MAKING DATA SCIENCE PRACTICES EXPLICIT IN A DATA INVESTIGATION PROCESS: A FRAMEWORK TO GUIDE REASONING ABOUT DATA

Gemma F. Mojica, Hollylynne S. Lee, Emily Thrasher, Zack Vaskalis and Greg C. Ray
NC State University
gmmojica@ncsu.edu

We propose a six-phase framework that identifies key practices, processes and dispositions of investigating data, building on the work of statistics education to include data science. We share overarching dispositions and considerations for a data investigator and unpack key considerations to frame "ways of thinking" about the practices and processes in three of the six phases.

INTRODUCTION

Data are everywhere. Societal demands require that individuals are able to make sense of data presented in media, and many careers and jobs are increasingly requiring skills with data. But what does it mean to think and do investigations like a data scientist? In this paper, we build on the work of statistics educators and researchers by expanding it to make fundamental practices, processes and dispositions from data science explicit. Like the work of Wild and Pfannkuch (1999) that examined the practices of statisticians, this work is built on examining the work and practices of data scientists to identify key practices, processes and dispositions to guide data investigations.

REVIEW OF LITERATURE

Key Practices and Processes of Statistics

Statistics educators and researchers have long put forward frameworks to describe foundational practices and processes of investigating data. Many involve a four-phase cycle for solving a statistical problem or engaging in a statistical investigation (e.g., Bargagliotti et al., 2020; Franklin et al., 2007; Friel, et al., 2006; Graham, 1987): Pose a question, Collect or consider data, Analyze data, and Interpret results. Others propose similar five-phase cycles that emphasize the importance of planning and exploring data (Watson et al., 2018; Wild and Pfannkuch, 1999). A key difference between these frameworks is that Watson et al. (2018) separates the Analyze phase into two phases, Data Representation and Data Reduction, highlighting the importance of data visualization.

These frameworks also identify other key aspects of investigating data, such as context, attention to variability, uncertainty, informal inference, and data as a distribution (Bargagliotti et al., 2020; Franklin et al., 2007; Friel et al., 2006; Lee & Tran, 2015; Wild & Pfannkuch, 1999). Lee and Tran (2015) identify other key statistical habits of mind that are essential for conducting a productive data investigation process: ensuring best measures of an attribute, attending to sampling issues, and using multiple visual and numerical representations to make sense of data. Additionally, some emphasize the value in being a skeptic (Lee & Tran, 2015; Wild & Pfannkuch, 1999). It is also important to be curious, creative, intuitive, persistent and resilient (Wild & Pfannkuch, 1999), and to communicate and collaborate (IDSSP, 2019; Wild & Pfannkuch, 1999).

Key Practices and Processes of Data Science

As a discipline, data science is more nascent than statistics, where practices and processes are evolving (Cao, 2017; Donoho, 2017). Recently, those that work in data science education (e.g., Gould et al., 2016) have begun to propose frameworks or models that are similar or build on the work in statistics education. Processes used by data scientists often involve six or seven-phases which begin with understanding and/or defining a problem and context (Agarwal, 2018; EDC, 2014; Goldstein, 2017). This includes identifying central goals and attributes (Agarwal, 2018). The next phase involves gathering, cleaning, transforming and/or managing data. Goldstein (2017) characterizes this as collecting raw data and processing it, whereas others describe wrangling data which requires data collection and cleaning (EDC, 2014). In the next phase, there is an emphasis on understanding data through exploration (Agarwal, 2018; Goldstein, 2017) and/or analyzing data (EDC, 2014; 2016). Finally, the last phase involves finalizing a project (EDC, 2014) and communicating findings (EDC, 2014; Goldstein, 2017) which may involve visualization of data (Agarwal, 2018).

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-Sept 2021, Online conference. ©2021 ISI/IASE iase-web.org/Conference_Proceedings.php

EMPIRICAL AND THEORETICAL GROUNDING

A phenomenological study of the work of data scientists was conducted by the second author. In 2018-2019, H. S. Lee was immersed in the everyday work of data scientists for 9 months. To understand the daily work of data scientists, she attended meetings, presentations, and engaged in informal discussions with a diverse group of data scientists, taking field notes about what she was observing and learning. Lee worked closely with a data scientist to design an interview protocol, and she conducted interviews with five data scientists and collected seven additional interviews from publicly posted interviews with data scientists. Participants were eleven males and one female from various companies (e.g., SAS, RTI International, Pivotal Data Labs, Insights Association, Home Depot). Data included field notes, interview transcripts, and analytic notes that were open coded to identify emergent themes related to key practices, processes and dispositions of data scientists' work. Themes that emerged from the phenomenological study about critical aspects of the work of data scientists are described in Table 1. More details about this study are in Lee et al. (accepted).

Table 1. Critical aspects of the work of data scientists

Critical Aspects	Main Findings
Role of context/phenomenon	Data scientists emphasized that their work is nonlinear and approached holistically and always situated within a larger phenomenon. Context is not just vital to posing an investigative question and then used at the end to interpret and explain results; understanding and making sense of the context is woven throughout the entire process.
Immersion in data	Data scientists described their job as being immersed in data or even being overwhelmed by data. Much time is spent searching for data, making data useful and wrangling data, where they rely on various data sources.
Communication	Communication is a key skill for data scientists, especially among team members and clients. Data storytelling and visualizing is one of the most important ways they communicate with partners, clients and other stake holders.
Skepticism/flexibility	Data scientists explain that they approach tasks with flexibility, curiosity and skepticism.
Persistence/resilience	Common personality traits that were used to describe data scientists highlighted the importance of being persistent and resilient. They pointed out that projects often span months of work, and they "chip away" at a problem.
Broad Toolkit	Data scientists use a wide variety of technology tools to support all aspects of their work. They are often learning new software applications and techniques to apply immediately to their work.

In addition to the empirical research study, we examined the literature to identify key practices, processes and dispositions proposed and discussed by statistics educators and researchers. and those highlighted by professionals who work with big data. We attended to places where overlap exists among practices and processes and those that are novel or nuanced. Based on themes that emerged from the study and our examination of the literature, we created an initial framework, a Data Investigation Process, to identify key practices, processes and dispositions of investigating data. We further described key considerations through the Data Investigation Process, to expound upon the practices, processes and dispositions of each of the six phases. Content validity of this newly proposed framework was established through feedback from various experts. These experts included statistics educators, data scientists, mathematics and science teacher educators, middle and high school mathematics and statistics teachers, and a data software developer. We revised our framework based on feedback from experts.

FRAMEWORK FOR DATA INVESTIGATION PROCESS

We propose a six-phase Data Investigation Process that encompasses the following six phases (see Figure 1): Frame the Problem, Consider & Gather Data, Process Data, Explore and Visualize Data, Consider Models, and Communicate & Propose Action. For an in-depth explication of this framework, see Lee et al. (accepted). Like a puzzle, the phases fit together emphasizing that engaging in a productive data investigation involves revisiting and refining work within phases and making connections among phases. This constant movement back and forth and amongst phases is often dynamic in nature, although it can be linear and cyclic. While the investigator may enter at any phase, it is essential that solving a problem or answering a question within a context using data is at the heart of the process.

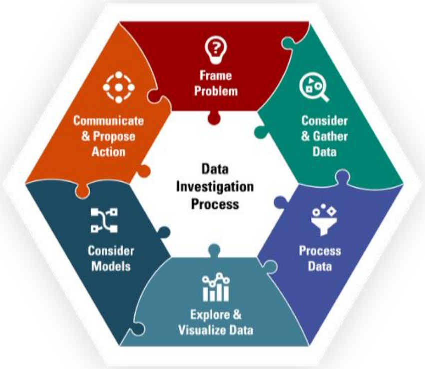

Figure 1. A Data Investigation Process

THINKING THROUGH A DATA INVESTIGATION PROCESS

Further, we identified key considerations for each phase in our Data Investigation Process, unpacking the practices and processes to guide thinking and actions for engaging in a data investigation, where the goal is to answer a statistical question within a context to communicate approaches and solutions using data-based evidence. We first highlight important ideas that should be considered throughout the ***entire*** investigation that are pertinent to all phases. Additionally, we identify dispositions that lead to productive data investigations. The following key considerations and dispositions are relevant to all phases in a data investigation:

- How are you making sense of data with respect to real-world phenomena/context and investigative questions (i.e., engaging in interpretation throughout the process)?
- What is the role of technology? How can it be best used to facilitate your work?
- Are you attending to variability in data and uncertainty in models and claims?
- What biases may be in your data and what biases, experiences, or perspectives do you bring to an investigation that could enhance or negatively impact your work within a context? What privacy or ethical issues need to be considered?
- Do you need to seek out additional expertise or find information about the context to inform your work and interpretations?
- Are your approaches based on being curious, creative and intuitive?
- Are you being skeptical as you examine data and the claims and actions that can be proposed?
- How are you communicating and collaborating with team members and/or clients or stakeholders?
- Are you being persistent and resilient in your problem solving when you need to overcome difficult obstacles in an investigation?

While our full framework includes key considerations for each of the six phases of a Data Investigation Process, our focus in this paper is to highlight aspects of data science that may not be explicit in other frameworks from statistics education, or may be altogether missing. Therefore, we will delineate the practices and processes for the phases Frame the Problem, Process Data, and Communicate and Propose Action.

Frame the Problem Phase

Almost all frameworks in statistics education start with posing or asking an investigative question that is statistical rather than mathematical in nature (e.g., Graham, 1987; Franklin et al., 2007; Wild & Pfannkuch, 1999), where context is emphasized. While many advocate for engaging learners in investigations that involve real or authentic data (e.g., Ben-Zvi et al., 2018), data scientists' work is based in solving real-world problems. Thus, real-world phenomena and broader issues related to a problem should also be a focus of work in this phase. When ***considering the context of the problem***, it is important to attend to the following:

- What is the context?
- What is the issue of interest within this context?
- What background information is needed? What resources are available to better understand the context of the problem?
- What is the broader purpose of the investigation? Why is this problem important to consider?
- What kind of data is available in this discipline/context?

When ***posing an investigative question***, the following should be taken into account:

- What statistical questions are you addressing?
- Is the statistical question appropriate for the context of the problem?
- Do these questions anticipate variability?
- Will the question(s) lead to a productive investigation?
- What strategies could potentially be used to answer the question(s)? What types of data are needed to use these strategies?
- What model assumptions should be considered about underlying populations or processes related to the context?

Process Data Phase

While many statisticians would describe the work of processing data as an important part of their work, these practices and processes have not been historically heavily emphasized in frameworks from statistics education as compared to frameworks from data science. One of the most important aspects of the work of a data scientist involves processing data (Agarwal, 2018; EDC 2016), and they spend much of their time engaged in these activities. In the Process Data phase, the investigator should ***consider strategies for processing and structuring data***, including:

- What strategies or techniques are most useful for obtaining or sourcing data?
- Where and how will data be stored and protected?
- Are there any issues with the ways data were entered? What will you do about possible erroneous/invalid data entries?
- What decisions will be made about missing data?
- What strategies or techniques will help process (e.g., clean messy data, organize, transform, etc.) and structure data in a consistent and usable format? Which are the most efficient? Which are the easiest to use? Do you have the necessary skills and resources to carry these out?
- Should you merge multiple data tables or other structures?

Additionally, the investigator should ***consider processes that may help focus the investigation*** by considering:

- Do new cases need to be added to the data set?
- Is it helpful to sort, group or filter the data?
- Is it useful to create new variables based on the available variables?
- Do measurement units need to change?
- Is it useful to recode data values (e.g., change No/Yes to 0/1 to easily sum 1's)?

Communicate and Propose Action Phase

Communication and the ability to communicate with others is an important aspect of the work of a both a statistician and data scientist. While some frameworks from statistics education emphasize communicating results, data scientists draw attention to the importance of communicating evidence-based claims to solve real-world problems and propose actions. In the Communicate and Propose Action phase, the investigator should ***devise a strategy for communication*** by considering:

- What are the important issues within the problem context that stakeholders are interested in?
- Who is the audience? What information do they need to inform their decision-making?
- What are the best formats, media, and language for communicating findings and suggested actions?

Further, the investigator should d***evelop and support an argument for claims and proposed actions*** by considering:

- How should you convey the problem, investigative question(s), methods, and analysis?
- Is it appropriate to discuss alternative approaches, models, or past results?
- What claims can be made from the data? What evidence is there to support these claims?
- What data visualizations could best support the claims? How are these visualizations interpreted? Do these visualizations need to be enhanced to be clearer to the audience?
- What statistical measures could best support the claims? How are these measures interpreted?
- Has uncertainty in findings been conveyed?
- What are the limitations, constraints, and potential biases of your data or analysis?
- What proposed actions within the context of the problem follow from the data investigation?
- What further data investigations should be recommended?
- Does the data story convey insights about the problem to your audience?

DISCUSSION

Our work is clearly built on the work of others in the field of statistics and data science. Many key practices, processes and dispositions of investigating data that we draw attention to are identified in this work. However, we believe that our framework makes some of the practices and processes more explicit or that we highlight nuances. For example, while context is considered within statistical frameworks, our Data Investigation Process highlights the importance of understanding the broader context, bringing the context of the problem to the forefront. The goal of the data investigation is not merely answering a statistical question but solving a real-world problem. This goal necessitate communicating results and proposing action to resolve the problem based on data-based findings. With the incorporation of Communicate and Propose Action as one of the six phases, this practice is more clearly visible. Additionally, other aspects of our work bring practices, processes, and dispositions from data science to the forefront. For example, previous cycles have included ideas on collecting and considering data (e.g., Bargagliotti et. al., 2020), but our process includes processing data as a separate phase to emphasize the work of wrangling data. Separating ideas of gathering and collecting data from the work of processing data, highlights the energy expended within this phase and the importance of work dedicated to organizing data into a structure that supports analysis.

Because this framework includes and highlights the key practices and processes of data scientists, researchers, curriculum developers, and practitioners, across various educational contexts, can use this framework to support learning and better understand student work with data. Researchers can use practices and dispositions in the framework to guide studies of students' work with data to attend to their thinking. It can help teachers reflect on the types of opportunities they provide their students to engage in different aspects of the data investigation process. Additionally, teachers can use the questions to consider (given in the example processes above) to help plan and implement data rich tasks in classrooms and courses.

REFERENCES

Agarwal, S. (2018, February 9). Understanding the data science lifecycle. *http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/*

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 Guidelines for assessment and instruction in statistics education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Ben-Zvi, D., Gravemeijer, K., & Ainley, J. (2018). Design of statistics learning environments. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 473-502). Springer. https://doi.org/10.1007/978-3-319-66195-7_16

Cao, L. (2017). *ACM Computing Surveys*, *50*(3), 1-42. Data science: A comprehensive overview. https://doi.org/10.1145/3076253

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734

Education Development Center. (2014). *Profile of a Big-Data-Enabled Specialist*. http://oceansofdata.org/our-work/profile-big-data-enabled-specialist.

Education Development Center (2015). *Call for Action to Promote Data Literacy*. Oceans of Data Institute. http://oceansofdata.org/call-action-promote-data-literacy

Education Development Center. (2016). *Profile of a Data Practitioner*. http://oceansofdata.org/our-work/profile-data-practitioner

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) Report: A Pre-K-12 curriculum framework*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK-12_Full.pdf

Friel, S., O'Connor, W., & Mamer, J. (2006). More than "Meanmedianmode" and a bar graph: What's needed to have a statistical conversation? In G. Burrill and P. Elliott (Eds.), *Thinking and reasoning with data and chance: Sixty-eighth Yearbook* (pp. 117–137). National Council of Teachers of Mathematics.

Goldstein, A. (2017, January 14). *Deconstructing data science: Breaking the complex craft into it's simplest parts*. Mission.org. https://medium.com/the-mission/deconstructing-data-science-breaking-the-complex-craft-into-its-simplest-parts-15b15420df21

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L., & Zanontian, L. (2016). Teaching data science to secondary students: The mobilize introduction to data science curriculum. In J. Engel (Ed.), *Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE)*. https://iase-web.org/documents/papers/rt2016/Gould.pdf

Graham, A. T. (1987). *Statistical investigations in the secondary school*. Cambridge University Press.

Lee, H. S., Mojica, G. F., Thrasher, E., & Baumgartner, P. (accepted). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*.

Watson, J., Fitzallen, N., Fielding-Wells, J., & Madden, S. (2018). The practice of statistics. In D. Ben-Zvi, K., Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 105-138). Springer.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x