# INVESTIGATING DATA LIKE A DATA SCIENTIST: KEY PRACTICES AND PROCESSES

HOLLYLYNNE S. LEE
*North Carolina State University*
*hollylynne@ncsu.edu*

GEMMA F. MOJICA
*North Carolina State University*
*gmmojica@ncsu.edu*
EMILY P. THRASHER
*North Carolina State University*
*epthrash@ncsu.edu*

PETER BAUMGARTNER
*Explosion AI*
*peter@explosion.ai*

**ABSTRACT**

*With a call for schools to infuse data across the curriculum, many are creating curricula and examining students' thinking in data-intensive problems. As the discipline of statistics education broadens to data science education, there is a need to examine how practices in data science can inform work in K–12. To better understand how to frame data science practices in K–12, we synthesize literature about statistics investigation processes, data science as a field, and practices of data scientists. Further, we provide results from a phenomenological study of the work of data scientists. Together, these inform a new framework to support data investigation processes. We explicate the practices and dispositions needed and offer a glimpse of how the framework can be used to move data science education forward.*

***Keywords:*** *Statistics education research; Data science education; Data investigation framework; Literature review; Industry ethnography*

## 1. INTRODUCTION

The ability to make sense of data and graphs is essential—from elections to COVID-19 to personal fitness trackers. In 2015, professionals from data-intensive industries and K–16 education endorsed a Proclamation of the Need for Data Literacy that included a call "for a revolution in education, placing data literacy at its core, integrated throughout K–16 education nationwide and around the world" (Education Development Center [EDC], 2015). The National Academies of Science, Engineering and Medicine (NASEM, 2018) made a strong call for undergraduate programs in data science and suggested that enhancements in middle and high school instruction and curriculum were needed to prepare students for such degree programs. Engel (2017) urged secondary and tertiary statistics education to focus on learning about complex data, and Bargagliotti et al. (2020a) called for and outlined learning outcomes for undergraduate cross disciplinary data pathways.

There has been long standing support for including statistics and data in K–12 education in U.S. curricula (e.g., *Principles and Standards for School Mathematics*, 2000; *Common Core State Standards for Mathematics*, 2010; *Next Generation Science Standards*, 2012) and efforts by the American Statistical Association to have guidelines and support for K–12 statistics instruction (e.g., Franklin et al., 2007; Bargagliotti et al., 2020b). Recent efforts around the globe included data science courses in high schools through such groups as the International Data Science in Schools Project (IDSSP) and a push to reconsider elements of mathematics curriculum necessary for 21st century learners, with data skills a top priority (e.g., Boaler & Levitt, 2019). Efforts in the U.S. not only included a focus on a high school course in data science (e.g., Gould et al., 2016), but ways that teachers and students in a variety

*Statistics Education Research Journal, 21(2), Article 3. https://doi.org/10.52041/serj.v21i2.41*

of disciplines in K–12 curriculum (e.g., mathematics, social sciences, science, computer science) could integrate a greater focus on making sense of real-world phenomena through and with data.

While the IDSSP (2019) provided a framework for a high school course in data science, we aim to provide a descriptive framework about key practices and processes in data investigations that could be used by K–12 teachers, curriculum developers, teacher educators, and researchers in data science education. To do this, we bring together theoretical perspectives from literature on research in K–12 statistics and data science education, and professional descriptions about the practices and processes used by data scientists. Our framework is also informed by empirical results from a phenomenological study to better understand the authentic work of data scientists.

## 2. THEORETICAL PERSPECTIVES FROM LITERATURE

In this section, we discuss frameworks that identify key practices, processes, and dispositions of investigating data. For decades, statistics educators and researchers have proposed frameworks that describe approaches to solving problems using data. While some frameworks suggest four-phase cycles (e.g., Franklin et al., 2007; Graham, 1987) and others suggest five-phase cycles (e.g., Watson et al., 2018; Wild & Pfannkuch, 1999), there is general agreement regarding central practices and processes for productively investigating data. Thus, our work builds on these well-established theoretical and empirical perspectives. As theory has begun to develop in the multidisciplinary field of data science, and professionals who work in this area have begun to document how they work with data to solve problems, it is crucial to identify practices, processes and dispositions that are shared and to expand work in statistics education. First, we highlight the theoretical and empirical work conducted in statistics education from which we draw upon. Next, we discuss literature from data science. Since theory in this multidisciplinary field is still developing, we also draw on the work of those who identify as data scientists or who use data to solve problems in their profession.

### 2.1. K–12 STATISTICS AND DATA PRACTICES

Across all levels of education, attempts have been made to include "data-driven approaches, more emphasis on data production and the measuring and modeling of variability (Moore, 1997), real data and contexts, and generally a more holistic approach that reflects the practice of statistics" (MacGillivray & Pereira-Mendoza, 2011, pp. 109-110). Many have recommended learners actively engage in real data investigations using a variety of technology tools (e.g., Ben-Zvi et al., 2018; Finzer, 2013). Others recommend learners investigate messy, large data requiring technology tools for actions such as in preparing, collecting, exploring and visualizing, and summarizing (e.g., Engel, 2017; Gould et al., 2018; Grimshaw, 2015).

For decades, statistics educators and researchers have proposed frameworks that describe key practices of statistics as one investigates data. Researchers in data science education (e.g., Gould et al., 2016) have recently begun to describe frameworks, often building on these models. Table 1 shows practices and processes across selected frameworks. Several describe a four-phase cycle for solving a statistical problem (e.g., Franklin et al., 2007; Friel, et al., 2006; Graham, 1987), which involves: Posing a question, Collecting data, Analyzing data, and Interpreting results. Building from this earlier work, Gould et al. (2016) and Bargagliotti et al. (2020b) emphasize Considering data, in addition to Collecting data. Wild and Pfannkuch (1999), IDSSP (2019), and Watson et al. (2018) propose five-phase cycles with similarities to these cycles but with greater attention to planning and exploring data. While many frameworks are cyclic, they emphasize the dynamic nature of a data investigation. Some describe a back-and-forth movement among phases (e.g., Bargagliotti, 2020b), yet others describe the process as simultaneously attending to various phases (e.g., Friel et al., 2006; Wild & Pfannkuch, 1999). Wild and Pfannkuch also highlight that engagement in an investigative cycle often motivates a new investigation.

*Table 1. Practices and processes from selected literature in statistics and data science education*

| Title | Source | Major Practices and Processes |
|---|---|---|
| PCAI Model of Statistical Investigation | Graham (1987) | Model: Pose question, Collect data, Analyze data, and Interpret results. |
| A 4-dimensional framework for statistical thinking in empirical enquiry | Wild & Pfannkuch (1999) | Four Dimensions:<br>*Investigative Cycle:* Problem, Plan, Data, Analysis, and Conclusions.<br>*Types of Thinking:* General (Strategic, Seeking Explanations, Modelling, and Applying techniques) and Statistical (Recognition of need for data, Transnumeration, Consideration of variation, Reasoning with statistical models, and Integrating the statistical and contextual)<br>*The Interrogative Cycle*: Generate, Seek, Interpret, Criticise, Judge<br>*Dispositions:* Skepticism, Imagination, Curiosity and awareness, Openness, A propensity to seek deeper meaning, Being logical, Engagement, and Perseverance |
| Process of Statistical Investigation | Friel et al. (2006) | *Process of Statistical Investigation*: Pose questions, Collect data, Analyze distributions, Interpret results.<br>*Other critical aspects*: data as a distribution and focus on variability and center. |
| Statistical Problem Solving as an Investigative Process and Developmental Levels of Statistical Literacy | Franklin et al. (2007) | Guidelines for Assessment and Instruction in Statistics Education (in K–12)<br>*Three Dimensions:* Problem-solving Process components, attention to variability, and three developmental levels (A, B, & C) based on statistical literacy.<br>*Problem Solving Process Components*: Formulate question, Collect data, Analyze data, and Interpret results. |
| Statistical Investigation Cycle and Statistical Habits of Mind | Lee & Tran (2015) | *Investigation Cycle*: Pose questions, Collect data, Analyze data, and Interpret results.<br>*Habits of Mind*: Always consider context; ensure best measures of attribute; anticipate, look for, and describe variability; attend to sampling issues; use several visual and numerical representations to make sense of data; embrace uncertainty but build confidence in interpretations; and, be a skeptic throughout investigation. |
| Mobilize Introduction to Data Science Curriculum | Gould et al. (2016) | *The Data Cycle*: Ask questions, Consider data, Analyze data, Interpret data |
| Practices of Statistics | Watson et al. (2018) | *Practices of statistics*: Problem posing, planning for and collecting data, data analysis (devising and presenting visual representations and summarizing and reducing data), and drawing conclusions.<br>*Other critical aspects*: Level of uncertainty and informal inference, as well as variation. |
| IDSSP model | IDSSP Curriculum Team (2019) | *Cycle of learning from data*: Problem elicitation and formulation, Getting the data, Exploring the data, Analyzing the data, Communicating the results |
| Statistical Problem-solving Process and Developmental Levels of Statistical Literacy | Bargagliotti et al. (2020b) | Guidelines for Assessment and Instruction in Statistics Education II (in K–12)<br>*Dimensions*: Statistical problem-solving process components and three developmental levels (A, B, & C) based on statistical literacy.<br>*Statistical Problem-solving Process*: Formulate statistical investigative questions, collect/consider data, analyze data, and interpret results. |

In describing the practice of statistics, Watson et al. (2018) splits the Analysis phase into two parts (i.e., Data Representation and Data Reduction), emphasizing experiences with visual representations. This may involve engaging in Exploratory Data Analysis [EDA] (Tukey, 1977), which is characterized by exploring data to summarize main characteristics. EDA is the "art of making sense of data by organizing, describing, representing, and analyzing data, with a heavy reliance on informal analysis methods, visual displays" (Ben-Zvi & Ben-Arush, 2014, p. 197). Although approaches often use visual methods, statistical measures are sometimes calculated to make sense of data.

Several multi-dimensional frameworks describe other important aspects of investigating data such as attention to variability, uncertainty, informal inference, and data as a distribution (Bargagliotti et al., 2020b; Franklin et al., 2007; Friel et al., 2006; Lee & Tran, 2015; Wild & Pfannkuch, 1999). They highlight the important role of context in a data investigation and that context should be considered throughout various phases. Additionally, Lee and Tran underscore other statistical habits of mind such as ensuring best measures of an attribute, attending to sampling issues, and using multiple visual and numerical representations to make sense of data. Both Lee and Tran and Wild and Pfannkuch point to the importance of being a skeptic throughout. Wild and Pfannkuch identify additional dispositions that are crucial to productively investigating data: imagination, curiosity and awareness, openness, engagement, being logical, propensity to seek deeper meaning, and perseverance. Further, Wild and Pfannkuch and IDSSP (2019) elucidate the significance of communication and collaboration in an investigative cycle.

## 2.2. DESCRIPTIONS OF DATA SCIENCE

Unlike statistics with its rich history as a discipline, data science is a newer field still being defined (Cao, 2017; Donoho, 2017; NASEM, 2018), but often referred to as multidisciplinary. The field has grown from industry's need to utilize and make sense of vast amounts of data (Cao, 2017) and a recognition in academia for new theories and methods for data analysts (Cleveland, 2001; Donoho, 2017). Most agree that data science is "the science of learning from data" (Donoho, 2017, p. 748) or using "data to solve problems" (Carmichael & Marron, 2018, p. 1). Data science is described as a multidisciplinary field (Barber, 2018; Cao, 2017; Conway, 2010; Geringer, 2014; Tierney, 2012), drawing "on individual skills and concepts from a wide spectrum of disciplines that may not always overlap with one another—a truly multidisciplinary field" (NASEM, 2018, p. 8). In general, most definitions (see Table 2) describe data science as an overlap between expertise in a field/business, mathematics and statistics knowledge, and computational/programming skills, usually shown in a Venn diagram (e.g., Barber, 2018; Conway, 2010; Geringer, 2014). The sources in Table 2 illustrate a variety of diagrams and descriptions used to depict the professional work of data science. Some add skills such as communication (e.g., Cao, 2017; Kolassa, 2014), and others break down larger domains (i.e., programming) into overlapping subdomains (e.g., Tierney, 2012).

*Table 2. Descriptions of data science from selected sources*

| Title | Source | Descriptions of Data Science |
|---|---|---|
| The Data Science Venn Diagram | Conway (2010) | Data Science as the intersection of [Computational] Hacking Skills, Math & Statistics Knowledge, and Substantive Expertise.<br>Conway emphasizes the need for understandings in all three areas to solve problems with data and particularly calls out a "danger zone" as a person working at the intersection of hacking skills and substantive expertise *but without* math and statistics knowledge to guide model assumptions and implications. |
| Data Science is Multidisciplinary | Tierney (2012) | Data Science is the intersection of many disciplines and skills: Machine learning, neurocomputing, AI, Data Mining, Knowledge Discovery in Databases, Data and Data Processing, visualizations, Statistics, and Pattern Recognition. Additionally, data science requires other general knowledge: domain knowledge, communications, presentation, inquisitiveness, problem-solving, business analysis, and business strategy. |
| Data Science Venn Diagram 2.0 | Geringer (2014) | This Venn diagram shows Data Science as the field that incorporates all activities within Computer Science, Math and Statistics, and Subject Matter Expertise.<br>This blogger places Unicorn in the middle of this Venn Diagram to highlight the uniqueness of a single individual having all of these skills. |
| The Data Scientist Venn Diagram | Kolassa (2014) | Data Scientist roles are grouped into 4 overlapping ovals that create a Venn diagram representing the different kinds of data science roles. These 4 parts are: Hacking skills/Programming, Math and Statistics Knowledge and Substantive Expertise/Business, Communication. The overlap of all 4 of these parts is the "perfect" Data Scientist. |
| Data Science: A Comprehensive Overview | Cao (2017) | Gives three common ways that data science is defined: high-level statement, disciplinary perspective, and data products.<br>*High-level*: data science is the science of data<br>*Disciplinary perspective*: data science = statistics + informatics + computing + communication + sociology + management conditional on data + environment + thinking<br>*Data products:* data science creates deliverables from data or a product that is enabled or driven by data. |
| Data science concepts you need to know! Part 1 | Barber (2018) | A Venn diagram where Data Science is the unique overlap of the disciplines: Computer Science/IT, Math and Statistics, and Domains/Business Knowledge.<br>This description more specifically mentions machine learning, software development, and traditional research within these overlapping disciplines as coming together to outline data science. |

Another approach to defining data science is to describe processes used by data scientists or the activities of a data scientist (see Table 3 for sample descriptions). Generally, these processes have six or seven parts. While most of these descriptions imply a cycle or connection between the parts (Agarwal, 2018; EDC, 2014, 2016; Goldstein, 2017, Saltz, 2020), others list different activities that comprise data science work (Donoho, 2017). Most begin with understanding/defining the problem and business/context (Agarwal, 2018; EDC, 2014; Goldstein, 2017; Saltz, 2020). This involves "identifying the central objectives of your project by identifying the variables" (Agarwal, 2018). The next steps vary between different descriptions, but all involve gathering, cleaning, transforming and/or managing data. For example, Goldstein (2017) described steps around collecting raw data and processing data, while

the *data practitioner* (EDC, 2016) "wrangles data" requiring data collection and cleaning among others. Donoho (2017) combined wrangling the data and exploration and states that "80% of the effort devoted to data science is expended by *diving into* or *becoming one* with one's messy data to learn the basics of what's in them" (p. 755). Donoho also included another activity of Data Representation and Transformation that deals with data structure. For others, data exploration is a separate step (Agarwal, 2018; Goldstein, 2017) and is combined with the step of performing the duty of "analyzes data" (EDC, 2014, 2016). The Cross Industry Standard Process for Data Mining (CRISP-DM) framework (Saltz, 2020; Shearer, 2000, as cited in Saltz et al., 2017) called the analyze phase "Modeling", which includes selecting modeling techniques, building a model and testing the model, while Donoho (2017) separated out what many consider the analysis phases into two activities: computing with data and data modeling. Finally, a data scientist "closes out the project" (EDC, 2014) and communicates findings (EDC, 2014; Goldstien, 2017) possibly through data visualization (Agarwal, 2018; Donoho, 2017). The CRISP-DM framework delineates two phases of "Evaluation" and "Deployment" to round out its process. Donoho (2017) also described the activity of "Science about Data Science", which is the activity of understanding the patterns and methods within the field of data science to better the field. Saltz and colleagues (2017) contributed to this understanding of the field and illustrated how big data projects may involve phases, but that actions were done in coordination across large teams where individuals had specific roles in the process. In 2020, Saltz and Hotz reported that the six-phase CRISP-DM was the most used framework from their survey of 109 industry professionals in data science.

*Table 3. Selected sources that describe practices and processes of data science*

| Title | Source | Major Practices and Processes |
|---|---|---|
| Profile of a Big-Data-Enabled Specialist | EDC (2014) | *Duties*: Defines the Problem, Wrangles Data, Manages Data Resources, Develops Methods and Tools, Analyzes Data, Communicates Findings, Engages in Professional Development |
| Profile of the Data Practitioner | EDC (2016) | *Duties*: Initiates the Project, Sources the Data, Transforms the Data, Analyzes the Data, Closes Out the Project, Engages in Professional Development |
| Six Divisions of Greater Data Science | Donoho (2017) | *Aspects of work in data science:* Data Gathering, Preparation and Exploration, Data Representation and Transformation, Computing with Data, Data Modeling, Data Visualization and Presentation, Science about Data Science |
| Data Science Deconstructed | Goldstein (2017) | *The Data Science Process*: Frame the Problem, Collect Raw Data, Process the Data, Explore the Data, Perform In-Depth Analysis, and Communicate Results. |
| Data Science Lifecycle | Agarwal (2018) | *7 steps of the Data Science Lifecycle*: Business Understanding, Data Mining, Data Cleaning, Data Exploration, Feature Engineering, Predictive Modeling, and Data Visualization. |
| CRISP-DM | Saltz (2020) | *Cross Industry Standard Process for Data Mining*. Six high-level phases that can be used to frame projects in data science: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. |

## 3. EXAMINING THE WORK OF DATA SCIENTISTS

In this section, we describe the research study and results that provide the empirical basis for our framework development. In order to authentically understand the work of data scientists, a phenomenological approach was utilized (Creswell, 2013) to focus on the common lived experiences of engaging in data science work. Phenomenological studies typically include interviews with those who have first-hand knowledge of the phenomena under study, as well as observations of different aspects of the experiences within a context.

### 3.1. METHODS AND PARTICIPANTS

To dig into the day-to-day work of data scientists, the first author immersed herself for nine months in 2018–19 to attend meetings, presentations, and have informal conversations with data scientists working on a diverse set of projects at the Center for Data Science at RTI International. Field notes and memoing were used to document observations and impressions of the work the group of data scientists were engaged in. The fourth author, as a member of that group, engaged with the first author to discuss observations and wonderings and to clarify purposes or details about practices observed. This form of member checking ensured that field notes and memos were accurately capturing aspects of phenomena of data science work (Creswell, 2013).

The first and fourth authors designed an interview protocol using many of the sources in Tables 2 and 3, as well as craft knowledge about data science work and questioning techniques that elicit rich descriptions (see protocol in Appendix). For example, we felt it was important to have the interviewees describe details about their work on a current project, including aspects that were challenging, more simplistic, and experiences that may have evoked emotional responses (see questions in Part 1 of protocol). In addition, we also situated the data science professionals in contexts where they were asked to describe the work of a data scientist to different audiences (see Part 2) and to engage in a critique of diagrams and descriptions of the work of data scientists created by others to describe the profession (see Part 3).

A list of 13 potential interviewees was generated, with the intent to interview 8–10. Potential interviewees were chosen based on the following: recommendations from author four identifying others he considered to be working in data science in a few different organizations, a university professor working closely with projects at the Center for Data Science at RTI International, contacts of the first author who had connections or experiences with education and were currently working as data scientists, and two data scientists involved in the Women in Statistics and Data Science committee and conference. Only one recruited participant declined participation with a justification they felt they did not have as much direct experience with data science work to effectively contribute. After extensive attempts at scheduling common times for interviews over a two-month period, the first author conducted 45 to 60-minute interviews with five data scientists. Due to the small number of interviews, the first author did an internet search to locate conversations and interviews with data scientists that were posted publicly online. Seven such publicly posted interviews with data scientists were collected. Data scientists worked at a variety of companies (e.g., SAS, RTI International, Pivotal Data Labs, Insights Association, Home Depot) and one was a university professor working in data science within a computer science department. Eleven were male, and one was female. Most had undergraduate degrees in a quantitative field (mathematics, statistics, economics, engineering, experimental psychology, computer science) and worked in another discipline (e.g., business, engineering, high school mathematics teacher) before transitioning to a career in data science. A few had a masters degree in data analytics or data science.

Field notes, researcher memos, interview transcripts ($n$ = 5), and public interviews ($n$ = 7) were organized and coded using Atlas.Ti. Aligned with analytic methods used for phenomenological approaches (Creswell, 2013), the data corpse was read, reread, and open coded for issues and perspectives about the work of data scientists. As new codes formed, data were revisited to refine and recode and write descriptions of each code. Codes were grouped into categories. By examining descriptions within each category, emergent themes were identified and characterized.

### 3.2. RESULTS

Results about the work of data scientists are organized based on four themes that emerged from analysis. Within each theme, examples are included to illustrate the findings.

***Data science as a growing field.*** Several data scientists noted they have been doing data science before having that title, highlighting that the field has evolved to encompass many different fields. Two key background areas came up often: statistics and computational fluency/programming, though several emphasized that statistics and programming can be taught if a person had strong general problem-solving and communication skills. Regarding statistics, one data scientist noted:

*As far as statistical preparation, there's a lot of understanding that you have to know about what a dataset has to look like for a certain type of modeling algorithm ... and then there's just the practice of knowing what level of aggregation makes sense for this context.*

The data scientists indicated they used programming skills and computing tools such as SQL, R, Python, SAS, Excel, Hadoop, and GitHub to accomplish different tasks. They often learned new tools and techniques and were continually engaging in self-directed learning to improve their skills. The field of data science is heavily dependent on computing tools, and these were ever evolving, where each has strengths that can be useful for different aspects of their work. Several data scientists discussed how some projects involved developing specific applications to process, analyze, and display data. For example, one data scientist developed a tool for text mining and analysis, another developed an application to assist in processing geographic satellite images to extract data, and another developed applications for automating data extraction and analytics from video streaming platforms.

***Characteristics of data.*** All data scientists discussed the volume of data they used and how they were often managing, cleaning, transforming, and merging several data sources (e.g., tables, text, images, videos). They spent a lot of time combing through data sources, digging into values and meaning of measurements (e.g., "Is this zero a true value or did the sensor malfunction?"), and framing and reframing a problem to ensure they used the right data to answer questions posed and meet a client's needs. One data scientist summarized the feeling of working with big data:

*I'm not sure how to describe the feeling behind this, but when we have a dataset this big, I mean you just can't ever know everything about it for sure other than easy descriptives, like how many records are in it. So, it's this feeling of looking into the unknown, looking into the ocean. You're like, "I really can't conceive of what is in there."*

The data scientists often used data generated or collected by another entity or process. They considered ethical issues concerning access and use of data, privacy policies in applications that generated data, and the provenance and quality of data. They carefully thought about biases that may be represented in data or biases that are introduced because of what is *not* represented in data (e.g., biases in Twitter data because of characteristics of users and non-users of Twitter). However, sometimes projects involved creating applications to produce and capture data to investigate a phenomenon (e.g., health sensors, geographic satellite images, or video streaming platforms). In this way, they designed observational data collection methods and defined measurements needed for their problem.

***General skills and ways of thinking.*** One skill was emphasized over and over again: the ability to communicate. The data scientists needed to communicate with team members, their clients, and other domain experts. Communication involved listening carefully to a client's needs, asking questions about the meaning, purpose, and measurements in a dataset, documenting procedures for reproducibility, describing strategies to others, and getting and giving feedback.

Communication was also tied to another highly discussed aspect of data science, that of storytelling. Data scientists are required to deeply understand the purpose of their data investigation project and know how to communicate to a variety of audiences (e.g., "selling skills") so there is usable and actionable output from their projects (e.g., "value proposition"). Storytelling with data included data visualizations and easy-to-understand diagrams, brief written papers or "quick guides", or data dashboards. Large teams often had data visualization specialists who assisted with final products for strong communication.

Several personality traits or soft skills discussed by data scientists include creativity, curiosity, passion, persistence, and resilience. Many data scientists emphasized the creative aspects of their job, that solving big data problems was not a cut-and-dry application of specific techniques. They often drew upon their curiosity and passion for making sense of a vast amount of data and looked at problems in new or out-of-the-box ways. As one data scientist noted,

*... passion also plays an important role in data scientists' life. Are you excited about getting valuable insights out of messy big data via creative ways of applying machine learning models, scaling up your algorithm to petabyte scale?*

In team meetings, individual work, and presentations, data scientists demonstrated persistence and resilience in problem-solving. Their projects span months, and often restart or shift in direction. Throughout, the data scientists persisted and "chipped away" at the problem and talked about "not giving up" or building resilience to not get discouraged when they must throw out early work or spend days or weeks wrangling data and making sense of measurements to move forward effectively.

As mentioned above, the data scientists were skilled problem solvers who approached tasks with logical reasoning, flexibility, a "hacker" mindset, intuition, inquisitiveness, and caution and skepticism—all general ways of thinking mentioned in interviews. Being inquisitive, cautious and skeptical were brought up as data scientists described aspects of their work. For example, one data scientist noted:

> *Now that we have the data, we have further questions about it. There's this kind of trepidation about making assumptions about the data, but also kind of really needing to know the facts behind it. There definitely is that moment of clarity when you do get to ask those questions* [to experts] *and feel like, "okay, yeah, we definitely understand this."*

Time management, efficiency, and being a team player were essential for keeping data projects on track. They were acutely aware of how long certain tasks may take and what they needed to do daily to move work forward across projects. Some data scientists that worked on larger teams also consulted with other team members for solution strategies and knew certain tasks were interdependent and required time coordination.

***Key practices and processes*** In interviews, presentations, and team meetings, it was evident that data science work was rarely done in isolation, was nonlinear, and was approached holistically, always situated within the larger phenomena. The data scientists did not blindly apply data analytic techniques. Instead, they worked on developing a deep understanding of the problem and context that encompassed a project or needs of a client. They were immersed in context throughout a project and always had an eye towards the value of their work for clients, business, or a discipline. A data science project always started with:

> *... how you frame the problem, and then you can figure out the aspects of the dataset that are relevant to that problem, and maybe do some filtering to get it down to a more manageable amount. And, then, you can really start to get a picture of what the data contains, with respect to that problem.*

They spent time searching for datasets, combing through them to decide which data were useful, making sense of data, and processing, cleaning, or wrangling data. Their data sources varied and often existed in many different tables and formats and needed to be parsed and merged. As mentioned, they often engaged in agile software development that generated data from various sources or used web scraping techniques to obtain data from larger online sources.

In making sense and exploring data, data scientists emphasized working with purpose, guided by a larger problem and in productive ways. They spent so much time immersed in data, one data scientist explained they,

> *... have a natural curiosity to explore things and dig into things, but are you really digging into the right things? ... if you don't get those questions right at the beginning, you can just waste so much time and energy.*

Another expressed that,

> *When you get a dataset and someone says to play around with the data but doesn't give you a goal. It's kind of a useless exercise.*

Exploration of data is grounded and purposeful. Some data scientists talked about exploring and sense making of data by creating visualizations of distributions and relationships, descriptive statistics, and sometimes just scrolling and searching through large CSV files in a spreadsheet or database tables to understand the structure of data, missing values, or anomalies. This indicated that one purpose of exploring data was to discover what aspects of data needed to be cleaned or transformed, and there seemed to be a back and forth between exploring data, processing data, and even collecting data as they saw a need for additional data to supplement existing data.

The work of data scientists involved a lot of modelling. In the beginning of a project, they considered which models or modeling techniques (e.g., predictive modeling, machine learning, agent-based modeling/simulation) may be useful and the data and its format needed. One data scientist said you start with,

> *... some hypotheses about the data, because model building is basically a hypothesis and that you see how your hypothesis is changing after you revisit and add data.*

They built models, ran simulations, used samples of data to train their models, and revised models often through iteratively revisiting data in exploratory ways, and sometimes needed to find additional data sources or do further data wrangling. Statistical modeling was a strong aspect of data science work and they recognized the uncertainty in their models. They discussed how it is hard to communicate uncertainty to clients. As one said,

> *... in data science the output is not something that's easily verified ... if you're building a model, or you're communicating something, there's no judge of whether it's right or not;*

so, there needed to be an evidence-based argument for a model and results to build a level of confidence.

The data scientists continually revisited their larger problem and client needs. They were solution-oriented and intended final products be used for improvement, and many saw their work specifically aimed at finding solutions for the greater social good. A final product could be a software application, a data visualization dashboard, a report or presentation with key findings, or a guidebook for a client or business. The data scientists made their work actionable and meaningful.

## 4. FRAMEWORK FOR DATA INVESTIGATIONS

Our framework development process drew upon the results from the empirical study of the work of data scientists described in Section 3 as well as the theoretical perspectives described in Section 2 to identify practices, processes and dispositions used in statistics education and data science (see Section 2 and Tables 1 and 3), as well as blogs and other media resources illustrating depictions of data scientists' work (see Tables 2 and 3). After agreeing on overarching phases of a data investigation, we delineated the specific work within each phase. Once a framework was drafted, we got feedback from statistics educators, data scientists, mathematics and statistics teachers of grades 6-12, mathematics and science teacher educators, and a data software developer. Feedback was used to refine the framework.

We propose a Data Investigation Process that involves six phases (Figure 1). Although this process may be linear and cyclic, it is often nonlinear and dynamic in nature. While some describe investigators as simultaneously working within phases (e.g., Friel et al., 2006), we describe a process that involves revisiting and refining work within phases and making connections among phases. The phases fit together like pieces of a puzzle, emphasizing a holistic and productive approach to data investigations. By engaging in and connecting phases, investigators are able to make sense of a real-world issue through data and make evidence-based claims and inferences to propose solutions to a problem. Investigators are situated in the center of the diagram (white puzzle piece) and enter the appropriate phase as needed. This movement is fluid and may be messy at times.

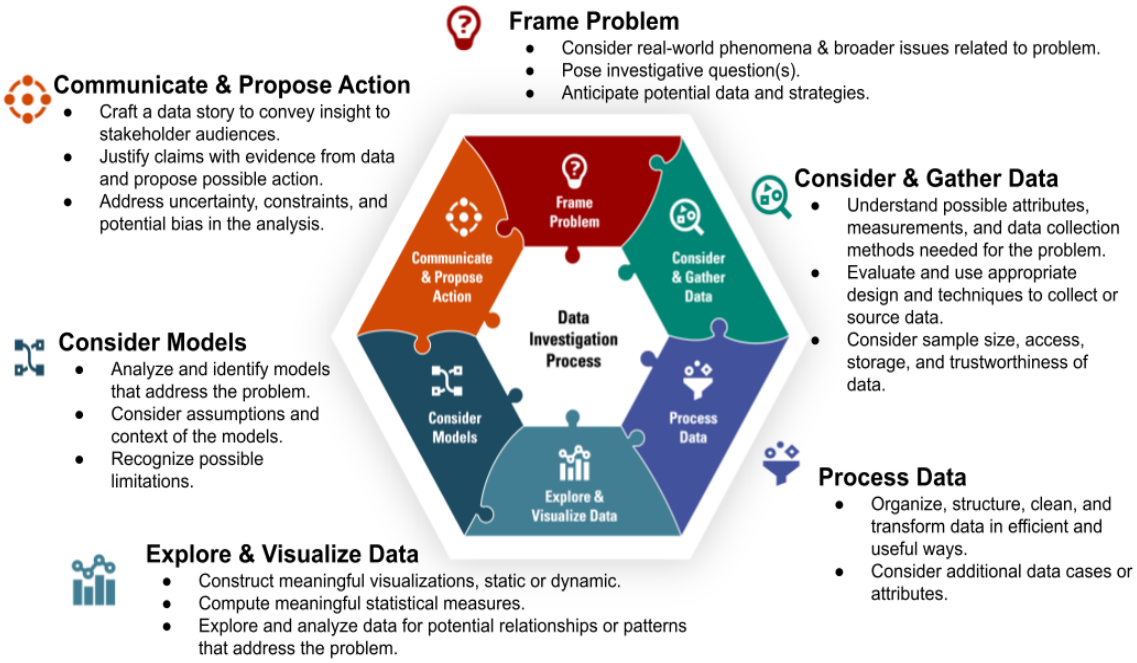


*Figure 1. Data Investigation Process framework with example practices*

Our framework builds on the work of statistics educators and researchers. While we expand this work by making fundamental practices and processes from data science explicit, it is not to say these practices and processes are not a part of doing statistics or are completely absent from other frameworks. Rather, our purpose is to elevate specific practices and processes since they may be implicit in other frameworks.

### 4.1. FRAME THE PROBLEM

Data investigations typically begin by considering real-world phenomena and broader issues related to a problem. Most cycles in statistics education begin with posing an investigative question (e.g., Graham, 1987; Wild & Pfannkuch, 1999) within a context. Wild and Pfannkuch point out that problem-solving is grounded in a *real* problem to change a system for the better. Data scientists' work is rooted in solving real-world problems, and they emphasize the importance of understanding broader issues before focusing on a specific question (e.g., EDC, 2014). Broader issues related to the problem include identifying: necessary background information, importance of the problem, and available data within the context or discipline. In considering variability inherent in context, one or more investigative questions are posed that could use statistical approaches to answer question(s) after the broader issues are understood. According to one data scientist, framing a problem includes,

> *… figure[ing] out what the focus is and it kind of goes back to what I was talking about reframing the problem. You know, the client gives us one problem, but maybe a different problem is easier to solve.*

Throughout a data investigation, an investigator should continually be revisiting the Frame the Problem phase in order to keep the context of the real world problem at the fore to inform work in other phases.

### 4.2. CONSIDER AND GATHER DATA

*Consider and Gather Data* involves considering types of data needed to answer an investigative question. As emphasized in the interviews with data scientists, sometimes data has already been collected, and other times data will need to be collected or gathered. Design and methods for data collection or programming techniques should be considered. The investigator should understand what

data is relevant and useful to addressing the problem, as well as its attributes, and how these attributes are measured.

Key considerations should be given related to the data source, how it is or will be stored and accessed, and whether additional data needs to be collected (e.g., larger sample, different or additional attributes). Other issues related to bias and ethical concerns should also be considered: potential bias related to collection or source, who is represented in the data and who is left out, an investigator's personal connection to and knowledge about a data source, trustworthiness of data source(s), and the purpose and in whose interest the data was collected or will be collected.

### 4.3. PROCESS DATA

Considering strategies for processing and structuring data is a necessary aspect of data investigations. Data scientists spend much of their time processing big data (e.g., Agarwal, 2018; EDC, 2016; Donoho, 2017; Saltz et al., 2017). Attention should be paid to considering strategies and techniques that are most useful for accomplishing the investigator's goals, and should take into account: efficiency, ease, expertise, and available resources. Work in this phase involves obtaining data in a usable and consistent format which may include merging data that may or may not be structured similarly. It also involves cleaning data to identify and make decisions about possible erroneous/invalid and missing data. Data may need to be transformed through normalizing, creating new attributes from existing attributes, converting measurements or recoding data values. It also involves processes that may help focus an investigation, such as sorting, grouping or filtering data. One data scientist explained,

> *I actually look at the problem and these ten different sources, I really only need these seven ... now it's about aggregating and combining data and getting it all into one place. The next step is really about transformation … there's missing values, there's things that are coded incorrectly in the data. I have this one that's recorded by day but I actually need it summed up to the week level, or I need the log transformation of it.*

### 4.4. EXPLORE AND VISUALIZE DATA

In Figure 1, the phases of Explore and Visualize Data and Consider Models are two shades of the same color. This is purposeful to highlight the connectedness between these phases. While other frameworks combine them into one Analyze data phase (e.g., Friel et al., 2006), we emphasize the importance of exploring data and creating data visualizations, often using dynamic data visualization and analysis tools, in accordance with Watson et al. (2018). Engaging in exploratory data analysis (i.e., EDA) is grounded in the goals of an investigation and understanding of the context. EDA may involve multiple visualization and analysis techniques that give an investigator insight into the data and can lead to more purposeful refinement of questions, call for a need for additional data, or refinement of hypotheses and models.

*Explore and Visualize Data* involves creating data visualizations (e.g., graphs, images, diagrams), which may be dynamic in nature, and statistical measures to explore and reason about data in relation to an investigative question and context. While this includes looking for relationships among attributes, patterns and trends, we draw explicit attention to the role of exploration, visualization, and modeling as key aspects of analyzing data. Data scientists highly value visualizations during their own exploration as well as for communicating results. Different visualizations can inform other phases of an investigation: a diagnostic plot can help with model selection, a distribution plot can help identify outliers or data quality issues that mean revisiting data collection or processing, and a dashboard built to communicate results could inform actions taken as a result of analysis or even reconsideration of the problem framing. As one interviewee noted, visualizations are "a really underrepresented part of the data science art" and that early emphasis on visualizing data would be good for young learners.

### 4.5.CONSIDER MODELS

Investigating data involves exploring and selecting models (e.g., statistical measures, data visualizations, predictive models, distribution models) that address a problem and answer an

investigative question, taking into account variability and uncertainty. As mentioned, it is highly connected to Explore and Visualize Data, where the focus was on exploration and visualization, although data may have also been summarized as statistical measures. The key distinction is that the investigator chooses specific models as evidence to support claims that address an investigative question, often discarding models that do not help answer a question. One data scientist illustrated this iterative process, including revisiting processing data:

> *… there's the modeling which is taking this data, and sometimes those things go back and forth where I'm modeling and then I say, "oh, I actually need to go back, I need to transform more stuff. I need to go get more data. I need to go combine more data."*

Another aspect of considering models involves deciding whether the objective is inference or prediction—or understanding how important model interpretability is to answer the investigative question (Breiman, 2001). In a standard research study aiming at understanding some phenomena, simplicity and model interpretability may be paramount, but if the task is to automate the classification of documents, performance matters above interpretability. In addition, both philosophies can be used in concert to improve each other: opaque, black-box models can be used to understand the "upper bound" of predictive performance, while more interpretable methods can help determine areas of improvement in data quality or model performance.

### 4.6. COMMUNICATE AND PROPOSE ACTION

*Communicate and Propose Action* involves connecting and interpreting results and models to context and making evidence-based claims in relation to a broader problem and specific investigative question. Additionally, it involves proposing actions to solve problems. When devising a strategy for communicating recommendations and proposed actions that are supported by evidence, the investigator should take into account the stakeholders, as well as the audience. It is important to craft a data story to convey insight about the problem to the stakeholders/audience so they can make informed decisions. Bargagliotti et al. (2020b) emphasize that data should be used to tell a story, which was also emphasized by data scientists that were interviewed. Lastly, the investigator should consider the most effective mechanism, format and language for communication. A data scientist stated:

> *The most powerful stuff that we do for our clients is the vessel of delivery.* [On a project] *we just put together what we called a "quick start guide" and this encapsulated all of the cool models and insights and findings from all of 2018 workstreams and put it into tangible, actionable, consumable insights for the people who are actually on the frontline of this business.*

Like work throughout the process, where the investigator is constantly revisiting and refining various phases, sometimes recommendations and proposed actions may include revisiting data from a new perspective or collecting additional data. Often, the result of work in this phase may lead to uncovering additional problems and motivate new investigative questions, in accordance with Wild and Pfannkuch (1999).

### 4.7. KEY CONSIDERATIONS AND DISPOSITIONS THROUGHOUT

There are several key considerations that should be attended to throughout a data investigation, as well as important dispositions.

***Make sense of data with respect to context.*** It is crucial to engage with context throughout the entire process. This means continually making sense of data with respect to real-world phenomena and context, as well as the broader problem and investigative question. Contexts should be meaningful and relevant to the investigator, where the problem-solving process has a purpose. Keeping the context of data at the forefront should strengthen all decisions, actions, and interpretations made within each of the phases.

***Take advantage of technology tools.*** One could argue it is impossible to investigate big data without technology. At each phase, the investigator should consider which technology tool is most appropriate to facilitate the work at hand as several *different* tools may be needed throughout the process since each

tool will have a different purpose. When framing the problem, tools (e.g., videos, images, reports, social media, maps) can help investigators situate a problem in a real-world phenomenon. In considering and gathering data, an investigator may need to identify appropriate tools to collect data. Different tools may be needed to design and invite participants to complete a survey, for gathering data from sensors or personal devices, or scraping websites to collect data in usable formats (e.g., CSV or JSON). While some tools are useful in processing data (e.g., spreadsheets, Python, R), other tools are useful for exploring and visualizing data and considering models (e.g., Tableau, CODAP). Finally, additional tools (e.g., video-makers, dashboards, presentation tools) may be needed to communicate findings to support recommendations and proposing actions. Data science project teams typically have different individuals who may have expertise (or a willingness to learn) in each of the different tools needed.

***Consider common biases and ethical issues.*** An investigator should understand the data collection process and identify potential limitations in the generalizability of their analysis. All data and models contain potential biases and investigators must ethically interrogate sources of these biases. Investigators have an obligation to mitigate interpretations of results with data and do not perpetuate inequities or overgeneralizations in how others use results from an investigation. There may be missing data from certain contexts or groups of people (e.g., health data collected on personal fitness devices are likely not inclusive of many groups of people), and missing or limited measurements that impose choices not reflective of individuals or phenomena (e.g., gender identity, familial living and housing structures across cultures), especially in curated data where an investigator has little control over what had been measured. Investigators have an ethical responsibility to consider the final "stakeholders" of their analysis and how underrepresented groups may be impacted by actions taken as a result of the analysis (e.g., D'Ignazio & Klein, 2020). When available, models and visualizations should be explored by relevant demographic variables to identify potential areas of concern. Finally, investigators should be aware of how their own biases, experiences and perspectives enhance or negatively impact work during an investigation, and ideally work within a team with diverse backgrounds and experiences. Other ethical issues like privacy should also be considered (e.g., protecting identity, restricting sharing of information). Investigators should be transparent about what data is collected, how data is stored, and if and with whom it is shared.

***Seek expertise and information.*** Throughout investigations an investigator should evaluate whether they have the skills needed to carry out processes and seek appropriate expertise when needed. For example, an investigator may need data scraping to gather data from a specific source and may need additional assistance to obtain this data and structure it in a useful format. Taking a team approach to data investigations can ensure a wider range of skill sets and perspectives on a project (e.g., Saltz et al., 2017). It may also be necessary to reach out external stakeholders or experts to find additional information about the data context or phenomena to inform work throughout an investigation.

***Communicate.*** Wild and Pfannkuch (1999) and IDSSP (2019) emphasize the importance of communication and collaboration throughout an investigation. This was also emergent in the interviews with data scientists. This may take place between a client or stakeholder and the investigator. Many data investigations involve teams of individuals working together, relying on different expertise, to solve a problem, where communication and collaboration is a part of the entire process.

***Engage in phases as needed.*** While investigations may proceed linearly and in a cycle, all investigations do not emerge and proceed in this way. As noted earlier, expert data investigators likely move among the different phases in fluid ways that may seem messy to a novice. Novice investigators will likely need support in understanding how and when their investigative work moves among the phases. For example, one may begin with a set of data that has already been collected and do some preliminary exploration and visualization of data (EDA). From what is noticed, one may go back to Consider and Gather Data to consider the data source, make sense of different measures, and decide to use different strategies to Process Data in meaningful ways. One may then dive into resources to Frame the Problem by making sense of the bigger context the data represent and pose a targeted statistical question involving only a few attributes in the dataset. From there, the appropriate attribute(s) of interest would be selected, and one may proceed to Consider Models and require additional work in the Explore

and Visualize Data or Process Data phase. Deciding how to Communicate and Propose Actions may spark new or additional questions to require further investigation with data at hand or require additional data collection and processing.

***Dispositions.*** Statistics education frameworks and models from data science education (e.g., IDSSP, 2019; Lee & Tran, 2015; Wild & Pfannkuch, 1999) identify dispositions that are key in engaging productively in data investigations. Data scientists also identified these characteristics which included: creativity, curiosity, intuition, passion, persistence, perseverance, and resilience. These dispositions towards solving problems with and through data will not develop in a single data experience and must be grown through extended and repeated opportunities.

***Experience & Expertise.*** As investigators gain more experience with using the framework to more productively investgate data, they can develop the ability to use the framework more fluidly (i.e., knowing which phase applies at any given time and which phase to do next), as well as contextually (i.e., working on a problem where data might already be collected, so there's less focus on the Consider and Gather stage) (National Research Council, 2000). A beginner benefits from explicitly and consciously working through the steps of the framework to build their skills and experience in each phase to eventually gain the fluidity and contextual-based reasoning, which leads to a fluency that can result in more subconscious or implicit transitions between phases in data investigations. In reality, those working on larger data science projects work in teams where individuals bring strengths in various aspects of the work needed.

## 5. DISCUSSION

In our discussion, we first highlight the ways we have built on and expanded the work in statistics education and data science to provide a descriptive framework about key practices and processes in data investigations that could be used by K–12 teachers, curriculum developers, teacher educators, and researchers in data science education. Next, we discuss potential ways this framework can inform efforts of those who work in data science education. We identify some limitations of using such a framework, as well as implications throughout.

### 5.1. EXPANDING FRAMEWORKS FOR INVESTIGATING DATA

In order to prepare today's students for potential careers in data science or other STEM careers that focus on solving problems with data, it is crucial to provide data-intensive experiences throughout K–12+. Whether students choose data-intensive careers, a goal of the K–12+ schooling experience should be to prepare all students to be data literate, which includes the ability to evaluate and make data-based decisions that impact their everyday lives. While current efforts have begun to focus on developing data science courses at the secondary and tertiary levels (e.g., IDSSP, 2019) where the emphasis is on solving problems using messy, large data (e.g., Gould et al., 2018), it is imperative to develop frameworks that describe the practices, processes and dispositions that guide productively investigating this type of data.

In 1999, Wild and Pfannkuch's framework of statistical practices and dispositions, based on observations of statistics students and statisticians, had major influence on the next two decades of statistics education research, curriculum development and statistics teacher education. While our framework development was inspired by and builds from this collective work, we purposely include critical aspects of the work of data science to create a framework to depict processes, practices, and dispositions that support investigating data in the modern age, where work with data often involves solving problems using large, complex data.

Work in statistics education often emphasizes that investigative questions should be statistical rather than mathematical in nature and should be posed within a context (e.g., Franklin et al., 2007) using real or *realistic* data. We highlight the importance of solving real-world problems within a context using *real* data to solve a real problem as underscored in the work of data scientists (e.g., EDC 2014, 2016). This makes this process more relevant to other disciplines (e.g., Social Sciences, Science), and provides other disciplines guidance on how to use data to solve problems. Using data to tackle real problems means students and teachers must develop deep understandings of the context of data, as the context is

what drives the purpose of the investigation. Some disciplines, such as mathematics, are often taught devoid of context, or with minimal attention to the rich aspects of a context. Being a critical data investigator, and developing students' literacy for using data in their everyday lives, means bringing the rich, and often difficult, aspects of a real problem that impacts society (e.g., climate change) and daily lives of our students (e.g., environmental racism) into the classroom to develop critical consciousness and use of data to understand their world (e.g., Lee & Campbell, 2020; Lesser, 2007; Pangrazio & Selwyn, 2021).

Like frameworks from statistics education and data science discussed in Section 2, our framework is composed of multiple phases. While processing data likely occurs in enacting frameworks from statistics education, the work in this phase (i.e., cleaning, transforming, managing data) often accounts for much of data scientists' work (e.g., Agarwal, 2018; Saltz et al., 2017). In fact Donoho (2017) suggests that data wrangling involves 80% of data scientists' efforts. An emphasis on these practices and processes are likely due to the type of data that data scientists often encounter in their work (i.e., large, complex, messy data). In accordance with Agarwal (2018) and Goldstein (2017), we emphasize the importance of exploring data and creating data visualizations with dynamic technology by delineating Exploring and Visualizing Data and Considering Models, whereas most statistics education frameworks combine these into one phase labeled Analysis (e.g., Bargagliotti et al., 2020b). Many frameworks in statistics education (e.g., Graham, 1987) end with the investigator making interpretations about the results from an analysis process. Our work, influenced by the importance of solving real problems and communication, goes a step further by suggesting that the end goal should not just be about justifying claims with evidence from data but should also focus on *proposing possible actions* where a data story and possible solutions to the problem are conveyed to stakeholders. As highlighted by Wild and Pfannkuch (1999), this can bring to light new problems, motivating new questions or bringing to the surface unresolved questions that become the focus of a new data investigation. Many aspects of our work are influenced by a shift in work with small, tidy datasets to work with large, complex datasets that are a hallmark of our time.

A distinction in our work is that we identify key considerations that should be attended to throughout the entire problem-solving process when investigating data. For example, we emphasize the importance of considering bias and ethical issues throughout. Due to the collaborative nature of solving real problems on teams, we draw attention to the importance of seeking the right expertise and communicating with colleagues and stakeholders throughout. One of the most significant aspects of our work is emphasizing that work with data to solve problems does not proceed in a linear, cyclical fashion. Rather the process is nonlinear, dynamic and fluid, where the investigator consistently returns to previous phases, refines work and makes connections among phases. It is understanding the big picture or a holistic view of all the pieces of the puzzle (Figure 1) and their connections to one another that supports the investigator in solving a problem. Finally, our work highlights key dispositions that support investigating data. While some were identified in the work of Wild and Pfannkuch (1999), they have not been as heavily emphasized in the work of statistics educators as in contrast to the work of data scientists (e.g., IDSSP, 2019).

Even though our work builds and expands the work of others, it does have limitations. Many of the frameworks from statistics and data science education focus on four or five-phase cycles (e.g., Bargagliotti et al., 2020b; Gould et al., 2016). Some might argue that these frameworks are less complex and offer an easier pathway for investigators, especially those in a K–12 setting. We would argue that although we describe more phases, our framework provides descriptions that more specifically delineate the authentic work of those who work with data to solve problems and provides other disciplines with guidance about how data science can support their work. Additionally, teachers themselves need to understand this complexity in order to engage their students with authentic data investigations. We caution that there should not be an expectation that learners in K–12+ settings engage in every phase of the Data Investigation Process *every time* they work with data. We recognize the time and complexity involved in using such an approach and constraints in K–12+ settings. While we advocate for engaging in the entire process and there is support for this (Watson et al., 2018), it is also appropriate to provide experiences with separate aspects of the framework to develop reasoning about various ideas relating to statistics and data throughout school experiences.

## 5.2. IMPLICATIONS OF THE FRAMEWORK AND CHALLENGES

Our framework provides a next step forward in *data science education.* For K–12+ education to meet the urgency to prepare data literate students and students that are ready for data careers, understanding the activities and dispositions of data scientists that are illuminated in our framework can guide teachers, classroom experiences and curricula development. Additionally, the Data Investigation Process framework provides new opportunities for research into the teaching and learning of data science.

The Data Investigation Process framework can support teacher educators and curricula developers in designing data intensive learning experiences for teachers and assisting them in developing investigations for use in their classrooms. As mentioned previously, this framework includes or emphasizes novel activities that are absent from or implicit in previous data cycles. Our framework highlights how modern work with data requires a greater emphasis on processing data, exploring data and creating visualizations with the goal of solving a real-world problem through the use of data. While researchers have shown early evidence that secondary students *can* successfully manage, wrangle, visualize, and model large data (e.g., Kahn & Jiang, 2021; Lee & Wilkerson, 2018; Rosenberg et al., 2020), teachers still expose students to small datasets (Rubin, 2020). Data science educators and curricula developers can work to develop learning opportunities that match experiences reflected in our framework and introduce students to working with messy, complex data. This includes the appropriate technology supports since different types of technology tools are needed throughout a data investigation, and different tools (e.g., R, Tableau, CODAP) are appropriate given the educational context. Continuing to identify and develop freely accessible data tools that afford opportunities to dynamically visualize and analyze data should be a priority for software developers, as well as understanding how students reason about data using such tools.

Teachers will need learning experiences to support the development of their professional growth in teaching and learning data investigations. Many teachers, even those who specialize in mathematics and statistics, may lack the confidence, content knowledge, and pedagogical content knowledge (Lee & Harrison, 2021; Lovett & Lee, 2018) to successfully provide opportunities for data-intensive experiences and incorporate the Data Investigation Process framework in instruction. This framework provides a structure for these professional learning experiences. Additionally, the framework can support teachers in identifying and adapting tasks for their classroom, as well as serve as a reflective tool to examine their own classroom implementation.

Researchers can use practices and dispositions in the Data Investigation Process framework to guide studies of students' work with data to attend to their thinking, successes, and struggles. Many questions remain about how secondary students understand structures and measurements in large and complex data, and what is needed in curricula and teachers' knowledge for teaching data science in different content domains. Across the K–12 curriculum, we need learning progressions describing how students may develop sophistication in data investigations and what the processes and practices may look when integrated into various content domains. The practices, processes and dispositions described in the Data Investigation Process framework can be used to guide the development of assessment tools that assist researchers, curriculum developers, and teachers to evaluate the effectiveness of data science learning experiences and students' development of expertise as data investigators.

While we often think about inclusion of data into mathematics classrooms, the multidisciplinary nature of data science provides an opportunity for other disciplines to incorporate data science. There are obvious aspects of data science that naturally fit in other disciplines, such as data processing in a computer science course. The Data Investigation Process can also be incorporated into many other disciplines as a method to better understand phenomena (e.g., cultural events or climate change). Investigating data within multiple disciplines can help build disciplinary expertise, which is emphasized in data science. Our purpose is to provide a framework that can be useful in K–12+ settings where students are learning and applying data to investigate issues that may be within, or cut across, various domains and curricula strands (e.g., mathematics, statistics, sciences, social sciences, humanities, engineering). We acknowledge that for students to become data literate, exposure to data science concepts needs to occur in mathematics and statistics classrooms, as well as disciplines outside of mathematics and statistics. Although this is an opportunity, it also presents challenges. Historically,

integration of ideas across disciplines has been met with resistance, and it will take a creative approach to successfully incorporate data science into multiple disciplines within K–12.

Finally, another challenge to the incorporation of data science in K–12 classrooms and use of the Data Investigation Process framework within these classrooms is the resistance to changing standards and curriculum. With already packed standards and curriculum, historically, teachers have been hesitant to embrace more and/or novel material. Additionally, the nature of standards-based instruction and emphasis on standardized testing presents a challenge to incorporating data science into the existing school structure. As one data scientist, a former high school mathematics and statistics teacher, said in an interview,

> *... standardized testing is pretty much the exact opposite of doing what I described and what you're gonna have to do in the real-world.*

He believed, and our Data Investigation Process framework supports, that teachers should give students more experience with

> *... open ended, project type approaches that encourage students to find their own way and come up with their own solution.*

More work is needed to inform K–12 teachers, curriculum developers, teacher educators, and researchers in data science education to enact these types of experiences in K–12 settings.

## ACKNOWLEDGEMENTS

The research reported here was partially supported by the RTI University Scholars Program and funding from the National Science Foundation (DRL 1908760) awarded to North Carolina State University. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Science Foundation or RTI International. We thank several individuals for their critical contributions and feedback: Alex Dreier, Zachary Vaskalis, Greg Ray, Christine Franklin, Anna Bargagliotti, Rebecca Nichols, Donna LaLonde, Dennis Pearl, Andee Rubin, Susan Peters, Michelle Wilkerson, Bill Finzer, Tim Erickson, several middle and high school teachers, and anonymous reviewers for SERJ.

## REFERENCES

Agarwal, S. (2018, February 9). *Understanding the data science lifecycle*. Sudeep.co. https://www.sudeep.co/data-science/2018/02/09/Understanding-the-Data-Science-Lifecycle.html

Barber, M. (2018, January 14). *Data science concepts you need to know! Part 1*. Towards Data Science. https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Bargagliotti, A., Binder, W., Blakesley, L., Eusufzai, Z., Fitzpatrick, B., Ford, M., Huchting, K., Larson, S., Rovetti, R., Seal, K., & Zachariah, T. (2020a). Undergraduate learning outcomes for achieving data acumen. *Journal of Statistics Education*. https://doi.org/10.1080/10691898.2020.1776653

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020b). *Pre-K–12 Guidelines for assessment and instruction in statistics education II (GAISE 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., & Ben-Arush, T. (2014). EDA instrumented learning with TinkerPlots. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Using tools for learning mathematics and statistics* (pp. 193–208). Springer Spektrum, Wiesbden. https://doi.org/10.1007/978-3-658-03104-6_15

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

Boaler, J., & Levitt, S. (2019, October 23). Modern high school math should be about data science: Not Algebra 2. *Los Angeles Times*. https://www.latimes.com/opinion/story/2019-10-23/math-high-school-algebra-data-statistics

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 16*(3),

199–231.
Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, *50*(3), 1–42. https://doi.org/10.1145/3076253
Carmichael, I., & Marron, J. S. (2018). Data science vs. statistics: Two cultures?. *Japanese Journal of Statistics and Data Science*, *1*(1), 117–138. https://doi.org/10.1007/s42081-018-0009-3
Creswell, J. W. (2013). *Qualitative inquiry & research design: Choosing among the five approaches.* SAGE Publications.
Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26. https://doi.org/10.1111/j.1751-5823.2001.tb00477.x
Conway, D. (2010, September 30). *The data science venn diagram*. Drew Conway Data Consulting. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
D'Ignazio, C., & Klein, L. F. (2020) *Data feminism*. MIT Press.
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.* EDC 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
Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1), 44–49. https://doi.org/10.52041/serj.v16i1.213
Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, *7*(2). https://doi.org/10.5070/T572013891
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.
Geringer, S. (2014, January 6). *Data science venn diagram v2.0*. Steve's Machine Learning Blog. http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html
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)*, Berlin, Germany. https://iase-web.org/documents/papers/rt2016/Gould.pdf
Gould R., Wild C. J., Baglin J., McNamara A., Ridgway J., & McConway K. (2018). Revolutions in teaching and learning statistics: A collection of reflections. In Ben-Zvi D., Makar K., Garfield J. (Eds), *International handbook of research in statistics education* (pp. 457–472). Springer. https://doi.org/10.1007/978-3-319-66195-7_15
Graham, A. T. (1987). *Statistical investigations in the secondary school*. Cambridge University Press.
Grimshaw, S. D. (2015). A framework for infusing authentic data experiences within statistics courses. *The American Statistician*, *69*(4), 307–314. http://dx.doi.org/10.1080/00031305.2015.1081106
International Data Science in Schools Project Curriculum Team. (2019). *Curriculum frameworks for Introductory Data Science*. http://idssp.org/files/IDSSP_Frameworks_1.0.pdf.
Kahn, J., & Jiang, S. (2021). Learning with large, complex data and visualizations: Youth data wrangling in modeling family migration. *Learning, Media and Technology*, *46*(2), 128–143. https://doi.org/10.1080/17439884.2020.1826962

Kolassa, S. (2014, November, 5). The data scientist venn diagram [Comment on the blog post "Data Science without knowledge of a specific topic, is it worth pursuing as a career?"]. Stack Exchange. https://datascience.stackexchange.com/questions/2403/data-science-without-knowledge-of-a-specific-topic-is-it-worth-pursuing-as-a-ca

Lee, H. S. , & Tran, D. (2015). Framework for supporting students' approaches to statistical investigations: A guiding framework for the Teaching Statistics through Data Investigations. In *Teaching Statistics Through Data Investigation MOOC-Ed*. Friday Institute for Educational Innovation, NC State University. https://s3.amazonaws.com/fi-courses/tsdi/unit_3/SASI%20Framework.pdf

Lee, H. S., & Harrison, T. R. (2021). Trends in teaching Advanced Placement Statistics: Results from a national survey. *Journal of Statistics and Data Science Education, 29*(3), 317–327. https://doi.org/10.1080/26939169.2021.1965509

Lee, O., & Campbell, T. (2020). What science and STEM teachers can learn from COVID-19: Harnessing data science and computer science through the convergence of multiple STEM subjects. *Journal of Science Teacher Education*, *31*(8), 932–944. https://doi.org/10.1080/1046560X.2020.1814980

Lee, V. R., & Wilkerson, M. (2018). Data use by middle and secondary students in the digital age: A status report and future prospects. Commissioned Paper for the *National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6–12*. https://digitalcommons.usu.edu/itls_facpub/634/

Lesser, L. M. (2007). Critical values and transforming data: Teaching statistics with social justice. *Journal of Statistics Education*, *15*(1). https://doi.org/10.1080/10691898.2007.11889454

Lovett, J. N., & Lee, H. S. (2018). Preservice secondary mathematics teachers' statistical knowledge: A snapshot of strengths and weaknesses. *Journal of Statistics Education*, *26*(3), 214–222. https://doi.org/10.1080/10691898.2018.1496806

MacGillivray, H., & Pereira-Mendoza, L. (2011). Teaching statistical thinking through investigative projects. In C. Batanero, G. Burrill, and C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 109–120). Springer. https://doi.org/10.1007/978-94-007-1131-0_14

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistics Review, 65*(2), 123–165. https://doi.org/10.2307/1403333

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. The National Academies Press. https://doi.org/10.17226/25104

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Author.

Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For states, by states*. National Academies Press. https://www.nextgenscience.org/standards/standards

National Governors Association Center for Best Practice & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. http://www.corestandards.org/Math/

National Research Council. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition.* The National Academies Press. https://doi.org/10.17226/9853

Pangrazio, L., & Selwyn, N. (2021). Towards a school-based "critical data education." *Pedagogy, Culture & Society 29*(3), 431–44. https://doi.org/10.1080/14681366.2020.1747527

Rosenberg, J., Edwards, A., & Chen, B. (2020). Getting messy with data. *The Science Teacher, 87*(5), 30–34. https://www.nsta.org/science-teacher/science-teacher-january-2020/getting-messy-data

Rubin, A. (2020). Learning to reason with data: How did we get here and what do we know?. *Journal of the Learning Sciences*, *29*(1), 154–164. https://doi.org/10.1080/10508406.2019.1705665

Saltz, J. S. (2020, May 29). *CRISP-DM for data science teams: 5 actions to consider.* Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-for-data-science-teams-5-actions-to-consider

Saltz, J. S., Shamshurin, I., & Connors, C. (2017). A framework for describing big data projects. In W. Abramowicz, R. Alt, & B. Franczyk (Eds), *Business Information Systems Workshops. BIS 2016. Lecture Notes in Business Information Processing,* Vol 263. Springer. https://doi.org/10.1007/978-3-319-52464-1_17

Saltz, J. S., & Hotz, N. (2020). Identifying the most common frameworks data science teams use to structure and coordinate their projects. *Proceedings of the 2020 IEEE International Conference on Big Data* (pp. 2038–2042). https://doi.org/10.1109/BigData50022.2020.9377813

Tierney, B. (2012, June 13). *Data science is multidisciplinary*. Oralytics. https://oralytics.com/2012/06/13/data-science-is-multidisciplinary/

Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley.

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

HOLLYLYNNE S. LEE
502C Poe Hall
Campus Box 7801
NC State University
Raleigh, NC 27695
USA

## APPENDIX

### Thinking Like a Data Scientist Interview Protocol

**Part 1: Your Job and Working with Data**

1. Please briefly describe your job. (follow-up: what is your official job title?)
2. What preparation or education did you need to be able to work with data in the ways you do?
3. What are some of the key skills you need in your daily work?
4. What are some of the key tools you use to do your work?
5. I'd like to know a bit more about the ways you work with data. Can you tell me about a current or recent project you have been engaged in? (Follow-up questions may include what is your role in the project? Are there other team members on the project? What is their role?
6. Thinking about this project, what were some aspects of the project that were particularly challenging? What made this challenging? How did you and the team overcome those challenges?
7. What were some of the more simplistic aspects of the project that were relatively easy to achieve. Why were they easier than others? Do you have experiences in your career where you learned something that then made something more simplistic, or expected in future projects?
8. Over the course of a project, how would you describe the ways you interact with the data? Does your relationship with the data change over the course of a project? [follow-up Were there ever strong emotional responses you had at different points in a project?]

**Part 2: Explaining Data Science to Others**

1. If you were going to describe what a data scientist does to a fifth grader, how would you explain it? (follow-up: I noticed you characterized data science as [fill in with something they said]. Why do you think that characterization is helpful to explain what a data scientist is?]
2. How would you explain the field of data science to a high school student who is considering different career pathways? (follow-ups: I noticed you characterized data science as [fill in with something they said]. Why do you think that characterization is helpful to explain what a data scientist is? Why was your explanation similar or different for a high schooler as opposed to the fifth grader? What advice would you have for a high schooler who may be interested in a data-intensive career?)
3. How do you explain what you do in your job to an adult you meet in a social setting? (follow-up: what do you think is important to communicate to all three of these audiences? why?)
4. What would you tell middle or high school teachers to include in their curriculum or instruction to get students experienced with data and prepare them for pursuing a data-intensive career?

**Part 3: Frameworks for Data Science**

In this portion of the interview, several images and descriptors that have been used in publications to describe Data Science will be presented. For each image, I will ask you 2 questions:

1. Comment on the information presented and if you think this accurately depicts or represents what you conceive of as data science and how you use data in your career.
2. What would you change or add to the diagram or list of characteristics?

[Note: Images and descriptions are presented on separate slides and ordered based on the specific job of the interviewee and the interviewer's perception of which images and descriptors may be best suited for the interviewee to comment on.]

The following is a list of URLs where the publicly posted diagrams used in the interviews were retrieved in September 2018 from:

https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

https://www.kdnuggets.com/2017/03/data-science-data-scientist-do.html

https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining
https://commons.wikimedia.org/wiki/File:DataScienceDisciplines.png
http://www.kiwidatascience.com/
http://www.cellstrat.com/2018/05/25/data-science-deconstructed-2-of-3/
https://medium.com/@YvesMulkers/how-to-become-data-scientist-f2b5b3d2a73a
https://searchenterpriseai.techtarget.com/definition/data-scientist