Running Head: DATA SCIENCE AND MSW CURRICULUM

Accepted for publication by the
Journal of Social Work Education

Data Science in the MSW Curriculum: Innovating Training in Statistics and Research Methods

Brian E. Perron[a*]

Bryan G. Victor[b]

Barbara S. Hiltz[a]

Joseph Ryan[a]

[a] University of Michigan, School of Social Work
[b] Indiana University, School of Social Work

[*] Brian E Perron, PhD, University of Michigan, School of Social Work, 1080 S. University Avenue, Ann Arbor, MI. 48103. beperron@umich.edu

Data Science in the MSW Curriculum: Innovating Training in Statistics and Research Methods

Abstract

Recent and rapid technological advances have given rise to an explosive growth of data, along with low-cost solutions for accessing, collecting, managing, and analyzing data. Despite the advances in technology and availability of data, social work organizations routinely encounter data-related problems that impact their opportunities for making data-driven decisions. While training in research methods and statistics are important for social work students, these courses often do not address the needs organizations face in collecting, managing, and using data for data-driven decision making. In this teaching note, we propose innovating social work curriculum using a data science framework as a way to address the day-to-day challenges organizations face regarding data. We provide a description of data science, along with four examples of MSW student projects that were based in a data science framework.

Recent and rapid technological advances have given rise to an explosive growth of data, along with low-cost solutions for accessing, collecting, managing, and analyzing data. These advances have led to innovative solutions to a variety of problems in the areas of law (e.g., (Hildebrandt, 2016), business (Chen, Chiang, & Storey, 2012), and medicine (Krumholz, 2014). The field of social work is attuned to the opportunities that these new technologies provide with the recent Grand Challenges put forth by the American Academy of Social Work and Social Welfare, with a key priority being *Harnessing Technology for Social Good* (Berzin, Singer, & Chan, 2015; Coulton, Goerge, Putnam-Hornstein, & de Haan, 2015). The associated policy recommendations focus on developing better data systems for communities, and unlocking government data to drive solutions to social problems (Berzin et al., 2016).

Despite the advances in technology and availability of data, organizations routinely encounter data-related problems that impact their opportunities for making data-driven decisions. The recent work of Bopp, Harmon, and Voida (2017) supports this supposition, showing that nonprofits and other mission-driven organizations face many challenges in designing information systems that can help inform decisions that are unique to their own needs and stakeholder groups. While training in research methods and statistics are important for social work students, these courses often do not address the needs organizations face in collecting, managing, and using data for data-driven decision making. For example, an adage in the data science literature is that 80% of the resources for data projects is devoted to accessing, managing and preparing data for analysis (Wickham, 2014). This encompasses a wide range of activities including (but not limited to) extracting or exporting data from different software environments, reshaping data to ensure that values are organized in rows and columns, screening and cleaning data files, converting file formats, merging data files, increasing data access or security, developing data documentation and codebooks, and implementing data quality improvement initiatives. These are essential activities for using data in a way to promote data-driven decision making but are often missing from the social work curriculum.

In this teaching note, we recognize a practical need for developing social work students' capacity to leverage existing data to solve problems common to social work settings. We propose innovating social work technical training using a data science framework, which is broadly defined as an area of study that draws on the fields of statistics, computer science, and information theory to manage and extract insights from massive collections of structured and unstructured data sources (Dhar, 2013). We first provide an overview of data science, with an emphasis on how the skills and knowledge are unique from what is traditionally taught in statistics, research methods, program evaluation, and policy analysis. Then, we provide four examples of data science projects completed by social work students as part of independent studies in data science. Finally, we offer suggestions for curricular development to support skill development for addressing complex data problems.

Overview of Data Science

The field of data science has emerged from a variety of technical fields due to the ongoing need for making data-driven decisions from data stores that, thanks to continual technological advancements, are rapidly increasing in size and complexity. Whereas traditional research methods and statistical training emphasize study design and hypothesis testing, data science focuses on dealing with practical data problems to promote actionable insights. This makes the field, and its tools, particularly useful for organizations operating outside of academia, where data-driven decision-making, and not hypothesis testing, is key.

Traditional research methods in social work are often built around the concept of evidence-based practices, with an emphasis on addressing service-related questions like, *What works for whom?* The concept and implementation of evidence-based practice (EBP) are not the same as data-driven decision making, for which data science is uniquely situated. The steps of EBP are built around informing practice decisions using the professional literature (McCracken & Marsh, 2008). However, whether service disparities or other problems exist within a given service organization requires local data collected by the organization itself. Social work organizations may possess important data that are critical for making data-driven decisions, but if the data are stored in an inaccessible manner, they are rendered unusable. Data science offers strategies and tools for extracting and making use of these data, which is not the focus of traditional research methods training.

Whereas research methods focus on the design of studies, data science emphasizes solving problems across the data lifecycle. The data lifecycle represents all the stages of data, from their creation to their final distribution and reuse (Goben & Raszewski, 2015; Pouchard, 2016). Data science gives careful attention to discovering, acquiring, processing, using, distributing, archiving, repurposing, and managing data. In turn, data science addresses the development and maintenance of systems to improve data quality, sustainable storage systems that are responsive to future technology and organizational changes, strategies that promote the use of data, and policies that promote the effective and ethical use of organizational data. Data science, therefore, can be regarded as an adjunct to traditional research methods training that is responsive to the rapid changes observed in the current technology environment.

The following student examples were selected to illustrate various types of problems that students have addressed by integrating their existing social work training with data science. These are intended to highlight the diversity of data-related problems in social work and the corresponding skills needed to address them.

Student Examples

Data Acquisition

Data acquisition involves extracting or obtaining data that already exist but are held in an environment that does not make the data immediately available for analysis. For example, social work agencies collect and manage large amounts of data, but much of the data are in an unstructured format, especially written reports and case notes. Other data may be provided to an organization, but the data are contained in an electronic document across a series of tables. In absence of having a core set of tools for dealing with these data, social workers have to invest considerable effort manually coding and constructing data files for analysis. As background to this example, a social work student participated in the development of a social enterprise that promoted activities and skill development for inner-city youth. The social enterprise involved teaching the youth woodworking skills for constructing various handicraft products that were sold on a major online store for handmade crafts. The profits from the sales directly supported the operations of the social enterprise, including educational and social supports for the youth, tools, and staff.

The specific decision-making problem encountered by this social enterprise was setting a price that would be competitive with similar products sold by other retailers. Setting a price too high could thwart the sales needed to be profitable. Ultimately, the social enterprise was interested in knowing the range of prices of the highest rated products across the different product categories. These data are readily available online, but manually extracting the data from the website proved to be a labor-intensive task that could not be supported by the enterprise's limited resources. As part of the project, the social work student learned how to write a series of computer scripts in the statistical programming language R (R Core Team, 2018) to *scrape* the data from the online store. This involves automating the process by having the computer copy and store the data in a format that can be analyzed using traditional software procedures. In this example, the data already existed, so this was a problem of data acquisition rather than primary data collection. These are not secondary data, because they were never designed specifically for research purposes and were not organized in a format for conducting analysis. Being aware of this distinction is important because we can see that social workers work with data that do not fit in the traditional primary and secondary data framework. Such skills for acquiring data are increasingly important for social work organizations. For example, an organization may need to acquire and summarize data from a governmental website, but the summary data are saved in many different tables across many webpages or are embedded in a large collection of electronic documents stored as PDFs. In these cases, manual extraction is very costly. Strategies and tools for dealing with these kinds of problems are central to data science and would serve social workers well.

Data Preparation

Data preparation involves a wide range of procedures that makes existing data usable. These include cleaning, converting, merging, and linking data files. Such data preparation problems are common in social work organizations, especially when data are saved across various platforms and software environments. For data to be useful, they need to be saved in a *tidy* format, following specific rules about how values are carefully organized into rows and columns (Wickham, 2014).

In this second example, a social work student worked with a social work organization to improve the efficiency of constructing monthly reports from data that were contained in separate systems, including a proprietary financial record keeping system, a proprietary donor database, and a freely available cloud-based spreadsheet. The organization's original approach involved constructing each report manually because of the complications involved with exporting data from each system, creating a common file structure, merging the data, and finally conducting an analysis.

To solve this problem, the student learned and applied various software procedures that automated the processes that had, to date, been done manually. This was an ideal opportunity for automation because the procedures for constructing the report were exactly the same during each iteration and relied only on updated data. The student developed an electronic *pipeline* using specialized software that performed each task for preparing the data files that could then be quickly converted into the necessary monthly report. These types of data preparation problems are very common, but social work organizations often lack the internal capacity for solving them in an efficient manner. Few such organizations possess the financial resources to outsource these problems to data consultants, and those that do build dependency rather than capacity.

Therefore, a data science approach can be regarded as a capacity-building strategy for social work organizations.

Data Usage

Many social work organizations collect and analyze various types of data and ultimately present that data via written reports. However, the presentation of the data is often "output friendly" rather than "reader friendly." That is, readers must integrate information across multiple data tables in order to extract insights that are relevant to their unique information needs. Ultimately, such static tables are not the most effective delivery mechanism for communicating data and helping stakeholders derive meaning. Recent technology advances allow increased opportunities for providing data that is tailored to the information needs of the end user.

The third student project focused on using interactive data visualizations to help a research group work more effectively and efficiently with financial data. More specifically, this student was completing a field placement at a university-based research center that received financial support from a variety of foundations, private donors, and government agencies. The diversity of the funding sources resulted in challenges for project planning and resource allocation. Thus, administrative meetings often involved ongoing reviews of budgets, reports, spreadsheets, and timelines. The process of achieving a holistic understanding of finances required a careful review of different types of records and documents. As an alternative approach, the student developed an interactive data visualization tool developed specifically to meet the financial information needs of the research group. The visualization did not require modifications to the existing financial data; instead, it was built on top of the data, which allowed the visualization to be automatically updated when the data were modified. The data visualization tool increased both the effectiveness and efficiency of decision-making because the student worked actively to understand the specific information needs of the research group. The unique contributions of data science in this example focus on the use and communication of data around the information needs of the end user.

Data Governance

Data governance refers to an organization's practices around its capability and capacity for managing data across the data lifecycle. As data-capturing technology advances, organizations are collecting more and more data. With more data comes the need to manage those data in a way that renders them useful for organizational decision making while preserving their integrity and adhering to privacy rules and regulations. While administrators may recognize their existing stores of data as organizational assets, many do not have a holistic view of the various data collected and managed by their organizations, the quality of the data, or the range of inferences that can be made. This type of problem relates to the need for training in how to manage data and set policies through data governance.

As a final example, a student worked with a social work organization to develop a governance plan. One of the most important parts of doing this involves taking a complete inventory of all existing data being captured and managed within the organization. At this particular organization, this inventory included data stored in electronic and hard copy form, as well as summaries of the data-generating mechanisms, file types, storage locations, how data were being used, who was using the data, and data quality ratings (e.g., completeness, accuracy, age). The data inventory became part of the organization's board-level strategic planning discussions. By consulting the data inventory during its planning discussions, the board recognized the need to increase investments in their approach to managing and using data, while considering potential direct and indirect returns on the investments.

Next Steps in Curricular Development

This teaching note provides examples of the knowledge and skills from a data science framework that social work students used to solve problems that are relevant to organizational functioning and decision-making. This framework is consistent with the Social Work Grand Challenges and also has direct relevance to the competencies of the Council on Social Work Education's 2015 Educational Policy and Accreditation Standards (2015). While many schools of social work focus exclusively on competency 4, *Engage in Practice-informed Research and Research-informed Practice*, data science applications would strengthen several other critical competencies, including assessment (competency 7), intervention (competency 8) and evaluation (competency 9).

Preparing social workers to effectively integrate data science skills will require modifications to the social work curriculum. We think there are two important steps to this process for building such capacity within schools of social work. The first step is to help faculty gain a better understanding of data science, emphasizing how this framework is distinct from traditional training in research methods and statistics. The second step is to build a more comprehensive understanding of the data-related needs that organizations encounter when trying to make data-driven decisions. This can help determine which skills are most essential for social work students, given that data science represents a vast array of skills that cannot be mastered by a single individual.

In our experience, integrating data science in the social work curriculum requires formal recognition and faculty support for the relevance and potential for leveraging data to address social problems. We consider data science a specialty area that would complement rather than replace any of the core curriculum. We think skill development in coding, either in R or Python is among the most important skills for accessing, managing, and maximizing the value of different types of data. This is in contrast to traditional statistical training, which focuses on the analysis of prepared data. We also think that training in presenting data, grounded in strong principles of data visualization (see Tufte, 2001) is critical for promoting the effective and ethical use of data.

Students who carried out these projects were motivated to develop skills in data science were influenced by campus-wide data science initiatives and the attention data science has received in the popular media. These projects were carried out through independent coursework and field placements. We see other strategies for offering students to opportunities for learning about data science. For example, specialty coursework in data science could be implemented within the curriculum through cross-listings with other departments. Data science skills can also be integrated into existing courses of research methods and statistics, although this approach may not give the students a clear framework for understanding and building additional skills in data science. Ultimately, integrating data science into the social work curriculum requires faculty with the relevant background and experience who can promote capacity building within a social work program. Integration of data science training at the doctoral level is therefore also encouraged to ensure the availability of faculty members with the ability to help students develop a robust set of data-related skills that are responsive to the current technological environment.

References

Berzin, S.C., Coulton, C. J., Goerge, R., Hitchcock, L., Putnam-Hornstein, E., Sage, M., & Singer, J. (2016). Policy recommendations for meeting the grand challenge to harness technology for social good. *Center for Social Development Research*. https://doi.org/10.7936/K7930SPJ

Berzin, S.C., Singer, J., & Chan, C. (2015, October*)*. Practice innovation through technology in the digital age: A grand challenge for social work- working paper No. 12. Retrieved from http://grandchallengesforsocialwork.org/uncategorized/practice-innovation-through-technology-in-the-digital-age-a-grand-challenge-for-social-work-working-paper-no-12/

Bopp, C., Harmon, E., & Voida, A. (2017). Disempowered by data: Nonprofits, social enterprises, and the consequences of data-driven work. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3608–3619. https://doi.org/10.1145/3025453.3025694

Chen, H., Chiang, R.H.L., & Storey, V.C. (2012). Business intelligence and analytics. *Management Information Systems, 36,* 1165-1188.

Coulton, C. J., Goerge, R., Putnam-Hornstein, E., & de Haan, B. (2015, July). Harnessing big data for social good: A grand challenge for social work- working paper No. 11. Retrieved from http://grandchallengesforsocialwork.org/publications/harnessing-big-data-for-social-good-a-grand-challenge-for-social-work-working-paper-no-11/

Council on Social Work Education (CSWE) - 2015 EPAS. (2015, July). Retrieved from https://www.cswe.org/Accreditation/Standards-and-Policies/2015-EPAS

Dhar, V. (2013). Data science and prediction. *Commun. ACM*, *56*, 64–73. https://doi.org/10.1145/2500499

Experian Data Quality. (2013, September). The state of data quality. Retrieved from https://www.experian.com/assets/decision-analytics/white-papers/the%20state%20of%20data%20quality.pdf?offer=at_eiwt100&br=exp&showDisclaimer=true&pc=sem_exp_adnet&cc=sem_exp_adnet_ad_070971408122109319147242468729933303175

Goben, A., & Raszewski, R. (2015). The data life cycle applied to our own data. *Journal of the Medical Library Association : JMLA*, *103*, 40–44. https://doi.org/10.3163/1536-5050.103.1.008

Hildebrandt, M. (2016). Law as information in the era of data-driven agency. *The Modern Law Review*, *79*, 1–30. https://doi.org/10.1111/1468-2230.12165

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking , training , and tools needed for a learning health system. *Health Affairs; Chevy Chase*, *33*, 1163–1170. http://dx.doi.org.proxy.lib.umich.edu/10.1377/hlthaff.2014.0053

McCracken, S. G., & Marsh, J. C. (2008). Practitioner expertise in evidence-based practice decision making. *Research on Social Work Practice*, *18*, 301–310. https://doi.org/10.1177/1049731507308143

Pouchard, L. (2016). Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation*, *10*, 176–192. https://doi.org/10.2218/ijdc.v10i2.342

R Core Team. (2018). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press USA.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*, 1–23.