Teaching with Data in the Social Sciences at Michigan State University

Scout Calvert, Data Librarian, calvert4@msu.edu
Andrew Lundeen, Digital Projects Librarian, alundeen@msu.edu
Shawn Nicholson, Associate Dean for Digital Initiatives, nicho147@msu.edu
Amanda Tickner, GIS Librarian, atickner@msu.edu

Introduction

Data, data analysis, and data literacy are growing in importance in society and industry. With increased attention to data matters, there are also enhanced expectations that higher education will prepare students to use and understand data in their lives and work. This report is part of a broader project by Ithaka S+R, mobilizing teams at 20 universities to investigate, through interviews and qualitative analysis, how quantitative data is being used in social sciences instruction for undergraduates. At Michigan State University, we interviewed 10 faculty members over fall and winter 2020-2021 from a range of departments that fit a broad definition of social sciences (for Ithaka S+R's list of in-scope departments/disciplines, see Appendix B), using an interview instrument provided by Ithaka S+R for all teams (Appendix A). All in all, eight unique departments, schools, and programs were represented across four MSU colleges (Appendix C). The purpose of this study is to gain insight into how data is being used in teaching and learning and to identify areas where support may be provided by MSU Libraries or through other campus partnerships.

Planning for this study was initiated before the beginning of the COVID-19 pandemic in early 2020. Midway through spring semester of that year, MSU closed its campus, and like many universities, asked faculty to find ways to continue instruction remotely, with online collaboration and meeting tools. Remote instruction was the norm through the fall semester of 2020, when we spoke with instructors. In our interviews, we invited respondents to reflect on how changes in their instructional practices for remote teaching did or did not influence how their courses used data.

The COVID-19 pandemic was a major driver of change in how faculty taught courses involving quantitative analysis of data. The "pivot to remote" necessitated a host of modifications in course design, impacting software selection and data collection methods. Our interviews

showed that faculty are resourceful, and this resourcefulness was on display in these adaptations.

The resilience and adaptability of instructors comes at the cost of time, however, as preparing data to support student learning was sometimes labor intensive, compounding steadily increasing class sizes and decreasing availability of graduate teaching assistants. Additionally, instructors worked to meet undergraduates at their level in terms of comfort and familiarity with mathematical and statistical concepts and software. This suggests areas for intervention and support by MSU Libraries and other partners in teaching and learning.

In what follows, we explore participant responses in the three main topic areas: getting data, working with data, and training and support. We then consider the reported impacts of the COVID-19 pandemic on these topic areas, as appropriate. Finally, we briefly consider the parameters of supportive interventions for resourceful faculty members and provide targeted recommendations.

Getting Data

The first hurdle that instructors must overcome in teaching with data is obtaining appropriate datasets for their classes. Our initial set of questions to instructors probed this, asking where the data used in their courses were found and how they were provided to students. Were datasets provided to students by instructors? Were students expected to collect or generate datasets themselves or search for pre-existing datasets?

In some courses, data were collected by students individually or as a class, but more typically, data were provided by instructors. Instructors located data from well-known sources such as the U.S. Census, American Community Survey, the Pew Research Center, and the World Bank, as well as targeted sources like the National Historical Geographic Information System (NHGIS) or the Global Terrorism Database (GTD). A common reported data source was the Decennial U.S. Census, either from its source via the census.gov website or via library databases that provide historical census data. Subject matter repositories (for example, tDAR for archeological data) and open government data portals (such as data.gov) were other places that instructors directed students to and used themselves to find data to provide to students. Other reported resources included "*data clearing houses*," "*the library*," "*the US spatial data*," and "*USGS Data Explorer [and] National Map*" (See Appendix D for a complete list). For most instructors, the exact process of finding data was not as relevant to learning goals for students as the datasets themselves: "*I usually have a couple of links in my class materials to*

clearinghouses, data clearing houses where they can find things. Sometimes we talk in class or office hours about a student's specific interests and we try to locate something. And I have sent students to the library."

When instructors provided data to students, the data often required cleaning and preparation for use in class. Some instructors wanted to provide it in a form that students could easily use for specific exercise analysis. One instructor described their data preparation process as follows: "*I basically cleaned the data, created scales for them, got the data in pretty good shape. For some of the variables that were really highly skewed I did things like median splits on them so that they had a relatively clean data set.*" Data cleaning activities were sometimes labor intensive, but the need to not overwhelm students' skill levels and comfort with statistical analysis justified the effort in the view of instructors. Sometimes, datasets were provided because the specific types of data would be too difficult for students to find or access themselves, for example large transaction datasets. In that case, the instructor also had to provide instruction on using the data in its computational environment (MSU's high performance computing center or HPCC), and students had to sign a data use agreement.

When instructors did require students to find data, they often pointed students to specific databases, including some that were provided by the MSU Libraries. In such cases, instructors often offered guidance to help navigate the database portals, which ran the gamut of institutional, disciplinary, or commercial repositories, on classroom websites (in the case of instructor provided data), library-provided databases, or government websites. As one instructor explained: "*If I don't kind of start pointing them in the right direction and give them a couple examples – from that standpoint, I think they would never get out of first gear. What I find interesting, though, is as soon as you start pointing them to some repositories of data, I think they start searching out and start branching out on their own.*" There was a reported benefit, though, when students were assigned to obtain data themselves: "*they get an area that they're more interested in. The problem with that is that often good data are hard to come by.*" One instructor observed that students sometimes struggle to find appropriate data because of their lack of experience, but often it is because the appropriate data are not well documented. Left to their own devices, students do use internet searches to find data as well, which may be a source of frustration, as one instructor described: "*I think more often than not they're hoping that if they put in the right Google search term, it just comes up, and if it doesn't, then all of a sudden it doesn't exist. So I think more often than not... their search methods are informal.*"

Social science instructors rely on a broad range of data types, described topically, by file type, or by the tool used to process the data. For example, these can include spatial data,

historical data, sensor data, survey data, digital trace data, or transaction data on the one hand, and R libraries, .csv files, or spreadsheet data on the other hand. Most frequently, the data used and provided by instructors was tabular data in Excel or .csv format. In some disciplines, spatial and GIS (Geographic Information System) data were used as part of undergraduate instruction. Data and file type were interrelated with needs for software that allowed disciplinary-relevant kinds of analyses. For most instructors, the size of files and a place to store them was not a concern; however, at least one instructor supplied data for students with size, storage, and computational concerns: "*I'm requiring them to use these cell phone data and this grocery store data that I have and there's terabytes of data there.*"

Working with Data

Once appropriate datasets were found and/or made available, the ways in which instructors integrated those data into their classes were as varied as the sources of data, depending on discipline, course level, instructor background, and the focus of any individual class. Instructors used data to teach a number of different approaches or analyses, including machine learning, data visualization, descriptive statistics, basic correlation, t-tests, bivariate analysis, models, logistic progression, and using high performance computing to conduct statistical analysis. These varied approaches or analyses influenced the choice of software used in class, although the ultimate driving goal behind most data instruction seemed to be providing students with a general sense of data literacy or numeracy to prepare them for future coursework and careers. Instructors also identified several issues or challenges involved in teaching with data in the social sciences, including a wide range of ethical considerations, increasing class sizes, a "digital divide" between students, and the use of "big data" in the classroom.

Software Tools

Instructors identified a range of specific software applications and technical tools that their students use to manipulate, analyze, or interpret data in their undergraduate courses. These varied by both course level and discipline, and included statistical analysis programs such as STATA and SPSS, geospatial data tools such as ArcGIS and QGIS, and programming languages and related software packages such as R, Python, or JavaScript (see Appendix D for comprehensive list).

Ease of use and ease of access were often cited as reasons for software selection. As one instructor put it, "*I think the software packages I have used have been good because they're easy, they're user-friendly. Yes, students can kind of get a feel for how to run these tests without having to also do a lot of learning like writing syntax or code.*" The ubiquity or availability of software applications was also cited as a reason for selection. For example, Microsoft Excel was one of the most frequently mentioned pieces of software used in the classroom setting. One instructor remarked, "*the main reason why I stick with Excel and don't go beyond into statistical software packages is because I don't know which statistical software package your organization is going to have, but everyone is going to have Excel. So I try and keep it basic, and they try and give them that overview.*" Instructors also expressed an appreciation for free and/or open source software, with one remarking, "*It's free, right? That's the 'F' part of the OSS or FOSS [Free and Open-Source Software]. So they're aware of it and so I am seeing a greater understanding of open source software.*"

Several instructors made a clear distinction between teaching the tools and teaching the fundamental methods or concepts behind the tools, and emphasized that both are important: "*So we very strictly differentiate between teaching the methodological and statistical concepts on them, from the technical, practical application, through a software platform in the labs. So we do both.*" Another instructor, when asked "To what extent are the tools or software students use to work with data pedagogically important," replied, "*I would say they're vitally important. So for example, part of my goal is really never to just teach them the button pushing aspects of using software. My goal is to train them to be critical computational thinkers*." One instructor indicated that the software they used in the classroom was merely a means to an end. In response to the question "Is the software you choose really advancing the core purpose of your course, or is the software incidental to what you're actually trying to accomplish," they said, "*I want them to be able to apply the statistical tests and to understand those statistical tests and what they mean*," and that "*almost any statistical software package would do what it needs to do.*"

Data Literacy

Data literacy is a multifaceted set of skills and proficiencies that can include both knowledge of specific tools and methodologies and a general conceptual understanding of how data can be collected, evaluated, and used. Several aspects of data literacy were mentioned by instructors as critical driving factors behind much of their data-based instruction in the social sciences. For a sense of the various skills and competencies that make up this type of literacy from the teacher and learner perspectives, see *Figure 1* below.

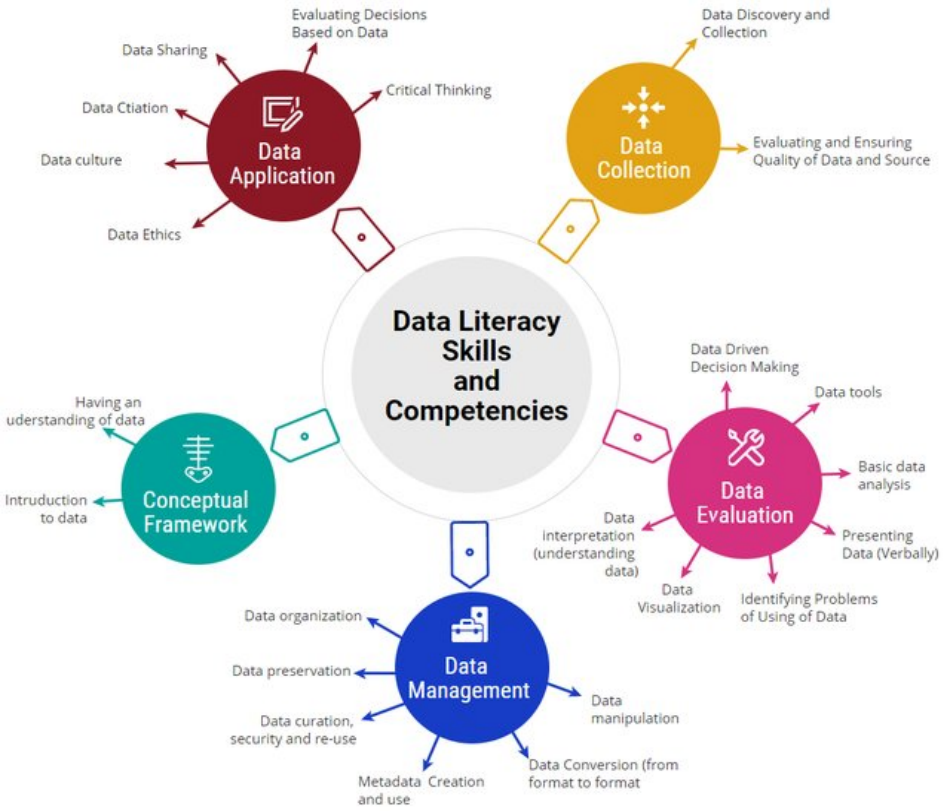

Figure 1. Schematic representation of data literacy skills and competencies (Guler, 2019)

For upper-level courses with prerequisites, instructors expect students to have some baseline level of experience upon which they can build. However, outside of these specific requirements, most instructors noted that they expect students to have very little prior knowledge of either the data tools or theoretical constructs necessary for the work. Several instructors mentioned that while they used to assume students had more of this foundational knowledge, they have come to realize over time that students are less prepared than they would have liked. Many instructors note that they have to spend at least some time teaching the basics, as their class might be the first time some students have had to work with data at all.

Some instructors cited students' apparent "fear" of numbers or statistics as an important factor in course design. Others used less charged language, describing students' lack of basic familiarity with statistical or numerical concepts, or their uncertainty about their abilities to work with numbers, data, and statistics. One instructor mentioned that some of their students lack an understanding that the object of study is something that can be amenable to quantitative analysis: "*I feel like half of that class... the data analysis class students come in knowing that they're going to be doing mathy things, but socio-linguistics students don't come in thinking that*

they're going to do math and it's basically like half the course design is like, that's cool that you like language. Surprise! Now you have to also like math." This influenced how instructors prepared students for analytic work or couched instruction to account for student anxiety or avoidance: "*So even students rarely have the ability to critically look at data*" and "*Not everybody is comfortable with numbers. Some of them just start seeing a big spreadsheet and start freaking out.*" This could make a difference to disciplinary understanding as well. As one instructor put it, "*Sometimes it's very intensively, statistically, data analytical, and sometimes it's not quite so intensively data analytical, but data analysis is at the core of what we do. And often, data analysis is actually… that process is framed as, is theoretical in archeology.*"

Data literacy instruction is impacted by several other factors, including increasing class sizes. As one instructor lamented, "*I mean, I cannot ask a hundred students to write a data-rich research paper.*" A few also mentioned that some students' lack of access to fast, reliable internet connections at home can be an issue, especially when it comes to remote learning during the COVID-19 pandemic. This digital divide between students also impacts the amount of prior experience students have coming into the classroom, and their general level of data literacy. However, one instructor compared their experience at MSU to their prior institution, and indicated that this disparity is much less of an issue at MSU: "*MSU students compared to my old students are much more – they're better at problem solving than [previous institution] students who have a big problem with the Internet. There's a big digital divide... Data literacy and just the ability to use the Library's resources was nonexistent compared to MSU's.*"

Instructors are sensitive to using data and approaches that resonate with real world phenomena and career goals. Some recognize that the students will use what they learn to answer research questions that interest them in future coursework, while others suggested an outcome that students leave with skills to critically analyze material and evaluate evidence. One instructor said, "*it's important for students to understand the limits and the restrictions associated with data analysis. There's some questions we're not good at answering.*" Some instructors replaced analytical methods with data visualizations as a means of engaging students with data: "*So a good example of that might be they have a data visualization exercise where, you know, it's pretty basic from the standpoint of ... tak[ing] some simple Excel data that might be in a 2x2 format and just create some data visuals to get them to understand that the data will tell us the story.*"

The ability to work with data is viewed as a crucial job skill, and most instructors considered how students would use what they learned in the course later, either in graduate school or in a job after the bachelor's degree, viewing data competence as "*absolutely critical*"

for student career success and financial outcomes. One instructor explained, "*An awful lot of what those courses are about is professional development and being at a place where they can land a position because they understand data well.*" This could also motivate students: "*connecting data analysis skills to career development, professional development… I think that's an easy way to get the students involved.*" In addition to general facility with data, some curricular choices were made or continued because of the potential benefit to student job prospects, as when one professor explained that a student's ability to do advanced data analysis in Excel impressed a potential employer: the student "*brought up in a job interview that [they] can do this stuff in Excel. And it stopped the interview. And they all started to ask [the student] how to do it. Yeah. So it has changed my perspective [on the value of Excel skills] a little bit.*"

Another aspect of data literacy that some instructors commented on was the need for good data management practices, even if they did not include it explicitly in their course content. There was sensitivity to documenting how data will be collected, structured, analyzed, stored, and shared. As one instructor explained, "*I do not teach any data work, however you want to characterize that, without management. And again, management expresses itself differently given the data.*" For instructors who brought it up, data sharing involved ethical concerns as well as standardization and documentation concerns. As one instructor said, "*And then thinking about, okay, you have produced a dataset, you've generated an original dataset. How do you share that? How do you give that? How do you do that ethically?*"

Data and Research Ethics

Most instructors noted that there were significant ethical considerations involved in teaching with data in their disciplines. When asked whether instructors in their fields faced any ethical challenges in working with data, one replied, "*I cannot teach any data related stuff – and in this case, I'm using the term data very inclusively and very broadly – without giving equal and equitable standing to the ethical implications of all of that.*" However, not everyone identified ethical challenges in their data-centric instruction. In response to the question "Do instructors in your field face any ethical challenges working with data," one instructor said, "*Not in the classroom setting, no.*" Another explained that while they did incorporate ethics into their instruction, these considerations were not unique to teaching with data: "*No, I don't think so. Ethics is a big thing in any field and it is in planning as well… but it doesn't deal with data… I think the same thing would go for anybody who teaches anything to do with numbers and data – you know, giving credit to where you actually find the data, make sure it's a credible source and*

giving credit. So that's not only with data, it's for anything that you do in academics, at least. The main thing that you teach them is don't plagiarize… I think it's common in any field, not just teaching with data." Still, several specific ethical considerations were mentioned.

Several instructors raised issues related to privacy in data collection. One instructor, who said that they use data from student surveys in class, mentioned the need to clean these datasets to remove personally identifying information: "*I do try with our surveys since the students are kind of analyzing data from their own class, I do remove all of the demographic variables... So, I don't give them individualized demographic data because that could potentially identify individual students in the class. So I have thought about ethics of identification and those kinds of things.*" Another instructor expressed a concern about privacy in the use of surveillance data, saying, "*I'm also a little bit concerned too about surveillance data… I have my students often read an article, but I often see, it's often about surveillance data. This idea that students are completely surrounded and completely enmeshed and enveloped by data, just because they have a smartphone in their pocket. And that data is being used in ways that they don't know about and that they have no control over.*"

Diversity, equity, and inclusion in data collection were also mentioned as important ethical considerations. As one instructor put it, "*And then one of the questions to look at is who's in the sample. How is the sample representative? What is missing and how do the conclusions represent the findings? And are they overstating a case, or who do they think this applies to? Because I think those are questions that we don't ask very often. And it's still a very, very contentious issue within science and data collection, where there isn't enough representation of under-represented groups, minority groups. So I think that's one ethical issue that I think most of us need to look more into and emphasize more, to be more critical about... diversity, equity, and inclusion.*" In addition to who is represented in datasets, how those subjects are represented was also mentioned as an ethical concern. One instructor said that they try to avoid using data that could be taken out of context or used to paint subjects unfairly: "*It is really important for me that I never used data that could frame participants in any negative light.*"

The use of data from indigenous populations was also identified as an area rife with ethical concerns, especially in certain disciplines. One instructor raised concerns about data sovereignty in the field of archeology, for example:

> "*Archeology isn't unique by any stretch of the imagination for the ethical overburden of a lot of what we do... Living peoples, indigenous peoples, descent communities... the legacy of colonialization, the need to decolonialize... So that ethical imperative is woven into everything we do. Everything we do and*

> *everything that we teach, whether it is specifically data related or sort of data adjacent... But also data sovereignty in indigenous communities and descent communities, right? So yes, it is part and parcel of everything that we do. I mean, not all of us do it, and archeology still... that the legacy of colonialism and lack of ethics is really close to the surface."*

Another instructor mentioned the importance of complying with the Native American Graves Protection and Repatriation Act (NAGPRA), saying, "*I talk about NAGPRA in almost every class I teach, and I talk about issues of Repatriation. I've done NAGPRA compliance work in my job and I talk about those issues and we don't work with that kind of data for a good reason. So yes, we talk about ethics of data and how we don't use certain data because it requires a level of project planning that can't happen in just a semester.*"

A few instructors mentioned using an ethics module or delivering a targeted lecture on the subject of ethical challenges in their field. Several indicated that data used in their classes were collected following approval from MSU's Institutional Review Board (IRB), and that the data was cleaned (e.g. removing demographic variables) to mitigate ethical issues, while others said that they make use of IRB training materials in order to incorporate ethical considerations into their instruction. Some instructors indicated that there has been an increased focus on ethical issues in their fields in recent years, with one saying, "*when I started teaching this, I relegated ethics to the last class of the semester. Not anymore. There's a full module in there – the ethics of research.*"

Big Data

Working and teaching with "big data" was another factor brought up by several instructors. While "big data" is more strictly understood to mean "extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis" (NIST, 2015, p. 5), we did not press instructors to explicitly define the term in their responses. However, they identified several challenges related to big data and its changing stature in teaching, research, and society that bear reporting: the availability of large datasets for research and teaching, the required hardware and software infrastructure to work with big data, and ethical considerations related to methods for using big data.

Three instructors recognized the need for students to understand the computational and software requirements of big data because some sectors of industry are moving in the direction of having bigger, more heterogeneous datasets. As one instructor put it, "*You always will need*

to know the fundamental statistical approaches but the hands-on application is moving at light speed to the era of big data."

One instructor described the challenges for students in working with especially large datasets, in this case terabytes of transaction data, noting that "*many of them don't have the skill sets to not only analyze the data, you know, the statistics tools, but also just in terms of handling the data [computationally], the big data that I deal with.*" Utilizing big data in the classroom involves additional preparation to teach students to navigate the required computational environment, which entails some tradeoffs in what can be taught in any given semester: "*to be honest with you, what I'd like to do is have them use Python, because the private sector, it's better for manipulating and handling big data.*"

Another instructor expressed concern that relatively few social scientists are involved with high performance computing at MSU, and that this was of broader concern for the discipline and data instruction within it: "*they're just not looking at problems that are that big. But I think also part of it is because it's hard to conceive of how you'd even start on a problem that big. So yeah, I think there could be a role for greater sort of institutional advice, training, consulting, pointers on where to look. That would help across the board with modernization of courses that involve data.*"

The last aspect of "big data" that was brought up was the issue of surveillance data, and the concern that "*data collection has become so ubiquitous and so universal that people are allowing data about their personal day-to-day activities to be captured and used*." One instructor said, "*I caution my students more with every passing semester—maybe caution is too strong a word, but I bring it up—you guys know how much data is being collected about you on a day to day basis? Well, new sources of data, global positioning data and micro-targeting advertising data.*" This suggests that including big data topics like machine learning and algorithmic literacy in data literacy instruction could counteract a potential "*blind allegiance to [the notion that] all data is good data, all data is usable data.*"

Training and Support

Support for Students

Although there are many challenges involved in teaching data-intensive undergraduate courses, faculty do not have to go it alone. When asked about additional training or support their students receive in obtaining or working with data, respondents identified two major sources of

this instruction: teaching assistants (TAs) and campus units/partners outside the classroom such as the MSU Libraries.

Several instructors mentioned that teaching assistants play an important role in many of their courses, whether teaching face-to-face or in a fully online environment. TAs were identified as providing students first-line guidance and support, while also learning themselves from the primary instructor or other TAs. Teaching assistants often led the lab portion of courses by teaching specific technology tools. Instructors addressed the importance of TAs gaining familiarity and comfort with these tools and concepts, with one noting, "*TAs must attend every class; we meet twice a week. We touched base on a very, very regular basis,*" and suggesting that instructors should create a structure whereby "*having a senior and a junior teaching assistant that overlap so they can teach each other the techniques*" can set pedagogical expectations and establish continuity. However, one instructor expressed a concern about the ability of teaching assistants to reinforce content and provide students with similar opportunities to apply concepts and practice skills, noting, "*I usually have a TA for the course and I always worry about the answer they give relative to what I give… I'm kind of a control freak on that stuff.*"

Places on campus that instructors described as useful resources include the MSU Libraries,[1] particularly the Libraries' Digital Scholarship Lab,[2] MSU's Institute for Cyber-Enabled Research (ICER),[3] the Lab for the Education and Advancement in Digital Research (LEADR),[4] the Digital Humanities program at MSU (DH@MSU),[5] the Writing Center,[6] and MSU's Information Technology Services (ITS).[7] Some instructors were also aware of the Center for Statistical Training and Consulting (CSTAT),[8] but there was a lack of clarity as to whether it was available as a resource for undergraduate students. The Libraries were lauded as being useful for providing workshops and information sessions specific to classes, expertise via librarians, help with data collection, access to data and databases for use in the classroom, and for having powerful computers available to students via the lab in the Digital Scholarship Lab. The HPCC at ICER was mentioned as being helpful for a specific class that was working with very large datasets. One instructor mentioned that ITS helped provide access to remote desktops for remote classes that needed computer lab access. DH@MSU was mentioned as providing

[1] https://lib.msu.edu/
[2] https://lib.msu.edu/dslab/
[3] https://icer.msu.edu/
[4] http://leadr.msu.edu/
[5] https://digitalhumanities.msu.edu/
[6] https://writing.msu.edu/
[7] https://tech.msu.edu/
[8] https://www.cstat.msu.edu/

useful workshops, and LEADR was mentioned as providing support for "*small digital class projects.*"

When asked about other ways in which students might be learning to work with data outside their formal coursework, some instructors indicated that they point students to online tutorials and other extracurricular resources that they can pursue on their own time, usually to support instruction in specific software or tools. On the subject of peer instruction/study groups, however, most instructors seemed to believe that their undergraduate students were not participating much in these activities, unless it was explicitly a part of the lesson plan. For example, when asked about extracurricular tutorials or peer learning, one instructor said, "*Certainly some of that stuff happens within the context of the formal learning in the class. So I'll say, here's a tutorial. Do it, right? It's happening outside, but it is part of the curriculum of the class… usually the outside resources, tutorials, videos on YouTube, whatever, they're doing it on their own, but it's within the framework of the pedagogy itself.*"

Training for Faculty

When asked about training that they themselves have received in teaching with data, instructors overwhelmingly reported both a lack of and a desire for more professional development and ongoing training in this area. Said one, "*I think a big challenge is that most people who do teach technical things learn most of what they learn in graduate school. And so the longer they're out of school, the more critical this gets.*" While some instructors do engage in informal training, they still cite a need for more formal training in teaching with data. However, there are perceived barriers to such training, such as a lack of free time to enroll in structured, scheduled courses: "*For my GIS, when Esri puts out MOOCs [massive open online courses] and things like that, I take them up just to see what these new things are. So I do those, but no formal training. Because I just don't have the time.*" Another barrier mentioned is a lack of institutional or research benefits for ongoing professional training: "*The issue is, what skills am I going to develop… to do my research?"* Some instructors reported being unsure what training opportunities were available for teaching with data: "*If there was maybe some training that I would use, I would take it. I've taken lots of training on how to use different technologies to gather data. Tons of training on my own volition, but not about how to teach with [data]. I don't think I've ever taken anything with how to teach data and maybe there's stuff out there I don't know about, but yeah, it's simply because I don't know of anything that I should be taking to learn.*"

One instructor did express a desire for venues where ideas about teaching could be shared, saying, "*I think for me just having access to and hearing how other people do teach with data to get new and fresh ideas for things that they could do with the students to get them engaged. I think that's always helpful,*" and another mentioned Facebook groups and conference networking as being useful. Places where instructors could share ideas for teaching in addition to technical skills were mentioned as opportunities for training and sharing: "*I've become a member of other kinds of groups, I guess, like Facebook groups from other… instructors, that we can kind of exchange ideas. So we kind of have our own community – it's all nationwide. And so I also picked up some there. But it's not really formal, you know, a formal learning process, rather informal.*" Many instructors also reported participating in workshops of various sorts, but again cited a lack of formal, institutionally recognized training or accreditation: "*So I've done some workshops on data against some software carpentry training that was focused on training trainers, which was actually pretty useful. And I've done any number of workshops at conferences, since I got a lot of conferences that are sort of data oriented, right before the pandemic. But in terms of like formal training, teaching or certificate kind of thing now…*"

One avenue for informal peer collaboration that was mentioned is the sharing of course syllabi or other instructional materials between instructors. Instructors consistently expressed a pronounced willingness to both share and borrow instructional materials, though in practice most indicated that they have never been asked explicitly to share their materials. When asked "Do you use any data sets, assignment plans, syllabi, or other instructional resources that you receive from others, and do you make your own resources available to others," one instructor succinctly replied, "*Yes and yes.*" Another elaborated: "*I would be happy to share assignments or lecture slides or those kinds of things. I'm pretty open and believe in sharing pedagogical materials. Nobody's asked me specifically for materials about this particular course, but actually, one of the instructors for this course did ask me if I had done some stuff on open science and I shared that I had done little bit.*"

The types of instructional materials most often borrowed include syllabi, specific class assignments, and tutorials. This type of material borrowing is typically for pragmatic reasons, as one instructor stated: "*It's usually a framework for them explaining the material better than I currently do. And just using that and modifying my slides or stealing theirs,*" and another explained, "*Yes, yes. I have benefited from that, from those interactions in terms of resources, and there's some interesting projects or ways to present material, like supporting videos, for instance.*"

Teaching with Data During a Global Pandemic

In our interviews, we invited instructors to share their experiences teaching with data, whether before the COVID-19 pandemic or during it, and to discuss the pandemic's impact on their teaching practices as appropriate. Responses show that the effects of the pandemic were wide-reaching. Instructors reported that the pandemic and the shift to remote learning completely changed the way that they taught with data in both spring of 2020 and the 2020-21 school year.

Instructors employed a range of strategies to cope with the cascade of impacts from campus closures, including (perhaps most critically) the closure of computer labs. Instructors' immediate focus was on simply making it through the semester "*without having that luxury of being there to go over all the nuances about data.*" One instructor reported that without access to on-campus computer labs, they were "*worried that students might not get the application piece, the interpretation. So... well, the remainder of the spring was just basically trying to survive.*"

The strategies employed by instructors included changing the software used in classes, simplifying datasets, altering data collection assignments, and providing video tutorials from other sources to complement their asynchronous instruction. One instructor noted that "*with the change to distance learning during COVID, that actually completely changed the way I've worked with the data and the students.*" After the spring semester, some faculty took advantage of teaching and learning resources to redesign their courses for the online environment, and provided new ways to support students, sometimes with graduate teaching assistants. This was vital, as some instructors remarked that the challenges of remote learning seem to have exacerbated existing uncertainty or fear of numbers or statistics, at least for some students: "*They still come in with the fear, and even more so with distance learning.*"

The impact of remote instruction had a significant impact on instructors' choice of software and other tools used to manipulate and analyze data. By and large, instructors leaned even more heavily on software that was free and/or openly available during the COVID-19 pandemic, since visiting campus-based computing labs was not a viable option for students. For example, one instructor commented, "*This year because we are completely remote, SPSS wasn't going to work because students don't have access to it and it's costly. So I am using a software package called JASP, which is free and can be downloaded from the internet and actually has a lot of the same utility as SPSS.*"

At least one instructor was aware of MSU's Virtual Desktop Initiative (VDI), a service that provides remote access to software packages normally available only in physical computing labs. When discussing the VDI, they stated, "*[students] could access it for free because normally they could go to a computer lab and it would be available. But because it's remote now, they can access it via… it's called VDI. But it's basically a remote desktop where they sign onto the university server and then it identifies them as an internal user. And so they do have access to it that way. So that's what they use.*" However, this service was not always the perfect solution. The same instructor went on to say, "*Even the remote desktop doesn't always work… There was actually one case, where even working with MSU IT, apparently their computer is not compatible for some reason.*"

Some changes brought about by the shift to remote learning appeared to have benefitted both students and instructors, and instructors speculated about whether these changes would persist once in-person classes become the norm again: "*I'd be very interested to see how we go back to on-campus computer labs, the same old way I've taught for 25 years. Or will we incorporate some of the online characteristics that we have had to incorporate in the last two semesters, will we continue to do that?*"

Recommendations

Faculty are resourceful and determined. We identified time as the crucial resource needed to teach the data skills necessary for social science education and workforce preparation. Libraries are adept at identifying, providing, and creating resources that "save the time of the user" (Ranganathan's Fourth Law of Library Science), and most of our recommendations intervene in this way.

Resources

- Create a regularly scheduled help room (physical and virtual) dedicated to social science data. The College of Social Science's current Economics Help Rooms can serve as a model/exemplar for this service.
    - Responsible parties: CSTAT and MSU Libraries
- Provide curated, cleaned, and well-documented data from commonly used sources that are ready to support a variety of standard statistical analyses and use cases.
    - Responsible party: MSU Libraries

- Streamline pointers and access to general data resources for undergraduates and instructors.
    - Responsible party: MSU Libraries
- Create and offer data cleaning workshops.
    - Responsible parties: CSTAT and MSU Libraries
- Create and provide a data cleaning service.
    - Responsible party: CSTAT
- Provide targeted funds for Teaching Assistants for data-intensive class instruction, to manage course loads, teach labs and tutorials, and make process-product assessments like data-rich research possible.
    - Responsible parties: University Administration and academic departments

Technologies

- Survey Social Science departments with focus on needs relating to computing, physical space, and remote access to common software and platforms. Advocate for financial resources to expand access through MSU Virtual Desktop Initiative (VDI).
    - Responsible party: MSU IT
- Provide an enterprise-wide calendar application for short-duration workshops, and increase cross-unit coordination and promotion.
    - Responsible party: MSU IT
- Promote the re-use of MSU-generated data by searching research data repositories (including MSU's Dryad holdings) and preparing curated extracts for teaching.
    - Responsible party: MSU Libraries

Training

- Create an introductory data literacy skills seminar series aimed at first and second year students in the Social Sciences.
    - Responsible parties: CSTAT and MSU Libraries
- Launch an Academic Advancement Network (AAN)-sponsored learning community on the topic of teaching with data.
    - Responsible parties: Professors and instructors

- Promote and provide funding for data-centric professional development for librarians (e.g. Data Carpentry), so that librarians can be credible and valued partners in teaching data in the social sciences.
    - Responsible party: MSU Libraries administration
- Create a data privacy and ethics workshop for undergraduates, drawing on the student data working group report produced in 2020, and partnering with participants in the group.
    - Responsible parties: MSU Libraries, Human Research Protection Program (HRPP), Undergraduate Research and Creative Activity (URCA) Office
- Offer disciplinary-specific instructional sessions aimed at undergraduates on finding data, using download tools, and getting and creating documentation, including using research data repositories.
    - Responsible party: MSU Libraries
- Provide support for graduate student teaching assistants in the social sciences, through specific outreach and promotion of the resources described above. Consider embedding a librarian to work alongside TAs.
    - Responsible parties: University administration and academic departments

Promotion/Marketing

- Add relevant offerings to MSU's Research Technology and Support Service catalog: https://tech.msu.edu/service-catalog/research-technology-collaboration/
    - Responsible parties: MSU Libraries and Research Facilitation Network members
- Identify extant MSU Libraries resources germane to teaching with data (e.g., SAGE Research Methods with "practice data," STATISTA, and Gallup) and create a new research guide to be shared widely.
    - Responsible party: MSU Libraries

Conclusion

It should come as no surprise that faculty are resourceful and largely self-sufficient. They typically find and prepare data for use in teaching themselves, using a mix of well-known sources, library-provided sources, industry sources, data collection, and web searching, and they are generally able to point students to good sources of data. However, in our interviews, instructors expressed interest in working with partners both on and off-campus, both in the

context of providing students with extra assistance as well as attending targeted training workshops themselves. Desire for physical space for computing and collaboration was also mentioned, though tempered and reevaluated due to the COVID-19 pandemic.

Faculty streamlined processes to make course concepts less intimidating and to focus attention on teaching analysis, and when applicable, relevant software tools. The importance of specific software tools varied between instructors, but all of them worked to balance the weight they gave to the tool with the exigencies of teaching, especially during the pandemic. Other impacts of the pandemic on teaching included supplying resources, such as video tutorials, to augment or take the place of face-to-face classroom training and office hours. While shifting to an online environment did involve significant effort, the work appears to have been largely successful and had positive ripple effects.

A theme throughout our interviews was student readiness (or lack thereof) to work with data. When courses are not advanced or upper division courses, instructors typically have low expectations for how prepared students will be to find and work with data and data analysis tools. However, taking the time to address the data literacy of students at the beginning of the semester can reduce the amount of time students have to work with and practice new data skills over the term. Balancing the need to teach basic data literacy skills with the desire to cover specific, discipline-relevant data analysis methods or tools was one of the most critical elements in course design mentioned by instructors.

Another important finding was the number and variety of ethical issues mentioned by instructors in data-based instruction in the social sciences. Most instructors were keenly aware of the ethical considerations faced in collecting and using data in their respective disciplines, and took steps to address those challenges with their classes. Some instructors cleaned or prepared data for their classes to mitigate any potential ethical issues, while others devoted parts of their syllabi to ethics and/or incorporated IRB training modules into their instruction.

While faculty identified several challenges in teaching with data, they did not mention campus resource providers as particularly lacking. Rather, they pointed to a scarcity of time, teaching assistance, and professional development incentives as barriers to course improvements. On the contrary, campus partners were spoken of very highly, and several instructors had stories to tell of specific times when librarians or other data professionals were especially helpful in supporting students in finding or using data in their classes. In general, faculty expressed both a willingness and a readiness to work with the Libraries and other campus resource providers.

Libraries excel at providing access to data and databases, offering assistance in finding appropriate datasets, providing physical spaces for individual and group study, and providing instruction in topics such as information literacy, data analysis, and data visualization. The unique confluence of mediated and unmediated access to data-centric resources offered by the MSU Libraries holds promise. We envision the potential for the Libraries and other campus partners to continue to assist with and advance teaching with data, in the social sciences and beyond. With the growth and pressure on using data in undergraduate instruction, we anticipate more need, and can move from a position of strength to provide what faculty need before it becomes a pain point or major gaps appear.

References

Guler, Gulsen. “Data Literacy from Theory to Reality: How Does It Look?” Master Thesis, Vrije Universiteit Brussel, 2019. https://www.researchgate.net/publication/335620777_Data_literacy_from_theory_to_reality_How_does_it_look

NIST Big Data Public Working Group, Definitions and Taxonomies Subgroup. *NIST Big Data Interoperability Framework: Volume 1, Definitions*. NIST Special Publication (SP) 1500-1. Gaithersburg, MD, National Institute of Standards and Technology, 2015. http://dx.doi.org/10.6028/NIST.SP.1500-1

Ranganathan, S. R. (Shiyali Ramamrita). *The Five Laws of Library Science*. Ranganathan Series in Library Science. Bombay, Asia Pub. House, 1963.

Appendix A: Interview Instrument

Interview Guide: Teaching with Data in the Social Sciences

Note regarding COVID-19 disruption I want to start by acknowledging that teaching and learning has been significantly disrupted in the past year due to the coronavirus pandemic. For any of the questions I'm about to ask, please feel free to answer with reference to your normal teaching practices, your teaching practices as adapted for the crisis situation, or both.

Background

Briefly describe your experience teaching undergraduates.

- How does your teaching relate to your current or past research?
- In which of the courses that you teach do students work with data?

Getting Data

In your course(s), do your students collect or generate datasets, search for and select pre-existing datasets to work with, or work with datasets that you provide to them?

If students collect or generate datasets themselves Describe the process students go through to collect or generate datasets in your course(s).

- Do you face any challenges relating to students' abilities to find or create datasets?

If students search for pre-existing datasets themselves Describe the process students go through to locate and select datasets.

- Do you provide instruction to students in how to find and/or select appropriate datasets to work with?
- Do you face any challenges relating to students' abilities to find and/or select appropriate datasets?

If students work with datasets the instructor provides Describe the process students go through to access the datasets you provide. *Examples: link through LMS, instructions for downloading from database*

- How do you find and obtain datasets to use in teaching?
- Do you face any challenges in finding or obtaining datasets for teaching?

Working with Data

How do students manipulate, analyze, or interpret data in your course(s)?

- What tools or software do your students use? *Examples: Excel, online platforms, analysis/visualization/statistics software*
- What prior knowledge of tools or software do you expect students to enter your class with, and what do you teach them explicitly?

- » To what extent are the tools or software students use to work with data pedagogically important?
- » Do you face any challenges relating to students' abilities to work with data?

How do the ways in which you teach with data relate to goals for student learning in your discipline?

- » Do you teach your students to think critically about the sources and uses of data they encounter in everyday life?
- » Do you teach your students specific data skills that will prepare them for future careers?
- » Have you observed any policies or cultural changes at your institution that influence the ways in which you teach with data?

Do instructors in your field face any ethical challenges in teaching with data?

- » To what extent are these challenges pedagogically important to you?

Training and Support

In your course(s), does anyone other than you provide instruction or support for your students in obtaining or working with data? *Examples: co-instructor, librarian, teaching assistant, drop-in sessions*

- » How does their instruction or support relate to the rest of the course?
- » Do you communicate with them about the instruction or support they are providing? If so, how?

To your knowledge, are there any ways in which your students are learning to work with data outside their formal coursework? *Examples: online tutorials, internships, peers*

- » Do you expect or encourage this kind of extracurricular learning? Why or why not?

Have you received training in teaching with data other than your graduate degree? *Examples: workshops, technical support, help from peers*

- » What factors have influenced your decision to receive/not to receive training or assistance?
- » Do you use any datasets, assignment plans, syllabi, or other instructional resources that you received from others? Do you make your own resources available to others?

Considering evolving trends in your field, what types of training or assistance would be most beneficial to instructors in teaching with data?

Wrapping Up

Is there anything else from your experiences or perspectives as an instructor, or on the topic of teaching with data more broadly, that I should know?

Appendix B: ITHAKA S+R's Guidelines for Disciplinary Inclusion in the Social Sciences

From the *TDSS Project Scope and Recruitment* document:

Instructors may be interviewed for this project if they teach "teach with data" in one or more undergraduate courses offered through one of the departments below, even if their research is affiliated with another department.

Courses listed or cross-listed in the following departments are considered in-scope for the purposes of this study:

Anthropology
Archaeology
Area Studies
Communications
Cultural Studies (e.g. African American, Gender)
Economics
Education
Environmental Science
Geography
History
Law/Criminal Justice
Linguistics
Political Science
Psychology
Public Health/Public Administration/Social Work
Sociology
Urban Studies/Urban Planning

Some institutions are moving toward nontraditional or cross-disciplinary academic department scoping. Use your best judgement as to which of the above descriptors best match the departments at your institution. In general, we are using a broad definition of the social sciences in order to seek out rich examples of teaching with data.

Appendix C: Summary of Participants

Total number of participants: 10

Participant Rank

- Professor (2)
- Associate Professor (6)
- Assistant Professor (2)

Unique Departments/Programs Represented (8)

- Department of Anthropology (2)
- Department of Economics
- Department of Geography, Environment, and Spatial Sciences
- Department of Linguistics, Languages, and Cultures
- Department of Psychology
- Department of Sociology
- School of Social Work
- Urban & Regional Planning Program (School of Planning, Design and Construction)

Unique Colleges Represented (4)

- College of Agriculture and Natural Resources
- College of Arts & Letters
- College of Social Science (7)
- James Madison College

Appendix D: Sources of Data and Software Identified by Respondents

Sources of Data

- American Community Survey (ACS)
- The Digital Archaeological Record (tDAR)
- Fake/made up/toy data
- General Social Survey (GSS)
- Government contacts (e.g. planning departments)
- Industry contacts or partners
- MSU Libraries
- Open Context
- Penn Museum of Archaeology and Anthropology
- Repositories, data clearinghouses, open data portals (e.g. data.gov)
- Sanborn Maps
- US Decennial Census
- USGS EarthExplorer
- USGS National Map
- YouTube

Software

- ArcGIS
- Excel
- Google Workspace
- GRASS GIS
- GSS (Guided Statistics for Scientists)
- JASP
- JavaScript data libraries
- KOBO Toolbox
- MAXQDA
- OpenRefine
- Python
- QGIS
- R
- SPSS
- STATA
- Survey123
- Tableau