10-1-2021

Teaching with Data in the Social Sciences: The Purdue Report

Jane Yatcilla
Purdue University, janeyat@purdue.edu

Clarence Maybee
Purdue University, cmaybee@purdue.edu

Bethany McGowan
Purdue University, bmcgowa@purdue.edu

Gang Shao
Purdue University, gshao@purdue.edu

D Trevor Burrows
trevor.burrows@gmail.com

TEACHING WITH DATA IN THE SOCIAL SCIENCES: THE PURDUE REPORT

PURDUE UNIVERSITY

LIBRARIES AND SCHOOL OF INFORMATION STUDIES

Jane Kinkus Yatcilla, Associate Professor, Health & Life Sciences Division
Clarence Maybee, Professor, W. Wayne Booker Chair in Information Literacy
Bethany S. McGowan, Associate Professor, Health & Sciences Division
Gang Shao, Assistant Professor, Research Data Division
Trevor Burrows, Post-doctoral Fellow, Humanities, Social Sciences, Business & Education Division

October 1, 2021

EXECUTIVE SUMMARY

Purdue University Libraries and School of Information Studies participated in a multi-university qualitative study about how instructors teach with data in social sciences courses. The study was organized by ITHAKA S+R, a not-for-profit organization that organizes research studies to help academic libraries and their parent organizations understand evolving trends within their communities, and provide data to inform their decision making.

RESULTS !

Thirteen semi-structured interviews were conducted with social sciences faculty who teach undergraduate students to use data. The interview questions focused on discerning how faculty 1) connect students to data, 2) have them work with data, and 3) what types of training and support faculty have received that informs their approach and efforts to teaching undergraduates about using data. Participating faculty members were from 12 different departments located across 5 different colleges/schools at Purdue. The qualitative data collected through the interviews were analyzed using thematic analysis. The analysis of the transcripts revealed four themes:

- Theme 1: Array of Purposes and Practices for Teaching about Data
- Theme 2: Instructors' Experiences, Needs, and Perspectives
- Theme 3: Managing Wide-Ranging Student Confidence and Ability
- Theme 4: Technical Concerns When Teaching With Data

Key findings related to Theme 1 are that instructors' work with students ranged from exercises intended to help students begin to understand the concept of 'data' and working with it to engaging in research projects in which the students learn several aspects of managing and analyzing data in line with social science research. Most instructors emphasized that the primary purpose of teaching about using data is that it is relevant to the students' understanding of the discipline or the social sciences more generally.

Key findings related to Theme 2 reveal that most instructors learned how to teach both their disciplinary content, as well as teach with data, in graduate school. Interviewees expressed a need for professional development on how to teach with data; training on software packages; statistical methods; and how to develop students' critical thinking skills with regard to data. Most reported difficulty in finding data appropriate for their teaching, describing challenges like finding data in the best format for their needs, finding data with all desired variables, and finding current rather than outdated data. While several interviewees talked about campus data-related initiatives like AMAP, IDSI, or the DataMine, almost half did not admit awareness of any such initiatives.

Key findings related to Theme 3 are that while teaching with data in the social sciences is a growing trend and that students' data literacy competencies are improving, difficulty managing

wide-ranging student confidence and abilities recurred as a key challenge for instructors. Ranges in student confidence and ability impact instructors differently and seem to be influenced by course level, with faculty in upper-level courses expecting mastery of some basic data literacy competencies as a prerequisite. Instructors list several curricular strategies for managing wide-ranging student confidence and ability, and also mentioned extracurricular and co-curricular strategies.

Key findings related to Theme 4 suggest that technical concerns mainly focused on data accessibility and software logistics. Most of the interviewees prefer to share data with students using platforms that students are familiar with (like Purdue's LMS, Brightspace). When big and complex data are used in class, external resources like IT support and online data portals would be considered for data access. Instructors also need to balance the desire to make access as easy as possible with the size and complexity of the data when preparing and selecting datasets for their courses. Logistical questions related to software used in class are complex and diverse. Some instructors choose to use particular software in their courses based on the availability and accessibility of the software on campus, while some choose the software based on their own experience or established curriculum. Software versioning, documentation, and user support are all important factors to be considered as some software is technically available on campus computers but does not always run as expected.

RECOMMENDATIONS !

The findings from the study suggest actions to be taken to better support teaching with data in the social sciences. Campus level recommendations include:

- Campus-wide recognition that teaching with data in social science areas raises its own unique set of challenges;
- Continual assessment of the specific needs of instructors working with data to better understand and address those challenges; and
- Programming and resources that support instructors in their work.

The Libraries has been supporting campus data needs for several years, through the collections, services, and courses it offers. The results of this study can inform Libraries' services and programming in several ways. Specific recommendations include:

- Market data-related Libraries courses to students in the social sciences;
- Improve campus access to open and proprietary data-related resources;
- Work with CIE to address instructors' pedagogical needs; and
- Conduct a campus-wide needs assessment.

INTRODUCTION

In fall 2019 ITHAKA S+R[1] invited Purdue University Libraries and School of Information Studies (the Libraries, hereafter) to participate in a multi-university qualitative study about how instructors teach with data in social sciences courses. ITHAKA S+R is a not-for-profit organization that organizes research studies to help academic libraries and their parent organizations understand evolving trends within their communities, and provide data to inform their decision making.

Purdue University is a large, publicly-funded land-grant research university in the Midwest, with over 34,000 undergraduate students, 10,000 graduate students, nearly 1,000 professional students, and over 16,000 FTE faculty and staff. Purdue has been classified by the Carnegie Classification of Institutions of Higher Education as R1, i.e., a doctoral degree granting institution with a high level of research activity. Purdue has an extensive tradition in teaching with data and research with data, especially in STEM-related disciplines. In the fall of 2017, Purdue University held open forums across four campuses to shape the data vision at the university level. A series of actions were then undertaken to build a campus-wide data science education ecosystem.

UNIVERSITY LEVEL INITIATIVES !

Purdue has engaged in several initiatives to advance data science across the university. The Integrative Data Science Initiative (IDSI) was created by the Provost's Office to advance campus-wide data science research and education. The Advanced Methods at Purdue (AMAP) program focuses on increasing data expertise in the behavioral, health, and social sciences. There is also a large-scale learning community for undergraduate students from all majors, the Data Mine. In 2021, a project was conducted called IMPACT Data Science Education, in which librarians worked with faculty across campus to integrate data science into undergraduate coursework. An active and multidisciplinary community of scholars focused on Critical Data Studies also supports programming related to data ethics and critique.

The Integrative Data Science Initiative (IDSI) was launched in 2018 to establish a "Data Science for All" education ecosystem, strengthen students' data literacy and fluency, and prepare students for the rapidly growing data-driven future. IDSI's education thrust is to train students with fundamental knowledge of data science and build deep analytical thinking skills for big data challenges. IDSI embeds data science into domain curricula. IDSI is a campus-wide

[1] Refer to Appendix 1 for a list of links used throughout this report.

collaboration focusing on students in all majors. Detailed curriculum and educational components can be found via IDSI Education Ecosystem Website.

Advanced Methods at Purdue (AMAP) is a multi-college initiative focusing on behavioral, health, and social sciences, with the goal of increasing methodological expertise and collaborative opportunities. AMAP facilitates interdisciplinary collaboration by affiliated faculty from different colleges, hosting regular brown bag seminars and workshops, and inviting external speakers to share experience and exchange knowledge on research and graduate education. AMAP also offers a graduate interdisciplinary methods certificate and provides services to support students and faculty with regard to grants.

The Data Mine is the first large-scale living learning community on campus for undergraduates from all majors. It's a living, learning and research-based community, where students will live in the same dorm on campus, Hillenbrand Hall, attend regular data seminars, collaborate with other students, faculty, and industry partners to solve real-world data science problems. There are also 20 sub-learning communities covering a wide range of study fields available on campus. Students within each sub-learning community share course curriculum, participate in research and professional development programs offered by departments, research centers, and colleges throughout the university.

The IMPACT Data Science Education project was created to support Purdue instructors to develop and integrate data-focused student projects into undergraduate courses.[2] Funded by an Integrative Data Science Initiative grant, the project was led by three faculty members from the Libraries, a mathematics professor who directs the Data Science Consulting Service, and a staff person from the Center for Instructional Excellence. Nine instructors participated in the project in spring 2020 to develop coursework for eight courses with an approximate overall enrollment of 1890 students. The student projects created by the instructors spanned a range from basic data literacy exercises to having students learn computational methods for analyzing different types of datasets. Seven more instructors are participating in the project in fall 2021.

Critical Data Studies is a multidisciplinary community, whose disciplines range across many Purdue colleges, of scholars examining the ethical, legal, social, cultural, political and epistemological aspects of data science, big data and digital infrastructures. CDS provides a monthly seminar series, a public lecture series, and undergraduate and graduate courses offered by multiple departments, including the Libraries.

[2] Maybee, C., Lin. G., Zakharov, W. & Cai, C., Fitzsimmons, J. & Sun, Y. (2021). Building undergraduate data literacy through faculty development. In J. Bauder (Ed.). *Teaching Critical Thinking with Numbers: Data Literacy and the Framework for Information Literacy for Higher Education.* Chicago: ALA Editions.

COLLEGE/SCHOOL LEVEL INITIATIVES !

Purdue University offers majors, minors, and certificates in data science from various colleges/schools, including

- The College of Science offers a Data Science Major to undergraduate students in the departments of Computer Science and Statistics, and a Master's degree in Geodata Science in the department of Earth, Atmospheric, and Planetary Sciences.
- Purdue Polytechnic Institute offers a Bachelor's Degree in Data Visualization from the Computer Graphics Technology Program in the department of Computer Graphics Technology.
- Krannert School of Management offers a Master of Science Degree in Business Analytics and Information Management to graduate students.
- The College of Agriculture offers a Data Driven Agriculture Minor to undergraduate students, and a Spatial Data Science Certificate at the graduate level.
- The College of Engineering offers Engineering in the World of Data for first-year students who are interested in data science.
- The Libraries offers a Graduate Certificate in Geospatial Information Science (GIS).

PURDUE LIBRARIES AND RESEARCH DATA !

The Libraries support of research data needs on campus was ramped up as soon as Dean James L. Mullins was hired in 2004. Mullins began by conducting a campus-wide data needs assessment in August 2004 and rapidly responding to the survey results. He created a new administrative position of associate dean for research, which was filled in November 2005, and developed descriptions of additional faculty and staff positions for the unit. Over the next several years, as the unit came into existence, The Libraries rolled out numerous data-related services, including:

- Purdue e-Pubs, an institutional repository for documents, such as preprints or white papers, authored by Purdue faculty, staff, or students.
- The Distributed Data Curation Center (D2C2), a virtual center highlighting research in data curation by Purdue's library science faculty and their collaborators
- Purdue University Research Repository (PURR), an online space for collaboration and data sharing for Purdue researchers and their collaborators
- Databib, a tool for helping researchers identify and locate online repositories of research data, which was eventually merged with the international Registry of Research Data Repositories (Re3Data).
- Data Management Toolkit, a collection of Open Access resources to help Library and Information Science professionals work with researchers to craft a data management plan aligned with their specific needs as well as with funding agency requirements.

As of July 2021, the Research Data unit is comprised of four faculty and three professional staff members (a data curator, data repository outreach specialist, and digital library software developer) who consult with members of the Purdue community on various aspects of research data management, including data management and curation education; data management planning; metadata and date documentation; working with big data; publishing and sharing data sets, including adding a Digital Object Identifier for citation purposes.

In addition to services provided by Data Services, the Libraries faculty teach several for-credit courses about data. These include

- Introduction to Data Lifecycle Management (1 credit)
- Data Management at the Bench (3 credits)
- Data Management and Curation for Qualitative Research (3 credits)
- Data Sharing and Publication (3 credits); Critical Data Studies (3 credits)
- Data Science and Society: Ethical, Legal, and Social Issues (3 credits)
- Special Topics in Information and Data Science, a variable course title that can be used to address various facets of data science depending on student or disciplinary needs (1-4 credits).

Additional Libraries courses that work with data, but do not specifically address general data management, include

- DeepAI for Information Sciences (1 credit)
- GIS for Public Health Research (3 credits)
- GIS Project Experience (3 credits)
- Digital Humanities Foundations (3 credits)
- Computational Text Analysis (3 credits).

METHODS

The research plan provided by Ithaka S+R, outlining methods for participant selection, data collection and data analysis, aims to contribute to our understanding of how faculty in the social sciences teach undergraduate students to use data. This section describes how the research process was carried out at Purdue University. Different members of the research team participated in different parts of the research process. One member was primarily responsible for participant recruitment activities, although other members of the team contributed to identifying potential participants. Three members of the team conducted the interviews, and 4 members conducted the data analysis. Approval from Purdue's Institutional Review Board was obtained before the research was undertaken.

PARTICIPANT RECRUITMENT !

A pool was created of potential participants who met the criteria for participation, i.e., taught undergraduate students to use data in the social sciences. To identify potential participants, 2 members of the research team conducted informational interviews with two members of the Purdue community who are leaders and supporters of data in higher education. One interview was with a faculty member who created a larger undergraduate learning community for students from all majors to learn about and use data. The other was with an associate dean who is known for supporting data projects in the College of Liberal Arts. Potential participants were also drawn from Purdue's Integrative Data Science Initiative's web page listing faculty affiliates, which provides a list of faculty engaged with data that includes their field and if they work with undergraduate students. Additionally, members of the research team added names of social science faculty they believed taught undergraduate students about using data.

An email was sent to each of the faculty in the original pool seeking their participation in the study. While some agreed to participate, others declined. Despite the information gathered by the research team, the primary reason faculty stated for not participating was that they did not work directly with undergraduate students. However, many of those who declined suggested the names of colleagues that they recommended the research team contact. From the 42 instructors identified through these various efforts, 14 were interviewed. Owing to a technical difficulty, 13 completed interviews were videoed and transcribed. As shown in Table 1, while almost half of the participants (n = 6) were assistant professors, associate professors and professors participated as well. Participating faculty members were from 12 different departments located across 5 different colleges/schools at Purdue.

Table 1. Participant demographics

Faculty rank	College	Primary department
Assistant professor	Liberal Arts	Anthropology (2)
	Liberal Arts	Communication
	Liberal Arts	Sociology
	Health & Human Sciences	Human Development & Family Studies
	Health & Human Studies	Psychological Sciences
Associate professor	Education	Educational Studies
	Liberal Arts	Linguistics
	Libraries & Information Studies	Libraries
Clinical associate professor	Honors	Honors College (teaching in a social sciences content area
Professor	Education	Curriculum & Instruction
	Liberal Arts	Archaeology
	Liberal Arts	Political Science

DATA COLLECTION !

Data were collected through semi-structured interviews. Three members of the research team conducted the interviews with the 13 faculty participants. An information sheet outlining the study and highlighting potential risks was shared with the interviewees before meeting for the interview. All the interviews, approximately 60 minutes in length, were held over Zoom. The member of the research team conducting the interview would begin by explaining the purpose of the study. Acknowledging that the COVID-19 changed teaching strategies during the past year, the participants were apprised that during the interview they could share their regular approaches to teaching about data and/or how it may have changed during the pandemic. The interview questions focused on discerning how faculty 1) connect students to data, 2) have them work with data, and 3) what types of training and support faculty have received that informs their approach and efforts to teaching undergraduates about using data. Once the interviews were completed, the Zoom recordings were stored in the 'Box,' a password protected file repository subscribed to by Purdue. The videos were transcribed by a Libraries' staff person. The anonymized transcripts were made available to the 4 members of the research team conducting data analysis.

DATA ANALYSIS !

The qualitative data collected through the 13 participant interviews were analyzed using thematic analysis. As trained by Ithaka S+R, the four members of the research team engaged in 4 stages of the analysis process (outlined in Figure 1).

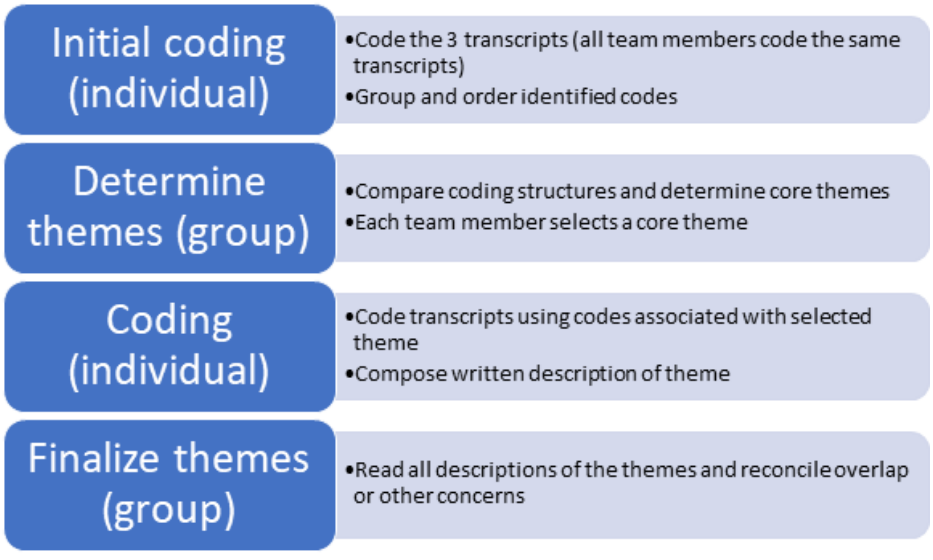

Figure 1. Four stages of coding process

The coding team began by coding the same three transcripts individually and grouping the codes

they identified. Then they met on Microsoft Teams and shared their results. An initial effort was made to categorize the codes using the 'whiteboard' feature of the Teams software. Two team members further refined the codes into 8 potential themes, which included 1) instructor experience, 2) course planning, 3) technical issues, 4) data management, 5) pedagogic practices/challenges, 6) practices/challenges related to student experience, 7) student perspectives, and 8) supporting social science students. The 8 themes were confirmed by the research team at a follow up meeting.

Each of the four coders selected 2 themes that were closely related (such as student perspectives and supporting social science students) and corresponding sets of codes, which were used to code the remaining 10 transcripts. Three of the researchers used NVivo software to code the transcripts, while one coded manually. After completing the coding and analysis, each researcher wrote a description of one theme, which was a consolidation of the two themes identified in the early part of the analysis. This consolidation of themes was not a planned effort, but was an outcome of the analytical work of each of the members of the research team. The four resulting themes are described in the 'Results and Discussion' section.

RESULTS AND DISCUSSION

As detailed in this section, the analysis of the transcripts revealed four themes:

- Theme 1: Array of Purposes and Practices for Teaching about Data
- Theme 2: Instructors' Experiences, Needs, and Perspectives
- Theme 3: Managing Wide-Ranging Student Confidence and Ability
- Theme 4: Technical Concerns When Teaching With Data

THEME 1: ARRAY OF PURPOSES AND PRACTICES FOR TEACHING ABOUT DATA !

The 13 instructors interviewed in this study shared a variety of ways they have students engage with data. The instructors' work with students ranged from exercises intended to help students begin to understand the concept of 'data' and working with it to engaging in research projects in which the students learn several aspects of managing and analyzing data in line with social science research. Ten of the 13 instructors emphasized that the primary purpose of teaching about using data is that it is relevant to the students' understanding of the discipline or the social sciences more generally. One instructor working with students from various social science majors on interdisciplinary research projects emphasized how working with social science data may help students begin to identify as social science researchers:

> …I've seen so many students change their minds and like how they think about themselves as a social scientist once they like get a little bit more into the actual coding of things and actual manipulation of data and stuff. (Interview #1)

Another instructor who was teaching an honors course on geospatial analysis identified that the data analysis they can do in the course may enable students to recognize the significance of the work of the discipline:

> …we had a week in class where we looked at satellite imagery—satellite image analysis or archeologically looted sites—where they were learning how to do kind of a remote sensing visual analysis of sites that had been looted over time. … when I teach about archeology and they're [students] not innately interested in archeology, I have hard time convincing them of the relevance of archeology to everyday life. And this was an example of well here are things happening on the ground right now that are impacting people's ability to preserve archeological sites, to understand the past and that sort of thing. (Interview #13)

An instructor teaching an introductory undergraduate course on statistics for psychology emphasized the importance of students that were intending to pursue graduate degrees to recognize the importance of data usage and tools in the social sciences:

> …most students in my class don't go on to grad school in the social sciences, right? But a subset do and so I want to make it clear to them why this is relevant and why they will continue to build on these kind of mathematical skills, these computational skills, and potentially be using this software for many years to come. (Interview #8)

A few instructors also focused specifically on using data to learn disciplinary content. For example, undergraduate students may enroll in a lower-level graduate course that focuses on analyzing speech acoustics to learn about phonological structures in different languages. A couple of the instructors emphasized that learning about using data would be useful to students in their future studies. Five of the instructors linked learning to use data to future career prospects. An instructor teaching a course on science education to future teachers recognized that the approach taken to teaching them about analyzing and visualizing data could act as a model for the students to follow when teaching children to work with data in elementary schools. An archeologist expressed that the students in the course do not always intend to become archeologists, but learning to use technologies, such as ArcView, provides the students with transferable skills that may be useful in various professional contexts.Most of the instructors who participated in the study introduced their students to software or online tools to examine data. The instructors largely taught the students to use these tools rather than having the students engage with online tutorials or other training options outside of the course. A wide range of tools

were used across the various courses. For example, in a political sociology course the instructor introduces students to basic ideas about data by having them use Google Trends to explore a public concern in relation to time or geographic location. In some cases, the software used is specific to the discipline. In the course focused on speech acoustics students learn to use 'Praat,' a software used specifically to analyze speech.

Of course, some instructors introduce their students to more commonly used data software. One instructor works with undergraduates to do independent studies in which the students use SAS to analyze survey data, and QDA Miner and Excel to code qualitative data. In a political science research methods course the students learn Stata. Although the students may elect to use another software unsupported by the instructor, the use of Stata is encouraged because of the instructor's ability to provide support. Some instructors felt more comfortable teaching students with tools they had prior experience with, such as tools they learned to use as a graduate student or post-doc, rather than searching for new or open tools.

There was wide variation in what students are taught about using data once they have access to it. Most of the instructors did not expect their students to be familiar with using data beyond perhaps the basics of Excel when the course begins. One instructor noted that some students, especially those from engineering or computer science, enter the course with a lot of data experience, while others are novices. Therefore, the instructor must work with students at whatever level of experience they bring to the course. Another instructor noted that students' lack of experience at the beginning of the class is not necessarily predictive of their ability to use data:

> I can start with a student that says, "What does code mean?" and then two weeks later they just can't get enough of it and they're just like, "Oh man, this is great!" And there are others who are in the same boat at the beginning and just at the end they are just sort of pulling their hair and going like, "I can't do this!" (Interview #4)

There was a range of ways that instructors had students engage with data. This typically aligned with goals for the course. A few instructors mentioned having their students clean or organize data. Only one instructor specifically mentioned having students transcribe interview data. Another instructor mentioned having students create a data plan. While a few courses, such as the political science research methods course, had students conducting statistical analysis, other courses only provided students with an initial foray into data analysis. The instructor teaching the political sociology course that has students work with Google Trends or other data explained that the purpose of the exercise is to have students think about aspects of data analysis:

> …the exercise being getting them thinking actively about operationalization, essentially. You know, how certain questions are asked on a survey and what that

> looks like when we compare those answers to another variable or to think carefully about what we're actually measuring when we do something like trends in search terms across cities or across years or something like that. (Interview #6)

Data visualization was focused on in 8 of the courses described by the interviewees. While some instructors had students create visualizations a few only taught students 'about' different aspects of visualization:

> …we look at examples of … good and bad published graphs, but the students aren't making the graphs. It's more like why was it made this way? What's the rationale? How should you do this for your own papers? (Interview #8)

Three instructors mentioned conceptualizing data as something they wanted their students to be able to do. The instructor of the course that focuses on analyzing speech acoustics explained:

> …one interesting thing about working with language data is that you get to approach a familiar source from an unfamiliar perspective, right? Because we all…speak a language and we never think of language as data. …It's just part of being human. So they have to kind of reverse their thinking a little bit. …Stop looking at it not as just something they do in everyday life, but as data so they can analyze. (Interview #9)

Data evaluation was focused on by 8 of the instructors. While the primary focus is on evaluating data in a research context, instructors noted this may enable students to become more informed consumers of statistical information that students may come across in their daily lives. The instructor of the introductory courses on statistics for psychology indicated that students would be able to apply what they learned in the course to statistical data they encounter in the media:

> I try to get them to be good consumers of science and I always use the example of how if you watch like an infomercial late at night, they will inevitably throw some kind of bar graph or something. It's usually like wildly misleading, right? Like it's not real science. But if you don't know better, it can look very convincing and so I try to teach them like how to be good kind of critical consumers of science whether it's an actual scientific published article or just something you see in the news or something you see online. (Interview #8)

By contrast, an instructor who teaches geographic information system (GIS) mapping in in various courses observed that some students do not apply a critical lens when engaging with data, and notes that more should be done to address this in future:

> …they seem very excited about the data they find.…especially if it's a federal agency's database. They don't really question like, is there any underlying issues? Is they any estimation? What's the accuracy about the data? Is it time sensitive or

> a couple years ago? When I ask the question I think I never saw students say the data has some underlying issues or something. They say, "I find from this website and it's a federal website. It should be very reliable. I should trust it." …I always have a concern after I read those. I think I should spend more time to explain the data source. (Interview #11)

Five instructors mentioned interpretation of data as something that students learn in the undergraduate courses that they teach. For example, the instructor teaching the science education course focused on having the students look for trends and patterns and determine the claims that could be made. Ethical concerns were not a major focus of instruction for most of the interviewees. Some instructors described enabling their students to understand ethical concerns as part of the research process, such as considering how to minimize bias in the collection of data.

THEME 2: INSTRUCTORS' EXPERIENCES, NEEDS, AND PERSPECTIVES !

Another broad theme we coded addressed the instructors' experiences, needs, and perspectives. Most of the interviewees reported having received little to no formal instruction in teaching with data. When asked whether they had had any formal training on how to teach with data outside of their graduate school experience, a majority also referred to their time as graduate students or graduate teaching assistants as where they learned to both teach their disciplinary content as well as teach with data. One comment that seemed to represent many of the responses is

> …I'm from a generation where even as a graduate student as part of your graduate education you probably weren't taking a class on how to do any kind of teaching as a political scientist. You just did it. So our grad students now, they actually do have some teaching workshops. But when I was in graduate school there weren't even…at least the grad school I went to…they weren't teaching workshops and those sorts of things. (Interview #2)

Another instructor stated that their field is very specialized, and therefore few professional development opportunities exist.. They continued,

> ..most of the practitioners in my area are self-taught essentially...we teach ourselves and we exchange our expertise...we learn from each other and we learn ourselves and that's pretty much how we do things. (Interview #9)

One interviewee spoke enthusiastically about the Summer Institute for Computational Social Scientists (sicss.io) as a formative learning experience, describing SICSS as an environment where they were able to meet other social scientists, talk about teaching, and learn about new methods.

When talking about the needs of instructors in their discipline, answers were varied, and ranged from the need for workshops or tutorials on how to teach with data; training on software packages; statistical methods; and how to develop students' critical thinking skills with regard to data. Another instructor described a need for an easily navigable online resource that aggregates many types of data-related resources, including data sets, software packages and tutorials, and other instructional modules. One interviewee thought it would be useful if new instructors had the opportunity to shadow someone else who has already been teaching with data. Another expressed interest in a local working group for social scientists teaching with data.

When questioned about course development, most of the interviewees reported at least looking for publicly available syllabi or data sets related to their course; only four responded that they had not re-used other faculty's syllabi in their course design. A couple referred to friends from graduate school or their current network of collaborators with whom they discuss ideas related to their syllabi, methods, or data sets. With regard to sharing their own syllabi, assignments or data sets, most of the instructors expressed willingness to do so. One interviewee stated they put their syllabi and data sets on a public web page with open licensing, while another had shared only within their own department. It seems that time may be a factor in why at least a couple instructors have not yet shared their course materials.

All thirteen instructors reported that they provided students with data sets to use for coursework. Most reported at least some amount of difficulty in finding data appropriate for their teaching. Challenges described by the interviewees included finding data in the best format for their needs, finding data with all desired variables, and finding current data rather than outdated data.

Having data in the best format for the topic being taught was a common challenge. An instructor whose teaching focuses on social networks noted that although the network data they needed was generally available, such datasets were often in non-standard formats that were difficult for students to use, making "getting them in a format that's easy for students to import […] the hardest part." An instructor working with GIS data shared a similar experience, saying that the challenge was finding data that was ready to be used by students who weren't already familiar with the software, not necessarily finding the data itself: "We know where the data is. It's just difficult getting it to manifest in the ways that we would like" (Interview #3).

Another common challenge involved finding data that served the particular needs of the course. For some, such as an instructor in Anthropology, finding a single dataset with all the necessary variables concerning the topic in question proved difficult. Similarly, instructors who required students to find data for their own work reported that students often assumed data for their particular projects would be readily available, when in reality the data either does not exist, does not precisely fit their needs, or needs to be transformed into other formats for the student's use case. And many instructors noted that while finding data itself was not difficult, the data was frequently outdated, obsolete, or not well-maintained, or that it was difficult to find enough data

to keep their teaching fresh in each iteration of the course. In many cases, these challenges could take up quite a bit of time and effort, as one instructor explained:

> Sometimes, I would say the biggest challenge, the one I encounter most frequently is just websites not keeping their data updated, whether their storage has run out or whatever it is. Because a lot of times when I'm doing sort of searches […] You know, I'll find something that looks promising only to realize that this is no longer being hosted. So I think that that is sort of the biggest time challenge in developing course material. Most people in the areas that I work in or I would say there's a general sort of push towards having open data. So usually it's not an issue of data, for example, being like behind a paywall or being siloed in somebody's personal repository. So I would say it's not so much a problem of the data not being accessible as it is sort of not being maintained. (Interview #13)

Instructors shared many methods for working through these challenges. Several reported the usefulness of textbooks in finding the best datasets for their ready availability and, for one instructor, in part because the data was "real." As that instructor explained, the data provided in their textbook was useful because:

> …it's like "real data" in a sense, right? So it's messy enough to kind of approximate the kind of data you actually see in the real world, right? I think sometimes fake data is so fake and so perfect that it ends up like being misleading to students, right? So I like that and so I don't have any complaints there. (Interview #8)

Other sources for data used in the classroom included the instructor's own research, research data shared at conferences or in publications, data used by other instructors either for the same course or for similar courses, and conversations with colleagues.

External data set named by instructors included the National Election Studies, the General Social Survey, Federal Judicial Data Center, data available through Purdue's membership in the Inter-university Consortium for Political and Social Research, Centers for Disease Control and Prevention, National Institute of Mental Health, the American Community Survey, GIS databases available through the Libraries, and publicly available data from online communities and social media.

When asked what cultural or policy changes they have observed at Purdue that influence the way they teach with data, there was a mixture of responses. Three instructors referred to the Advanced Methods at Purdue (AMAP) as significant to their teaching with data experience, with one interviewee who is actively involved in AMAP stating,

> We have a whole community that cares about understanding data and how to leverage different kinds of data and how to collaborate through methods

> across very diverse disciplines and really be able to use data as a common language that can help me talk to a political scientist or an anthropologist and understand what their perspectives have to offer for my own fields and my own research. And I just think that's super cool. (Interview 1)

Three instructors referred to the Integrative Date Science Initiative (IDSI) as a cultural change they have observed, with one mentioning their membership on the IDSI curriculum committee. Two instructors referred to the student learning community the Data Mine. One comment about Data Mine was neutral, while the other described the Data Mine as being too STEM-oriented, and not applied enough to be relevant for students who are interested in doing human subjects research.

Six instructors responded that they were unaware of any policy changes or other influences at the university level, with two stating they have been at Purdue for only a few years and therefore have not been here long enough to see these types of changes, even though initiatives like AMAP and IDSI have been active and publicized in the past few years. One instructor acknowledged that their department is very supportive of teaching with data, while another noted the importance the university administration places on teaching with data, especially "big data," but stated that they did not perceive any additional support trickling down to them at the departmental level.

THEME 3: MANAGING WIDE-RANGING STUDENT CONFIDENCE AND ABILITY !

Findings from our interviews suggest that teaching with data in the social sciences is a growing trend and that students' data literacy competencies are improving.

> I imagine…..I don't know for a fact…..but I imagine that out of all the different students I've got that there's probably a nontrivial number of them who are taking courses where there's some sort of data driven research involved. You know that certainly would be the case for the CS students and Engineering students. But even the Liberal Arts students that...I'd have to sit down and look at what my colleagues this semester are teaching but my guess is that at least a couple of them are teaching classes that have some sort of data analysis paper associated with the class. (Interview #2)

However, difficulty managing wide-ranging student confidence and abilities recurred as a key challenge for instructors teaching with data in the social sciences. An increasing number of students have data experience from other courses or participation in research projects that make them confident and competent data users. While others, especially liberal arts and social science majors, may have actively avoided math, coding, stats, etc. and have greater anxiety about their ability to work with data.

> …it's a requirement for our majors and there's a lot of students who are political science majors because they like to talk about politics and they like to read about politics. They don't want to do math and the first thing that they approach….the first thought they have when they come into the research methods class is oh my god this is a math class. So they….they're a little frightened by it, maybe. They're a little put off by having to take it because, again in their minds, they're not studying political science to figure out to estimate a regression line. (Interview #2)

> Students in CLA [College of Liberal Arts] are not necessarily used to working with data as far as the students that I've come across. If you take science students they are much more used to it so it can be challenging depending on the department that they are coming from. (Interview #4)

Ranges in student confidence and ability impact instructors differently and seem to be influenced by course level. Instructors in lower level courses report less of an impact, likely because they have low expectations of students' prerequisite knowledge and tend to focus on teaching basic data principles and/or specific software. Faculty in upper level courses expect a mastery of some basic competencies as a prerequisite--e.g. basic Excel competency, basic algebra, basic probability and statistics, basic data management--but students' mastery of these basic competencies ranges significantly.

> The other thing the students struggle with is I think we spend a fair amount of time reviewing just basic probability. Like this isn't a probability course. They don't need like advanced understanding of probability but all of the statistical models that we do are probabilistic and so they need, you know, if our conclusion is that we're 95% certain that this result is real they need to have some understanding of like what does 95% certain mean, right? And so we spend a fair amount of time just reviewing base….real base probability. And that trips up…..that's hard for the students, honestly. (Interview #8)

Instructors list several curricular strategies for managing wide-ranging student confidence and ability, including integrating activities that allow for hands-on practice, reassuring students that concepts will take time to learn, repeating and reinforcing concepts, and exuding a contagious enthusiasm. And, instructors also mentioned extracurricular and co-curricular strategies such as encouraging students to participate in data-centered learning communities, leading independent studies, and working with students on undergraduate research projects. Some faculty, especially those working on research projects or in labs with undergraduates, rely on graduate students or undergraduate proctors to teach basic data competencies. Other support mechanisms included one-on-one conversations and similar in-person support, general lab meetings, Q&A discussion in the learning management system (i.e. Blackboard or Brightspace), and email correspondences.

THEME 4 - TECHNICAL CONCERNS WHEN TEACHING WITH DATA !

In their interviews, many instructors highlighted technical issues that either pose challenges to teaching with data or that have to be taken into consideration and thus shape how they choose to teach with data. Two prominent issues were the problem of how to provide access to data used for students in their coursework, and logistical questions related to providing the software that students need to complete their work.

Most of the interviewees used Purdue's LMS, Brightspace (or previously Blackboard), to facilitate student access to the data used in their courses. In some cases, the instructors upload the data files themselves to the LMS and students can download them to their local computer. In cases where the data being used is freely available online, some instructors provided links to those websites and had the students download the data there, while others still uploaded the data to the LMS for direct access. Of those interviewed, only three provided data through other platforms: one created a course-specific Wiki and posted files and links there; another shared data through their own server; and one instructor used a cloud-based storage system, Box, to provide and manage access to data for her students' work.

In choosing how to provide students access to needed data, instructors balanced the desire to make access as easy as possible with the size and complexity of the data. On the one hand, using the LMS or a course-specific website provides a single point of access. To make things easier for students, for instance, one instructor uploaded data to the LMS for student access even though the data was available through their textbook because they "just assume" that having students go through the textbook's platform "is going to cause problems" (Interview #8). Even the decision of whether to provide links to data hosted elsewhere or upload the files directly to the LMS could be guided by whether data-providing websites were easy to navigate. As one instructor explained:

> Sometimes datasets are housed really nicely and are in a very convenient format and for those I would just link to the website with instructions of how to access the data. And so there's an example. There's a project about redlining in the United States that has historical maps and basically I linked the website and it's really nicely laid out where you can go down and choose which city you want to look at. So I had students pick which city that they wanted to do so they were instructed to choose the city that interests them. Other times if that's not as nice and intuitive, I will pull things out and link them directly in the……or load them directly into the LMS. (Interview #13)

At the same time, some instructors highlighted how the size and complexity of the data, along with what the instructor is trying to teach, is taken into account. Several interviewees noted that they sometimes simplified complicated datasets to make it easier for students to access and

import into their software. In other cases, where datasets are simply too large or complicated to share through the LMS, providing links to other sites that host the data is the best option. And one instructor used the General Social Survey's exploration tool and the Google Trends platform in their teaching, rather than files of raw data, in part because they made accessing and using the data easier for students with little technical experience.

One instructor whose courses center on GIS noted that providing students access to large data, such as image files needed for GIS work, could be a real challenge, one that they knew other instructors faced as well. Because some of this data is shared by many instructors on campus, they looked into using university IT resources to provide a centralized portal for providing student access to shared GIS data but financial concerns made it difficult to find a solution.

One interviewee shared a unique factor in their decisions about how data was stored and shared with students: confidentiality and access management. Because this instructor's students worked in the instructor's research lab with real health data, they store all the data on a Box account. In addition to offering remote access to the data, which allowed students to work with it anywhere, the instructor felt this solution offered greater control of who had access what data at any given time. This was also the only instructor to specifically require their students to sign a data use agreement that outlined clear instructions regarding how to access data and keep it secure.

Many instructors commented on how technical considerations such as student access to software and institutional IT support factored into their teaching with data. Several instructors noted that they chose to use particular software in their teaching because they were more accessible to students either in cost or in campus availability. One instructor, for instance, teaches Stata because Purdue distributes student licenses for the software for free through its central software hub, and it is also available on many lab computers on campus. Another instructor teaches with ArcGIS Online for similar reasons, as all students have access to the software for free through Purdue and it is supported by the university.

Although campus availability or support of software could encourage instructors to teach with certain tools, several instructors noted that this did not always go according to plan. One instructor teaches with SPSS because, although student licenses are expensive, the software is used in their research lab and students can use the software there. When COVID arrived, this proved more complicated than expected:

> In the past, that's worked fine because most of my students are in the lab a lot. They're in the lab a few hours a week. I think all of that has been made much harder with the current pandemic, right? Since we're all kind of working from home, working remotely I have not really come up with a great solution for that. So ideally they would have…..ideally there would be a way to get SPSS installed on their personal computer at home, right? Like a student license, or a temporary license or something like that but as far as I know Purdue doesn't have that and it

> costs, I think, a couple hundred bucks to buy SPSS. So that's prohibitively expensive for my students. (Interview #8)

And in another instance, an instructor who teaches with R noted that it is technically available on campus computers but did not always run as expected:

> …my original plan was to have students use university computers and we just had trouble. We had trouble installing the extra libraries in R, we had trouble just getting things to work as smoothly as we liked and so I switched to having students just use their own laptops. And it worked fine with the students that I have but it made me worry a little bit. I would have liked to be able to have at least as a back-up option that they could use university resources. (Interview #3)

LIMITATIONS

The results from this study describe the experiences of faculty at Purdue University and may not reflect the experiences of faculty at other universities. Additionally, the research team experienced some difficulty recruiting social sciences faculty to participate in the study. There may be many reasons for this, including the possibility that the stress of teaching and conducting research during the COVID-19 pandemic made some faculty less willing to participate in additional activities. Another factor is the disciplines represented by the interviewees. In defining the social sciences, ITHAKA S+R provided participants with a list of academic departments to target for recruiting interviewees. Ten of our interviewees taught within these departments. However, we went further afield than the list, and included instructors from the Honors College, Human Development & Family Studies, and Libraries & Information Studies, after confirming that their courses were aligned with the social sciences.

RECOMMENDATIONS

The interviews and analysis conducted for this study provide an informative snapshot of how and why Purdue University faculty in the social sciences currently teach with data, as well as some of the unique challenges posed by incorporating data into their instruction. Although many conclusions can be drawn from the study, one overarching result of these interviews may be the extent to which instructors approach teaching with data in a largely independent manner. This can be seen in three related observations.

First, although most interviewees agreed that teaching with data did raise unique challenges, nearly none of them received any formal instruction or support in teaching with data and learning to navigate those challenges. Instead, most come to teaching with data through a combination of

experience, such as serving as a teaching assistant in graduate school, and independent learning. Similarly, awareness or use of formal resources for instruction involving data, or of initiatives that might speak to such concerns such as AMAP, was relatively low. Instead, many of those interviewed relied on a personal network of colleagues or on departmental connections for support in tasks such as finding appropriate datasets or developing course syllabi, perhaps reflecting less engagement with resources beyond disciplinary borders. Finally, the instructors interviewed for this study primarily articulated their goals for teaching with data in individual, course-level terms that rarely connected their own teaching goals to larger frameworks such as departmental or college curricula, or institutional programs such as the Integrative Data Science Initiative. Yet despite being largely independent in their teaching with data, many had no trouble articulating tools and programming that might help them better incorporate data into their classroom, with a number of their ideas being distinctly collaborative in nature, such as platforms for shared resources or local working groups.

To turn the results of the study into practical local initiatives that better support teaching with data in the social sciences, we recommend:

- Campus-wide recognition that teaching with data in social science areas raises its own unique set of challenges;
- Continual assessment of the specific needs of instructors working with data to better understand and address those challenges; and
- Programming and resources that support instructors in their work.

The Libraries has been supporting campus data needs for several years, through the collections, services, and courses it offers. The results of this study can inform Libraries' services and programming in several ways. Specific recommendations include:

- Market data-related Libraries courses to students in the social sciences;
- Improve campus access to open and proprietary data-related resources;
- Work with CIE to address instructors' pedagogical needs; and
- Conduct a campus-wide needs assessment.

The Libraries is already actively involved in the campus-wide effort to increase undergraduate student data education through the data-related courses taught by Libraries' faculty. These efforts should be maintained, and Libraries instructors should make sure their courses are marketed to students in social sciences disciplines as well as STEM disciplines. The Libraries offers web-based finding aids (known as LibGuides) related to many subjects and courses across the university. The LibGuides have been authored by numerous Libraries faculty and staff, who created each LibGuide in response to a specific user need they were trying to address. This has resulted in many disparate and possibly overlapping LibGuides on data services, data sets, data management practices, software tutorials, etc., that the end user may find daunting to navigate.

Creating a central guide that indexes these data-related LibGuides could make them more accessible to both teachers and students, and it would be easier and more efficient to market one data LibGuide to the community. Doing so also would directly address the comments by interviewees who expressed needing help finding data sets and other resources for their courses.

While Libraries faculty and staff facilitate student learning through data-focused workshops and credit courses, these learning opportunities focus on *students* rather than instructors, and focus on developing specific data skills rather than on the pedagogical aspects of teaching with data. While taking on the teaching of teachers of data is not completely outside the domain of an academic library, in this case a better strategy might be to leverage Libraries's partnership with Purdue's Center for Instructional Excellence (CIE), especially through their long collaboration on the IMPACT course redesign project, to address instructors' pedagogical needs. One interviewee participated in the IMPACT Data Science Education, which was limited to a cohort of six instructors, and stated that it was the only real education in teaching with data they had received. The results of study imply that there is a real need for pedagogical support for instructors who teach with data, especially in the social sciences. To this end, CIE and the Libraries could conduct a campus-wide needs assessment for teaching with data, and plan additional programming as appropriate.

This study suggests that although many initiatives successfully promote data science and literacy across Purdue University, aside from IMPACT Data Science Education, this programming has largely neglected to address *teaching* with data as a practice that poses unique pedagogical challenges and thus require specialized attention. As data science, data literacy, and related topics continue to attract attention in higher education, units that work in these areas may provide a great service by helping to conceptualize teaching with data as a distinct practice in this way, and by offering opportunities for collaboration across disciplines that focus on sharing best practices and resources in this area.

APPENDIX 1. WEB LINKS USED THROUGHOUT THE REPORT

Advanced Methods at Purdue (AMAP): https://www.purdue.edu/amap/

AMAP Graduate Interdisciplinary Methods Certificate: https://www.purdue.edu/amap/curriculum/Certificate.php

AMAP Resources for Grants: https://www.purdue.edu/amap/Grants%20Resources.php

College of Agriculture's Data Driven Agriculture Minor: https://ag.purdue.edu/digital-ag-resources/purdue-offers-data-driven-agriculture-minor/

College of Agriculture's Spatial Data Science Certificate: https://online.purdue.edu/programs/agriculture/spatial-data-science-certificate

College of Engineering's course Engineering in the World of Data: https://www.purdue.edu/learningcommunities/profiles/engineering/engineering_data.html

College of Science Undergraduate Degree in Data Science: https://www.cs.purdue.edu/undergraduate/curriculum/data-science.html

College of Science Master's Degree in Geodata Science: https://www.eaps.purdue.edu/gdsp/

Critical Data Studies: https://www.purdue.edu/critical-data-studies/index.php

Data Management Toolkit: https://docs.lib.purdue.edu/dcptoolkit/

Data Mine residential learning community: https://datamine.purdue.edu/

Databib: http://databib.org

IDSI Education Ecosystem Website: (https://www.purdue.edu/data-science/education/index.php)

IMPACT Data Science Education: http://sites.lib.purdue.edu/dse/

Integrative Data Science Initiative (IDSI): https://www.purdue.edu/data-science/index.php

ITHAKA S+R: https://sr.ithaka.org

Krannert School of Management's Master's Degree in Business Analytics and Information Management: https://polytechnic.purdue.edu/degrees/data-visualization

Libraries Graduate Certificate in Geospatial Information Science: https://blogs.lib.purdue.edu/news/2021/05/20/introducing-the-new-graduate-certificate-in-geospatial-information-science-gis-from-purdue-libraries-and-school-of-information-studies/

Purdue University e-Pubs: https://docs.lib.purdue.edu/

Purdue University Research Repository (PURR): https://purr.purdue.edu/

Purdue Polytechnic Institute's Bachelor's Degree in Data Visualization: https://polytechnic.purdue.edu/degrees/data-visualization

APPENDIX 2. RECRUITMENT LETTER

Dear ,

Researchers from Purdue Libraries and School of Information Studies are conducting a study titled, "Teaching with Data in the Social Sciences" (IRB protocol # 2020-818), on the practices of social science instructors who teach undergraduate students to use quantitative data, such as by conducting research using quantitative methods, analyzing or visualizing datasets, or learning to use specific tools or software to work with data. The purpose of the study is to inform the improvement of Libraries' support services for teaching with data.

Would you be willing to participate in a one-hour interview to share your unique experiences and perspective?

Our local study at Purdue is part of a suite of parallel studies at 23 other institutions of higher education in the US, coordinated by Ithaka S+R, a not-for-profit research and consulting service. The information gathered at Purdue will also be included in a landmark capstone report by Ithaka S+R and will be used for the Libraries and School of Information Studies to further understand how the support needs of social science instructors are evolving more broadly.

If you have any questions about the study, please don't hesitate to contact the project's Principal Investigator, Clarence Maybee by email at cmaybee@purdue.edu or by phone at 765-494-7603.

Thank you so much for your consideration.

Sincerely,

Clarence Maybee

APPENDIX 3. INTERVIEW QUESTIONS

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?