Knology

Big Data Innovation Hubs National Coordination Committee: Appendix

Advancing Data Science Research: An Integrated Analysis

July 31, 2020

Rebecca Joy Norlander & John Voiklis

Advancing Data Science Research: An Integrated Analysis is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License.

Recommended Citation Norlander, R.J. & Voiklis, J. (2020). Big Data Innovation Hubs National Coordination Committee: Appendix: Advancing Data Science Research: An Integrated Analysis. Knology Publication # NSF.159.648.03-A. 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 research 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-A

Table of Contents

Appendix A: Microlab Breakout Group Discussion Notes

The following data have been aggregated across all Microlab breakout groups and edited slightly for clarity and to remove all individually identifying information.

What types of successes have you had, or seen, in collaborative data science projects? i.e. between groups with disparate expertise for example?

- Academic interests v. practical problems (i.e., industry). Industry issues not being academically interesting/complex enough; Notwithstanding, they need to be addressed Mixed expectations across professional and what counts as "research"
- As a data scientist working on inverse material design using machine learning approaches 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. My collaborator suggested our group use Pymatgen, and we found that by using Pymatgen, we can save a lot of work for data preparation, as many of the things we try to do have already been done in Pymatgen.
- Being able to work together
- Building shared vocabulary
- Choose the same programming platform (Big Data Hubs have found success through Jupyter Hub)
- Collaboration involving graduate students working with multiple PI's.
- Continual communication and openness. Do a lot of roundtables and structured inquiry. Listen to what people care about and what their concerns are; then, try to build programs to meet those needs. The best we can do is establish trust and to like the people on the team. Makes it important to invest in relationships, especially when/if there are difficulties among the team members.
- Convening in-person
- Critical is the willingness to step outside of your zone of comfort, for example in our project none but one PI knows anything about corals, yet we have all read up on the literature and share interesting new articles. We also bring different degrees of what we consider important. For example, standardization of experiments as a reliable setup to test hypotheses derived from computational predictions has been traditionally more important to molecular scientists but is equally important to all collaborators now.
- Cybertraining in atmospheric data science
- DataONE: we have scientist subject experts, tech experts, data scientists, information scientists and social scientists working together to build the cyberinfrastructure to make it useful for all the groups as they may collaborate moving forward. We used a working group approach with teams carefully organized to bring expertise and connections to the group. We also had specified liaisons to cross the groups. This was multi-institutional across the US and globe.

- Different structures for data science education for project members at different career levels (senior, early career, graduate, undergraduate)
- Early on, we invested a substantial effort in tutorials and developing a common challenge. This may have slowed down the early progress, but is helping over time.
- Exposing students to other disciplines
- Exposure to thinking about social impact, social good, ethical issues, rather than just using the coolest, latest tool
- Find a shared problem that may not necessarily be the same in the details, but is the same in its fundamental structure; for example, image informatics. Bring the right people across fields together in a match-making workshop who would not normally interact. We have found interesting shared challenges, and resulting discussions, doing this in a symposium in early February with domain researchers from biology and materials science, and data scientists across fields ranging from robotics to mathematical modeling for predictive experiments. Make sure there is good representation from both data and domain scientists; discussions about both the structures and approaches (what they can and cannot do) as well as the domain applications.
- Find a team where everyone has expertise in data science tools (Jupyter notebooks, cloud, Python stack, etc.). Then it is easier to find a common language around those tools. For example, let's use deep learning to analyze some set of images from astronomy and neuroscience. Once there is a common language, it is easier to innovate.
- Finding partners who can adapt their science to different topics
- Frameworks: math and chemists and stats; jargon challenges; YouTube webinars to share understanding across groups; Gateway with ed materials; workflows topological analysis, Exede; Sharing beyond main group; Also DOE experience with mixed data; domain variance in scales and between experimental; making things more friendly
- Had worked for several years with material scientist, had several papers, but never got any funding. Subject Matter Experts (SME) did not have sufficient data for a data scientists' needs. Realized that this would ultimately not be productive. Need an SME with a lot of data. Many times, simple methods in our world are very amazing for the SME.
- Has a lot to do with the mindset of the people involved, that they are willing to work with other people. Successful groups tend to be those where each person is looking for an answer from someone else.
- Have domains present in interactive manner (e.g., Jupyter notebooks) to entire team and provide resource after the meeting for individuals from other domains to work with directly (and build); Interact via the data technologies
- Host small interdisciplinary workshops to give different disciplines a chance to interact closely (50-60 people); Where do we find support for these workshops? What resources are there for this support? Perhaps a specific route for such funding that has quick turnaround (for anywhere between 20-60 people)?

- I've had successes with collaborative data projects outside of HDR. With our HDR project, we are successfully meeting remotely and conducting activities, but also facing some challenges.
- IDEAS Lab Autonomous Computing Materials and IDEAS Lab Accelerating Synthetic Biology Discovery: We have bi-weekly meetings amongst all groups via Zoom, PIs and students and postdocs, each week 2 groups present, rotating alphabetically. Data scientist is facilitating conversations between groups focused on sub-domains, and the dataset of interest, plus data science approaches. Sub-groups of 2-3 research groups/PIs are focused on sub-areas/thrusts, to make progress in early days without being limited by availability of data or understanding how to process it. Enabling interactions between different domain specialists in ensuring that the teams are focused on using data to do the research effectively. Slack has been helpful as well to increase rate/frequency/ease of communication between PIs-students/postdocs and students-postdocs/students. Pre-existing datasets also very helpful to pilot data analysis approaches.
- IDIRSE Power Systems Project (Framework Project ASU+TAMU team): Great success so far between computer scientists, data scientists, and power engineers meeting weekly to discuss how to tackle storage of 100TBs of streaming PMU data and what power engineers would actually like to see in visualizing a very large data set in real time. Data Science aspects have consumed the researchers.
- 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.
- It's also important to have a large and well-connected community. Such that people can work together in different subgroups as their interests evolve.
- Juypter Notebook from Domain Scientists to allow data colleagues to build on talks
- Knowledge Guided Machine Learning for Modeling Complex Systems (Frameworks), Biology Guided Neural Networks for Identifying Phylogenetic Traits (IDEAS Lab): New methodologies unfolding on how to bring together traditional scientific knowledge and paradigms and machine learning methods across physical and biological sciences, through the paradigm of knowledge (science)-guided machine learning.
- Learning/ed oriented framework grant HDR: common workflows; data repositories made available to get folks up and running quickly; what measures success? Includes social scientists, data folks and LDI to share with K-12 teaching and educators for both teaching and analysis; some data sharing restrictions
- Like each other; SME value the data scientists; structured & planned communication; SME needs to have a lot of data;
- Machine learning person working with particle physicists – developed project that got funded; growing community
- NASA Frontiers Development Lab: Data + Domain scientists together for a specific project where they are effectively able to communicate, and actually produce successful outcomes. The idea is to put the data and domain scientists in the

same space and give them a problem to solve in a sprint setting, to come up with solutions.

- Neuroscientists, computer scientists, mathematicians – developing understanding between disciplines
- NSF research -focused on Materials but a panel on Data Science Education; Clear data transfer rules; half is using Dropbox or university Google Drive; starter kit for teams; Toolbox dialogue initiative - change the pace of science.
- One participant has 10 PIs on her award. Has weekly calls with everyone on the award. Will have 30-45 min someone will present. Finds that in an interdisciplinary setting, they can learn from things being used in other fields. Can borrow from what's been done in other fields. Having diverse team (5 men & 5 women) is working well.
- Our team has been successful for small-group collaborations among two PIs. Some projects are done and some are ongoing. And the next step is to integrate efforts from all 5 PIs.
- Power grid: industry data (NDA restricted) sharing is very tricky. Reproducibility could be an issue going forward
- Projects must be innovative on both domain and data science side, which requires adaptability, change mechanism, and lots of communication
- Slack for quick communication, sharing short reviews and progress items
- Slack site- 50 workshop very supportive; funding with quick turnaround for workshop for HDR projects
- Still figuring out blueprint of successful projects, diverse ecosystem of attempts
- Team science is critical - Convergence Accelerator brought in Michigan State University Toolbox Dialogue Initiative was hugely helpful on a different data science project
- TRIPODS: students must have shared advisors across disciplines; overcoming jargon gap; joint seminars; 4 departments; Institute for cross-disciplines
- University of Idaho, as a data scientist working on inverse material design using machine learning approach. 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. As for software, we try to homogenize. Our group use Pymatgen, and we found that by using Pymatgen, we can save a lot of work for data preparation, as many of the things we try to do have already be done in Pymatgen (Python Materials Genomics)
- We collaborate with computational materials science to develop algorithms for generating new materials. Our interaction has generated many interesting ideas that take advantage of latest deep learning techniques and has potential to solve the new materials discovery problem.
- We identified that in order to have a successful collaborative data science project, the team would need the following:
    - (1) develop common language so that people from different discipline can talk with one another;
    - (2) team members should have "soft skills" in networking and communication;

- (3) team members should be open minded to work on new projects and be willing to learn;
- (4) there should be a well-defined project goal;
- (5) the team can draw from the disciplinary networks of each participants to get a broad grasp of the individual fields;
- (6) clear leadership and sustained effort towards a well-defined project goal

What are the opportunities you see as most pivotal for elevating collaborative research within the Harnessing the Data Revolution Ecosystem and beyond?

- Collect data from successful PIs this round to help identify mechanisms to help project teams work on soundly tested ways to communicate, interact, and speak the same language
- Create a platform or channel (e.g., slack) for different topics, either a specific domain science question, or a specific type of data type, or a general type of methods.
- Create data repositories; Connections within the HDR community; Understanding structure of data and domain questions; Where can students go for future training; More and better ways to disseminate our work both within the statistics world and the SME field
- Difficult to send only two representatives to PI meeting – (or is it three?) – need a way for everyone to share what they are working on – each group could make a one-slide run-down of what they are working on.
- Difficulty in building infrastructure (dealing with large microscope images – how to collaboratively build the dataset) – resorted to sending around hard drives because downloading is so time-consuming
- Ensuring sound mechanisms to store and access datasets across multiple research groups/subgroups
- Help determine and communicate what the specific goals or sub-goals of the collaborative research are, so that teams can effectively move towards accomplishing those goals.
- Hosting workshops with common themes to facilitate discussion
- How to formulate a problem in another discipline into a data science problem. In materials science, the structure of the data is not what we usually encounter. The "design" problems…figuring out what new material that is not known. We are trying to understand what the SME questions/goals are.
- How to train early career people (students/postdocs)?
- I am eager to start the process of forming trial institutes involving PIs from different areas of the HDR ecosystem.
- Identify a shared space (Google Drive?) for info to share with potential collaborators at the institute level, opportunity to see if there are overlapping interest areas
- If data repositories or databases could be made available, funded by a federal agency, then that could be very helpful to push the field forward. Another opportunity is that there are many talented people, but it's difficult to figure out

who is doing what. We submitted one-slide, but these were never distributed. The April meeting will be very helpful to forge connections.

- It will be good to have general best practices (guidebook) for collaborative research spelt out: Don't use jargon, be mindful of biases etc.
- It will be nice to also have the NSF decide 1-2 themes — this will help focus and push people into more productive collaboration mode
- Large NIH dataset -- reconciling different data types from 12,000 brain scans
- Look at how certain algorithms in viz/data science is applicable in diverse applications. (Reuse is really important).
- Maybe several small My Discipline 101.
- Need (shared) training modules for interdisciplinary training
- Need a small group of people "devoted" to a more focused theme
- Need to communicate to broader community aspects of their projects that could be extended to /benefit other groups. Discover connections between groups.
- NSF can help increase interaction/communication among the funded projects. Each team can broadcast gaps and see who can help fill the gaps.
- NSF can help to organize and ask the domain scientists to summarize their key research questions to answer and what data they have: the challenge list.
- One of the challenges is they don't have enough data (ground penetrating radar) to see sub-surface objects. You can collect data but don't know the truth of the subsurface. Don't have ground truth data. One opportunity is to develop computational tools to augment the data.
- Opportunities for looking for common core techniques that are transferable to other disciplines; It would be useful to develop data science training modules that are customized to fit the needs of individual domain disciplines (e.g. data science 101 for other disciplines); Linking data science people with application groups and vice versa; Explore the application of algorithms to many different application fields; Possible mechanisms for students/postdocs to rotate/learn in other labs/projects; Coherent effort devoted to focused themes.
- Possible setup a mechanism where students/postdocs can rotate through different IDEA labs/HDR Institutes.
- Provide long-term support for graduate students. the 2-year frame for the NSF-HDR Ideas Lab is an unfortunate limit.
- Reusability of methods, reduce reinventing the wheel, pop-up labs, centralized effort and sharing in data science education
- Suggest having conversations in advance of the meeting in April to make that meeting most effective
- Team Science built in; money for workshops on a quicker turn around; Platform to look up what other types of projects are funded to group together similar teams; Database of capstone projects to have ideas from domain science and to save them and record impact after the fact -what was the dataset used, lessons learned.
- Training for SME to understand their own data and the kinds of questions that they are asking
- TRIPODS: good at connecting application folks with theory folks; suggestion that two PIs across programs must collaborate

- Use April meeting to establish cross-grant pollination for all kinds of HDR awardees
- We can identify specific technical problems such as, for example, “image analytics”. We can bring teams together around those technical problems. Perhaps there are astronomers and biologists all working on applying deep learning to image analytics. Perhaps another group is working on hosting image databases in the cloud. Another group yet perhaps is working on sharing image databases, etc.
- We need to have the scientific community to decide themes.

What are the gaps you see as most challenging for collective action within the Harnessing the Data Revolution Ecosystem?

- Aligning research goals within groups towards publication/academic productivity on shorter time-scales, with overall longer-term mission of the HDR, particularly on a 2-year timeline.
- Another gap is that of the HDR Frameworks projects being too short a time to allow sufficient time for ensuring the data engineering, research, as well as educational components.
- Awareness of other projects; Educating in a way others can understand; Outputs are not equally rewarded - publications vs hack-a-thons, software, organize tutorials in Artificial Intelligence (AI); give data and access to tools; measure impact (i.e. #downloads of software or commit edits) and get credit; transfer into social impact; connector course for Machine Learning, AI, + domain. Platform to compare courses such as: Domain Science + Material Science
- Big Data Hubs have found success through JupyterHub - do they provide tutorials and or resources to help teams? A 'starter kit' and corresponding teacher (maybe culled from Big Data Hubs community) for new teams in terms of data science resources and technologies guidance
- Challenge when individuals are using different programming platforms/languages - clear rules for data transfer can mitigate to some extent; Often the problem arises when different institutions attempt to interact (e.g., one university supports Box and other support Google Drive)
- Change in scholarly habits and incentives to encourage what we need to have a healthy ecosystem
- Coordinate activities for data science tutorials/demos/interactive talks. How do we make domain scientists aware of the tools and data scientists aware of the problems? How do these communities interact?
- Data analysis results require domain experts to examine to validate if they are meaningful. Conversely, domain experts might not be aware of the data analysis techniques used to generate the results (e.g. assumptions made, parameters used, etc.), which can be rate-limiting.
- Data engineering from storage, compression, access, and the architecture design of this entire pipeline
- Data security, especially if data sets are public or to be made public.
- Data sharing, not just data repositories or metadata. It is often that data are proprietary. How do we get people to share data that they have spent months/years collecting? There are major silos in education.

- Different expertise, language, publication strategies in different domain and AI area
- Encourage student exchange working in areas of different expertise among different groups & teams
- Ensuring sound mechanisms to store and access datasets across multiple research groups/subgroups
- Find the domain science problems that are of interest for both domain scientists and data scientists.
- Finding 'framework problems' to start the discussion. We've been talking about image informatics in this breakout session - what are examples of other 'framework problems' that we share and could spur collective action on data science approaches and their applications?
- Guaranteeing multi-year support for graduate students within the HDR framework to support training of multi-disciplinary Ph.D. overlapping domain and data science.
- Have a small team build a repository that each project team contributes. Data structure should be pushed by the funders.
- HDR is a risky venture, some short-term planning activities would really help to structure the likelihood of success for bigger ventures (institute awards). Some groups already have a strength in this area, but for groups who either do not have that background or would like to try a more experimental approach to organization there should be some flexibility built-in
- Help determine and communicate what the specific goals or sub-goals of the collaborative research are, so that teams can effectively move towards accomplishing those goals.
- Hierarchical multi-modal data is not like images/vision, but is crucial to many scientists, so how to overcome language barriers to get people interested in these problems? How do we see commonality even within our sciences? These are very intensive processes of community building, but some projects addressing?
- How to get data to be shared? i.e., privacy & cybersecurity; There are education silos everywhere, no standard approach; Provide standards for data collection; Publicizing and disseminating work
- If these diverse initiative PIs are only brought together once per year, they are unlikely to form deep collaborations, right now it seems reliant on things happening organically; Ideas lab was very planned and orchestrated to foster collaboration
- In some cases data are not yet available or ready for use in machine learning techniques.
- It would be helpful to organize workshops that are open to any other outside group or team in the HDR ecosystem
- Lack of availability of standard metadata across the domain, and agreed upon metadata standards across sub-domains, so that scientists from even within the same domain, but different sub-domains can communicate/collaborate effectively.
- Learning resources (maybe graduate courses or Massive Open Online Courses). Can NSF sponsor the creation of new MOOCs?

- More cloud compute credits upfront from the NSF; specialized access specialized compute for projects with less-than-top resources; sharing of best practices for dealing with managing big-data compute resources; Exede resources good model; Science gateway resource leveraging; minimize the number of sub-applications and shopping around for resources (need one-stop shopping)
- Not everyone collects their data the same way; usually it is the funders who provide the guidance/standards for data collection. Surprised that NSF has not implemented any efforts in this area.
- One challenge I see is in receiving credit. Because the application domain is or may be new to the data scientists there is a risk that credit for discovery will be attributed to the domain experts.
- One challenge is that we are all focusing on getting our teams going together. So there are few cycles available to identify other HDR teams to partner. On top of this there is the issue of the disruption due to the coronavirus coupled with the short time-frame of the NSF HDR grants.
- Organize PI meetings as less of a show/tell but more to facilitate collaboration formation
- Our biggest challenge in HDR is with infrastructure limitations with accessing the kinds of data we need.
- 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. Need better platforms for dissemination.
- Reproducibility in research, standard way to compare results
- Roadmap to transform academic impact into societal impact
- Sustainable model for data science education, fitting in data science into different traditional STEM majors (pre-requisites and avoiding repetitions), very high diversity in research background may slow down progress
- The organized building of data infrastructure is critical to the success and should be supported with funding from NSF. The success examples include the pubmed.org (funded by NIH) and the materialsproject.org funded by DOE and NSF.
- The size of the groups can be challenging to coordinate coherent efforts within the teams as well as through coordination with other outside groups and teams
- The timeline is short, very multidisciplinary teams must deliver in 2 years.
- Timing is a big challenge, especially for the short-term Ideas Lab projects; The size of group is also overwhelming, it is challenging to reach a coherent effort within our groups and coordinate with other groups in the bigger HDR community; Gap in thinking and/or expertise, communication could be challenging sometimes; Coordinate with different levels of expertise on a common topic.
- Traditional views about the pace of projects and the incentives (papers vs. products, such as publishing a data set or a software package). Can these types of projects pioneer a broader (and needed) change of incentive? Data scientists and domain scientists typically have different incentive structures. How do we measure impact? Software downloads and edits?
- What are the characteristics of problems that data scientists find interesting? There is a large energy bump to learn about a domain science application (and

vice versa) - how do we increase the probability on the front end that we are accurately matching the interest and expertise of the correct data scientists, and increase the percentage of engagement?

In your experience what are the most successful mechanisms for achieving bigger scientific grand challenges with data that involve coordination /collaboration?

- Access to computing; open, transparent code.
- Access to data; in the Chem context modelling and simulation data sharing is rare (too much work, not enough incentive); Sometimes data sharing not done because of possible value to creators; Some folks do want to share as means to collaborate in the future.
- Being charged to think big. Opportunity to ask a grand question.
- Building a vocabulary that is shared across the domains and technological approaches being converged to address the grand challenge. Too often words are used across a team that all assume has a specific meaning but it varies across domain or system. The consistent use of metadata standards is one example but there are similar problems even with unstructured data.
- Challenge in figuring out what other Ideas labs are doing -- more sharing of information would be good. Find out who is in a similar space.
- Changing the pace of projects. There is a traditional perspective on the pace of academic projects. However, these types of projects require acceleration (e.g., hackathons and accelerators/incubators) and new thinking (e.g., Moon Shots).
- Clearly delineate/decide whether you and your team/sub-team is in an exploratory, divergent, brainstorming phase, versus a phase of applying tools/methods in a specific direction towards publishable/deployable output, since modes of interacting/collaborating are very different in these different phases.
- Collecting the data in form that other people can use. Good data hygiene. Standards in data collection.
- Common data standards; clear documentation of data and metadata; good data hygiene, analysis ready data for benchmarking of computational tools; clear documentation of software and assumptions in algorithms; data quality, approach for dealing with heterogeneous data; dedicated support for data curation and data clean up.
- Create incentives for sharing
- Embracing curiosity. People will work on what they are passionate about. How do we match projects/ideas to the talent via aligned curiosity?
- Encourage recombination of groups to try to make connections and better determine underlying structures. How do we create a 'cocktail party' structure where people go from group to group in a very interconnected way, until the groups that have the most synergy naturally combine?
- External board to get active external big picture feedback, Open senior leadership that is not biased toward one domain; Go beyond the NSF (i.e. DOE or others) some key data scientists should work closely interrogate with domain experts to come up with the key challenge problems and available data sources and formulate it into a data science problem.

- Find ways to minimize risk in electricity grid (don't be too reliant on a specific electricity source). Start at a smaller scale, say a state, and then figure out the best way to optimize
- Generating data is a problem in some fields, other communities (e.g. bio-research) have been able to get funding to generate data (e.g. NIH Common Funds).
- Having data scientists and computer engineers as part of the ultimate team is essential to success given the major challenges associated with big data.
- I think we have to experiment with forming institutes to learn what the successful mechanisms are.
- Including our customer in design of any software being developed for them to use for collaboration or data analysis rather than assuming we know what they need or will use.
- Make sure that the assumption behind data analysis methods are clear to everyone.
- Making datasets talk to each other is something that everyone seems to need
- Maybe just host a Zoom call at a fixed time per week (or rotating time to avoid teaching)
- Obstacles to success, learning each other's languages, SME often think that they have a lot of data but they don't. Sharing data among groups. Funding agencies have to fund these projects, work that can be used for several different groups and not just specific projects. Find a way to understand who is doing what.
- PI meeting: Take advantage of the upcoming virtual PI meeting to initiate the conversation between groups, including short presentation by different groups, and 'academic speed dating'. We would likely need a few of those meetings.
- Possibly have a meeting before the proposal deadline where proposers can present what they are thinking of doing, what they can bring to the project, to help find collaborators (DOD does this for some mechanisms)
- Resources for 'getting started' - there are many resources and much information about the buzz term of 'data science', such that it is difficult for anyone to know where to get started. Guided, personalized recommendations for getting going is critical
- Spend time learning the SME language; Funding projects that can be used across multiple groups; How different sectors' data may then interact with each other; Thinking about different frequencies of data collection
- Team management skills that encourage inclusivity (domains etc.), transparency (of workflows, constructions etc.), communication
- The NSF should invest in going beyond XSEDE (NSF program for data storage) to pass along to PIs tools, file-systems, storage systems, access strategies/etc. on the fly as well as offline.
- Trying to build data science institute -- would like to see a more experimental funding mechanism -- a short-term planning mechanism -- don't want to waste a lot of money on a collaboration that won't end up working out. New kinds of interdisciplinary collaborations are risky -- want to combine HDR Ideas labs
- Two bases to understand our work: data science methods basis and scientific problem basis. Sometimes figuring out that mapping is tricky for seeing commonalities.

- Two steps: can give short talk or write an abstract to see who is interested, then can see who shows up for longer talk. The first should be like scientific speed dating by pitching the problem and science more generally. Like talking to the general public, but talk more about the problem and structure of problem rather than too big picture. Mainly just no jargon.
- Virtual meeting where different teams present their ongoing research: domain challenges, the data science methods, and the current status. Useful for identifying teams to collaborate with.
- Virtual visits between groups? Virtual 'site visit' since cannot go to one another? Need to clarify what the appropriate size is.

Appendix B: Stakeholder Survey Questions

Q1. <Funded_HDR> Are you currently working on a project being funded by NSF's Harnessing the Data Revolution (HDR) grant mechanism?

- ❍ Yes
- ❍ No

Q2. <Resources_Create> 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.

- ❑ Curricula
- ❑ Software
- ❑ Publications
- ❑ Data Sets
- ❑ Other <*text box*>
- ❑ Project will not produce tools/resources <make answer exclusive>

<*Display for all responses to Q2 except the last option. Do not display if Q2 is skipped*>

Q2.1 <Resources_Share> What plans does your project have for sharing those tools / resources with others in the data science field? Please select all that apply.

- ❑ Mailing Lists
- ❑ Web Portals
- ❑ Live Events
- ❑ Web Events
- ❑ Other <*text box*>
- ❑ Project will not share tools/resources <make answer exclusive>

Q3. <Resources_Find> Where do you look if you need to find, access, and use tools / resources in your area of expertise? Please select all that apply.

- ❑ Code Repositories
- ❑ Publications
- ❑ Online Forums
- ❑ Conversations with peers
- ❑ Emailing colleagues
- ❑ Regional Big Data Hub

❑ Other <*textbox*>

Q3.1. <Resources_Ease> Please move the slider to indicate how easy it is to search and find the tools / resources you need.

very easy < ---------- > very challenging <*bipolar slider -1 to 1 with 2 decimals*>

Q4. <Activity_Interest_1/Activity_Interest_2> How helpful would it be to do the following in order to advance the impact of your work across the data science field? Please move the slider to indicate your opinion.

<A slider will appear for each of the following items. Those respondents who selected "Yes" to Q1 will see the HDR-funded project(s) text, those who responded "No" will see the generic data science project(s) text.>

Not at all helpful < ---------- >Very helpful < unipolar slider 0 to 1 with 2 decimals>

Making the **tools and resources** from my [HDR-funded project(s) | data science project(s)] more easily **findable** and **usable.**

Hearing about the **results** of other [HDR-funded projects | data science projects]

Hearing about the **tools and resources** produced by other [HDR-funded projects | data science projects]

Hearing about the **challenges faced** by other [HDR-funded projects | data science projects]

Hearing about the **methods being used** by other [HDR-funded projects | data science projects]

Hearing about **learning and workforce development methods/materials** developed by other [HDR-funded projects | data science projects]

Learning about the **best practices** for disseminating my work to broader communities.

Networking with others in my field **online**

Networking with others in my field **in person**

Connecting with experts in **evaluation** to understand the impact of my work.

Collaborating with others in my field **online**

Collaborating with others in my field **in person**

Q5. <Functions_Ranking> Here are some of the various functions of a coordinating entity. Please select three functions and rank them in order of importance to you in the box to the right. *<Ttems show up in alphabetical order for researchers to assess if they were changed.>*

- Conducting research on grant outcomes

- Facilitating networking across disciplinary fields
- Managing an ongoing discussion forum for current grantees
- Promoting collaboration within the network and beyond
- Promoting the success of network members
- Providing engagement mechanisms to broader communities underrepresented in data science
- Providing evaluation tools or recommended metrics
- Serving as a vehicle for sharing best practices and resources to prepare the research workforce for careers in data science
- Strategic coordination of network activity
- Strengthening a common vision for network members

Q6. <Functions_Other> Are there any functions missing from this list that you think are important to add?

❍ Yes <text box>

❍ No additional functions

Q7. <Other_Entity> Have you been involved with other professional networks that have a central coordinating entity?

❍ Yes <*continue with Q8*>

❍ No <*skip to Q9*>

<*If respondent does not provide an answer for question 7, skip to Q9.*>

Q8. <Other_Entity_Role> Please describe the main role of the coordinating entity.

<*text box*>

Q8.1 <Other_Entity_Impact> What was the main impact that the coordinating entity had upon advancing the field it was intended to support?

<*text box*>

Q8.2 <Other_Entity_Manage> Did you have a leadership / management role in the coordinating entity?

❍ Yes

❍ No

Q9. <Advice> 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?

<*text box*>

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