections
Colleagues at my community college
Conference or workshop (x3)
Courses or tutorials (x2)
Data dissemination
Data library
Direct sharing between partner institutions
Docker containers at Dockerhub
FTP and Databases
Github/Git repository (x 4)
Journal Publications (x2)
Professional Presentations (x2)
Public access hosting
Publication (x6)
Relationships with relevant projects
SVN
Twitter
Use but not produce tools
Youtube channel

Note. Respondents were able to include multiple answers in an open-ended question format and each response was listed separately. As a result, the total number of rows in the table may be more than the number of respondents. Duplicate responses (where more than one person said the same thing) are tallied in the above table using parentheses.

Finding Tools & Resources

On average respondents reported that finding tools/resources is neither *"very easy"* nor *"very challenging"* (*M* = 0±.06). Most respondents said that they look for tools and resources in three of the six specified outlets – *Code Repositories*, *Publications*, and *Conversations with peers*. We got mixed responses from participants when they were asked about looking for tools and resources on *Online forums* and by *emailing colleagues*. Overall, participants did not mention regional Big Data Hubs (BD Hubs) frequently as places where they look for tools and resources, but individuals who were involved with another coordinating entity appear much more likely to report looking to the BD Hubs (Figure 3).

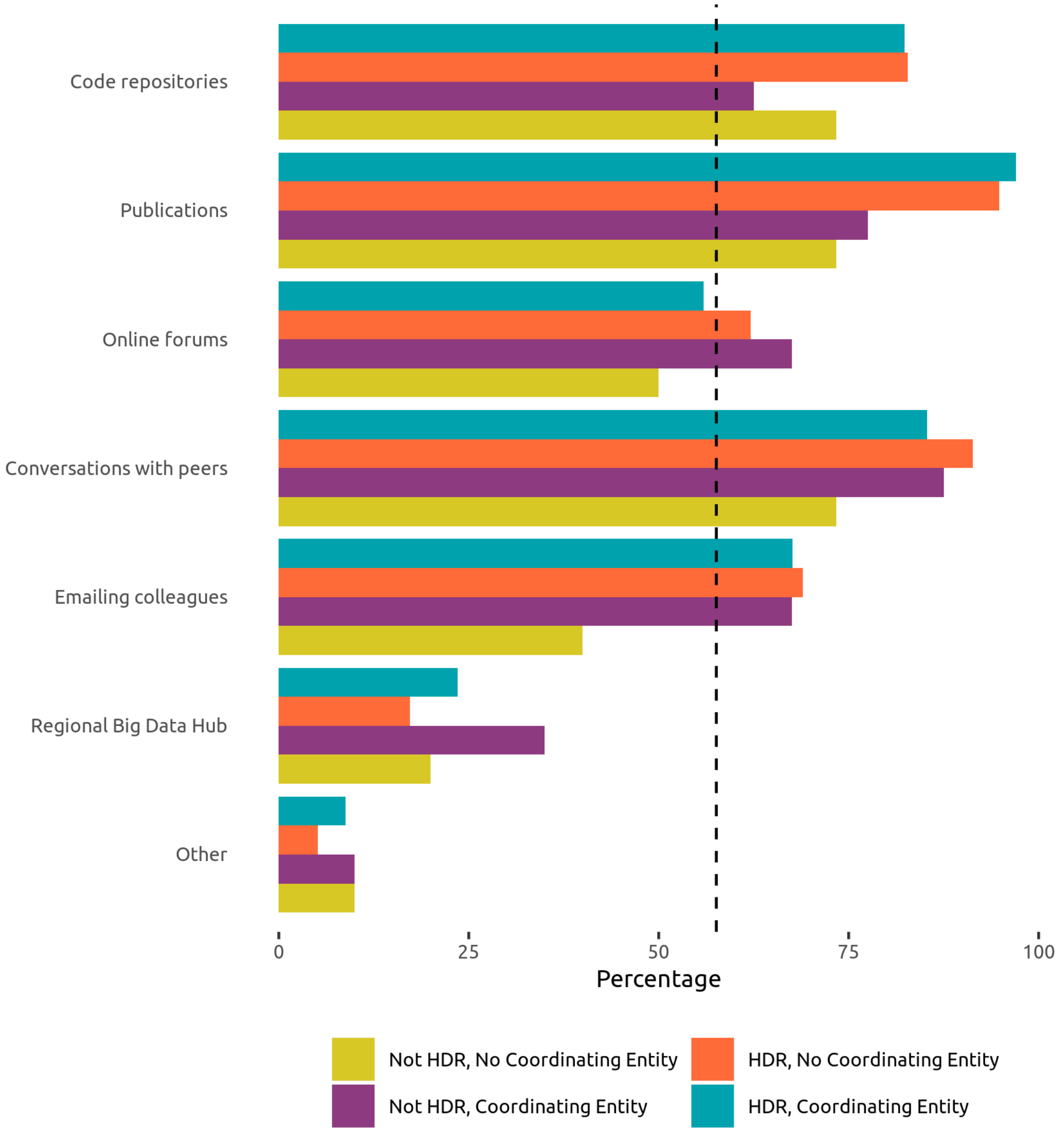

Figure 3 Distribution of responses to "Where do you look if you need to find, access, and use tools/resources in your area of expertise?"

The analysis found that HDR grantees were more likely to look for tools and resources from a larger number of outlets (χ^2 = 5.47, *p* = .01). In contrast, non-HDR grantees who were involved with a coordinating entity were more likely than other groups to look to Regional Big Data Hubs for tools and resources (χ^2 = 20.26, *p* < .01). Table 4 details other places that respondents said they turn to when they need to find tools for their data science projects.

Table 4. Other places that respondents look when they need to find, access, and use tools for their data science projects.

Where do you look if you need to find, access, and use tools / resources in your area of expertise? Please select all that apply. - Other - Text
BIG Math
Chemistry and Materials Community Data Project teams and websites
Conferences and workshops
Databases on web, repositories, publications/books
Github
Google (x2)
Government data repositories eg. Genbank
IASSIST list
Internet searches
Twitter

Note. Respondents were able to include multiple answers in an open-ended question format and these are listed separately. As a result, the total number of rows in the table may be more than the number of respondents. All duplicate responses (where more than one person said the same thing) are tallied in the above table using parentheses.

Advancing Data Science

In general, respondents rated all options quite close to *"very helpful"* on a sliding scale when asked what action would advance the impact of your work across the data science field. On average, HDR grantees tended to rate items slightly higher than other respondents ($\eta_p^2 < .01$, $F = 4.56$, $p = .03$). In most cases, this effect could be attributed to HDR grantees who were previously involved with a coordinating entity ($\eta_p^2 = .02$, $F = 46.12$, $p < .001$). While these differences exceeded chance, they were quite small (Figure 4).

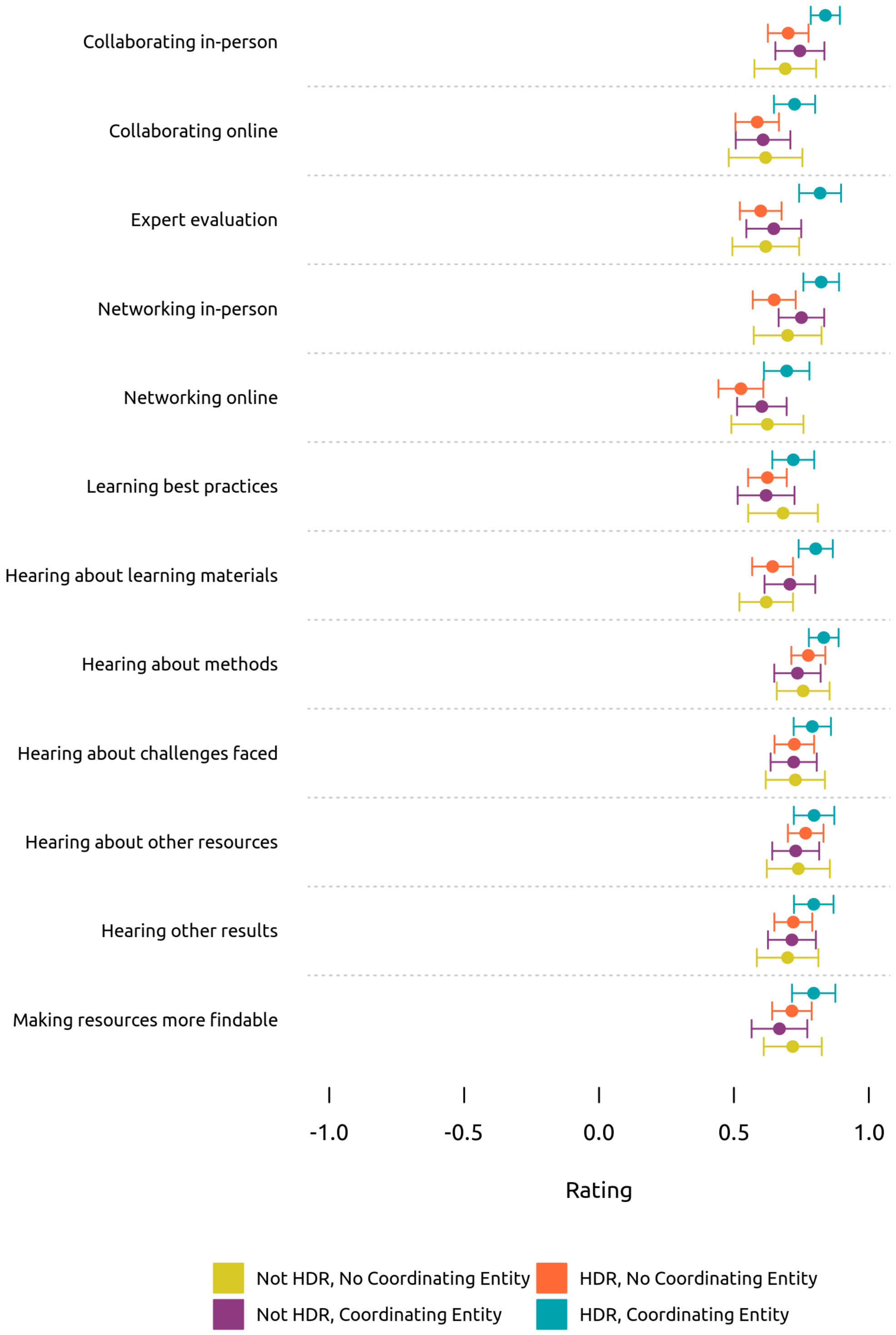

Figure 4 Distribution of responses to “How helpful would it be to do the following in order to advance the impact of your work across the data science field?”

Functions of Coordinating Entities

Most respondents selected four functions as important for a coordinating entity to perform. Eighty-five respondents chose *Promoting collaboration within the network and beyond*, with an average rank of 1.82. Seventy-nine respondents chose *Facilitating networking across disciplinary fields*, with an average rank of 1.87. Sixty-four respondents chose *Providing engagement mechanisms to broader communities underrepresented in data science*, with an average rank of 2.08. Sixty-three respondents chose *Serving as a vehicle for sharing best practices and resources to prepare the research workforce for careers in data science*, with an average rank of 2.06. Any apparent group-level differences in choosing and ranking functions of a coordinating entity did not exceed chance occurrence (all _p_s > .2). Other functions that respondents wrote in are presented in Table 5.

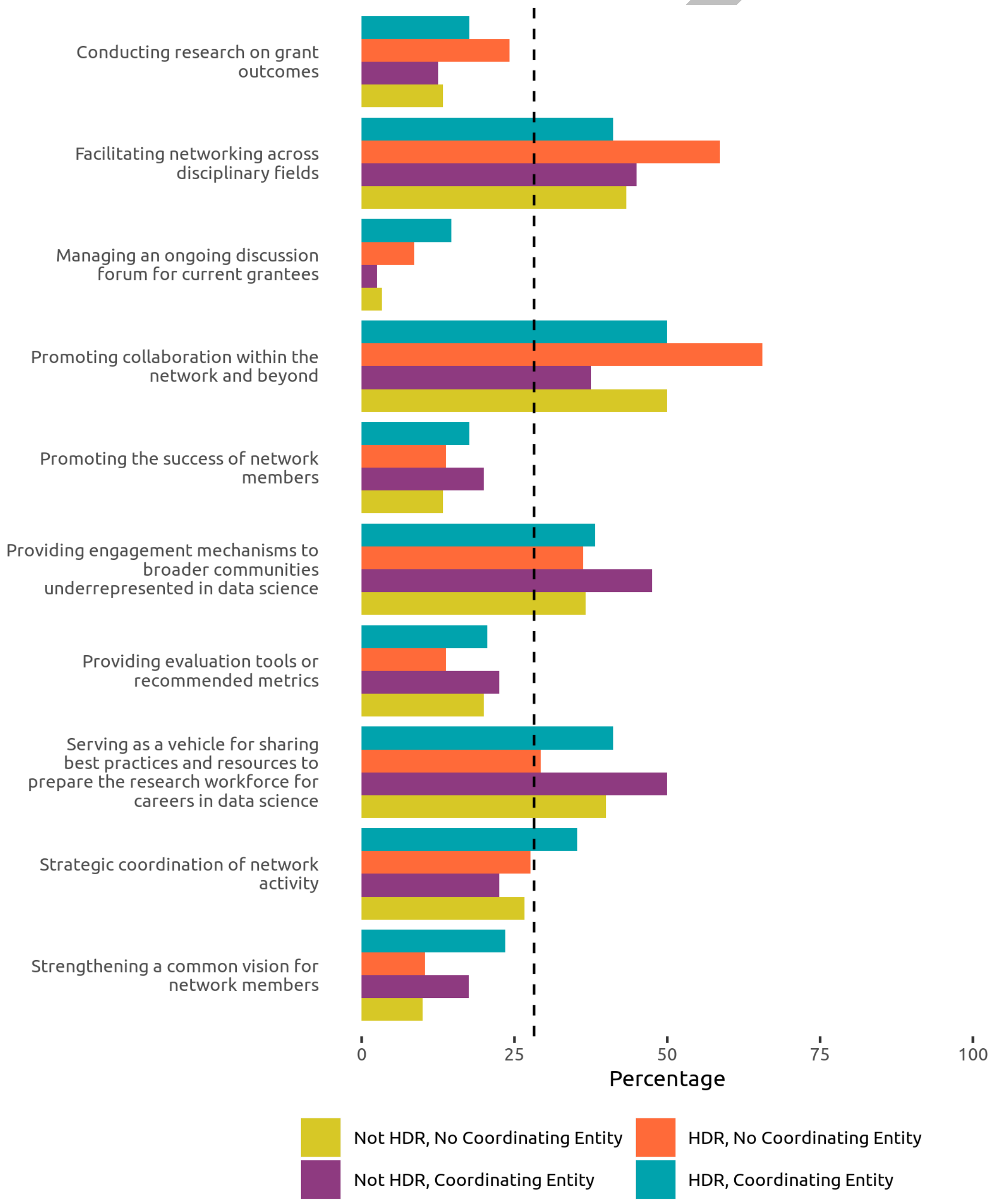

Figure 5 Distribution with which participants selected various functions of a coordinating entity selected as important.

Table 5. Other functions that respondents felt were missing from the above lists.

Are there any functions missing from this list that you think are important to add? - Yes - Text
Advising providers of resources on priorities and opportunities
Coordinating grant applications
Developing network funding mechanisms
Don't just focus on those who have grants -- it will be all men (based on the NSF Big Data PI meetings); if we want to increase diversity, we need to include those who haven't yet received a grant too, so they can learn what is needed to be successful at receiving grants! Building the collaborations they need to write proposals.
Encouraging and facilitating data sharing
Engaging directly with professional organizations central to data science, such as the ASA and others
Given the current system, anything that helps non-grantees information that helps them become grantees and helps grantees do their own work better could help.
Introducing researchers at various stages (including students and postdocs) to successful other projects: people, tools and outcomes
Making it easy to access/share data resources, and access to compute resources
Promoting collaboration across disciplinary fields
Providing funding
Providing reviews of evaluation tools or recommended metrics
Providing tools and recommendations for researchers from different backgrounds communicating more effectively.
Technical resources
The details of implementing the functions listed above can make a big difference. For sustaining systematic collaboration over a longer period of time is challenging without clear incentives from funding agencies and home universities/organizations of researchers/experts
Training, internship opportunities for faculty
Understanding what lessons/tools/techniques/practices/etc. are common across different disciplines in data science, and what's discipline-specific.

Textual Responses

Methods

We did a textual analysis of responses from participants who indicated in the survey that they had been involved in professional networks with a central coordinating entity. In the survey, we asked two open-ended questions about existing coordinating entities. The analysis for these questions was informed by the same categories of functions that we used as response options in the survey. These categories provided an initial framework used to code the responses to the questions. Additional codes were added as themes emerged from the data.

Due to the specific nature of the categories used in the survey (*e.g., promoting collaboration within the network and beyond*), some codes were used more extensively than originally intended. We examined potential differences in responses based on whether or not respondents had a leadership role in the additional coordinating entity with which they were involved, but ultimately this did not prove an important distinction to factor into the analysis as differences were negligible. All the codes for the question asking respondents if they had advice for NSF were emergent, but the results largely fit into the overall categories in the instrument. In this report, much of the analysis for this question intentionally includes quotes from survey respondents.

Results

Other Coordinating Entities & Their Impact

Survey respondents who indicated prior involvement with some other coordinating entity were asked about the main role and impact of that other entity. Both questions were open-ended allowing respondents to respond with specificity and nuance.

Sixty-four people answered the first question about the entity's role. A small number of respondents named the coordinating entity (*n* = 13). The list of entities included the Long-Term Ecological Research Coordination Office, Environmental Data Initiative, Earth Science Information Partnership, NSF Cybersecurity Center of Excellence, Earthcube, and Harvard Catalyst.

Nearly half (*n* = 30) of the respondents described the main role of the coordinating entity as **enabling social interactions** between members, either by promoting collaboration (*n* = 18) or facilitating interactions more generally among network members (*n* = 12). This aligns with what most survey respondents chose as the most important function of a coordinating entity (see results for *Promoting collaboration within the network and beyond* in Figure 5 above). Specific collaborative activities that respondents mentioned included working together on proposals and publications, as well as creating working groups. The second most common role mentioned by survey respondents was **sharing of best practices and resources** among members (*n* = 21). Respondents largely described resources in general terms, but a few specified details about research studies, use cases, data, software, and hardware.

The same number of respondents said the main role of the coordinating entity is to **organize events for members**, and also **creating products and deliverables** (*n* = 17 in both cases). The products they described included grant proposals, publications, and research projects. The events that respondents listed included in-person conferences and workshops, and virtual events such as webinars. The former code overlapped to some extent with the pre-existing code describing **coordinating network activity** (*n* = 15), when the language implied coordination was aimed at organizing events.

Survey respondents also wanted a coordinating entity to act as a **hub or a repository** for resources (*n* = 8) and to focus on **providing services** such as consulting, education, guidance, and feedback (*n* = 7). The pre-determined codes about **providing engagement mechanisms for underrepresented communities** and **strengthening a common vision** received almost no mention among responses to this question. However, emergent codes revealed that respondents envision several other roles for coordinating entities. These include creating knowledge, providing research support, promoting a specific area of study/discipline, providing communications generally, providing funding, managing membership, and providing an infrastructure or framework.

We note here that **Providing engagement mechanisms to broader communities underrepresented in data science** was ranked highly in the list of important functions of a coordinating entity, though almost no respondent mentioned it as part of their experience with existing entities. This indicates a chasm between current and desired practice.

Sixty-three people responded to the question: *What was the main impact that the coordinating entity had upon advancing the field it was intended to support?* Nearly all the respondents described positive impacts from being engaged in the coordination entity. Two respondents mentioned negative impacts, which suggests that they need stronger support in their future engagements with the entity. Four respondents were unsure of the impact at the time they completed the survey.

We saw a similar pattern in participants reflections on the impact of coordinating entities. About half (*n* = 29) highlighted the **social interactions** that the coordinating entity enabled. Roughly equal numbers among this group appreciated the entity's efforts to facilitate network interactions (*n* = 15) and promote collaborations (*n* = 14).

Most respondents offered general descriptions of the collaborations that the coordinating entity fostered but some did highlight specific products that resulted. Fifteen respondents reported **creating products and deliverables** such as new grants, publications, patents, and research projects. Some respondents said that the entity facilitated the **sharing resources and best practices** (*n* = 11) including ideas, tools, and lessons learned as well as helped them access and disseminate different resources.

Nine survey respondents could articulate **perceived impact on a field or area of study** of a coordinating entity. They described how it had grown the field, increased its reach, or built a specific community. A similar number of respondents felt that the entity provided a novel **organizing framework or infrastructure** that unified similar areas/fields of interest (*n* = 8).

Another group said that the entity helped grow the **professional capacity of members** by building their skills and competences (*n* = 6). Specific skills mentioned included becoming more politically savvy to know how to get funding and being better at doing self-assessments.

The pre-determined codes about **providing engagement mechanisms for underrepresented communities**, **coordinating network activity**, and **strengthening a common vision** were not as useful for coding the responses to this question, suggesting coordinating entities' impact did not lie in these areas, yet a handful of respondents highlighted additional categories of impact. Specifically, they described the ability of the entity to facilitate communication, create knowledge, organize events, provide funding and training services, manage membership, and create a central hub/repository.

Advice for NSF

We received 84 responses to the question, *"What additional advice do you have for the creation of a new coordinating entity for data science research grantees? For example, what could a coordinating entity do to better facilitate collaboration?"* Most of the responses

(60%) were from those currently working on HDR projects (*n* = 50). Fifty-four respondents had previously been involved with other professional networks that have a central coordinating entity. The two groups (current HDR awardees and those with prior coordinating entity experience) were not necessarily overlapping – only 27 respondents who gave advice belonged to both groups. In this section, we have reported the respondents' feedback in significant detail to provide the clearest possible picture of the community's current feelings and perspectives.

Responses to this question largely mirrored what participants had already said when asked about the various functions of a coordinating entity. We coded responses according to six main areas: Pursuit of a Shared Mission, Events, Communication, Collaboration, External Sector or Community Engagement, and Access to / Sharing of Resources. A final code designated comments that described the management or operations of an entity, rather than its ideal function. (Often comments fell into more than one category, so the total number of comments about each topic is greater than the number of respondents.)

Nineteen respondents wrote about the need for an entity that pursued a shared mission. The word "mission" seemed more apt for this code than the word "vision" (which had been pre-selected and used elsewhere in the survey instrument) since what was meant is less of an abstract ideal and more a set of actions that would advance an agenda. Indeed, many of the responses included specific tasks that respondents felt would support that mission. For example, respondents said that the entity should pursue concrete solutions to real-world problems that would have a meaningful impact on the data science field, and end users. Respondents also wanted the coordinating entity to be involved in establishing best practices and success metrics, identifying and coordinating funding opportunities, and promoting the field to young professionals with an overall commitment to diversity and inclusion.

Fourteen responses gave feedback related to events that the coordinating entity could offer to the community. All but two comments came from current HDR awardees. A few suggested doing in-person events like *"face-to-face meetings"* or an "*annual symposium"* but others preferred the efficiency of online options for webinars and training. A midway point between these was the suggestion to host *"relatively local gatherings"* to minimize the travel burden on participants while still allowing them to benefit from in-person interactions.

Sixteen survey respondents gave advice related to communications – both communication from the entity to its stakeholders, and between stakeholders themselves. A recurring theme in these responses was a need for cross-sector or interdisciplinary communication that reduced domain-specific jargon in an effort to include researchers from a diverse range of backgrounds. Comments here focused both on the *content* of communication (e.g., existing work, best practices) as well as the *means* of communicating (e.g., a rotating author blog, Slack channels, or an open web portal). One respondent noted that the current system of emailing PIs and expecting them to distribute important messages to their team is not working, and another expressed a strong preference for a listserv over a periodic newsletter.

The nineteen comments coded as collaboration were closely connected to themes that surfaced about mission, events, and communication. Respondents highlighted a need for NSF to bridge across domains, and expand opportunities for collaboration with new partners, including past and potential future grantees. One respondent noted that *"HDR projects are multidisciplinary [and] a challenge is that communication between collaborators from different backgrounds (i.e. a materials scientist and a data scientist) can be difficult, delaying progress. A coordinating entity could play a role in bridging the important points across domains."*

One obstacle to collaboration that respondents noted was a system that perpetuates the *"insularity"* of current awardees. Also, one respondent lamented the inability to collaborate because of limited resources. To mitigate these challenges, respondents suggested establishing grant opportunities that require a collaborative component for funding. The corollary to collaboration that some respondents (*n* = 12) suggested was external sector or community engagement. Participants suggested moving from largely disciplinary-specific pursuits toward engaging academic and other institutions more broadly, alongside industry, commercial entities, funders, and publishers.

Some respondents also called for broader engagement with community members. The term "community" seemed to imply persons who would in some way benefit from data science research projects, such as users of software tools. Respondents offered three concrete ideas to engage with the wider community: having a *"charter that facilitates interactions with the broader community,"* incentivizing community participation through *"projects that provide funding for project participants,"* and having the coordinating entity *"provide timely feedback about community needs and interests."*

We received thirteen responses suggesting ways a coordinating entity can facilitate access to or sharing of resources. Comments reflected sentiments shared elsewhere in survey feedback – that there is a clear need but no clear example of something working well (e.g., *"github is a morass of bad info and practices"*). Specifically, the coordinating entity would collect and categorize resources, learnings, and tools. It would then facilitate access by providing the "best search algorithm possible." Respondents suggested that the proposed resource portal should include short descriptions of tools and methods being developed by HDR awardees, methods that could have useful applicability elsewhere, seminar videos, and software.

The final code for the question about advice designing a coordinating entity focused on its operations and management. This code category received the most responses (*n* = 30). Most responses came from people that were experienced with other types of coordinating entities (*n* = 21). In general, survey respondents want an entity that is **simple, efficient, flexible** and **transparent** – not bogged down by bureaucracy, placing an undue burden, or duplicating efforts. As one respondent said, *"Keep it lightweight...An over-eager entity trying to do too much will just be a nuisance."* Respondents also wanted the efficiency that cloud-based systems can provide, and leadership who are comfortable with an online work environment. Participants also wanted diversity leadership roles, and to avoid an imposed *"top-down"* structure – they prefer one that is *"bottom-up"* and community driven.

Many responses highlighted a need for clarity across the board. Respondents want the coordinating entity to have a clearly articulated purpose and focus, and the expectations of awardees need to be clear as well. Furthermore, awardees participation in the entity's activities should provide both short- and long-term value.

Some responses highlighted ways that the coordinating entity could foster collaboration. As one respondent said, *“Collaboration is harder than working solo, but it's rare for organizations to reward the extra effort. If you want to facilitate collaboration, find ways to provide significant, meaningful, positive incentives and rewards for collaboration.”* Moreover, respondents also highlighted a need for an accountability mechanism to ensure that collaboration is actually taking place.

A number of respondents expressed skepticism that a coordinating entity would add value to the field without becoming overly burdensome. One respondent said, “*The challenge will be to provide value without creating too much extra work for everyone involved.*” At least one respondent offered an alternative to the current system of Regional Hubs, which they thought were *“spread too thin across too many topics, and don't seem to coordinate well with each other,”* suggesting the creation of *“Synthesis Centers”* instead for key areas within HDR. “*Creat[ing] Synthesis Centers… may result in greater re-analysis / re-use of the data*.” Time management was also a pervasive concern for multiple respondents with at least one individual noting that they “*would have to be convinced of the usefulness of a coordinating entity. We are all too busy otherwise*.”

Some participants suggested that it may be more fruitful to partner with organizations that already have coordinating infrastructure in place. One respondent said, *“Engage with professional organizations that have been coordinating and facilitating extensively, rather than continuously reinventing the wheel. It is tiresome to always have yet another ‘coordinating entity’”.*

Whichever approach NSF adopts, the feedback from this survey suggests a coordinating entity should be developed in response to the needs of the people it is intended to serve. We recommend that the ongoing input collection process – through virtual and in-person events – build on the results of this survey and zero in on specific stakeholder needs. As one respondent said, *“The coordinating entity has to solve a problem that the grantees feel they actually have. Do they need some kind of mechanisms for accessing and storing common data sets? Do they need help with communications about their work? Can the coordinating group help build community by working with those who are not grantees to write successful proposals? Depending on what is needed by the community, any of these could be helpful.”*

Microlab Virtual Discussion

The Microlab Virtual Discussion, held on March 16, 2020, was designed to facilitate conversation among small breakout groups of HDR stakeholders. Participants discussed topics related to the idea of collaboration, responding to a series of four prompts that were developed based on survey findings:

1. What types of successes have you had, or seen, in collaborative data science projects? i.e. between groups with disparate expertise for example?
2. What are the opportunities you see as most pivotal for elevating collaborative research within the Harnessing the Data Revolution Ecosystem and beyond?
3. What are the gaps you see as most challenging for collective action within the Harnessing the Data Revolution Ecosystem?
4. In your experience what are the most successful mechanisms for achieving bigger scientific grand challenges with data that involve coordination /collaboration?

Participants were organized into 16 virtual breakout groups for discussion. Breakout groups took notes in a Google Form. These notes were exported into a .csv file and then compiled as an appendix to this report. (See Appendix A) This section of the report contains no standalone analysis, but the Microlab notes were used to inform the development of the Conference and were considered for analysis in developing the synthesis presented in the synthesis section of this report.

Conference Summary Themes

The following themes emerged from the Harnessing the Data Revolution PI Conference, conducted virtually April 28-30, 2020. In breakout sessions, participants discussed sub-topics and took notes on their conversations. The researcher based the summaries presented in this document directly on the notes from the conversations.

The summaries are organized into two main sections in this document. The first section focuses on scientific research – in particular opportunities for collaboration. The second section illustrates the various processes needed to support a robust data science ecosystem.

Section 1. Opportunities for Collaboration in Data Science

Conference participants worked in self-selected breakout groups according to interest in various areas of research and application. They were provided with the following questions to guide conversation, though each conversation followed a unique course and discussants often did not provide answers to these prompts.

- What's the problem/opportunity?
- What's your idea?
- What's the potential impact?
- What are your first steps?
- What expertise is needed, if any?

Theme 1.1 Science-guided machine learning

In this conversation, the group attempted to determine what data science approaches participants are using for their application domain(s) and find commonalities across domains. The conversation included open questions or roadblocks in combining scientific knowledge with machine learning, that could be explored together, and the potential for a new machine learning paradigm that could incorporate different types of scientific knowledge (ensuring that the vocabularies are inclusive of different science domains). The group explored types of application domains conference participants are working on, possible types of domain knowledge (e.g., ontologies / taxonomies), the ways domain knowledge types can be characterized (e.g., deterministic / probabilistic or theoretical / empirical), and the different types of formulations for integrating scientific knowledge in machine learning (e.g., feature engineering). Participants also shared examples of success stories of science-guided machine learning, and papers or other resources they would like to share, which are relevant to science-guided machine learning.

Theme 1.2 Online data science education best practices and evaluation

The group conversation acknowledged that, due to COVID-19, many training activities that are usually done in person are being forced online. Challenges with online training include ensuring adequate technology and internet, how to do soft skills training, communicating about Capstone projects, making online content accessible to students with disabilities, assessment approaches, and limited capacity of faculty who are being pulled in many directions. The group discussed opportunities for aligning efforts in this area and shared strategies to mitigate challenges, for example, pooling best practices for a) online teaching b) online meeting moderation, and c) managing an online capstone project. In a second session focused on the same topic, the group produced the following: a list of resources that would be useful to everyone teaching data science online, and evaluation questions that participants would want to explore across awards.

Theme 1.3 Learning from multi-modal data; Fusing heterogeneous and multiscale data

This group considered how to develop a domain-independent (generalizable) framework for causal modeling and prediction of future system behavior. All domains impacted by time have issues related to the problem of finding a unified framework (e.g., humans, corals, and climate). The first step toward such an effort would be to start with dimensionality reduction.

Theme 1.4 Disparate datasets

Participants discussed how data scientists often need to connect or reconcile disparate datasets, in particular higher-resolution data sets and old historical legacy data sets that are not interoperable. This would result in the ability to capture data at very different scales, resolutions and time periods. The first steps toward accomplishing this goal would be to define all the potentially relevant datasets applicable to a particular domain area, then reverse engineering complex datasets. Participants emphasized the need to be very clear about data extraction methods for generating any kind of data, and that expertise in data mining, data curation, data integration, and data visualization would be required for success.

Theme 1.5 Integrative data equity

Breakout group participants considered the term "Integrative data equity" to capture the technical and societal challenges of building models from diverse heterogeneous sources that may be untrusted, noisy, low-quality, and biased in complicated ways. They noted that irresponsible use of data science techniques and technologies can reinforce inequities and introduce biases, while giving the illusion of higher accuracy and better performance. This can happen by obfuscating the context in which data was collected or by introducing new sources of bias as a side effect of integration itself. This group's idea was to develop a framework for integrative data equity systems that aim to bring these issues to the forefront at all stages of design, development, deployment, and monitoring. The components of the framework can be software-enabled services for unbiased data discovery, combining datasets to make them more representative, measuring fitness for use for particular tasks, warning labels for bias. This would result in general and operationalized techniques for exposing and controlling equity issues at all stages of the data pipeline will be broadly useful across fields and problems. Better exposure to equity issues can help educate students and practitioners about the potential harms, leading to more advanced and nuanced discussions about the role of technology.

Theme 1.6 Organizing different aspects of materials science data

Conference participants talked about how to integrate data from different fields, deal with sparse or missing data, and connect machine learning algorithms towards a more uniform platform in order to increase accuracy. To do this, they suggested creating an open source platform to collect, label, store, and expand materials data, and create a digital twin for virtual experiments. This would result in a range of impacts, with benefits across multiple domains and on industry and society, such as the acceleration of commercialization of new technology.

Theme 1.7 Identifying general data issues unique and associated with biological data

There is a critical need to link and integrate diverse and heterogeneous data streams to answer questions across biological scales. However, biological data are ‘dirty,’ heterogeneous, and have multiple issues that limit science. Discussants identified a need for a common tool kit, approaches, and methods to integrate biological data, and the ability to finally address multiple grand challenges in biology from genomes to the biosphere. To do this successfully, expertise will be needed in data science, computer science, taxonomy, comparative biology, and macroecology.

Theme 1.8 Ecological forecasts, protecting biodiversity and human well-being

The key questions addressed by this conversation were: How do we make research integrated with and relevant to policy making? How do we ensure the general public is educated to understand the implications of research? How much diversity do you need for ecosystem survival and success? The main idea of the conversation was to integrate social science with environmental science: use the analogy of hidden factors that contribute to human diversity dependent on socioeconomic / environmental factors and experiences versus the factors that influence the survival of an ecosystem. This would result in the cross-fertilization of disciplines, such as researchers learning from ecology and applying it to human health. The first step would be to define quantifiable traits and their interdependencies.

Theme 1.9 Balancing machine learning with other data science methods

Participants in this discussion were concerned with the dominance of machine learning approaches and worked to identify the ‘decision points’ for choosing a particular method, noting that sometimes simple classical algorithms are more clear, intuitive, and garner more physical insight. The complex systems community has a heuristic that in the absence of robust information (e.g., quality and representative data) then you must use the simplest model. The group was interested in the possibility of creating a ‘controlled vocabulary’ for machine learning and saw potential for the future HDR ecosystem to include a working group of individuals across HDR projects to develop and maintain a controlled vocabulary to be used across the cohort. Ultimately, this would result in better collaboration across the HDR cohort and set an example for transdisciplinary data-driven projects to structure more effective interaction.

Theme 1.10 Critical indicators in complex (trans-domain/human-natural/adaptive) systems

The natural-human world is characterized by highly interconnected systems, in which a single discipline is not equipped to identify broader signs of systemic risk and mitigation targets. HDR can be a cohort/organizational structure to share tools that can more capably identify the times, data, and periods where a system is behaving abnormally. The risk and collapse of our critical systems can be identified from newly capable means to quantify the unusual behavior - i.e., identify the risks. This would result in a better understanding of critical risks and the protection of systems across society. The first step in this initiative would be to develop a 'database' of ideas for identifying risk.

Theme 1.11 Uncertainty quantification in data sciences

Breakout group participants asked: How can we develop data science to statistically link the incomplete/imperfect information available from both experiments and theoretical models to deepen our understanding of the behavior of physical, (bio)chemical and materials systems?

Section 2. Building a Robust Data Science Ecosystem

Conference participants were asked to envision ideas deemed as important to supporting an HDR data science research ecosystem. Each participant worked in a self-selected breakout group according to interest and each thematic area represents multiple breakout group conversations. Discussants were provided with the following questions to guide conversation, though each conversation followed a unique course and discussants often did not provide answers to these prompts.

- What is the main short-term actionable idea?
- How does it advance or enable the HDR ecosystem?
- What are the ideal outcome(s)?
- What is the opportunity for long-term HDR impact?
- How might we maximize inclusivity and accessibility?
- What motivates participation?

Conference participants were then asked to cluster their work which resulted in seven broad themes emerging as important to support an HDR ecosystem. Due to the volume of ideas generated by participants, the themes below are not comprehensive and do not include all ideas, merely examples to illustrate some of the important points.

Theme 2.a Data sharing & stewardship

Conference participants talking about data sharing and stewardship proposed a range of ideas, such as holistic repository services that would help with acquiring missing data. NSF investments in data repositories has been significant, but uptake, maintenance, and use vary across domains. Services to add value to these repositories would have broad reach. This idea highlights the need to move beyond a repository as just a collection of independent datasets and instead consider it a single data resource that can be queried and analyzed as a whole. In other words, a generational shift from "dead" repositories with little more than keyword search to "live" repositories with active services for integration, evaluation, visualization, learning. Another breakout group proposed a similar idea, noting that data structure should be pushed by the funders.

Theme 2.b Education, training, and pathways to careers in data science

Breakout groups discussed ways to engage with the public and private sectors to develop applied projects, and how to share best practices for teaching data science (in particular through hands-on learning). Providing real-world projects allows students to transfer and apply skills learned to develop a professional portfolio, and allows for networking for future job possibilities. Additionally, this benefits the project partner by helping them address their business needs and solve problems. One group suggested a data science education repository, the core of which would be a collection of "great stories in data science," similar to case studies in business schools. Another envisioned a summer undergraduate data science research online community, to help grow connections among members of the future workforce, and expose them from the very beginning to the plurality of experiences that are required for data science. It will also encourage collaboration between the different sites running educational experiences for undergraduates and results in sharing of best practices.

Theme 2.c Interdisciplinary collaboration and team science

The ideas that surfaced in these breakout groups largely focused on the communication needed for successful collaboration. For example, the groups suggested developing a common language between domain experts and one that communicates with the common language of data scientists. This idea extended to agreeing on a unified language and framework for datasets, so that data analysis tools are universally applicable across datasets. Bridging the language gap between what domain scientists need and what the data scientists can use hack weeks as a model for data science education and collaboration. Other ideas that surfaced included a platform for coordinating conversations between HDR awards, developing a collaborative community within HDR to identify and document best practices and cutting-edge innovations in leveraging data science for domain science, and shared mentorship responsibilities (enabling teams to be greater than the sum of their parts). Implementing these various ideas would have broad and significant implications, including the creation of lasting relationships between students and investigators across institutional boundaries, and extended professional networks to support development rapid transfer of ideas and approaches across boundaries. The ability to solve problems faster without "reinventing the wheel" could be a motivating factor for participation, as could the ability to solve complex domain science challenges and address societal needs (e.g. COVID).

Theme 2.d Impact through community engagement

Discussants agreed upon the immense value to identifying and understanding a project's community of stakeholders, since the results of a project affect not just the PI team, but also collaborators, taxpayers (who should benefit from federally sponsored research), and people who don't identify as scientists or know what NSF is. This group proposed systematically enabling and training PI teams to do stakeholder mapping as a way of advancing HDR as well as supporting equity, diversity, and inclusion -- and help drive outcomes that are relevant to a broader community. Another group thought that including someone who studies human behavior in the project teams to mediate the connection between technology developers and stakeholders could be an important part of building a network of stakeholder types and examples of functioning collaborations. Currently most HDR ecosystems are missing this connecting human who is willing to play the role of mediating between needs of stakeholders and student / faculty capabilities. This "human integrator" would shift the focus from what data analysts could do to something that they should do based on the greatest societal and community impact, thereby saving resources and increasing impact. A related idea that surfaced was outreach – even if stakeholders are identified, many scientists are not trained to effectively translate their work into useful insights or actual impact to the broader public. A focus on outreach could facilitate deeper insights about collective capacity. Under steady state or crisis situations (such as COVID-19), it would be easier to see how we might leverage the power of the broader scientific community.

Theme 2.e The best practice hub

Conference participants reiterated that HDR is all about connecting domain scientists with appropriate tools, expertise, and resources. A best practice hub could help guide researchers in the selection of algorithms or machine learning methods in the same way that recipes are rated for their suitability for novice versus expert chefs. Additionally, a website or framework could serve as a "dating app" for data scientists to find each other. Systematic adoption and application of the best practices and methods would enable more effective collaboration and transparency, lower entry barriers for graduate students and new researchers, lead to better understanding of algorithms for addressing different problems, and provide training pathways and a new workforce of data science-literate scientists.

Theme 2.f Incentivizing convergence

This theme highlights what conference attendees identified as a disparity in incentives across projects, since data scientists and domain scientists typically have different incentive structures. They thought NSF may wish to fund professional curation to create access to "results" of funded projects – perhaps supporting an institute to curate data, create keyword searches, create a large dictionary of keywords, and provide researchers with a way to identify areas of synergy. They also suggested that NSF can help ensure that data management plans are followed up on (and not just ensure data are posted, but is good quality and reusable), and use previous follow through on data management plans to make future funding decisions. They would also appreciate knowing what good administrative management and team management means for an HDR project, and the consideration of soft metrics to ensure collaboration. Soft metrics could reward, for example, collaborations of domain and data scientists improving the overall ecosystem. These steps would contribute to building sustainable infrastructure that supports long-term goals.

Theme 2.g HDR Directory

This breakout session focused on developing a "Matrix of Competency," as a central coordination unit for the HDR group to coordinate expertise and activities. The matrix was envisioned with domains along one axis and methods on the other, and participants placed themselves as points on this matrix. The tool for building this matrix needs to be clear, flexible, and editable by anyone in this community; it also needs to make use of searchable and accessible “tags” (e.g., domain expertise, data type, or problem tags).

Synthesis, Key Findings & Discussion

This chapter integrates the raw data presented and summarized in the previous three chapters. Each activity (Stakeholder Survey, Microlab, and PI Conference) contributed valuable input into this research process, capturing critical feedback from a breadth of data science researchers. The analysis presented in this section responds to the question: **Where do we go from here to further advance a robust data science research ecosystem?** We note again that at the outset of this research process, the goal was to inform the creation of a national coordinating entity. Over time, however, this transitioned into documenting the many opportunities available to support the field. The breakout group sessions from Day 3 of the HDR PI Conference feature prominently in this analysis, as conference participants were asked explicitly to think about how to advance the field by identifying and stewarding *"thematic needs of the HDR community."* In this exercise, thematic needs were defined as *"something that will advance the HDR community, or something - without which - data science research cannot meet its full potential."*

Methods

To perform this integrated analysis, we used a grounded theory approach that iterated between the raw data and emergent themes. In other words, a researcher reviewed all data and noted key ideas as they appeared. As those ideas were refined through aggregation and synthesis, the researcher continually revisited the original raw data to validate the resulting analysis. We note that very specific "Actionable Ideas" generated by Conference breakout groups have been captured in Google Docs. As a complement to the ideas presented in this synthesis, we recommend further review of each of the Actionable Ideas as a basis for determining next steps. While this analysis surfaces overarching themes, it does not attempt to systematically organize the many specific recommendations participants had for advancing the data science research field.

Key Findings

The key findings below are a blend of what is currently happening and a "wish list" that, if operationalized, would greatly increase the impact of data science research efforts. We also note that no one idea described below exists in isolation. These ideas are and should be treated as parts of an interrelated whole.

Collaboration between data scientists & subject matter experts

A recurring theme across all three events is that effective collaboration underpins long term success of the data science field. Indeed, the top-ranked choice selected by survey respondents when asked to prioritize various potential functions of a hypothetical HDR coordinating entity was ***"Promoting collaboration within the network and beyond"*** (*n* = 85).

The data revealed a clear need for – and the challenges of – work occurring between so-called "domain" scientists or Subject Matter Experts (SMEs) and data science researchers. The two have an inextricably linked yet sometimes fraught relationship. SMEs are needed to validate data analysis results, yet they might be unfamiliar with the analysis techniques used to generate the results. In the words of one Microlab participant, *"I found the most successful aspect of the collaboration is that the domain scientists are able to give us some deep insight on feature selection and data representation,"* indicating that they successfully use Python Materials Genomics, or Pymatgen, to reduce the work involved in data preparation. Data science-SME collaboration was related to another similar theme: Interdisciplinarity or a "team science" approach, where people collaborate across scientific domains. Respondents to the Stakeholder survey selected "networking across disciplinary fields" as the second most important function of coordination efforts for the field. Data scientists, once they understand the needs of a particular domain, can effectively guide practice and apply applications or solutions to additional domains.

Participants had various thoughts on how to promote collaborative efforts. These often centered on more effective communication. In particular, this means developing a shared vocabulary across domains and technological approaches, one that is not beholden to the jargon of any single discipline. One group labeled their Actionable Idea as *"Create a concrete approach to making an HDR 'dictionary' to facilitate interactions,"* noting that *"too often words are used across a team that all assume has a specific meaning but it varies across domain or system."*

Framing education and training opportunities

One of the keys to a long-term sustainable data science research field is ensuring that viable pathways exist for students, trainees, early career researchers, and academic scholars to advance. Reimagining how this occurs can facilitate interdisciplinary collaborative work. As described by one breakout group: *"Involvement of graduate students - including framing research problems that combine domain science and data science and are tractable within a Ph.D. thesis - has been a very efficient way to speed up the knowledge transfer between scientists from different areas."* Interdisciplinary collaboration can be achieved in education programs that integrate aspects of multiple domains as well as data science, or when students work with multiple PIs. For this to occur more widely, participants expressed a need for interdisciplinary training modules. They also suggested that students and post-docs could invite faculty to serve as advisors or mentors, participate in student committees that cross institutional boundaries, or participate in lab rotations outside of their field (domain versus data science). These efforts will facilitate the transfer of concepts and break down communication barriers by providing a specific context for exchange to occur – resulting in what one group called *"polyglot"* trainees.

Data resulting from the three activities in this research showed that an important aspect of education and training is cultivating an external orientation – i.e., data science students should be encouraged early on to think about how the profession aims at real-world impact through problem solving. Engaging with the public and private sector to develop applied projects allows students to transfer and apply their skills and develop a professional portfolio, along with networking vital to future job prospects. It simultaneously provides benefits to the external project partner by addressing their business needs or solving a problem. Respondents also felt that real world experience would help students engage with new types of data and apply some of the skills they are developing in school. One group noted that for this approach to work, students or trainees need funded opportunities, especially those at minority serving institutions.

A major theme in the conference data – likely due at least in part to the conference timing during a wave of Covid-19 infections and the cancellation of in-class instruction at universities across the country – is ambivalence toward online education. Conference participants noted a host of issues that merit further consideration:

- How to adapt activities for a virtual format, how to ensure equitable access to instruction and resources for all students (including fundamentals such as a stable internet connection),
- How to determine what "effective" teaching practices are in an online learning environment, and
- How to assess student learning outcomes.

One breakout group noted some particular challenges for professors, such as having to learn new platforms so as to be able to assist students, and coordinating Capstone projects online. They noted that professors are feeling taxed and overwhelmed. Conference participants had some thoughts about how to mitigate some of these challenges. This includes supporting both synchronous and asynchronous learning for students as well as pooling resources for best practices in online teaching, meeting moderation, and managing online capstone projects. However, additional efforts are required as online learning becomes a more substantial part of the data science ecosystem.

Re-thinking the data

Unsurprisingly, participants across all three events were concerned with data and thought the field would be served by deliberately considering how data can be used most effectively. It was noted that data sharing can be especially problematic when those who collect the data are incentivized to keep it private. Furthermore, data are collected, stored, and used in different ways, and as such there is a need for guidance or standards around overall data management. The field currently struggles with the lack of standardized approaches for "*data engineering from storage, compression, access, and the architecture design of this entire pipeline*." Specific areas meriting deeper consideration that Microlab participants noted were clear documentation of data and metadata, good data hygiene, analysis-ready data for benchmarking computational tools, clear documentation of software and assumptions in algorithms, and dedicated support for data curation and clean up. Participants also spoke about the need to consider approaches to deal with heterogeneous data, connect or reconcile disparate datasets, and integrate data from different fields. These are all aimed at ensuring data are usable and accessible by various research groups. Data interoperability was noted as a major issue for interdisciplinary work. As one breakout group noted *"making datasets talk to each other is something that everyone seems to need."* According to one group, operationalizing efforts to improve data quality means creating measures for assessing data completeness and quality, automated quality checks, support for professional data curation, and a discussion forum for data quality and improvement that provides a "*two-way street between repositories and users.*" Participants also advocated for developing incentives and assessment strategies to encourage and reward open access of datasets.

Identifying best practices and creating repositories

Survey results indicated that coordinating efforts should prioritize sharing best practices and resources to prepare the research workforce for careers in data science. Other findings indicate that best practices are needed not just for workforce preparedness, but for virtually every aspect of ensuring a robust data science research ecosystem. One breakout group labeled these best practice compilations "*guidebooks*," and added that they must avoid jargon and be mindful of biases. One specific area for developing best practices that came up in multiple groups was ***assessment*** or ***measurement***. Essentially, they felt it was important to create a mechanism for the field to collectively define and evaluate its impact. Respondents felt that an important part of establishing evaluation best practices is acknowledging the need to re-think the meaning of "*success.*" This could include, for example, recognizing interdisciplinary work or factoring in real-world impact metrics (e.g., policy implementation) for tenure review, rather than more traditional academic metrics.

Broader cross-sector engagement

The word sector here refers to all of the different potential stakeholders and audiences involved in data science research and its outputs. Just as students benefit from projects with a "real world" application, this finding acknowledges that for data science research to have societal impact, it must ultimately transcend institutional boundaries in a consideration of roles and identities.

The third most frequent coordination function identified in the survey was "***Providing engagement mechanisms to broader communities underrepresented in data science***." (*n* = 64). This means diversifying the field of those doing data science research. It also means translating academic work into societal impact – a major focus for the regional Big Data Hubs. Broader engagement also means developing pathways to seek investment and attention from industry or government partners, both of which will serve to further legitimize and extend the influence of the field.

Another idea that surfaced in multiple conversations was a need for stakeholder mapping. The breakout group discussing the theme of "***Impact through community engagement***" considered a wide array of potential stakeholders (extending, ultimately, to all who benefit from federally funded research). They proposed systematically enabling and training PI teams to do stakeholder mapping as a way of advancing HDR as well as supporting equity, diversity, and inclusion. They also felt that this would help drive outcomes that are relevant to a broader community.

As with the key finding on collaboration, respondents felt that effective communication is central to cross-sector engagement, yet scientists are often not explicitly trained in how to communicate the results of their work to various audiences. Current vectors for dissemination and the output reward system further limit broader community engagement.

Specifically, respondents noted that peer-reviewed journal publications – the standard for communicating findings to the field and the benchmark for how researchers are judged – have some problems. In the words of one group, "*Publicizing work is very challenging because the subject matter journals often think that the work is too technical, and the data scientists feel that the work is not technical enough.*"

Discussion

Initially, this research project was designed to inform the creation of a national coordinating entity. Over time, the project transitioned into documenting the many opportunities available to support and expand the data science field. The information synthesized in the current chapter highlights several avenues for continued growth and improvement including concrete suggestions for possible next steps. A recurring theme through all the data and responses collected for this project was the need for greater collaboration and effective communication between the different stakeholders. Participants said repeatedly that both of these are crucial for sustainable success in the data science field.

Along these lines, participants suggested ways to frame education and training opportunities to improve communication and foster greater cross-disciplinary collaboration. The list included creating interdisciplinary training modules, providing opportunities for students and post-docs to work with faculty as advisors or mentors. They felt that efforts like these will facilitate the transfer of concepts across domains and help break down communication barriers. They also noted that implementing these efforts, even if students initiate them, will likely need administrative or professor support. Respondents also felt that providing opportunities for data science students to participate in ***"real-world"*** projects would be beneficial for the field as a whole. These projects would provide space for students to practice the skills they are developing in school, and expose them to new types of data perhaps increasing the students' enthusiasm for continued learning. Some participants noted that for this approach to work, students or trainees need more funded opportunities.

Equally important are standardized ways of doing data science and communicating information and concepts across disciplines. Various participants highlighted the creation of repositories as central to the HDR ecosystem to ensure collaboration, eliminate the need to "reinvent the wheel," and maximize federal resources. Furthermore, respondents often discussed best practices in relation to creating various types of repositories. For example, a guide to best practice in evaluation could be accompanied by a repository of validated instruments housed in a separate GitHub repository dedicated to evaluation. Participants also suggested creating repositories for completed Capstone projects with lessons learned, as well as for PhD theses, story-based case studies in data science, and one that would bridge tools / capabilities and applications/needs across domain areas.

In terms of communicating their results more broadly, respondents felt that academic publications are not the only type of useful research output – indeed, a vast array of possibilities exists – and should not be the only standard for benchmarking researchers' performance. Breakout groups recommended developing standards and processes for attributing and crediting research outputs other than journal articles. They also suggested developing identifiers and standard citation practices for all types of research products. They noted that creating various kinds of research products is fundamental to the pursuit of broader cross-sector engagement.

Response from the Field

This section to be completed following the open comment period that allows HDR stakeholders to provide feedback on the findings in this report.

Knology

Behaviors

Biosphere

Culture

Media

Wellness

Systems

Knology.org
info@knology.org

tel: (442) 222-8814
3630 Ocean Ranch Blvd.
Oceanside, CA 92056

tel: (347) 766-3399
40 Exchange Pl. Suite 1403
New York, NY 10005