Making Mountains out of Molehills: Challenges for Implementation of Cross-Disciplinary Research in the Big Data Era

Daniel Andresen and Eugene Vasserman
Department of Computer Science, Kansas State University
{dan, eyv} @ksu.edu

We present a "Researcher's Hierarchy of Needs" (loosely based on Maslow's Hierarchy of Needs) in the context of interdisciplinary research in a "big data" era. We discuss multiple tensions and difficulties that researchers face in today's environment, some current efforts and suggested policy changes to address these shortcomings and present our vision of a future interdisciplinary ecosystem.

Big data, as noted by Dr. Francine Berman of the San Diego Supercomputer Center, is crucial to maintaining competitiveness in today's research environment. She notes, "More scientists will depend on exabyte data than on exaflop machines." Big data is also a new strategic advantage, and the new shared environments for scientists and researchers to explore. Virtually all of the NSF's *10 Big Ideas for 2019* relate to interdisciplinary research on big data.

Several overarching issues come to mind when considering interdisciplinary research centered around big data. For example, interdisciplinary communication is challenging, as similar terms may mean different things, and each party is to develop a sufficiently sophisticated vocabulary to be able to communicate across the research areas. Also, finding the right people with the right skills is particularly challenging. For example, each of the authors is in the department of computer science, yet even with our own discipline we had difficulty finding collaborators with the skills necessary to deal with large quantities of data in an efficient, effective manner. In addition, that funding agency has emphasized the importance of interdisciplinary collaboration, even going so far as insisting on it in any calls for proposals, yet adding significant barriers in the form of legislation such as CUI, ITAR, and other regulations. Cybersecurity and privacy also come heavily into play, as large datasets become increasingly likely to contain data which is plausibly personally identifiable, leading to dangers both in researchers learning information which is supposed to be hidden, and hackers causing potentially embarrassing and financially ruinous data leaks. Finally, comprehensive institutional support is frequently lacking, where the infrastructure is inadequate to support the desired scale and types of research, and the rewards systems for researchers fails to incentivize interdisciplinary research.

In examining interdisciplinary research, particularly in the academic environment, one must take into account the whole environment, and not simply focus on skills, Cyberinfrastructure, or other more easily tackled subproblems.

In this paper we first present a model for interdisciplinary research in a big data environment. We then discuss two overarching issues, interdisciplinary communication (both at the data and interpersonal levels), and the effects of the need for cybersecurity and privacy. We finish by offering several suggestions and observations for changes at the institutional level.

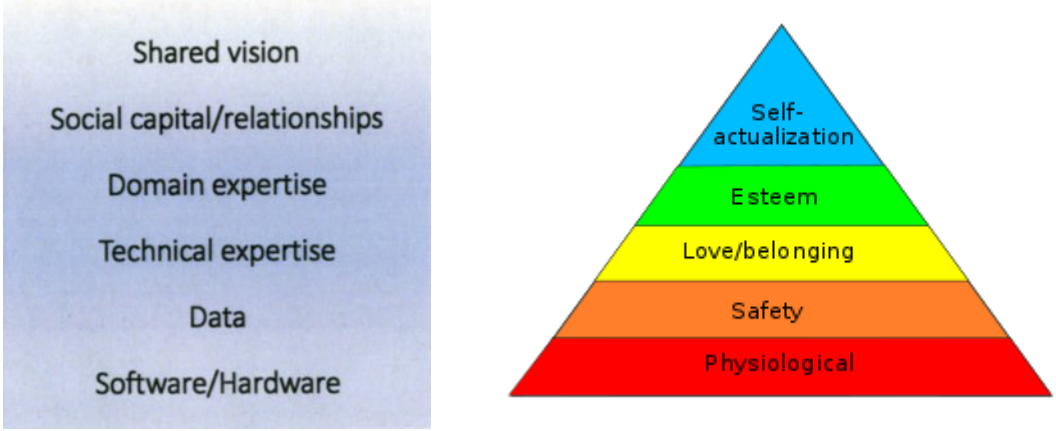

Figure 1. A Research Hierarchy of Needs (left), and Maslow's Hierarchy of Needs (right).

A Research Hierarchy of Needs

That realization drives us to introduce a Research Hierarchy of Needs loosely modeled on Maslow's Hierarchy of Needs[1]. And Maslow's hierarchy of needs, needs which are underserved at the lower levels of the hierarchy prevent the full expression of needs at higher levels in the hierarchy. So for instance if a person is in danger, the need for self esteem is substantially diminished. Similarly, in research, if the two researchers have a strong shared vision but lack the data on which to base their research, they will be unsuccessful.

1. *Shared Vision* The end goal for a relationship among researchers is that moment when they agree on the shared vision containing the research goal, outline the general plan, and realize they have the resources to accomplish it.
2. *Social capital/relationships* The shared vision is built on the ability to work together, to communicate well, and develop sufficient trust in the other to warrant risking valuable resources (e.g., time, effort, and expertise). Such relationships typically evolve over time, motivated by mutual respect, shared interests, and shared goals – frequently developing from loose, unofficial ties. Conversely, they can develop quickly if sufficient incentives (positive or negative) are introduced.
3. *Domain expertise* Frequently, particularly for larger projects, much of the actual work will be accomplished by a pool of experts – typically post-doctoral associates or graduate students – who are well-versed in the theory and experience in the associated research areas. However, in a big-data environment, we have noticed a serious shortage of qualified personnel with sufficient experience and knowledge of big data tools, leaving significant data analysis either unaccomplished or unattempted.
4. *Technical expertise* The domain experts cannot accomplish their goals, however, without a rich cyberinfrastructure ecosystem.[2] It is impossible to accomplish rich, large-scale data analysis on a typical academic IT infrastructure consisting of personal computers, Wi-Fi, and an Internet connection. Specialized hardware and

software environments that can easily handle multi-terabyte- to petabyte-level data analysis (both storage and computation) are required – and these require support staff (frequently referred to as "HPC Facilitators"[3]) who are experts at helping researchers effectively use these tools at scale. These facilitators are frequently domain scientists with doctoral degrees in related fields who have accumulated sufficient training and expertise to be effective in their hybrid role.

5. *Data* Clearly the basis for any research with the foundation in big data is the data itself. There are multiple issues that arise typically when acquiring the data needed to accomplish the research in question. The first issue is, of course, getting permission to acquire the data. As noted below, security, privacy, and tradition can all conspire to make getting permission difficult. Also, as the data is acquired, frequently sufficient metadata is not collected to make the easily searchable and curatable. Finally, after the data is collected, metadata attached (if this vital step is not effectively ignored – many researchers consider naming a directory 'lab3_day4_ir_rat_exp1' to be sufficient labeling), the cost of storing the data and making it available can prove prohibitive.
6. *Software/hardware* In the end, all data and analysis must occur in a sufficiently-scaled software and hardware environment. With local HPC compute resources available at every Top-100 research institution[4], and the benefits in productivity and prestige accrued[5], most institutions have cyberinfrastructure in place for previous-generation science. However, given the massive increase in scale required by today's research, institutions are scrambling to find the right mix of local and cloud resources, while also seeking effective funding models to support this vital expansion. Cyberinfrastructure is the new laboratory environment, and resources allocations (such as F&A) need to reflect this on par with more traditional efforts.

Interdisciplinary Communication

My colleague at Kansas State University, Dr. Doina Caragea, noted when asked about the toughest part of interdisciplinary research, "**I can't understand your problems, and you can't understand my possibilities**." Building the vocabulary and depth of understanding for sound interdisciplinary research is generally extraordinarily challenging, requiring significant investments in time and energy. However, university environments tend to be heavily siloed, typically by department, and as Kleinbaum notes, "…pairs of individuals that are in the same business unit, subfunction, and office location communicate at an estimated rate that is 1,000 times higher."[6] Overcoming a 1000x communication disadvantage is challenging, and workers even 30 meters apart have a perceived 1KM of distance between them.[7] Overcoming the siloes to build the social capital is needed to overcome the remorseless logic that in general, the short-term, annual-report-driven return on incremental efforts invested is likely to be better as extensions on existing domain successes rather than risky large-scale new projects. Building an understanding of each disciplines' vocabulary, workflows, rewards/

priorities, and arranging these into a mutually-beneficial project structure flows significantly more easily when the principals have an established foundation of trust and respect to build upon.

Cybersecurity and Privacy

Data sharing, and even data coexistence is challenging, especially as the data volumes increase and the amount of trust and collaboration between research groups decreases. The data may contain all kinds of sensitive information, from personally identifiable information, to financial records, to intellectual property, sensitive but unclassified, export controlled, etc. Careful design at the domain expertise level, and special security controls at the software level are required. An extreme case is that of complete separation (no group is allowed to interact with another) or even allow data analysis processes to coexist on the same physical hardware, which makes data sharing impossible but provides an excellent level of information leakage protection. The level of special controls and the associated difficulty in navigating them must be agreed upon at the level of shared vision (in terms of benefits and trade-offs).

In general, the more data sharing is allowed, the higher the risk of unforeseen information leakage. Pre-processing data before it is shared is an effective way to preserve privacy, but it is highly work-intensive, and is sometimes difficult to reuse once prepared: different methods of pre-processing are required depending on the types of analyses that collaborators would like to run on the data set. An example of pre-processing is an algorithm that adds noise to the original data while preserving desired statistical properties. Data redaction as pre-processing can also be effective, but requires significant domain knowledge to perform correctly, as the privacy properties of the redacted data set are heavily dependent on the type of data being shared. For instance, simple de-identification (removing names and identifiers) is usually not enough, and additional work is required to provide better anonymization[8] as demonstrated in practice by assigning names to user IDs in the Netflix challenge data set, which only contained film star ratings and randomly assigned user identifiers.[9] This was made possible by comparing the Netflix dataset (public but deidentified) to the IMDb data set (public but **not** deidentified) and inferring the Netflix user identities through similarities in watched films and star rankings.

Shared databases are fundamentally vulnerable to data extraction through the combination of multiple queries.[10] Some modern alternatives allow non-experts to query a database while the framework enforces privacy constraints.[11] This appears to be one of the more usable alternatives as of the time of writing, and allows the database to be fully shared, as long as it can only be queried using the PINQ platform.[4]

Institutional Support

We find the current research environments lacking in providing the type of comprehensive, universal support needed for today's interdisciplinary research. Steve Blank comments, "[I]nnovation is not a point activity, it's an end-to-end process. You need a pipeline."[12] We also see a strong competitive advantage for those institutions who can best enable their researchers (both professionals and students) in discovery, funding, and reputation. As such, we have several specific recommendations:

1. *Continue deliberately building opportunities for bridges across disciplines.* Funding – e.g., new NIH/NSF interdisciplinary CFPs – can certainly be a strong motivator, but build-

ing other opportunities for ties to organically form can pay off. One company deliberately reduced the number of coffee machines by a factor of 20 to force interaction among different groups – and sales increased by 20%![13] While this approach may not work for faculty (we suspect there would be a sudden explosion in in-office coffeemakers), variations like shared coffee machines between departments or university-funded coffeeshop accounts could help relationships develop.

2. *Make big data competence universal.* Given the need for competence (or at least familiarity) with big data tools in virtually every research area and job occupation today, from astronomy to zoology, we suggest that every student and postdoc be trained early in their tenure at the university. For example, at Kansas State University, we have made use of remote workshops offered through XSEDE which last two days, require no prior programming experience, and introduce participants to tools like Hadoop/Spark and TensorFlow.[14] These have been popular across multiple colleges and departments, and we have suggested that they form a basis for requiring every student in the College of Engineering at least be exposed to these tools. Longer-term workshops (typically one week) through Data Carpentry can offer another quality option.[15]
3. *Build the community of experienced interdisciplinary researchers.* We recommend that institutions incentivize interdisciplinary research. At present, there represents an implicit penalty for an interdisciplinary grant – e.g., given that most departments are assessed based on research expenditures, a grant with a colleague in the same department is better (at least in the short term) than a grant with a colleague outside the department. Similarly, adding young researchers to a grant proposal may strategically grow the number of experienced interdisciplinary researchers, in the short term it may decrease the odds of an individual proposal's success. Institutions may find it advantageous to weight interdisciplinary achievements (papers, grants, and other artifacts) more heavily in evaluations for tenure or promotion; they may also want to set an expectation that larger interdisciplinary efforts will have a certain percentage (say, 10%) of faculty who are relatively new to these environments.

Conclusion

In light of our "Research Hierarchy of Needs", we suggest that both researchers and their institutions recognize both the difficulty and rewards in interdisciplinary research, and the need to adapt in the modern research era to the size, speed, and scale of data and its contribution to science. We need to convert the "molehills" to "mountains" of resources at every level of the hierarchy: hardware/storage/cyberinfrastructure should be well resourced with dedicated staff; human capital with training in big data should be ubiquitous across disciplines; and there should be an institutional commitment to replacing a culture of scarcity with planned, systemic infrastructure on par with traditional science environments like laboratories and other physical resources.

References

1 Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), *370-396*.

2 Lowndes, Julia S. Stewart, et al. "Our path to better science in less time using open data science tools." *Nature ecology & evolution* 1.6 (2017): 0160.

3 Neeman, H.. "A Case Study for HPC Workforce Development and Workforce Meta-Development." (2015).

4 Neeman, H., personal communication, June, 2018.

5 Apon, Amy W., et al. "Assessing the effect of high performance computing capabilities on academic research output." *Empirical Economics* 48.1 (2015): 283-312.

6 Kleinbaum, Adam M., Toby Stuart, and Michael Tushman. *Communication (and coordination?) in a modern, complex organization*. Boston, MA: Harvard Business School, 2008.

7 Kraut, Robert E., Carmen Egido, and Jolene Galegher. "Patterns of contact and communication in scientific research collaborations." *Intellectual teamwork*. Psychology Press, 2014. 163-186.

8 Clete A. Kushida, Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. (2012) "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies." *Medical care* vol. 50 Suppl: S82-101.

9 Arvind Narayanan and Vitaly Shmatikov. (2008) "Robust De-anonymization of Large Sparse Datasets." In *IEEE Symposium on Security and Privacy (S&P)*.

10 Dorothy Elizabeth Robling Denning. (1982) "Cryptography and Data Security." Addison-Wesley.

11 Frank D. McSherry. (2009) "Privacy integrated queries: An extensible platform for privacy-preserving data analysis." In ACM SIGMOD International Conference on Management of data (SIGMOD).

12 Blank, Steve, and Pete Newell. "What your innovation process should look like." *Harvard Business Review* (2017).

13 Waber, Ben, Jennifer Magnolfi, and Greg Lindsay. "Workspaces that move people." *Harvard business review* 92.10 (2014): 68-77.

14 Maiden, T., "XSEDE HPC Workshop: BIG DATA," https://psc.edu/hpc-workshop-series/big-data , accessed 9/9/19.

15 Teal, Tracy K., et al. "Data carpentry: workshops to increase data literacy for researchers." *International Journal of Digital Curation* 10.1 (2015): 135-143. https://datacarpentry.org/