Training for Cross-Disciplinary Research and Science as a Team Sport

Jennifer L. Clarke, PhD, Professor, Food Science and Technology, Statistics
Bob Wilhelm, PhD, Vice Chancellor for Research and Economic Development
University of Nebraska-Lincoln

This article is dedicated to the memory of David R. Swanson, Ph.D. (8/13/65 - 8/12/19), former Director of the Holland Computing Center at the University of Nebraska. David was a wonderful person and a great colleague. He will be missed.

As members of a land-grant highly research active university, we recognize the growing importance of data and computing across all disciplines. We are also aware that addressing most, if not all, societal problems will require knowledge from multiple disciplines. This brings to mind the recent work by Peter Watson in which he makes a compelling argument that many diverse scientific branches are converging on the same truths [1]. Hence training faculty members, postdoctoral scholars, and students to excel in cross-disciplinary environments and leverage advances in data and computing to further research goals is critical to future institutional success. Only by working together and leveraging intellectual resources can we make significant discoveries for the benefit of humankind.

There are multiple, high profile examples in science of large successful collaborative projects. These include The Cancer Genome Atlas [2], the Laser Interferometer Gravitational-Wave Observatory (*LIGO*) Scientific Collaboration [3], ELIXIR (an intergovernmental organization that brings together life science resources from across Europe)[4], and the French Conseil Européen pour la Recherche Nucléaire (CERN) [5]. All of these projects leverage human resources from multiple disciplines as well as the latest in data and computing resources. A primary reason for the ongoing societal impact of these projects is their willingness to share resources with the broader community.

This way of thinking - **scientific research as a team sport for communal benefit** - represents several challenges for most faculty researchers. Foremost, many faculty lack the knowledge and human resources required to plan for the use of their data and/or code outside of their own projects. This means that data and computational pipelines are generated to meet immediate or short-term needs, often associated with a specific project. Once the project is completed, data and code that could benefit other researchers (and even future projects within the same faculty group) languish.

The current guidelines for proper data sharing follow the **F.A.I.R. principles** - Findable, Accessible, Interoperable, and Reusable [4] (see Figure 1). There are existing web-accessible platforms on which data may be shared in a reproducible manner, e.g., Cyverse [6], the Dryad Digital Repository [7], and resources from the National Center for Biotechnology Information (NCBI) [8]. These are complemented by professional organizations that focus on research data management, e.g., the Research Data Alliance [9]. On our campus one of the key resources for information about F.A.I.R. is the **Uni-**

versity Libraries who have embraced the digital age of archiving [10]. Even with these resources, however, faculty time and effort are required to prepare either data or code for further scientific use. The National Institutes of Health and the National Science Foundation both encourage reproducibility and sharing through data sharing policies, yet it is difficult to secure financial support for long-term data storage and maintenance of data repositories. Most educational institutions view federal agency requirements to store and maintain data generated using public funds as an unfunded mandate. At this time the federal agencies have responded with some support via web repositories and databases. Processing and uploading data and associated code, however, remains largely unsupported. This indicates that a more sustainable model for cross-disciplinary data services and management is needed.

Associated with the challenges around proper data sharing and maintenance is computational toolkit development and maturity. As mentioned above, as with data, research groups often develop code for a specific purpose where the priority is one-time use. Once the individual responsible for the code (usually a student or postdoctoral scholar) moves to another project or leaves the university, the code is usually lost. Other research groups who could benefit from the toolkit and associated knowledge are forced to effectively start from scratch.

As with research data, there are web-accessible resources for hosting and sharing code that support efforts toward reproducibility [12,13]; see Figure 2. These include Github, the Science

Figure 1: F.A.I.R. principles (left column) and ways to support implementation [11]

Gateways Community Institute (SGCI) [14], and Cyverse (mentioned previously) where code can be linked with Jupyter notebooks and other tools for documentation and ease of reuse. Publishers are starting to take notice of the need to properly document and test code. For example, De Gruyter (publisher of more than 700 journals in the humanities, social sciences, and law), SPIE (the international society for optics and photonics), and BMC Bioinformatics allow authors to share working code associated with their publications through Code Ocean [15], a platform for code and data sharing to improve research reproducibility. The Nature Publishing Group insisted as a condition of publication in a Nature Research journal that authors make data, code, and associated protocols available in a timely fashion to readers without undue qualifications [16].

or recruitment of **application specialists.** These are individuals with advanced training (i.e., hold or are earning graduate degrees) in a discipline related to data science (e.g., computer science, engineering, physics, statistics, bioinformatics) who serve as catalysts for cross-disciplinary research. They span disciplinary boundaries and can manage multiple projects simultaneously. This is an extension to cross-disciplinary contexts of the basic concept behind the NSF project in Advanced Cyberinfrastructure Research and Education Facilitators [18], the Carpentries [10,19], and the SGCI. As can be seen in Figure 3, effectiveness in data science requires a taxonomy of skills and the idea is to match these with disciplinary knowledge and the ability to communicate within interdisciplinary environments. These individuals would reside within research core facilities or Centers

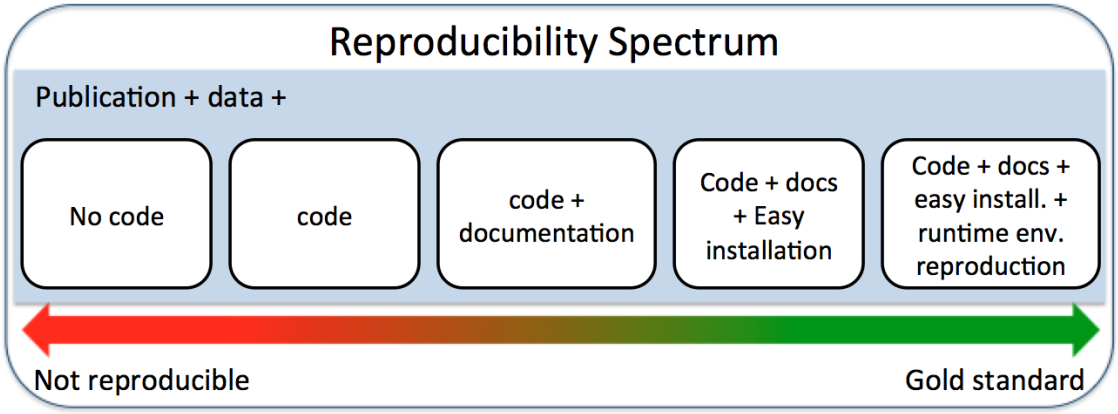

Figure 2. Reproducibility Spectrum for Research and Publication [17].

Hence faculty researchers are discovering that proper documentation and sharing of code are a requirement for publication in many highly respected venues. The challenge for institutions is how to support researchers who should or must meet this requirement.

A proposal on our campus that would provide this support is the development and be tasked with facilitating transdisciplinary research through knowledge of data and advanced cyberinfrastructure. Some are heroically technical while others emphasize the translation of scientific problems to computational solutions. These specialists often serve an `on-boarding' role for new staff/faculty/students to orient them to data and

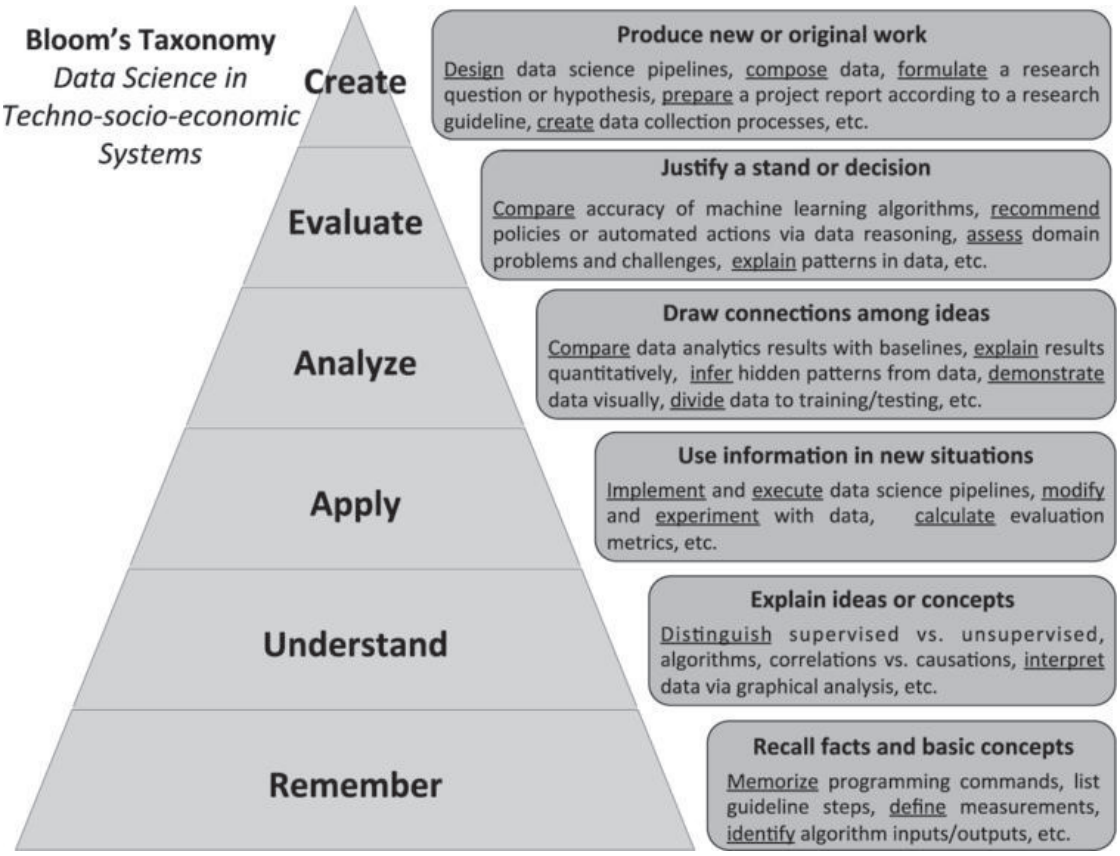

Figure 3. Bloom's Taxonomy applied to Data Science and Computing [20].

computing resources and current interdisciplinary projects. One of their most important roles is as repositories of institutional memory; in other words, they enable the use and reuse of data and code from university research projects. Even if they are graduate students, part of their job as application specialists is to document institutional projects and associated resources as part of building this memory. This leads to efficiencies in research that cannot be gained elsewhere.

It is important to note that even with the resources, tools, and willingness to meet the gold standard for reproducibility, there are additional challenges to conducting research in a transdisciplinary environment. We define **transdisciplinary research** as research that results in new knowledge formed via the integration of those domains that contribute to them; see Figure 4 [21]. Building an effective transdisciplinary team requires strong communication, not only of scientific concepts and ideas but also disciplinary expectations in terms of research output and recognition. The team must have a clearly defined goal and be able to articulate how each contributing discipline is expected to benefit. Faculty members who are willing (and able) to work in such environments need institutional support, as it takes considerable time and effort to build a team and realize the tangible benefits. Fortunately, federal funding agencies, in particular the NSF with their Big 10 Idea of **Growing Convergence Research** [22], are willing to support efforts in this direction. In effect, transdisciplinary research is a ``high risk, high reward'' endeavor. It is the role of the institution to mitigate the risk for fac-

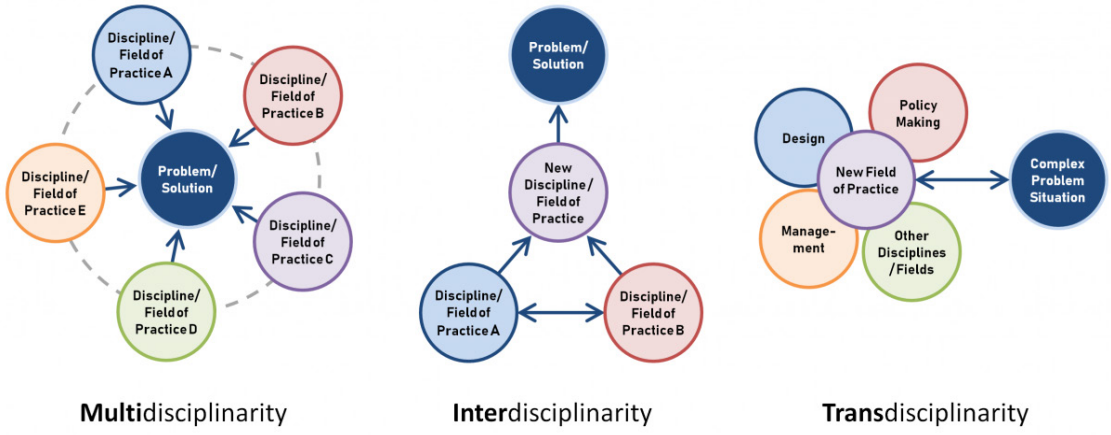

Figure 4. Multidisciplinary, interdisciplinary, and transdisciplinary approaches to research [21].

ulty members so that significant rewards can be realized.

On our campus we provide training opportunities for students and faculty to support acquisition of data and computing skills, reproducibility, and convergent research. These include data and software carpentry workshops, digital commons, and data archiving by the Libraries and the High Performance Computing Center. We also support an interdisciplinary **PhD program in Complex Biosystems** through the Office of Graduate Studies [23]. This program prepares doctoral students to conduct research that requires knowledge of both the data and life sciences. Each student is mentored by a pair of faculty advisors, one from the data and computing disciplines and one from the life sciences. Students in this program have earned predoctoral awards from several agencies including NIH, NSF, and the Foundation for Food and Agricultural Research (FFAR).

In summary, research has evolved into a team sport with members from multiple disciplines, working together toward a shared goal, enabled by continual advances in data science and cyberinfrastructure. As institutions of higher education, our role is to enable convergent research. We have articulated some avenues for support: reproducibility of data and code, University Libraries, application specialists, and strategic investments in transdisciplinary teams research. Convergence research is the future of science as solving some of society's largest challenges, from rural economic vitality to feeding our growing population, requires expertise in data, computing, and multiple scientific disciplines. The institutions represented at this year's Merrill Conference are well placed to play a leading role in the growth of convergence research to address societal challenges.

References

1. Watson, Peter (2017). Convergence: The Idea at the Heart of Science. Simon & Schuster.
2. Hutter, C. and Zenklusen, J. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 173(2):283-285. doi:10.1016/j.cell.2018.03.042
3. The Laser Interferometer Gravitational-Wave Observatory (*LIGO*) Scientific Collaboration. Web Page. https://dcc.ligo.org/LIGO-M980279/public Accessed 2019-09-01

4. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18
5. The French Conseil Européen pour la Recherche Nucléaire (CERN). Web page. https://home.cern/ Accessed 2019-09-01
6. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, et al. (2016). The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLOS Biology 14(1): e1002342. https://doi.org/10.1371/journal.pbio.1002342
7. The Dryad Digital Repository. Web Page. https://datadryad.org/ Accessed 2019-09-01
8. The National Center for Biotechnology Information (NCBI). Web Page https://www.ncbi.nlm.nih.gov/ Accessed 2019-09-01
9. Research Data Alliance (2019) "About RDA". Web Page. https://rd-alliance.org/about-rda Accessed 2019-09-01
10. Hart E, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, Poisot T, Woo K, Zimmerman N, Hollister J. (2016). Ten simple rules for digital data storage. PeerJ Preprints 4:e1448v2 doi:10.7287/peerj.preprints.1448v2
11. Australian National Data Service. Web Page. https://www.ands.org.au/working-with-data/fairdata/training Accessed 2019-09-01
12. Wilson G. et al. (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. doi: 10.1371/journal.pcbi.1005510
13. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, et al. (2014) Best Practices for Scientific Computing. PLOS Biology 12(1): e1001745. doi: 10.1371/journal.pbio.1001745
14. Science Gateways Community Institute (SGCI). Web Page. https://sciencegateways.org/home Accessed 2019-09-01
15. Code Ocean (2018). De Gruyter Partners with Code Ocean to Improve Research Reproducibility. Press Release of 2018-07-10. Web Page. https://codeocean.com/press-release/de-gruyter-partners-with-code-ocean-to-improve-research-reproducibility Accessed 2019-09-01
16. Nature Publishing Group (2019). Nature Research Editorial Policies. Web Page. https://www.nature.com/nature-research/editorial-policies Accessed 2019-09-01
17. Akalin, A. (2018). Scientific Data Analysis Pipelines and Reproducibility. Towards Data Science. Web Page. https://towardsdatascience.com/scientific-data-analysis-pipelines-and-reproducibility-75ff9df5b4c5 Accessed 2019-09-01
18. Advanced CyberInfrastructure - Research and Education Facilitators (ACI-REF) (2019). Web Page. https://aciref.org/ Accessed 2019-09-01
19. The Carpentries. Web page. https://carpentries.org/ Accessed 2019-09-01.
20. Pournaras, E (2017). Cross-disciplinary higher education of data science – beyond the computer science student. Data Science, vol. 1, no. 1-2, pp. 101-117. doi: 10.3233/DS-170005
21. McPhee, C., Bliemel, M., & van der Bijl-Brouwer, M. (2018). Editorial: Transdisciplinary Innovation. Technology Innovation Management Review, 8(8): 3-6. doi:10.22215/timreview/1173
22. The National Science Foundation (NSF) Big 10 Ideas (2018). Web Page. https://www.nsf.gov/news/special_reports/big_ideas/ Accessed 2019-09-01.
23. PhD Program in Complex Biosystems (2015). University of Nebraska-Lincoln. Web Page. https://bigdata.unl.edu/phd-program Accessed 2019-09-01.