A pilot "big data" education modular curriculum for engineering graduate education:

Development and implementation

Megan Sapp Nelson
Purdue University Libraries
Purdue University
West Lafayette, IN USA
Corresponding Author: msn@purdue.edu

Line Pouchard
Computational Science Initiative Directorate
Center for Data Driven Discovery
Brookhaven National Laboratory
Upton, NY USA

***Abstract*— Engineering higher education increasingly produces data in the volume, variety, velocity, and need for veracity such that the output of the research is considered "Big Data". While engineering faculty members do conceive of and direct the research producing this data, there may be gaps in faculty members' knowledge in training graduate and undergraduate research assistants in the management of Big Data. The project described herein details the development of a Big Data education module for a group of graduate researchers and undergraduate research assistants in Electrical and Computer Engineering. This project has the following objectives: to document and describe current data management practices; to identify gaps in knowledge that need to be addressed in order for research assistants to successfully manage Big Data; and to create curricular interventions to address these gaps. This paper details the motivation, relevant literature, research methodology, curricular intervention, and pilot presentation of the module. Results indicate that, generally, students involved in Big Data projects need comprehensive introduction to the topic, which will be most effective when contextualized to the work that they are performing in the research or classroom environment.**

Keywords—research data management; Big Data; Graduate education; Undergraduate education; curriculum development

I. Motivation

Projects collecting Big Data present many challenges for research data management (RDM), including organizing real time data from large numbers of data sources, managing data from many disparate sources (e.g., where each data stream may have its own data usage agreement) [1], understanding the training needs of a mix of graduate and undergraduate students in the research and programming team, and addressing the need for clear public as well as internal documentation.. This paper identifies support services that could be used to increase the efficiency and efficacy of research data management skills among Big Data(BD) research team members This article highlights a possible protocol for the development of BD education that is developed in situ with a research team. The research team involved collects and organizes visual data from the web.

Based upon initial conversations with this BD team, the authors developed an understanding of current RDM practices within the specific research group and proposed a set of initial learning objectives to address through educational interventions.

These learning objectives included:

"…Students and researchers will run cost analyses for their code prior to submission to the cloud in order to understand the financial ramifications of their code.

Students and researchers will develop a test bed subset of streams in order to test their analysis code for accuracy and efficiency prior to submission to the cloud.

Students and researchers will identify potential legal and ethical implications of their code in order to make their advisor aware of potential problems [2]."

The learning objectives were intended to create a starting place for response and to clarify the priorities for data management education for students and researchers.

With priorities identified that focused on the "Big Data" nature of the project, skills needed to manage and understand best practices for working with BD[3] were identified as a priority for education interventions.

II. Relevant Literature

Data information literacy, the ability of individuals to manage data appropriately and successfully has emerged in the literature as a necessary component of graduate education, and a crucial pre-cursor to the success of a research enterprise [4-6]. Simultaneously, the requirement by US federal granting agencies that principle investigators detail data management plans for all funding proposals has brought awareness of data management to faculty as well[7-13].

In response to the emergence of managing data as a crucial aspect of the research endeavor, academic libraries have collaborated with faculty to provide RDM support services. These services include developing data repositories; providing consultations on data management plans; assisting with the

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

development of metadata for research projects; and working with researchers to develop data management workflows [5, 14-18].

One component of this support that has grown significantly in the past six years is the development of data management curricula. These curricula span from online tutorials to semester long courses [5, 11, 19-22]. Generally, the curricula are built around the same basic competencies. These competencies include the basics of databases and data formats as well as data management and organization, which includes planning for documentation, sharing, and organizational planning for the management of collected data. The competencies also include the fundamentals of metadata and applications of metadata to RDM; and data curation and re-use, including identifying dissemination strategies for data sets [11].

Skills education needed to work with and manage Big Data is emerging as a subset of data information literacy, primarily as stand-alone courses, whether at the undergraduate or graduate level [23-25]. These courses primarily focus on the analysis of data using commonly available tools such as R or scripting languages such as Python. These courses also focus on effectively communicating about findings and reproducibility of research on BD. Few courses intermingle undergraduate and graduate students in the same curricula due to the advanced skills needed to work with BD and the storage technologies needed to handle data of that size.

III. Research Methodology

Information regarding current data practices of the team was gathered through reflective exercises and structured interviews. The reflective exercises focused on lessons that the group participants had already learned about managing BD.

Those reflections highlighted several issues that were a requirement in the curriculum. The first, the heterogeneity of the data being managed by the team, required that the students think critically about the variety of BD that was flowing through the data collection system. The second was the importance of the cloud to the research endeavor. Due to the large amount of data and the distributed sources of data, cloud storage technology wass an essential part of the project. The use of the cloud impacted the efficiency of the system, as well as the design strategies used to program the tools and work with the data. No curriculum could effectively teach programmers in this team if cloud computing were not an integral component.

After completing the reflection exercise, Professor Sapp Nelson and Dr. Pouchard interviewed the advisor and graduate students using a structured interview based upon the interview protocol described in [5] and [26]. The interviews were about one hour in length and focused on describing individuals' understanding of the data set[s] they were managing, the data management skills that they felt would be most important to successfully completing their research endeavors, and describing their current data management practices.

In total, four interviews were completed, transcribed and analyzed using the qualitative coding and analysis software, Nvivo. The transcripts were coded according to the data information literacy competencies found in [4]. The code book for the analytical project can be found at [27]. Two researchers coded the transcripts according to the node structure.

A. Findings From Transcript Analysis

All interviewees agreed that data documentation was the highest priority for data education and the skill that most needed development. This skill was present in all interviews completed. The graduate students identified sharing and reuse as the second most important competency for instruction. Interestingly, the team leader indicated that data management and organization was the second highest priority for the educational intervention. Data documentation was the fourth most important competency for the graduate students, following tools in third. Sharing and reuse is important for the students in the lab, as sharing with individuals in the research lab is a driving motivator for improving data management overall. Data management and organization presents a specific issue for the research advisor who must expend considerable effort introducing succeeding groups of students to an insufficiently organized data set (due to lack of training in previous students in the management and organization of the data set.

Tools for research data management were identified as an area where individuals would like further instruction by all interviewees. Students were looking for tools that will streamline the very messy nature of BD, whereas the advisors' perception was that many of the competencies were equally important for his students to master. After data documentation and data management, data processing and analysis was slightly more important to the advisor than the other nine competencies.

B. Gap Analysis

After analyzing the transcripts, analysis of themes that emerged across all interviewees and a gap analysis was performed. Generally, the students were focused on data practices that slowed down or in some other way harmed the research project that they were working on. This focus did not have a long term perspective on the management of data. In fact, the graduate students were very much focused on the immediate consequences of data management decisions. Identifying short cuts or tools that will make work more efficient came up repeatedly.

For the advisor, the skills necessary to support the management of the project for the long term was more highly valued. Short cuts were not a priority. Instead a RDM protocol that everyone in the research laboratory would be held accountable to, regardless of the sunk cost of getting everyone on board or fixing existing problems, was a high priority for the advisor.

C. Self Assessment

The remaining group of students who had yet to provide input were the undergraduate research assistants. These students were less likely to have completed a course on data management or even data analysis than their graduate student peers. In order to measure the baseline of research data

management experience. Professor Sapp Nelson and Dr.. Pouchard, along with Professor Nastasha Johnson, created a self-assessment tool that was intended to measure the skills set of the undergraduate students. The assessment requests that survey participants identify research data management activities that they are already integrating within their work flows. The survey instrument wording can be found at [27].

The undergraduate students were least likely to be familiar with and least likely to have integrated skills having to do with managing research data over the long term. Skills that encourage preservation or that make the data attributable for reuse were notably lacking in the survey response.

Top five learning objectives to be taught		
Competency	*Affirmative responses (have integrated competency in work flow)*	*Negative responses (have not integrated competency in workflow)*
Developing a preservation policy	0	17
Talking with a preservation professional about what is entailed in preserving a data set	1	16
Making data citeable, attributable	3	14
Developing a tagging schema to make documents findable for others	3	14
Creating a standardized file structure that makes clear where raw, analyzed, and final data files should be kept	3	14

Table 1. Top five competencies to focus on, as indicated by lack of integration within existing workflows (n=17)

Table 1. Top five competencies to focus on, as indicated by lack of integration within existing workflows (n=17)

Interestingly, the students believe that "the amount of documentation and description that your research team members provide [is] sufficient for you to be able to understand and make use of the data (12 affirmative responses, n=17). This is in direct contradiction to the perspective of the research advisor (who again, is the permanent member of the research team and the individual who is most likely to deal with the consequences of insufficient documentation as the team transitions between members periodically.)

D. Presentation of curricular objectives for feedback

Once all of the data collected from the individuals in the research group had been analyzed as described above, Professor Sapp Nelson and Dr.. Pouchard created a list of prioritized learning objectives. The proposed learning objectives can be found at [27]. The learning objectives (LOs) were collocated into three groups: LOs the students indicated that they already possess; LOs indicated for further instruction and training; and LOs indicated for new instruction and training. After review by the research advisor, a curricular intervention was planned that focused on the implications of data size and heterogeneity for the project.

IV. CURRICULAR INTERVENTION

Working with the advisor, a five hour, one day workshop was planned. The topic of the workshop focused on the implications of data size by exploring the 4 Vs of big data: volume; variety; velocity and veracity[28, 29]. Active learning was chosen as the mode of delivery for the educational content.

One activity was developed for each of the 4 Vs. These activities focused on integrating the Big Data conceptual information within a framework built on existing practices of this Big Data team. In order to do that, activities explicitly made reference to existing work flows and data management practices within the research laboratory, and drew upon student experience with their research.

A. Details of the Pilot Presentation

Resources for the pilot presentation were collected on a single website[31].The five hour workshop began with a self-assessment of students' perception of their own skills in data management. This self-assessment is available for download [27]. This self-assessment was used to cognitively prime the students for the activities that followed. The curriculum then moved into a series of modules focusing on the Four Vs of Big Data (i.e. variety, velocity, volume and veracity).

To teach variety, students were asked to consider the implications of data use agreements for the streaming video data that is analyzed by this team. Students completed a jigsaw activity in which they read either of a set of two brief policy documents or one long policy document. The students were then asked to critically consider what the policy contained, what the policy asked of the end user, what the policy forbade, and what the implications of the policy were.

The teams reported out on their discussions and then the class as a whole discussed the implications if the data tool they are developing must simultaneously manage video streams with three very different constraints.

Students then focused on considering impacts from a much larger array of potential sources of variety. The students were asked to complete a table that described the implications for storage, metadata, security, access, quality control and analytical methods. They had to consider a variety of constraints including multiple data types, security requirements access requirements and multiple other factors as well. The table made visual the complex array of constraints that each data feed may introduce into the design of the system. The students used critical thinking to consider what types of data could be included in their system.

To investigate velocity and the implications on the project, members of the project team were asked to lead a discussion/panel that considered how the rate of data accumulation impacts coding decisions, how bandwidth impacts research plans or code design and the practical implications of data accumulation for those on the team.

To think deeply about volume, the participants were asked to brainstorm (using post it notes) and contrast the implications of managing a single data stream, multiple data streams from a single source, and multiple data streams from multiple sources. The participants were asked to compare and contrast the implications of increasing volume on backups; finding/sharing data; naming files; documentation; the interface a programmer must use; and the hardware that must be available to be successful. Then the group as a whole discussed how the volume of data changes data management practices.

Veracity was the final module covered during the workshop. In order to add a component of veracity that was not hugely technical, a thought experiment was conducted wherein the impact of adding text annotation to the lab's data analysis tool was considered. The participants were asked to identify where the text annotations originated from, how reliable that data was, who the data authors were, who had permissions to add that data to their system, and what controls needed to be in place for that type of data.

The participants then drew a spectrum of messiness of data on the whiteboards. They were asked to place video and textual data on the spectrum of messiness and then explain their rationale. The discussion then turned to the decision points for when to clean messy data (if the data is cleaned at all?) The participants were asked to consider how the decision is made whether to clean data; what criteria are used to include or exclude data from analysis; what algorithms are preferred for cleaning data; and whether enough similarities exist between their current data set and textual data to transfer knowledge about data management from one type of data to the other.

The workshop then concluded with summative assessment regarding what the participants learned, what was useful, and what was relevant to their work.

V. Results

To collect the summative data, Mentimeter crowd polling software was used. The participants were asked "What activities in today's workshop were most relevant to your work?

Most relevant portion of curricular activities to participants' work			
Volume	*Variety*	*Velocity*	*Veracity*
10	5	5	6

Table 2. What part of today's activities were most relevant to your work? n=21

Table 2. What activities in today's workshop were most relevant to your work? n=21

Two participants indicated all parts were equally helpful. One participant declined to answer. Four individuals selected two modules as most relevant.

When asked about the practical implications of the speed of data accumulation, the students generally agreed that planning ahead for storage is the primary concern (8 responses, n=21)

When asked how adding volume adds complexity to managing data, participants responded with a variety of themes. These include: more testing is required for the algorithms; runtime for the code increases; the amount of metadata correspondingly increases; increased search time is needed to find specific data; manual checks will not work so algorithms need to be in place for quality control; and ways to insert, access, and delete data need to be thoroughly planned prior to code execution.

VI. Discussion

The participants generally responded well to the instruction. The primary feedback to improve the curricula was to provide materials in advance so that students could familiarize themselves with unfamiliar terms prior to the workshop. Students also requested more technical information and protocols about how to approach some of the problems of data management, as well as more hands on experience managing data.

This curriculum development project was largely successful due to the close coordination of the instruction materials to the research laboratory's workflows. A key to the success of the workshop was the mirroring of real world practice within the theoretical examples. Interviewing multiple members of a lab group is time intensive, but also provides far richer insight than a single conversation with a research advisor could provide. Given that the research advisor and the students had different priorities concerning skills to be taught and refreshed, and different perceptions of the longevity of their work, a single interview will not address the needs of the group. Instead, multiple interviews and reflective exercises help to build not only a fuller understanding of the research project and workflows but also of the instructional gaps present within the individuals comprising the research group.

The use of reflection exercises and interviews provided the curriculum designers with a wealth of insight into the needs and capabilities of the individuals who participated in the workshop. The curriculum developers were able to articulate prior to the beginning of the workshop what the likely levels of mastery would be among the participants prior to instruction. The curriculum developers then had the added advantage of being able to articulate an appropriate learning goal and be confident that the learning goal would meet the needs of everyone in the workshop.

VII. Limitations and Future Work

In order to transfer the curricula from one research laboratory to another, a significant re-write of the curricula would be needed. The method articulated in the protocol could transfer to developing training for other Big Data research groups, as could the learning objectives. However, the activities would require specific attention to bring them into alignment with a different research project or group's needs. Similar reflection exercises, interviews, and self-assessment would need to be conducted in order to tailor the curricula to a given new situation.

VIII. Conclusions

Data information literacy succeeds when it is embedded within the RDM enterprise that the participants are engaged in. The use of real world examples, experts from within the research enterprise, and specific context to introduce theoretical constructs provides a framework through which learners can articulate concepts of data management in a "real world" application. The use of information gathering tools such as interviews and written personal reflections gives the background necessary to create the highly integrated curricula.

ACKNOWLEDGMENT

This manuscript has been authored in part by employees of Brookhaven Science Associates, LLC under Contract No. DESC0012704 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

[1] L. C. Pouchard, M. S. Nelson, and L. Yung-Hsiang, "Data sharing and re-use policies for webcam video feeds from international sources," IASSIST Quarterly, vol. 39, pp. 14-23, Winter2015 2015.

[2] M. Sapp Nelson, "Learning objectives to [excised] lab," L. Pouchard, Ed., ed. West Lafayette, IN: Purdue University Libraries 2014, p. 1.

[3] A. De Mauro, M. Greco, M. Grimaldi, G. Giannakopoulos, D. P. Sakas, and D. Kyriaki-Manessi, "What is big data? A consensual definition and a review of key research topics," in AIP conference proceedings, 2015, pp. 97-104.

[4] J. Carlson, M. Fosmire, C. C. Miller, and M. S. Nelson, "Determining data information literacy needs: A study of students and research faculty," portal: Libraries and the Academy, vol. 11, pp. 629-657, 2011.

[5] Data Information Literacy: Librarians, data, and the education of a new generation of researchers. West Lafayette, IN: Purdue University Press, 2014.

[6] D. E. Keil, "Research data needs from academic libraries: The perspective of a faculty researcher," Journal of Library Administration, vol. 54, pp. 233-240, 2014.

[7] A. R. Diekema, A. Wesolek, and C. D. Walters, "The NSF/NIH effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories," Journal of Academic Librarianship, vol. 40, pp. 322-331, 2014.

[8] M. A. Haendel, N. A. Vasilevsky, and J. A. Wirz, "Dealing with data: A case study on information and data management literacy," PLoS biol, vol. 10, p. e1001339, 2012.

[9] J. M. Scaramozzino, M. L. Ramírez, and K. J. McGaughey, "A study of faculty data curation behaviors and attitudes at a teaching-centered university," College & Research Libraries, pp. crl-255, 2011.

[10] J. Mervis, "Agencies rally to tackle big data," Science, vol. 336, pp. 22-22, 2012.

[11] M. Sapp Nelson, "A pilot competency matrix for data management skills: A step toward the development of systematic data information literacy programs," Journal of eScience Librarianship, vol. 6, p. e1096, 2017 February 15 2017.

[12] K. G. Akers and J. Doty, "Disciplinary differences in faculty research data management practices and perspectives," International Journal of Digital Curation, vol. 8, pp. 5-26, 2013.

[13] E. Verbaan and A. M. Cox, "Occupational sub-cultures, jurisdictional struggle and Third Space: Theorising professional service responses to research data management," Journal of Academic Librarianship, vol. 40, pp. 211-219, 2014.

[14] S. M. Samuel, P. F. Grochowski, L. N. Lalwani, and J. Carlson, "Analyzing data management plans: Where librarians can make a difference," in ASEE Annual Conference and Exposition, Conference Proceedings, 2015.

[15] K. G. Akers, F. C. Sferdean, N. H. Nicholls, and J. A. Green, "Building support for research data management: Biographies of eight research universities," International Journal of Digital Curation, vol. 9, pp. 171-191, 2014.

[16] S. Pinfield, A. M. Cox, and J. Smith, "Research data management and libraries: relationships, activities, drivers and influences," PLoS One, vol. 9, p. e114734, 2014.

[17] L. Lyon, "Librarians in the lab: Toward radically re-engineering data curation services at the research coalface," New Review of Academic Librarianship, vol. 22, pp. 391-409, 2016.

[18] A. H. Poole, "How has your science data grown? Digital curation and the human factor: a critical literature review," Archival Science, vol. 15, pp. 101-139, 2015.

[19] J. Calzada Prado and Á. Marzal Miguel, "Incorporating data literacy into information literacy programs: Core competencies and contents," in Libri vol. 63, ed, 2013, p. 123.

[20] J. Qin and J. D'ignazio, "The central role of metadata in a science data literacy course," Journal of Library Metadata, vol. 10, pp. 188-204, 2010/08/31 2010.

[21] A. Wanner, "Data literacy instruction in academic libraries: best practices for librarians," See Also, vol. 1, 2015.

[22] J. Carlson and M. Stowell Bracke, "Planting the seeds for data literacy: lessons learned from a student-centered education program," International Journal of Digital Curation, vol. 10, pp. 95-110, 2015.

[23] J. Saltz and R. Heckman, "Big Data science education: A case study of a project-focused introductory course," Themes in Science and Technology Education, vol. 8, pp. 85-94, 2015.

[24] M. O'Neil, "As data proliferate, so do data-related graduate programs," The Chronicle of Higher Education, 2014 February 03 2014.

[25] N. J. Horton, B. S. Baumer, and H. Wickham, "Setting the stage for data science: integration of data management skills in introductory and second courses in statistics," arXiv preprint arXiv:1502.00318, 2015.

[26] J. Carlson, M. R. Sapp Nelson, M. S. Bracke, and S. J. Wright, "The Data Information Literacy Toolkit," 2015.

[27] M. R. Sapp Nelson and L. C. Pouchard, "A pilot "big data" education module curriculum for engineering graduate education: Development and implementation" (2017). *Libraries Faculty and Staff Scholarship and Research.* Paper 171. http://docs.lib.purdue.edu/lib_fsdocs/171

[28] F. X. Diebold, "On the origin(s) and development of the term "Big Data"," Penn Institute for Economic Research Philadelphia, PA2012 September 21 2012.

[29] L.C. Pouchard, "Revisiting the data life cycle with Big Data curation," International Journal of Digital Curation, vol 10(2), pp. 176-192, 2016.

[30] M. Sapp Nelson and L. Pouchard. (2016, March 22). Applied Big Data Workshop. Available: http://guides.lib.purdue.edu/bigdata

[31] M. Sapp Nelson and L. Pouchard. (2016, March 22). Applied Big Data Workshop: resources. Available: http://guides.lib.purdue.edu/bigdata/Resources