STUDYING INTERDISCIPLINARY THINKING ABOUT COMPLEX REAL-WORLD DATA AT DATAFEST

Traci Higgins, Jessica M. Karch, and James K.L. Hammerman
TERC, Cambridge, MA 02140
Traci_Higgins@terc.edu

In the 21st century with the rise of computing power, it has become increasingly important to create opportunities for students to learn to work with large, authentic, complex (LAC) data across multiple disciplines. DataFest, a hackathon style undergraduate event, creates a space for such inquiry due to the collaborative, data-driven, open-problem, real-world relevant nature of the challenge it presents. We present preliminary findings from research that explores how teams at DataFest leverage and integrate multidisciplinary tools and domain knowledge to engage productively with the data investigation process. Implications for statistics and data science education are discussed.

INTRODUCTION

The data revolution is upon us, impacting every facet of human activity (Bargagliotti et al, 2020; Biehler et al, 2022; Murtagh & Devlin, 2018; Ridgway, 2016). At the heart of the data revolution is the increasing importance of being able to work with large, authentic, complex (LAC) datasets (Engel, 2017; Erickson, 2022). Across many domains there is a need to make sense of and derive meaningful, actionable insight from LAC datasets. Course work within many academic fields can provide opportunities for students to develop tools that support work with data, but creating opportunities for students to experience being awash in data (Erickson, 2022), develop productive questions to ask with and of the data (Arnold & Franklin, 2021), and translate findings into meaningful and relevant action can take a backseat in the classroom to developing technical skills. This paper explores how teams of students with multidisciplinary knowledge, different skill sets, and diverse experiences draw on these in an interdisciplinary way as they navigate being awash in data, generating a productive framing of the problem, and creating a storyline that communicates the real-world relevance of their findings or product. We center our investigation on DataFest, a hackathon-style undergraduate event that is an ideal space for such an inquiry because of the collaborative, data-driven, open-problem, real-world relevant nature of the challenge it presents.

Modern data science has emerged as a field in response to the challenges of working with and deriving insights from LAC data (Biehler, et al, 2022; Gould, 2017; Hardin et al, 2015). Data science is widely described as an interdisciplinary field that pulls from computer science, statistics and mathematics, and disciplinary domains directly relevant to a given dataset. Engel (2017) endorses this view and explicitly notes the importance of skills and techniques used in data mining, visualizing data, coding, and communication. The interdisciplinary nature of data science is captured repeatedly throughout the literature using various versions of a Venn diagram with three overlapping circles roughly mapping out these different disciplinary competencies, with data science situated in the overlap between them (Conway, 2010; Engel, 2017; Lee et al, 2022).

Erickson (2022) suggests that many would describe "data science as living in the untamed frontierland between statistics and computer science." Although this captures something important about the field, he suggests that defining the field may be less important than recognizing ways of engaging with data that are indicative of *doing* data science: the experience of being "awash in data," using "data moves" (Erickson, et al, 2019) to tame the data, and needing to communicate about data, especially using visualizations to share insights. Erickson's work highlights aspects of data science that come to the fore when moving beyond the traditional statistics curriculum and embracing the complexities of working with LAC data.

Lee and colleagues (2022) have examined the key practices and processes professional data scientists engage in when carrying out an investigation. They identify six distinct components of the data investigation process (DIP): Frame the Problem, Consider and Gather Data, Process Data, Explore and Visualize Data, Consider Models, and Communicate and Propose Action. Although this process is easiest to conceptualize as linear and cyclical, professionals move through these phases in various sequences and in a fluid way. For example, when working with a given set of LAC data, the data scientist may begin by conducting exploratory data analysis and using simple visualizations to

examine the characteristics of the data at hand. Based on what is noticed during preliminary analyses, the data scientist might then shift to considering the data, gathering additional sources of data, processing the data to prepare to conduct additional analyses, begin framing the problem, or engaging in further exploration and visualization. Alternatively, when given very messy datasets, the data scientist might go back and forth between considering the data and processing the data before conducting any exploratory analysis. Or, if the data scientist has been given a directive or has strong content knowledge for understanding the context of the data, he or she may begin by considering ways to frame the problem, working to articulate a good investigative question.

The question remains open as to how students leverage different disciplinary skills and domain relevant knowledge to cope with being awash in the data and to navigate through the DIP. As LAC data become ubiquitous beyond statistics, and data literacy becomes a central skill across disciplines (e.g., Schultheis & Kjelvik, 2020), it is increasingly important to understand how domain knowledge and statistical and computational thinking intersect. Although data scientists may not necessarily be domain experts, domain experts must now be able to navigate using computational tools to make sense of and think statistically about LAC– i.e., engage in data science practices. Moreover, professional data scientists work within a collaborative setting. This raises the question of if and how teams of students with diverse disciplinary skills and background knowledge find ways to integrate them in ways that help them navigate the DIP when working collaboratively.

THEORETICAL FRAMING AND RESEARCH QUESTION

DataFest, as an event that may invite students from a range of disciplines and backgrounds to collaboratively make sense of LAC data, provides a fertile ground to explore some of these complexities. To investigate the role domain thinking plays in data investigations, we present an exploratory analysis that focuses on what kinds of knowledge and skills students bring to bear in these investigations and how they are leveraged and integrated into their processes. We draw on multiple theoretical perspectives to help elucidate these processes. First, we conceptualize "domain thinking" broadly as "tools", encompassing formal pieces of domain knowledge (e.g., concepts), and ways of thinking (e.g., methodologies or epistemologies) that students may draw on that are not formally computational or statistical in nature. Second, we conceptualize the process of utilizing these tools as potentially multi- or interdisciplinary in nature. By this, we mean that for students to use these tools in their data investigations, they may apply domain knowledge and make sense of it parallel to their computational thinking, i.e., do multidisciplinary work, or they may have to negotiate different norms, epistemologies, and methodologies in order to integrate them and create something new, i.e., do interdisciplinary work (Collin, 2009; Tripp & Shortlidge, 2019). Finally, we use the DIP as a lens to operationalize the data investigation in which the students are engaged (Lee, et al., 2022). The overarching research question guiding our exploration is: How do students leverage and integrate disciplinary and domain knowledge as tools while working with LAC data at DataFest?

METHODS

We collected multiple streams of data at six DataFest (DF) sites in 2023, including pre-event and post-event surveys, retrospective interviews, and artifacts (e.g., copies of teams' final presentations). DF is a weekend-long cocurricular data science competition sponsored by the American Statistical Association (ASA) which takes place at over 50 sites, with over 2000 participating students. Each year the ASA seeks a sponsor from industry or the public sector to provide LAC data that has real-world relevance as well as special significance to the sponsoring industry or organization that can be communicated to students. Students sign up as teams or as individuals to be placed in teams of up to five students. On Friday evening, the datasets are revealed. Teams of students have until roughly midday on Sunday to extract meaning and insight from the LAC dataset with the goal of addressing an open-ended real-world problem (Gould, 2014). On Sunday afternoon, the teams present their work in very short presentations. Teams vie for several awards determined by a panel of judges, such as Best Use of Outside Data. In Spring 2023, teams were tasked with finding insights to improve the effectiveness of the American Bar Association's (ABA) online platform that is designed to provide pro bono answers to low-income users' questions within the realm of civil law. The problem space is purposefully vague, forcing teams to define the problem and generate their own statistical questions. The data sets were messy and contained different types of

information, e.g., demographic and time data. One especially messy and complex dataset included text from posts to the platform.

Our findings are based on a preliminary analysis of the retrospective interviews. Teams were invited based on their post-survey responses, where they answered questions pertaining to their DF experience, including the types of skills and knowledge they used (e.g., knowledge from computer science, knowledge from life outside of school, etc.) and which were especially helpful, how successful they thought their team was, and whether they experienced the event as supportive and welcoming, as well as demographic information. Respondents were also invited to elect whether they would be interested in participating in a retrospective interview. From those who expressed interest, we invited 2-4 teams per site to be interviewed and offered each interviewee a $20 gift card as a token of appreciation. This was all done with IRB oversight and approval. In selecting teams, we prioritized those who demonstrated potential for interdisciplinarity, e.g., by having a range of formal academic majors or lived experiences or who indicated that their team members differed in the knowledge and skills they brought to the challenge. These teams were selected in order to collect rich data that would help us better understand how various tools impacted students' approach to the DF challenge.

These interviews typically lasted 60 to 70 minutes and included questions about the coursework and background knowledge teammates brought to their work, the timeline of their work, how they engaged with each phase of the investigative process, the focus of their presentation, and how they made decisions about what to include in their presentation. In addition to the streams of data described above, we also observed the Sunday activities and recorded presentations at three sites, and at one site researchers sat with two teams through the entire data investigation process taking field notes and recording interactions. These additional observations provided a framework for understanding and contextualizing the interview data.

FINDINGS & DISCUSSION

Our research is in a preliminary stage, but we will present two vignettes that illustrate our approach to examining how multidisciplinary tools were leveraged by teams to cope with being awash in the data and make progress within different phases of the DIP. To shape these vignettes, we will describe (1) the primary tools the students drew on within their team, and (2) how integrating these tools and ways of grappling with the data shaped the direction of the group's data investigation. The purpose of these vignettes is to begin to explore how interdisciplinary reasoning with domain and other forms of knowledge impact students' reasoning with LAC data throughout the data investigation process, in particular the framing of the problem phase of the DIP (Lee et al., 2022). We will occasionally demarcate our analysis with italics, to indicate when the students are engaging in practices that may support interdisciplinary work (Tripp & Shortlidge, 2019).

Vignette 1: Finding "Whales" in the Data

Team Whale (pseudonym) had a breadth of disciplinary knowledge that proved especially helpful given the context of the challenge and the nature of the LAC datasets they were to work with. While everyone on the team majored in computer science, several students were double majors, in both STEM (math and statistics) and non-STEM (e.g., English) fields. One team member even had specialized expertise in law. Right after the data for the challenge were revealed, the team immediately began considering the nature of the data, in particular the vast amount of text data. Considering the data led them to identify relevant tools that they could use for analyzing text data and they turned toward the forms of language processing they were familiar with, e.g., sentiment analysis, and considered what they could make out of that. Their initial exploration of the data was largely driven by analytical and data-centric concerns, e.g., "What kinds of data do we have? What statistical tools do we have that can do something interesting with these kinds of data?" This led them to do topic modeling, a text analysis technique to identify word clusters, in an early exploratory phase.

A breakthrough in their investigation occurred when the team was faced with making sense of an anomalous finding in the topic model, specifically a strange semantic cluster. To figure this out, they shifted from a purely machine learning approach, to drawing on multidisciplinary and *different research methods* (Tripp & Shortlidge, 2019) to identify insights that they could *integrate*. One team member, who was an English major, was "very good at digging out the stories, the narratives, and finding patterns." Drawing on this way of knowing and making sense of data, she pulled up the CSV

file and started reading conversations and realized that some lawyers sign off their posts in very distinctive ways. At the same time, other team members started doing statistical analyses to figure out how many hours each lawyer contributed to the site from state to state and found that states with a high number of average hours per attorney were skewed by outliers–lawyers who answered an exceptionally large number of questions. By *collaborating across disciplinary perspectives* to *integrate* these two observations, the team conceptualized these disproportionate contributors as "whales", a term from the mobile gaming sphere where the majority of the game's income comes from a small subset of players who spend a large amount of money. The concept of a whale turned out to be incredibly productive, because they could reframe their problem from a broad exploration of text to a targeted investigation of how whales impacted the ways client questions were answered across the site.

Through our theoretical lens, we can see that leveraging other ways of knowing served as a way for the students to find meaning in an anomalous piece of data they identified as salient, but confusing. By taking different methodological approaches grounded in different epistemologies, i.e., the "English" approach of reading the data and interrogating what stories the data told, and the statistical approach of asking how pro bono hours were distributed across attorneys across states, they were able to gain insight into the anomaly, and integrate through the concept of the "whale," a concept pulled from their background knowledge as gamers. In this way, disciplinary tools and lived experiences were integrated to gain traction within the DIP and develop a meaningful storyline.

Vignette #2: Becoming Data Scientists

Team Computers (pseudonym) entered the competition with similar backgrounds and strong computational grounding, but limited experience with LAC data or the sort of open-ended problem space presented by the DF challenge. In the group of four, three were computer scientists and one was studying data science within an engineering department, but still quite early in his coursework. Their prior experience with data mainly consisted of being given data and told what to find in a class setting. Although they started with considering and familiarizing themselves with the data, they were uncomfortable working in the open problem space and quickly shifted into framing the problem phase, before exploring the data further. This first framing was based on background knowledge about the needs of prisoners that one team member had developed when researching and writing a paper. However, as they got deeper into the data investigation, they found that the data was not well suited to their question, and they decided to abandon the approach.

Team Computers was now well into Saturday and still had no focus. They leveraged their data skills to get them unstuck and began using "the data to guide [their] direction," searching for trends within a data exploration phase. During this phase, they noticed that the Family and Children category had the most posts, creating graphs to explore demographic characteristics and scanning through text data to get a sense of clients' questions. They found that many clients were low income, female, single or divorced. As they grappled with how to connect these findings to something meaningful that would help them establish a storyline with real-world relevance, rather than presenting "just random graphs here and there," one member took up the task of looking for outside data. He wondered whether there had been an increase in single mother households below the poverty line and found evidence of this in the Federal Reserve Economic Database. These external data drove additional explorations of the ABA data and helped them develop a storyline about single mother households. Although they let the data drive their investigation (Higgins et al., 2021), they recognized the need to find *domain grounding* outside of the computational sciences to give meaning and relevance to their findings and that could support proposed action.

Through our theoretical lens, we see that they used domain knowledge and the data themselves to reciprocally shape and constrain their data investigation as they framed and reframed the problem. As they navigated the problem space, they experienced a tension between what the data could show them, and what connections they needed to create meaning and real-world relevance for their work. Engaging with this tension required them to embrace framing the problem, something they had little experience with. According to the team, in computer science, "you're usually given the problem and we need to find the solution." They had strong experience being creative in applying their skills to solve a problem, but not in creating the problem space themselves. At DataFest, the problem was not predefined, however, they came to realize, it was constrained by characteristics of the

data at hand. Awash in the data and searching for a foothold to give focus to their investigation, they sought connection to real-world topics and relevant background knowledge. By seeking a way to integrate domain knowledge, statistical and critical thinking about the data, and their strong computational skills, they began to engage in the interdisciplinary activity of doing data science, working within the overlap of the data science Venn Diagram (Conway, 2010).

CONCLUSION

This preliminary analysis explores how teams at DF leverage domain knowledge and experiences in different ways to approach the DF task and work to integrate these forms of knowledge to make sense of and gain insight from LAC data. We found that these tools can be used to provide the students entry points to grapple with thorny issues in the data investigative process. Both vignettes were fundamentally about the teams trying to frame the problem and searching for domain knowledge to create a relevant storyline. However, the two teams also varied in the range of disciplinary and domain knowledge available within the team and how they made use of this knowledge. To unpack their anomaly, Team Whale engaged in a deeply interdisciplinary process to support their specific data investigation, drawing on different epistemologies and methods in order to advance through integration by applying the concept of whale to their problem. These interdisciplinary methods were a core part of their data investigation. Team Computers, on the other hand, had less disciplinary diversity to draw on, and struggled to identify domain knowledge they could *integrate* as they worked through the DIP. This became especially salient when they worked on framing the problem and communicating and proposing action. However, they moved toward interdisciplinary practices as they leveraged their real-world knowledge and sought out other resources to learn more about the domain. In sum, Team Whale consisted of students already comfortable working within the overlap of the data science Venn Diagram, who integrated domain knowledge as an additional tool in their investigation, whereas Team Computers consisted of students with limited experience with open-ended data, who were just beginning to find ways to cross between their computational skills, domain thinking, and statistical thinking in order to develop the interdisciplinary practice of data science.

In our ongoing work, we are trying to unpack the roles disciplinary and domain knowledge play in different phases of the DIP for groups with various [inter]disciplinary compositions. A limitation of this early stage is we have only just begun to explore these complexities. Team Whale, an interdisciplinary group with data science experience, drew on both informal disciplinary epistemologies (the "English" approach) and shared cultural touchpoints (the concept of whale), while Team Computers, a more homogenous group with less data science experience drew on more formal disciplinary knowledge in the form of prior knowledge and skills. Our ongoing work will grapple with the roles and origins of these different tools and how they impact the data investigative process.

We argue that attending to these ways of knowing and practicing data science is vital in deepening our understanding of how to engage students in LAC data for three reasons. First, although working with LAC data is an increasingly important part of statistics education (Engel, 2017; Ridgway, 2016; Schultheis & Kjelvik, 2020), it can often be daunting for lower-level students, i.e., those with less formal statistics experience, to engage in data competitions where they can gain practical experience working with LAC data (Dalzell & Evans, 2023). Identifying and emphasizing how non-statistical knowledge can serve as an asset in making sense of LAC data can be used to make data competitions like DataFest more accessible and open to students from a broad range of backgrounds and levels of experience. Second, developing data literacy with LAC data is becoming more and more important in many disciplines other than statistics and data science, ranging from STEM disciplines like biology (e.g., Schultheis & Kjelvik, 2020) to the emerging field of digital humanities (Sun, Yu, and Tian, 2022). Developing an understanding of how students with different disciplinary orientations engage with LAC data is vital to expand insights from statistics education to other fields where that information is important. Finally, our work contributes to furthering a fundamental understanding of the role domain knowledge plays in working with LAC data, an aspect of working with real world data that is currently understudied in data science education.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2216023. We are grateful to the DataFest organizers who graciously invited us to collect data at their sites, and to Teams Whale and Computers (among others) for sharing their processes with us.

REFERENCES

Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education*, *29*(1), 122–130.

Bargagliotti, A., Binder, W., Blakesley, L., Eusufzai, Z., Fitzpatrick, B., Ford, M., Huchting, K., Larson, S., Miric, N., Rovetti, R., Seal, K., & Zachariah, T. (2020). Undergraduate learning outcomes for achieving data acumen. *Journal of Statistics Education*, *28*(2), 197–211.

Biehler, R., Veaux, R. D., Engel, J., Kazak, S., & Frischemeier, D. (2022). Editorial: Research on data science education. *Statistics Education Research Journal*, *21*(2), 1-4.

Collin, A. (2009). Multidisciplinary, interdisciplinary, and transdisciplinary collaboration: Implications for vocational psychology. *Int. J. Educ. Vocat. Guidance*, *9*, 101-110.

Conway, D. (2010). The data science Venn diagram. URL Http://Drewconway. Com/Zia/2013/3/26/the-Data-Science-Venn-Diagram.

Dalzell, N.M. & Evans, C. (2023). Increasing student access to and readiness for statistical competitions. *Journal of Statistics and Data Science Education*, 1-6.

Engel, J. (2017). Statistical literacy for active citizenships: A call for data science education. *Statistics Education Research Journal, 16*(1), 44-49.

Erickson, T., Wilkerson, M., Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education*, *12*(1), 1-24.

Erickson, T. (2022, July). *Awash in data*. Common Online Data Analysis Platform (CODAP). https://codap.xyz/awash/

Gould, R. (2014). Datafest: Celebrating data in the data deluge. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (pp. 1-4). International Association for Statistical Education (IASE).

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22-25.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., & Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "think with data." *The American Statistician*, *69*(4), 343–353.

Higgins, T., Mokros, J., Rubin, A., Sagrans, J., & Ren, A. (2021). When the data drive the learning. In R. Helenius & E. Falck (Eds.), *Statistics Education in the Era of Data Science: Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE),* Aug. 30-Sept.4, hosted online. https://www.doi.org/10.52041/iase.hmtse

Lee, H., Mojica, G., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*, *21*(2), 1-23

Murtagh, F., & Devlin, K. (2018). The development of data science: Implications for education, employment, research, and the data revolution for sustainable development. *Big Data and Cognitive Computing, 2*(2), 1-16.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, *84*(3), 528–549.

Schultheis, E. H., & Kjelvik, M. K. (2020). Using messy, authentic data to promote data literacy & reveal the nature of science. *The American Biology Teacher*, *82*(7), 439–446.

Sun, L., Yu, J., & Tian, J. (2022). Practice of social science digital humanities education based on big data analysis technology. In Z. Chan, D. Zhou, & H. Wu (Eds.), *Proceedings of the 2022 2nd International Conference on Education, Information Management and Service Science* (p. 514-522). Atlantic Press.

Tripp, B., & Shortlidge, E. E. (2019). A framework to guide undergraduate education in interdisciplinary science. *CBE—Life Sciences Education*, *18*(2), 1-12.